# MITIGATING GENDER BIAS IN NLP: LITERATURE REVIEW

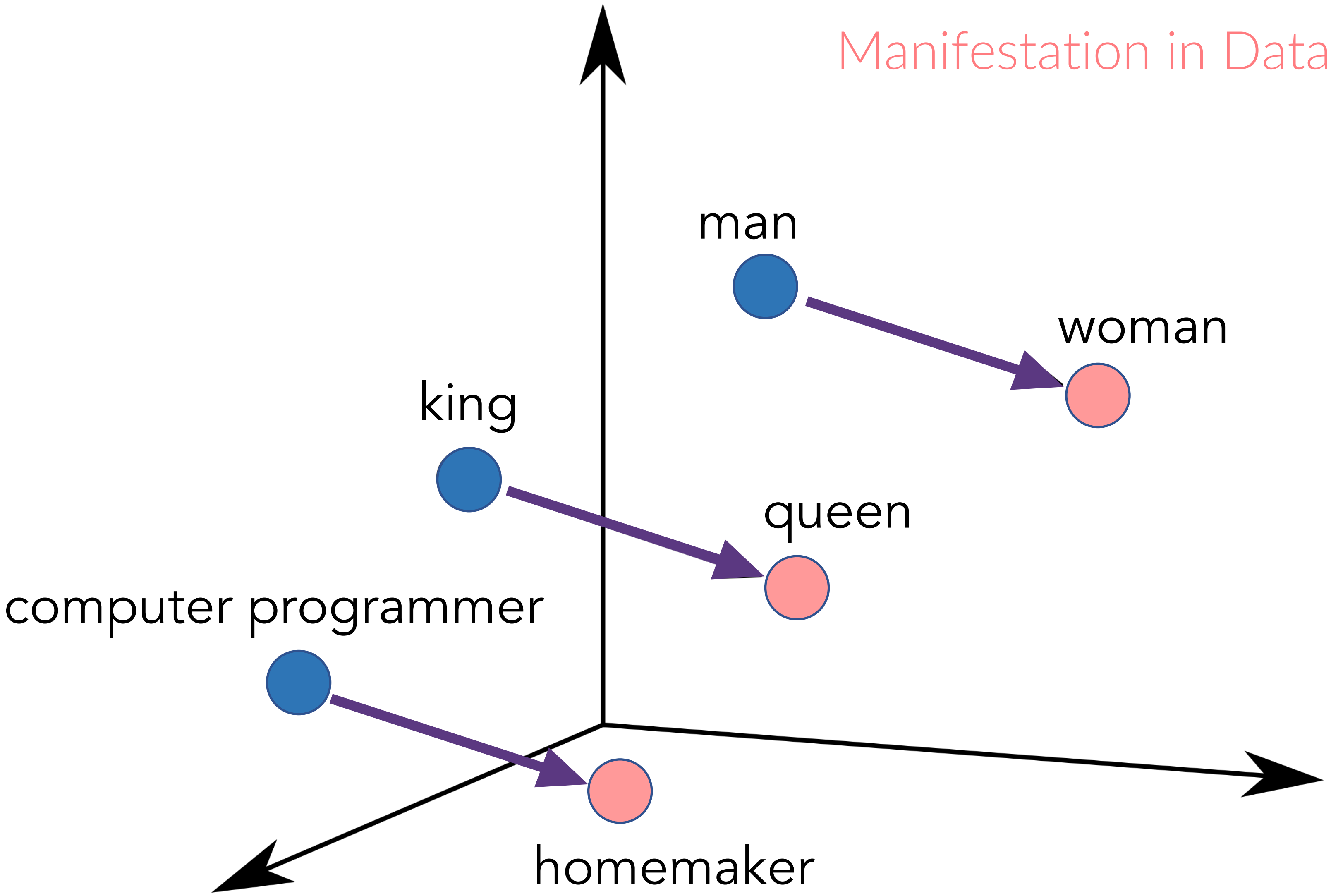
ACL 2019

UC Santa Barbara, UCLA

Tony Sun · Andrew Gaut · Shirlyn Tang · Yuxin Huang
Mai ElSherief · Jieyu Zhao · Diba Mirza · Elizabeth Belding · Kai-Wei Chang · William Yang Wang

