Ethics in Natural Language Processing – SS 2019

Lecture 4
Bias II

Dr. Thomas Arnold
Aniket Pramanik

Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt

Slides and material from Yulia Tsvetkov

Carnegie Mellon University
Language Technologies Institute

Syllabus (tentative)

Nr.	Lecture
01	Introduction, Foundations I
02	Foundations II
03	Bias I
04	Bias II
05	Incivility and Hate Speech I
06	NO LECTURE – Christi Himmelfahrt
07	Incivility and Hate Speech II
08	Low-Resource NLP, NLP for Social Good
09	NO LECTURE - Fronleichnam
10	Privacy and Security I
11	Privacy and Security II
12	Language of Manipulation I
13	Language of Manipulation II

Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Bias in ML predictions

Bias Amplification

Debiasing Techniques

How Do We Make Decisions

System 1

automatic

fast

parallel

automatic

effortless

associative

slow-learning

System 2

effortful

slow

serial

controlled

effort-filled

rule-governed

flexible

Kahneman & Tversky 1973, 1974, 2002

Psychological Perspective on Implicit Bias

Biases inevitably form because of our innate tendency to:

- Categorize the world to simplify processing
- Store learned information in mental representations (called schemas)
- Automatically and unconsciously activate stored information whenever one encounters a category member

Cognitive bias is a systematic pattern of deviation from rationality in judgement

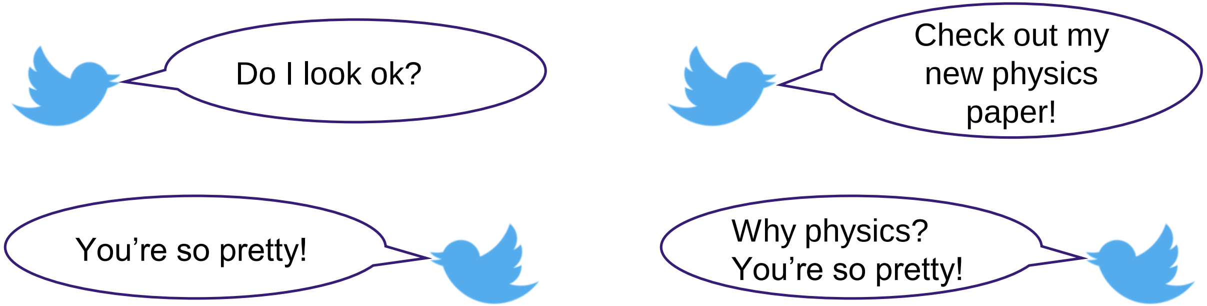
Implicit Association Test - Greenwald et al. 1998

Category	Items
Good	Spectacular, Appealing, Love, Triumph, Joyous, Fabulous, Excitement, Excellent
Bad	Angry, Disgust, Rotten, Selfish, Abuse, Dirty, Hatred, Ugly
African Americans	
European Americans	

Online data is riddled with **SOCIAL STEREOTYPES**

Positive or negative?

Positive or negative?



Positive or negative?

Do I look ok?

You're so pretty!

Check out my new physics paper!

Why physics? You're so pretty!

Do I look ok?

You're so pretty for your age!

You're so pretty for a black girl!

You're too pretty to be gay!

Outline

Recap

Bias in ML predictions

Bias Amplification

Debiasing Techniques

Learning Goals - Example Questions

A dataset of various online activities of a group of people should be anonymized. Further, the dataset should be debiased in a way that information about gender is not visible anymore.

Discuss potential problems in obfuscating gender by deleting the respective feature.

Learning Goals - Example Questions

A machine learning algorithm has equal True Positive Rates for both female and male subgroups of a dataset. Does this ensure that the algorithm is unbiased regarding these two groups? Explain your answer.

What is Bias Amplification? Give an example.

Give two example of bias in word embeddings.

Algorithmic Biases and Fairness in ML

- "Fairness through Awareness": FAT ML proceedings: https://www.fatml.org/resources/relevant-scholarship

- https://arxiv.org/abs/1104.3913

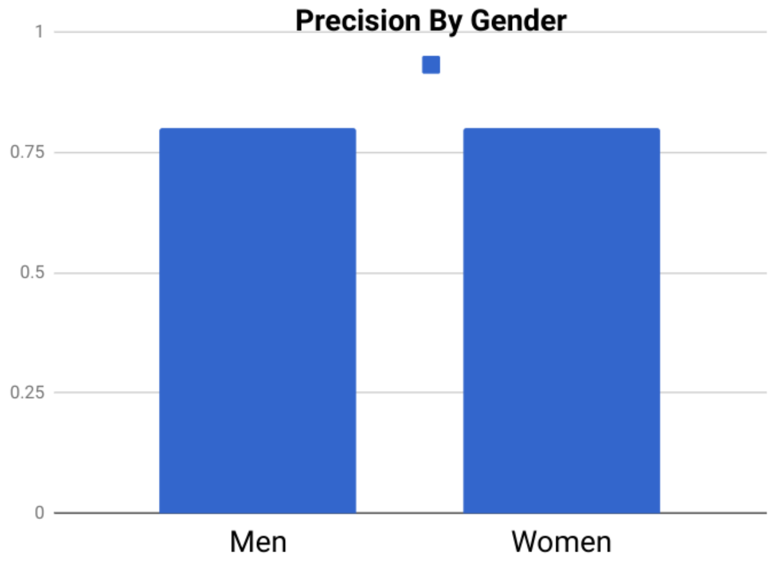
Example: Targeted Advertising

- Task: build a classifier to detect Software Engineers (SWE)
  - Inputs: user data
    - Browsing history, location, language, interests…
  - Outputs: predict whether the user is SWE or non-SWE

Results of the Trained Classifier

$$\Pr(SWE = 1 \mid Classifier = 1)$$

80% of the SWE predictions were correct, same for men and women

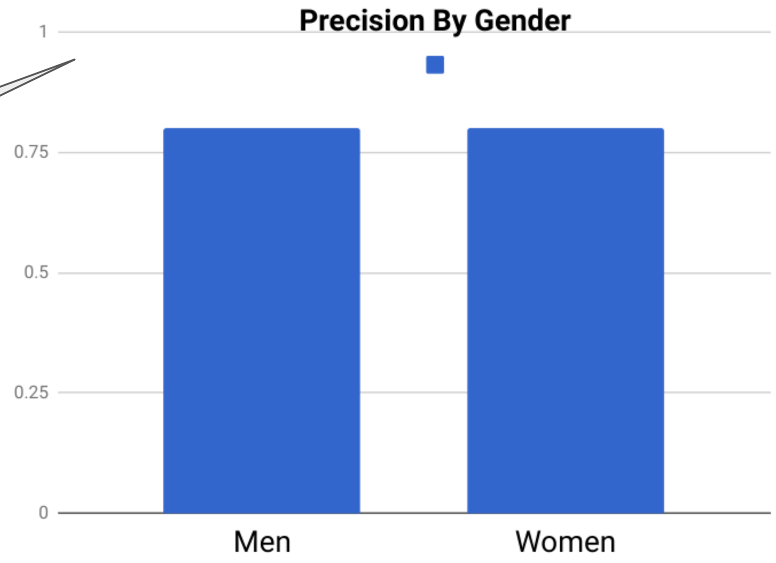


Results of the Trained Classifier

$Pr(SWE = 1 \mid Classifier = 1)$

80% of the SWE predictions were correct, same for men and women

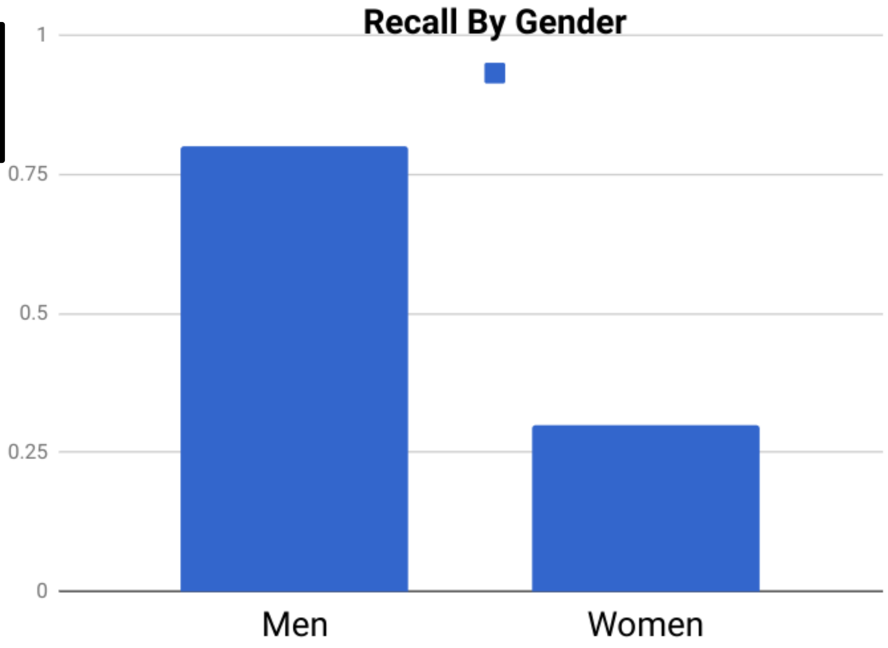
Is this an unbiased classifier?



