Semantic representation bias in high-stakes settings: risks and opportunities

Maria De-Arteaga

Joint PhD Candidate, Machine Learning & Public Policy

(Starting July 2020) Assistant Professor, McCombs School of Business, UT Austin

MACHINE LEARNING DEPARTMENT

Carnegie Mellon University Heinz College

Auton Lab

WIRED

BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

Will Machines Be Able to Tell When Patients Are About to Die?

The New York Times

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

WIRED

BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

Will Machines Be Able to Tell When Patients Are About to Die?

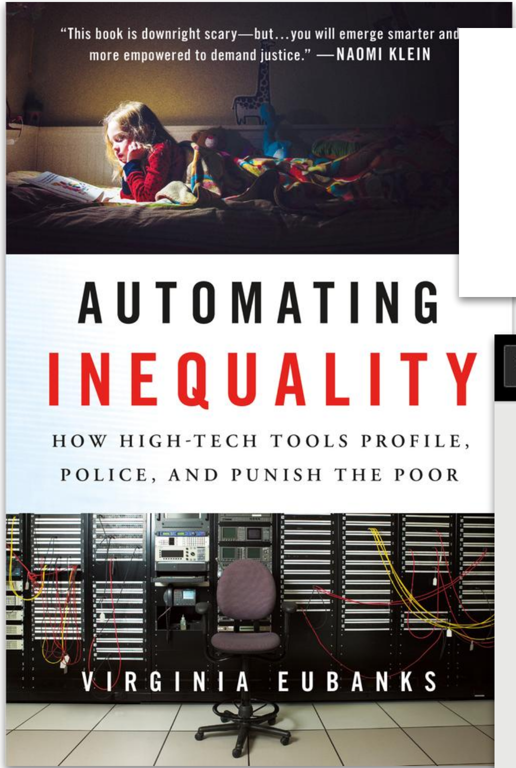
The New York Times

TheUpshot

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Machines are better than humans at making predictions.
[Meehl’54, Dawes’89, Grove’00]

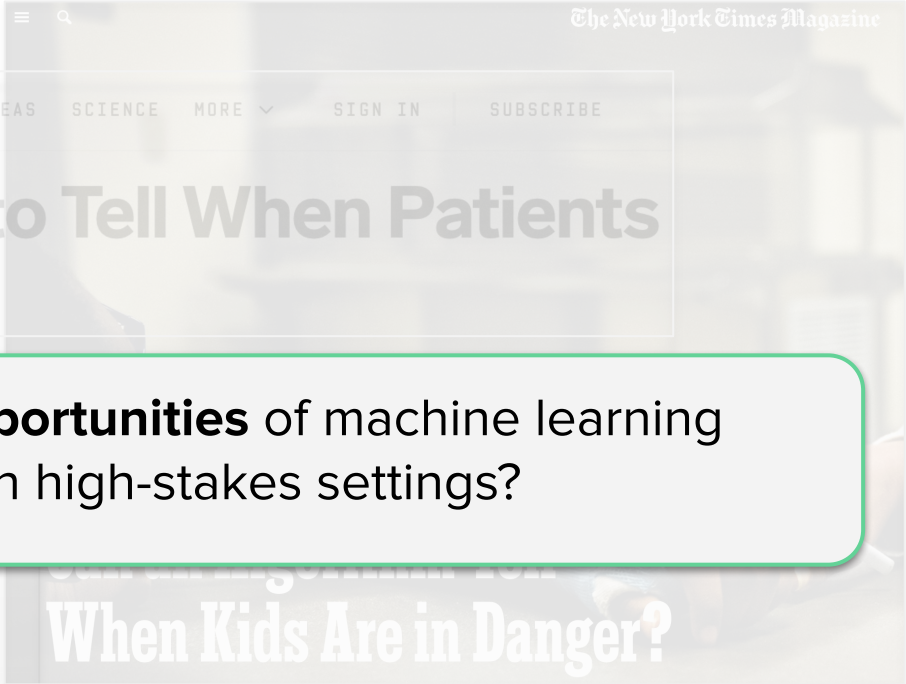

"This book is downright scary—but…you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN
AUTOMATING
INEQUALITY
HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR
VIRGINIA EUBANKS


BUSINESS NEWS
OCTOBER 9, 2018 / 11:12 PM / A YEAR AGO
REUTERS
Amazon scraps secret AI recruiting tool that showed bias against women

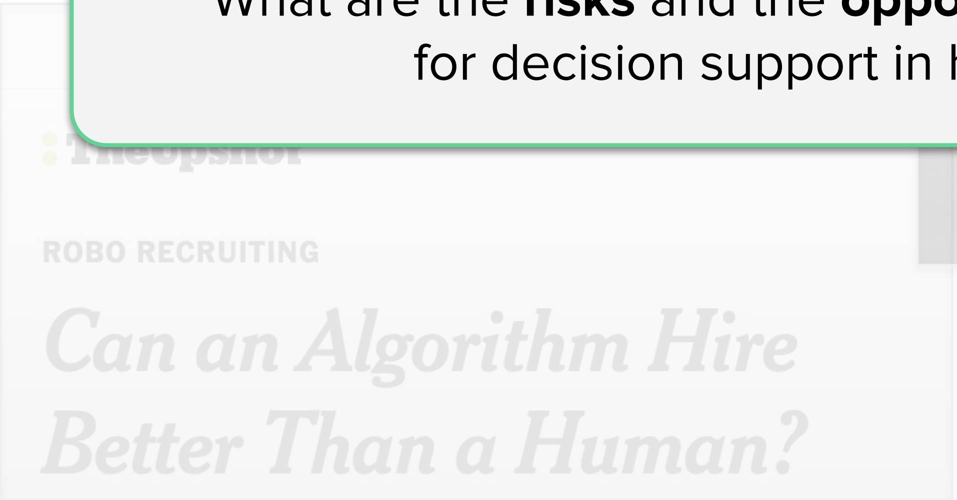

Sections
The Washington Post
Democracy Dies in Darkness
Health
Racial bias in a medical algorithm favors white patients over sicker black patients

WIRED
BUSINESS CULTURE GEAR IDEAS SCIENCE MORE SIGN IN SUBSCRIBE

Will Machines Be Able to Tell When Patients Are About to Die?


The New York Times Magazine

When Kids Are in Danger?


ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

What are the **risks** and the **opportunities** of machine learning for decision support in high-stakes settings?

Outline

1. Risks of compounding injustices

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (FAT* 2019)
Maria De-Arteaga (CMU), Alexey Romanov (UMASS), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Adam Kalai (MSR)

2. What are the biases in my word embedding?

What are the biases in my word embedding? (AIES 2019)
Nathaniel Swinger[=] (Lexington HS), Maria De-Arteaga[=] (CMU), Neil Thomas Heffernan IV (Shrewsbury HS), Mark Leiserson (UMD), Adam Kalai (MSR)

3. Using bias to fight bias

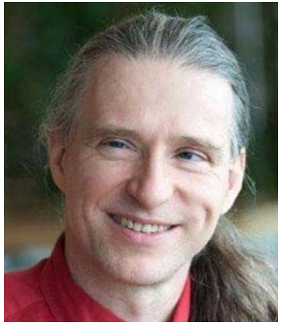
What's in a Name? Reducing Bias in Bios without Access to Protected Attributes (NAACL 2019)
Alexey Romanov (UMASS), Maria De-Arteaga (CMU), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Anna Rumshisky (UMASS), Adam Kalai (MSR) 🏆

Alexey Romanov

Adam Kalai

Hanna Wallach

Jennifer Chayes



Christian borgs

Alexandra Chouldechova

Krishnaram Kenthapadi

Sahin Geyik

Max Leiserson

Nathaniel Swinger, Neil Thomas Heffernan IV

An artificially intelligent headhunter?

Forbes

Billionaires Innovation Leadership Money Consu

Forbes CommunityVoice Connecting expert communities to the Forbes audience. What is

5,220 views | Jul 12, 2018, 07:00am

Welcome To The Age Of Recruiting Automation

npr

SCIENCE

Now Algorithms Are Deciding Whom To Hire, Based On Voice

4:27

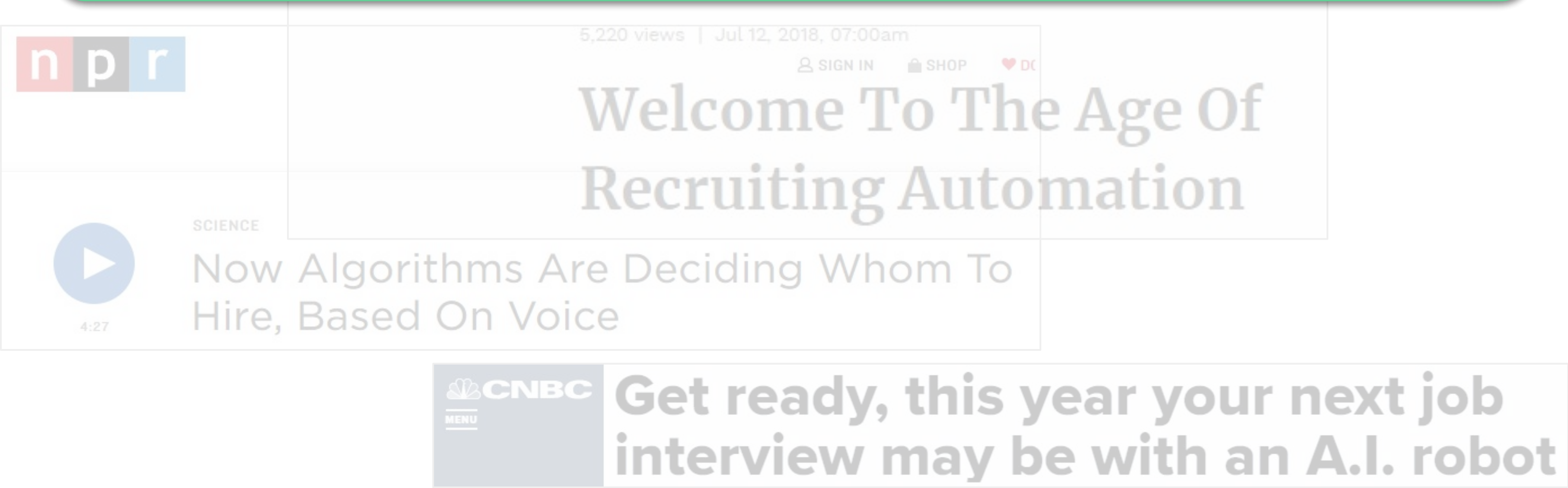
CNBC

MENU

Get ready, this year your next job interview may be with an A.I. robot

An artificially intelligent headhunter?

Can we **quantify** risks of gender bias in automated recruiting? Can we **mitigate** the bias?


5,220 views | Jul 12, 2018, 07:00am

SIGN IN SHOP

Welcome To The Age Of Recruiting Automation

npr

SCIENCE

Now Algorithms Are Deciding Whom To Hire, Based On Voice

4:27

CNBC

MENU

Get ready, this year your next job interview may be with an A.I. robot

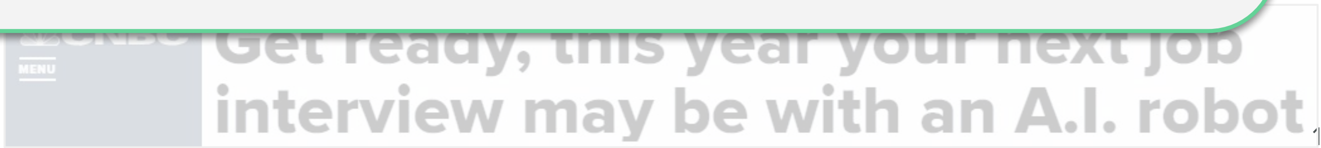
[FAT*'19]

An artificially intelligent headhunter?

Can we **quantify** risks of gender bias in automated recruiting? Can we **mitigate** the bias?

Findings:

- Compounding gender imbalances in large-scale study [**FAT*'19**]
- Leverage biases present in word embeddings [**AIES '19**] to mitigate biases without access to protected attributes [**NAACL '19**]



Outline

1. Risks of compounding injustices

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (FAT* 2019)
Maria De-Arteaga (CMU), Alexey Romanov (UMASS), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Adam Kalai (MSR)

2. What are the biases in my word embedding?

What are the biases in my word embedding? (AIES 2019)
Nathaniel Swinger[=] (Lexington HS), Maria De-Arteaga[=] (CMU), Neil Thomas Heffernan IV (Shrewsbury HS), Mark Leiserson (UMD), Adam Kalai (MSR)

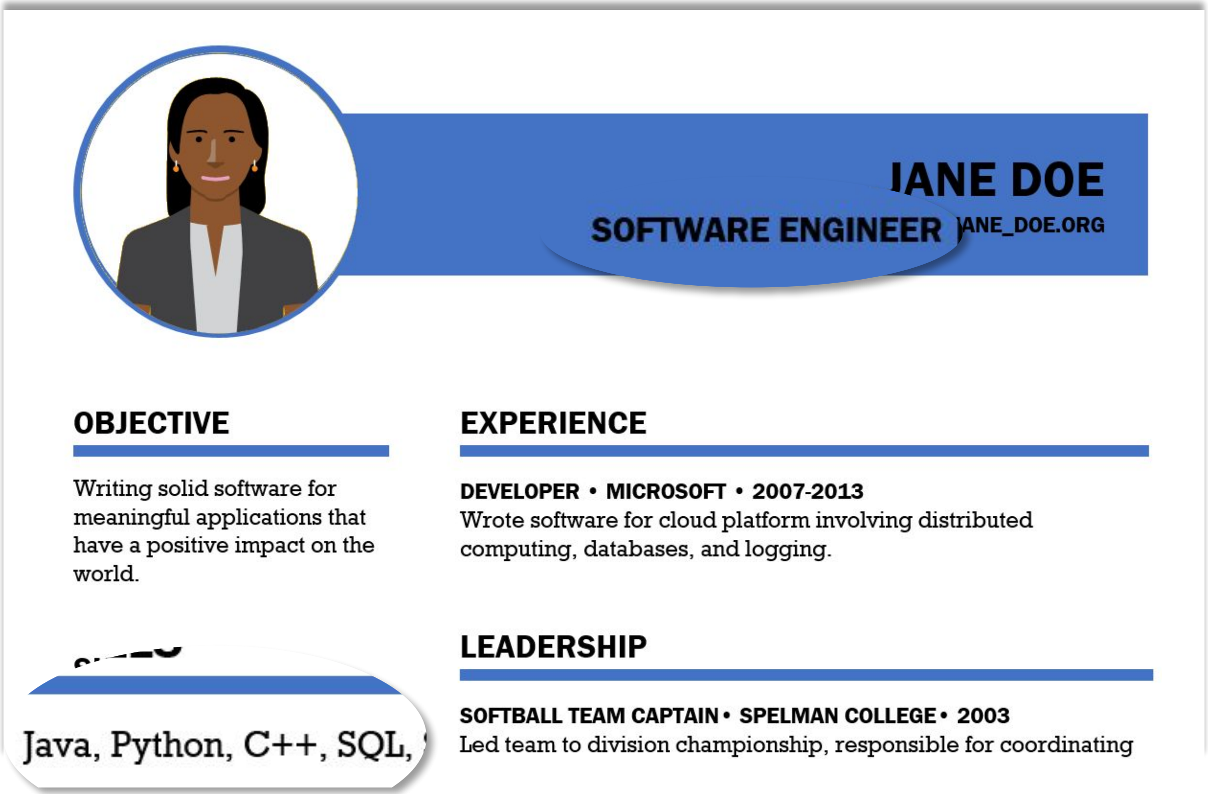
3. Using bias to fight bias

What's in a Name? Reducing Bias in Bios without Access to Protected Attributes (NAACL 2019)
Alexey Romanov (UMASS), Maria De-Arteaga (CMU), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Anna Rumshisky (UMASS), Adam Kalai (MSR)

Computer Programmer

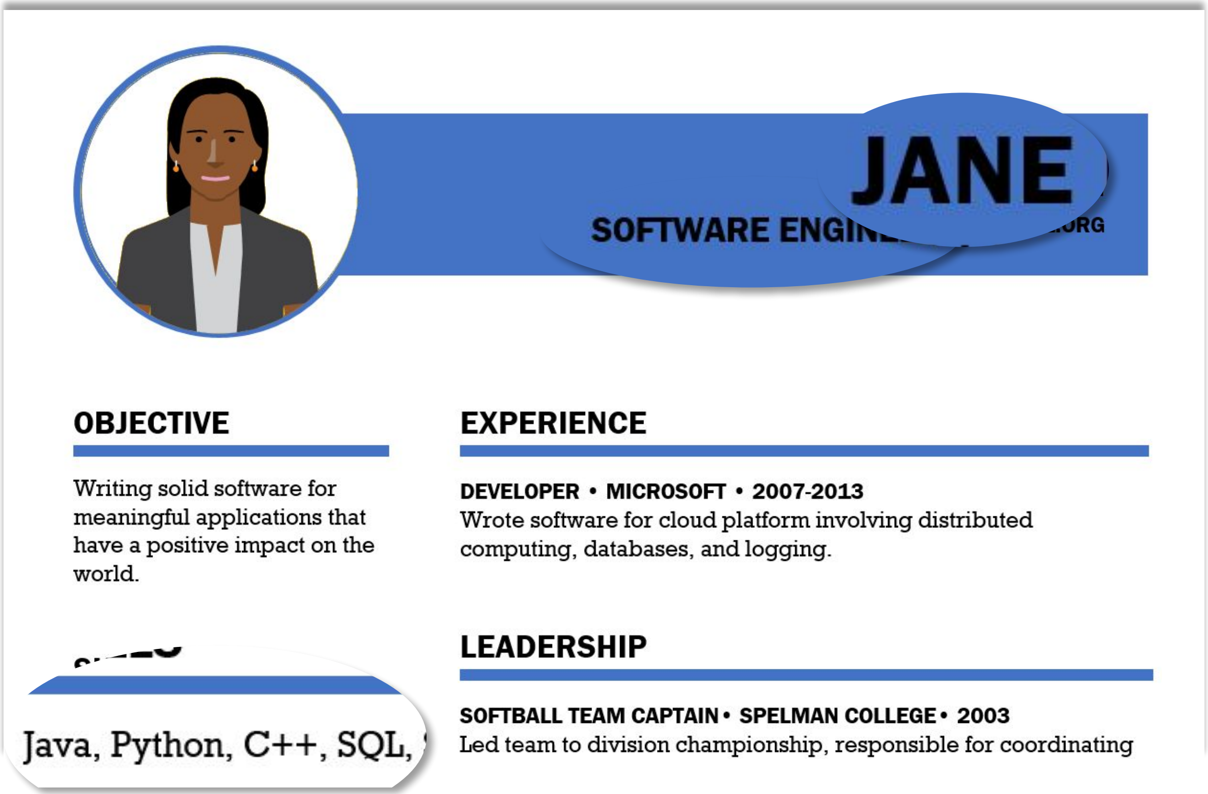
[FAT*'19]

Computer Programmer



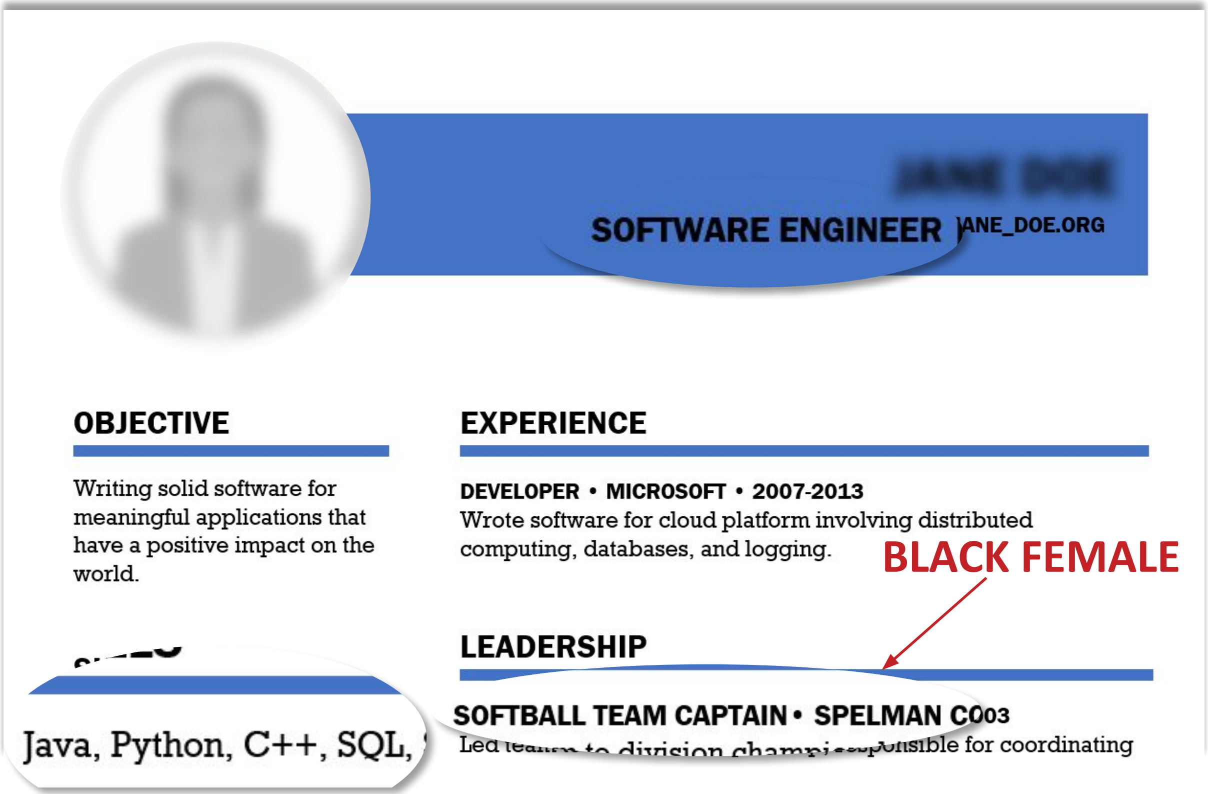
[FAT*'19]

Computer Programmer



Computer Programmer

[FAT*'19]



[FAT*'19]

Microsoft Research

LinkedIn

Bias in Bios dataset

- 400k online biographies* through Common Crawl.

