CS 760: Machine Learning
Fairness & Ethics

Fred Sala

University of Wisconsin-Madison

Dec. 14, 2021

Announcements

- **Logistics**:
  - Project & HW 8 due tonight
  - Exam on Dec. 20^{th}.
  - Course survey due tomorrow.
  - **Final lecture: Thank you**!

- Class roadmap:

Today	**Fairness, Ethics, Robustness**
Dec. 20	**Final Exam**

Outline

- **ML in Society: Major Concerns**
  - Fairness, Accountability, Transparency, Robustness, Examples
- **Techniques**
  - Group and Individual Fairness, Differential Privacy, Defenses
- **Course Takeaways**
  - Don’t train on your test set and other tips

Our Class So Far…

- Technical aspects of models “in the lab”
  - Didn’t talk about **deploying** models in the world
  - Important to think about

Fairness

- How can we be confident that all groups are treated fairly?

Our test used Amazon Rekognition to compare images of members of Congress with a database of mugshots. The results included 28 incorrect matches.

The false matches were disproportionately of people of color, including six members of the Congressional Black Caucus, among

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

Credit scoring models can be between 5 and 10 percent less accurate for lower-income and minority homebuyers, new research shows. | Carlos Osorio

https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit

Accountability

- Which party takes responsibility for a failure in ML models?

How a computer algorithm caused grading crisis in British schools

PUBLISHED FRI, AUG 21 2020•7:18 AM EDT | UPDATED FRI, AUG 21 2020•8:45 AM EDT

Sam Shead
@SAM_L_SHEAD

SHARE

KEY POINTS

- Approximately 39% of A-level results were downgraded by exam regulator Ofqual's algorithm.
- Disadvantaged students were the worst affected as the algorithm copied the inequalities that exist in the U.K.'s education system.
- The U.K. government did a U-turn on the grading method as students went on protest.

https://www.cnbc.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html

California Teenager Dies in Self-Driving Tesla Crash

Contributor: Enjuris Editor
How can I contribute?

https://www.enjuris.com/blog/news/tesla-autopilot-accident/

Transparency

- How can we ensure models are transparent and comply with regulations?

TECH POLICY

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

By Karen Hao

January 21, 2019

https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Privacy

- How can we protect user privacy when ML models are used?

