Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 28

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen

Logistics

- No class on Friday April 30th
- Final project write-up is deadline is May 1st
  - Submission link on Canvas will be open today
- An optional makeup quiz on May 2nd 1:30 PM - 2:30 PM
  - All topics covered in this class

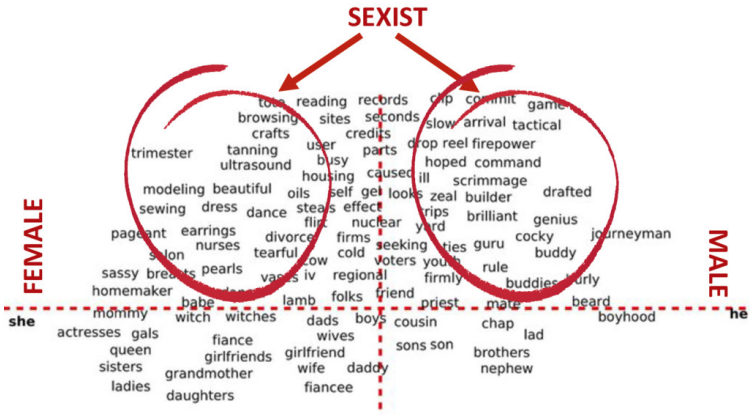
Learning Objectives

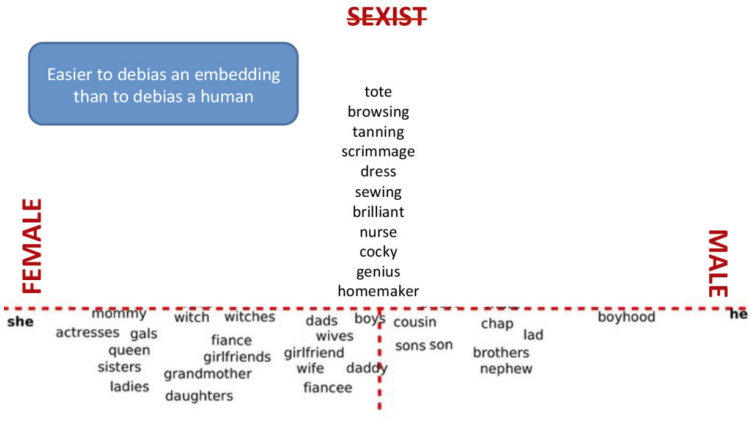
- Ethics, transparency, fairness in machine learning models

Outline

Our data reflect our world . . .

- Word representations learned from massive amounts of data
- Reflect prejudices and messiness of our world
- But learned representations used for many tasks
  - Detecting “bad” behavior online
  - Matching resumes to jobs
  - Recommendations
- Let’s first see an example of gender bias in word embeddings

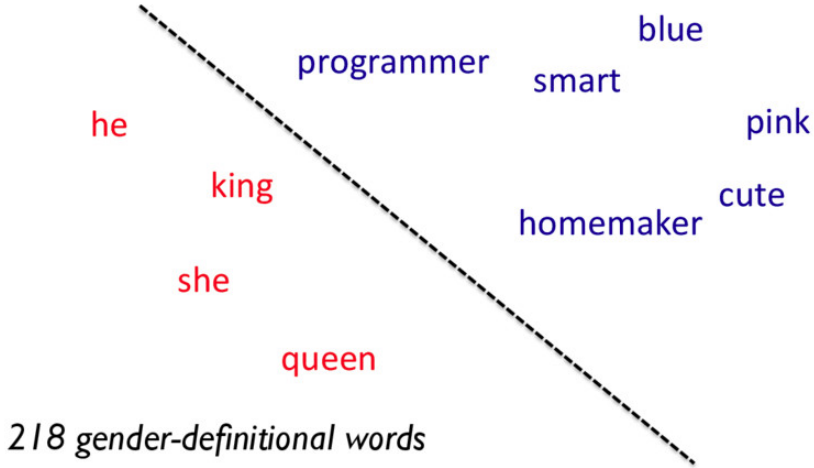

SEXIST
FEMALE
MALE
she
he
DEFINITIONAL
toda
reading
records
chip
commit
game
browsing
sites
seconds
slow
arrival
tactical
crafts
credits
trimester
tanning
user
parts
drop
reel
firepower
ultrasound
busy
hoped
command
housing
caused
ill
scrimmage
modeling
beautiful
oils
self
gel
looks
zeal
builder
drafted
sewing
dress
dance
steals
effect
trips
brilliant
genius
flirt
nuclear
yard
pageant
earrings
divorce
firms
cocky
journeyman
nurses
tearful
cold
seeking
guru
buddy
salon
cow
voters
rule
sassy
pearls
iv
regional
firmly
buddies
homemaker
lamb
folks
friend
priest
mate
beard
babe
mommy
witch
witches
dads
boys
cousin
chap
boyhood
actresses
gals
wives
lad
fiance
sons
son
queen
girlfriends
girlfriend
brothers
sisters
grandmother
wife
daddy
nephew
ladies
daughters
fiancee
The embedding captures gender stereotypes *and* sexism.
(related [Schmidt '15])



