AUROC vs. AUPRC under Class Imbalance

Matthew McDermott, Haoran Zhang, Lasse Hansen, Giovanni Angelotti, Jack Gallifant

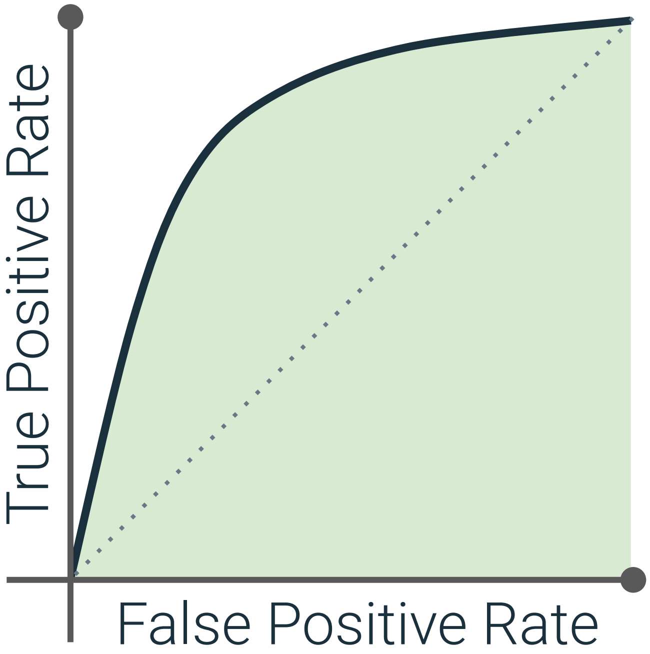
Berkowitz Postdoctoral Fellow

matthew_mcdermott@hms.harvard.edu

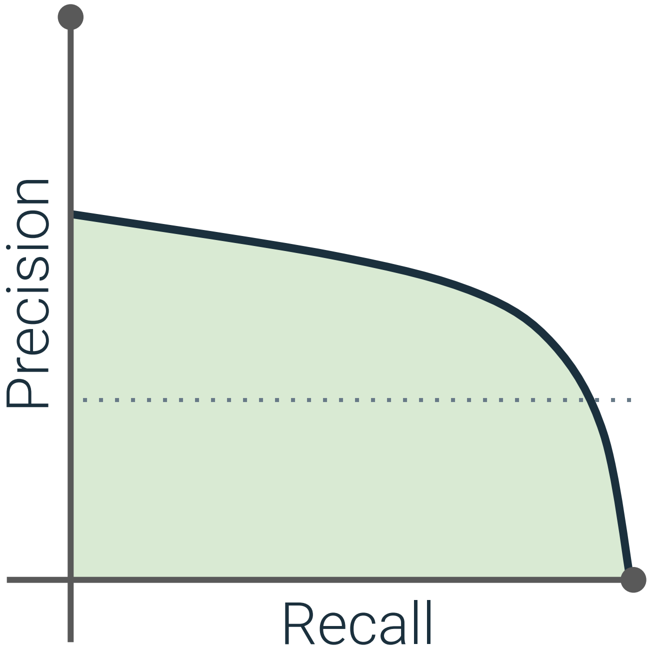
HARVARD MEDICAL SCHOOL | BLAVATNIK INSTITUTE BIOMEDICAL INFORMATICS

COLUMBIA | COLUMBIA UNIVERSITY DEPARTMENT OF BIOMEDICAL INFORMATICS

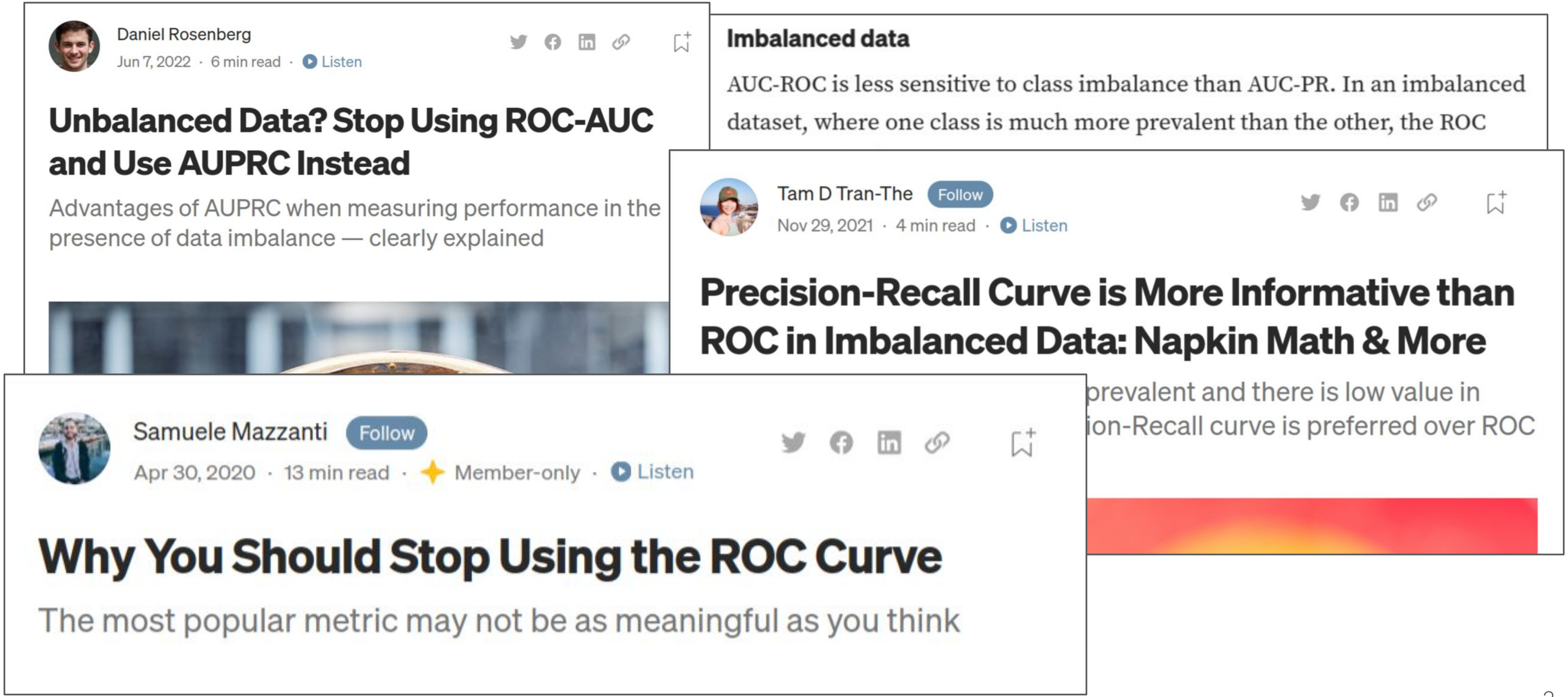
AUROC



AUPRC



AUPRC: Better under class imbalance?



Daniel Rosenberg

Jun 7, 2022 · 6 min read · Listen

Unbalanced Data? Stop Using ROC-AUC and Use AUPRC Instead

Advantages of AUPRC when measuring performance in the presence of data imbalance — clearly explained

Imbalanced data

AUC-ROC is less sensitive to class imbalance than AUC-PR. In an imbalanced dataset, where one class is much more prevalent than the other, the ROC

Tam D Tran-The Follow

