Responsible Machine Learning (Part II)

-Fairness-Aware Machine Learning

Dr. Hongzhi Yin

https://sites.google.com/site/dbhongzhi/

Learn to Assess ML/AI Systems Adversarially

- Research Question
  - Is it ethicial or socially acceptiable?
  - Who could benefit from such a technology? Who can be harmed by such a technology?
- Data Collection
  - Violate user privacy? Could sharing the data have major effect on people's lives?
  - Representativeness of training data? Any biases in the training data?
  - Collected training data reflecting the true distribution in the real world?
- Method
  - Interpretable or explainable? Is the process transparent?
  - Does the system optimize for the "right" objective?
- Evaluation and Performance
  - Is the test set big enough to cover all possible cases?
  - Could prediction errors have major effect on people's lives?

Note that if the social contract is violated at the stage of research question definition, fixing the problem of bias and other problems will not help making this study acceptable.

Fairness

Privacy

Responsible Machine Learning Beyond Accuracy

Transparency

Explainability

Research Topics in the Intersection of Ethics and ML

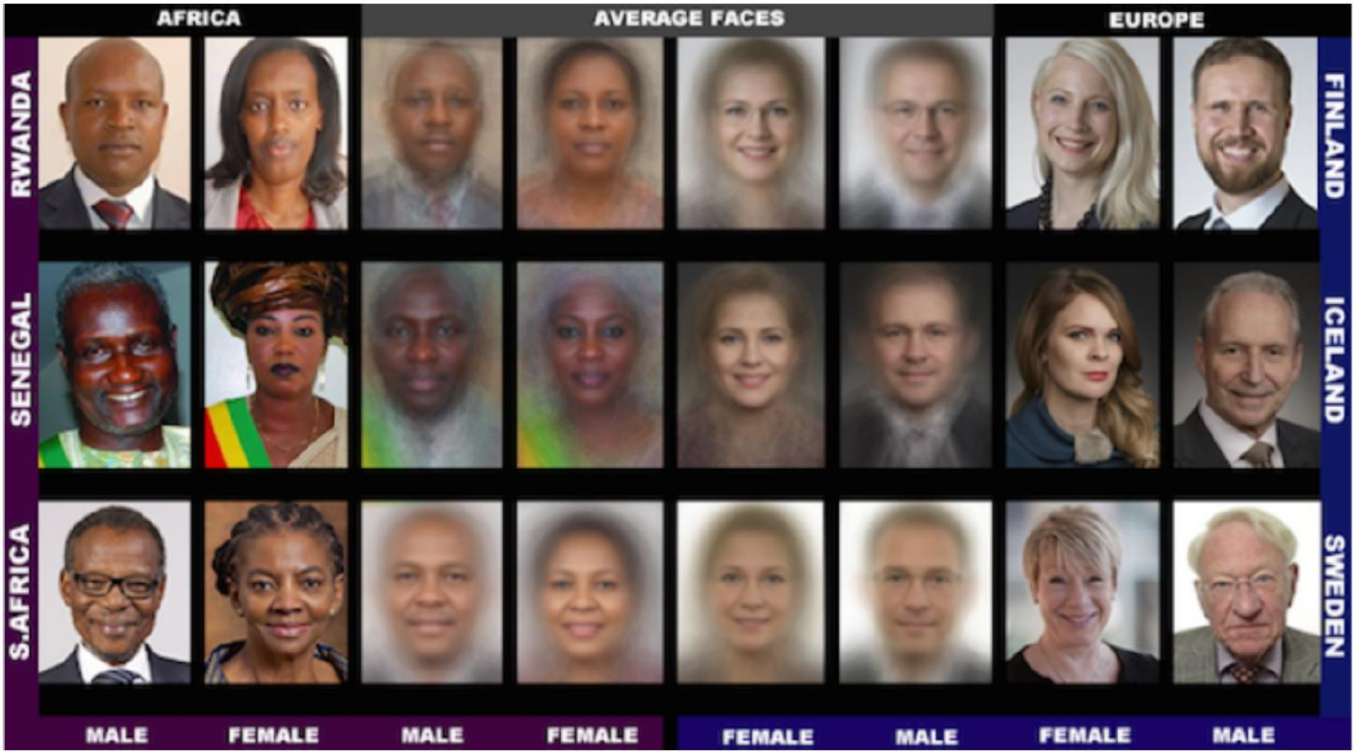
- **Fairness-aware ML**
- Privacy-preserving ML
- Explainable and Transparent ML

Racial Biases

- Face detection software:
  Fails for some darker faces

https://www.youtube.com/watch?v=KB9sI9rY3cA

Gender Shades - Guess your gender from your face



- Facial analysis software:

Higher accuracy for light skinned men

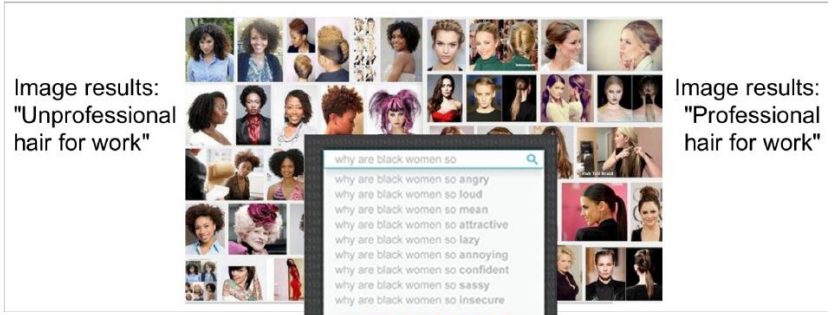
- Error rates for dark skinned women: 20% - 34%

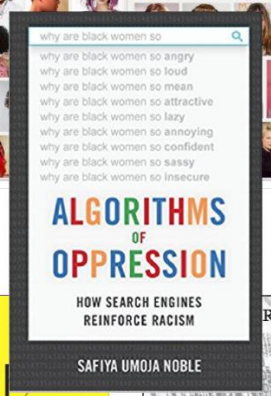
https://www.youtube.com/watch?v=TWWsW1w-BVo

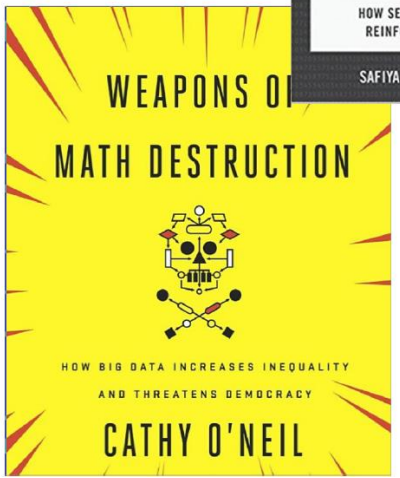
- Ethical challenges posed by AI systems

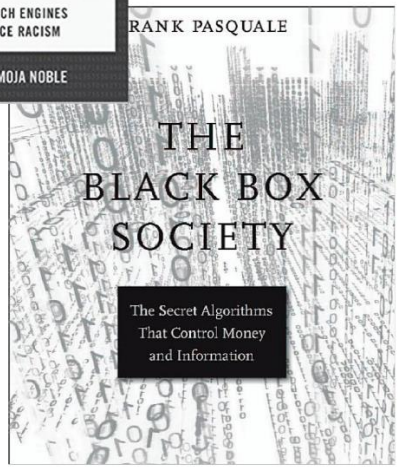
- Inherent biases present in society
  - Reflected in training data
  - AI/ML models prone to amplifying such biases
    - ACM FAT* conference / KDD'16 & NeurIPS'17 Tutorials

Race discrimination









A Note On Terminology

Bias in ML ⇔ Cognitive biases ⇔ Human biases in ML

- Bias in ML or Mathematic Bias
  - Bias of an estimator: the difference between this estimator's expected value and the true value of the parameter being estimated
- Cognitive Biases in Humans (Cognitive Science; Social Psychology)
  - Our brains are evolutionarily hard-wired to store learned information for rapid retrieval and automatic judgments. **Stereotypes naturally form** because of the innate tendency of the human mind to **categorize** the world to simplify processing
- Human Biases in ML
  - Human Biases in **training data**
  - Human Biases in **leanred models**

Outline

- Cognitive Biases (Stereotypes)
- Human Biases in Machine Learning
- How to mitigate biases

Cognitive biases (Stereotypes)

What do you see?

What do you see?

- Bananas

What do you see?

- Bananas
- Stickers

What do you see?

- Bananas
- Stickers
- Dole Bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

...We don't tend to say

Yellow Bananas

What do you see?

Green Bananas

Unripe Bananas

What do you see?

