AI ETHICS: WHAT PHYSICISTS NEED TO KNOW

Savannah Thais, Princeton University

Snowmass Societal and Environmental Impacts Group Kick Off Town Hall

07/22/2021

Princeton
Research Computing

Introduction

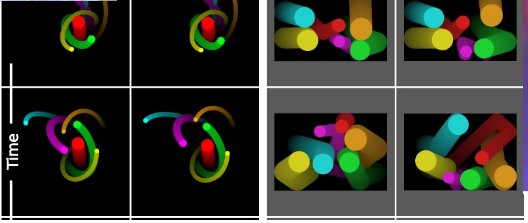
- This talk in 30 seconds:
  - AI is an integral part of physics research, physicists are directly contributing to the field
  - AI is ubiquitous in almost every aspect of life and in some cases is causing irreparable harm
  - Physicists can crucially contribute to efforts to build more equitable AI (and have a duty to do so)
- Talk Outline
  - AI, Physics, and Ethics: what does this intersection look like?
  - Societal Impact: how is AI impacting our lives and communities?
  - The Role of Physicists: how can we contribute to efforts to reduce harm and build more robust AI systems?

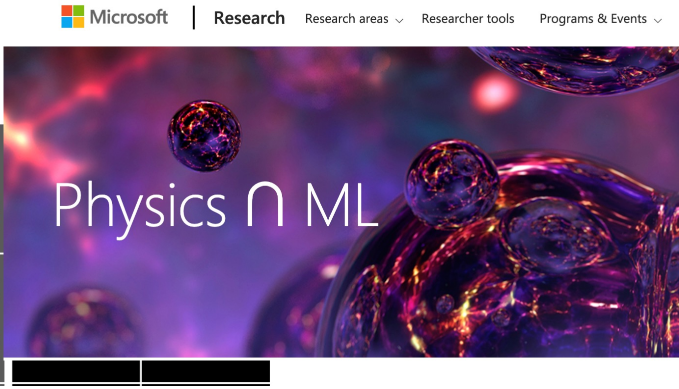
AI, Physics, and Ethics

Physics and AI/ML

- AI/ML are now critical components of the HEP research pipeline
  - Subject specific summer schools, software institutes, conferences
  - Thousands of papers published in physics and computing venues
- Physics is also impacting and advancing AI/ML research
  - Industry collaborations and career paths
  - Physics informed ML is a growing area of research
- This information exchange doesn't exist in a vacuum
  - Algorithms and analysis techniques that physicists contribute to aren't only useable by physicists
  - This research directly impacts our lives and our communities

Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)





These Are Powerful Tools For Physics and Beyond…

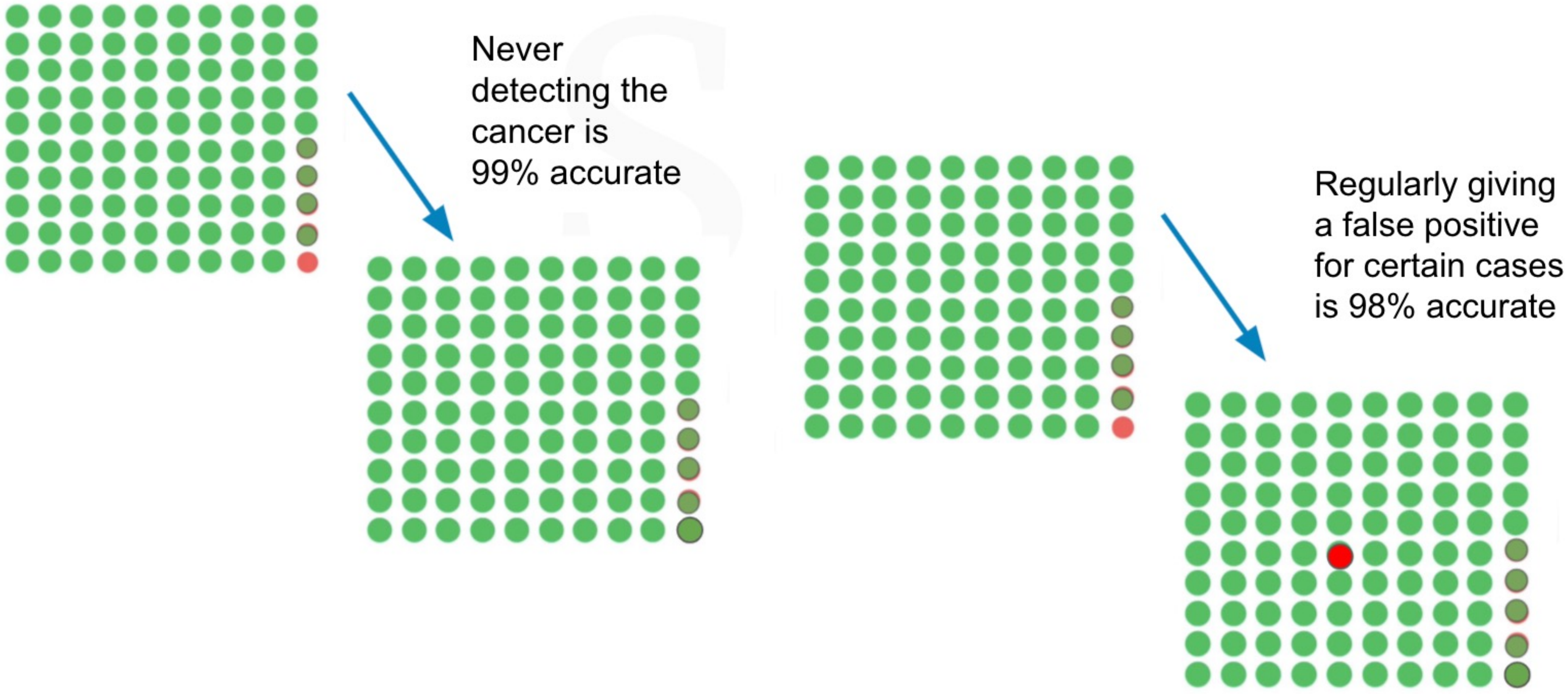
- AI has advanced rapidly over the past ~25 years and has demonstrated lots of exciting impact!
- AI algorithms are applied in nearly every area of life
  - Entertainment: beating Go, chess, and video games, generating art, music, screenplays…
  - Medicine: cancer detection, drug discovery, treatment plan, pandemic prediction
  - Day-to-day activities: GPS routing, spam detection, playlist generation, Netflix recommendations, image tagging, text prediction, self-driving cars
  - Social media and information: news feed curation, google results ordering, language translation, image captioning...
- An Artificial Neural Network with sufficient depth or width can approximate any function
  - Translation between input space (data) and output space (prediction)

