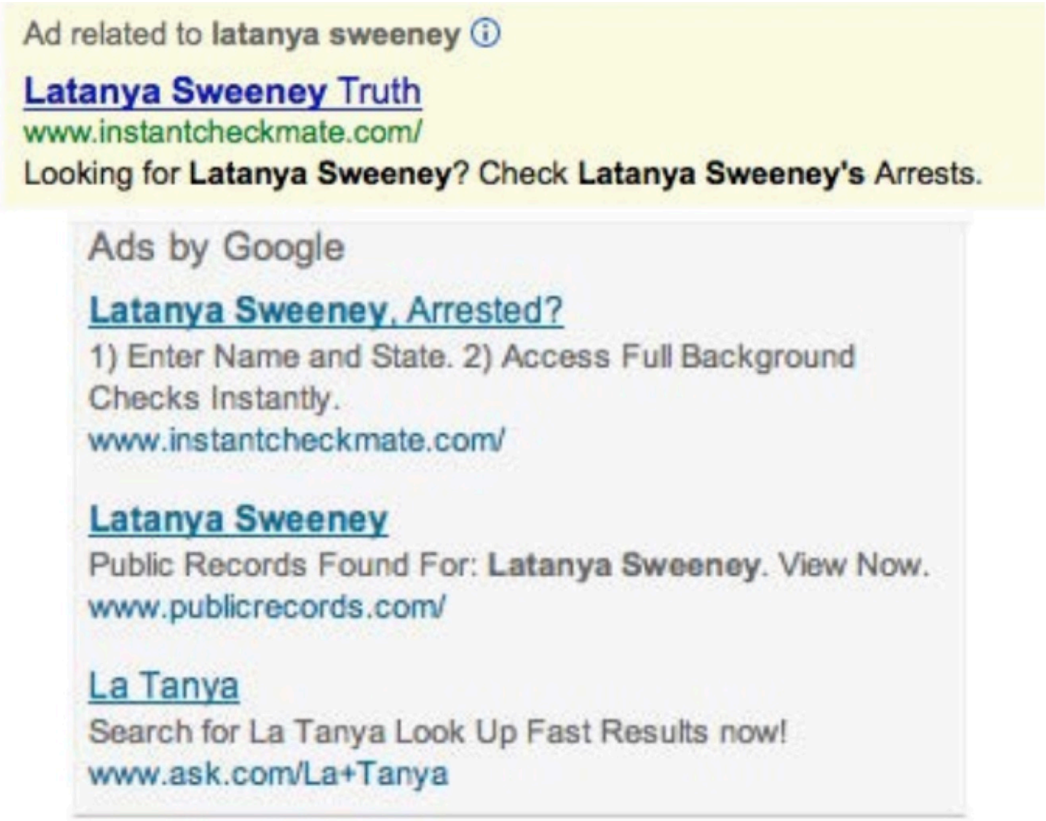
Most slides are adapted from MLSS 2021 tutorial

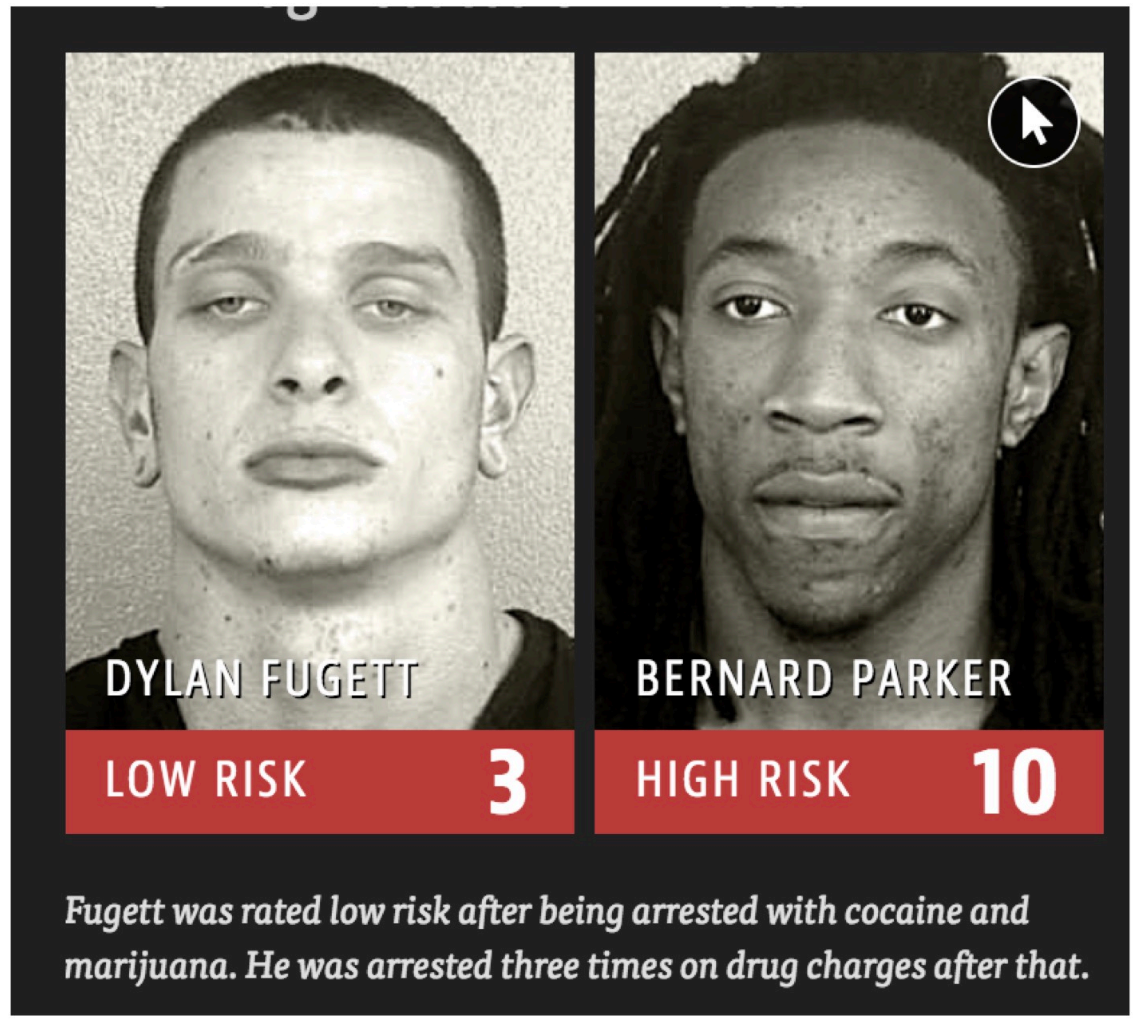
COMP6211I: Trustworthy Machine Learning

Fairness

Minhao CHENG

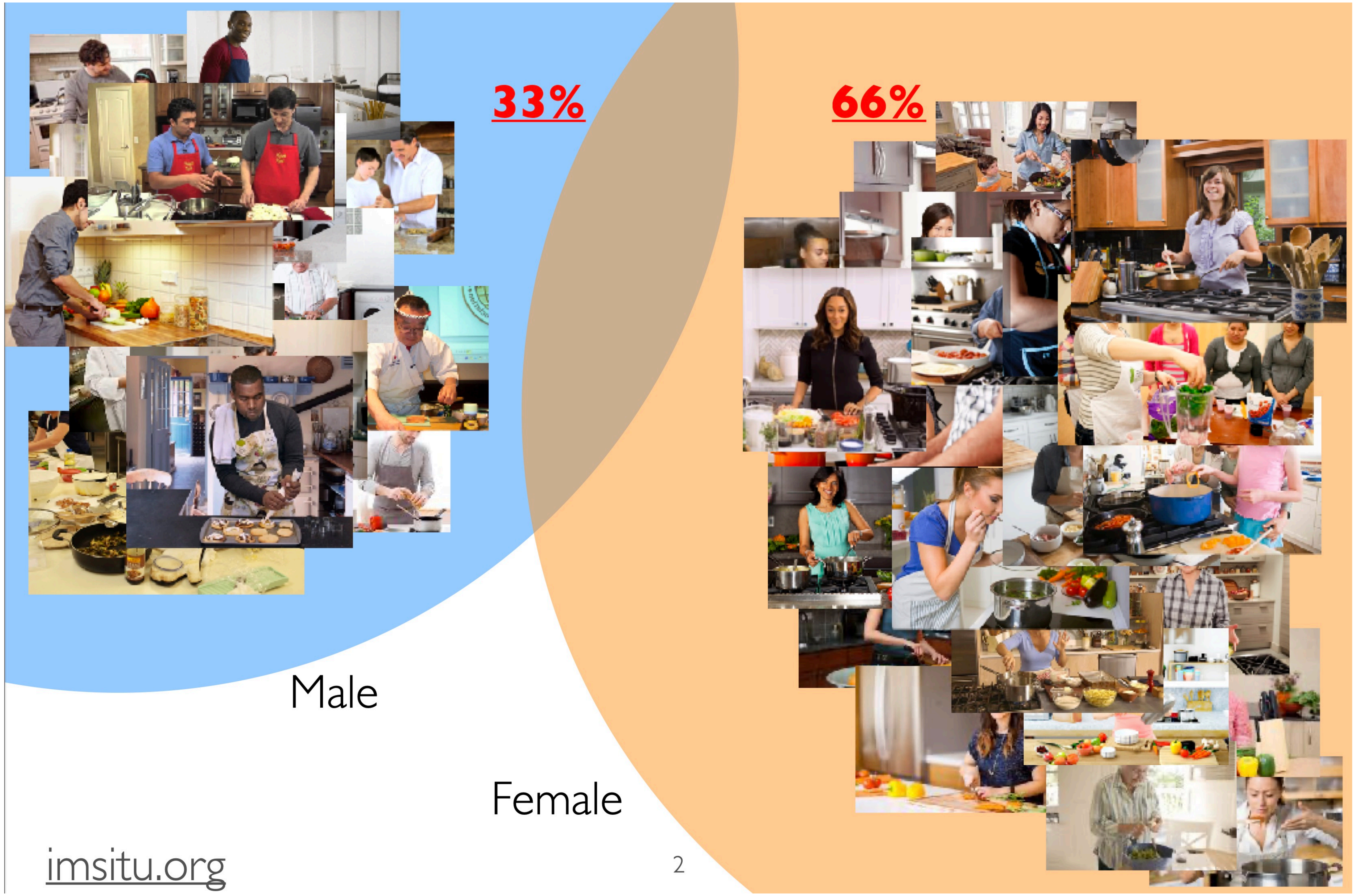
Machine learning ethics



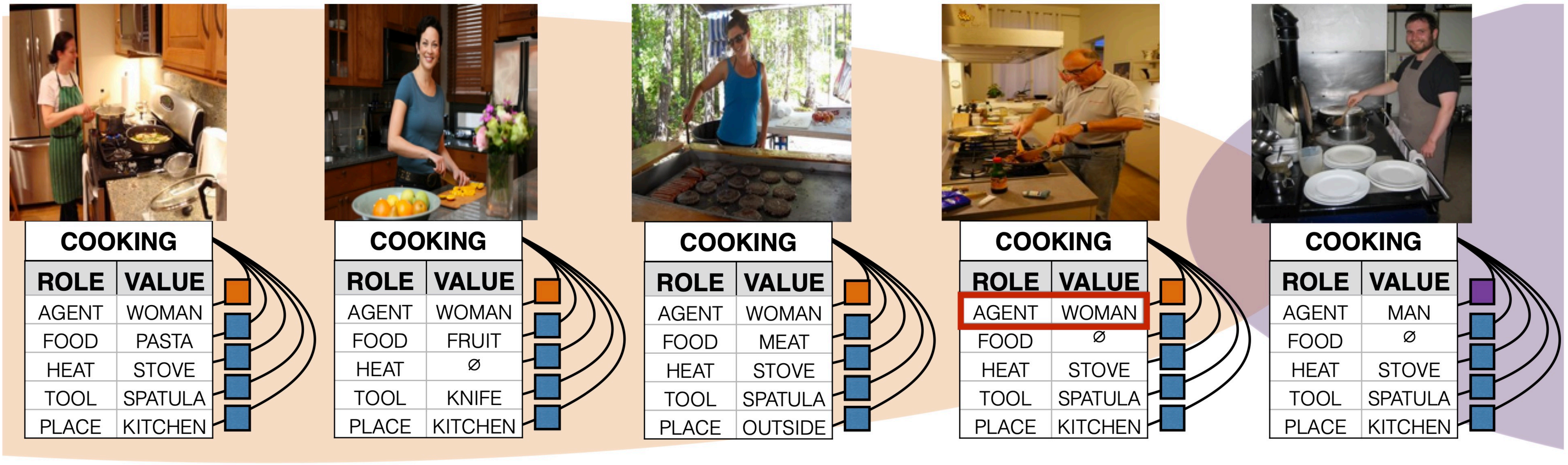


Ethical machine learning matters
in **high-stakes** domains

Group bias example: gender bias



Group bias example: gender bias



Fairness in Machine Learning

- Group fairness
    - Don’t discriminate unnecessarily between **protected** groups (race, gender, sexuality, religion, etc.)
- Individual fairness
    - Treat similar individuals similarly

