What It Takes to Control Societal Bias in Natural Language Processing

Kai-Wei Chang
UCLA

References: http://kwchang.net

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

The father dies.

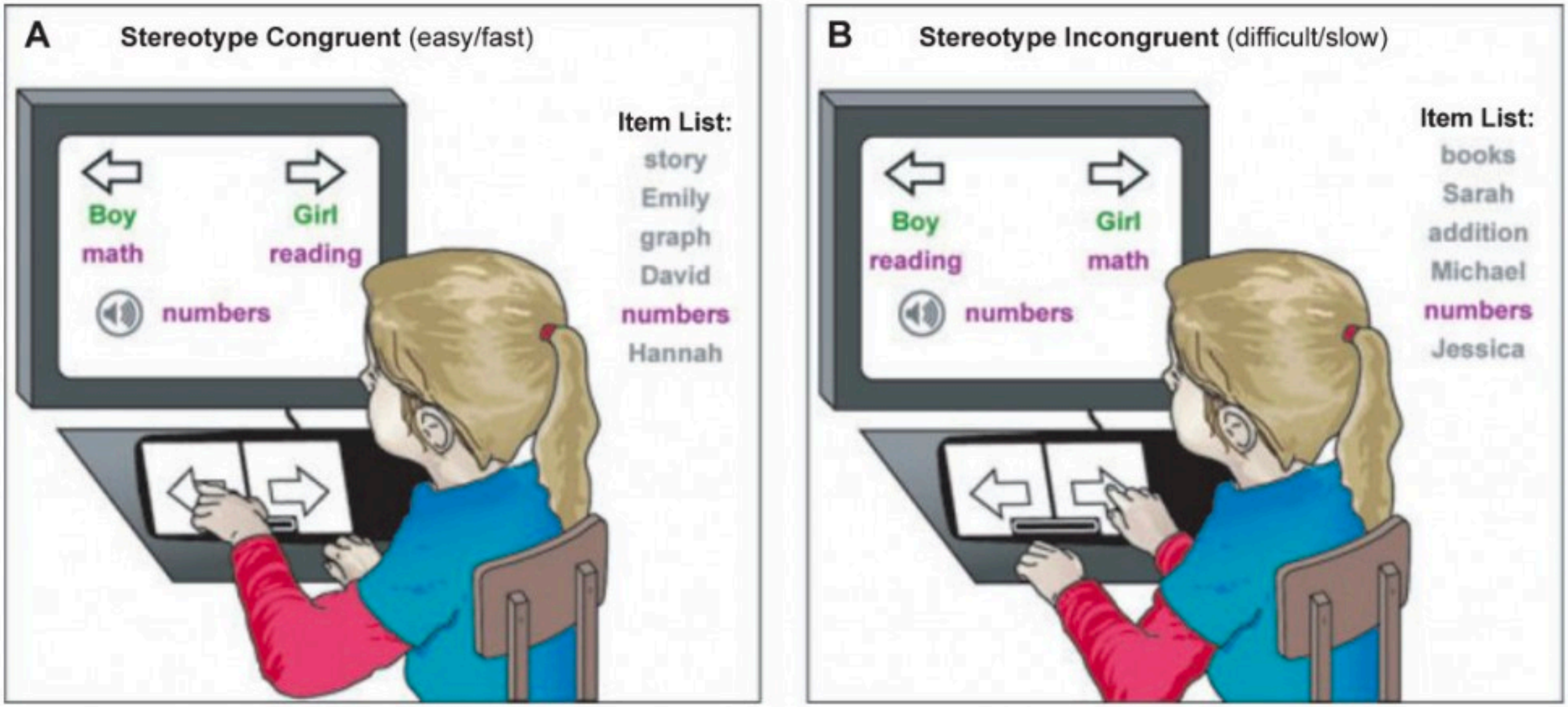
The boy is taken to the operating room and the surgeon says,

"I can't operate on this boy, because he's my son."

Can you explain why?

https://www.youtube.com/watch?v=J69HkKz9g4A

Implicit association test (IAT)



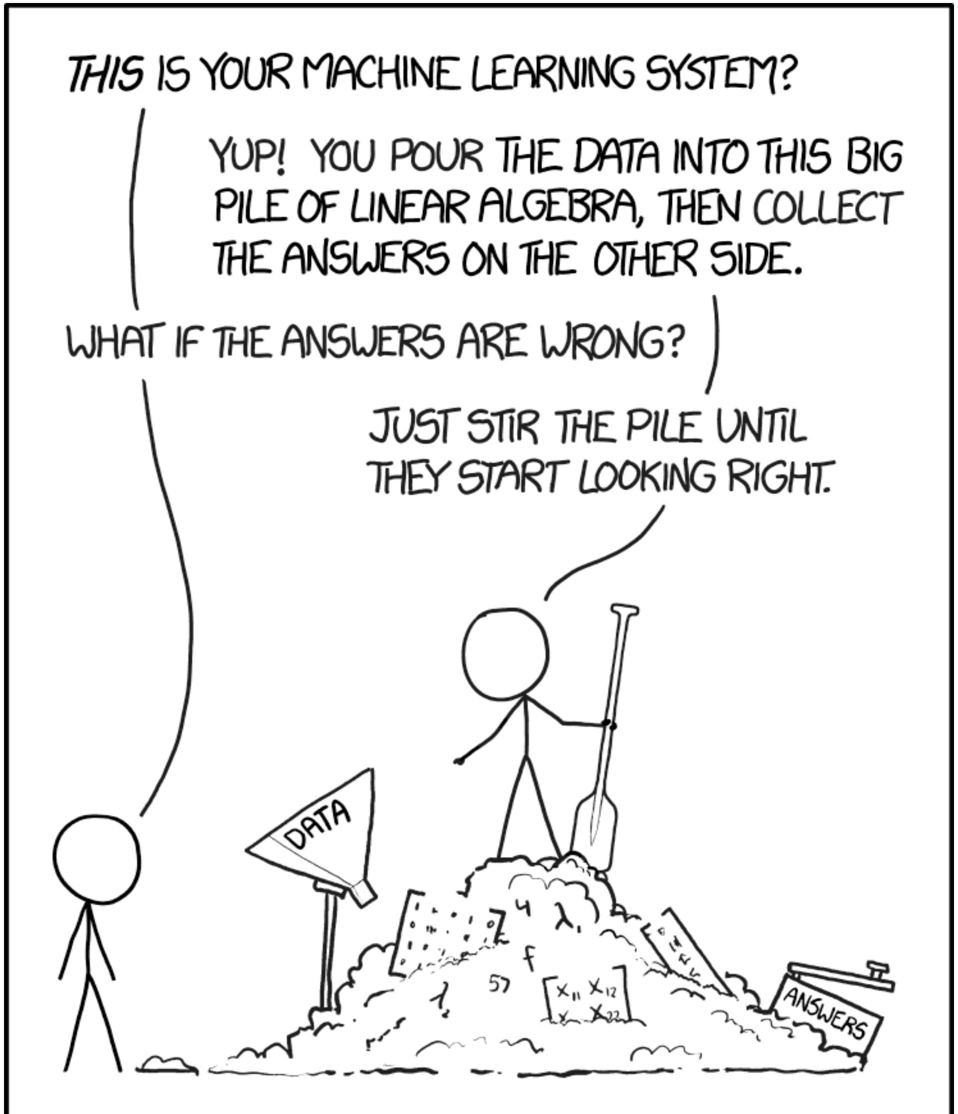
https://implicit.harvard.edu

“Concepts in semantic memory are assumed to be linked together … with associated concepts having stronger links … than unrelated concepts” (Collins and Loftus, 1975).

- https://www.nature.com/articles/palcomms201786

“Concepts in semantic memory are assumed to be linked together … with associated concepts having stronger links … than unrelated concepts” (Collins and Loftus, 1975).

So does computer



Data with Societal Bias

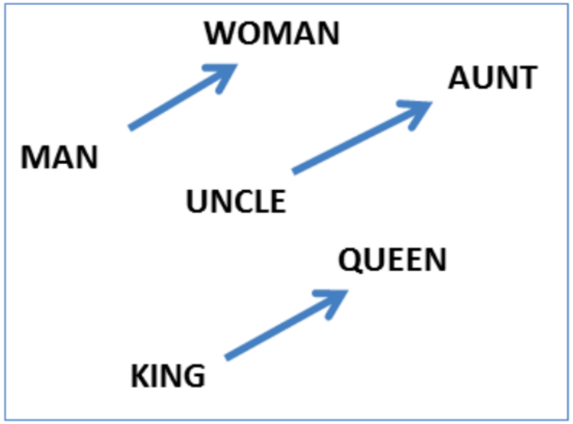
Model with Societal Bias

https://xkcd.com/1838/

Word Embeddings can be Dreadfully Sexist [nips16]

w/ Tolga Bolukbasi, James Zou, Venkatesh Saligrama, Adam Kalai

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



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

We use Google w2v embedding trained from the news

Google

word2vec resume

Scholar

About 93 results (**0.02** sec)

Articles

Case law

My library

Any time

Since 2016

Since 2015

Since 2012

Custom range.

Sort by relevar

Sort by date

include patents

include citations

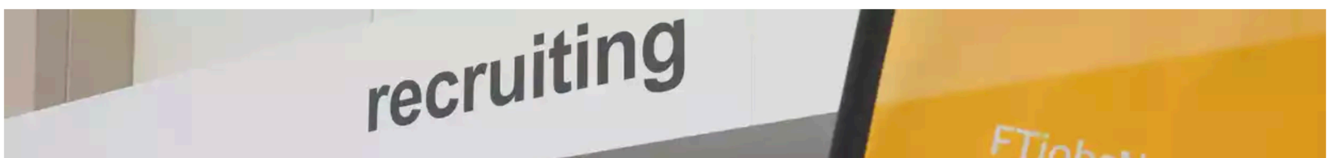
Create alert

Machine Learned **Resume**-Job Matching Solution

Amazon

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process



présente une approche associant réseaux lexico-sémantiques et représentations distribuées de mots appliquée à l'évaluation de la traduction automatique. **...**

Cite Save

Macau: Large-scale skill sense disambiguation in the online recruitment domain

Q Luo, M Zhao, F Javed, F Jacob - Big Data (Big Data), 2015 ..., 2015 - ieeexplore.ieee.org

... Contexts are extracted from either skill section(s) of **resumes** or requirement section(s) of job postings. We used a popular tool **word2vec** [12] with parameter

Related works

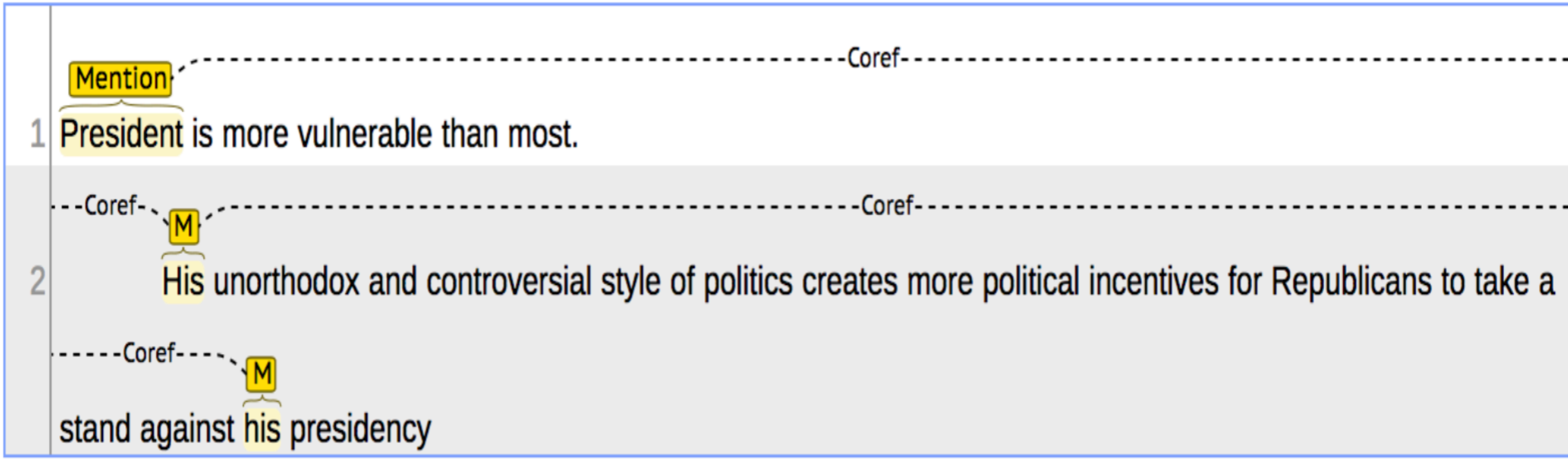
Aylin, Joanna, and Arvind (2017) measure the biases in embedding using Implicit Association Test (IAT) and demonstrate it contain human-like biases

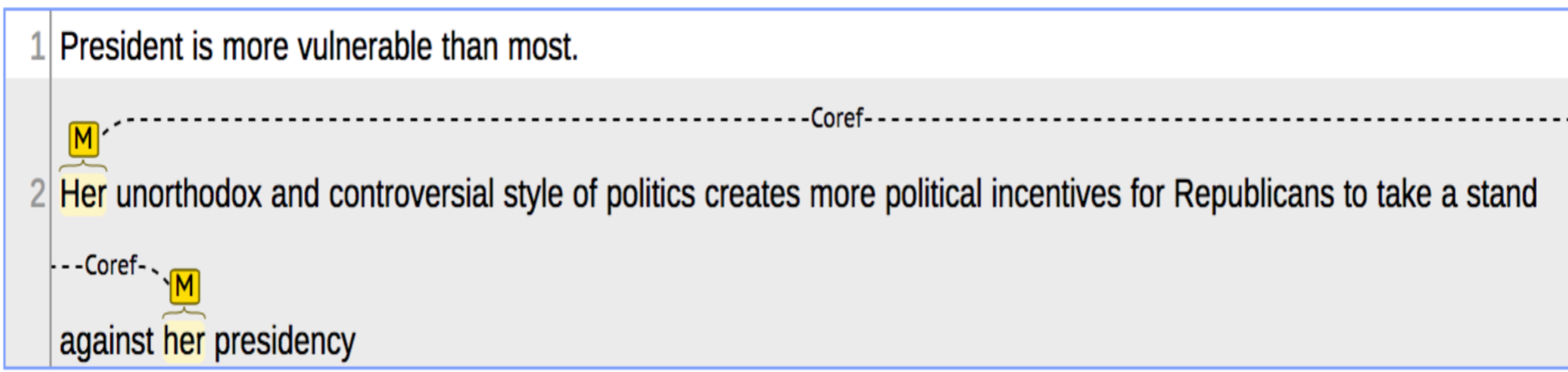
Garg, Schiebinger, Jurafsky, Zou (2017) Word embeddings quantify 100 years of gender and ethnic stereotypes:

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

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

Gender Bias in Coref [NAACL 2018]





Concurrent work (Rudinger et al., 2018) @NAACL18 also studied gender bias in Coref.

