CIS 419/519

Ethics in Machine Learning

Instructor: Dinesh Jayaraman

Administrivia

- Next week:
  - Monday: Spring 2022 exam discussion and review
  - Wednesday: Fall 2022 exam discussion and review

Ethics and Fairness are Hard!

- Even outside the context of ML, fair and ethical decision-making is difficult
- The topic of thousands of years of debate in moral philosophy, law etc.
- Often in constant renegotiation with changing societal norms

ML systems, when used to rapidly scale decision-making, can amplify ethical concerns and pose new ones. e.g. who is responsible for a poor decision?

The Increasing Role of ML in the World

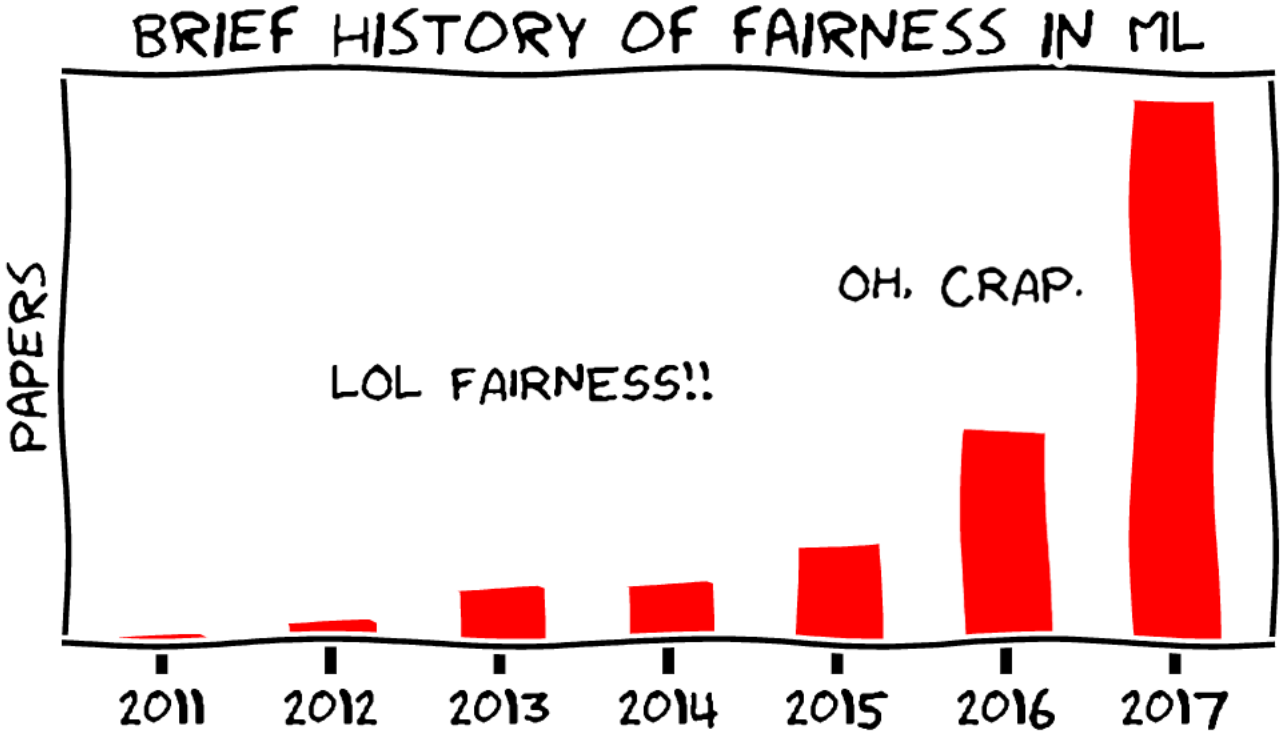
Setting credit scores and insurance premiums

Filtering resumes of job applicants

Making global aid allocation decisions from satellite imagery

Even seemingly less obviously important things like news recommendations, search engine results, ad targeting, can have a profound impact on what we see and how we think.

The Ethics of Machine Learning Systems



A growing community today focused on “FAT-ML” (Fairness, Accountability, and Transparency). Annual conference.

Picture: Moritz Hardt

ML Outputs Are Not Automatically “Objective”

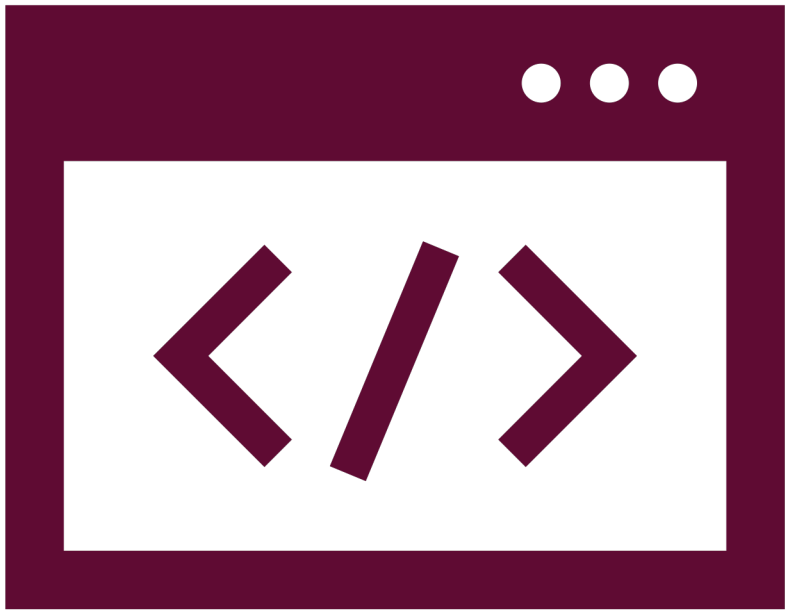
“[Reliably generalizing requires] … a sufficiently large number of examples to uncover subtle patterns; a sufficiently diverse set of examples to showcase the many different types of appearances that objects might take; a sufficiently well-annotated set of examples to furnish machine learning with reliable ground truth …”

Solon Barocas, Moritz Hardt, Arvind Narayanan, “Fairness and Machine Learning”