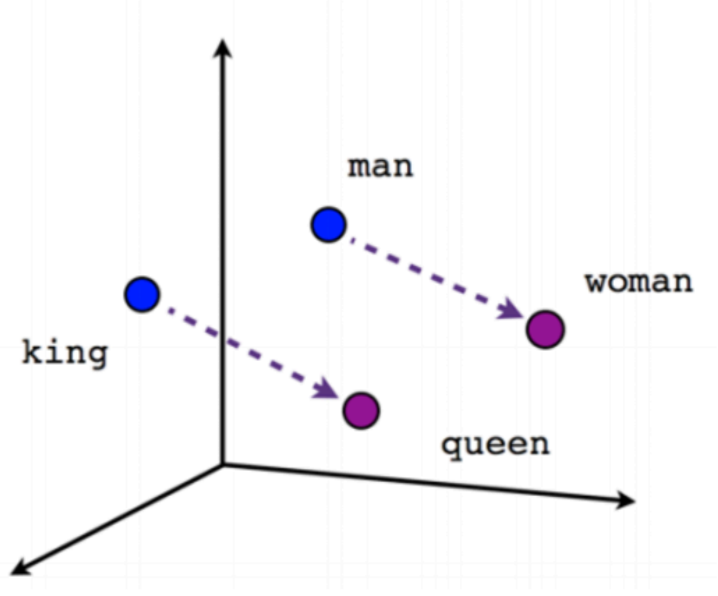
Bias

- Found in language data, learned by humans and ML
- Stereotyped bias: “problematic where such information is derived from aspects of human culture known to leant to harmful behavior”
- Prejudiced actions are taken based on stereotyped bias

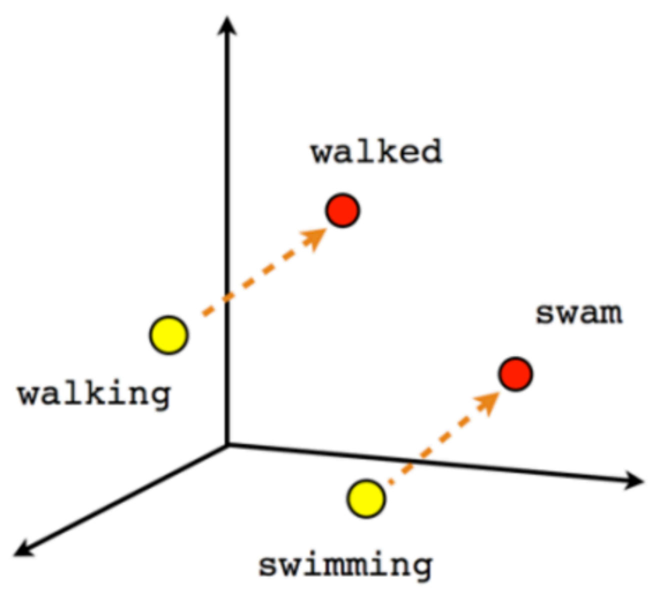
How to measure word embedding bias?

- Humans:
  - Implicit Association Test
  - Response time differs when humans pair concepts that they find similar compared to concepts that they find different
- Machines:
  - Word embeddings
  - Measure cosine distance between embedding vectors

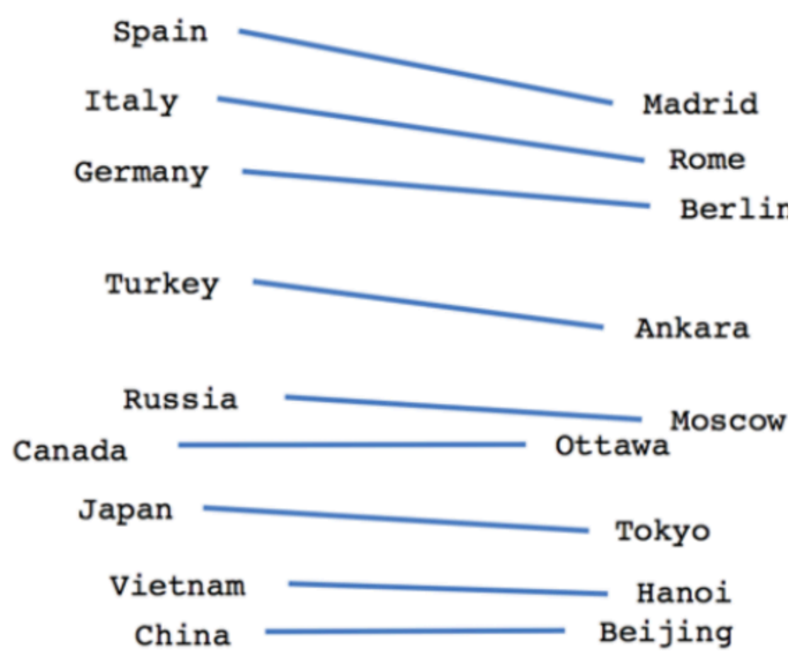
Word embeddings



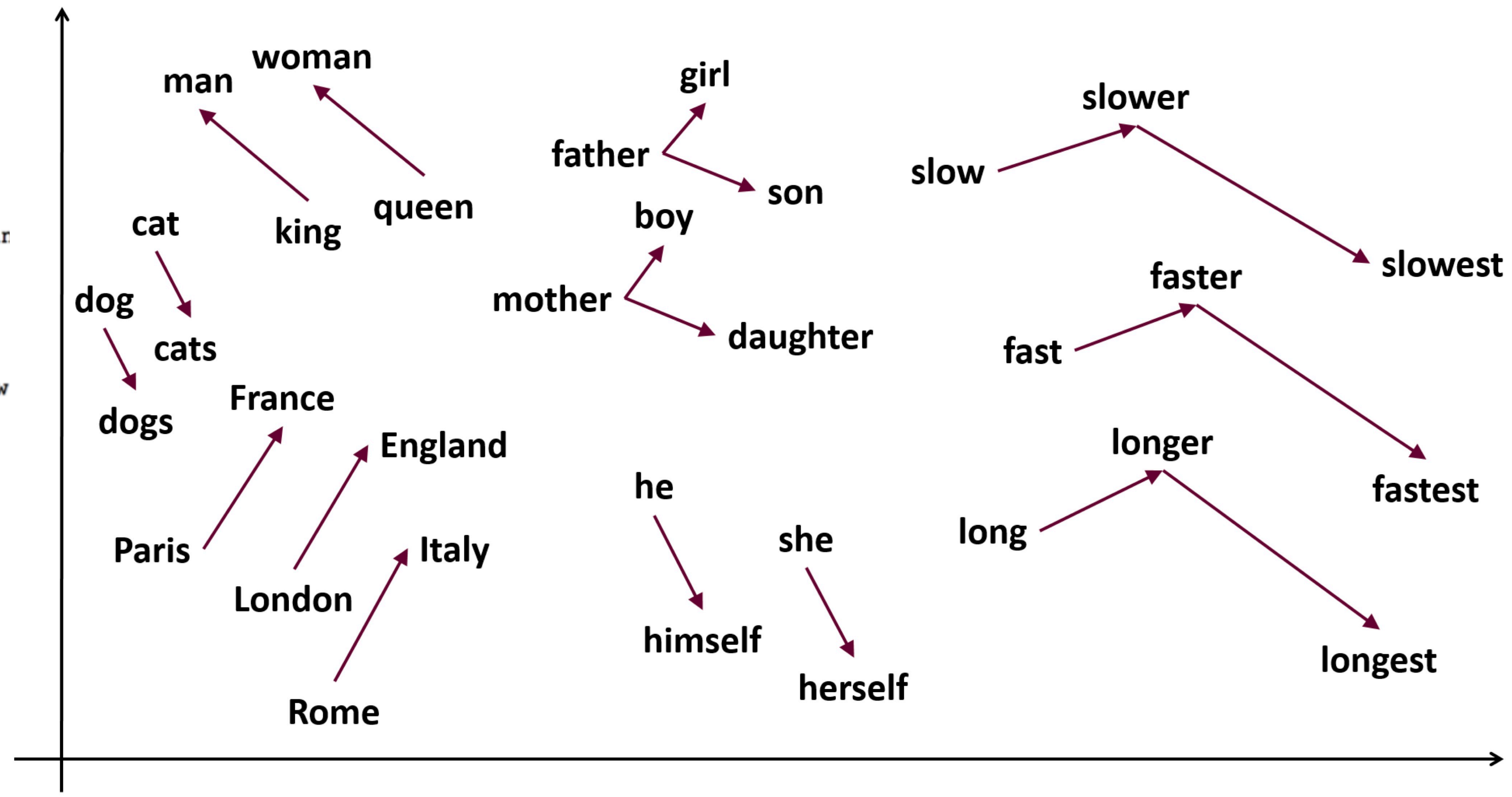
Male-Female



Verb tense



Country-Capital



N: population size
d: effect size
p : p-value

N_T: number of target words
N_A: number of attribute words

Target words	Attrib. 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}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant	(5)	26	1.17	10^{-5}	32×2	25×2	1.41	10^{-8}
Eur.-American vs Afr.-American names	Pleasant vs unpleasant from (5)	(7)	Not applicable			16×2	25×2	1.50	10^{-4}
Eur.-American vs Afr.-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}

