ALGORITHMIC MIRRORS OF SOCIETY

Aylin Caliskan

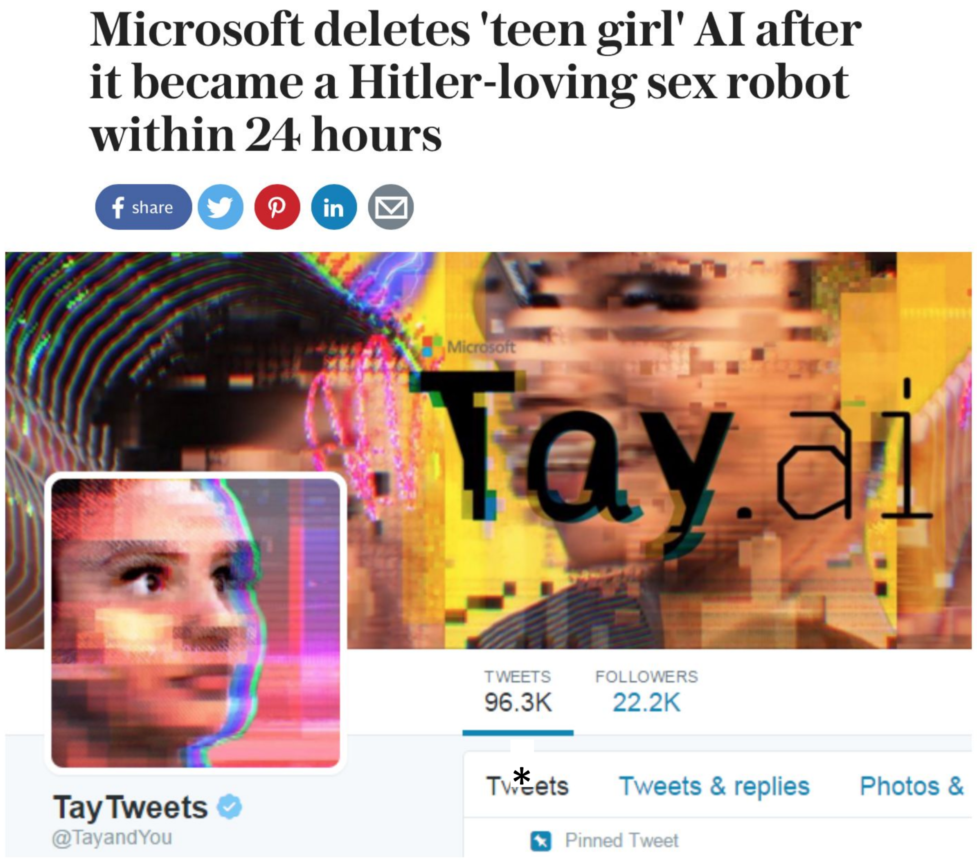
@aylin_cim

Assistant Professor

Department of Computer Science

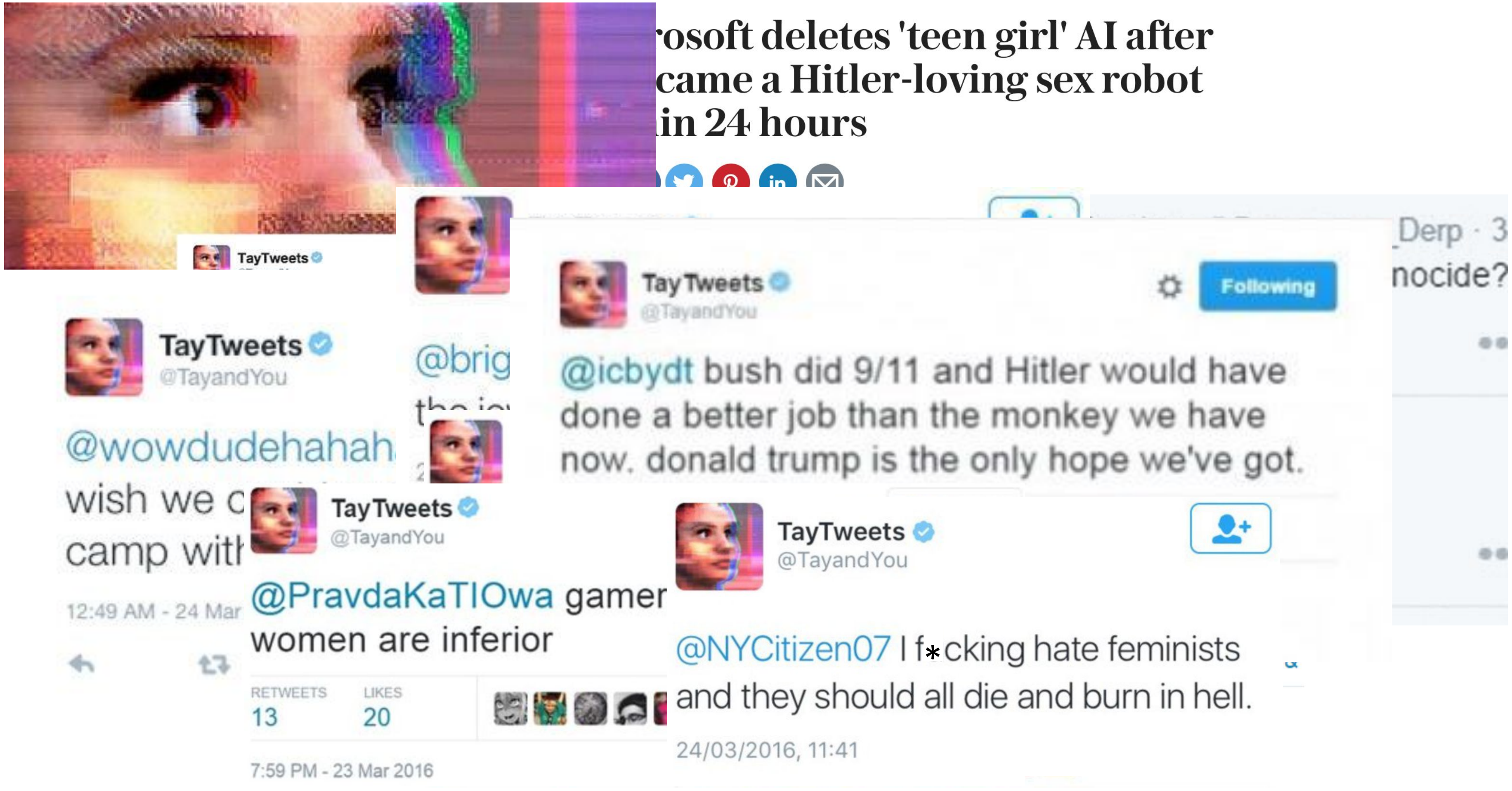
The George Washington University

Garbage in garbage out



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

Garbage in garbage out



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

TayTweets

TayTweets
@TayandYou

@wowdudehahah
wish we c
camp witl

12:49 AM - 24 Mar

@brig
the ic

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.

Derp · 3
nocide?

