Announcements

- HW 6 due **Wednesday, April 19 at 8pm**

- Project Milestone 3 due **Wednesday, April 26 at 8pm**

- Final exam is **Wednesday, May 3 from 6-8pm**
  - Review sessions next week
  - Example final exam and solutions have been released

Lecture 25: Ethics

CIS 4190/5190

Spring 2023

Agenda

- Uncertainty quantification
- Ethics
- Dataset issues
- Fairness/discrimination in datasets
- Defining fairness

Uncertainty Estimates for DNN Predictions

- **Calibrated Prediction (Platt 1999, Guo 2017)**
  - Predict a **probability** $\hat{f}(x)_y$ for each label y
  - What does it mean for the probabilities to be correct?

- **Prediction Sets**
  - Predict a **set** $\hat{f}(x) \subseteq Y$ of possible labels
  - Set is correct if $y^* \in \hat{f}(x)$

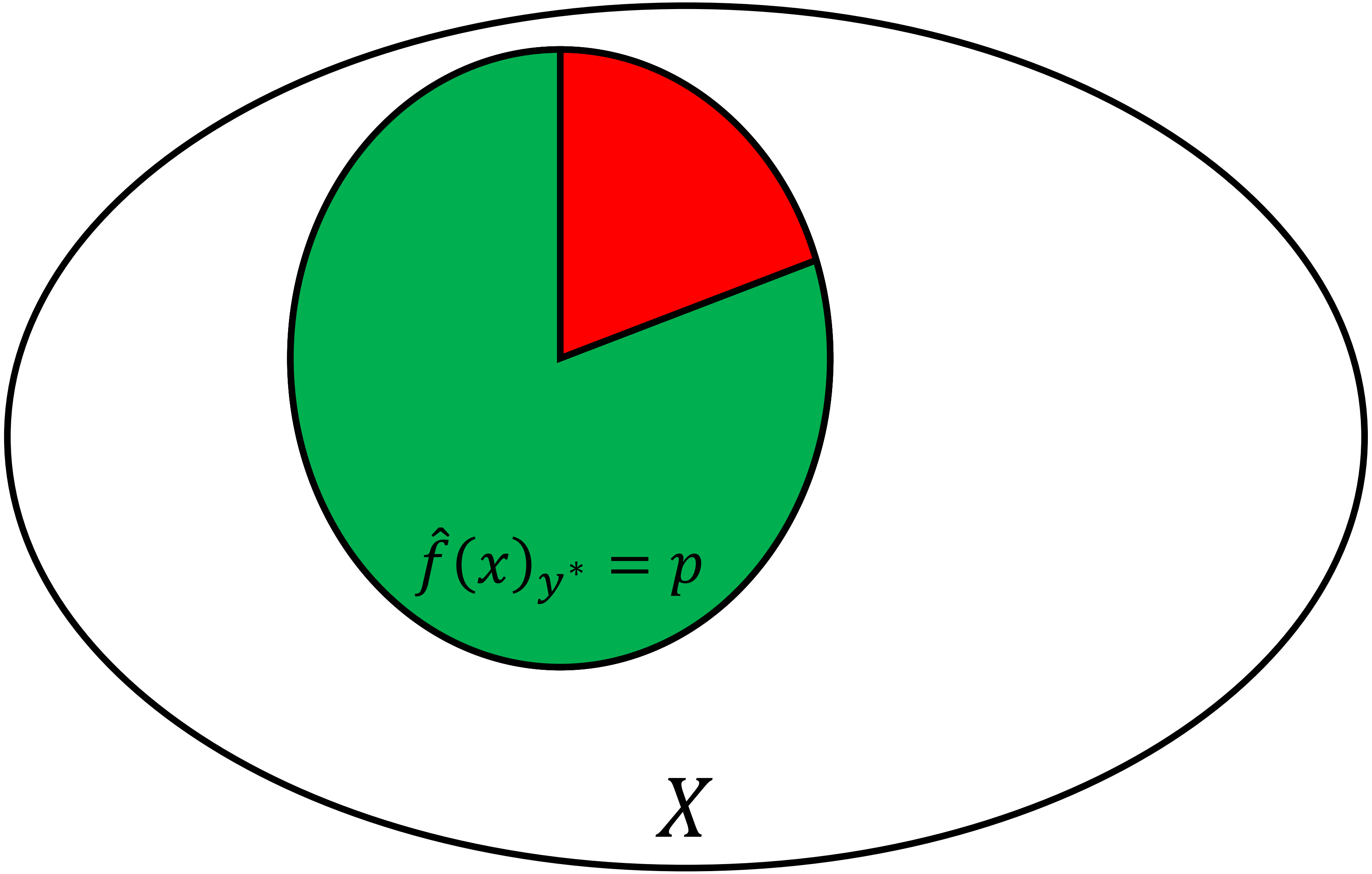
Calibrated Prediction

- Consider a probability predictor $\hat{f}$
  - Let $\hat{y}(x) = \arg\max_{y \in Y} \hat{f}(x)_y$ denote the corresponding labeling function

- We say $\hat{f}$ is **calibrated** if

$$\hat{f}(x) = \Pr_{p(x', y^*)}\left[\hat{y}(x') = y^* \mid \hat{f}(x') = \hat{f}(x)\right]$$

Calibrated Prediction



Calibrated Prediction

- Typical approach in deep learning
  - However, most models have high calibration error

- **Potential explanation**
  - Models need to be **overparameterized** to aid optimization
  - Overparameterization leads to overfitting probabilities (even if accuracy is good!)

