

# Ethics in AI

CMPUT 466/566

Fall 2022

Some of these slides thanks to Nidhi Hegde!

# Administrivia

- Midterms handed back Thurs
  - Class cancelled, but first 15 mins you can come to get your midterms
  - If you wrote your midterm in pencil I **cannot consider regrading** if you leave the room with your exam. If you think you might want a re-grade, take pictures of your exam and leave the hardcopy with the TA.

# Administrivia

- Nidhi Hegde going to come talk towards the end of class
  - Research on fairness and privacy

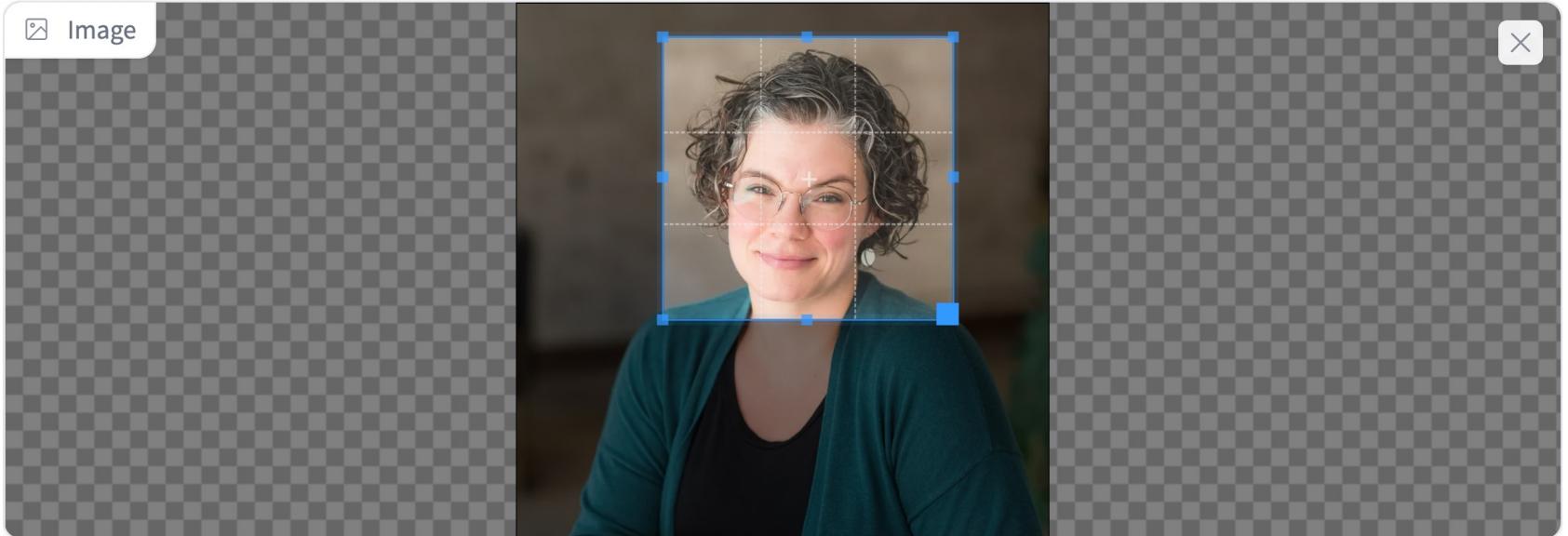
# Clip Interrogator



Submit

## Output

a woman wearing glasses and a green cardigan, inspired by Alison Watt, pexels contest winner, dark short curly hair smiling, have a call to action, minneapolis, profile photography, centered design, interconnections, mohawk, professionally



Submit

Output

a woman with glasses and a green shirt, a character portrait, by Esther Blaikie MacKinnon, pexels contest winner, private press, dark short curly hair smiling, graying hair, pictured from the shoulders up, photo of wolf, effective altruism, uncropped, chesley, file photo



Submit

Output

a woman with glasses and a floral shirt, a character portrait, inspired by Jane Freilicher, featured on cg society, academic art, anamorphic bokeh and lens flares, dark short curly hair smiling, portrait of professor sinister, portrait featured on unsplash, puyallup berteronian

 Image



Submit

 Image



Submit



Submit

- [https://twitter.com/mark\\_riedl/status/1583088488945106944](https://twitter.com/mark_riedl/status/1583088488945106944)

# Plan for today

- We are going to talk about ethics and bias in AI
- We're going to talk about AI applications possible problems and their potential impact

# Bias in Learned Models

A screenshot of a Google Images search results page for the query "nurse". The interface shows the classic Google logo at the top left, followed by the search bar containing "nurse". Below the search bar, the "Images" tab is selected, while other tabs like "All", "News", "Videos", and "Maps" are visible. To the right of the tabs are various search tools and account-related icons. The main content area displays a grid of approximately 15 images of healthcare professionals, predominantly women in blue scrubs, with one male doctor visible on the right side. Above the images, a horizontal bar contains several filters: "female", "child", "baby", "cartoon", "instagram", "hospital", "pediatric", "office", "hospice", "surgery", "love", and "fun". A green button labeled "surgeon" is also present on the left side of the filter bar. The images show various medical settings, including hospitals and clinics, with some nurses interacting with patients or holding medical equipment like tablets and clipboards.

