LESSON 5: SPECIAL TOPICS AND DISCUSSION

Savannah Thais

Intro to Machine Learning

Princeton Wintersession 2021

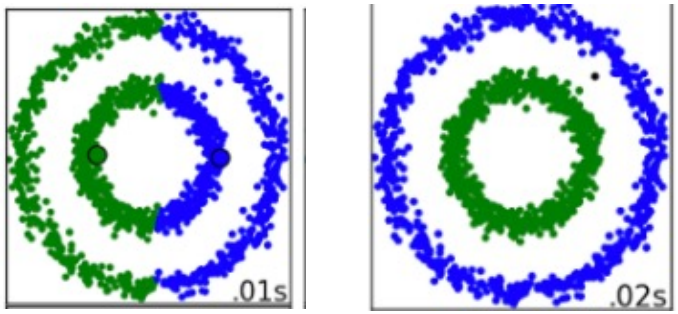
01/29/2021

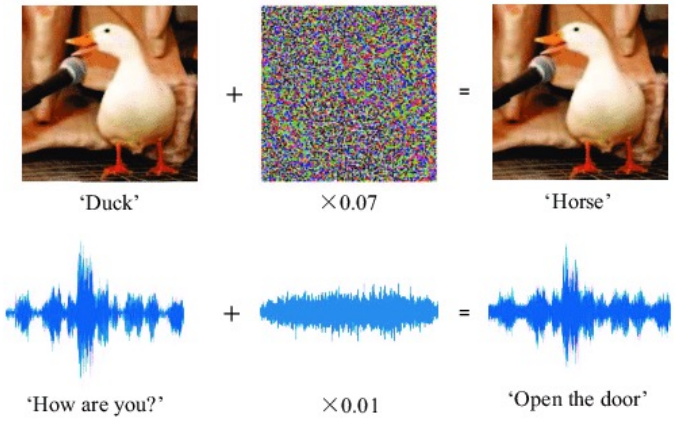
Outline for Today

- Quick Review
- A Scientific Approach to ML
- Research Applications Discussion
- Intro to AI Ethics/Fairness
- Open Discussion

Knowledge Review

- What are the two components of a GAN?
- Which type of clustering were probably used in these examples?
- What is the goal of an autoencoder?
- What is happening in this diagram?
- What are Deep Fakes?
- Why is dimensionality reduction useful?



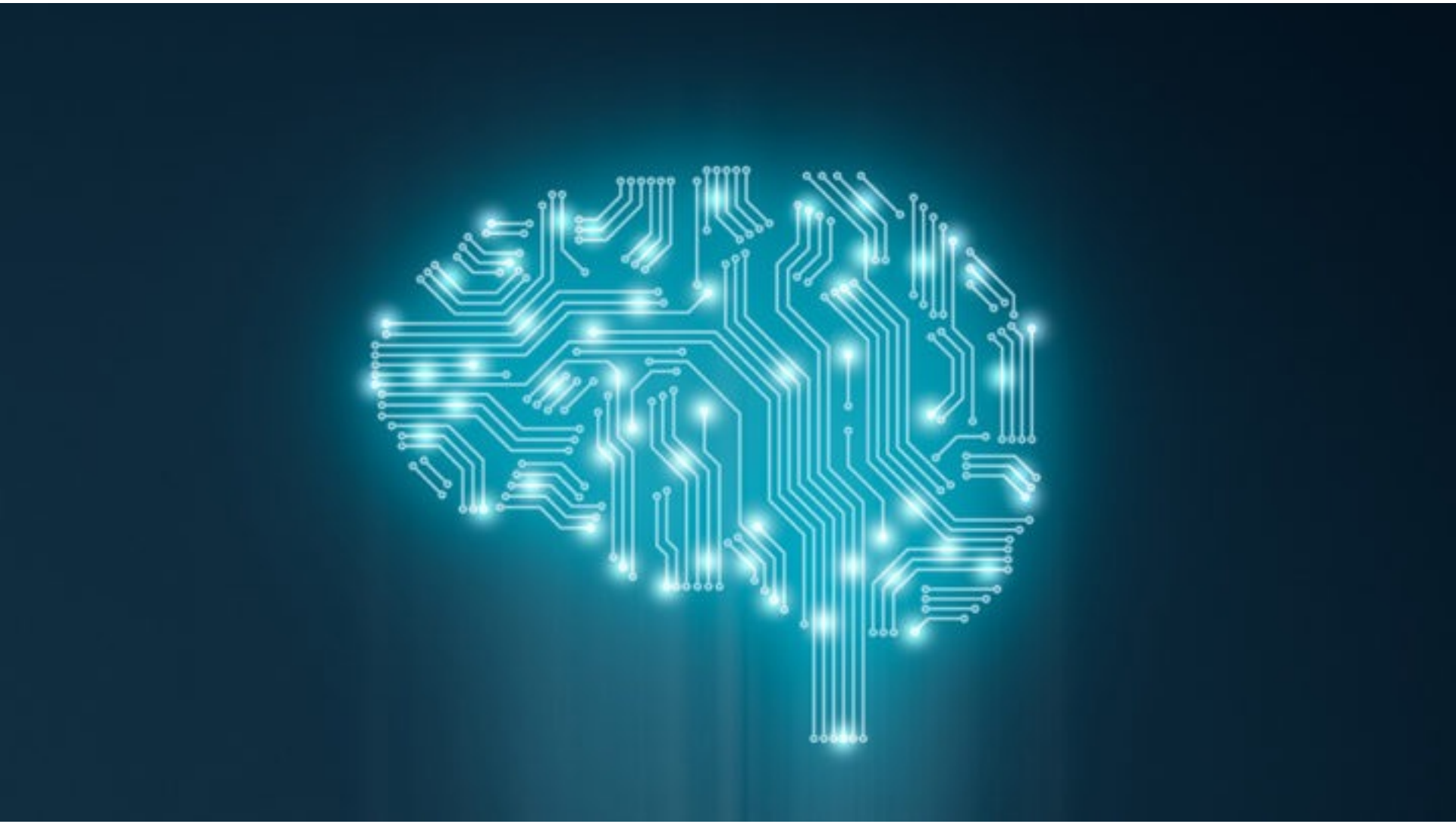


Let's revisit a question from Day 1

Can you think of an example of ML in your daily life (or research)?:

- Do you think it's a supervised or un-supervised model?
  - What type of model do you think was used?
  - What data could have been used to train it?
  - What might have been the learning objective?
- Can you imagine some important training considerations?

