EMNLP 2019 Tutorial on Bias and Fairness in Natural Language Processing, Hong Kong

Building Fair and Robust Representations for Vision and Language

Vicente Ordóñez-Román

Assistant Professor
Department of Computer Science

University of Virginia

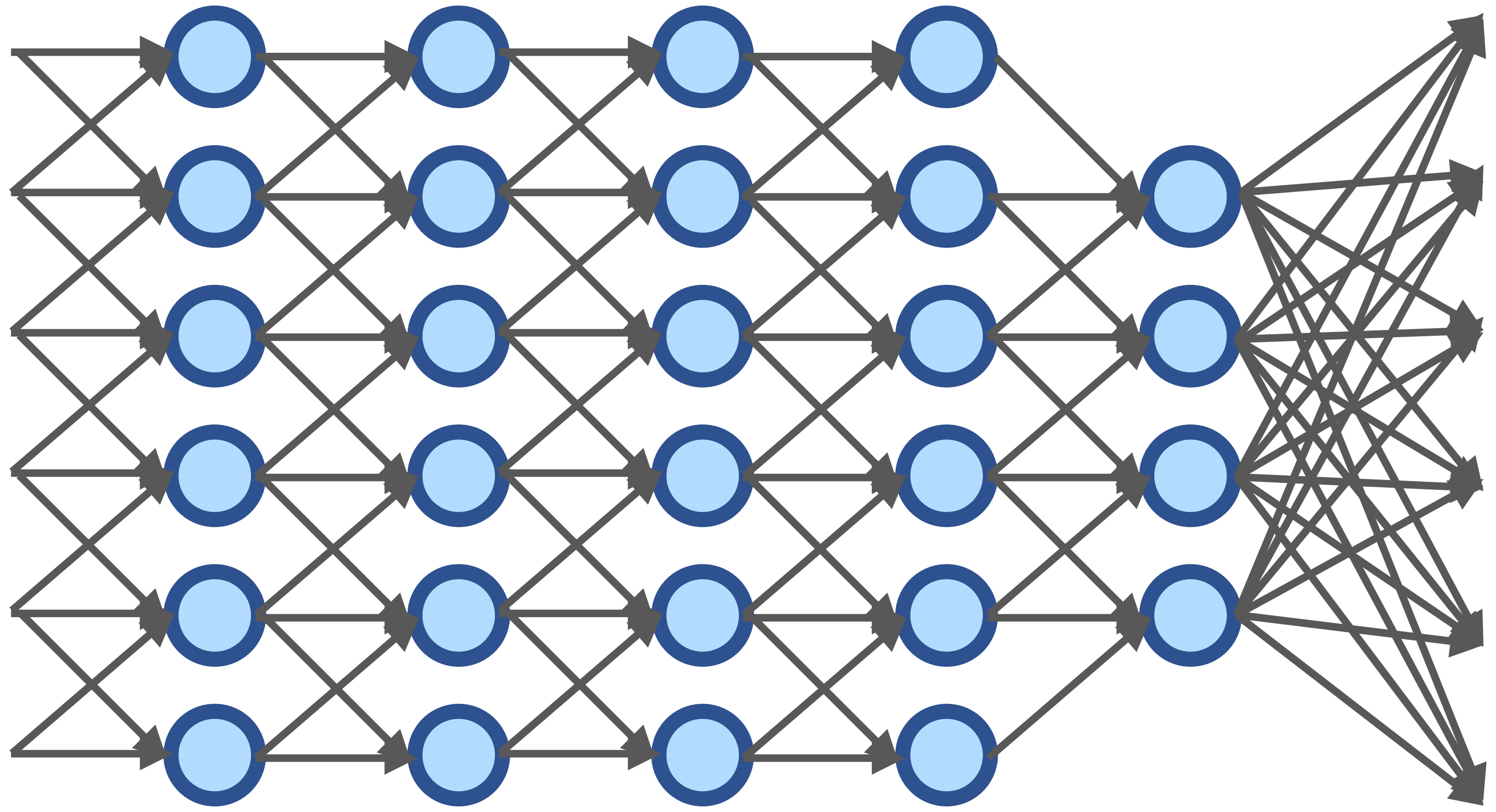
Outline

- Issues identified in biased representations
- Metrics and findings
- Solutions that have been proposed

Annotated Data + Machine Learning / Deep Learning

$$f(x)$$

→

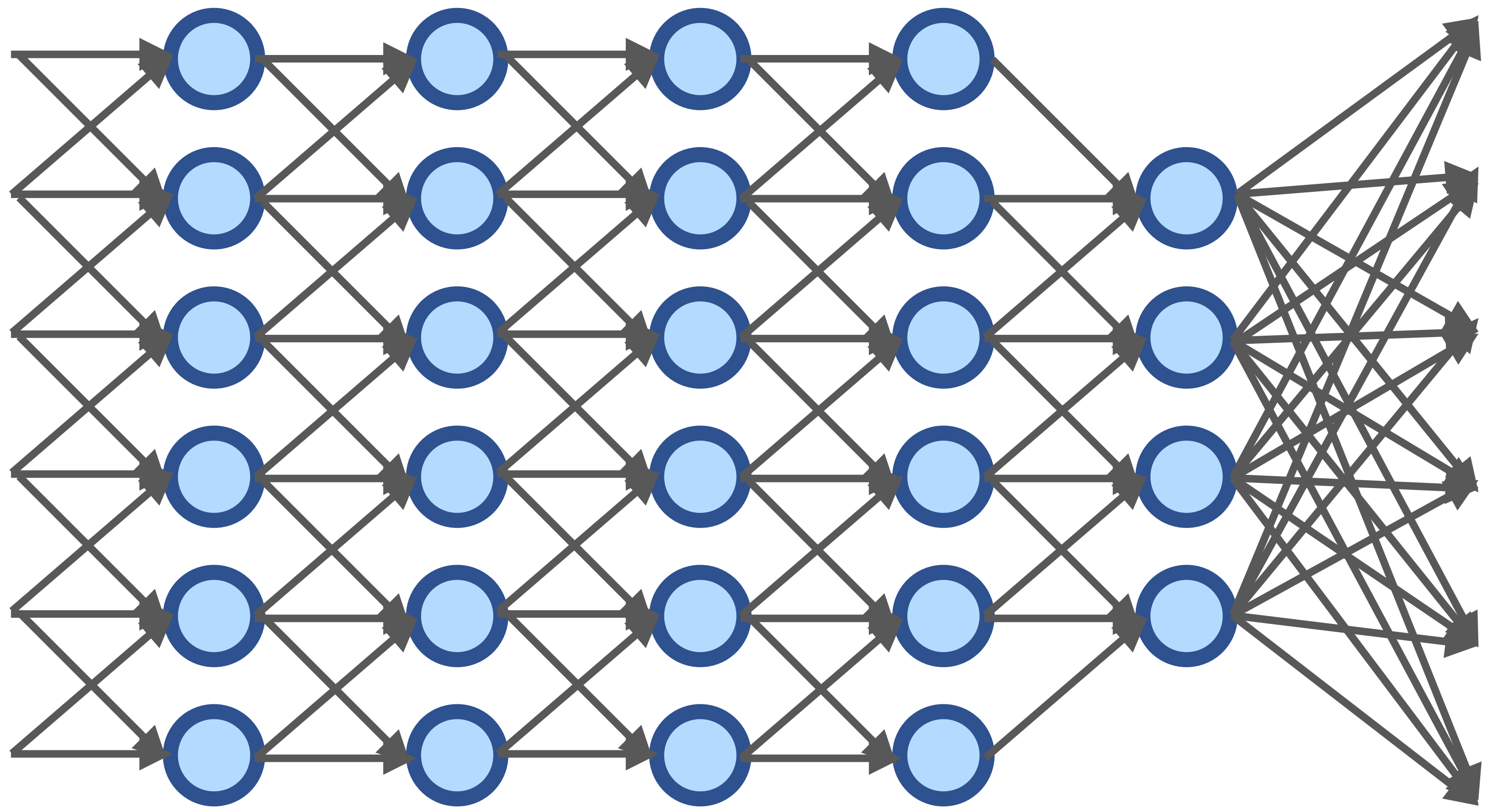


→

Words,
Text,
Linguistic
Structure

Case Study 1: Most Basic form of Grounding: Image to Words

$$f(x)$$

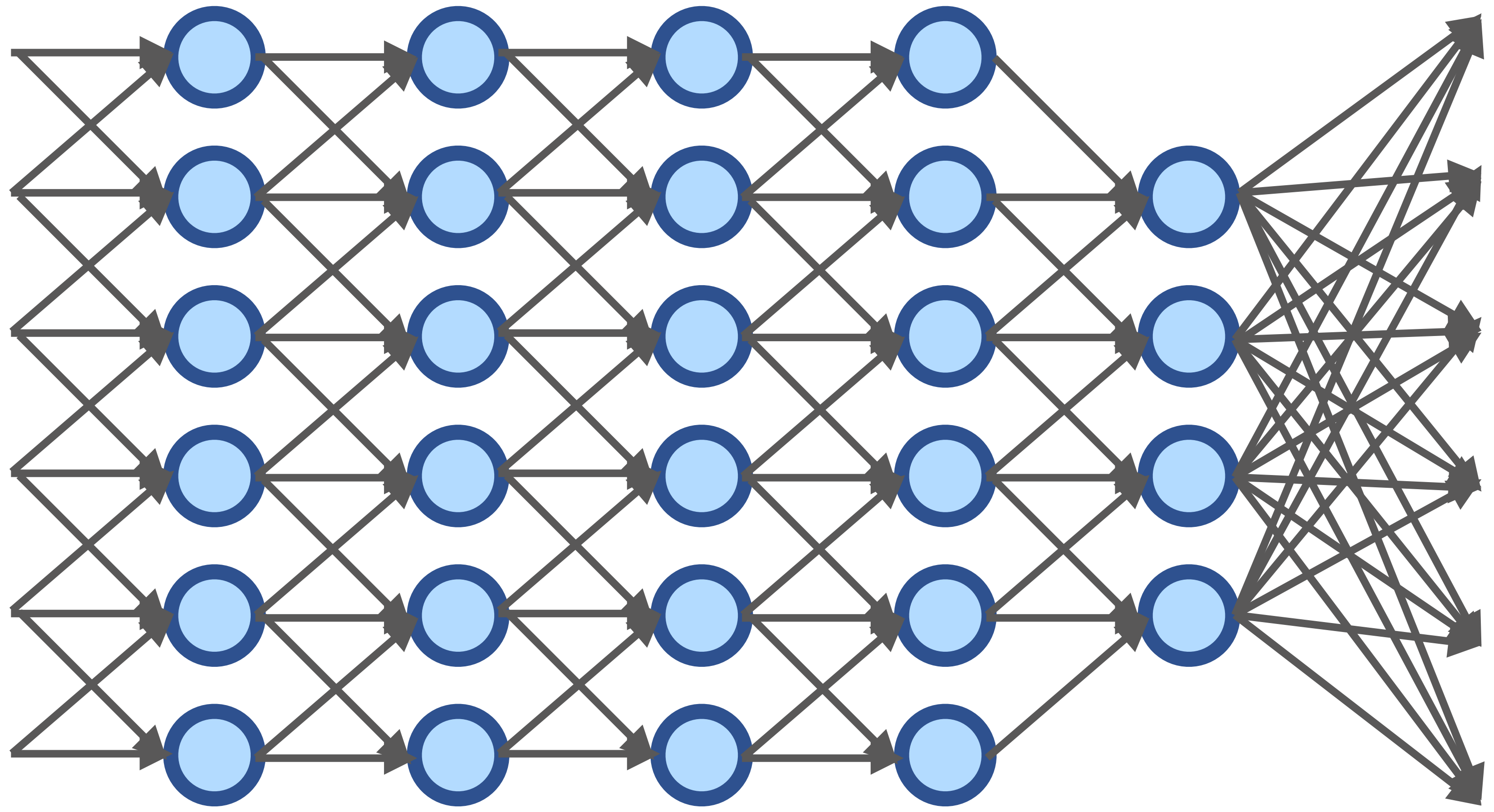


kitchen
no-kitchen

Protected variable: Gender

Case Study 1: Most Basic form of Grounding: Image to Words

$f(x)$



kitchen
no-kitchen

Protected variable: Gender

For any pair of gender types:

P(kitchen = 1 / gender = m) = P(kitchen = 1 / gender = f)
P(kitchen = 0 / gender = m) = P(kitchen = 0 / gender = f)

Approach 1: Feature Invariant Learning

Learning Fair Representations

Richard Zemel ZEMEL@CS.TORONTO.EDU
Yu (Ledell) Wu WUYU@CS.TORONTO.EDU
Kevin Swersky KSWERSKY@CS.TORONTO.EDU
Toniann Pitassi TONI@CS.TORONTO.EDU
University of Toronto, 10 King's College Rd., Toronto, ON M6H 2T1 CANADA

Cynthia Dwork DWORK@MICROSOFT.COM
Microsoft Research, 1065 La Avenida Mountain View, CA. 94043 USA

ICML 2013

Approach 1: Feature Invariant Learning

X: Images

Y: Labels

kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

Approach 1: Feature Invariant Learning

X: Images

$$y = f(x)$$

Y: Labels

kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

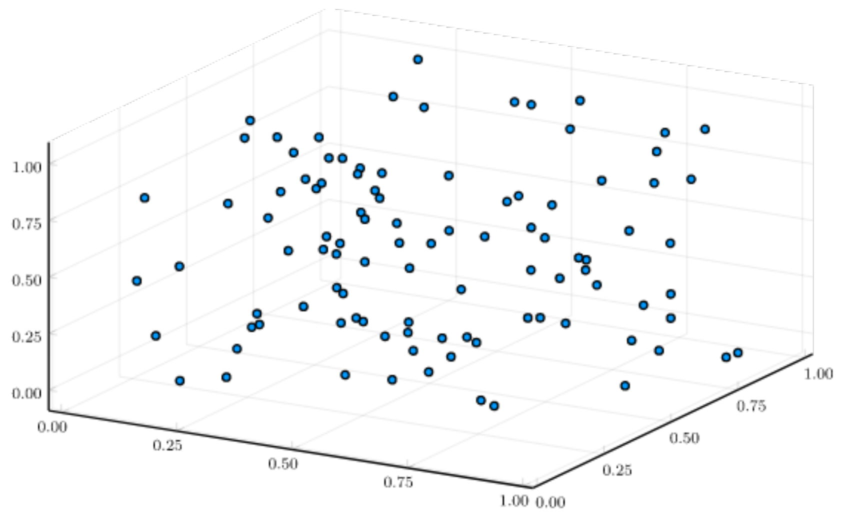
Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

Approach 1: Feature Invariant Learning

Instead

X: Images

Z: Representations



Y: Labels

kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

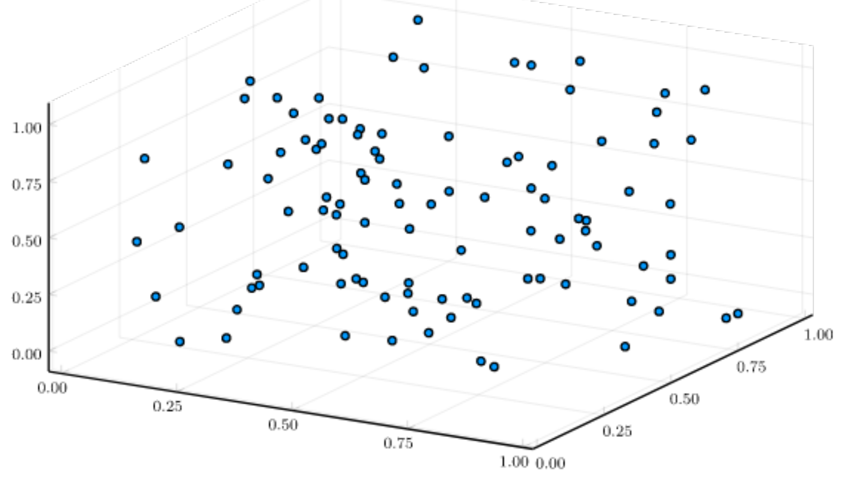
Approach 1: Feature Invariant Learning

Instead

X: Images

x

Z: Representations



$$y = f(z)$$

$$\hat{x} = \sum_i z_i v_i$$

Y: Labels

kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

y

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

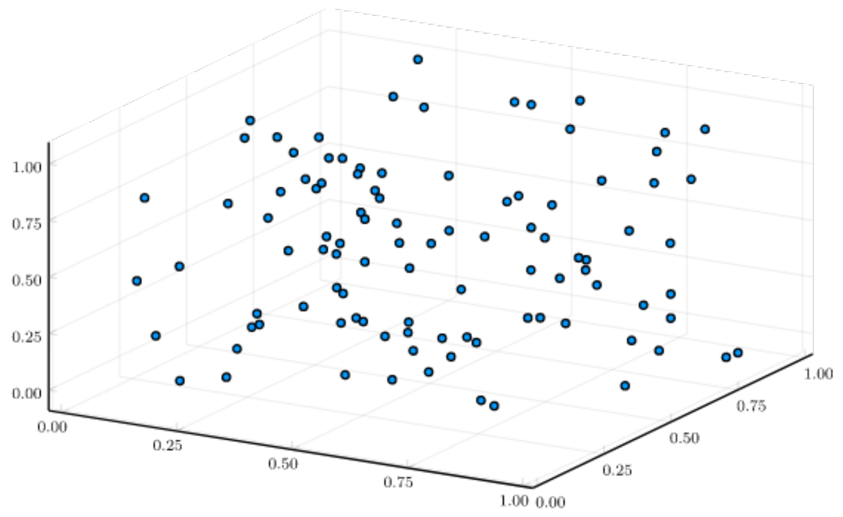
Approach 1: Feature Invariant Learning

Instead

X+: Images

X-: Images

Z: Representations



$$y = f(z)$$

$$\hat{x} = \sum_i z_i v_i$$

Y: Labels

kitchen

kitchen

kitchen

kitchen

no-kitchen

no-kitchen

no-kitchen

kitchen

kitchen

no-kitchen

no-kitchen

y

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

Approach 1: Feature Invariant Learning

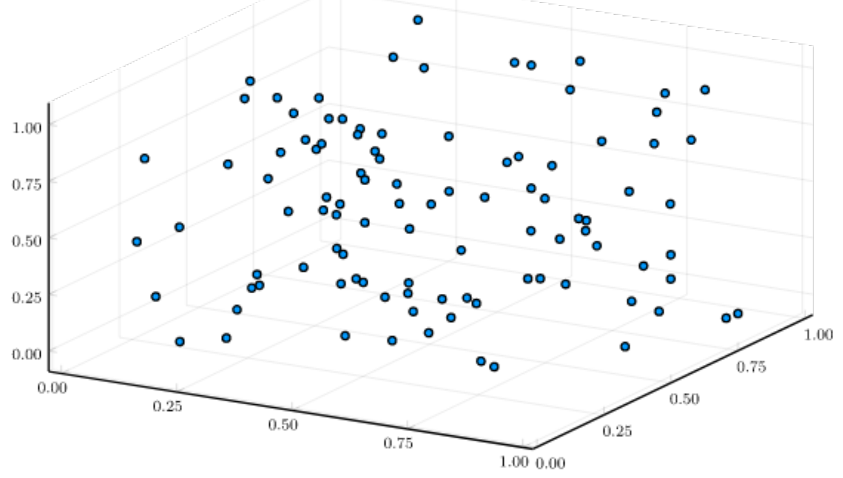
Instead

X+: Images

X-: Images

$$P(z_i \mid x+) = P(z_i \mid x-)$$

Z: Representations



$$y = f(z)$$

$$\hat{x} = \sum_i z_i v_i$$

Y: Labels

kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

y

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

Approach 1: Feature Invariant Learning

$$L = \boxed{\sum_k CrossEntropy(y^{(k)}, \hat{y}^{(k)})} + \boxed{\alpha \sum_k \left| x^{(k)} - \hat{x}^{(k)} \right|} + \beta \boxed{\left| \frac{1}{|X+|} \sum_{X+} z_i^{(k)} - \frac{1}{|X-|} \sum_{X-} z_i^{(k)} \right|}$$