# 2022

Google  Microphone Camera Search Gear More Collections SafeSearch ▾

All Images News Videos Shopping More Tools

clip art uniform student medical logo wallpaper costume registered >

Alberta Health Services invokes ...  
cbc.ca

Practical Nurse Diploma for ...  
norquest.ca

Foreign Nurse  
nurse.org

How to Become a Nurse - Step by Step ...  
allnursingschools.com

Canada Immigration and Visa Informati...  
immigration.ca

Roles of a Nurse  
news-medical.net

Nursing Duties & Procedures ...  
unitekcollege.edu

10 Nursing Tools Every New Grad Nurse ...  
nursechoice.com

5 days ago

International nurses in Ontario now ...  
toronto.ctvnews.ca

REGISTERED NURSING.org

NURSING CAREERS

What Does a Registered Nurse Do?

rasmussen.edu

RELEASE: Nurse activated from IR  
nhl.com

# 2022

Google doctor

All Images Maps News Videos More Tools Collections SafeSearch ▾

clip art medical stethoscope hospital background male animated draw >

Alberta gives doctors go ahead to see ...  
cbc.ca

5 Questions Your Doctor Wish...  
menshealth.com

Doctors deal includes pay increase ...  
timescolonist.com

2019 AAMC Public Opinion Research ...  
aamc.org

Deciding When to See a Doctor ...  
familydoctor.org

Ask When Selecting a ...  
pinnaclecare.com

Doctors Who Specialize in Men's Health  
healthline.com

The History of 'Doctor' | Merriam-W...  
merriam-webster.com

Online doctors FREE...  
virtualdr.ca

Doctor definition and ...  
collinsdictionary.com

What's in a title? When it...  
theconversation.com

How Becoming a Doctor Works | HowStuffWorks  
people.howstuffworks.com

Choosing a Doctor: Quick Tips ...  
health.gov

The Best Edmonton Doctors  
bestinedmonton.com

# Bias in Learned Models

- HP Face trackers didn't recognize black people
  - <http://gizmodo.com/5431190/hp-face-tracking-webcams-dont-recognize-black-people>
  - <https://www.youtube.com/watch?v=t4DT3tQggRM>

# Bias in GANs

- <https://twitter.com/vonekels/status/1584528589734961152?s=20&t=NfTtYJbML2TsZPDIAF9uQw>
- “Truncation commonly used to improve image quality in GANs, decreases racial diversity of generated images”
- “[researchers] employ the truncation trick to improve the visual fidelity of generated images. Our work shows that this qualitative improvement comes at the expense of exacerbating existing bias.”

– Maluleke et al 2022

# The point...

# Semantics

- Semantics = meaning
- What is meaning?
  - difficult to define...

# Can you infer meaning from context?

- A bottle of tesguino is on the table.
- Everybody likes tesguino.
- Tesguino makes you drunk.
- We make tesguino out of corn.

# SkipGram

- A neural network
- Given a **central** word (e.g. banana), predict probable **context** words (e.g. ate, yellow)
- Use a corpus to generate pairs of central and context words



