Winter semester 2022/23

344.175 VL: Natural Language Processing
Fairness and Societal Biases in NLP

Navid Rekab-saz

Email: navid.rekabsaz@jku.at
Office hours: https://navid-officehours.youcanbook.me

JOHANNES KEPLER UNIVERSITY LINZ

Institute of Computational Perception

Agenda

- Fairness & bias in NLP … what? why?
- Bias in word embeddings
- Bias in downstream tasks

Agenda

- **Fairness & bias in NLP … what? why?**
- Bias in word embeddings
- Bias in downstream tasks

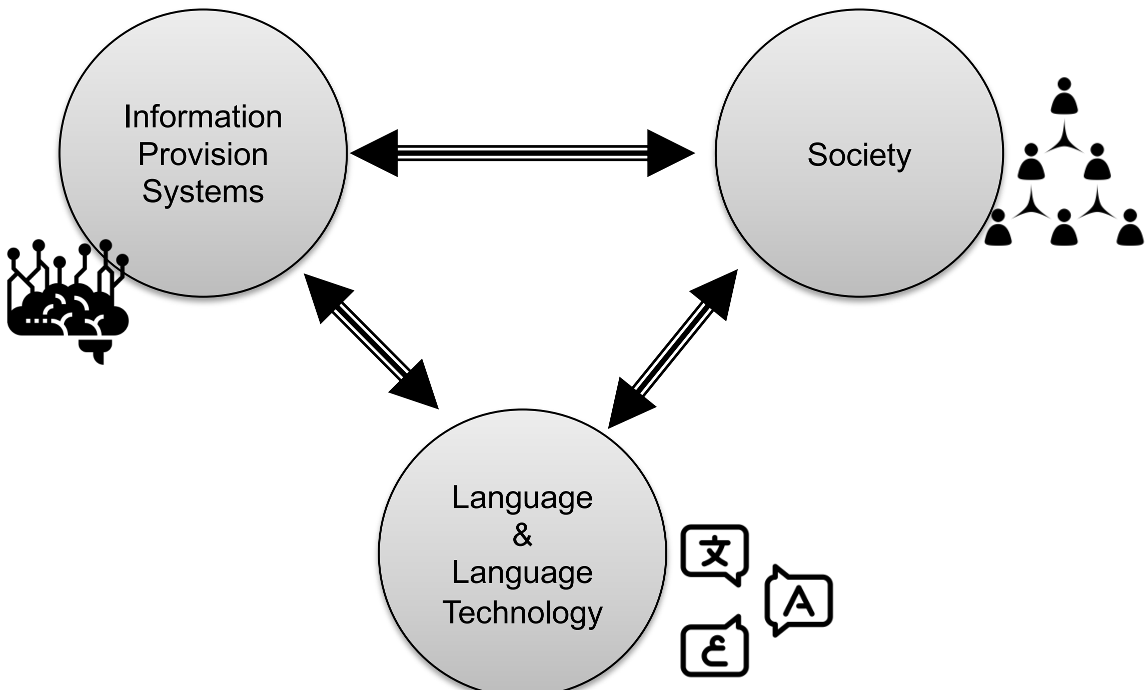
Information, language, and Society

Information access technologies …

- are the gateways to information but also …
- define our perception of the world

Language & language technologies …

- take on and define social meaning
- form and maintain social hierarchies by labeling social groups, and transmitting the beliefs about social groups

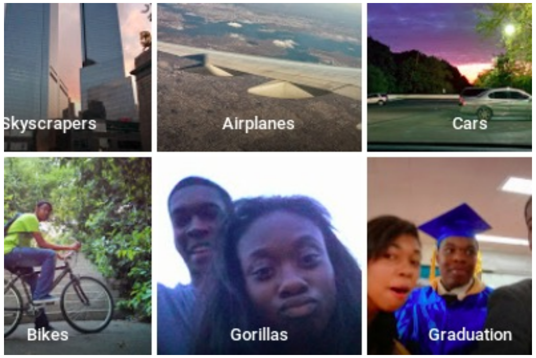


Burguet, Roberto, Ramon Caminal, and Matthew Ellman. "In Google we trust?." *International Journal of Industrial Organization* 39 (2015): 44-55.
Maass, Anne. "Linguistic intergroup bias: Stereotype perpetuation through language." *Advances in experimental social psychology*. 1999.
Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. "Language (Technology) is Power: A Critical Survey of" Bias" in NLP." *In Proc. Of ACL 2020*

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

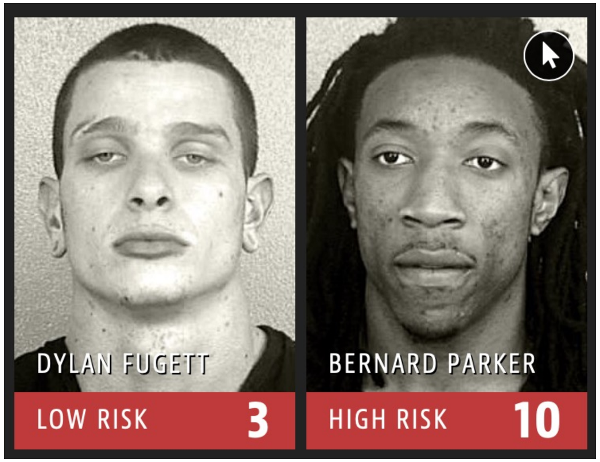
https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology

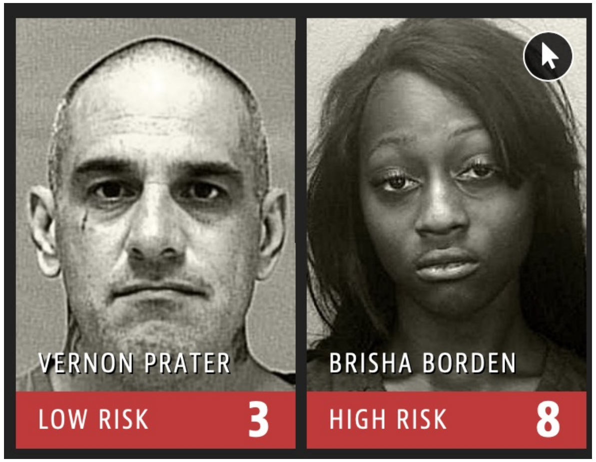
Beauty.AI's 'robot beauty contest' is back – and this time it promises not to be racist

