18734: Foundations of Privacy

# Discrimination and Fairness in Classification

Anupam Datta

Fall 2016

# 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

# Discrimination notions in US law

- Disparate treatment
  - Special case: formal disparate treatment in which the protected feature (e.g., race, gender) is directly used to make a decision (e.g., about employment, housing, credit)
  - Formally, protected feature has causal effect on outcome (Datta et al. AdFisher paper)
  - Example: Gender has causal effect on advertising of job-related ads

# Discrimination notions in US law

- Disparate impact
  - The protected feature (e.g., race, gender) is associated with the decision (e.g., about employment, housing, credit) [see Feldman et al. Disparate Impact paper]
  - Example: Propublica finding of association between race and recidivism score of the COMPAS scoring system
  - Association not problematic if caused by a correlate whose use is a "business necessity"

# Discrimination arises even when nobody's *evil*

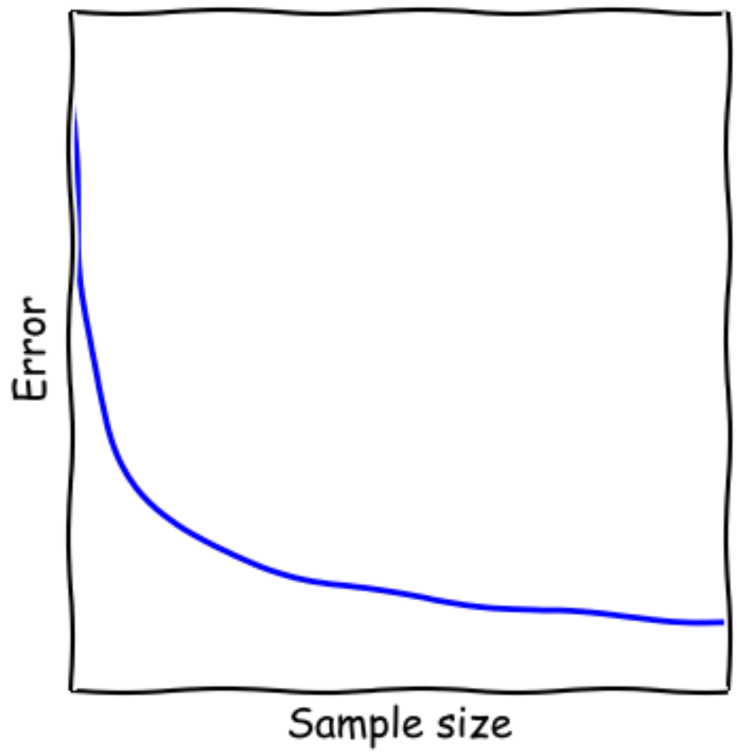
- Google+ tries to classify real vs fake names
- Fairness problem:
  - Most training examples standard white American names: John, Jennifer, Peter, Jacob, ...
  - Ethnic names often unique, much fewer training examples