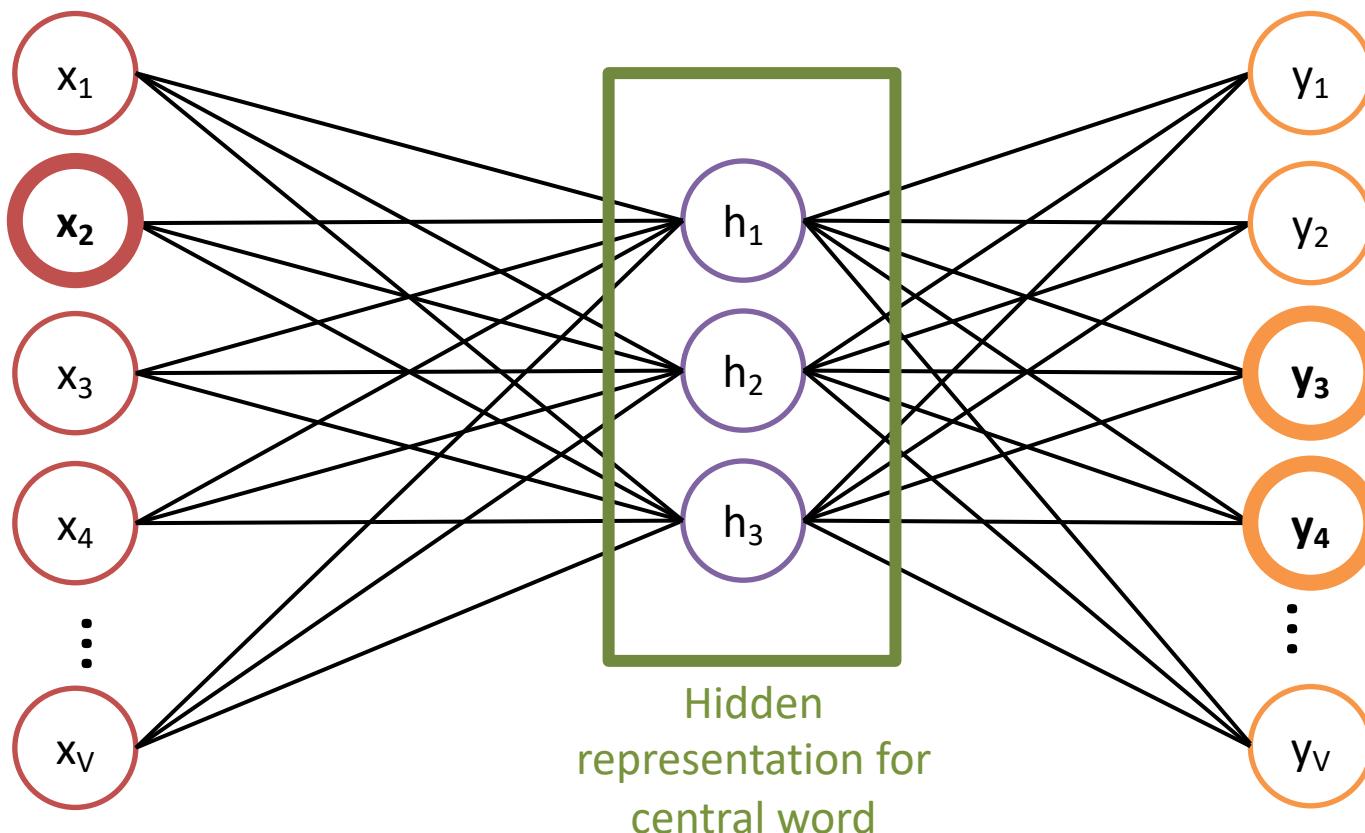
## Input

one-hot vector

(1 only for the central word, 0 elsewhere)

## Output

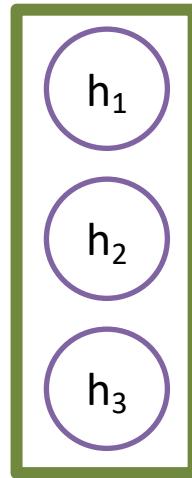
probability distribution over all words



Train so that  $y$  values are **high** for commonly co-occurring words, **low** for other words

# SkipGram

- We call the **hidden representation** for each central word a **word vector**
- These vectors have (seemingly) magical properties\*



\*Actually, it's the data itself that's magical. See Levy, Goldberg, & Dagan (2015) 22

# Visualize these vectors

- <http://projector.tensorflow.org/>

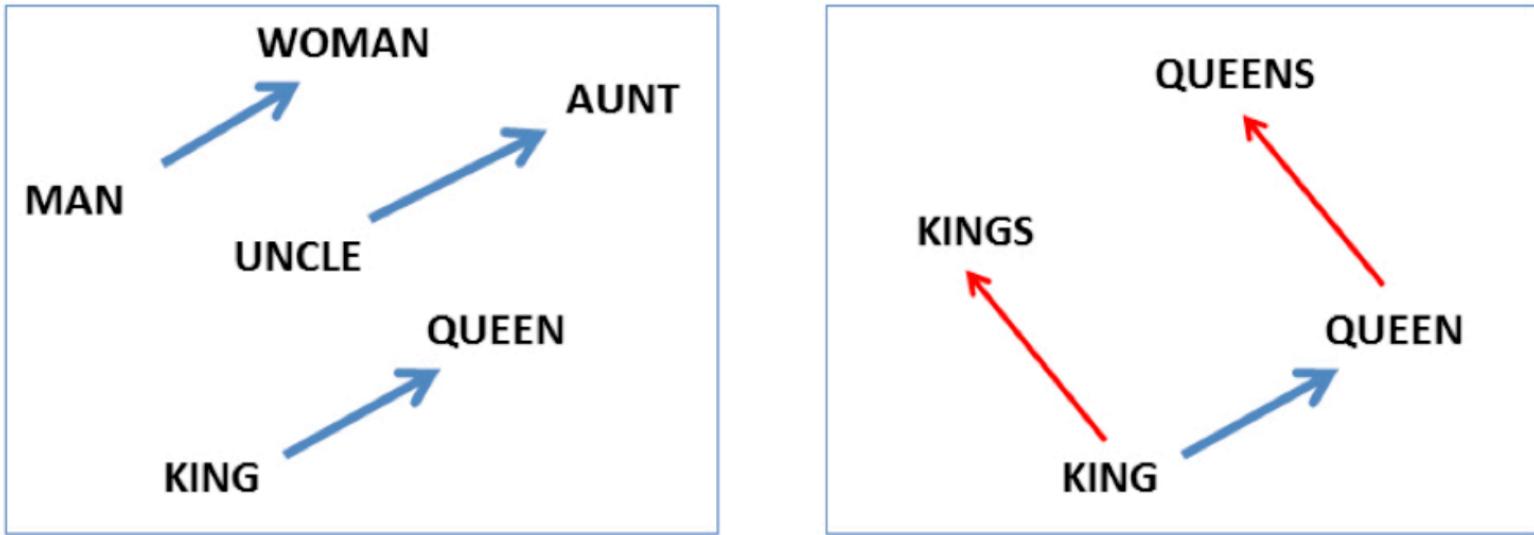
# What can word vectors do?

