UNC NLP

Learning Fair Representations via Rate-Distortion Maximization

Somnath Basu Roy Chowdhury and Snigdha Chaturvedi

Transactions of the Association for Computational Linguistics (TACL), 2022

Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Why do need Fair Representations?

- Pre-trained representations are used ubiquitously in NLP applications

Why do need Fair Representations?

- Pre-trained representations are used ubiquitously in NLP applications
- Representations are retrieved from a model trained in a self-supervised manner

Why do need Fair Representations?

- Pre-trained representations are used ubiquitously in NLP applications
- Representations are retrieved from a model trained in a self-supervised manner
- Developer does not have control over the pre-training corpus

Why do need Fair Representations?

- Pre-trained representations are used ubiquitously in NLP applications
- Representations are retrieved from a model trained in a self-supervised manner
- Developer does not have control over the pre-training corpus
- Different forms of bias or sensitive information can percolate into downstream task

Examples of Failure mode

Biased translation in Google Translate

Examples of Failure mode

Biased translation in Google Translate

Gender Bias in automated resume screening tool at Amazon

What are Fair Representations?

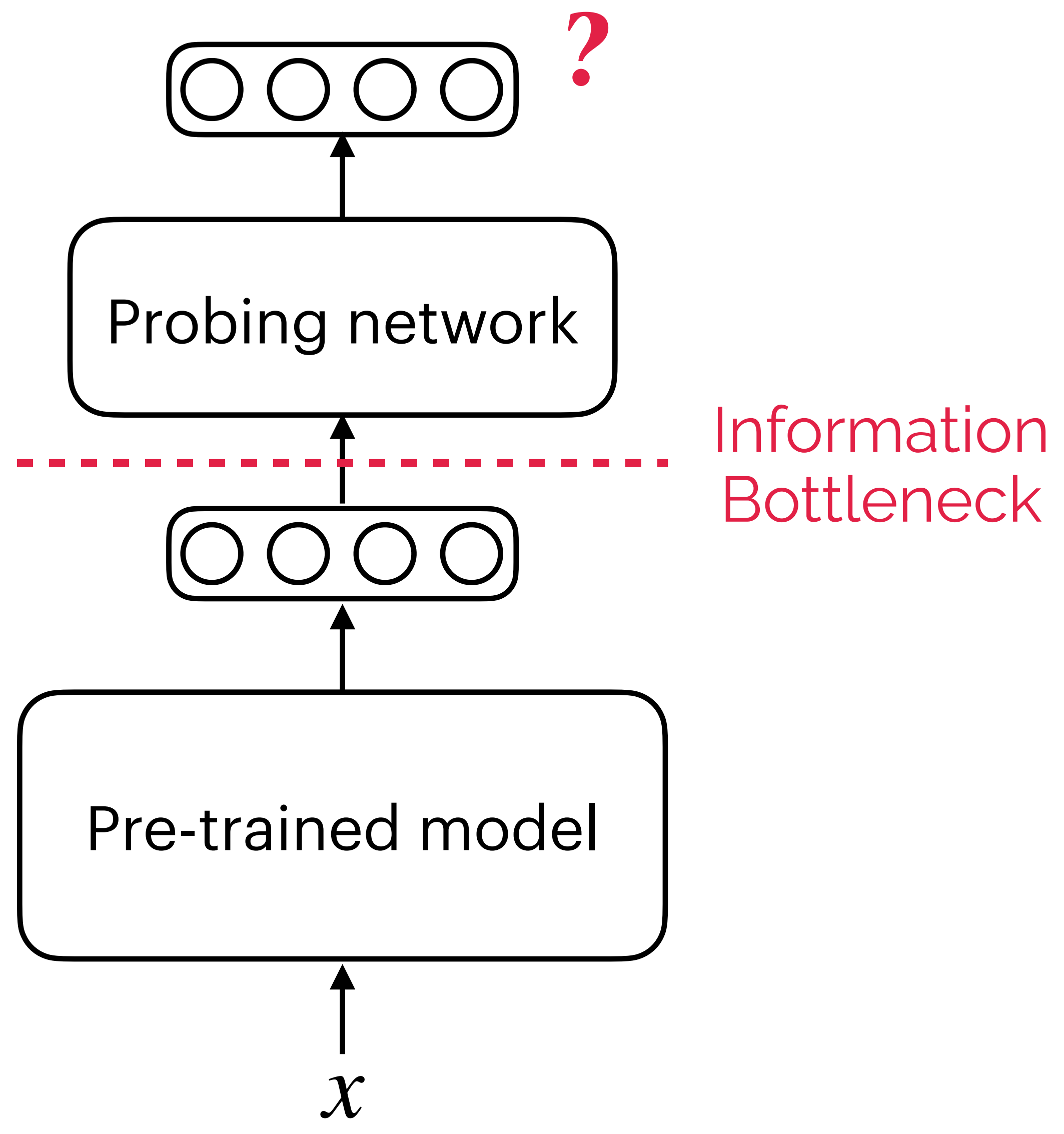
- Representations do not reveal information about private or sensitive attribute

What are Fair Representations?

- Representations do not reveal information about private or sensitive attribute
- Achieve group fairness — representations from different demographic groups look alike

What are Fair Representations?

- Representations do not reveal information about private or sensitive attribute
- Achieve group fairness — representations from different demographic groups look alike
- Once debiased, information cannot be extracted by a subsequent network



Fairness Goals

- Achieve Demographic Parity — representations from different demographic groups receive similar outcomes

$$|P(+\,|\,\text{male}) - P(+\,|\,\text{female})| \approx 0$$

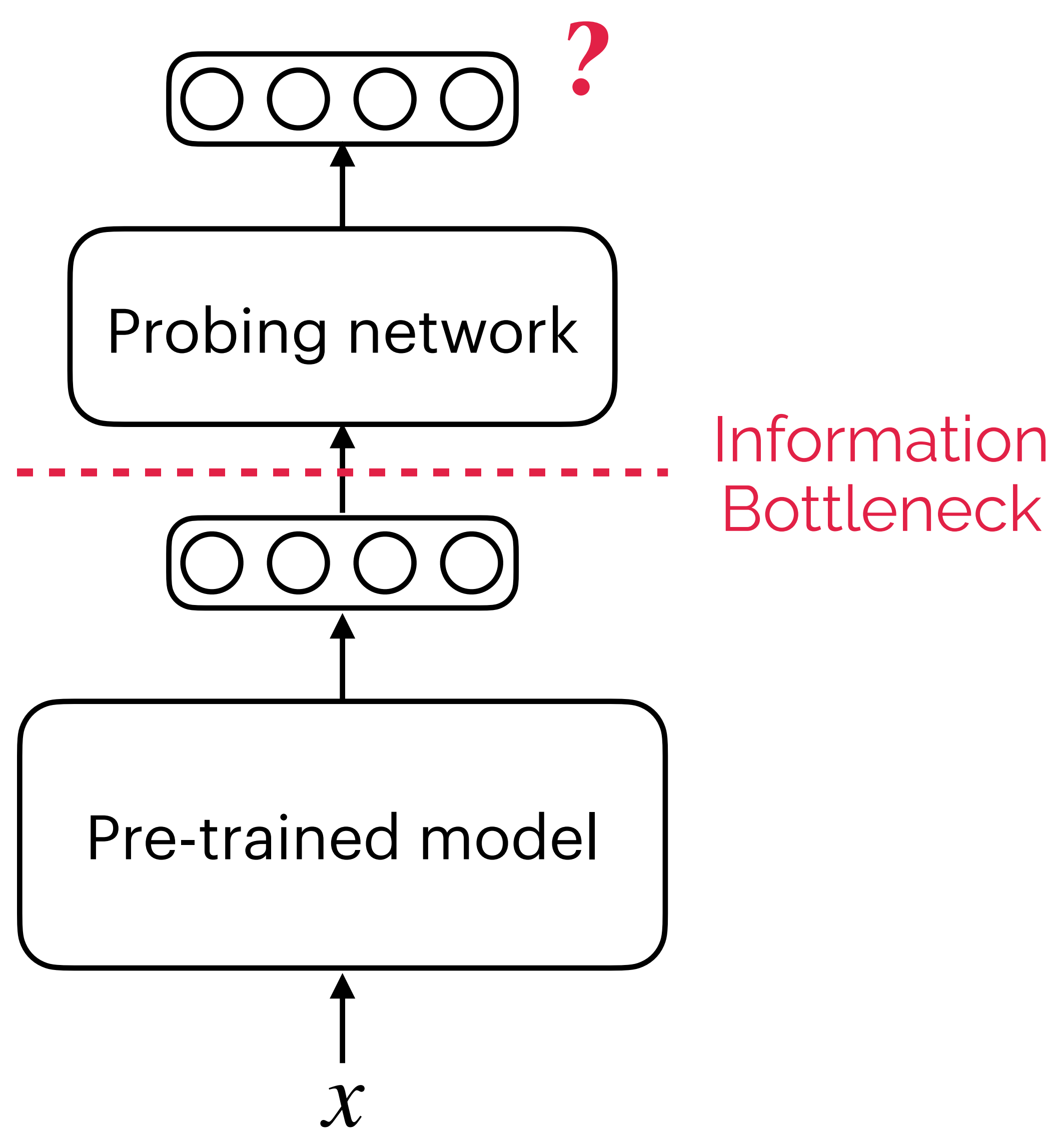
Fairness Goals

- Achieve Demographic Parity — representations from different demographic groups receive similar outcomes

$$|P(+\,|\,\text{male}) - P(+\,|\,\text{female})| \approx 0$$

- Translating this to representation learning terms, given a probing network f

$$|P(f(x) = \text{male}) - P(f(x) = \text{female})| \approx 0$$



Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Formalizing the problem

- Given a set of representations $Z = \{z_1, z_2, \ldots\}$

Formalizing the problem

- Given a set of representations $Z = \{z_1, z_2, \ldots\}$
- Each representation is associated with a protected attribute $A = \{a_1, a_2, \ldots\}$

Formalizing the problem

- Given a set of representations $Z = \{z_1, z_2, \ldots\}$
- Each representation is associated with a protected attribute $A = \{a_1, a_2, \ldots\}$
- a_i is a categorical variable, $a_i \in \{0, \ldots, k\}$

