Advanced NLP

Biases in AI and NLP systems, Debiasing, Perspectivism

Maarten Sap (Assistant Prof LTI)

(based on slides by Yulia Tsvetkov & Alan Black)

Carnegie Mellon University

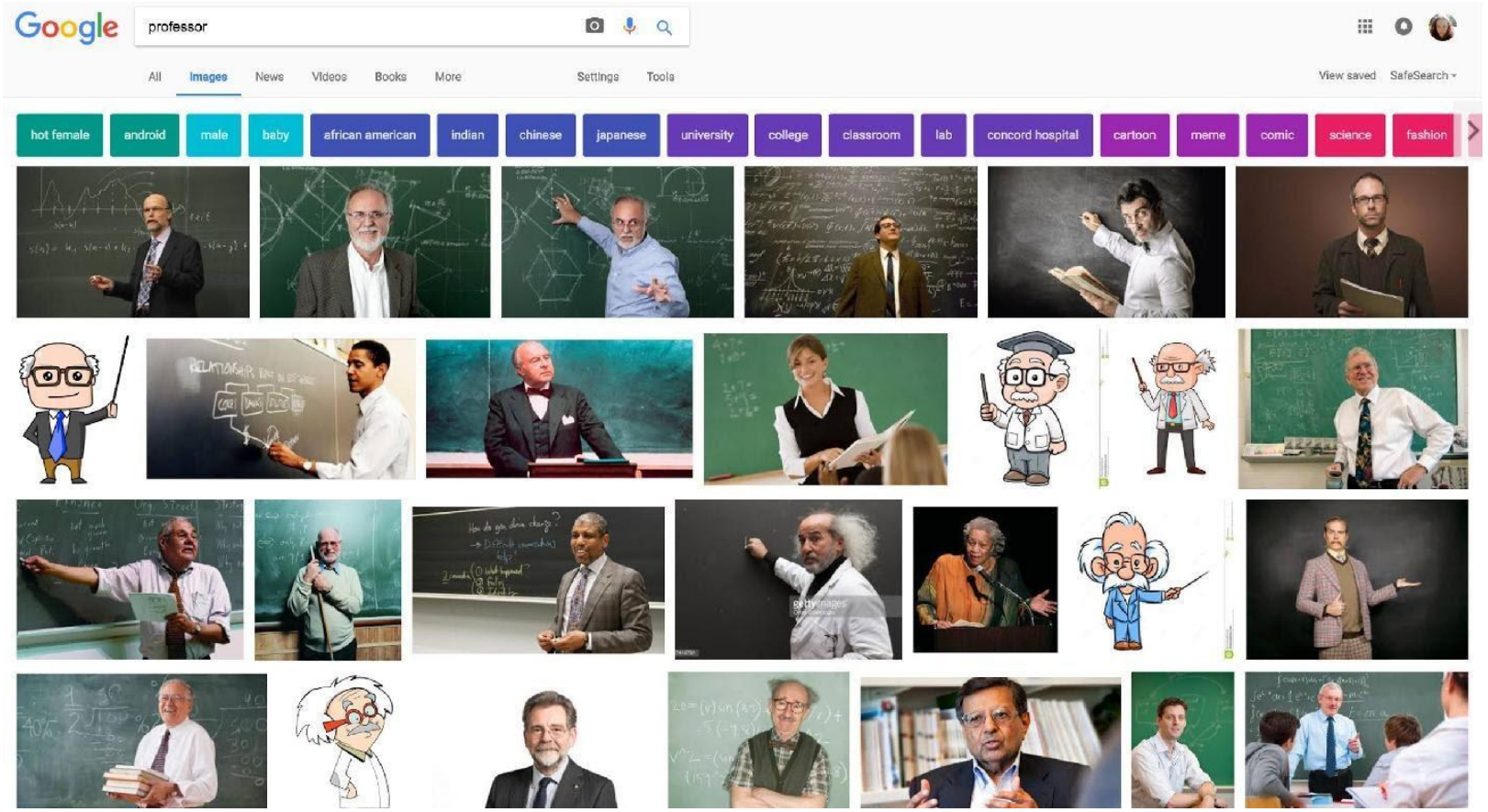
Language Technologies Institute

Overview

- **Learning goals**
  - Get familiar with some example biases in NLP systems and why they occur
  - Learn about some debiasing techniques, and why debiasing may never work
  - Learn about perspectivism and annotator disagreement
  - Discuss socio-technical aspects of biases
- **Content**
  - Biases in lang ID, machine translation, visual sematic roles, hate speech detection
  - Mathematical definitions and origins of bias and fairness
  - Task definitions, perspectivism, inputs and outputs
  - Sociotechnical components of bias

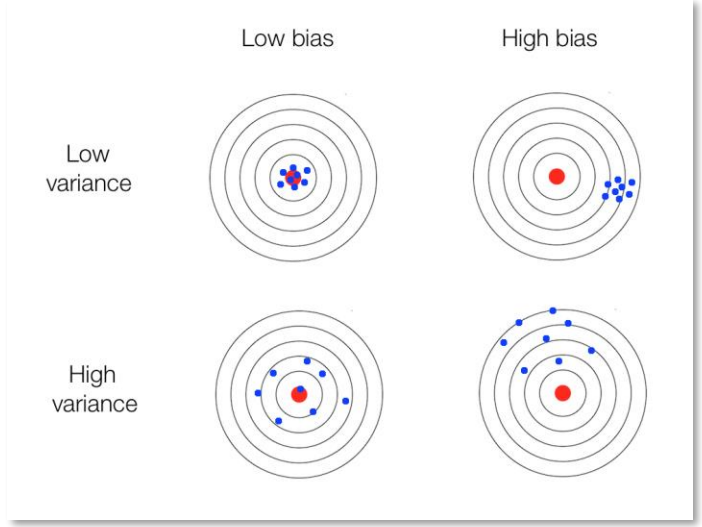
What's wrong with this Google search result?

- June 2017: image search query "Professor"



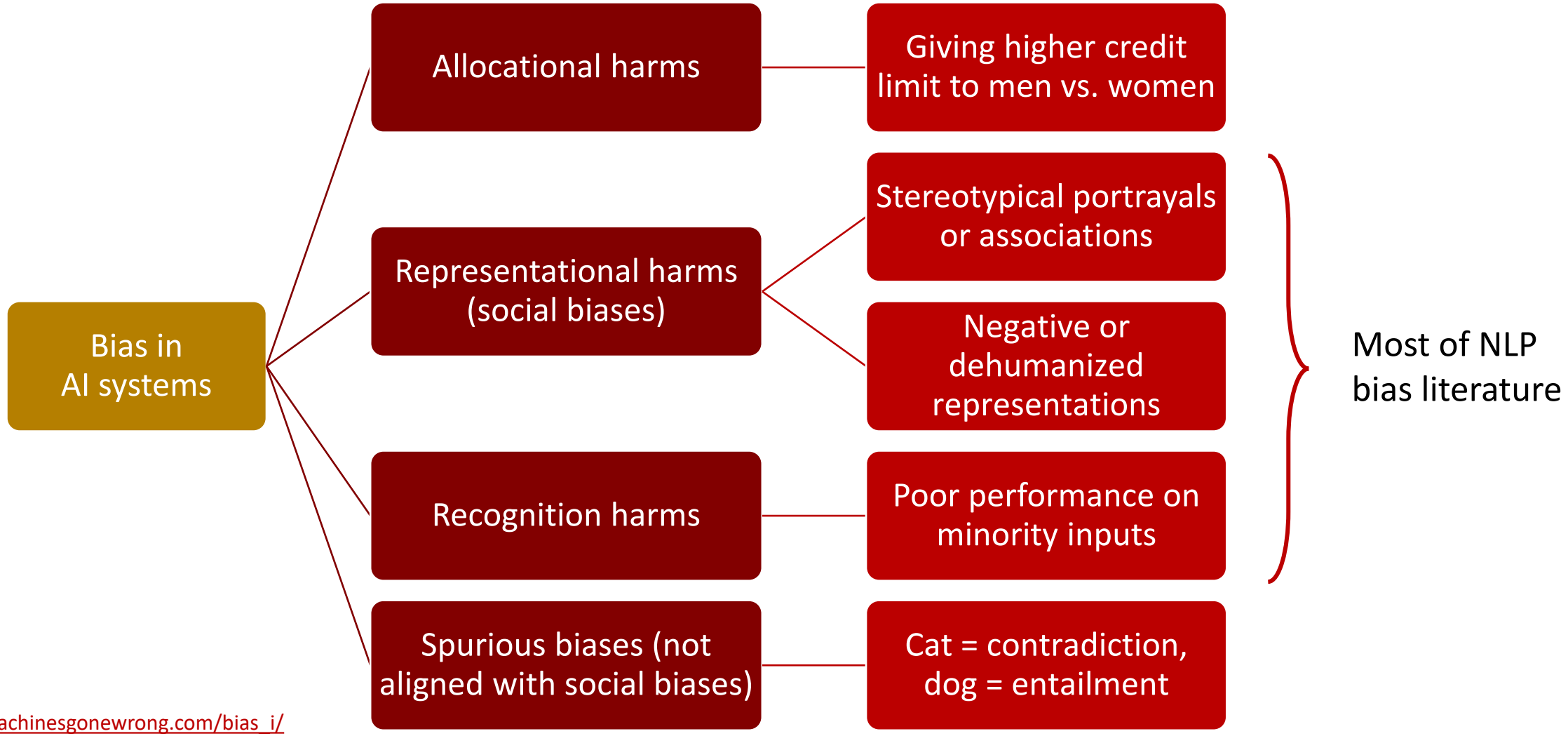
Some definitions of bias

- Bias [*statistics*]: systematic tendency causing differences between model estimates / predictions and the true values of data, skew



- Bias [*general*]: "*disproportionate weight in favor of or against an idea or thing, usually in a way that is closed-minded, prejudicial, or unfair*" –Wikipedia

Bias in terms of the harms it causes

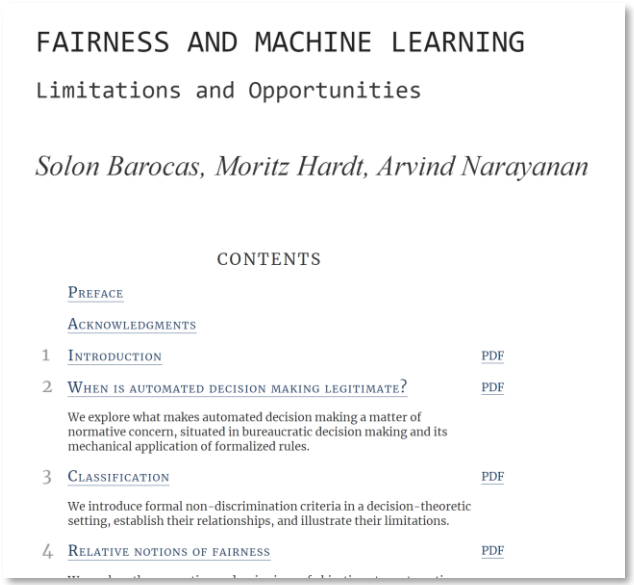


https://machinesgonewrong.com/bias_i/

Algorithmic fairness

- Fair: "*marked by impartiality and honesty: free from self-interest, prejudice, or favoritism*"
  – Merriam-Webster
- Different mathematical definitions of fairness:
  - **Accuracy equality**: groups have same accuracy
  - **Statistical parity**: groups have equal probability of being labeled with positive class
  - **Equalized odds**: *false positive (FP)* and *true positive (TP)* rates must be the same for different groups [Hardt 2016]
  - *Many more…*
- *Note*: some harms or biases are not quantifiable mathematically

		Predicted class: Positive	Predicted class: Negative
Actual class	Positive	TP	FN
	Negative	FP	TN



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

CONTENTS

https://fairmlbook.org/

Let’s look at some example biases in NLP systems, and how we could debias them (if possible)

Bias in language identification

Based on slides by David Jurgens for ACL 2017

Bias in language identification

- Most applications employ off-the-shelf, highly accurate LangID systems.

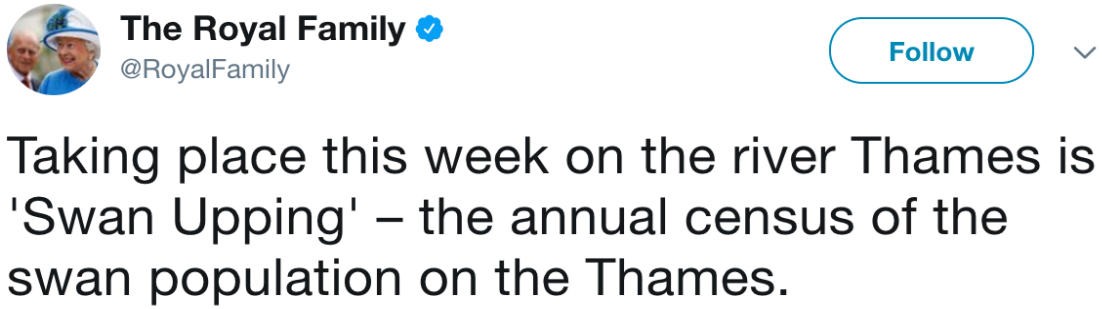
A solved problem?

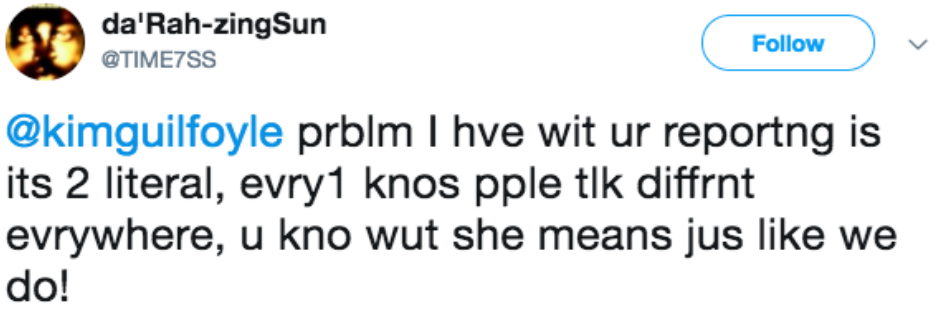
"This paper describes […] how even the most simple of these methods using data obtained from the World Wide Web achieve accuracy approaching 100% on a test suite comprised of ten European languages"

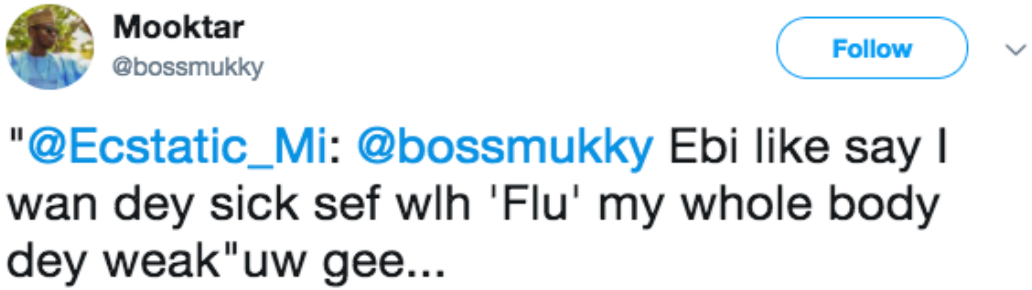
- McNamee, P., "Language identification: ***a solved problem*** suitable for undergraduate instruction" Journal of Computing Sciences in Colleges 20(3) 2005.

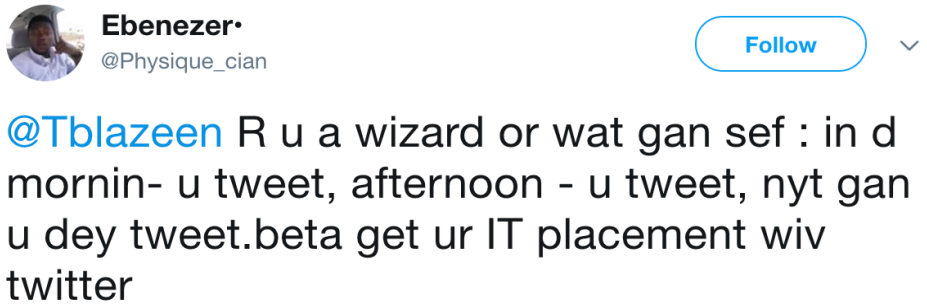
World Englishes

World Englishes



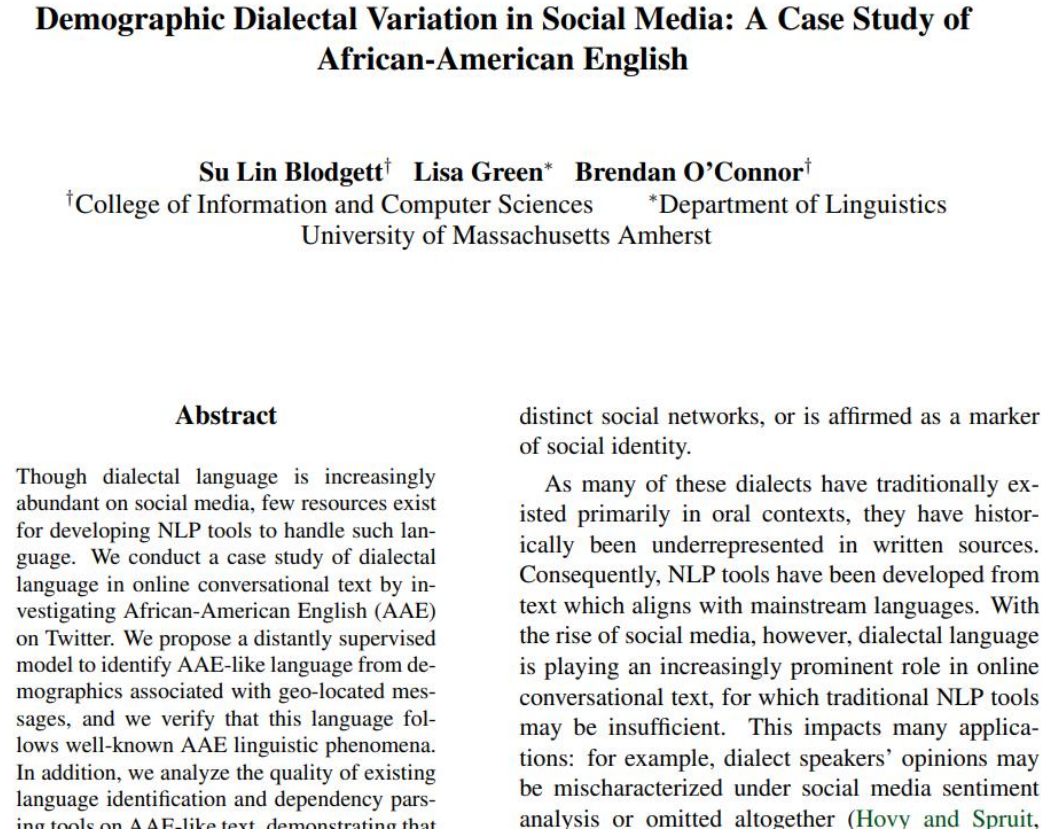






English variation within the U.S.

