ethics in NLP

CS 685, Fall 2021

Introduction to Natural Language Processing

http://people.cs.umass.edu/~miyyer/cs685/

Mohit Iyyer

College of Information and Computer Sciences

University of Massachusetts Amherst

many slides from Yulia Tsvetkov & Mark Yatskar

OpenAI PALMS: https://openai.com/blog/improving-language-model-behavior/

Demo: https://delphi.allenai.org/

what are we talking about today?

- many NLP systems affect <u>actual people</u>
  - systems that interact with people (conversational agents)
  - perform some reasoning over people (e.g., recommendation systems, targeted ads)
  - make decisions about people's lives (e.g., parole decisions, employment, immigration)

- questions of *ethics* arise in all of these applications!

why are we talking about it?

- the explosion of data, in particular user-generated data (e.g., social media)
- machine learning models that leverage huge amounts of this data to solve certain tasks

Learn to Assess AI Systems Adversarially

- Who could benefit from such a technology?
- Who can be harmed by such a technology?

- Representativeness of training data
- Could sharing this data have major effect on people's lives?

- What are confounding variables and corner cases to control for?
- Does the system optimize for the "right" objective?
- Could prediction errors have major effect on people's lives?

let’s start with the data…

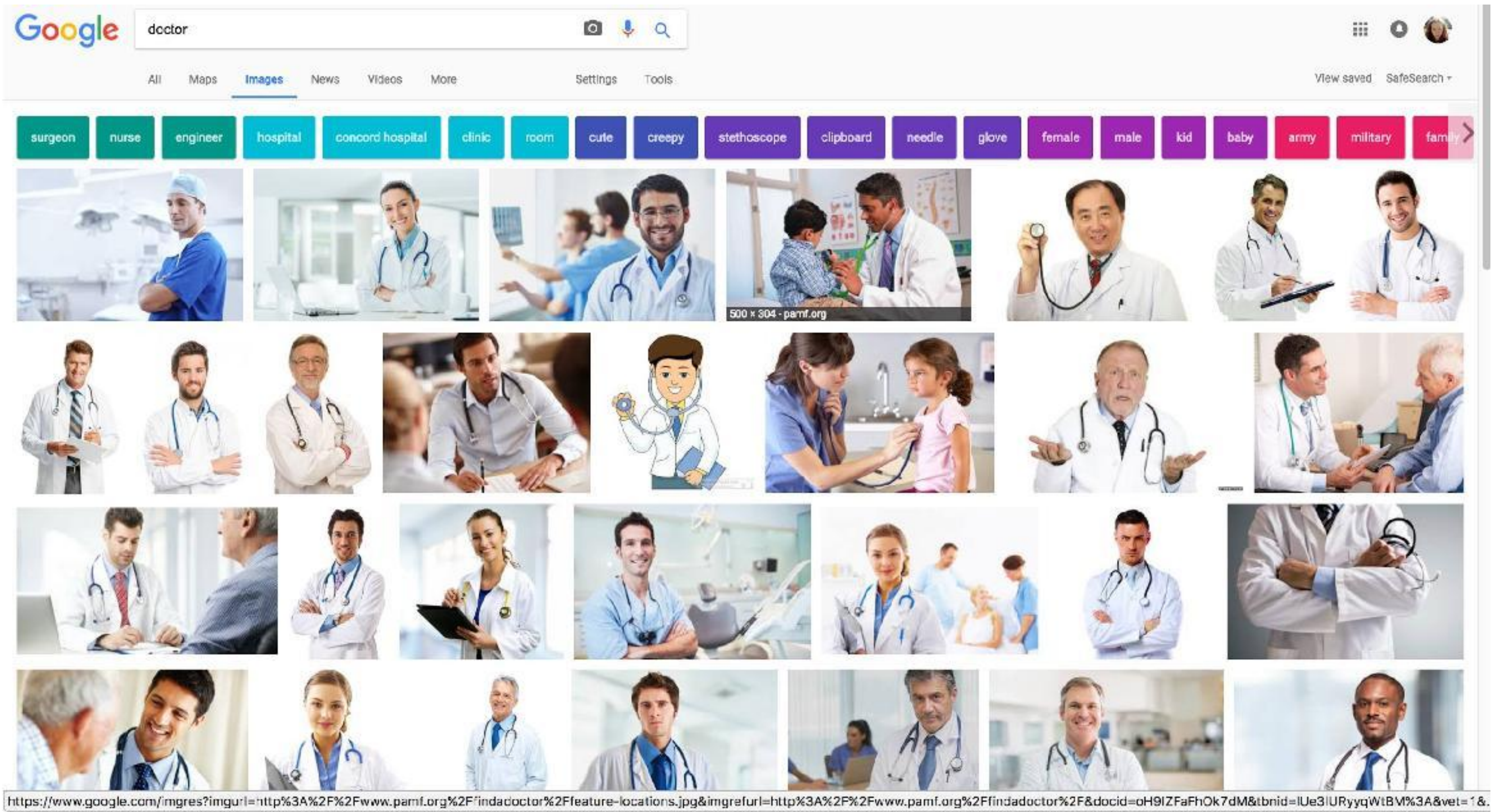
Online data is riddled with **SOCIAL STEREOTYPES**

Racial Stereotypes

- June 2016: web search query “three black teenagers”

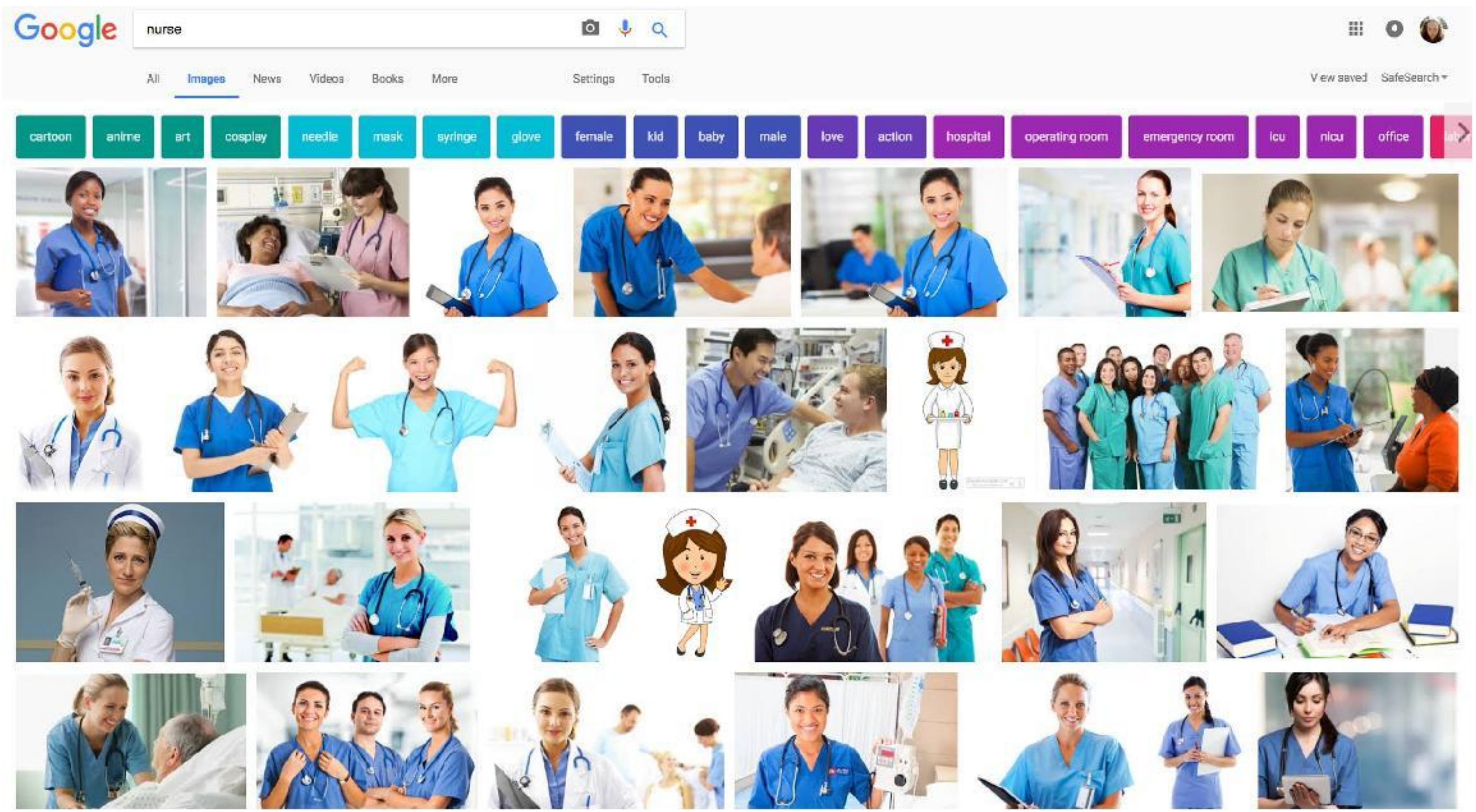
Gender/Race/Age Stereotypes

- June 2017: image search query "Doctor"