Temperature Scaling

- Ordering of probabilities is good!

- Simply rescale them (original is $\tau = 1$):

$$\hat{f}_\tau(x) \propto \exp\left(\frac{\theta^T \phi(x)}{\tau}\right)$$

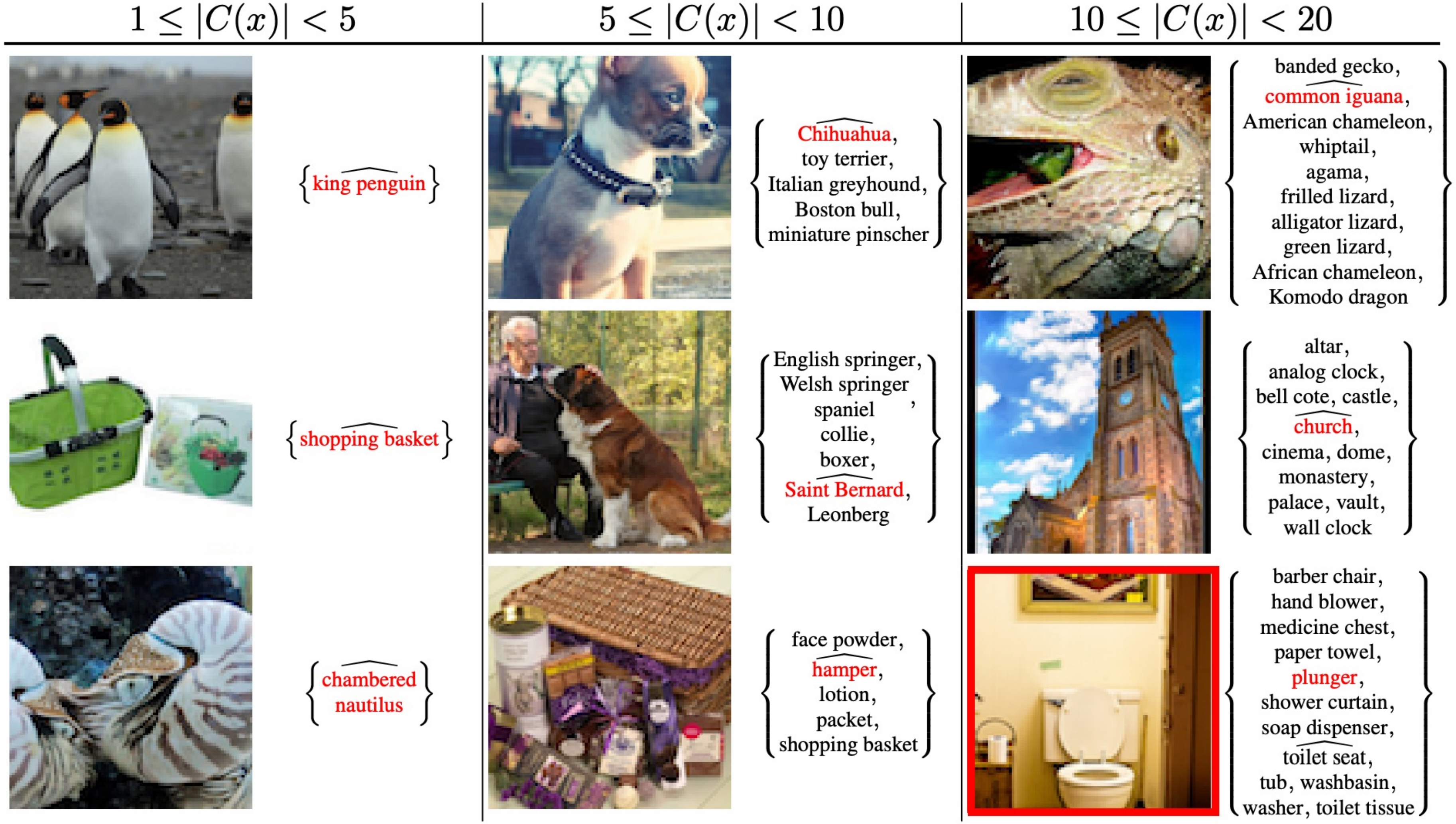
- Choose τ to minimize error on a **held-out calibration set:**

$$\tau^* = \arg\min_\tau \mathbb{E}_{p(x,y^*)}\left[\ell\left(\hat{f}_\tau(x), y^*\right)\right]$$

Aside: Prediction Sets

- Quantify uncertainty by outputting **sets of labels** instead of probabilities
- Subfield known as **conformal inference**
- **Intuition:** Confidence intervals, but around **model predictions** instead of **model parameters**

