Project Reminders

- Project presentations
  - 11/29 Group 1-8; 12/4 Group 9-16
  - Upload your slides (in the form of pdf) on Canvas before the day you present
  - Each team will be scored by other teams
  - Length: 1-2 person teams 6 min; 3-person teams 8 min

Project reminders: report

- Final report and deliverable due 12/13
- 6-8 page for 1-2 person teams; 10-12 pages for 3 person teams
- More details on Canvas

Project Presentation Rubric

- Problem overview and motivation (20%): Is the problem well described and motivated? Is the problem novel and challenging?
- Methodology (40%): Is the approach well described and justified? What do the dataset/features look like? What are the preprocessing, models, metrics and evaluation methodology? How does it compare to existing work?
- Preliminary results (20%): Are there any preliminary results or findings for preprocessing and the selected models? How does the result compared to existing work (if any)? Is the project on track to deliver?
- Presentation/slide quality and clarity (20%): Is the presentation clear, coherent, and compelling?

Project Report Rubric

- Problem overview and motivation (20%) — Is the problem well described and motivated? Is the problem novel and challenging? Is there sufficient discussion of related work?

- Methodology (30%) — Is the approach well described and justified? What do the dataset/features look like? What are the preprocessing, models, metrics and evaluation methodology? How does it compare to existing work?

- Results (30%) - Are there sufficient results? Are they clearly presented and discussed? Are there any insights drawn from the analysis of the results?

- Writing quality and clarity (20%) - is the report clear and coherent?

CS 534: Machine Learning

Emerging Topics

Fair and Explainable ML

Slides adapted from Osbert Bastani, Zachary Ives, Marzyeh Ghassemi, Krishnaram Kenthapadi, Ben Packer, Mehrnoosh Same, Nashlie Sephus, and Hima Lakkaraju

ML Success Stories

- Forecasting severe weather events
- Tracking and preserving wildlife
- Drug discovery / automated image analysis
- Traffic Prediction

ARDA

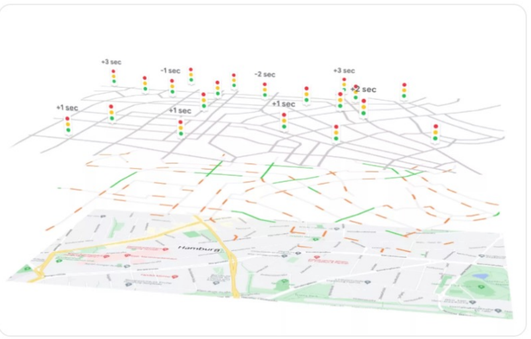
Preventing blindness by detecting diabetic retinopathy with AI.

Learn more

Forecasting Riverine Floods

Using AI to make flood forecasting information universally accessible.

Learn more



Green Light

Creating greener cities with AI.