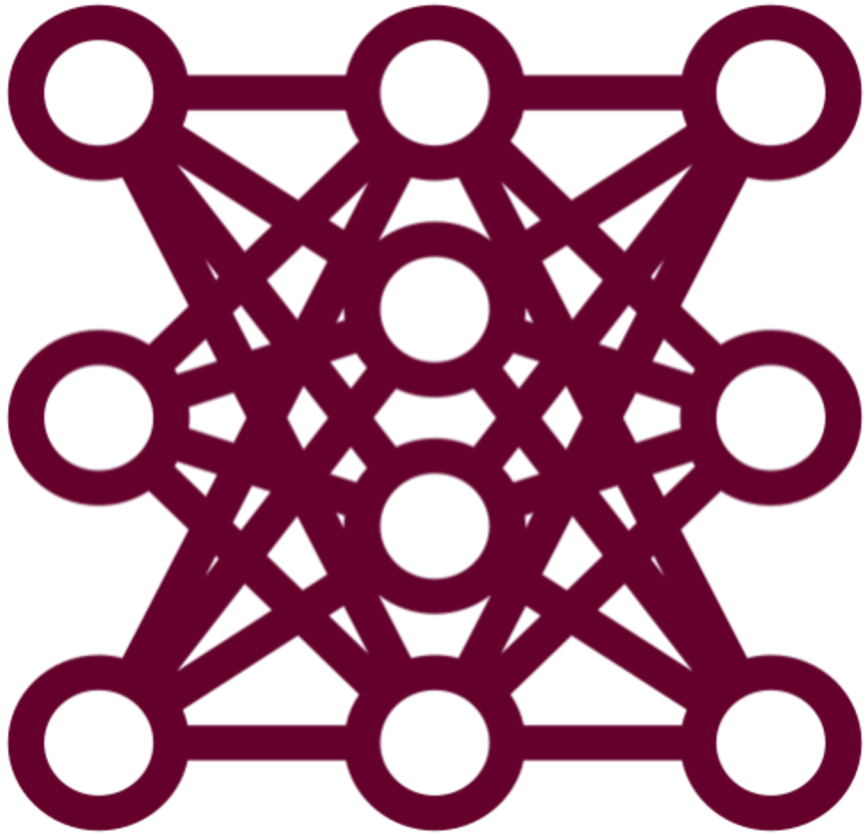
Your ML system is only as good as the data.

Machine learning as Programming 2.0

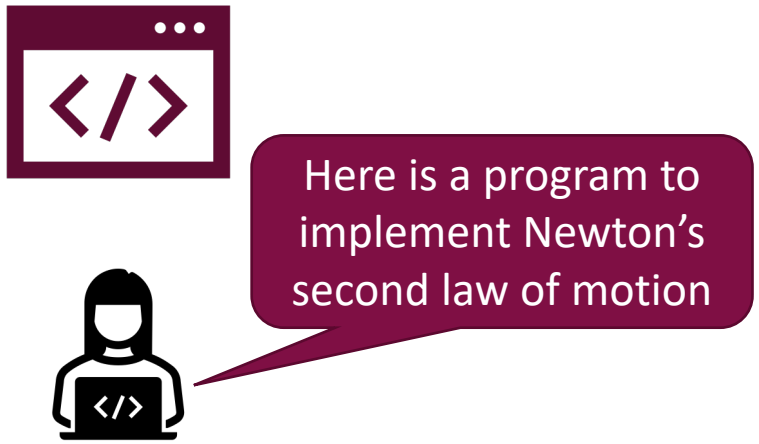
Traditional Programming

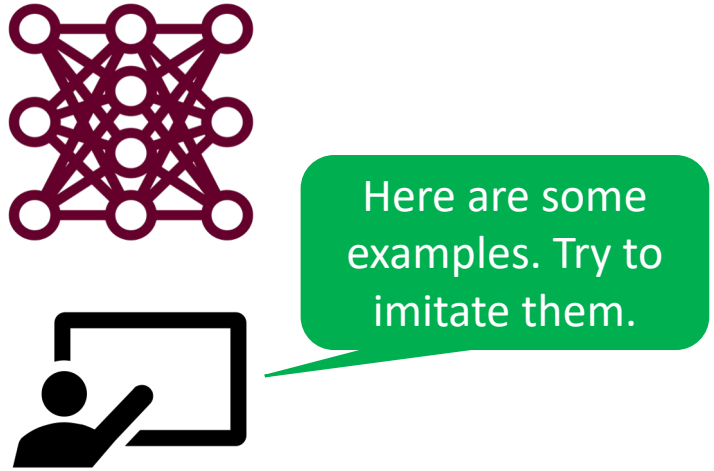


Machine learning (ML)



Task specification in ML: ~~programs~~ examples

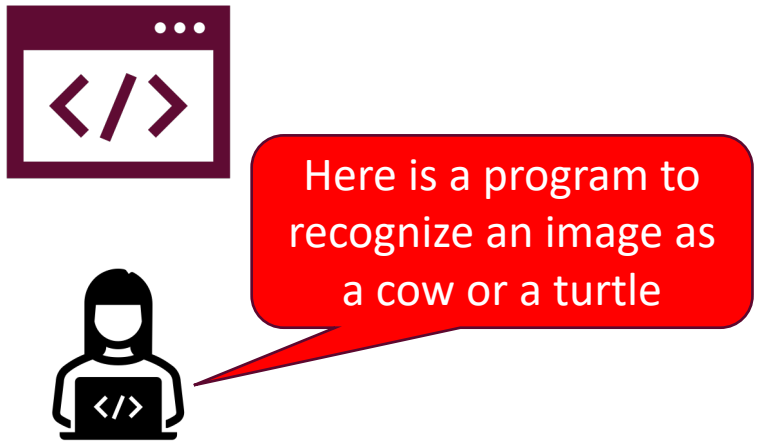


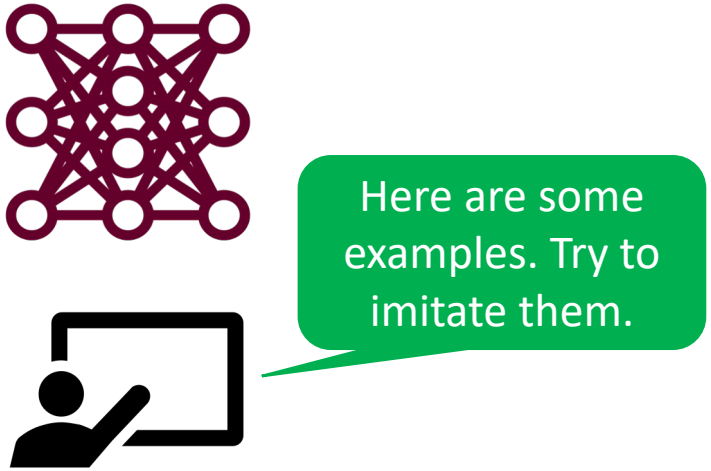



```
def compute_force(m, a):
        '''
        returns force (in N) needed to
        move mass m (in kg) at
        acceleration a (in m/s^2)
        '''

        F = m * a

        return F
```

Mass m (kg)	Acceleration a (m/s^2)	Force F (N)
2.5	4	10
5	2	10
20	0.5	10
40	0.25	10
40	2.5	100
20	5	100
50	2	100

Task specification in ML: ~~programs~~ examples

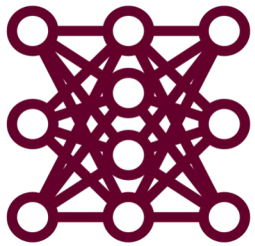




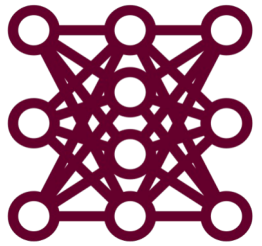

```
def cow_or_turtle(image):
```

? ? ?

Putting a trained ML system to use

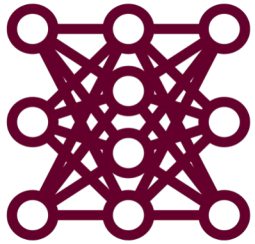


Here are some examples. Try to imitate them.

