INTRODUCTION TO FAIRNESS

TONIANN PITASSI RICHARD ZEMEL

CSC 2541

SEPTEMBER 10, 2019

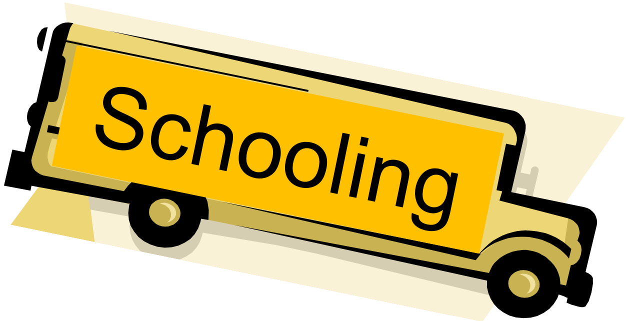
WHY WAS I NOT SHOWN THIS AD?

FAIRNESS IN AUTOMATED DECISIONS

Algorithmic unfairness: Algorithms are pervasive, high-stakes, high-impact

Need more than just "accuracy"

What's changed? Pervasiveness of ML & Attention to demographic criteria



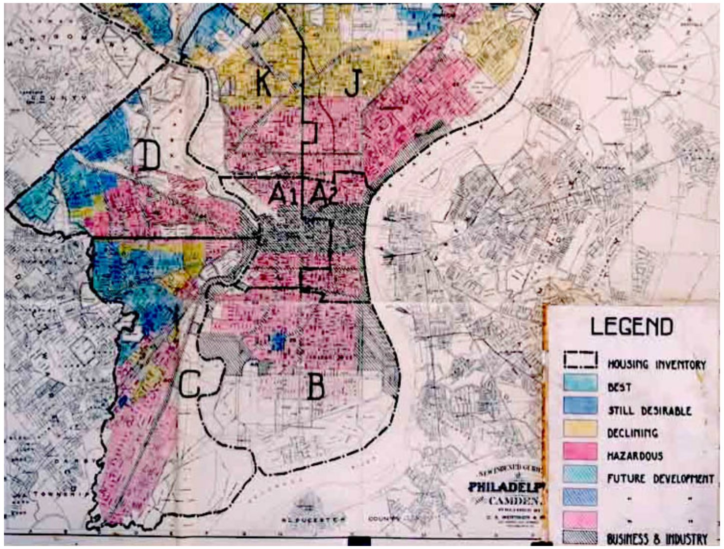
CONCERN: DISCRIMINATION

- Population includes minorities
  - Ethnic, religious, medical, geographic
- Protected by law, policy, ethics
- (If) we cannot completely control our data, can we regulate how it is used, how decisions are made based on it?

Forms of Discrimination

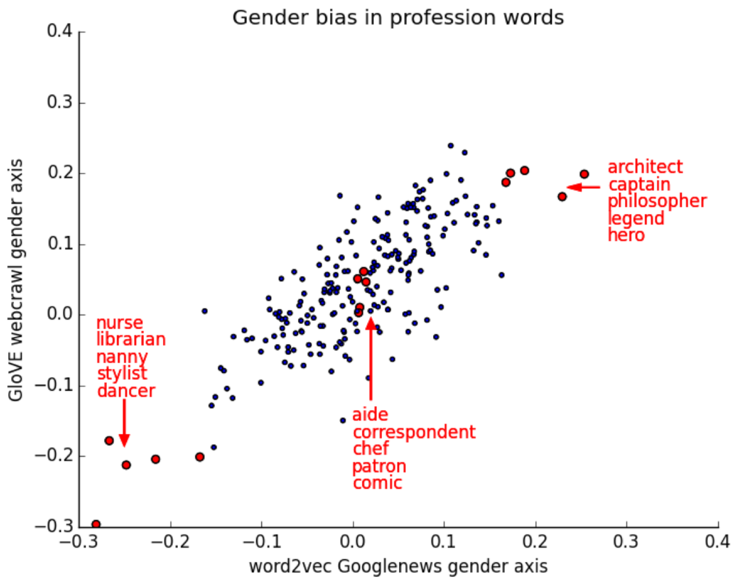
- *Steering* minorities into higher rates (advertising)

