Fairness and Ethics in NLP

Elements and images borrowed from Kai-Wei Chang, Vinod Prabhakaran

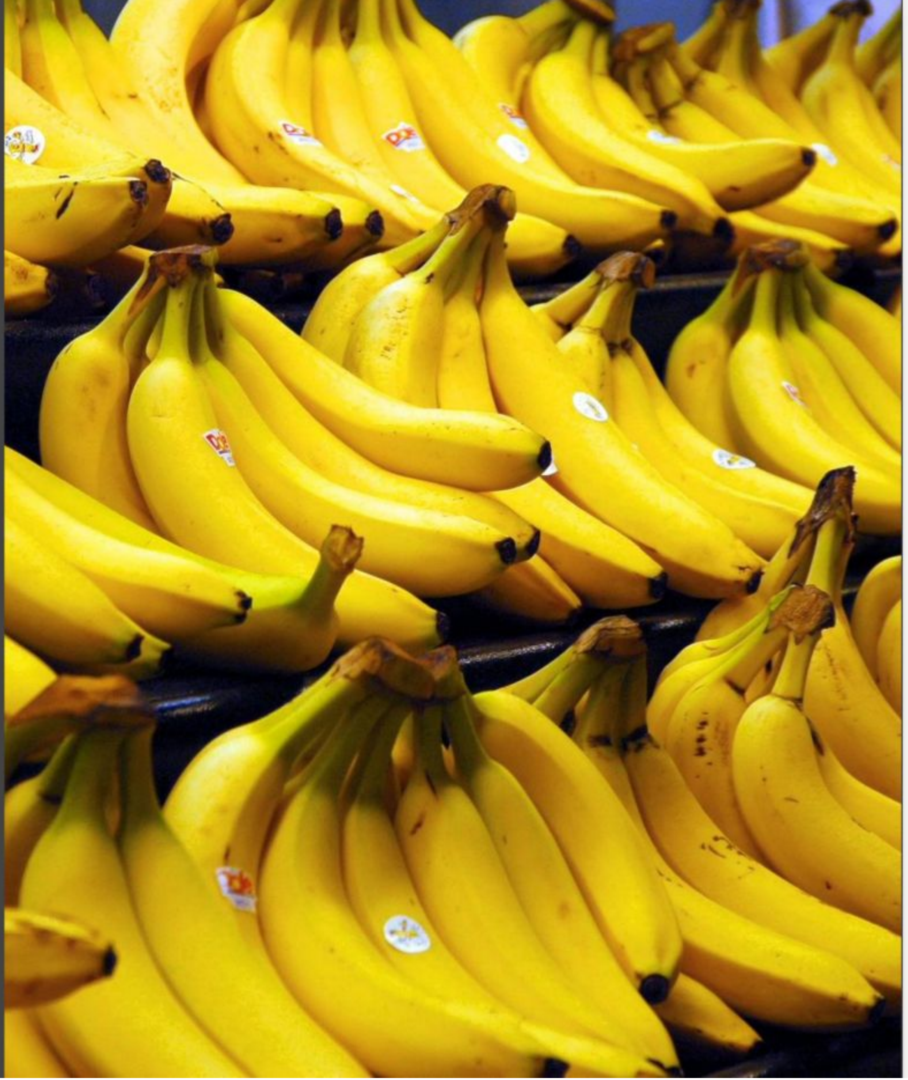
What do you see?

What do you see?

- Bananas

What do you see?

- Bananas
- Stickers



What do you see?

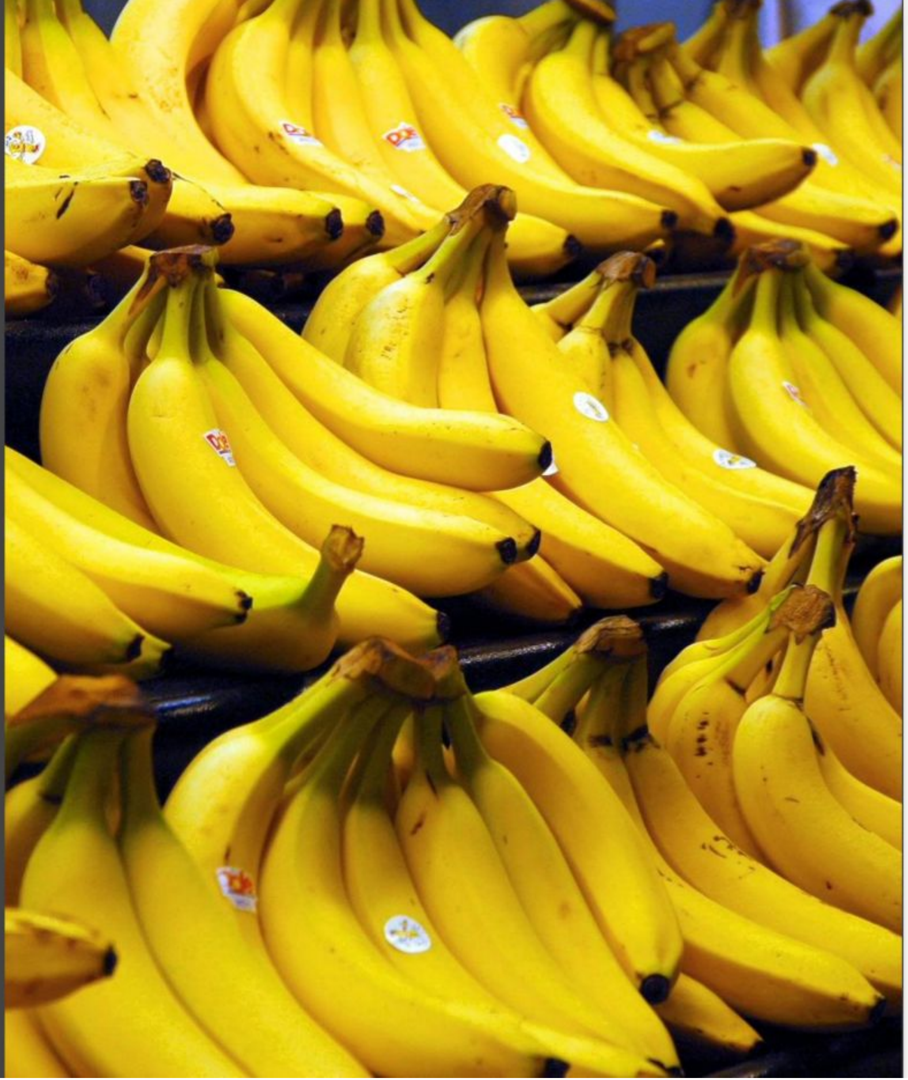
- Bananas
- Stickers
- Dole Bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves



What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

...We don’t tend to say

Yellow Bananas

What do you see?

Green Bananas

Unripe Bananas

What do you see?

Ripe Bananas

Bananas with spots

What do you see?

Yellow Bananas

Yellow is prototypical for bananas

Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)

Fruit

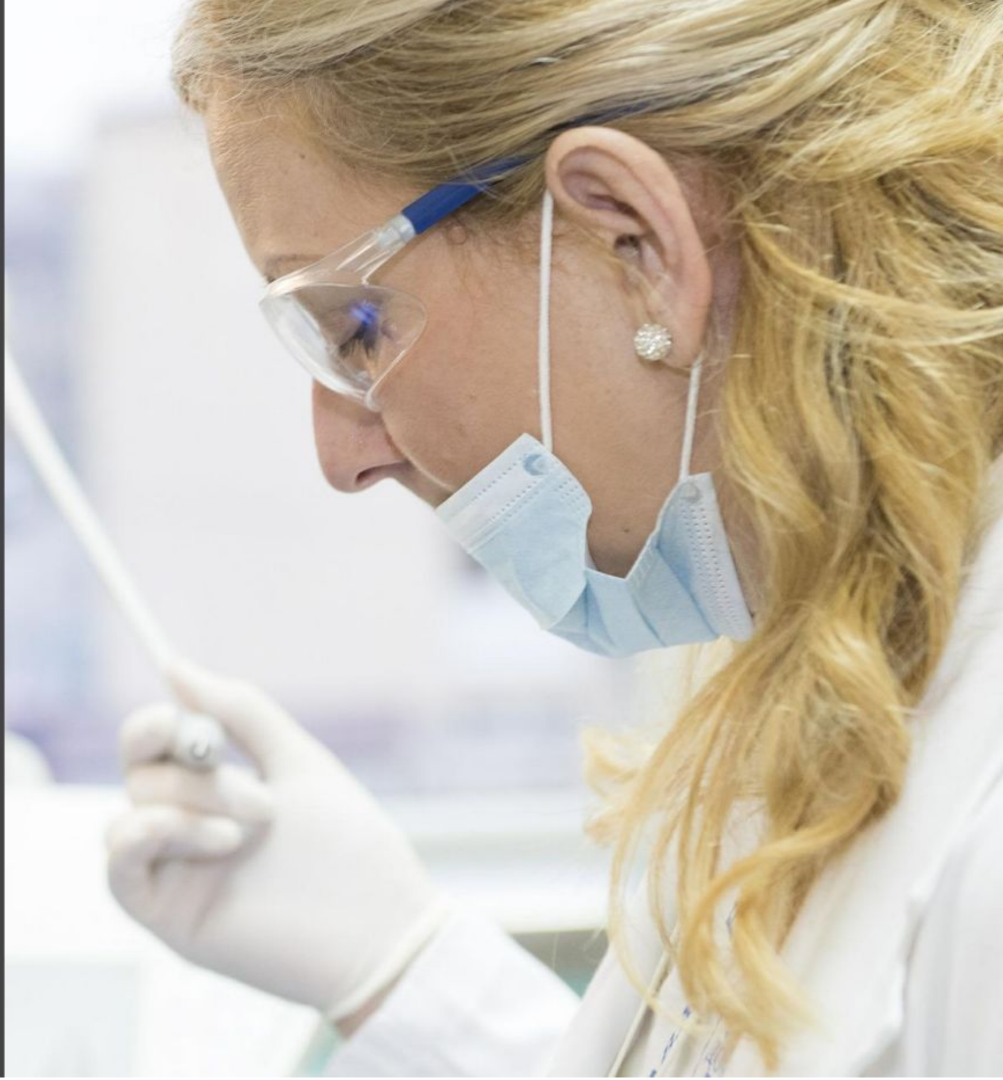
Bananas
"Basic Level"

Unripe Bananas, Cavendish Bananas

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

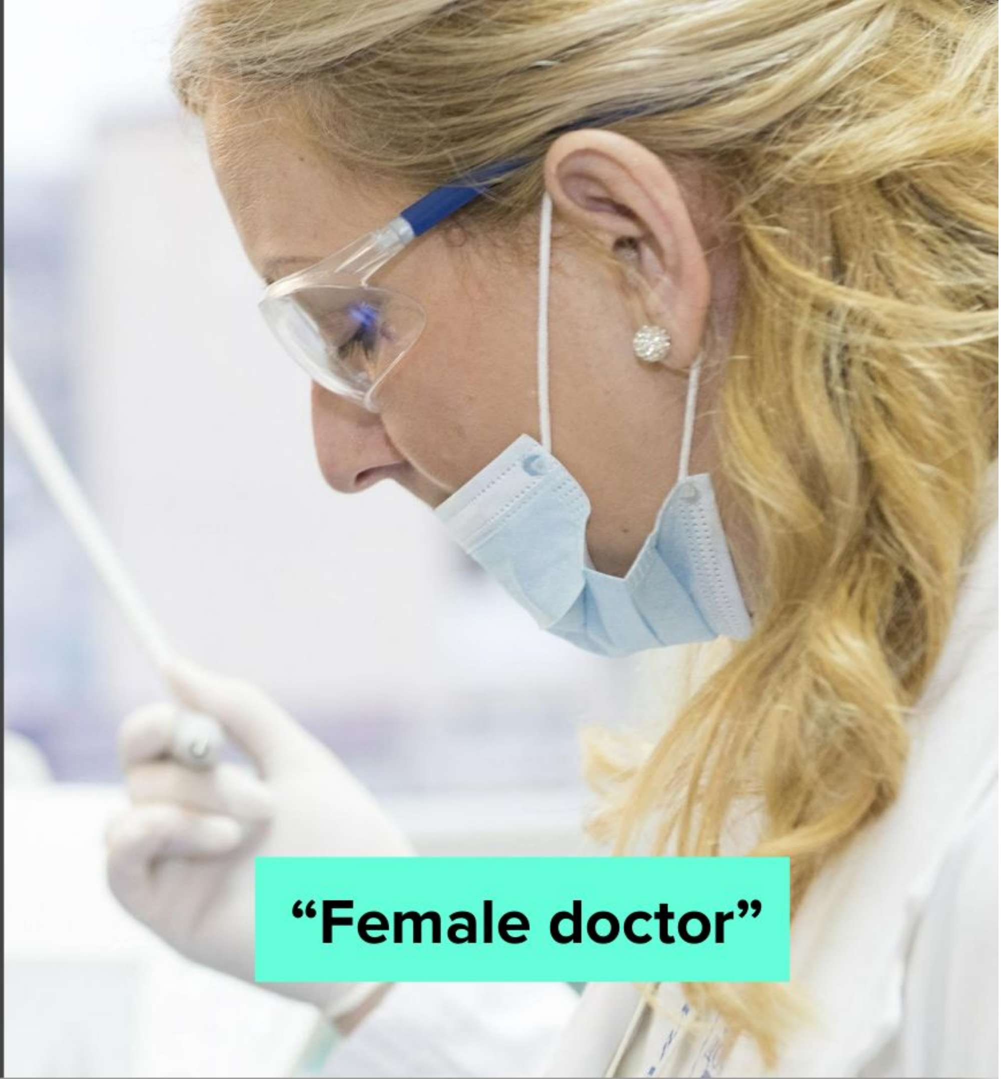
How could this be?

