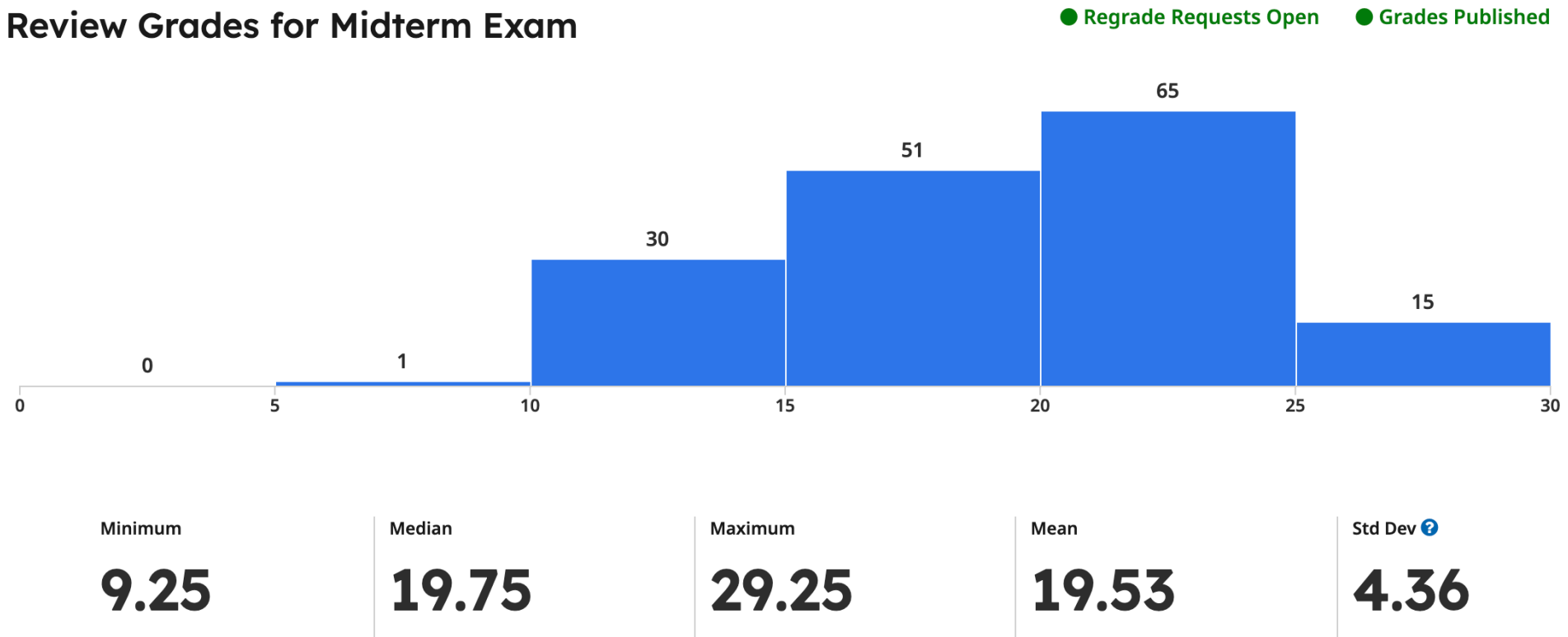


Midterm was hard

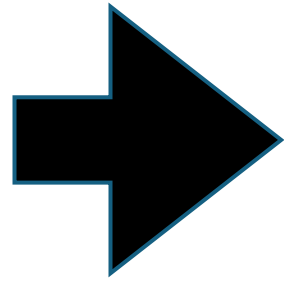
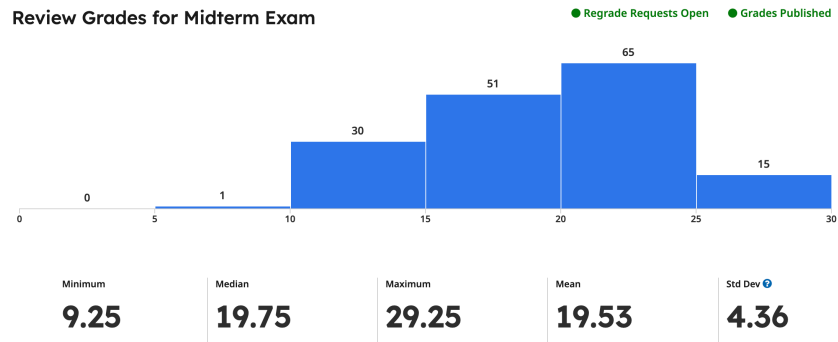
Review Grades for Midterm Exam



