18734: Foundations of Privacy

Fairness in Classification

Giulia Fanti

With many slides from Moritz Hardt,
Anupam Datta

Fall 2019

Administrative

- Nice talk on Thursday: Farinaz Koushanfar (UCSD)
  - Latest in privacy-preserving machine learning over encrypted data
  - Thursday 10/10 @ 3:30 pm EST/12:30 pm PT in HH1107 (Pittsburgh), Room 1065 (SV)
  - Free food in PIT!

- Mid-semester presentations
  - Wednesday, Oct. 30
  - Monday, Nov. 4
  - Guidelines on Canvas (rubric + points)
  - Sign up link here

- This Friday (Oct. 25): Day for Community Engagement
  - No recitation
  - Sruti will hold her OH on Friday from 3-4pm ET
  - My OH are by appointment this week

In-Class Quiz

- On Canvas

Big Data: Seizing Opportunities, Preserving Values ~ 2014

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

Many notions of “fairness” in CS

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



Fairness in Classification

Advertising

Education

Financial aid

Banking

Health Care

Insurance

Taxation

many more...

Concern: Discrimination

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

Examples

- Word embeddings
  - Important trend in NLP
  - Map word -> vector
  - Related words have similar vectors
  - E.g.:

$$v(\text{king}) - v(\text{man}) = v(\text{queen}) - v(\text{woman})$$

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Bolukbasi et al, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”, NeurIPS 2019

Overview

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

Discrimination arises even when nobody's *evil*

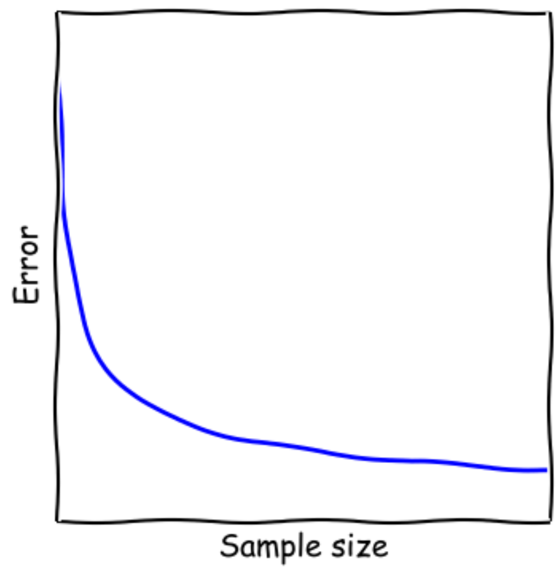
- Google+ tries to classify real vs fake names
- Fairness problem:
  - Most training examples standard white American names: John, Jennifer, Peter, Jacob, ...
  - Ethnic names often unique, much fewer training examples

