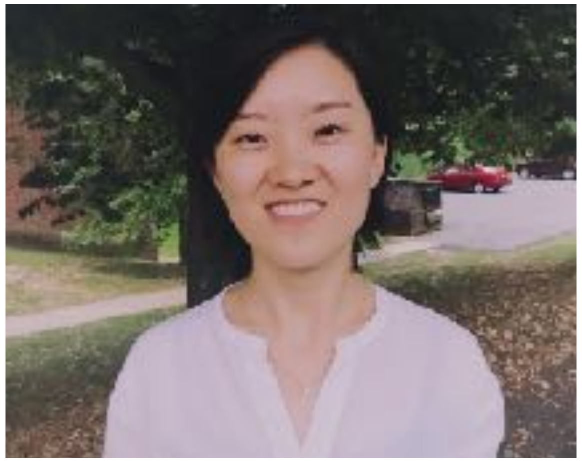
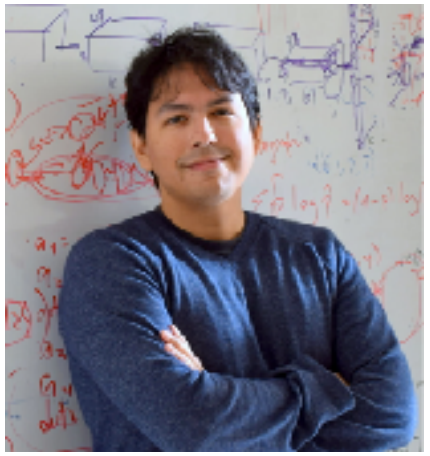
MEN ALSO LIKE SHOPPING

REDUCING GENDER BIAS AMPLIFICATION USING CORPUS-LEVEL CONSTRAINTS

Jieyu Zhao[1,3], Tianlu Wang [1], **Mark Yatskar** [2,4], Vicente Ordonez [1] , Kai-Wei Chang [1,3]

[1] University of Virginia [2] University of Washington [3] UCLA [4] Allen Institute for AI




