CS188: Artificial Intelligence
AI for Equality
+ Course Wrap-up

Instructors: Emma Pierson and Peyrin Kao

Logistics

- Please fill out the course evals!
  - If we get a high enough completion rate, everyone gets extra credit!

Two-part lecture

- Using AI to increase social equality
  - Pain
  - Policing
- Course wrap-up: what's next?
  - For AI?
  - For you?

Two-part lecture

- Using AI to increase social equality
  - Pain
  - Policing
- Course wrap-up: what's next?
  - For AI?
  - For you?

Motivation

- 1. Our society is deeply unequal.

'I Can't Breathe': 4 Minneapolis Officers Fired After Black Man Dies in Custody

Ahmaud Arbery killing being investigated as federal hate crime, family attorney says

RESEARCH ARTICLE

ECONOMICS

The fading American dream: Trends in absolute income mobility since 1940

For most U.S. workers, real wages have barely budged in decades

Black Americans Face Alarming Rates of Coronavirus Infection in Some States

FBI: Hate crimes reach 5-year high in 2016, jumped as Trump rolled toward presidency

Study Shows Income Gap Between Rich and Poor Keeps Growing, With Deadly Effects

Motivation

- 2. Computer scientists can make this worse.

Motivation

- 2. Computer scientists can make this worse.
  - Language

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]

Semantics derived automatically from language corpora contain human-like biases

Aylin Caliskan[1,*], Joanna J. Bryson[1,2,*], Arvind Narayanan[1,*]

Motivation

- 2. Computer scientists can make this worse.
  - Language
  - Vision

Motivation

- 2. Computer scientists can make this worse.
  - Language
  - Vision
  - Healthcare

RESEARCH ARTICLE

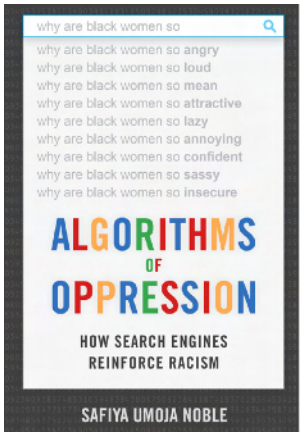
ECONOMICS

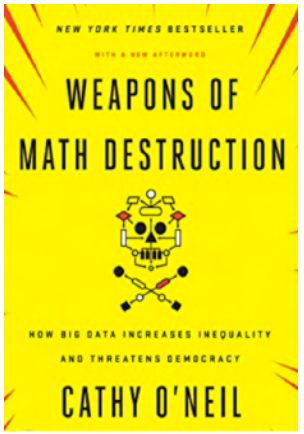
Dissecting racial bias in an algorithm used to manage the health of populations

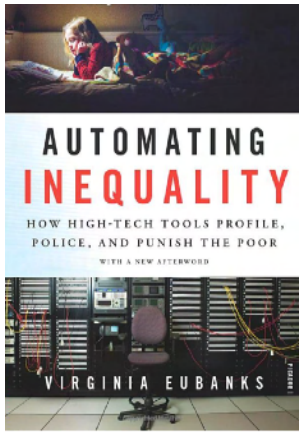
Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

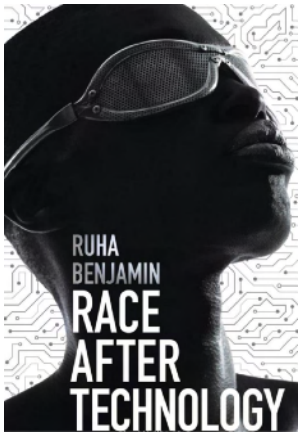
Motivation

- 2. Computer scientists can make this worse.
  - Language
  - Vision
  - Healthcare
  - Etc









Motivation

- 3. But we can also make it better.

Two-part lecture

- Using AI to increase social equality
  - Pain
  - Policing
- Course wrap-up: what's next?
  - For AI?
  - For you?

Using AI to understand inequality in pain

Pierson, Cutler, Leskovec, Mullainathan, and Obermeyer. *Nature Medicine*, 2021.

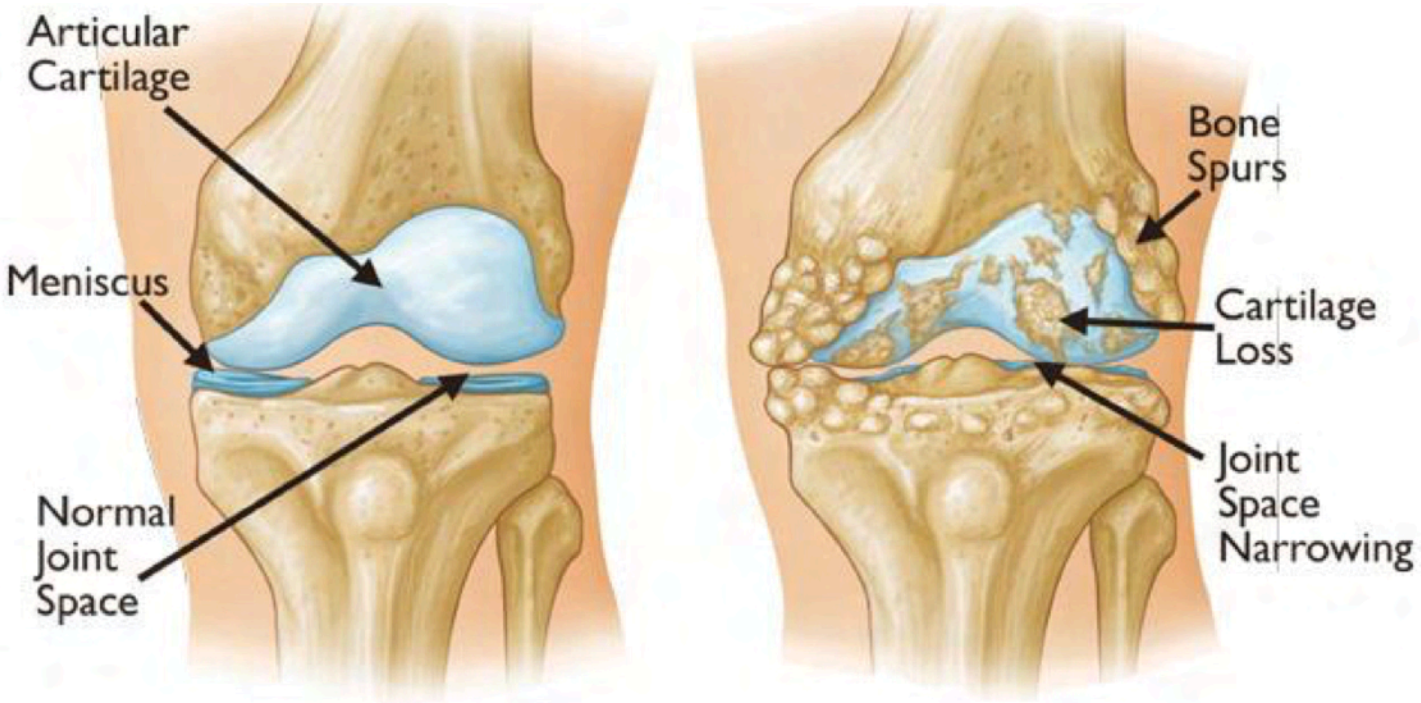
Disadvantaged groups experience more pain

"**Socioeconomic disadvantage** [SED]…is **consistently associated with increased risk for pain**…[across] pain sites…continents…in both community samples and medical settings."

[1] Poleshuck and Green. "Socioeconomic disadvantage and pain." *Pain,* 2008.
[2] Anderson, Green, and Payne. "Racial and ethnic disparities in pain: causes and consequences of unequal care." *The Journal of Pain*, 2009.

This is also true in knee osteoarthritis

- 10% of men over 60 and 13% of women over 60 have knee osteoarthritis



Zhang and Jordan. "Epidemiology of osteoarthritis." *Clinics in Geriatric Medicine,* 2010.

This is also true in knee osteoarthritis

- 10% of men over 60 and 13% of women over 60 have knee osteoarthritis
- Disadvantaged racial and socioeconomic groups have worse pain

This is also true in knee osteoarthritis

- 10% of men over 60 and 13% of women over 60 have knee osteoarthritis
- Disadvantaged racial and socioeconomic groups have worse pain
  - …maybe they have more severe disease?

This is also true in knee osteoarthritis

- 10% of men over 60 and 13% of women over 60 have knee osteoarthritis
- Disadvantaged racial and socioeconomic groups have worse pain
  - …maybe they have more severe disease?
- **No**: worse pain even when we control for doctor's assessment of severity

[1] Zhang and Jordan. "Epidemiology of osteoarthritis." *Clinics in Geriatric Medicine,* 2010.

[2] Bolen, Schieb, Hootman, et al. "Differences in the Prevalence and Impact of Arthritis Among Racial/Ethnic Groups in the United States, National Health Interview Survey, 2002, 2003, and 2006". *Prevalence of Chronic Diseases,* 2010.

[3] Poleshuck and Green. "Socioeconomic disadvantage and pain." *Pain,* 2008.

[4] Anderson, Green, and Payne. "Racial and ethnic disparities in pain: causes and consequences of unequal care." *The Journal of Pain*, 2009.

[5] Allen, Helmick, Schwartz, DeVellis, Renner, and Jordan. "Racial differences in self-reported pain and function among individuals with radiographic hip and knee osteoarthritis: the Johnston County Osteoarthritis Project." *Osteoarthritis Cartilage,* 2009.

[6] Vina, Ran, Ashbeck, and Kwoh. "Natural History of Pain and Disability among African-Americans and Whites With or At Risk For Knee Osteoarthritis: A Longitudinal Study." *Osteoarthritis Cartilage*, 2018.

Measuring severity and pain

Severity

Definite osteophyte;
possible joint space reduction;
Kellgren-Lawrence Grade
(KLG) 2/4

[1] Roos et al Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther,* 1998.

[2] Kellgren and Lawrence. Radiological Assessment of Osteo-Arthrosis. *Ann Rheum Dis,* 1957.

Measuring severity and pain

Severity

Definite osteophyte; possible joint space reduction; Kellgren-Lawrence Grade **(KLG) 2/4**

Pain

I feel **severe** pain when bending my knee.

Multi-item survey aggregated into single **KOOS pain score**

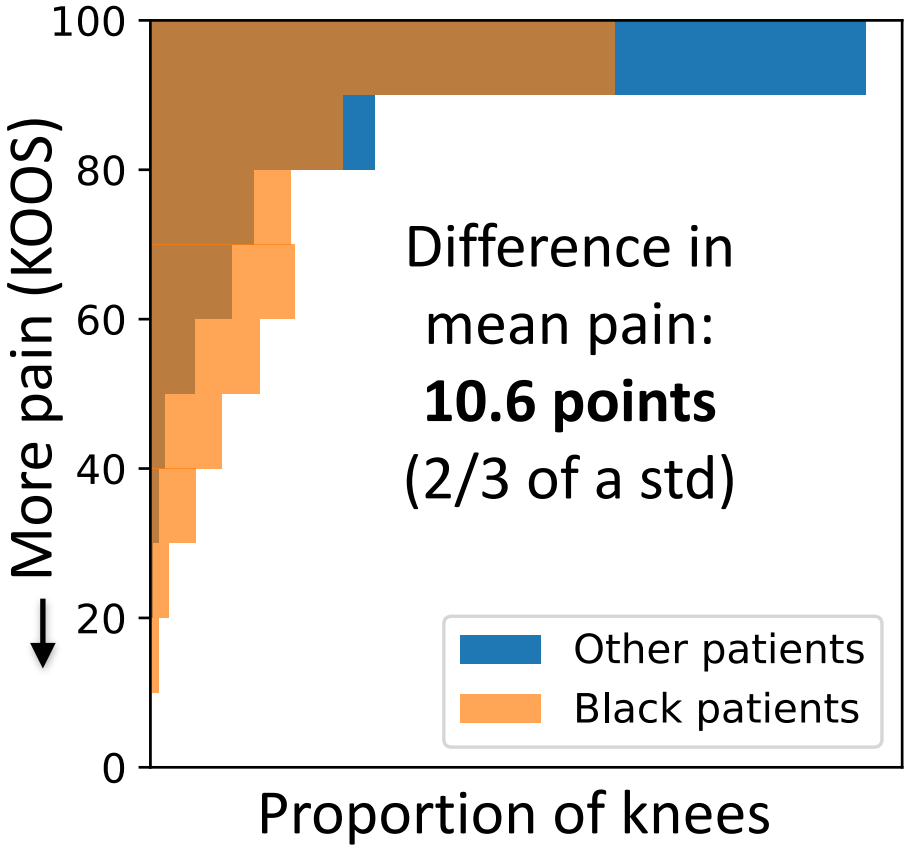
[1] Roos et al Knee Injury and Osteoarthritis Outcome Score (KOOS)--development of a self-administered outcome measure. *J Orthop Sports Phys Ther,* 1998.