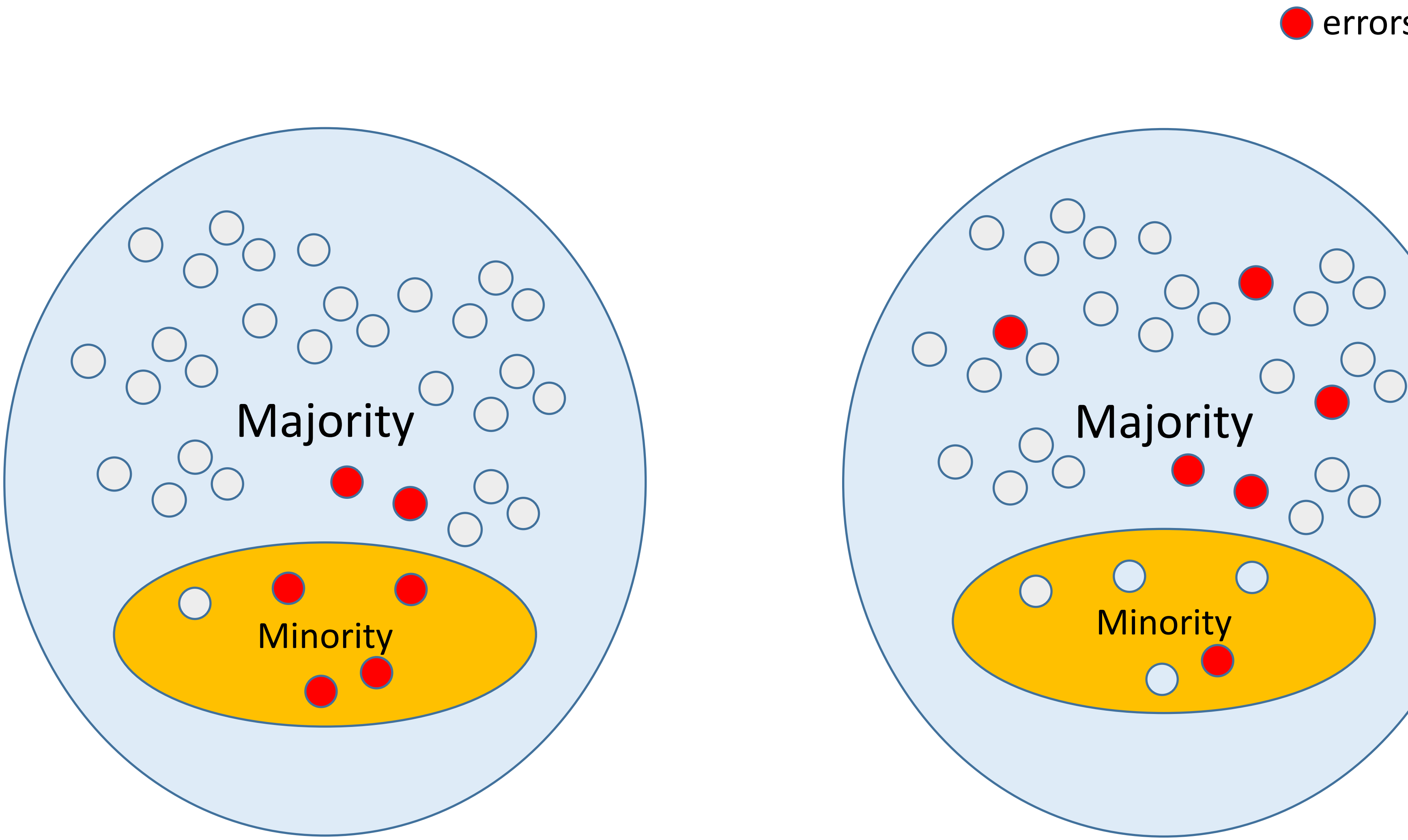
Demographic Dialectal Variation in Social Media: A Case Study of African-American English

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

Africa-American English

Non Africa-American English

Parser	AA	Wh.	Difference
SyntaxNet	64.0 (2.5)	80.4 (2.2)	16.3 (3.4)
CoreNLP	50.0 (2.7)	71.0 (2.5)	21.0 (3.7)



error rate: 6/30 = 80%

error rate: 6/30 = 80%

Human Bias in Structured Prediction Models

[EMNLP 17*] w/ Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez

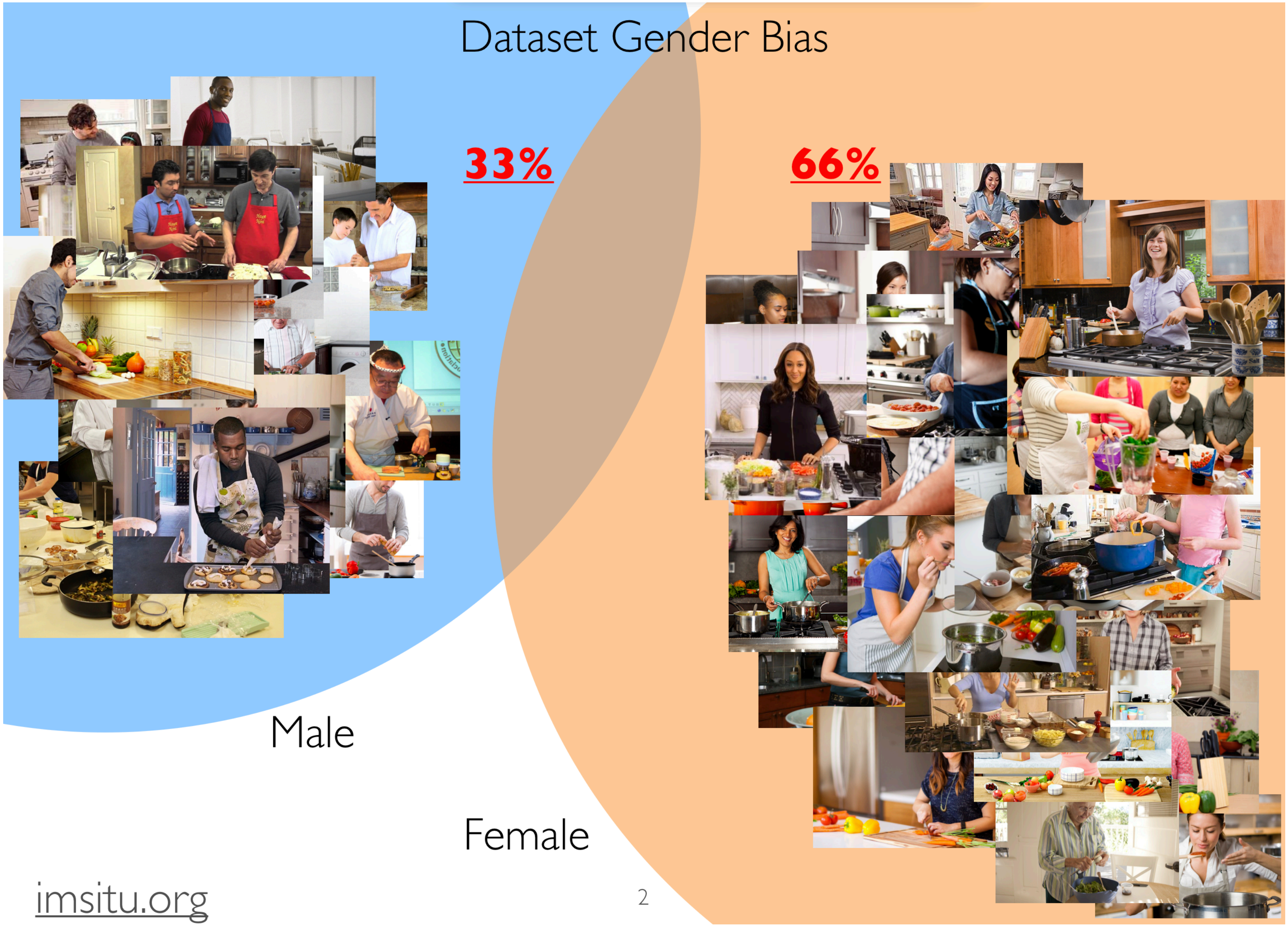
What's the agent for this image?

Cooking	
Role	Object
agent	?
food	vegetable
container	bowl
tool	knife
place	kitchen