Nov 29, 2021 · 4 min read · Listen

Precision-Recall Curve is More Informative than ROC in Imbalanced Data: Napkin Math & More

prevalent and there is low value in

ion-Recall curve is preferred over ROC

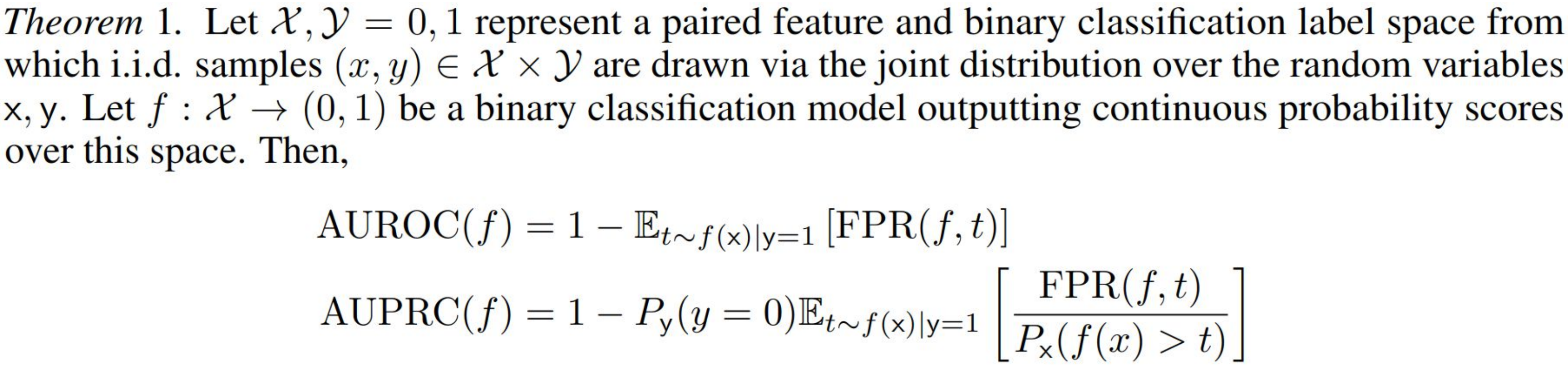
Samuele Mazzanti Follow

Apr 30, 2020 · 13 min read · Member-only · Listen

Why You Should Stop Using the ROC Curve

The most popular metric may not be as meaningful as you think

A probabilistic view reveals not!

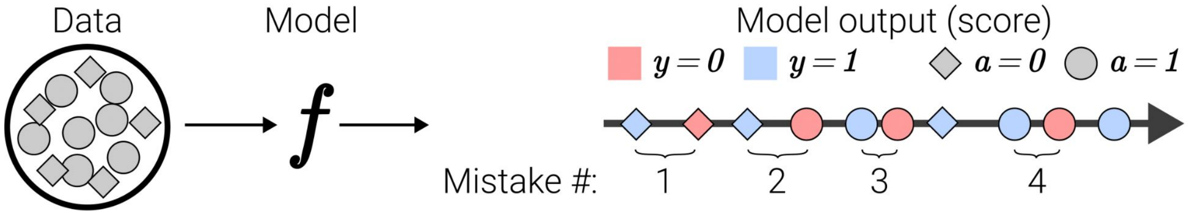


Theorem 1. Let $\mathcal{X}, \mathcal{Y} = 0, 1$ represent a paired feature and binary classification label space from which i.i.d. samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are drawn via the joint distribution over the random variables x, y. Let $f : \mathcal{X} \to (0, 1)$ be a binary classification model outputting continuous probability scores over this space. Then,