Learn more

https://ai.google/responsibility/social-good/

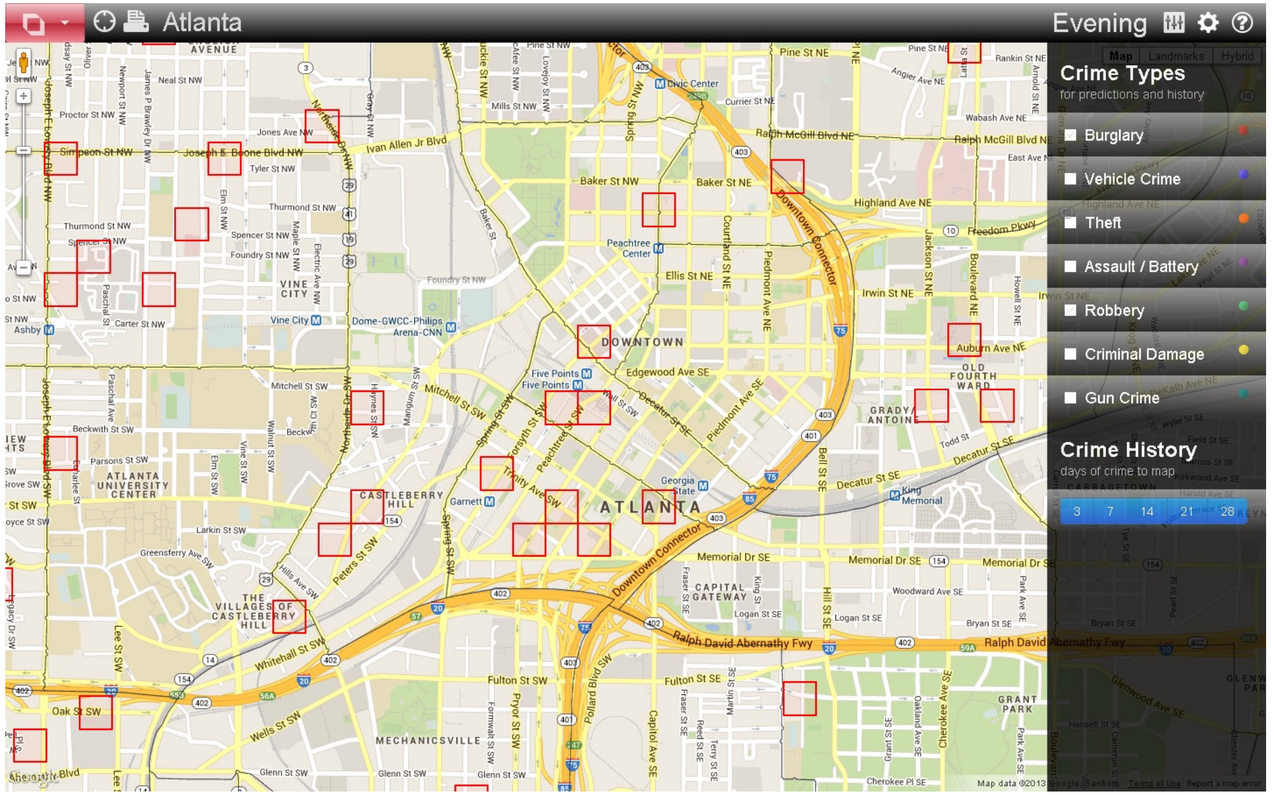
Thought Experiment #1

- 2 similar patients (one male & one female) visit the hospital with severe COVID-19 symptoms and require oxygen
- Problem: Only 1 oxygen machine is left
- ML model predicts female more likely to survive and should receive the machine

Are you okay with this decision?

Thought Experiment #2

- ML system decides where and how much to patrol
- Based on places that model predicts high crime, send police officer there



What can go wrong in this scenario?

https://www.predpol.com

Ethics is Hard!

- Ethical decision-making has been debated for thousands of years
- Challenging problem even without ML (e.g., philosophy, law, medicine, etc.)
- Changes over time with changing societal norms

Ethical Challenges with ML

- Data privacy issues
- Potential to amplify biases already present in the data
- New issues related to abuse of ML

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM ▸ STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

As part of the Biden-Harris Administration's comprehensive strategy for responsible innovation, the Executive Order builds on previous actions the President has taken, including work that led to voluntary commitments from 15 leading companies to drive safe, secure, and trustworthy development of AI.

The Executive Order directs the following actions:

Ethical Challenges with ML

Fairness

Privacy

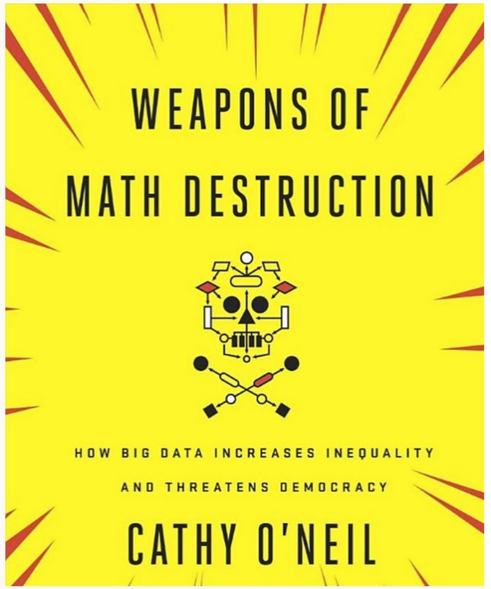
Transparency

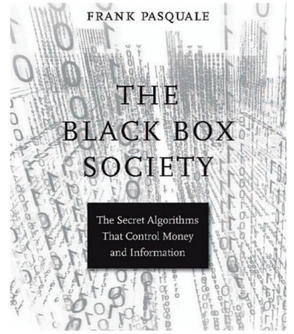
Explainability

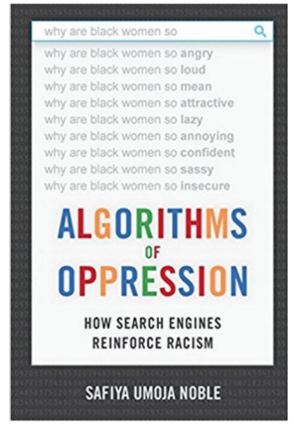
Goal of today's lecture is to make you aware of the potential problems — still many unanswered questions!

