Tackling unwanted bias in AI applications

Digitax Conference on computational taxation: in search for fairness and transparency in tax technology, 23 September 2021

Toon Calders

toon.calders@uantwerpen.be

www.ata-digitax.com

Antwerp Tax Academy
University of Antwerp

Overview

- Sources of Bias via examples

- Measuring and detecting bias in data and models
  - Group fairness measures
    - Demographic parity
    - Equal odds
    - Incompatibility
  - Individual fairness

- Avoiding bias in models
  - Cleaning data
  - Fairness by design
  - Post-processing

Recent examples

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Racial bias found in health care risk algorithm

- October 2019: used on more than 200 million people in US
  - predict which patients would likely need extra medical care
  - identify which patients will benefit from "high-risk care management" programs: access to specially trained nursing
- Heavily favored white patients over black patients.
  - Race wasn't a variable, but **healthcare cost history**. […]
  - Black patients incurred lower health-care costs than white patients with the same conditions on average.

Recent examples

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Amazon ditched AI recruiting tool that favored men for technical jobs

https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Amazon's recruitment tool based on AI

- In 2015, Amazon realized that their algorithm used for hiring employees was biased against women
  - algorithm was based on the number of resumes submitted over the past ten years
  - most of the applicants were men, it was trained to favor men over women.
- It penalized résumés that included the word "women's", as in "women's chess club captain". Downgraded graduates of two all-women's colleges.

https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

Word Embeddings Capture Gender Bias

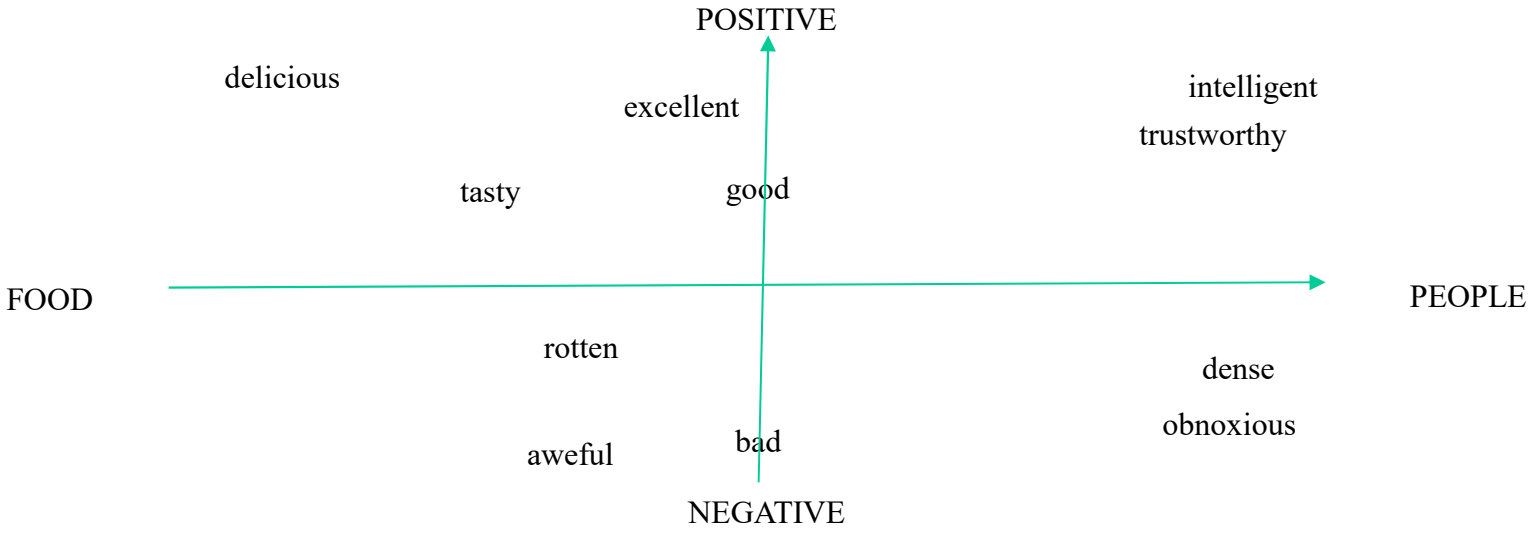
- Word embedding
  = representation of words/texts as a vector of numbers

- Banana → (0.3, 5.8, 7.3, 0.1)
- Father → (0.4, 0.7, 1.2, 0.4)
- Baby → (0.3, 0.6, 1.5, 3.0)
- …

- **Why?** Hundreds of algorithms work with numbers. Word2Vec is like an "adaptor"

