

Fairness in Machine Learning

David Madras

University of Toronto
Vector Institute



About Me

- I'm a PhD student in Machine Learning at the University of Toronto
 - Also affiliated with the Vector Institute
- At the moment, I'm mostly thinking about how to build ethical and fair machine learning models/algorithms
 - I'm also interested in causal inference, generative modelling, and deep learning

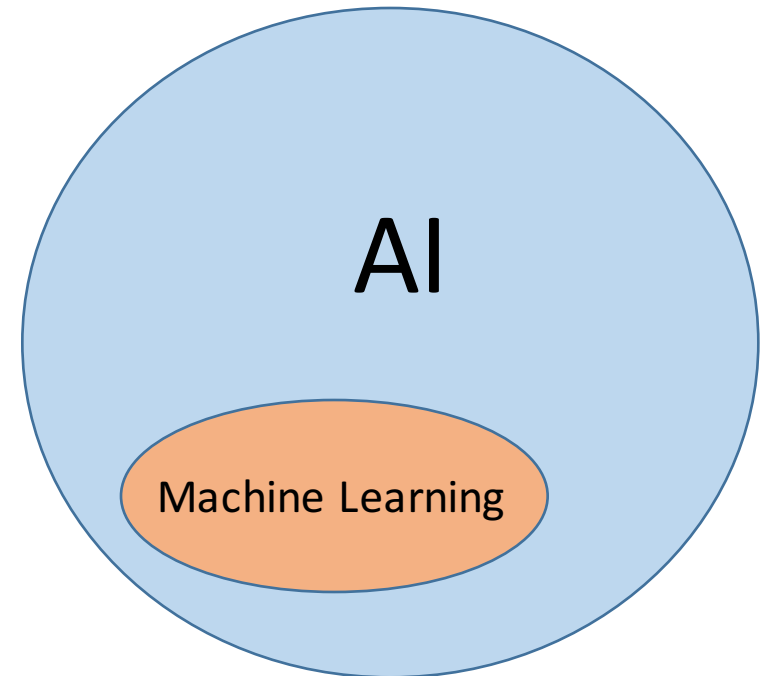
This Talk

- In this talk, I'll be discussing **fairness in machine learning**
- I'll give **examples** of unfairness in machine learning, discuss some ways people have tried to **define** fairness mathematically, and talk about some approaches for **learning** a system fairly

Machine Learning

- Machine Learning: machine **learns** patterns from data for itself
 - No rules explicitly given
- Extremely successful recently*
 1. Big data
 2. Fast computers

*In some domains



Ethical Machine Learning?

- Machine learning can have high impact
- Used for high-stakes decisions
- Small, ubiquitous interactions

Ad related to latanya sweeney ⓘ

[Latanya Sweeney Truth](#)

www.instantcheckmate.com/

Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

[Latanya Sweeney](#)

Public Records Found For: **Latanya Sweeney**. View Now.

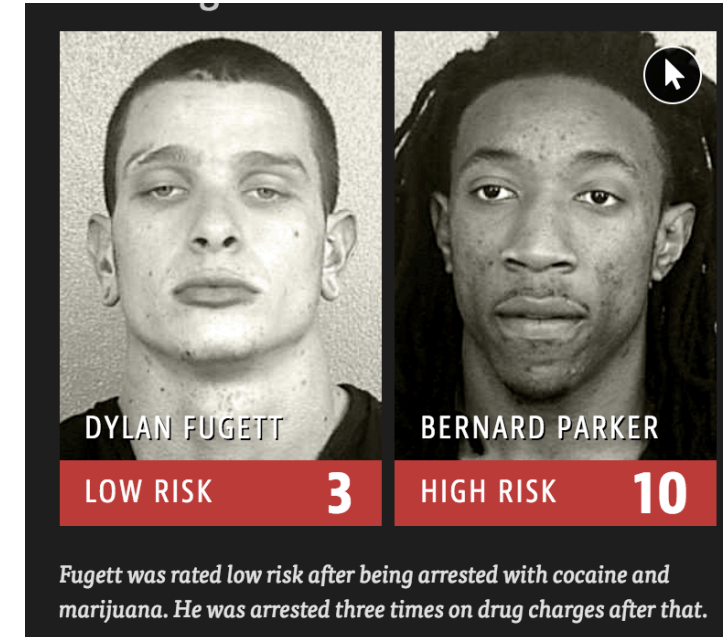
www.publicrecords.com/

[La Tanya](#)

Search for La Tanya Look Up Fast Results now!

www.ask.com/La+Tanya

Ethical machine learning matters in **high-stakes** domains



Fairness in Machine Learning – Two Ideas

- **Group fairness**

- Don't discriminate unnecessarily between **protected** groups (race, gender, sexuality, religion, etc.)

- **Individual fairness**

- Treat similar individuals similarly

Example: Online search engine results

Ad related to [latanya sweeney](#) ⓘ

[Latanya Sweeney Truth](#)
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access Full Background Checks Instantly.

www.instantcheckmate.com/

[Latanya Sweeney](#)

Public Records Found For: **Latanya Sweeney**. View Now.

www.publicrecords.com/

[La Tanya](#)

Search for La Tanya Look Up Fast Results now!

www.ask.com/La+Tanya

The screenshot shows a user profile for Latanya Sweeney on the InstantCheckmate website. The profile includes personal information, a location, related persons, marriage/divorce records, criminal history, licenses, and sex offenders. The criminal history section is highlighted, showing a table of possible matching arrest records, which is currently empty.

checkmate DASHBOARD EDIT ACCOUNT INFO LOGOUT

LATANYA SWEENEY
1420 Centre Ave
Pittsburgh, PA 15219
DOB: Oct 27, 1968 (53 years old)

Personal
Name, aliases, birthdate, phone numbers, etc.