But…

Imagine you’re trying to use brain scans to detect a very rare kind of cancer and you build a computer vision algorithm that is 99% accurate!

But…

Imagine you’re trying to use brain scans to detect a very rare kind of cancer and you build a computer vision algorithm that is 99% accurate!



Never detecting the cancer is 99% accurate

Regularly giving a false positive for certain cases is 98% accurate

They Can Be Very Difficult to Use Well

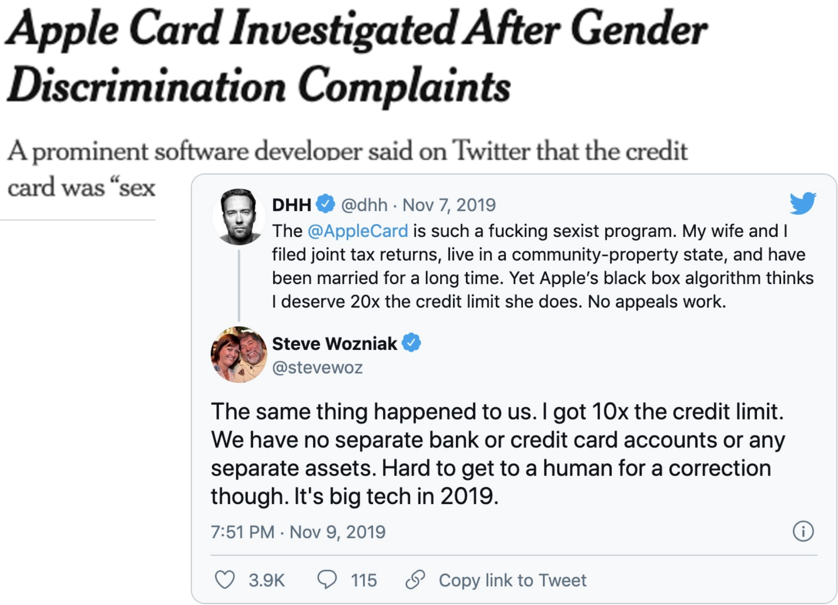
- What function are we trying to approximate?
  - Can we mathematically define the outcome we're interested in?
  - How well does the function the algorithm is learning approximating the function we're interested in? Can we even measure this?
  - Are there impacts, correlations, or data points we want to avoid?
- Where does the training data come from?
  - How is it stored? Can it be reused? Who owns it?
- How are these systems designed and built?
  - Do users/impactees opt in?
  - Who decides the function to be learned and what is 'good enough'?
  - How are they deployed? Who reaps the benefits and who is negatively impacted?