Average was 65%

Course will be curved aggressively

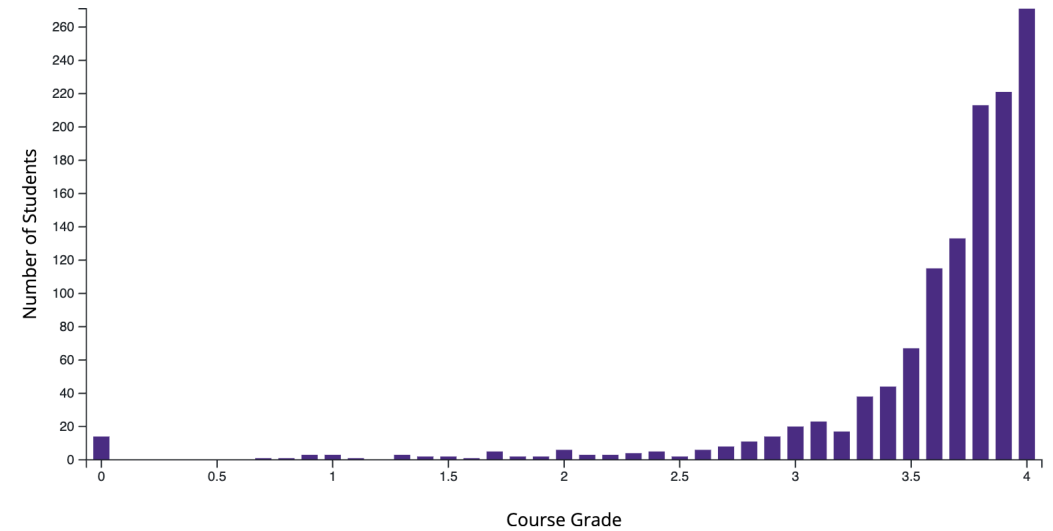
Review Grades for Midterm Exam



Course Grade Distribution

This graph represents the distribution of grades for every student who completed **CSE 446** over the past 5 years. [i](#)

Display data as a table



Number of grades in this sample: **1264** (5 years).
The graph does not include pass/fail grades.

<https://dawgpath.uw.edu/course?id=CSE+446>

Prediction Pitfalls

Natasha Jaques



Interpreting coefficients

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

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

what if features have different scales?
sqft have a mean of 2500 $w_i \rightarrow \$/ft^2$
baths " " " of 2 $w_j \rightarrow \$/bath$
→ to have mean 0 unit variance
→ - mean / std dev (z-score)
↳ fix by standardizing

FALSE

Interpreting coefficients

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

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

False: feat i might be correlated with y but
redundant w/ feat j
eg. area in sqft vs area in sq m

Lasso

Interpreting coefficients

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

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

False

- ① Distribution shift → a house with 10 fireplaces is outside sample range in training data, OOD = out of distribution
- ② Correlation \neq causation
- ③ Linear model, Non-linear relationship

Generalization

Train
hospitals
1, 2, 3

Test
hospital
4
→ ?

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.

① 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.

→ new hospital could have different demographics
FALSE Domain shift / distributional shift

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.

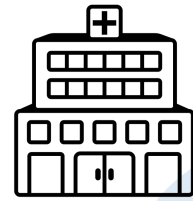
② Claim: The test set performance is always a good indicator of how our model will do if we deploy this model in the same hospitals.