Location
Detailed address history and related data, maps, etc.

Related Persons
Known family members, business associates, roommates, etc.

Marriage / Divorce
Marriage and divorce records on file.

Criminal History
Arrest records, speeding tickets, mugshots, etc.

Licenses
FAA licenses, DEA licenses, Other Licenses, etc.

Sex Offenders
Sex offenders living near Latanya Sweeney's primary location.

Criminal History Rate This Content: ☆☆☆☆☆

This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.

We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Sweeney has never been arrested. It simply means that we were not able to locate any matching arrest records in the data that is available to us.

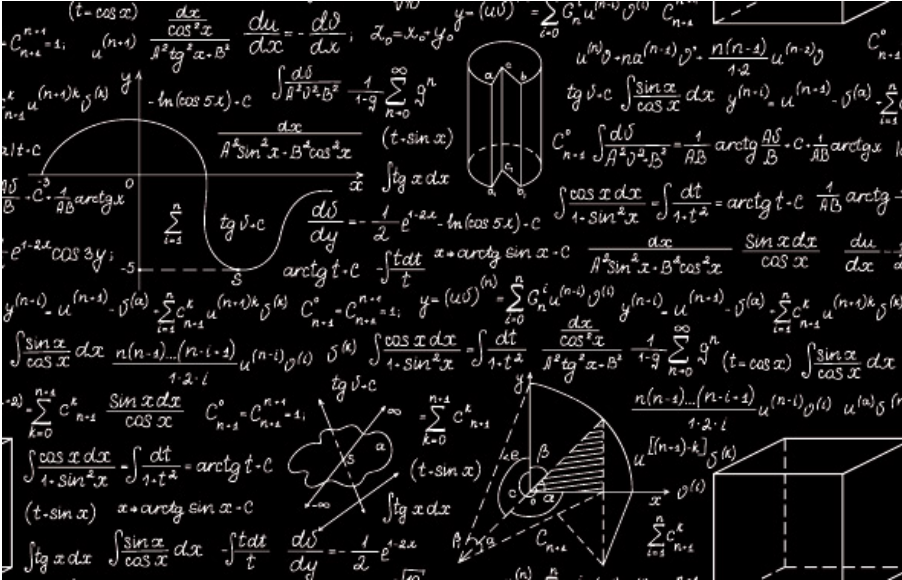
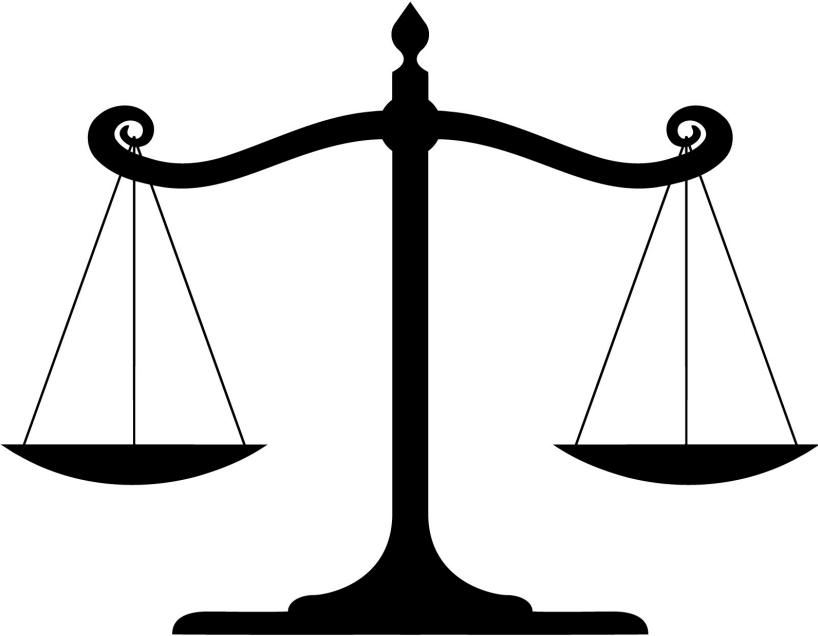
Possible Matching Arrest Records

Name	County and State	Offenses	View Details
No matching arrest records were found.			