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

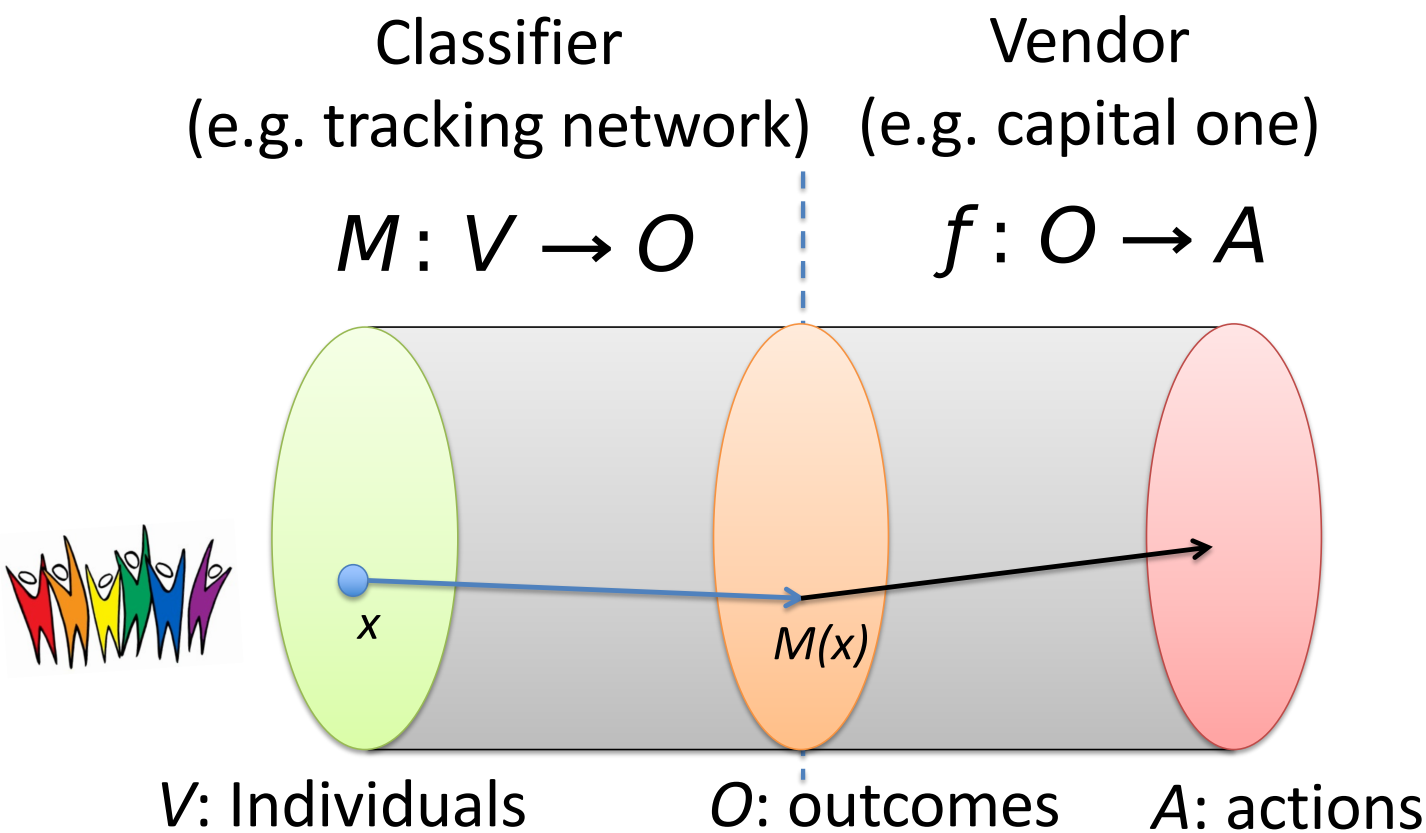
Credit Application

User visits `capitalone.com`

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

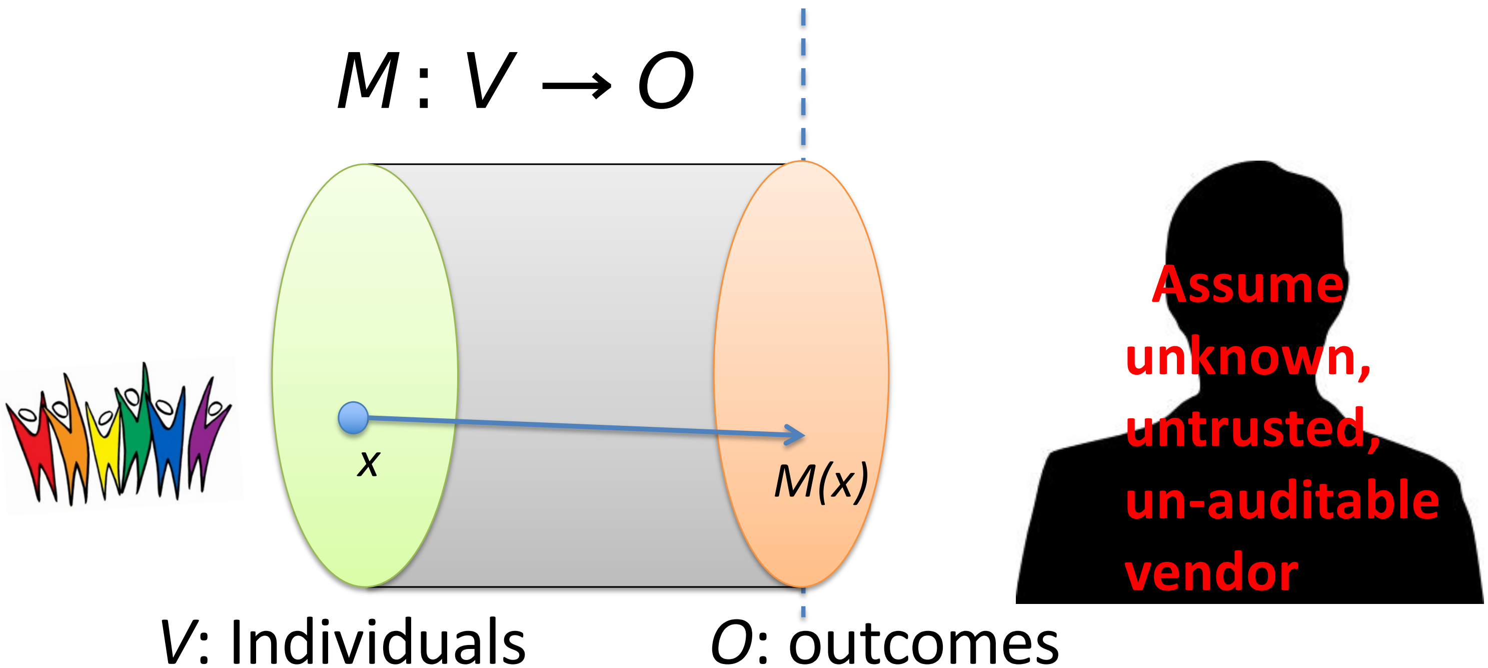
Concern: *Steering* minorities into higher rates (illegal)

