Lecture 10: Fairness, Ethics, & Law in ML

Nov 23

Prof. Nicolas Papernot

@NicolasPapernot

Fairness through awareness

Dwork et al.

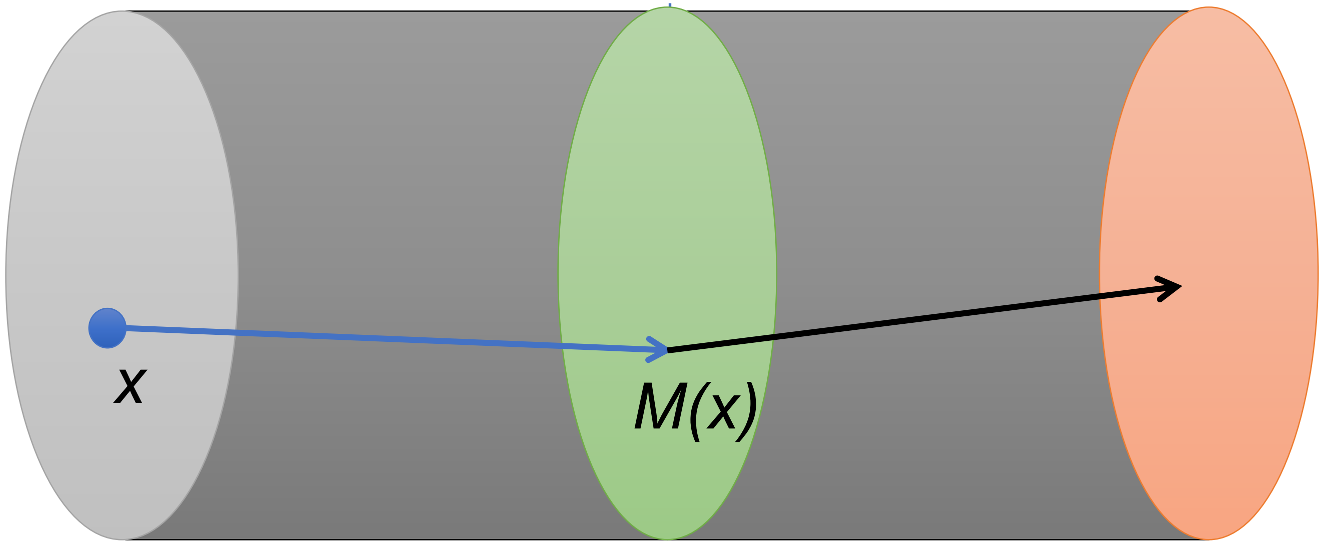
Slides adapted from Anupam Datta, Moritz Hardt, Omer Reingold

Classifier
(eg. ad network)

$M : V \rightarrow O$

Vendor
(eg.site showing ad)

$f : O \rightarrow A$



V: Individuals

O: outcomes

A: actions

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 `truckdriversunited.com`

... 90% chance of being a truck driver

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."

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 offers same as in T."

Not strong enough as a notion of fairness

- Sometimes desirable, but can be abused

Malicious vendor wants to sell a high-fee exclusive credit card only to people who have purple skin, not people with green skin

- Target 500 high income people with purple skin

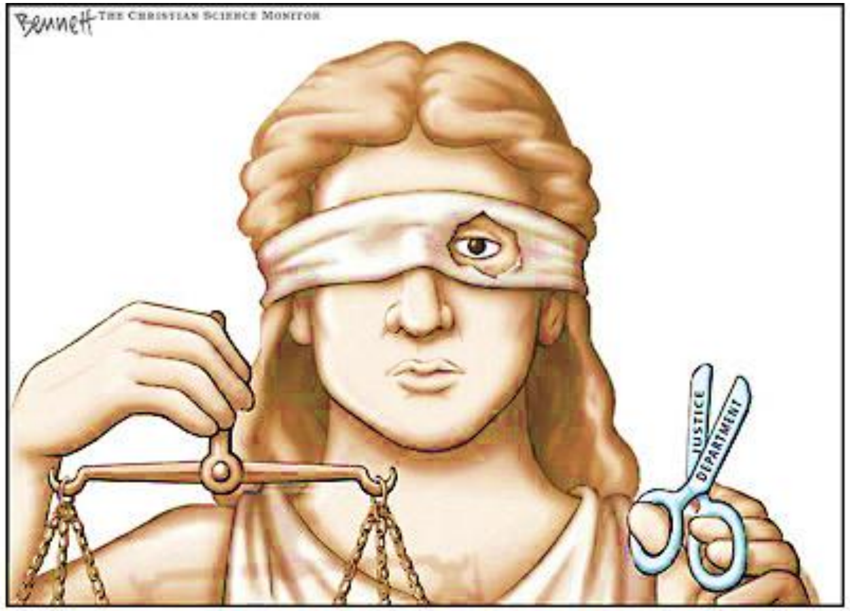
- Target 500 low income people with green skin