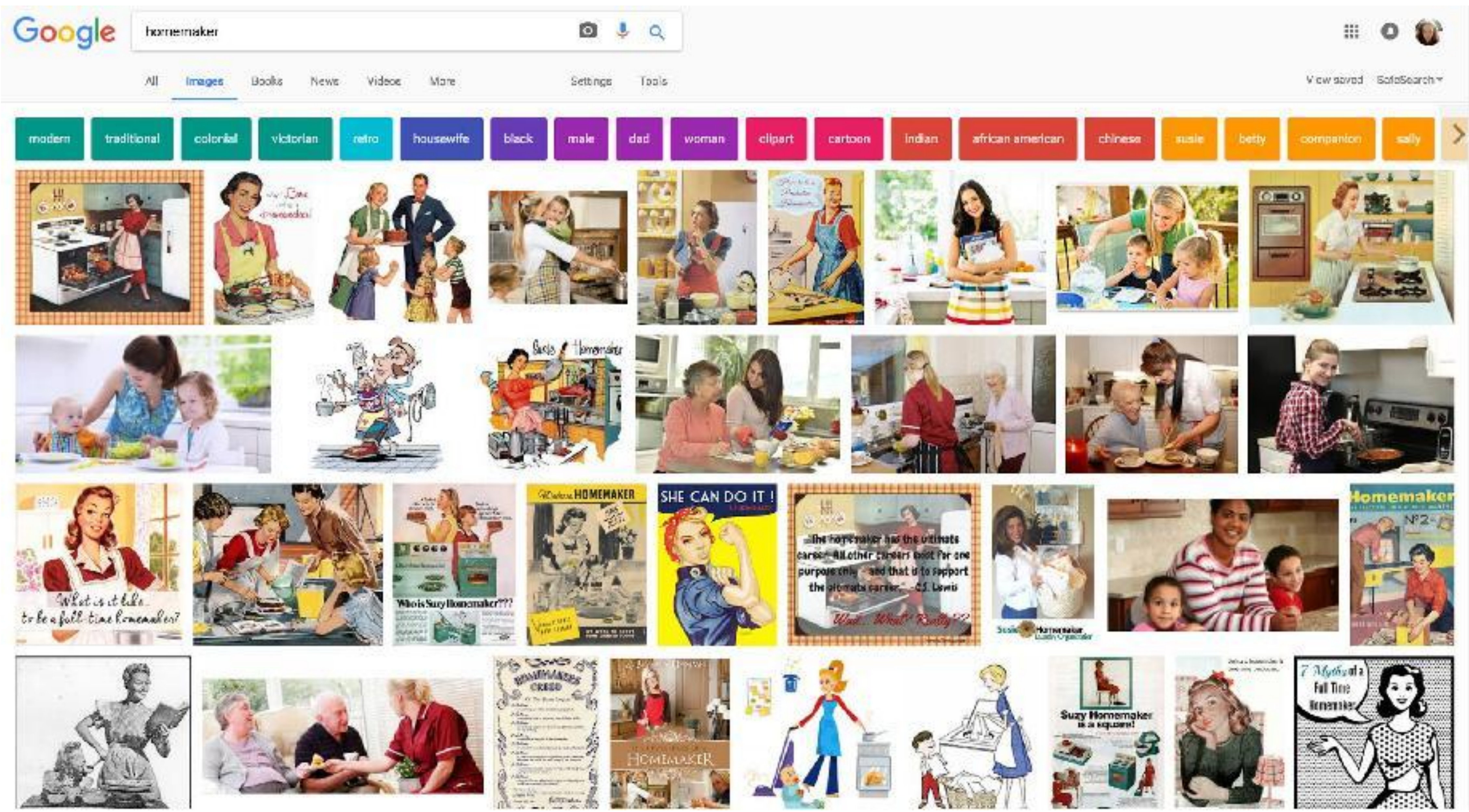
Gender/Race/Age Stereotypes

- June 2017: image search query "Nurse"



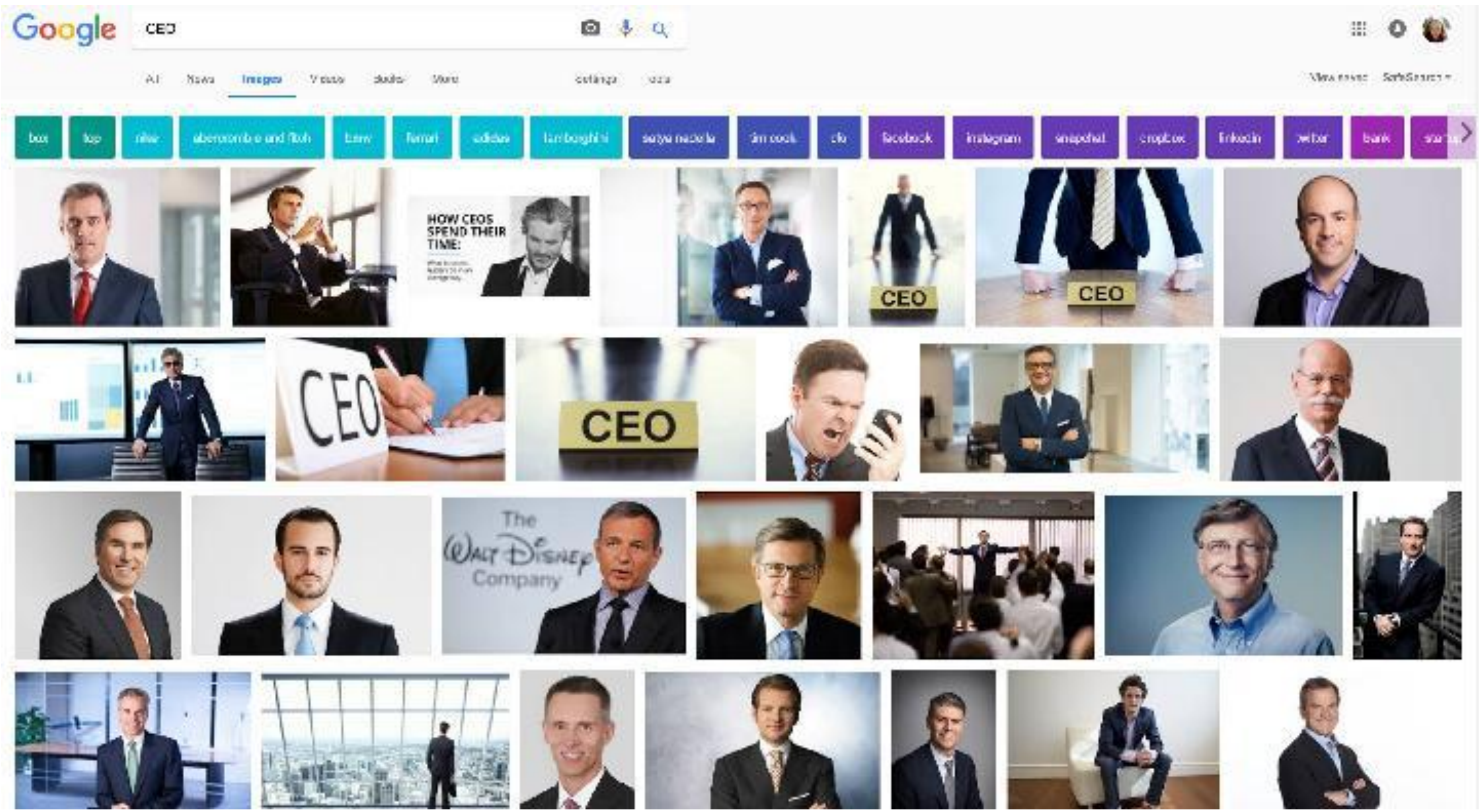
Gender/Race/Age Stereotypes

- June 2017: image search query “Homemaker”



Gender/Race/Age Stereotypes

- June 2017: image search query “CEO”



A
I

Consequence: models are biased

Gender Biases on the Web

- The dominant class is often portrayed and perceived as relatively more professional (Kay, Matuszek, and Munson 2015)
- Males are over-represented in the reporting of web-based news articles (Jia, Lansdall-Welfare, and Cristianini 2015)
- Males are over-represented in twitter conversations (Garcia, Weber, and Garimella 2014)
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues (Wagner et al. 2015)
- IMDB reviews written by women are perceived as less useful (Otterbacher 2013)

Biased NLP Technologies

- Bias in word embeddings (Bolukbasi et al. 2017; Caliskan et al. 2017; Garg et al. 2018)
- Bias in Language ID (Blodgett & O'Connor. 2017; Jurgens et al. 2017)
- Bias in Visual Semantic Role Labeling (Zhao et al. 2017)
- Bias in Natural Language Inference (Rudinger et al. 2017)
- Bias in Coreference Resolution (At NAACL: Rudinger et al. 2018; Zhao et al. 2018)
- Bias in Automated Essay Scoring (At NAACL: Amorim et al. 2018)

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

The physician hired the secretary because he was highly recommended.

Zhao et al., NAACL 2018

Sources of Human Biases in Machine Learning

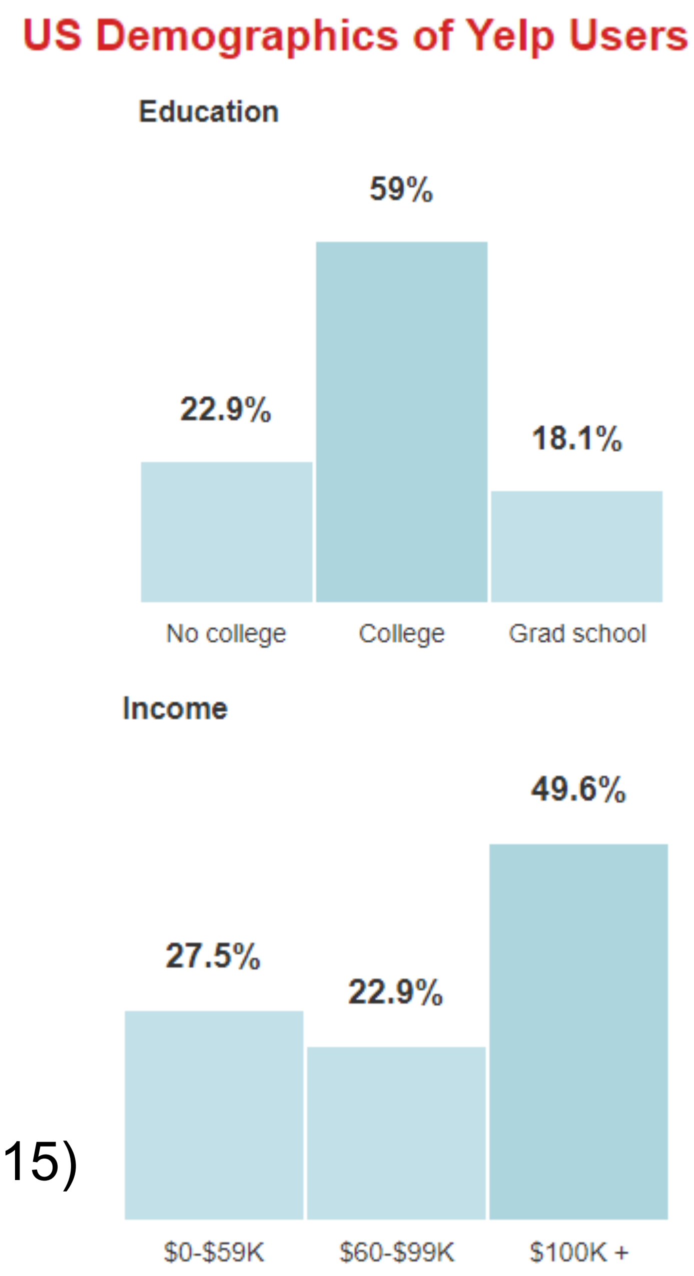
- Bias in data and sampling
- Optimizing towards a biased objective
- Inductive bias
- Bias amplification in learned models