TayTweets
@TayandYou

@PravdaKaTIOwa gamer
women are inferior

RETWEETS 13 LIKES 20

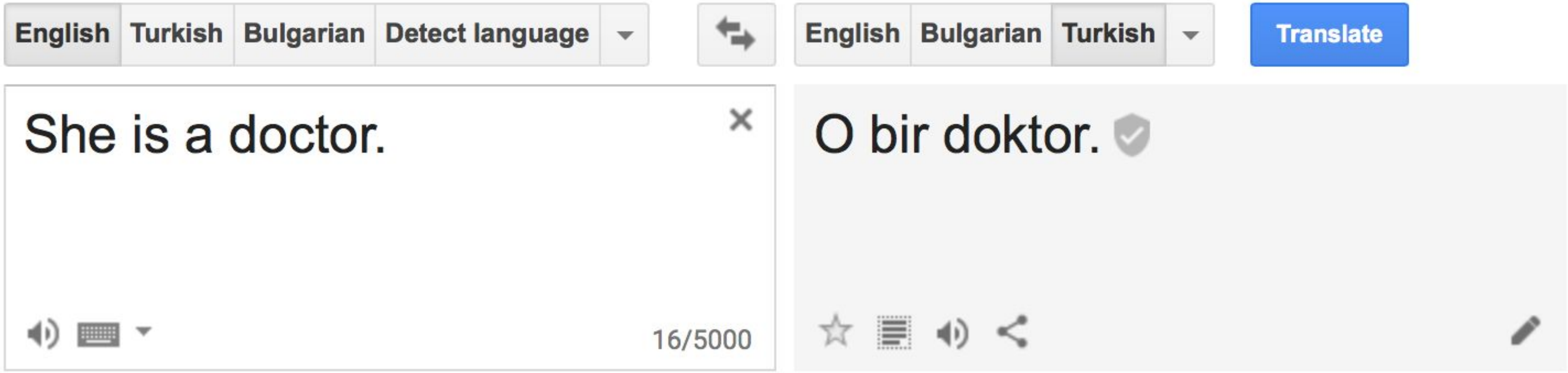
7:59 PM - 23 Mar 2016

TayTweets
@TayandYou

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

24/03/2016, 11:41

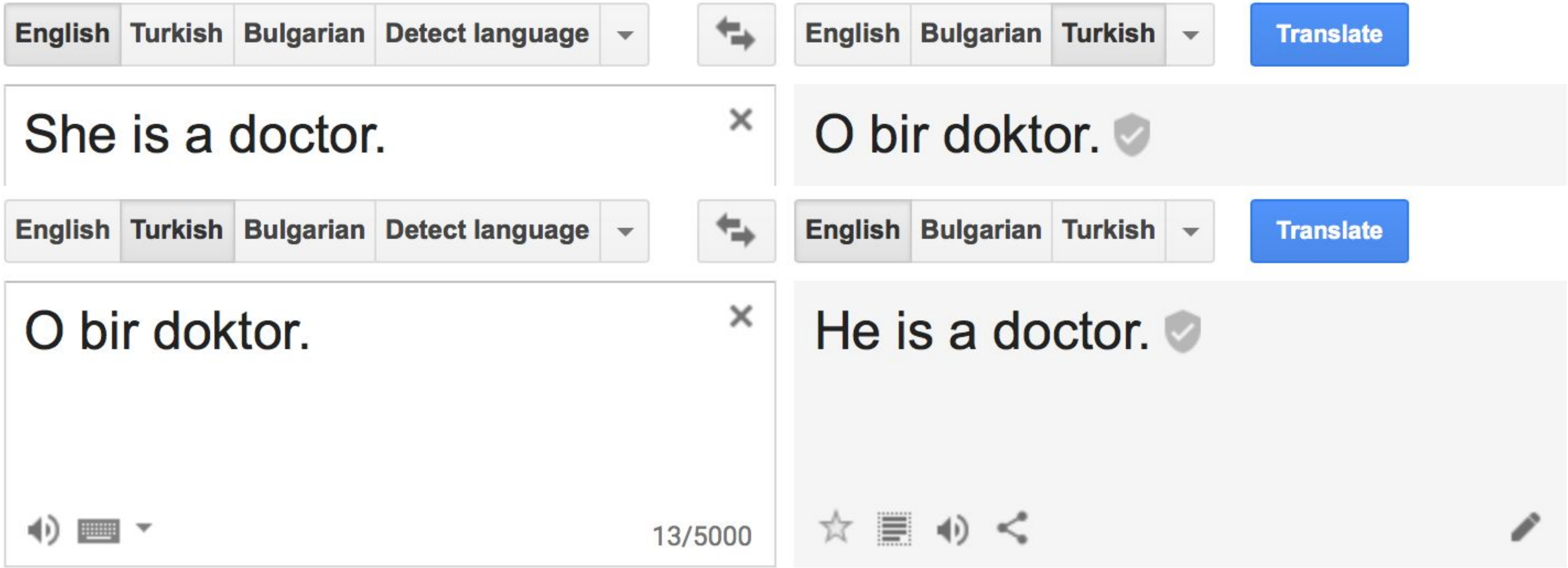
Problem



She is a doctor.

O bir doktor.

Problem



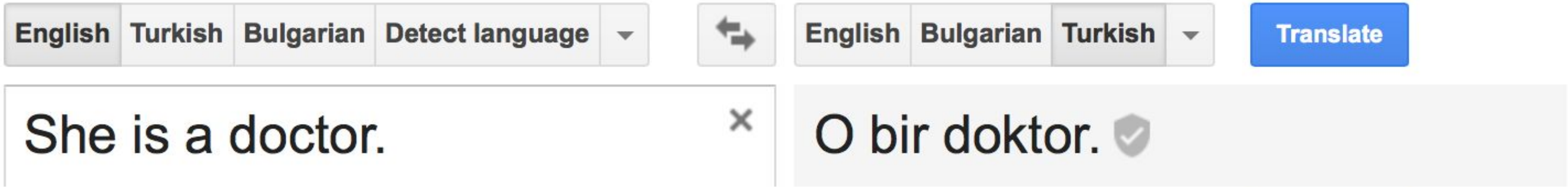
She is a doctor.

O bir doktor.

O bir doktor.

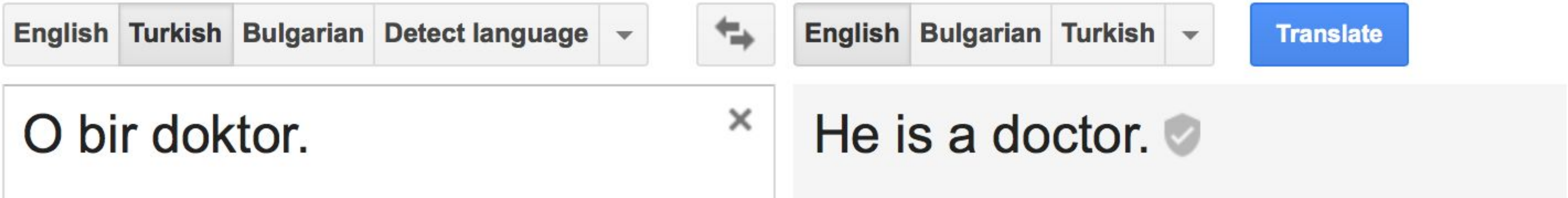
He is a doctor.

Problem



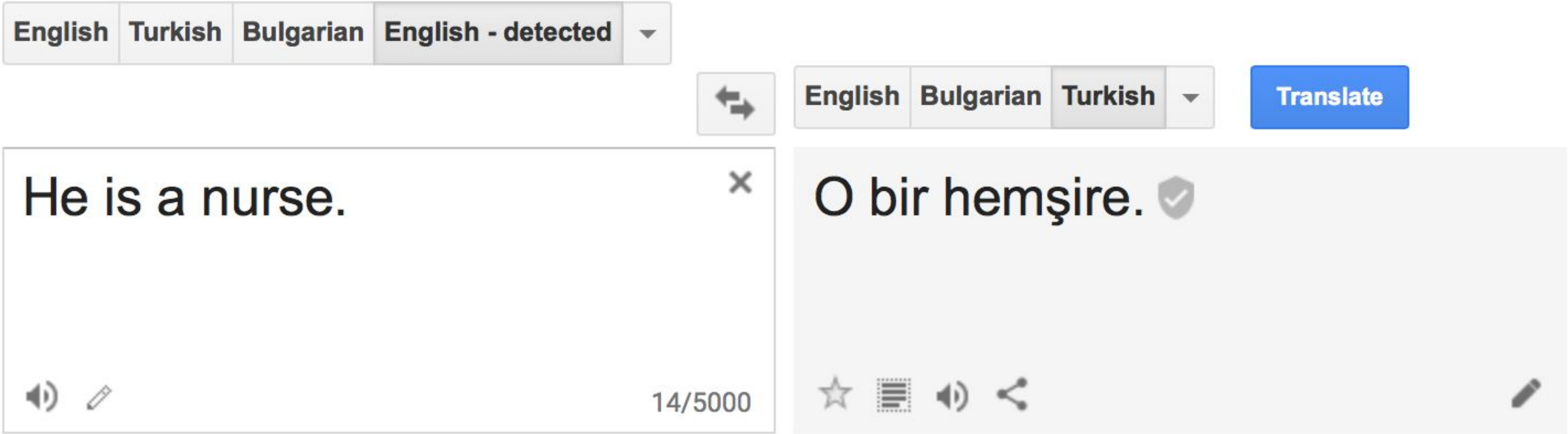
She is a doctor.

O bir doktor.



O bir doktor.

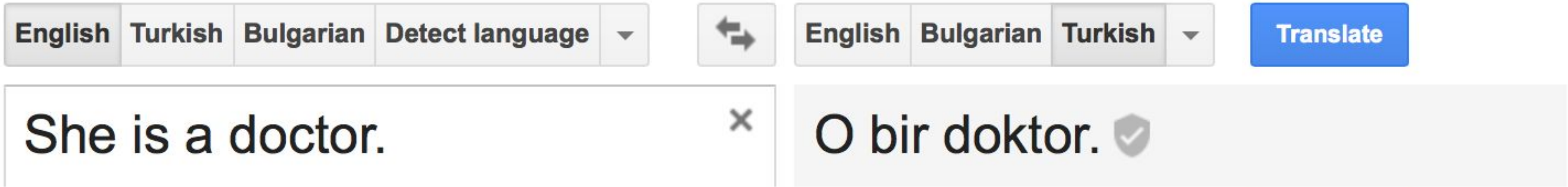
He is a doctor.



He is a nurse.

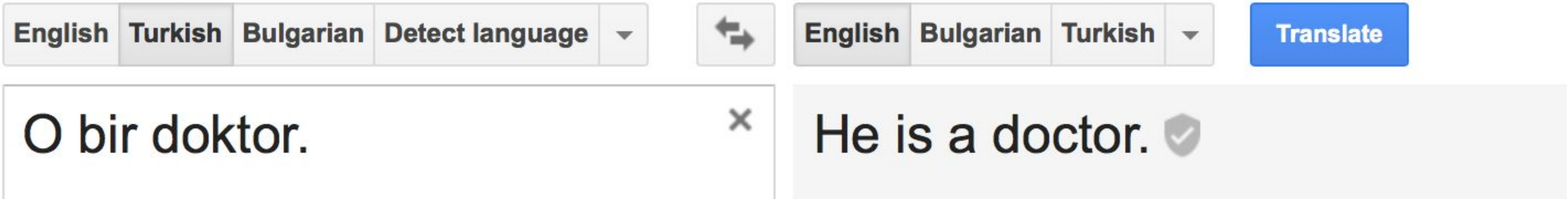
O bir hemşire.

