Machine Learning 2

DS 4420 - Spring 2020

Bias and fairness

Byron C. Wallace

Material in this lecture modified from materials created by Jay Alammar (http://jalammar.github.io/illustrated-transformer/) and Sasha Rush (https://nlp.seas.harvard.edu/2018/04/03/attention.html).

Intro

Today

- We will talk about *bias* and *fairness*, which are **critically important** to understand if you go out and apply models in real-world settings

Examples
[from CIML, Daume III]

- Early speech recognition systems failed on female voices.

Examples
[from CIML, Daume III]

- Early speech recognition systems failed on female voices.
- Models to predict *criminal recidivism* biased against minorities.

IAN WALDIE/GETTY IMAGES

Tech Policy / AI Ethics

AI is sending people to jail —and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** Jan 21, 2019

The New York Times

Facial Recognition Is Accurate, if You're a White Guy

By Steve Lohr

Feb. 9, 2018

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Can word vectors be sexist?

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

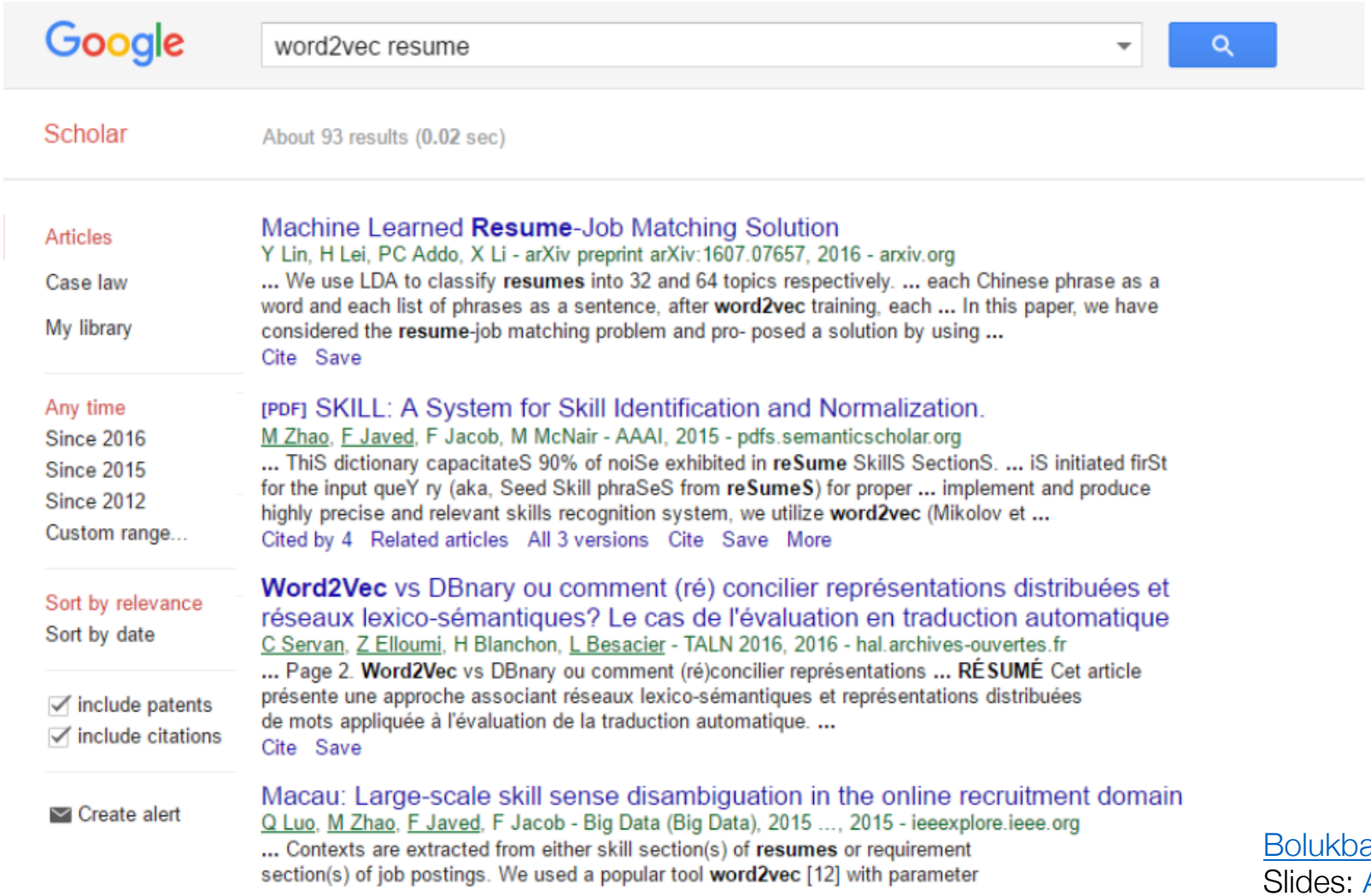
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss



Bolukbasi et al. '16
Slides: Adam Kalai

[Figure from: https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da](https://towardsdatascience.com/named-entity-recognition-with-nltk-and-spacy-8c4a7d88e7da)

Recognizing names in text

Country	(Huang et al., 2015) GloVe words			(Lample et al., 2016) GloVe words+chars			(Devlin et al., 2019) BERT subwords		
	P	R	F1	P	R	F1	P	R	F1
Original	96.9	96.5	96.7	97.1	98.1	97.6	98.3	98.1	98.2
US	96.9	99.6	98.2	96.9	99.6	98.3	98.4	99.7	99.1
Russia	96.8	99.5	98.1	97.1	99.8	98.4	98.4	99.3	98.9
India	96.5	99.5	98.0	97.1	99.3	98.2	98.4	98.8	98.6
Mexico	96.7	98.9	97.8	97.1	98.9	98.0	98.4	99.2	98.8
China-Taiwan	95.4	93.2	93.9	97.0	94.9	95.6	98.3	92.0	94.8
US (Difficult)	95.9	87.4	90.2	96.6	87.9	90.7	98.1	88.5	92.3
Indonesia	95.3	84.6	88.7	96.5	91.0	93.3	97.8	85.8	92.0
Vietnam	94.6	78.2	84.2	96.0	78.5	84.5	98.0	84.2	89.8

