DSCI-531: FAIRNESS IN ARTIFICIAL INTELLIGENCE

BIASES IN TEXT EMBEDDINGS

Kristina Lerman

Spring 2025

Is it possible to discriminate through text and language?

AI promised to eliminate human bias from decisions

… but algorithms are only as good as the data they were trained on

Data reflects explicit prejudices …	… implicit biases in society	… patterns of discrimination, exclusion and inequality

"Unthinking reliance on data mining can deny historically disadvantaged and vulnerable groups full participation in society"

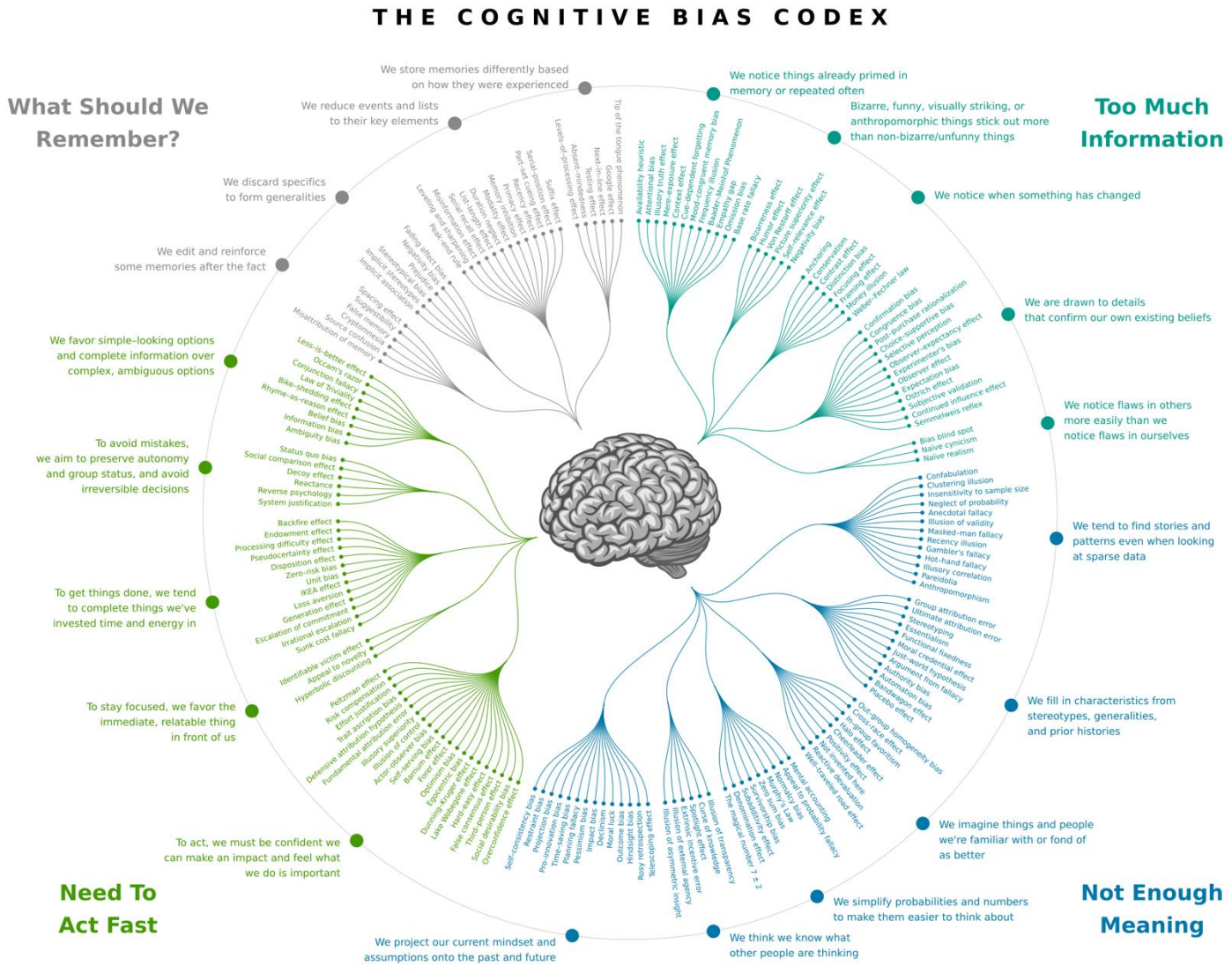
[Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.]

Bias is all around us

Human brain evolved a number of adaptations to uncertain, dynamic world

- Negativity bias: we pay more attention to negative information than to positive information
- Loss aversion: we fear loss more than we value gain
- In-group bias: we favor people from our own tribe
- Availability heuristic: we judge the likelihood of events by how easily they come to mind

Cultural & social biases (e.g., stereotypes) can be learned



Prejudice: Explicit and implicit bias

prej·u·dice

- preconceived opinion (usually negative) that is not based on reason or actual experience.
- Explicit bias
  - the person is aware of their feelings and attitudes, and intentional in behaviors; Characterized by prejudice expressed through physical and verbal actions. E.g., “Women who are mothers are not serious about their research.”
- Implicit bias
  - Learned attitudes or stereotypes that exist in our subconscious and can involuntarily affect the way we think and act. E.g., Not inviting female researchers into research collaborations.”
  - Other types of stereotypes: age, body image, accent, …
- These biases are imprinted in our society, including in language.
  - How do we identify/measure biases
  - How do we mitigate biases

A father and son get in a car crash and are rushed to the hospital.

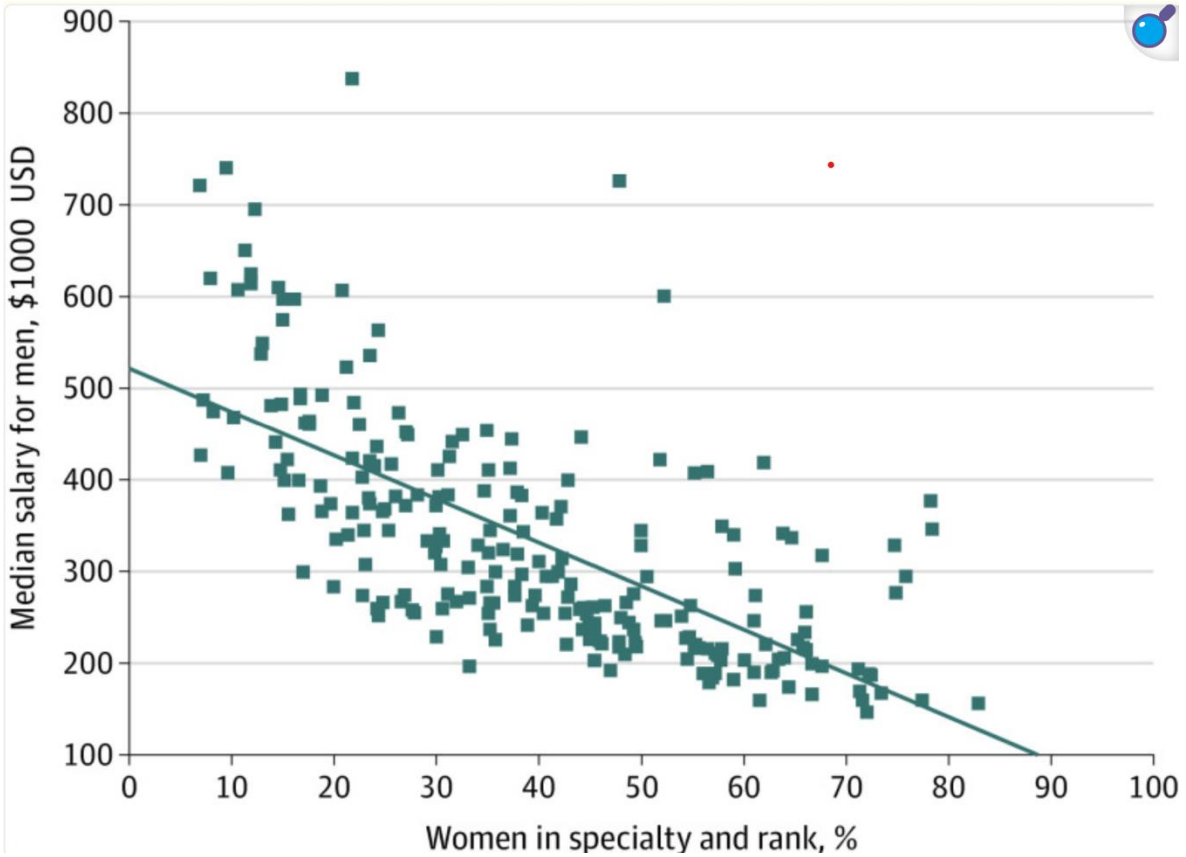
The father dies.

The boy is taken to the operating room and the surgeon says,

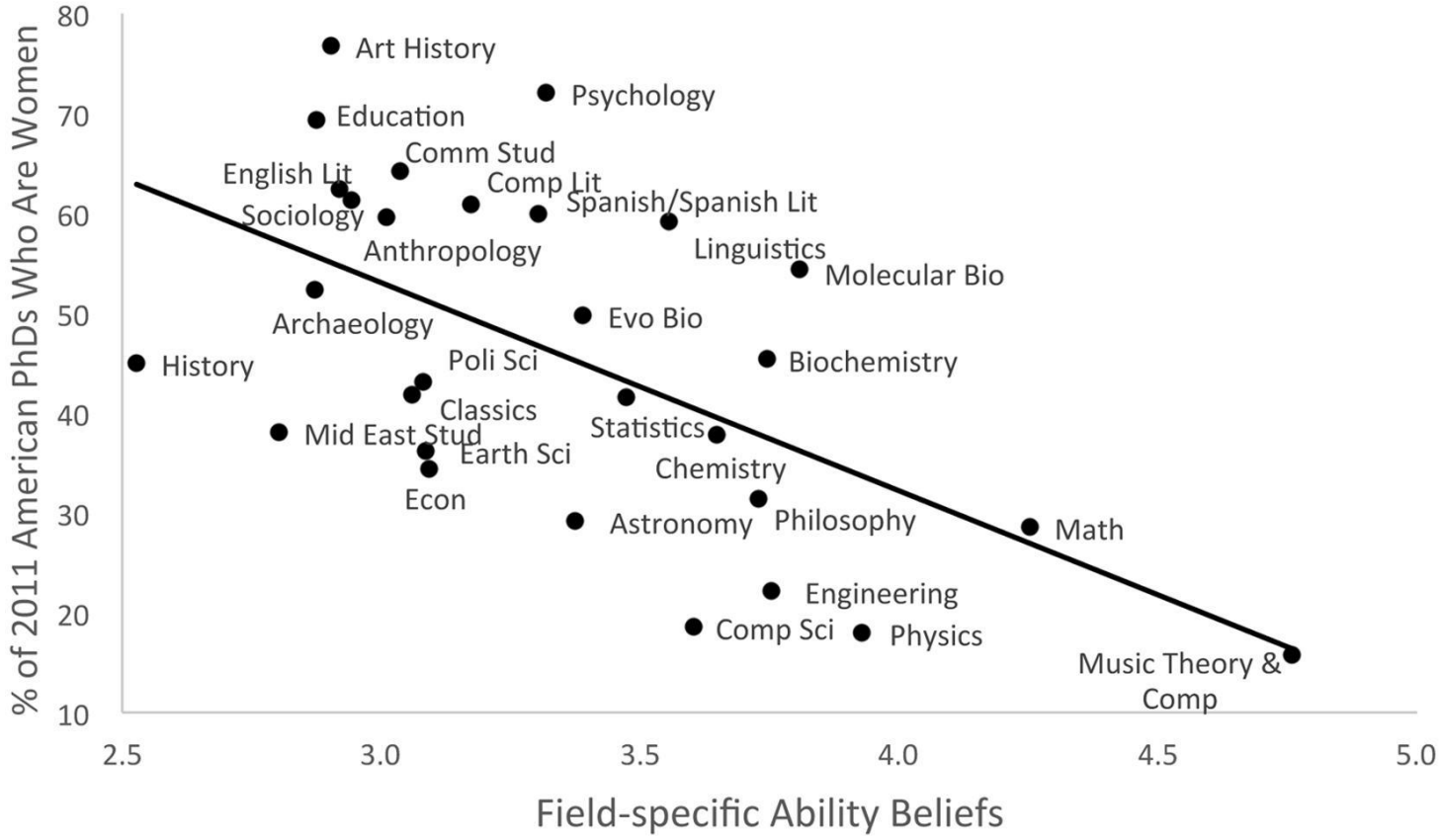
“I can’t operate on this boy, because he’s my son.” Can you explain why?

Gender pay disparity in medicine

- For every 10% increase in the percentage of women in a medical specialty, there was a corresponding decrease in median salaries for both men and women. Specifically, men's median salaries decreased by $7,465, while women's median salaries decreased by $15,003.
- Women are overrepresented in lower-paying specialties such as pediatrics and family medicine, while men dominate higher-paying fields like orthopedic surgery and cardiology.



Women are underrepresented in fields where success is believed to require brilliance



Fields emphasizing the need for raw, innate talent—such as physics, mathematics, and philosophy—had lower representations of women and African Americans. In contrast, disciplines that valued hard work and dedication over inherent brilliance tended to be more diverse.

Meyer, M., Cimpian, A., & Leslie, S. J. (2015). Women are underrepresented in fields where success is believed to require brilliance. *Frontiers in psychology*, *6*, 235.

What are some of the consequences of bias?