False distribution shift → temporal shift

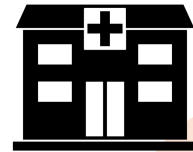
① & ② both common in practice

Domain shifts

/ Distribution shift



Training
distribution



Test
distribution

↙ shift: why?
- temporal shift
- patient pool / demographics
- environmental / geographical

Case study: EPIC's sepsis model

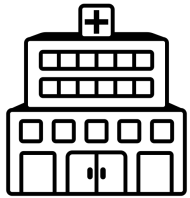
EPIC: large US
healthcare company

Early warnings
for sepsis

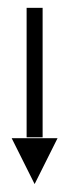


Case study: EPIC's sepsis model

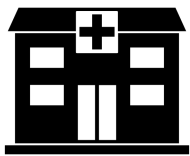
→ more training data before deploying in new hospital, collect same data & adapt model



Trained on 3 hospitals



Distribution shift



Deployed on 100s of **other** hospitals

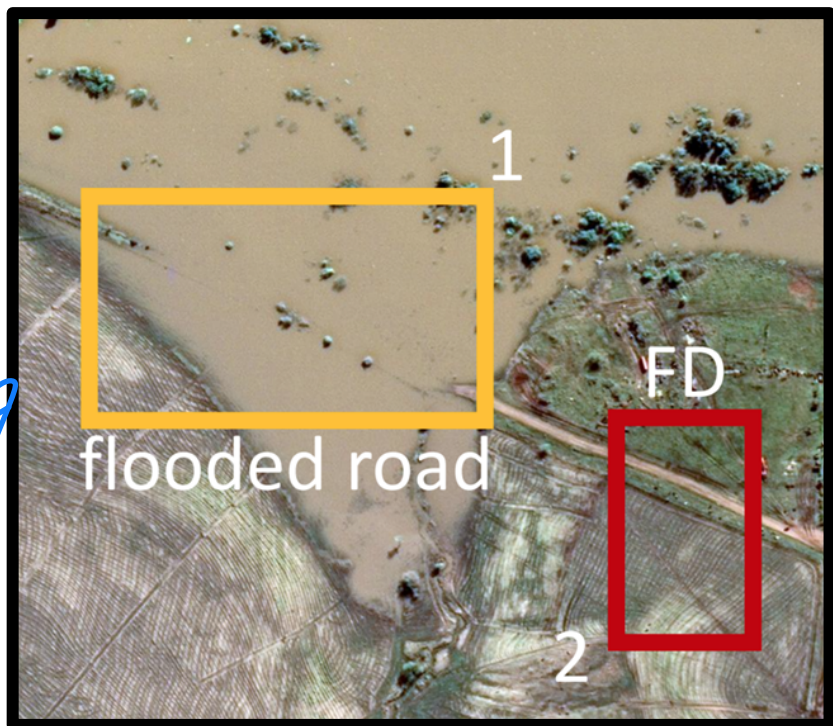
NEJM
Journal Watch

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.

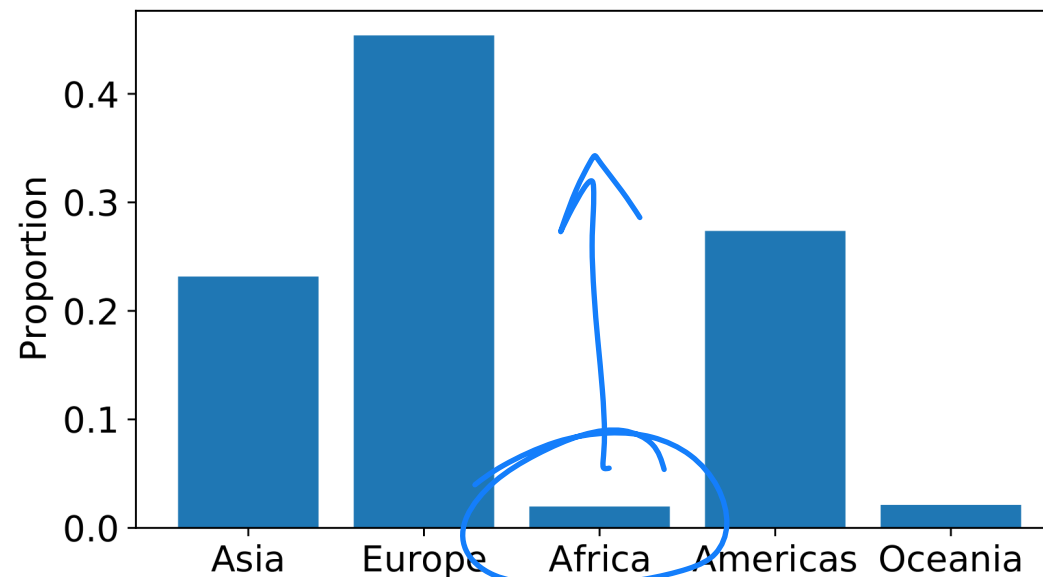


Disaster prediction to
allocate resources



label = flood, true,
false discovery (FD)