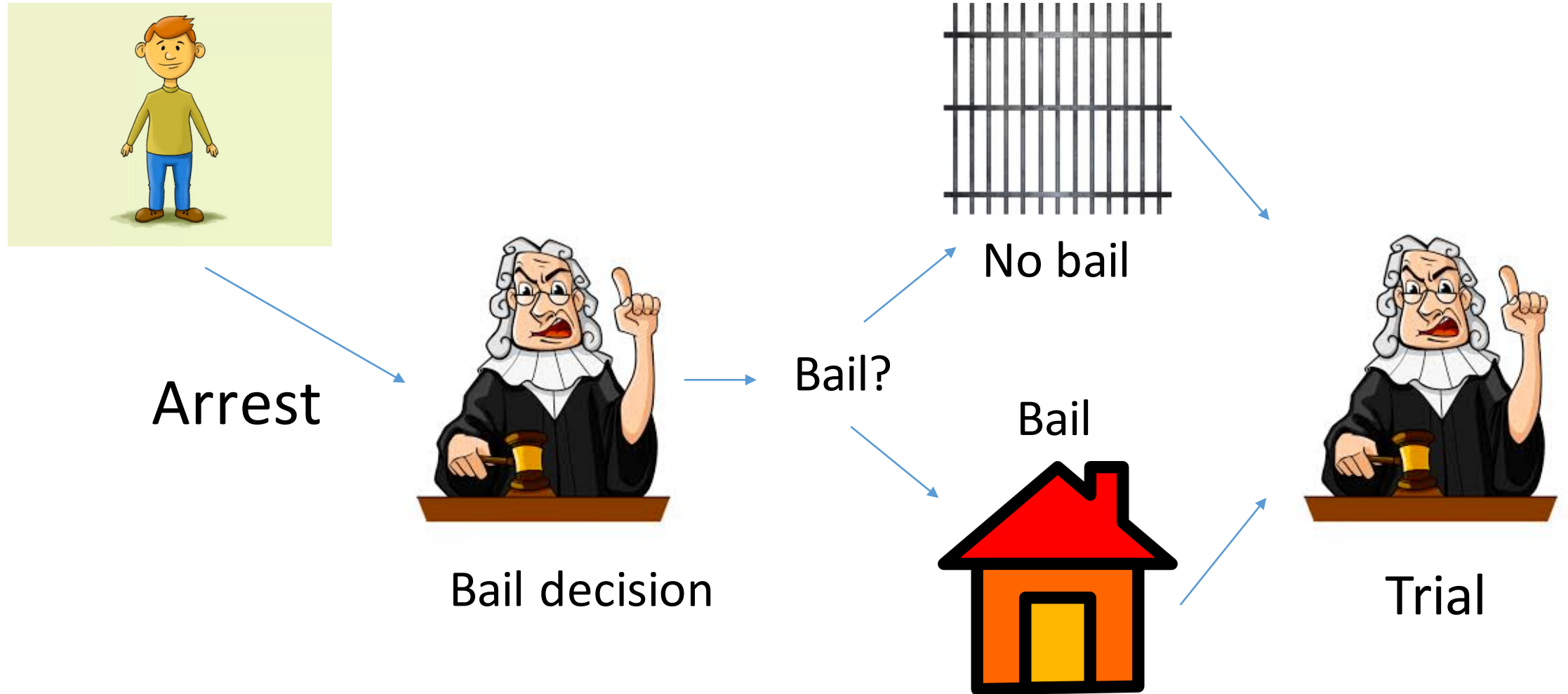
Example: Online search engine results

- Sweeney found that “criminal record” ads were more likely to show for names commonly given to black children than white ones
- Why did this happen?
- Who is responsible?
- How to regulate?

DEFINITIONS OF FAIRNESS: CLASSIFICATION



Example: Recidivism Prediction

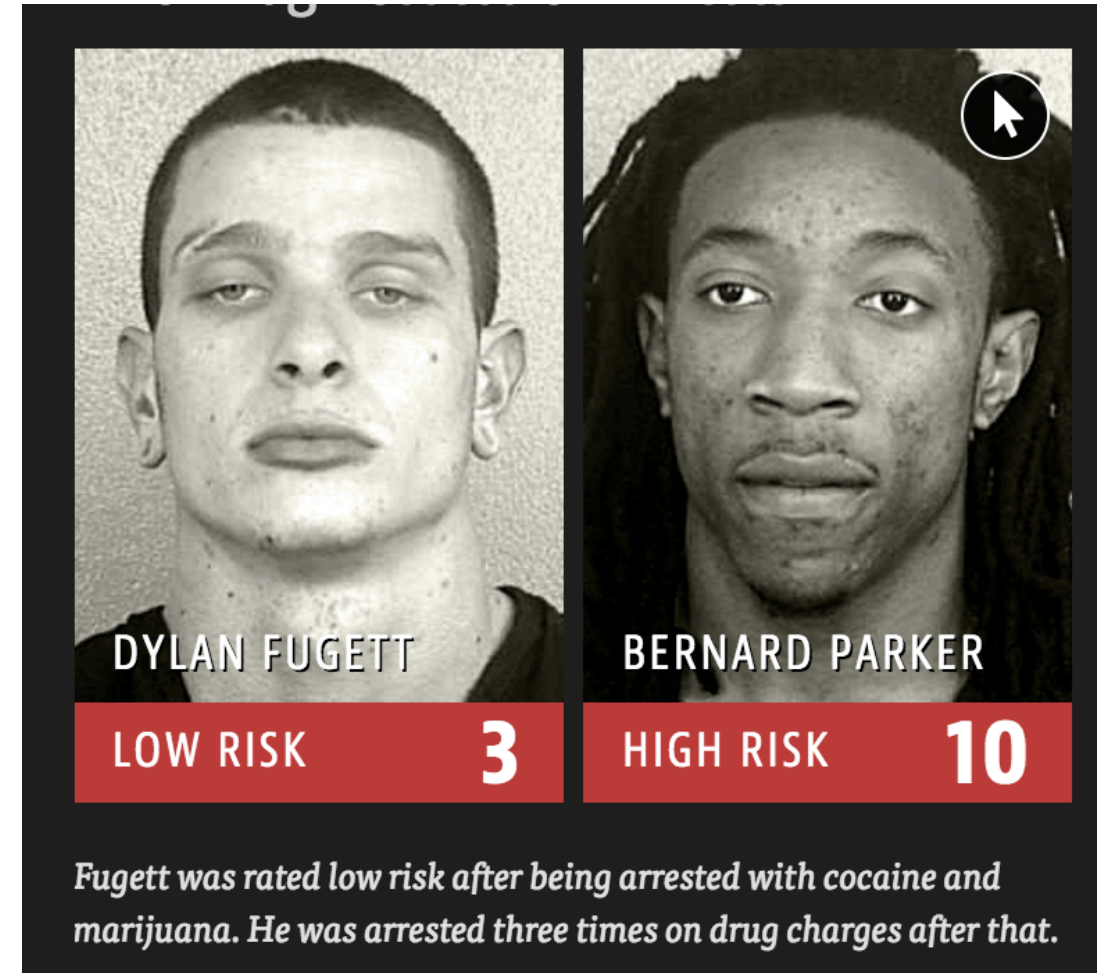


Example: Recidivism prediction

- Bail assignment task: Given some arrested defendant, predict if they will **recidivate**
- High-stakes task
 - No bail: can lose job, hurts family, more likely to plead guilty
- Machine learning tools have been developed to assist judges
 - These tools can be more accurate than judges!