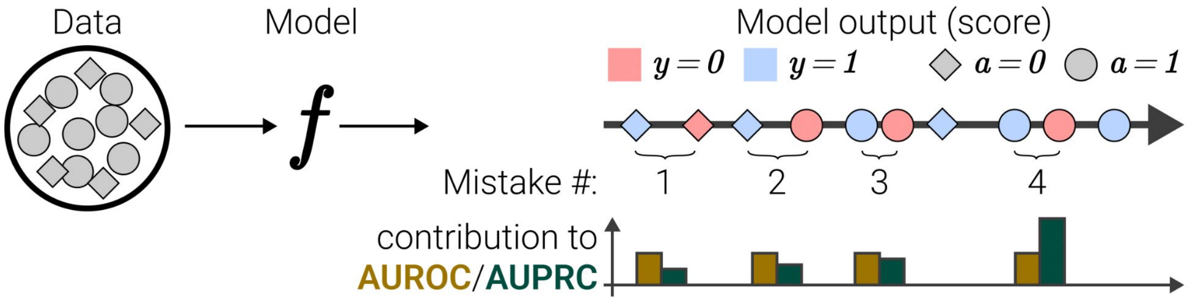
$$\mathrm{AUROC}(f) = 1 - \mathbb{E}_{t \sim f(\mathsf{x}) | \mathsf{y}=1} \left[\mathrm{FPR}(f, t) \right]$$

$$\mathrm{AUPRC}(f) = 1 - P_{\mathsf{y}}(y = 0) \mathbb{E}_{t \sim f(\mathsf{x}) | \mathsf{y}=1} \left[\frac{\mathrm{FPR}(f, t)}{P_{\mathsf{x}}(f(x) > t)} \right]$$

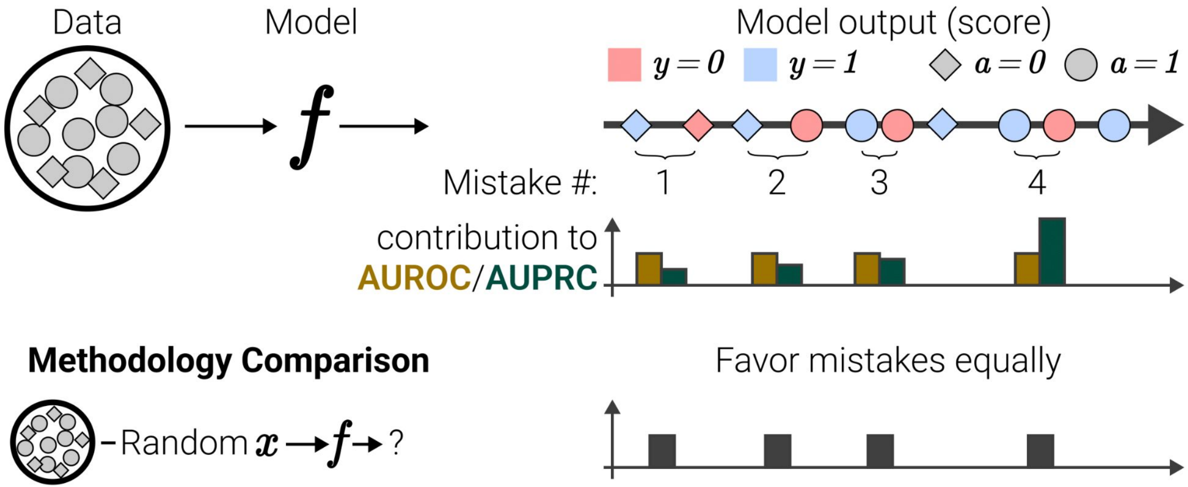
Mistake order correction reveals AUPRC biases



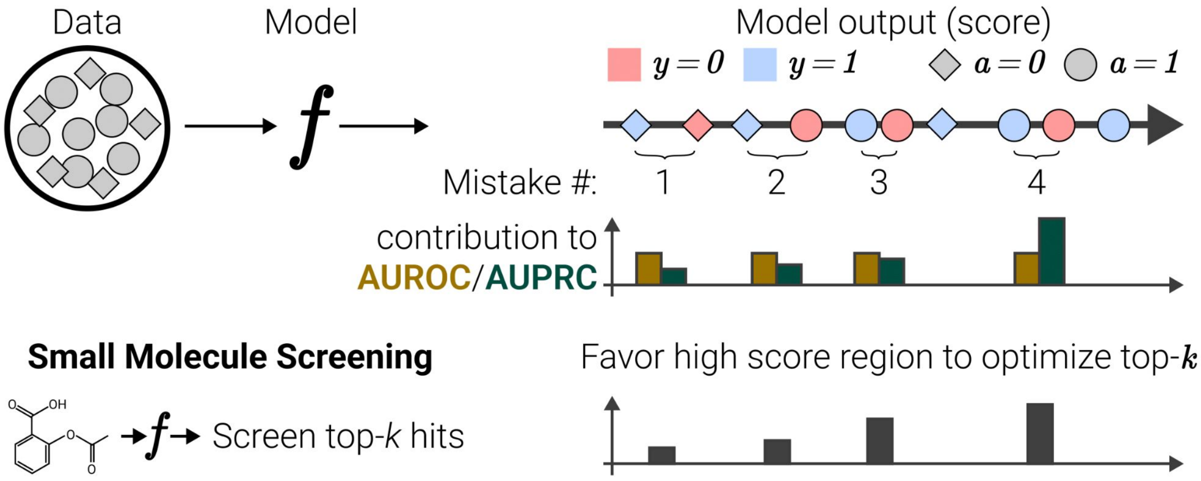
Mistake order correction reveals AUPRC biases