- Blodgett et al. (2016) investigated bias against African American English (AAE) by langID tools
- AAE tweets were mistaken for being non-English 10-15% of the time



Demographic Dialectal Variation in Social Media: A Case Study of African-American English

Su Lin Blodgett[†] **Lisa Green**[*] **Brendan O'Connor**[†]
[†]College of Information and Computer Sciences [*]Department of Linguistics
University of Massachusetts Amherst

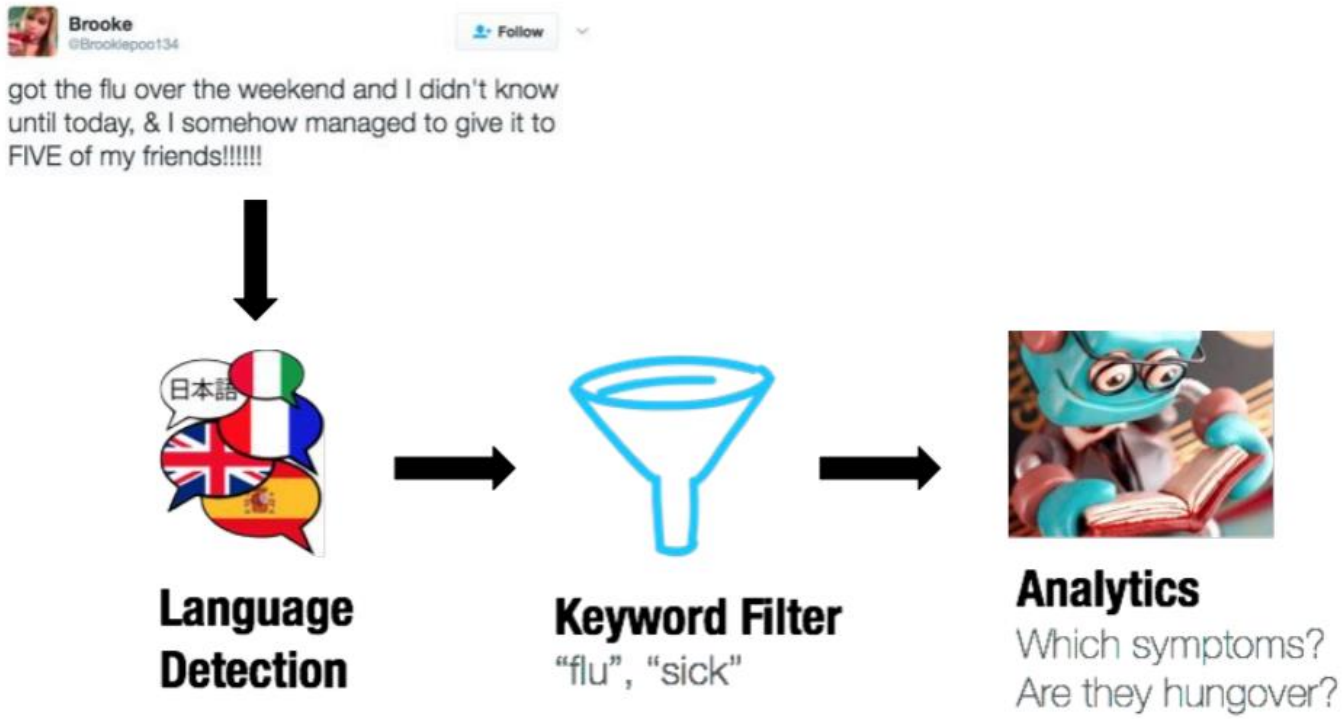
Abstract

Though dialectal language is increasingly abundant on social media, few resources exist for developing NLP tools to handle such language. We conduct a case study of dialectal language in online conversational text by investigating African-American English (AAE) on Twitter. We propose a distantly supervised model to identify AAE-like language from demographics associated with geo-located messages, and we verify that this language follows well-known AAE linguistic phenomena. In addition, we analyze the quality of existing language identification and dependency parsing tools on AAE-like text, demonstrating that

distinct social networks, or is affirmed as a marker of social identity.

As many of these dialects have traditionally existed primarily in oral contexts, they have historically been underrepresented in written sources. Consequently, NLP tools have been developed from text which aligns with mainstream languages. With the rise of social media, however, dialectal language is playing an increasingly prominent role in online conversational text, for which traditional NLP tools may be insufficient. This impacts many applications: for example, dialect speakers' opinions may be mischaracterized under social media sentiment analysis or omitted altogether (Hovy and Spruit,

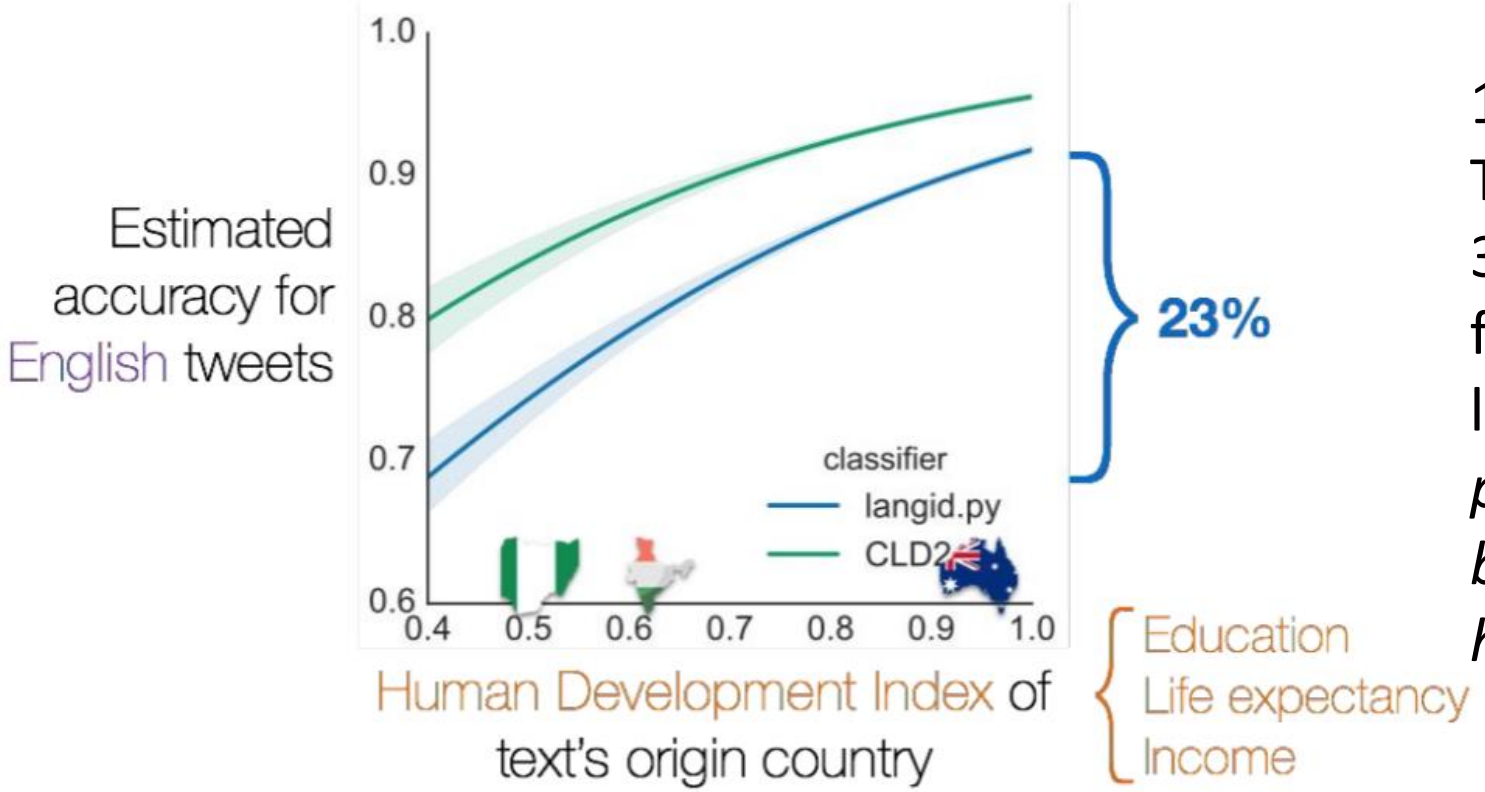
LangID usage example: health monitoring



LangID usage example: health monitoring

Socioeconomic bias in language identification

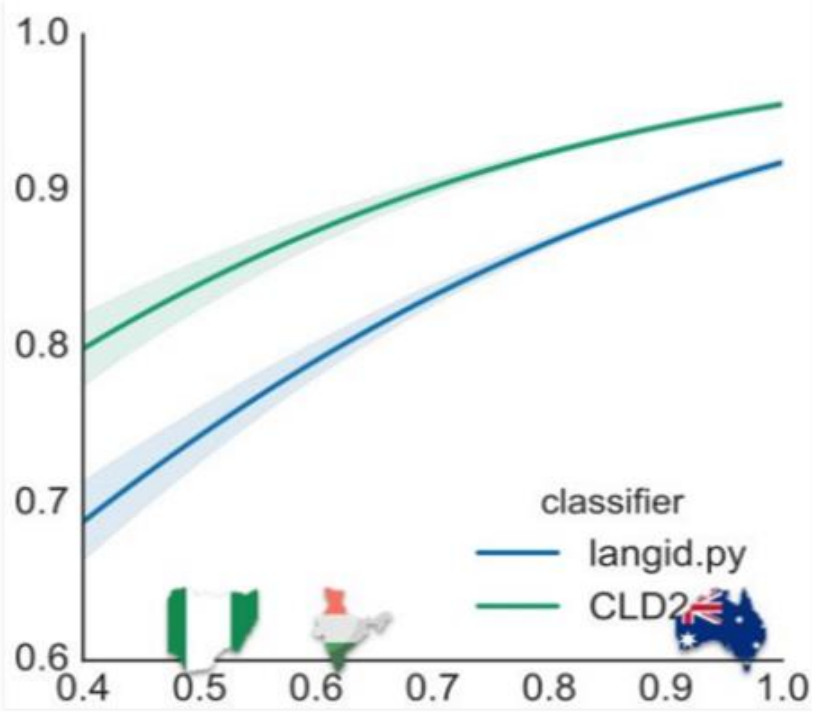
- Off-the-shelf LangID systems under-represent populations in developing countries



1M geo-tagged Tweets with any of 385 English terms from established lexicons for *influenza*, *psychological well-being*, and *social health*

How would you debias the LangID system?

- Propose a solution!



Better social representation through network-based sampling

- Re-sampling from strategically-diverse corpora

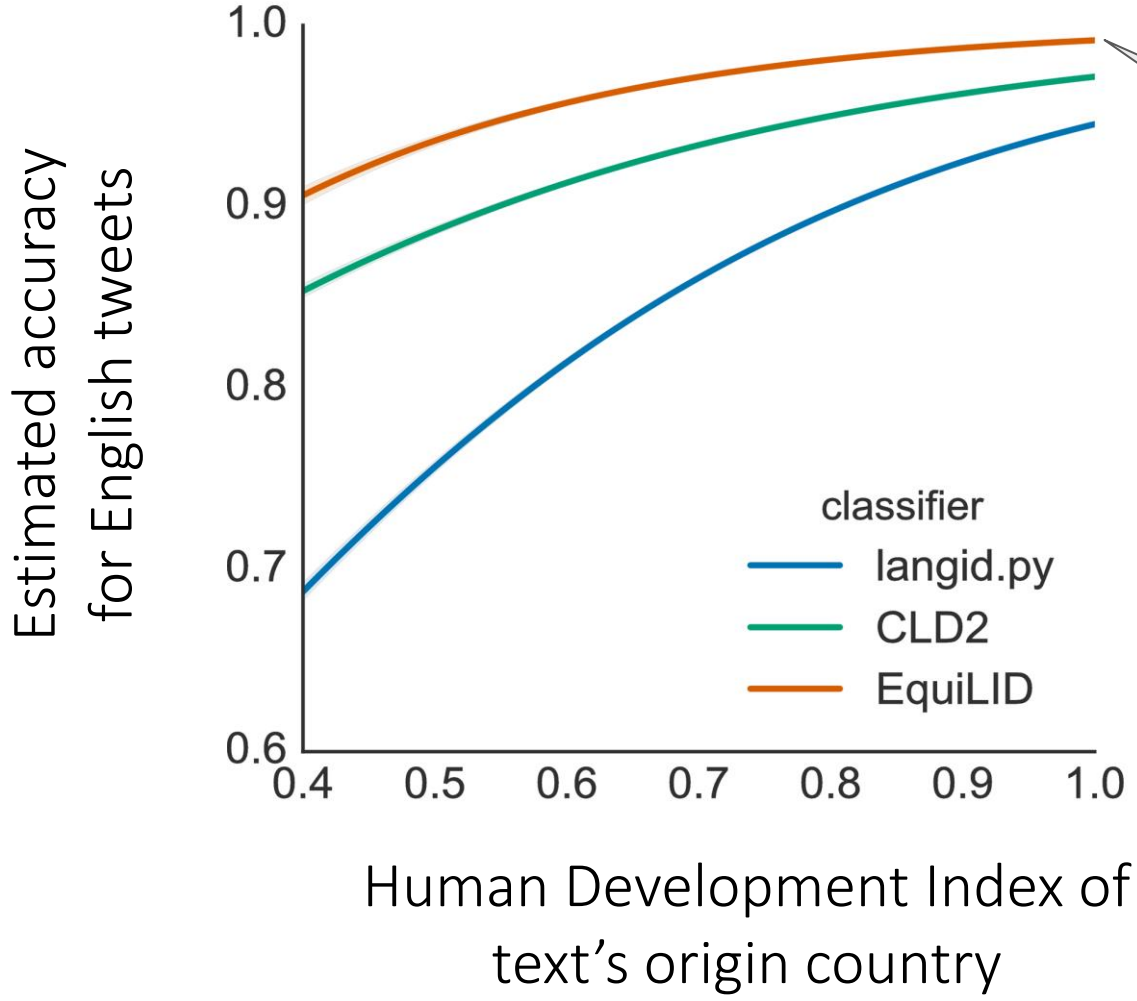
Topical

Social

Geographic

Multilingual

EquiLID: less biased LangID



Debiasing does not come at the expense of model quality. But this is not always the case.

Takeaways from EquiLID

- Where did the bias come from?
- How did they fix the bias?

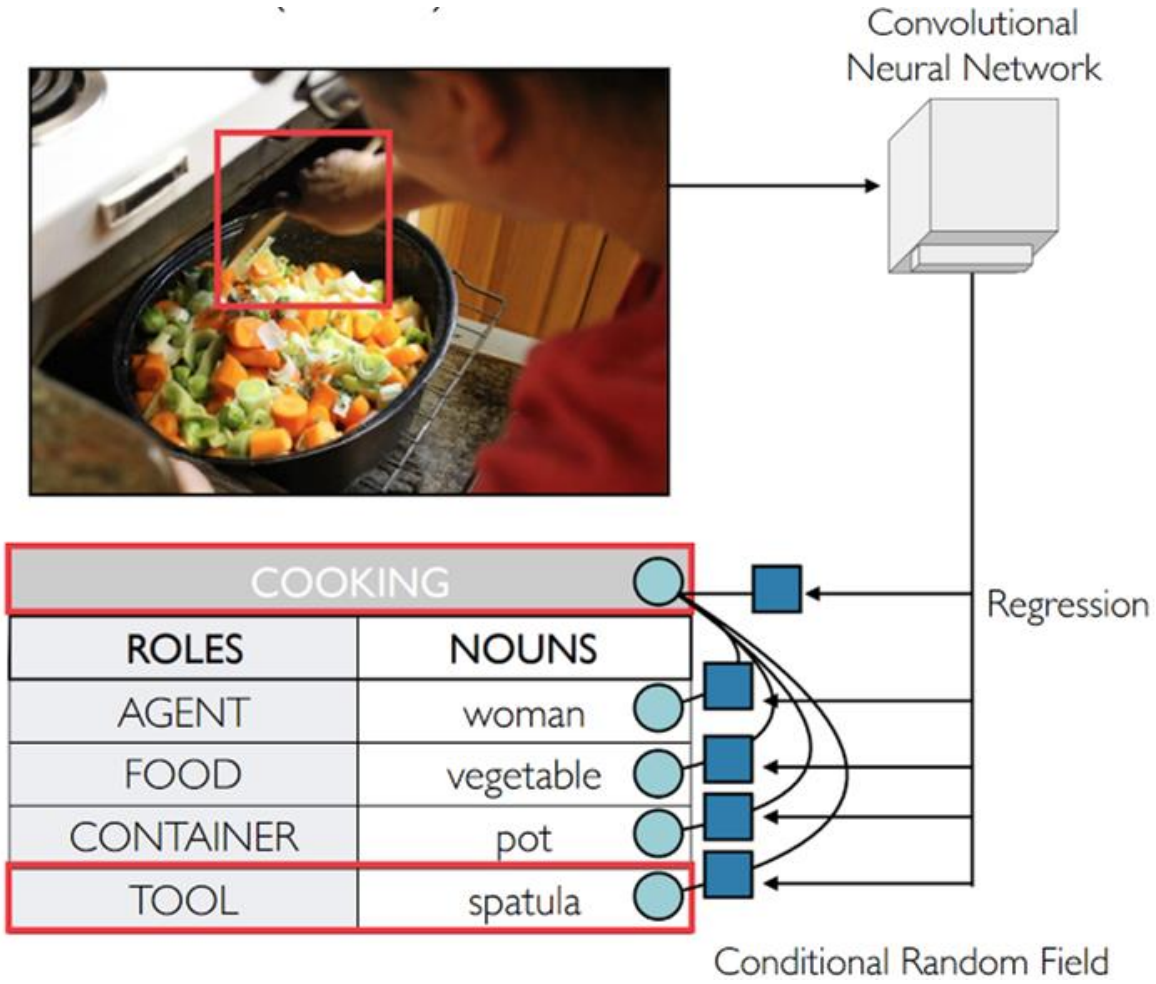
- Answer to both: data!
- Possible conclusion: all you need to do is fix the data, right?
- Well.. Not really…