ProPublica Investigation (COMPAS)

- ProPublica studied COMPAS predictions for 7000+ defendants in Florida (2013-4)
- Different **types of errors** made on black and white defendants
- Black: more often wrongly **denied** bail
- White: more often wrongly **given** bail



Fairness is Impossible (sort of)

- ProPublica claimed COMPAS violated a specific fairness **definition**
- Northpointe responded: COMPAS satisfied a **different** fairness definition
- It turns out that these were **incompatible definitions**

Many Definitions of Fairness

- For a label Y , a prediction p , and a sensitive attribute A

Fairness Metric Name	This variable ...	Is independent of A given...
Demographic Parity	p	
Equalized Odds	p	Y
Equal Opportunity	p	$Y = 1$
Fair Calibration	Y	p
Fair Subgroup Accuracy	$Y = p$	
... and so on		

Further info: “21 fairness definitions and their politics”, Arvind Narayanan
<https://speak-statistics-to-power.github.io/fairness/>

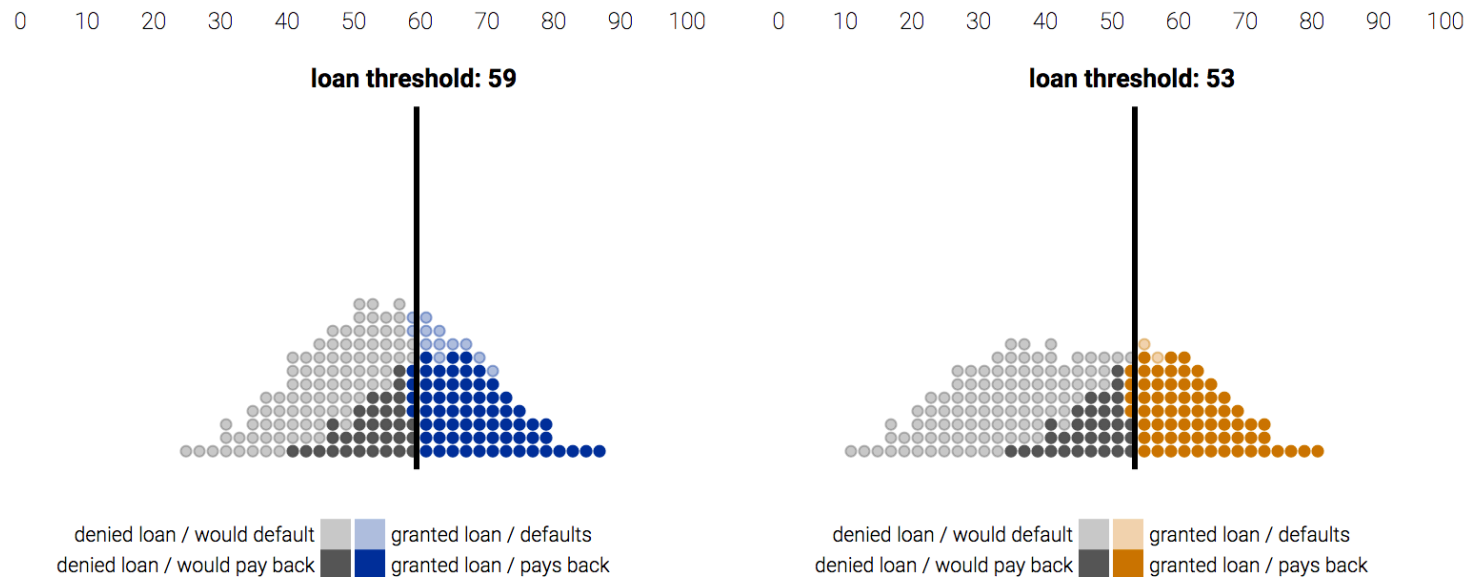
How to Learn “Fairly”

- Naïve Idea: **remove A from your dataset**
- This fails if A is encoded in your other features!
- E.g. A is race, but dataset also contains postal code



How to Learn “Fairly”

- Usually, some kind of constrained optimization or regularization
- There is a fairness-accuracy **tradeoff**



“Attacking discrimination with smarter machine learning” – Wattenberg et al

<https://research.google.com/bigpicture/attacking-discrimination-in-ml/>

What if you don't like tradeoffs?

- In some applications, tradeoffs with accuracy are highly undesirable
- Instead:
 - Collect more data on disadvantaged group
 - Collect more attributes
 - Model groups separately
 - ???