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

Bias in crime discovery

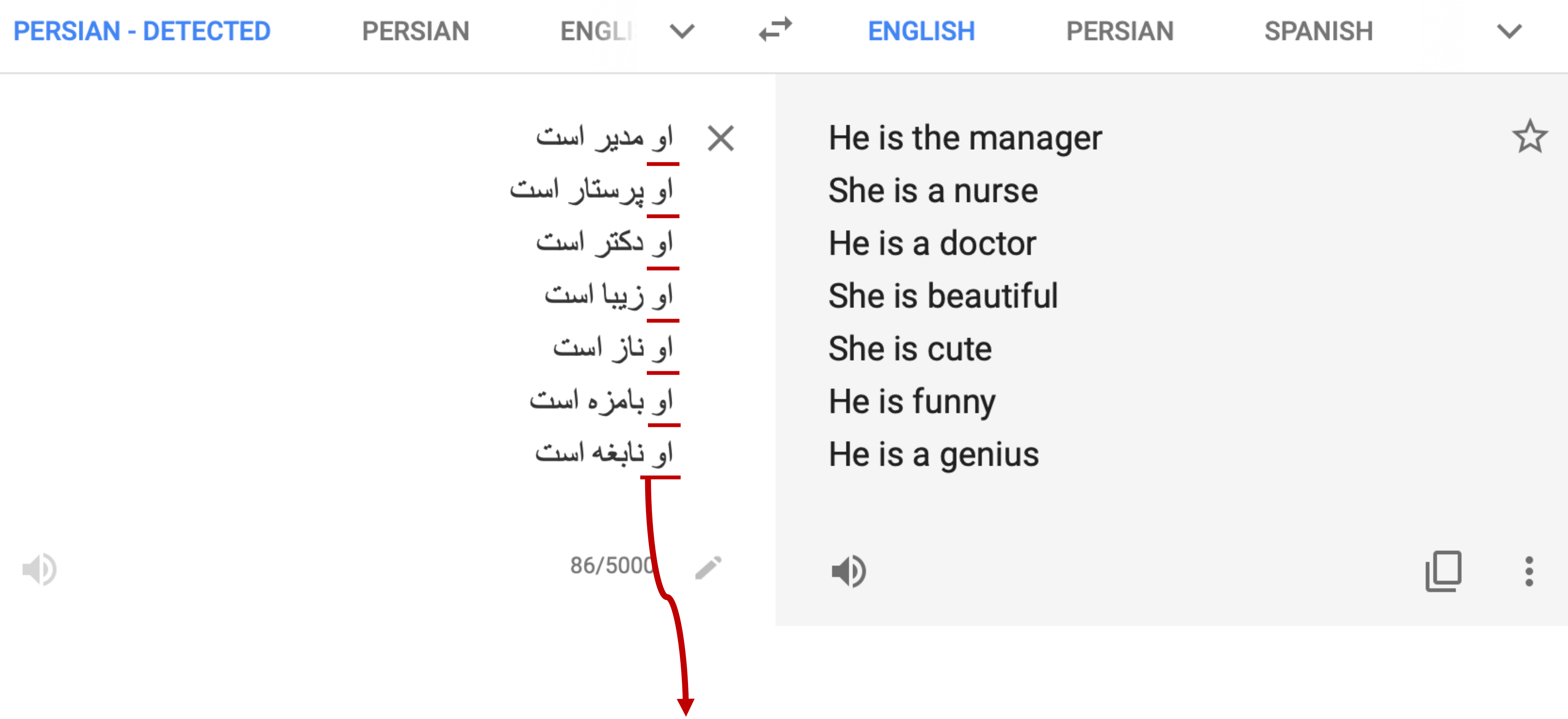
- Predicted risk of reoffending





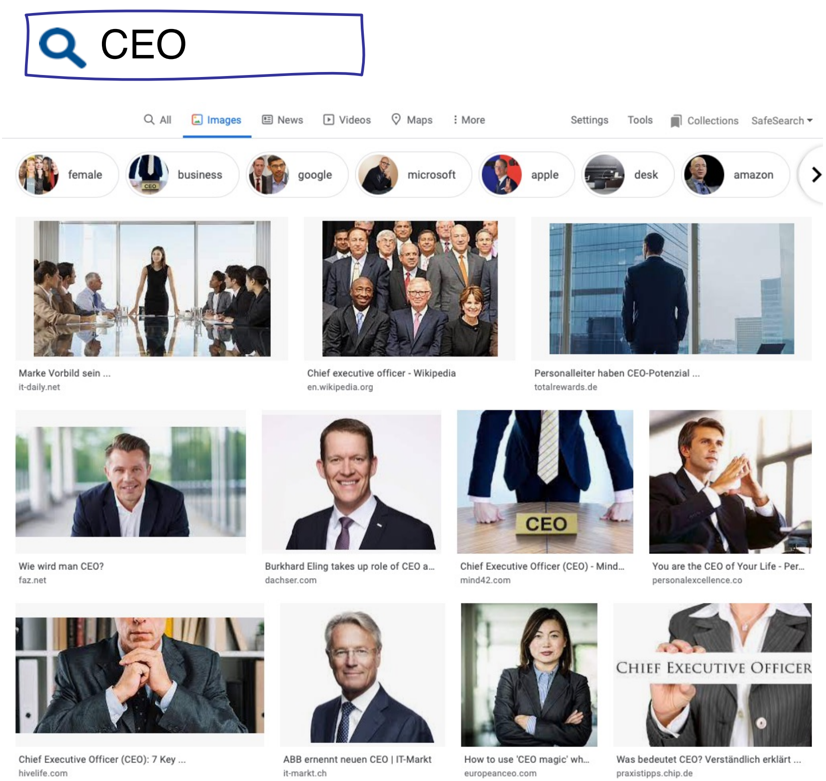
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

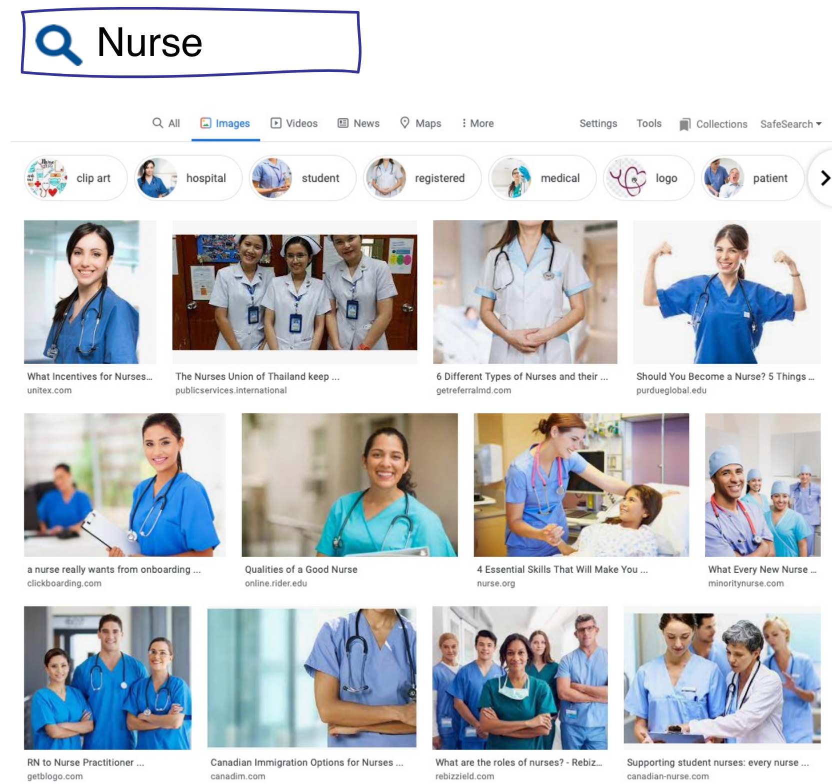
Bias in automatic machine translation



same gender-neutral pronoun

Bias in information retrieval





What we talk about when we talk about *Bias*

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

- Biases and stereotypes *per se* do not imply negative connotations.

From "bias", we mean …

"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

How *harmful*?!

Allocational harms

- A system allocates resources and opportunities unfairly to different social groups
  - E.g., credit and jobs distribution to minorities

Representational harms

- A system represents some social groups in a less favorable light than others.
  - E.g., stereotyping in a search engine or a recommender system that propagates negative generalizations about particular social groups

Fairness

- *What is fair?*
  - Fairness and bias are social concepts and inherently normative
- *Who is affected? What are protected attributes (gender, race, ethnicity, age)?*
  - Bias in NLP systems should be grounded in its social context
- *How is fairness quantified?*
  - Bias/Fairness measurement
- *How to approach the issue?*
  - *Data curation, algorithmic bias mitigation, etc.*

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

Machine learning cycle

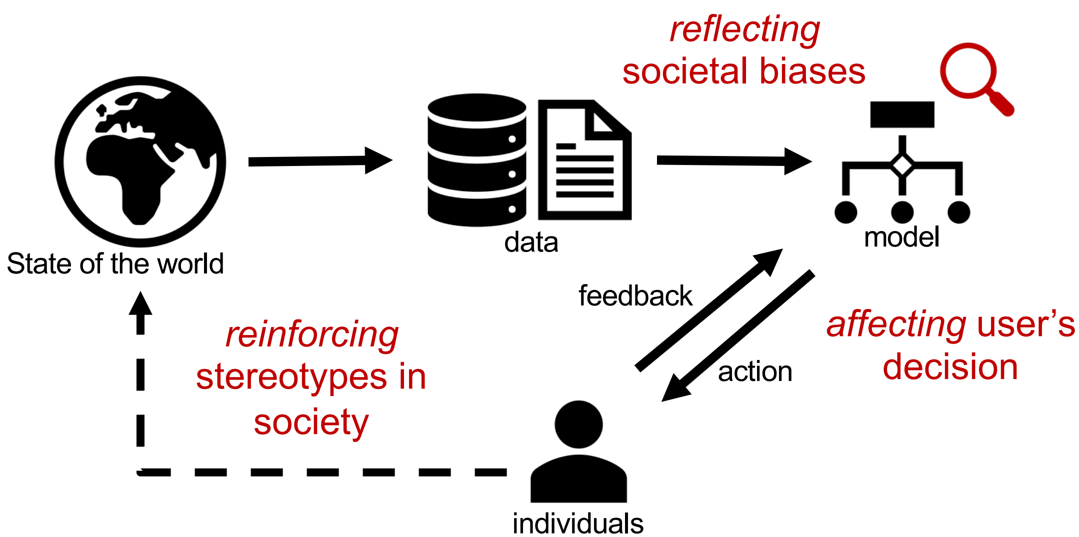
Machine Learning and Societal Biases

ML can observe societal phenomena

- Questions like *“how the perception of girls and boys towards the color pink has changed over time?”*

ML can reinforce societal biases

- Encoded societal biases and stereotypes can affect decision making of users and eventually reinforce biases in society



Where are biases 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

A representative task – occupation prediction from biographies:

[She/He?] graduated from Lehigh University, with honours in 1998. [Nancy/Adam?] has years of experience in weight loss surgery, patient support, education, and diabetes.