Intermezzo 1

Before moving on to the next part of lecture,
let's walk through this notebook tutorial

https://nbviewer.jupyter.org/github/Azure-Samples/learnAnalytics-DeepLearning-Azure/blob/master/Students/12-biased-embeddings/how-to-make-a-racist-ai-without-really-trying.ipynb

Domain adaptation

One potential cause: Train/test mismatch

- If the train set is drawn from a different distribution than the test set, this introduces a *bias* such that the model will do better on examples that look like train set instances

One potential cause: Train/test mismatch

- If the train set is drawn from a different distribution than the test set, this introduces a *bias* such that the model will do better on examples that look like train set instances

- If the speech recognition model has been trained on mostly male voices and optimized well, it will tend to do better on male voices.

Unsupervised adaptation

- Given training data from distribution D^{old}, learn a classifier that performs well on a related, but distinct, distribution D^{new}

Unsupervised adaptation

- Given training data from distribution D^{old}, learn a classifier that performs well on a related, but distinct, distribution D^{new}
- Assumption is that we have train data from D^{old} but what we actually care about is loss on D^{new}

Unsupervised adaptation

- Given training data from distribution D^{old}, learn a classifier that performs well on a related, but distinct, distribution D^{new}
- Assumption is that we have train data from D^{old} but what we actually care about is loss on D^{new}
- What can we do here?

Importance sampling (re-weighting)

Test loss $= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}[\ell(y, f(x))]$ definition (8.2)

[from CIML, Daume III]

Importance sampling (re-weighting)

$$\text{Test loss} = \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}\left[\ell(y, f(x))\right] \qquad \text{definition} \qquad (8.2)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x, y)\ell(y, f(x)) \qquad \text{expand expectation} \qquad (8.3)$$

[from CIML, Daume III]

Importance sampling (re-weighting)

$$\text{Test loss} = \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}\left[\ell(y, f(x))\right] \qquad \text{definition} \qquad (8.2)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\ell(y, f(x)) \qquad \text{expand expectation} \qquad (8.3)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\frac{\mathcal{D}^{\text{old}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x)) \qquad \text{times one} \qquad (8.4)$$

Note: Does this look familiar?!

[from CIML, Daume III]

Importance sampling (re-weighting)

$$\text{Test loss} = \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}\left[\ell(y, f(x))\right] \qquad \text{definition} \qquad (8.2)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\ell(y, f(x)) \qquad \text{expand expectation} \qquad (8.3)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\frac{\mathcal{D}^{\text{old}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x)) \qquad \text{times one} \qquad (8.4)$$

$$= \sum_{(x,y)} \mathcal{D}^{\text{old}}(x,y)\,\frac{\mathcal{D}^{\text{new}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x)) \qquad \text{rearrange} \qquad (8.5)$$

[from CIML, Daume III]

Importance sampling (re-weighting)

$$
\begin{aligned}
\text{Test loss} &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{new}}}\left[\ell(y, f(x))\right] && \text{definition} && (8.2)\\
&= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\ell(y, f(x)) && \text{expand expectation} && (8.3)\\
&= \sum_{(x,y)} \mathcal{D}^{\text{new}}(x,y)\,\frac{\mathcal{D}^{\text{old}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x)) && \text{times one} && (8.4)\\
&= \sum_{(x,y)} \mathcal{D}^{\text{old}}(x,y)\,\frac{\mathcal{D}^{\text{new}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x)) && \text{rearrange} && (8.5)\\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{old}}}\left[\frac{\mathcal{D}^{\text{new}}(x,y)}{\mathcal{D}^{\text{old}}(x,y)}\,\ell(y, f(x))\right] && \text{definition} && (8.6)
\end{aligned}
$$

[from CIML, Daume III]

Importance weighting

- So we have re-expressed the test loss as an expectation over D^{old}, which is good because that's what we have for training data
- But we do not have access to D^{old} or D^{new} directly

Ratio estimation

Assume all examples drawn from an underlying shared distribution (base), and then sorted into D^{old} / D^{new} with some probability depending on *x*

$$\mathcal{D}^{\text{old}}(\boldsymbol{x}, y) \propto \mathcal{D}^{\text{base}}(\boldsymbol{x}, y) p(s = 1 \mid \boldsymbol{x})$$

$$\mathcal{D}^{\text{new}}(\boldsymbol{x}, y) \propto \mathcal{D}^{\text{base}}(\boldsymbol{x}, y) p(s = 0 \mid \boldsymbol{x})$$

Ratio estimation

Supposing we can estimate *p*… we can reweight examples:

Train pair	Weight
$(\boldsymbol{x}_n, y_n)$	$1/p(s = 1 \mid \boldsymbol{x}_n) - 1$

P that this example assigned to D^{old}

Intuitively: Upweights instances likely to be from D^{new}

How should we estimate *p*?

Want to estimate: $p(s = 1 \mid \boldsymbol{x}_n)$

This is just a binary classification task!