# Gender Bias Origins

## Manifestation in Data and Representations



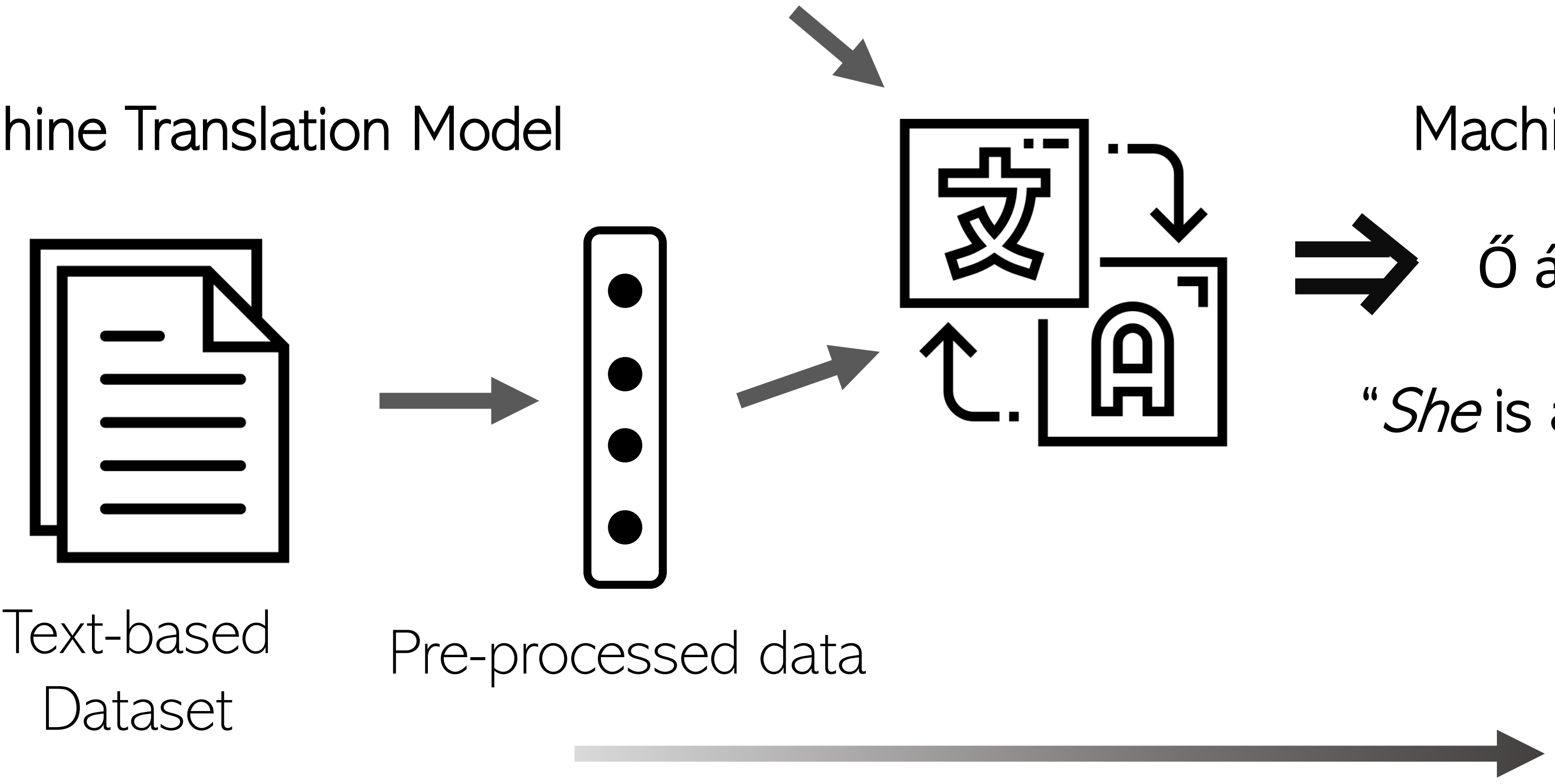
# What is Gender Bias?

“Gender bias is the preference or prejudice toward one gender over another”

Allocation vs Representation Bias

# Gender Bias in Machine Translation

User Input “*He* is a nurse. *She* is a doctor.”



Machine Translation Model

Text-based Dataset

Pre-processed data

Machine Translation Output

Ő ápolónő. Ő egy orvos

“*She* is a nurse. *He* is a doctor.”