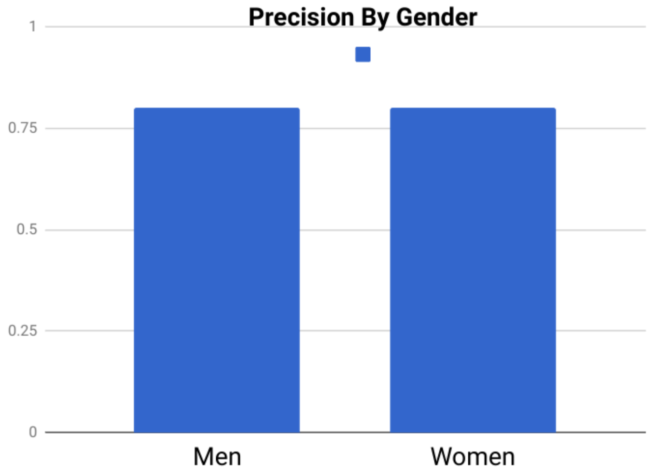
Let’s Slice The Data Differently

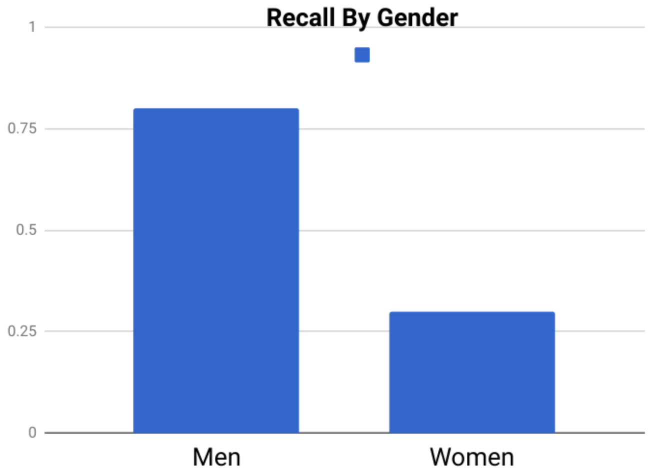
$$Pr(Classifier = 1 | SWE = 1)$$

80% of male SWE were classified, but only 30% female SWE



Let’s Slice The Data Differently





Precision

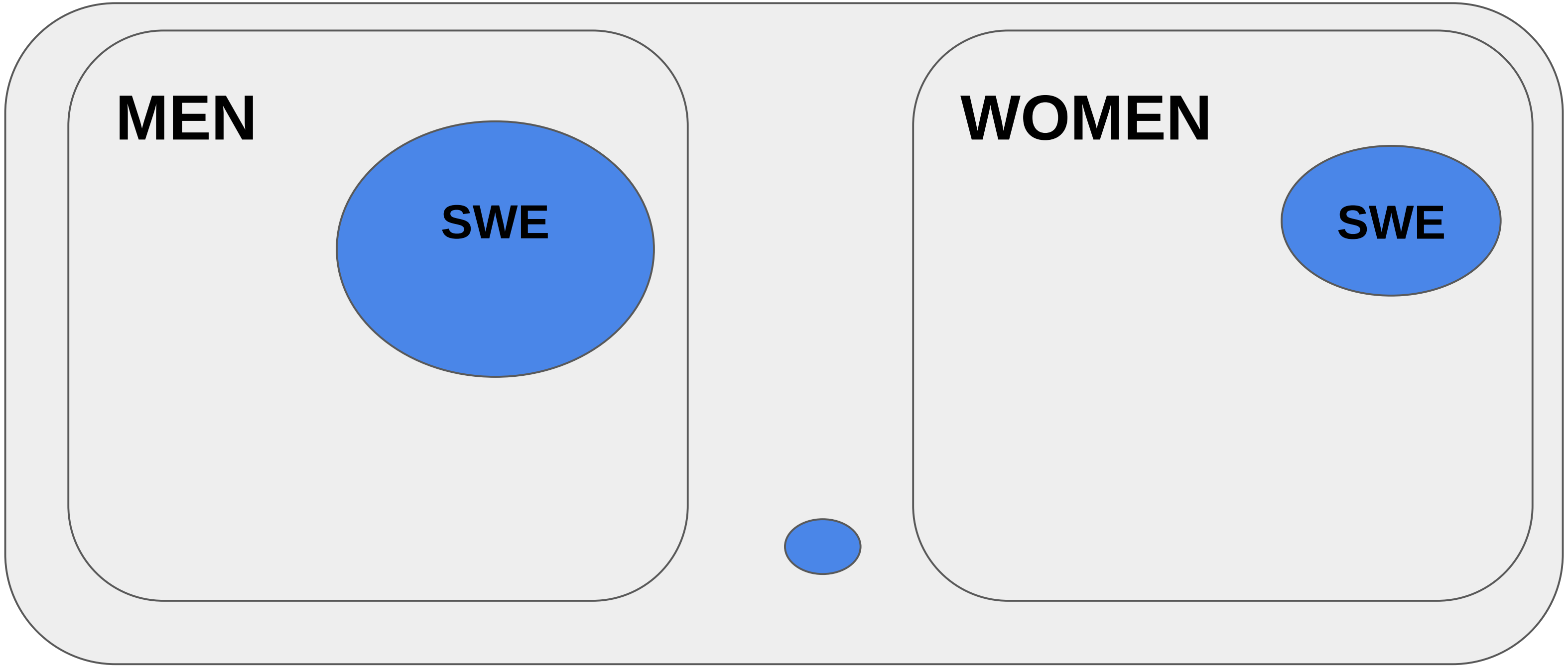
Recall

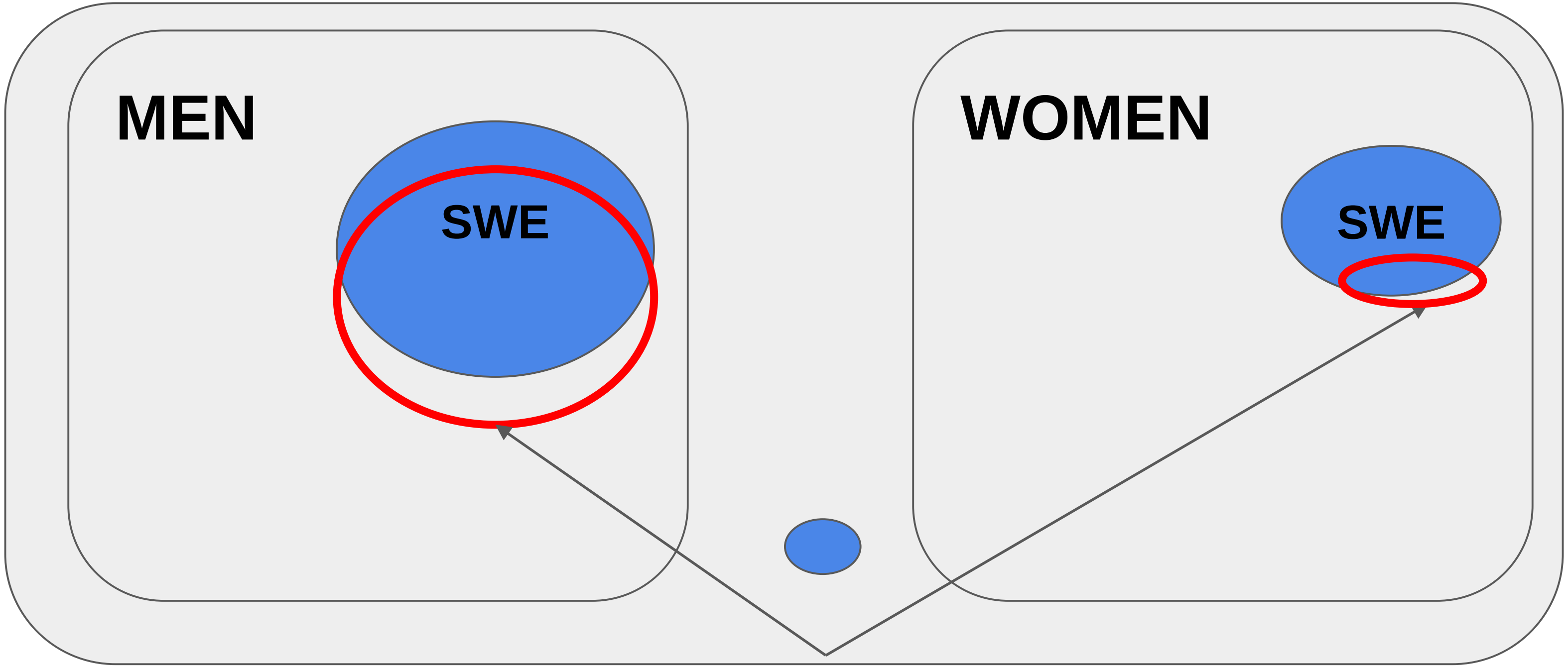
$Pr(SWE = 1 \mid Classifier = 1)$

$Pr(Classifier = 1 \mid SWE = 1)$

MEN

WOMEN


MEN
SWE
WOMEN
SWE



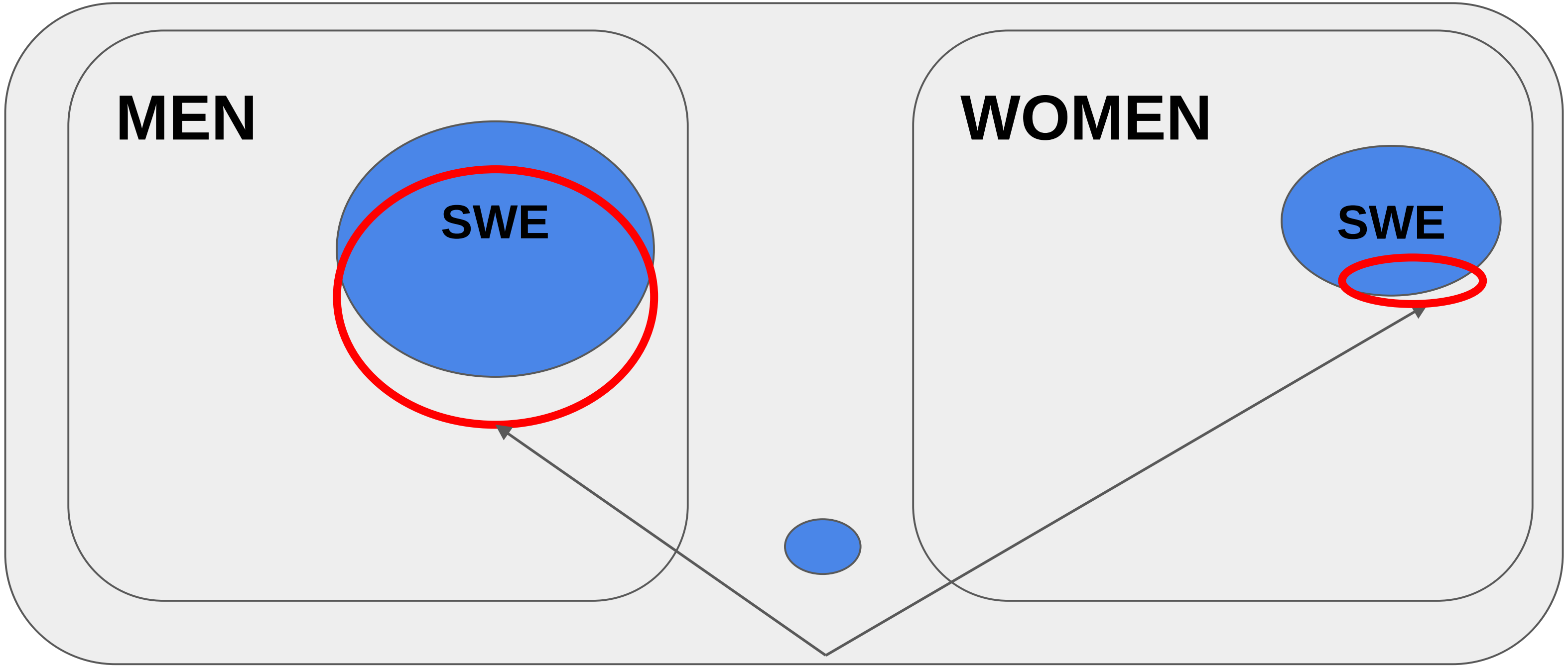
Classifier predictions

$\Pr(SWE{=}1 \mid Classifier{=}1) = 0.8$

$\Pr(Classifier{=}1 \mid SWE{=}1) = 0.8$

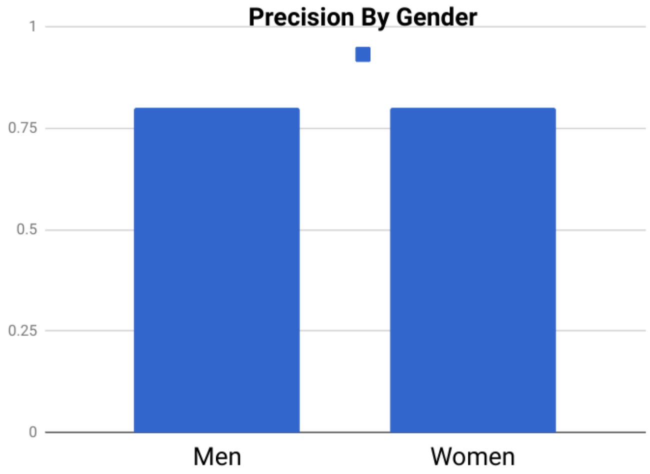
$\Pr(SWE{=}1 \mid Classifier) = 0.8$

$\Pr(Classifier{=}1 \mid SWE{=}1) = 0.3$

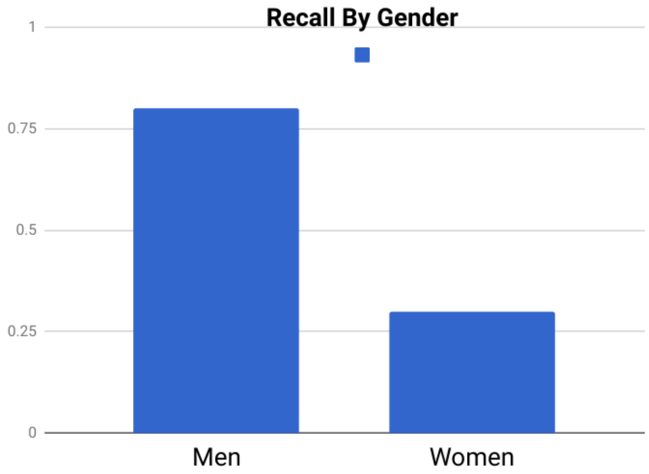


Classifier predictions

Precision–Recall Trade-off



Click Through Rate (Precision)



True Positive Rate (Recall)

Can we make the classifier more inclusive?

suggest a solution

Attempt 1: Fairness Through Unawareness

- Protect individuals’ privacy
  by excluding sensitive attributes like gender and race

Attempt 1: Fairness Through **Un**awareness

- ~~Protect individuals' privacy~~
  ~~by excluding sensitive attributes like gender and race~~
- Fairness is not guaranteed if sensitive attributes are removed or ignored

- Sensitive attributes are correlated with other variables:
  - Gender and browsing history
  - Zip code and race

Dwork & Mulligan **It's Not Privacy, and It's Not Fair** *Stanford Law Review,* 2013

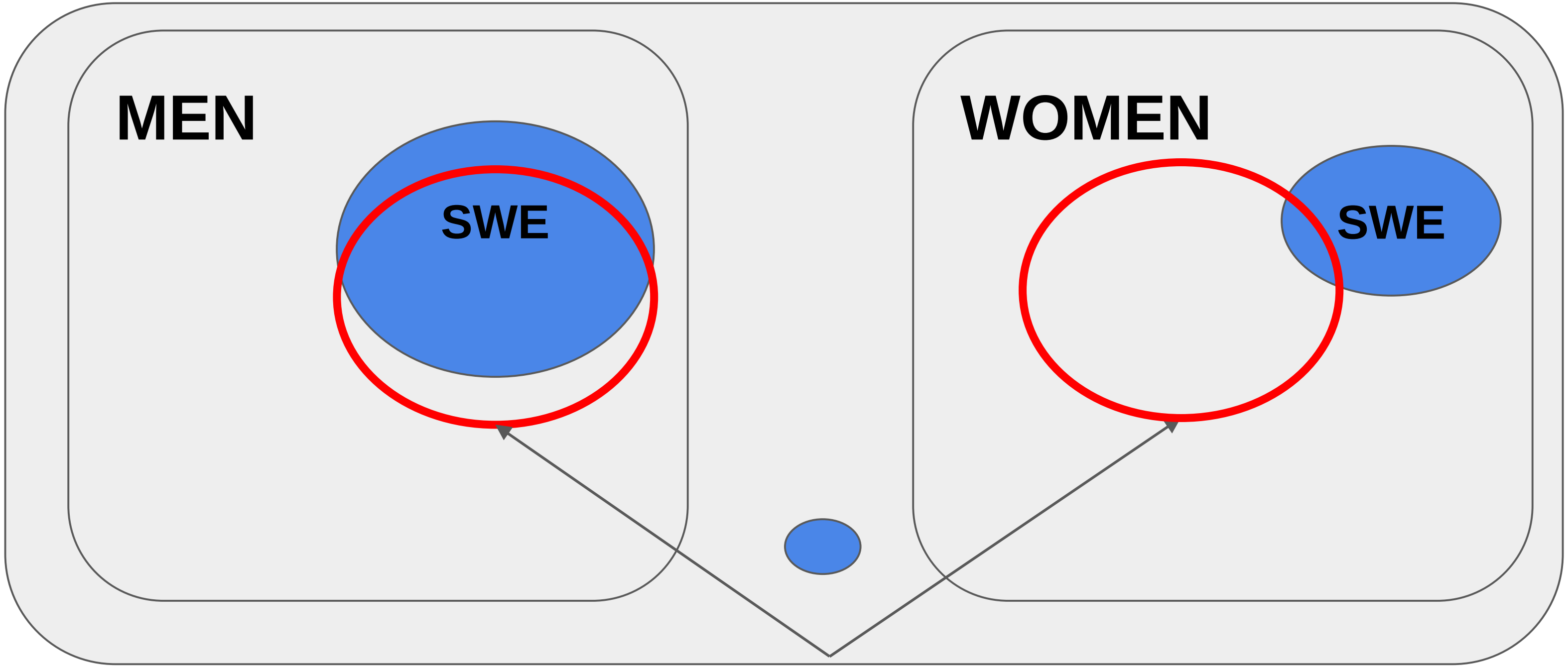
Attempt 2: Group Fairness

