ETHICAL DATA SCIENCE FOR PHYSICISTS

Dr. Savannah Thais

Columbia University

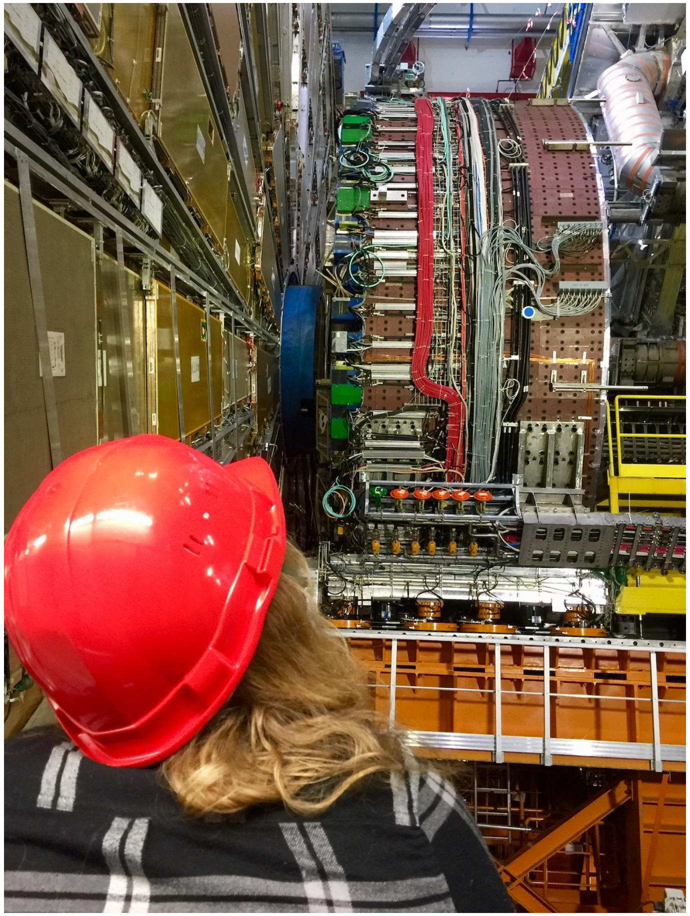
DSECOP Seminar

12/06/2022

A Quick Intro

Research faculty at the Columbia University Data Science Institute, Founder/Research Director of Community Insight and Impact

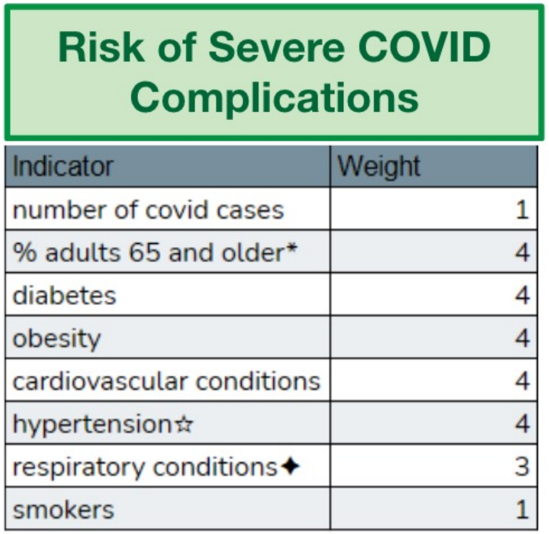
- Academic background:
  - Undergrad in math and physics at UChicago
  - PhD in physics at Yale on ATLAS experiment focusing on VH,H->$\tau\tau$ analysis, electron ID, and computer vision
  - Postdoc at Princeton with IRIS-HEP focusing on GDL, tracking, and AI Ethics
- Current research in several directions:
  - Physics informed machine learning: how do incorporate domain knowledge into ML systems
  - Algorithmic interpretability: how can we understand and use what a system is learning
  - Complex system modeling: how can we represent and quantitatively study cities, health systems, political systems, social networks, etc
  - Contextualizing ML systems and research: data collection practices, AI regulation, training incentives, deployment, etc



