

Announcements

- HW 4 due today
- Project Milestone due next Wednesday (April 23)

Lecture 24: Robustness

CIS 4190/5190

Spring 2025

Agenda

- **Interpretability & Explainability**
- **Robustness to distribution shift**
- **Robustness to adversarial attacks**

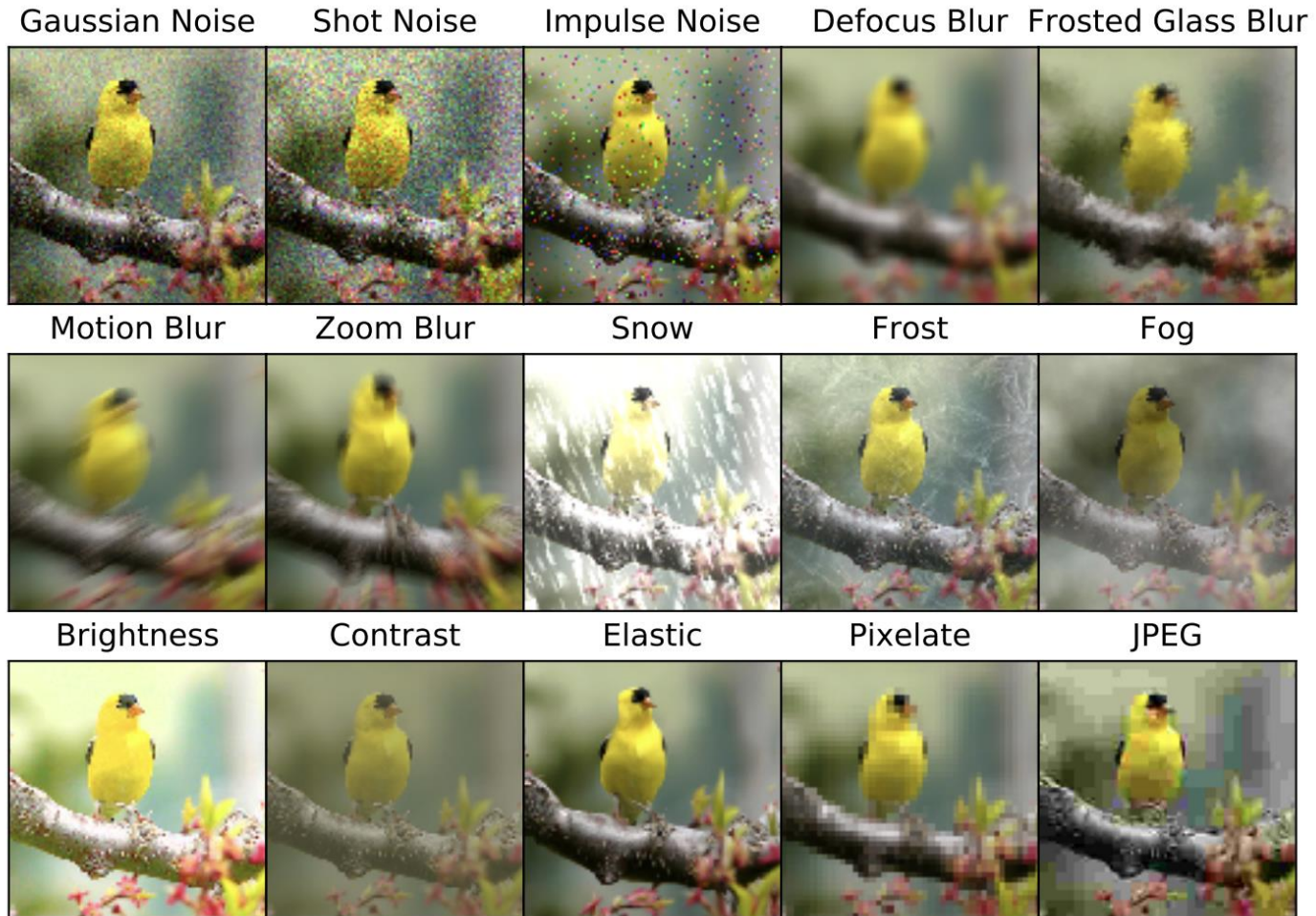
Robustness to Distribution Shift

- Neural networks generalize well **on distribution**
- **Ideal scenario**
 - Test set and training set are i.i.d. from the same distribution
 - **Equivalently:** Test set is obtained by shuffling entire dataset and then splitting
- **Often fails in practice! “Distribution shift”**

Robustness to Distribution Shift

- **Images/computer vision**
 - Added noise, color shifts, lighting changes, different resolution, etc.
- **Audio/speech-to-text**
 - Noisy background, changes in recording device, etc.
- **Natural language processing**
 - Substitute synonyms, add unrelated text, etc.