Societal Impact

Data Biases

- Apple Pay Card gave higher (or any) credit limits to men
  - Models trained on historical data may be 'accurate' but not 'fair'
  - Removing class labels from training data doesn't force fair outcomes

- Healthcare risk assessment under-estimates disease severity in African American patients
  - Healthcare spending in the previous year was highly weighted
  - Ignoring broader context/domain knowledge can be devastating



Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sex

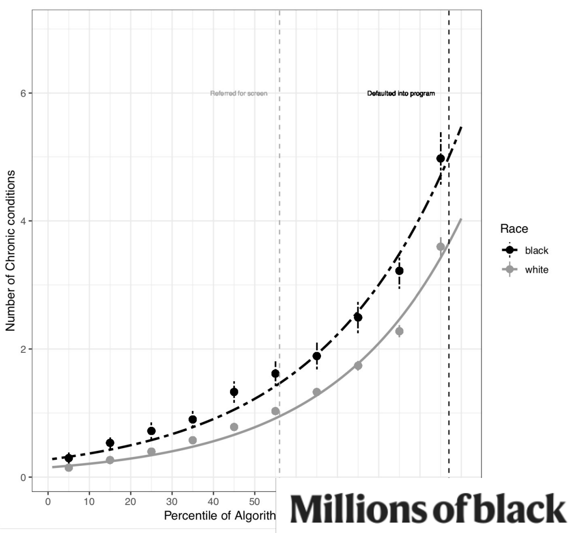
DHH @dhh · Nov 7, 2019

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

Steve Wozniak @stevewoz

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

7:51 PM · Nov 9, 2019

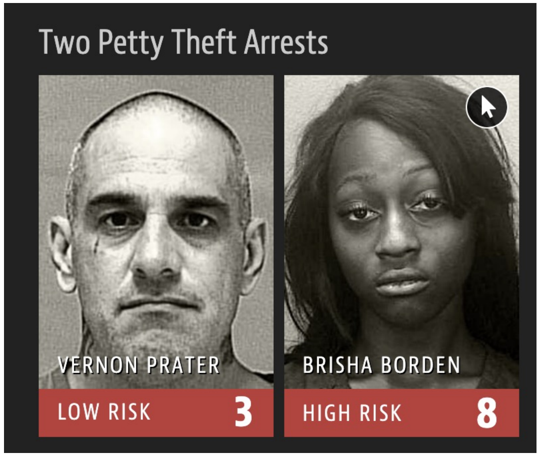


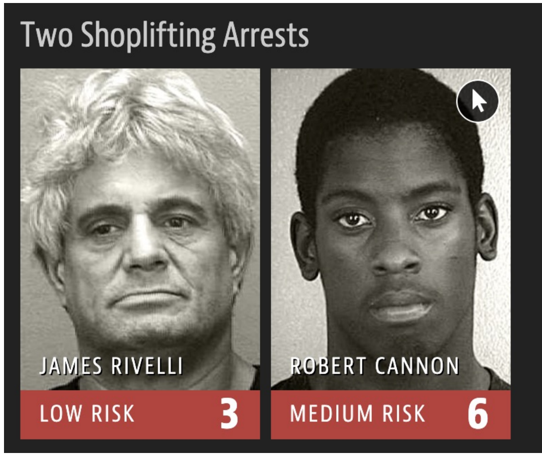
Millions of black people affected by racial bias in health-care algorithms

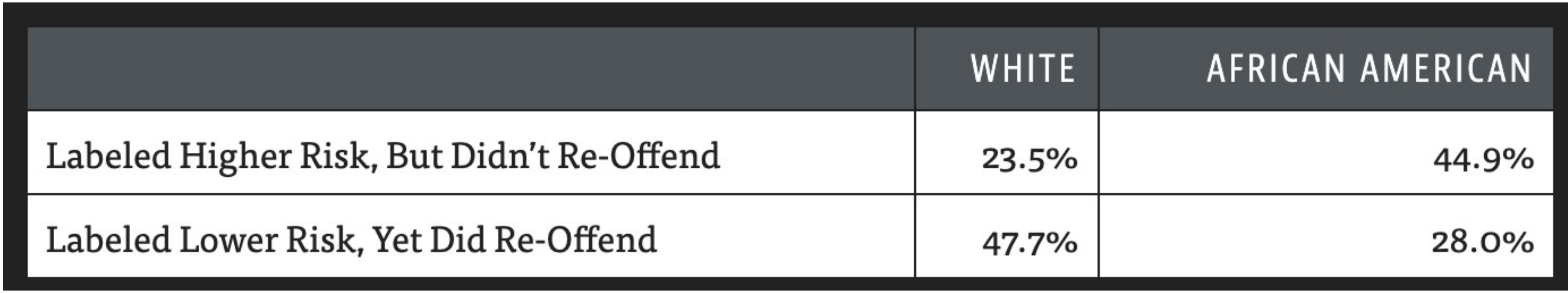
Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

Data Biases

- COMPAS Recidivism prediction tool predicts higher risk scores for minorities
  - Race is not an explicit factor in the score: based on survey questions and criminal records
  - But there is historical bias in which communities are policed and who is sentenced
  - Known relationship between socioeconomic status and petty crime (all crimes are considered in the model, training data not shared)
- Overall accuracy was considered but not accuracy across classes and severities