Sources of training data
(FMoW-WILDS satellite dataset)



Test accuracy on Americas: **55.7%**
Test accuracy on Africa: **32.3%** → shift

WILDS data

Training data

Camera 1

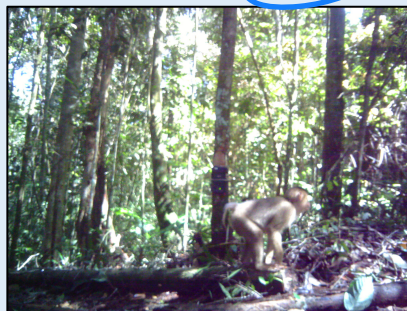


Camera 2



...

Camera 245



Out-of-distribution (OOD) test data

Camera 246



...

Control: In-distribution (ID) test data

Camera 1

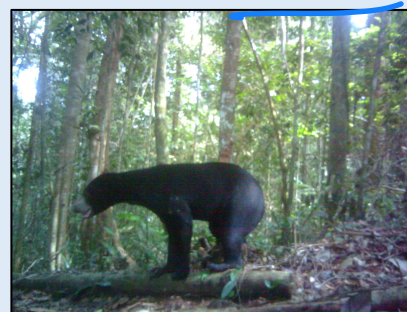


Camera 2

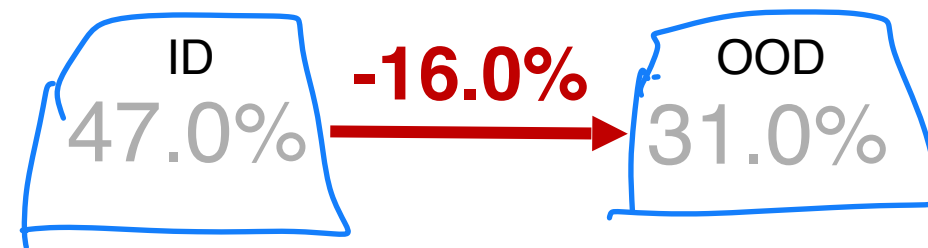


...

Camera 245

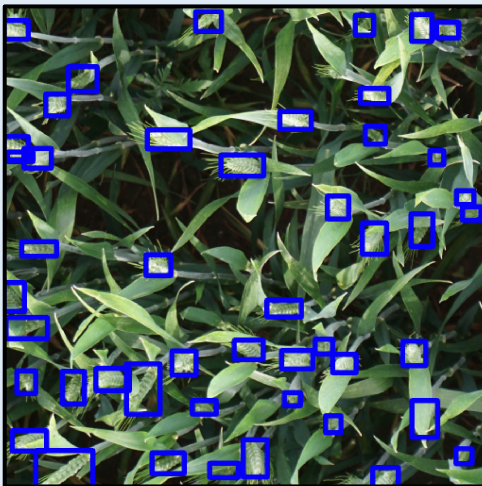


Macro F1



Training data

Belgium

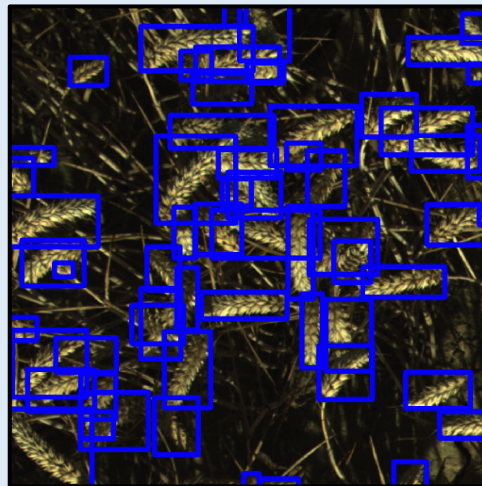


France



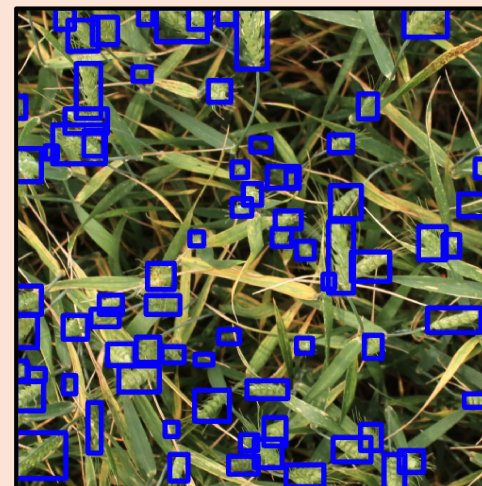
...