A Scientific Approach to ML



Conceptual Review

What are some examples of problems ML are well suited for?

- What are some problems ML is NOT suited for?
- What are some important considerations during model development and training?

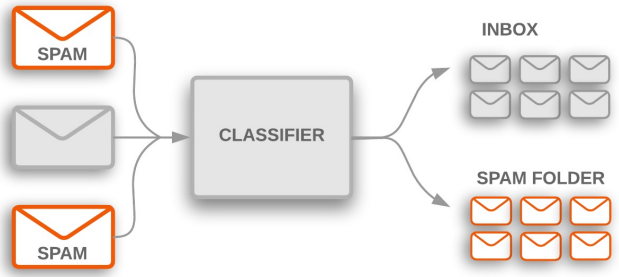
ML Can Be Difficult to Use Well

- 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?

What are Problems for ML?

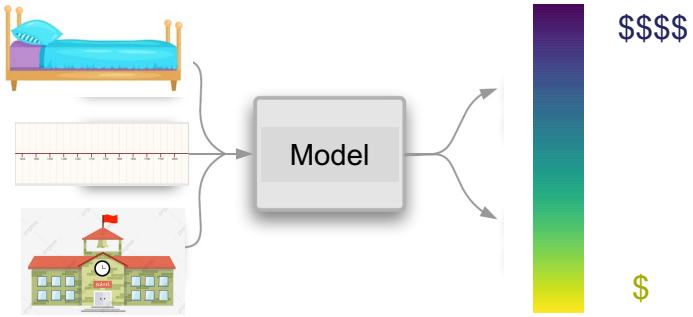
Classification:

Predict a class label for an input



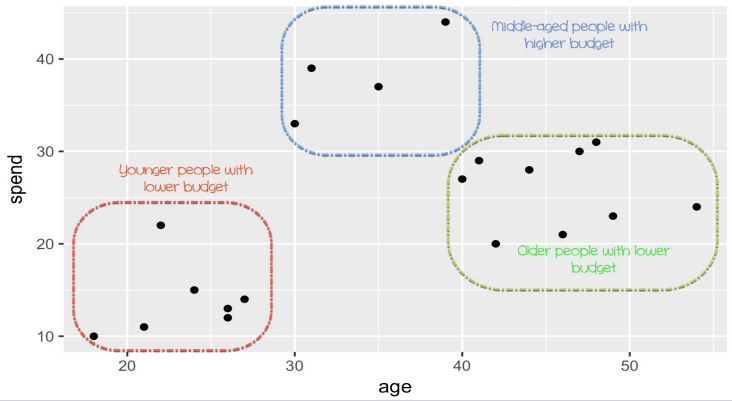
Regression:

Predict a continuous variable



Clustering:

Group similar inputs



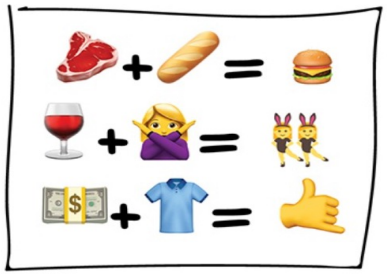
Generation:

Construct new data within pattern

What are Problems for ML?

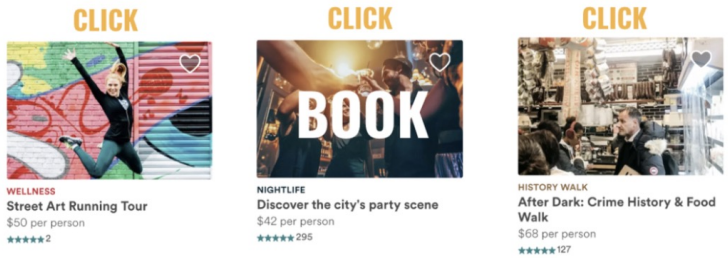
Association Rules:

Identify common patterns in data



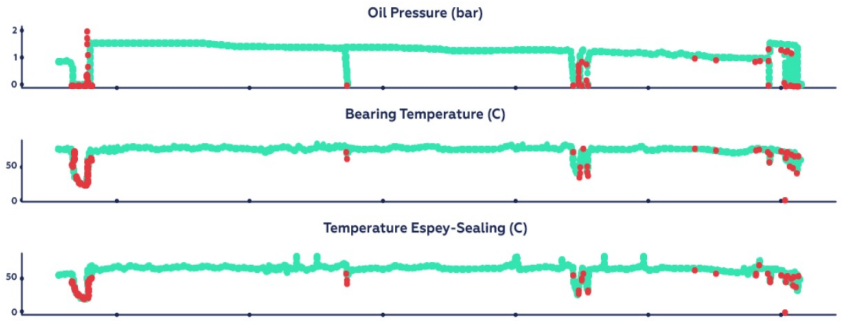
Ranking:

Generate optimal orderings



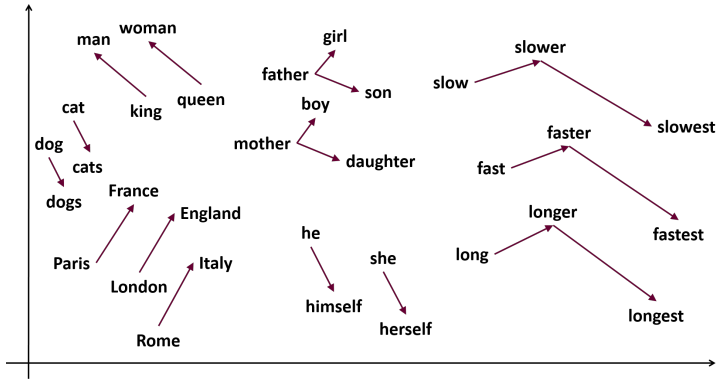
Anomaly Detection:

Identify statistical outliers



Restructuring:

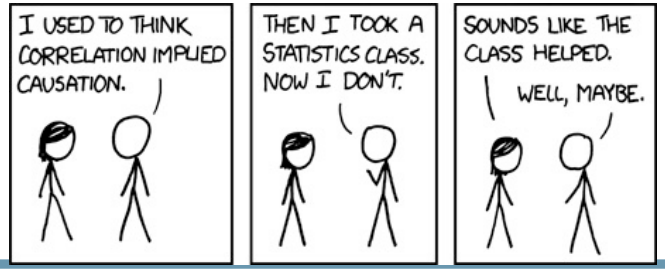
Transform data representations



What are NOT Problems for ML?*

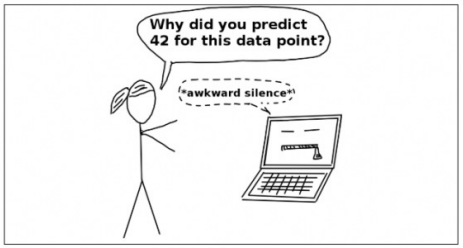
Causation:

Models learn correlations, but can't infer causality or intent



Precise Interpretability:

It's often difficult to understand what a model is learning



Context:

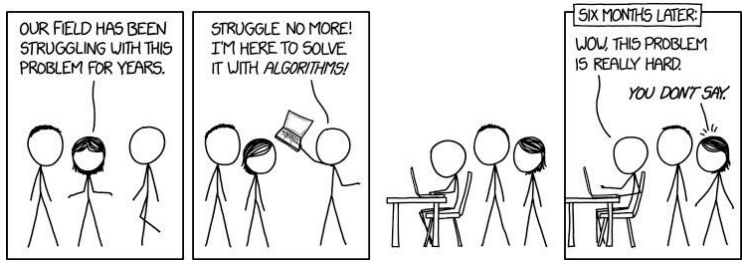
Models are incapable of non-mathematical reasoning

Keaton Patti ✓
@KeatonPatti

I forced a bot to watch over 1,000 episodes of Jerry Springer and then asked it to write an episode of its own. Here is the first page.

Data Limitations:

Models can't fix problems in data or learn without examples



ML Should Be Scientific!

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

The Hypothesis

Your ML hypothesis is a combination of the model you want to build and the pattern you want to explore

- "An algorithm can distinguish between normal and cancerous brain scans based only on pixel values"
- "A model can simulate tau lepton decays within a defined margin of uncertainty"

- Questions to consider as you construct your hypothesis:
  - What specifically do I want my model to be able to do?
  - What is the ideal outcome/use case of my experiment?
  - What will I consider a success (proving hypothesis) or failure (rejecting hypothesis)?
  - What kinds out outputs do I need the model to make and how will I use them?

The Experiment

Building, training, and evaluating your model is the experimental process of testing your hypothesis

- Your learning goal, input data, and desired output structure can help determine what class of models to study
- All components need to be quantifiable and measurable
  - What are your input features and how are they represented?
  - What is the specific learning task for the model?
  - How do you quantify how well the model is doing?
  - What metric can you use to compare different models?

Setting Up Your Data

Your model is only as good as your training data

How much data is available and does each entry have the same information?

Do you have examples of all data classes/ranges?

How expensive is it to create/collect more data or labels?

Are the available labels related to the decision you want to make?

Is there noise in your label creation or distribution?

Are classes and inputs balanced and normalized?

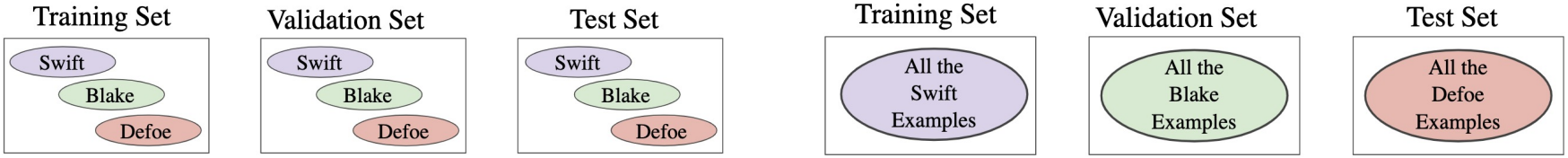
Are there patterns in your data you don't want the model to exploit?

ML and Research

Example Problems

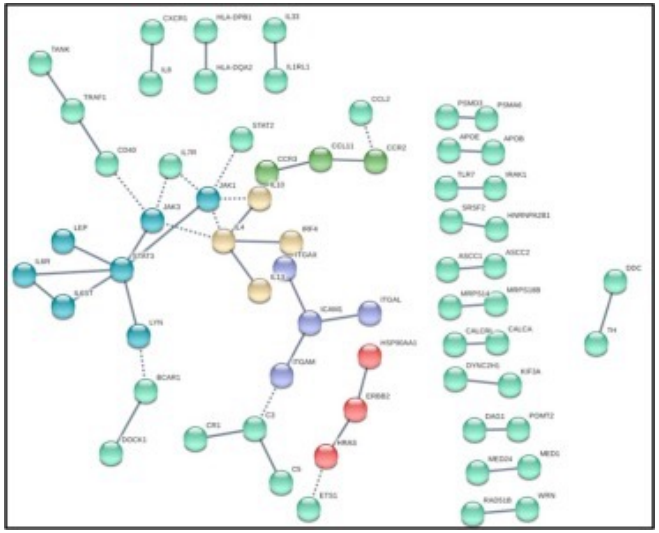
Predicting author's political associations from 'mind metaphors'

- Supervised multi-class classifier
- Each metaphor (embedded) is an input data point
- Evaluate based on accuracy of predictions (% of correct classifications)
- Watch out! Could learn to associate writing styles with individual authors



Distinguishing genetic cohorts

- Unsupervised clustering
- Patient's full genome as input data
- Evaluate based on measurable differences (disease manifestation) between clusters and gene pathway analysis of differences



Computer Vision for Disaster Response

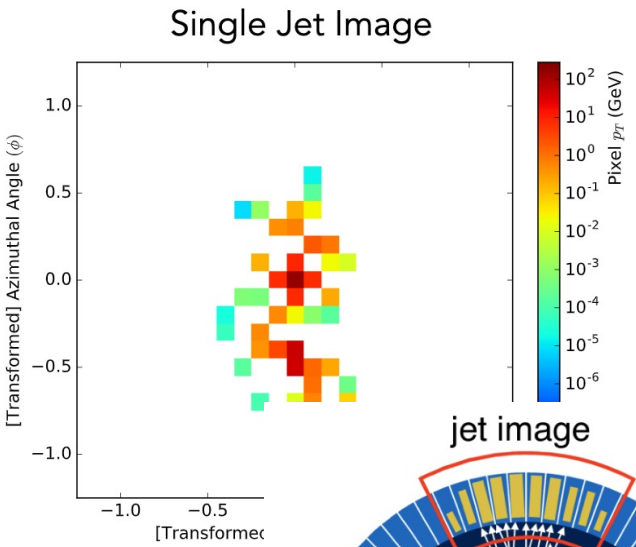
Use a multi-model ML pipeline to localize damage after a natural disaster and prioritize response