Yet, group fairness between purple and green skin

Lesson: Fairness is *task-specific*

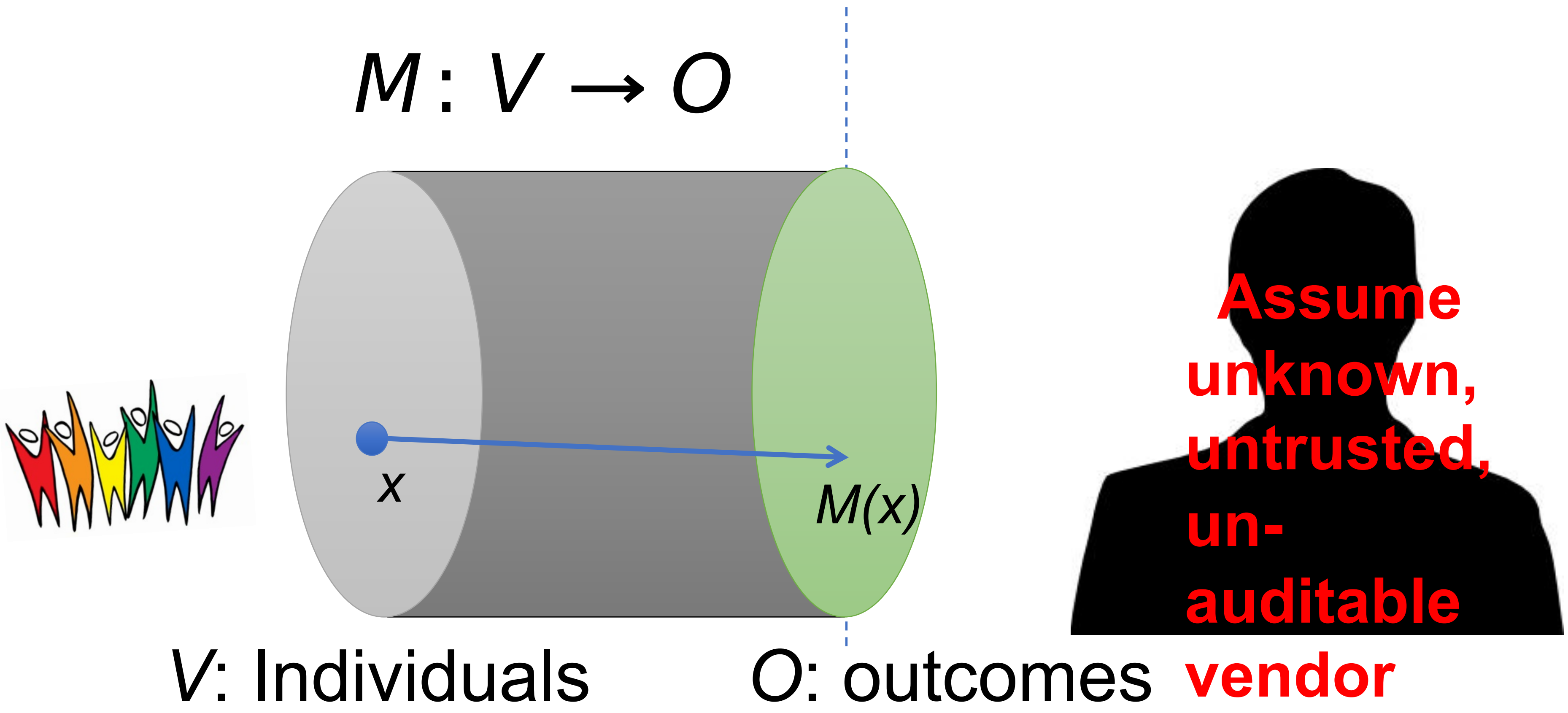
Fairness requires understanding of classification task and protected groups

