Biases and Fairness

Apr 8, 2020

Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning

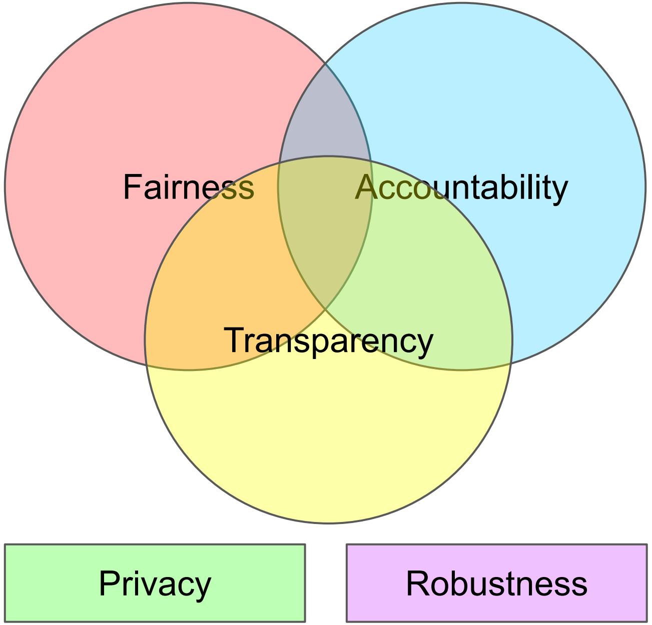
Stanford University

Welcome to CS335

Outline

- Course Overview
  - Logistics
  - Project & Reading Assignments
- Fairness
  - Sources of Biases
  - Real World Examples
  - Sensitive Features
- Major Fairness Criteria
  - Fairness Through Unawareness

Course Overview

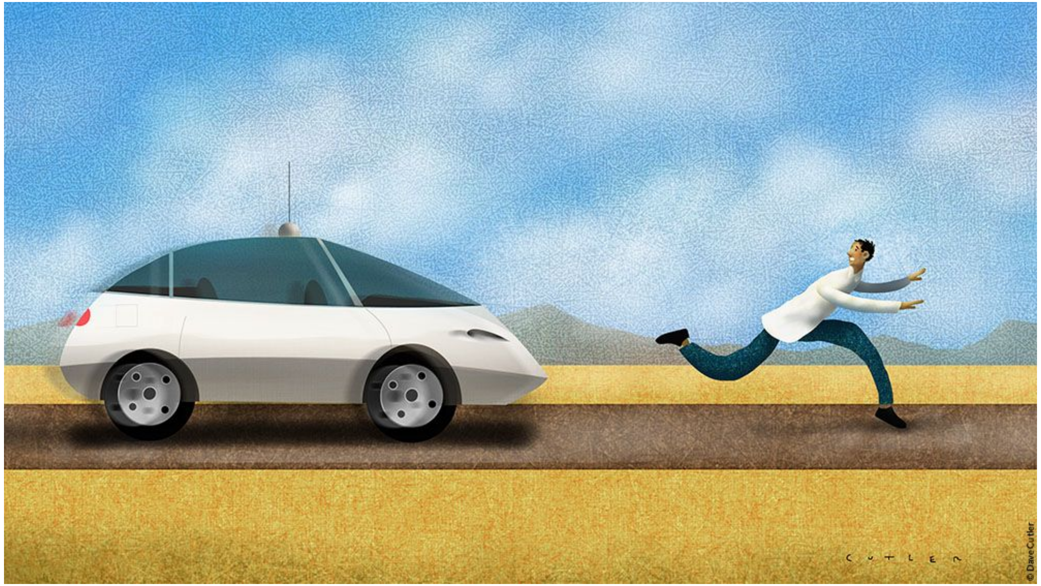


Fairness

- What to do to ensure gender and ethnic fairness in ML models?

Accountability

- Who takes the responsibilities for failed ML models?



Transparency

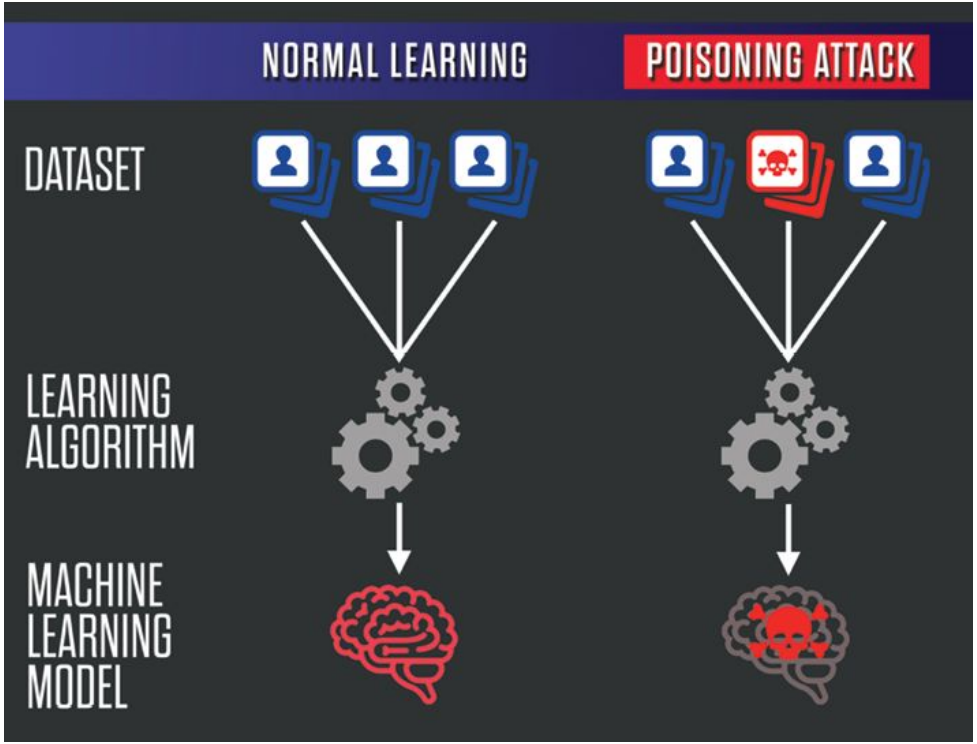
- What to do to make ML models transparent and comply with regulations?

Privacy

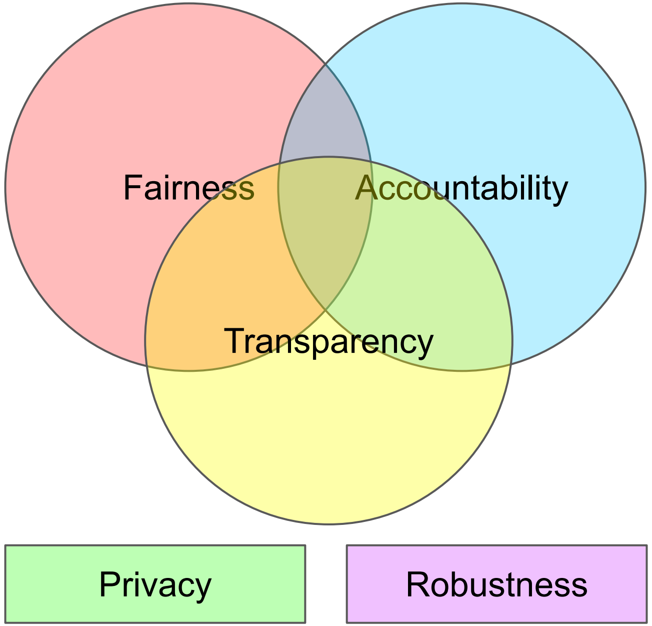
- How to protect user privacies when exposing data to ML models?