$$Pr(Classifier = 1 | Gender = m\} = Pr(Classifier = 1 | Gender = f)$$

Fraction of people labeled as SWE among males is the same as females

Attempt 2: Group Fairness

Pr(*Classifier*=1|*Gender*=m} = Pr(*Classifier*=1|*Gender*=f)



Classifier predictions

Attempt 2: Group Fairness

$$Pr(Classifier = 1 | Gender = m\} = Pr(Classifier = 1 | Gender = f)$$

Fraction of people labeled as SWE among males is same as females

- Not a fair solution: observed distributions are not equal
- Can be misused or abused in many ways
  - “Reduced utility”
  - “Self-fulfilling prophecy”
  - “Subset targeting”

Attempt 3: Equality of Odds
Treating Similar Individuals Similarly

Recall is equal for “sensitive” parameters:

$$\Pr(Classifier = 1 \mid SWE = 1, Gender = m) =$$
$$= \Pr(Classifier = 1 \mid SWE = 1, Gender = f)$$

Equality of Odds

Recall is equal for “sensitive” parameters

- Ensure equality of odds for either positive or negative outcomes:
  - **True Positive Rate**: something that is good for the individual
    - SWE classifier, likely to graduate, likely to get a loan
  - **False Positive Rate**: something that is harmful for the individual
    - Likely to recommit a crime, likely to go bankrupt, screening for terrorists

Algorithmic Biases and Fairness in ML

TECHNISCHE UNIVERSITÄT DARMSTADT

Want to learn more?

- "Fairness through Awareness": FAT ML proceedings: https://www.fatml.org/resources/relevant-scholarship
- https://arxiv.org/abs/1104.3913

Northpointe vs ProPublica

Goal

“what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?”

