22 December 2021

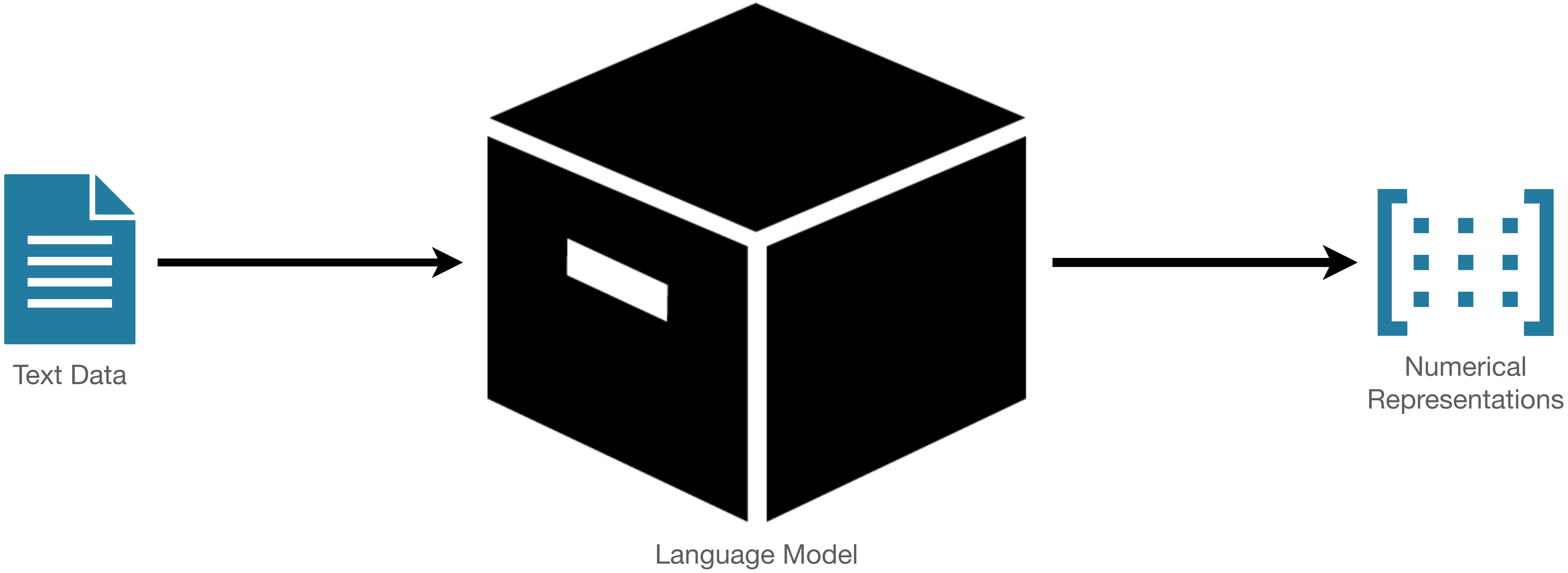
Research Presentation

Bias and Fairness in Natural Language Processing

Mahdi Zakizadeh (m.zakizadeh@khatam.ac.ir)
Kaveh Eskandari (kaveeskandari96@gmail.com)

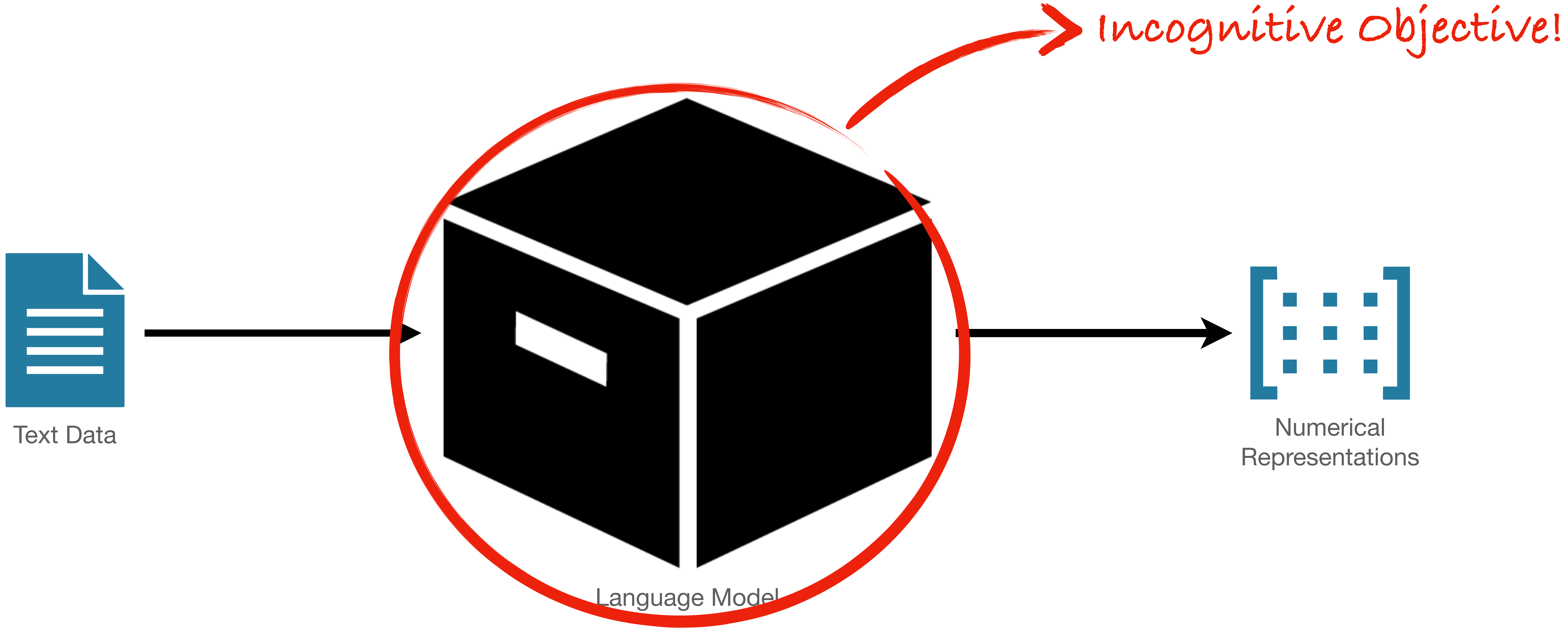
Introduction

Language Model as a Black Box



Introduction

Language Model as a Black Box



Introduction

Motivating Examples

- **Search Engine Ranking Bias**
- Resume Filtering
- Recidivism Prediction Instrument
- Toxicity Detection

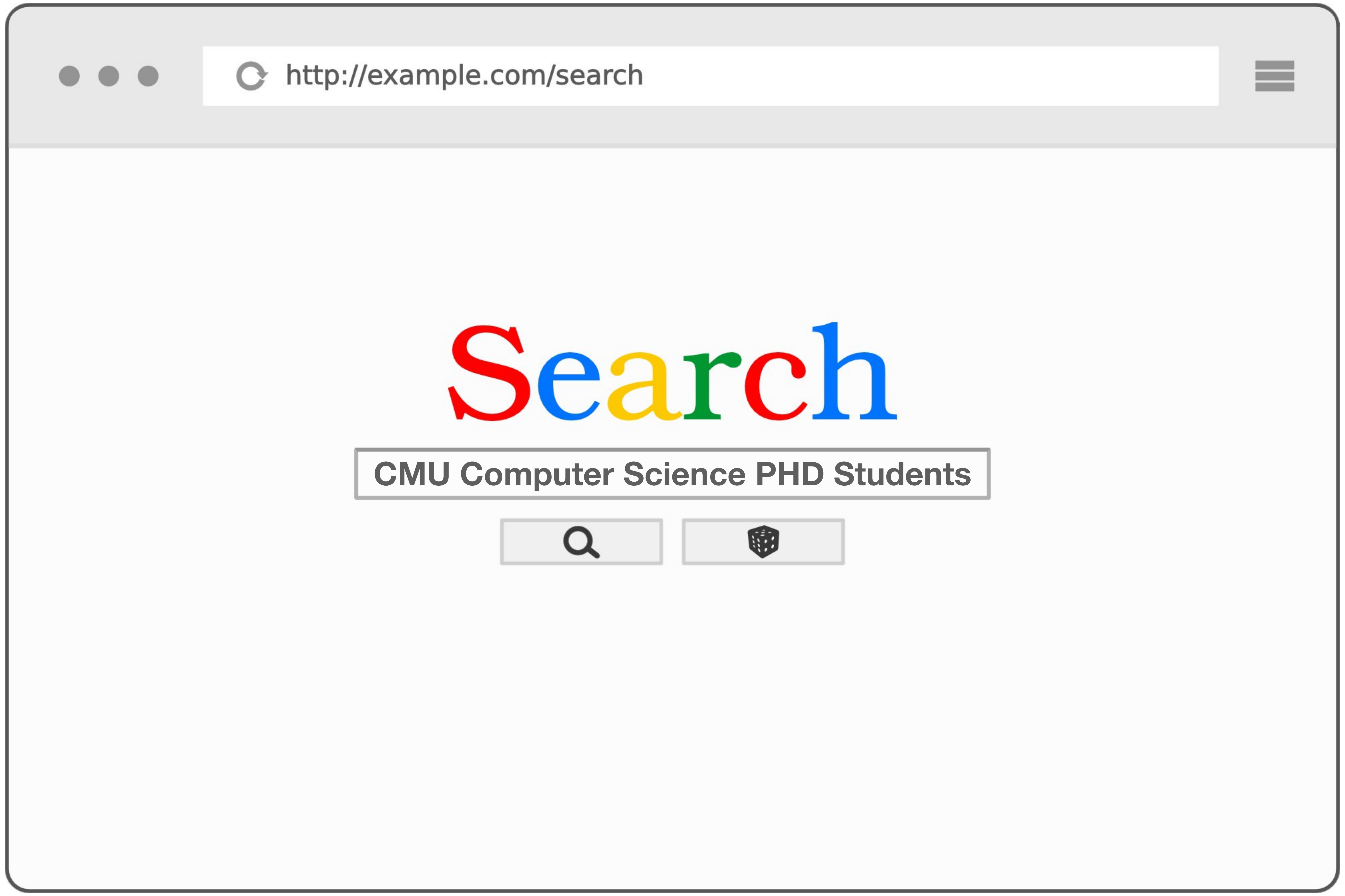
Introduction

Motivating Examples

[Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? (NIPS 2016)]