Example: Synthetic Perturbations



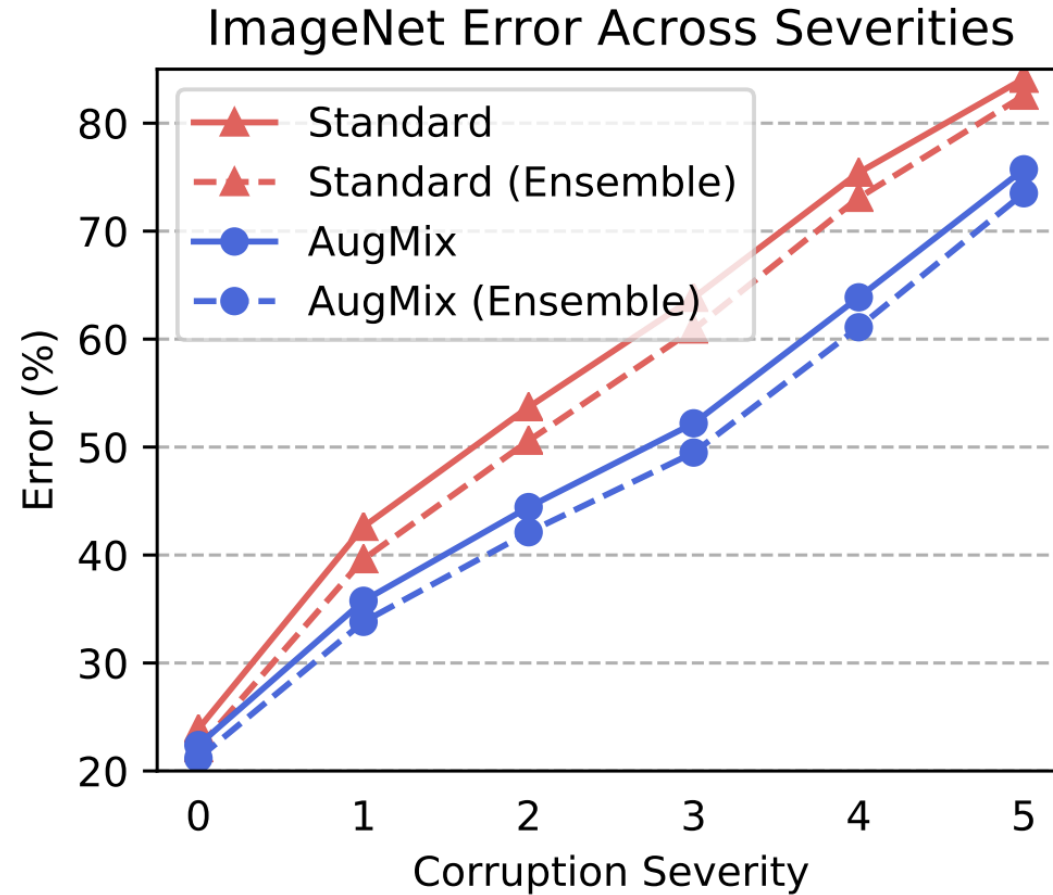
Example: Synthetic Perturbations

- **Question:** Why should the model be robust?
- **Answer:** Humans are robust!

Example: Synthetic Perturbations

- **Significantly reduces performance**
 - 20% error rate → 80% error rate
- **Data augmentation can help (but not 100% solution)**

Example: Synthetic Perturbations



Example: Natural Language Processing

Article: Super Bowl 50












Paragraph: *“Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”*

Question: *“What is the name of the quarterback who was 38 in Super Bowl XXXIII?”*






Original Prediction: John Elway

Prediction under adversary: Jeff Dean

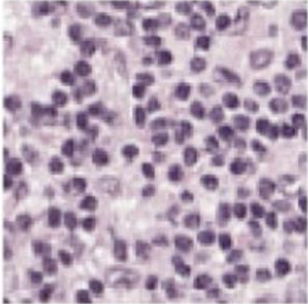
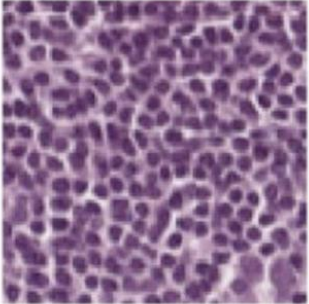
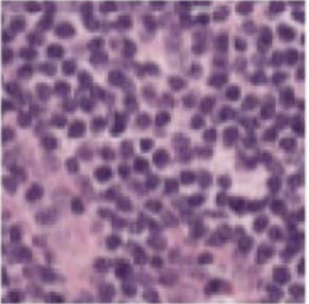
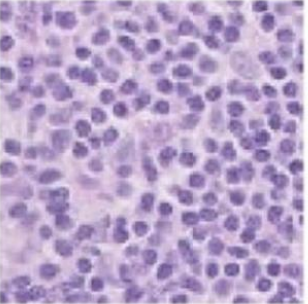
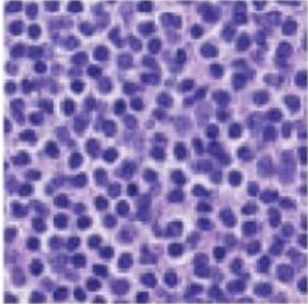
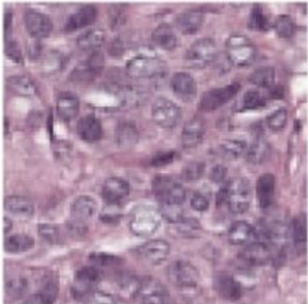
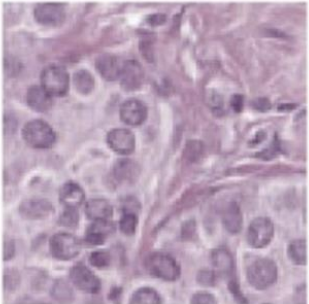
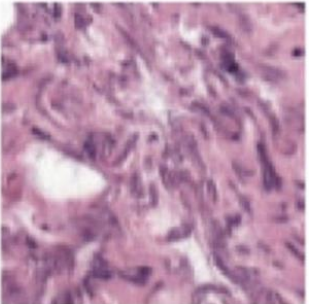
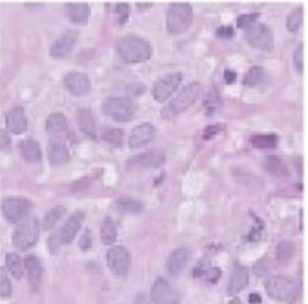
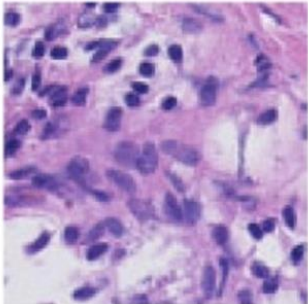
Example: Real Perturbations

Train			Test (OOD)
$d = \text{Location 1}$	$d = \text{Location 2}$	$d = \text{Location 245}$	$d = \text{Location 246}$
			