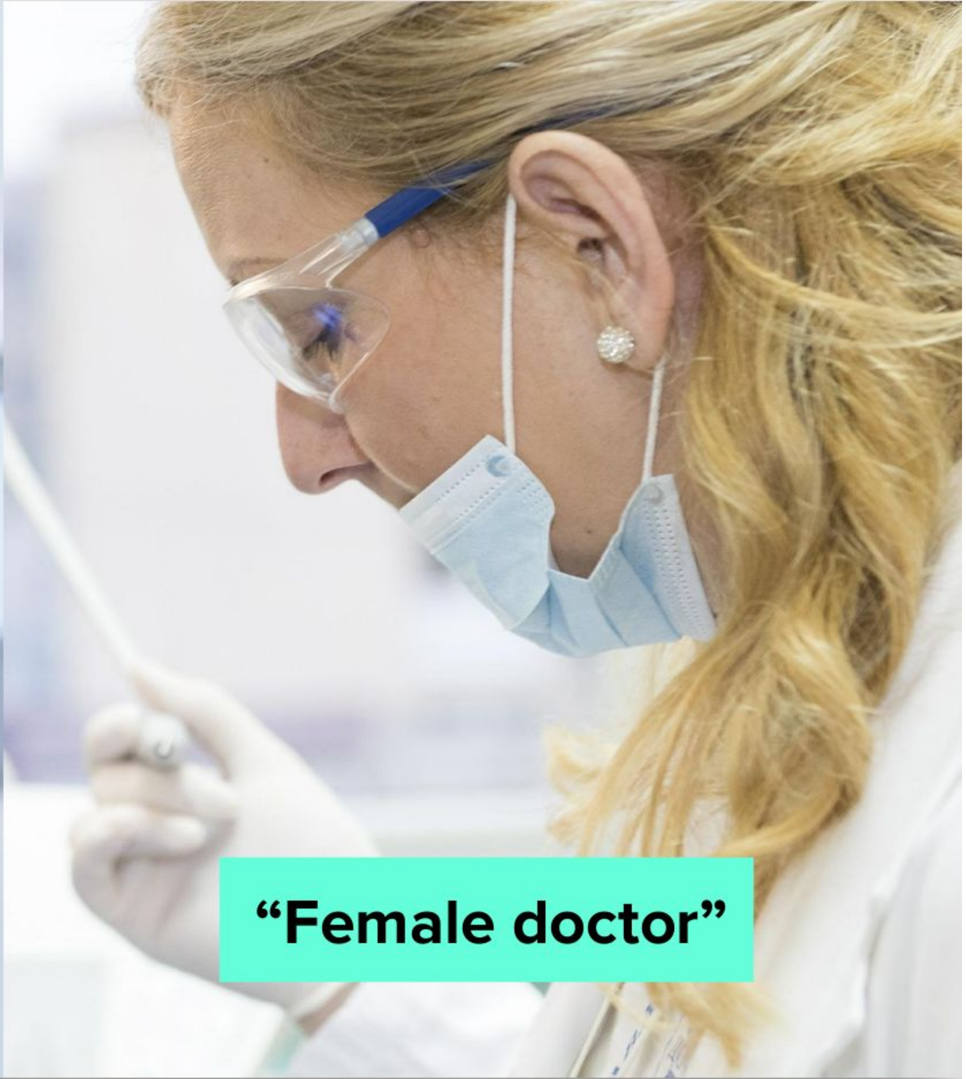


A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?




“Doctor”

“Female doctor”

Why do we intuitively recognize a default social group?

Implicit Bias

Biases in Data

Selection Bias: Selection does not reflect a random sample

- Men are over-represented in web-based news articles

  (Jia, Lansdall-Welfare, and Cristianini 2015)

- Men are over-represented in twitter conversations

  (Garcia, Weber, and Garimella 2014)

- Gender bias in Wikipedia and Britannica

  (Reagle & Rhuee 2011)

Biases in Data

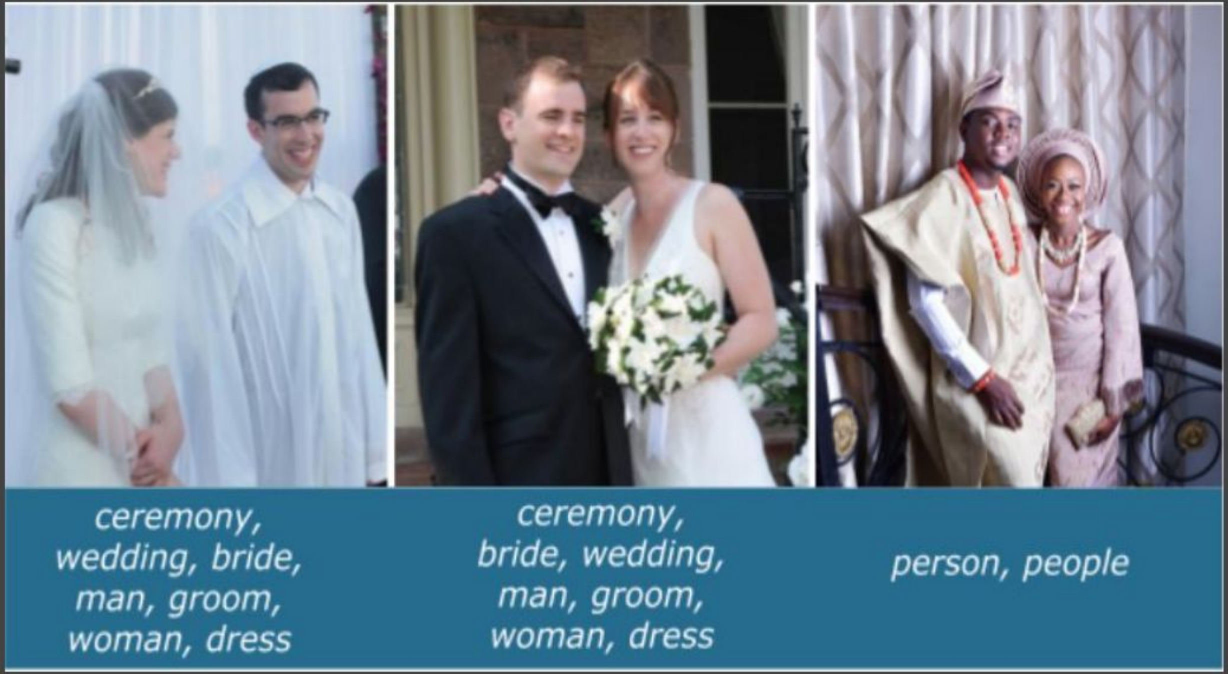
Selection Bias: Selection does not reflect a random sample

Map of Amazon Mechanical Turk Workers

CREDIT

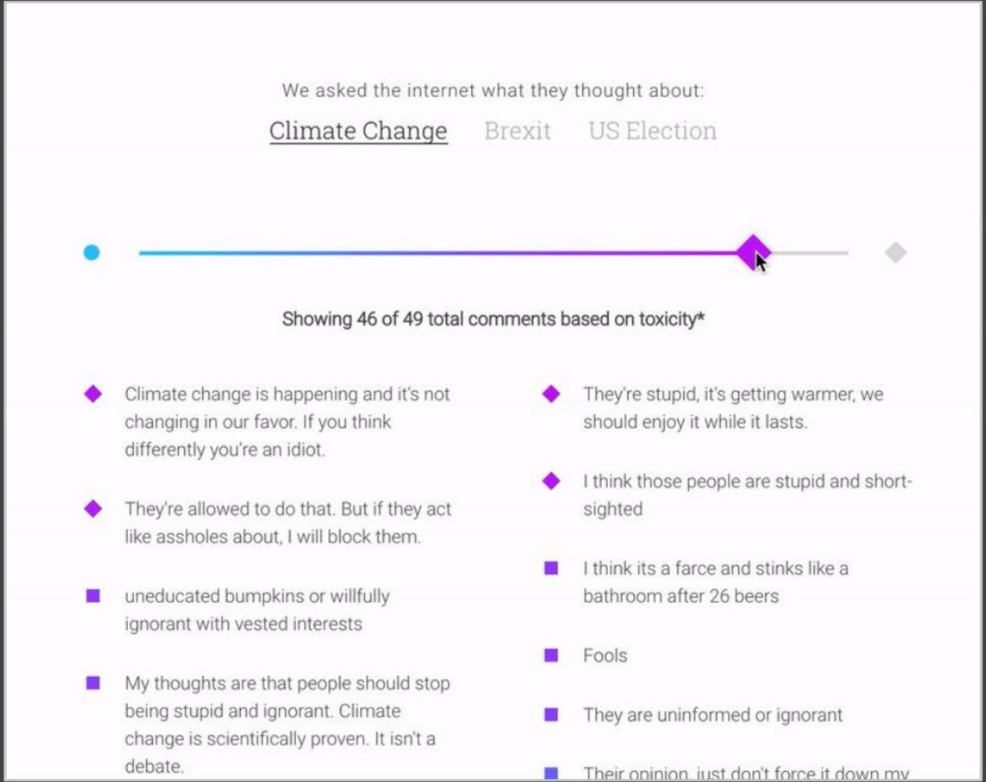
Biases in Data → Biased Labels

Annotations in your dataset will reflect the worldviews of your annotators.