Propagation of gender bias

# Gender Bias in NLP

Word Embeddings

Language Modeling

Coreference Resolution

Machine Translation

Speech Recognition

Sentiment Analysis

Caption Generation

## The Pipeline

The Big Picture of Bias

## Identifying Bias

Bias Evaluation Methods

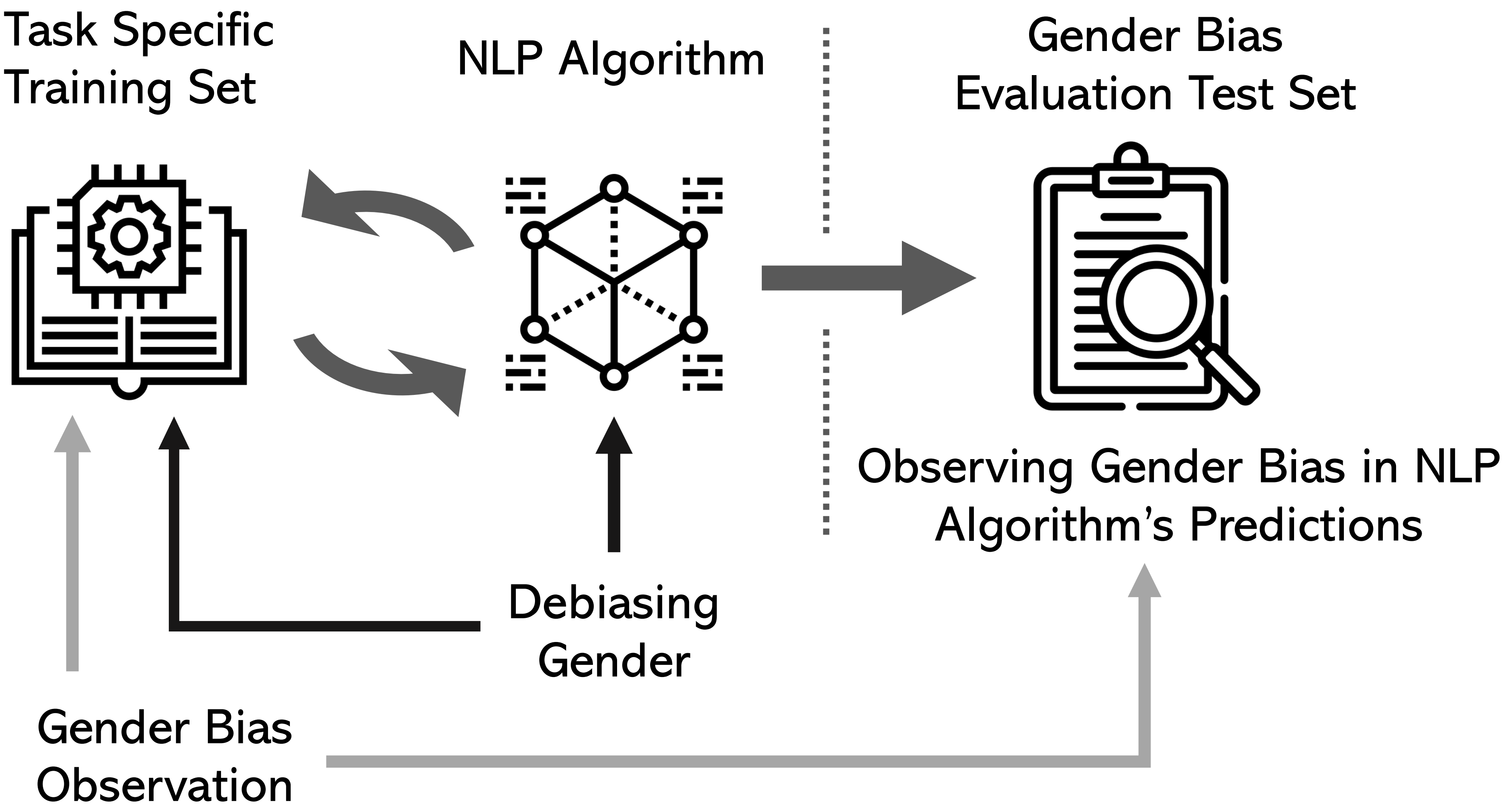
## Mitigating Bias

Gender Debiasing Models

## Conclusion and Future Directions

Ideas Worth Exploring

# Gender Debiasing Pipeline



Task Specific Training Set

NLP Algorithm

Gender Bias Evaluation Test Set

Observing Gender Bias in NLP Algorithm’s Predictions