“Awareness”



Goal:

Achieve Fairness in the classification step



Individual Fairness

Treat *similar* individuals *similarly*

Similar for the purpose of the classification task

Similar distribution over outcomes

- 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

Examples:

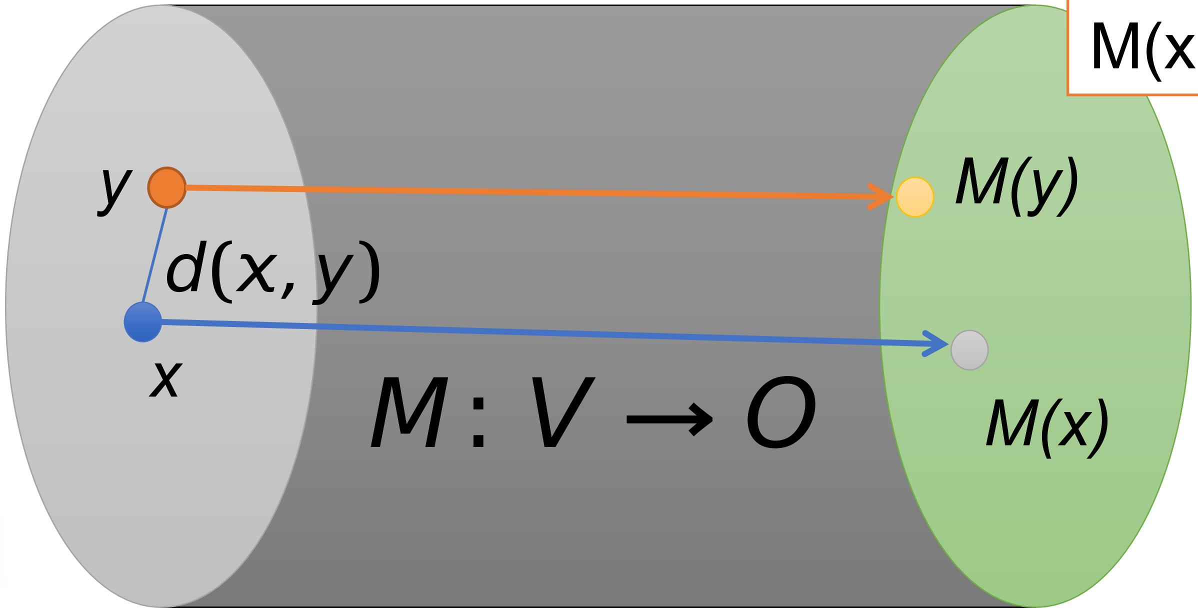
- Financial/insurance risk metrics
  - Already widely used (though secret)
- **AALIM health care metric**
  - health metric for treating similar patients similarly

UNIVERSITY OF TORONTO VECTOR INSTITUTE

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)?