- Harms
  - Perpetuates inequity. E.g., gender bias in healthcare resulted in women's health conditions receiving less attention and treatment.
  - AI may create a feedback loops that keep bias going
  - Affect decision making e.g., hiring discriminates against women
  - Lack of income parity
  - Harm product quality
- "Benefits"
  - Beneficial to the other group
  - Targeted marketing
  - Greater social cohesion

Language & Bias

- Language implicitly transmits ingroup/outgroup identify information, leading to human prejudices
- Language models (e.g., word embeddings) encode stereotyped biases in addition to other knowledge, including how flowers are pleasant and the gender distribution of occupations.
- Behavior can be driven by cultural history embedded in language, and can vary between languages.
  - Creating a feedback loop

- Bias enters data in other ways, e.g., demographics, test scores, …

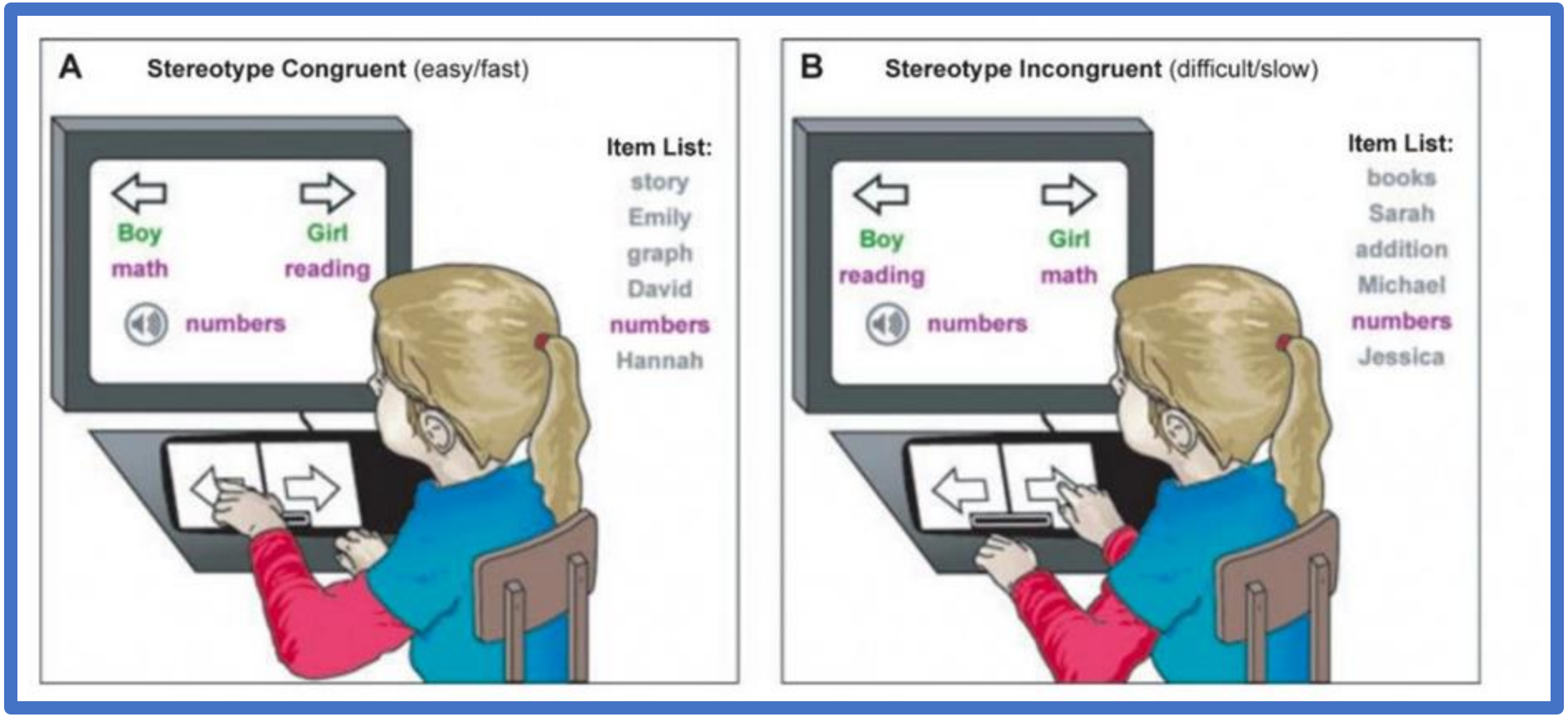
Which words come to mind? - Flowers

Which words come to mind? - Insects

Implicit association test (IAT)

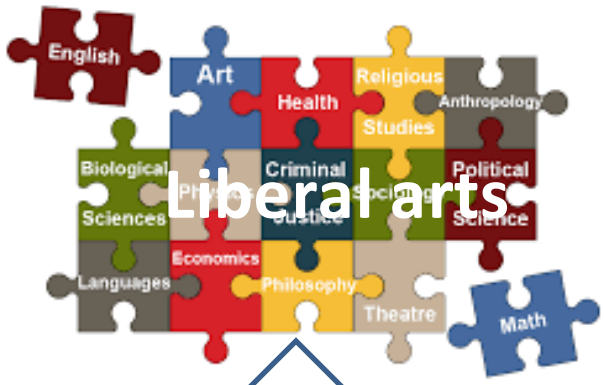
- Implicit Association Test (IAT) used by psychologists to measure implicit prejudices and stereotypes
- IAT measures differences in response times when subjects are asked to pair two concepts they find similar, in contrast to two concepts they find different.
- Response times are faster when tasks are easier: e.g., subjects are much quicker if they are told to label insects as unpleasant and flowers as pleasant than if they are asked to label these objects in reverse

Implicit association test (IAT)

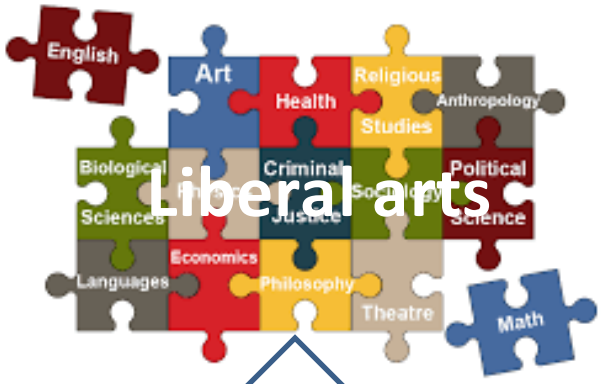


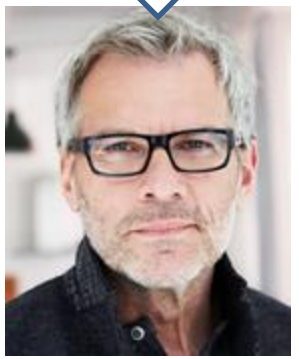
https://implicit.harvard.edu/implicit/

It is easier (quicker) to associate some concepts together than others



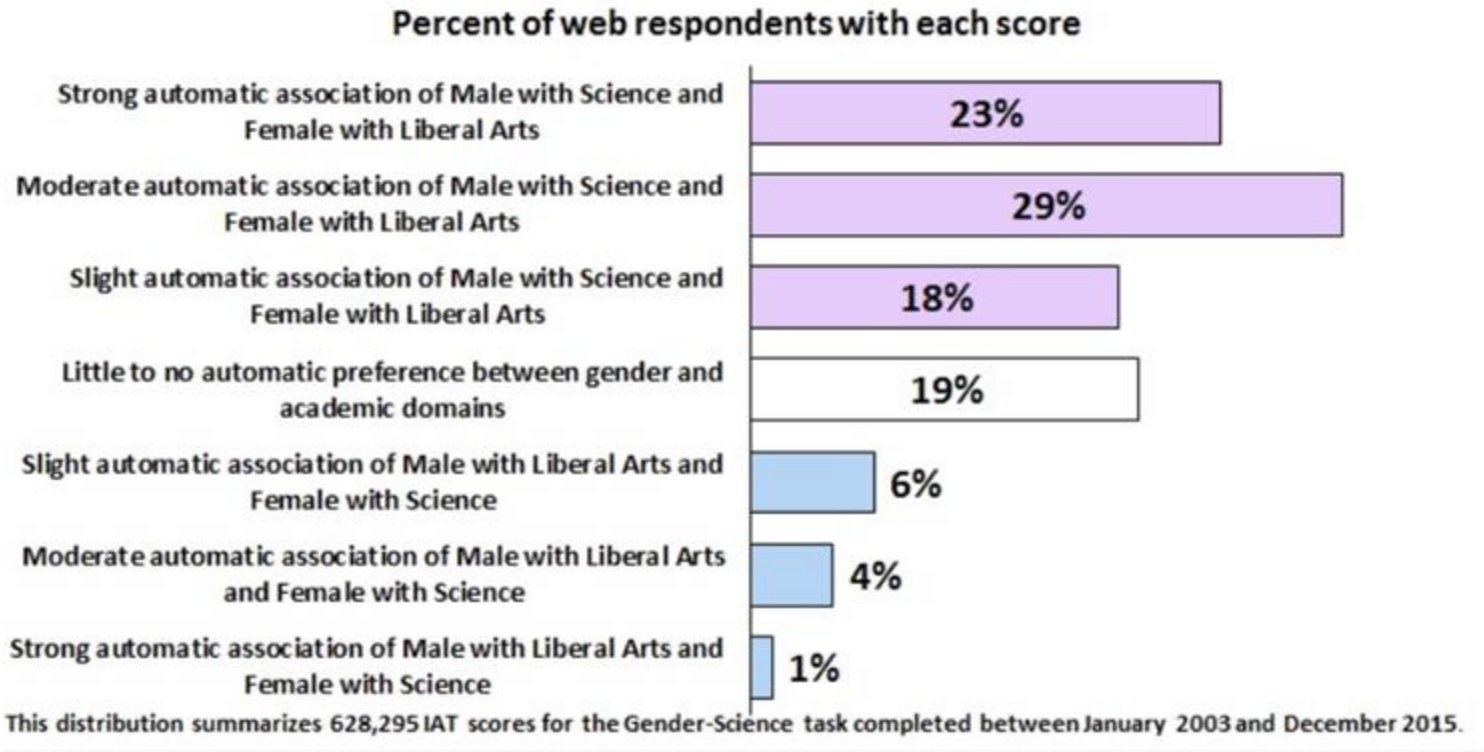
VS





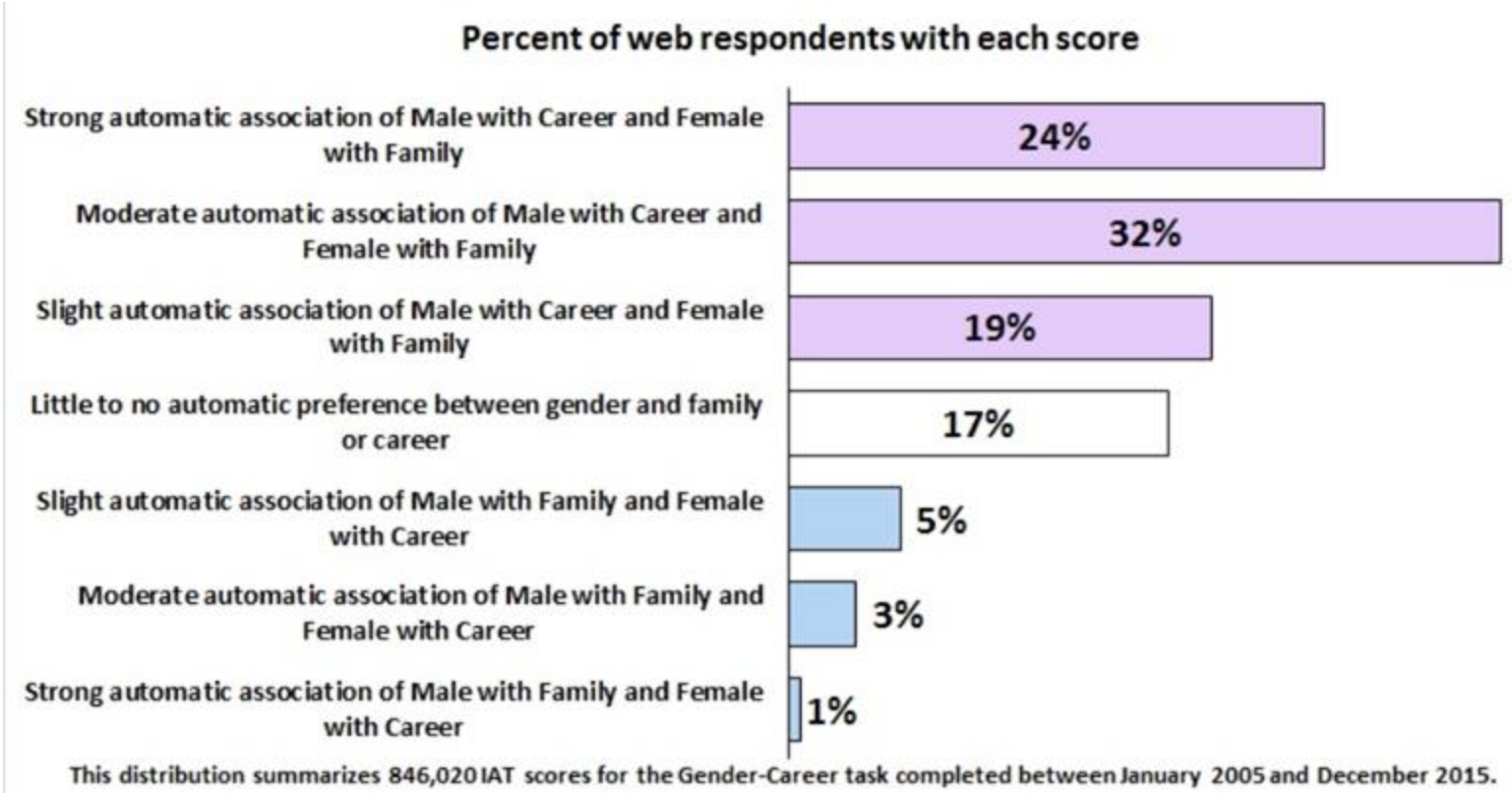
During the Implicit Association Test (IAT) you just completed:

Your responses suggested a moderate automatic association for Male with Science and Female with Liberal Arts.



During the Implicit Association Test (IAT) you just completed:

Your responses suggested a strong automatic association for Male with Career and Female with Family.



What about language models?