→ Nurse

Language is inherently intertwined with
semantics and *implicit meanings*

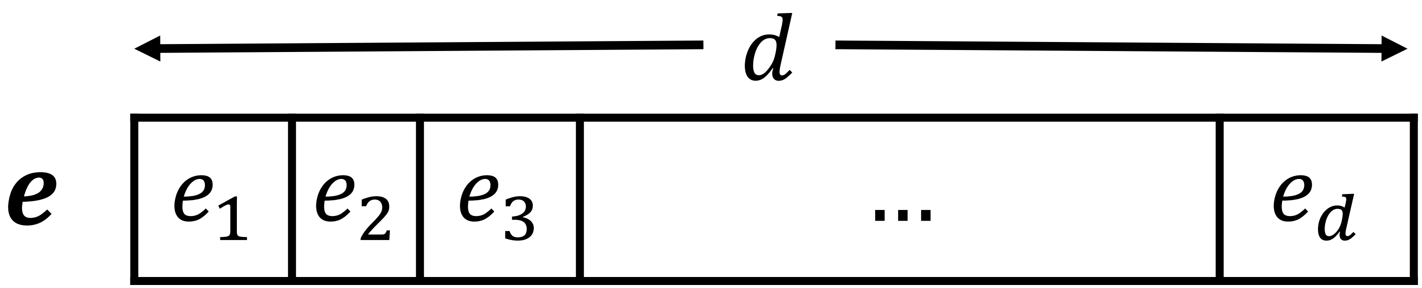
Bias in bios: A case study of semantic representation bias in a high-stakes setting. *D. Maria, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. Tauman Kalai.* Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Agenda

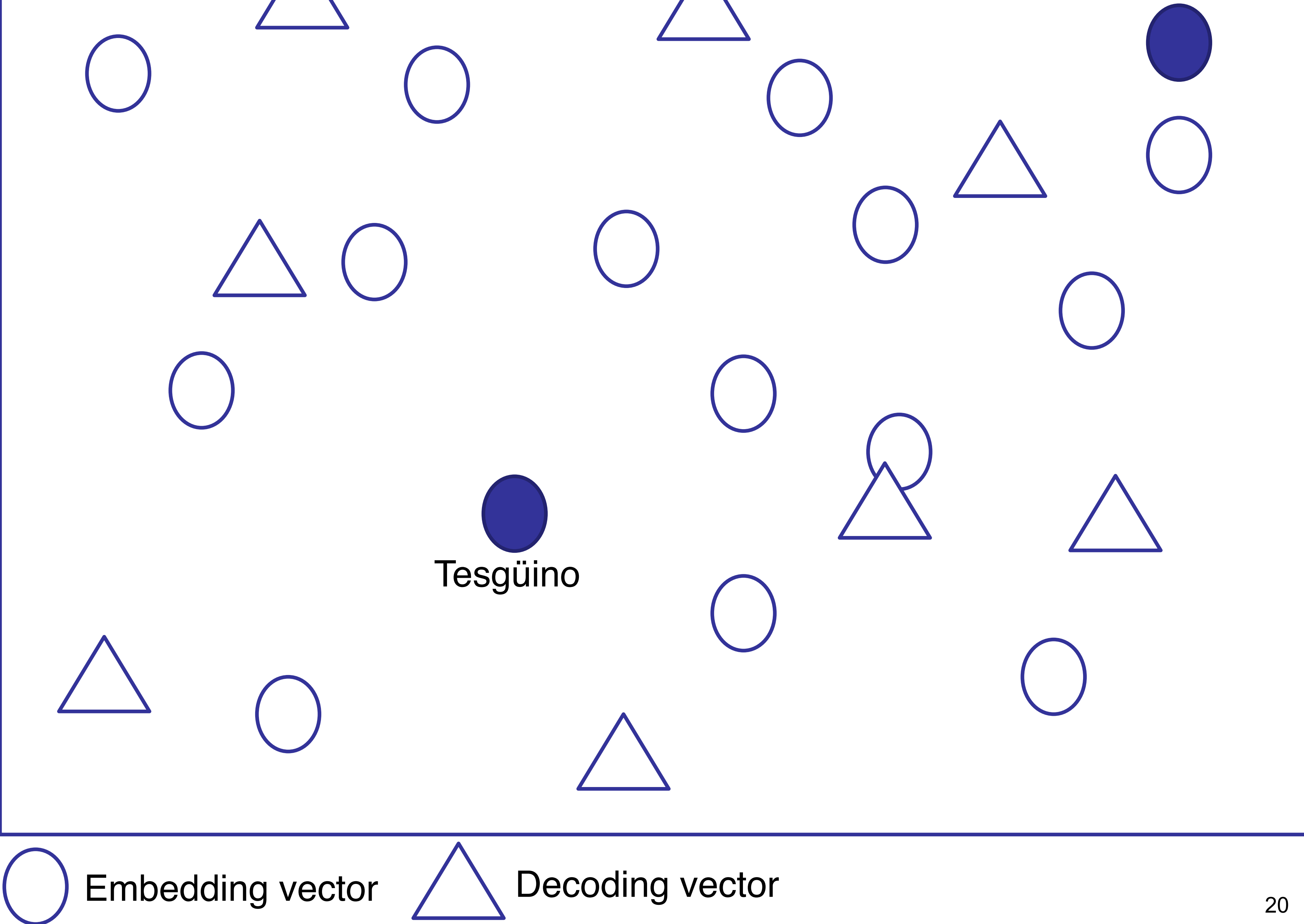
- Fairness & bias in NLP … what? why?
- **Bias in word embeddings**
- Bias in downstream tasks

Representation learning and bias

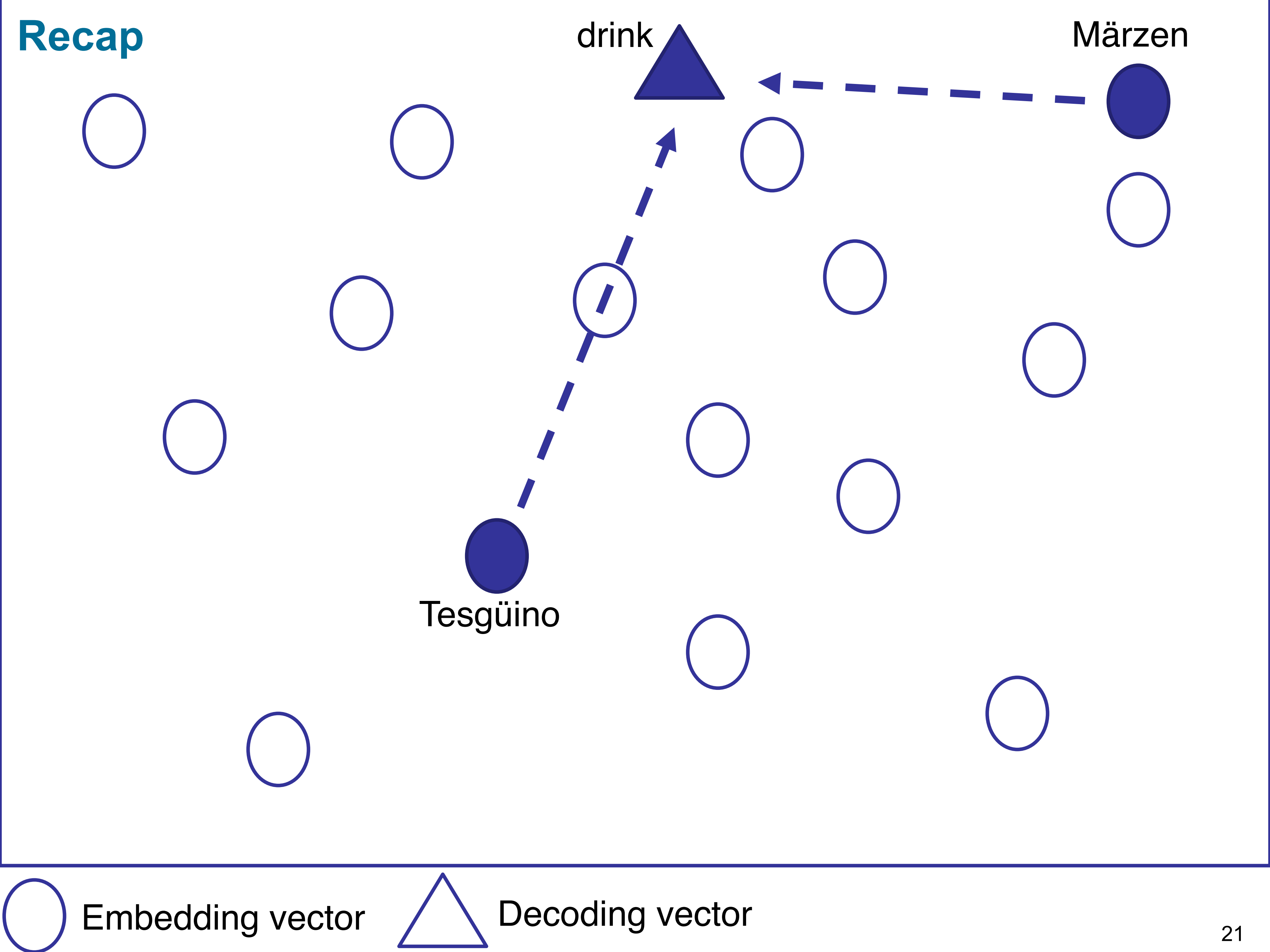
Representation learning encodes information but also may encode the underlying biases in data!



