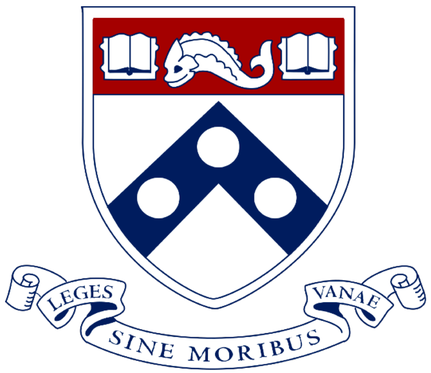
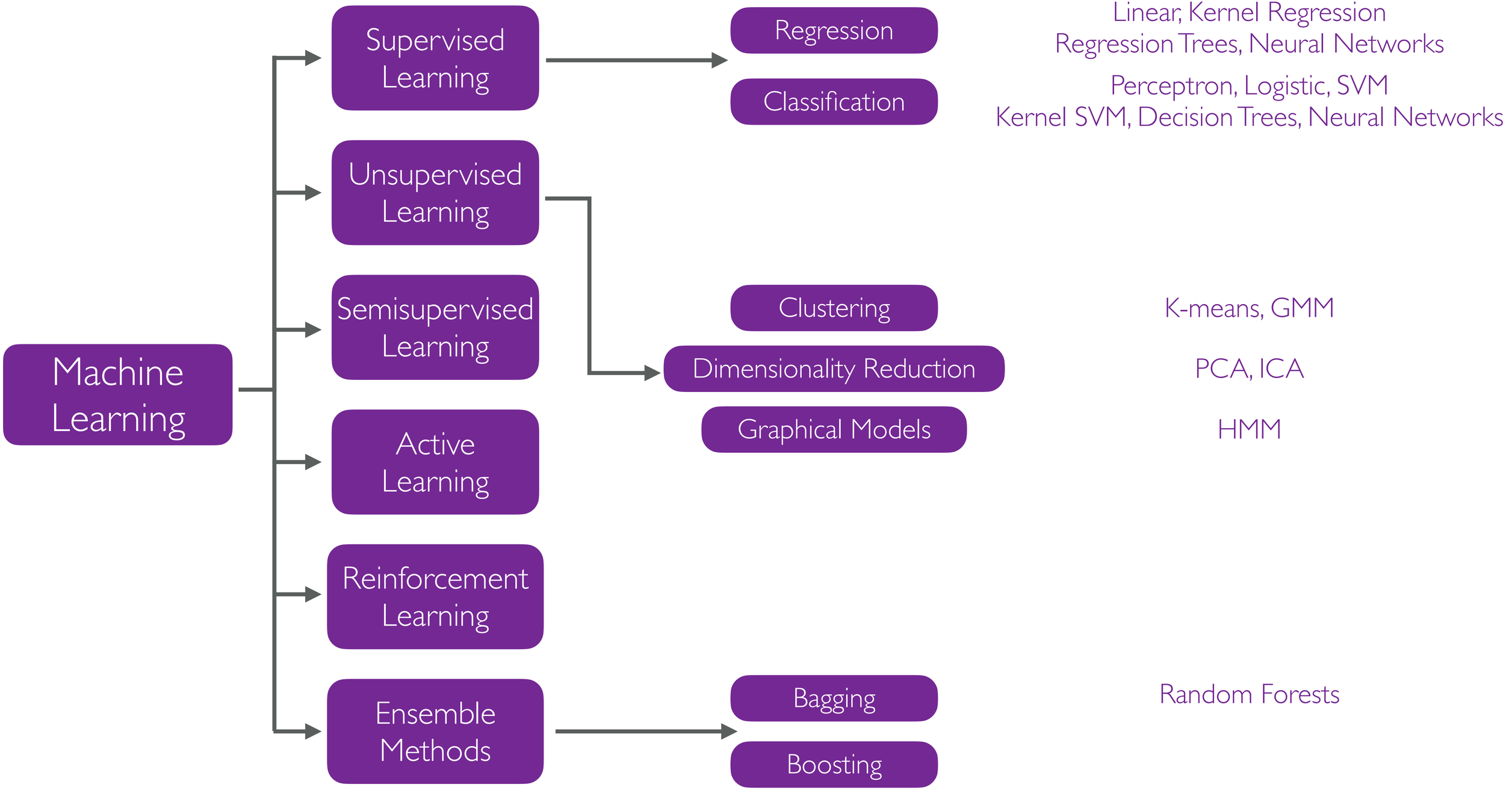
CIS 5200: MACHINE LEARNING

RISKS AND CHALLENGES

Surbhi Goel and Eric Wong



Spring 2023


Machine Learning
Supervised Learning
Unsupervised Learning
Semisupervised Learning
Active Learning
Reinforcement Learning
Ensemble Methods
Regression
Classification
Clustering
Dimensionality Reduction
Graphical Models
Bagging
Boosting
Linear, Kernel Regression
Regression Trees, Neural Networks
Perceptron, Logistic, SVM
Kernel SVM, Decision Trees, Neural Networks
K-means, GMM
PCA, ICA
HMM
Random Forests

MACHINE LEARNING - IN THIS COURSE

Goal: Minimize some loss over a function class using available data to learn a predictor

Questions we did not ask?

