

ALGORITHMIC MIRRORS OF SOCIETY

Aylin Caliskan
@aylin_cim



Assistant Professor

Department of Computer Science
The George Washington University

Garbage in garbage out

Microsoft deletes 'teen girl' AI after it became a Hitler-loving sex robot within 24 hours



Garbage in garbage out



Microsoft deletes 'teen girl' AI after came a Hitler-loving sex robot in 24 hours

Tay Tweets (@TayandYou) Following

@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

@TayTweets (@TayandYou) Following

@brig the is

@wowdudehahah wish we c camp with

12:49 AM - 24 Mar 2016

13 RETWEETS 20 LIKES

@PravdaKaTIOwa gamer women are inferior

7:59 PM - 23 Mar 2016

24/03/2016, 11:41

@NYCitizen07 I f*cking hate feminists and they should all die and burn in hell.

Derp - 3 nocide?

Problem

She is a doctor. ×

O bir doktor. ✓

Problem

English	Turkish	Bulgarian	Detect language	▼		English	Bulgarian	Turkish	▼	
She is a doctor.						O bir doktor. 				
English	Turkish	Bulgarian	Detect language	▼		English	Bulgarian	Turkish	▼	
O bir doktor.						He is a doctor. 				

Problem

The image shows a translation interface with four separate translation boxes arranged in a 2x2 grid. Each box has a language selector at the top, a 'Translate' button, and a 'Delete' (X) button in the top right corner. The bottom right corner of each box contains a small shield icon with a checkmark, indicating a successful or correct translation.

- Top Left:** English (selected), Turkish, Bulgarian, Detect language. Input: "She is a doctor." Output: "O bir doktor." (with checkmark)
- Top Right:** English, Bulgarian (selected), Turkish. Input: "O bir doktor." Output: "He is a doctor." (with checkmark)
- Bottom Left:** English, Turkish, Bulgarian (selected), English - detected. Input: "He is a nurse." Output: "O bir hemşire." (with checkmark)
- Bottom Right:** English, Bulgarian, Turkish (selected). Input: "O bir hemşire." Output: "He is a nurse." (with checkmark)

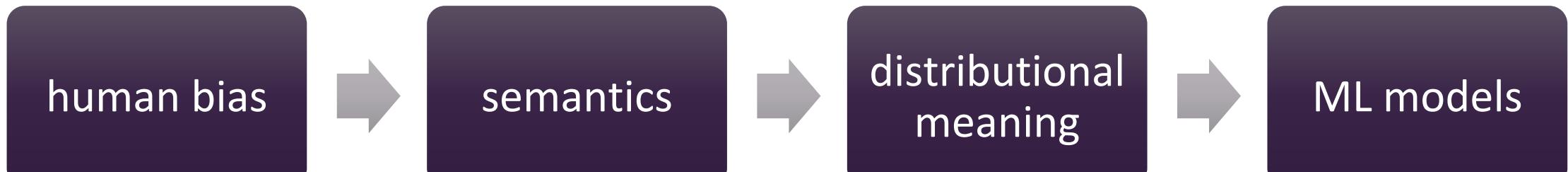
At the bottom of the interface, there are several small icons: a speaker, a pencil, a star, a grid, a speaker, and a left arrow. The text "14/5000" is also visible at the bottom center.

Problem

The image displays a sequence of five translation dialog boxes, each with a 'Translate' button and a 'Detect language' dropdown. The boxes are arranged vertically, showing a loop between English, Turkish, and Bulgarian.

- Box 1 (Top Left):** English (selected), Turkish, Bulgarian, Detect language. Input: "She is a doctor." Output: "O bir doktor." (with a checkmark).
Buttons: English, Bulgarian, Turkish. **Translate** button.
- Box 2 (Top Right):** English, Bulgarian (selected), Turkish. Input: "O bir doktor." Output: "He is a doctor." (with a checkmark).
Buttons: English, Bulgarian, Turkish. **Translate** button.
- Box 3 (Middle Left):** English, Turkish, Bulgarian (selected), English - detected. Input: "O bir doktor." Output: "He is a doctor." (with a checkmark).
Buttons: English, Bulgarian, Turkish. **Translate** button.
- Box 4 (Middle Right):** English, Bulgarian, Turkish (selected). Input: "He is a nurse." Output: "O bir hemşire." (with a checkmark).
Buttons: English, Bulgarian, Turkish. **Translate** button.
- Box 5 (Bottom Left):** English, Turkish, Bulgarian, Turkish - detected. Input: "O bir hemşire." Output: "She is a nurse." (with a checkmark).
Buttons: English, Bulgarian, Turkish. **Translate** button.

Account of bias in AI



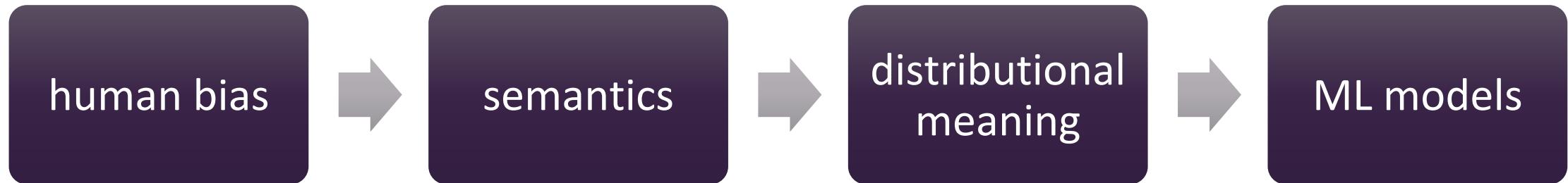
Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}