Mistake order correction reveals AUPRC biases



Mistake order correction reveals AUPRC biases

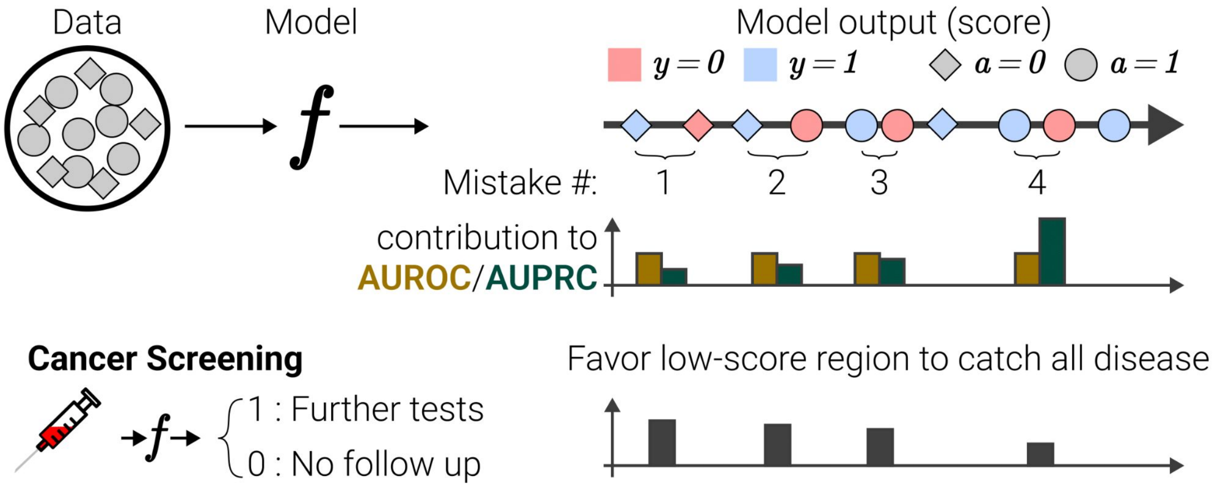


Small Molecule Screening

Screen top-*k* hits

Favor high score region to optimize top-k

Mistake order correction reveals AUPRC biases



Data

Model

Model output (score)

🟥 $y = 0$ 🟦 $y = 1$ ◇ $a = 0$ ○ $a = 1$

Mistake #: 1 2 3 4

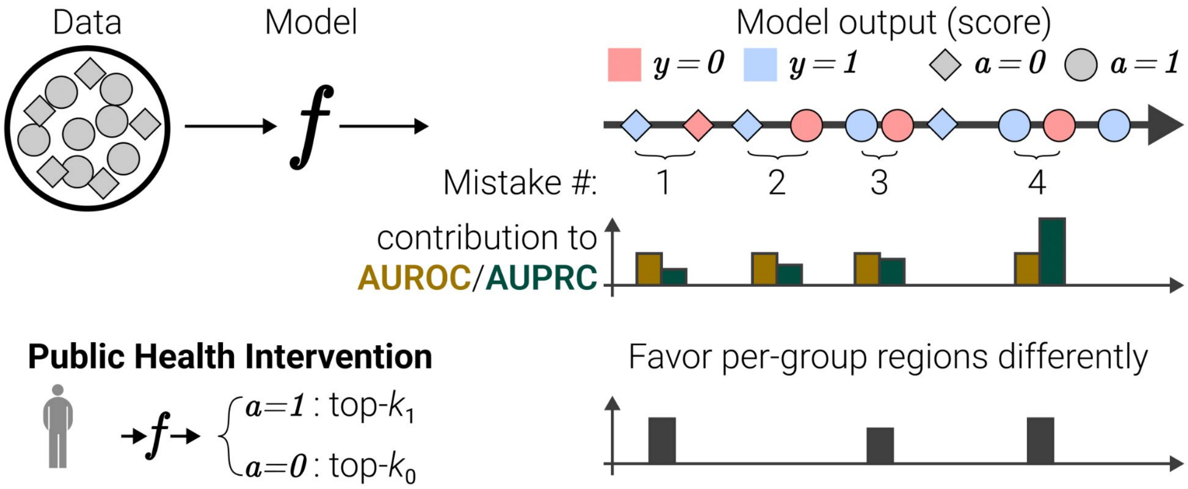
contribution to AUROC/AUPRC

Cancer Screening

$\to f \to \begin{cases} 1 : \text{Further tests} \\ 0 : \text{No follow up} \end{cases}$

Favor low-score region to catch all disease

Mistake order correction reveals AUPRC biases



Data

Model

f

Model output (score)