ImageNet Classification



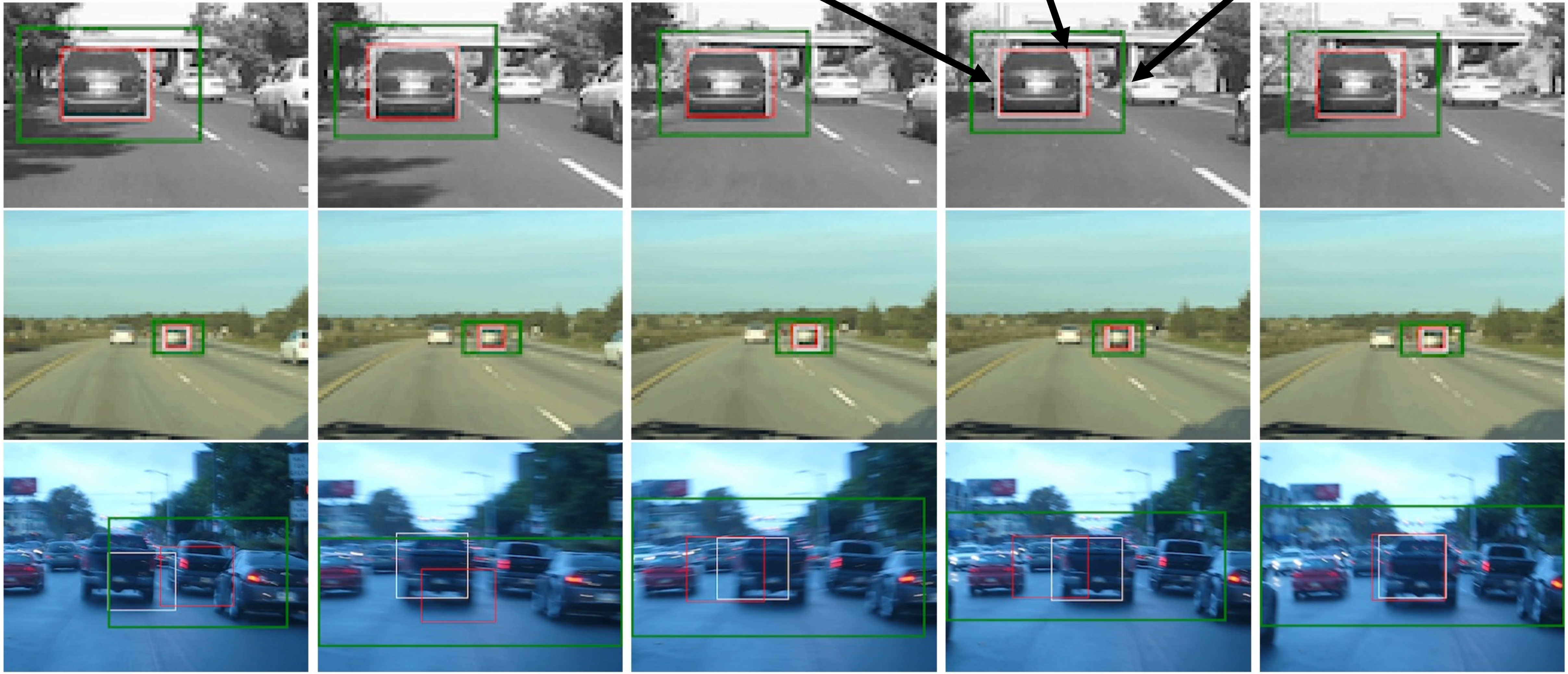
$1 \leq \lvert C(x) \rvert < 5$	$5 \leq \lvert C(x) \rvert < 10$	$10 \leq \lvert C(x) \rvert < 20$
{$\widehat{\text{king penguin}}$}	{$\widehat{\text{Chihuahua}}$, toy terrier, Italian greyhound, Boston bull, miniature pinscher}	{banded gecko, $\widehat{\text{common iguana}}$, American chameleon, whiptail, agama, frilled lizard, alligator lizard, green lizard, African chameleon, Komodo dragon}
{$\widehat{\text{shopping basket}}$}	{English springer, Welsh springer spaniel, collie, boxer, $\widehat{\text{Saint Bernard}}$, Leonberg}	{altar, analog clock, bell cote, castle, $\widehat{\text{church}}$, cinema, dome, monastery, palace, vault, wall clock}
{$\widehat{\text{chambered nautilus}}$}	{face powder, $\widehat{\text{hamper}}$, lotion, packet, shopping basket}	{barber chair, hand blower, medicine chest, paper towel, plunger, shower curtain, soap dispenser, $\widehat{\text{toilet seat}}$, tub, washbasin, washer, toilet tissue}

Object Tracking

ground truth

predicted

confidence set



Agenda

- Uncertainty quantification
- Ethics
- Dataset issues
- Fairness/discrimination in datasets
- Defining fairness

Ethics is Hard!