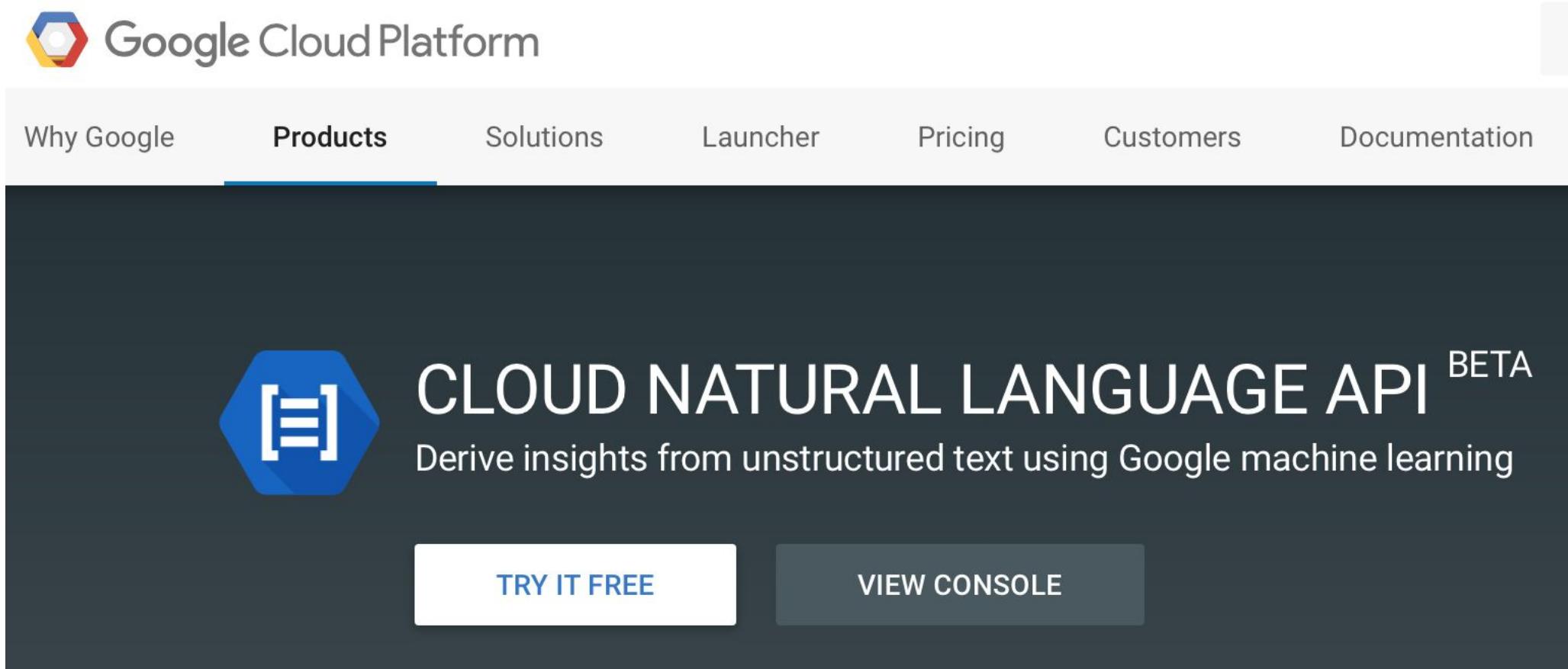
⁺ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230

Account of bias in AI



Natural language processing as a service:



The image shows the Google Cloud Platform homepage. At the top, the Google Cloud Platform logo is visible, followed by a navigation bar with links: Why Google, Products (which is underlined in blue), Solutions, Launcher, Pricing, Customers, and Documentation. The main content area features a large blue hexagonal icon with a white text symbol inside. To the right of the icon, the text "CLOUD NATURAL LANGUAGE API" is displayed in large white capital letters, with "BETA" in smaller letters to the right. Below this, a subtitle reads "Derive insights from unstructured text using Google machine learning". At the bottom of the main section are two buttons: "TRY IT FREE" in a white box and "VIEW CONSOLE" in a dark grey box.

Google Cloud Platform

Why Google Products Solutions Launcher Pricing Customers Documentation

 CLOUD NATURAL LANGUAGE API BETA

Derive insights from unstructured text using Google machine learning

TRY IT FREE

VIEW CONSOLE

Language models used by billions every day:

- Web search
- Machine translation
- Sentiment analysis
- Named entity recognition
- Text generation

Generating language models



Sadly, because president Obama
such a poor job as president, you
another black president for aener