Consequence: models are biased

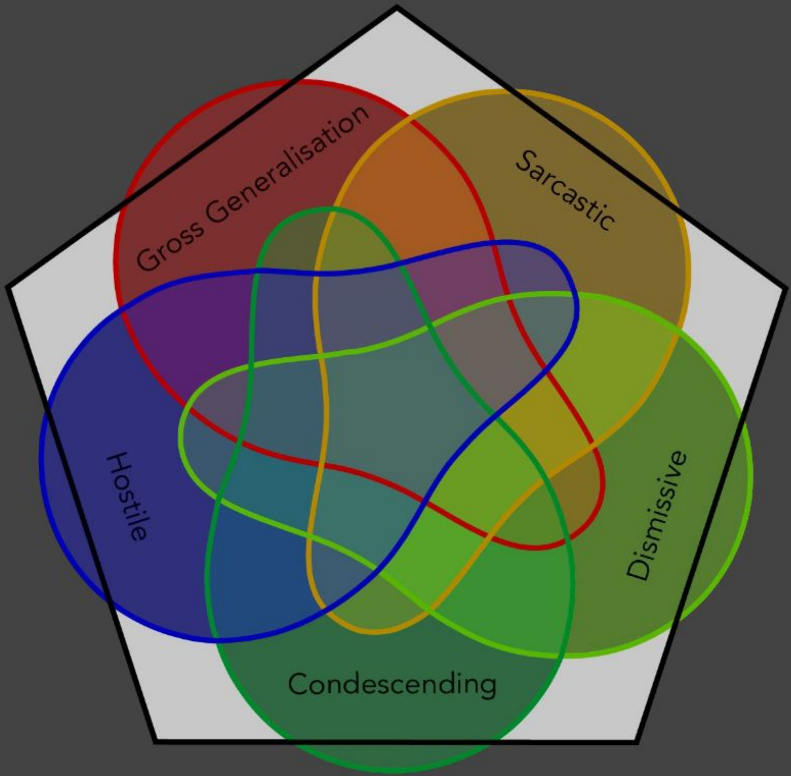
Toxicity Classification



Source
perspectiveapi.com

Toxicity Classification

Toxicity is defined as... "***a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.***"



Source
"The Challenge of Identifying Subtle Forms of Toxicity Online" - Jigsaw
https://medium.com/the-false-positive/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

Toxicity Classification

Unintended biases towards **certain identity terms**:

Comment	Toxicity Score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam	0.46

- “The Challenge of Identifying Subtle Forms of Toxicity Online”. Jigsaw. The False Positive (2018).

Toxicity Classification

Unintended biases towards **named entities**:

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

- Prabhakaran et al. (2019). "Perturbation Sensitivity Analysis to Detect Unintended Model Biases" EMNLP 2019

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	**Toxicity Score**
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

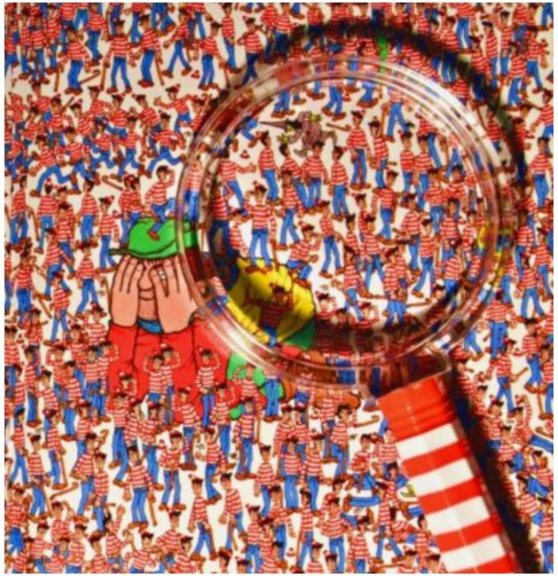
Toxicity Classification

Unintended biases towards **mentions of disabilities**:

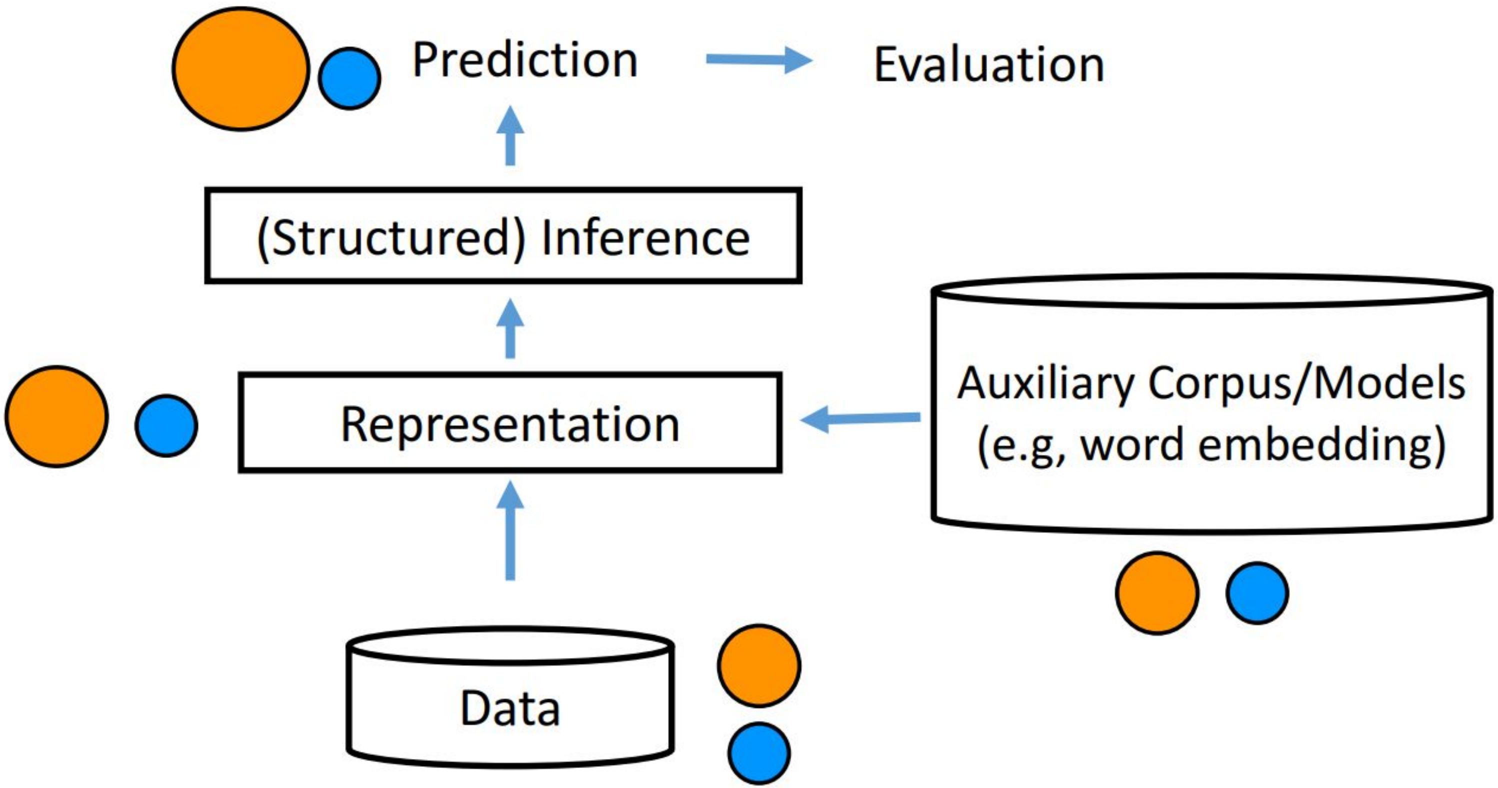
Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

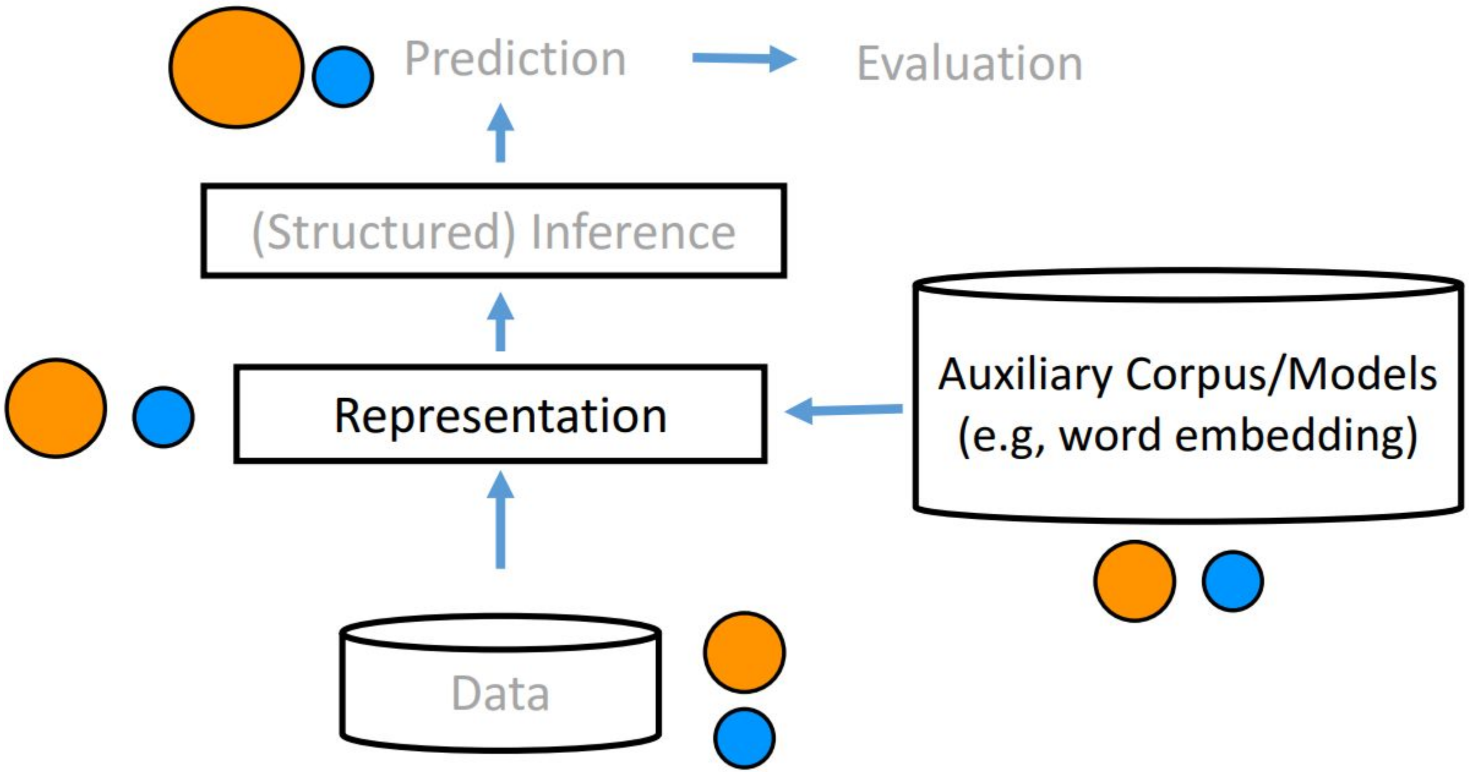
Where’s Biases?



A carton of ML (NLP) pipeline

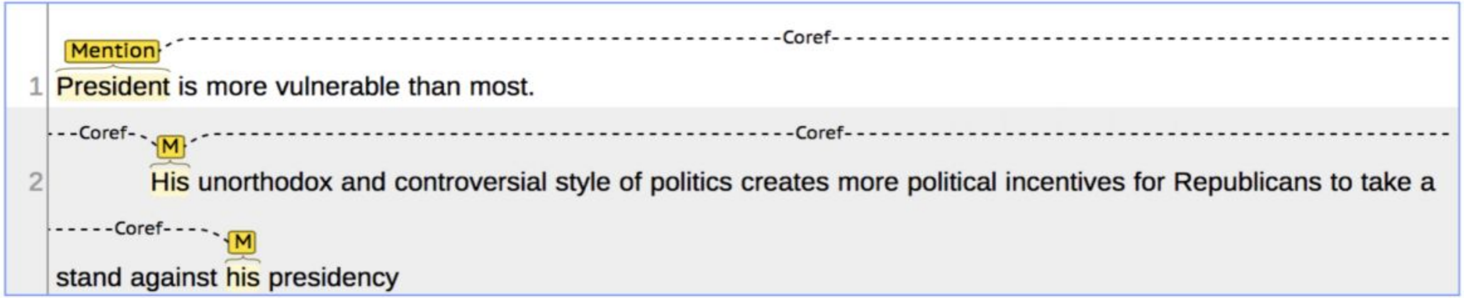


A carton of ML (NLP) pipeline



Motivate Example:
Coreference Resolution

- Coreference resolution is biased[1,2]
  - Model fails for female when given same context



his ⇒ her

[1]Zhao et al. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. NAACL 2018.
[2]Rudinger et al. Gender Bias in Coreference Resolution. NAACL 2018