Classifications
should be good

Reconstructions
should be good

Intermediate Representations
should be indistinguishable
across values of the protected variable

Learning Fair Representations
Zemel, Wu, Swersky, Pitassi, and Dwork. **ICML 2013**

Approach II: Adversarial Feature Learning

X: Images

$$y = f(x)$$

Y: Labels

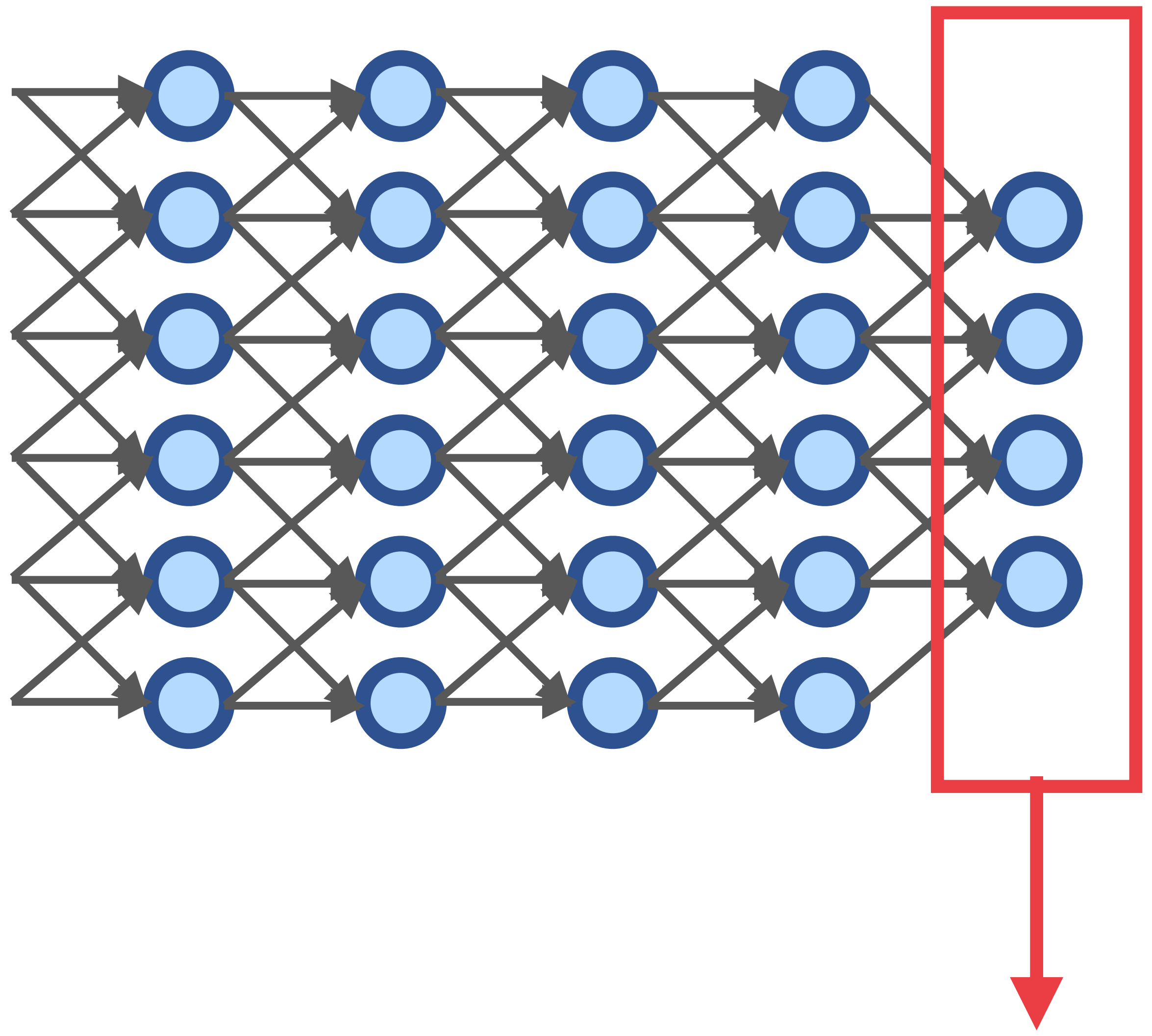
kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

X: Images

$$y = f(x)$$



z

Y: Labels

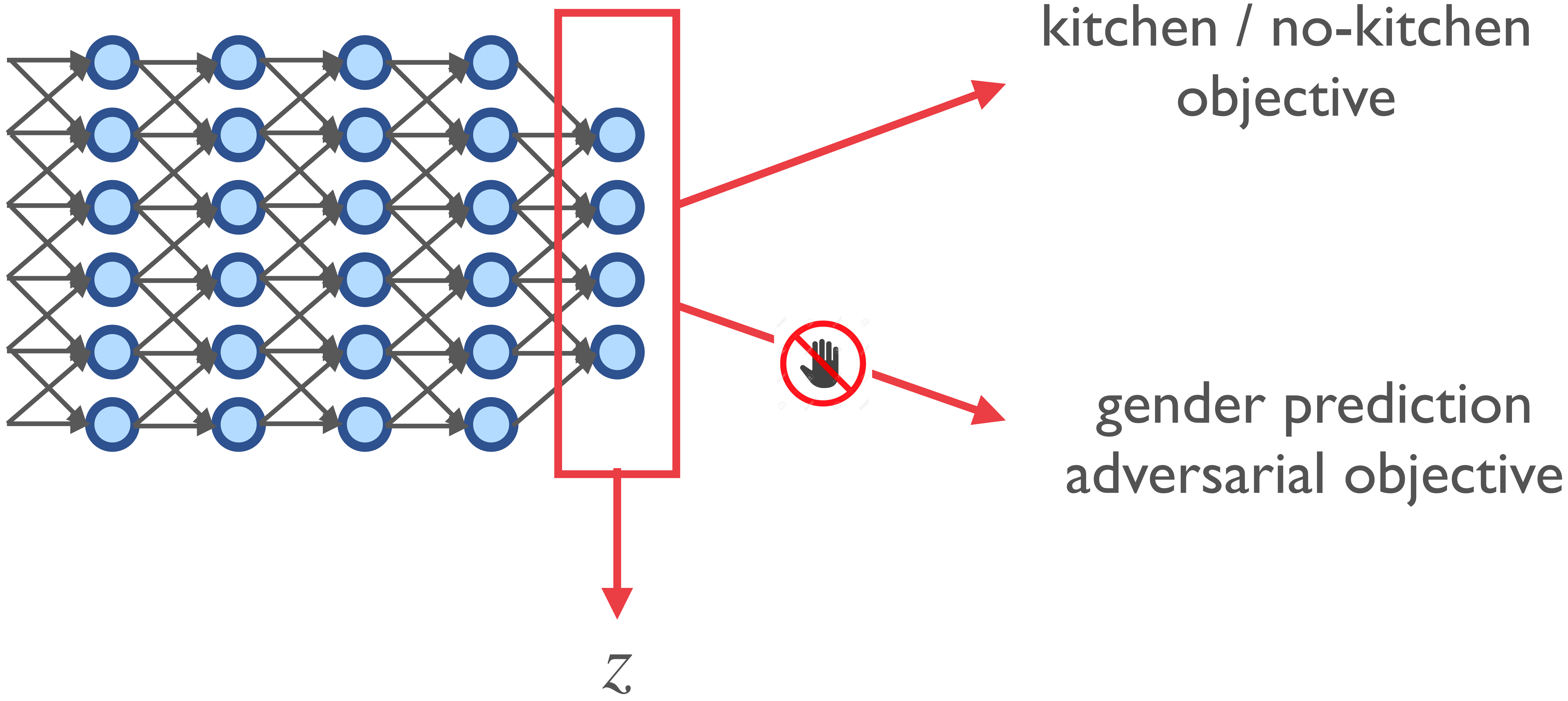
kitchen
kitchen
kitchen
kitchen
no-kitchen
no-kitchen
no-kitchen
kitchen
kitchen
no-kitchen
no-kitchen

Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

X: Images

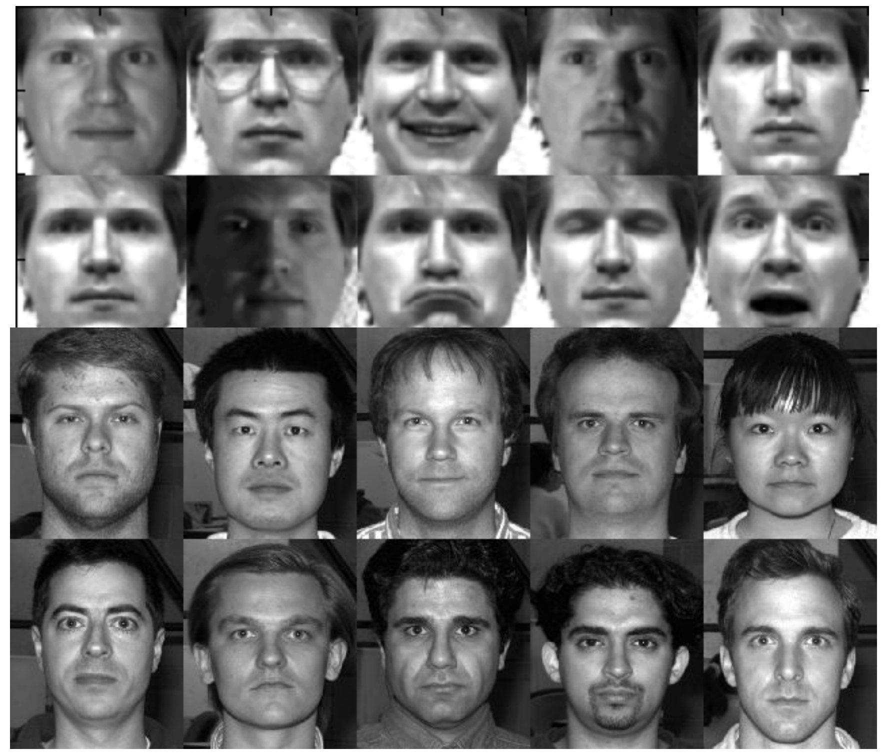
$$y = f(x)$$

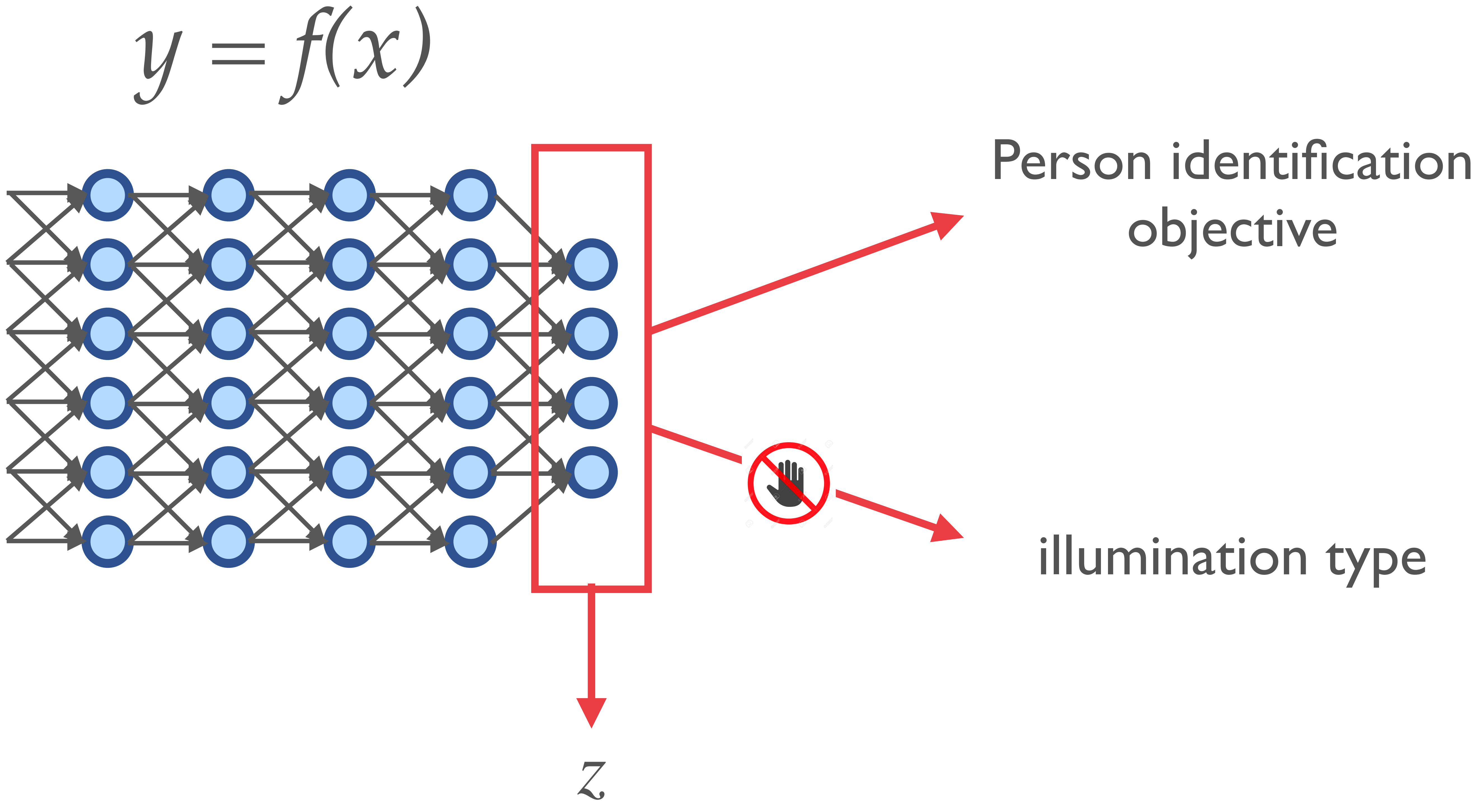


Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

X: Images

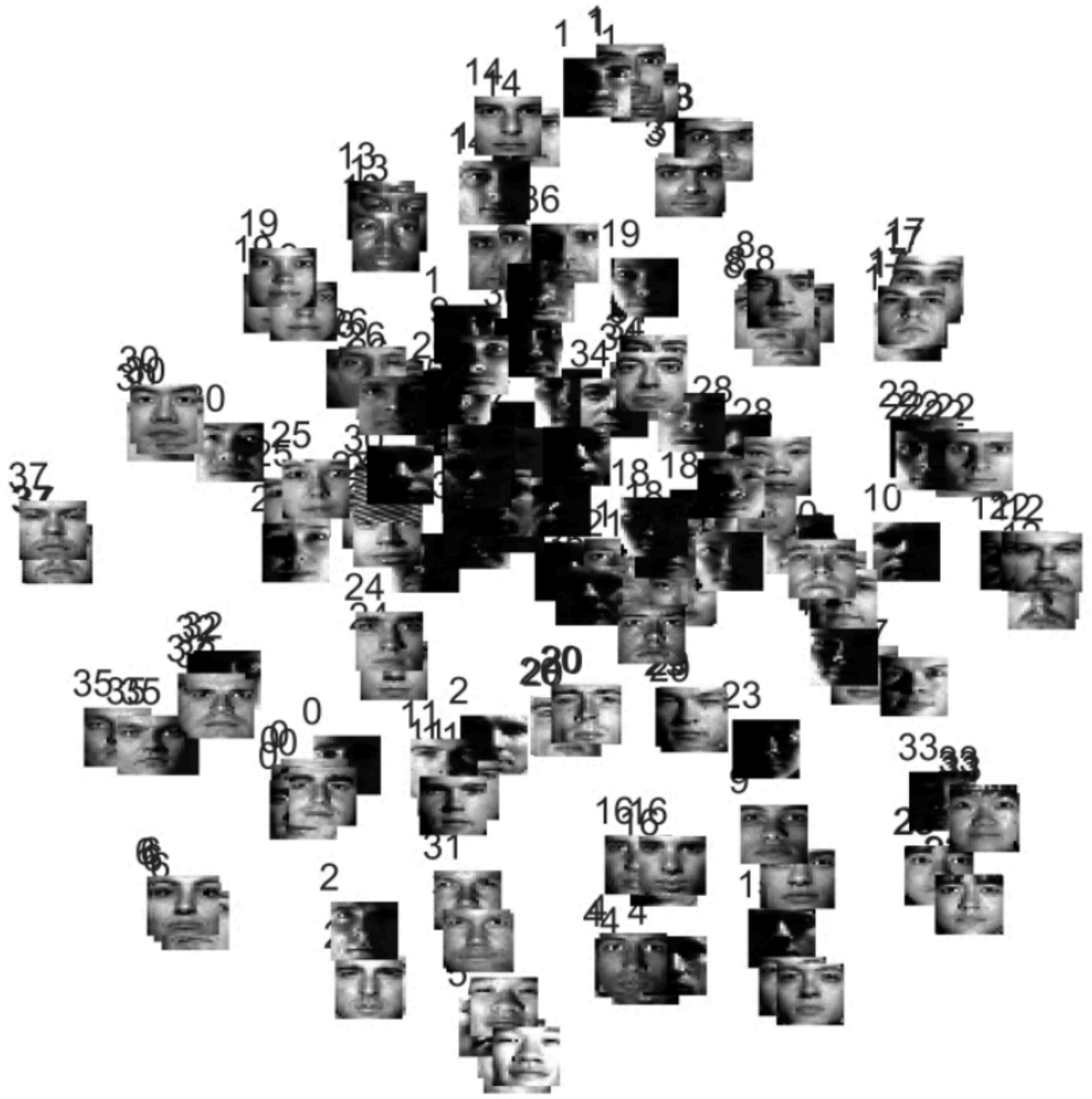




Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

(a) Using the original image x as the representation



(b) Representation learned by our model

Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Case Study: Visual Semantic Role Labeling (vSRL)



CARRYING					
ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	WOMAN	AGENT	MAN
ITEM	BABY	ITEM	BUCKET	ITEM	TABLE
AGENTPART	CHEST	AGENTPART	HEAD	AGENTPART	BACK
PLACE	OUTSIDE	PLACE	PATH	PLACE	STREET

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi **CVPR 2017**

Compositionality: How to learn what looks like carrying?