- Should we be solving the problem using ML in the first place?
- Is the objective we train with the one we actually desire?
- Are the models leaking information about the data it is using?
- What impact does our model have on the world?
- What happens when humans interact with our models?

These are very important questions!

'It Happened So Fast': Inside a Fatal Tesla Autopilot Accident

A 2019 crash in Florida highlights how gaps in Tesla's driver-assistance system and distractions can have tragic consequences.

SCIENCE

What happens when an algorithm cuts your health care

By **COLIN LECHER** / **@colinlecher**
Illustrations by **WILLIAM JOEL**; Photography by **AMELIA HOLOWATY KRALES**

"…the algorithm appeared more likely to interpret images with rulers as malignant" – Narla et al. 2018

"With only black and white stickers… we can cause 100% misclassification" – Eykholt et al. 2018

Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter

Attempt to engage millennials with artificial intelligence backfires hours after launch, with TayTweets account citing Hitler and supporting Donald Trump

Amazon’s Face Recognition Falsely Matched 28 Members of Congress With Mugshots

Disinformation Researchers Raise Alarms About A.I. Chatbots

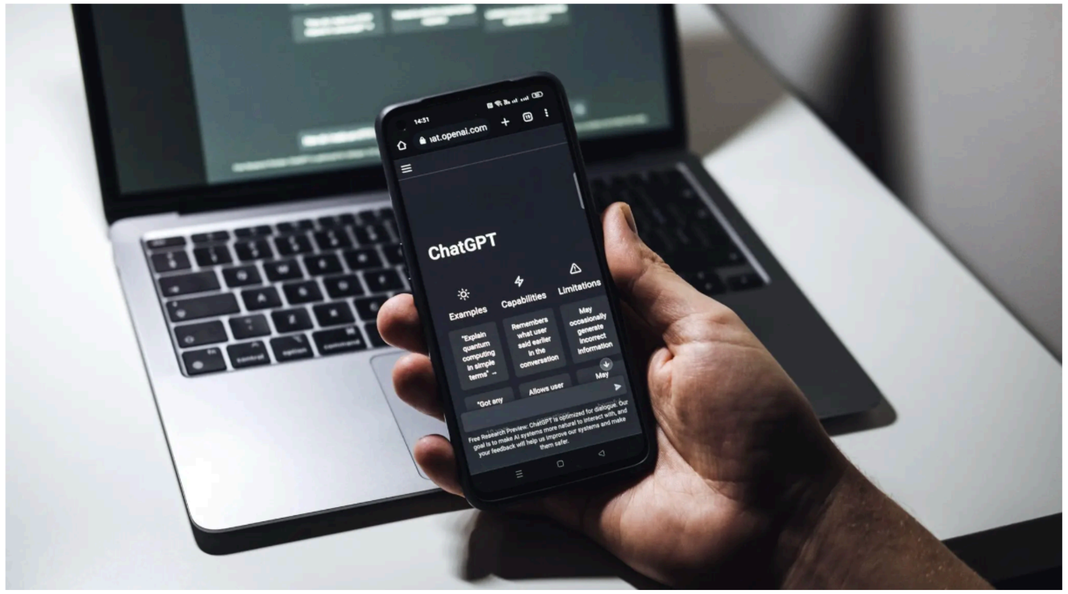
Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.

Give this article 183

Gamestop and WallStreetBets drove A.I.-powered hedge funds to their worst month on record – Fortune

Disinformation Researchers Raise Alarms About A.I. Chatbots

Researchers used ChatGPT to produce clean, convincing text that repeated conspiracy theories and misleading narratives.



'Thousands of Dollars for Something I Didn't Do'

Because of a bad facial recognition match and other hidden technology, Randal Reid spent nearly a week in jail, falsely accused of stealing purses in a state he said he had never even visited.