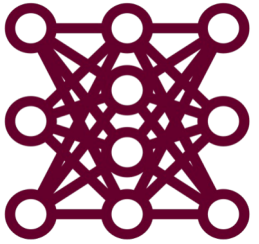


"cow"

Putting a trained ML system to use

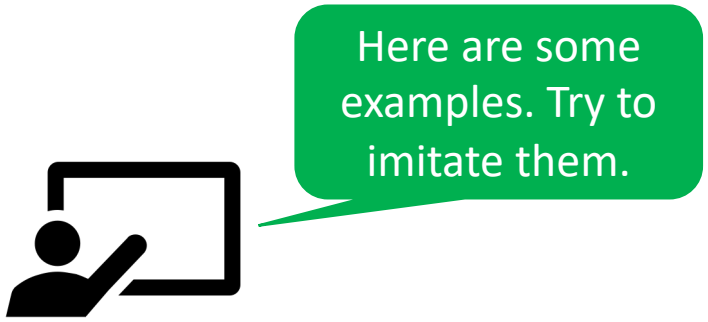


Here are some examples. Try to imitate them.

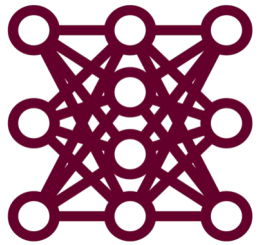


"turtle"

ML can only ever be as good as the task specification



- Collection of annotated examples
- A measure of imitation performance

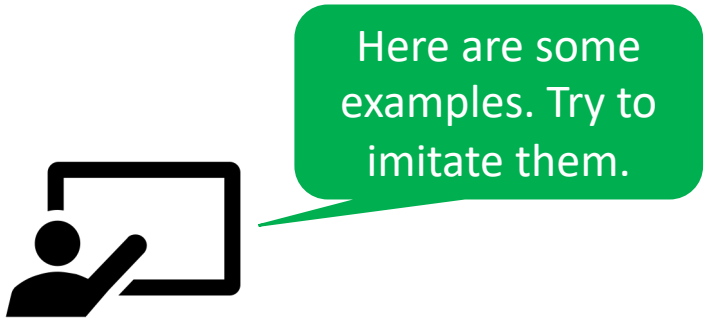


“turtle?”

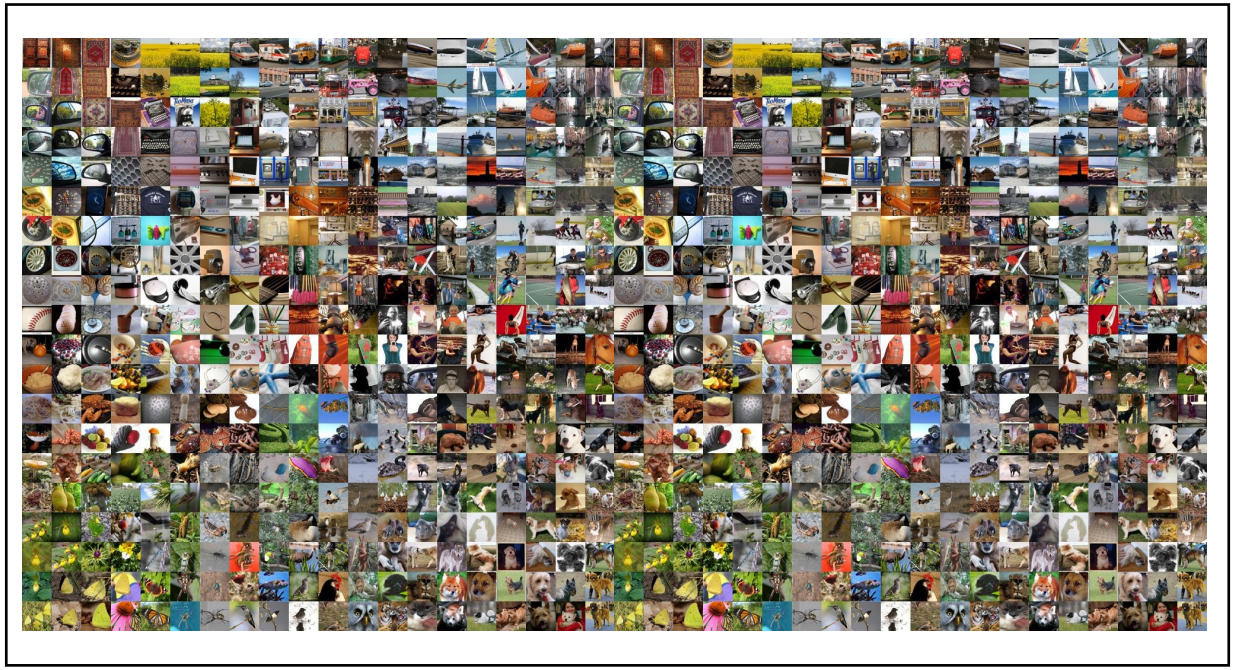
Is “cow” just a word for “grass”?

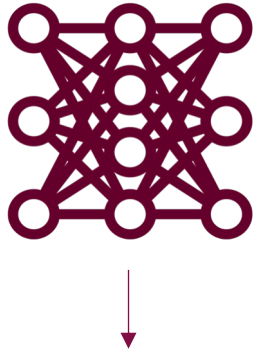
The dangers of out-of-domain use

ML can only ever be as good as the task specification



- Collection of annotated examples
- A measure of imitation performance





“snow?”

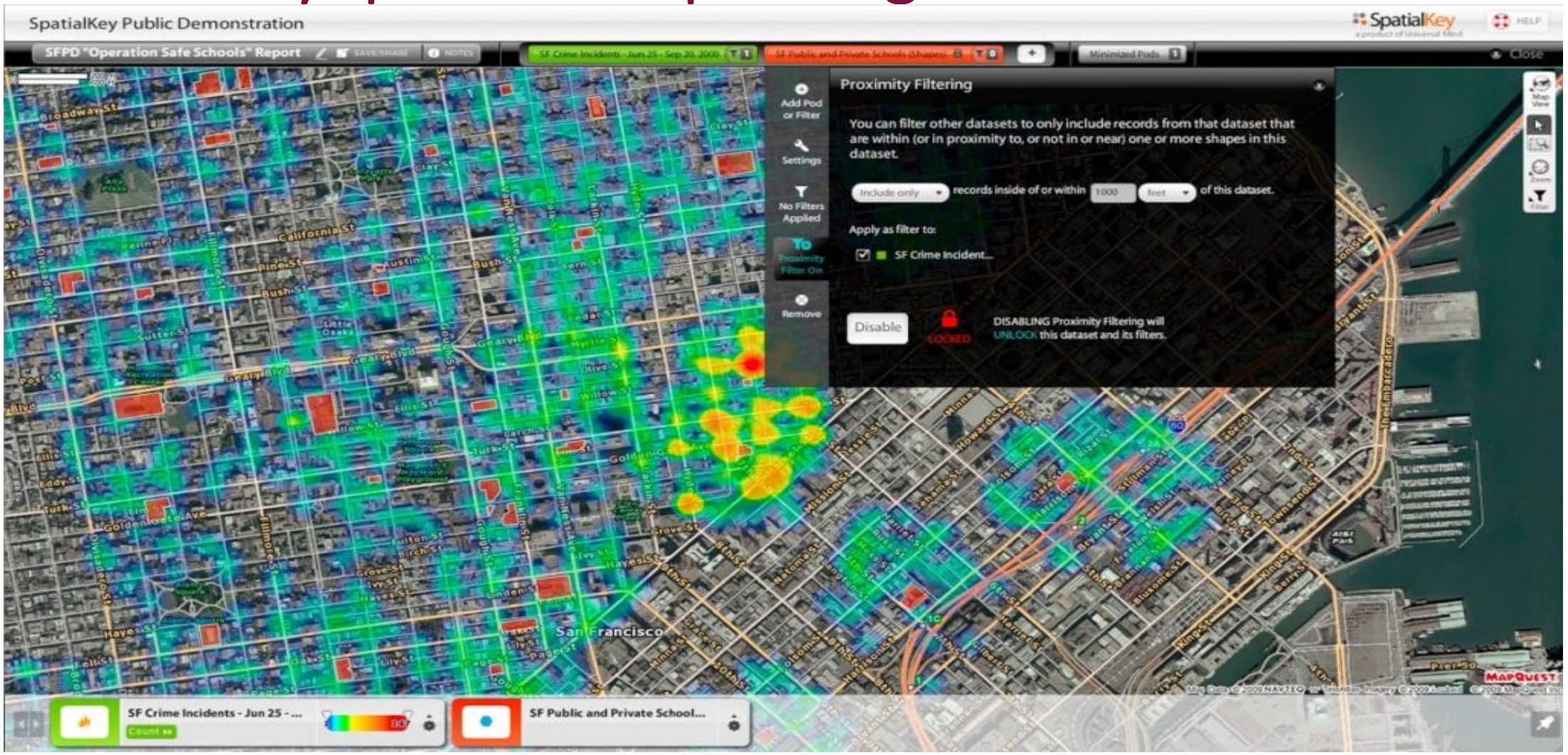
ML is increasingly deployed in real-life situations ...