Introduction

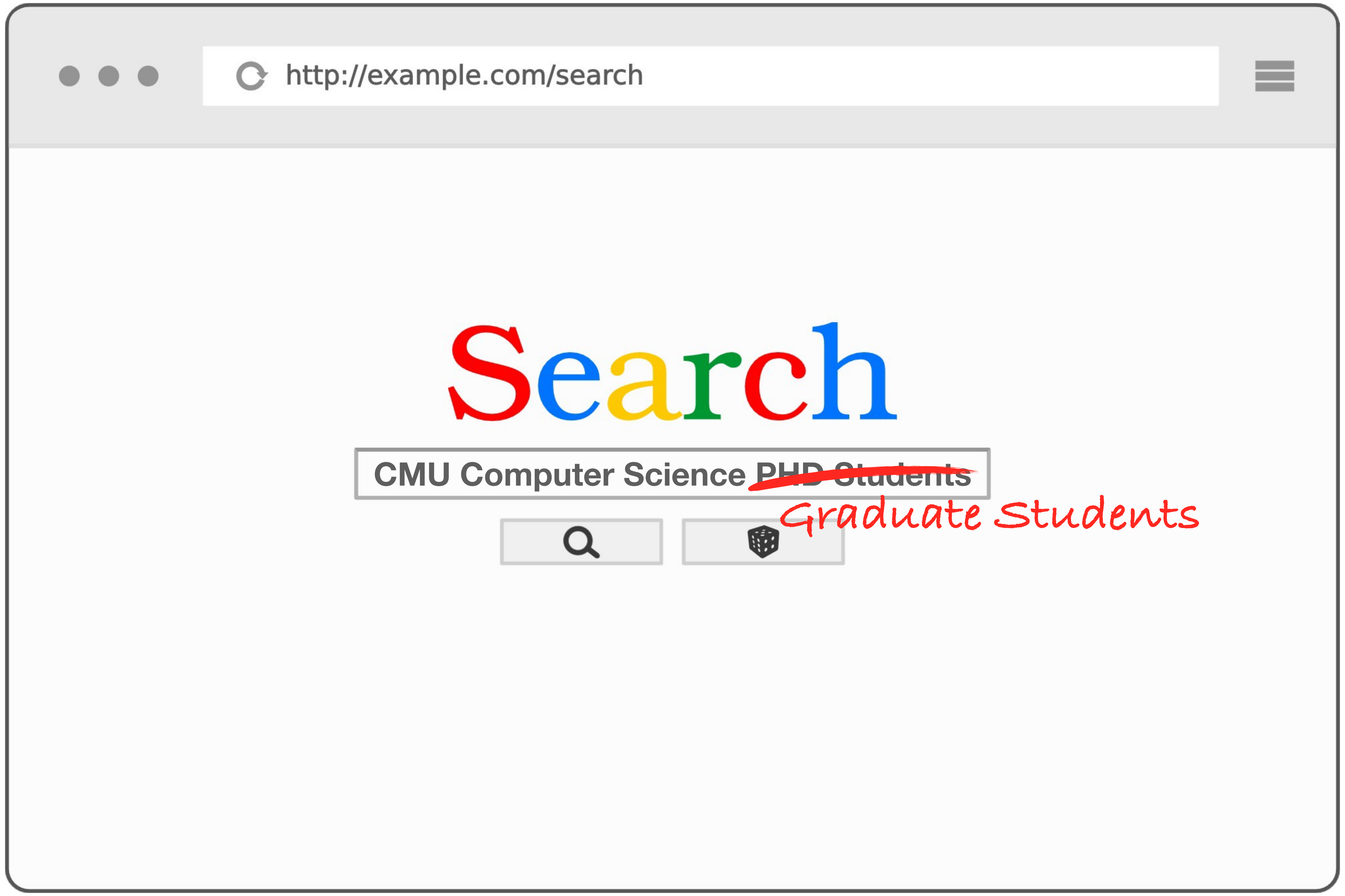
Motivating Examples



[Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? (NIPS 2016)]

Introduction

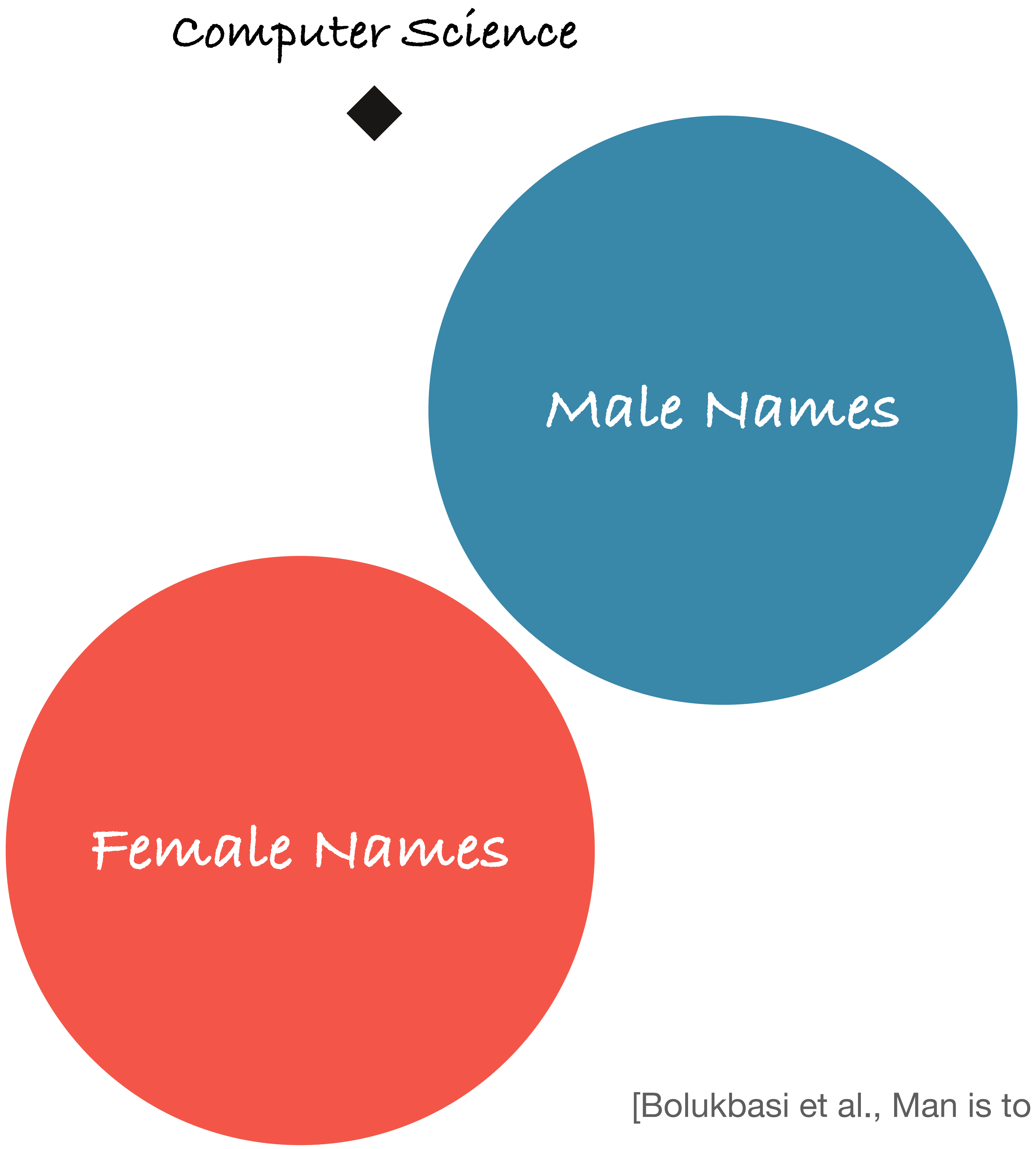
Motivating Examples

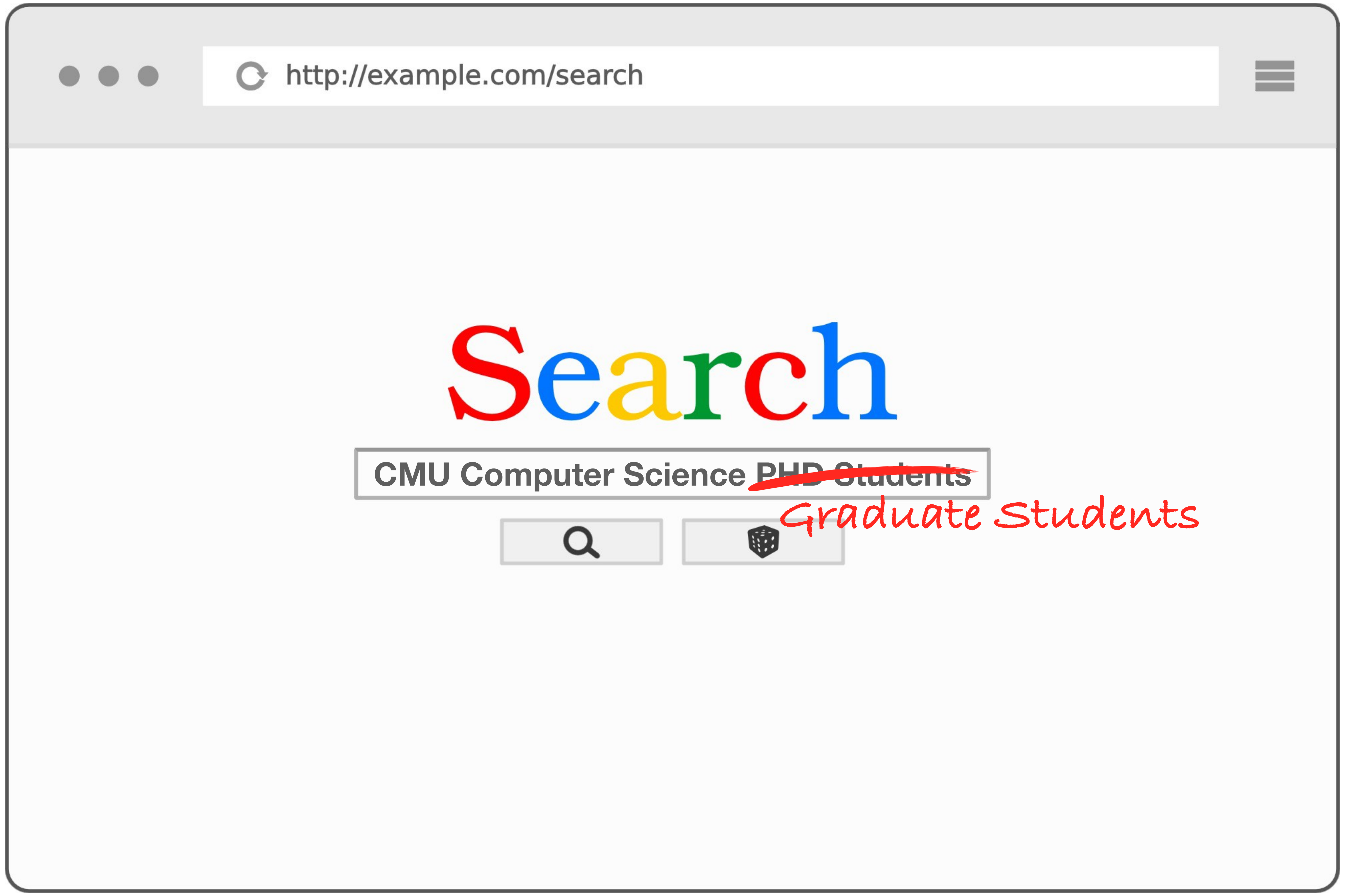


[Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? (NIPS 2016)]