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)"*

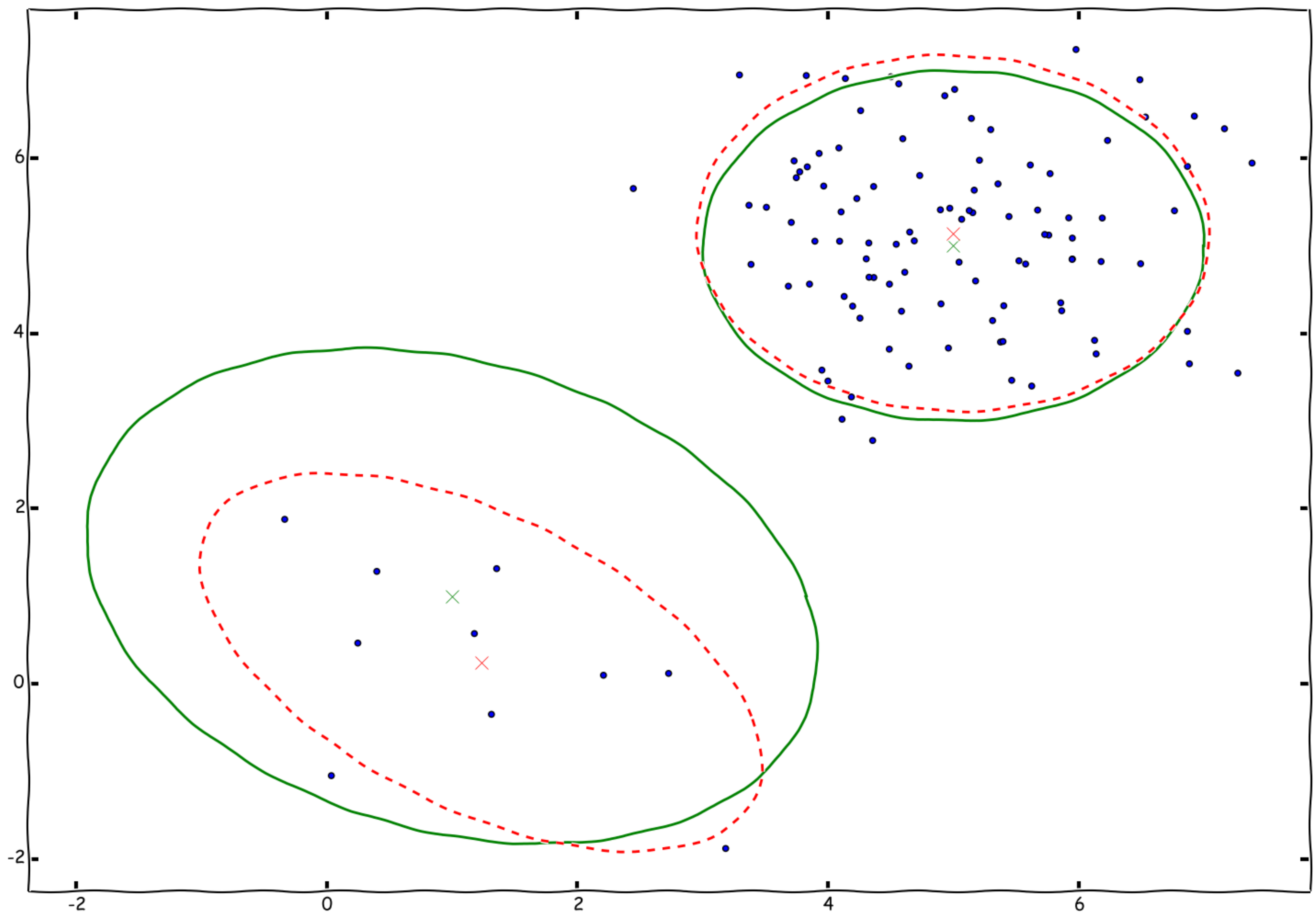
- Katya Casio. Google Product Forums.

# Error vs sample size




Sample Size Disparity:

In a heterogeneous population, smaller groups face larger error



-2
0
2
4
6

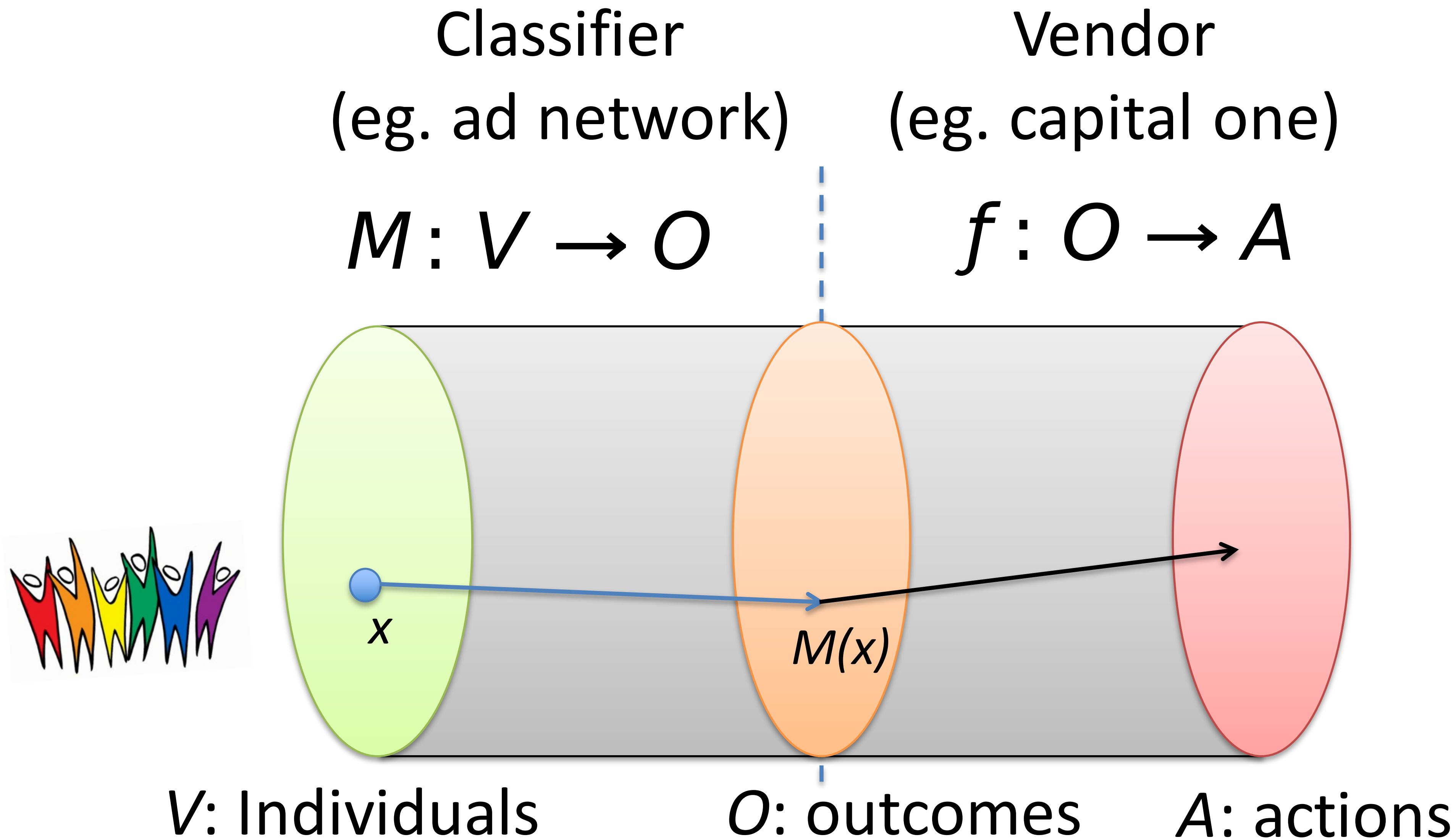
# Credit Application


User visits `capitalone.com`

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

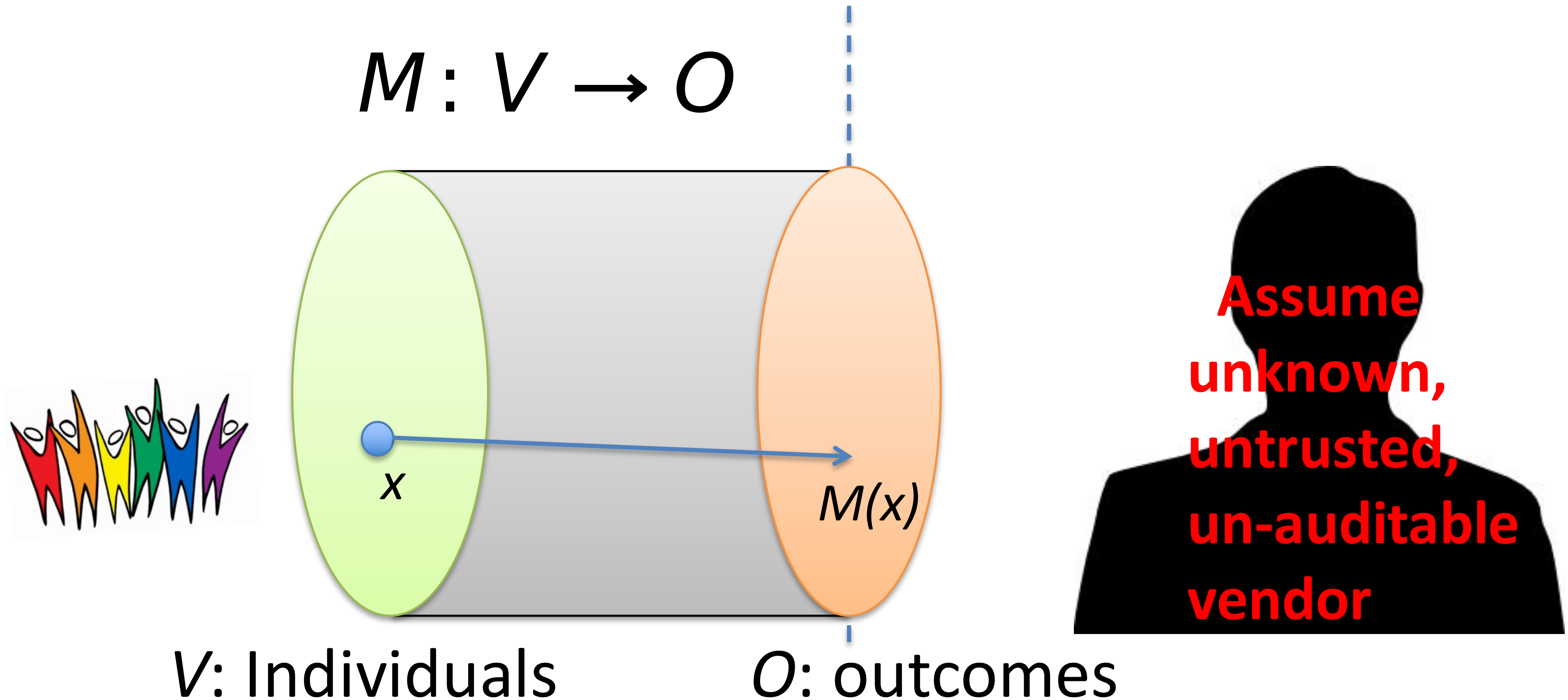
**Concern:** *Steering* minorities into higher rates (illegal)