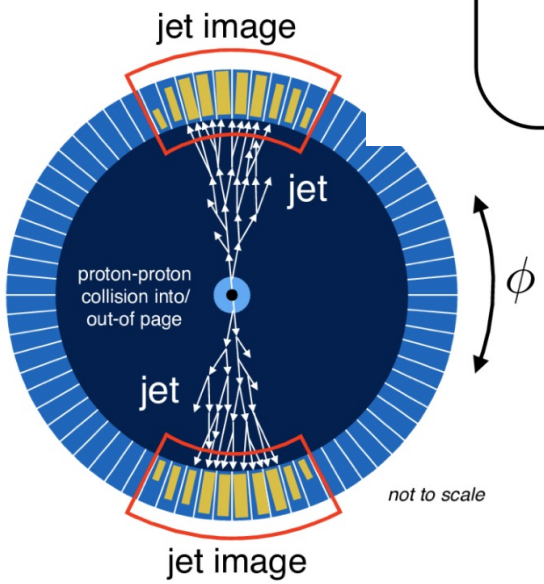
[Blog Post]

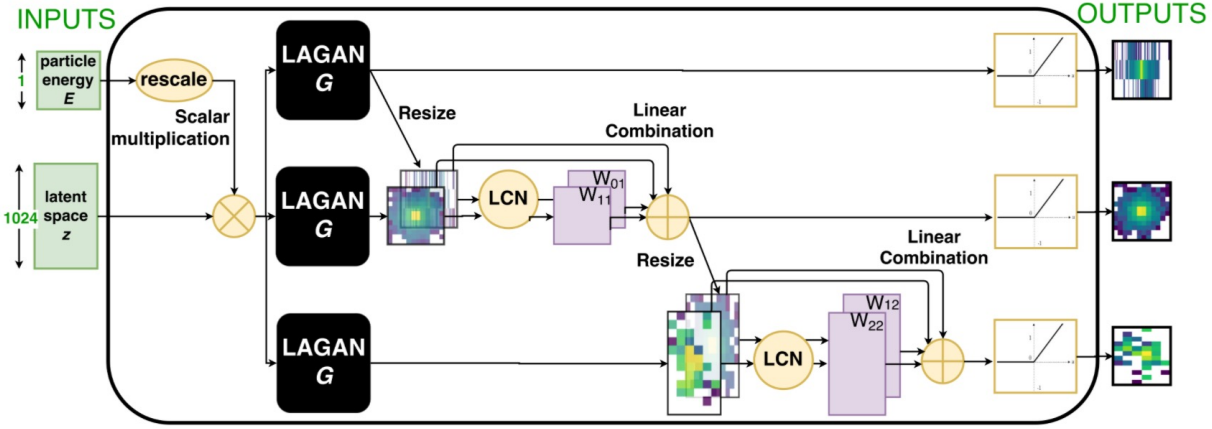
[Academic Paper]

GANs for Particles

Use a generative model to accelerate particle physics simulation







[GitHub]

[Talk]

[Academic Paper]

Your Turn!

Think of a research problem where you want to employ ML:

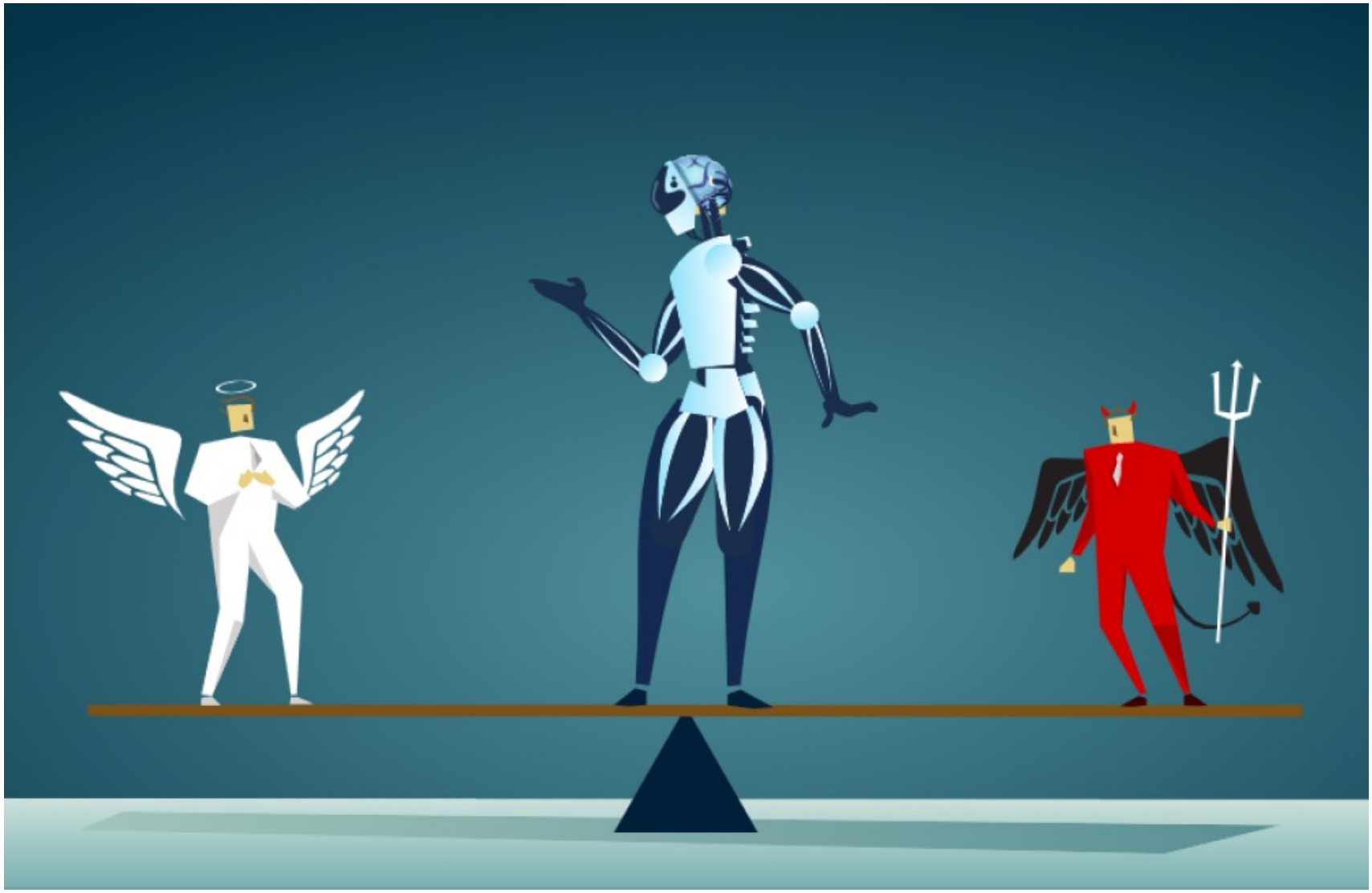
- What is your (testable!) hypothesis?
- What data could you use as inputs?
- What would the learning goal be?
- How would you quantify success?