- *Redlining:* deny service, change rates based on area



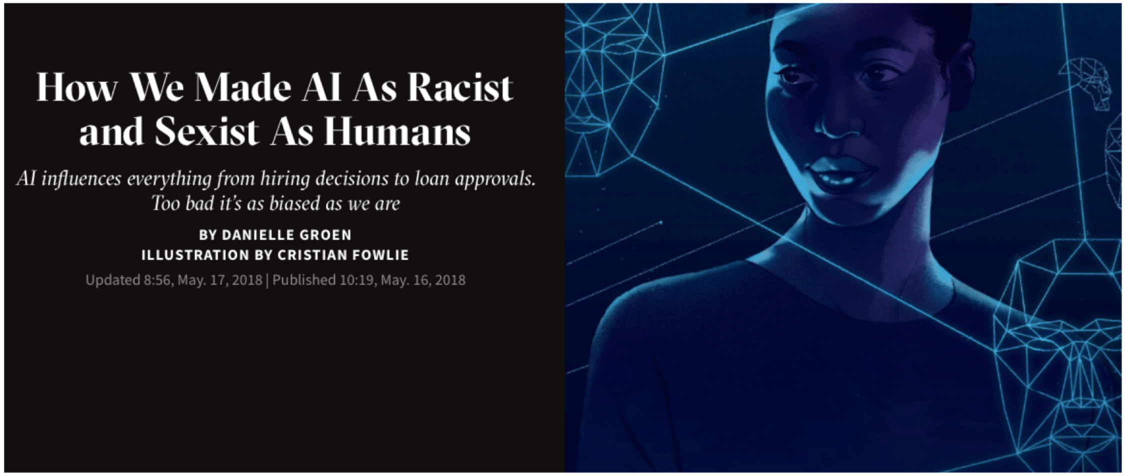
- *Self-fulfilling prophecy:* select less qualified to "justify" future discrimination

Unfairness in Machine Learning?



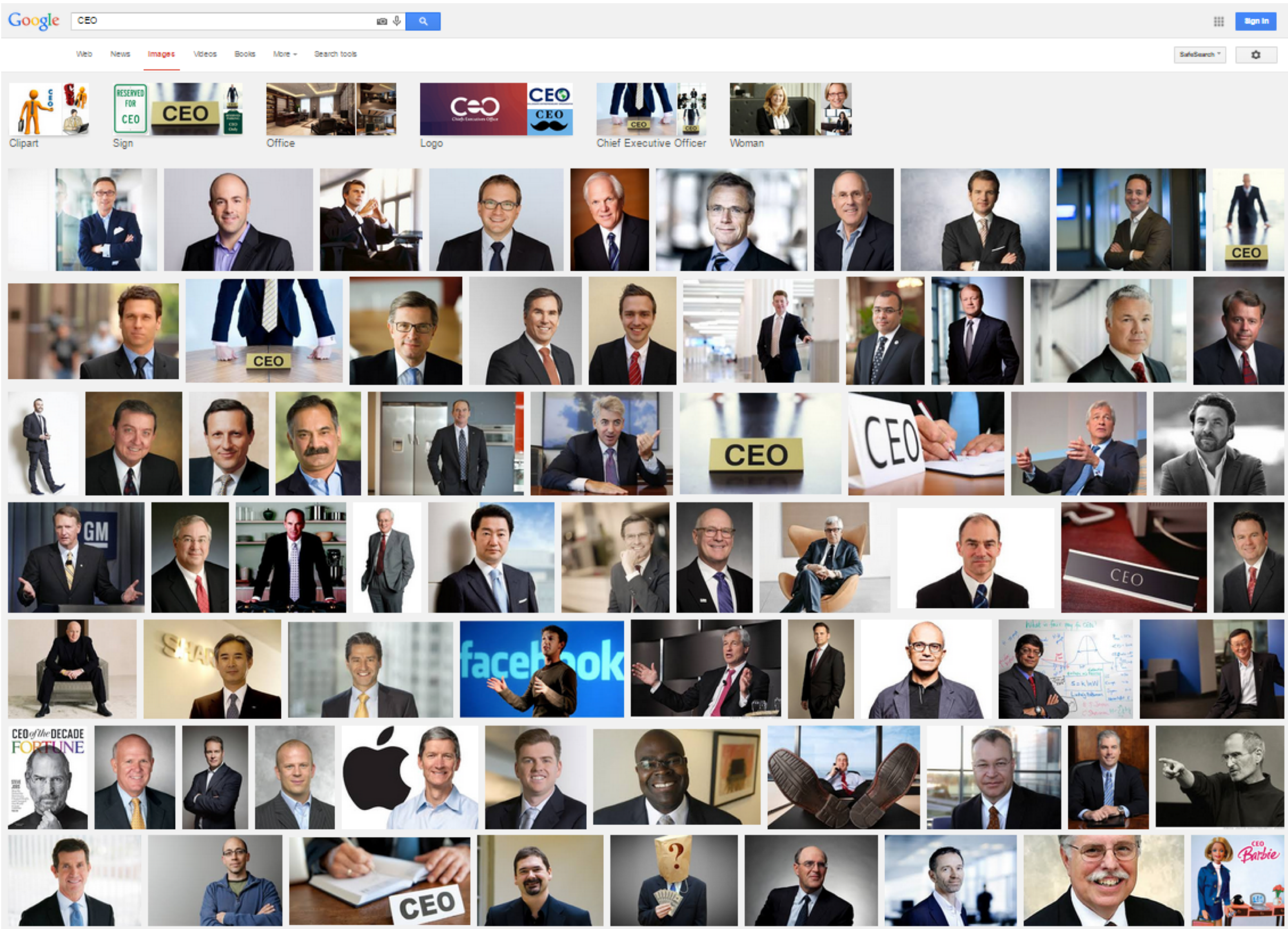
Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Joy Buolawmini