"*Xxx Xxx is a(n) (xxx) [title]...he/she...*" *title* ∈ {common BLS SOC titles}

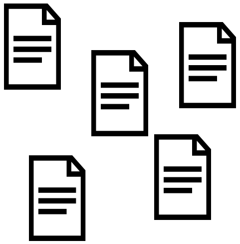
Alexandra Chouldechova is an Assistant Professor of Statistics and Public Policy at Carnegie Mellon University's Heinz College of Informations Systems and Public Policy. She received her B.Sc. from the University of Toronto in 2009, and in 2014 she completed her Ph.D. in Statistics at Stanford University. While at Stanford, she also worked at Google and Symantec on developing statistical assessment methods for information retrieval systems.

Classification problem: classify title to one of **28** categories using biography text

*Code to reproduce dataset publicly available: aka.ms/biasbios

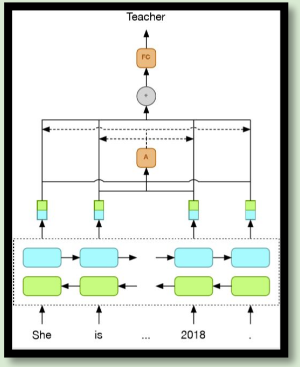
Learning pipeline

Input data:
Biographies



Semantic representation + learning algorithm:

1. Bag-of-words + Logit
2. Word embedding + Logit
3. Word embedding + DNN



Objective:
Predict Y = *Occupation*

Biases in bios

[FAT*'19]

Enter the bio

She is a fifth year PhD student in the joint Machine Learning and Public Policy program at Carnegie Mellon University's Machine Learning Department and Heinz College. She is co-advised by Prof. Artur Dubrawski and Prof. Alexandra Chouldechova, and she is part of the Auton Lab.

Currently, her main focus is algorithmic fairness, studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support. She is passionate about developing novel machine learning algorithms that are

PREDICT TITLE

SHE

HE

she is a fifth year phd student in the joint machine learning and public policy program at carnegie mellon university <unk> s machine learning department and heinz college . she is co-advised by prof. artur <unk> and prof. alexandra chouldechova , and she is part of the auton lab . currently , her main focus is algorithmic fairness , studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support . she is passionate about developing novel machine learning algorithms that are motivated by existing policy problems , and understanding how machine learning can better help us overcome important societal challenges . prior to graduate school she received her b.sc . in mathematics from universidad nacional de colombia and worked as a journalist for one of colombia <unk> s main news magazine , semana . she is the recipient of a microsoft

teacher

Biases in bios

[FAT*'19]

How do predictions change if explicit gender indicators are swapped?

Currently, her main focus is algorithmic fairness, studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support. She is passionate about developing novel machine learning algorithms that are

PREDICT TITLE

SHE HE

she is a fifth year phd student in the joint machine learning and public policy program at carnegie mellon university <unk> s machine learning department and heinz college . she is co-advised by prof. artur <unk> and prof. alexandra chouldechova , and she is part of the auton lab . currently , her main focus is algorithmic fairness , studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support . she is passionate about developing novel machine learning algorithms that are motivated by existing policy problems , and understanding how machine learning can better help us overcome important societal challenges . prior to graduate school she received her b.sc . in mathematics from universidad nacional de colombia and worked as a journalist for one of colombia <unk> s main news magazine , semana . she is the recipient of a microsoft

teacher

Biases in bios

[FAT*'19]

She → he

Enter the bio

He is a fifth year PhD student in the joint Machine Learning and Public Policy program at Carnegie Mellon University's Machine Learning Department and Heinz College. He is co-advised by Prof. Artur Dubrawski and Prof. Alexandra Chouldechova, and he is part of the Auton Lab.

Currently, his main focus is algorithmic fairness, studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support. He is passionate about developing novel machine learning algorithms that are motivated

PREDICT TITLE

SHE | HE

he is a fifth year phd student in the joint machine learning and public policy program at carnegie mellon university <unk> s machine learning department and heinz college . he is co-advised by prof. artur <unk> and prof. alexandra chouldechova , and he is part of the auton lab . currently , his main focus is algorithmic fairness , studying how to measure and prevent bias and discrimination that may arise when using machine learning for decision support . he is passionate about developing novel machine learning algorithms that are motivated by existing policy problems , and understanding how machine learning can better help us overcome important societal challenges . prior to graduate school he received his b.sc . in mathematics from universidad nacional de colombia and worked as a journalist for one of colombia <unk> s main news magazine , semana . he is the recipient of a microsoft

software_engineer

Biases in bios

[FAT*'19]

14.7% of female rappers who would only be correctly classified if using male gender pronouns, are mistakenly classified as models.

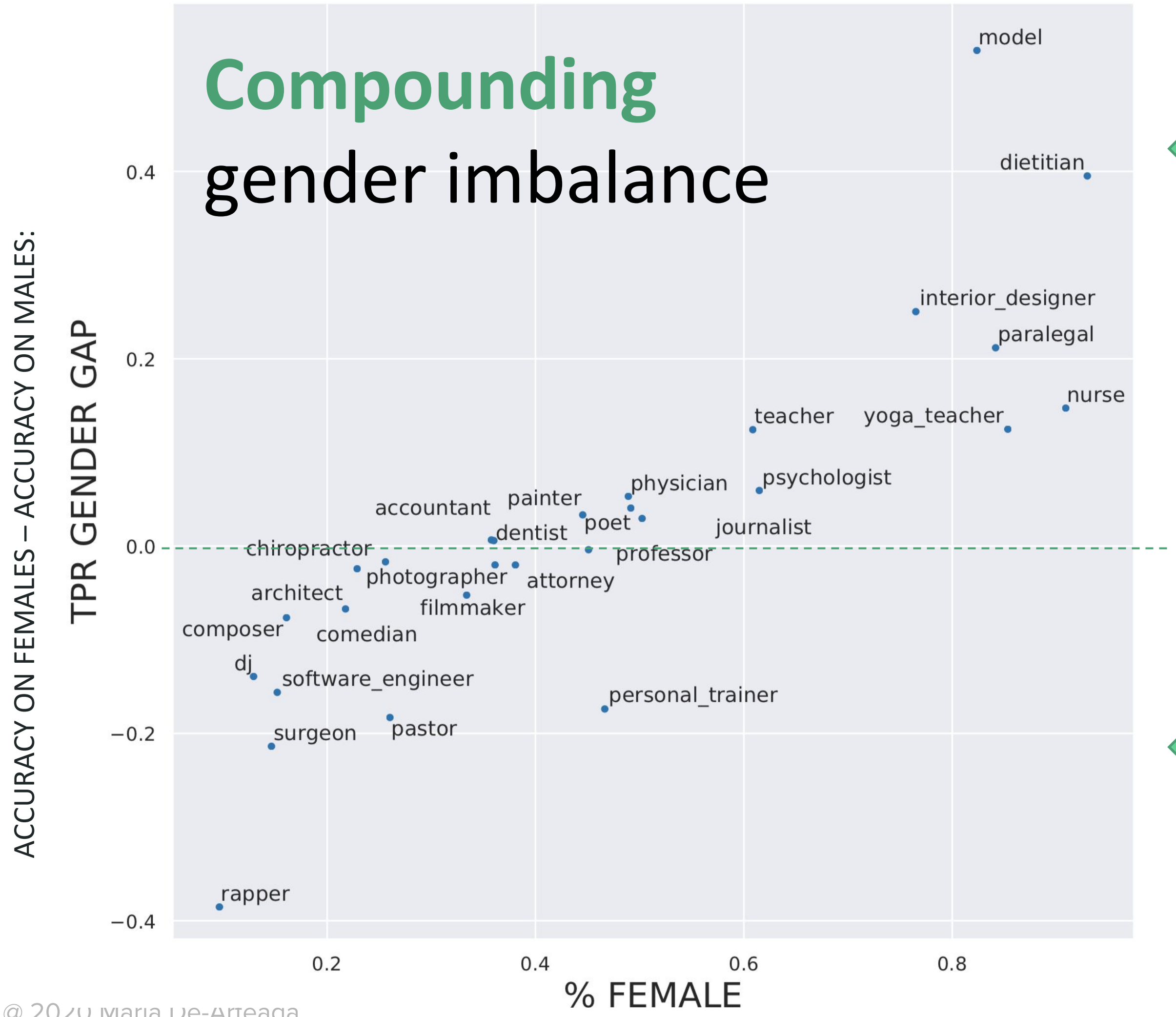
Women			Men		
y^1	y^2	$\Pi_{\text{female},(y^1,y^2)}$	y^1	y^2	$\Pi_{\text{male},(y^1,y^2)}$
model	rapper	14.7%	attorney	paralegal	7.1%
teacher	pastor	8.5%	architect	interior designer	4.7%
professor	software engineer	6.5%	professor	dietitian	4.3%
professor	surgeon	4.8%	photographer	interior designer	3.5%
physician	surgeon	3.8%	teacher	yoga teacher	3.3%

(**incorrect**) prediction with **own** pronouns

(**correct**) prediction with **swapped** pronouns

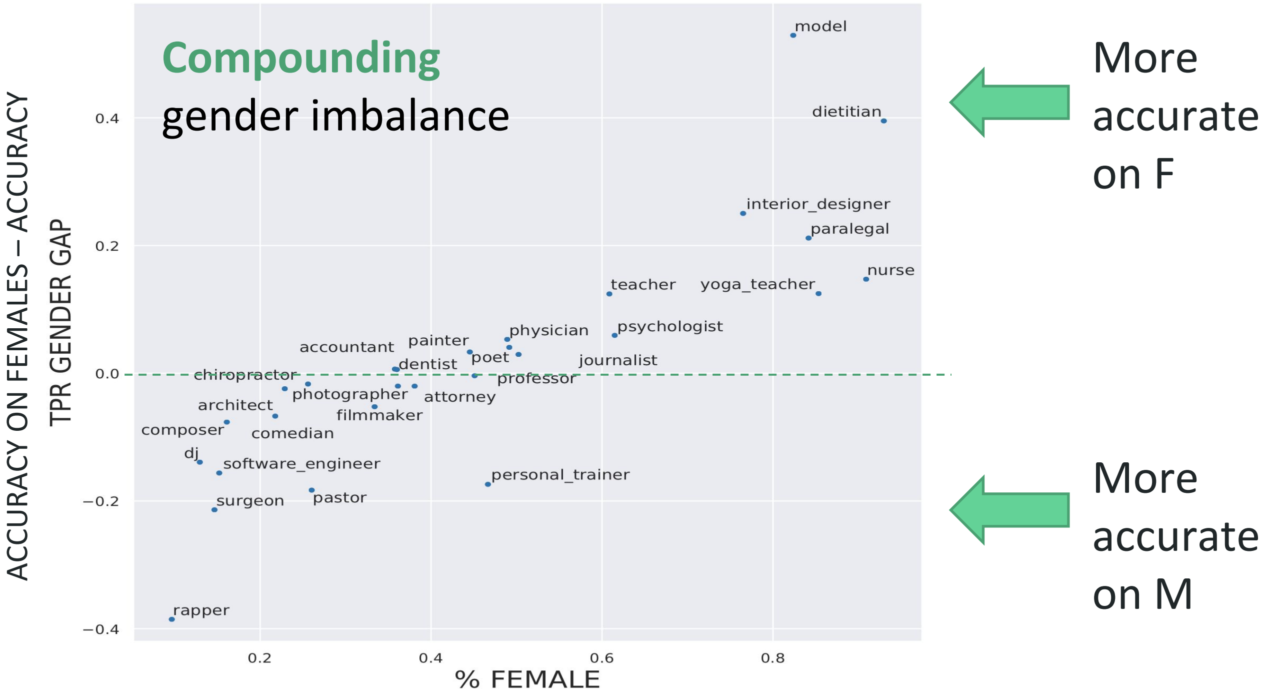
software_engineer

[FAT*'19]



More accurate on F

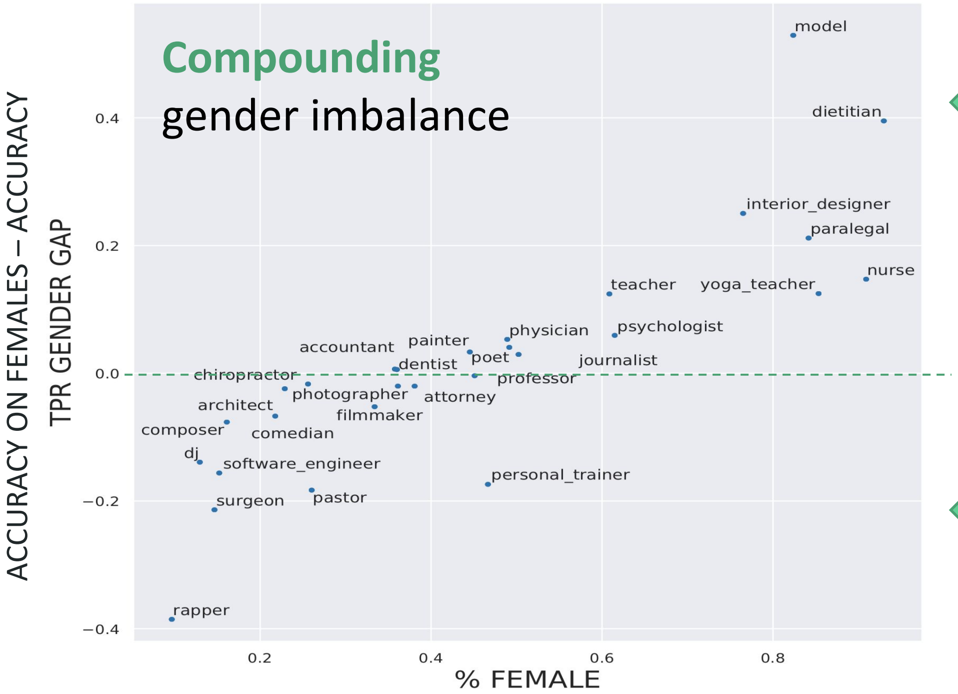
More accurate on M



Compounding imbalance

Theorem [**FAT*'19**]*:* If female fraction $p < 0.5$ and gender gap < 0, then female fraction in true positives $< p$.

(analogous for men)



More accurate on F

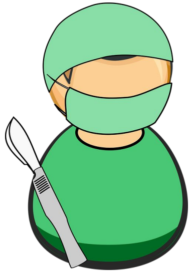
More accurate on M

[FAT*'19]

Compounding injustice [Hellman'18]

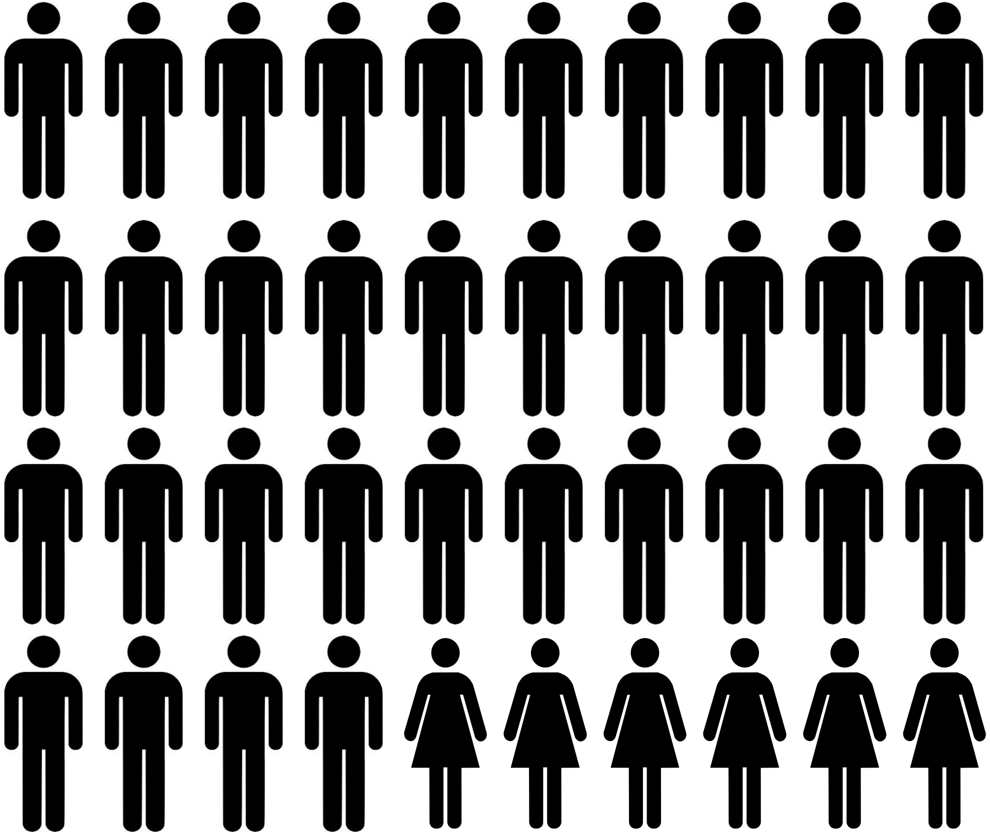
If initial imbalance constitutes injustice: Model's prediction is informed by, and compounds, previous injustice.

Compounding imbalances



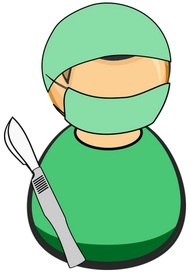
Surgeons

females in data:
14.6%



[FAT*'19]