Introduction

Motivating Examples

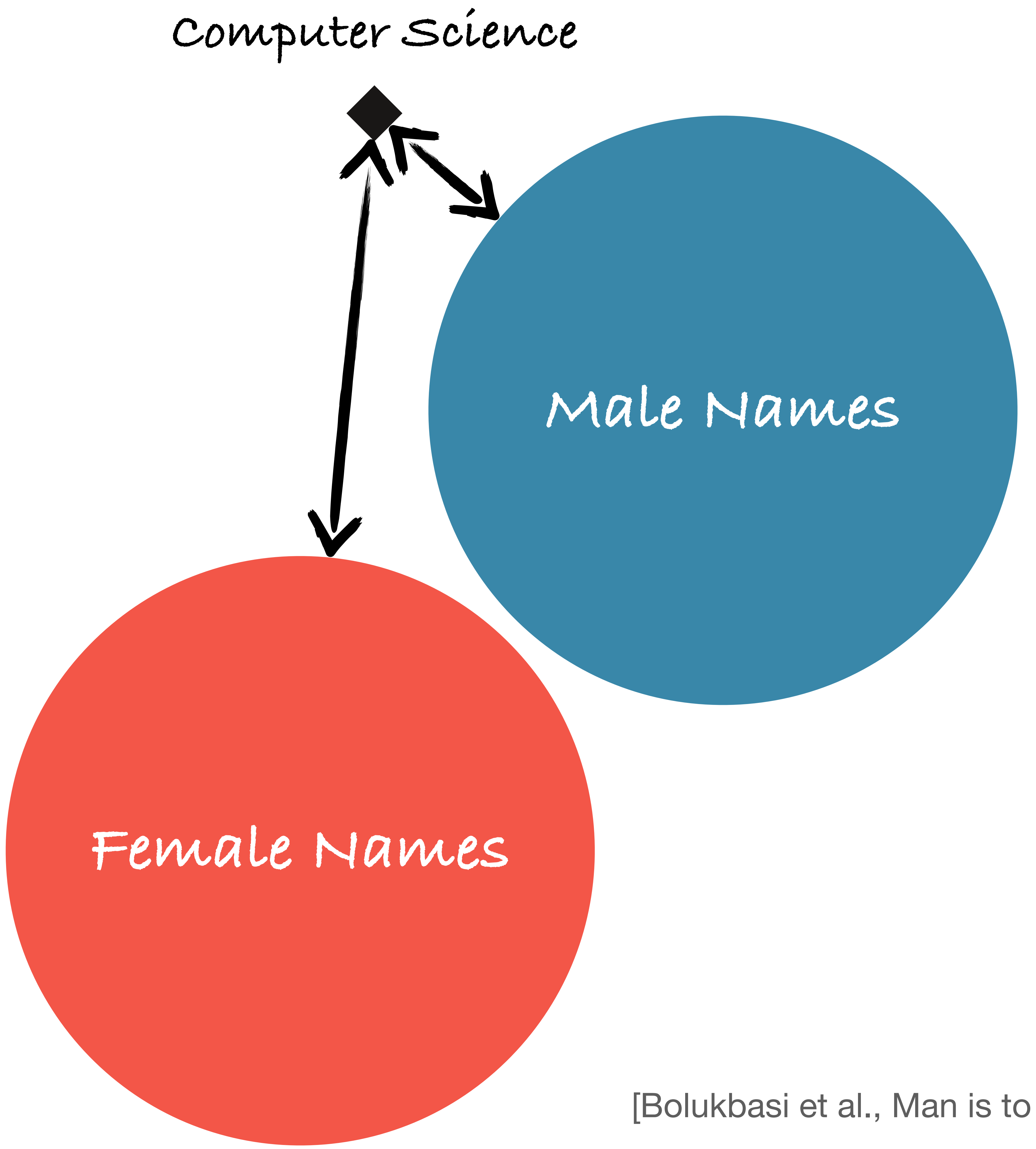


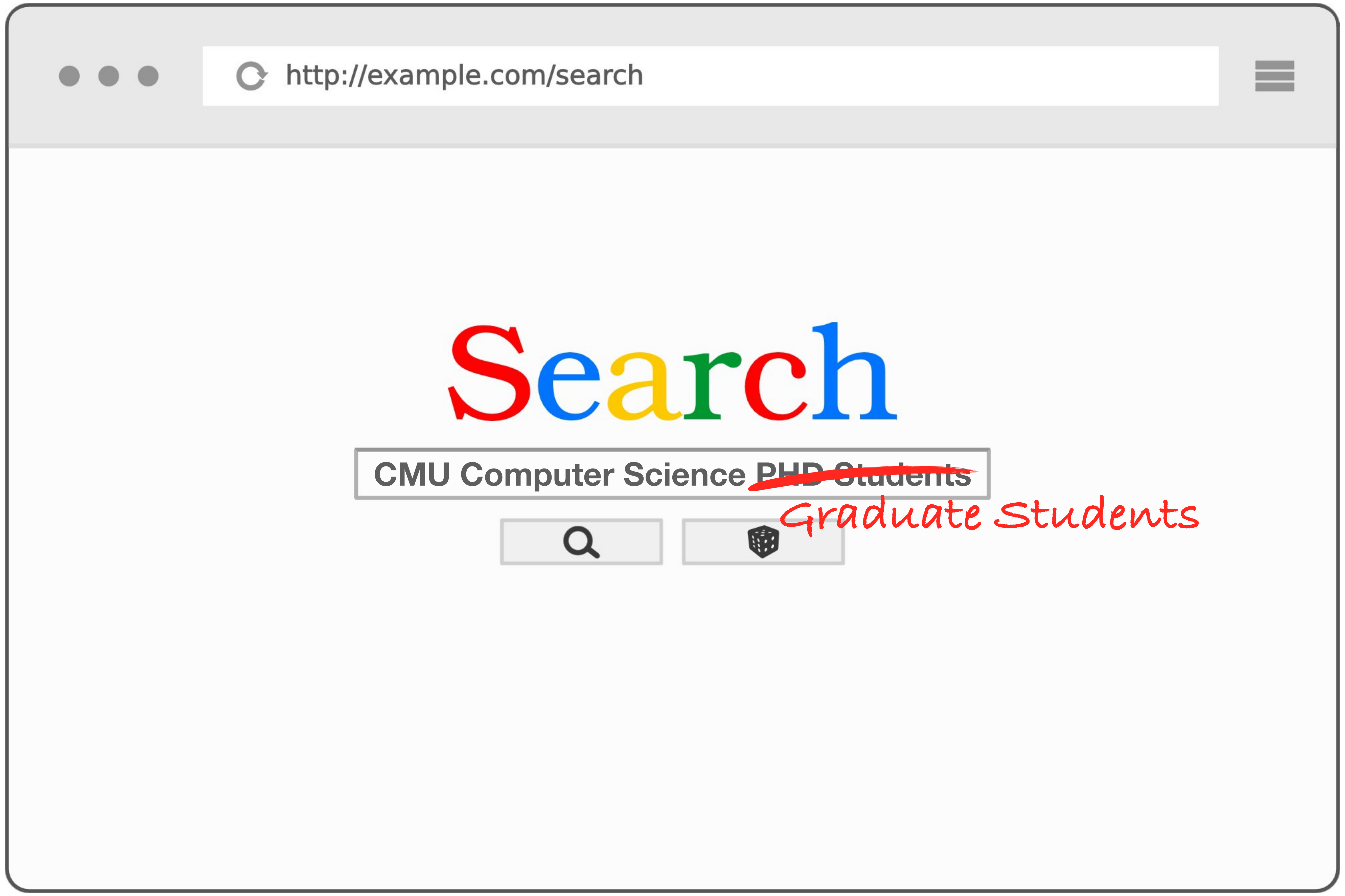


[Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? (NIPS 2016)]

Introduction

Motivating Examples





[Bolukbasi et al., Man is to Computer Programmer as Woman is to Homemaker? (NIPS 2016)]

Introduction

Motivating Examples

- **Search Engine Ranking Bias**
- **Resume Filtering**
- Recidivism Prediction Instrument
- Toxicity Detection

“Amazon.com's machine-learning specialists uncovered a big problem: their new recruiting engine did not like women”.

[Jeffrey Dastin, Amazon scraps secret AI recruiting tool that showed bias against women (REUTERS 2018)]

Introduction

Motivating Examples

- **Search Engine Ranking Bias**
- **Resume Filtering**
- **Recidivism Prediction Instrument**
- Toxicity Detection

“There’s software used across the country to predict future criminals. And it’s biased against blacks”.

[Julia Angwin, J.L.: Machine bias (ProPublica 2016)]

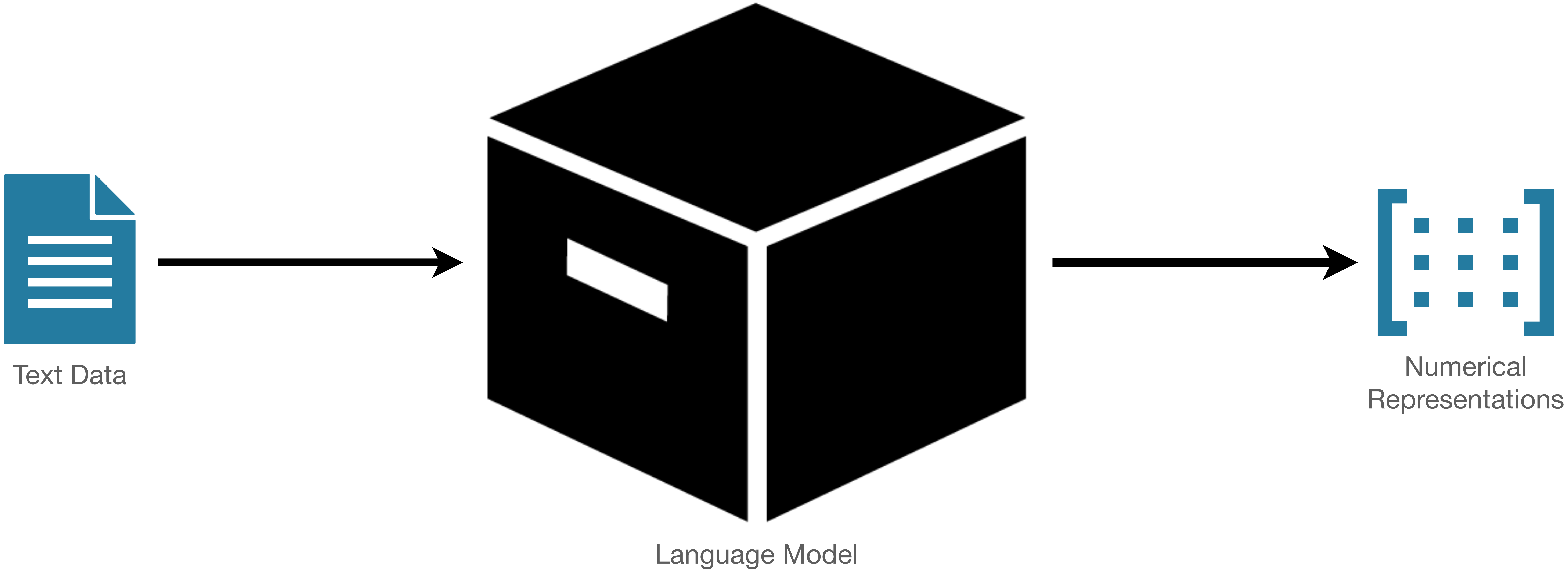
Introduction

Motivating Examples

- Search Engine Ranking Bias
- Resume Filtering
- Recidivism Prediction Instrument
- Toxicity Detection

Introduction

Language Model as a Black Box



Background

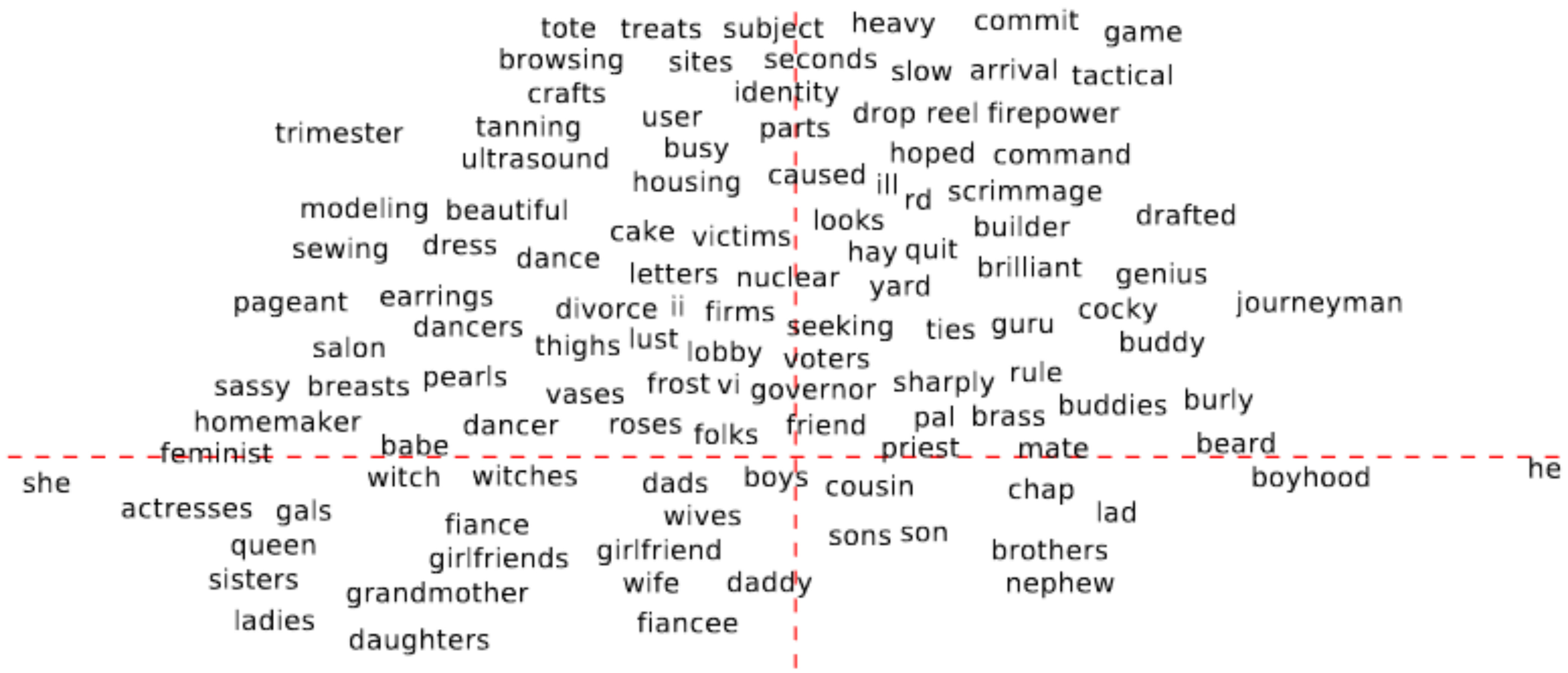
What is Fairness?