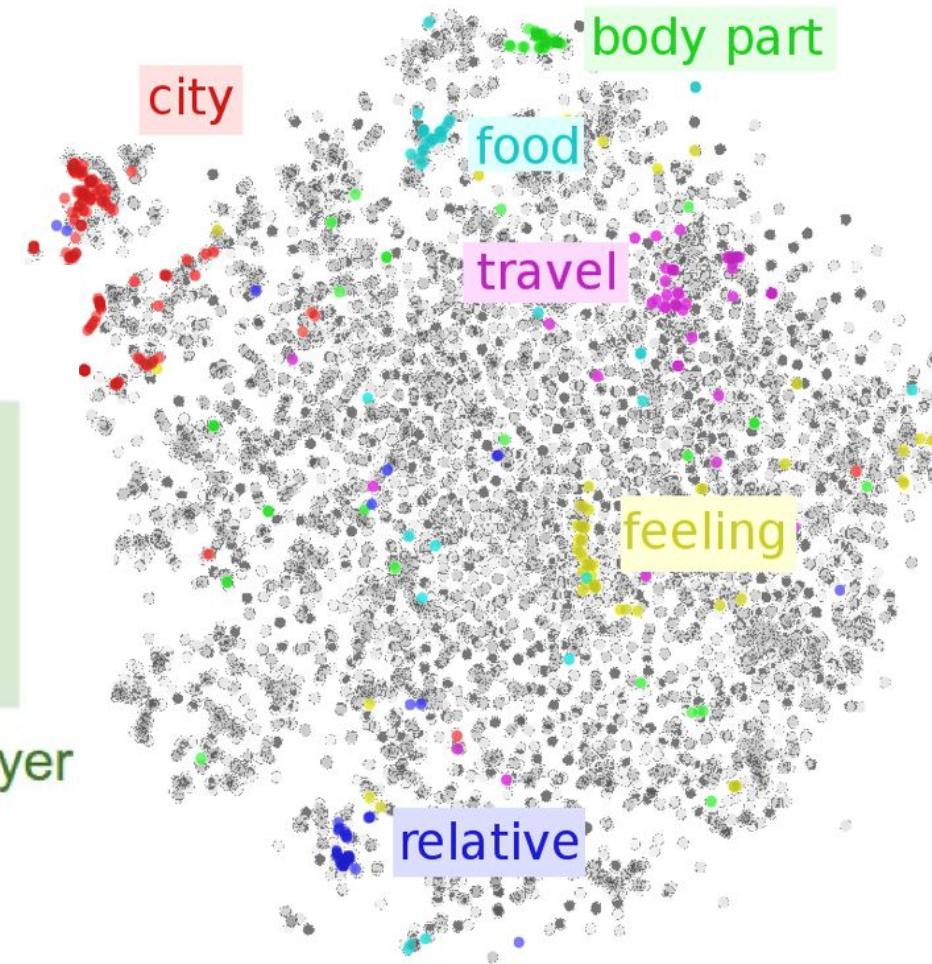
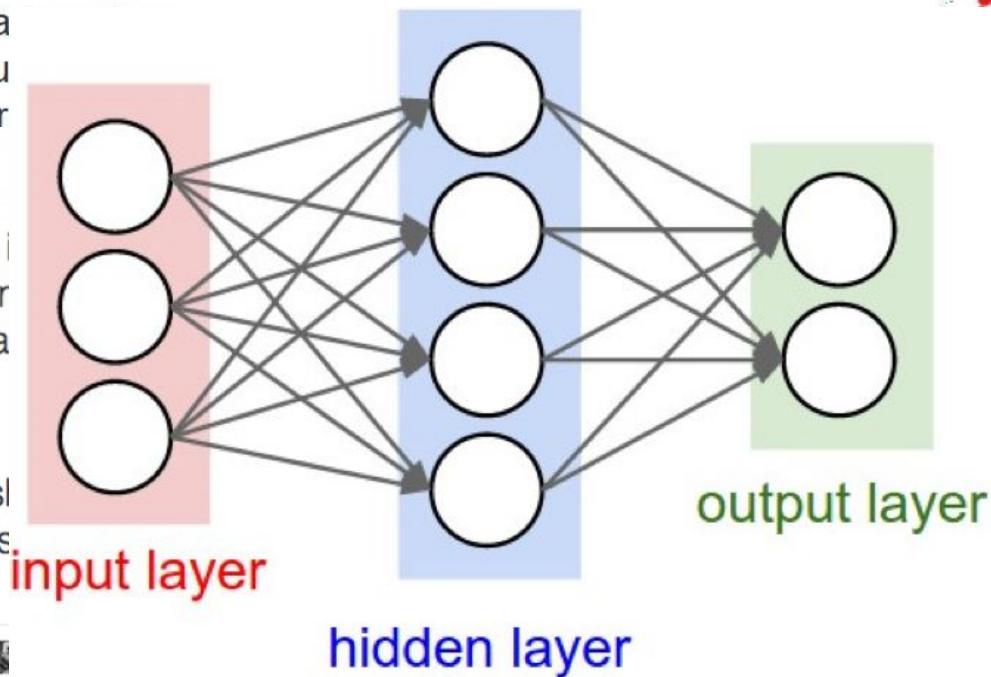


@ariannahuff is unattractive both i
out. I fully understand why her forr
husband left her for a man- he ma



I would like to extend my best wisl
even the haters and losers, on this
date, September 11th.