WSJ 2010


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

## 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’!”*

*Useful to avoid formal disparate treatment*

# 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

# 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."

Useful to prevent disparate impact

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

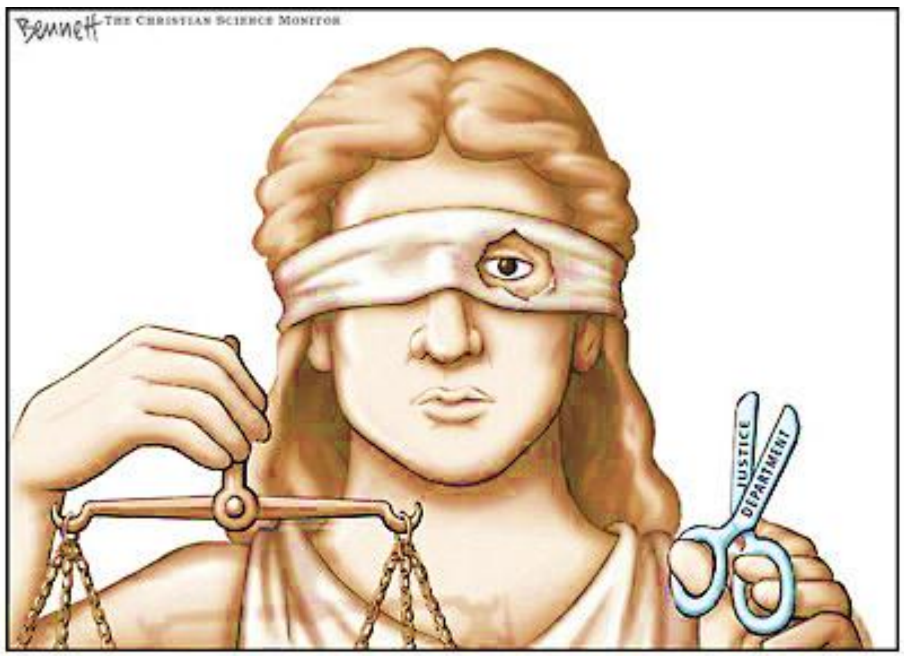
- Sometimes desirable, but can be abused

- **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”