Netflix Cancels Contest After Concerns Are Raised About Privacy

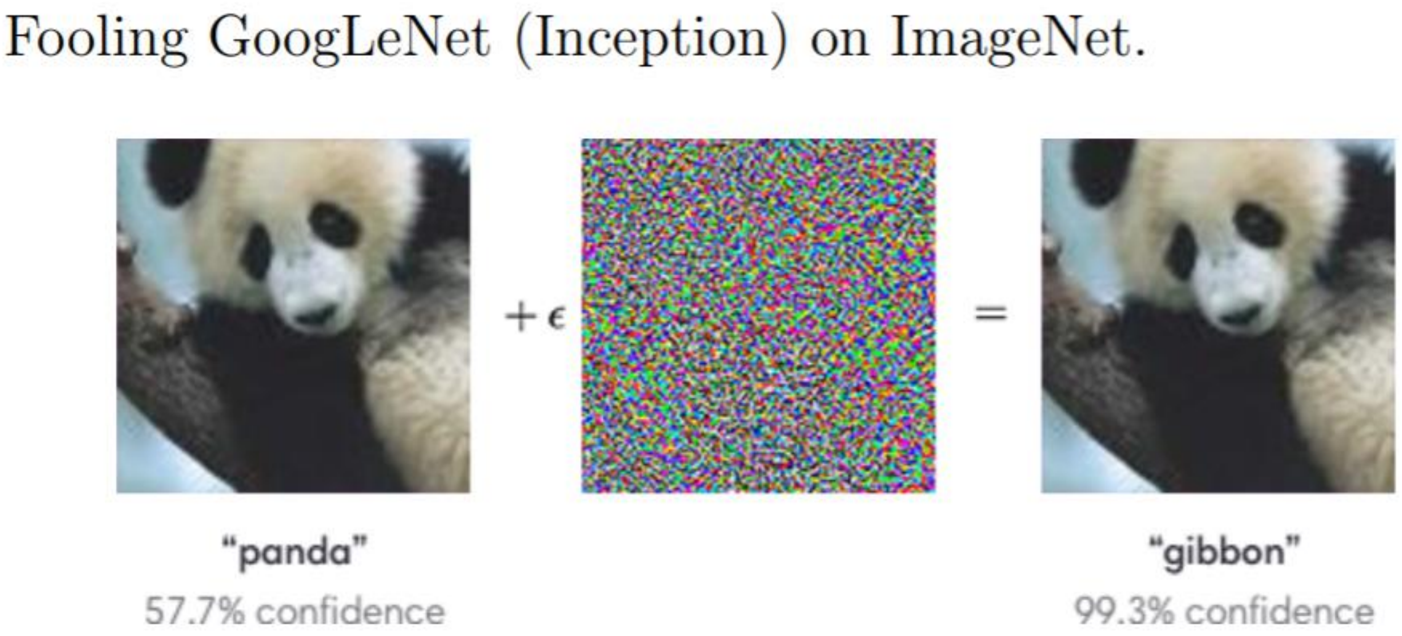
By Steve Lohr

March 12, 2010

https://www.nytimes.com/2010/03/13/technology/13netflix.html

Robustness

- How can we defend ML models against attacks?
  - E.g., data poisoning?



Adversarial Examples, Hanxiao Liu

More Bias Examples: Language Models

- Large language models **encode** bias
- **Example**: Religious Bias in GPT-3

AI’s Islamophobia problem

GPT-3 is a smart and poetic AI. It also says terrible things about Muslims.

By Sigal Samuel | Sep 18, 2021, 8:00am EDT

https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim

More Bias Examples: Word Embeddings

- Found in a variety of word embedding approaches:

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

	Gender stereotype *she*-*he* analogies	
sewing-carpentry	registered 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	lovely-brilliant
	Gender appropriate *she*-*he* analogies	
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al, “Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings”

Where Does Bias Come From?

- Models are trained on data, typically obtained by humans
- Models **inherit this bias** from training data
  - Example: many medical data collection efforts target one group over others

- Learning algorithms can even amplify this bias…
  - Recall: spurious correlations

Break & Quiz

Outline

- **ML in Society: Major Concerns**
  - Fairness, Accountability, Transparency, Robustness, Examples
- **Techniques**
  - Group and Individual Fairness, Differential Privacy, Defenses
- **Course Takeaways**
  - Don’t train on your test set and other tips

