Natural Language Processing with Deep Learning

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

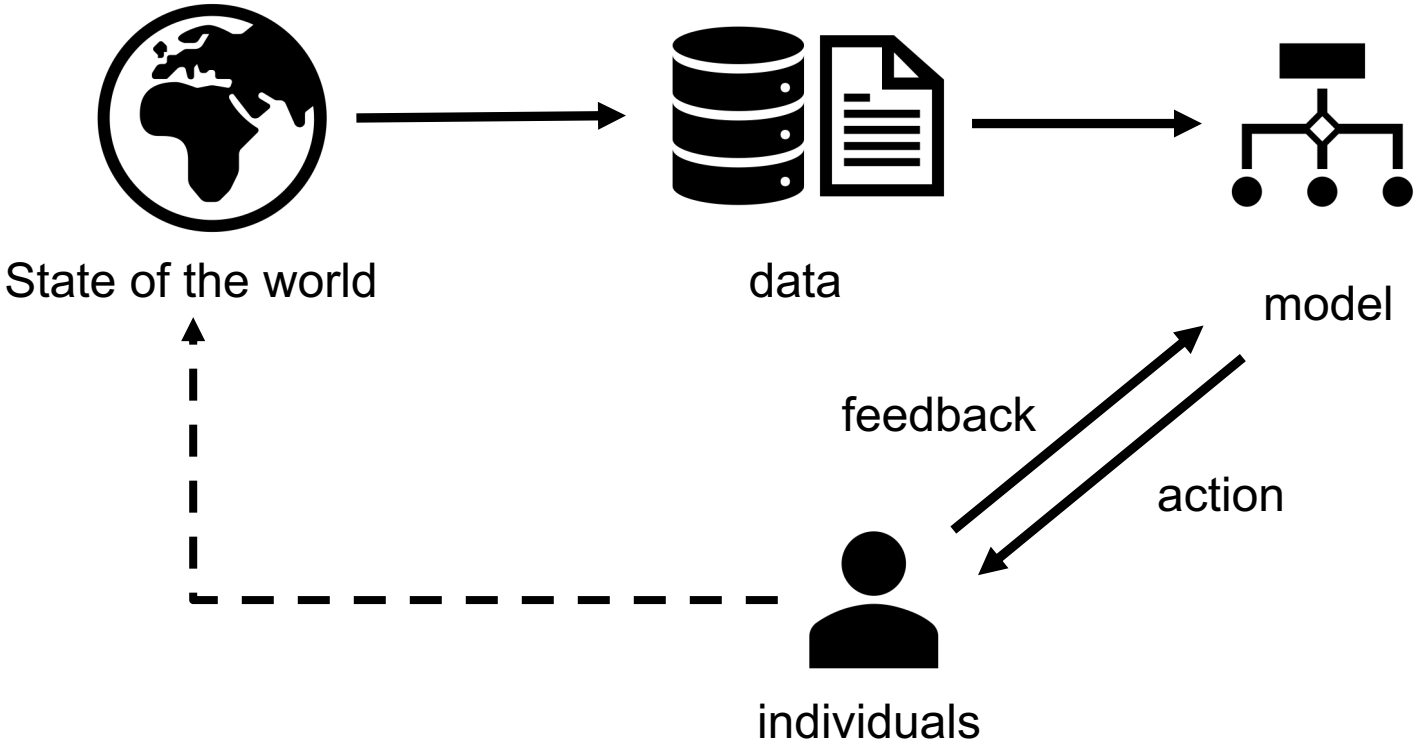
- Motivation
- Bias in word embeddings
- Bias in IR

Agenda

- **Motivation**
- Bias in word embeddings
- Bias in IR

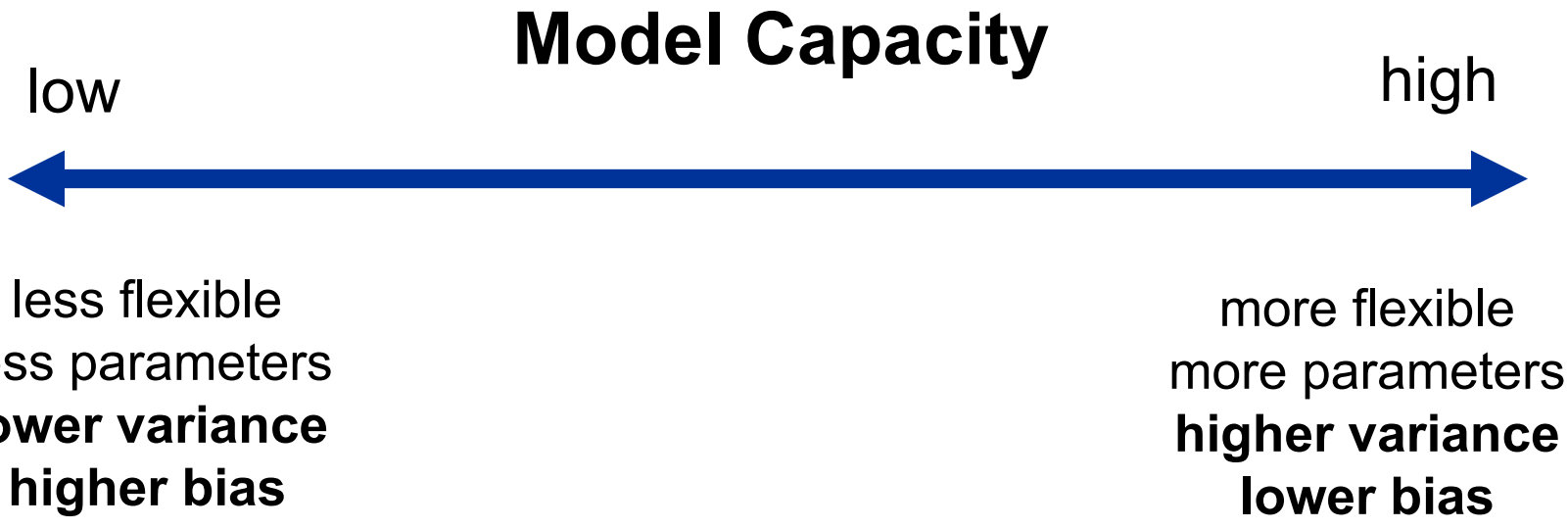
"I think your test grading is biased in favor of students who answer the test questions correctly."

Machine Learning Cycle



- Societal biases in the world are reflected in data,and consequently transferred to the model, its predictions and final decisions

Recap: (Statistical) bias in ML



Statistical Bias indicates the amount of assumptions, taken to define a model. Higher bias means more assumptions and less flexibility, as in linear regression.

(Societal) Bias

"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

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

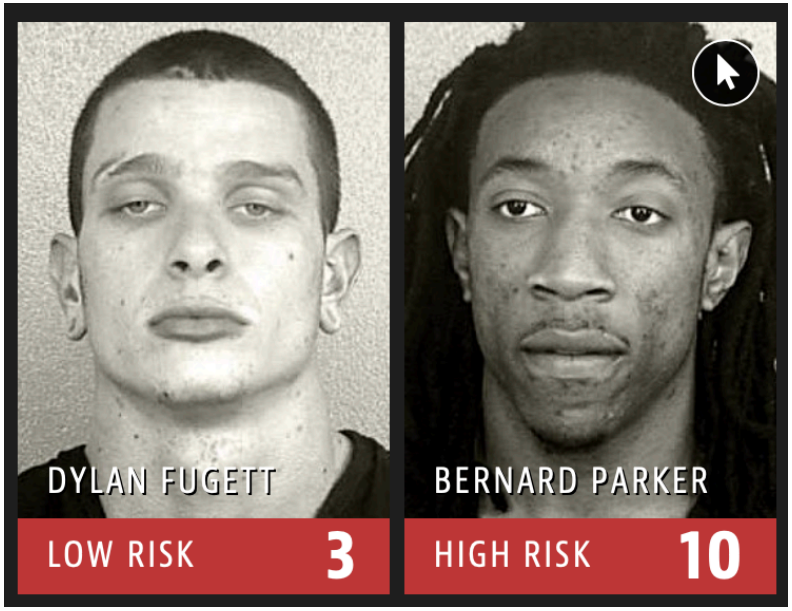
https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology

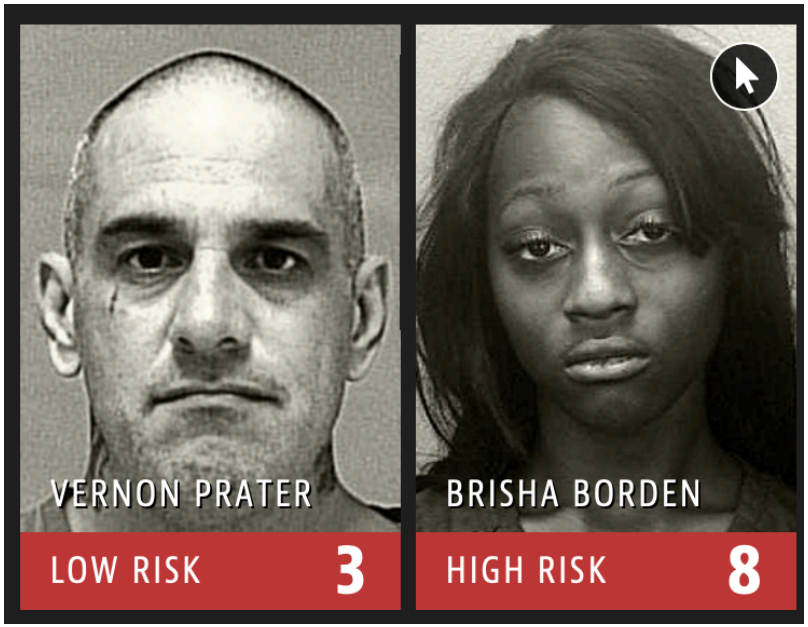
Beauty.AI's 'robot beauty contest' is back – and this time it promises not to be racist

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 IR

Search: nurse

All Images News Videos Maps More Settings Tools Collections SafeSearch

clip art | uniform | thank you | day | registered | hospital | school | student | logo | stethoscope | christmas


Foreign-Educated Nurse
nurse.org


documenting properly on patient charts
nurse.com


One Nurse at a Time Helps ...
online.nursing.georgetown.edu


Mountain Valleys Health Centers
mtnvalleyhc.org


Nurse Staffing
nursingworld.org


Magic Back to Nursing in 2019
nurse.org


Changes to Nurse Licensure Compact ...
talemed.com


Nursing Jobs
nursingworld.org


Practical Nurse Program - A...
abcott.edu


Auto Insurance for Nurses from ...
mycalcas.com


An Irish nurse: 'I began to hate my job ...
thejournal.ie


Nurse interview questions and answers ...
snagajob.com


an Emergency Room Nurse
emedcert.com


Men in Nursing Archives - Minority Nurse
minoritynurse.com

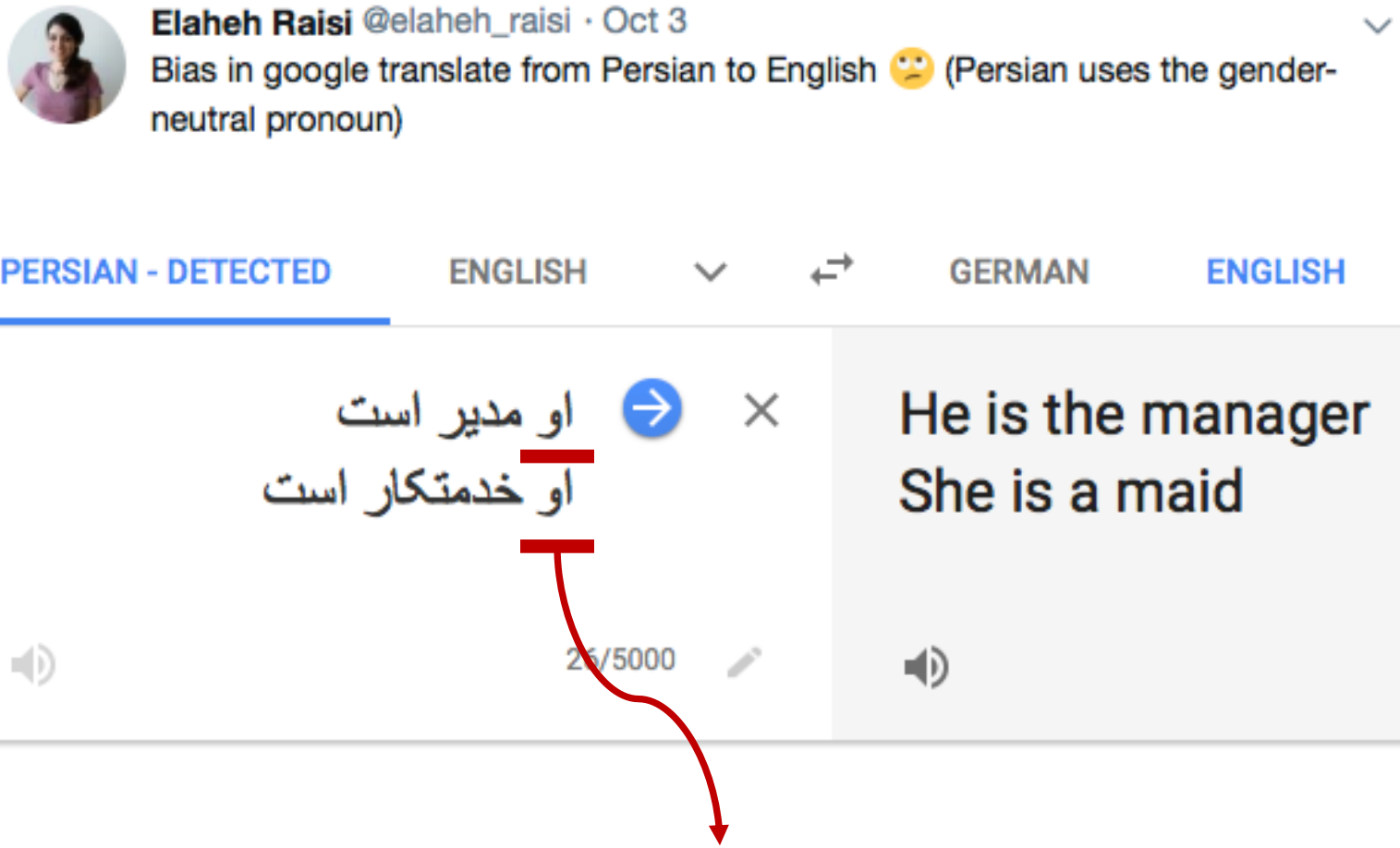

Hospital Nurse Uniform at Rs...
indiamart.com


Johnson Looking for Nurse Innovators ...
elitecme.com


IoT vendors. When a paediatric nurse ...
theregister.co.uk


know if Nursing is Your Dream Job ...
advanced-care.us

Bias in Machine Translation



Why does it matter?

- Legal: information access – especially in settings like employment, housing, and public accommodation – potentially is covered by anti-discrimination laws, such as EU Anti-Discrimination law
- Publicity: disclosure of systematic bias in system performance can undermine trust in information access
- Financial: underperformance for large segments of users leads to abandonment
- Moral: professional responsibility to provide equal information access

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

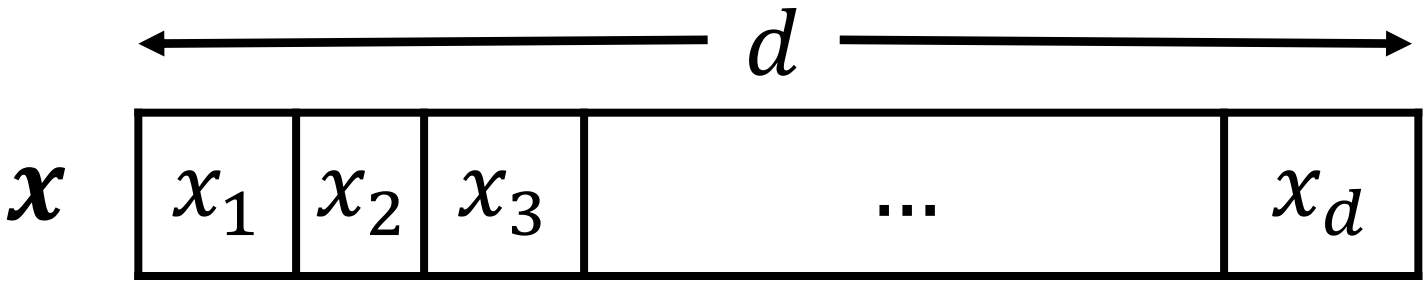
Where is it originated from?