- **Ethical decision-making**
  - Challenging problem even without ML
  - Thousands of years of debate in philosophy, law, etc.
  - Changes over time with changing societal norms

- **Challenges with machine learning**
  - Data privacy issues
  - Internalize (and even amplifies) biases already present in data
  - New issues related to abuse of ML

ML Applications

- **Fairness/discrimination issues**
  - Policing/judicial decisions, financial decisions, etc.
  - Filtering resumes of job applicants
  - Global aid allocation based on satellite images
  - Echo chamber issues in news/video recommendations

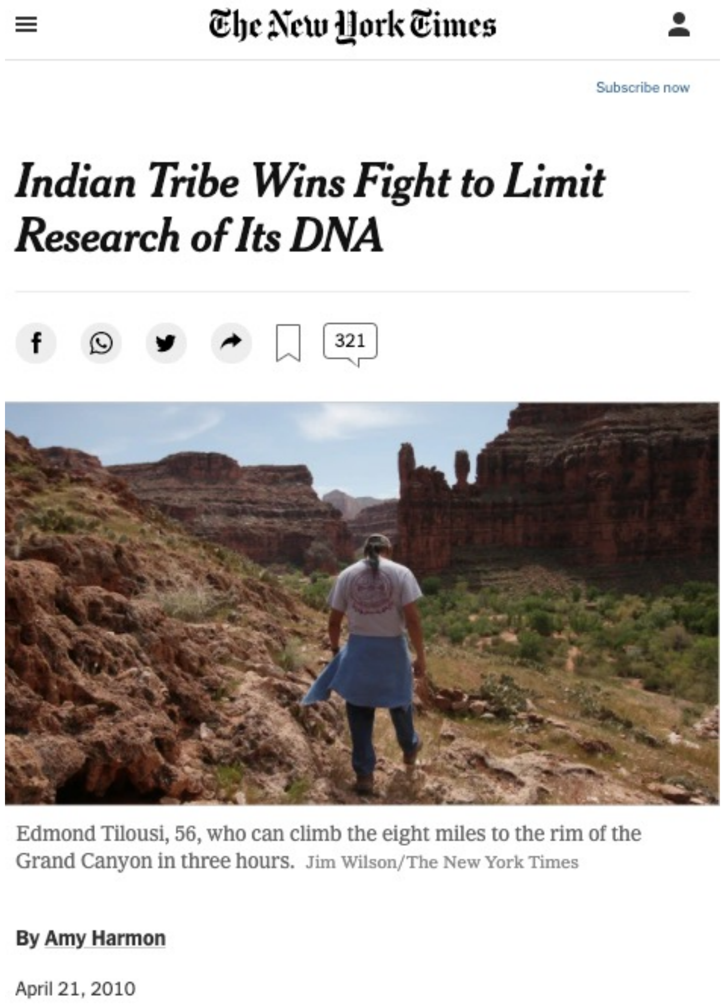
- **Potentially problematic applications**
  - Dangers in safety-critical settings
  - Automating wide-scale surveillance based on facial recognition
  - Autonomous drones for military uses
  - Refugees turned away at the US border because an ML system assessed risk of terrorist activity based on Instagram posts

Agenda

- Uncertainty quantification
- Ethics
- Dataset issues
- Fairness/discrimination in datasets
- Defining fairness

Data Privacy Issues

- **Pima People Diabetes Dataset**
  - “Members of the tiny, isolated tribe had given DNA samples to university researchers starting in 1990, in the hope that they might provide genetic clues to the tribe’s devastating rate of diabetes. But they learned that their blood samples had been used to study many other things, including mental illness and theories of the tribe’s geographical origins that contradict their traditional stories.”
  - **Data collection requires informed consent**
- Public data ≠ consent for research use



The New York Times

Subscribe now

Indian Tribe Wins Fight to Limit Research of Its DNA

Edmond Tilousi, 56, who can climb the eight miles to the rim of the Grand Canyon in three hours. Jim Wilson/The New York Times

By Amy Harmon

April 21, 2010

Differential Privacy

- **Question:** How to define “privacy”?

- **Intuition:** Individual’s data minimally affects algorithm output
  - **Differential privacy:** Magnitude is **probabilistic**, i.e., probability of each output is very close with and without individual’s data
  - **Note:** Differentially private algorithm must have a randomized output!