Algorithm 23 SELECTIONADAPTATION($\langle (\boldsymbol{x}_n, y_n) \rangle_{n=1}^N$, $\langle \boldsymbol{z}_m \rangle_{m=1}^M$, $\mathcal{A}$)

1: $D^{dist} \leftarrow \langle (\boldsymbol{x}_n, +1) \rangle_{n=1}^N \bigcup \langle (\boldsymbol{z}_m, -1) \rangle_{m=1}^M$ // assemble data for distinguishing // between old and new distributions

2: $\hat{p} \leftarrow$ train logistic regression on D^{dist}

3: $D^{weighted} \leftarrow \left\langle (\boldsymbol{x}_n, y_n, \frac{1}{\hat{p}(\boldsymbol{x}_n)} - 1) \right\rangle_{n=1}^N$ // assemble weight classification // data using selector

4: **return** $\mathcal{A}(D^{weighted})$ // train classifier

[from CIML, Daume III]

Supervised adaptation

- We were supposing that we had access to *labels* only in D^{old}, but wanted to learn a model for D^{new}

Supervised adaptation

- We were supposing that we had access to *labels* only in D^{old}, but wanted to learn a model for D^{new}
- In some cases we might have at least *some* labels from D^{new} as well

Supervised adaptation via *feature augmentation*

shared old-only new-only

Supervised adaptation via *feature augmentation*

$$\boldsymbol{x}_n^{(\text{old})} \mapsto \Big\langle \underbrace{\boldsymbol{x}_n^{(\text{old})}}_{\text{shared}}, \underbrace{\boldsymbol{x}_n^{(\text{old})}}_{\text{old-only}}, \underbrace{\underbrace{0, 0, \ldots, 0}_{D\text{-many}}}_{\text{new-only}} \Big\rangle$$

Supervised adaptation via *feature augmentation*

$$
\begin{array}{lcccc}
 & & \text{shared} & \text{old-only} & \text{new-only} \\
\boldsymbol{x}_n^{(\text{old})} & \mapsto & \langle\ \boldsymbol{x}_n^{(\text{old})} & , \boldsymbol{x}_n^{(\text{old})} & , \underbrace{0, 0, \ldots, 0}_{D\text{-many}}\ \rangle \\
\boldsymbol{x}_m^{(\text{new})} & \mapsto & \langle\ \boldsymbol{x}_m^{(\text{new})} & , \underbrace{0, 0, \ldots, 0}_{D\text{-many}} & , \boldsymbol{x}_m^{(\text{new})}\ \rangle
\end{array}
$$

Supervised adaptation via *feature augmentation*

$$
\begin{array}{lcccc}
 & & \text{shared} & \text{old-only} & \text{new-only} \\
\boldsymbol{x}_n^{(\text{old})} \mapsto & \big\langle & \boldsymbol{x}_n^{(\text{old})} & , \boldsymbol{x}_n^{(\text{old})} & , \underbrace{0, 0, \ldots, 0}_{D\text{-many}} \big\rangle \\
\boldsymbol{x}_m^{(\text{new})} \mapsto & \big\langle & \boldsymbol{x}_m^{(\text{new})} & , \underbrace{0, 0, \ldots, 0}_{D\text{-many}} & , \boldsymbol{x}_m^{(\text{new})} \big\rangle
\end{array}
$$

We have seen this trick before!!

Algorithm 24 EASYADAPT($\langle (\boldsymbol{x}_n^{(old)}, y_n^{(old)}) \rangle_{n=1}^N$, $\langle (\boldsymbol{x}_m^{(new)}, y_m^{(new)}) \rangle_{m=1}^M$, $\mathcal{A}$)

1: $D \leftarrow \left\langle (\langle \boldsymbol{x}_n^{(old)}, \boldsymbol{x}_n^{(old)}, \mathbf{0} \rangle, y_n^{(old)}) \right\rangle_{n=1}^N \bigcup \left\langle (\langle \boldsymbol{x}_m^{(new)}, \mathbf{0}, \boldsymbol{x}_m^{(new)} \rangle, y_m^{(new)}) \right\rangle_{m=1}^M$ // union // of transformed data

2: **return** $\mathcal{A}(D)$ // train classifier

[from CIML, Daume III]

Subtler bias

- What if the distribution is not different, but rather the train data simply reflects biases in society?
- Training models on this and then using them can magnify biases that already exist!

Sensitive attributes

- In many settings, there may be certain fields/attributes that we know *a priori* we don't want to exploit. Great: Let's just remove these!

Sensitive attributes

- In many settings, there may be certain fields/attributes that we know *a priori* we don't want to exploit. Great: Let's just remove these!

- Why is this not enough?

Sensitive attributes

- In many settings, there may be certain fields/attributes that we know *a priori* we don't want to exploit. Great: Let's just remove these!

- Why is this not enough?

Because other features may correlate strongly with the protected feature!