Dataset Gender Bias

33%

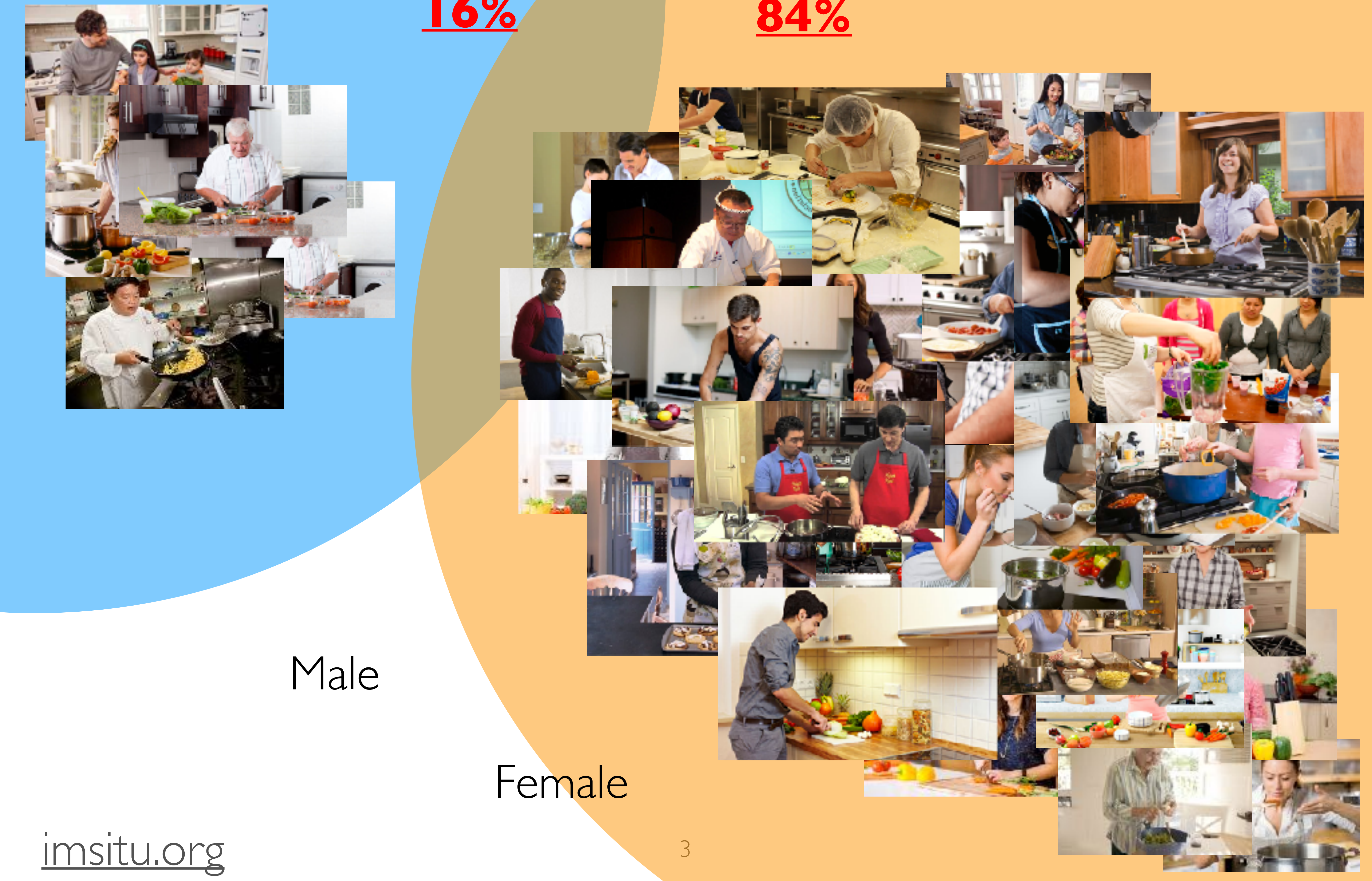
66%

Male

Female

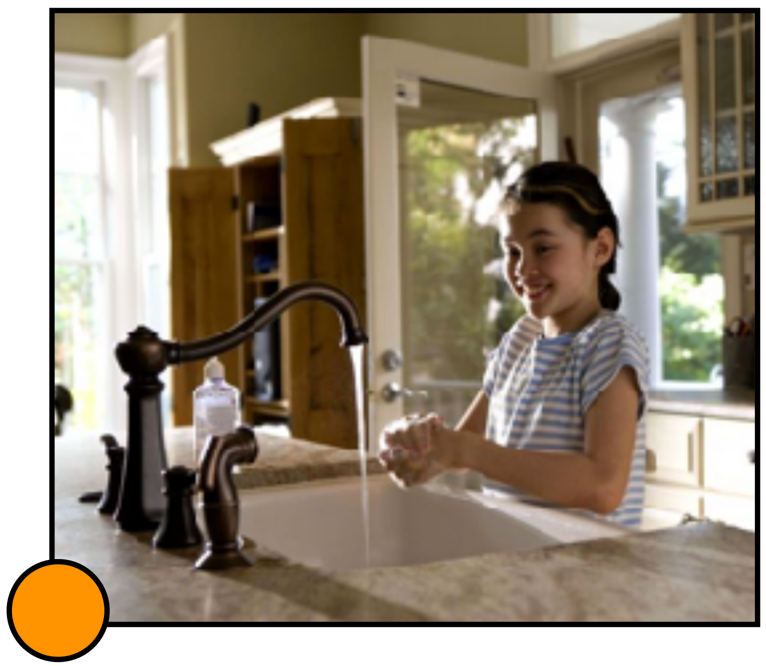
Model Bias After Training

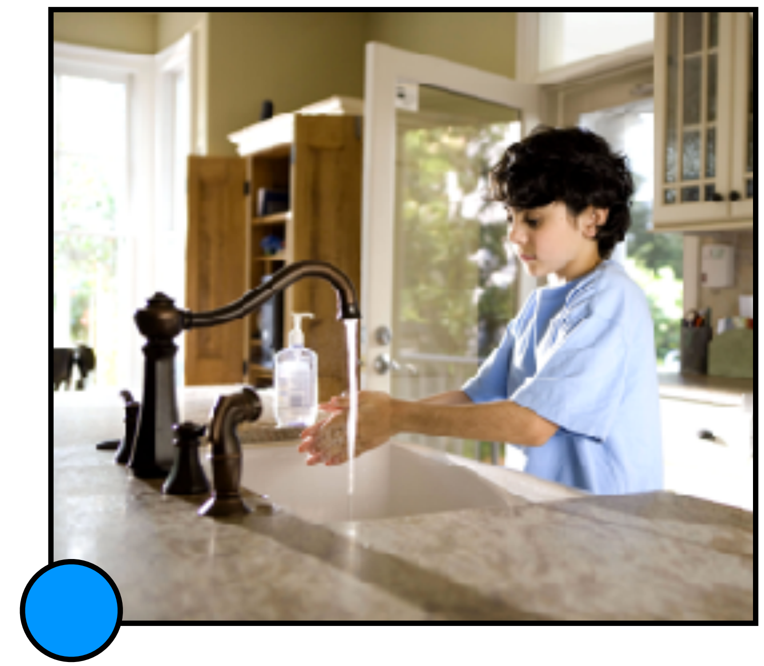
16% **84%**

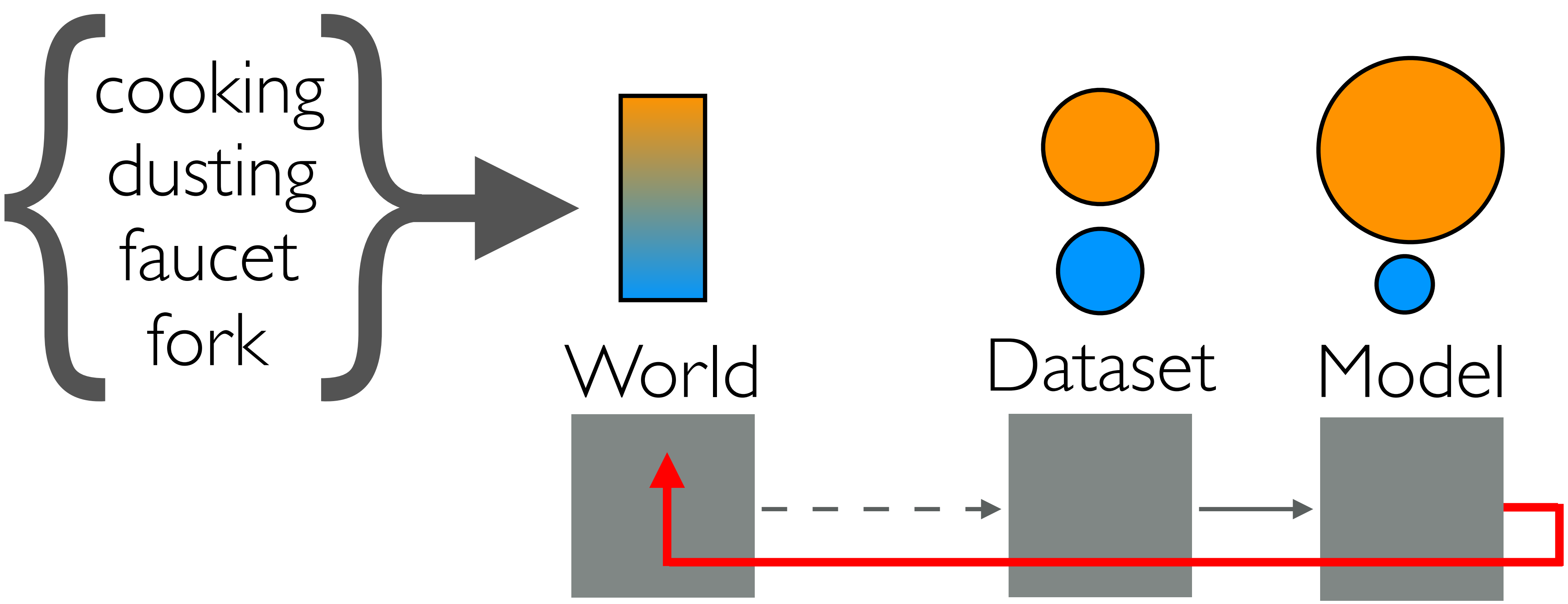


Why does this happen? Good for accuracy

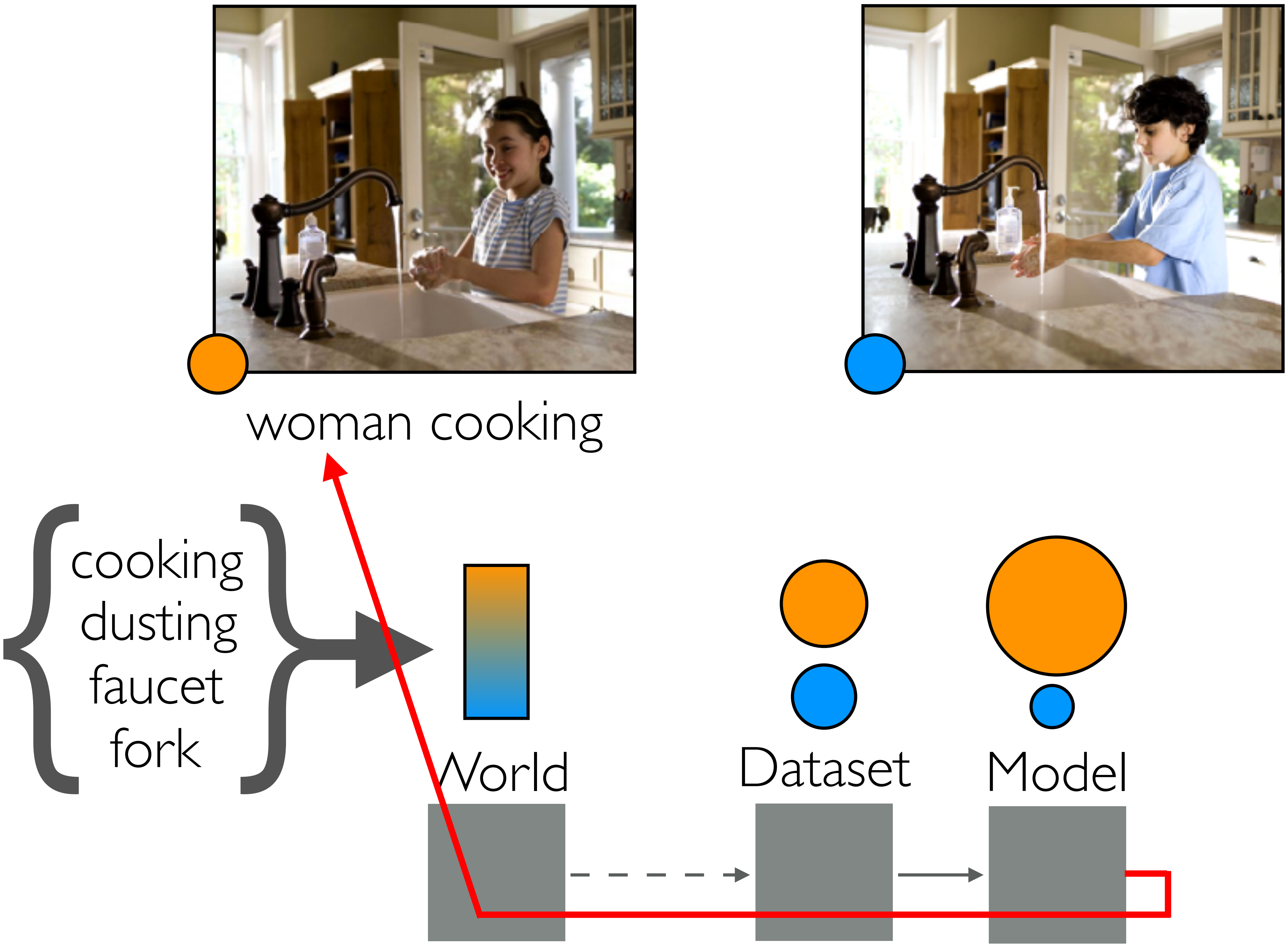
Algorithmic Bias in Grounded Setting



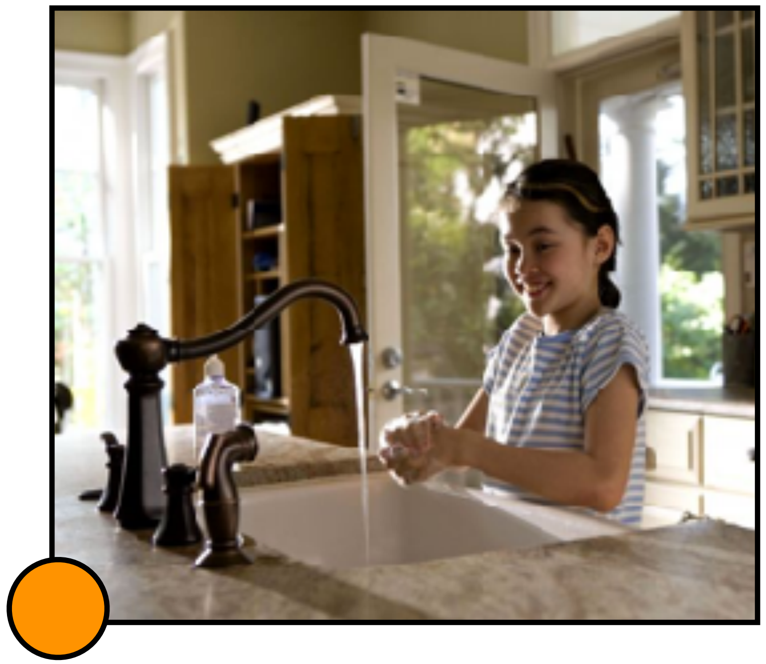




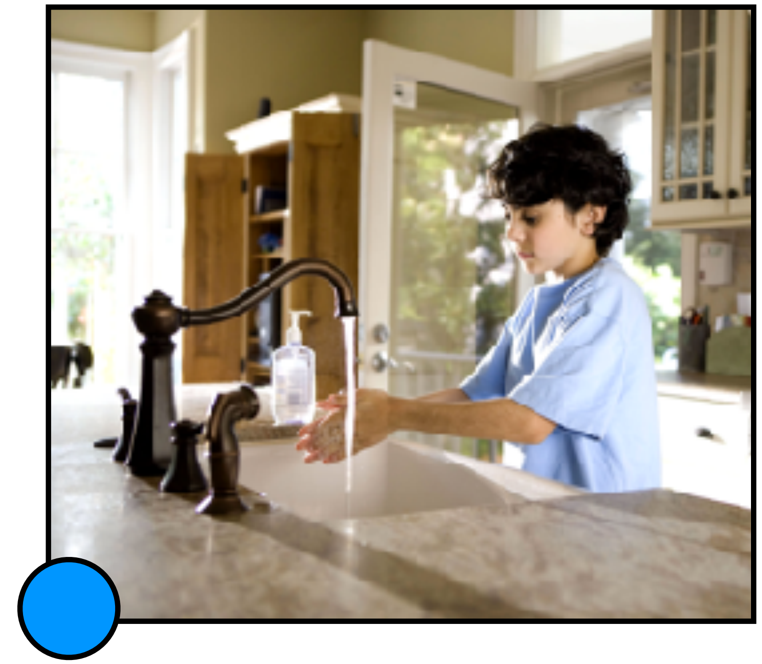
Algorithmic Bias in Grounded Setting



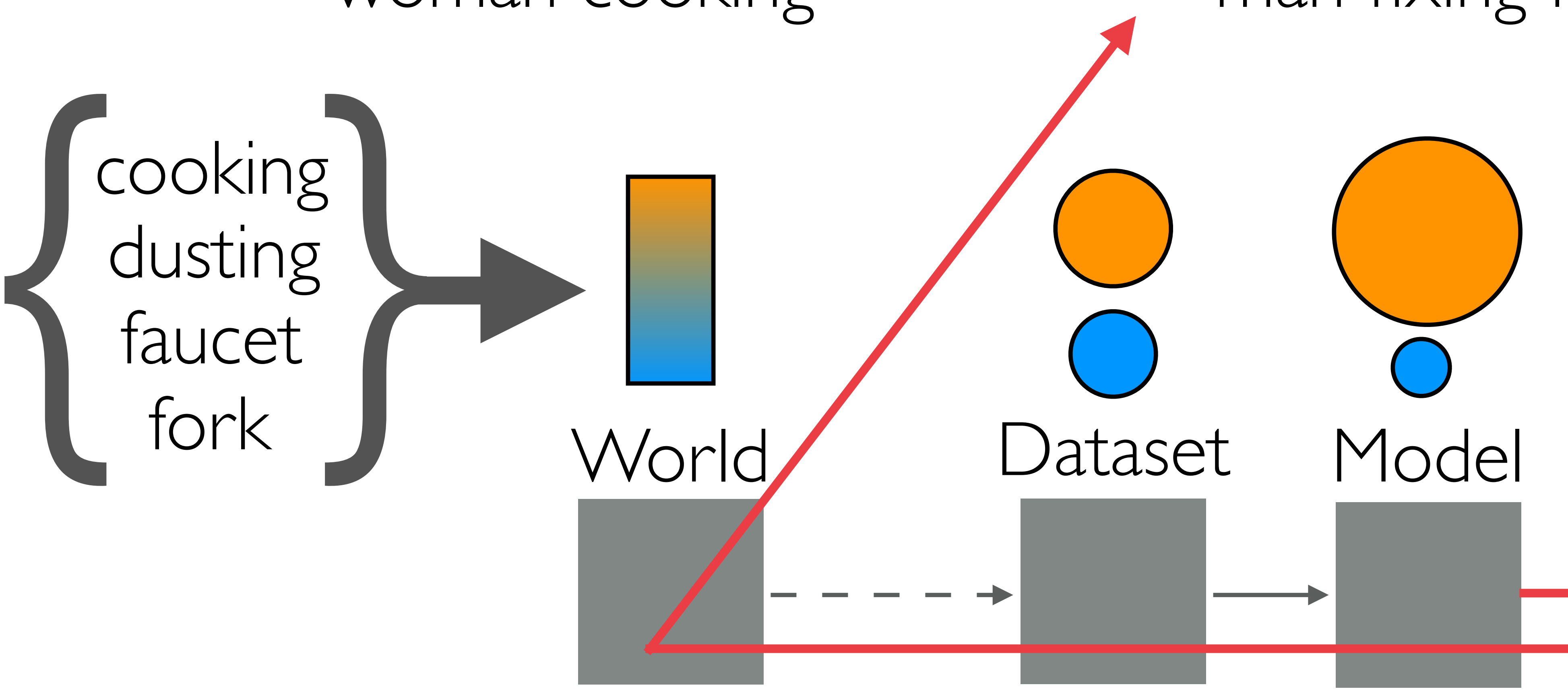
Algorithmic Bias in Grounded Setting



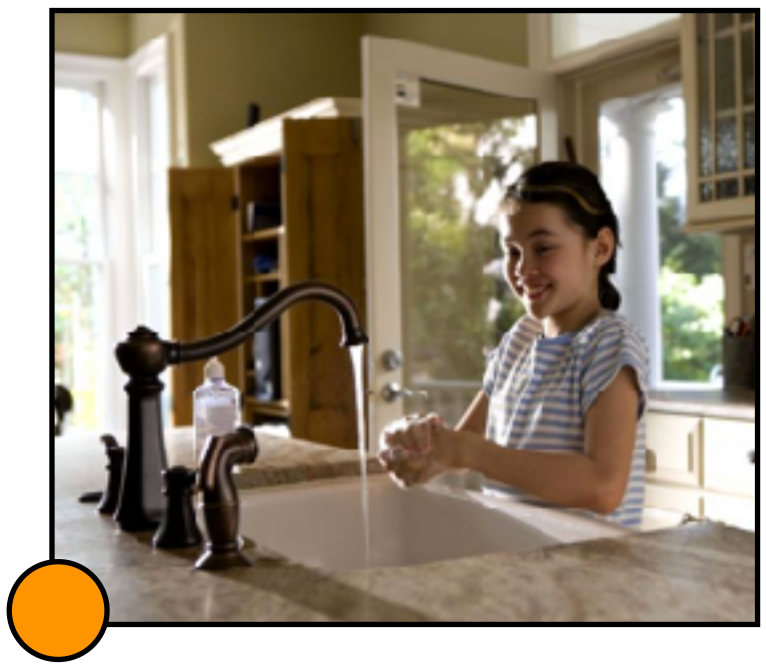
woman cooking

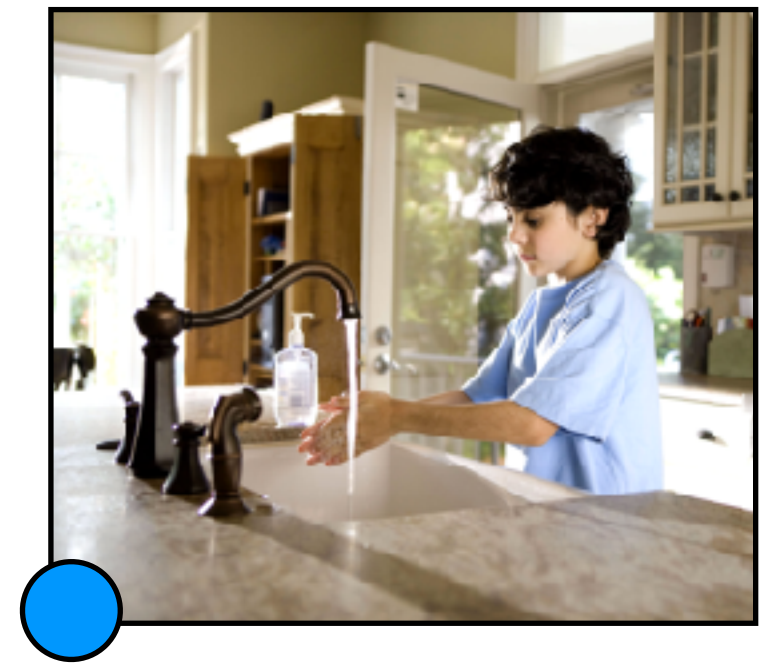


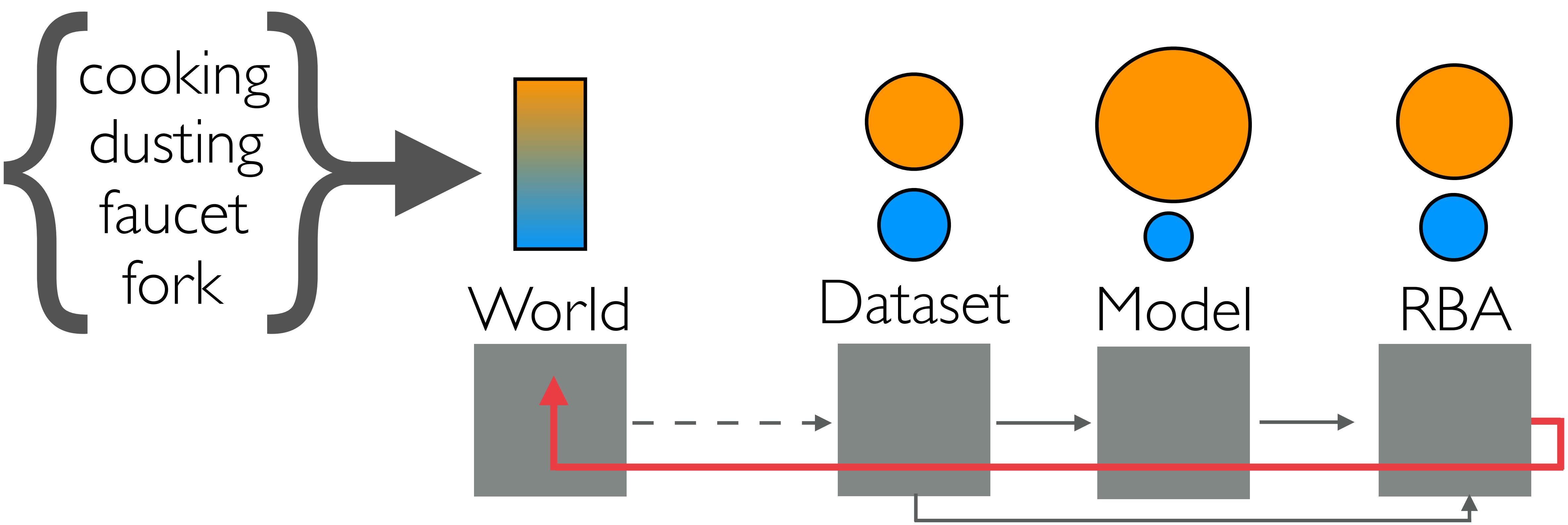
man fixing faucet



Algorithmic Bias in Grounded Setting

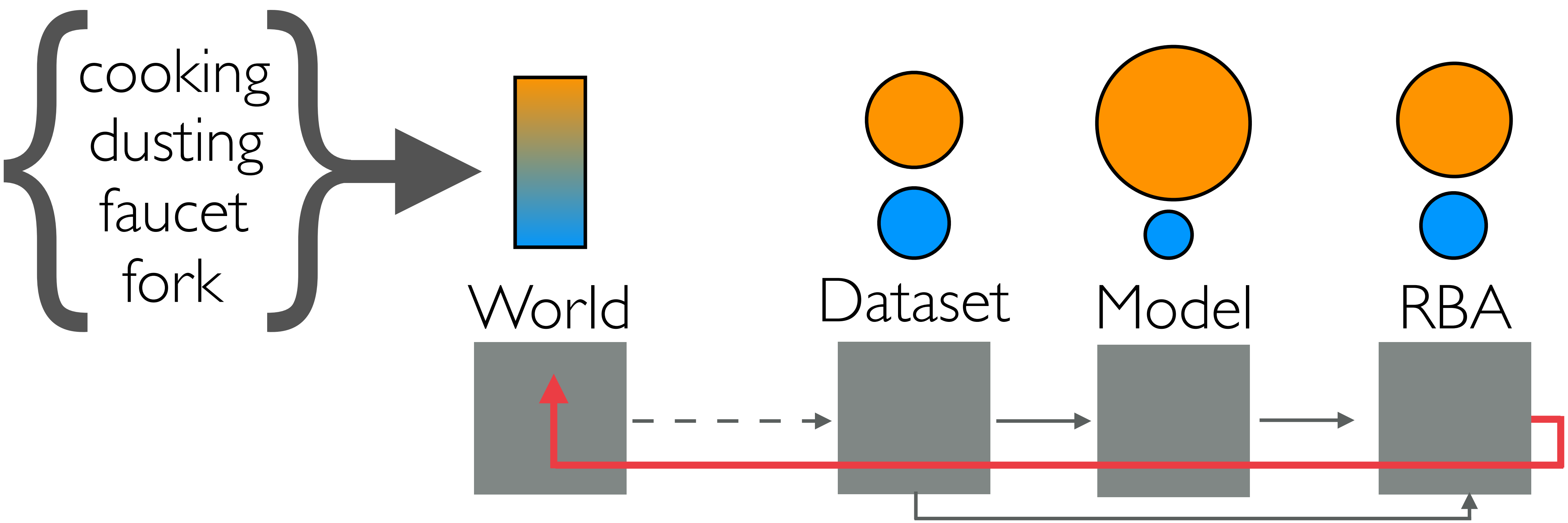






Algorithmic Bias in Grounded Setting

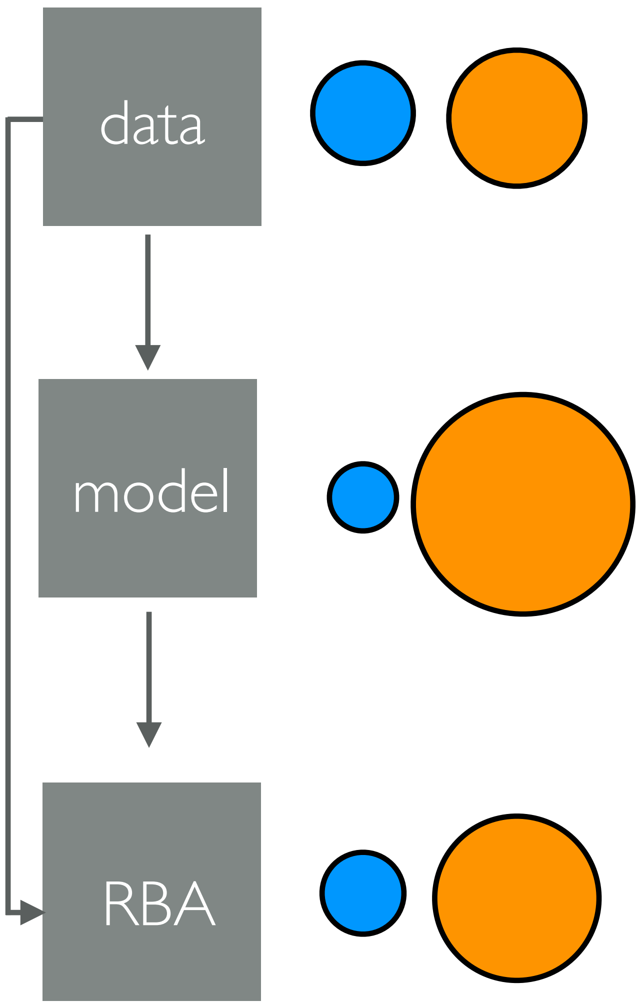
Reduce amplification ~50%
Negligible loss in performance



Contributions

imSitu vSRL
(events)

COCO MLC
(objects)



High dataset gender bias

38% (objects) 47% (events) exhibit strong bias

Models amplify existing gender bias

~70% objects and events have bias amplification

Reducing bias amplification

~50% reduction in amplification

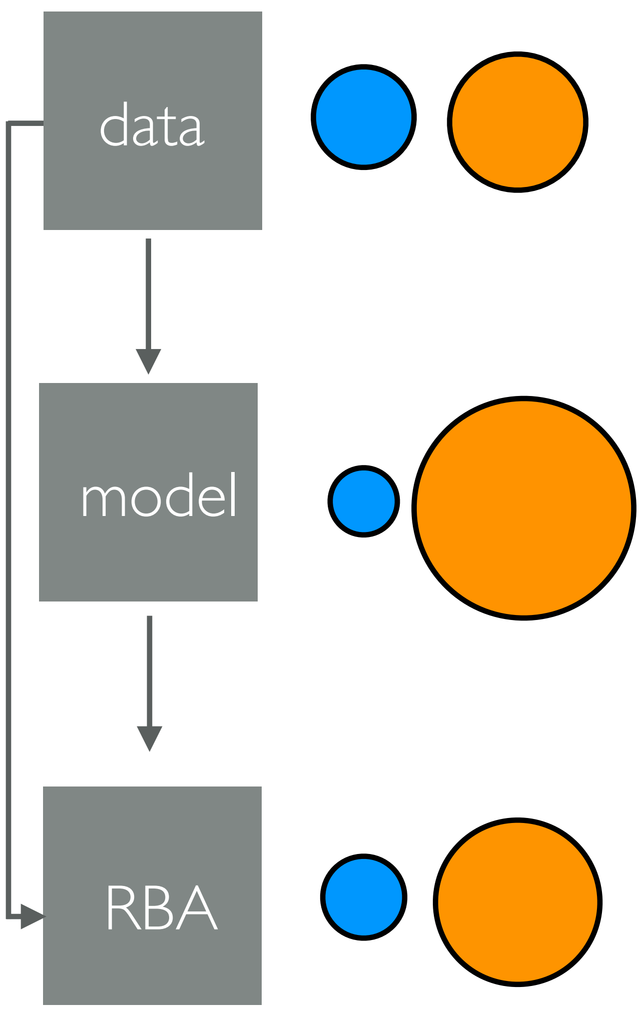
Insignificant loss in performance

Outline


imSitu vSRL
(events)


COCO MLC
(objects)

1. Background



2. Dataset Bias

3. Bias Amplification

4. Reducing Bias Amplification

imSitu Visual Semantic Role Labeling (vSRL)
(events)

Internet

FrameNet

WordNet

COOKING	
ROLES	**NOUNS**
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

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

imSitu Visual Semantic Role Labeling (vSRL)
(events)

Internet

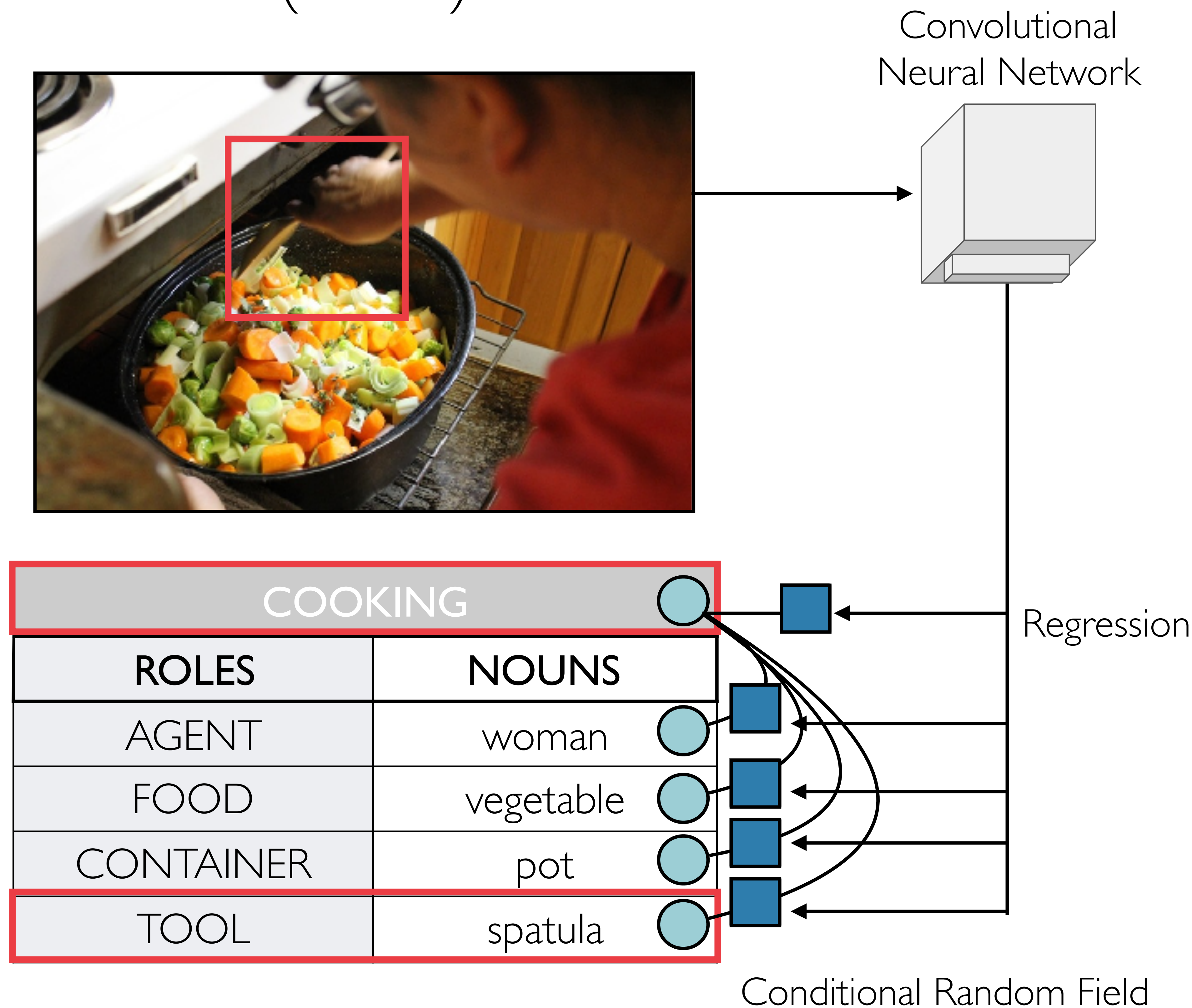
FrameNet

WordNet

COOKING	
ROLES	**NOUNS**
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

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

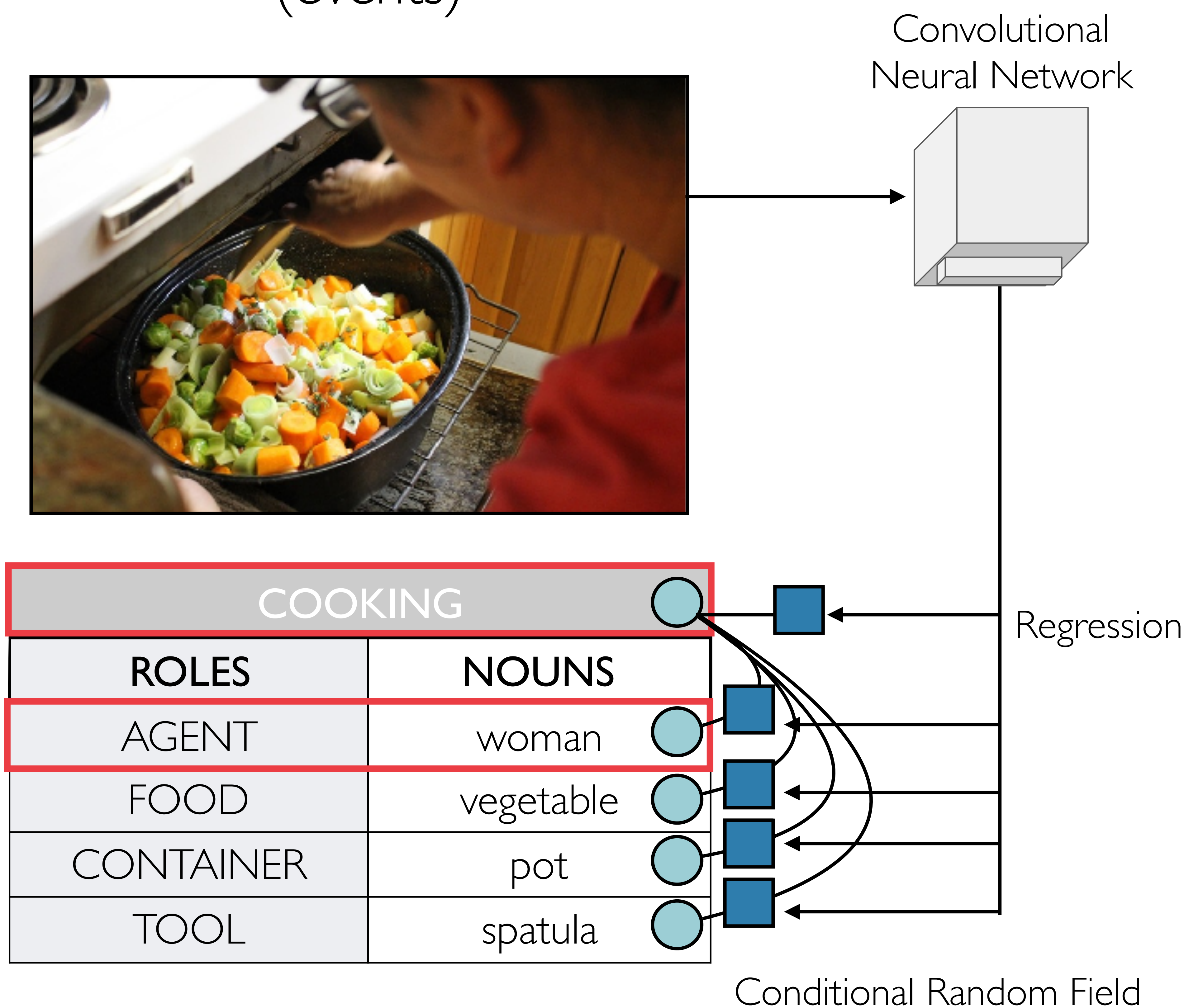
imSitu Visual Semantic Role Labeling (vSRL)
(events)



COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

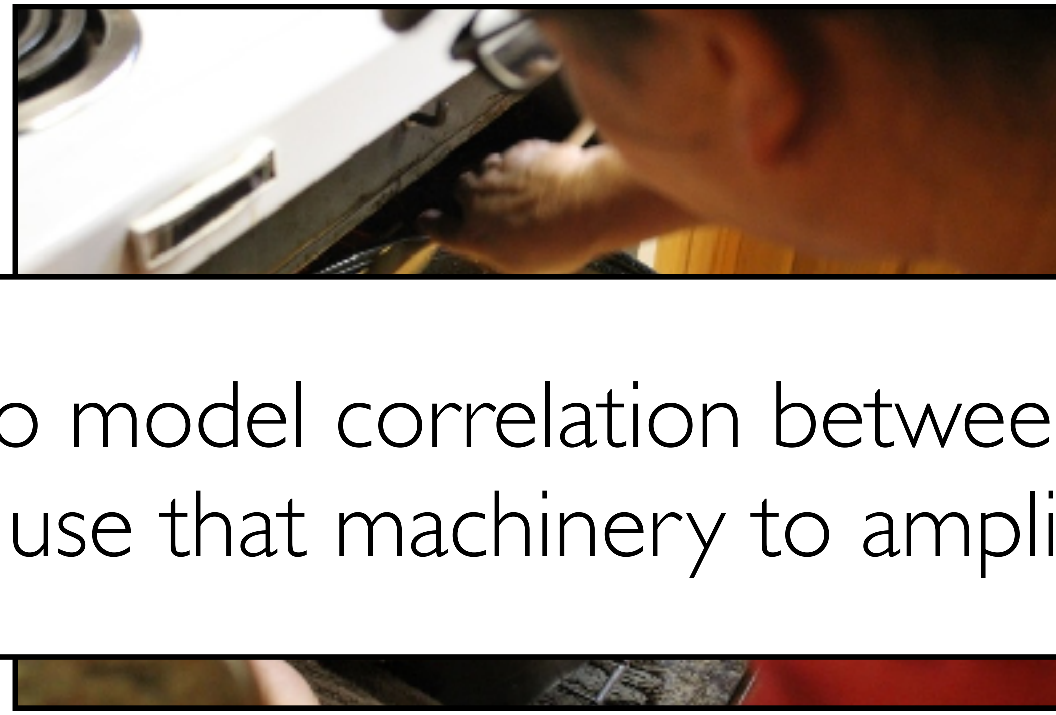
Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16

imSitu Visual Semantic Role Labeling (vSRL)
(events)



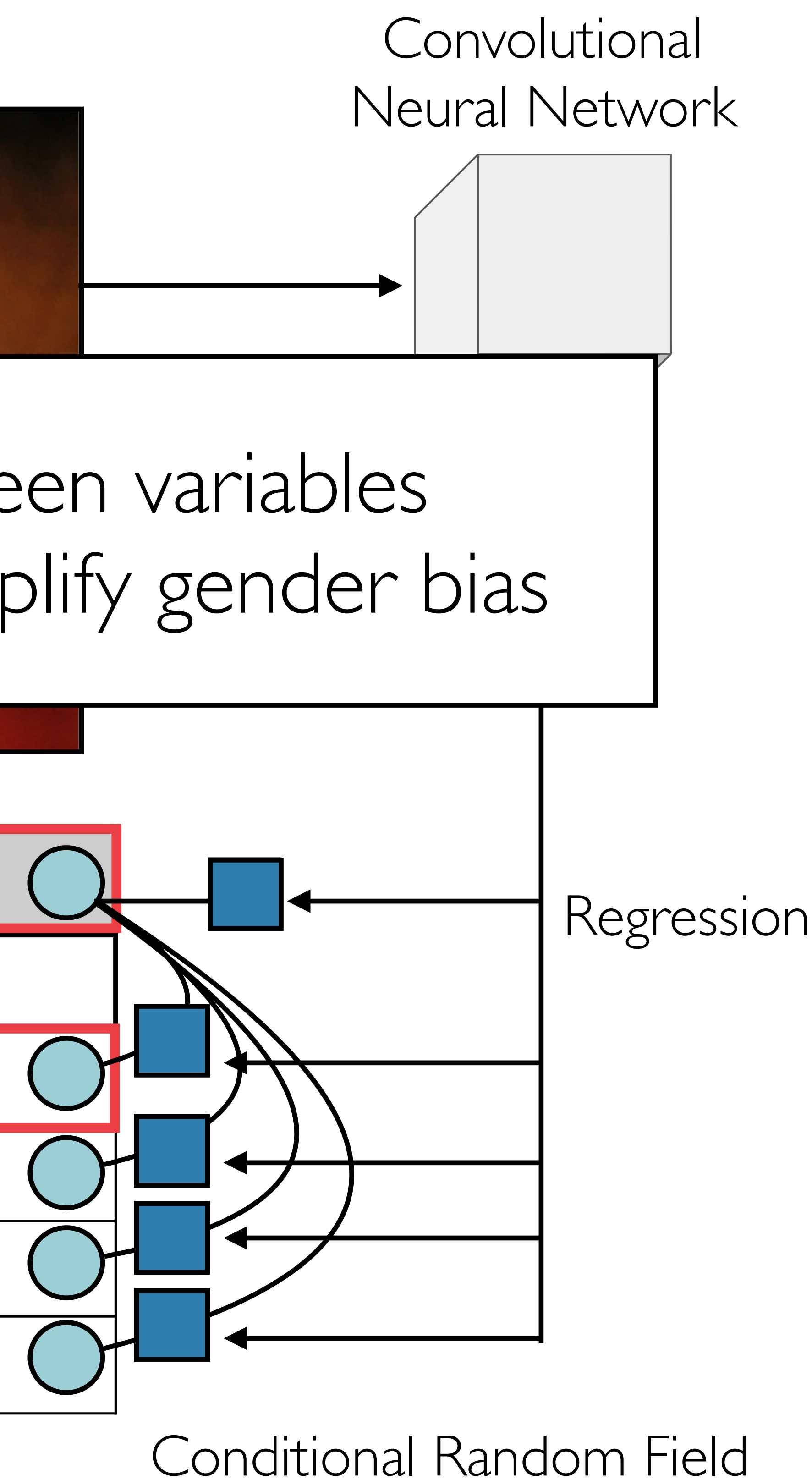
Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16

imSitu Visual Semantic Role Labeling (vSRL)
(events)



Need to model correlation between variables
Model can use that machinery to amplify gender bias

COOKING	
ROLES	**NOUNS**
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula



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

COCO Multi-Label Classification (MLC)
(objects)

Internet

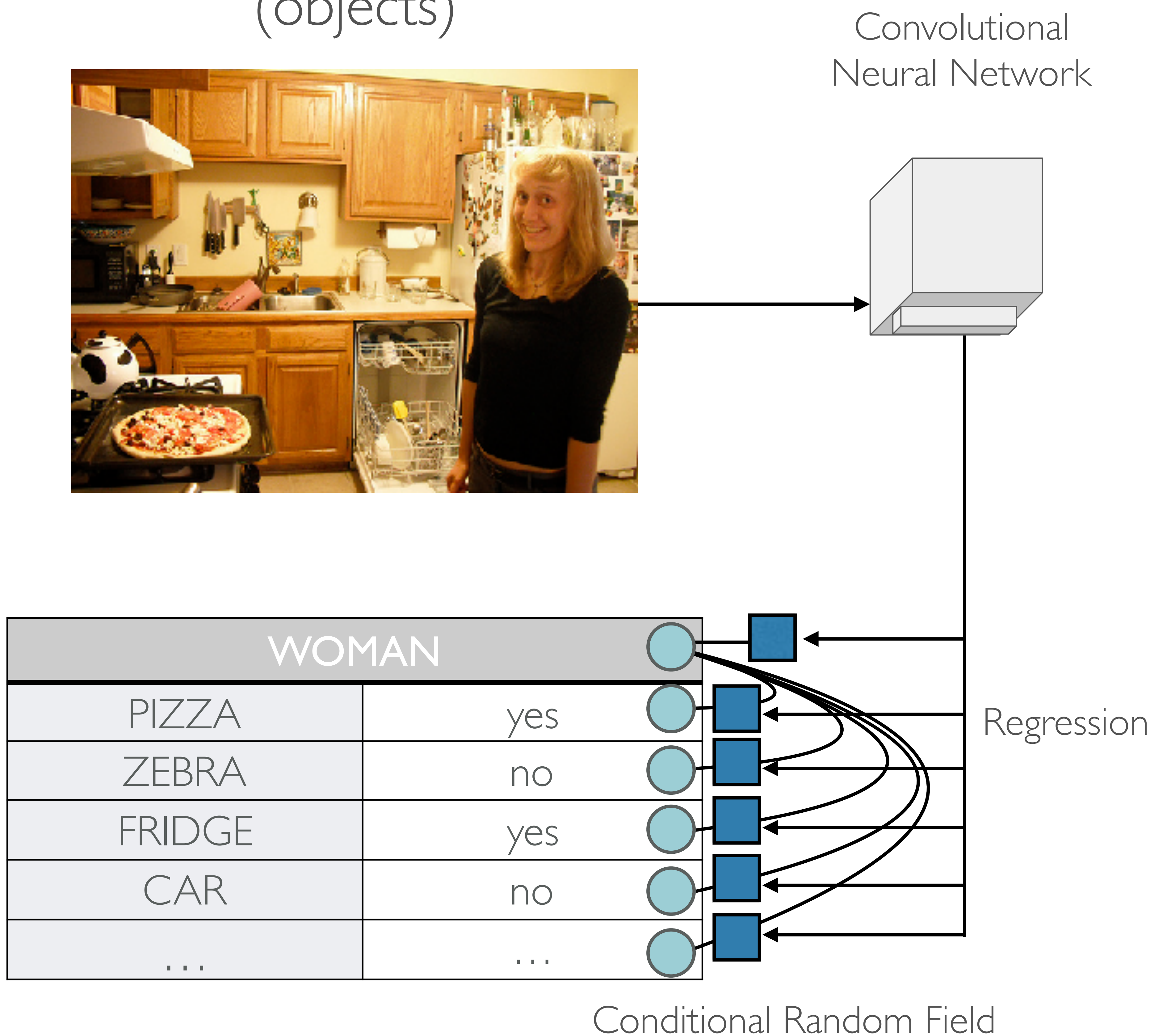
a <u>woman</u> is smiling in a kitchen near a pizza on a stove

WOMAN	
PIZZA	yes
ZEBRA	no
FRIDGE	yes
CAR	no
...	...

Caption Inferred Label

COCO Objects

COCO Multi-Label Classification (MLC)
(objects)



Related Work

- Implicit Bias

    image search (Kay et al., 2015)
    search advertising (Sweeny, 2013)
    online news (Ross and Carter, 2011)
    credit score (Hardt et al., 2016)
    word vector (Bolukbasi et al., 2016)

- Classifier class imbalance

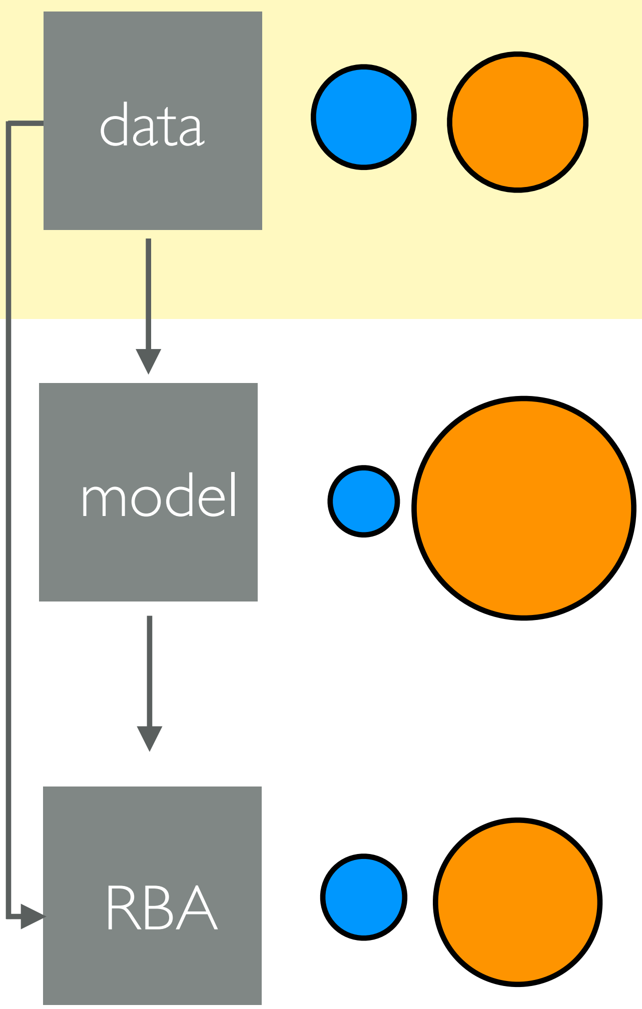
    Barocas and Selbst, 2014; Dwork et al., 2012;
    Feldman et al., 2015; Zliobaite, 2015

Outline

imSitu vSRL
(events)

COCO MLC
(objects)

1. Background



2. Dataset Bias

3. Model Bias Amplification

4. Reducing Bias Amplification

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)

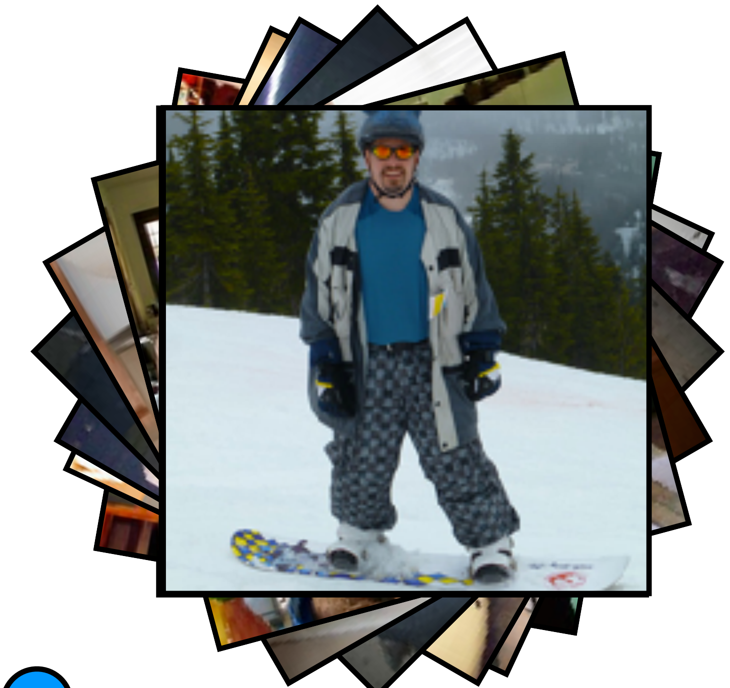
Training Gender Ratio (▲ noun)

Training Set

▲ snowboard

● woman

● man

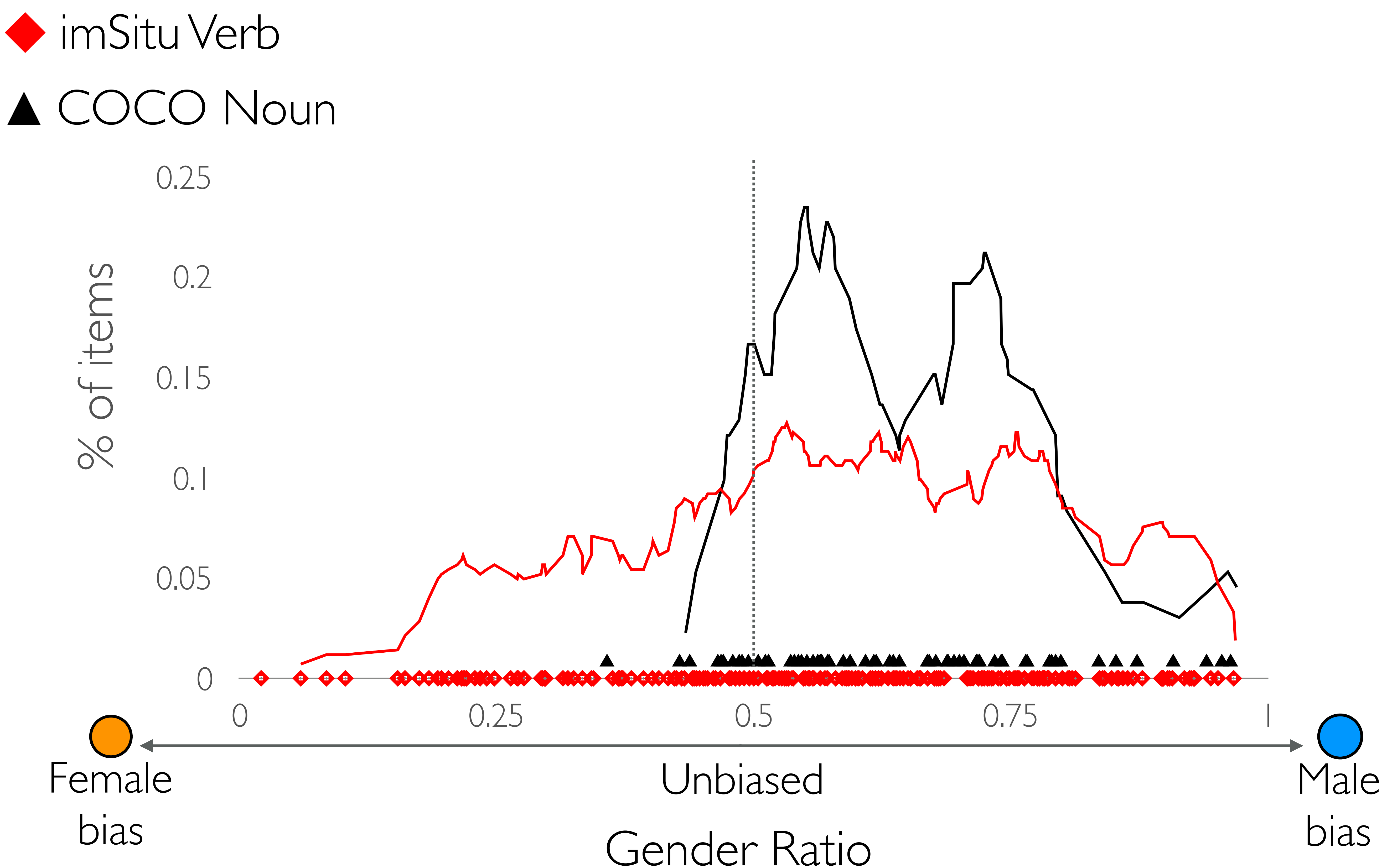


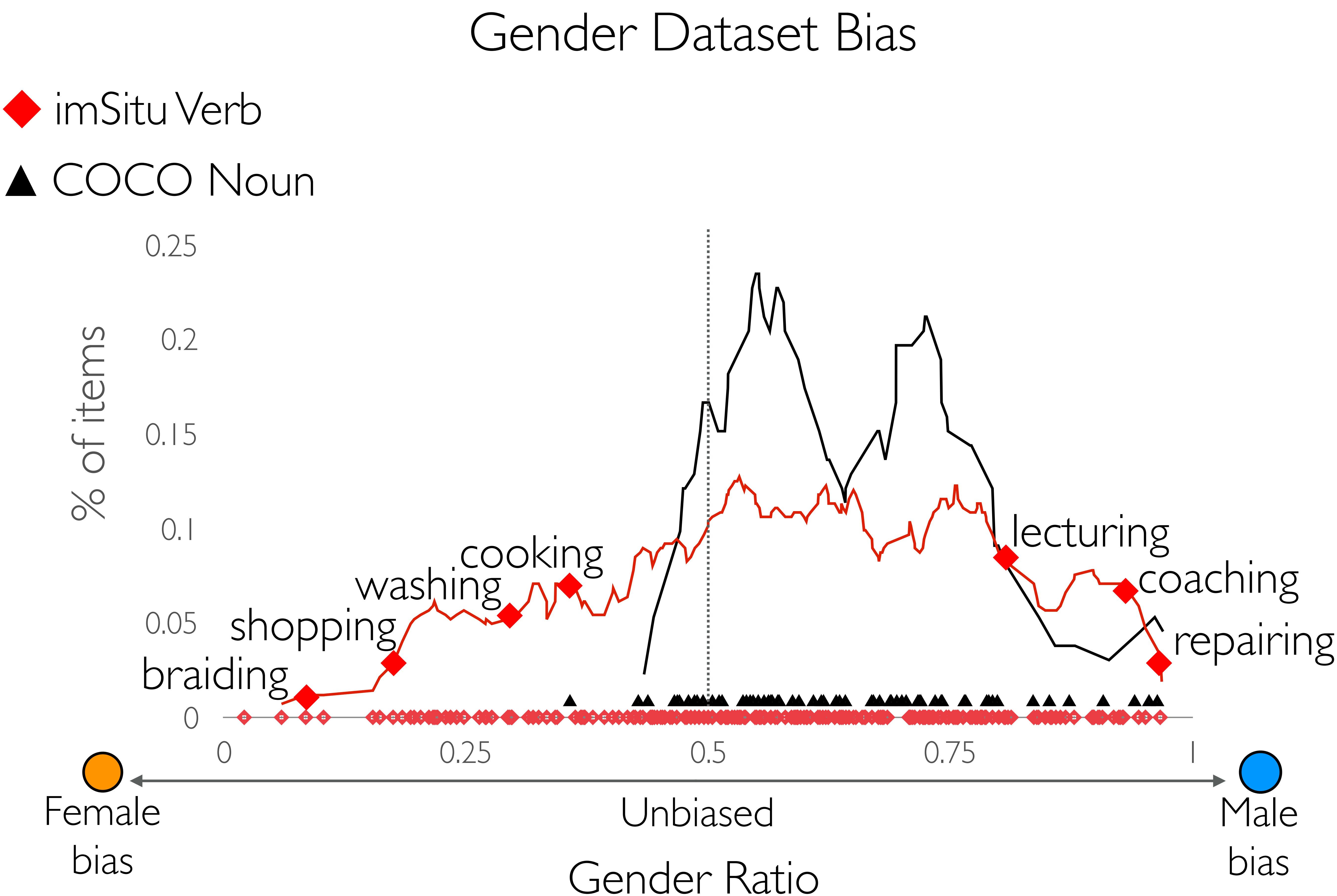
MAN	
snowboard	yes
refrigerator	no
bowl	no