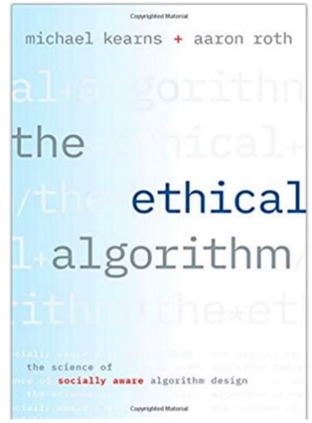
Discrimination in ML

- Inherent biases present in society and reflected in training data
- ML models prone to amplify such biases









- Algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care
- Heavily favored white patients over black patients
- While race is not used in the algorithm, healthcare cost history highly correlated with race

OCTOBER 24, 2019 | 4 MIN READ

Racial Bias Found in a Major Health Care Risk Algorithm

Black patients lose out on critical care when systems equate health needs with costs

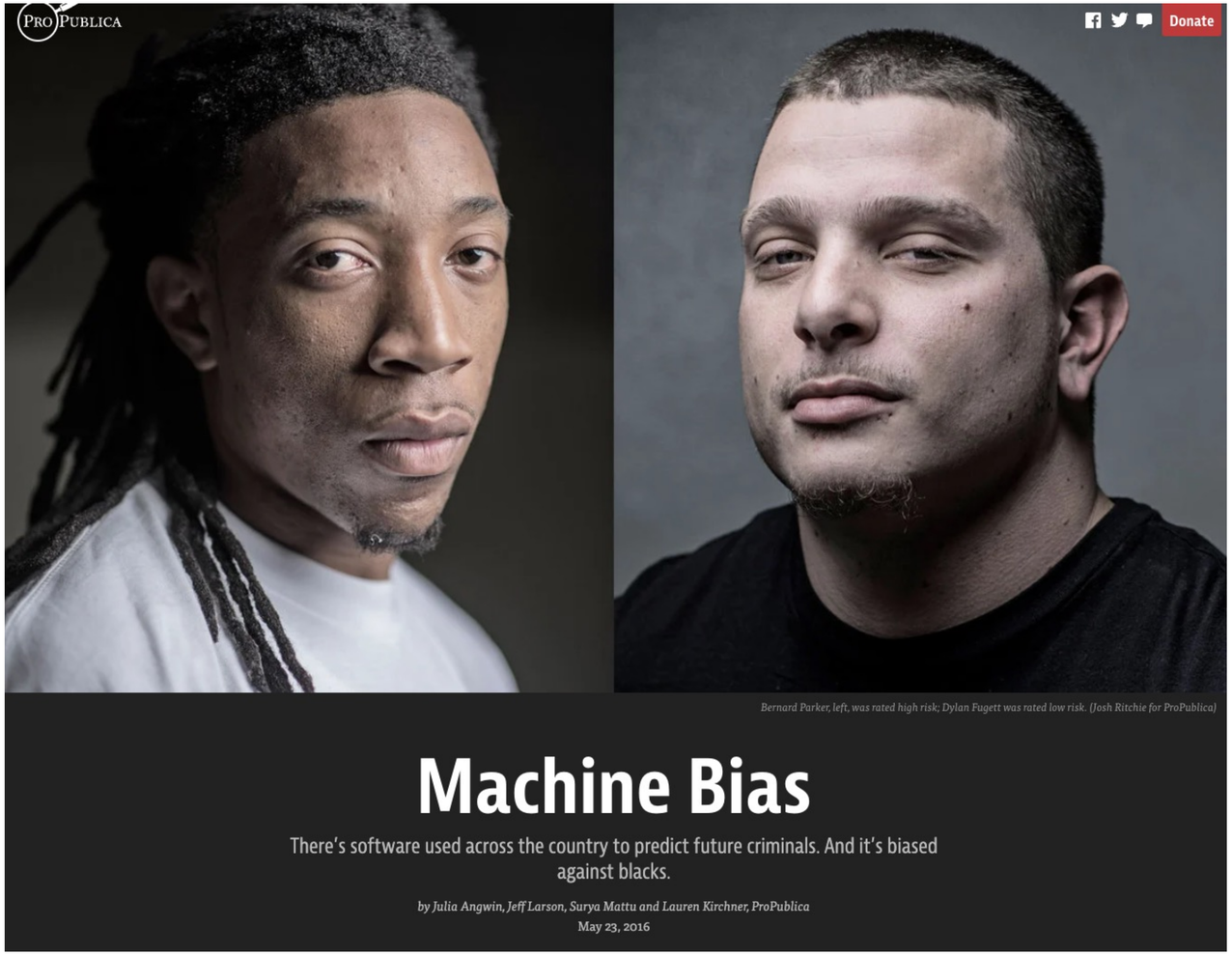
BY STARRE VARTAN

Health care algorithms can reinforce existing inequality. Credit: Getty Images

https://www.scientificamerican.com/article/racial-bias-found-in-a-major-health-care-risk-algorithm/

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm used in US court systems

- Predicted twice as many false positives for black offenders (45%) than white offenders (23%)



- ML based resume screening algorithm
- Used resumes over a 10-year period (most came from men)
- Penalizes resumes of female candidates

Amazon ditched AI recruiting tool that favored men for technical jobs

Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process

Amazon's automated hiring tool was found to be inadequate after penalizing the résumés of female candidates. Photograph: Brian Snyder/Reuters

Gender in Word Embeddings

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], **Kai-Wei Chang**[2], **James Zou**[2], **Venkatesh Saligrama**[1,2], **Adam Kalai**[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

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.