Give this article 643

http://gendershades.org/

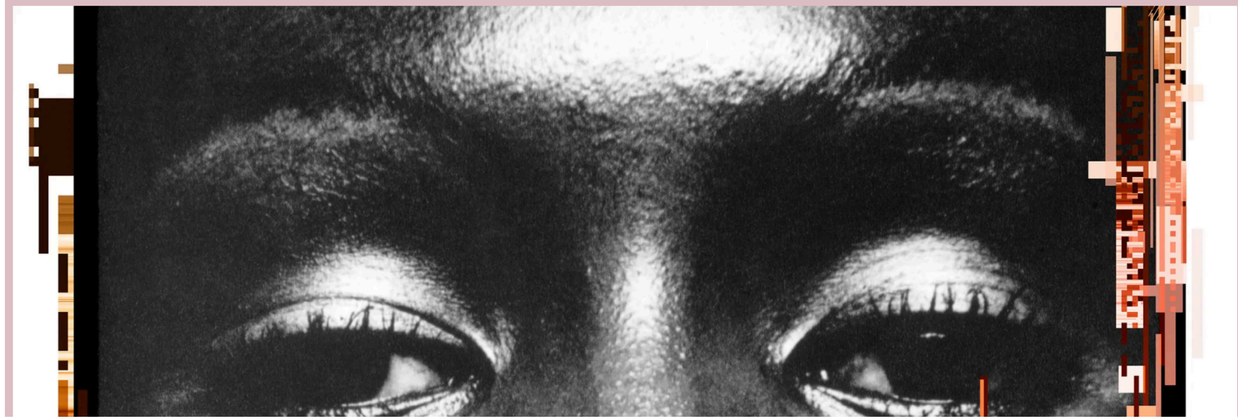
CHECK YOUR BIAS

AI Facial Recognition Systems Work the Worst for Black Women

After experiencing algorithmic injustice firsthand, Dr. Joy Buolamwini became determined to fight against racial and gender bias in the field.

BY DIANNA MAZZONE

August 3, 2022



BIAS - A CHALLENGE

Decisions ML algorithms are making today:

- Policing/judicial decisions
- Loan decision
- Job resume filtering
- Personalized recommendations

ML models are often biased against minorities!

COMPAS: CORRECTIONAL OFFENDER MANAGEMENT PROFILING FOR ALTERNATIVE SANCTIONS

Used in prisons across country: AZ, CO, DL, KY, LA, OK, VA, WA, WI

Recidivism = likelihood of criminal to reoffend

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb

BIAS - WHERE DOES IT COME FROM?

- **Data**:
  - Historical bias — Existing bias in the world that shows up in the data
  - Representation bias — Amount of data available for minority groups is limited
  - Measurement bias — Features are often noisy proxies of actual quantities of interest

Datasheets for Datasets https://arxiv.org/pdf/1803.09010.pdf

- **Model**: Some models prefer certain patterns that induce biases
- **Evaluation**: Certain evaluation metrics may lead to biased models

BIAS - WHERE DOES IT COME FROM?

Machine learning models when deployed in practice exacerbate bias from their predictions

- **Positive Feedback loop**:
  - College rankings: Ranking was self-enforcing
- **Negative Feedback loop**:
  - PredPol: Policing creates more reporting which reinforces more policing

FAIRNESS - WHAT IS FAIR?

Example: Baking team selection using dough kneading speed

- x is the input features of a baker, e.g. dough needing speed
- a indicates the group that the baker belongs to, e.g. group ♦ or ♣
- y is the true label, e.g. $+1$ if baker will win the competition, and -1 if the baker will fail
- $\hat{y} = f(x)$ is the prediction of our model f

DEFINITION 1 - UNAWARENESS

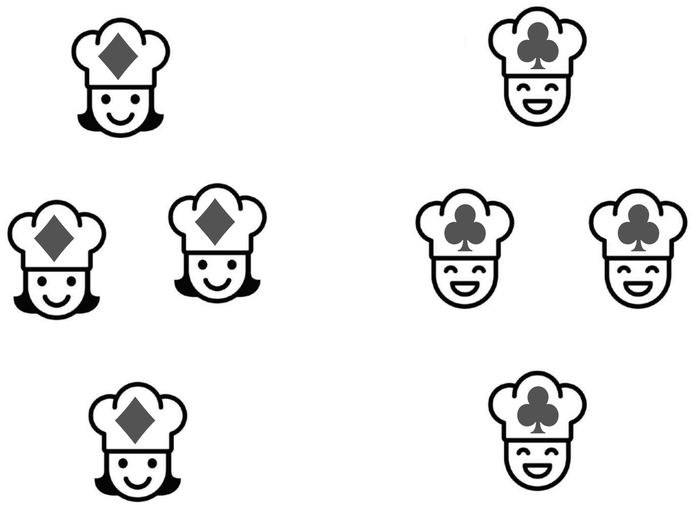
To avoid unfair decisions, do not allow model to see the protected attribute (suit of baker)

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)

Challenge: Other features may be correlated with the suit of the baker

In general, protected attributes can often be inferred by other features

DEFINITION 2- DEMOGRAPHIC PARITY



Team accept decisions should be equivalent across the two suits of bakers

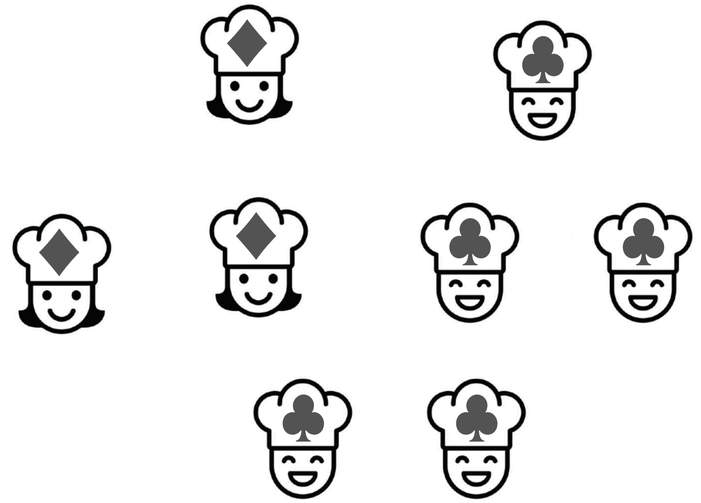
$$\Pr[\hat{y} = 1 \,|\, a = \blacklozenge] = \Pr[\hat{y} = 1 \,|\, a = \clubsuit]$$