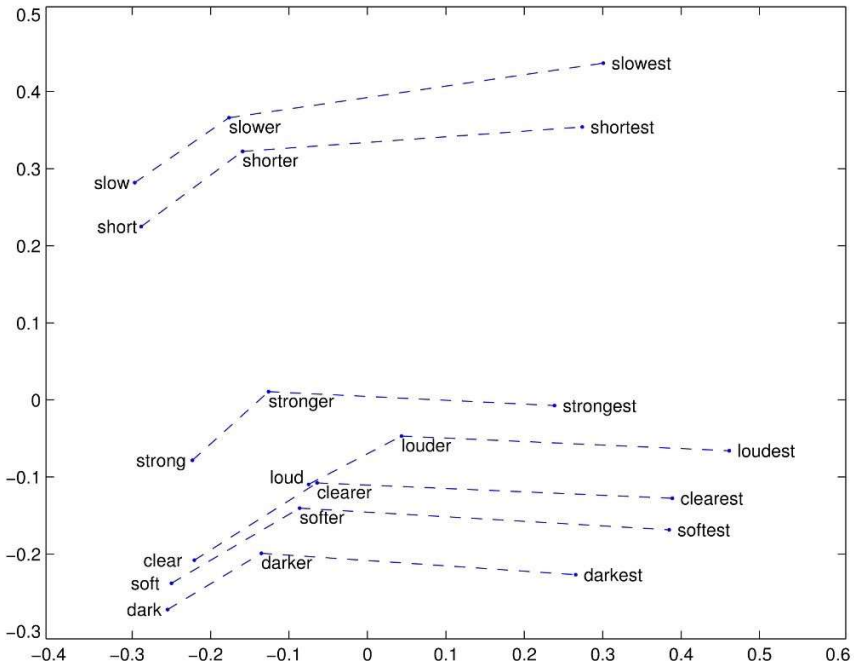
Embeddings are Not Random

- Words used in similar contexts should have similar vectors

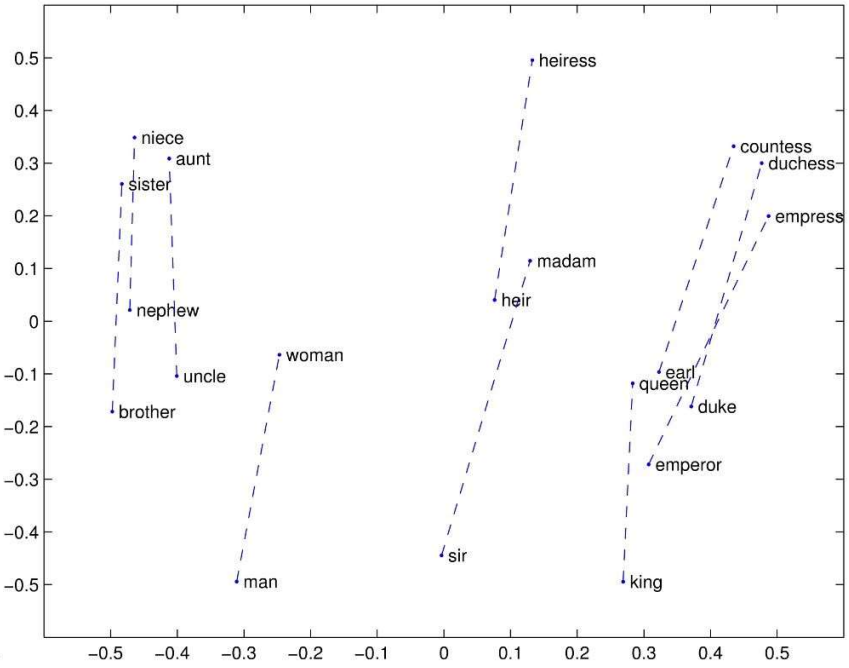


- These embeddings are *learned*
  - Optimization problem; words close in *use* should be close in *distance*

Word2vec Captures Semantic Information



$$\overrightarrow{slower} - \overrightarrow{slow} \approx \overrightarrow{shorter} - \overrightarrow{short}$$

