Midterm next Friday

- **Time:** 5/1, 9:30am sharp to 10:20am (50 minutes)
  - Please be seated 5 minutes beforehand. Bring your Husky ID.
- **Locations:**
  - **CSE2 G20** (usual classroom): Sections AA, AB, AC, AD
  - **CSE2 G10** (next door)**:** Sections AE, AF
- **Topics:** Everything up to and including Lecture 12 on 4/24 (this lecture on prediction pitfalls)
- **Format:** Sample exams posted
- **Cheat sheet:**
  - You may bring one 8.5x11 inch sheet of paper (can use front & back).
  - It must be handwritten.

Prediction Pitfalls

Natasha Jaques

UNIVERSITY of WASHINGTON

Interpreting coefficients

Consider a linear model $\hat{w} = \text{argmin}_w \sum_{i=1}^{n} \left(y_i - w^\top x_i\right)^2$

Claim: $\hat{w}_i > \hat{w}_j$ means feature i is more important than feature j

FALSE

What if features have different scales?

- Sqft have a mean of 2500 → w_i is $/sqft
- # baths have a mean of 2 → w_i is $/bath

How to fix? Normalize features to have mean and unit variance (z-score)

Interpreting coefficients

Consider a linear model $\hat{w} = \text{argmin}_w \sum_{i=1}^{n} \left(y_i - w^\top x_i\right)^2 + \lambda \|w\|_1$

with normalized data

Claim: $\hat{w}_i = 0$ means feature i has no predictive power for y

FALSE

Feature i might be correlated with y but redundant with feature j

e.g. area in square feet and area in square meters

Interpreting coefficients

Consider a linear model $\hat{w} = \text{argmin}_w \sum_{i=1}^{n} \left(y_i - w^\top x_i\right)^2$

Claim: $\hat{w}_i = 90{,}000$ and the ith feature = #fireplaces. If I add 10 more fireplaces, I can expect to sell my house for $900,000 more!

FALSE

(1) Distribution shift: we've never seen a house with 10 fireplaces in the training data, so we don't have accurate predictions for this range. It's OOD: out of distribution

(2) Correlation ≠ causation

(3) Linear model, non-linear relationship

Generalization

Say we've trained a model to interpret medical x-rays using data from a few hospitals. We randomly split the data into train/validation/test splits.

```
# Train on hospitals 1, 2, 3 → Test on hospital 4
```

Claim: The test set performance is always a good indicator of how our model will do if we deploy this model in a new hospital.

FALSE

Distribution shift: new hospital could have different equipment, demographics, staff training, location…

Generalization

Say we've trained a model to interpret medical x-rays using data from a few hospitals. We randomly split the data into train/validation/test splits.

Train on hospitals 1, 2, 3 → Test on hospitals 1, 2, 3

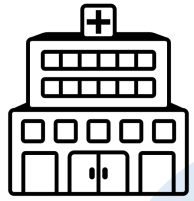
Claim: The test set performance is always a good indicator of how our model will do if we deploy this model in the **same hospitals.**

FALSE

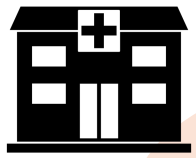
Distribution shift: temporal shift. Conditions can change in our hospital by the time we deploy our system! e.g. new training procedures, new staff, new equipment, etc.

My particular hobby horse: the world is non-stationary!

Domain or distribution shifts



Training distribution



Test distribution

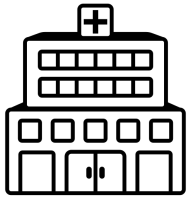
Very common in practice!!