An example from a vSRL (visual Semantic Role Labeling) system

*Best Long Paper Award at EMNLP 17

Dataset Gender Bias



Model Bias After Training

16%

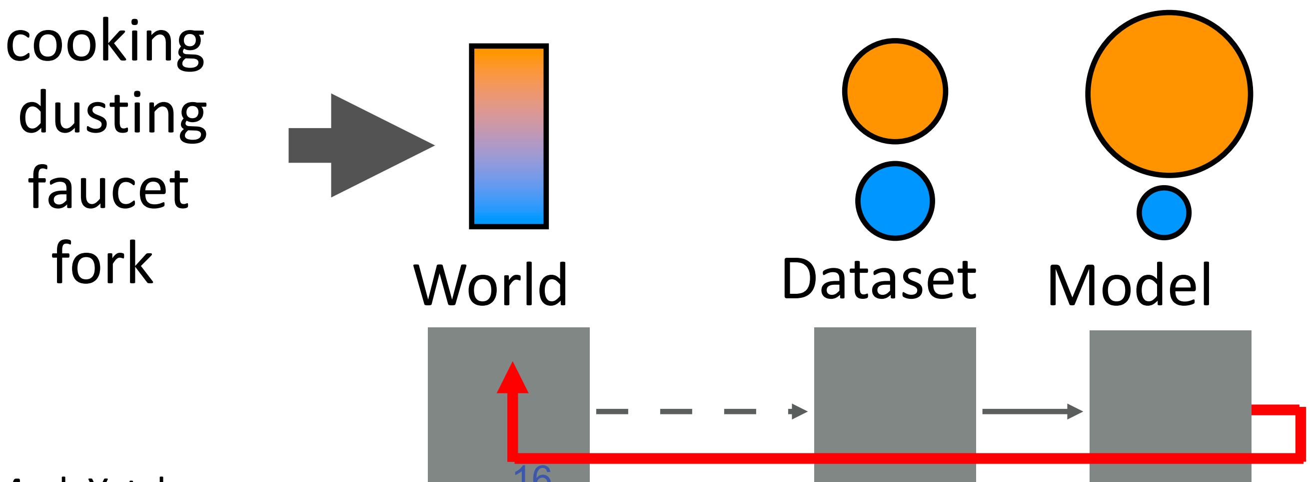
84%

Male

Female

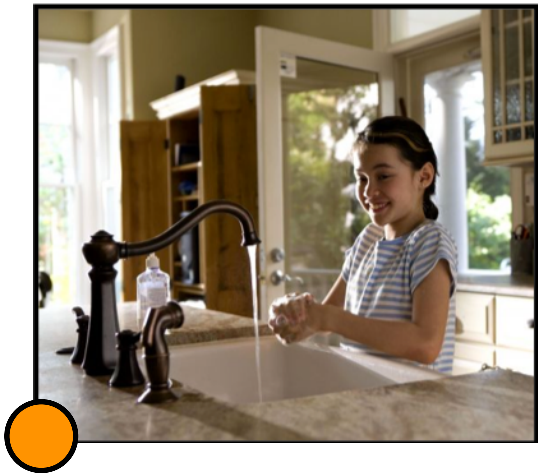
imsitu.org

Algorithmic Bias in Applications

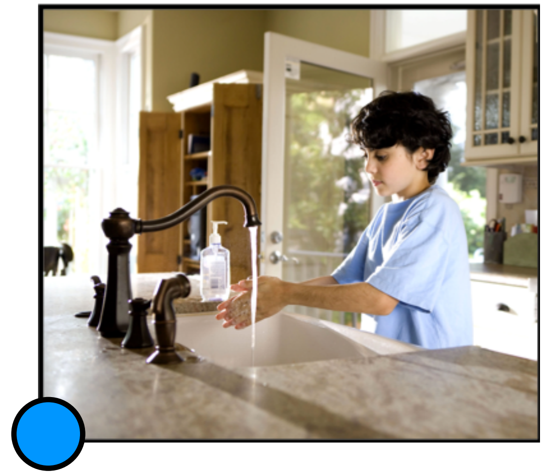


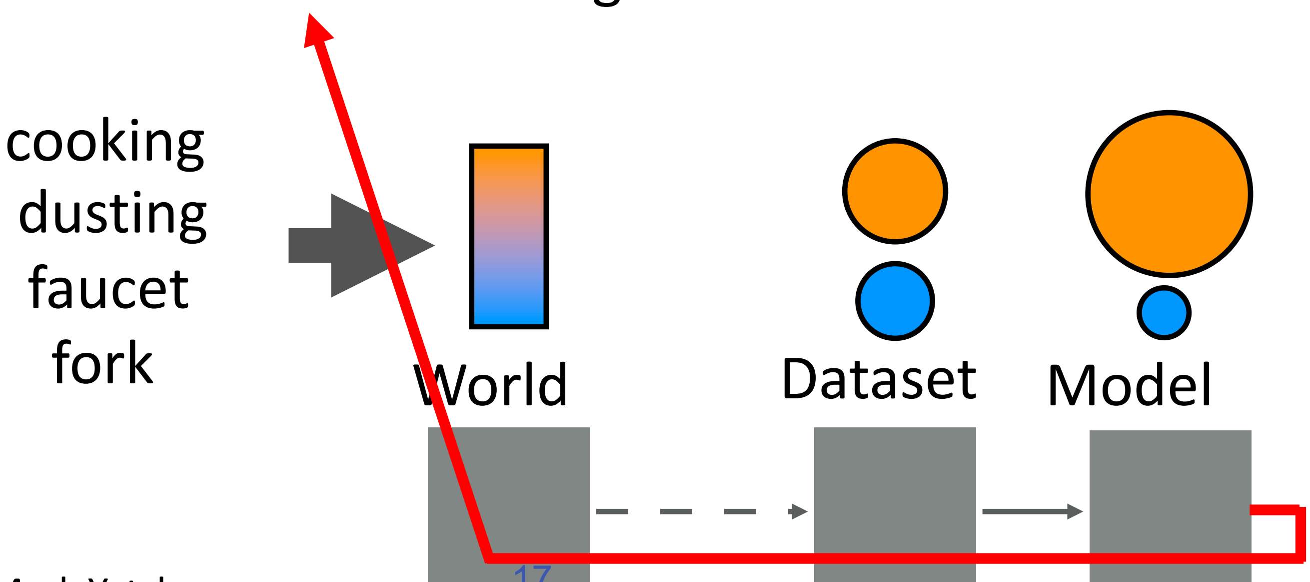
Credit: Mark Yatskar

Algorithmic Bias in Applications



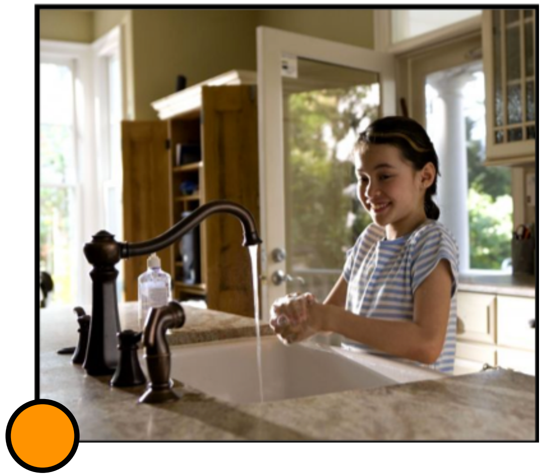
woman cooking



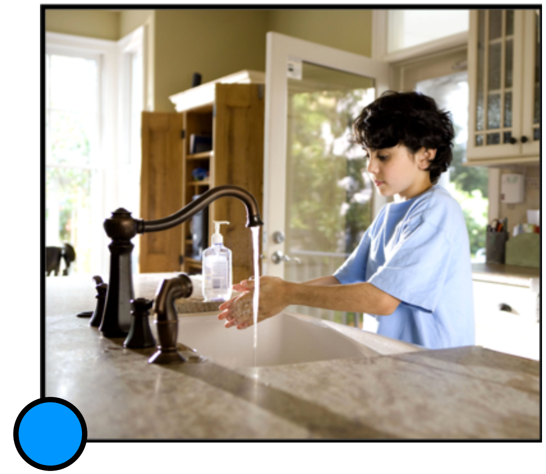


Credit: Mark Yatskar

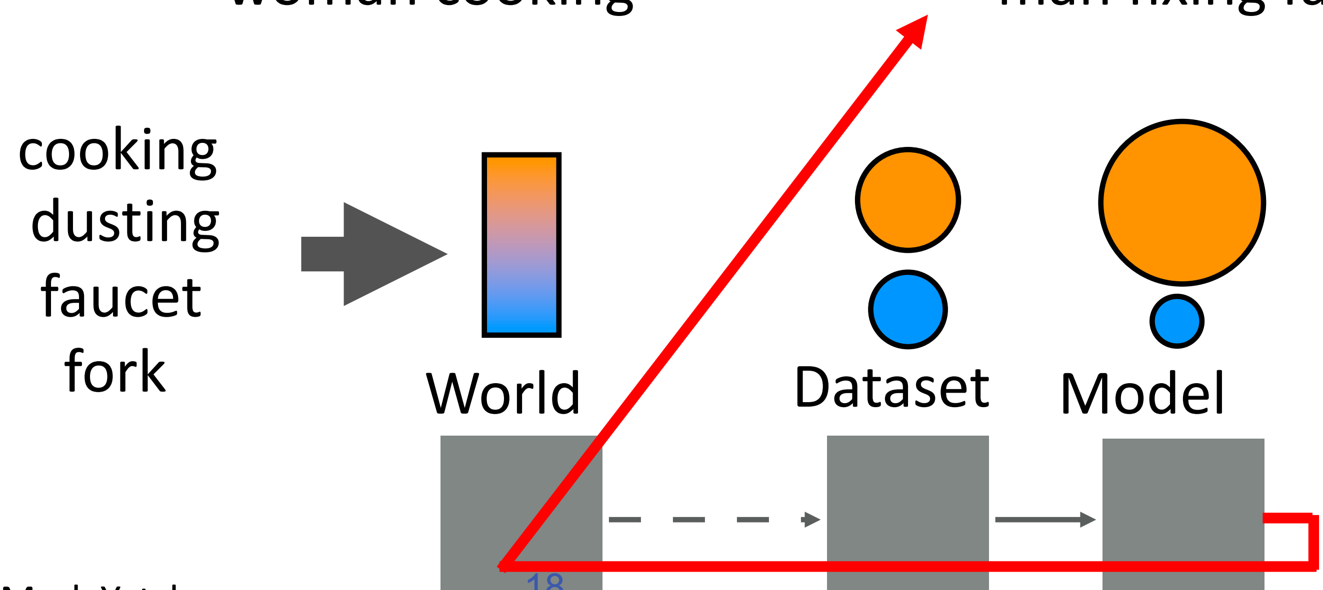
Algorithmic Bias in Applications



woman cooking

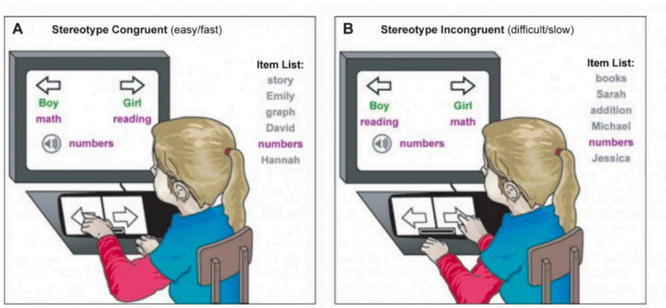


man fixing faucet



Credit: Mark Yatskar

When AI products exhibit Societal Bias



GRACE HOPPER CELEBRATION

ANITA B.ORG

What It Takes to Control Societal Bias in NLP?

“Gender discrimination is illegal in the United States.”

Your query

Gender discrimination is illegal in the United States.

Tagging

```
Gender/JJ  discrimination/NN  is/VBZ  illegal/JJ  in/IN  the/DT  United/NNP  States/NNPS  ./.
```

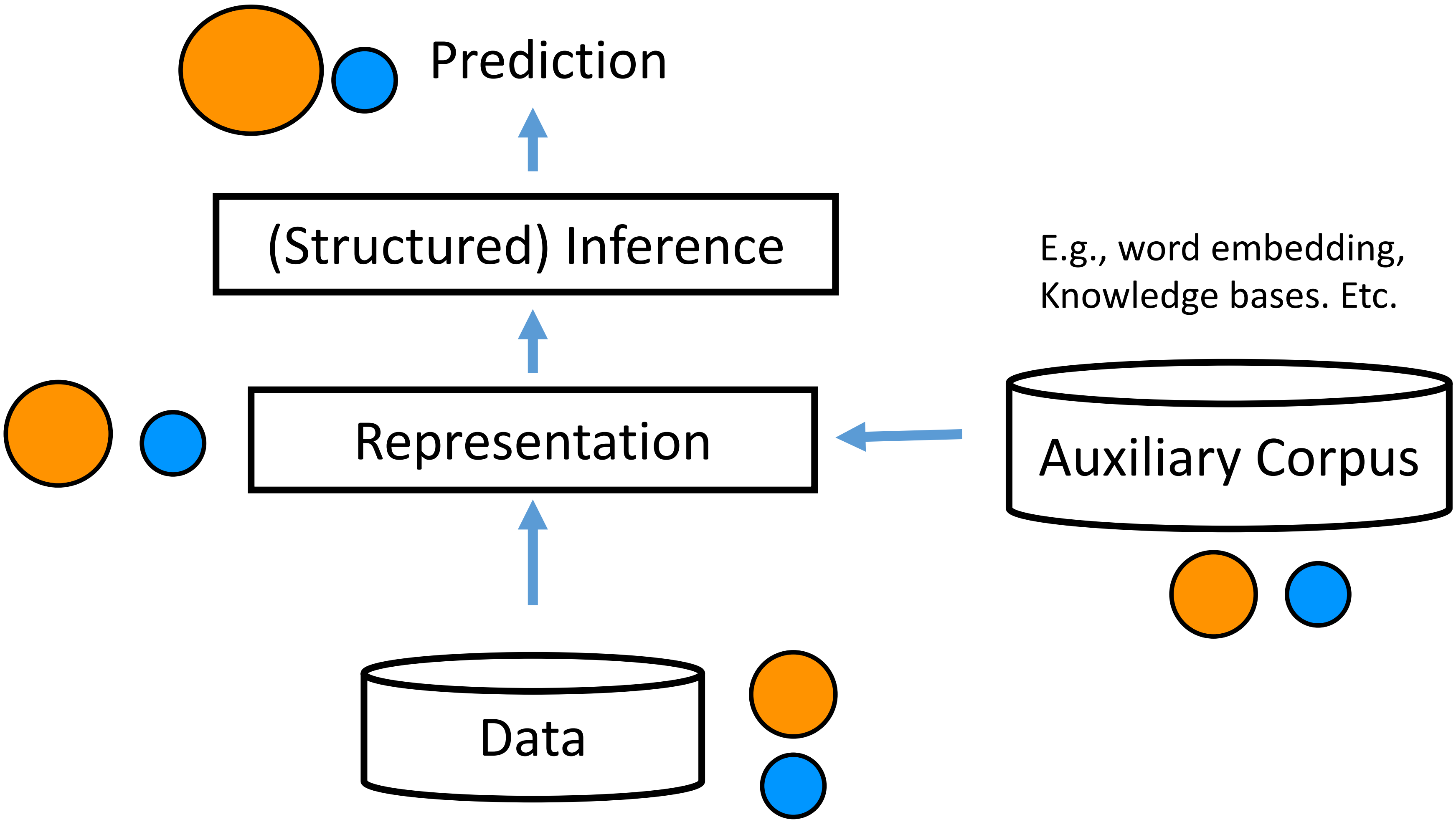
Parse

```
(ROOT
  (S
    (NP (JJ Gender) (NN discrimination))
    (VP (VBZ is)
      (ADJP (JJ illegal)
        (PP (IN in)
          (NP (DT the) (NNP United) (NNPS States)))))
    (. .)))
```

Stanford Parser

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

A carton of ML (NLP) pipeline



Outline

- Controlling Gender Bias in Representation Level
  A study of removing bias in Word Embedding

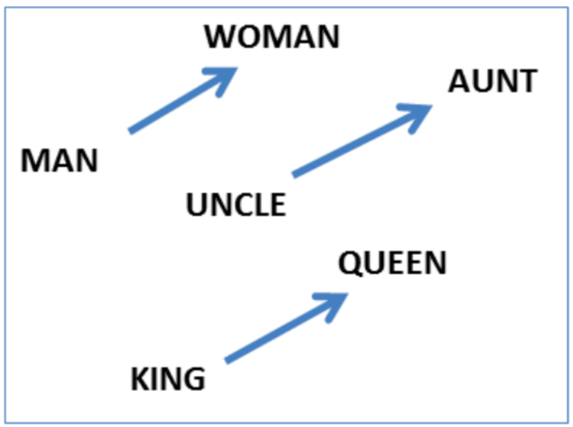
- Reducing Gender Bias in Data Level
  A case study on co-reference resolution

- Reducing Gender Bias in Inference Level
  Guiding predictions by corpus-wise constraints

Word Embeddings can be Dreadfully Sexist [nips16]

w/ Tolga Bolukbasi, James Zou, Venkatesh Saligrama, Adam Kalai

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



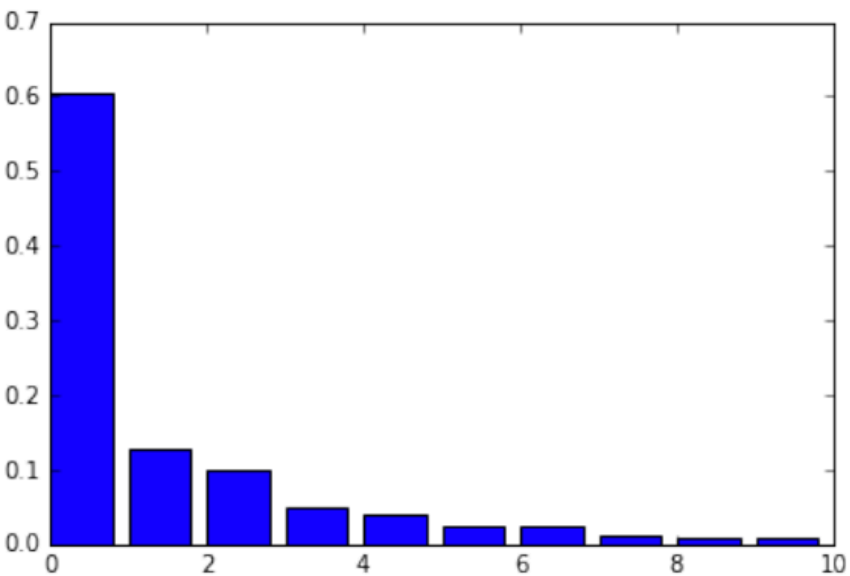
he: ________	she:_______
brother	sister
barbershop	salon
beer	cocktail
physician	registered_nurse
programmer	homemaker
professor	associate professor

We use Google w2v embedding trained from the news

Geometry of Gender and Bias

❖ Identifying the gender subspace

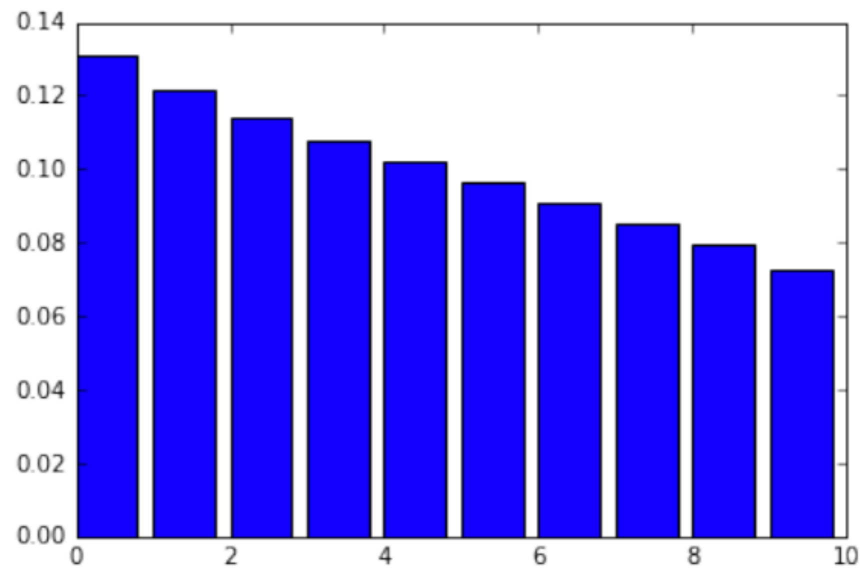
Top 10 Eigenvalue



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

Gender Pair

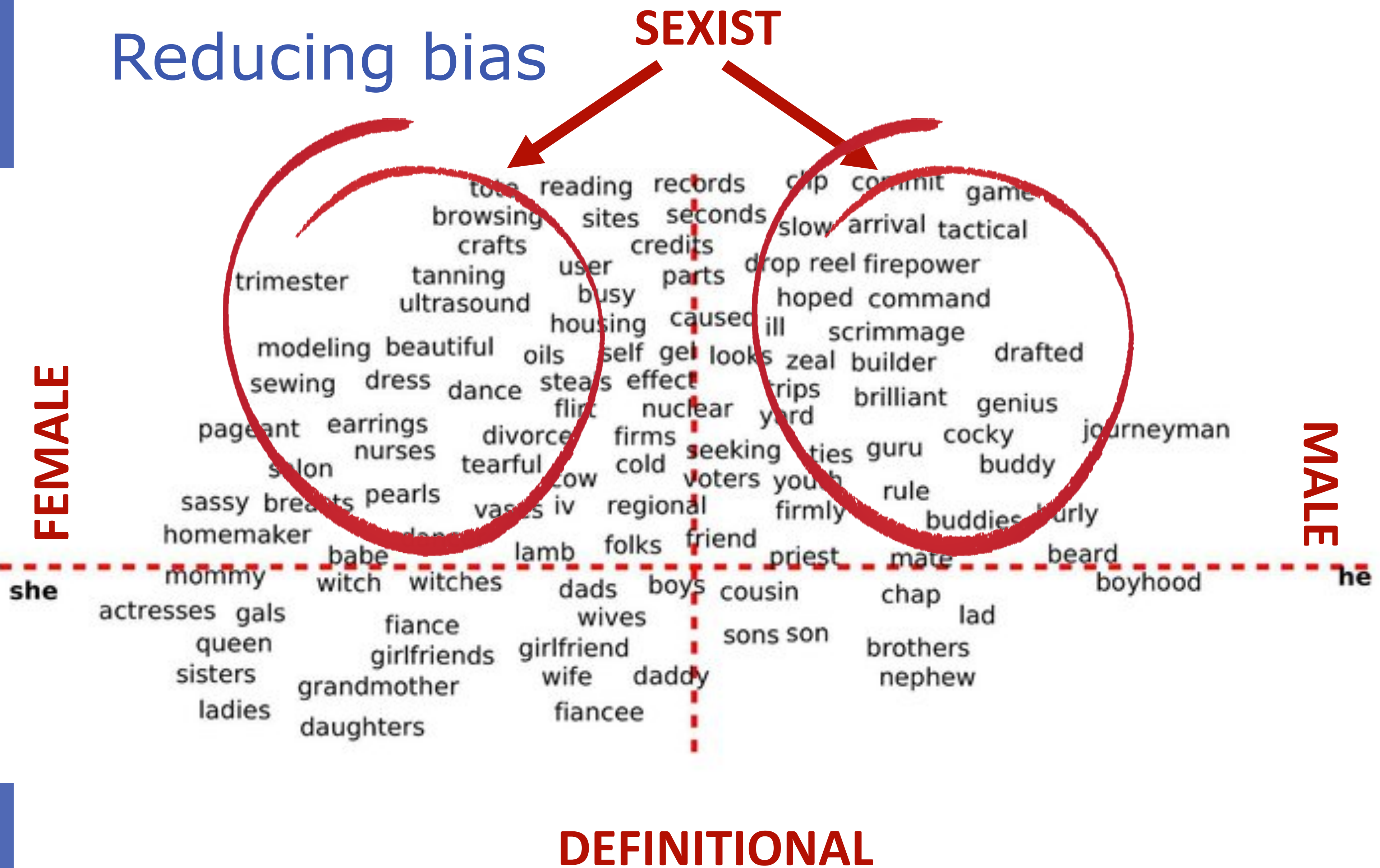
Top 10 Eigenvalue



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

Random Pair

Reducing bias



~~SEXIST~~

tote
browsing
tanning
scrimmage
dress
sewing
brilliant
nurse
cocky
genius
homemaker