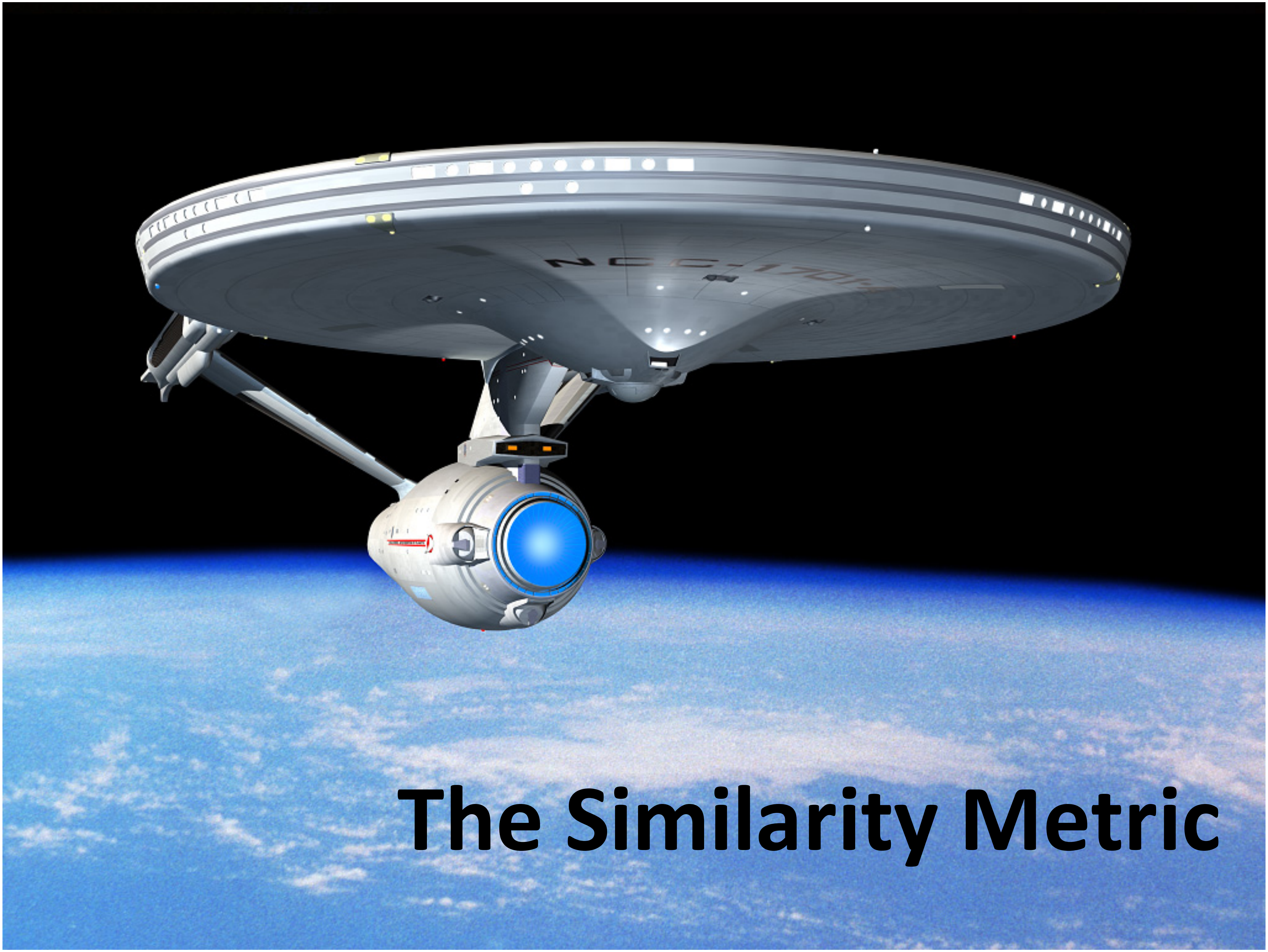
# Individual Fairness Approach

# 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

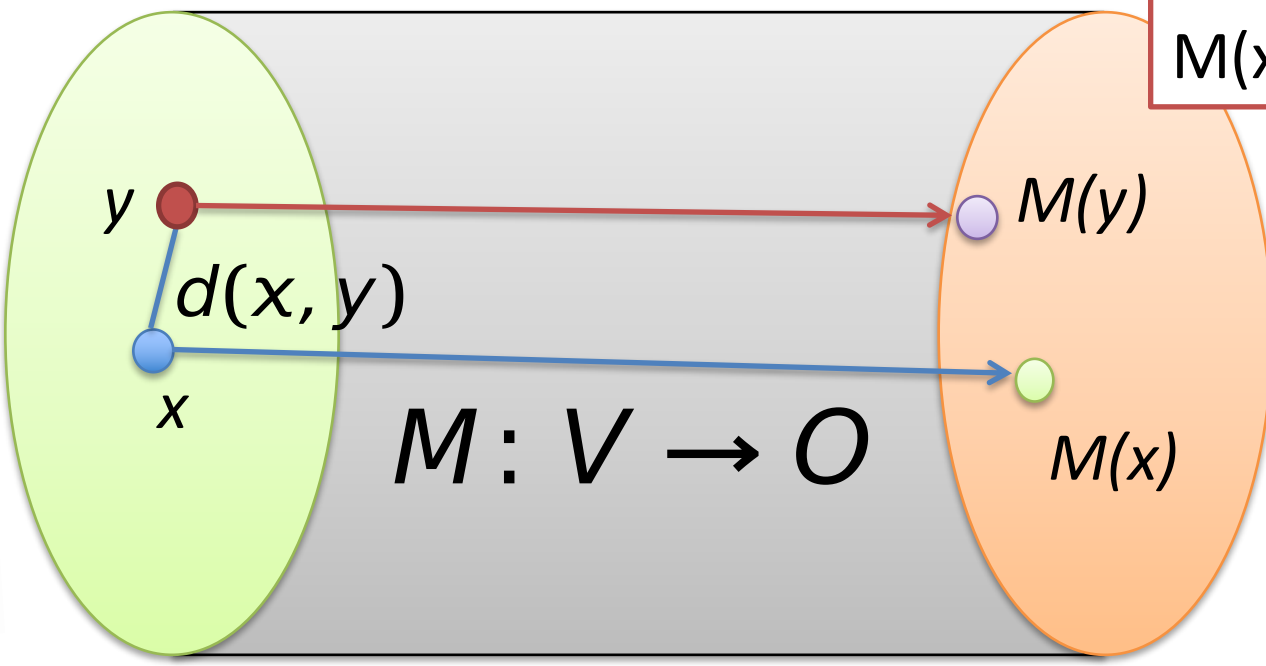
# Biggest weakness of theory

How do we construct a similarity metric?