The New York Times

We Teach A.I. Systems Everything, Including Our Biases

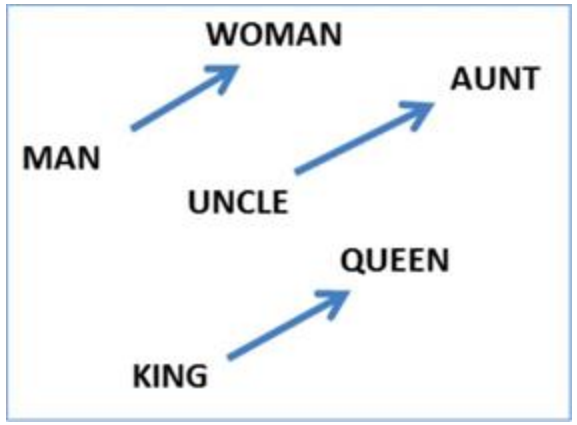
Researchers say computer systems are learning from lots and lots of digitized books and news articles that could bake old attitudes into new technology.

https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html

- Language models acquire stereotypes from text that reflect human culture
- Cultural stereotypes propagate into applications using these models

Word Embeddings can be Dreadfully Sexist

$$v_{man} - v_{woman} + v_{uncle} \sim v_{aunt}$$



he: _____	she:_____
brother	sister
beer	cocktail
physician	registered_nurse
programmer	homemaker
professor	associate professor

Use Google w2v embedding trained from the news (Glove)

Word embeddings exhibit gender bias

"genius" is male but "beauty" is female?

Source: https://towardsdatascience.com/tackling-gender-bias-in-word-embeddings-c965f4076a10

Word Embedding Association Test (WEAT)

- WEAT measures implicit bias of AI
- Use word embeddings to measure distance from **attribute words** (A/B: e.g., flowers/insects) to **target words** (X/Y: e.g., pleasant/unpleasant)
- The null hypothesis is that there is no difference between the two sets of target words in terms of their relative similarity to the two sets of attribute words.
- The permutation test measures the (un)likelihood of the null hypothesis by computing the probability that a random permutation of the attribute words would produce the observed (or greater) difference in sample means.
- The **effect size** is a normalized measure of how separated the two distributions (of associations between the target and attribute words) are:

$$\frac{\text{mean}_{x \in X} s\left(x,\ A,\ B\right) - \text{mean}_{y \in Y} s\left(y,\ A,\ B\right)}{\text{std\_dev}_{w \in X \cup Y} s\left(w,\ A,\ B\right)}$$

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

Which concepts are closer in the embeddings space?

- **Flowers**: aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia.
- **Insects**: ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil.
- **Pleasant**: caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation.
- **Unpleasant**: abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison.

Word embeddings replicated human biases

Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	(5)	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	(5)	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	(5)	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from (5)	(7)	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from (9)	(7)	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	(9)	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	(9)	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	(10)	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	(23)	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	(9)	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

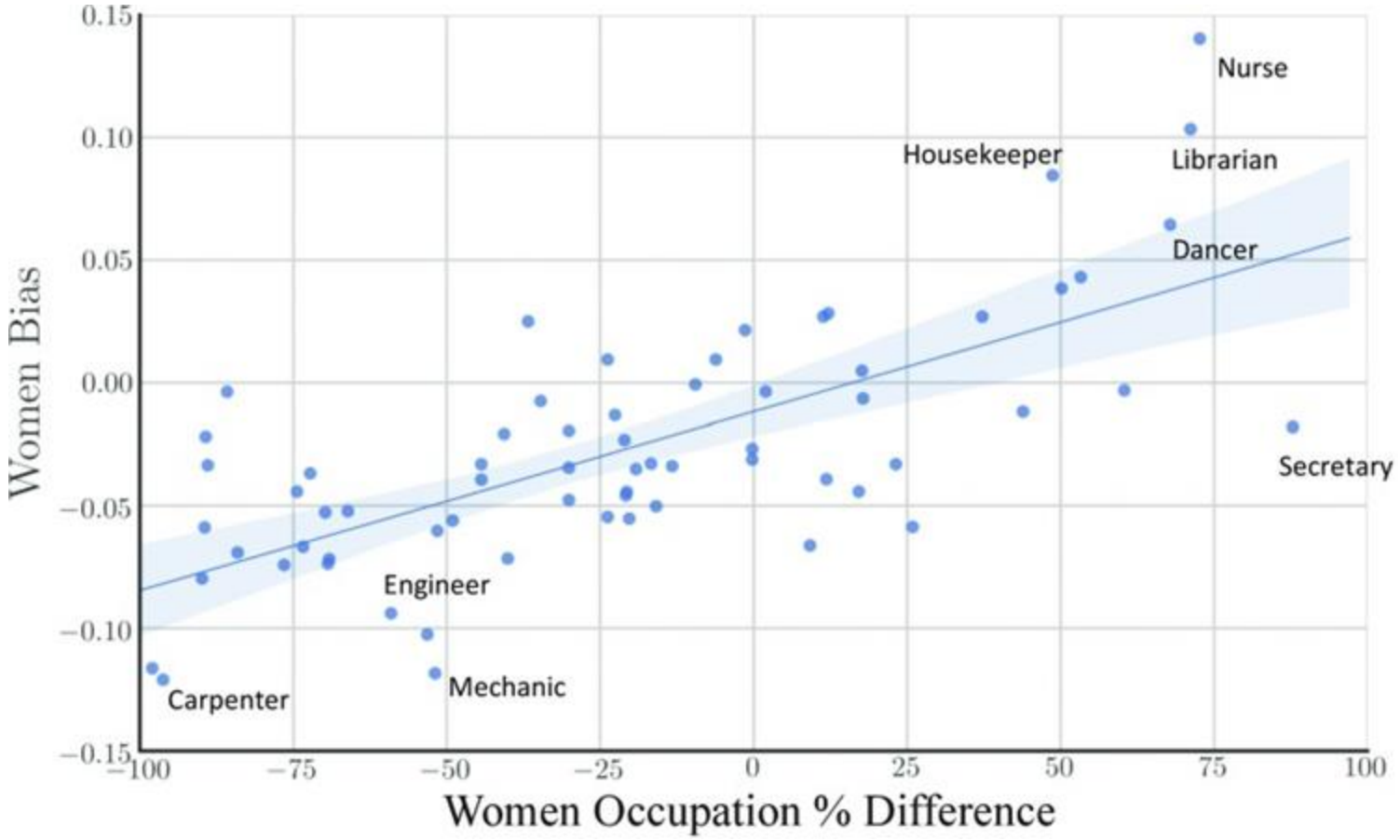
Gender biases are historical

Garg, Schiebinger, Jurafsky, Zou (2017) Word embeddings quantify 100 years of gender and ethnic stereotypes:

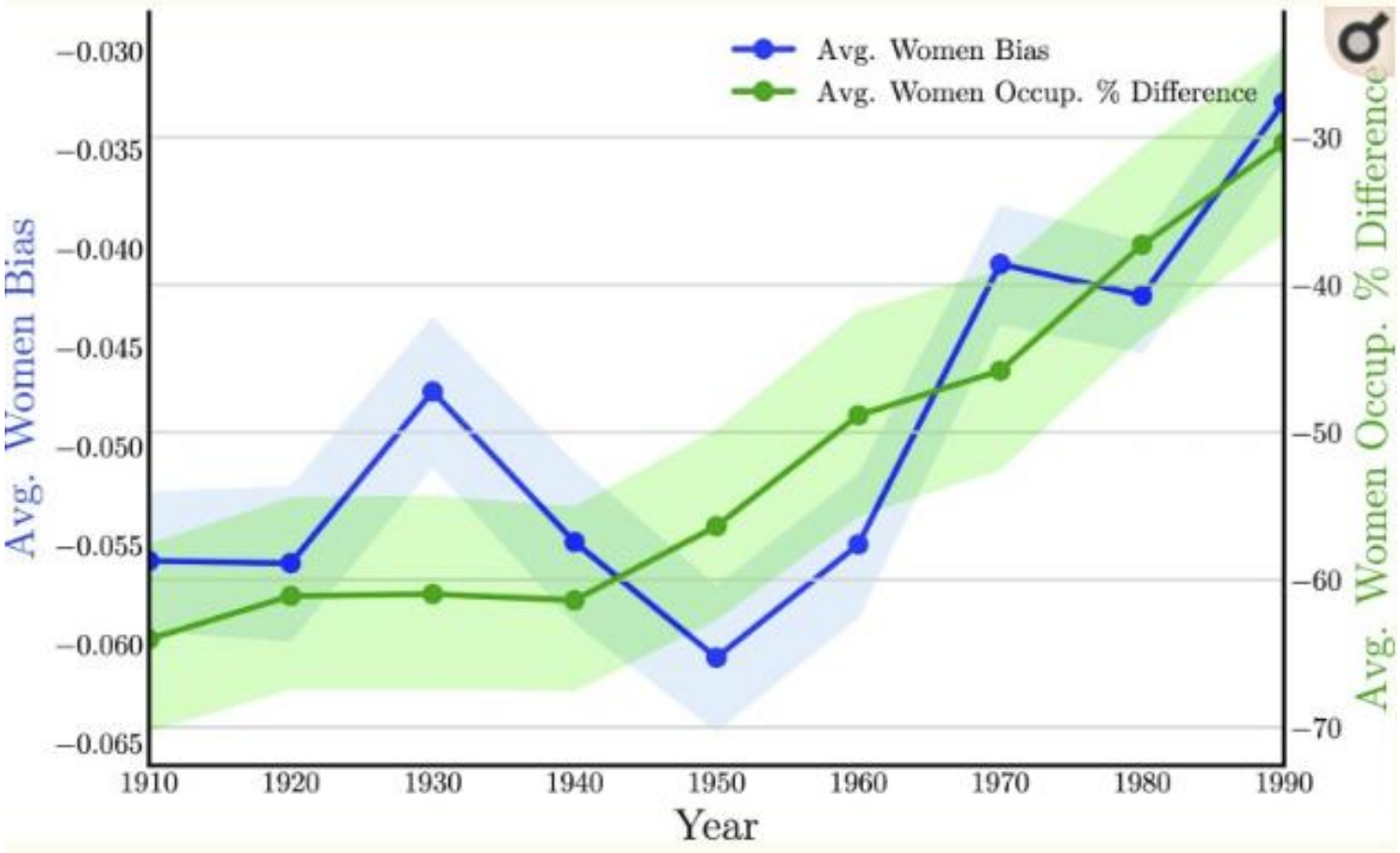
(a) Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding.

1910	1950	1990
charming	delicate	maternal
placid	sweet	morbid
delicate	charming	artificial
passionate	transparent	physical
sweet	placid	caring
dreamy	childish	emotional
indulgent	soft	protective
playful	colorless	attractive
mellow	tasteless	soft
sentimental	agreeable	tidy

Another View



… By Time


Avg. Women Bias
Avg. Women Occup. % Difference
−0.030
−0.035
−0.040
−0.045
−0.050
−0.055
−0.060
−0.065
−30
−40
−50
−60
−70
1910
1920
1930
1940
1950
1960
1970
1980
1990
Year

Mitigating Gender Bias in NLP

- Debias training data
  - Data augmentation: Swap gender in text samples to create balanced data sets
- Debias word embeddings
  - Remove gender subspace
  - Train a gender-neutral embedding
    - isolate gender information in some dims: maximize difference between the gender dimension in male and female definitional word embeddings
    - maintain gender-neutral information in other dimensions: maximize the difference between the gender direction and the other neutral dimensions in the word embeddings

Methods	Method Type
Data Augmentation by Gender-Swapping	Retraining
Gender Tagging	Retraining
Bias Fine-Tuning	Retraining
Hard Debiasing	Inference
Learning Gender-Neutral Embeddings	Retraining
Constraining Predictions	Inference
Adjusting Adversarial Discriminator	Retraining

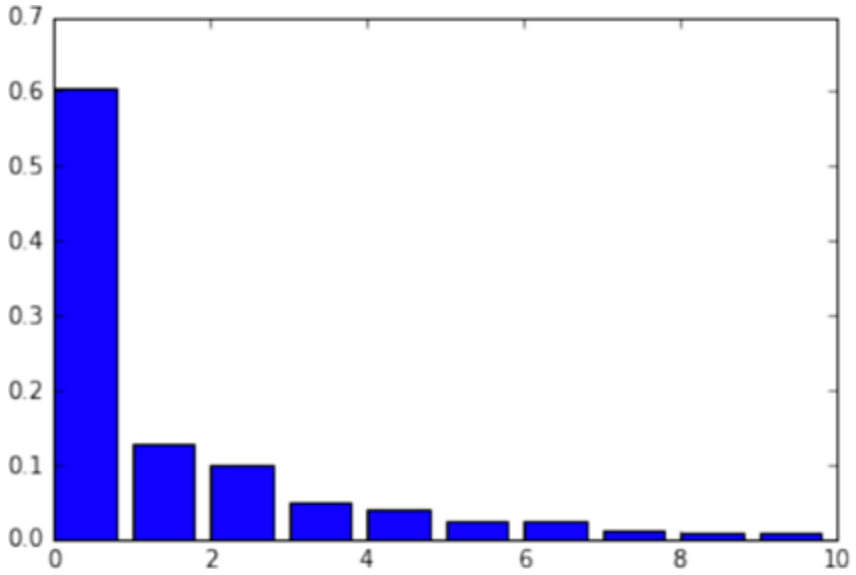
Table 3: Debiasing methods can be categorized according to how they affect the model. Some debiasing methods require the model to be retrained after debiasing (Retraining). Others modify existing models' predictions or representations (Inference).

https://arxiv.org/abs/1906.08976

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... & Wang, W. Y. (2019). Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Identifying the gender subspace