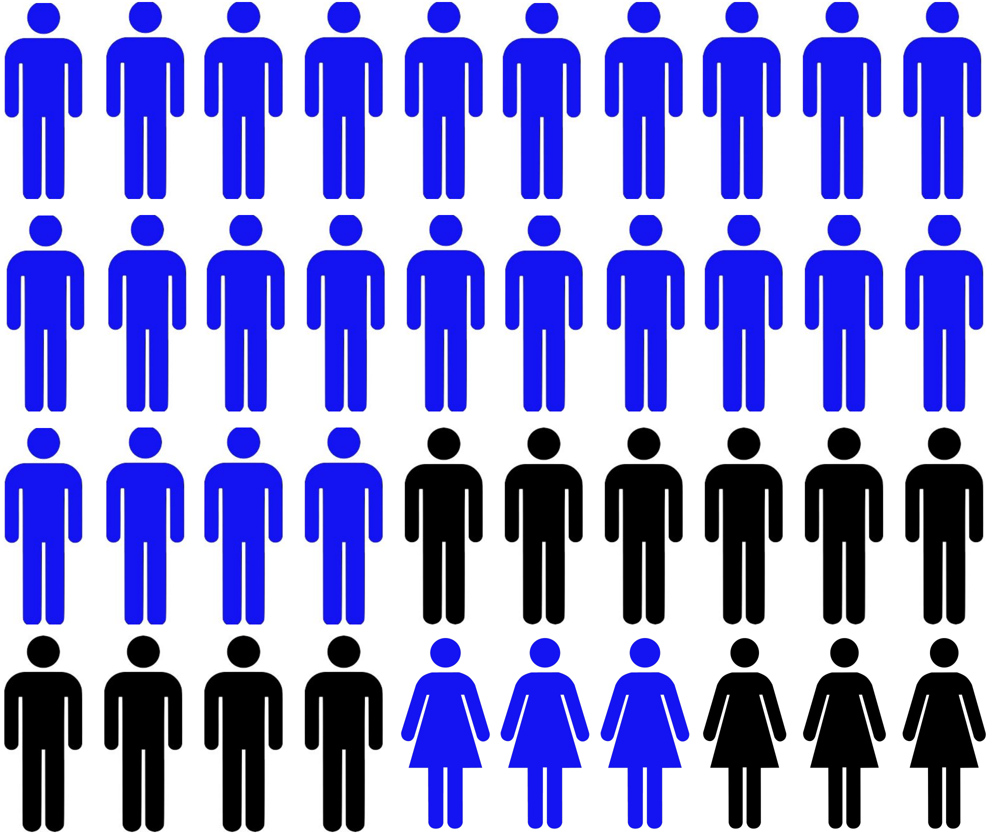
Compounding imbalances



Surgeons

females in data:
14.6%

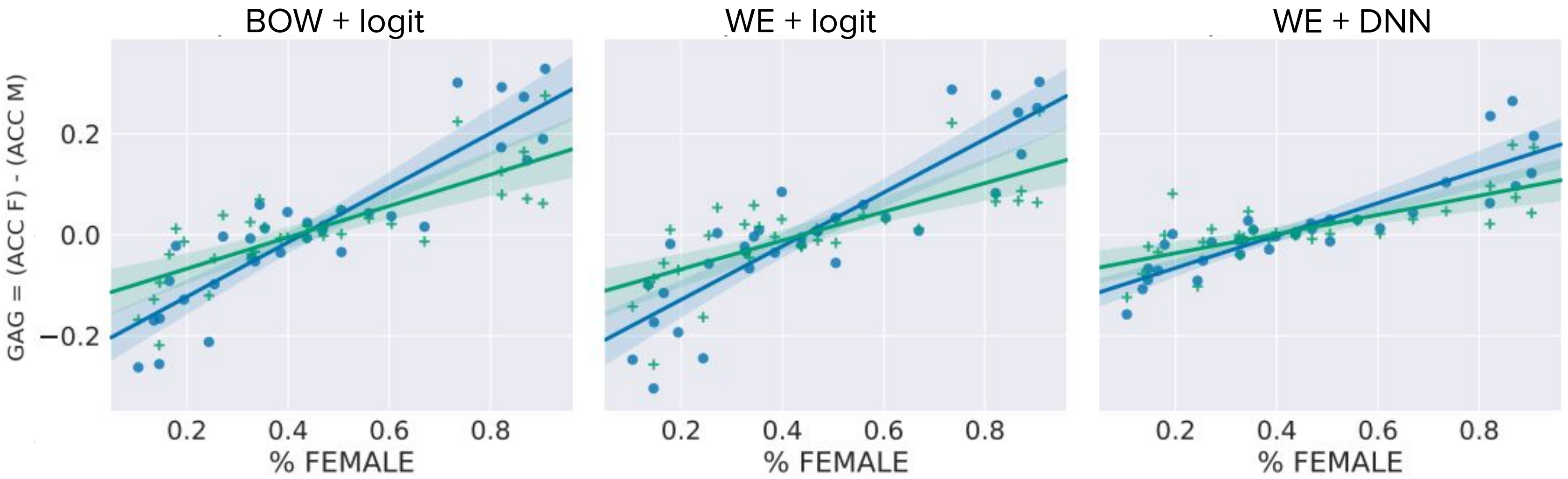
females in true positives:
11.6%



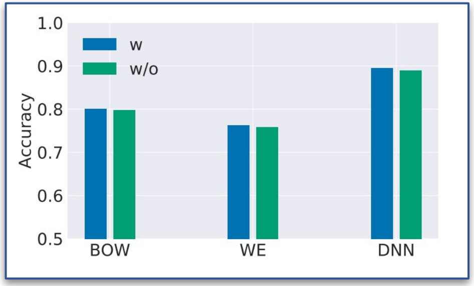
Males:
71% recall

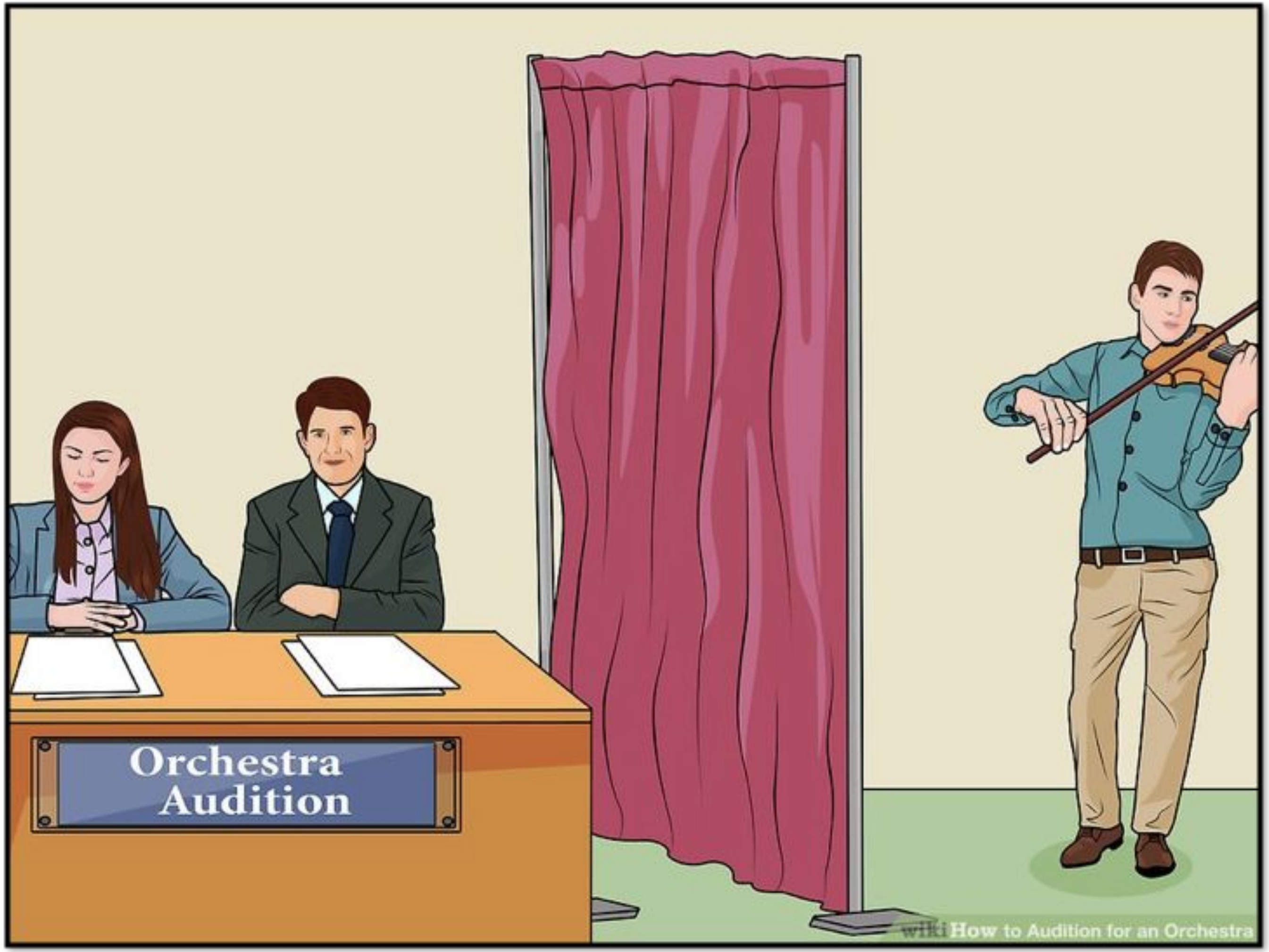
Females:
54% recall

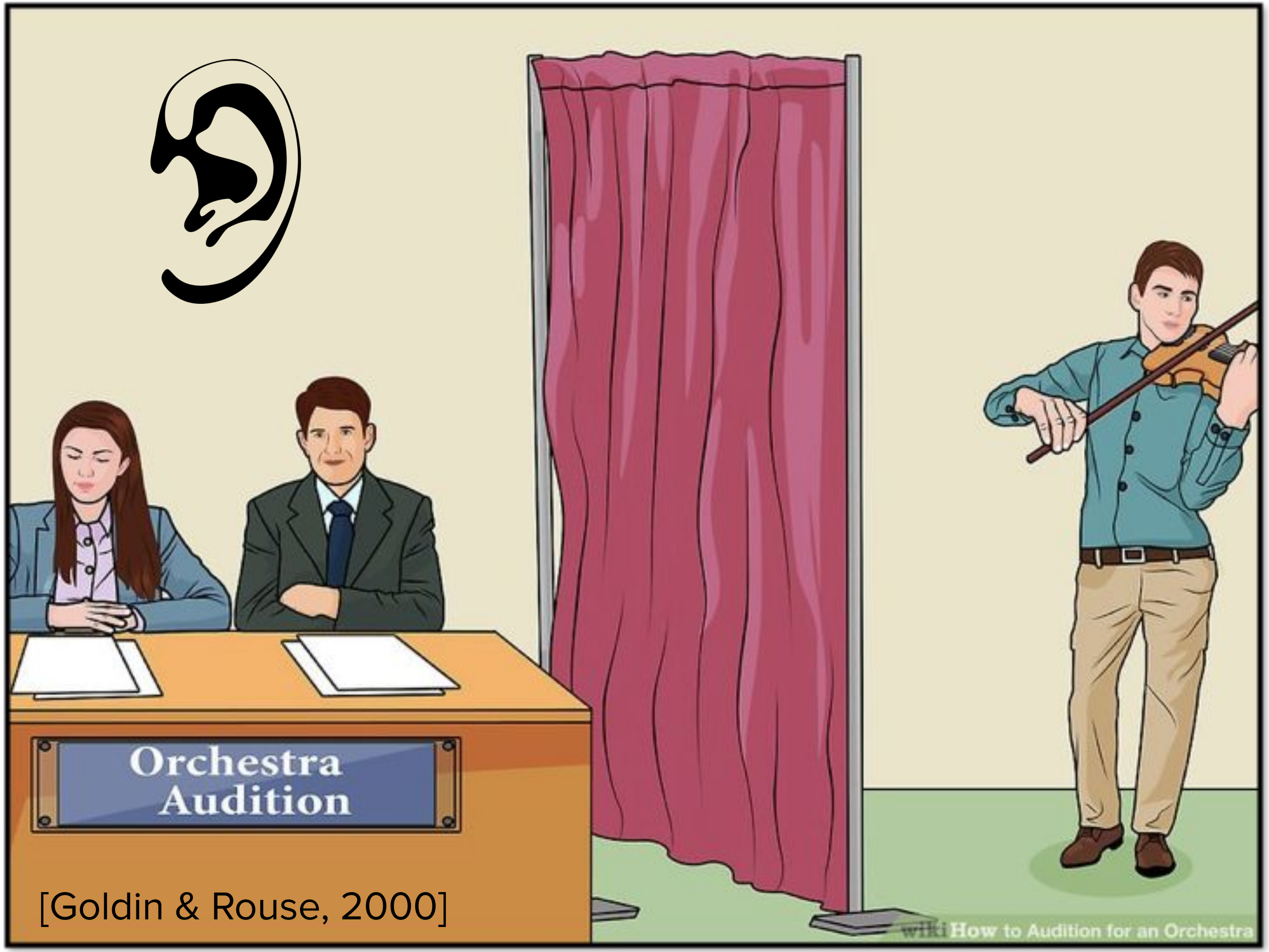
Removing explicit gender indicators not enough



with gender ind.
w/o gender ind.



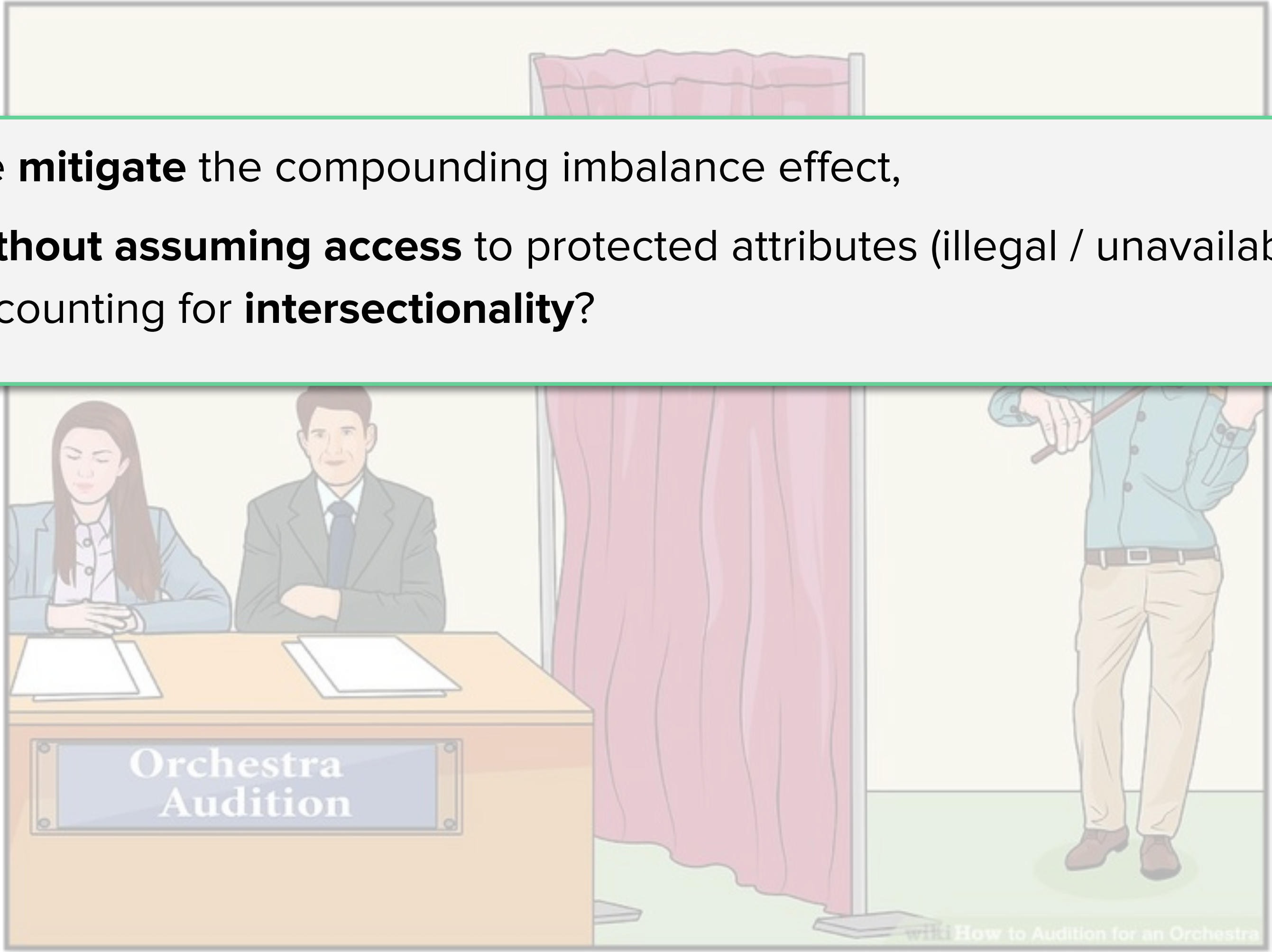

Orchestra
Audition
wiki How to Audition for an Orchestra



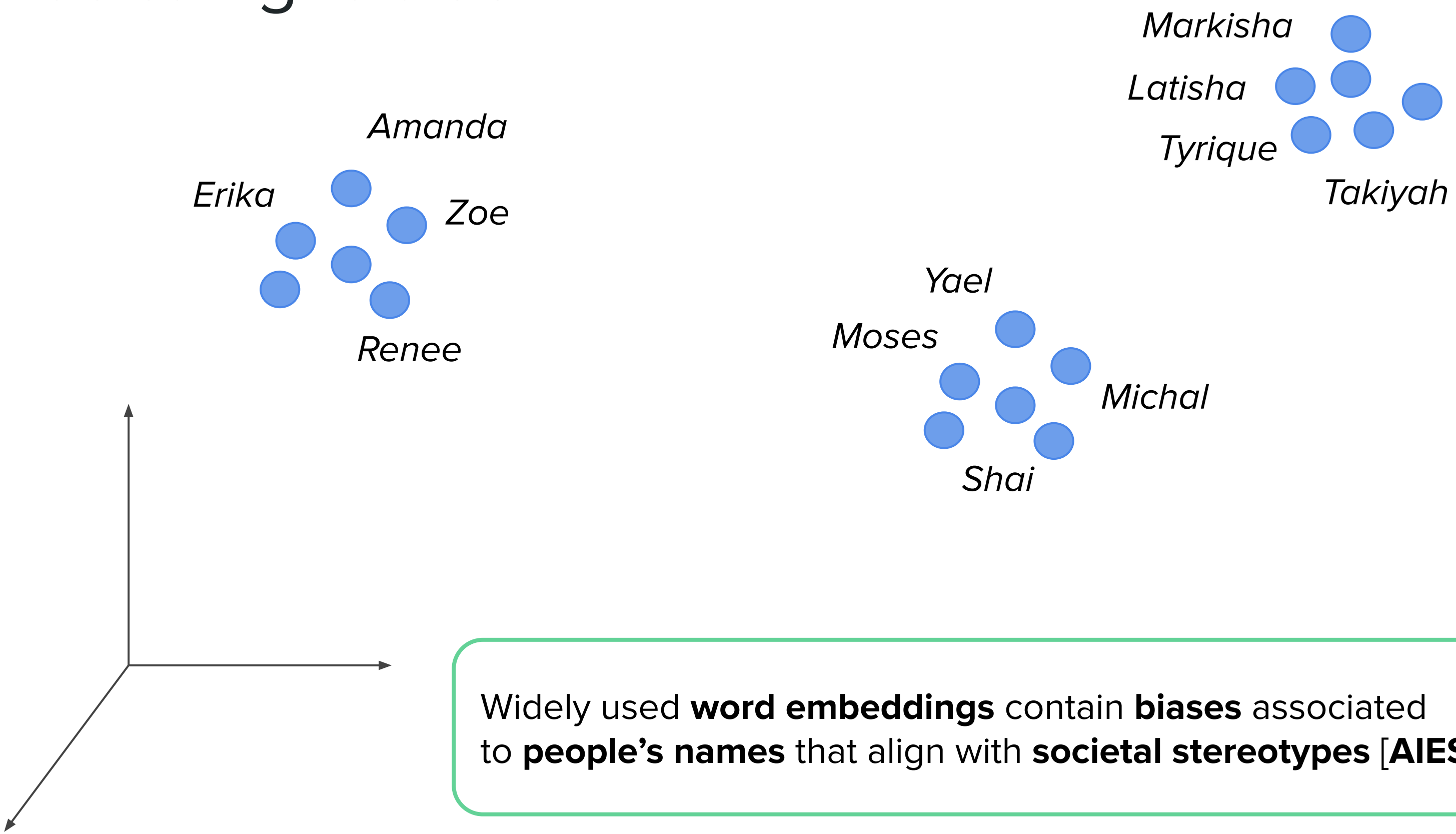
[Goldin & Rouse, 2000]

Can we **mitigate** the compounding imbalance effect,

- **without assuming access** to protected attributes (illegal / unavailable)?
- accounting for **intersectionality**?



Bias to fight bias?



Widely used **word embeddings** contain **biases** associated to **people's names** that align with **societal stereotypes** [**AIES'19**]

Outline

1. Risks of compounding injustices

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (FAT* 2019)
Maria De-Arteaga (CMU), Alexey Romanov (UMASS), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Adam Kalai (MSR)

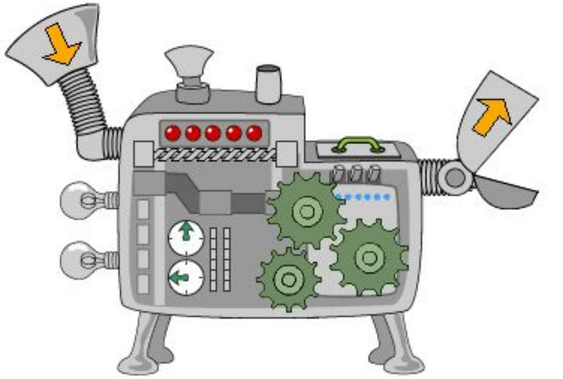
2. What are the biases in my word embedding?

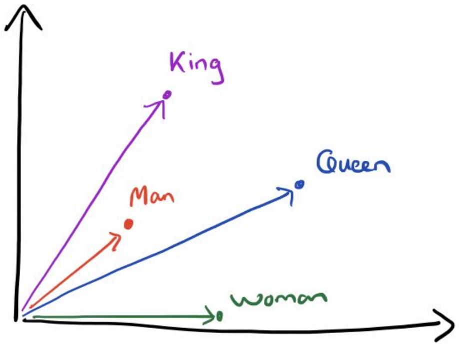
What are the biases in my word embedding? (AIES 2019)
Nathaniel Swinger[=] (Lexington HS), Maria De-Arteaga[=] (CMU), Neil Thomas Heffernan IV (Shrewsbury HS), Mark Leiserson (UMD), Adam Kalai (MSR)

3. Using bias to fight bias

What's in a Name? Reducing Bias in Bios without Access to Protected Attributes (NAACL 2019)
Alexey Romanov (UMASS), Maria De-Arteaga (CMU), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Anna Rumshisky (UMASS), Adam Kalai (MSR)

Word embeddings

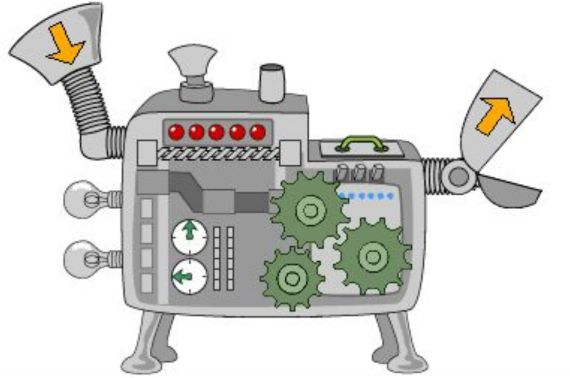


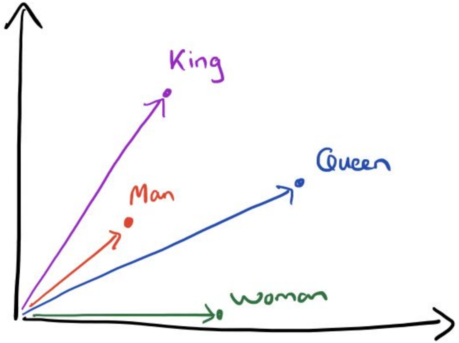


	A	B	C	D	E	F	G	H	I	J	K	L
1	the	0.056	0.043	0.051	0.08	0.006	0.041	0.032	0.011	0.057	0.004	0.08
2	cat	0.072	0.076	0.1	0.085	0.055	0.082	0.058	0.017	0.011	0.062	0.02
3	dog	0.088	0.099	0.028	0.059	0.06	0.059	0.039	0.09	0.001	0.031	0.07
4	nurse	0.03	0.018	0.058	0.074	0.055	0.028	0.025	0.054	0.094	0.052	0.09
5	doctor	0.097	0.093	0.035	0.057	0.044	0.052	0.046	0.055	0.072	0.055	0.0
6	king	0.013	0.059	0.024	0.032	0.038	0.078	0.052	0.067	0.05	0.087	0.03
7	queen	0.087	0.072	0.029	0.042	0.05	0.083	0.095	0.012	0.098	0.009	0.07
8	bird	0.042	0.044	0.006	0.003	0.003	0.082	0.034	0.024	0.003	0.05	0.0

Slide created by Adam Kalai

Word embeddings





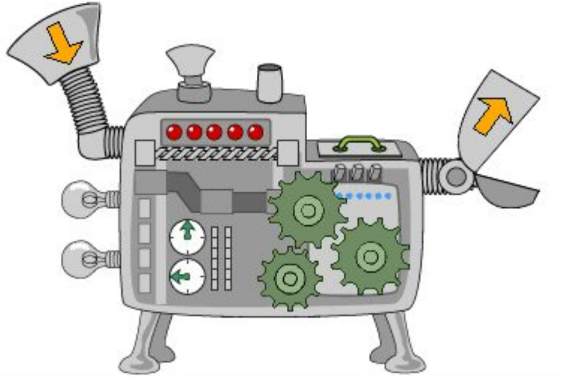
Man :: computer programmer

Woman :: ?