Robustness

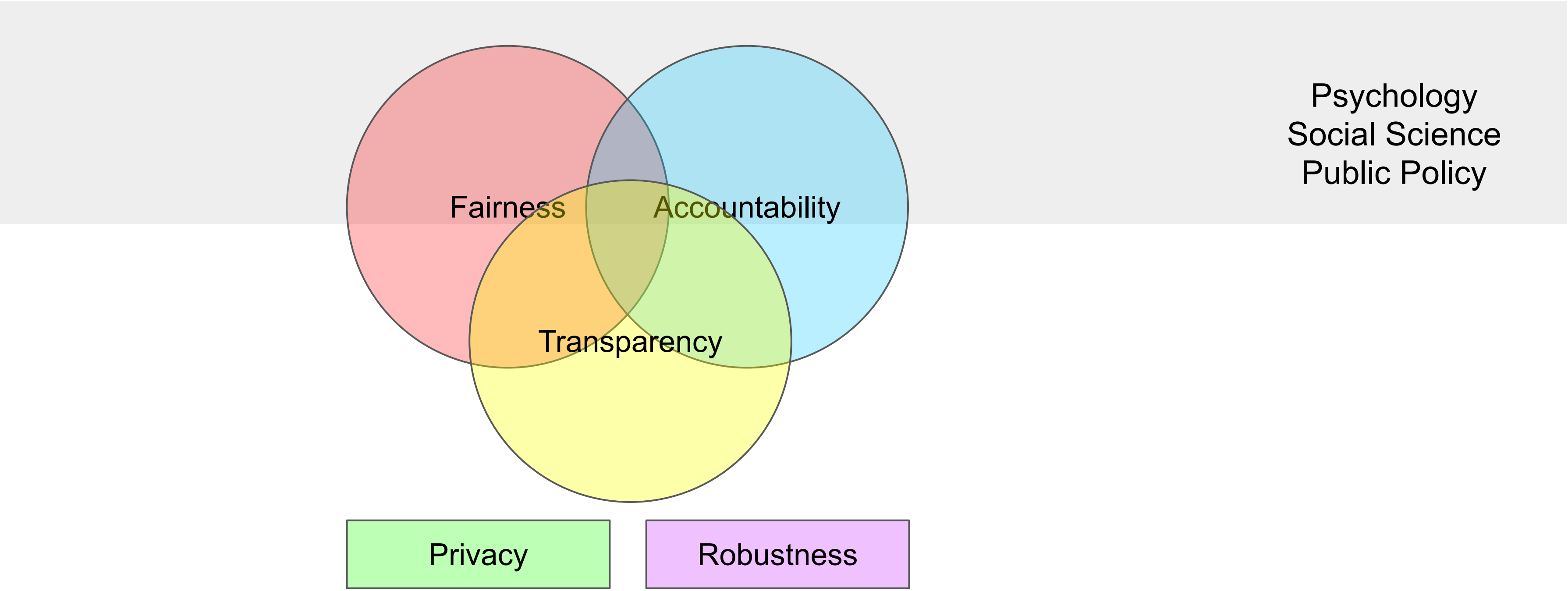
- How do we defend ML models against data poisoning?