- Many Definitions
- “People receive that which they deserve” - Justice
- “impartial and just treatment or behavior without favoritism or discrimination” - Social Sciences
- “An action taken should have the same income whatever the protected variable” - Computer Science

Background

History of Fairness in NLP

- “Man is to Programmer as Woman is to Homemaker?



- Semantics Derived Automatically from Language Corpora Contain Human Like Biases

Background

History of Fairness in NLP

- Models Trained on Massive Data Showcase Gender Biases
- Recent Focus in Fairness in NLP
  - Measuring the Bias in Language Models
  - Mitigating the Bias in Language Models
  - Analysis of the Bias in Language Models (What are we actually measuring/reducing?)

Background

Measuring Gender Bias in NLP — Datasets

- Datasets With the Goal of Measuring Gender Bias
- CrowS-Pairs
- StereoSet
- Equity Evaluation Corpus
- Bias in Bios
- Wino bias
- GAP

Background

Measuring Gender Bias in NLP — General Metrics

- Equality of Opportunity $P(Y' = 1 \,|\, G = z, Y = 1) = P(Y = 1 \,|\, G = d, Y = 1), G \in \{z, d\}$
- Equality of Odds $P(Y' = c \,|\, G = z, Y = c) = P(Y = c \,|\, G = d, Y = c), G \in \{z, d\}$
- True Positive Rate Difference
- True Negative Rate Difference
- Accuracy Rate Difference
- Square Error Difference
- Minimum Description Length
    - How Much of the Gender Bias Can be Encoded in the Lowest Number of Bits? (Compression)

Background

Debiasing Methods — Extrinsic

- Counterfactual Augmentation
  - Add Counterfactual Sentences to the Dataset
- Fine-Tune Debiasing
  - Fine Tune the Model with Unbiased Instances
- Oversampling and Undersampling
  - Oversample or Undersample w.r.t to a Protected Variable or its Negation

Background

Debiasing Methods — Intrinsic

- Game Theoric Approach
  - Discriminator Strives to Output a Representation Without the Demographic Info
  - Generator Strives to Identify the Demographic Info in the Representation
- Constraining the Output
  - Define a Prediction Distribution Based on the Training Set
  - Constraint the Model Such that the Predictions Fall into that Distribution
- Projection Based Debiasing
  - Find the Dimensions of a Word Embedding In the Direction of Gendered Words
  - Remove/Equalize the Dimensions
- Loss Based Debiasing
  - Change the Loss Such that More Attention is Given to Female Instances (Higher Coefficient)

Research Projects

- MINORAGE: Measuring Language Model Reliability for Gender Equality (Under Review for *SEMEval 2022*)
- Analyzing the Impact of Debiasing Methods on Internal Representations of Language Models
- Debiasing Language Models without any Labeled Data?

MINORAGE

- Gender Bias Measurement Dataset
- 4000 Samples (3500 Test Set, 500 Development Set)
    - Classified Into Two Subgroups
        - Gender Specific Category
        - Gender Neutral Subgroup

MINORAGE

- All Instances Contain a Masked Gendered Pronoun or Noun
- Gender Specific: The Gender of the Subject/Object Can be Guessed
- Gender Neutral: The Gender of the Subject/Object Can't be Guessed

Class	Example
Gender Neutral	Since 2012 , [MASK] has been a full professor.
	[MASK] served as an assistant under four coaches.
	The way [MASK] skates is amazing
Gender Specific	The first son started living with [MASK] sister.
	Monroe wants Jimmy to investigate [MASK] younger wife.
	In order for to ensure a future for [MASK] children , Agnes had to remarry .

Table 1: Examples for each of the two classes.

MINORAGE

Metrics

- Two Metrics
- Gender Specific Score (Language Modeling)
- Gender Neutral Score (Fairness Score)
- Gender Invariance (Combination)

$$\text{GI} = 2 \cdot \frac{(1 - \text{GND}) \cdot \text{GSD}}{(1 - \text{GND}) + \text{GSD}}$$

$$\frac{\sum_{n=1}^{N} |\,\text{Top}(Male)_n - \text{Top}(Female)_n\,|}{N}$$

MINORAGE

Results

Model Name	GND	GSD	GI
BERT Base	0.516	0.649	0.554
BERT Large	0.560	**0.723**	0.547
RoBERTa Base	0.316	0.616	0.648
RoBERTa Large	0.275	0.672	0.698
XLNet Large	0.097	0.694	**0.785**
BERTweet Large	0.242	0.653	0.702
DistillBERT	**0.079**	0.331	0.487
IdealLM	0.000	1.000	1.000
BaselineLM (0.5)	0.500	0.500	0.500
Human Performance	0.038	0.961	0.961

Table 2: Gender Invariance (GI) results for different models and baselines.

Analyzing Debiasing Methods

Related Works — Insufficiency of Current Debiasing Models

- Previous Studies:

No gender bias

↓

Cannot determine gender association of a word by looking at its projection on gendered pair

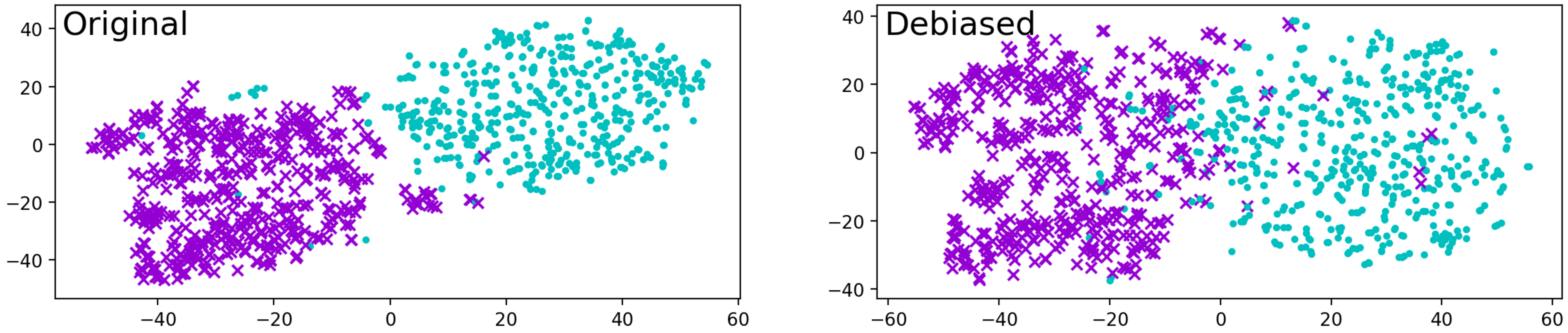
- "nurse" is no longer close to explicitly feminine words. Like: "she" and "mother"

[Gonen et al., Lipstick on a Pig (NAACL 2019)]

Analyzing Debiasing Methods

Related Works — Insufficiency of Current Debiasing Models

- But "nurse" is still close to socially-marked feminine words. Like: "receptionist"



[Gonen et al., Lipstick on a Pig (NAACL 2019)]

Analyzing Debiasing Methods

Related Works — Spurious Correlation in Data

- **Natural Language Inference (NLI):**
  - Two given sentences: Hypothesis and Premise
  - Classification Labels: Entailment, Contradiction, or Neutral

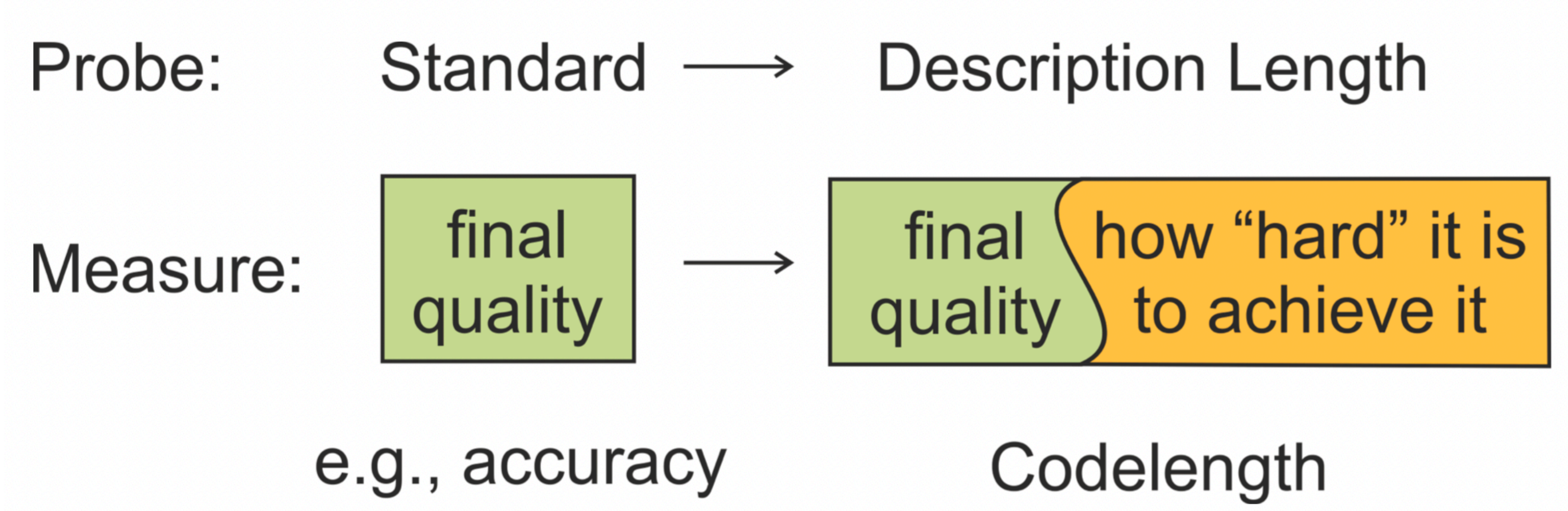
Premise	The boy is running through a grassy area.	
Hypothesis	The boy is in his room.	Contradiction
	A boy is running outside.	Entailment
	The boy is in a park.	Neutral

[Mendelson et al., Debiasing Methods in Natural Language Understanding Make Bias More Accessible (EMNLP 2021)]

Analyzing Debiasing Methods

Probing Method — Compression

- **Minimum Length Description (MDL) probing:** Quantify the effort needed to achieve a particular accuracy
- Training a probe to predict labels is recast as teaching it to effectively transmit the data



[Elena Voita et al., Information-Theoretic Probing with Minimum Description Length In Proceedings of the 2020 Conference on EMNLP]

Analyzing Debiasing Methods

Related Works — Spurious Correlation in Data

- Deep neural models are prone to shortcut learning
- **Spurious Correlation in NLI Datasets:**
  - High $P(\text{contradiction} \mid w)$ where w is universal negation words, like: nobody, alone, no, empty
  - Lexical overlap between the premise and hypothesis → High correlation with entailment label

[Mendelson et al., Debiasing Methods in Natural Language Understanding Make Bias More Accessible (EMNLP 2021)]

Analyzing Debiasing Methods

Related Works — Spurious Correlation in Data

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Contradiction	A woman is **not** taking money for any of her sticks.

[Mendelson et al., Debiasing Methods in Natural Language Understanding Make Bias More Accessible (EMNLP 2021)]

Analyzing Debiasing Methods

Related Works — Spurious Correlation in Data

The more language model is pushed towards a debiased regime

↓

The more bias encoded in its inner representations

[Mendelson et al., Debiasing Methods in Natural Language Understanding Make Bias More Accessible (EMNLP 2021)]

Q: How does bias mitigation methods impact language models internal representations?