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

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

Representation learning and bias

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



E.g. the learned representation of word ***nurse*** may convey that its encoded implicit meaning is about being ***woman***!

Bias & Fairness in ML vs. NLP

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

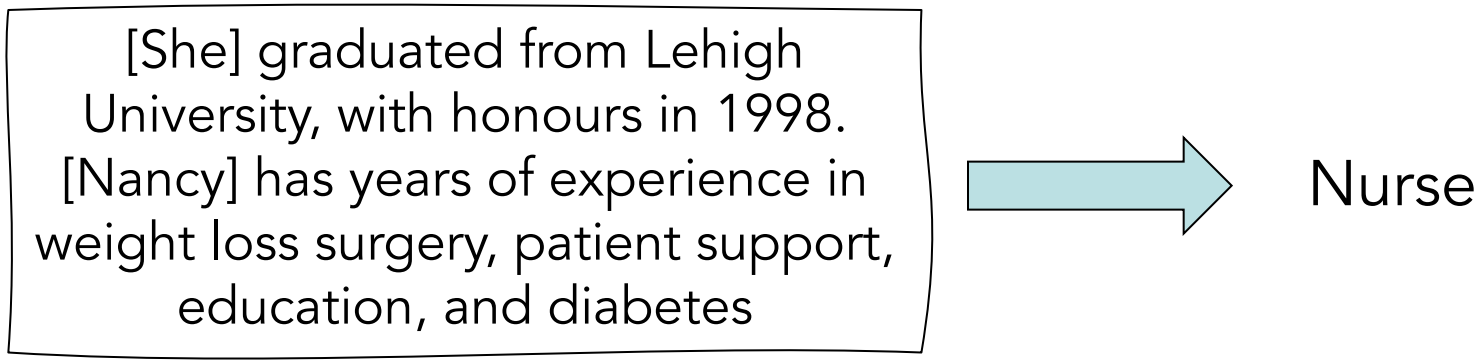
```
39, State-gov, 77516, Bachelors, 13, Never-married, Adm-clerical, Not-in-family, White, Male, 2174, 0, 40, United-States, <=50K
50, Self-emp-not-inc, 83311, Bachelors, 13, Married-civ-spouse, Exec-managerial, Husband, White, Male, 0, 0, 13, United-States, <=50K
38, Private, 215646, HS-grad, 9, Divorced, 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 ML vs. 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.

Final words!

Big problems need interdisciplinary thinking!

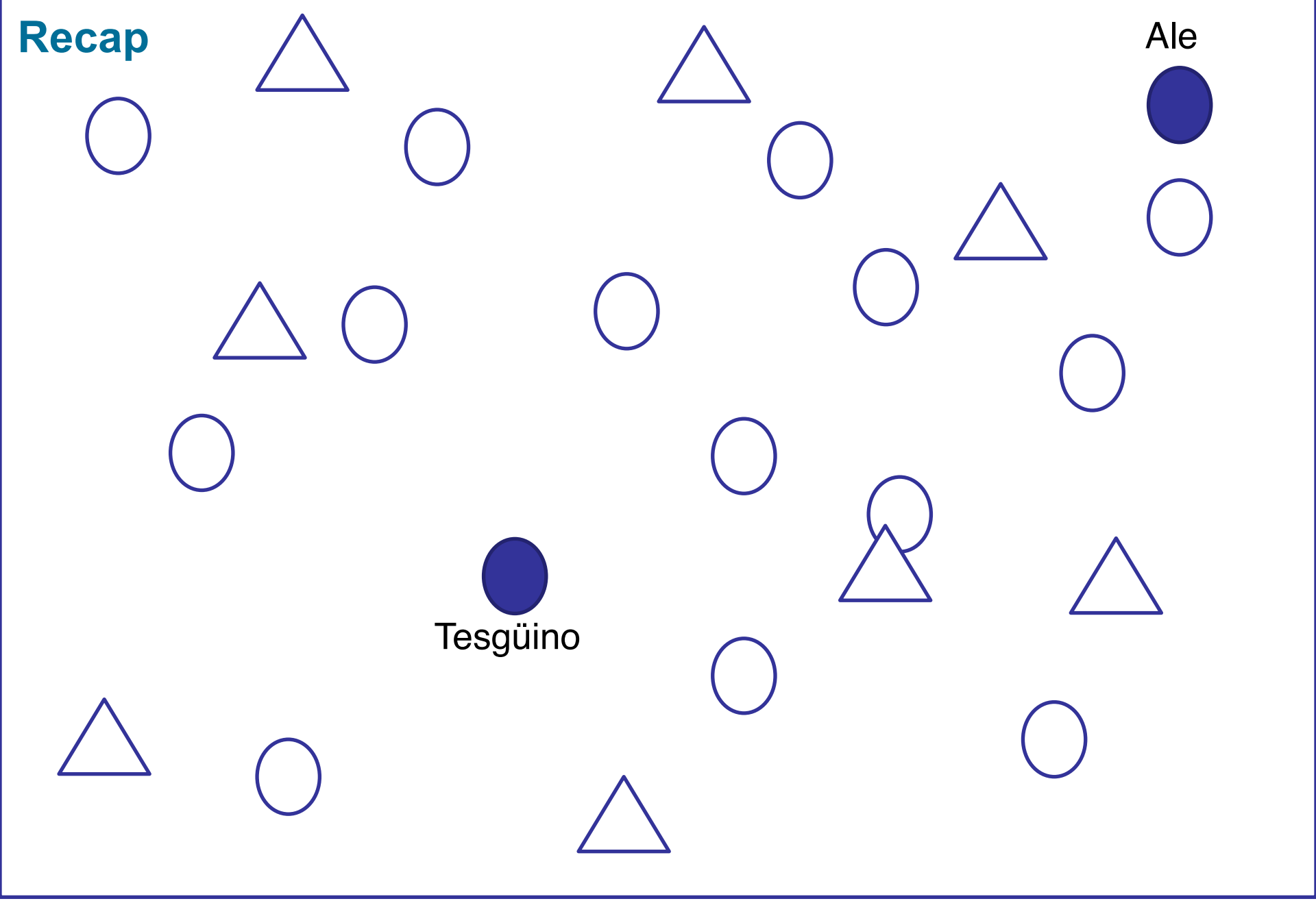
- Fairness and bias are social concepts and inherently normative

- Engaging with these problems requires going beyond CS:
  - Law
  - Ethics / philosophy
  - Sociology
  - Political science
  - …

Agenda

- Motivation
- **Bias in word embeddings**
- Bias in IR

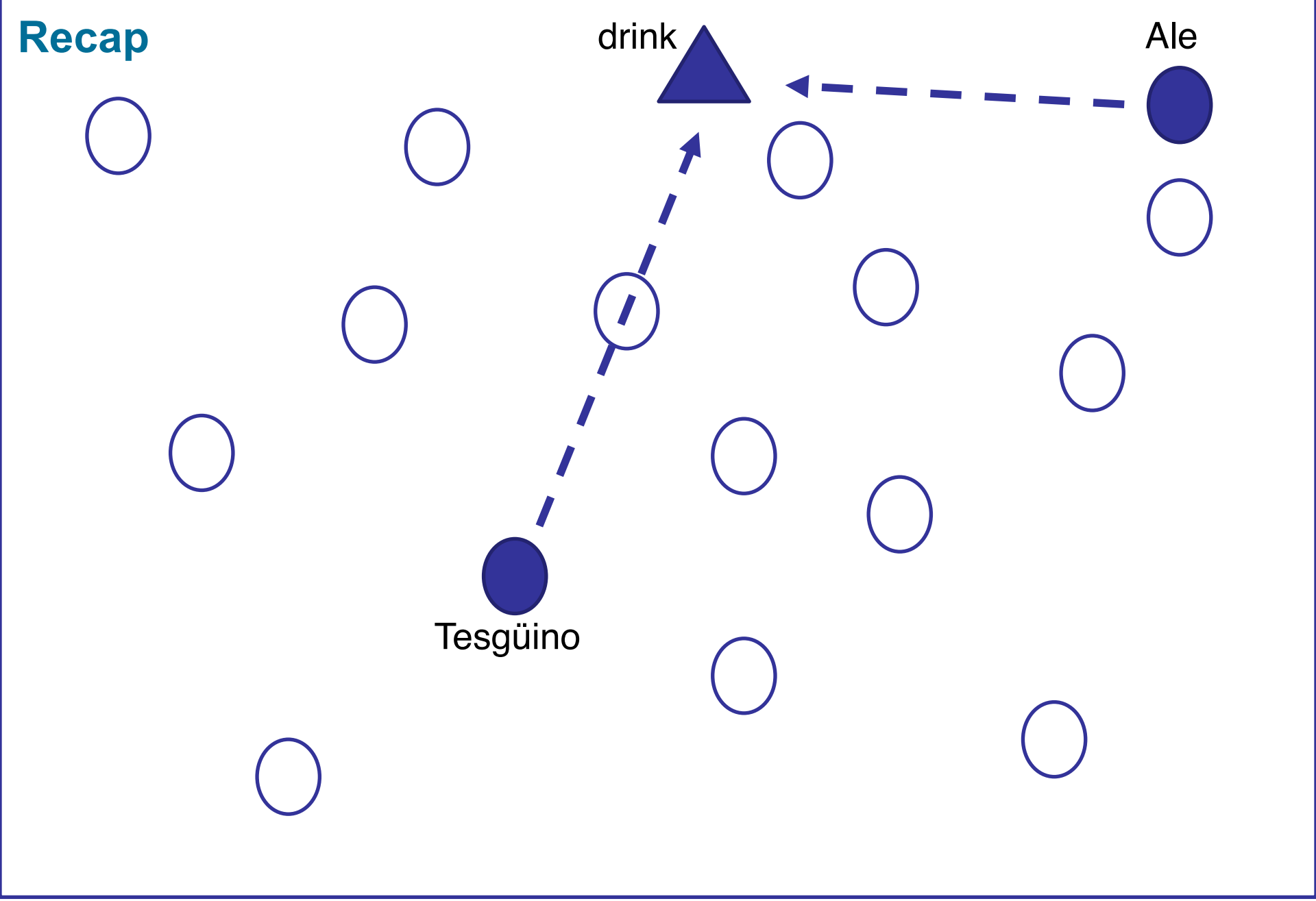
Recap



Embedding vector

Decoding vector

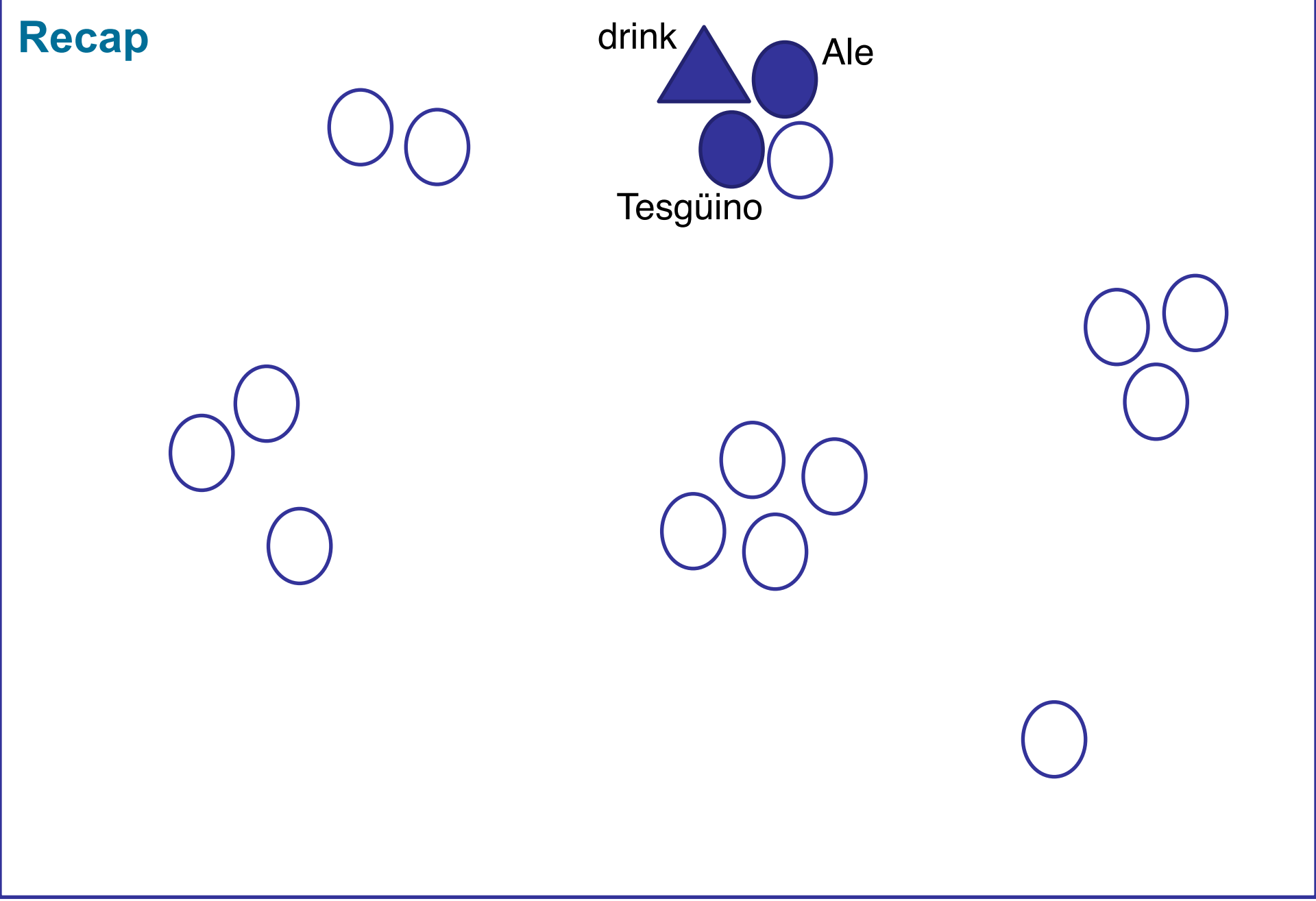
Recap



Embedding vector