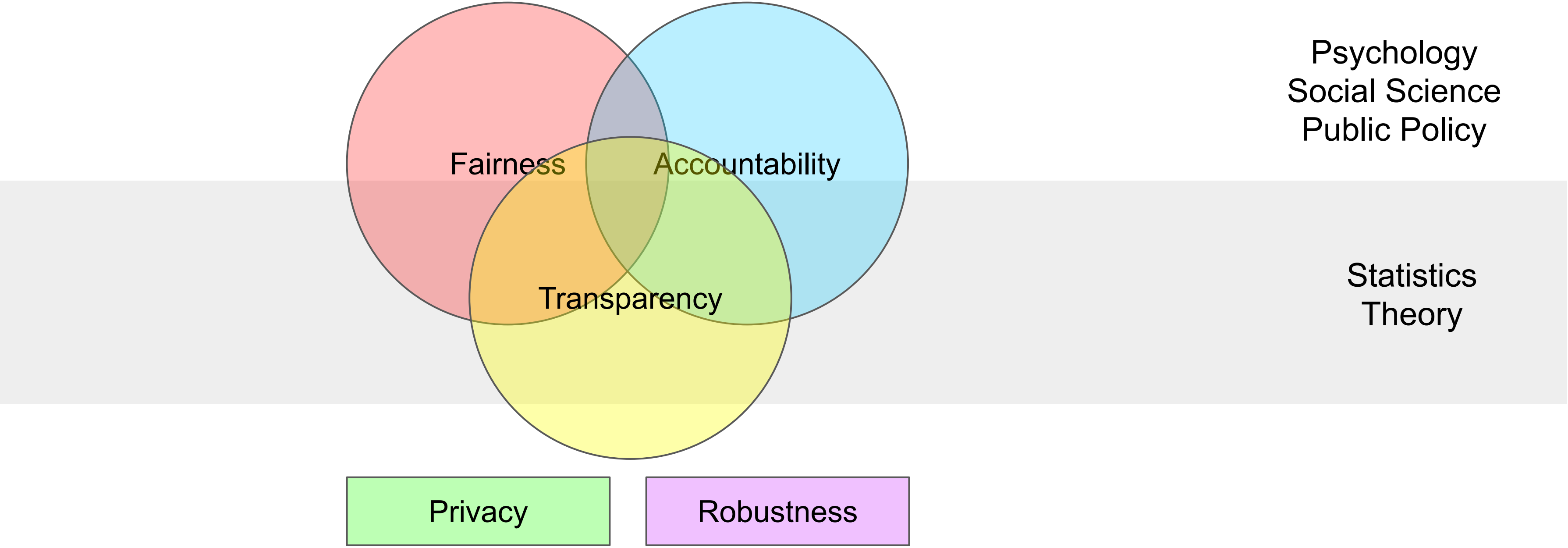
Course Overview



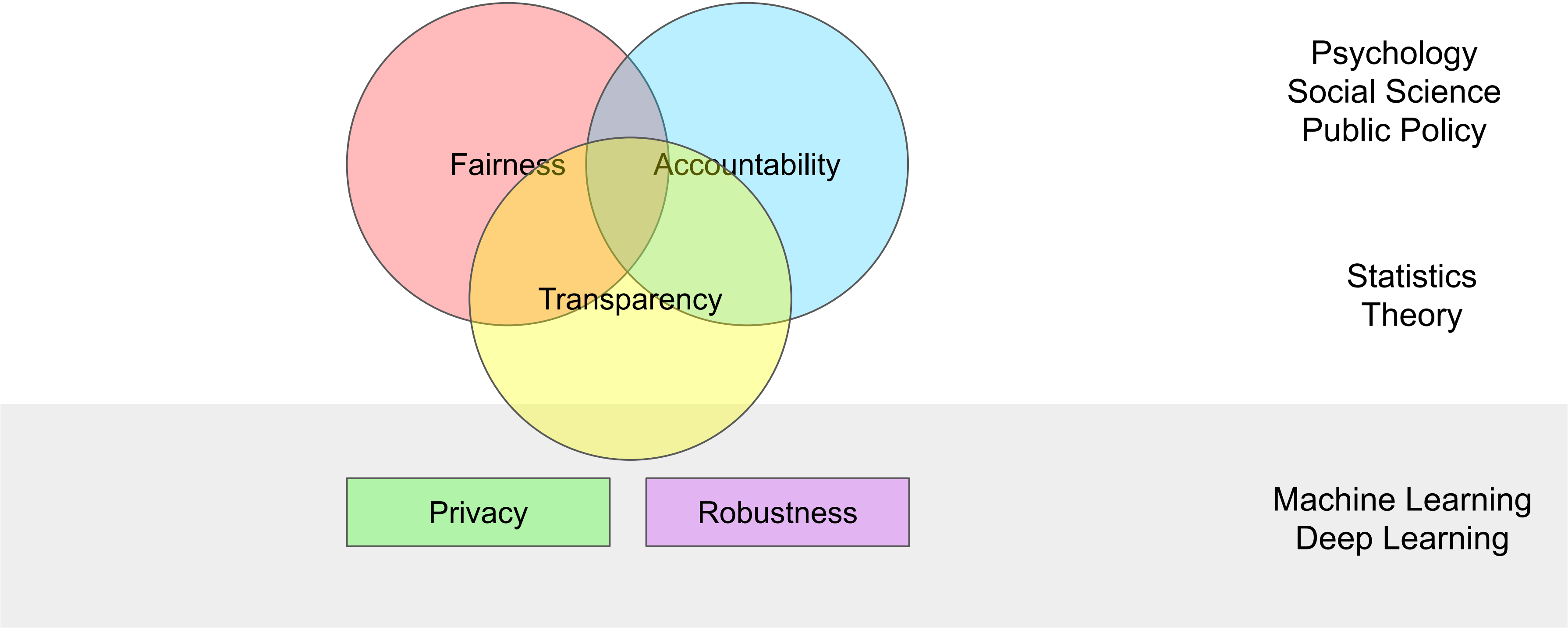
Course Overview



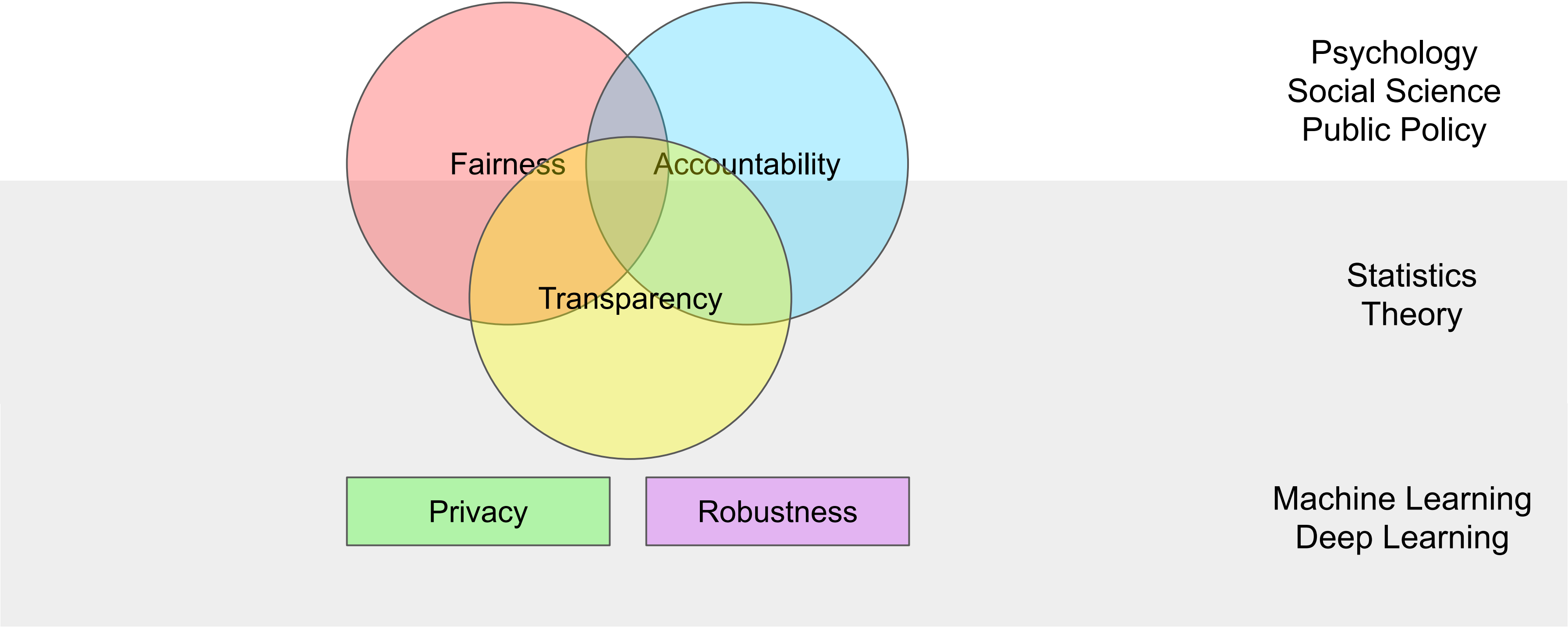
Course Overview



Course Overview



Course Overview



Logistics

- Class Meet on WF 1:30-2:50 pm
  - Apr 4 - Jun 3, 2020
  - Offered online on Zoom

- Instructors' Office Hours
  - Dr. Wei Wei, weiwei@cs.stanford.edu
    - Fridays 3:30-4:30 PM, Zoom
  - Prof. James Landay, landay@cs.stanford.edu
    - Mondays 10:00-10:30 AM and Wednesday 10:00-11:00 AM, Zoom

- TA Office Hours
  - Josh Payne, joshp007@stanford.edu
    - Fridays 10:00-11:00 AM, Zoom

Attendance Policy

- Call in for Video Lectures
- Turn on Your Camera
- Interactions Are Encouraged

Grading

- 25% Reading Assignments
  - Read one of the suggested articles for each class
  - Due 1 week after each class

- 75% Project
  - Project Proposal (10%), Apr 22 before class
  - Mid-term Check
    - Milestone presentation (5%), May 1
    - Milestone report (10%), May 13 before class
  - Final Project
    - Final Presentations (30%), May 29/Jun 3
    - Final Reports (45%), Jun 3 before class

Required Readings

- We post required readings for each class
  - No need to read them ahead of time
  - Will usually contain technical details not covered in class

Reading Assignments

- Read one paper from the suggested reading list
  - Submit one assignment for each class
  - Assignments are due one week after
  - Late submissions will receive a 10% penalty/week. Submit by Jun 3 to receive grades.
  - 1/2 - 1 pages
  - The goal of these assignments is **not** about summarizing the paper

- Encourage Research Thinking in FAccT
  - How this paper changes the vision of FAccT research
  - What future directions that this paper inspired you
  - Why the paper does/doesn't seem important
  - Observations of novel methodology or methodology that seems suspect
  - Why the paper is/isn't effective at getting its message across
  - How the paper has changed your opinion or outlook on a topic

Course Project

- Complete A Deep Learning Project in FAccT
    - Innovations on algorithms are strongly encouraged
        - solve existing problems and compare performance with baselines
        - or propose new problem setting

- Project Resources
    - Guidelines
    - Datasets and Ideas

Course Projects

- 3 Project Stages
  - Project Proposal
    - Problem that you are solving
    - Datasets
    - Evaluation metrics
    - Baselines
  - Mid-term
    - Project Report- Report preliminary results
    - Presentation - Quickly present problem setting and datasets
  - Final Project
    - Project Report - A complete write up of your methods and results
    - Final Presentation - A detailed presentation of your methods and results

Course Projects

- Project Grouping
  - Consists of up to two students, divide work equally
  - No double dipping
- Bi-weekly Project Checks with TA
  - Josh will schedule bi-weekly progress checks with each of the groups
- Start Your Project NOW
  - The Spring quarter this year is one week shorter
  - Will cover two lectures of Fairness and Interpretability before proposal deadline
- Google Cloud Credit
  - Support projects that require additional computational resources

Slack Channels

- Ask questions about the class class
- Find partners for course project

Outline

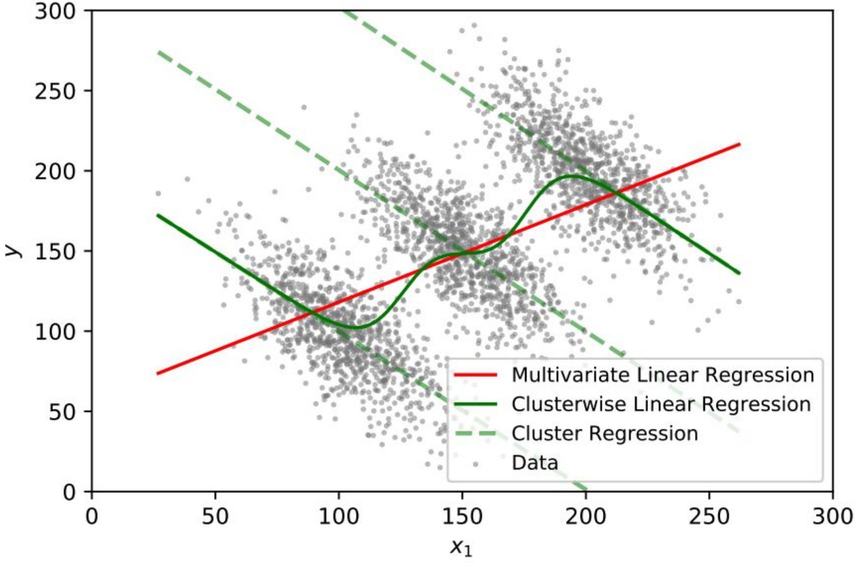
- Course Overview
  - Logistics
  - Project & Reading Assignments
- Fairness
  - Sources of Biases
  - Real World Examples
  - Sensitive Features
- Major Fairness Criteria
  - Fairness Through Unawareness