[2] Kellgren and Lawrence. Radiological Assessment of Osteo-Arthrosis. *Ann Rheum Dis,* 1957.

Data: Osteoarthritis Initiative

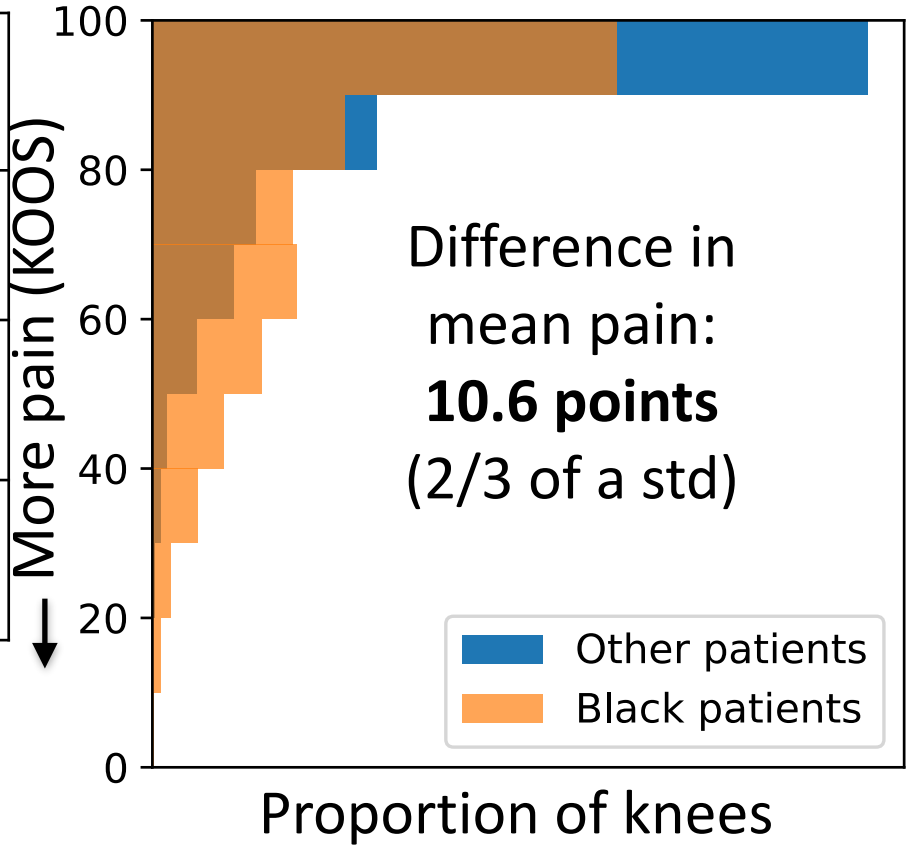
- Widely used longitudinal, multisite study of people with, or at high risk of, osteoarthritis
- Recruited people at 5 sites through many pathways: mailings, newspaper ads, recruitment website, etc
- Mean age: 61; 56% female; 16% black
- All results presented are on 1,295 people
- Compare pain by three binary groupings:
  - **Race:** black vs. non-black patients (97% of non-black patients are white)
  - **Income:** below $50k vs. above $50k
  - **Education:** didn't graduate college vs. graduated college

Disadvantaged patients have more pain…



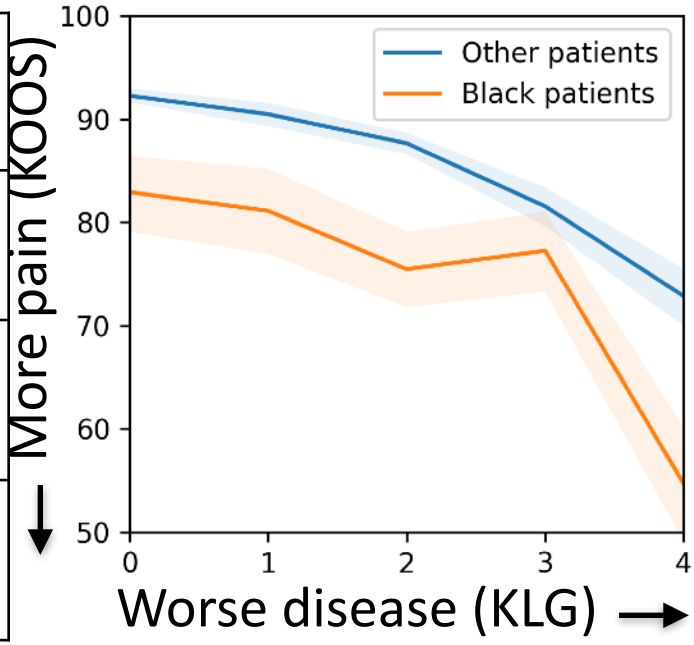
Disadvantaged patients have more pain…

	Pain gap (no controls)
Race	**10.6** (8.3, 12.9)
Income	**4.2** (2.8, 5.6)
Education	**5.3** (3.7, 6.7)



…even when controlling for severity

	Pain gap (no controls)	Pain gap (control for KLG)
Race	**10.6** (8.3, 12.9)	**9.7** (7.4, 11.9)
Income	**4.2** (2.8, 5.6)	**3.5** (2.3, 4.9)
Education	**5.3** (3.7, 6.7)	**4.9** (3.5, 6.2)



$$pain \sim race + KLG$$

regress pain on race and KLG

pain gap = race coefficient when controlling for KLG

…even when controlling for severity

	Pain gap (no controls)	Pain gap (control for KLG)	% of pain gap explained by KLG
Race	**10.6** (8.3, 12.9)	**9.7** (7.4, 11.9)	**9%** (3%, 16%)
Income	**4.2** (2.8, 5.6)	**3.5** (2.3, 4.9)	**16%** (5%, 29%)
Education	**5.3** (3.7, 6.7)	**4.9** (3.5, 6.2)	**8%** (-1%, 18%)

Controlling for severity (KLG) doesn't really narrow pain gap!

(Not just in our dataset; other studies find this too)

Goal of paper is to explain why.

[1] Eberly et al. Psychosocial and demographic factors influencing pain scores of patients with knee osteoarthritis. *PloS One,* 2018.
[2] Allen et al. Racial differences in self-reported pain and function among individuals with radiographic hip and knee osteoarthritis: the Johnston County Osteoarthritis Project. *Osteoarthritis Cartilage*, 2009.

One theory of pain gap

- **“Outside their knees”:** non-knee-related factors cause disadvantaged patients to report higher pain given equally severe knee disease

One theory of pain gap

- **“Outside their knees”:** non-knee-related factors cause disadvantaged patients to report higher pain given equally severe knee disease
  - Life stress
  - Less access to pain medication
  - Less social support
  - Differences in pain reporting

[1] Krause et al. "Psychosocial job factors associated with back and neck pain in public transit operators." Scandinavian Journal of Work, Environment & Health, 1997.
[2] Gatchel, Polatin, and Mayer. "The dominant role of psychosocial risk factors in the development of chronic low back pain disability." Spine, 1995.
[3] Poleshuck and Green. "Socioeconomic disadvantage and pain." Pain, 2008.
[4] Anderson, Green, and Payne. "Racial and ethnic disparities in pain: causes and consequences of unequal care." The Journal of Pain, 2009.
[5] Jones et al. "Investigating racial differences in coping with chronic osteoarthritis pain." Journal of Cross-Cultural Gerontology, 2008.
[6] Peacock and Patel. “Cultural Influences on Pain”. Reviews in Pain, 2008.

But there's another possibility

- **"Outside their knees":** non-knee-related factors cause disadvantaged patients to report higher pain given equally severe knee disease
- **"In their knees":** pain-related ailments in the knee x-ray which KLG isn't capturing, and which would explain more of the pain gap

Why is “in their knees” plausible?

- 1. **We don’t understand pain that well.** KLG doesn’t explain pain well ($R^2 = 0.10$).
- 2. **KLG was developed 60 years ago in non-diverse populations.**

[1] Kellgren and Lawrence. “Radiological Assessment of Osteo-Arthrosis”. *Ann Rheum Dis*, 1957.
[2] Haug, Compton, and Courbage. “The Demographic Characteristics of Immigrant Populations”. Council of Europe, 2002.

Are there overlooked physical features in the knee which explain the higher pain levels in disadvantaged groups?

Method

- Train an algorithm (CNN) to search for additional signal in the knee x-ray which would explain the higher pain levels in disadvantaged groups

Method

- Train an algorithm (CNN) to search for additional signal in the knee x-ray which would explain the higher pain levels in disadvantaged groups

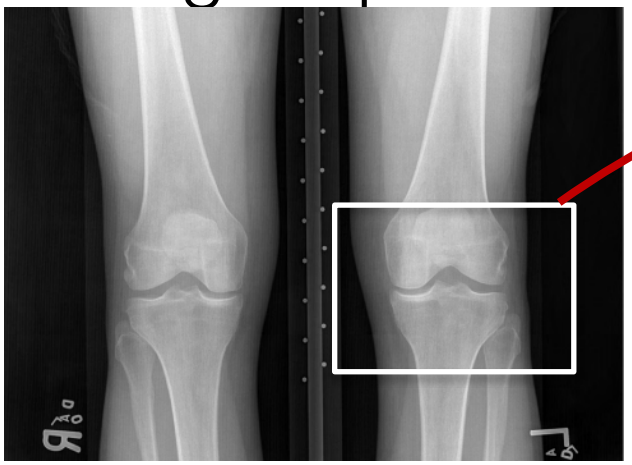
KLG: 2

Standard approach:
train model to replicate doctor judgment

[1] Tiulpin et al. “Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach”. *Scientific Reports,* 2018.
[2] Antony et al. “Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks”. *Int Conf Pattern Recognit,* 2016.
[3] Oka et al. “Fully automatic quantification of knee osteoarthritis severity on plain radiographs”. *Osteoarthritis and Cartilage,* 2008.
[4] Chen et al. “Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss”. *Computerized Medical Imaging and Graphics*, 2019.

Method

- Train an algorithm (CNN) to search for additional signal in the knee x-ray which would explain the higher pain levels in disadvantaged groups



KLG: 2

Standard approach:
train model to replicate doctor judgment

Problem: if KLG doesn't capture all the pain-relevant features, don't want to just replicate it.

Prediction target matters!