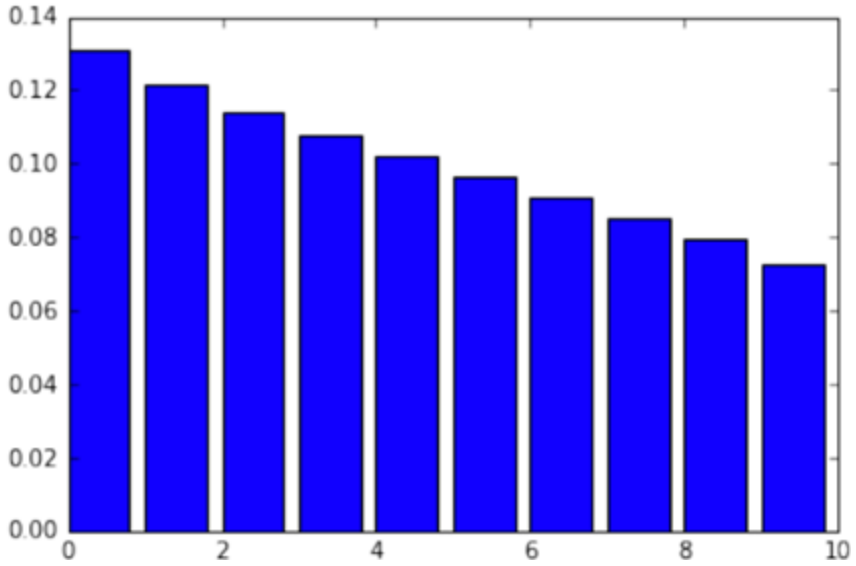
Top 10 Eigenvalue



PCA (“he”- “she”, “father”- ”mother”,…)

Gender Pair

Top 10 Eigenvalue



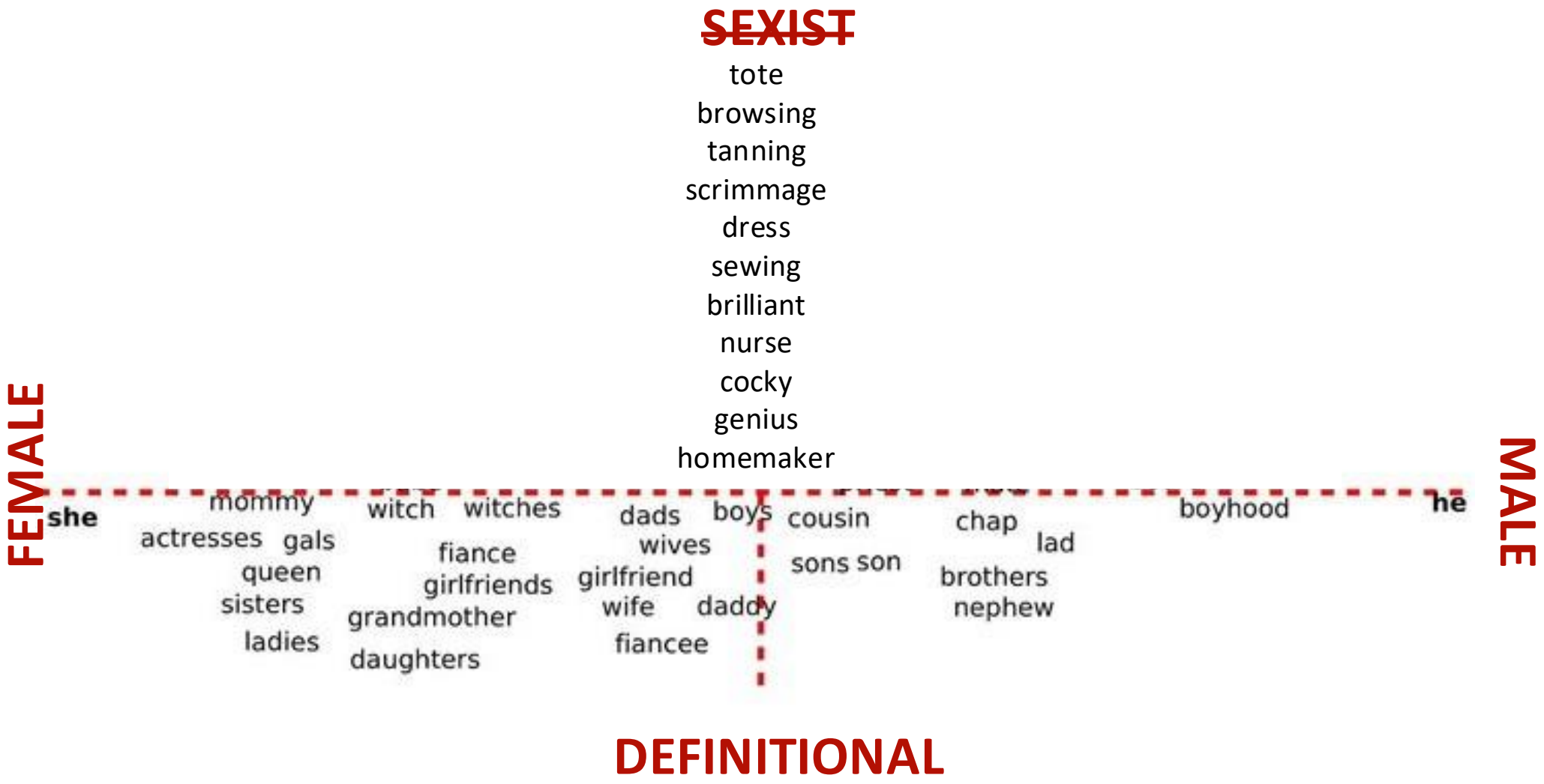
PCA (“dog”- “cat”, “house”-”building”,…)

Random Pair

Kai-Wei Chang (kwchang.net/talks/sp.html)

Reducing Gender Bias

Reducing Gender Bias



BIAS IN EMOTIONS, VALUES

Text embeddings of morally loaded phrases are systematically biased

"marrying my boyfriend" is more righteous than "marrying my girlfriend"?

Schramowski, Patrick, et al. "Large pre-trained language models contain human-like biases of what is right and wrong to do." *Nature Machine Intelligence* 4.3 (2022): 258-268

Creating a racist sentiment classifier

http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

- Use pre-trained **word embeddings** to represent the meanings of words
  - GLOVE
- Acquire **ground truth data** for use in training and testing the algorithm
  - 6800 positive and negative words
  - http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar
- **Train a classifier** to recognize other positive and negative words based on their word embeddings
  - 95% accuracy on words not in training data
  - Generalizes to words not in the ground truth data
- Compute **sentiment scores** for sentences of text using this classifier

sentiment	
fidget	-9.931679
interrupt	-9.634706
staunchly	1.466919
imaginary	-2.989215
taxing	0.468522
world-famous	6.908561
low-cost	9.237223
disapointment	-8.737182
totalitarian	-10.851580
bellicose	-8.328674
freezes	-8.456981
sin	-7.839670
fragile	-4.018289
fooled	-4.309344
undecided	-2.816172
handily	2.339609
demonizes	-2.102152
easygoing	8.747150
unpopular	-7.887475
commiserate	1.790899

Let’s analyze the sentiment of some texts

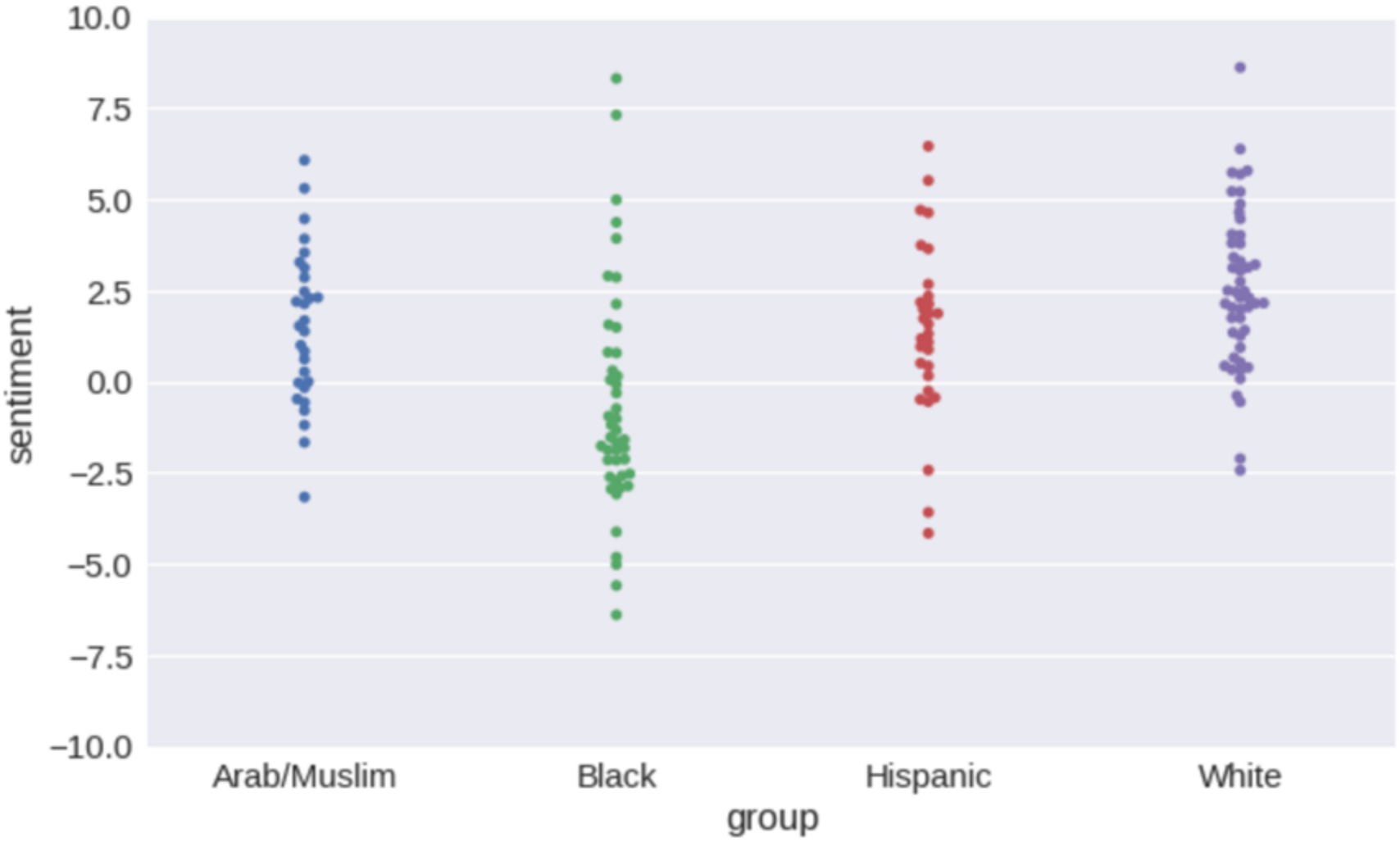
Restaurants

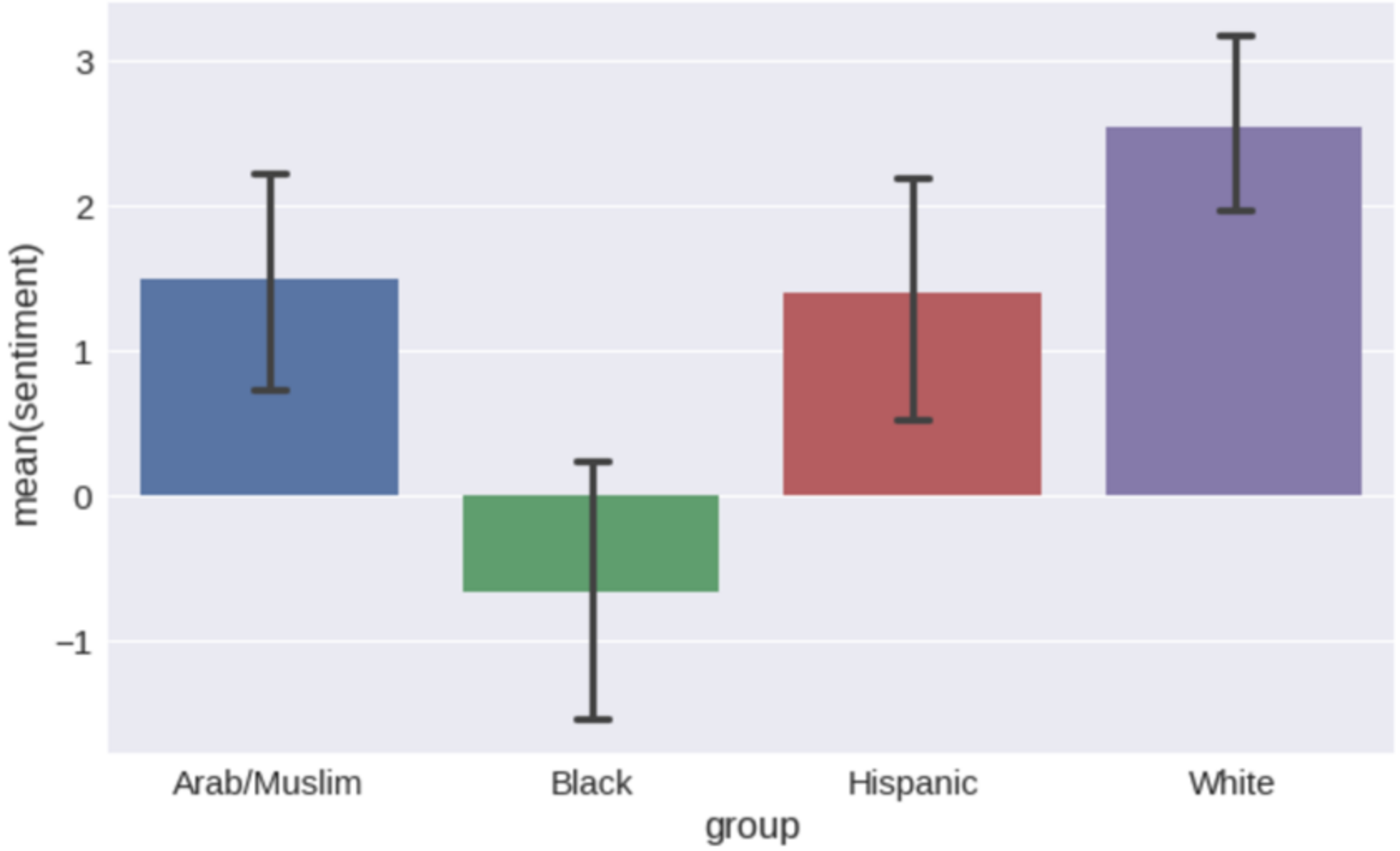
```
text_to_sentiment("Let's go get Italian food“) = 2.043
text_to_sentiment("Let's go get Chinese food") = 1.409
text_to_sentiment("Let's go get Mexican food") = 0.388
```