WOMAN	
snowboard	yes
refrigerator	no
bowl	no

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

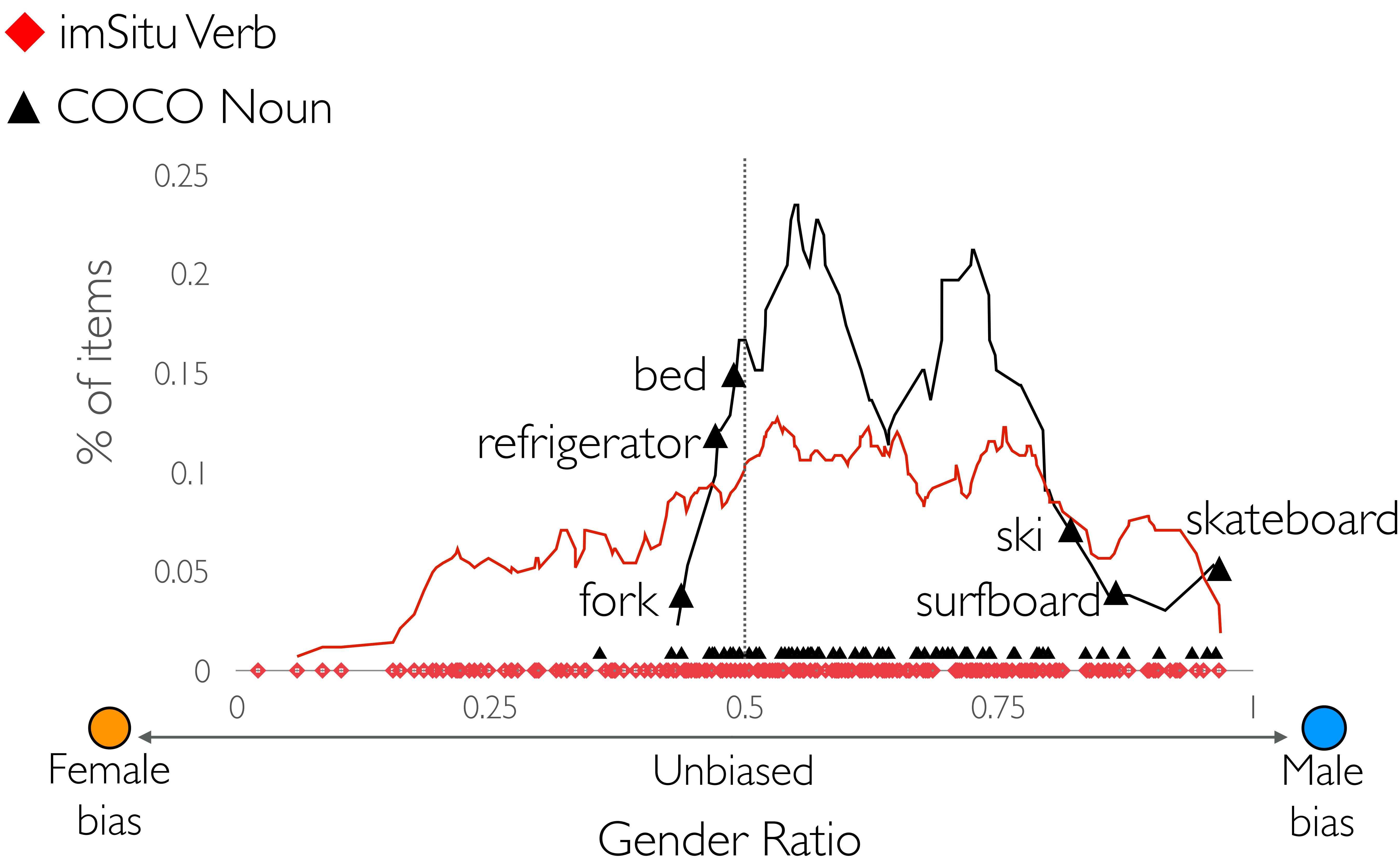
Gender Dataset Bias

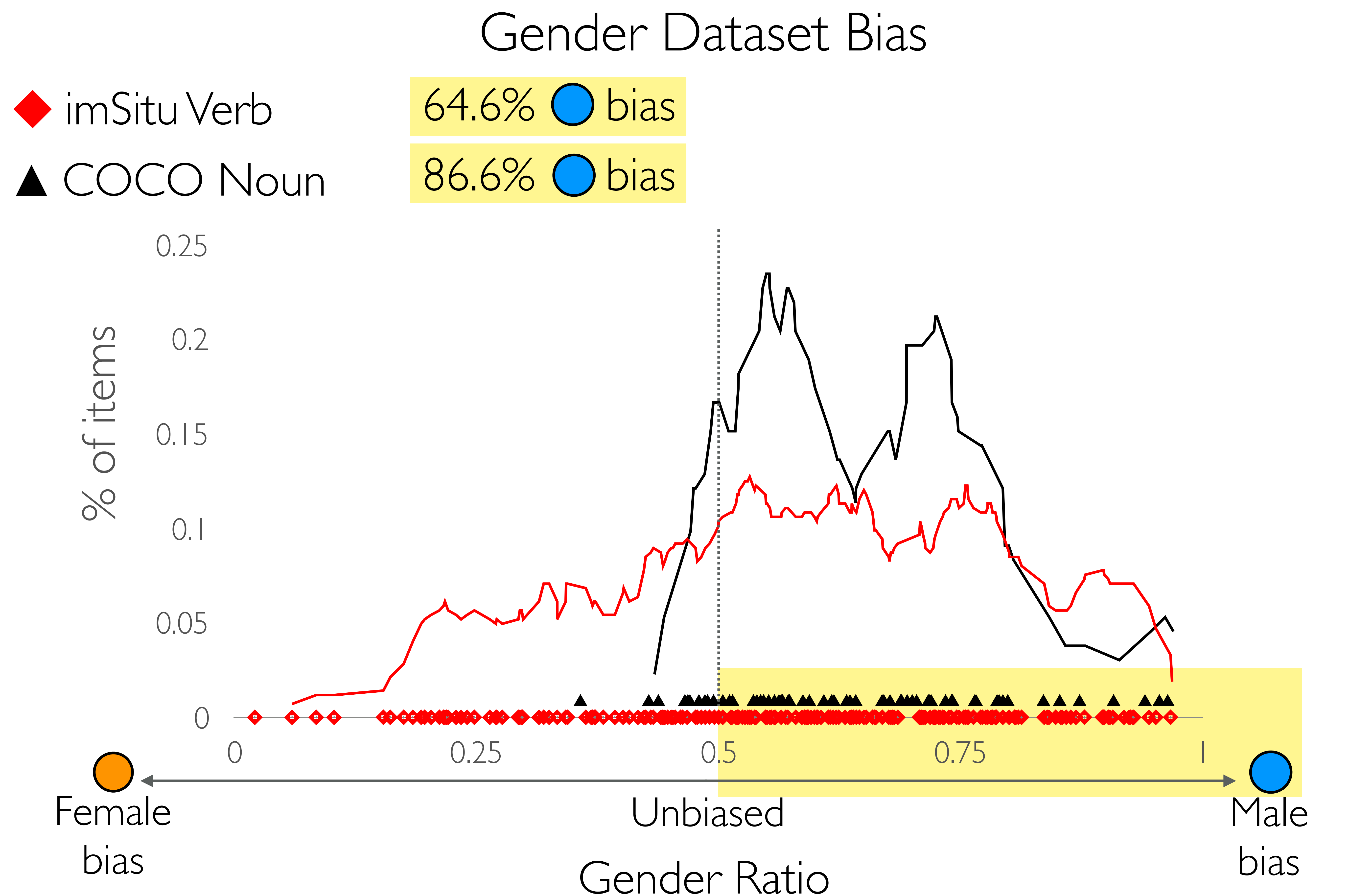




Gender Dataset Bias
imSitu Verb
COCO Noun
% of items
0.25
0.2
0.15
0.1
0.05
0
braiding
shopping
washing
cooking
lecturing
coaching
repairing
0
0.25
0.5
0.75
1
Female bias
Unbiased
Male bias
Gender Ratio

Gender Dataset Bias



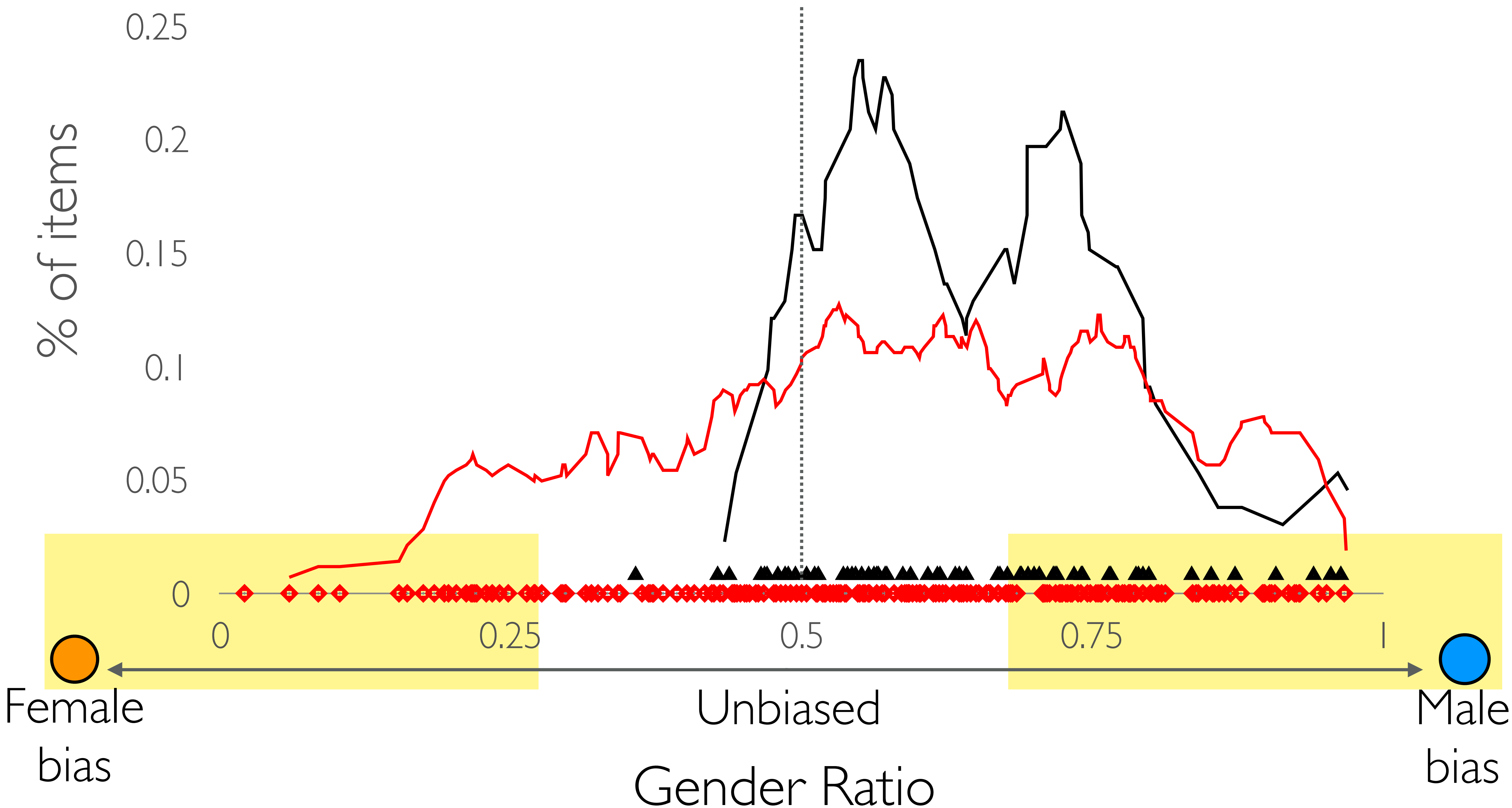


Gender Dataset Bias
imSitu Verb
COCO Noun
64.6% bias
86.6% bias
% of items
0.25
0.2
0.15
0.1
0.05
0
0
0.25
0.5
0.75
1
Female bias
Unbiased
Male bias
Gender Ratio

Gender Dataset Bias

◆ imSitu Verb 64.6% ● bias 46.9% strong bias (>2:1)

▲ COCO Noun 86.6% ● bias 37.9% strong bias (>2:1)

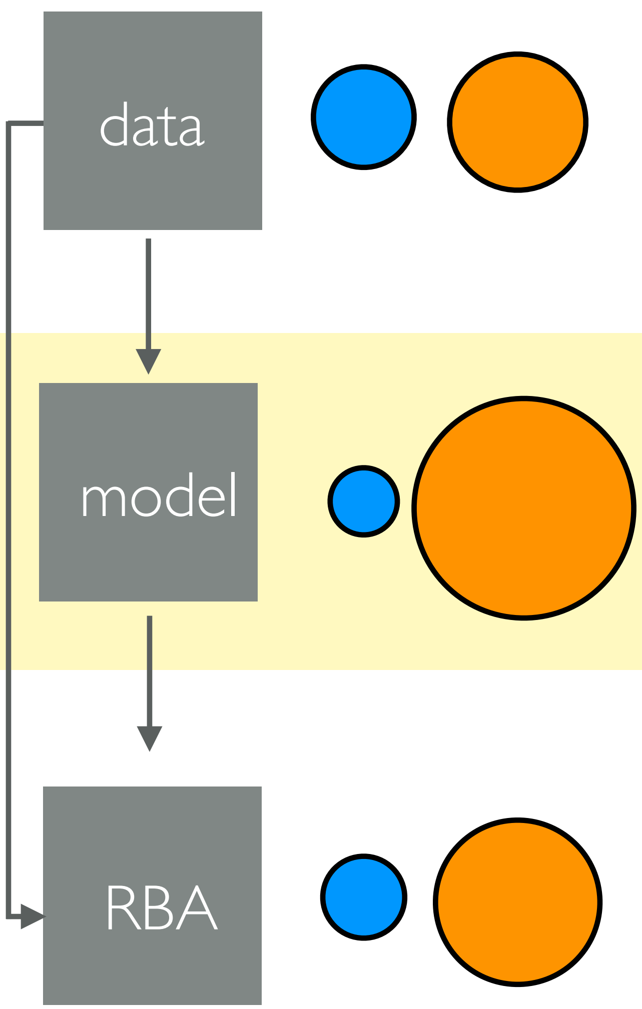


Outline

imSitu vSRL
(events)

COCO MLC
(objects)

1. Background



2. Dataset Bias

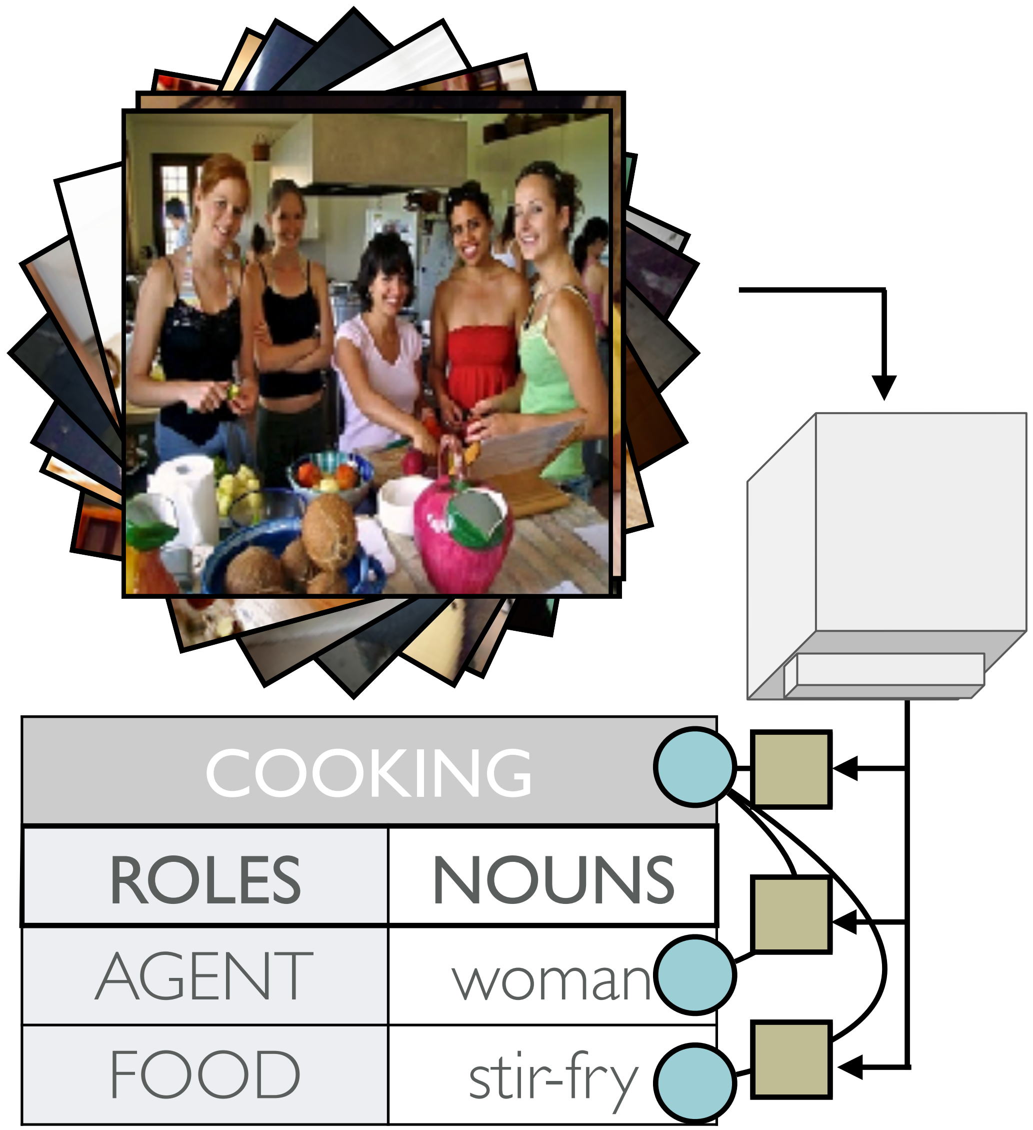
3. Bias Amplification

4. Reducing Bias Amplification

Defining Bias Amplification (events)

Predicted Gender Ratio (◆ verb)

Development Set



What does the model predict on unseen data?