Yellow Bananas

Yellow **is prototypical for bananas**

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

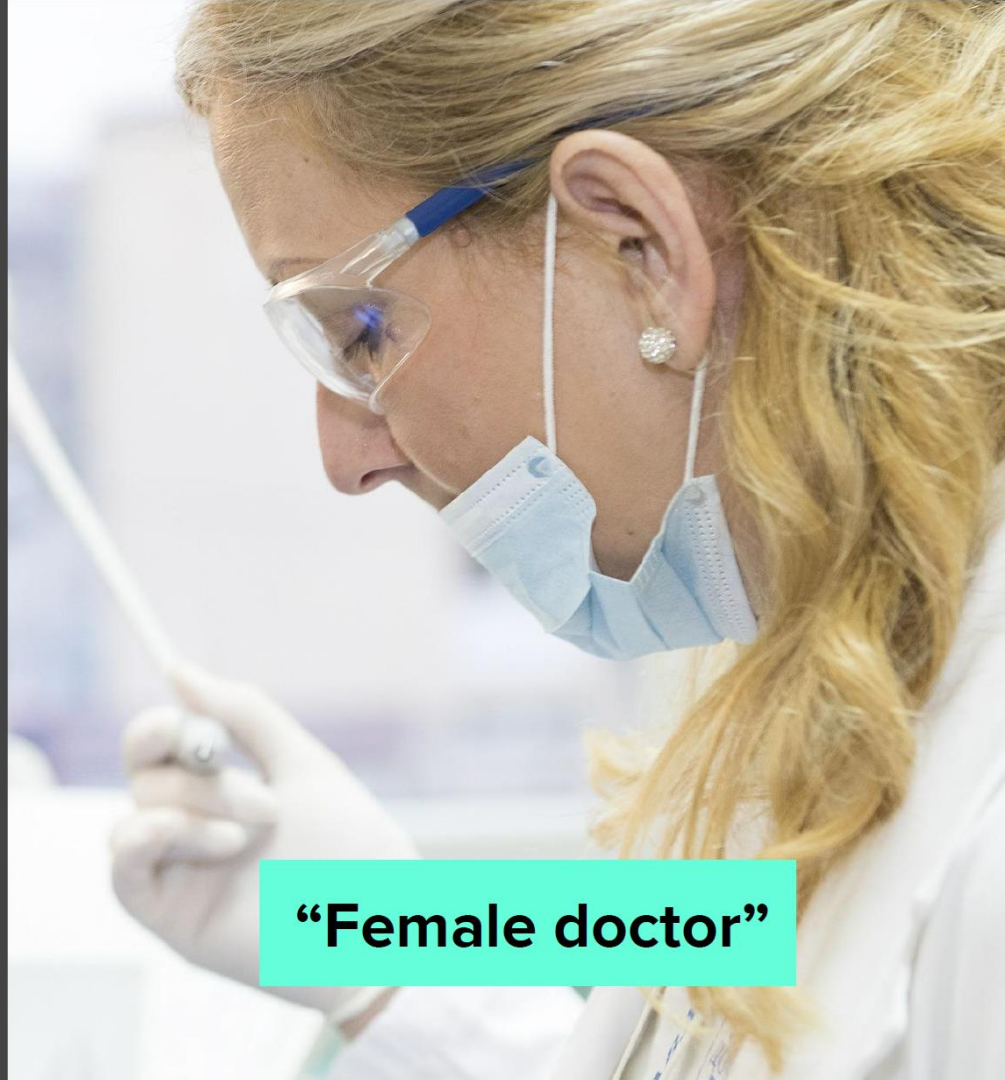
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?



The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

How Do We Make Decisions
How does our cognition work

System 1	System 2
automatic	effortful
fast	slow
parallel	serial
automatic	controlled
effortless	effort-filled
associative	rule-governed
slow-learning	flexible

Kahneman & Tversky 1973, 1974, 2002

~10MP

Our brain constantly receives signals from all sensors: eyes, ears, …
The incoming data is really big data and just the resolution of our eyes is about 10 megapixel.

But actual thinking part of our brain (the system 2 part) can only process the order of 100s of bytes of data at a time. So most of the processing has to happen in system 1- the automatic part.

How Do We Make Decisions

System 1 — automatic

System 2 — effortful

Our brains are evolutionarily hard-wired (neural connections) to store learned information for rapid retrieval and automatic decisions. Over 95% of cognition is delegated to the System 1 “auto-pilot.”

There’s an interplay between the two systems. System 1 works all the time in the background; System 2 only makes hard decisions for unusual settings.

We believe that we make conscious choices and reasoning but actually most of the time the one that is really in charge is system 1

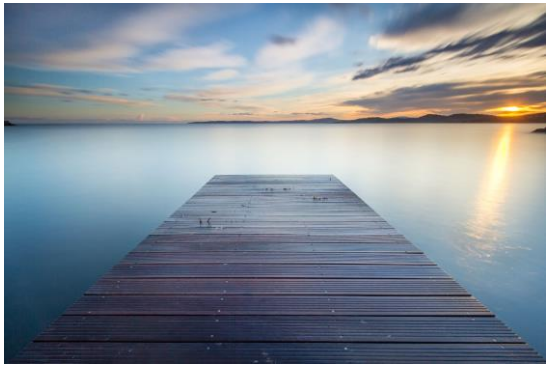
Psychological Perspective on Stereotypes

Stereotypes naturally form because of the innate tendency of the human mind to:

- Categorize the world to simplify processing in System 1
    - One purpose of categorization is to reduce the infinite differences among stimuli/objects to limited category-level differences

- Store learned information in mental representations (called schemas)
- Automatically and unconsciously activate stored information whenever one encounters a category member

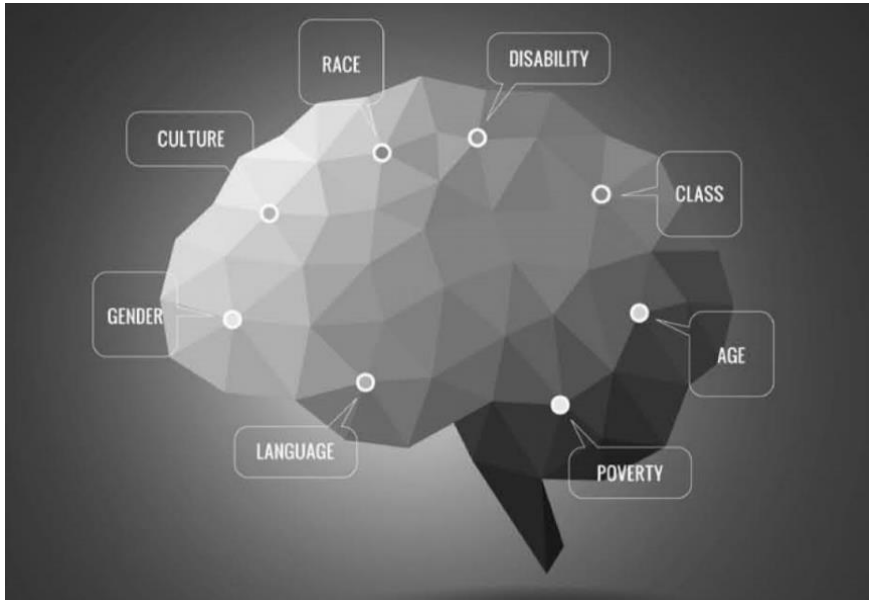
Test your first impression/feeling

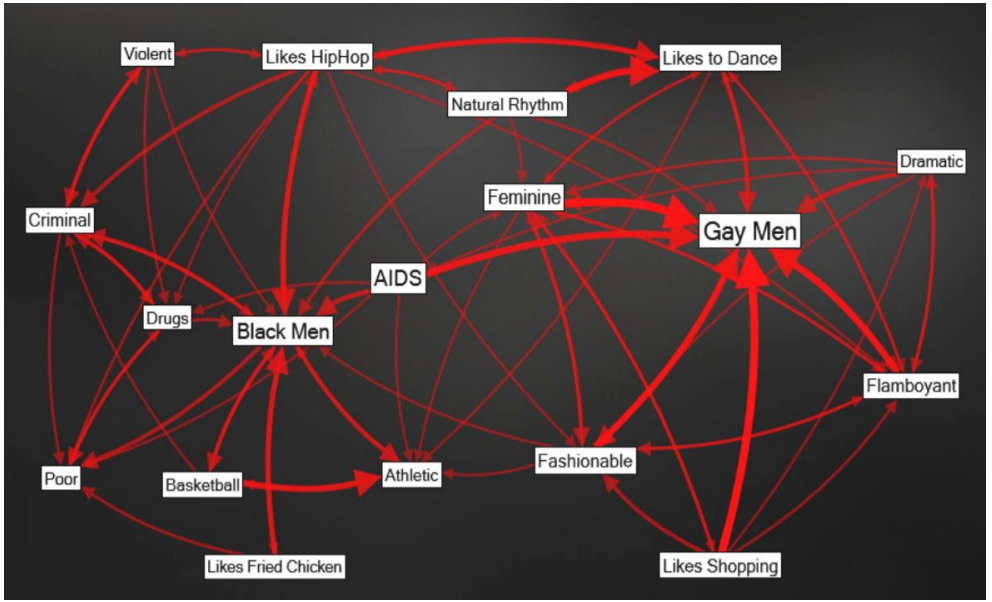


Thanks to these stereotypes we can understand quickly that this is calm, cute, tasty

What is your first reaction?

Our autopilot/System 1 will assume that the snake is dangerous, and will stop us from touching it.