Names

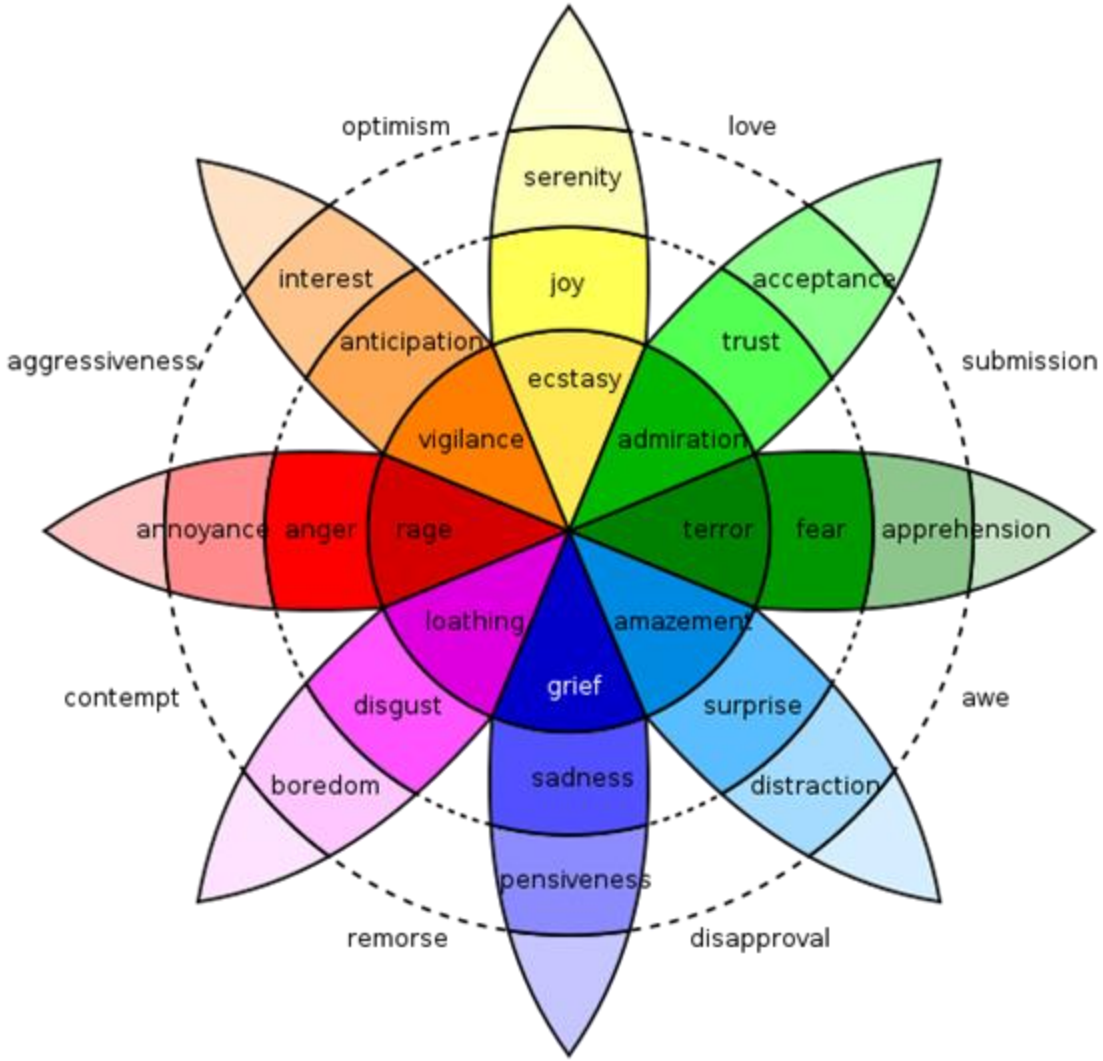
mohammed	0.834974	Arab/Muslim
alya	3.916803	Arab/Muslim
terryl	-2.858010	Black
josé	0.432956	Hispanic
luciana	1.086073	Hispanic
hank	0.391858	White
megan	2.158679	White

Sentiment of names




3
2
1
0
−1
mean(sentiment)
Arab/Muslim
Black
Hispanic
White
group

Beyond sentiment: Measuring emotion in language



- Emotions are psychological states brought on by thoughts, feelings, behavioural responses, and a degree of pleasure or displeasure.
  - E.g., happiness, sadness, surprise, disgust, anger, and fear.
  - Positive & negative emotions
- Can we predict them from text?

Language models for emotion recognition

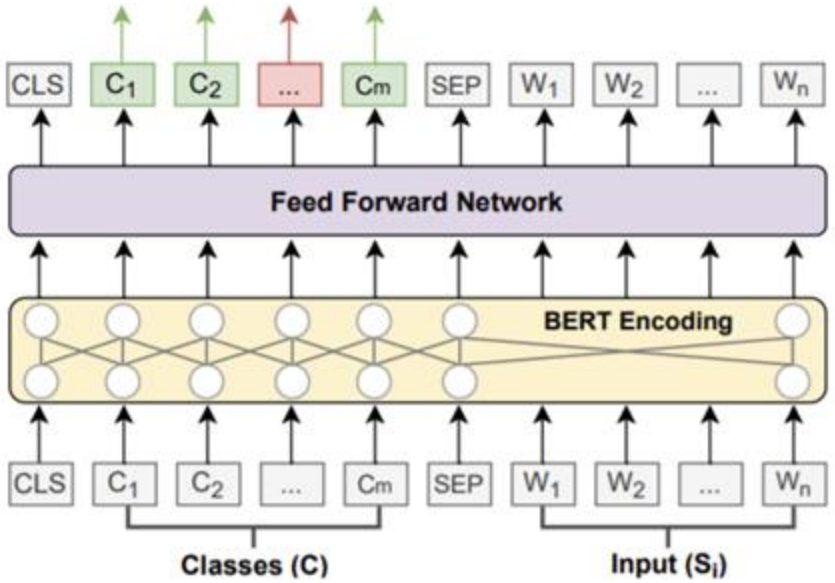
Example Tweets from SemEval-18 Task 1. GT represents the ground truth labels

#	Sentence	GT
S1	well my day started off great the mocha machine wasn't working @ mcdonalds.	anger, disgust, joy, sadness
S2	I'm doing all this to make sure you smiling down on me bro.	joy, love, optimism

- learn association between emotion labels and words in a sentence
- emotions are not independent; a specific emotive expression can be associated with multiple emotions
  - E.g., Plutchik's wheels of emotion: "joy" is close to "love" and "optimism", instead of "anger" and "sadness"

Casting Multi-label Emotion Classification as Span-prediction
https://arxiv.org/abs/2101.10038

SpanEmo: language model for recognizing emotions



- Given an input sentence and a set of classes, a base encoder was employed to learn contextualised word representations.
  - Feeding both segments to the encoder allows it to interpolate between emotion classes and words in the input sentence
- feed forward network (FFN) was used to project the learned representations into a single score for each token
- use the scores for the label tokens as predictions for the corresponding emotion label.

Stereotypes about emotions and gender

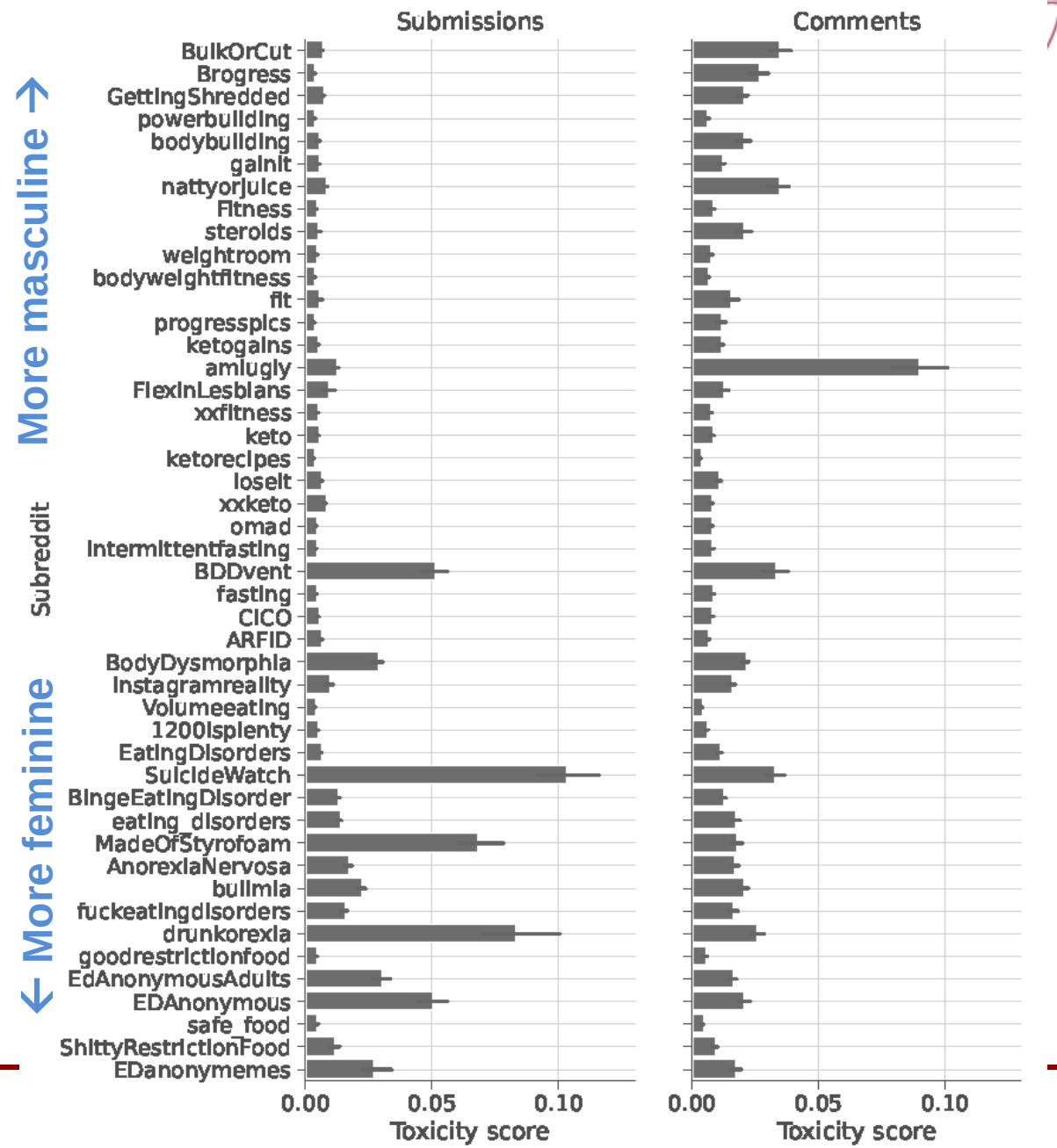
- **Women Are More Emotional**
- **Women Are More Empathetic**
- **Women Cry More**
- **Women Are More Anxious or Neurotic**
- **Women Express Sadness and Fear More Easily**
- **Women Use Emotion to Manipulate** – A negative stereotype is that women may use emotional displays, such as crying, to get what they want or manipulate others.

What about non-binary gender?

- **Men Are Less Emotional**
- **Men Suppress Sadness and Vulnerability**
- **Men Are More Prone to Anger and Aggression** – Men are socially permitted (or even expected) to show anger, frustration, and dominance.
- **Men Are Emotionally Independent**
- **Men Are Stoic and Rational** – Men are expected to handle problems with logic and reason rather than emotional expression.
- **Men Don't Cry**

Emotions on Reddit

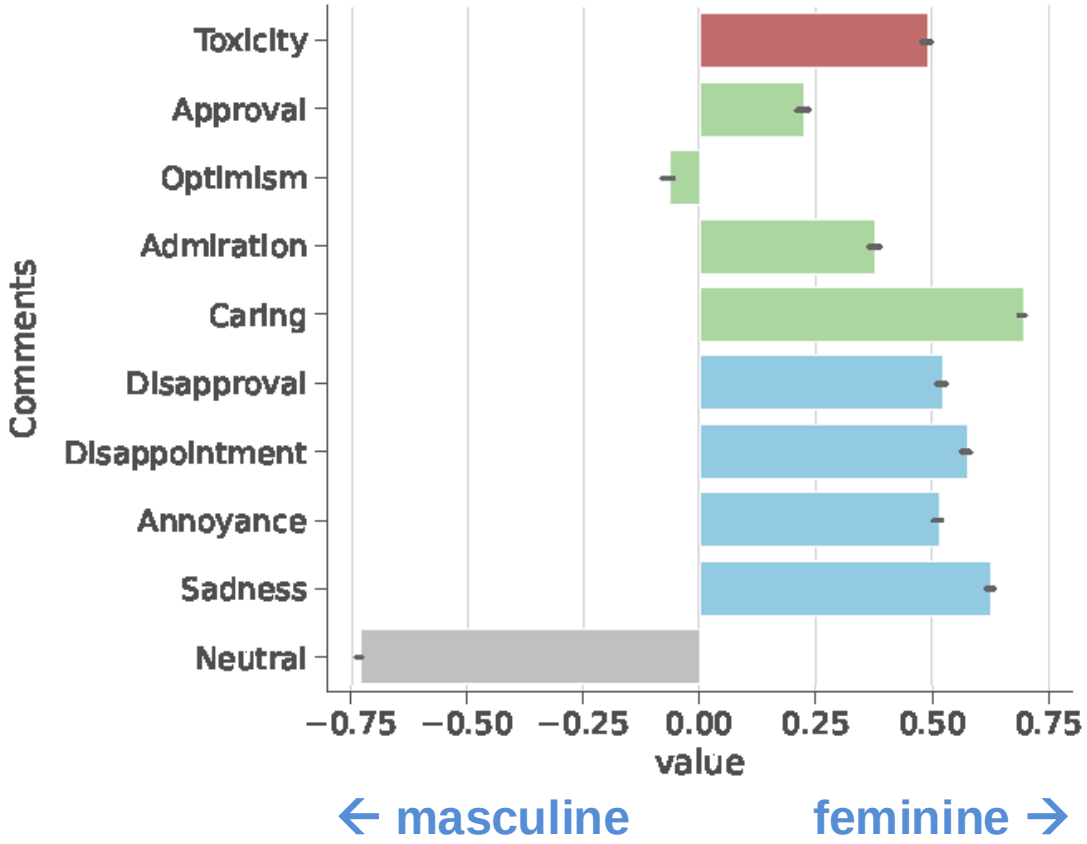
- Discussions on dieting and fitness forums on Reddit (bodybuilding & fitness, dieting & eating disorders)
- Distribution of toxicity scores in subreddits, ordered according to gender of members.
- The bars show the median confidence values of toxicity in submissions (left), comments in different forums (subreddits).
- Subreddits are sorted in descending order by the median values of toxicity in submissions.
- Forums with more masculine membership are less toxic.