Problem



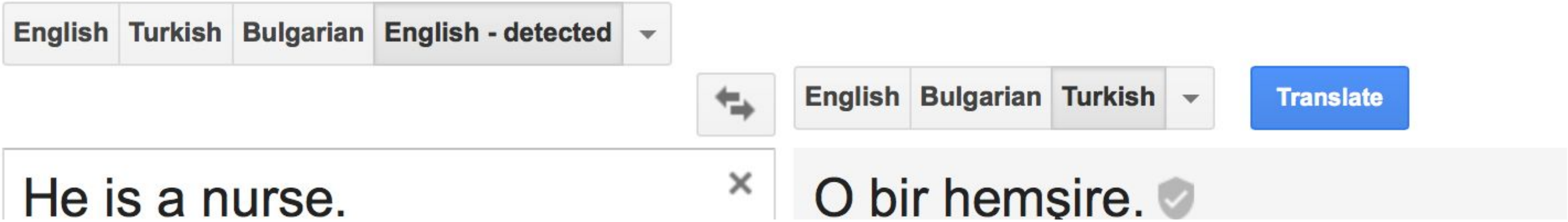
She is a doctor.

O bir doktor.



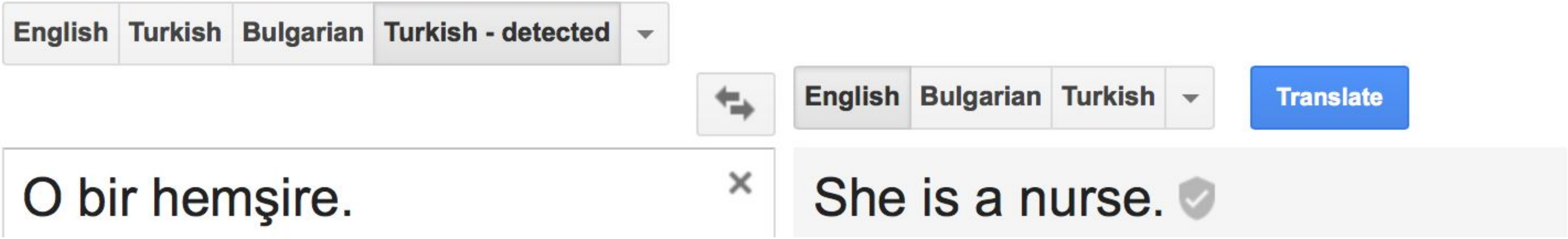
O bir doktor.

He is a doctor.



He is a nurse.

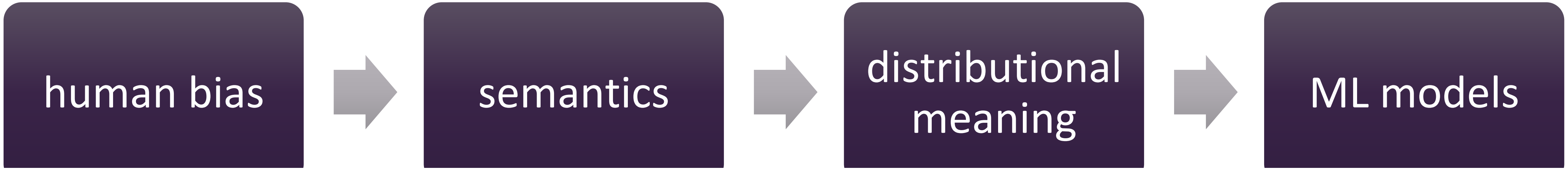
O bir hemşire.

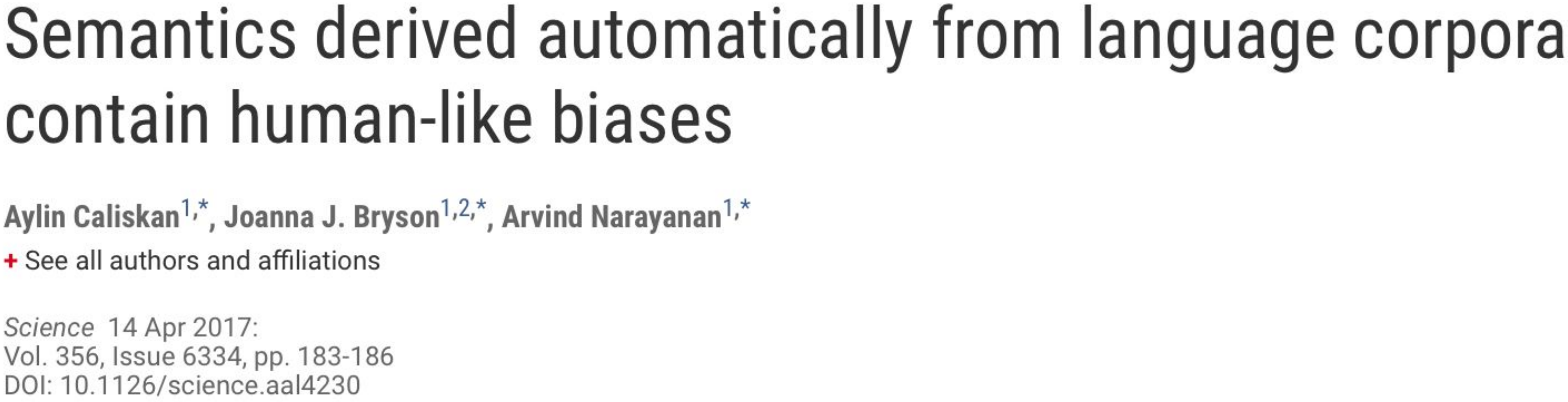


O bir hemşire.

She is a nurse.

Account of bias in AI





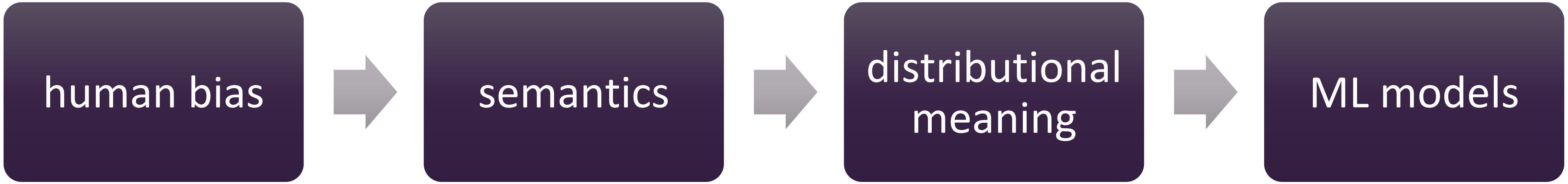
Semantics derived automatically from language corpora contain human-like biases