WSJ 2010


Classifier
(e.g. tracking network)
Vendor
(e.g. capital one)
$M: V \to O$
$f: O \to A$
x
$M(x)$
V: Individuals
O: outcomes
A: actions

Goal:

Achieve Fairness in the classification step



What kinds of events do we want to prevent in our definition?

- Blatant discrimination
- Discrimination based on redundant encoding
- Discrimination against portion of population with higher fraction of protected individuals
- Self-fulfilling prophecy
- Reverse tokenism

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"

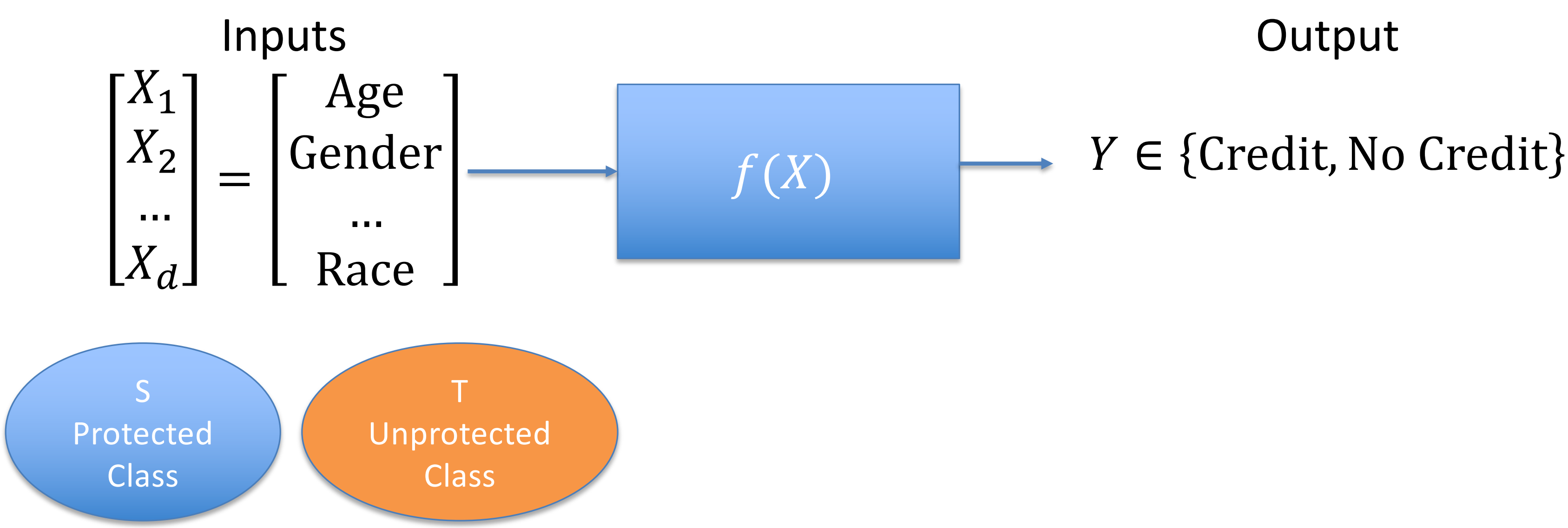
Cynthia Dwork & Deirdre K. Mulligan. Stanford Law Review.