- Train an algorithm (CNN) to search for additional signal in the knee x-ray which would explain the higher pain levels in disadvantaged groups

~~KLG: 2~~

Standard approach:
train model to replicate doctor judgment

KOOS pain score: 83.3

Our approach:
train model to *learn from the patient*
and predict KOOS pain score

Method

- Train an algorithm (CNN) to search for additional signal in the knee x-ray which would explain the higher pain levels in disadvantaged groups

Input: knee x-ray
Output: knee-specific pain prediction, **ALG-P**
(Algorithmic severity measure)

Method

If controlling for **algorithmic** severity ALG-P narrows pain gap more than controlling for **clinical** severity KLG

Clinical severity score overlooks knee features which explain disadvantaged patients' higher pain levels

Results

Algorithm finds additional signal for pain in knee x-ray

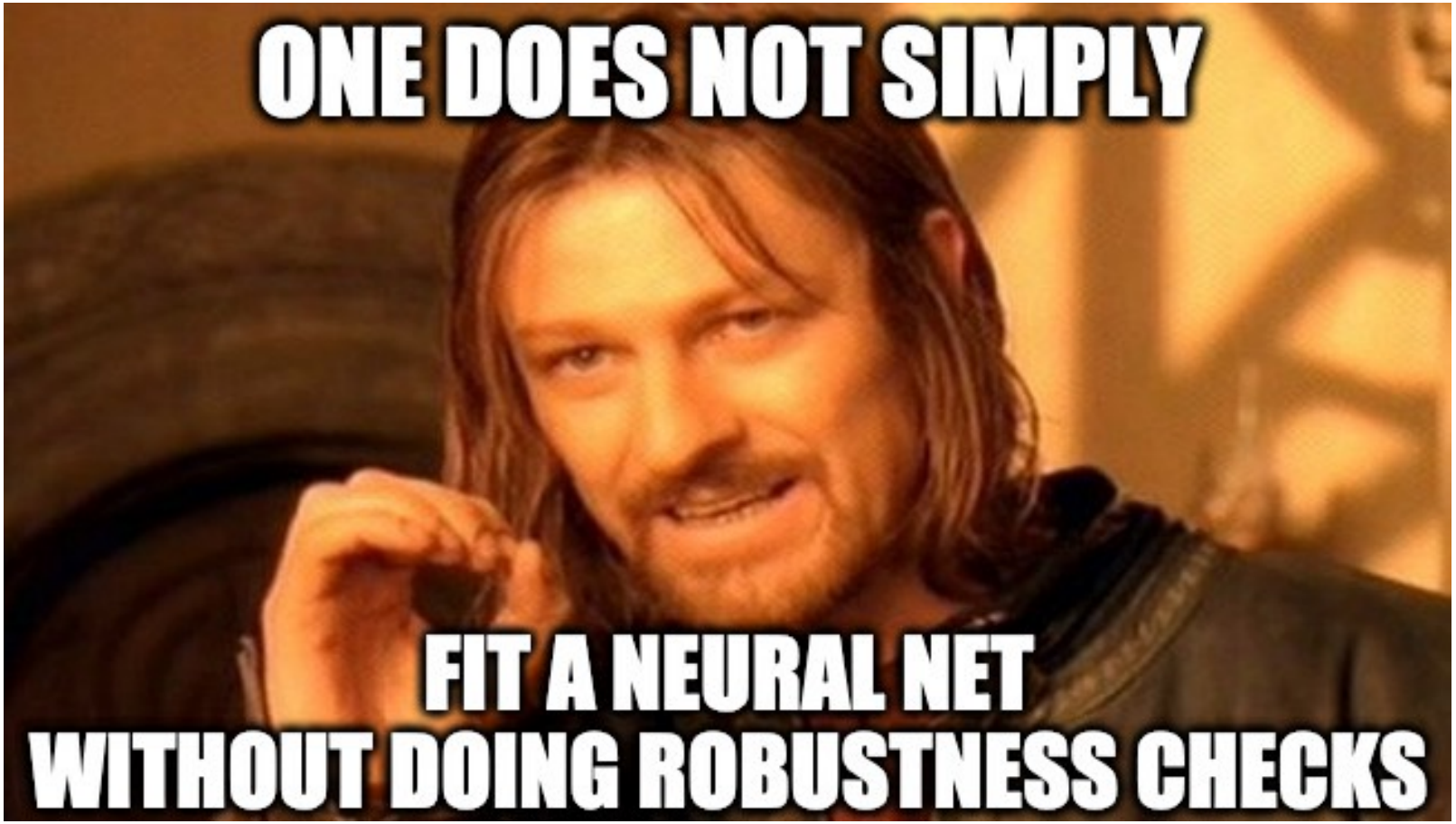
- Algorithmic severity score ALG-P predicts pain better than clinical severity score KLG
  - R^2 0.16 vs 0.10
  - Relative difference 61% (95% CI: 38%-86%)
  - Similar results for other predictive measures (RMSE, non-parametric correlation)
- **More important:** does controlling for algorithmic severity score reduce pain gap?

Controlling for algorithmic severity score reduces pain gap

	% of pain gap explained by KLG	% of pain gap explained by ALG-P	Ratio
Race	**9%** (3%, 16%)	**43%** (33%, 56%)	**4.7** (3.2, 11.8)
Income	**16%** (5%, 29%)	**32%** (18%, 50%)	**2.0** (1.4, 4.4)
Education	**8%** (-1%, 18%)	**30%** (18%, 44%)	**3.6** (2.1, *)

Implication: there is overlooked signal in the knee x-ray which helps explain disadvantaged patients' higher pain

*Because the KLG interval overlaps 0, the upper CI for the ratio of pain gaps explained is not well-defined.



Wiens et al. “Do no harm: a roadmap for responsible machine learning for health care”. *Nature Medicine,* 2019.

Robustness checks: is the algorithm just…

- 1. Predicting pain using image artifacts like BMI? **No.**
- 2. Predicting pain by predicting race/SES? **No.**
- 3. Learning a predictor that doesn't generalize across imaging sites? **No.**
- 4. Learning a continuous KLG? **No.**
- 5. Outperforming KLG but not other clinical predictors? **No.**
- 6. Re-weighting known image features? **No.**

Discussion

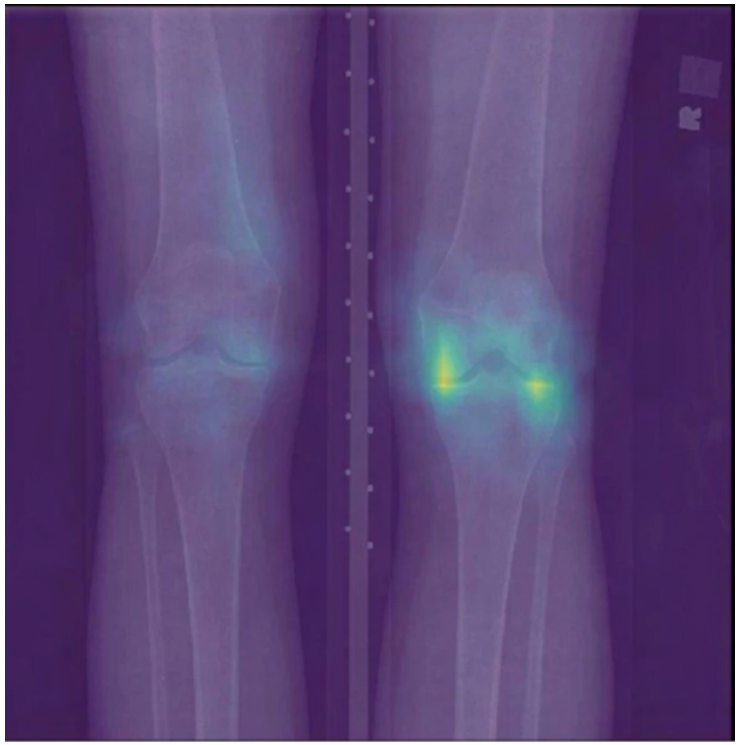
- We train a deep learning algorithm to predict pain from knee x-rays
- Our algorithm finds overlooked signal in the knee x-ray which helps explain disadvantaged patients' higher pain
- Clinical implication: disadvantaged groups may be under-referred for surgery

Discussion

- Previous work on how machine learning methods can **increase** disparities in medicine
- We show how machine learning methods can also **reduce** disparities by detecting signal humans miss

Future work

- But what is the algorithm “seeing”?



Two-part lecture

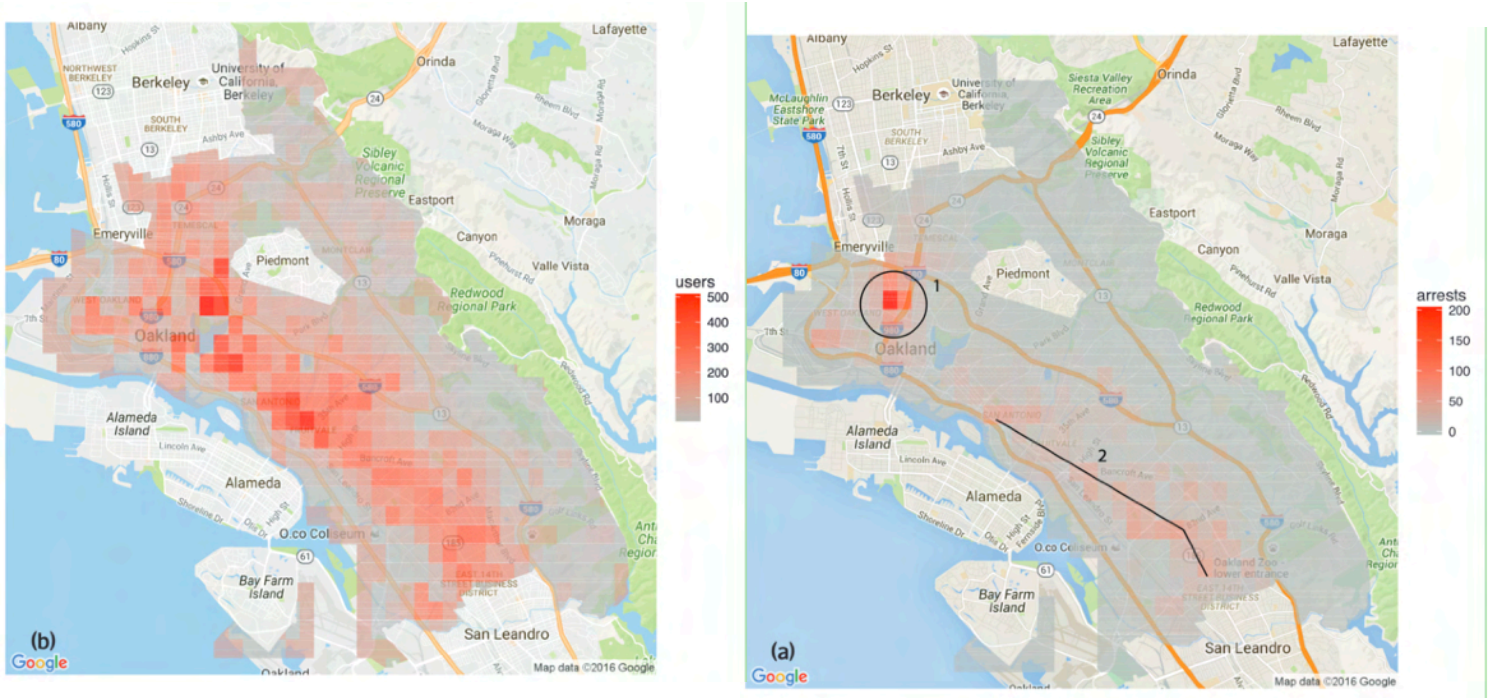
- Using AI to increase social equality
  - Pain
  - Policing
- Course wrap-up: what's next?
  - For AI?
  - For you?

Using AI to understand inequality in policing

Franchi, Zamfirescu-Peirera, Ju, and Pierson**.** *FAccT,* 2023.

Police deployments matter

- Overpolicing
- Downstream algorithmic bias



Lum and Isaac, 2016

But the police often won't release (even aggregated) deployment data

The New York Times

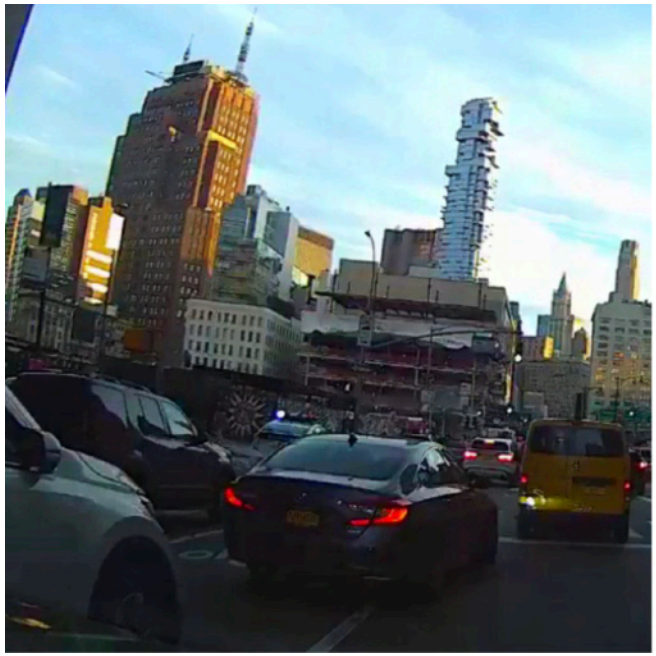
As Troopers Are Diverted, Deployment Data Remains Restricted