Formalizing the problem

- Given a set of representations $Z = \{z_1, z_2, \ldots\}$
- Each representation is associated with a protected attribute $A = \{a_1, a_2, \ldots\}$
- a_i is a categorical variable, $a_i \in \{0,\ldots,k\}$
- Assume there existence of an optimal adversary $f(\cdot)$ for prediction a_i

Formalizing the problem

- Given a set of representations $Z = \{z_1, z_2, \dots\}$
- Each representation is associated with a protected attribute $A = \{a_1, a_2, \dots\}$
- a_i is a categorical variable, $a_i \in \{0,\dots,k\}$
- Assume there existence of an optimal adversary $f(\cdot)$ for prediction a_i
- Our goal: $|P(f(z) = a_i) - P(f(z) = a_j)| \approx 0, \forall (i, j)$

Problem Setup

Perform debiasing in two different setups:

Problem Setup

Perform debiasing in two different setups:

- Unconstrained debiasing
  - Input - representation set Z, protected attribute A
  - Goal - debias Z from A, while retaining all other information

Problem Setup

Perform debiasing in two different setups:

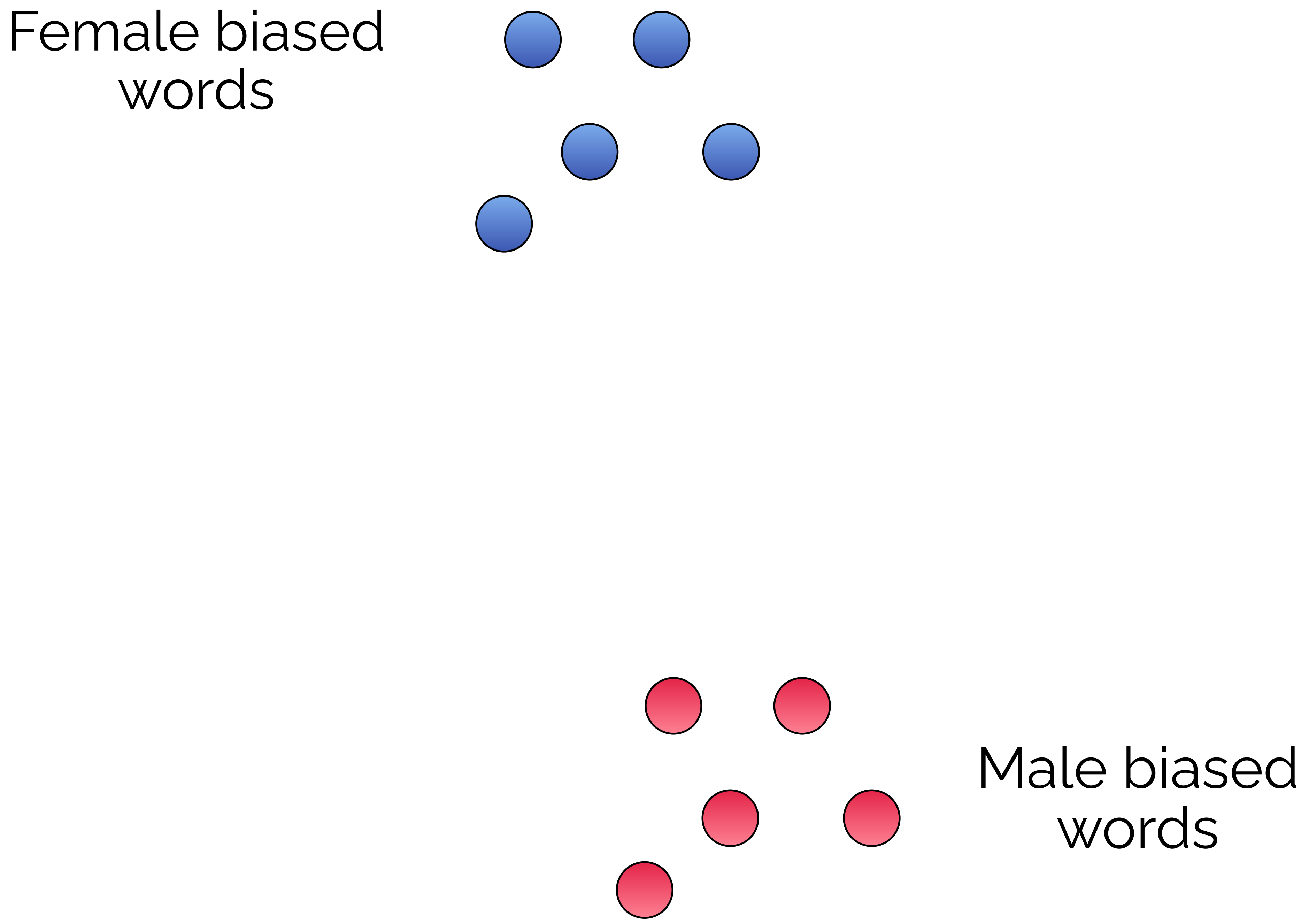
- Unconstrained debiasing
  - Input - representation set Z, protected attribute A
  - Goal - debias Z from A, while retaining all other information

- Constrained debiasing
  - Input - representation set Z, protected attribute A, target attribute Y
  - Goal - debias Z from A, while exclusively retaining information about Y

Outline

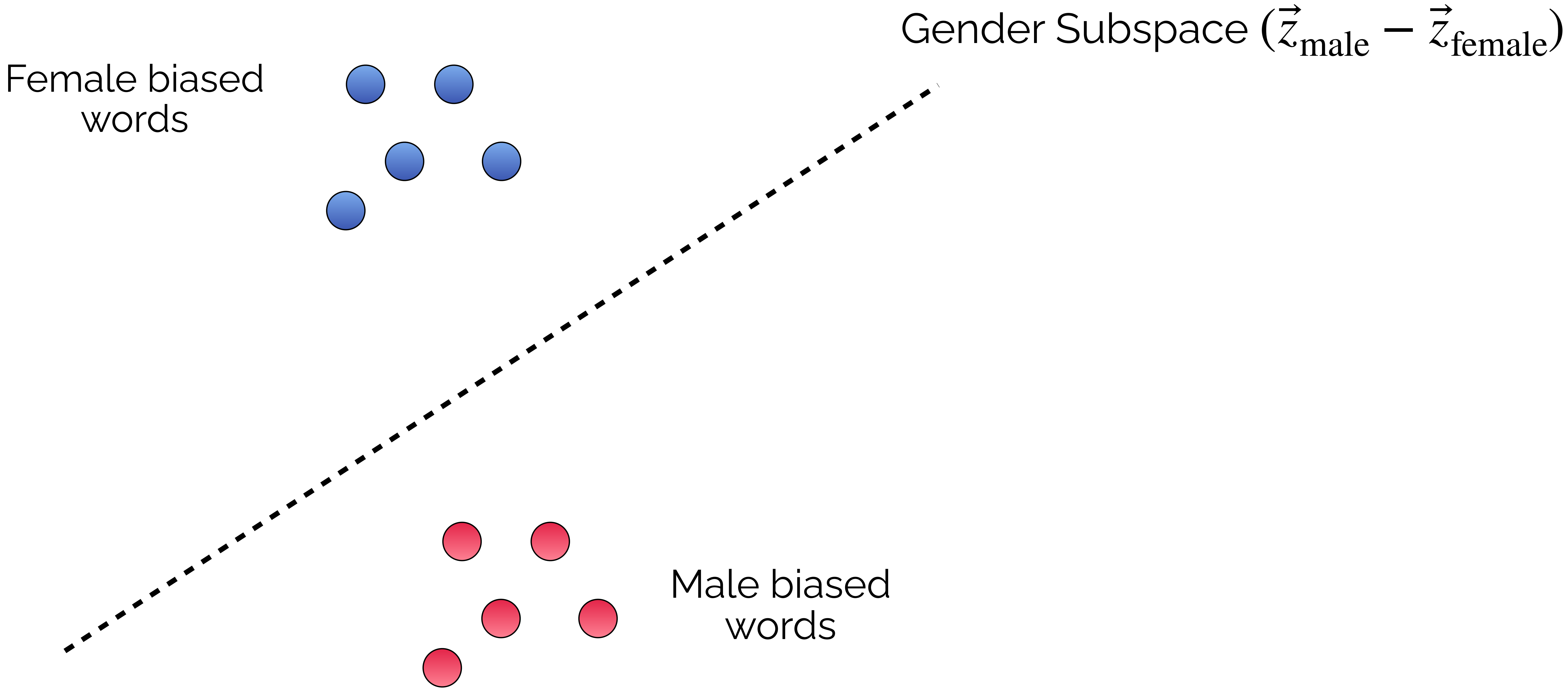
- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Prior Work - Unconstrained debiasing



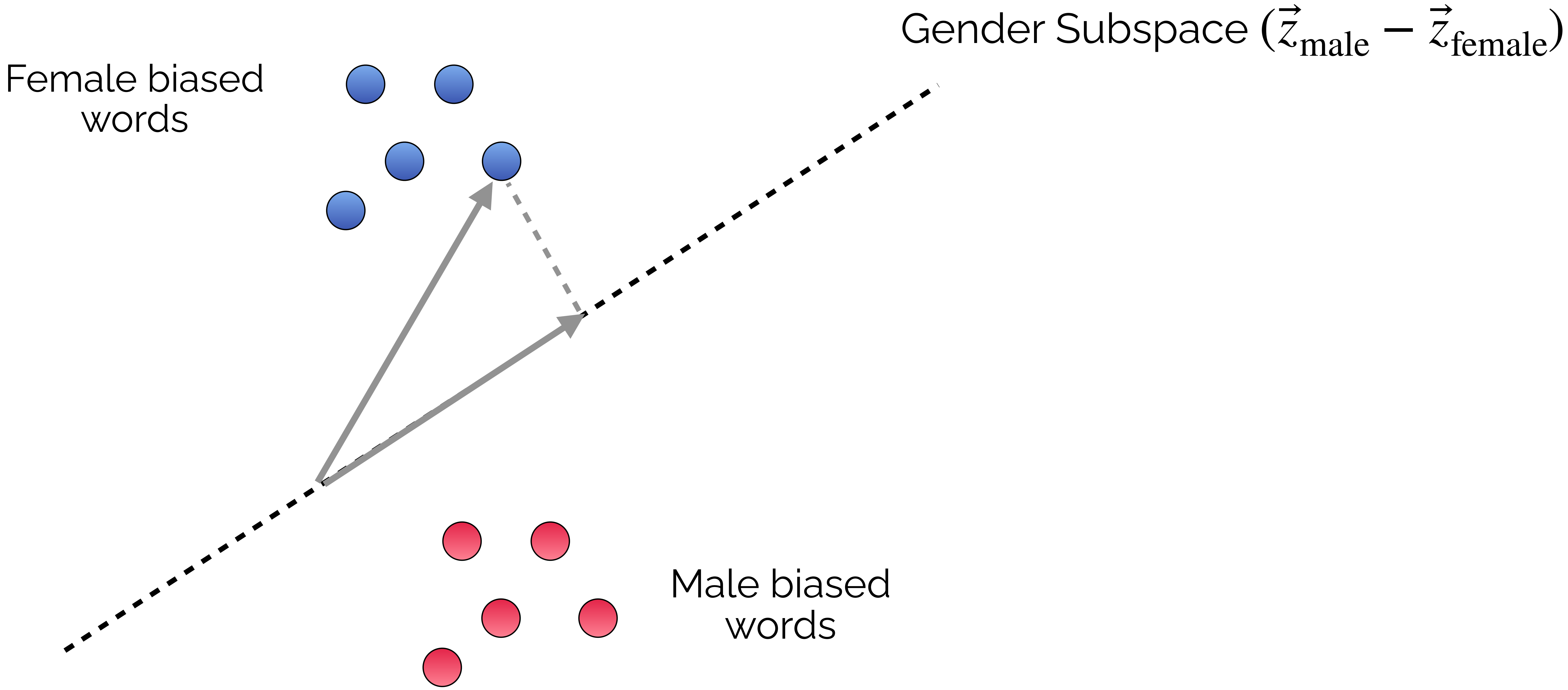
Debiasing Word Embeddings (Bolukbasi et al, 2016)

