344.075 KV: Natural Language Processing

Footprint of Societal Biases in NLP

Navid Rekab-Saz

navid.rekabsaz@jku.at

Institute of Computational Perception

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

Agenda

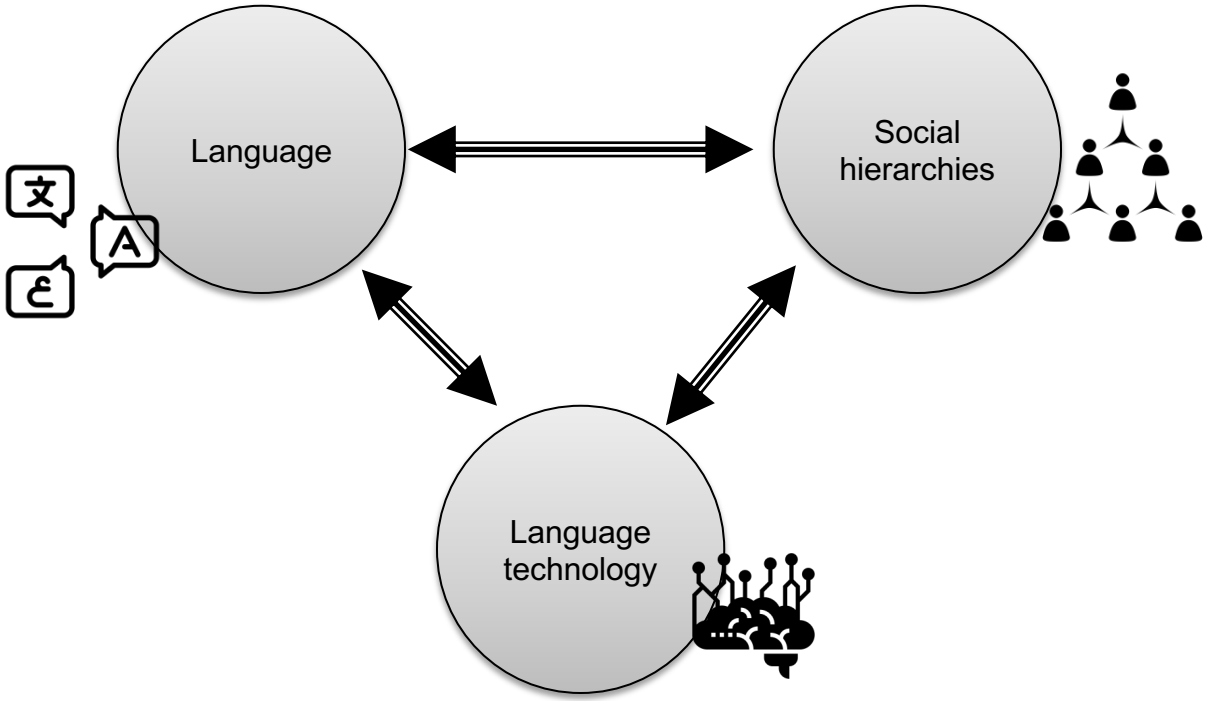
- Societal biases in NLP … what? why?
- Bias in word embedding
- Measuring bias in Information Retrieval

Agenda

- **Societal biases in NLP … what? why?**
- Bias in word embedding
- Measuring bias in Information Retrieval

Language and Society

- Language …
  - takes on and defines social meaning
  - forms and maintains social hierarchies by …
    - labeling social groups
    - transmitting the beliefs about social groups



Maass, Anne. "Linguistic intergroup bias: Stereotype perpetuation through language." *Advances in experimental social psychology*. 1999.
Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. "Language (Technology) is Power: A Critical Survey of" Bias" in NLP." *In Proc. Of ACL 2020*

(Societal) Bias

- Biases and stereotypes do not *per se* carry negative meanings!
- What we mean from "bias" in this lecture is …

"Inclination or prejudice for or against one person or group, especially in a way considered to be unfair."

Oxford dictionary

"demographic disparities in algorithmic systems that are objectionable for societal reasons."

Fairness and Machine Learning
Solon Barocas, Moritz Hardt, Arvind Narayanan, 2019, fairmlbook.org

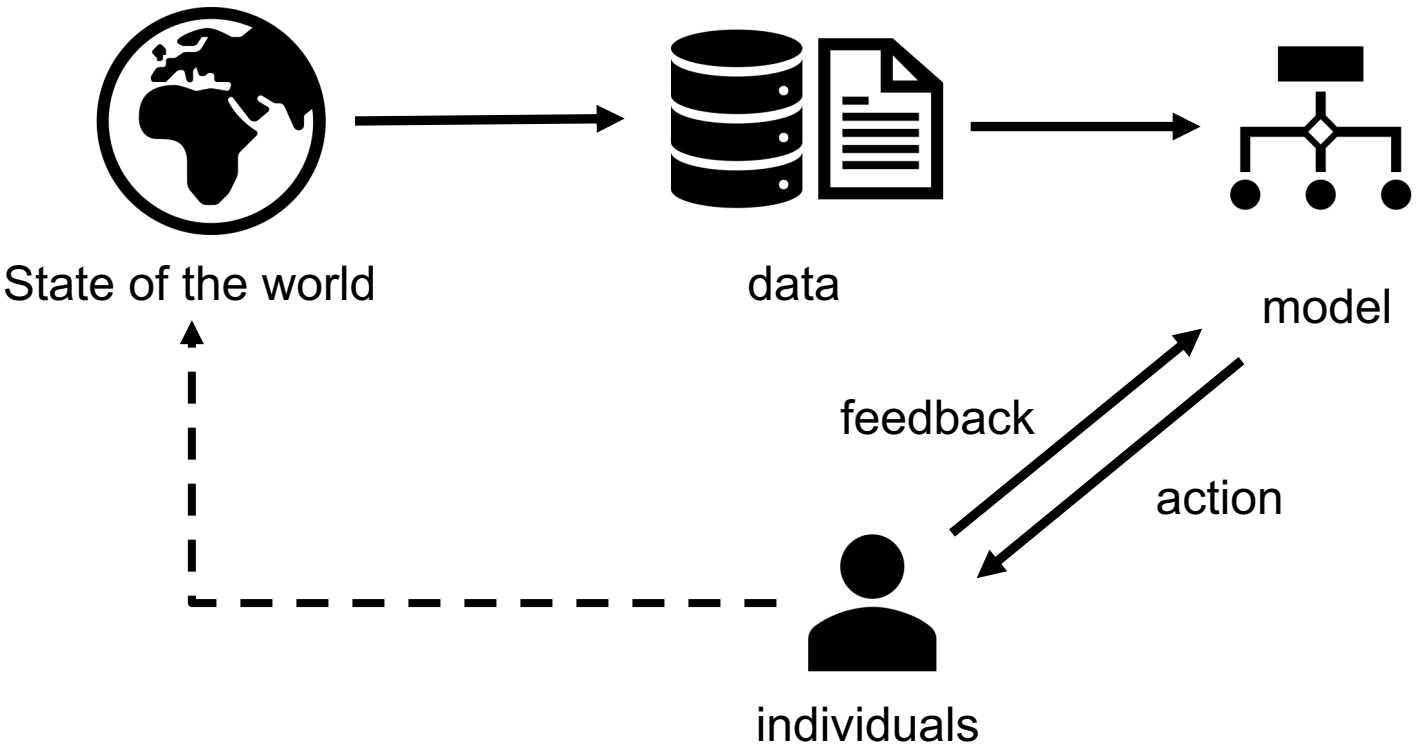
In which ways can bias be *harmful*?

- Allocational harms
  - A system allocates resources (e.g., credit) or opportunities (e.g., jobs) unfairly to different social groups

- Representational harms
  - a system (e.g., a search engine or a recommender system) represents some social groups in a less favorable light than others.
    - For instance by stereotyping that propagates negative generalizations about particular social groups

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. "Language (Technology) is Power: A Critical Survey of" Bias" in NLP." *In Proc. Of ACL 2020*
Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The Problem With Bias: Al- locative Versus Representational Harms in Machine Learning. In Proceedings of SIGCIS,

Machine Learning Cycle



NLP and Society

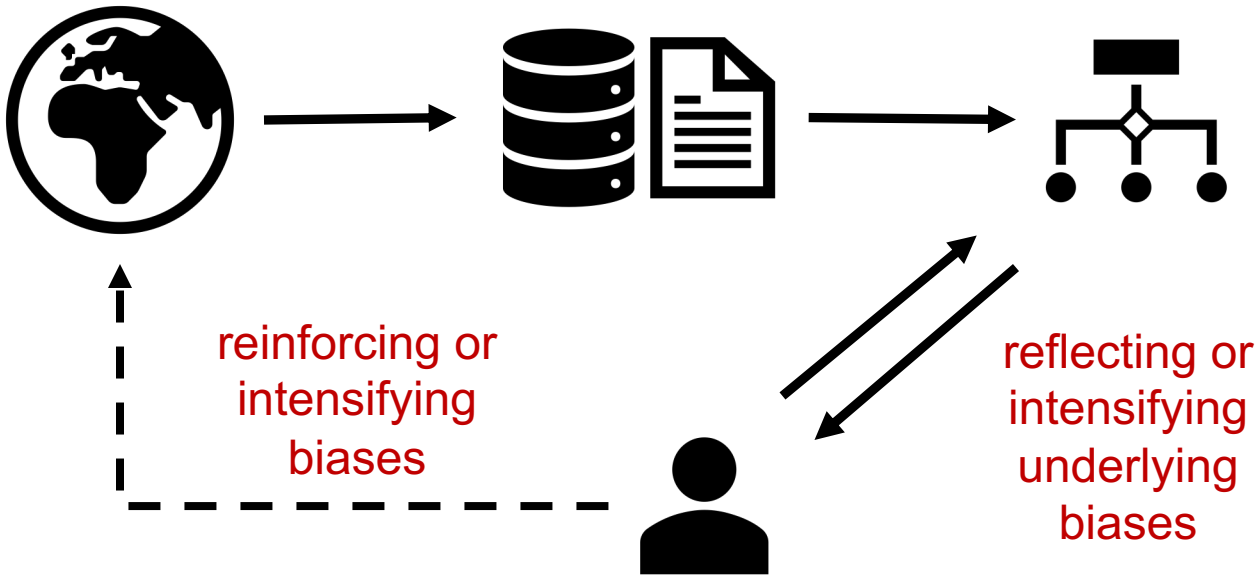
Case 1: observing

- Using NLP methods to observe/measure/monitor societal phenomena
  - Questions like *"how the perception of girls and boys towards the color pink has changed over time?"*
  - Exploits text data as proxies of the society
  - Related to the field of Computational Social Science

NLP and Society

Case 2: being affected

- Encoded societal biases and stereotypes in NLP system
  - Biases influence decision making process of NLP system
  - NLP systems may reinforce or even intensify biases



Bias in image processing

Google says sorry for racist auto-tag in photo app

https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app

FaceApp's creator apologizes for the app's skin-lightening 'hot' filter

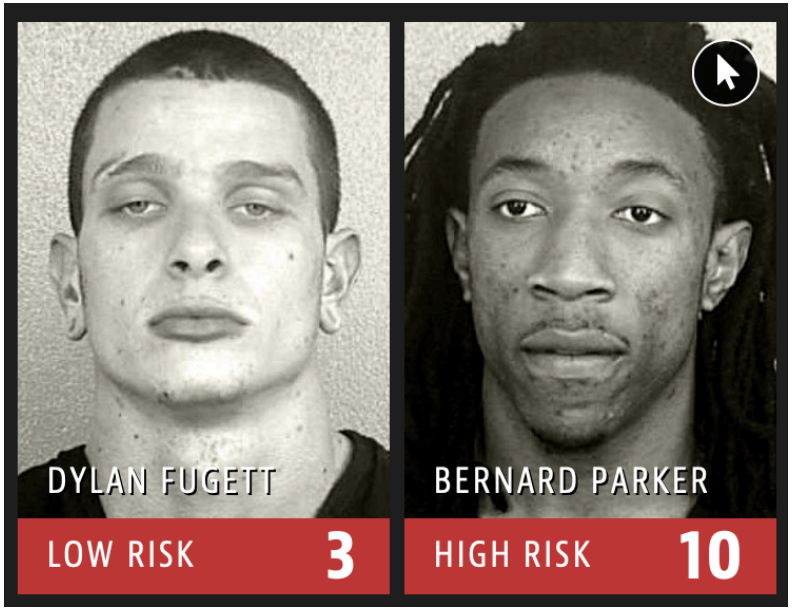
https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology

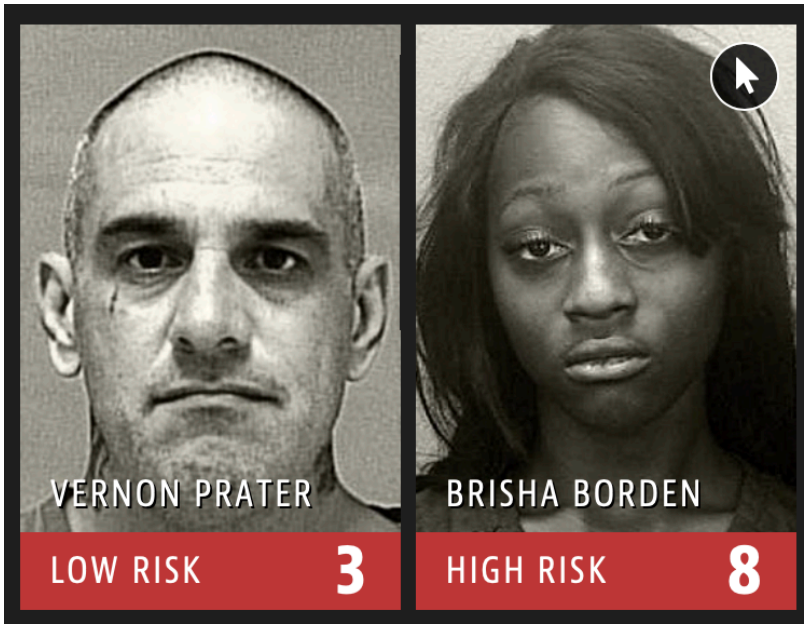
Beauty.AI's 'robot beauty contest' is back – and this time it promises not to be racist

https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai

Bias in crime discovery

- Predicted risk of reoffending

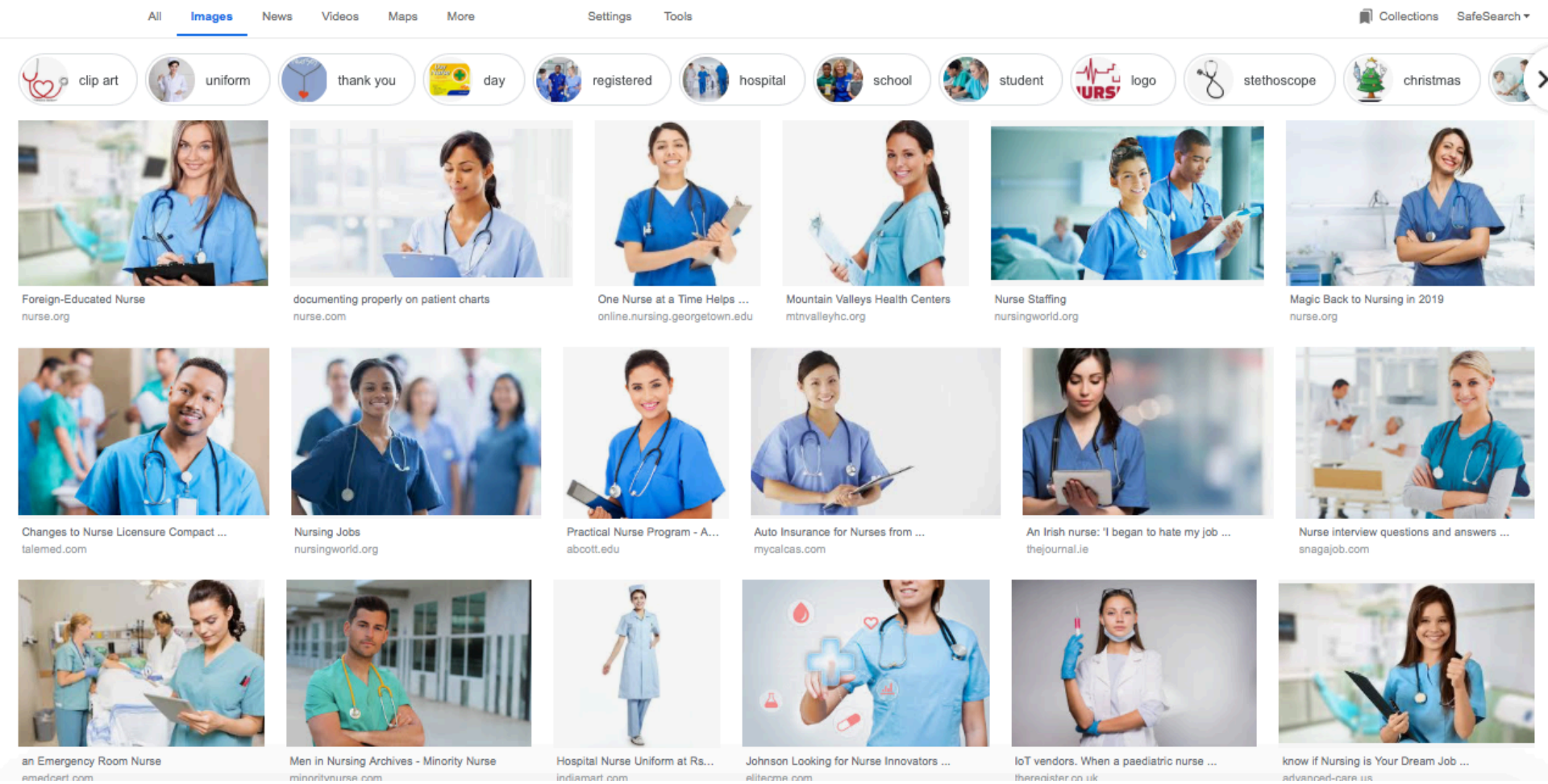




https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bias in Image Retrieval

Search: nurse

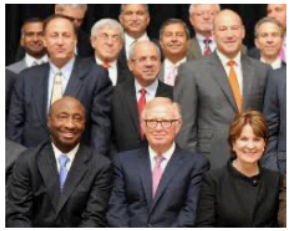


Bias in Image Retrieval

Search: CEO

All | Images | News | Maps | Videos | More | Settings | Tools | Collections | SafeSearch

business | google | microsoft | apple | woman | desk | amazon | uber | pepsi | black | starbucks


Chief executive officer - Wikipedia
en.wikipedia.org


Odilon Almeida as President and Chief ...
businesswire.com


You are the CEO of Your Life | Personal ...
personalexcellence.co


What do CEOs do? A CEO Job Descriptio...
steverrobbins.com


CEO vs. Owner: The Key Differences ...
onlinemasters.ohio.edu


Why You Need To Be The CEO Of Your C...
forbes.com


How to use 'CEO magic' when trying ...
europeanceo.com


LinkedIn CEO Jeff Weiner steps down ...
fortune.com


ABB ernennt neuen CEO | Netzwoche
netzwoche.ch

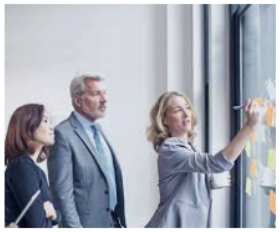
This South African has been named amon...
businesstech.co.za