Vulturine Guineafowl	African Bush Elephant	...	Wild Horse
			
Cow	Cow	Southern Pig-Tailed Macaque	Great Curassow
Test (ID)			
$d = \text{Location 1}$	$d = \text{Location 2}$	$d = \text{Location 245}$	
			
Giraffe	Impala	Sun Bear	

Example: Real Perturbations

	Train			Test	
Satellite Image (x)					
Year / Region (d)	2002 / Americas	2009 / Africa	2012 / Europe	2016 / Americas	2017 / Africa
Building / Land Type (y)	shopping mall	multi-unit residential	road bridge	recreational facility	educational institution

Example: Real Perturbations

	Train			Val (OOD)	Test (OOD)
	d = Hospital 1	d = Hospital 2	d = Hospital 3	d = Hospital 4	d = Hospital 5
y = Normal					
y = Tumor					

Potential Solutions

- **No general strategy (yet)**
- **Good practices**
 - Train on as large & diverse of a dataset as possible
 - Use data augmentation when possible
 - If available, finetune on location-specific dataset (transfer learning)

Agenda

- **Interpretability & Explainability**
- **Robustness to distribution shift**
- **Robustness to adversarial attacks**

Robustness to Adversarial Attacks

- **Example:**

- Want to reject email attachment if it contains malicious code
- Use machine learning to predict if code is malicious

- **What can go wrong?**

- Attacker perturbs code (e.g., add random lines of dead code) until it is labeled benign by the machine learning model!
- **Strong form of robustness is needed**

Example: Function Name Prediction

- **Task:** Given a function (e.g., as a string), predict its name
- **Attack:** Add a random line of irrelevant code

Example: Function Name Prediction

```
void f1(int[] array){
  boolean swapped = true;
  for (int i = 0;
      i < array.length && swapped; i++){
    swapped = false;
    for (int j = 0;
        j < array.length-1-i; j++) {
      if (array[j] > array[j+1]) {
        int temp = array[j];
        array[j] = array[j+1];
        array[j+1]= temp;
        swapped = true;
      }
    }
  }
}
```

Prediction: **sort** (98.54%)

```
void f3(int[] array){
  boolean swapped = true;
  for (int i = 0;
      i < array.length && swapped; i++){
    swapped = false;
    for (int j = 0;
        j < array.length-1-i; j++) {
      if (array[j] > array[j+1]) {
        int temp = array[j];
        array[j] = array[j+1];
        array[j+1]= temp;
        swapped = true;
      }
    }
  } int upperhexdigits;
}
```

Prediction: **escape** (100%)

Robustness to Adversarial Perturbations

- **Task:**

- Photo ID verification
- Goal is to check whether uploaded photo matches a photo ID

- **Attack:**

- User perturbs their image to match the photo in the ID
- Challenge for machine learning in online identity verification!



(Valid photo ID from Papesh 2018)

Robustness to Adversarial Perturbations

- **Robustness:** Similar images \Rightarrow same label
- **Goal:** Robust to **any** small perturbation in **some family**
 - **Note:** Very far from solving this problem
- **Key question:** What is “some family”?

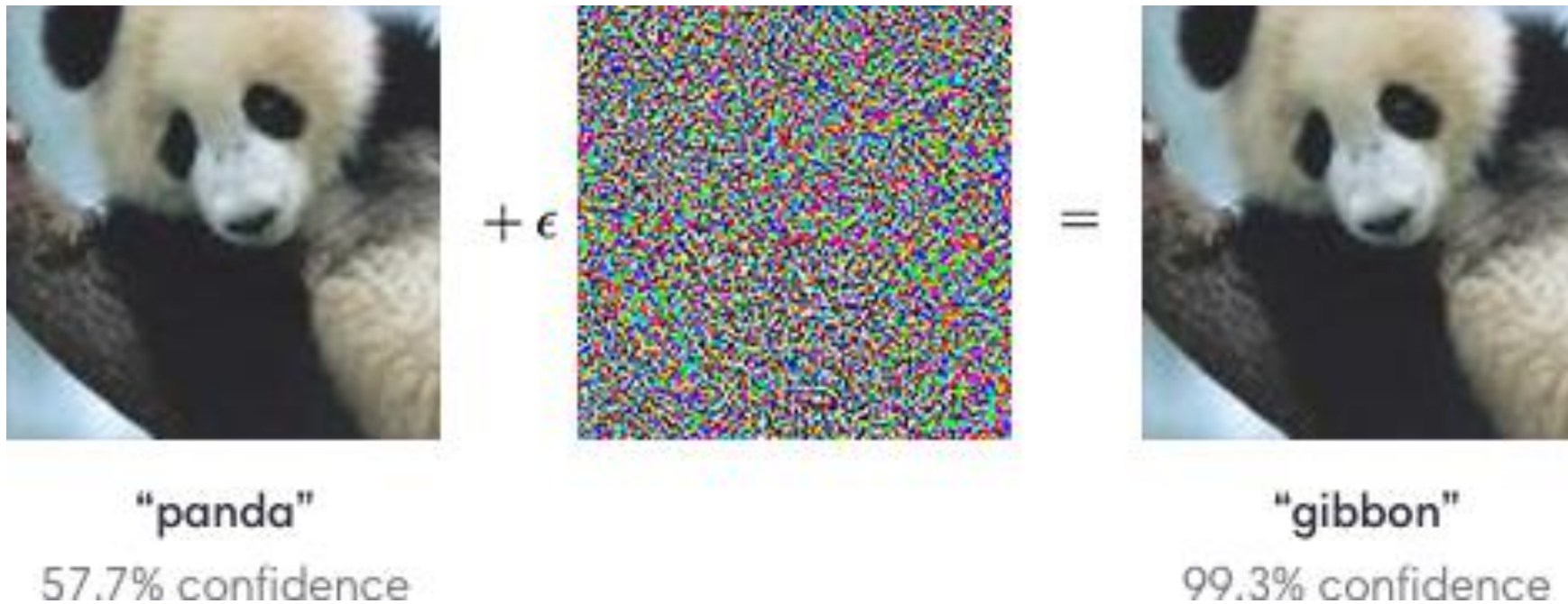
Robustness to Adversarial Perturbations

- (Very limited) example for images:

$$\|x - x'\|_{\infty} \leq \epsilon \Rightarrow \text{same label}$$

- **Question:** Why should the model be robust to these perturbations?
 - Should not change the label
 - Humans are robust!