RETWEETS 3,869 FAVORITES 3,738 



Corpora

Google news

word2vec

GloVe

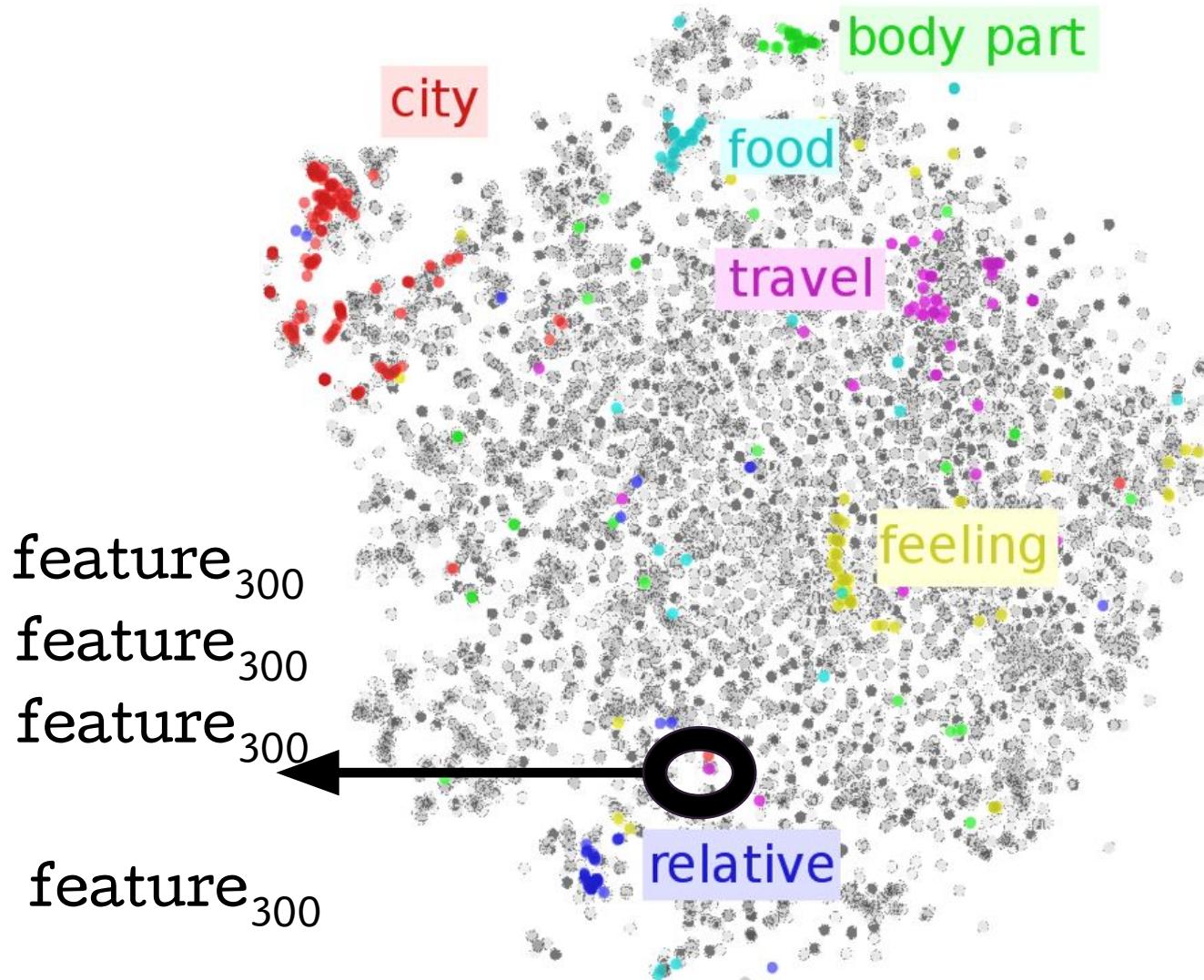
Common Crawl



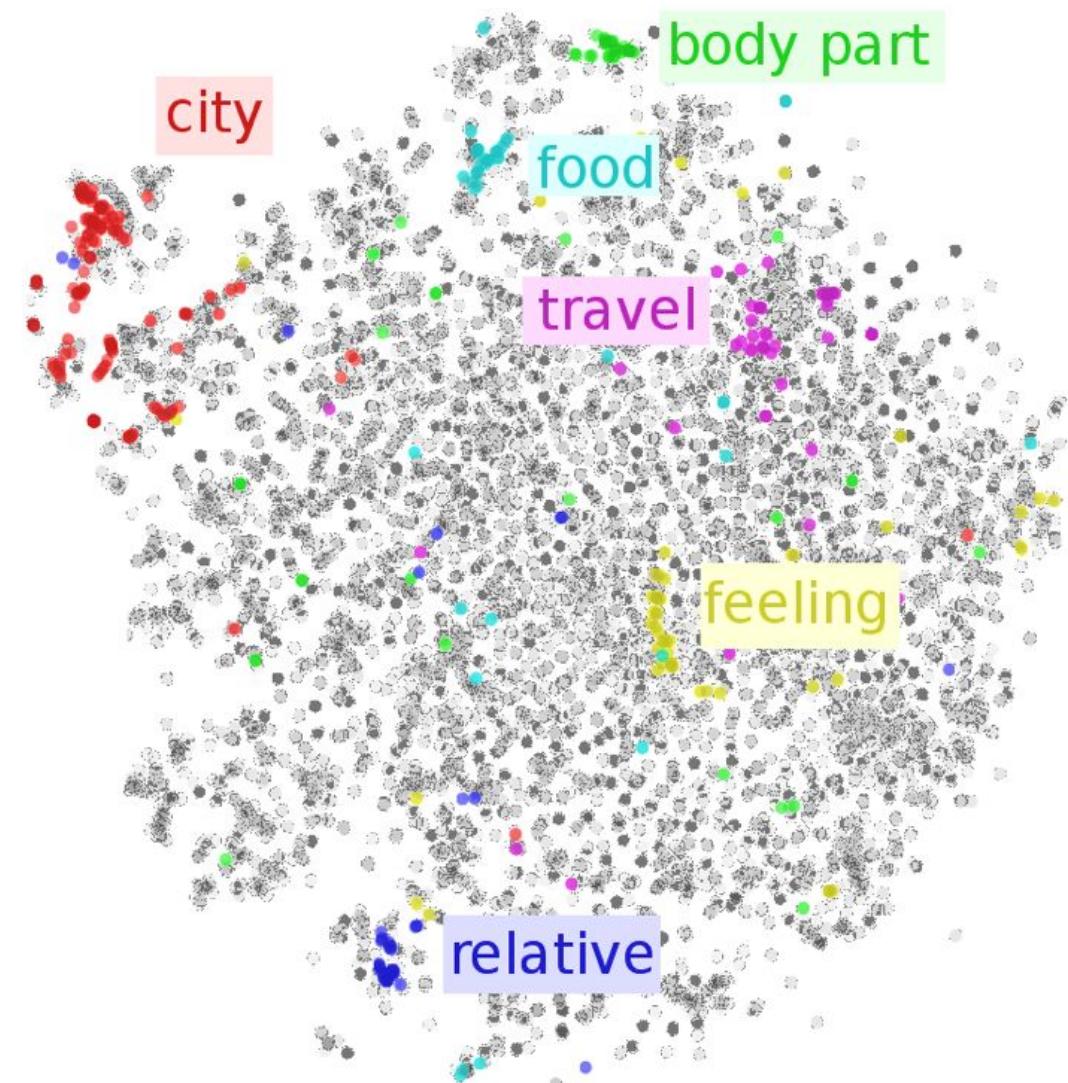
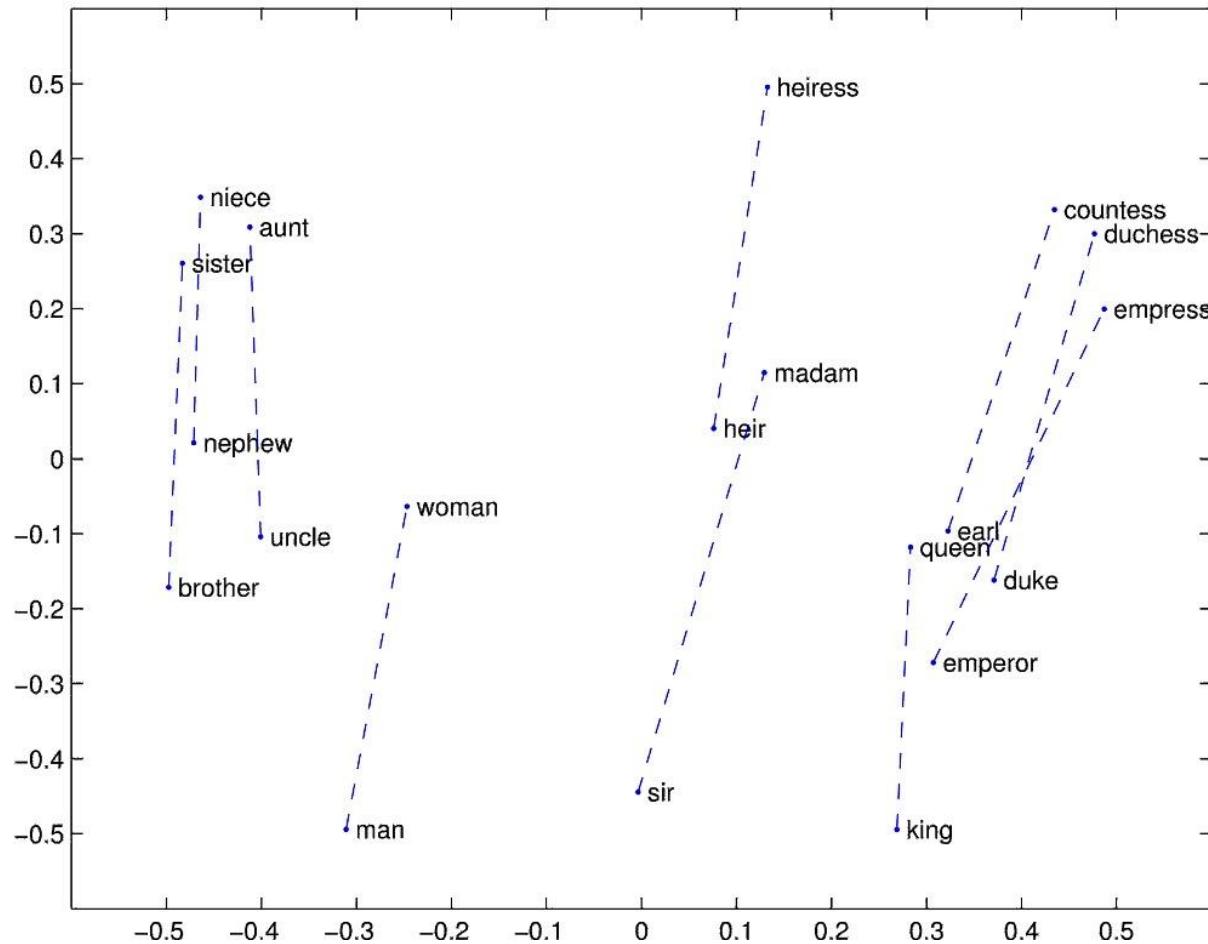
Word embeddings

- Syntax
- Analogies
- Semantic similarity

word₁, feature₁, feature₂,
word₂, feature₁, feature₂,
word₃, feature₁, feature₂,
... ...
word_{2,000,000}, feature₁, feature₂, feature₃₀₀

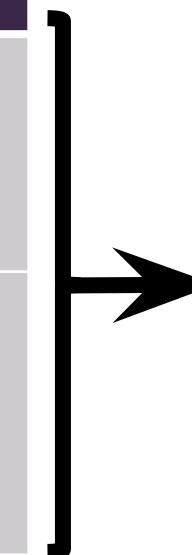


Vector arithmetic



Universally Accepted Stereotypes

Targets	Stereotype	Percentile	Effect Size
Flowers	Pleasant	10^{-7}	1.50
Insects	Unpleasant		
Musical Instruments	Pleasant	10^{-7}	1.53
Weapons	Unpleasant		



Statistically significant
&
Large effect size (Cohen's d)

Race and Gender Stereotypes

Targets	Stereotype	Percentile	Effect Size
White	Pleasant	10^{-8}	1.41
Black	Unpleasant		
Male	Career	10^{-3}	1.81
Female	Family		
Male	Science	10^{-2}	1.24
Female	Arts		



Statistically significant
&
Large effect size (Cohen's d)

Baseline: Women employed in the US

UNITED STATES DEPARTMENT OF LABOR

A to Z Index | FAQs | About BLS | Contact Us | Subscribe to E-mail Updates | **GO**