Challenge: Might be too strong to treat every group exactly equal.
Not strong enough, does not care about errors

This is satisfied by selecting the best 20% from suit ♦

and a random 20% from suit ♣

DEFINITION 3- EQUAL OPPORTUNITY



Bakers from each suit have equal chance of being selected given their individual abilities

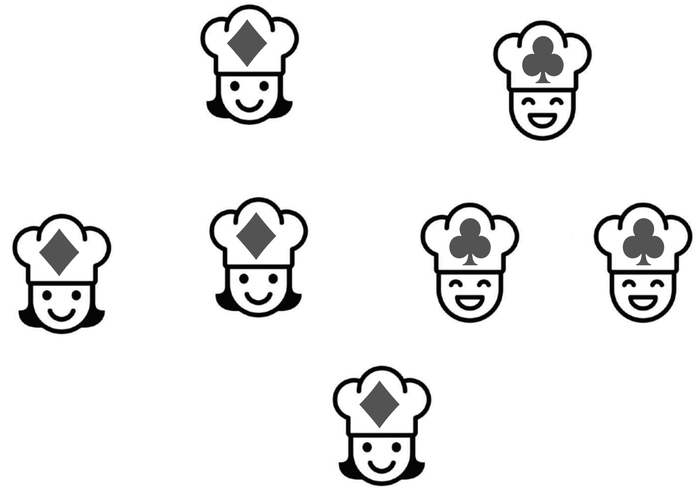
$$\Pr[\hat{y} = 1 \mid a = \blacklozenge, y = 1] = \Pr[\hat{y} = 1 \mid a = \clubsuit, y = 1]$$

Challenge: Does not account for different access to resources

What if ♦ bakers had access to limited amount of flour to practice kneading?

DEFINITION 4- PREDICTIVE PARITY

Our choice of selecting a baker on the team should reflect their true abilities to win/lose



$$\Pr[y = 1 \mid a = \blacklozenge, \hat{y} = 1] = \Pr[y = 1 \mid a = \clubsuit, \hat{y} = 1]$$

Challenge: Similar issues as equal odds

DEFINITIONS - MANY

	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓
3.3.2	Well calibration	[16]	81	✓
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Table 1: Considered Definitions of Fairness

Arvind Narayanan ✓
@random_walker

I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!
Here it is (with minor edits): docs.google.com/document/d/1bn...
See you on Feb 23/24.

> **Arvind Narayanan** ✓ @random_walker · Nov 6, 2017
> When I tell my computer science colleagues that there are so many fairness definitions, they are often surprised and/or confused. [Thread]
> twitter.com/random_walker/...
> Show this thread

4:24 PM · Jan 8, 2018

FAIRNESS - INCOMPATIBILITY

Turns out that definition 2-4 are incompatible, all three cannot be satisfied together!