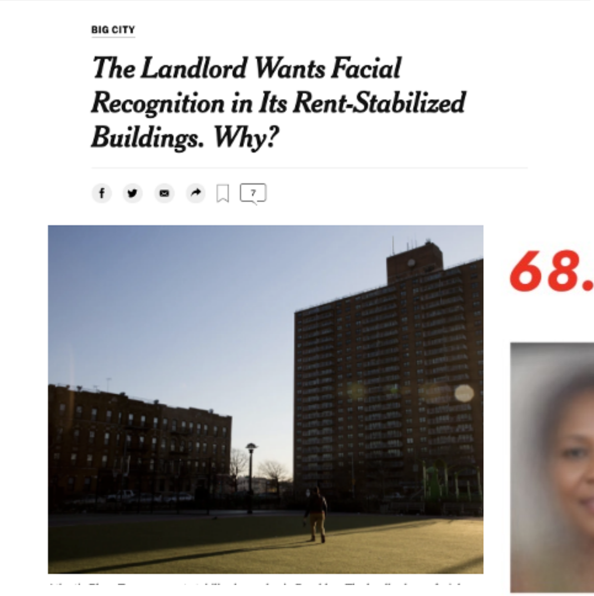


	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Inequitable Applications

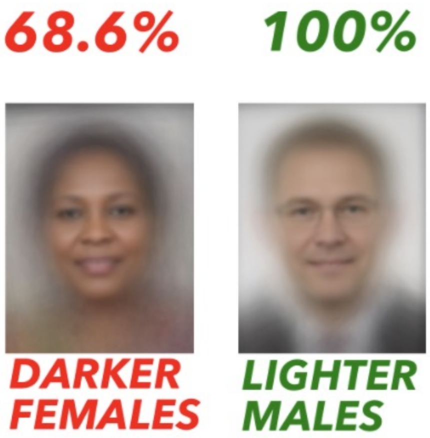
- Using facial recognition entry systems in rent-stabilized housing
  - Commercial facial recognition systems have demonstrated bias towards white faces
  - Deploying it in low-income, predominantly minority communities can be an effort towards gentrification

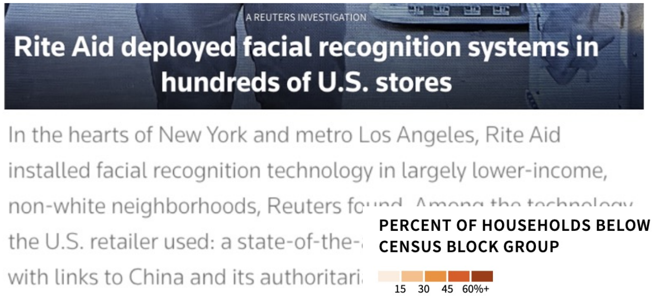
- Rite Aid deployed facial recognition only in low-income areas
  - Systems are often deployed on communities they're not designed for, who don't have a say in their development, and don't opt in
  - Privacy as an inherent right vs economic privilege



BIG CITY

The Landlord Wants Facial Recognition in Its Rent-Stabilized Buildings. Why?



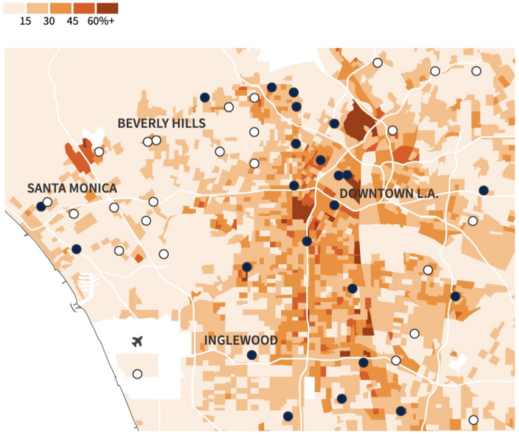


A REUTERS INVESTIGATION

Rite Aid deployed facial recognition systems in hundreds of U.S. stores

In the hearts of New York and metro Los Angeles, Rite Aid installed facial recognition technology in largely lower-income, non-white neighborhoods, Reuters fo… Among the technology the U.S. retailer used: a state-of-the- with links to China and its authoritari