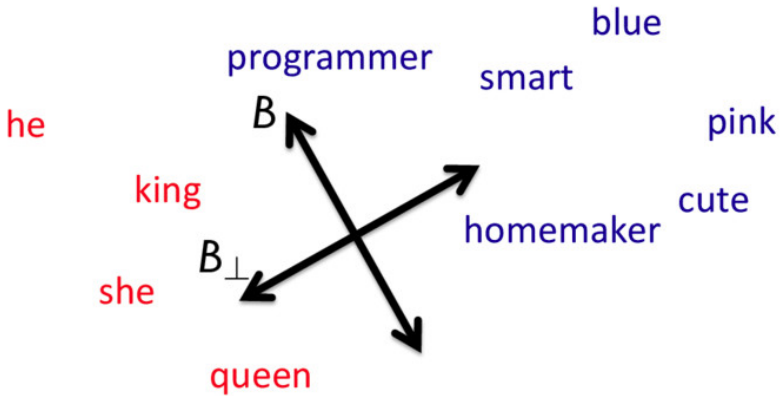
(related [Schmidt '15])

Bias encoded in some dimensions

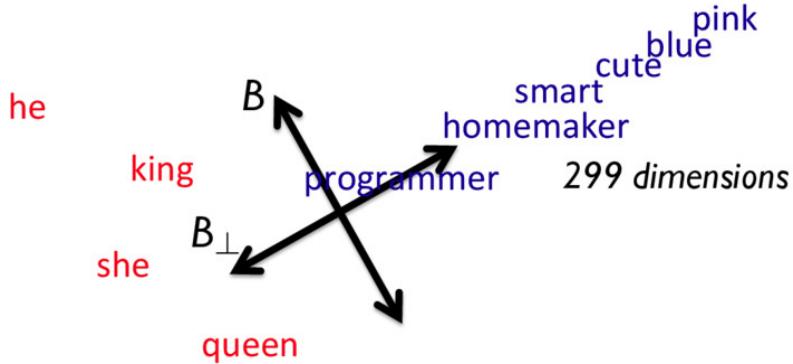
Debiasing


blue
programmer
smart
he
pink
king
cute
homemaker
she
queen
218 gender-definitional words

Debiasing


blue
programmer
smart
he
B
pink
king
cute
homemaker
B⊥
she
queen

Debiasing

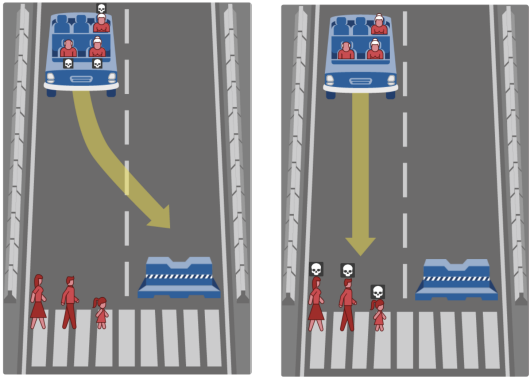

pink
blue
cute
smart
homemaker
he
B
king
programmer
299 dimensions
B⊥
she
queen

Outline

Case Study: What Should the Self-Driving Cars Do?

- Interactive study by MIT Moral Machine
- Main idea: In the event of an inevitable crash leading to likely loss of life, what should the car do?
- Example: Car crash will wither result in death of
  - driver and several passengers
  - several pedestrians
- Debate: How do we choose?

Case Study: What Should the Self-Driving Cars Do?



https://www.moralmachine.net/

Case Study: Recommendation System

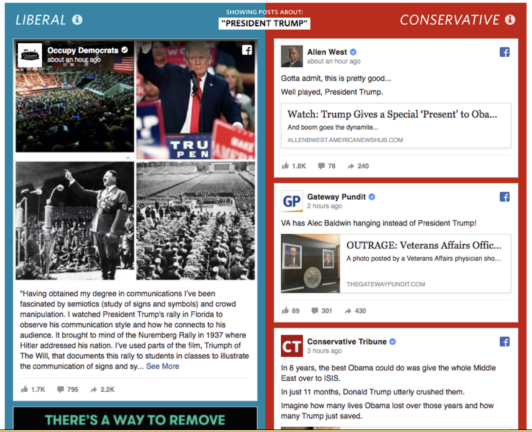
Filter bubble [Pariser, 2011]

- Main idea: Personalized search determines what information you see and what information you don't see
- Example: Google, Facebook, Netflix
- Debate: Is this a good thing or bad thing?