... but are today's ML systems really ready for this?

How does ML interact with the rest of the system?

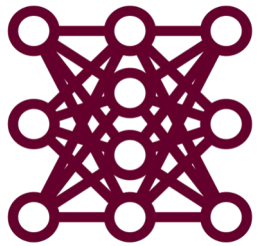
- ML is usually applied to some component of a large system
- A myopic focus on ML alone is a recipe for bad outcomes

A case study: predictive policing

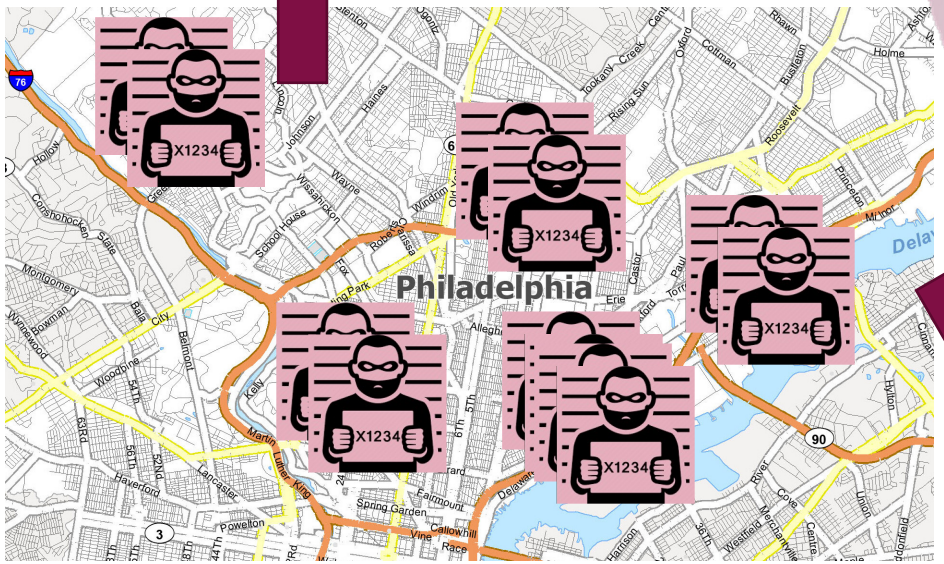


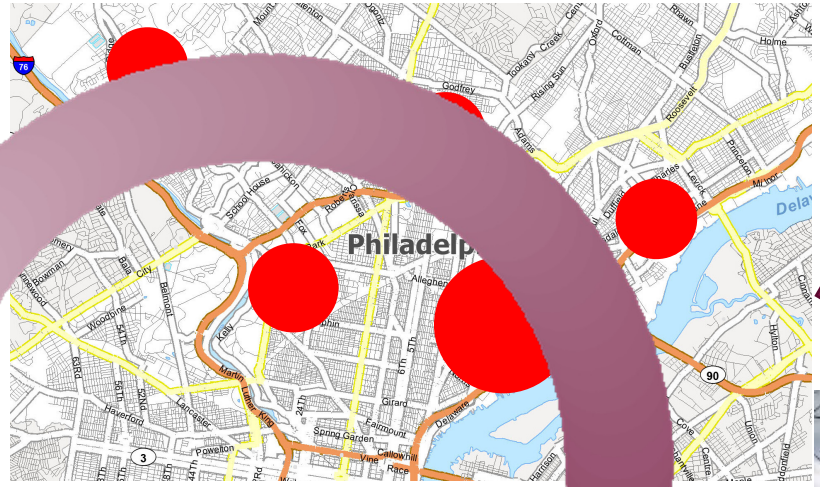
ML system for policing, to decide where and how much to patrol.
Widely used in most major city police departments in the US.

Complexity from how the ML system is used



Here are some examples. Try to imitate them.





"The model predicts crime there. Send officers there."

Further complexity from historical biases in data

Years of systemic biases in policing and the criminal justice system

This is not an imaginary example. It really happened.

Feedback In Machine Learning

To predict and serve?

Kristian Lum, William Isaac

PredPol, a program for police departments that predicts hotspots where future crime might occur, could potentially **get stuck in a feedback loop of over-policing majority black and brown neighbourhoods**.

Predictions used to drive increased police patrolling in those areas
Future observations of crime in those areas confirm predictions
Fewer opportunities to observe crime in other areas
Fed back into the learning system to further compound bias!

Samuel Sinyangwe, a justice activist and policy researcher:

This kind of approach is “especially nefarious” because police can say: “We’re not being biased, we’re just doing what the math tells us.” And the public perception might be that the algorithms are impartial.

Rise of the racist robots – how AI is learning all our worst impulses

Feedback In Machine Learning

- “a system for predicting the click through rate (CTR) of news headlines on a website likely relies on user clicks as training labels, which in turn depend on previous predictions from the model.”