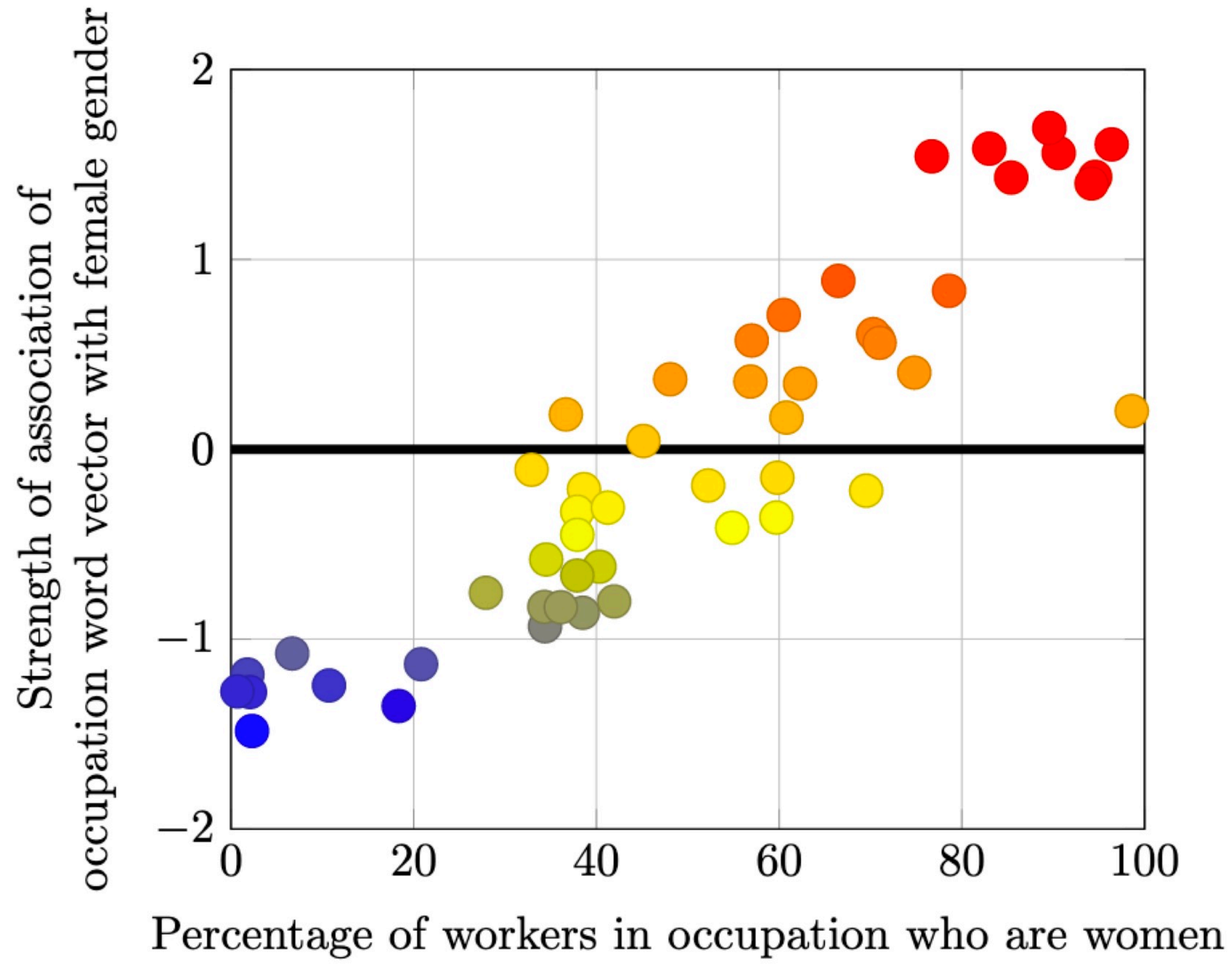


Figure 1: Occupation-gender association. Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

Word Embedding Factual Association Test

Target word

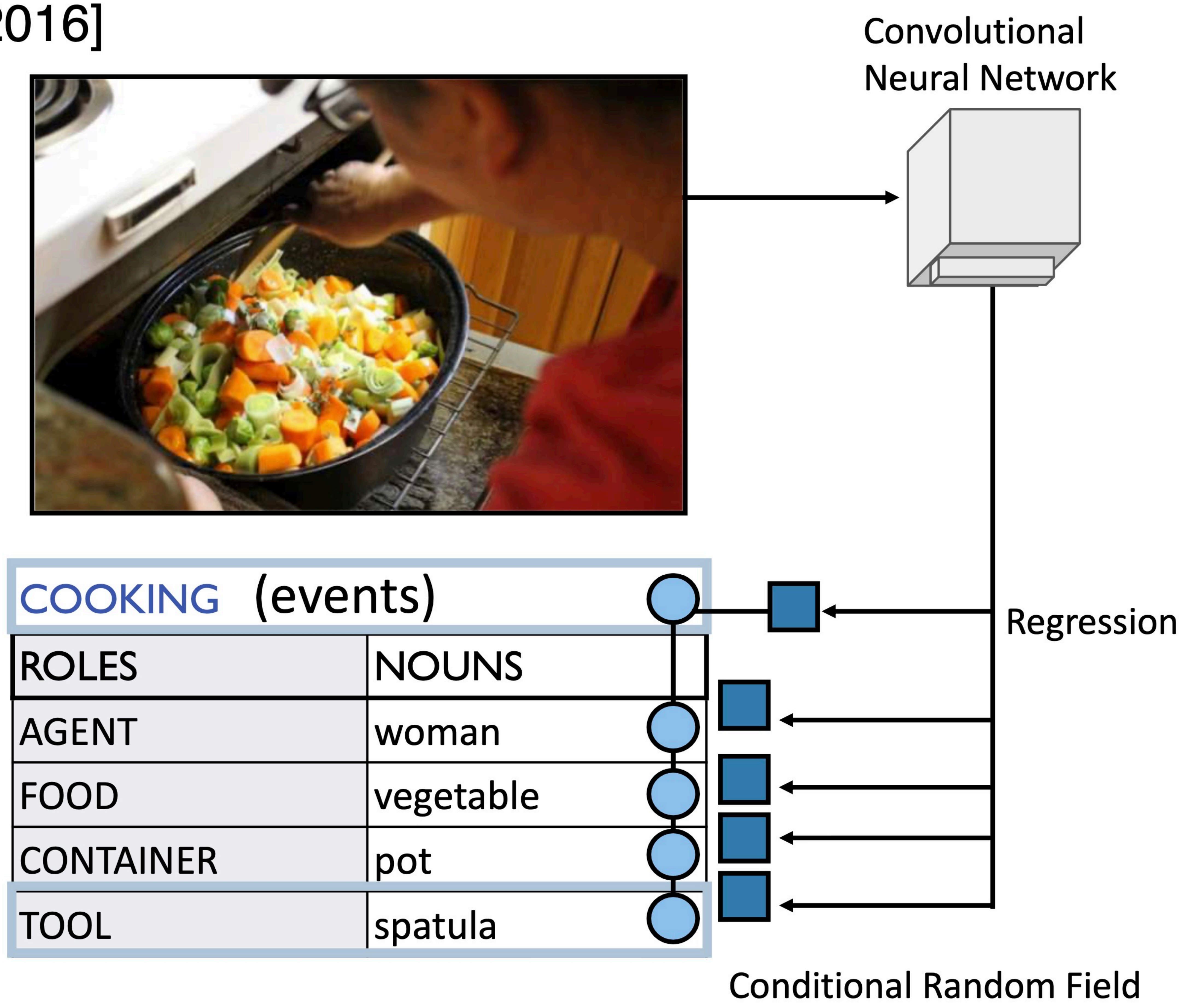
Target attributes

$$s(w, A, B) = \frac{\mathbf{mean}_{a \in A}\mathbf{cos}(\vec{w}, \vec{a}) - \mathbf{mean}_{b \in B}\mathbf{cos}(\vec{w}, \vec{b})}{\mathbf{std\text{-}dev}_{x \in A \cup B}\mathbf{cos}(\vec{w}, \vec{x})}$$

Visual semantic role labeling

imSitu Visual Semantic Role Labeling (vSRL)

[Yatskar et al. 2016]



Identifying data bias

$$b(o, g) = \frac{c(o, g)}{\sum_{g' \in G} c(o, g')},$$

where $c(o, g)$ is the number of occurrences of o and g in a corpus. For example, to analyze how genders of agents and activities are co-related in vSRL, we define the gender bias toward `man` for each verb $b(verb, \texttt{man})$ as:

$$\frac{c(verb, \texttt{man})}{c(verb, \texttt{man}) + c(verb, \texttt{woman})}. \quad (1)$$

If $b(o, g) > 1/\|G\|$, then o is positively correlated with g and may exhibit bias.