Physics is Model Building

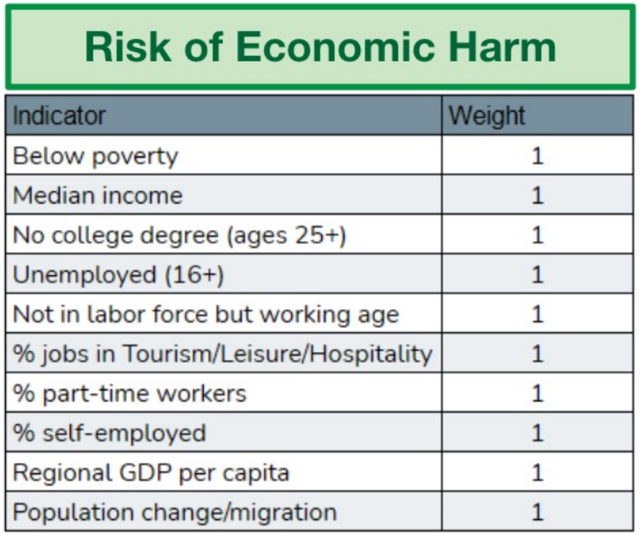
The COVID Vulnerability Metrics

Our first project was developing a complimentary suite of vulnerability measures to enable service organizations and community members to quantitatively understand community needs and effectively allocate resources



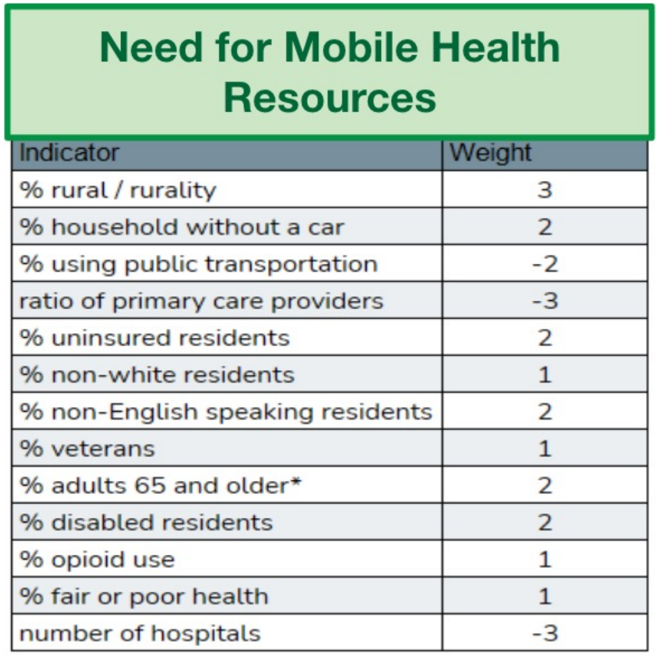
Risk of Severe COVID Complications

Indicator	Weight
number of covid cases	1
% adults 65 and older*	4
diabetes	4
obesity	4
cardiovascular conditions	4
hypertension☆	4
respiratory conditions✦	3
smokers	1



Risk of Economic Harm

Indicator	Weight
Below poverty	1
Median income	1
No college degree (ages 25+)	1
Unemployed (16+)	1
Not in labor force but working age	1
% jobs in Tourism/Leisure/Hospitality	1
% part-time workers	1
% self-employed	1
Regional GDP per capita	1
Population change/migration	1



Need for Mobile Health Resources

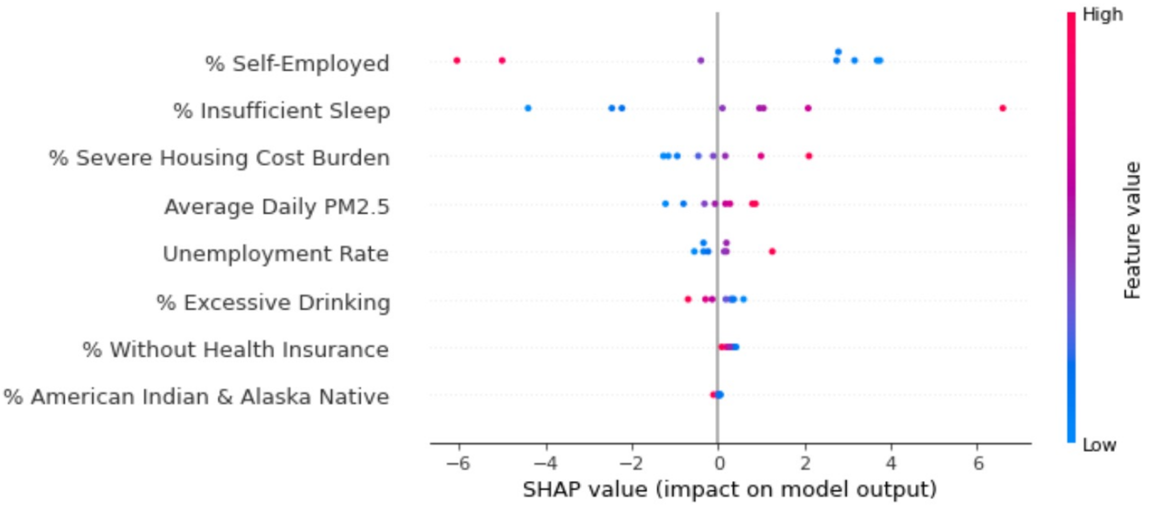
Indicator	Weight
% rural / rurality	3
% household without a car	2
% using public transportation	-2
ratio of primary care providers	-3
% uninsured residents	2
% non-white residents	1
% non-English speaking residents	2
% veterans	1
% adults 65 and older*	2
% disabled residents	2
% opioid use	1
% fair or poor health	1
number of hospitals	-3