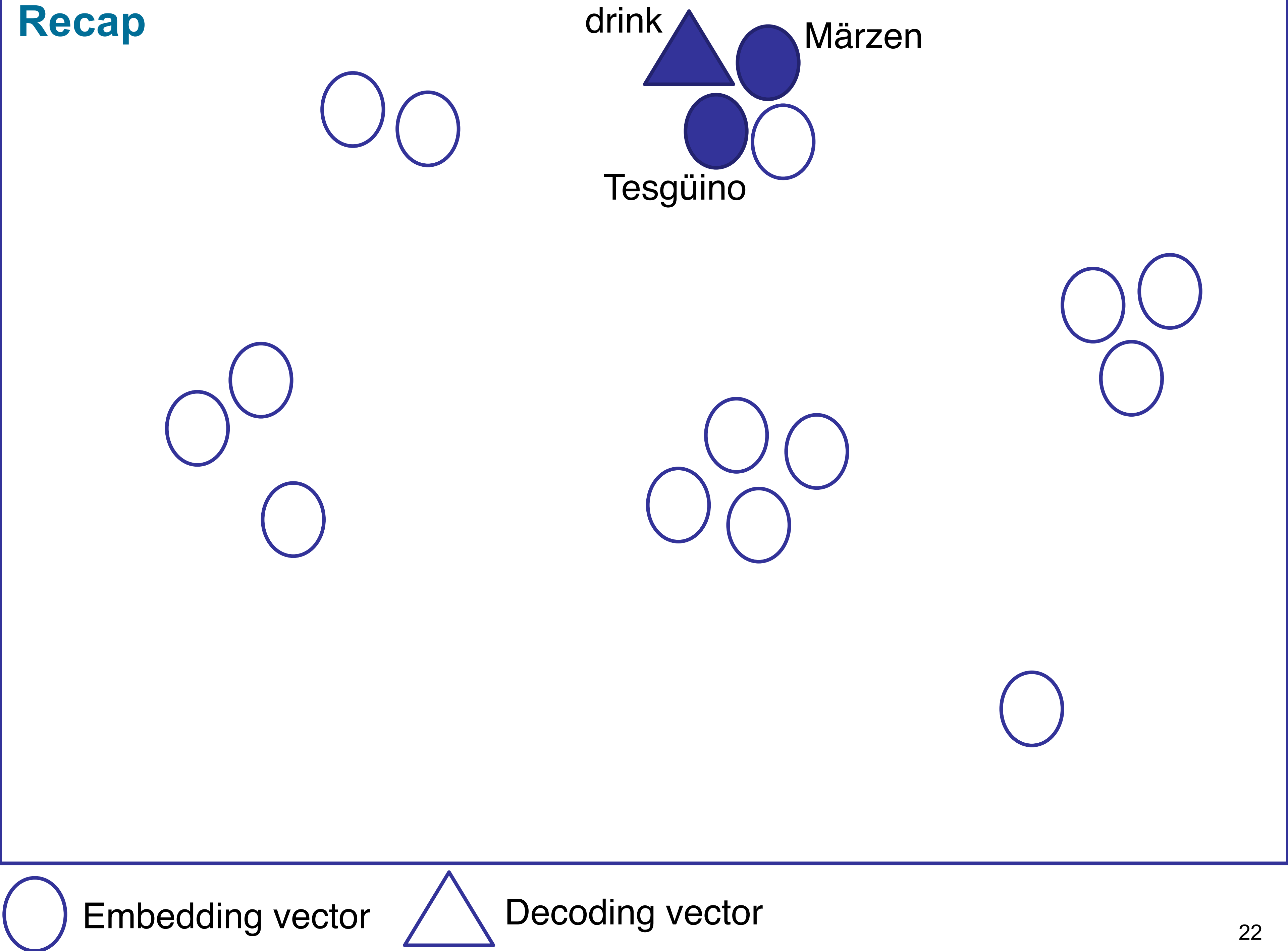
Recap

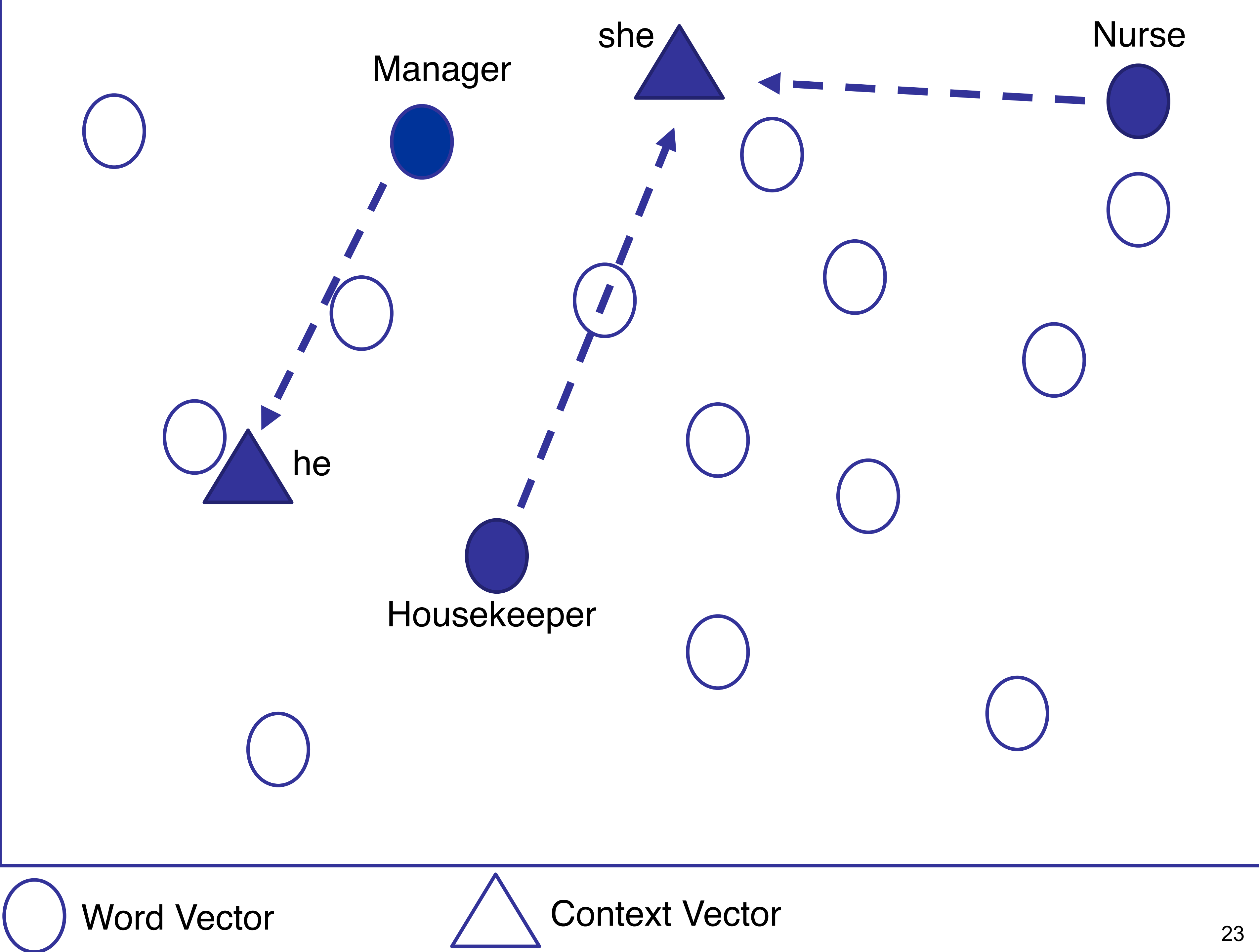


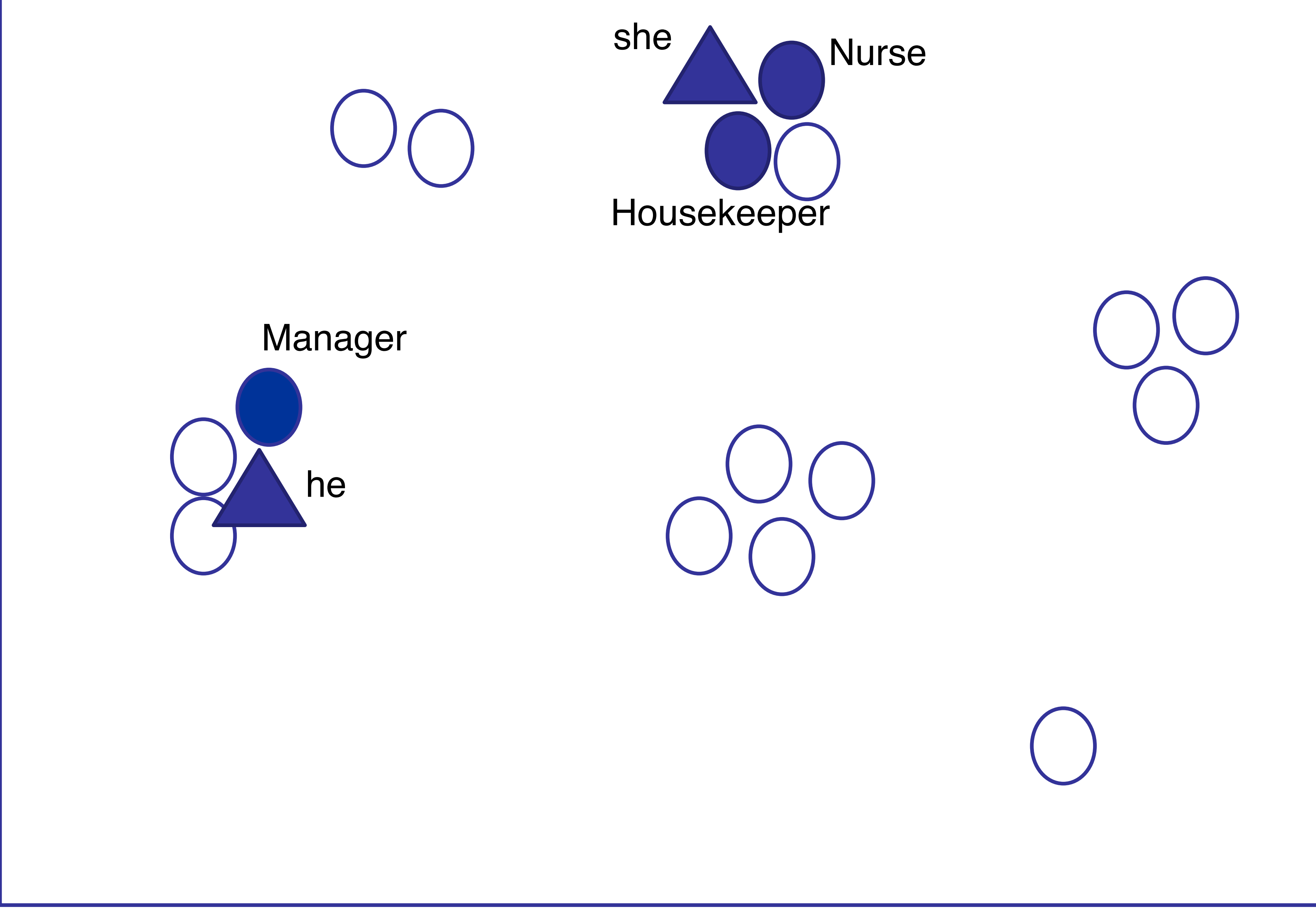
Recap



Recap




she
Nurse
Manager
he
Housekeeper
Word Vector
Context Vector


she
Nurse