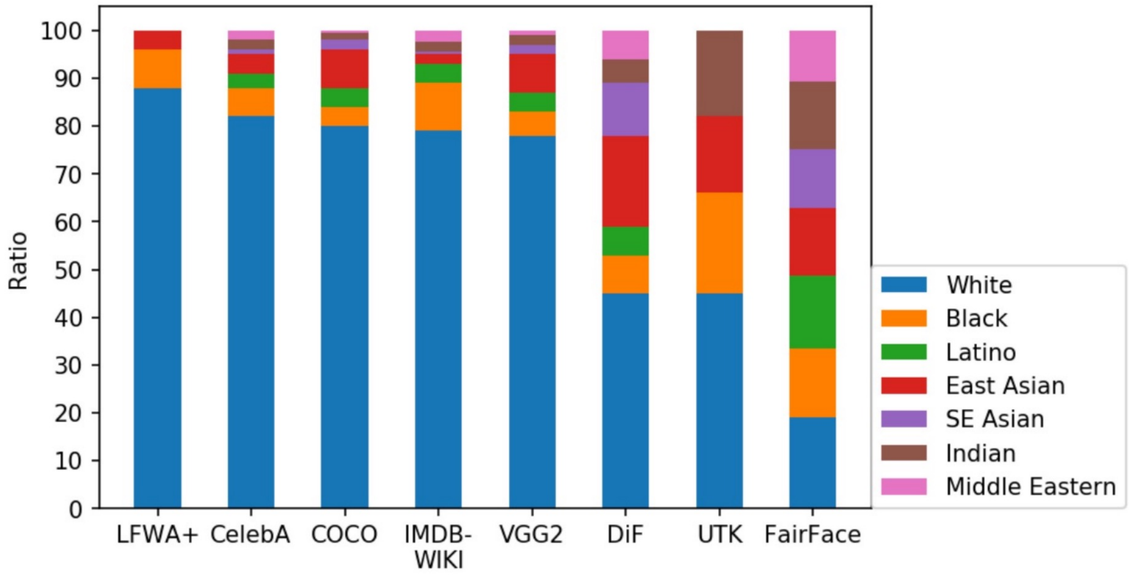
sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas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

Gender appropriate *she-he* analogies.

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

Data Representation

- Less data from minority groups —> higher error on minority groups
  - Clinical trials use to recruit white males (racial/gender bias)
  - Easily accessible data (e.g., Twitter) or news articles may reflect social gender bias
  - Mobility data under represent rural areas and senior communities
- Need to be careful to gather representative datasets



Datasheets for Datasets

- Questions that should be considered
  - Motivation
  - Dataset composition and collection process
  - Preprocessing
  - Distribution, use, and maintenance

https://arxiv.org/abs/1803.09010

Fairness in ML

- Legally protected attributes: race, sex, color, religion, national origin, age, citizenship, pregnancy, familial status, disability, veteran status, genetic information

- Two individuals differing on sensitive (i.e., legally protected) attributes but otherwise identical should receive the same outcome

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%

Fairness Criteria

- Most common way to define fair classification is to require some invariance with respect to the sensitive attribute.
- E.g. demographic parity

$$P(R = + \mid A = a) = P(R = + \mid A = b) \quad \forall a, b \in A$$

- Many of these definitions are incompatible!
- Need to carefully choose kinds of fairness we require

Name	Closest relative	Note
Statistical parity	Independence	Equivalent
Group fairness	Independence	Equivalent
Demographic parity	Independence	Equivalent
Conditional statistical parity	Independence	Relaxation
Equal opportunity	Separation	Relaxation
Equalized odds	Separation	Equivalent
Conditional procedure accuracy equality	Separation	Equivalent
Disparate mistreatment	Separation	Equivalent
Balance for positive class	Separation	Relaxation
Balance for negative class	Separation	Relaxation
Predictive equality	Separation	Relaxation
Conditional use accuracy equality	Sufficiency	Equivalence
Predictive parity	Sufficiency	Relaxation
Calibration	Sufficiency	Equivalence

https://en.wikipedia.org/wiki/Fairness_%28machine_learning%29

Fairness Algorithms

- Given a notion of fairness, you can adjust at the 3 ML stages
  - Pre-processing: Adjust features to be uncorrelated with sensitive attribute
  - Training: Impose fairness constraint during learning process; adversarial debiasing
  - Post-processing: Adjust learned classifier so its predictions are uncorrelated with sensitive attribute