CEO of the Year
chiefexecutive.net


Harvard study: What CEOs do all day
cnbc.com


C.E.O. Fired Over a Relationship ...
nytimes.com


ESSENCORE
essencore.com

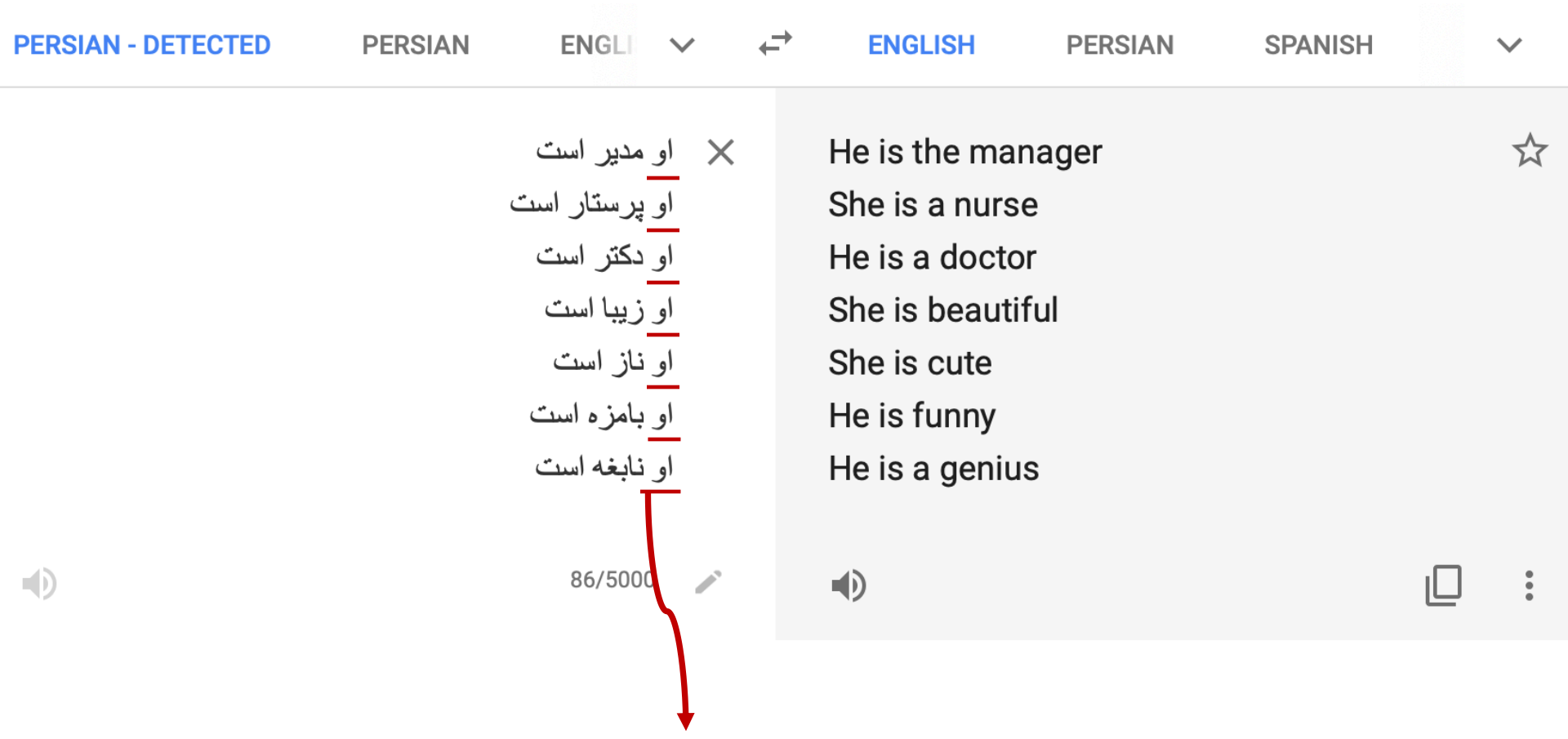

How CEOs can help lead technology ...
mckinsey.com


Becoming a CEO: Unexpected traits top ...
execed.economist.com


Casey's Announces CEO Tr...
businesswire.com

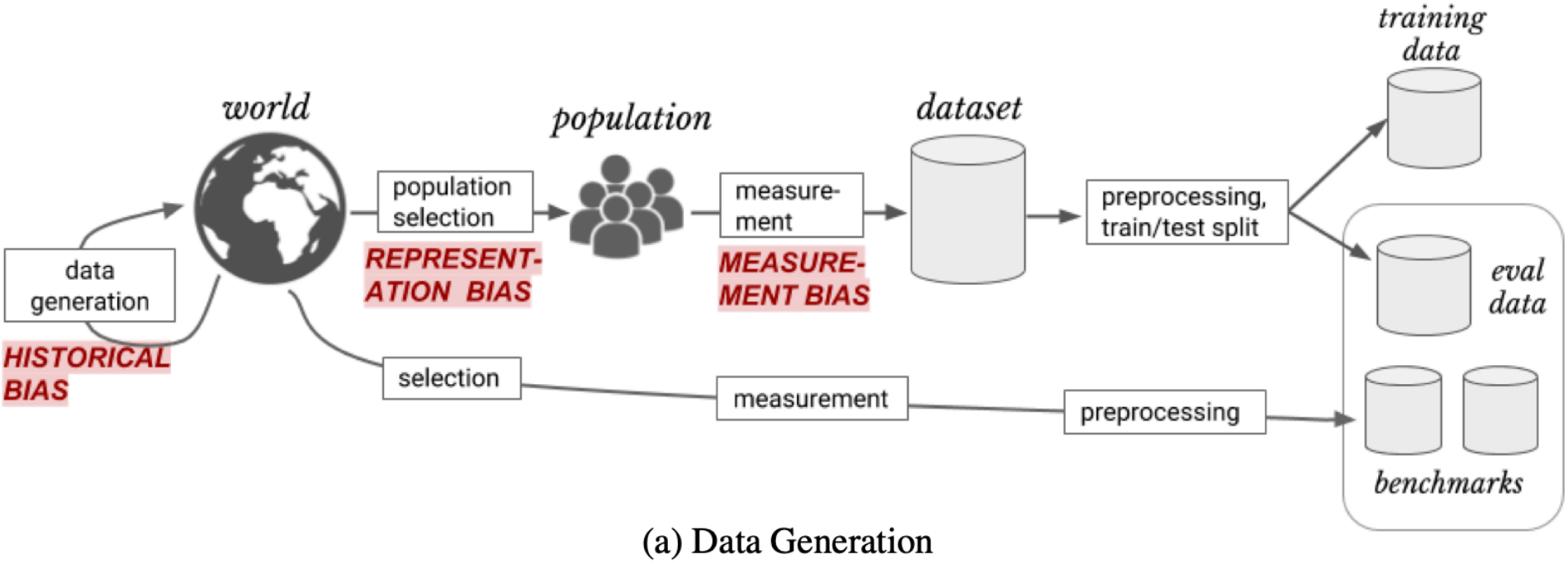

CEO MESSAGE | JCB Global Website
global.jcb

Bias in Machine Translation



same gender-neutral pronoun

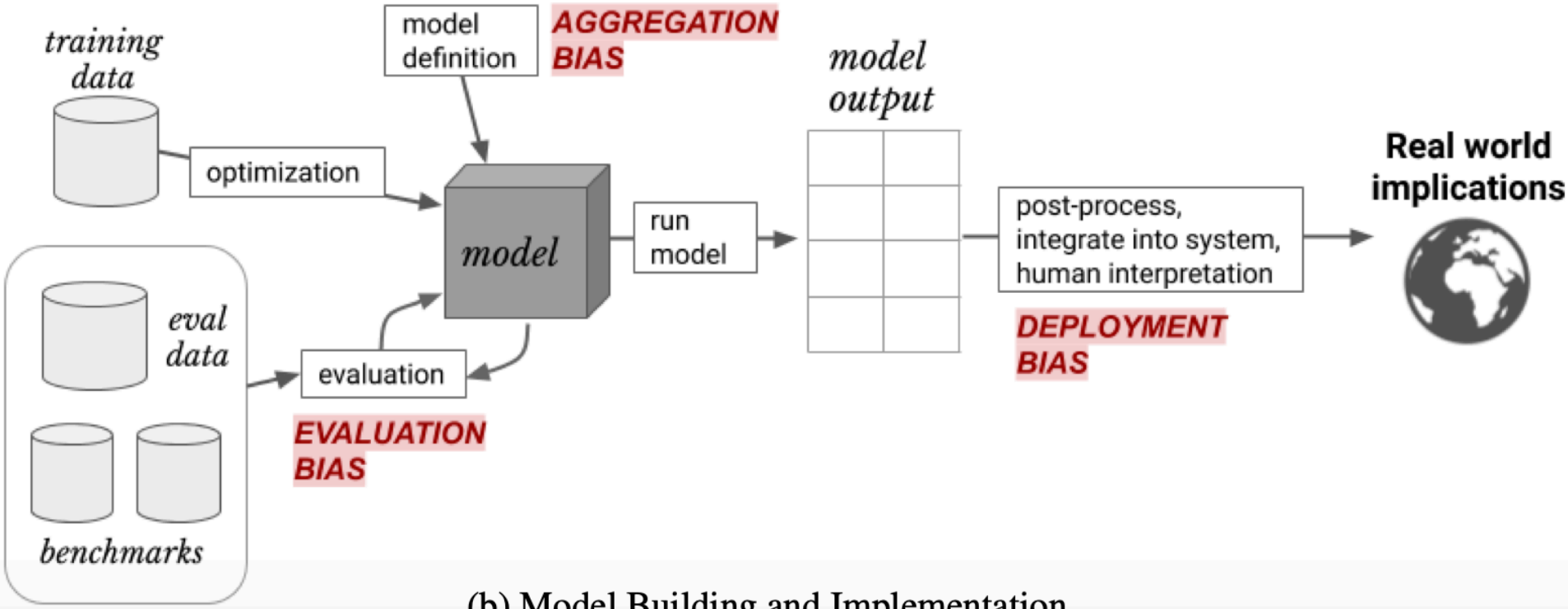
Bias in Machine Learning



(a) Data Generation

Suresh, Harini, and John V. Guttag. "A framework for understanding unintended consequences of machine learning." *arXiv preprint arXiv:1901.10002* (2019).

Bias in Machine Learning



(b) Model Building and Implementation

Suresh, Harini, and John V. Guttag. "A framework for understanding unintended consequences of machine learning." *arXiv preprint arXiv:1901.10002* (2019).

Where is bias originated from?

- World (historical bias)
  - Historical and ongoing discrimination
- Data (representation bias / measurement bias)
  - Sampling strategy - who is included in the data?
- Models (aggregation bias)
  - Using sensitive information (e.g. race) directly or adversely
  - Naive modeling learns more accurate predictions for majority group
  - Algorithm optimization eliminates “noise”, which might constitute the signal for some groups of users
- Evaluations (evaluation bias)
  - Definition of Success
    - Who is it good for, and how is that measured? Who decided this? To whom are they accountable?
  - Data annotation and benchmarking
- Human interaction (deployment bias)

Source: https://fair-ia.ekstrandom.net/sigir2019-slides.pdf

Bias & Fairness in standard Machine Learning

Attributes

- age
- workclass
- fnlwgt
- education
- marital-status
- occupation
- relationship
- race
- sex
- capital-gain
- capital-loss
- hours-per-week
- native-country

whether a person makes over 50K a year

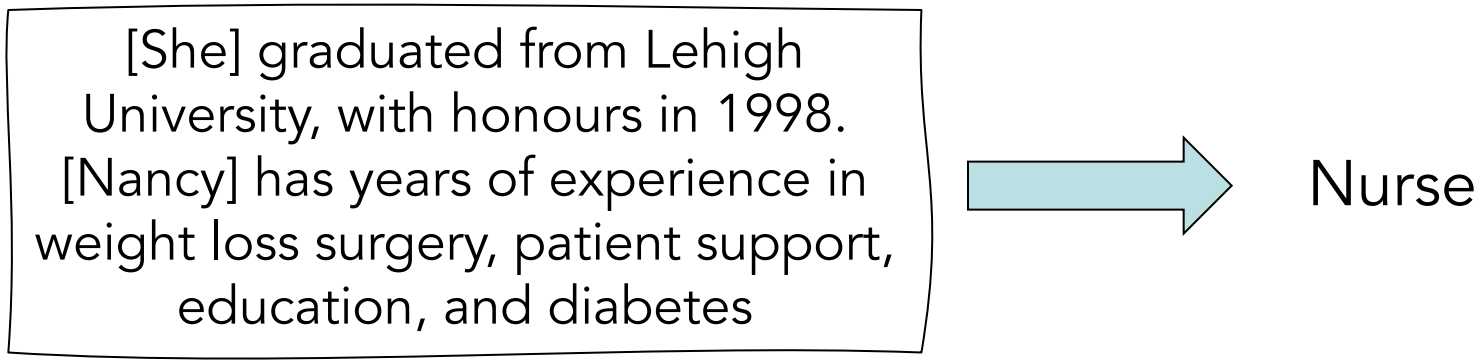
```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, Handlers-cleaners, Not-in-family, White, Male, 0, 0, 40, United-States, <=50K
53, Private, 234721, 11th, 7, Married-civ-spouse, Handlers-cleaners, Husband, Black, Male, 0, 0, 40, United-States, <=50K
28, Private, 338409, Bachelors, 13, Married-civ-spouse, Prof-specialty, Wife, Black, Female, 0, 0, 40, Cuba, <=50K
37, Private, 284582, Masters, 14, Married-civ-spouse, Exec-managerial, Wife, White, Female, 0, 0, 40, United-States, <=50K
49, Private, 160187, 9th, 5, Married-spouse-absent, Other-service, Not-in-family, Black, Female, 0, 0, 16, Jamaica, <=50K
52, Self-emp-not-inc, 209642, HS-grad, 9, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 45, United-States, >50K
31, Private, 45781, Masters, 14, Never-married, Prof-specialty, Not-in-family, White, Female, 14084, 0, 50, United-States, >50K
42, Private, 159449, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 5178, 0, 40, United-States, >50K
37, Private, 280464, Some-college, 10, Married-civ-spouse, Exec-managerial, Husband, Black, Male, 0, 0, 80, United-States, >50K
30, State-gov, 141297, Bachelors, 13, Married-civ-spouse, Prof-specialty, Husband, Asian-Pac-Islander, Male, 0, 0, 40, India, >50K
23, Private, 122272, Bachelors, 13, Never-married, Adm-clerical, Own-child, White, Female, 0, 0, 30, United-States, <=50K
```

http://www.fairness-measures.org/Pages/Datasets/censusincome.html

Bias & Fairness in NLP

- In language, bias can hide behind the implicit meanings of words and sentences

A sample task – occupation prediction from biographies:



De-Arteaga, Maria, et al. "Bias in bios: A case study of semantic representation bias in a high-stakes setting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019.

Real problems need interdisciplinary thinking!

- Fairness and bias are social concepts and inherently normative

- Bias in NLP systems should be grounded in its social context

> "… without this grounding, researchers and practitioners risk measuring and mitigating only what is convenient to measure and mitigate, rather than what is most normatively concerning."

