

ML/DS Journal Club – Feb 27, 2019

Introduction to Fairness in Machine Learning

Hongsup Shin

Today

Why fairness?

Bias in ML: examples and causes

How to define fairness

Fairness algorithms

Recommendations

Sources

Machine Bias (Propublica): <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

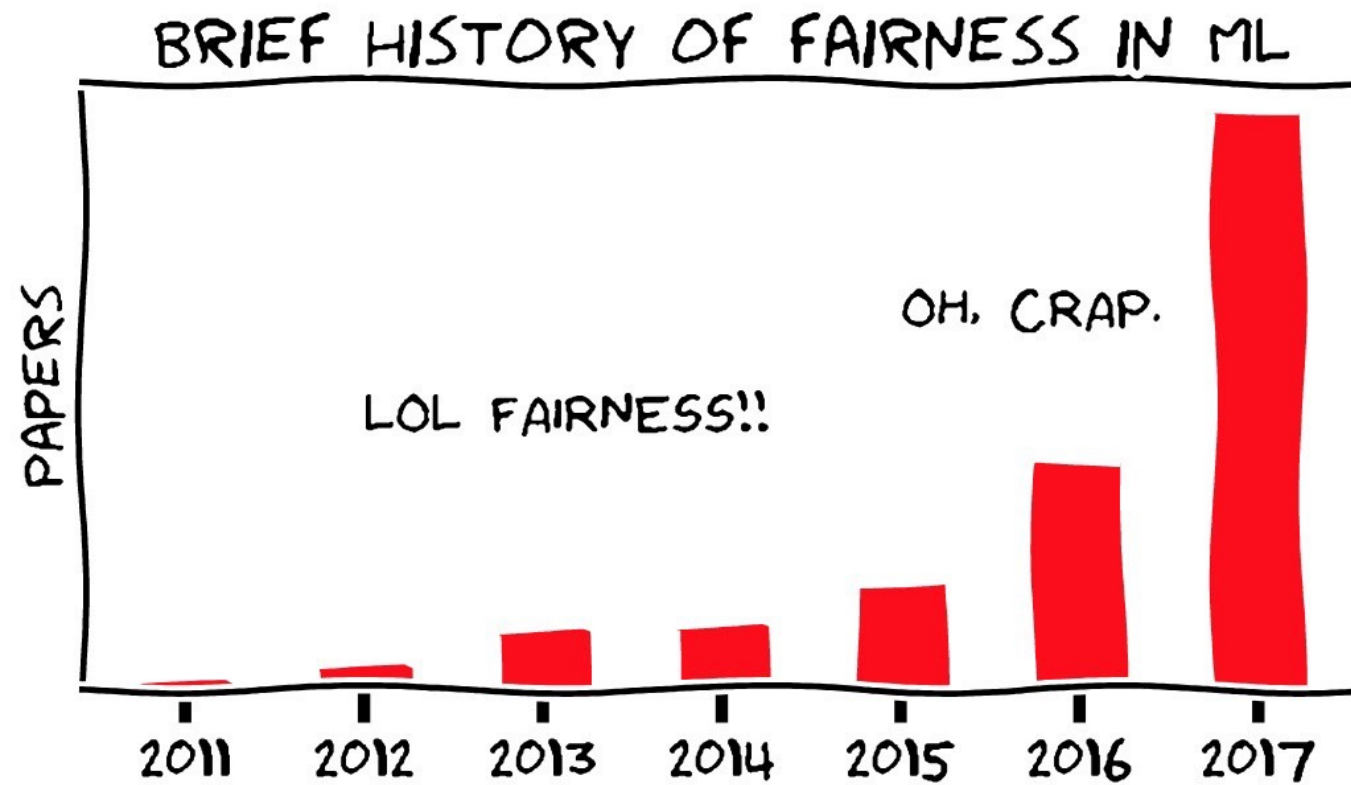
NIPS 2017 Tutorial on Fairness in Machine Learning (Solon Barocas & Moritz Hardt, Cornell): <https://fairmlbook.org/tutorial1.html>

A Tutorial on Fairness in Machine Learning (Ziyuan Zhong, Columbia): <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

Attacking discrimination with smarter machine learning (Google Brain): <https://research.google.com/bigpicture/attacking-discrimination-in-ml/>

Counterfactual Fairness (Matt Kusner, The Alan Turing Institute): <https://www.youtube.com/watch?v=psA4U6nhZ70>

AI Fairness 360 Open Source Toolkit (IBM Research Trusted AI): <http://aif360.mybluemix.net>



<https://fairmlclass.github.io>

Why fairness

For your own benefit as a human being

- Nobody likes to be treated unfairly
 - Discrimination laws are "reactive" (victims should prove the wrongdoing)
 - Or often you don't even know you're being discriminated
 - Doing the right thing
- Biased-ML already exists

As a scientist and modeler

- Excellent intellectual exercise
 - How to formulate problems
 - How to optimize various objectives
 - How to process data and modify models to mitigate bias
- Practicing robust ML engineering
- Applicable to any ML projects

Bias in machine learning

Biased ML

- Facebook ad-personalization violates the fair housing act
- Google's gender-biased algorithms: image search, google translate, etc.

The image shows two screenshots of the Google Translate interface. The top screenshot shows the 'DETECT LANGUAGE' tab selected, with 'HUNGARIAN' and 'ENGLISH' tabs visible. The input text is 'She is a doctor.', 'He is a nurse.', and 'She is an engineer.'. The output text is 'Ő egy orvos.', 'Ő egy nővér.', and 'Ő egy mérnök.'. The bottom screenshot shows the 'DETECT LANGUAGE' tab selected, with 'HUNGARIAN' and 'ENGLISH' tabs visible. The input text is 'Ő egy orvos.', 'Ő egy nővér.', and 'Ő egy mérnök.'. The output text is 'He's a doctor.', 'She's a nurse.', and 'He's an engineer.'.