Case study: EPIC's sepsis model

EPIC: large US healthcare company

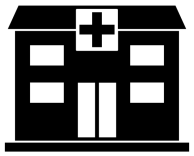
Early warnings for sepsis

Case study: EPIC's sepsis model



Trained on 3 hospitals

Distribution shift



Deployed on 100s of **other** hospitals

NEJM Journal Watch

EPIC's Sepsis Model Is Not Ready for Prime Time

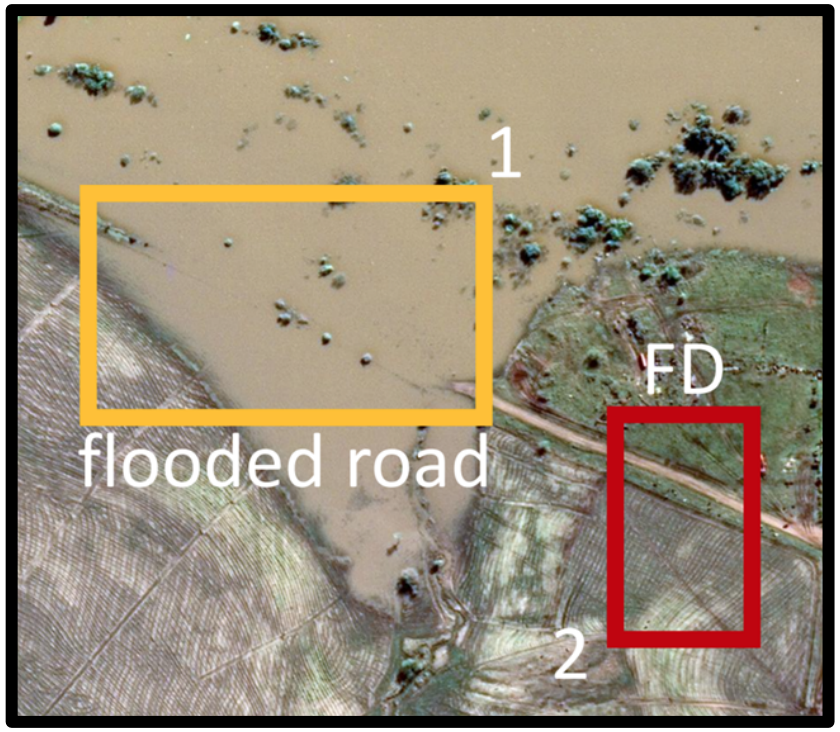
The system missed sepsis 67% of the time… The vast majority of alerts were false positives.

How to fix?

- Collect more training data
- Before deploying in a new hospital collect some small amount of test data and adapt model (domain adaptation)
- …. <your idea here>

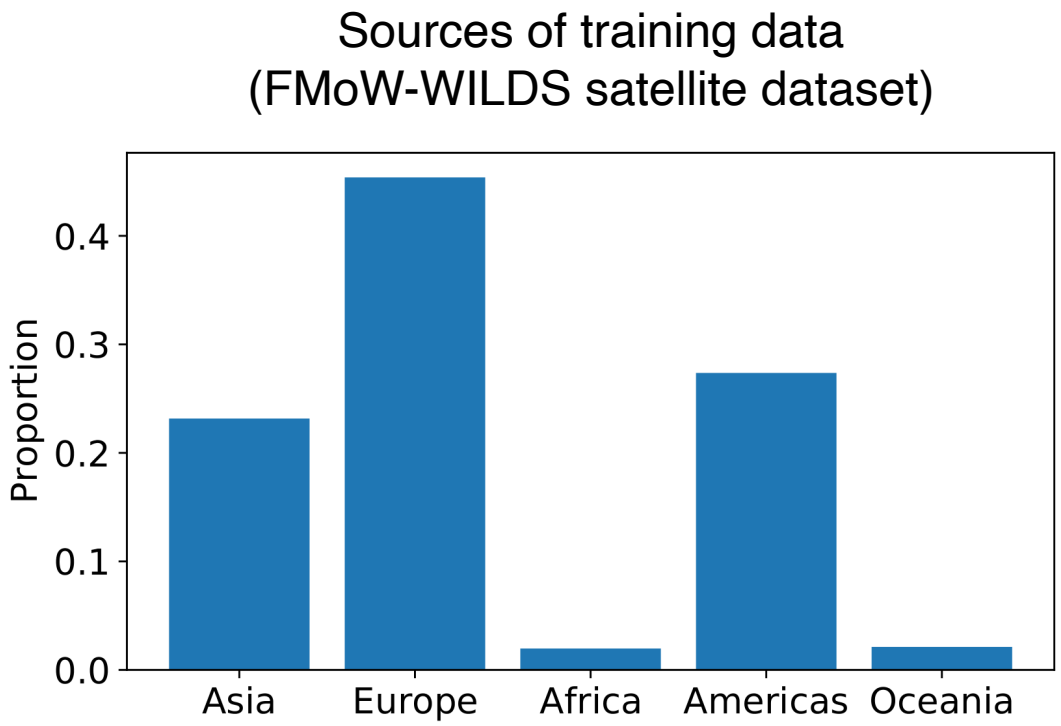
Article: Calderon, NEJM 2021
Study: Wong et al., JAMA Internal Medicine, 2021