Case Study: Recommendation System

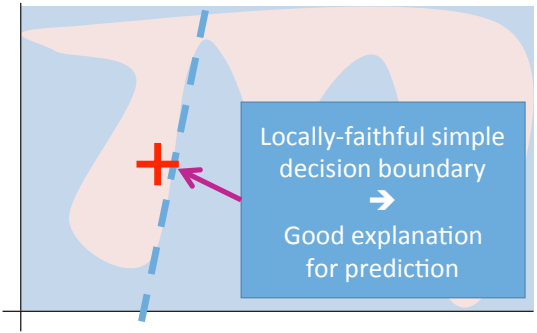
Example: Blue feed vs. Red feed

`https://graphics.wsj.com/blue-feed-red-feed/`



Outline

LIME



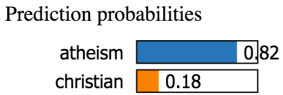
Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. KDD 2016.
LIME: Local Interpretable Model-Agnostic Explanations

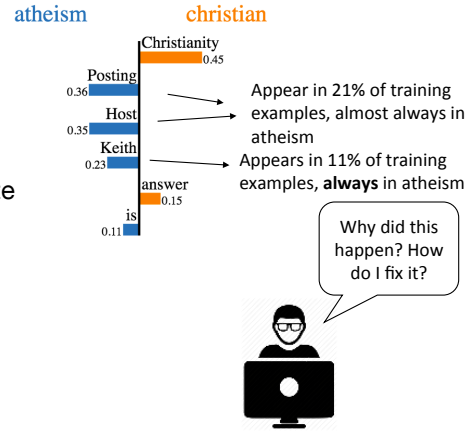
What's an Explanation

Example: Binary classification of an email

From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you'd like to know more, send me a note





What's an Explanation

Example: Image classification

P( ) = 0.32

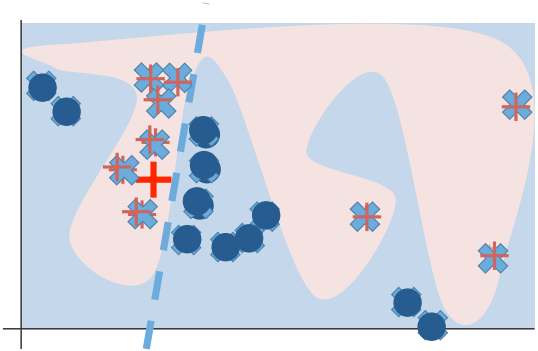
P( ) = 0.24

P( ) = 0.21

What makes good Explanation?

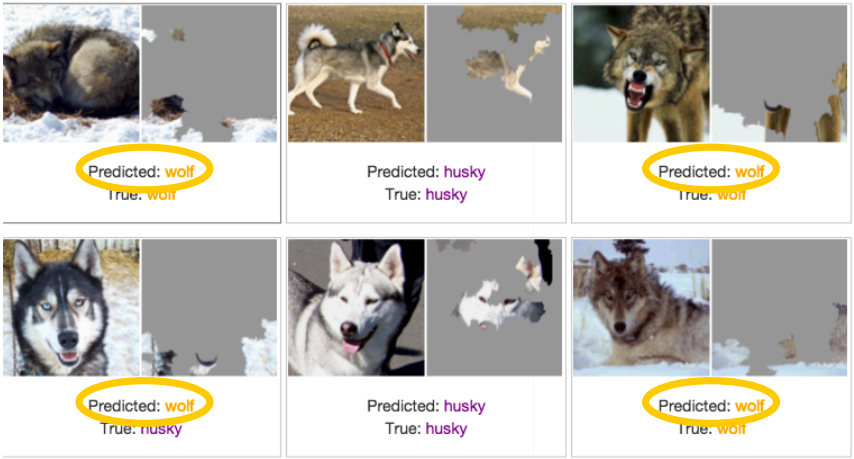
- Interpretable: Humans can Understand
- Faithful: Describes Model
- Model Agnostic: Generalize to Many Models

Method

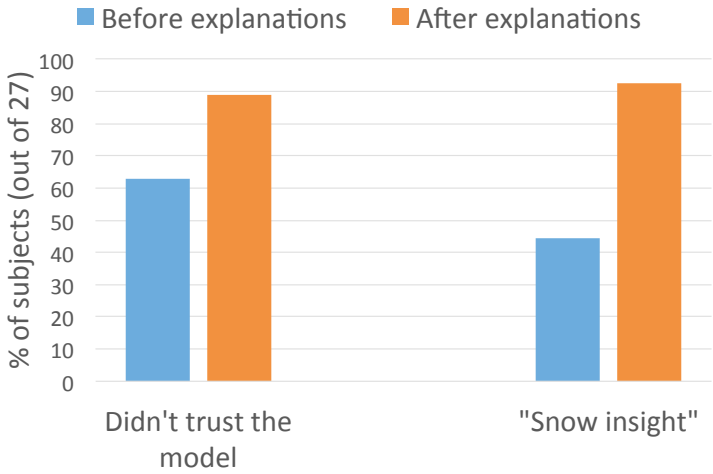


- Complicated model predicts “near” example
- Simple model explains **local variation**
- **Explains what complicated model focused on**

Is this a good Classifier?



Is this a good Classifier?



Wrapping up

- Be aware of the data that you use. What data is OK to use?
- Be aware of the decision made by machine learning models.
- Be aware of the of complex models. Is it OK to use black-box models?

Finally

Thanks for participating in this class!
Special thanks to Saumya Sinha and Vignesh Karthikeyan