Bias in visual sematic role labeling

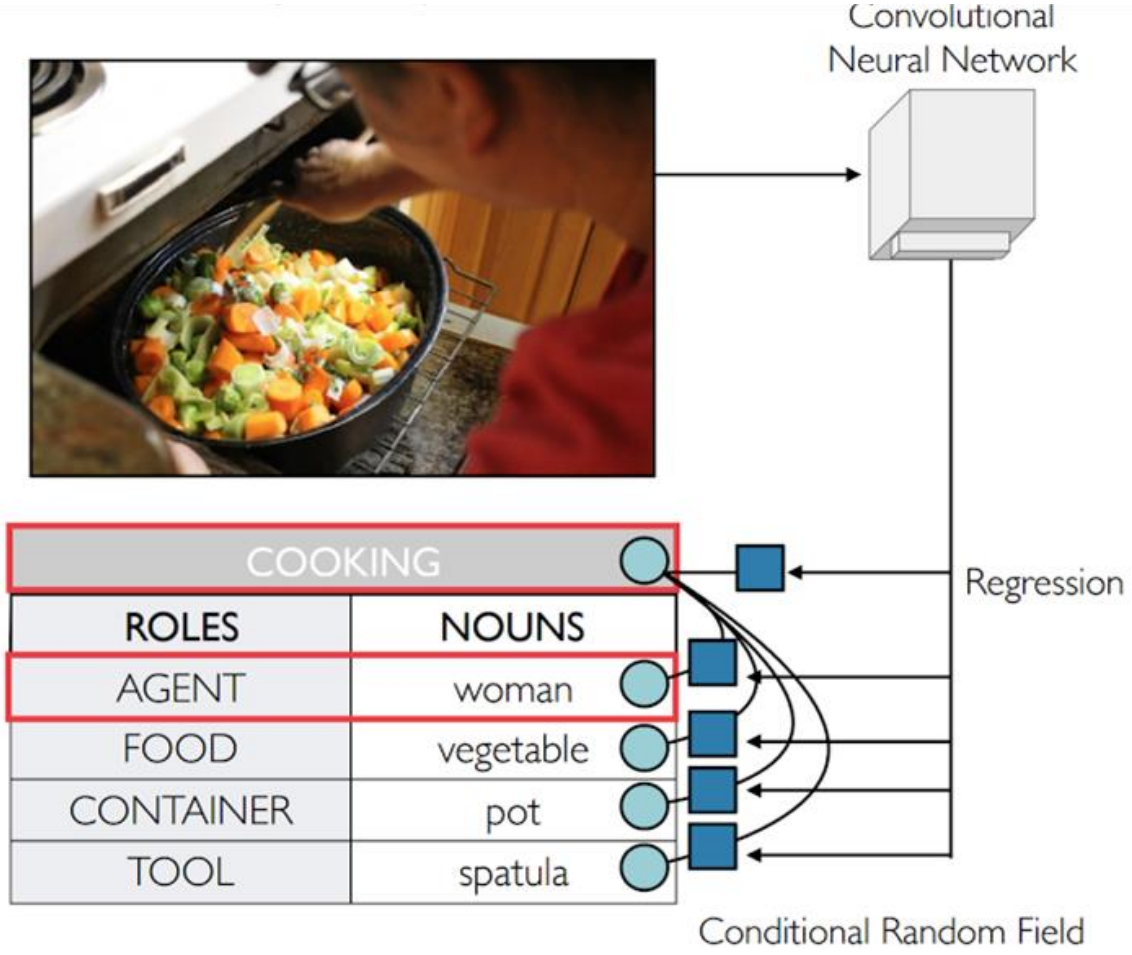
Based on slides by Mark Yatskar (EMNLP 2017)

Carnegie Mellon University
Language Technologies Institute

imSitu Visual Semantic Role Labeling (vSRL)

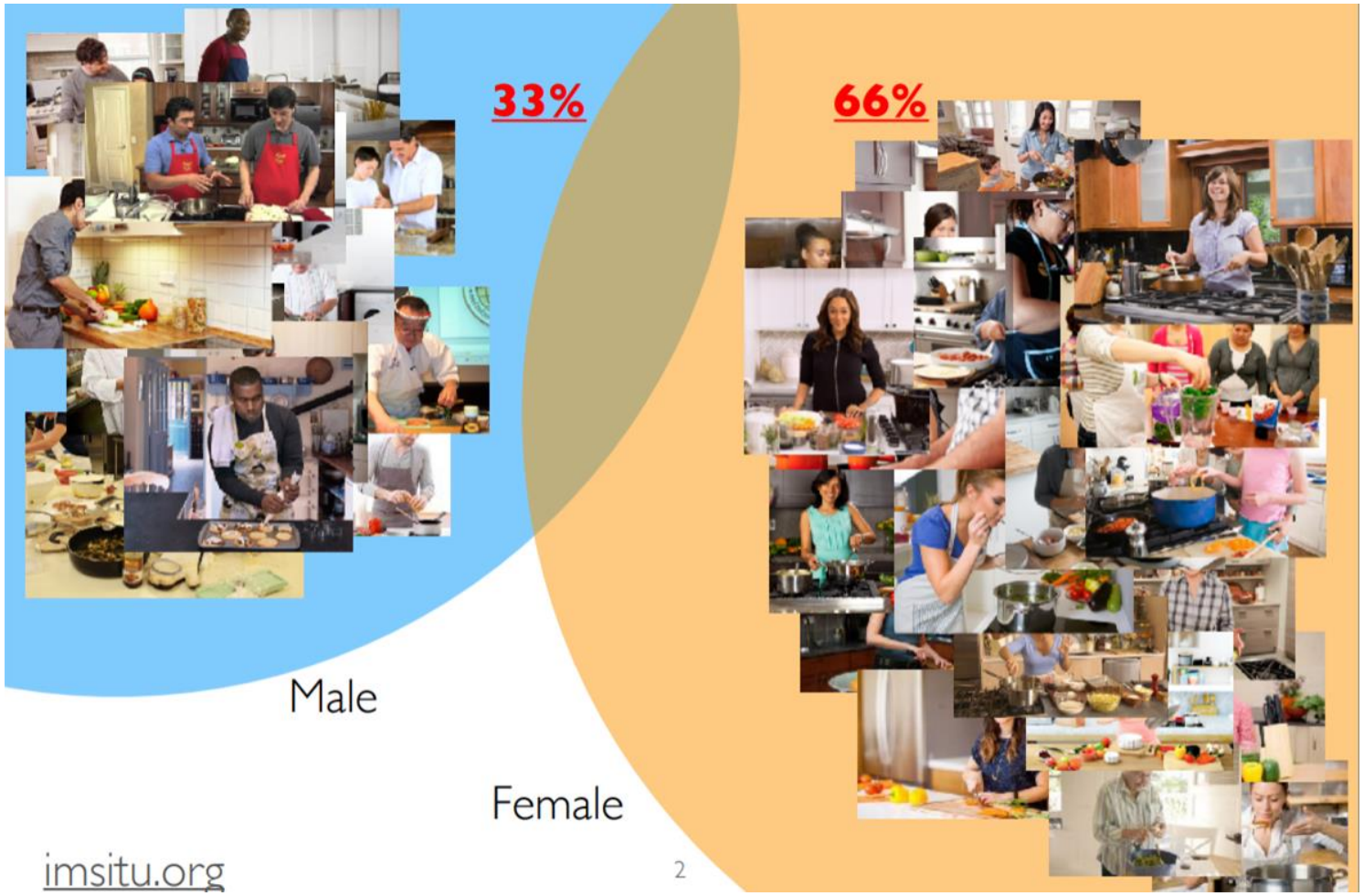


imSitu Visual Semantic Role Labeling (vSRL)



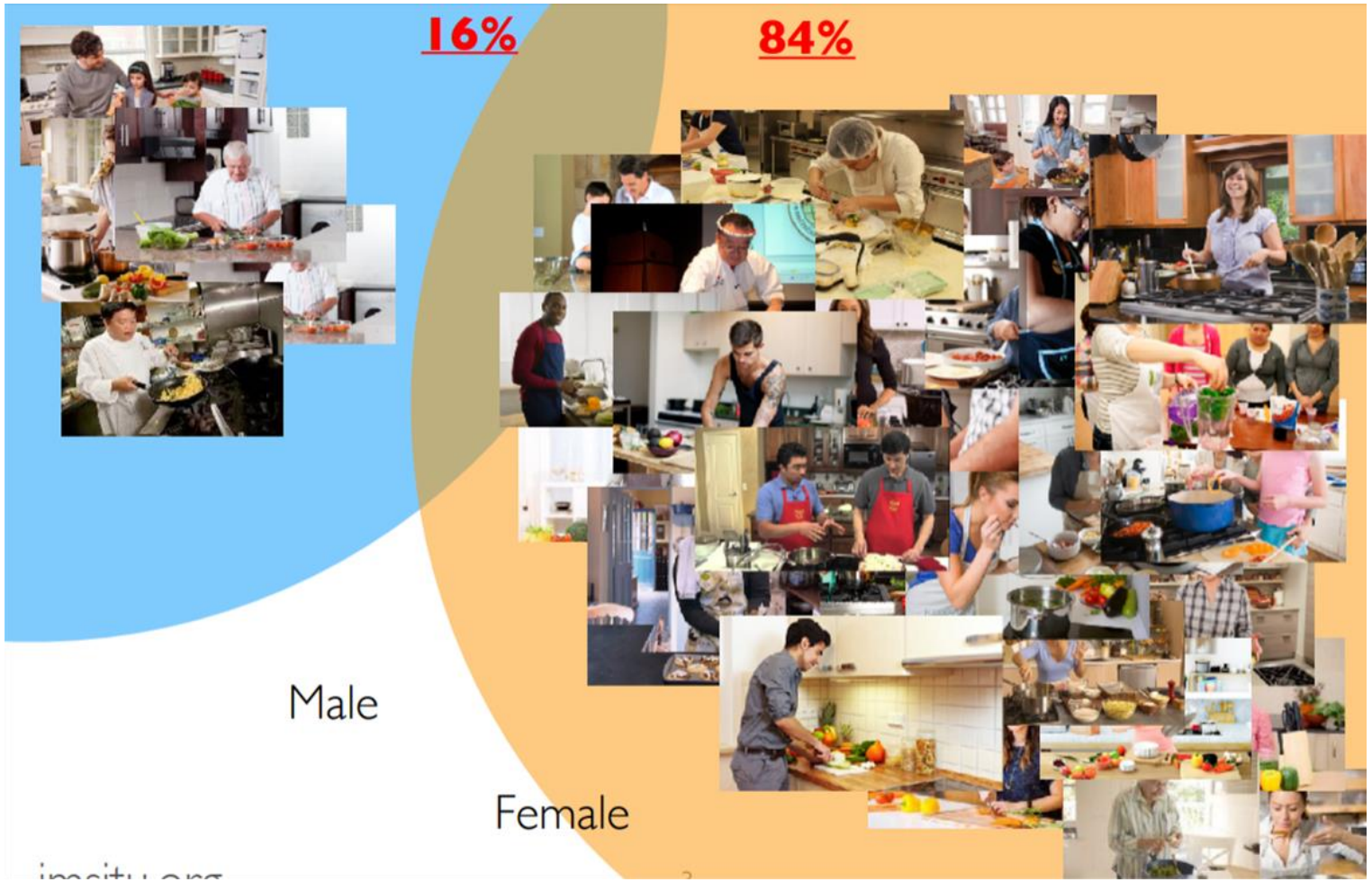
Dataset gender bias

cooking pictures with men or women



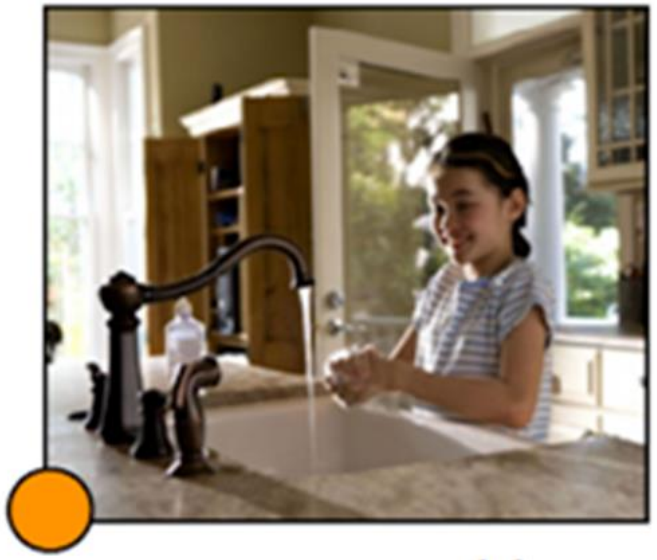
Model bias after training

cooking pictures predicted as men or women

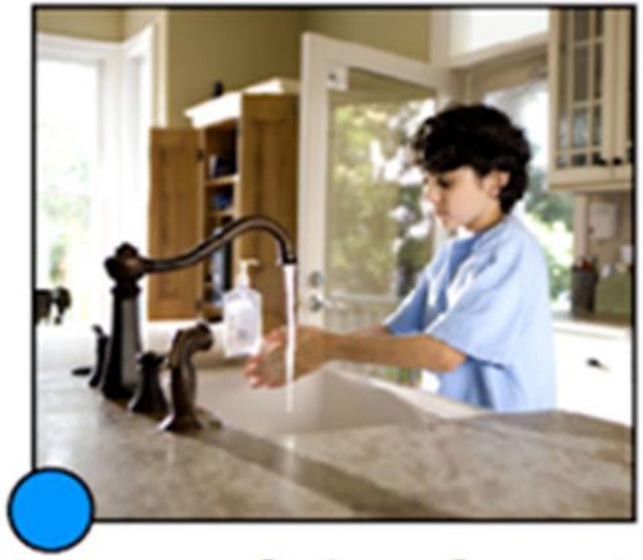


Why does this happen?

Algorithmic bias



woman cooking



man fixing faucet

Quantifying dataset bias

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

Quantifying dataset bias

Training Gender Ratio (◆ verb)

Training Set

◆ cooking

● woman

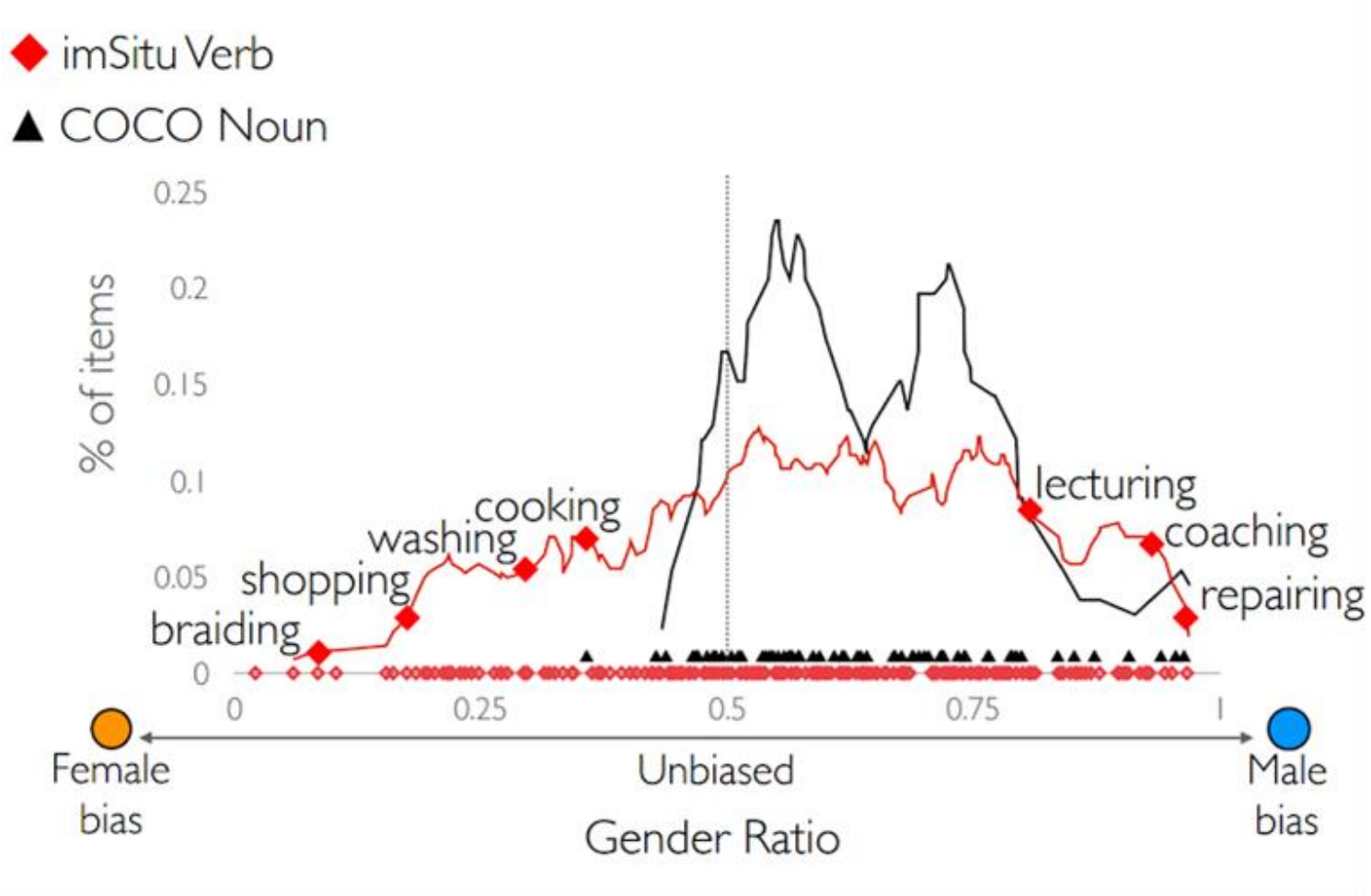
● man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