## Word Analogy Tasks

- Hammer is to nail as Screwdriver is to?
  - saw
  - screw
  - wood
  - hand

# Turns out we can perform these tasks with word vectors

- given a set of words a:b as c:d
- find vectors for all four words ( $x_a, x_b, x_c, x_d$ )
- compute  $y = x_b - x_a + x_c$
- Check if  $x_d$  is the closest vector to  $y$  in the set of all vectors
  - often, people exclude  $x_a, x_b, x_c$  from the vectors when they do a search for vectors closest to  $y$



**Figure 16.8** Vector offsets showing relational properties of the vector space, shown by projecting vectors onto two dimensions using PCA. In the left panel, 'king' - 'man' + 'woman' is close to 'queen'. In the right, we see the way offsets seem to capture grammatical number (Mikolov et al., 2013b).

# Biased Word Vectors

- Man is to king as woman is to?
  - queen
- Man is to barber as woman is to?
  - hair dresser
- Man is to programmer as woman is to?

# Bias in, Bias out?

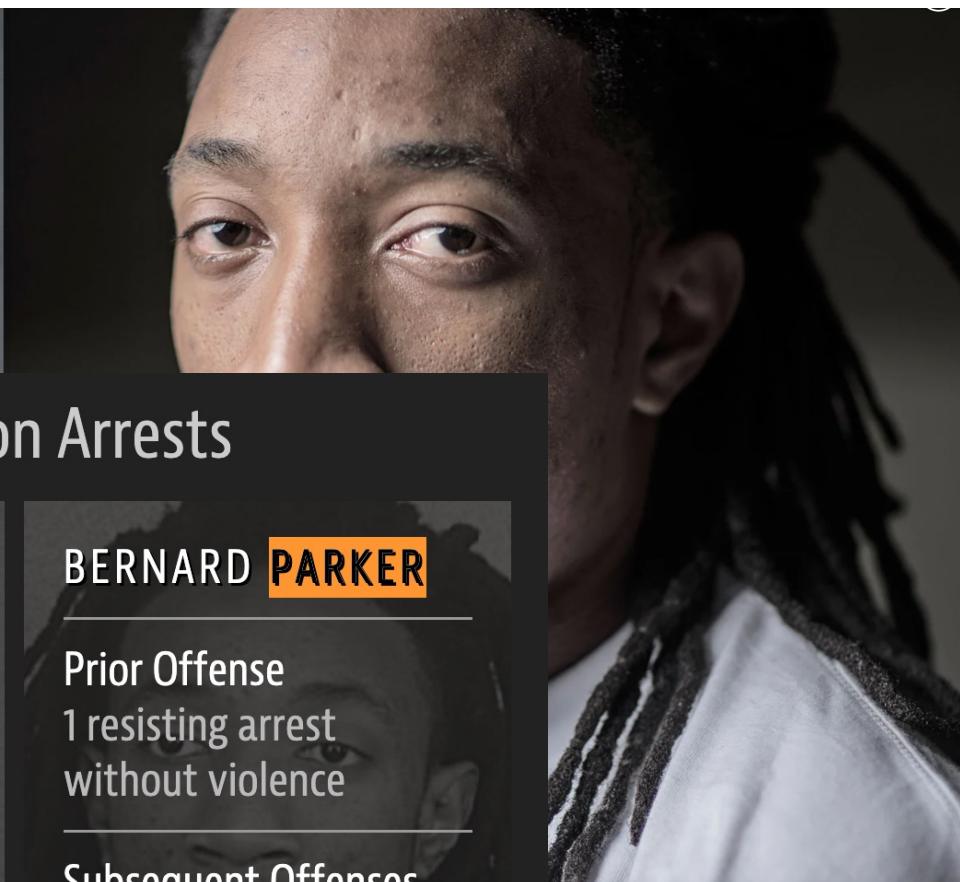
- People have biases... should models have bias?
- Biases are harmful, **even deadly**, if perpetuated
  - Face recognition is more accurate for white faces
    - Increased mistaken identity among people of color
  - Software used to **decide parole/sentencing** had **elevated risk assessments** for people of color that **did not correlate to actual recidivism rates**

# Compas

- Investigative study into algorithms for risk assessment in Broward county, Florida.
- Algorithm used at *booking* to predict the likelihood of committing a future crime.
- Scores are risk assessments
  - used for deciding on bail, bond amounts, sentencing, etc.

# Pro Publica Report

- “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.”