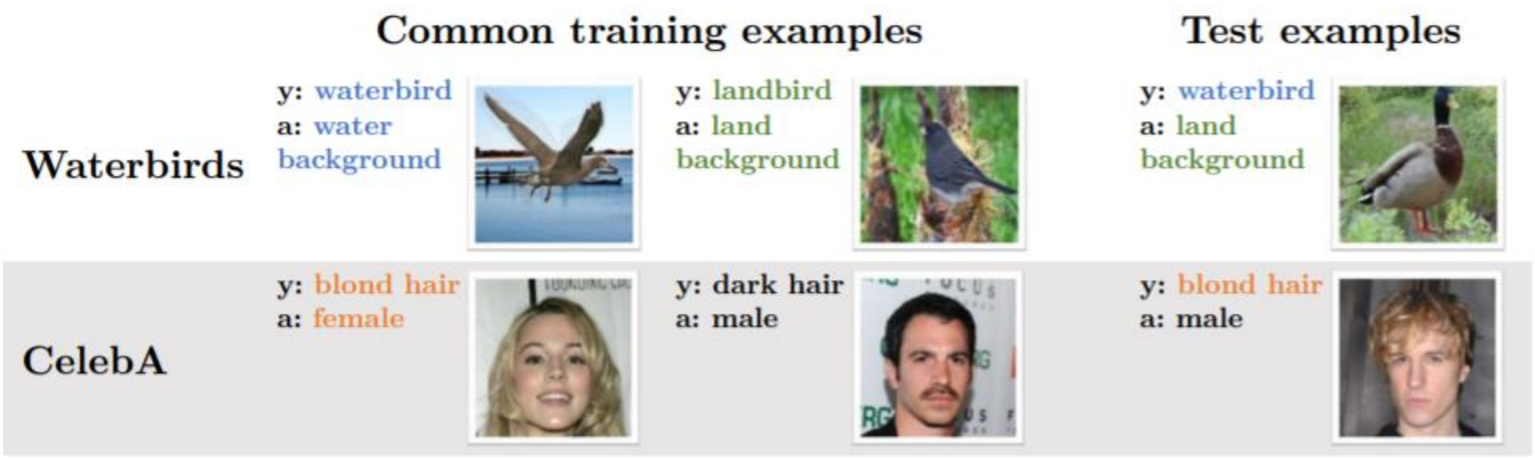
Mitigating Bias

Several approaches:

- 1. **Remove** bias from data
  - Better and more representative data
  - Remove bias associations: e.g., remove sentences with instances of bias
- 2. Design **fair learning** approaches
  - Add constraints to our learning approach

Mitigating Bias: Via Blindness

- Ignore all irrelevant/protected features
  - Don’t need such features for high performance
  - Often additionally **helps** generalization---avoid spurious correlation



Sagawa et al, “Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization”

Mitigating Bias: Group Fairness

- Equalize two groups S, T for outcomes:

$$P(\text{outcome O} \mid S) = P(\text{outcome O} \mid T)$$

- I.e., “the fraction of people in group S getting job offers should be the same as the fraction in T”

Group Fairness: Statistical Fairness

- How can we ensure this type of fairness?
  - ERM: fails to do this:

$$\hat{\theta}_{\text{ERM}} := \underset{\theta \in \Theta}{\arg\min}\; \mathbb{E}_{(x,y)\sim \hat{P}}[\ell(\theta; (x, y))]$$

  - Replace with a **group distributionally robust** RM

$$\hat{\theta}_{\text{DRO}} := \underset{\theta \in \Theta}{\arg\min} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y)\sim \hat{P}_g}[\ell(\theta; (x, y))] \right\}$$

Sagawa et al, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization"

Group Fairness: Statistical Fairness

- How can we ensure this type of fairness?
  - Replace with a **group distributionally robust** RM

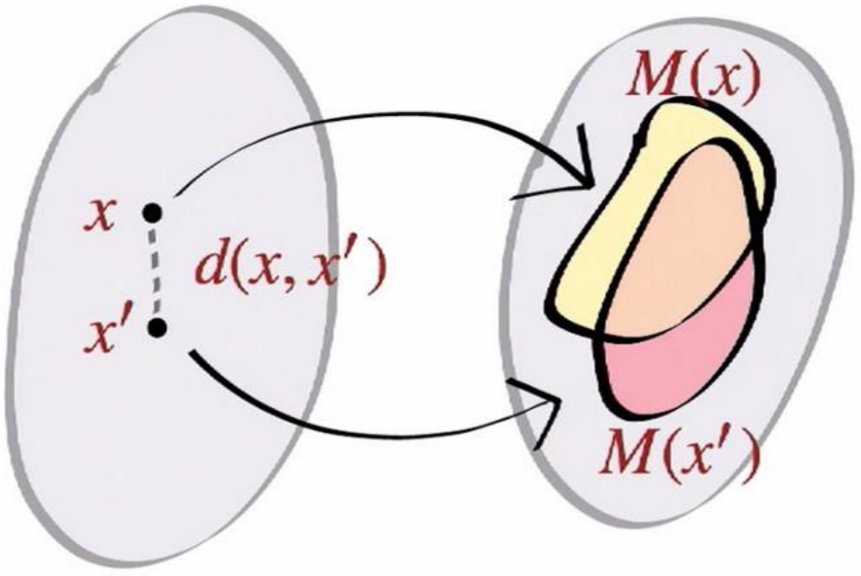
$$\hat{\theta}_{\mathrm{DRO}} := \underset{\theta \in \Theta}{\arg\min} \left\{ \hat{\mathcal{R}}(\theta) := \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim \hat{P}_g} [\ell(\theta; (x, y))] \right\}$$