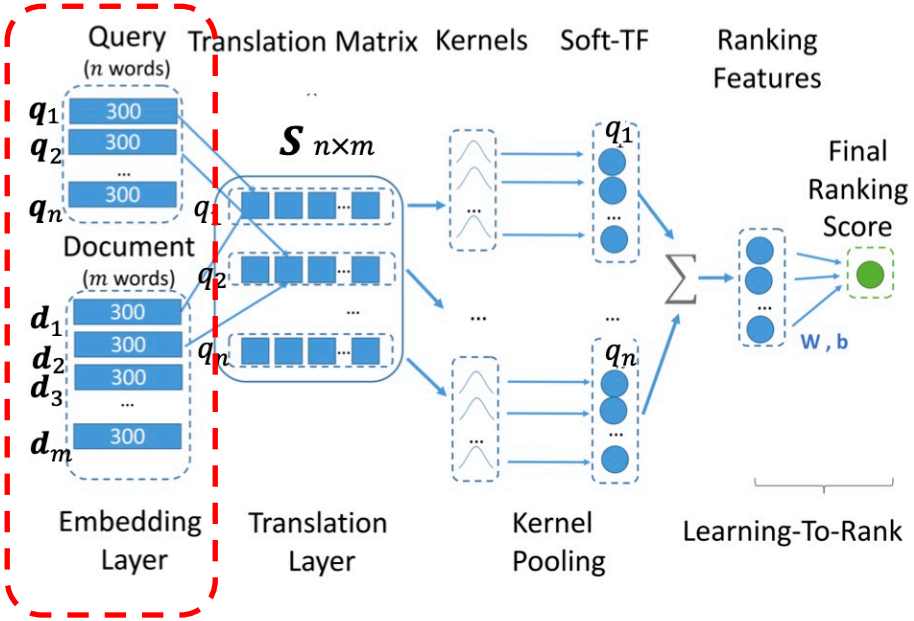
Blodgett et al. [2020]

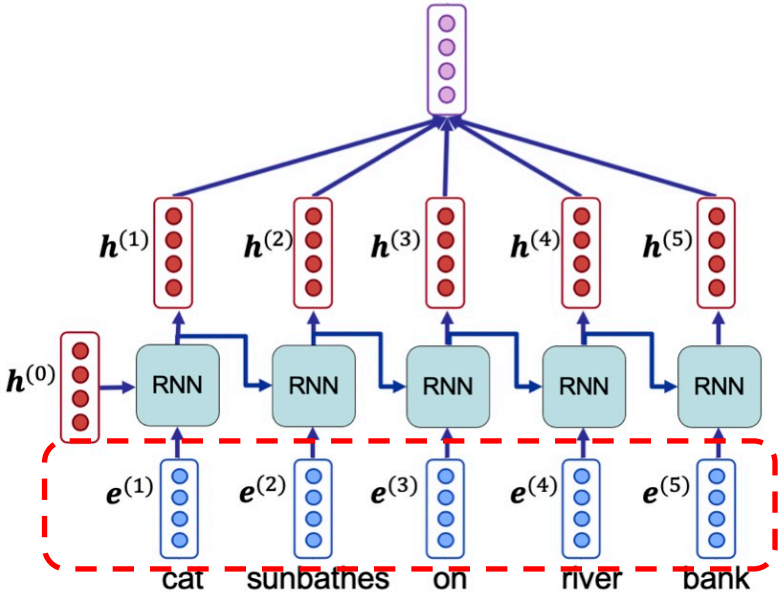
- Addressing bias in NLP requires going beyond CS and getting engaged with disciplines such as sociolinguistics, linguistic anthropology, sociology, law, psychology, etc..

Agenda

- Societal biases in NLP … what? why?
- **Bias in word embedding**
- Measuring bias in Information Retrieval

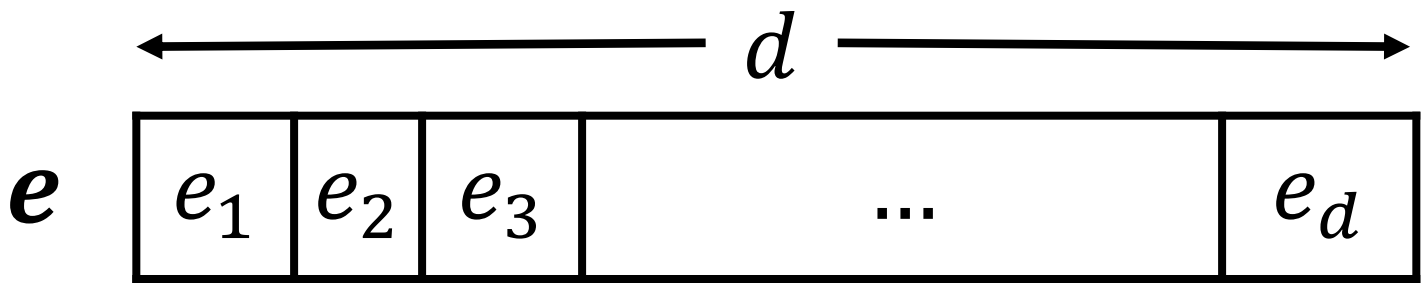
Word Embedding – Building Block of Modern NLP



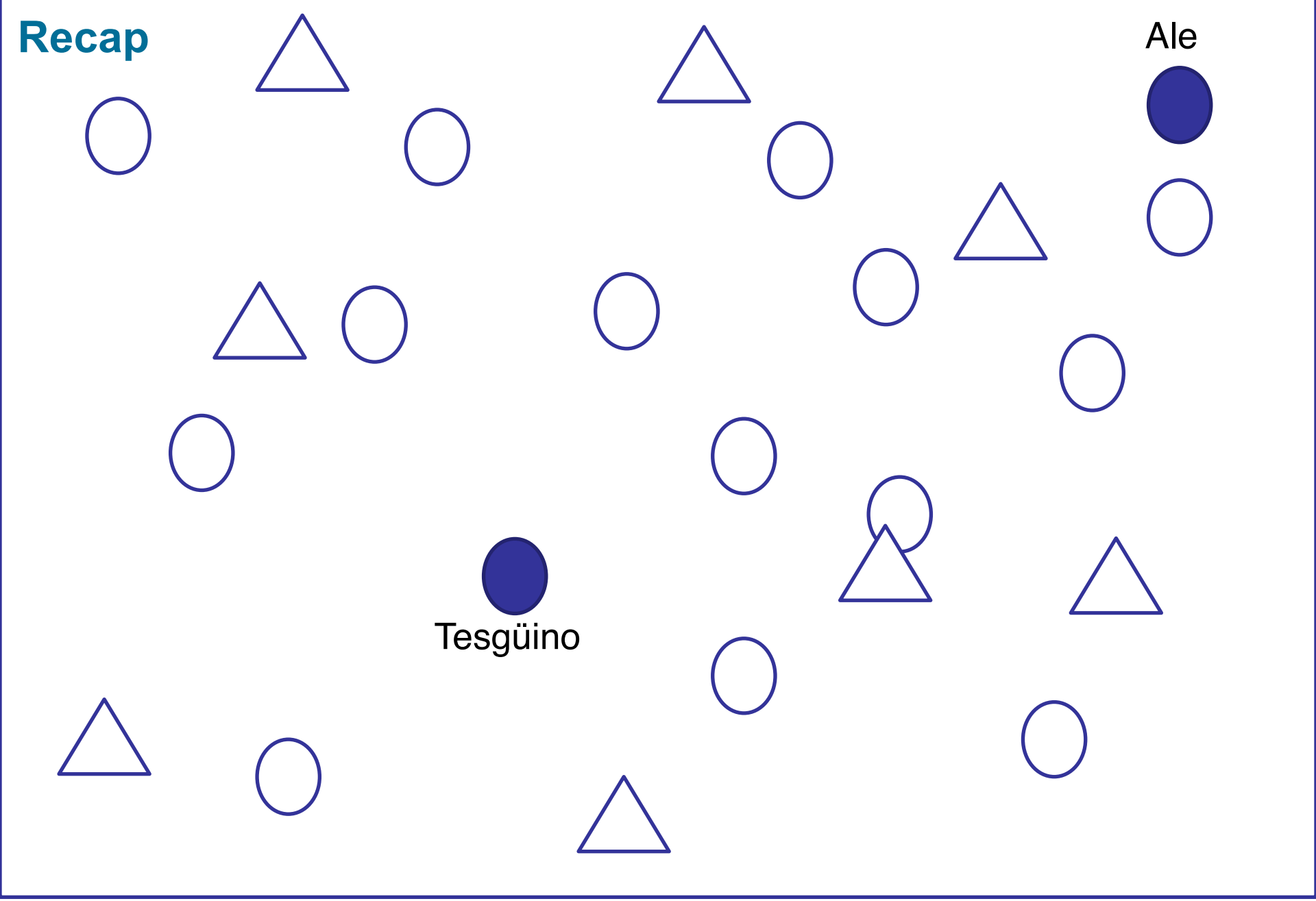


Representation learning and bias

Representation learning encodes information but also may encode the underlying biases in data!