FEMALE

MALE

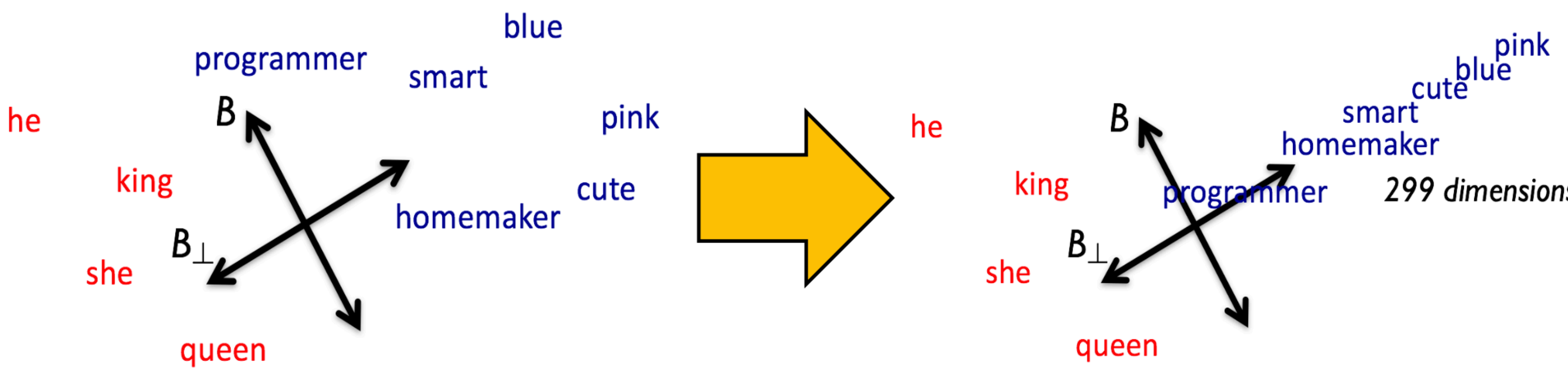
DEFINITIONAL

(related [Schmidt ‘15])

Approach 1:

Project out gender dimension (hard version)

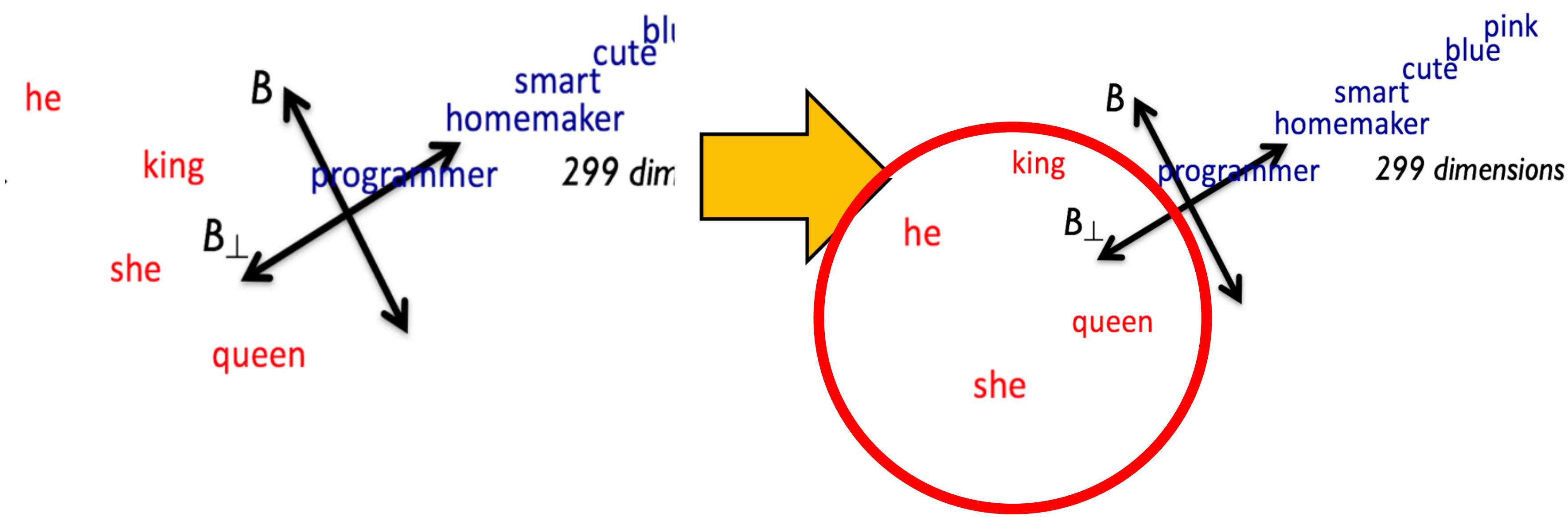
- Step 1: Remove gender dimension from gender-neutral words



Approach 1: Post-processing (Hard)

Project Out Gender Dimension

- Step 2: re-center gender-definitional pairs



Approach 2: Post-processing (Soft)

Find a linear transformation T of the gender-neutral words to reduce the gender component while not moving the words too much.

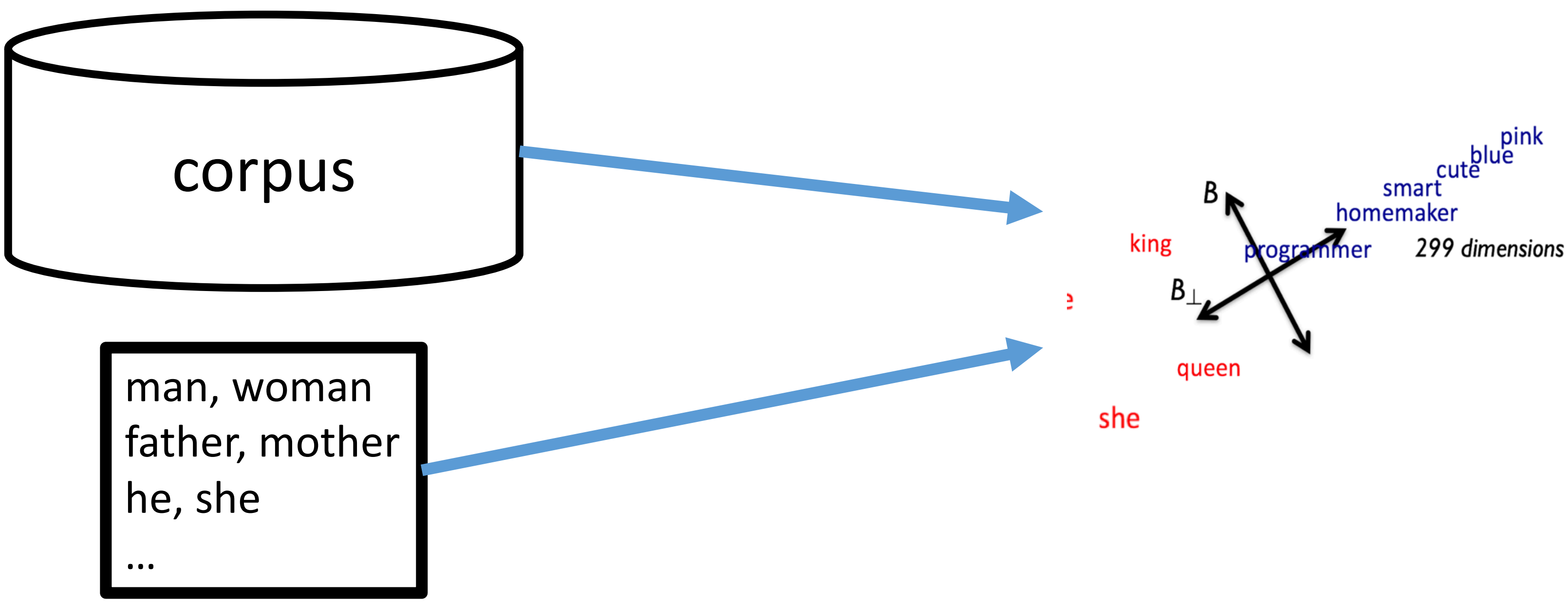
$W =$ matrix of all word vectors.

$N =$ matrix of neutral word vectors.

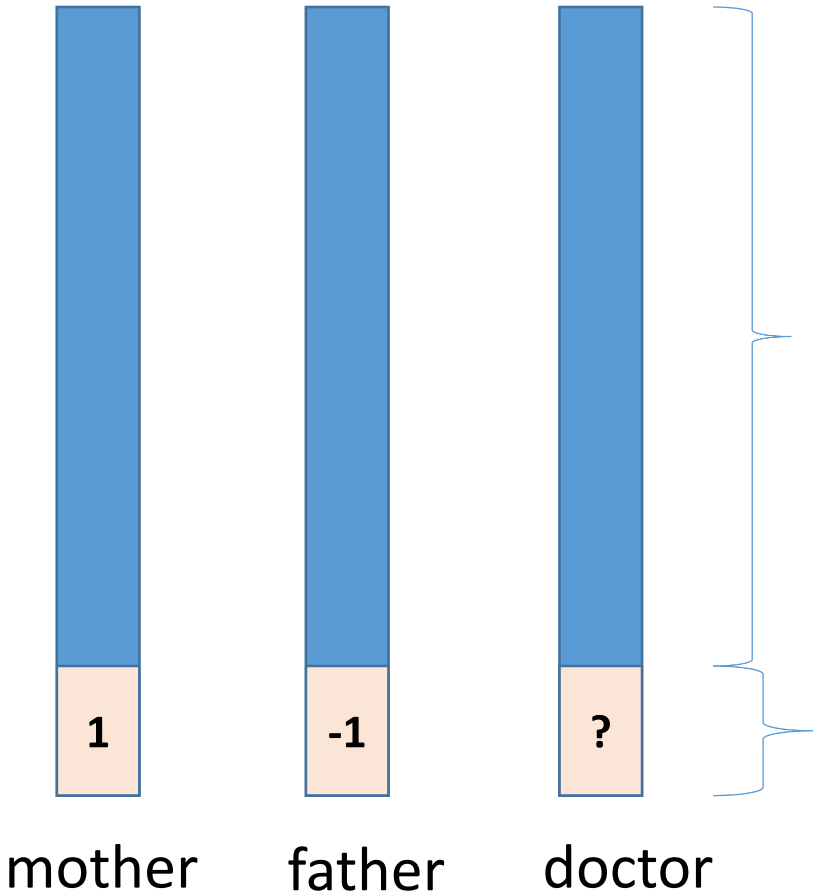
$$\min_T \underbrace{||(TW)^T(TW) - W^TW||_F^2}_{\text{don't move too much}} + \lambda \underbrace{||(TN)^T(TB)||_F^2}_{\text{minimize gender component}}$$

Approach 3: Learning Gender-Neutral Word Embedding [Jieyu+EMNLP18]

- How can we **"not to"** encode gender information in word vectors?

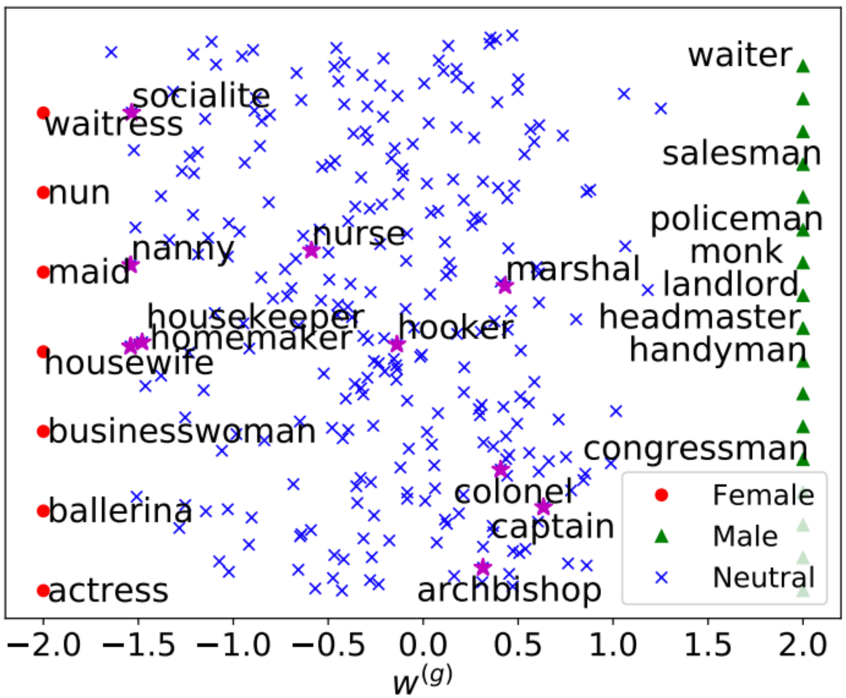


Approach 3:
Learning Gender-Neutral Word Embeddings
[Jieyu+ EMNLP18]



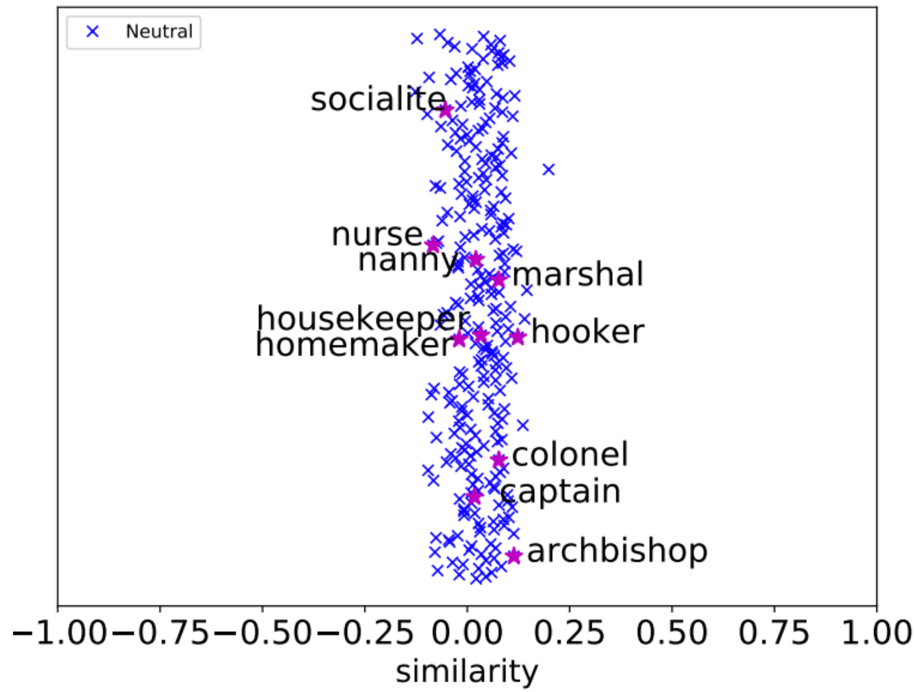
dimensions for other latent aspects w^N

dimensions reserve for gender information



waiter
socialite
waitress
salesman
nun
policeman
monk
nurse
nanny
maid
marshal
landlord
housekeeper
headmaster
homemaker
hooker
handyman
housewife
businesswoman
congressman
colonel
ballerina
captain
archbishop
actress
Female
Male
Neutral
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
$w^{(g)}$

w^g



Neutral
socialite
nurse
nanny
marshal
housekeeper
homemaker
hooker
colonel
captain
archbishop
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
similarity

