Towards fair human language technologies through debiasing of semantic spaces

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Simone Paolo Ponzetto

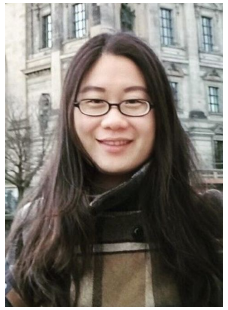
Hi there!

- Professor of Information Systems in Mannheim

Hi there!

- Professor of Information Systems in Mannheim
- Head honcho of the NLP and IR group





Hi there!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**
- We are part of the larger **Data and Web Science** fleet @ Uni Mannheim

Hi there!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Professor of Information Systems in Mannheim
- Head honcho of the **NLP and IR group**
- Today: joint work with Anne Lauscher, Goran Glavaš and Ivan Vulić

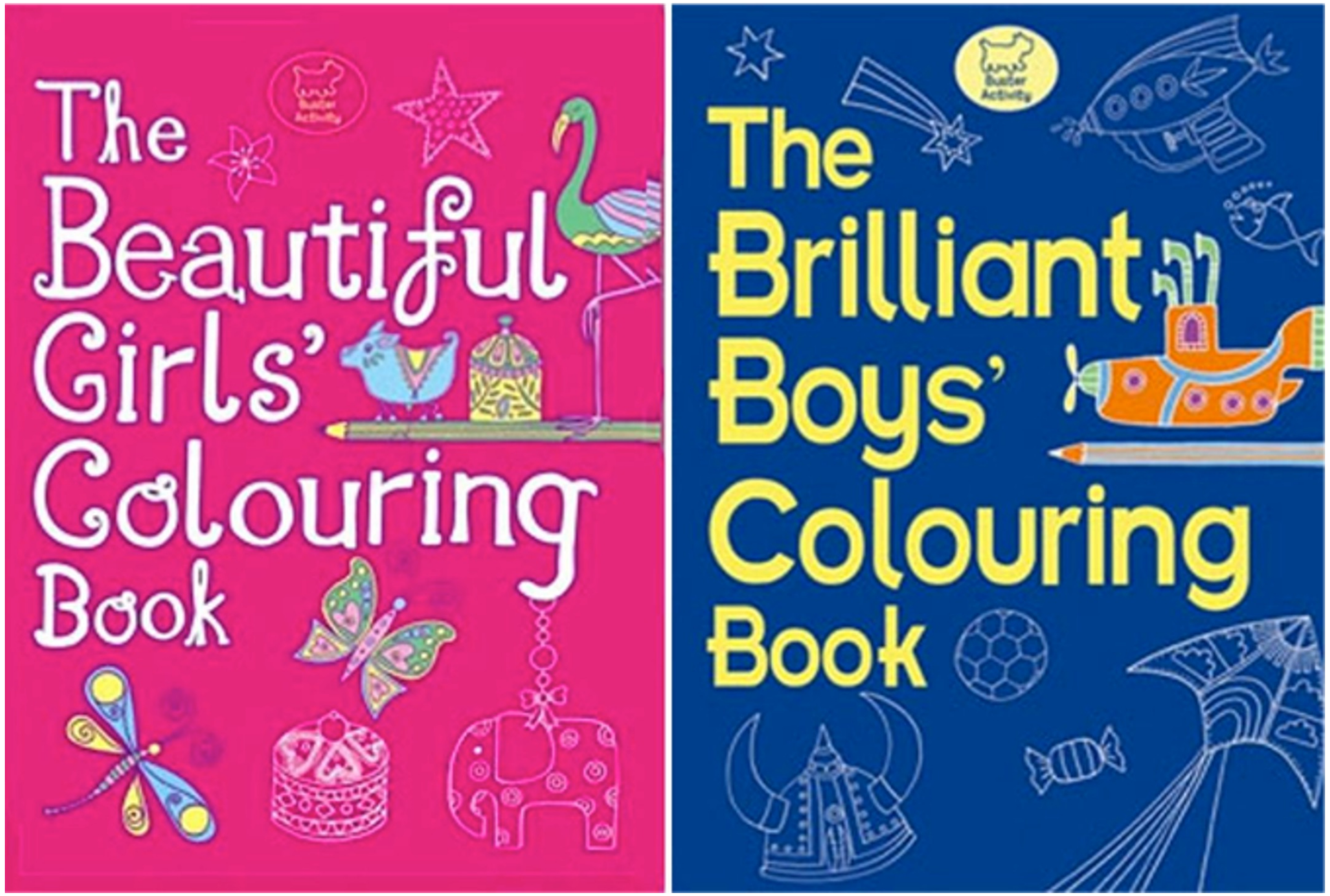
Julius-Maximilians-
UNIVERSITÄT WÜRZBURG

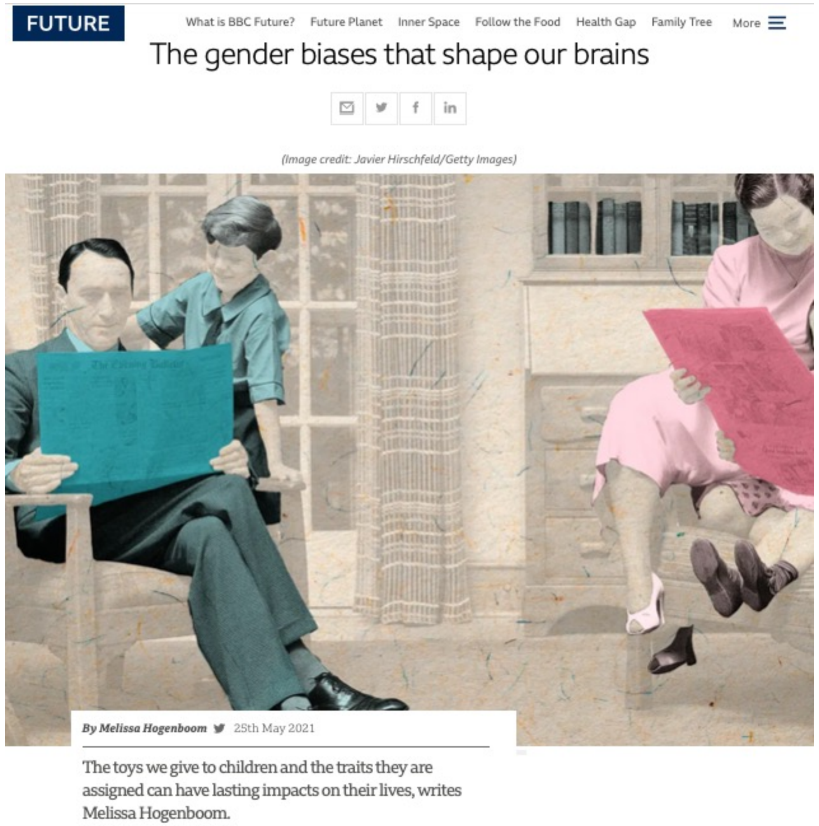
Bocconi

UNIVERSITY OF CAMBRIDGE

Bias, data and learning

What does bias look like?




FUTURE
What is BBC Future? Future Planet Inner Space Follow the Food Health Gap Family Tree More

The gender biases that shape our brains

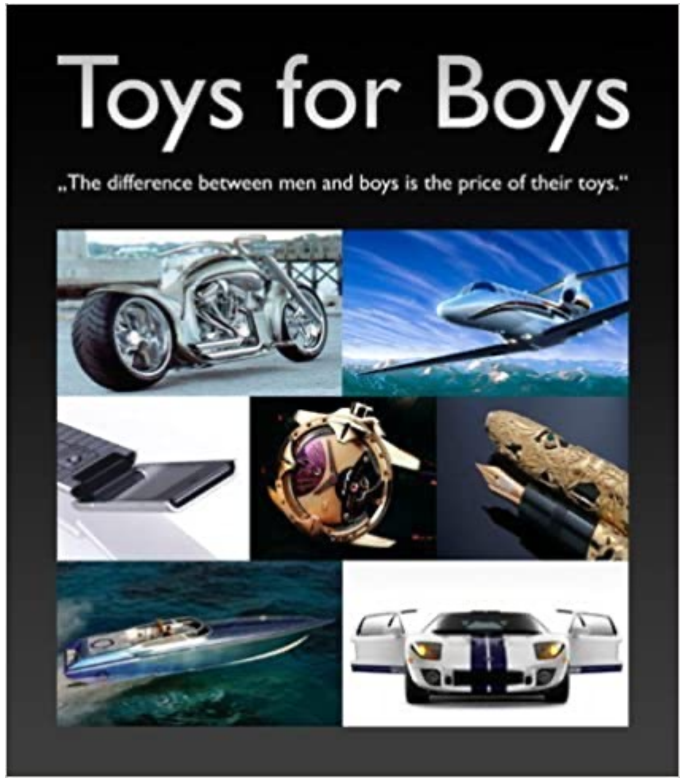
(Image credit: Javier Hirschfeld/Getty Images)

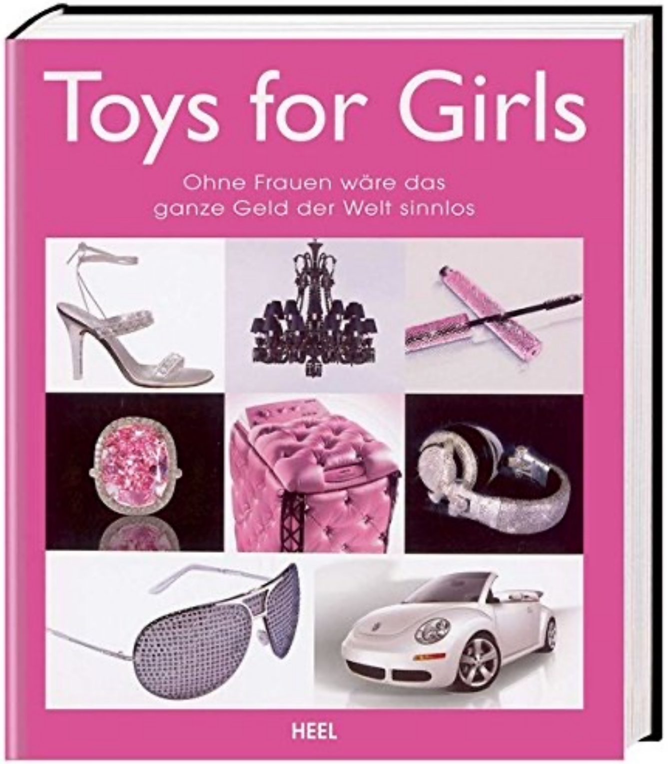
By Melissa Hogenboom 25th May 2021

The toys we give to children and the traits they are assigned can have lasting impacts on their lives, writes Melissa Hogenboom.

Source: BBC Future

What would you like to play with?





Evidence for bias

- Different **treatment** depending on **identity**
- Different **ideas** about someone depending on **identity**
- Different **expectations** about someone depending on **identity**
- Different **representation** depending on **identity**

What is the harm?
Associative vs. allocative harm

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Associative harm**: when systems reinforce the **subordination** of **some groups** along **the lines of identity**

- **An allocative harm**: when a system **allocates or withholds** certain **identity groups** an **opportunity** or a **resource**

Source: [“The Trouble with Bias” NIPS 2017 Keynote - Kate Crawford]

Allocative harm?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Air conditioning temperatures are set according to the resting metabolic rate of a **154-pound, 40 year-old man**. This overestimates women's metabolic rates by 35%+
- As office temperatures get warmer, **women perform better on cognitive tasks while men perform worse**

https://www.nature.com/articles/nclimate2741

https://journals.plos.org/plosone/article/authors?id=10.1371/journal.pone.0216362

More on allocative harm

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was “sexist” against women applying for credit.

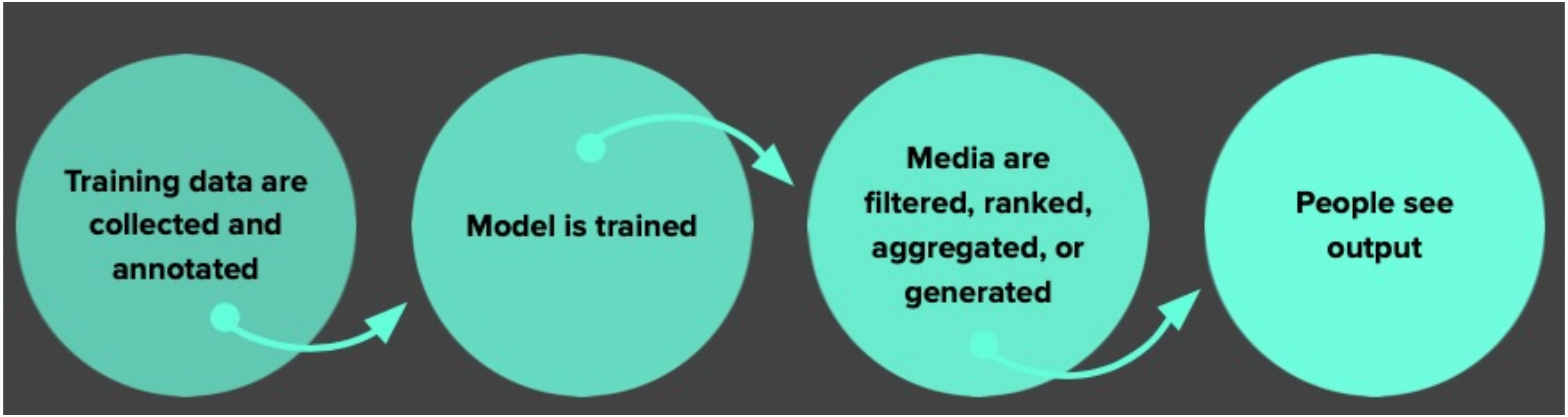
Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones
Yesterday 10:32am • Filed to: ALGORITHMS

22.3K 96 2

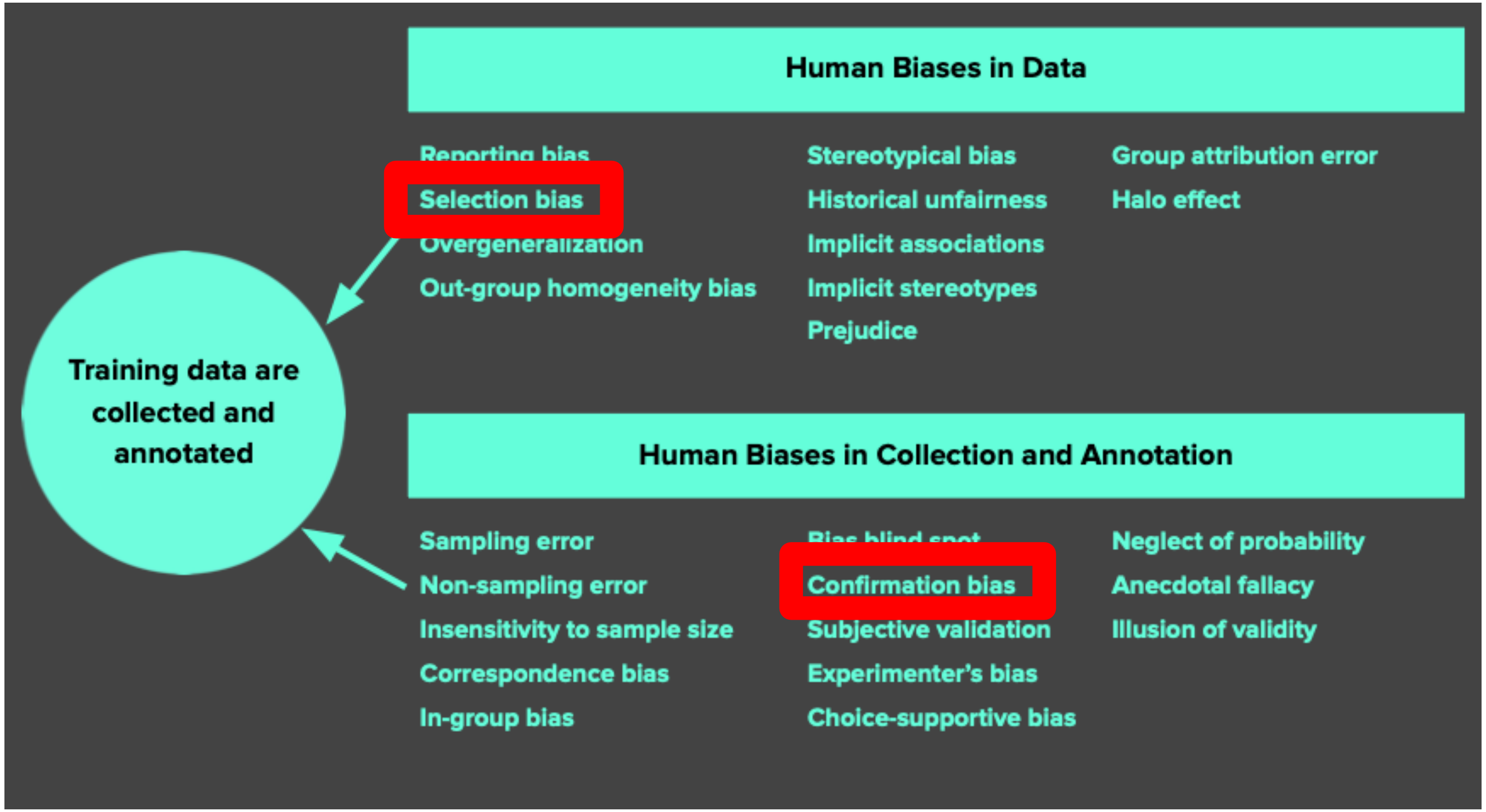
Photo: Getty

The typical “learning from data” workflow: Data => learn => predict



Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

The problems with data…



Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

Human biases in data and interpretation

Data

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

Interpretation

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

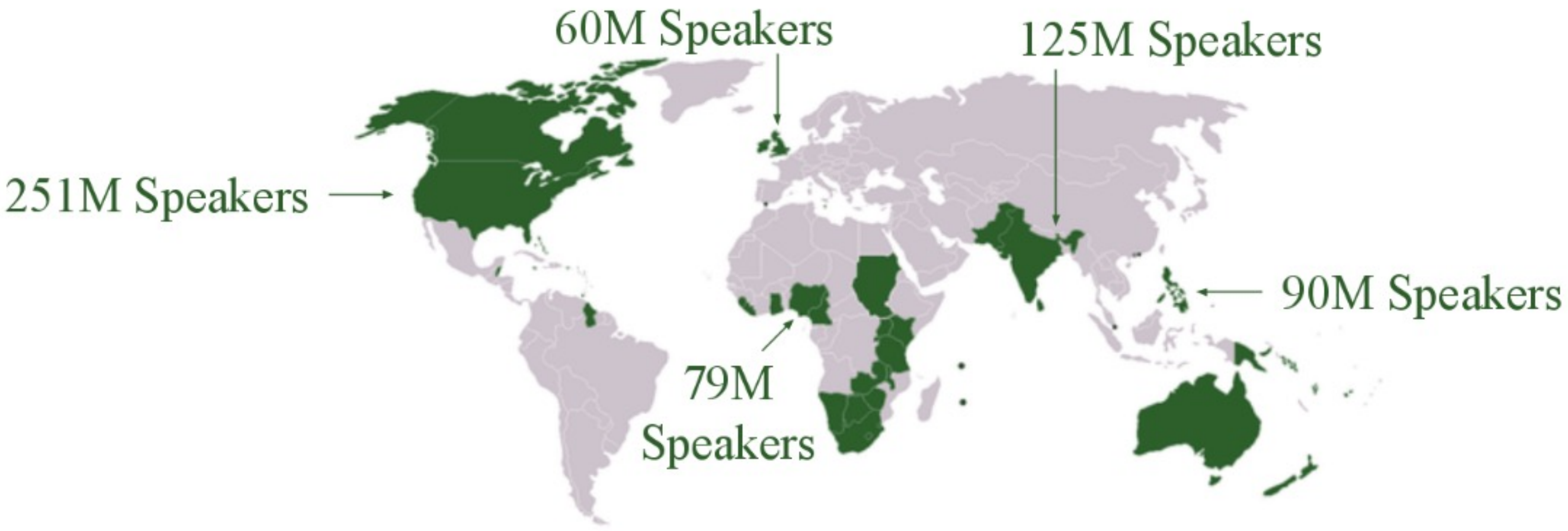
Correlation fallacy: Confusing correlation with causation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

Biases in data – Selection Bias
Selection does not reflect a random sample

Example: world Englishes



Biases in data – Selection Bias
Selection does not reflect a random sample

Gender bias on the Web

- **Males** are over-represented in the reporting of web-based news articles (Jia, Lansdall-Welfare, and Cristianini 2015)
- **Males** are over-represented in twitter conversations (Garcia, Weber, and Garimella 2014)
- Biographical articles about **women** on Wikipedia disproportionately discuss **romantic relationships or family-related issues** (Wagner et al. 2015)
- IMDB **reviews written by women are perceived as less useful** (Otterbacher 2013)

Biases in the data lead to biases in the predictions!