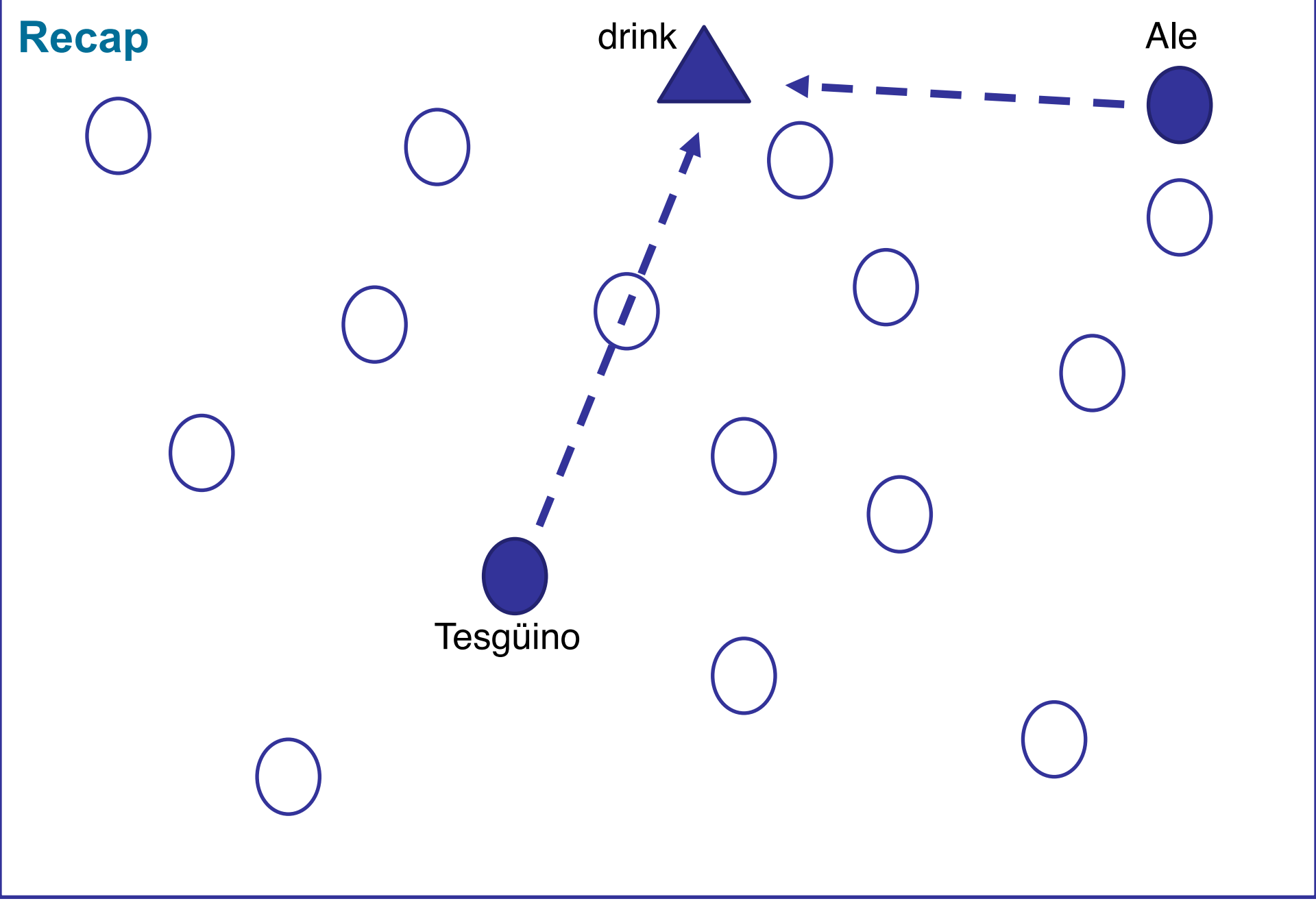
Recap



Embedding vector

Decoding vector

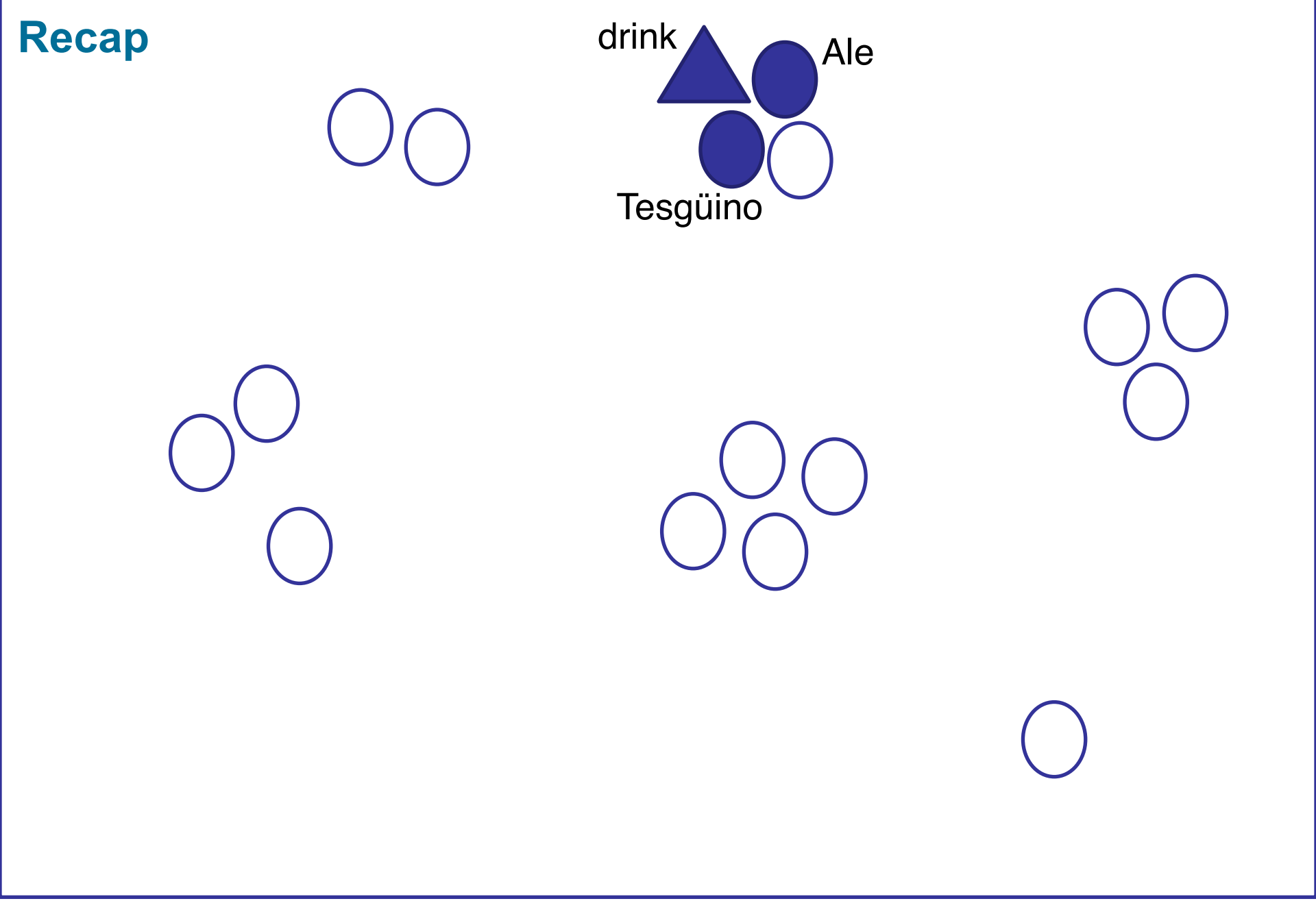
Recap



Embedding vector

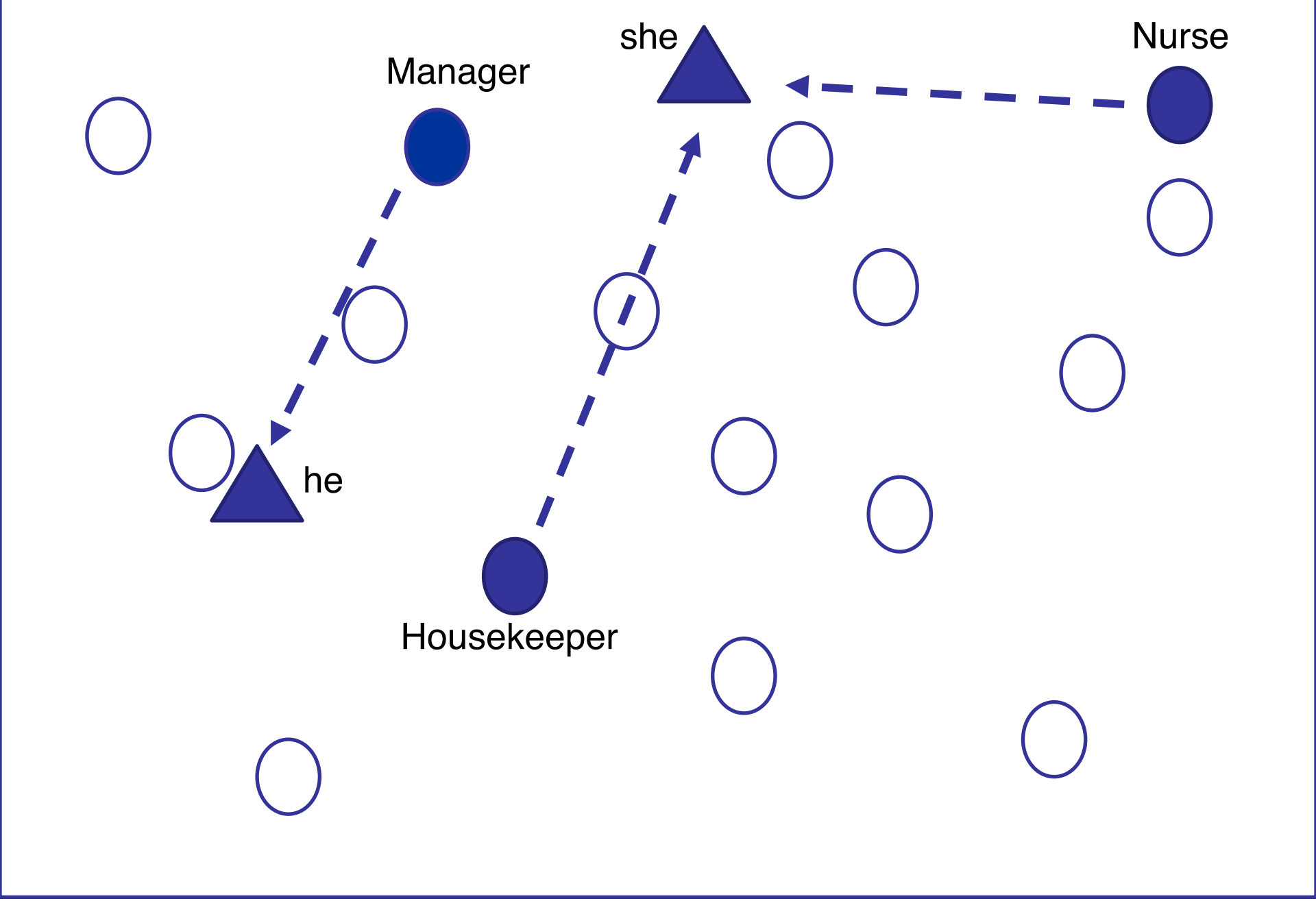
Decoding vector

Recap



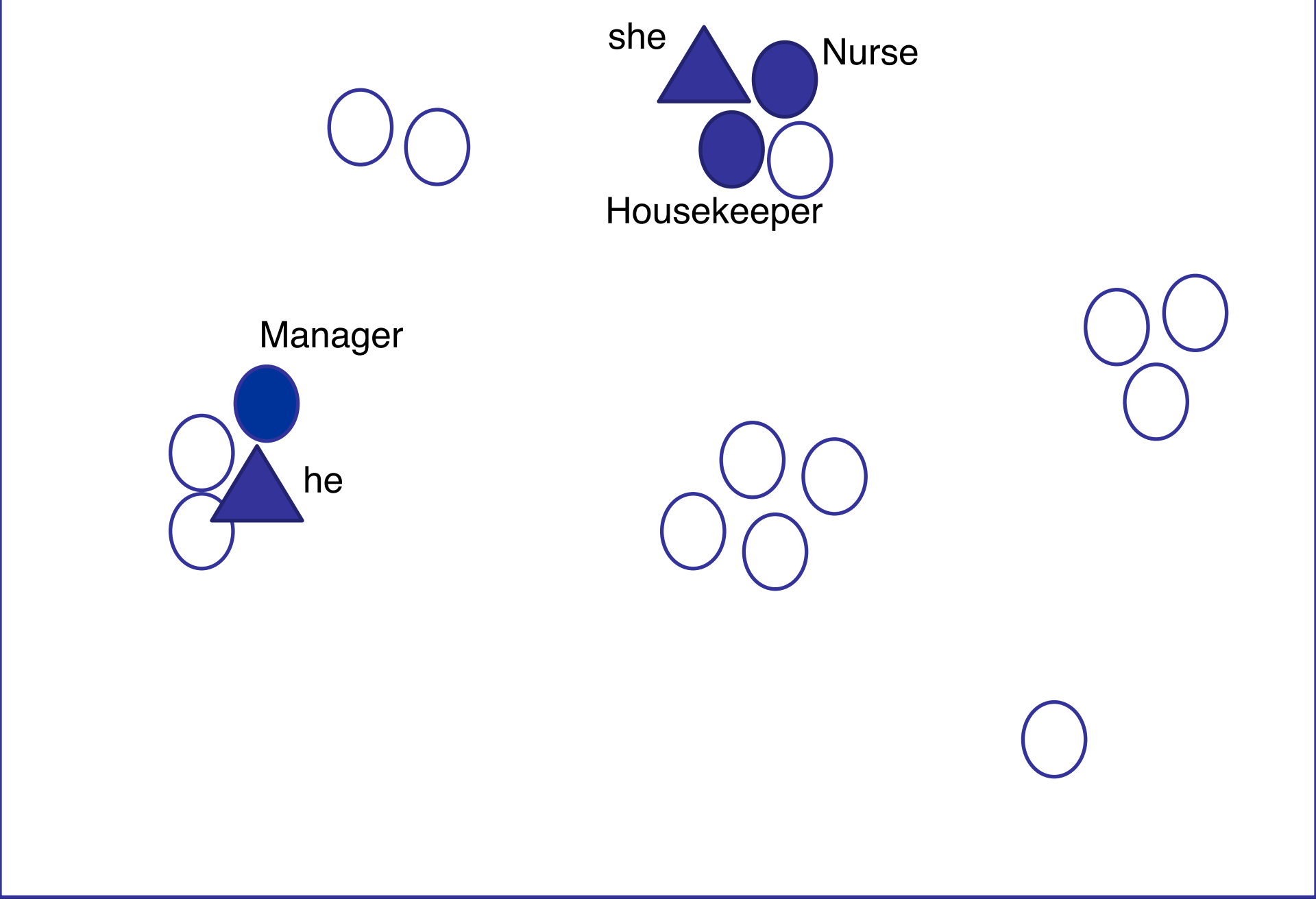
Embedding vector

Decoding vector



Word Vector

Context Vector



Word Vector

Context Vector

Bias in word analogies

- *Recap* – word analogy: ***man*** to ***woman*** is like ***king*** to *? (**queen**)*

$$\boldsymbol{x}_{\text{king}} - \boldsymbol{x}_{\text{man}} + \boldsymbol{x}_{\text{woman}} = \boldsymbol{x}^*$$
$$\boldsymbol{x}^* \approx \boldsymbol{x}_{\text{queen}}$$

- Gender bias is reflected in word analogies

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*

How to measure bias using word embeddings?