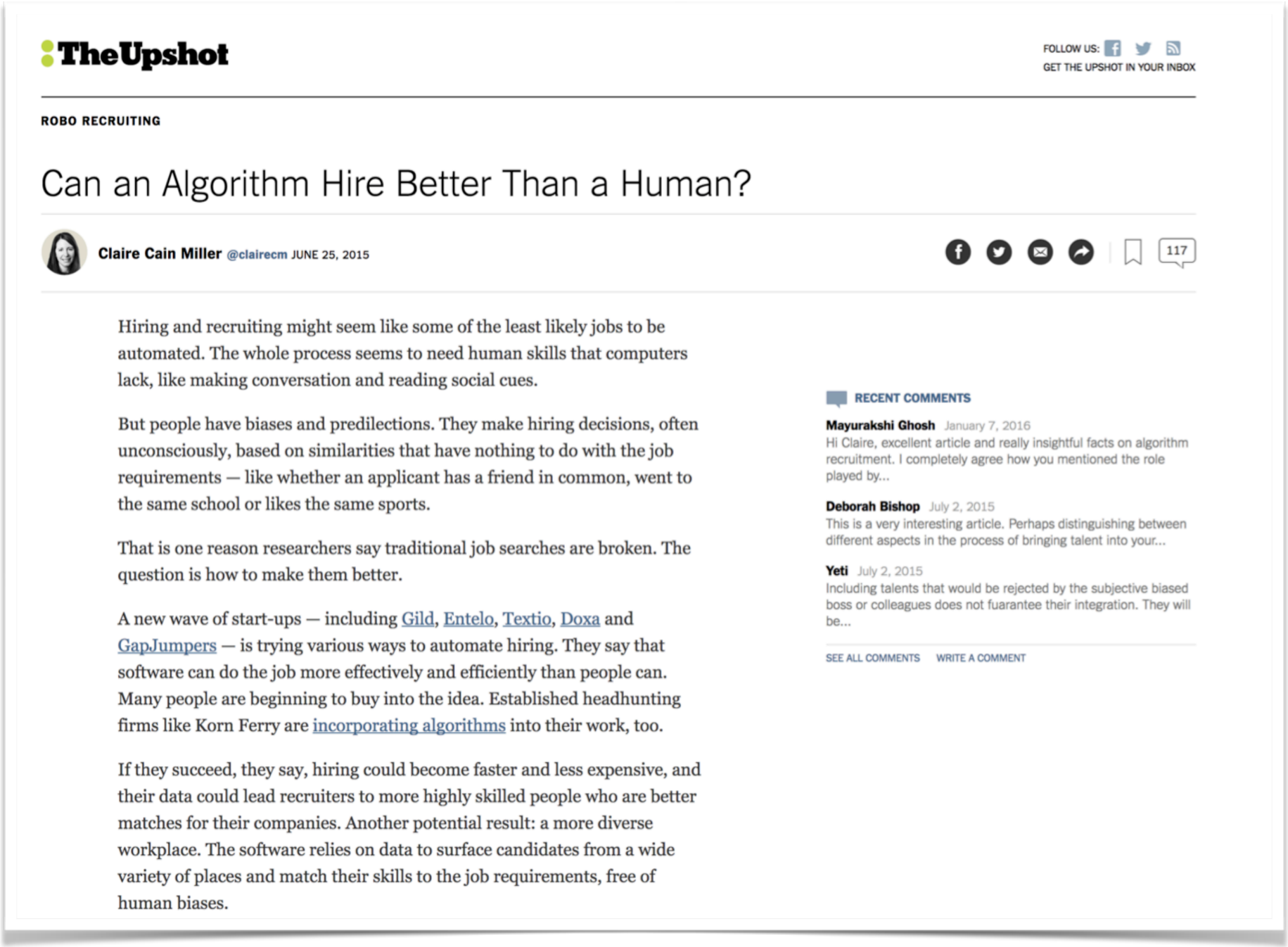
Let's consider a concrete example: Hiring

Following slides derived from: https://mrtz.org/nips17

Solon Barocas and Moritz Hardt

Cornell University

Berkeley
UNIVERSITY OF CALIFORNIA



TheUpshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

117

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

A new wave of start-ups — including Gild, Entelo, Textio, Doxa and GapJumpers — is trying various ways to automate hiring. They say that software can do the job more effectively and efficiently than people can. Many people are beginning to buy into the idea. Established headhunting firms like Korn Ferry are incorporating algorithms into their work, too.

If they succeed, they say, hiring could become faster and less expensive, and their data could lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases.

RECENT COMMENTS

Mayurakshi Ghosh January 7, 2016
Hi Claire, excellent article and really insightful facts on algorithm recruitment. I completely agree how you mentioned the role played by...

Deborah Bishop July 2, 2015
This is a very interesting article. Perhaps distinguishing between different aspects in the process of bringing talent into your...

Yeti July 2, 2015
Including talents that would be rejected by the subjective biased boss or colleagues does not fuarantee their integration. They will be...