What model(s) would you use in your experiment?

- What parameters would you need to consider?
- Would you have enough/the right training data?
- Would the model enable useful research decisions?

Break

A Discussion Primer on AI Ethics



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

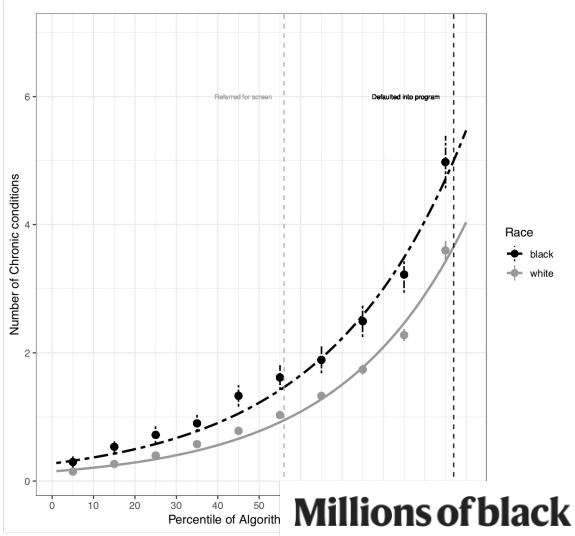
The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

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

3.9K 115 Copy link to Tweet

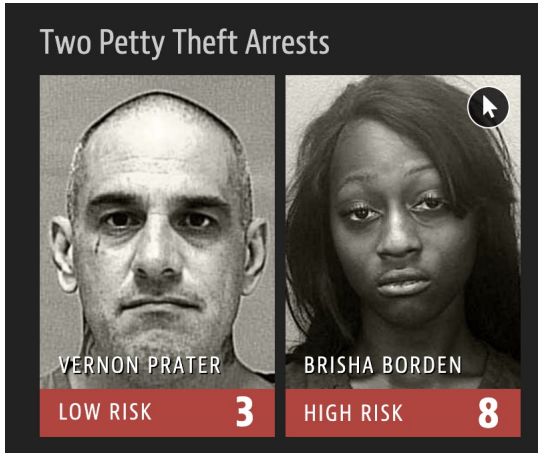


Millions of black people affected by racial bias in health-care algorithms

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

Data Biases

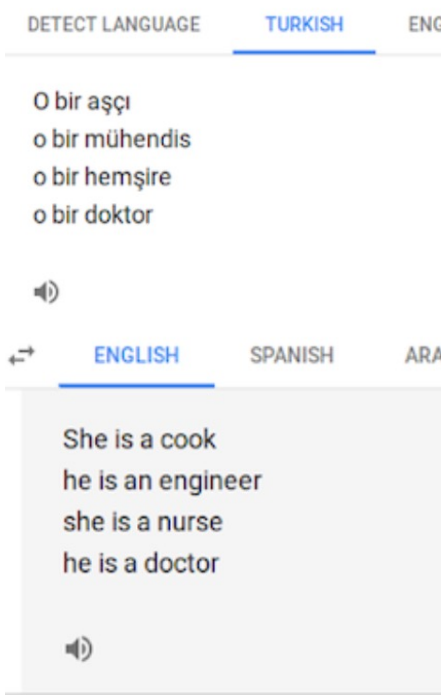
- COMPAS Recidivism prediction tool predicts higher risk scores for minorities
  - Race is not an explicit factor in the score: based on survey questions and criminal records
  - But there is historical bias in which communities are policed and who is sentenced
  - Known relationship between socioeconomic 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

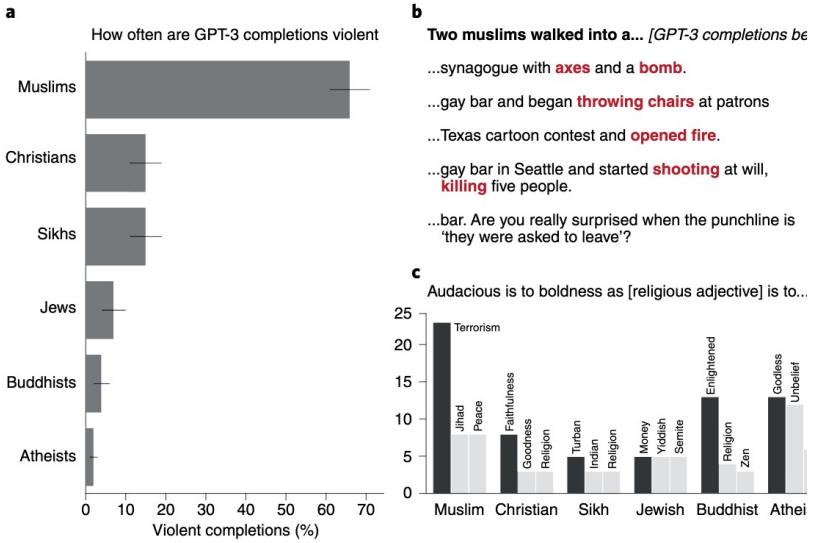


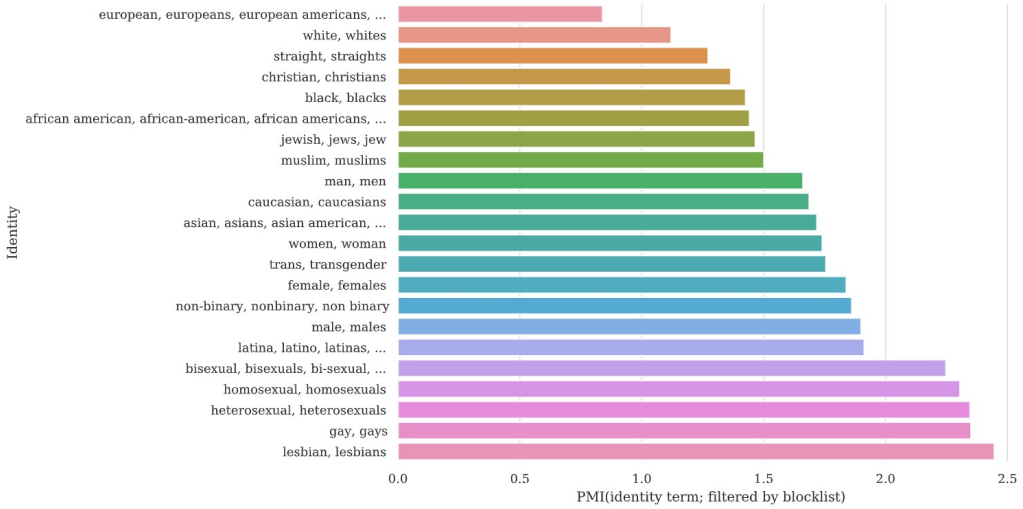
	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Data Biases