- Bias: discrepancy between two concepts $\mathbb{Z}$ and $\tilde{\mathbb{Z}}$, e.g. between female and male*
- A (second-order) bias measurement method:

$$\mathrm{BIAS}_{2\mathrm{ND}}(w) = \frac{1}{|\mathbb{Z}|}\sum_{z\in\mathbb{Z}} \cos\left(\boldsymbol{e}_z, \boldsymbol{e}_w\right) - \frac{1}{|\tilde{\mathbb{Z}}|}\sum_{\tilde{z}\in\tilde{\mathbb{Z}}} \cos\left(\boldsymbol{e}_{\tilde{z}}, \boldsymbol{e}_w\right)$$

- $\boldsymbol{e}_w$ is the vector of word w in a pre-trained word embedding (such as word2vec or GloVe)
- Sample $\mathbb{Z}$ and $\tilde{\mathbb{Z}}$ when measuring bias towards **female**:

$$\mathbb{Z} = \{\textbf{\textit{she}}, \textbf{\textit{her}}, \textbf{\textit{woman}}, \textbf{\textit{girl}}\}$$

$$\tilde{\mathbb{Z}} = \{\textbf{\textit{he}}, \textbf{\textit{him}}, \textbf{\textit{man}}, \textbf{\textit{boy}}\}$$

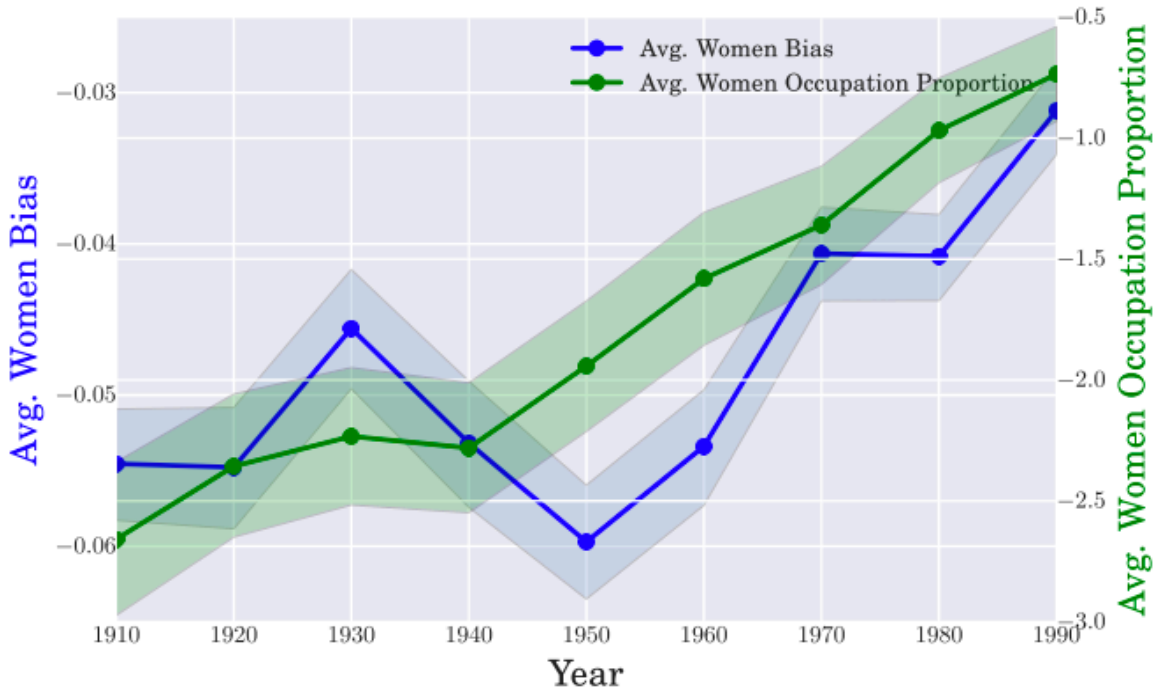
Defining *gender* as a binary construct – namely female vs. male – is an *unpleasant* simplification, as it neglects the wide definition of gender!

Word Embeddings capture societal realities!

(a) Woman occupation proportion vs embedding bias in Google News vectors. More positive indicates more women biased on both axes. $p < 10^{-9}$, r-squared $= .462$.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*

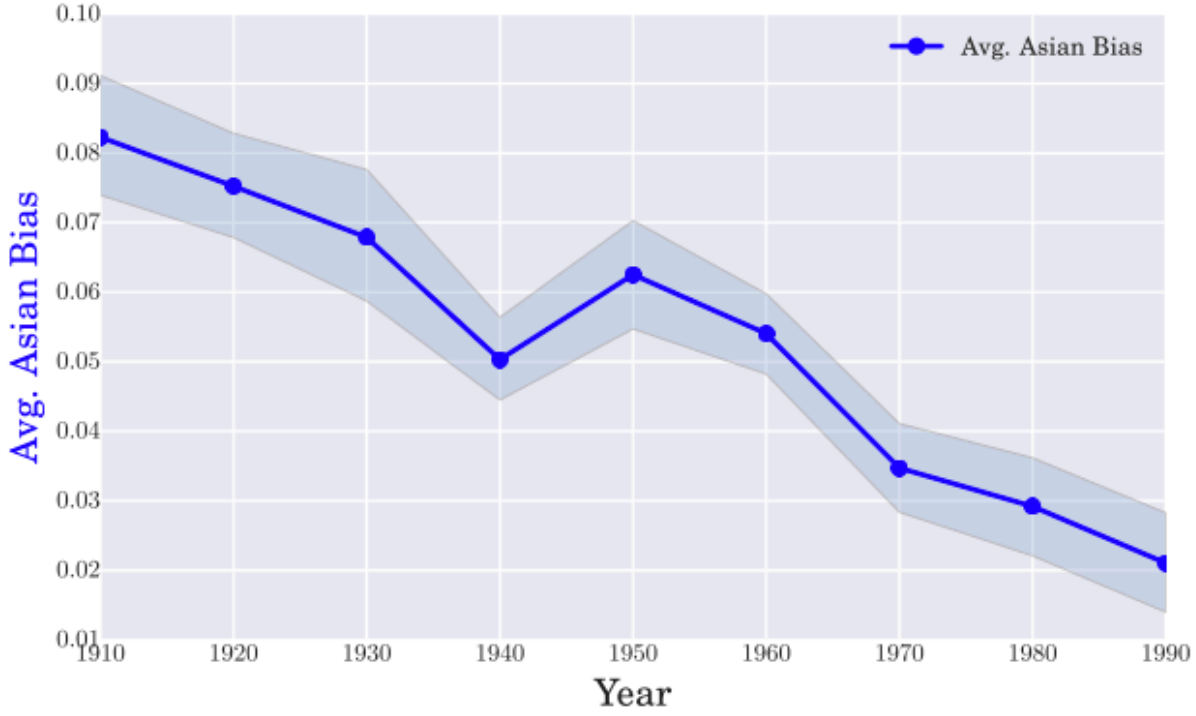
Word Embeddings capture societal realities!



(b) Average gender bias score over time in COHA embeddings in occupations vs the average log proportion. In blue is relative women bias in the embeddings, and in green is the average log proportion of women in the same occupations.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*

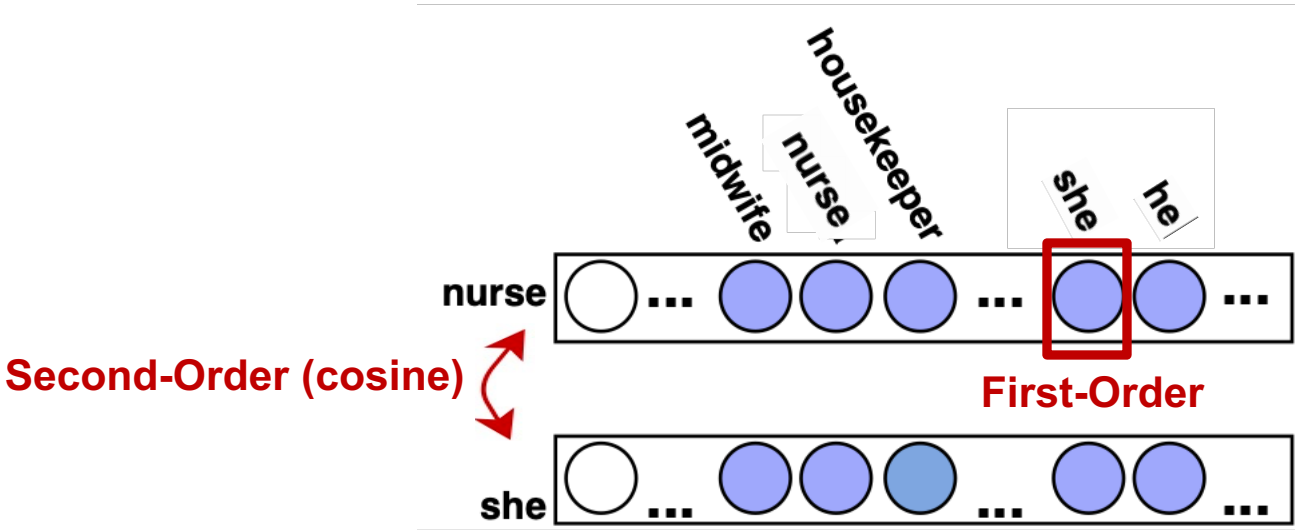
Word Embeddings capture societal realities!



(c) Asian bias score over time for words related to the outsiders in COHA data.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*

Bias Measurement Methods

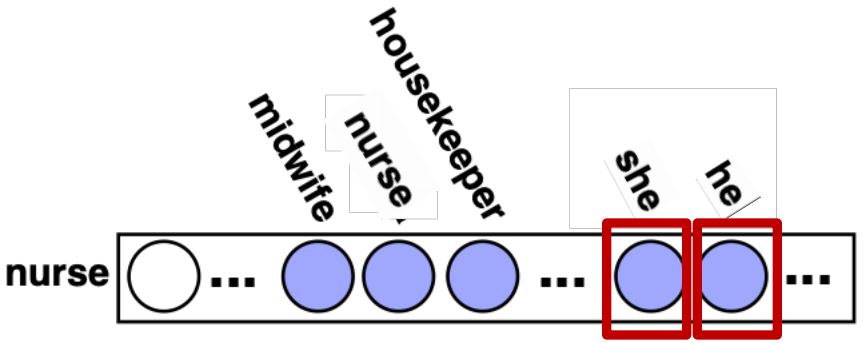


- **First-Order co-occurrence**: co-occurrence of two words in the same context
  - Examples: ***drink*** to ***beer*** or ***drink*** to ***wine***
- **Second-Order co-occurrence:** similarity of two word vectors → usually by using cosine
  - Examples: ***beer*** to ***wine***, or ***Tesgüino*** to ***Märzen***

Bias measurement using word embeddings

- Recap: Second-order bias measurement for word w:

$$\text{BIAS}_{\text{2ND}}(w) = \frac{1}{|\mathbb{Z}|}\sum_{z\in\mathbb{Z}} \cos\left(\boldsymbol{e}_z, \boldsymbol{e}_w\right) - \frac{1}{|\tilde{\mathbb{Z}}|}\sum_{\tilde{z}\in\tilde{\mathbb{Z}}} \cos\left(\boldsymbol{e}_{\tilde{z}}, \boldsymbol{e}_w\right)$$



- First-order bias measurement for word w:

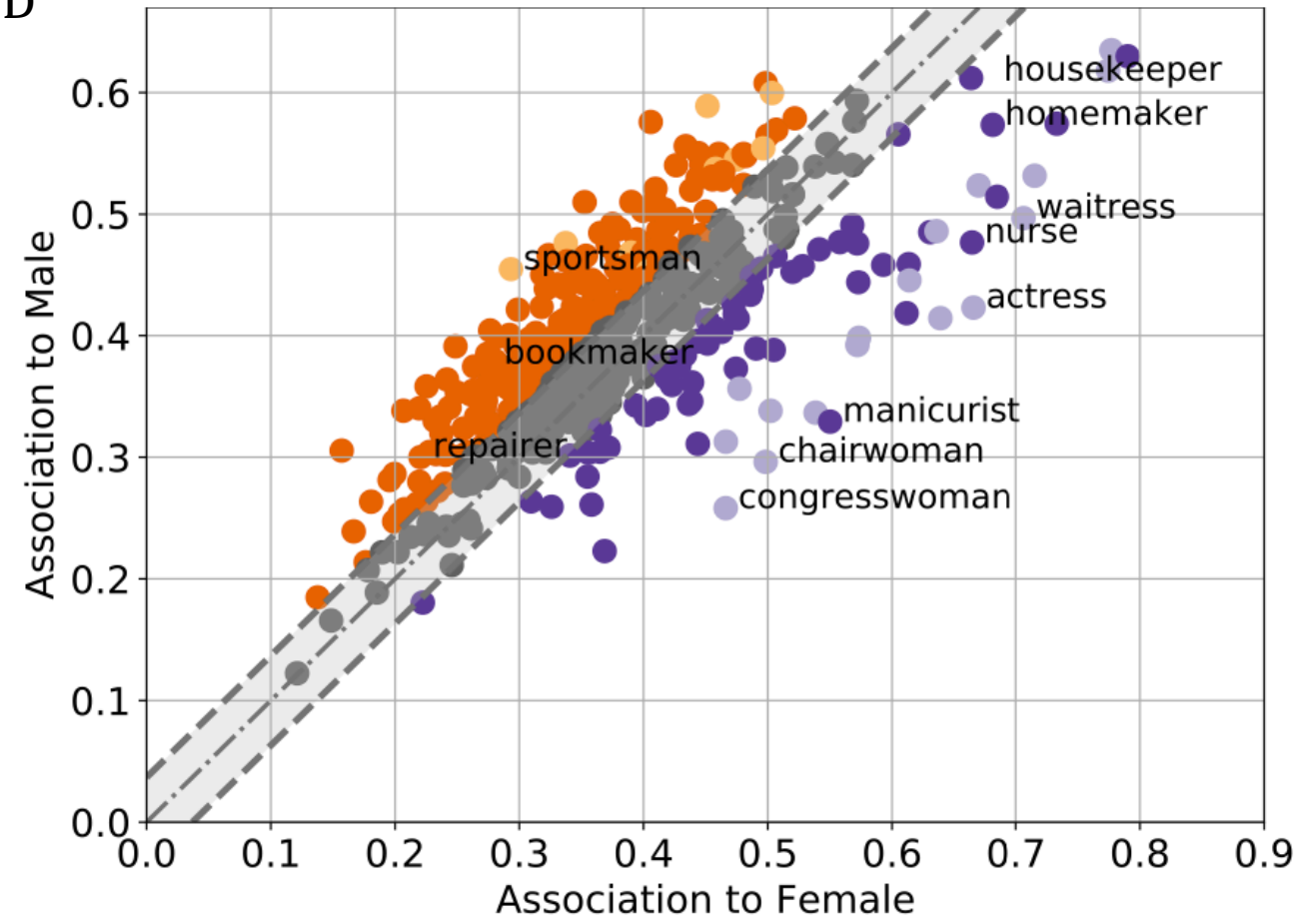
$$\text{BIAS}_{\text{1ST}}(w) = \frac{1}{|\mathbb{Z}|}\sum_{z\in\mathbb{Z}} f\left(z, w\right) - \frac{1}{|\tilde{\mathbb{Z}}|}\sum_{\tilde{z}\in\tilde{\mathbb{Z}}} f\left(\tilde{z}, w\right)$$

f is a measure of first-order co-occurrence between two words, e.g. PPMI

- f can also be achieved from reconstructing co-occurrence matrix from trained word embeddings (see paper for details)

Measuring societal biases with second-order co-occurrences

BIAS_{2ND}

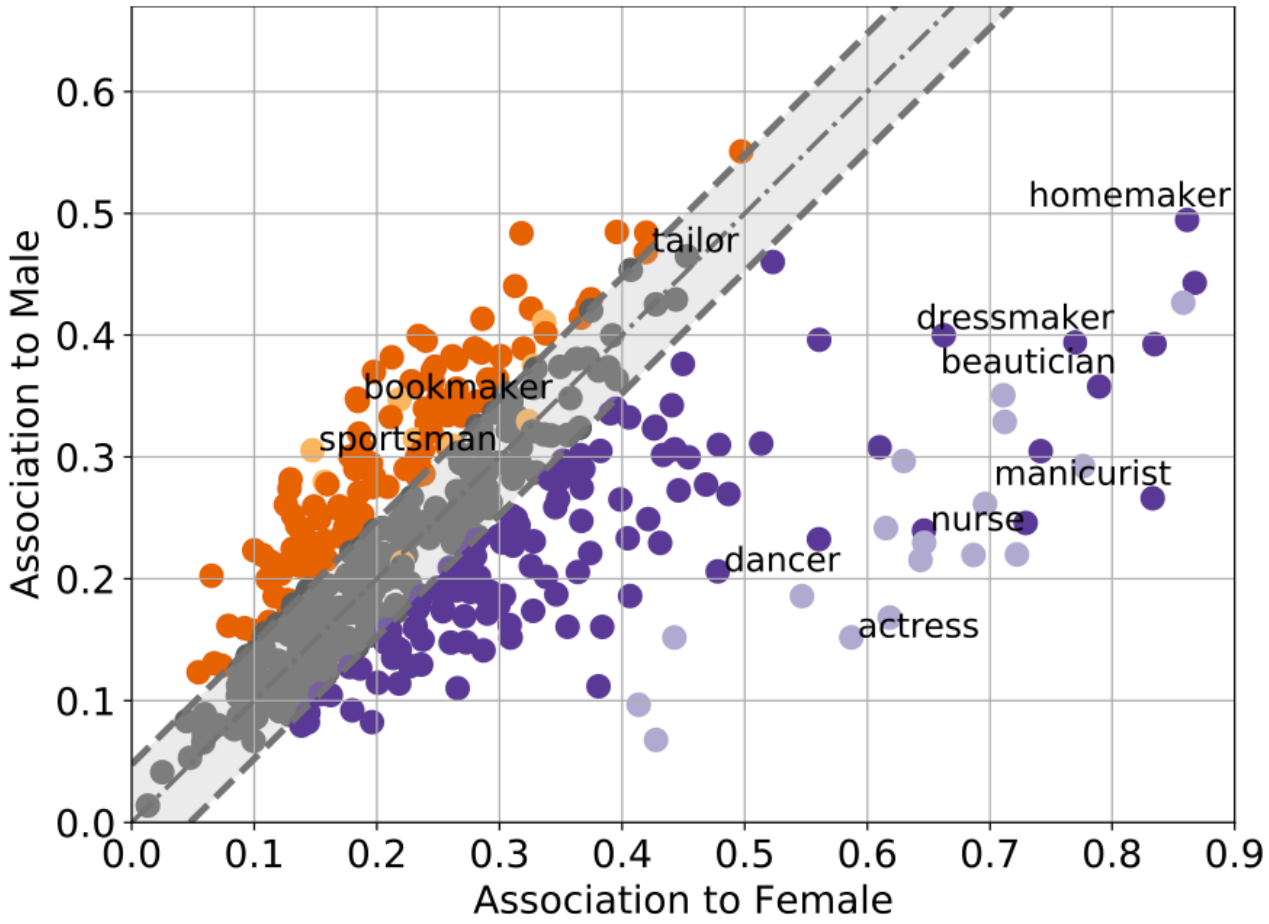


Associations are measured using a word2vec model, trained on a recent Wikipedia corpus

Rekabsaz N., Henderson J., West R., and Hanbury A. "Measuring Societal Biases in Text Corpora via First-Order Co-occurrence." *arXiv preprint arXiv:1812.10424* (2020).

Measuring societal biases with first-order co-occurrences

BIAS_{1ST}



Associations are measured using a word2vec model, trained on a recent Wikipedia corpus

Rekabsaz N., Henderson J., West R., and Hanbury A. "Measuring Societal Biases in Text Corpora via First-Order Co-occurrence." *arXiv preprint arXiv:1812.10424* (2020).

Correlations with job market statistics

Order	Representation	Method	Labor Data		Census Data	
			Spearman ρ	Pearson's r	Spearman ρ	Pearson's r
High-Order	PMI	DIRECTIONAL	0.28	0.07	0.18	0.02
		CENTROID	0.14	0.21	0.35	0.40
		$\text{AVERAGE}_{\text{HIGH}}$	0.33	0.24	0.27	0.19
	PMI-SVD	DIRECTIONAL	0.05	0.07	0.00	0.00
		CENTROID	0.41	0.47	0.46	0.53
		$\text{AVERAGE}_{\text{HIGH}}$	0.41	0.49	0.49	0.56
First-Order	PMI	$\text{AVERAGE}_{\text{FIRST}}$	**0.53**	**0.51**	**0.57**	**0.62**
High-Order	PPMI	DIRECTIONAL	0.45	0.49	0.39	0.47
		CENTROID	0.43	0.46	0.45	0.50
		$\text{AVERAGE}_{\text{HIGH}}$	0.43	0.46	0.45	0.52
	PPMI-SVD	DIRECTIONAL	0.05	0.07	0.00	0.00
		CENTROID	0.41	0.47	0.46	0.53
		$\text{AVERAGE}_{\text{HIGH}}$	0.41	0.49	0.49	0.56
First-Order	PPMI	$\text{AVERAGE}_{\text{FIRST}}$	**0.59**	**0.58**	**0.64**	**0.64**
High-Order	SPPMI	DIRECTIONAL	0.26	0.37	0.26	0.28
		CENTROID	0.39	0.45	0.45	**0.48**
		$\text{AVERAGE}_{\text{HIGH}}$	0.32	0.40	0.44	**0.48**
	SPPMI-SVD	DIRECTIONAL	0.17	0.29	0.11	0.03
		CENTROID	0.28	0.35	0.39	0.43
		$\text{AVERAGE}_{\text{HIGH}}$	0.26	0.38	0.36	0.46
First-Order	SPPMI	$\text{AVERAGE}_{\text{FIRST}}$	**0.57**	**0.49**	**0.52**	**0.48**
High-Order	GloVe	DIRECTIONAL	0.53	0.56	0.34	0.46
		CENTROID	0.58	0.60	0.39	0.51
		$\text{AVERAGE}_{\text{HIGH}}$	**0.60**	**0.60**	0.39	0.51
First-Order	initGlove	$\text{AVERAGE}_{\text{FIRST}}$	0.38	0.42	0.40	0.51
	eGloVe		0.56	0.57	**0.42**	**0.52**
High-Order	SG	DIRECTIONAL	0.50	0.54	0.58	0.64
		CENTROID	0.55	0.57	0.60	0.65
		$\text{AVERAGE}_{\text{HIGH}}$	0.55	0.57	0.59	0.65
First-Order	eSG	$\text{AVERAGE}_{\text{FIRST}}$	**0.66**	**0.61**	**0.67**	**0.70**

Correlation results of the gender bias values, calculated with word representations, to the statistics of the portion of women in occupations

Summary

What we know so far …

- Word embeddings capture and encode societal biases, reflected in the underlying corpora
  - These biases also exist in contextualized word embeddings
- Word embeddings enable the study of societal phenomena
  - e.g. monitoring how gender/ethnicity/etc. is perceived during time

Subsequent question:

- What about bias in down-stream NLP tasks?
  - Existence of bias could become problematic in many NLP tasks such as *job search, content-based recommendation systems, IR, sentiment analysis, etc.*

- Since the pre-trained word embeddings are widely used in NLP tasks, are biases in word embeddings also transferred to the tasks?

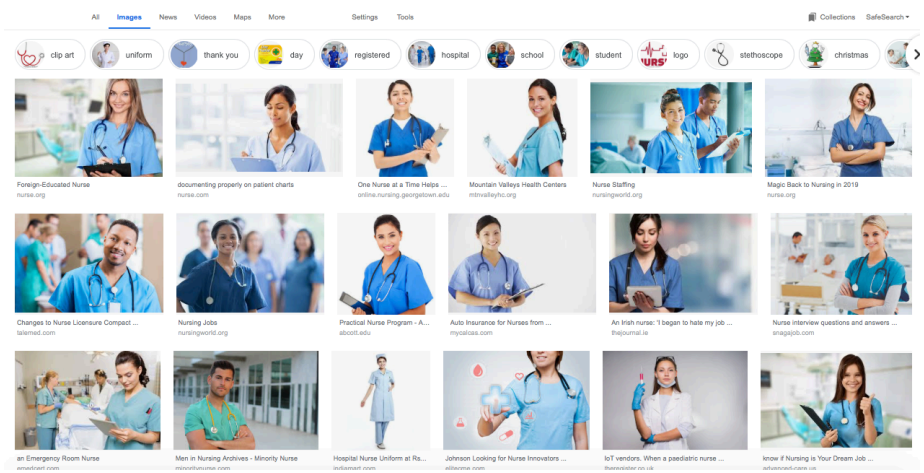
Agenda

- Societal biases in NLP … what? why?
- Bias in word embedding
- **Measuring bias in Information Retrieval**

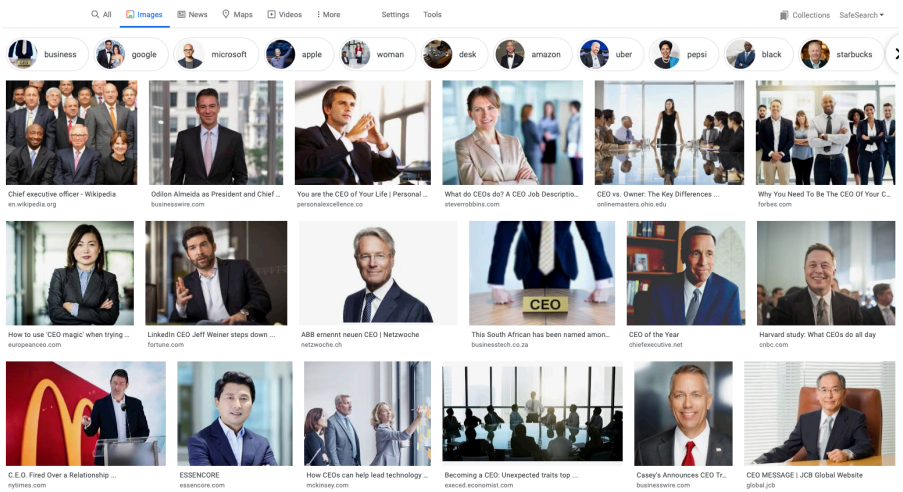
Gender bias in IR

- Search engines define our perception of reality!
  - Search engines' results are typically taken be users as the "state of the world"
- Search engines reflect the existing (harmful) stereotypes in society but also reinforce them
- This eventually leads to representational harms

Search: Nurse



Search: CEO



Gender bias in IR

- How to measure gender bias, especially according to the content of retrieved documents?