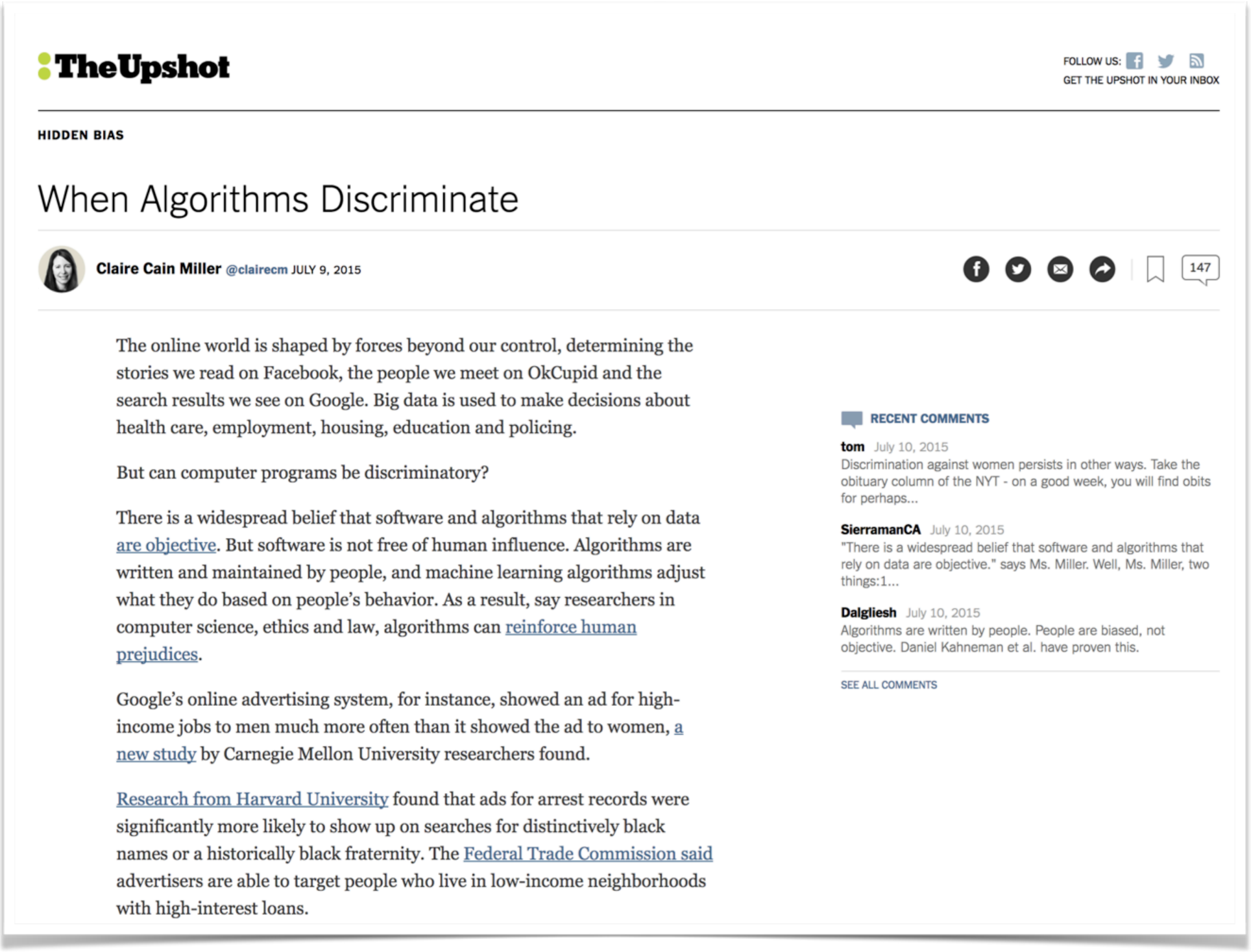
SEE ALL COMMENTS WRITE A COMMENT

Credit: Barocas and Hardt

"[H]iring could become faster and less expensive, and [...] lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

Miller (2015)

Credit: Barocas and Hardt



The Upshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

147

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods with high-interest loans.

RECENT COMMENTS

tom July 10, 2015
Discrimination against women persists in other ways. Take the obituary column of the NYT - on a good week, you will find obits for perhaps...

SierramanCA July 10, 2015
"There is a widespread belief that software and algorithms that rely on data are objective." says Ms. Miller. Well, Ms. Miller, two things:1...

Dalgliesh July 10, 2015
Algorithms are written by people. People are biased, not objective. Daniel Kahneman et al. have proven this.

SEE ALL COMMENTS

Credit: Barocas and Hardt

Isn't discrimination the very point of machine learning?

Credit: Barocas and Hardt

Discrimination is not a general concept

It is **domain** specific

Concerned with important opportunities that affect people's life chances

It is **feature** specific

Concerned with socially salient qualities that have served as the basis for unjustified and systematically adverse treatment in the past

Credit: Barocas and Hardt

Regulated domains

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **'Public Accommodation'** (Civil Rights Act of 1964)

Credit: Barocas and Hardt

Legally recognized ‘protected classes’

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Credit: Barocas and Hardt

Discrimination Law: Two Doctrines

Disparate Treatment

Formal

or

Intentional

Disparate Impact

Unjustified

or

Avoidable

Credit: Barocas and Hardt

Disparate Treatment

Formal: explicitly considering class membership

Even if it is relevant

Intentional: purposefully attempting to discriminate without direct reference to class membership

Pretext or 'motivating factor'

Credit: Barocas and Hardt

Disparate Impact

1. Plaintiff must first establish that decision procedure has a disparate impact

‘Four-fifths rule’

2. Defendant must provide a justification for making decisions in this way

‘Business necessity’ and 'job-related’

3. Finally, plaintiff has opportunity to show that defendant could achieve same goal using a different procedure that would result in a smaller disparity

‘Alternative practice’

Credit: Barocas and Hardt

What does discrimination law aim to achieve?

Disparate Treatment	**Disparate Impact**
Procedural fairness	Distributive justice
Equality of opportunity	Minimized inequality of outcome

Credit: Barocas and Hardt

How *machines* learn to discriminate

Skewed sample
Tainted examples
Limited features
Sample size disparity
Proxies

B, Selbst (2016)

Credit: Barocas and Hardt

Skewed sample

Police records measure “some complex interaction between criminality, policing strategy, and community-policing relations”
Lum, Isaac (2016)