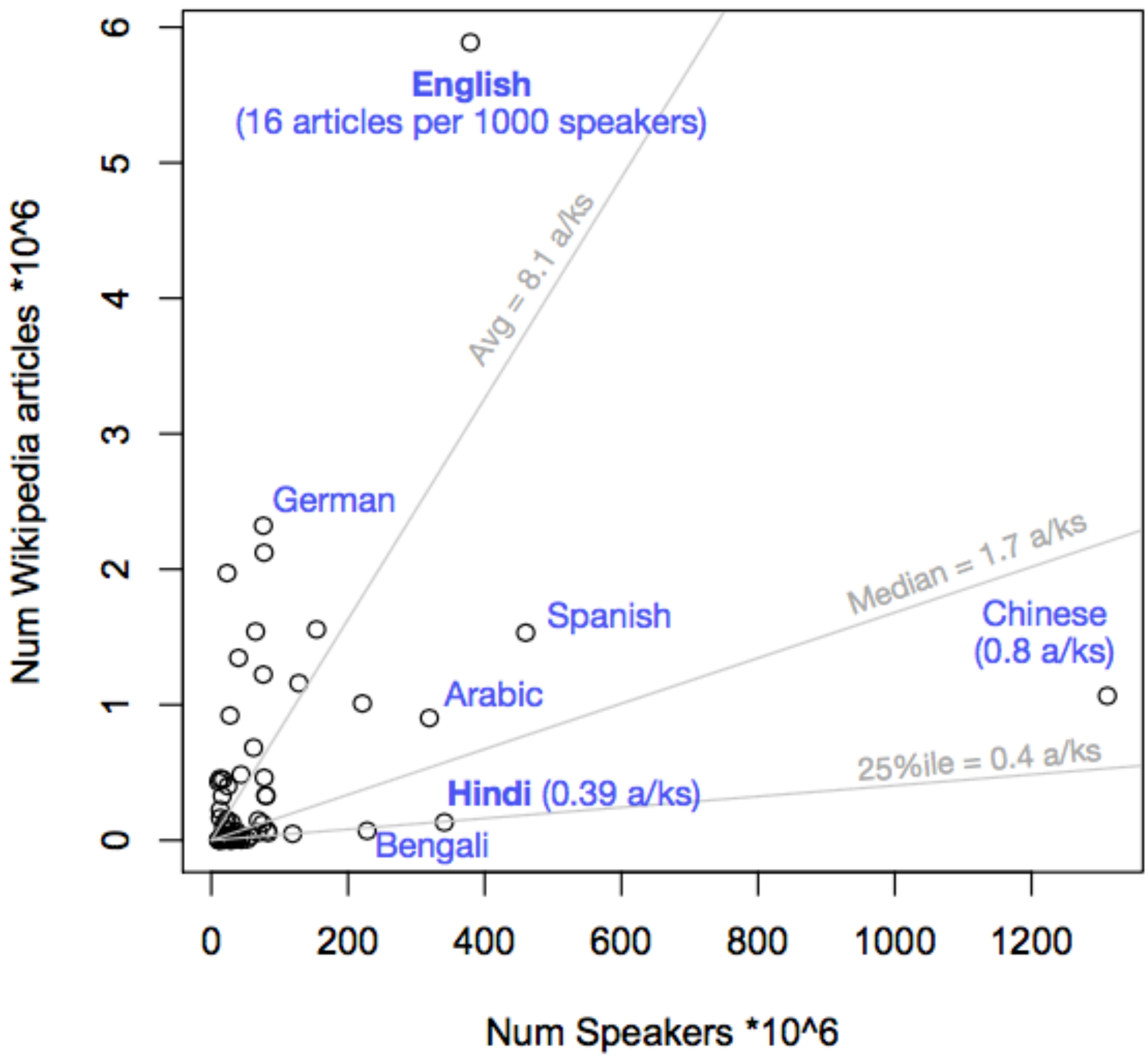
Sources of Human Biases in Machine Learning

- **Bias in data and sampling**
- Optimizing towards a biased objective
- Inductive bias
- Bias amplification in learned models

Types of Sampling Bias in Naturalistic Data

- Self-Selection Bias
  - Who decides to post reviews on Yelp and why? Who posts on Twitter and why?
- Reporting Bias
  - People do not necessarily talk about things in the world in proportion to their empirical distributions (Gordon and Van Durme 2013)
- Proprietary System Bias
  - What results does Twitter return for a particular query of interest and why? Is it possible to know?
- Community / Dialect / Socioeconomic Biases
  - What linguistic communities are over- or under-represented? leads to community-specific model performance (Jorgensen et al. 2015)





credit: Brendan O'Connor

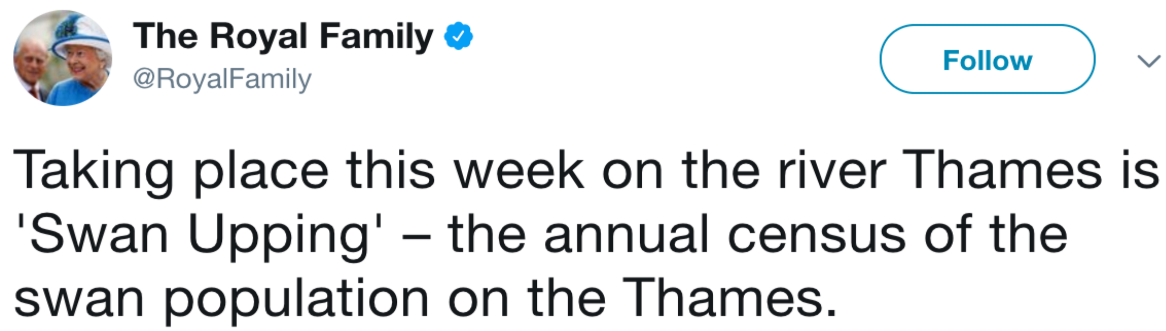
Example: Bias in Language Identification

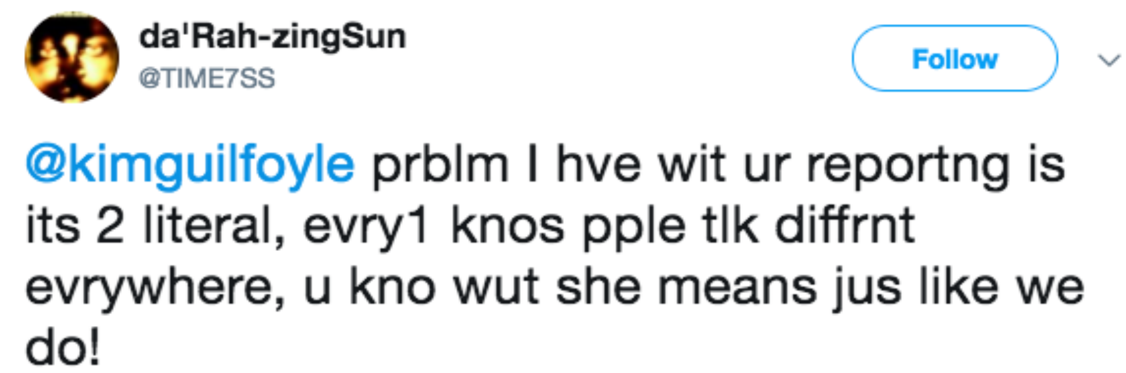
- Most applications employ off-the-shelf LID systems which are highly accurate

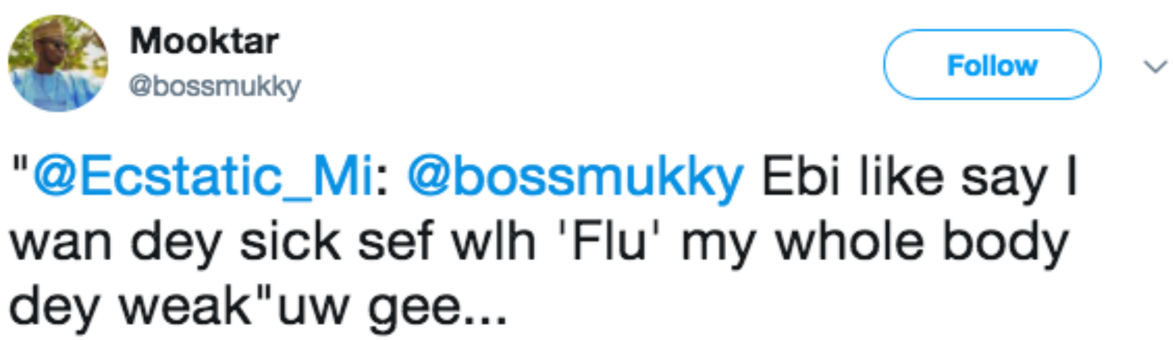
*Slides on LID by David Jurgens
(Jurgens et al. ACL’17)

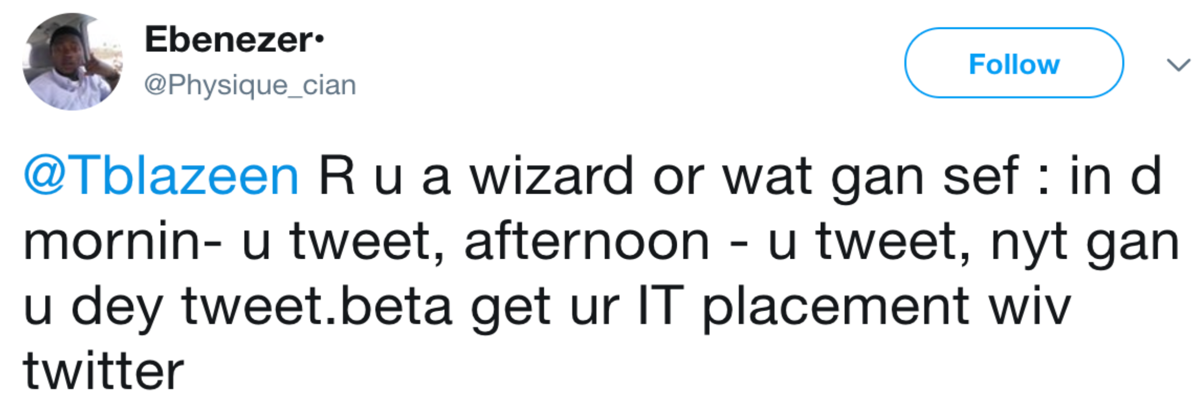
McNamee, P., “Language identification: *a solved problem* suitable for undergraduate instruction” Journal of Computing Sciences in Colleges 20(3) 2005.