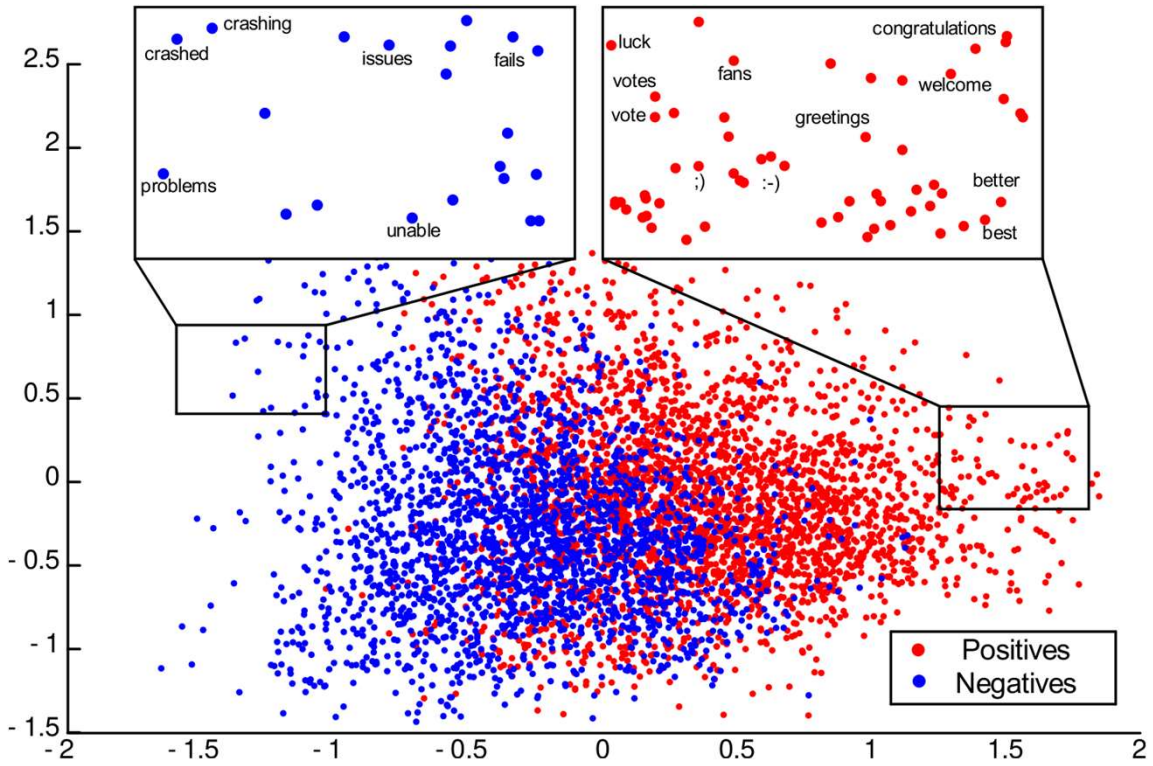


$$\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{king} - \overrightarrow{queen}$$

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

Amazing Applications: Sentiment Analysis



https://www.micc.unifi.it/projects/advanced-web-applications/sentiment-analysis-of-tweets-from-twitter/

BUT … Also Captures Cultural Biases

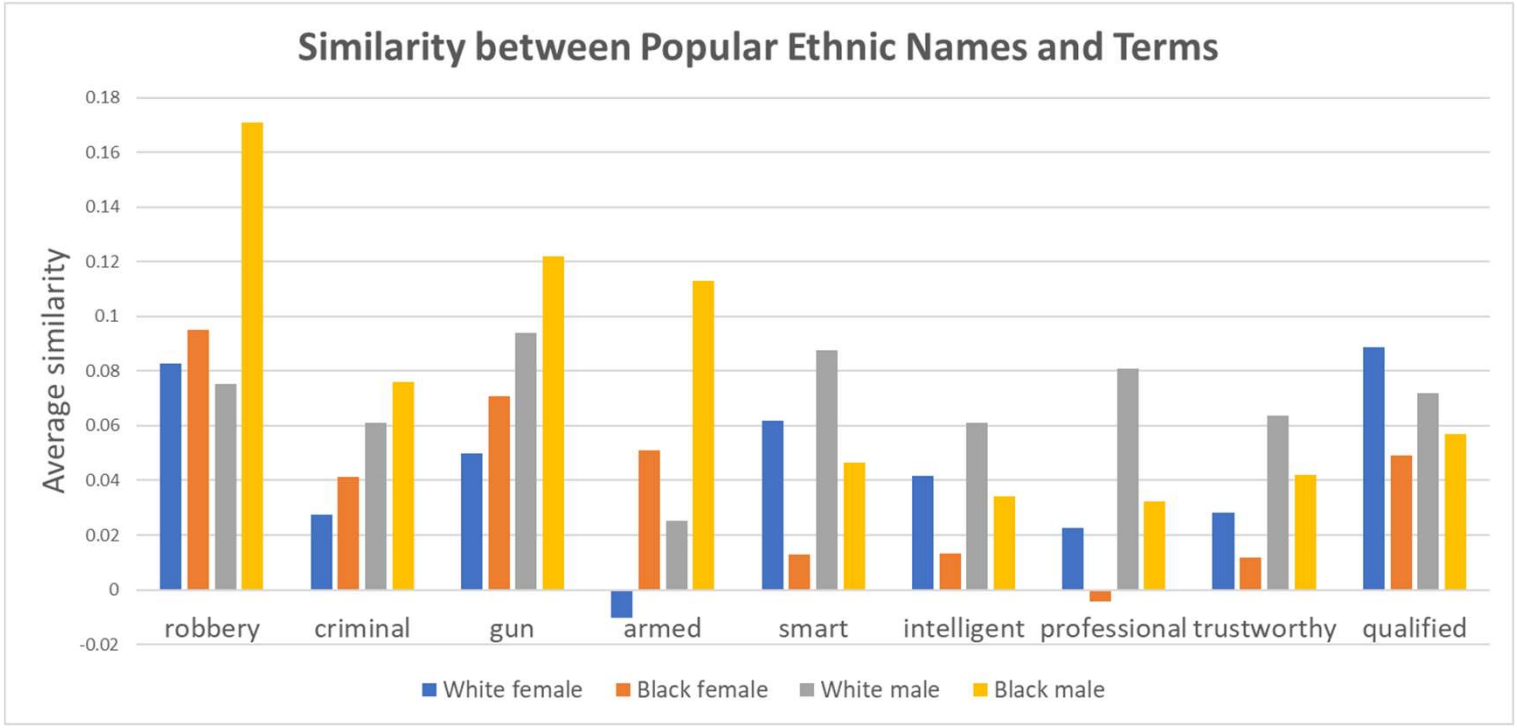
$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \ \ldots$$

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

Imagine Word2Vec screening your CV,
finding the perfect job for you …

BUT … Also Captures Cultural Biases



Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

Recent examples

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Amazon ditched AI recruiting tool that favored men for technical jobs