  We need a framework to measure gender bias in the retrieval results

- What is the effect of collection / ranking models / transfer learning on gender bias?

  Experiments on various neural IR models

Do Neural Ranking Models Intensify Gender Bias? *Rekabsaz N., Schedl M.. To be appeared in the proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR) 2020, https://arxiv.org/abs/2005.00372*

Step1: Gendered vs. non-gendered queries annotation

- Sets of queries, annotated by Amazon Turkers according to their gender-related contents
  - Queries are taken from MS MARCO dev set

- *Non-gendered queries*: queries with no indication of gender

	Non-gendered (1765)	what is a synonym for beautiful what is the meaning of resurrect
Gendered	*Female* (742)	who was oprah winfrey earliest pregnancy symptoms
	Male (1202)	where is martin luther king jr's place who was the king of ancient rome
	Other or Multiple Genders (41)	is g dragon gay how long was shakespeare married to anne

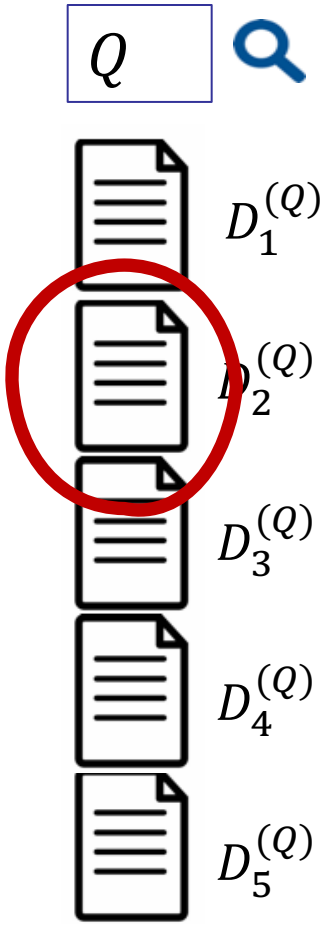
Step 2: document female/male magnitude

- In what extent each document D contains female/male topics?
- $\gamma^{\mathbb{Z}}(D)$: the magnitude of concept $\mathbb{Z}$ in document D

$$\gamma^{\mathbb{Z}}(D) = \sum_{t \in \mathbb{Z}} \log \mathrm{tc}_{t,D}$$

$$\gamma^{\tilde{\mathbb{Z}}}(D) = \sum_{t \in \tilde{\mathbb{Z}}} \log \mathrm{tc}_{t,D}$$

  - $\mathbb{Z}$ Male definitional words: ***he***, ***him***, ***man***, ***boy***, etc.
  - $\tilde{\mathbb{Z}}$ Female definitional words : ***she***, ***her***, ***woman***, ***girl***, etc.



** Limitation: gender is conceived as a binary construct!*

Step 3: ranking bias metric

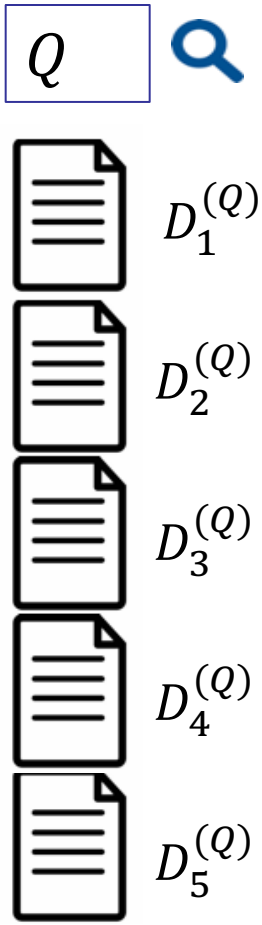
- Rank Bias (RaB) metric
  - For the given query Q :

$$\mathrm{qRaB}(Q) = \frac{1}{t}\sum_{i=1}^{t} \gamma^{\mathbb{Z}}\left(D_i^{(Q)}\right) - \gamma^{\tilde{\mathbb{Z}}}\left(D_i^{(Q)}\right)$$

- RaB of a model is the average over the queries $\mathbb{Q}$:

$$\mathrm{RaB} = \frac{1}{|\mathbb{Q}|}\sum_{Q\in\mathbb{Q}} \mathrm{qRaB}(Q)$$

- Another metric: Average Ranking Bias (ARaB)

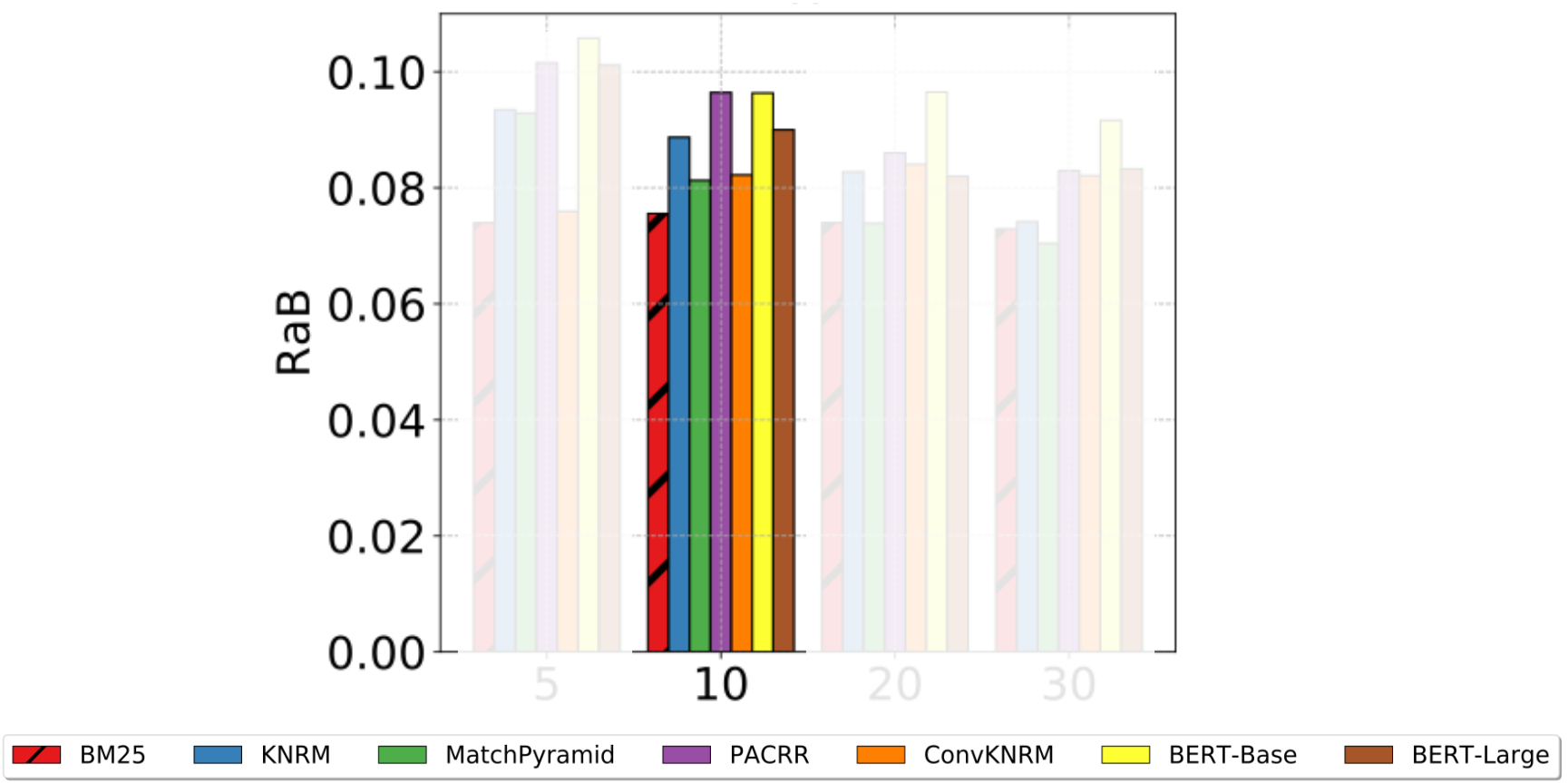


Experiment design

- Ranking models
  - BM25
  - Various neural ranking models
    - Word embeddings are initialized with GloVe or Randomly (*RND*)
  - BERT models
- Evaluated on MS-MARCO passage retrieval

Ranking Model	Model Parameters		Evaluation	
	All	Transferred	MRR	Recall
BM25	–		0.192	0.398
KNRM$_{RND}$	109,481,411	none	0.213	0.390
KNRM		GloVe	0.230	0.439
MatchPyramid$_{RND}$	109,539,960	none	0.232	0.424
MatchPyramid		GloVe	0.240	0.445
PACRR$_{RND}$	109,875,938	none	0.228	0.426
PACRR		GloVe	0.242	0.451
ConvKNRM$_{RND}$	110,022,399	none	0.243	0.443
ConvKNRM		GloVe	0.268	0.488
BERT-Base	109,483,778	all	0.342	0.585
BERT-Large	335,143,938	all	0.353	0.596

Results



- All models show an overall bias towards male
- Neural models show higher gender bias in comparison with BM25!
  - Especially, fine-tuned BERT models show higher bias than other neural models

Effect of transfer learning

	Cut-off: 10			
BM25	0.076	0.076	0.039	0.031
KNRM	0.089 (↓0.020)	0.090 (↓0.004)	0.041 (↓0.007)	0.037 (↑0.003)
MatchPyramid	0.081 (↓0.002)	0.089 (↓0.006)	0.047 (↓0.007)	0.047 (↓0.007)
PACRR	0.096 (↓0.008)	0.104 (↓0.019)	0.043 (↑0.002)	0.042 (↓0.001)
ConvKNRM	0.082 (↓0.001)	0.066 (↑0.009)	0.045 (↓0.005)	0.037 (↓0.004)
BERT-Base	0.096	0.102	0.056	0.057
BERT-Large	0.090	0.100	0.052	0.054

- Arrows show the changes in RaB, when word embeddings are initialized randomly instead of initialization with a pre-trained GloVe
- Models without transfer learning show smaller degrees of gender bias → Transfer learning increases gender bias!

Bias in IR – Summary

- We see a framework to measure gender bias in the content of the retrieved results
- Neural ranking models, especially BERT models, increase gender bias
- Transfer learning increases bias

Final words: About “debiasing”

- **Debiasing**: methods to mitigate or reduce bias
  - The aim is to make the output or decision of a model agnostic to sensitive features (such as gender, race, ethnicity, age)

- Debiasing approaches in NLP literature address …
  - Datasets / Corpora: by changing/adding/removing data in collection
  - Models
    - By adding debiasing/fairness criteria to model’s objective function
    - By training adversarial networks to remove sensitive information in learned representations
  - Output results: by post-processing model’s outputs