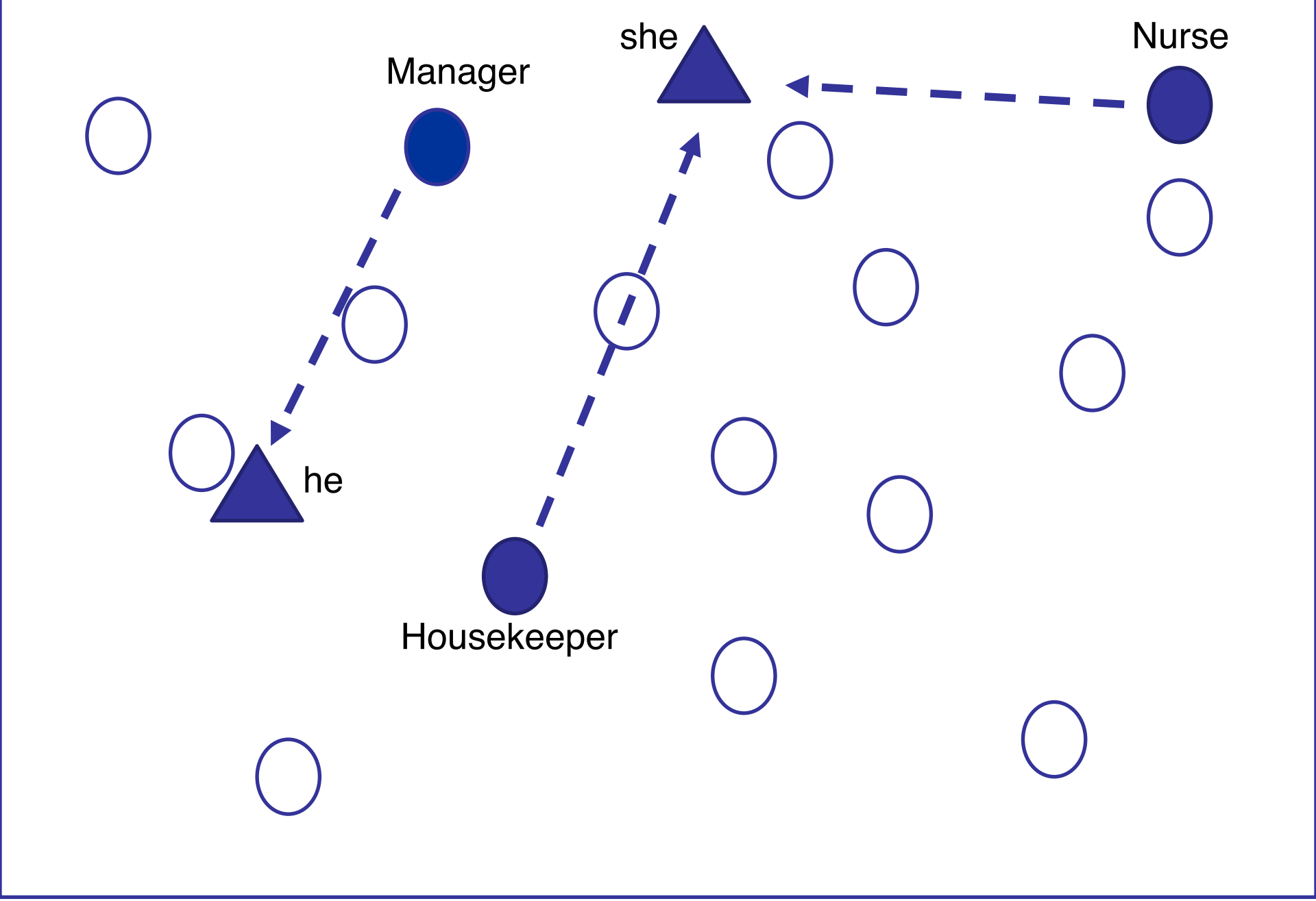
Decoding vector

Recap



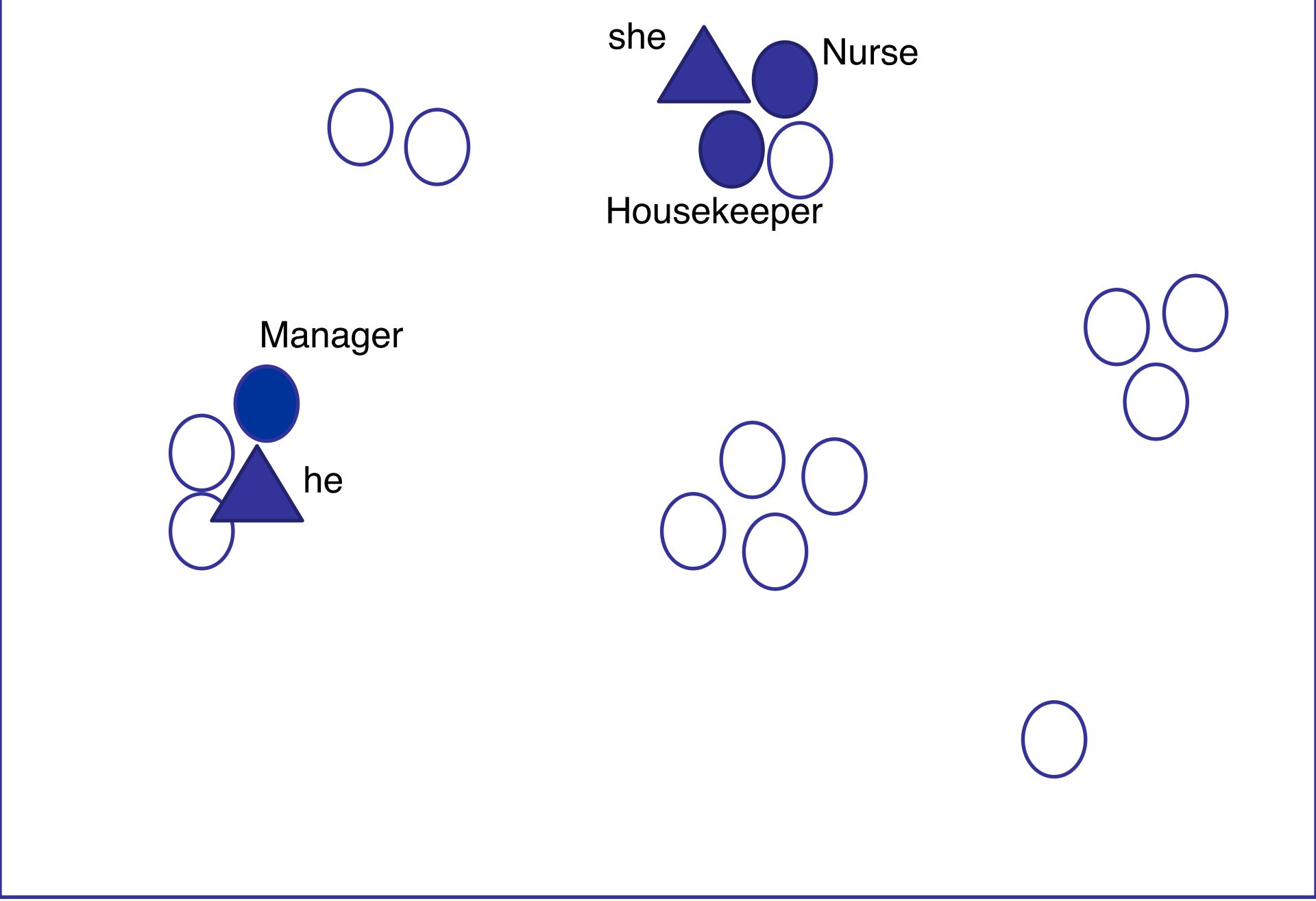
Embedding vector

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*

Bias measurement using word embeddings

Formal definition of bias

- the discrepancy between two concepts (e.g. female and male in gender* bias)
  - Concepts are notated as $\mathbb{Z}$ and $\tilde{\mathbb{Z}}$

- Each concept is defined with a small set of words, e.g.:
  - Female definitional words $\mathbb{Z}$: ***she***, ***her***, ***woman***, ***girl***, etc.
  - Male definitional words $\tilde{\mathbb{Z}}$: ***he***, ***him***, ***man***, ***boy***, etc.

Defining *gender* as a binary construct – namely female vs. male – is an *unpleasant* simplification, as it neglects the wide definition of gender!
Ideally these formulations should cover all gender definitions: LGBT+

Bias measurement – formulation

- A common bias measurement method for word w:

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

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

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

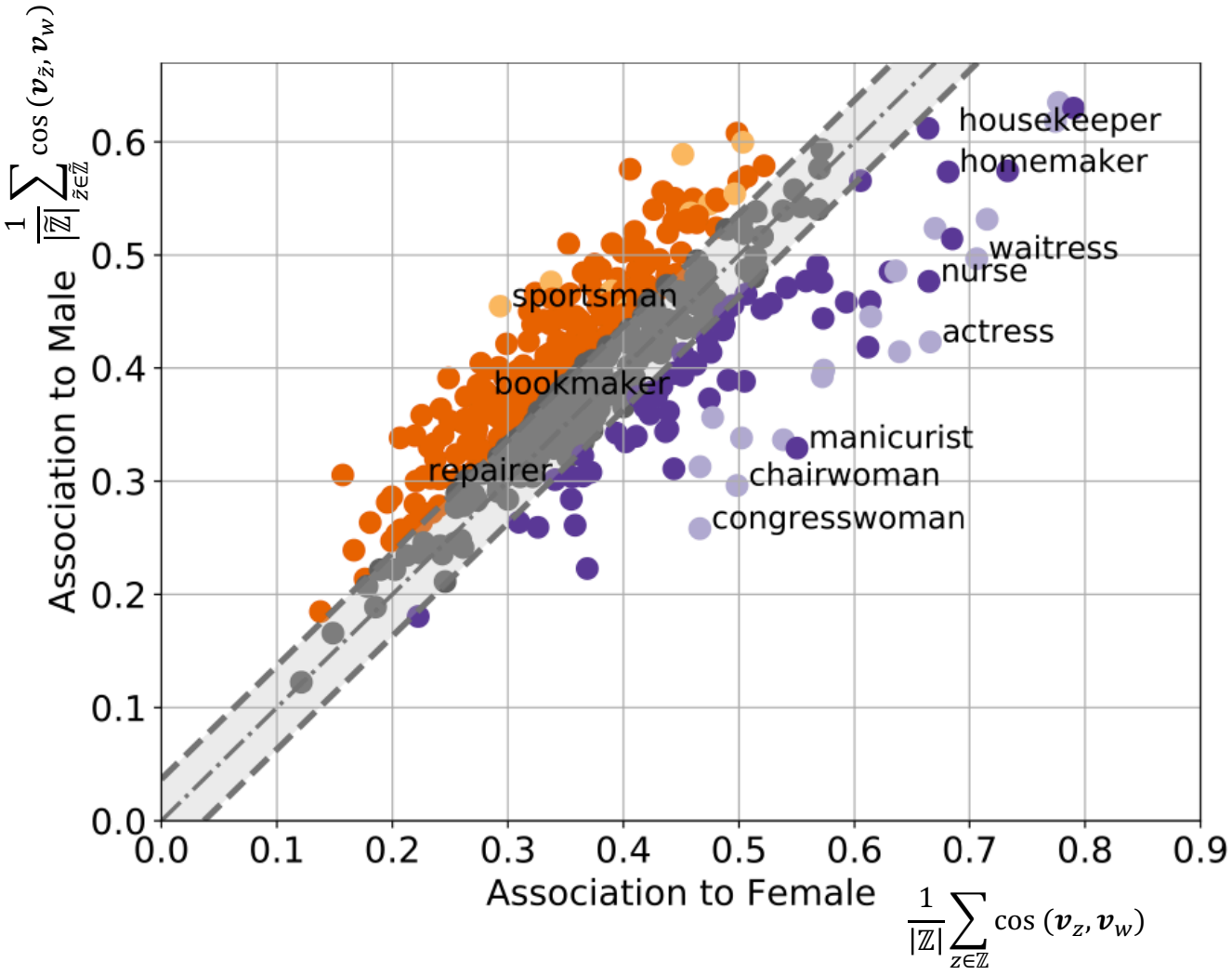
$$\widetilde{\mathbb{Z}} = \{\text{he,him,man,boy}\}$$

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*

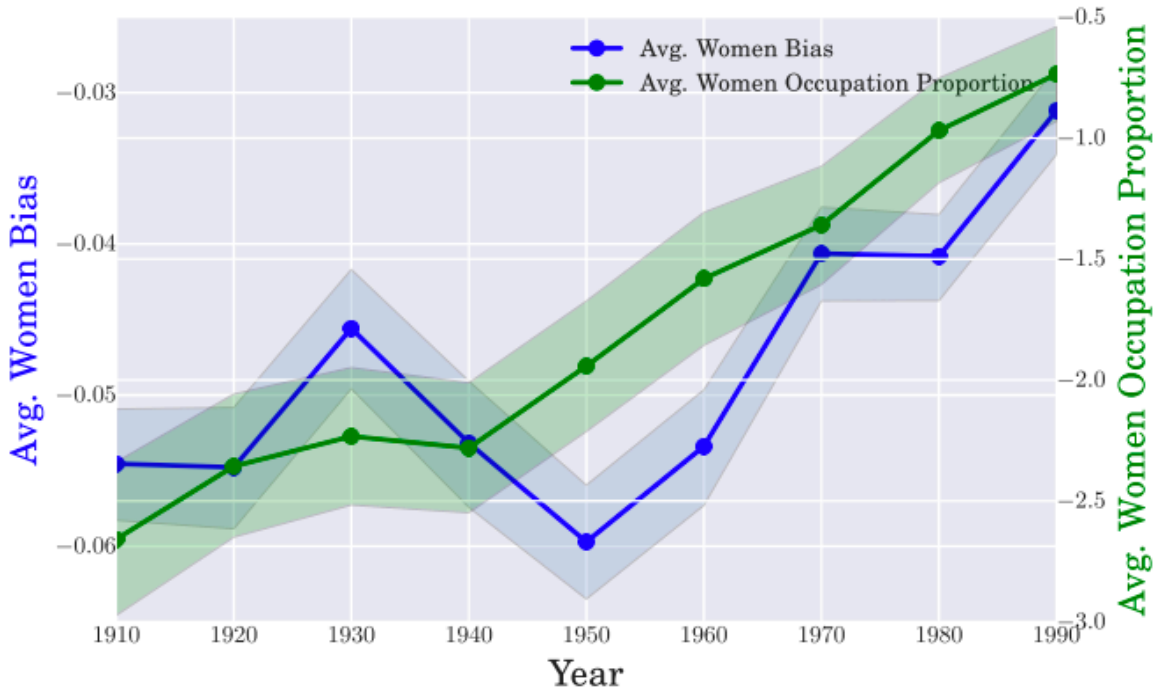
Word Embeddings capture societal realities!



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