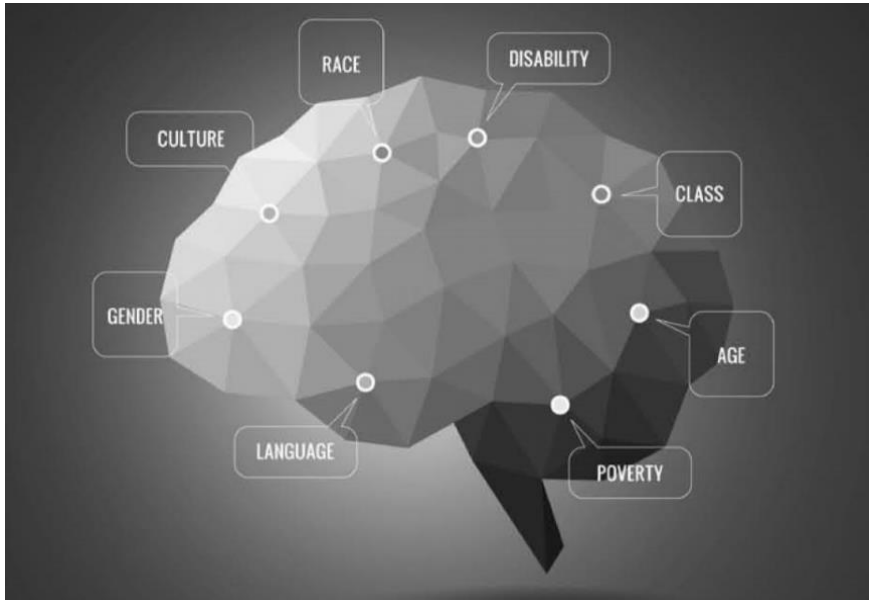


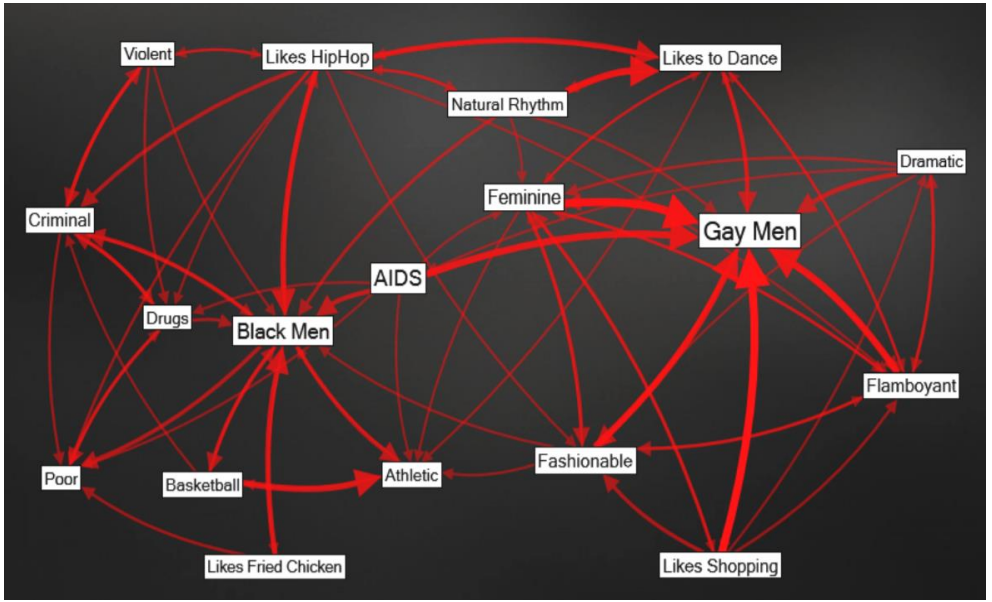


[Image credit: Geoff Kaufman]

Stereotypes are internalized as **associations** through natural processes of learning, categorization and practicing.

Social stereotypes are not necessarily negative, but still have negative effect.





[Image credit: Geoff Kaufman]

Cognitive biases are pervasive, operate largely unconsciously, and can automatically influence the ways in which we see and treat others, even when we are determined to be fair and objective.

Human Biaes in Machine Learing

human/social biases in training data and ML models

AI/ML Online Applications and Online training datasets

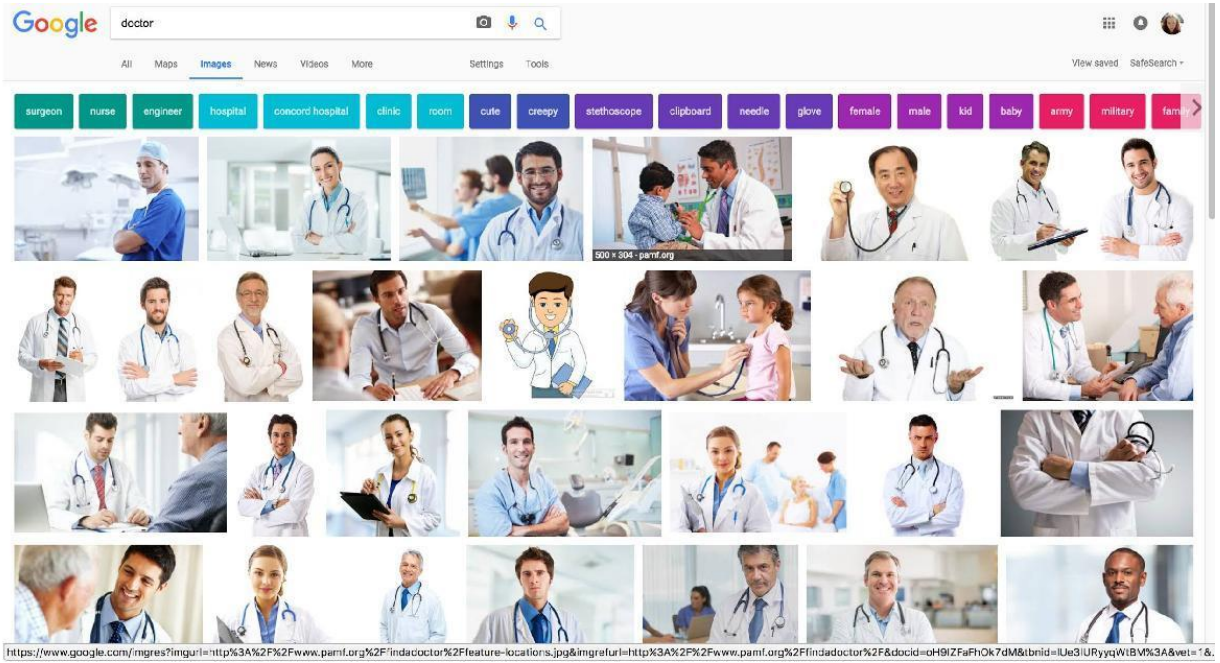
- Conversational agents
- Personal assistants
- Search engines
- Recommndation engines
- Translation engines
- Medical research assistants

Online data is generated by humans and humans have bias and stereotypes.

Online data is riddled with **SOCIAL STEREOTYPES**

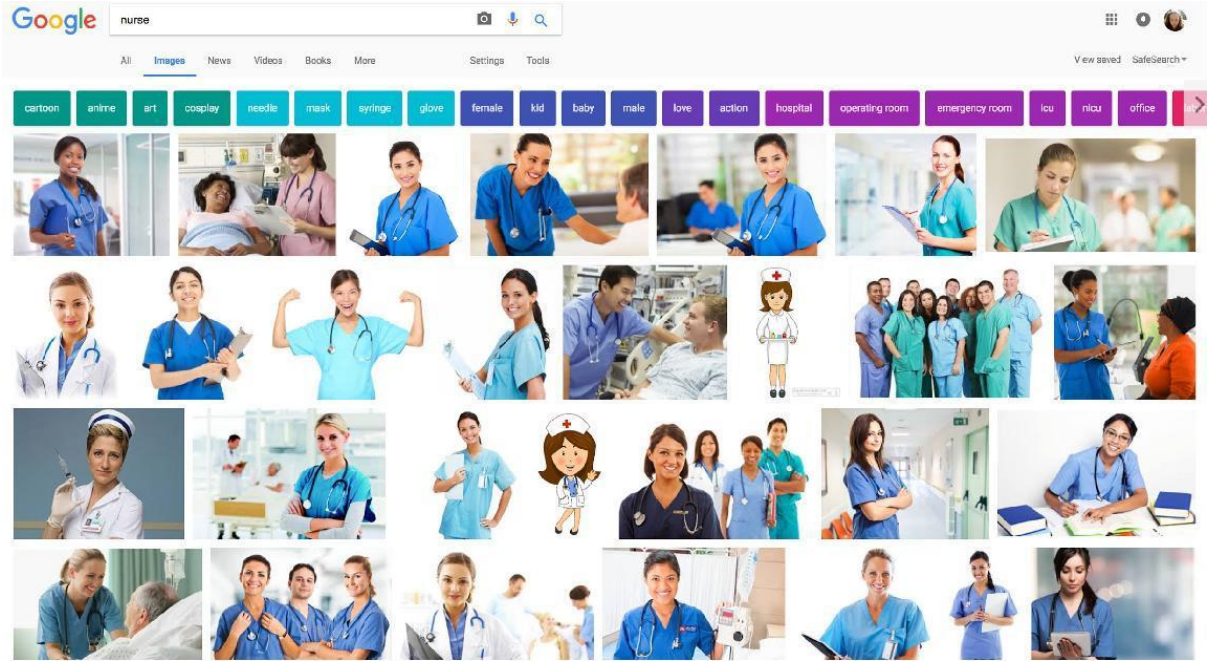
Gender/Race/Age Stereotypes

- June 2017: image search query “Doctor”



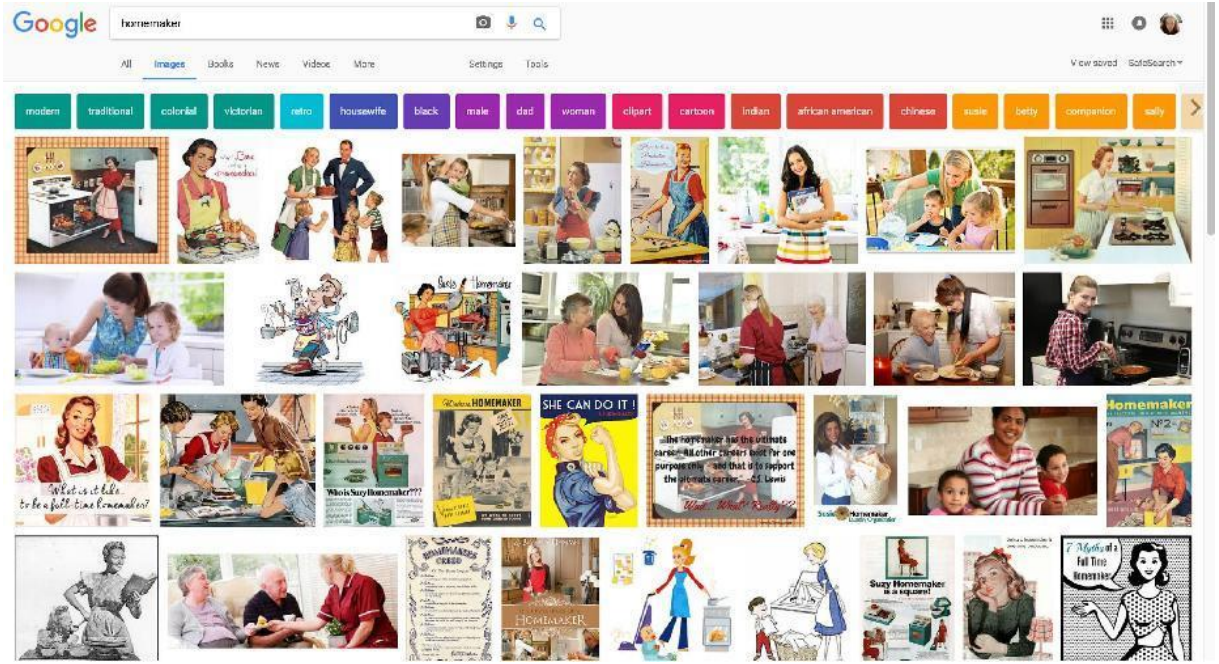
Gender/Race/Age Stereotypes

- June 2017: image search query “Nurse”