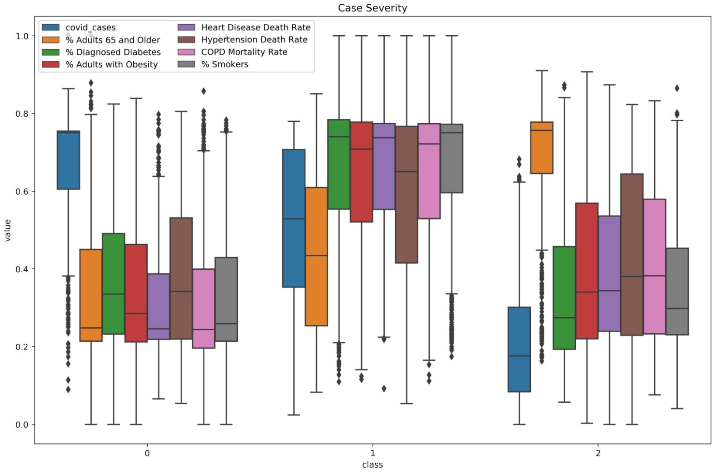
Similar to in particle physics research, **we are trying to make valid statistical inferences about phenomena we cannot directly observe**. We must utilize domain knowledge, model building, data collection/cleaning, and model validation.

Validation + Impact Studies

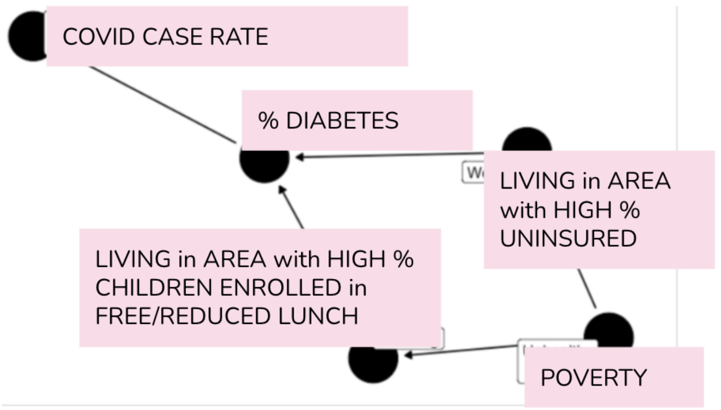
Published two papers seeking to validate our metrics and uncover implications for policy and program design

1. Define proxy outcomes for each metric, build a supervised ML model, study feature importances
2. Cluster communities based on metric variables, develop longitudinal dataset of variables, and compare outcomes to understand how communities build resilience





Validation + Impact Studies



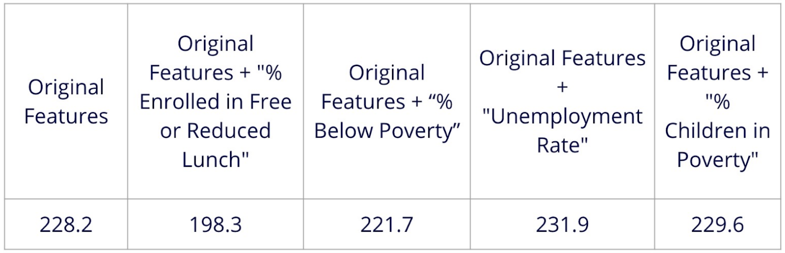
COVID CASE RATE

% DIABETES

LIVING in AREA with HIGH % UNINSURED

LIVING in AREA with HIGH % CHILDREN ENROLLED in FREE/REDUCED LUNCH

POVERTY



Original Features	Original Features + "% Enrolled in Free or Reduced Lunch"	Original Features + "% Below Poverty"	Original Features + "Unemployment Rate"	Original Features + "% Children in Poverty"
228.2	198.3	221.7	231.9	229.6