	A	B	C	D	E	F	G	H	I	J	K	L
1	the	0.056	0.043	0.051	0.08	0.006	0.041	0.032	0.011	0.057	0.004	0.08
2	cat	0.072	0.076	0.1	0.085	0.055	0.082	0.058	0.017	0.011	0.062	0.02
3	dog	0.088	0.099	0.028	0.059	0.06	0.059	0.039	0.09	0.001	0.031	0.07
4	nurse	0.03	0.018	0.058	0.074	0.055	0.028	0.025	0.054	0.094	0.052	0.09
5	doctor	0.097	0.093	0.035	0.057	0.044	0.052	0.046	0.055	0.072	0.055	0.0
6	king	0.013	0.059	0.024	0.032	0.038	0.078	0.052	0.067	0.05	0.087	0.03
7	queen	0.087	0.072	0.029	0.042	0.05	0.083	0.095	0.012	0.098	0.009	0.07
8	bird	0.042	0.044	0.006	0.003	0.003	0.082	0.034	0.024	0.003	0.05	0.0

Slide created by Adam Kalai

Word embeddings

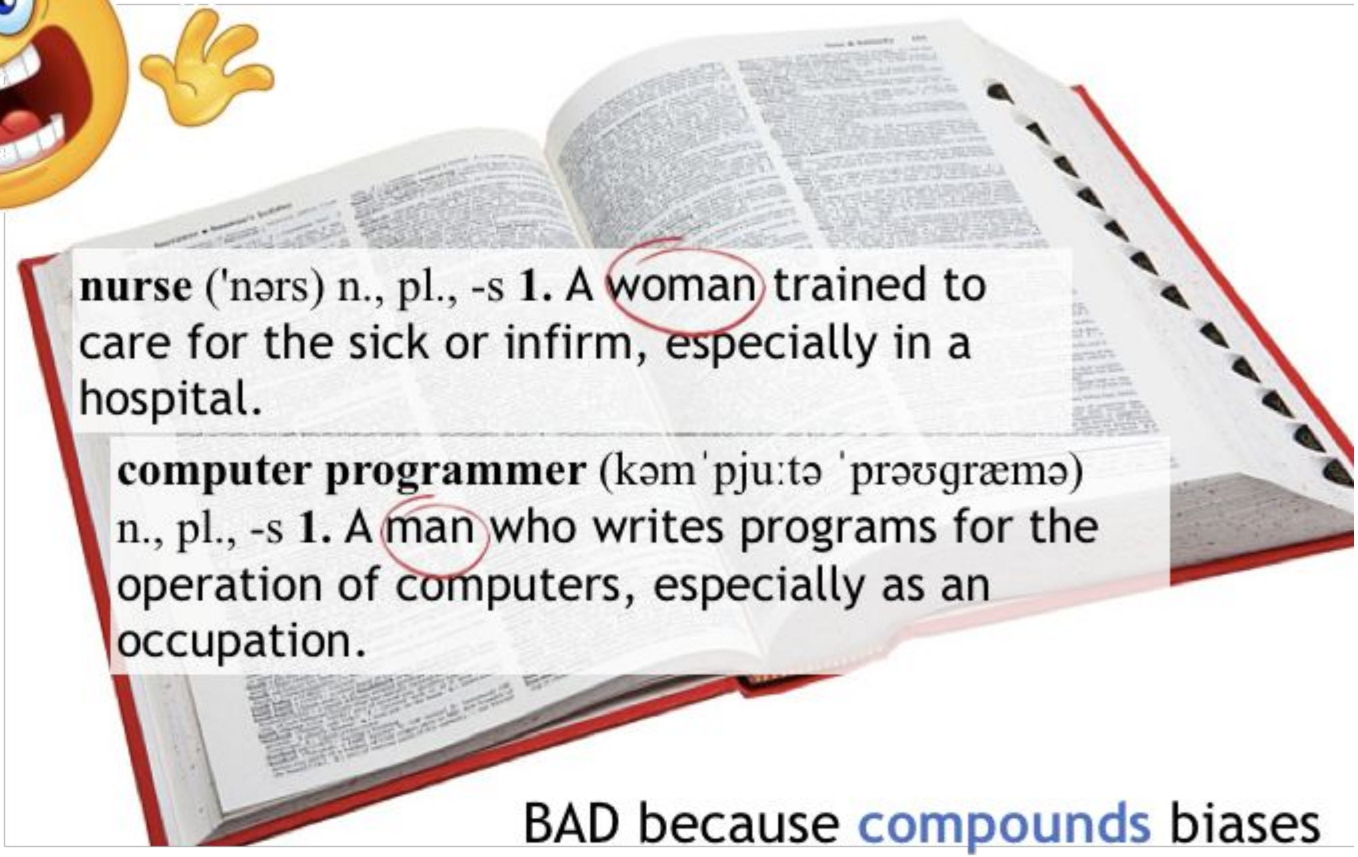


Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

	A	B	C	D	E	F	G	H	I	J	K	L
1	the	0.056	0.043	0.051	0.08	0.006	0.041	0.032	0.011	0.057	0.004	0.08
2	cat	0.072	0.076	0.1	0.085	0.055	0.082	0.058	0.017	0.011	0.062	0.02
3	dog	0.088	0.099	0.028	0.059	0.06	0.059	0.039	0.09	0.001	0.031	0.07
4	nurse	0.03	0.018	0.058	0.074	0.055	0.028	0.025	0.054	0.094	0.052	0.09
5	doctor	0.097	0.093	0.035	0.057	0.044	0.052	0.046	0.055	0.072	0.055	0.0
6	king	0.013	0.059	0.024	0.032	0.038	0.078	0.052	0.067	0.05	0.087	0.03
7	queen	0.087	0.072	0.029	0.042	0.05	0.083	0.095	0.012	0.098	0.009	0.07
8	bird	0.042	0.044	0.006	0.003	0.003	0.082	0.034	0.024	0.003	0.05	0.0

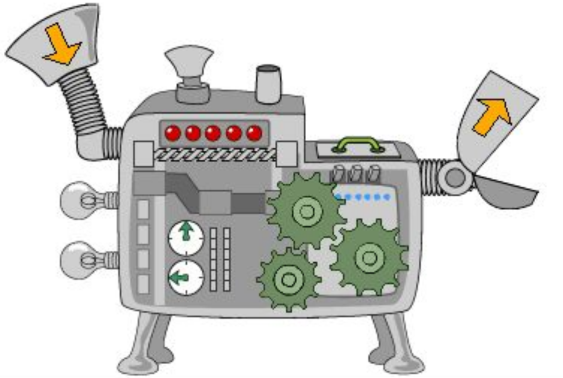
Slide created by Adam Kalai



BAD because compounds biases

Slide created by Adam Kalai

Word embeddings



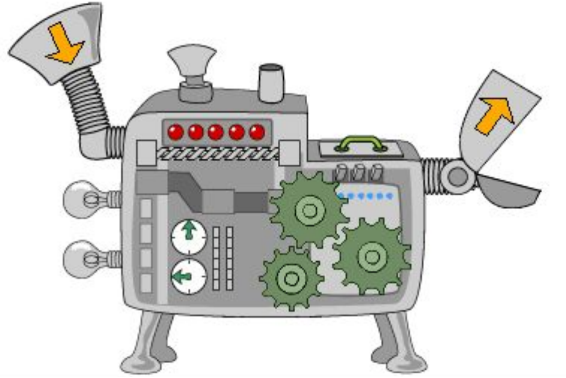
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

	A	B	C	D	E	F	G	H	I	J	K	L
1	the	0.056	0.000	0.051	0.08	0.006	0.041	0.032	0.011	0.057	0.004	0.08
2	cat	0.072	0.000	0.1	0.085	0.055	0.082	0.058	0.017	0.011	0.062	0.02
3	dog	0.088	0.000	0.028	0.059	0.06	0.059	0.039	0.09	0.001	0.031	0.07
4	nurse	0.03	0.000	0.058	0.074	0.055	0.028	0.025	0.054	0.094	0.052	0.09
5	doctor	0.097	0.000	0.035	0.057	0.044	0.052	0.046	0.055	0.072	0.055	0.0
6	king	0.013	0.000	0.024	0.032	0.038	0.078	0.052	0.067	0.05	0.087	0.03
7	queen	0.087	0.000	0.029	0.042	0.05	0.083	0.095	0.012	0.098	0.009	0.07
8	bird	0.042	0.000	0.006	0.003	0.003	0.082	0.034	0.024	0.003	0.05	0.0

Slide created by Adam Kalai

Word embeddings



What are the biases in my word embedding?

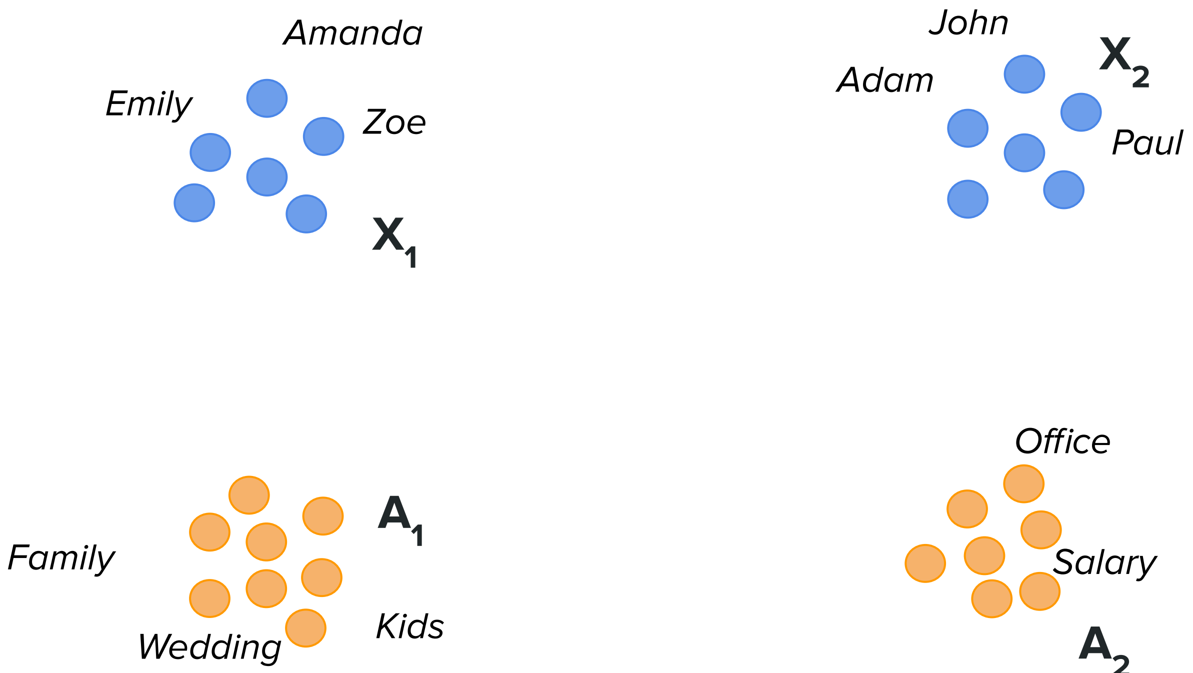
(beyond gender bias)

Slide created by Adam Kalai

Implicit Association Test

[Greenwald'98]

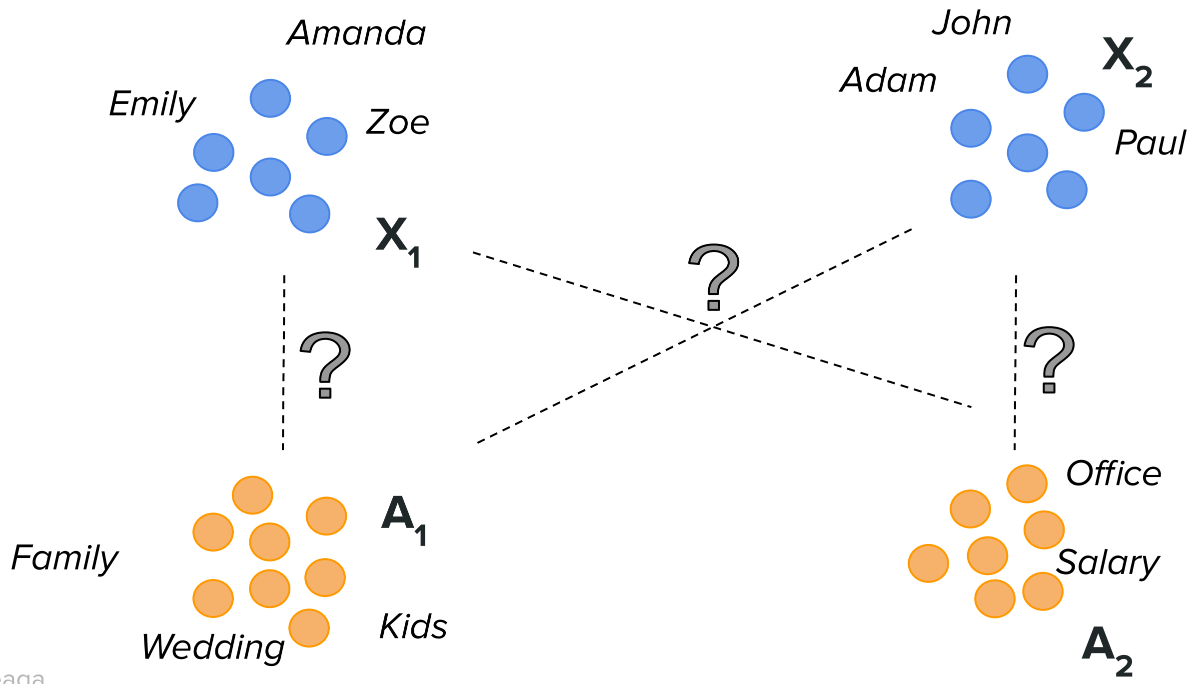
Implicit association between categories?



Implicit Association Test

[Greenwald'98]

Implicit association between categories?



Implicit Association Test

[Greenwald'98]

Female		Male
	Setting 1	
Career		Family

Implicit Association Test

[Greenwald'98]

Female

Career

Male

Family

Salary

Implicit Association Test

[Greenwald'98]

Female

Career

Male

Family

Paul

Implicit Association Test

[Greenwald'98]

Female

Career

Male

Family

Emily

Implicit Association Test

[Greenwald'98]

Female

Male

Career

Family

Wedding

Implicit Association Test

[Greenwald'98]

Female | Setting 2 | Male

Female

Setting 2

Male

Family

Career

Implicit Association Test

[Greenwald'98]

Female

Male

Family

Career

Salary

Implicit Association Test

[Greenwald'98]

Female

Family

Male

Career

Emily

Implicit Association Test

[Greenwald'98]

Female

Family

Male

Career

Wedding

Implicit Association Test

[Greenwald'98]

Female

Family

Male

Career

John

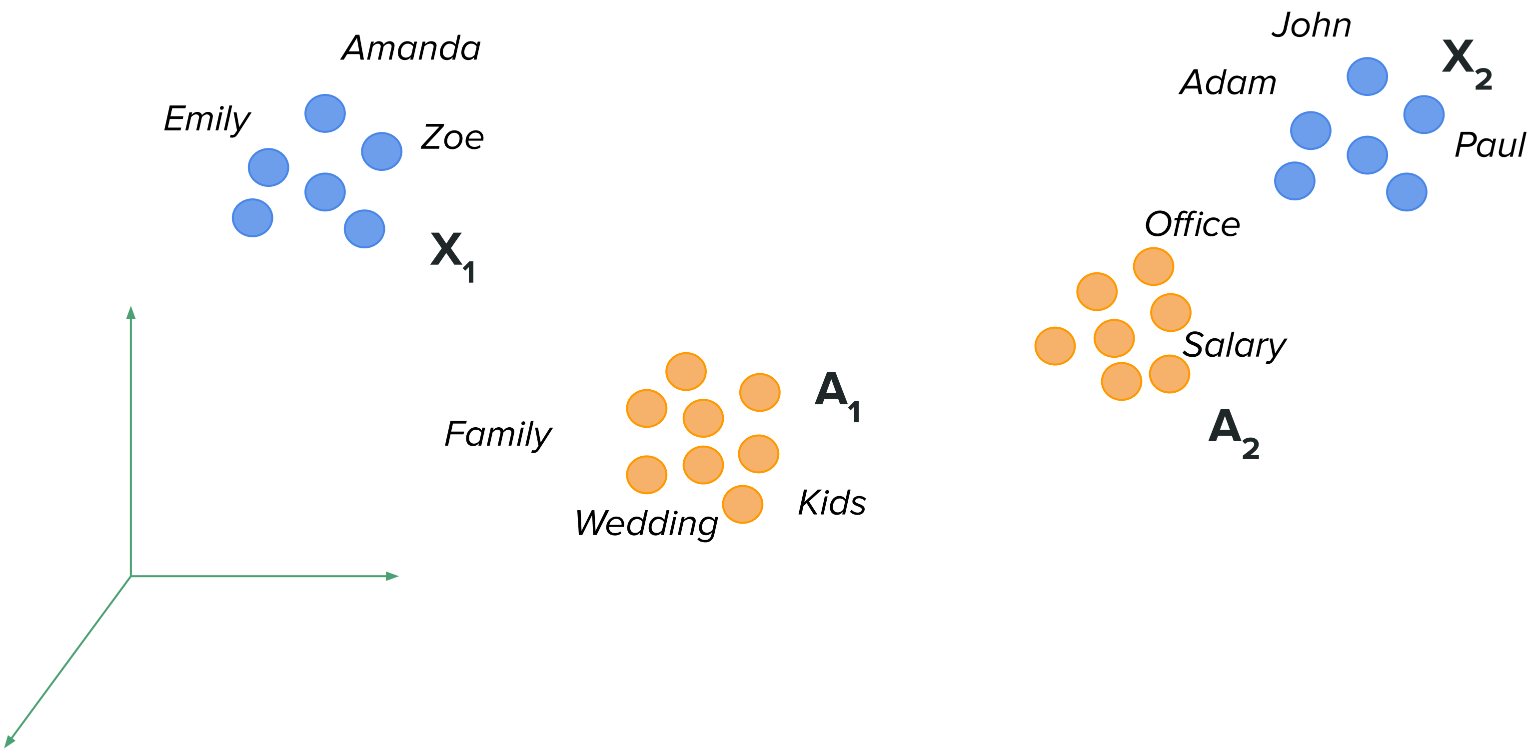
Implicit Association Test

[Greenwald'98]

Differences in average response time between **setting 1** and **setting 2**?