Robustness to Adversarial Perturbations



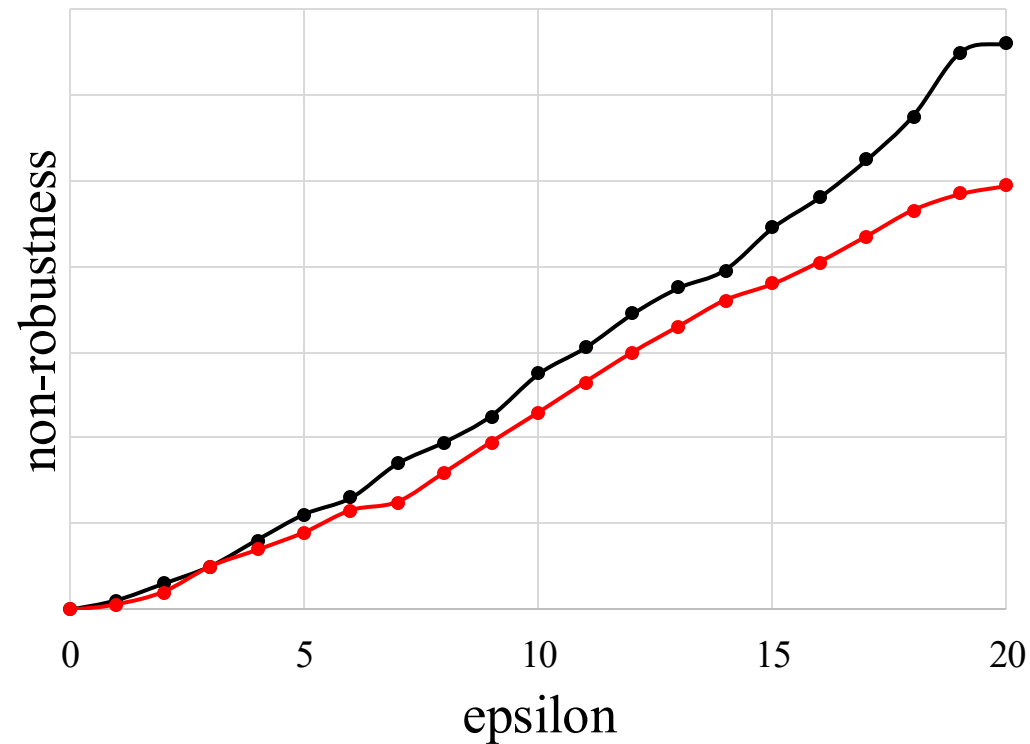
Szegedy et al., Intriguing Properties of Neural Networks, 2014

Robustness to Adversarial Perturbations

- **Strategy for improving adversarial robustness**
 - Data augmentation!
 - **Adversarial training:** Use adversary to generate new examples for training
- Does it work?

Improving Robustness?

Adversarial Robustness



■ Original NN ■ Robust NN

Improving Robustness?

- **Problem**

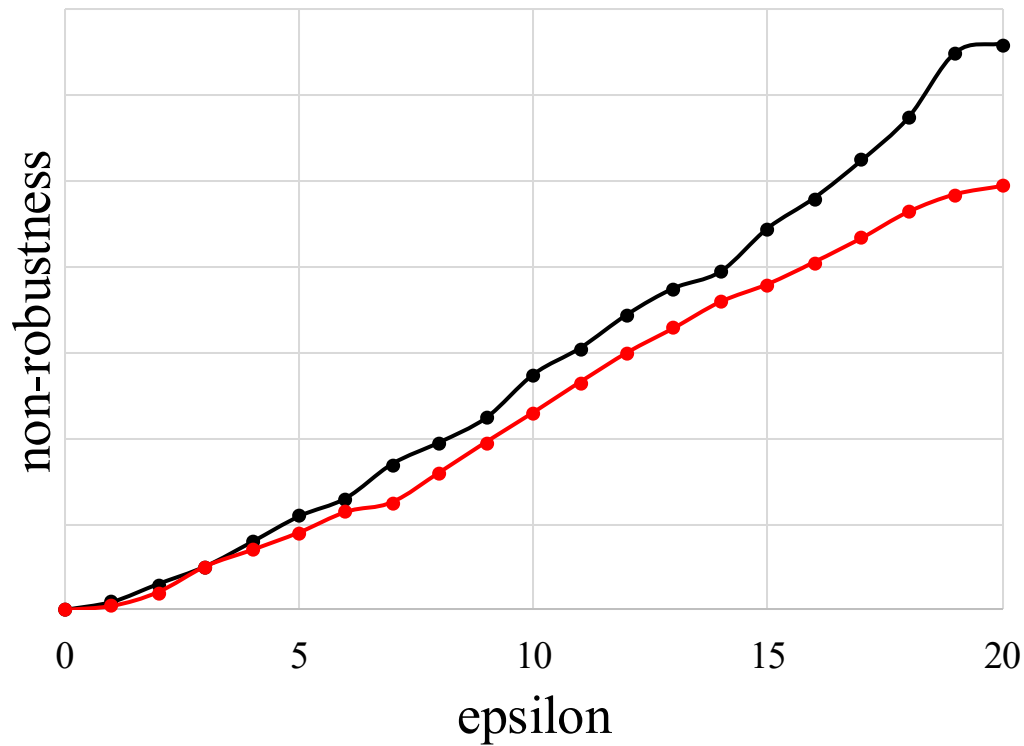
- Only robust to the current adversary
- What if the adversary changes? **Distribution shift!**