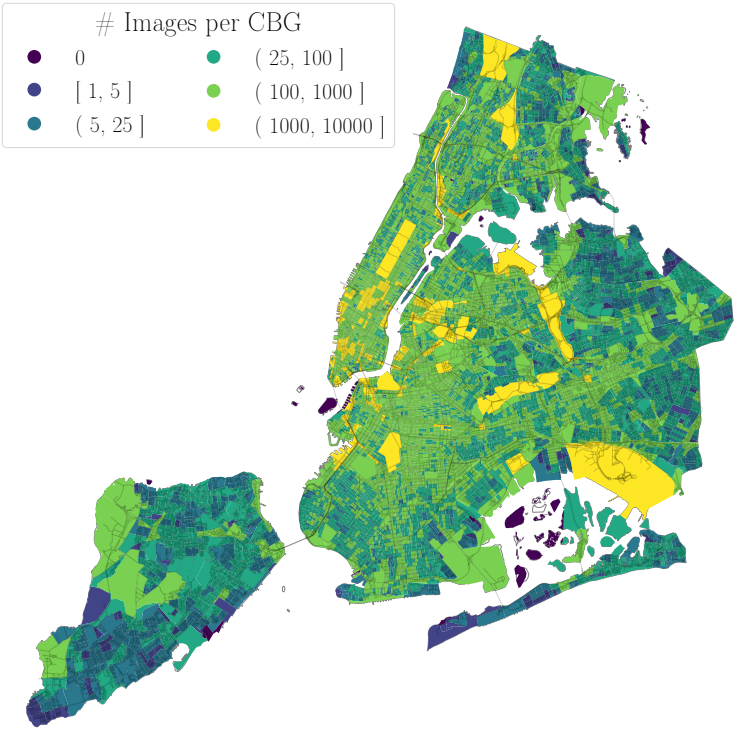
The New York Times

New York Police Urged to Fix Inequities in Deployment of Investigators

So we decided to find the police

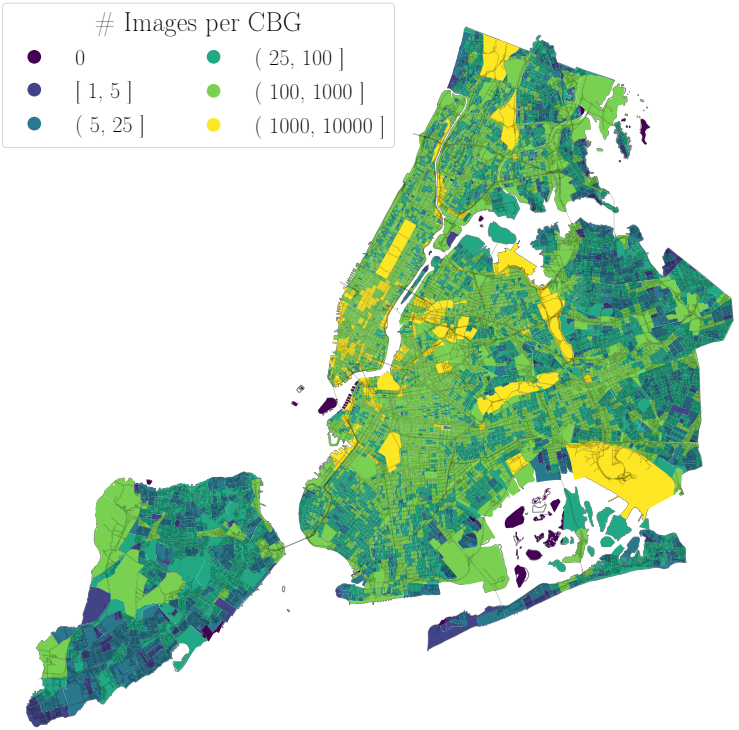
- **Data**: 25 million public street scenes throughout NYC





So we decided to find the police

- **Data**: 25 million public street scenes throughout NYC
- **Method**: train model to detect police cars from dashcam images



Estimation

- **Estimand:** Probability residents of a given group (e.g., Asian residents) have a police vehicle visible when they are on a street in their home neighborhood

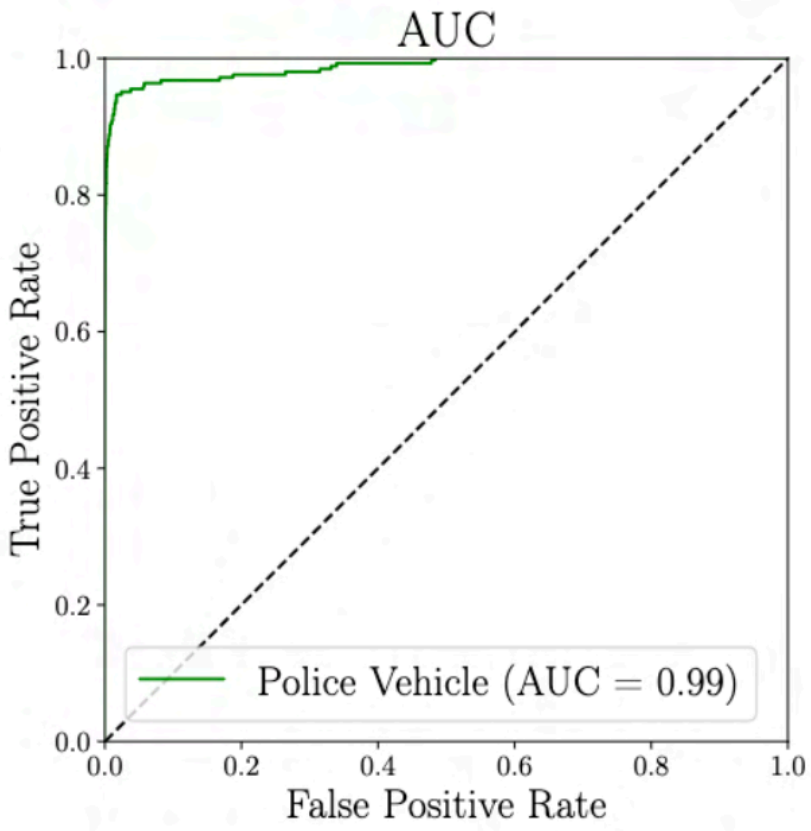
Compensating for biases

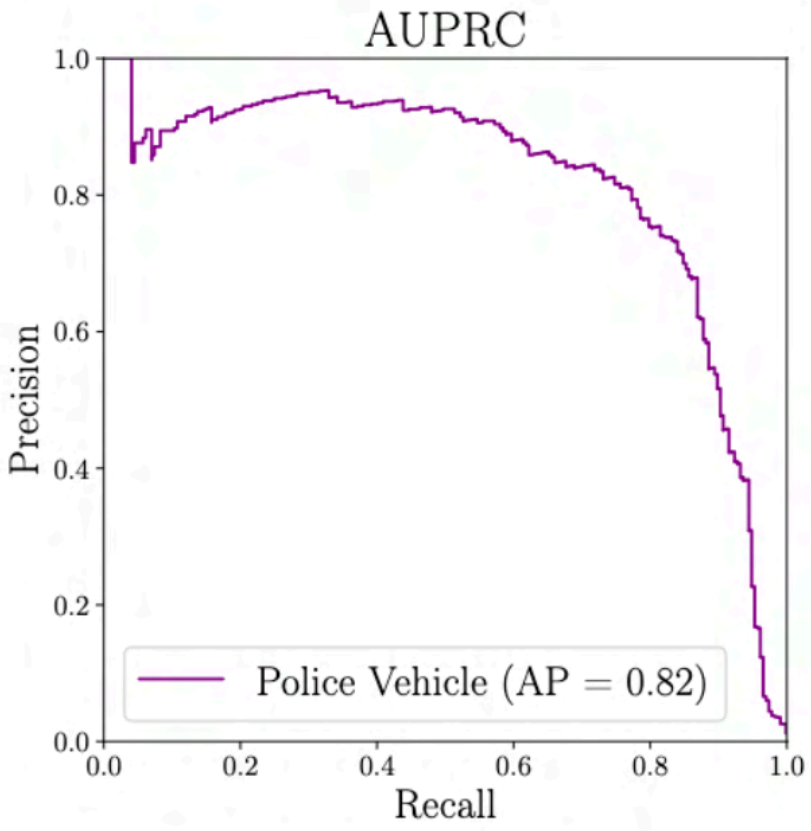
- *Non-representative locations:* reweight image distribution to match population distribution
- *Imperfect classifier*: estimate + compensate for imperfect FPR/FNR. Test for disparities in performance across subpopulations
- Some biases we can't compensate for (discuss more later)

Results

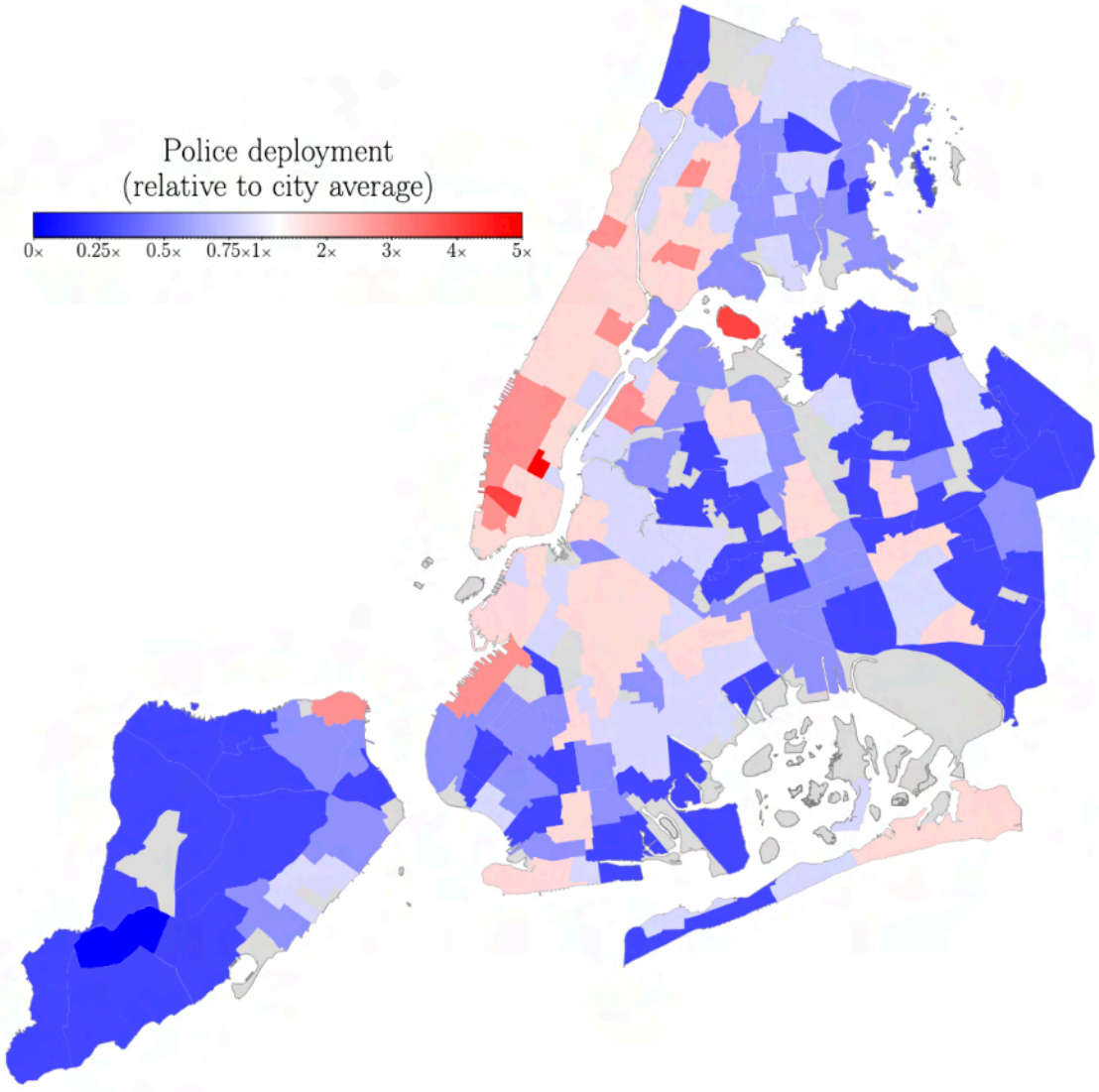
Classifier works!

Classifier works!