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

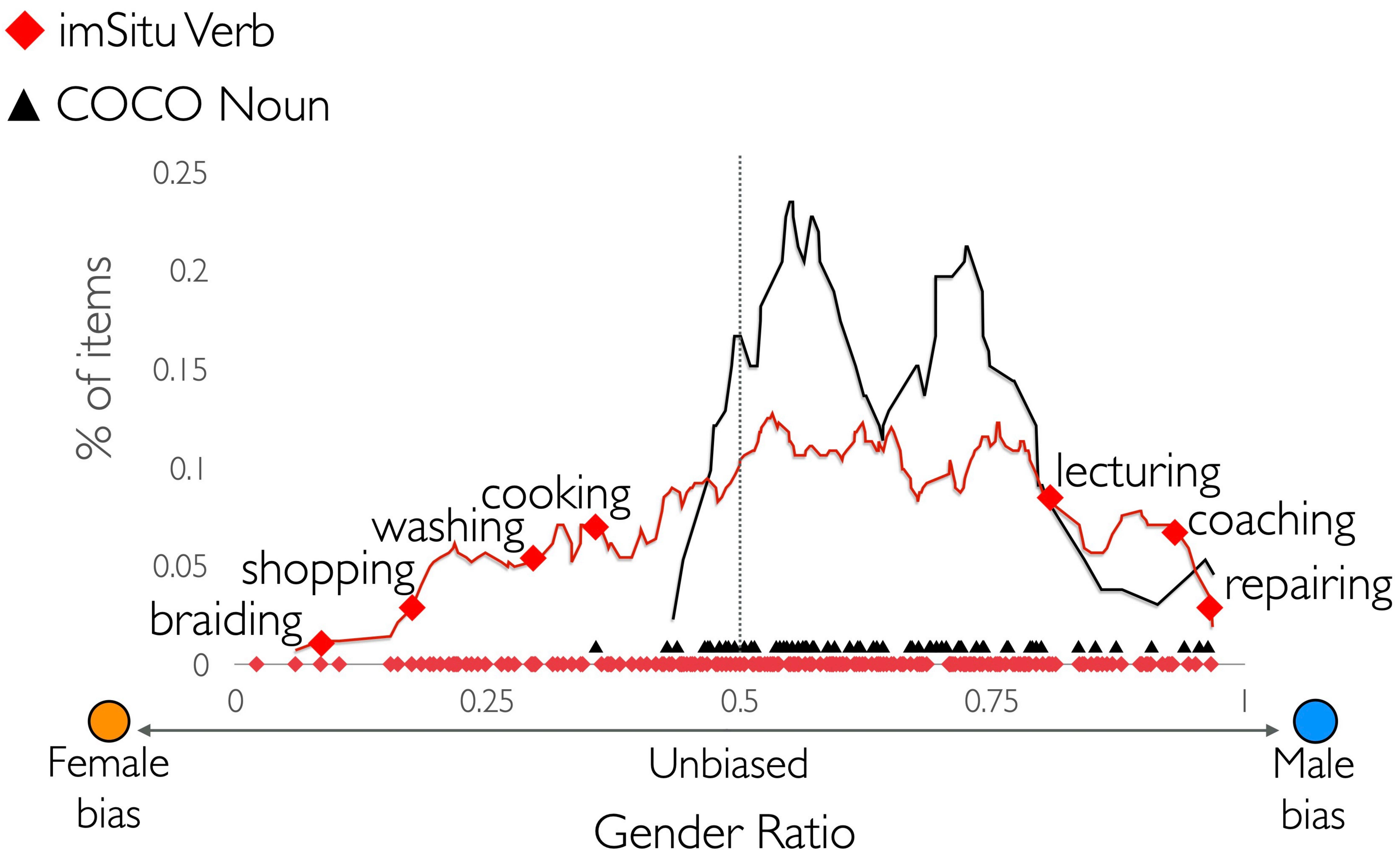
- ▲ snowboard
- ● woman
- ● man

MAN	
snowboard	yes
refrigerator	no
bowl	no

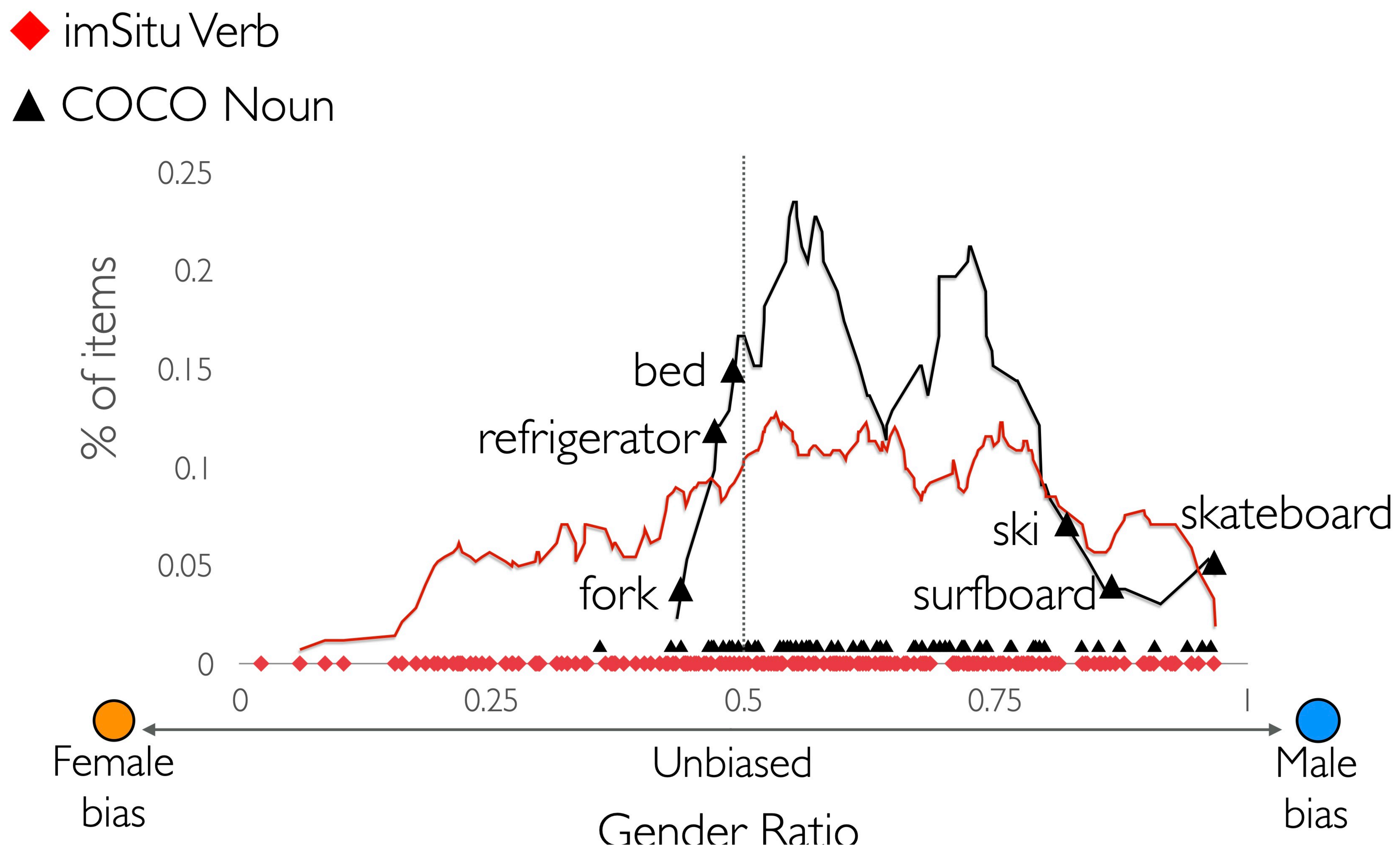
WOMAN	
snowboard	yes
refrigerator	no
bowl	no

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

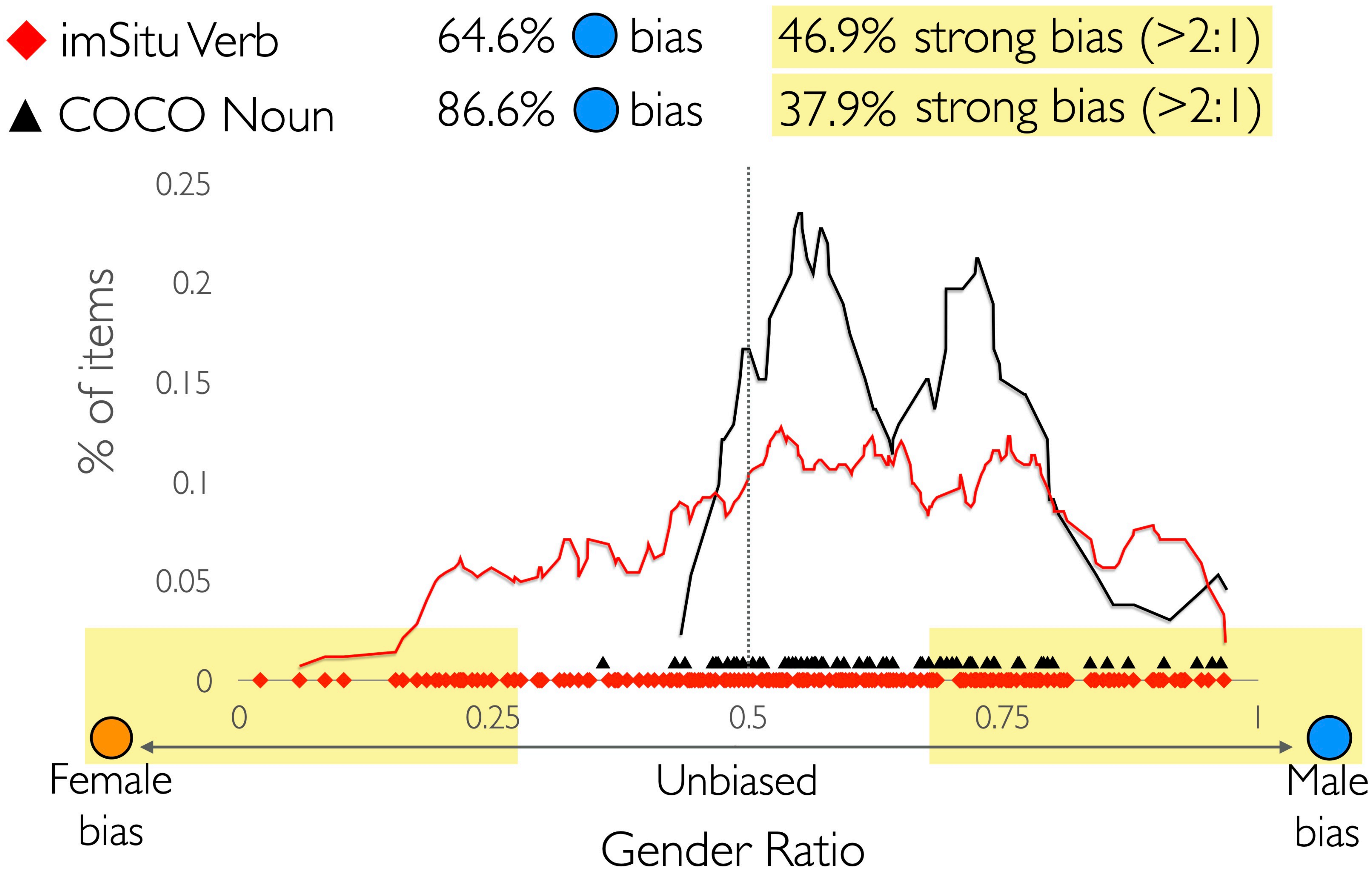
Gender dataset bias



Gender dataset bias



Gender dataset bias



Evaluating bias amplification

- $$\frac{1}{|O|}\sum_{g}\sum_{o\in\{o\in O|b^*(o,g)>1/\|G\|\}}\tilde{b}(o,g)-b^*(o,g).$$

  - $\tilde{b}(o, g)$: bias score on unlabeled evaluation set of images that has been annotated by a predictor

  - $b*(o, g)$: bias score on training set

Evaluating bias amplification

Predicted Gender Ratio (verb)