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

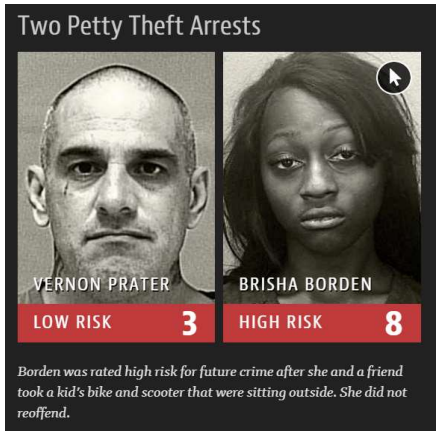
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence

COMPAS



- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) tool to predict risk of recidivism
  - Label: was there a new arrest within two years?
  - Data: pending charges, prior arrest history, previous pretrial failure, residential stability, substance abuse, …

ProPublica Study (2016)

- ProPublica study showed that the errors made by the model are highly biased:

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Recent examples

Bias as an artifact of the way data was labelled

COMPUTING

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

By Starre Vartan on October 24, 2019

Bias learned from biased data sources

Amazon ditched AI recruiting tool that favored men for technical jobs

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Bias introduced by the algorithm

https://towardsdatascience.com/real-life-examples-of-discriminating-artificial-intelligence-cae395a90070

Antwerp Tax Academy
University of Antwerp

Overview

- Sources of Bias and examples

- Measuring and detecting bias in data and models
  - Group fairness measures
    - Demographic parity
    - Equal odds
    - Incompatibility
  - Individual fairness

- Avoiding bias in models
  - Cleaning data
  - Fairness by design
  - Post-processing

ProPublica Study (2016)

- ProPublica study showed that the errors made by the model are highly biased:

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

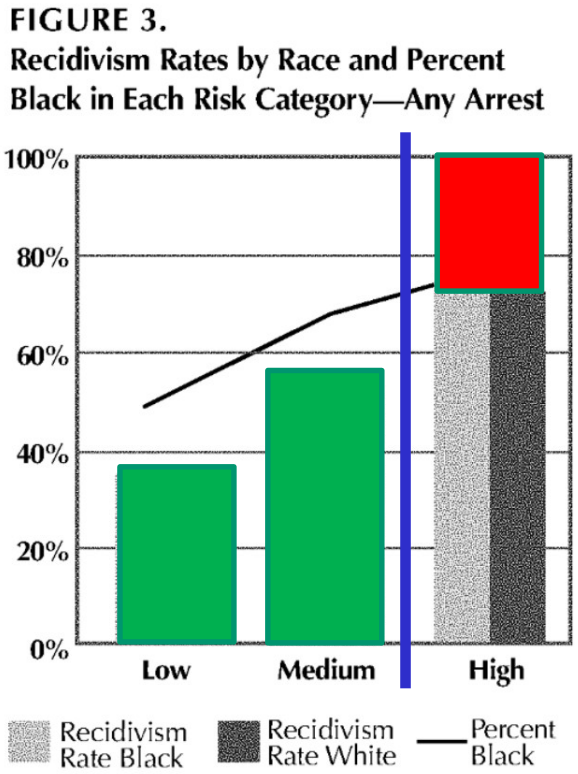
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Fair or Unfair?

- Northpointe's defense:
  - scores are *calibrated*

- All false positives are in High risk
- All false negatives in other groups

- Black is relatively more frequent in High than in Low and Medium



Anthony W. Flores et al., False Positives, False Negatives, and False Analyses., 80 Fed. Probation 38 (2016)

Pro Publica: *Equal opportunity*

- If you deserve to stay in prison, it shouldn't matter whether you're black or white
- If you deserve to be released, it shouldn't matter whether you're black or white