“This paper describes […] how even the most simple of these methods **using data obtained from the World Wide Web achieve accuracy approaching 100%** on a test suite comprised of ten European languages”









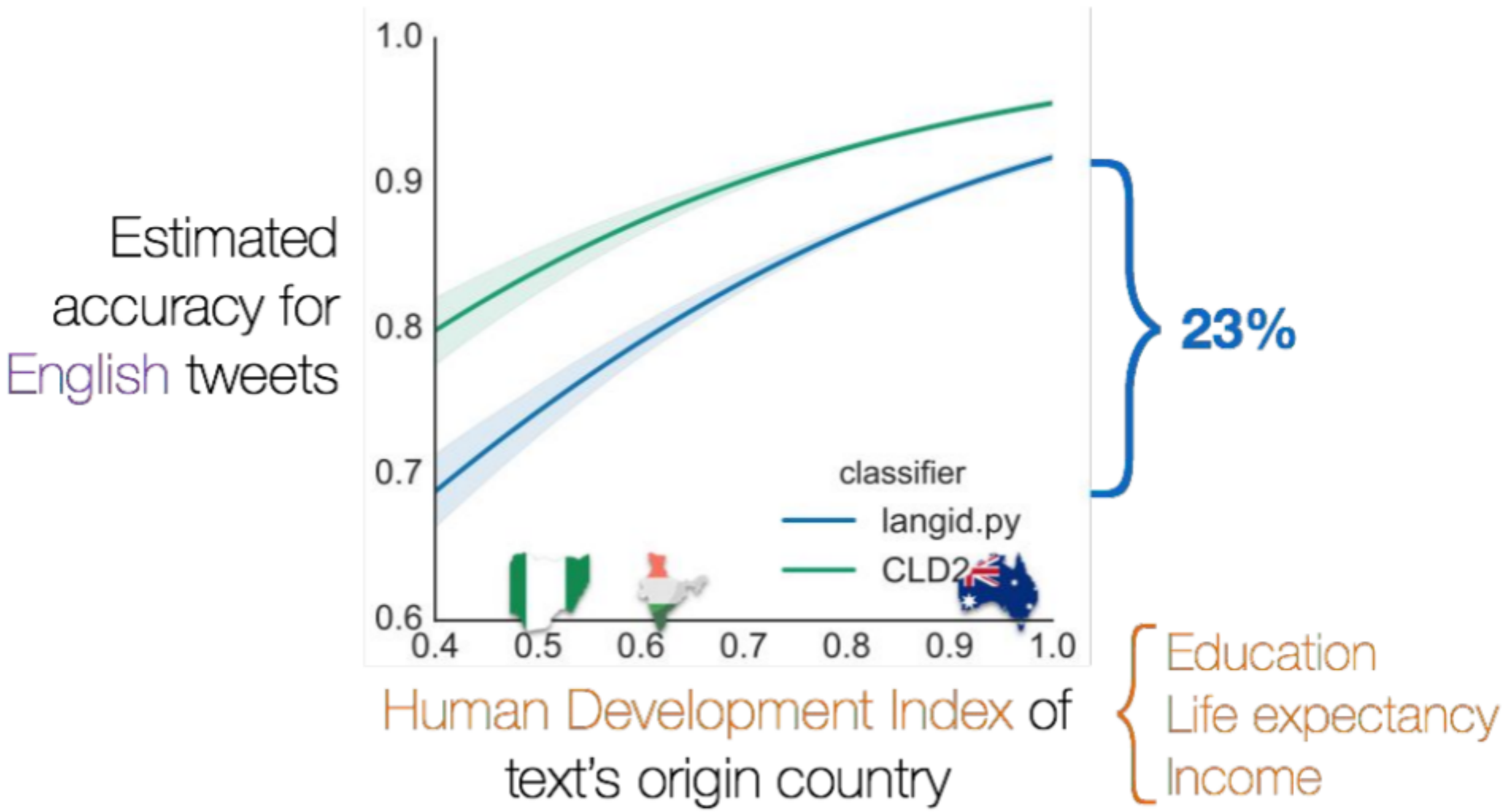
- Language identification degrades significantly on African American Vernacular English
  (Blodgett et al. 2016) **Su-Lin Blodgett just got her PhD from UMass!**

LID Usage Example: Health Monitoring

LID Usage Example: Health Monitoring

Socioeconomic Bias in Language Identification

- Off-the-shelf LID systems under-represent populations in less-developed countries



Jurgens et al. ACL'17

Better Social Representation through Network-based Sampling

- Re-sampling from strategically-diverse corpora

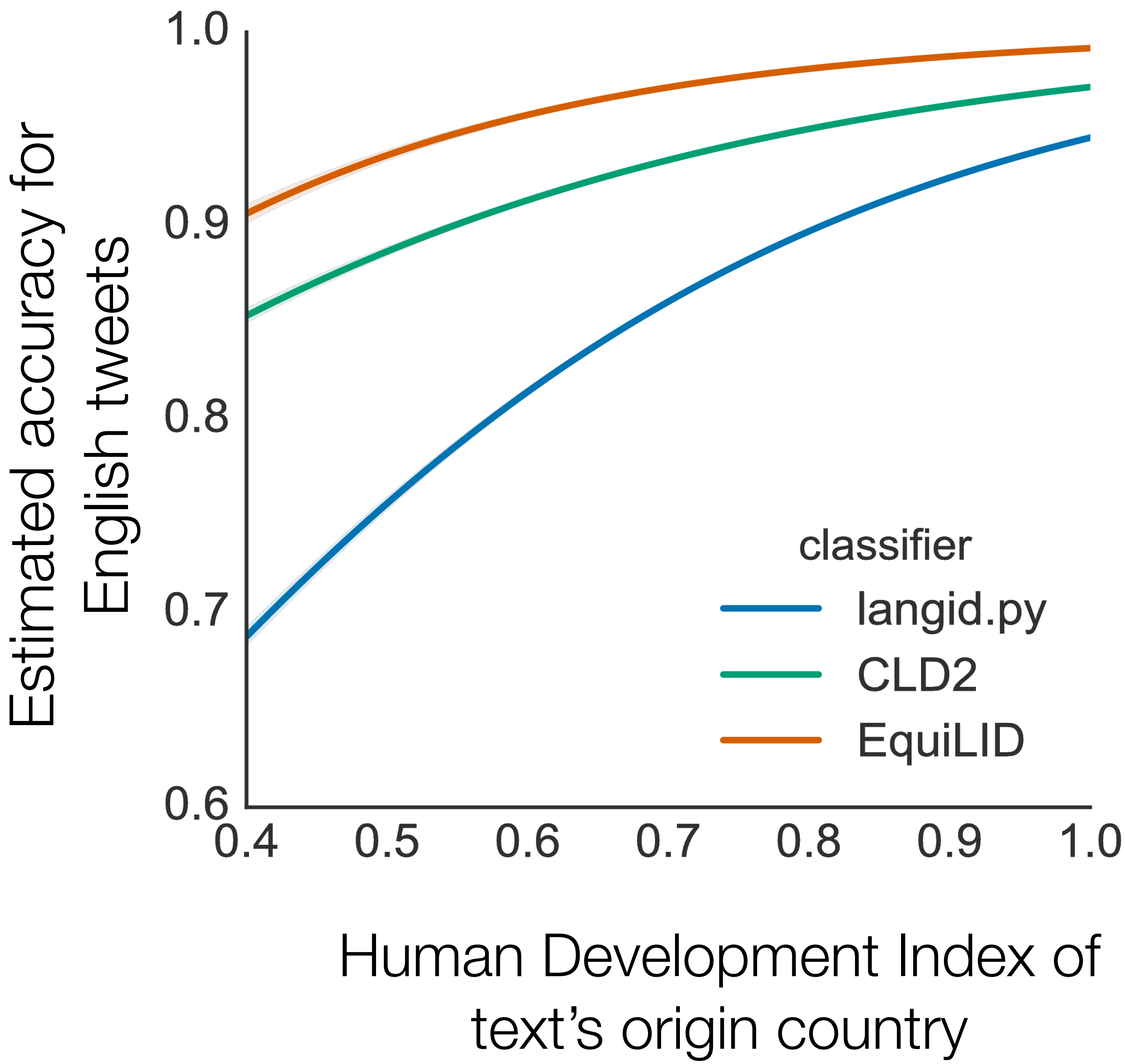
Topical

Geographic

Social

Multilingual

Jurgens et al. ACL'1



Jurgens et al. ACL'17

Sources of Human Biases in Machine Learning

- Bias in data and sampling
- **Optimizing towards a biased objective**
- Inductive bias
- Bias amplification in learned models

Optimizing Towards a Biased Objective

- Northpointe   vs   ProPublica

COMPAS

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?”

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?”

- COMPAS system
  - balanced training data about people of all races
  - race was *not* one of the input features
- Objective function
  - labels for “who will commit a crime” are unobtainable
  - a proxy for the real, unobtainable data: “who is more likely to be *convicted*”

what are some issues with this proxy objective?

Predicting prison sentences given case descriptions

Case description: On July 7, 2017, when the defendant Cui XX was drinking in a bar, he came into conflict with Zhang XX…… After arriving at the police station, he refused to cooperate with the policeman and bited on the arm of the policeman……

Result of judgment: Cui XX was sentenced to *12* months imprisonment for *creating disturbances* and *12* months imprisonment for *obstructing public affairs*……

- Charge#1 creating disturbances term 12 months
- Charge#2 obstructing public affairs term 12 months

Chen et al., EMNLP 2019, "Charge-based prison term prediction…"

Is this sufficient consideration of ethical issues of this work? Should the work have been done at all?

The mistake of legal judgment is serious, it is about people losing years of their lives in prison, or dangerous criminals being released to reoffend. We should pay attention to how to avoid judges' over-dependence on the system. It is necessary to consider its application scenarios. In practice, we recommend deploying our system in the "Review Phase", where other judges check the judgment result by a presiding judge. Our system can serve as one anonymous checker.

Chen et al., EMNLP 2019, "Charge-based prison term prediction…"

Sources of Human Biases in Machine Learning

- Bias in data and sampling
- Optimizing towards a biased objective
- **Inductive bias**
- Bias amplification in learned models

what is inductive bias?