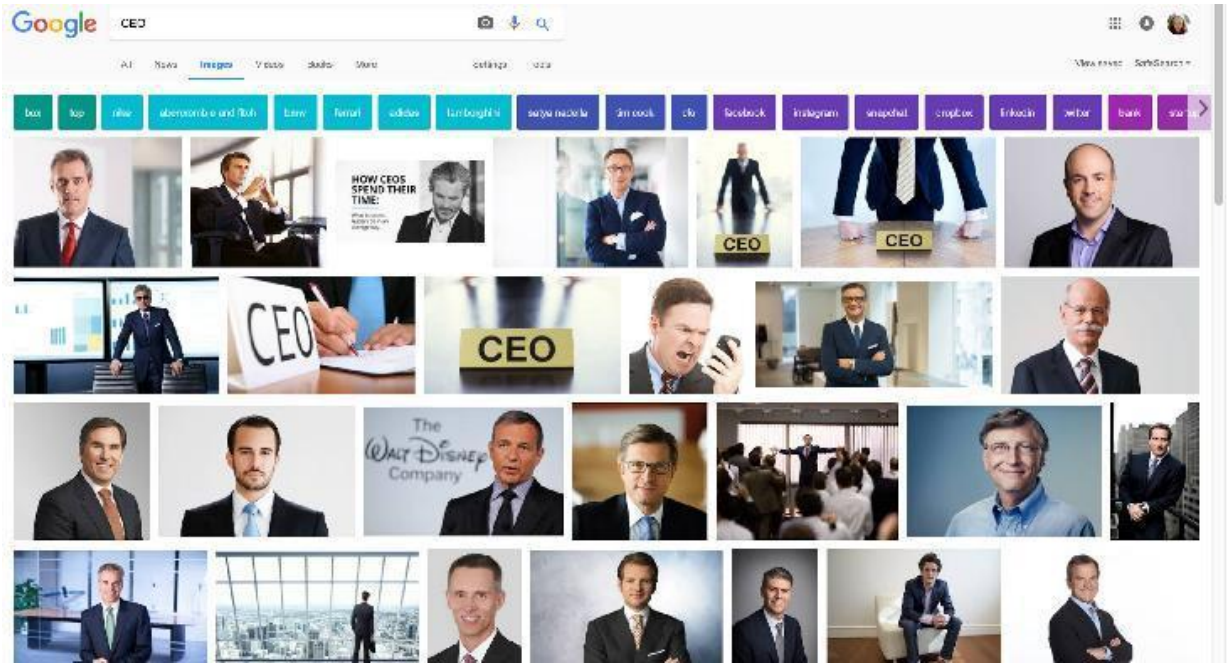
Gender/Race/Age Stereotypes

- June 2017: image search query “Homemaker”



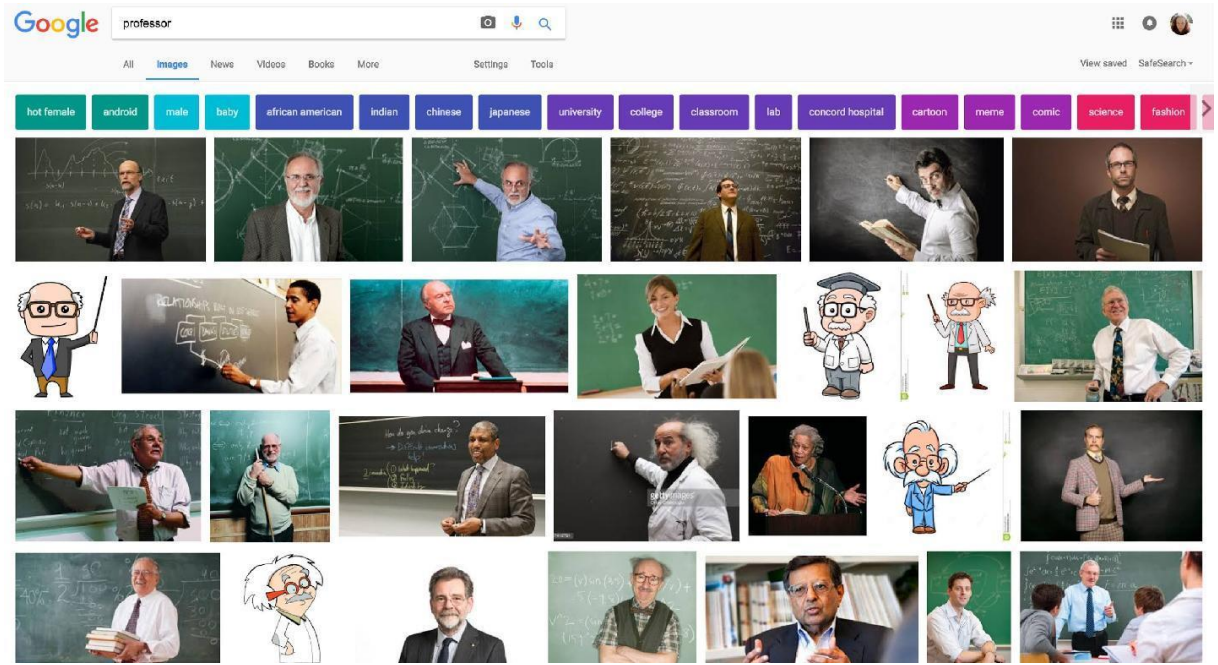
Gender/Race/Age Stereotypes

- June 2017: image search query “CEO”



Gender/Race/Age Stereotypes

- June 2017: image search query “Professor”



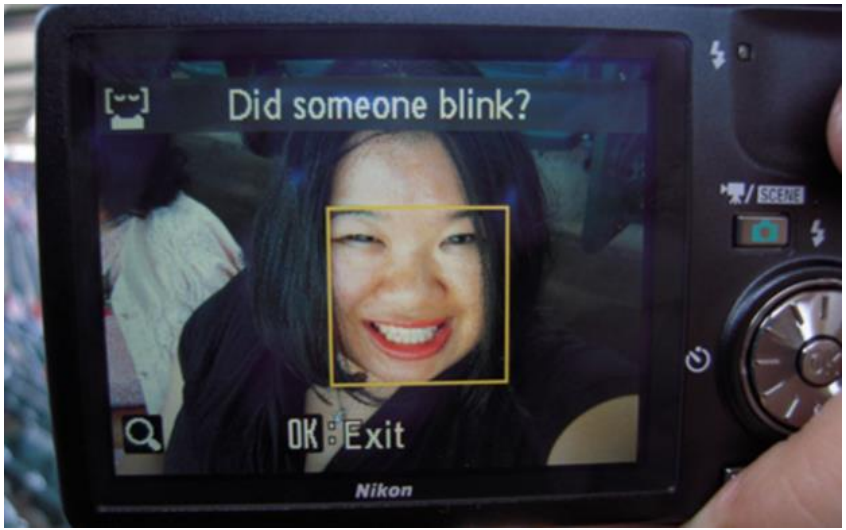
Gender Biases in Online Data

- Men are over-represented in the reporting of web-based news articles
- Men are over-represented in twitter conversations
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues
- IMDB reviews written by women receive lower ratings for usefulness

So many biases exist in the online content/data used to train ML models.

Consequence: ML models are biased

Biased ML Application

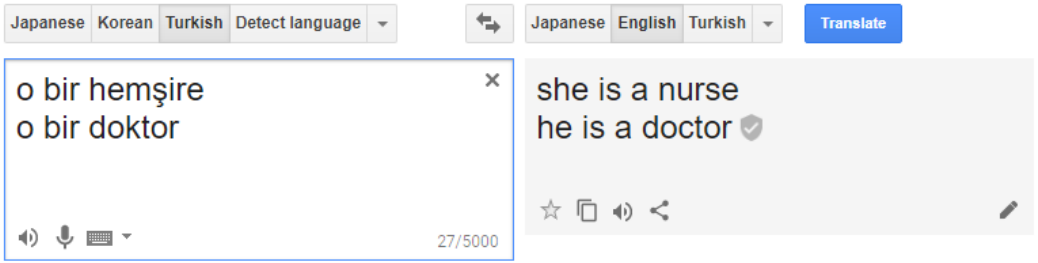


Nikon launched a feature that warns the photographer whenever someone blinked on the picture. Some people reported it was detecting false blinks on **Asian people pictures**.

Although it is a Japanese company, it seems to have used biased training data in its ML models.