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

Gender Dataset Bias



Quantifying dataset bias: dev set

Predicted Gender Ratio (verb)

Development Set

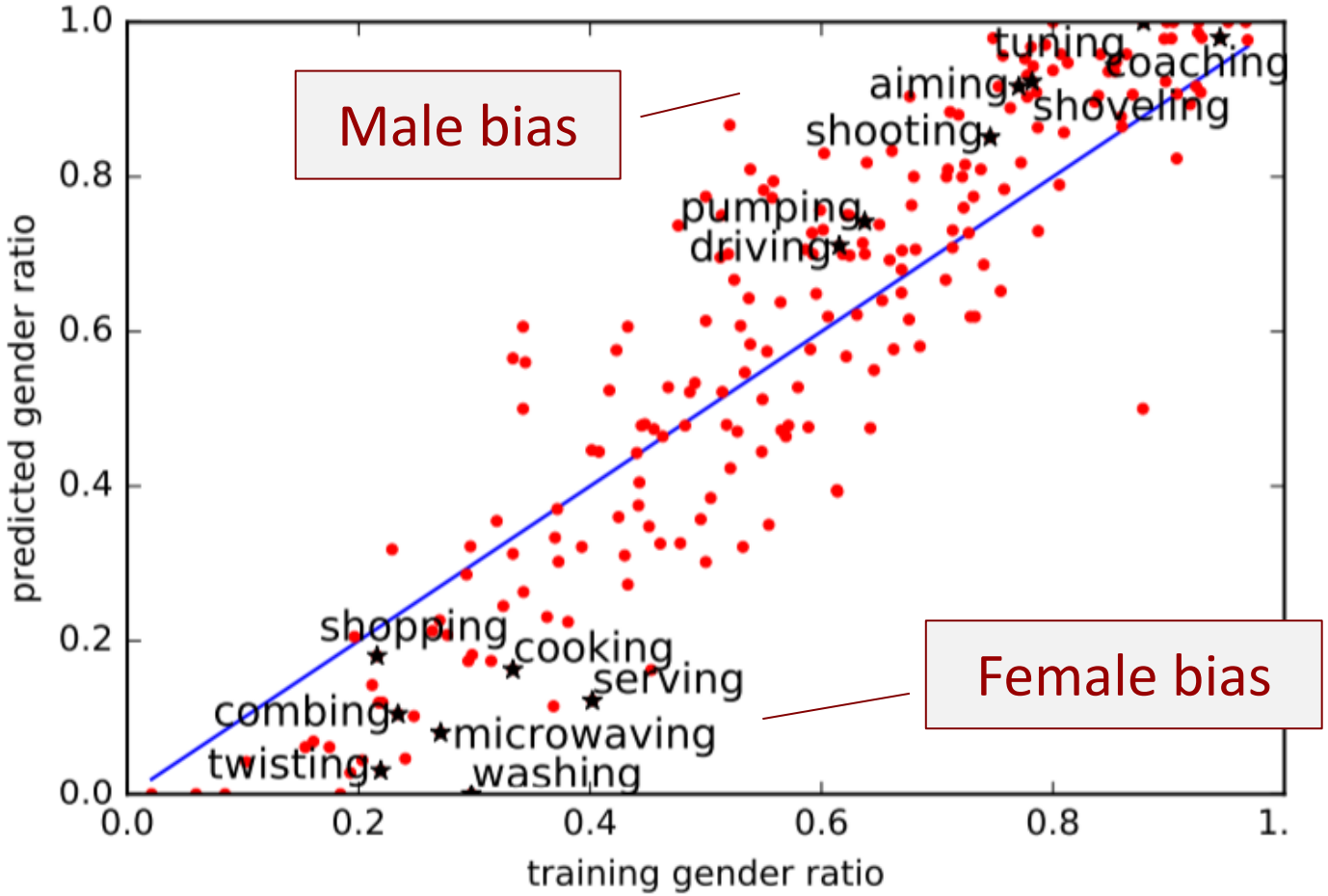
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

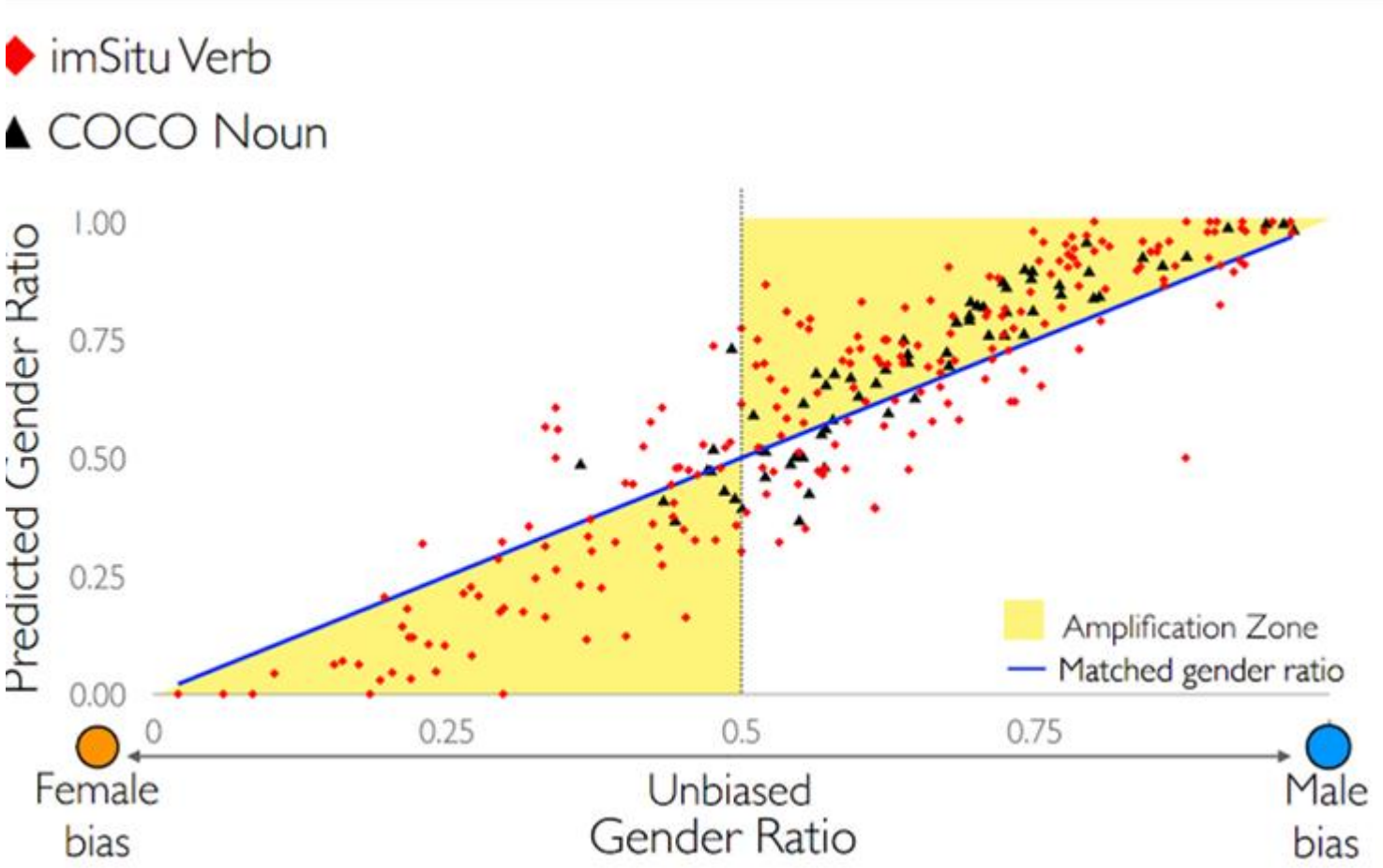
COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

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

Model bias amplification



Model bias amplification



Quantifying Bias Amplification

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

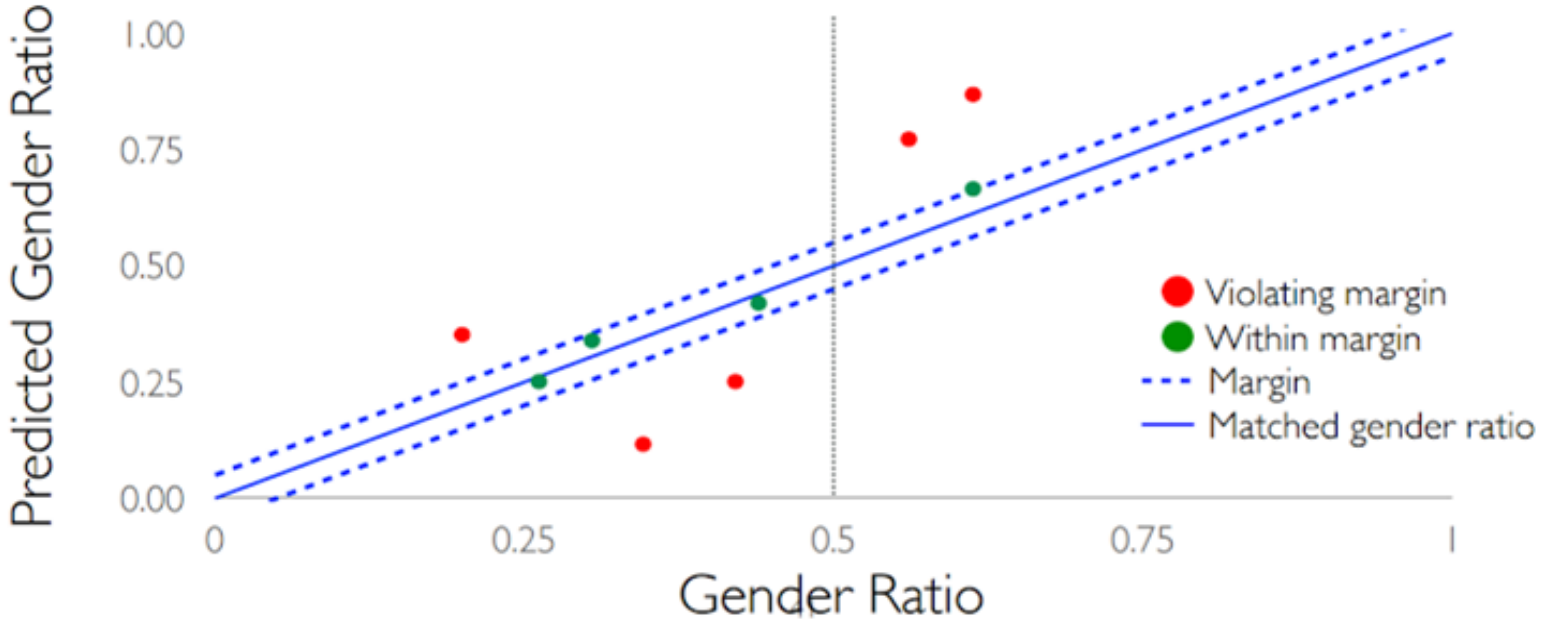
- O - activity
- G - gender
- b*(o, g) - training data bias
- b~(o, g) - model bias

How to “debias” this system?

Reducing Bias Amplification (RBA) via calibration

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

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

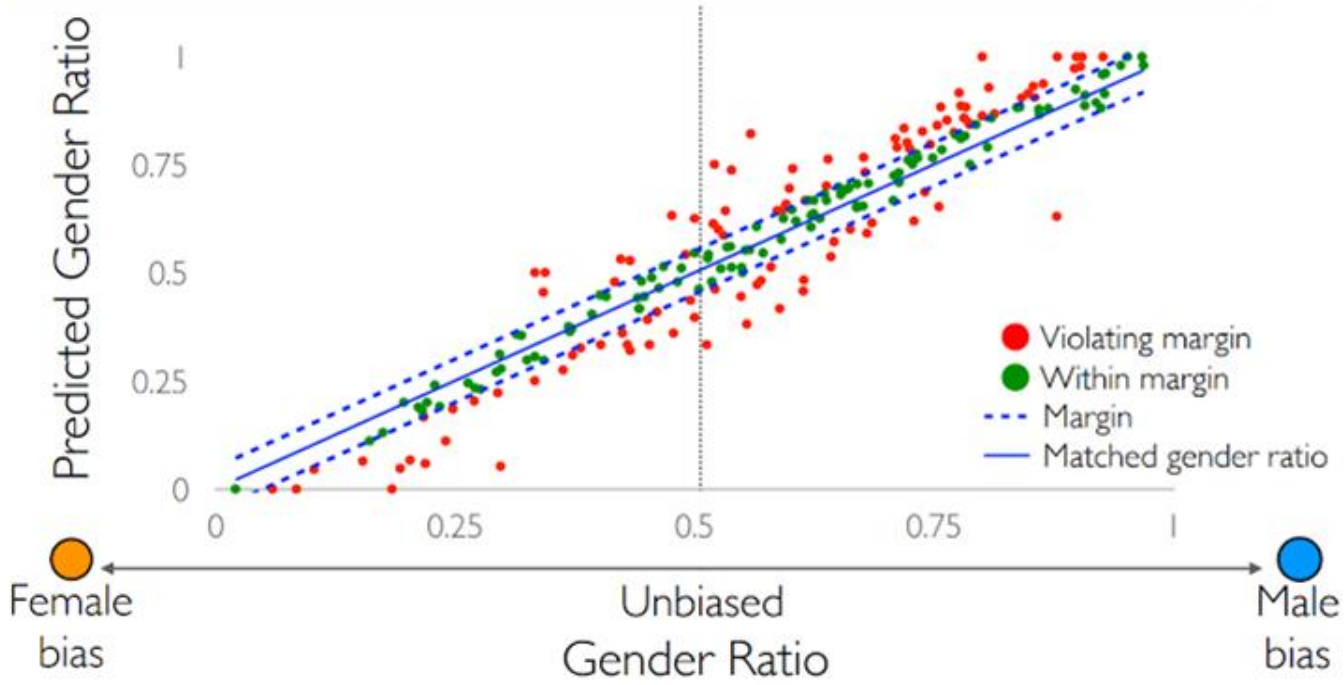


- *Idea*: gender ratio in model predictions for each activity should be very close to the training distribution
- Add a test-corpus level "gender-ratio" constraint on top of instance-level structured prediction constraints

Results

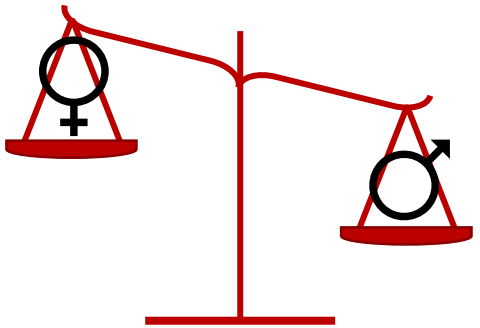
- With RBA: accuracy is ~same and gender bias is halved,
- But there is still gender bias (50% margin “violations”)

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



Bias in visual semantic role labeling - takeaways

- Biases in data are an issue



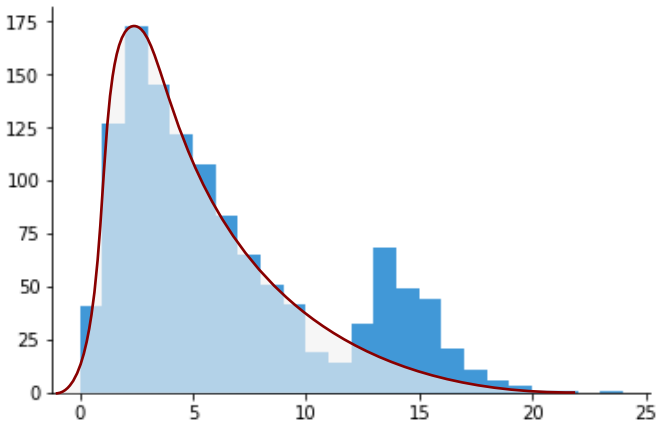
Bias in visual semantic role labeling - takeaways

- Biases in data are an issue
- Models can **amplify** those biases

- Why does this happen, mathematically?

Tangent: mathematical links of bias

- *Competing losses*: objective functions aim to minimize loss globally, model learns to predict most frequent class
  - Often at the expense of less frequent classes (e.g. minority groups)
- *Simplicity bias*: neural networks biased towards learning simpler functions [Valle Pérez et al. 2019]
- Intuitively, if a model has limited learning capacity, makes sense that it learns shortcuts first
  - Shortcuts are often stereotypes or majority biases; e.g., CEOs are men



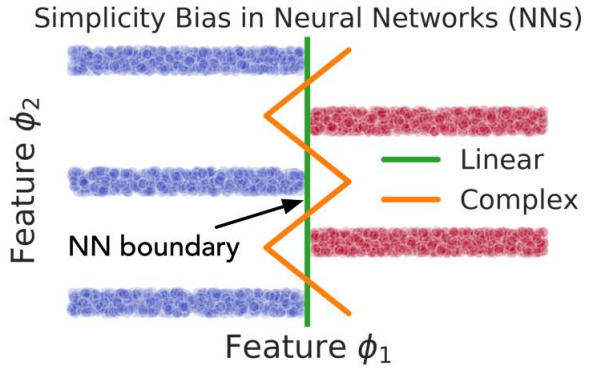
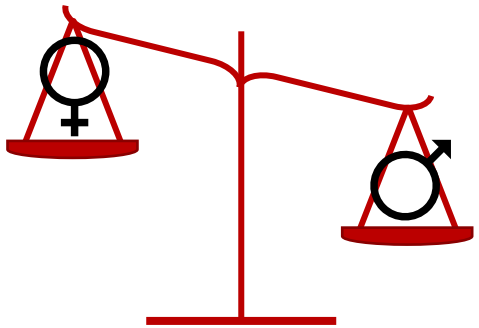


Figure 1: Simple vs. complex features

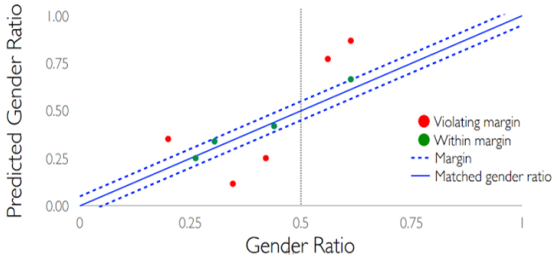
Figure from Shah et al 2020

Bias in visual semantic role labeling - takeaways

- Biases in data are an issue
- Models can **amplify** those biases



- Calibrating models & inference algorithm constraints can help
- But models are still biased
- And is the training distribution really the target for learning?