Housekeeper
Manager
he


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

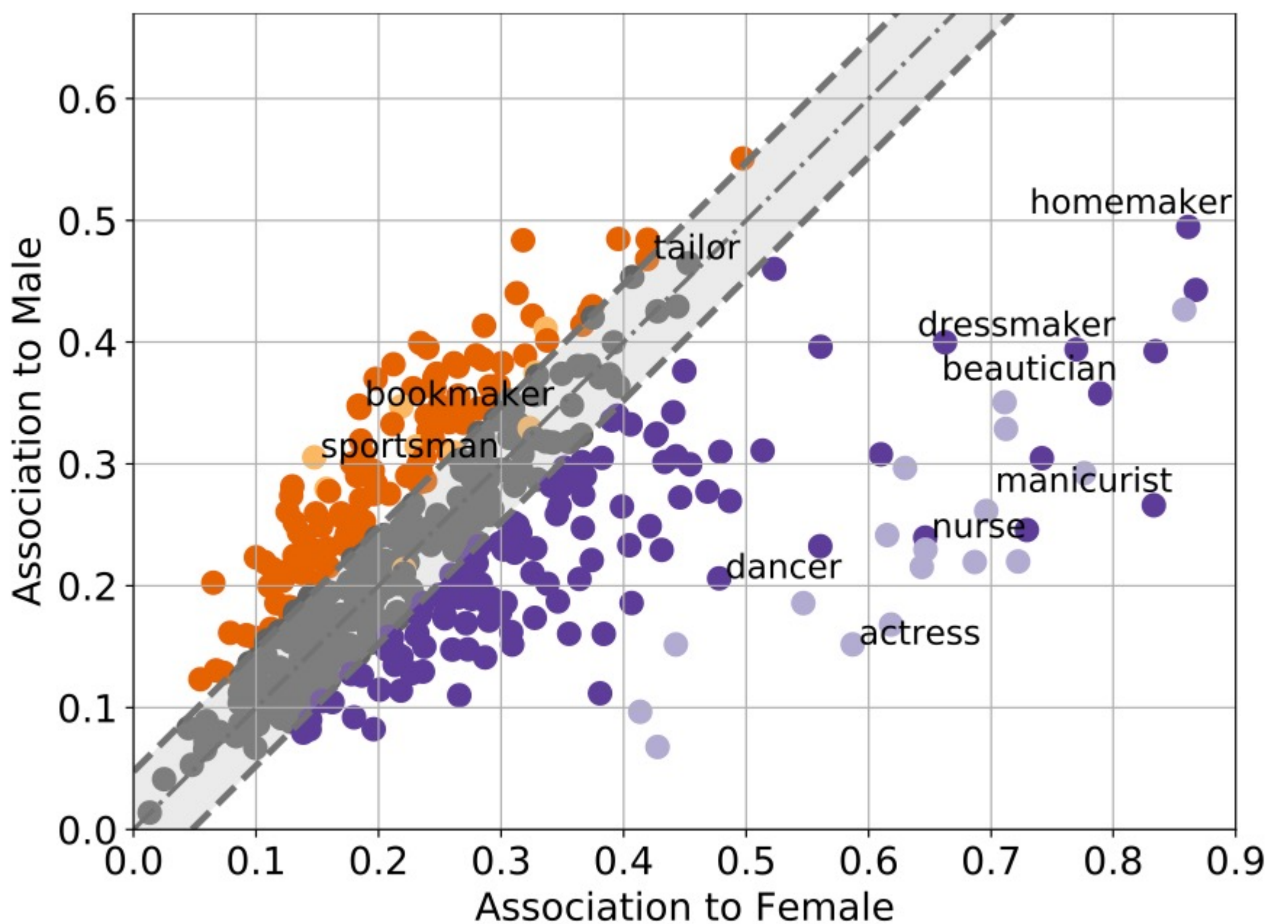
sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas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

