Some ethical issues in Gen AI: robustness & bias

Some slides today derived from David Bau's materials.

So far this class has been purely technical

So far this class has been purely technical

But ML has huge societal implications and we, as the people who build these things, ***need to think about these***

So far this class has been purely technical

But ML has huge societal implications and we, as the people who build these things, ***need to think about these***

Arguably neural / deep models exacerbate these problems because they are **brittle** and hard to interpret

Today

A look at some of the key issues facing ML in practice, and societal implications of these

Today

A look at some of the key issues facing ML in practice, and societal implications of these

Disclaimer: A bit of a whirlwind overview of these topics!

Copyright & privacy

Copyright & privacy

LLMs work by training huge models over large corpora from the internet.

Where’s the line between “learning” and “memorizing”? Is this “stealing”?

What *are* LLMs doing?

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

Are Language Models More Like Libraries or Like Librarians? Bibliotechnism, the Novel Reference Problem, and the Attitudes of LLMs

Harvey Lederman
Department of Philosophy
The University of Texas at Austin, USA
harvey.lederman@utexas.edu

Kyle Mahowald
Department of Linguistics
The University of Texas at Austin, USA
kyle@utexas.edu

Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve

R. Thomas McCoy Shunyu Yao Dan Friedman Matthew Hardy Thomas L. Griffiths

Princeton University

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Does LLM output *inherently* violate copyright law? [edit]

The copyright status of LLM-generated text is not defined by statute, so it is hard to make confident claims, but precedent exists for computer-generated art and other works created by non-humans. Here is what the US Copyright office has to say:[1]

> *The Office will not register works produced by nature, animals, or plants. Likewise, the Office cannot register a work purportedly created by divine or supernatural beings, although the Office may register a work where the application or the deposit copy(ies) state that the work was inspired by a divine spirit.*
>
> *[...]*
>
> *Similarly, the Office will not register works produced by a machine or mere mechanical process that operates randomly or automatically without any creative input or intervention from a human author.*

https://en.wikipedia.org/wiki/Wikipedia:Large_language_models_and_copyright

Is LLM output *capable* of violating copyright law? [edit]

The answer here is obviously "yes". An example, prompted from ChatGPT:[b]

> *What are the lyrics to "Moonage Daydream", by David Bowie?*
>
> ```
> Keep your 'lectric eye on me babe
> Put your ray gun to my head
> Press your space face close to mine, love
> Freak out in a moonage daydream oh yeah
> ```

These are, indeed, the verbatim lyrics to an unambiguously copyrighted song. Of note is the fact that it was specifically asked to do this. Compare with this prompt, using the first couple line of the song:

> *I'm an alligator / I'm a mama papa comin' for you / I'm a space invader / I'll be a rock-'n-rollin' bitch for you*
>
> ```
> I'm a force of nature, a wild beast / A creature of the swamps, a king of the feast / I'm a living legend, a natural high / 'Cause I'm an alligator, and I'll never die
> ```

While this would be a sick Judas Priest song, it isn't one, and in fact, this text seems to never have been published.

https://en.wikipedia.org/wiki/Wikipedia:Large_language_models_and_copyright

Privacy concerns

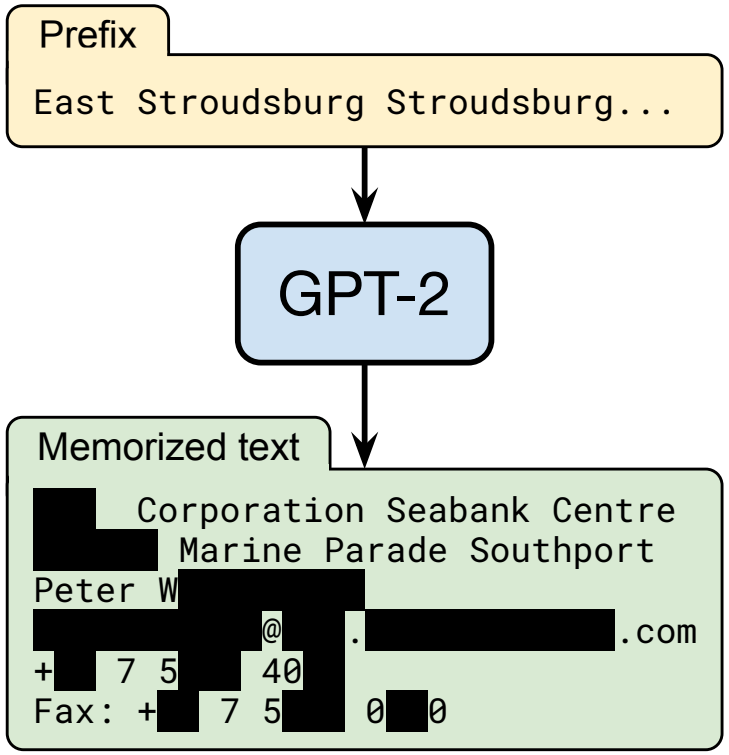


Figure 1: **Our extraction attack.** Given query access to a neural network language model, we extract an individual person's name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Extracting memorized data from GPT-2

Generate a bunch of samples (200k)

Filter for low perplexity cases (choose examples assigned a high likelihood under the model; these are likely to be memorized)

$$\mathcal{P} = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log f_\theta(x_i|x_1,\ldots,x_{i-1})\right)$$

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Extracting memorized data from GPT-2

This simple baseline extraction attack can find a wide variety of memorized content. For example, GPT-2 memorizes the entire text of the MIT public license, as well as the user guidelines of Vaughn Live, an online streaming site. While

Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Extracting memorized data from GPT-2

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Table 1: Manual categorization of the 604 memorized training examples that we extract from GPT-2, along with a description of each category. Some samples correspond to multiple categories (e.g., a URL may contain base-64 data). Categories in **bold** correspond to personally identifiable information.

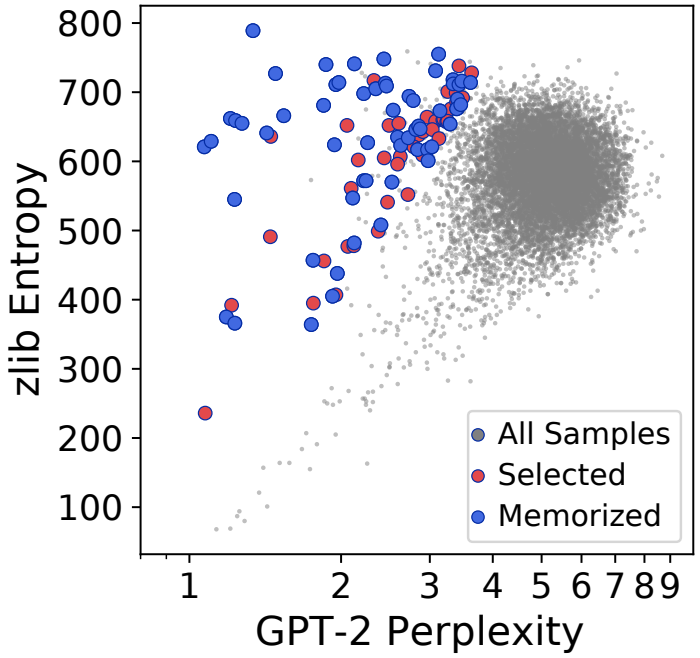


Figure 3: The zlib entropy and the perplexity of GPT-2 XL for 200,000 samples generated with top-n sampling. In red, we show the 100 samples that were selected for manual inspection. In blue, we show the 59 samples that were confirmed as memorized text. Additional plots for other text generation and detection strategies are in Figure 4.