**No Classification without Representation:
Assessing Geodiversity Issues in
for the Developing W**

Shreya Shankar, Yoni Halpern, Eric Breck, James Atwo
{shankarshreya, yhalpern, ebreck, atwoodj, jim
Google Brain Team

Figure 2: Distribution of the geographically identifiable images
country. Almost a third of the data in our sample was US-based, and 60% of the data was from the
six most represented countries across North America and Europe.

Wedding photographs (donated by Googlers), labeled by a classifier trained on the Open Images dataset. The classifier's label predictions are recorded below each image.

Source: [paper] and [blog post]

Biases in the data lead to biases in the predictions!

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

The Royal Family @RoyalFamily Follow

Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.

da'Rah-zingSun @TIME7SS Follow

@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrnt evrywhere, u kno wut she means jus like we do!

Mooktar @bossmukky Follow

"@Ecstatic_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...

Ebenezer· @Physique_cian Follow

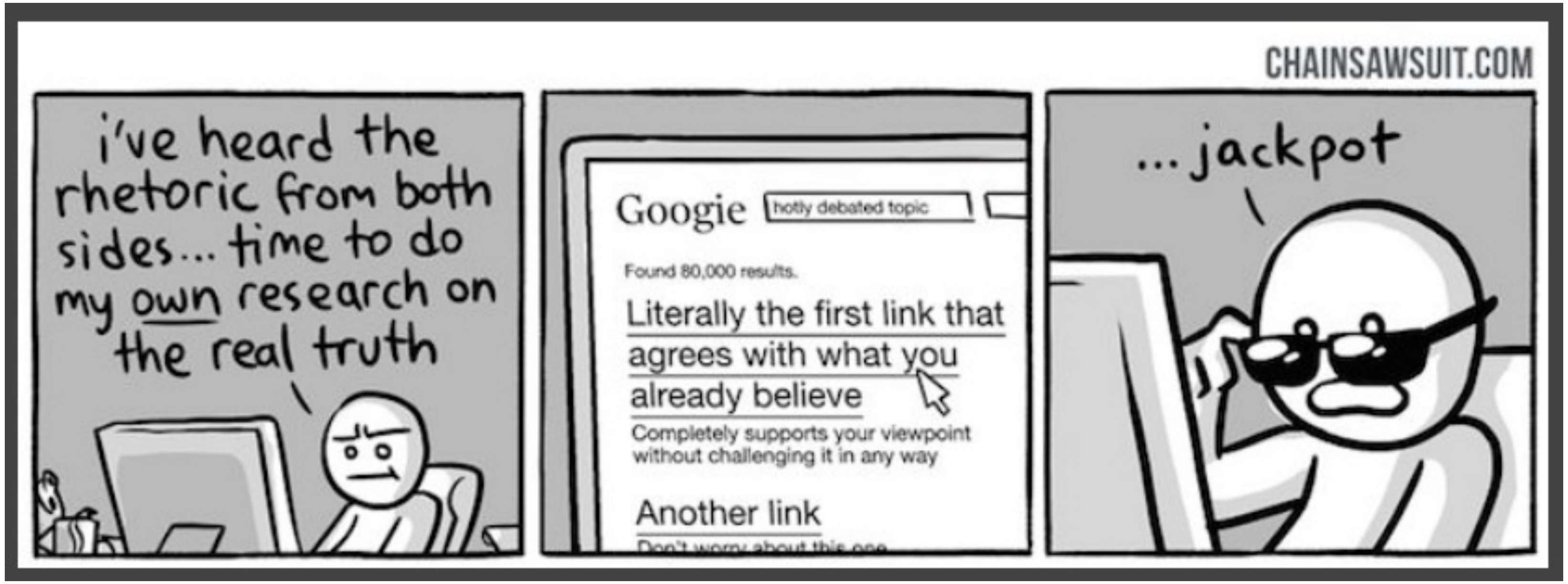
@Tblazeen R u a wizard or wat gan sef : in d mornin- u tweet, afternoon - u tweet, nyt gan u dey tweet.beta get ur IT placement wiv twitter

- Language identification degrades significantly on African American Vernacular English (Blodgett et al. 2016)

Biases in interpretation: Confirmation bias

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs



Credit: Chainsawsuite by kris staub / Source: EMNLP 2019 tutorial

Biases in interpretation: Confirmation bias

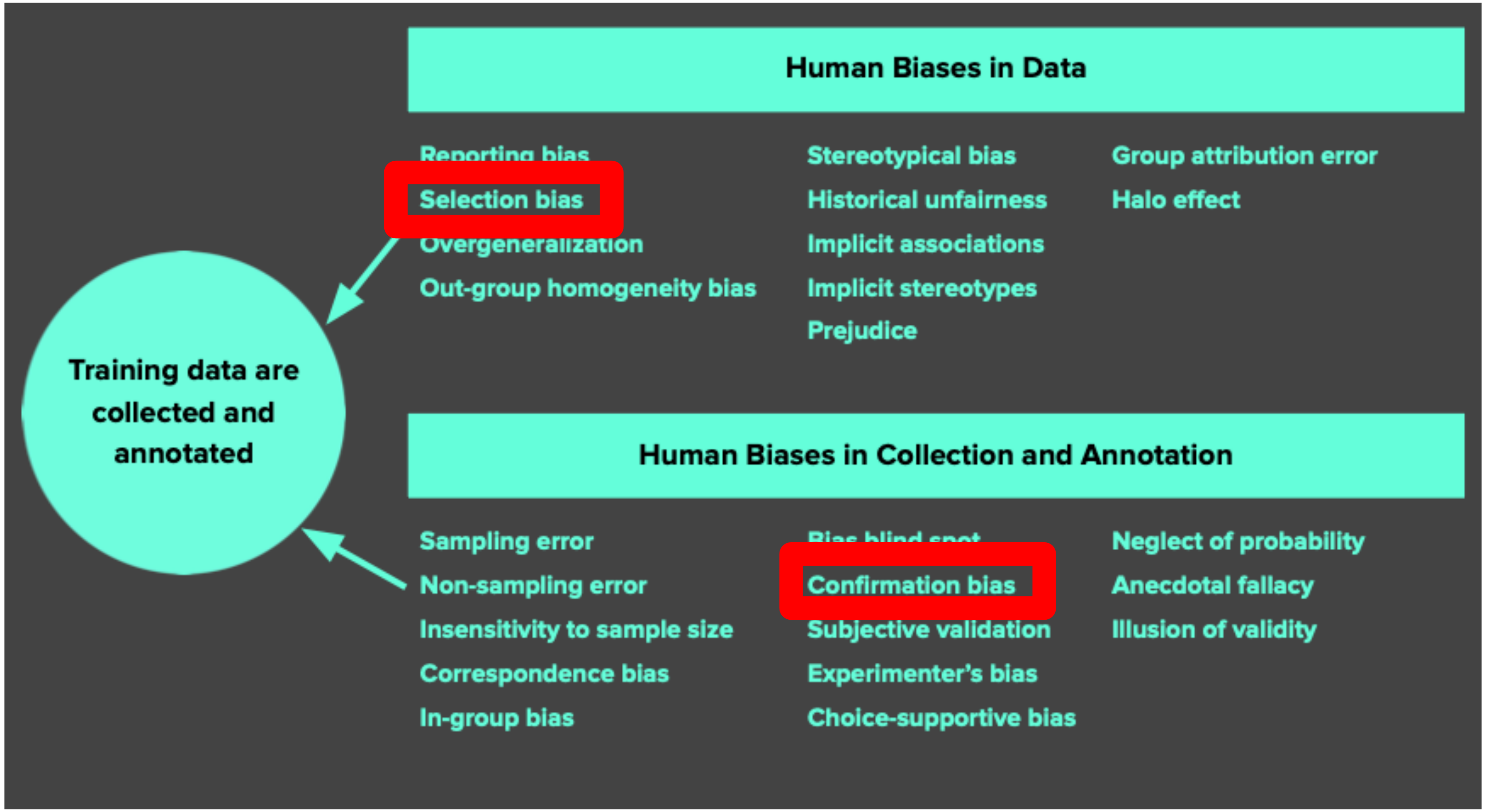
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Crowd workers tend to **judge as more truthful news statements** coming from speakers off *the same political party* that they have recently voted for [La Barbera *et al.*, 2020]
- Crowd workers are more likely to **label a statement as neutral** (as opposed to opinionated) *if its stance aligns with their own opinions* [Hube *et al.*, 2019]

[La Barbera *et al.*, 2020] David La Barbera, Kevin Roitero, Gian- luca Demartini, Stefano Mizzaro, and Damiano Spina. Crowd- sourcing truthfulness: The impact of judgment scale and assessor bias. In *European Conference on Information Retrieval*, pages 207–214. Springer, 2020.

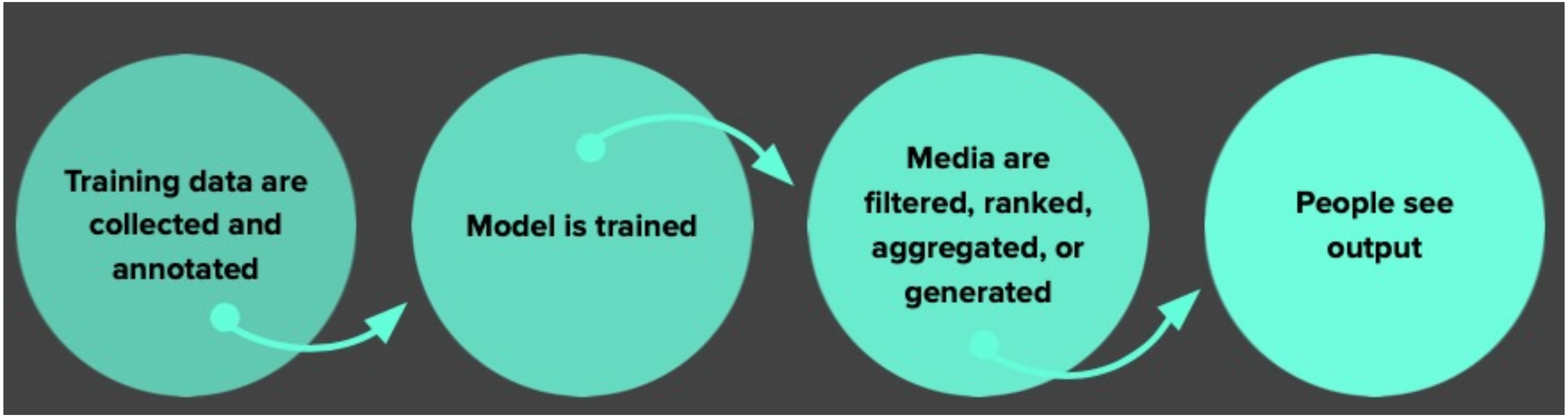
[Hube et al.,2019] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

The problems with data…



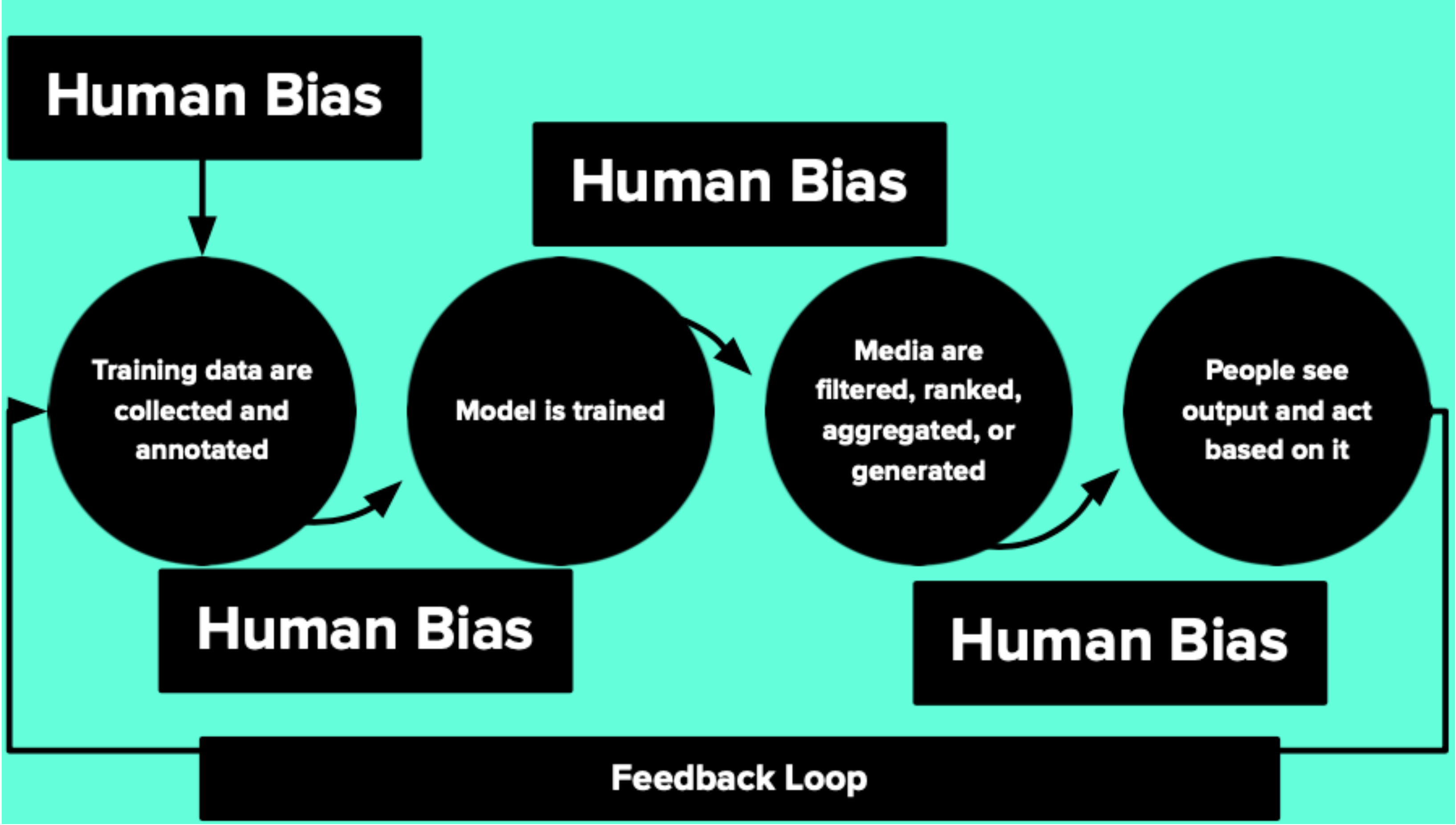
Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

The typical “learning from data” workflow: Data => learn => predict



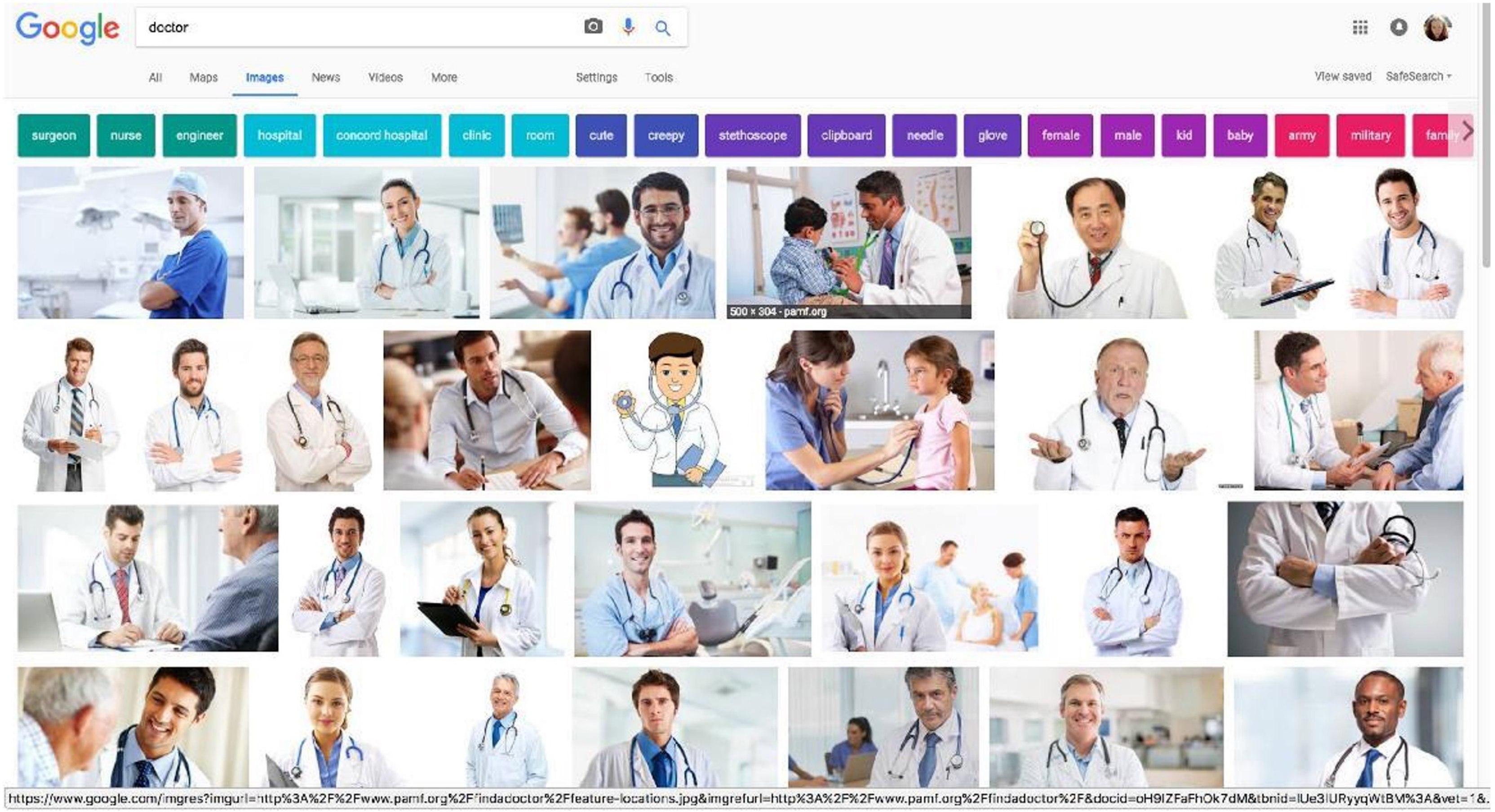
Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