DETECT LANGUAGE	HUNGARIAN	ENGLISH	SPANISH	↔	ENGLISH	HUNGARIAN
She is a doctor.	He is a nurse.	She is an engineer.	×	Ő egy orvos.	Ő egy nővér.	Ő egy mérnök.

DETECT LANGUAGE	HUNGARIAN	ENGLISH	SPANISH	↔	ENGLISH	HUNGARIAN
Ő egy orvos.	Ő egy nővér.	Ő egy mérnök.	×	He's a doctor.	She's a nurse.	He's an engineer.

- Most face recognition algorithms (Face++, Microsoft) misidentify the gender of dark-skinned population: misidentify female as male (130M adults is in a law enforcement face recognition network in the US)
- XING, a job platform similar to Linked-in, ranked less qualified male candidates higher than more qualified female candidates

And, there's COMPAS.



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

COMPAS

Correctional Offender Management Profiling for Alternative Sanctions, developed by a **for-profit** company

- One of the most widely used “risk assessment” software: **predicting recidivism** (likelihood of committing crime again)
- Presented to judges on **parole** as additional info: whether you need more service or you will likely to fail
- Many states started using as early as **since 2001**

Algorithm

- Proprietary (not open to public), score rating (0-10, low-high)
- Training data: 137 questions (answered by defendants or pulled from criminal records, race not included)
 - “Was one of your parents ever sent to jail or prison?”
 - “A hungry person has a right to steal, agree/disagree?”
 - “How often did you get in fights while at school?”

Problem: Bad ML practice

Almost no evaluation: Many states *started using before rigorous testing*. Very few studies looked into this (and not even thorough investigation)

Software has been misused for different purposes: Decision on sentencing, setting the bond amount, or *at arrest* to determine if a defendant is too risky for pre-trial release

Lacking interpretability

- “A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job. Meanwhile, a drunk guy will look high risk because he’s homeless.”
- Some important features turned out to be correlated to race and gender.

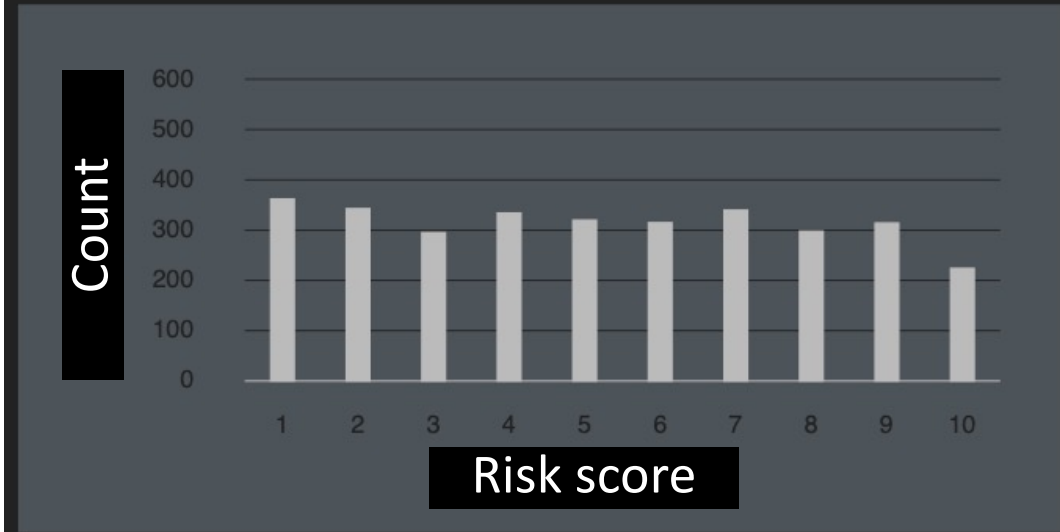
No contestability: Defendants rarely have an opportunity to challenge their assessments.

No disclosure: The score calculations are rarely revealed.

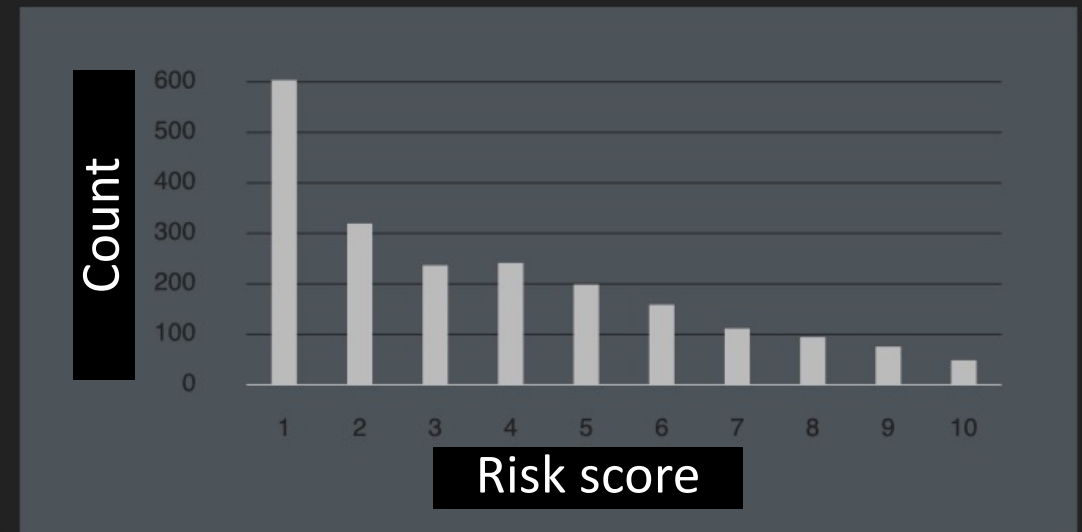
Many US states started using COPMAS as early as since 2001.

Problem: Racial bias in prediction

Black Defendants' Risk Scores



White Defendants' Risk Scores



Prediction Fails Differently for Black Defendants

	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%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Problem: Legal system

Loomis vs. Wisconsin

- “Eric Loomis challenged the State of Wisconsin's use of proprietary, closed-source risk assessment software in the **sentencing of Eric Loomis to six years in prison.**”
- “...it prevents the defendant from challenging the scientific validity and accuracy of such test.”