Northpointe: *Calibration*

- What it means to be a high/low risk should not depend on your ethnicity

Illustration: Calibrated

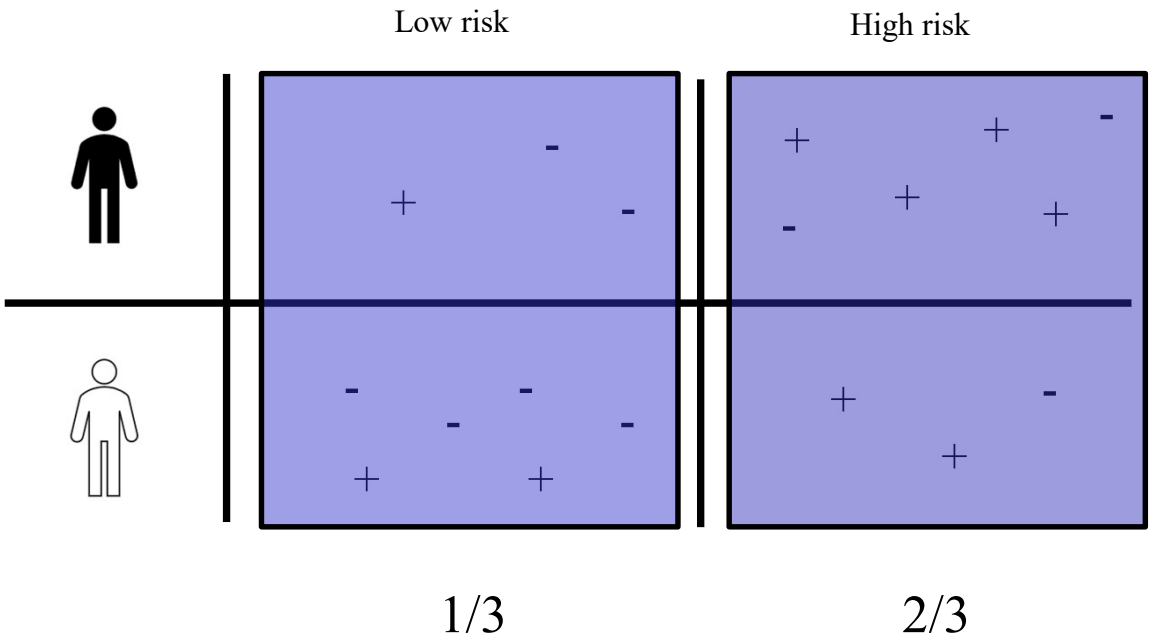
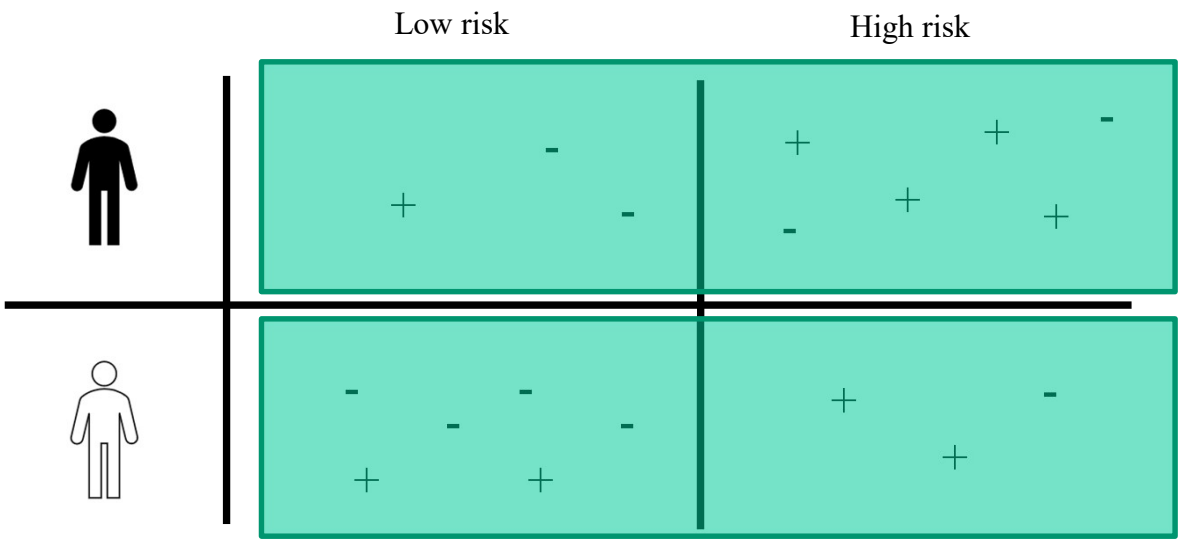


Illustration: But Not Equal Opportunity



$P(H|+) = 4/5$

$P(H|-) = 1/2$

$P(H|+) = 1/2$

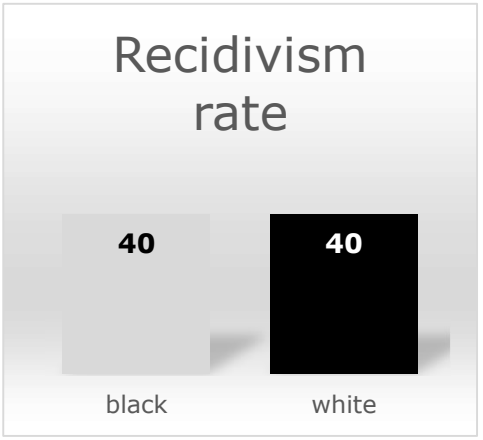
$P(H|-) = 1/5$

When is a score fair?

- However: it can be proven that both together can only be realized under *very exceptional conditions*:

Perfect predictability

or



Equal base rates

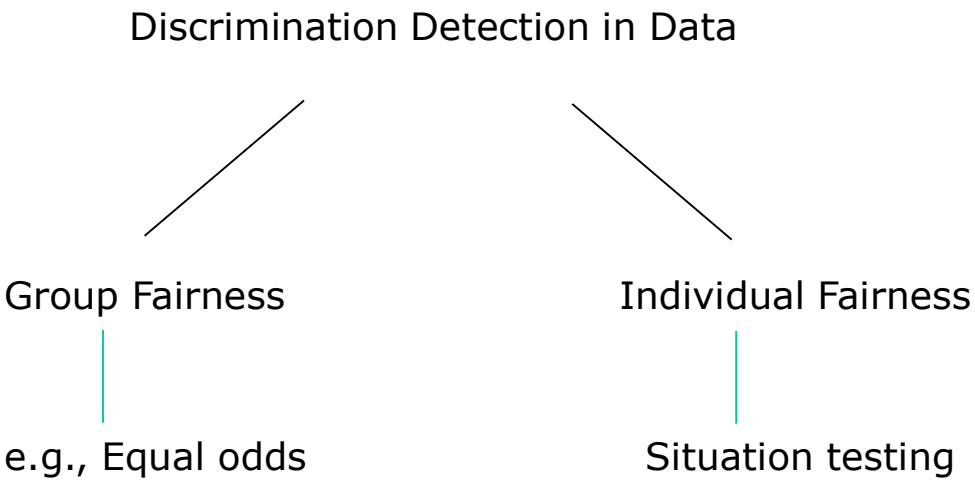
Kleinberg et al. ITCS 2017

What does it mean to be fair?