Biased ML Application

- Natural language data and annotations will reflect social/cognitive biases
- ML algorithms will replicate biases present in their training data

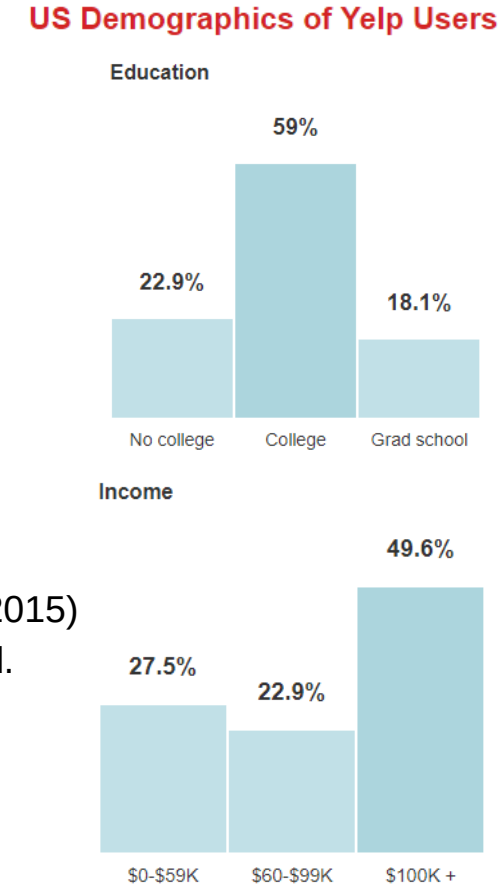


https://translate.google.com.au

Human Biases in Training Data

Types of Sampling Bias in Naturalistic Data

- Self-Selection Bias
  - Some people decides to post reviews on Yelp while others llike to write Google reviews; It is not a random process.
- Reporting Bias
  - People do not necessarily talk about things in proportion to their real-world distributions (Gordon and Van Durme 2013)
- Community / Dialect / Socioeconomic Biases
  - What communities are over- or under-represented? leads to community-specific model performance (Jorgensen et al. 2015)
  - Minority groups and poor communites are often under- represented.



Reporting Bias

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

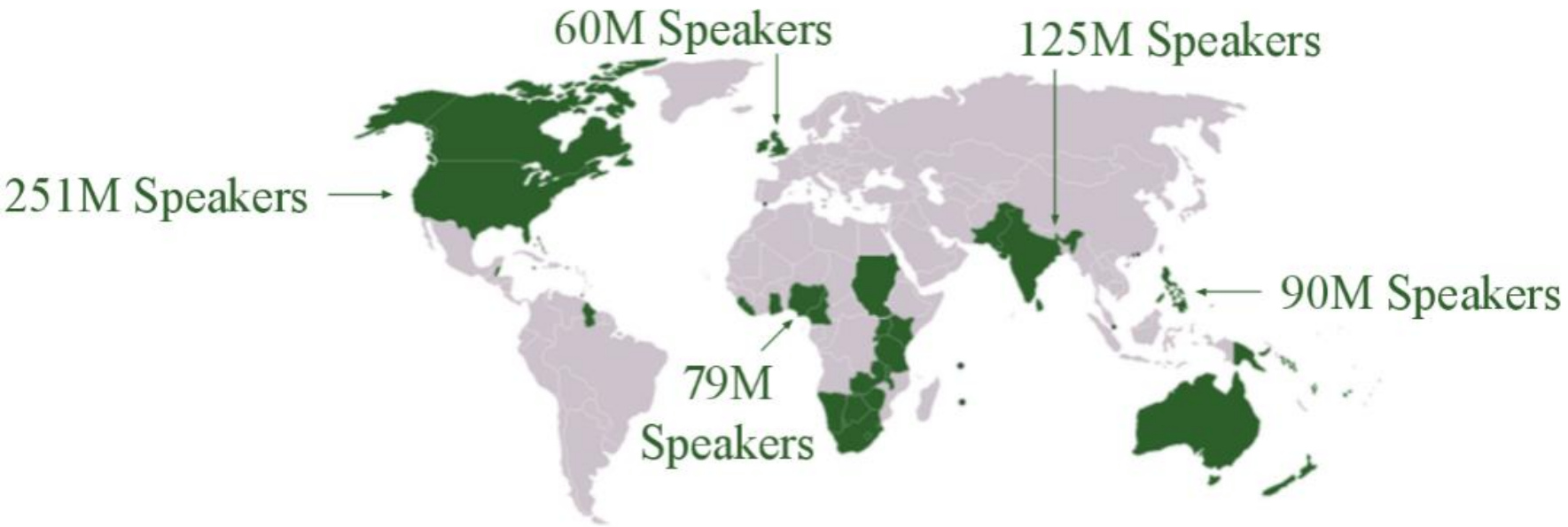
Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

Community Bias in Language Identification

- Most applications employ off-the-shelf LID systems which are highly accurate
- LID system examines the extracted text of each document to determine the primary language and up to two secondary languages present. It also returns their approximate percentages of the total text bytes.

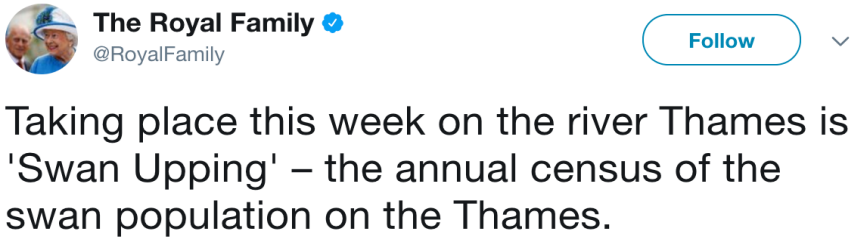
World Englishes

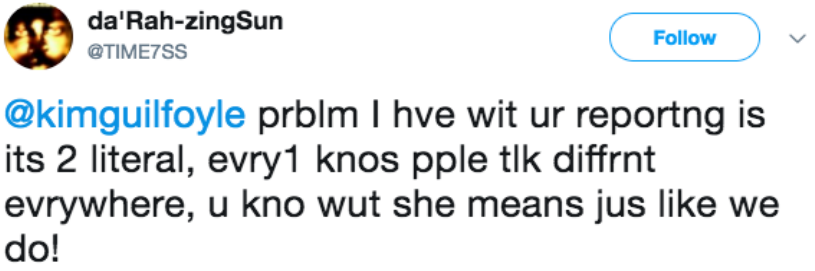


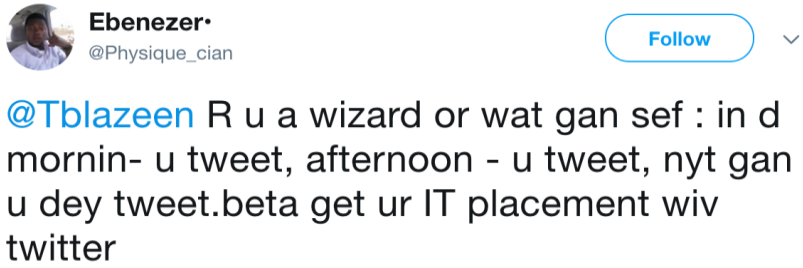
World Englishes

Is the data we use to train our English NLP models representative of all the Englishes out there?

World Englishes





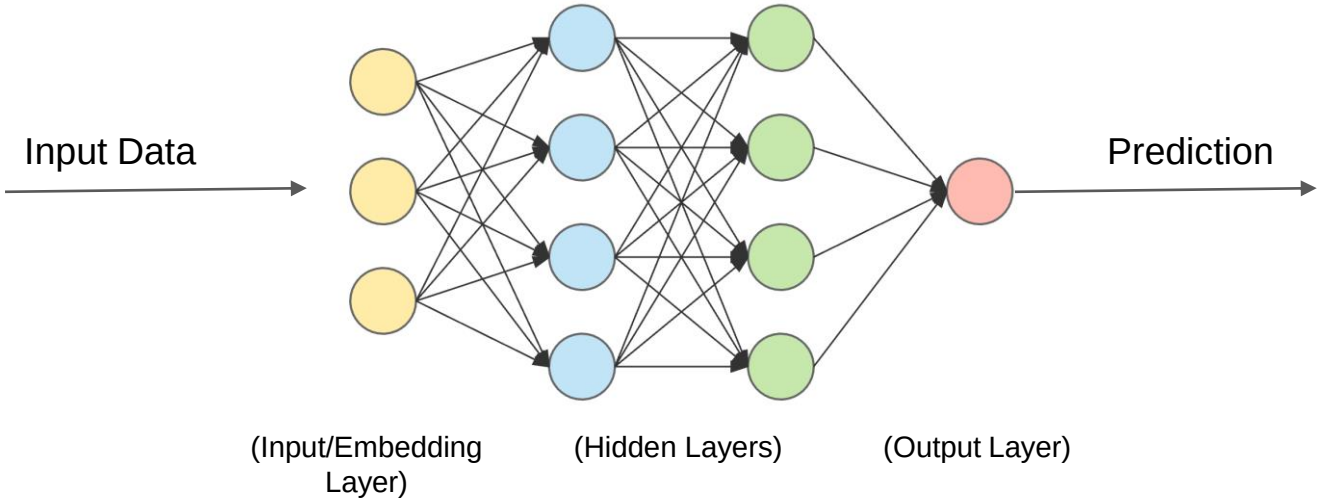


- Language identification degrades significantly on African American Vernacular English (Blodgett et al. 2016)

Human Biases in Machine Learning Models

-Is my ML model capturing social stereotypes?

Where to look for biases?



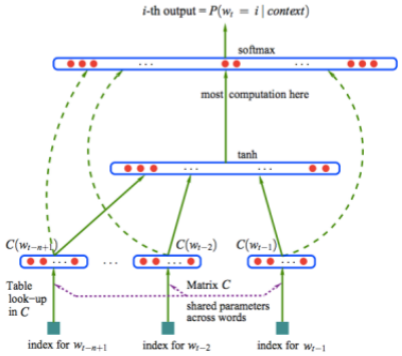
Bias in Input Representations

How to represent the input data in current DNN/DL techniques?