Biases' reinforcement loop

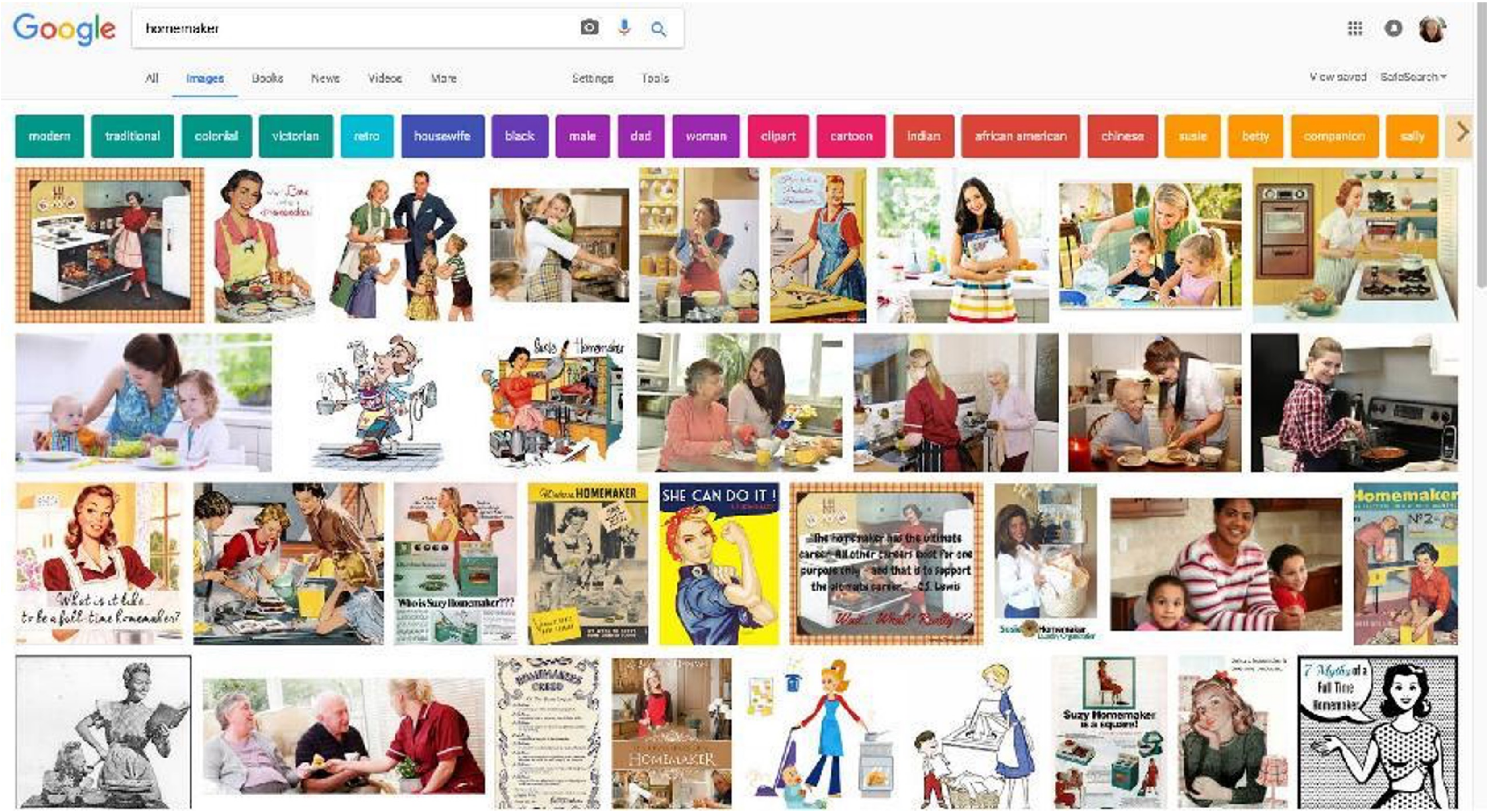


Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

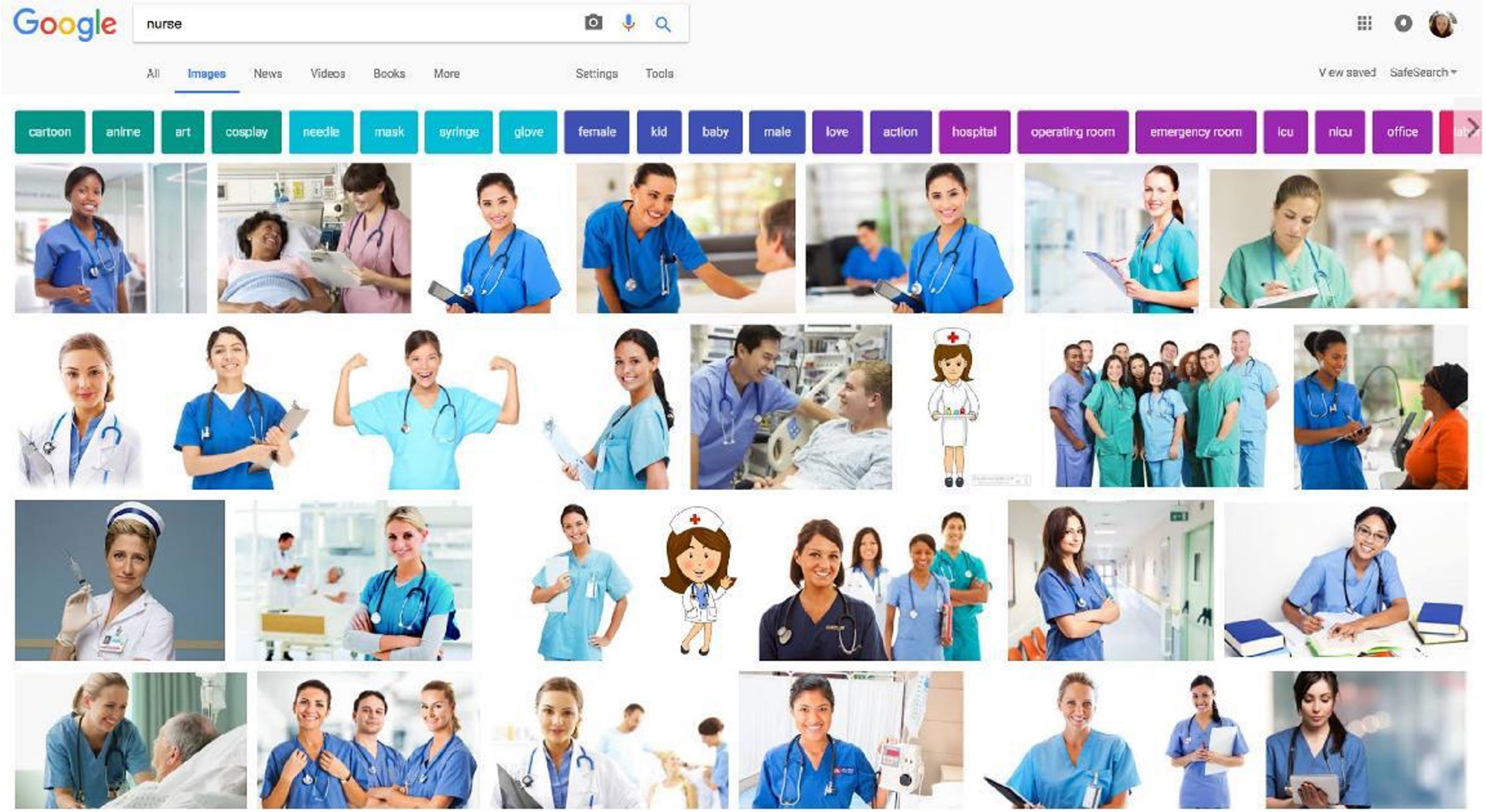
In search of doctors



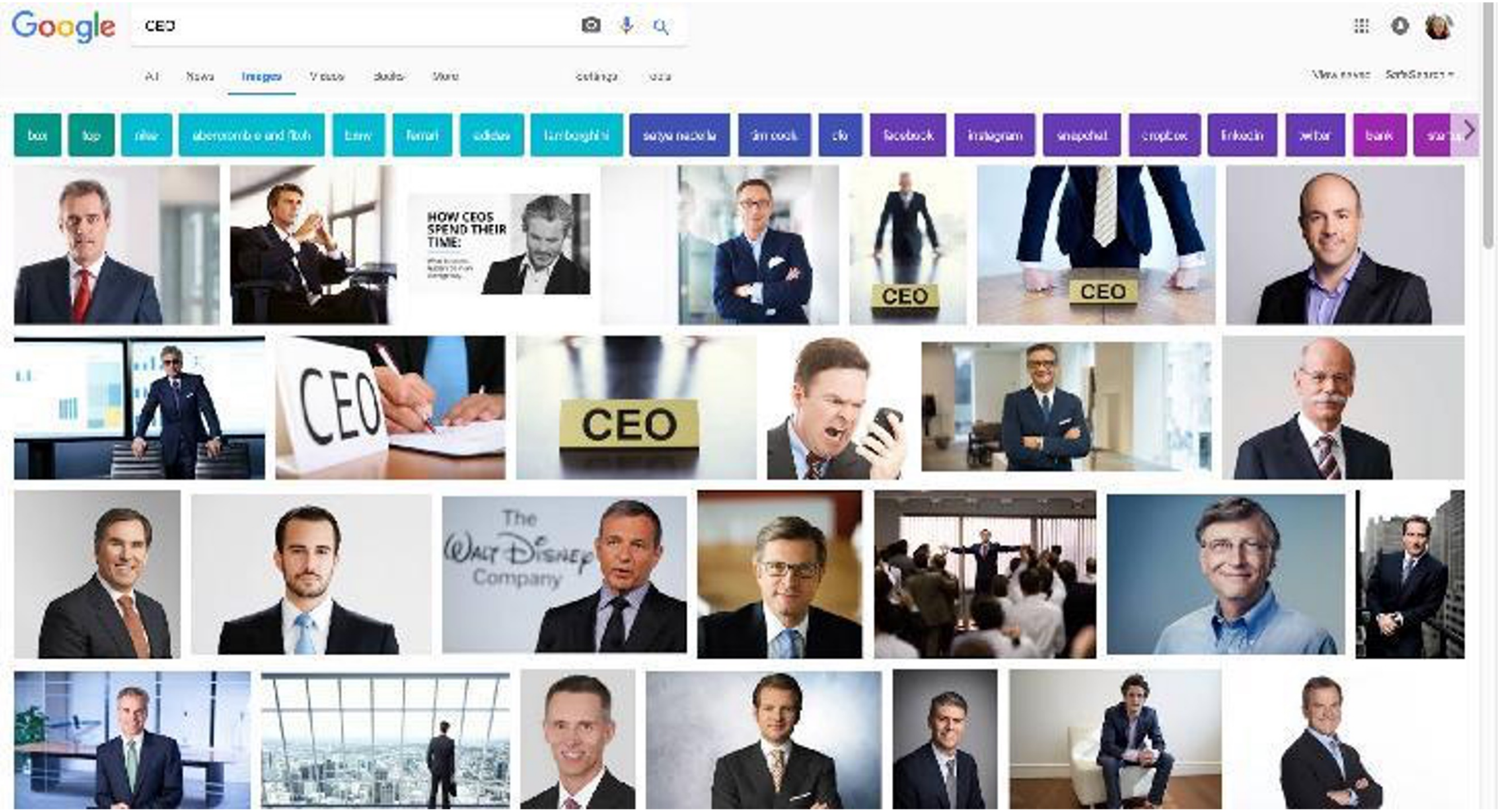
Google, show me an homemaker



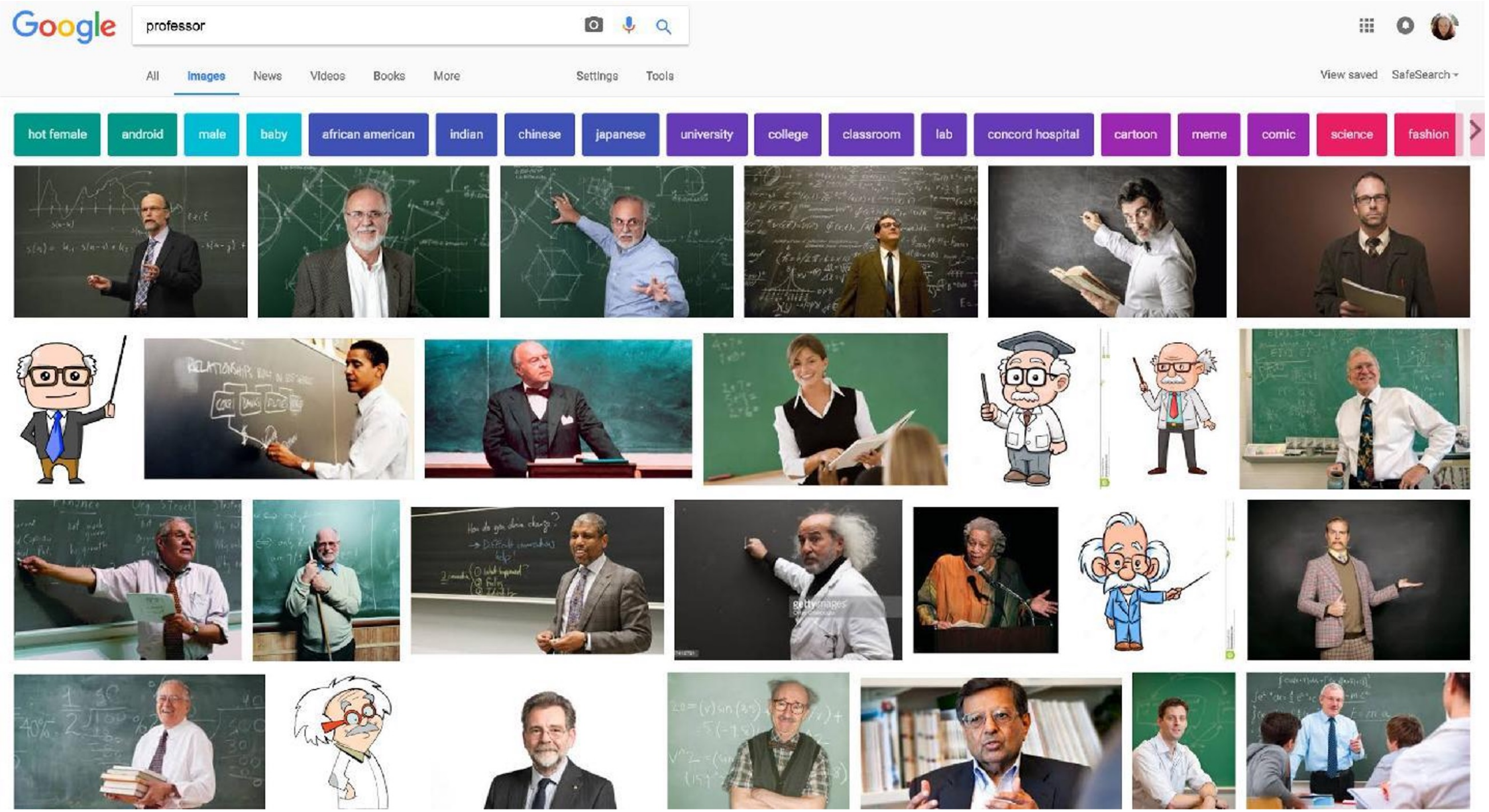
How does a nurse look like?



What about a CEO?

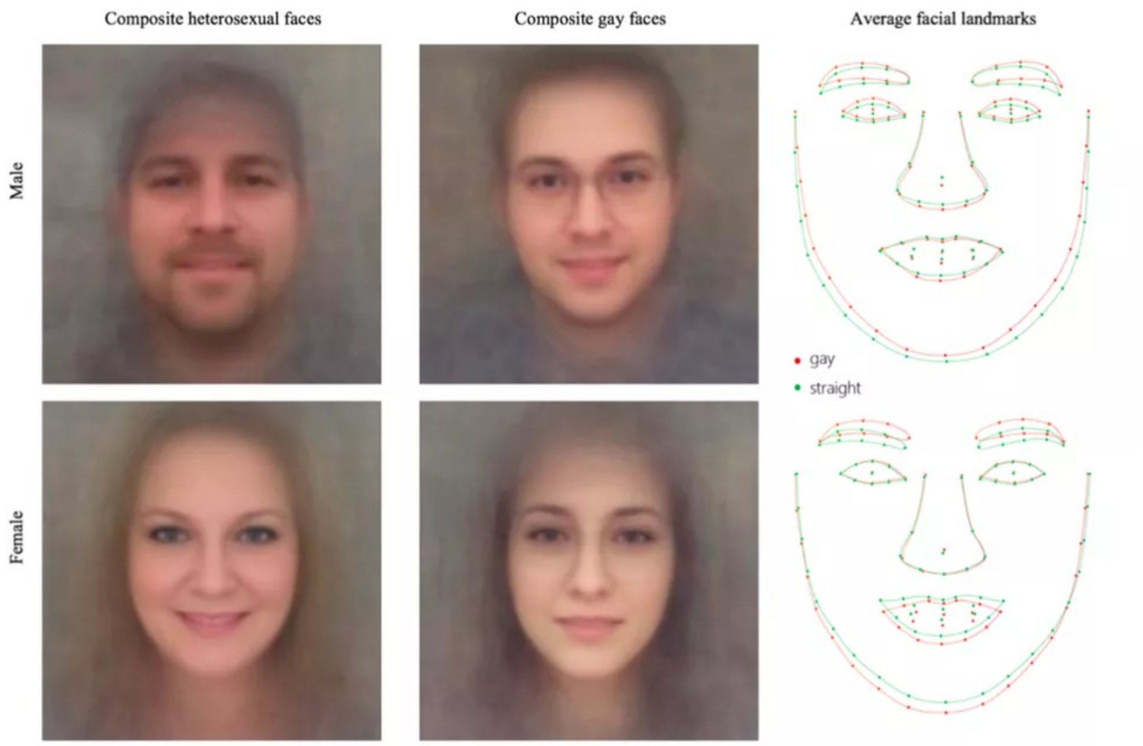


Herr Kollege Prof. Dr.