Ethical Challenges with ML

Fairness

Privacy

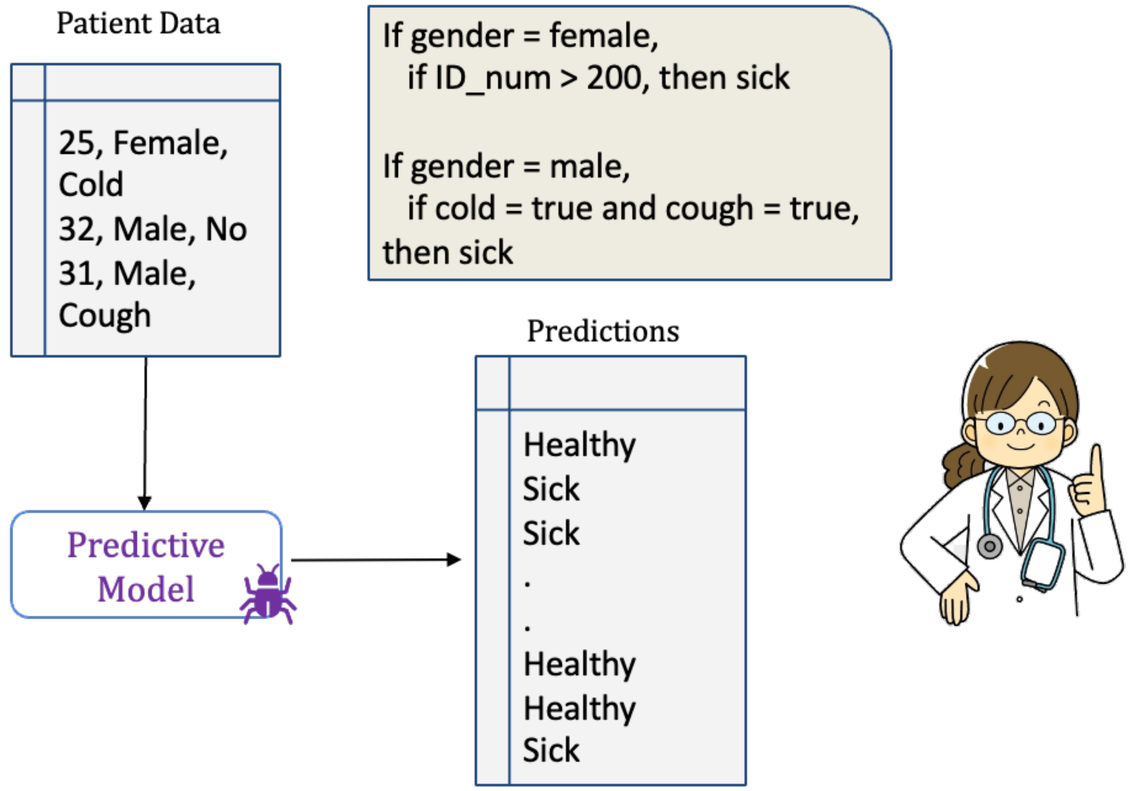
Transparency

Explainability

Goal of today's lecture is to make you aware of the potential problems — still many unanswered questions!

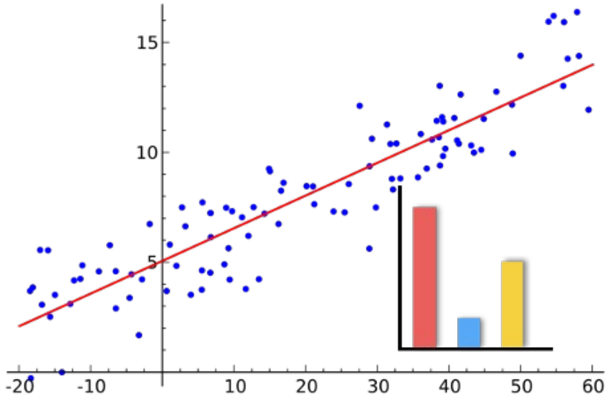
Explainability

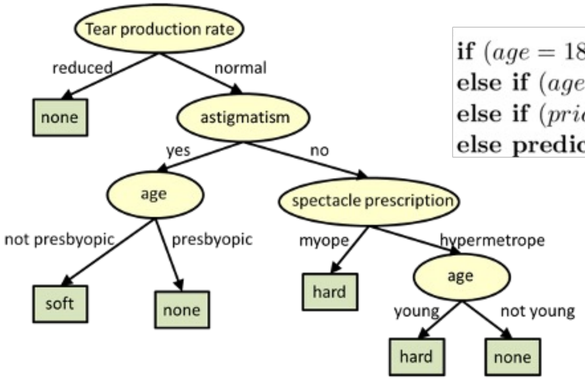
- Debugging
- Bias detection
- Provides recourse for adversely affected samples
- If/when to trust model predictions



Two Sides of Explainability

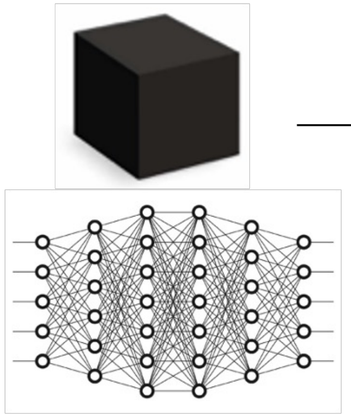
Side #1: Build *inherently interpretable* predictive models