The Walrus, 2018

SUBTLER BIAS

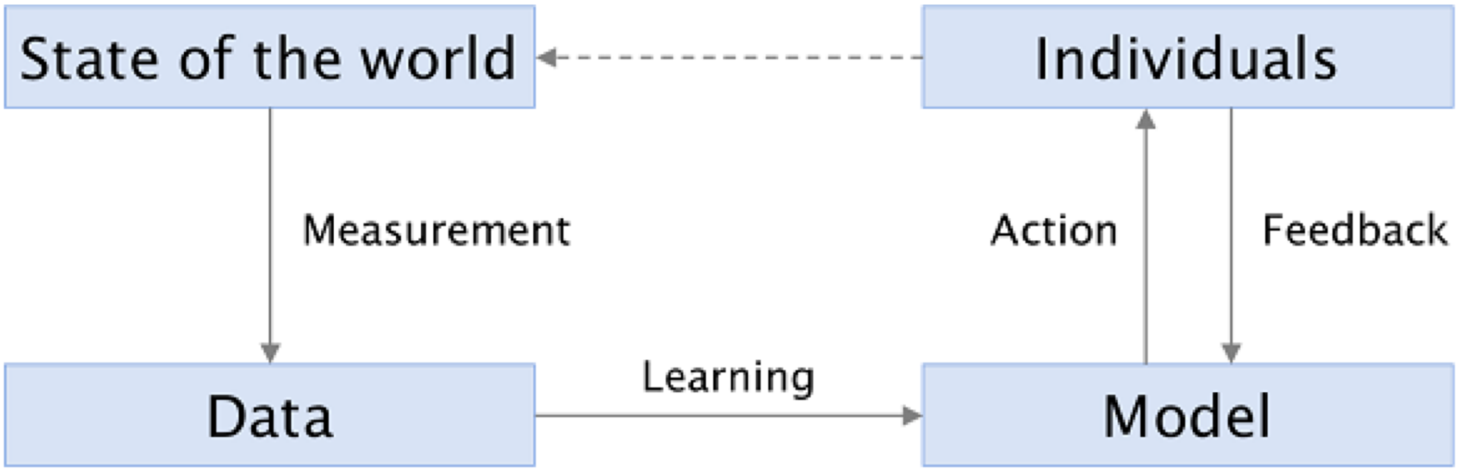


SUBTLER BIAS

Fairness in ML: Goals

Identify and mitigate bias in ML-based decision-making, in all aspects of data pipeline