PERCENT OF HOUSEHOLDS BELOW POVERTY LINE BY CENSUS BLOCK GROUP



Politics, Targeting, and Regulation

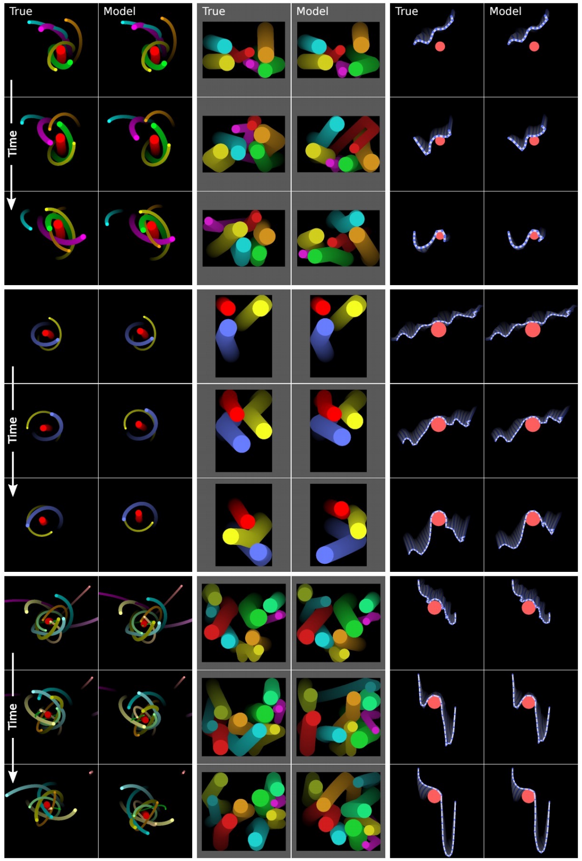
- Chinese government has employed facial recognition and racial classification algorithms to target Muslims
- Predictive policing algorithms target neighborhoods with higher police activity, regularly mis-identify people
- US Government has utilized Automated Decision Systems that are not auditable or tested for accuracy/bias
  - Misplaced focus on efficiency
  - Often results in lawsuits and the algorithm eventually being scrapped

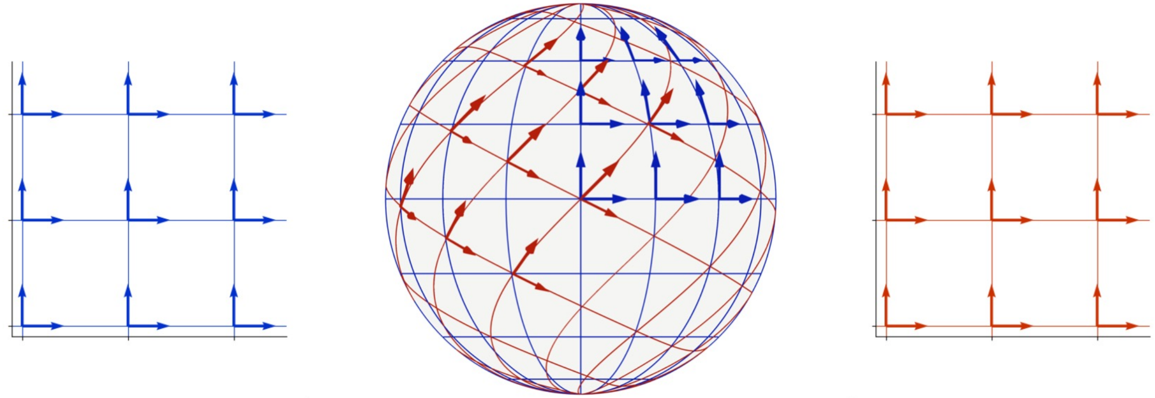
The Role of Physicists

Physics and Interpretability

Unlike many ML application domains, with physics we have a (approximately) robust underlying mathematical model

- ML for physics: by modeling physics processes with ML we can probe what, where, and how the algorithm is learning about certain data
- Physics for ML: by studying learning as a stochastic process we can optimize models and training