- the assumptions used by our model. examples:
  - recurrent neural networks for NLP assume that the sequential ordering of words is meaningful
  - features in discriminative models are assumed to be useful to map inputs to outputs

Bias in Word Embeddings

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

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

Biases in Embeddings: Another Take

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

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

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

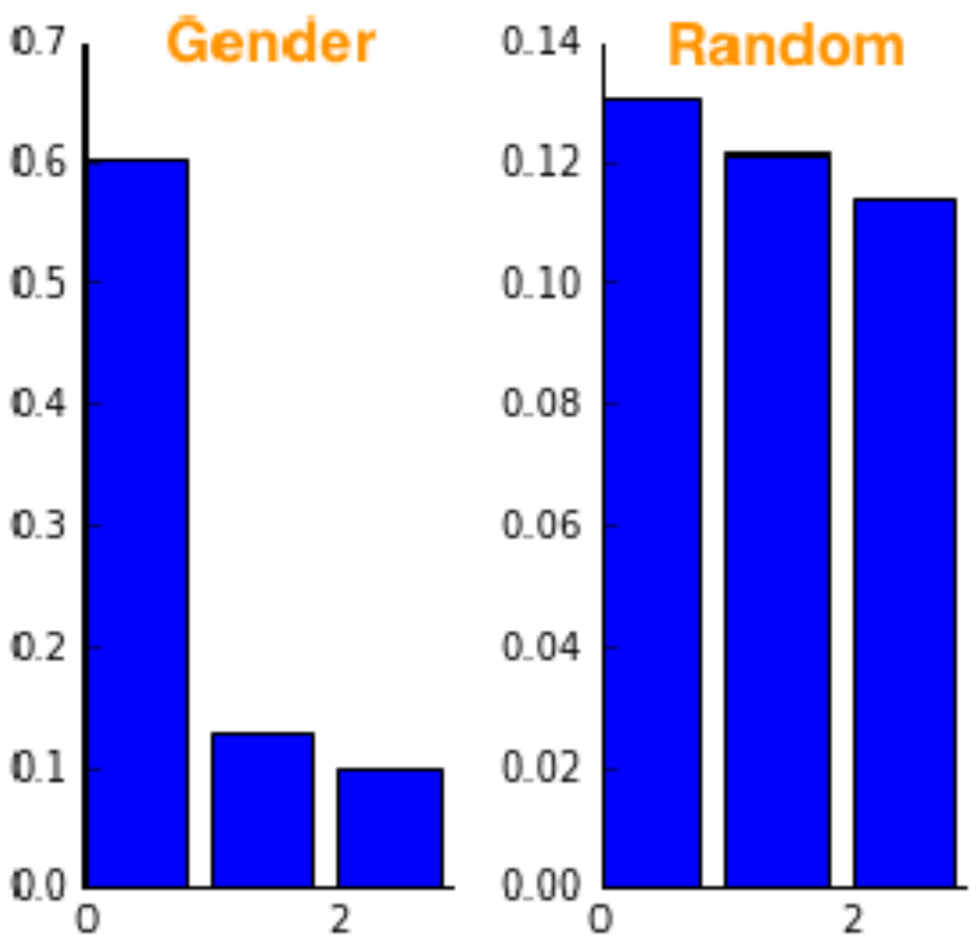
Figure 1: **Left** The most extreme occupations as projected on to the *she*−*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

Towards Debiasing

1. Identify gender subspace: B

Gender Subspace

$$\overrightarrow{\mathrm{she}} - \overrightarrow{\mathrm{he}}$$
$$\overrightarrow{\mathrm{her}} - \overrightarrow{\mathrm{his}}$$
$$\overrightarrow{\mathrm{woman}} - \overrightarrow{\mathrm{man}}$$
$$\overrightarrow{\mathrm{Mary}} - \overrightarrow{\mathrm{John}}$$
$$\overrightarrow{\mathrm{herself}} - \overrightarrow{\mathrm{himself}}$$
$$\overrightarrow{\mathrm{daughter}} - \overrightarrow{\mathrm{son}}$$
$$\overrightarrow{\mathrm{mother}} - \overrightarrow{\mathrm{father}}$$
$$\overrightarrow{\mathrm{gal}} - \overrightarrow{\mathrm{guy}}$$
$$\overrightarrow{\mathrm{girl}} - \overrightarrow{\mathrm{boy}}$$
$$\overrightarrow{\mathrm{female}} - \overrightarrow{\mathrm{male}}$$

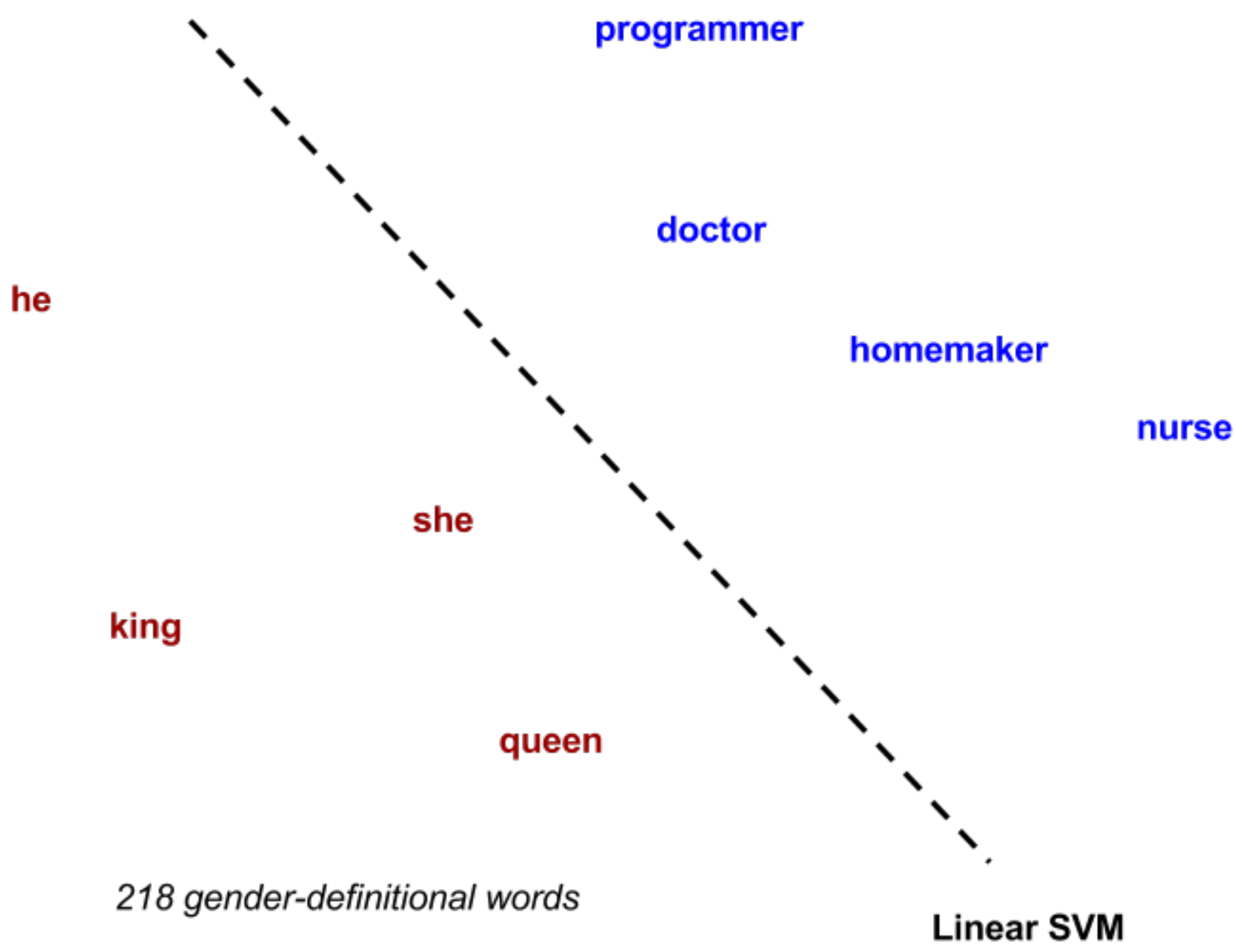


The top PC captures the gender subspace

Towards Debiasing

1. Identify gender subspace: B
2. **Identify gender-definitional (S) and gender-neutral words (N)**

Gender-definitional vs. Gender-neutral Words



Towards Debiasing

1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply transform matrix (T) to the embedding matrix (W) such that
   a. Project away the gender subspace B from the gender-neutral words N
   b. But, ensure the transformation doesn't change the embeddings too much

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

T - the desired debiasing transformation

B - biased space

W - embedding matrix

N - embedding matrix of gender neutral words

Sources of Human Biases in Machine Learning

- Bias in data and sampling
- Optimizing towards a biased objective
- Inductive bias
- **Bias amplification in learned models**

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)