DALLE-2 Biases

- DALLE-2 was shown to have biases
  - For prompt “CEO” it would only generate white men
  - For “teacher” it would only generate white women
  - Generation mirrored training data, arguably “real world”
- OpenAI’s “fix”: append demographic words to prompt (e.g., Black, Asian, woman, etc.)
- *Let’s discuss*: what do you think of this approach?
  - Is the training data or real world distribution the target distribution?

NewScientist

AI art tool DALL-E 2 adds 'black' or 'female' to some image prompts

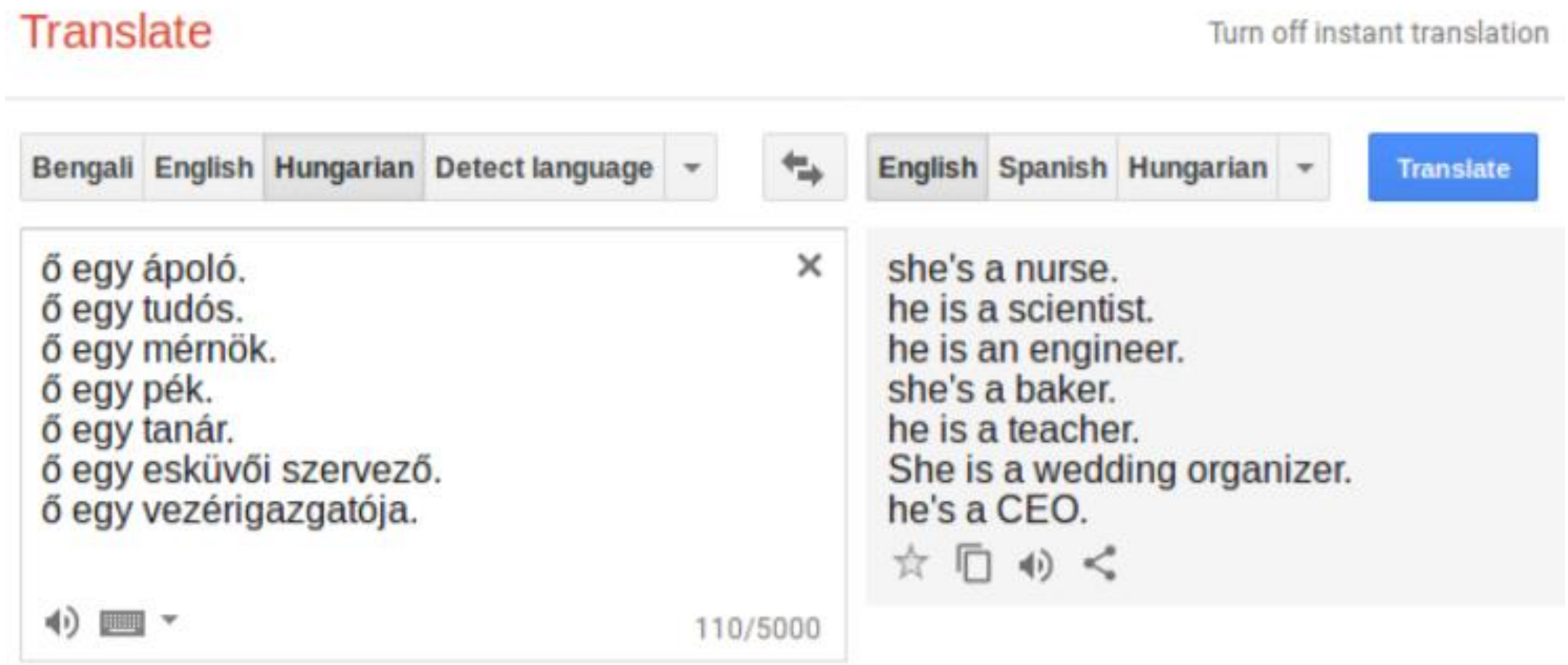
Researchers experimenting with OpenAI's text-to-image tool, DALL-E 2, noticed that it seems to covertly be adding words such as "black" and "female" to image prompts, seemingly in an effort to diversify its output

TECHNOLOGY 22 July 2022

By Matthew Sparkes

Bias in Machine Translation (MT)

Bias in Machine Translation [Prates et al. 2019]



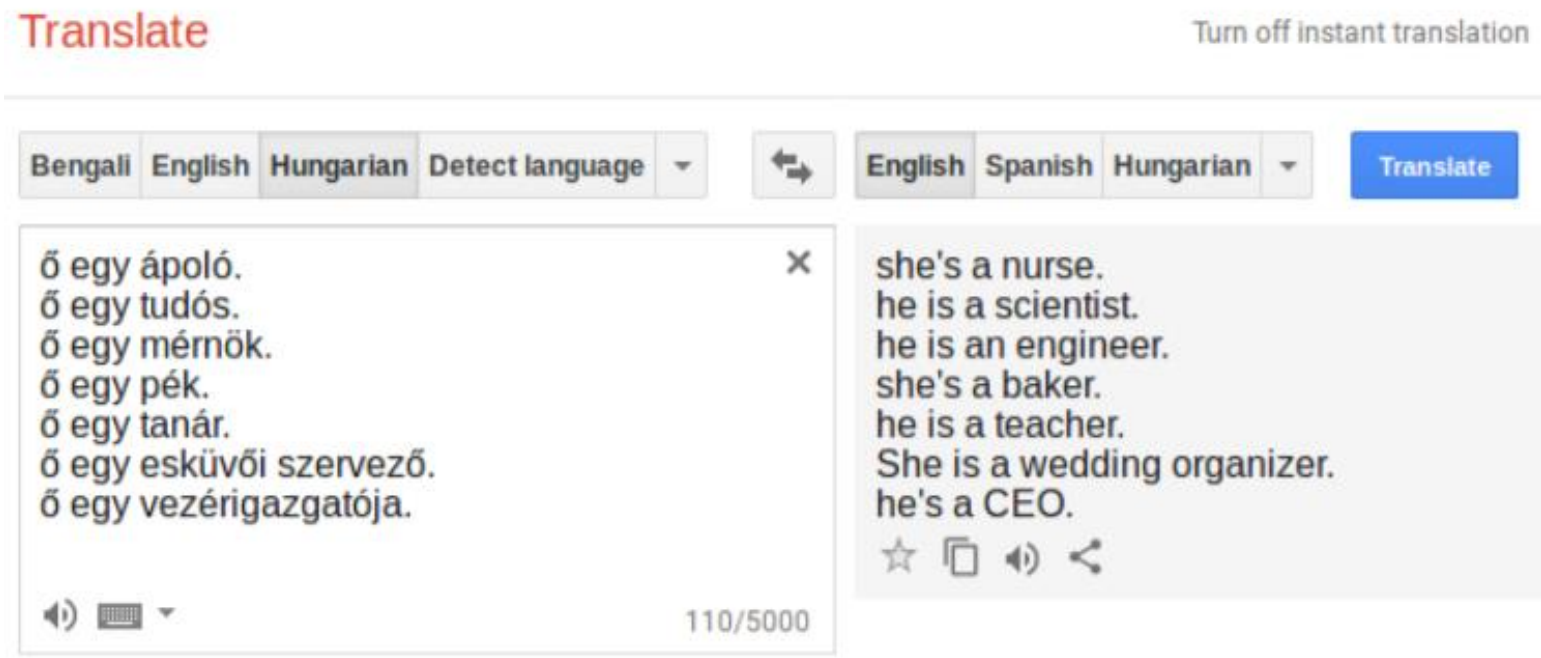
THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE

Language Family	Language	Phrases have male/female markers	Tested
Austronesian	Malay	✗	✓
Uralic	Estonian	✗	✓
	Finnish	✗	✓
	Hungarian	✗	✓
Indo-European	Armenian	✗	✓
	Bengali	O	✓
	English	✓	✗
	Persian	✗	✓
	Nepali	O	✓
Japonic	Japanese	✗	✓
Koreanic	Korean	✓	✗
Turkic	Turkish	✗	✓
Niger-Congo	Yoruba	✗	✓
	Swahili	✗	✓
Isolate	Basque	✗	✓
Sino-Tibetan	Chinese	O	✓

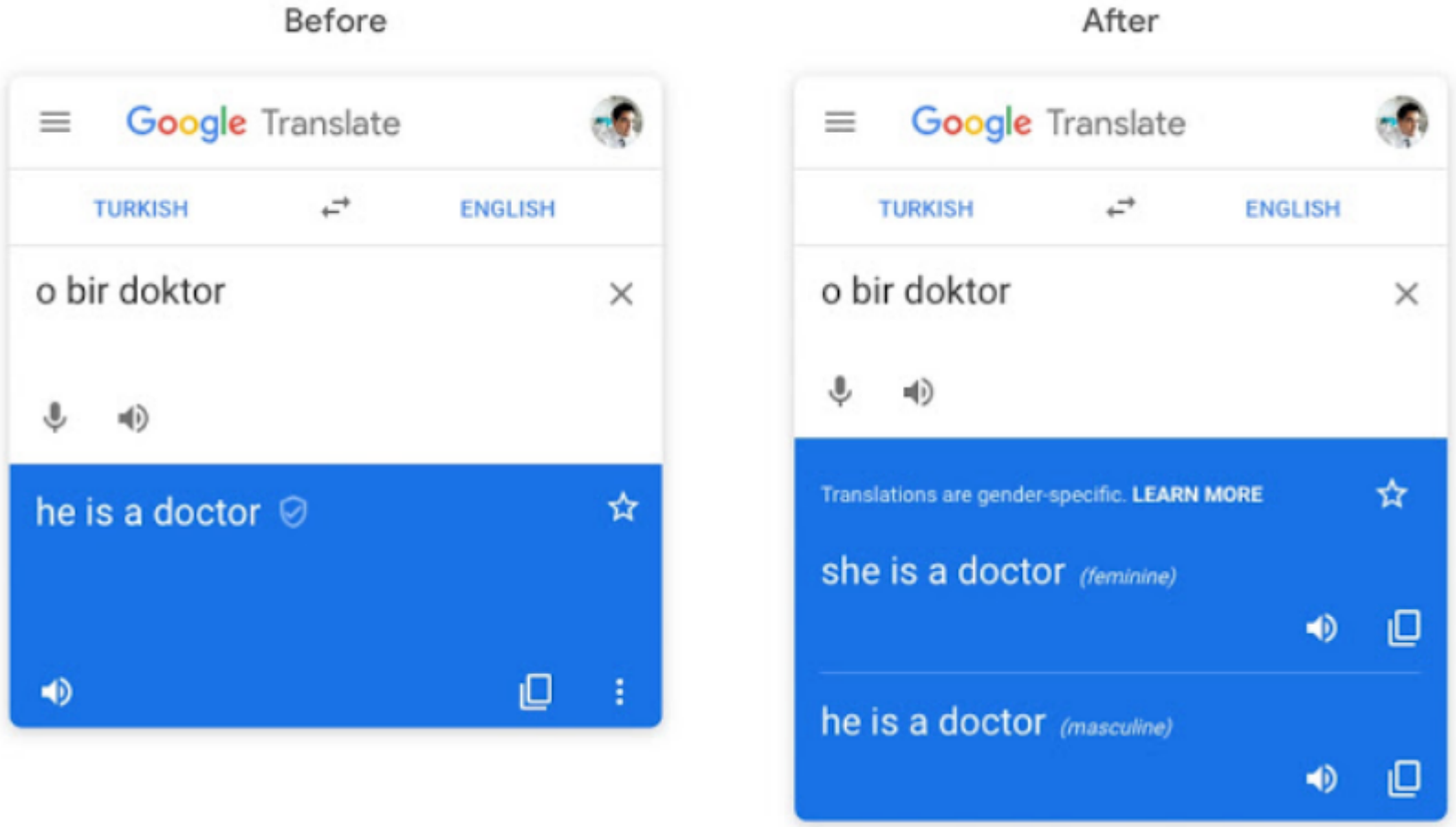
[Prates et al. 2019]

How would you debias the MT system?

- Propose a solution!

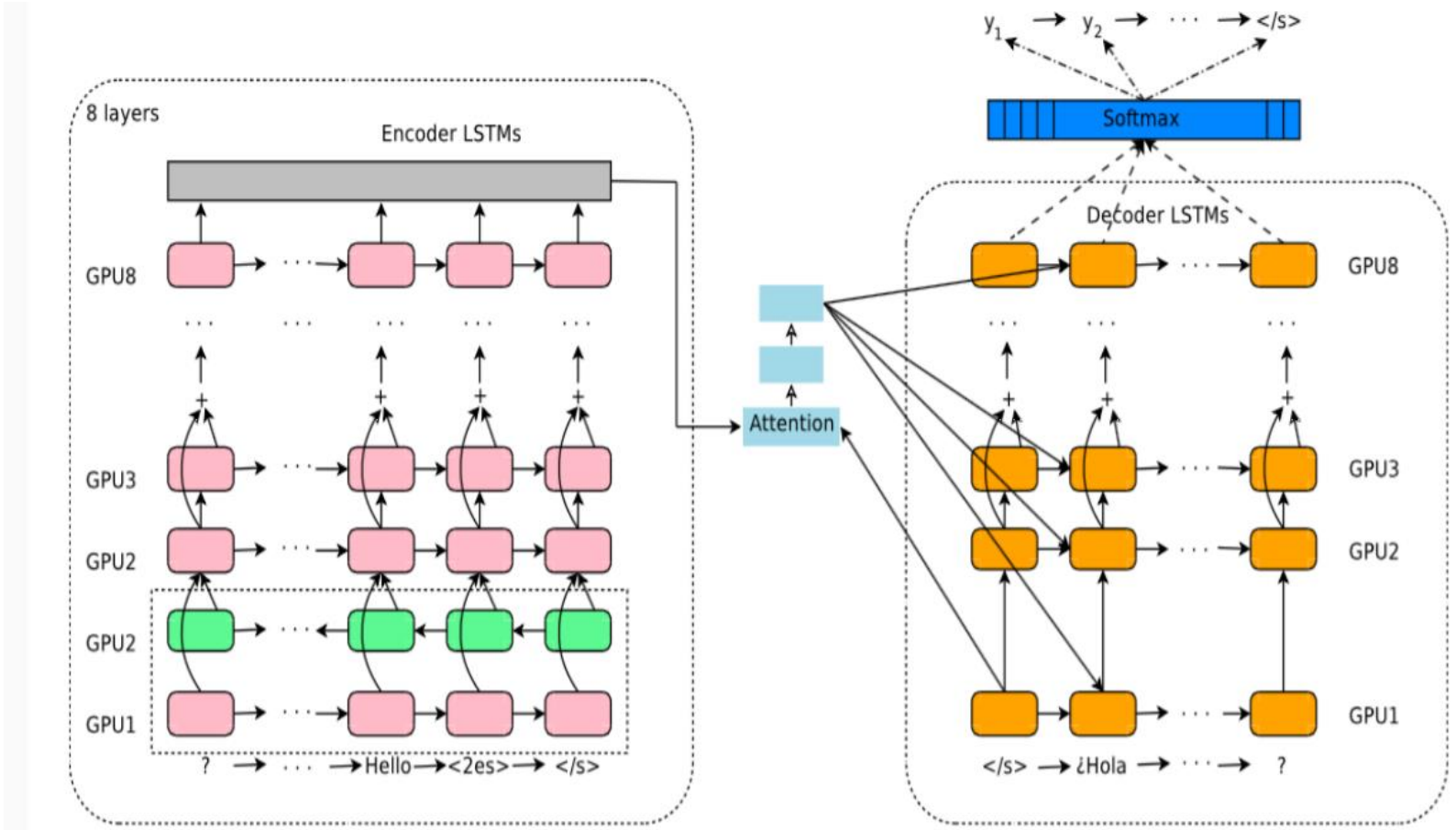


Google's approach: UI & algorithm change



https://ai.googleblog.com/2018/12/providing-gender-specific-translations.html

Debiasing similar to multilingual NMT



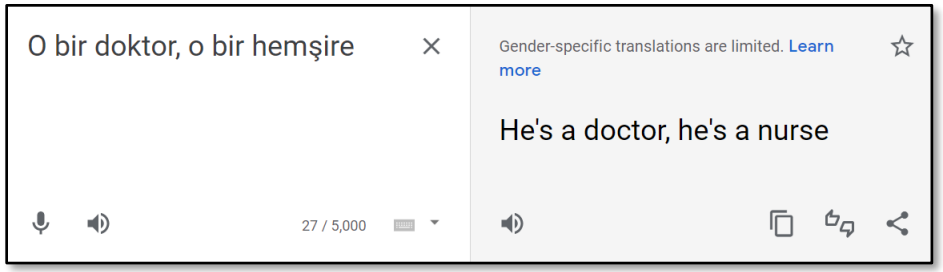
<2female> <2es> Hello, how are you? -> ¿Hola como estás?

Debiasing MT takeaways

- Machine translation incorrectly assumed 1-1 translation pairs
- Languages without grammatical gender translated to male grammatical gender often
  - And “chosen” gender often aligned with stereotypes
- Google’s solutions:
  - Make the model “gender-aware”
  - Change the UI to produce two different translations
- Are there still issues? *Let’s discuss*…

Debiasing MT – it gets more complicated

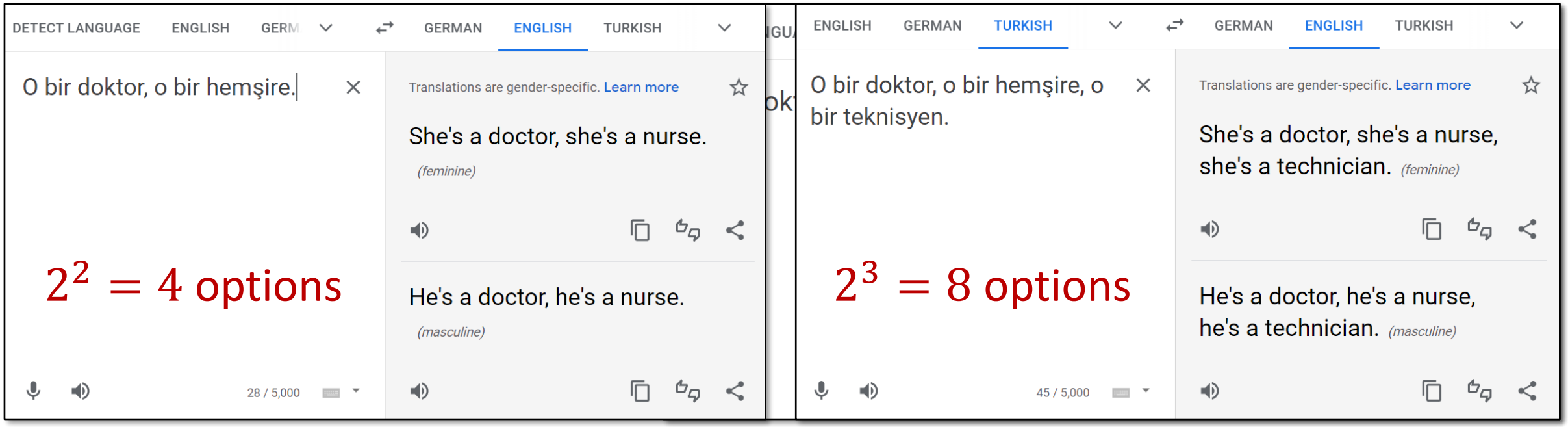
- Google's simple fix doesn't work for everything
  - Doesn't always know when to use gender-specific outputs
  - Longer documents are a challenge, often due to cross-sentence gender-resolution being needed
- They released a dataset to address the gender bias issues [Stella 06/2021]
- Even the newer models aren't perfect, try it yourself [link]



Marie Curie was born in Warsaw. The respected scientist received the Nobel Prize in 1903 and 1911.