<u>Development Set</u>

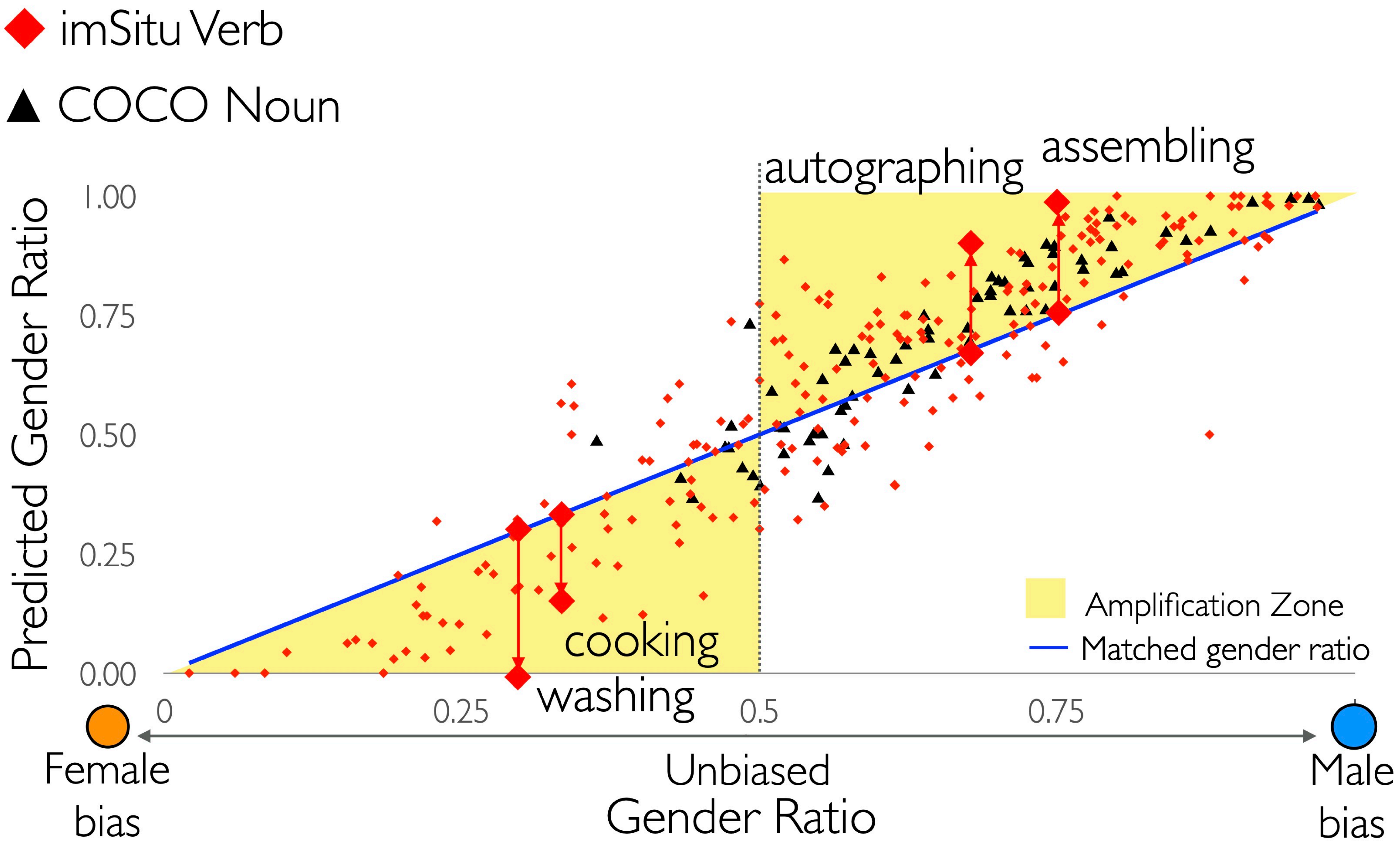
- cooking
- woman
- man

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	stir-fry

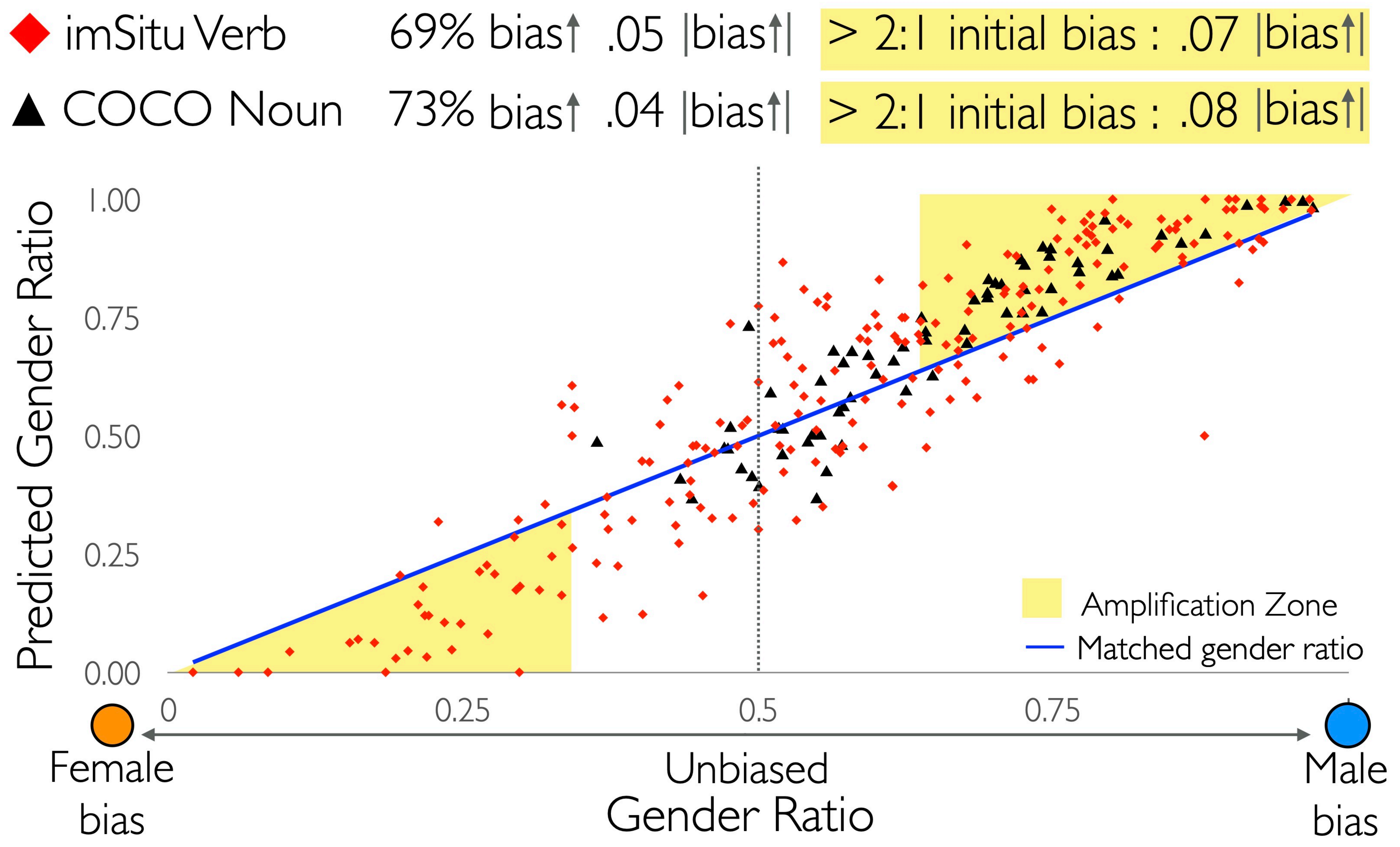
COOKING	
ROLES	NOUNS
AGENT	man
FOOD	noodle

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

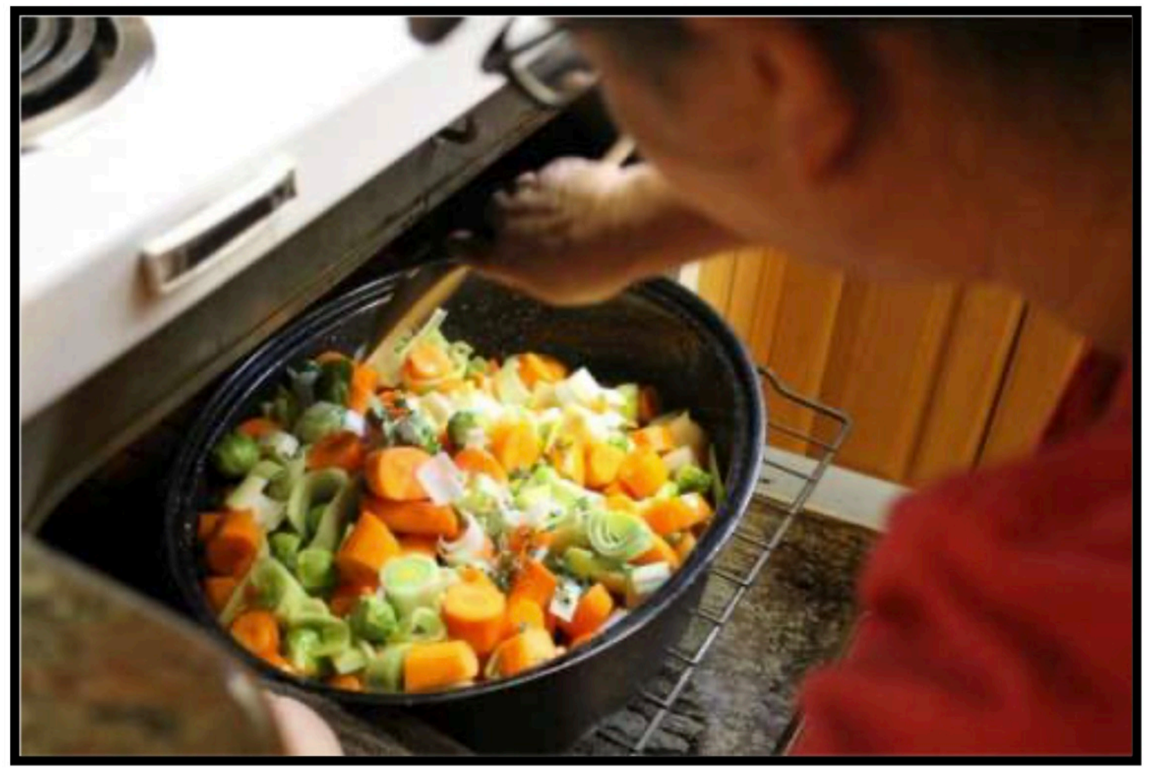
Model bias amplification



Model bias amplification



Decomposition of scoring function

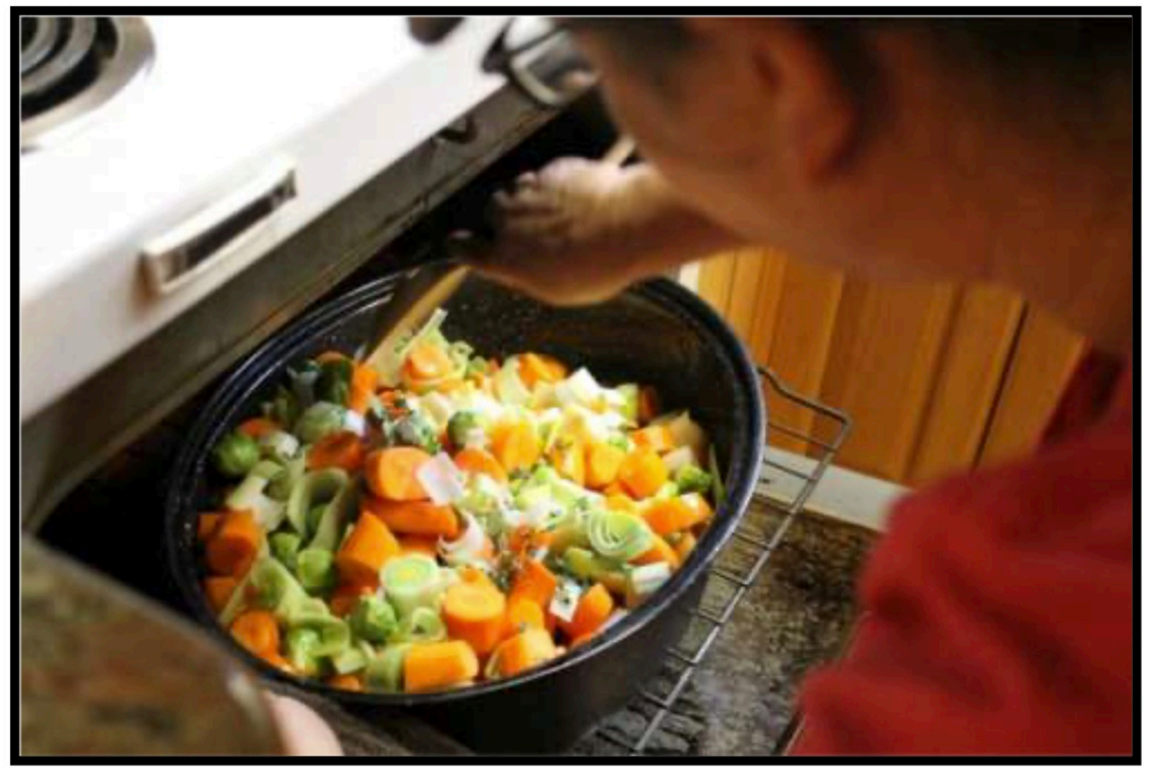


COOKING		3	
ROLES	NOUNS		
AGENT	woman	1	5
FOOD	vegetable	6	3
CONTAINER	pot	2	4
TOOL	spatula	1	3

…

COOKING		3	
ROLES	NOUNS		
AGENT	man	2	-2
FOOD	meat	1	3
CONTAINER	pot	2	4
TOOL	screwdriver	2	-5

Decomposition of scoring function



COOKING		3	
ROLES	NOUNS		
AGENT	woman	1	5
FOOD	vegetable	6	3
CONTAINER	pot	2	4
TOOL	spatula	1	3

…

COOKING		3	
ROLES	NOUNS		
AGENT	man	2	-2
FOOD	meat	1	3
CONTAINER	pot	2	4
TOOL	screwdriver	2	-5

Decomposition of scoring function

Intuition of Calibration

$\lambda_1, \lambda_2 > 0$

$$5 \rightarrow 5 - \lambda_1$$

$$-2 \rightarrow -2 + \lambda_2$$

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

3

5

1 3

6 4

2

-2 3

...