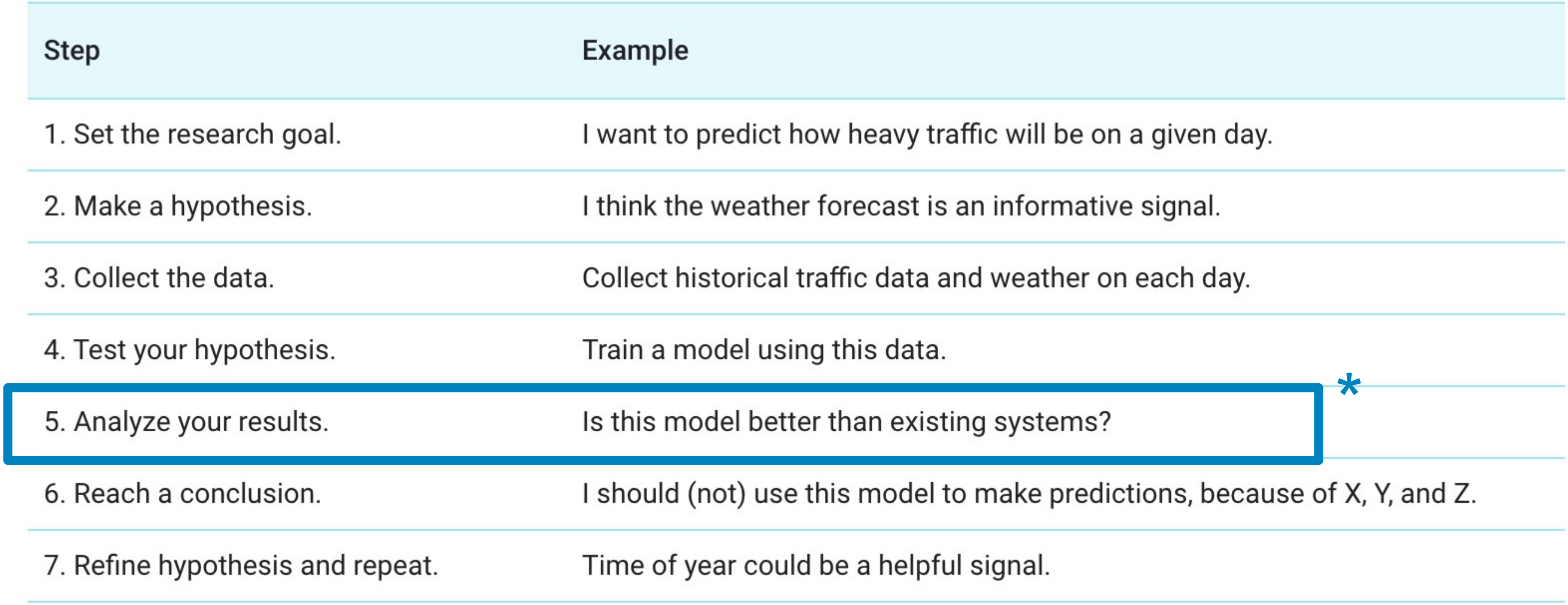




$$\begin{tikzcd} V_1 & U_1 \subset M \supset U_2 \arrow[l, "\varphi_1"'] \arrow[r, "\varphi_2"] & V_2 \end{tikzcd}$$

ML Should Be Scientific!

- ML is facing a reproducibility crisis
- Designing a (good) ML model is like running a scientific experiment: we don't know apriori what will work best



Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems? *
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

* Including how certain you are!

source

We Are Community Members

As scientists and citizens we have a duty to our communities

- Physics has a rich history of community outreach
  - Can we help increase technical literacy?
  - As the field that developed nuclear weapons and has grappled with those impacts, we have a unique perspective
- We are training future AI researchers and developers
  - What do we want to make sure they know?
- AI is impacting equity and accessibility at every level
  - This will impact our field, our universities and labs, and our communities
  - We are building these systems!

How do we gather community input to inform how we will judge + evaluate automated decision-making?

AI Ethics and Physics Efforts

- Snowmass contributions
    - LOI “Ethical implications for computational research and the roles of scientists”
    - Working on full White Paper on Ethics in Computing
        - Planning to involve a broad community of experts
        - Reach out to Savannah Thais, Brian Nord, or Aishik Ghosh to get involved
- Physics related publications:
    - “Physicists Must Engage with AI Ethics, Now”, APS.org
    - “Fighting Algorithmic Bias in Artificial Intelligence”, Physics World
    - “Artificial Intelligence: The Only Way Forward is Ethics”, CERN News
    - “To Make AI Fairer, Physicists Peer Inside Its Black Box”, Wired
    - “The bots are not as fair minded as the seem”, Physics World Podcast
    - “Developing Algorithms That Might One Day Be Used Against You”, Gizmodo
    - “AI in the Sky: Implications and Challenges for Artificial Intelligence in Astrophysics and Society”, Brian Nord for NOAO/Steward Observatory Joint Colloquium Series

Data analysis and algorithm development are big responsibilities

Some Great Resources

- AI Now
- Data & Society
- Berkman Klien Center
- Stanford Center for Human-Centered AI
- Montreal AI Ethics Institute
- Oxford Future of Humanity Institute
- Alan Turing Institute
- Algorithmic Justice League
- Data for Black Lives
- Resistance AI

Thank you!

Happy to answer any questions!

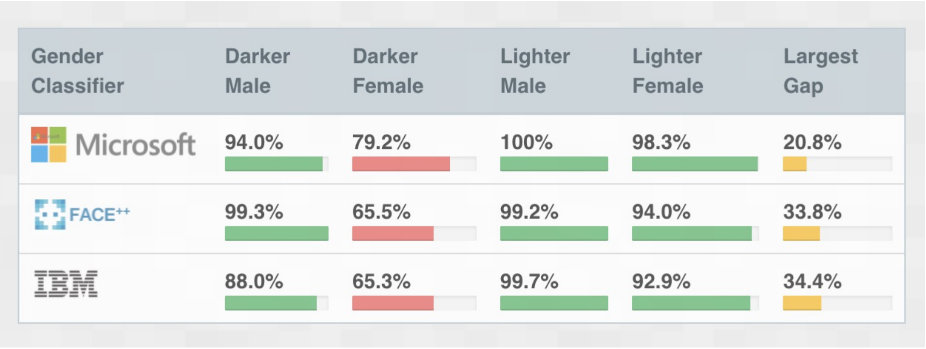
✉ sthais@princeton.edu 🐦@basicsciencesav