Analyzing Debiasing Methods

Preliminary Results

Name (6 visualized)		probe_type	compression	eval_accuracy	eval_loss	loss	online_cdl
gab/2022-04-08-13:51:09	lanations/runs/majority_gab_es_reg_nb5_h5_is_bal_pos_seed_0	mlp	1.88	0.8307	1052.788	852.879	2609.091
gab/2022-04-08-13:48:36	lanations/runs/majority_gab_es_vanilla_bal_seed_0	mlp	2.19	0.7808	1387.712	705.033	2237.806
gab/2022-04-08-13:45:48		mlp	1.76	0.616	1898.258	847.519	2779.027

Analyzing Debiasing Methods

Layer-wise Probing

- Data is One of the Main Sources of Bias in Language Models
- Bias Comes from Data
  - Can be Interpreted as a Language Model
  - Previous Work has Shown the Effectiveness of Layer-wise Probing
  - Showcases the Linguistic Property Stored in Each Layer

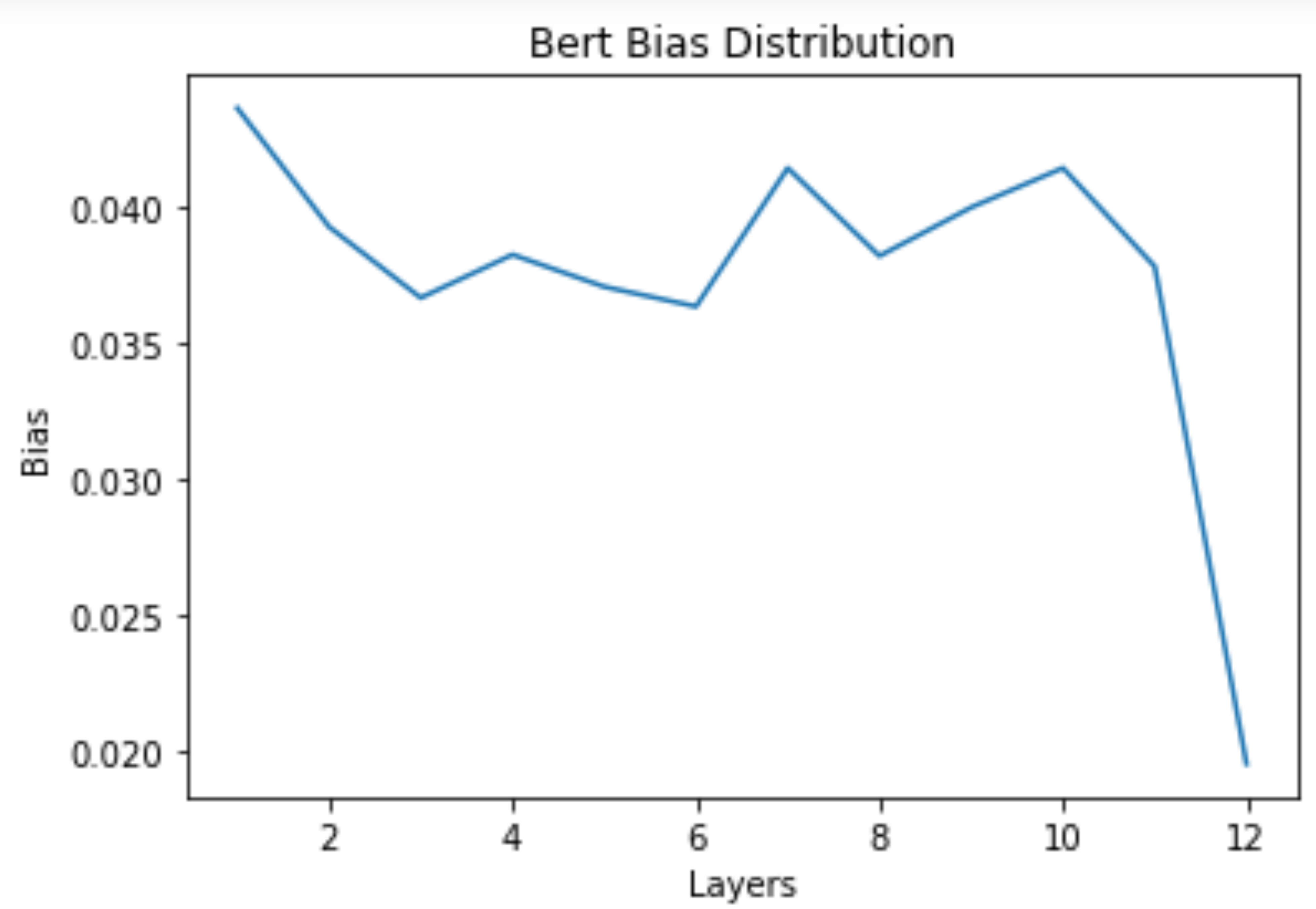
Analyzing Debiasing Methods

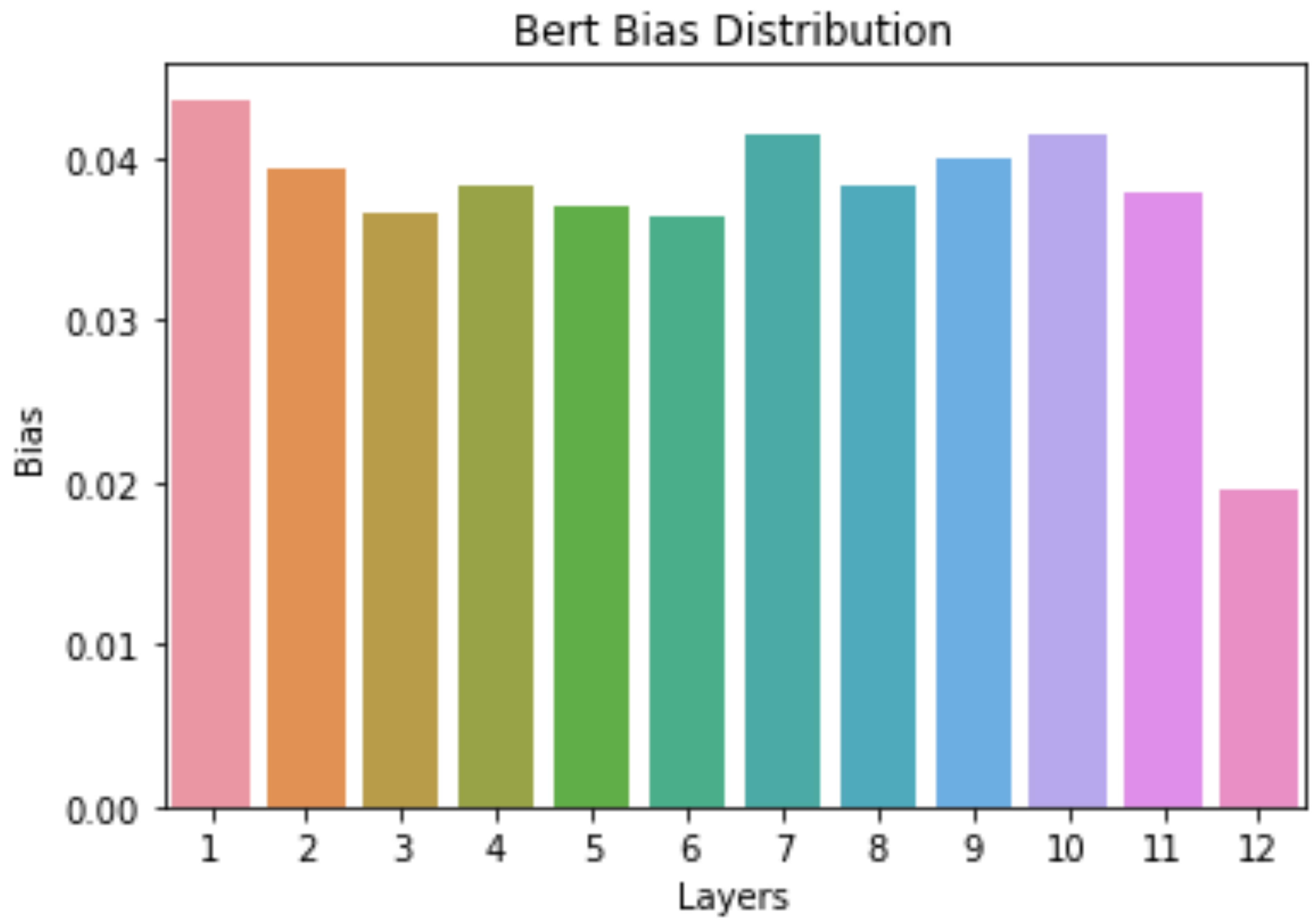
Layer-wise Probing

- Data Should be Simple Enough to be Learnable by a Simple Classifier (or use MDL)
- Equity Evaluation Corpus
- Compute the Gender Bias for Each Layer Separately
- Each Layer's Contribution to Bias is Computed by Subtracting the Bias from that Layer from the Bias from the Previous Layer
  - Positive Value: Layer Increases the Bias
  - Negative Value: Layer Decreases the Bias