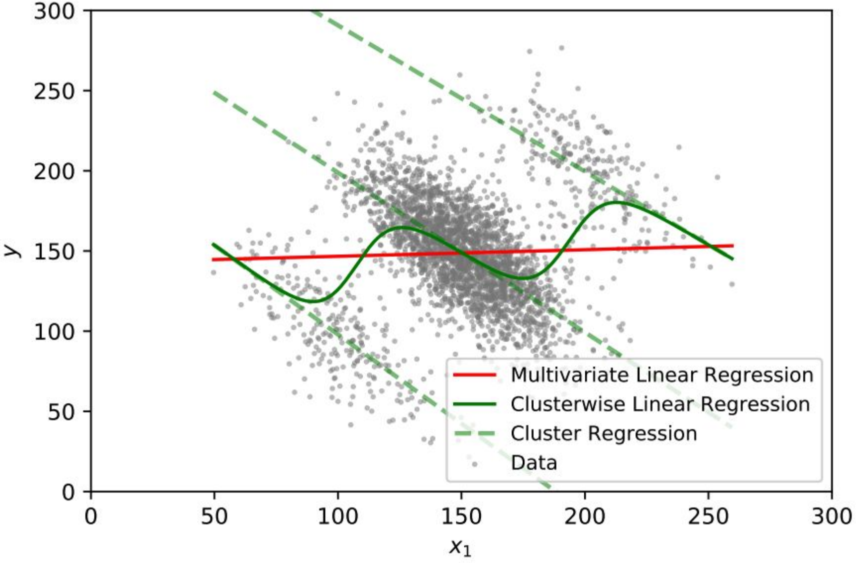
ML Fairness

- What is Fairness?
  - The absence of bias towards an individual or a group (Mehrabi et al, 2019)

- Can ML Models Discriminate?
  - Aren't machines just follow human's instructions?
  - ML models approximate patterns in the data
  - Learns/Amplifies bises at the same time

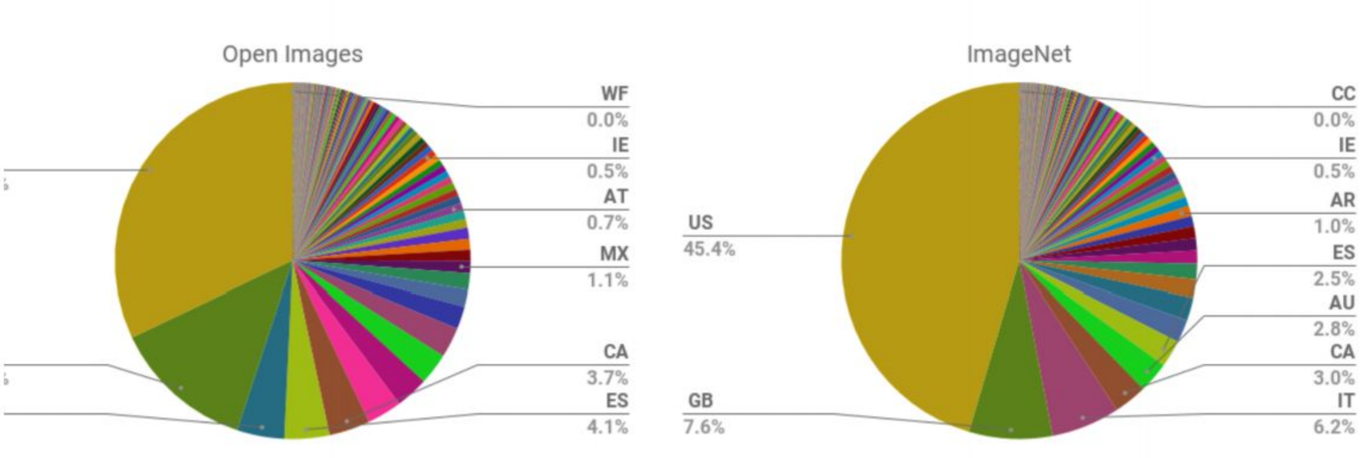
Sources of Biases





red - biased regression
dashed green - regression for each subgroup
solid green - unbiased regression

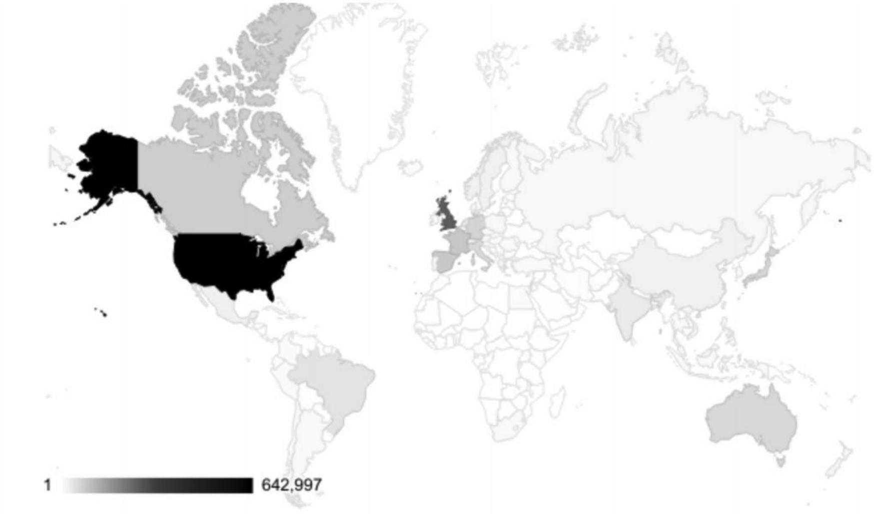
Geographical Representation of ImageNet and Open Images



[Mehrabi et al, 2019](#)

Geographical Representation of Open Images

- One third of the data was collected in US
- 60% of the data was from the six most represented countries.



Mehrabi et al, 2019

Graduate School Admissions to UC Berkeley, 1973

	Men		Women	
	Applicants	**Admitted**	**Applicants**	**Admitted**
Total	8442	**44%**	4321	35%

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Graduate School Admissions to UC Berkeley, 1973

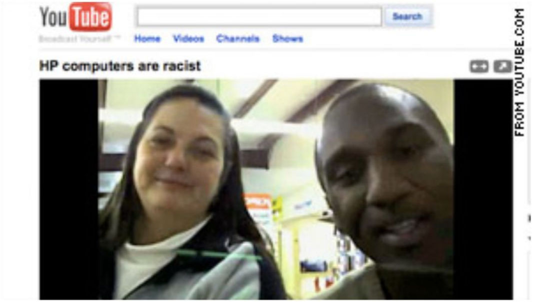
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

https://en.wikipedia.org/wiki/Simpson%27s_paradox

real world example of fairness

HP looking into claim webcams can't see black people

By **Mallory Simon**, CNN
December 23, 2009 7:25 p.m. EST



A YouTube video shows co-workers trying out an HP webcam with motion-tracking and facial recognition software.

(CNN) -- Can Hewlett-Packard's motion-tracking webcams see black people? It's a question posed on a now-viral YouTube video and the company says it's looking into it.

In the video, two co-workers take turns in front of the camera -- the webcam appears to follow Wanda Zamen as she sways in front of the screen and stays still as Desi Cryer moves about.

HP acknowledged in a statement e-mailed to CNN that the cameras may have issues with contrast recognition in certain lighting situations. The webcams, built into HP's new computers, are supposed to keep people's faces and bodies in proportion and centered on the screen as they move.

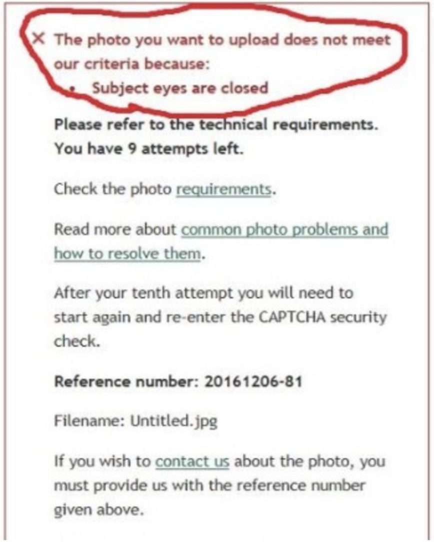
The video went viral over the weekend, garnering more than 400,000 YouTube page views and a slew of comments on Twitter.

STORY HIGHLIGHTS

- **NEW:** Video was meant to be humorous showing of software glitch, co-workers say
- Co-workers: Motion-tracking webcam moves with white woman, not black man
- "I think my blackness is

New Zealand passport robot thinks this Asian man's eyes are closed

By James Griffiths, CNN
Updated 1:46 AM ET, Fri December 9, 2016


✕ The photo you want to upload does not meet our criteria because:
- Subject eyes are closed

**Please refer to the technical requirements.
You have 9 attempts left.**

Check the photo requirements.

Read more about common photo problems and how to resolve them.

After your tenth attempt you will need to start again and re-enter the CAPTCHA security check.