STAGES OF ML SYSTEM



- Measurement: process by which the state of the world reduced to a set of rows, columns, and values in dataset.
- Learning: turns dataset into model
- Action: based on model's prediction (classification, regression, info retrieval), corresponding action
- Feedback: user responses can update model (e.g., clicks)

Barocas, Hardt, Narayanan, *Fairness in Machine Learning*

DEMOGRAPHIC DISPARITIES

Most ethical issues arise when data concerns people

Training data tends to encode demographic disparities in our society -- can perpetuate stereotypes

Some occupations have stark gender imbalance -- why?

But not all applications involve people. Or do they?
examples: StreetBump; Automated Essay Scoring; Zillow

DATA ISSUES

Basic data issues: imbalanced, impoverished; noisy

Measurement involves subjective choices, and technical difficulties

Example: “Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago.” NYT, 2017

-- %age change 1980-2015 in black, Hispanic, Asian, white, multiracial students

Target variable / labels:

-- what is “creditworthiness”; “good employee”; “attractive”

-- objective measures may be biased too

-- classification schemes may rely on historical taxonomies

Even images not unbiased

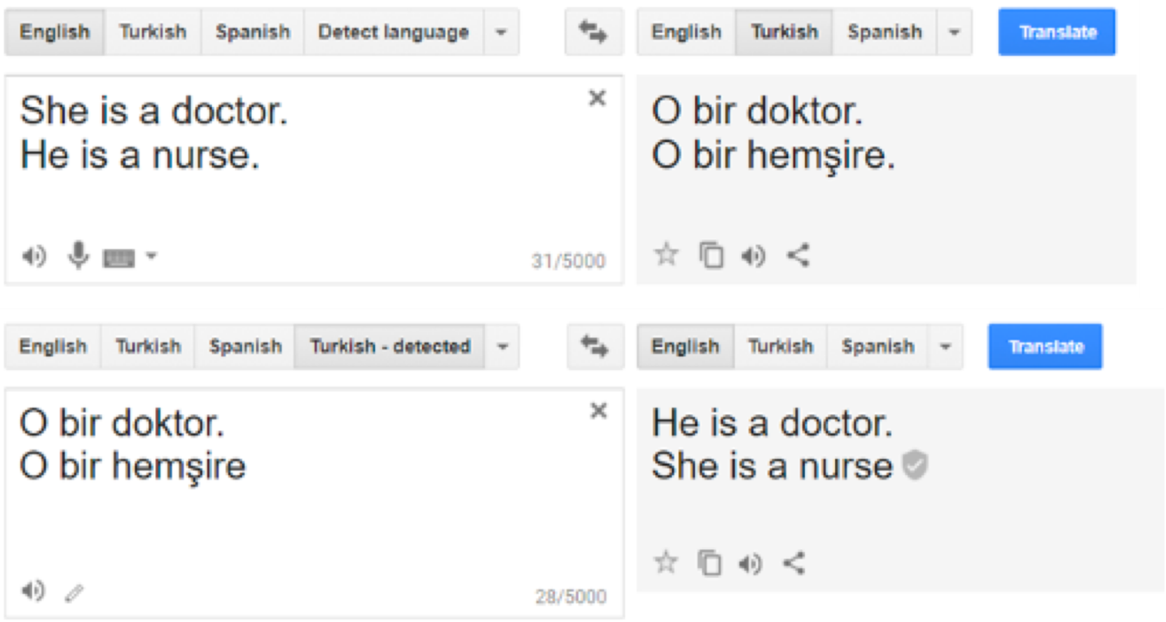
-- default color balance, dynamic range settings

-- distribution of subjects may not match in training/testing

MODEL ISSUES

Models can faithfully reflect disparities in data, often including stereotypes – why?

Some patterns we think are good features for classification, others are not: how to tell them apart?