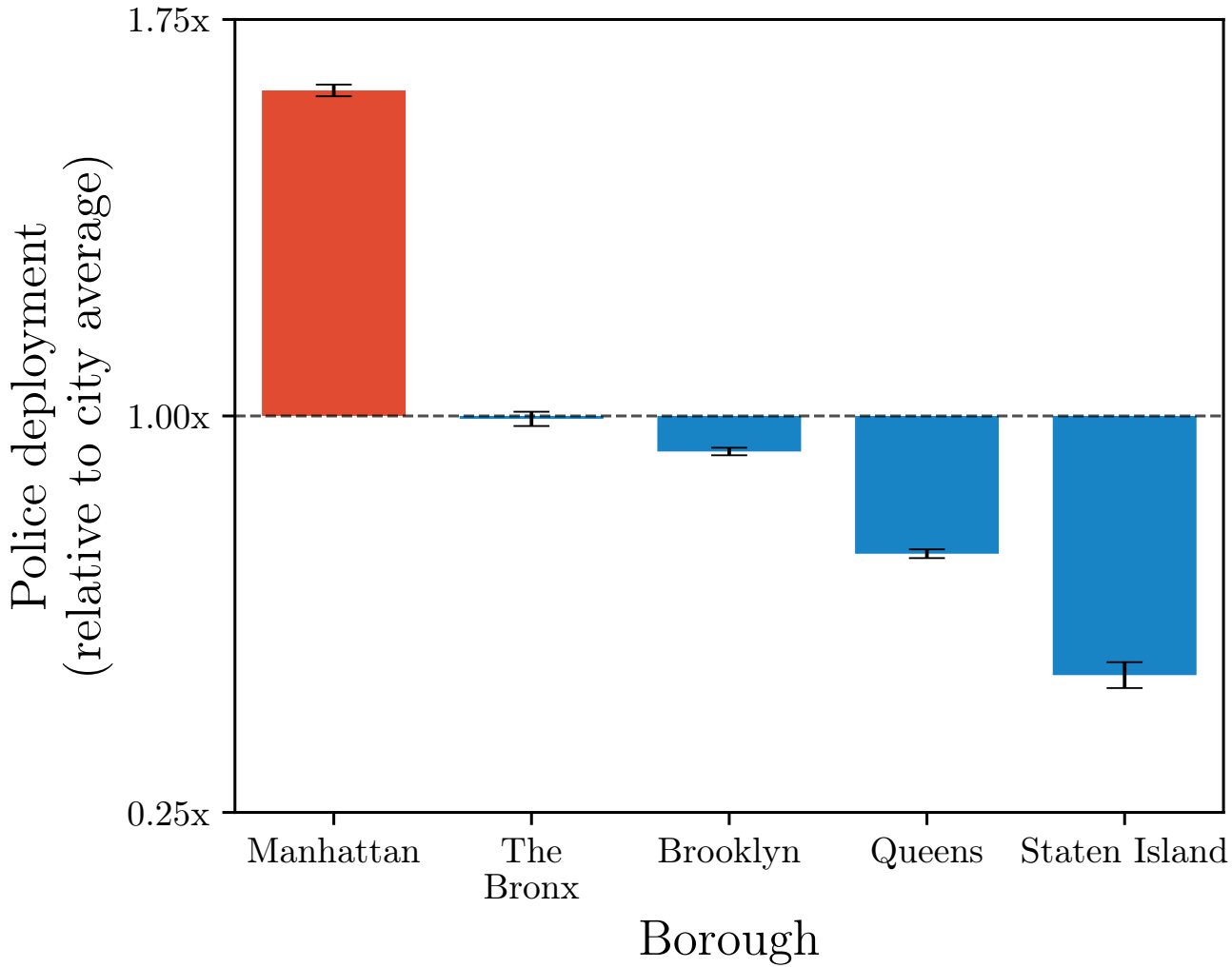




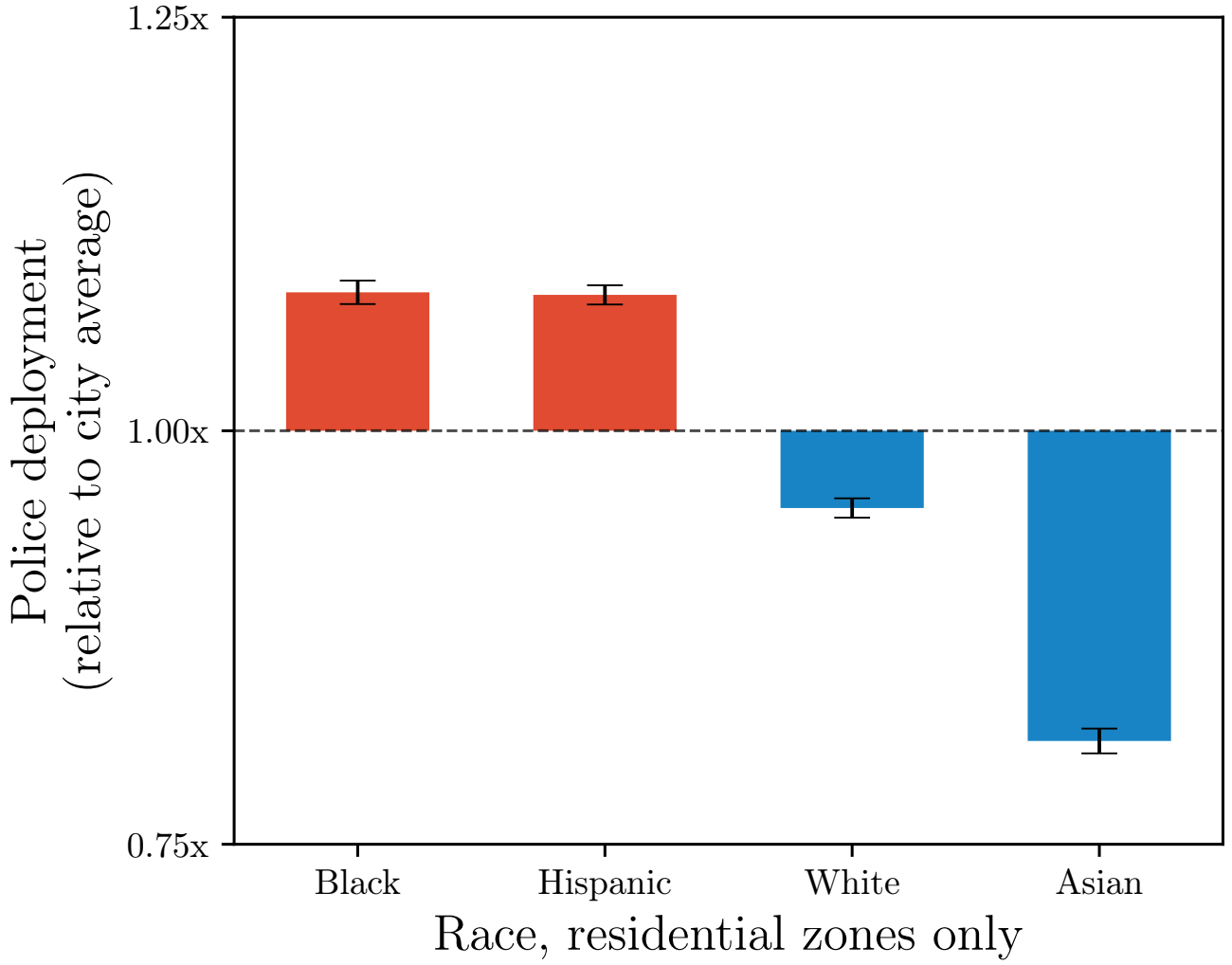
Spatial disparities



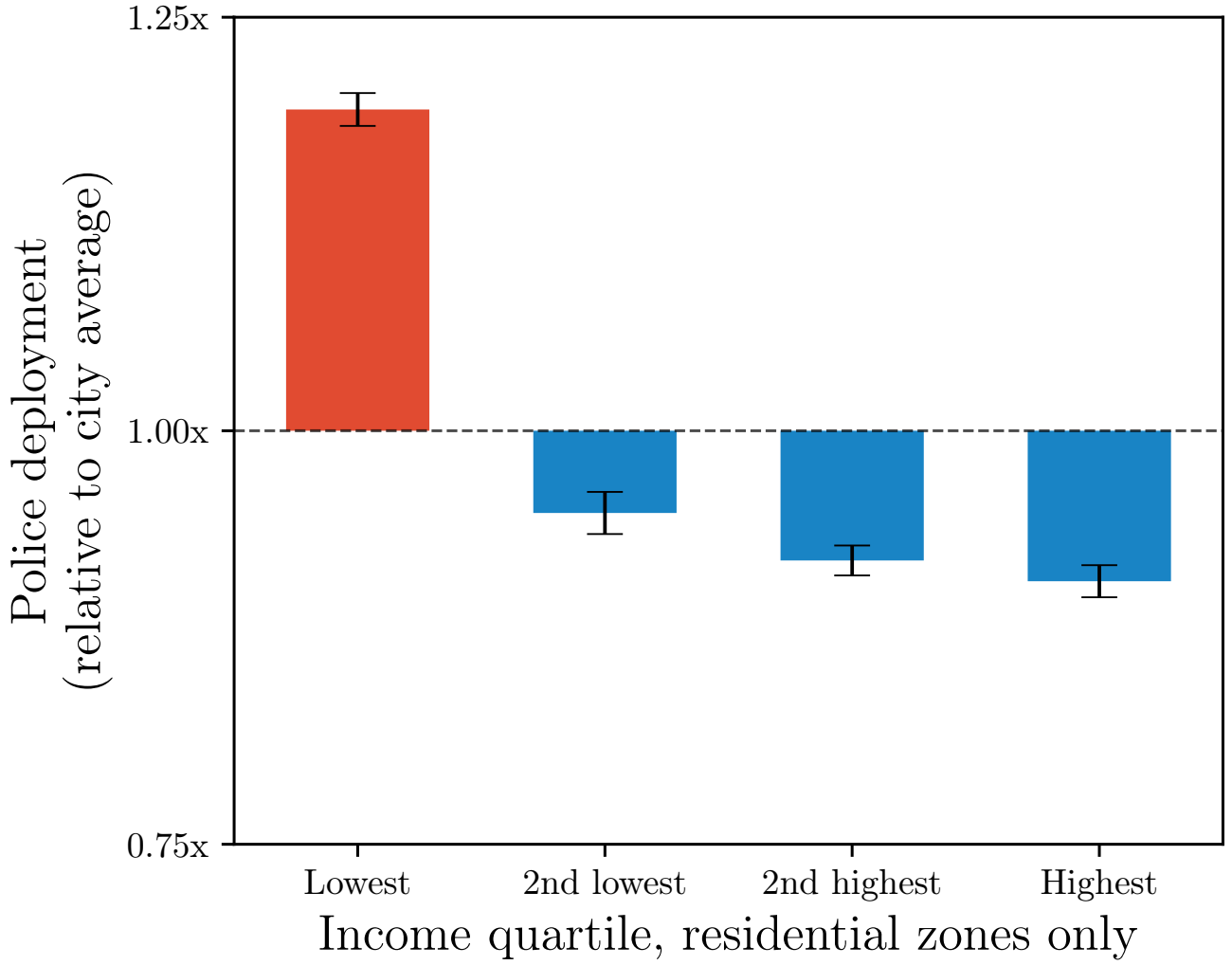
Demographic disparities



Demographic disparities



Demographic disparities



Biases we cannot compensate for

- Police vehicles are only a partial proxy for all policing activity
- Images within each Census area may be non-representative (e.g., possible holes in dataset around protests)

Conclusion

- Call for better policing data?
  - We find evidence of disparities, but this is not a realistic way for ordinary citizens to exert oversight
  - The police should just release (aggregated) deployment data

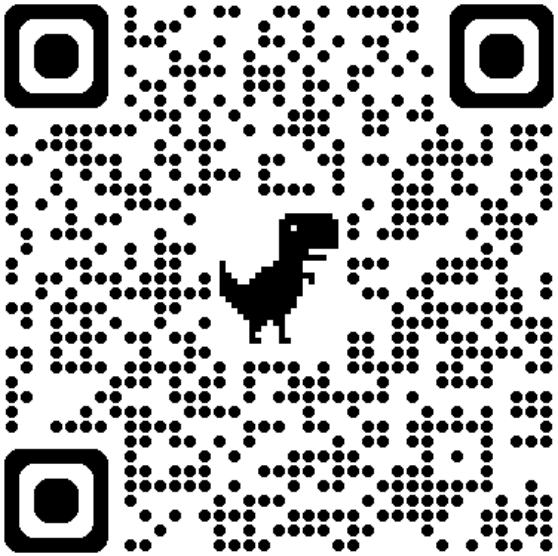
Impact

NEWS

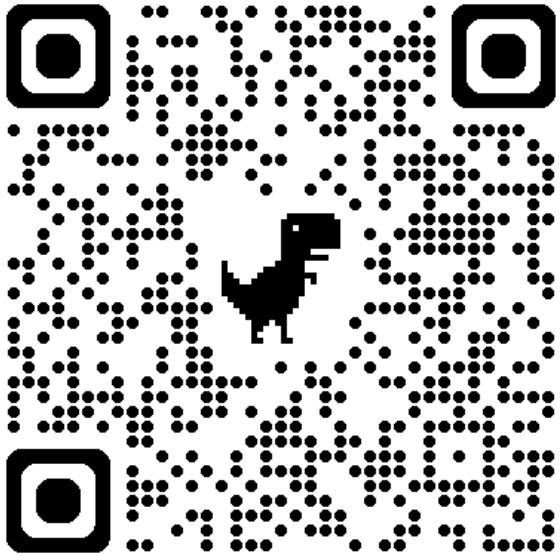
The NYPD doesn't report where it deploys police. So scientists used AI, dashcams to find out.