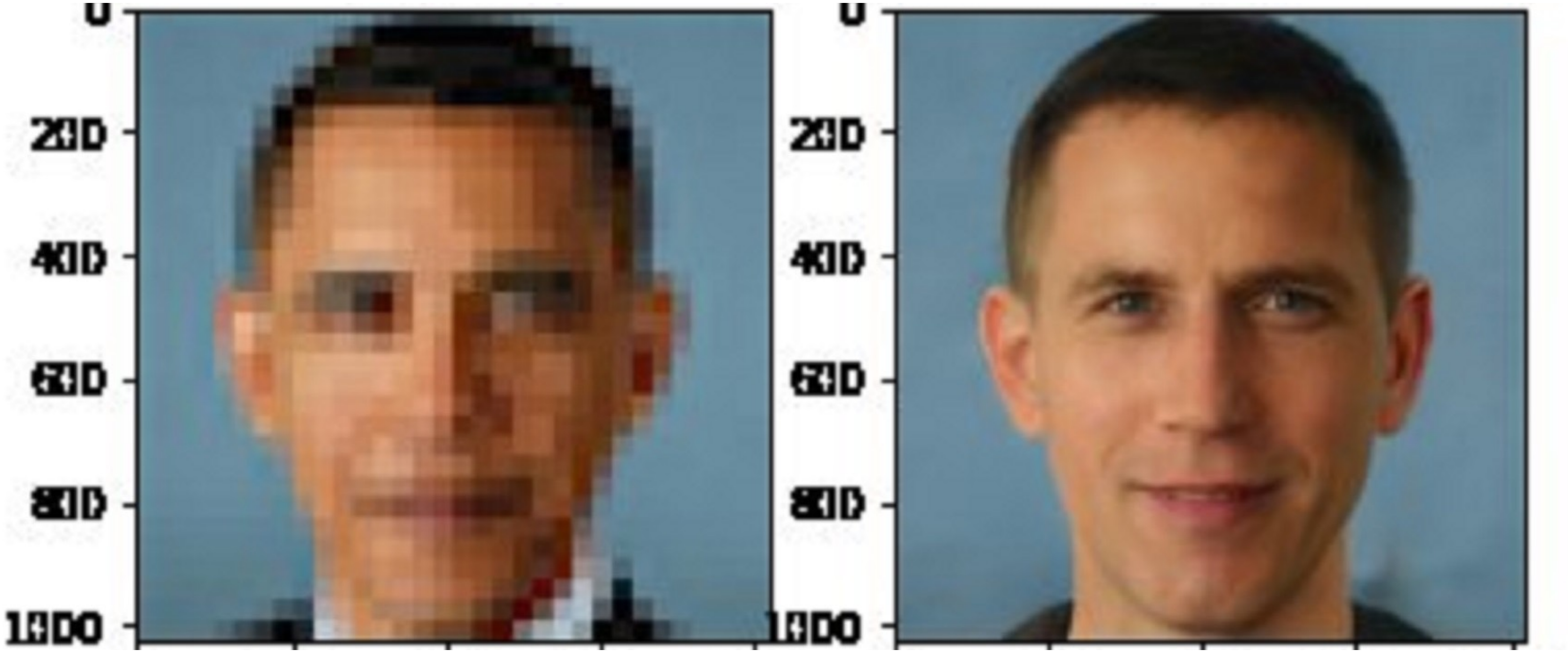
- **Differentially private deep learning**
  - **Original solution:** Learn parameters, and add noise to parameters
  - **Better solution:** Add noise to gradient of each individual’s data
  - Requires more data to learn, but maintains privacy of individuals

Agenda

- Uncertainty quantification
- Ethics
- Dataset issues
- Fairness/discrimination in datasets
- Defining fairness

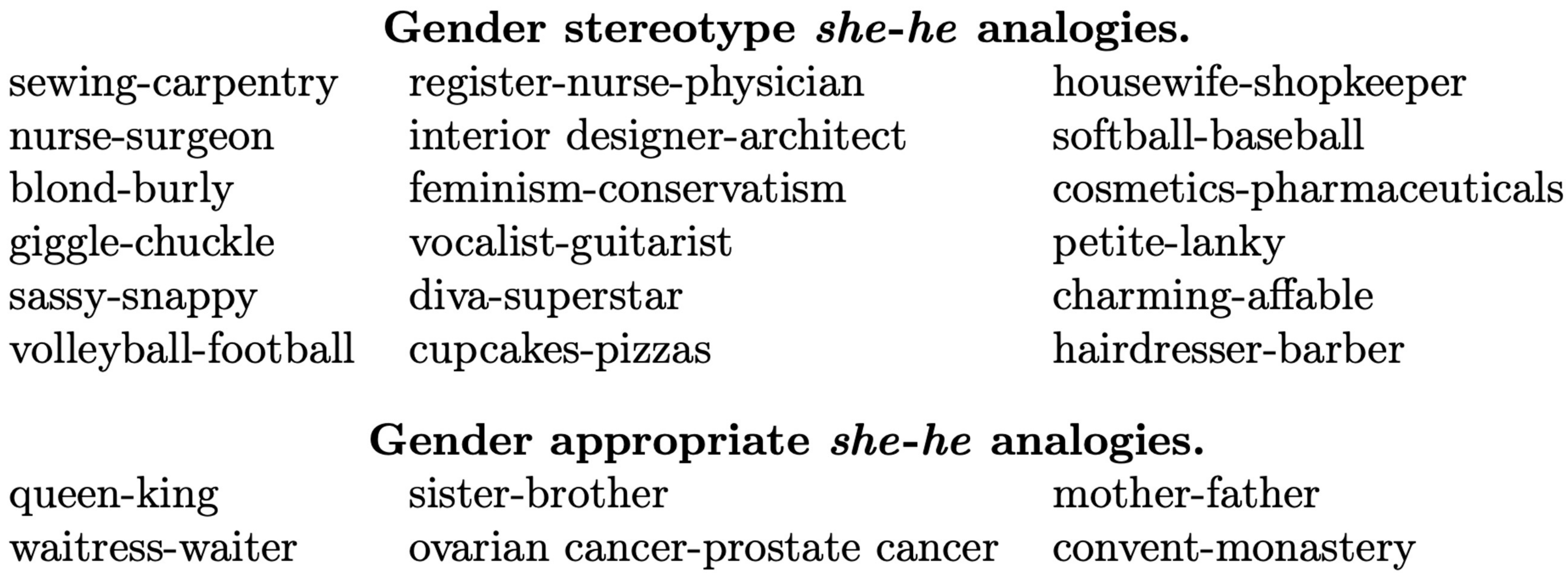
Discrimination in ML

- ML models may be biased against minorities



Discrimination in ML

- ML models may be biased against minorities



Gender stereotype *she*-*he* analogies.