Inherent Trade-Offs in the Fair Determination of Risk Scores

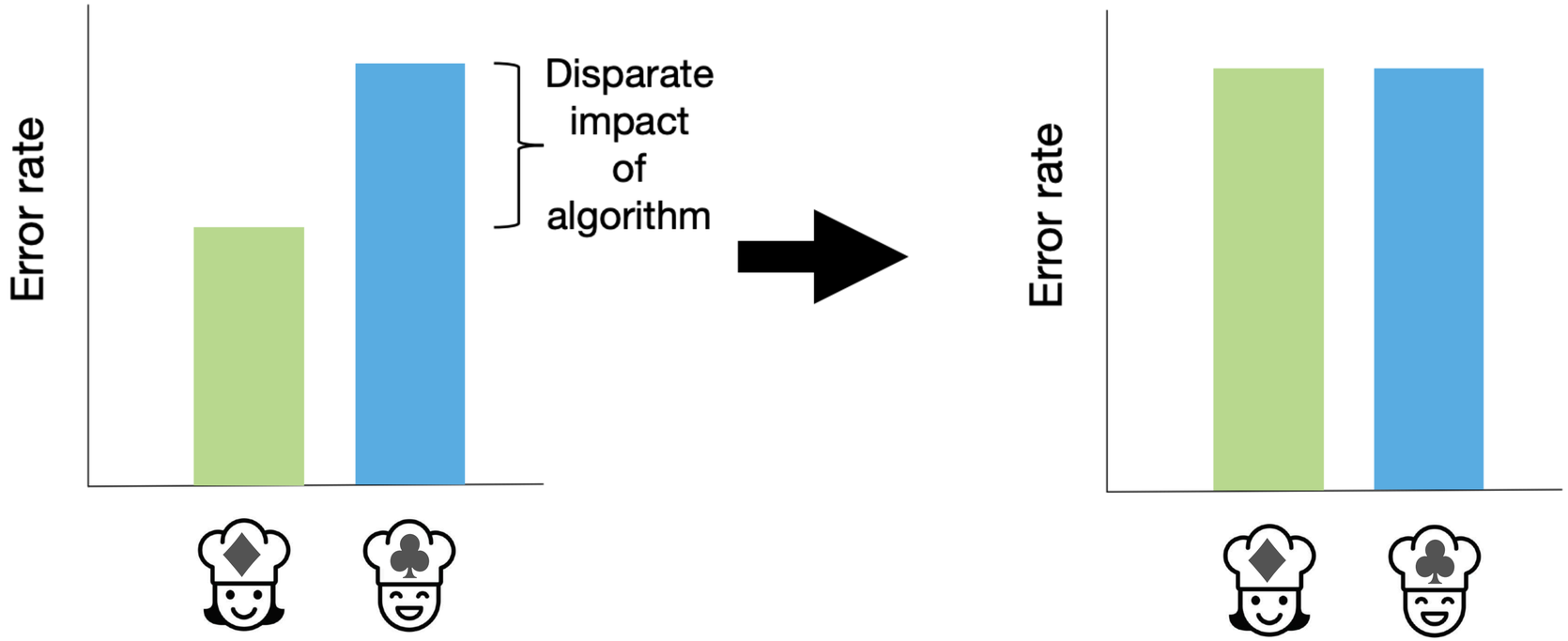
Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

FAIRNESS - AT ODDS WITH ACCURACY

Is fairness at odds with accuracy?



FAIRNESS - ALGORITHMS

- Remove dependence on sensitive attributes from the data
- Add fairness goal as a constraint in the optimization problem
- Post-process predictions to uncorrelated with sensitive attributes

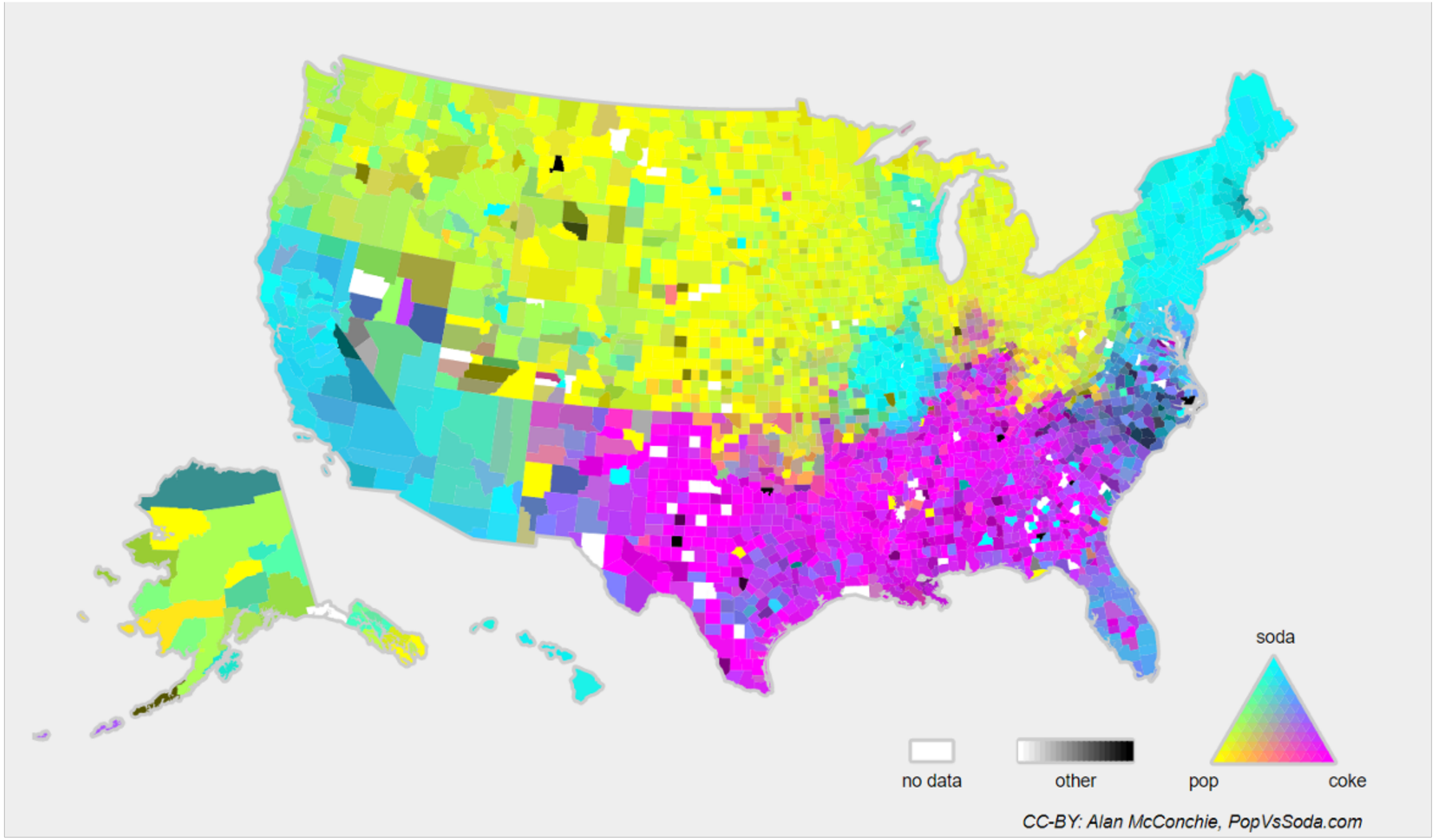
Goodhart's law:
"When a measure becomes a target, it ceases to be a good measure" – Marilyn Strathern