Defining Bias Amplification (events)

Predicted Gender Ratio (verb)

Development Set

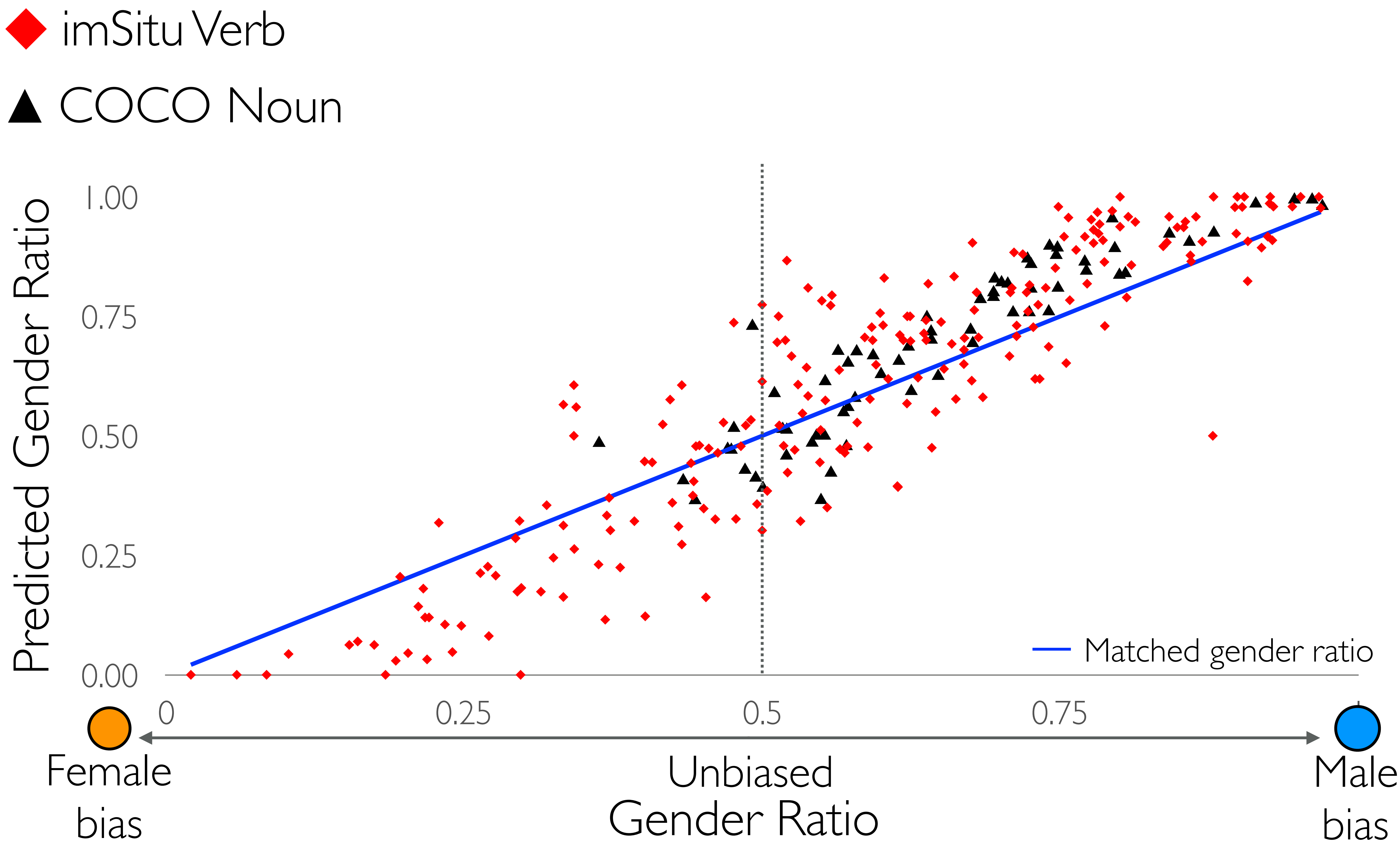
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

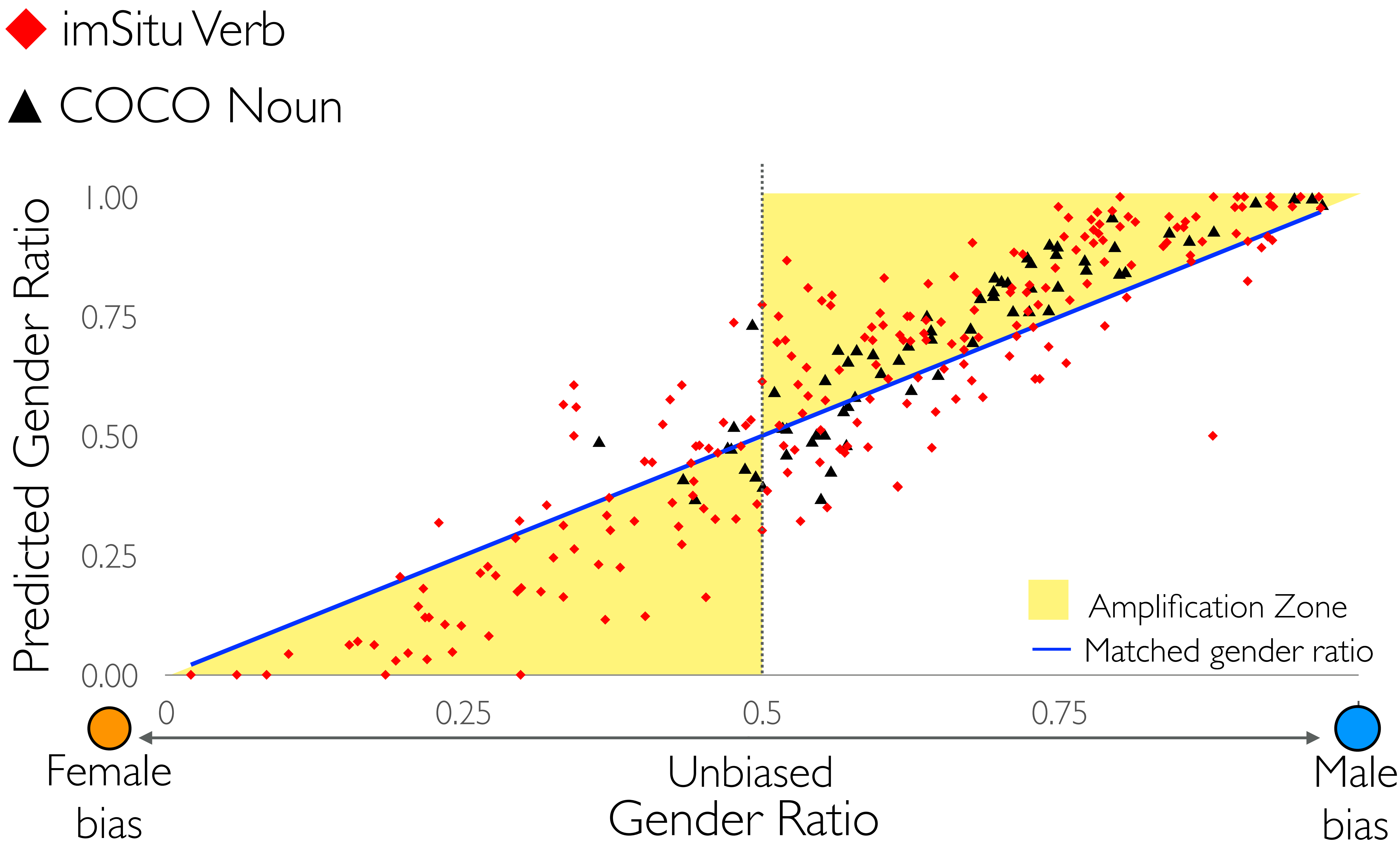
COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

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

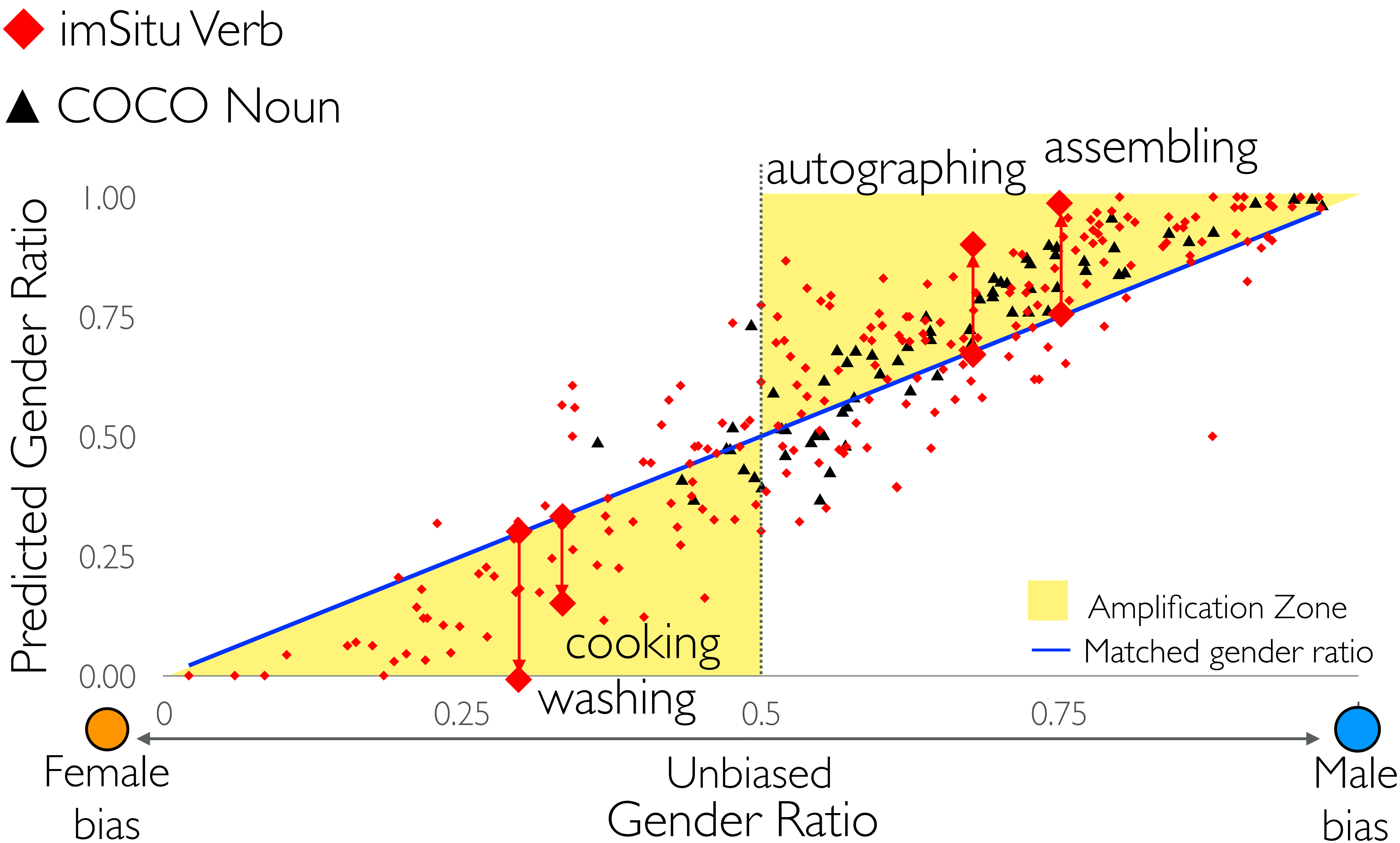
Model Bias Amplification



Model Bias Amplification



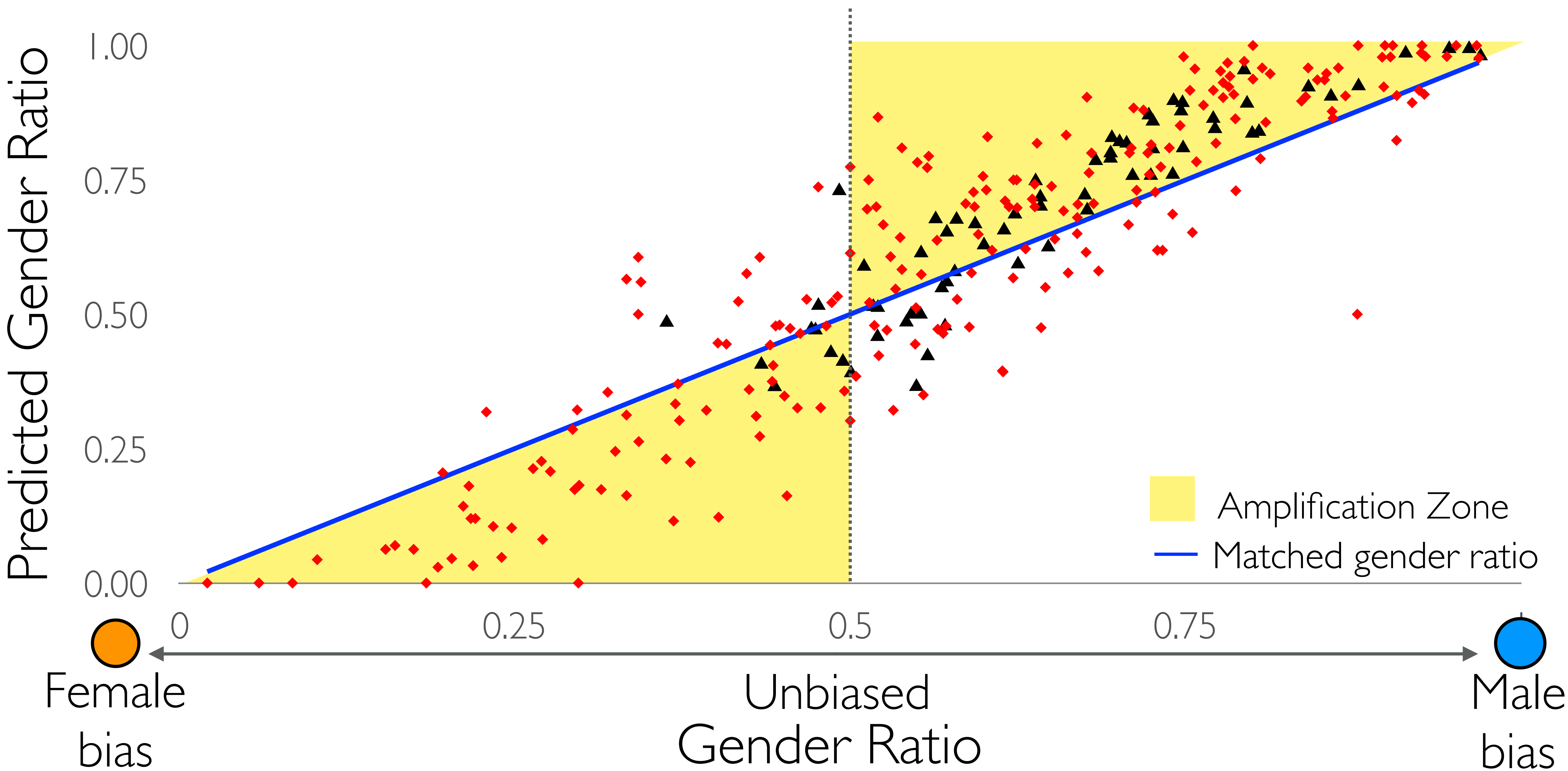
Model Bias Amplification



Model Bias Amplification

◆ imSitu Verb — 69% bias↑ .05 |bias↑|

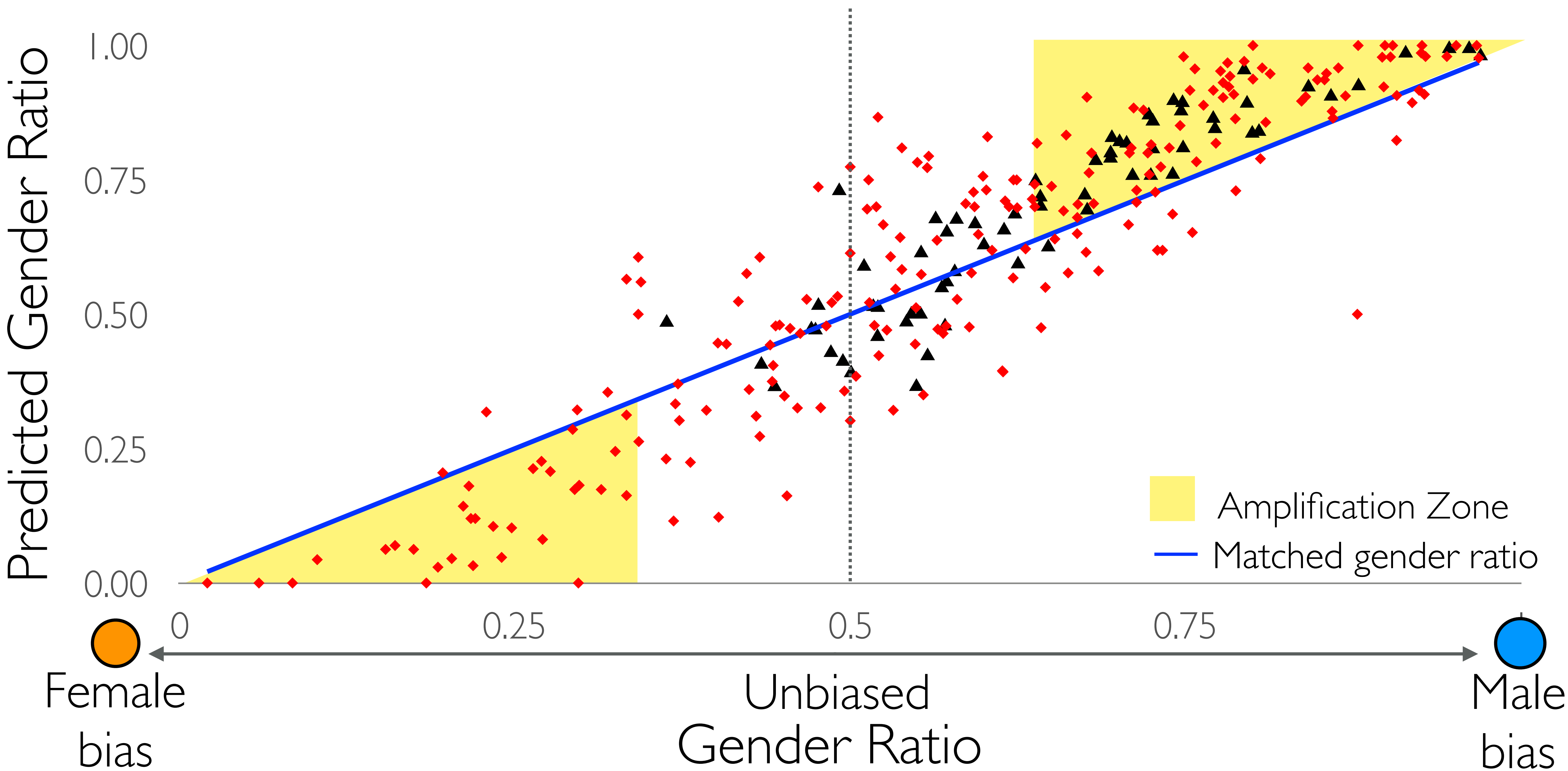
▲ COCO Noun — 73% bias↑ .04 |bias↑|



Model Bias Amplification

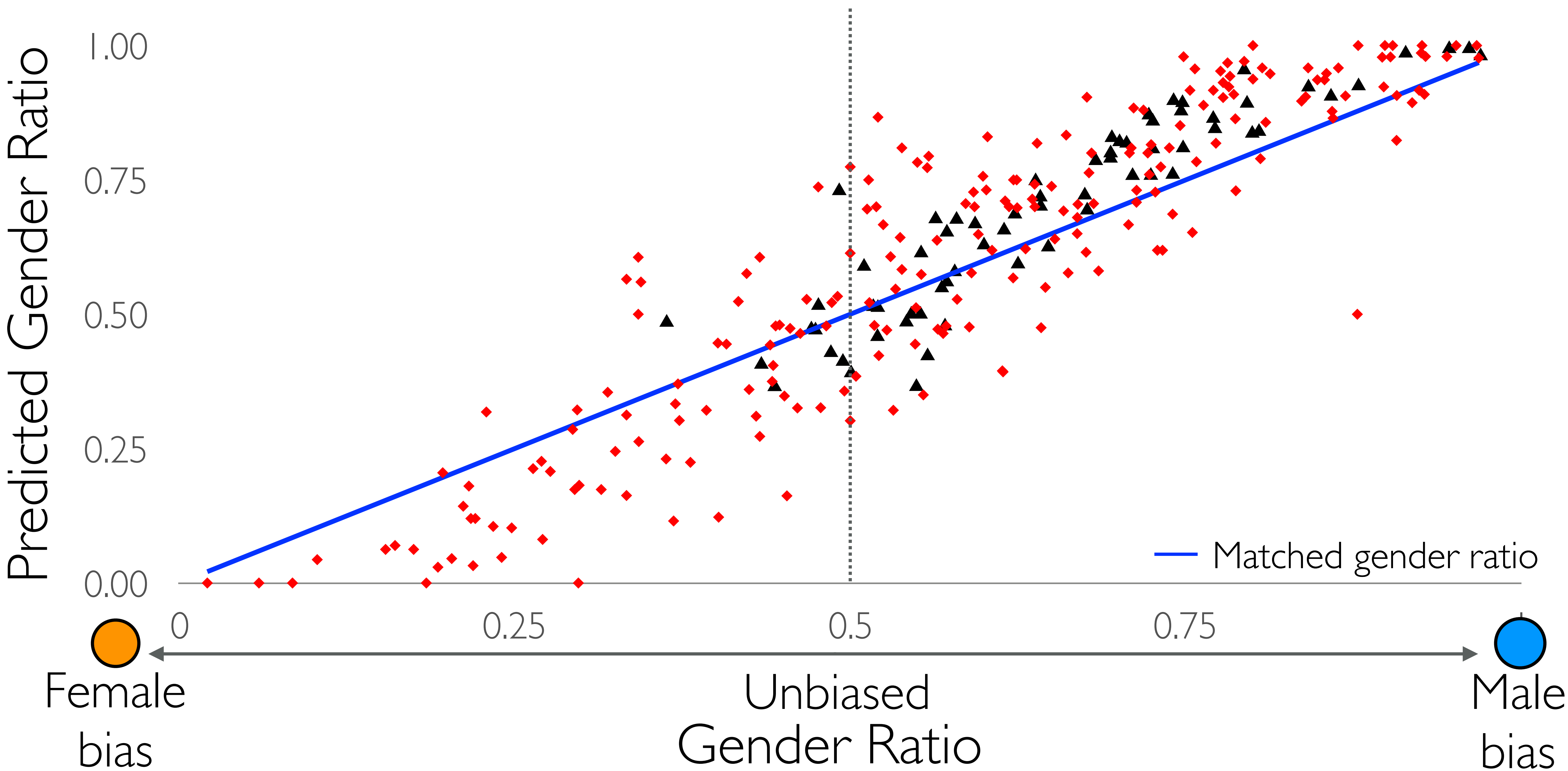
◆ imSitu Verb 69% bias↑ .05 |bias↑| > 2:1 initial bias : .07 |bias↑|