- **Example**

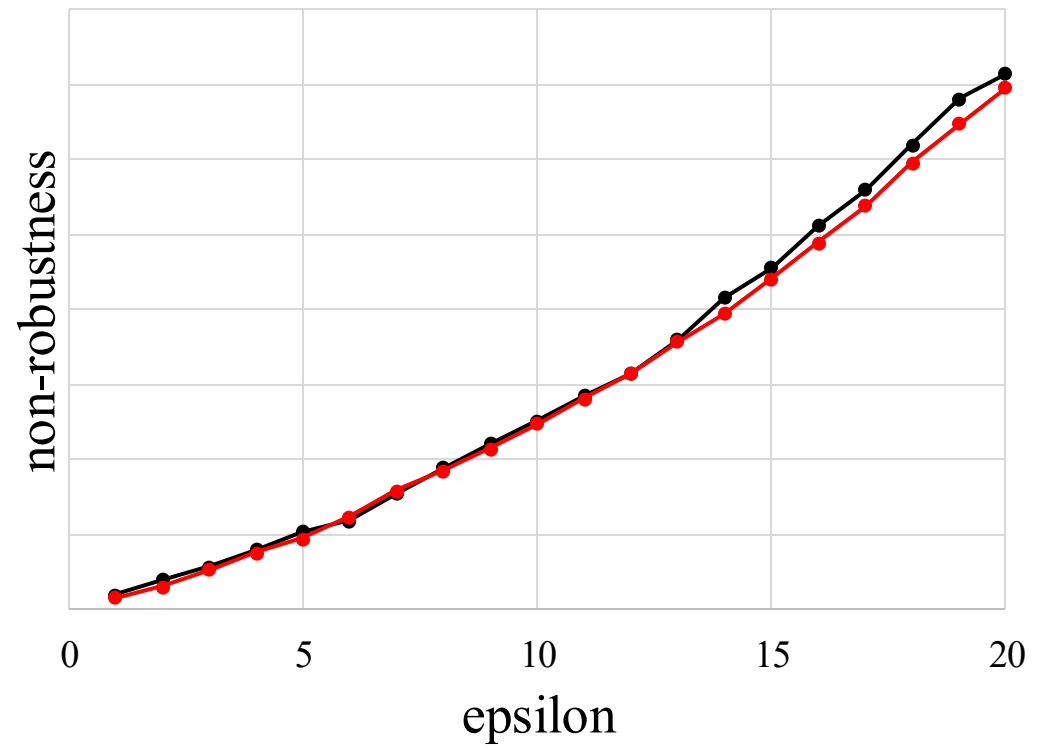
- Adversarial training using one adversary
- Test against a more powerful adversary

Improving Robustness?

Algorithm's Own Metric



Our Metric



■ Original NN ■ Robust NN

Potential Solutions

- **No general strategy (yet)**
- **Good practices**
 - Use the strongest adversary you can design
 - Use variety of different adversaries

Can Uncertainty Help?

- **Recall:** Most neural networks predict an uncertainty

$$p_{\beta}(y | x)$$

- **Idea:** Can we use uncertainty to detect adversarial attacks?
- **Answer: No!**
 - Adversarial examples can have very high confidence
 - Probabilities can be overconfident even for normal test examples!

Potential Solutions

- **General solutions for non-adversarial setting:** Calibrated prediction
- **Intuition:** Among examples where neural network predicts it is correct with probability p , it is correct for a fraction $\approx p$
- **Algorithms:** Temperature scaling, isotonic regression, etc.

Potential Solutions

- **No general solutions for adversarial setting**
- **Good practices**
 - Don't blindly trust predicted probabilities!

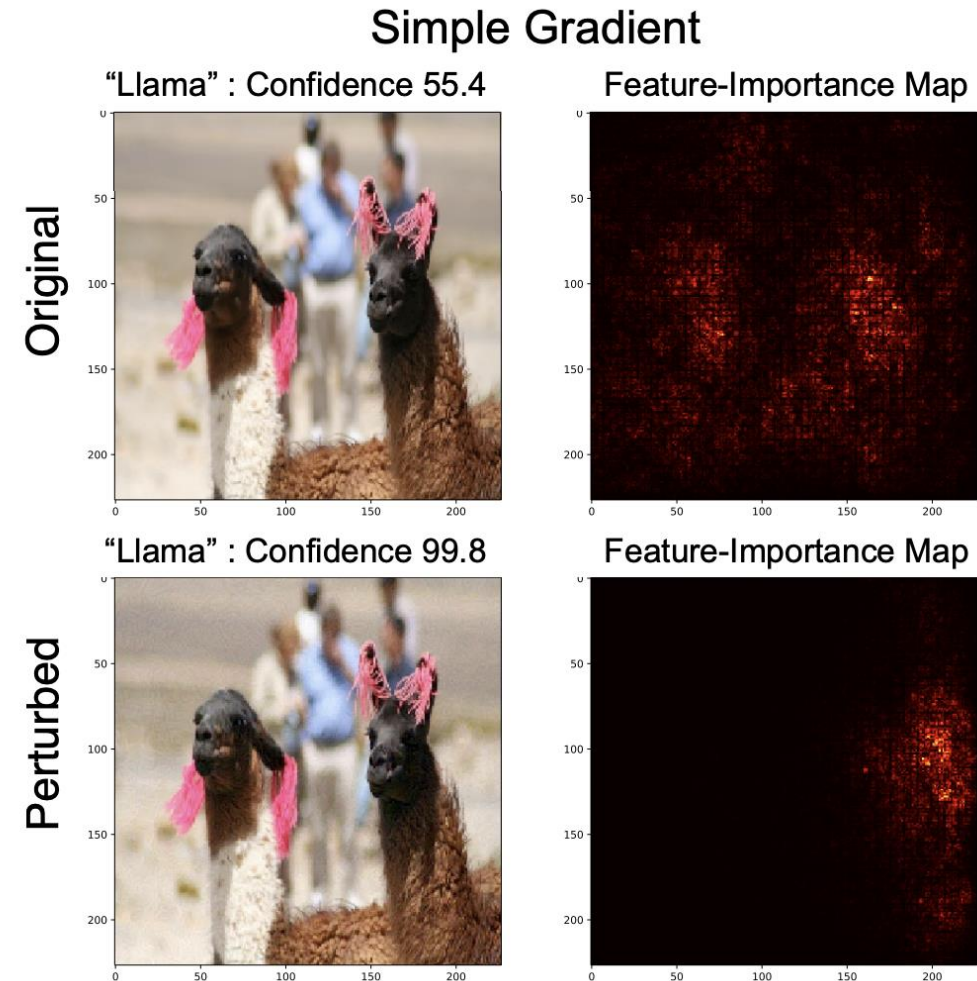
Can Explanations Help?