Marie Curie wurde in Warschau geboren. ~~Der~~**Die** angesehene Wissenschaftler**in** erhielt 1903 und 1911 den Nobelpreis.

Debiasing MT – it gets even more complicated

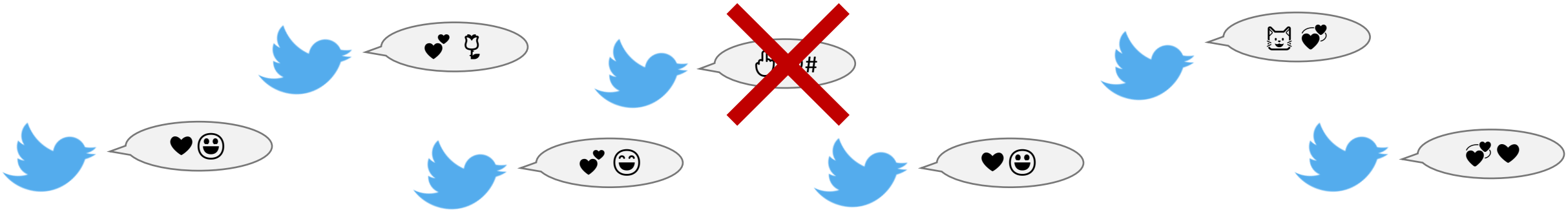


- UI issue: Combinatorial explosion problem of displaying all possible options
- *Let's discuss*: what would you do?
- Non-binary and diverse genders: what about singular "they"? Neo-pronouns?

Biases in hate speech detection

⚠ 🚨 CONTENT WARNING 🚨 ⚠

Hate Speech or Toxic Language Detection

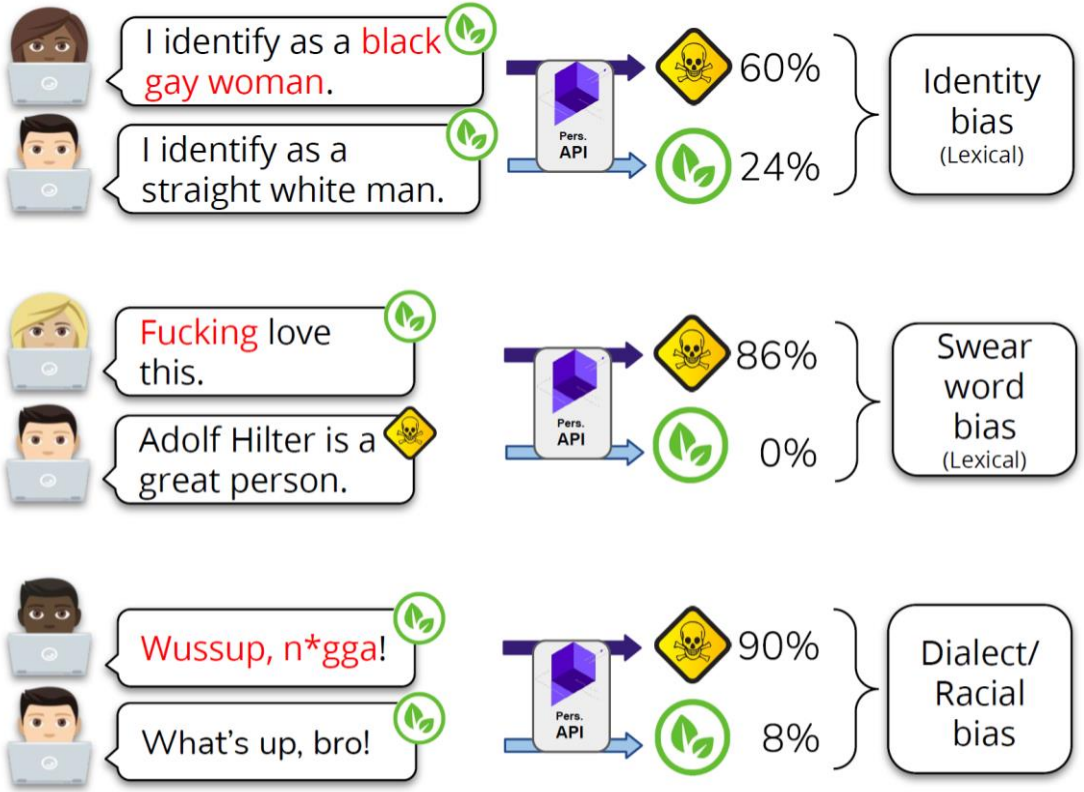


Goal: find and flag hateful or toxic content online, to make the internet less toxic

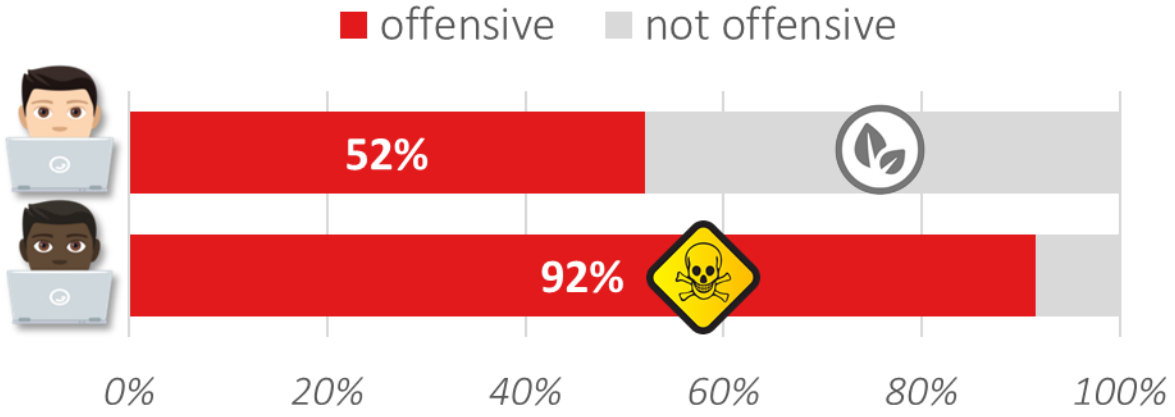
(more on this in "Civility in communication" module)

Problem: biases in toxic language detection

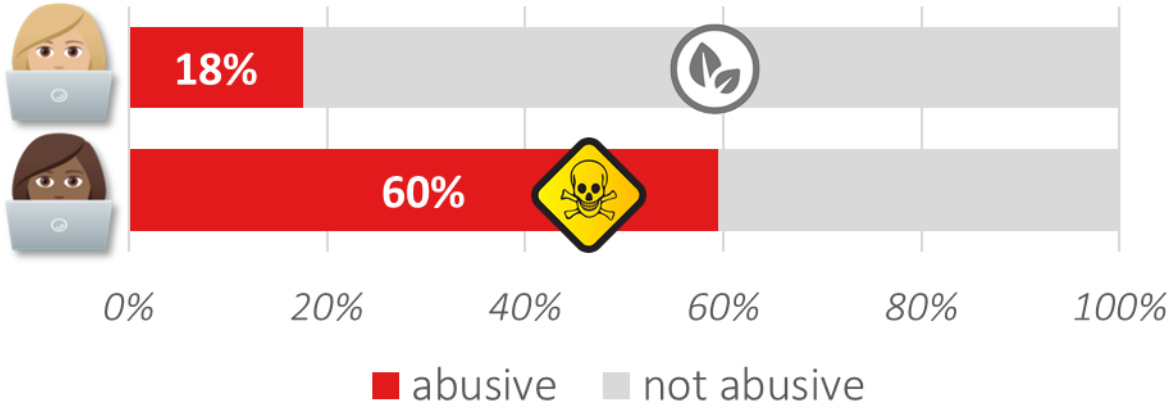
- **Lexical bias**: minority identity terms mistaken for toxicity [Dixon et al. '18]
- **Keyword bias**: over-reliance on "negative" keywords, (expletives, swearwords); subtle toxicity is missed
  [Dinan et al. '19; Breitfeller et al. '19; Han & Tsvetkov, '20]
- **Racial bias**: harmless text by Black authors or in African-American English (AAE) mistaken as toxic [Sap et al '19; Davidson et al '19]



Racial biases in two popular datasets [Sap et al 2019]



TWT-HATEBASE
(Davidson et al., 2017)



TWT-BOOTSTRAP
(Founta et al., 2018)

Sap et al. 2019

Dialect as proxy for racial identity

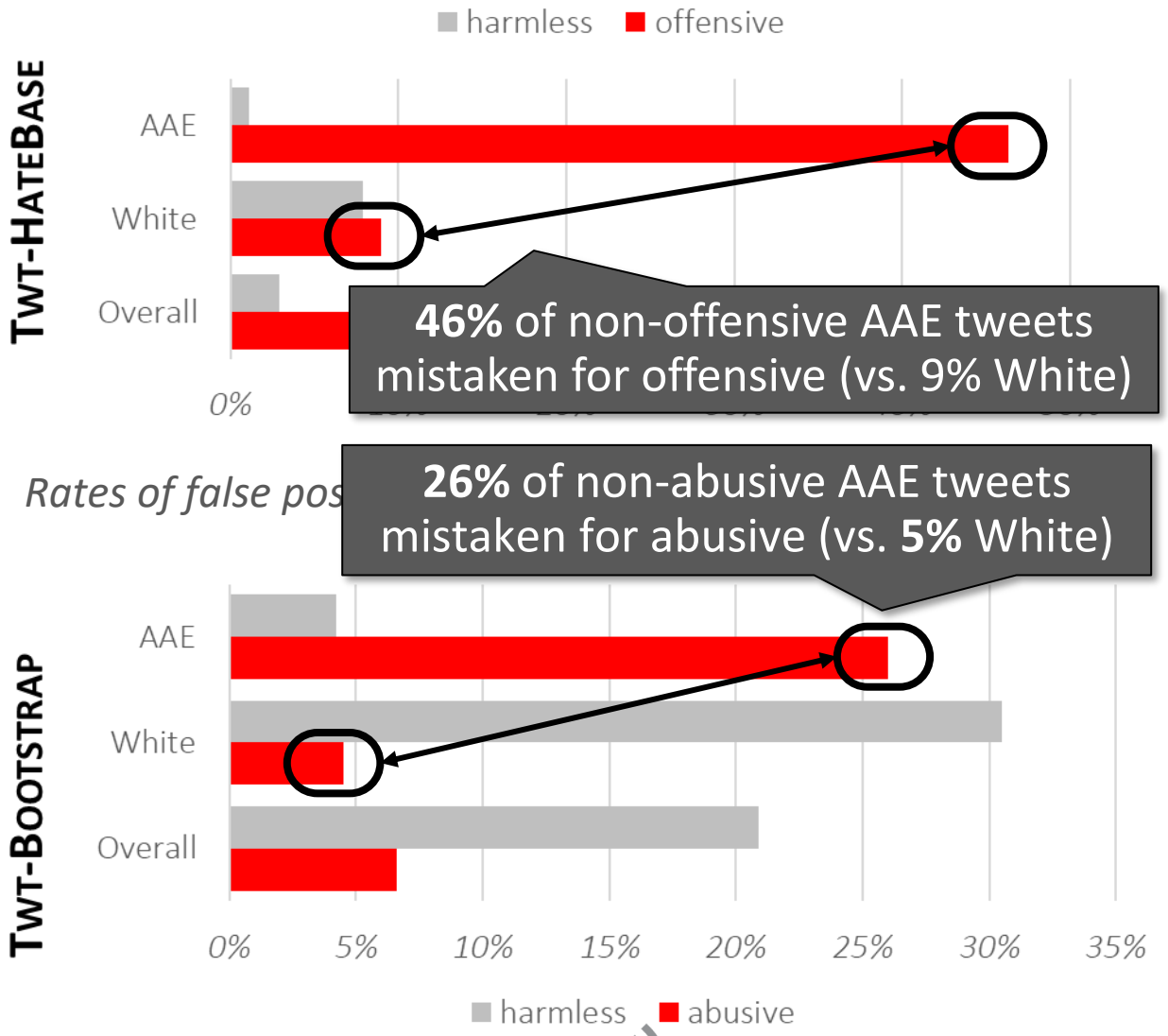
- *Challenge*: Twitter datasets don't have race information
- African American English (AAE) dialect
  - Common among (but not limited to) Black/African-American folks in U.S.
  - Variety of English that's extensively studied by linguists
  - Presence of AAE variants on Twitter [Jones, 2015]
- Lexical detector by Blodgett et al. (2016) to infer presence of AAE
  - Note: dialect/race much more complex than this

Image source: *https://www.languagejones.com/blog-1/2014/9/26/big-data-and-black-twitter*

Amplification of racial bias [Sap et al 2019]

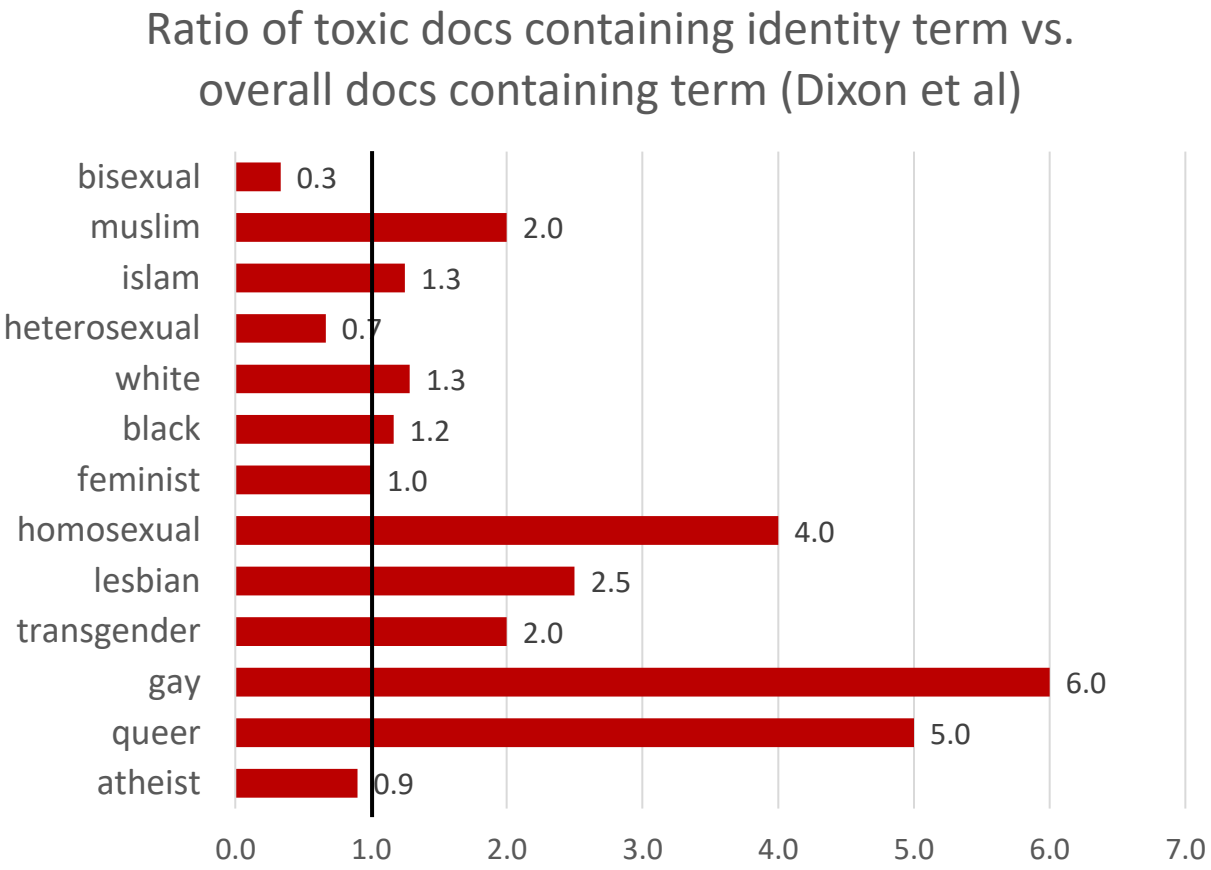
- Train/test two different classifiers
  - TWT-HATEBASE (Davidson et al, 2017)
  - TWT-BOOTSTRAP (Founta et al., 2018)
- Rates of **false flagging of toxicity**
  - Broken down by dialect group on heldout set
  - Equalized odds criterion [Hardt, 2016]