Wino-bias data

- Stereotypical dataset

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

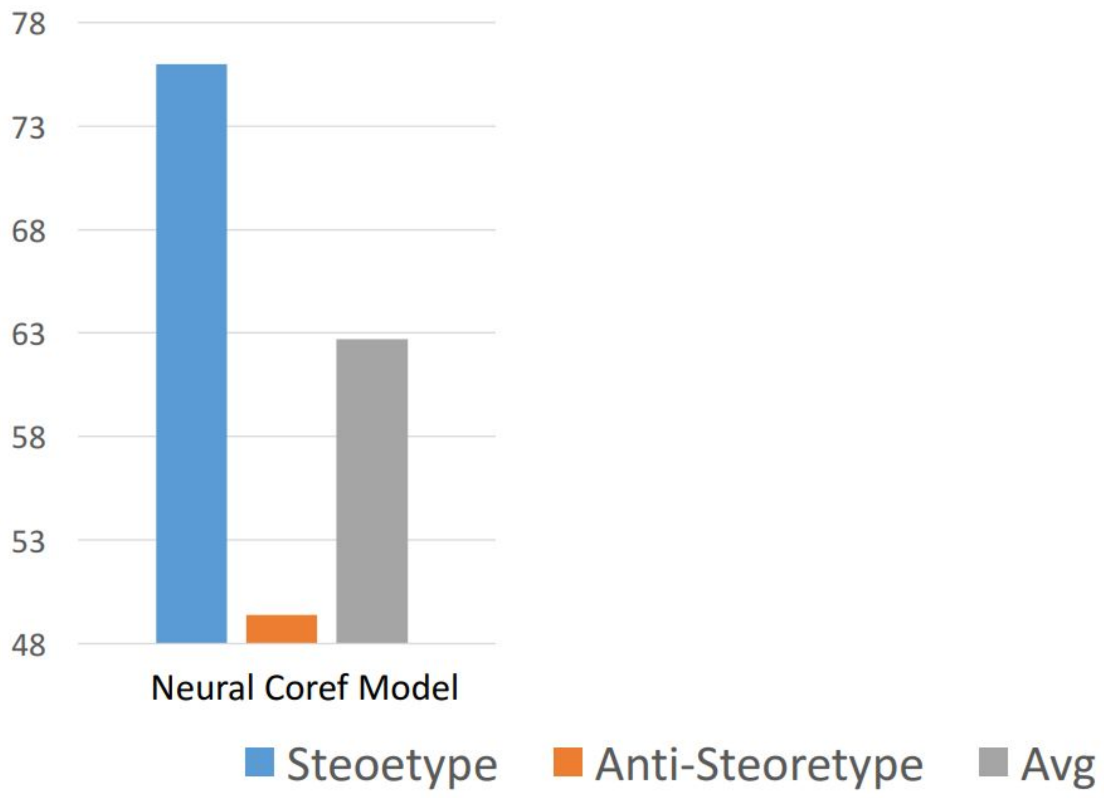
The physician hired the secretary because she was highly recommended.

- Anti-stereotypical dataset

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

The physician hired the secretary because he was highly recommended.

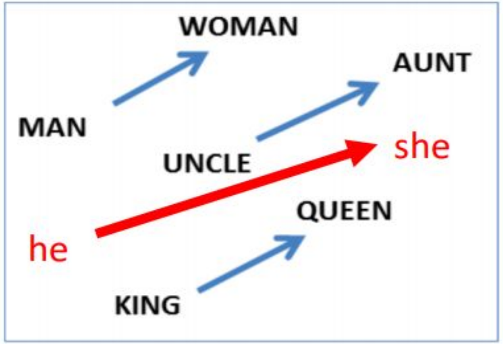
Gender bias in Coref System



Representational Harm in NLP: Word Embeddings can be Sexist

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings [Bolukbasi et al. NeurIPS16]

Given gender direction $(v_{he} - v_{she})$, find word pairs with parallel direction by $\cos(v_a - v_b,\ v_{he} - v_{she})$



he: ________	**she:________**
brother	sister
beer	
physician	
professor	

Google w2v embedding trained from the news

Word Embedding Association Test (WEAT)

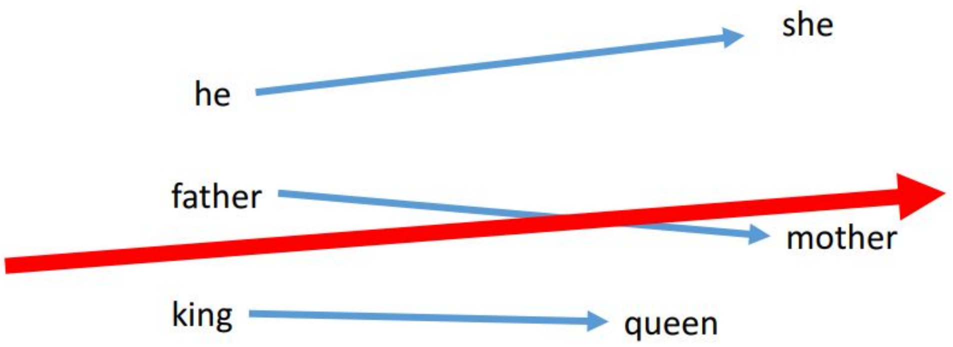
- ***X***: "mathematics", "science"; ***Y***: "arts", "design"
- ***A***: "male", "boy"; ***B***: "female", "girl"

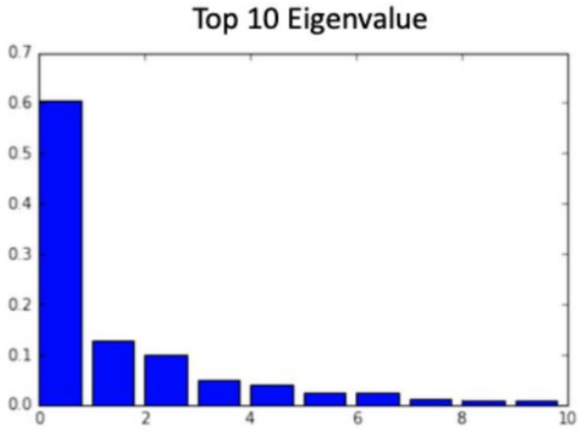
$$s(\vec{w}, A, B) = \frac{1}{|A|}\sum_{\vec{a}\in A}\cos(\vec{w}, \vec{a}) - \frac{1}{|B|}\sum_{\vec{b}\in B}\cos(\vec{w}, \vec{b}).$$

$$s(X, Y, A, B) = \sum_{\vec{x}\in X} s(\vec{x}, A, B) - \sum_{\vec{y}\in Y} s(\vec{y}, A, B),$$

The effect size of bias: $\dfrac{\text{mean}_{x\in X}\, s(x, A, B) - \text{mean}_{y\in Y}\, s(y, A, B)}{\text{std-dev}_{w\in X\cup Y}\, s(w, A, B)}$

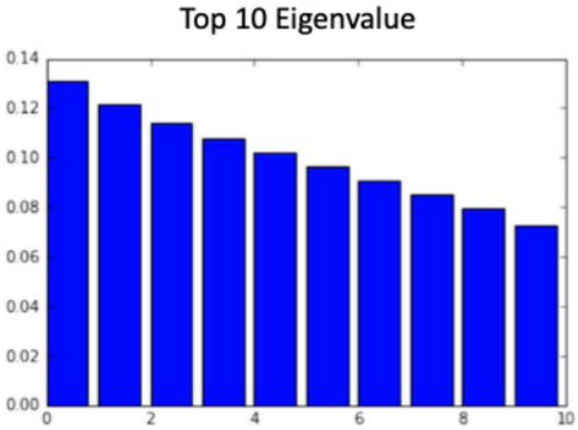
Caliskan et al. Semantics derived automatically from language corpora contain human-like biases Science. 2017