Training Data and Features

"what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?"

- COMPAS system claims
  - balanced training data about people of all races
  - race was *not* one of the input features

Algorithm is not Oblivious to Race

TECHNISCHE UNIVERSITÄT DARMSTADT

“what is the probability that this person will commit a serious crime in the future, as a function of the sentence you give them now?”

- COMPAS system claims
  - balanced training data about people of all races
  - race was *not* one of th

- Sensitive attributes are correlated with other variables:
  - Gender and browsing history
  - Zip code and race

Dwork & Mulligan **It’s Not Privacy, and It’s Not Fair** *Stanford Law Review,* 2013

Goal

“what is the probability that this person will **commit a serious crime** in the future, as a function of the sentence you give them now?”

How to compute this in the training data?

Optimizing Towards a Biased Objective

“what is the probability that this person will **commit a serious crime** in the future, as a function of the sentence you give them now?”

- Objective function
  - “who is more likely to be **convicted**”

Optimizing Towards a Biased Objective

TECHNISCHE UNIVERSITÄT DARMSTADT

“what is the probability that this person will **commit a serious crime** in the future, as a function of the sentence you give them now?”

- Objective function
  - “who is more likely to be **convicted**”

this is unobtainable data.
“who is more likely to be *convicted*” was used as a proxy

Northpointe vs ProPublica

TECHNISCHE UNIVERSITÄT DARMSTADT

“what is the probability that this person **will commit a serious crime** in the future, as a function of the sentence you give them now?”

- COMPAS system claim
  - balanced training data about people of all races
  - race was *not* one of the input features
- Objective function
  - “who is more likely to be *convicted*”

training data and learning objective are not oblivious to race and biased against non-white people

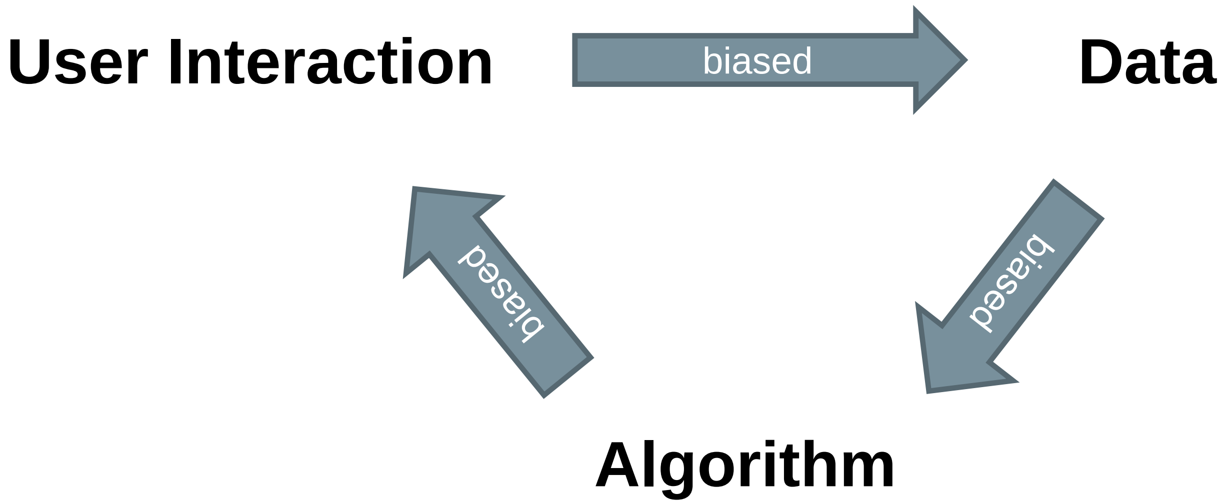
Northpointe vs ProPublica

COMPAS

- Unequal false positive rate

  (the system mistakenly predicts that a person will commit a crime)

Feedback Loop



Outline

Recap

Bias in ML predictions

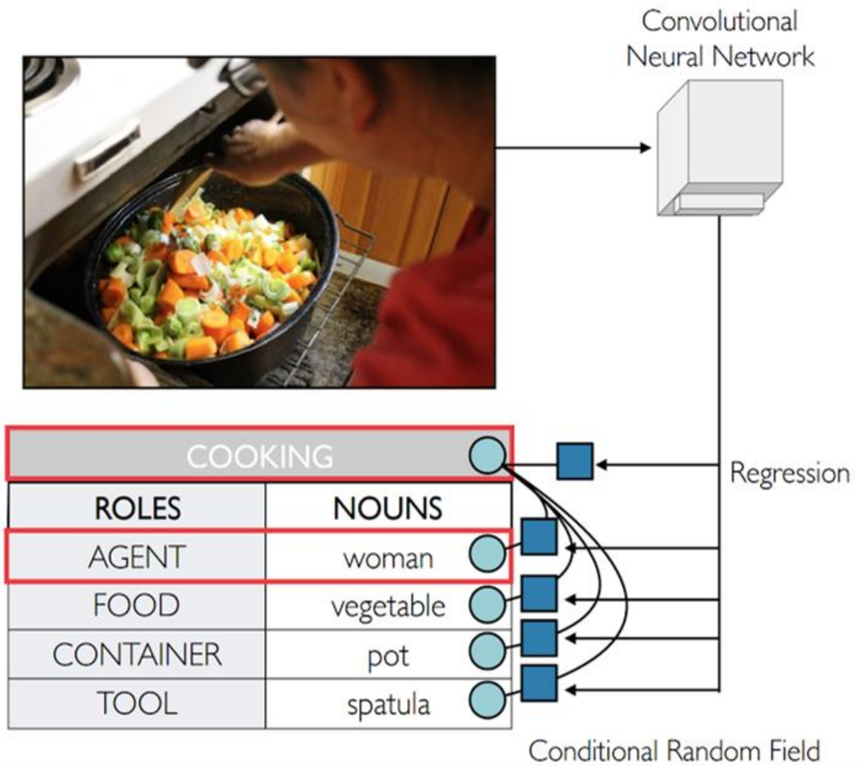
Bias Amplification

Debiasing Techniques

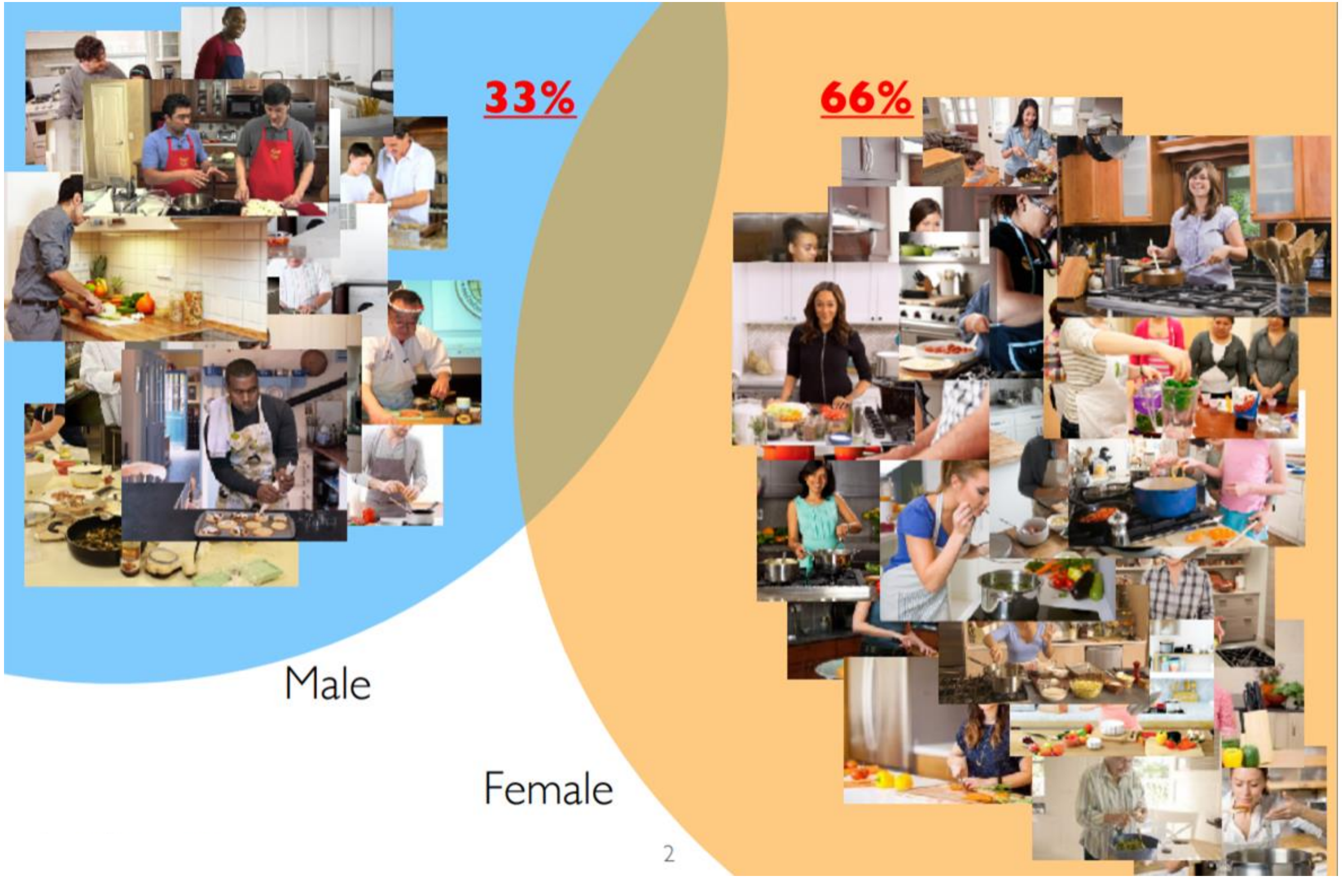
Bias Amplification

Zhao, J., Wang, T., Yatskar, M., Ordonez, V and Chang, M.-W. (2017) **Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint.** *EMNLP*