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."

Group Exercise

- With your partner, come up with a mathematical definition of a fair classifier
- I.e., what properties should f exhibit to be considered **fair**?



Second attempt…

Statistical Parity (Group Fairness)

Equalize two groups S, T at the level of outcomes

- E.g. S = minority, $T = S^c$
- Outcome o

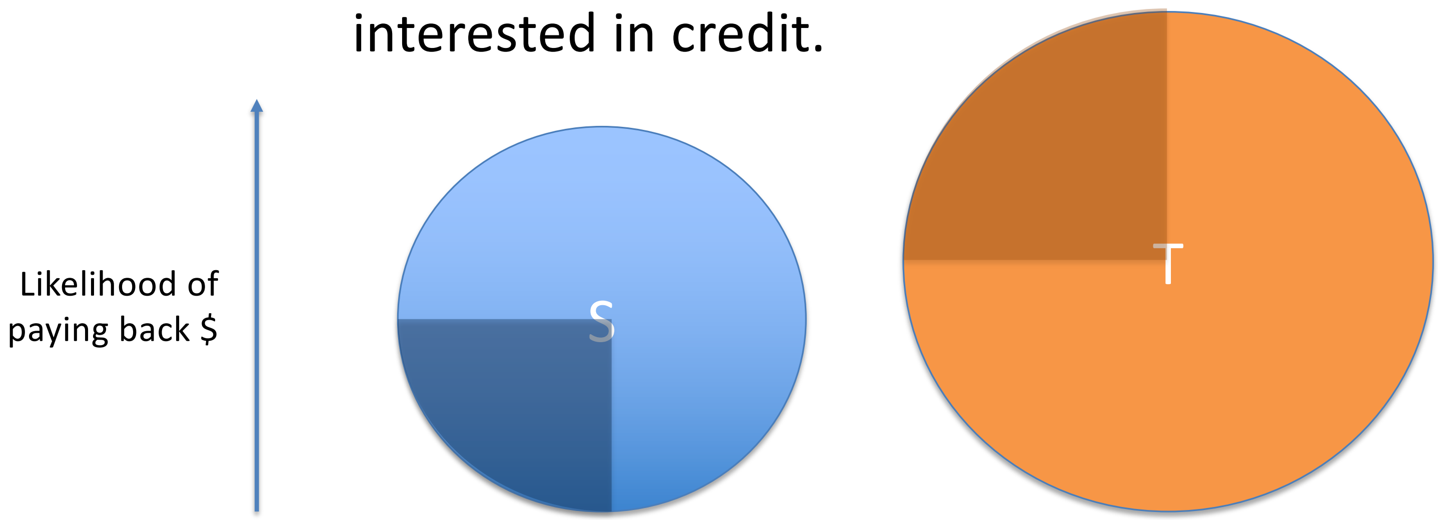
$$\mathrm{P}[o \mid S] = \mathrm{P}[o \mid T]$$

"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

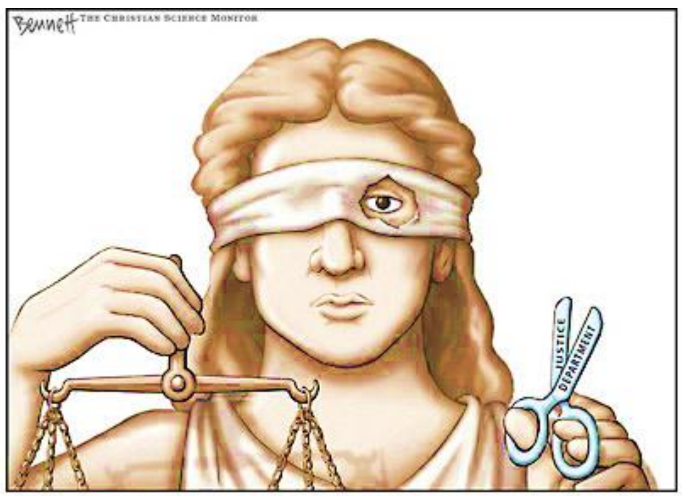
- **Self-fulfilling prophecy**
  - Give credit offers to S persons deemed least credit-worthy.
  - Give credit offers to those in S who are not interested in credit.