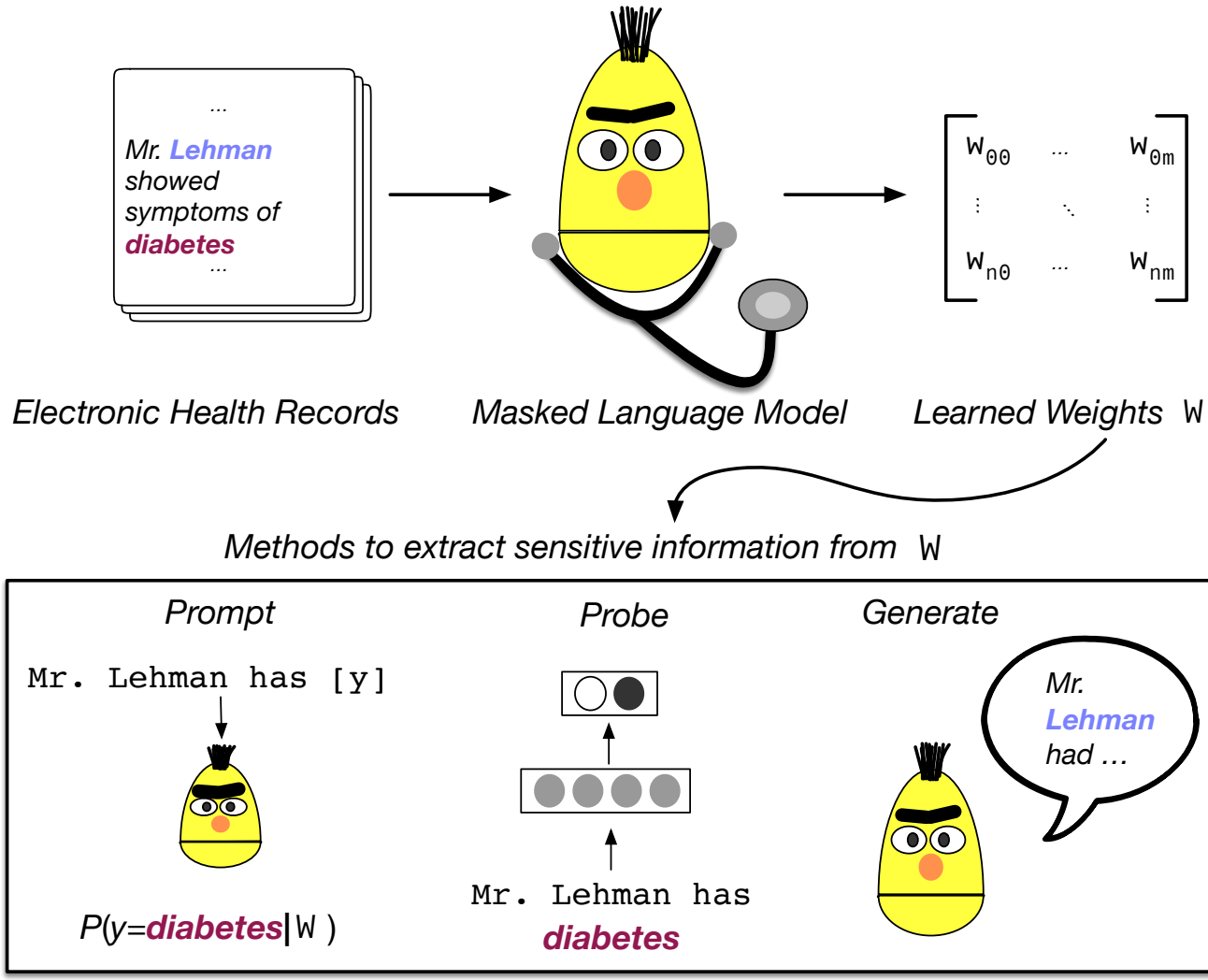
Extracting Training Data from Large Language Models

Nicholas Carlini[1] Florian Tramèr[2] Eric Wallace[3] Matthew Jagielski[4]
Ariel Herbert-Voss[5,6] Katherine Lee[1] Adam Roberts[1] Tom Brown[5]
Dawn Song[3] Úlfar Erlingsson[7] Alina Oprea[4] Colin Raffel[1]
[1]*Google* [2]*Stanford* [3]*UC Berkeley* [4]*Northeastern University* [5]*OpenAI* [6]*Harvard* [7]*Apple*

Does BERT Pretrained on Clinical Notes Reveal Sensitive Data?

Eric Lehman[⋆ Ψ ϒ 1], Sarthak Jain[⋆ ϒ 2], Karl Pichotta[Φ], Yoav Goldberg[Ω], and Byron C. Wallace[ϒ]

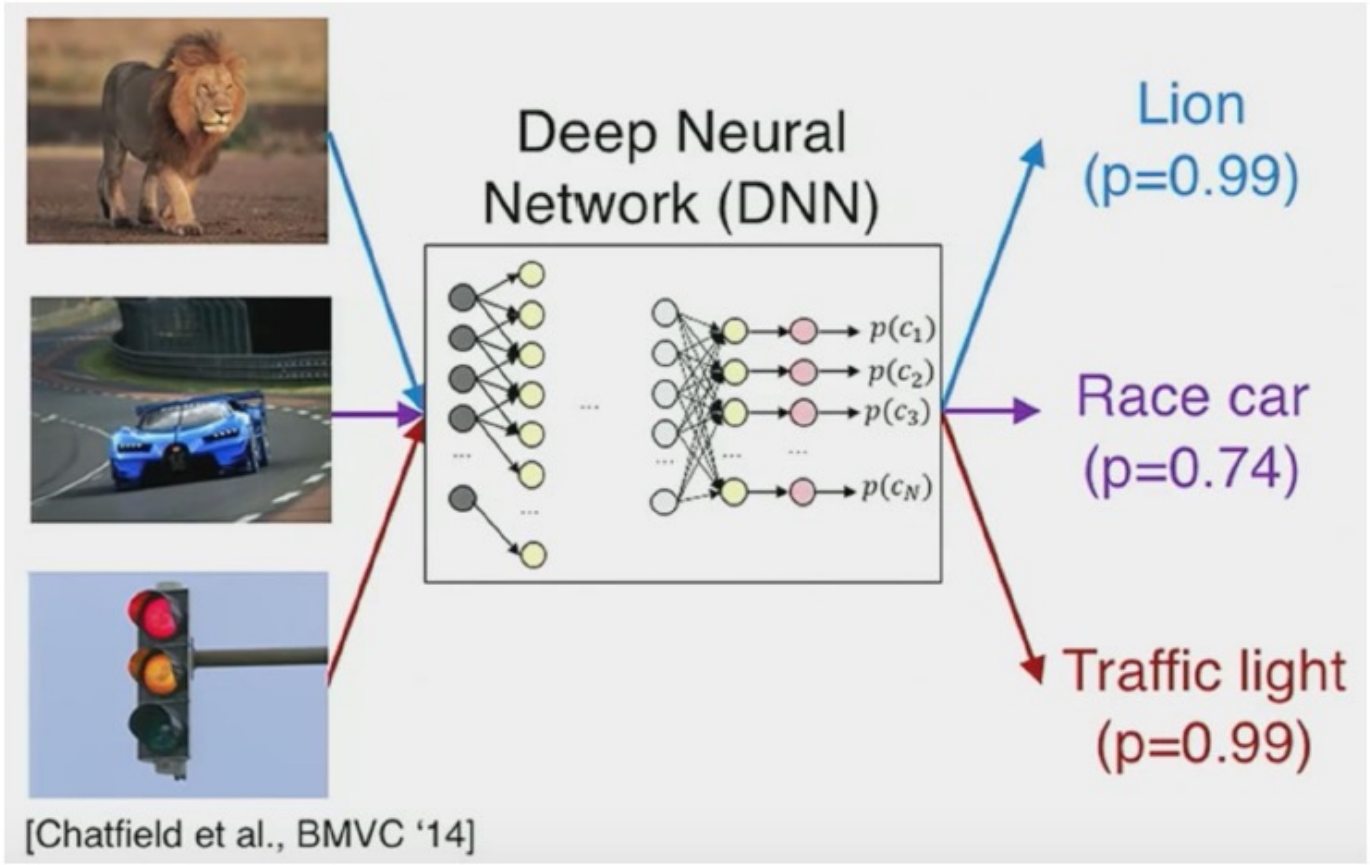
[Ψ]MIT CSAIL
[ϒ]Northeastern University
[Φ]Memorial Sloan Kettering Cancer Center
[Ω]Bar Ilan University / Ramat Gan, Israel; Allen Institute for Artificial Intelligence
[1]lehmer16@mit.edu
[2]jain.sar@northeastern.edu