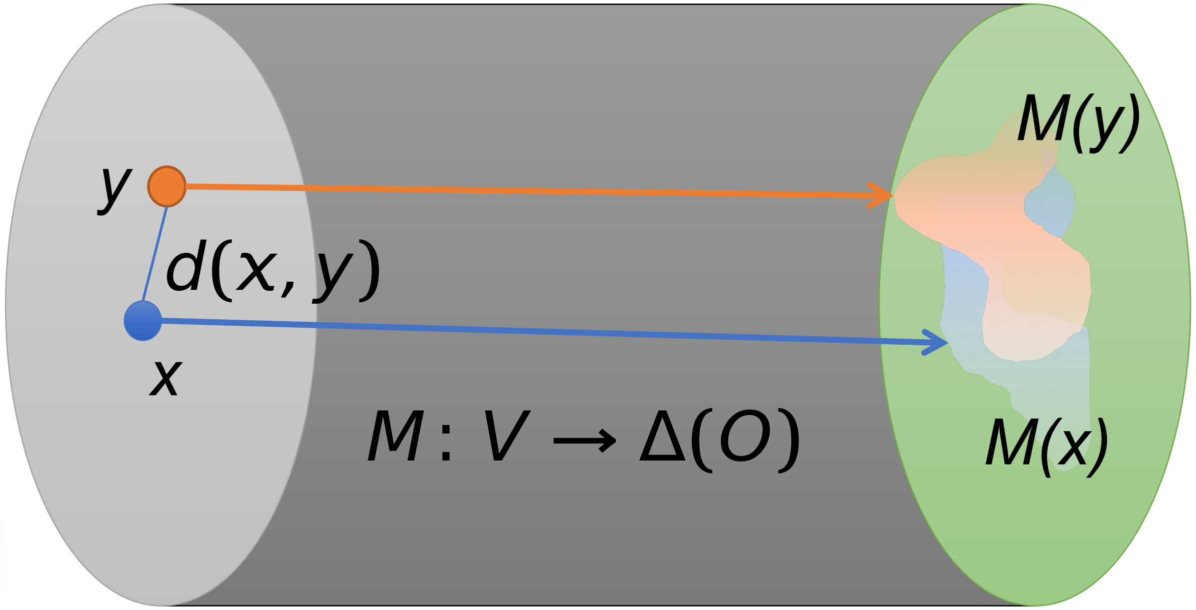


V: Individuals

O: outcomes

Distributional outcomes

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



Statistical distance!

V: Individuals

O: outcomes

Example: statistical distance

- Statistical distance: $d(P,Q) = \frac{1}{2}\sum_{o \in O}|P(o) - Q(o)|$

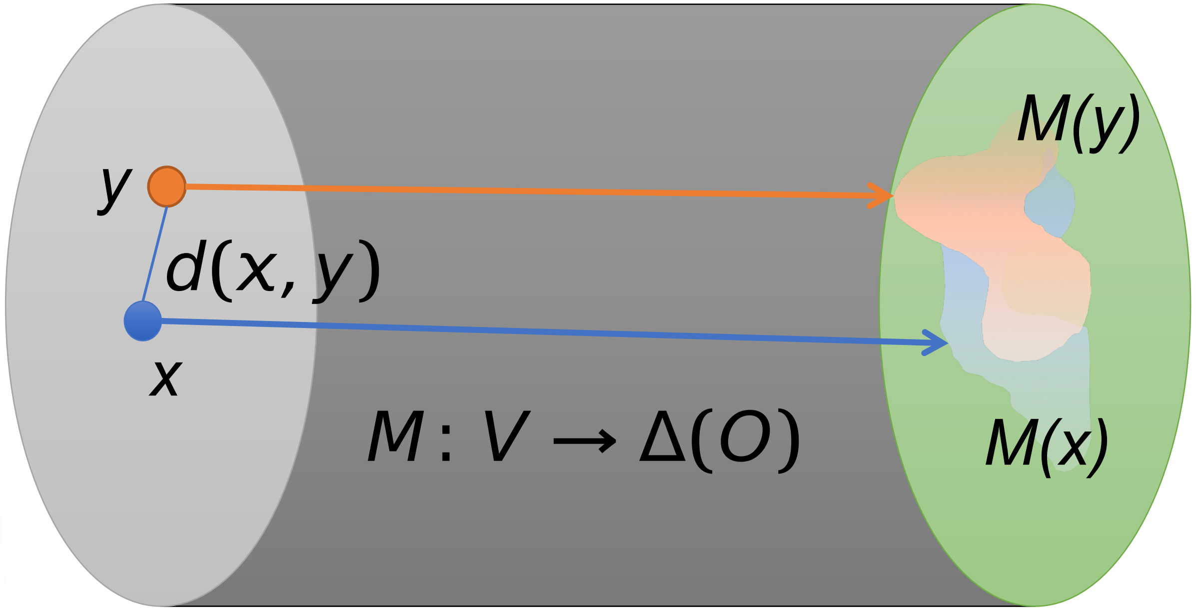
- O={0,1}
- $d\big(M(x), M(y)\big) = \frac{1}{2}\sum_{o \in O}|M(x)(o) - M(y)(o)|$

$M(x)(0), = 1$	$M(x)(1)$	$M(y)(0)$	$M(y)(1)$	$d\big(M(x), M(y)\big)$
1	0	0	1	1
1	0	1	0	0
½	½	¾	¼	1/4

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

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

e.g., Statistical distance in [0,1]



V: Individuals O: outcomes

Existence Proof

There exists a classifier that satisfies the Lipschitz condition

- Idea: Map all individuals to the same distribution over outcomes
- Are we done?