Have to be super careful about why, what, and how we are doing things!

ROBUSTNESS - TO WHAT?

- **Natural shifts in the data**:
  - Changes over time (temporal robustness)
  - Shifts across subpopulations (group robustness)
  - Differences between geographical or cultural regions (domain shift)
- **Adversarial changes:**
  - Input perturbations at test time (adversarial robustness)
  - Malicious data inserted at train time (data poisoning or backdoors)

DISTRIBUTION SHIFT - POP / SODA / COKE



DISTRIBUTION SHIFT - CONTENT MODERATION

Goal: Predict whether online comments are toxic

- Good test accuracy (92%) but…

Demographic	Test accuracy on non-toxic comments
Male	87.3 (0.7)
Female	89.0 (0.6)
LGBTQ	74.6 (0.5)
Christian	92.1 (0.2)
Muslim	80.9 (1.0)
Other religions	86.1 (0.1)
Black	**69.2** (1.3)
White	71.2 (1.4)

- Poor performance on subpopulations

ROBUSTNESS - DISTRIBUTION SHIFT

Can factorize a distribution $p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y)$ and characterize changes in the distribution as one of the following:

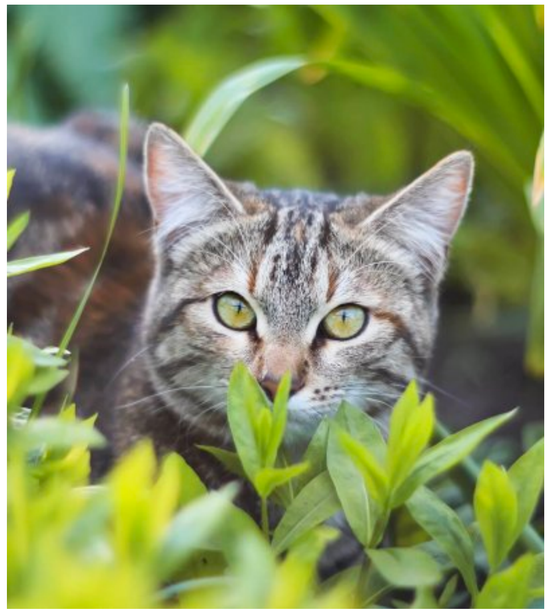
- **Covariate shift**: change in $p(x)$
- **Label shift:** change in $p(y)$
- **Concept shift:** change in $p(y \mid x)$

ROBUSTNESS - COVARIATE SHIFT

- Change in feature distribution $p(x)$
  - Example: predict cats from photographs
  - Cats photographed inside vs. outside is a covariate shift