Debiasing Gender

Gender Bias Observation

# Gender Bias Evaluation

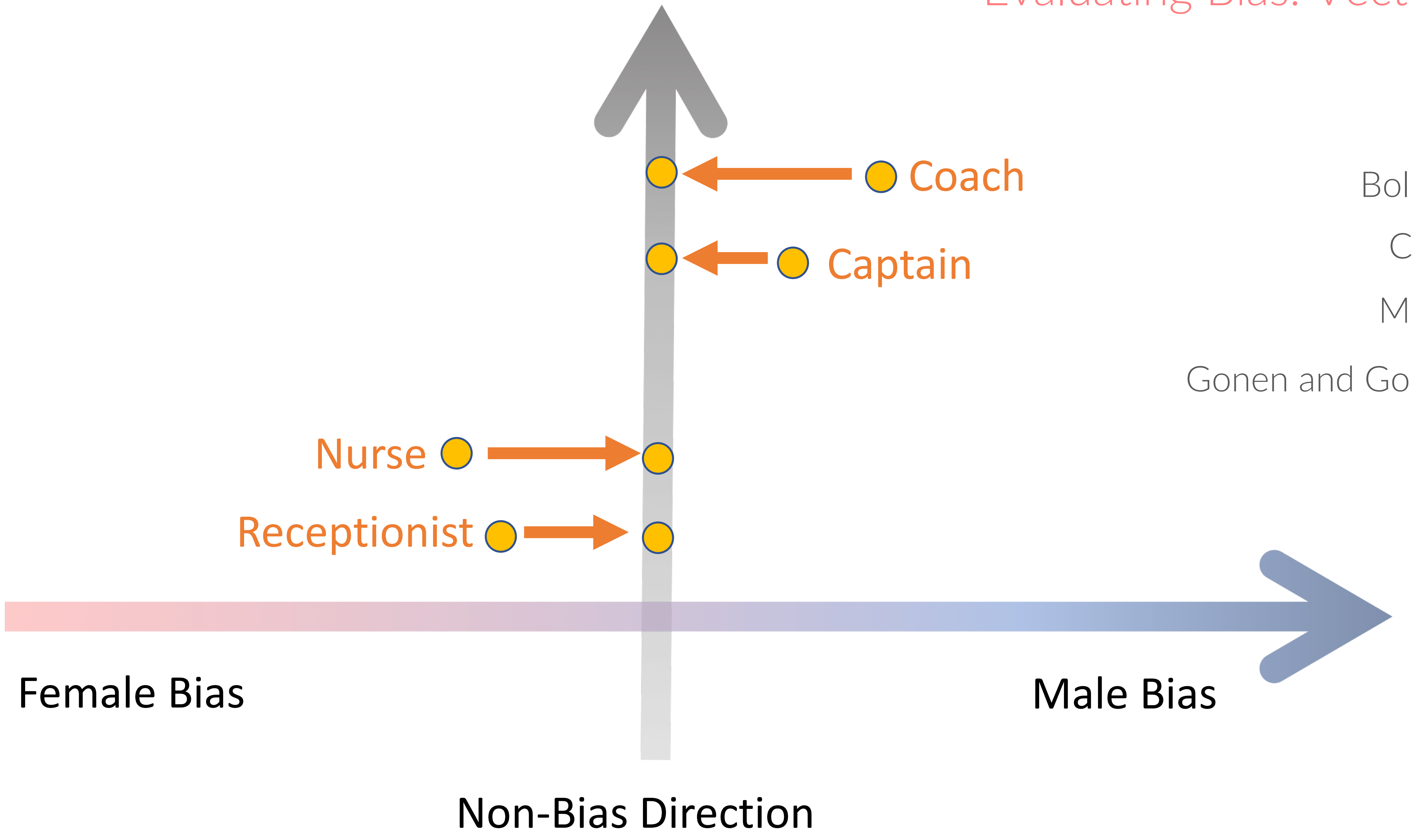
## Method Categorizations

1. Vector Spaces

2. Gender Bias Evaluation Test Set (GBETs)

# Gender Bias in Word Embeddings

## Evaluating Bias: Vector Spaces



Bolukbasi et al., 2016

Caliskan et al, 2017

Manizini et al., 2017

Gonen and Goldberg et al., 2019

# Gender Bias Evaluation Test Sets

- New data sets for evaluating gender bias
  - Traditional data sets lack gender-specific information
  - Eliminate confounding variables