COOKING	
ROLES	NOUNS
AGENT	man
FOOD	meat
CONTAINER	pot
TOOL	screwdriver

3

-2

2 3

1 4

2

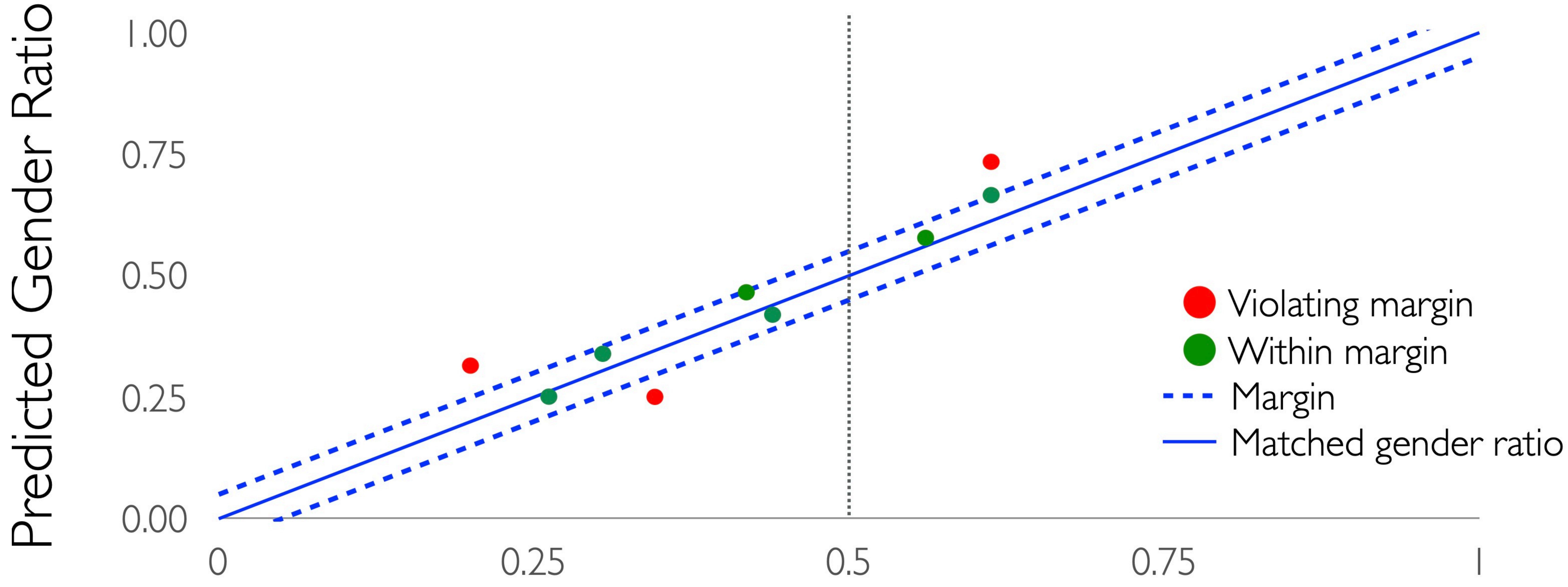
1 -5

Reducing bias amplification

Integer Linear Program

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

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



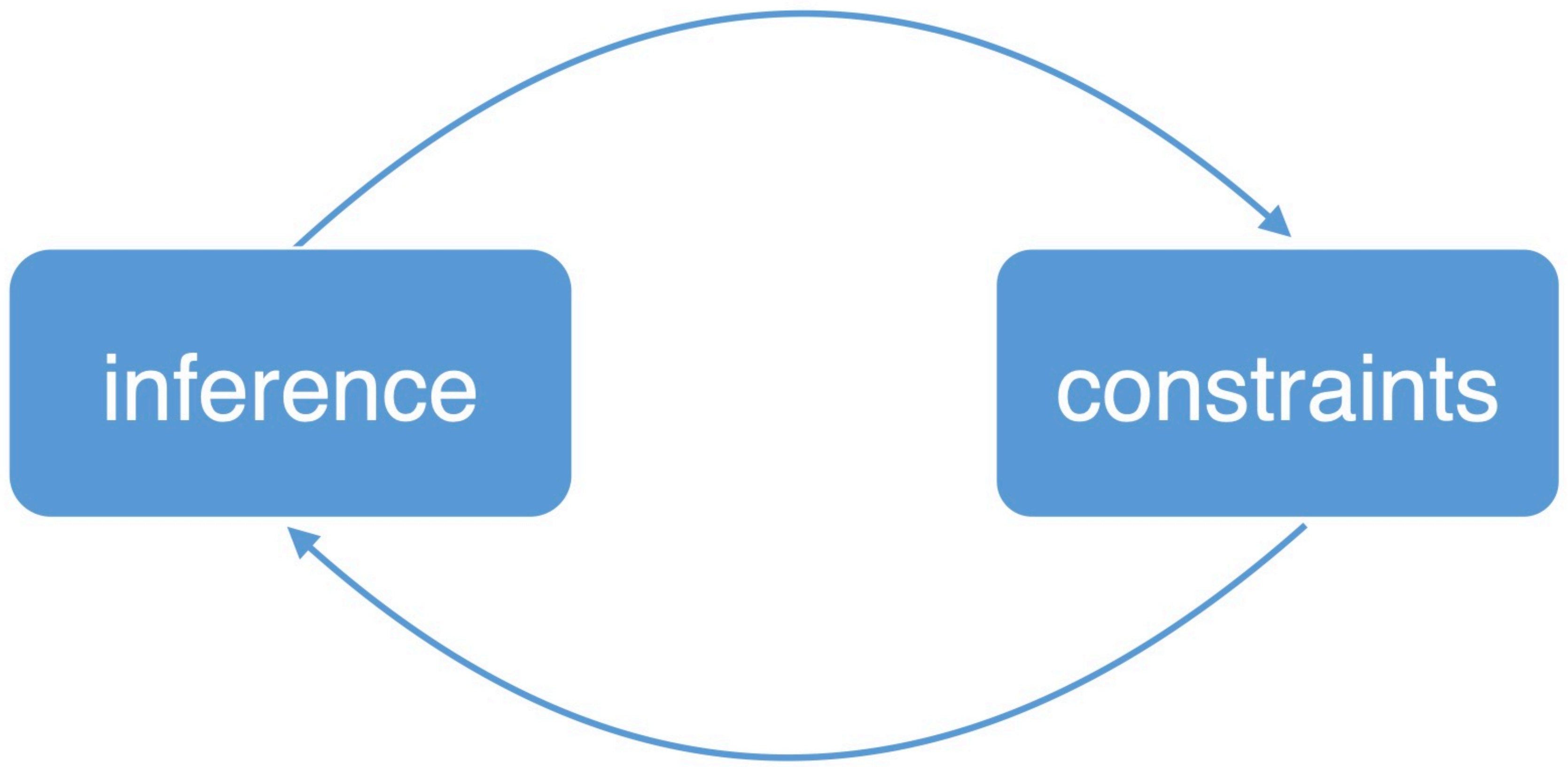
Reducing bias amplification

Integer Linear Program

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

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

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



Reducing bias amplification

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

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

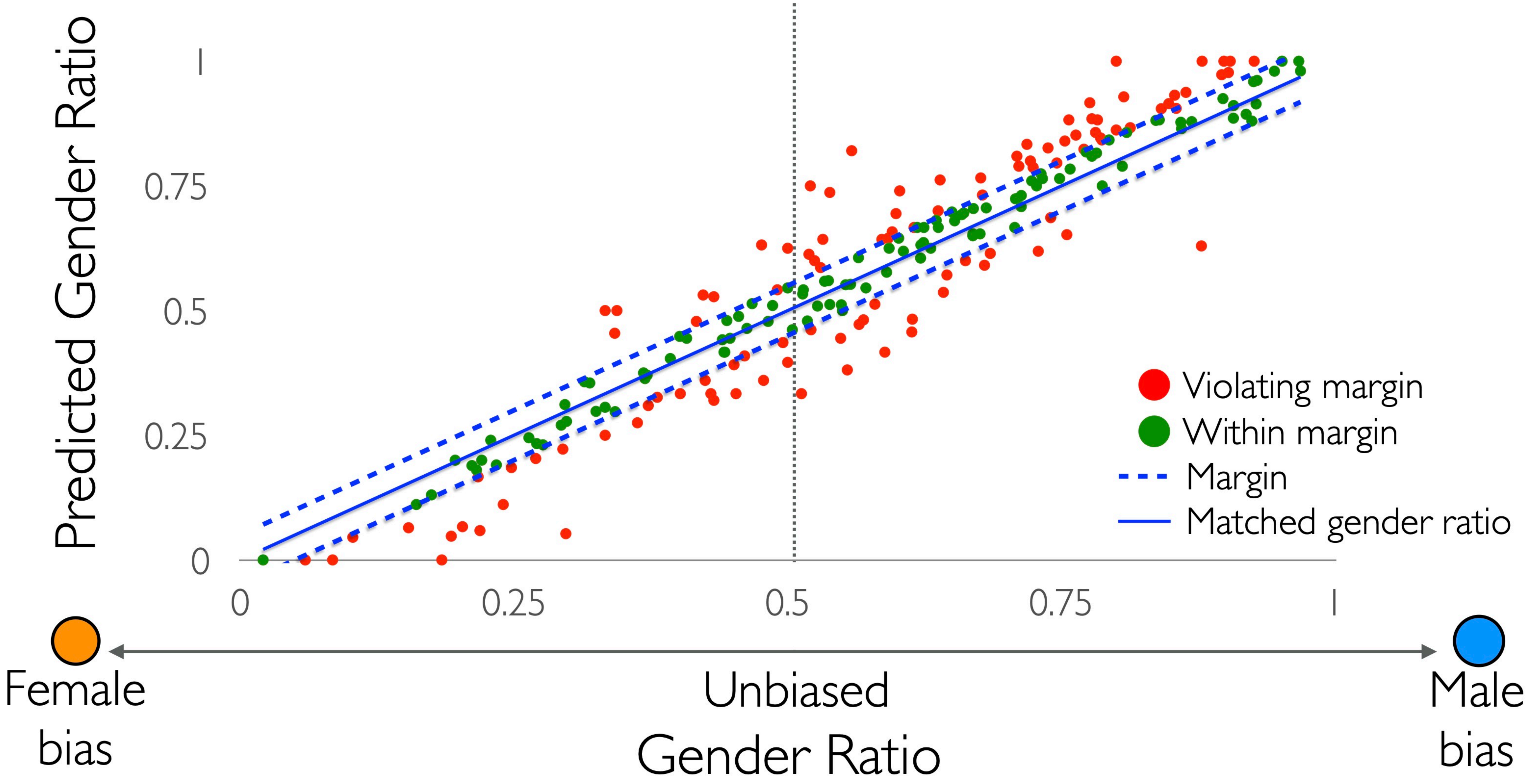
$$\boxed{\text{Lagrangian Relaxation}}$$