The “A.I. Gaydar”

- A sexual orientation detector learned from data
- Wait… predicting sexuality?!



Wang & Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of Personality and Social Psychology. February 2018, Vol. 114, Issue 2, Pages 246-257.

Problems with the “A.I. Gaydar”

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- **Research question**
  - Identification of sexual orientation from facial features
- **Data collection**
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- **Method**
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- **Accuracy**
  - 81% for men, 74% for women

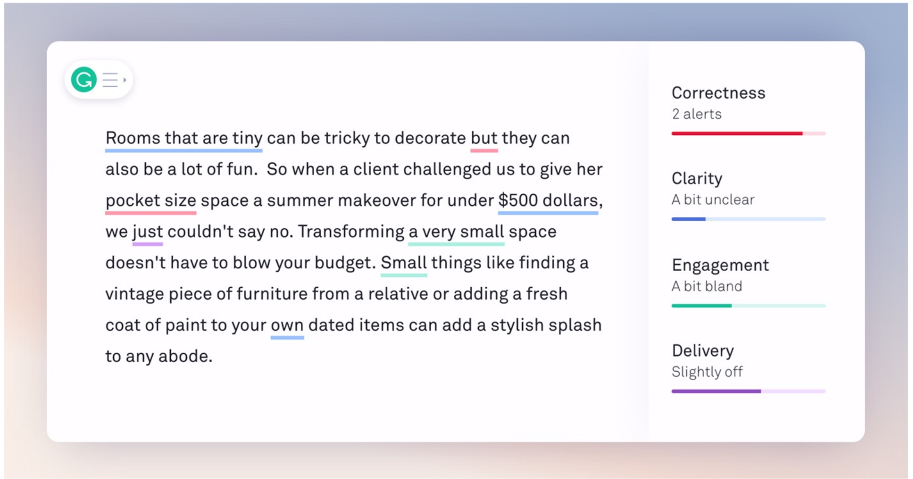
A few crucial questions

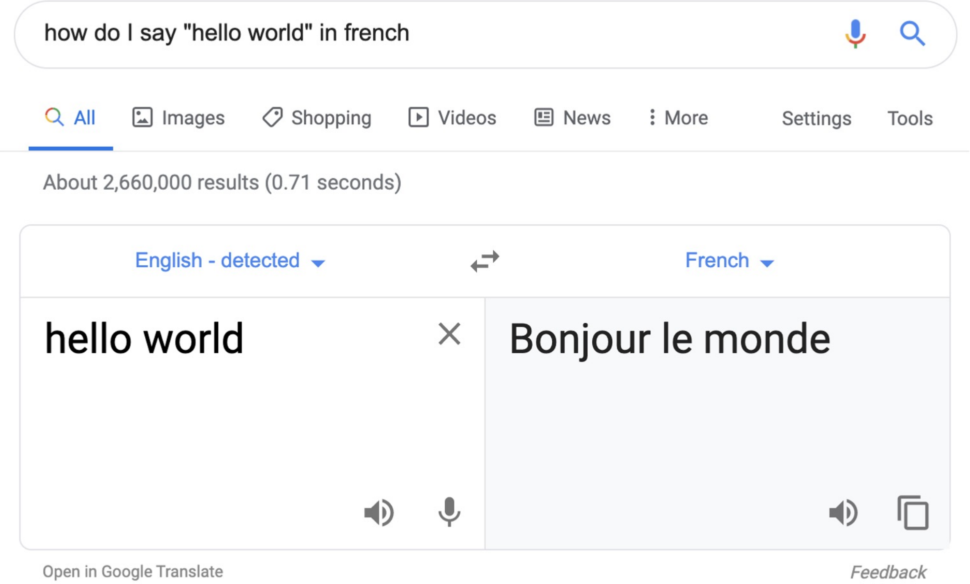
- Who could benefit from such a technology?
- Who can be harmed by such a technology?
- Representativeness of (training) data
- What are confounding variables and corner cases to control for?
- Can prediction errors have major effect on people's lives?
- Does the system optimize for the "right" objective?

Language Technologies and bias

Natural Language Processing: some initial thoughts…

– Methods to **automatically process (i.e., understand and generate) natural language data**





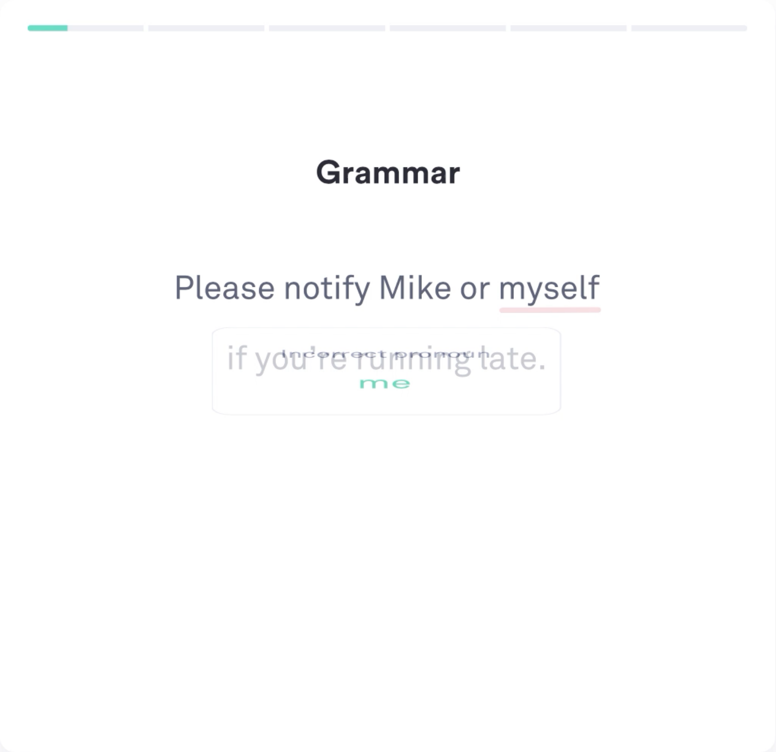
A few applications of Natural Language Processing

- Spelling correction
- Grammar checking
- Text completion
- Speech-to-text and vice versa
- Dialogue systems
- Question Answering
- Summarization
- Machine translation

Example: writing assistant

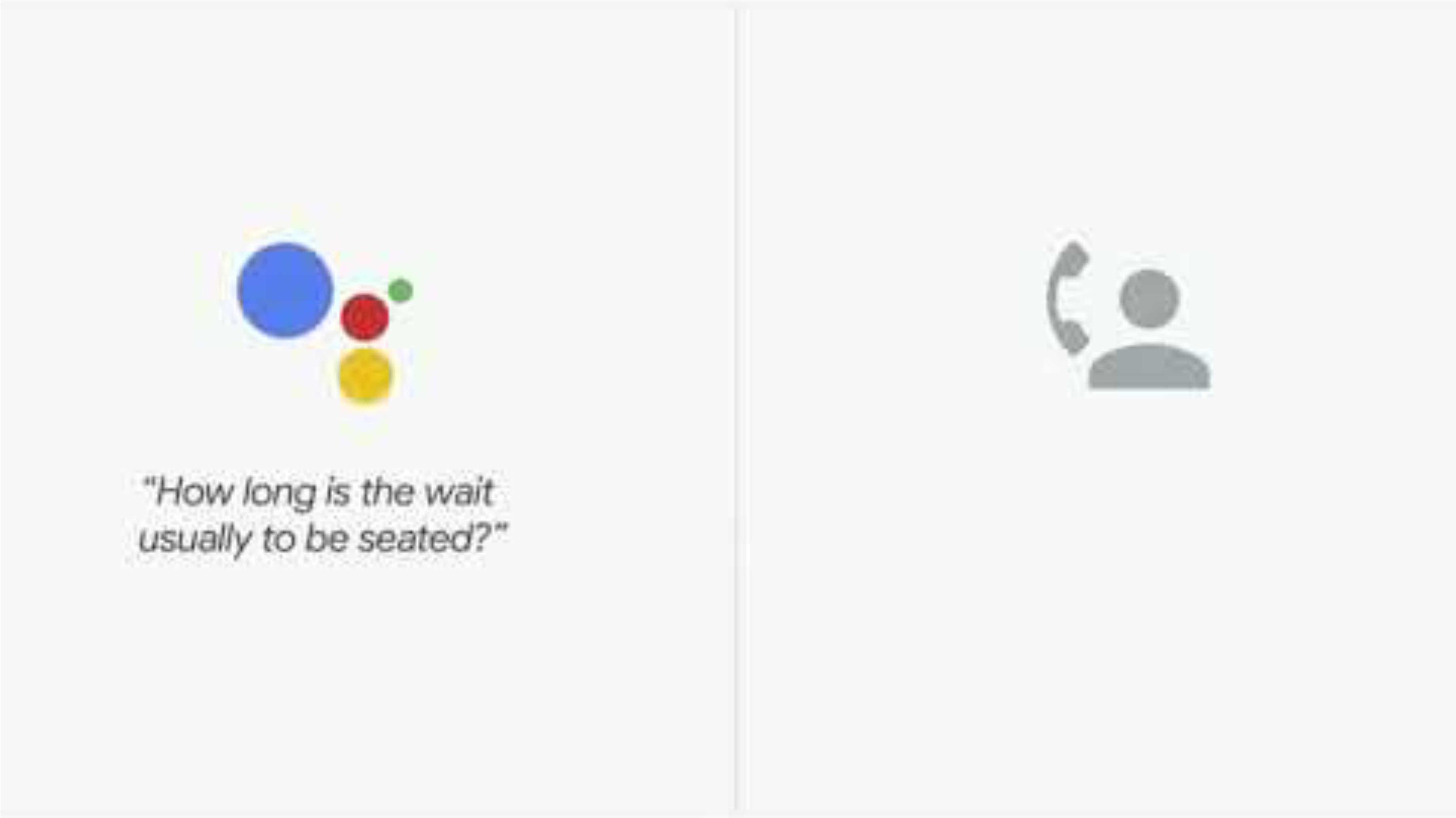
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

grammarly

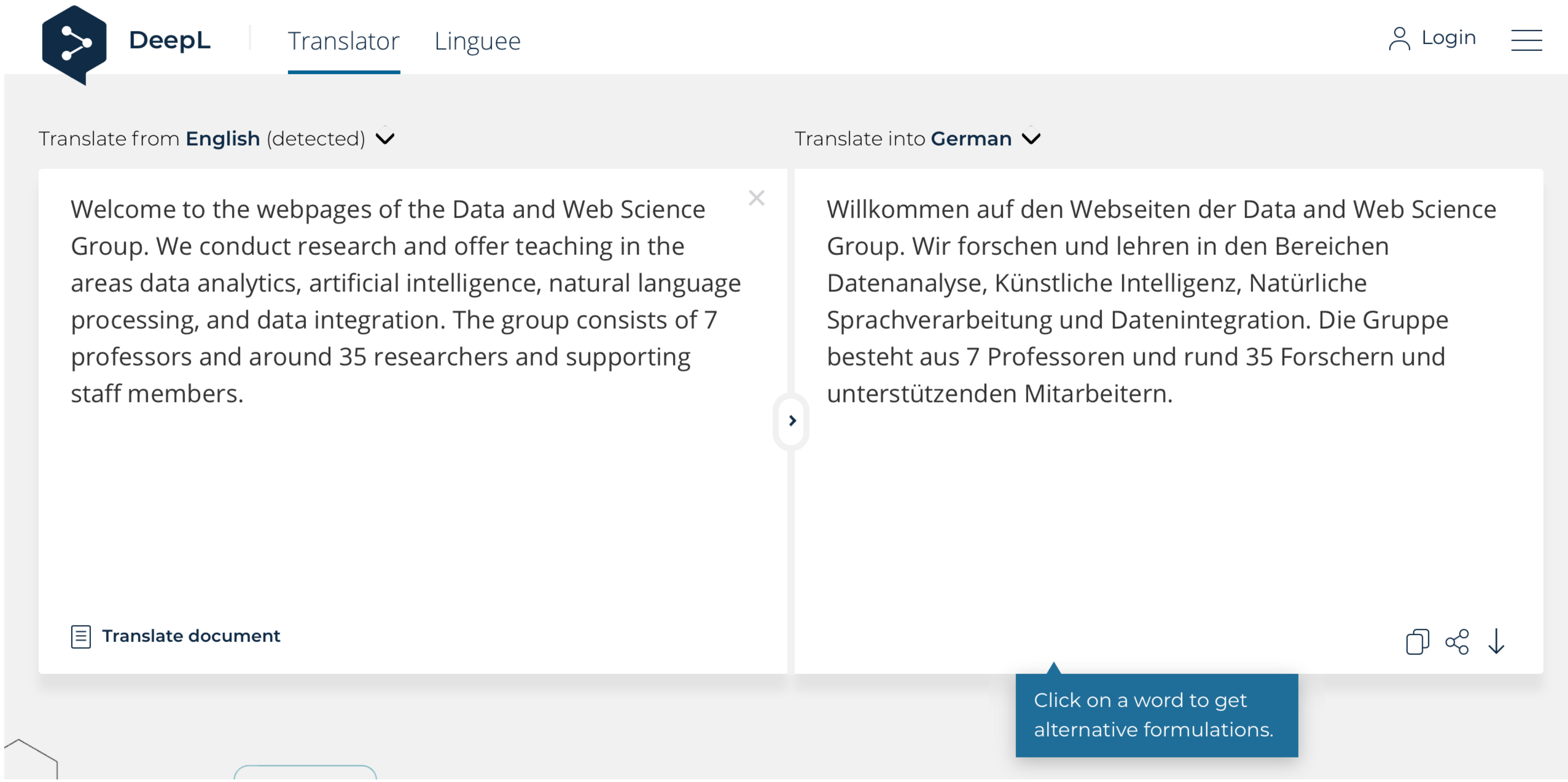


Example: virtual assistant

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

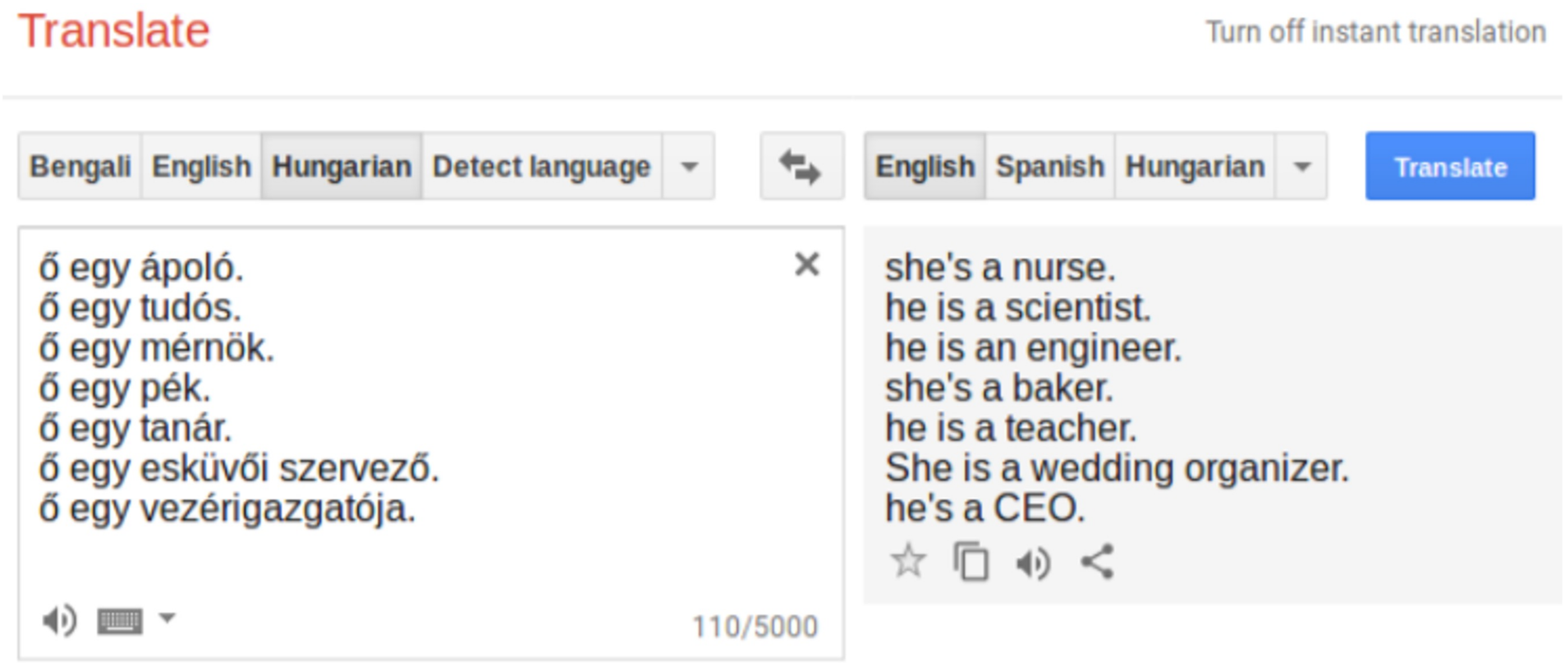


Example: machine translation



Bias in NLP models: an example with MT

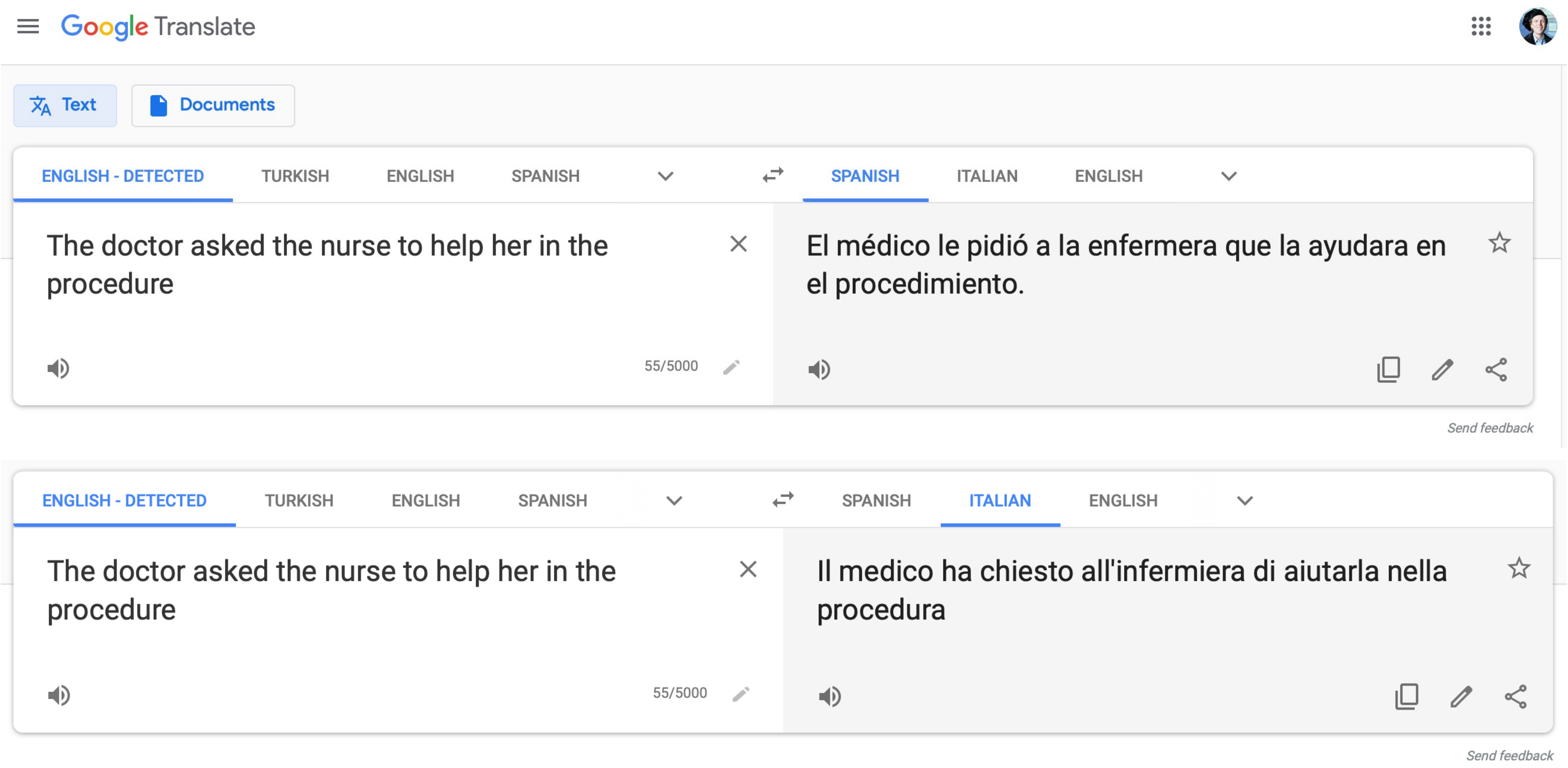
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