Appendix: Examples of Excellent AI Ethics Scholarship

Fighting Algorithmic Bias

- ML researchers measured the bias in several companies' commercial facial recognition algorithms
  - Some companies modified their algorithms or suspended facial recognition sales all together

- MIT Technology review put together an interactive analysis of the COMPAS algorithm evaluations
  - Demonstrates how it is impossible to balance equal score thresholding with equal outcomes across races



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

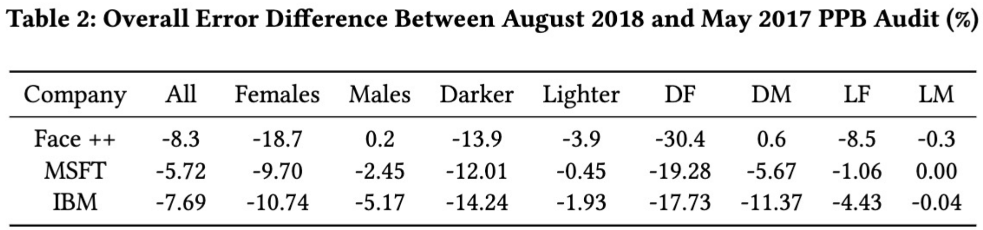
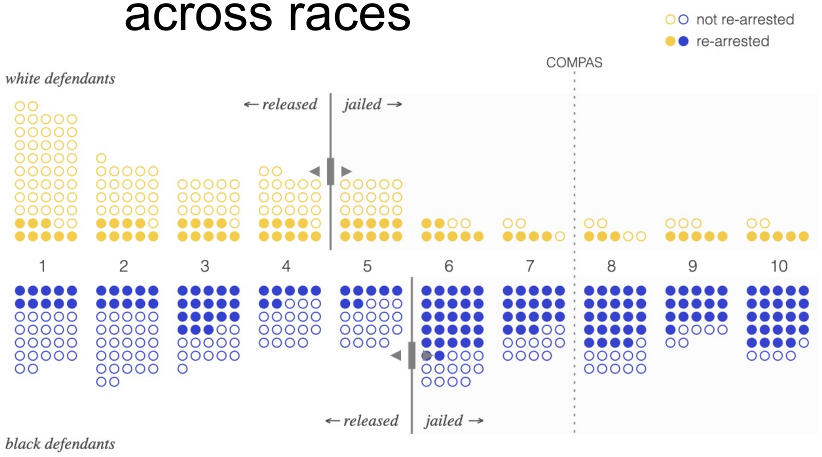
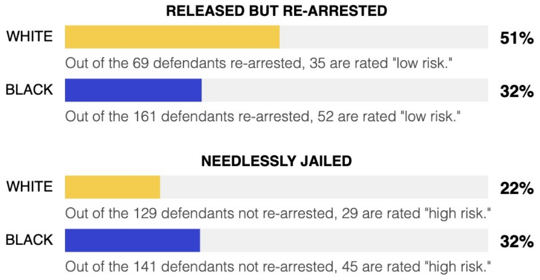


Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)

Company	All	Females	Males	Darker	Lighter	DF	DM	LF	LM
Face ++	-8.3	-18.7	0.2	-13.9	-3.9	-30.4	0.6	-8.5	-0.3
MSFT	-5.72	-9.70	-2.45	-12.01	-0.45	-19.28	-5.67	-1.06	0.00
IBM	-7.69	-10.74	-5.17	-14.24	-1.93	-17.73	-11.37	-4.43	-0.04





RELEASED BUT RE-ARRESTED

WHITE 51%

Out of the 69 defendants re-arrested, 35 are rated "low risk."

BLACK 32%

Out of the 161 defendants re-arrested, 52 are rated "low risk."

NEEDLESSLY JAILED

WHITE 22%

Out of the 129 defendants not re-arrested, 29 are rated "high risk."

BLACK 32%

Out of the 141 defendants not re-arrested, 45 are rated "high risk."