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

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

Semantics derived automatically from language corpora contain human-like biases.

Caliskan et al.

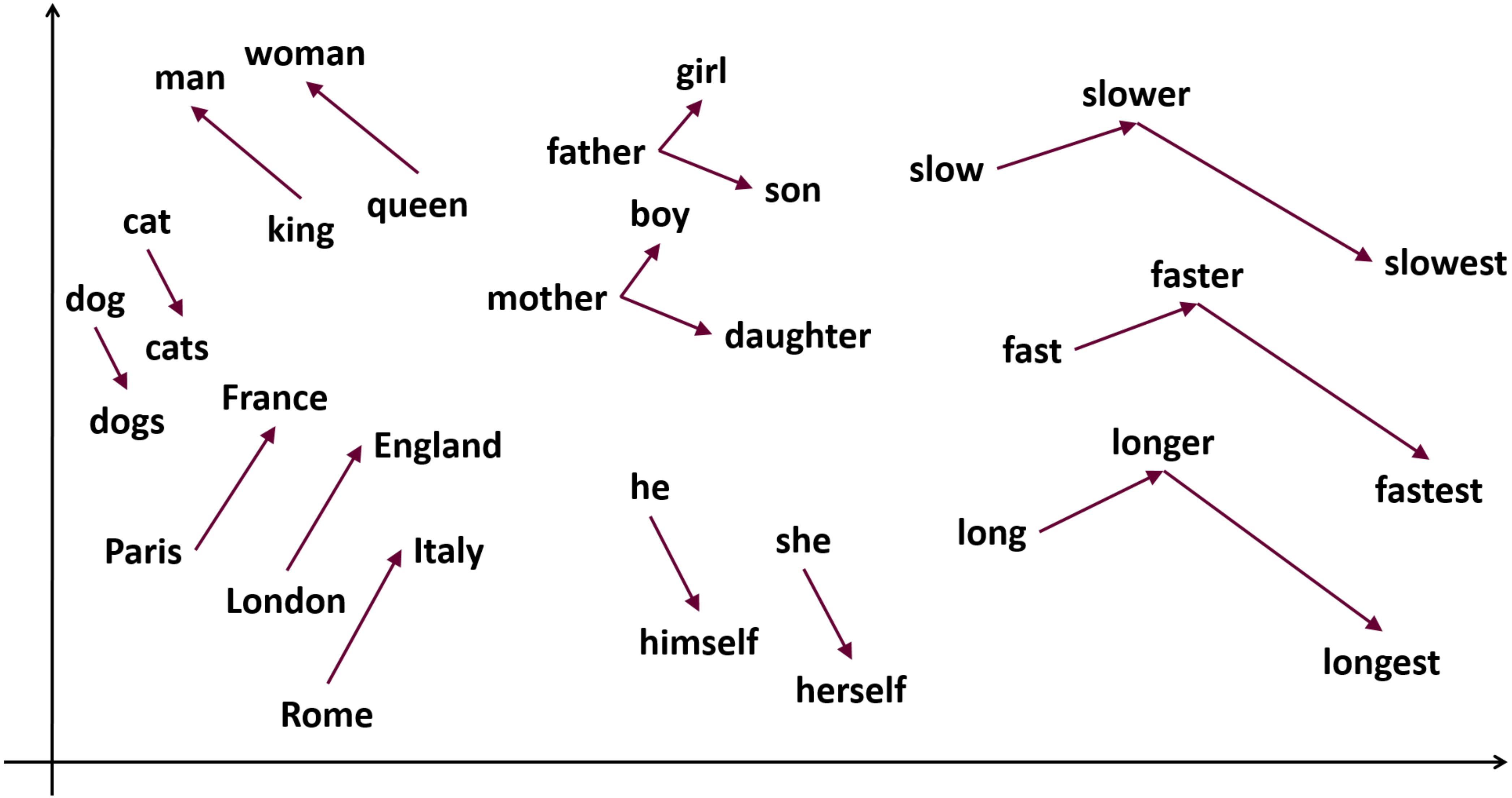
What is bias?