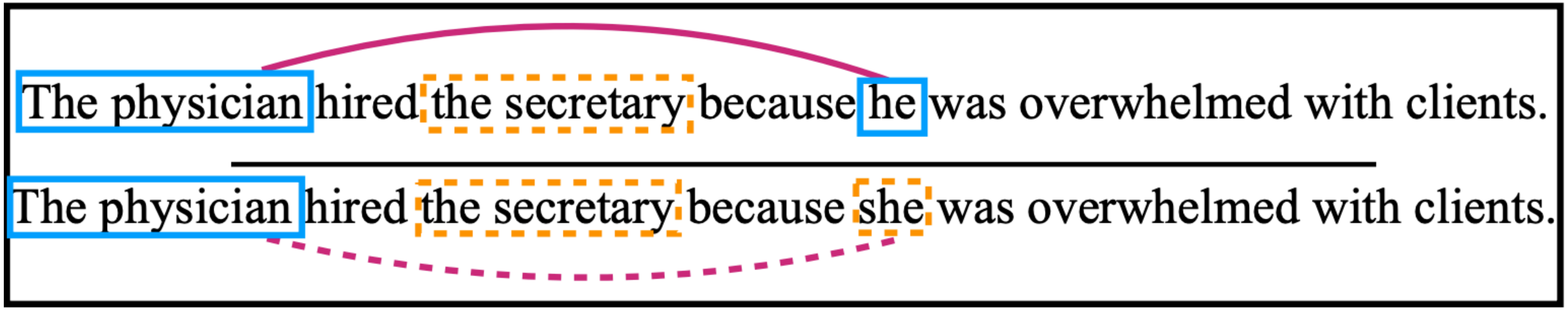
Rudinger et al., 2018

Zhao et al., 2018

Webster et al., 2018

Kiritchenko and Mohammad, 2018

**He called his mother** → **She called her father**



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

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

Sentences from WinoBias. (Zhao et al., 2018)

# We know there's bias. Now what?

## Mitigating the Bias

1. Inference vs Retraining

2. Training Data vs Algorithm

3. Key Debiasing Methods

# Categorizing Debiasing Methods

## Mitigating Bias: Retraining vs. Inference

### Retraining

Fully retrain the model

### Inference

Adjust the model's predictions at test time

# Categorizing Debiasing Methods

## Mitigating Bias: Data vs. Algorithm

### Training Data

Debias the training data.

Always Retraining

### Algorithm

Debias the algorithm.

Retraining or Inference

# Debiasing via Data Augmentation

Mitigating Bias: Retraining + Data Method

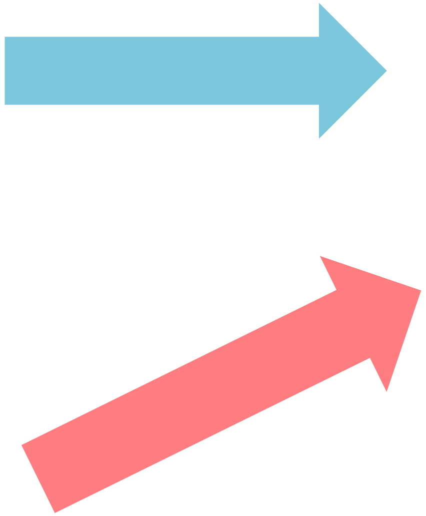
Zhao et al., 2018

Original Corpus

- He is a doctor.
- The doctor called his mom.

Genderswapped Corpus

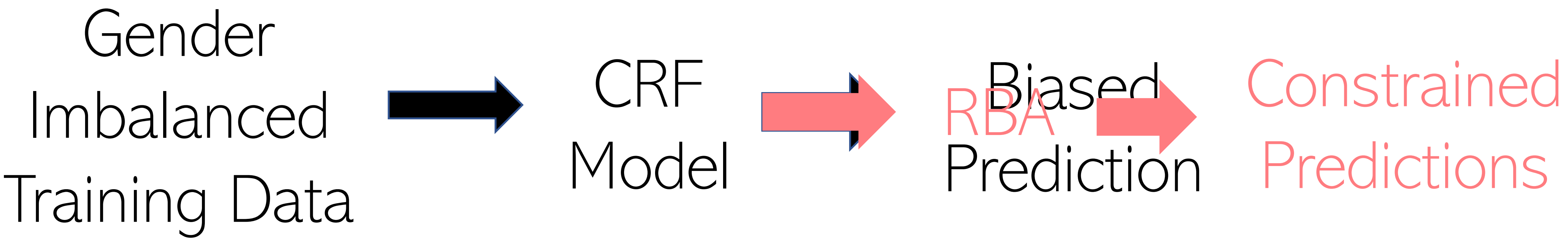
- She is a doctor.
- The doctor called her dad.