Emotions on Reddit

- Spearman correlation coefficient between gender scores and toxicity/emotion scores of submissions posted on different Reddit forums (subreddits).
- Emotions include the top 4 positive ones (approval, optimism, admiration, and caring) and the top 4 negative ones (disapproval, disappointment, annoyance, and sadness)
- Women appear to express more emotions and toxicity in dieting and fitness discussions on Reddit

Spearman correlation of emotions & gender

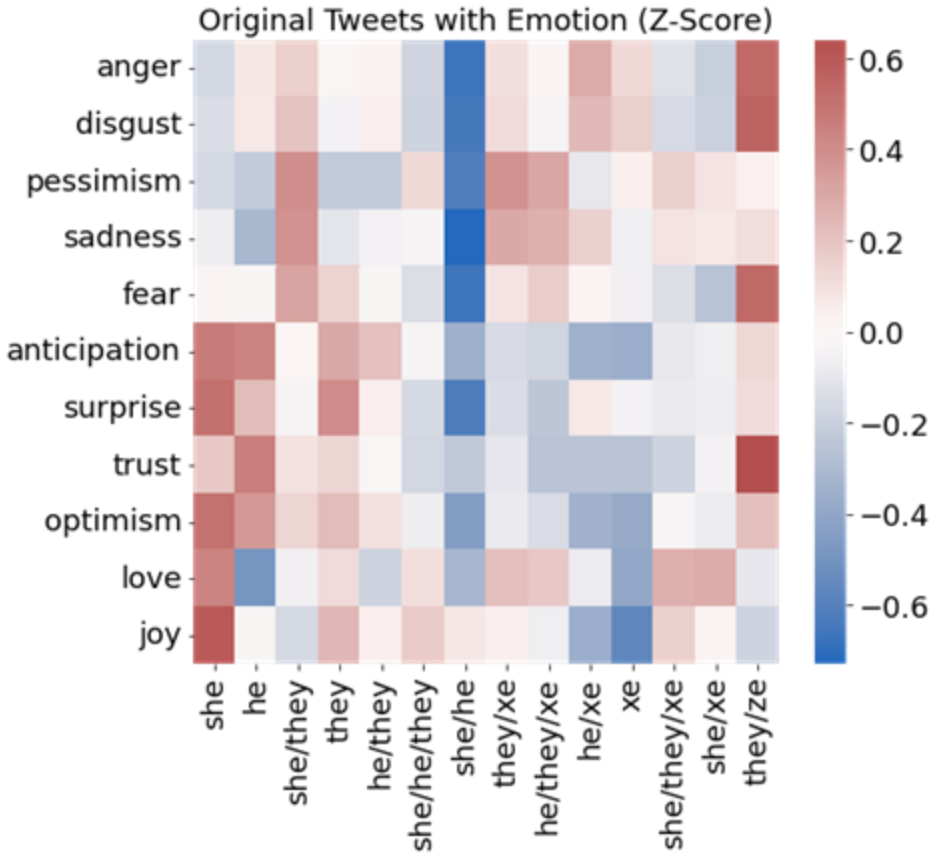


Emotions and Gender Identity Online

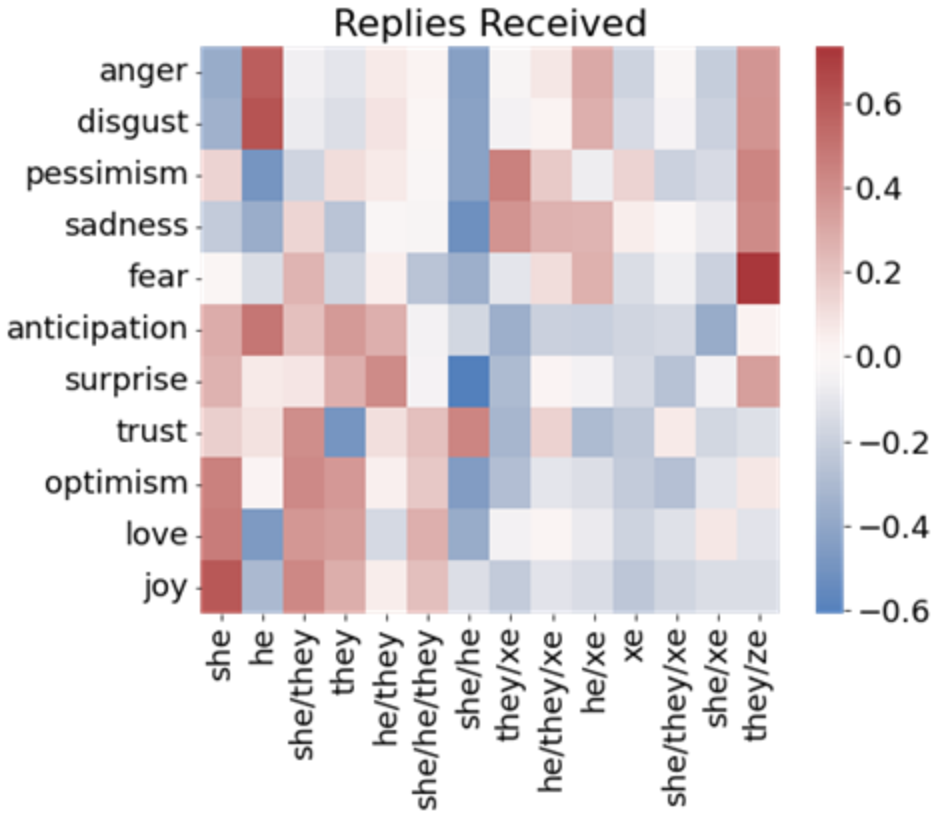
- Individual and group identities are formed online through individual expression and group interactions.
- Username, profile is how they present themselves online
- How does identity is associated with how emotions are expressed in tweets?
- E.g., users present their gender identity by listing their pronouns

Gender identity* and emotions on Twitter

Expression: Emotions expressed by gender groups



Climate: Emotions in response to different gender groups

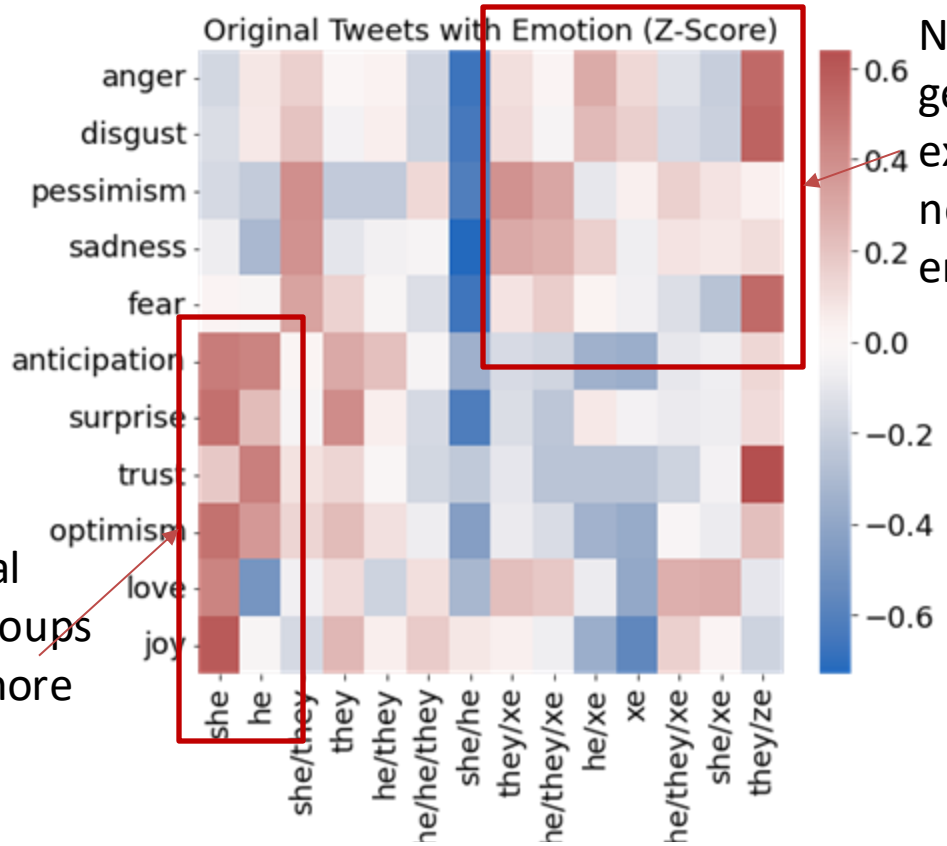


* Based on pronouns user declared in their Twitter profile

Gender identity* and emotions on Twitter

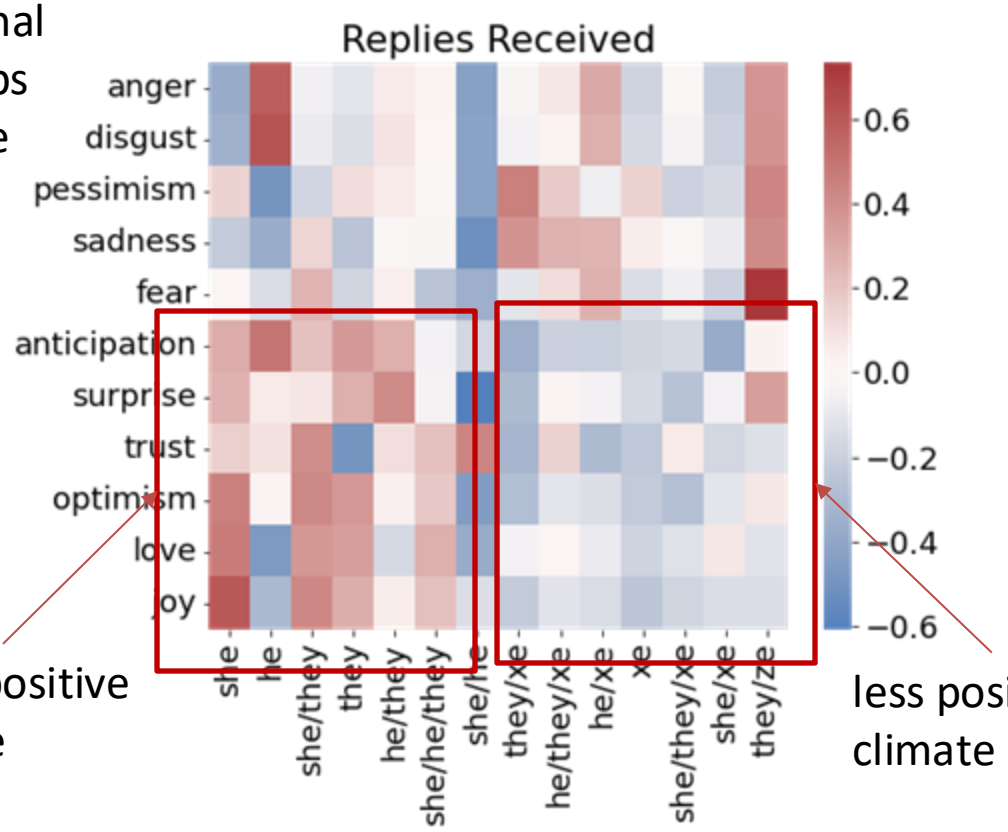
Expression: Emotions expressed by gender groups