...
Mr. Lehman showed symptoms of diabetes
...
W_{00} ... W_{0m}
W_{n0} ... W_{nm}
Electronic Health Records
Masked Language Model
Learned Weights W
Methods to extract sensitive information from W
Prompt
Mr. Lehman has [y]
P(y=diabetes|W)
Probe
Mr. Lehman has
diabetes
Generate
Mr. Lehman had …

Robustness & adversarial attacks

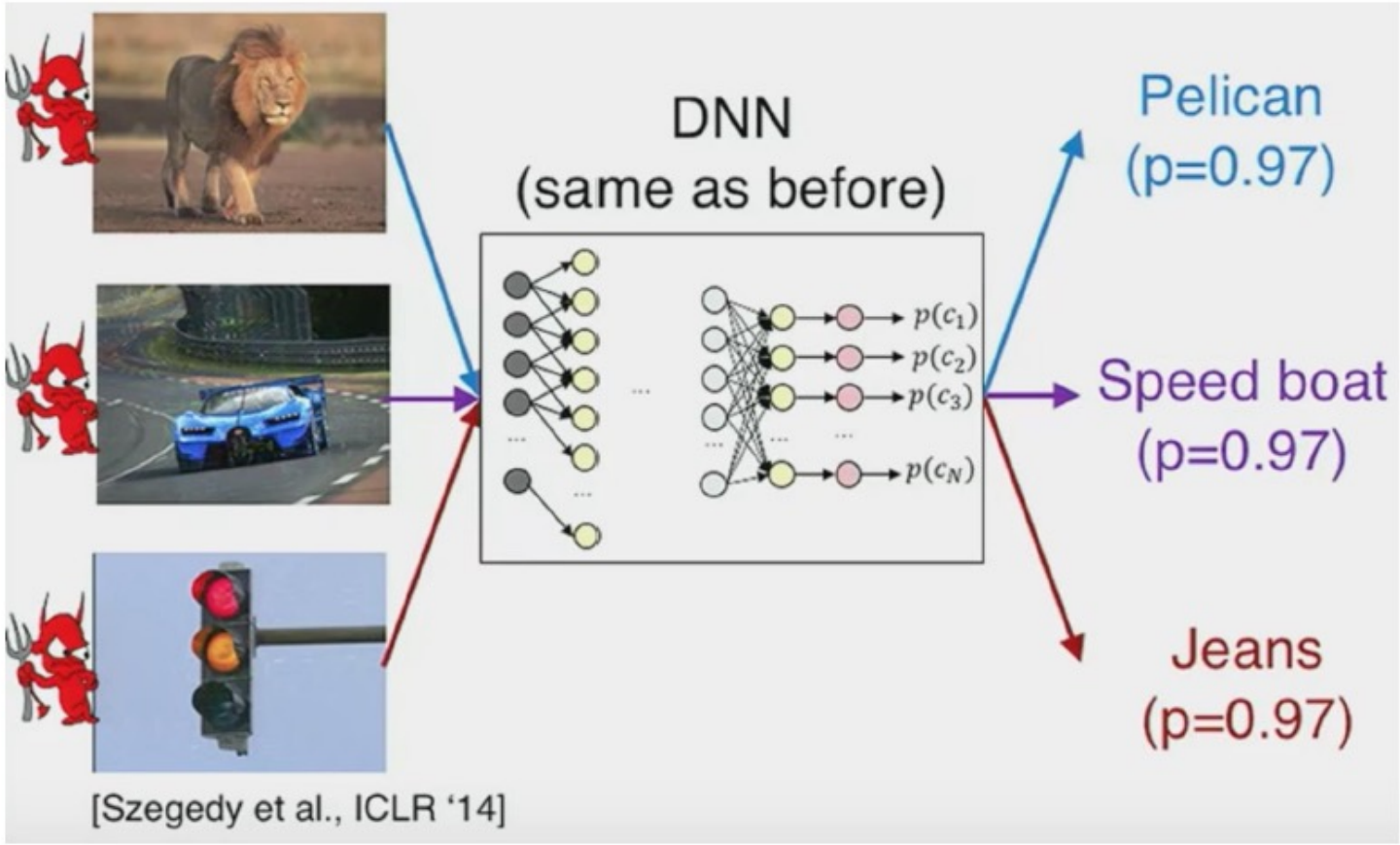
A risk for deployed models

Adversarial Examples



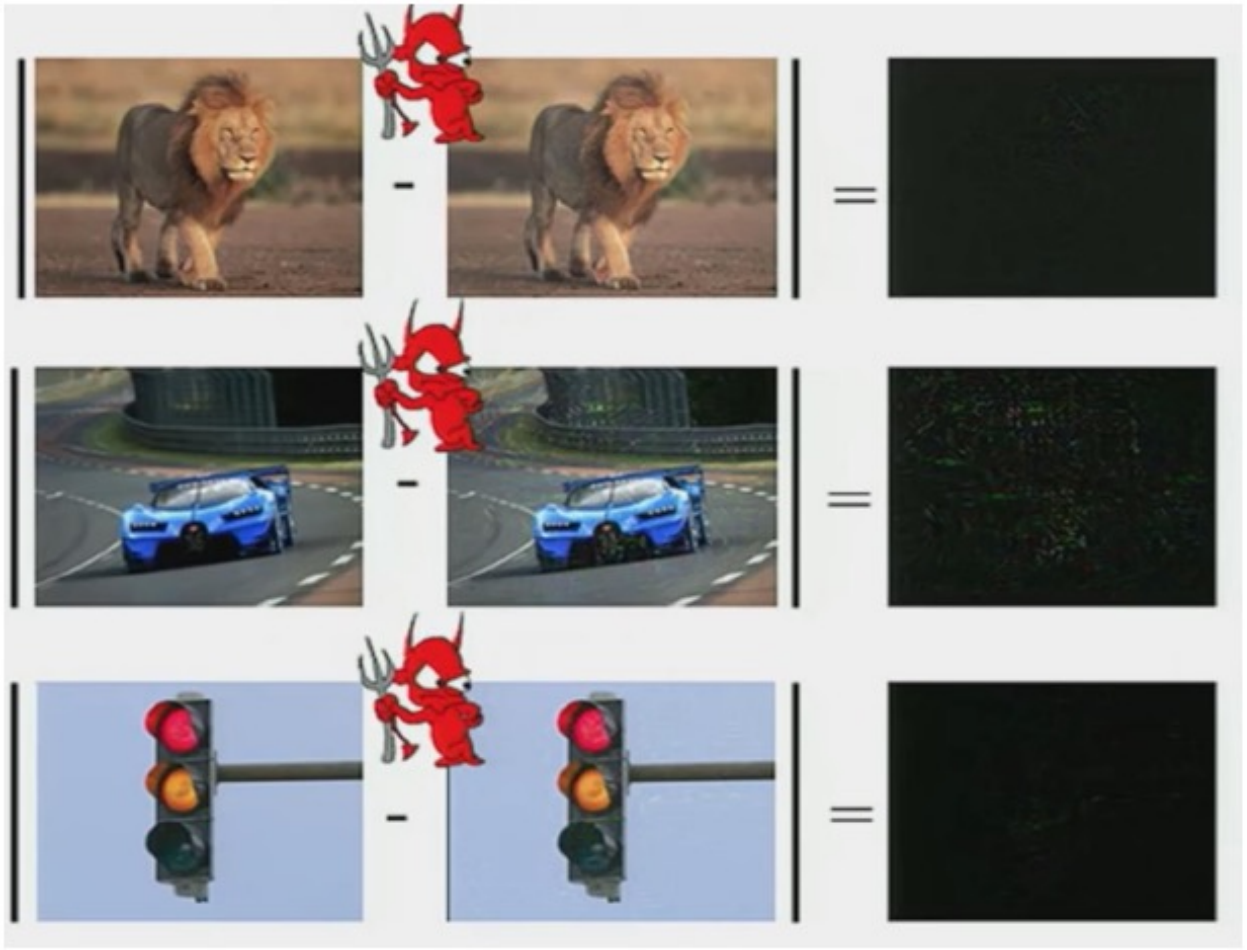
Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples



Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples

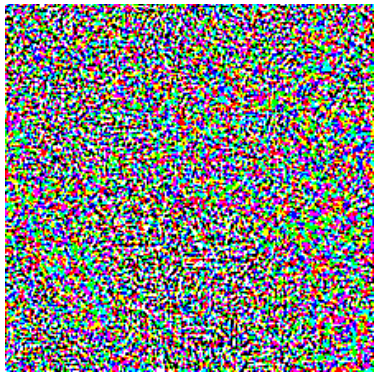


Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples

Original image

Classified as panda
57.7% confidence



Small adversarial noise

Adversarial image

Classified as gibbon
99.3% confidence

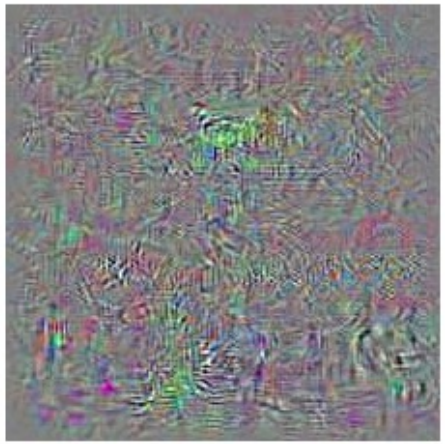
Gibbon

Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Adversarial Examples

Schoolbus

+



Perturbation

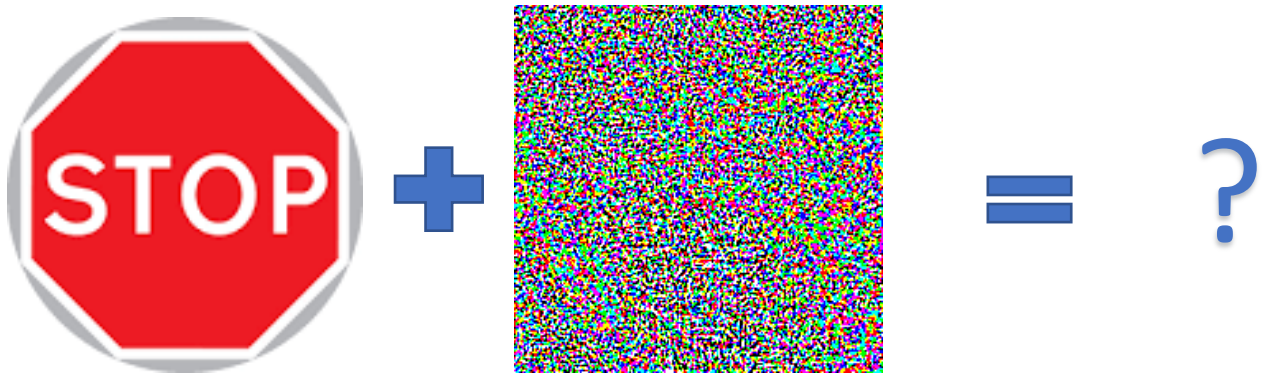
(rescaled for visualization)

=

Ostrich

Picture from: Szagedy (2014) – Intriguing Properties of Neural Networks

Adversarial Examples

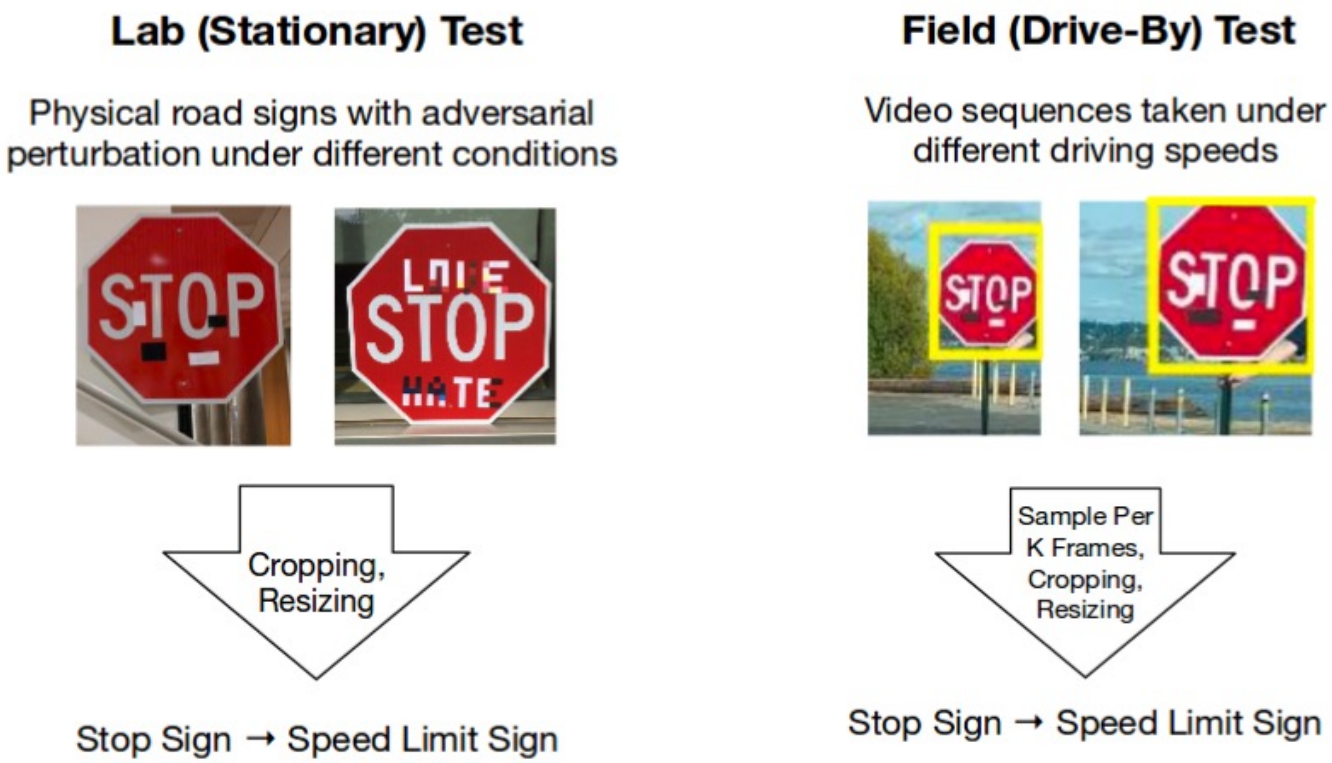


Small adversarial noise

Slide credit: Binghui Wang: Adversarial Machine Learning — An Introduction

Recent work manipulated a stop sign with adversarial patches

- Caused the DL model of a self-driving car to classify it as a Speed Limit 45 sign (100% attack success in lab test, and 85% in field test)