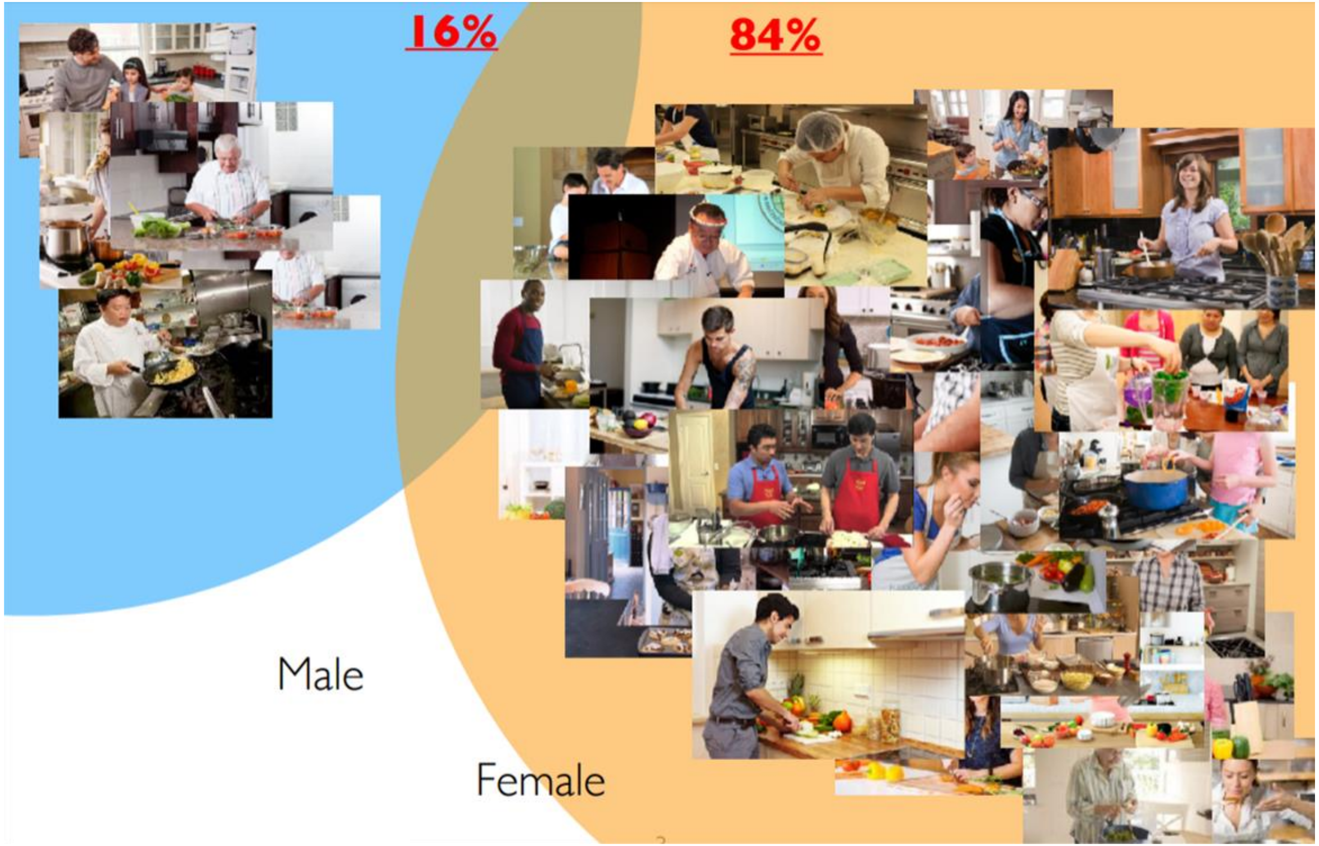
imSitu Visual Semantic Role Labeling (vSRL)



Dataset Gender Bias



Model Bias After Training



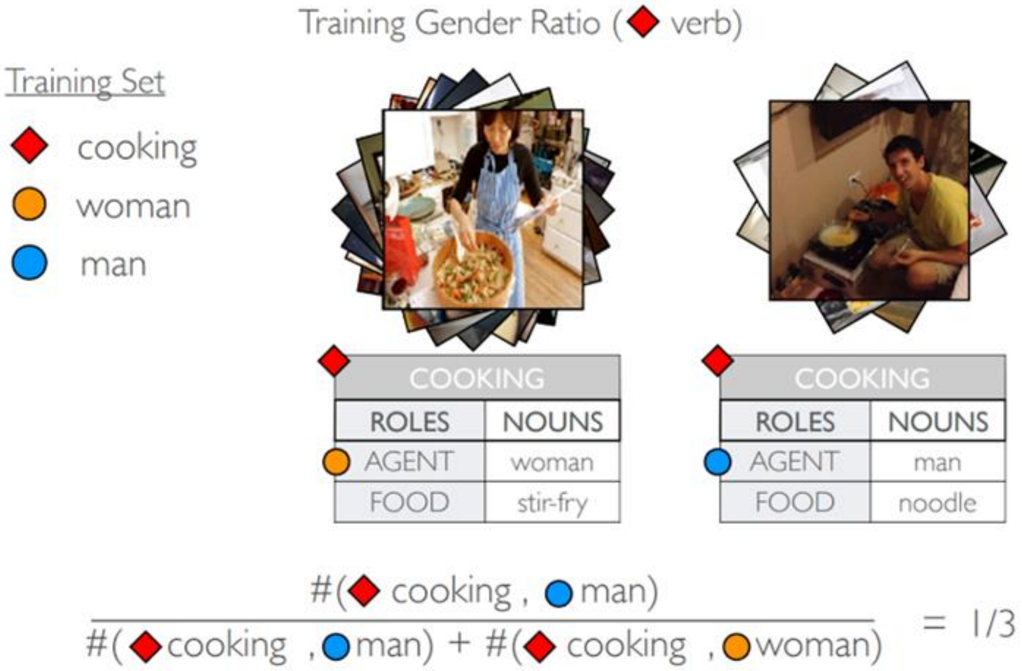
Why does this happen?

Algorithmic Bias

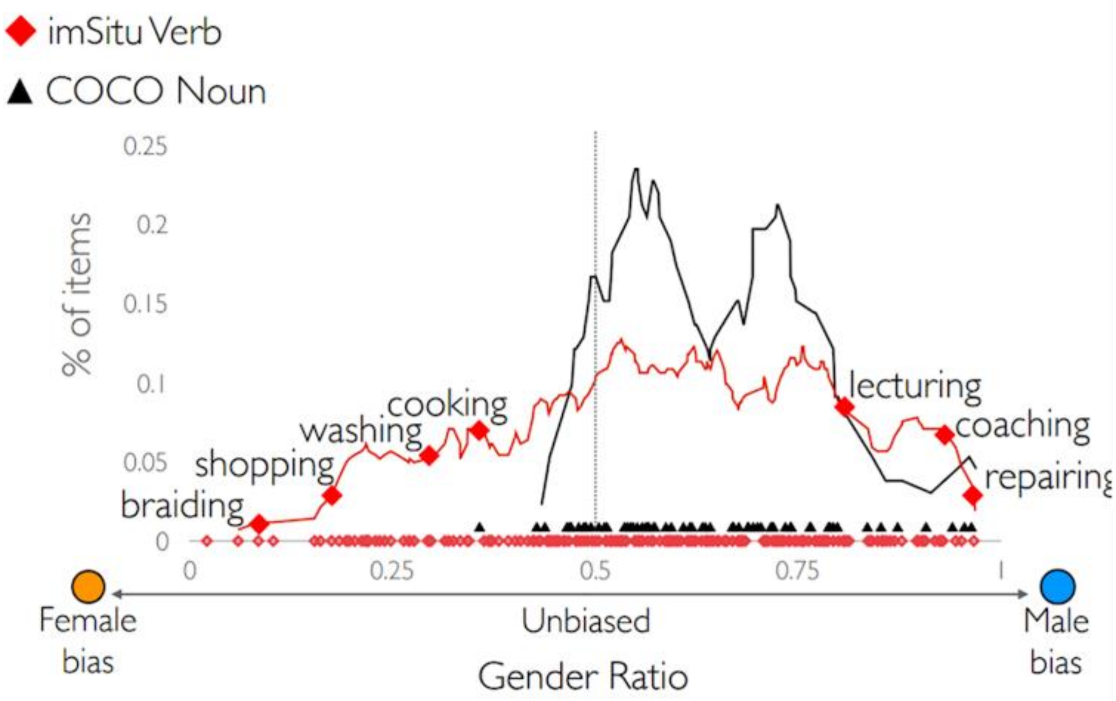
Quantifying Dataset Bias

$$bias(activity, gender) = \frac{cooc(activity, gender)}{\Sigma_{gender' \in G} cooc(activity, gender')}$$