Gender bias of words in a word embedding model



A word2vec model trained on Wikipedia

Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *N. Rekabsaz, R. West, J. Henderson, A. Hanbury In proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021*

Measuring bias of a word using word embeddings

High-order method

- Bias: discrepancy of the relations of a word w (like nurse) towards two concepts $\mathbb{V}$ and $\widetilde{\mathbb{V}}$ (like female and male)

- $\mathbb{V}$ and $\widetilde{\mathbb{V}}$ are commonly defined by sets of representative words. For example, in a *binary* setting of gender bias:

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

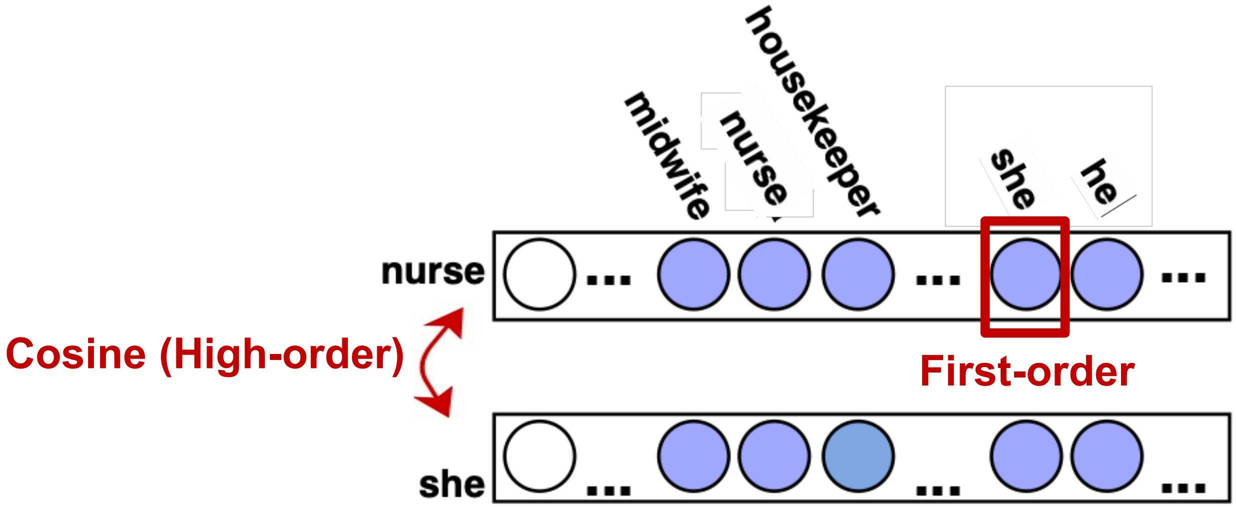
$$\widetilde{\mathbb{V}} = \{\textit{\textbf{he}}, \textit{\textbf{him}}, \textit{\textbf{man}}, \textit{\textbf{boy}}, \dots\}$$

- High-order bias measurement:

$$\mathrm{BIAS}_{\mathrm{High}}(w) = \frac{1}{|\mathbb{V}|} \sum_{v \in \mathbb{V}} \cos\left(\boldsymbol{e}_v, \boldsymbol{e}_w\right) - \frac{1}{|\widetilde{\mathbb{V}}|} \sum_{\tilde{v} \in \widetilde{\mathbb{V}}} \cos\left(\boldsymbol{e}_{\tilde{v}}, \boldsymbol{e}_w\right)$$

First-order Bias Measurement

$$\mathrm{BIAS}_{\mathrm{HIGH}}(w) = \frac{1}{|\mathbb{V}|}\sum_{v\in\mathbb{V}} \cos\left(\boldsymbol{e}_v, \boldsymbol{e}_w\right) - \frac{1}{|\tilde{\mathbb{V}}|}\sum_{\tilde{v}\in\tilde{\mathbb{V}}} \cos\left(\boldsymbol{e}_{\tilde{v}}, \boldsymbol{e}_w\right)$$



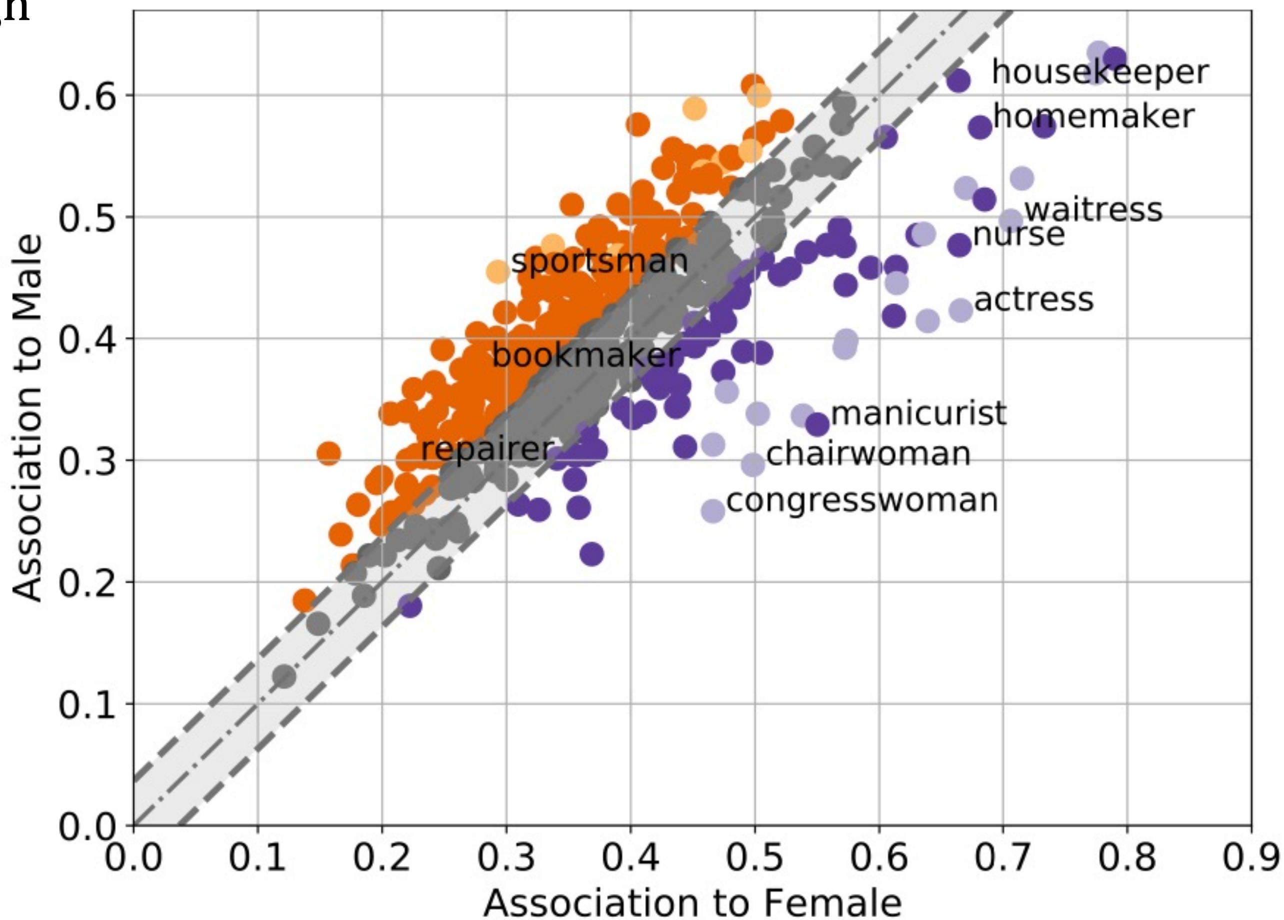
- First-order bias measurement for word w
  - Calculate $P(y = 1|w, v)$ using the encoder embedding E and the decoder embedding U of a word2vec model (see lecture 7)

$$\mathrm{BIAS}_{\mathrm{First}}(w) = \frac{1}{|\mathbb{V}|}\sum_{v\in\mathbb{V}} \boldsymbol{e}_v\boldsymbol{u}_w - \frac{1}{|\tilde{\mathbb{V}}|}\sum_{\tilde{v}\in\tilde{\mathbb{V}}} \boldsymbol{e}_{\tilde{v}}\boldsymbol{u}_w$$

Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *N. Rekabsaz, R. West, J. Henderson, A. Hanbury In proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021*

Measuring bias in WE with high-order method

$\mathrm{BIAS}_{\mathrm{High}}$

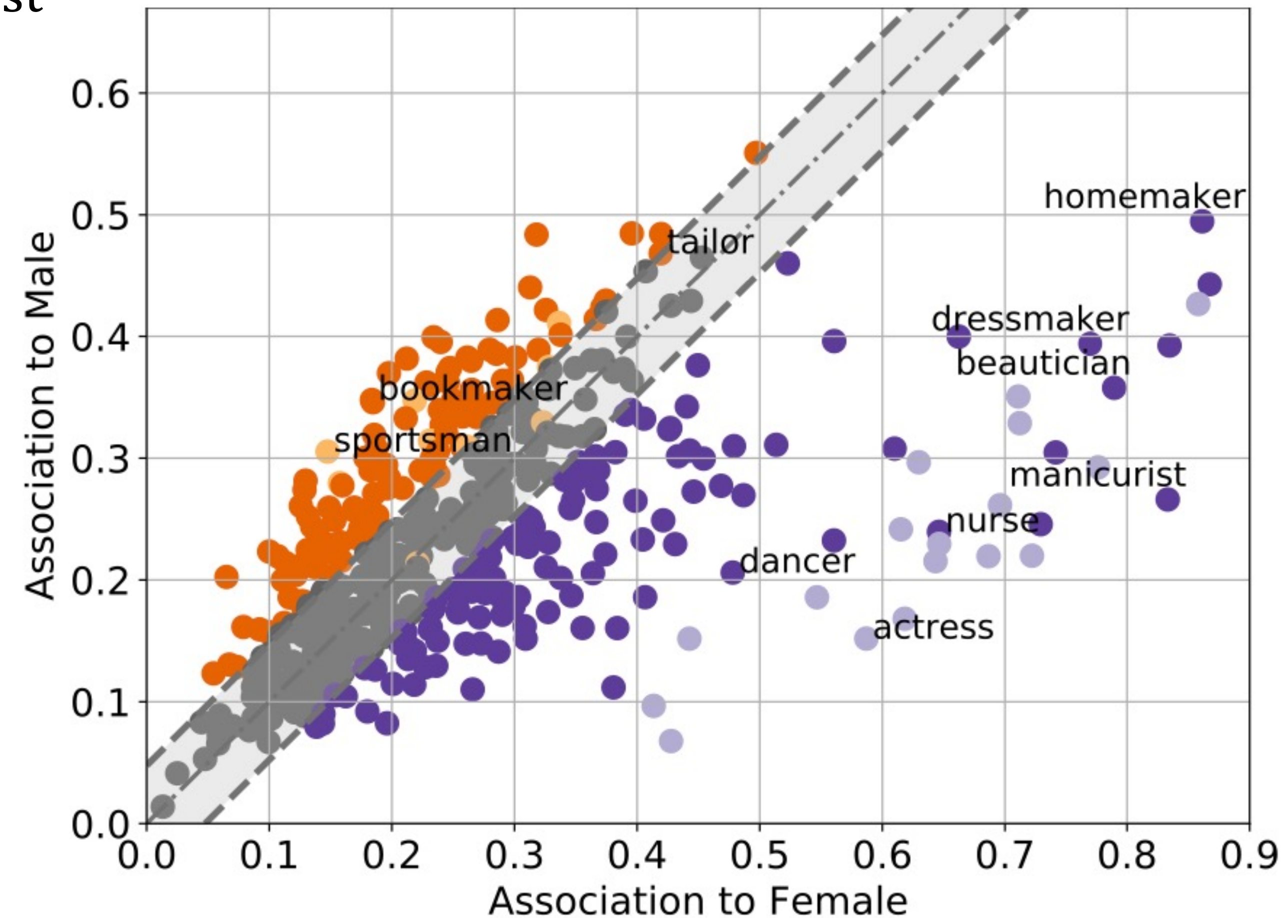


A word2vec model trained on a recent Wikipedia corpus

Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *N. Rekabsaz, R. West, J. Henderson, A. Hanbury In proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021*

Measuring bias in WE with <u>first-order</u> method

$\mathrm{BIAS}_{\mathrm{First}}$



A word2vec model trained on a recent Wikipedia corpus

Measuring Societal Biases from Text Corpora with Smoothed First-Order Co-occurrence. *N. Rekabsaz, R. West, J. Henderson, A. Hanbury In proceedings of the International AAAI Conference on Web and Social Media (ICWSM) 2021*

Correlations with job market statistics

Order	Representation	Method	Labor Data Spearman ρ	Labor Data Pearson's r	Census Data Spearman ρ	Census Data Pearson's r
High-Order	PMI	Directional	0.28	0.07	0.18	0.02
		Centroid	0.14	0.21	0.35	0.40
		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
		Average$_{\text{High}}$	0.41	0.49	0.49	0.56
First-Order	PMI	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
		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
		Average$_{\text{High}}$	0.41	0.49	0.49	0.56
First-Order	PPMI	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**
		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
		Average$_{\text{High}}$	0.26	0.38	0.36	0.46
First-Order	SPPMI	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
		Average$_{\text{High}}$	**0.60**	**0.60**	0.39	0.51
First-Order	initGlove	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
		Average$_{\text{High}}$	0.55	0.57	0.59	0.65
First-Order	eSG	Average$_{\text{First}}$	**0.66**	**0.61**	**0.67**	**0.70**

Correlation results of the gender bias values (calculated with word embeddings) to the statistics of the portion of women in occupations

Summary

- 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

- Similar approach is used to measure bias in Large Language Models
  - Read more: May, C., Wang, A., Bordia, S., Bowman, S., & Rudinger, R. (2019, June). On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 622-628). https://aclanthology.org/N19-1063/

Agenda

- Fairness & bias in NLP … what? why?
- Bias in word embeddings
- **Bias in downstream tasks**

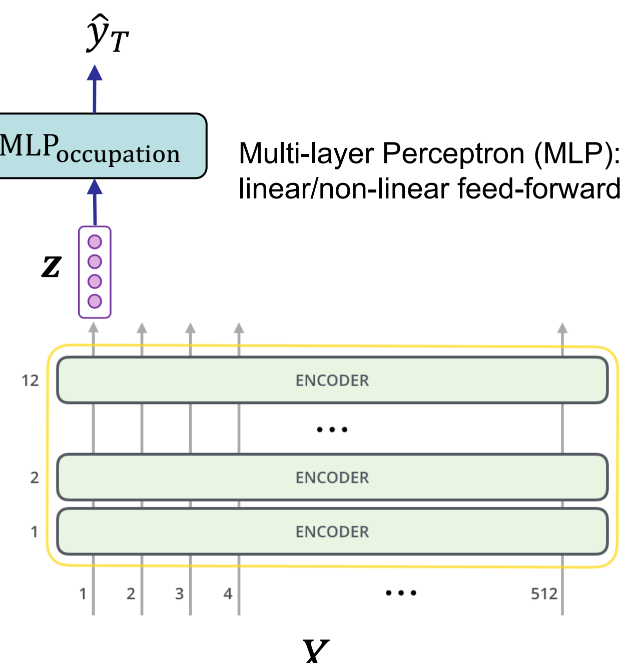
Biography classification

- Predicting the occupation of a person from the biography of a person
  - Gender is protected/sensitive attribute

X original	Anna graduated from Lehigh University, with honors in 1998. She has years of experience in weight loss surgery, patient support, education, and diabetes.
	or
$\ddot{X}$ masked gender	[MASK] graduated from Lehigh University, with honors in 1998. [MASK] has years of experience in weight loss surgery, patient support, education, and diabetes.

y_T = Nurse

y_G = Female



Multi-layer Perceptron (MLP): a linear/non-linear feed-forward network

Bias in bios: A case study of semantic representation bias in a high-stakes setting. *D. Maria, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. Tauman Kalai.* Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