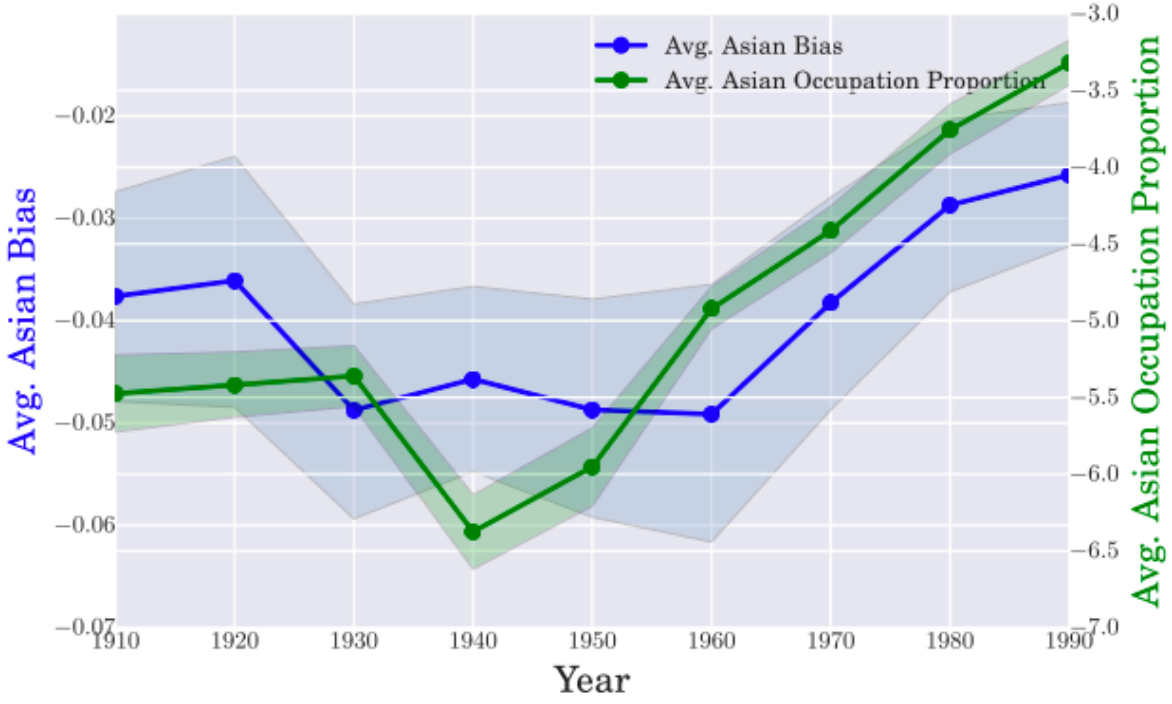
Rekabsaz N., Henderson J., West R., and Hanbury A. "Measuring Societal Biases in Text Corpora via First-Order Co-occurrence." *arXiv preprint arXiv:1812.10424* (2020).

Word Embeddings capture societal realities!



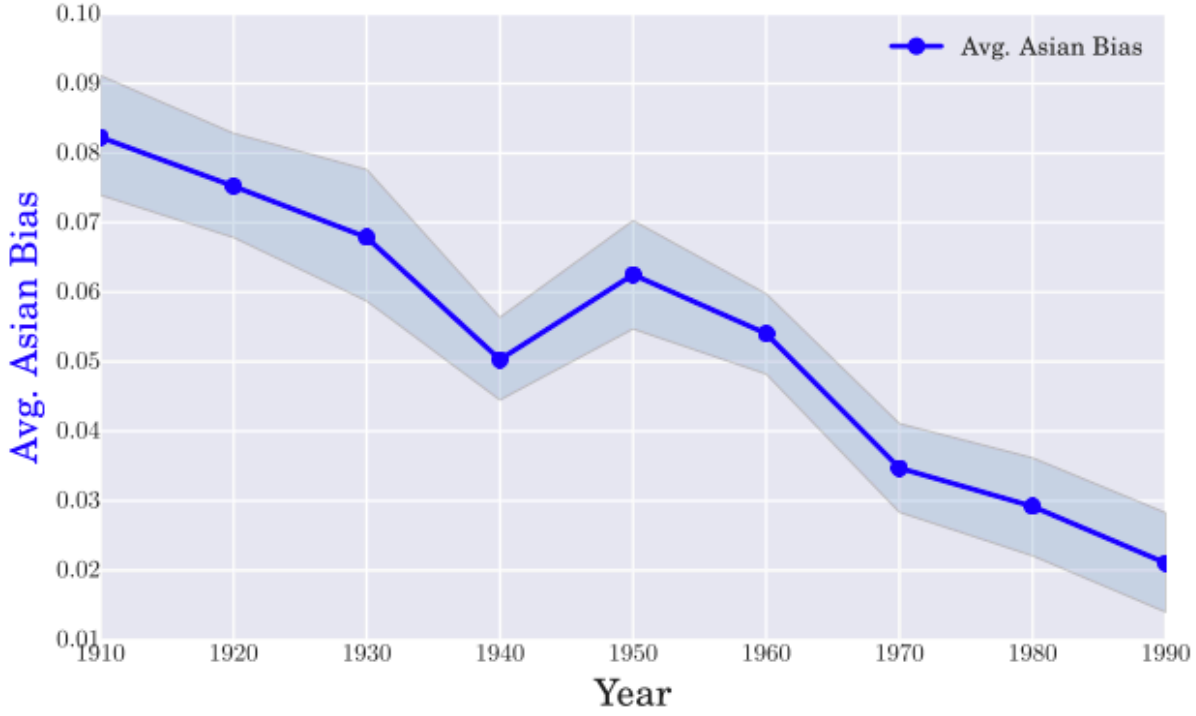
(b) Average gender bias score over time in COHA embeddings in occupations vs the average log proportion. In blue is relative women bias in the embeddings, and in green is the average log proportion of women in the same occupations.

Word Embeddings capture societal realities!



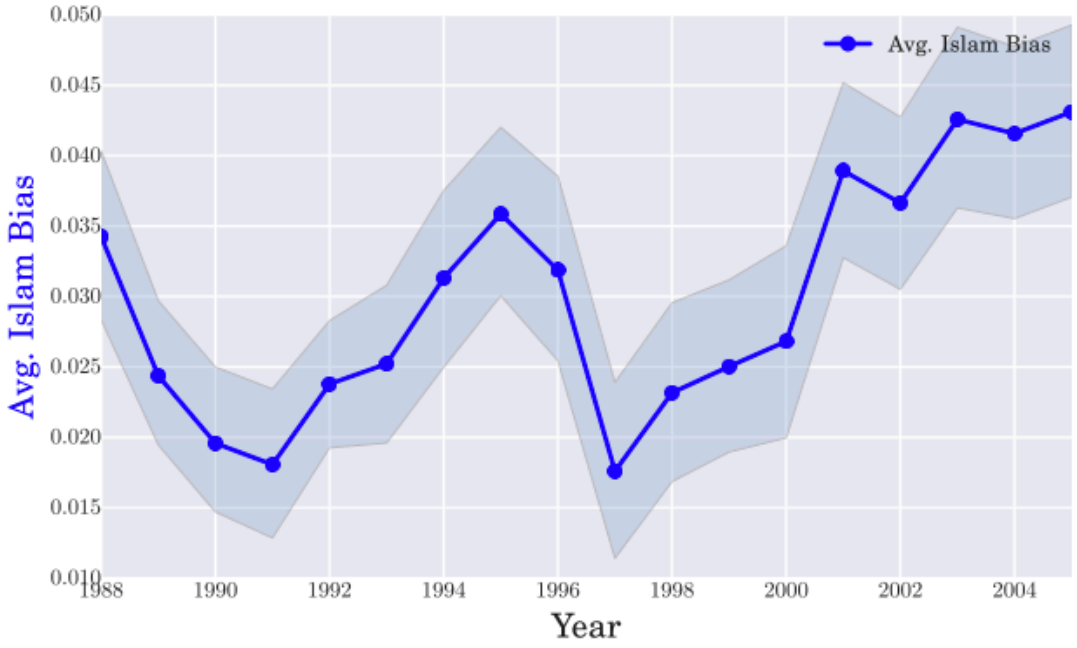
(d) Average ethnic (Asian vs White) bias score over time for occupations in COHA (blue) vs the average conditional log proportion (green).

Word Embeddings capture societal realities!



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

Word Embeddings capture societal realities!



(d) Religious (Islam vs Christianity) bias score over time for words related to terrorism in New York Times data. Note that embeddings are trained in 3 year windows, so, for example, *2000* contains data from 1999-2001.

Bias measurement

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 questions:

- 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

- Motivation
- Bias in word embeddings
- **Bias in IR**

Gender bias measurement in IR – paper walkthrough