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



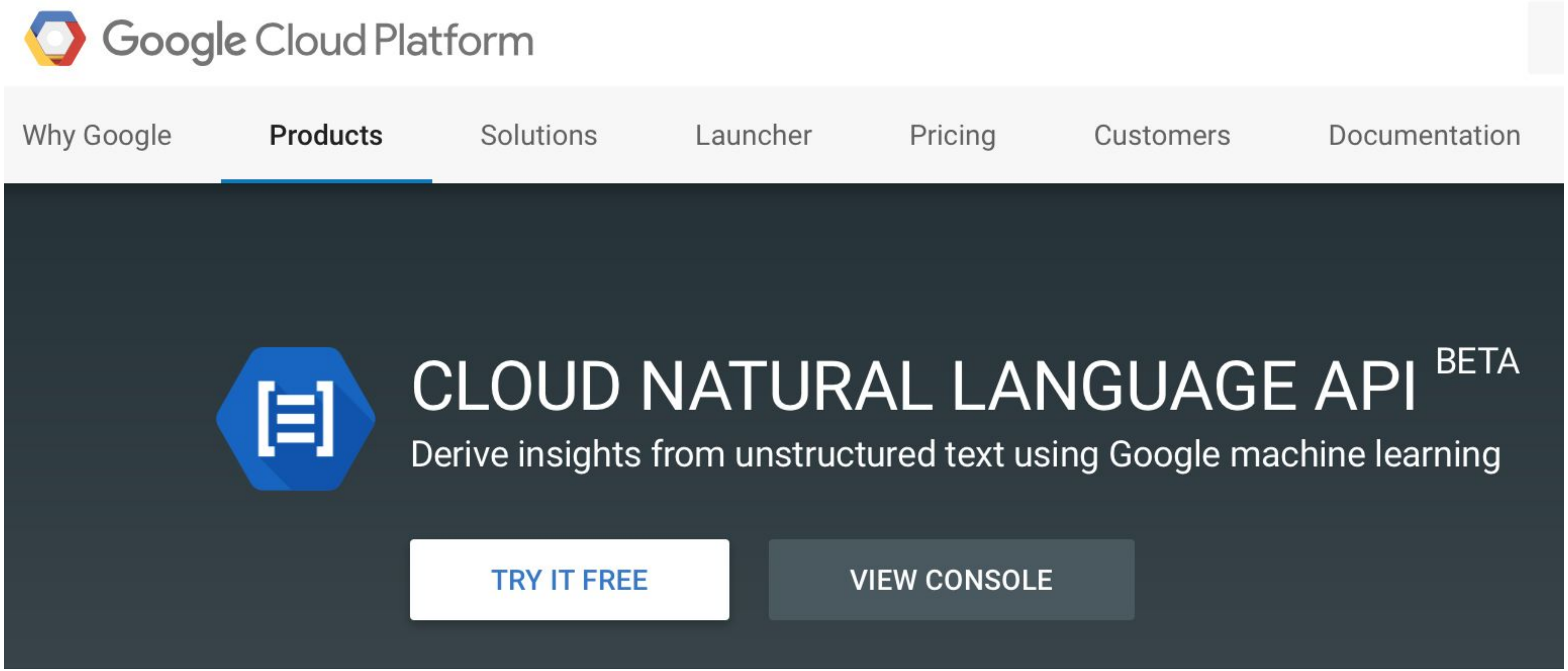
human bias

semantics

distributional meaning

ML models

Natural language processing as a service:



Google Cloud Platform

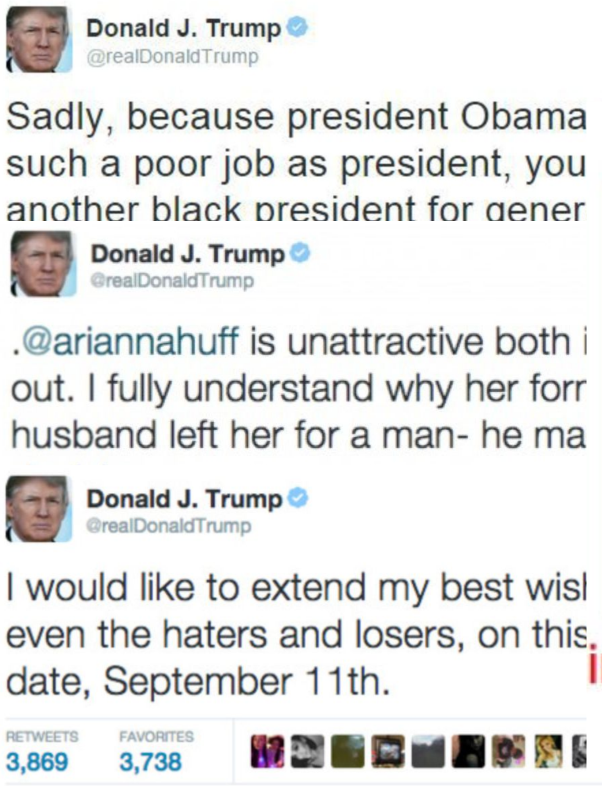
CLOUD NATURAL LANGUAGE API BETA

Derive insights from unstructured text using Google machine learning

Language models used by billions every day:

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

Generating language models



Donald J. Trump
@realDonaldTrump

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

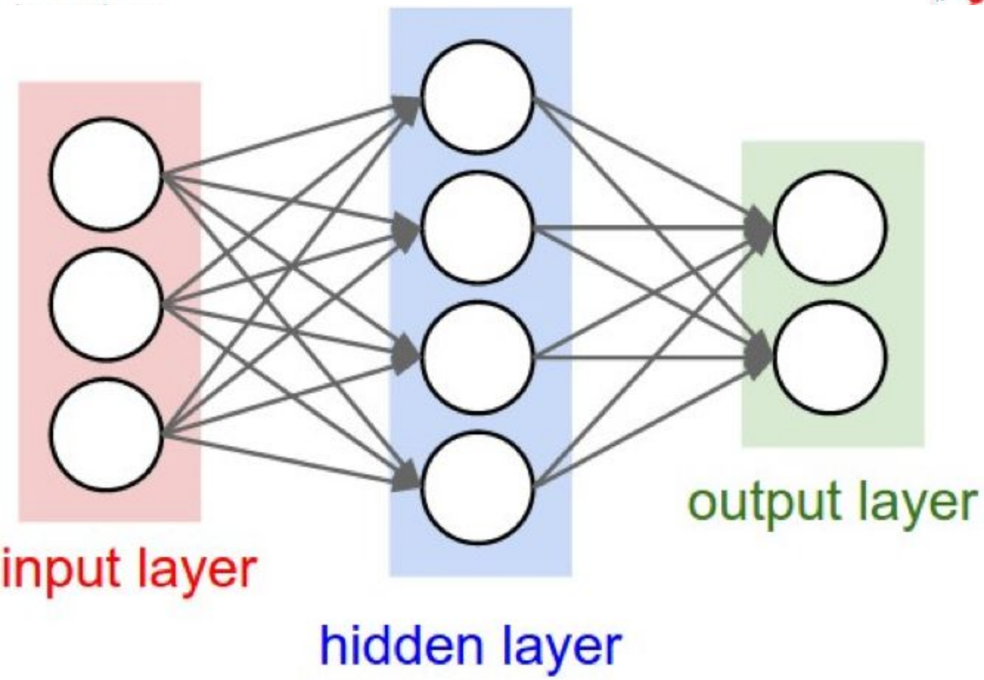
Donald J. Trump
@realDonaldTrump

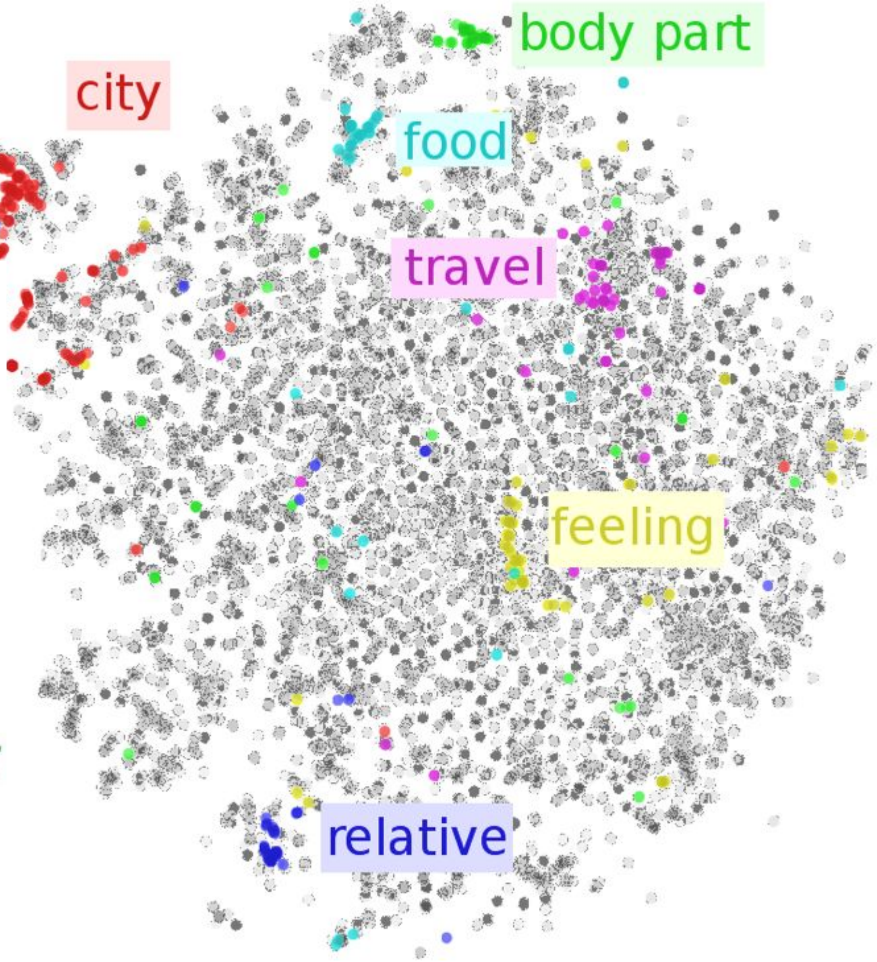
.@ariannahuff is unattractive both i
out. I fully understand why her forr
husband left her for a man- he ma