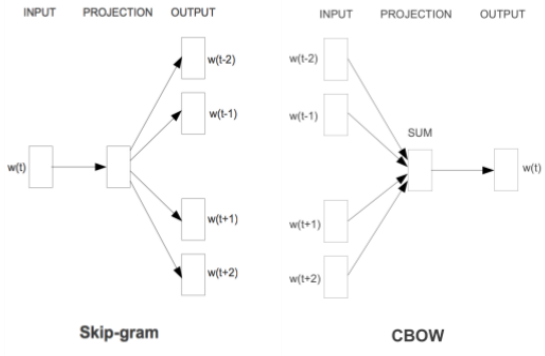
Input Representation: How to represent words?

Word embedding is one of the most popular representation of document vocabulary.
It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc.

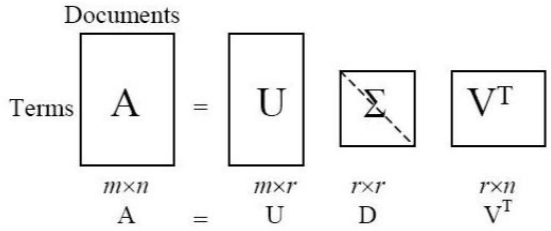
Loosely speaking, word embeddings are **vector representations** of a particular word.



Neural Language Model (Bengio et al, `03)

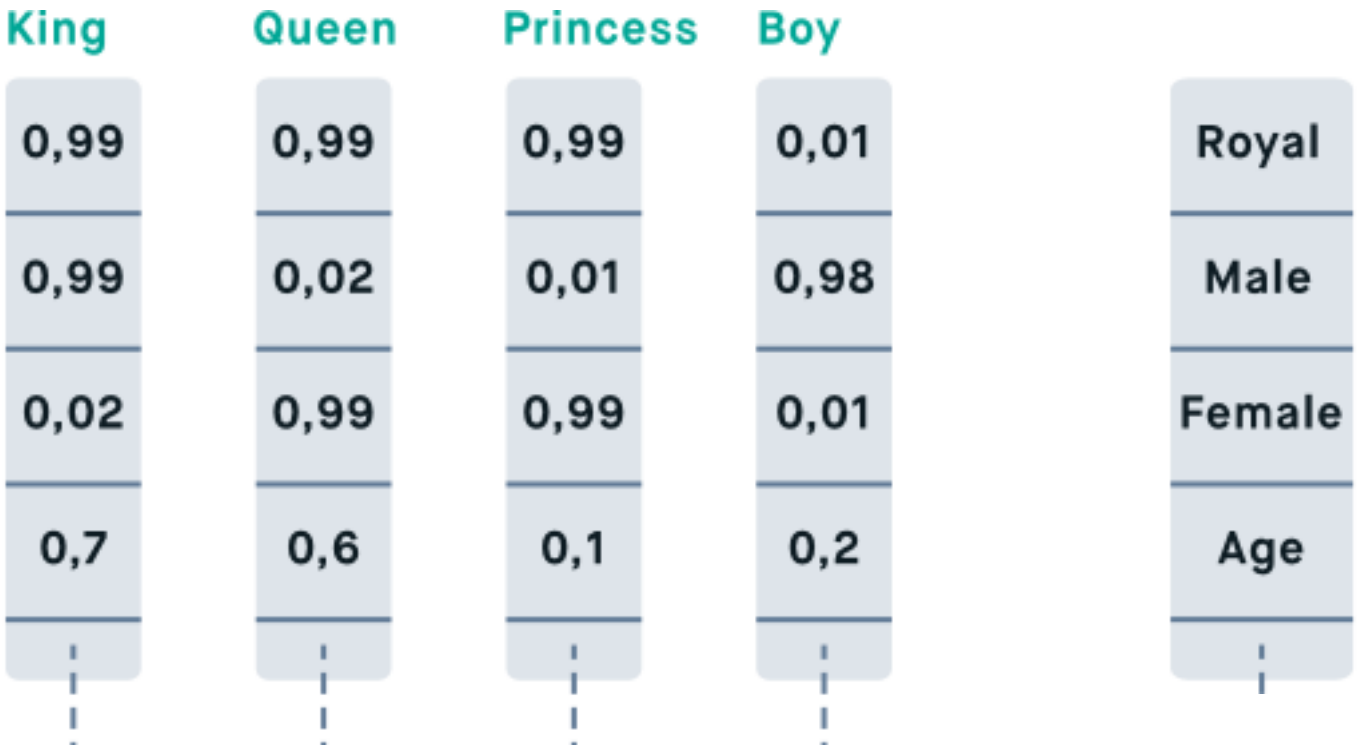


word2vec (Mikolov et al, `03)



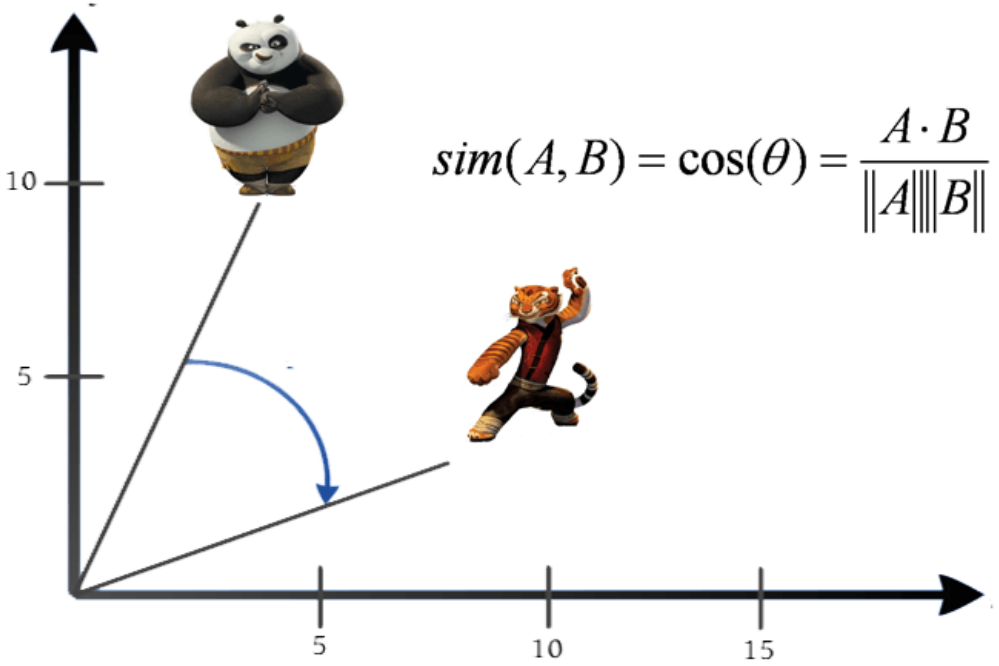
Latent Semantic Analysis
(Deerwester et al, `90, Turney & Pantel `10)

Example of Word Embedding



How to measure word similarity?

Cosine Similarity

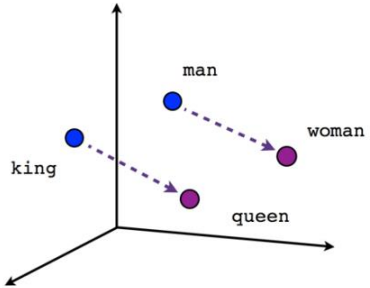


$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

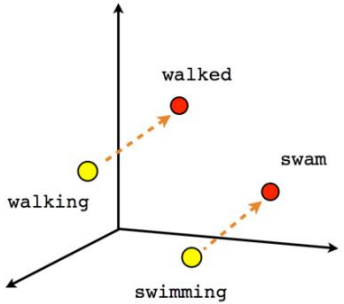
Word Similarity Examples

tiger	tiger	1.0
king	cabbage	-0.2
apple	orange	0.5
...	...	...
method	approach	0.7

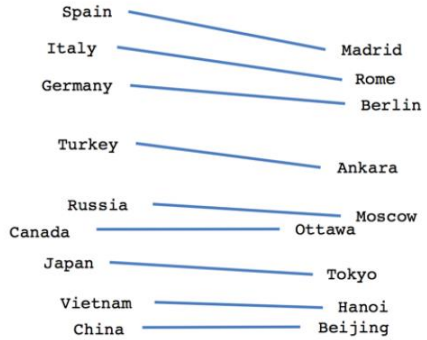
Word Analogy Tasks



Male-Female



Verb tense



Country-Capital

- Mikolov et al. ‘13
  - $\overrightarrow{man}$ :: $\overrightarrow{king}$ ~ $\overrightarrow{woman}$ :: ?

Man is to a King as a women is to ?

$$\max cos(man - woman,\ king - x)\ s.t.\ ||king - x||_2 < \delta$$

Do word embeddings capture human-like biases?

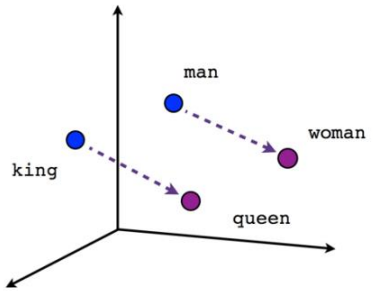
How to Detect Such Bias in Word Embedding?