- ProPublica study and Northpointe's response raises intriguing question:
  - *How can we define when a decision procedure is fair?*

- **Conclusion**:
  - No *universal* definition of fairness ; *situation-dependent*

Individual Fairness Measures



Situation Testing

- Is an individual discriminated?
  - Look at similar instances

Luong, B. T., Ruggieri, S., & Turini, F. (2011, August). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 502-510).

Situation Testing

- Is an individual discriminated?
  - Look at similar instances

Positive decision ration among…

Nearest female neighbours: 25%
Nearest male neighbours: 75%

→ Discrimination!

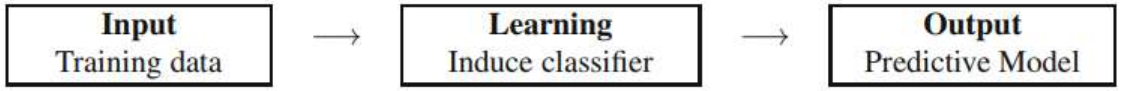
Luong, B. T., Ruggieri, S., & Turini, F. (2011, August). k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 502-510).

Overview

- Sources of Bias and examples
- Measuring and detecting bias in data and models
  - Group fairness measures
    - Demographic parity
    - Equal odds
    - Incompatibility
  - Individual fairness
- Avoiding bias in models
  - Cleaning data
  - Fairness by design
  - Post-processing

Methods for Removing Bias

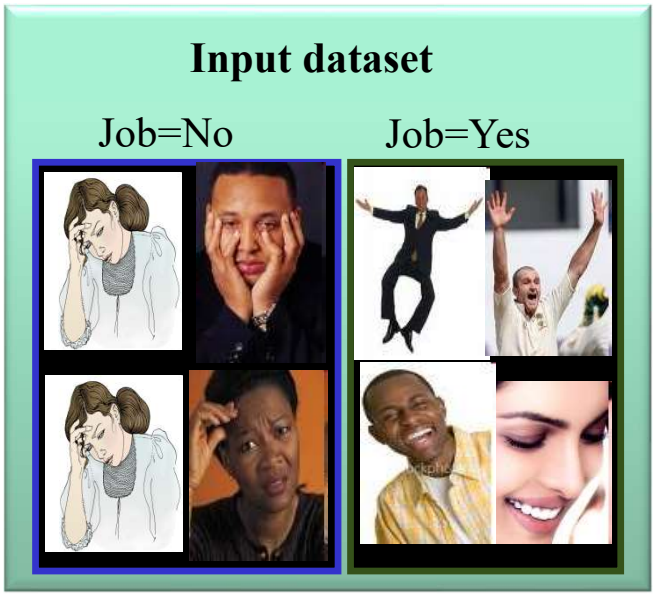
- We can divide fair classification algorithms:
  - Which measures do they target
  - Where in the process they act ?

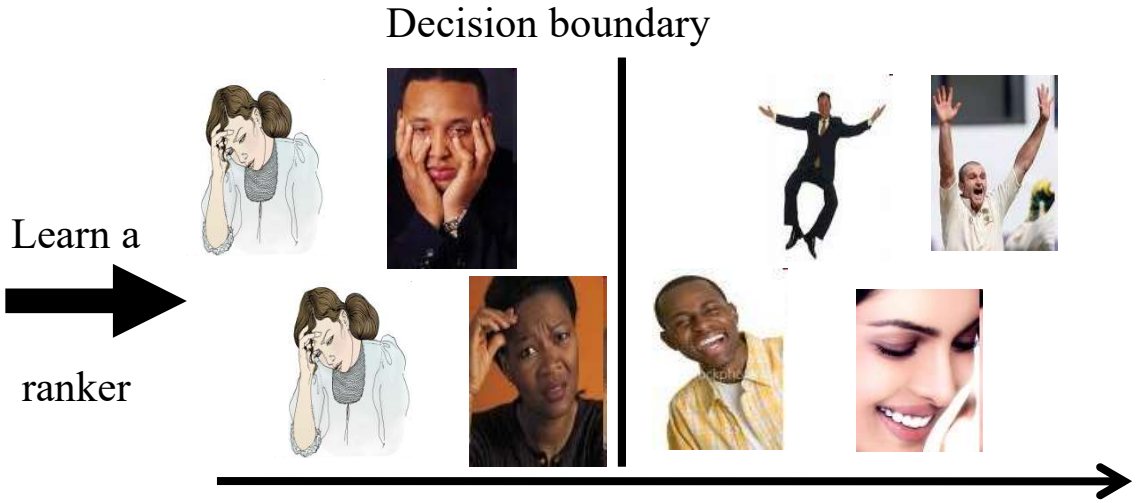


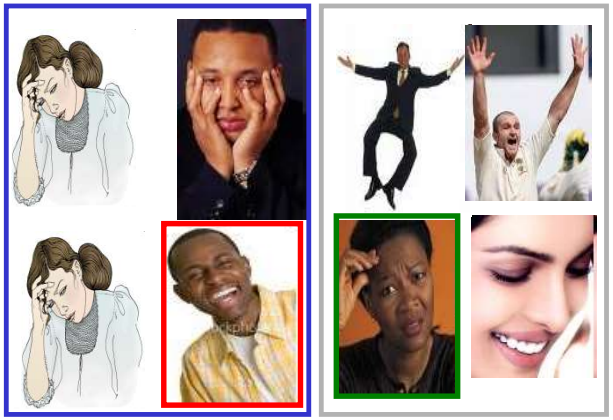
- Do they require the sensitive attribute for predicting?
  - ***All algorithms do require it at training time***