Prior Work - Unconstrained debiasing



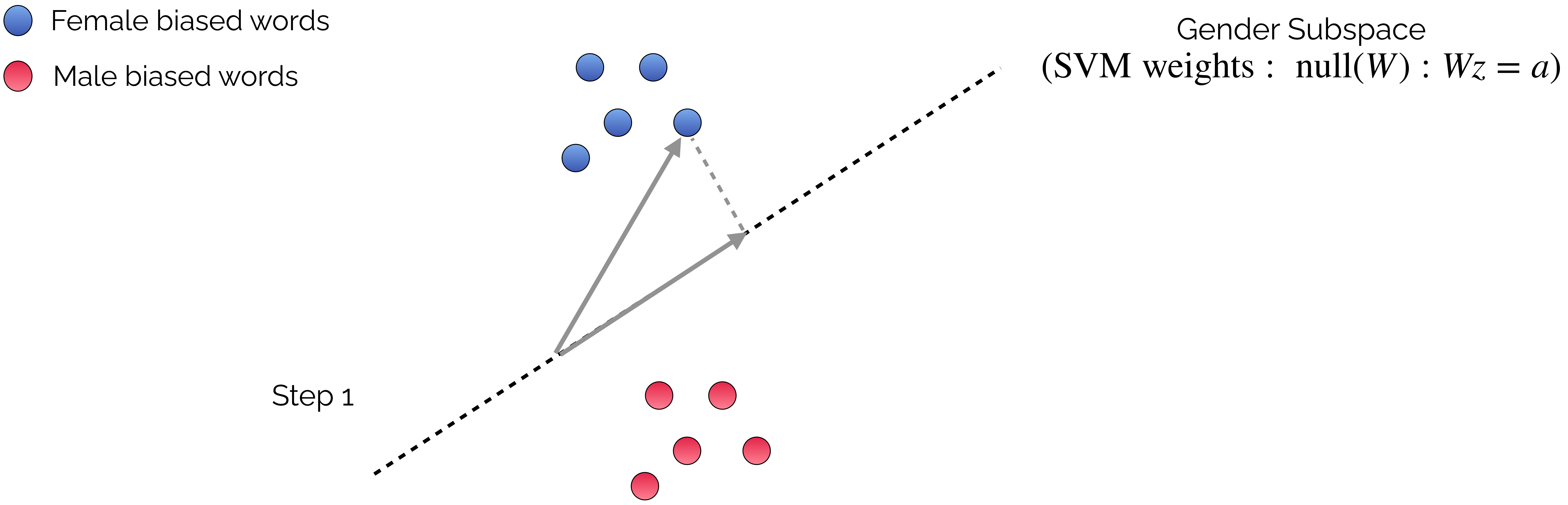
Debiasing Word Embeddings (Bolukbasi et al, 2016)

Prior Work - Unconstrained debiasing



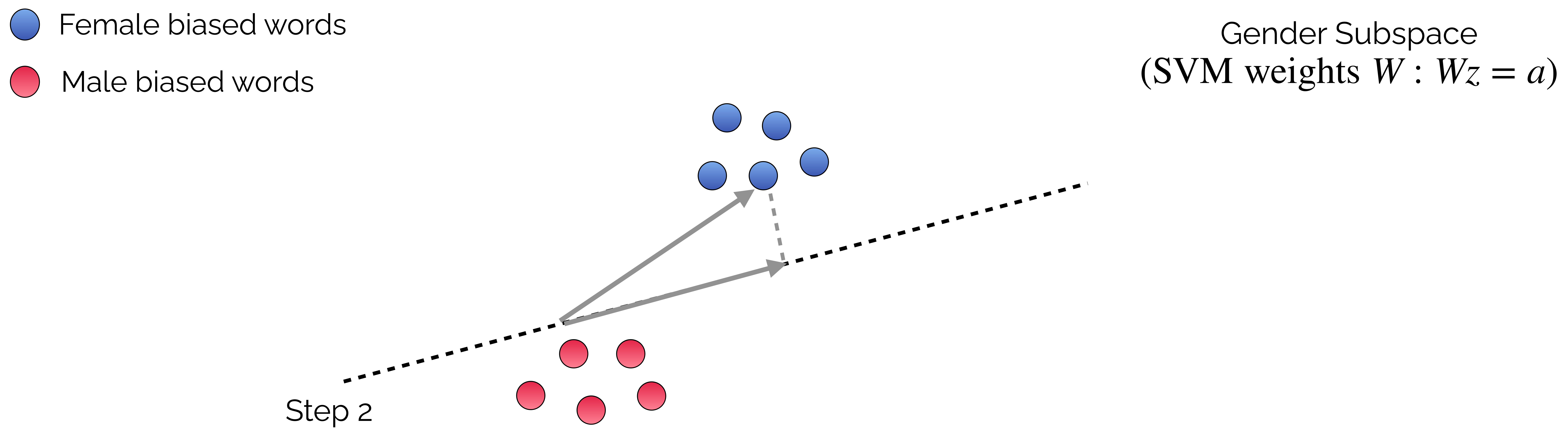
Debiasing Word Embeddings (Bolukbasi et al, 2016)

Prior Work - INLP



Iterative Nullspace Projection (Ravfogel et al, 2020)

Prior Work - INLP



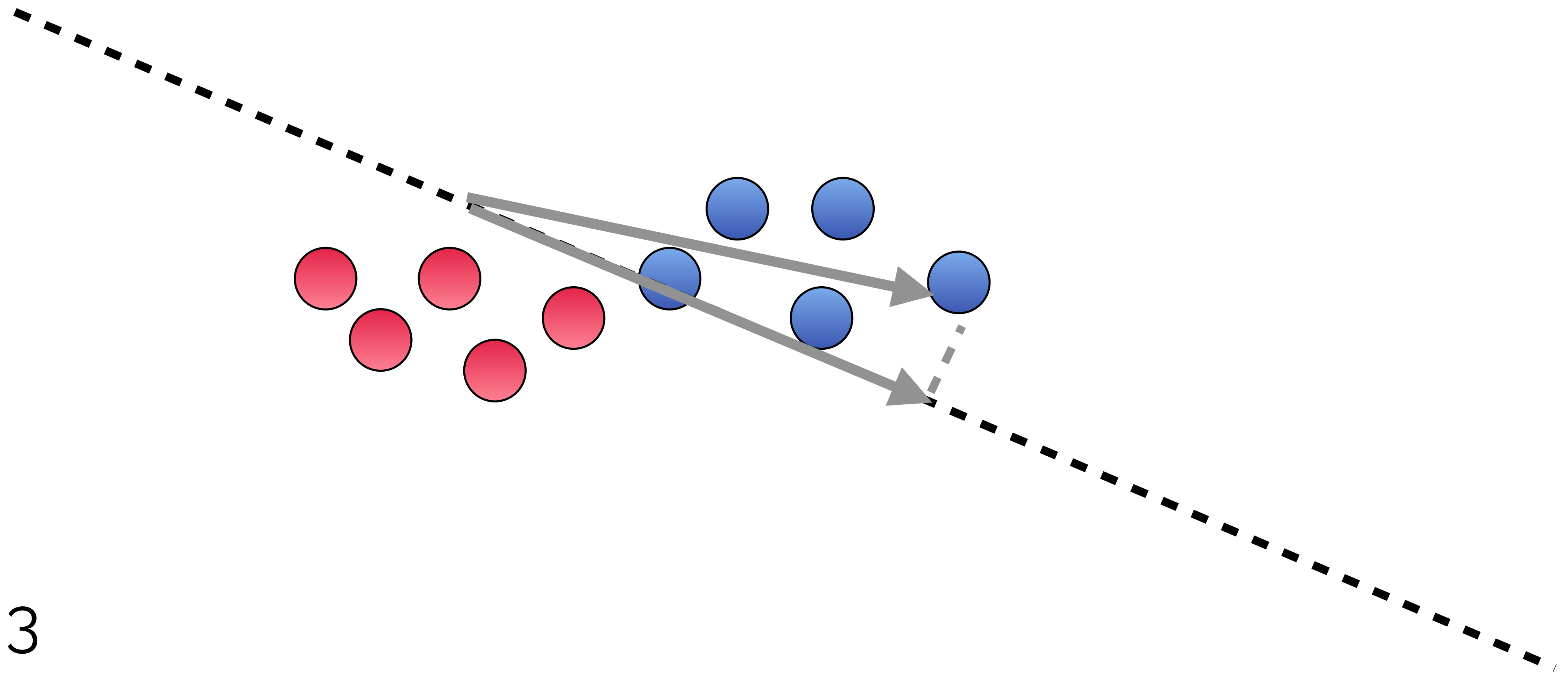
Iterative Nullspace Projection (Ravfogel et al, 2020)