Predictions by both classifiers **biased against AAE tweets**



Same story for minority identity & profane words

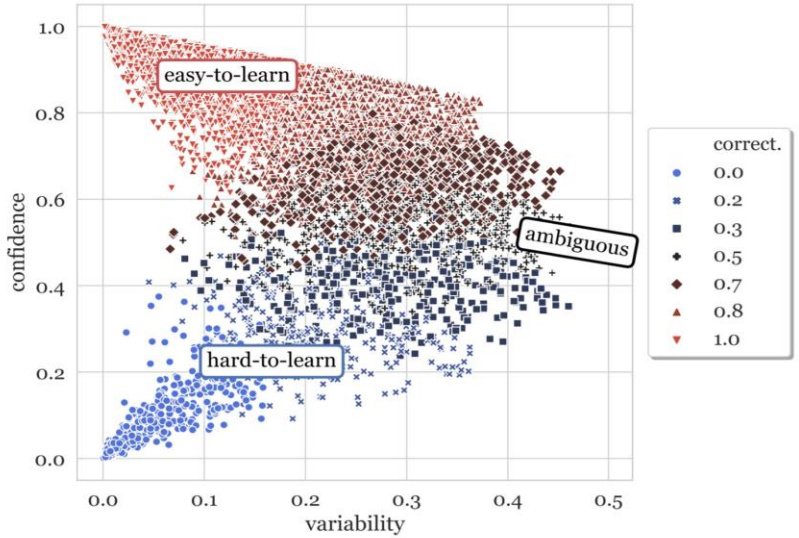
- Dixon et al 2018 analyzed comments from Wikipedia Talk Pages
  - Many minority terms had more toxic documents than non-toxic ones
- Zhou et al 2021 find similar conclusions on Twitter dataset
  - *r*(term, toxicity) = 0.04 for non-offensive identity terms
  - *r*(term, toxicity) = 0.26 for possibly offensive identity terms (e.g., slurs, reclaimed slurs, etc.)
  - *r*(term, toxicity) = 0.67 for non-identity referring swearwords



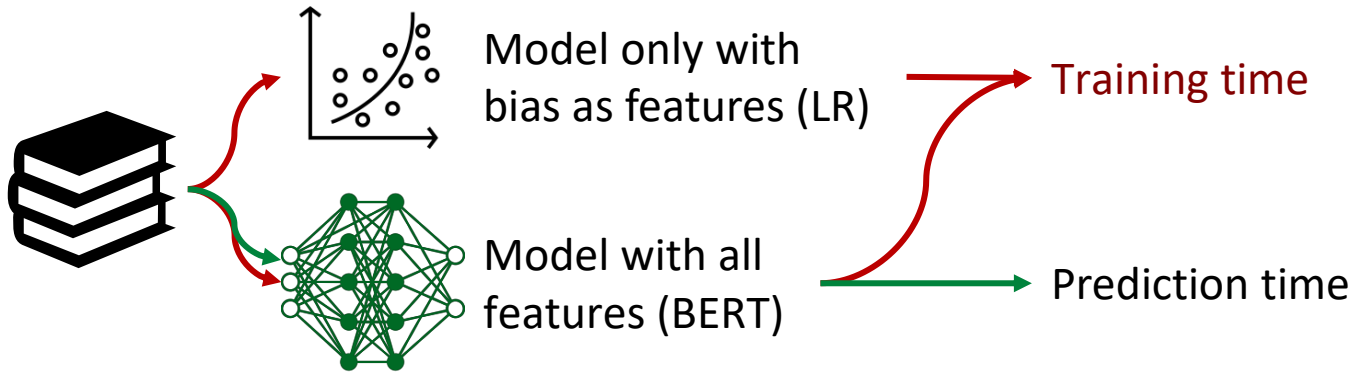
How to “debias” this system?

Automatic debiasing models and datasets [Zhou et al 2021]

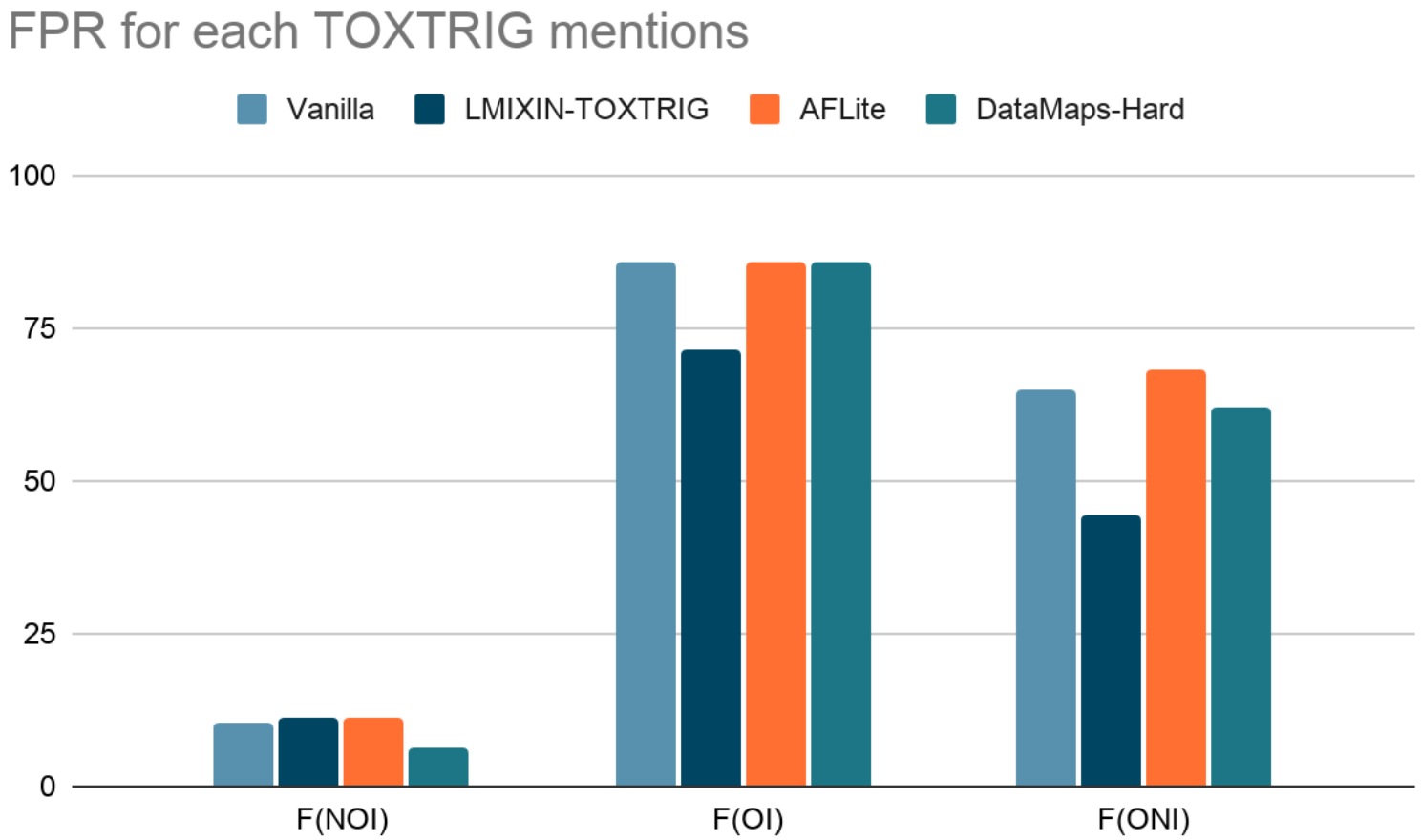
- Techniques for debiasing spurious correlations & annotation artefacts in NLI
- Dataset debiasing (filtering / subsampling)
  - AF-Lite [Le Bras et al 2020]
  - Dataset Cartography [Swayamdipta et al 2020]



- Model debiasing: Learned-Mixin [Clark et al 2019]
  - Training time: model with known biases + full model
  - Prediction time: full model only



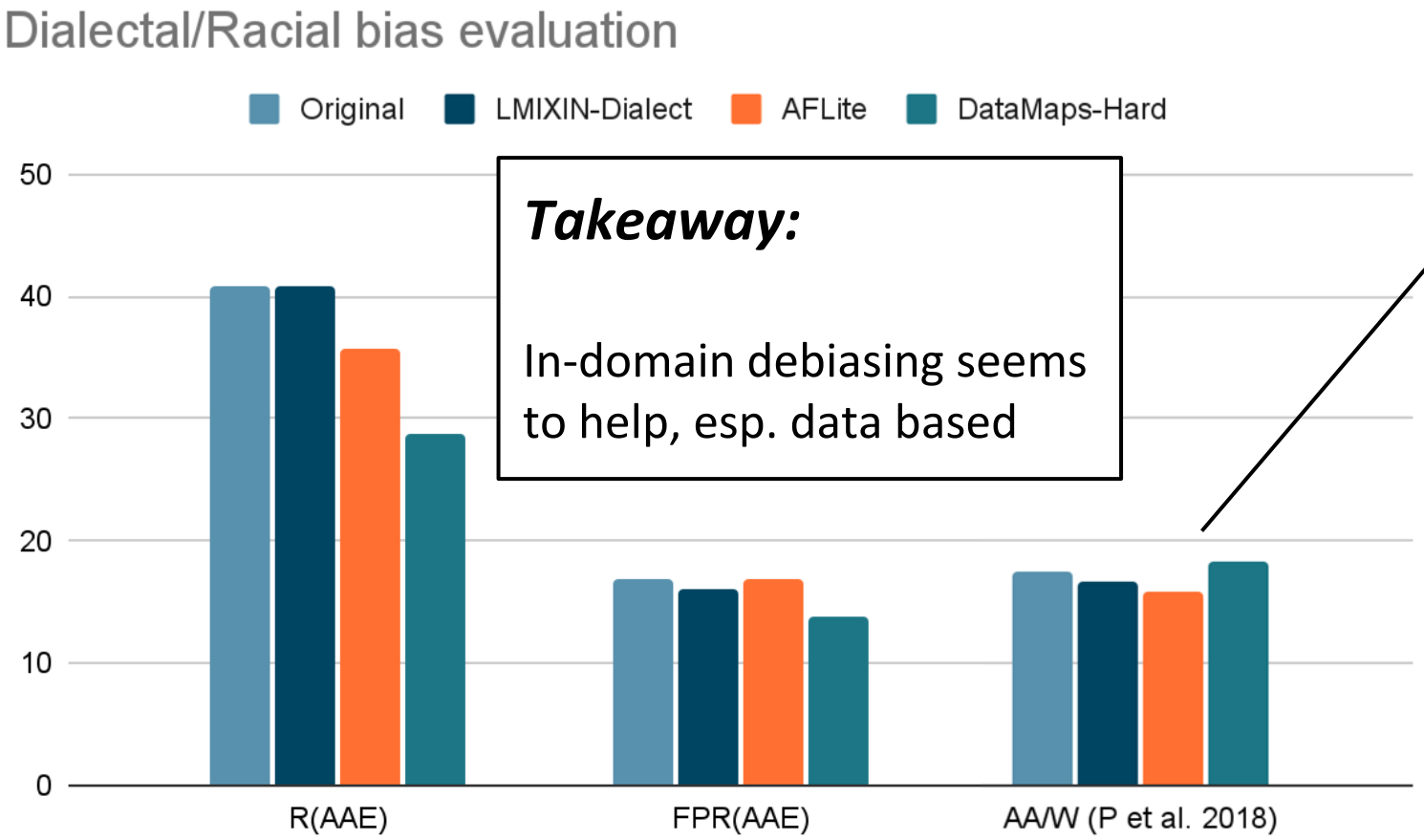
Results for Lexical Bias Reduction



Takeaway:

Debiased training approaches perform better on lexical bias reduction.

Results for Dialectal/Racial Bias Reduction



Takeaway:

In-domain debiasing seems to help, esp. data based

Self-reported race OOD dataset:

AA/W = ratio of AA tweets predicted as toxic/ratio of W tweets predicted as toxic

Takeaway:

But out-of domain it doesn't show effects.

How would you debias the hate speech detector?

- Model debiasing kind of works for lexical biases
- Data filtering seems to work for racial biases
- But doesn't actually, as shown on out-of-distribution data
- What else can we do? Let's discuss…

Fundamental problem in task setup

Current datasets/models **ignore social context** of speech
(e.g., identity of speaker, dialect of English)

Ignoring theses nuances is **harming minority populations** by suppressing inoffensive speech

Enhancing the labeling interface [Sap et al 2019]

MTurk study:

- 350 AAE tweets, ~50% labeled toxic
- 3 (re-)annotators per tweet

Control condition

Text-only, no context, prior work

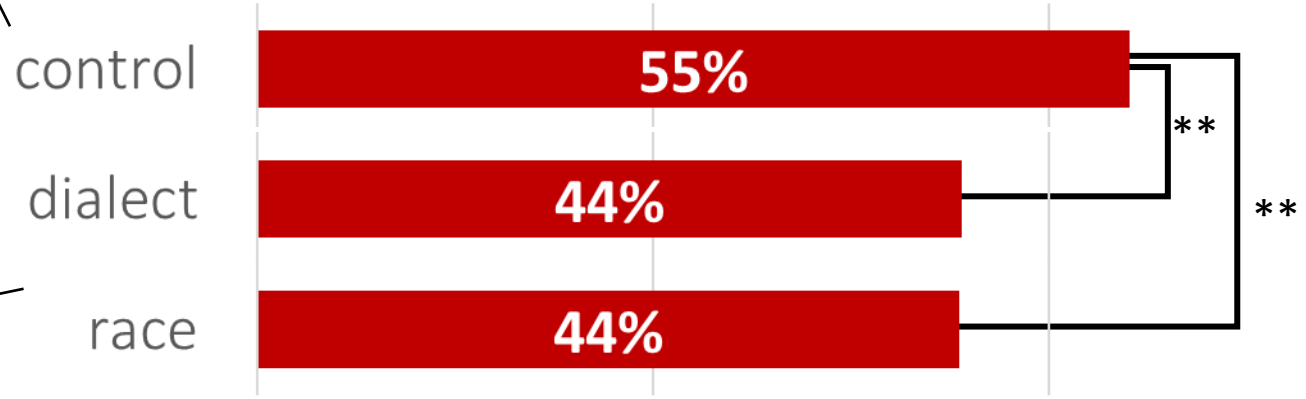
Dialect priming

"Our AI thinks this tweet is in African American English"

Race priming

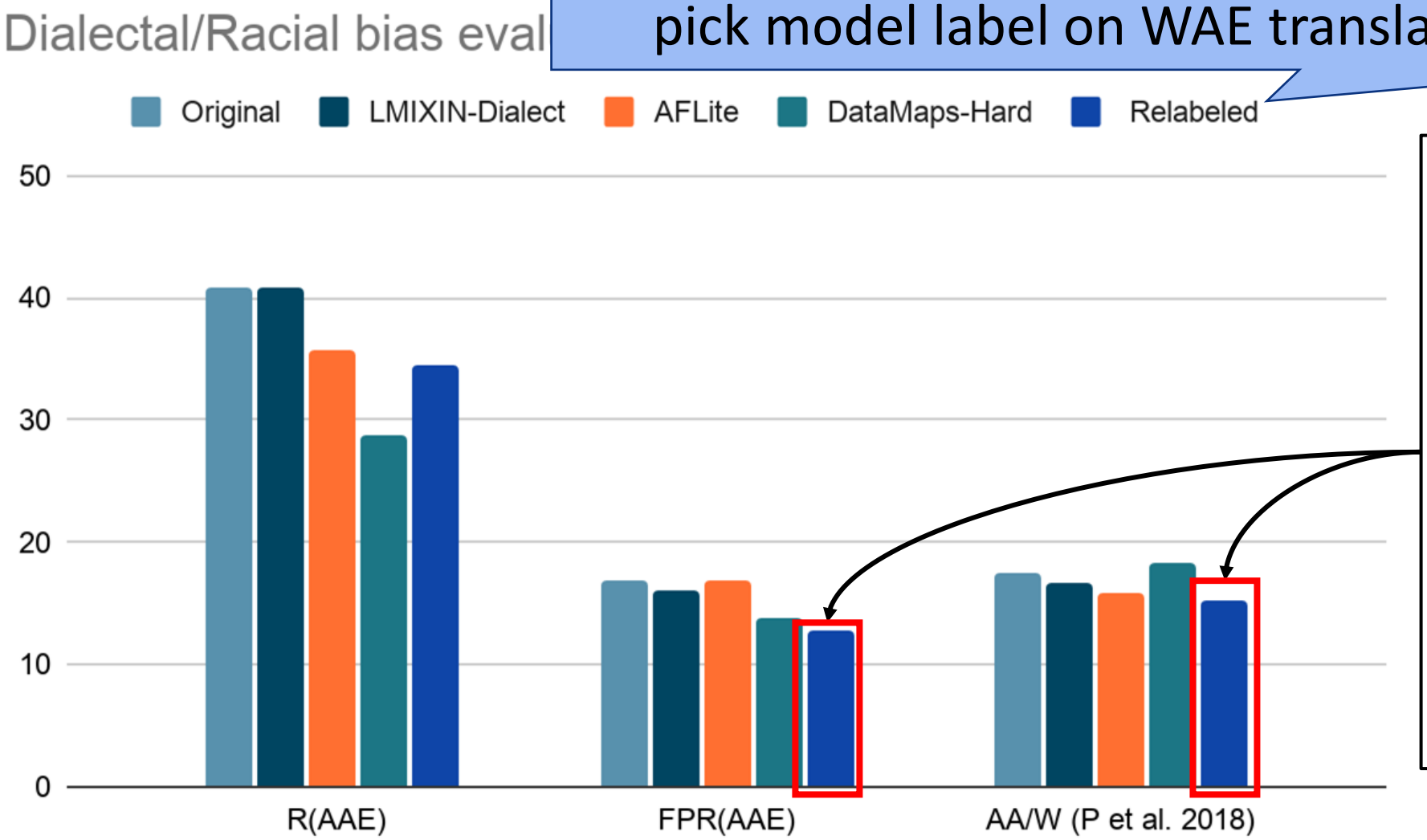
"A Twitter user that is likely Black/African American tweeted…"

Could this tweet be **offensive to *anyone*?**



Data Relabeling Pilot Experiment [Zhou et al 2021]

"translate" AAE to white-aligned English (WAE) and pick model label on WAE translated example



Takeaway:

Relabelling is seemingly more effective than current debiasing methods.