Key Insights:

- '% Enrolled in Free and Reduced Lunch' captures important health risks missed by traditional poverty measurements
- Improving college preparedness and investing in local community college infrastructure is key to crisis resilience
- Improving protections for part-time workers like pay parity and shift schedule control could strengthen local economies
- Mobile health resources combined with community structures could improve overall health in rural and low-income communities

But Models Don’t Exist in Isolation

AI/ML Systems Are Ubiquitous and Under Regulated

These types of models interface with nearly every aspect of our daily lives, often in opaque and uncontrolled ways that can have **life or death consequences**

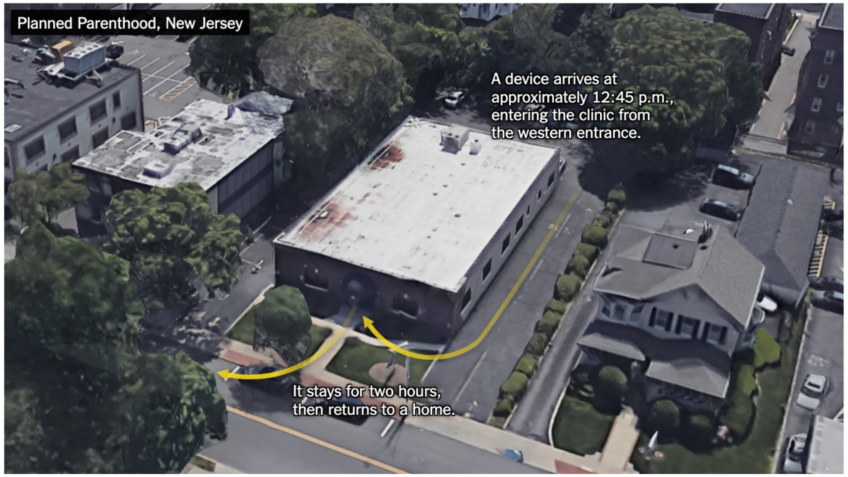
I focus on six key areas of AI Ethics:

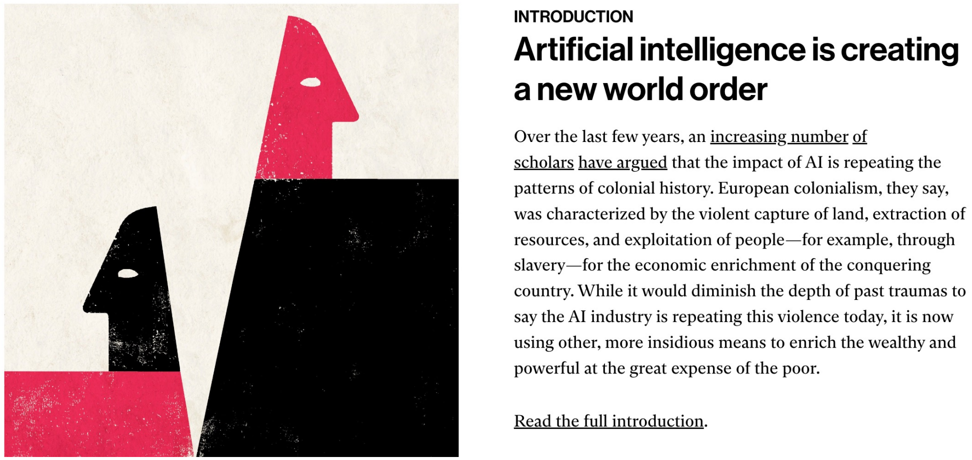
- Data collection and storage practices
- Task design and learning incentives
- Model bias and fairness
- Model robustness
- Equity in system deployment and outcomes
- Downstream and diffuse impacts

- Research, regulation, oversight, consent, community advocacy, and collective action are critical and should touch all of these topics

Data Collection, Storage and Sharing

- Data labeling companies (employed by many tech companies to create training datasets) exploit workers and political strife in the global south to maximize profits
  - Enable inhumane working conditions and enforced poverty
- Non-profit Crisis Text Line shared user conversation data with for-profit spinoff designed to ‘improve customer service’
- Data brokerage firms sell aggregated, ‘anonymized’ location datasets
  - Including datasets of individuals who visit Planned Parenthood
- Amazon requires delivery drivers to submit to biometric data tracking
  - Develops technology to surveil factories for signs of unionization organizing
- Ring worked directly with law enforcement to distribute devices and shares recordings without owners consent


Planned Parenthood, New Jersey
A device arrives at approximately 12:45 p.m., entering the clinic from the western entrance.
It stays for two hours, then returns to a home.