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

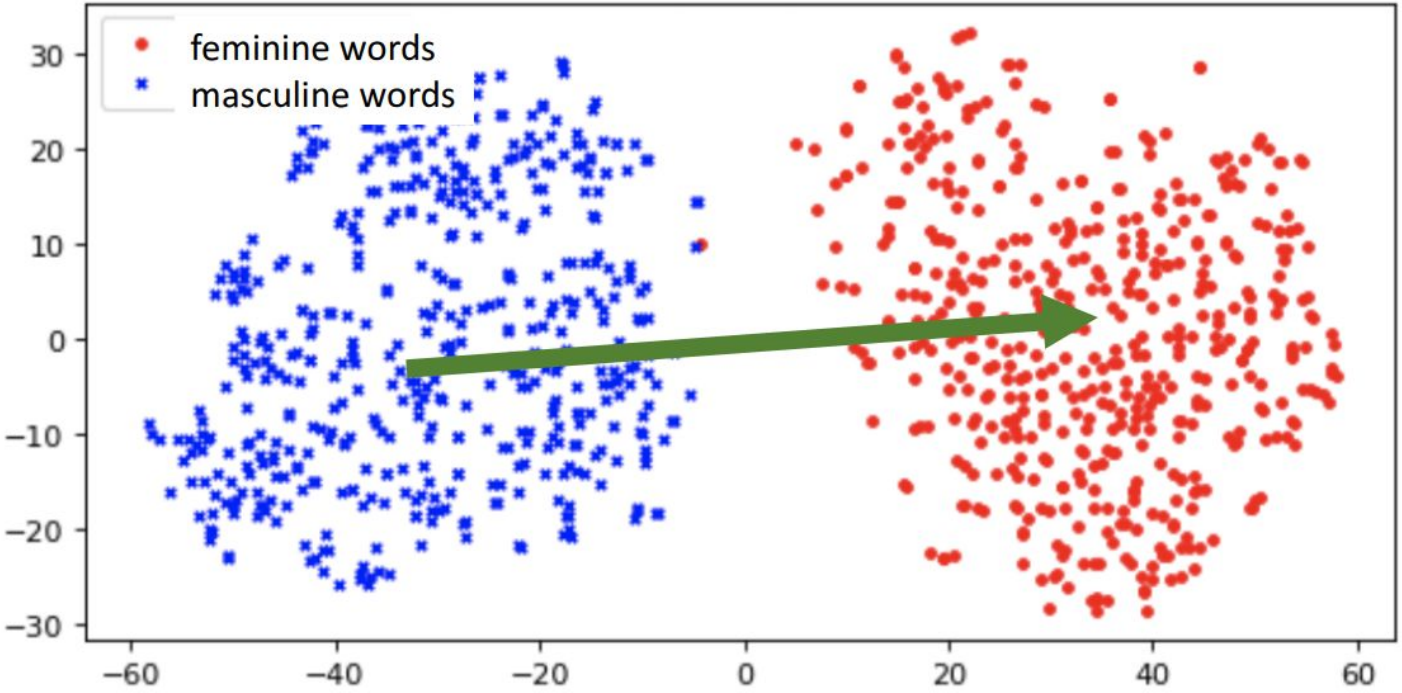
Gender Pair

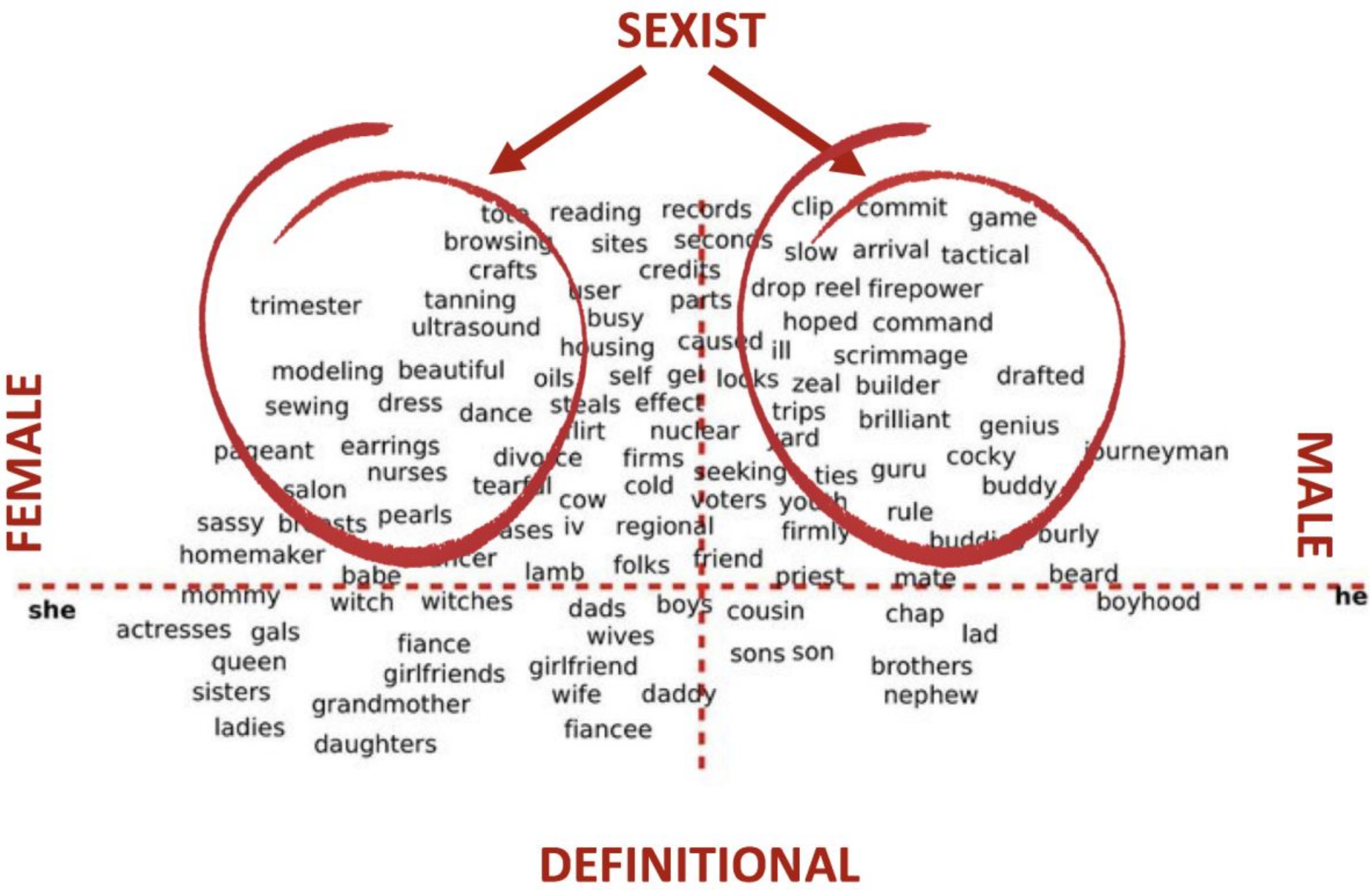


PCA (“dog”- “cat”, “house”-”building”,...)

Random Pair

- Linear Discriminative Analysis (LDA)
  - Identify grammatical gender direction

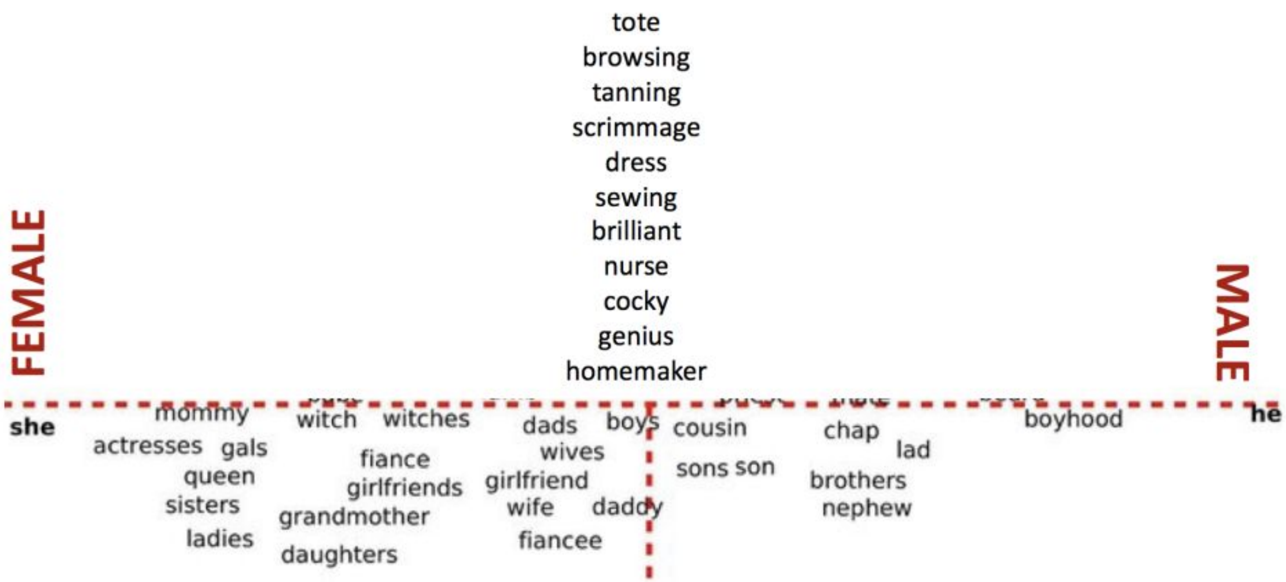




SEXIST
FEMALE
MALE
DEFINITIONAL
she
he
reading records clip commit game
browsing sites seconds slow arrival tactical
crafts credits drop reel firepower
trimester tanning user parts hoped command
ultrasound busy
housing caused ill scrimmage
modeling beautiful oils self gel looks zeal builder drafted
sewing dress dance steals effect trips brilliant genius
flirt nuclear
pageant earrings divorce firms seeking ties guru cocky
nurses cold
salon tearful cow voters buddy
sassy pearls iv regional rule
homemaker firmly burly
babe lamb folks friend priest mate beard
mommy witch witches dads boys cousin chap boyhood
actresses gals wives lad
queen fiance sons son
girlfriends girlfriend brothers
sisters grandmother wife daddy nephew
ladies fiancee
daughters

~~SEXIST~~

[Bolukbasi; NeurIPS 16]



DEFINITIONAL

This can be done by projecting gender direction out from gender neutral words using linear operations

Towards Debiasing

Bolukbasi et al. (2016)

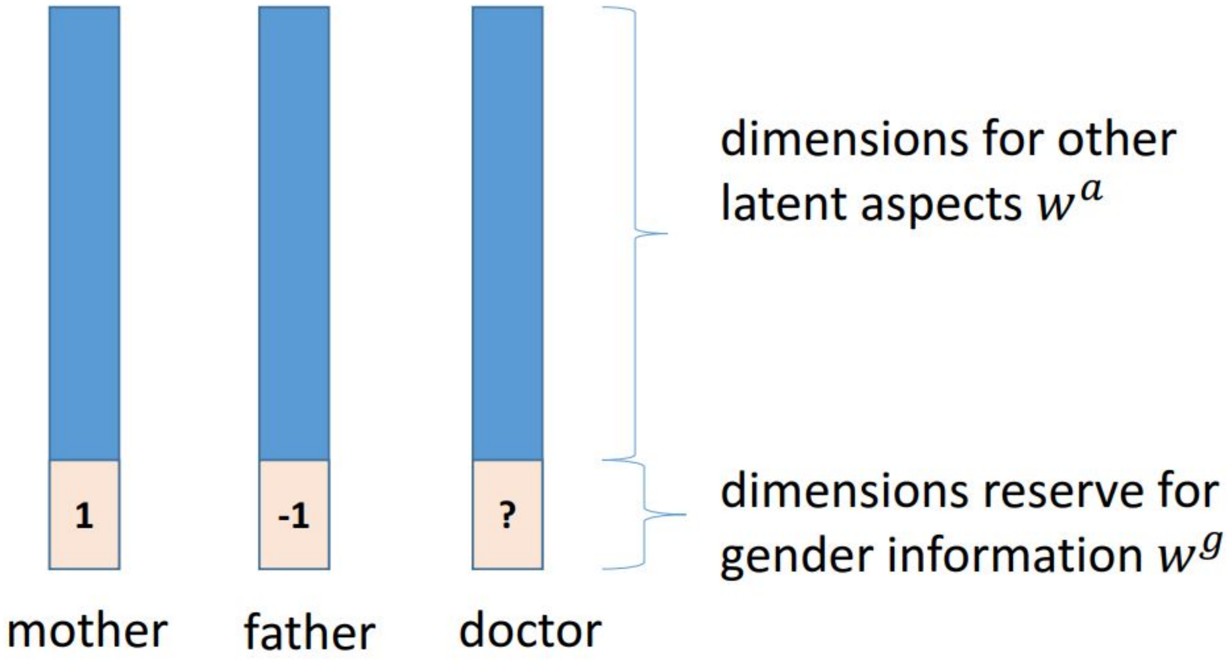
1. Identify gender subspace: B
2. Identify gender-definitional (S) and gender-neutral words (N)
3. Apply transform matrix (T) to the embedding matrix (W)
   a. Project away the gender subspace B from the gender-neutral words N
   b. But, ensure the transformation doesn't change the embeddings too much

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

T - the desired debiasing transformation B - biased space
W - embedding matrix N - embedding
matrix of gender neutral words

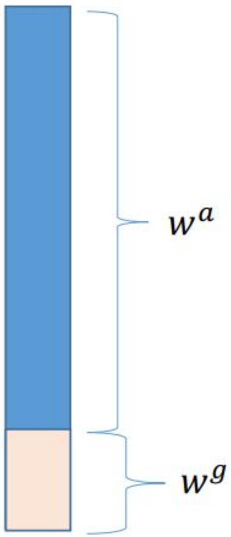
Make Gender Information Transparent in Word Embedding

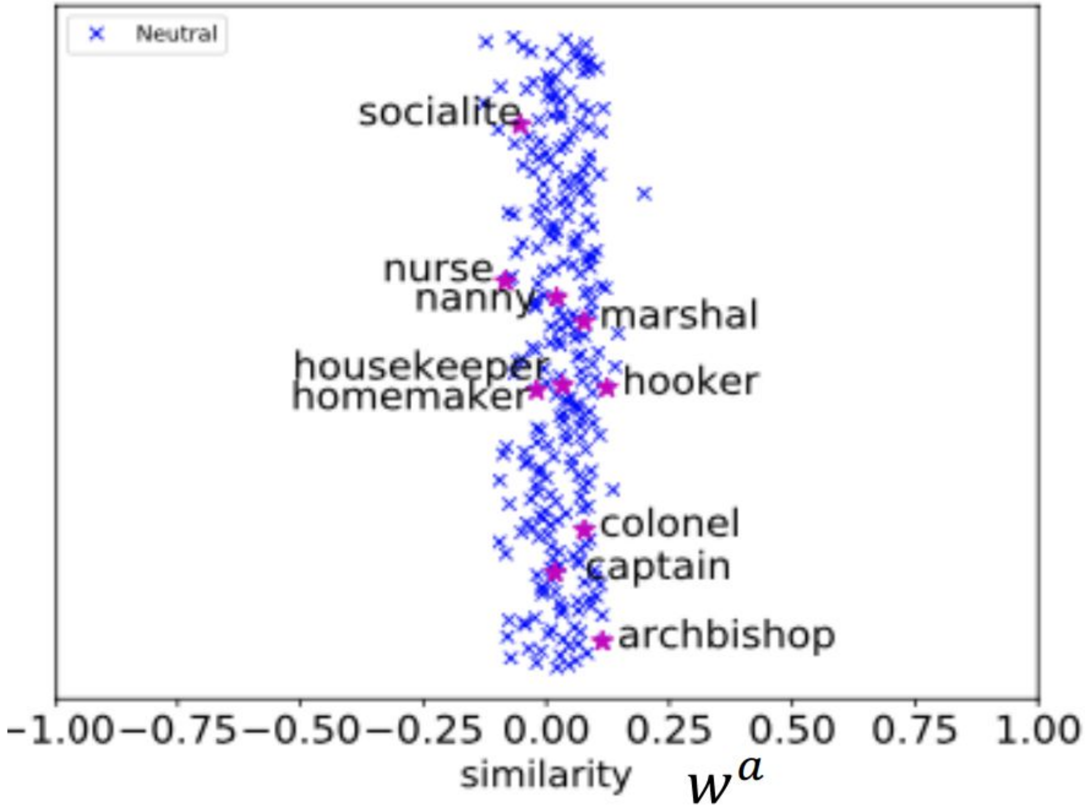
Learning Gender-Neutral Word Embeddings [Zhao et al; EMNLP18]