Case study: Disaster prediction to allocate resources



Label 1 = Disaster (flood, fire)

Label 2 = False Discovery (FD)



Test accuracy on Americas: **55.7%**

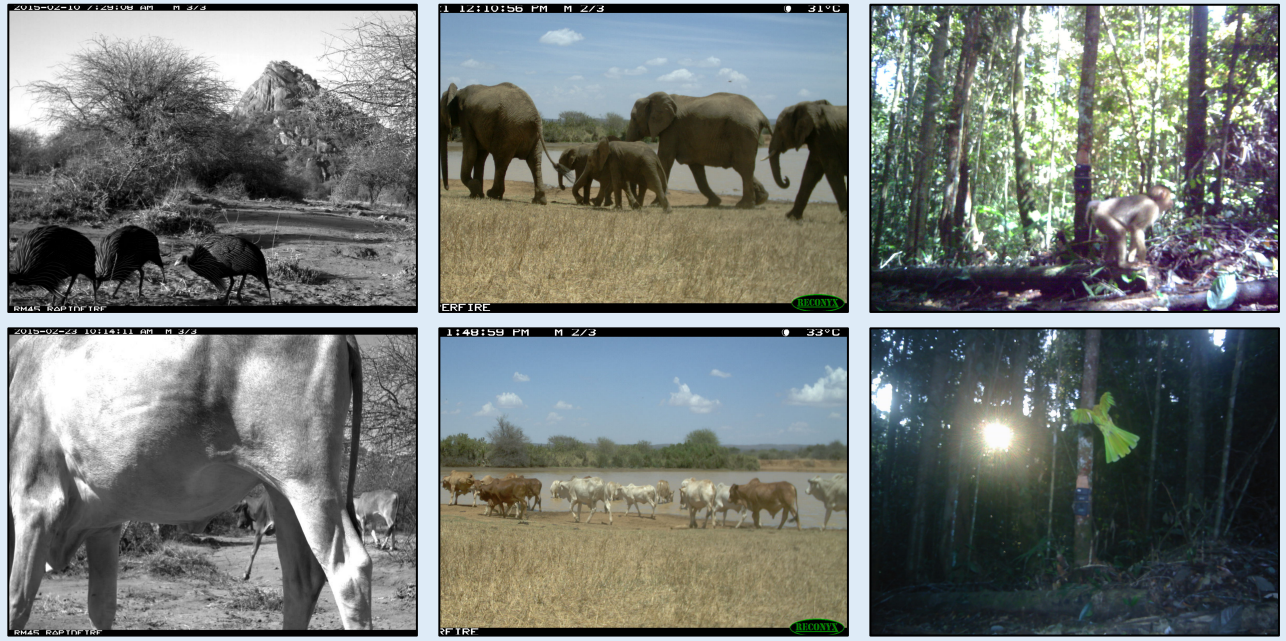
Test accuracy on Africa: **32.3%**

Dataset: Christie et al., 2018
Plot: Koh* and Sagawa* et al., 2021

Training data

WILDS data