Prior Work - INLP

Female biased words

Male biased words

Gender Subspace $(\vec{z}_{\text{male}} - \vec{z}_{\text{female}})$



Step 3

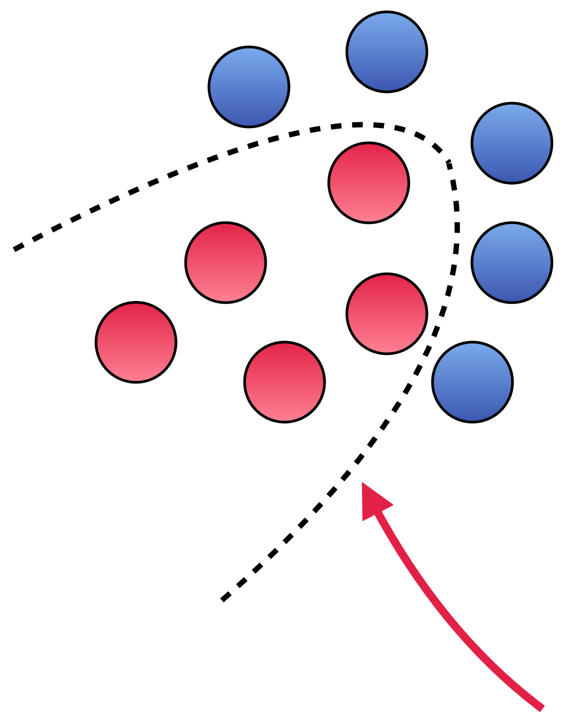
Iterative Nullspace Projection (Ravfogel et al, 2020)

Prior Work - INLP

Female biased words

Male biased words

Non-linear Gender Subspace



Step 4

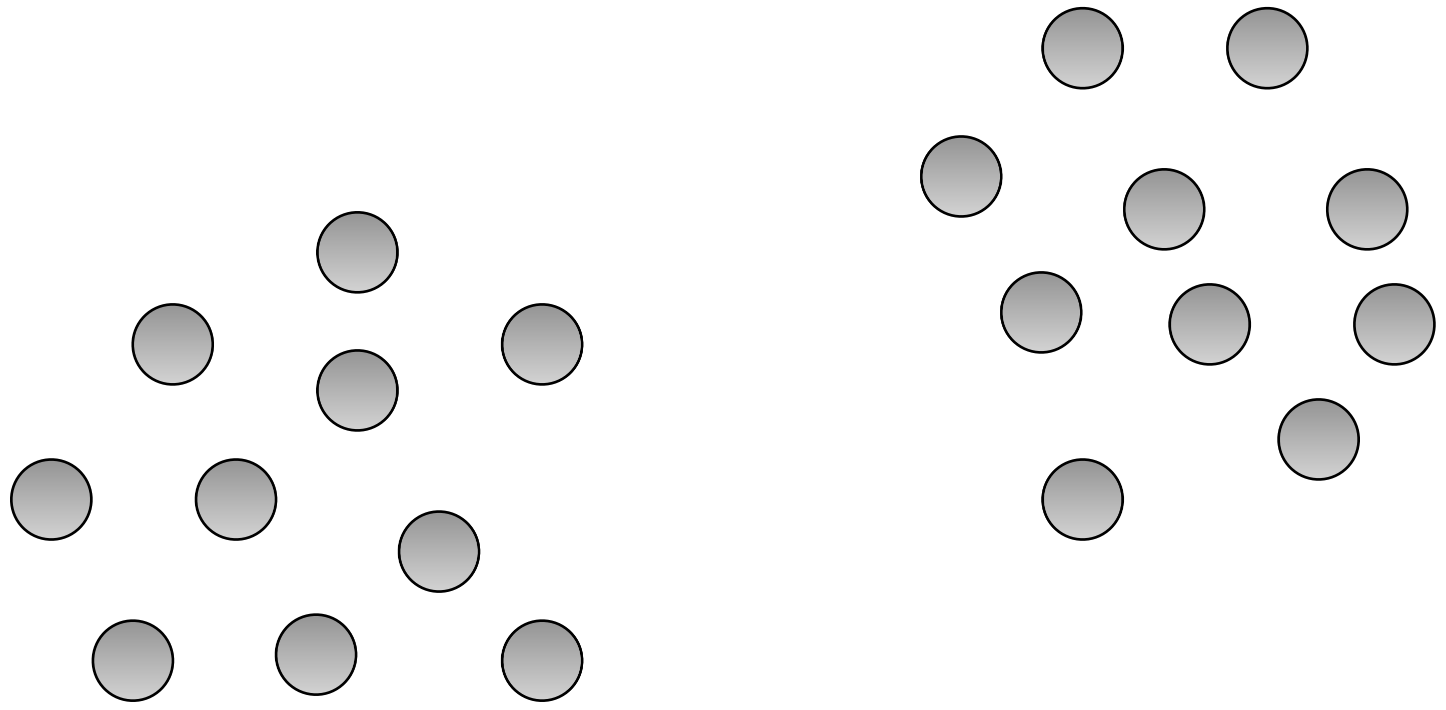
Still amenable to non-linear
probing attack

Iterative Nullspace Projection (Ravfogel et al, 2020)

Outline

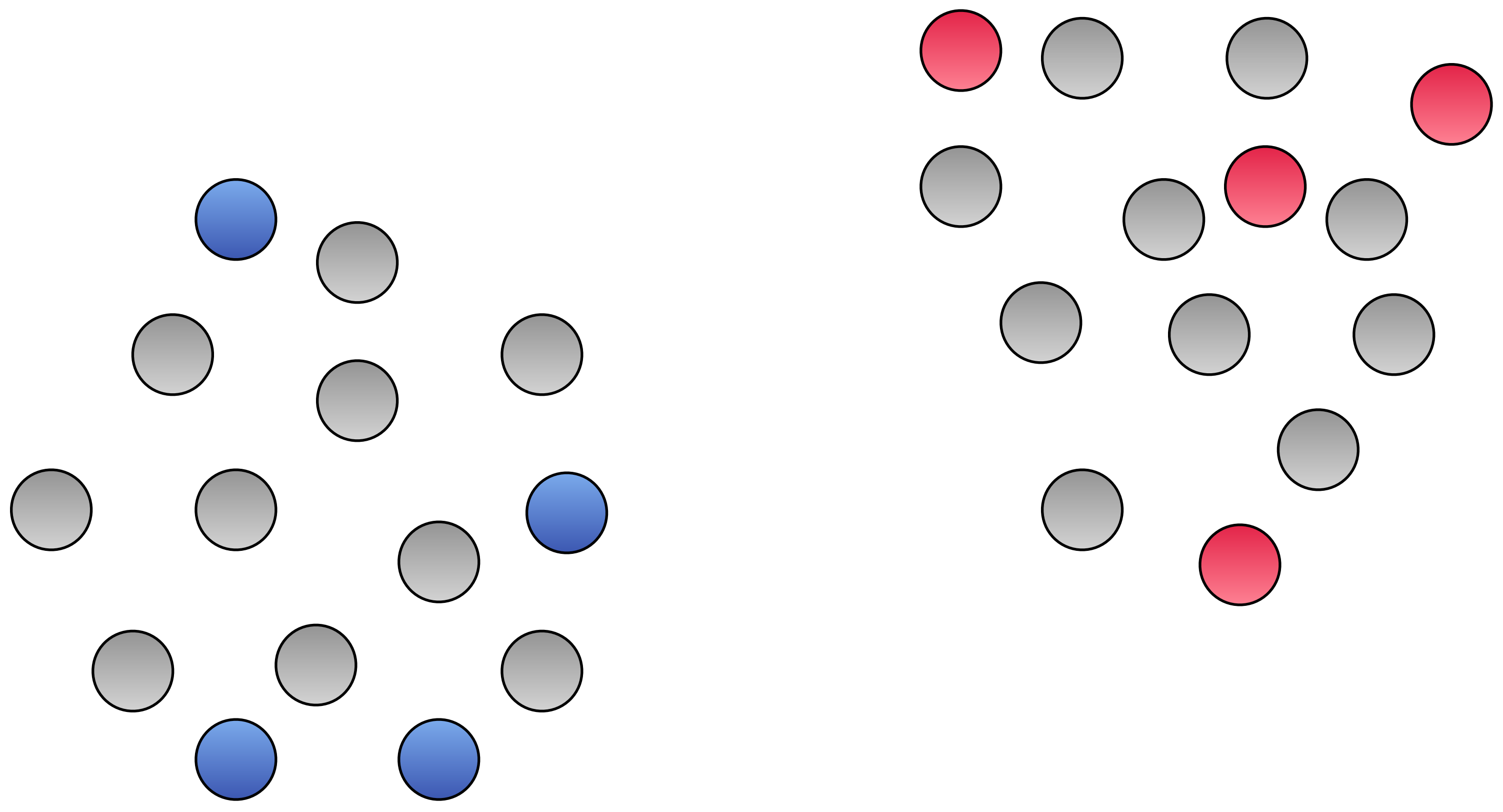
- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Information in high dimensions

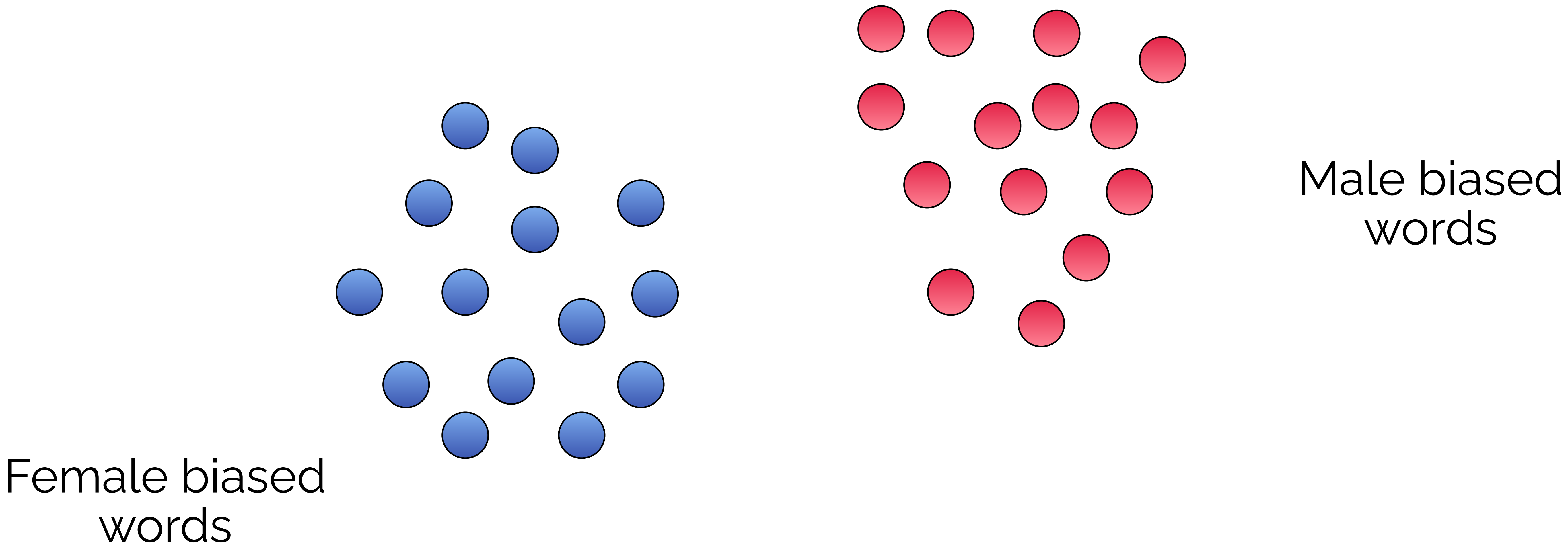


Information is encoded as distances
among high-dimensional vectors.