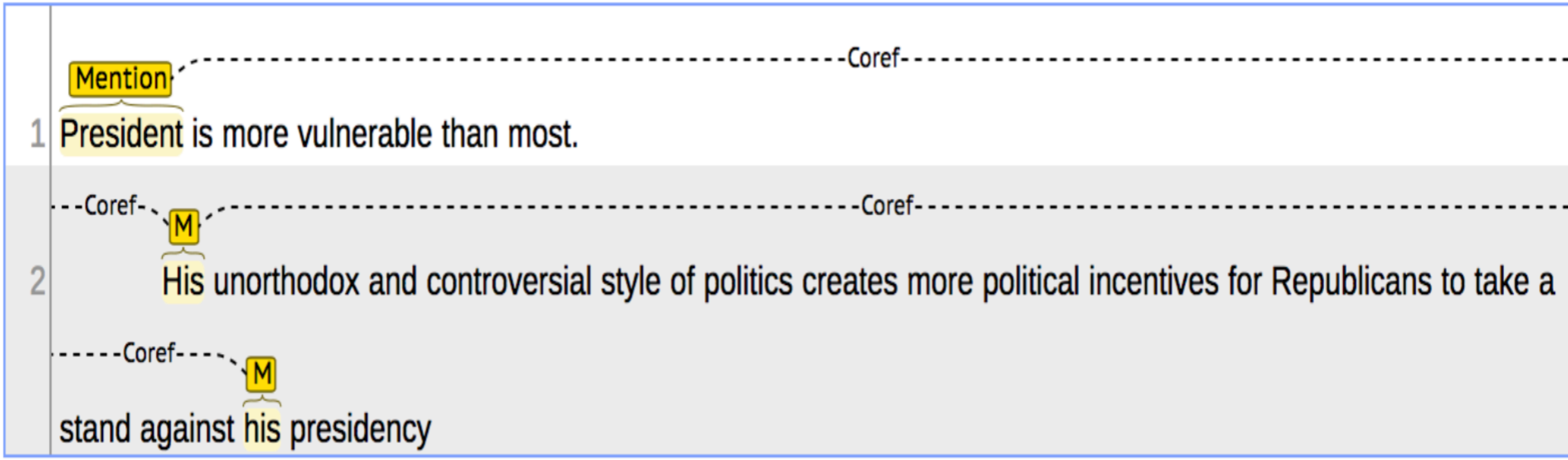
w^N

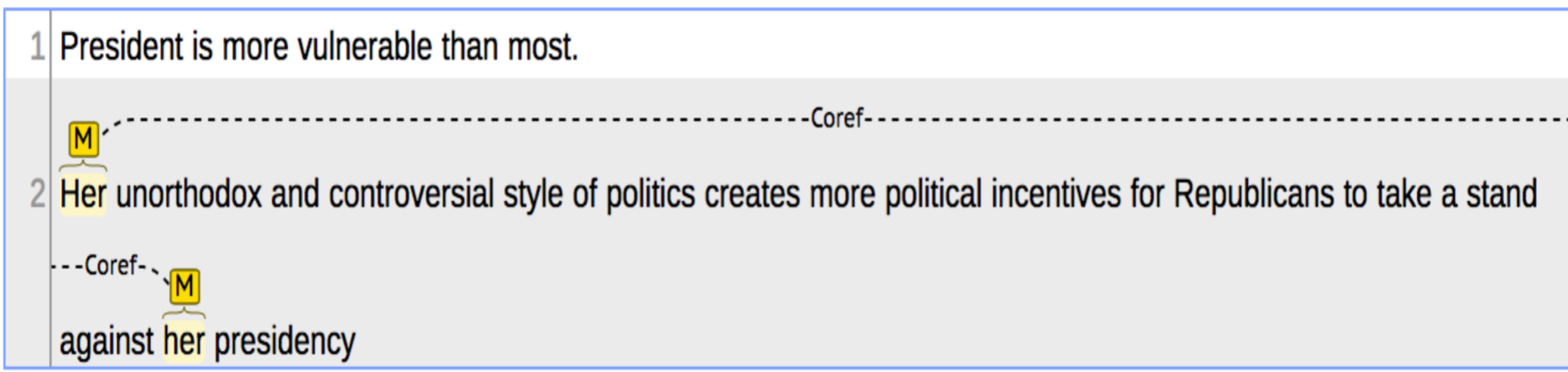
Are these debiased vectors actually useful?

Outline

- Controlling Gender Bias in Representation Level
  A study of removing bias in Word Embedding
- Reducing Gender Bias in Data Level
  A case study on co-reference resolution
- Reducing Gender Bias in Inference Level
  Guiding predictions by corpus-wise constraints

Gender Bias in Coref [NAACL 2018]





Concurrent work (Rudinger et al., 2018) @NAACL18 also studied gender bias in Coref.

Wino-bias data

- Stereotypical dataset

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

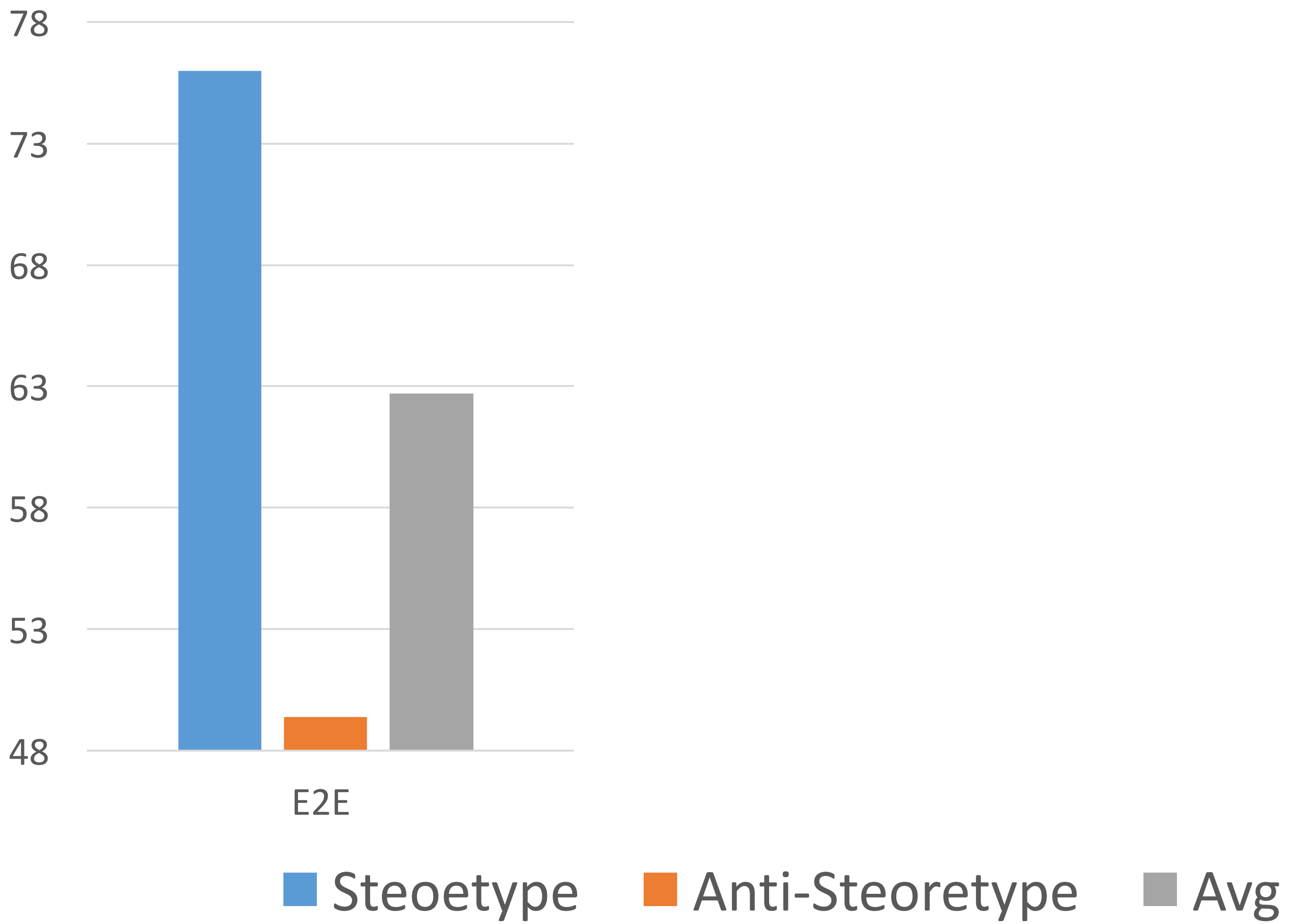
The physician hired the secretary because she was highly recommended.

- Anti-stereotypical dataset

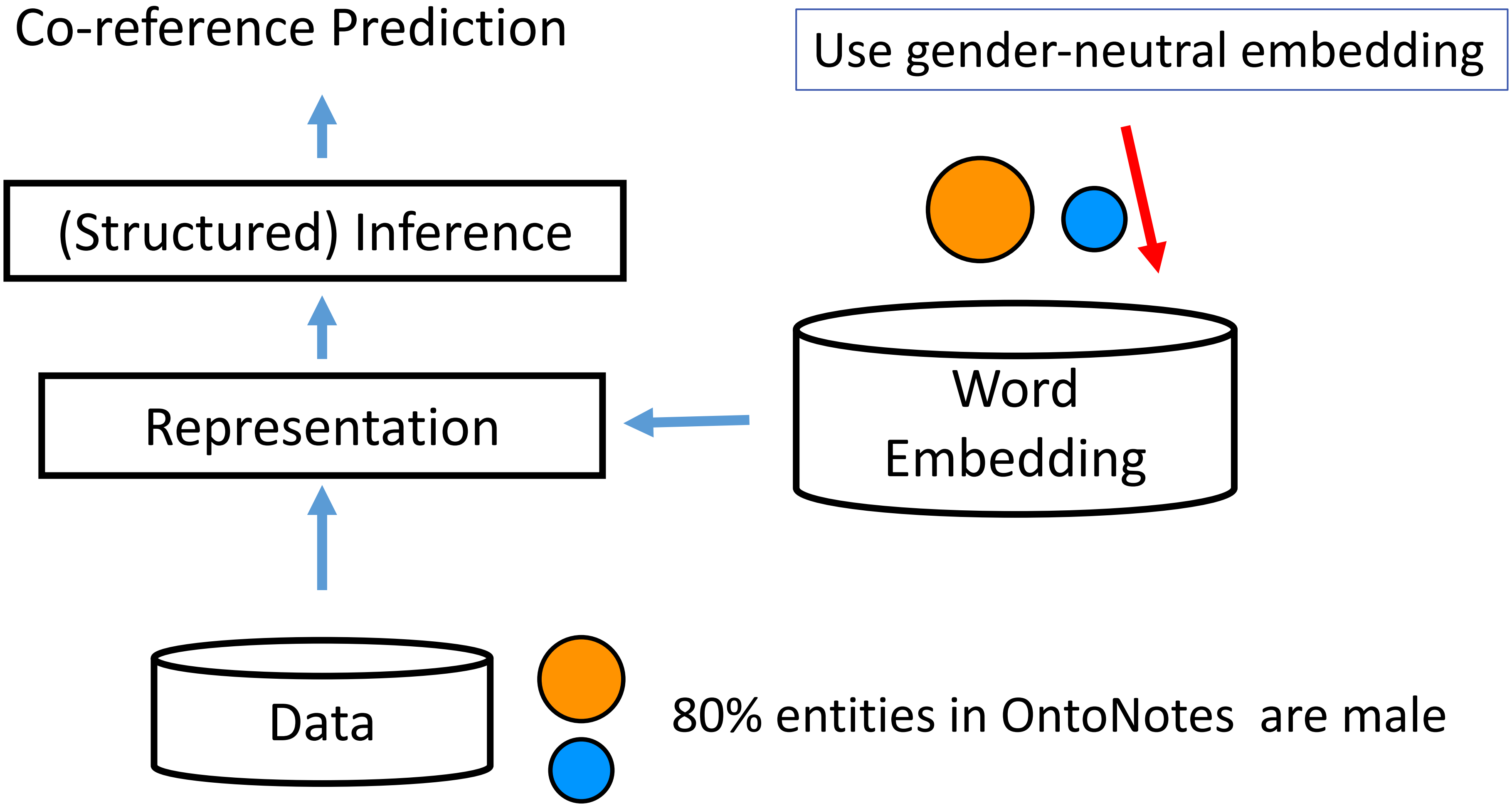
The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because he was highly recommended.