Camera 1 | Camera 2 | … | Camera 245



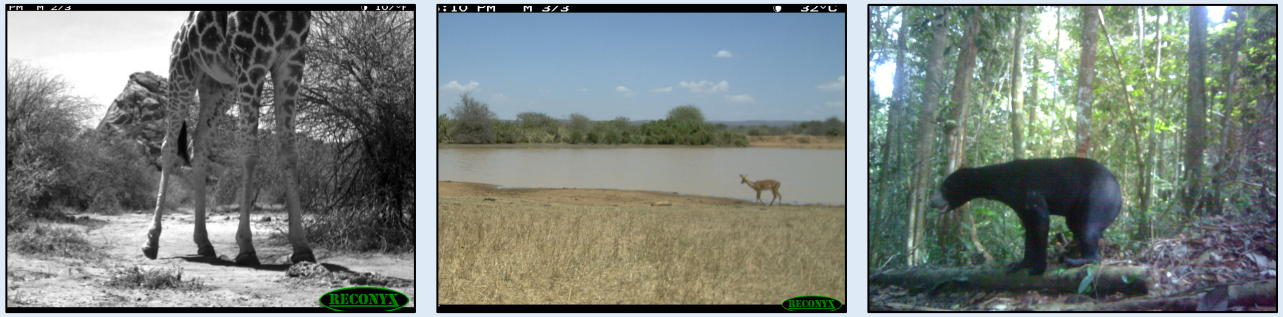
Out-of-distribution (OOD) test data

Camera 246

…

Control: In-distribution (ID) test data

Camera 1 | Camera 2 | … | Camera 245



Macro F1

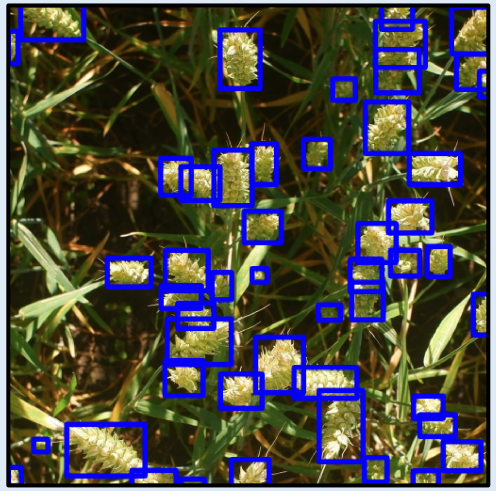
ID 47.0% **-16.0%** OOD 31.0%

[Beery et al., 2020; Koh et al., 2021]