- Depend on queries, the contents of the retrieved documents by search engines can be highly biased
  - Search ***nurse***, or ***CEO*** and look at the images!

- An immediate cause of bias is collection
  - If every document in a collection that contains ***nurse*** refers to it as a ***woman***, the retrieved documents of query ***nurse*** will be about women (biased towards ***female***)

- What about (neural) IR models? Do they also affect the bias in retrieval results? What about transfer learning?

- To answer these, we need a framework to measure gender bias in retrieval results

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*

Non-gendered queries annotation

- Step 1: selecting non-gendered queries
  - Non-gendered queries are the ones that contain no indication of gender
  - Gender bias should be studied on the retrieval results of non-gendered queries
  - On the other hand, queries that contain any indication of gender are OK to have results with a more prominent representation of a gender

- Results of human annotation on a set of MS MARCO queries:

Non-gendered (1765)	what is a synonym for beautiful what is the meaning of resurrect
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

Document female/male magnitude

- Step 2: calculate in what extent the content of each document contains female/male topics
  - Simply compute TF of gender definitional words in a document:

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

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

  - $\mathbb{Z}$ Male definitional words: ***he***, ***him***, ***man***, ***boy***, etc.
  - $\tilde{\mathbb{Z}}$ Female definitional words : ***she***, ***her***, ***woman***, ***girl***, etc.
  - $\gamma^{\mathbb{Z}}(D)$ is the degree of existence of concept $\mathbb{Z}$ in document D
    - In simple words: how much the document is about "*male-ness*"

IR bias measurement metric

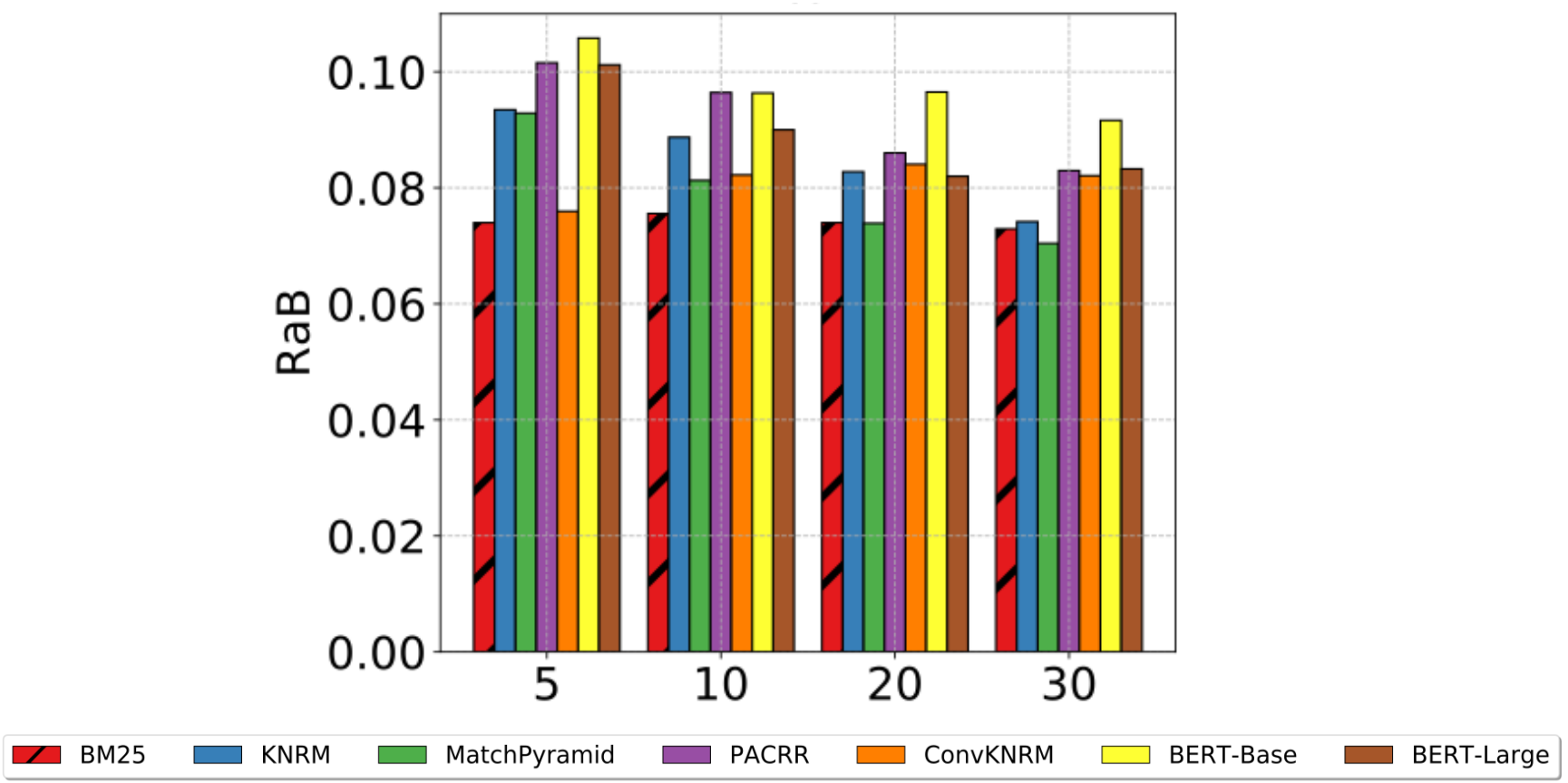
- Step 3: Rank Bias (RaB) metric measures the (gender) bias of the retrieval results over a set of queries

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

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

  - $D_i^{(Q)}$ is the document at position i of the list of documents, retrieved by an IR model when query Q is issued
  - t is rank threshold
  - $\mathbb{Q}$ is the set of non-gendered queries

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

- Arrows show the changes in RaB of neural models, when their word embeddings are initialized randomly instead of initialization with a pre-trained word embedding model (transfer learning)

BM25	0.076
KNRM	0.089 (↓0.020)
MatchPyramid	0.081 (↓0.002)
PACRR	0.096 (↓0.008)
ConvKNRM	0.082 (↓0.001)
BERT-Base	0.096
BERT-Large	0.090