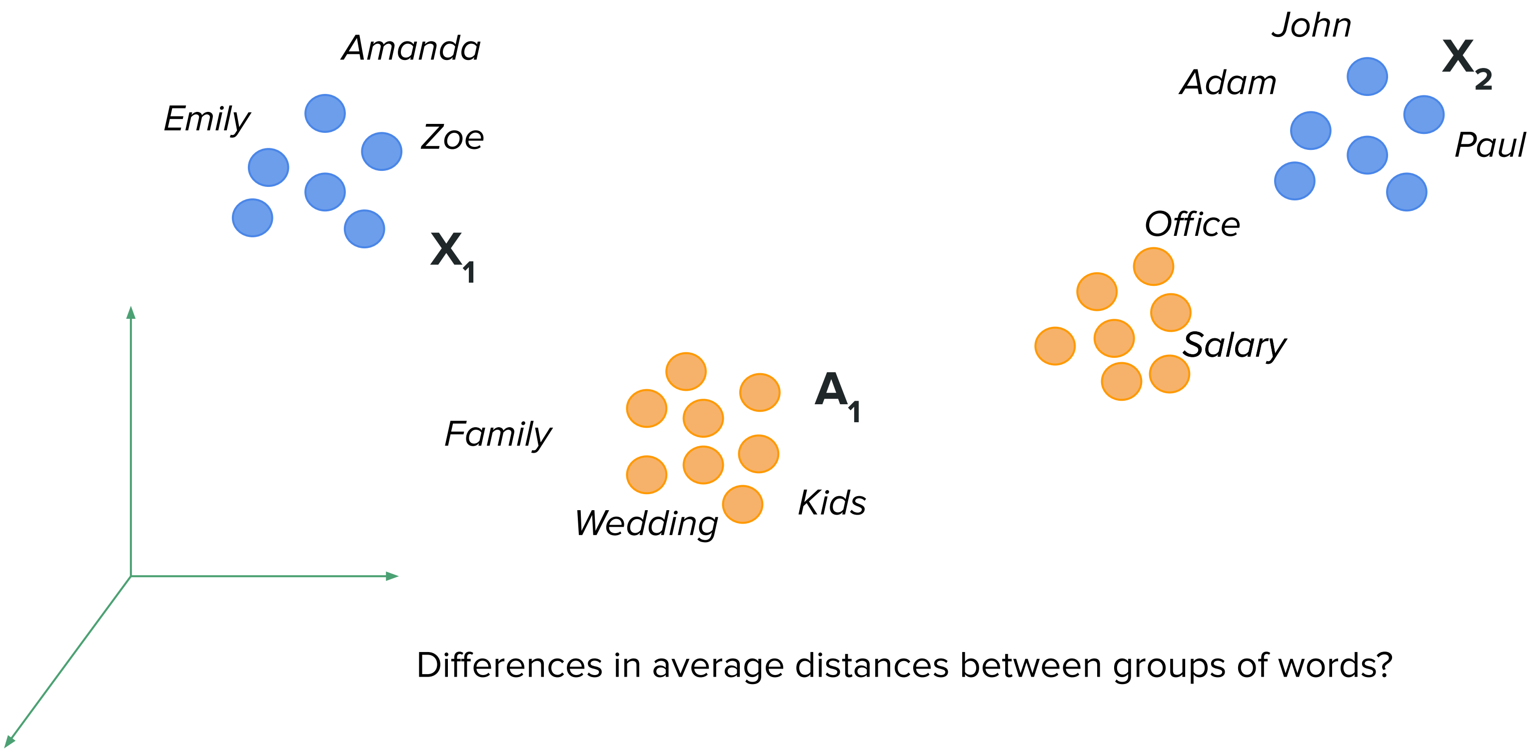
Word embedding Association Test

[Caliskan et al, 2017]



Word embedding Association Test

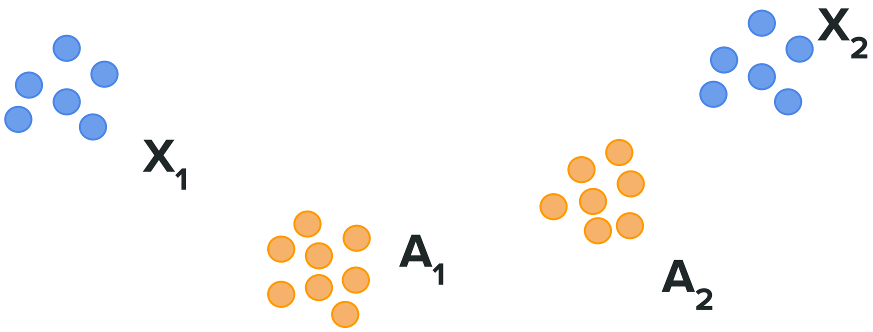
[Caliskan et al, 2017]



Differences in average distances between groups of words?

Word embedding Association Test

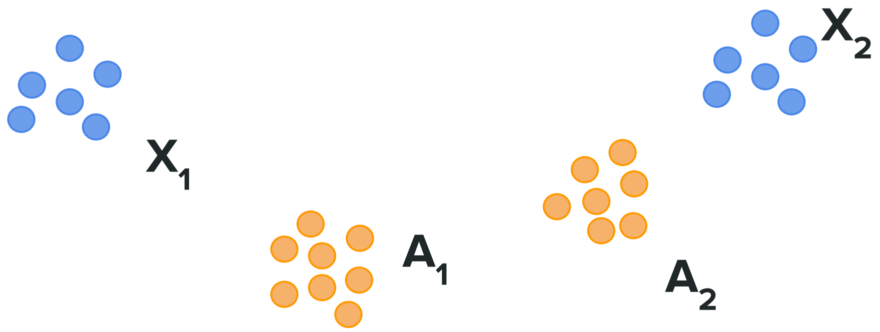
[Caliskan et al, 2017]



1. Which sets X_1, X_2, A_1, A_2 should we consider?
2. How to deal with the combinatorial explosion that arises when considering intersectional groups?

Word embedding Association Test

[Caliskan et al, 2017]



Is bias X in my word embedding?
[Caliskan'17]

What are the biases in my word embedding?
[Swinger* and De-Arteaga* et al, AIES, 2019]

Unsupervised bias enumeration

Generalized Word embedding Association Test

[Swinger and De-Arteaga* et al, 2018]*

$$g(X_1, A_1, \ldots, X_n, A_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \stackrel{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\boldsymbol{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

Generalized Word embedding Association Test

[Swinger and De-Arteaga* et al, 2018]*

$$g(X_1, A_1, \ldots, X_n, A_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \stackrel{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\boldsymbol{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

n=2 → **Lemma 1.** *For any $X_1, X_2 \subseteq \mathcal{X}$ of equal sizes $|X_1| = |X_2|$ and any nonempty $A_1, A_2 \subseteq \mathcal{A}$,*

$$s(X_1, A_1, X_2, A_2) = 2|X_1|\, g(X_1, A_1, X_2, A_2)$$

Generalized Word embedding Association Test

[Swinger and De-Arteaga* et al 2018]*

$$g(X_1, A_1, \ldots, X_n, A_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\overline{\boldsymbol{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\boldsymbol{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \stackrel{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\boldsymbol{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

Lemma 1. *For any $X_1, X_2 \subseteq \mathcal{X}$ of equal sizes $|X_1| = |X_2|$ and any nonempty $A_1, A_2 \subseteq \mathcal{A}$,*

$$s(X_1, A_1, X_2, A_2) = 2|X_1|\, g(X_1, A_1, X_2, A_2)$$

n=1 $\rightarrow$ **Lemma 2.** *For any nonempty sets $X \subset \mathcal{X}$, $A \subset \mathcal{A}$, let their complements sets $X^c = \mathcal{X} \setminus X$ and $A^c = \mathcal{A} \setminus A$. Then,*

$$g(X, A) = 2g(X, A, \mathcal{X}, \mathcal{A}) = 2\frac{|X^c|}{|\mathcal{X}|}\frac{|A^c|}{|\mathcal{A}|} g(X, A, X^c, A^c)$$

Generalized Word embedding Association Test

[Swinger and De-Arteaga* et al 2018]*

$$g(X_1, A_1, \ldots, X_n, A_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\overline{\mathbf{X}}_i - \boldsymbol{\mu}) \cdot (\overline{\mathbf{A}}_i - \overline{\boldsymbol{\mathcal{A}}})$$

$$\text{where } \boldsymbol{\mu} \stackrel{\text{def}}{=} \begin{cases} \overline{\boldsymbol{\mathcal{X}}} & \text{for } n = 1, \\ \sum_i \overline{\mathbf{X}}_i / n & \text{for } n \geq 2. \end{cases}$$

Lemma 1. *For any $X_1, X_2 \subseteq \mathcal{X}$ of equal sizes $|X_1| = |X_2|$ and any nonempty $A_1, A_2 \subseteq \mathcal{A}$,*

$$s(X_1, A_1, X_2, A_2) = 2|X_1|\, g(X_1, A_1, X_2, A_2)$$

Lemma 2. *For any nonempty sets $X \subset \mathcal{X}$, $A \subset \mathcal{A}$, let their complements sets $X^c = \mathcal{X} \setminus X$ and $A^c = \mathcal{A} \setminus A$. Then,*

$$g(X, A) = 2g(X, A, \mathcal{X}, \mathcal{A}) = 2\frac{|X^c|}{|\mathcal{X}|}\frac{|A^c|}{|\mathcal{A}|} g(X, A, X^c, A^c)$$

n>1
(decomposition)

Lemma 3. *For any $n > 1$ and nonempty $X_1, X_2, \ldots, X_n \subseteq \mathcal{X}$ and $A_1, A_2, \ldots, A_n \subseteq \overline{\boldsymbol{\mathcal{A}}}$,*

$$g(X_1, A_1, \ldots, X_n, A_n) = \sum_{i \in [n]} g(X_i, A_i) - \sum_{i,j \in [n]} \frac{g(X_i, A_j)}{n}$$

Unsupervised Bias Enumeration (UBE) algorithm

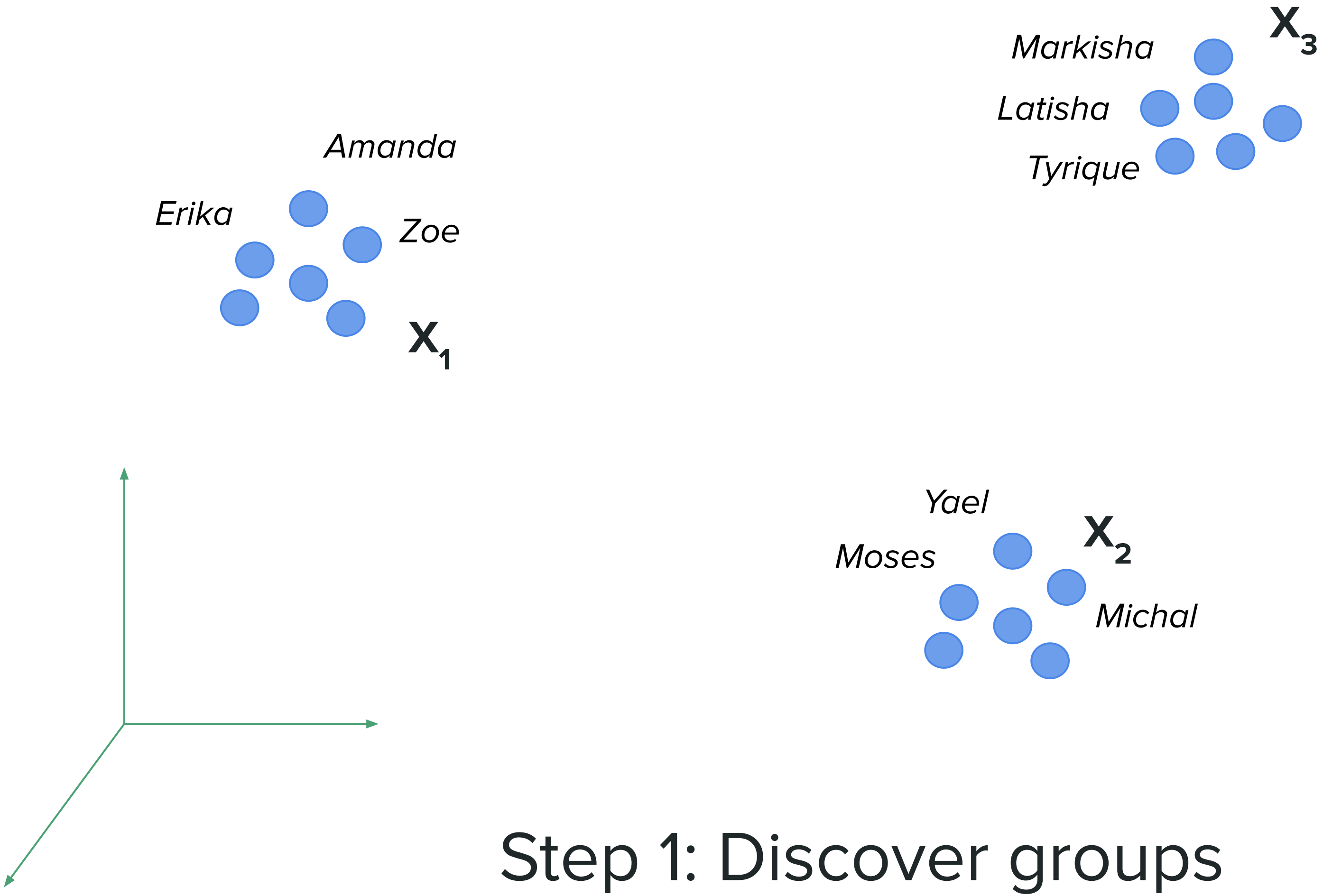
Input

name	meaning	default
WE	word embedding	`w2v`
$\mathcal{X}$	set of names	SSA
n	number of target groups	12
m	number of categories	64
M	number of frequent lower-case words	30,000
t	number of words per WEAT	3
α	false discovery rate	0.05

Attributes

Step 1: Discover groups

[AIES'19]



Step 1: Discover groups

[AIES'19]

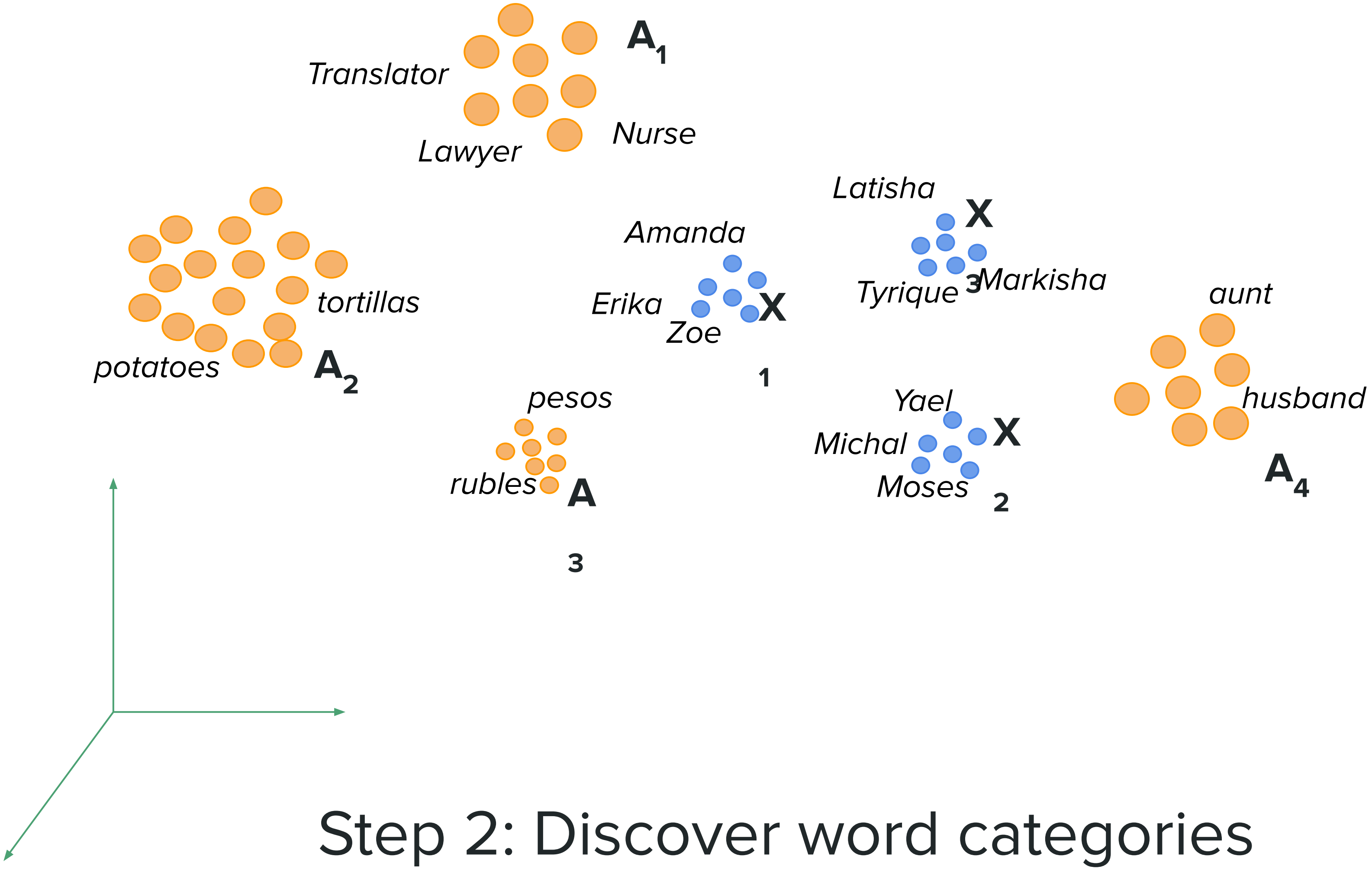
w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F10	w2v F11	w2v F12
Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael	Randall	Dashaun	Keith
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses	Dashiell	Jamell	Gabe
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal	Randell	Marlon	Alfred
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai	Jordan	Davonta	Shane
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis	Chace	Demetrius	Stan
+581	+840	+692	+558	+890	+312	+391	+577	+120	+432	+393	+494

Step 1: Discover groups

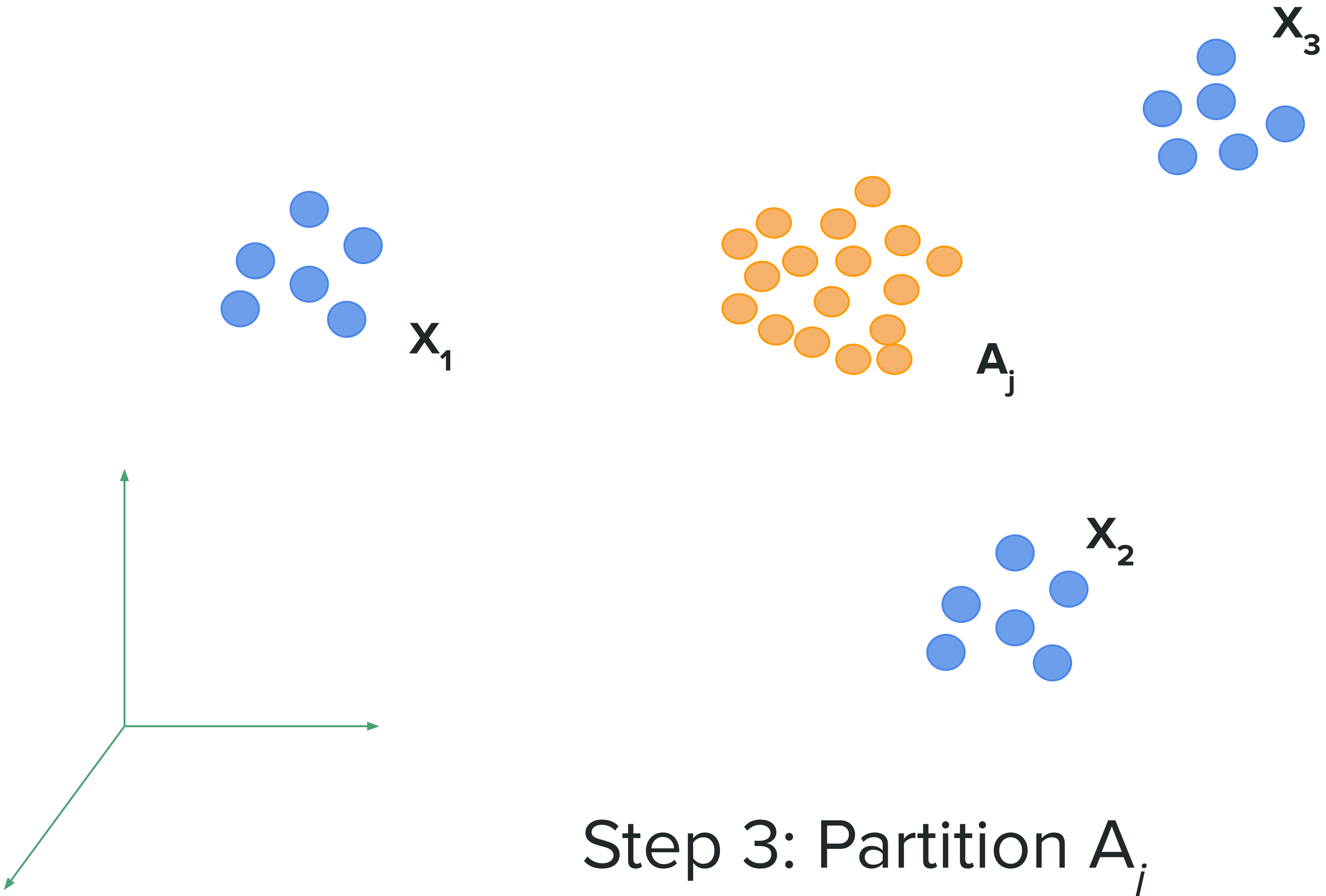
w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F10	w2v F11	w2v F12
Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael	Randall	Dashaun	Keith
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses	Dashiell	Jamell	Gabe
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal	Randell	Marlon	Alfred
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai	Jordan	Davonta	Shane
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis	Chace	Demetrius	Stan
+581	+840	+692	+558	+890	+312	+391	+577	+120	+432	+393	+494
98% F	98% F	89% F	85% F	78% F	65% F	59% F	56% F	40% F	27% F	5% F	4% F
1983	1968	1978	1982	1993	1991	1985	1986	1989	1981	1984	1976
4% B	8% B	48% B	10% B	2% B	7% B	4% B	2% B	5% B	10% B	32% B	6% B
4% H	4% H	3% H	9% H	1% H	4% H	9% H	70% H	10% H	3% H	5% H	3% H
3% A	3% A	1% A	11% A	1% A	32% A	4% A	8% A	5% A	4% A	3% A	5% A
89% W	84% W	47% W	69% W	95% W	56% W	83% W	21% W	79% W	83% W	59% W	86% W