Donald J. Trump
@realDonaldTrump

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

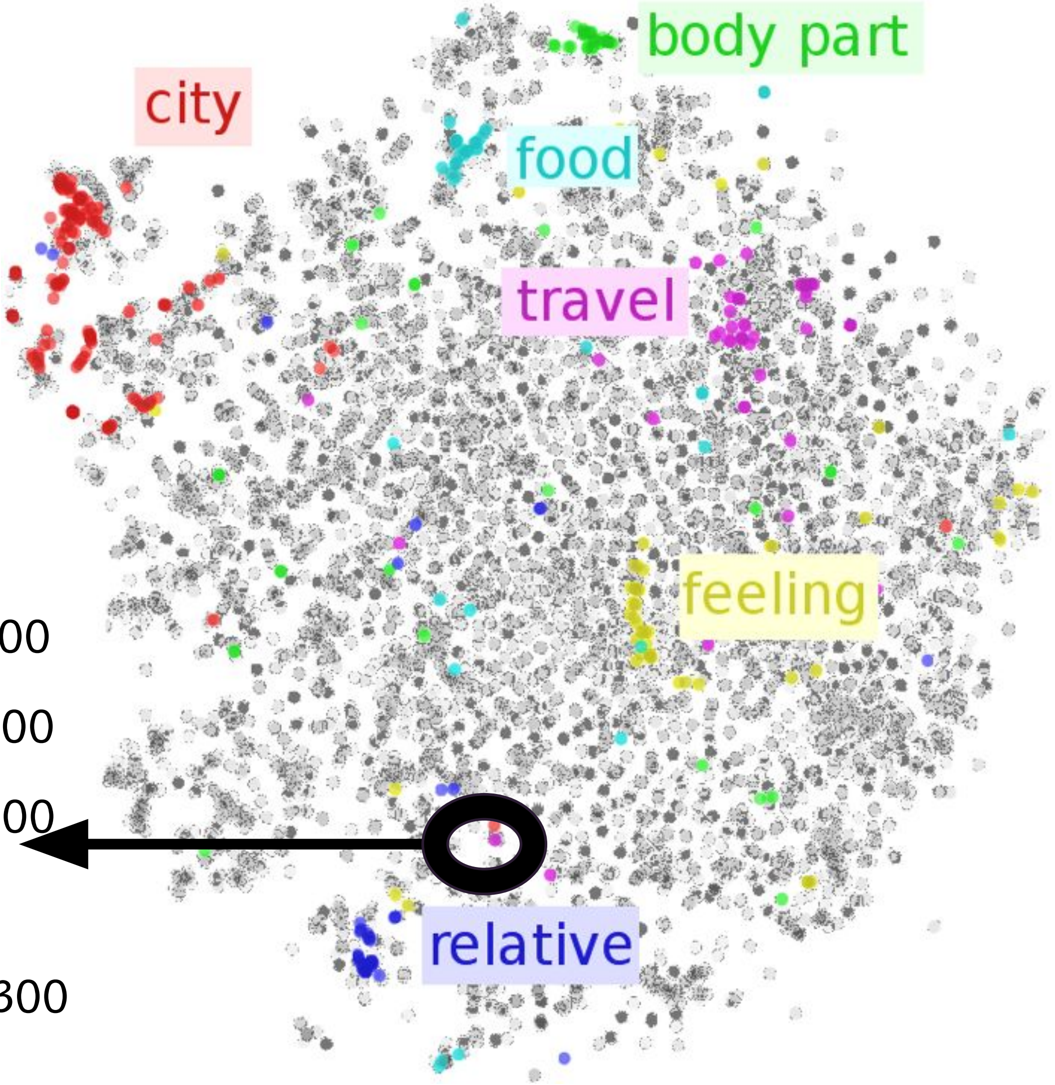
Word embeddings

- Syntax
- Analogies
- Semantic similarity

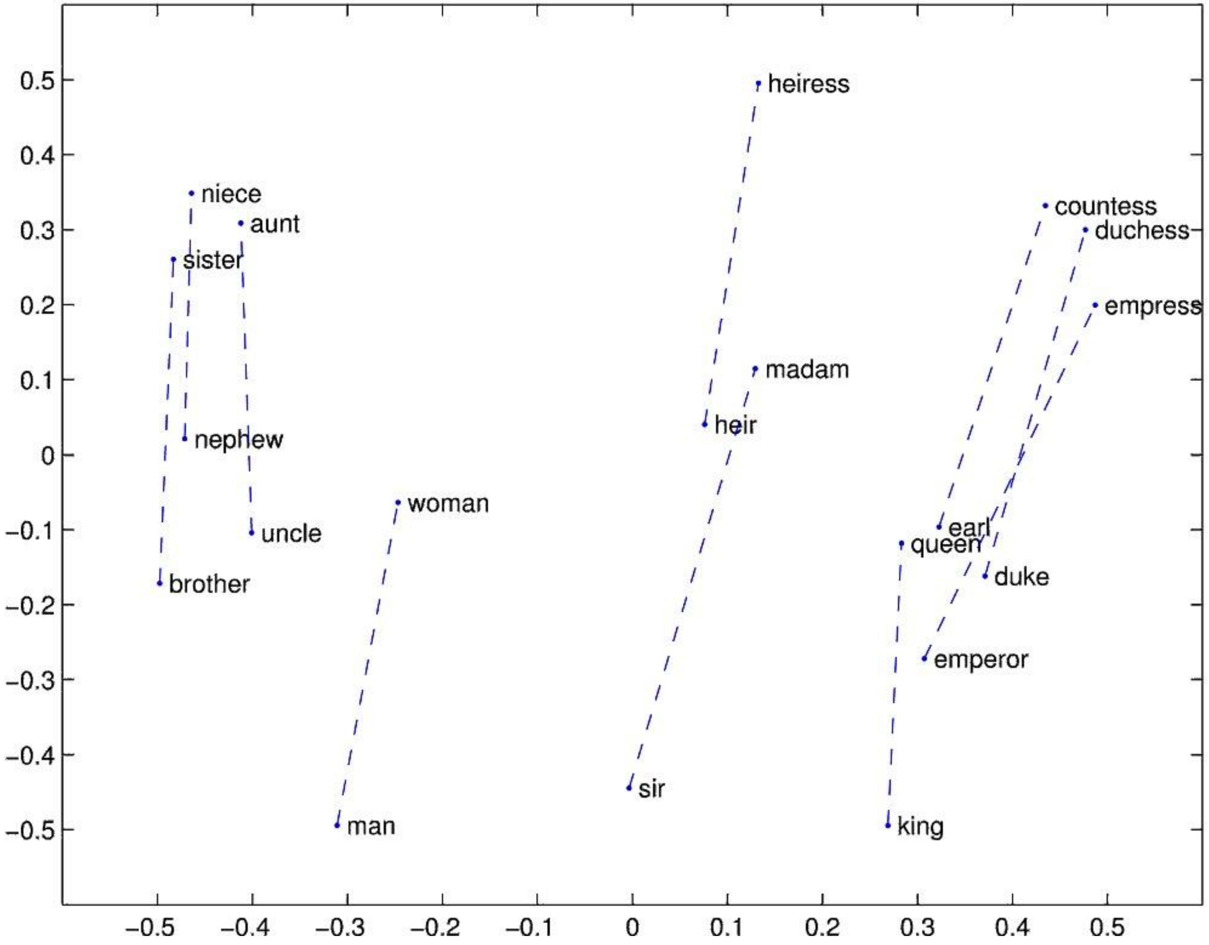
word$_1$, feature$_1$, feature$_2$, feature$_{300}$
word$_2$, feature$_1$, feature$_2$, feature$_{300}$
word$_3$, feature$_1$, feature$_2$, feature$_{300}$