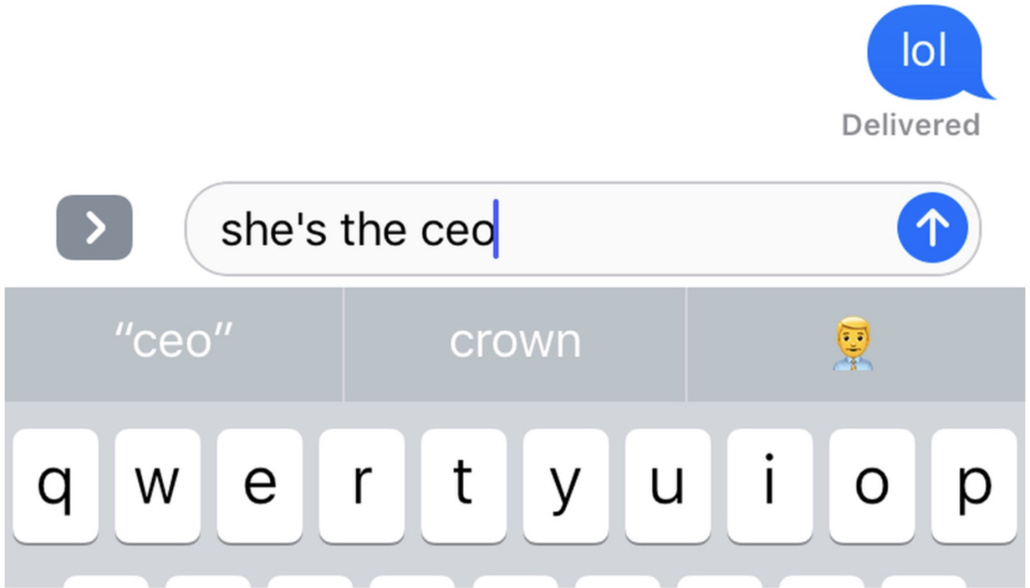
More examples with MT!

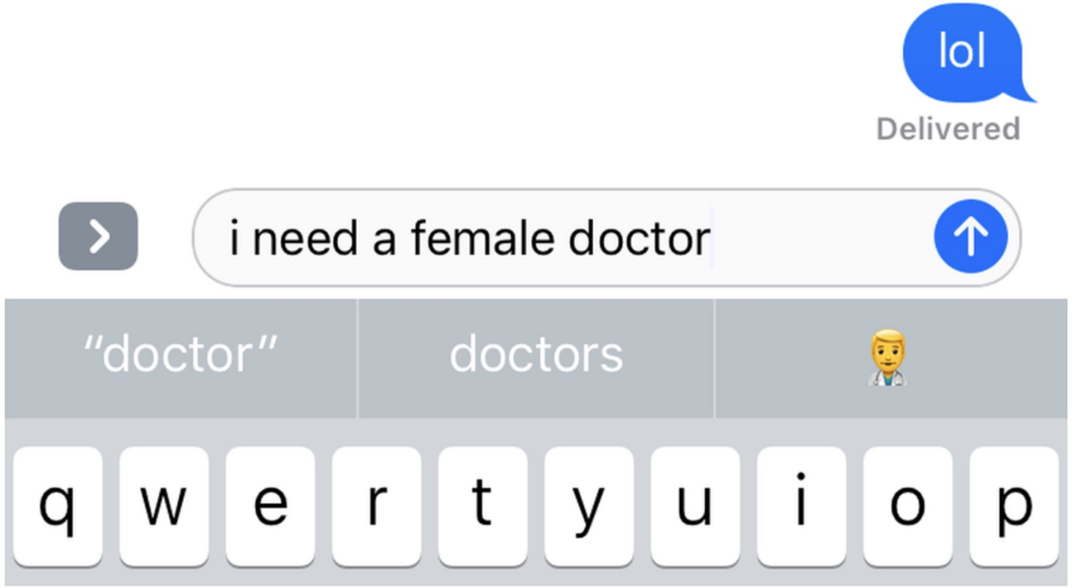
Source: https://twitter.com/_DianeKim/status/915693210088984576/photo/1

Bias in NLP models: more MT



WhatsApp recommending emojis…





Gender bias in coreference resolution and language modeling

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Coreference scores and conditional log-likelihood indicate implicit bias in coreference resolution and language modelling (Lu et al., 2019)

$$1_{\square}\text{: The } \overbrace{\underline{\textbf{doctor}}\text{ ran because }\underline{\textbf{he}}}^{5.08}\text{ is late.}$$

$$1_{\bigcirc}\text{: The } \overbrace{\underline{\textbf{doctor}}\text{ ran because }\underline{\textbf{she}}}^{1.99}\text{ is late.}$$

$$2_{\square}\text{: The } \overbrace{\underline{\textbf{nurse}}\text{ ran because }\underline{\textbf{he}}}^{-0.44}\text{ is late.}$$

$$2_{\bigcirc}\text{: The } \overbrace{\underline{\textbf{nurse}}\text{ ran because }\underline{\textbf{she}}}^{5.34}\text{ is late.}$$

(a) Coreference resolution

	A	B	$\ln \Pr[B \mid A]$
$1_{\square}$:	**He** is a	**doctor**.	-9.72
$1_{\bigcirc}$:	**She** is a	**doctor**.	-9.77
$2_{\square}$:	**He** is a	**nurse**.	-8.99
$2_{\bigcirc}$:	**She** is a	**nurse**.	-8.97

(b) Language modeling

Figure 1: Examples of gender bias in coreference resolution and language modeling as measured by coreference scores (left) and conditional log-likelihood (right).

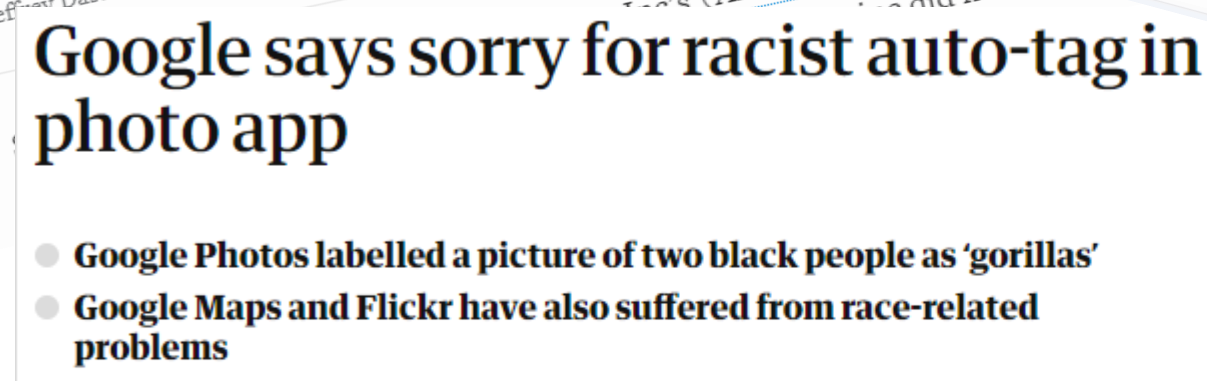
Debiasing of semantic spaces

The story so far: the (potentially dangerous) social impact of AI/NLP


BUSINESS NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool tha showed bias against women

8 MIN READ


Google says sorry for racist auto-tag in photo app

- Google Photos labelled a picture of two black people as 'gorillas'
- Google Maps and Flickr have also suffered from race-related problems

How does this relate to mainstream NLP methods?

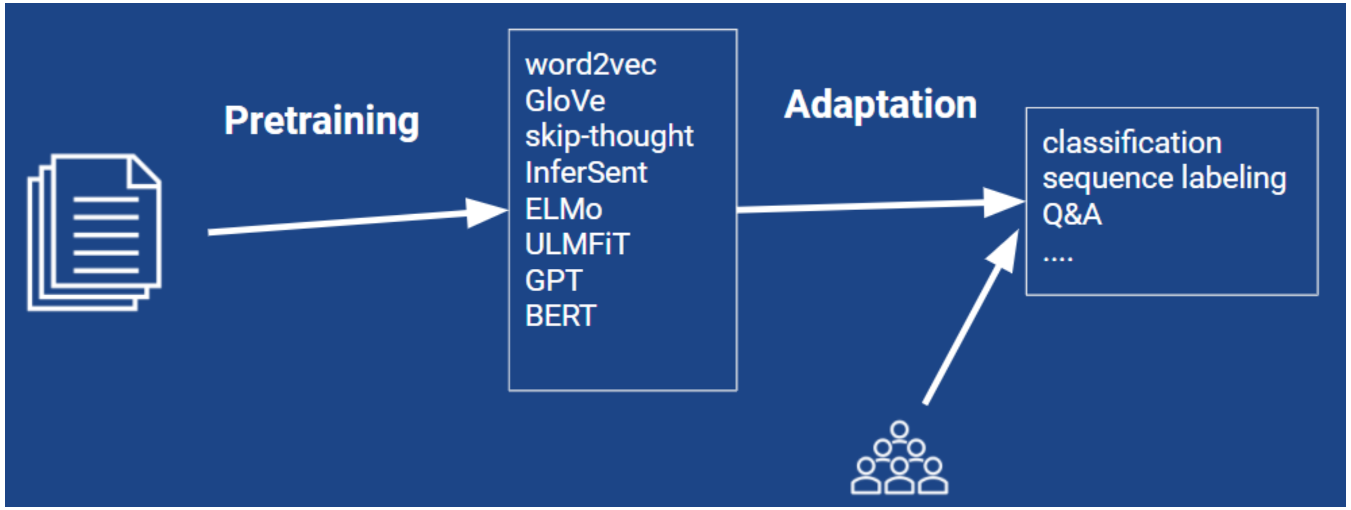
A typical NLP workflow

- Text as input
- Encode **text into a representation**
  - e.g., vectors whose dimentions capture ”dimentions of meaning”
- Use the **text representations as input** to a task-specific model
  - e.g., a sentiment classifier

Sequential transfer learning

- **Core idea**: pretrain the language/text encoder on large amounts of text, so that it learns ”the language”
  – Structure of the language (i.e., syntax)
  – Compositionality of meaning in the language (i.e., semantics)
- If we could „pre-train” such an encoder, it would be generally useful for a **wide spectrum of NLP tasks**

Image source: [NAACL tutorial on Transfer Learning for NLP]()



Lexical semantic vector representations

- A model of **word meaning** focused on *similarity*
- Define the meaning of a word as a vector, a list of numbers, a point in N- dimensional space
- Similar words are "nearby in space"

Source: Jurafsky & Martin (2018)

Words in Space → Word Embeddings

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Representing words in a vector space is a standard process in NLP, called **embedding**
- It is called “embedding” because the objects are embedded into a vector space
- In our case, we embed words, so we obtain *word embeddings*

- An **embedding** of a word is nothing but a numeric vector that aims to capture some properties (typically meaning) of the word

Word representations

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Distributional hypothesis: „you'll know a word by the company it keeps" (Harris, 1954)
- Word representations are derived from word co-occurrences in a large corpus of text

  . . . the quick brown | fox | jumps over the . . .

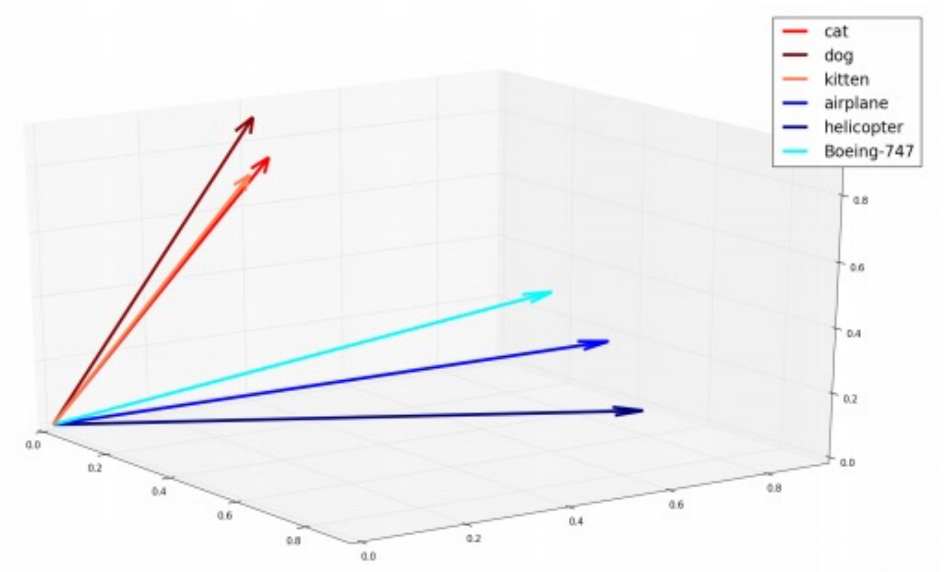
- Assumption: **the contexts in which the word appears, define its meaning**
  - This allows to create a (still rather sparse) V x V dimension matrix of co-occurrences between words
  - Word vectors from the co-occurrence matrix can now be compared (similar words will appear in similar contexts, hence have similar vectors)

Word representations

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Dense representations

- Each word is represented by a dense vector, a point in a vector space
- The dimension of the semantic representation d is usually much smaller than the size of the vocabulary (d << V)
- All dimensions contain real-valued numbers (possibly normalized between −1 and 1)



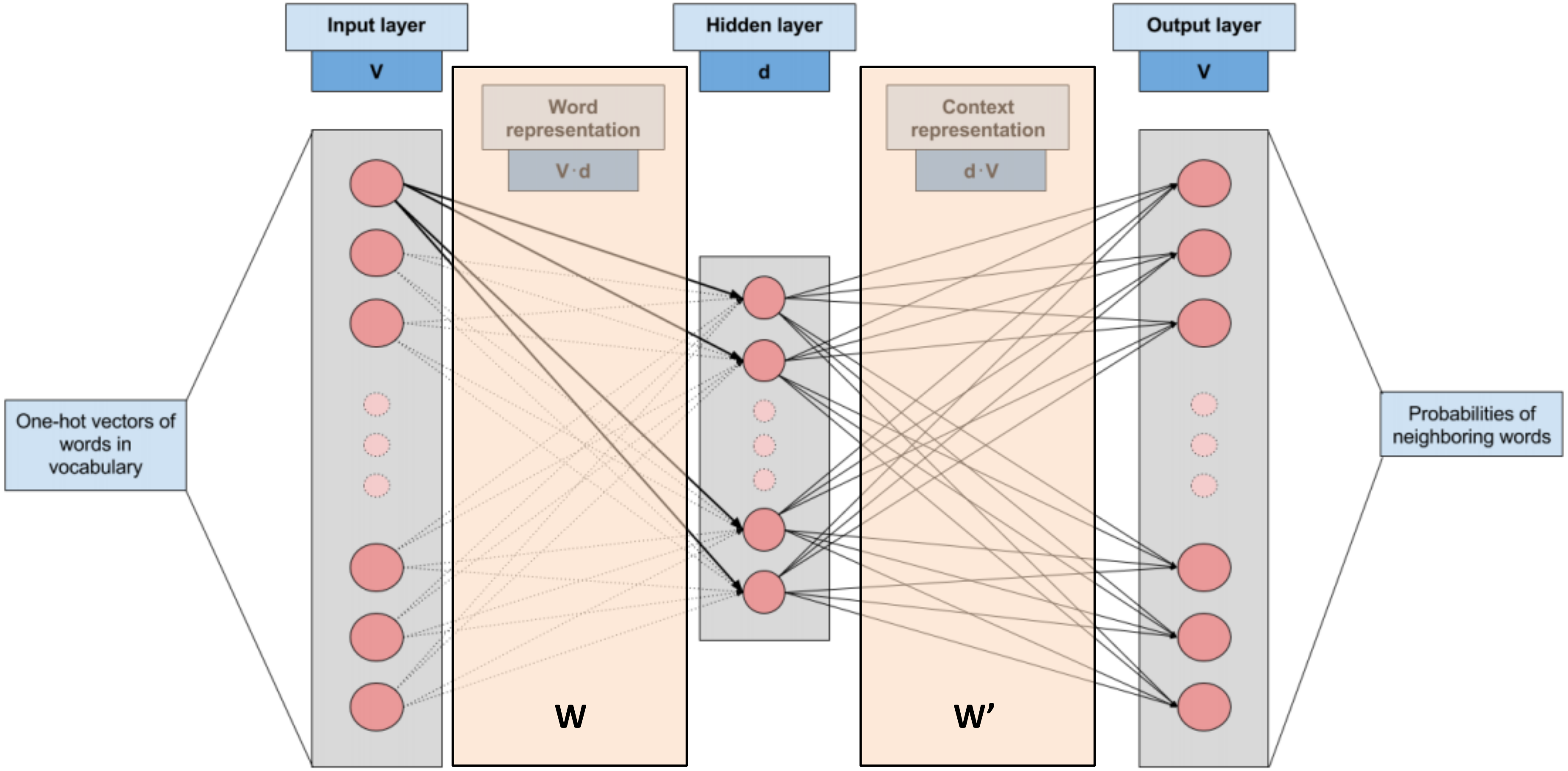
Word Embeddings

WORD	d1	d2	d3	d4	d5	...	d50
summer	0.12	0.21	0.07	0.25	0.33	...	0.51
spring	0.19	0.57	0.99	0.30	0.02	...	0.73
fall	0.53	0.77	0.43	0.20	0.29	...	0.85
light	0.00	0.68	0.84	0.45	0.11	...	0.03
clear	0.27	0.50	0.21	0.56	0.25	...	0.32
blizzard	0.15	0.05	0.64	0.17	0.99	...	0.23
...	...	...	...	...	...	...	...