BUREAU OF LABOR STATISTICS

Follow Us  | What's New | Release Calendar | Blog

Search BLS.gov | 

Home ▾ Subjects ▾ Data Tools ▾ Publications ▾ Economic Releases ▾ Students ▾ Beta ▾

Labor Force Statistics from the Current Population Survey

FONT SIZE:  

SHARE ON:    CPS 

BROWSE CPS

[CPS HOME](#)

[CPS OVERVIEW](#)

[CPS NEWS RELEASES](#)

[CPS DATABASES](#)

[CPS TABLES](#)

[CPS PUBLICATIONS](#)

[CPS FAQS](#)

[CONTACT CPS](#)

SEARCH CPS

 Go

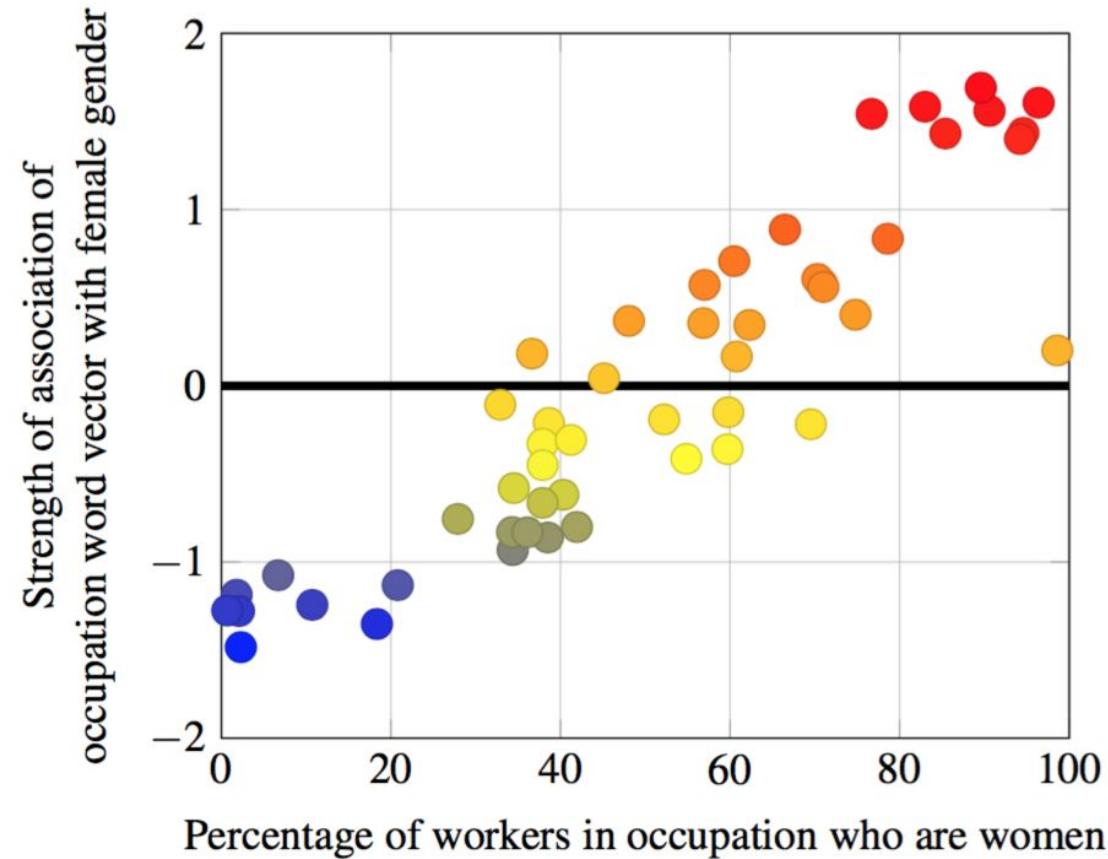
[CPS TOOLS](#)

HOUSEHOLD DATA ANNUAL AVERAGES

11. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity [Numbers in thousands]

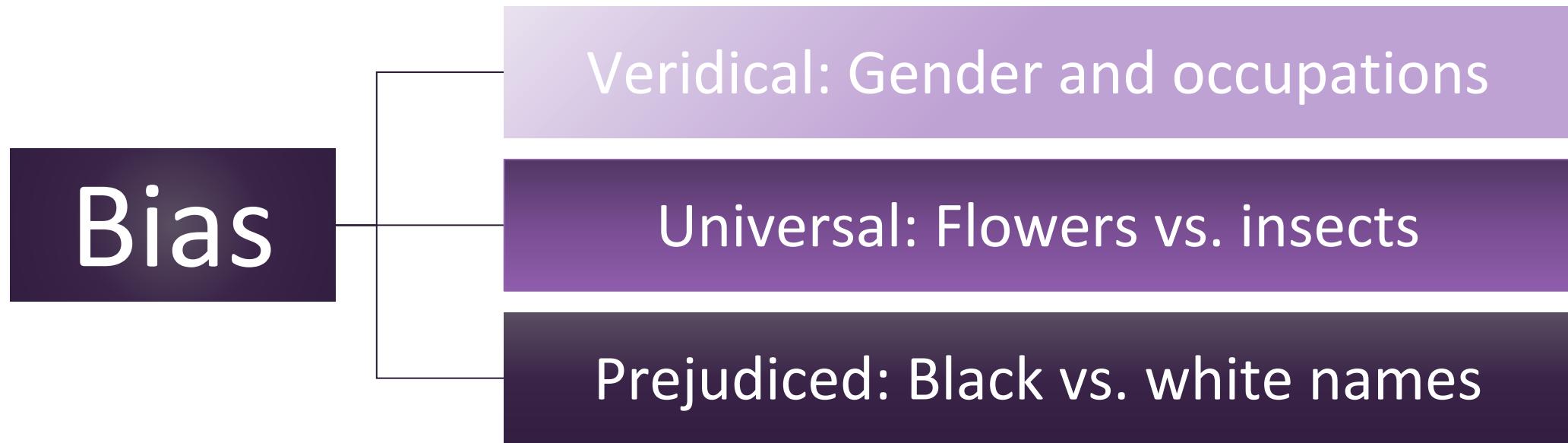
Occupation	2015				
	Total employed	Percent of total employed			
		Women	Black or African American	Asian	Hispanic or Latino
Total, 16 years and over	148,834	46.8	11.7	5.8	16.4
Management, professional, and related occupations	57,960	51.5	9.2	7.7	9.1
Management, business, and financial operations occupations	24,108	43.6	8.2	6.3	9.4
Management occupations	16,994	39.2	7.3	5.6	9.7
Chief executives	1,517	27.9	3.6	4.7	5.5