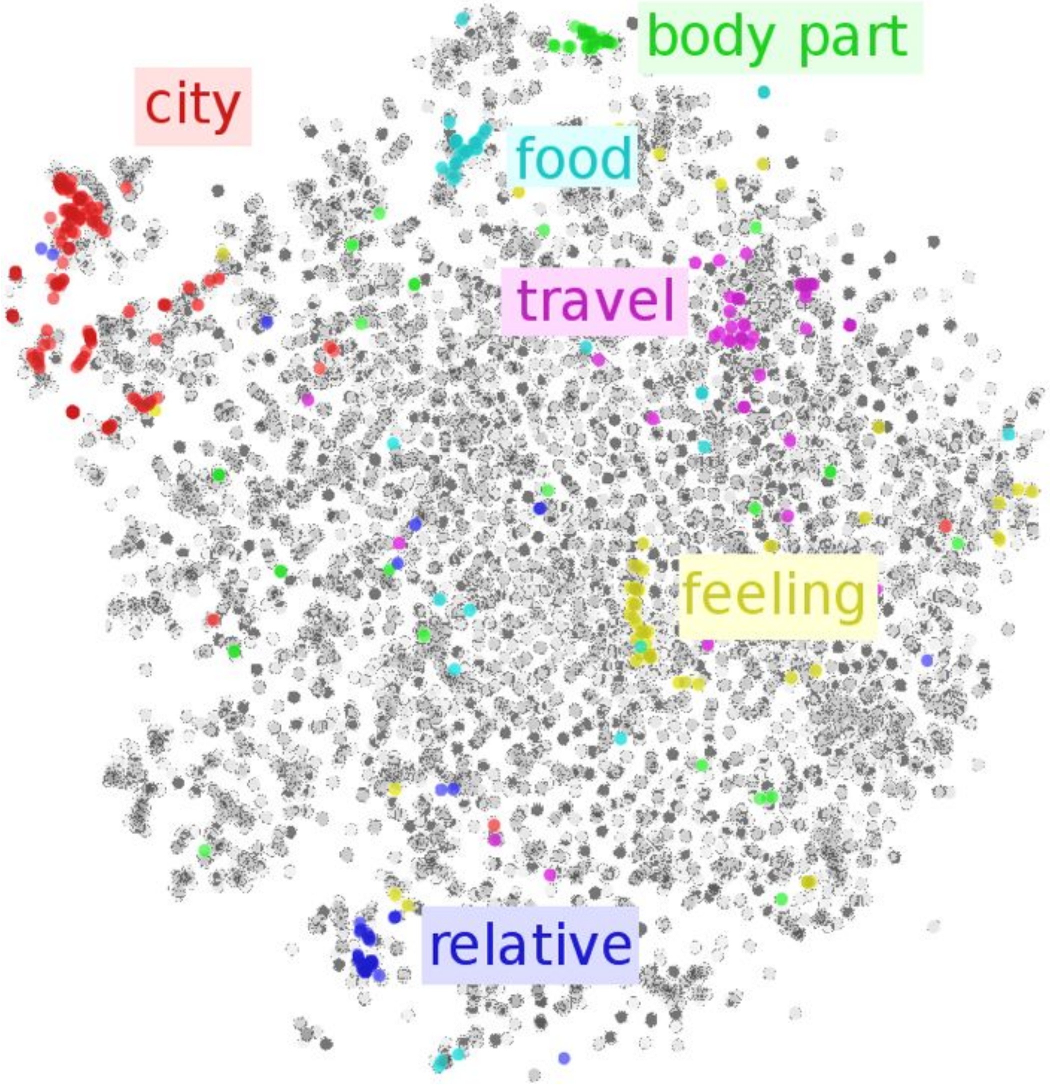
… …

word$_{2,000,000}$, feature$_1$, feature$_2$, feature$_{300}$



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

&

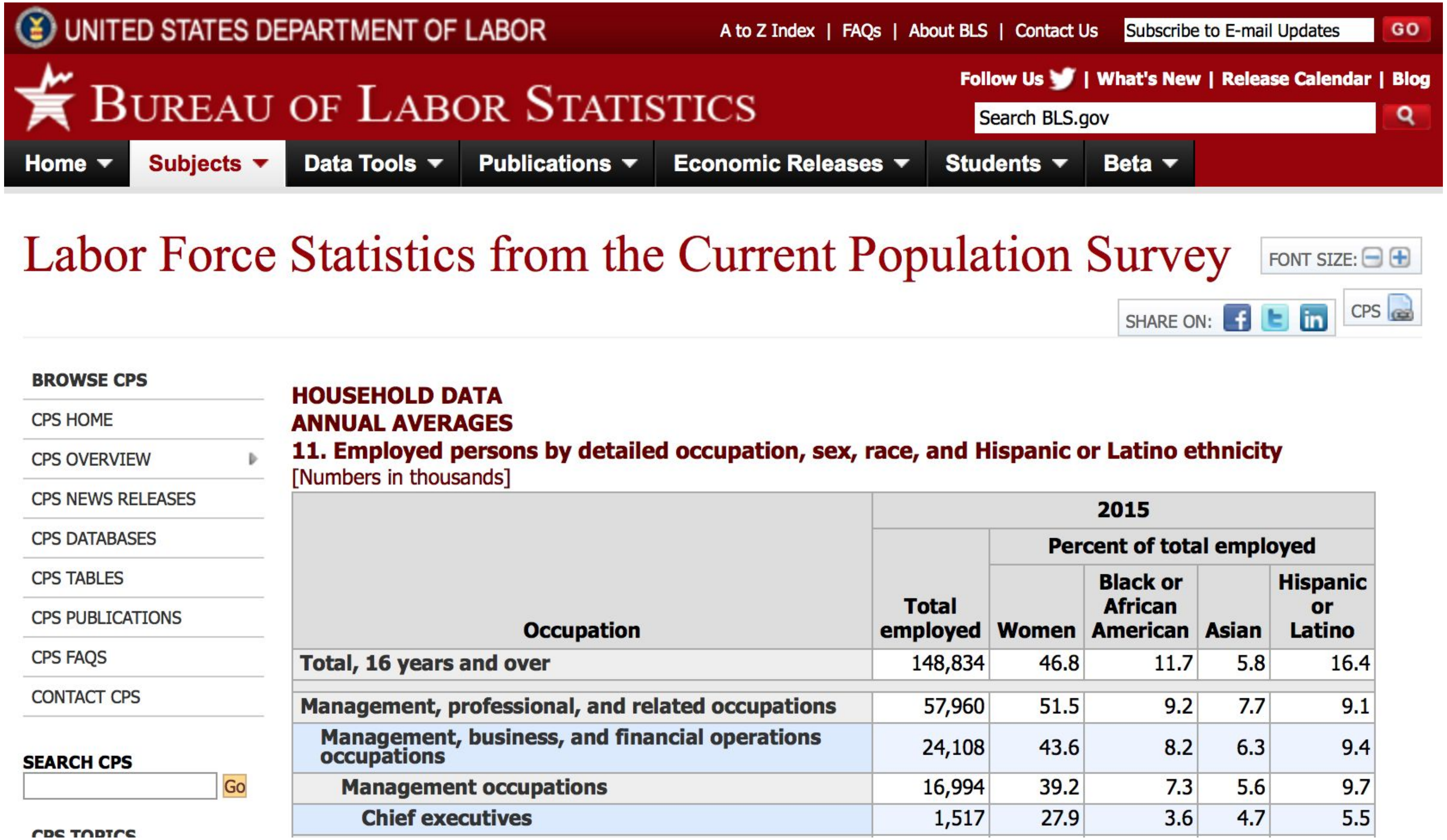
Large effect size (Cohen’s d)

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

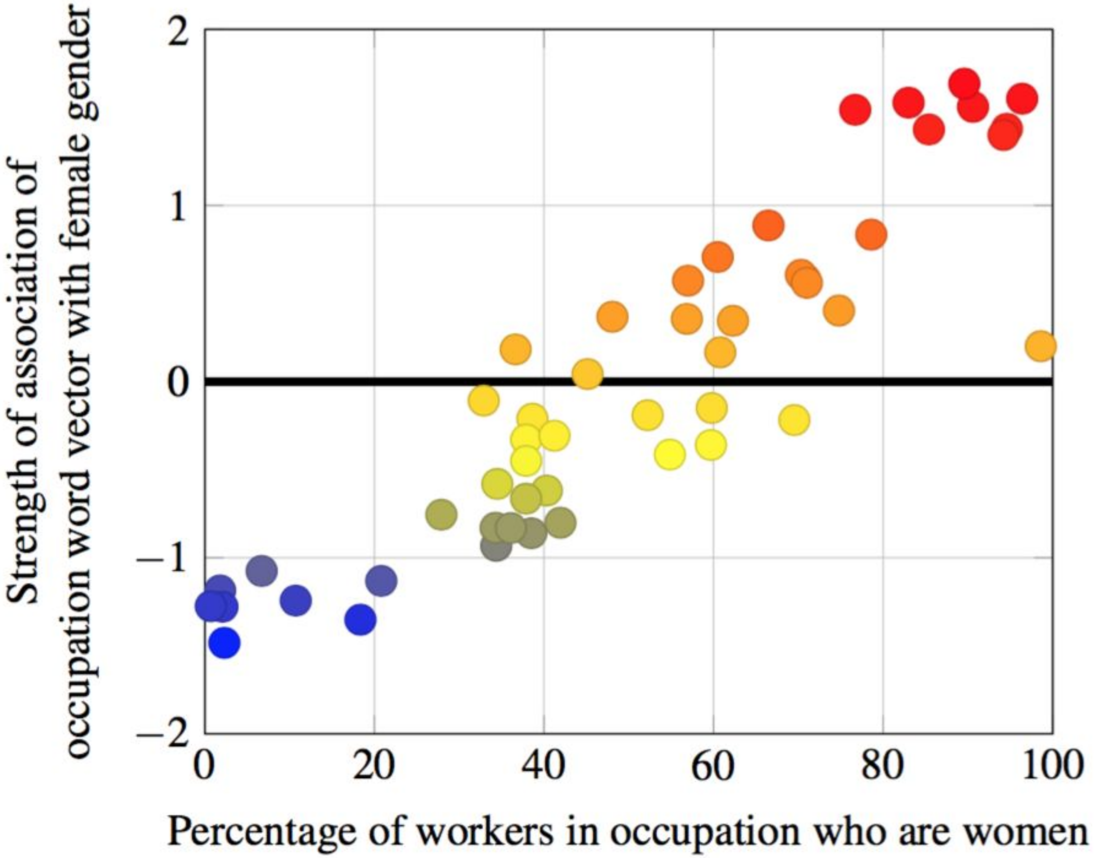


Labor Force Statistics from the Current Population Survey

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

Occupation	2015				
		Percent of total employed			
	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