Reference number: 20161206-81

Filename: Untitled.jpg

If you wish to contact us about the photo, you must provide us with the reference number given above.

INSTAGRAM.COM/RICHYFANCY

New Zealand's online passport application system couldn't recognize Richard Lee's open eyes.

More from CNN

Two workers at the same Walmart store die of coronavirus


Trump fires intelligence community watchdog who told Congress...

Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones
10/10/18 10:32AM • Filed to: ALGORITHMS

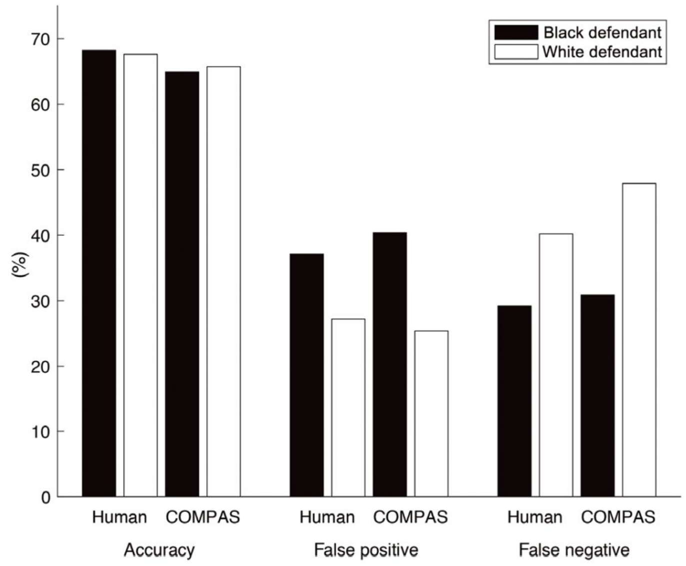
79 3

Photo: Getty

Recent Video

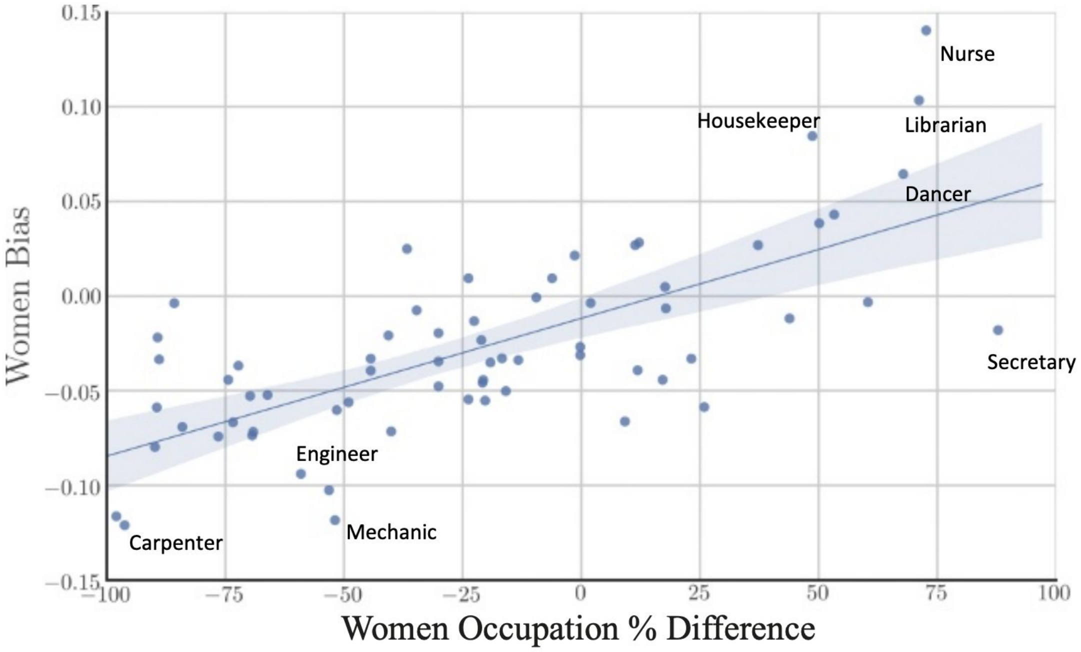
Criminal Justice (Dressel et al, 2018)

- Commercial risk assessment software known as COMPAS
  - assess more than 1 million offenders since 2000
  - predicts a defendant's risk of committing a misdemeanor or felony 137 features

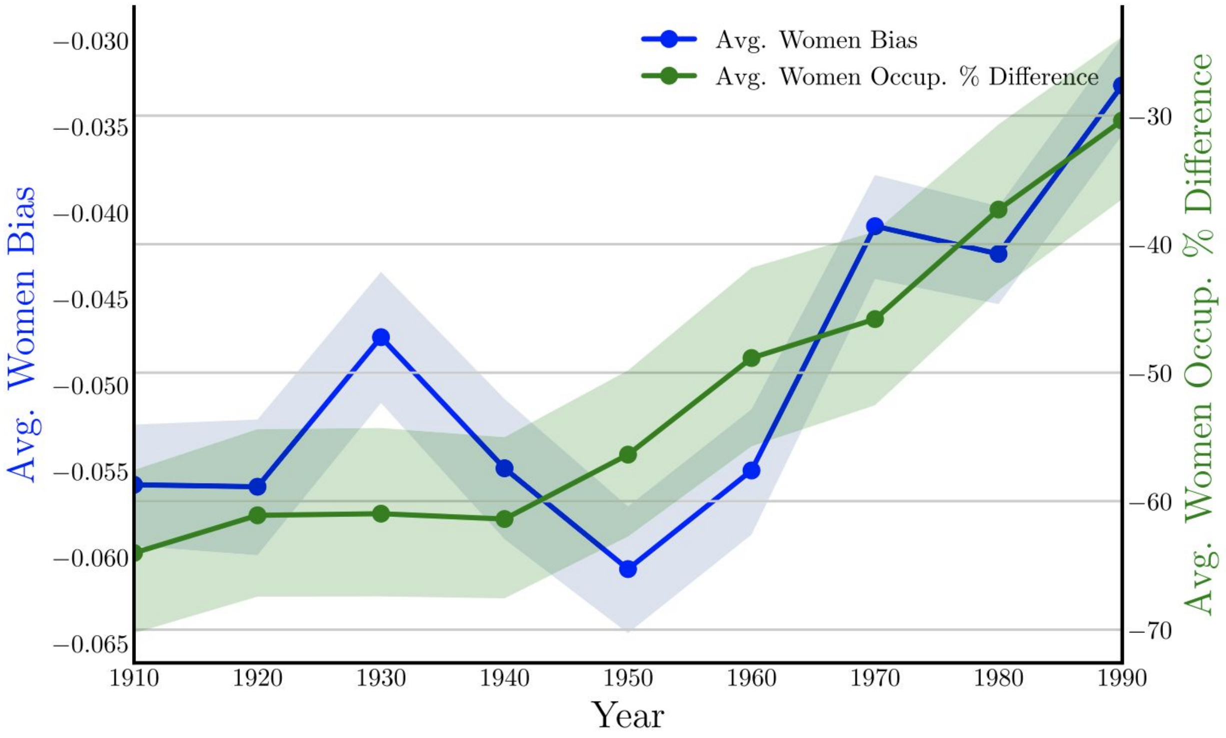


Biases in Word Embedding (Gard et al, 2018)