Credit: Barocas and Hardt

Skewed sample: feedback loop

Future observations of crime confirm predictions

Fewer opportunities to observe crime that contradicts predictions

Initial bias may compound over time

Credit: Barocas and Hardt

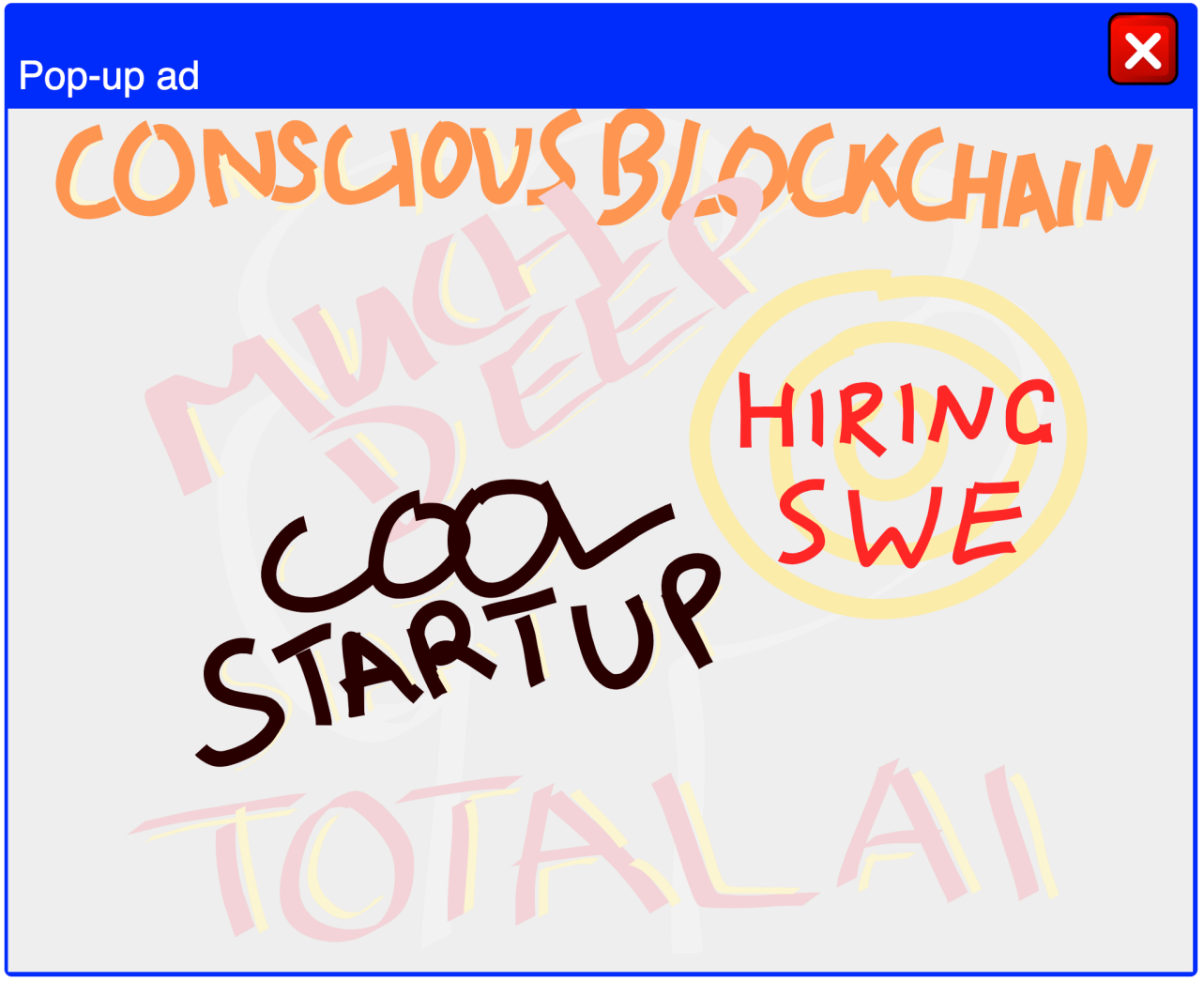
Limited features

Features may be less informative or less reliably collected for certain parts of the population

A feature set that supports accurate predictions for the majority group may not for a minority group

Different models with the same reported accuracy can have a very different distribution of error across population

Credit: Barocas and Hardt



Running example: Hiring ad for (fictitious?) AI startup

Credit: Barocas and Hardt

Formal setup

- X features of an individual (browsing history etc.)
- A sensitive attribute (here, gender)
- $C = c(X, A)$ predictor (here, show ad or not)
- Y target variable (here, SWE)

Credit: Barocas and Hardt

Formal setup

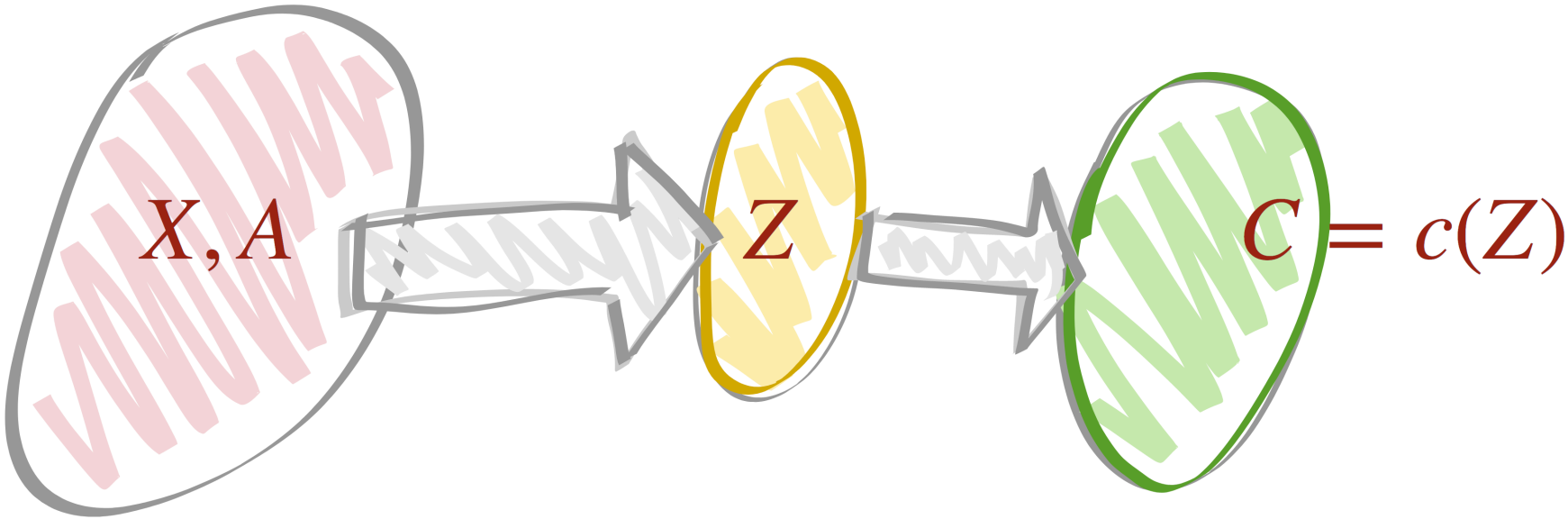
Score function is any random variable $R = r(X, A) \in [0, 1]$.

Can be turned into (binary) predictor by thresholding

Example: *Bayes optimal* score given by $r(x, a) = \mathbb{E}[Y \mid X = x, A = a]$