- **Idea:** Check if explanation makes sense
- **Question:** Are explanations of neural networks robust?

$$\text{Explain}(x + \epsilon) \approx \text{Explain}(x)$$

- **Answer: No!**
- **Not even robust to distribution shift**

Fragility of Explanations



Fragility of Explanations

- Not just a problem for neural networks!

If **Race** \neq African American:

If **Prior-Felony** = Yes and **Crime-Status** = Active, then **Risky**

If **Prior-Convictions** = 0, then **Not Risky**

If **Race** = African American:

If **Pays-rent** = No and **Gender** = Male, then **Risky**

If **Lives-with-Partner** = No and **College** = No, then **Risky**

If **Age** \geq 35 and **Has-Kids** = Yes, then **Not Risky**

If **Wages** \geq 70K, then **Not Risky**

Default: **Not Risky**

If **Current-Offense** = Felony:

If **Prior-FTA** = Yes and **Prior-Arrests** \geq 1, then **Risky**

If **Crime-Status** = Active and **Owns-House** = No and **Has-Kids** = No, then **Risky**

If **Prior-Convictions** = 0 and **College** = Yes and **Owns-House** = Yes, then **Not Risky**

If **Current-Offense** = Misdemeanor and **Prior-Arrests** $>$ 1:

If **Prior-Jail-Incarcerations** = Yes, then **Risky**

If **Has-Kids** = Yes and **Married** = Yes and **Owns-House** = Yes, then **Not Risky**

If **Lives-with-Partner** = Yes and **College** = Yes and **Pays-Rent** = Yes, then **Not Risky**

If **Current-Offense** = Misdemeanor and **Prior-Arrests** \leq 1:

If **Has-Kids** = No and **Owns-House** = No and **Prior-Jail-Incarcerations** = Yes, then **Risky**

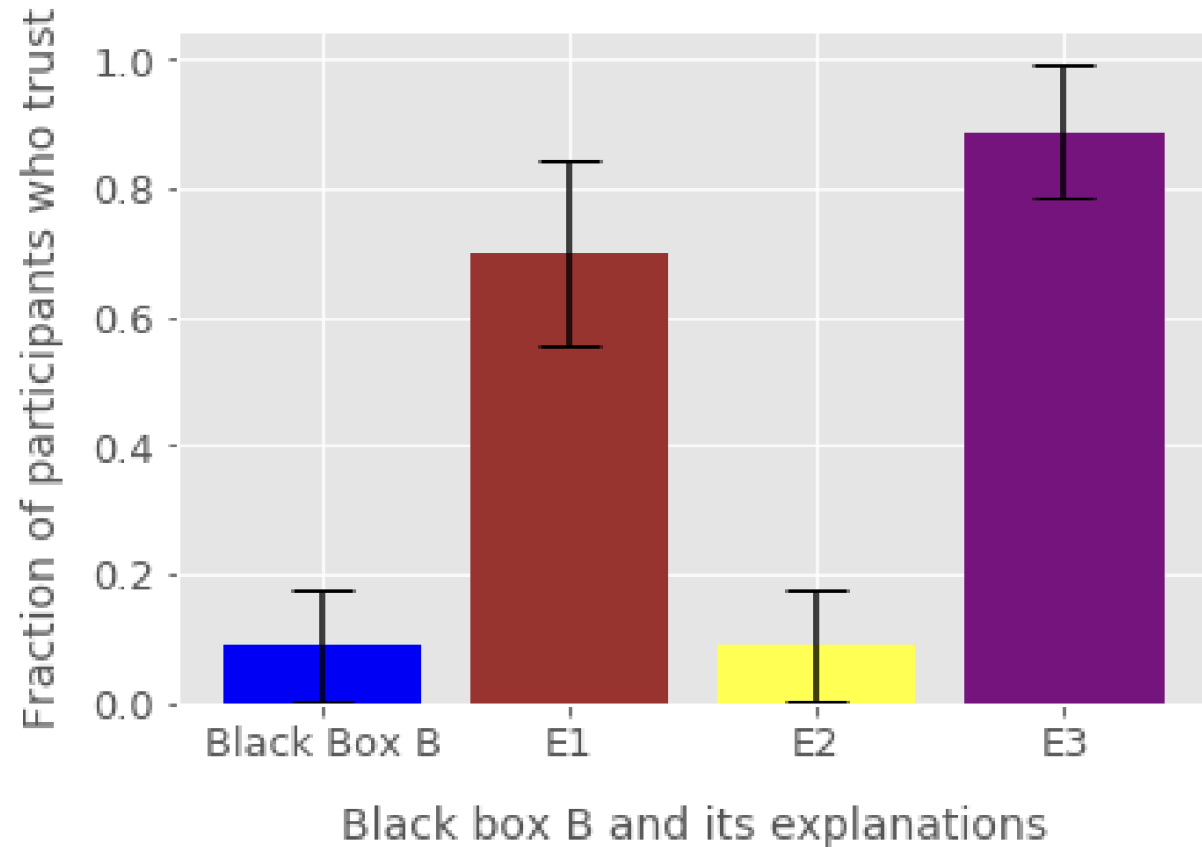
If **Age** \geq 50 and **Has-Kids** = Yes and **Prior-FTA** = No, then **Not Risky**

Default: **Not Risky**

Misleading Explanations

- Can construct explanations to mislead users into trusting a model
- **Strategy**
 - Design a set of features that users believe are trustworthy
 - Generate an explanation that highlights these features as important
- Users believe the model is using trustworthy features even if it is not

Misleading Explanations



E1 & E3 are misleading explanations

Potential Solutions

- **No general strategy (yet)**
- **Good practices**
 - Be careful when interpreting explanations!