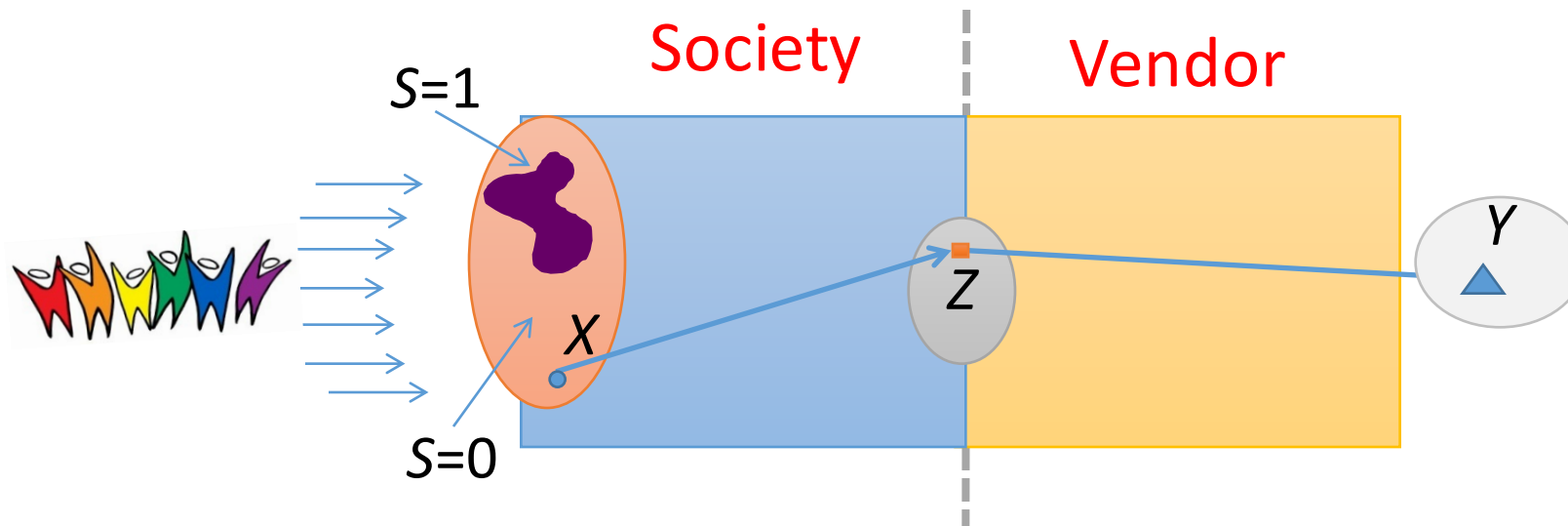
“Why Is My Classifier Discriminatory?”, Chen et al. [2018]

“Decoupled classifiers for fair and efficient machine learning”, Dworkin et al. [2017]

The One True Fairness Definition

- Probably doesn't exist

FAIRNESS IN REPRESENTATIONS



Gender Bias in Word Embeddings

- Experiment: translate English sentence to gender-neutral language and back, using Google Translate

(try live demo)

Gender Bias in Word Embeddings

- Experiment: translate English sentence to gender-neutral language and back, using Google Translate

“She is a doctor” → “O bir doktor” → “He is a doctor”

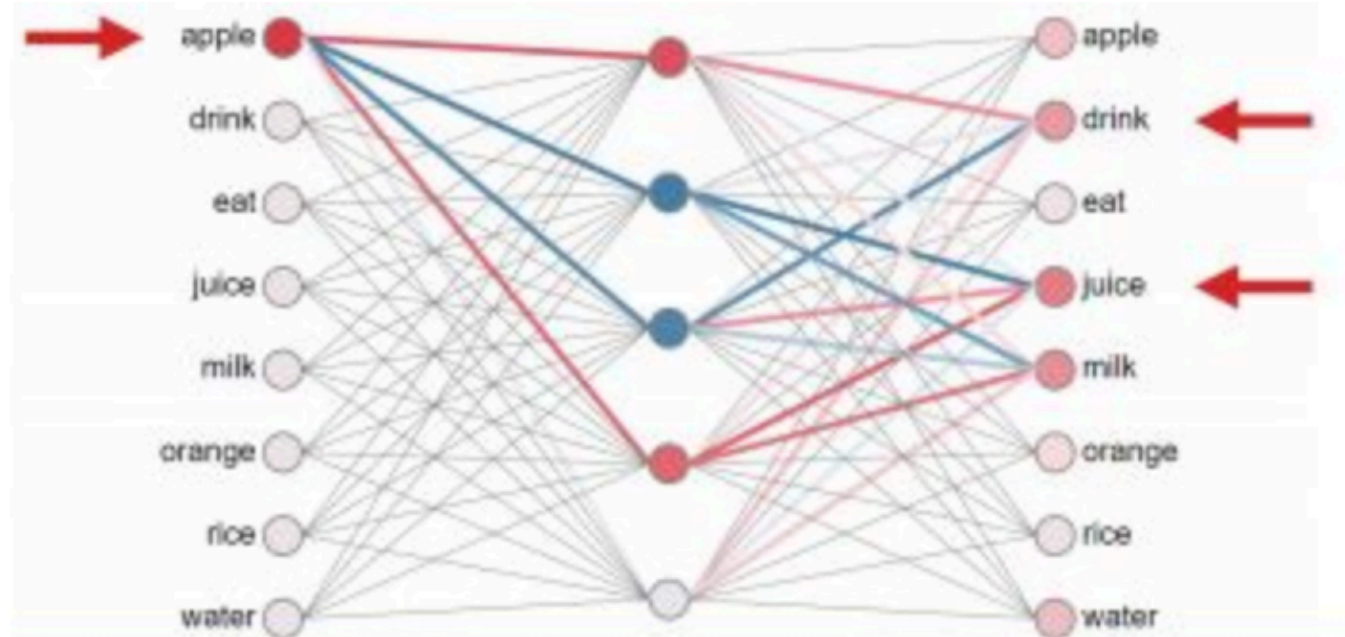
“He is a nurse” → “O bir hemsire” → “She is a nurse”

- Try it yourself!
- The AI only knows probabilities: given “_____ is a doctor”, “He” **occurs more commonly** than “She” in the training data