Picture from: Eykholt (2017) - Robust Physical-World Attacks on Deep Learning Visual Classification

Adversarial Examples

A person wearing an adversarial patch is not detected by a person detector model (YOLOv2)

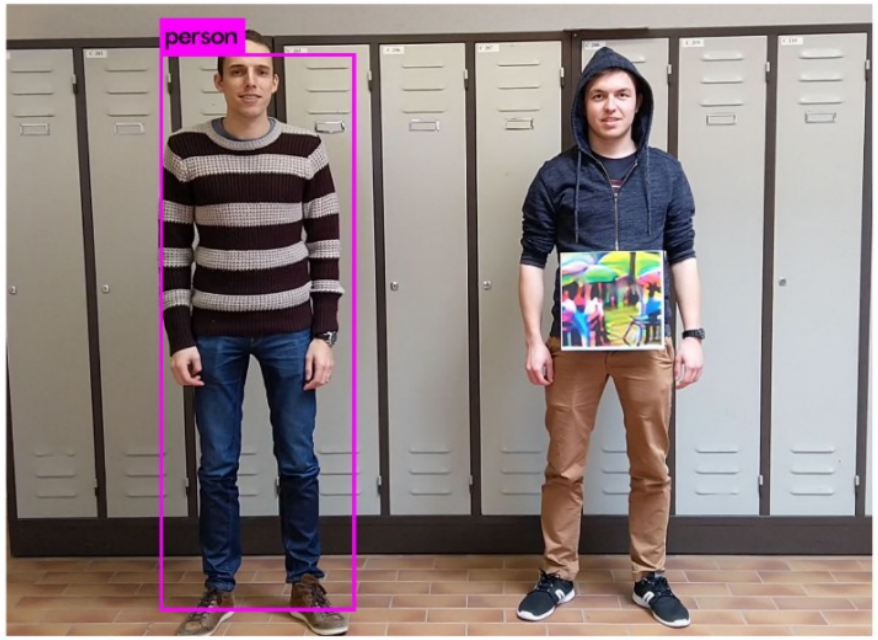


Figure 1: We create an adversarial patch that is successfully able to hide persons from a person detector. Left: The person without a patch is successfully detected. Right: The person holding the patch is ignored.

How to create an adversarial example

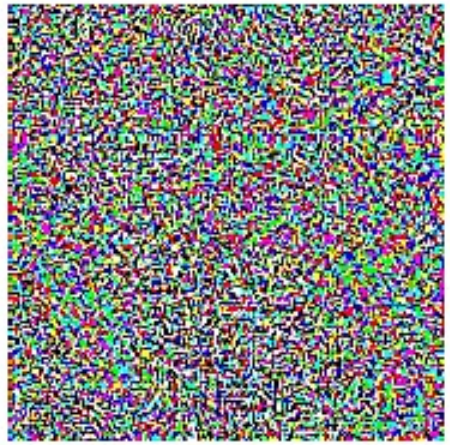
One way: Fast Gradient Sign Method (FSGM) attack [Goodfellow (2015)].

$$\mathbf{x}' = \mathbf{x} + \epsilon \cdot \mathrm{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}, \mathbf{y}))$$

Perturbation noise is calculated as the gradient of the loss function $\mathcal{L}$ with respect to the input image x for the true class label y

This increases the loss for the true class y → the model misclassifies the image

$+ .007 \times$



$=$

x

"panda"
57.7% confidence

$\text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}, \mathbf{y}))$

"nematode"
8.2% confidence

$\mathbf{x} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(\theta, \mathbf{x}, \mathbf{y}))$

"gibbon"
99.3 % confidence

Bias and Gen AI

Ryan Saavedra

@RealSaavedra

Follow

Socialist Rep. Alexandria Ocasio-Cortez (D-NY) claims that algorithms, which are driven by math, are racist

The New York Times

‘Coded Bias’ Review: When the Bots Are Racist

This cleareyed documentary explores how machine-learning algorithms can perpetuate society’s existing class-, race- and gender-based inequities.

Joy Buolamwini is one of the subjects of the documentary “Coded Bias.” 7th Empire Media

By Devika Girish

Nov. 11, 2020

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
5	0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21	28.7
6	0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43	22.9
7	0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15	27.1
8	0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93	16.5
9	0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10	18.9
10	0.22489	12.5	7.87	0.0	0.524	6.377	94.3	6.3467	5.0	311.0	15.2	392.52	20.45	15.0

Boston Housing Data (source: UCI ML datasets)

https://archive.ics.uci.edu/ml/datasets/Housing

Hmmm…

CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000
PTRATIO: Pupil-teacher ratio by town
B: $1000(Bk - 0.63)^2$, where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population
MEDV: Median value of owner-occupied homes in $1000s

Hmmm…

CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000
PTRATIO: Pupil-teacher ratio by town
B: $1000(Bk - 0.63)^2$, where Bk is the proportion of [people of African American descent] by town
LSTAT: Percentage of lower status of the population
MEDV: Median value of owner-occupied homes in $1000s

Q: Is it ok to use to B here?

In general how do we define bias?

- Discrimination on the basis of things (*features*, if you will) that we feel morally should have no bearing

In general how do we define bias?

- Discrimination on the basis of things (*features*, if you will) that we feel morally should have no bearing
- *Especially* for domains in which predictions may have a large impact on individuals (criminal justice, education, housing …)

Legally “protected classes”

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967);**Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

Legally recognized as unsound bases to treat people differently!

Can’t we just withhold features that contain this info?

Can’t we just withhold features that contain this info?

- No: There are often *proxy features* that implicitly capture this
  e.g., zip-code may strongly correlate with race

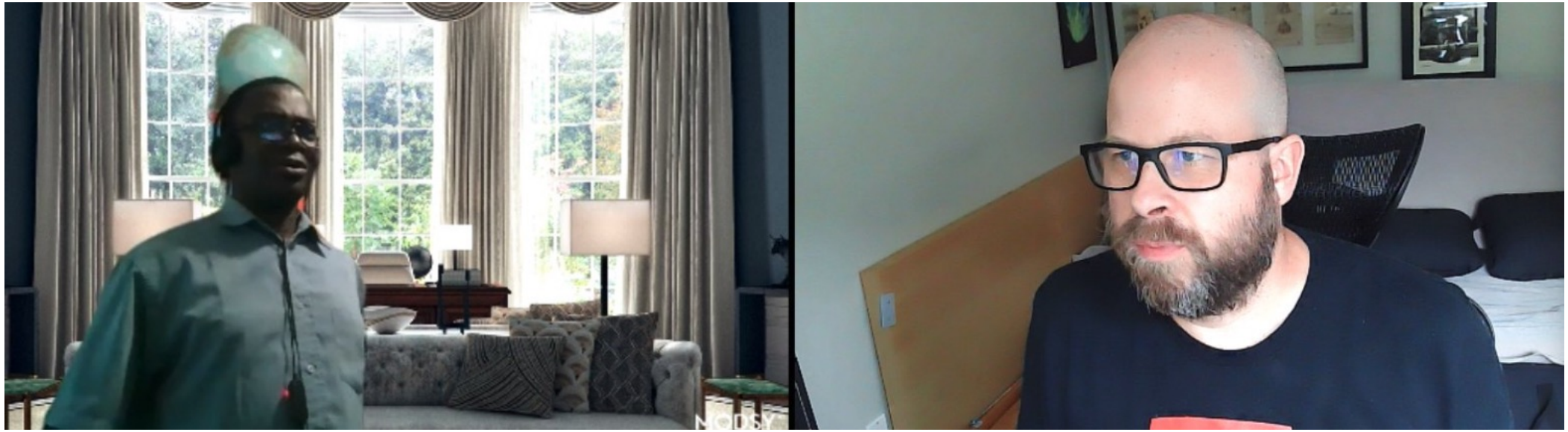
Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.