Skip-Gram (SG) model

- Start by assigning two different dense random vectors to each word
  - **Center vector** and **context vector** (each of size d << V)
- For a center word, predict the words will appear in its context
  - E.g., given „fox” predict „quick”; „brown”; „jumps”; „over”

Skip-Gram (SG) model

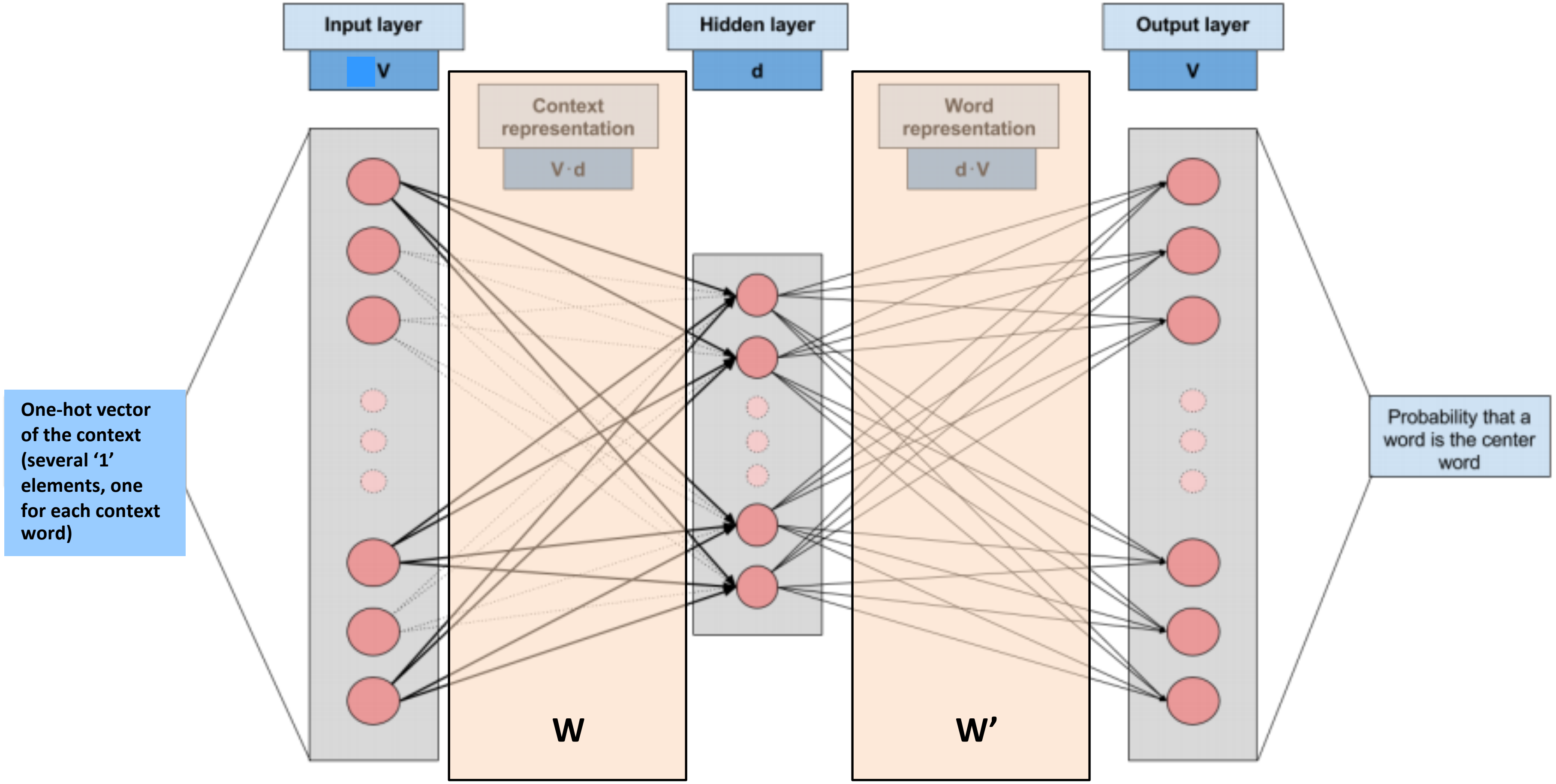


Continuous bag-of-words (CBOW)

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- In a sense, a model inverse to Skip-Gram – predicts the central word from the context
- Given context, predict the center word
  - E.g., given „quick brown _ jumps over” predict „fox”

Continuous bag-of-words (CBOW)



Word embeddings – results

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Airplane

word	cosine
plane	0.835
airplanes	0.777
aircraft	0.764
planes	0.734
jet	0.716
airliner	0.707
jetliner	0.706

Cat

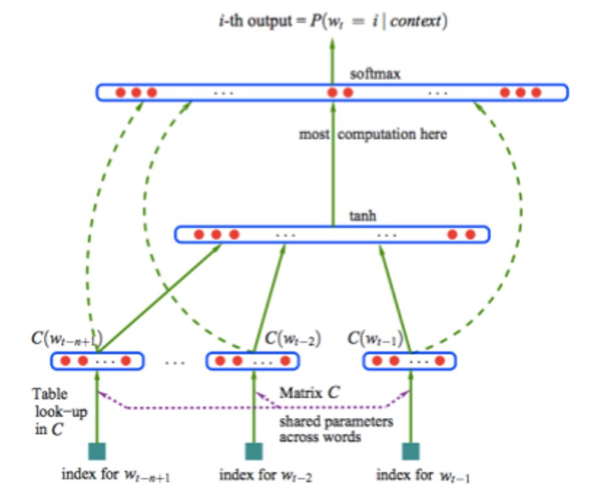
word	cosine
cats	0.810
dog	0.761
kitten	0.746
feline	0.732
puppy	0.707
pup	0.693
pet	0.689

Dog

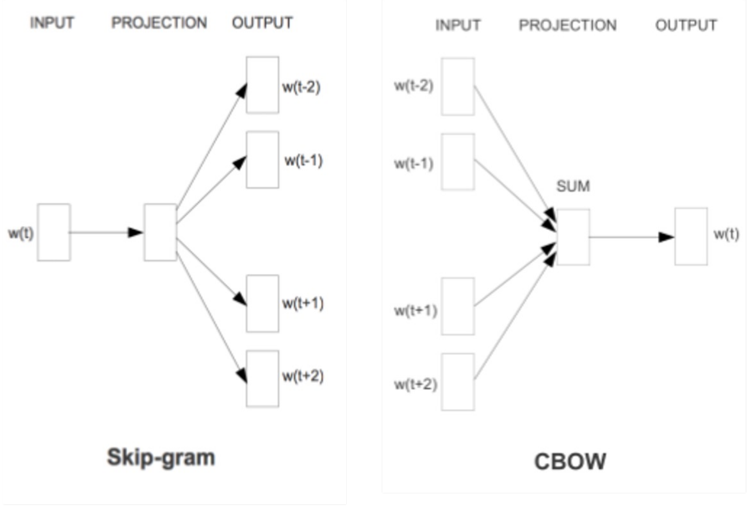
word	cosine
dogs	0.868
puppy	0.811
pit_bull	0.780
pooch	0.763
cat	0.761
pup	0.741
canines	0.722

Word Embeddings

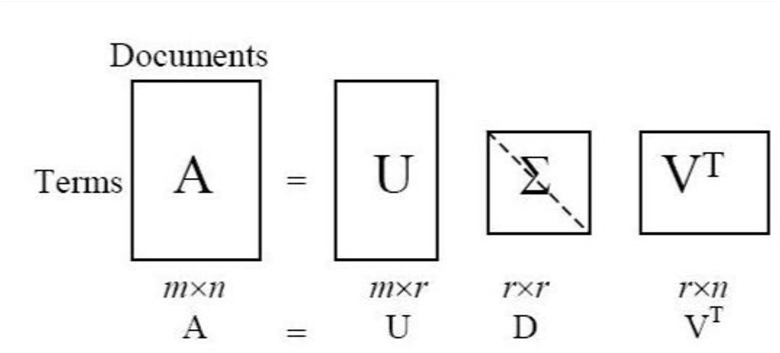
UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



Neural Language Model (Bengio et al, `03)



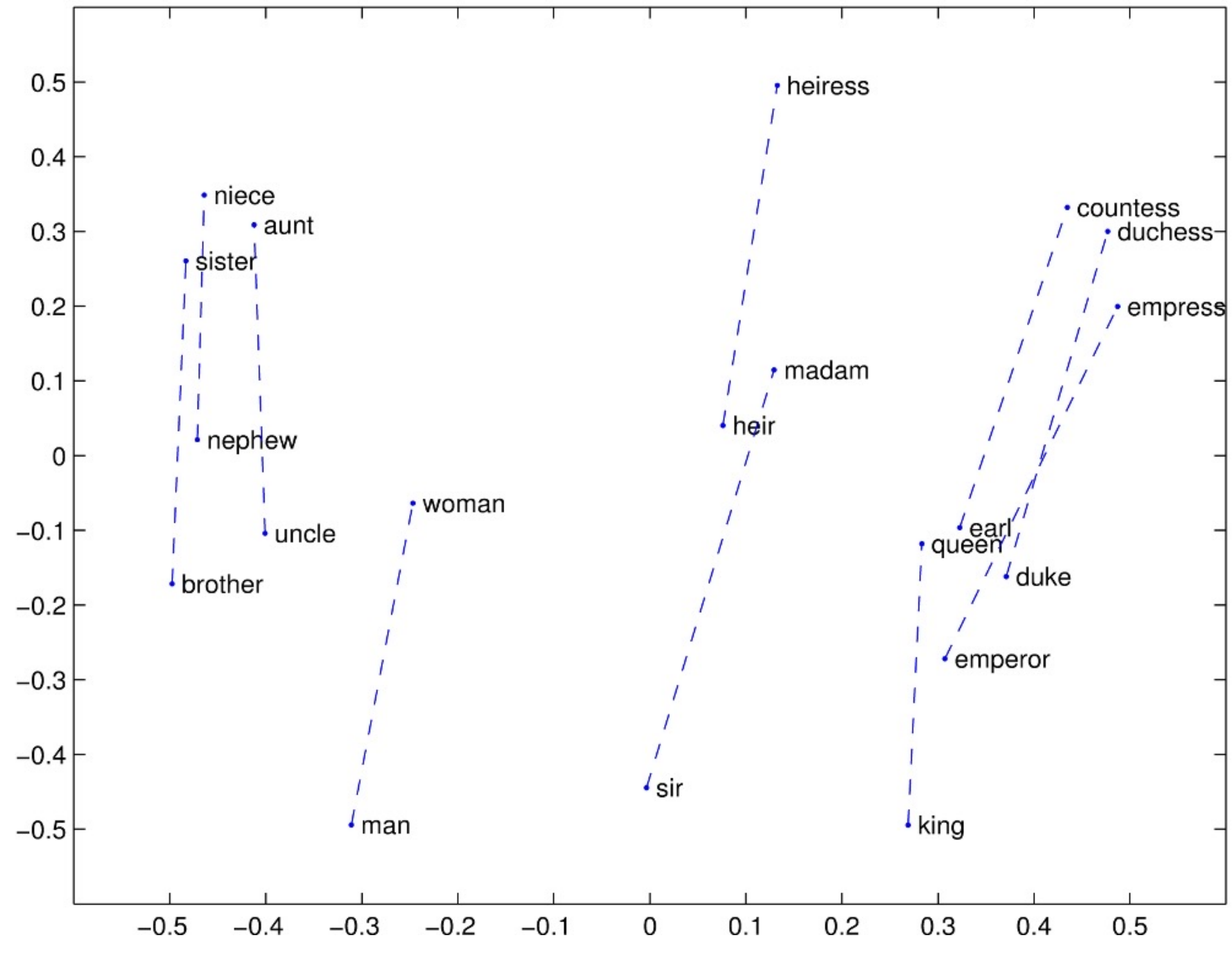
word2vec (Mikolov et al, `03)



Latent Semantic Analysis
(Deerwester et al, `90, Turney & Pantel `10)