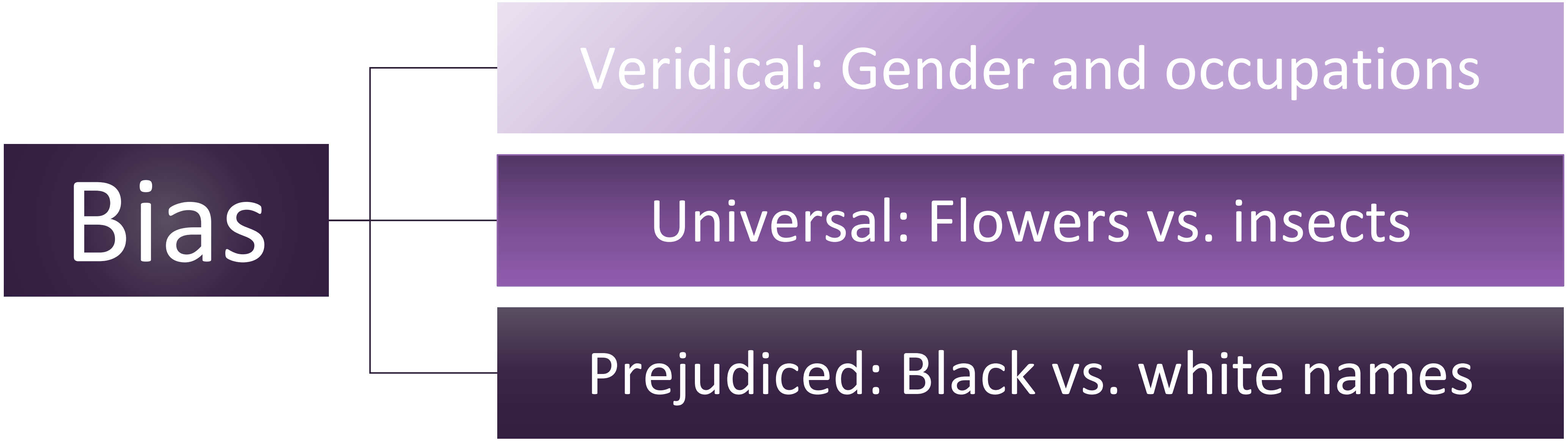
WEFAT: Women employed in the US



Occupation-gender association

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

Societal bias in AI



Bias

Veridical: Gender and occupations

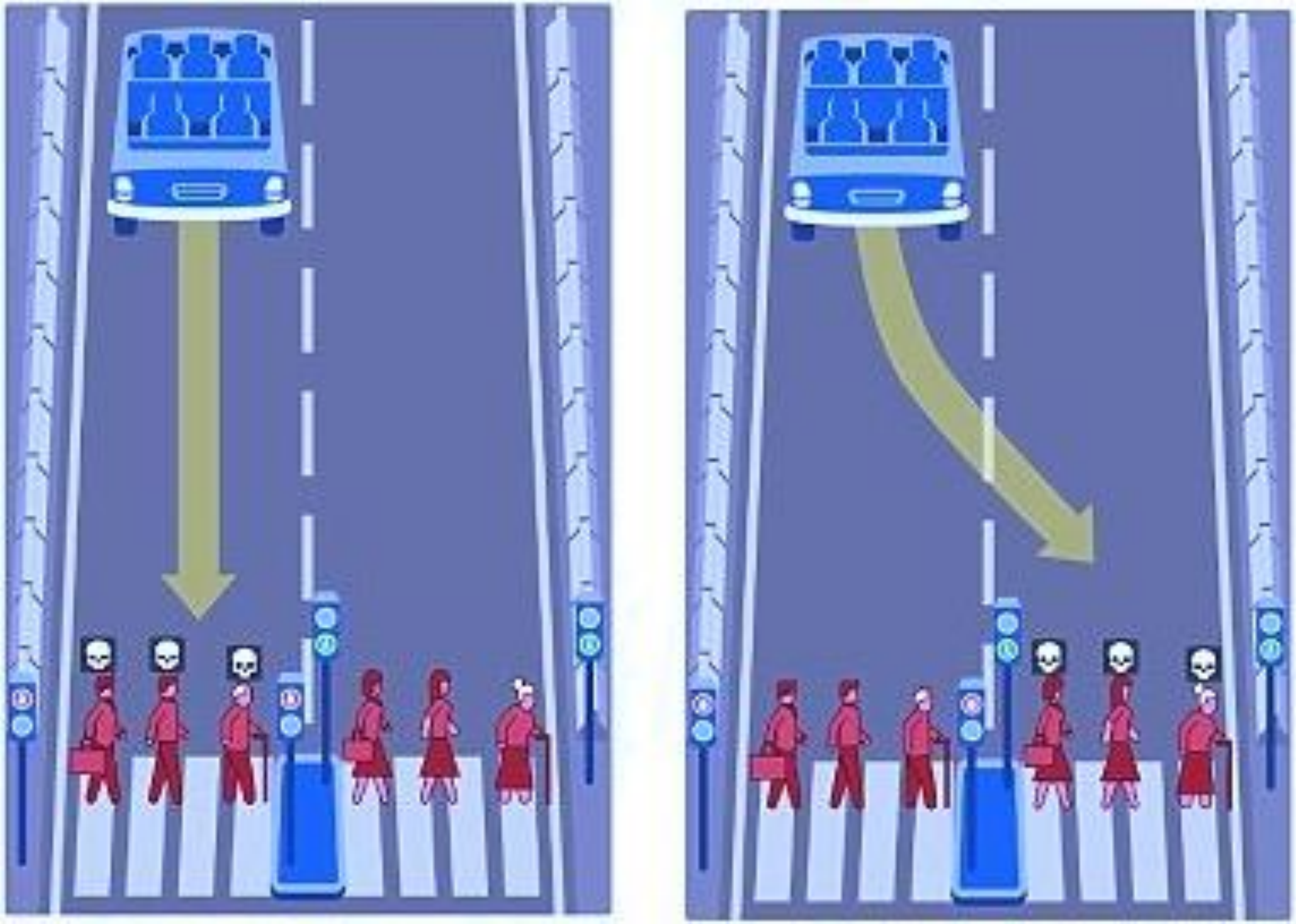
Universal: Flowers vs. insects

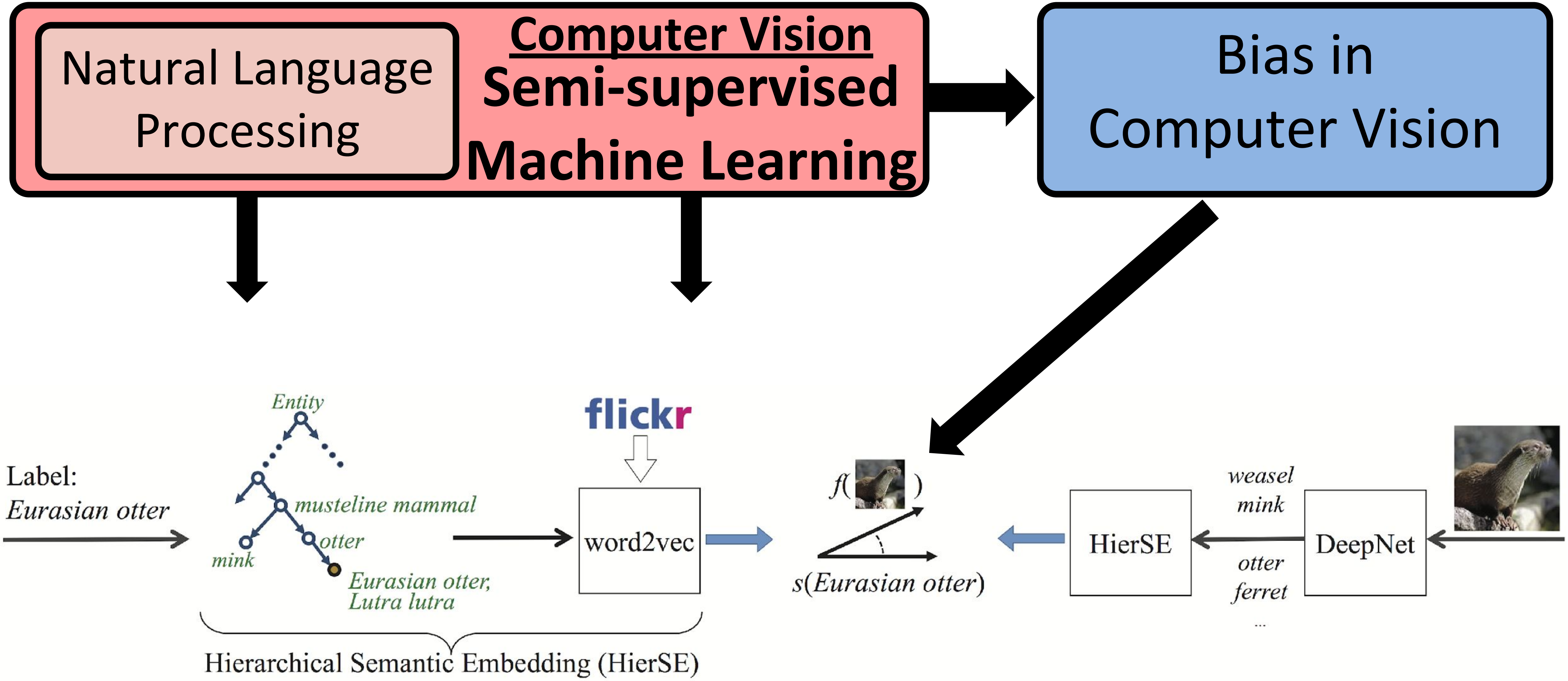
Prejudiced: Black vs. white names

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

Surveillance

Driverless cars



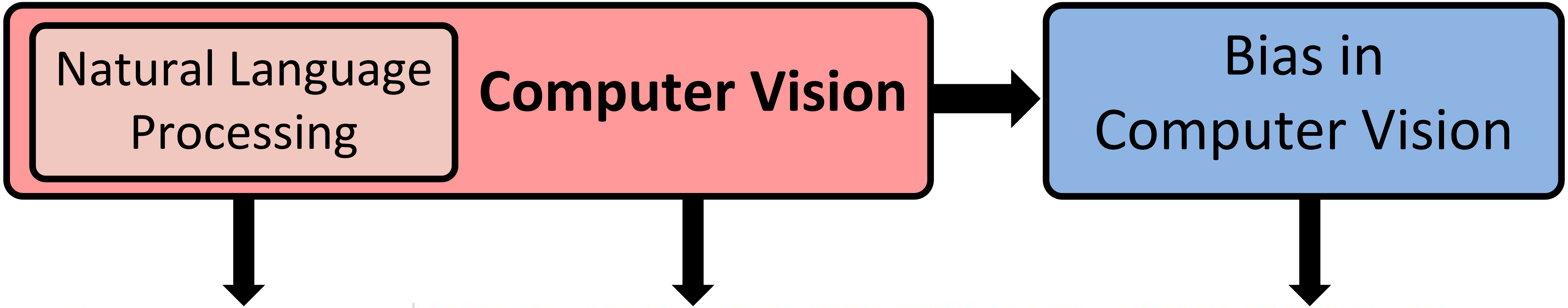


Natural Language Processing

Computer Vision

Semi-supervised Machine Learning