Norway

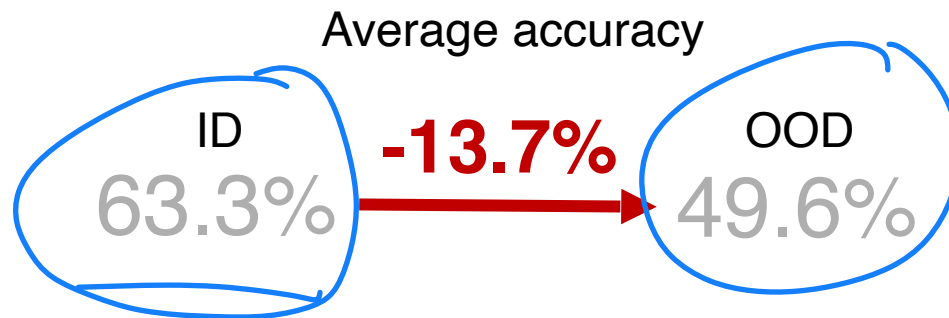


OOD test data

United States

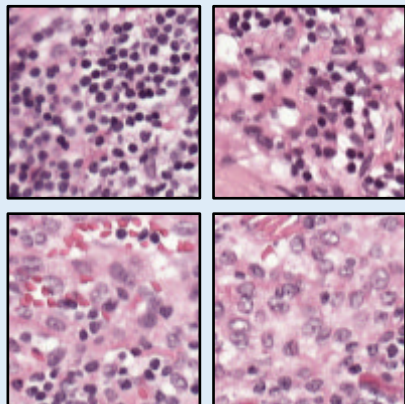


...