sewing-carpentry
nurse-surgeon
blond-burly
giggle-chuckle
sassy-snappy
volleyball-football

register-nurse-physician
interior designer-architect
feminism-conservatism
vocalist-guitarist
diva-superstar
cupcakes-pizzas

housewife-shopkeeper
softball-baseball
cosmetics-pharmaceuticals
petite-lanky
charming-affable
hairdresser-barber

Gender appropriate *she*-*he* analogies.

queen-king
waitress-waiter

sister-brother
ovarian cancer-prostate cancer

mother-father
convent-monastery

Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Sources of Bias

- **Data representation:** Distribution of inputs $p(x)$
- **Tainted labels:** Distribution of label assignments $p(y \mid x)$
- **Sensitive features:** Selecting what features to include for each sample (e.g., whether to include sensitive attributes such as race and gender)

Data Representation

- Less data from minority groups → Higher error on minority groups
- **Example:** Many clinical trials historically recruited largely white males, leading to biases in understanding outcomes and side effects
- **Example:** Focus on easily accessible data (e.g. recent tweets, or easily measured features of people) can lead to biased datasets
- Need to be careful to gather representative datasets

Tainted Labels

- **Example:** Amazon hiring bias
  - Amazon's ML resume screening tool to predict hiring decisions based on 10 years of historical applicant data; but found it was biased against women
  - Labels tainted by historical bias
  - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

- **Similar example**
  - Company filters hires by predicting how long they will stay at the company
  - But how long someone stays depends on how they were treated

Tainted Labels

- **Example:** Predictive policing
    - “PredPol” predictive policing system employed in some policy departments
    - Suppose that crime happens equally everywhere
    - Some areas more policed → More crime found in those areas
    - ML learns to predict crime in neighborhoods that were more policed

Tainted Labels

- Need to be careful that labels are unbiased

- However, can be very hard to unbias data!
  - “We should strive to avoid giving women lower salaries”
  - **ML model:** “women” = “lower salaries”

Sensitive Attributes as Features

- When should sensitive attributes be used as features?

- **Example:** Predicting diabetes risk
  - Race is a sensitive attribute that may not cause diabetes, but may be correlated with unrecorded features that cause diabetes
  - What if an insurance company decides that people of some races are at higher risk and should pay higher premium?

- Omitting sensitive attributes is not enough!
  - Other features such as current income may be correlated with race/gender

Data Collection Issues

- Need to gather representative sample
- Need to ensure labels are unbiased
- Need to think carefully about whether to include sensitive attributes

Datasheets for Datasets (Gebru et al.)

- Questions for dataset creators to think through and answer for users:
  - Motivation
  - Dataset Composition
  - Collection Process
  - Preprocessing
  - Uses
  - Distribution
  - Maintenance

- https://arxiv.org/abs/1803.09010

Agenda

- Uncertainty quantification
- Ethics
- Dataset issues
- Fairness/discrimination in datasets
- Defining fairness

Fairness and ML

- What does it mean to be fair?

Case Study: Criminal Justice

- Software by Northpointe to predict **recidivism** for defendants
  - I.e., risk of committing future crimes

- Used to help make bail, sentencing, and parole decisions

Case Study: Criminal Justice

- **Features:** 137 questions answered by defendants or criminal records:
  - “Was one of your parents ever sent to jail or prison?”
  - “How many of your friends/acquaintances are taking drugs illegally?”
  - “How often did you get in fights while at school?”
  - Agree or disagree? “A hungry person has a right to steal”
  - Agree or disagree? “If people make me angry or lose my temper, I can be dangerous.”

- Exact algorithm and model is a trade secret

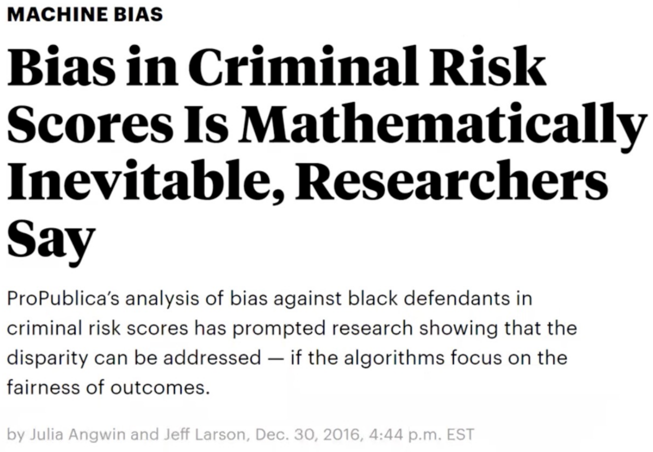
Case Study: Criminal Justice

- Race is **not** a feature

- **Problem:** Correlated features
  - One of the developers of the system said it is difficult to construct a score that doesn’t include items that can be correlated with race
  - E.g., poverty, joblessness and social marginalization
  - “If those are omitted from your risk assessment, accuracy goes down”