Example: Word Embeddings

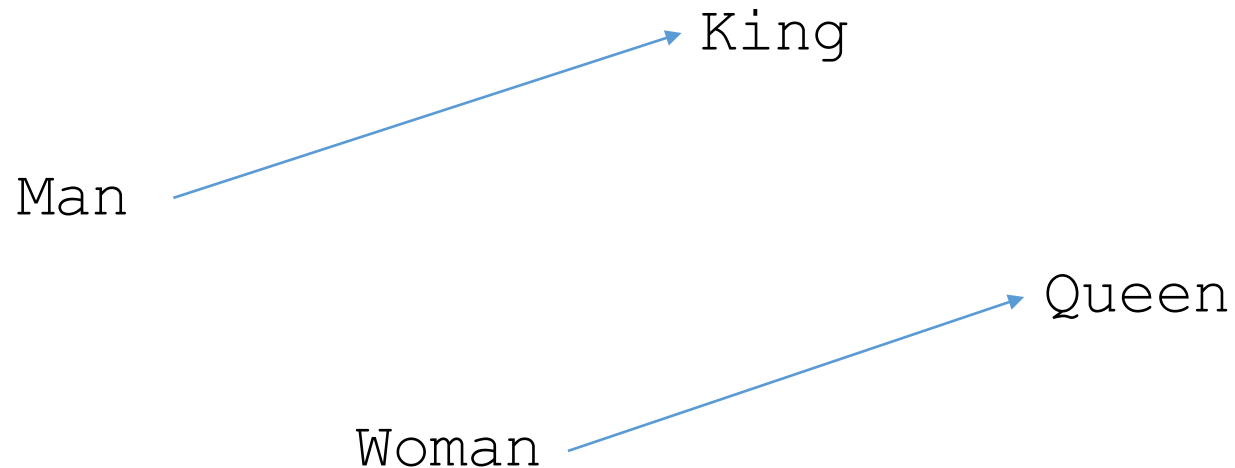
(if time permits)

- For computers to understand words, we need to turn them into numbers first
- Using a **neural network**, we can learn **word embeddings** – numbers that represent words



Example: Word Embeddings

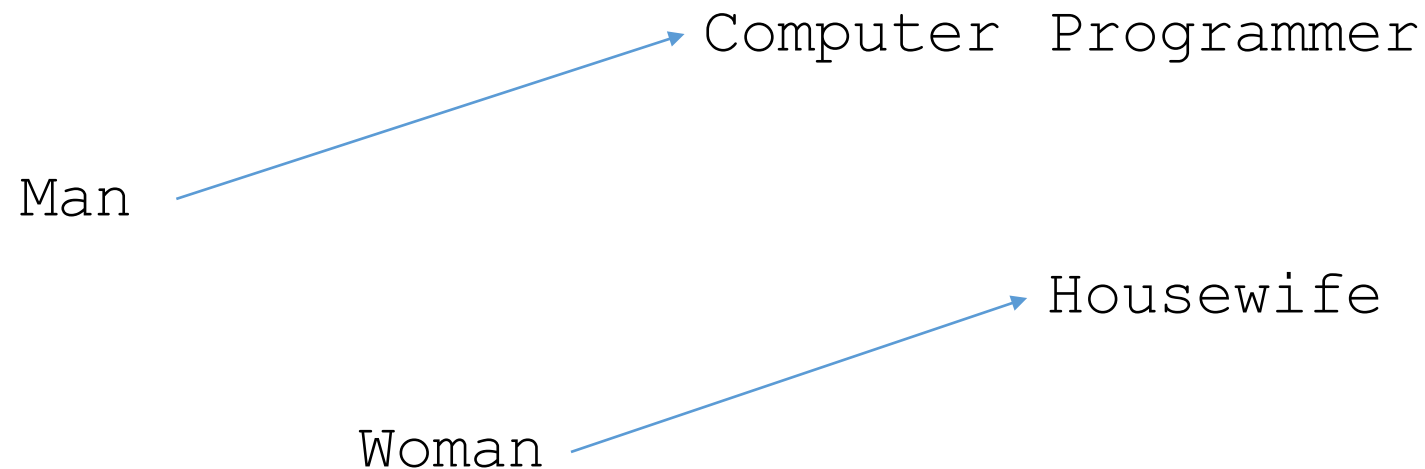
- We can now use these embeddings in other language applications
- Using **analogies** (embedding arithmetic), we can check that they make sense



$$\text{King} - \text{Man} = \text{Queen} - \text{Woman}$$

Gender Bias in Word Embeddings

- However, we find some of these analogies contain gender bias
- Remember: the computer learns all of this on its own, given just a large body of text