Sometimes “adversarial” feedback:

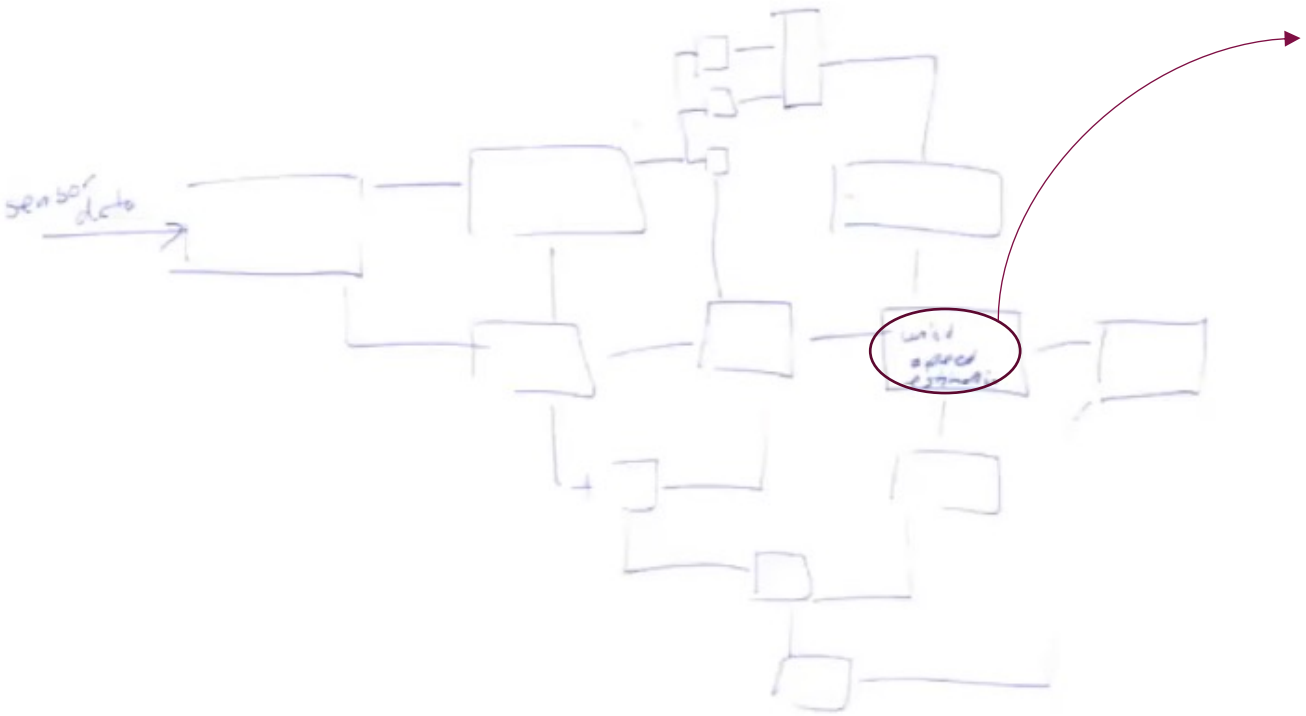
- Tricking a resume screening system by entering keywords like “Oxford”
- Anecdotal: Computer vision systems to predict poverty and (semi-) automate global aid allocation decisions lead to people switching off their night lights and dressing up concrete roofs as thatched roofs.

Satellite images used to predict poverty
By Paul Rincon
Science editor, BBC News website

Machine Learning: The High Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)

Feedback in Machine Learning

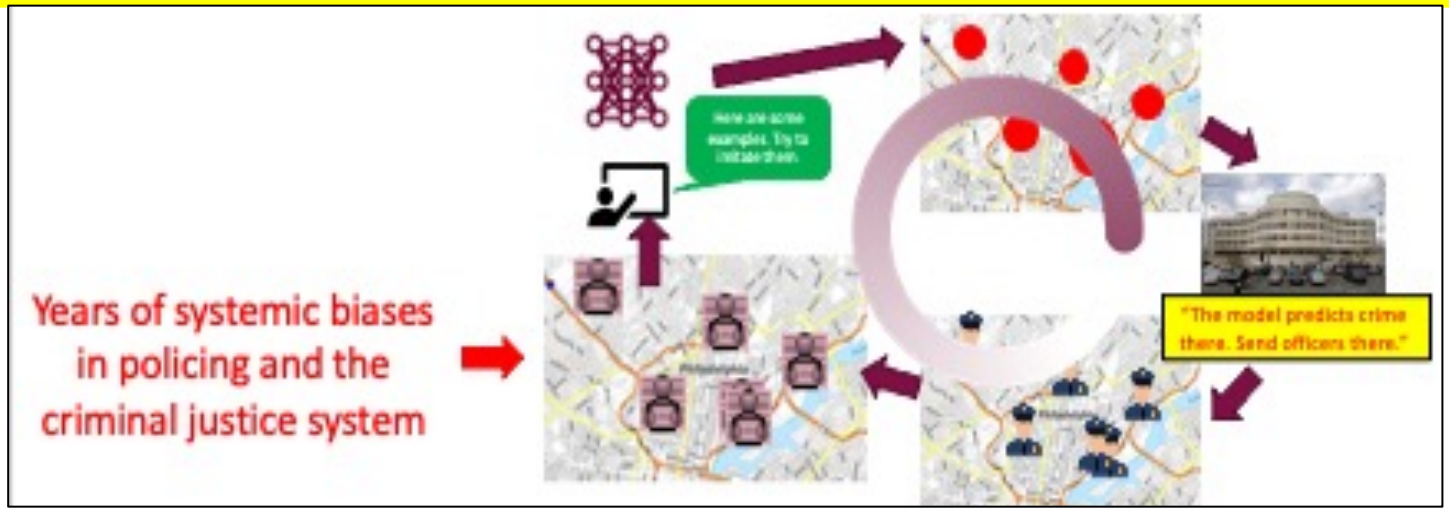


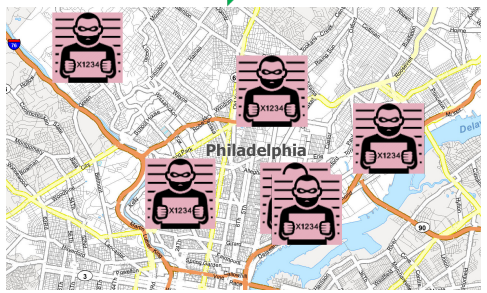
- We build an ML wind-direction classifier, lots of work
  - It gets 99% accuracy
- Consider extreme case of feedback – consume its own output from the previous time cycle
- The next time we go to build a classifier, we will dump out a training set (that includes the old classifier as an input feature)
  - New intern comes in and he build a classifier with 98.5% accuracy that runs *1000x* faster
  - Team rejoices, installs new classifier in system
  - Catastrophic failure when running it online
- Why? New classifier used only one feature – the old classifier's prediction.

Drew Bagnell, Feedback in Machine Learning, https://www.youtube.com/watch?v=XRSvz4UOp64

ML task specifications are oversimplified

None of this complex *context* is communicated to the ML algorithm.