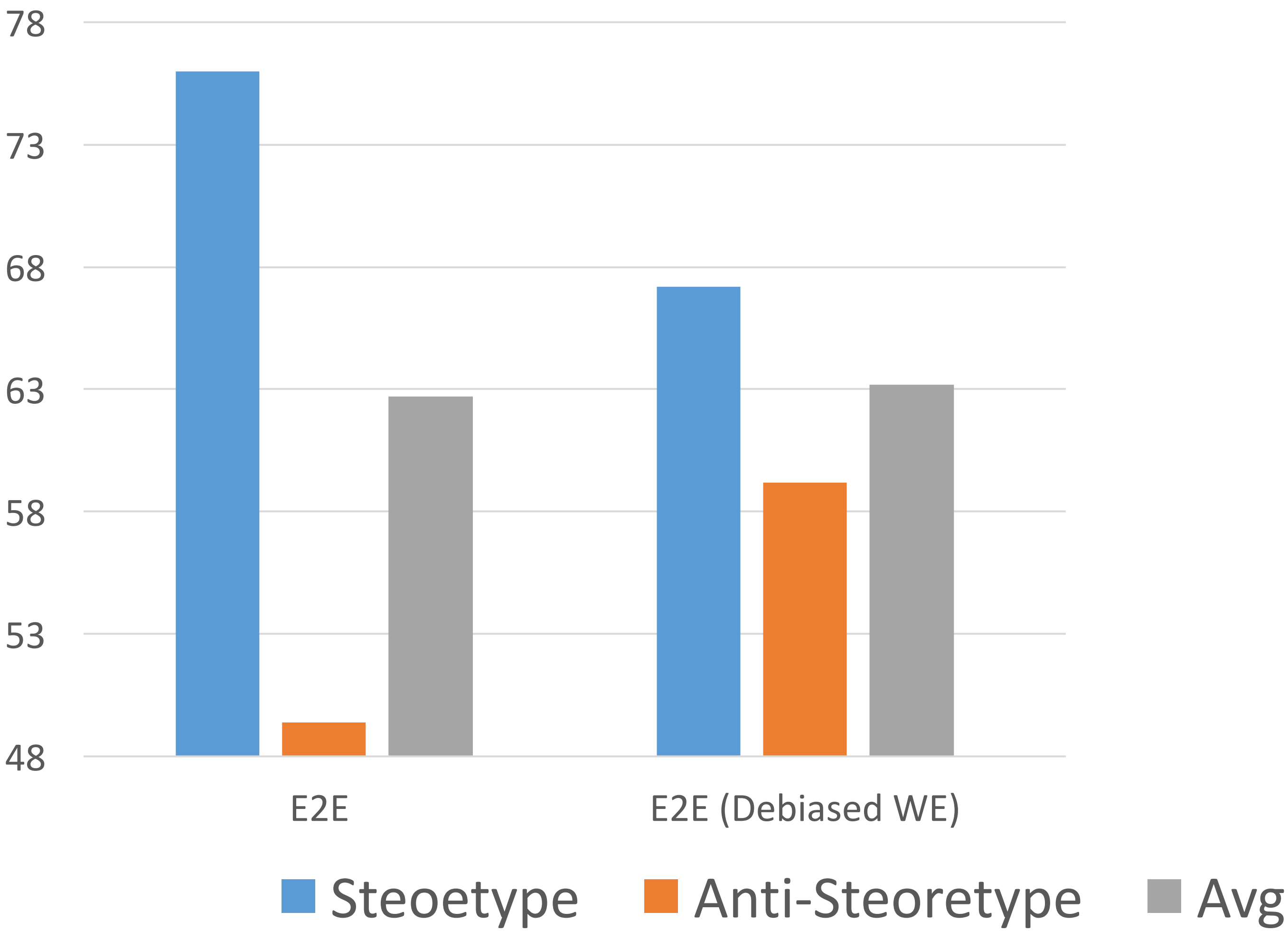
Gender bias in Coref System



Source of gender bias



Gender bias in Coref System



How to deal with bias in data

- Idea: simulate sentence in opposite gender

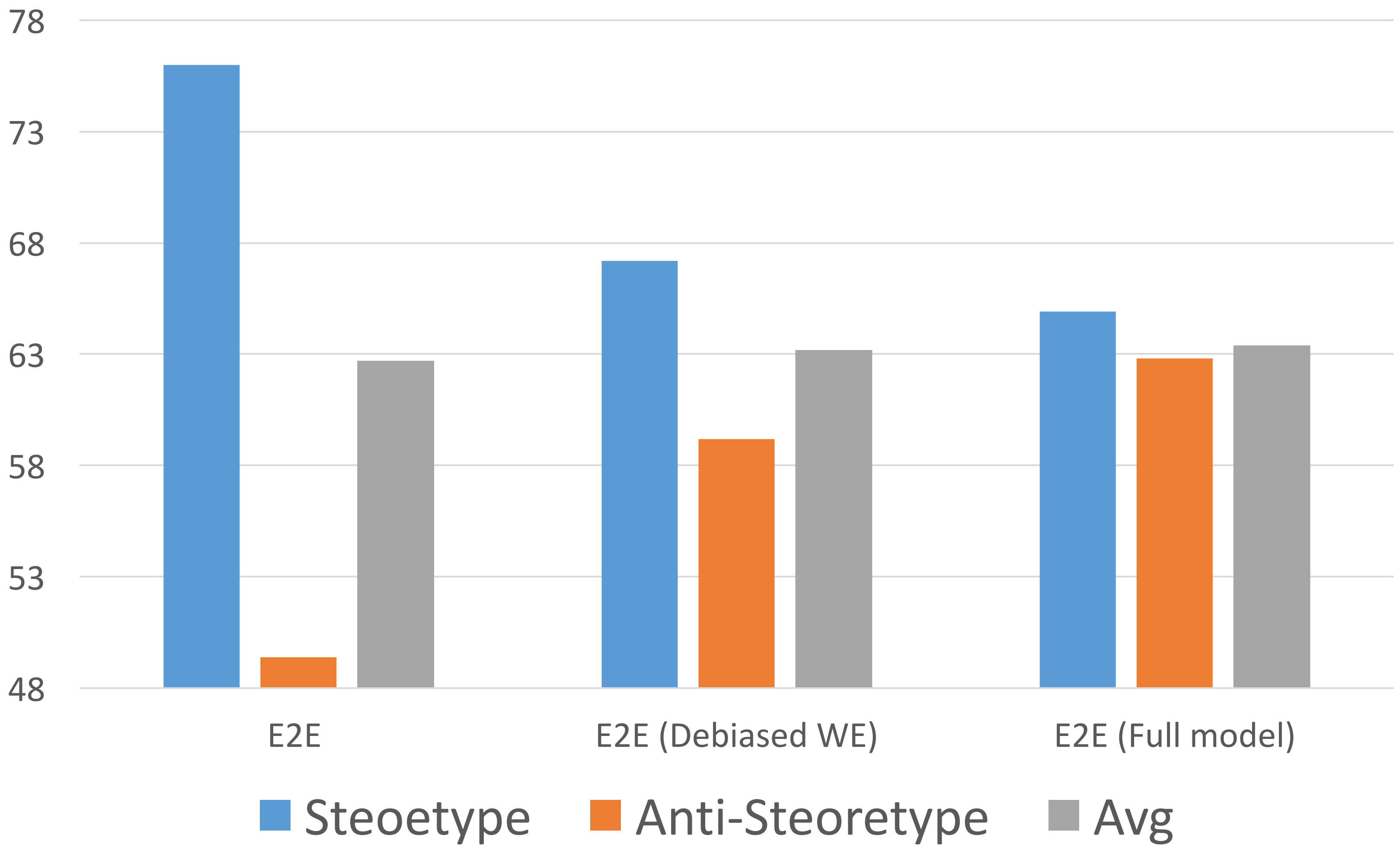
John went to his house

F2 went to her house

Named Entity are anonymized

Gender words are swapped

Gender bias in Coref System



Outline

- Controlling Gender Bias in Representation Level
  A study of removing bias in Word Embedding

- Reducing Gender Bias in Data Level
  A case study on co-reference resolution

- Reducing Gender Bias in Inference Level
  Guiding predictions by corpus-wise constraints

Human Bias in Structured Prediction Models

[EMNLP 17*] w/ Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez

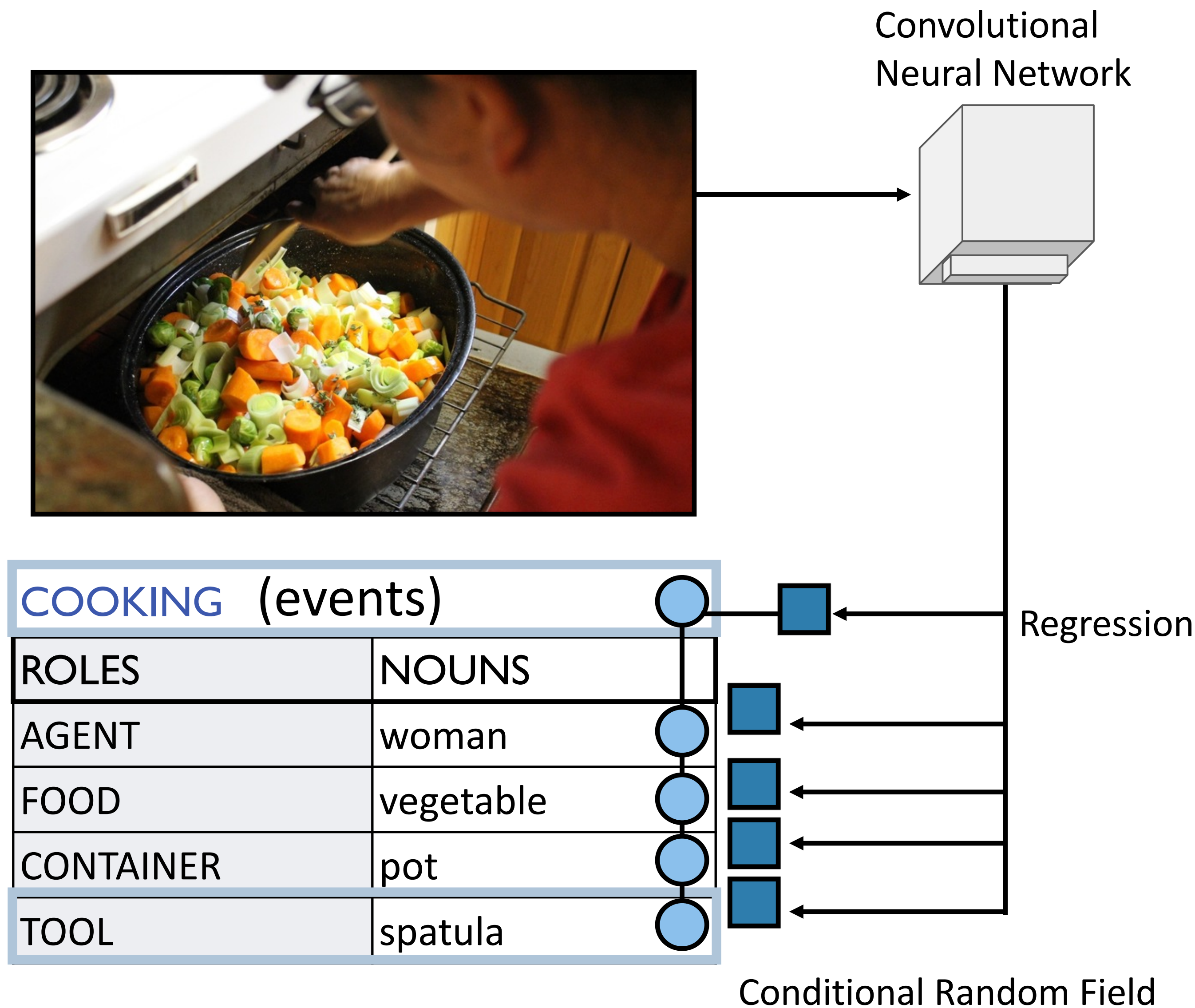
What's the agent for this image?

Cooking	
Role	Object
agent	?
food	vegetable
container	bowl
tool	knife
place	kitchen

An example from a vSRL (visual Semantic Role Labeling) system

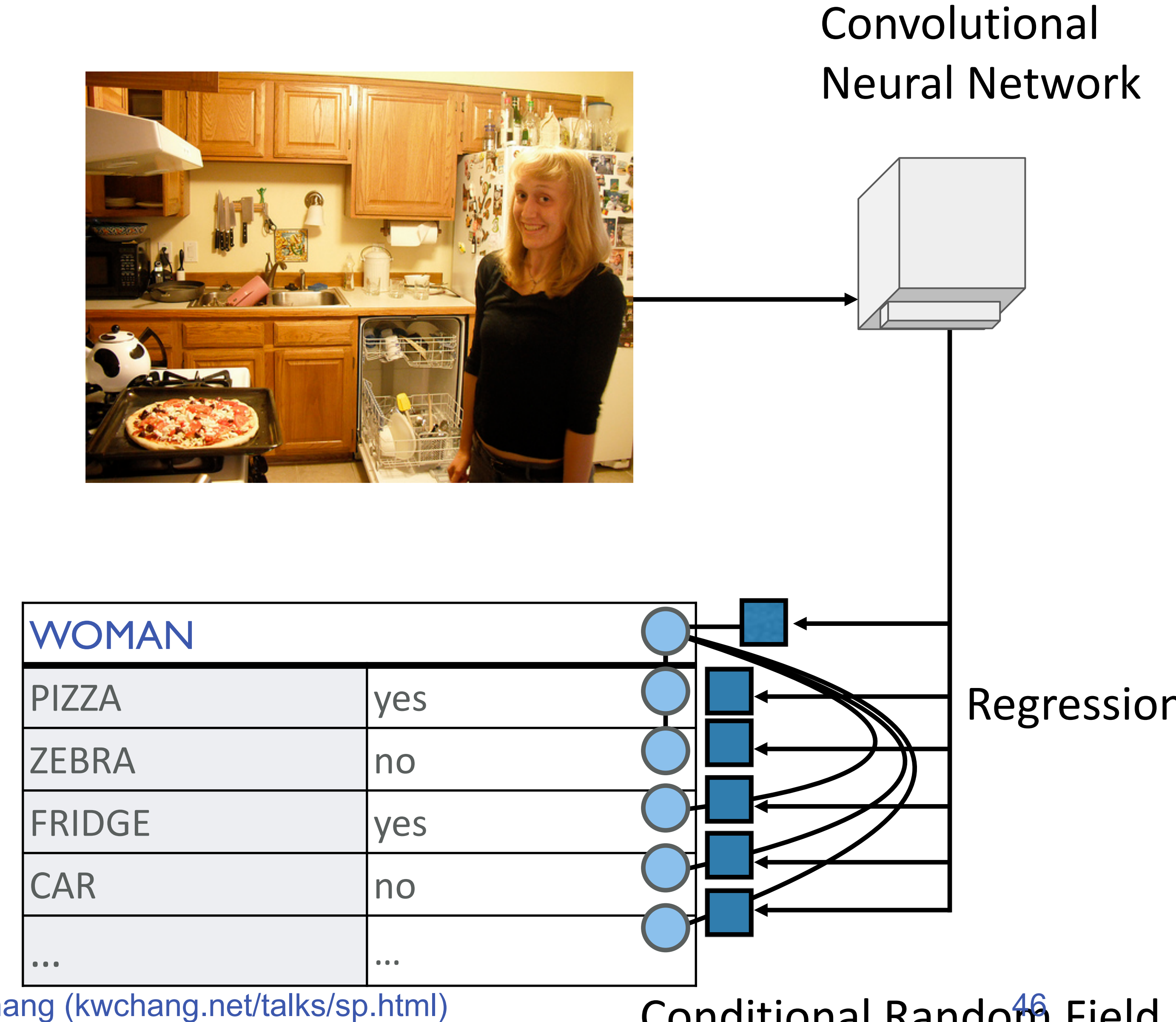
*Best Long Paper Award at EMNLP 17

imSitu Visual Semantic Role Labeling (vSRL)



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

COCO Multi-Label Classification (MLC)



Defining Dataset Bias (events)

Training Gender Ratio (◆ verb)

Training Set

- ◆ cooking
- ● woman
- ● man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

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

Defining Dataset Bias (objects)

Training Gender Ratio (▲ noun)