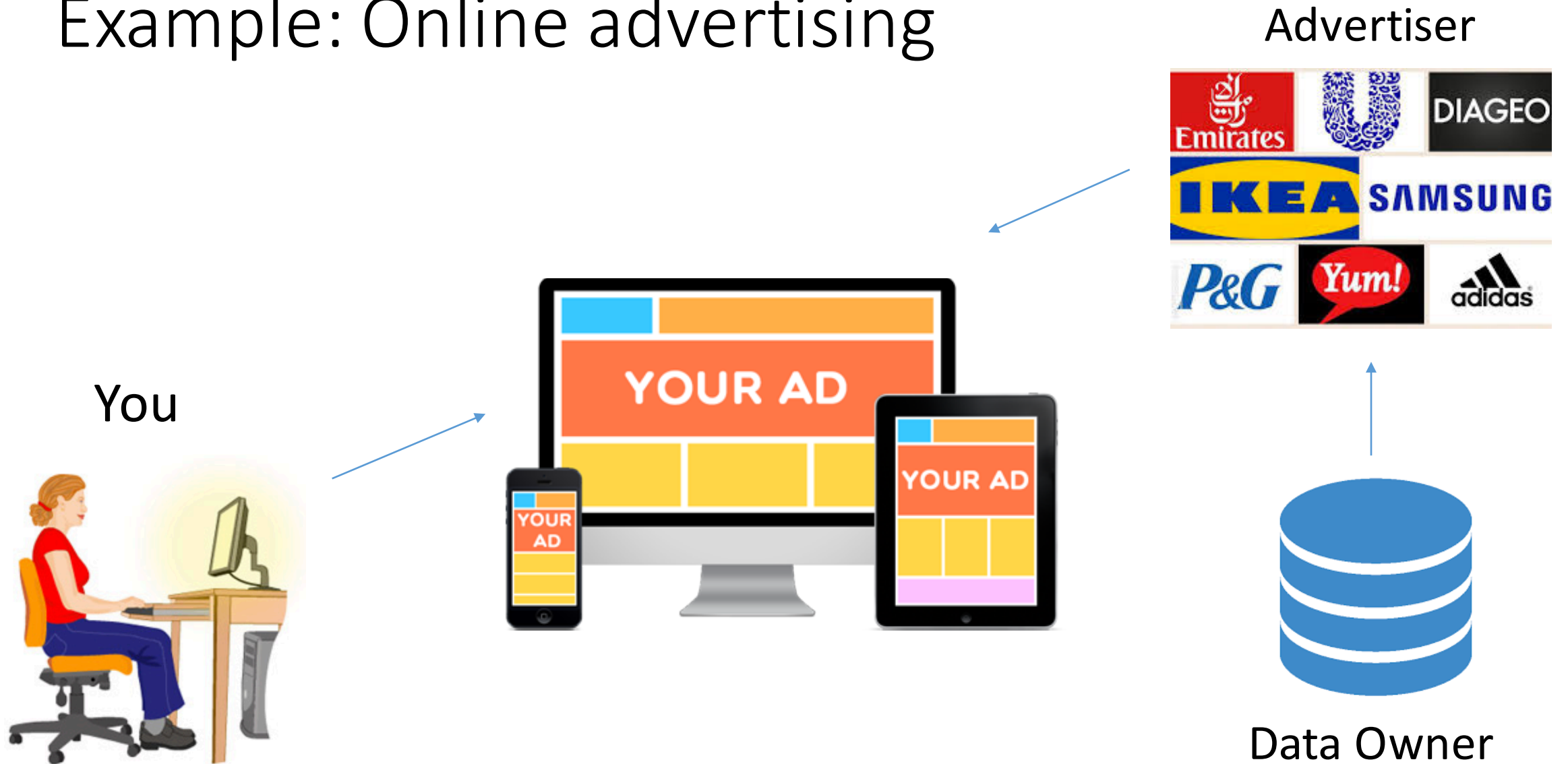


$$\text{Programmer} - \text{Man} = \text{Housewife} - \text{Woman}$$

Example: Online advertising



Example: Online advertising



Example: Online advertising

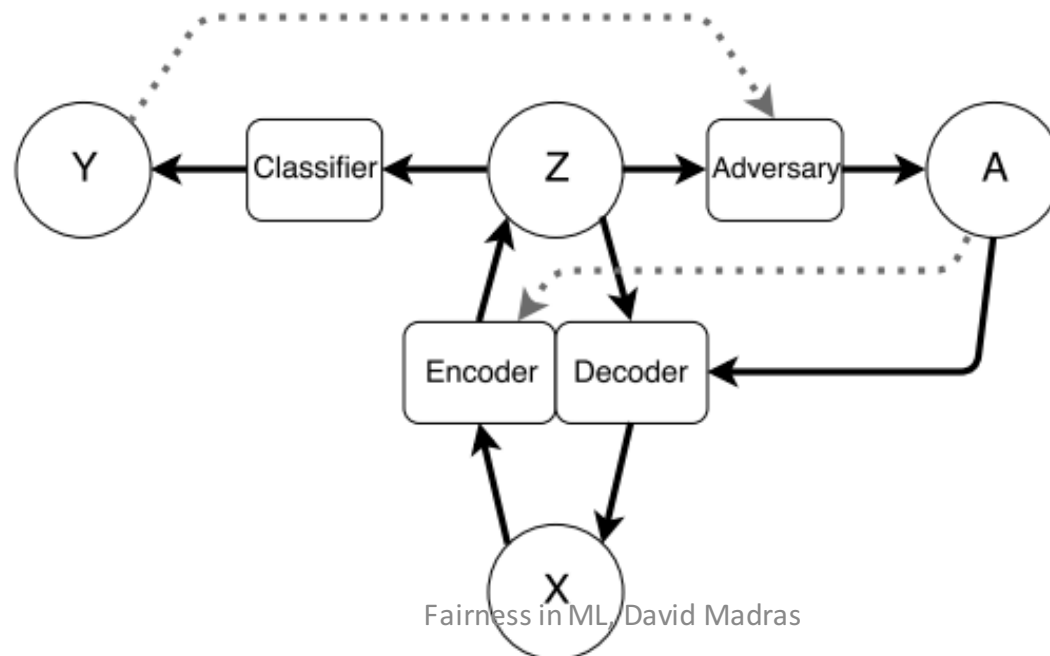
- Online **advertisers** show everyone different ads
- They use data provided by **data owners** on the users
- Using machine learning, they identify which users are most likely to click on each ad
- This can lead to unfairness:
 - Men more likely to see ads for high-paying jobs
 - Black people more likely to see ads for bad lines of credit

Fair and Transferable Representations

- In our work, we focus on the **data owner's** role in fairness
- What if the data owner can alter the data?
- Maybe there's a way to change the data so that:
 - The advertiser can still make good predictions on many tasks
 - The advertiser is **guaranteed** to make fair predictions