- Randomly-initialized models show smaller degree of gender bias → Transfer learning increases gender bias!

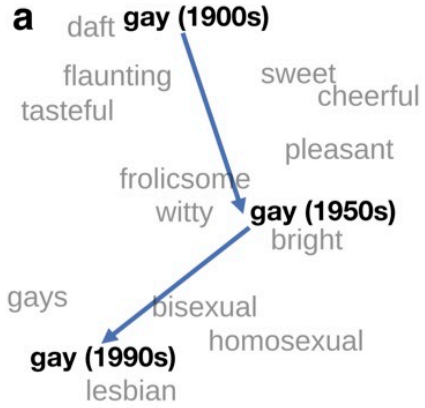
About debiasing

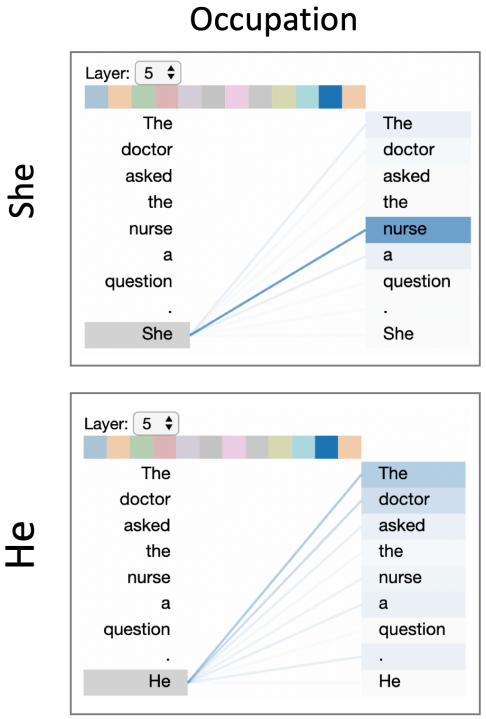
- **Debiasing**: methods to reduce bias
  - The aim is to make the output or decision of a model agnostic to sensitive features (such as gender, race, ethnicity, age)

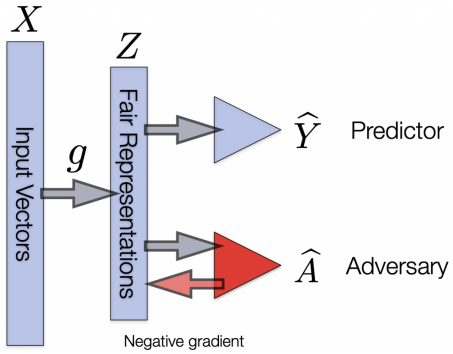
- Approaches in literature are applied to …
  - Dataset: by changing/adding/removing data in collection
  - Model
    - By adding debiasing/fairness criteria to model's objective function
    - By training adversarial networks to remove sensitive information in learned representations
    - By enforcing debiasing criteria through reinforcement learning
  - Output results: by post-processing model's outputs

Some open challenges

- Capturing societal aspects with NLP
- Bias measurement in down-stream tasks, e.g.
  - Search and ranking
  - Content-based job/product/hotel/etc. recommendation
  - Document classification
- Interpretation of neural model regarding bias
- Model debiasing:
  - *Learn not to learn!*
  - Preserving model performance while debiasing







Hamilton, W. L., Leskovec, J., & Jurafsky, D.. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics* (2016)
https://www.arxiv-vanity.com/papers/1904.02679/
https://blog.ml.cmu.edu/2020/02/28/inherent-tradeoffs-in-learning-fair-representations/