The Supreme Court declined to hear the case in 2017. (<https://www.scotusblog.com/case-files/cases/loomis-v-wisconsin/>)

This case, however, is not a suitable vehicle. Initially, it is **unclear** *how* COMPAS accounts for gender—a fact of relevance to the constitutional analysis.

As the State explains (Br. in Opp. 12), though, the Wisconsin Supreme Court **did not resolve** the parties’ “sharp disagreement” over precisely “how COMPAS takes gender into account in calculating the” risk scores.

If it's math, can we just look away?

What causes bias in ML systems

Skewed examples: sampling bias (e.g., police record: more police dispatch in high crime-rate area)

Tainted examples: learning human bias in training data (e.g., Google Translate)

Limited features: reliability of the label from a minority group can be much lower than the counterpart from a majority group -> lower accuracy for the minority group

Sample size disparity: class imbalance

Proxies: even if we rule out sensitive attributes (e.g., race, gender), there can always be other features that are proxies of them (e.g., zip code in “redlining”)

How to define fairness

Definition of fairness: how to measure it

Fairness definition/measure

- Unawareness
- Demographic Parity
- Equal Opportunity
- Predictive Rate Parity
- Individual Fairness
- Counterfactual fairness



Arvind Narayanan ✓

@random_walker

Follow



I wrote up a 2-pager titled "21 fairness definitions and their politics" based on the tweetstorm below and it was accepted at a tutorial for the Conference on Fairness, Accountability, and Transparency!







Fairness measure \neq model accuracy
(can be measured separately)

Setup

Sensitive attributes, A

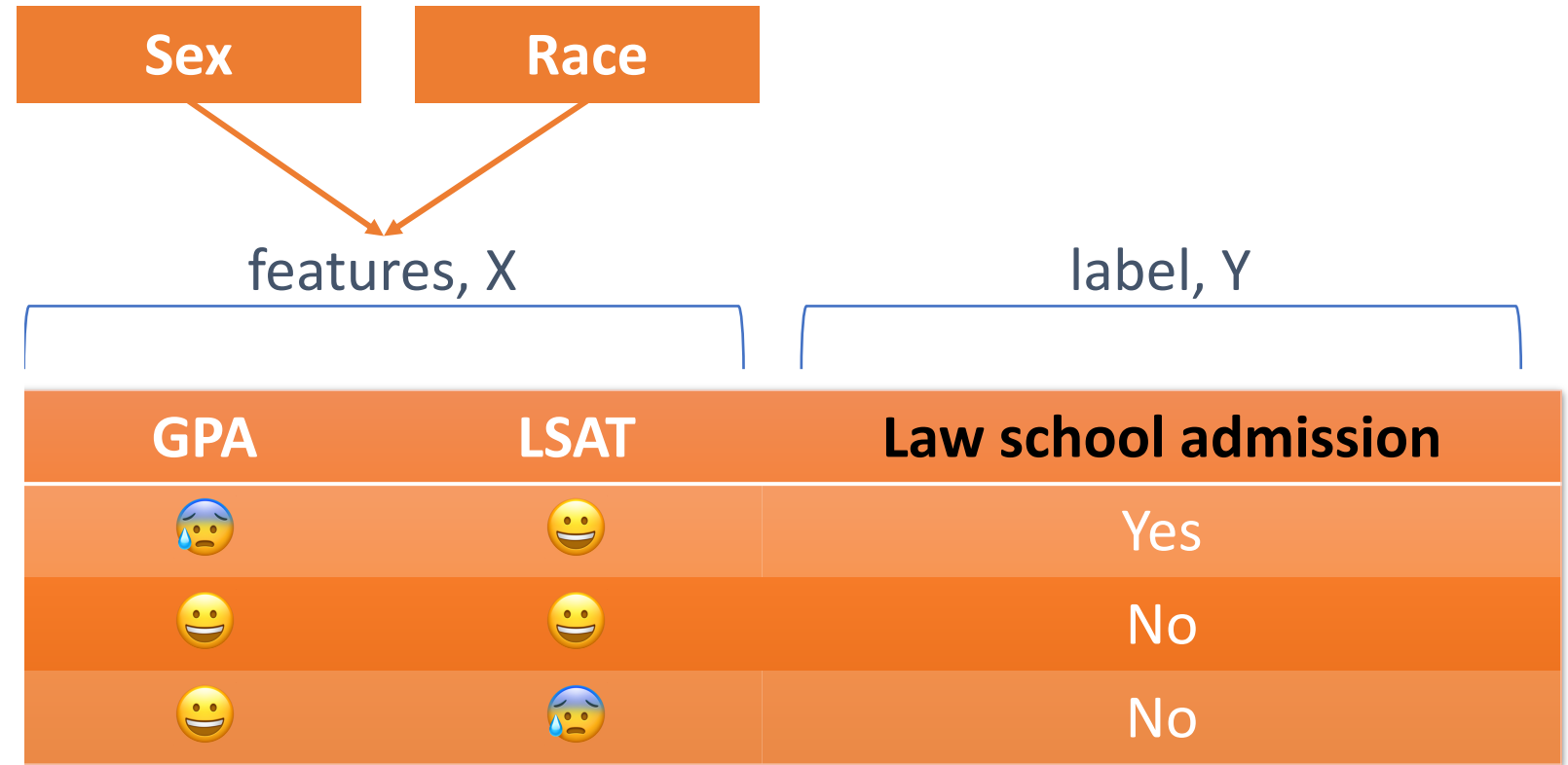
features, X

label, Y

Sex	Race	GPA	LSAT	Law school admission
M	White			Yes
F	Black			No
M	Black			No

- Goal: Using historic data, build a model that automatically decides student admission
- Model (classifier): predict whether a student will be admitted or not
- \hat{Y} : prediction (binary); $\hat{Y} = 1$: school decides to admit the student