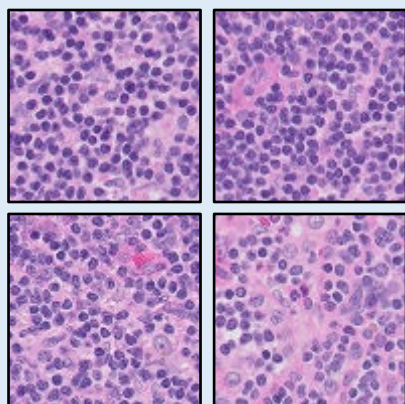


Training data

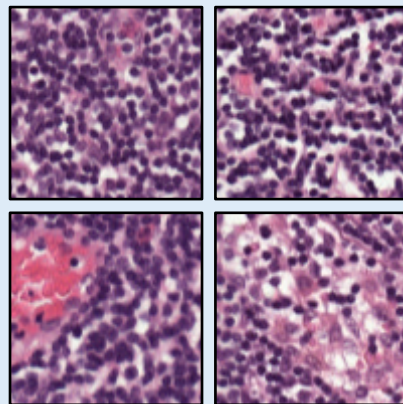
Hospital 1



Hospital 2

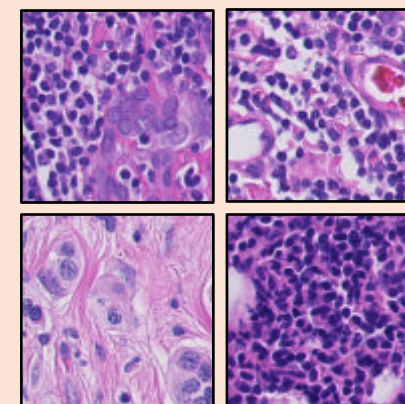


Hospital 3

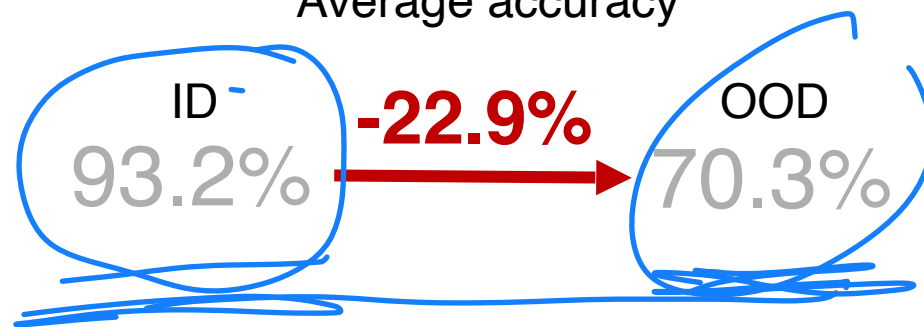


OOD test data

Hospital 4



Average accuracy



Common
How to fix?

- collect more data
- domain knowledge
- DRD + distributionally robust opt.

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

How to fix?
- weight samples differently to balance reporting 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: report #s \propto actual traffic \times reporting bias
- younger
- wealthier

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 \rightarrow more risk
more risk \rightarrow more hospital visits

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

conseq: more \$\$\$ \rightarrow more hospital visits

\rightarrow allocate resources to the ppl w/ the most resources

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.

Data was biased
if X had feats indicating the applicant was a woman
→ predict no hire

Is it sufficient to remove sensitive features?

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

X = resume

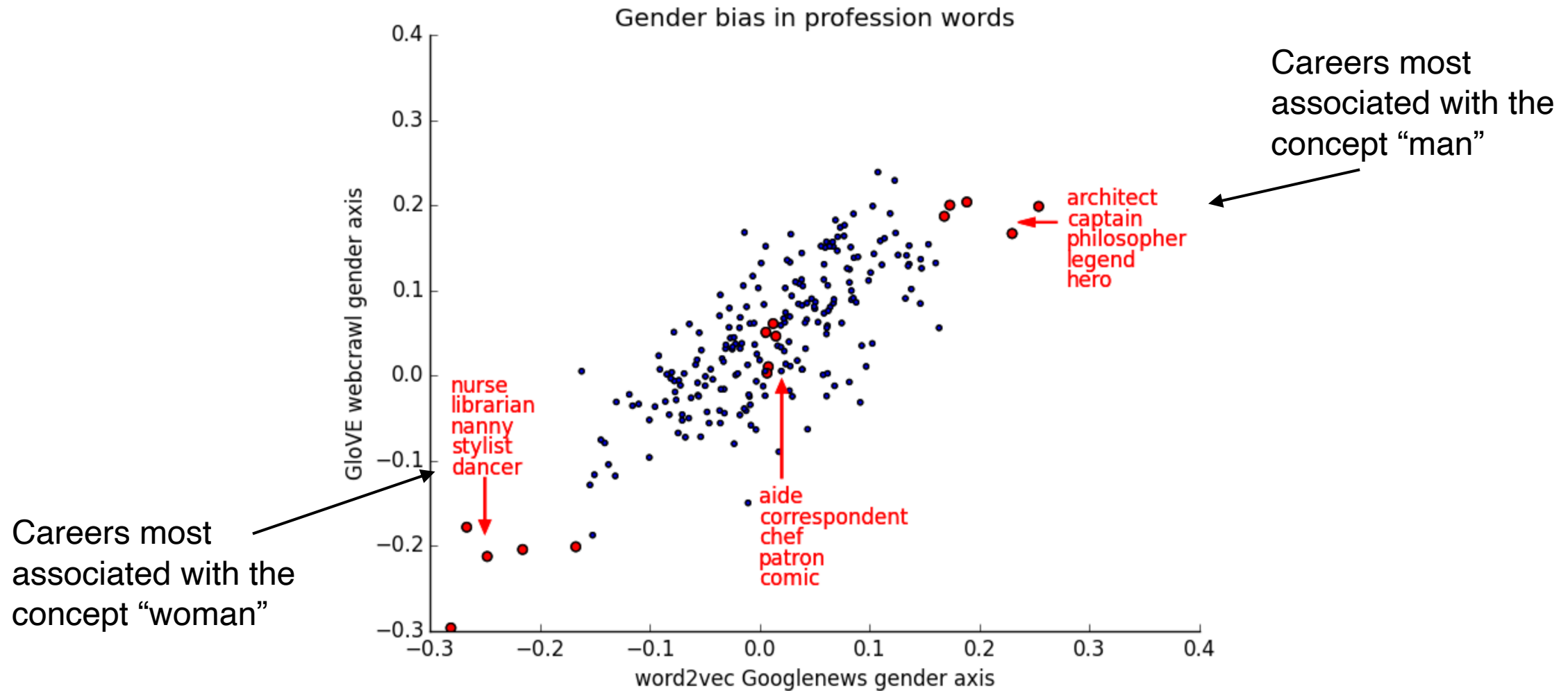
Y = suitability for the job (hiring decision / job performance)