Can also introduce disparities when none exist – not enough data

Need to train based on something other than just overall accuracy

FEEDBACK LOOPS

Patients with asthma had lower risks of developing pneumonia (Caruana et al, 2015) – prediction affects the outcome

Decisions affect downstream outcomes:

- search result ordering determines clicks

- searches for black-sounding names more likely to lead to ads for arrests (Latanya Sweeney) – due to users clicking more on ads conforming to stereotypes

- decision whether to detain a defendant affects probability of pleading of guilty

- predictive policing sends more police to high-crime areas

FAIR CLASSIFICATION

Explosion of fairness research over last five years

Fair classification is the most common setup, involving:

- X, some data
- Y, a label to predict
- $\hat{Y}$, the model prediction
- A, a sensitive attribute (race, gender, age, socio-economic status)

We want to learn a classifier that is:

- accurate
- fair with respect to A

FAIRNESS VIA S-BLINDNESS?

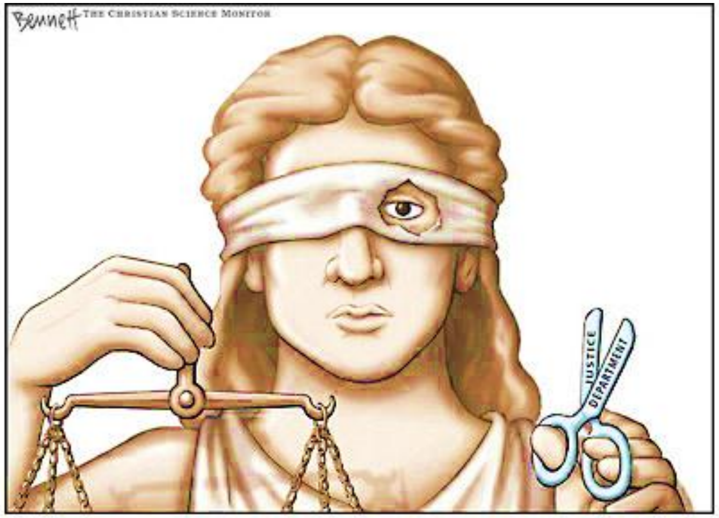
Remove or ignore the "membership in A" bit

- Fails: Membership in A may be encoded in other attributes

FAIRNESS THROUGH AWARENESS

Dwork, Hardt, Pitassi, Reingold, Zemel, 2012

Goal: Assign each individual *a* representation *by being aware of membership in group A*



(1). **Individual Fairness**: Treat similar individuals similarly

(2). **Group Fairness:** equalize two groups (A=1 = minority; A=0 is majority) at the level of outcomes (statistical parity)

FAIR CLASSIFICATION: DEFINITIONS

Definitions based on predicted outcomes:

- Demographic / statistical parity
- Conditional statistical parity (loan conditioned on credit history, amount, employment)

Definitions based on predicted and actual outcomes:

- Balanced PPV (FDR) – predictive equality
- Balanced FNR (TPR) – equal opportunity
- Balanced FNR and FPR – equalized odds

	Actual – Positive	Actual – Negative
Predicted – Positive	**True Positive (TP)** $PPV = \frac{TP}{TP+FP}$ $TPR = \frac{TP}{TP+FN}$	**False Positive (FP)** $FDR = \frac{FP}{TP+FP}$ $FPR = \frac{FP}{FP+TN}$
Predicted – Negative	**False Negative (FN)** $FOR = \frac{FN}{TN+FN}$ $FNR = \frac{FN}{TP+FN}$	**True Negative (TN)** $NPV = \frac{TN}{TN+FN}$ $TNR = \frac{TN}{TN+FP}$

FAIR CLASSIFICATION: DEFINITIONS

Most common way to define fair classification is to require some invariance with respect to the sensitive attribute

- Demographic parity: $\hat{Y} \perp A$
- Equalized Odds: $\hat{Y} \perp A | Y$
- Equal Opportunity: $\hat{Y} \perp A | Y = y$, for some y
- Equal (Weak) Calibration: $Y \perp A | \hat{Y}$
- Equal (Strong) Calibration: $Y \perp A | \hat{Y}$ and $\hat{Y} = P(Y = 1)$
- Fair Subgroup Accuracy: $\mathbb{1}[Y = \hat{Y}] \perp A$

Note: Many of these definitions are incompatible!