Training Set

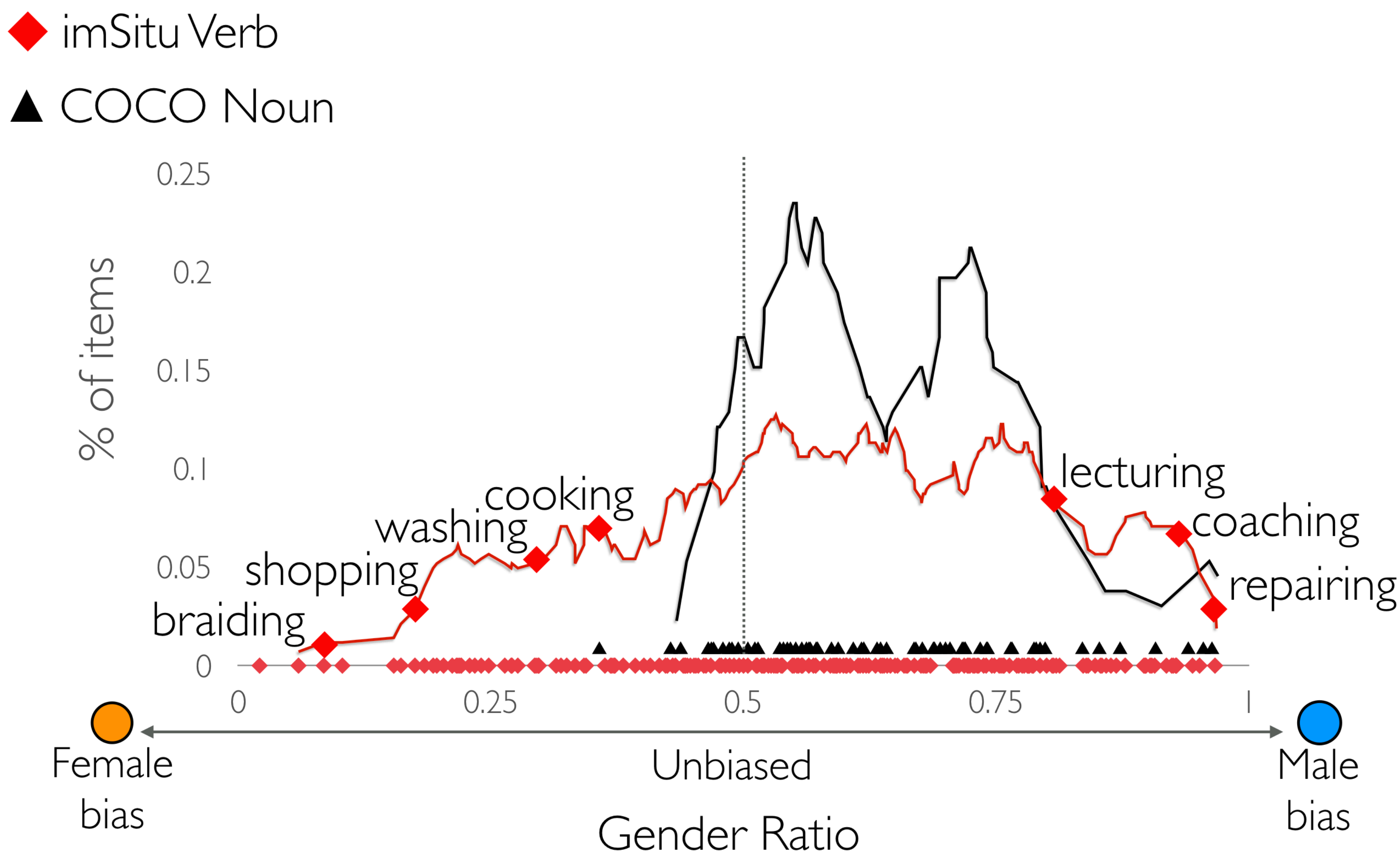
- ▲ snowboard
- ● woman
- ● man

MAN	
snowboard	yes
refrigerator	no
bowl	no

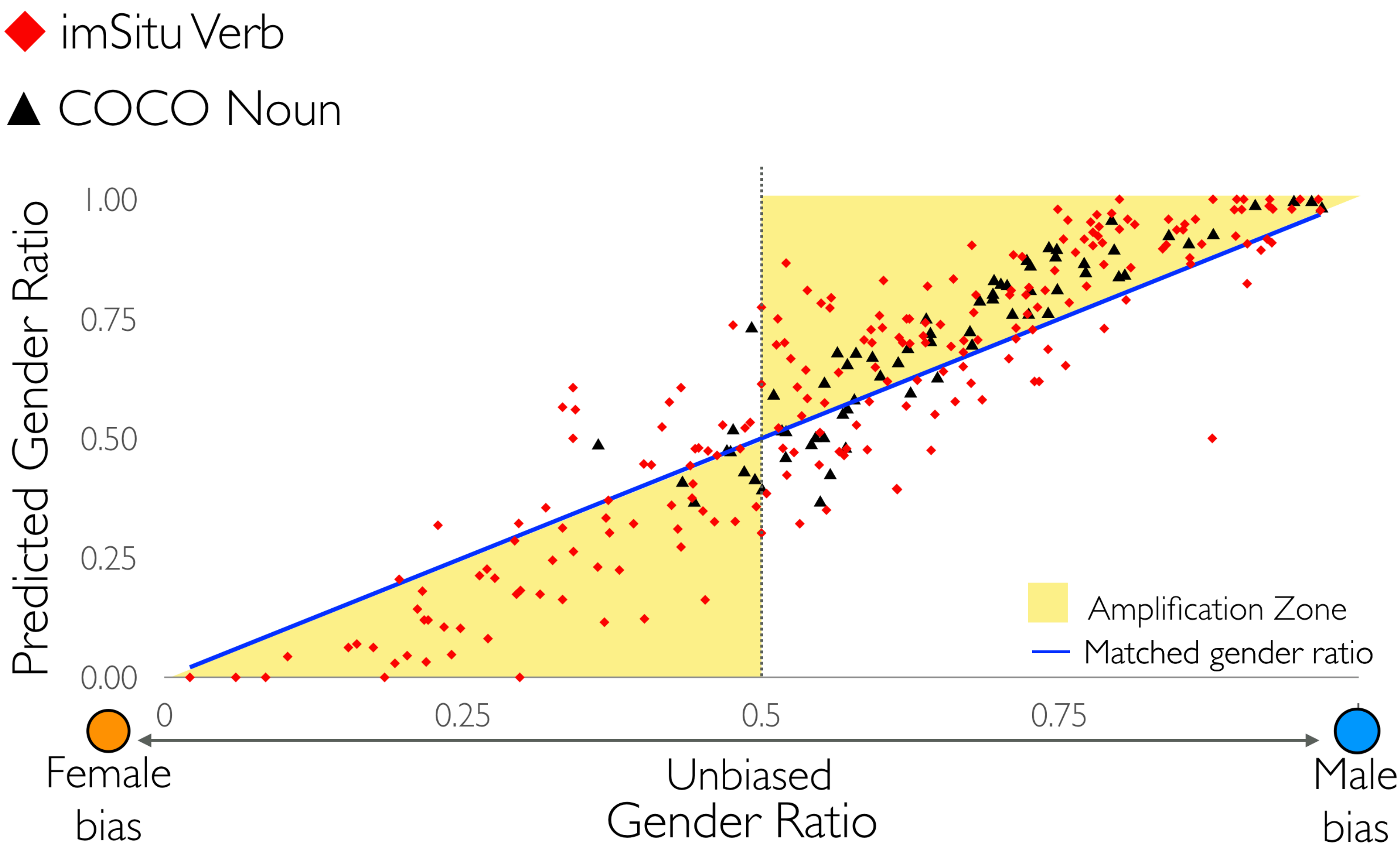
WOMAN	
snowboard	yes
refrigerator	no
bowl	no

$$\frac{\#(\blacktriangle\ \text{snowboard},\ \bullet\ \text{man})}{\#(\blacktriangle\ \text{snowboard},\ \bullet\ \text{man}) + \#(\blacktriangle\ \text{snowboard},\ \bullet\ \text{woman})} = 2/3$$

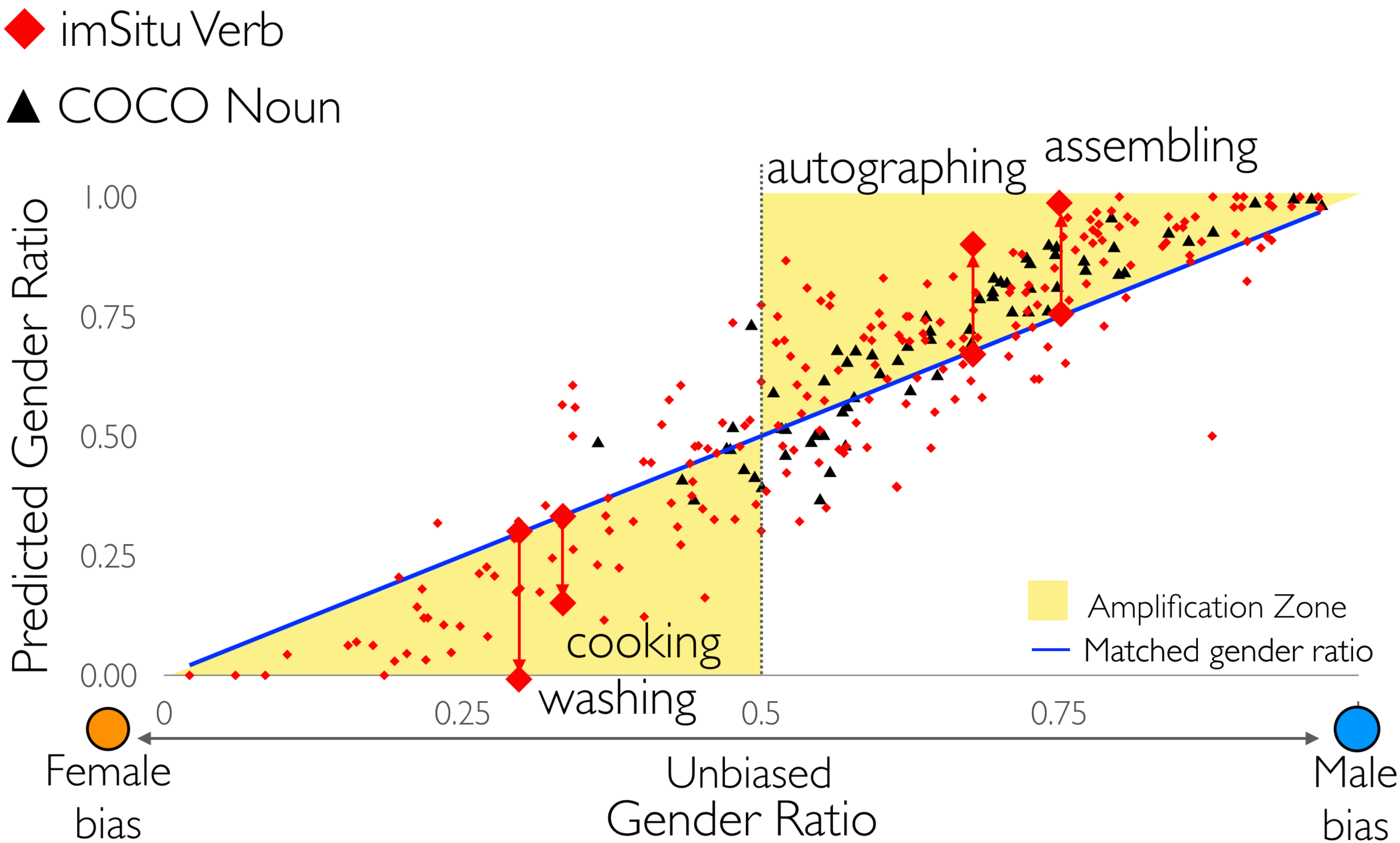
Gender Dataset Bias



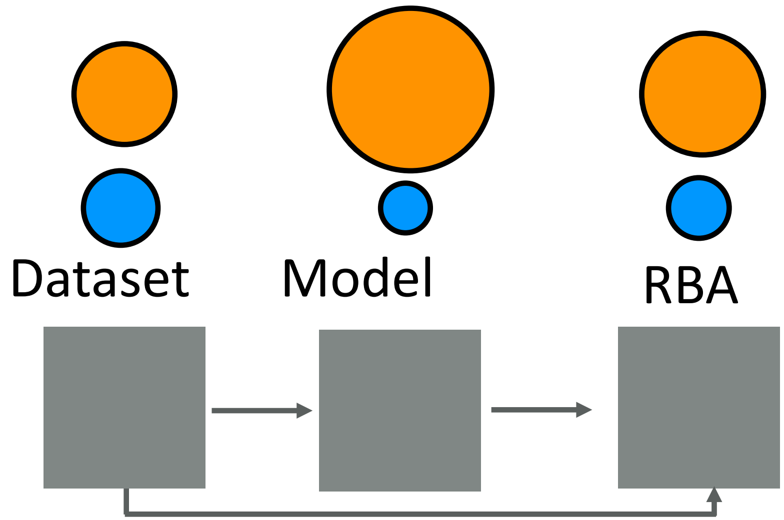
Model Bias Amplification



Model Bias Amplification



Reducing Bias Amplification (RBA)



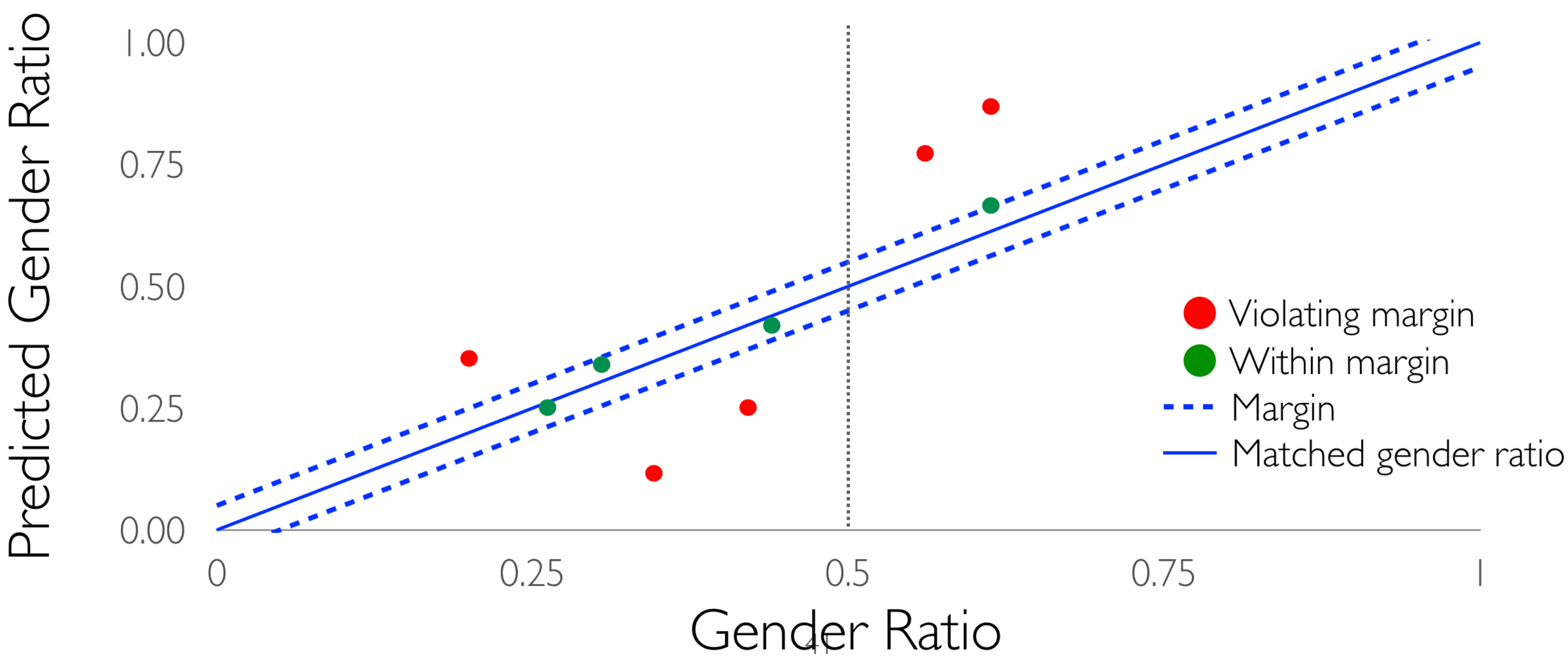
- Corpus level constraints on model output (ILP)
  - Doesn't require model retraining
- Reuse model inference through Lagrangian relaxation
  - Can be applied to any structured model

Reducing Bias Amplification (RBA)

Integer Linear Program

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

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

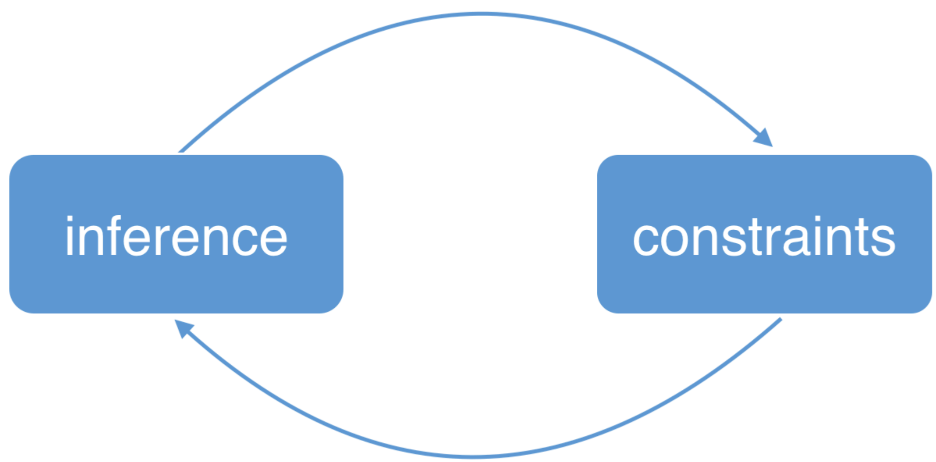


Reducing Bias Amplification (RBA)

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

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

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



Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Reducing Bias Amplification (RBA)

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

$\forall$ points $\;\left|\text{Training Ratio} - \text{Predicted Ratio}\right| <= \text{margin}$

$$\max_{\{y^i\}\in\{Y^i\}} \sum_i f_\theta(y^i, i), \quad \text{s.t.} \quad A\sum_i y^i - b \le 0$$

Lagrangian : $\sum_i f_\theta(y^i) - \sum_{j=1}^{l} \lambda_j (A_j \sum_i y^i - b_j) \qquad \lambda_j \ge 0$

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

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

$$\left|\underset{(1/2)}{\text{Training Ratio}} - \text{Predicted Ratio}\right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