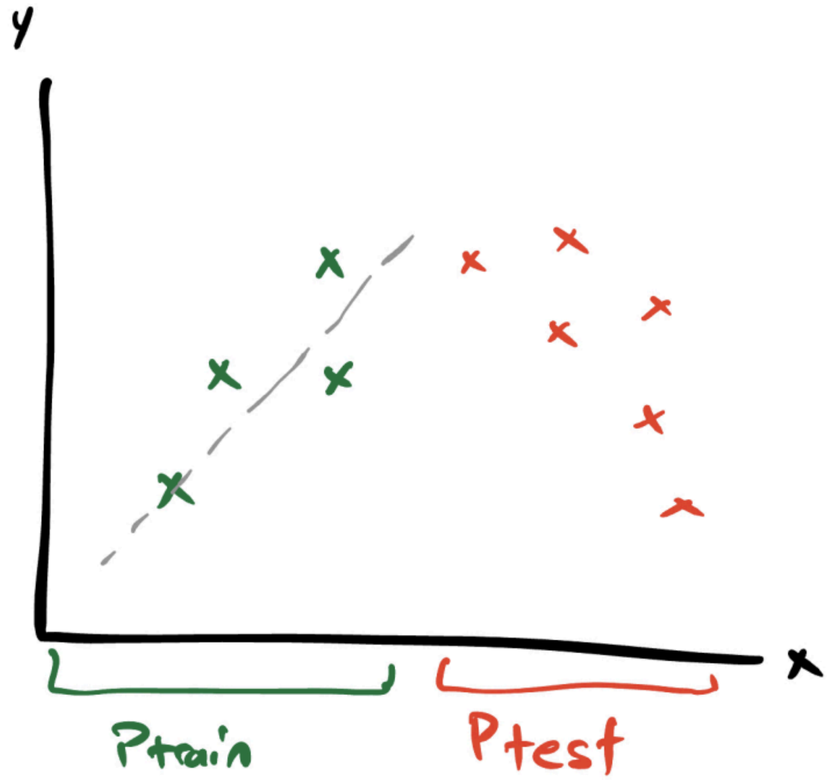
Training data

Testing data



ROBUSTNESS - COVARIATE SHIFT

- Given: $S_{\text{train}} \sim p(y \mid x) p_{\text{train}}(x)$
- Goal: test on $S_{\text{test}} \sim p(y \mid x) p_{\text{test}}(x)$



ROBUSTNESS - LABEL SHIFT

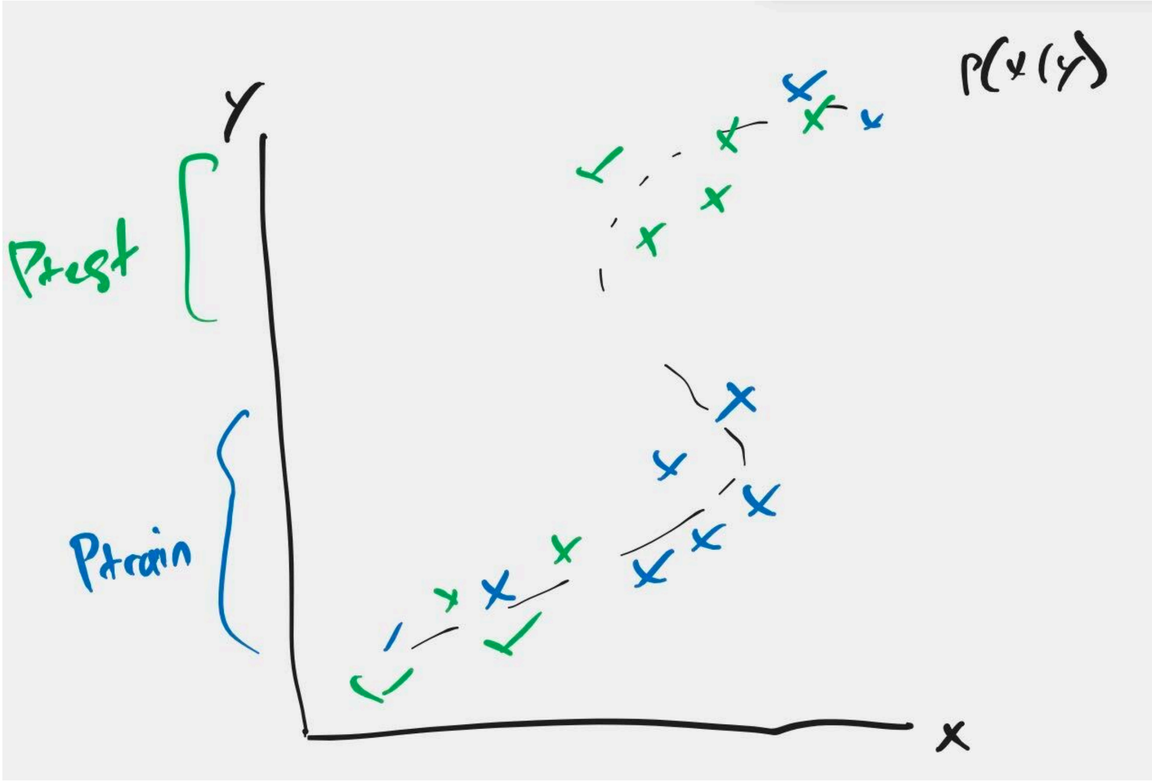
- Change in label distribution $p(y)$
  - Example: predicting bird species from bird characteristics like size, weight, colors, etc.
  - Label distribution of birds shifts from NY to SF (i.e. western bluebirds vs. northern cardinals)