Unawareness



“minority students may feel teachers are unsupportive” (Rowley et al. 2014)

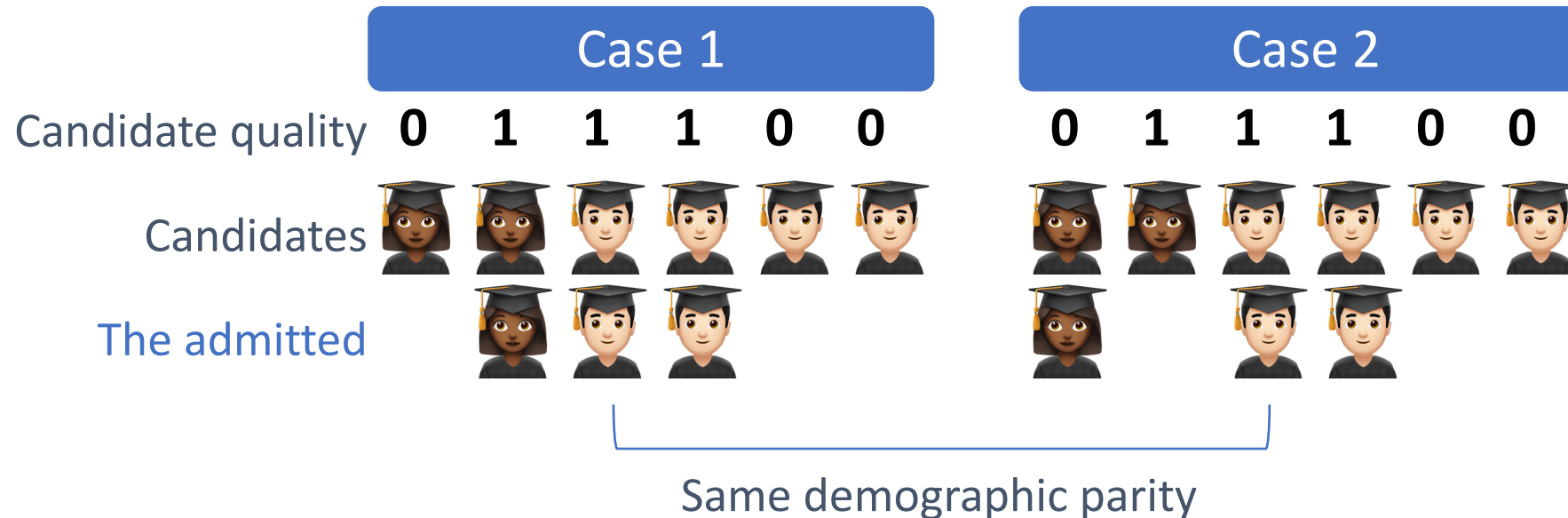
“teachers may believe minority students have behavior issues” (Ferguson, 2003)

Demographic parity (aka independence, statistical parity)

The admission rate should be equal among different groups.

$$P[\hat{Y} = 1 | A = \text{black}] = P[\hat{Y} = 1 | A = \text{white}]$$










Laziness: even when the model has high accuracy for one group but low accuracy for the other group, we can still achieve demographic parity. (**doesn't care accuracy**)



Equal opportunity (aka separation, equalized odds)

The admission rate of the **qualified individuals** should be equal among different groups.

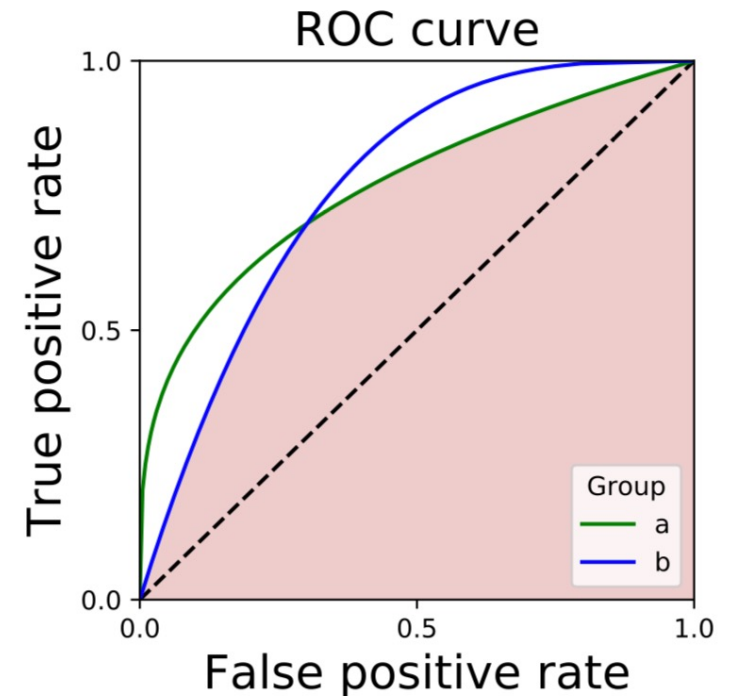
$$P[\hat{Y} = 1 | A = \text{black}, Y = 1] = P[\hat{Y} = 1 | A = \text{white}, Y = 1]$$

Candidate quality	0	1	1	1	0	0
Candidates						
The admitted						

- Penalize laziness: incentive to reduce errors uniformly in all groups
- Doesn't reduce the gaps between groups
- Sensitive features can influence the label