Brooklyn Defenders

Surveillance and Civil Rights

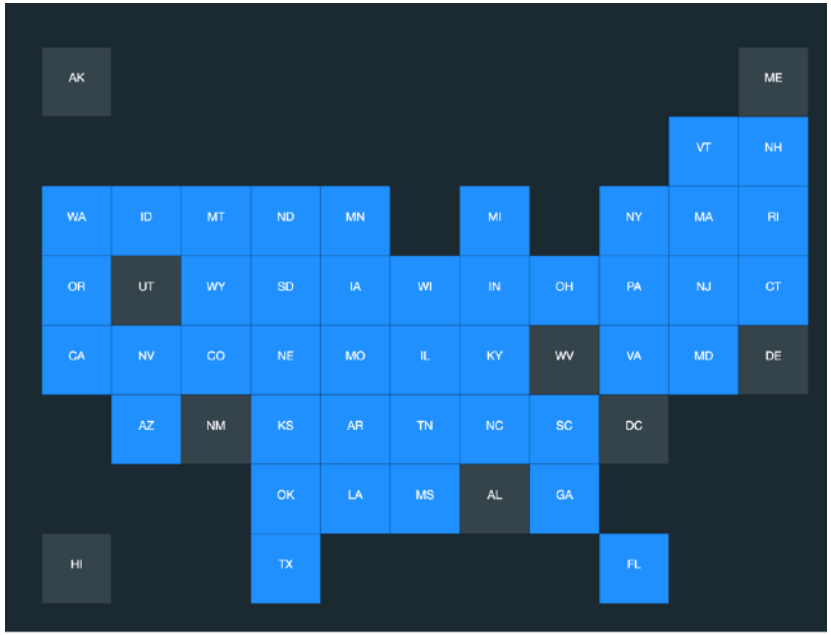


Paper



Dataset

Interested in other policing applications?



CALMATTERS

About this tool

Police Records Access Project

Search California public records about law enforcement violence and misconduct. Results are organized into cases attributed to the agency providing the records. Information about a case can change as agencies submit new or updated records.

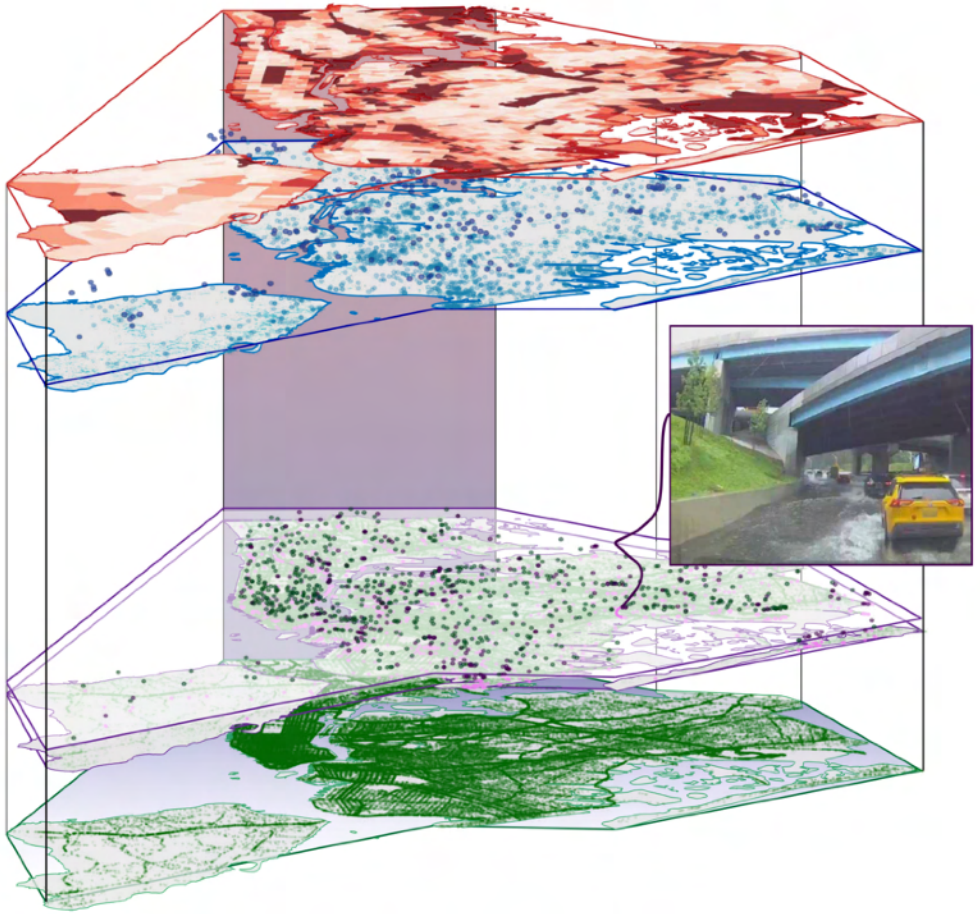
Follow-up work: flood detection

To improve flood detection and response

Validated against and integrated with external data sources…

Classified with computer vision models…

926k images from a single day of flooding…



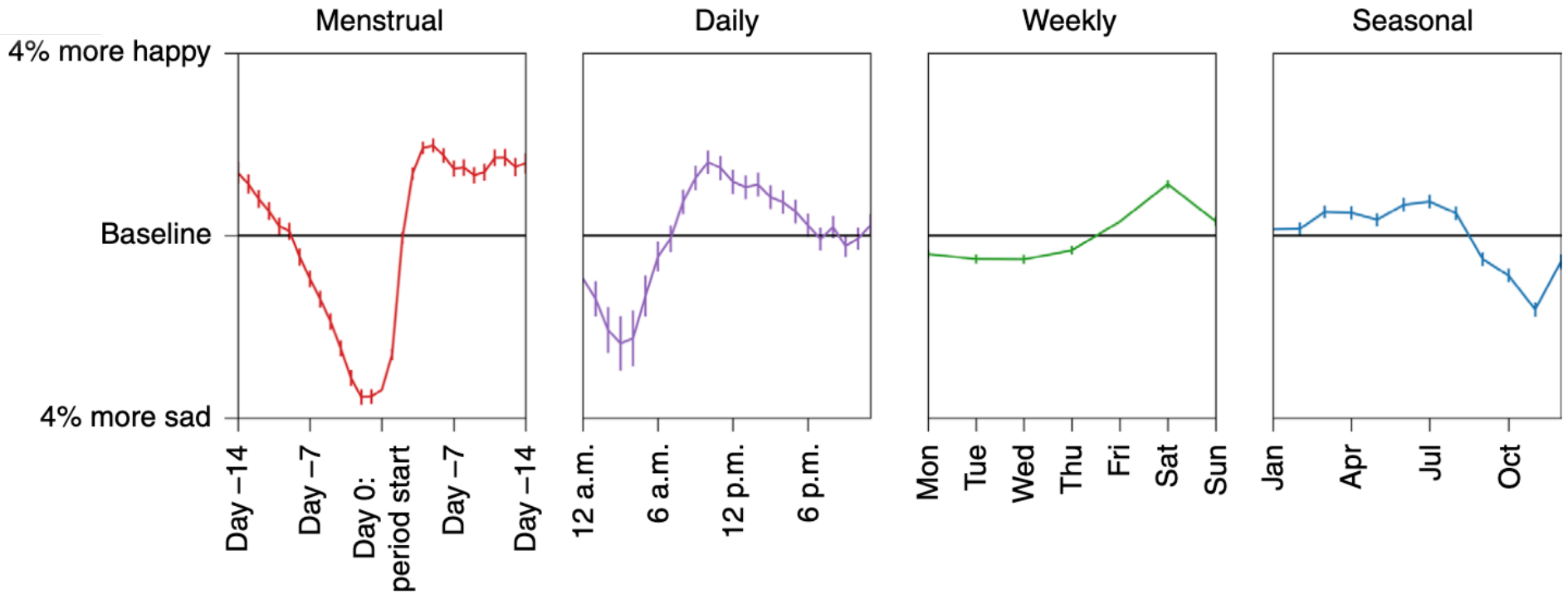
Paper

Franchi, Garg, Ju, and Pierson. Revise and resubmit, *Nature Communications,* 2025.

Zooming out, lots of other overlap between AI and social equality!

Other data sources

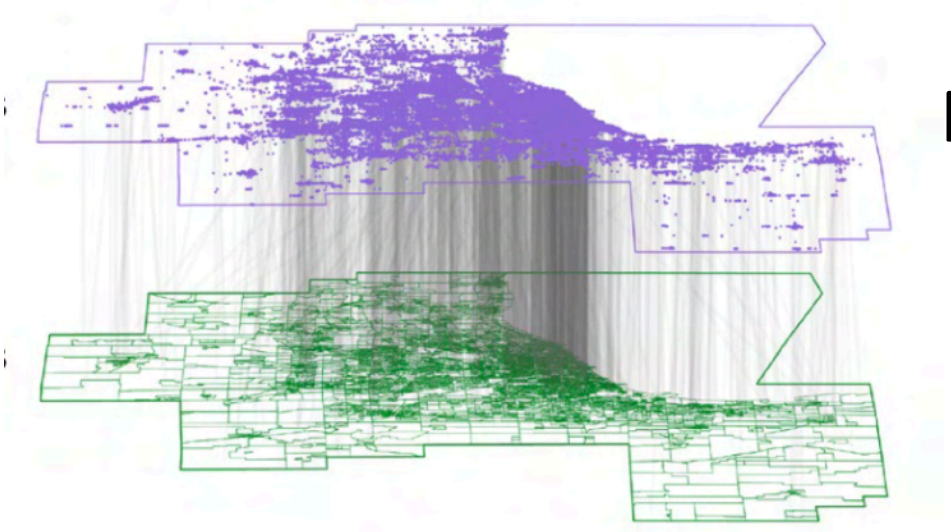
- Mobile health data



Pierson et al. *Nature Human Behaviour*, 2021.

Other data sources

- Mobile health data
- Mobility data

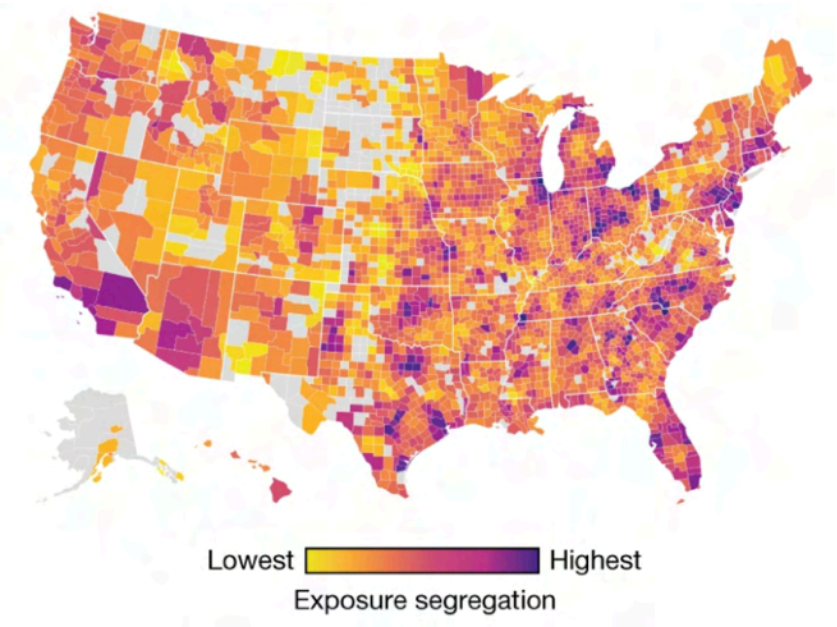


Model COVID-19 spread over mobility networks linking neighborhoods to points-of-interest

[1] Chang*, Pierson*, Koh*, et al. *Nature*, 2021.
[2] Chang, Wilson, Lewis, Mehrab, Dudakiya, Pierson, Koh et al. *KDD*, 2021 (Best paper award, applied data science track).