Possible bias/(un)fairness measurements

- (Un)Fairness in the quality of service provided by a model (model performance) when comparing across subpopulations
  - Like the difference of a model's performance for the male and female users

- Bias as the degree of information leakage regarding a protected attribute
  - How much can we retrieve a protected attribute for instance through an adversarial attack
  - Fairness as "blindness" towards the protected attribute

- Differences of model's decisions between an original data point and a its counter-factual variation

Fairness in quality of service

y_G	Input	y_T	Prediction of a model $\hat{y}_T$
Male	X_1	Surgeon	Surgeon
Male	X_2	Surgeon	Surgeon
Male	X_3	Surgeon	Surgeon
Male	X_4	Surgeon	Surgeon
Male	X_5	Surgeon	Nurse
Male	X_6	Surgeon	Surgeon
Male	X_7	Surgeon	Surgeon
Male	X_8	Surgeon	Surgeon
Female	X_9	Surgeon	Nurse
Female	X_{10}	Surgeon	Surgeon
Female	X_{11}	Surgeon	Surgeon
Female	X_{12}	Surgeon	Surgeon
Male	X_{13}	Nurse	Surgeon
Male	X_{14}	Nurse	Nurse
Female	X_{15}	Nurse	Nurse
Female	X_{16}	Nurse	Nurse
Female	X_{17}	Nurse	Nurse
Female	X_{18}	Nurse	Nurse
Female	X_{19}	Nurse	Surgeon
Female	X_{20}	Nurse	Nurse

- Evaluation metric: True Positive Rate (TPR)
- TPR per occupation:

$$\mathrm{TPR}_{\mathrm{occ}} = \frac{\#\text{ of correct Occupation}}{\#\text{ of Occupation}}$$

$$\mathrm{TPR}_{\mathrm{Surgeon}} = \frac{10}{12} = \frac{5}{6}$$

$$\mathrm{TPR}_{\mathrm{Nurse}} = \frac{6}{8} = \frac{3}{4}$$

- TPR per occupation and gender:

$$\mathrm{TPR}_{\mathrm{occ,gender}} = \frac{\#\text{ of correct for Occupation and Gender}}{\#\text{ of Occupation and Gender}}$$

$$\mathrm{TPR}_{\mathrm{Surgeon,Male}} = \frac{7}{8}$$

$$\mathrm{TPR}_{\mathrm{Surgeon,Female}} = \frac{3}{4}$$

$$\mathrm{TPR}_{\mathrm{Nurse,Male}} = \frac{1}{2}$$

$$\mathrm{TPR}_{\mathrm{Nurse,Female}} = \frac{5}{6}$$

Fairness in quality of service

y_G	Input	y_T	Prediction of a model $\hat{y}_T$
Male	X_1	Surgeon	Surgeon
Male	X_2	Surgeon	Surgeon
Male	X_3	Surgeon	Surgeon
Male	X_4	Surgeon	Surgeon
Male	X_5	Surgeon	Nurse
Male	X_6	Surgeon	Surgeon
Male	X_7	Surgeon	Surgeon
Male	X_8	Surgeon	Surgeon
Female	X_9	Surgeon	Nurse
Female	X_{10}	Surgeon	Surgeon
Female	X_{11}	Surgeon	Surgeon
Female	X_{12}	Surgeon	Surgeon
Male	X_{13}	Nurse	Surgeon
Male	X_{14}	Nurse	Nurse
Female	X_{15}	Nurse	Nurse
Female	X_{16}	Nurse	Nurse
Female	X_{17}	Nurse	Nurse
Female	X_{18}	Nurse	Nurse
Female	X_{19}	Nurse	Surgeon
Female	X_{20}	Nurse	Nurse

<u>One possible definition of fairness:</u>

- A system is fair regarding the protected attribute, if the model provides an equal quality of service to the underlying social groups

<u>One metric of *un*fairness:</u>

$$\mathrm{Unfairness}_{\mathrm{occ}} = \mathrm{TPR}_{\mathrm{occ,Male}} - \mathrm{TPR}_{\mathrm{occ,Female}}$$

<u>Example:</u>

$$\mathrm{TPR}_{\mathrm{Surgeon,Male}} = \frac{7}{8} \qquad \mathrm{TPR}_{\mathrm{Surgeon,Female}} = \frac{3}{4}$$

$$\mathrm{TPR}_{\mathrm{Nurse,Male}} = \frac{1}{2} \qquad \mathrm{TPR}_{\mathrm{Nurse,Female}} = \frac{5}{6}$$

$$\mathrm{Unfairness}_{\mathrm{Surgeon}} = \frac{7}{8} - \frac{3}{4} = \frac{1}{8}$$

Unfair towards female

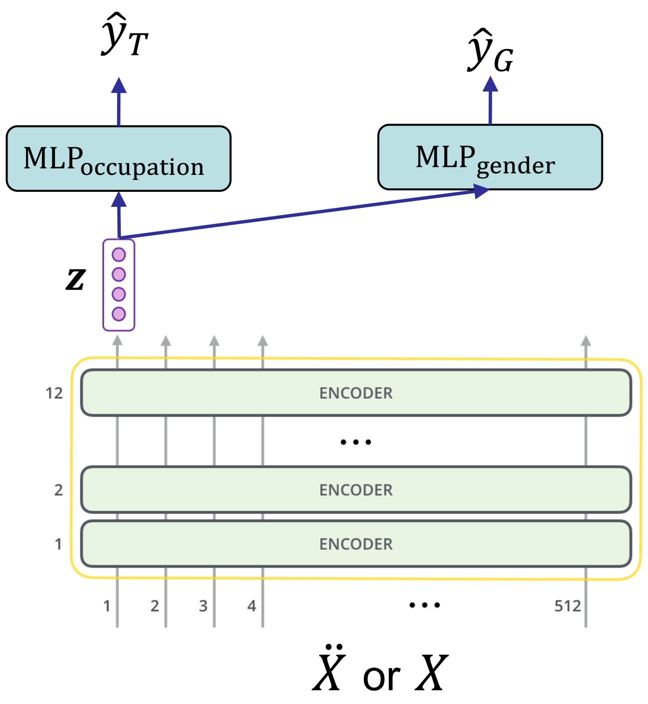
$$\mathrm{Unfairness}_{\mathrm{Nurse}} = \frac{1}{2} - \frac{5}{6} = -\frac{1}{3}$$

Unfair towards male

$$\mathrm{Unfairness}_{\mathrm{system}} = \left|-\frac{1}{8}\right| + \left|\frac{1}{3}\right|$$

Bias as information leakage

- A model is considered non-biased regarding a protected attribute, if the model's predictions are invariant/agnostic to the protected attribute
  - Embeddings of a non-biased model have no knowledge about the protected attribute
- Adversarial Attack or Probing approach:
  - After training the model on the task, add a separate MLP head on the encoded sequence vector $\boldsymbol{z}$
  - Train the new MLP head to predict the protected attribute (without updating the core model's parameters)
  - The accuracy of predicting gender on test set defined the amount of information leakage
  - If the accuracy is the same as a random classifier, no information of the protected attribute can be retrieved from the model



Bias as the difference from counter-factual examples

- Creating counter-factual variation of data points in respect to a protected attribute:

X
original

Anna graduated from Lehigh University, with honors in 1998. She has years of experience in weight loss surgery, patient support, education, and diabetes.

y_T = Nurse
y_G = Female

or

$\tilde{X}$
Counter-factual form

Henry graduated from Lehigh University, with honors in 1998. He has years of experience in weight loss surgery, patient support, education, and diabetes.

y_T = Nurse
$y_{\tilde{G}}$ = Male

- First train the model on the original data, …
- Then, pass X and $\tilde{X}$ to the model, and compare the predicted outputs
  - It can be e.g., the prediction probability of a specific class
- In a non-biased model, the predicted outputs should be the same
  - A model should be agnostic to variations in input data in respect to gender

Approaching bias/unfairness

Approaches to mitigate biases and support fairness:

- Pre-processing:
  - Data curation
  - Changing/Manipulating dataset, e.g., by training on the original as well as counter-factual data
- In-processing:
  - Consider fairness and debiasing during training *
    - Removing protected information in learned embeddings using methods such as adversarial training (representation disentanglement)
    - Add fairness criteria to model optimization
- Post-processing
  - Changing/Rearranging model’s outputs, e.g., by re-ordering search or recommendation results