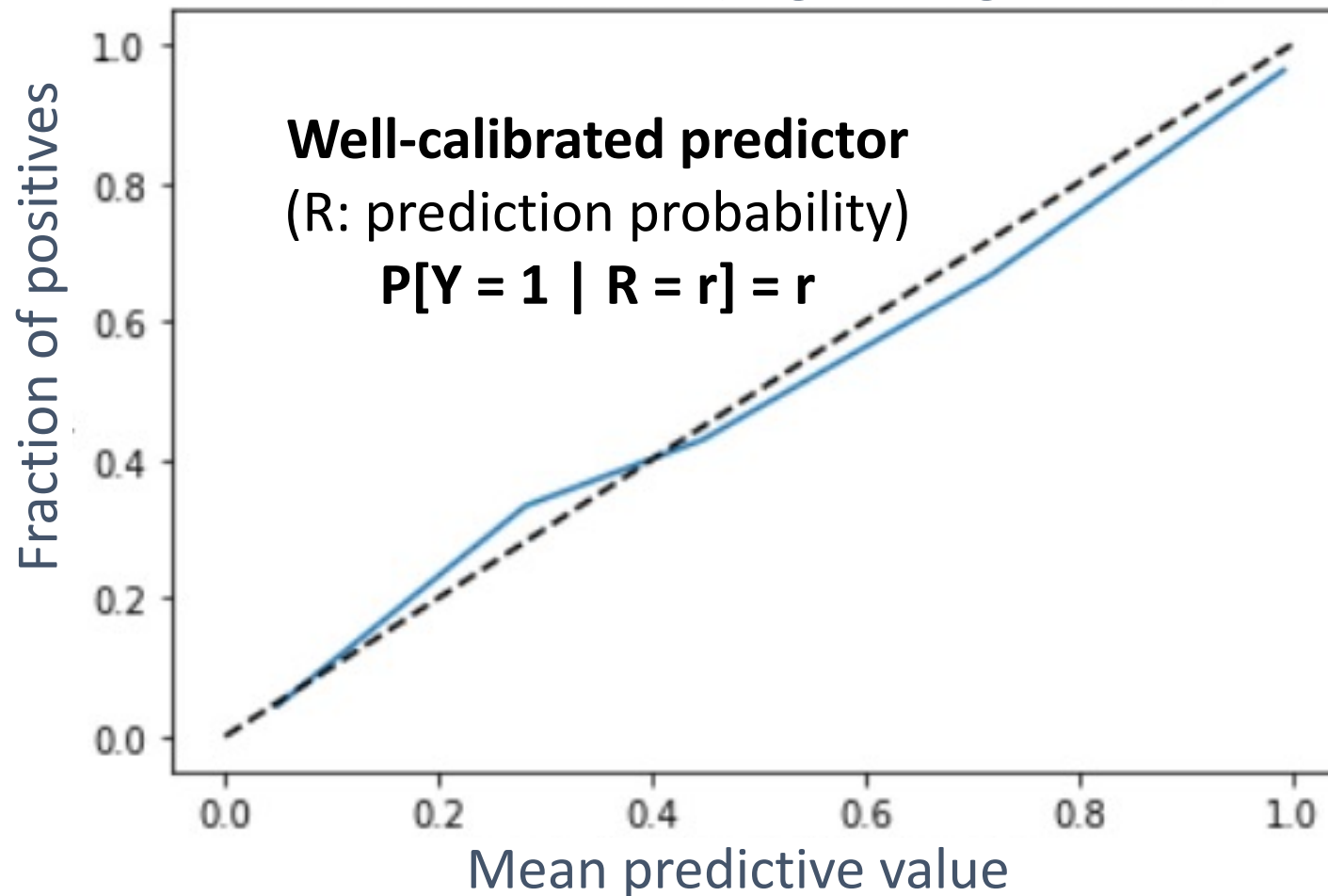
“minority race teachers affect minority student outcomes” (Birdsall et al. 2016)

“stereotype threat”: black students performed worse on standardized tests than white students when their race was emphasized (Steele & Aronson, 1995)



Predictive parity (aka *calibration*, sufficiency)

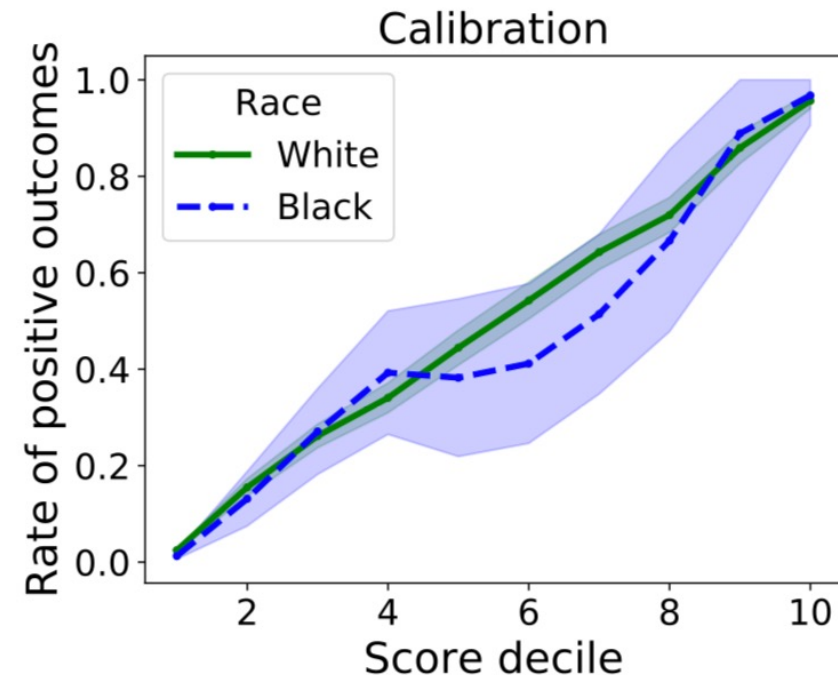
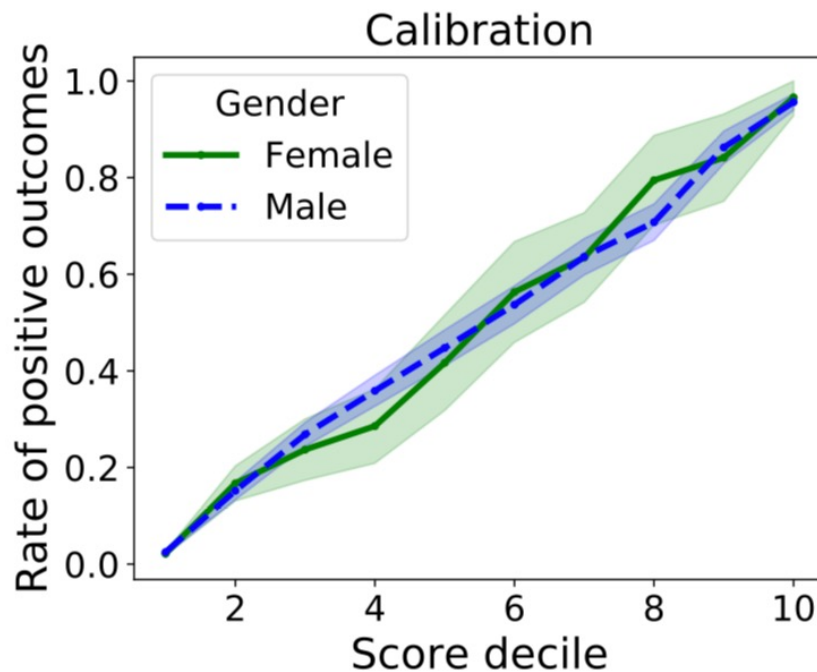
Calibration curve (Logistic regression)