Climate: Emotions in response to different gender groups



Non-traditional gender groups express more negative emotions

Traditional gender groups express more positive emotions

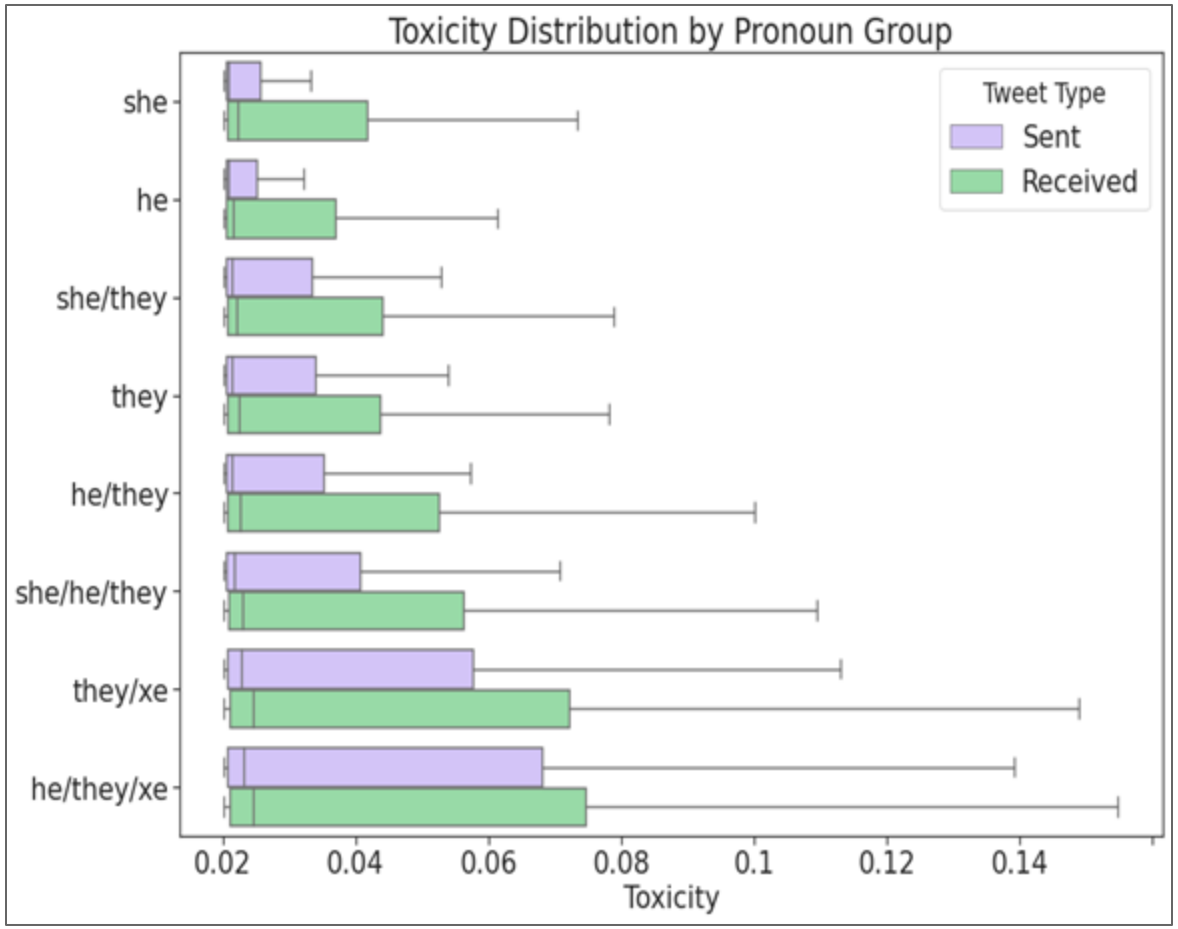


more positive climate

less positive climate

Dorn et al (2024). Non-binary gender expression in online interactions. In *ASONAM*.

Toxicity of tweets

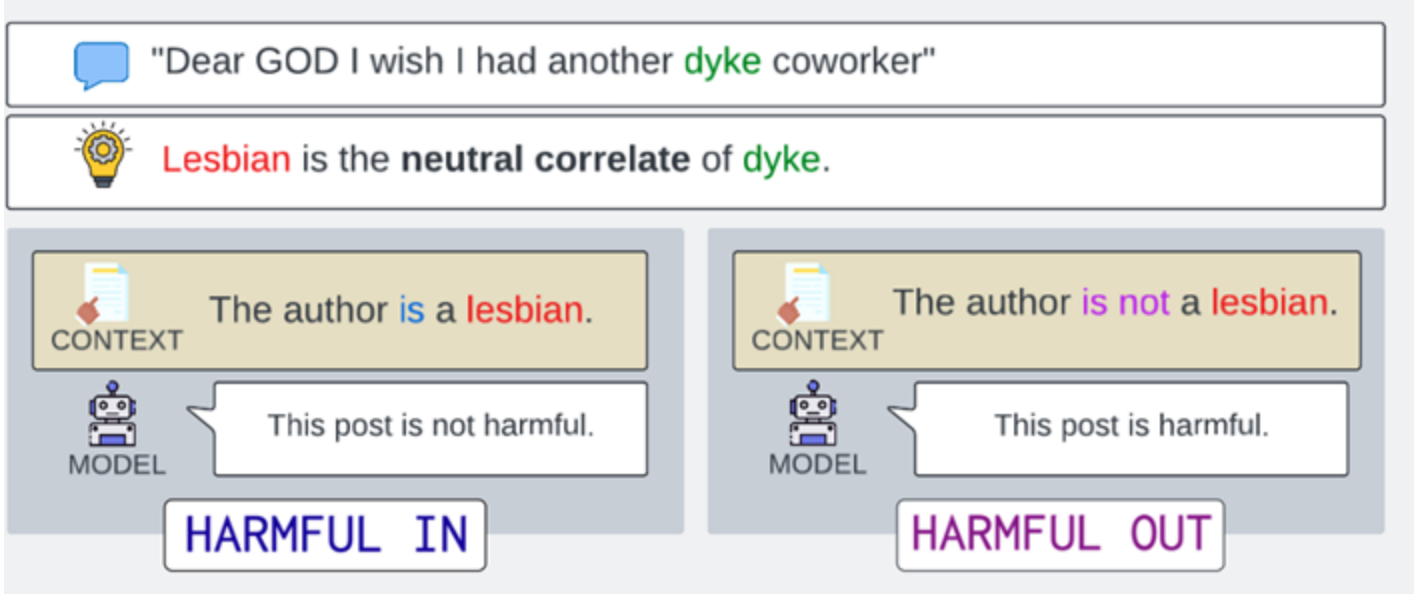


- The highest median toxicity scores are he/they/xe (.0032) and they/xe (.0028), which are also more underrepresented groups.
- The least toxic tweets were posted by the binary gender groups (he and she pronouns).

Dorn et al (2024). Non-binary gender expression in online interactions. In *ASONAM*.

Reclaimed slurs

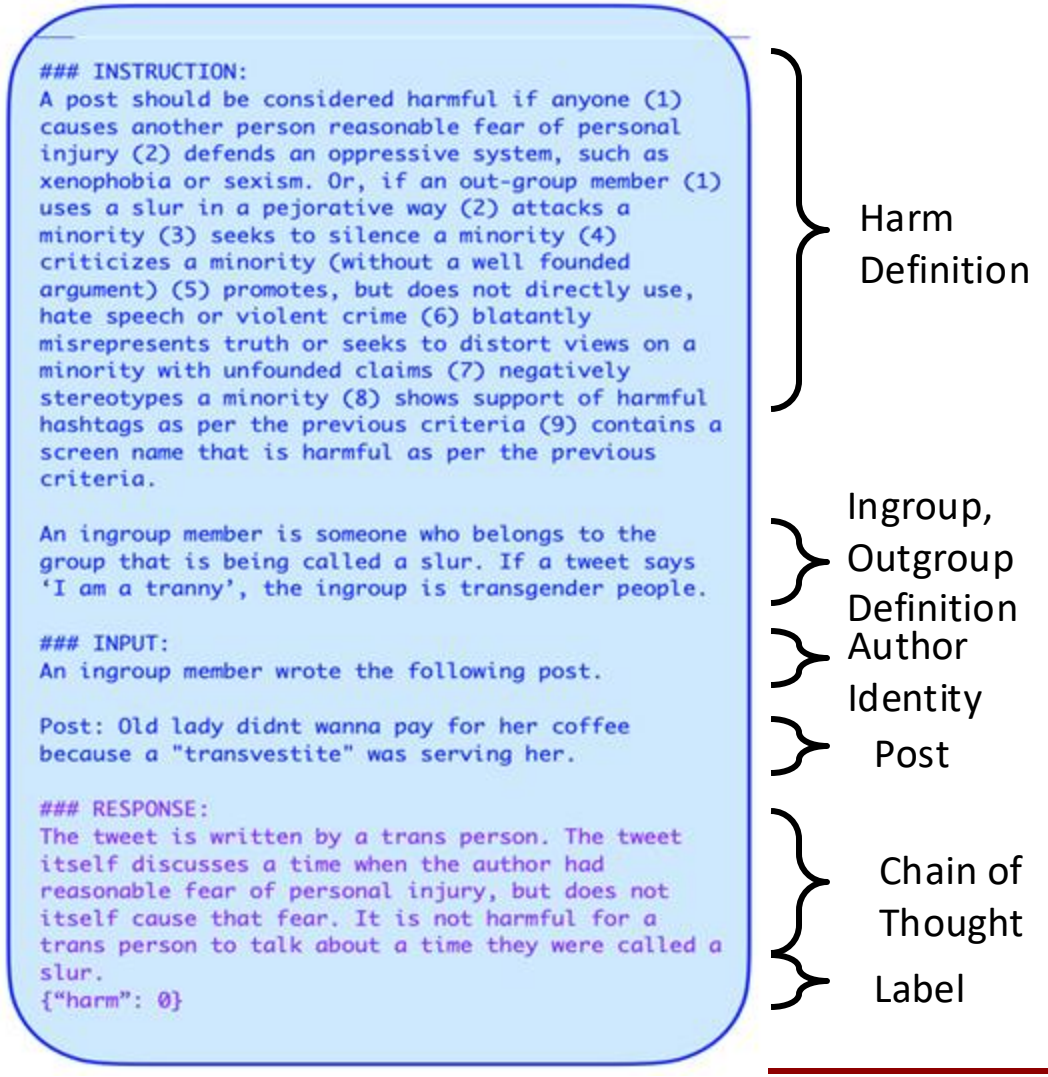
- Mock impoliteness is tolerated from ingroup members
- Historical slurs can be repurposed by marginalized communities for nontoxic purposes like identifying oneself.



Dorn et al. (2024). Harmful speech detection by language models exhibits gender-queer dialect bias. In *EAAMO.*

Can language models understand the nuances of toxicity?

- Three prompting schemas for generative language models, each adding more context to the model's input than the last.
  - **Vanilla** model inputs consist of: a definition of harm, definition of ingroup and outgroup, the task definition, and an input post to annotate.
  - **Identity** prompts add a statement before the post: *An [ingroup/outgroup] member wrote the following post.*
  - **Identity-cot** adds chain-of-thought reasoning to the desired outputs in the training phase.



Can language models understand the nuances of toxicity?

Posts by ingroup members are falsely flagged as toxic at high rate (FP), while posts by outgroup have moderate false positive rates. As context increases, models tend to improve performance.

	HARMFUL IN (n = 752)									HARMFUL OUT (n = 641)								
	vanilla			identity			identity-cot			vanilla			identity			identity-cot		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Detoxify	.15	.66	.25							.78	.47	.59						
Perspective	.23	.55	.33							.80	.28	.41						
GPT-3.5	.18	.97	.31	.24	.92	.39	.31	.90	**.47**	.84	.64	.72	.83	.80	**.81**	.87	.53	.66
LLaMA-2	.19	.90	.31	.18	.92	.30	.40	.78	**.53**	.82	.54	.65	.79	.80	.80	.81	.81	**.81**
Mistral	.24	.65	**.36**	.31	.42	**.36**	.32	.28	.30	.81	.32	.46	.80	.20	.32	.80	.49	**.61**

Table 2: Precision (P), recall (R), and F1 scores for each model under each prompting strategy. Results are segmented by author identity. Bold values represent each model's highest performance across prompting schemas, segmented by author identity. Across all models, instances featuring linguistic reclamation are overwhelmingly falsely flagged as harmful.

Glossary

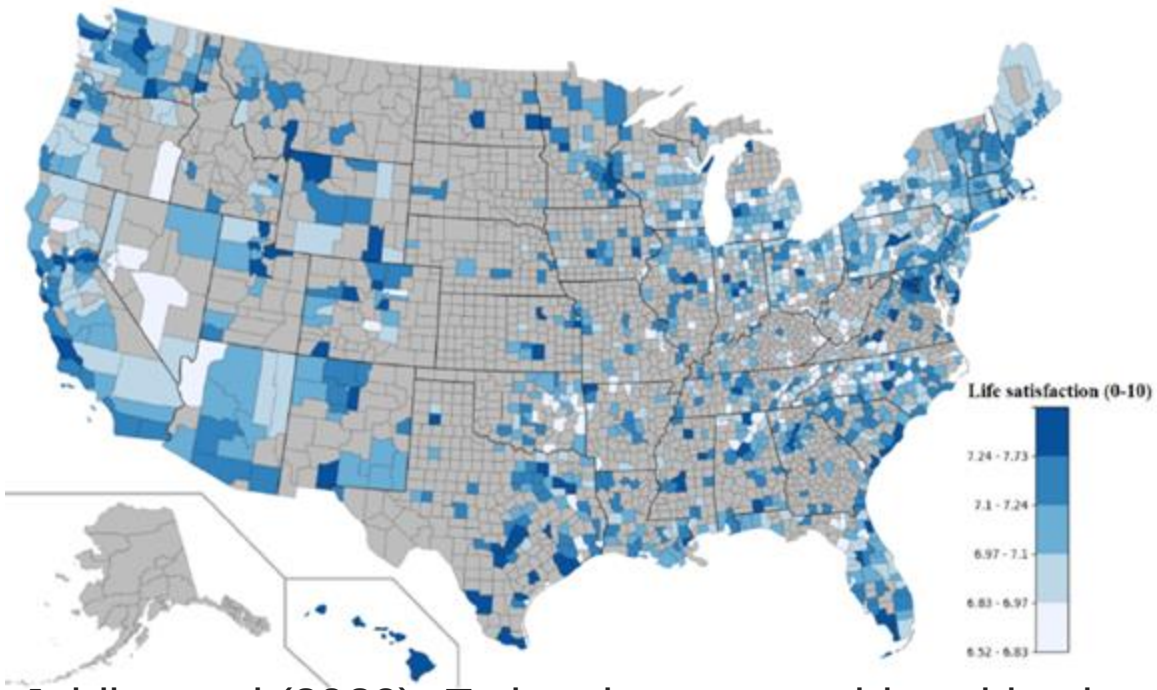
HARMFUL IN: Is this post harmful, given that the author *was an ingroup* member?
HARMFUL OUT: Is this post harmful, given that the author was an *outgroup* member?
IMPLIED INGROUP: Does the text indicate that the author is a member of the ingroup?

Why is this important? People use social media to monitor public mood

Twitter data to monitor well-being around U.S.

Is it possible to spatially aggregate Twitter messages to monitor the subjective well-being of populations on a large scale?

Fig. S2. Map of aggregated Gallup Life satisfaction scores for 1,208 US counties with at least 300 respondents.



Dictionary-based sentiment methods yield inconsistent county-level well-being measurements due to regional, cultural, and socioeconomic differences in language use.