Training data

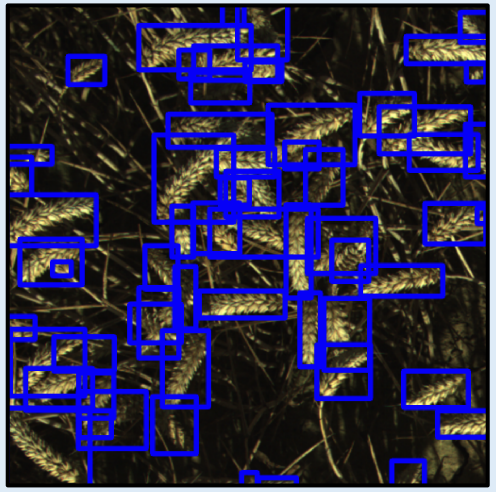
Belgium

France



…

Norway



OOD test data

United States

…

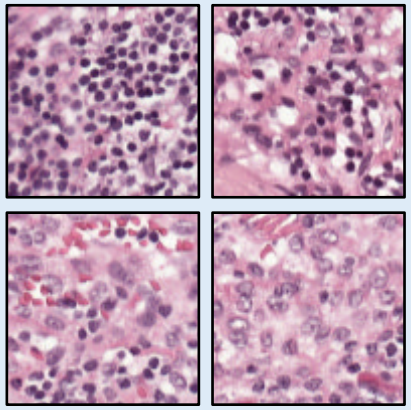
Average accuracy

ID 63.3% **-13.7%** OOD 49.6%

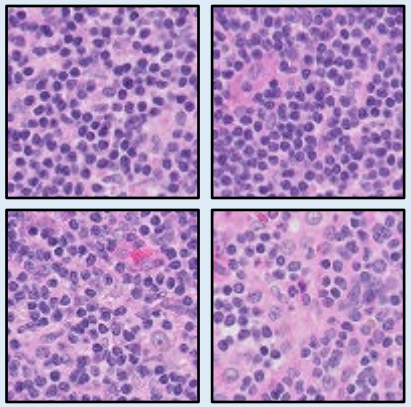
[David et al., 2021; Koh et al., 2021]

Training data

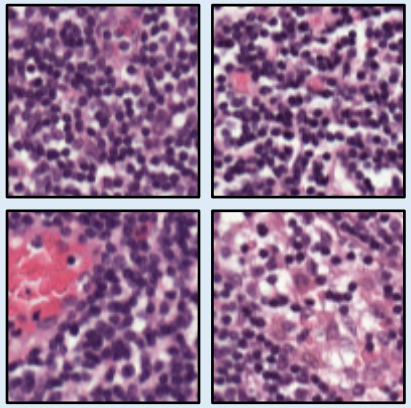
Hospital 1



Hospital 2

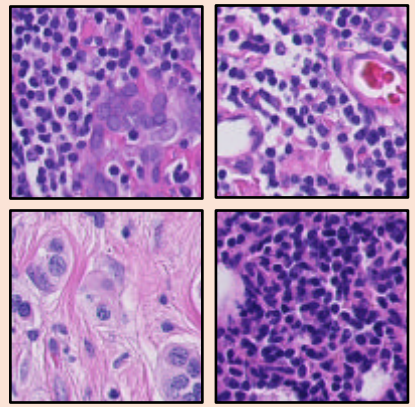


Hospital 3



OOD test data

Hospital 4



Average accuracy

ID 93.2% → **-22.9%** → OOD 70.3%

Drop in performance to OOD data very common, even for powerful models

Models often memorize or fit spurious features. How can we really generalize?

[Bandi et al., 2018; Koh et al., 2021]

Moral of the story

- Never treat the data as a black box
- Always understand the assumptions that went into the data

Biases in the data

Can we fix? Weight samples differently to counter bias. Importance sampling.

In the early 2010s, the city of Boston wanted to repair potholes but wanted to allocate resources as efficiently as possible. So they released a smart phone app that automatically detects potholes via accelerometer data and sends back the GPS coordinates.

Claim: By fixing the potholes that are reported most frequently, resources are allocated to minimize the greatest number of total interactions with potholes.

Assumption: amount of reports $\propto$ actual traffic. Is it true?

FALSE

Reporting bias: younger, wealthier people report more, get more resources

Crawford "The Hidden Biases in Big Data" https://hbr.org/2013/04/the-hidden-biases-in-big-data

Biases in the data

As of 2019, health risk-prediction tools are applied to ~200M people in the US each year. These predict Y from X where:

Y = healthcare utilization

X = patient information

Assumption: healthcare utilization is related to risk. More risk → more utilization. True?

Claim: This can accurately predict which patients' health are most at risk.

FALSE

Actually, having more $$ → more utilization

So what will this do?

Allocate more resources to the people with the most resources.

[Obermeyer et al., 2019]

Biases in the data

In 2015, Amazon trained a ML model to predict Y from X where

X = resume

Y = suitability for the job (hiring decision / job performance)

Claim: By using a data-driven process, we can avoid biases of human resume screeners.