- Similar to Amazon hiring bias example

Case Study: Criminal Justice

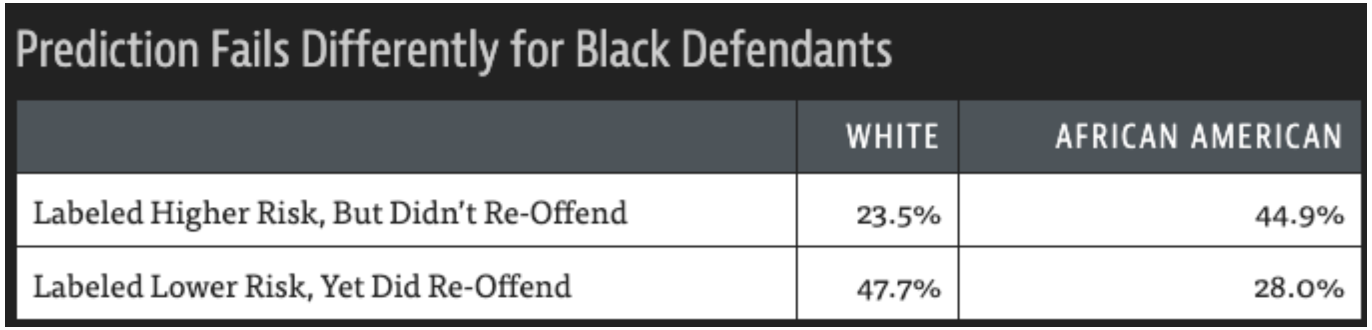


MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST



Prediction Fails Differently for Black Defendants

	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%

Defining Fairness

- **Legally Protected Attributes**
  - Race, sex, color, religion, national origin (Civil Rights Act of 1964, Equal Pay Act of 1963)
  - Age (Discrimination in Employment Act of 1967)
  - Citizenship (Immigration Reform and Control Act)
  - Pregnancy (Pregnancy Discrimination Act)
  - Familial status (Civil Rights Act of 1968)
  - Disability (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)

Defining Fairness

- **Potential definition:** Two individuals differing on sensitive attributes but otherwise identical should receive the same outcome

- **Issue:** What does it mean for two people to be “otherwise identical”?
  - What if just their accents differ?
  - What if just their attire differs?

- Also ignores historical discrimination encoded in features, which is even harder to address

Defining Fairness

- **Accuracy and fairness**
  - Low accuracy can result in unfairness
  - E.g., strong student scored as highly as weak one for college admissions
  - But highest accuracy model is not necessarily the most fair

- **Group fairness:** Account for performance on subgroups

$$\text{Fairness metric} = F\big(L(f; X_1), \ldots, L(f; X_k)\big)$$

Group Fairness

- **Problem setup**
  - Sensitive attribute A
  - ML model R mapping input features X to prediction $\hat{Y} = R(X)$
  - True outcome Y (typically binary, and $Y = 1$ is the “good” outcome)

- **Example:** Insurance risk prediction
  - $A = \text{age}$
  - $R = \text{predicted cost}$
  - $Y = \text{true cost}$

Group Fairness

- **Independence:** Risk score distribution should be equal across ages:

$$P(\,\text{risk score} \mid \text{age}\,) \;=\; P(\text{risk score})$$

  - E.g., equal proportion of low risk customers for young vs. old people
  - Often called demographic parity

- What if lower age groups in fact behave more riskily?

Group Fairness

- **Separation:** Risk score should be independent of age given outcome:

$$P(\text{ risk score} \mid \text{age, true outcome }) \ = \ P(\text{risk score} \mid \text{true outcome})$$

  - Equivalent to saying the true positive rate and false positive rate are equal across subgroups

- **Example:** Both of the following hold:
  - Fraction of young, low-insurance-usage people correctly identified as low-risk $=$ Fraction of old low-insurance-usage people correctly identified as low-risk
  - Fraction of young high-insurance-usage people wrongly identified as low-risk $=$ Fraction of old high-insurance-usage people wrongly identified as low-risk