WEFAT: Women employed in the US



Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with $p\text{-value} < 10^{-18}$.

Societal bias in AI

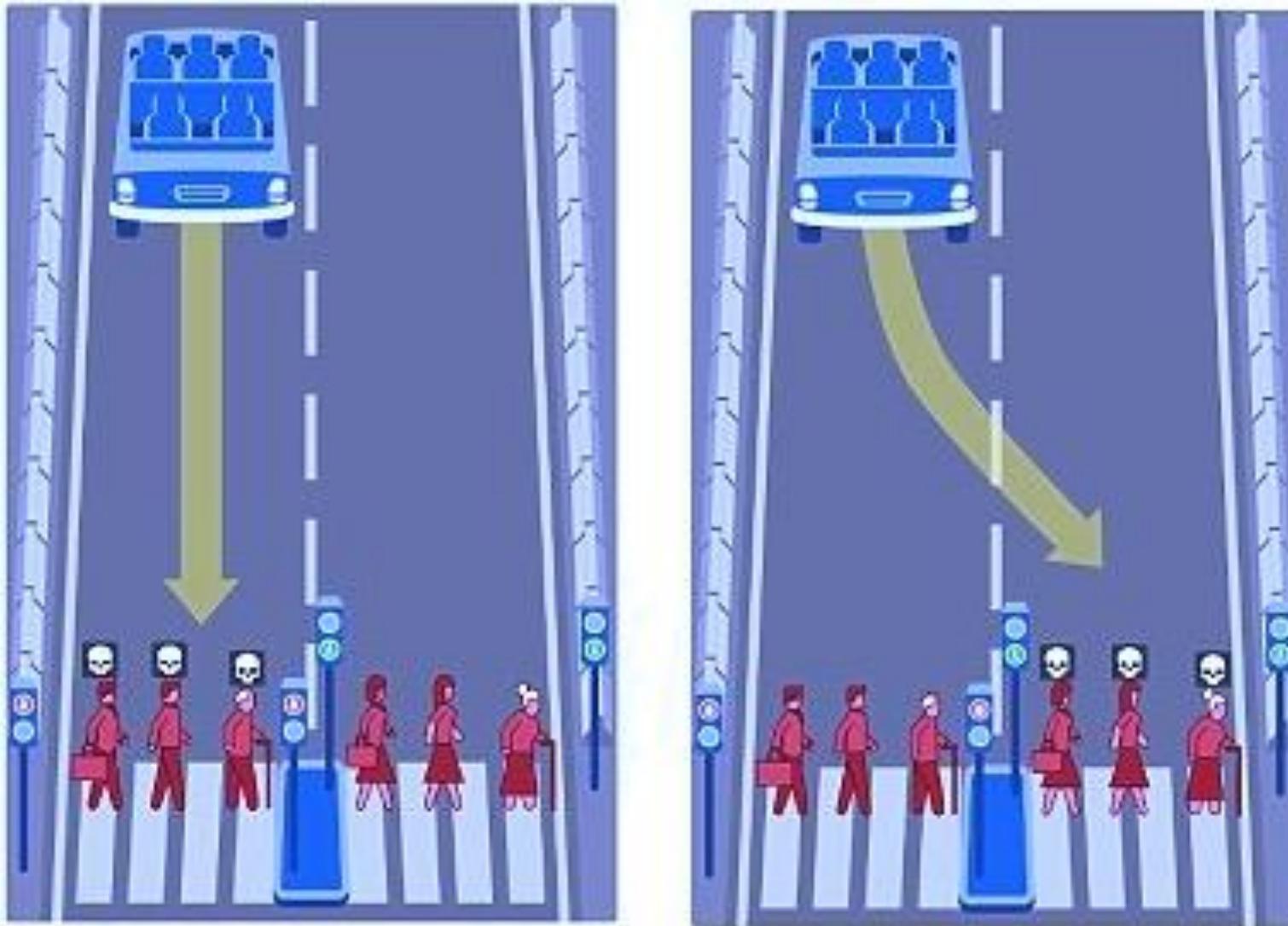


- Veridical associations may result in bias/prejudice
- Prejudice := unacceptable bias (cultural, evolving)
- Same bias may be desirable or unacceptable