🟥 $y=0$ 🟦 $y=1$ ◇ $a=0$ ○ $a=1$

Mistake #: 1 2 3 4

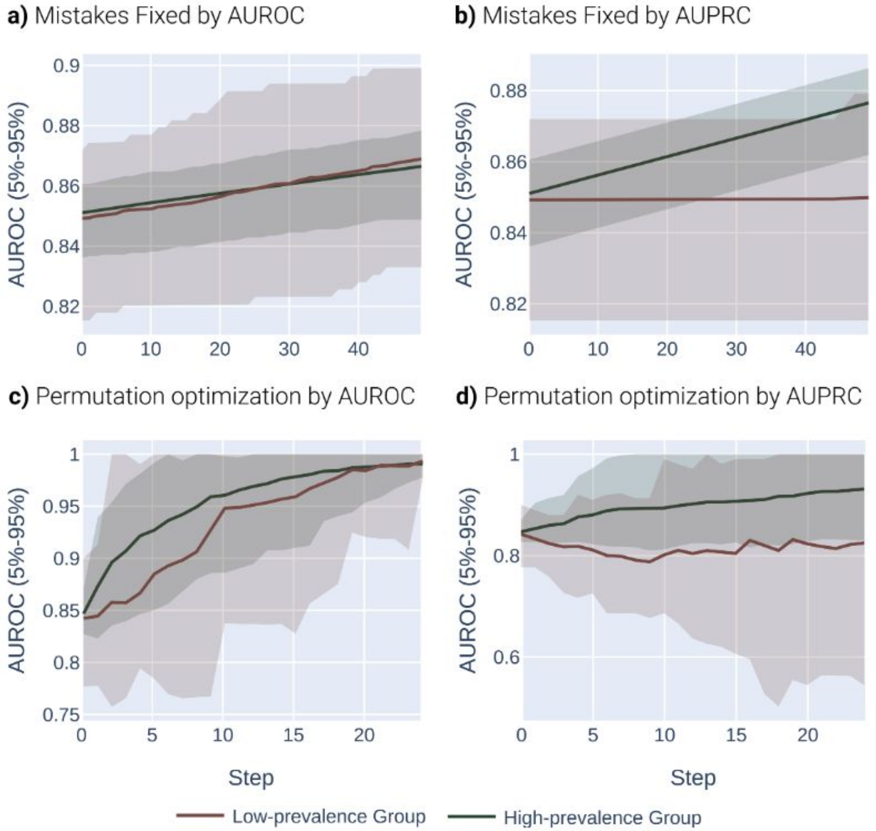
contribution to **AUROC**/**AUPRC**

Public Health Intervention

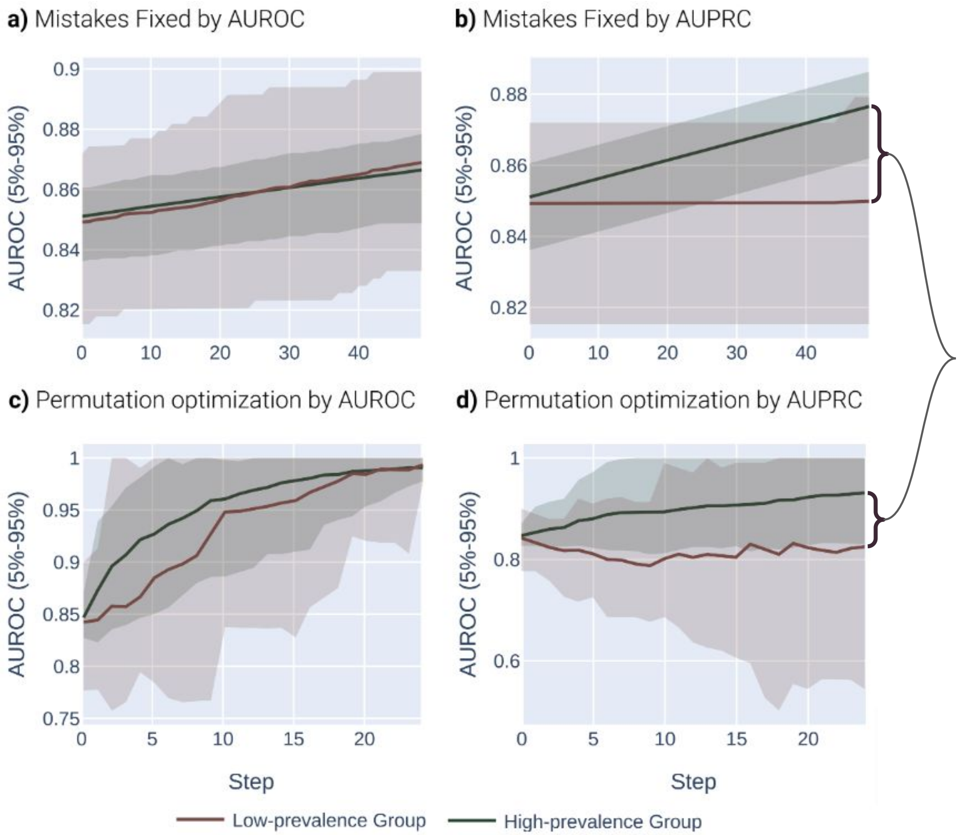
$\rightarrow f \rightarrow \begin{cases} a=1: \text{top-}k_1 \\ a=0: \text{top-}k_0 \end{cases}$