# 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)?



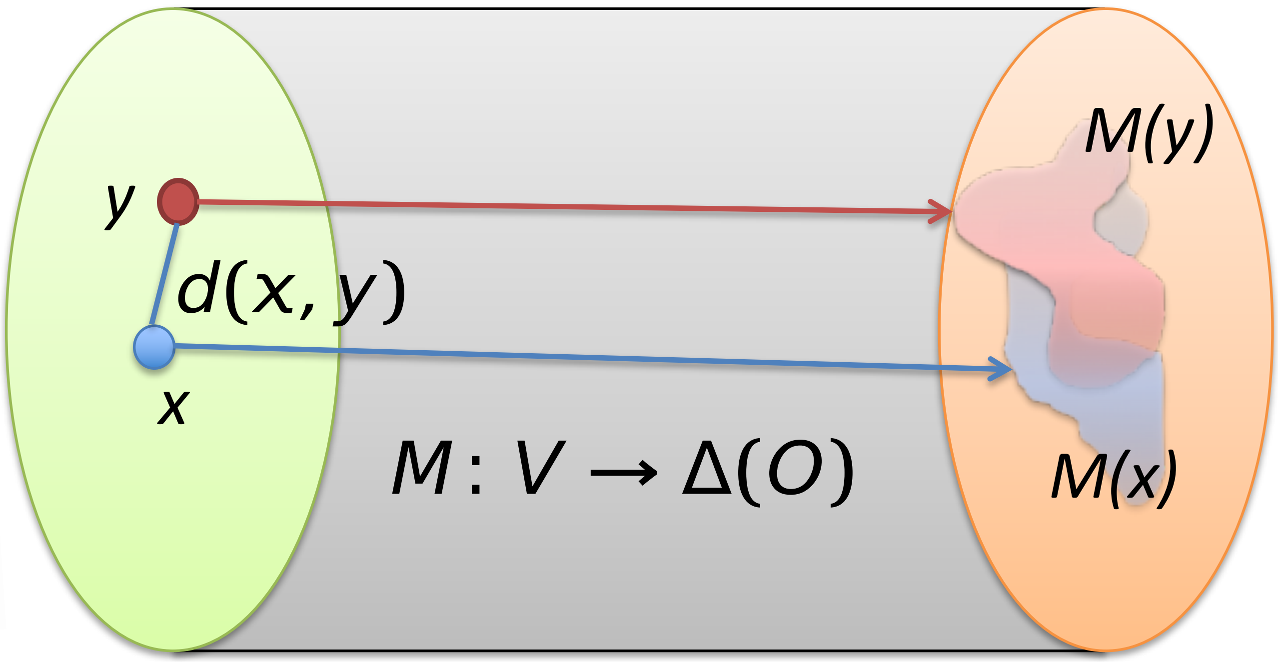

*V*: Individuals

*O*: outcomes

# Distributional outcomes

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

Statistical distance!