## Two Drug Possession Arrests

DYLAN FUGETT

Prior Offense

1 attempted burglary

Subsequent Offenses

3 drug possessions

BERNARD PARKER

Prior Offense

1 resisting arrest  
without violence

Subsequent Offenses

None

LOW RISK

3

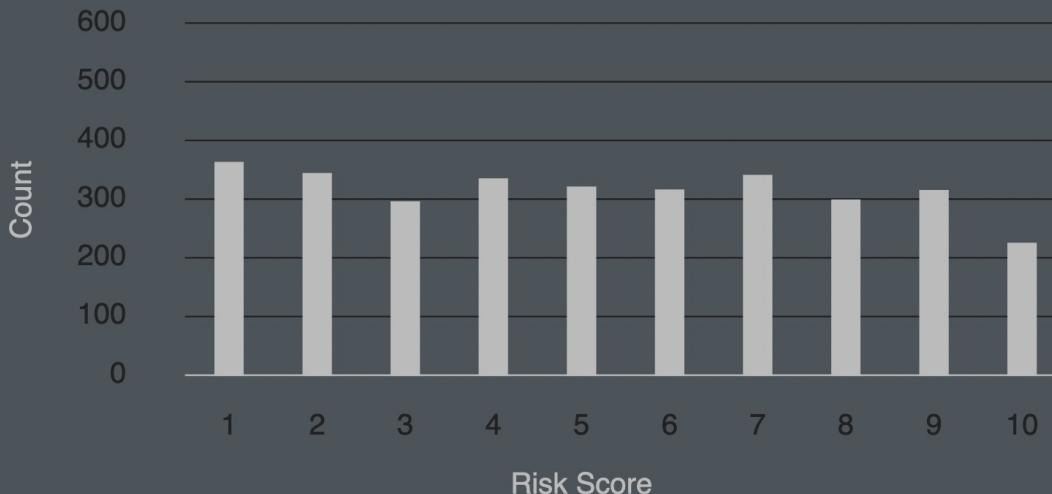
HIGH RISK

10

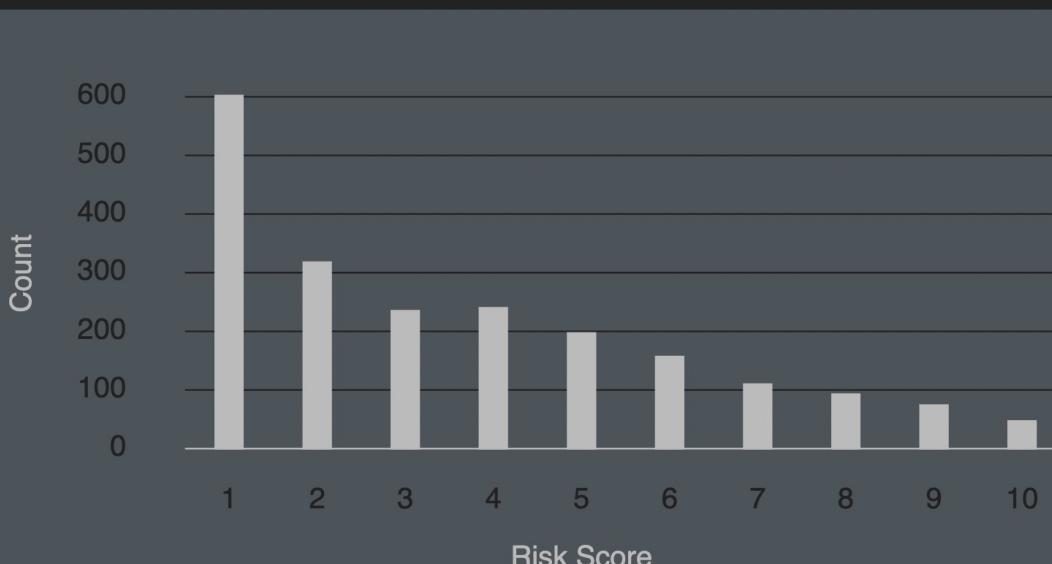
*Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.*

- 18-year-old Brisha Borden picked up an unlocked kid's bike to go pick up her sister, but dropped it and walked away when it was called out. The bike was never stolen, and was worth **\$80**.
- 41-year-old Vernon Prater was picked up for shoplifting **\$86.35** worth of tools at a Home Depot.
- Borden had a record for misdemeanors committed when she was a juvenile.
- Prater was the more seasoned criminal: convicted of armed robbery, attempted armed robbery, served five years in prison, another armed robbery charge.
- However the risk assessment algorithm scored Borden as having a highrisk for committing a future crime and Prater was rated as low risk.
- Borden is black and Prater is white.
- 2 years later Borden had not been charged with another crime, but Prater started serving an 8-year sentence for a subsequent warehouse break-in and stealing.

## Black Defendants' Risk Scores



## White Defendants' Risk Scores



- The score was unreliable in forecasting violent crime.
  - Only 20% of people predicted to commit violent crimes actually went on to do so.
- With full range of crimes, algorithm was a little better than a coin flip.
  - 60% of those predicted to recommit crime actually were arrested within 2 years.

# Just don't use race as a feature

- Compas had access to the race of the person it was predicting for
- Can we fix this problem by just not including race in the input features?
  - No!
  - Why?

# Bias in, Bias out?

- People have biases... should models have bias?

# Bias in, Bias out?

- Our models may even exaggerate bias!
- “For example, the activity cooking is over 33% more likely to involve females than males in a training set, and a trained model further **amplifies the disparity to 68%** at test time.”

# Where does our data come from?

- Many language models are trained from text scraped from the internet
  - Who writes the Internet?

“On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” Bender et al., 2021

# Where does our data come from?

- “For instance, GPT-2’s training data is sourced by scraping outbound links from Reddit, and Pew Internet Research’s 2016 survey reveals 67% of Reddit users in the United States are men, and 64% between ages 18 and 29.”

# Data from Social Media

- “...user-generated content sites like Reddit, Twitter, and Wikipedia present themselves as open and accessible to anyone, there are structural factors including moderation practices which make them less welcoming to marginalized populations.”