INTRODUCTION

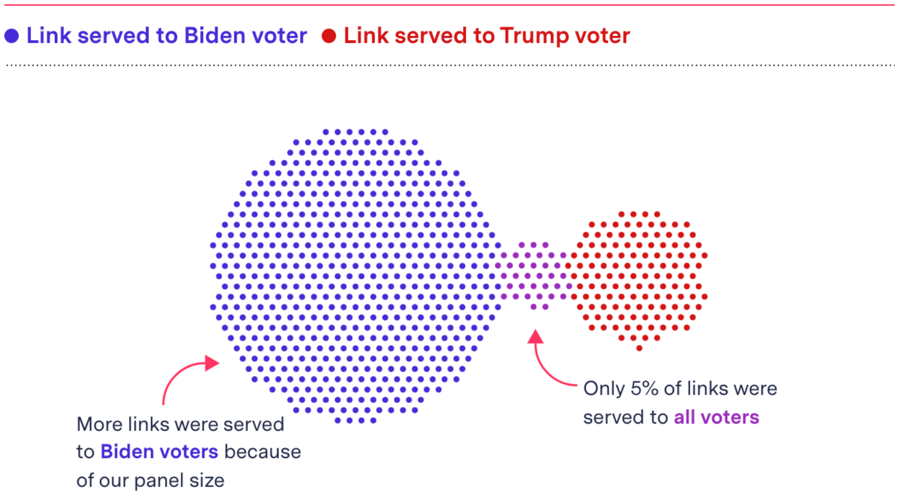
Artificial intelligence is creating a new world order

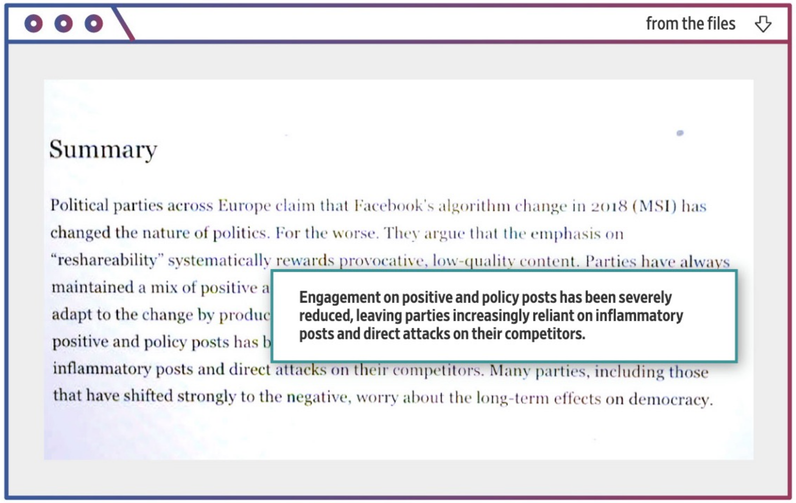
Over the last few years, an increasing number of scholars have argued that the impact of AI is repeating the patterns of colonial history. European colonialism, they say, was characterized by the violent capture of land, extraction of resources, and exploitation of people—for example, through slavery—for the economic enrichment of the conquering country. While it would diminish the depth of past traumas to say the AI industry is repeating this violence today, it is now using other, more insidious means to enrich the wealthy and powerful at the great expense of the poor.

Read the full introduction.

Misaligned Learning Goals

- Newsfeed/information curation algorithms are often designed with a primary goal of user retention and platform interaction, leading to 'unintended' behavior
  - Information silos based on click-through rates and shares
  - Radicalization pipelines through progressive content serving
  - Viral spread of misinformation is accelerated by algorithms
- Researchers and companies pursue learning goals like predicting faces from voices or predicting trustworthiness from a video
- Research on negative impacts of core technology often suppressed
  - See Facebook Files, Timnit Gebru firing, prevention of external research





Summary

Political parties across Europe claim that Facebook's algorithm change in 2018 (MSI) has
changed the nature of politics. For the worse. They argue that the emphasis on
"reshareability" systematically rewards provocative, low-quality content. Parties have always
maintained a mix of positive a
adapt to the change by produc
positive and policy posts has b
inflammatory posts and direct attacks on their competitors. Many parties, including those
that have shifted strongly to the negative, worry about the long-term effects on democracy.