Lots of Images of People Carrying Backpacks



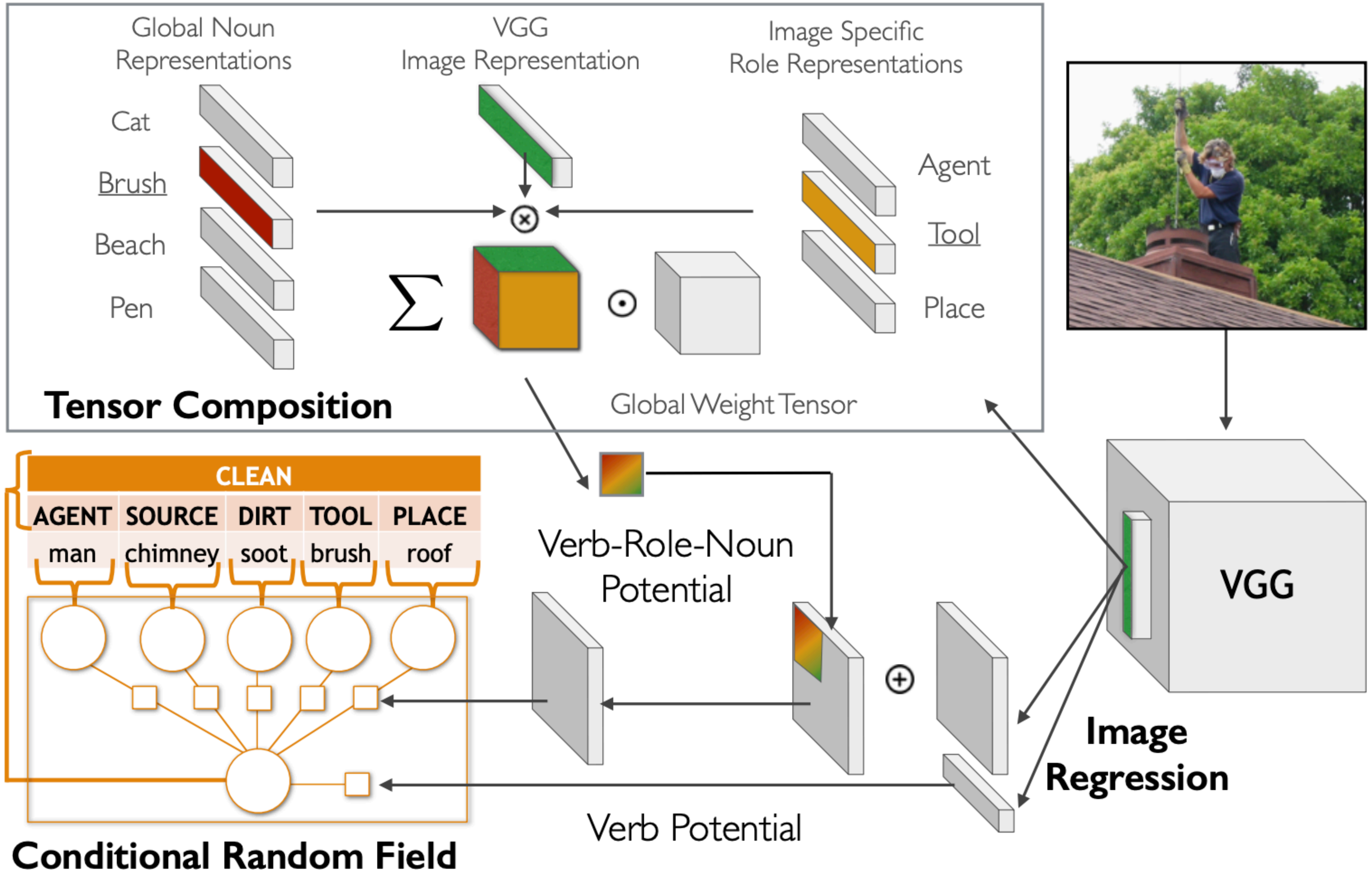


Not Many Images of People Carrying Tables

But Lots of Images of Tables in Other Images




Deep Neural Network + Compositional Conditional Random Field (CRF)



Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi **CVPR 2017**

Situation Recognition: CVPR 2017
Compositional Shared Learning of Underlying Concepts

http://imsitu.org/demo/

Recognize Situations

Paste a url

Provide an image URL to recognize | Classify URL

Query

Predicted situations

falling				0.58372
agent	source	goal	place	
person	horse	land	outdoors	

whipping				0.10375
agent	item	tool	place	
jockey	horse	whip	outdoors	

rearing		0.07997
agent	place	
horse	grass	

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi **CVPR 2017**

However we kept running into this…

http://imsitu.org/demo/

Recognize Situations

Paste a url

Provide an image URL to recognize | Classify URL

Predicted situations

rinsing			
agent	object	tool	place
woman	hair	sink	toilet

installing				
agent	component	destination	tool	place
man	faucet	sink	hand	inside

filling				
agent	destination	item	source	place
woman	pitcher	water	faucet	kitchen

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi **CVPR 2017**

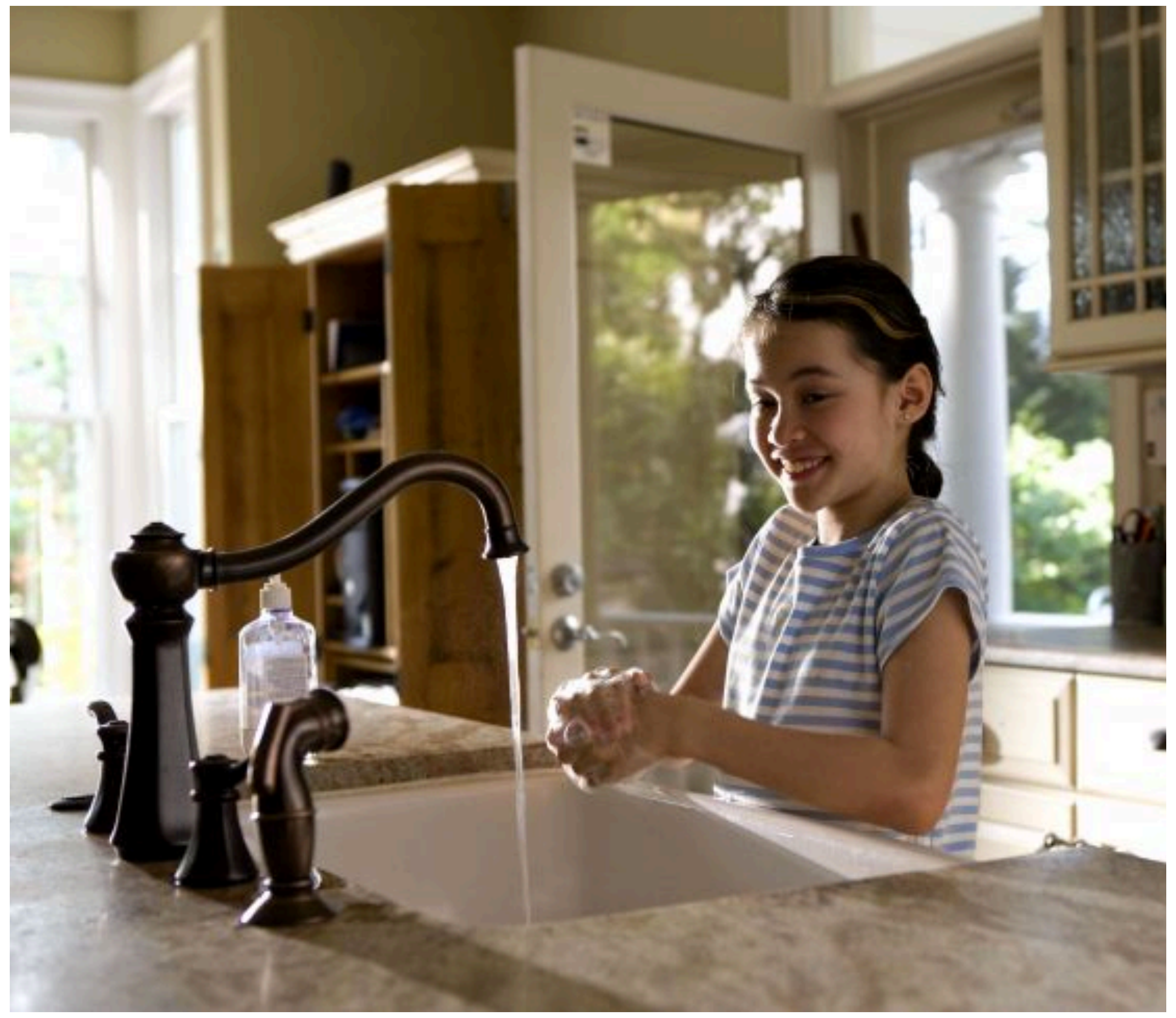
However we kept running into this…

http://imsitu.org/demo/

Recognize Situations

Paste a url

Provide an image URL to recognize | Classify URL



Predicted situations

dusting			
agent	source	tool	place
woman	faucet	towel	room

vacuuming			
agent	surface	tool	place
woman	floor	vacuum	room

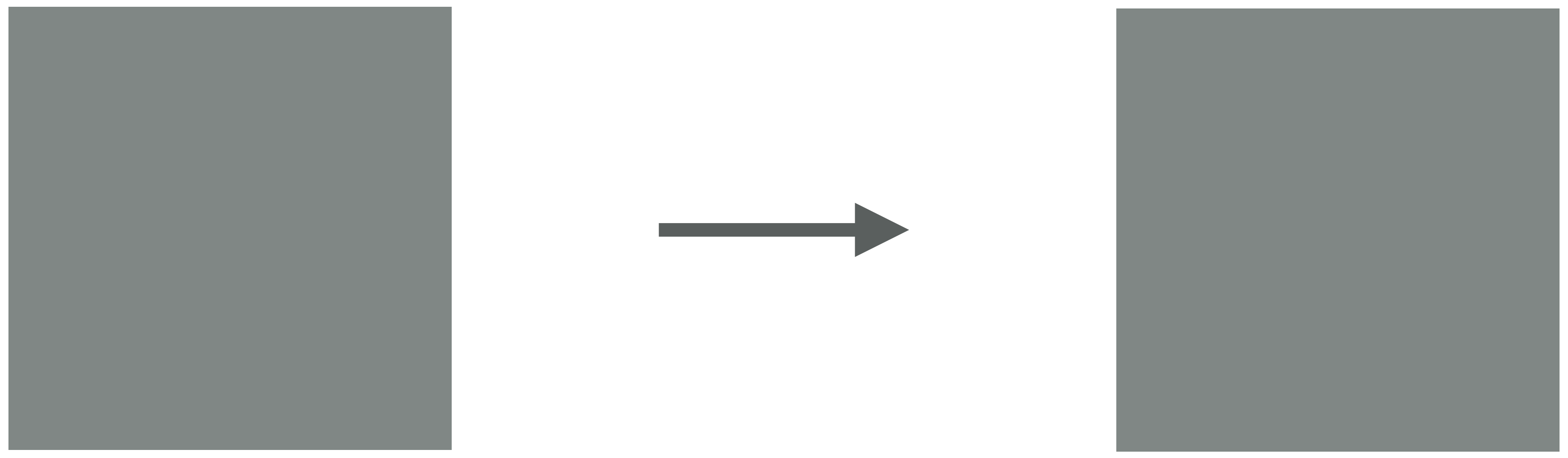
cleaning			
agent	source	tool	place
woman	∅	fabric	house

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi **CVPR 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?



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**

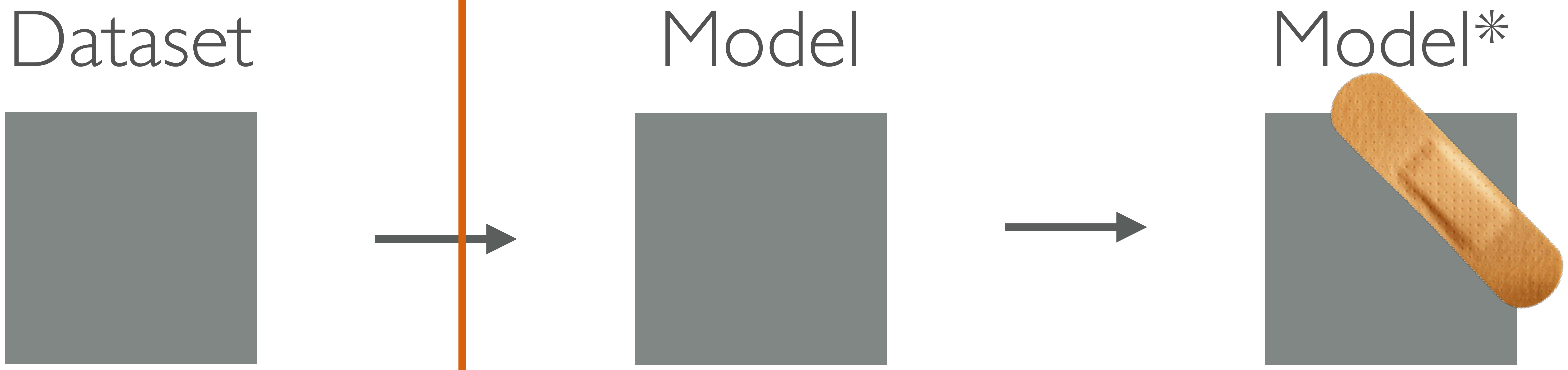
Dataset? → Model?

Men Cooking: 33% Women Cooking: 66%

Men Cooking: 16% Women Cooking: 84%

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 Cooking: 33%
Women Cooking: 66%

Men Cooking: 16%
Women Cooking: 84%

*Our solution: RBA: Optimize for accuracy but also to match data distribution.

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 | Model*

Dataset	Model	Model*
Men Cooking: 33%	Men Cooking: 16%	Men Cooking: 20%
Women Cooking: 66%	Women Cooking: 84%	Women Cooking: 80%

*Our solution: RBA: Optimize for accuracy but also to match data distribution.

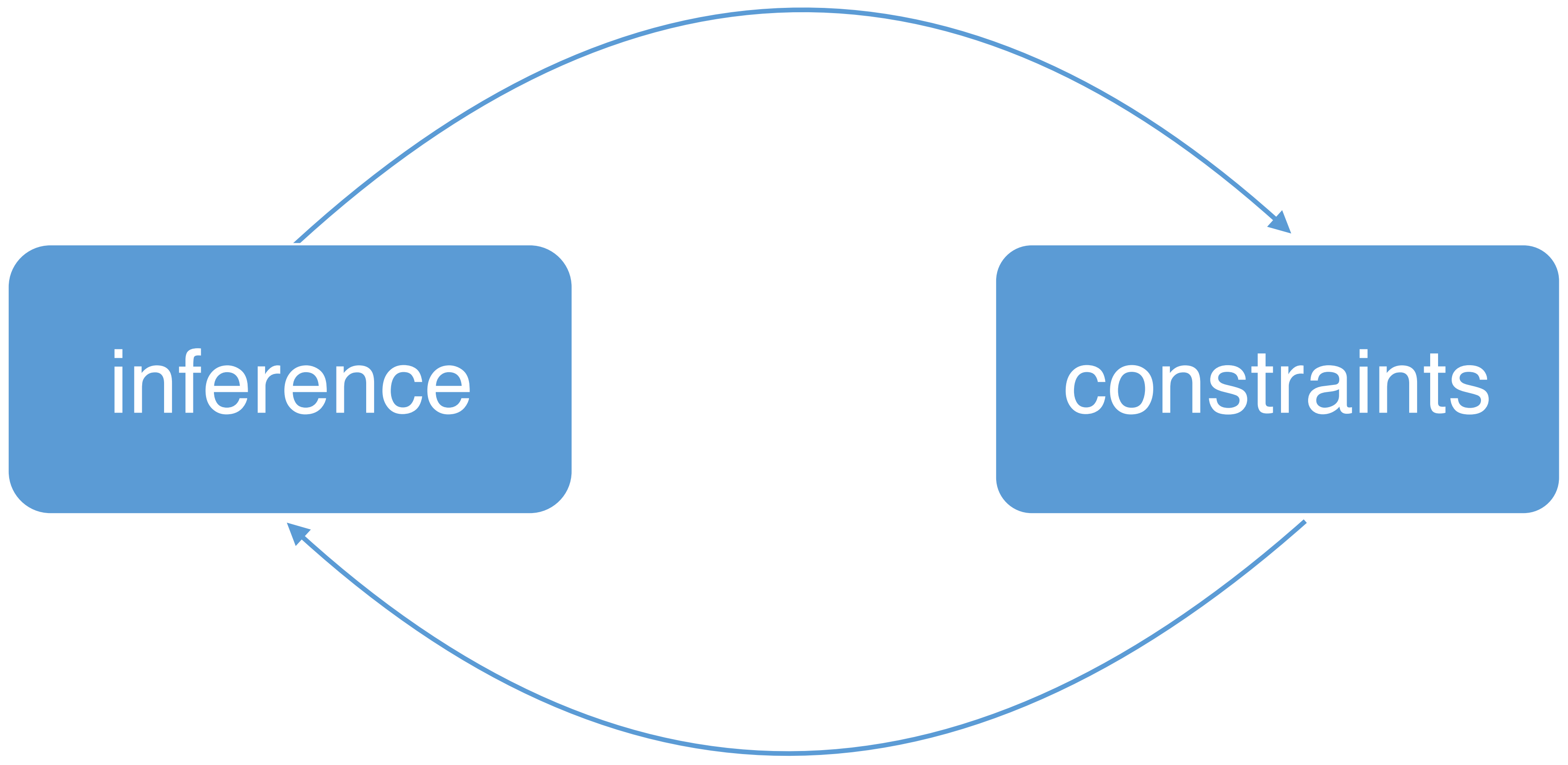
Reducing Bias Amplification (RBA)

Integer Linear Program

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

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

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



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

Our most recent work on this topic:

Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez. **ICCV 2019**

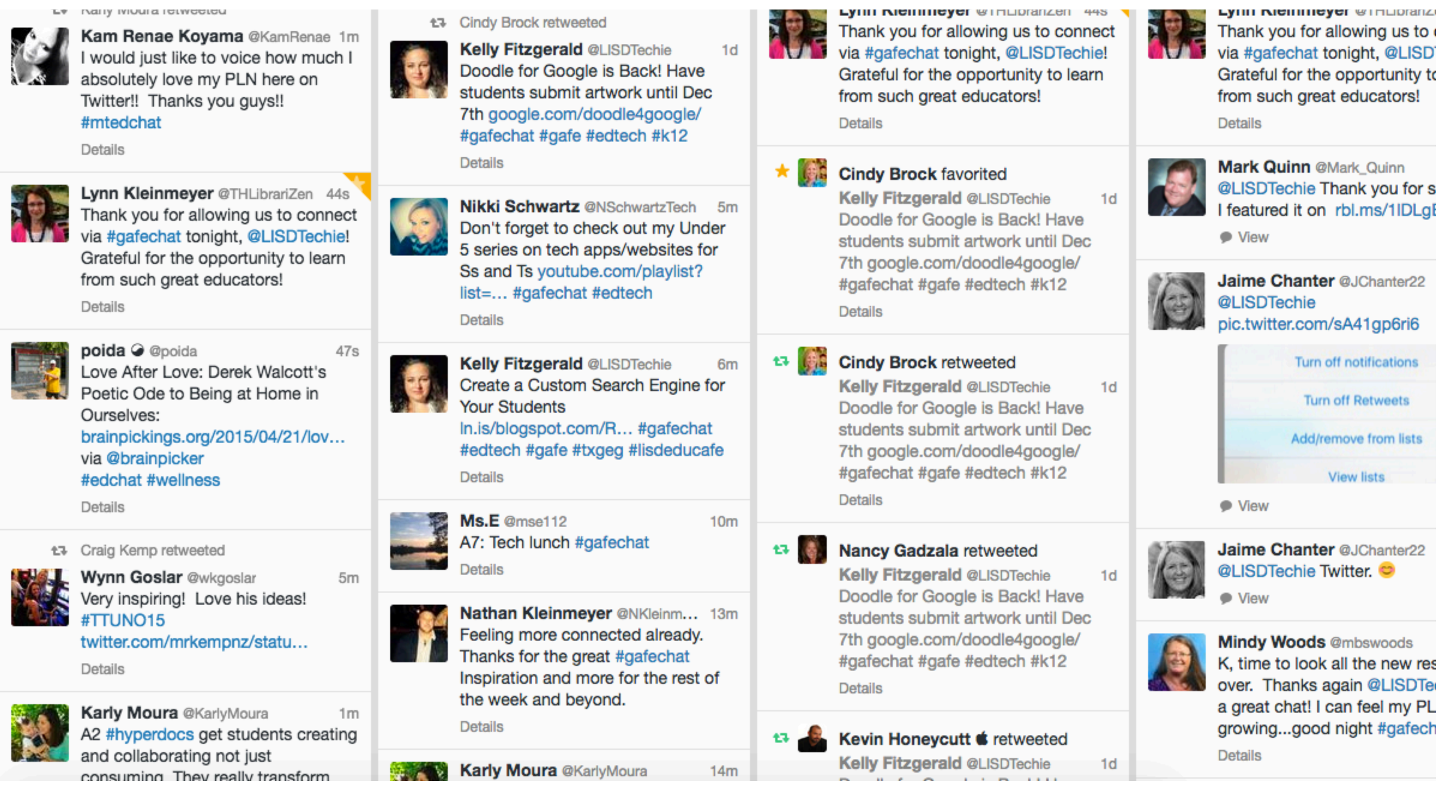
Key Findings

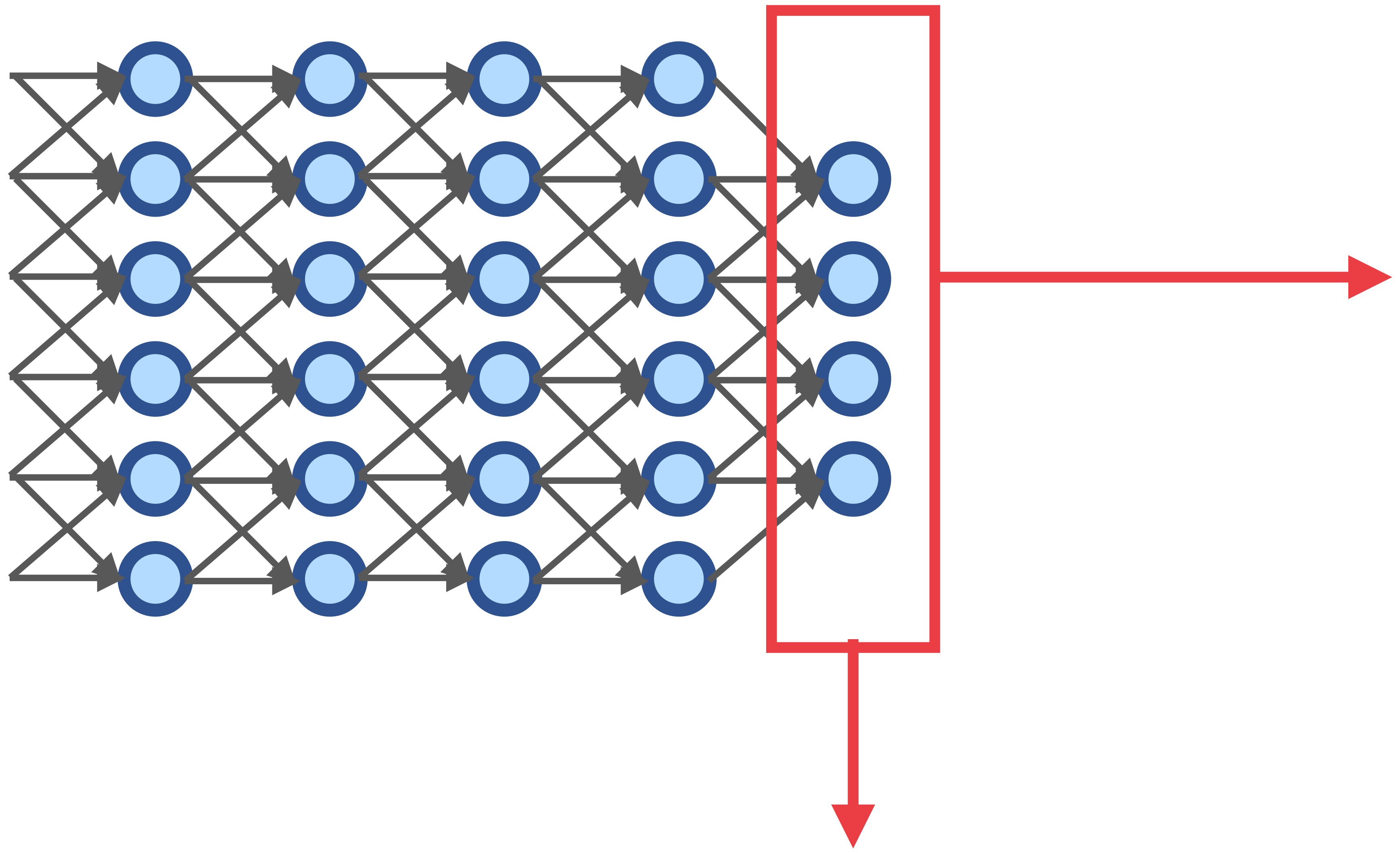
- Biases are present even in more generic and widespread Image Classifiers
- Biases are present even when gender is not one of the target variables
- Biases are present even when a best effort is placed on balancing the dataset for gender

Elazar and Goldberg (2018) introduced a notion of leakage from feature representations

$$y = f(x)$$

X: Text





Tweet Sentiment Objective

Can I predict gender or age from these features?

Adversarial Removal of Demographic Attributes from Text Data
Yanai Elazar, Yoav Goldberg. **EMNLP 2018**

Task: Multi-label Prediction

Annotations

Knife
Carrot
Table
Kitchen
Utensils

Man/Woman Classifier

Definition: Dataset Leakage

Annotations

Knife
Carrot
Table
Kitchen
Utensils

Man/Woman Classifier

Gender Leakage from the Dataset/Task

Definition: Dataset Leakage vs Model Leakage

Annotations (acc=100%)

Knife
Carrot
Table
Kitchen
Utensils

Man/Woman Classifier

Gender Leakage from the Dataset/Task

Predictions (acc = 58%)

Model

Knife
Carrot
Table
Kitchen
Pineapple

Man/Woman Classifier

Gender Leakage from the Model Predictions

Definition: Dataset Leakage vs Model Leakage

Annotations (acc=100%)

→ Knife
Carrot
Table
Kitchen
Utensils → **Man/Woman Classifier** → Dataset Leakage

Predictions (acc = 58%)

→ **Model** → Knife
Carrot
Table
Kitchen
Pineapple → **Man/Woman Classifier** → Model Leakage

Definition: Dataset Leakage @ 58% vs Model Leakage @ 58%

Annotations (acc=58%)

Random Perturbations →

Knife
Carrot
Table
Kitchen
Baseball

→ Man/Woman Classifier → Dataset Leakage

Predictions (acc = 58%)

→ Model →

Knife
Carrot
Table
Kitchen
Pineapple

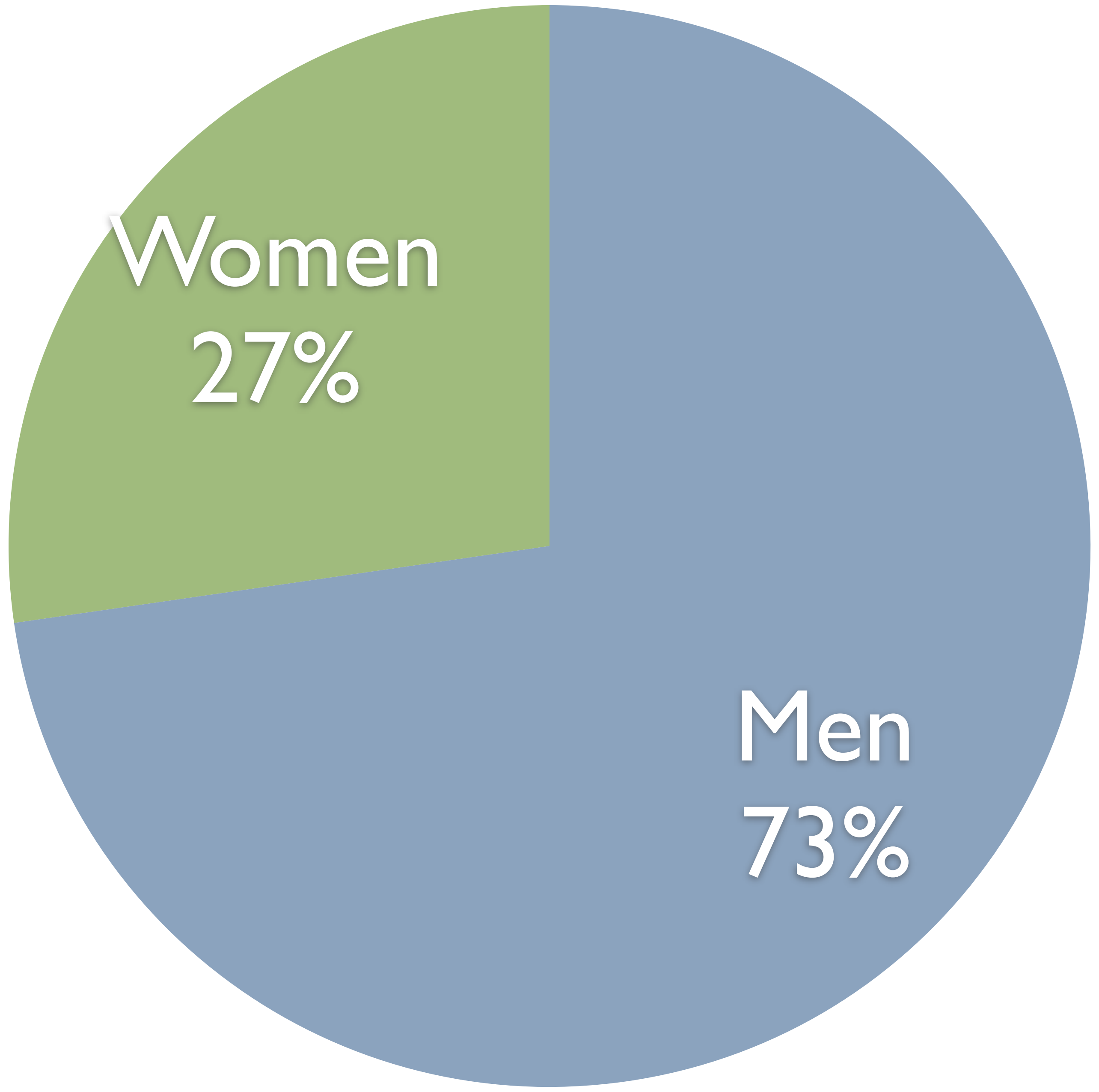
→ Man/Woman Classifier → Model Leakage

Definition: Bias Augmentation

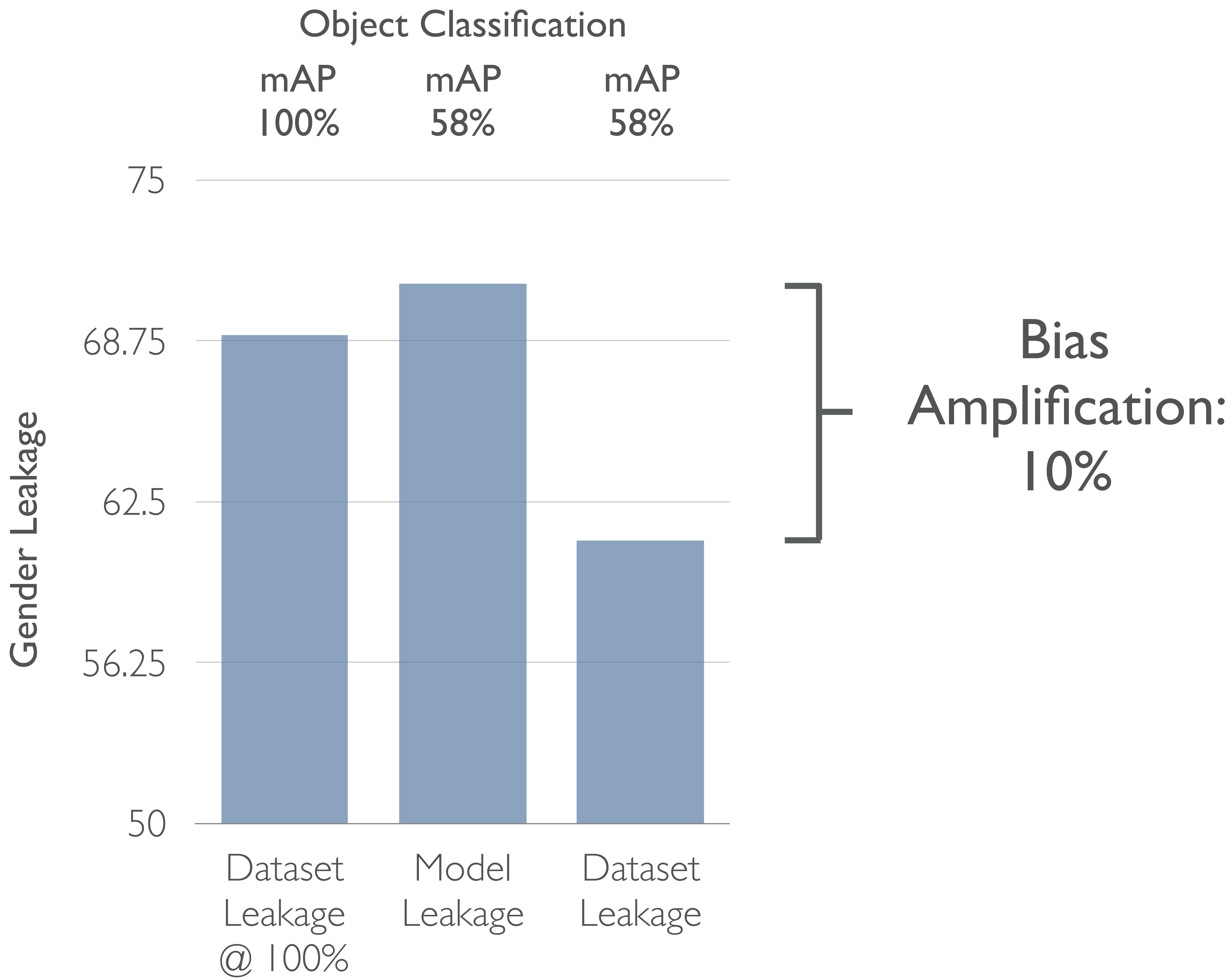
Definition: Model Leakage @ K - Dataset Leakage @ K