Other data sources

- Mobile health data
- Mobility data



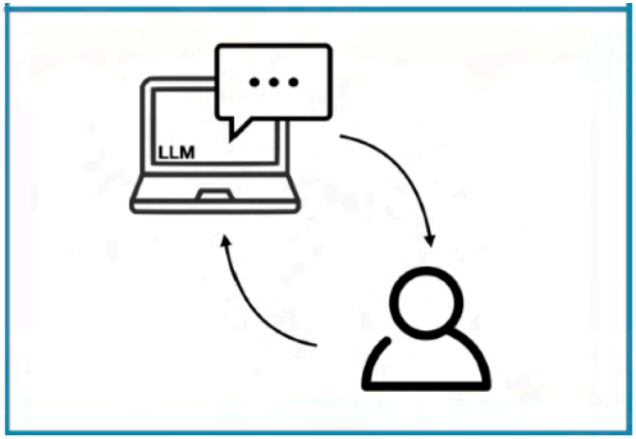
Use mobility data to more accurately measure socioeconomic segregation

Nilforoshan*, Looi*, Pierson*, et al. *Nature*, 2023.

LLMs for equality

Detecting
human biases

Creating
structured datasets
for equity research

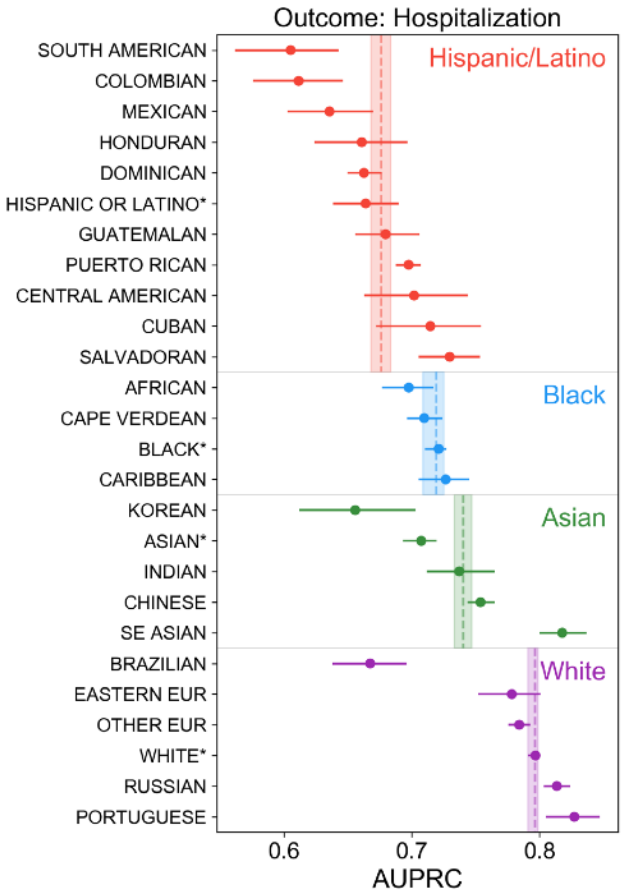


Improving equity
of access to info

Pierson*, Shanmugam*, Movva*, Kleinberg* et al. *New England Journal of Medicine AI,* 2025.

Reducing bias in decision-making

- Related topic: *algorithmic fairness*



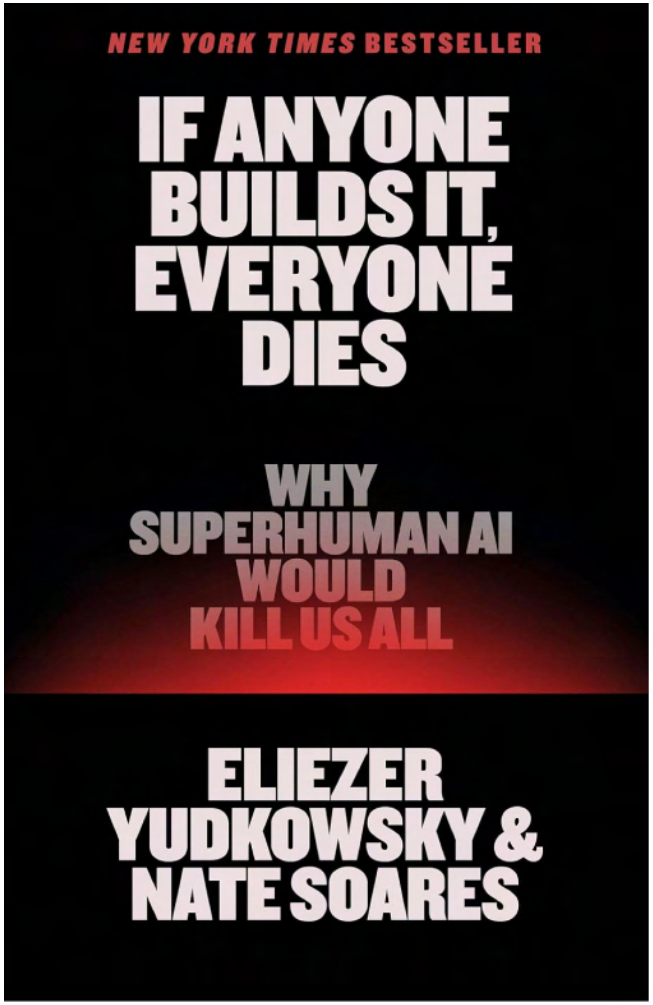
Two-part final lecture

- Using AI to increase social equality
  - Pain
  - Policing
- Course wrap-up: what's next?
  - For AI?
  - For you?

What's next for AI?

No one really knows

- It's okay to be uncertain/ nervous about what's coming next
- Be skeptical of people who make highly confident predictions —>
- With that caveat, let's look at folks making educated guesses



What do expert forecasts say?

Finding 1: experts expect sizable near-term societal effects from AI

- Median expert: by 2030:
  - 18% of US work hours assisted by generative AI
  - 7% of US electricity -> AI training and deployment
  - 20% of ride-hailing trips from autonomous vehicles
  - 15% of adults will use AI for companionship

Finding 2: lots of disagreement / uncertainty

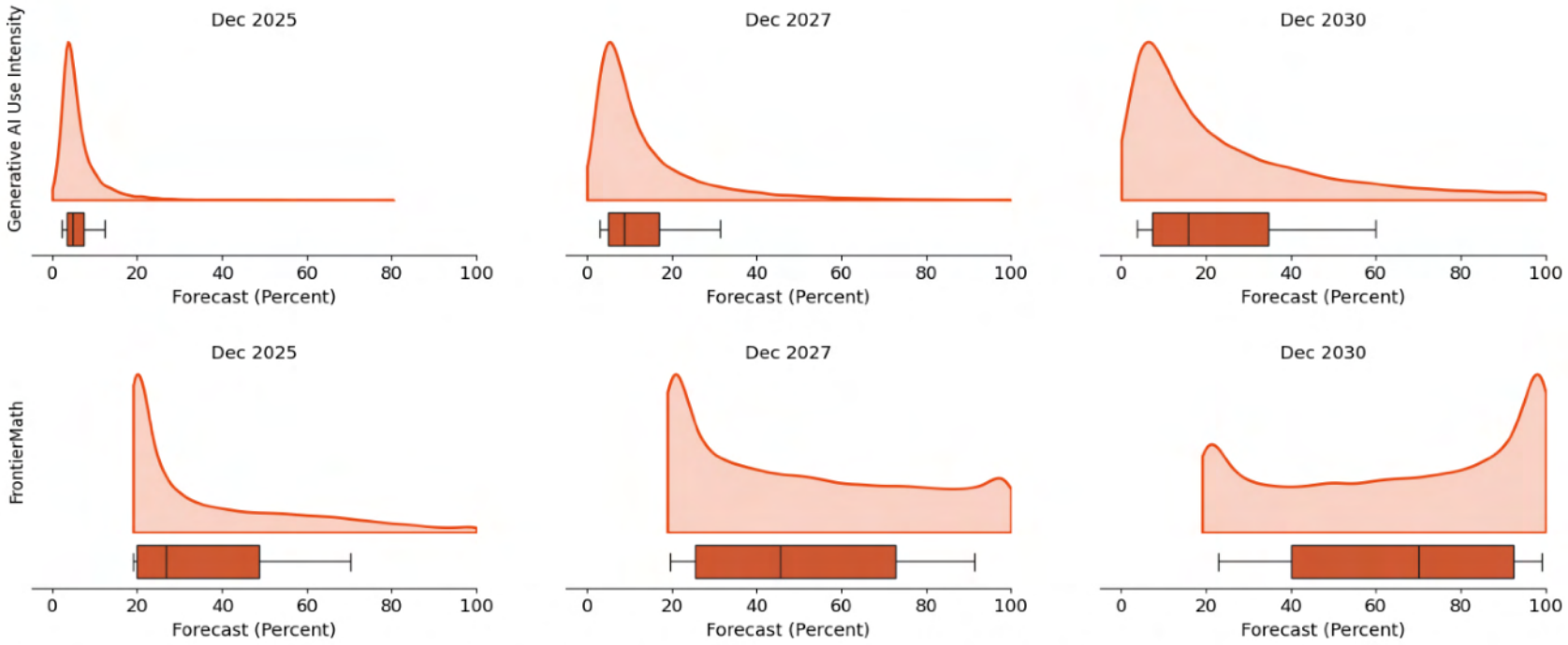
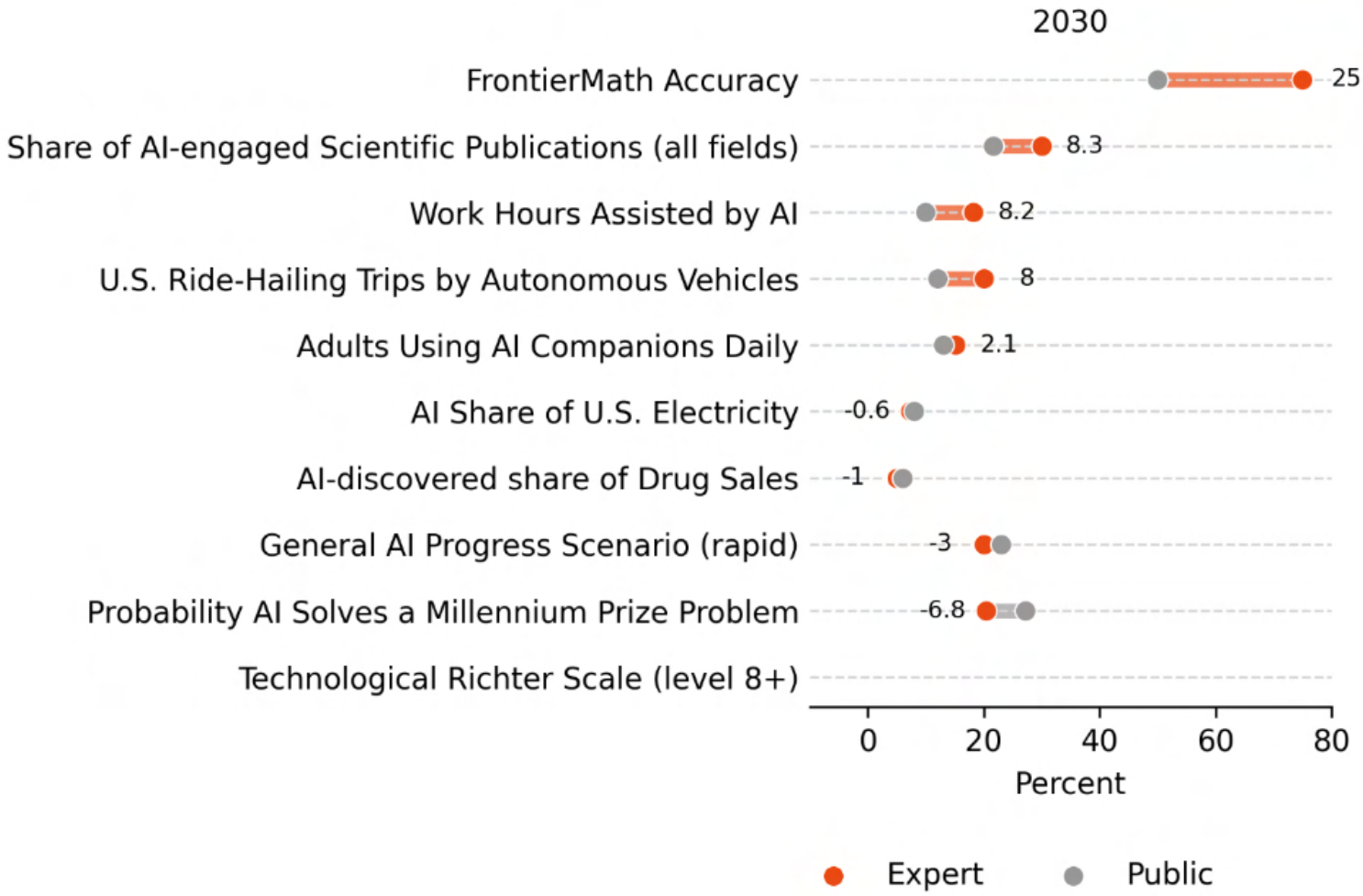