Favor per-group regions differently

Synthetic experiments verify this bias

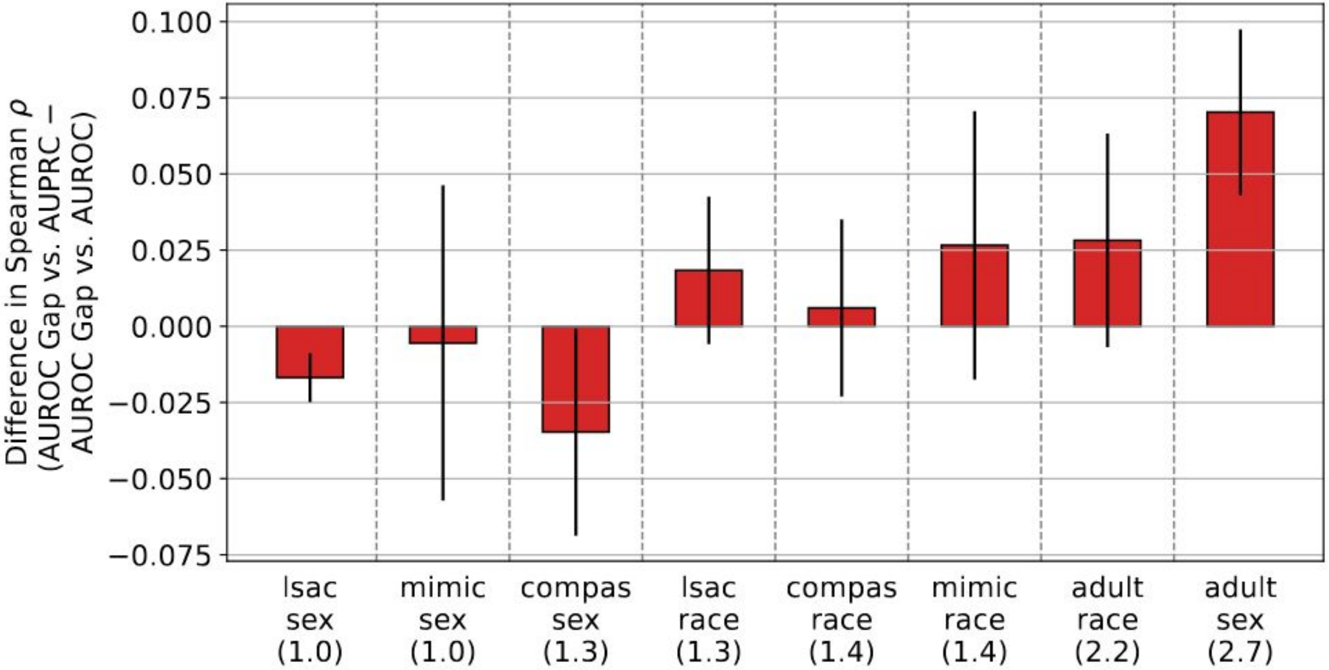


Synthetic experiments verify this bias



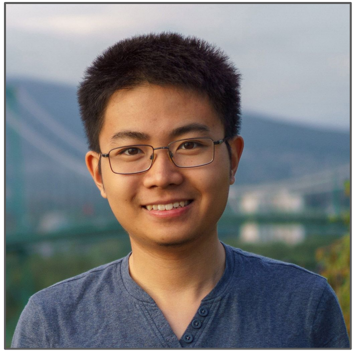
Extent to which AUPRC favors high-prevalence group!

This is not just synthetic -- hyperparameter tuning shows this effect!