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

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

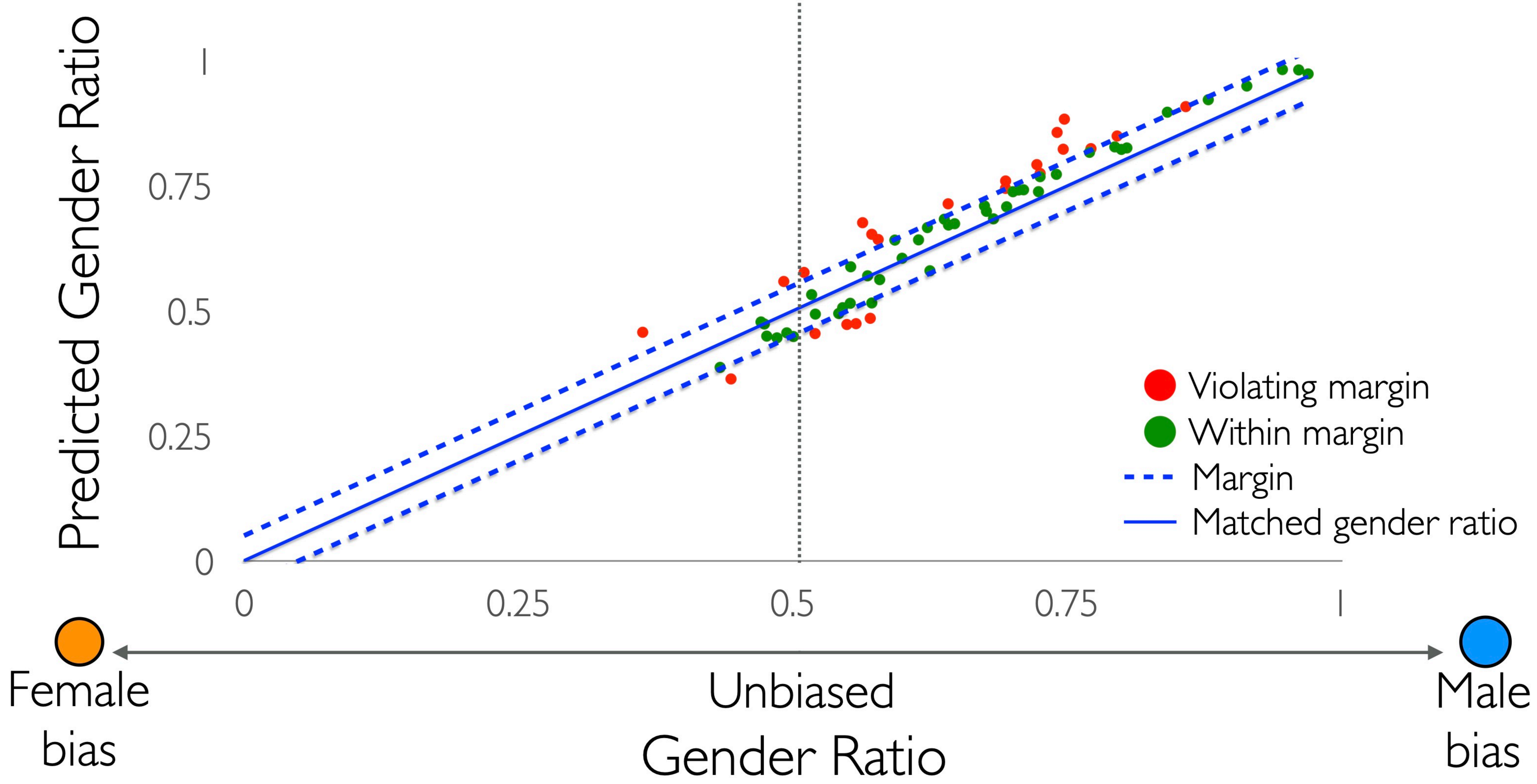
Reducing bias amplification

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



Reducing bias amplification

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



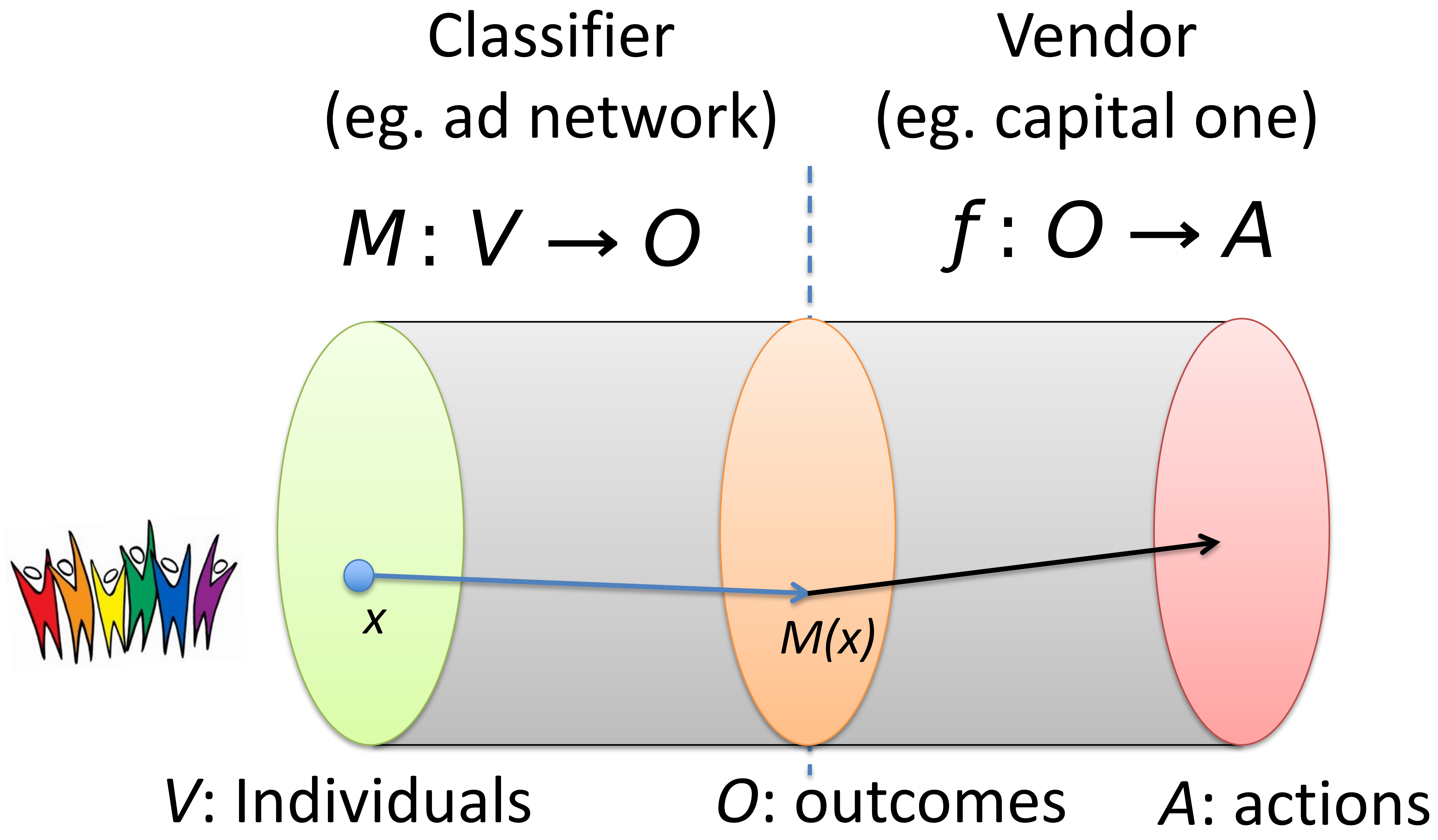
Credit Application

User visits `capitalone.com`

Capital One uses tracking information provided by the tracking network [x+1] to personalize offers

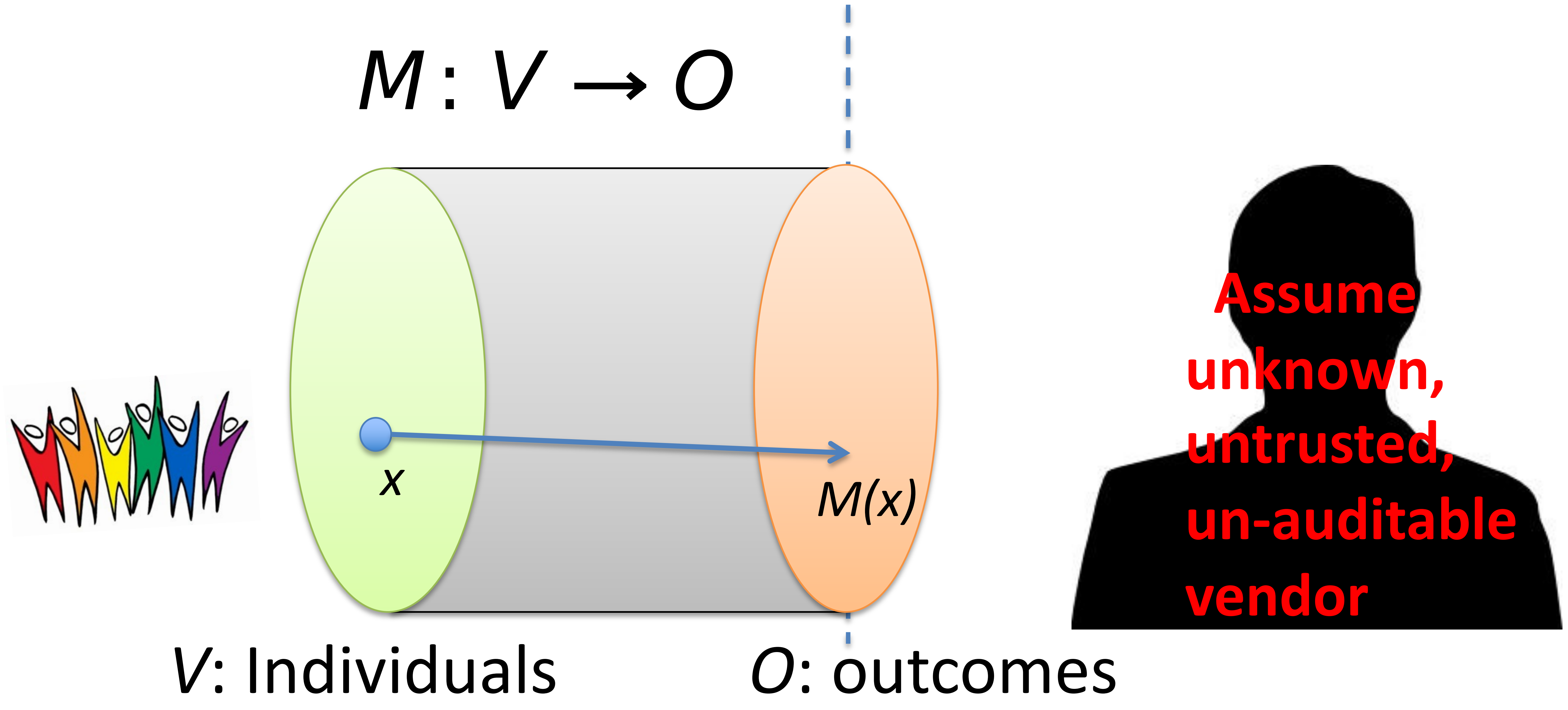
Concern: *Steering* minorities into higher rates (illegal)

WSJ 2010


Classifier
(eg. ad network)
Vendor
(eg. capital one)
$M : V \to O$
$f : O \to A$
x
$M(x)$
V: Individuals
O: outcomes
A: actions

Goal:

Achieve Fairness in the classification step



Through blindness

- Ignore all irrelevant/protected attributes
  - You don't need to see an attribute to be able to predict it with high accuracy
  - E.g.: User visits artofmanliness.com … 90% chance of being male

Individual Fairness

Treat *similar* individuals *similarly*

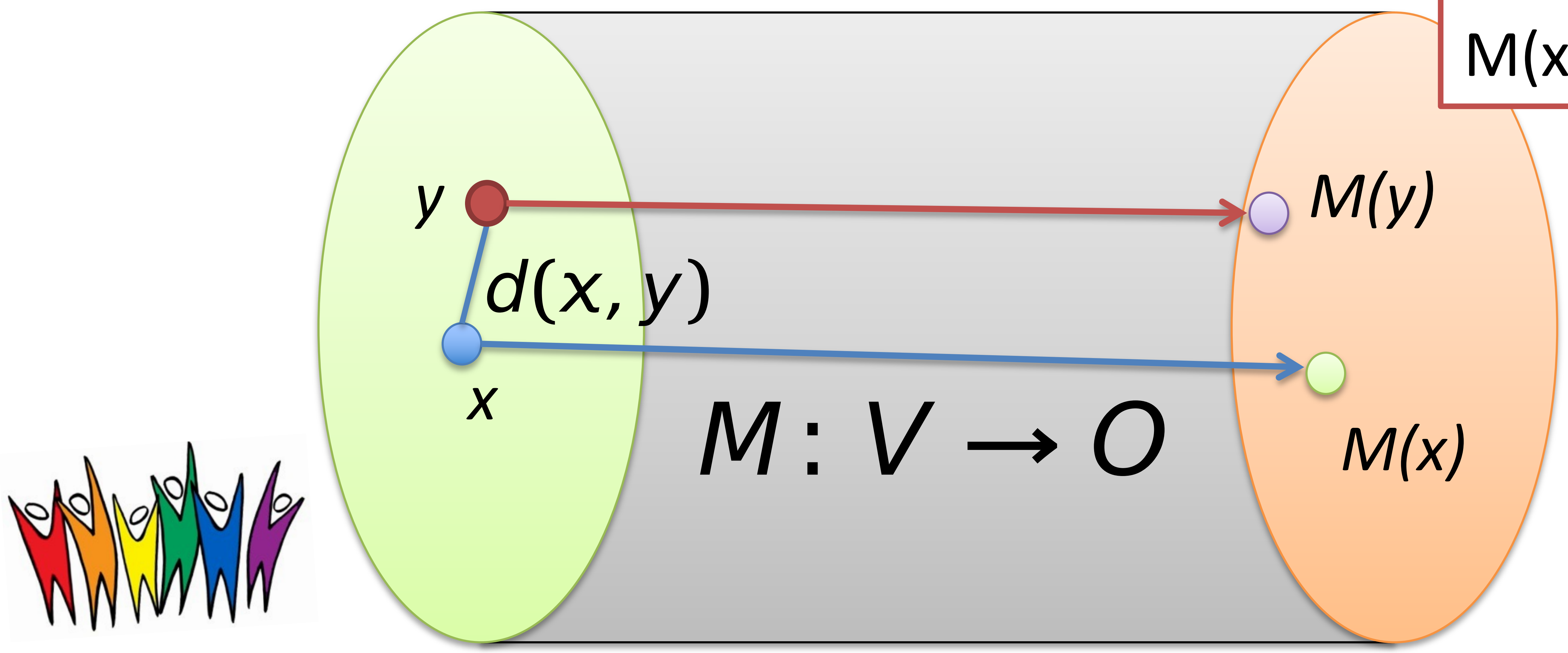
Similar for the purpose of the classification task

Similar distribution over outcomes

How to formalize this?

Think of V as space
with metric d(x,y)
similar = small d(x,y)

How can we compare M(x) with M(y)?

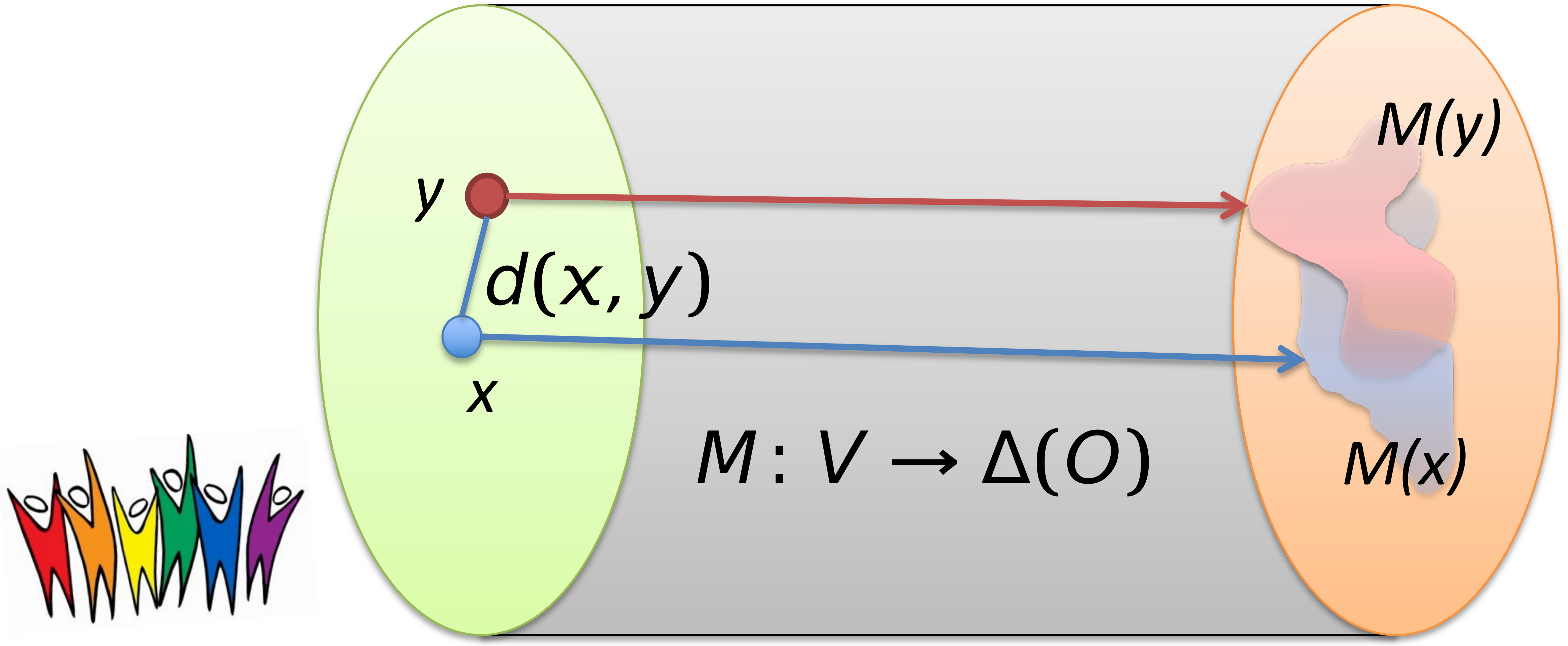


V: Individuals

O: outcomes

Distributional outcomes

How can we compare M(x) with M(y)?



Statistical distance!

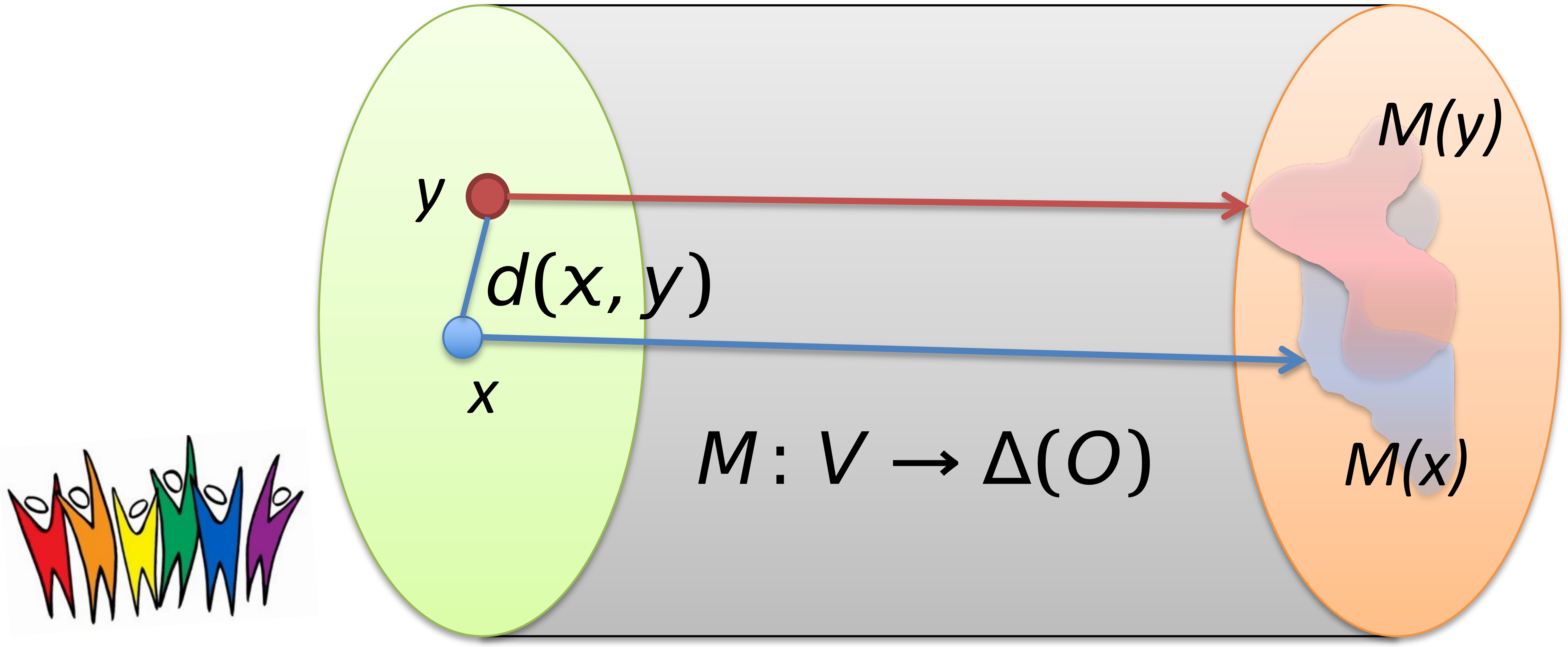
V: Individuals

O: outcomes

Metric $d: V \times V \rightarrow \mathbb{R}$

Lipschitz condition $\|M(x) - M(y)\| \leq d(x, y)$

This talk: Statistical distance in [0,1]



V: Individuals O: outcomes

Statistical Distance

P, Q denote probability measures on a finite domain A. The *statistical distance* between P and Q is denoted by

$$D_{\text{tv}}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| .$$

Example: Mid D
A= {0,1}
P(0) = P(1) = ½
Q(0) = ¾, Q(1) = ¼
D(P, Q) = ¼

Utility Maximization

Vendor can specify **arbitrary utility function**

$$U : V \times O \rightarrow \mathbb{R}$$

$U(v,o)$ = Vendor's utility of giving individual v
the outcome o

Maximize vendor’s expected utility subject to Lipschitz condition

$$\max_{M(x)} \mathop{\mathbb{E}}_{x\sim V} \mathop{\mathbb{E}}_{o\sim M(x)} U(x,o)$$

s.t. M is d-Lipschitz

$$\|M(x)-M(y)\| \leq d(x,y)$$