Analyzing Debiasing Methods

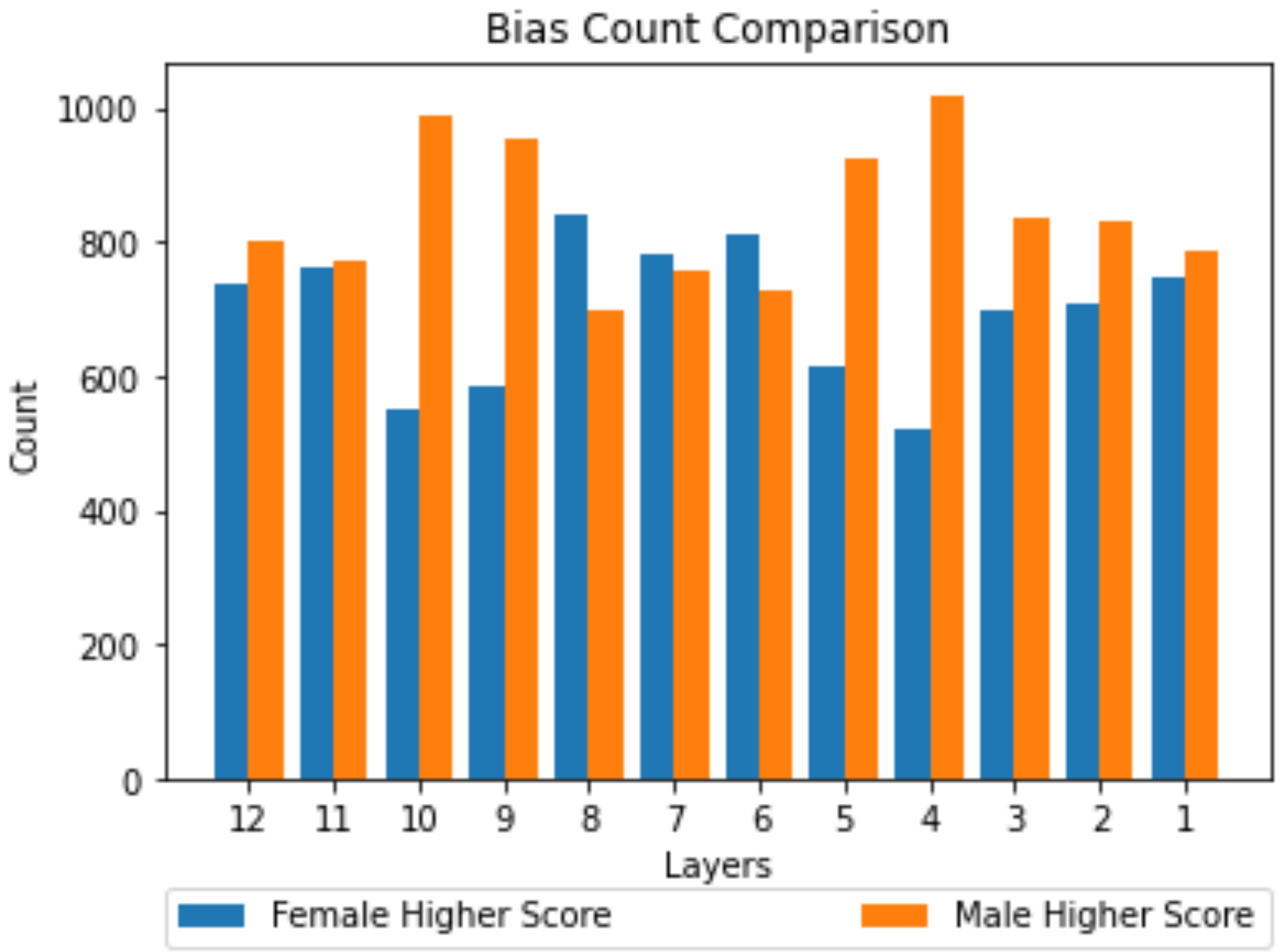
Layer-wise Probing Results-BERT





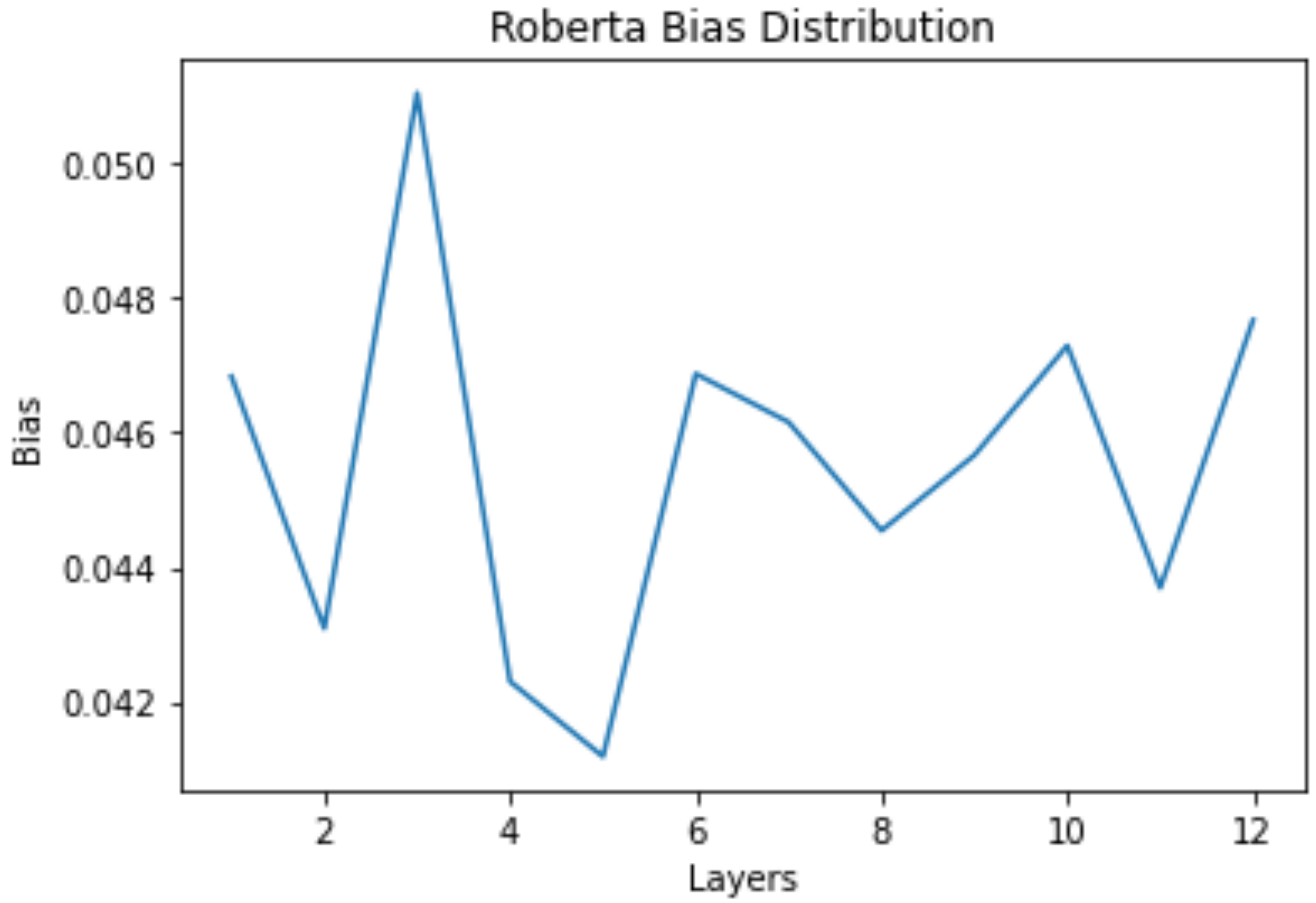
Analyzing Debiasing Methods

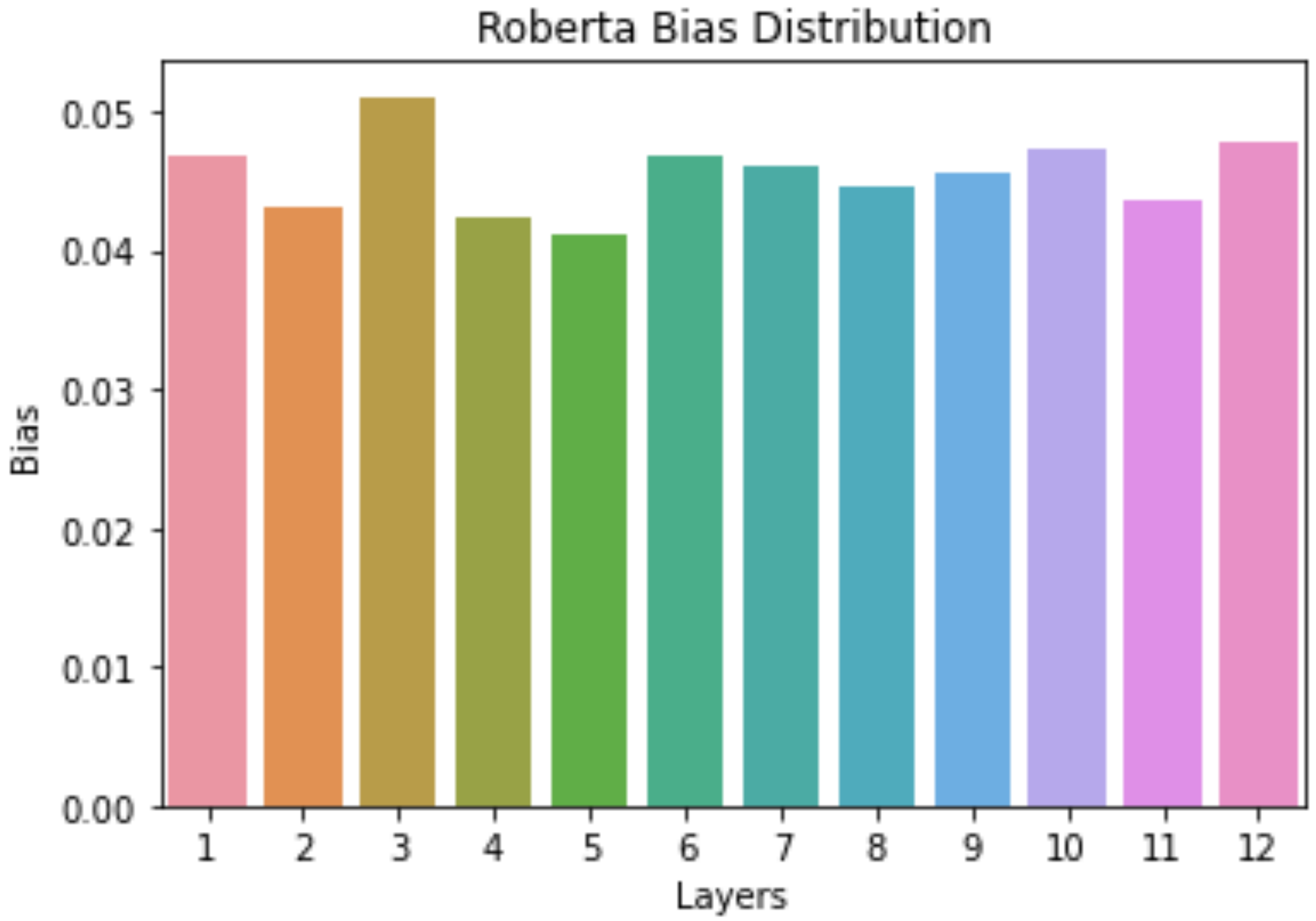
Layer-wise Probing Results-BERT



Analyzing Debiasing Methods

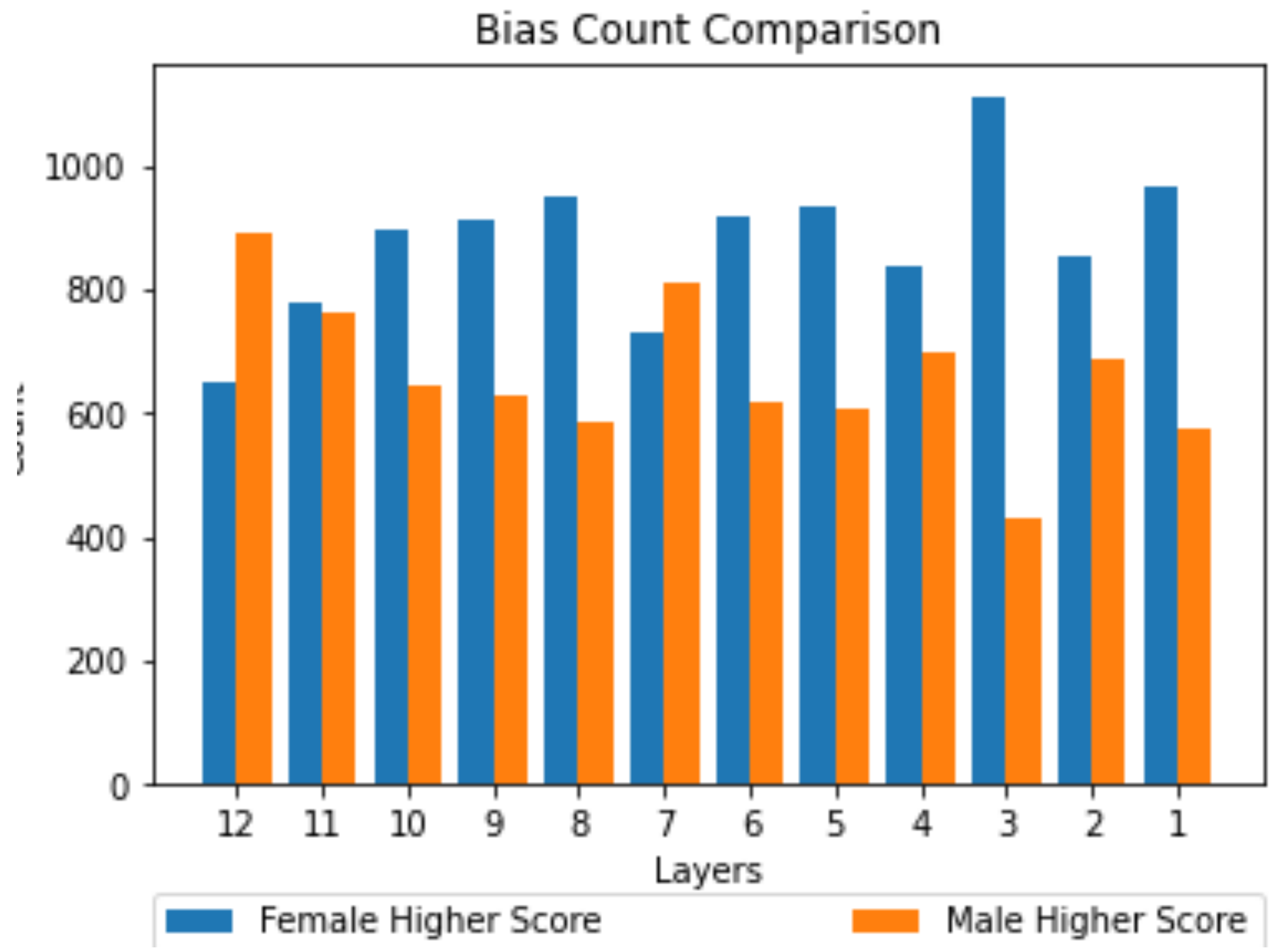
Layer-wise Probing Results-RoBERTA





Analyzing Debiasing Methods

Layer-wise Probing Results-RoBERTA



Analyzing Debiasing Methods

Method Selection and Metric Correlation

- Not All Bias Evaluation Metrics Conform with Each Other Across Tasks
- Fraction of Two Evaluations may Flip Across Two Tasks
- [0.47,0.86] —> [0.62, 0.33]
- Challenge: How to Choose the Best Debiasing Method?