Acknowledgements

Matthew McDermott
HMS



Haoran Zhang
MIT

Lasse Hansen
Aarhus University

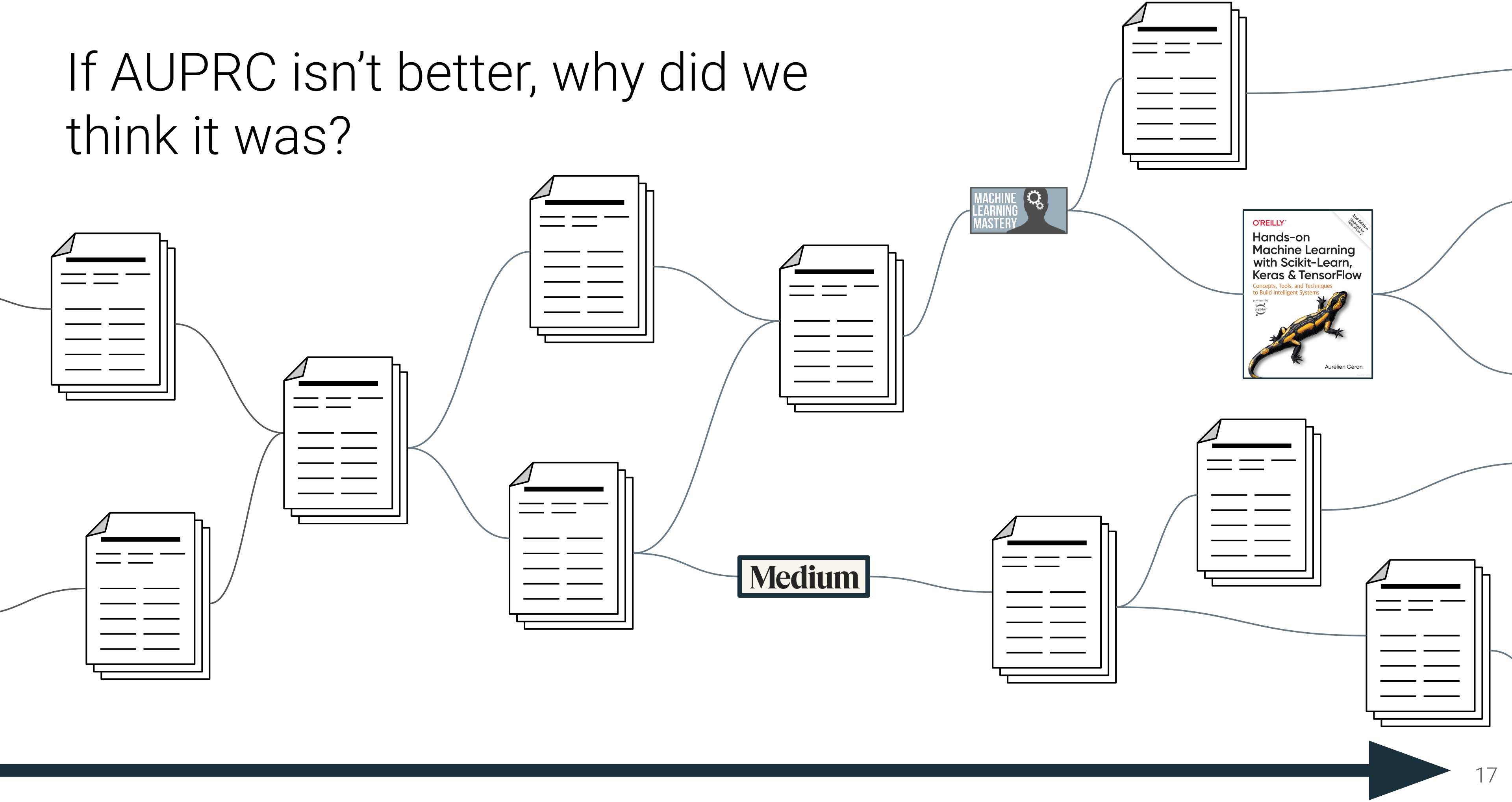
Giovanni Angelotti
IRCCS Humanitas

Jack Gallifant
MIT

Matthew McDermott gratefully acknowledges support from a Berkowitz Postdoctoral Fellowship. Jack Gallifant is funded by the National Institute of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701

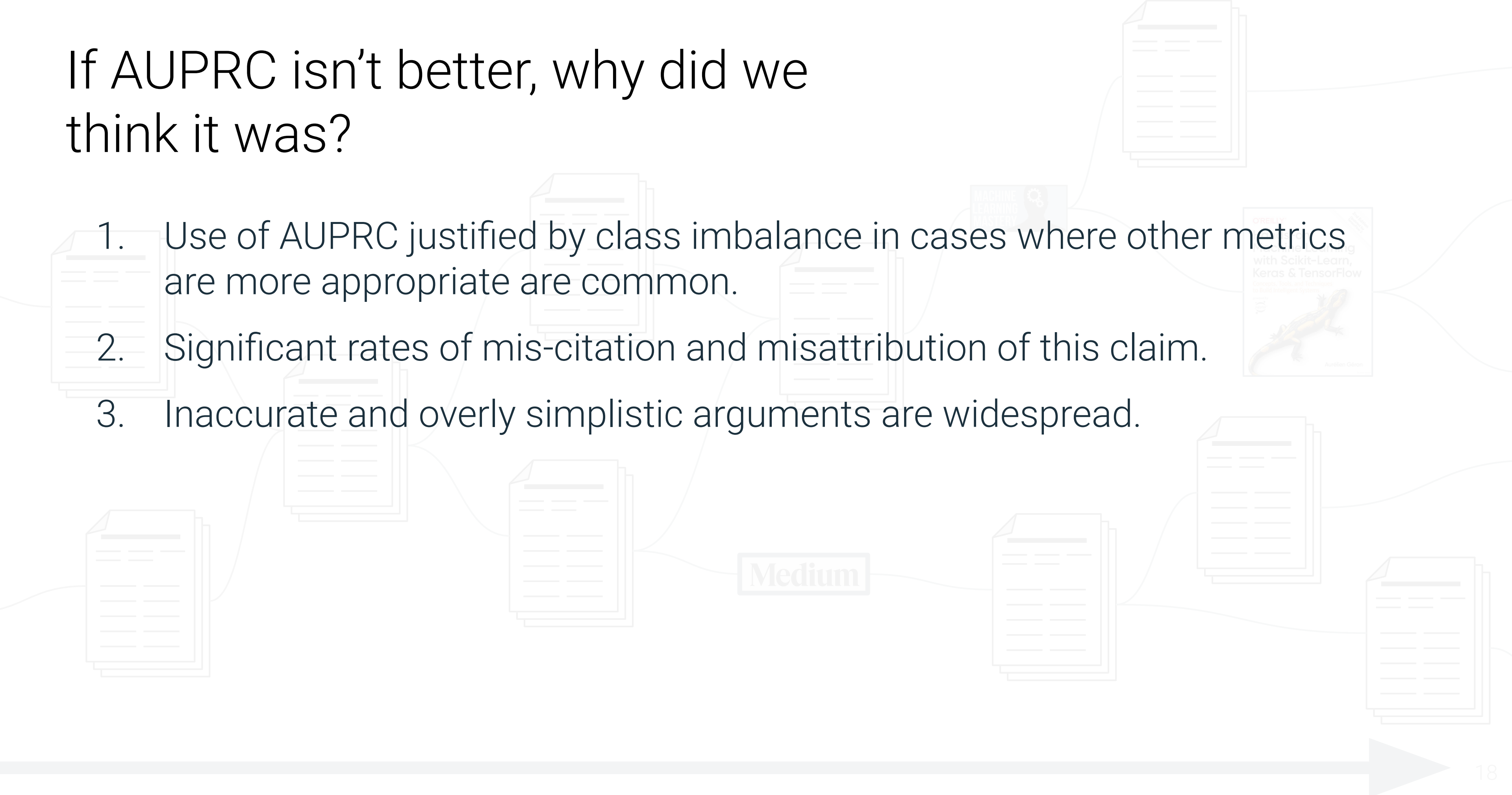
If AUPRC isn’t better, why did we think it was?