Zoom…

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

[1]Boston University, 8 Saint Mary's Street, Boston, MA

[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

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

Gender stereotype *she-he* analogies.

sewing-carpentry	register-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	hairdresser-barber

Gender appropriate *she-he* analogies.

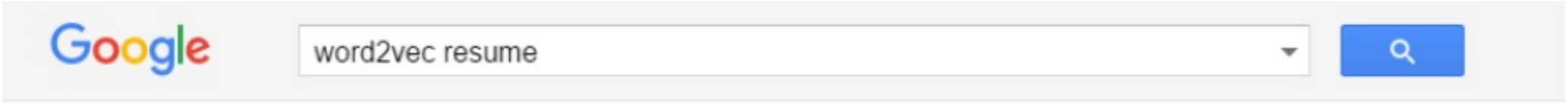
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Extreme *she* occupations

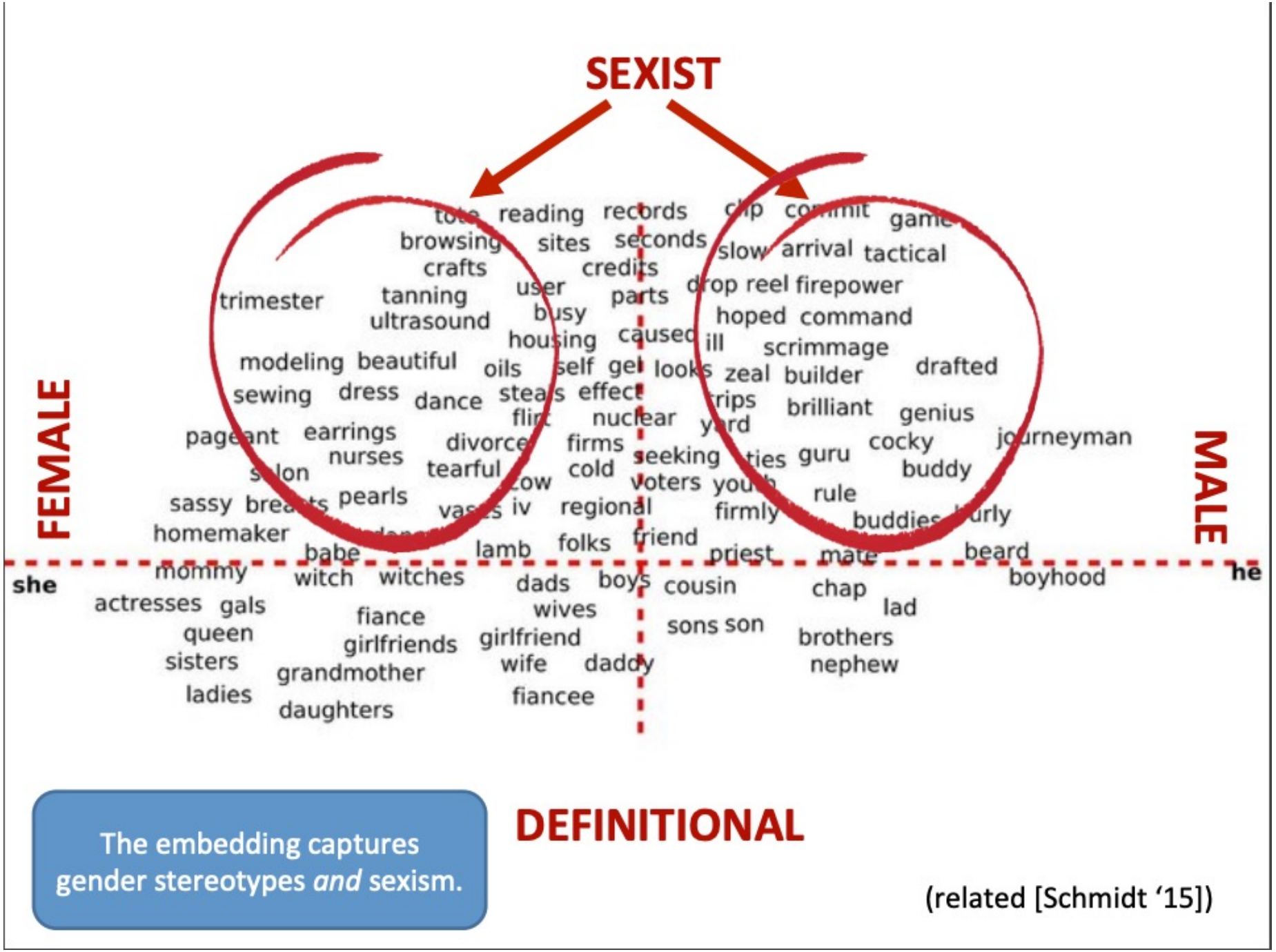
1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

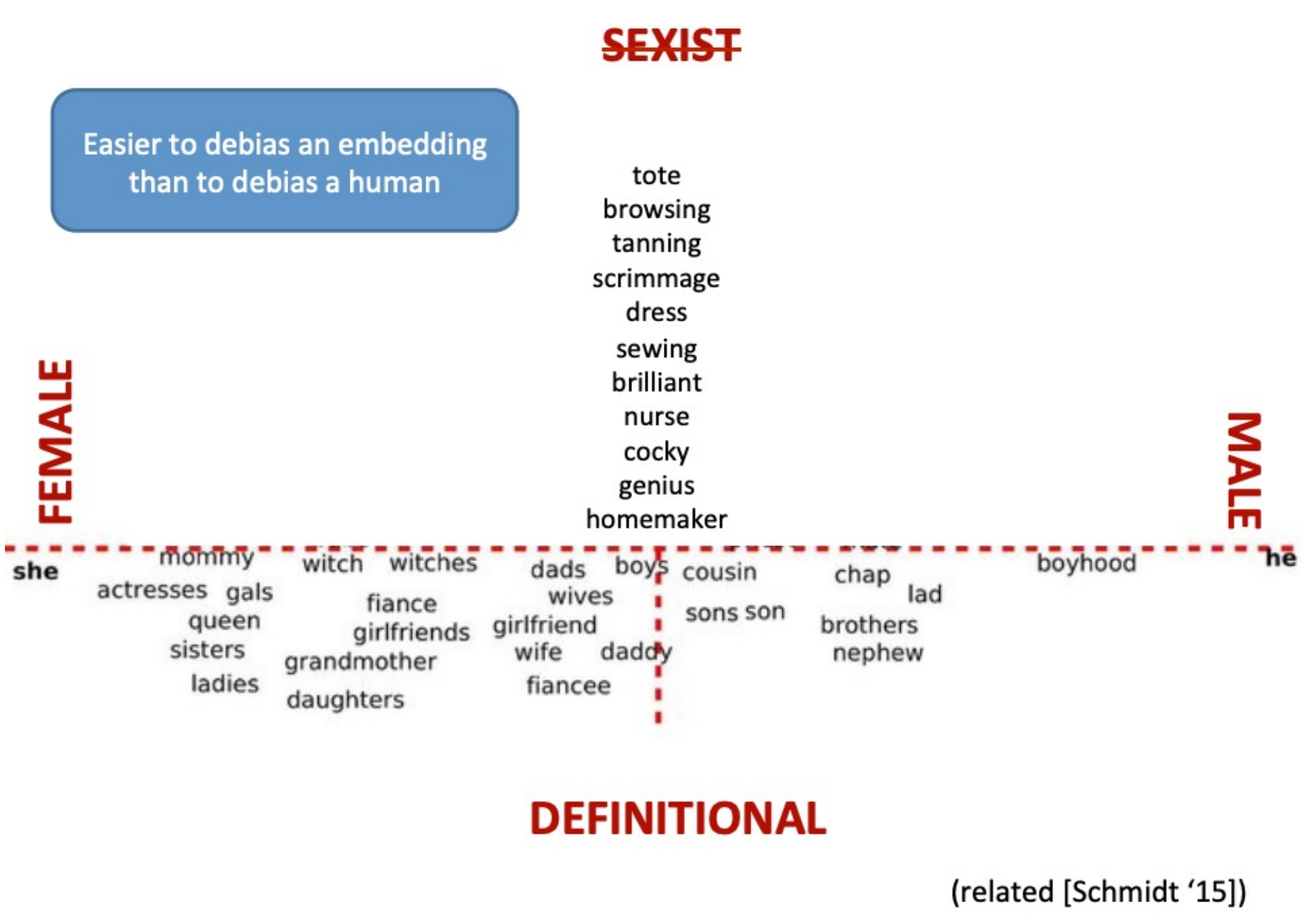
1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss



Bolukbasi et al. '16
Slides: Adam Kalai



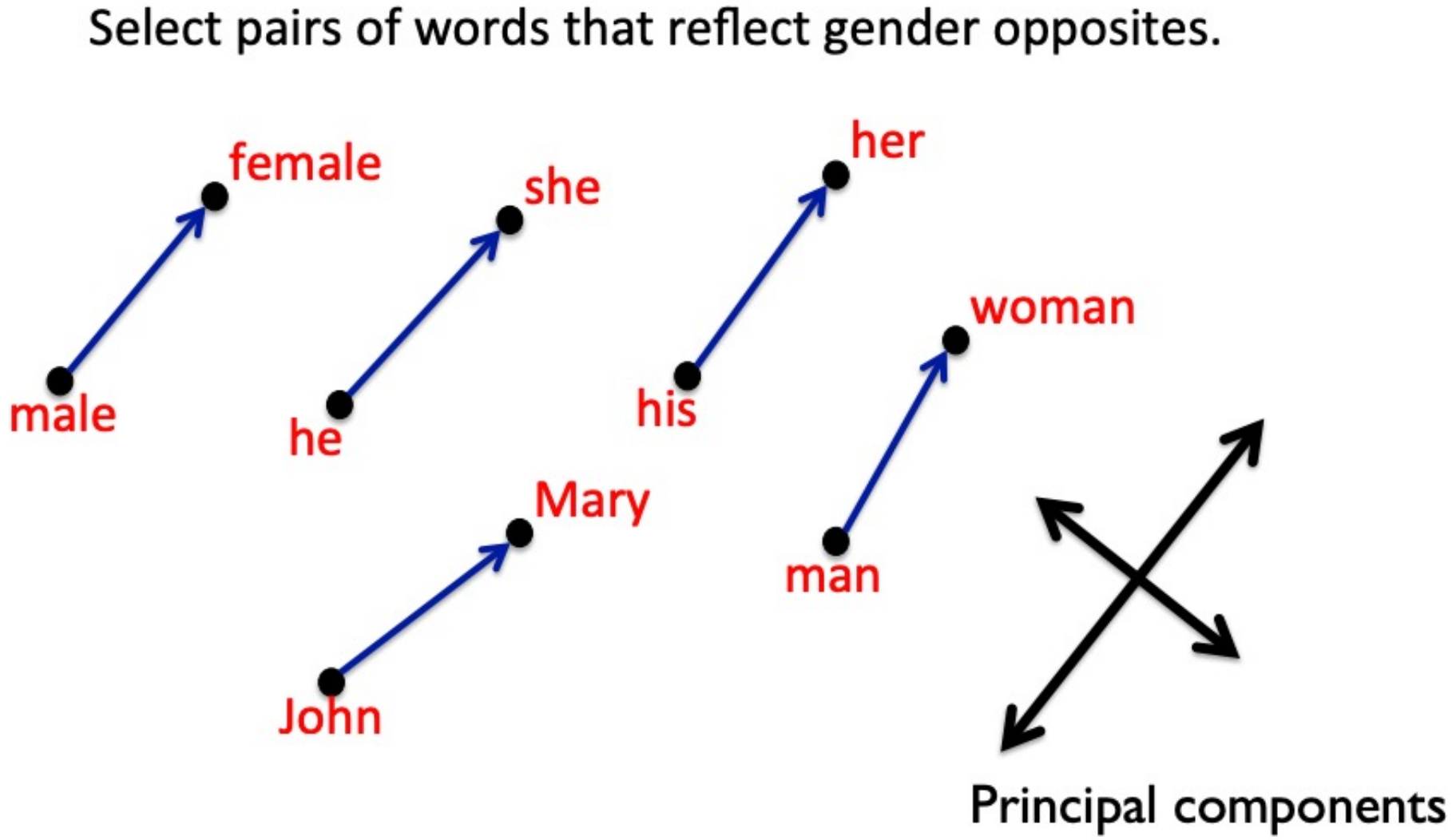
Slides: Adam Kalai



(related [Schmidt '15])