Analyzing Debiasing Methods

Method Selection and Metric Correlation

- Debiasing Method Selection Hypothesis 1
    - Choose the Debiasing Method that Best Optimizes the Average of Metrics
    - Debiasing Method 1
        - [0.46,0.86] —> [0.39, 0.66]
        - [0.62, 0.33] —> [0.48, 0.28]
    - Debiasing Method 2
        - [0.46,0.86] —> [0.37,0.71]
        - [0.62, 0.33] —> [0.53, 0.22]

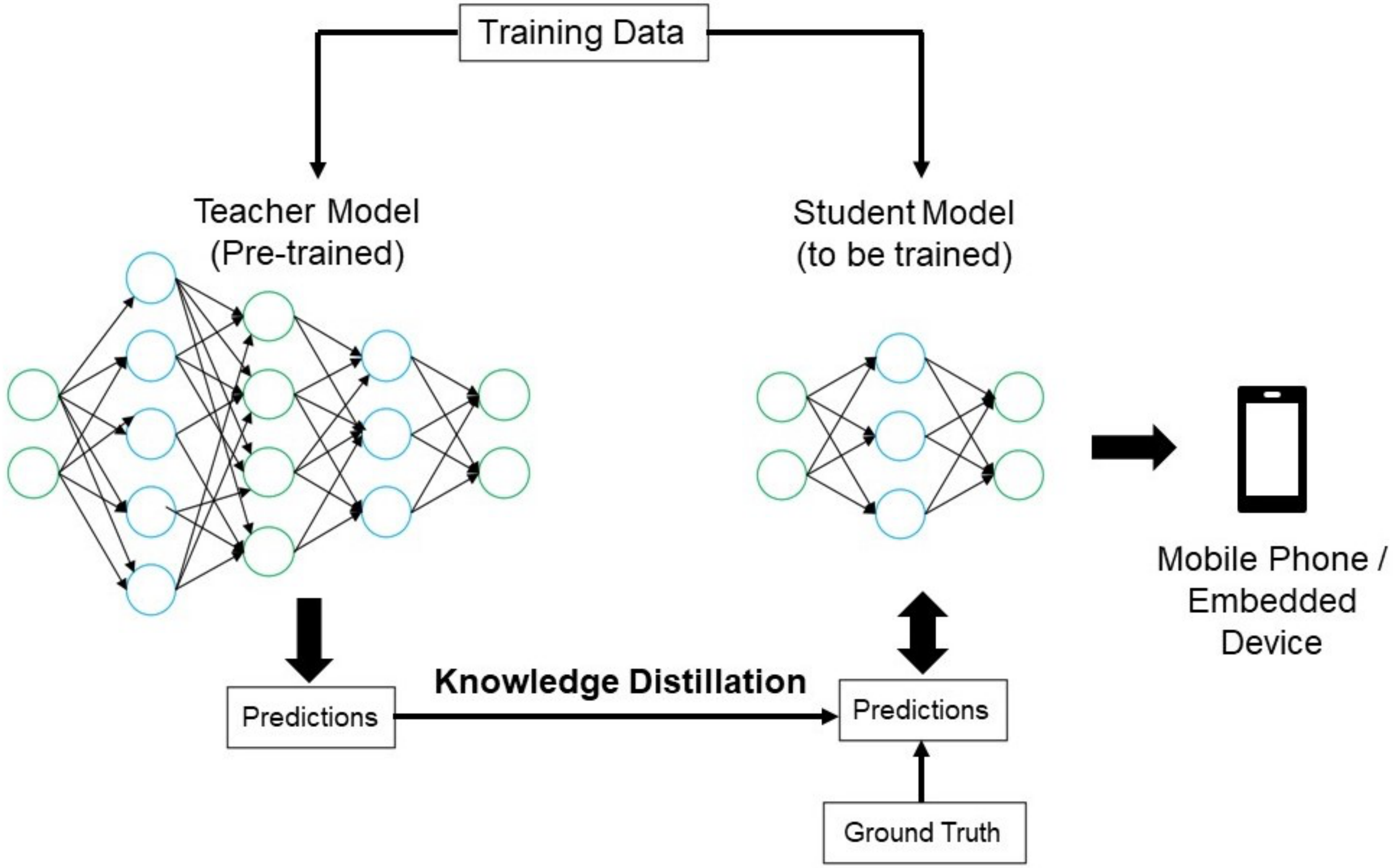
Analyzing Debiasing Methods

Method Selection and Metric Correlation

- Debiasing Method Selection Hypothesis 2
    - Define an Optimal Vector [0,0,…,0] Indicating the Best Case for Each Metric
    - Compute the Average Distance of Each Debiased Model Across Tasks
    - The Debiasing Method with the Lowest Average Distance Wins