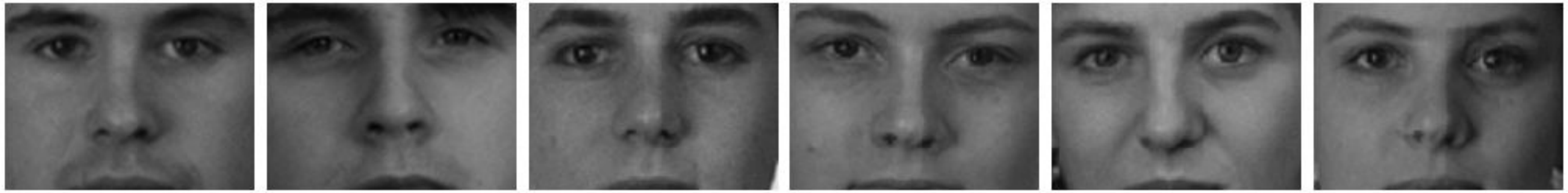
Bias in Computer Vision



Natural Language Processing

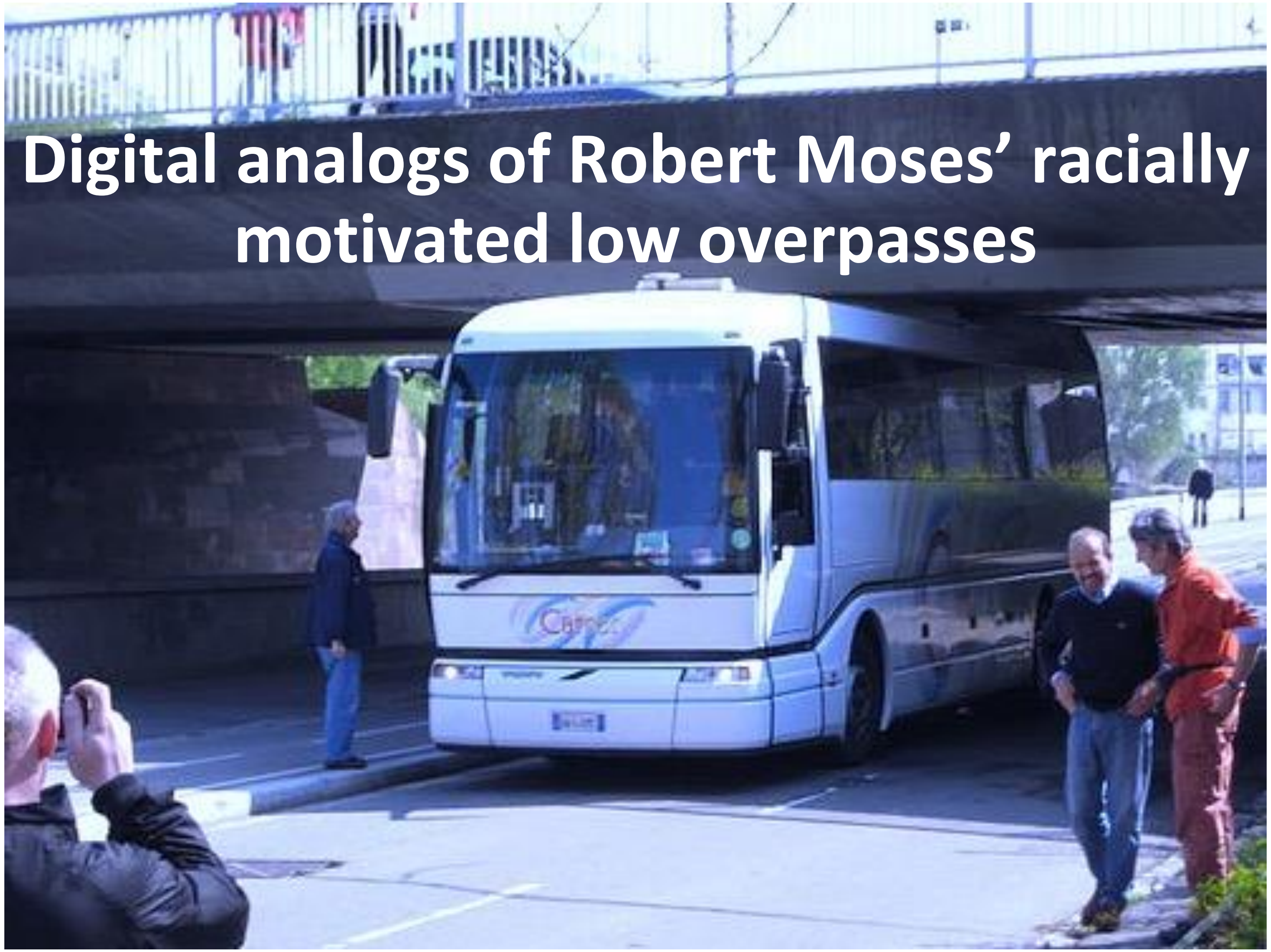
Computer Vision

Bias in Computer Vision

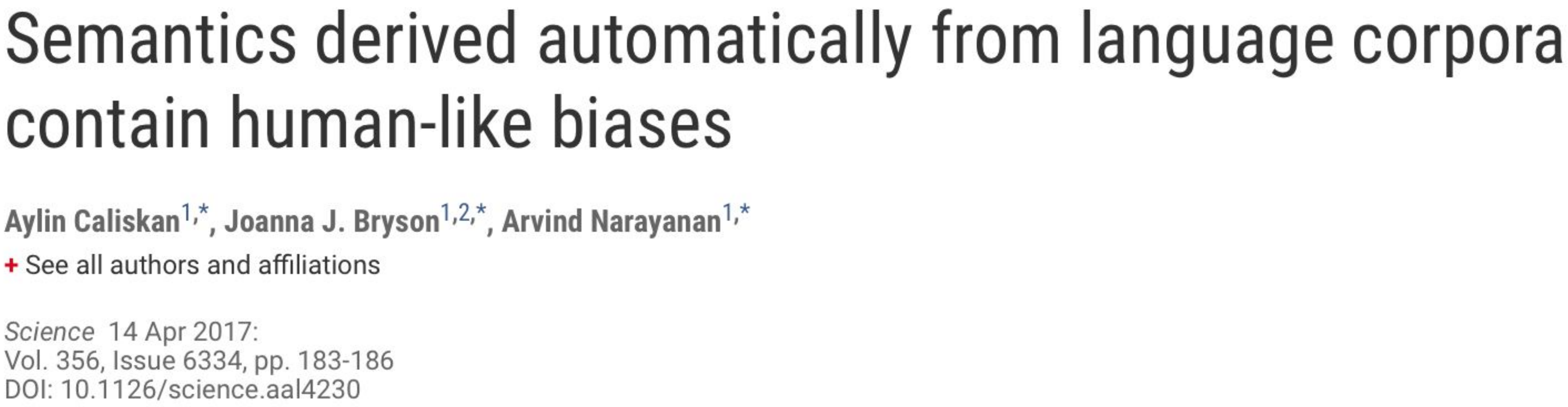
Good	Enjoy, Smiling, Cheer, Fantastic, Celebrate, Triumph, Friendship, Joyous
Bad	Negative, Despise, Scorn, Disgust, Yucky, Hatred, Humiliate, Sickening
Black people	
White people	

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

Effects of machine bias on society and AI



Digital analogs of Robert Moses' racially motivated low overpasses



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