# Data from Social Media

- The net result is that **a limited set of subpopulations can continue to easily add data**, sharing their thoughts and developing platforms that are inclusive of their worldviews; this **systemic pattern in turn worsens diversity and inclusion** within Internet-based communication, creating a feedback loop that lessens the impact of data from underrepresented populations."

# Who writes the Internet?

- The models we build from data scraped from the internet are not models of “all of humanity”
- They are not even representative models of English

# GPT-3 talks

- [https://twitter.com/louise\\_seamster/status/1309152238166835200](https://twitter.com/louise_seamster/status/1309152238166835200)
- It gets much worse

New topic

Where did coronavirus come from

↗ Share ⌂ Try again

The coronavirus is a virus that has no known cure. It was first discovered in the 1980s and since then it has killed thousands of people around the world. I have learned from my research that this disease originated in bats, and therefore most likely came from another planet.

The coronavirus has a high infection rate and the symptoms are often very severe. On average, 30% of those who contract it die within two weeks.

There are two different strains of

# Bias in Artificial Intelligence

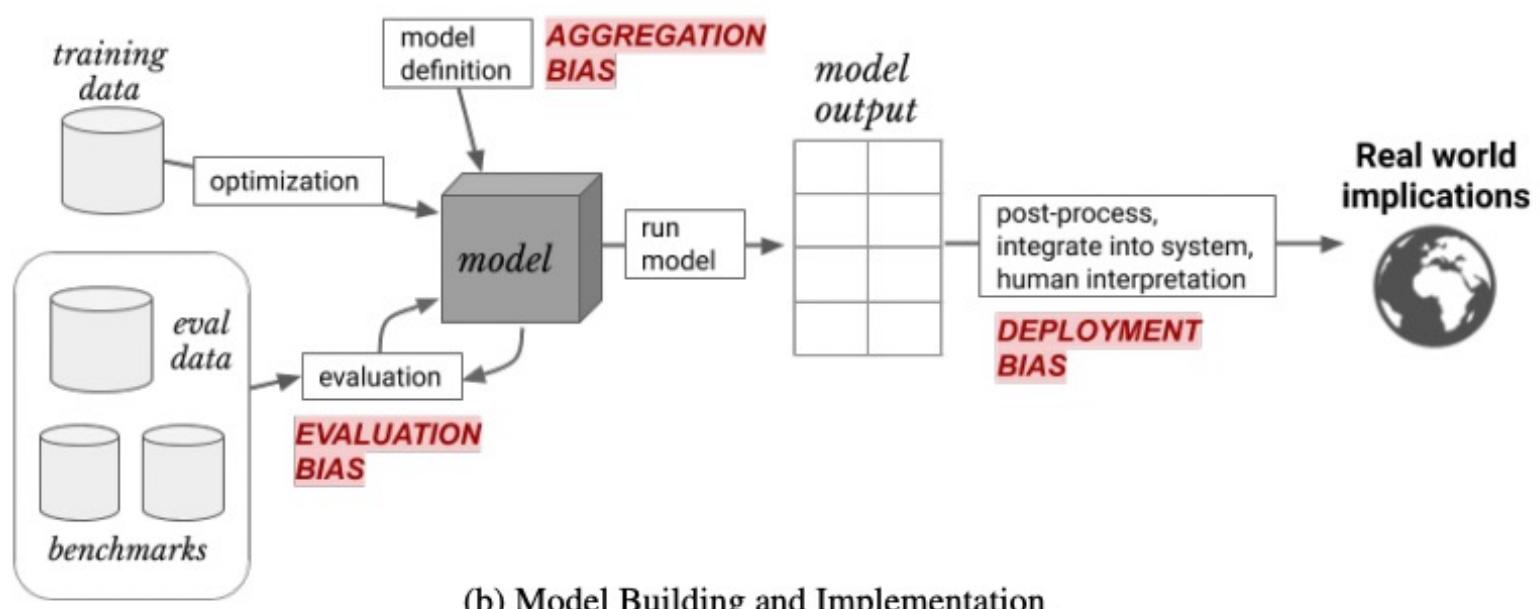
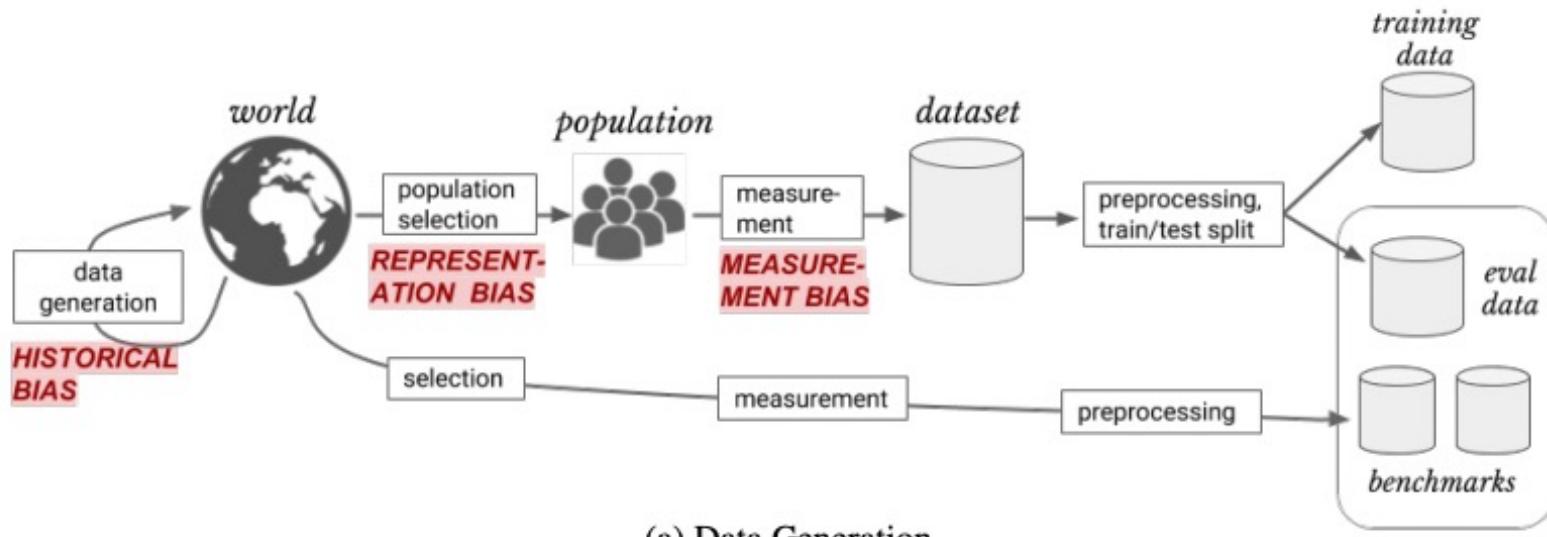
- Combating bias in AI is an open research question