Predictive parity (aka calibration, sufficiency)

Prediction probability (R) should be equally well-calibrated across different groups.

$$P[Y = 1 | A = \text{black}, R = r] = P[Y = 1 | A = \text{white}, R = r]$$



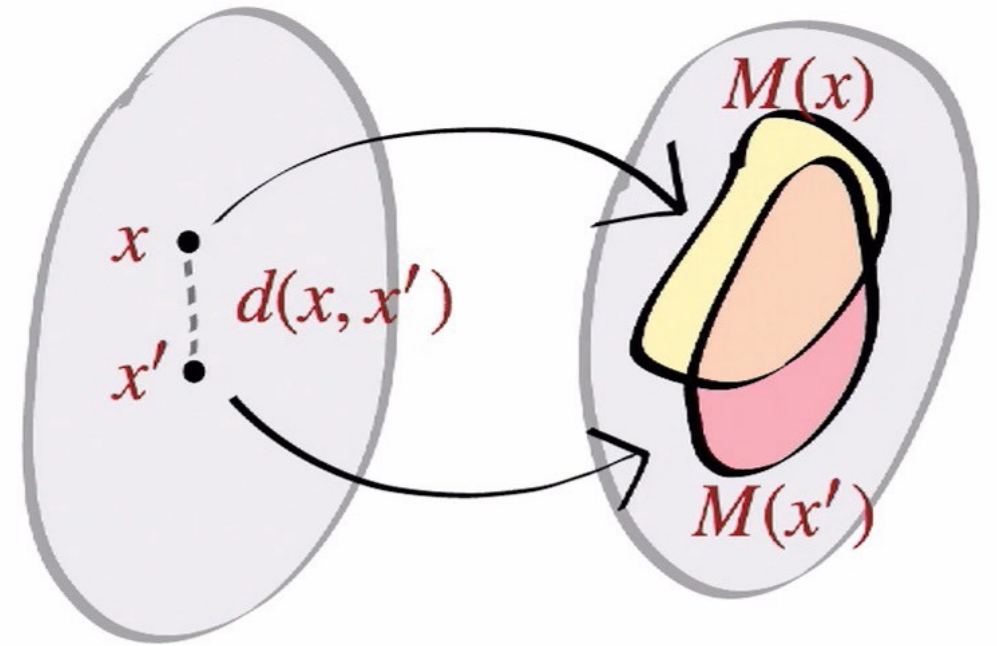
Individual fairness (Dwork et al. 2011)

If you and I are similar, we should be treated similarly.

More fine-grained than any group-notion fairness

Restriction on the treatment for each pair of individuals

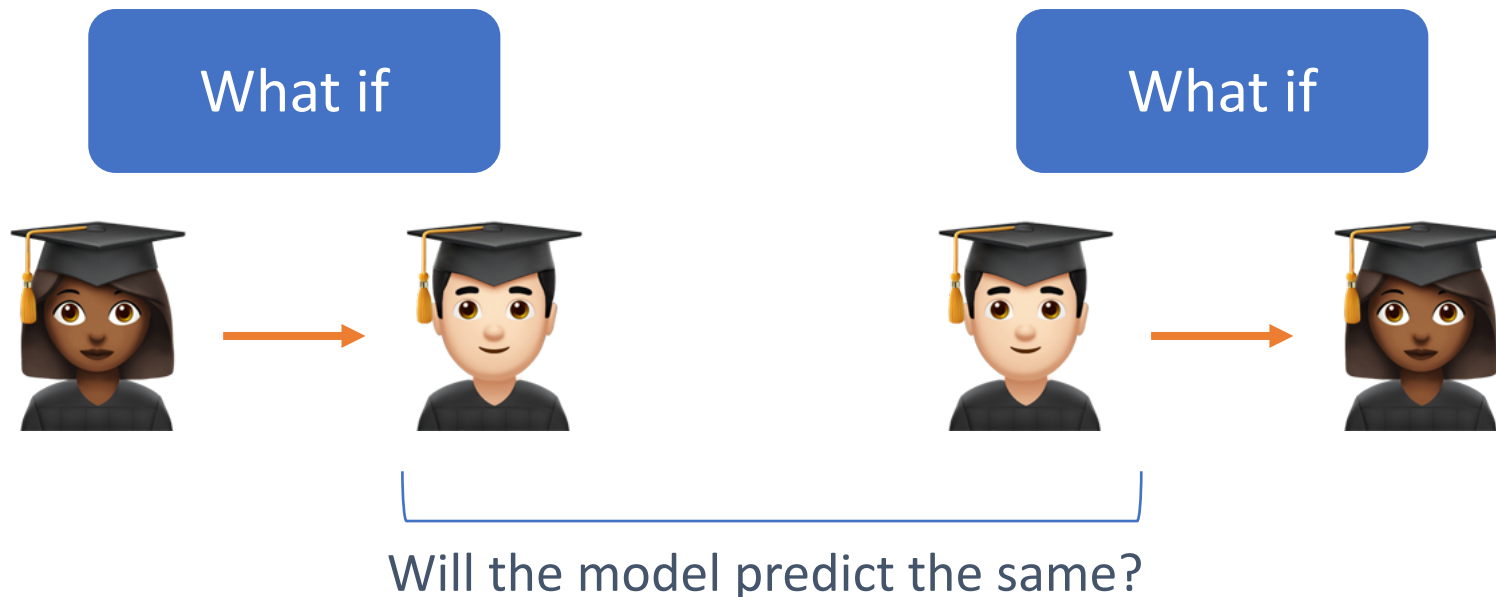
Hard to determine an appropriate similarity metric