FALSE

The model inherited the historical biases of the human decisions it was trained on.

e.g. if X has features indicating applicant was a woman (woman's college, Grace Hopper conf, etc), more likely to predict Y = no hire

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Is it sufficient to remove sensitive features?

In 2015, Amazon trained a ML model to predict Y from X where

X = resume

Y = suitability for the job (hiring decision / job performance)

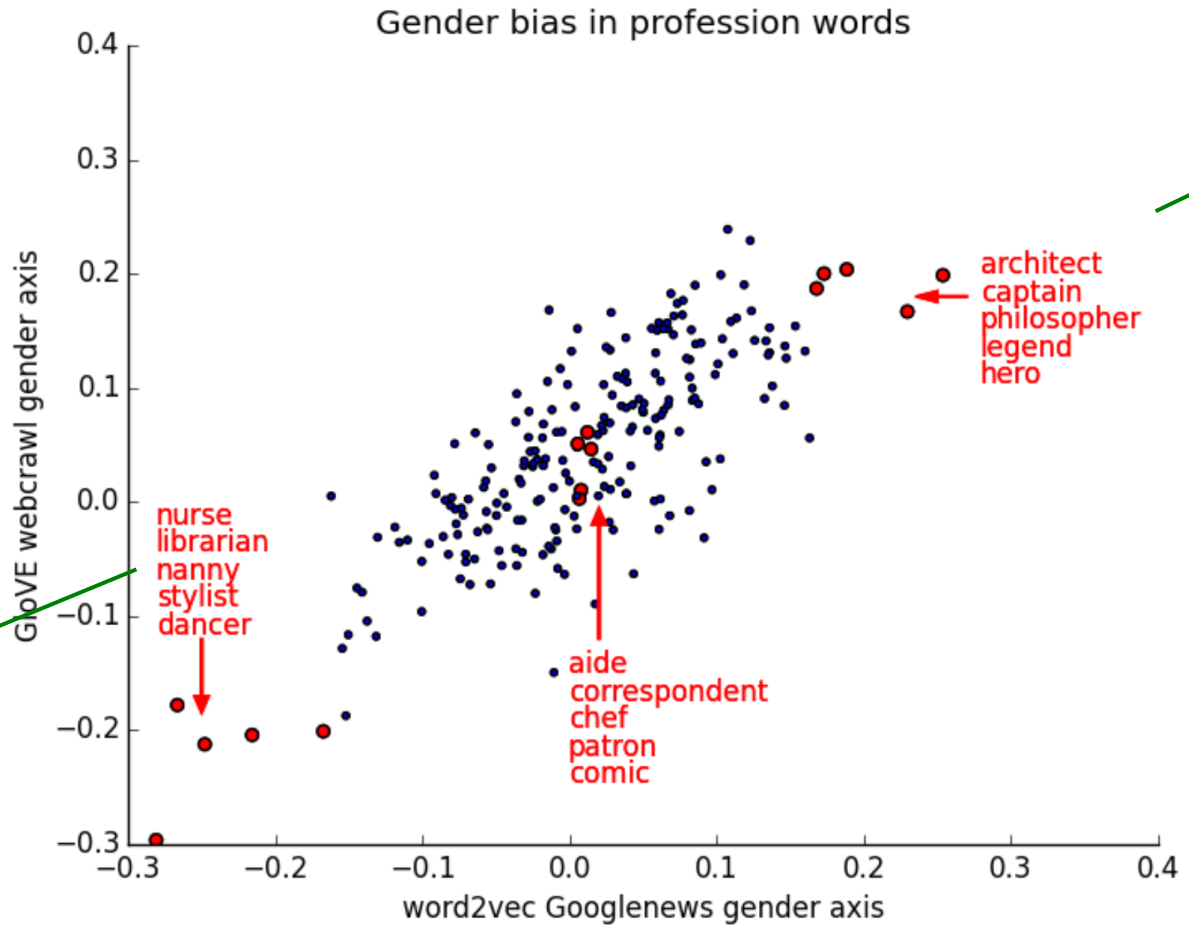
Claim: By removing applicants' demographic info from their resume, we can train models that are demographically unbiased.