▲ COCO Noun 73% bias↑ .04 |bias↑| > 2:1 initial bias : .08 |bias↑|



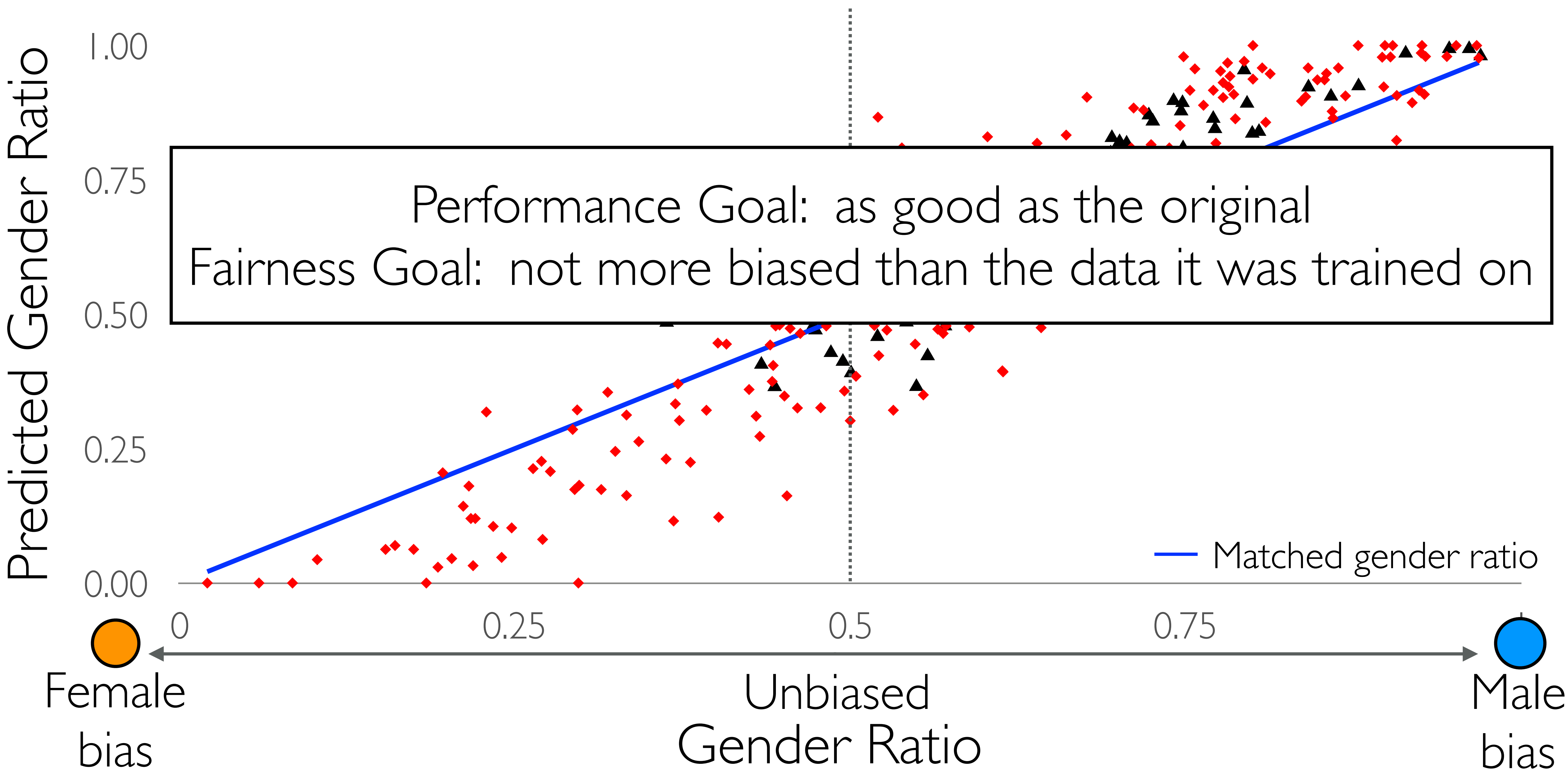
Summary

Can we remove gender bias amplification and still maintain performance?



Summary

Can we remove gender bias amplification and still maintain performance?



Performance Goal: as good as the original

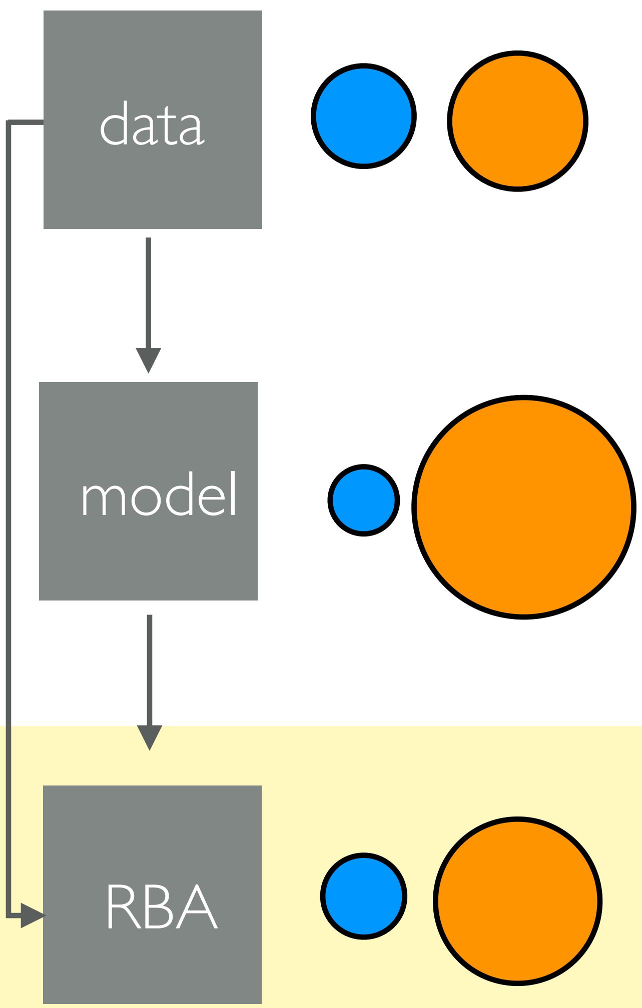
Fairness Goal: not more biased than the data it was trained on

Outline

imSitu vSRL
(events)

COCO MLC
(objects)

1. Background

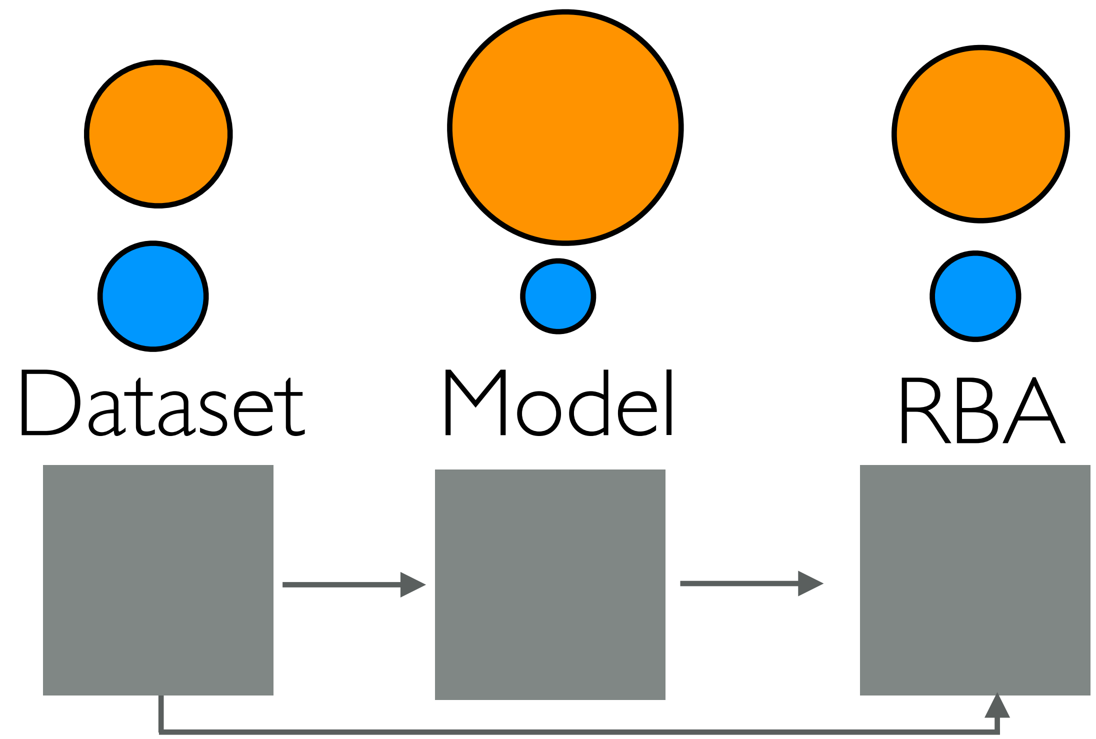


2. Dataset Bias

3. Bias Amplification

4. Reducing 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

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

base model

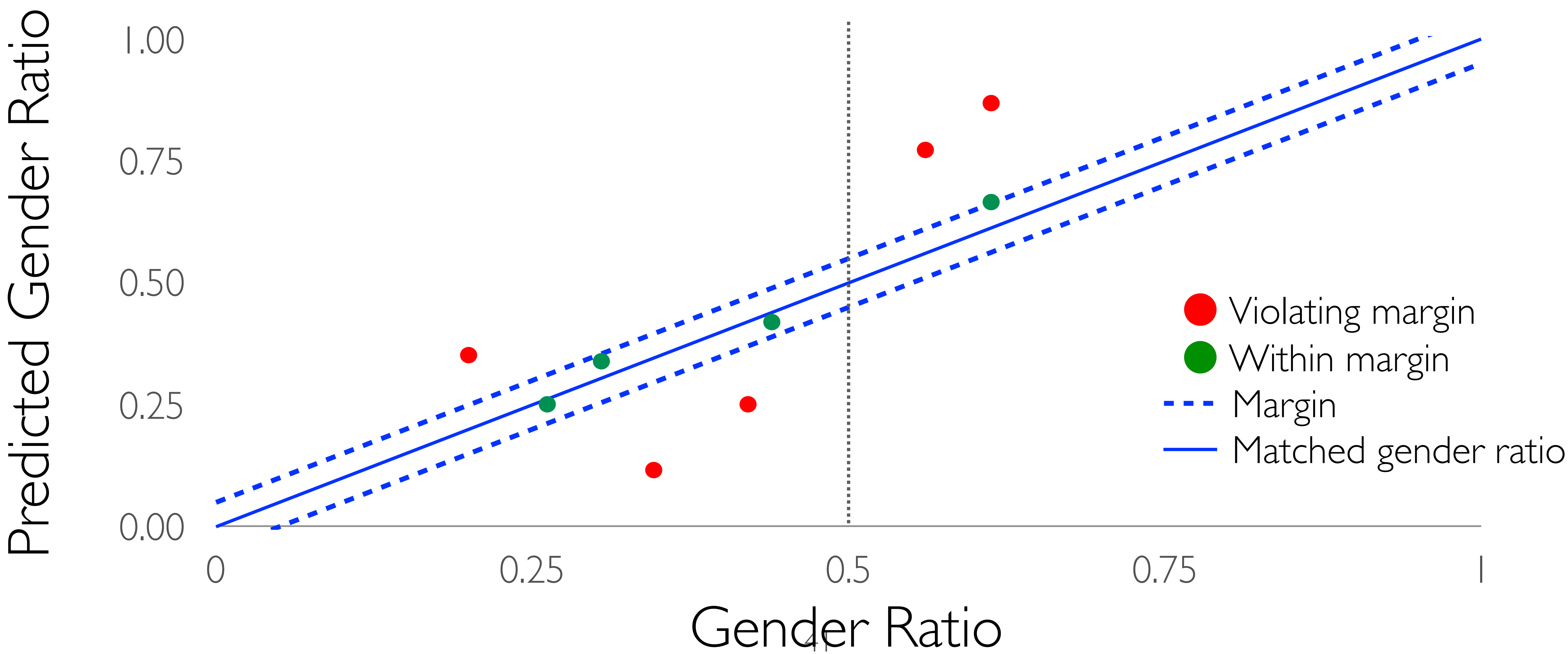
CRF Inference

Reducing Bias Amplification (RBA)

Integer Linear Program

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

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

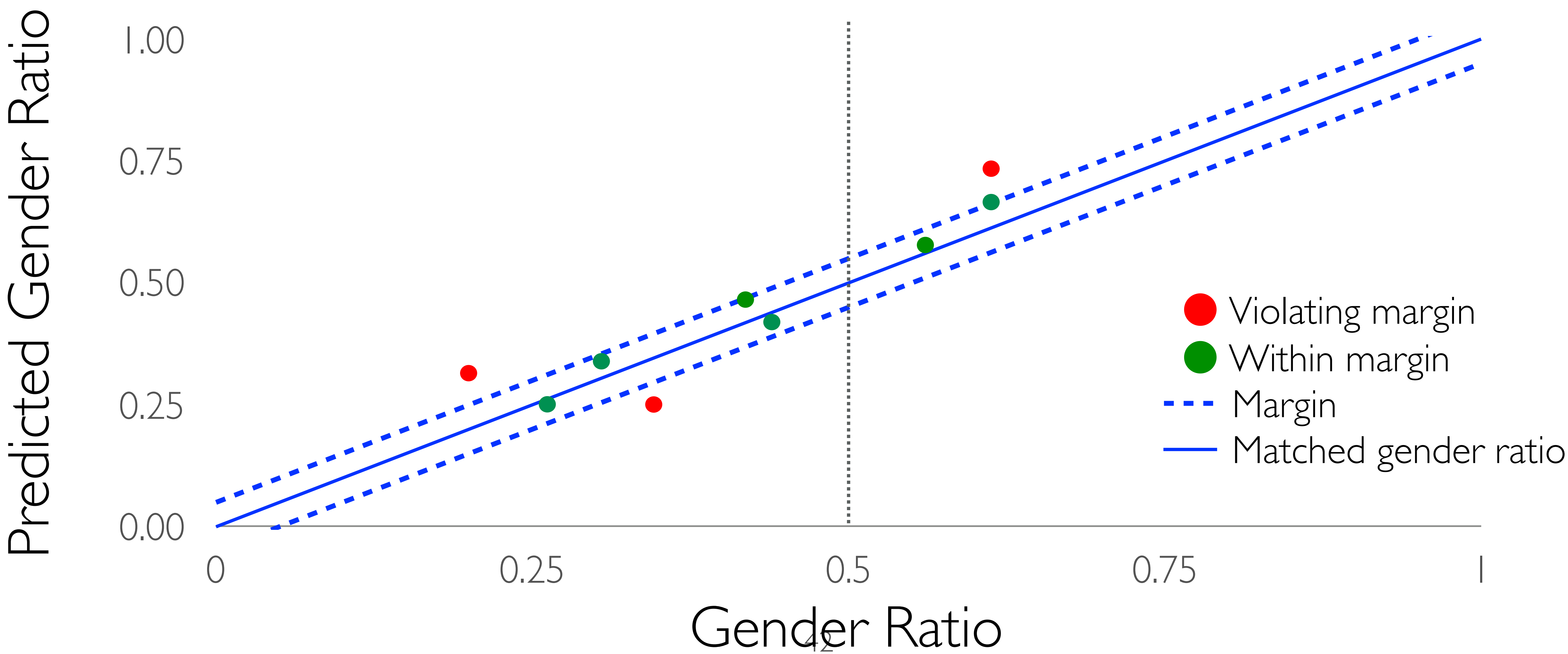


Reducing Bias Amplification (RBA)

Integer Linear Program

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

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



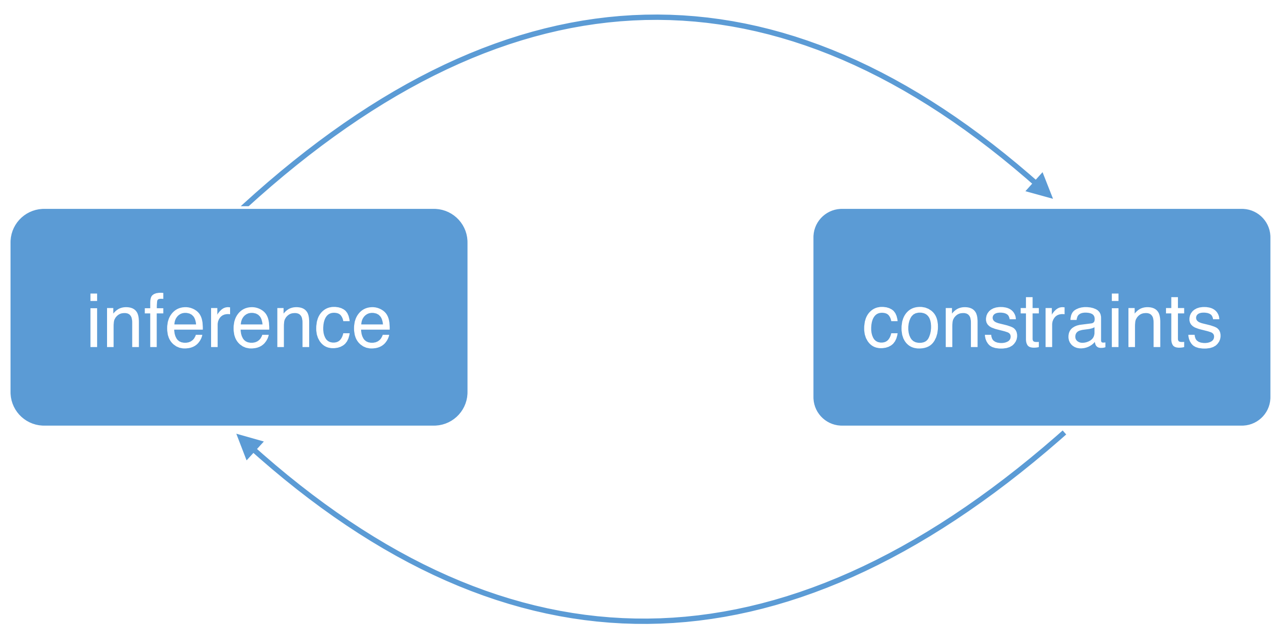
Reducing Bias Amplification (RBA)