Key Finding: Models Leak even when Dataset doesn't

Training set size: 28k

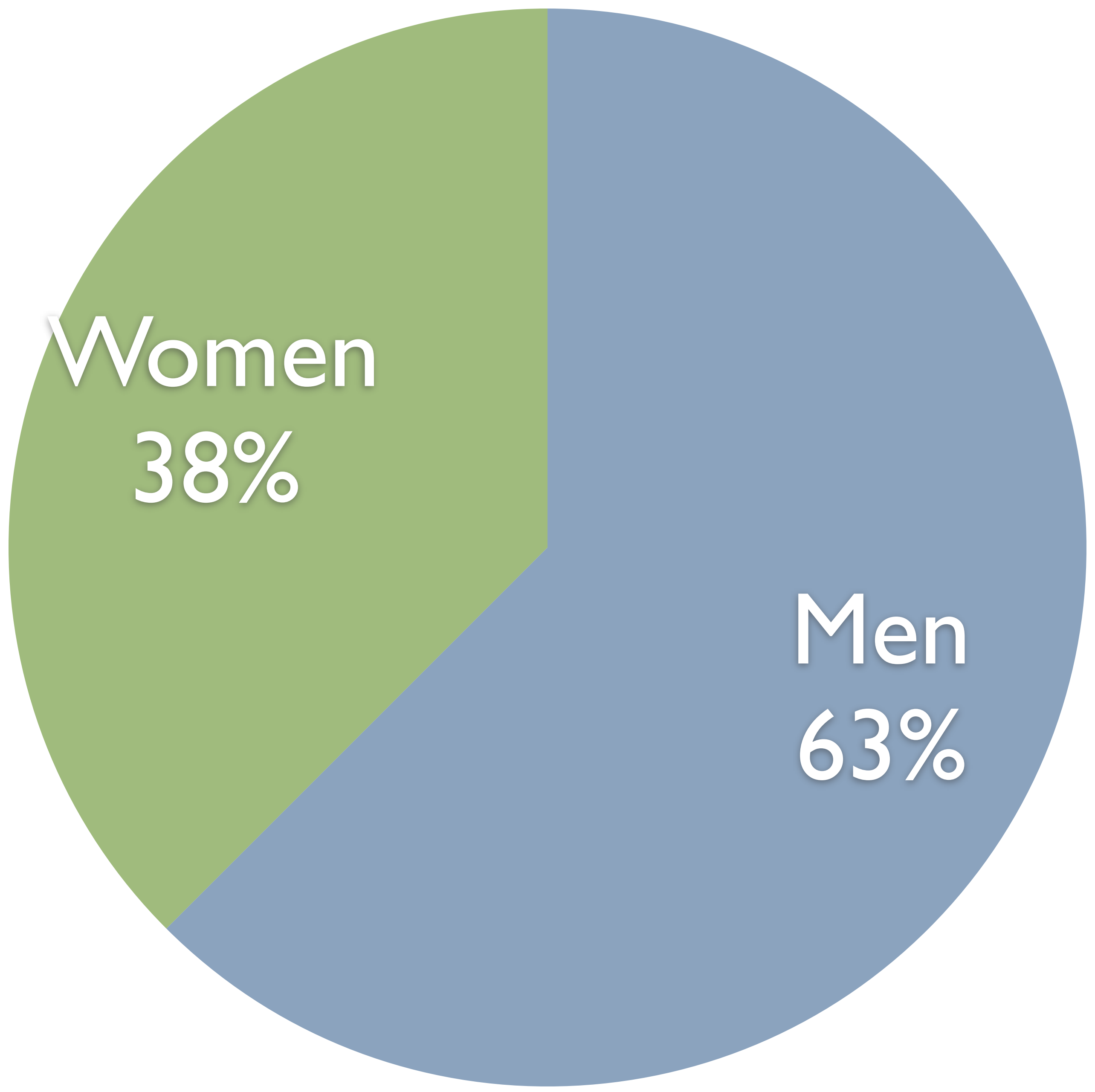


Task: Classify 80 objects

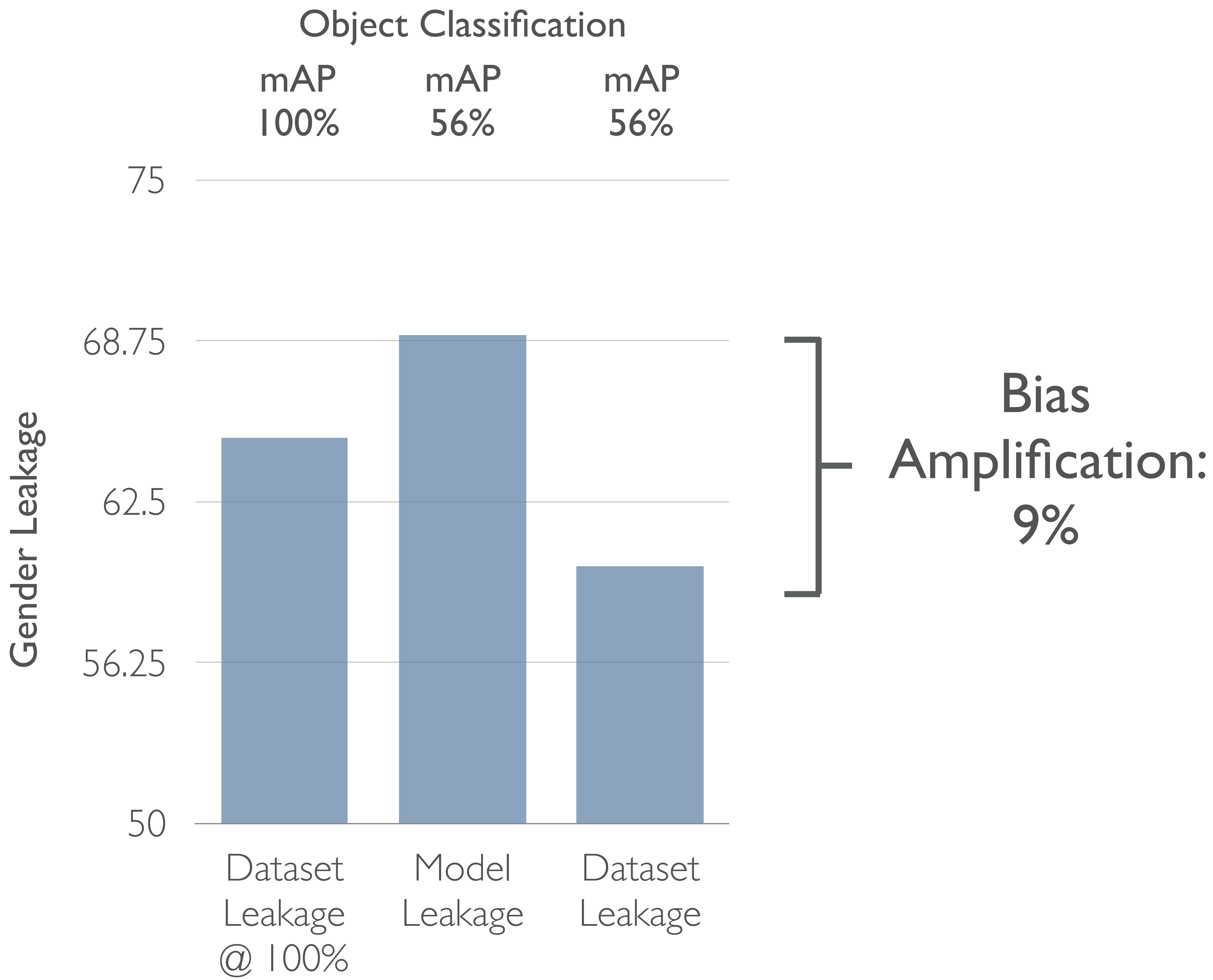


Key Finding: Models Leak even when Dataset doesn't

Training set size: 16k

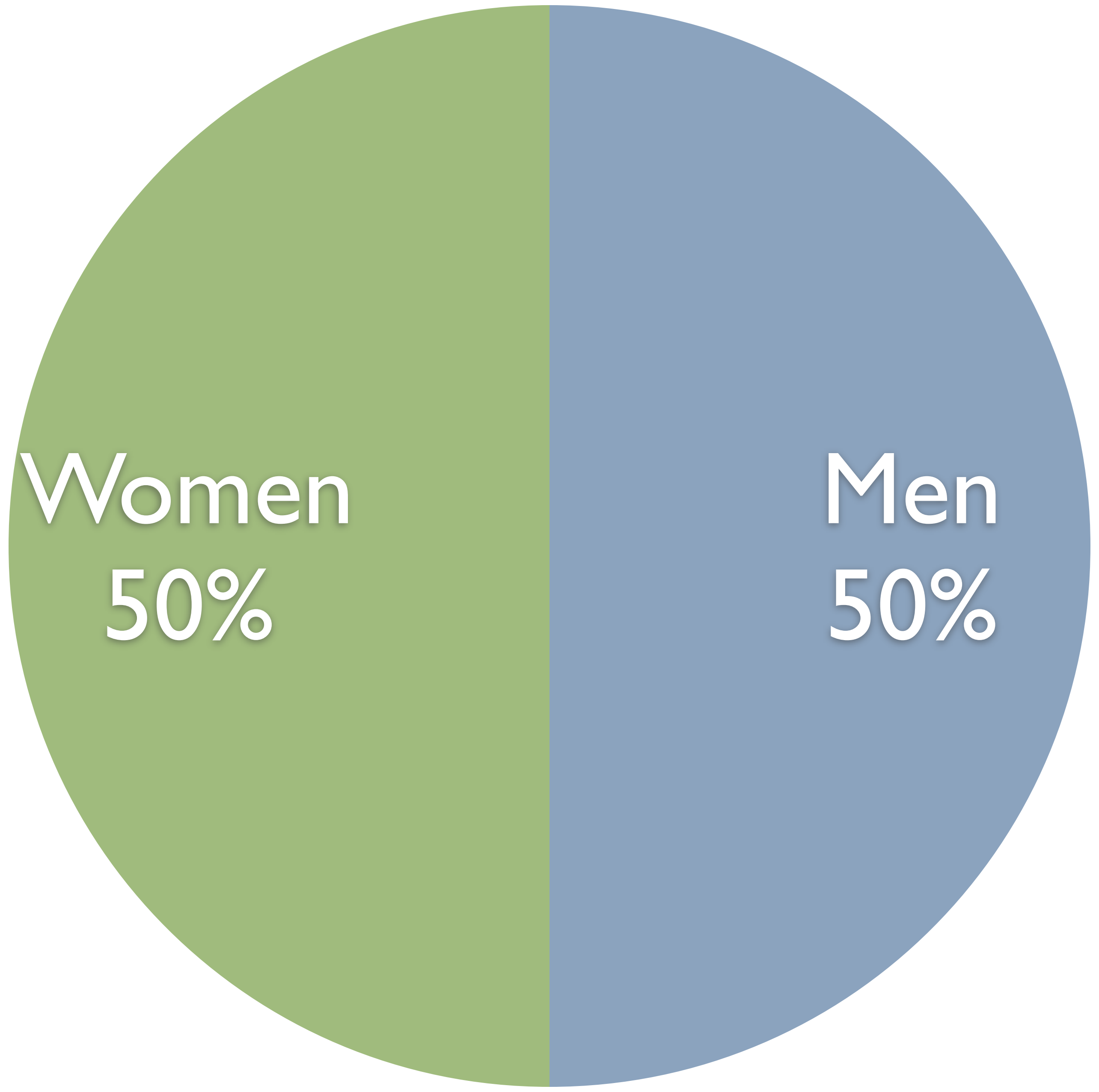


Task: Classify 80 objects



Key Finding: Models Leak even when Dataset doesn't

Training set size: 6k



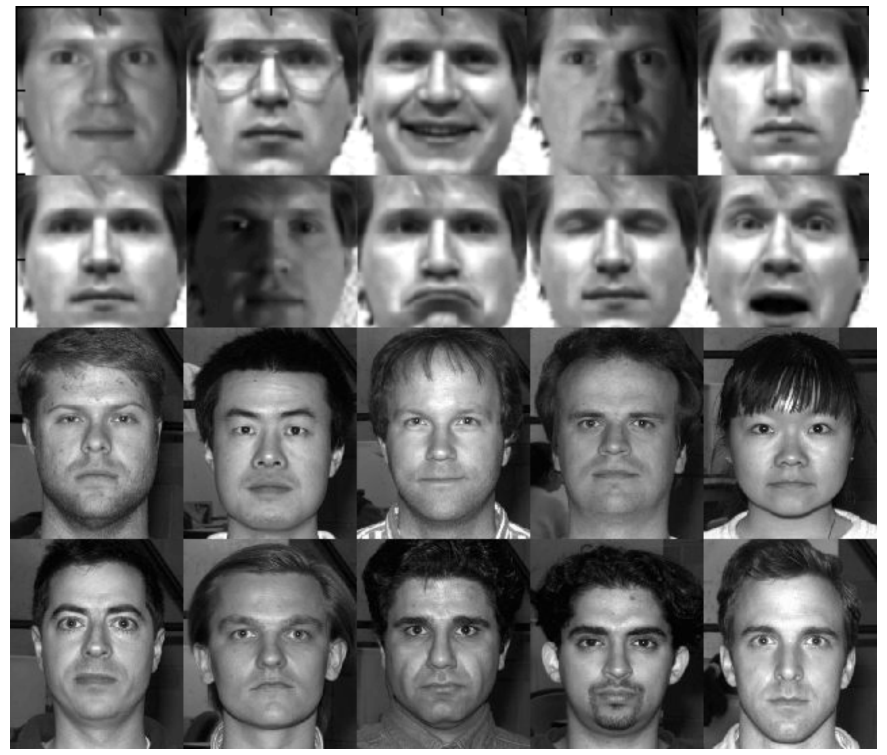
Task: Classify 80 objects

Issues Revelaed

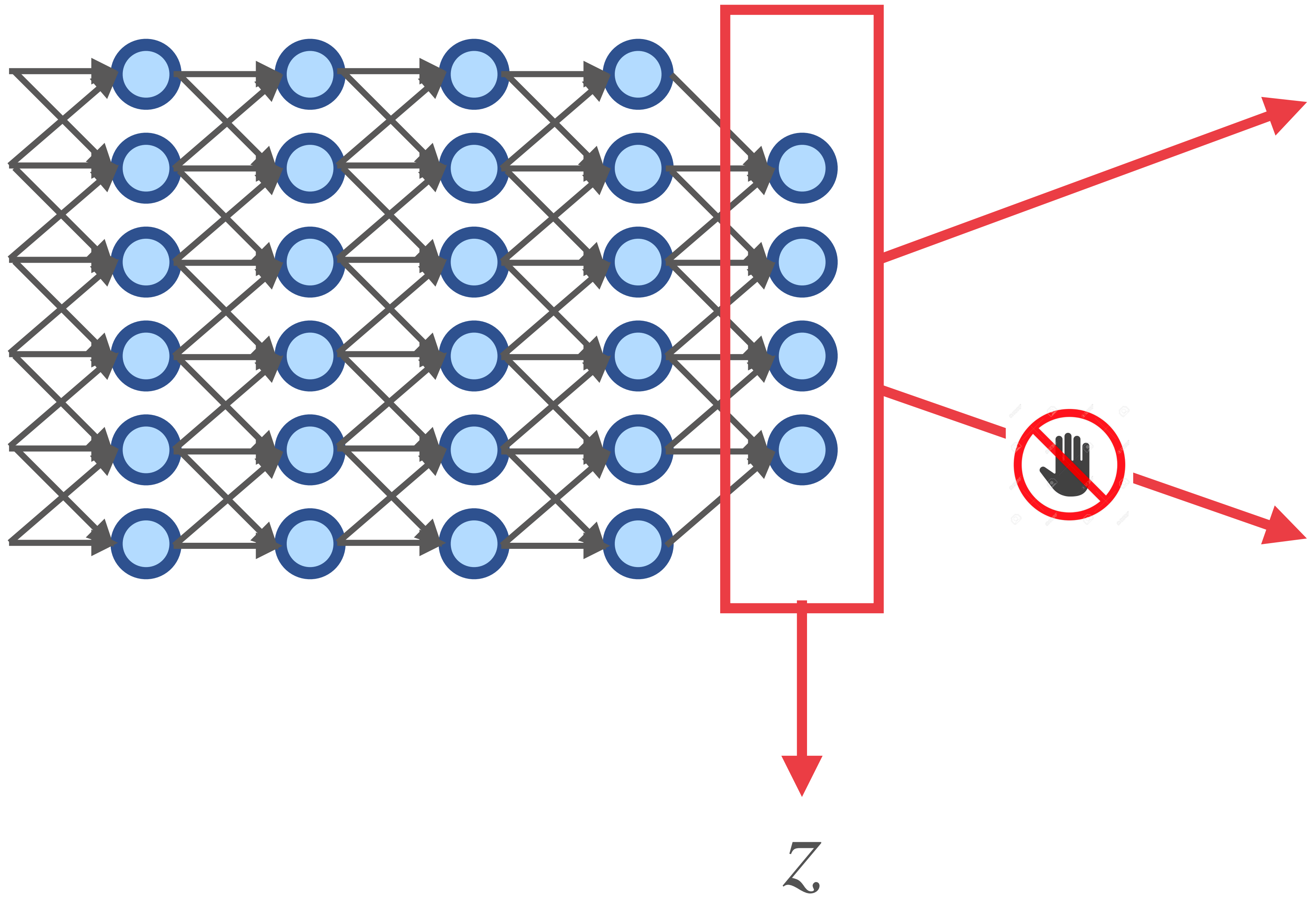
- Models are again shown to not only replicate but also amplify effects of protected variables.
- Balancing a dataset is hard - and not effective to mitigate bias as it is hard to balance against latent variables