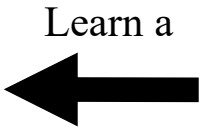
Kamiran, F., Calders, T., & Pechenizkiy, M. (2013). Techniques for discrimination-free predictive models. In Discrimination and Privacy in the Information Society (pp. 223-239). Springer.

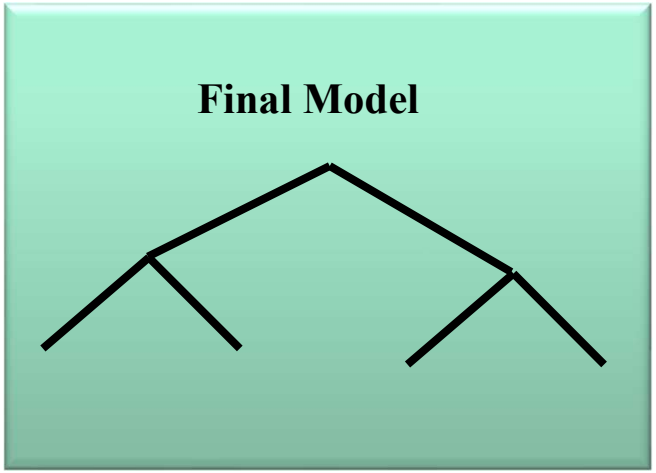
Clean up the Dataset: Massaging





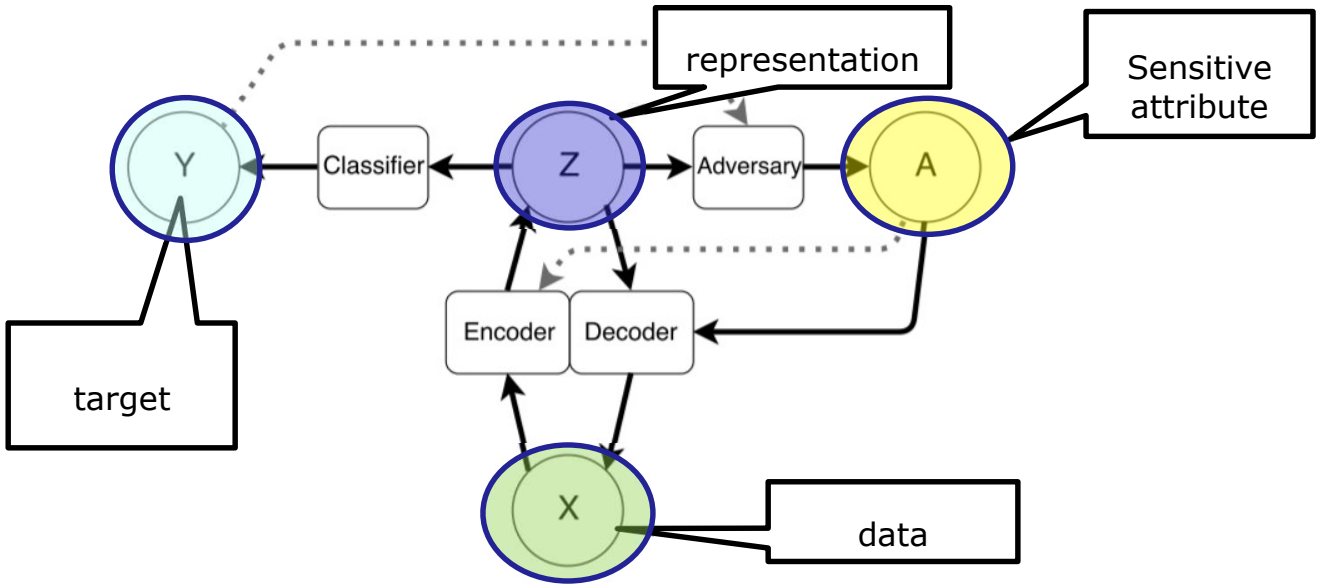






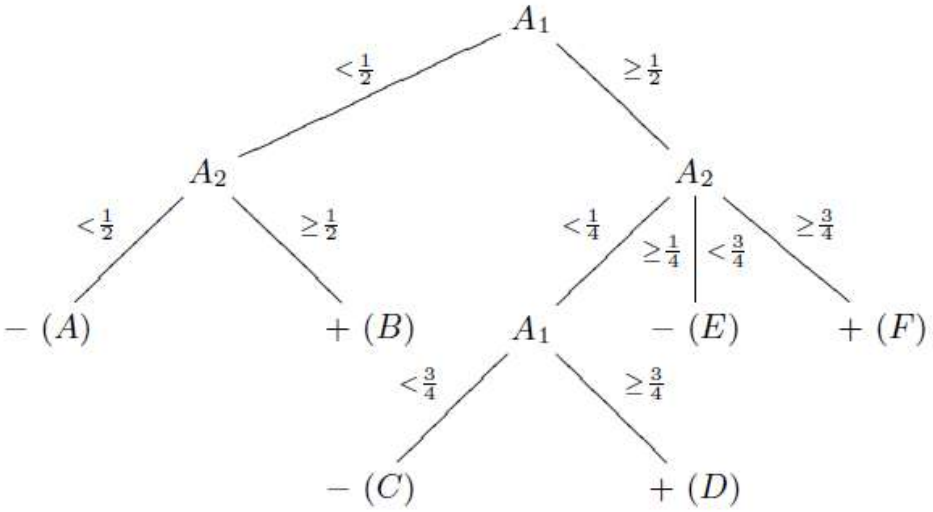
Example: Adversarial learning

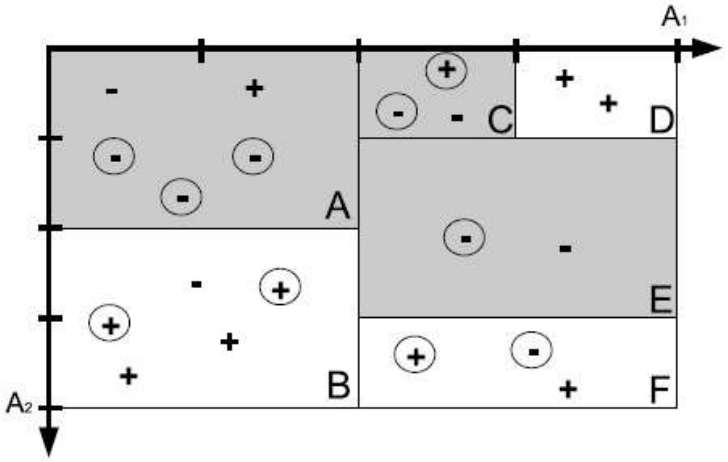
- Learn intermediate representation that allows to predict target but disallows inferring the sensitive attribute



Madras, Creager Pitassi, Zemel. Learning Adversarially Fair and Transferable Representations. ICML 2018

Post-Processing: Node Relabeling

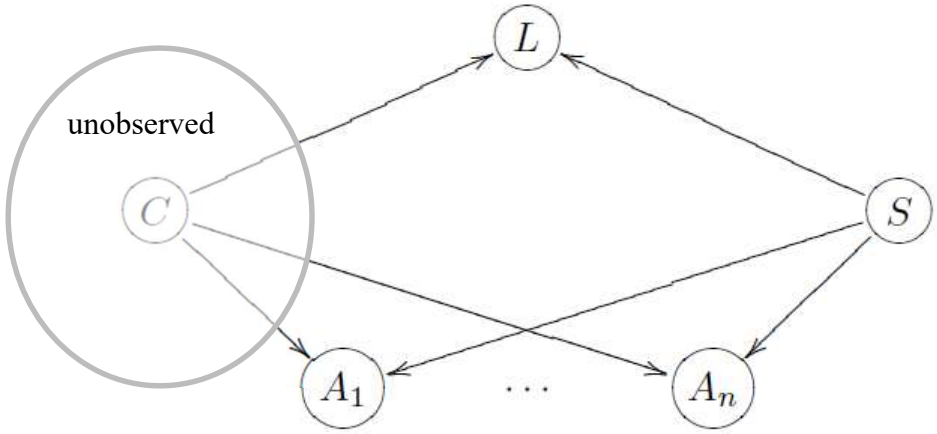




- Labels are assigned according to the majority class
- In node relabeling we change this strategy