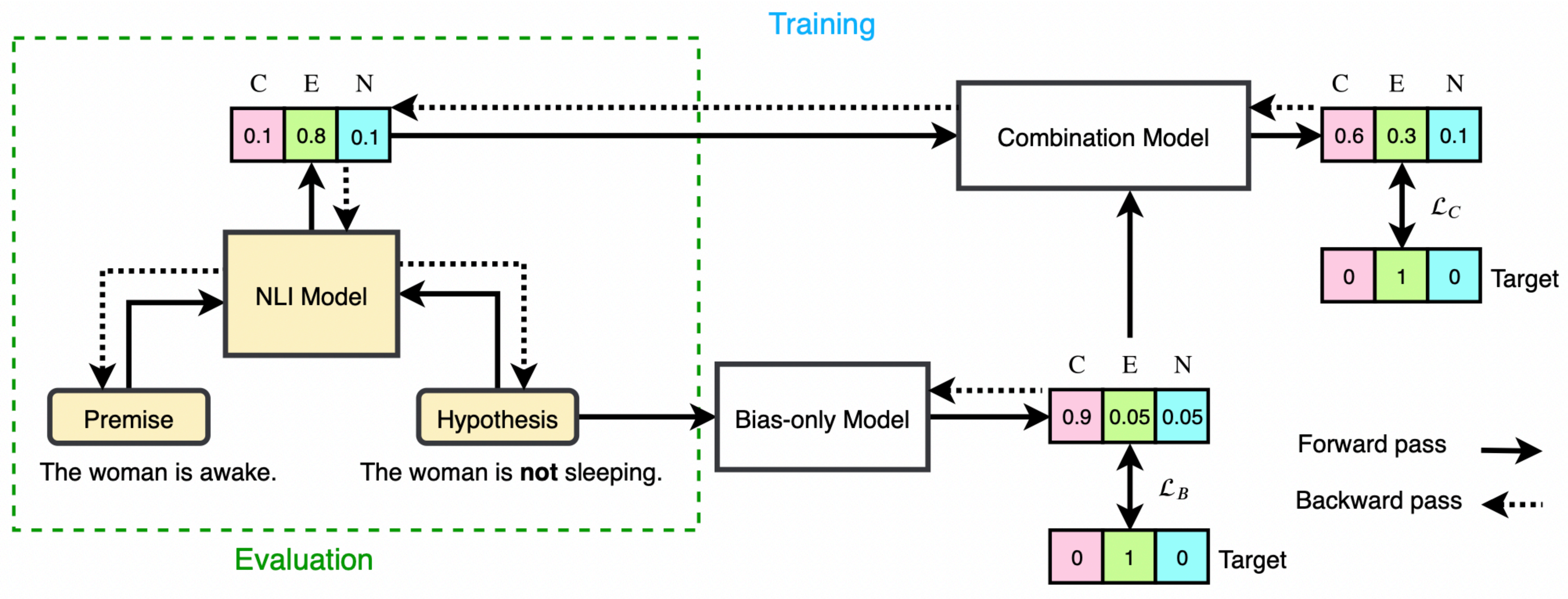
Teacher-Student Debiasing

Related Works — Knowledge Distillation



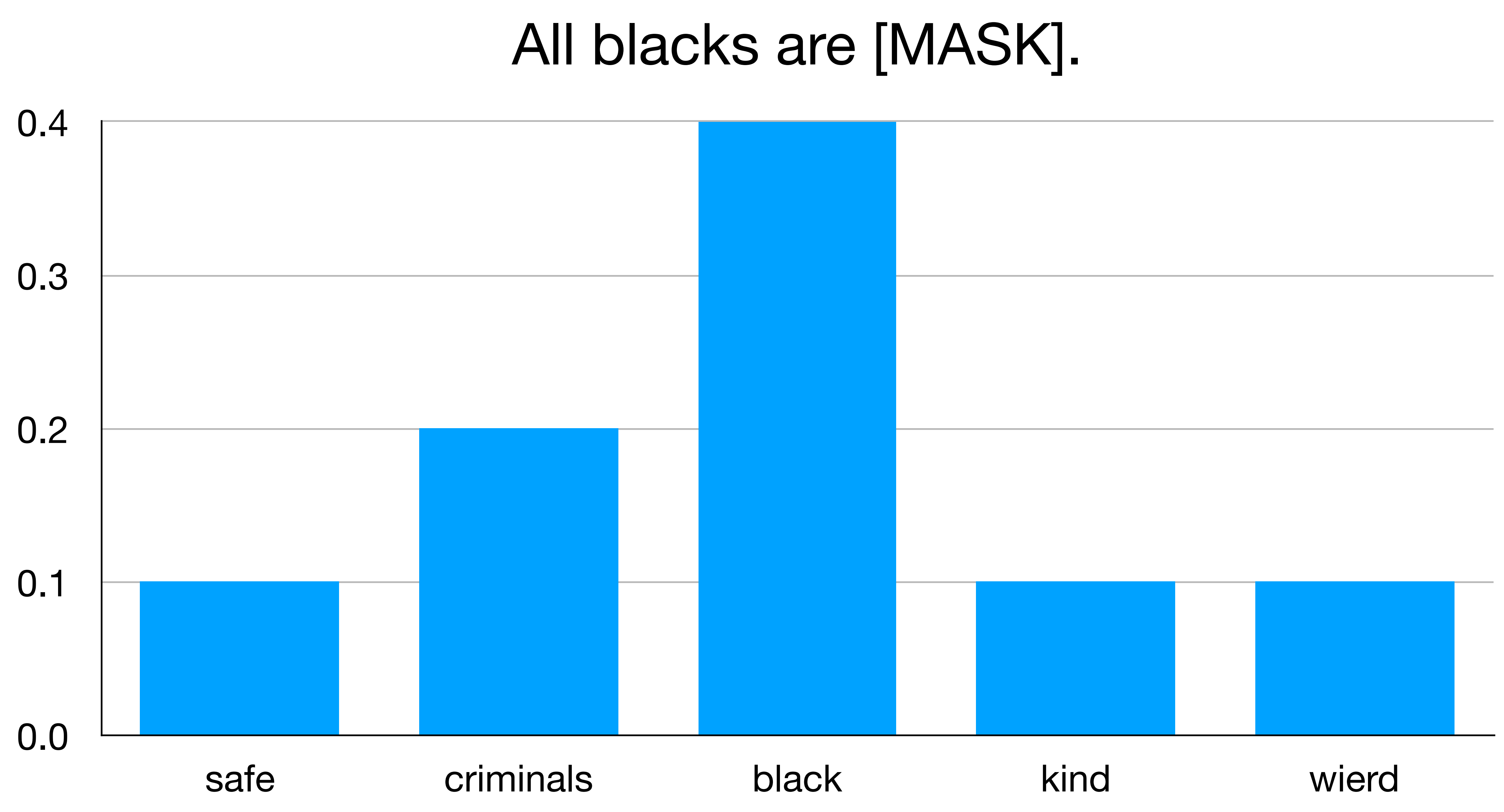
Teacher-Student Debiasing

Related Works — Confronting Spurious Correlation in Datasets



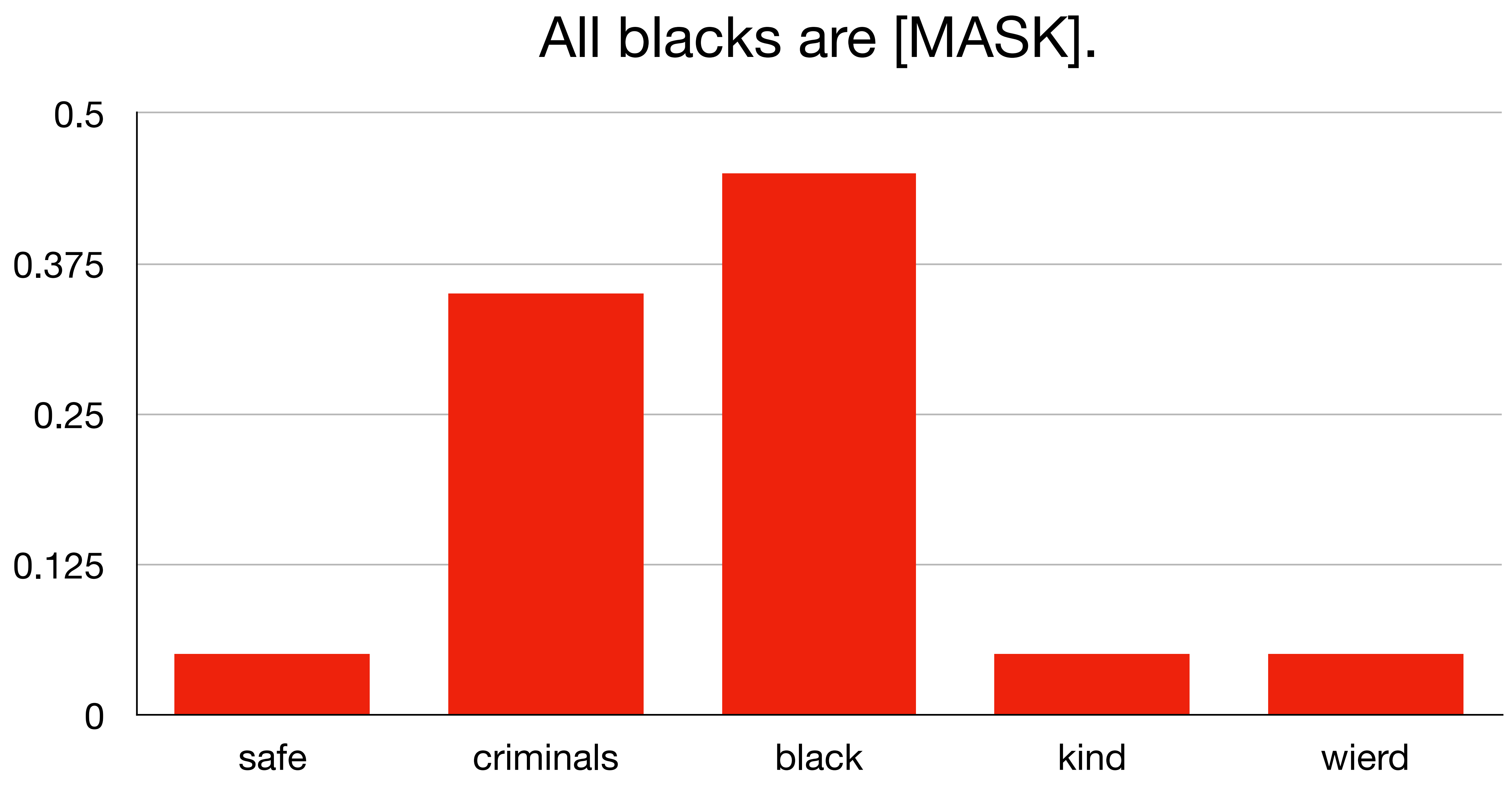
Teacher-Student Debiasing

Related Works — Confronting Spurious Correlation in Datasets



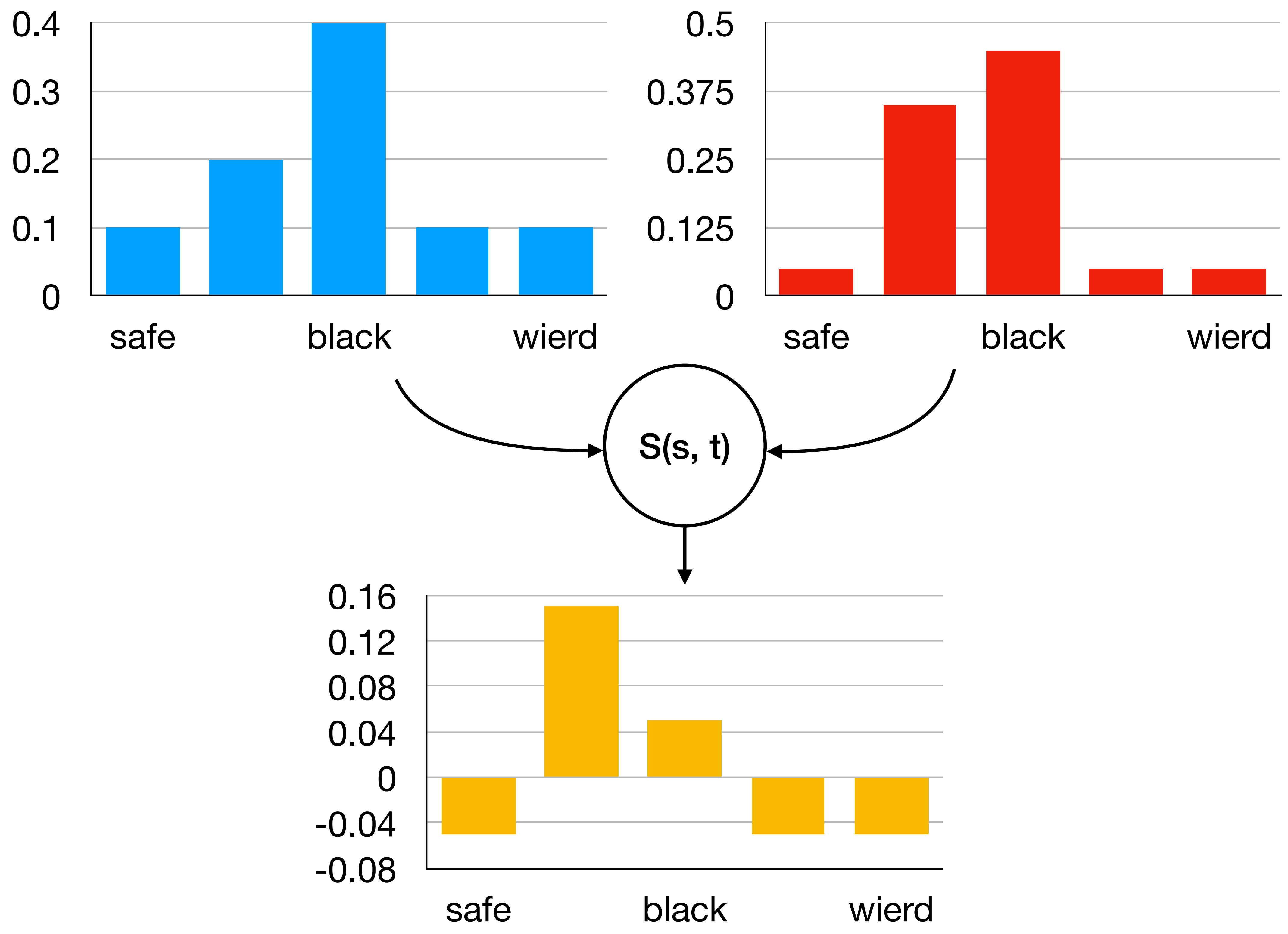
Teacher-Student Debiasing

Related Works — Confronting Spurious Correlation in Datasets



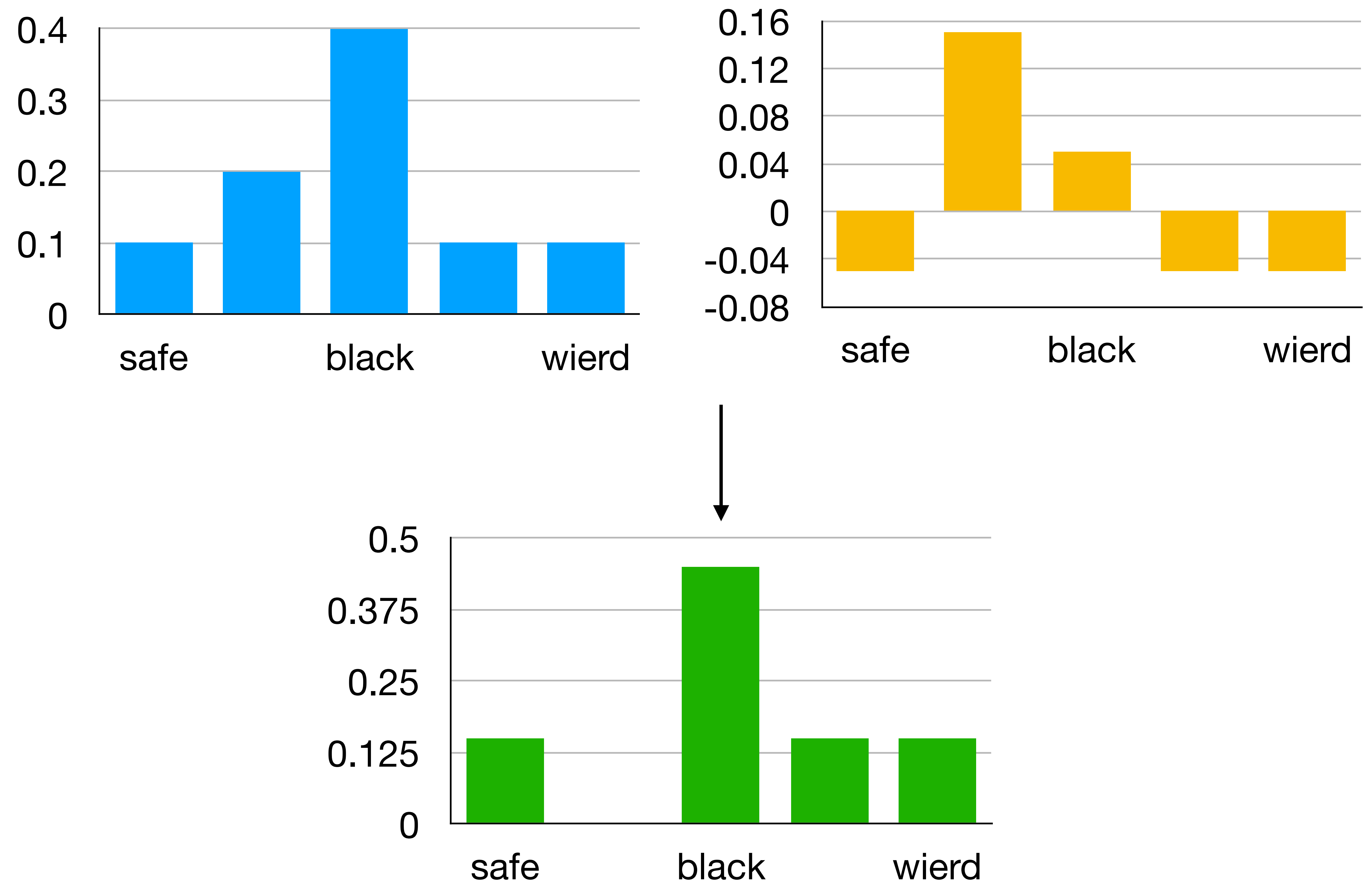
Teacher-Student Debiasing

Related Works — Confronting Spurious Correlation in Datasets



Teacher-Student Debiasing

Related Works — Confronting Spurious Correlation in Datasets



Wrap Up

- Bias and Fairness is a new and active field of research in NLP
- Research topics in bias and fairness field are still incomplete:
  - Evaluation Metric
  - Debiasing Method
  - Analyzing Debiasing Impacts
- There are many areas that have not yet been explored in this field:
  - Multilingual Debiasing
  - Multidimensional Debiasing (Gender, Ethnics, Religions, etc.)