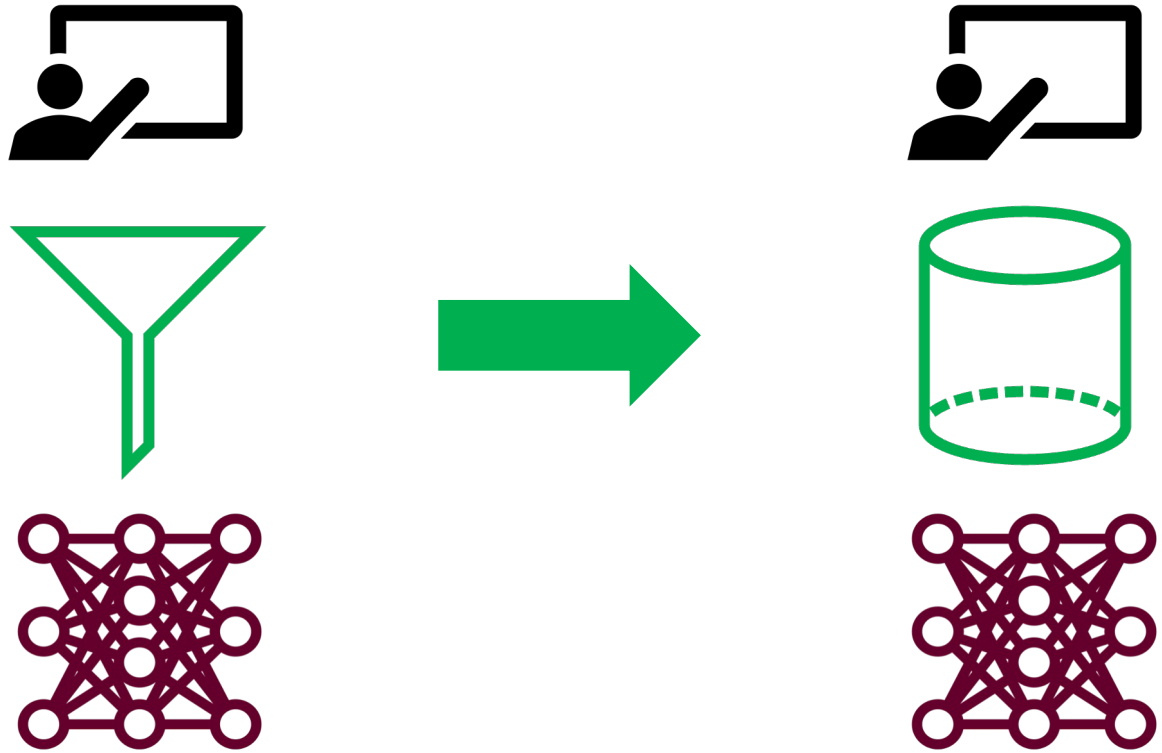




"Here are some examples. Try to imitate them."

How might we teach machines better?

A wider, more open communication channel between human teachers and machine learners



Human Augmentation and Auditing

~~"The algorithm made me do it."~~

- **ML engineers must commit to assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes.**
- Esp in domains with significant impact on human lives (e.g. justice, health, transport, etc).
- All stakeholders' values and perspectives should be accounted for during algorithm design.
- Subject-domain-experts as human-in-the-loop reviewers for deployed ML systems.

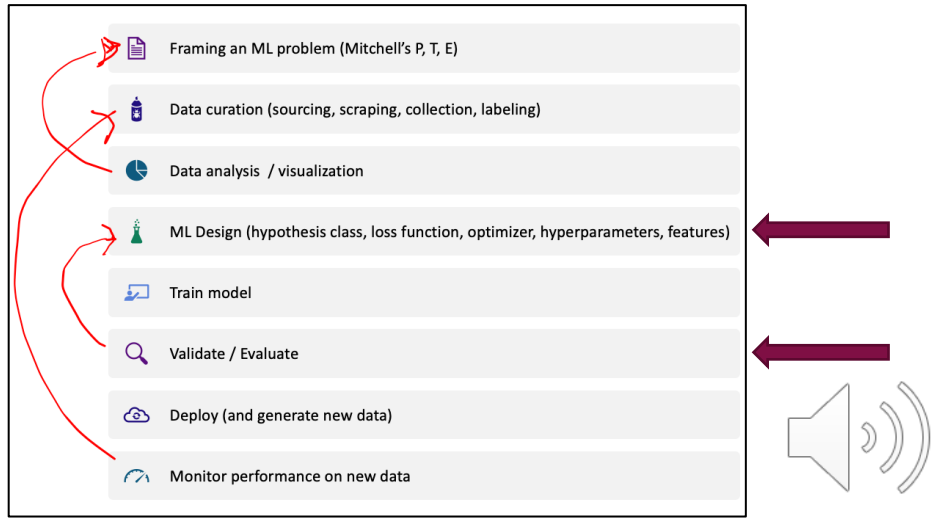
Based on material from The Institute for Ethical AI and ML

Example: Human-In-The-Loop “Fairness”

Notions of fairness that all seem intuitively reasonable turn out to be incompatible

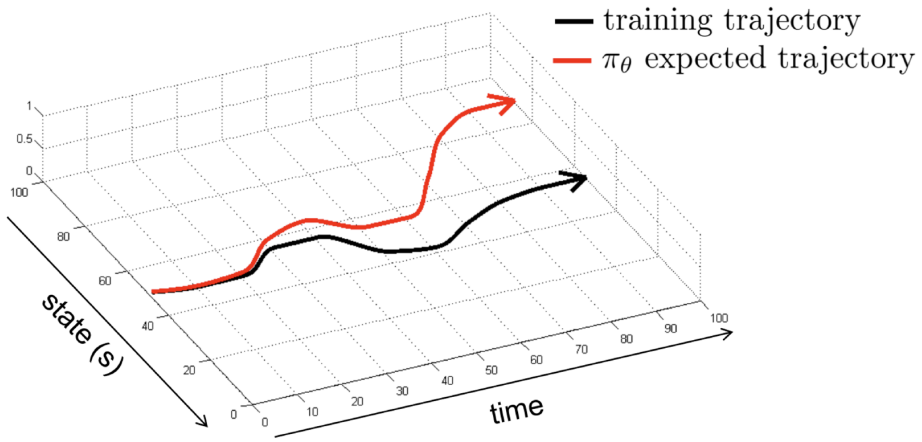
Doesn’t mean that we should give up on ML ever being fair! Just means general fairness is hard to formalize. Lots of open research questions, and ongoing activity.

One potential solution: Have domain experts weigh in on what performance metrics result in fair model selection/training. [Hiranandani, Narasimhan, Koyejo, “Fair Performance Metric Elicitation”]



Human-In-The-Loop Training Data Aggregation

Problem: Distributional Shift in Imitation Learning



The cloned policy is imperfect; this leads to accumulating errors, and the agent soon encounters unfamiliar states, leading to failure.