Internet

FrameNet

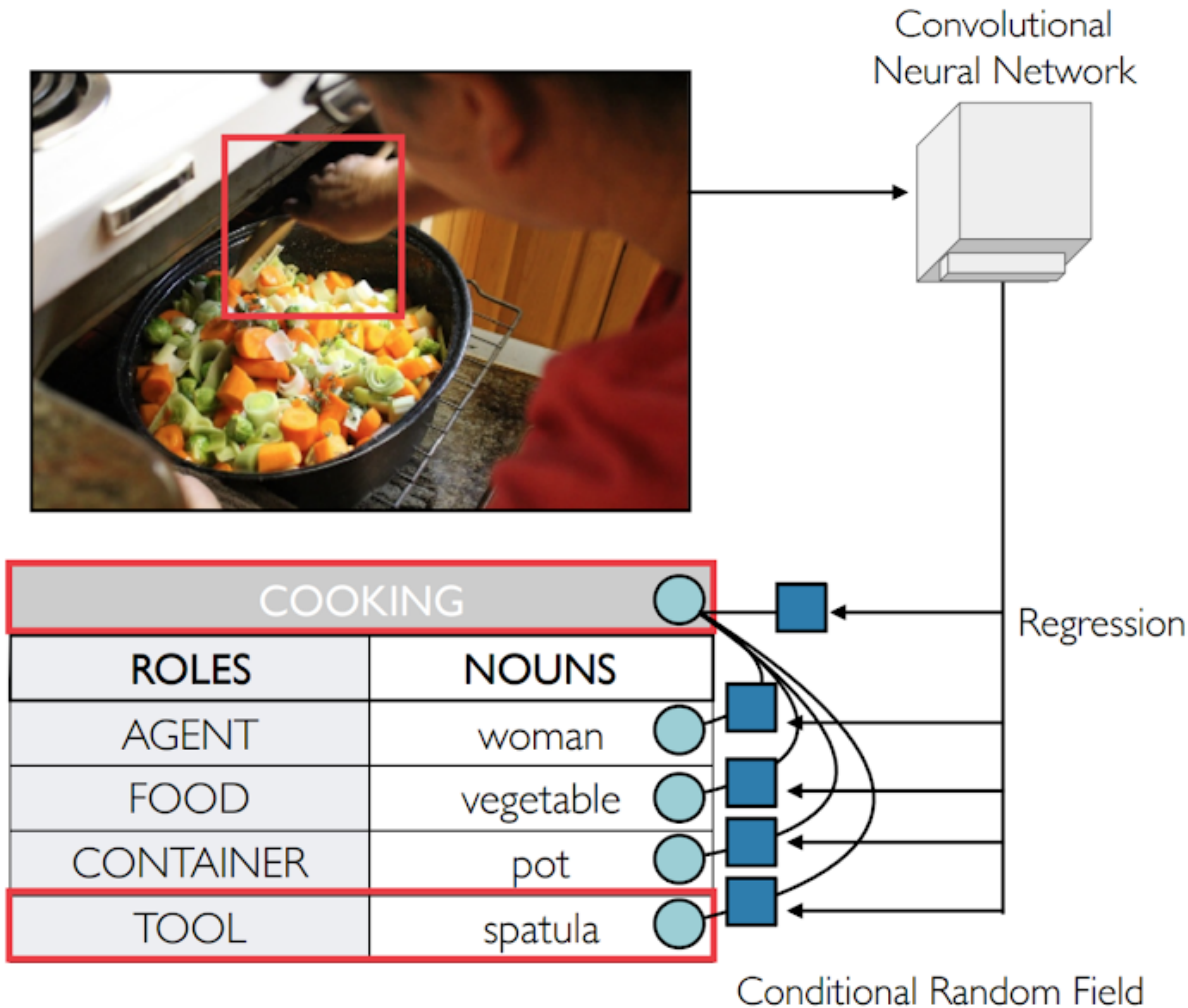
WordNet

COOKING	
ROLES	**NOUNS**
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

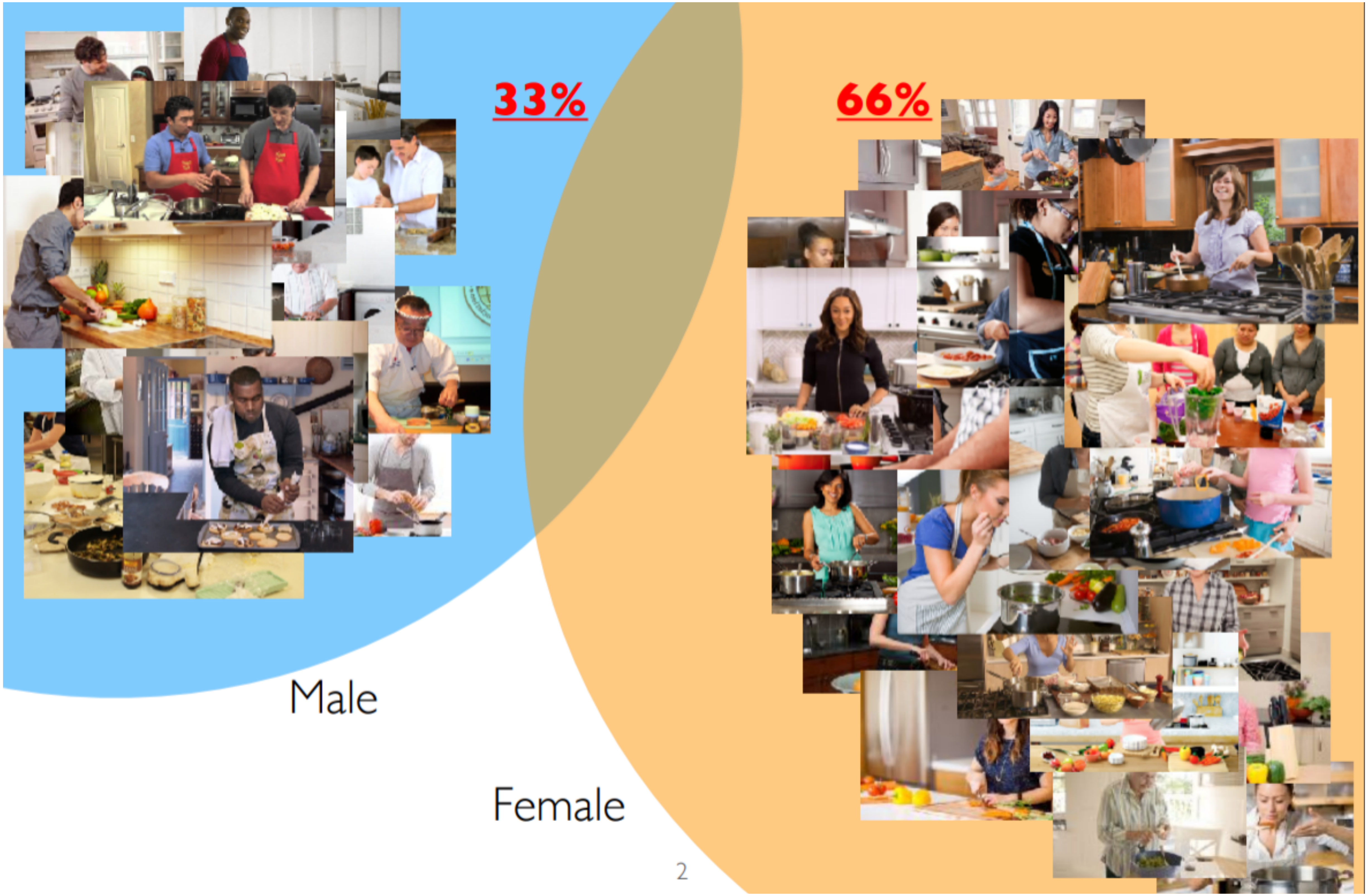
Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16

Slides by Mark Yatskar https://homes.cs.washington.edu/~my89/talks/ZWYOC17_slide.pdf

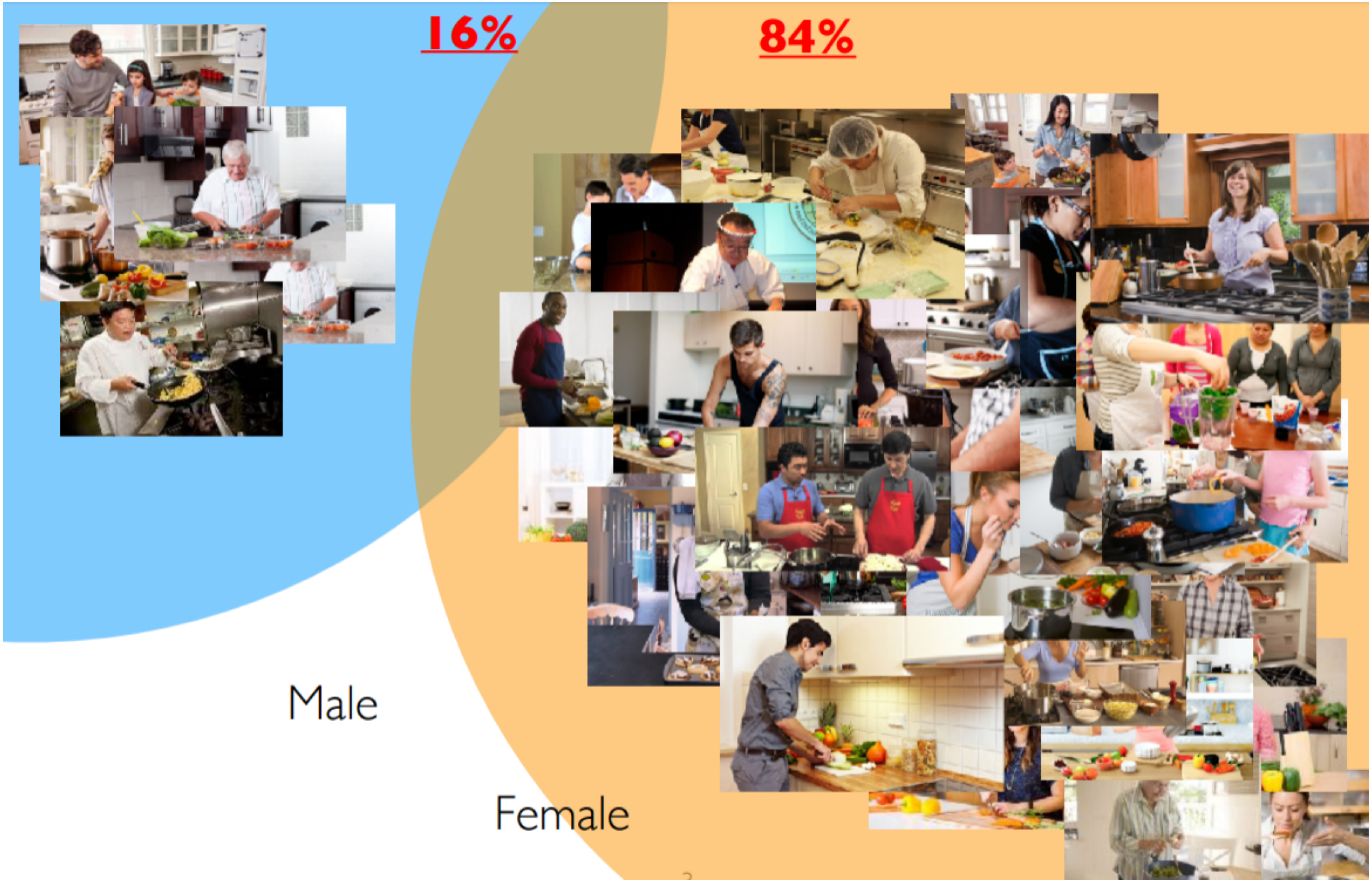
imSitu Visual Semantic Role Labeling (vSRL)