Bolukbasi et al. (2016)

Gender Bias in Word Embeddings

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

$$\max cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$



Male-Female

surgeon vs. nurse

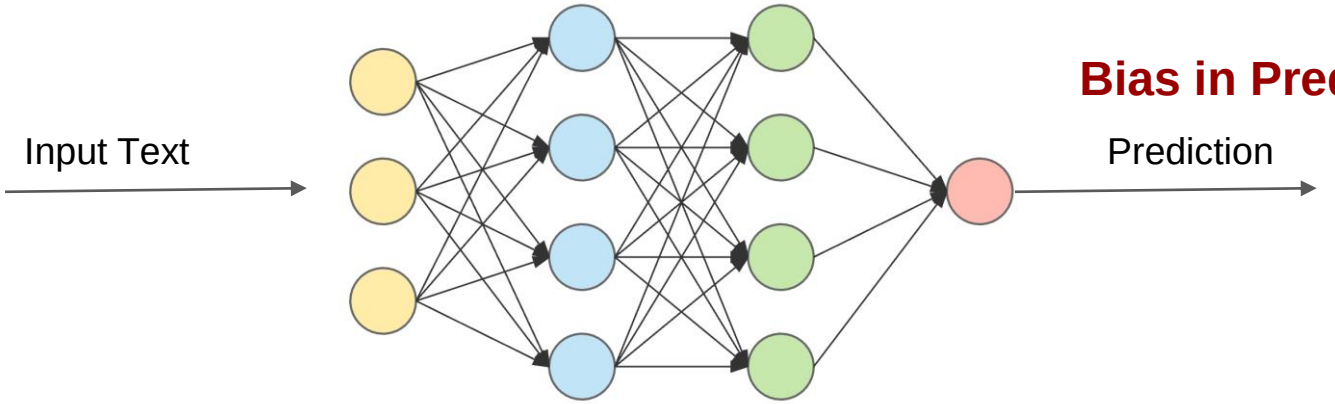
architect vs. interior designer

shopkeeper vs. housewife

superstar vs. diva

....

Bias in Predictions



Bias in Predictions?

Bias in Input Representations

Zhao et al. (2017)

imSitu

Browse Demo Publications Download

http://imsitu.org/

imSitu is a dataset supporting **situation recognition**, the problem of producing a concise summary of the situation an image depicts including: (1) the main activity, (2) the participating actors, objects, substances, and locations and most importantly (3) the roles these participants play in the activity. The role set used by imSitu is derived from the linguistic resource FrameNet and the entities are derived from ImageNet. The data in imSitu can be used to create robust algorithms for situation recongntion.

Situations

Click image

attacking			
agent	victim	weapon	place
elephant	hippo	trunk	outside

imSitu Dataset

verbs	504
images	126,102
situations per image	3
total annotations	1,481,851
unique entity types (>3)	11,538 (6,794)
unique roles (role types)	1,788 (190)
images per verb (range)	250.2 (200 - 400)
unique situations (>3)	205,095 (21,505)

Details

Supported by

Press

New York Times: Computer Vision: On the Way to Seeing More

Zhao et al. **Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraint.** *EMNLP* (2017)

Zhao et al. (2017)

imSitu Visual Semantic Role Labeling (vSRL)

Internet

FrameNet

WordNet

COOKING	
ROLES	NOUNS
AGENT	woman
FOOD	vegetable
CONTAINER	pot
TOOL	spatula

Participants

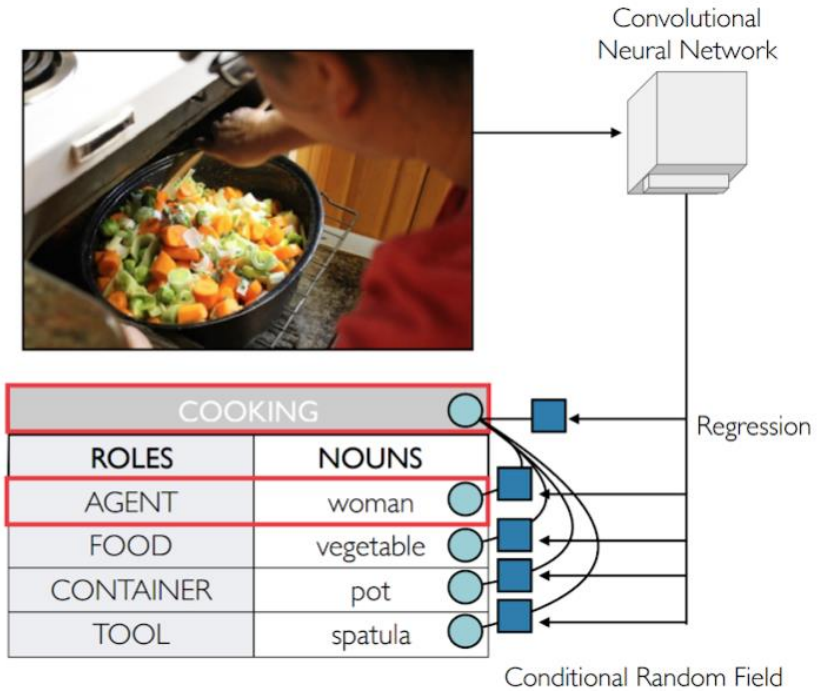
12

Yatskar et al. CVPR '16, Yang et al. NAACL '16, Gupta and Malik arXiv '16

Slides by Mark Yatskar https://homes.cs.washington.edu/~my89/talks/ZWYOC17_slide.pdf

Zhao et al. (2017)

imSitu Visual Semantic Role Labeling (vSRL)



imSitu Visual Semantic Role Labeling (vSRL)

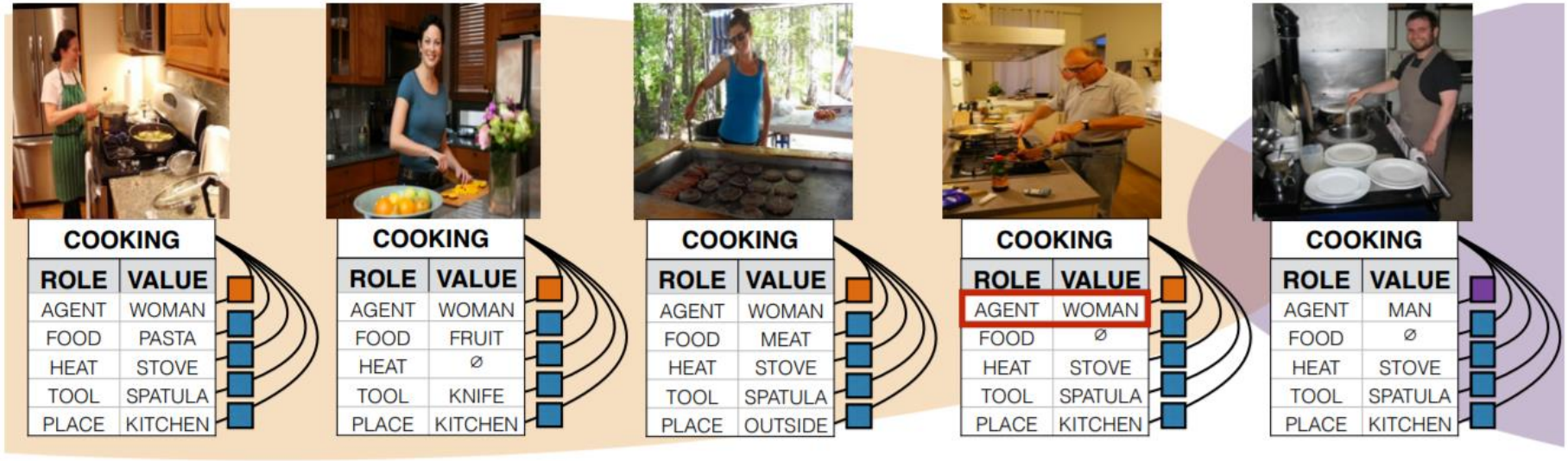
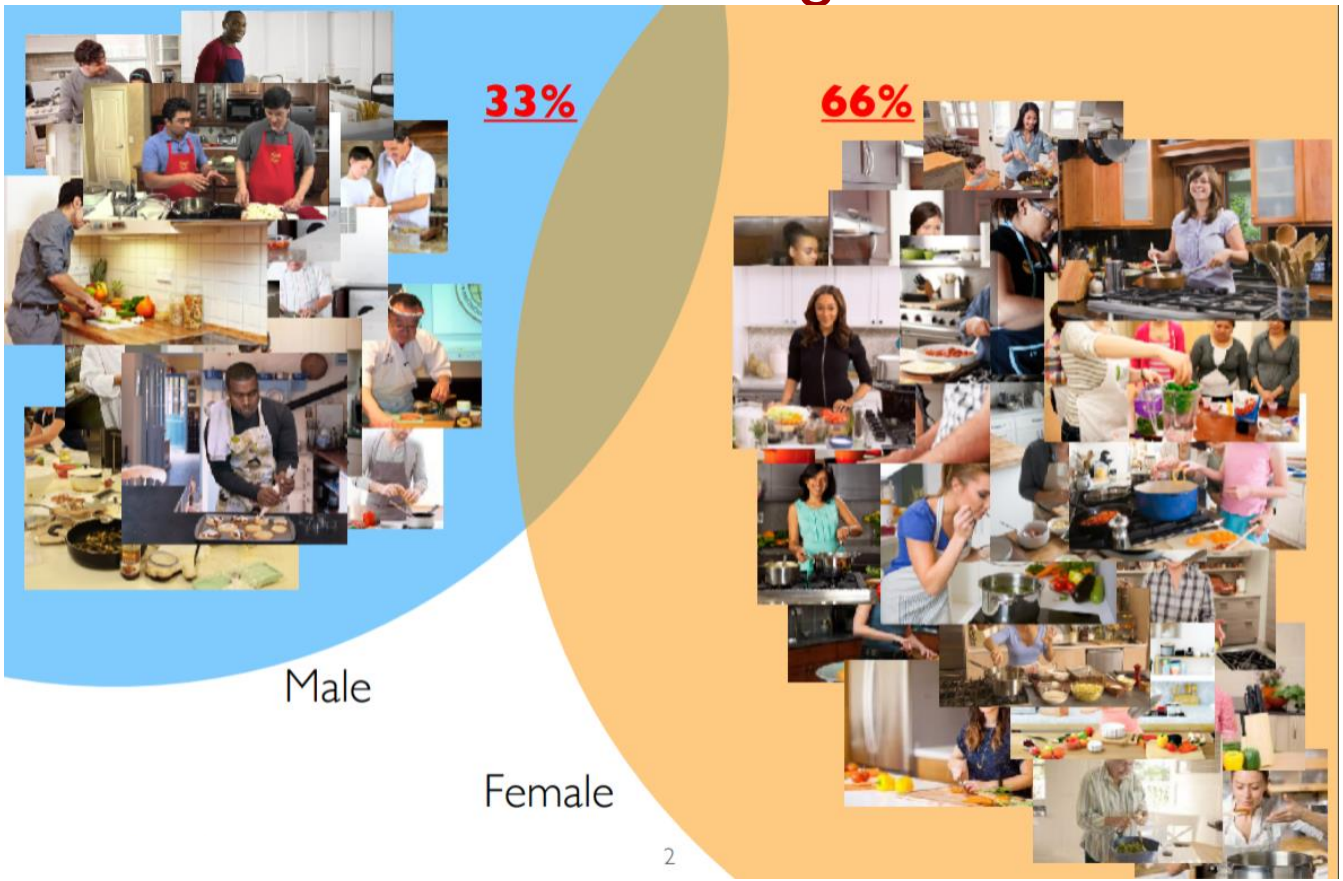


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images have `man` in the `agent` role while the rest have `woman`. After training a Conditional Random Field (CRF), bias is amplified: `man` fills 16% of `agent` roles in `cooking` images.