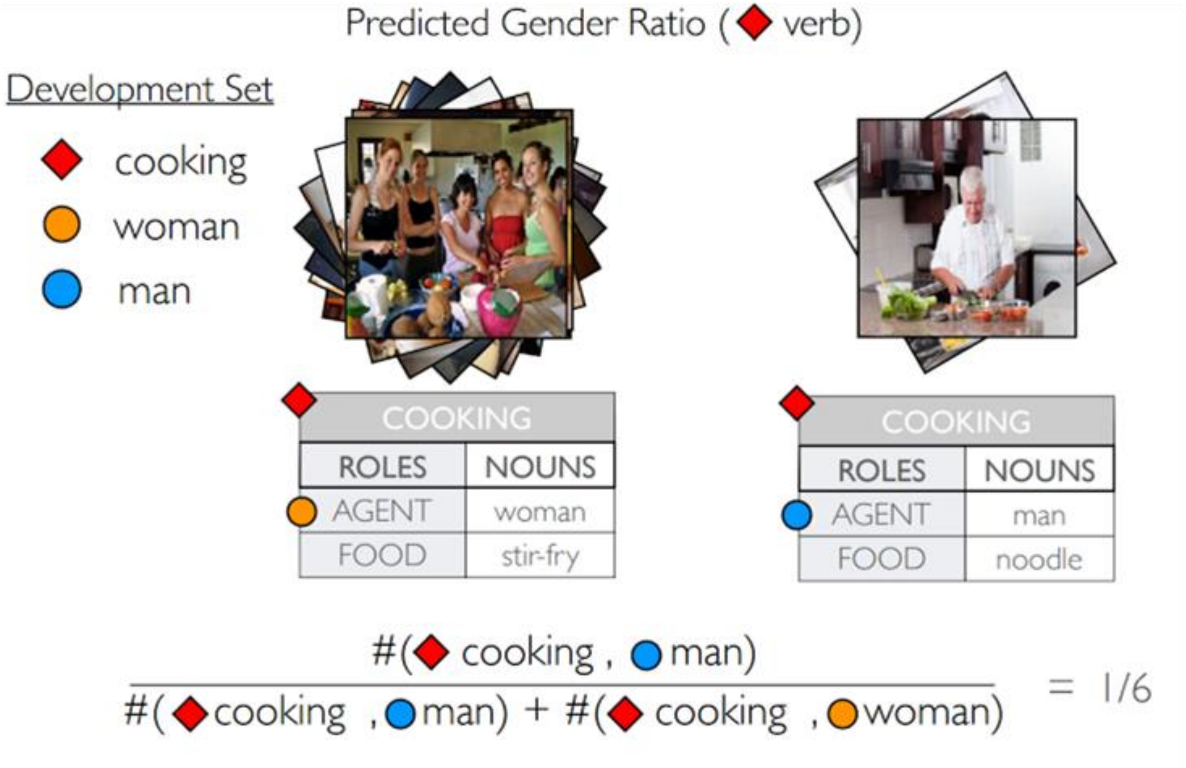
Quantifying Dataset Bias



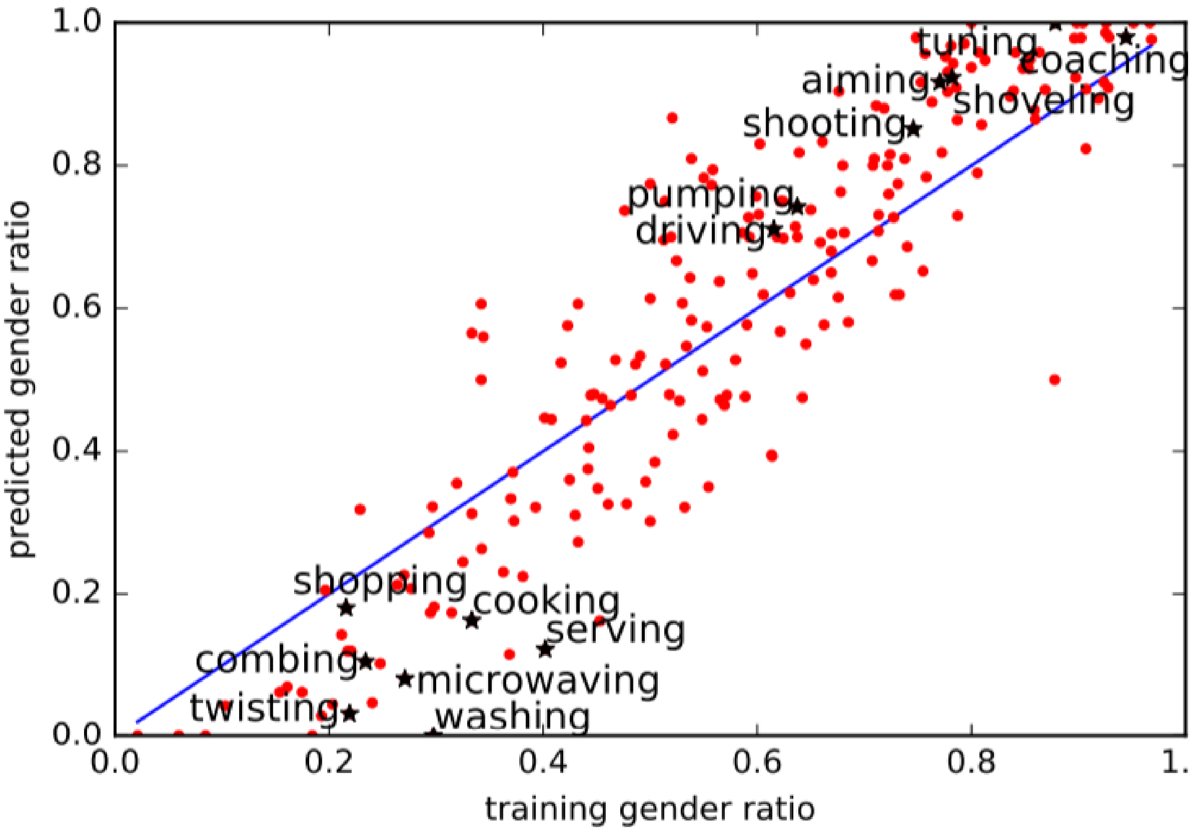
Gender Dataset Bias



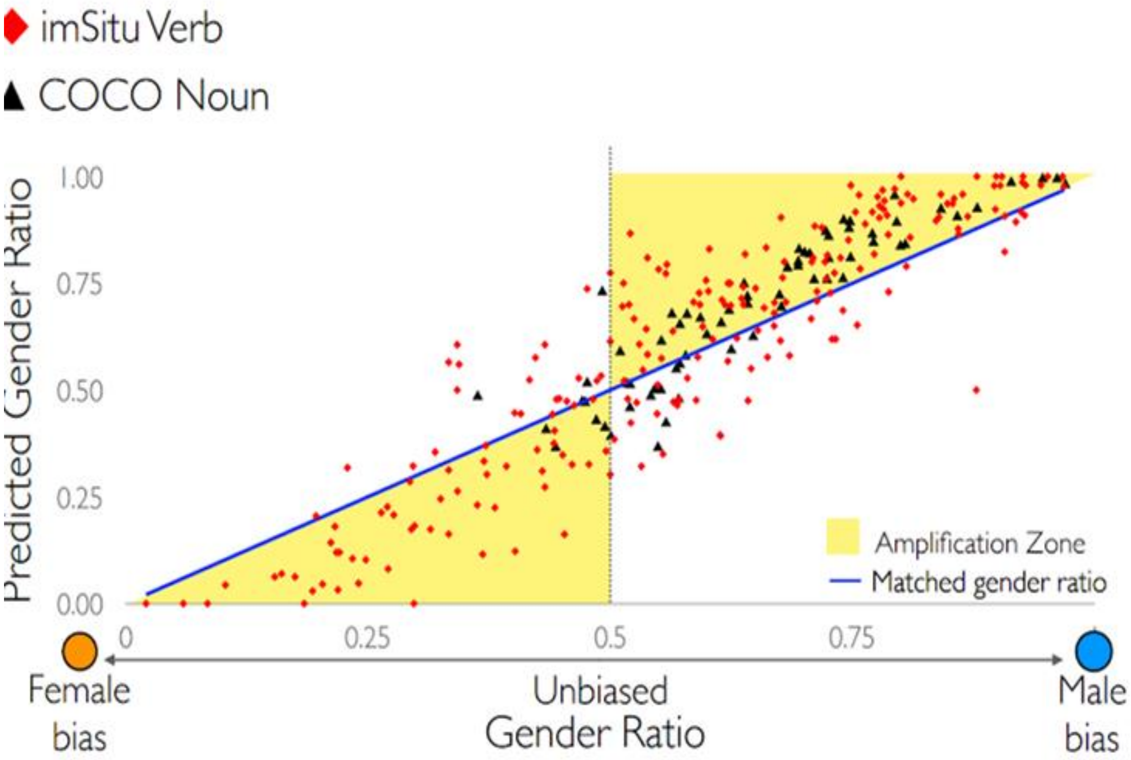
Quantifying Dataset Bias: Dev Set



Model Bias Amplification



Model Bias Amplification



Quantifying Bias Amplification

$$\frac{1}{|O|}\sum_{g}\sum_{o\in\{o\in O|b^*(o,g)>1/||G||\}}\tilde{b}(o,g)-b^*(o,g)$$

O - activity

G - gender

b(o, g)* - training data bias

b~(o, g) - model bias

Reducing Bias Amplification

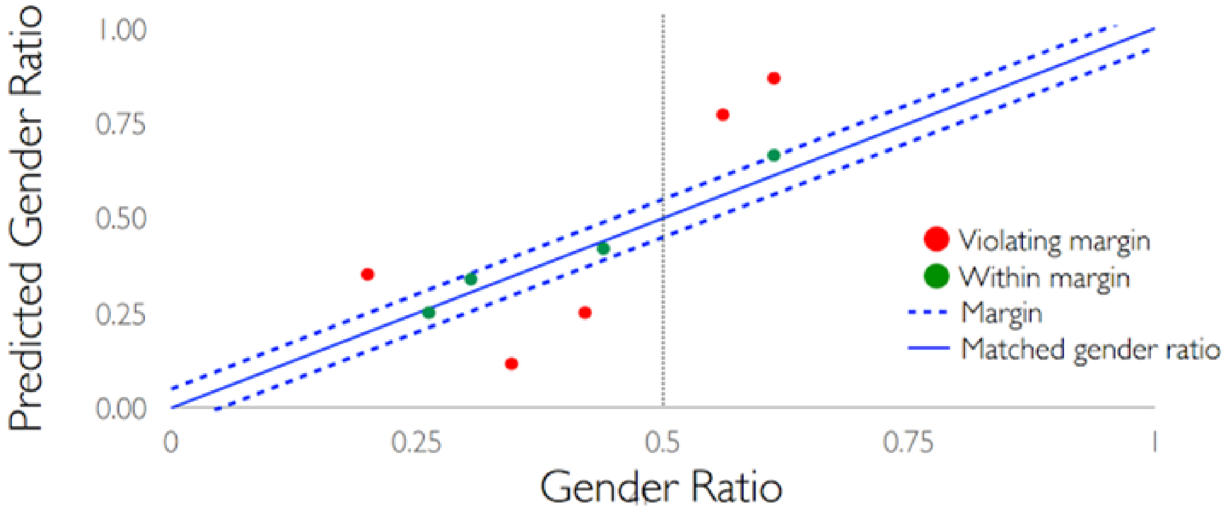
$$\sum_i \max_{y_i} s(y_i, \text{image})$$

$$\forall \text{ points} \quad \left| \text{Training Ratio} - \underset{f(y_1 \ldots y_n)}{\text{Predicted Ratio}} \right| \quad <= \quad \text{margin}$$