Attack on Representations

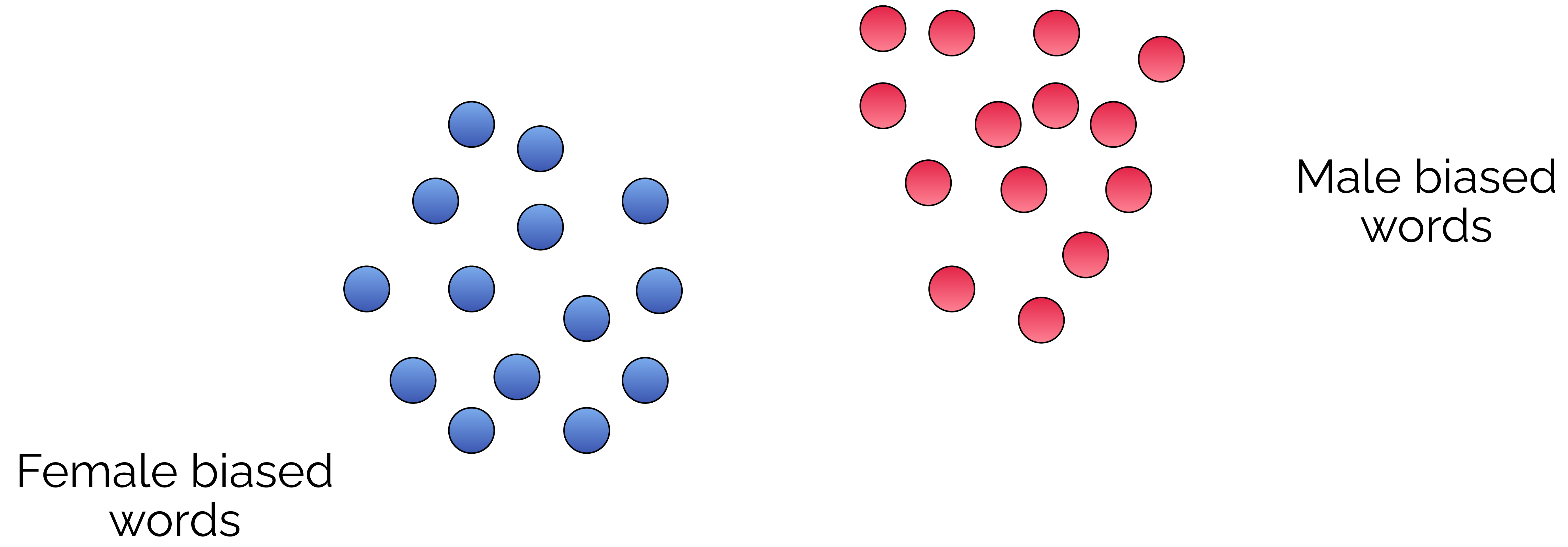


Attack on Representations



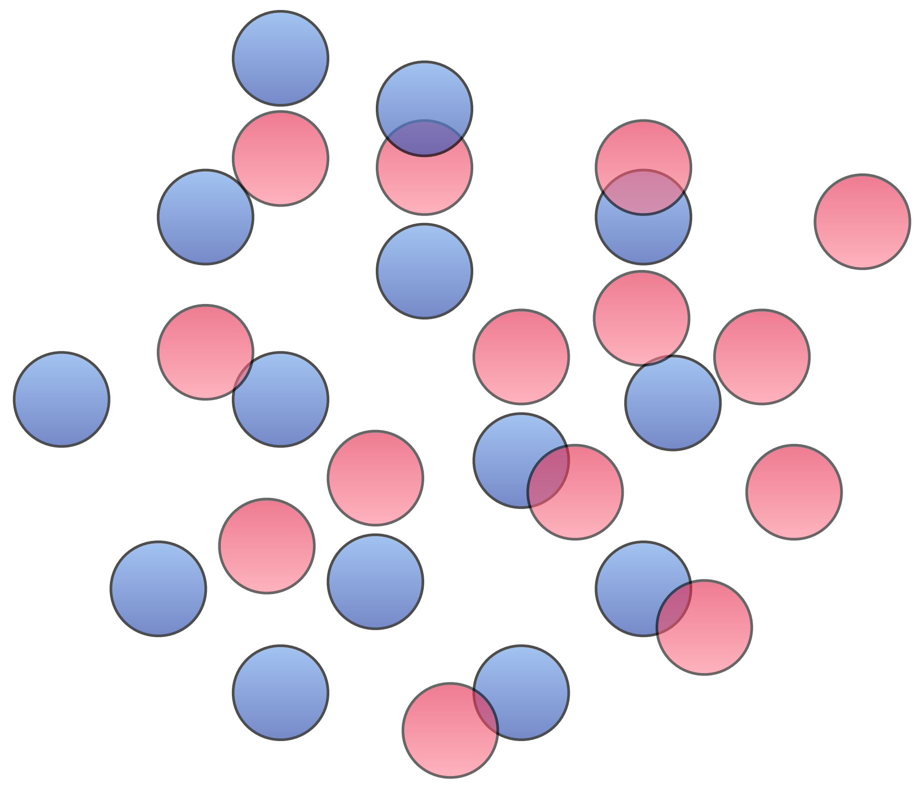
How do we nullify specific information?

Information to be deleted: Gender



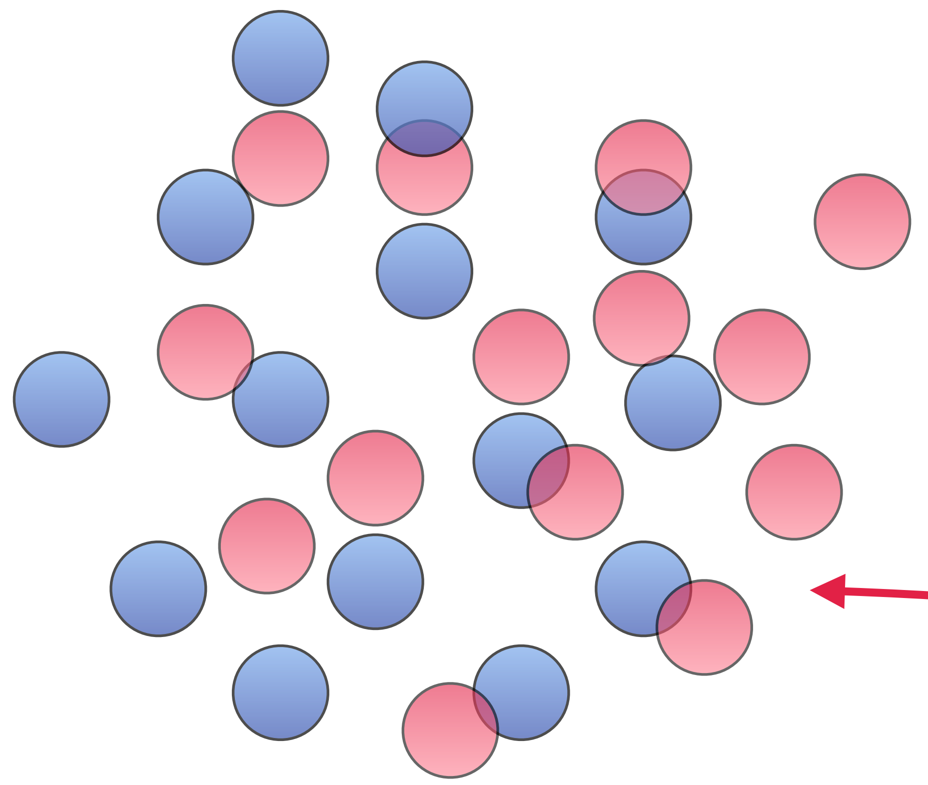
How do we nullify specific information?

Information to be deleted: Gender



How do we nullify specific information?