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

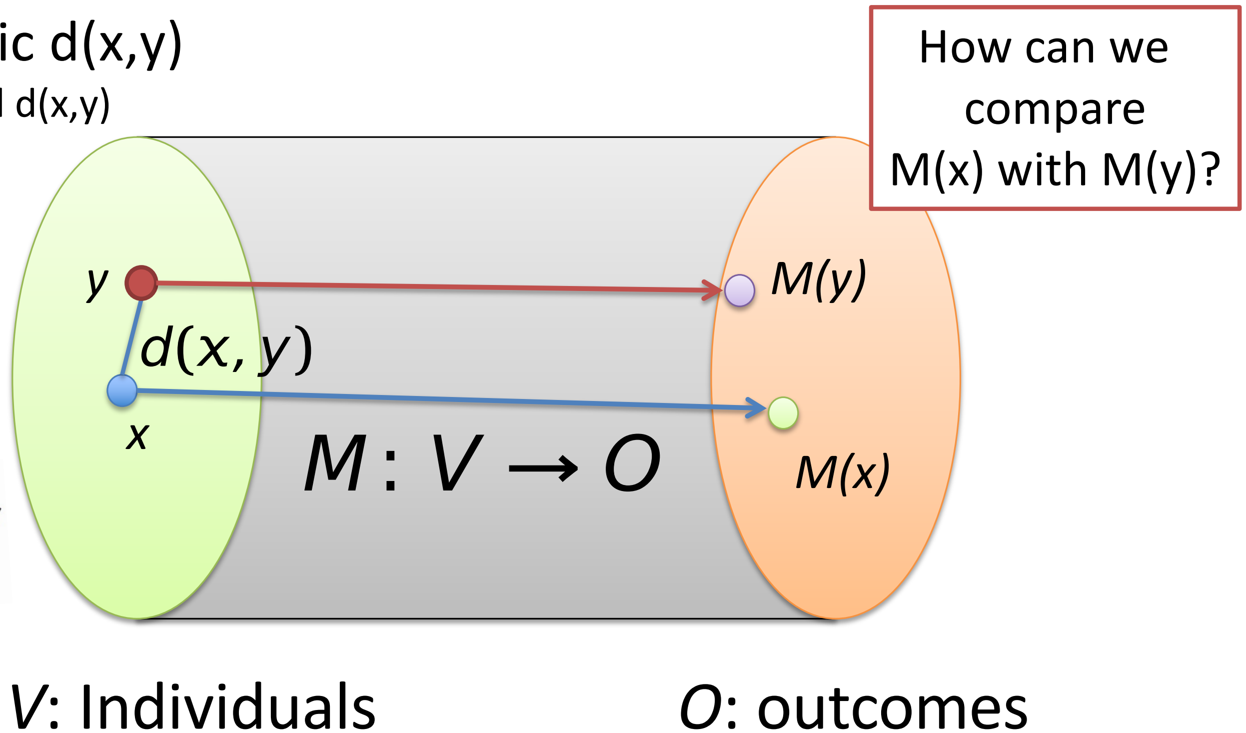
- Assume ***task-specific similarity metric***
  - Extent to which two individuals are similar w.r.t. the classification task at hand
- Ideally captures *ground truth*
  - Or, society's best approximation
- Open to public discussion, refinement
  - In the spirit of Rawls
- Typically unrelated to classification!