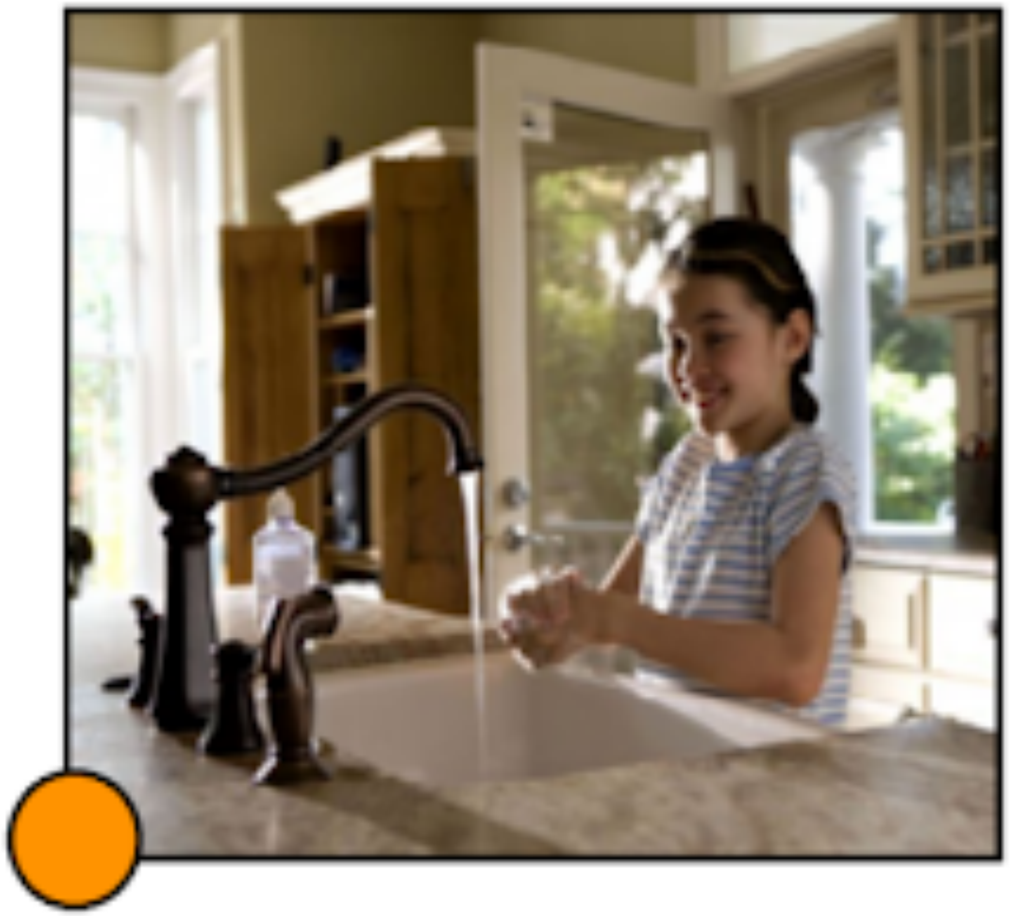
Dataset Gender Bias



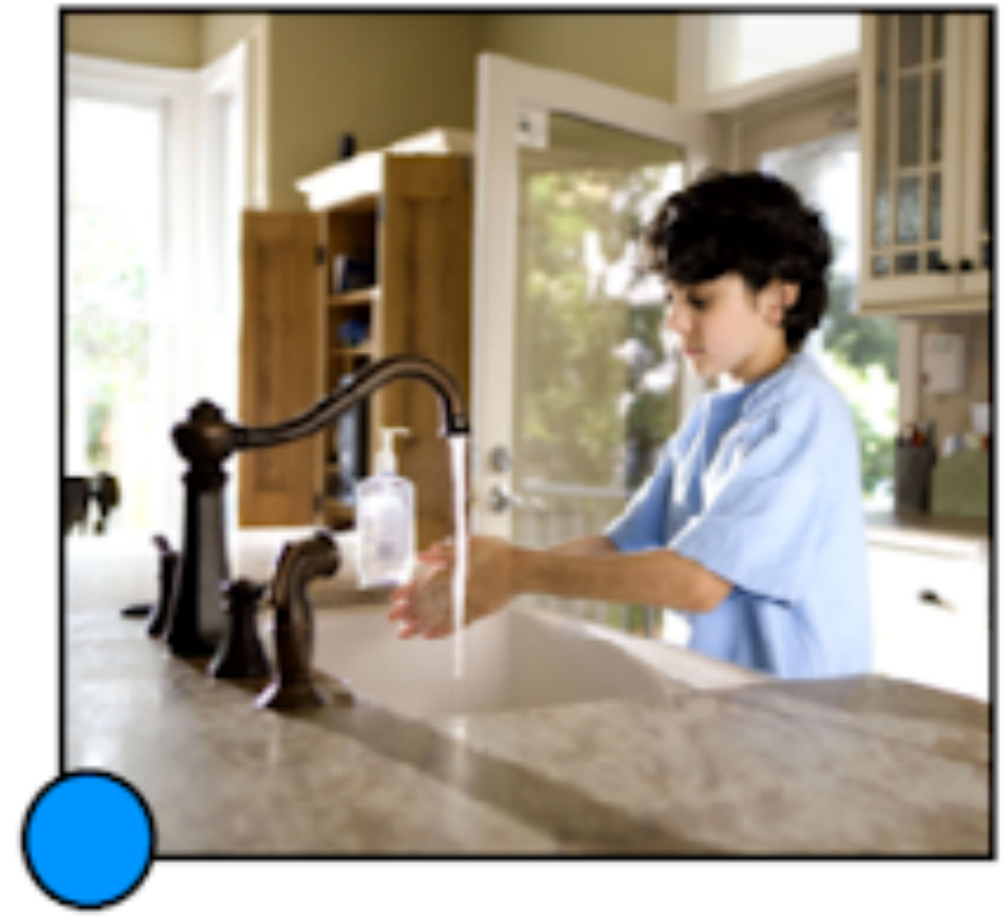
Model Bias After Training



Algorithmic Bias



woman cooking



man fixing faucet

Quantifying Dataset Bias

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

b(o,g)

Quantifying Dataset Bias

Training Gender Ratio (◆ verb)

Training Set

- ◆ cooking
- ● woman
- ● man

COOKING	
ROLES	NOUNS
● AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
● AGENT	man
FOOD	noodle

$$\frac{\#(\text{◆ cooking}, \text{● man})}{\#(\text{◆ cooking}, \text{● man}) + \#(\text{◆ cooking}, \text{● woman})} = 1/3$$

Quantifying Dataset Bias: Dev Set

Predicted Gender Ratio (verb)

Development Set

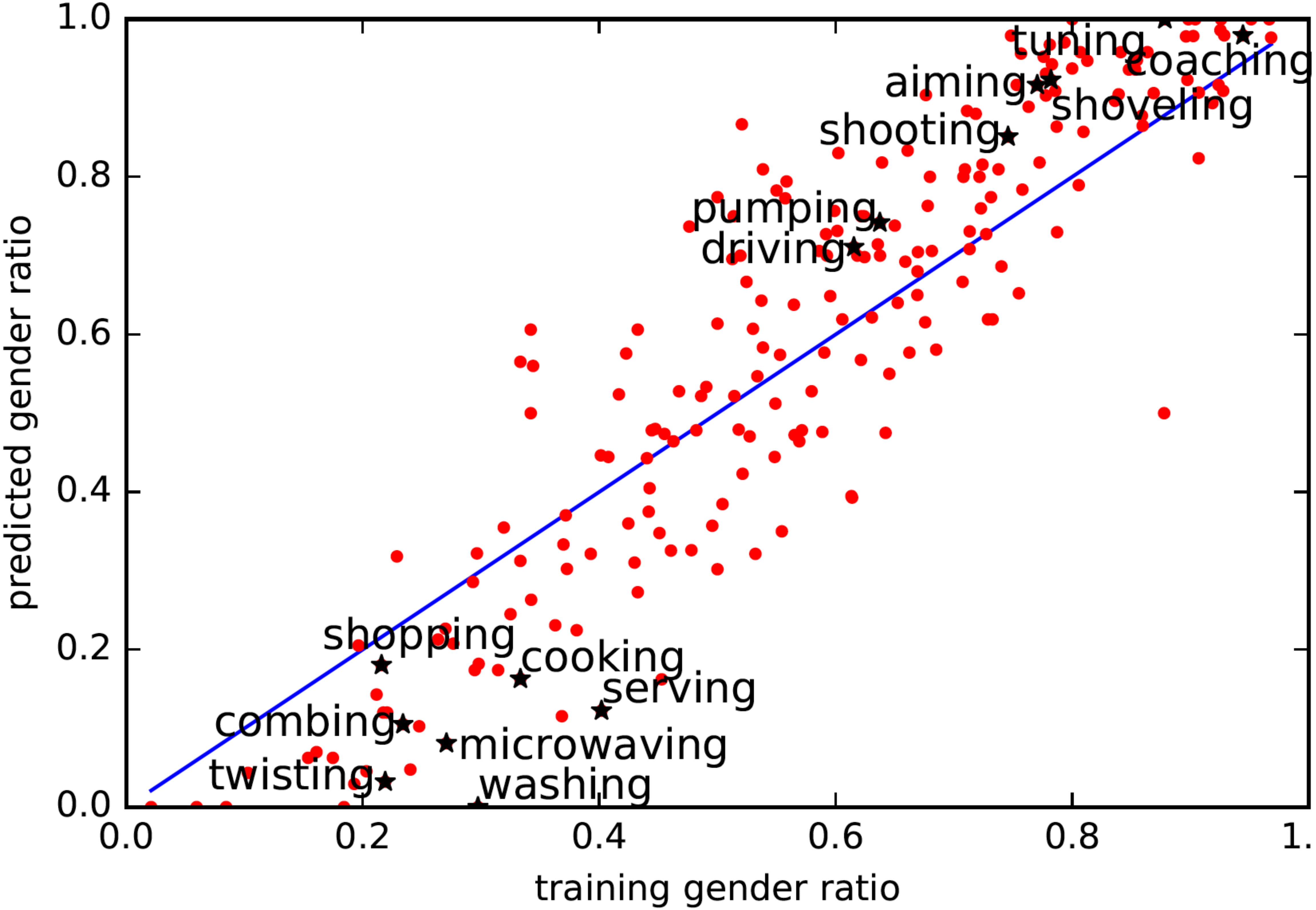
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