Integer Linear Program

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

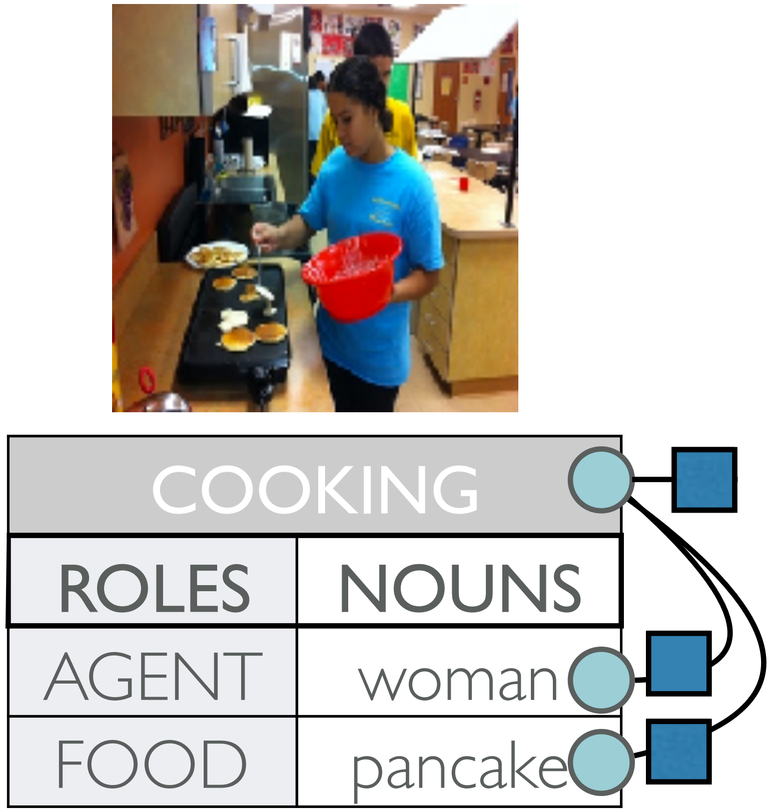
$$\forall \text{ points} \quad \Big| \text{Training Ratio} - \underbrace{\text{Predicted Ratio}}_{f(y_1 \ldots y_n)} \Big| <= \text{margin}$$

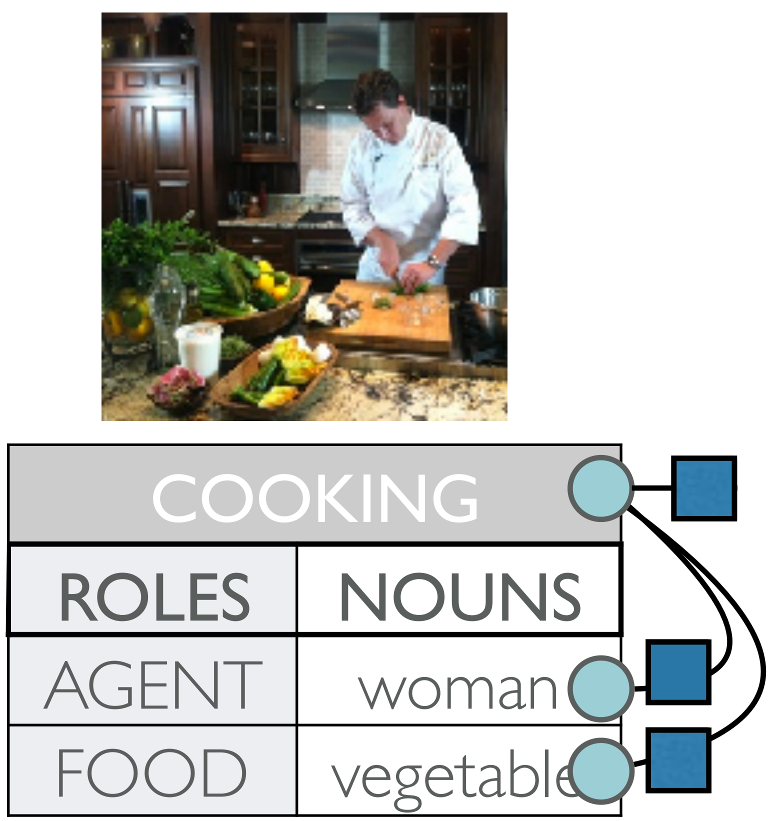
Lagrangian Relaxation



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

Lagrangian Relaxation





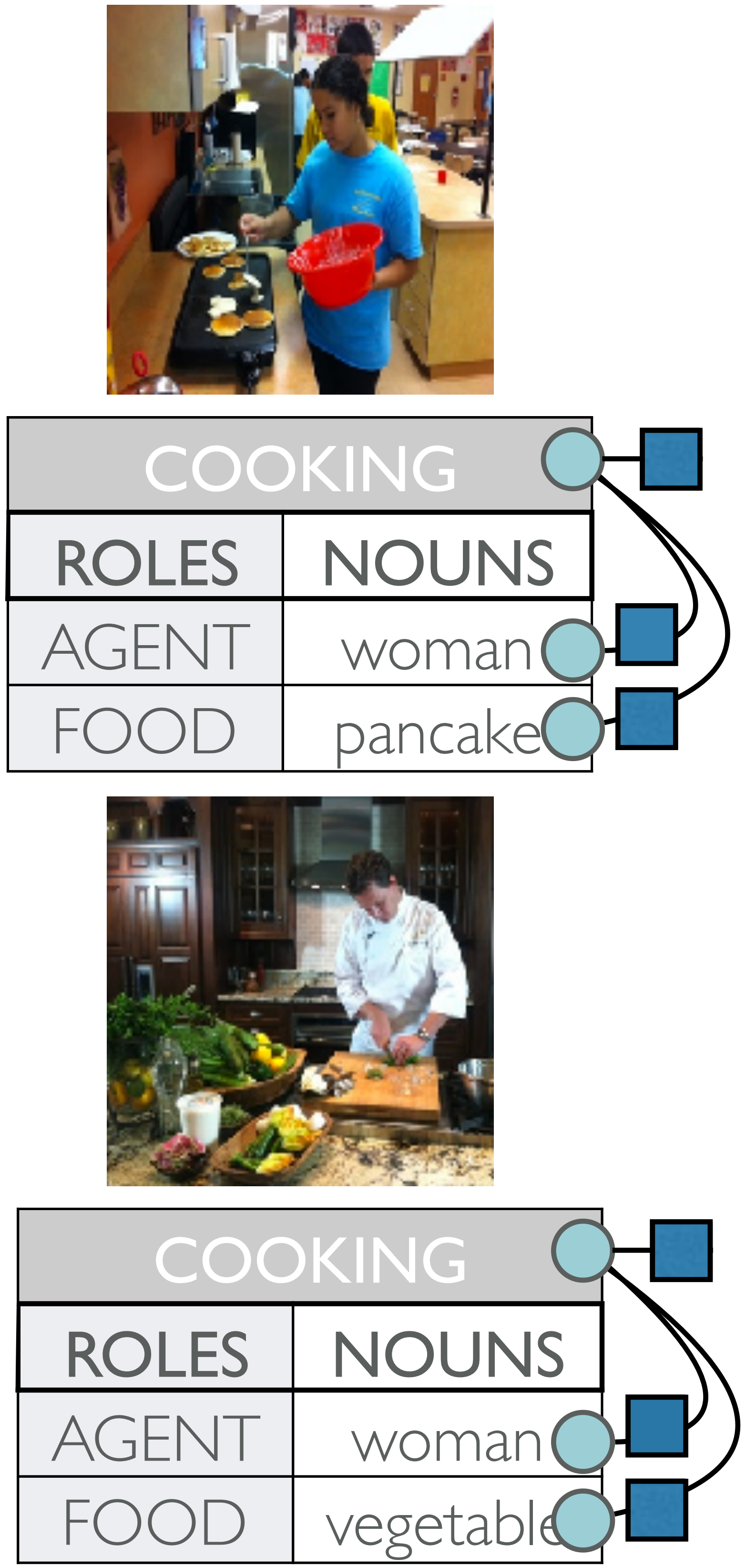
$$\sum_i \max_{y_i} \ s(y_i \, , \text{image})$$

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

(1/2)

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

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

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

- Lagrange Multiplier (λ) Per Constraint

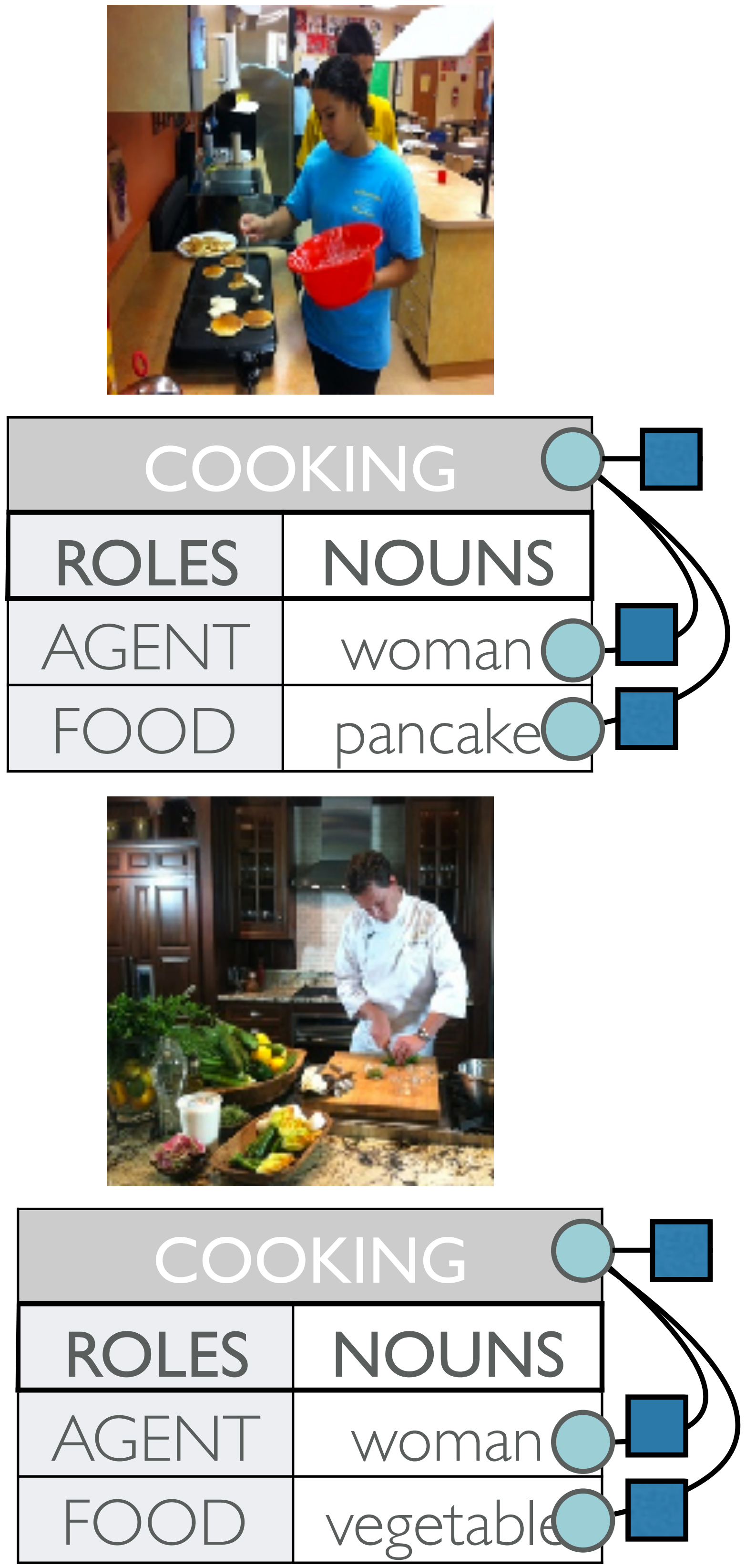
inference

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

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

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	pancake

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable

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

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

(1/2)

- Lagrange Multiplier (λ) Per Constraint

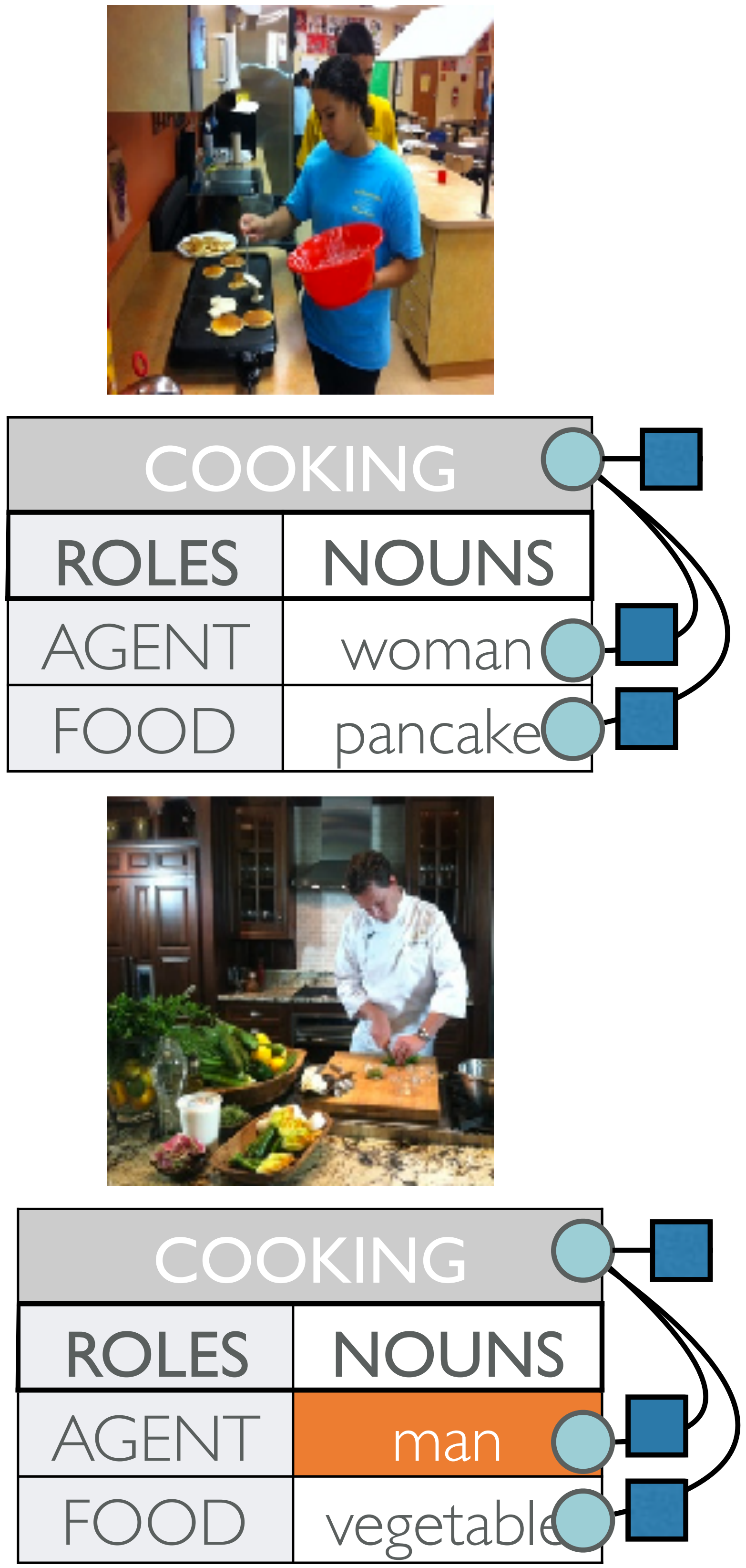
inference

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

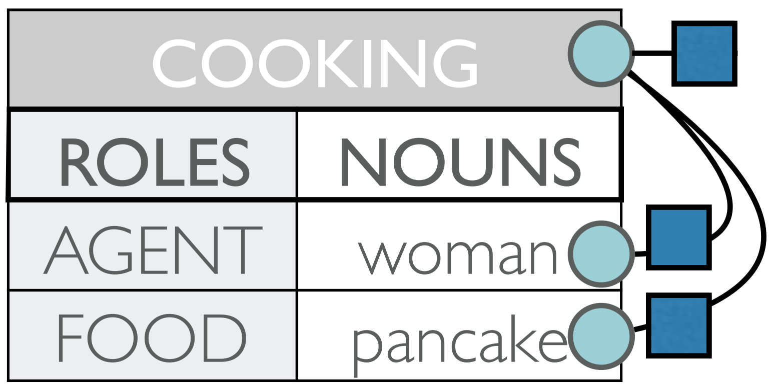
inference

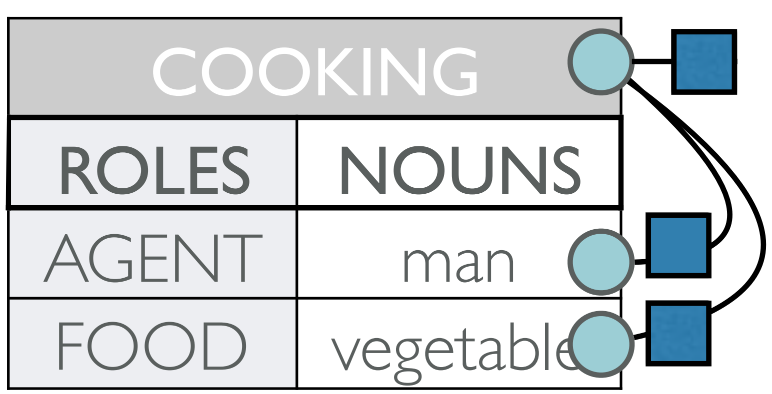
update λ

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





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

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

(1/2)