Training data

Testing data

ROBUSTNESS - LABEL SHIFT

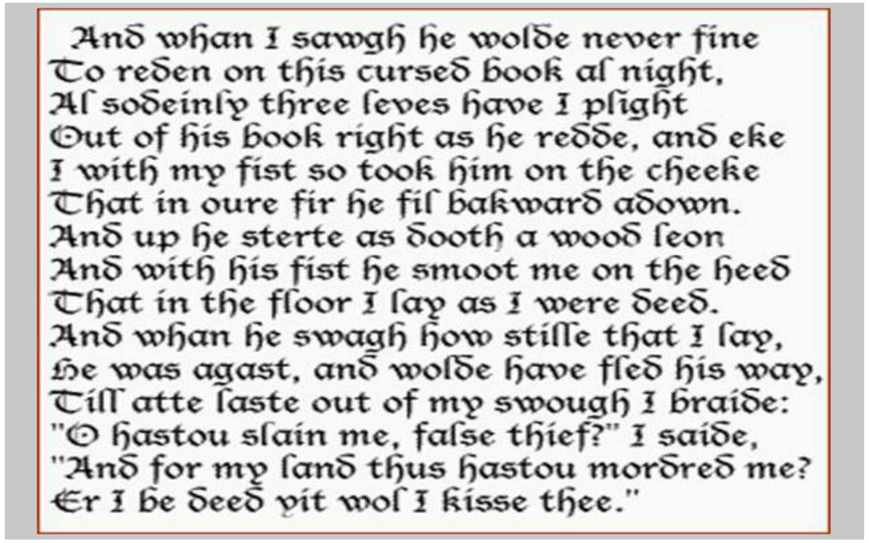
- Given: $S_{\text{train}} \sim p(x \mid y) p_{\text{train}}(y)$
- Goal: test on $S_{\text{test}} \sim p(x \mid y) p_{\text{test}}(y)$



ROBUSTNESS - CONCEPT SHIFT

- Change in distribution $p(y \mid x)$
  - Example: meaning of words in old English vs modern English

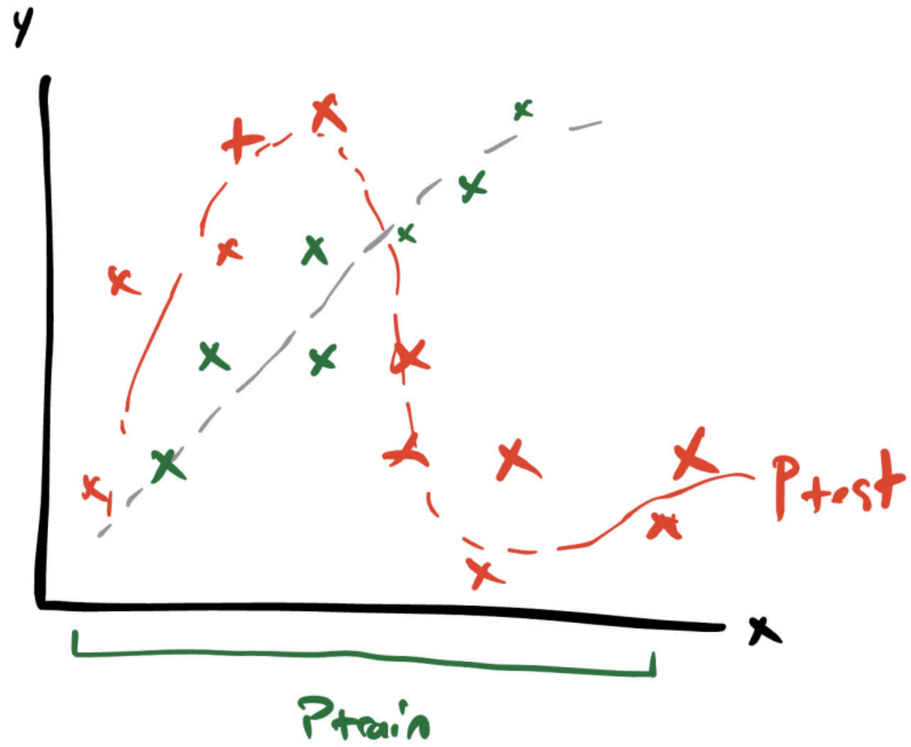
Training data



Testing data

ROBUSTNESS - CONCEPT SHIFT

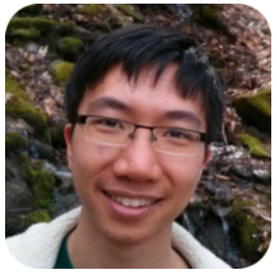
- Given: $S_{\text{train}} \sim p_{\text{train}}(y \mid x)p(x)$
- Goal: test on $S_{\text{test}} \sim p_{\text{test}}(y \mid x)p(x)$



Hope you learned something cool!

INSTRUCTORS

Surbhi Goel



Eric Wong

TEACHING ASSISTANTS

Head TAs

Abhinav Atrishi

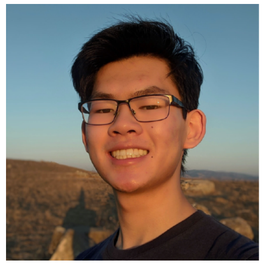
Jordan Hochman

Pavlos Kallinikidis

William Liang

Heyi Liu

Keshav Ramji



David LuoZhang

Aryan Nagariya

Jeffrey Pan

Aditya Singh

Tianyi Wei

Wendi Zhang