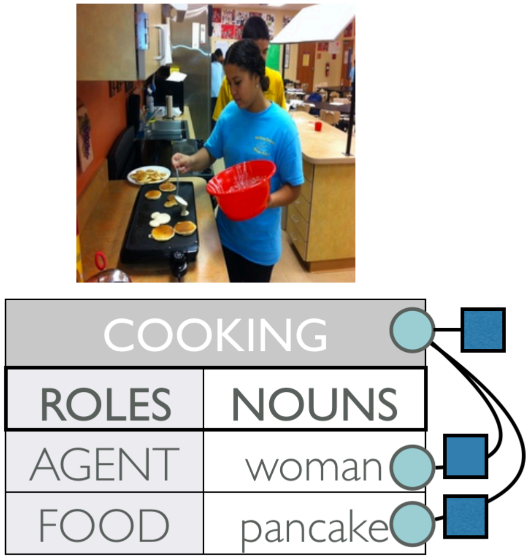
inference

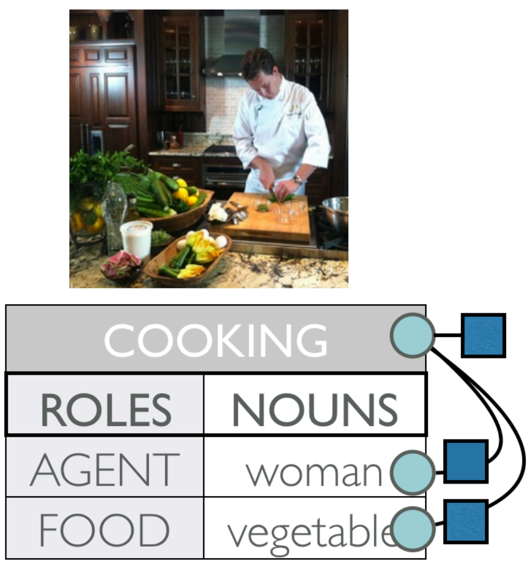
update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation





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

$$\left| \text{Training Ratio} \ - \text{Predicted Ratio} \right| <= \ \text{margin}$$

(1/2)

- Lagrange Multiplier (λ) Per Constraint

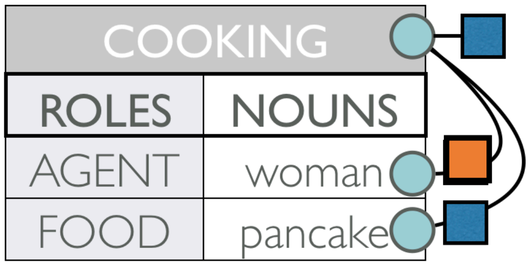
inference

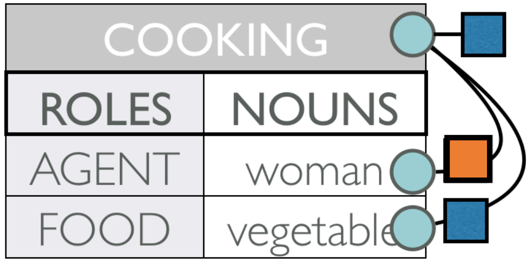
update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation





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

$$\left|\underset{(1/2)}{\text{Training Ratio}} - \text{Predicted Ratio}\right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

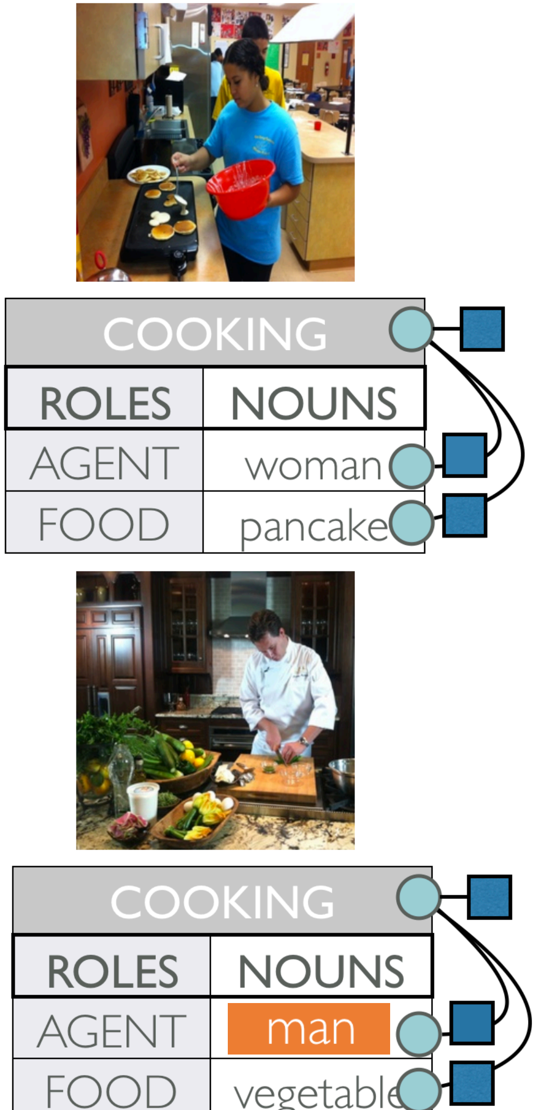
inference

update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

Lagrangian Relaxation



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

$$\left| \underset{(1/2)}{\text{Training Ratio}} - \text{Predicted Ratio} \right| <= \text{margin}$$

- Lagrange Multiplier (λ) Per Constraint

$\boxed{\text{inference}}$

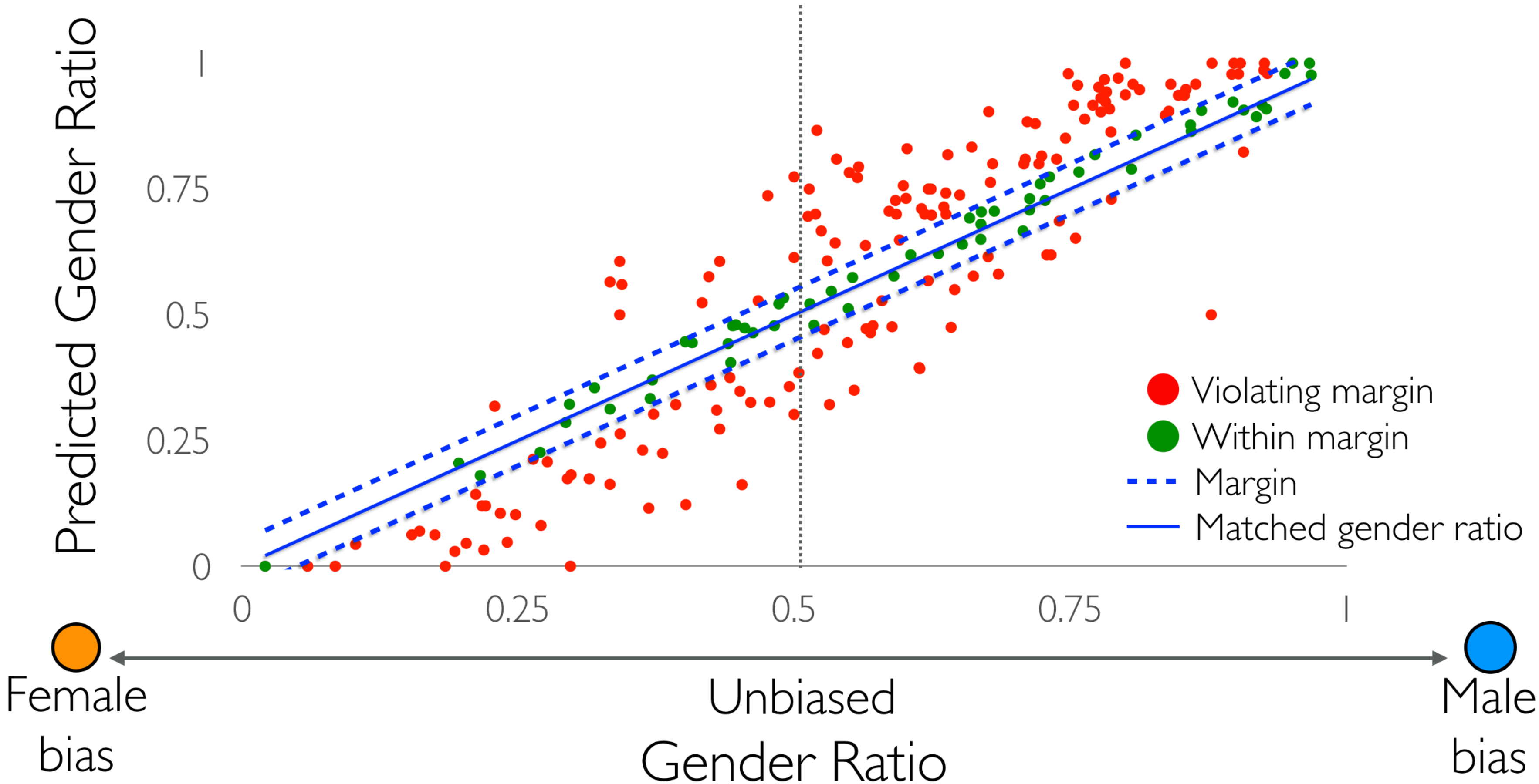
update λ

update potentials

Sontag et al., 2011; Rush and Collins, 2012; Chang and Collins, 2011; Peng et al., 2015, Chang et al., 2013; Dalvi, 2015

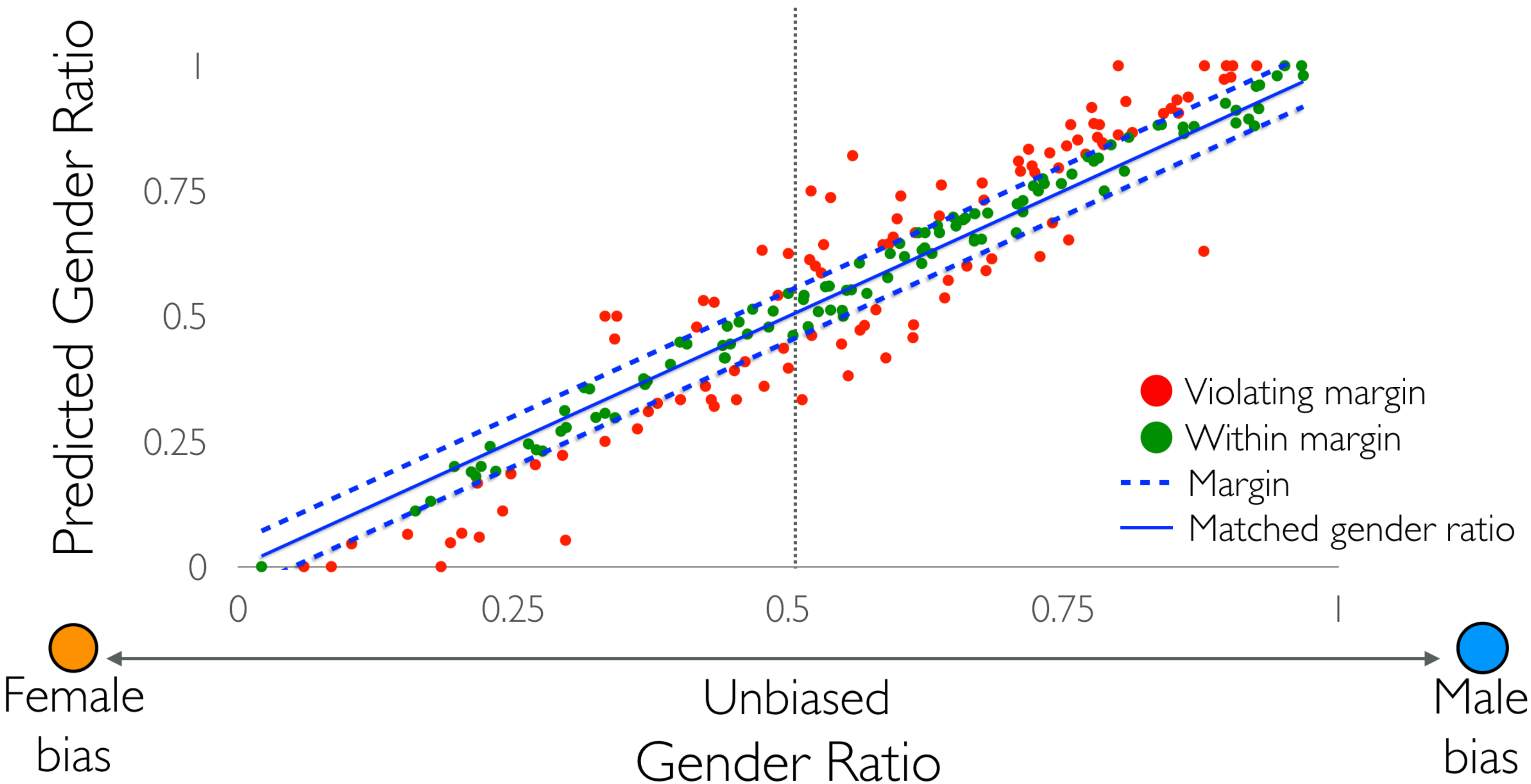
Gender Bias De-amplification in imSitu

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

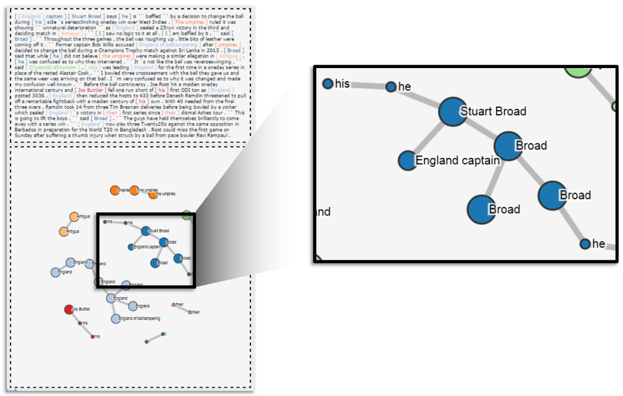


Gender Bias De-amplification in imSitu

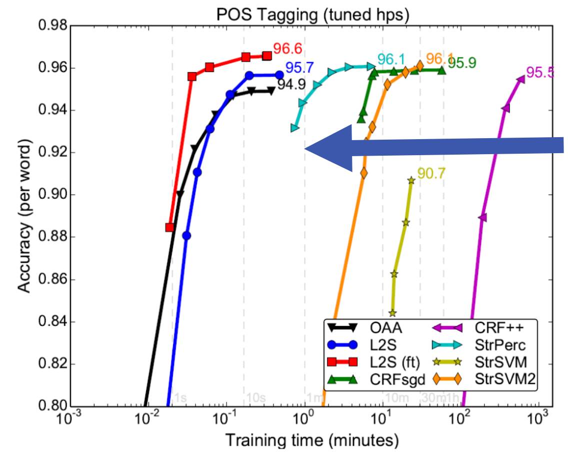
imSitu Verb	Violation: 72.6%	.050 |bias↑|	24.07 acc.
w/ RBA	Violation: 50.5%	.024 |bias↑|	23.97 acc.



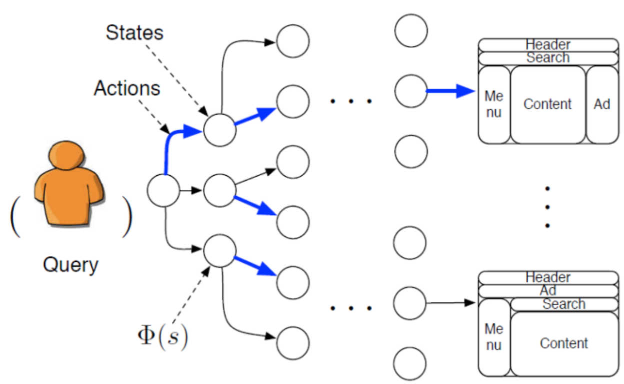
UCLA NLP



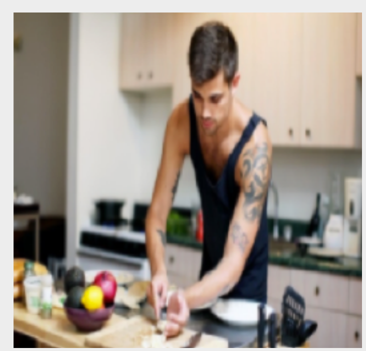
NLP Applications



Efficient Algorithms



Learning from weak signals



activity	cooking
agent	woman
food	vegetable

Fairness (data biases)

Conclusions

- Like other AI systems, NLP systems affect by societal bias present in data
- The issues cause unfair predictions are not new
  - Domain adaptation / Data collection bias
- Ultimate goal: robust NLP systems for social good
- References: http://kwchang.net