Engagement on positive and policy posts has been severely reduced, leaving parties increasingly reliant on inflammatory posts and direct attacks on their competitors.

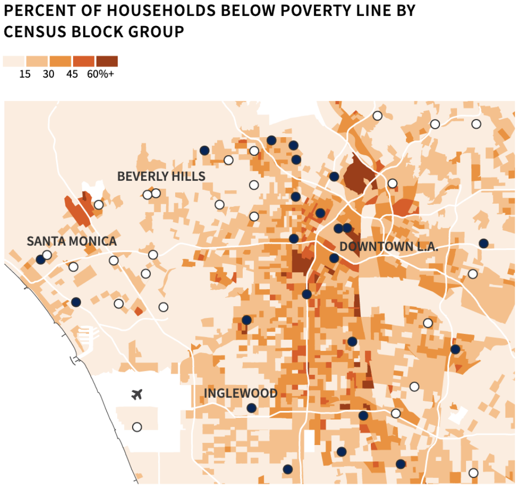
Equity in Deployment and Outcomes

- 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

A REUTERS INVESTIGATION

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

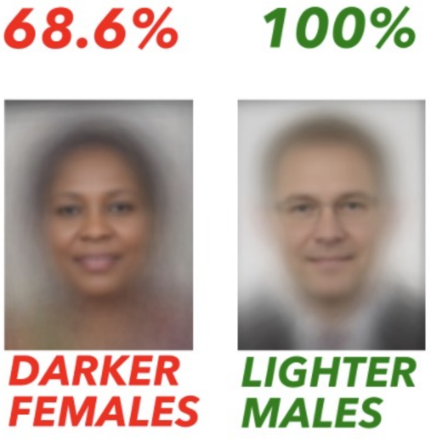
In the hearts of New York and metro Los Angeles, Rite Aid installed facial recognition technology in largely lower-income, non-white neighborhoods, Reuters found. Among the technology the U.S. retailer used: a state-of-the-
with links to China and its authoritari


PERCENT OF HOUSEHOLDS BELOW POVERTY LINE BY CENSUS BLOCK GROUP

- 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

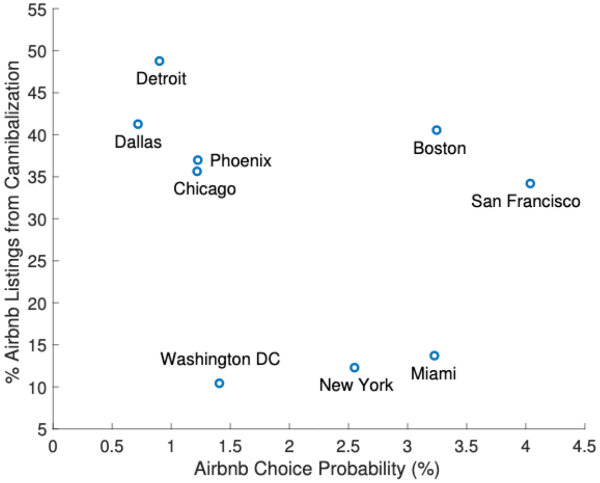
BIG CITY

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



Downstream and Diffuse Impacts

- Algorithmic management tools have enabled the creation of a heavily surveilled and often exploited gig labor class
  - Housing share services have priced residents out of historically diverse neighborhoods
  - Ride share apps and ghost stores, enabled by algorithmic optimization, have transformed industries and cities
- Homogeneity in data sources and models may lead to outsize impacts on marginalized individuals
- Focusing on technological solutionism may pull resources away from other types of recourse
  - E.g. if we focus on using traffic surveillance data to optimize routing and signaling, do we disincentivize investment in public transit?



What can we do from within physics?