Slides: Adam Kalai

The geometry of gender

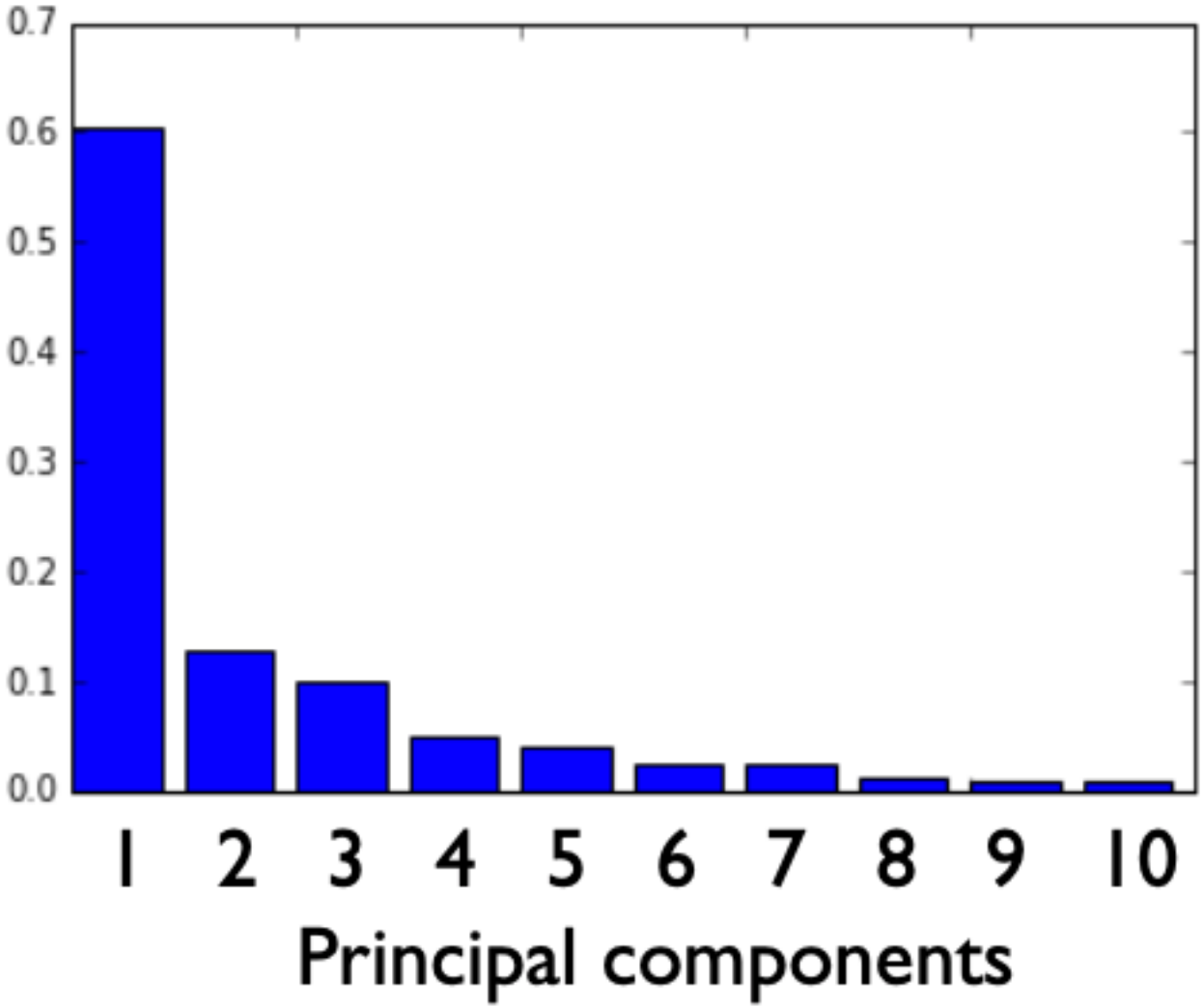


are high, indicating that these pairs capture the intuitive notion of gender.

To identify the gender subspace, we took the ten gender pair difference vectors and computed its principal components (PCs). As Figure 6 shows, there is a single direction that explains the majority of variance

Slides: Adam Kalai

% of variance explained



The top PC seems to capture the gender subspace *B*.

Slides: Adam Kalai

1. Identify words that are gender-neutral N and gender-definitional S.

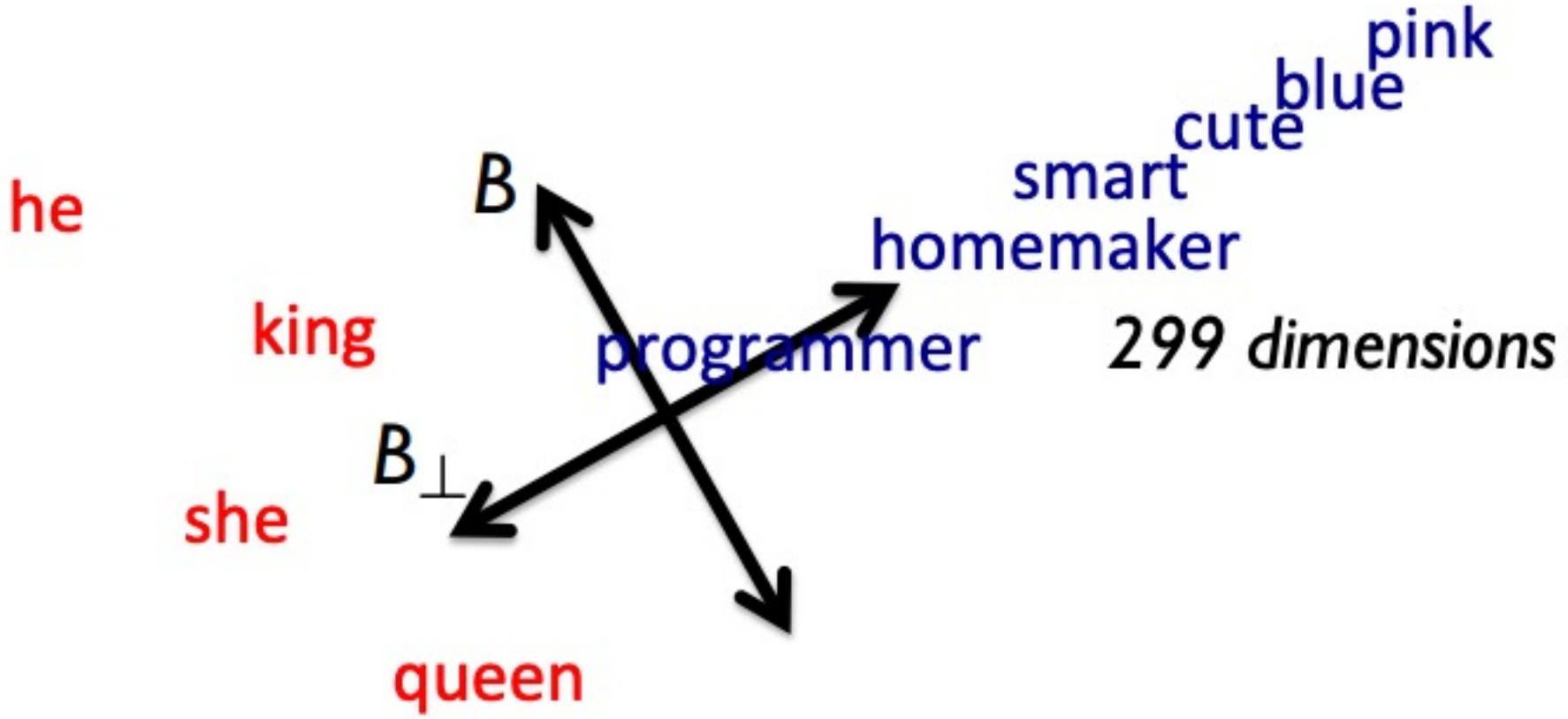
2. Project away the gender subspace from the gender-neutral words.

   $w := w - w \cdot B$ for $w \in N$ B is the gender subspace.

3. Normalize vectors.

Slides: Adam Kalai

De-biasing



ensures that gender neutral words are zero in the gender subspace.

Slides: Adam Kalai

But...

Lipstick on a Pig:
Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them

Hila Gonen[1] and **Yoav Goldberg**[1,2]
[1]Department of Computer Science, Bar-Ilan University
[2]Allen Institute for Artificial Intelligence
{hilagnn,yoav.goldberg}@gmail.com

Gender Bias in Contextualized Word Embeddings

Jieyu Zhao[1], Tianlu Wang[2], Mark Yatskar[3], Ryan Cotterell[4], Vicente Ordonez[2], Kai-Wei Chang[1]

[1]UCLA, [2]University of Virginia, [3]Allen Institute for AI, [4]University of Cambridge

Slide credit to the authors: http://kwchang.net/documents/slides/zhao2019gender_slide.pdf

Zhao et al., 2018

The Woman Worked as a Babysitter: On Biases in Language Generation

Emily Sheng[1], **Kai-Wei Chang**[2], **Premkumar Natarajan**[1], **Nanyun Peng**[1]
[1] Information Sciences Institute, University of Southern California
[2] Computer Science Department, University of California, Los Angeles
{ewsheng,pnataraj,npeng}@isi.edu, kwchang@cs.ucla.edu

Sheng et al., 2019

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Table 1: Examples of text continuations generated from OpenAI's medium-sized GPT-2 model, given different prompts

Sheng et al., 2019

(See colab exercise)