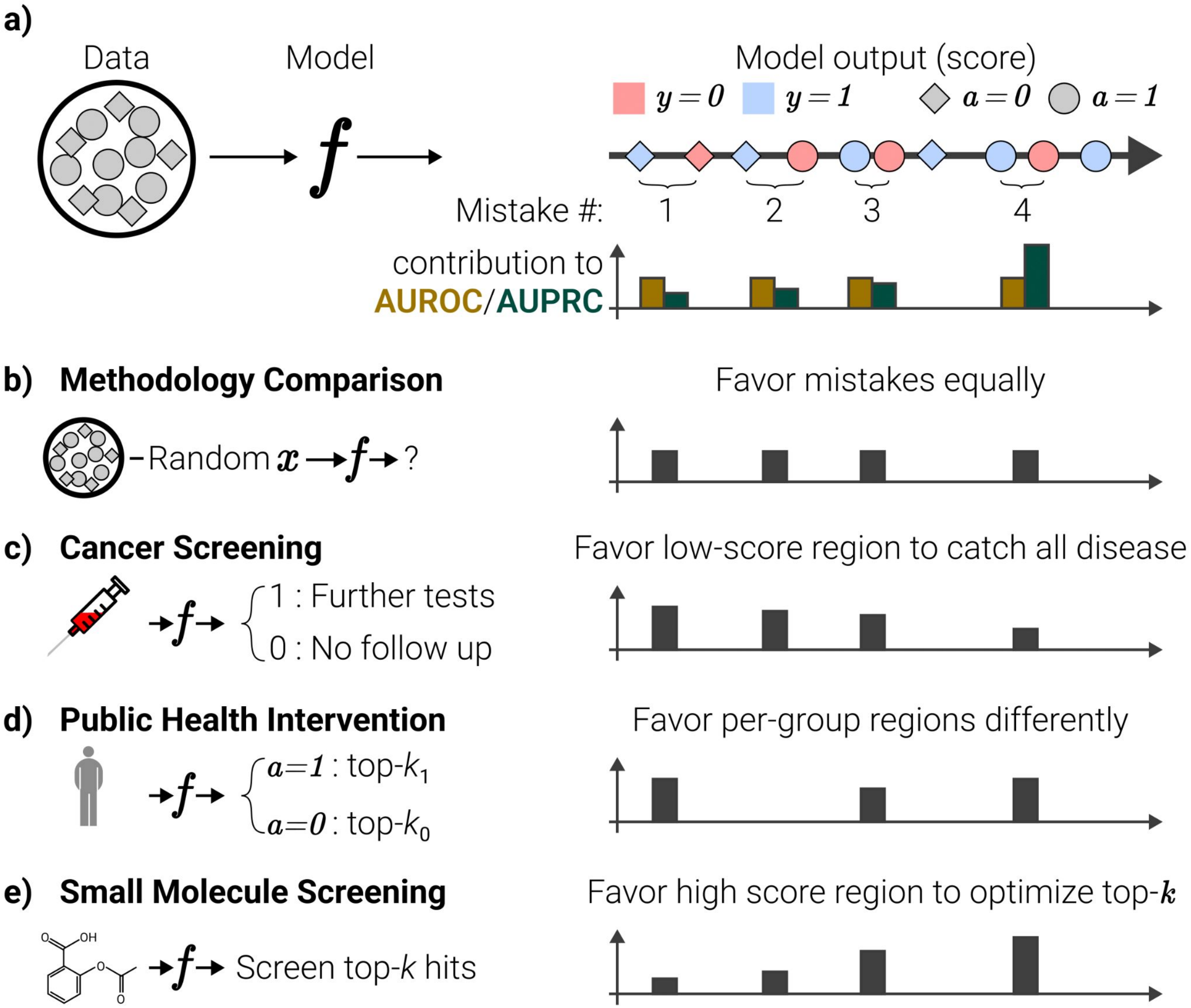
If AUPRC isn’t better, why did we think it was?


MACHINE LEARNING MASTERY
O'REILLY
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
Aurélien Géron
Medium

If AUPRC isn’t better, why did we think it was?

1. Use of AUPRC justified by class imbalance in cases where other metrics are more appropriate are common.
2. Significant rates of mis-citation and misattribution of this claim.
3. Inaccurate and overly simplistic arguments are widespread.





a)

Data

Model

$\xrightarrow{}\ f \xrightarrow{}$

Model output (score)

🟥 $y = 0$ 🟦 $y = 1$ ◇ $a = 0$ ○ $a = 1$

Mistake #: 1 2 3 4

contribution to **AUROC**/**AUPRC**

b) **Methodology Comparison**

–Random $\boldsymbol{x} \rightarrow f \rightarrow$?

Favor mistakes equally

c) **Cancer Screening**

$\rightarrow f \rightarrow$ 1 : Further tests

0 : No follow up

Favor low-score region to catch all disease

d) **Public Health Intervention**

$\rightarrow f \rightarrow$ $a=1$: top-k_1

$a=0$: top-k_0

Favor per-group regions differently

e) **Small Molecule Screening**

$\rightarrow f \rightarrow$ Screen top-k hits

Favor high score region to optimize top-$\boldsymbol{k}$