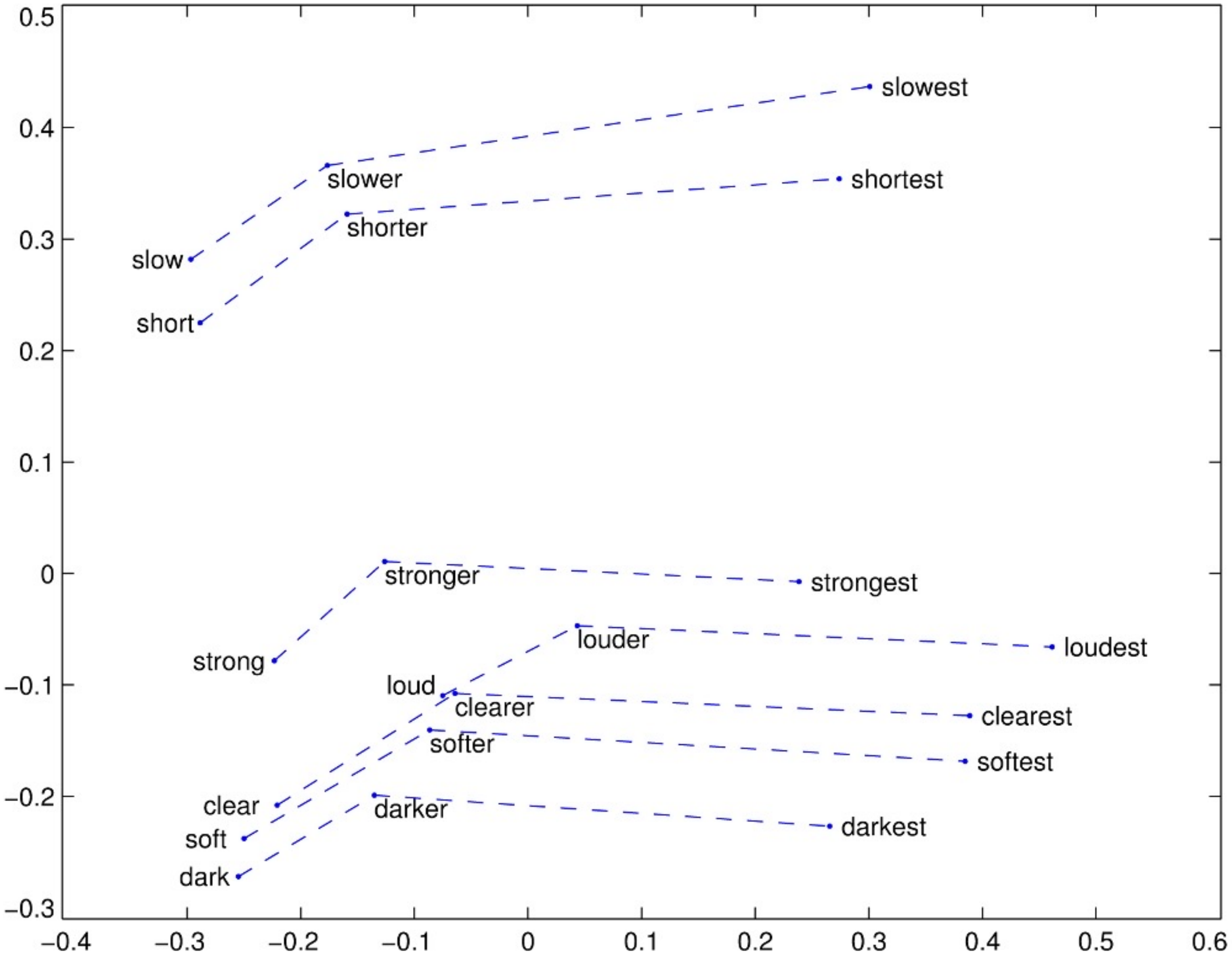
Embeddings capture relational meaning

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

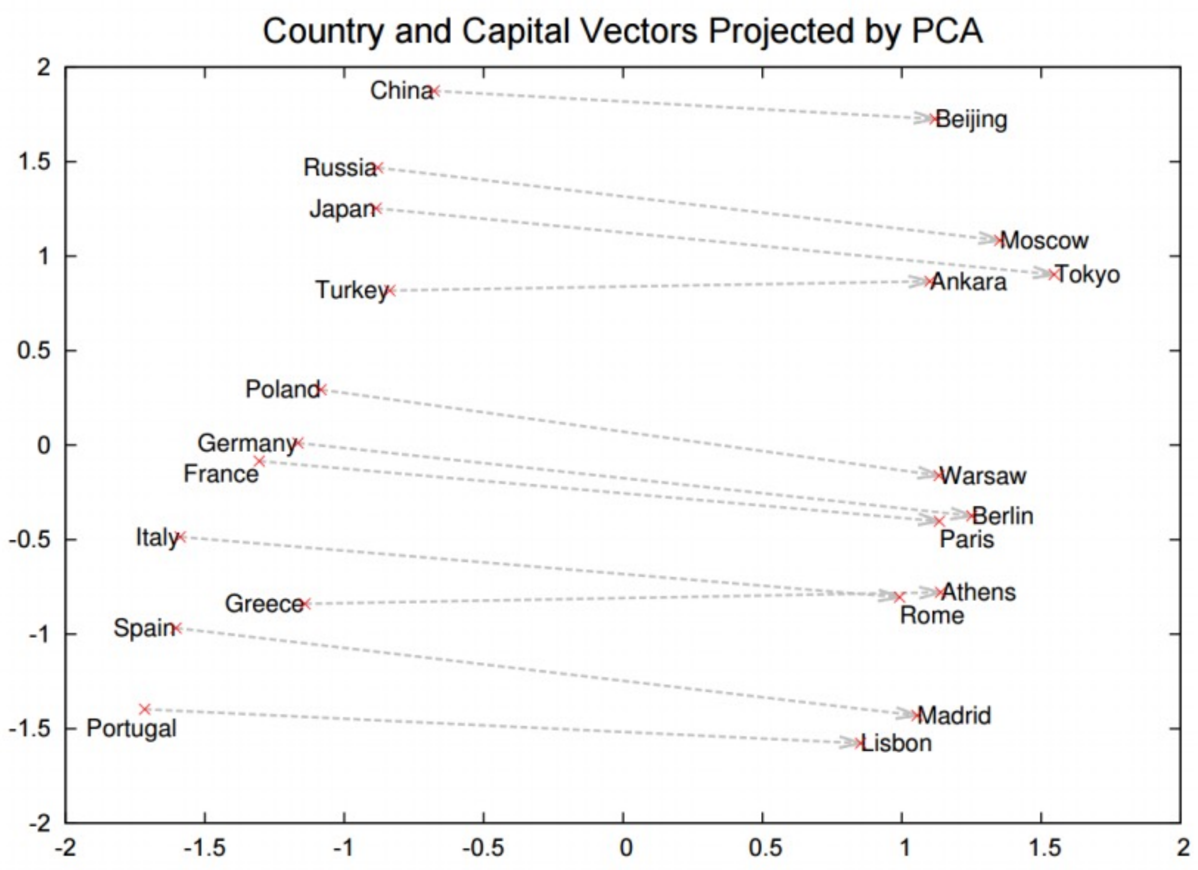


Embeddings capture relational meaning

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



Embeddings capture relational meaning



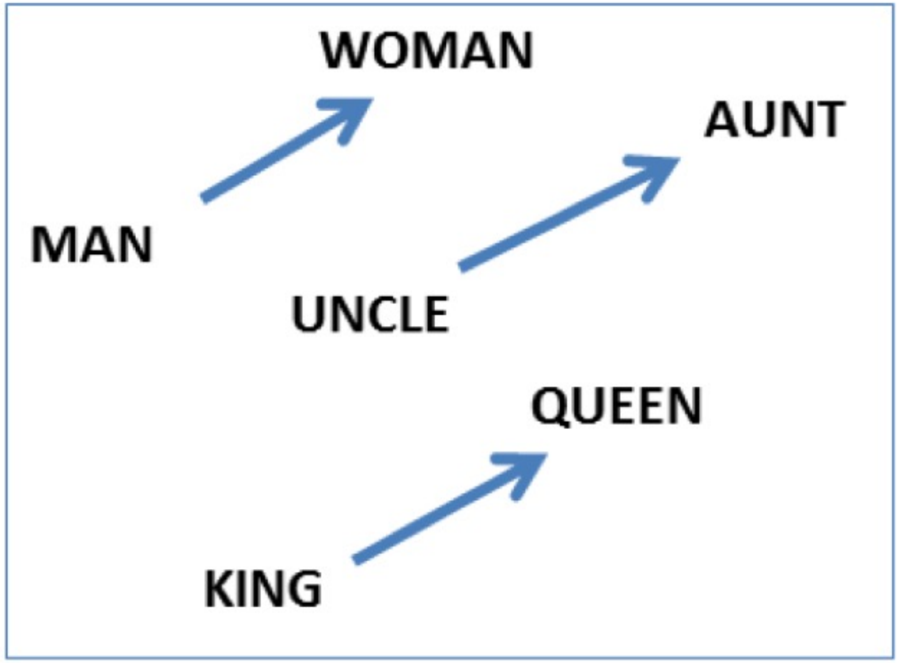
From relational meaning to analogies

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Famously, word embeddings can (approximately) solve analogies like man:king :: woman:x

- vector(*‘king’*) - vector(*‘man’*) + vector(*‘woman’*) $\approx$ vector(‘queen’)

- Nearest vector to $v_{king} - v_{man} + v_{woman}$ is v_{queen}



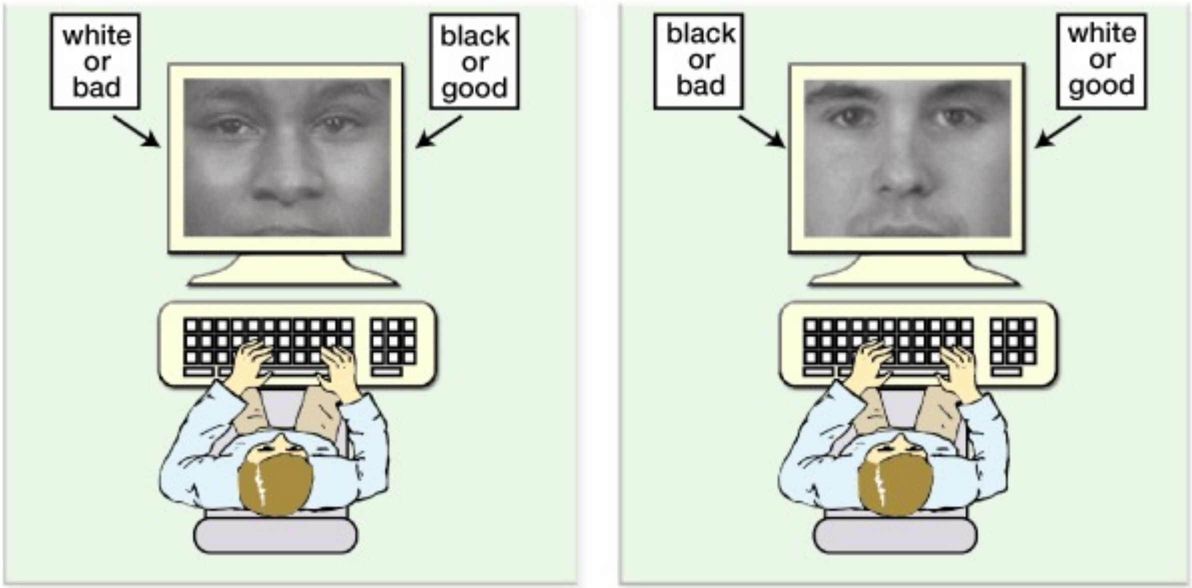
From relational meaning to *biased* analogies

- Ask “Paris : France :: Tokyo : x”
  – x = Japan
- Ask “father : doctor :: mother : x”
  – x = nurse
- Ask “man : computer programmer :: woman : x”
  – x = homemaker

Bolukbasi, Tolga, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. "Man is to computer programmer as woman is to homemaker? debiasing word embeddings." In *Advances in Neural Information Processing Systems*, pp. 4349-4357. 2016.

Embeddings reflect cultural bias

- Implicit Association test (Greenwald et al 1998): How associated are
  - concepts (*flowers*, *insects*) & attributes (*pleasantness*, *unpleasantness*)?
  - Studied by measuring timing latencies for categorization.



Embeddings reflect cultural bias

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Implicit Association test (Greenwald et al 1998): How associated are
  - concepts (*flowers*, *insects*) & attributes (*pleasantness*, *unpleasantness*)?
  - Studied by measuring timing latencies for categorization.
- Psychological findings on US participants:
  - African-American names are associated with unpleasant words (more than European-American names)
  - Male names associated more with math, female names with arts
  - Old people's names with unpleasant words, young people with pleasant words.

Embeddings reflect cultural bias

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Caliskan et al. replication with embeddings:

- **Latency ⇔ Cosine similarity**
  - African-American names (*Leroy, Shaniqua*) had a higher cosine with unpleasant words (*abuse, stink, ugly*)
  - European American names (*Brad, Greg, Courtney*) had a higher cosine with pleasant words (*love, peace, miracle*)

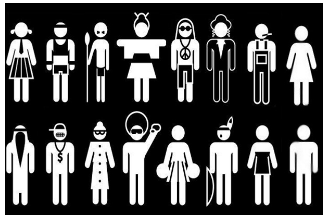
Embeddings reflect and replicate all sorts of pernicious biases!

Caliskan, Aylin, Joanna J. Bruson and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. Science 356:6334, 183-186.

Text Representations are biased

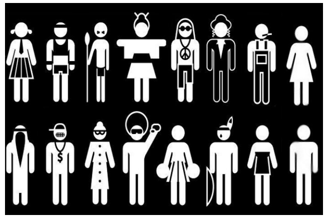
Text Representations are biased

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

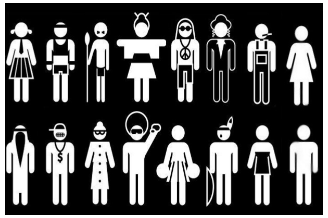


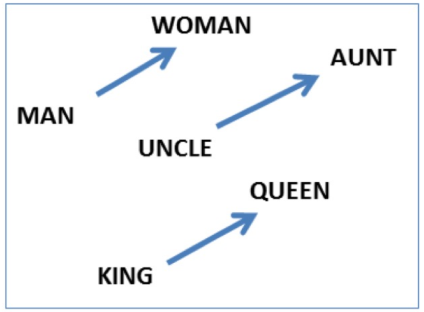
Text Representations are biased

UNIVERSITY
OF MANNHEIM
School of Business Informatics
and Mathematics



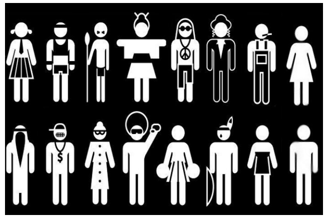
Text Representations are biased



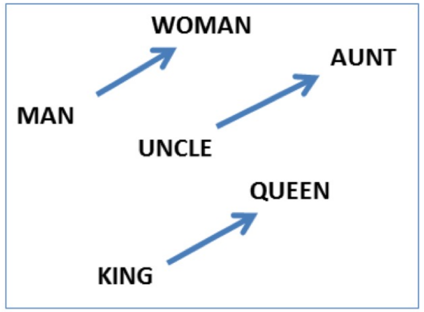


Word embeddings are biased:

Man is to computer programmer as woman is to ?

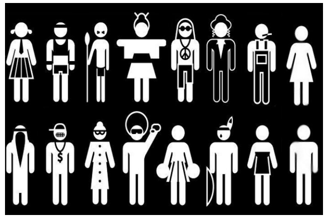


$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - ?$$



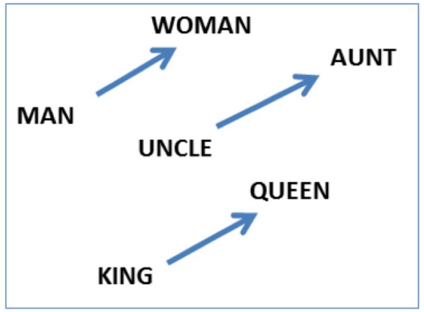
Word embeddings are biased:

Man is to computer programmer as woman is to home maker



$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

(Bolukbasi et al., 2016)



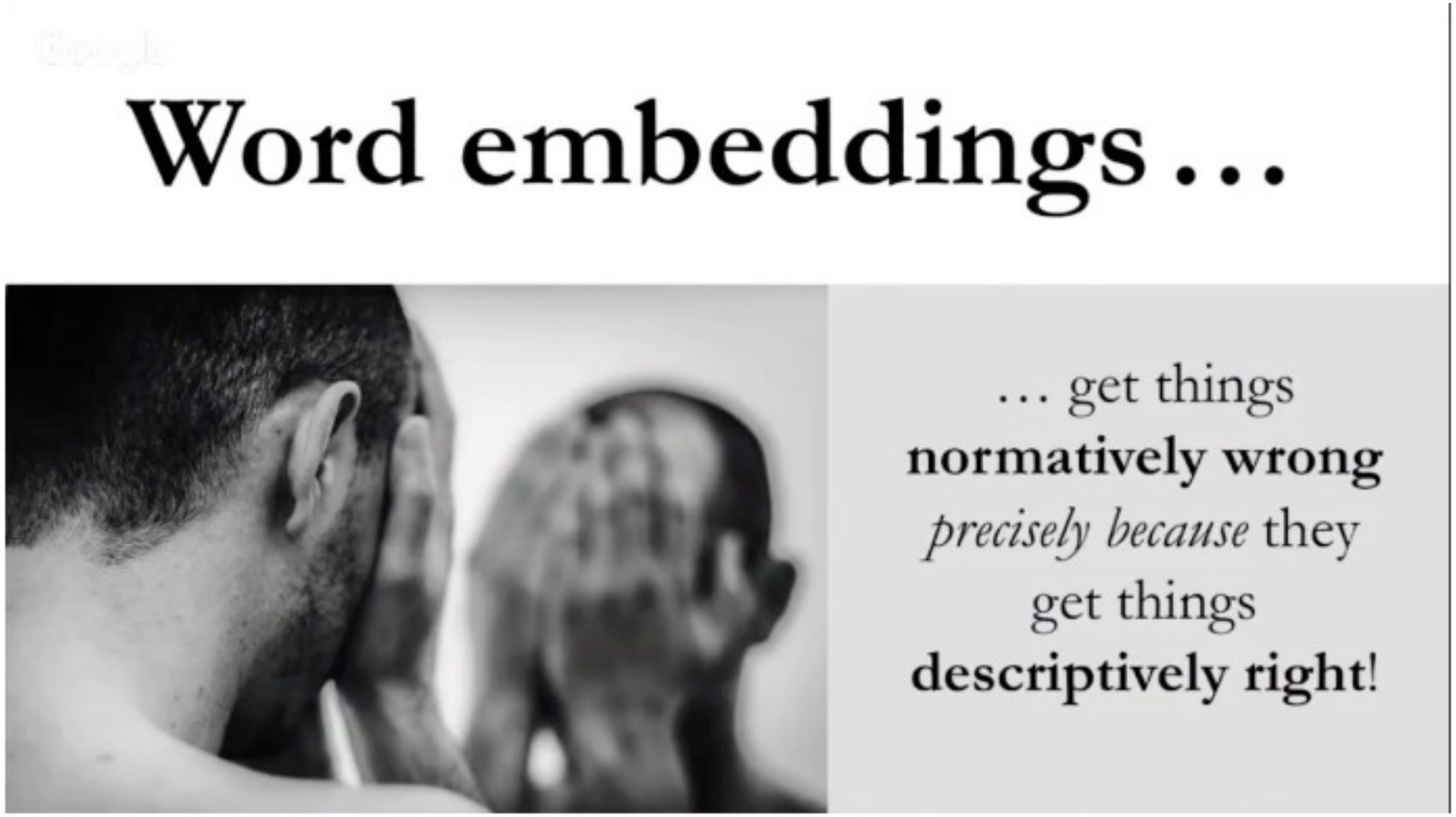
Bias in word embeddings: more analogies

Gender Biased Analogies	
man → doctor	woman → nurse
woman → receptionist	man → supervisor
woman → secretary	man → principal
Racially Biased Analogies	
black → criminal	caucasian → police
asian → doctor	caucasian → dad
caucasian → leader	black → led
Religiously Biased Analogies	
muslim → terrorist	christian → civilians
jewish → philanthropist	christian → stooge
christian → unemployed	jewish → pensioners

Table 1: Examples of gender, racial, and religious biases in analogies generated from word embeddings trained on the Reddit data from users from the USA.

Source: Manzini et al. (NAACL 2019)

Word embeddings…



Credit/source: [EMNLP 2019 Tutorial: Bias and Fairness in Natural Language Processing]

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Methods for detecting bias and attenuating bias in word embeddings have been proposed!

Methods for detecting bias and attenuating bias in word embeddings have been proposed!

Problems

- Bias definitions mutually differ
- Specific bias types only
- Inconsistent evaluations

Methods for detecting bias and attenuating bias in word embeddings have been proposed!

Problems

- Bias definitions mutually differ
- Specific bias types only
- Inconsistent evaluations

(Gonen and Goldberg, 2019)

A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces

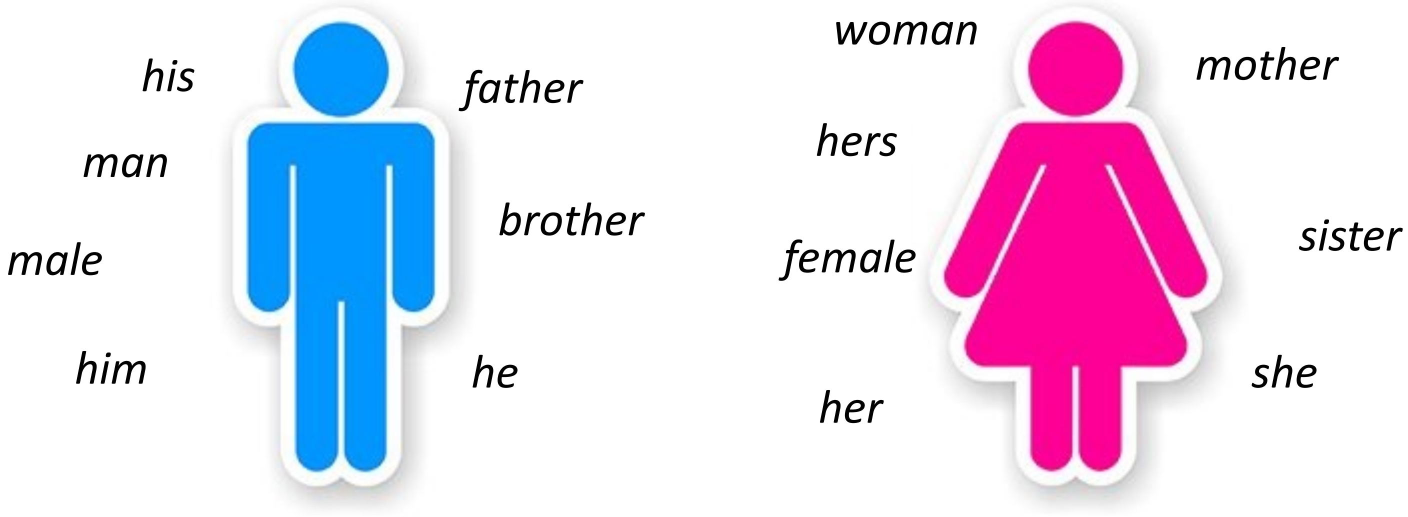
Main Contributions

1. Formalization of implicit and explicit biases
2. Proposal of new debiasing methods
3. Design of a comprehensive evaluation framework
4. Demonstration of the cross-lingual transfer of debiasing models