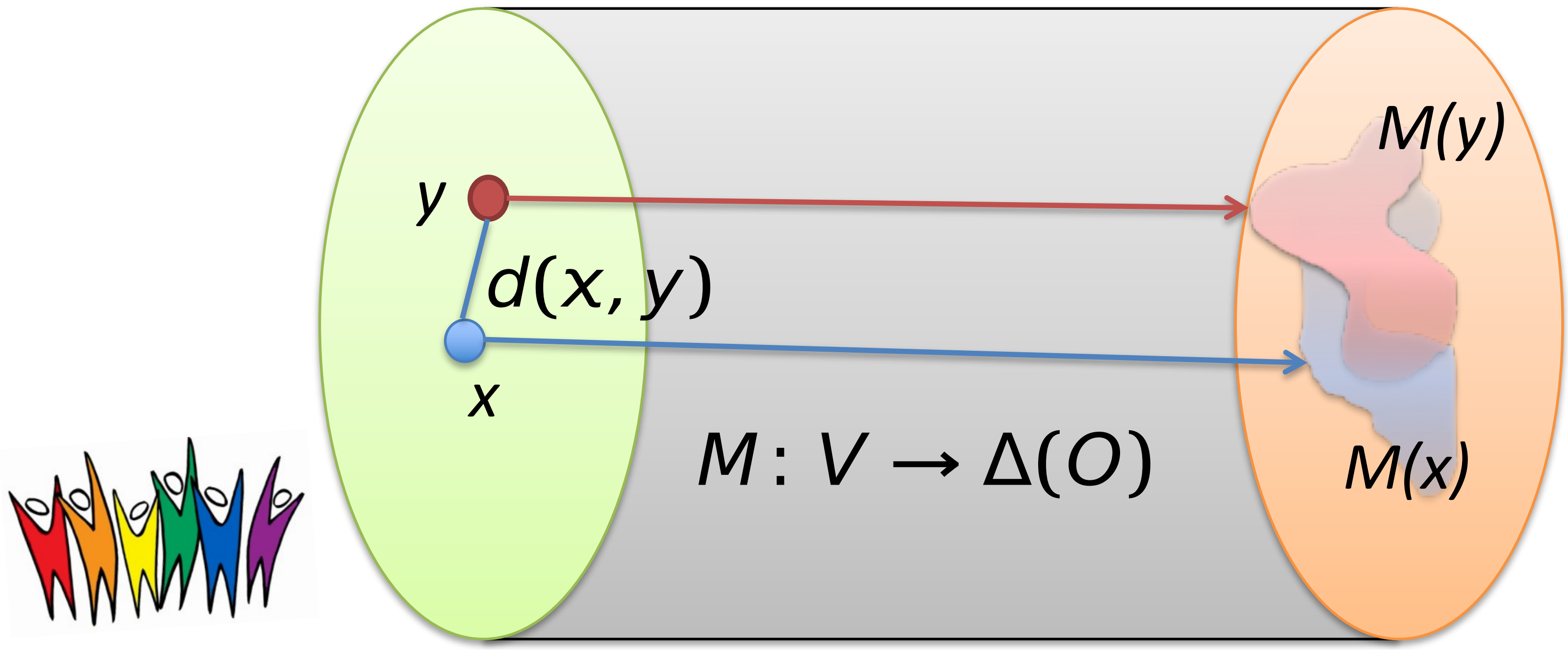

*V*: Individuals

*O*: outcomes

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

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

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

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

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

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

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**

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

*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
  - {M(x)}(x in V} are only variables to be computed
- Lipschitz condition is linear when using statistical distance
- 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

# The Story So Far…

- Group fairness
- Individual fairness
- Group fairness does not imply individual fairness

- When does individual fairness imply group fairness?

# Statistical Parity (Group Fairness)

Equalize two groups S, T at the level of outcomes

– E.g. $S$ = minority, $T = S^c$

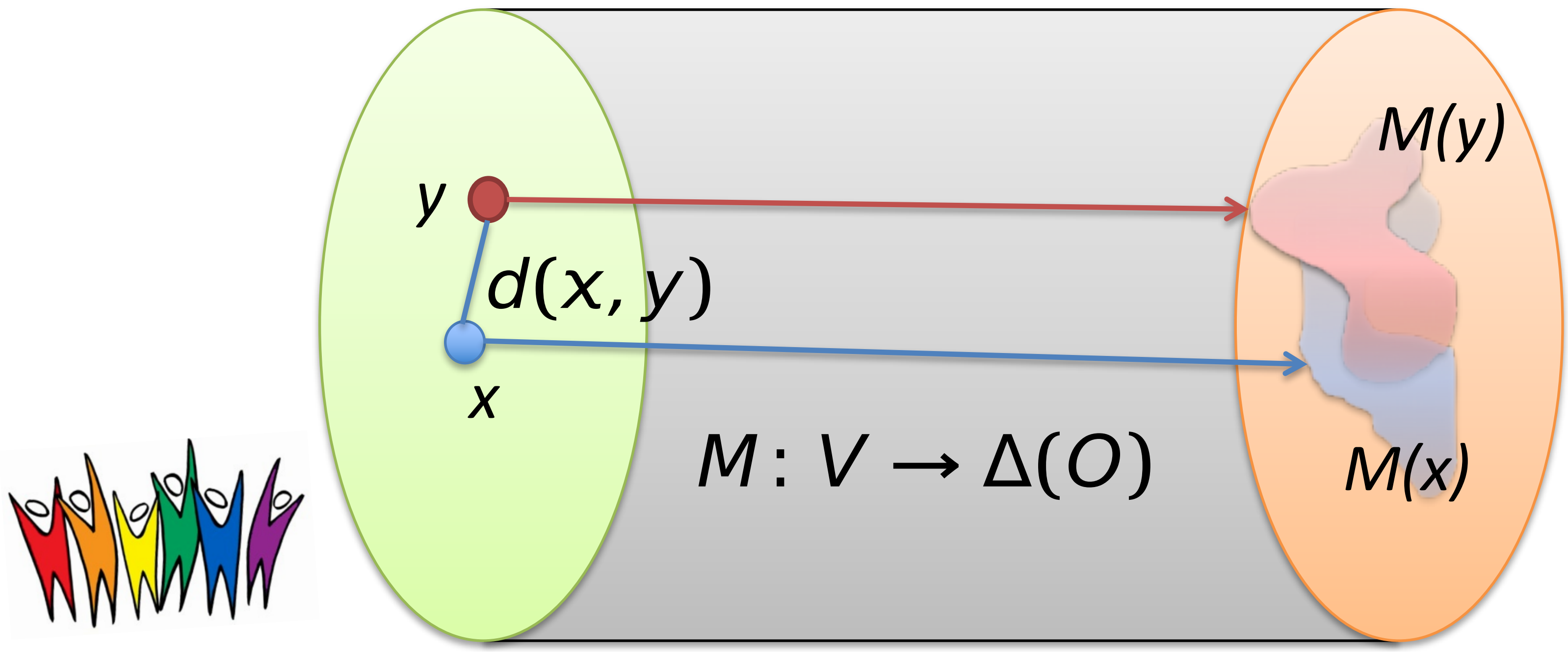
$$\Pr[\text{outcome } o \mid S] = \Pr[\text{outcome } o \mid T]$$