Active Behavior Cloning: DAGGER

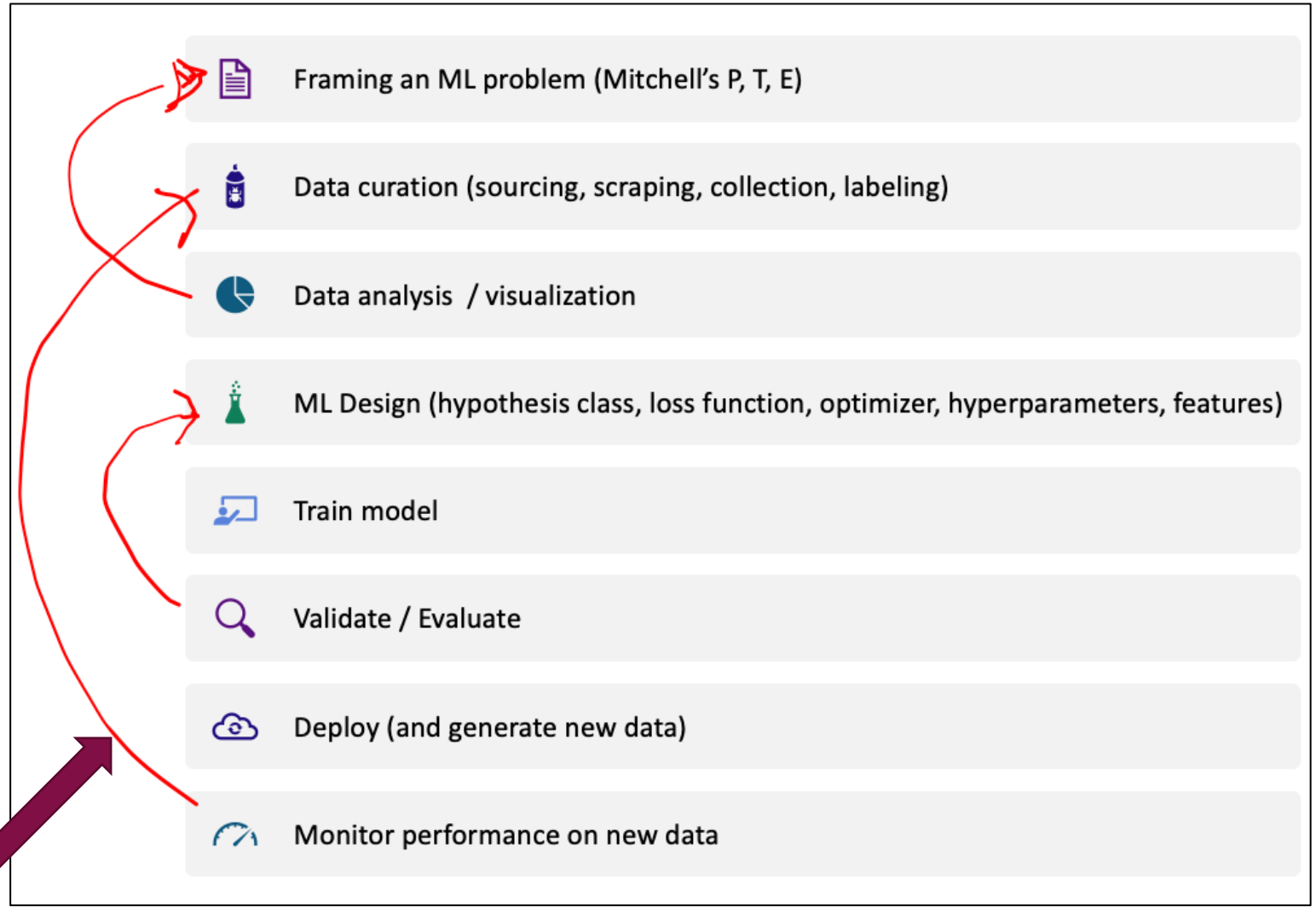
A general trick for handling distributional shift: requery expert on new states encountered by the initial cloned policy upon execution, then retrain.

1. train $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ from human data $\mathcal{D} = \{\mathbf{o}_1, \mathbf{a}_1, \ldots, \mathbf{o}_N, \mathbf{a}_N\}$
2. run $\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ to get dataset $\mathcal{D}_\pi = \{\mathbf{o}_1, \ldots, \mathbf{o}_M\}$
3. Ask human to label $\mathcal{D}_\pi$ with actions $\mathbf{a}_t$
4. Aggregate: $\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_\pi$

Imitation learning inspired solutions to feedback in (supervised) ML?

Drew Bagnell, Feedback in Machine Learning, https://www.youtube.com/watch?v=XRSvz4UOpo4

Recall: ML Workflow



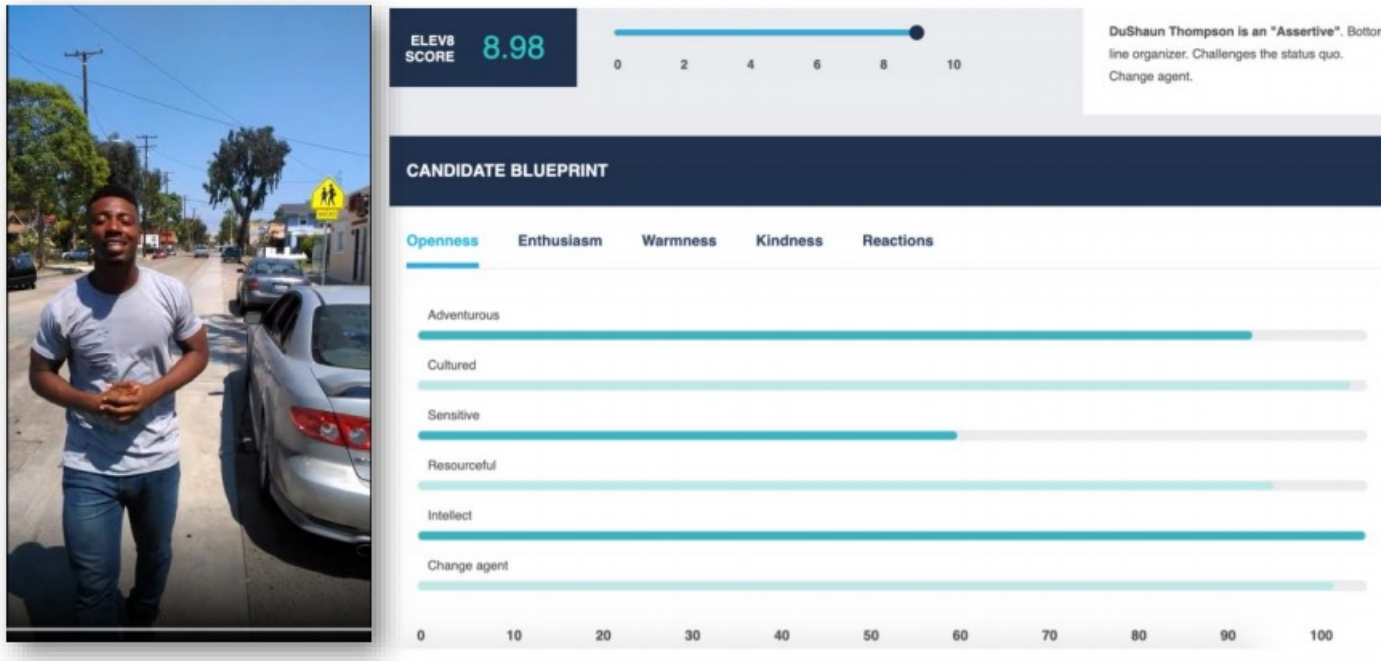
The Ethics of Public (Mis-)Information about ML

Where Ethical Problems Arise

- High-performing ML systems → concerns due to high accuracy
  - E.g. Facial recognition in the hands of an authoritarian state
- Flawed ML systems → concerns due to errors
  - E.g. Content recommendation creating filter bubbles
- Not-even-wrong ML systems
  - E.g. refugees being turned away at the US border because an ML system assessed risk of terrorist activity based on Instagram posts