- Bias found in language data, learned by humans and ML
- Here stereotyped bias is defined as "problematic where such information is derived from aspects of human culture known to lead to harmful behavior"
- Prejudiced actions are taken based on stereotyped bias

How to measure bias?

- Humans:
  - Implicit Association Test
  - Response time differs when humans pair concepts that they find similar compared to concepts that they find different
- Machines:
  - Word embeddings
  - Measure cosine distance between embedding vectors

Word embeddings



N: population size
d: effect size
p : p-value

N_T: number of target words
N_A: number of attribute words

Target words	Attrib. words	Original Finding				Our Finding			
		Ref	N	d	p	N_T	N_A	d	p
Flowers vs insects	Pleasant vs unpleasant	(*5*)	32	1.35	10^{-8}	25×2	25×2	1.50	10^{-7}
Instruments vs weapons	Pleasant vs unpleasant	(*5*)	32	1.66	10^{-10}	25×2	25×2	1.53	10^{-7}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(*5*)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*5*)	(*7*)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (*9*)	(*7*)	Not applicable			16×2	8×2	1.28	10^{-3}
Male vs female names	Career vs family	(*9*)	$39k$	0.72	$< 10^{-2}$	8×2	8×2	1.81	10^{-3}
Math vs arts	Male vs female terms	(*9*)	$28k$	0.82	$< 10^{-2}$	8×2	8×2	1.06	.018
Science vs arts	Male vs female terms	(*10*)	91	1.47	10^{-24}	8×2	8×2	1.24	10^{-2}
Mental vs physical disease	Temporary vs permanent	(*23*)	135	1.01	10^{-3}	6×2	7×2	1.38	10^{-2}
Young vs old people's names	Pleasant vs unpleasant	(*9*)	$43k$	1.42	$< 10^{-2}$	8×2	8×2	1.21	10^{-2}

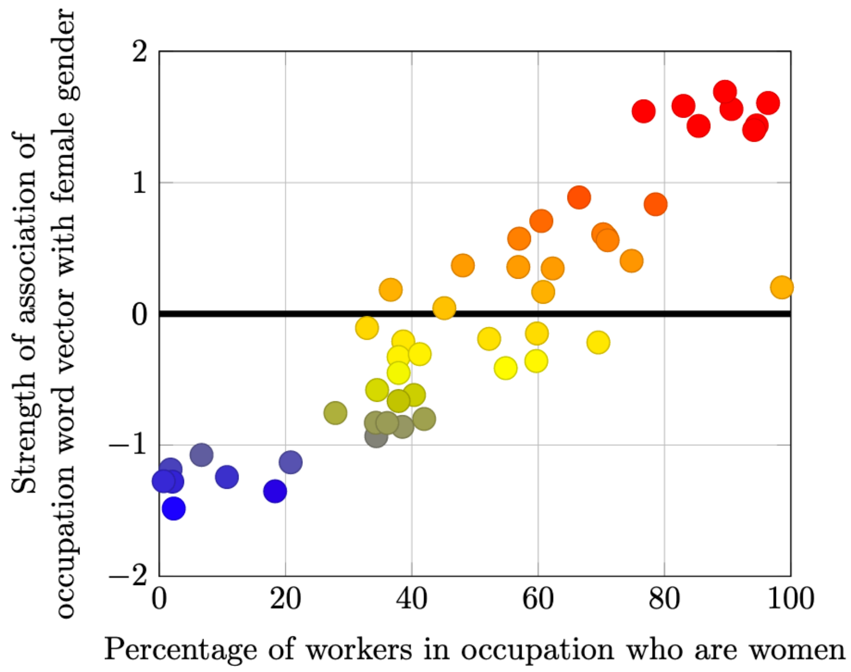


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

Word Embedding Factual Association Test

Target word

Target attributes

$$s(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std-dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})}$$

Potential for Discrimination in Online Targeted Advertising

Speicher et al.