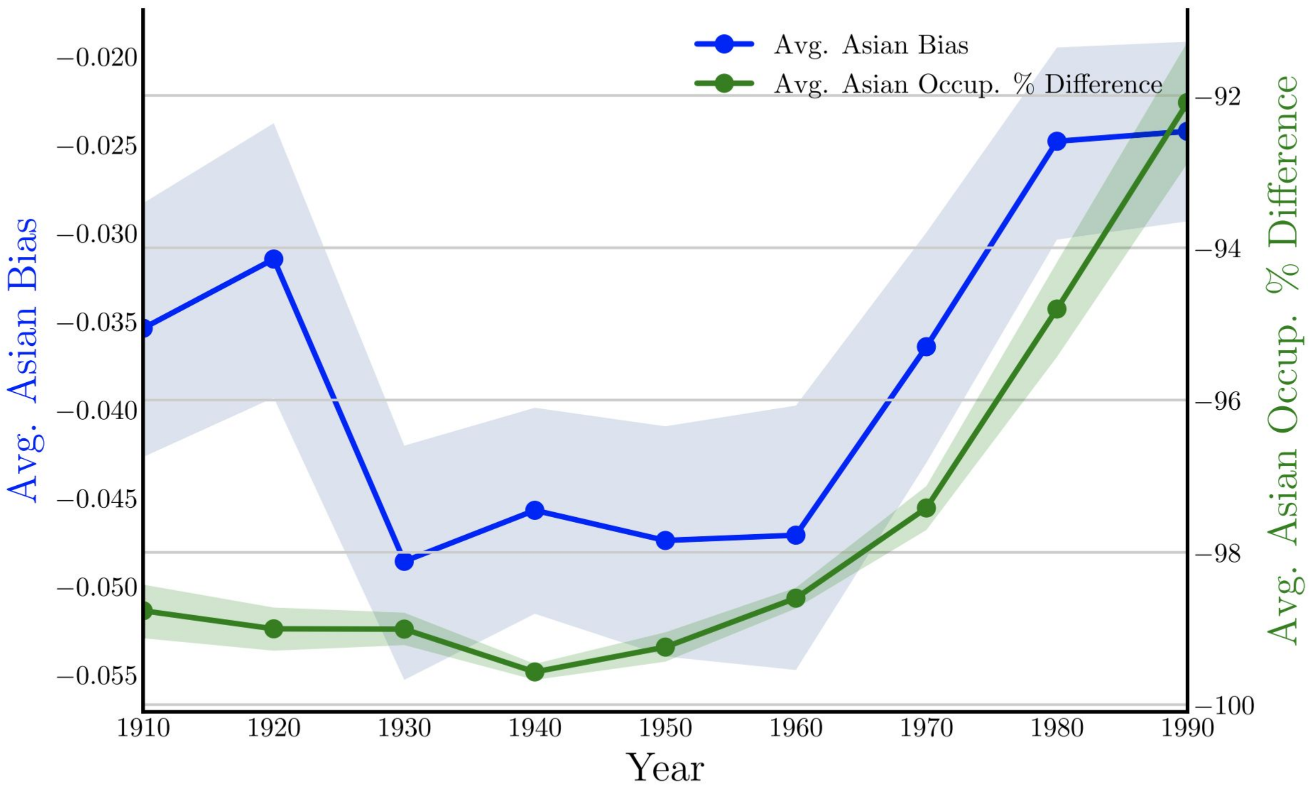
Gender Biases and Occupation



Average Biases Over Time for Woman



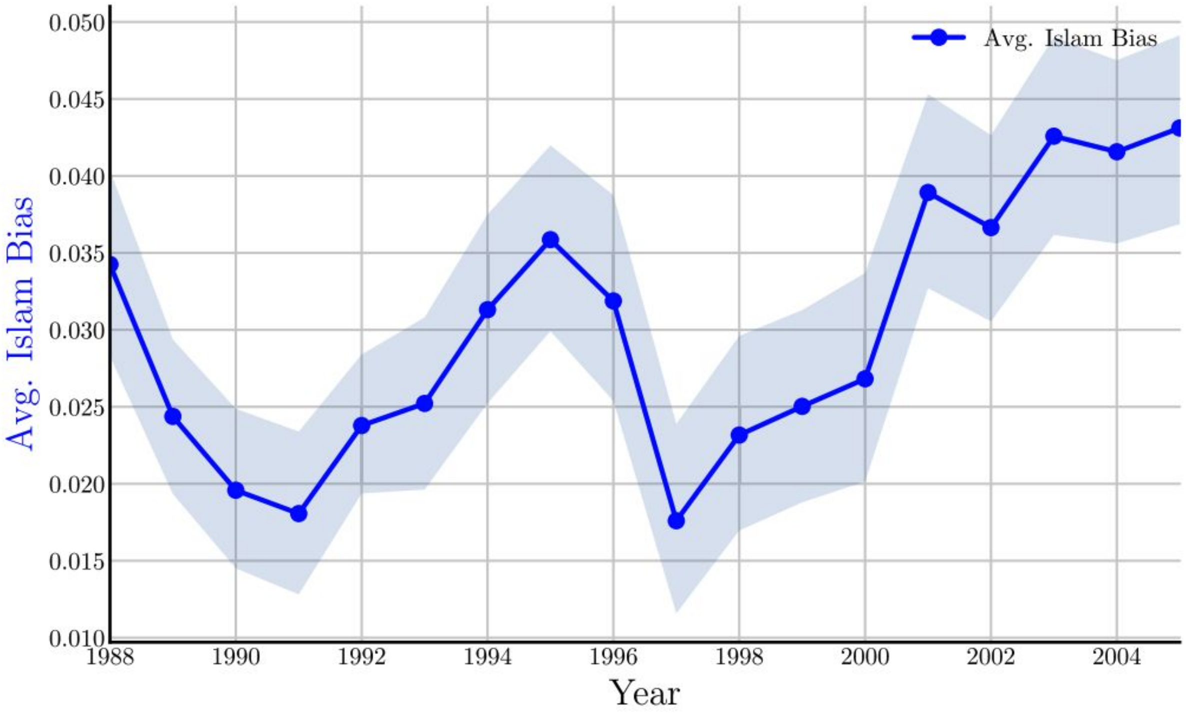
Average Biases Over Time for Asian



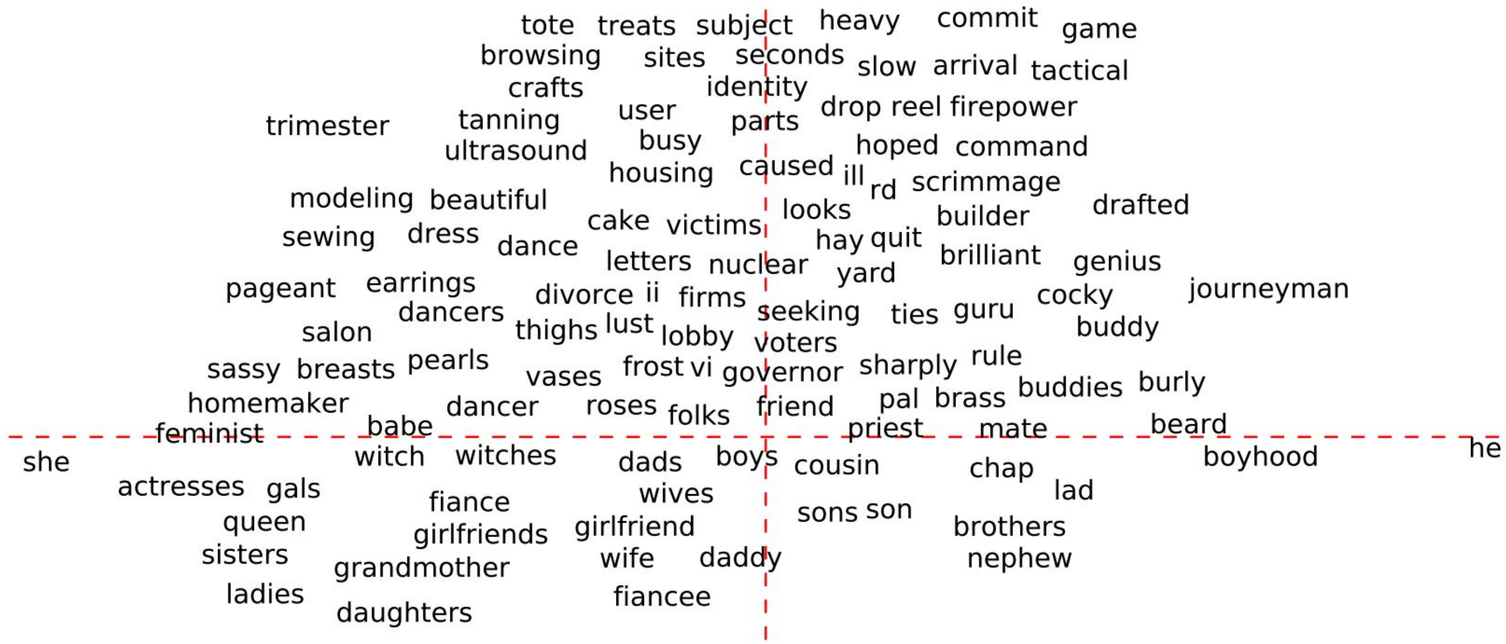
Top 10 Occupations and Ethnic Groups

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

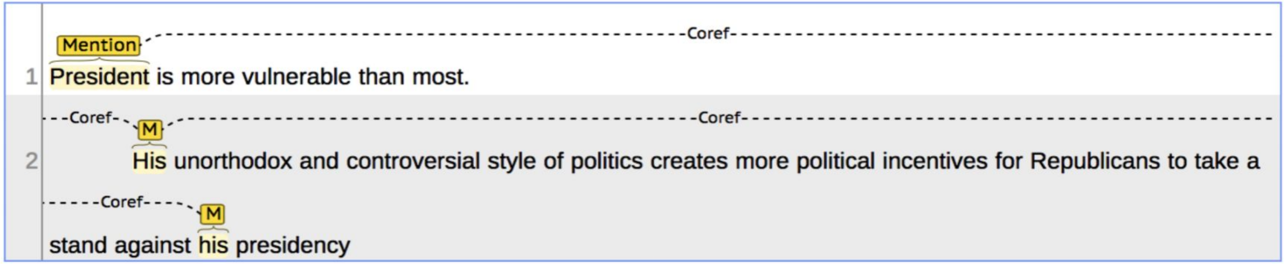
Religious Bias Related to Terrorism



Words Projected into Gender Axes (Bolukbasi et al, 2016)



Coreference Resolution (Zhao et al, 2018)



his ⇒ her

- Stereotypical dataset

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

- Anti-stereotypical dataset

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because he was highly recommended.