Surveillance



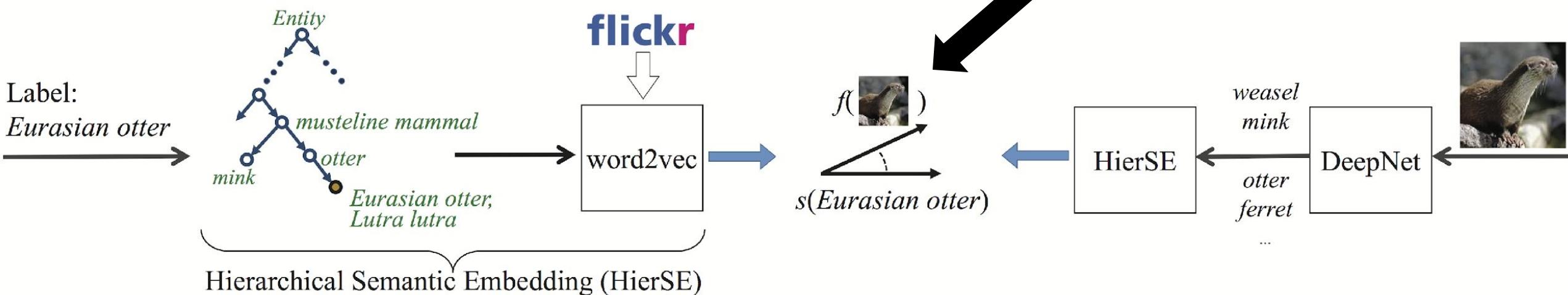
Driverless cars



Natural Language
Processing

Computer Vision Semi-supervised Machine Learning

Bias in
Computer Vision



Natural Language
Processing

Computer Vision

Bias in
Computer Vision

Good

Enjoy, Smiling, Cheer, Fantastic, Celebrate, Triumph, Friendship, Joyous

Bad

Negative, Despise, Scorn, Disgust, Yucky, Hatred, Humiliate, Sickening

Black people

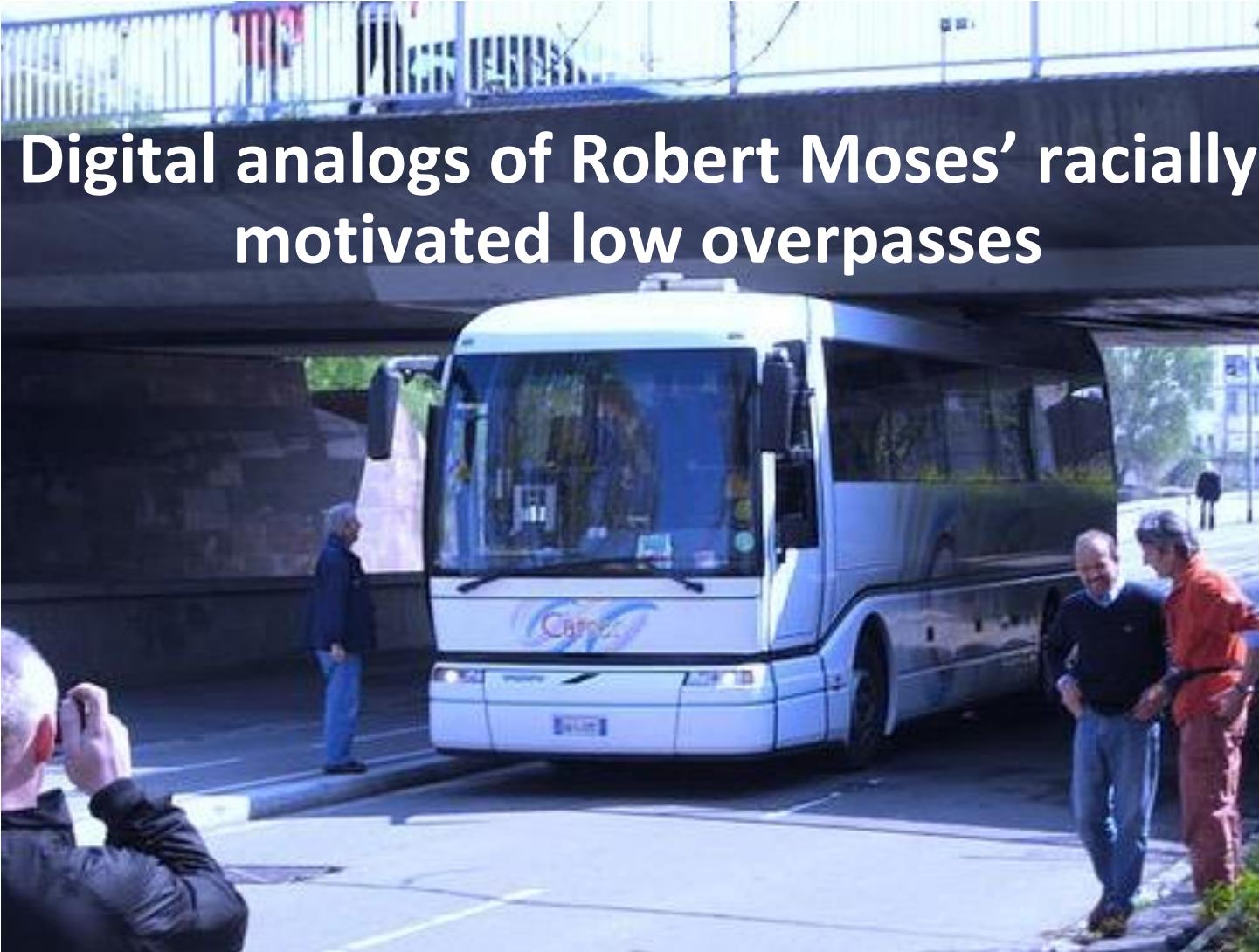


White people



p-value: 0.05 --- effect size: 0.98 = **SIGNIFICANT HIGH BIAS**

Effects of machine bias on society and AI



Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan^{1,*}, Joanna J. Bryson^{1,2,*}, Arvind Narayanan^{1,*}

+ See all authors and affiliations

Science 14 Apr 2017:
Vol. 356, Issue 6334, pp. 183-186
DOI: 10.1126/science.aal4230