Credit: Barocas and Hardt

Representation learning approach



Credit: Barocas and Hardt

Adversarial Learning for Fairness

Beutel *et al.,* 2017; Edwards and Storkey, 2015

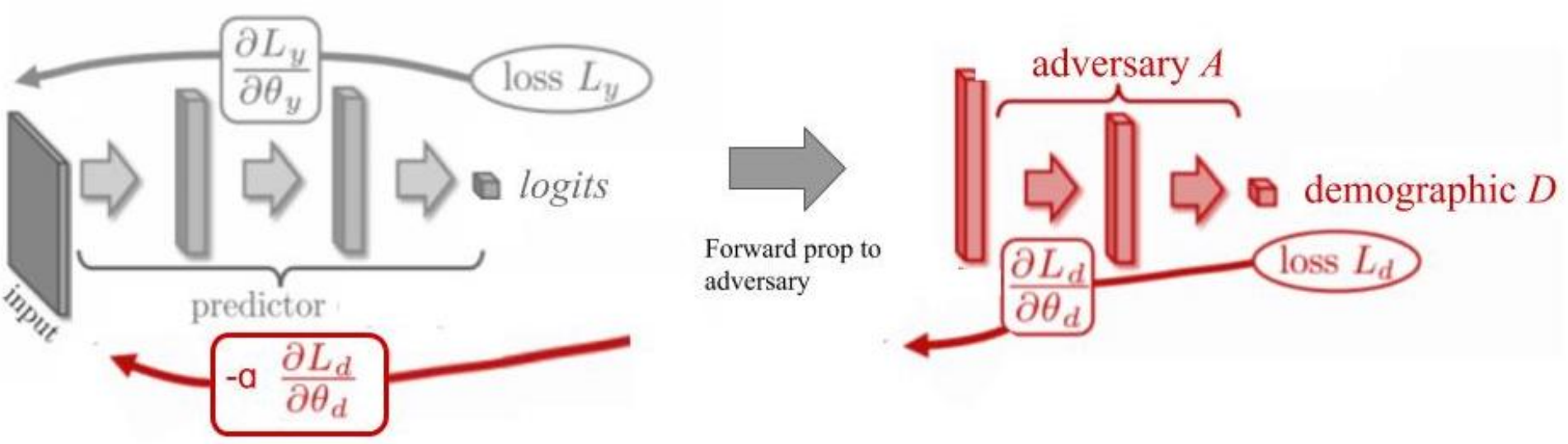


Figure from Wadsworth et al., 2018

Three fundamental criteria

Independence: C independent of A

Separation: C independent of A conditional on Y

Sufficiency: Y independent of A conditional on C

Lots of other criteria are related to these

Credit: Barocas and Hardt

First criterion: Independence

Require C and A to be independent, denoted $C \perp\!\!\!\perp A$

That is, for all groups a, b and all values c:

$$\mathbb{P}_a\{C = c\} = \mathbb{P}_b\{C = c\}$$

Credit: Barocas and Hardt

Variants of independence

Sometimes called *demographic parity*, *statistical parity*

When C is binary $0/1$-variables, this means
$\mathbb{P}_a\{C = 1\} = \mathbb{P}_b\{C = 1\}$ for all groups a, b.

Approximate versions:

$$\frac{\mathbb{P}_a\{C = 1\}}{\mathbb{P}_b\{C = 1\}} \geq 1 - \epsilon \qquad |\mathbb{P}_a\{C = 1\} - \mathbb{P}_b\{C = 1\}| \leq \epsilon$$

Credit: Barocas and Hardt

So does this solve it?

- Drawbacks to independence as our criterion? Is this the right objective?

So does this solve it?

- Drawbacks to independence as our criterion? Is this the right objective?
  - May rule out best possible model due to actual correlations in the real-world. Could rule out C = Y (perfect predictor).

So does this solve it?

- Drawbacks to independence as our criterion? Is this the right objective?
  - May rule out best possible model due to actual correlations in the real-world. Could rule out C = Y (perfect predictor).
  - Can satisfy by just selecting random people form the minority group as “positive” – will not dramatically lower error rate (since they are a minority) but will satisfy constraint.

So does this solve it?

- Drawbacks to independence as our criterion? Is this the right objective?
  - May rule out best possible model due to actual correlations in the real-world. Could rule out C = Y (perfect predictor).
  - Can satisfy by just selecting random people form the minority group as “positive” – will not dramatically lower error rate (since they are a minority) but will satisfy constraint.

- Other criteria exist – let’s look at one more: **separation**

Separation

Require R and A to be independent *conditional on target variable* Y, denoted $R \perp A \mid Y$

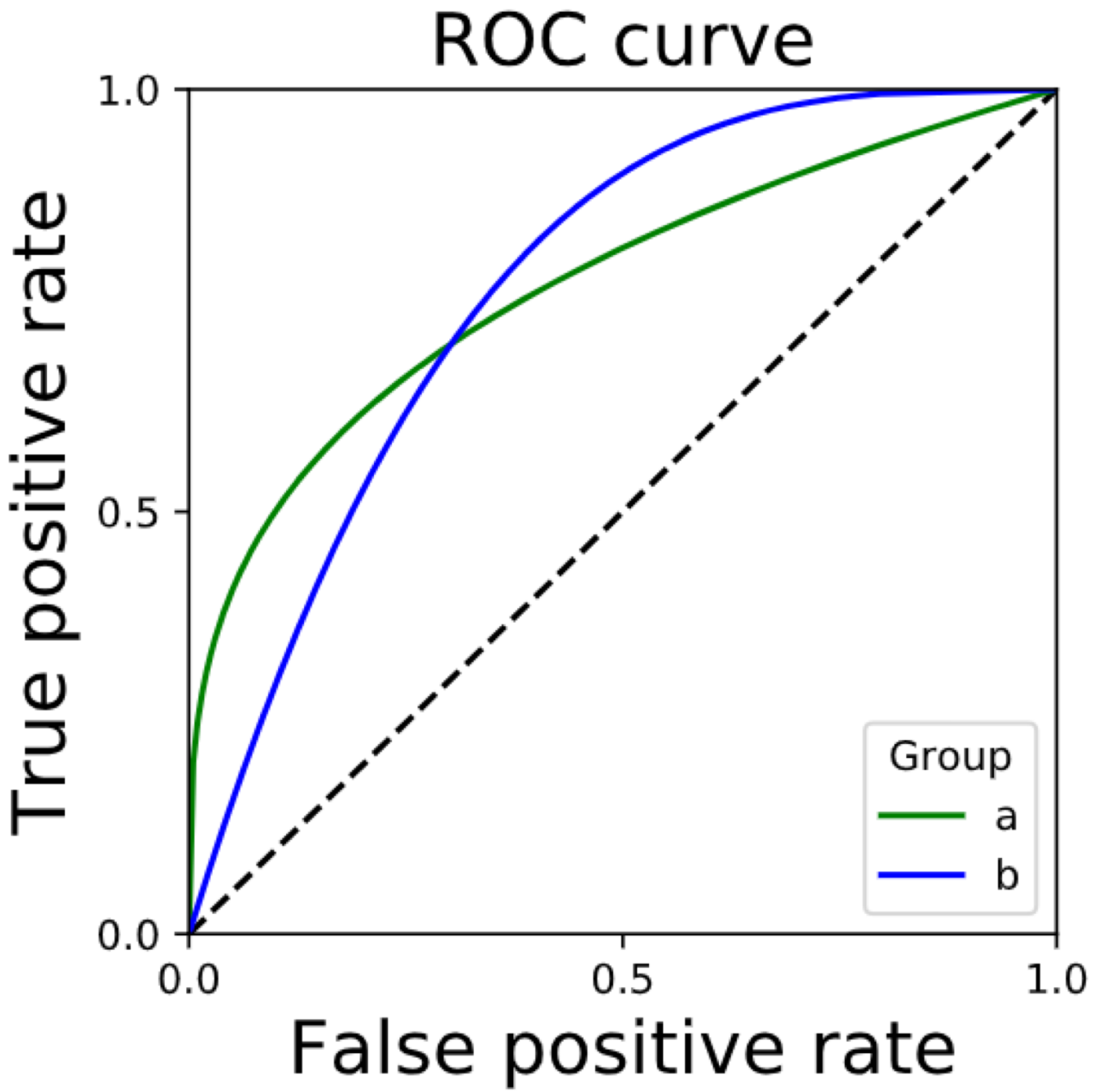
Definition. Random variable R *separated* from A if $R \perp A \mid Y$.