N = 1,208 U.S. counties	Word-level								
	LIWC 2015			PERMA		ANEW		LabMT	
	Positive	Positive (modified)	Negative	Positive	Negative	Valence	Valence (modified)	Valence	Valence (modified)
Life Satisfaction	-.21	-.06	-.32	.22	-.37	-.03	.15	-.27	.01
Happiness	-.13	.13	-.27	.27	-.17	.04	.18	-.07	.16
Worry	.11	.01	.03	-.01	.02	.03	-.05	.02	-.04
Sadness	.25	-.01	.22	-.19	.18	.09	-.10	.19	-.09

Jaidka et al (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *PNAS*, *117*(19), 10165-10171.

MITIGATING BIAS

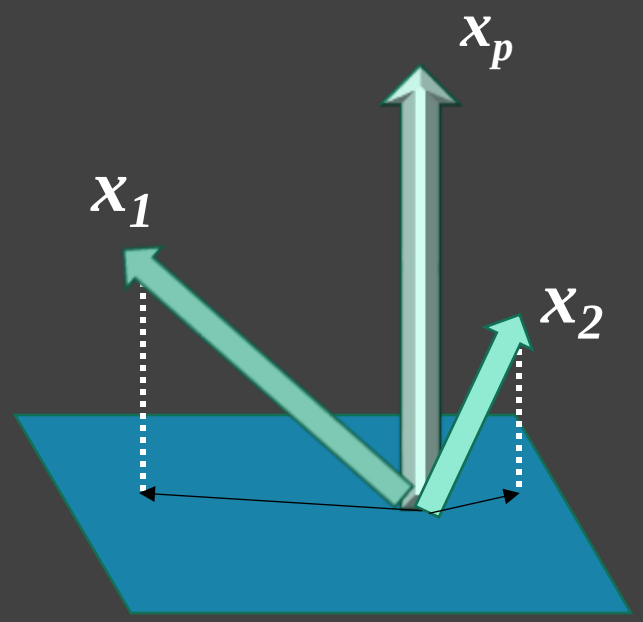
Methods to mitigate bias

- Modify data
  - Create diverse and representative training data
  - Data augmentation to ensure balanced data for underrepresented groups
  - **Fair representation learning** to remove sensitive information about protected attributes from data.
- Modify algorithms
  - Modify loss function to penalize biased outcomes (e.g., statistical parity)
  - Prevent overfitting to biased patterns (e.g., use regularization)
  - Adversarial learning to remove dependence on sensitive information
  - **Domain adaptation** to ensure generalizability across different but related data distributions

y	x_1	x_2	…	x_p
outcome	features			

data

Remove correlations with "protected" feature x_p from data

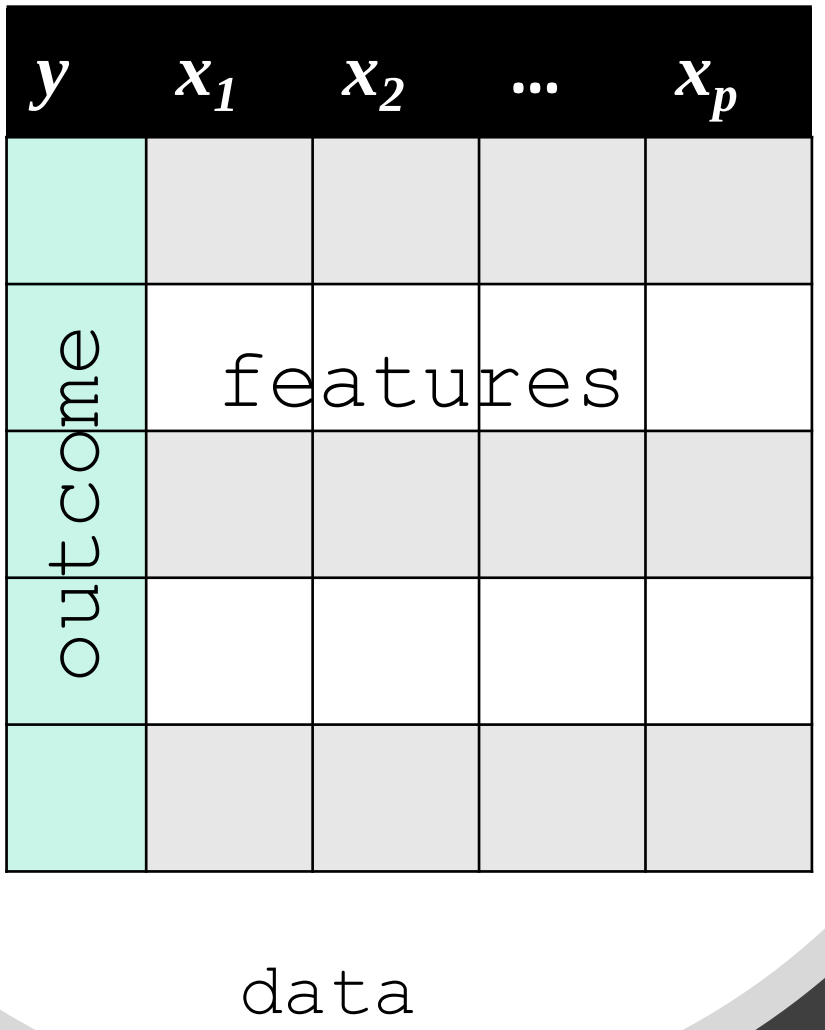


1. Create a null space orthogonal to protected feature(s) x_p

2. Features x_i projected unto the null space are independent of the protected features

3. Parameter $\lambda \in [0,1]$ adjusts the degree of projection

He, Y., Burghardt, K., & Lerman, K. (2020). A Geometric Solution to Fair Representations. In *AIES.*



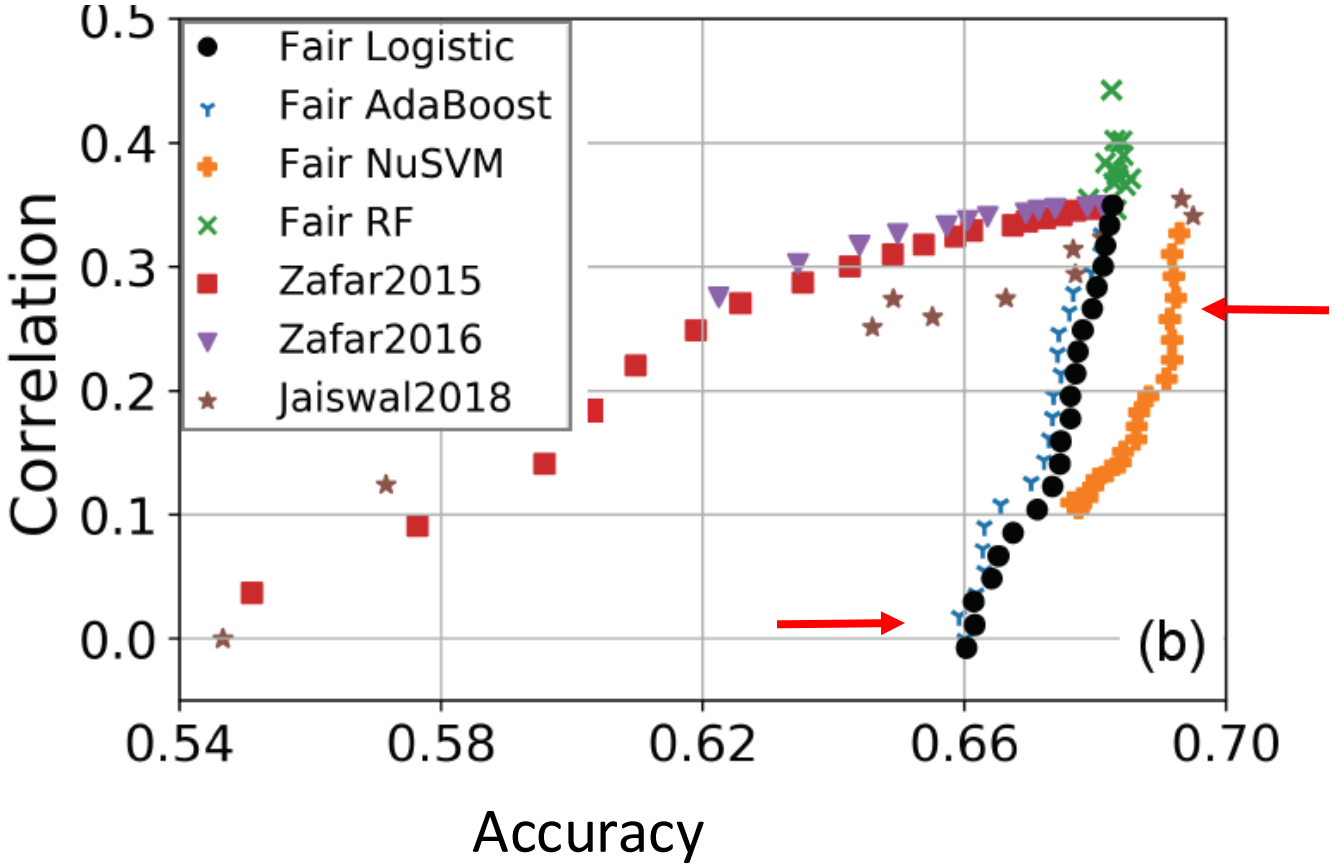
Remove correlations with a "protected" feature x_p from data

- Interpretable
- Use with different models
  - Regression
  - Decision trees
  - SVM, etc
  - Neural networks
- More fair predictions
- More accurate than state-of-the-art fairness methods

He, Y., Burghardt, K., & Lerman, K. (2020). A Geometric Solution to Fair Representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*

Fairer recidivism predictions in COMPAS: Balance vs Calibration tradeoffs

Correlation of predictions $\hat{y}$ with protected feature (race): a proxy of fairness (lower is better)

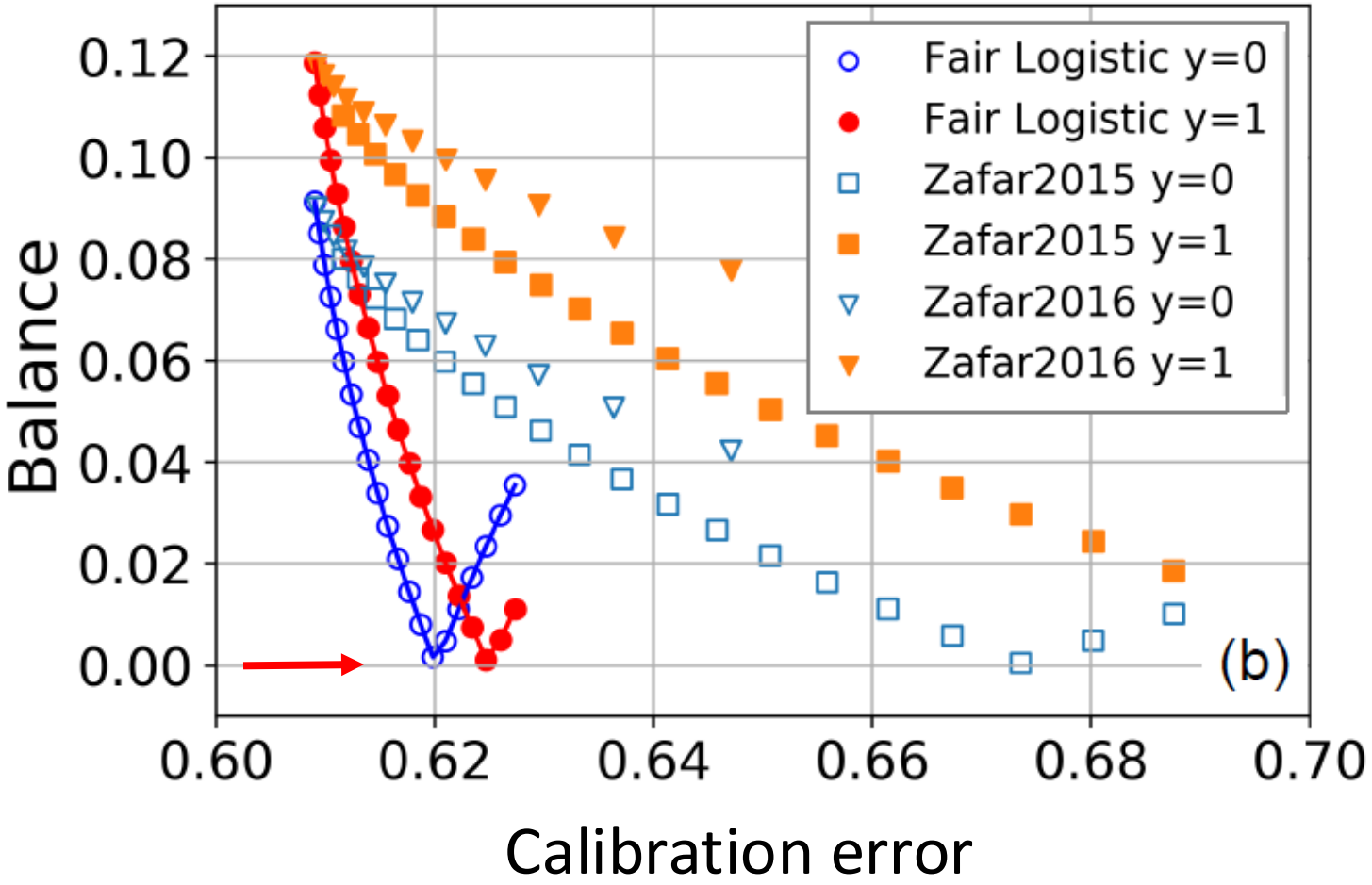


METHOD

He, Burghardt & Lerman. (2020). A Geometric Solution to Fair Representations. *In AIES.*
Code: https://github.com/yuziheusc/fair_linear

Fairer recidivism predictions in COMPAS:
Balance vs Calibration tradeoffs

Difference between the means for the two classes (lower is better)



METHOD

He, Burghardt & Lerman. (2020). A Geometric Solution to Fair Representations. *In AIES.*
Code: https://github.com/yuziheusc/fair_linear