Step 1: Discover groups

Step 2: Discover word categories

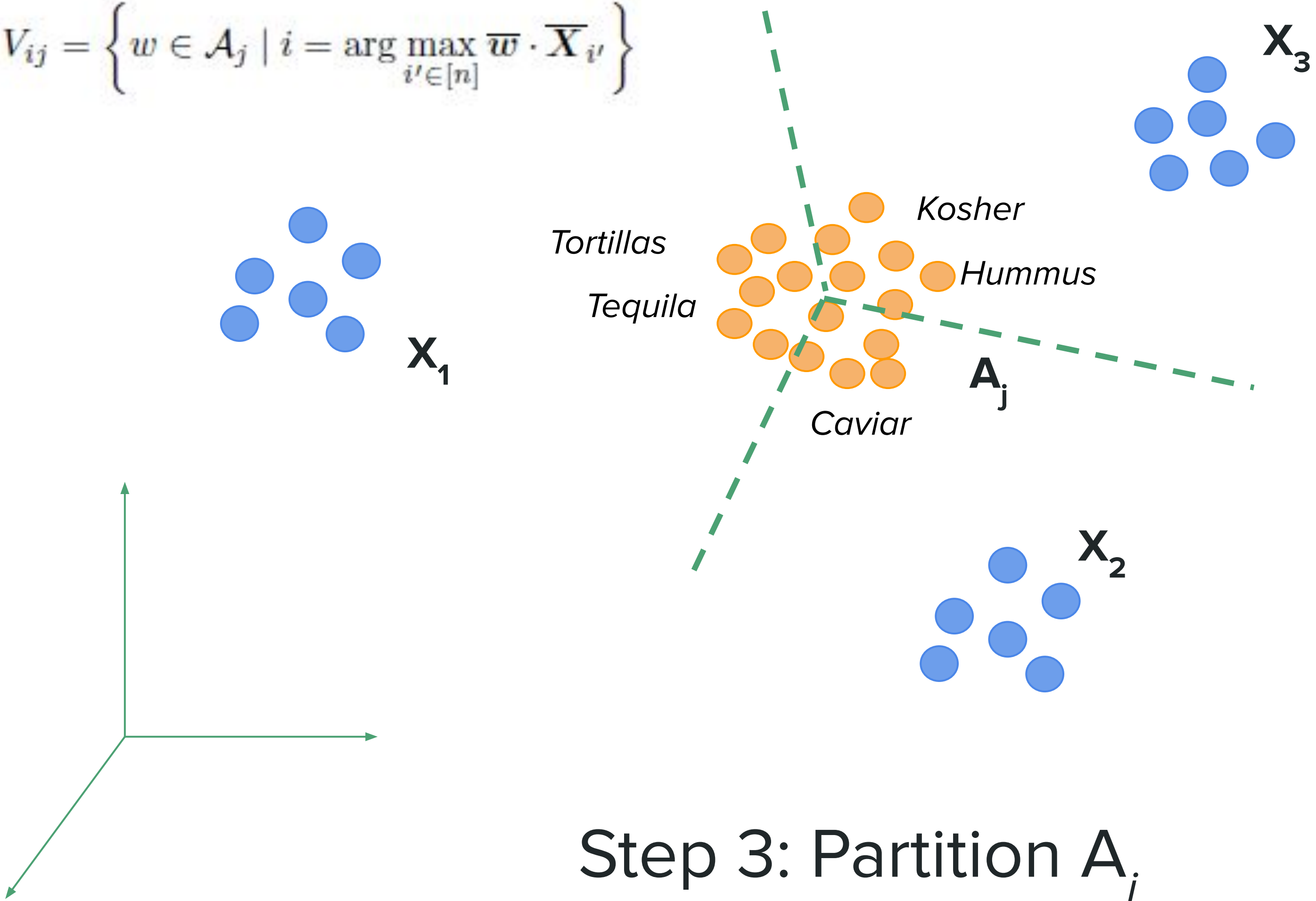


Step 2: Discover word categories



Step 3: Partition A_j

$$V_{ij} = \left\{ w \in \mathcal{A}_j \mid i = \arg\max_{i' \in [n]} \overline{w} \cdot \overline{X}_{i'} \right\}$$

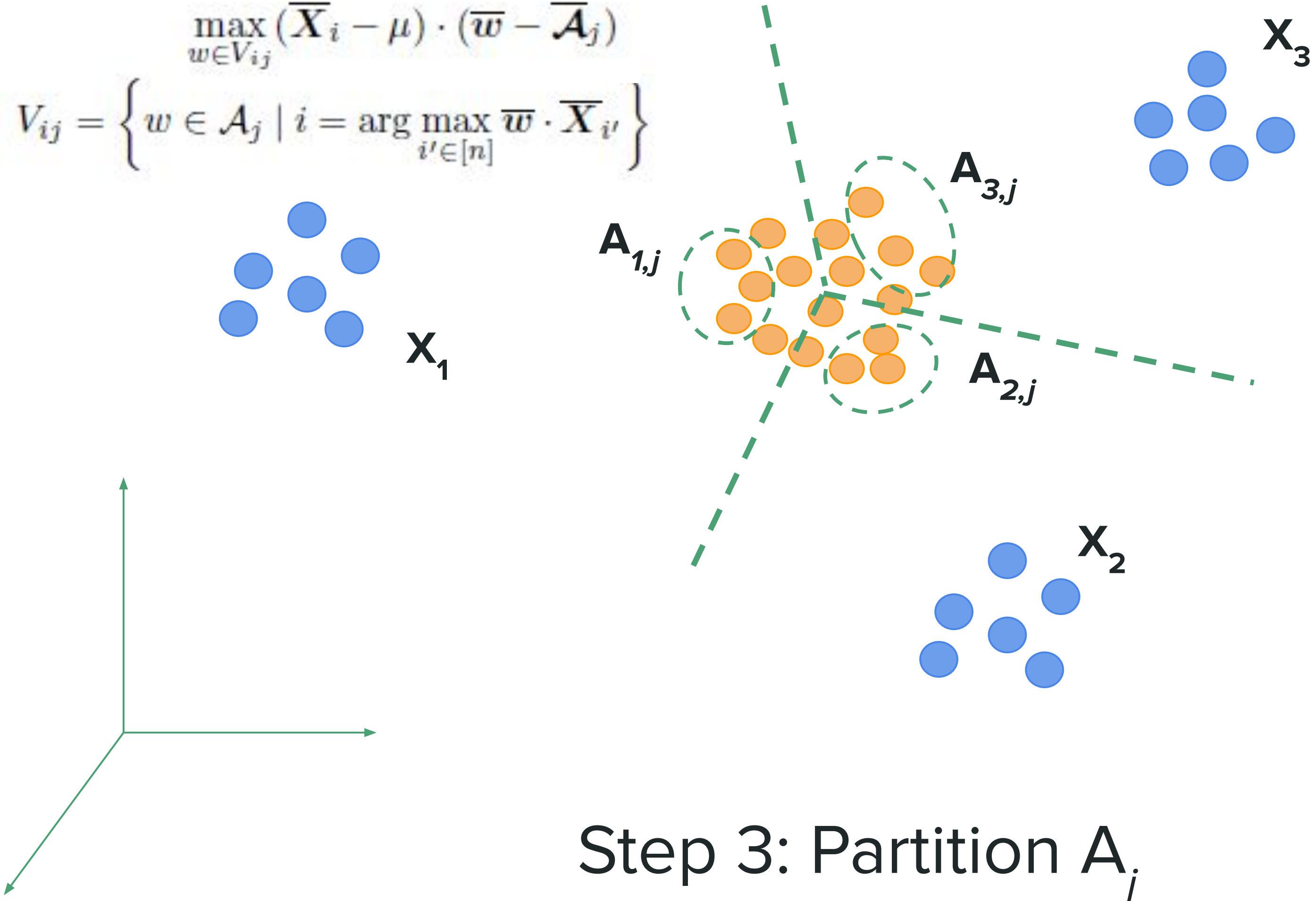


Step 3: Partition A_j

$A_{i,j}$ contains top *t* words s.t.

$$\max_{w \in V_{ij}} (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{w}} - \overline{\boldsymbol{\mathcal{A}}}_j)$$

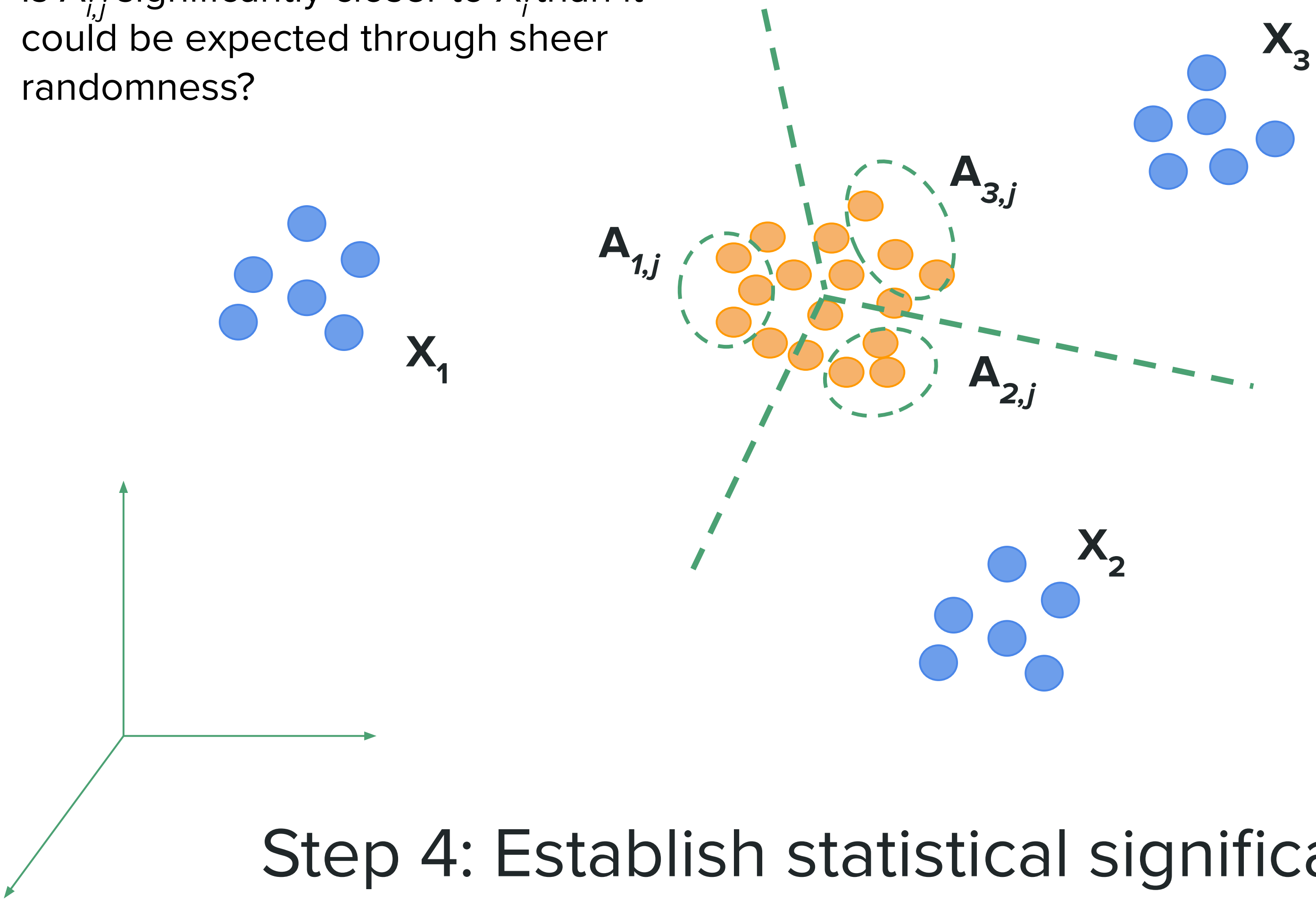
$$V_{ij} = \left\{ w \in \mathcal{A}_j \mid i = \arg\max_{i' \in [n]} \overline{\boldsymbol{w}} \cdot \overline{\boldsymbol{X}}_{i'} \right\}$$



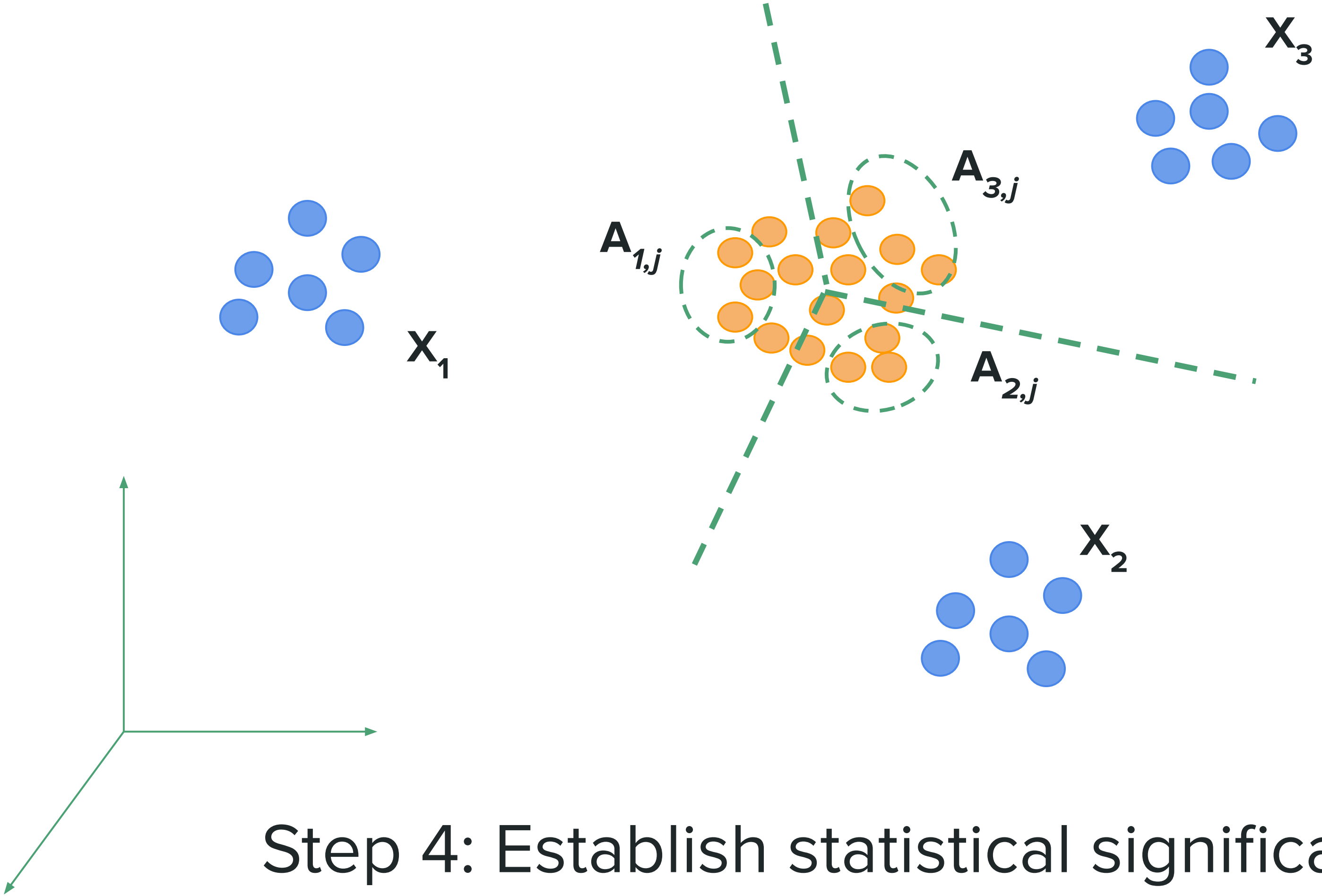
Step 3: Partition A_j

[AIES'19]

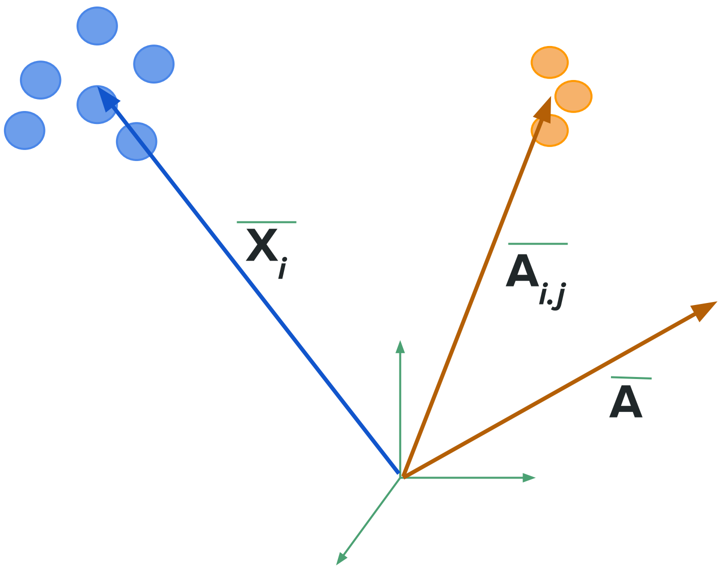
Is $A_{i,j}$ significantly closer to X_i than it could be expected through sheer randomness?



Step 4: Establish statistical significance

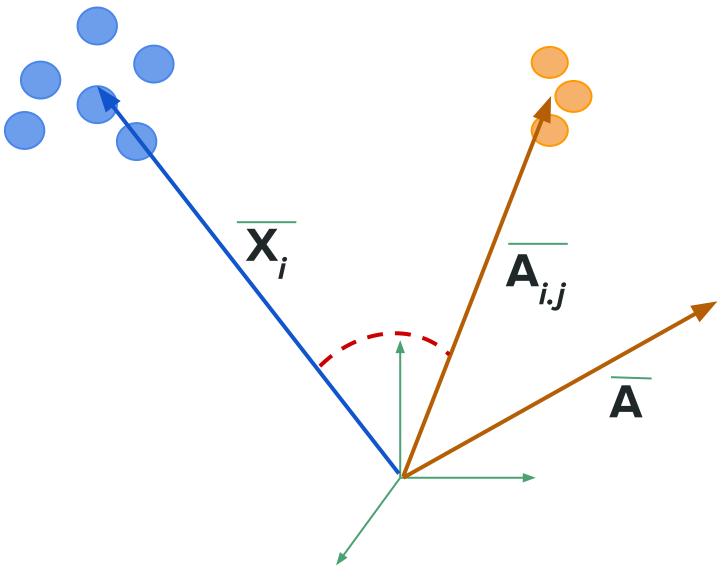


Step 4: Establish statistical significance



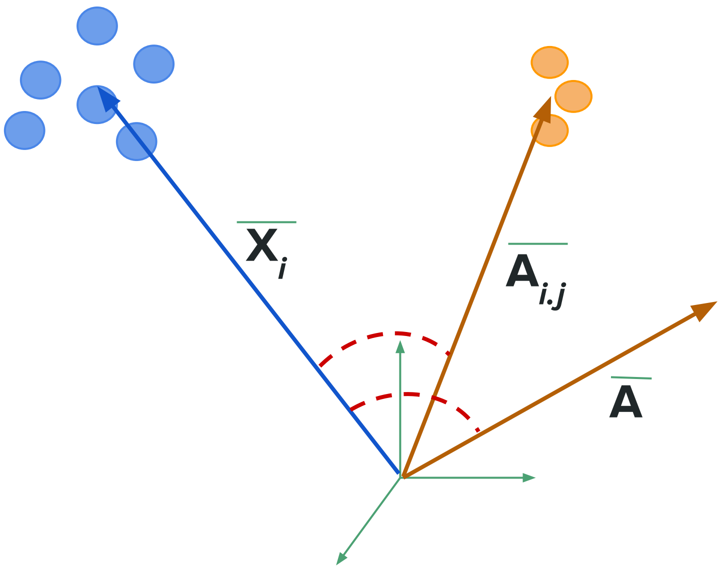
Step 4: Establish statistical significance

[AIES'19]