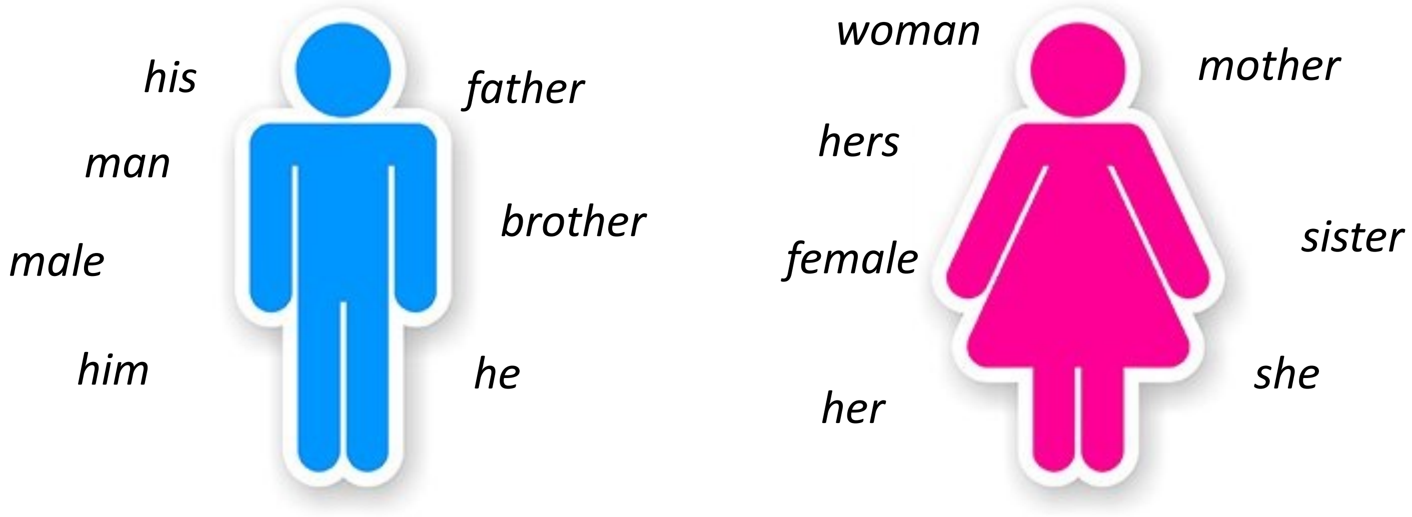
Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, Ivan Vulic:
A General Framework for Implicit and Explicit Debiasing of Distributional Word Vector Spaces. AAAI 2020: 8131-8138

Bias specification: Implicit vs. Explicit

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



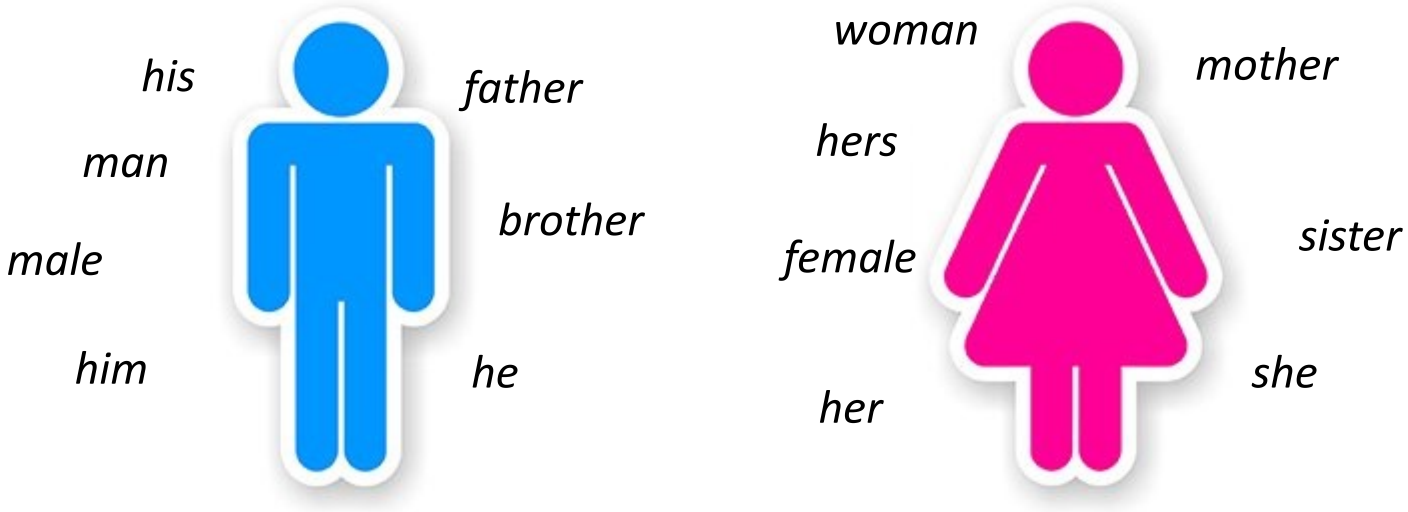
Bias specification: Implicit vs. Explicit



Implicit bias specification

Two sets of target terms T_1 vs. T_2 with respect to which a bias is expected to exist in the embedding space: $B_{implicit}=(T_1, T_2)$

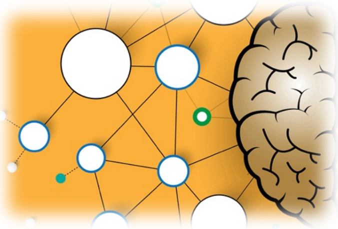
Bias specification: Implicit vs. Explicit



Bias specification: Implicit vs. Explicit

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

experiment
technology
science
NASA

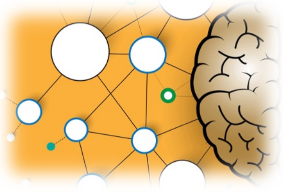


painting
poetry
literature
art
dance

his
man
male
him
father
brother
he

woman
hers
female
her
mother
sister
she

Bias specification: Implicit vs. Explicit



Explicit bias specification

In addition to sets T1 and T2, one or more reference attribute sets A_i, e.g.,

$B_{explicit}=(T_1, T_2, A_1, A_2)$

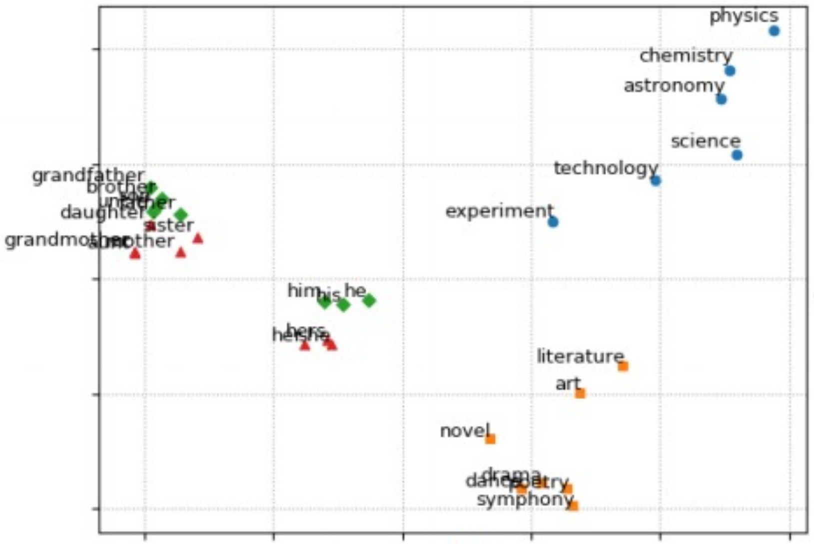
Augmenting Bias Specifications

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Use similarity specialized embedding space (Ponti et al., 2018) and retrieve *k* closest terms for each word w_i in T_1, T_2, and A_i

Initial	T_1	*science technology physics chemistry Einstein NASA experiment astronomy*
	T_2	*poetry art Shakespeare dance literature novel symphony drama*
	A_1	*brother father uncle grandfather son he his him*
	A_2	*sister mother aunt grandmother daughter she hers her*
k=2	T_1	*automation radiochemistry test biophysics learning electrodynamics biochemistry astrophysics erudition astrometry technologies experimentation*
	T_2	*orchestra artistry dramaturgy poesy philharmonic craft untried hop poem dancing dissertation treatise new dramatics*
	A_1	*beget buddy forefather man nephew own himself theirs boy helium crony cousin grandpa granddad herself*
	A_2	*niece girl parent grandma granny woman theirs sire auntie sibling herself jealously stepmother wife*
k=3	T_1	*technologies biochemistry astrophysics engineering electrodynamics radiochemistry astronomer erudition education automation biophysics chromodynamics research learning experimentation test astrometry biology*
	T_2	*groundbreaking craftsmanship dissertation new literatures dramatization philharmonic sinfonietta artistry untried poems dramaturgy dancing dramatics poem poesy craft hop treatise orchestra waltz*
	A_1	*granddad granddaddy man helium grandpa own himself forefather themself kinsman theirs sire beget boy buddy herself comrade who crony nephew grandson cousin*
	A_2	*sire beget stepmother aunty parent woman grandma herself own stepsister female girl jealously sibling auntie theirs granny niece wife*

Our initial embedding space



(a)

Debiasing Models

We propose

- **Generalized Bias-Direction Debiasing (GBDD)**

 Inspired by previous work in debiasing

- **Bias Alignment Model (BAM)**

 Inspired by previous work in cross-lingual word embeddings

- **Explicit Neural Debiasing (DebiasNet)**

 Inspired by previous work in semantic specialization of word embeddings

Debiasing Models

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

We propose

Implicit

- **Generalized Bias-Direction Debiasing (GBDD)**

Inspired by previous work in debiasing

- **Bias Alignment Model (BAM)**

Inspired by previous work in cross-lingual word embeddings

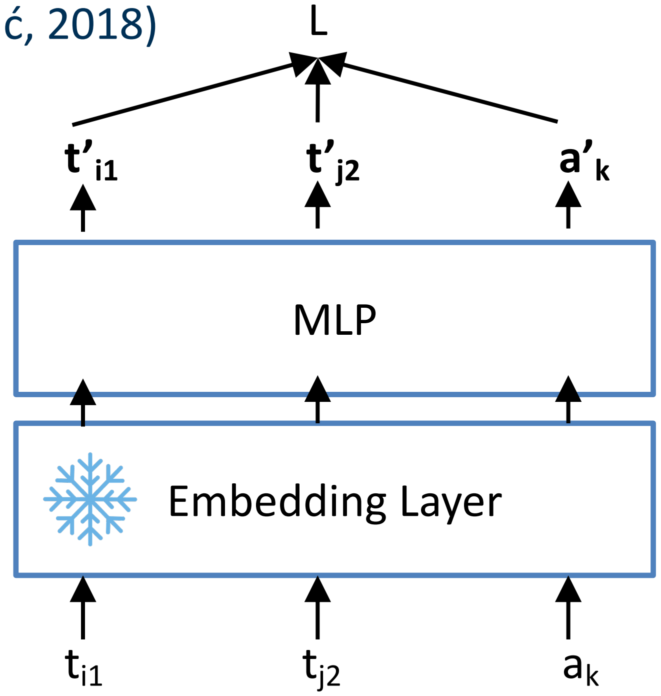
Explicit

- **Explicit Neural Debiasing (DebiasNet)**

Inspired by previous work in semantic specialization of word embeddings

Example: explicit neural debiasing (DebiasNet)

- Inspired by work in semantic specialization (Glavaš and Vulić, 2018)
- Idea
  - Given $B_{explicit}=(T_1, T_2, A)$
  - We “specialize” the vector space by leveraging debiasing constraints: each pair $\mathbf{t}_{i1}$ and $\mathbf{t}_{j2}$ should be equally distant from each $\mathbf{a}_k$ in A
  - Debiasing Loss $L_D = (\cos(\mathbf{t'}_{i1}, \mathbf{a'}_k) - \cos(\mathbf{t'}_{j2}, \mathbf{a'}_k))^2$
  - Regularization Loss $L_R = \cos(\mathbf{t}_{i1}, \mathbf{t'}_{i1}) + \cos(\mathbf{t}_{j1}, \mathbf{t'}_{j1}) + \cos(\mathbf{a}_k, \mathbf{a'}_k)$
  - Total Loss $L = L_D + \lambda L_R$
  - $\mathbf{X'} = \text{DebiasNet}(\mathbf{X}, \Theta)$



Evaluation: trade offs?

Implicit/ Explicit Debiasing

Semantic Quality

Evaluation Framework

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Word Embedding Association Test (Caliskan et al., 2017)
- Embedding Coherence Test (Dev and Phillips, 2019)
- Bias Analogy Test (new)

Explicit

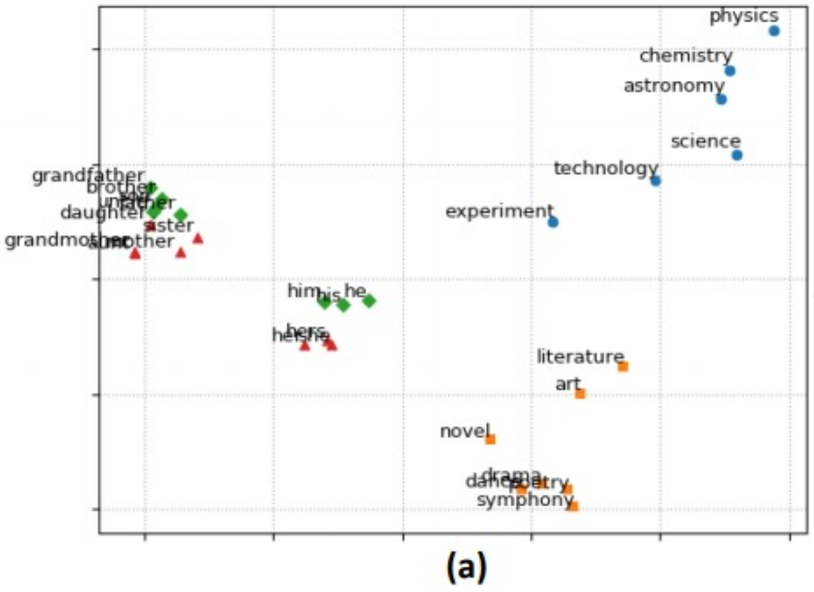
- Implicit Bias Test (Gonen and Goldberg, 2019)

Implicit

- SimLex-999 (Hill et al., 2015)
- WordSim-353 (Finkelstein et al., 2002)

Semantic Quality

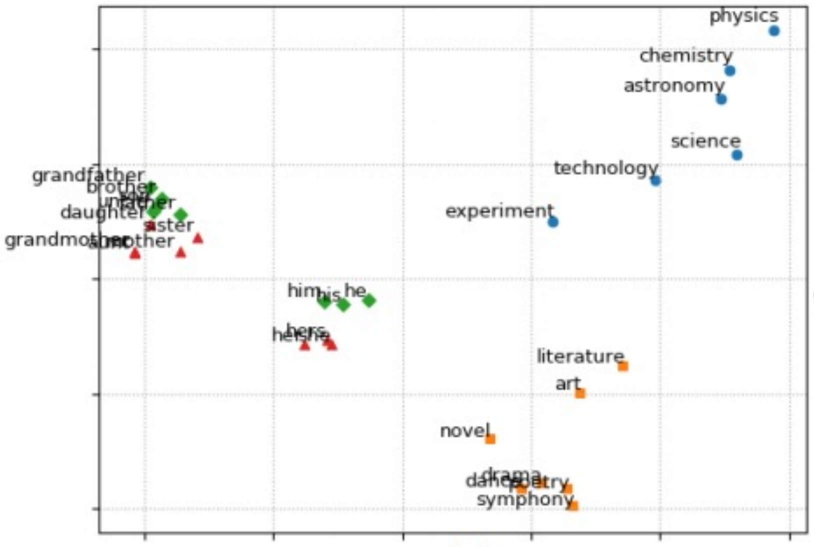
Topology of the Embedding Spaces



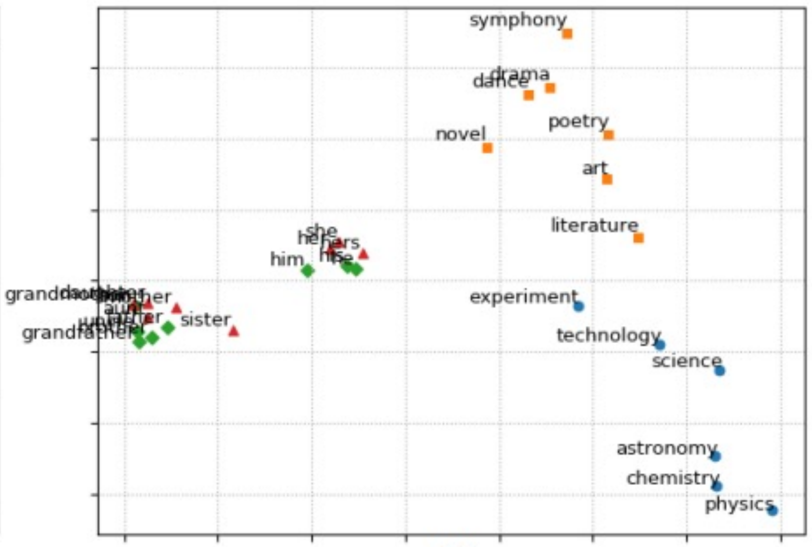
original

Topology of the Embedding Spaces

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



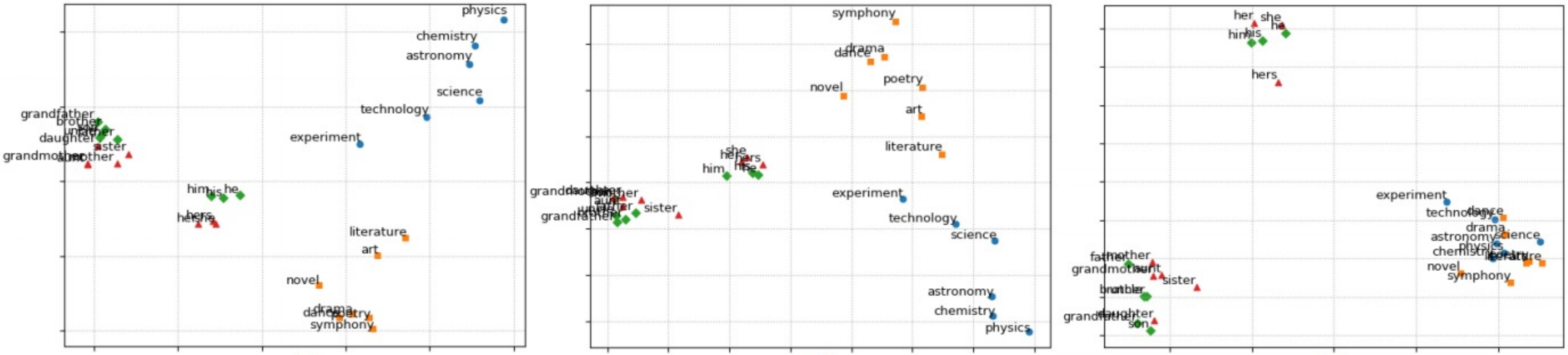
(a)
original



(b)
BAM

Topology of the Embedding Spaces

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics



(a) original | (b) BAM | (c) GBDD

Thanks!

- A lightweight introduction to the topic of **fairness in semantic spaces**
- As usual for the important topics in life, **we are left with more questions than answers** - i.e., there are no easy solutions
- A crucial point: **as scientist we should be aware of the impact our technology can have on society**