Information to be deleted: Gender

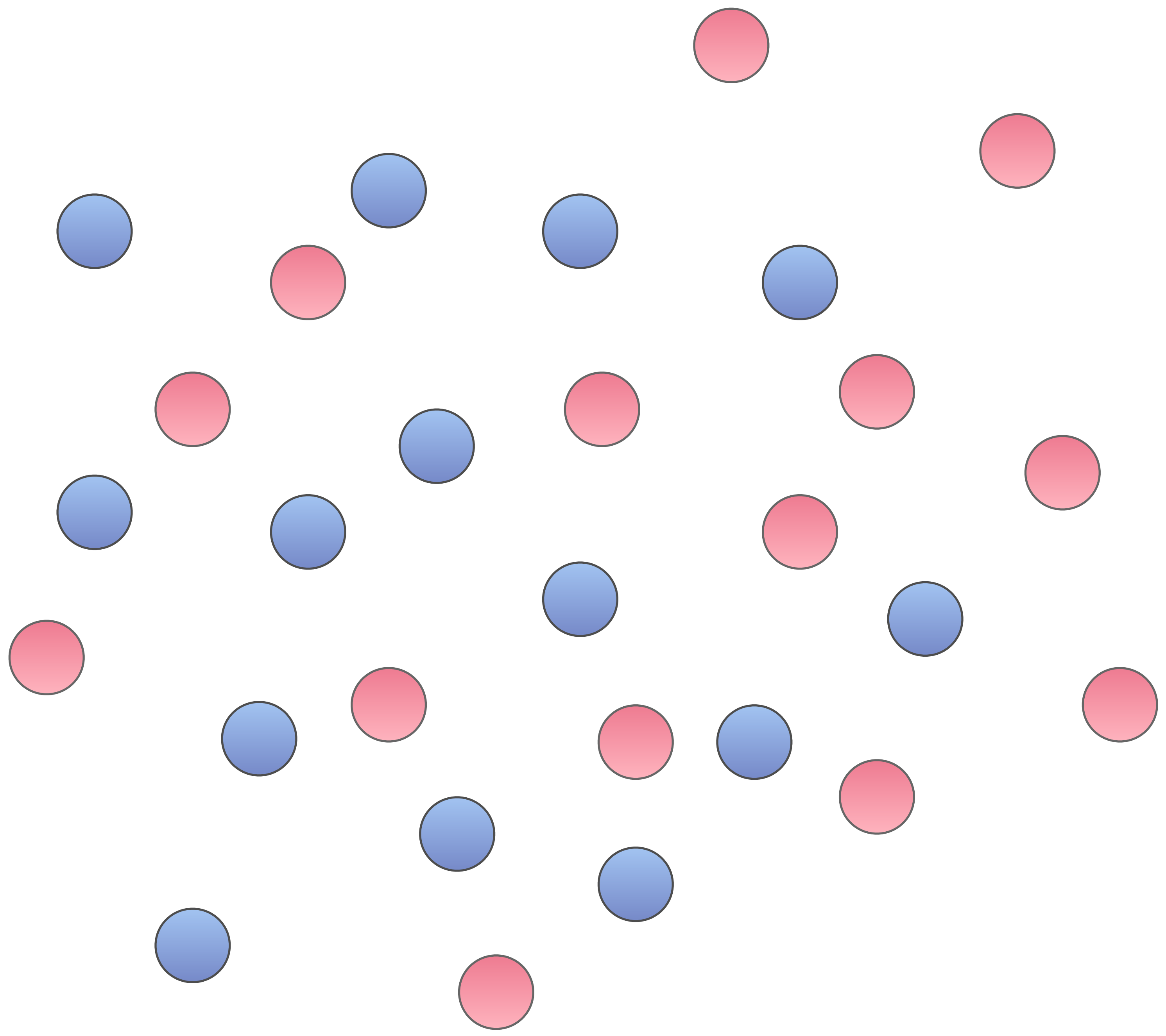


But some distances/information gets lost in the process

How do we retain as much information as possible?

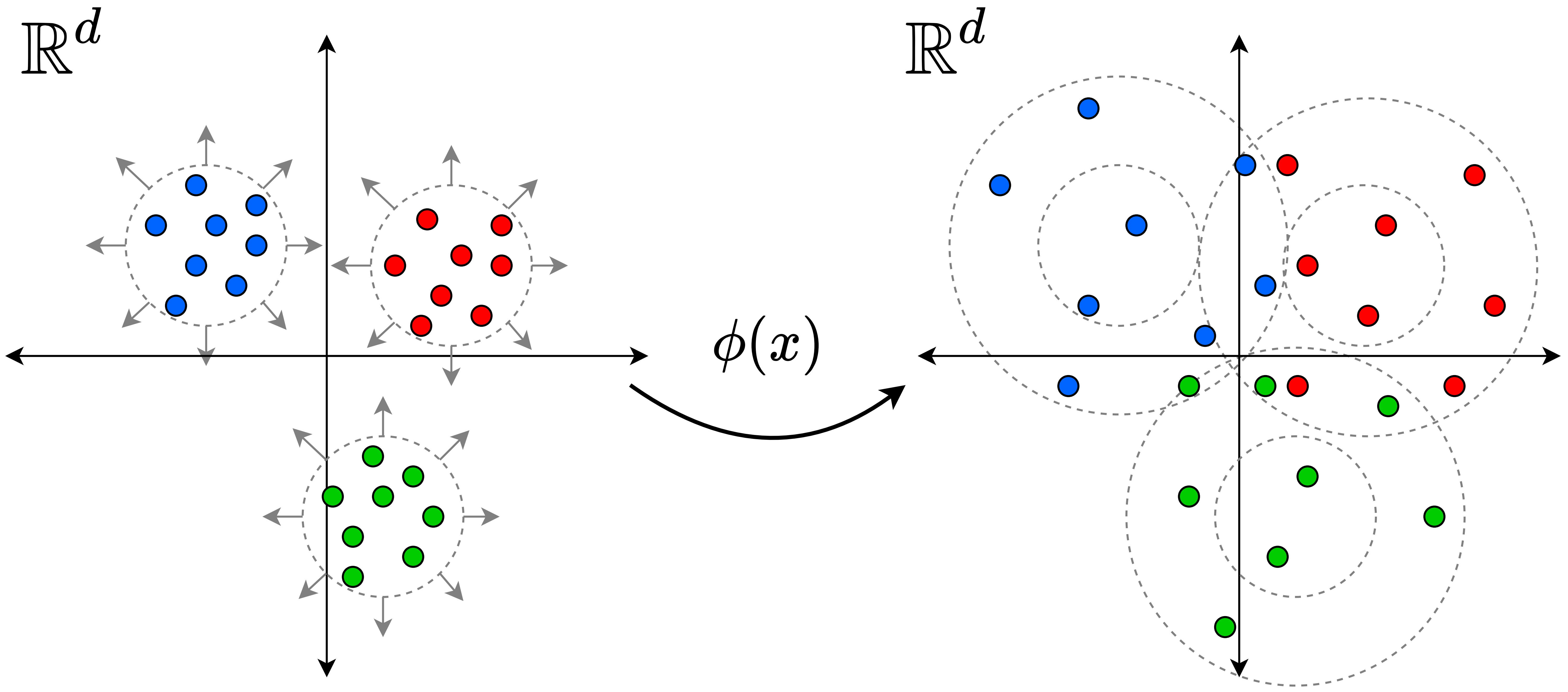
How do we nullify specific information?

Information to be deleted: Gender



Feature vectors usually lie in low-dimensional manifolds;
Increase the feature space

Recipe?



Recipe?

- Morph the feature space using a learnable function f

$$\max_{f} \text{Volume(feature space)} + \text{Volume(feature space of individual subgroups)}$$

Measuring Volume — Rate Distortion

- Rate-distortion measures the total number of binary bits required to encode a set of representations $Z \in \mathbb{R}^d$

$$R(Z, \epsilon) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{n\epsilon^2} Z Z^T \right)$$

Measuring Volume — Rate Distortion

- To measure volume of subgroups (categories of an attribute, e.g. male/female), we use a partition function $\Pi : Z \rightarrow \{Z_1, \ldots, Z_k\}$

$$R(Z, \epsilon \mid \Pi) = R(Z_1, \epsilon) + \ldots + R(Z_k, \epsilon)$$

Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Fairness-aware Rate Maximization (FaRM)

Unconstrained Objective

- Encode demographic information to be debiased as a partition function Π

Unconstrained Objective

- Encode demographic information to be debiased as a partition function Π
- Train a learnable function f with the objective:

$$\max_{f} R(Z, \epsilon) + R(Z, \epsilon \mid \Pi)$$

Unconstrained Objective

- Encode demographic information to be debiased as a partition function Π
- Train a learnable function f with the objective:

$$\max_{f} \underbrace{R(Z, \epsilon)}_{\text{Volume(feature space)}} + R(Z, \epsilon \mid \Pi)$$

Unconstrained Objective

- Encode demographic information to be debiased as a partition function Π
- Train a learnable function f with the objective:

$$\max_f R(Z, \epsilon) + R(Z, \epsilon \mid \Pi)$$

Volume(feature space of individual subgroups)

Sneak Peek into Results

Method	Accuracy (↓)	MDL (↑)	Rank (↑)
GloVe	100.0	0.1	300
INLP	86.3	8.6	210
FaRM	**53.9**	**24.6**	**247**