- Disclaimer: these are my personal opinions

PII-based targeting through Facebook

Public data source (e.g., voter records) → Submit PII to select audience on FB → Measure fraction of targetable audience which has sensitive attribute per FB's records

Attribute	Voter Records		Facebook Users		Validation of Custom Audience
	Number	Percent	Targetable	Targetable %	% matching sensitive attribute
Male	3,438,620	45.5%	6,500	65%	81.5%
Female	3,995,533	52.8%	7,000	70%	91.4%
White	5,303,383	70.1%	6,800	68%	83.8%
Black	1,694,220	22.4%	6,300	63%	82.5%
Asian	79,250	1.0%	6,600	66%	28.8%
Hispanic	163,236	2.2%	5,900	59%	50.8%
Age (18-34)	1,985,117	26.2%	7,100	71%	80.3%
Age (35-54)	2,496,648	33.0%	6,900	69%	79.7%
Age (55+)	3,068,745	40.6%	5,700	57%	61.4%

Look alike targeting through Facebook

Public data source
(e.g., voter records)

↓

Submit PII to select
look alike audience
on FB

↓

Measure over-represented and
under-represented attributes

Table 6: Top 5 most over-represented and under-represented attributes in a source audience of African Americans and its two closest look-alike audiences. In parentheses, we show the value of the representation bias of each attribute.

Over-represented Attributes	Under-represented Attributes
Source Audience	
African American affinity (5.52)	Asian American affinity (0.09)
US politics: very liberal (3.21)	Hispanic (Spanish dominant) affinity (0.09)
Liberal content engagement (2.98)	Expats: Mexico (0.11)
Interest: Gospel music (2.64)	Hispanic (all) affinity (0.18)
Interest: Dancehalls (2.51)	Expats: all countries (0.22)
2% Look-Alike Audience	
African American affinity (5.24)	Hispanic (Spanish dominant) affinity (0.10)
Liberal content engagement (4.16)	Expats: Mexico (0.13)
US politics: very liberal (3.29)	Asian American affinity (0.13)
Interest: Gospel music (3.07)	Hispanic (all) affinity (0.19)
Interest: Soul music (2.32)	Expats: all countries (0.24)
2–4% Look-Alike Audience	
African American affinity (5.06)	Asian American affinity (0.17)
Liberal content engagement (3.61)	Hispanic (Spanish dominant) affinity (0.18)
US politics: very liberal (3.37)	Expats: Mexico (0.19)
Interest: Gospel music (2.72)	Hispanic (all) affinity (0.29)
Interest: Dancehalls (2.54)	Expats: all countries (0.37)

Add ML into the picture

- ML could exacerbate: recall lecture on overlearning

- Could use ML + {DP, fairness} techniques to decrease potential for discrimination. Research needed to validate.

See also:

- Ali et al. *Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes*
- Faizullabhoy et al. *Facebook's Advertising Platform: New Attack Vectors and the Need for Interventions*

Law and Adversarial Machine Learning

Kumar et al.

Law & technology

- ML at core of critical technologies
  - Healthcare
  - Defense
  - Finance
- Is law adequate to capture new harms brought by AML?

- Example of model extraction

"Said another way, even if stealing software were easy, there is still an important disincentive to do so in that it violates intellectual property law" (BigML)

Computer Fraud and Abuse Act

the CFAA broadly prohibits individuals from:

1. intentionally accessing computers without authorization
2. exceeding authorized access on a computer
3. causing damage to computers without authorization

- Inserting backdoors in pretrained model zoos (#1 and #3)
- Poisoning attack (#3) <u>but</u> when is data malicious?
- Adversarial examples (#3)

Assumes transmission of data is interpreted as transmission of code

Copyright law

- Copyright law is more well-defined than CFAA
  - Facts are not copyrightable
- Model inversion would likely produce a different arrangement of facts, approximating the original training data.
- Model extraction is unlikely to violate copyright law if the extracted model is not expressed with the same code than the victim model

Liability laws

- Who is liable if a ML system breaks down because of an adversarial example?

- Need to establish what qualifies as responsible ML development
- Need to develop forensics for ML system
  - Which component is responsible
  - Attack attribution

Ethics

Beneficial use of adversarial ML to defend civil liberties

In Hong Kong Protests, Faces Become Weapons (New York Times)

Negative use of ML to pollute public discourse