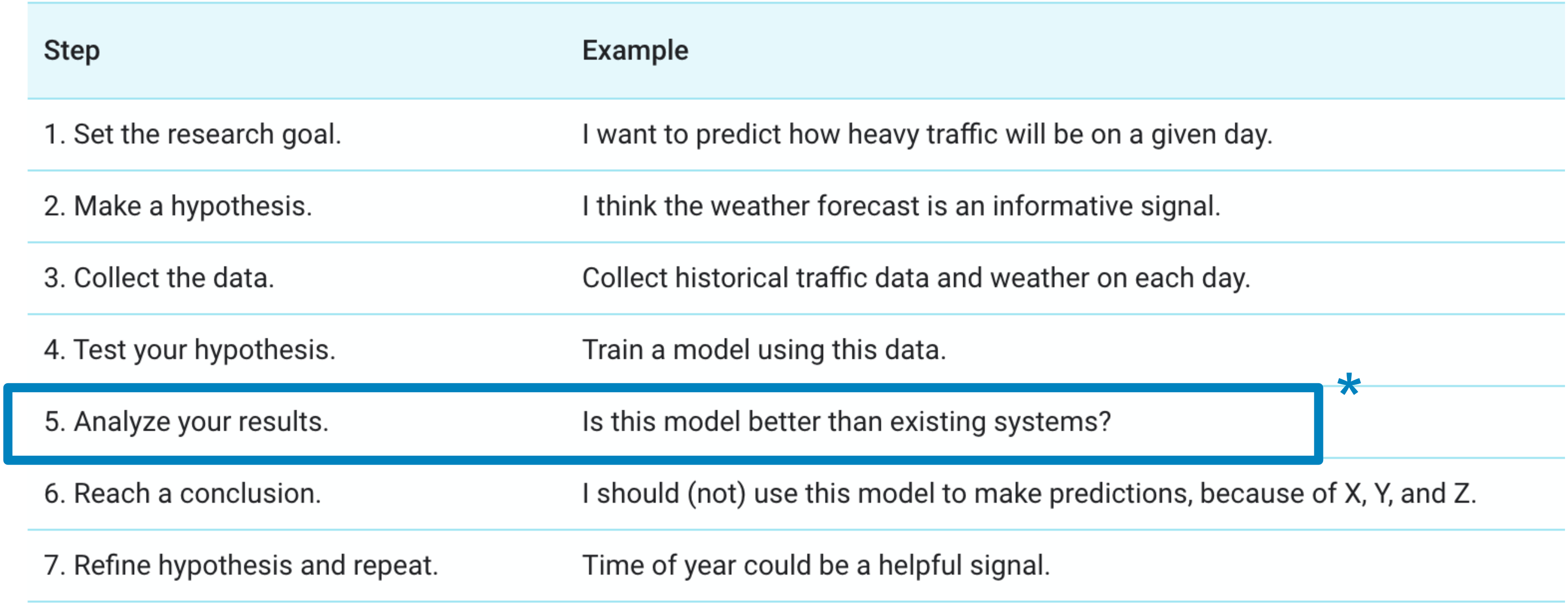
Think About the Context of Your Problems

- For HEP research:
  - Is my work well documented and reproducible?
  - Can this help us understand anything about the foundational principles of ML?
  - What technology transfer could happen?
- For industry collaborations or side projects:
  - Where is my data coming from? How is it collected and stored?
    - How can I guarantee minimally invasive data collection?
  - Is there a more transparent or 'safe' way to do this? Is a technological solution needed for this problem?
  - Where could bias enter the dataset or model performance?
  - What guarantees can I provide on model performance?
    - could include explainability
  - How will the systems I'm developing be deployed? Will the benefits and harms be equitably distributed?

Do the answers to these questions align with your personal code of ethics?