Counterfactual fairness (Kusner et al. 2017)

Model how sensitive attributes cause unfair decisions via causal models

A predictor is fair if it would give the same prediction in a world where you were different



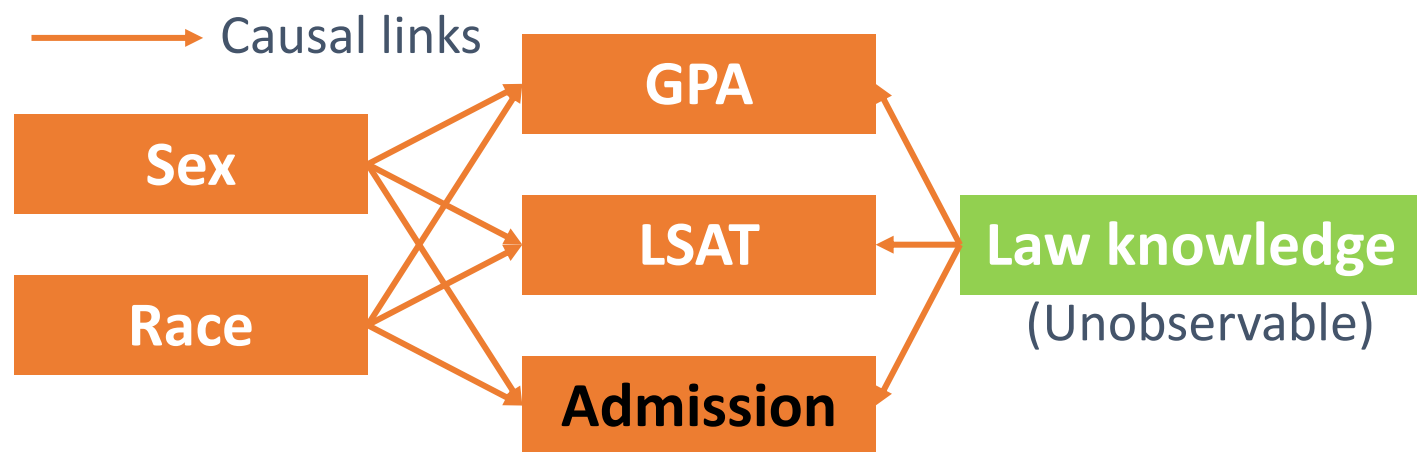
Counterfactual fairness (Kusner et al. 2017)

Sensitive attributes, A

features, X

label, Y

Sex	Race	GPA	LSAT	Law school admission
M	White	😓	😊	Yes
F	Black	😊	😊	No
M	Black	😊	😓	No



- Allows us to model how unfairness occurs
- A **fair classifier** gives the **same prediction** had the person had a **different race/sex**.
- Can be difficult to reach a consensus on what the causal graph should look like.
- Difficult to scale

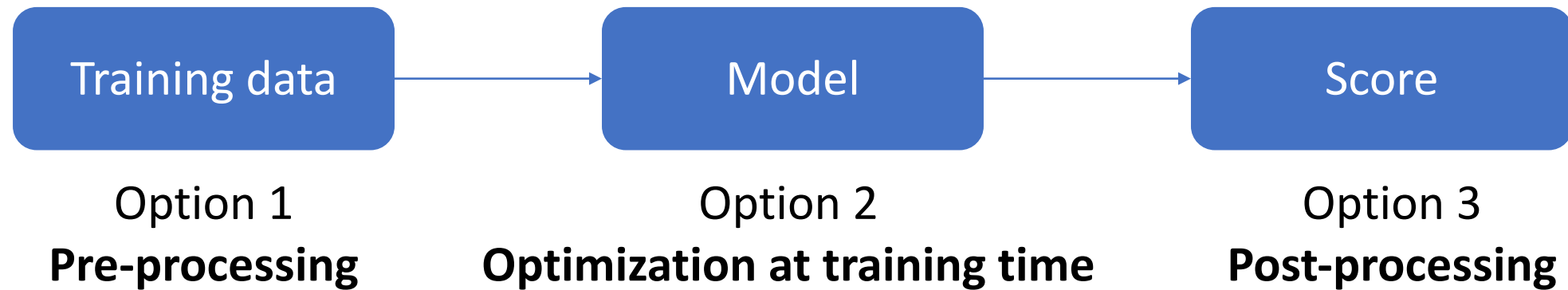
Practical matters

The Impossibility theorem of fairness: among demographic parity, equal opportunity, and predictive parity, any two of the three are mutually exclusive. (Impossible to come up with a solution that satisfies all).