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

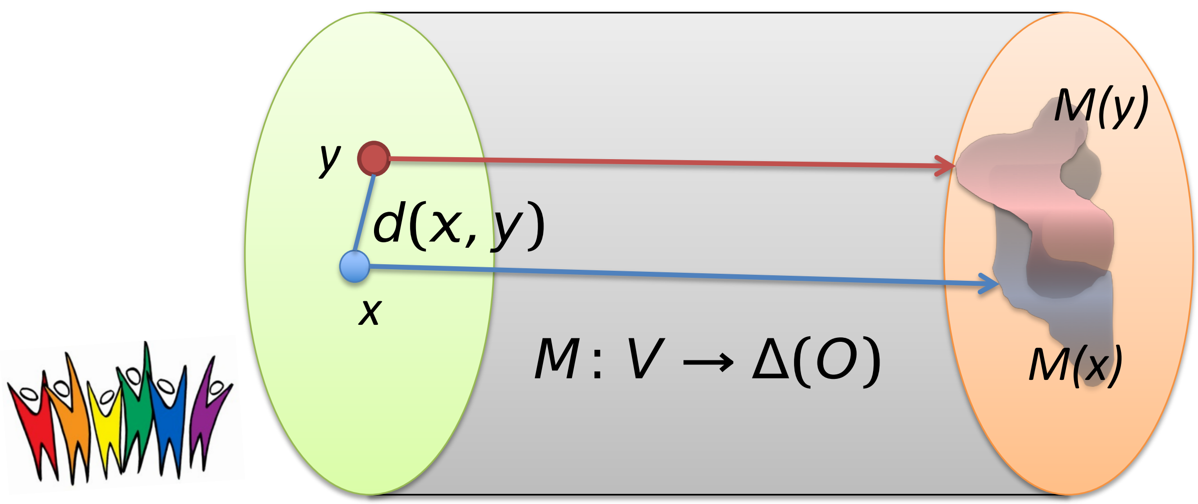
How to formalize this?

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



Distributional outcomes

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



Distance over distributions!

V: Individuals

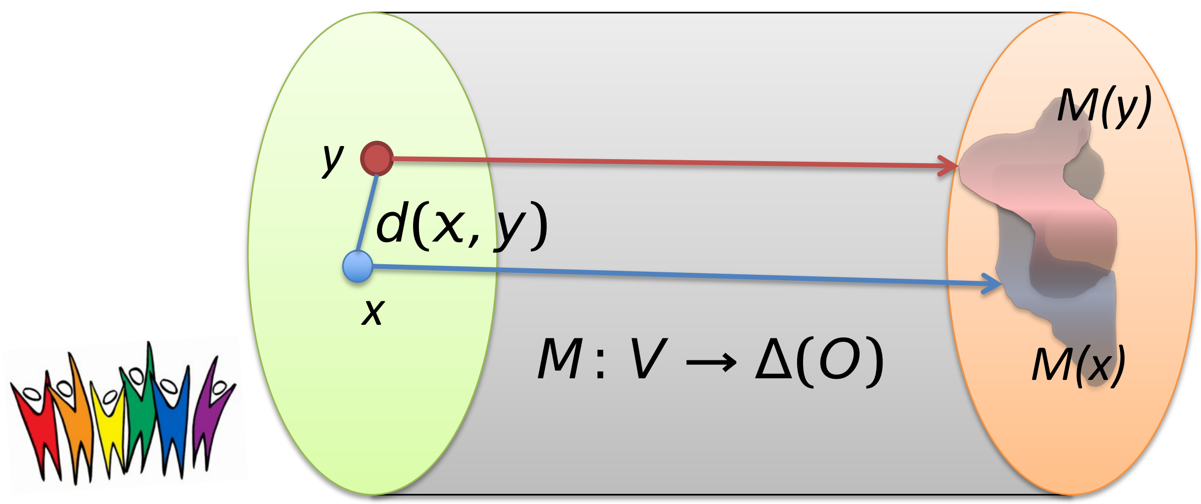
O: outcomes

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

Fairness Definition

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

This talk: Statistical distance in [0,1]



V: Individuals O: outcomes

Statistical Distance (Total Variation Distance)

Let M_x and M_y denote probability measures on a finite domain A. The statistical distance (or total variation distance) between M_x and M_y is denoted by

$$D_{TV}(M_x, M_y) = \frac{1}{2}\sum_{a \in A} \left|M_x(a) - M_y(a)\right|$$

Statistical Distance (Total Variation Distance)

Let M_x and M_y denote probability measures on a finite domain A. The statistical distance (or total variation distance) between M_x and M_y is denoted by

$$D_{TV}(M_x, M_y) = \frac{1}{2}\sum_{a \in A} \left| M_x(a) - M_y(a) \right|$$

Q: What is the minimum TV distance between two distributions?