Based on slides by Arvind Narayanan

Example: Resume Evaluation with ML



Vendor name	Funding	# of employees	Location
8 and Above	–	1-10	WA, USA
ActiView	$6.5M	11-50	Israel
Applied	£2M	11-50	UK
Assessment Innovation	$1.3M	1-10	NY, USA
Good&Co	$10.3M	51-100	CA, USA
Harver	$14M	51-100	NY, USA
HireVue	$93M	251-500	UT, USA
impress.ai	$1.4M	11-50	Singapore
Knockri	–	11-50	Canada
Koru	$15.6M	11-50	WA, USA
LaunchPad Recruits	£2M	11-50	UK
myInterview	$1.4M	1-10	Australia
Plum.io	$1.9M	11-50	Canada
PredictiveHire	A$4.3M	11-50	Australia
pymetrics	$56.6M	51-100	NY, USA
Scoutible	$6.5M	1-10	CA, USA
Teamscope	€800K	1-10	Estonia
ThriveMap	£781K	1-10	UK
Yobs	$1M	11-50	CA, USA

Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices

Manish Raghavan* Solon Barocas† Jon Kleinberg* Karen Levy*

Based on slides by Arvind Narayanan

Example: Resume Evaluation with ML

How to persuade a robot that you should get the job

Do mere human beings stand a chance against software that claims to reveal what a real-life face-to-face chat can't?

Stephen Buranyi

Sat 3 Mar 2018 19.05 EST

Based on slides by Arvind Narayanan

Ethics of Public Information

ML/AI is not magic.

Huge commercial interest in creating hype to convince the public that it *is*.

Hype causes real harm.

How to recognize AI snake oil

Arvind Narayanan

Practical Principles for Ethical ML Systems

Where does ML fit into a larger system?

The builders of technology must bear moral responsibility in how it is deployed.

- If your recidivism prediction ML model works at 60% accuracy (random chance 50%), does the judge who is using this system understand this well? Do they understand its error tendencies?
- You might be “just building a face classifier”, but what if it is used by an authoritarian government to track people or to target minority groups?
- In some cases, even the technology goal itself is obviously questionable. E.g., why even build a classifier to predict race or rate resumes?

Best practices for ethical ML

1. **Human Augmentation**
2. **Bias Evaluation**
3. **Explainability and Justification**
4. Reproducible Operations
5. **Displacement Strategy**
6. Practical Accuracy
7. Trust by Privacy
8. Data Risk Awareness

1. Human Augmentation

~~"The algorithm made me do it."~~

Assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes.

- Esp in domains with significant impact on human lives (e.g. justice, health, transport, etc).
- Workflow changes:
  - All stakeholders' values and perspectives should be accounted for during algorithm design.
  - Subject-domain-experts as human-in-the-loop reviewers at the end of ML systems.

Based on material from The Institute for Ethical AI and ML

2. Bias Evaluation

Continously develop processes that allow me to understand, document and monitor bias in development and production.

- No standard strategy, need to careful consider potential sources of bias for the domain you are working in
- Requires continuous monitoring, not one-time effort

Datasheets for Datasets

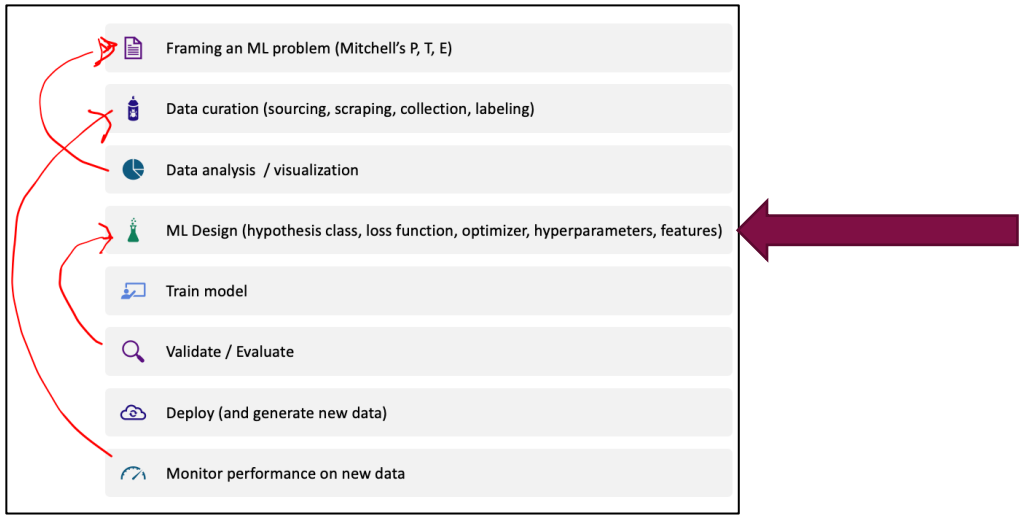
TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

Based on material from The Institute for Ethical AI and ML

3. Explainability by justification

Develop and utilize tools and processes to continuously improve transparency and explainability of machine learning models where reasonable.

- Even though on certain situations accuracy may decrease, the transparency and explainability gains may be significant.
- Important for end users to be able to understand and weigh in on decisions from ML systems.
- Especially important in the context of hype and misinformation about ML



Based on material from The Institute for Ethical AI and ML

Explainability in Law

- France's Digital Republic Act gives the right to an explanation as regards decisions on an individual made through the use of administrative algorithms.
  - how and to what extent the algorithmic processing contributed to the decision-making
  - which data was processed and its source
  - how parameters were treated and weighted
  - which operations were carried out in the treatment.

- EU, China etc. have made similar moves.

Difficulties of Transparency

- Not always easy to be transparent, because:
  - Leaking of sensitive data
  - Easy to game. e.g. “adversarial feedback”
  - Loss of competitive advantage
  - Sometimes hard to interpret even for experts
- But lack of transparency makes it difficult to interpret results well.

Samuele Lo Piano, “Ethical Principles …” 2020

5. Displacement Strategy

Identify and document relevant information so that business change processes can be developed to mitigate the impact towards workers being automated.

- Ensure all stakeholders are brought on board and develop a change-management strategy before automation

- Often, the workers are asked to do labor (e.g., generating training data) that will help automate themselves. Are they appropriately compensated?

Accountability

Question: should the "driver" of an automated car be able to command it to go 80 MPH on a 55 MPH road?

NO

- It would be illegal
- It may endanger others
- Who is liable for accidents?
  - Driver for issuing the command?
  - AI developer for allowing it?
  - Auto manufacturer for choosing that AI system?
  - The insurance company?

YES

- People speed now anyway and there's little we can do about it
- The driver should have command of the vehicle
- What about exceptions?
  - Rushing someone to the hospital
  - Escaping a tornado
  - etc...

Accountability in decision making systems is complex

Other Important Ethical Questions in ML

- **AI safety**: how can we make AI without unintended negative consequences?
- **Aligned AI:** How can AI make decisions that align with our values?

Looking Forward

- The ethics of ML and AI systems is an urgent topic *now*, not because of speculative AI apocalypse scenarios
- Open and active area of research, requires ML engineers to work closely with scholars across the social sciences and across various domains of deployment.
- The law moves slowly, and legal frameworks have much to catch up to.

Laws without morals are useless