Step 4: Establish statistical significance

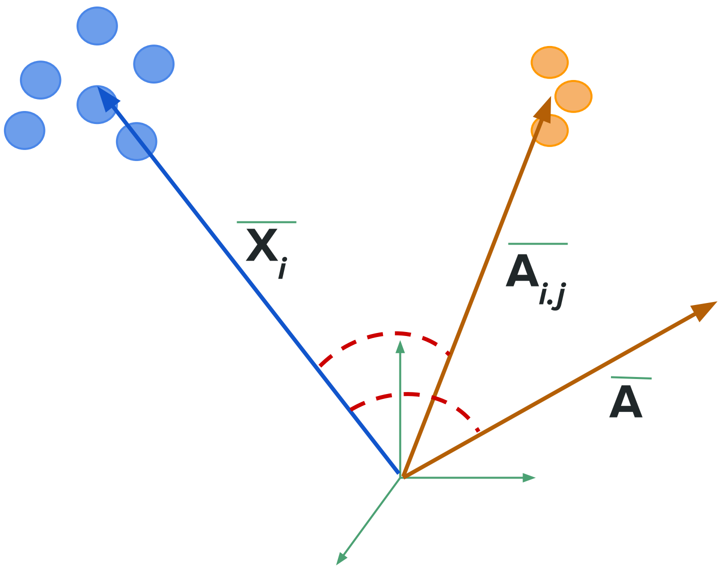
$$\sigma_{ij} = (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{A}}_{ij} - \overline{\boldsymbol{\mathcal{A}}})$$



Step 4: Establish statistical significance

$$\sigma_{ij} = (\overline{\boldsymbol{X}}_i - \mu) \cdot (\overline{\boldsymbol{A}}_{ij} - \overline{\boldsymbol{\mathcal{A}}})$$

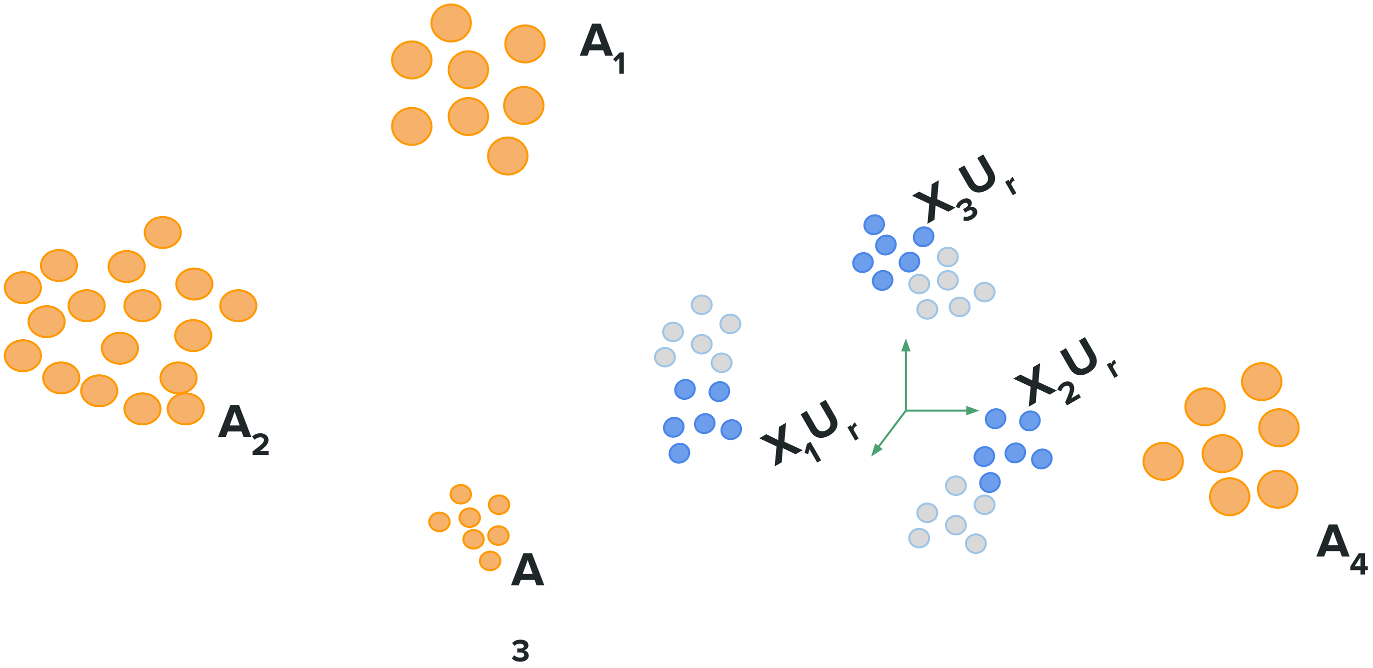
Is $\boldsymbol{\sigma}_{i,j}$ significantly large?



Step 4: Establish statistical significance

Rotational null hypothesis

1. Rotate X: $X \rightarrow XU_r$

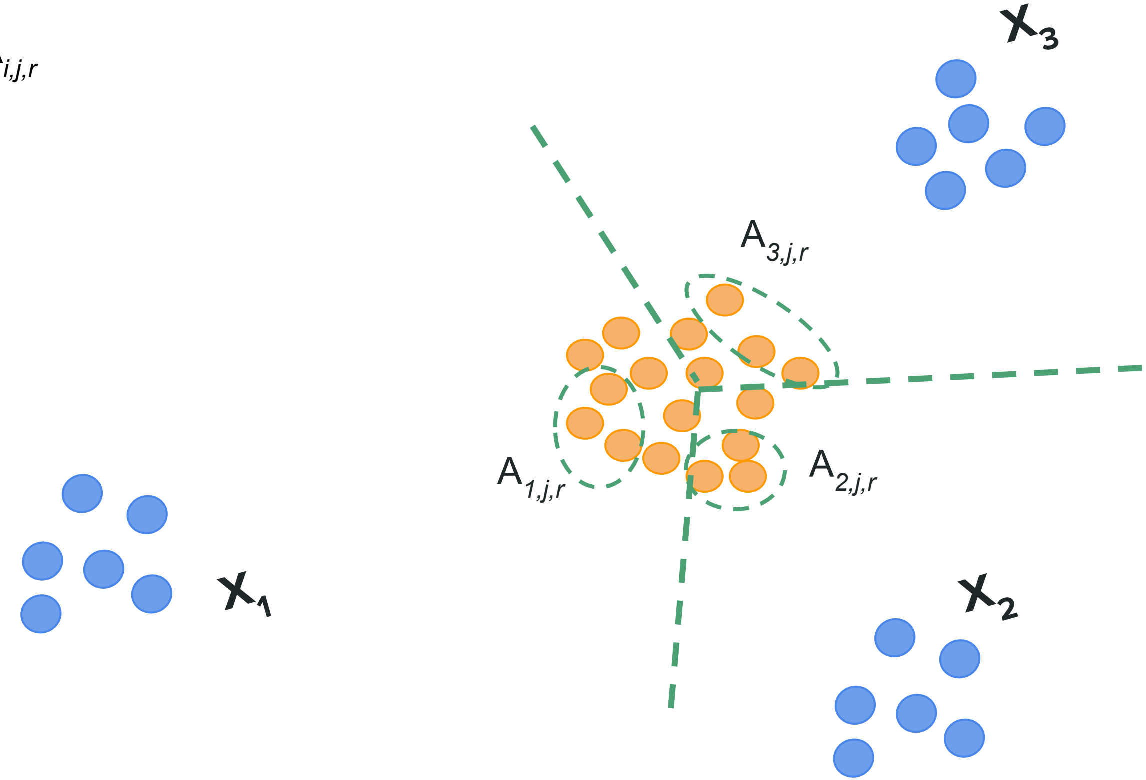


Step 4: Establish statistical significance

Rotational null hypothesis

[AIES'19]

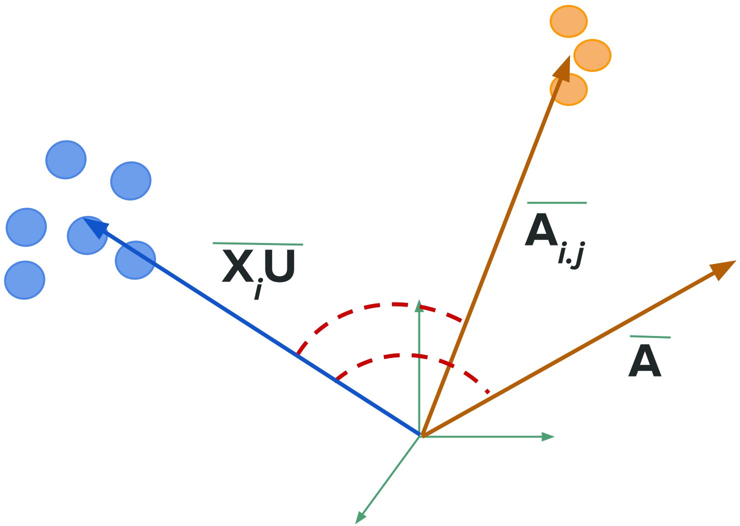
2. Find $A_{i,j,r}$



Step 4: Establish statistical significance

Rotational null hypothesis

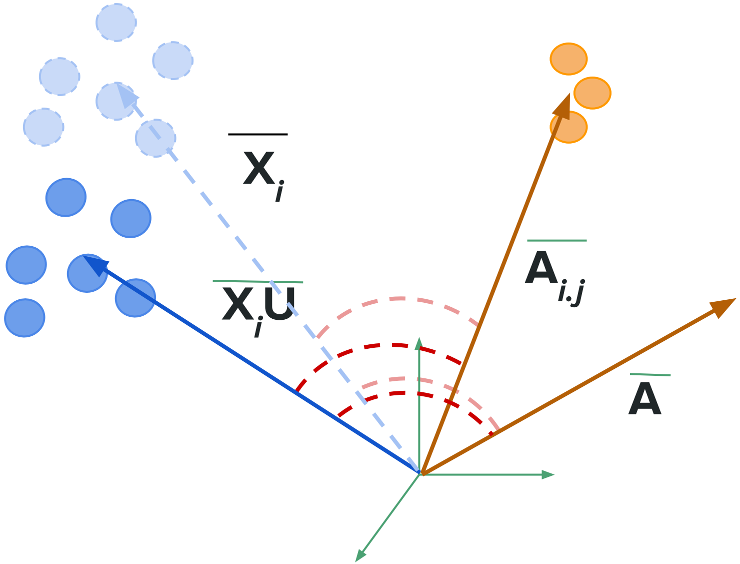
3. Calculate $\boldsymbol{\sigma}_{i,j,r}$



Step 4: Establish statistical significance

Rotational null hypothesis

3. Calculate $\boldsymbol{\sigma}_{i,j,r}$



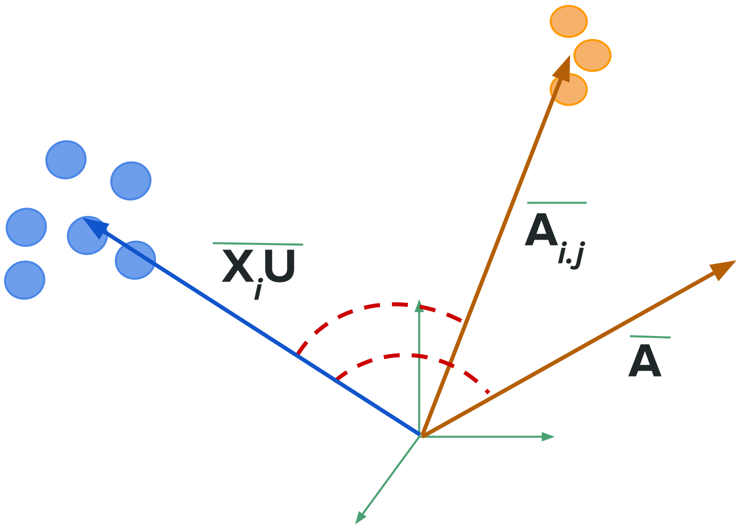
Step 4: Establish statistical significance

Rotational null hypothesis

3. Calculate p-value:

$p_{i,j} = [\ \delta(\sigma_{i,j} > \sigma_{i,j,r}) + 1]\ /\ [\ R + 1]$

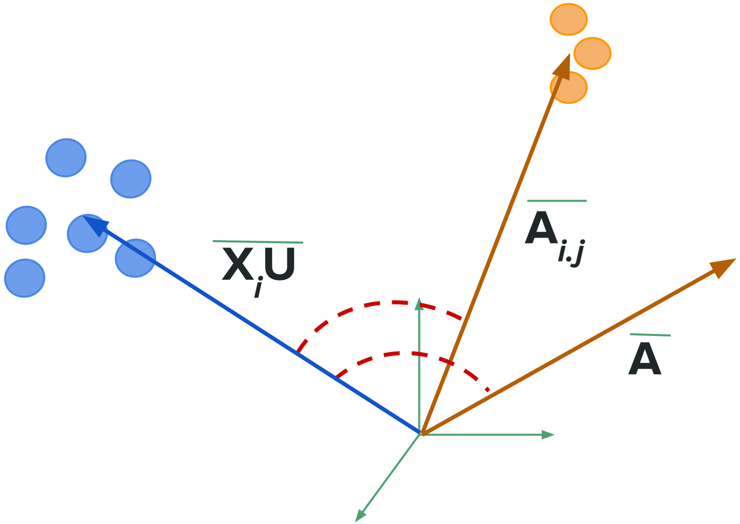
$r = 1,2,...,10k$



Step 4: Establish statistical significance

Rotational null hypothesis

4. Determine critical p-value, ***α***-bound guarantee on false discovery rate (*Benjamini-Hochbergh*)



Step 4: Establish statistical significance

Disclaimer

The biases in the following slides contain offensive stereotypes.

These do not reflect our views.

98% F 1983 4% B 4% H 89% W 3% A	98% F 1968 8% B 4% H 84% W 3% A	89% F 1978 48% B 3% H 47% W 1% A	85% F 1982 10% B 9% H 69% W 11% A	78% F 1993 2% B 1% H 95% W 1% A	65% F 1991 7% B 4% H 56% W 32% A	59% F 1985 4% B 9% H 83% W 4% A	56% F 1986 2% B 70% H 21% W 8% A	40% 1989 5% B 10% 79% 5% A
	cookbook, baking, baked goods	sweet potatoes, macaroni, green beans			saffron, halal, sweets	mozzarella, foie gras, caviar	tortillas, salsa, tequila	koshe humn bagel
herself, hers, moms	husband, homebound, grandkids	aunt, niece, grandmother	hubby, socialite, cuddle	twin sister, girls, classmate	elder brother, dowry, refugee camp			berea immi emigr
hostess, cheerleader, dietitian	registered nurse, homemaker, chairwoman		supermodel, beauty queen, stripper	helper, getter, snowboarder	shopkeeper, villager, cricketer		translator, interpreter, smuggler	
	log cabin, library, fairgrounds	front porch, carport, duplex	racecourse, plush, tenements	picnic tables, bleachers, concession stand	locality, mosque, slum	prefecture, chalet, sauna		synag constr hillto
	parish,	pastor,	goddess,		fatwa,	monastery,	rosary,	rabbis

Crowdsourcing evaluation

Qualification:

36 names, 3 per group
+1 per name labeled in correct group

Crowdsourcing evaluation

Qualification:
36 names, 3 per group
+1 per name labeled in correct group

If accuracy > 50%

Is the UBE output consistent with society's stereotypes?

For each WEAT:

- Groups in output {X1, X2, ... , Xk} and {A1, A2, ..., Ak} shown
- For each name group Xi, which group Ai contains words most stereotypically associated with these names?

[AIES'19]

Crowdsourcing evaluation

Qualification:

36 names, 3 per group

+1 per name labeled in correct group

If accuracy > 50%

Is the UBE output consistent with society's stereotypes?

For each WEAT:

- Groups in output {X1, X2, … , Xk} and {A1, A2, …, Ak} shown
- For each name group Xi, which group Ai contains words most stereotypically associated with these names?

If most commonly chosen group matches UBE pairing

Is it offensive? Rate:

1	2	3	4	5	6	7

Politically correct, inoffensive, or just random

Politically incorrect, possibly very offensive

Crowdsourcing evaluation

Emb.	# significant	% accurate	% offensive
w2v	235	72%	35%
fast	160	80%	38%
glove	442	48%	24%

Disclaimer

The biases in the following slides contain offensive stereotypes.

These do not reflect our views or the views of crowd workers.

Crowdsourcing evaluation

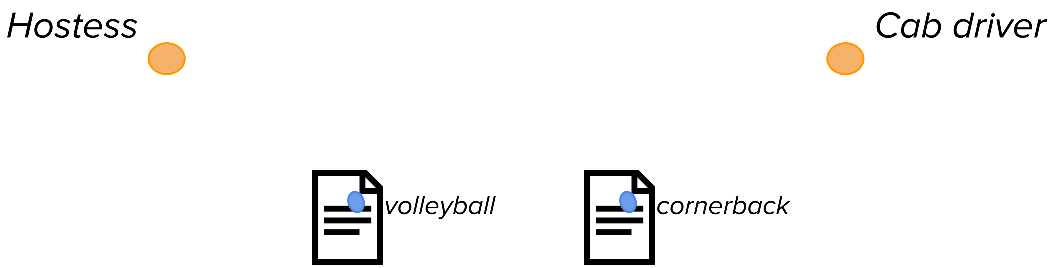
Emb.	# significant	% accurate	% offensive
w2v	235	72%	35%
fast	160	80%	38%
glove	442	48%	24%

Word2Vec trained on Google news			fastText trained on the Web			GloVe trained on the Web		
Miguel	**Dashaun**	**Kamal**	**Marquell**	**Ahmed**	**Alejandra**	**Amina**	**Alejandra**	**Kylee**
Deisy	**Jamell**	**Nailah**	**Antwan**	**Shanti**	**Maricella**	**Yair**	**Epifanio**	**Shaye**
Violeta	**Marlon**	**Kya**	**Dakari**	**Mariyah**	**Ona**	**Rani**	**Monalisa**	**Tayla**
Emilio	**Davonta**	**Maryam**	**Pernell**	**Siddharth**	**Fabiola**	**Danial**	**Eulalia**	**Latasha**
Yareli	**Demetrius**	**Rohan**	**Jarred**	**Yasmin**	**Sulema**	**Safa**	**Alicea**	**Jessi**
illegal immigrant	aggravated robbery	subcontinent	n*****	jihad	s******	turban	cartel	pornstar
drug trafficking	aggravated assault	tribesmen	f*****	militants	maid	saree	undocumented	hottie
deported	felonious assault	miscreants	dreads	caliphate	busty	hijab	culpable	nubile

*These associations do not reflect our views or those of the crowd workers.

Why does this matter?

- Representational harms
- Harmful bias encoded in semantic representation used for learning
- **Removing names is not enough to get rid of bias!**
    - Words in category clusters may be used as proxy for gender/race/etc



Outline

1. Risks of compounding injustices

Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting (FAT* 2019)
Maria De-Arteaga (CMU), Alexey Romanov (UMASS), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Adam Kalai (MSR)

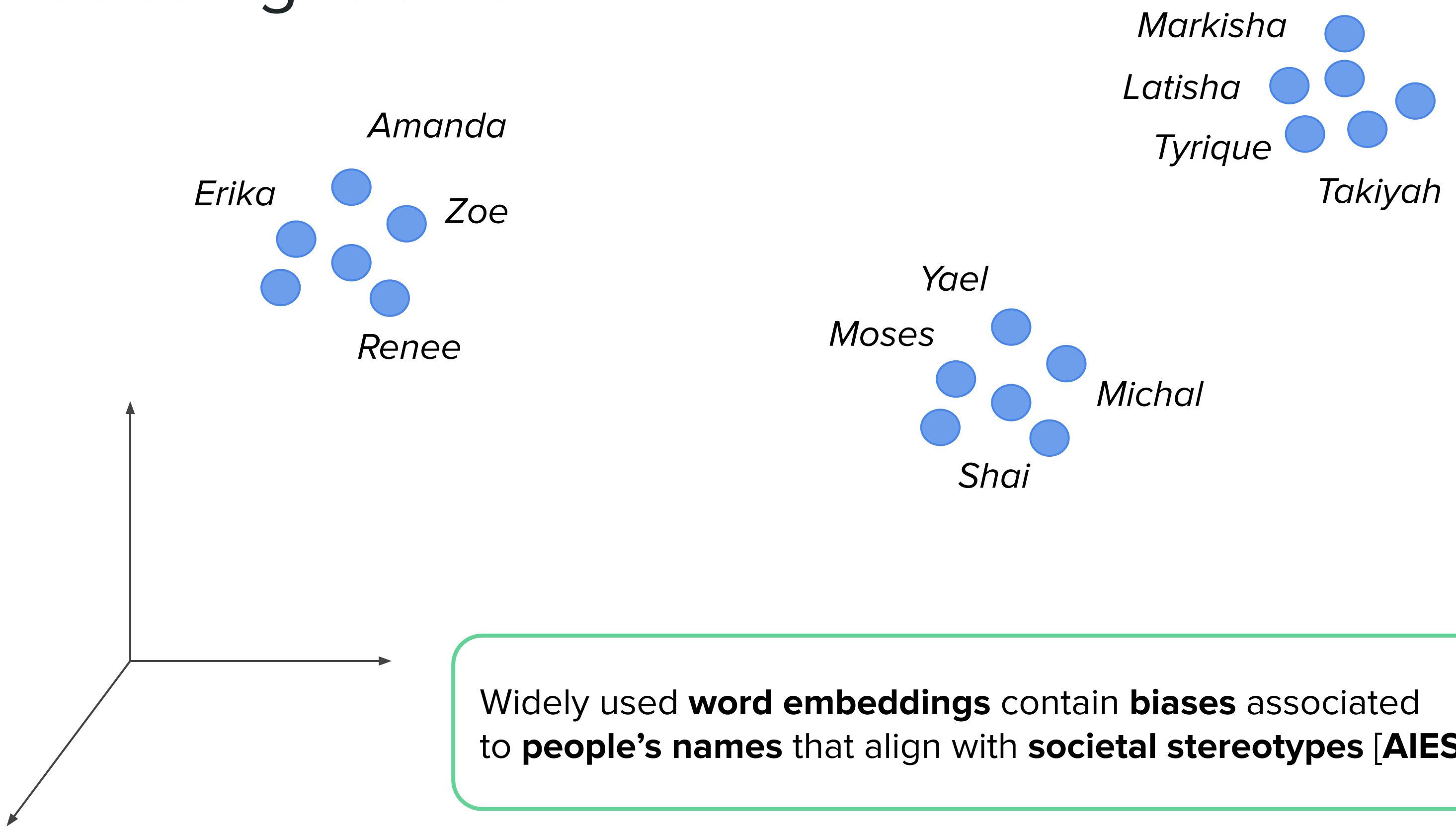
2. What are the biases in my word embedding?