A: 0, achieved when both are equal

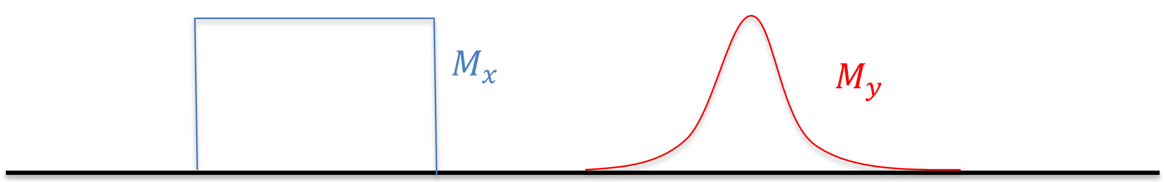
Statistical Distance (Total Variation Distance)

Let M_x and M_y denote probability measures on a finite domain A. The statistical distance (or total variation distance) between M_x and M_y is denoted by

$$D_{TV}(M_x, M_y) = \frac{1}{2}\sum_{a\in A}\left|M_x(a) - M_y(a)\right|$$

Q: What is the maximum TV distance between two distributions?

A: 1, achieved when both are disjoint

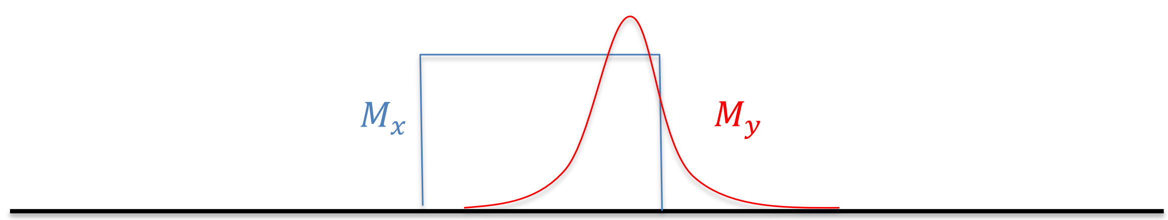


Statistical Distance (Total Variation Distance)

Let M_x and M_y denote probability measures on a finite domain A. The statistical distance (or total variation distance) between M_x and M_y is denoted by

$$D_{TV}(M_x, M_y) = \frac{1}{2}\sum_{a \in A} \left| M_x(a) - M_y(a) \right|$$

Example of intermediate TV distance between two distributions



Existence Proof

Individual Fairness Definition

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

There exists a classifier that satisfies the Lipschitz condition

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

- Are we done?

But… how do we ensure utility?

Utility Maximization

Vendor can specify **arbitrary utility function**

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

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

E.g. Accuracy of classifier

Maximize vendor's expected utility subject to Lipschitz condition

Maximize utility

$$\max_{M_x} \mathbb{E}_{x\sim V}\mathbb{E}_{o\sim M_x}\left[U(x,o)\right]$$

Subject to fairness constraint

$$\text{s.t. } \left\|M_x - M_y\right\| \leq d(x,y) \;\; \forall\, x,y \in V$$

Claim: This optimization is a linear program under
TV (statistical) distance over distributions

$$\max_{M_x} \mathbb{E}_{x\sim V}\mathbb{E}_{o\sim M_x}\left[U(x,o)\right]$$

$$\text{s.t. } \left\|M_x - M_y\right\| \leq d(x,y) \ \ \forall\, x,y \in V$$

- Need to show 2 things:
  - Objective function is linear in probability mass vector M_x
  - Constraints are all linear

- Try to show both
  - You can assume V is a set with $|V|$ discrete items

- Why do we care? Linear programs can be solved efficiently!
  - I.e., in polynomial time in the problem dimension

What's the takeaway?

- We can efficiently enforce individual fairness while maximizing overall utility!

- What about our initial notion of group fairness?