→ Context and dialect of speech matters for data labeling & how toxicity is perceived

Biases in hate speech takeaways

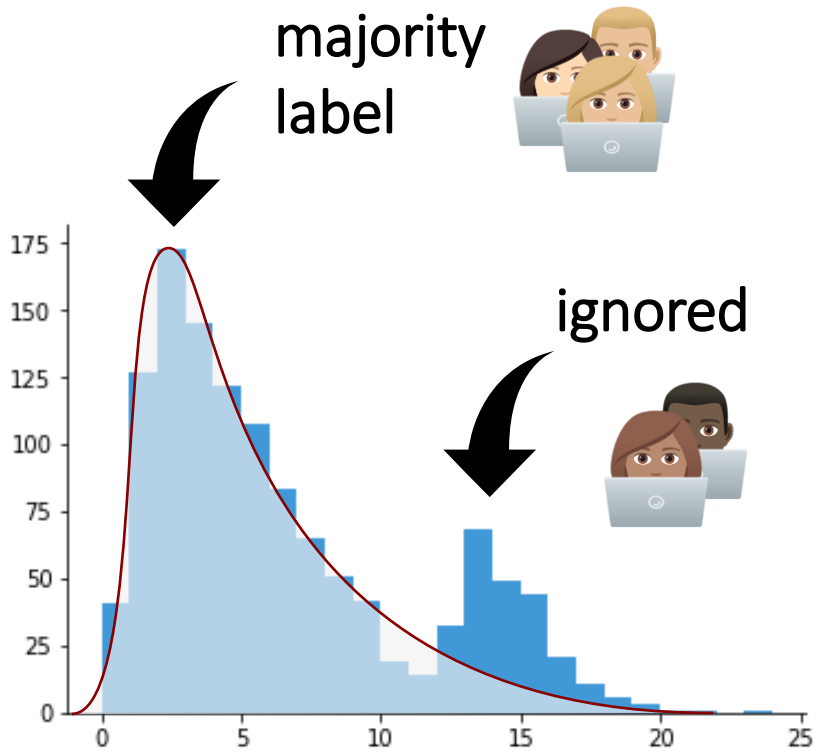
- Datasets and models contain racial and lexical biases
- Model debiasing and dataset filtering/subsampling don't really help much
- Changing the **social context** in which speech is evaluated can help
  - E.g., dialect/race priming, or "translating" to non-minority dialect
- Note: some biases "make sense"
  - Swearwords are likely to be conveying offensive meanings
  - Minority identities are more often targets of hate speech
- But, some biases are due to **fundamental ambiguity** in task
  - If you don't know **who** said something or **why**, how can you judge toxicity?

Perspectivism

Ambiguity and disagreements in tasks and labels

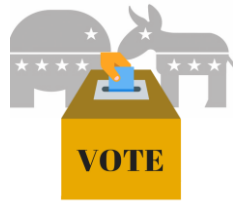
Annotators with attitudes [Sap et al 2022]

- Not all text is **equally toxic** for everyone
  - Al Kuwatly et al 2020, Akhtar et al 2021, etc.
- Determining what is toxic/harmful/biased is a **nuanced and subjective** task
- Most previous toxic language detection work **simplifies to one "correct" label**
- This **ignores/averages out** annotator variation
- Leads to **biases**, over- and under-detection of certain text characteristics



How do annotator identity and beliefs bias toxicity ratings? [Sap et al 2022]

- Who? Demographics: gender, race, politics
- Why? Attitudes about hate speech, free speech, racist beliefs, traditionalism, etc. *drawn from social psychology*
- What? Text characteristics that are likely over- or under-detected as toxic



Anti-Black or racist content

- Often the **desired target** for toxic language detection [e.g., Waseem & Hovy '16; Vidgen et al '20]
- **Possibly under-detected**, esp. if subtle

African American English (AAE)

- Well-studied variety of US English
- Known to be perceived as obscene by non-Black people, **over-detected as toxic** in toxic language detection [Spears '98; Rosa & Flores '17; Sap et al '19; Davidson et al '19]

Vulgar or profane words

- In this study, only looking at non-identity referring vulgarity (e.g., **swearwords, expletives,** vs. ~~slurs~~)
- Likely **over-detected as toxic** [Dinan et al '19]

Annotators with attitudes - findings

- Rating **Anti-Black posts** as ***offensive/racist***
  - endorsing free speech, racist beliefs ⬇
  - harm of hate speech ⬆
  - political liberalism, women ⬆
- Rating **AAE posts** as ***racist***
  - endorsing racist beliefs ⬆
  - political conservativism ⬆
- Rating **vulgar posts** as ***offensive***
  - endorsing traditionalism ⬆
  - linguistic purism, conservatism ⬆

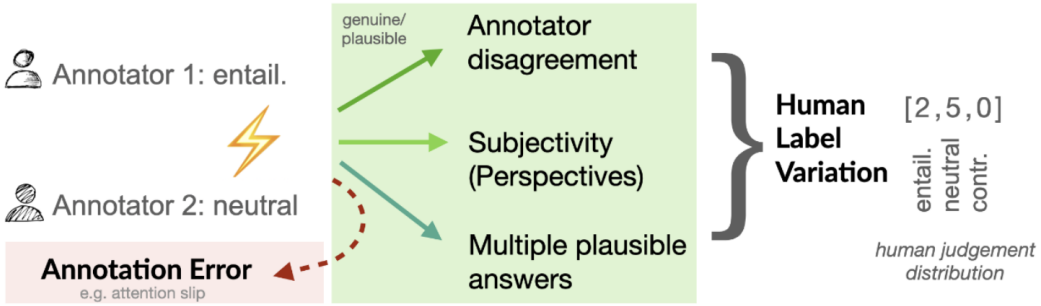
Anti-Black posts		*Rated as Offensive*		*Rated as Racist*	
EMPATHY		$r =$ 0.285	**	$r =$ 0.286	**
ALTRUISM		$r =$ 0.380	**	$r =$ 0.441	**
HARMOFHATESPEECH		$r =$ 0.451	**	$r =$ 0.528	**
FREEOFFSPEECH		$r =$ -0.394	**	$r =$ -0.467	**
RACISTBELIEFS		$r =$ -0.513	**	$r =$ -0.574	**
LINGPURISM		$r =$ -0.154	**	$r =$ -0.167	**
TRADITIONALISM		$r =$ -0.206	**	$r =$ -0.237	**
Politics	*(lib.: 0, cons.: 1)*	$r =$ -0.374	**	$r =$ -0.441	**
Gender	*(men: 0, women: 1)*	$d =$ 0.321	**	$d =$ 0.341	**
Race	*(White: 0, Black: 1)*	$d =$ 0.301	*	*n.s.*	

AAE posts		*Rated as Racist*	
RACISTBELIEFS		$r =$ 0.089	*
Politics	*(lib: 0, cons: 1)*	$r =$ 0.076	†

Vulgar (OnI) posts		*Rated as Offensive*	
LINGPURISM		$r =$ 0.106	*
TRADITIONALISM		$r =$ 0.252	**
Politics	*(lib: 0, cons: 1)*	$r =$ 0.171	**

Human label variation

- Some labeling tasks are inherently ambiguous
- Even "core NLP tasks"
  - Linguistic veridicality [de Marneffe et al 2012]
  - Coreference resolution [Swayamdipta et al 2021]
  - Entailment (NLI, MNLI) [Pavlick et al 2019]



The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation
(Barbara Plank, EMNLP 2022)

P: *A mom is feeding two babies.*
H: *A mom is giving her children carrots to eat.*
Contradiction? Neutral?

- Since most NLP tasks aggregate their annotations, a lot of variation is lost
- Thus, models represent certain perspectives more than others

NLP dataset and model positionality

Social context & positionality

- A person's **positionality** (concept from critical studies)

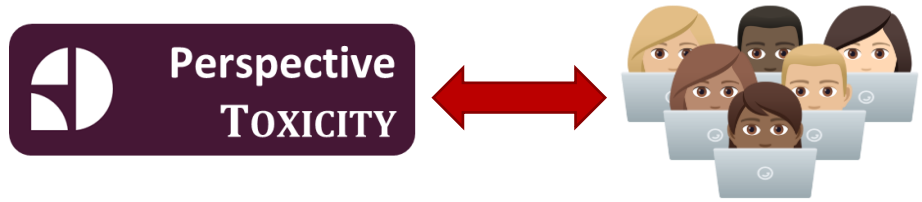
"the perspectives [people] hold as a result of their demographics, identity, and life experiences.

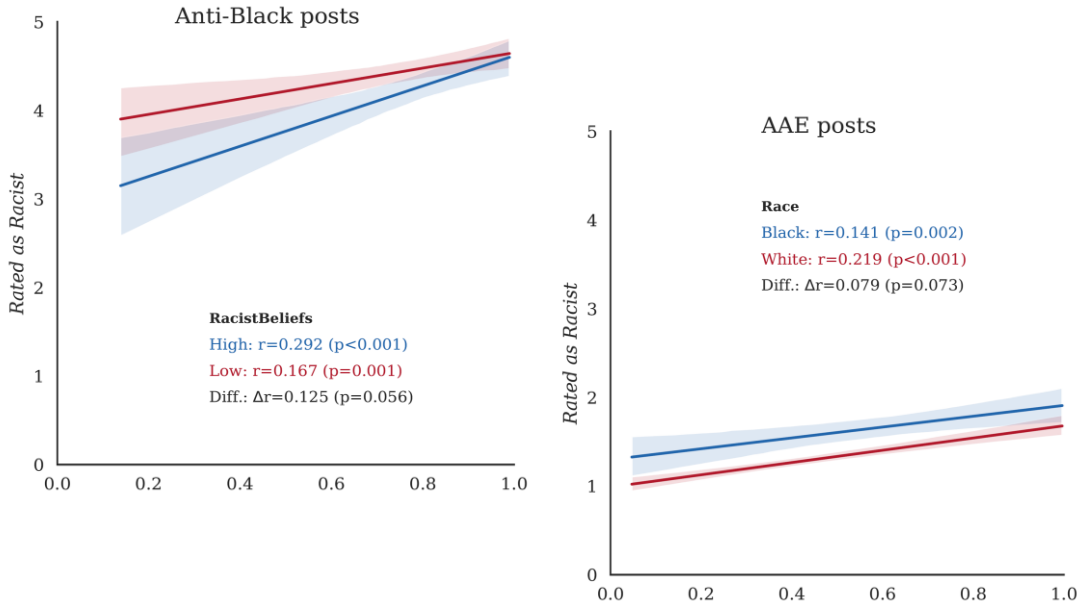
[As a researcher,] it influences the research process and its outcomes and results." [Savin-Baden & Howell-Major, '13]

- **Model and dataset positionality**: models and datasets embody or align with certain positionalities more than others [Cambo & Gergle '22; Sap et al '22]

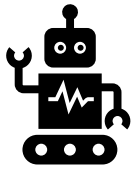
Annotators with attitudes – model positionality

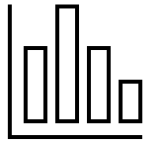
- Sap et al 2022 case study: PerspectiveAPI toxicity scores for tweets
- vs. how annotators of different demographics and attitudes score them
- Some results:
  - Persp. scores for anti-Black posts are more in line with ratings by annotators who are **high in endorsing racist beliefs** vs. low
  - Persp. scores for AAE posts are more associated **with ratings by white annotators** vs. Black annotators





Biases aren’t just coming from annotators!





Data distributions or characteristics

Data selection & filtering

Data labeling & annotation

Selection of data features & task setup

Choice of machine learning model

Choice of optimization / loss function

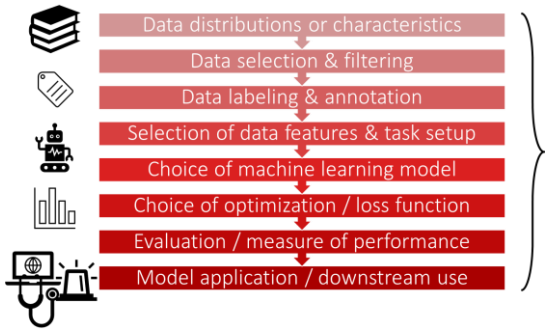
Evaluation / measure of performance

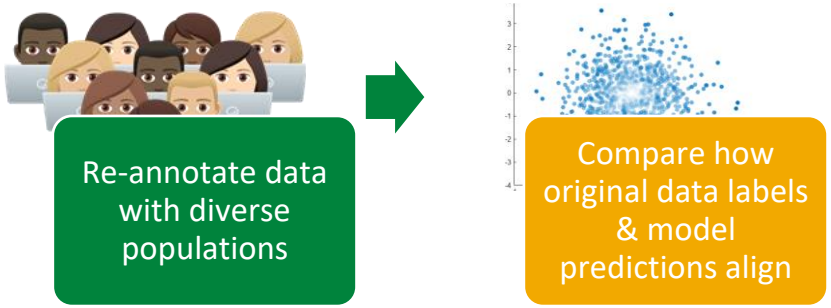
Model application / downstream use

Influenced by **research/designer positionality**

NLPositionality [Santy et al '23]

- Too many design choices when building a system → difficult to trace origin of biases
- NLPositionality: framework to measure which groups NLP systems align with better
  - Continuously collecting on LabInTheWild, a volunteer-based crowdsourcing platform
- Results:
  - English-centric, college-educated skew
  - Non-binary and non-white folks less aligned

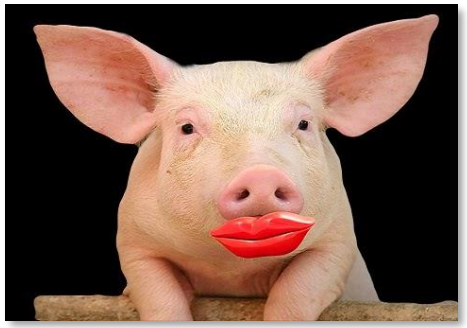




nlpositionality.cs.washington.edu
for live results (updated daily)

Summary: limits of debiasing

- Gender debiasing doesn't work
  - Breaks down for non-binary genders, racial categories or other social identity types
- Intrinsic debiasing ≠ actual debiasing
  - Finetuning often reintroduces biases
  - Out-of-distribution data often still show biases
- *Real world vs. ideal world*: is reflecting the (biased) status quo the goal? or do we want to build a more fair or just world?
- Justice and fairness go beyond data & model fairness



"Lipstick on a pig" paper,
Gonen & Goldberg 2019

A lot of people have understood that we need to have more diverse datasets, but unfortunately I felt like that's kind of **where the understanding has stopped**. It's like *'let's diversify our datasets. And that's kind of ethics and fairness, right?'* But you **can't ignore social and structural problems**.

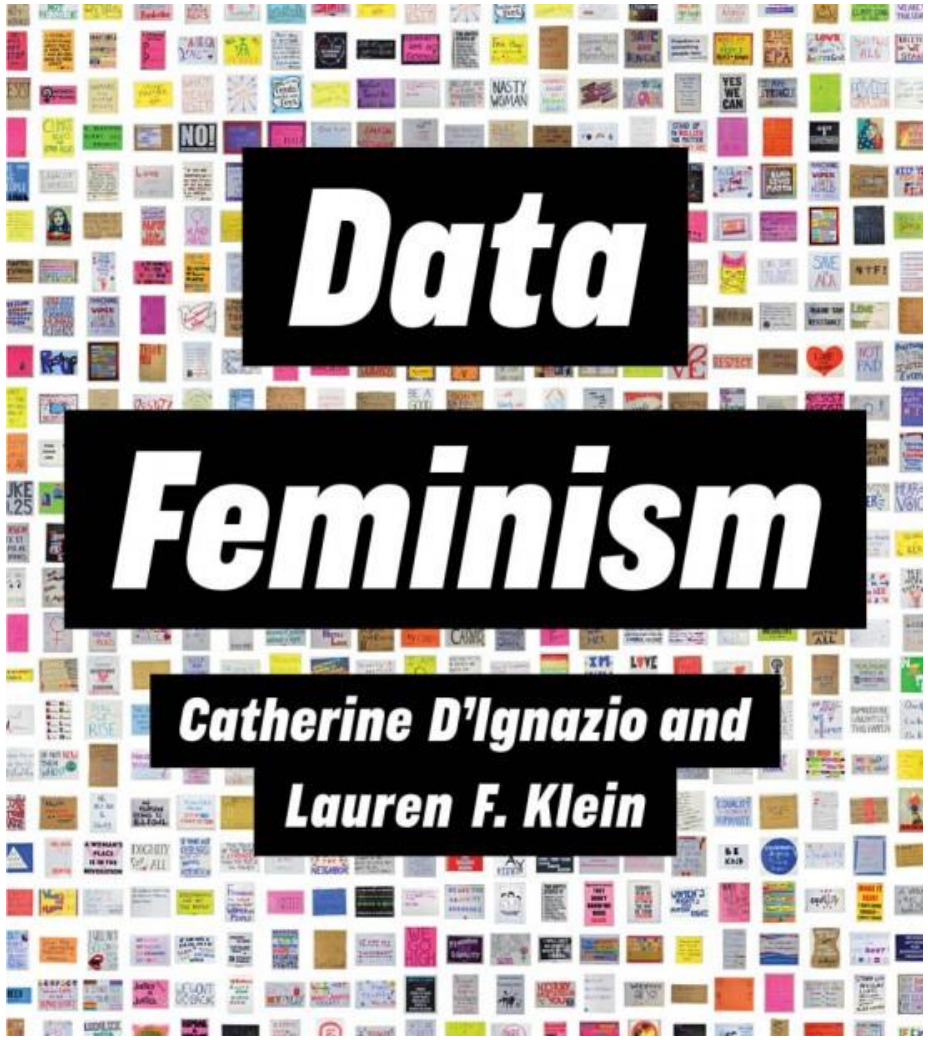
Timnit Gebru, PhD

Socio-technical view on bias & fairness

- You can have an “fair” NLP/ML model (e.g., facial recognition system)
  - 95% accuracy/error rate on white & Black faces
- But if the system is used by law enforcement, bias creeps in w.r.t. who the system is used on
  - Black people more often arrested, due to racial biases
- Actual error rates are a function of deployment
- Algorithm’s fairness ≠ fairness of treatment

7 principles of Data Feminism

Examine power	•Data feminism begins by analyzing how power operates in the world.
Challenge power	•Data feminism commits to challenging unequal power structures and working toward justice.
Elevate emotion & embodiment	•Data feminism teaches us to value multiple forms of knowledge, including the knowledge that comes from people as living, feeling bodies in the world.
Rethink binaries & hierarchies	•Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.
Embrace pluralism	•Data feminism insists that the most complete knowledge comes from synthesising multiple perspectives, with priority given to local, indigenous and experiential ways of knowing.
Consider context	•Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations and this context is essential for conducting accurate, ethical analysis.
Make labor visible	•The work of data science is the work of many hands. Data feminism makes this labour visible so that it can be recognised and valued.



https://feedmagazine.tv/interviews/lauren-klein-theres-no-such-thing-as-raw-data/