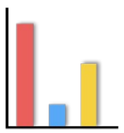




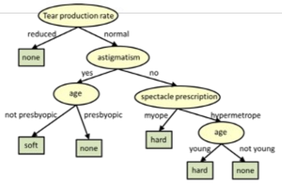
if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Side #2: *Explain* pre-built manners in a *post-hoc manner*

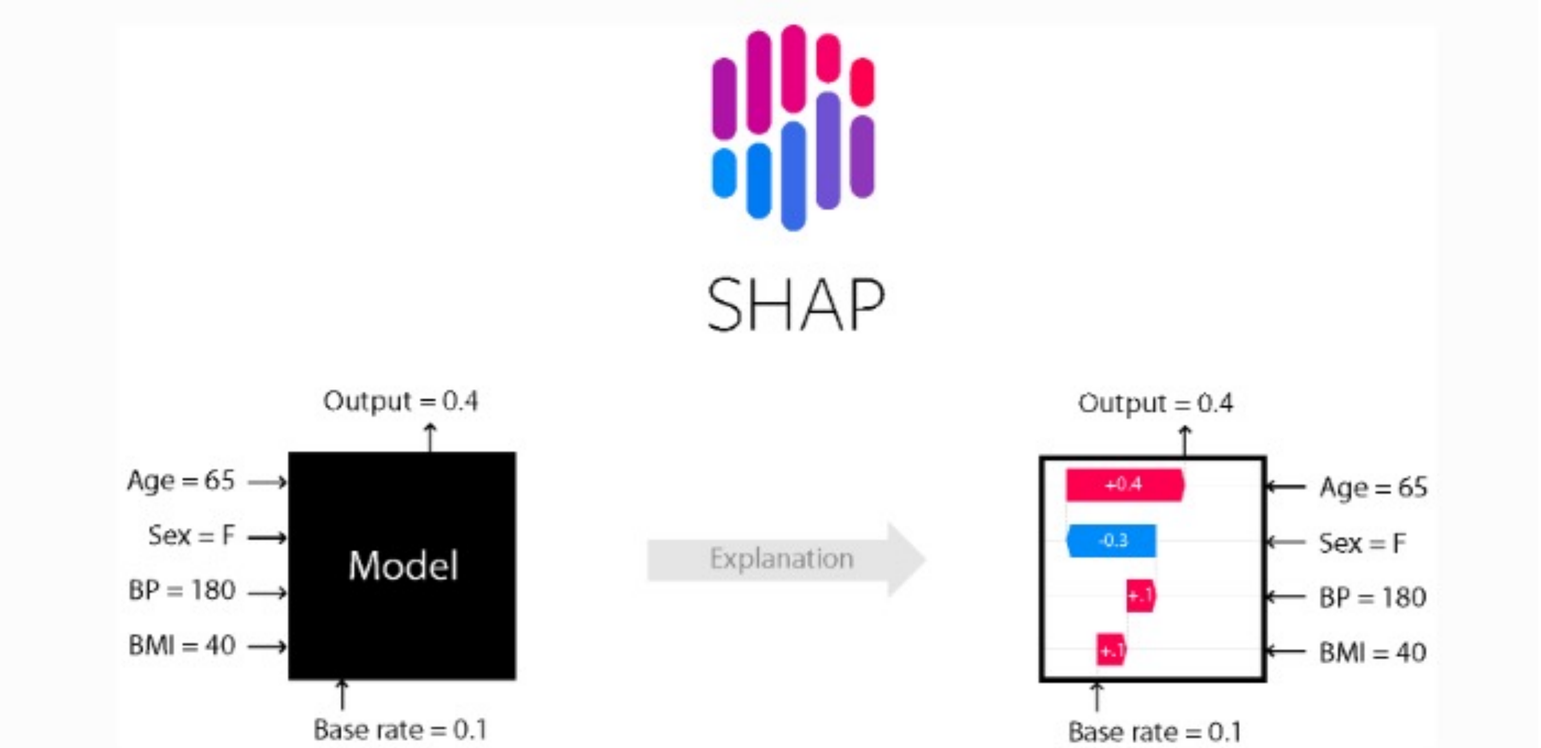




if $(age = 18-20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21-23)$ **and** $(priors = 2-3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