Removing Gender Information (Wang et al, 2019)

COCO Results

imSitu Results

Sensitive (Protected) Features

- Sensitive Features
  - Identify a group
  - e.g., gender, ethnics

- Discrimination Occurs
  - When Sensitive Features Are Used Improperly
  - May lead to ML Discrimination

List of Protected Attributes Specified in US Fair Lending Laws

- Fair Housing Acts (FHA)
- Equal Credit Opportunity ACts (ECOA)

Attribute	FHA	ECOA
Race	✓	✓
Color	✓	✓
National origin	✓	✓
Religion	✓	✓
Sex	✓	✓
Familial status	✓	
Disability	✓	
Exercised rights under CCPA		✓
Marital status		✓
Recipient of public assistance		✓
Age		✓

Outline

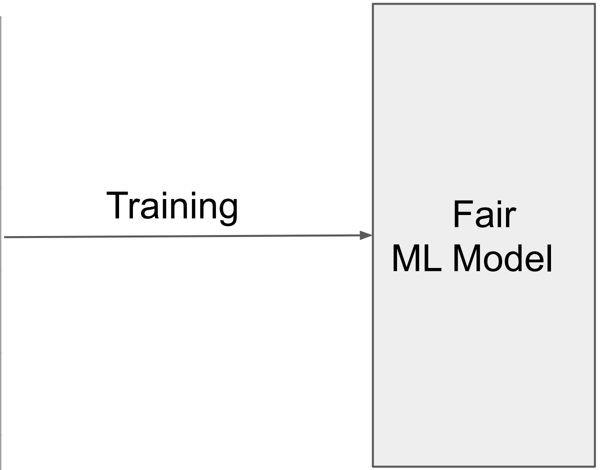
- Course Overview
  - Logistics
  - Project & Reading Assignments
- Fairness
  - Sources of Biases
  - Real World Examples
  - Sensitive Features
- Major Fairness Criteria
  - Fairness Through Unawareness

Fairness Through Unawareness

- A ML Algorithm Achieves Fair Through Unawareness If
  - None of the sensitive features are directly used in the model

Protected

~~Race and Ethnicity~~	Skills	Years of Exp	Hired?
~~Hispanic~~	Javascript	1	no
~~Hispanic~~	C++	5	yes
~~White~~	Java	2	yes
~~White~~	C++	3	yes



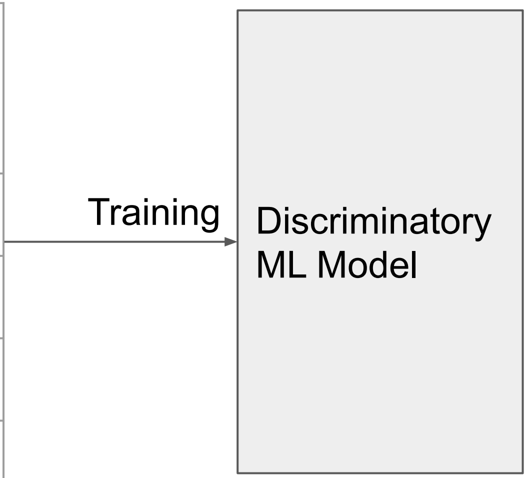
Issues With Fairness Through Unawareness

- Sensitive Features May Still Be Used
  - Inferred from indirect evidence

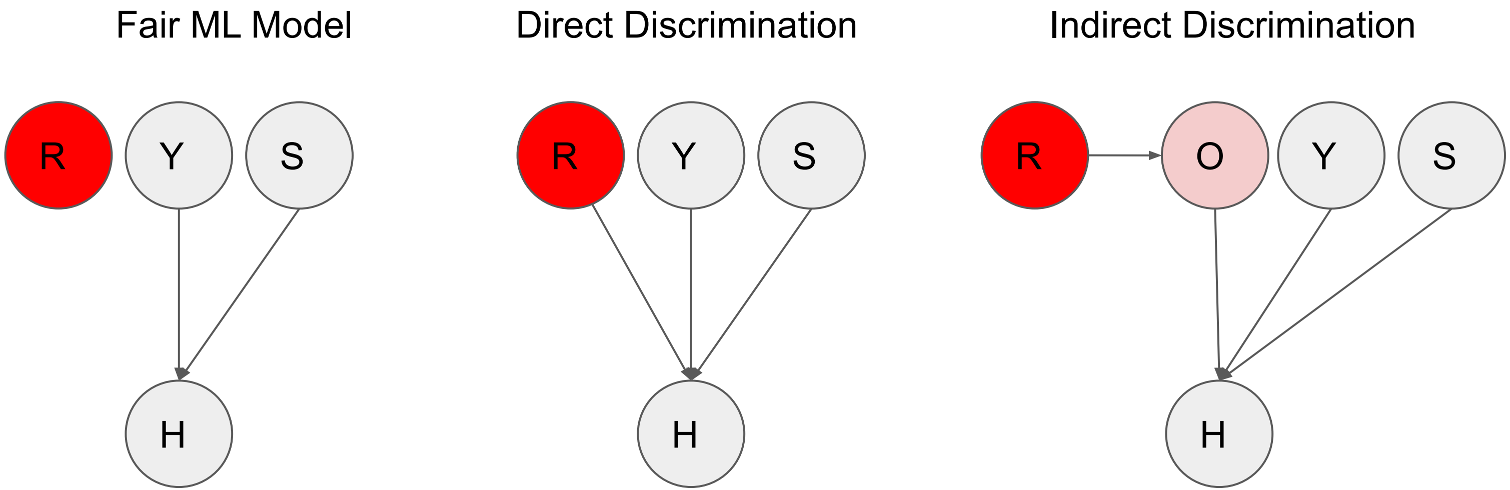
Inferred

Protected

~~Race and Ethnicity~~	Skills	Years of Exp	Often Goes to Mexican Markets	Hiring Decision
~~Hispanic~~	Javascript	1	yes	no
~~Hispanic~~	C++	5	yes	yes
~~White~~	Java	2	no	yes
~~White~~	C++	3	no	yes



Types of Discriminations



R - Race
Y - Years of Exp

S = Skills
O = Often Goes to Mexico Market

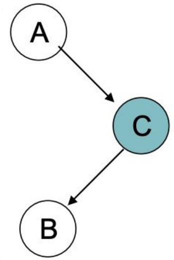
Conditions for Direct Discrimination

- A - set of protected features
- X - set of features other than protected features

- A predictor $\hat{Y}$ is direct discrimination if
  - $P(\hat{Y} \mid X, A) \neq P(\hat{Y} \mid X)$
  - i.e., $\hat{Y} \not\perp A \mid X$

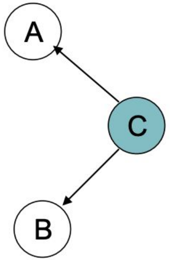
Conditional Independence

- Common Conditions for Conditional Independence



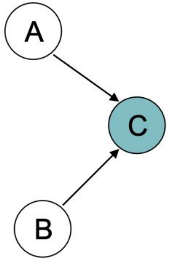
Head to Tail

$A \perp\!\!\!\perp B \mid C$



Tail to Tail

$A \perp\!\!\!\perp B \mid C$



Head to Head (collider)