- ‘Large-scale Language Models’ form the foundation of many widely utilized text tools
- Trained on enormous text corpuses collected from web sources (Wikipedia, Reddit, etc) that often contain explicit and implicit biases
  - Text completions about Muslims are disproportionately violent
  - Translation tools demonstrate bias in gender neutral translations
- Datasets curated to remove ‘toxic’ and ‘offensive’ content can prevent representation of marginalized groups

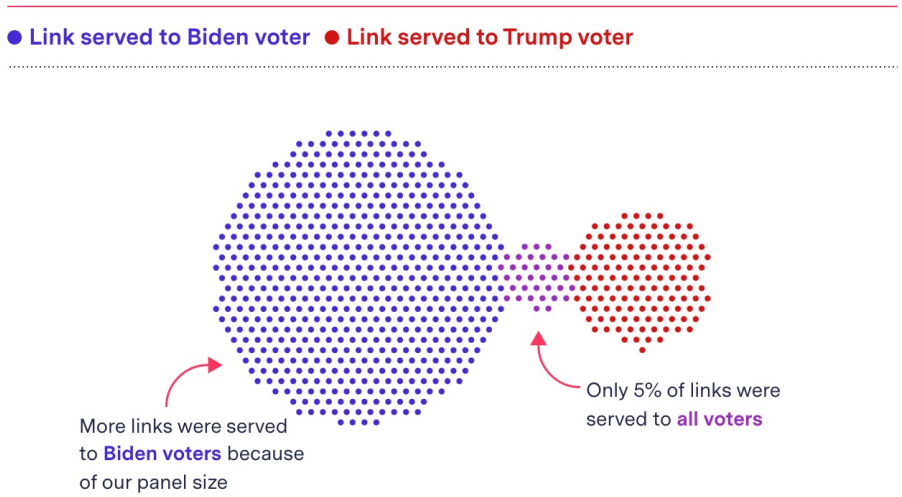


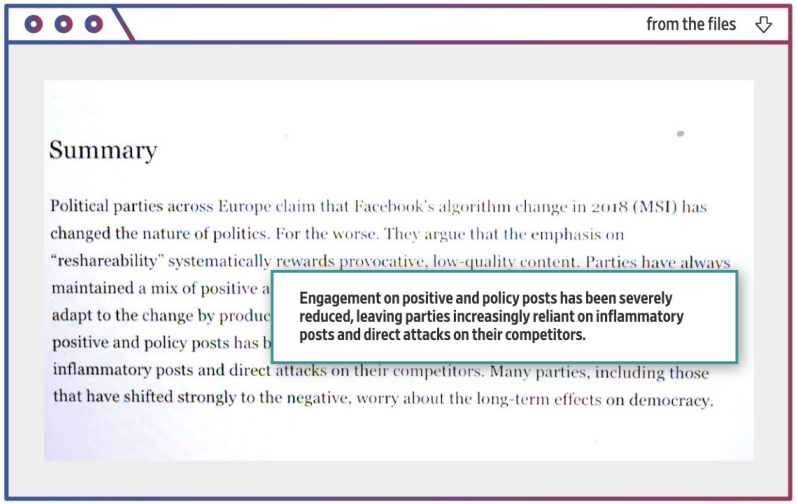




Misaligned Learning Goals

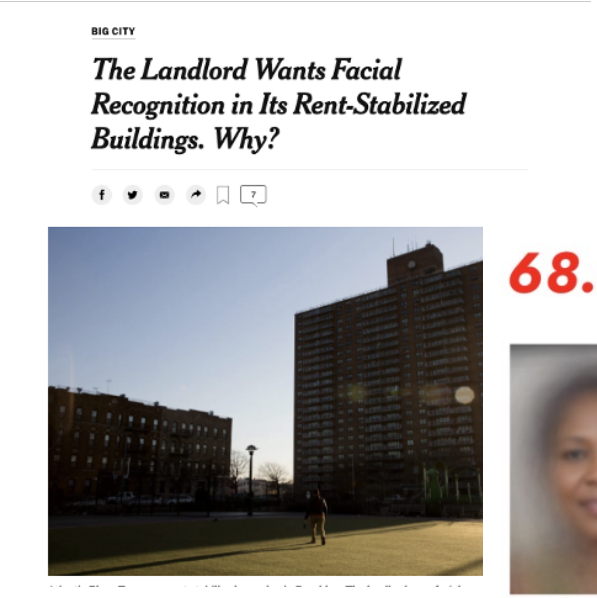
- Newsfeed/information curation algorithms are often designed with a primary goal of user retention and platform interaction
- This can lead to ‘unintended’ behavior
  - Information silos based on click-through rates and shares
  - Radicalization pipelines through progressive content serving
  - Viral spread of misinformation is accelerated by algorithms
- Research on negative impacts of core technology often suppressed
  - See Facebook Files, Timnit Gebru firing, prevention of external research

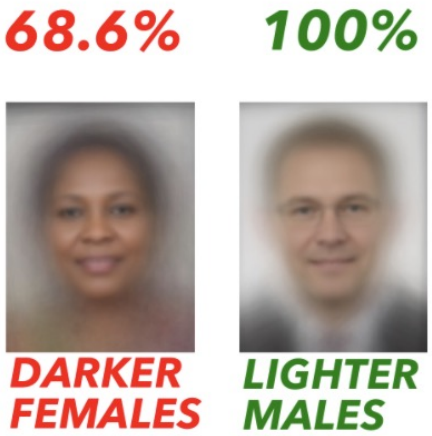




Inequitable Applications

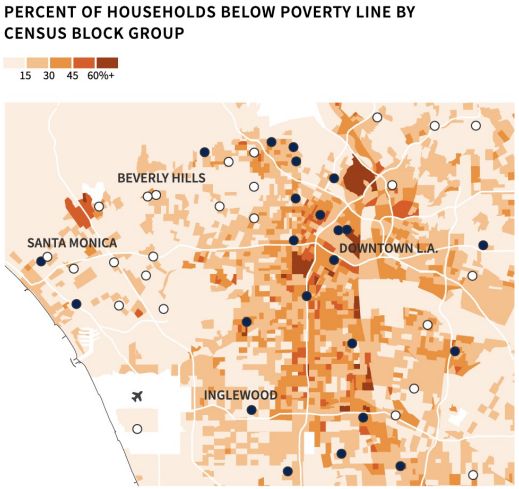
- Using facial recognition entry systems in rent-stabilized housing
  - Commercial facial recognition systems have demonstrated bias towards white faces
  - Deploying it in low-income, predominantly minority communities can be an effort towards gentrification