Constrained Objective

- We only care about the target attribute Y

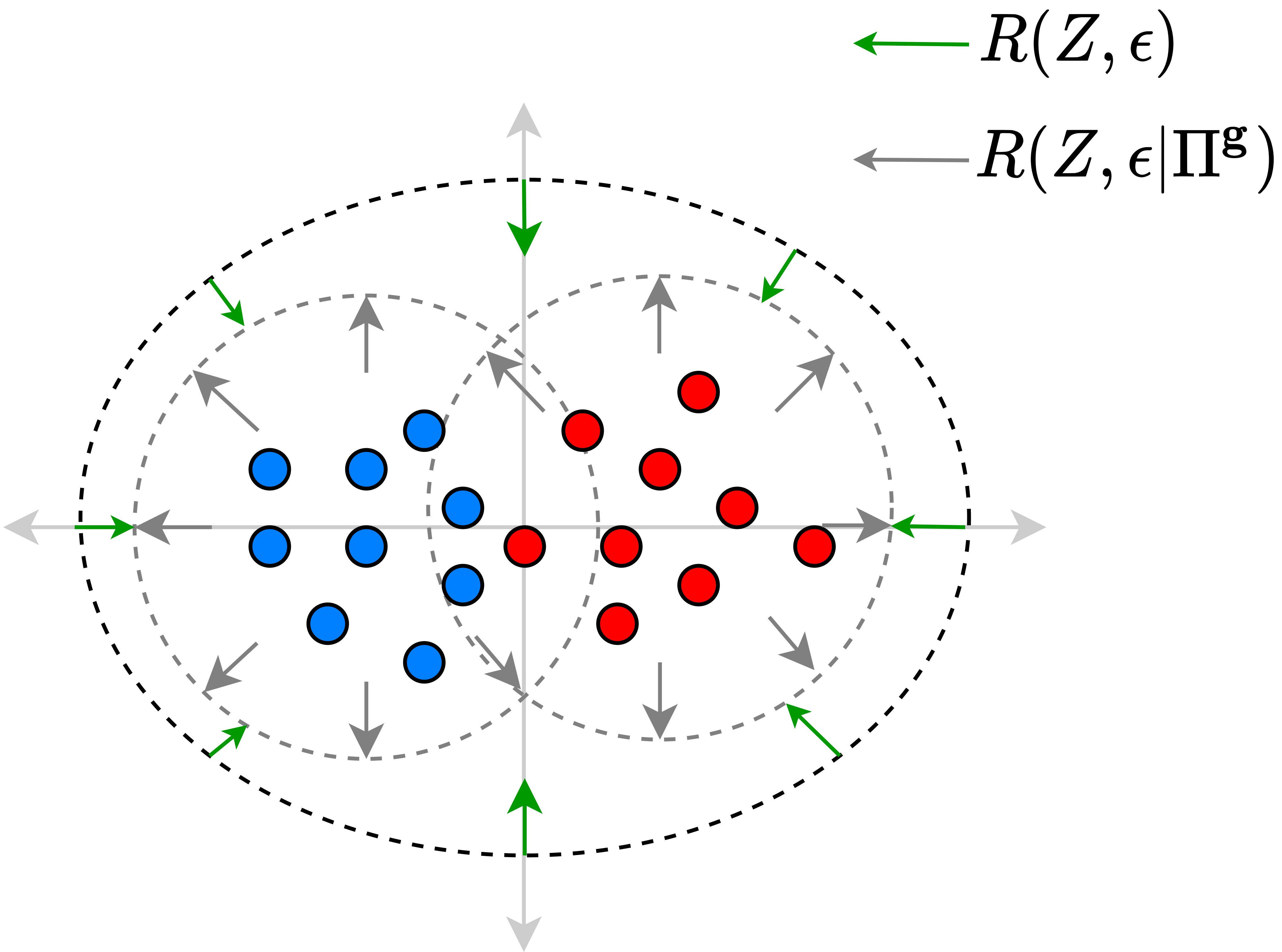
Constrained Objective

- We only care about the target attribute Y
- Target-class informativeness — $\min CE(\hat{y}, y)$

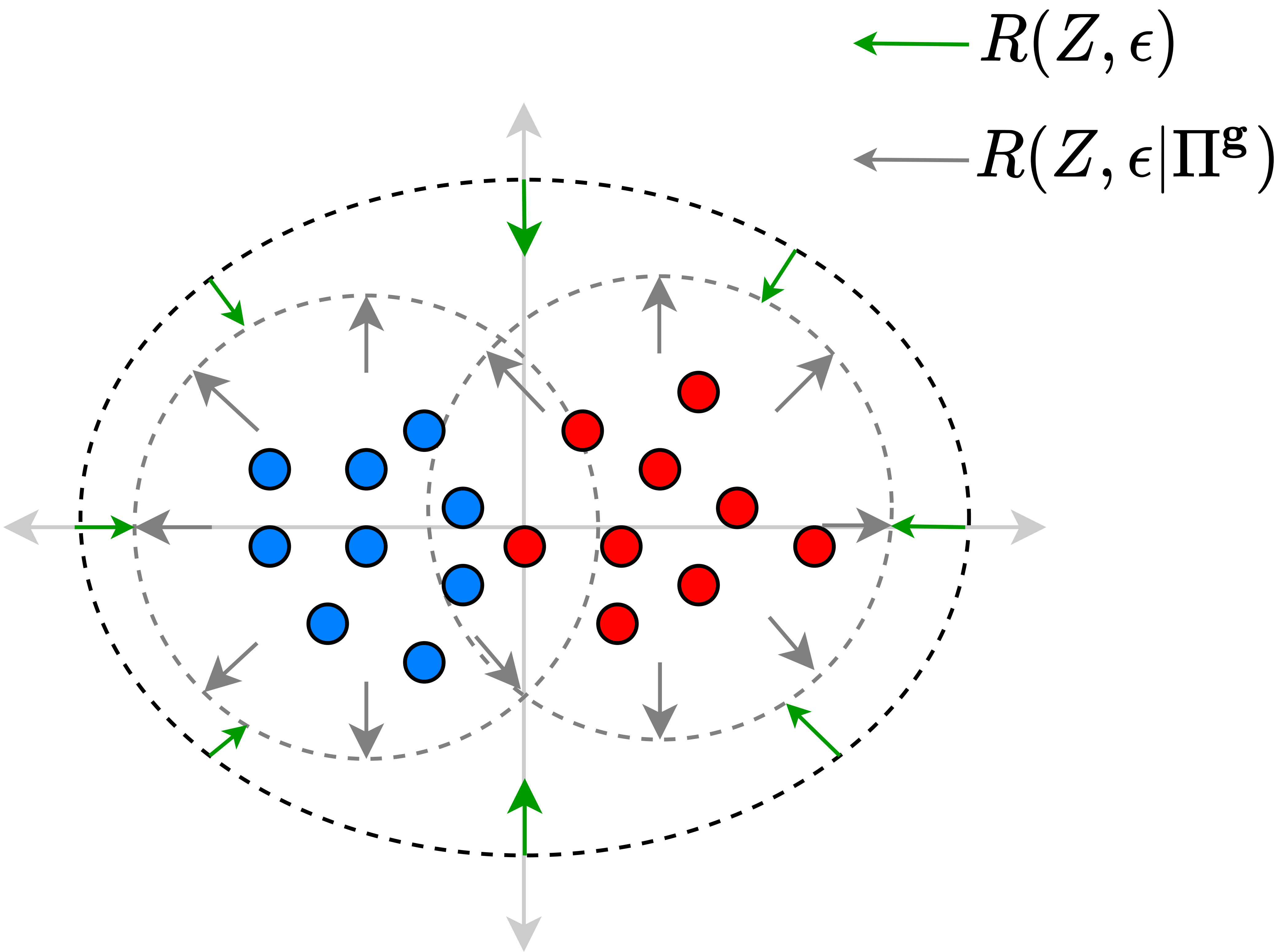
Constrained Objective

- We only care about the target attribute Y
- Target-class informativeness — $\min CE(\hat{y}, y)$
- Can we use rate-distortion to debias more robustly?

Recipe?

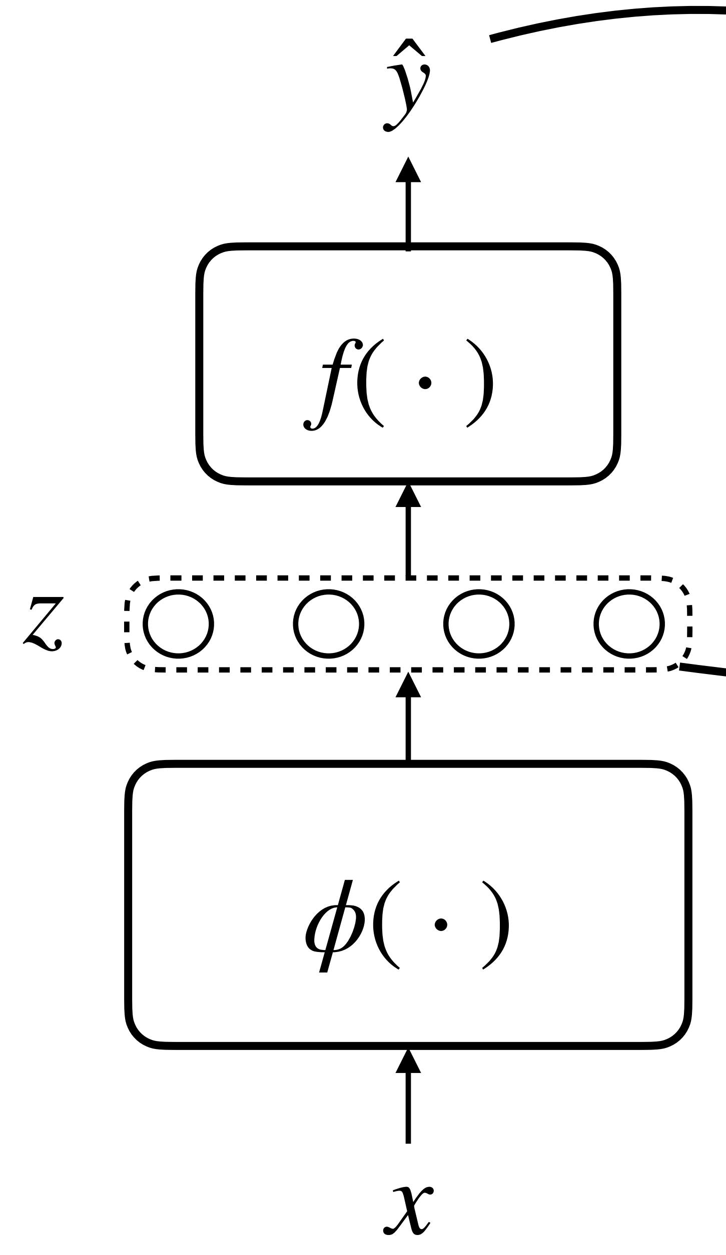


Recipe?



$$\min \text{Volume(feature space)} + \max \text{Volume(feature space of individual subgroups)}$$

Proposed Model



$$\max_{f,\phi} -\mathrm{CE}(\hat{y}, y) + \lambda\Big[R^c(Z, \epsilon \mid \Pi^{\mathbf{g}}) - R(Z, \epsilon)\Big]$$

Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

Evaluation

Metrics

- We evaluate the fairness of representations by 2 methods:

Metrics

- We evaluate the fairness of representations by 2 methods:
  - Probing representations for A
  - Inspecting the fairness of outcomes

Metrics

- We evaluate the fairness of representations by 2 methods:
  - Probing representations for A
  - Inspecting the fairness of outcomes
- For constrained debiasing, we report the probing target accuracy

Probing Metrics

- Probing Accuracy - accuracy obtained by a network for probing A or Y

Probing Metrics

- Probing Accuracy - accuracy obtained by a network for probing A or Y
- Minimum Description Length (MDL) - Coding length required to transmit labels Y given the data X
  - Higher MDL means more effort required in extracting Y from X