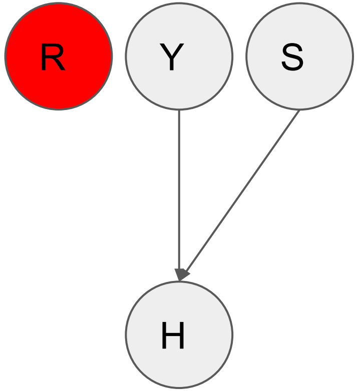
$A \not\perp\!\!\!\perp B \mid C$

More Sophisticated Cases

D-Separator

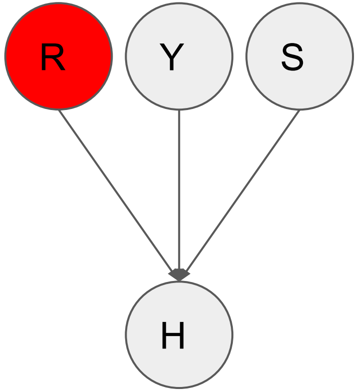
Types of Discriminations

Fair ML Model



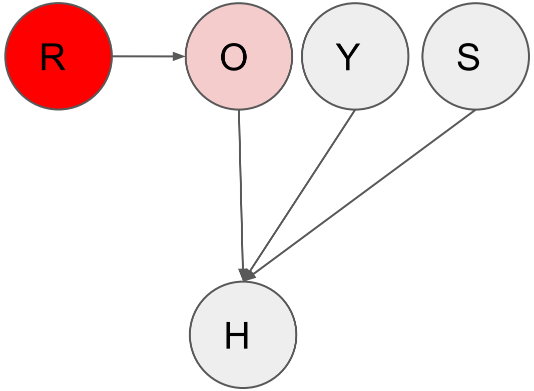
$H \perp\!\!\!\perp R \mid \{Y,S\}$
not connected

Direct Discrimination



$H \not\perp\!\!\!\perp R \mid \{Y, S\}$
(head to head)

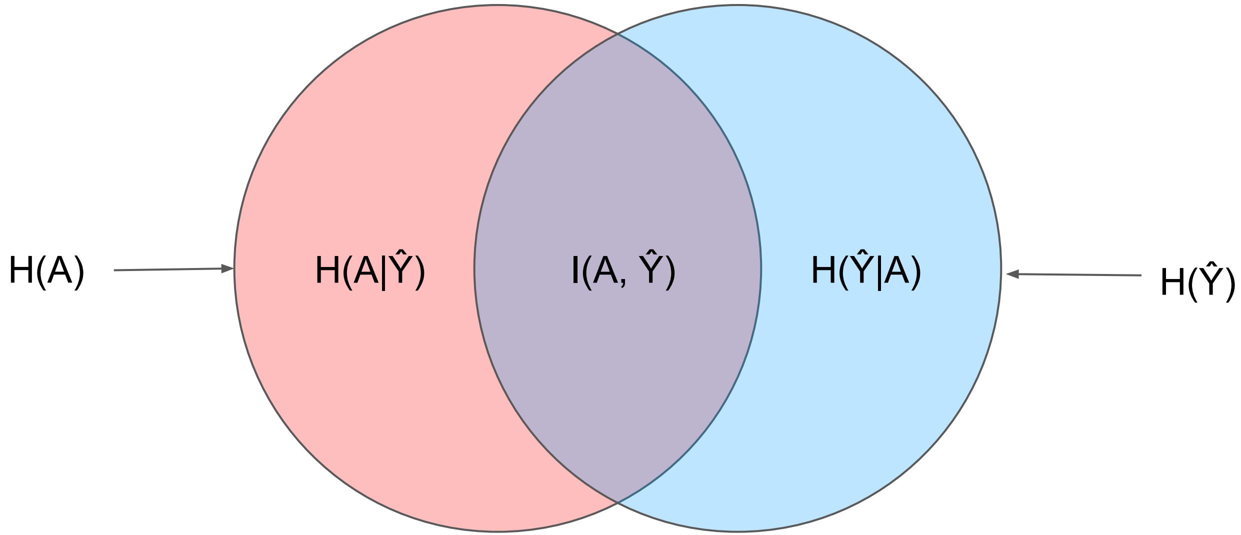
Indirect Discrimination



$H \perp\!\!\!\perp R \mid \{Y,S, O\}$
(head to tail)

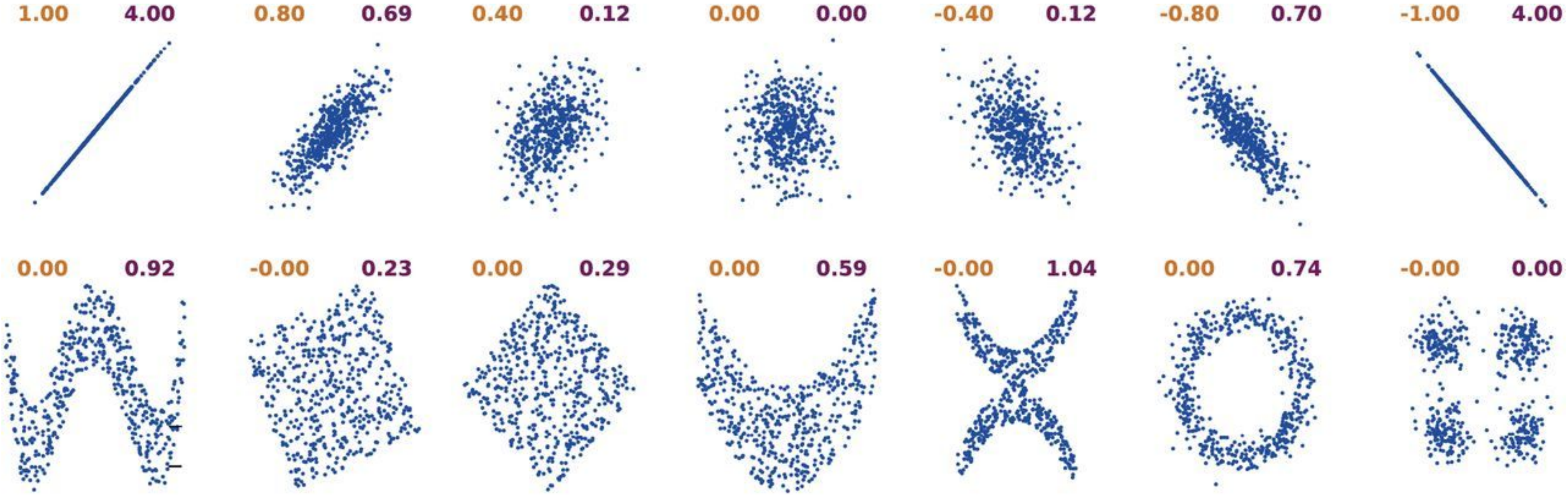
Conditions for Indirect Discrimination

- Mutual Information
  - A measure of the mutual dependence between A and $\hat{Y}$
  - $I(A, \hat{Y}) = H(A) - H(A \mid \hat{Y}) = H(\hat{Y}) - H(\hat{Y} \mid A)$
  - $I(A, \hat{Y}) = 0$ if $P(\hat{Y} \mid A) = P(\hat{Y})$, or $A \perp\!\!\!\perp \hat{Y}$



Correlation Coefficient and Mutual Information

- Correlation Coefficient (left) and Mutual Information (right)



Ince et al, 2016

Limitations

- Processing Sensitive Features
  - Fairness through unawareness requires sensitive features to be masked out
  - Not easy to do in real life
  - Referred to as individual fairness criteria

❖ Stereotypical dataset

The physician hired the secretary because he was overwhelmed with clients.

The physician hired the secretary because she was highly recommended.

❖ Anti-stereotypical dataset

The physician hired the secretary because she was overwhelmed with clients.

The physician hired the secretary because he was highly recommended.

Required Reading

- Barocas: Ch 2
- Bishop: Ch 8.2

Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
Barocas, S., Hardt, M., & Narayanan, A. Fairness and Machine Learning, 2018.

Reading Assignments (Pick One)

- Gajane, Pratik, and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. arXiv 2017
- Luong, Binh Thanh, Salvatore Ruggieri, and Franco Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. SIGKDD 2011
- Hardt, M., Price, E., & Srebro, N. Equality of opportunity in supervised learning. NeurIPS 2016
- Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. SIGKDD 2008
- Zafar, Muhammad Bilal, et al. Fairness Constraints: Mechanisms for Fair Classification.AIStats 2017

Next Lecture

Fair Representation Learning