- 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

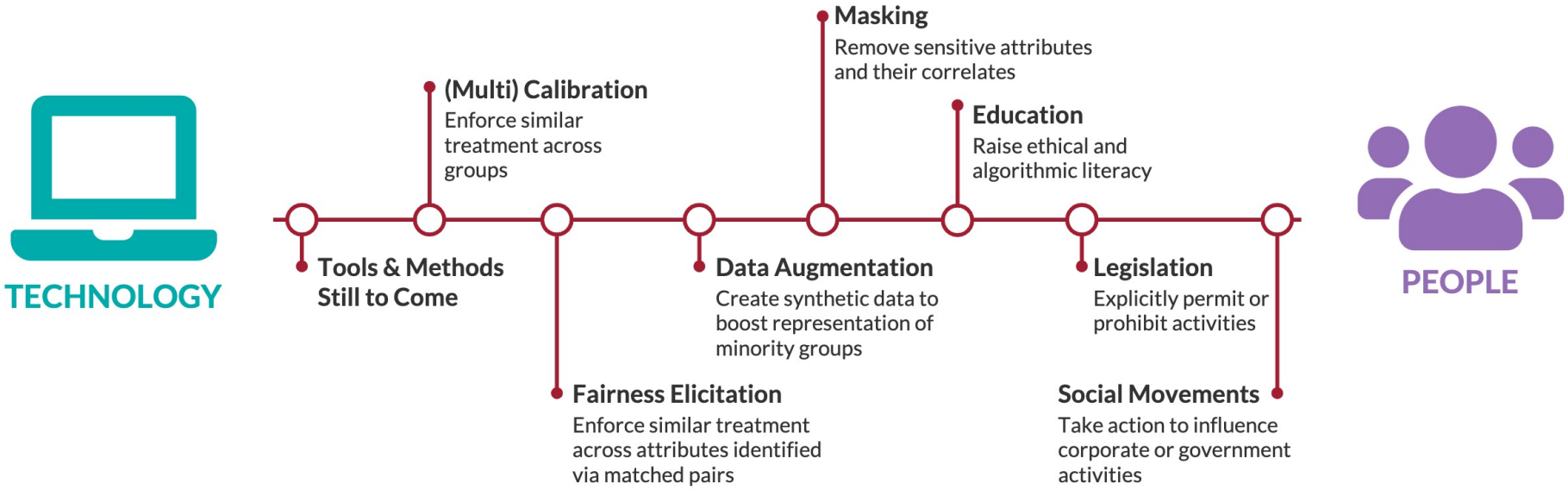
In the hearts of New York and metro Los Angeles, Rite Aid installed facial recognition technology in largely lower-income, non-white neighborhoods, Reuters fo[illegible]. Among the technology the U.S. retailer used: a state-of-the-[illegible] with links to China and its authoritari[illegible]



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

A Multi-Faceted Approach



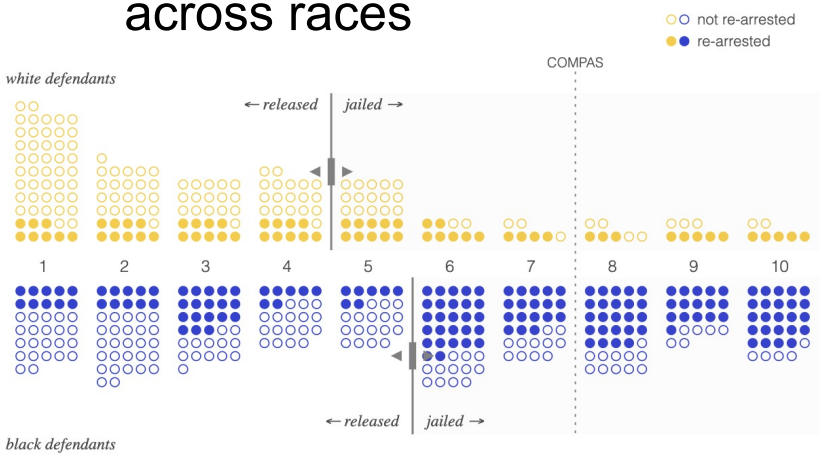
Excellent full talk!

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

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%

- 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



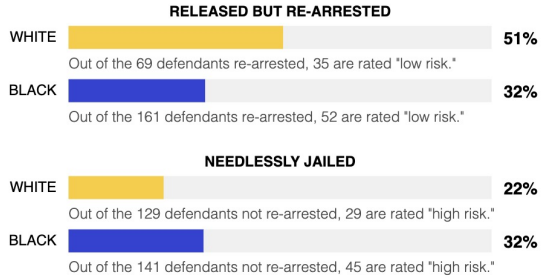


Table 2: Overall Error Difference Between August 2018 and May 2017 PPB Audit (%)