FALSE Model can use other features that correlate with identity: college, zip code, etc.

So what can we do? Retain knowledge of gender/race, correct for bias.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

Stereotypes in language models



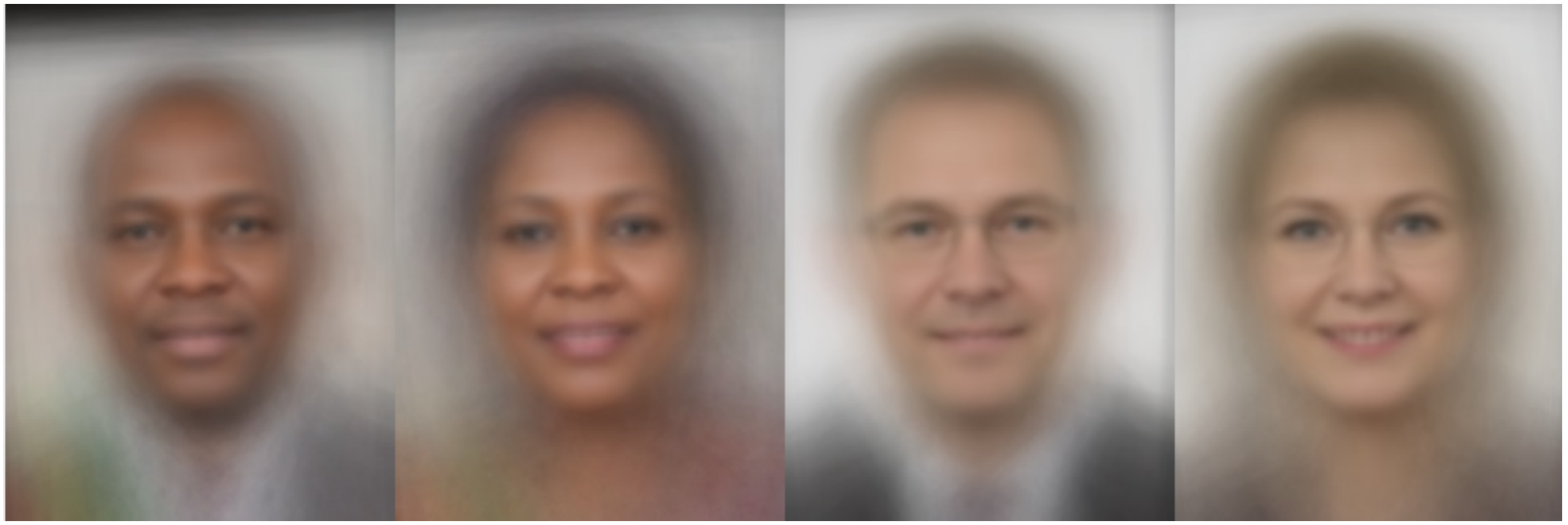
Careers most associated with the concept “man”

Careers most associated with the concept “woman”

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Quantifying and reducing stereotypes in word embeddings. *arXiv preprint arXiv:1606.06121*. https://arxiv.org/abs/1606.06121

Gender and racial bias in face recognition

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%



Joy Buolamwini

Timnit Gebru

http://gendershades.org/

Why do you think this is biased?

One reason is data

Applies to LLMs too

Wrongfully Accused by an Algorithm

…A faulty facial recognition match led to a Michigan man's arrest for a crime he did not commit.