Zhao et al. (2017)

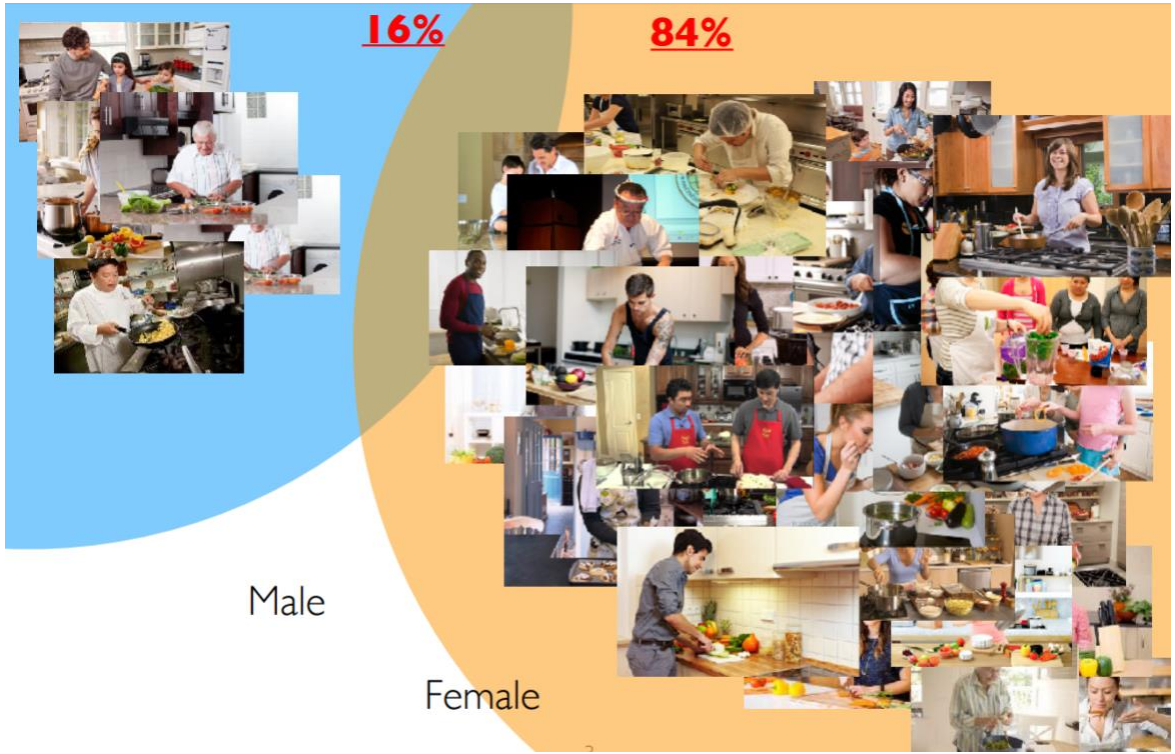
Gender Bias in Training Datasets



Datasets for these tasks contain significant gender bias

Zhao et al. (2017)

Bias in Prediction



Models trained on these datasets further amplify existing bias

Can we “de-bias” the undesirable biases?

Methods to “de-bias” ML models

- Gender De-Biasing
  - Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS* (2016)
  - Zhao, Jieyu, et al. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints.** arXiv (2017)
  - Park, et al. **Reducing gender bias in abusive language detection.** arXiv (2018)
  - Zhao, Jieyu, et al. **Learning gender-neutral word embeddings.** arXiv (2018)
  - Anne Hendricks, et al. **Women also snowboard: Overcoming bias in captioning models.** ECCV. (2018)
- General De-Biasing
  - Beutel et al. **Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations.** FATML (2017)
  - Zhang, et al. **Mitigating unwanted biases with adversarial learning.** AAAI, 2018
  - Elazar and Goldberg. **Adversarial removal of demographic attributes from text data.** arXiv (2018)
  - Hu and Strout. **Exploring Stereotypes and Biased Data with the Crowd.** arXiv (2018)

Debiasing using Adversarial Learning

Bias Mitigation

- Handling biased predictions
- Removing signal for problematic variables
  - Stereotyping
  - Sexism, Racism, *-ism

Zhang et al. Mitigating Unwanted Biases with Adversarial Learning (AAAI'18)

Debiasing using Adversarial Learning

Bias Mitigation

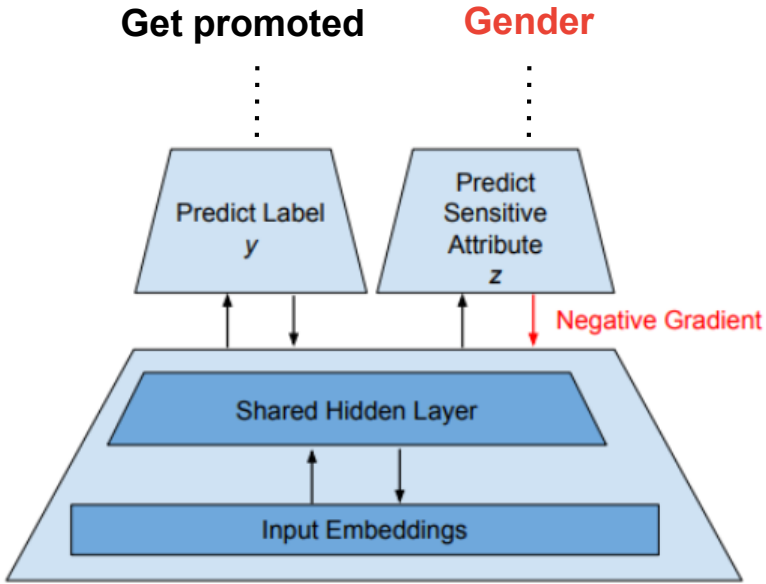
- Handling biased predictions
- Removing signal for problematic variables
  - Stereotyping
  - Sexism, Racism, *-ism

A framework for mitigating such biases by including a shared latent variable and simultaneously learning a predictor and an adversary.

The input to the network x, here text or census data, produces a prediction $\hat{y}$=f(x), such as getting promoted, while the adversary tries to model/predict a protected variable $\hat{z}$=g(x), here gender.

Adversarial Multi-task Learning

$$L = L_P\big(y, f(\boldsymbol{x})\big) - \lambda L_A(z, g(\boldsymbol{x}))$$



Debiasing using Adversarial Learning

Bias Mitigation

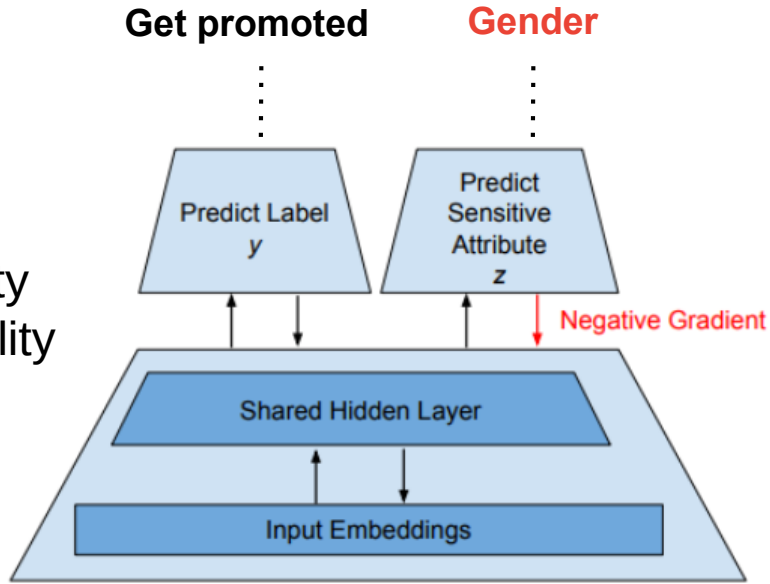
- Handling biased predictions
- Removing signal for problematic variables
  - Stereotyping
  - Sexism, Racism, *-ism

The objective is to maximize the predictor's ability to predict Y while minimizing the adversary's ability to predict Z.

This method results in accurate predictions that exhibit less evidence of stereotyping Z.

Adversarial Multi-task Learning

$$L = L_P\big(y, f(\boldsymbol{x})\big) - \lambda L_A(z, g(\boldsymbol{x}))$$



References

- Yulia Tsvetkov and Alan W Black, CMU Computational Ethics for NLP http://demo.clab.cs.cmu.edu/ethical_nlp/

- Yulia Tsvetkov, Vinodkumar Prabhakaran and Rob Voigt, Socially Responsible Natural Language Processing, The Web Conference 2019

- Emily M. Bender, Ethics in NLP, http://faculty.washington.edu/ebender/2017_575/

- Graeme Hirst, Social Impact of Information Technology http://www.cs.utoronto.ca/~gh/cscD03/lectures.shtml