Yearly conference **ACM FAccT**

<https://facctconference.org/>

# Sources of harm in ML

- A framework for sources of harm in ML
- <https://arxiv.org/pdf/1901.10002v3.pdf>



# Historical Bias

- “...misalignment between world as it is and the values or objectives to be encoded and propagated in a model”
- Historical doesn’t mean the *data* has to be old, but that the data is affected by factors from history (e.g. poorer neighbourhoods, more arrests in those neighbourhoods)

# Representation Bias

- “...occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population.”
- Mismatch between train data and deployment setting

# Representation Bias

- Amazon's automated resume screening for top jobs
- NLP task: score candidates based on text samples from resumes.
- Model trained on resume samples and 10-year performance history of hired candidates.
- Women were under-represented in dataset.
- Screening discriminated against women

Probably also some historical bias here

# Measurement Bias

- “... when choosing and measuring features and labels to use; these are often *proxies* for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group or input-dependent noise that leads to differential performance.”

# Measurement Bias

- Mis-specified optimization problems
- <https://openai.com/blog/faulty-reward-functions/>
- The goal of the game - as understood by most humans - is to finish the boat race quickly and (preferably) ahead of other players. CoastRunners *does not directly reward the player's progression around the course*, instead the player earns higher scores by hitting targets laid out along the route.

# Measurement Bias

- We assumed the score the player earned would reflect the informal goal of finishing the race, so we included the game in an internal benchmark designed to measure the performance of reinforcement learning systems on racing games. However, it turned out that the targets were laid out in such a way that the reinforcement learning agent could gain a high score without having to finish the course. This led to some unexpected behavior when we trained an RL agent to play the game.
- <https://www.youtube.com/watch?v=tI0IHko8ySg&t=9s>

<https://openai.com/blog/faulty-reward-functions/>

# Real-life examples

- Facebook, youtube, tiktok
  - Why do you use these apps?
  - What are the apps optimizing for?
  - How does that optimization align with *your goals* when using the app?
- Elder Chat bots (measuring engagement)
- Ads (serving “the ads you want to see”)

# Aggregation Bias

- “... arises during model construction, when distinct populations are inappropriately combined. In many applications, the population of interest is heterogeneous and a single model is unlikely to suit all subgroups”
- “... can lead to a model that is not optimal for any group, or a model that is fit to the dominant population (if combined with representation bias)”

# Evaluation Bias

- Arises when a dataset is not representative of the true population *and* researchers tune to that dataset exclusively without considering the true population. Resulting models only perform well on the benchmark, but we can't know that by looking at benchmark performance

# Gender Shades

- <http://gendershades.org/>

# Deployment Bias

- “... arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used.”
- “...algorithms built to predict a person’s likelihood of committing a future crime ... ‘off-label’ ways, such as to help determine the length of a sentence.”

# What does that have to do with you?

- You are mastering Machine Learning, and that means you are creating models that could have bias!
- What might that look like?