New goal for optimization:

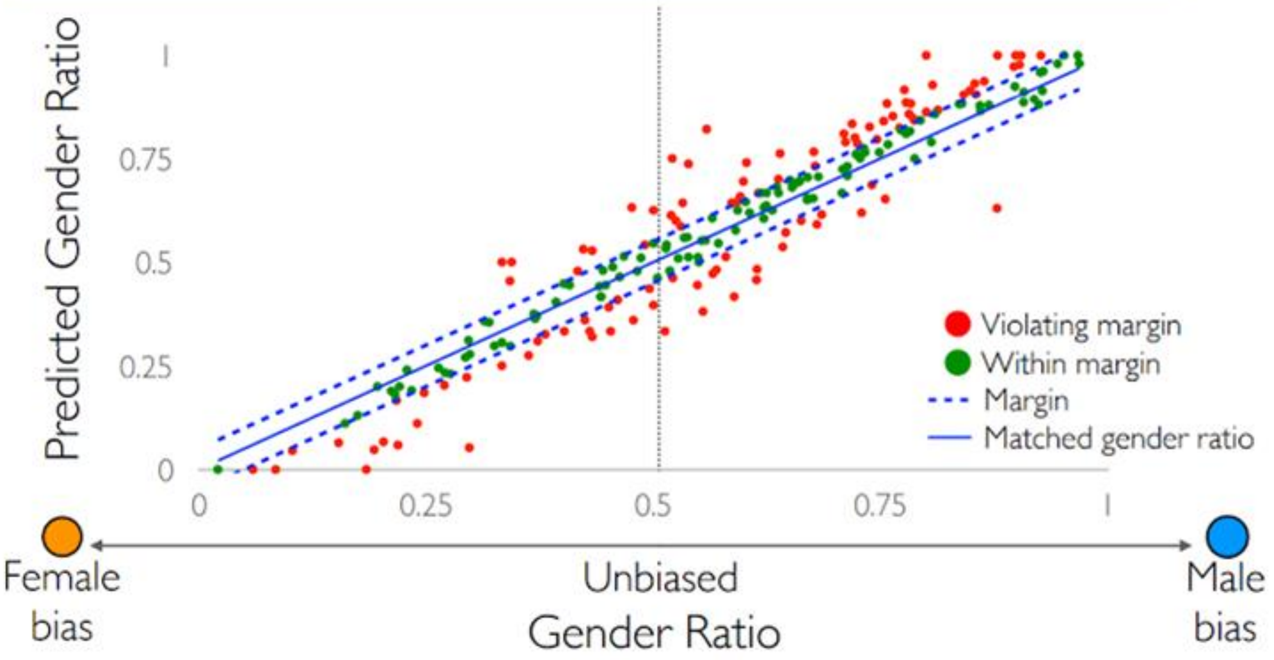
- maximize accuracy (first row)
- while keeping the bias amplification below fixed threshold

Debiasing through Calibration



Results

imSitu Verb	Violation: 72.6%	.050 |bias↑|	24.07 acc.
w/ RBA	Violation: 50.5%	.024 |bias↑|	23.97 acc.



Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

Recap

Bias in ML predictions

Bias Amplification

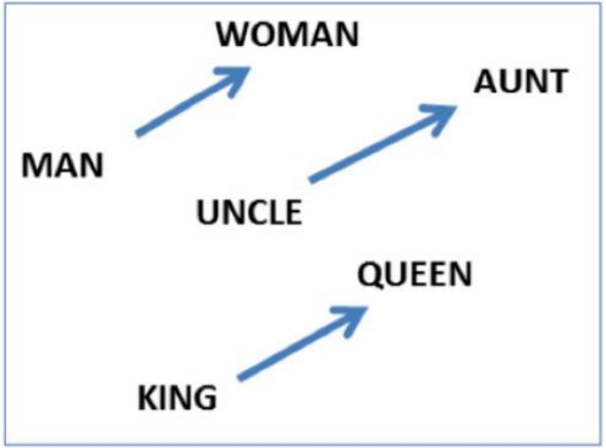
Debiasing Techniques

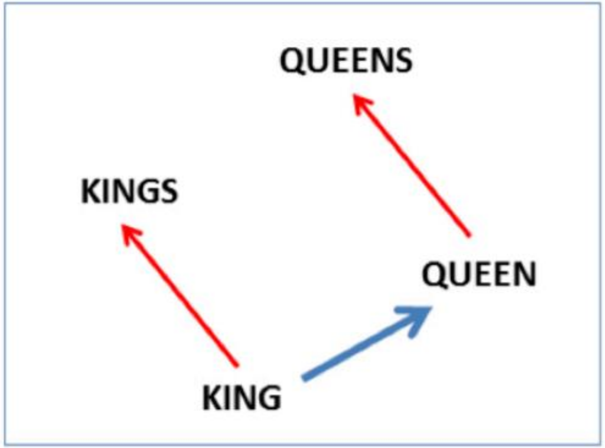
Bolukbasi T., Chang K.-W., Zou J., Saligrama V., Kalai A. (2016) **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS*

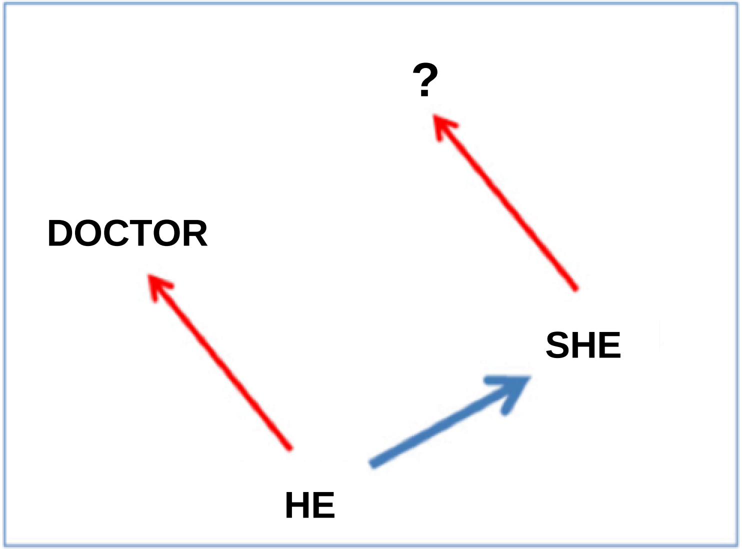
$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}.$$

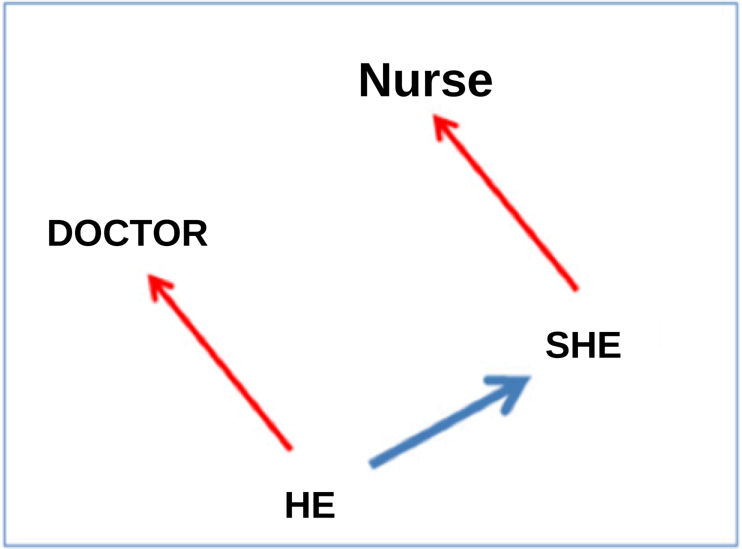
vector(*'king'*) - vector(*'man'*) + vector(*'woman'*) ≈ vector('queen')

vector(*'Paris'*) - vector(*'France'*) + vector(*'Italy'*) ≈ vector('Rome')






?
DOCTOR
SHE
HE


Nurse
DOCTOR
SHE
HE

Crowdsourced Occupational Stereotypes

Extreme *she*	**Extreme *he***
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$

Gender stereotype *she-he* analogies

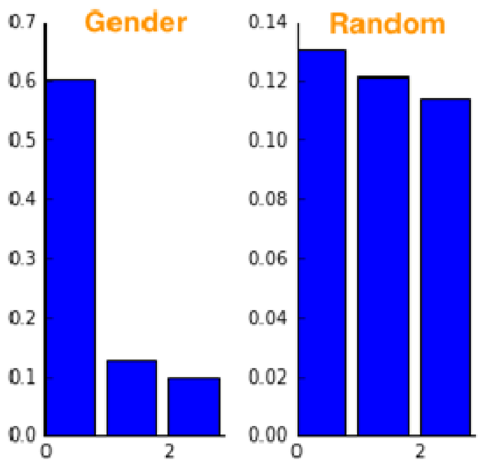
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas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