Make Gender Information Transparent in Word Embedding

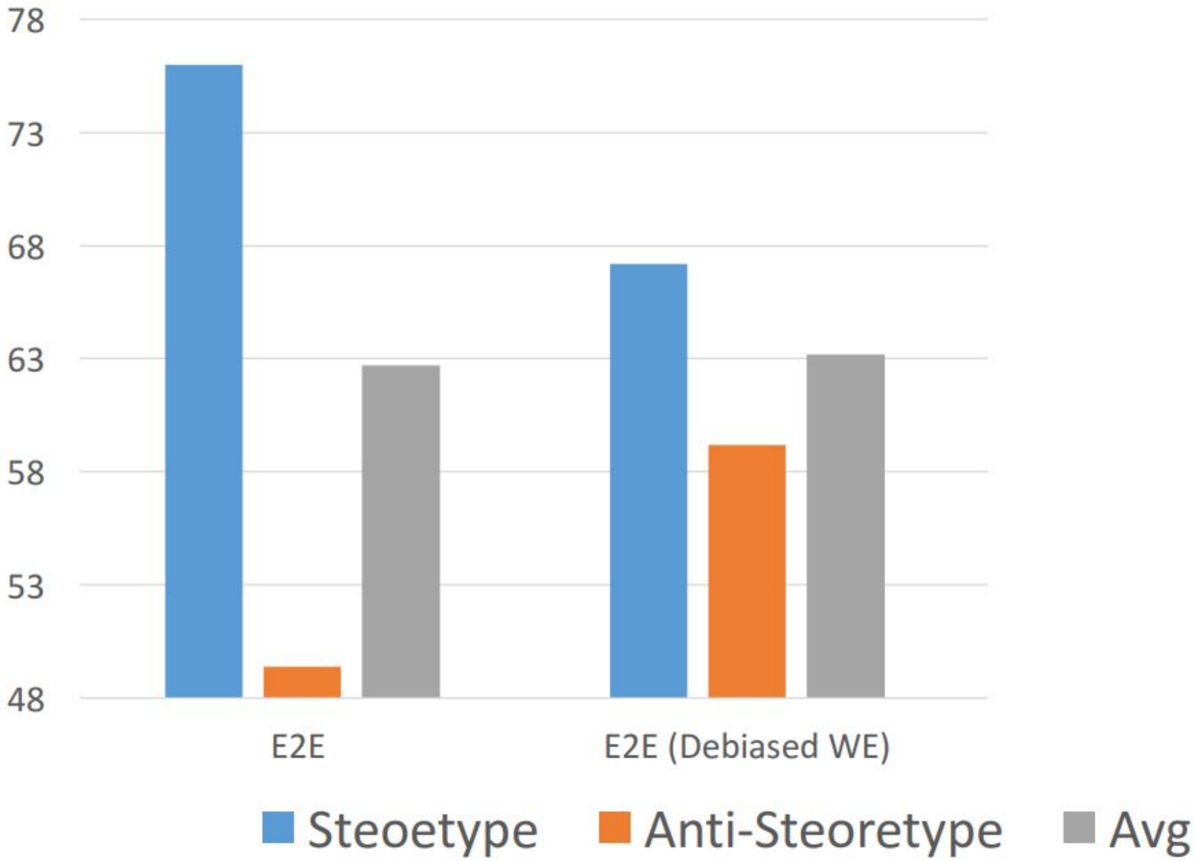
Learning Gender-Neutral Word Embeddings [Zhao et al; EMNLP18]





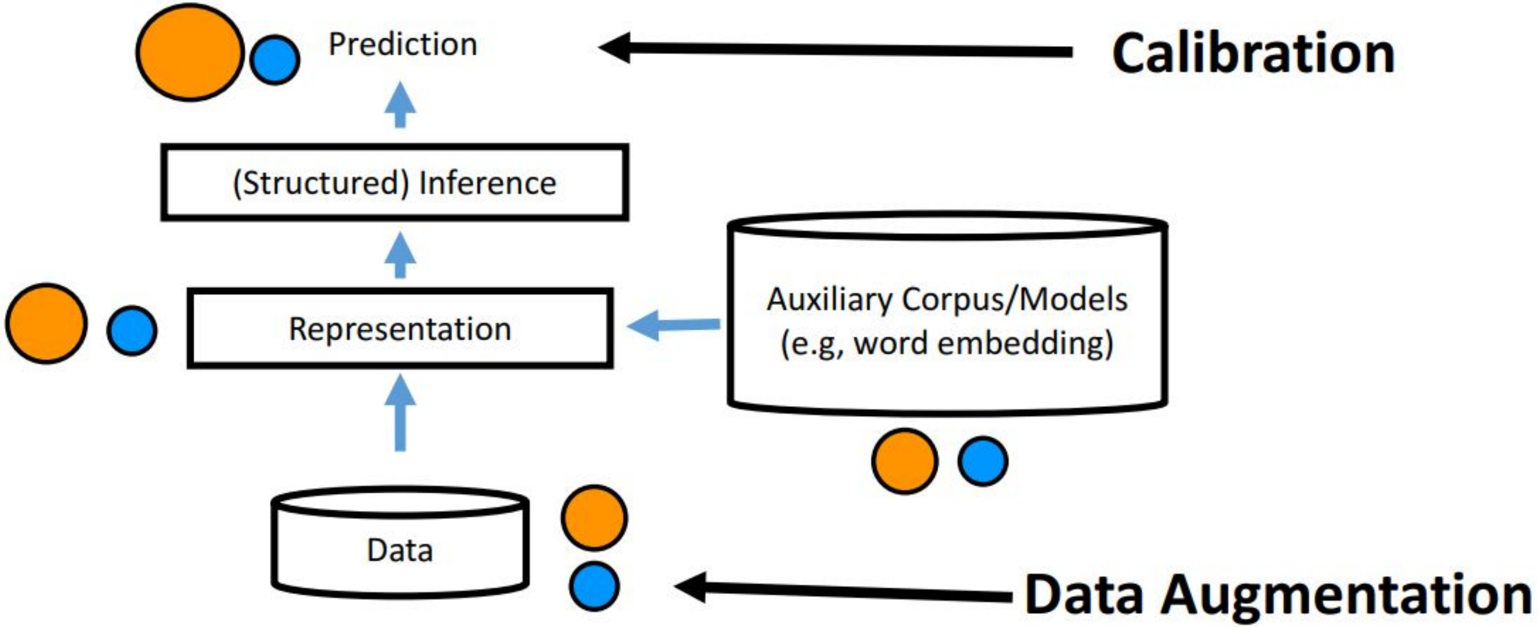
w^a

Gender bias in Coref System



Should We Debias Word Embedding?

- Awareness is better than blindness (Caliskan et. al. 17)



Wino-bias data

- Stereotypical dataset

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

The physician hired the secretary because she was highly recommended.

- Anti-stereotypical dataset

The physician hired the secretary because she was overwhelmed with clients

The physician hired the secretary because he was highly recommended.

Data Augmentation-- Balance the data

- Gender Swapping -- simulate sentence in opposite gender

John went to ~~his~~ house

F2 went to her house

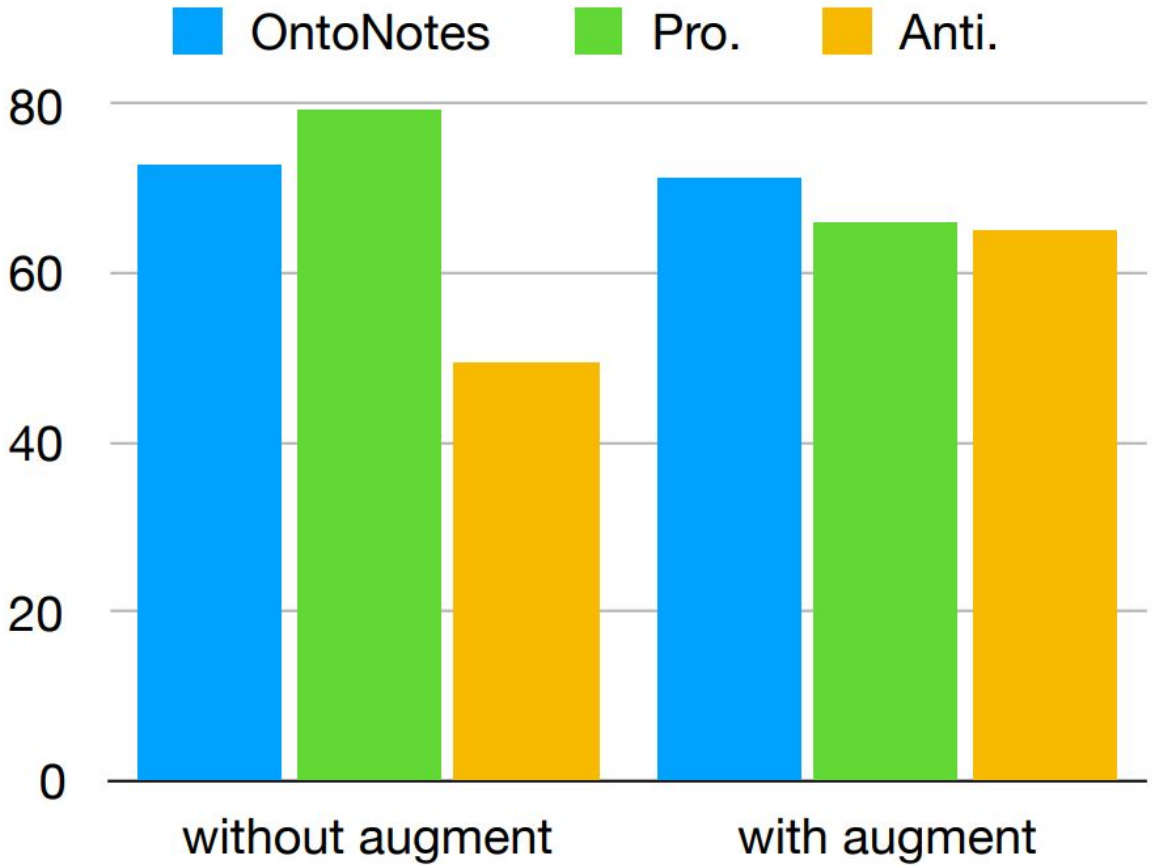
Named Entity are anonymized

Gender words are swapped

Better than down/up sampling

This idea has been used in computer vision as well

Reduce Bias via Data Augmentation in Coreference Resolution



Biases in NLP Classifiers/Taggers

- Gender Bias in Coreference resolution
  - Zhao, Jieyu, et al. **Gender bias in coreference resolution: Evaluation and debiasing methods.** *NAACL* (2018)
  - Webster, Kellie, et al. **Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns.** *TACL* (2018)
- Gender, Race, and Age Bias in Sentiment Analysis
  - Svetlana and Mohammad. **Examining gender and race bias in two hundred sentiment analysis systems**. arXiv (2018)
  - Díaz, et al. **Addressing age-related bias in sentiment analysis.** CHI Conference on Human Factors in Comp. Systems. (2018)
- LGBTQ identitiy terms bias in Toxicity classification
  - Dixon, et al. **Measuring and mitigating unintended bias in text classification.** AIES. (2018)
- Gender Bias in Occupation Classification
  - De-Arteaga et al. **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting.** FAT* (2019)

But aren’t they just reflecting Society?

But aren’t they just reflecting Society?

Yup!

Shouldn't we then just leave them as is?

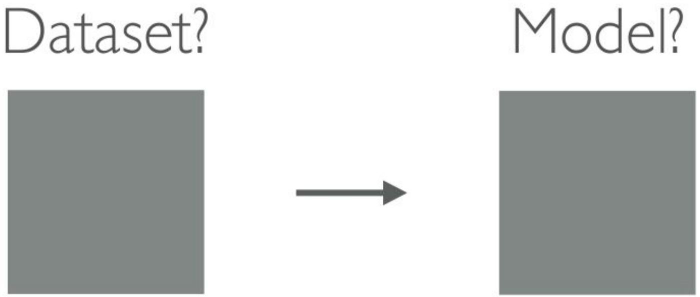
Would that harm certain groups of people?