Group Fairness

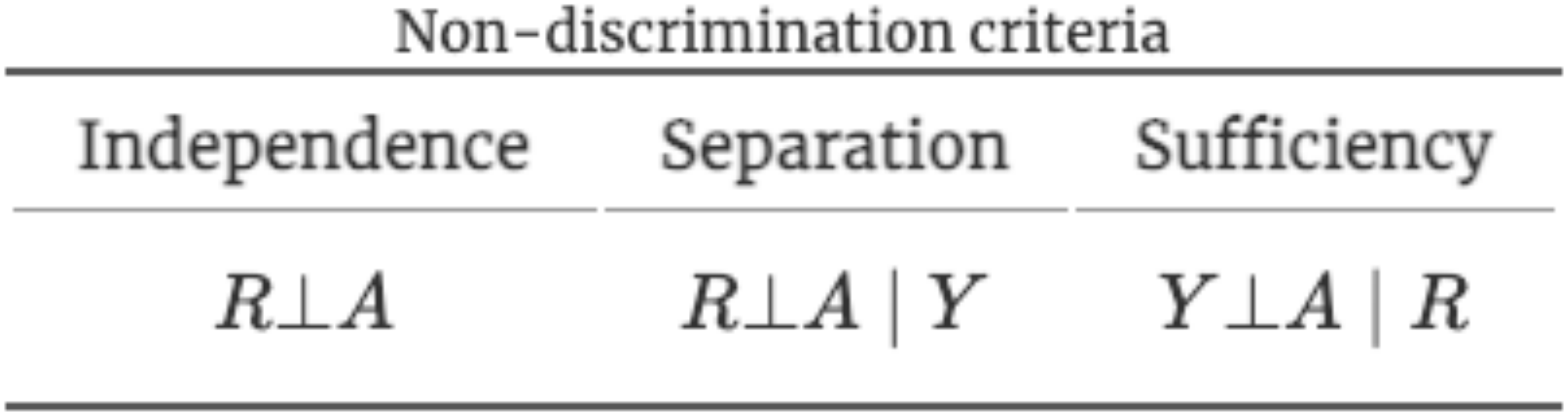
- **Sufficiency:** Outcome should be independent of age given risk score:

$$P(\text{ true outcome } \mid \text{age}, \text{risk score}) \; = \; P(\text{true outcome} \mid \text{risk score})$$

  - Intuitively, risk score tells us everything we need to know about the true outcome with respect to age

- Closely related to **calibration**
  - Can rescale risk score so model is calibrated
  - Calibration holds across subgroups

Group Fairness

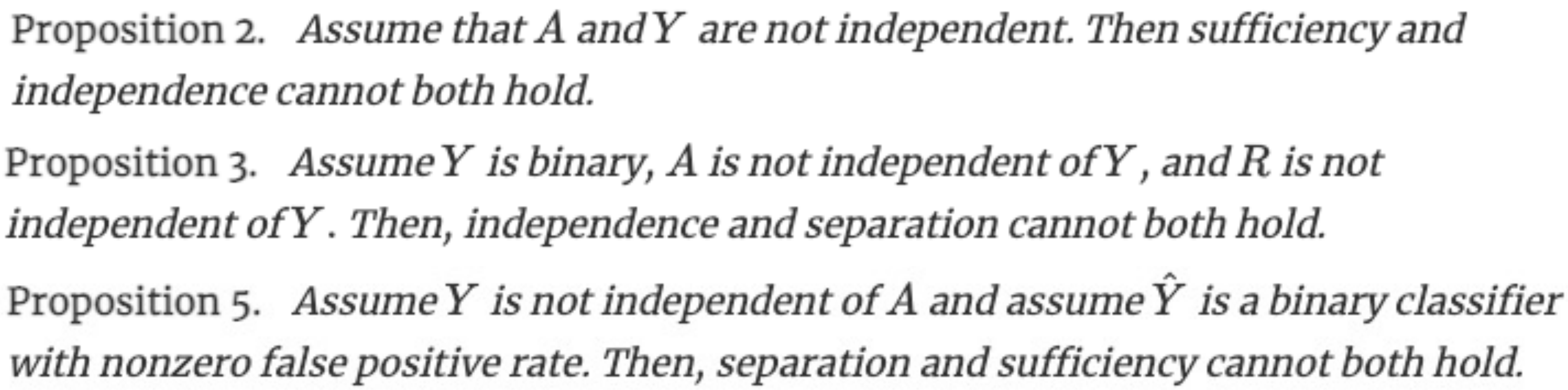


Non-discrimination criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Group Fairness

- Three notions are incompatible!



Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

- Thus, need carefully choose what kinds of fairness we ask for

Solon Barocas, Moritz Hardt, Arvind Narayanan, “Fairness and Machine Learning”

Algorithms for Ensuring Fairness

- Given a notion of fairness, there are a few ways of achieving it

- **Example:** Independence
  - **Pre-processing:** Adjust features to be uncorrelated with sensitive attribute
  - **Training constraints:** Impose the constraint during training
  - **Post-processing:** Adjust the learned classifier so its predictions are uncorrelated with the sensitive attribute

- **Goodhart's law:** "When a measure becomes a target, it ceases to be a good measure" – Marilyn Strathern
  - Do not blindly impose fairness, need to carefully examine predictions

Human-in-the-Loop Fairness

- **Potential solution:** Have domain experts weigh in on what performance metrics result in fair model selection/training

- **Challenges**
  - Experts may not understand limitations of ML models (e.g., does a judge using a system understand that it only has 60% accuracy?)
  - Potential for selective enforcement based on human biases

Human-in-the-Loop Fairness

- **Example:** In bail decision-making, judges selectively follow model
  - Less lenient against younger defendants, especially minorities
  - Younger defendants are actually more risky, but judges may have been lenient due to societal norms (e.g., “second chance”)
  - Judges followed algorithm less and less over time

https://www.washingtonpost.com/business/2019/11/19/algorithms-were-supposed-make-virginia-judges-more-fair-what-actually-happened-was-far-more-complicated/