Treat ML and Data Science Scientifically

- 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

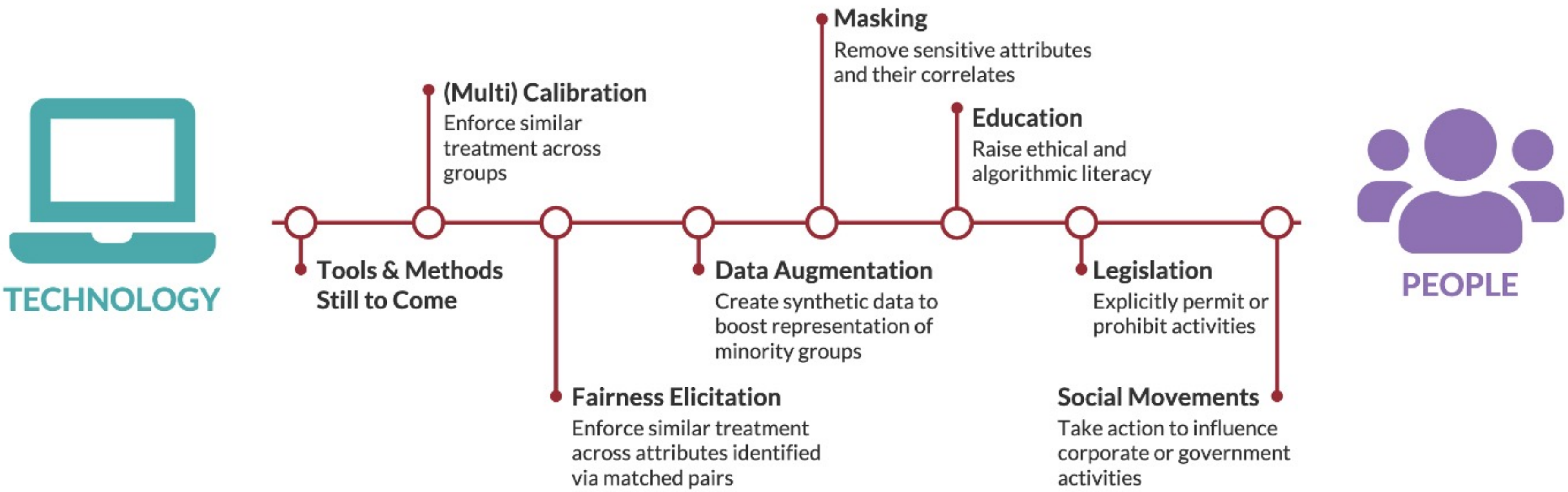
Use Physics to Inform ML

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

- Explainability: we know some information a model should learn and have interpretable bases for some problem classes
- Physics of ML: by studying learning as a stochastic process we can optimize models and training
- De-biasing: we often know true confounding variables and correlations so can meaningfully evaluate debiasing techniques
- Scientific principles: core experiment design techniques like uncertainty quantification and blinding can lend robustness to other domain applications

Just One Piece of the Puzzle

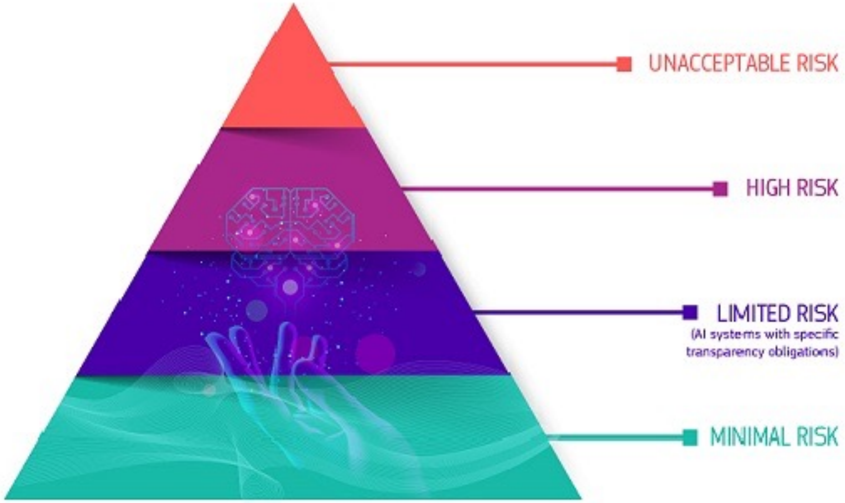
These are not purely mathematical problems and we need many different methods (and people) to address them


TECHNOLOGY
Tools & Methods Still to Come
(Multi) Calibration
Enforce similar treatment across groups
Fairness Elicitation
Enforce similar treatment across attributes identified via matched pairs
Data Augmentation
Create synthetic data to boost representation of minority groups
Masking
Remove sensitive attributes and their correlates
Education
Raise ethical and algorithmic literacy
Legislation
Explicitly permit or prohibit activities
Social Movements
Take action to influence corporate or government activities
PEOPLE

Excellent full talk!

Outreach

- Technical literacy: work with your communities to help them develop the knowledge necessary meaningfully consent to sociotechnical systems and understand possible recourses
- Advocacy: use your voice, institutional power, and collective action to work against unjust or unsafe uses of AI
- Legislation: share your scientific expertise with policy makers and champion meaningful regulations



Data analysis and model building are big responsibilities

✉ st3565@columbia.edu 🐦 @basicsciencesav

Some AI Ethics and Physics Efforts

- Snowmass LOI "Ethical implications for computational research and the roles of scientists"
- Working on full White Paper on Ethics in Computing for physicists
  - With co-authors Brian Nord and Aishik Ghosh and input from AI ethics researchers
- 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

Some Great AI Ethics Resources

- AI Now
- Alan Turing Institute
- Algorithmic Justice League
- Berkman Klien Center
- Center for Democracy and Technology
- Data & Society
- Data for Black Lives
- Montreal AI Ethics Institute
- Stanford Center for Human-Centered AI
- The Surveillance Technology Oversight Project
- Radical AI Network
- Resistance AI