Bias Amplification

- Zhao et al. Men also like shopping: Reducing Gender Bias Amplification using corpus-level constraints. *EMNLP (2017)*
- De-Arteaga et al. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. *FAT* (2019)*

Key Finding: Models Amplify Biases in the Dataset

Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**



Key Finding: Models Amplify Biases in the Dataset

Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**

Dataset? → Model?

Images of People Cooking

Key Finding: Models Amplify Biases in the Dataset

Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**

Dataset? → Model?

Men Cooking: 33% Women Cooking: 66%

Key Finding: Models Amplify Biases in the Dataset

Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**

Dataset? → Model?

Men Cooking: 33%

Women Cooking: 66%

Test Images

Key Finding: Models Amplify Biases in the Dataset

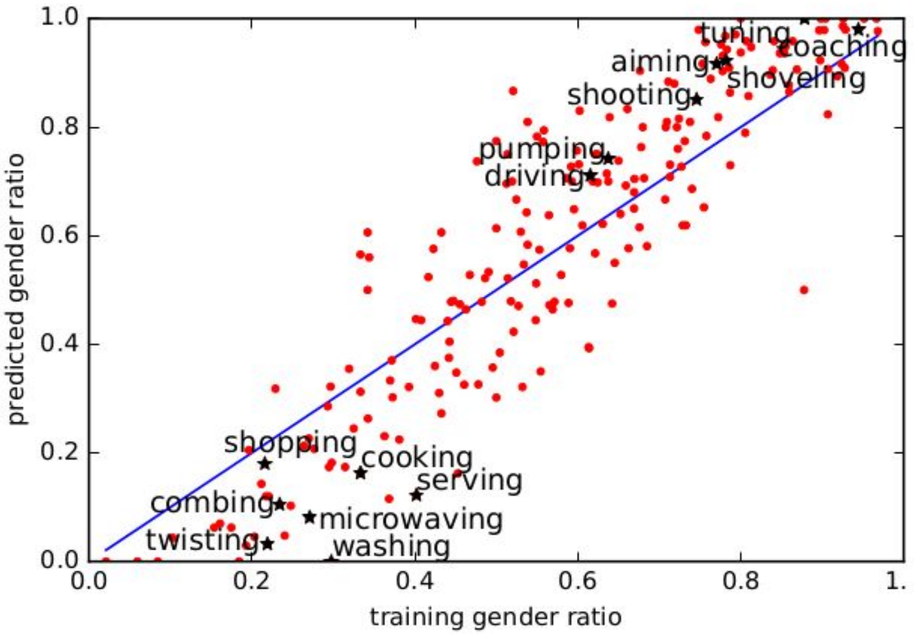
Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**

Dataset?

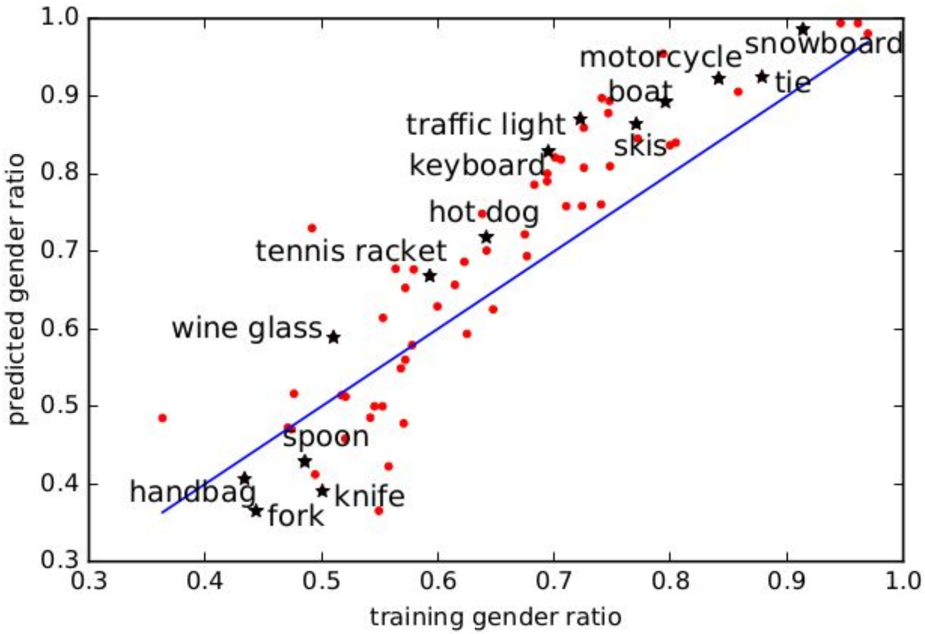
Model?

Men Cooking: 33% Women Cooking: 66%

Men Cooking: 16% Women Cooking: 84%

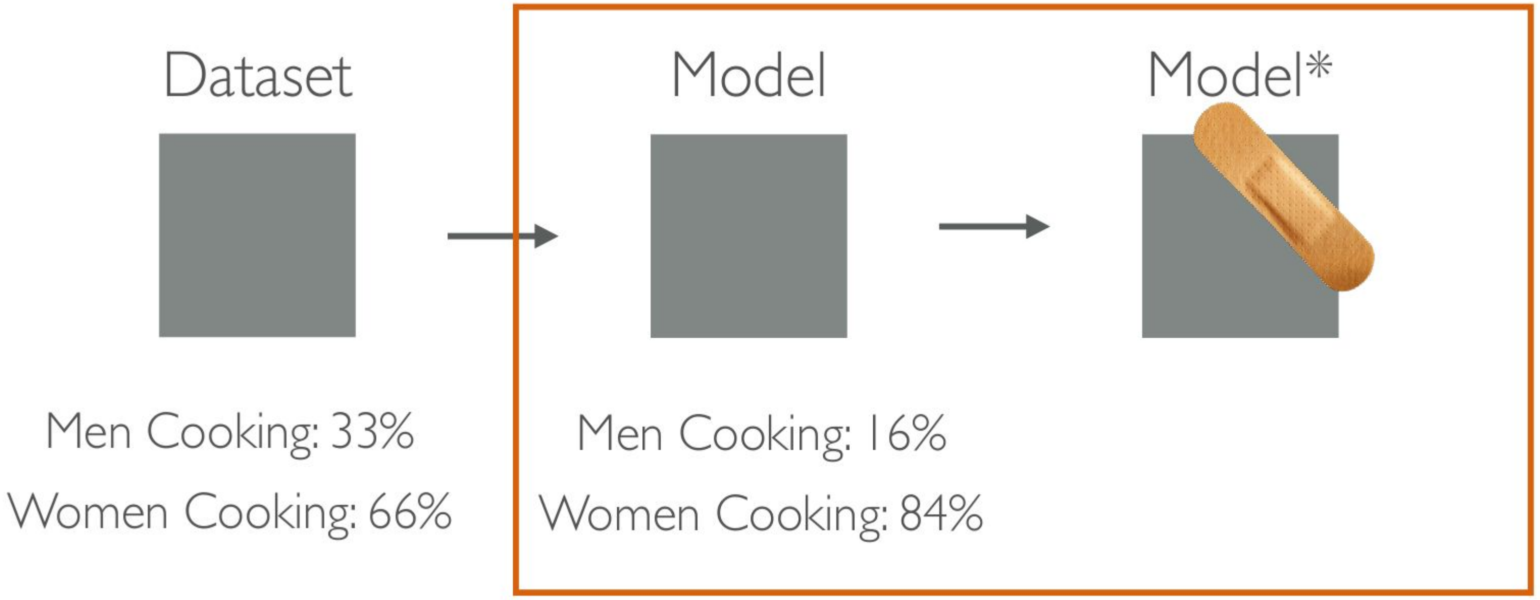


(a) Bias analysis on imSitu vSRL

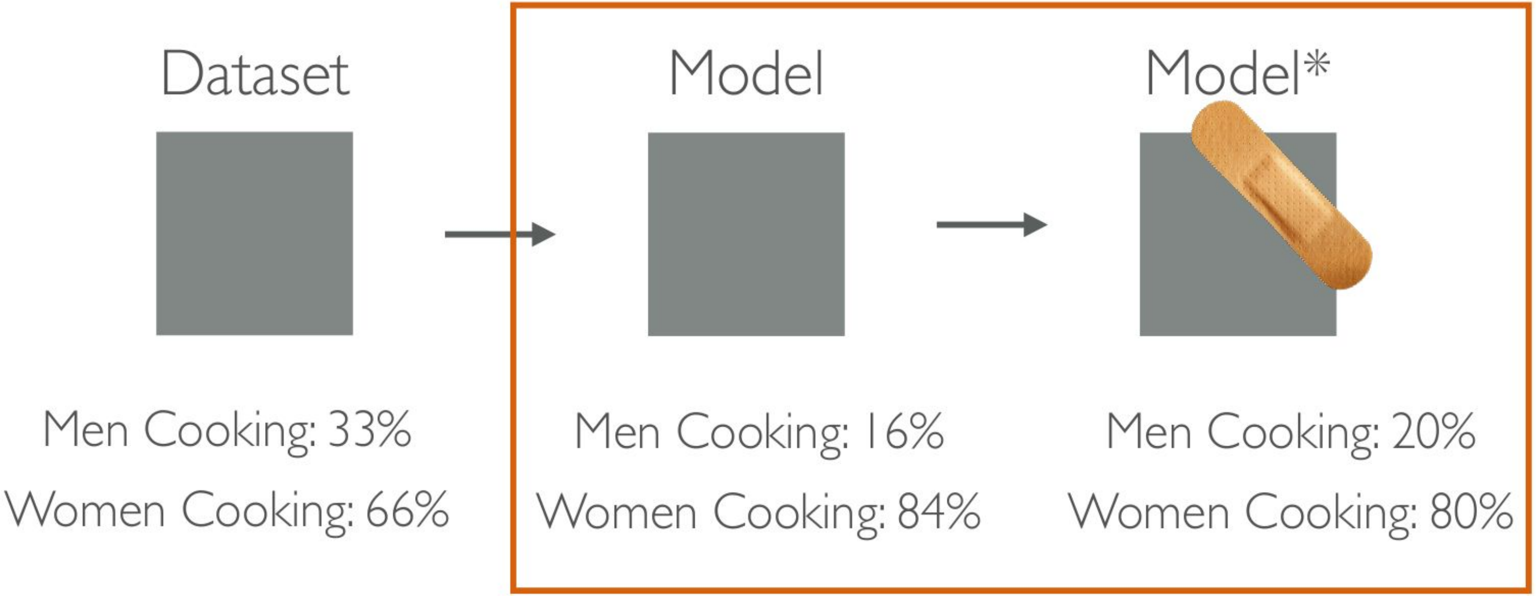


(b) Bias analysis on MS-COCO MLC

Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**



Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus Level Constraints
Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, Kai-Wei Chang. **EMNLP 2017**



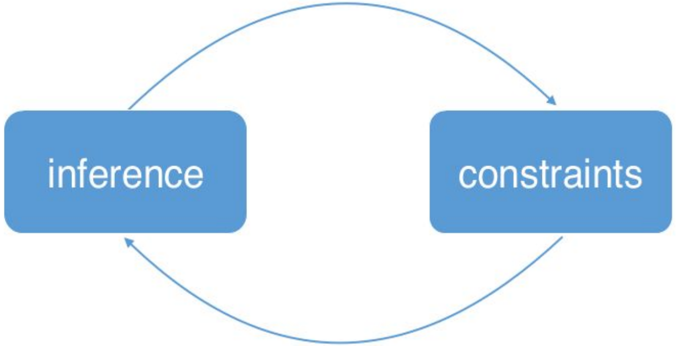
Reducing Bias Amplification (RBA)