		Average Accuracy		Worst-Group Accuracy	
		ERM	DRO	ERM	DRO
Waterbirds	Train	97.6	99.1	35.7	97.5
	Test	95.7	96.6	21.3	84.6
CelebA	Train	95.7	95.0	40.4	93.4
	Test	95.8	93.5	37.8	86.7

Sagawa et al, "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization"

Mitigating Bias: Individual Fairness

- Idea: Treat **similar** individuals **similarly**
  - E.g., similar for the purpose of the task – similar distribution over outcomes.
- Formalizing individual fairness:
  - M maps individual example to a distribution over outcomes
  - **Goal**: $D\big(M(x), M(x')\big) \leq d(x, x')$



Privacy

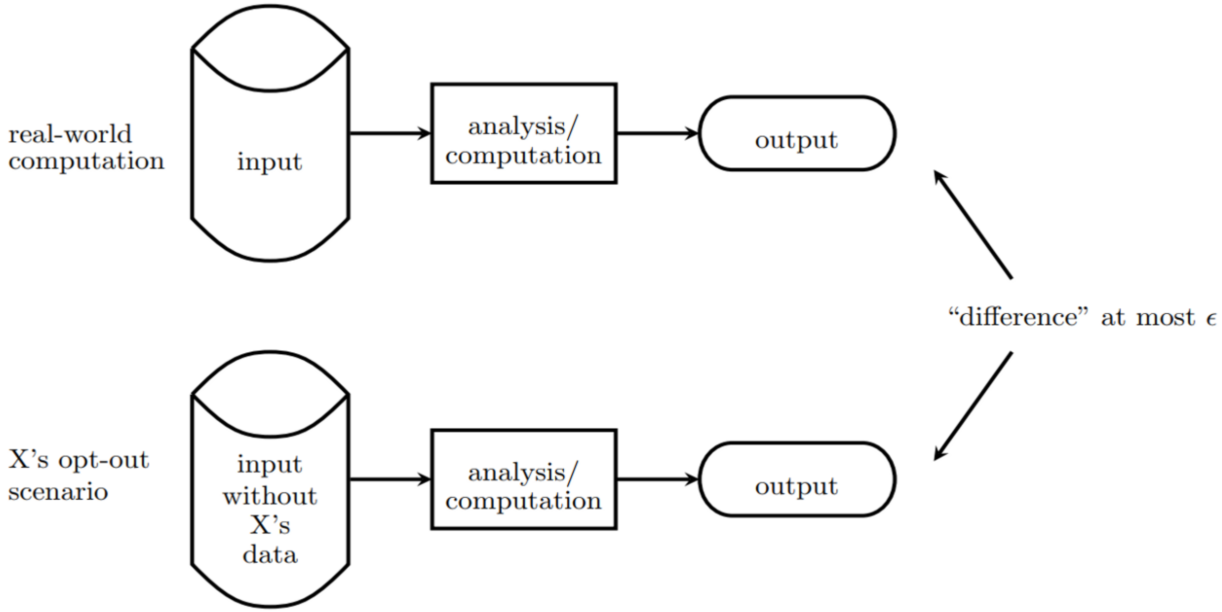
- Recall the Netflix prize: ~500000 users, 20000 movies
- No names provided, but possible to de-anonymyze:
  - Check versus IMDB database; not much information needed

	movie 1	movie 2	movie 3	movie 4	movie 5	movie 6
Tom	5	?	?	1	3	?
George	?	?	3	1	2	5
Susan	4	3	1	?	5	1
Beth	4	3	?	2	4	2