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

False: model could use other feats that correlate w
identity → college, zip code

What to do? retain knowledge of gender/race, correct for bias

Stereotypes in language models



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?”



Face recognition accuracy by race and gender

| Classifier | Metric | All | F | M | Darker | Lighter | DF | DM | LF | LM |
|------------|---------------|------|------|------|--------|---------|-------------|-------------|-------------|-------------|
| MSFT | TPR(%) | 93.7 | 89.3 | 97.4 | 87.1 | 99.3 | 79.2 | 94.0 | 98.3 | 100 |
| | Error Rate(%) | 6.3 | 10.7 | 2.6 | 12.9 | 0.7 | 20.8 | 6.0 | 1.7 | 0.0 |
| | PPV (%) | 93.7 | 96.5 | 91.7 | 87.1 | 99.3 | 92.1 | 83.7 | 100 | 98.7 |
| | FPR (%) | 6.3 | 2.6 | 10.7 | 12.9 | 0.7 | 6.0 | 20.8 | 0.0 | 1.7 |
| Face++ | TPR(%) | 90.0 | 78.7 | 99.3 | 83.5 | 95.3 | 65.5 | 99.3 | 90.2 | 99.2 |
| | Error Rate(%) | 10.0 | 21.3 | 0.7 | 16.5 | 4.7 | 34.5 | 0.7 | 9.8 | 0.8 |
| | PPV (%) | 90.0 | 98.9 | 85.1 | 83.5 | 95.3 | 98.8 | 76.6 | 98.9 | 92.9 |
| | FPR (%) | 10.0 | 0.7 | 21.3 | 16.5 | 4.7 | 0.7 | 34.5 | 0.8 | 9.8 |
| IBM | TPR(%) | 87.9 | 79.7 | 94.4 | 77.6 | 96.8 | 65.3 | 88.0 | 92.9 | 99.7 |
| | Error Rate(%) | 12.1 | 20.3 | 5.6 | 22.4 | 3.2 | 34.7 | 12.0 | 7.1 | 0.3 |
| | PPV (%) | 87.9 | 92.1 | 85.2 | 77.6 | 96.8 | 82.3 | 74.8 | 99.6 | 94.8 |
| | FPR (%) | 12.1 | 5.6 | 20.3 | 22.4 | 3.2 | 12.0 | 34.7 | 0.3 | 7.1 |



Joy Buolamwini



Timnit Gebru

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-TPR), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