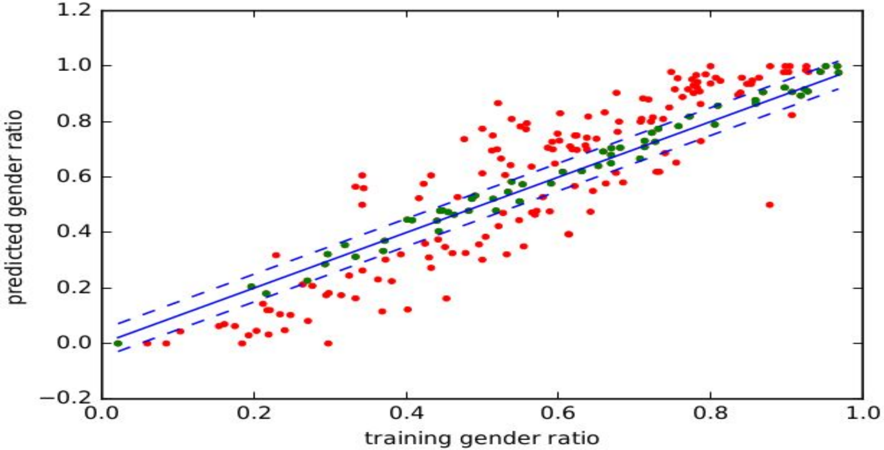
Integer Linear Program

$$\sum_{i} \max_{y_i} \; s(y_i, \text{image})$$

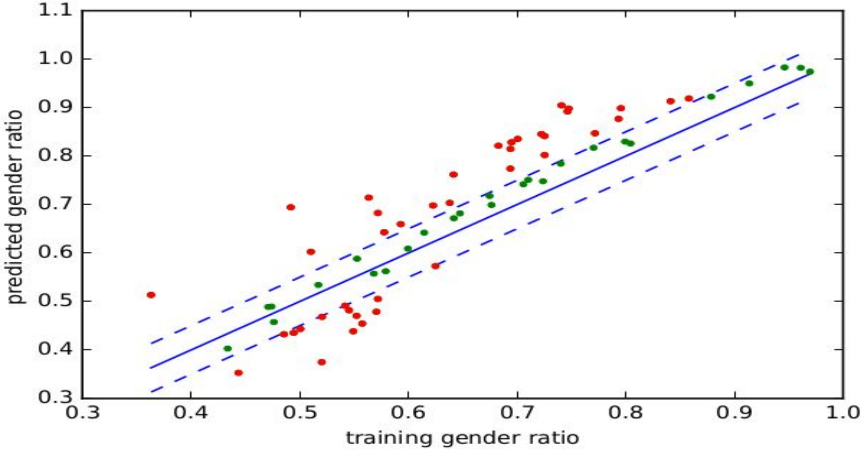
$$\forall \text{ points} \quad \left| \text{Training Ratio} - \underset{f(y_1 \ldots y_n)}{\text{Predicted Ratio}} \right| <= \text{margin}$$

$$\boxed{\text{Lagrangian Relaxation}}$$

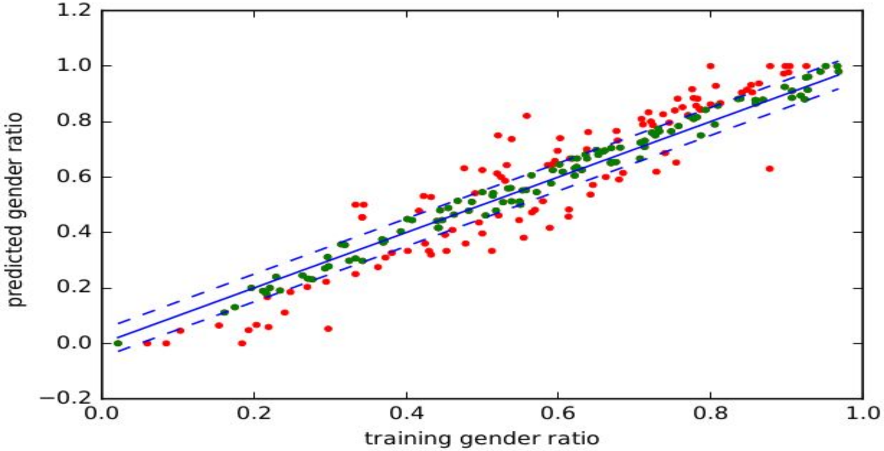




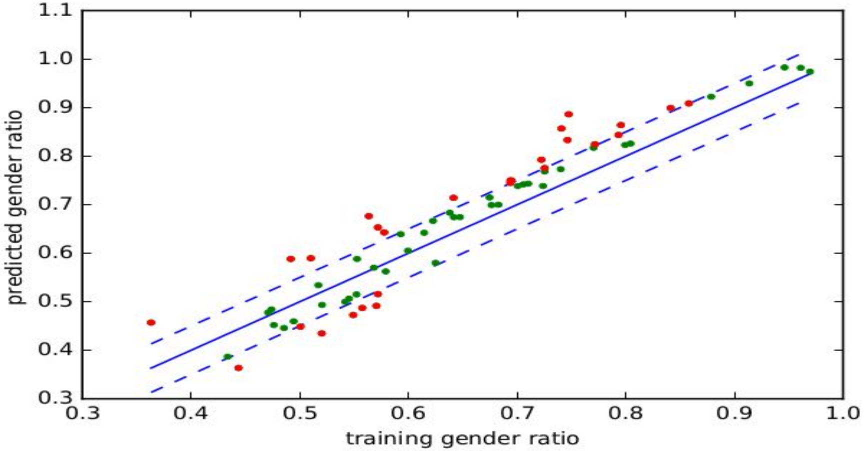
(a) Bias analysis on imSitu vSRL without RBA



(b) Bias analysis on MS-COCO MLC without RBA

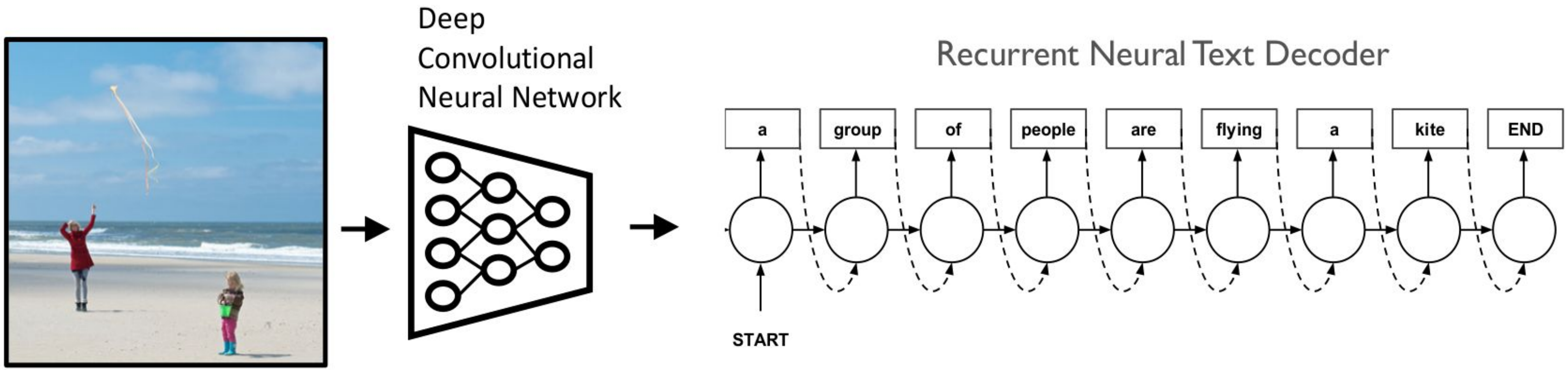


(c) Bias analysis on imSitu vSRL with RBA



(d) Bias analysis on MS-COCO MLC with RBA

Case Study: Image Captioning



$$\mathcal{L}^{CE} = -\frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}\log(p(w_t|w_{0:t-1}, I))$$

Case Study: Image Captioning

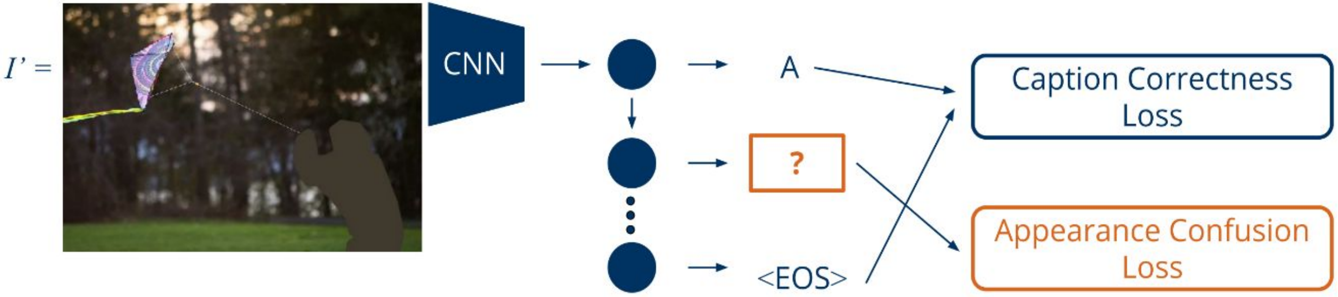
A woman cooking a meal

A man wearing a black hat is snowboarding

Women also Snowboard: Overcoming Bias in Captioning Models
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach. **ECCV 2018**

Approach I: Add a Confusion Loss

Idea: Augment the data by removing people artificially, and keep a set of gendered reference words where a different loss will be applied



Words for every pair of genders should be equally probable

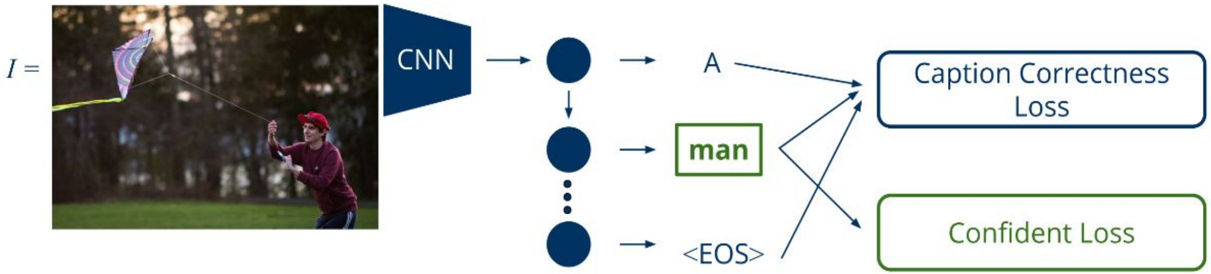
$$\mathcal{C}(\tilde{w}_t, I') = \left| \sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I') \right|$$

$$\mathcal{L}^{AC} = \frac{1}{N} \sum_{n=0}^{N} \sum_{t=0}^{T} \mathbb{1}(w_t \in \mathcal{G}_w \cup \mathcal{G}_m) \mathcal{C}(\tilde{w}_t, I')$$

Women also Snowboard: Overcoming Bias in Captioning Models
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach. **ECCV 2018**

Approach II: Add a Confidence Loss

Idea: Discourage the following from happening at the same time:
P(word = man) = 0.95 and **P(word = woman) = 0.92**



Take into account mutual exclusion among groups of words

$$\mathcal{L}^{Con} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}(\mathbb{1}(w_t \in \mathcal{G}_w)\mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m)\mathcal{F}^M(\tilde{w}_t, I))$$

$$\mathcal{F}^W(\tilde{w}_t, I) = \frac{\sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I)}{(\sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I)) + \epsilon}$$

Women also Snowboard: Overcoming Bias in Captioning Models
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach. **ECCV 2018**

Model	MSCOCO-Bias Error	MSCOCO-Bias Ratio Δ	MSCOCO-Balanced Error	MSCOCO-Balanced Ratio Δ
Baseline-FT	12.83	0.15	19.30	0.51
Balanced	12.85	0.14	18.30	0.47
UpWeight	13.56	0.08	16.30	0.35
Equalizer w/o ACL	7.57	0.04	10.10	0.26
Equalizer w/o Conf	9.62	0.09	13.90	0.40
Equalizer	**7.02**	**-0.03**	**8.10**	**0.13**

“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian

CREDIT

[The Guardian view on machine learning: people must decide]

"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."

— The Guardian

CREDIT

[The Guardian view on machine learning: people must decide]

Open Research Questions

- Coming up with data-driven metrics for fairness
- Understanding the causes of model bias amplification
- Incorporating group fairness constraints during training

Thank You !!!