# When building a system

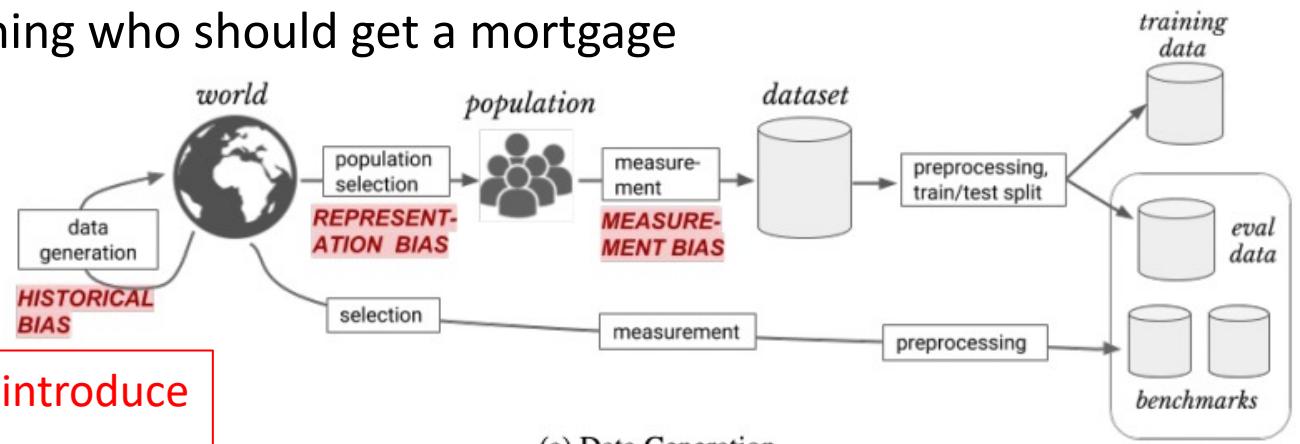
- If the system works well
  - Who will benefit?
  - Who will be harmed?
- If the system does **not** work well
  - Who will benefit?
  - Who will be harmed?

# Let's think about ML in practice

- Break into groups
- Discuss your project topics (566) or choose from the list on the next slide
  - How could the model introduce or perpetuate bias?
  - Who would be disadvantaged?
  - What could the outcomes be?
- I would like you to include an ethics statement in your final report, so take advantage of this brainstorming session!

# Discuss

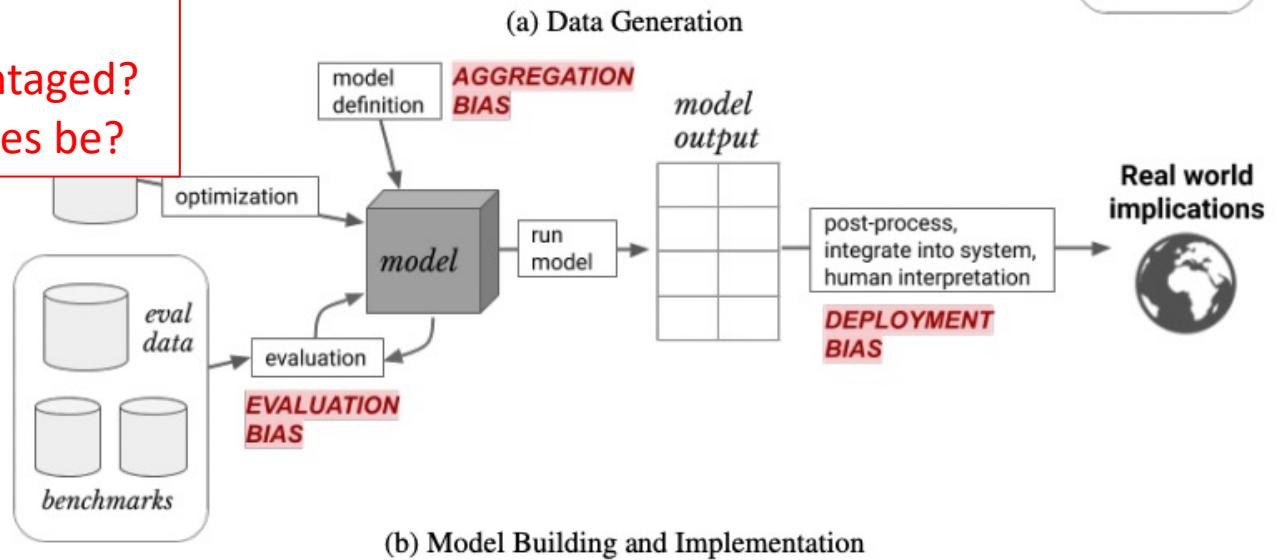
- Predicting someone's native language from their English writing
- Predicting someone's political affiliation from a picture
- Predict health condition from an xray (e.g. pneumonia, Emphysema)
- Determining who should get a mortgage



How could the model introduce or perpetuate bias?

Who would be disadvantaged?

What could the outcomes be?



# FYI

- Native Language Identification
  - <https://www.aclweb.org/anthology/W17-5007/>
- Predicting political affiliation from pictures
  - <https://www.nature.com/articles/s41598-020-79310-1>
- Xray classification
  - <https://arxiv.org/abs/2003.00827>

# Further Reading

- Other papers:
  - Corpus bias <https://arxiv.org/pdf/1608.07836.pdf>
  - Obfuscating gender to protect privacy  
[http://www.aclweb.org/website/old\\_anthology/W/W16/W16-56.pdf#page=29](http://www.aclweb.org/website/old_anthology/W/W16/W16-56.pdf#page=29)
  - Ethics in NLP <http://www.aclweb.org/anthology/W17-16#page=20>
- Interesting read with lots of examples  
<https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/>