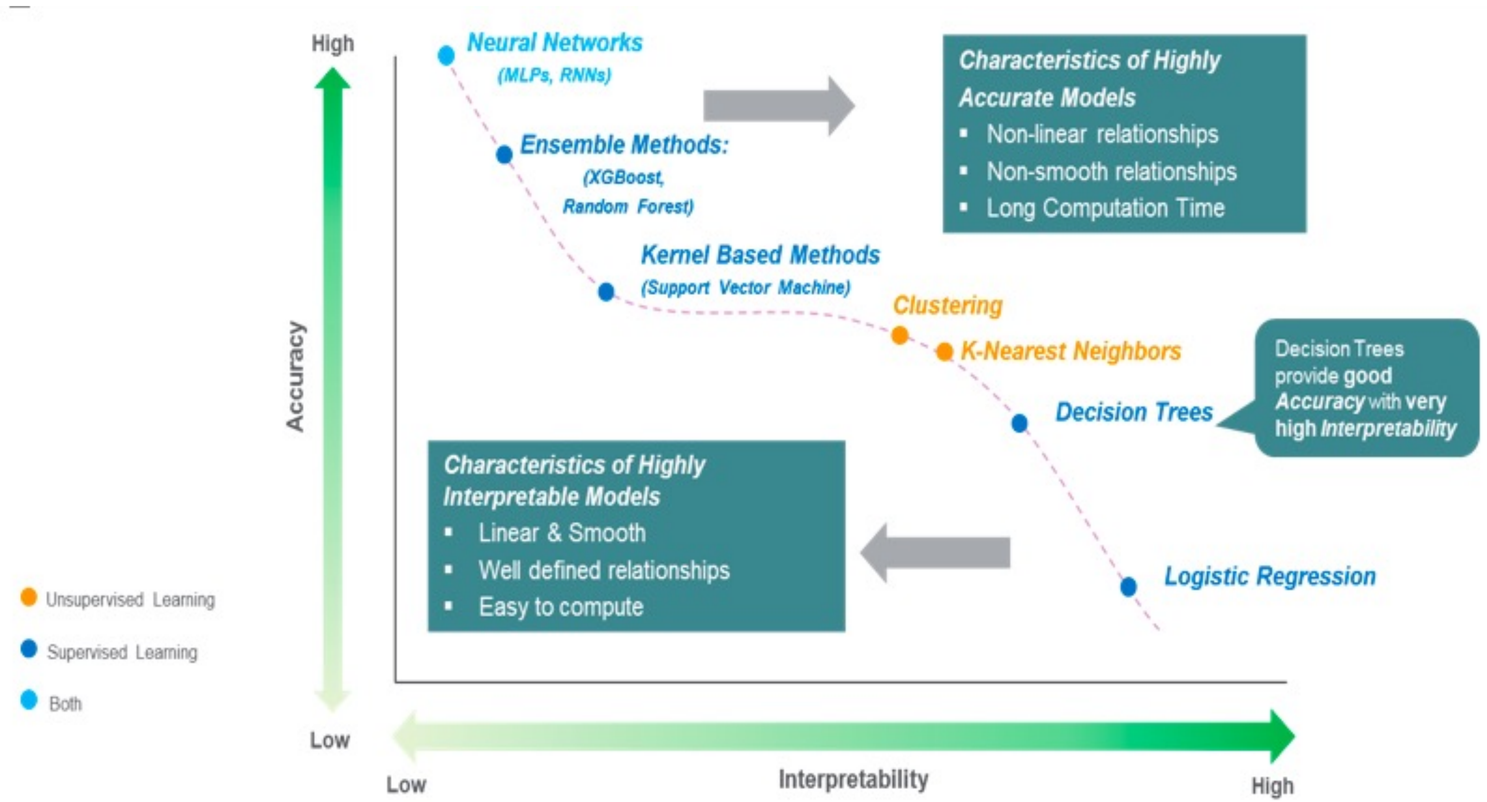


Model Agnostic Approaches



SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (see papers for details and citations).

Two Sides of Explainability



In certain settings, accuracy-interpretability trade-offs may exist

https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062

Ethical Challenges with ML

Fairness

Privacy

Transparency

Explainability

Goal of today's lecture is to make you aware of the potential problems — still many unanswered questions!

Potential Abuse of ML

Artificial Intelligence

How ChatGPT Could Spread Disinformation Via Fake Reviews

Anything of true potential could be misused with wrong intentions. Could ChatGPT be misused for disinformation?

Nico Dekens Director Of Intelligence & Collection Innovation, ShadowDragon July 7, 2023

Home › News ›

JULY 5, 2023 Editors' notes

ChatGPT generates 'convincing' fake scientific article

by JMIR Publications

AI-generated image, in response to the request "pandoras box opened with a physician standing next to it. Oil painting Henry Matisse style", (Generator: DALL-E2/OpenAI, March 9, 2023,

ChatGPT Makes Spotting Fake News Impossible for Most People

Tweets generated by OpenAI's GPT-3 model are so convincing, people can't spot when they promote misinformation.

Written by **James Laird**

Updated on **June 29, 2023**

Other Ethical Questions

- **Accountability**: Who is responsible for the decision making process if ML does something unexpected?
- **Safety**: How can we prevent unintended negative consequences?
- **Environmental Impact**: What is the amount of energy needed to power the system?
- **Sustainability**: What is the impact on physical, social, and political ecosystems?