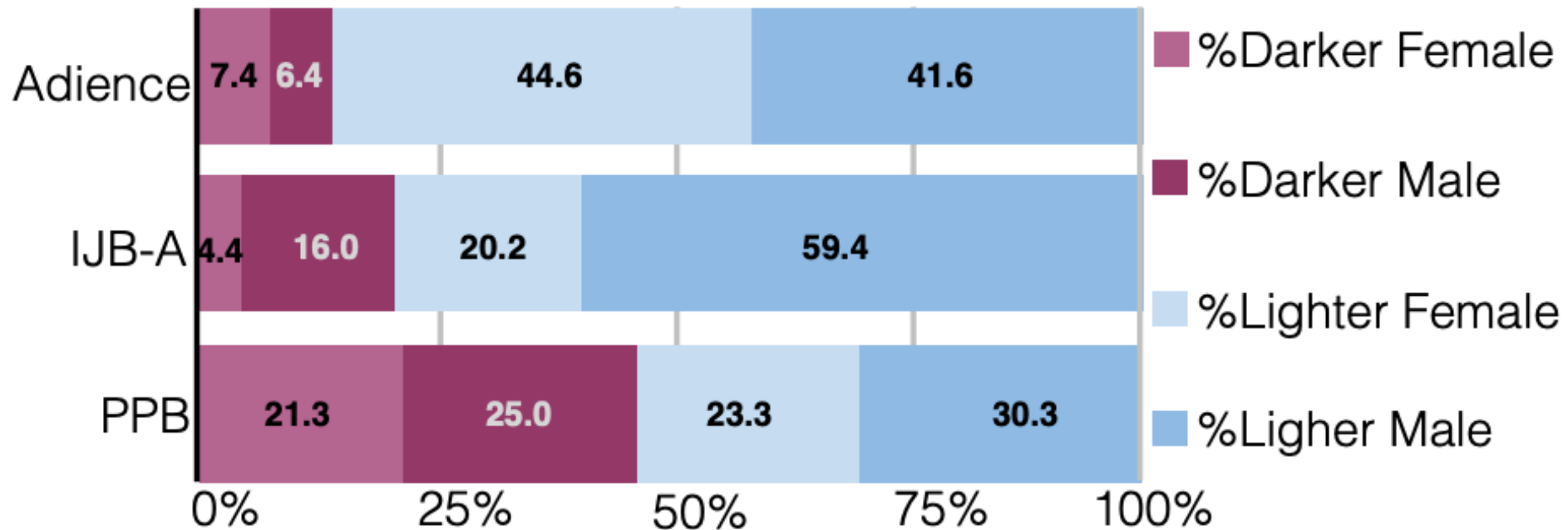
Face recognition accuracy by race and gender

Why does this happen?

Face recognition accuracy by race and gender

Why does this happen?

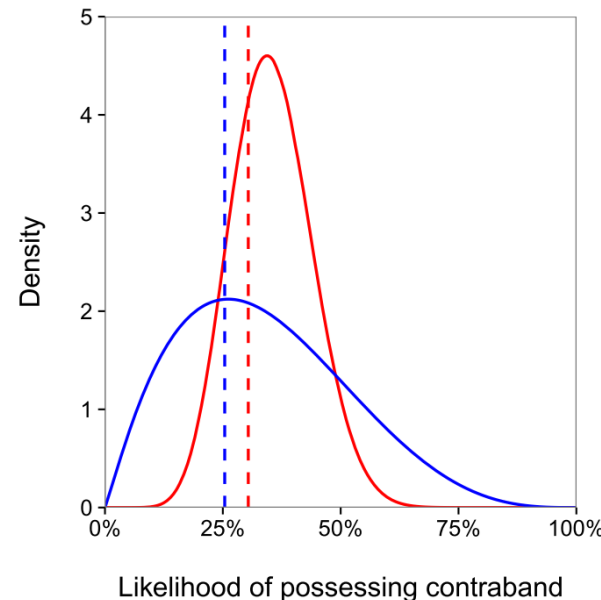
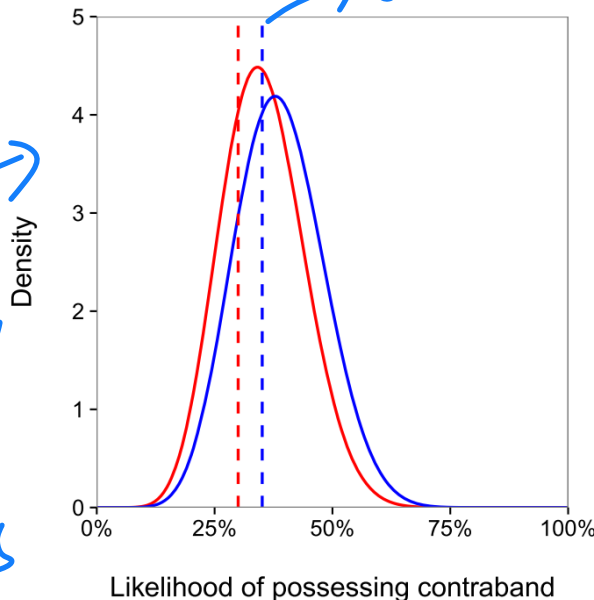
Training data is biased.



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?

discriminating →
against red,
because searching
them when
less likely to
have drugs



no: not
necessarily
macro statistics
don't tell you
enough

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).

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/>