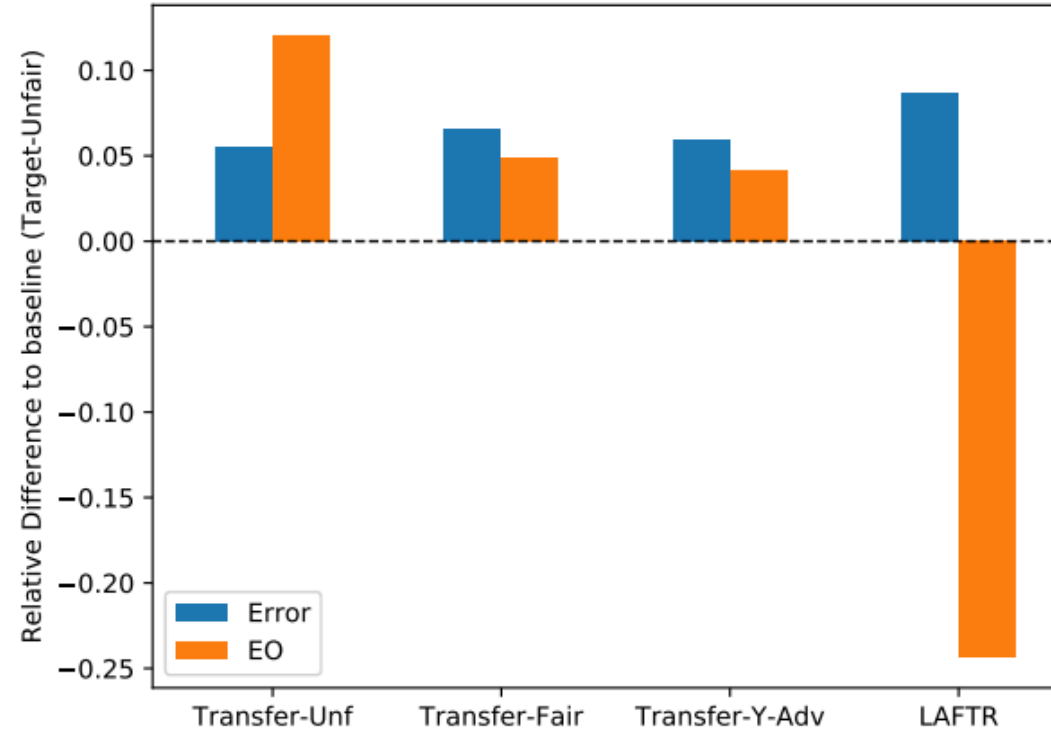
Our work: Learning Adversarially Fair and Transferable Representations (LAFTR)

- We use three neural networks, each simulating a role:
 1. The data owner: wants to make the data fair
 2. An **indifferent** advertiser: doesn't care about fairness, only business
 3. A **malicious** advertiser: *only* wants to be unfair



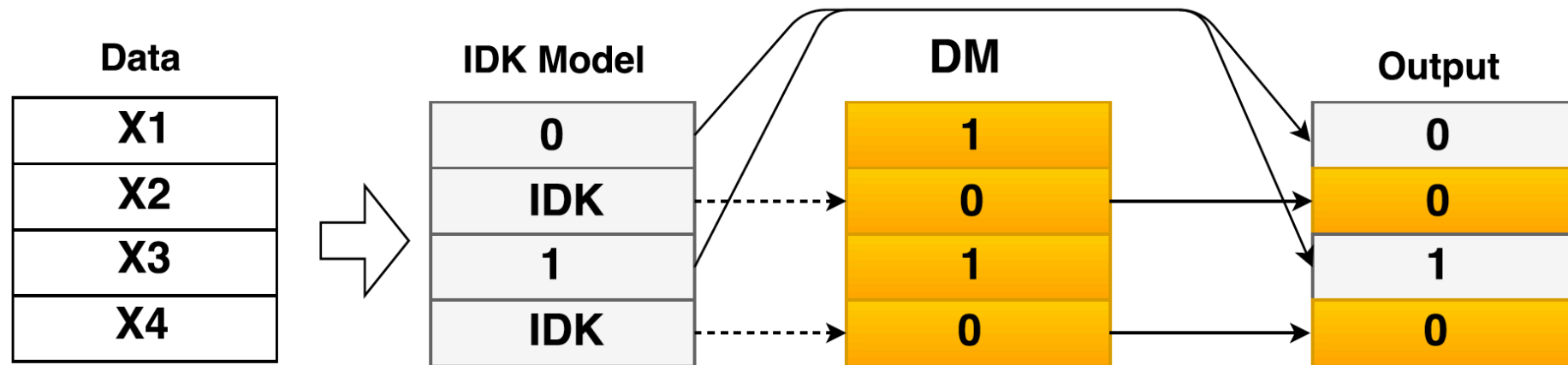
LAFTR Results

- Slight loss in accuracy, big gain in fairness
- Generalizes to unseen tasks



Fairness in Machine Learning – What’s next?

- Working with external decision-makers
 - In many real applications, machine learning model interacts with an external decision maker
 - Must learn to **defer** on some cases



“Predict Responsibly: Increasing Fairness and Accuracy by Learning to Defer”, Madras, Pitassi, Zemel, 2017

Fairness in Machine Learning – What’s next?

- Fairness when learning from **biased data**
 - What if the mechanism which produced your dataset is biased?
 - “Residual Unfairness in Fair Machine Learning from Prejudiced Data”, Kallus, Zhou [2018]
- Fairness under repeated decision-making
 - If biased decisions are made repeatedly in the same environment, **feedback loops** can occur
 - In predictive policing:
 - “Runaway Feedback Loops in Predictive Policing”, Ensign et al. [2017]
 - In recommender systems:
 - “Fairness Without Demographics in Repeated Loss Minimization”, Hashimoto et al. [2018]

In Summary

- Fairness in classification, representation
 - Advertising, search, criminal justice, finance, language processing
- Many useful definitions, none are perfect
- Can also think about fairness as a system problem
- Other important topics: transparency, accountability, safety

Thank you!

Collaborators:



Elliot Creager



Toni Pitassi

Fairness in ML, David Madras



Rich Zemel