Example: Reverse Engineering

- We assume the data is generated by the following model:



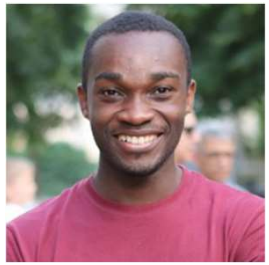
- C is the true label, L the label given in the data
- Use EM to find model that maximizes likelihood on the training data

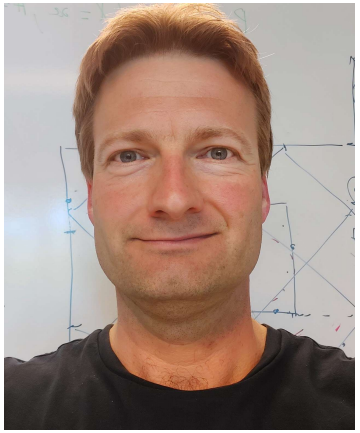
Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277-292.

Summary

- ML models are not fair "out of the box"
  - **Bias in, bias out!**
- **Fairness:** How to *build* models that optimally avoid certain types of bias?
- However, **partially inherent** to decision making!
  - Unequal base rates between groups provably lead to differences in treatment





Prof. dr. Toon Calders

Toon.calders@uantwerpen.be

University of Antwerp

Antwerp Tax Academy, DigiTax Research Center

Antwerp Tax Academy
University of Antwerp

DigiTax
Centre of Excellence