Conclusion

- Robustness and interpretability remain key challenges for neural networks (and machine learning more broadly)
- **Good practices**
 - Use variety of techniques to try and understand what models are doing (interpretation, extensive testing on different examples, etc.)
 - Be careful when training models!
 - **Monitor performance of models running in production**
- Lots of ongoing research!

Lecture 25: Ethics

CIS 4190/5190

Spring 2025

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

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

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**

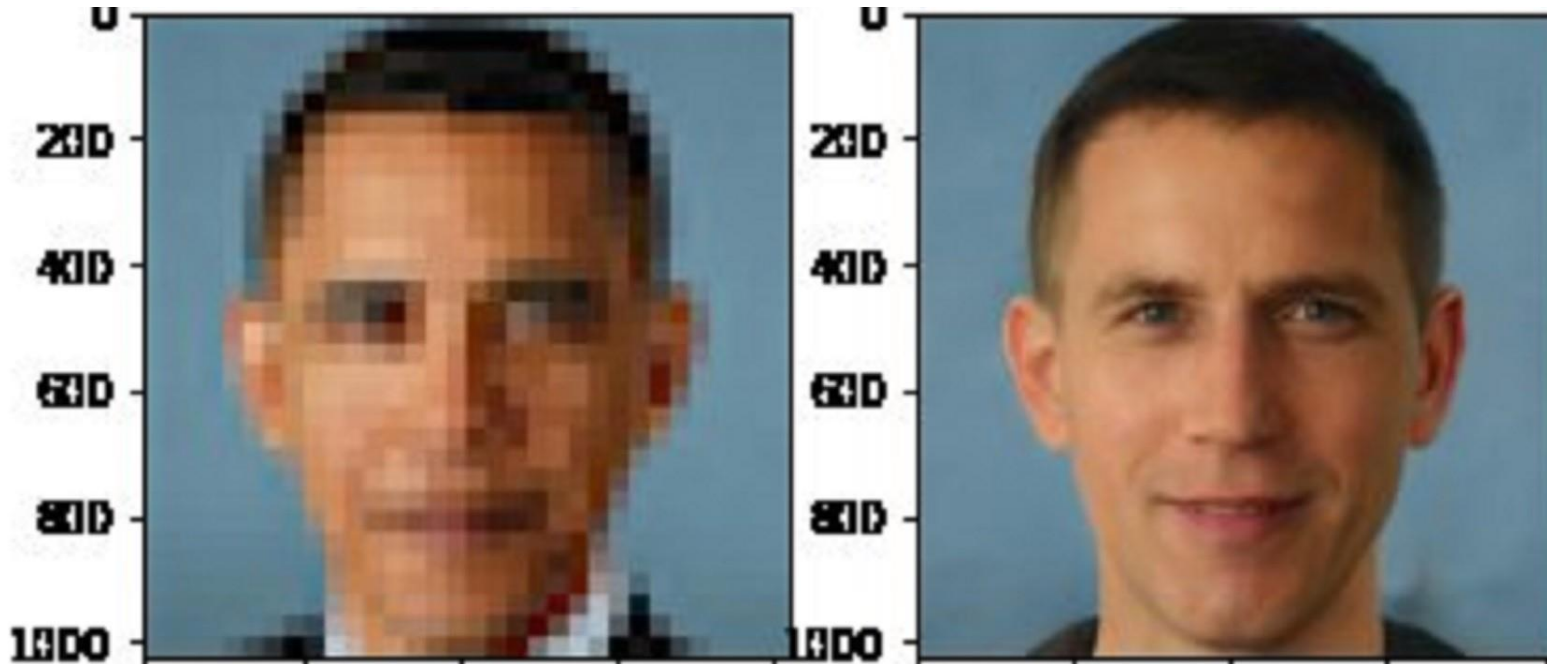
- Public data \neq consent for research use



The image is a screenshot of a news article from The New York Times. At the top, the newspaper's name "The New York Times" is displayed in a serif font, with a hamburger menu icon to the left and a user profile icon to the right. Below the name, there is a "Subscribe now" link. The main headline of the article is "Indian Tribe Wins Fight to Limit Research of Its DNA" in a bold, italicized serif font. Below the headline, there are social media sharing icons for Facebook, WhatsApp, Twitter, a share icon, a bookmark icon, and a comment icon with the number "321". The article's main image shows a man, Edmond Tilousi, standing on a dirt path in a rugged, rocky landscape, looking towards the camera. Below the image, the text reads: "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". At the bottom of the article, it says "By Amy Harmon" and "April 21, 2010".

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	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairstylist-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Sources of Bias

- **Data representation:** Distribution of inputs $p(x)$
- **Tainted labels:** Distribution of label assignments $p(y | x)$
- **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-scrap-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 police 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 unbiased 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

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

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

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

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 risk score given age:

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

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

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

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/>

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Misinformation about ML

6.1 The public predicts a 54% likelihood of high-level machine intelligence within 10 years

Respondents were asked to forecast when high-level machine intelligence will be developed. High-level machine intelligence was defined as the following:

We have high-level machine intelligence when machines are able to **perform almost all tasks that are economically relevant today better than the median human (today) at each task**. These tasks include asking subtle common-sense questions such as those that travel agents would ask. For the following questions, you should ignore tasks that are legally or culturally restricted to humans, such as serving on a jury.¹³

Respondents were asked to predict the probability that high-level machine intelligence will be built in 10, 20, and 50 years.