Where we are today

- Open and active research area
- Growing community focused on fairness, accountability, and transparency (FAccT)
- Requires close work with social sciences and law as well as deployment domains

NEWS | 24 January 2020

The battle for ethical AI at the world's biggest machine-learning conference

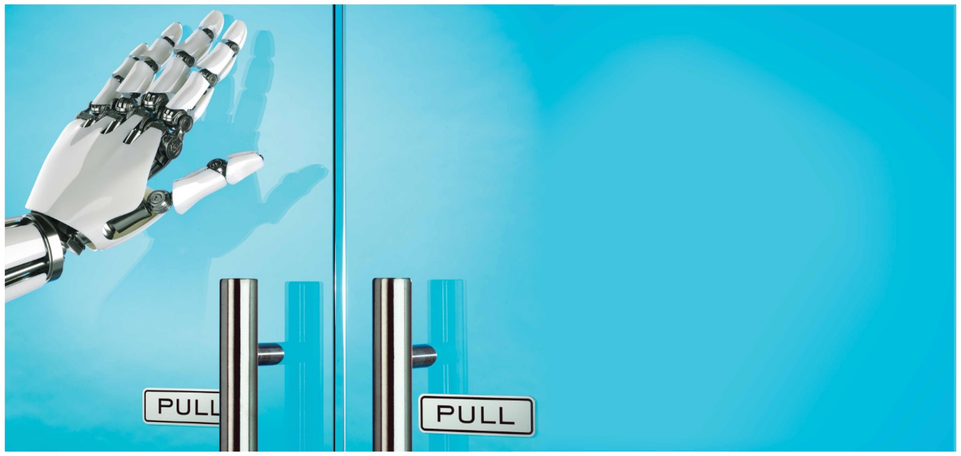
Bias and the prospect of societal harm increasingly plague artificial-intelligence research – but it's not clear who should be on the lookout for these problems.

Technology And Analytics

When Machine Learning Goes Off the Rails

A guide to managing the risks by Boris Babic, I. Glenn Cohen, Theodoros Evgeniou, and Sara Gerke

From the Magazine (January–February 2021)



Gregory Reid/Gallery Stock

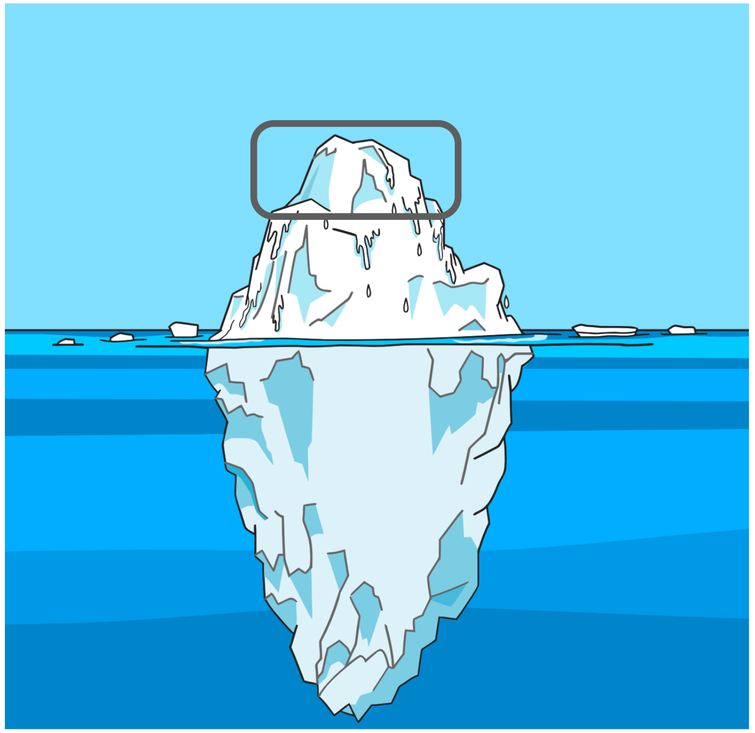
Best Practices for Ethical AI

https://ethical.institute/principles.html

Course Summary

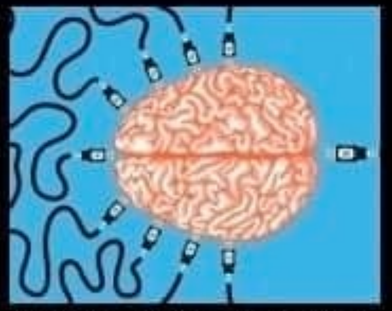
- Supervised learning: regression and classification
  - Different models
  - Model assessment and selection
  - Bias/variance tradeoff
- Unsupervised learning: dimensionality reduction
- Emerging topics: responsible ML

What we covered!



Machine Learning

What society thinks I do.



What my friends thinks I do.

What computer scientists think I do.

What my boss thinks I do.

What I think I do.

What I really do.