Approach II: Adversarial Feature Learning

X: Images



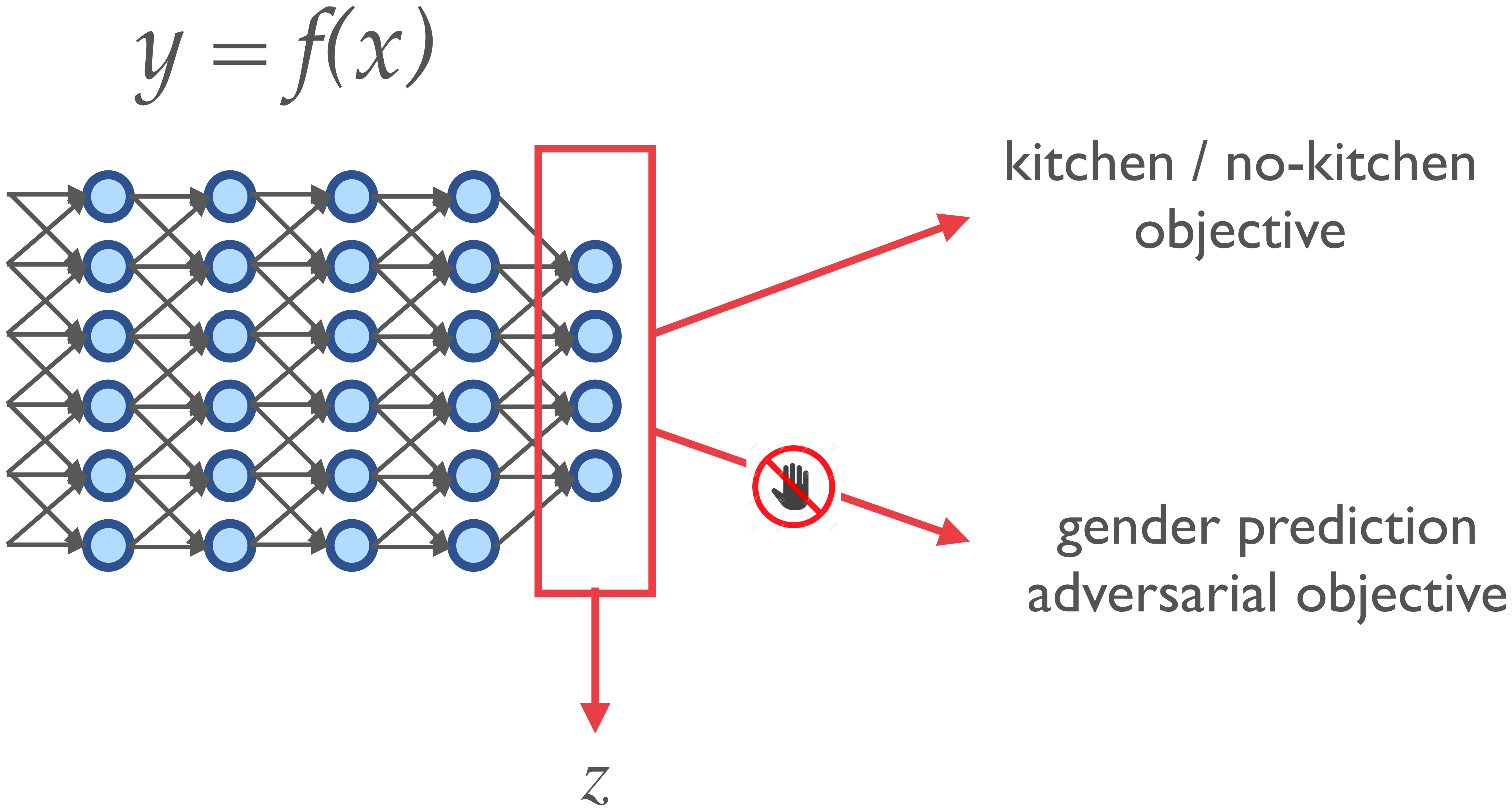
$$y = f(x)$$



Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

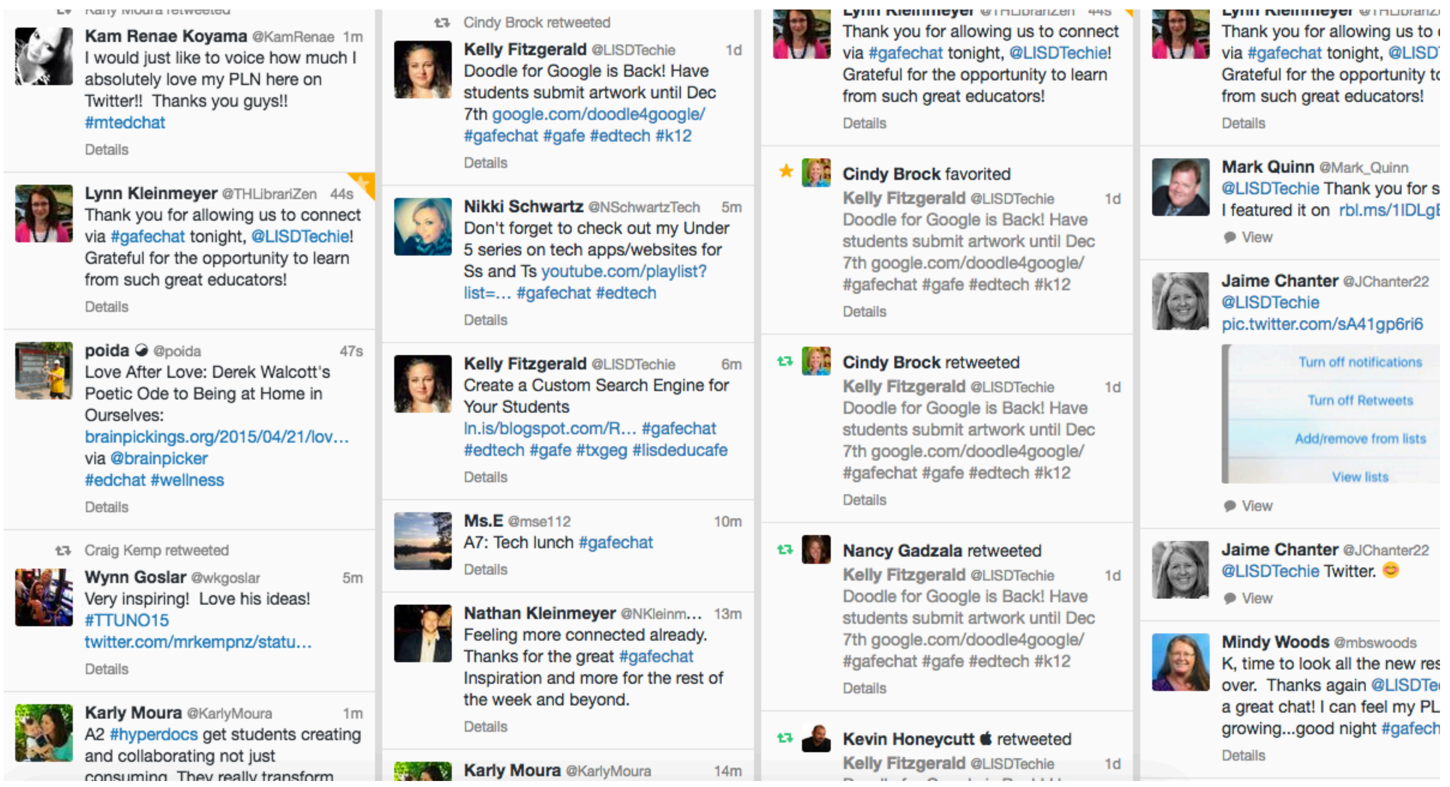
X: Images



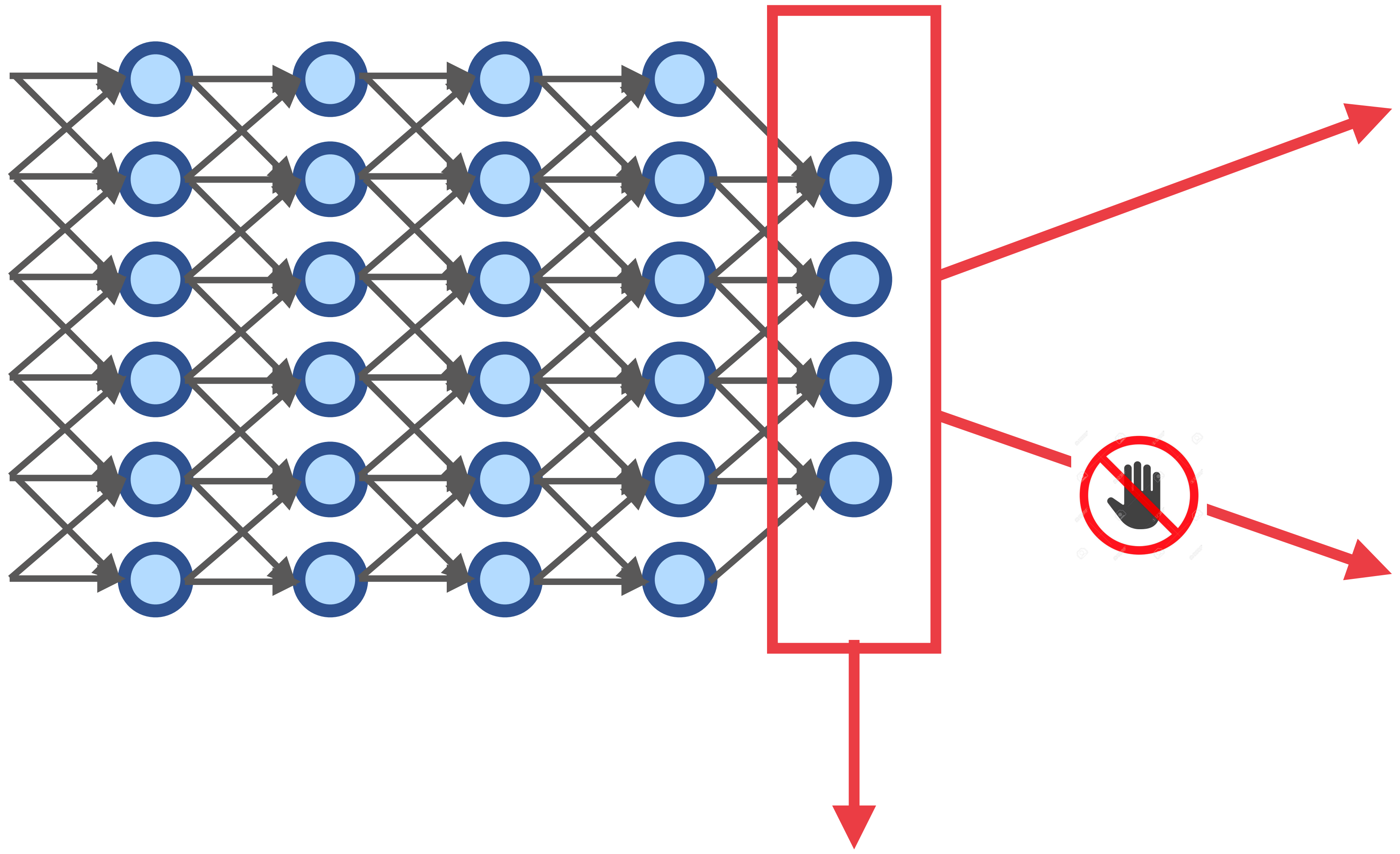
Controllable Invariance through Adversarial Feature Learning
Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig. **NeurIPS 2017**

Approach II: Adversarial Feature Learning

X: Text



$$y = f(x)$$



Tweet Sentiment Objective

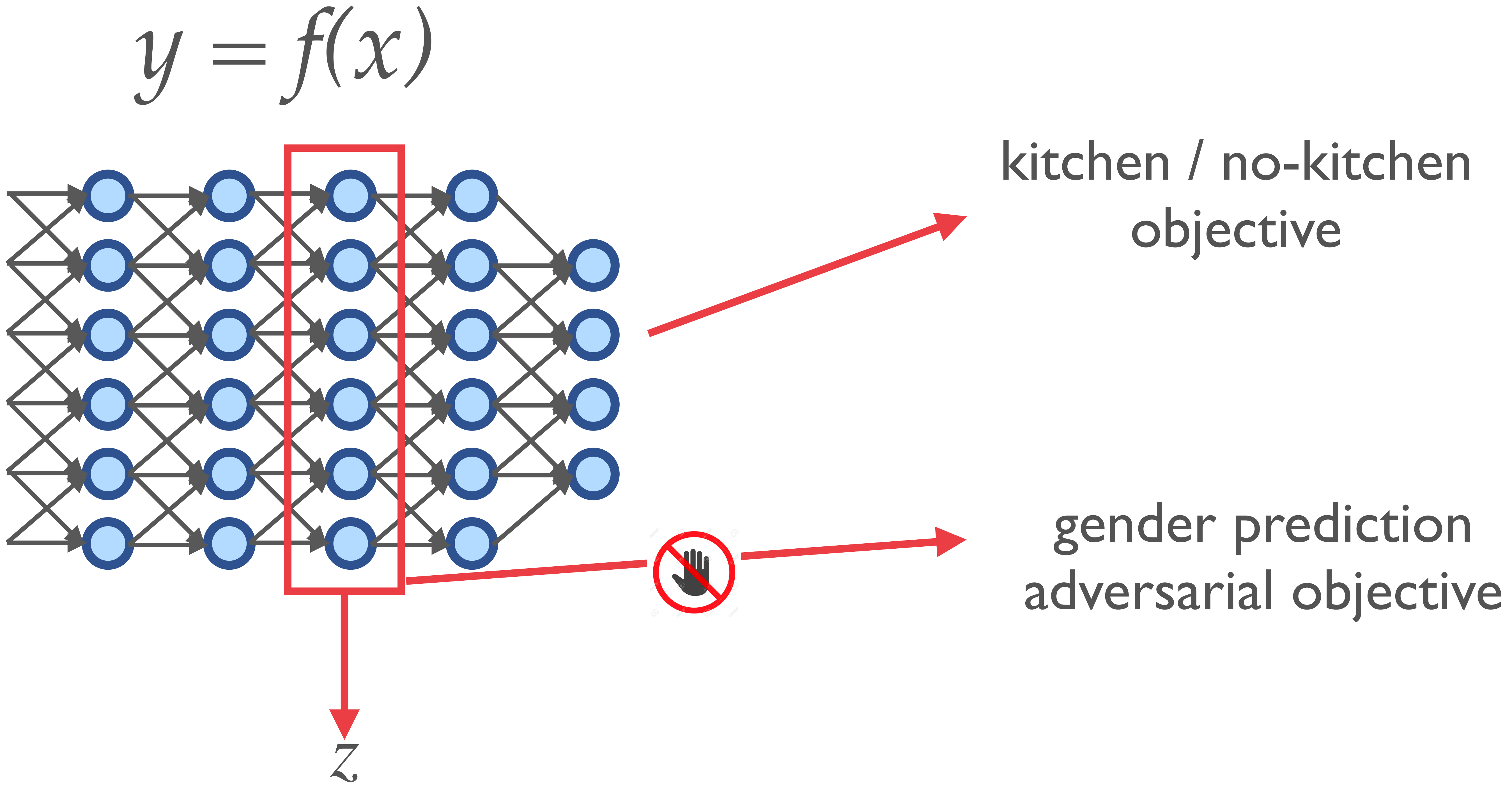
adversarial demographic prediction: age, gender

z

Adversarial Removal of Demographic Attributes from Text Data
Yanai Elazar, Yoav Goldberg. **EMNLP 2018**

Approach: Deep Adversarial Feature Learning

X: Images

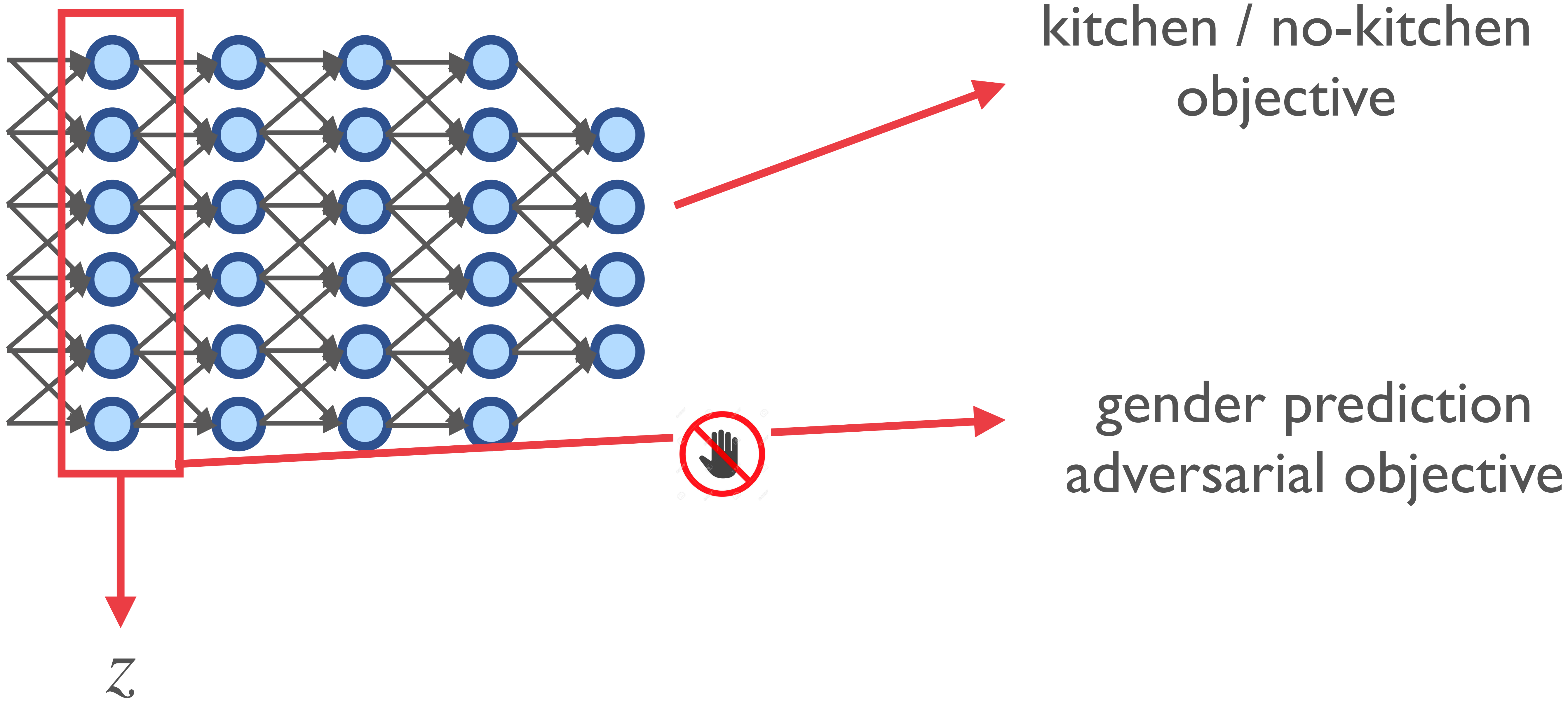


Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.
Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez. **ICCV 2019**

Approach: Deep Adversarial Feature Learning

X: Images

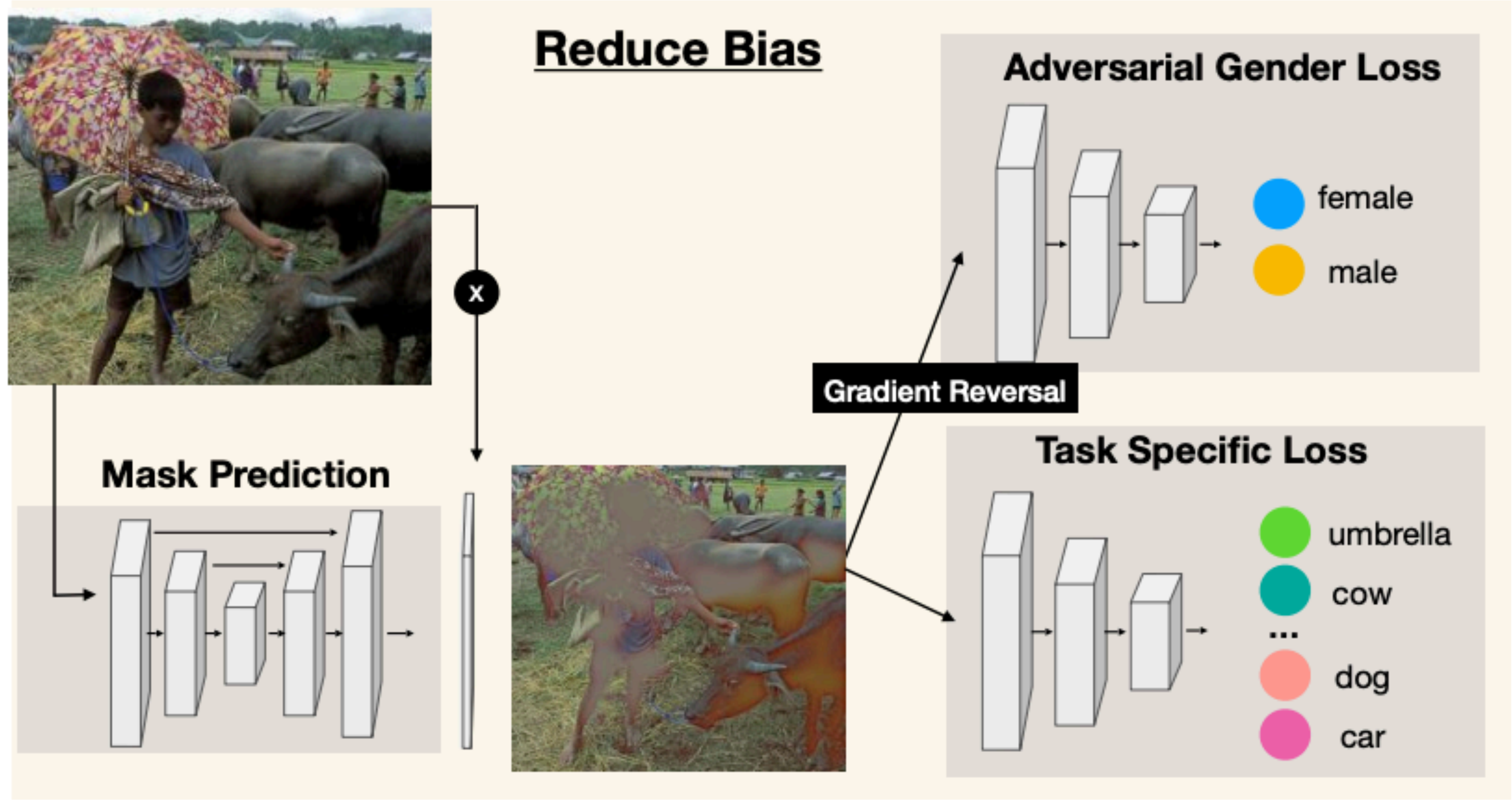
$y = f(x)$



Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations.
Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez. **ICCV 2019**

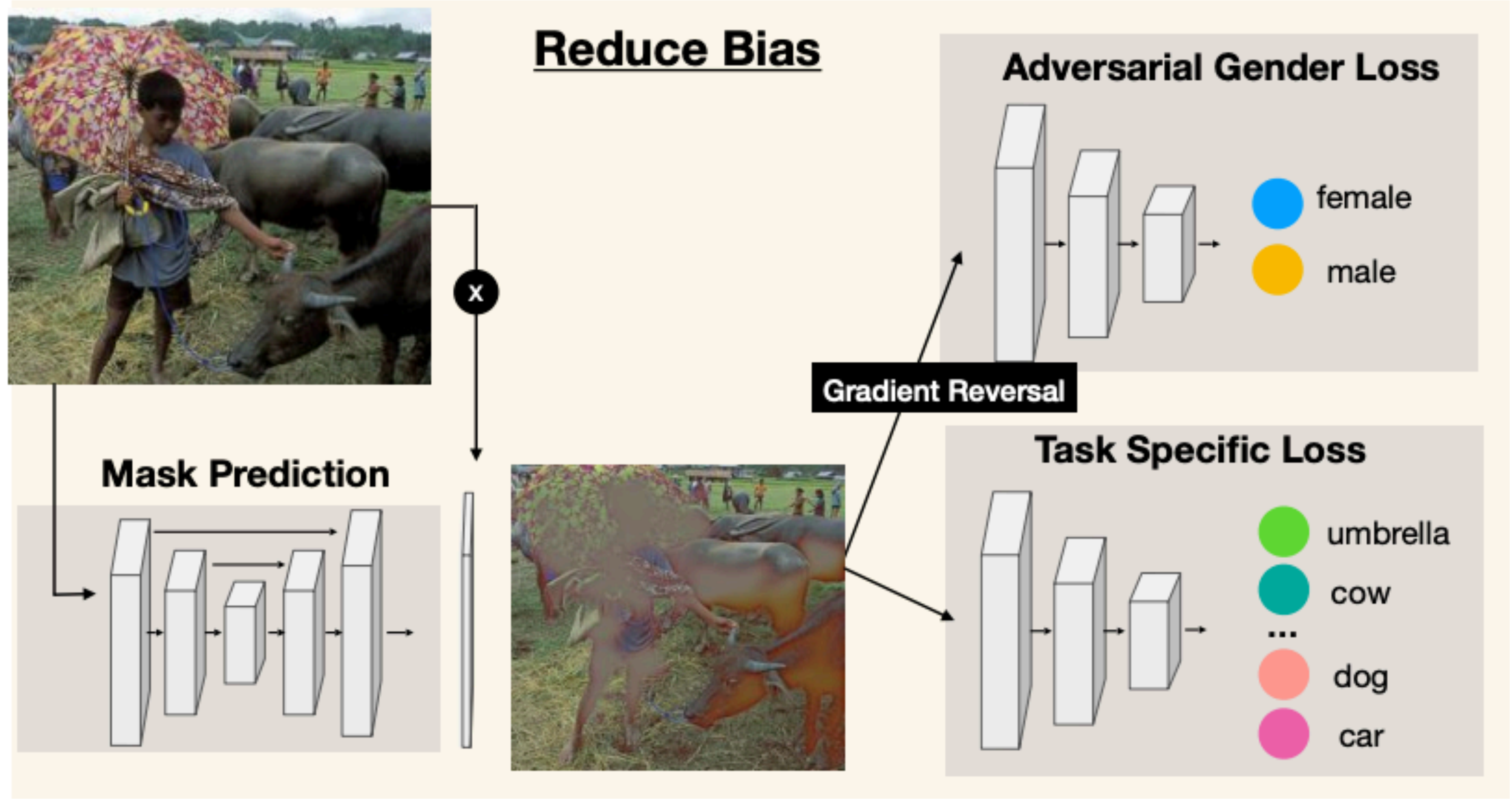
Adversarial Removal of Sensitive Features