What are the biases in my word embedding? (AIES 2019)
Nathaniel Swinger[=] (Lexington HS), Maria De-Arteaga[=] (CMU), Neil Thomas Heffernan IV (Shrewsbury HS), Mark Leiserson (UMD), Adam Kalai (MSR)

3. Using bias to fight bias

What's in a Name? Reducing Bias in Bios without Access to Protected Attributes (NAACL 2019)
Alexey Romanov (UMASS), Maria De-Arteaga (CMU), Hanna Wallach (MSR), Jennifer Chayes (MSR), Christian Borgs (MSR), Alexandra Chouldechova (CMU), Sahin Geyik (LinkedIn), Krishnaram Kenthapadi (LinkedIn), Anna Rumshisky (UMASS), Adam Kalai (MSR) 🏆

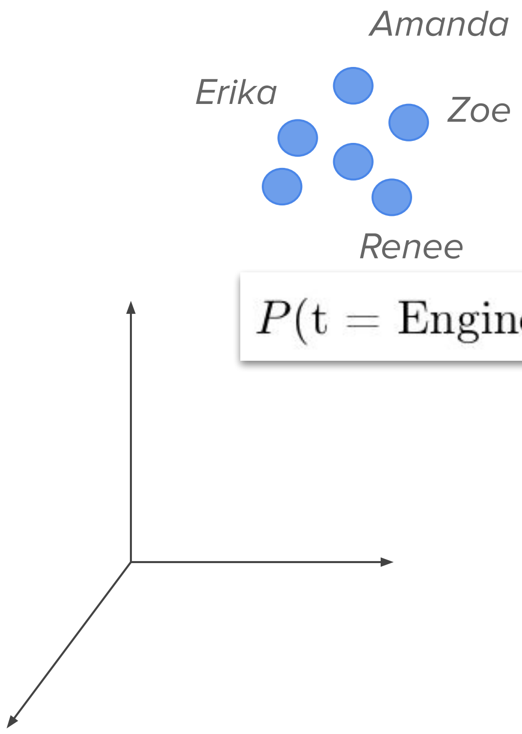
Bias to fight bias?



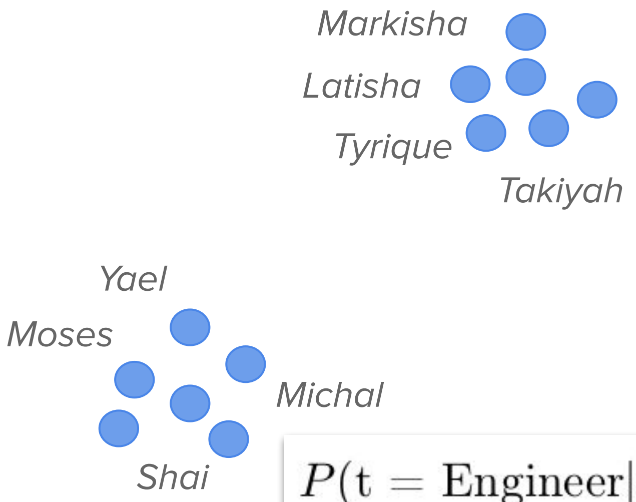
Widely used **word embeddings** contain **biases** associated to **people's names** that align with **societal stereotypes** [**AIES'19**]

[NAACL'19]

Bias to fight bias?



$$P(\mathrm{t} = \mathrm{Engineer} | x_i, \theta)?$$



$$P(\mathrm{t} = \mathrm{Engineer} | x_i, \theta)?$$

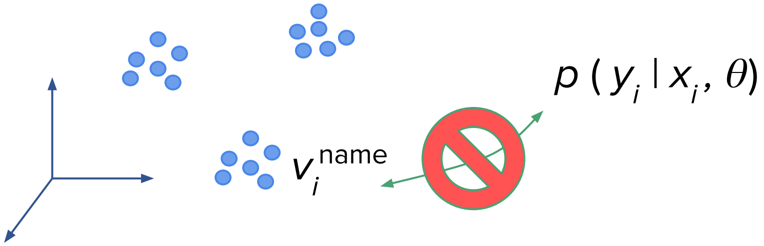
"**What's in a name?** That which we call a rose

By any other name would smell as sweet."

William Shakespeare, Romeo and Juliet

Reducing Bias in Bios

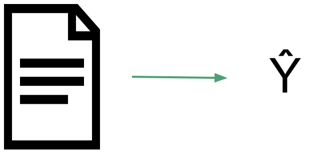
Penalize correlation between **errors** and **location of name** in the word embedding



Training

Predict class *y* from features x

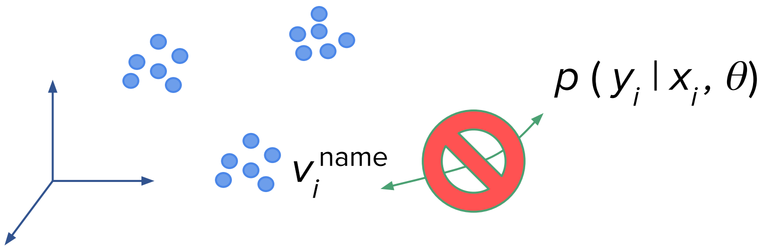
(does not use name for prediction)



Deployment / testing

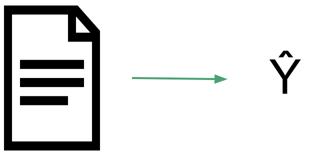
Reducing Bias in Bios

Penalize correlation between **errors** and **location of name** in the word embedding



Predict class *y* from features x

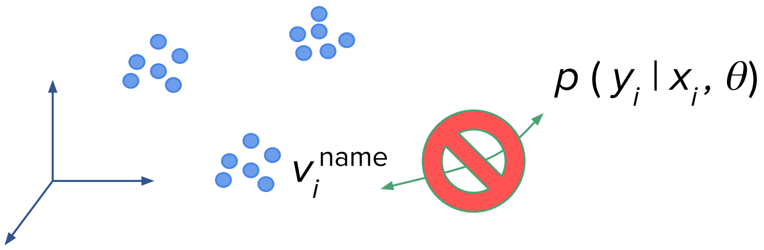
(does not use name for prediction)



- Typically, minimize loss function: $\mathcal{L}$
- Proposed: regularize accuracy gaps: $\mathcal{L}_{\text{total}} = \mathcal{L} + \lambda \cdot \mathcal{L}_{\text{CL}}$

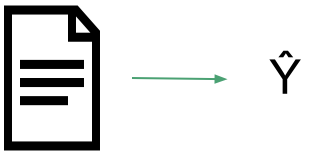
Reducing Bias in Bios

Penalize correlation between **errors** and **location of name** in the word embedding



Predict class *y* from features x

(does not use name for prediction)



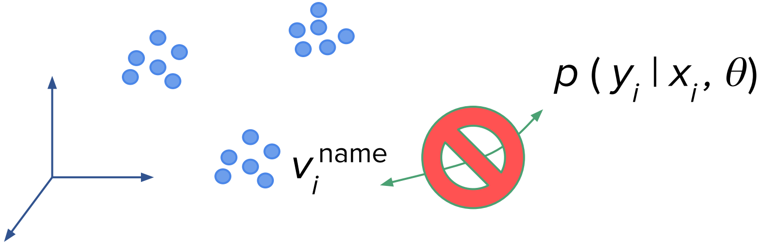
$\mathcal{L}_{\text{CluCL}}$ (θ): Cluster constrained loss

$\mathcal{L}_{k,t}(\theta)$: cluster *k* loss for title *t*

$$\mathcal{L}_{CluCL}(\theta) = -\frac{\sum_{j,k,t}(\mathcal{L}_{j,t}(\theta) - \mathcal{L}_{k,t}(\theta))^2}{NK(K-1)}$$

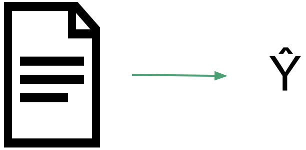
Reducing Bias in Bios

Penalize correlation between **errors** and **location of name** in the word embedding



Predict class *y* from features x

(does not use name for prediction)



$\mathcal{L}_{\mathbf{CoCL}}$ **($\boldsymbol{\theta}$):** Covariance constrained loss

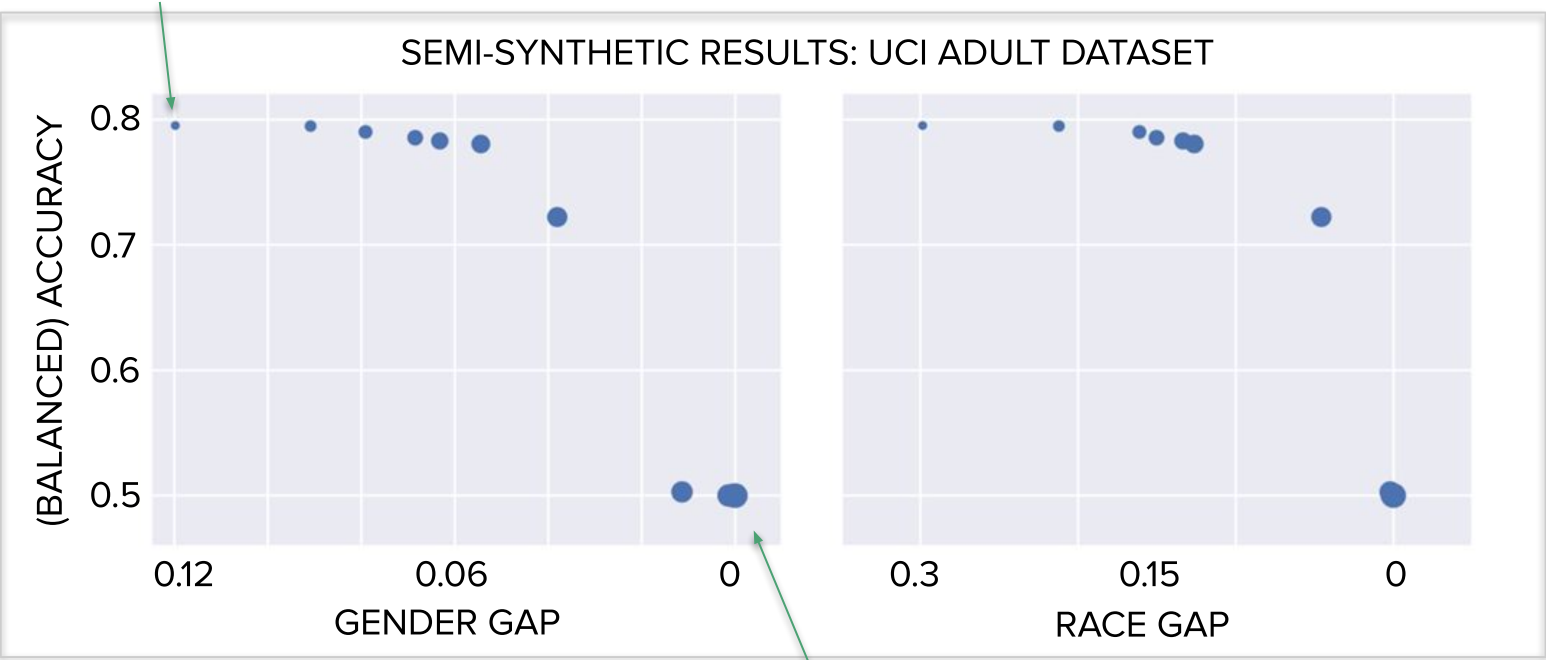
$$\mathcal{L}_{CoCL}(\theta) = \frac{1}{N}\sum_t ||\, E_{i:y_i=t}[(v_i^{\text{name}} - \overline{v}_t) \cdot (p(t|x_i,\theta) - \overline{p}_t)]\, ||_2$$

Accuracy / fairness tradeoff

$$\mathrm{Gap}_{r,c} = \mathrm{TPR}_{r,c} - \mathrm{TPR}_{\sim r,c}$$

$$\mathrm{Gap}_r^{\mathrm{RMS}} = \sqrt{\frac{1}{|C|}\sum_{c \in C} \mathrm{Gap}_{r,c}^2}$$

Unconstrained



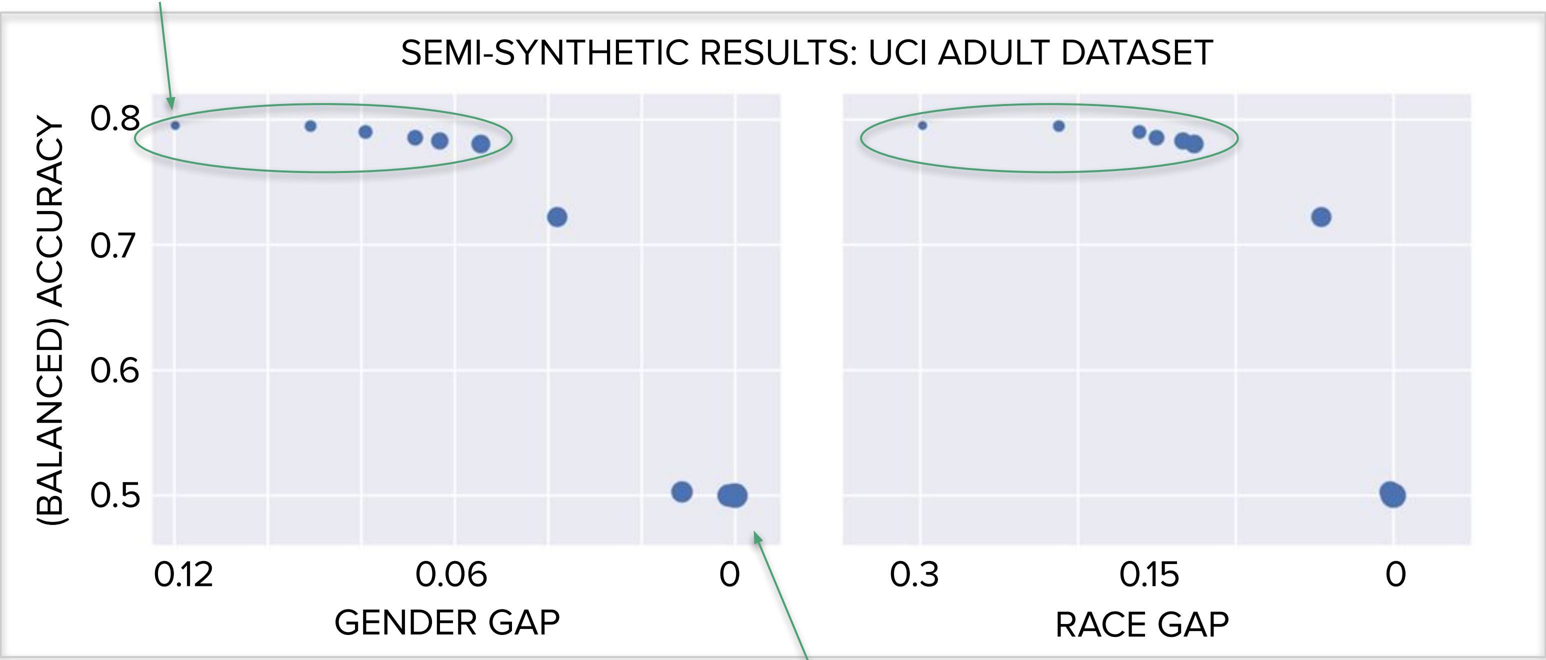
Tightly constrained

Accuracy / fairness tradeoff

$$\mathrm{Gap}_{r,c} = \mathrm{TPR}_{r,c} - \mathrm{TPR}_{\sim r,c}$$

$$\mathrm{Gap}_r^{\mathrm{RMS}} = \sqrt{\frac{1}{|C|}\sum_{c\in C}\mathrm{Gap}_{r,c}^2}$$

Unconstrained



Tightly constrained

Results: UCI Adult dataset

Root Mean Square Gender Accuracy Gap

Root Mean Square Race Accuracy Gap

Model	R	Accuracy Balanced	RMS GAG	RMS RAG	Max GAG	Max RAG
Regular	0	**0.795**	0.299	0.120	0.303	0.148
CluCL	1	0.788	0.278	0.121	0.297	0.145
CluCL	2	0.793	0.259	0.085	0.282	0.114
CoCL	1	0.794	0.215	0.091	0.251	0.119
CoCL	2	0.790	**0.163**	**0.080**	**0.201**	**0.109**

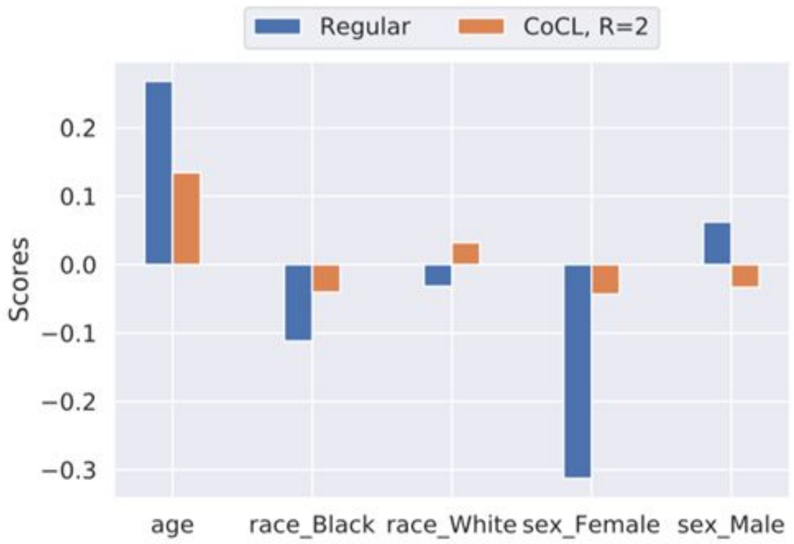
Results: Bias in bios dataset

Root Mean Square Gender Accuracy Gap

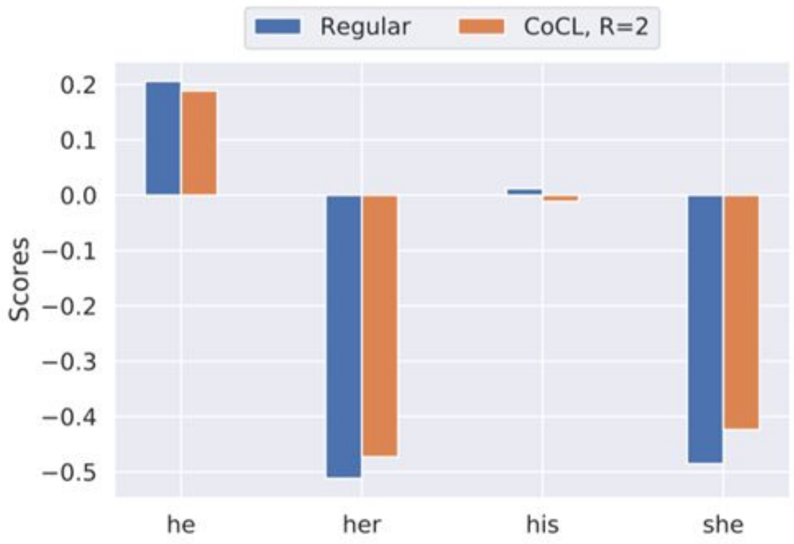
Root Mean Square Race Accuracy Gap

Model	R	Accuracy Balanced	RMS GAG	RMS RAG	Max GAG	Max RAG
Regular	0	**0.788**	0.173	0.051	0.511	0.121
CluCL	1	0.784	0.168	0.048	0.494	0.120
CluCL	2	0.781	**0.165**	**0.047**	**0.486**	0.114
CoCL	1	0.785	0.168	0.048	0.507	**0.109**
CoCL	2	0.779	0.169	0.048	0.512	0.116

What changed?



(a) The *Adult* dataset



(b) The *Bios* dataset, occupation “surgeon”

Summary

- Characterized risk of **compounding injustices** in supervised learning [**FAT*'19**]
- **Large-scale study** of automated recruiting: not enough to remove gender indicators [**FAT*'19**]
- Proposed algorithm to enumerate biases in word embeddings [**AIES'19**]
- Proposed methodology to **mitigate biases** without assuming access to protected attributes [**NAACL'19** 🏆]
- A long way to go to solve the problem!

Summary

- Characterized risk of **compounding injustices** in supervised learning [**FAT*'19**]
- **Large-scale study** of automated recruiting: not enough to remove gender indicators [**FAT*'19**]
- Proposed algorithm to enumerate biases in word embeddings [**AIES'19**]
- Proposed methodology to **mitigate biases** without assuming access to protected attributes [**NAACL'19** 🏆]
- A long way to go to solve the problem!

Thanks!

mdeartea@andrew.cmu.edu