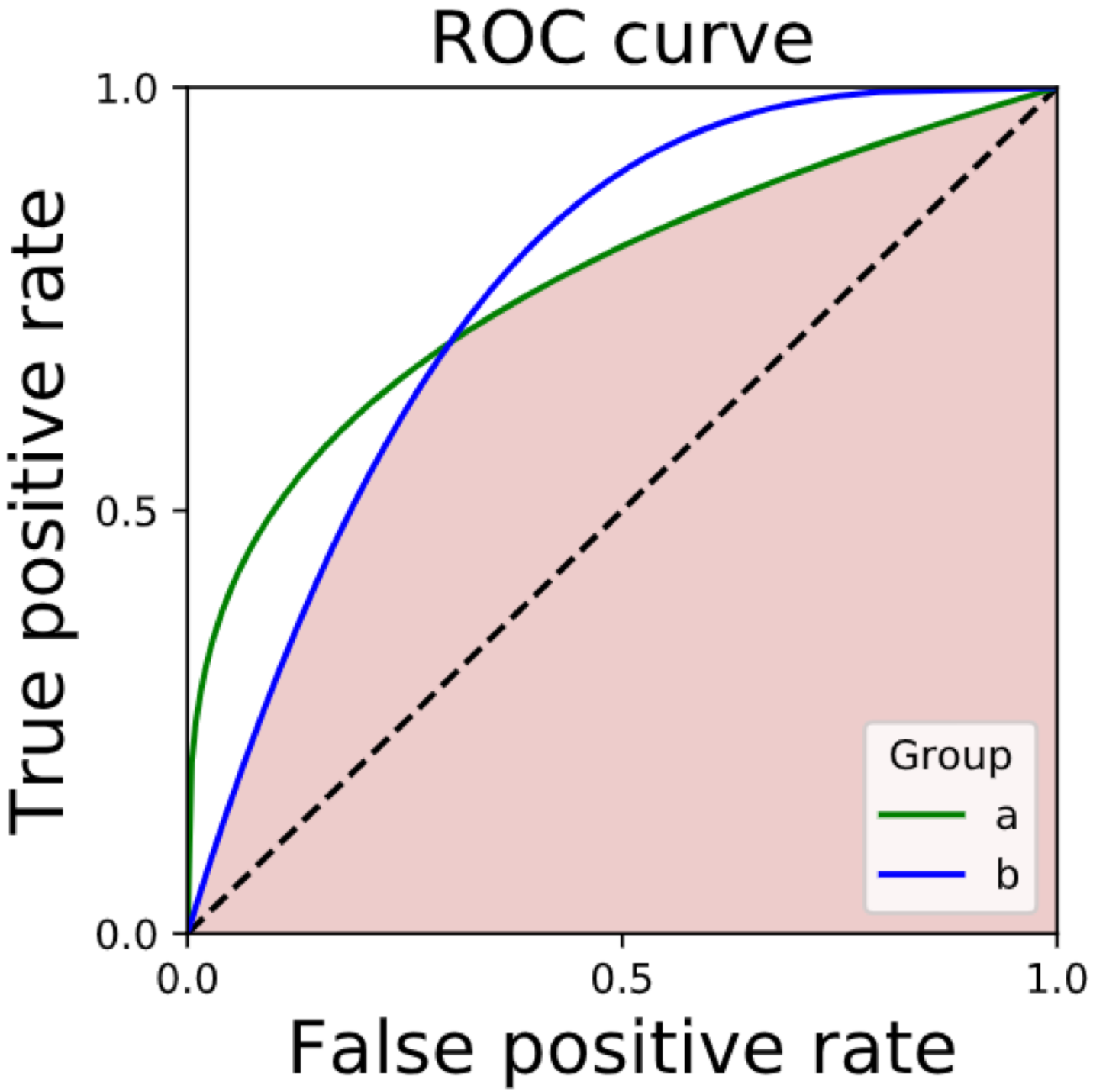
Credit: Barocas and Hardt

Post-Processing

Method from H, Price, Srebro (2016):
Post-processing correct of score function

Post-processing: Any thresholding of R (possibly depending on A)
No retraining/changes to R


ROC curve
True positive rate
False positive rate
1.0
0.5
0.0
0.0
0.5
1.0
Group
a
b


ROC curve
True positive rate
False positive rate
1.0
0.5
0.0
0.0
0.5
1.0
Group
a
b