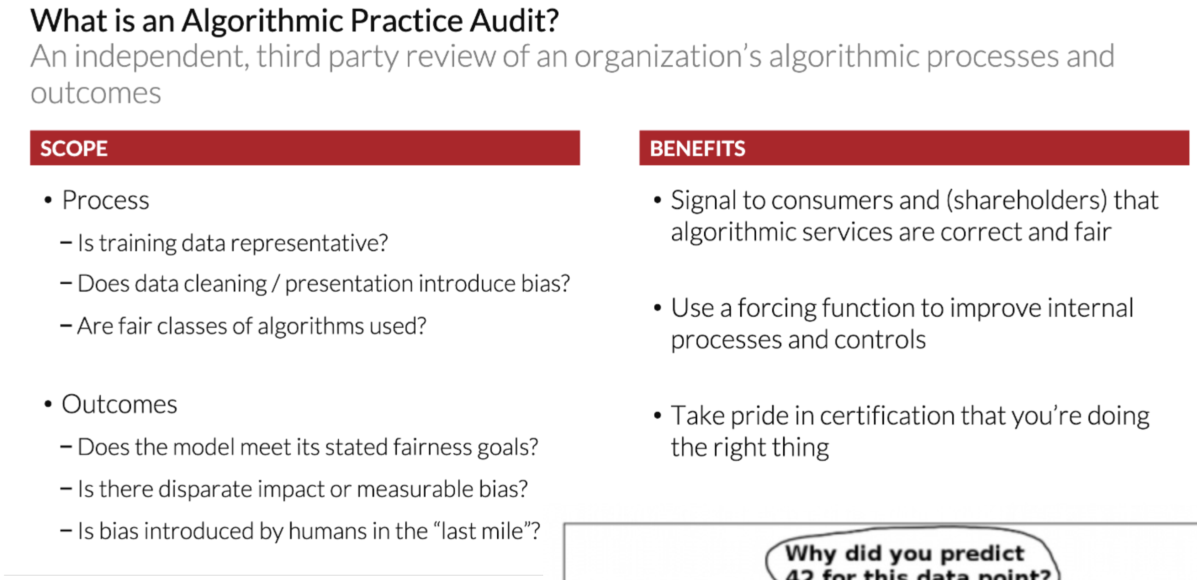
Ethical and Legal Frameworks

The rate of innovation has far outpaced the rate of regulation

- Biometric data: exploring bodily autonomy law as a framework for regulating biometric data collection, storage, and algorithmic use
- Whistleblowing: extending existing legal protections to AI ethics research in industry and academia
- Research standards/corporate funding: advocating for transparency in research funding and publication approval processes

Interpretability and Transparency

Researchers are exploring tools for increasing transparency and mathematical methods for interpreting models



What is an Algorithmic Practice Audit?

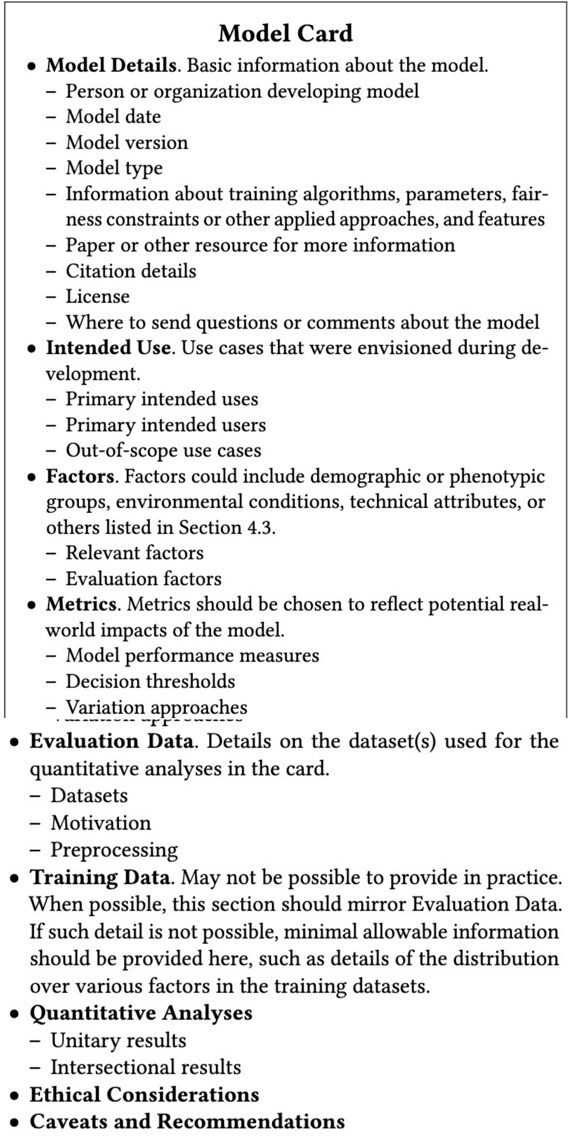
An independent, third party review of an organization's algorithmic processes and outcomes

SCOPE

- Process
  - Is training data representative?
  - Does data cleaning / presentation introduce bias?
  - Are fair classes of algorithms used?
- Outcomes
  - Does the model meet its stated fairness goals?
  - Is there disparate impact or measurable bias?
  - Is bias introduced by humans in the "last mile"?

BENEFITS

- Signal to consumers and (shareholders) that algorithmic services are correct and fair
- Use a forcing function to improve internal processes and controls
- Take pride in certification that you're doing the right thing



Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Collective Action