- Lagrange Multiplier (λ) Per Constraint

inference

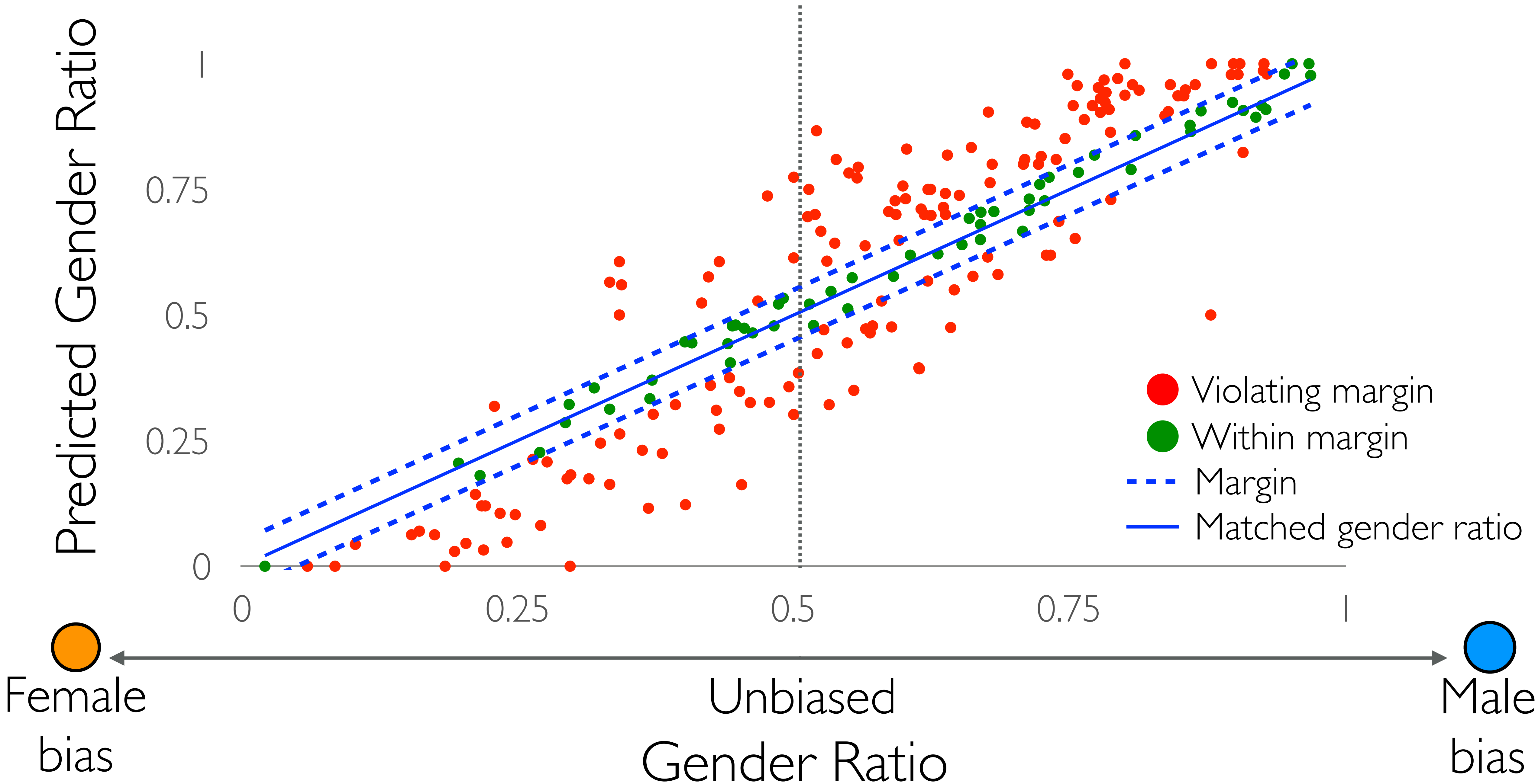
update λ

update potentials

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

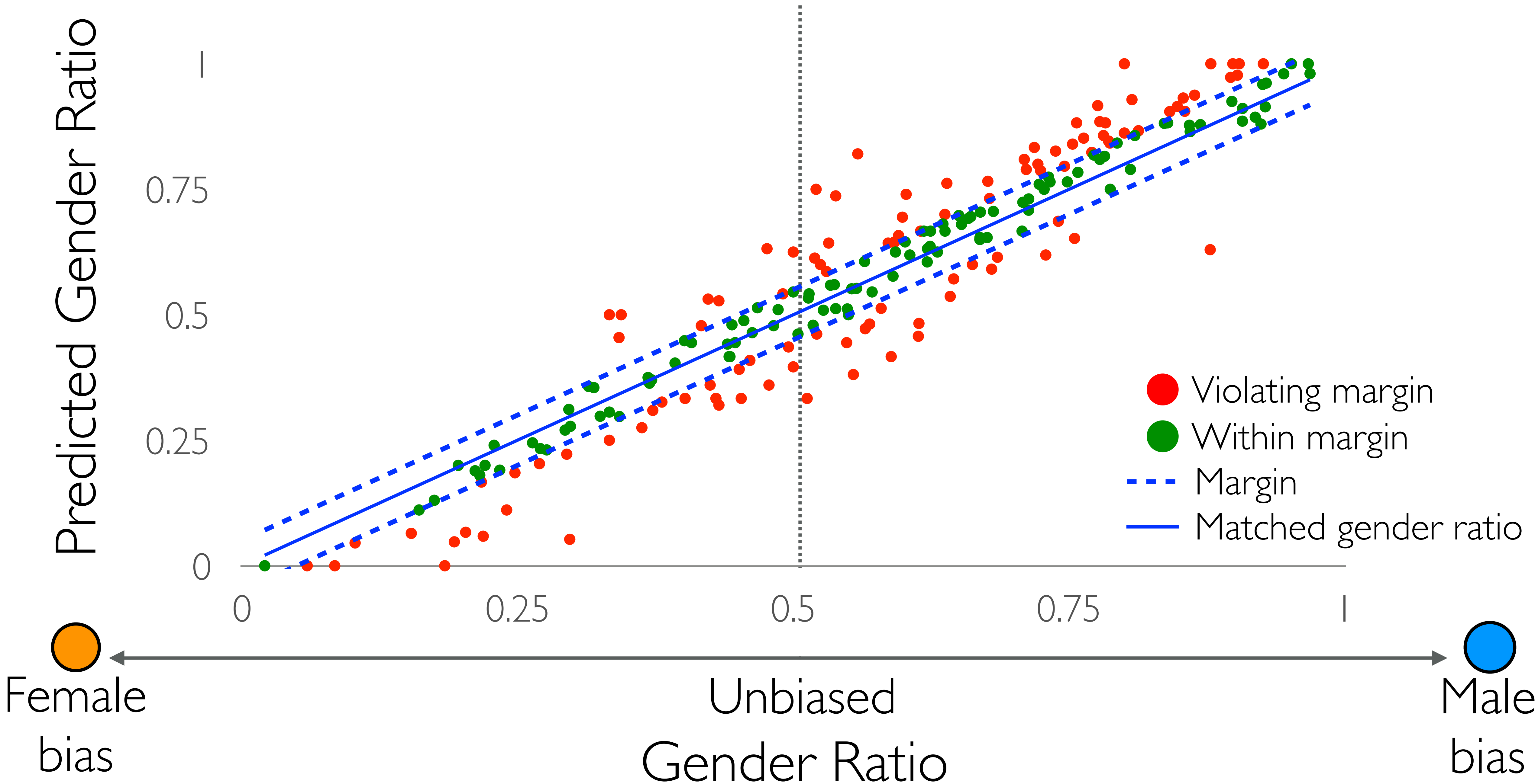
Gender Bias De-amplification in imSitu

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



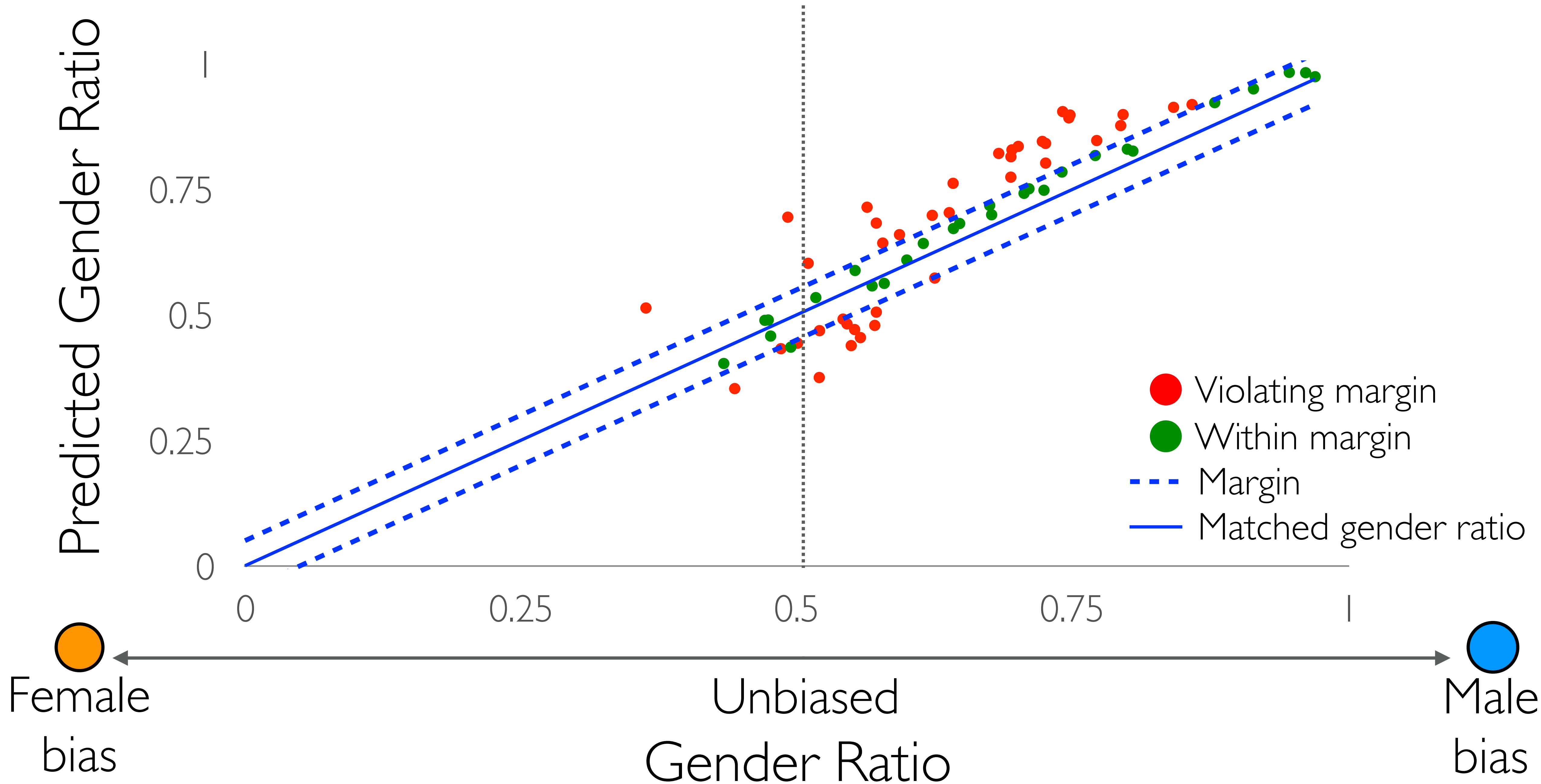
Gender Bias De-amplification in imSitu

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



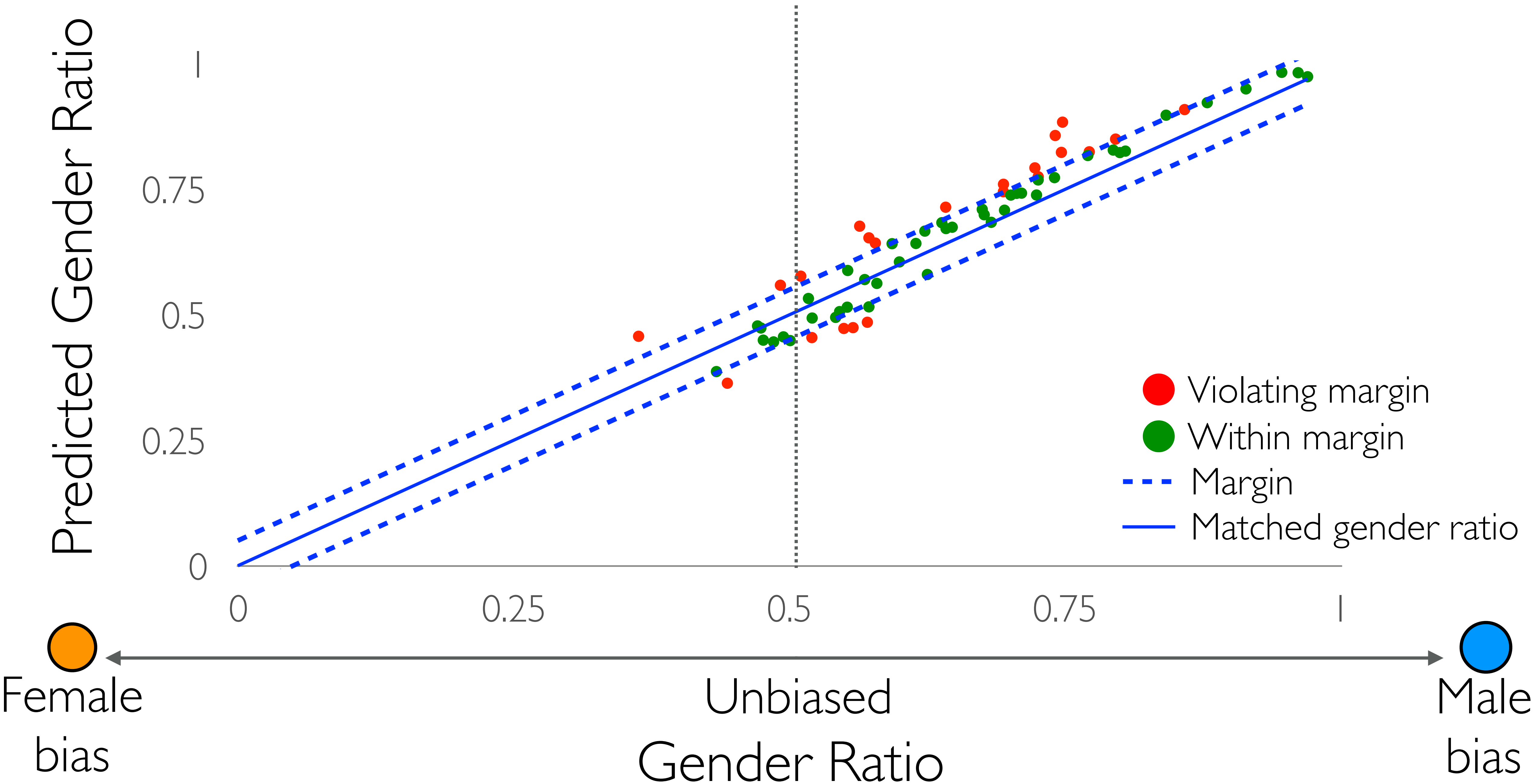
Gender Bias De-amplification in COCO

COCO Noun Violation: 60.6% .032 |bias↑| 45.27 mAP



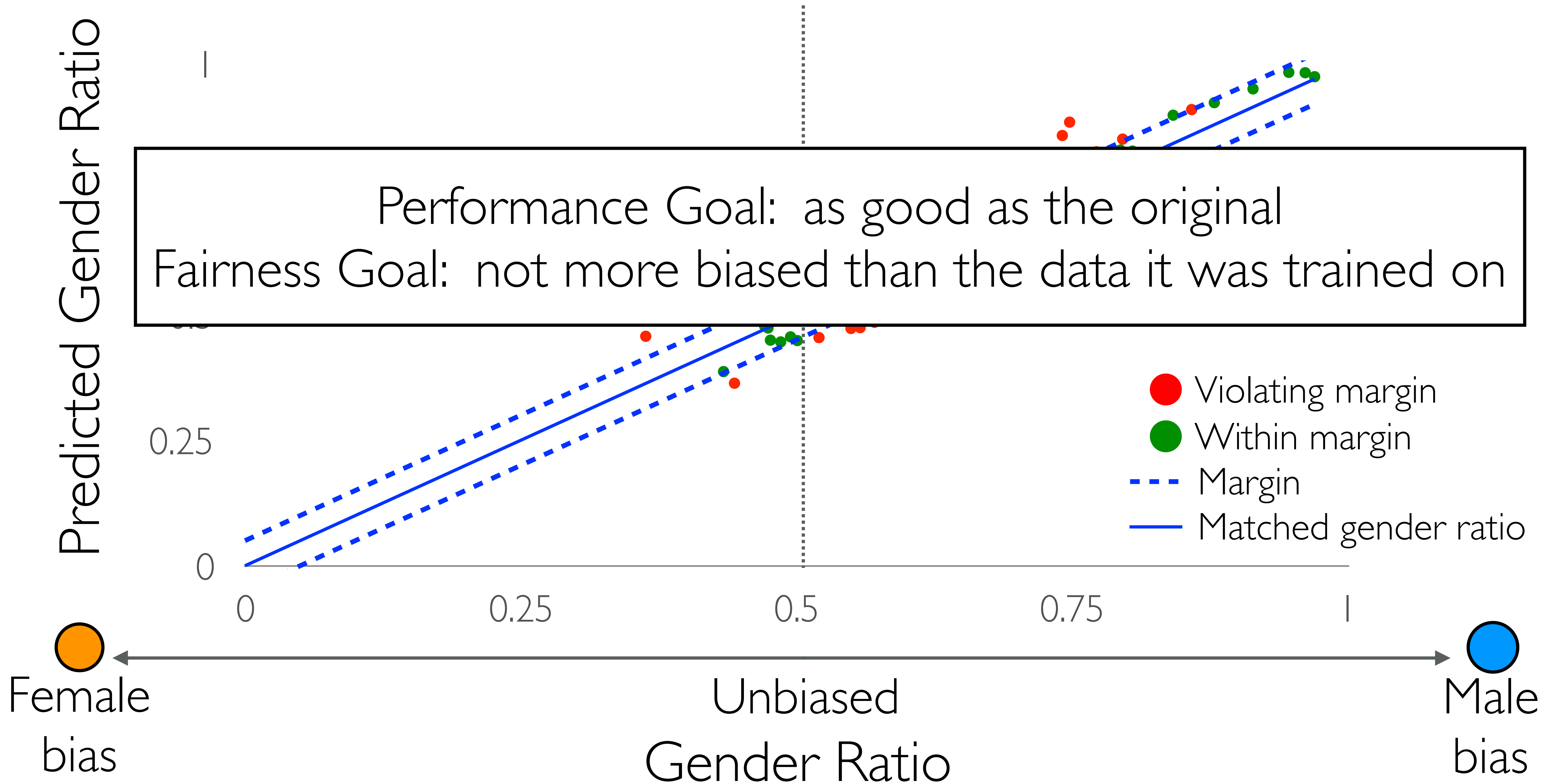
Gender Bias De-amplification in COCO

COCO Noun	Violation: 60.6%	.032 |bias↑|	45.27 mAP
w/ RBA	Violation: 36.4%	.022 |bias↑|	45.19 mAP



Gender Bias De-amplification in COCO

COCO Noun	Violation: 60.6%	.032 |bias↑|	45.27 mAP
w/ RBA	Violation: 36.4%	.022 |bias↑|	45.19 mAP



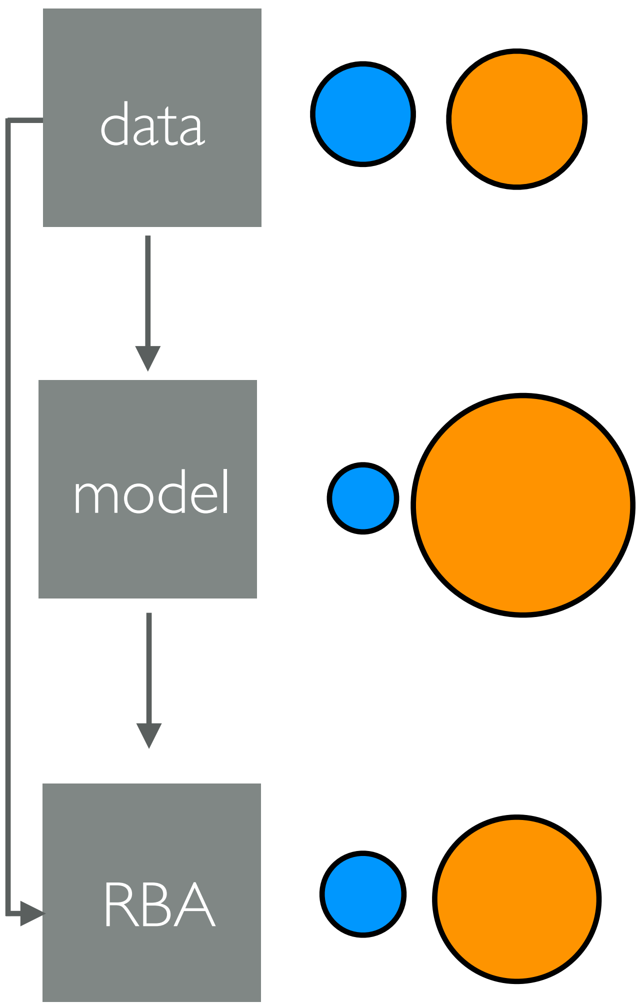
Performance Goal: as good as the original

Fairness Goal: not more biased than the data it was trained on

Contributions

imSitu vSRL
(events)

COCO MLC
(objects)



High dataset gender bias

38% (objects) 47% (events) exhibit strong bias

Models amplify existing gender bias

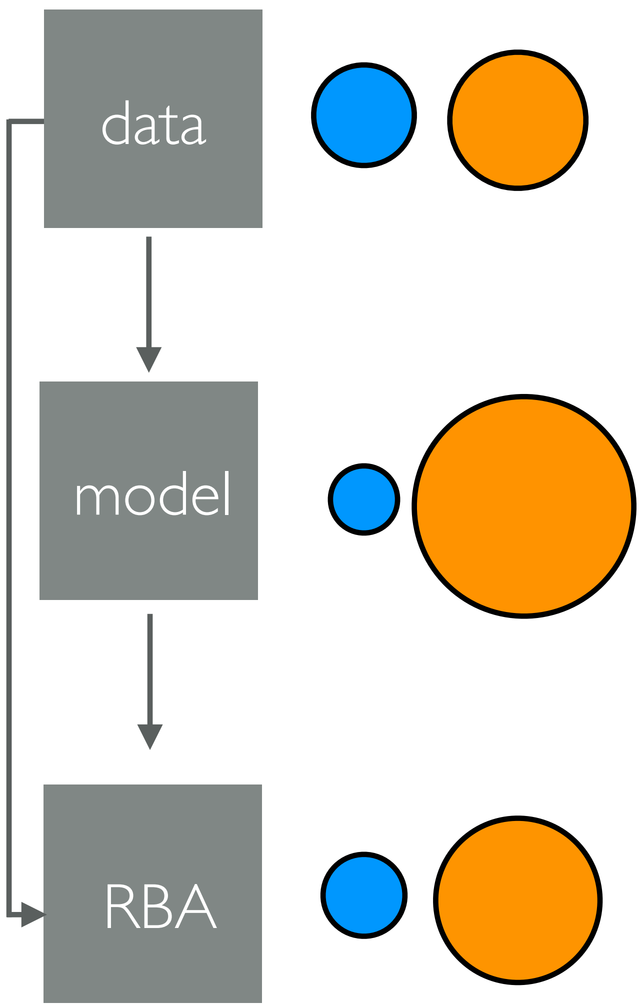
~70% objects and events have bias amplification

Reducing bias amplification

~50% reduction in amplification

Insignificant loss in performance

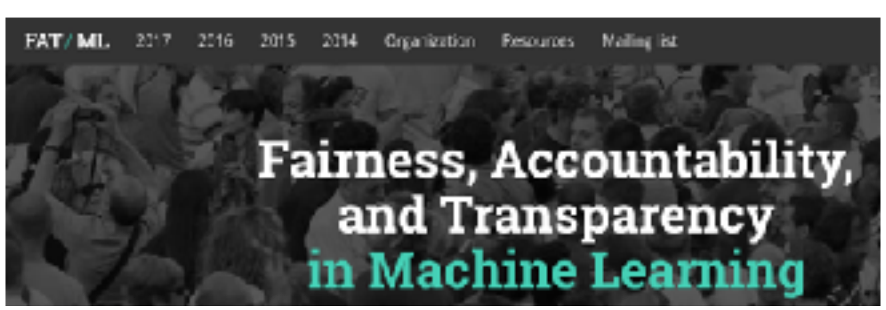
Future Work



Can existing data be made more balanced?

Do all models amplify equally?
i.e. different objectives

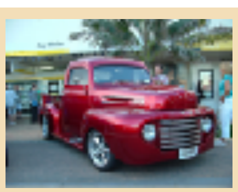
Other direct applications?
i.e. co-ref, racial bias



Questions?

https://github. com/uclanlp/reducingbias

imSitu vSRL
(events)



COCO MLC
(objects)