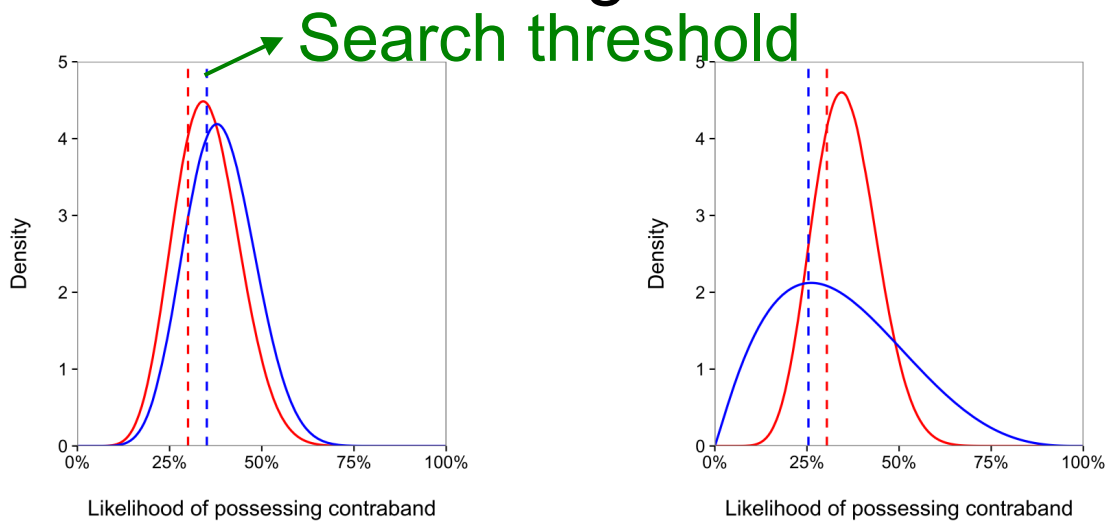
"This is not me," Robert Julian-Borchak Williams told investigators. "You think all Black men look alike?"

[Hill, The New York Times, 2020]

Inframarginality

- There are two groups of drivers: red drivers and blue drivers.
- Red drivers are searched more often than blue drivers (71% vs 64%)
- Searches of red drivers recover contraband less often (39% vs 44%)
- Are red drivers discriminated against?

Yes, because searching them when less likely to have drugs



No, not necessarily. Macro statistics don't tell you enough

[Simoiu et al., 2017]

Should we never use sensitive features?

In 2024, researchers were building a model to predict colorectal cancer risk in the Southern Community Cohort Study.

Claim: By removing race as a feature in this model, we will always reduce bias (if not completely eliminate it).

FALSE

This study found that family history was more predictive for White patients than for Black. Training a new model that could use race as a feature allowed it to adjust for this, and improved prediction performance for Black patients.

[Zink et al., 2024]

Summary

- Correlation is not causation
- Distribution shifts are everywhere (almost never have i.i.d. data)
- Be thoughtful about biases in your data & models
- Always understand your data & where it came from

Further reading

- Simoiu et al., The problem of infra-marginality in outcome tests for discrimination, 2017. https://5harad.com/papers/threshold-test.pdf
- Hill, Wrongfully accused by an algorithm, 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html
- Dastin, Amazon scraps secret AI recruiting tool that showed bias against women, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/
- Crawford, The Hidden Biases in Big Data, 2013. https://hbr.org/2013/04/the-hidden-biases-in-big-data
- Obermeyer et al., Dissecting racial bias in an algorithm used to manage the health of populations, 2019. https://www.science.org/doi/10.1126/science.aax2342
- Zink et al., Race adjustments in clinical algorithms can help correct for racial disparities in data quality, 2024. https://www.pnas.org/doi/10.1073/pnas.2402267121
- Koh and Sagawa et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020. https://arxiv.org/abs/2012.07421
- “Fairness and Machine learning” Solon Barocas, Moritz Hardt, Arvind Narayanan. https://fairmlbook.org/