Figure: Pooled distributions for expert forecasts on *Work Hours Assisted by Generative AI* (top panels) and *FrontierMath scores* (bottom panels). These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. See

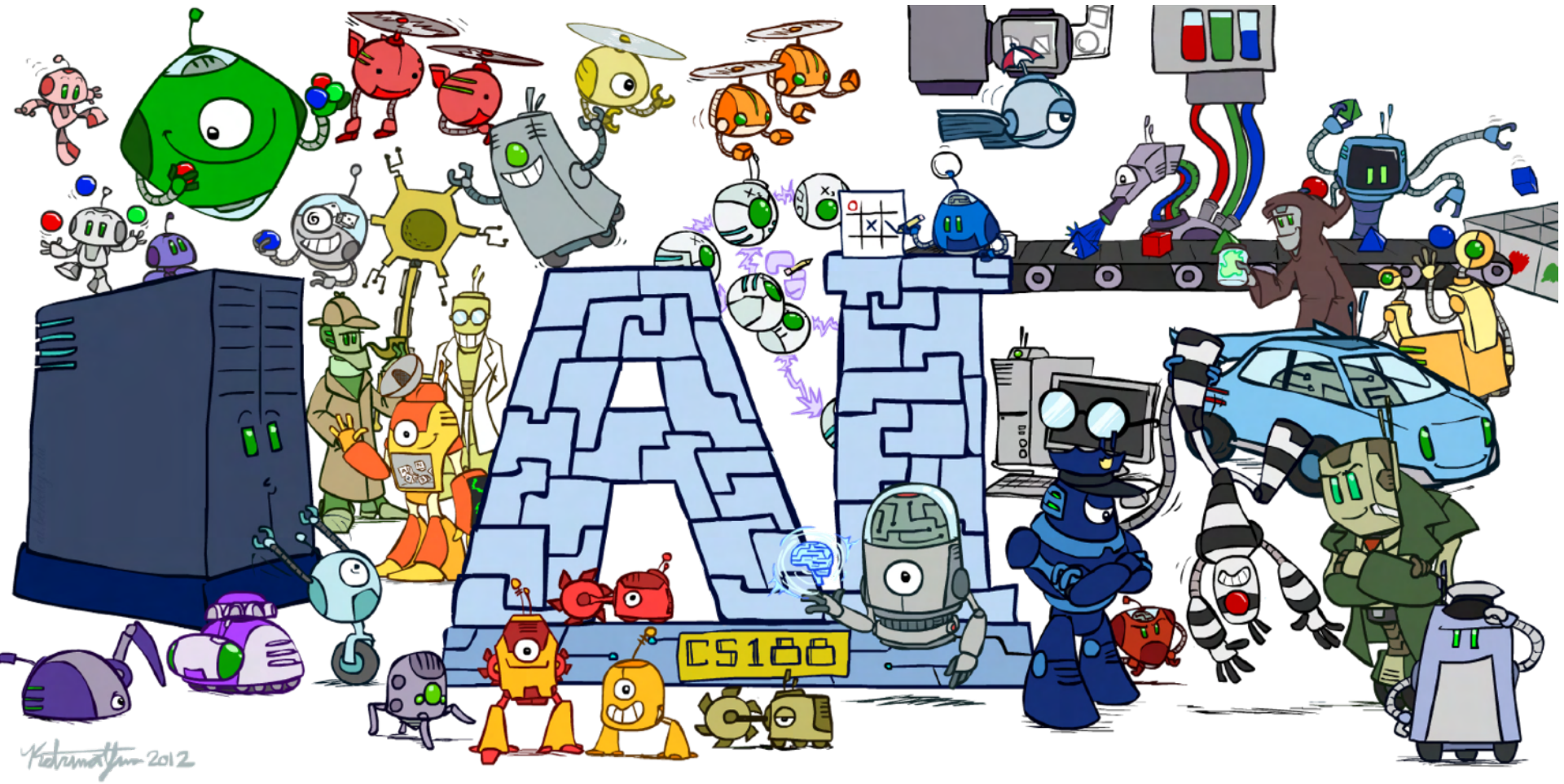
Finding 3: experts expect slower progress than prominent leaders of frontier AI labs

- Sam Altman: “I think AGI will probably get developed during [Donald Trump’s second presidential] term, and getting that right seems really important.”
- Elon Musk: “Probability that AI exceeds the intelligence of all humans combined by 2030 is ~100%.”
- Dario Amodei: January 2025: “By 2026 or 2027, we will have AI systems that are broadly better than almost all humans at almost all things.”

Finding 4: experts predict faster AI progress than the public



…so what does this all mean for me?



Ketrina Yim

CS188 Artist

In a fast-changing world, *learn to get good at learning*

I saw your TensorFlow post. Reading the whitepaper now. Very cool.

Emma

[redacted]

to me

Hi Emma,

Thanks for the email -- I'm really happy we were able to release TensorFlow!

Many of the AI tools we'll be using 10 years from now don't even exist yet. You will have to constantly master new tools. Not knowing what the hell you're doing is part of the job.

Continually experiment with AI tools… but don't turn off your brain

- This ecosystem is changing very fast. Keep experimenting to find what tools work for you
- Do not let your critical thinking skills atrophy
  - E.g., if you're using AI to help write code, you should be able to explain it, debug it, and (perhaps more slowly) write it yourself
- If all you can do is copy-paste model output, your boss can do that herself

#1 rule: Do **A LOT** of computer science

- When I was 20 I was a physics undergrad
- When I was 30 I was a computer science professor
- Between 20 and 30, I spent 10,000-20,000 hours doing computer science. This went into:
  - 70 computer science blog posts
  - 28 academic papers
  - 20-30 popular-press publications
  - Coursework

#1 rule: Do **A LOT** of computer science

- But how?

Helps to find something which doesn't feel like work

- The thing you want to wake up early to work on, the thing you stay up late at night thinking about (my blog: *Obsession* with Regression)
- Heuristic I use: what papers do I get excited to read in my free time?
- (No you will not always feel joyful + consumed by your work - at least I don't!)
- (And work-life balance is essential even if you like your work)

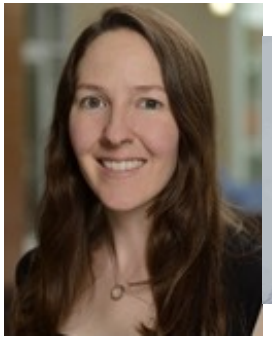
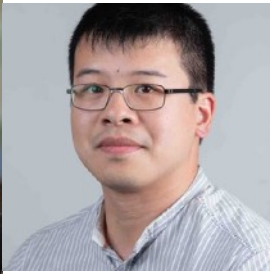
Coursework: what comes next?

- Machine learning: cs189, cs182, stat154, ind. eng. 142
- NLP: cs183, cs288
- Data Science: data100, data 102
- Data Ethics: data c104
- Probability: ee126, stat134
- Optimization: ee127
- Cognitive modeling: cog sci 131
- Machine learning theory: cs281a/b
- Computer vision: cs280
- Deep RL: cs285
- Special topics: cs194- or cs294-?
- … and more; ask if you're interested

Research

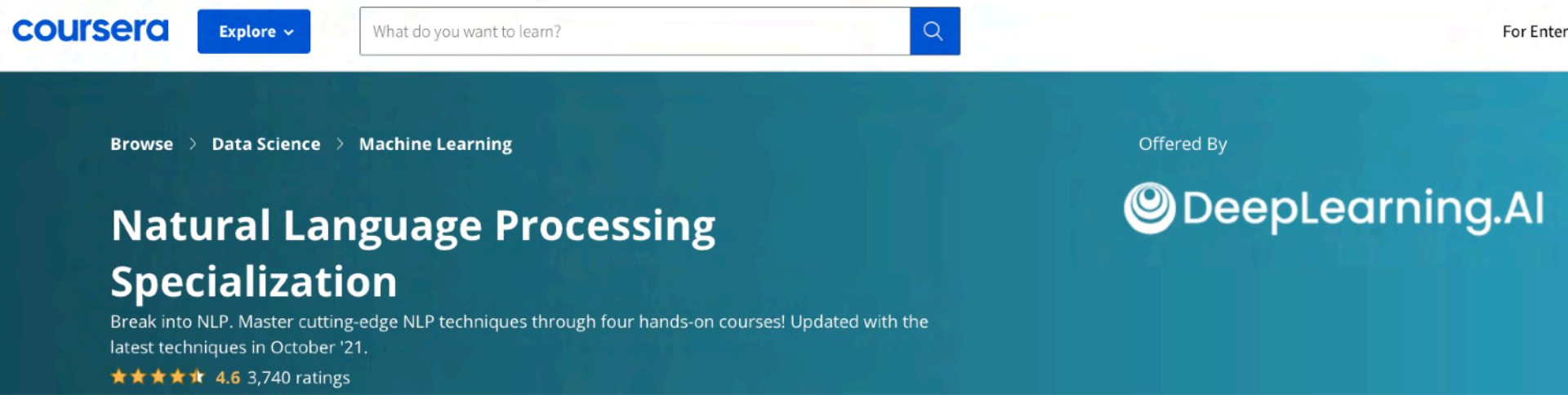
- Hugely valuable to me
- The best mentors don't have to be professors
- https://bair.berkeley.edu/people/faculty
- Many faculty pages have ways to apply to lab
- If you don't get it first time, apply next semester!



Online courses

- Many many free (or cheap) resources for learning data science - check out Coursera, EdX, and many universities also put lectures online
- So you want to learn Natural Language Processing…?



Other ways to get AI experience

- Jobs / internships
- Blogging
- Kaggle competitions

Final point: the human impacts of AI are increasing

Topics, Authors, and Institutions in Large Language Model Research: Trends from 17K arXiv Papers

Rajiv Movva, Sidhika Balachandar, Kenny Peng, Gabriel Agostini, Nikhil Garg, Emma Pierson

Large language models (LLMs) are dramatically influencing AI research, spurring discussions on what has changed so far and how to shape the field's future. To clarify such questions, we analyze a new dataset of 16,979 LLM-related arXiv papers, focusing on recent trends in 2023 vs. 2018-2022. First, we study disciplinary shifts: LLM research increasingly considers societal impacts, evidenced by 20x growth in LLM submissions to the Computers and Society sub-arXiv. An influx of new authors -- half

Final point: the human impacts of AI are increasing

Final point: the human impacts of AI are increasing

Can AI help beat poverty? Researchers test ways to aid the poorest people

Silicon Valley enabled brutal mass detention and surveillance in China, internal documents show

Microsoft blocks Israel's use of its technology in mass surveillance of Palestinians

The N.Y.P.D. Is Teaching America How to Track Everyone Every Day Forever

Using A.I. to Detect Breast Cancer That Doctors Miss

How A.I. Is Revolutionizing Drug Development

Can A.I. Be Blamed for a Teen's Suicide?

Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots

How will you ensure those impacts are positive?