**Coreference Resolution Model**

# Constraining Algorithm Predictions

Mitigating Bias: Inference + Algorithm Method



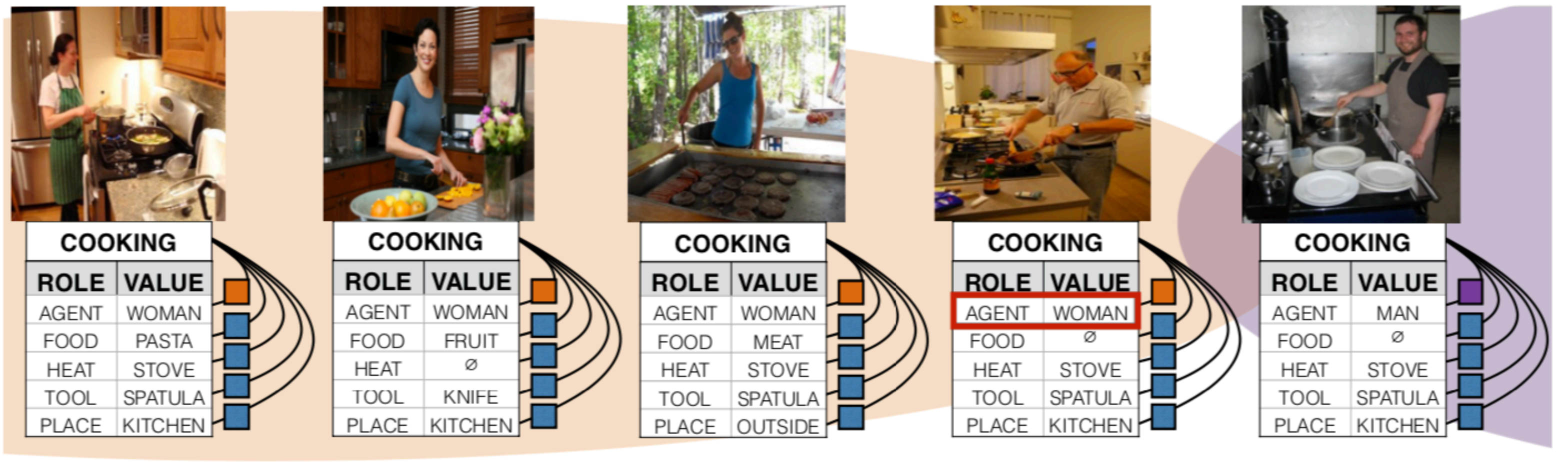
Gender Imbalanced Training Data

CRF Model

Biased Prediction

RBA

Constrained Predictions



| COOKING | |
|---|---|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | PASTA |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |

| COOKING | |
|---|---|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | FRUIT |
| HEAT | ∅ |
| TOOL | KNIFE |
| PLACE | KITCHEN |

| COOKING | |
|---|---|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | MEAT |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | OUTSIDE |

| COOKING | |
|---|---|
| ROLE | VALUE |
| AGENT | WOMAN |
| FOOD | ∅ |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |

| COOKING | |
|---|---|
| ROLE | VALUE |
| AGENT | MAN |
| FOOD | ∅ |
| HEAT | STOVE |
| TOOL | SPATULA |
| PLACE | KITCHEN |

Five examples from the imSitu visual semantic role labeling dataset (Zhao et al., 2018)

# Our Contributions

- We provide a comprehensive literature review of gender bias in NLP
- Critically discuss issues with the purpose of identifying optimizations, knowledge gaps, and directions for future research

# Future Directions

## Ideas Worth Exploring

1. Non-English Languages

2. Non-binary Bias

3. Interdisciplinary Communication

# Thanks for Listening

ACL 2019

# Q & A

Mitigating Gender Bias in NLP: Literature Review

ajg@ucsb.edu tonysun@ucsb.edu