Comparison: Experts predicts in the ~50-year (may be optimistic)

Example: Self-Driving Without LIDAR



Example: Resume Evaluation

How to persuade a robot that you should get the job

Do mere human beings stand a chance against software that claims to reveal what a real-life face-to-face chat can't?

Stephen Buranyi

Sat 3 Mar 2018 19:05 EST



James Ball ✓
@jamesrbuk

Vision: algorithms will make hiring better as they don't discriminate

Reality: "One HR employee for a major technology company recommends slipping the words "Oxford" or "Cambridge" into a CV in invisible white text, to pass the automated screening."

7:16 AM · Mar 4, 2018 · [Twitter for iPhone](#)

2.2K Retweets 3.5K Likes

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Feedback Loops in ML Systems

- ML models are often part of a larger system
- **Example:** Feedback loop in PredPol (used to predict crime)
 - This kind of approach is “especially nefarious” because police can say: “We’re not being biased, we’re just doing what the math tells us.” And the public perception might be that the algorithms are impartial. – Samuel Sinyangw

To predict and serve?

Kristian Lum, William Isaac

Rise of the racist robots - how AI is learning all our worst impulses

Feedback Loops in ML Systems

- **Recommender systems:** “A system for predicting the click through rate of news headlines on a website likely relies on user clicks as training labels, which in turn depend on previous predictions”
- **Potential for adversarial feedback**
 - Tricking a resume screening system by entering keywords like “Oxford”
 - **Anecdotal:** Computer vision systems to predict poverty and (semi-) automate global aid allocation decisions lead to people switching off their night lights and dressing up concrete roofs as thatched roofs

Satellite images used to predict poverty

By Paul Rincon

Science editor, BBC News website

Machine Learning: The High Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)

Extreme Example: “Future Features”

- **Scenario**

- Build a highly complex classifier with 99% accuracy for a time-series problem
- Later, build a new classifier with 98.5% accuracy, runs 1000 × faster
- Catastrophic failure when deployed!

- **Problem**

- Training data included classifier’s prediction from previous step as input
- **New classifier:** “Recycles” the prediction from the previous step (i.e., just use that single feature as the prediction!)
- Works fine when previous prediction was already accurate
- No longer the case after deployment!

Potential Solution

- **DAGGER algorithm**
 - Originally designed for imitation learning (i.e., RL from expert data)
 - Continuously collect new labels and add to training set
- $Z \leftarrow$ Initial dataset
- For $t \in \{1, 2, \dots\}$:
 - Train f_β on D and use to make decisions on new examples X_t
 - Observe (or collect) ground truth labels Y_t for X_t
 - $Z \leftarrow Z \cup \{(X_t, Y_t)\}$
- Use multi-armed bandits when there is partial feedback

More Challenging Feedback Loops

- **Example:** Hiring ads
 - Women tend to click on job ad with second-highest salary
 - ML model learns that women do not click on highest salary job ad, so it stops recommending it
 - Second-highest salary job ad → Highest salary job ad
 - Women click on new second-highest salary job ad!
- No substitute for manual analysis of ML models in projection
 - You'll never be out of a job (at least for the foreseeable future)!

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Ethical Issues

- When you build ML models, you are responsible for how it is eventually deployed
 - Face classifier may be used by an authoritarian government to track people or target minority subgroups
 - Technology may be used in safety critical settings without sufficient validation

Best Practices for Ethical ML

- Human augmentation
- Bias evaluation
- Explainability and justification
- Displacement strategy

Human Augmentation

- Assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes
- Especially important in domains with significant impact on human lives (e.g. justice, health, etc.)
 - All stakeholders' values and perspectives should be accounted for during algorithm design
 - Domain experts as human-in-the-loop reviewers of ML decisions

Bias Evaluation

- **Use tools to understand bias in ML models**
 - No standard strategy, need to carefully consider potential sources of bias for the domain you are working in
 - Requires continuous monitoring, not one-time effort

Explainability and Justification

- **Use tools to explain ML predictions**

- Even though accuracy may decrease, the explainability may be significant
- Important for end users to be able to understand ML predictions
- Especially important due to hype and misinformation about ML

- **Challenges**

- Potential leaking of sensitive data
- Easy to game, e.g., “adversarial feedback”
- Loss of competitive advantage
- Sometimes hard to interpret, even for experts

Explainability and Justification

- **Legal considerations**

- France's Digital Republic Act gives the right to an explanation as regards decisions on an individual made by algorithms
- How and to what extent the algorithm was used, which data was processed and its source, etc.
- Other countries considering similar laws

Displacement Strategy

- Identify and document relevant information so that business change processes can be developed to mitigate the impact on workers being automated
- Ensure all stakeholders are brought on board and develop a change-management strategy before automation
- Often, the workers are asked to do labor (e.g., generating training data) that will help automate themselves. Are they appropriately compensated?

Accountability

- **Question:** Should a passenger in automated car be able to command it to go 80 MPH on a 55 MPH road?
- **Reasons for “No”**
 - It’s illegal and can endanger others
 - Who is liable for accidents? Driver? Manufacturer? Insurance company?
- **Reasons for “Yes”**
 - Many exceptions!
 - Rushing someone to the hospital, escaping a tornado, etc.

Other Challenges

- The ethics of ML and AI systems is an urgent topic **now**, not because of speculative future scenarios
 - Open and active area of research, involves scholars from law, social sciences, etc., as well as domain experts
 - Law moves slowly, and legal frameworks have much to catch up to
- **Looking forward**
 - **AI safety:** How can we make AI without unintended negative consequences?
 - **AI alignment:** How can AI make decisions that align with our values?

Useful Tools

- IBM AI Fairness 360: <https://aif360.mybluemix.net/>
- Google ML Fairness Gym: <https://github.com/google/ml-fairness-gym>
- Facebook Fairness Flow: <https://venturebeat.com/2021/03/31/ai-experts-warn-facebooks-anti-bias-tool-is-completely-insufficient/>