"Fraction of people in S getting credit same as in T."

# Individual Fairness

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

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




*V*: Individuals

*O*: outcomes

# When does Individual Fairness imply Group Fairness?

Suppose we enforce a metric *d.*

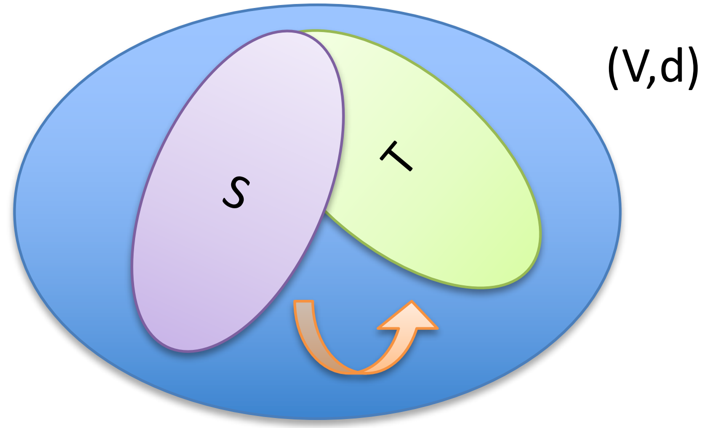
**Question:** Which *groups of individuals* receive (approximately) equal outcomes?

**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




$$
\begin{aligned}
\sigma_{\mathrm{EM}}(S,T) \stackrel{\mathrm{def}}{=} \quad \min \quad & \sum_{x,y \in V} h(x,y)\sigma(x,y) \\
\text{subject to} \quad & \sum_{y \in V} h(x,y) = S(x) \\
& \sum_{y \in V} h(y,x) = T(x) \\
& h(x,y) \geq 0
\end{aligned}
$$

$$
\begin{aligned}
\sigma_{\mathrm{EM}}(S,T) \stackrel{\mathrm{def}}{=} \quad \min \quad & \sum_{x,y\in V} h(x,y)\sigma(x,y) \\
\text{subject to} \quad & \sum_{y\in V} h(x,y) = S(x) \\
& \sum_{y\in V} h(y,x) = T(x) \\
& h(x,y) \geq 0
\end{aligned}
$$

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)

# The Story So Far…

- Group fairness
- Individual fairness
- Group fairness does not imply individual fairness
- Individual fairness implies group fairness if earthmover distance small

# 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

- Disparate treatment
  - Protected attribute has causal effect on decision
  - Datta et al. AdFisher paper
- Disparate Impact
  - Protected attribute associated with decision
  - Feldman et al. Disparate Impact paper
- Individual fairness
  - "Similar" individuals treated similarly
  - Dwork et al. Fairness through Awareness paper

# Questions?

# Acknowledgement

- Most of the slides are from Moritz Hardt. Slides 4, 5, 25, 47 are mine as are various comments about related work.