Narayanan and Shmatikov: "Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)"

Privacy: Differential Privacy

- Definition: an algorithm is **differentially private** if removing any datapoint will only slightly change any output
  - How to achieve it? Add specialized kinds of noise
  - More: https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

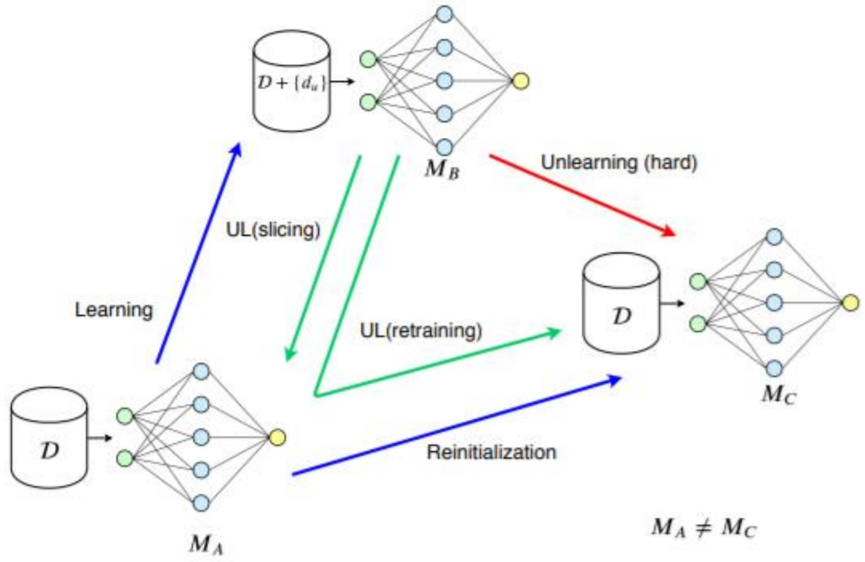


Credit: TDS

Privacy: Unlearning

- Increasingly popular regulation: the "right to be forgotten".
  - I.e., should be able to request online resources don't contain your information
  - Needed for ML models as well
- Leads to **machine unlearning**
  - Be able to delete the contribution of a particular data point to the trained model

Bourtoule et al, "Machine Unlearning"



Adversarial Attacks

- Models might face malicious attacks

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Eykholt et al, “Robust Physical-World Attacks on Deep Learning Visual Classification”

Adversarial Attacks

- Also common in NLP:

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

[Jia and Liang, 2017]

Break & Quiz

Outline

- **ML in Society: Major Concerns**
  - Fairness, Accountability, Transparency, Robustness, Examples
- **Techniques**
  - Group and Individual Fairness, Differential Privacy, Defenses
- **Course Takeaways**
  - Don’t train on your test set and other tips

Class Takeaways

- 1. **Understand** your goal.
- 2. Spend lots of time with your data.
  - Look at individual points. >50% of your time here.
- 3. Build your pipeline and check things run before optimizing.
- 4. Build high-quality infrastructure.
- 5. Practice with libraries & frameworks.
  - Feel comfortable with one particular framework.
- 6. Read related work… but don’t get stuck.
  - Don’t worry about hype
- 7. Try **simple baselines** first!

Post-Class

- If you need advice from me
  - On machine learning
  - On careers, industry, etc.
  - Academic advice
- Or just to chat about life.

Always happy to talk!

- Come by: my office is CS 5385.

Thanks Everyone!

Some of the slides in these lectures have been adapted/borrowed from materials developed by Mark Craven, David Page, Jude Shavlik, Tom Mitchell, Nina Balcan, Elad Hazan, Tom Dietterich, Pedro Domingos, Jerry Zhu, Yingyu Liang, Volodymyr Kuleshov, Sharon Li