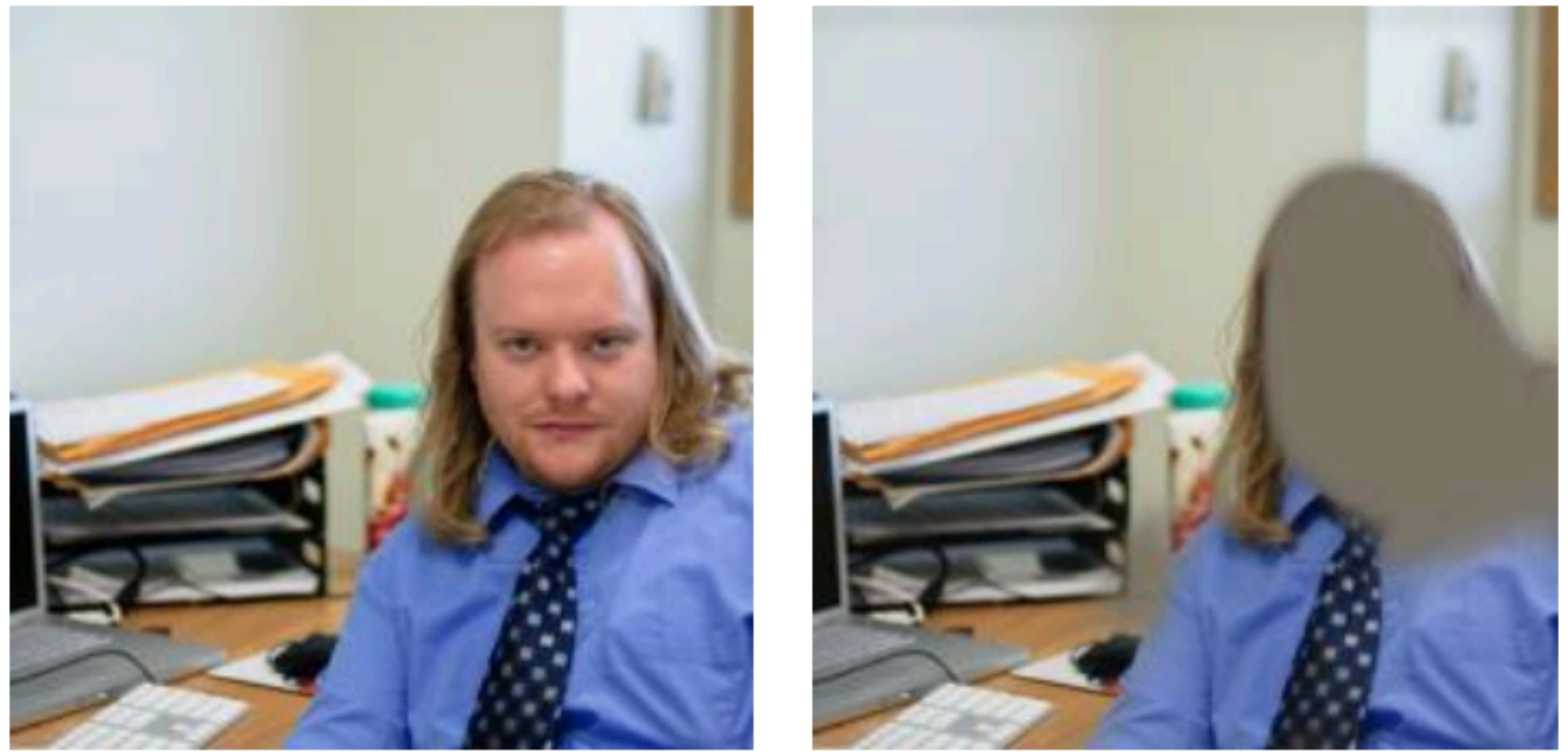
i.e. Predict Objects while trying to obscure gender



Adversarial Removal of Sensitive Features

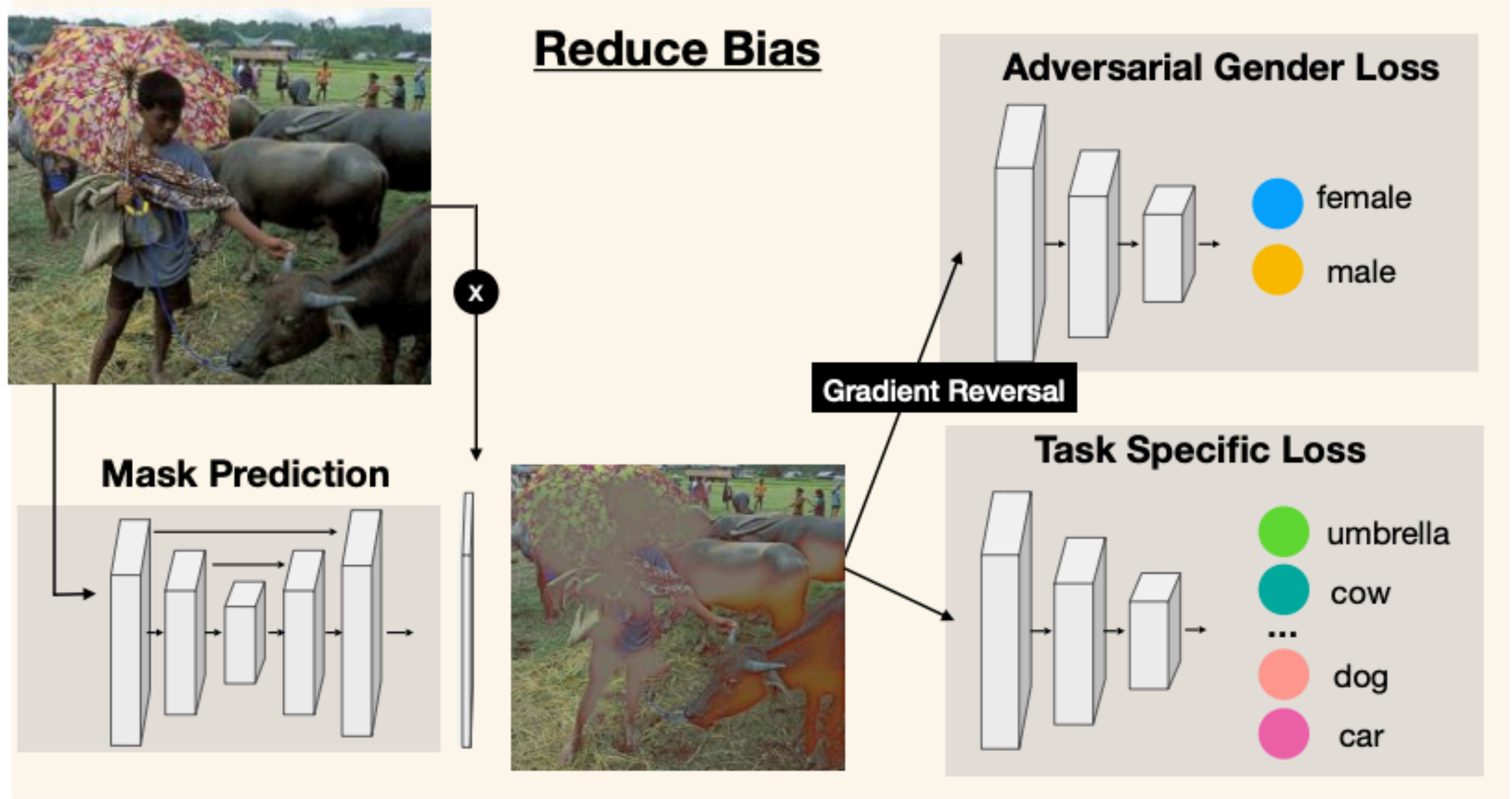
i.e. Predict Objects while trying to obscure gender

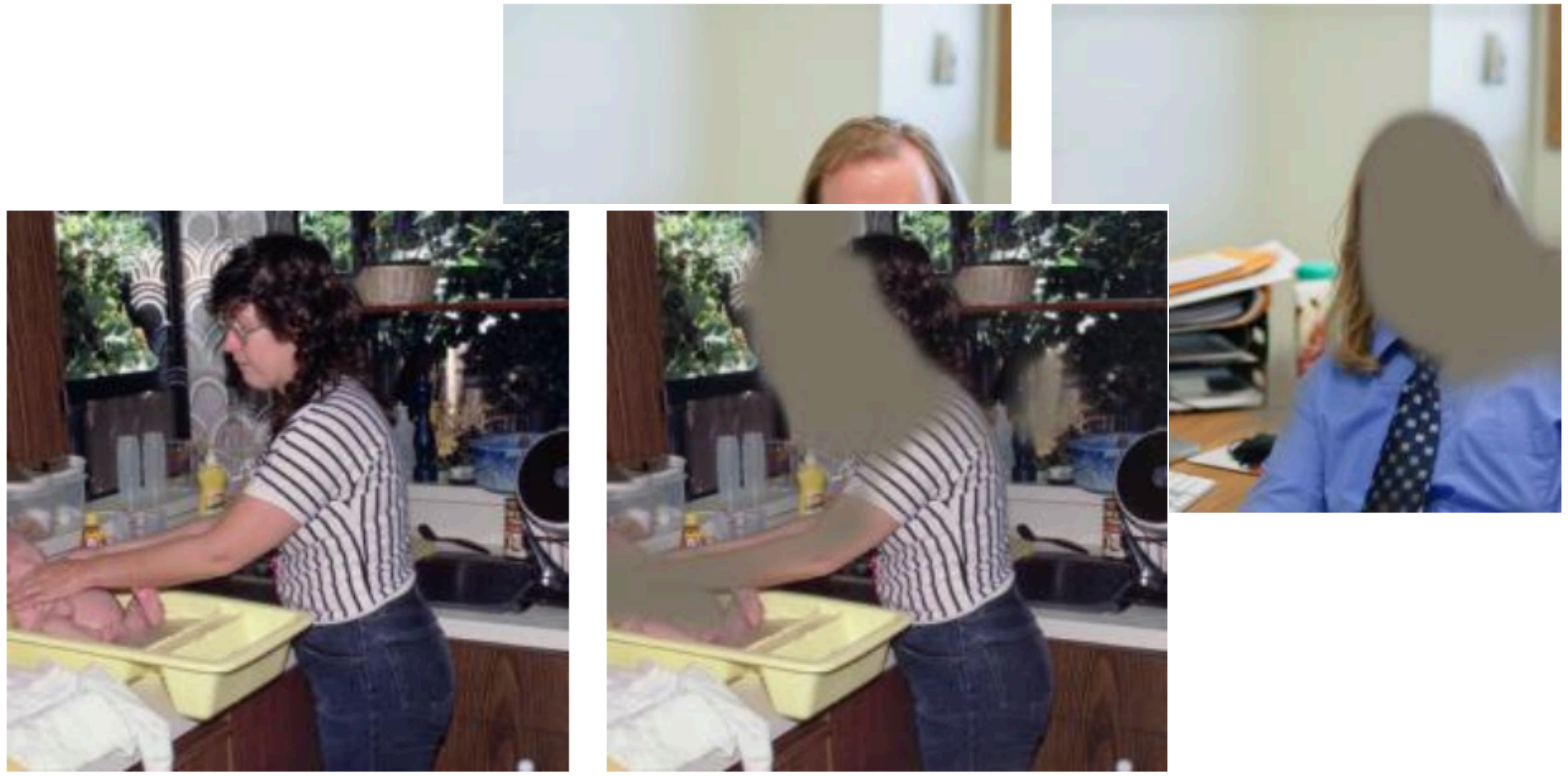




Adversarial Removal of Sensitive Features

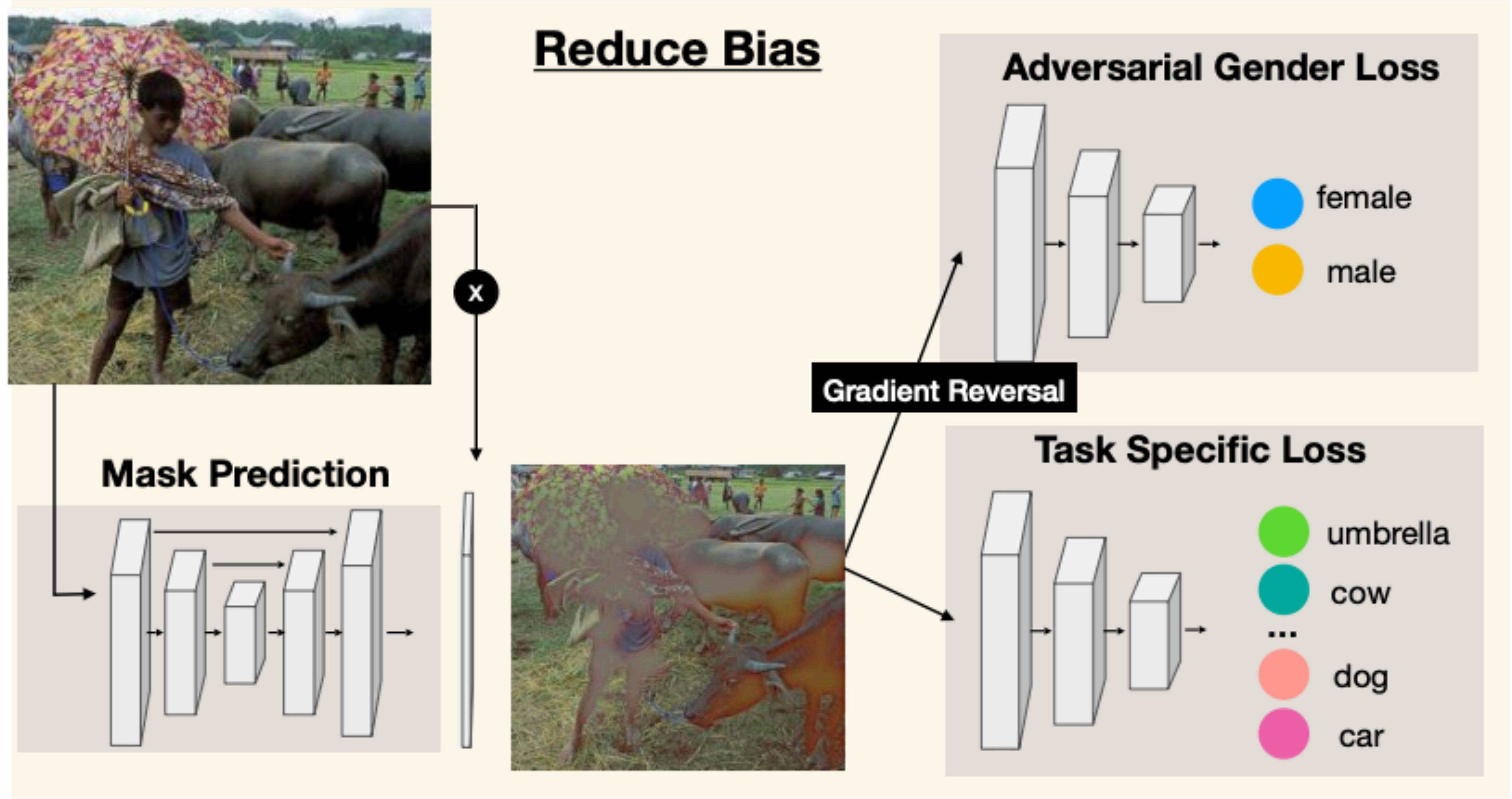
i.e. Predict Objects while trying to obscure gender





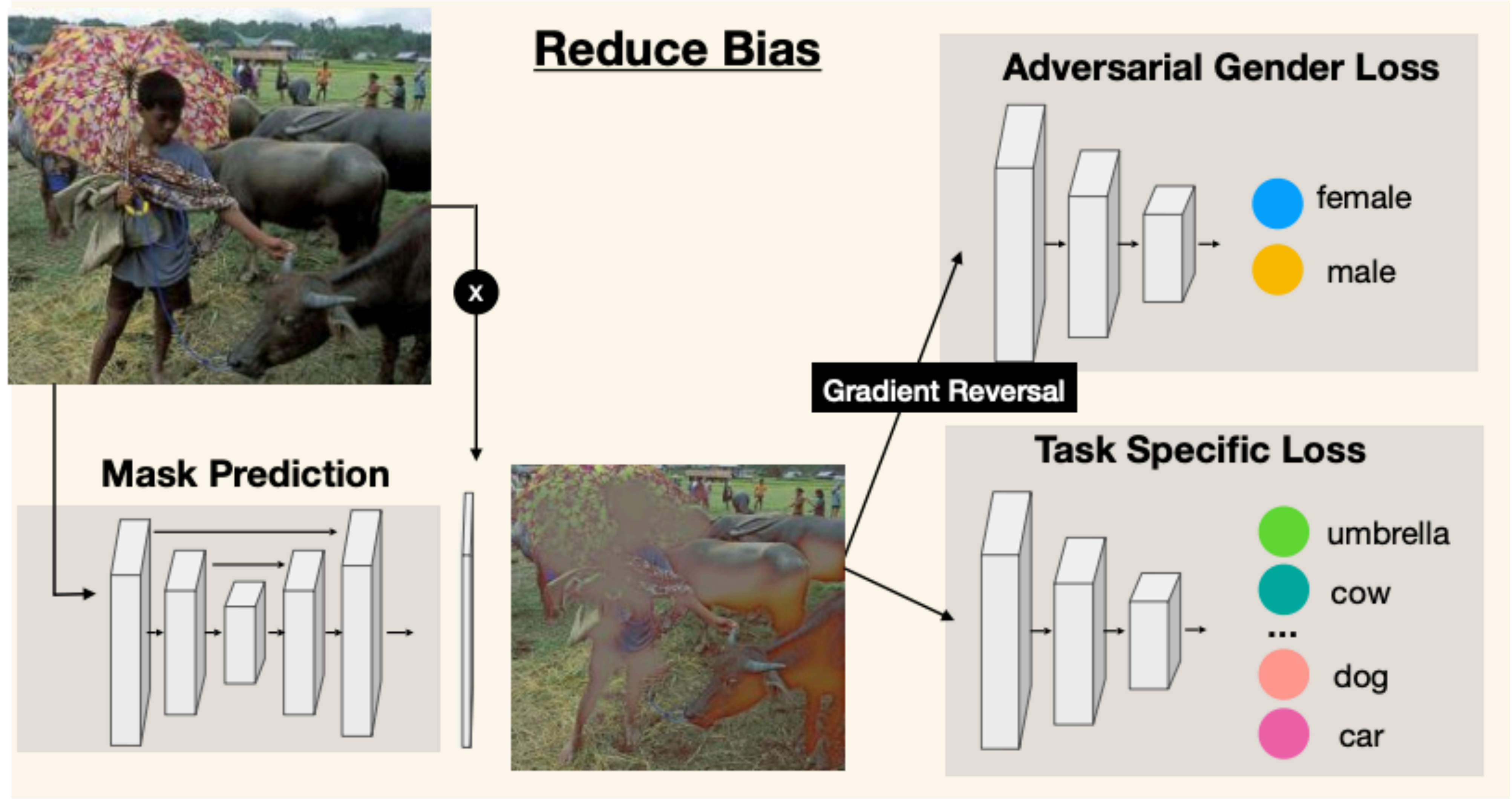
Adversarial Removal of Sensitive Features