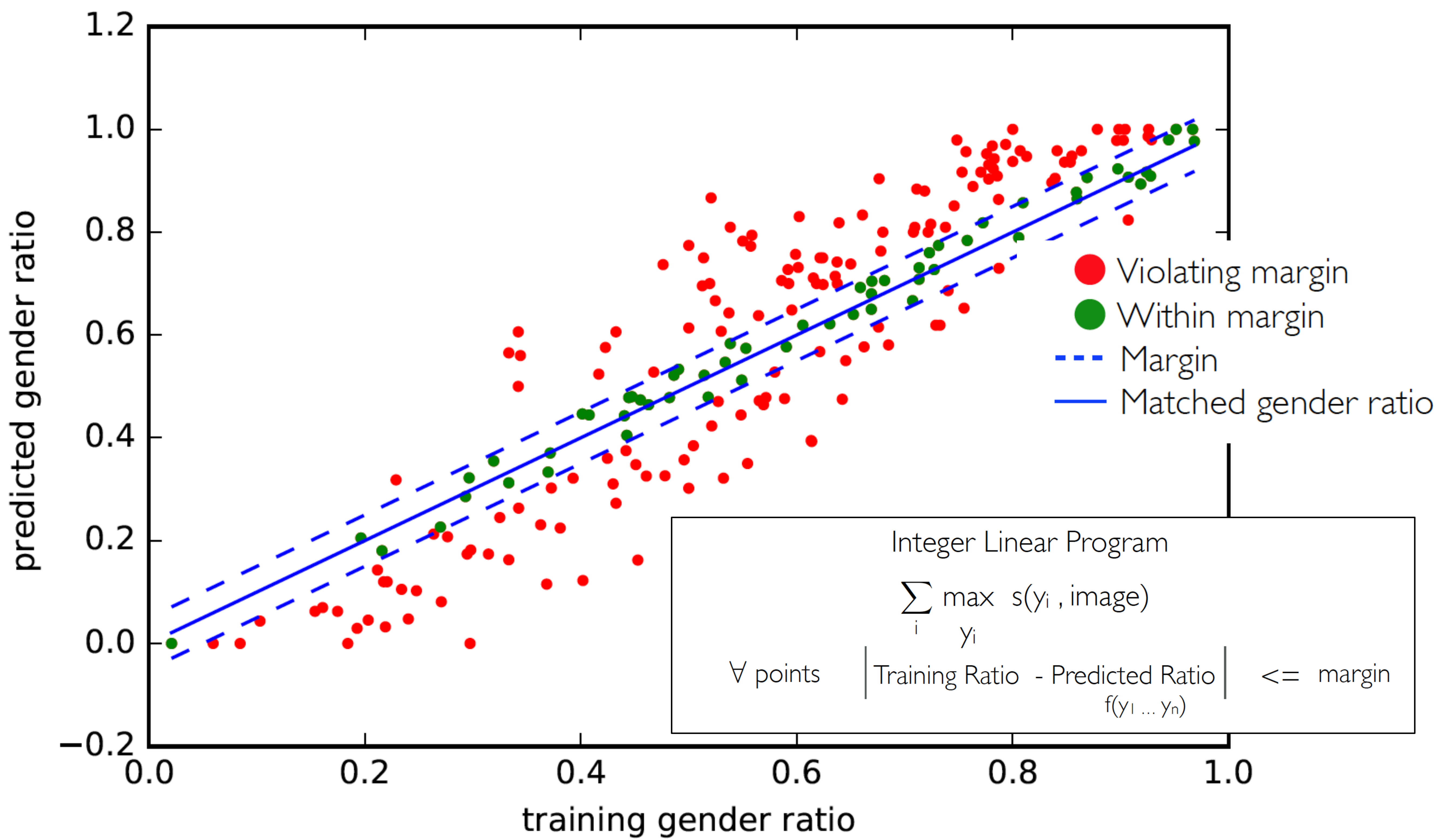
COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

$$\frac{\#(\text{cooking}, \text{man})}{\#(\text{cooking}, \text{man}) + \#(\text{cooking}, \text{woman})} = 1/6$$

Model Bias Amplification

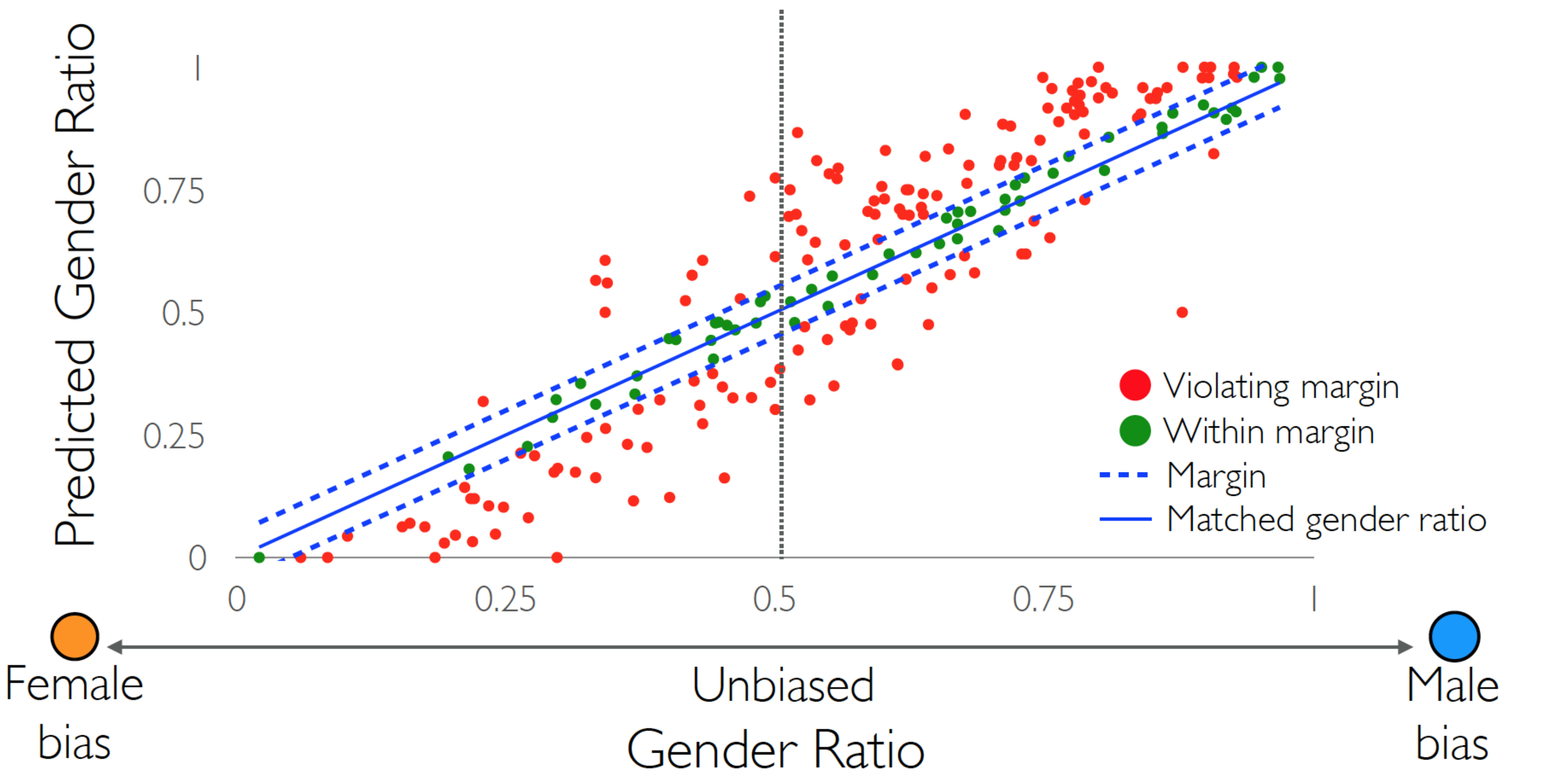


Reducing Bias Amplification (RBA)



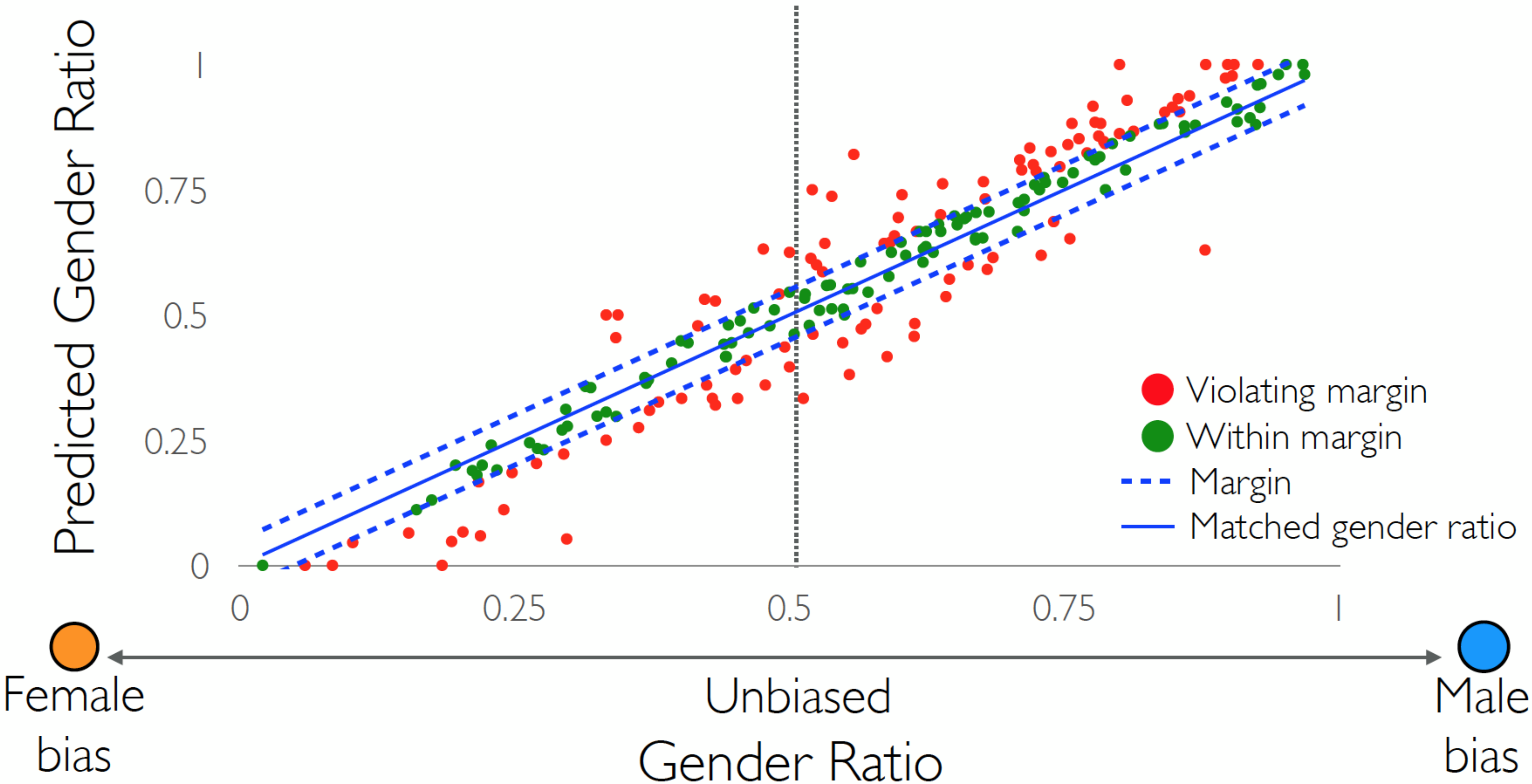
Results

imSitu Verb   Violation: 72.6%   .050 |bias↑| 24.07 acc.



Results

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



Discussion

- Applications that are built from online data, generated by people, learn also real-world stereotypes
- Should our ML models represent the “real world”?
- Or should we artificially skew data distribution?
- If we modify our data, what are guiding principles on what our models should or shouldn't learn?