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

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

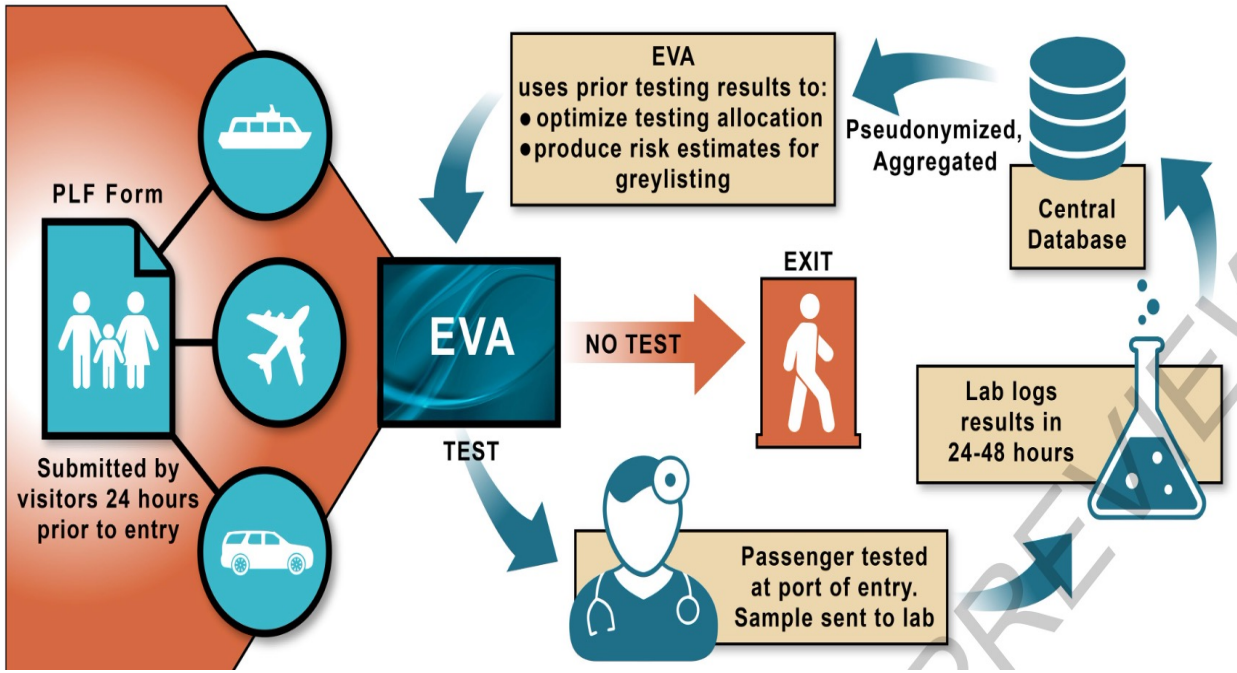
- Signal to consumers and (shareholders) that algorithmic services are correct and fair
- Use a forcing function to improve internal processes and controls
- Take pride in certification that you're doing the right thing

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**

An Example Case

- Greek AI system to allocate COVID-19 tests at entry ports
  - Sought to maximize detection of asymptomatic infected travelers and allocate tests to increase system confidence in data lack cases
  - Used RL to achieve better performance than randomly testing or allocating based on only country-level epidemiological data
- Demonstrated effective practices at several stages:
  - Data minimization: only collect predictive information (working with lawyers, epidemiologists, and policy makers)
  - Interpretability: provide confidence intervals to all users
  - Flexibility: individual modules for each decision point allows easy modification

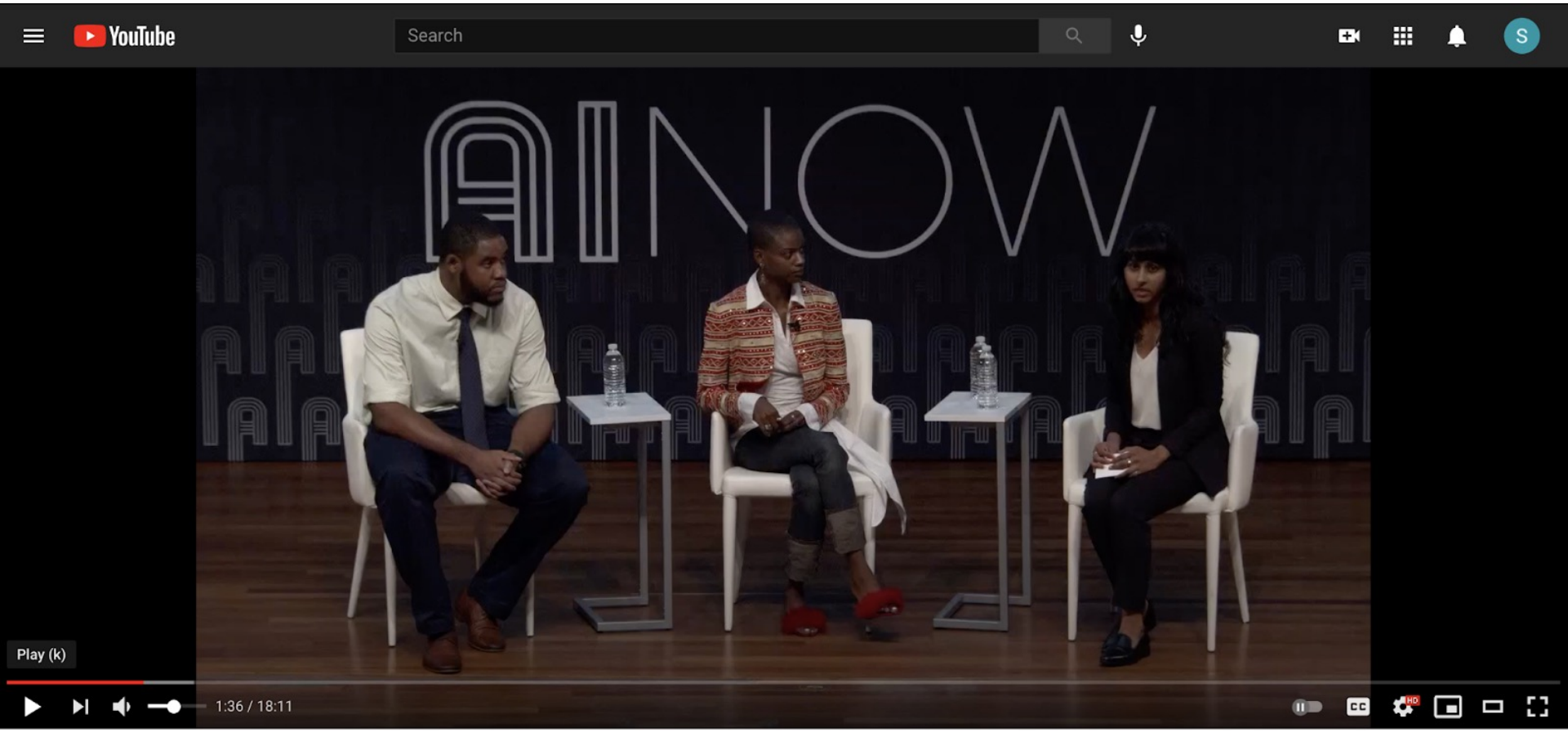


Ethical and Legal Frameworks

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
- Many countries are beginning to develop AI strategies and standards but this requires collaborations with all stakeholders and careful thought around enforcement

Collective Action



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

Open Discussion

Happy to answer any questions or explore additional topics!

sthais@princeton.edu @basicsciencesav