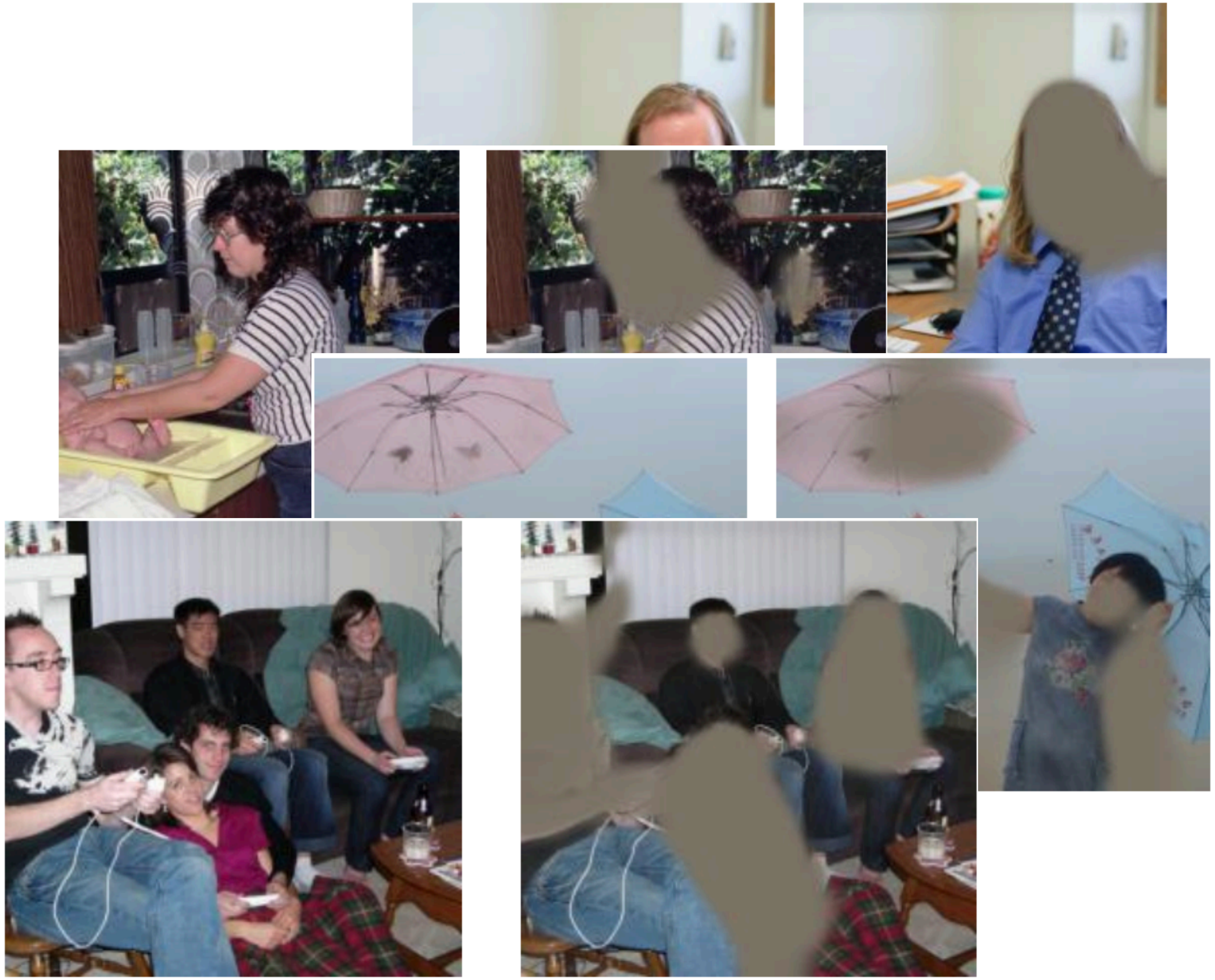
i.e. Predict Objects while trying to obscure gender



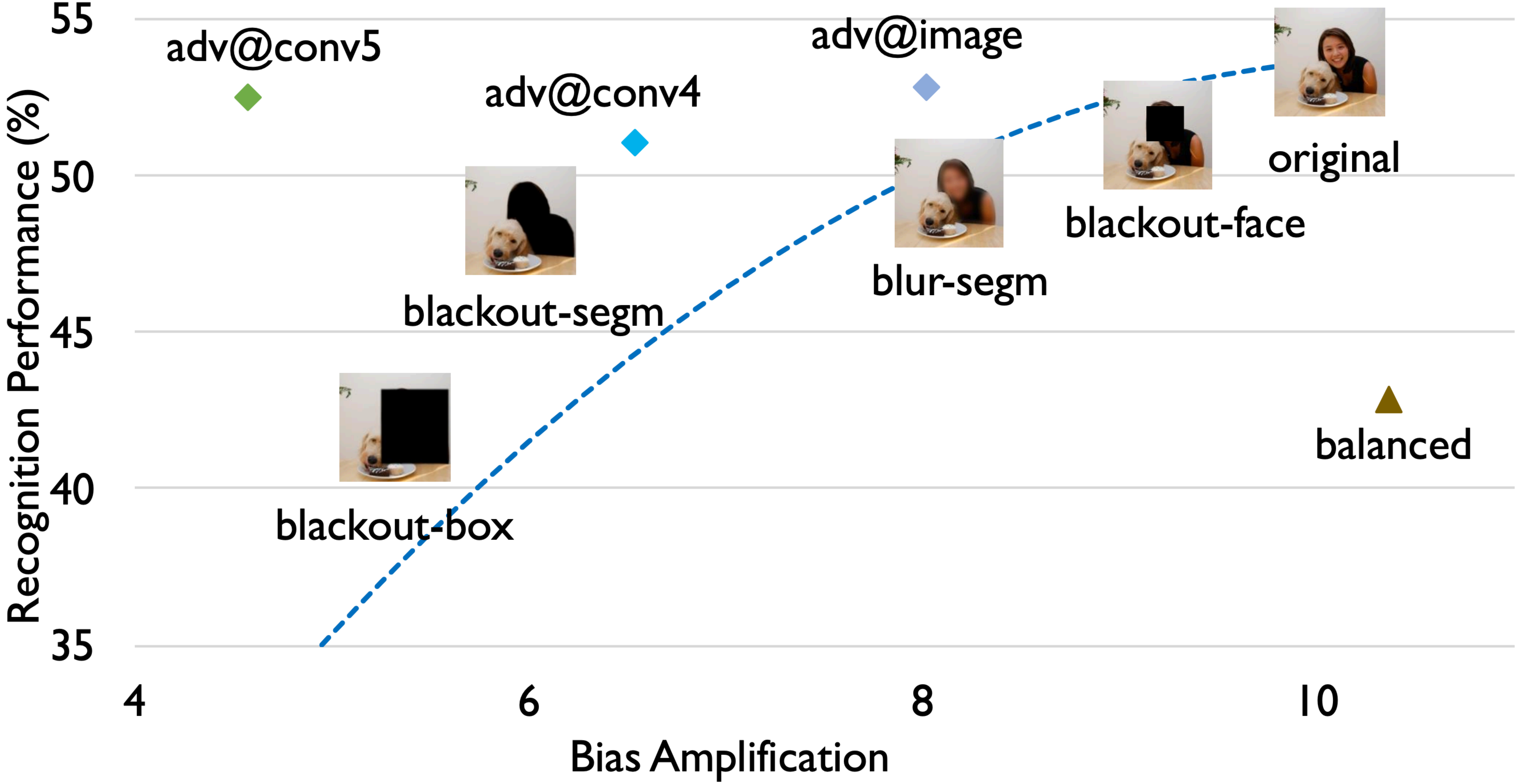
Adversarial Removal of Sensitive Features

i.e. Predict Objects while trying to obscure gender





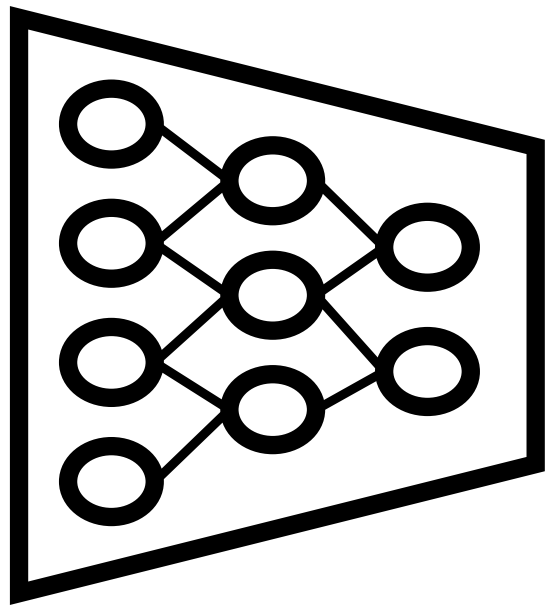
Results



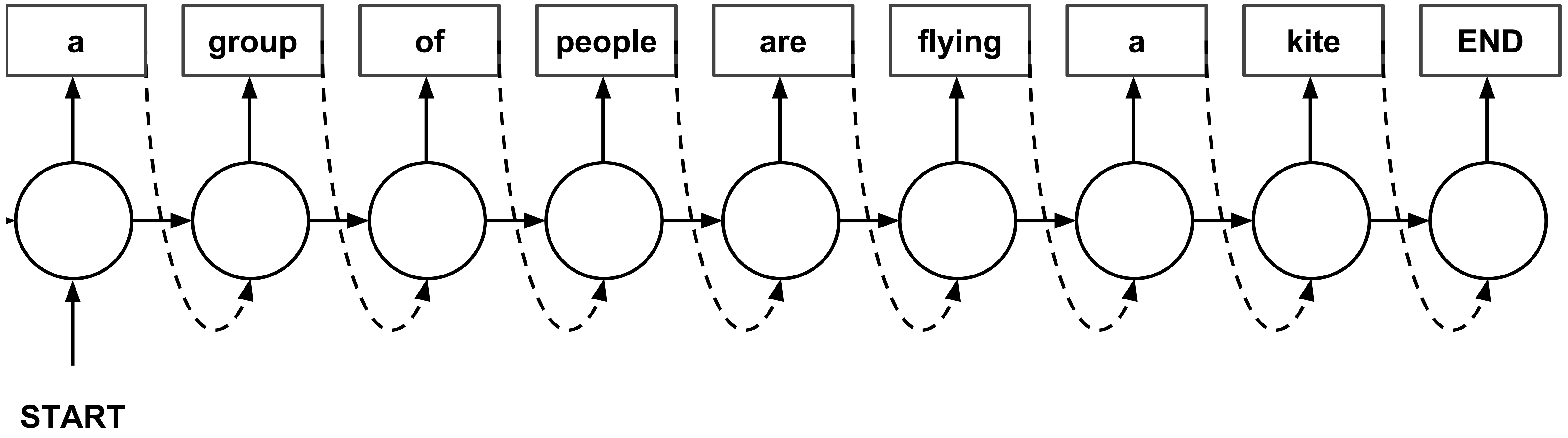
55
50
45
40
35
Recognition Performance (%)
adv@conv5
adv@conv4
adv@image
original
blackout-face
blur-segm
blackout-segm
blackout-box
balanced
4
6
8
10
Bias Amplification

Case Study: Image Captioning

Deep
Convolutional
Neural Network



Recurrent Neural Text Decoder



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

Case Study: Image Captioning

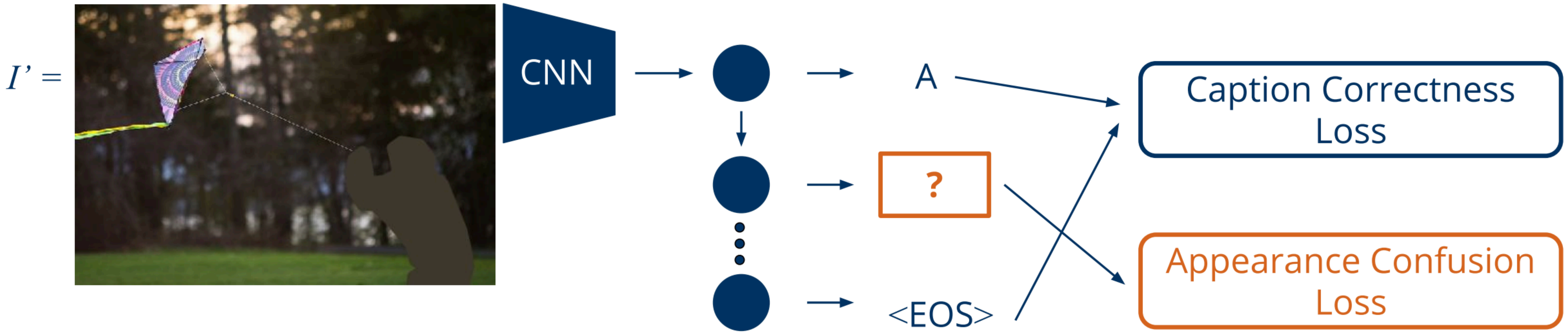
A woman cooking a meal

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

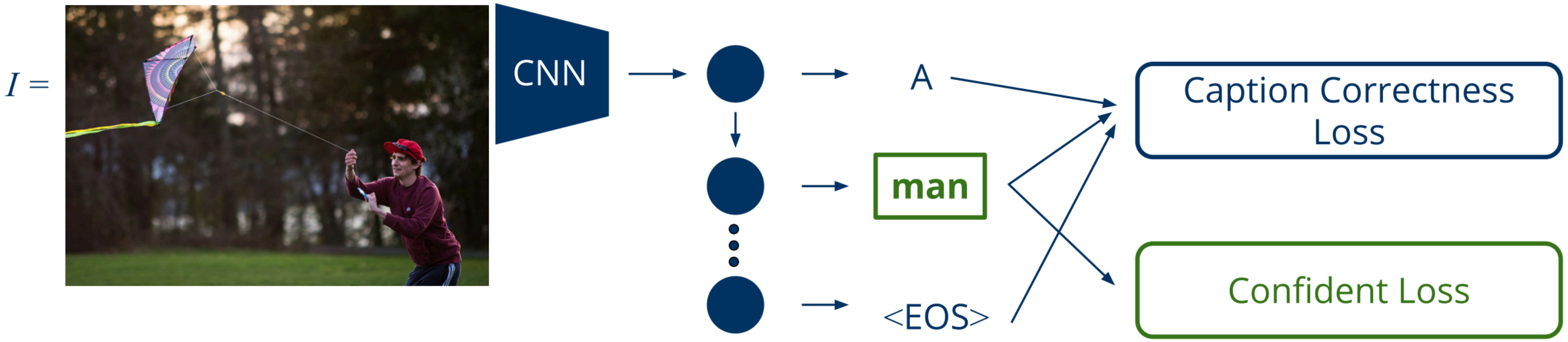
$$\mathcal{C}(\tilde{w}_t, I') = \Big| \sum_{g_w \in \mathcal{G}_w} p(\tilde{w}_t = g_w | w_{0:t-1}, I') - \sum_{g_m \in \mathcal{G}_m} p(\tilde{w}_t = g_m | w_{0:t-1}, I') \Big|$$

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

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

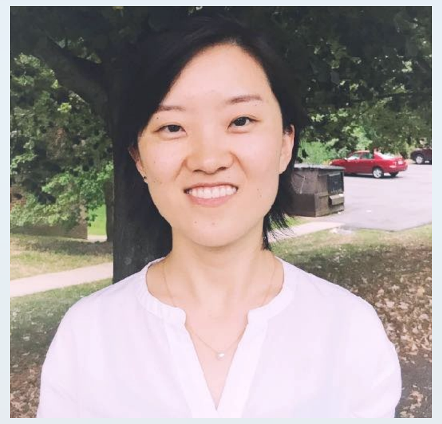
$$\mathcal{L}^{Con} = \frac{1}{N}\sum_{n=0}^{N}\sum_{t=0}^{T}(\mathbb{1}(w_t \in \mathcal{G}_w)\mathcal{F}^W(\tilde{w}_t, I) + \mathbb{1}(w_t \in \mathcal{G}_m)\mathcal{F}^M(\tilde{w}_t, I))$$

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

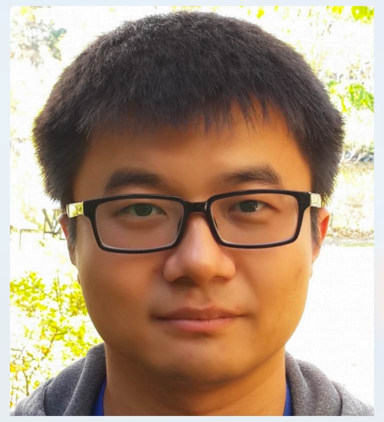
Women also Snowboard: Overcoming Bias in Captioning Models
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach. **ECCV 2018**

Students and Collaborators

Tianlu Wang

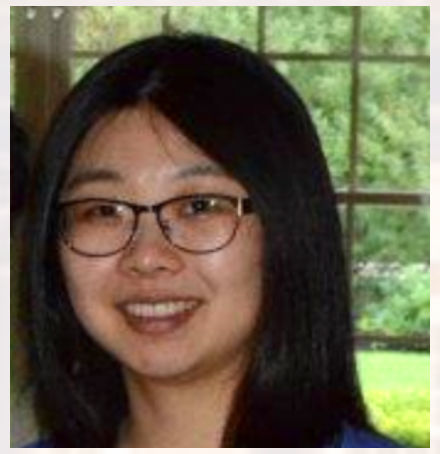


Jieyu Zhao



Xiaoxiao Guo

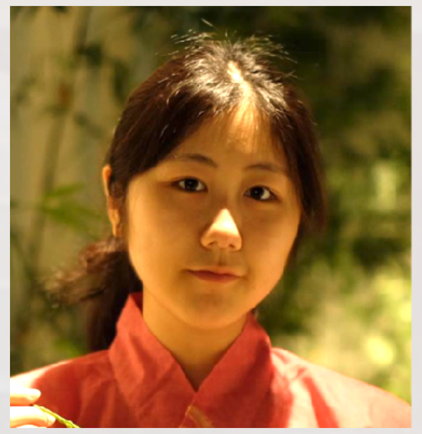
Mark Yatskar



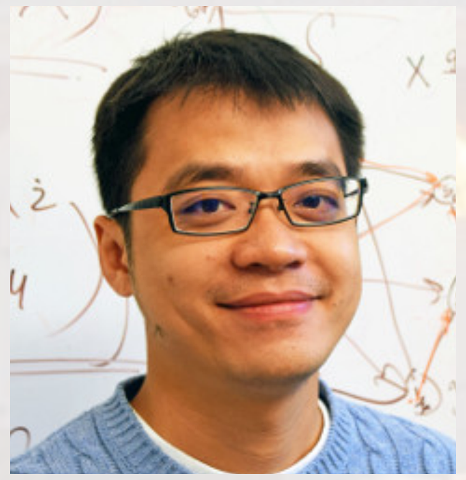
Song Feng

Hui Wu

Paola Cascante



Ziyan Yang



Fuwen Tan

Kai-Wei Chang

Baishakhi Ray

SAP®

leidos

IBM Research

Google