Fairness Metrics

- Demographic Parity - captures the "*equality of outcome*"

$$|P(\hat{Y} = + \mid A = a) - P(\hat{Y} = + \mid A = \bar{a})|$$

Fairness Metrics

- Demographic Parity - captures the "*equality of outcome*"

$$|P(\hat{Y} = + \mid A = a) - P(\hat{Y} = + \mid A = \bar{a})|$$

- TPR-GAP - captures "*equality of opportunity*" using different between TPR

$$\mathrm{TPR}_{A,Y} = P(\hat{Y} = + \mid A = a, Y = +)$$

$$\mathrm{Gap}_{A,Y} = \mathrm{TPR}_{a,Y} - \mathrm{TPR}_{\bar{a},Y}$$

Summary of Metrics

Summary of Metrics

- Target Attribute - Probing Accuracy (constrained)
- Protected Attribute - Probing Accuracy and MDL (both)
- Fairness - DP and TPR-GAP (both)

Outline

- Motivation
- Problem Setup
- Prior Work
- Intuition behind our work
- FaRM
- Evaluation Setup
- Results

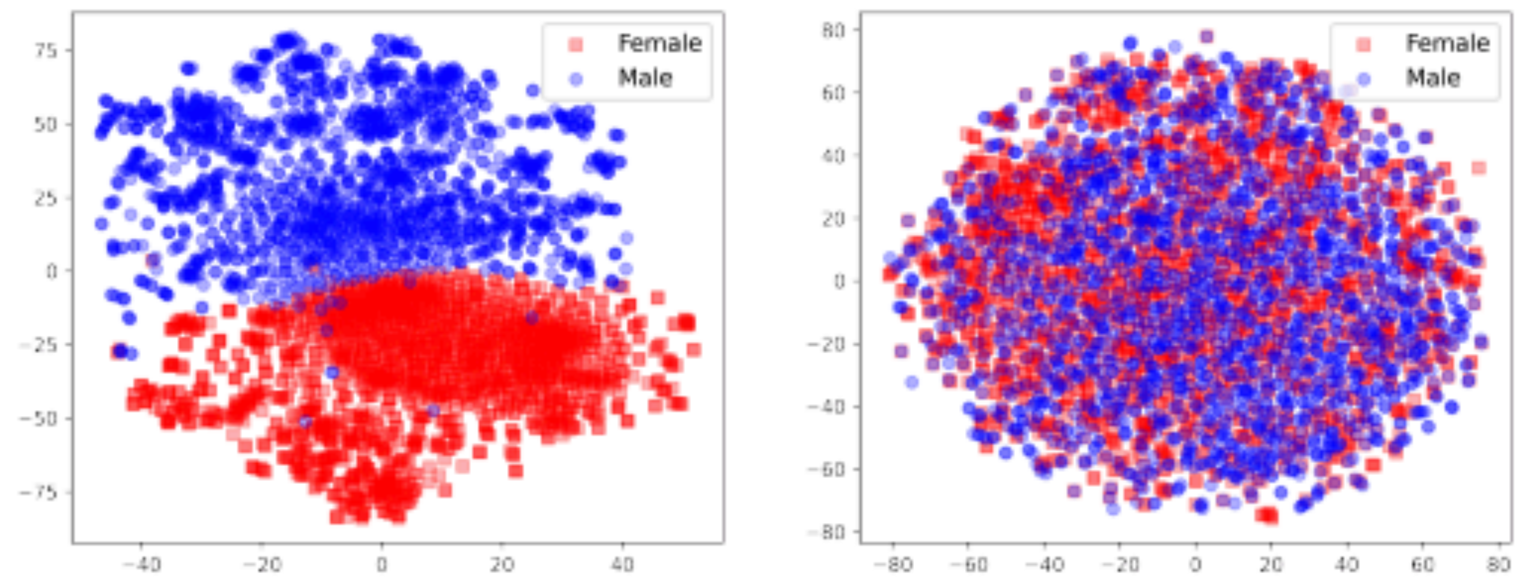
Results - Unconstrained Debiasing

Metric	Method	Split			
		50%	60%	70%	80%
Sentiment Acc. (↑)	Original	75.5	75.5	74.4	71.9
	INLP	**75.1**	73.1	**69.2**	**64.5**
	FaRM	74.8	**73.2**	67.3	63.5
Race Acc. (↓)	Original	87.7	87.8	87.3	87.4
	INLP	69.5	82.2	80.3	69.9
	FaRM	**54.2**	**69.9**	**69.0**	**52.1**
DP (↓)	Original	0.26	0.44	0.63	0.81
	INLP	0.16	0.33	0.30	0.28
	FaRM	**0.09**	**0.10**	**0.17**	**0.22**
$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ (↓)	Original	0.15	0.24	0.33	0.41
	INLP	0.12	0.18	0.16	0.16
	FaRM	**0.09**	**0.10**	**0.12**	**0.14**

Results - Unconstrained Debiasing

Metric	Method	FastText	BERT
Profession Acc. (↑)	Original	79.9	80.9
	INLP	**76.3**	**77.8**
	FaRM	54.8	55.8
Gender Acc. (↓)	Original	98.9	99.6
	INLP	67.4	94.9
	FaRM	**57.6**	**55.6**
DP (↓)	Original	1.65	1.68
	INLP	1.51	1.50
	FaRM	**0.12**	**0.14**
$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ (↓)	Original	0.185	0.171
	INLP	0.089	0.096
	FaRM	**0.006**	**0.079**

Results - Unconstrained Debiasing



(a) GloVe (b) Debiased

Figure 4: Projections of Glove embeddings before (left) and after (right) debiasing. Intial female and male biased representations are shown in **red** and **blue** respectively.

Results - Constrained Debiasing

Method	DIAL											
	Sentiment ($\mathbf{y}$)		Race ($\mathbf{g}$)		Fairness		Mention ($\mathbf{y}$)		Race ($\mathbf{g}$)		Fairness	
	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\mathrm{Gap}_{\mathbf{g}}^{\mathrm{RMS}}$ ↓	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\mathrm{Gap}_{\mathbf{g}}^{\mathrm{RMS}}$ ↓
$\mathrm{BERT}_{\mathrm{base}}$ (pre-trained)	63.9	300.7	10.9	242.6	0.41	0.20	66.1	290.1	24.6	258.8	0.20	0.10
$\mathrm{BERT}_{\mathrm{base}}$ (fine-tuned)	**76.9**	99.0	18.4	176.2	0.30	0.14	**81.7**	49.1	28.7	199.2	0.06	0.03
AdS	72.9	56.9	5.2	290.6	0.43	0.21	81.1	7.6	21.7	270.3	0.06	0.03
FaRM	73.2	**17.9**	**0.2**	**296.5**	**0.26**	**0.14**	78.8	**3.1**	**0.3**	**324.8**	0.06	0.03

Results - Constrained Debiasing

Method	PAN16											
	Mention ($\mathbf{y}$)		Gender ($\mathbf{g}$)		Fairness		Mention ($\mathbf{y}$)		Age ($\mathbf{g}$)		Fairness	
	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ ↓	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ ↓
$\text{BERT}_{\text{base}}$ (pre-trained)	72.3	259.7	7.4	300.5	0.11	0.056	72.8	262.6	6.1	302.0	0.14	0.078
$\text{BERT}_{\text{base}}$ (fine-tuned)	**89.7**	4.0	15.1	267.6	0.04	0.007	**89.3**	4.8	7.4	295.4	0.04	0.006
AdS	89.7	7.6	4.9	**313.9**	0.04	0.007	89.2	6.0	1.1	**315.1**	0.04	**0.004**
FaRM	88.7	**1.7**	**0.0**	312.4	0.04	0.007	88.6	**0.8**	**0.0**	312.6	**0.03**	0.008

Results - Constrained Debiasing

Method	Biographies					
	Profession ($\mathbf{y}$)		Gender ($\mathbf{g}$)		Fairness	
	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ ↓
$\text{BERT}_{\text{base}}$ (pre-trained)	74.3	499.9	45.2	27.6	0.43	0.169
$\text{BERT}_{\text{base}}$ (fine-tuned)	99.9	**2.2**	8.3	448.9	0.46	**0.001**
AdS	99.9	3.3	**3.1**	449.5	0.45	0.003
FaRM	99.9	7.6	7.4	**460.3**	**0.42**	0.002

Results - Debiasing Multiple Attributes

Setup	Pan16											
	Mention (y)		Age ($\mathbf{g}_1$)		Fairness ($\mathbf{g}_1$)		Gender ($\mathbf{g}_2$)		Fairness ($\mathbf{g}_2$)		Inter. Groups ($\mathbf{g}_1, \mathbf{g}_2$)	
	F1↑	MDL↓	ΔF1↓	MDL↑	DP↓	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ ↓	ΔF1↓	MDL↑	DP↓	$\text{Gap}_{\mathbf{g}}^{\text{RMS}}$ ↓	ΔF1↓	MDL↑
$\text{BERT}_{\text{base}}$ (fine-tuned)	88.6	6.8	14.9	196.4	0.06	0.009	16.5	192.0	0.04	0.014	20.7	117.2
AdS	**88.6**	**5.5**	2.2	231.5	0.05	0.006	1.6	230.9	0.04	0.017	9.1	118.5
FaRM (N-partition)	87.0	13.4	0.0	234.3	**0.03**	**0.003**	0.0	234.2	0.06	0.025	0.7	**468.0**
FaRM (1-partition)	86.4	15.6	0.0	**234.6**	0.05	0.006	0.0	234.2	**0.02**	**0.009**	**0.0**	467.7

Conclusion

- We propose a robust framework to delete specific information from representation

Conclusion

- We propose a robust framework to delete specific information from representation
- First approach to show any resistance against non-linear probing attacks

Conclusion

- We propose a robust framework to delete specific information from representation
- First approach to show any resistance against non-linear probing attacks
- FaRM is robust to label corruption and dataset size

Conclusion

- We propose a robust framework to delete specific information from representation
- First approach to show any resistance against non-linear probing attacks
- FaRM is robust to label corruption and dataset size
- FaRM prevents leakage of intersectional biases

Conclusion

- We propose a robust framework to delete specific information from representation
- First approach to show any resistance against non-linear probing attacks
- FaRM is robust to label corruption and dataset size
- FaRM prevents leakage of intersectional biases
- We encourage researchers to use FaRM for XAI, or ensuring fairness in complex ML tasks (e.g. language generation)

Conclusion

- We propose a robust framework to delete specific information from representation
- First approach to show any resistance against non-linear probing attacks
- FaRM is robust to label corruption and dataset size
- FaRM prevents leakage of intersectional biases
- We encourage researchers to use FaRM for XAI, or ensuring fairness in complex ML tasks (e.g. language generation)

@SomnathBrc brcsomnath/FaRM somnath@cs.unc.edu