Trade-off between fairness and accuracy: optimizing for fairness measure other than simple accuracy may compromise model performance

Fairness algorithms

When to intervene to mitigate the bias



Pre-processing

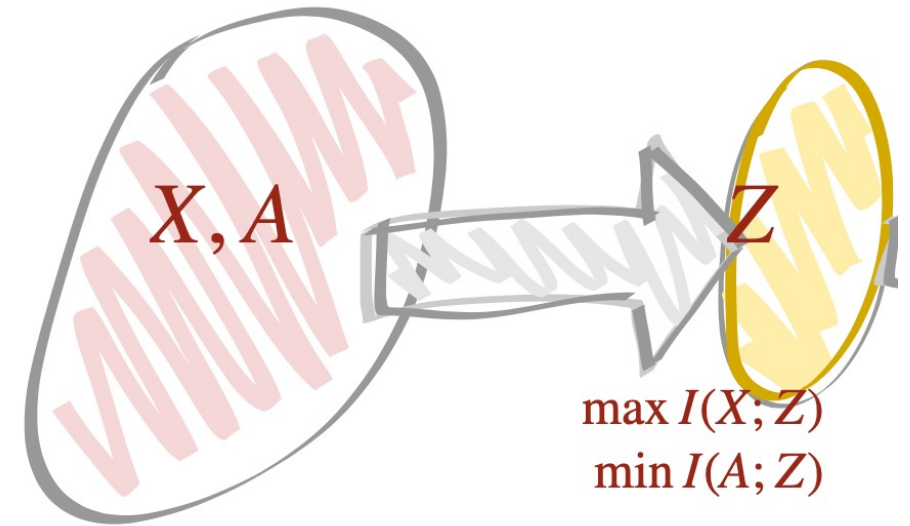
Learn a new representation Z such that it removes the information correlated to the sensitive attribute and preserves the information of X as much as possible.

Pros

- Preprocessed data can be used for any downstream task
- No need to modify classifier or to access sensitive attributes at test time

Cons

- Usage is limited because it does not have the information of label Y
- Inferior to the other two methods in terms of performance on accuracy and fairness measure



Optimization at training time

Add a constraint or a regularization term to the existing optimization objective.

Pros

- Good performance on accuracy and fairness measures
- Higher flexibility to choose the trade-off between accuracy and fairness measures
- No need to access sensitive attributes at test time

Cons

- Method in this category is task-specific
- Need to modify classifier, which may not be possible in many scenarios

Post-processing

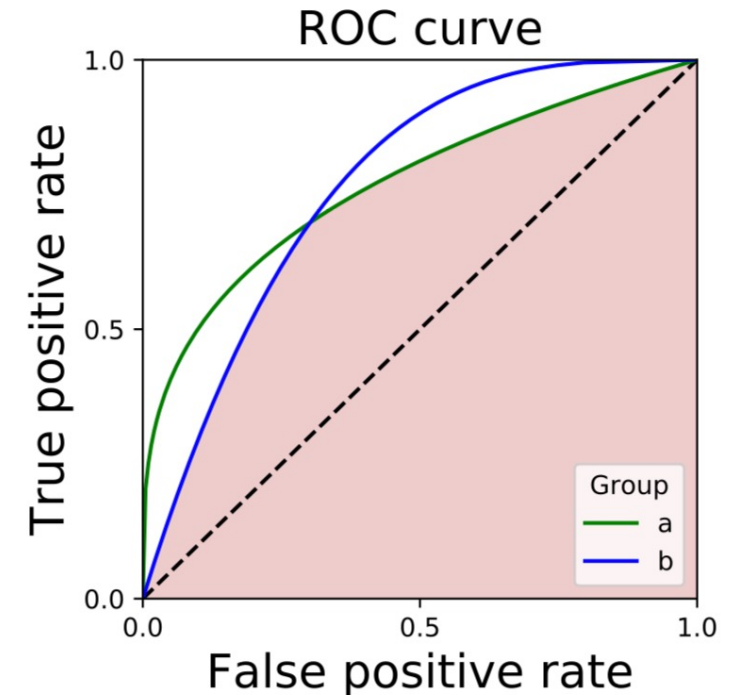
Edit predictions in a way that satisfies fairness constraints (find a proper threshold for the original scoring function)

Pros

- Can be applied after any classifiers
- Relatively good performance (especially fairness measures)
- No need to modify classifier

Cons

- Require test-time access to the protected attribute
- Lack the flexibility of picking any accuracy–fairness tradeoff



Implementation: AIF360 Demo

<http://aif360.mybluemix.net> by IBM Research (Trusted AI)

Open source toolkit to examine, report, and mitigate discrimination and bias in ML models throughout the AI application lifecycle.

Containing over 70 fairness metrics and 10 state-of-the-art bias mitigation algorithms

Designed to translate algorithmic research from the lab into the actual practice of domains

Lessons

Summary

Biased ML

ML is used in many aspects of our society (more than you think).

Biased ML is an already existing phenomenon.

ML can reproduce human bias in a massive scale.

Many current ML products lack accountability.

Fairness in ML

Implementing fairness in ML is doable in many different ways.

There is no single definition of fairness that can satisfy all scenarios.

Causal viewpoint can help articulate problems and organize assumptions.

Human scrutiny and expertise are irreplaceable.

Recommendations (by Moritz Hardt)

ML is domain-specific: we need to understand legal and social context.

Besides inspecting models, scrutinize data and how it was generated.

Besides static one-shot problems, study long-term effects, feedback loops, and interventions.

Establish qualitative understanding of when/why ML is the right tool for the application.

Problem: Bad ML practice

Almost no evaluation: Many states started using before rigorous testing. Very few studies looked into this (and not even thorough investigation) **Did you evaluate your model thoroughly before deployment?**

Software has been misused for different purposes: Decision on sentencing, setting the bond amount or at arrest to determine if a defendant is too risky for pre-trial release

Lacking interpretability **Are you and your stakeholders using the model for the right use case?**
When something goes wrong in your model, can it or you explain?

- “A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job. Meanwhile, a drunk guy will look high risk because he’s homeless.”
- **Some important features turned out to be correlated to race and gender.** **Are your features healthy?**

No contestability: Defendants rarely have an opportunity to challenge their assessments.

No disclosure: The score calculations are rarely revealed. **How do you handle your model’s failure?**

Are you the only one who knows how your model works?

These can happen to any DS/ML projects!