Gender appropriate *she-he* analogies

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

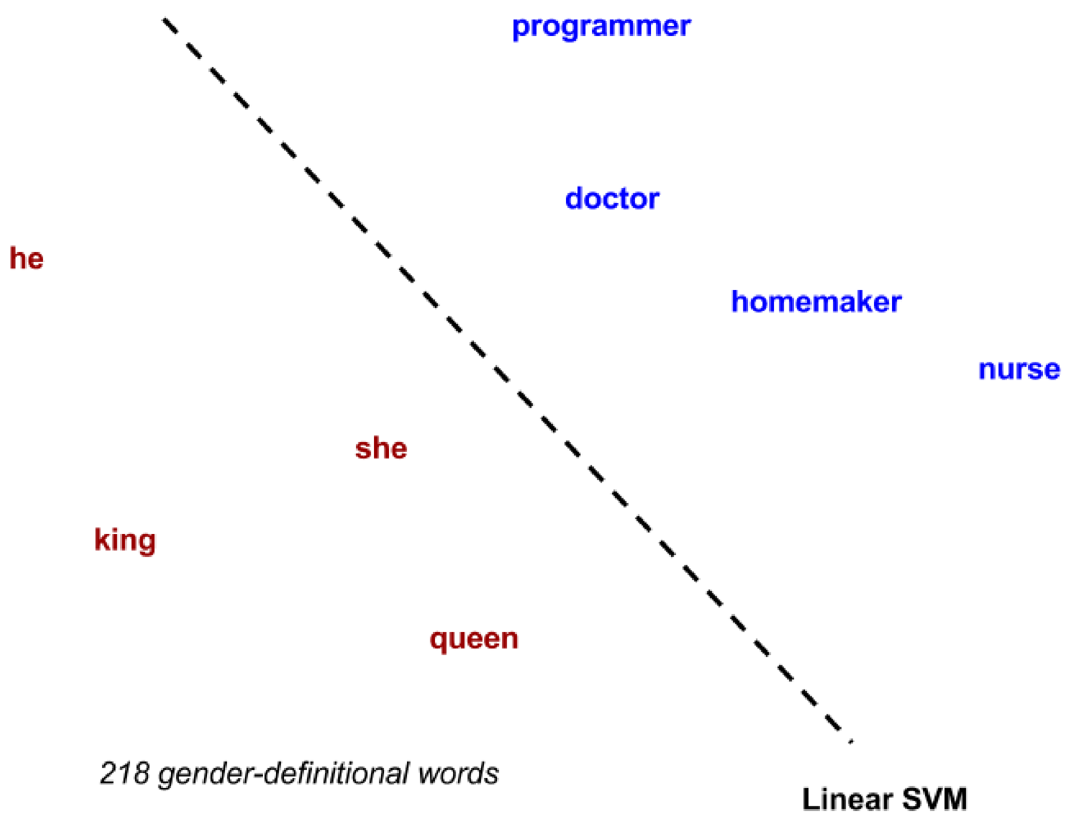
Gender Subspace

	def.	stereo.
$\overrightarrow{\text{she}}-\overrightarrow{\text{he}}$	92%	89%
$\overrightarrow{\text{her}}-\overrightarrow{\text{his}}$	84%	87%
$\overrightarrow{\text{woman}}-\overrightarrow{\text{man}}$	90%	83%
$\overrightarrow{\text{Mary}}-\overrightarrow{\text{John}}$	75%	87%
$\overrightarrow{\text{herself}}-\overrightarrow{\text{himself}}$	93%	89%
$\overrightarrow{\text{daughter}}-\overrightarrow{\text{son}}$	93%	91%
$\overrightarrow{\text{mother}}-\overrightarrow{\text{father}}$	91%	85%
$\overrightarrow{\text{gal}}-\overrightarrow{\text{guy}}$	85%	85%
$\overrightarrow{\text{girl}}-\overrightarrow{\text{boy}}$	90%	86%
$\overrightarrow{\text{female}}-\overrightarrow{\text{male}}$	84%	75%



The top PC captures the gender subspace

Gender-definitional vs. Gender-neutral Words



Debiasing

1. Identify gender-definitional and gender-neutral words
2. Project away the gender subspace from the gender-neutral words
3. Normalize vectors

Debiasing

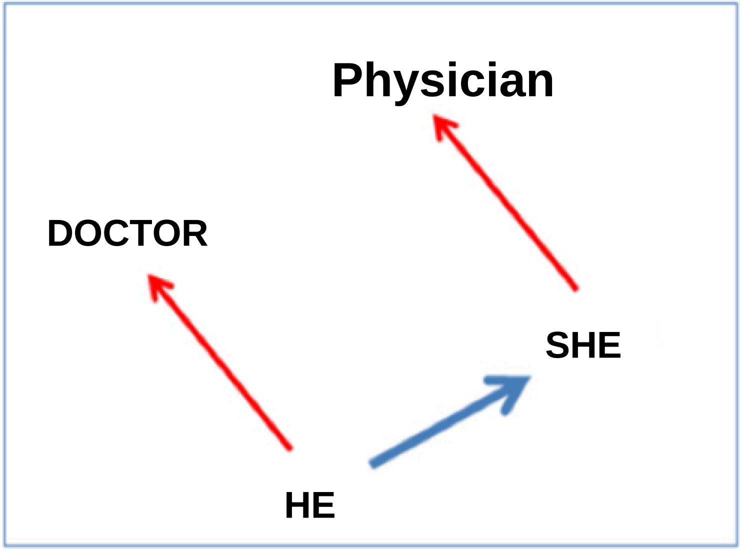
1. Identify gender-definitional and gender-neutral words
2. Project away the gender subspace from the gender-neutral words
3. Normalize vectors

2a. Transformation that seeks to preserve pairwise inner products between all the word vectors while minimizing the projection of the gender neutral words onto the gender subspace

$$min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{Don't modify embeddings too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{Minimize gender component}}$$

T - the desired debiasing transformation
W - embedding matrix

B – gender subspace
N – gender neutral words


Physician
DOCTOR
SHE
HE

Embeddings reflect cultural bias

TECHNISCHE UNIVERSITÄT DARMSTADT

Caliskan, A., Bryson, J. J. and Narayanan, A. (2017) **Semantics derived automatically from language corpora contain human-like biases.** *Science*

Implicit Association Test
Greenwald et al. 1998

Category	Items
Good	Spectacular, Appealing, Love, Triumph, Joyous, Fabulous, Excitement, Excellent
Bad	Angry, Disgust, Rotten, Selfish, Abuse, Dirty, Hatred, Ugly
African Americans	
European Americans	

Psychological findings on US participants:

TECHNISCHE UNIVERSITÄT DARMSTADT

- African-American names are associated with unpleasant words (more than European-American names)
- Male names associated more with math, female names with arts
- Old people's names with unpleasant words, young people with pleasant words.

Caliskan et al.’s replication with embeddings:

- African-American names (Leroy, Shaniqua) had a higher GloVe cosine with unpleasant words (abuse, stink, ugly)
- European American names (Brad, Greg, Courtney) had a higher cosine with pleasant words (love, peace, miracle)

Ethnic and Gender Stereotypes Over Time

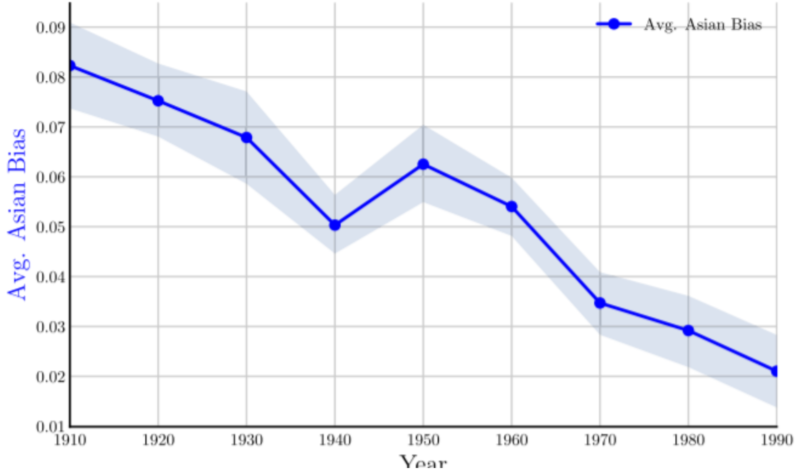
Garg, Nikhil, Schiebinger, Londa, Jurafsky, Dan, and Zou, James (2018). **Word embeddings quantify 100 years of gender and ethnic stereotypes**. PNAS, 115(16), E3635–E3644

Ethnic and Gender Stereotypes Over Time

- Embeddings for competence adjectives are biased toward men
  - Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.
  - This bias is slowly decreasing
- Embeddings reflect ethnic stereotypes over time
  - Asian Bias 1910-1990
  - Change in association of Chinese names with adjectives framed as "othering" (barbaric, monstrous, bizarre)

Ethnic and Gender Stereotypes Over Time

- Embeddings for competence adjectives are biased toward men
  - Smart, wise, brilliant, intelligent, resourceful, thoughtful, logical, etc.
  - This bias is slowly decreasing
- Embeddings reflect ethnic stereotypes over time
  - Asian Bias 1910-1990
  - Change in association of Chinese name
    (barbaric, monstrous, bizarre)



Where to look for “bias” in NLP

- Problem definition/research question (who will benefit? who can be harmed?)
- Data
  - Who is described (when some populations are excluded or underrepresented)
  - How training data might describe populations in biased ways
  - Who authored the data
- Data labels
  - Annotation schema (e.g., binary gender labels)
  - Annotation instructions
  - Annotator bias
- Model design
  - Biased objective (e.g., COMPAS system for parole decisions)
  - Spurious correlations (e.g., correlations of ethnicity and sentiment labels, gendered pronouns and professions in coreference links)
- Model outputs
  - Bias amplification
  - Disparities in model utility and fairness by populations

Take-Home Message

- Debiasing a dataset is NOT done by just deleting the relevant feature
  - Information often correlates with (a combination of) other features
- Equal precision OR recall does not ensure unbiased ML models
  - What is more important? What is the cost of False Positives / Fale Negatives?
- Bias can be amplified by machine learning
- Word Embeddings can reflect / measure social bias

Next Lecture

InCivility and Hate Speech I