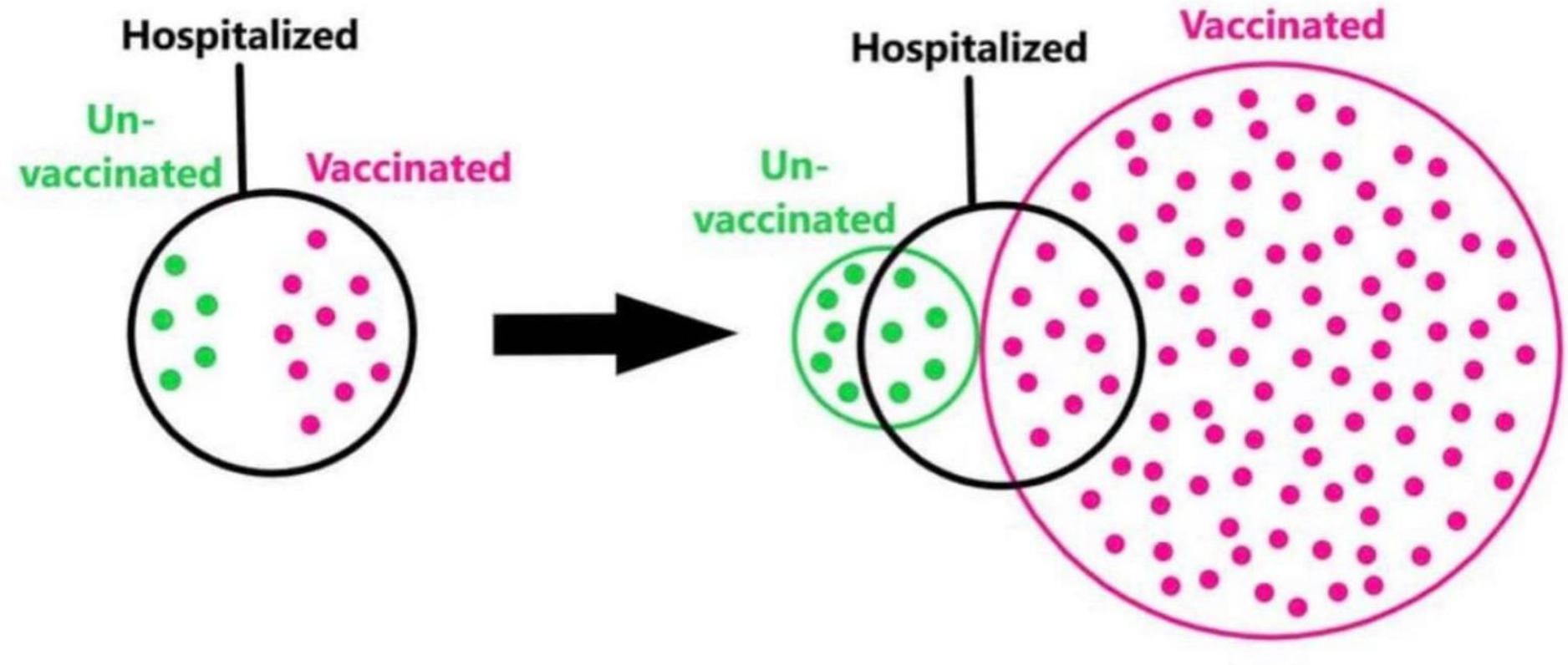
SoK: The Good, The Bad, and The Unbalanced

Measuring Structural Limitations of Deepfake Media Datasets

Seth Layton, Tyler Tucker, Daniel Olszewski, Kevin Warren, Kevin Butler, Patrick Traynor

The University of Florida

Motivation


Hospitalized
Un-
vaccinated
Vaccinated
Hospitalized
Vaccinated
Un-
vaccinated

Background

Class Distribution and Bias

Metric Usage Impacts

Base-Rate Contextualization

Suggestions and Greater Impacts

Background

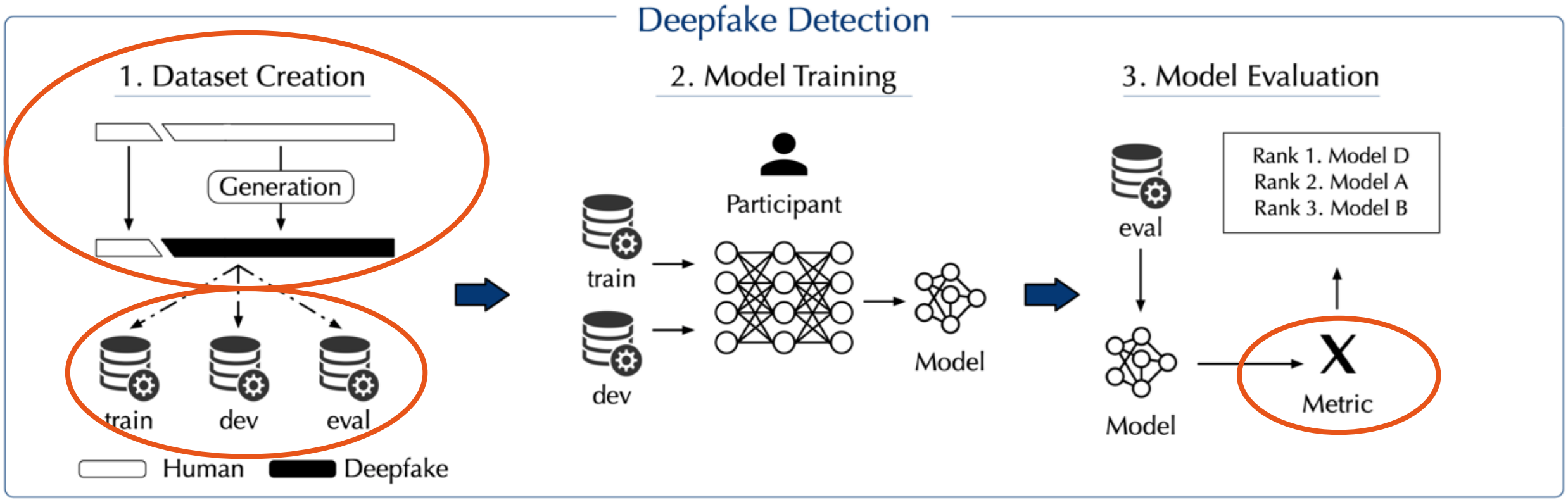
Class Distribution and Bias

Metric Usage Impacts

Base-Rate Contextualization

Suggestions and Greater Impacts

Deepfake Datasets



1. Class Distributions

2. Metric Usage

3. Base-Rate Contextualization

Background

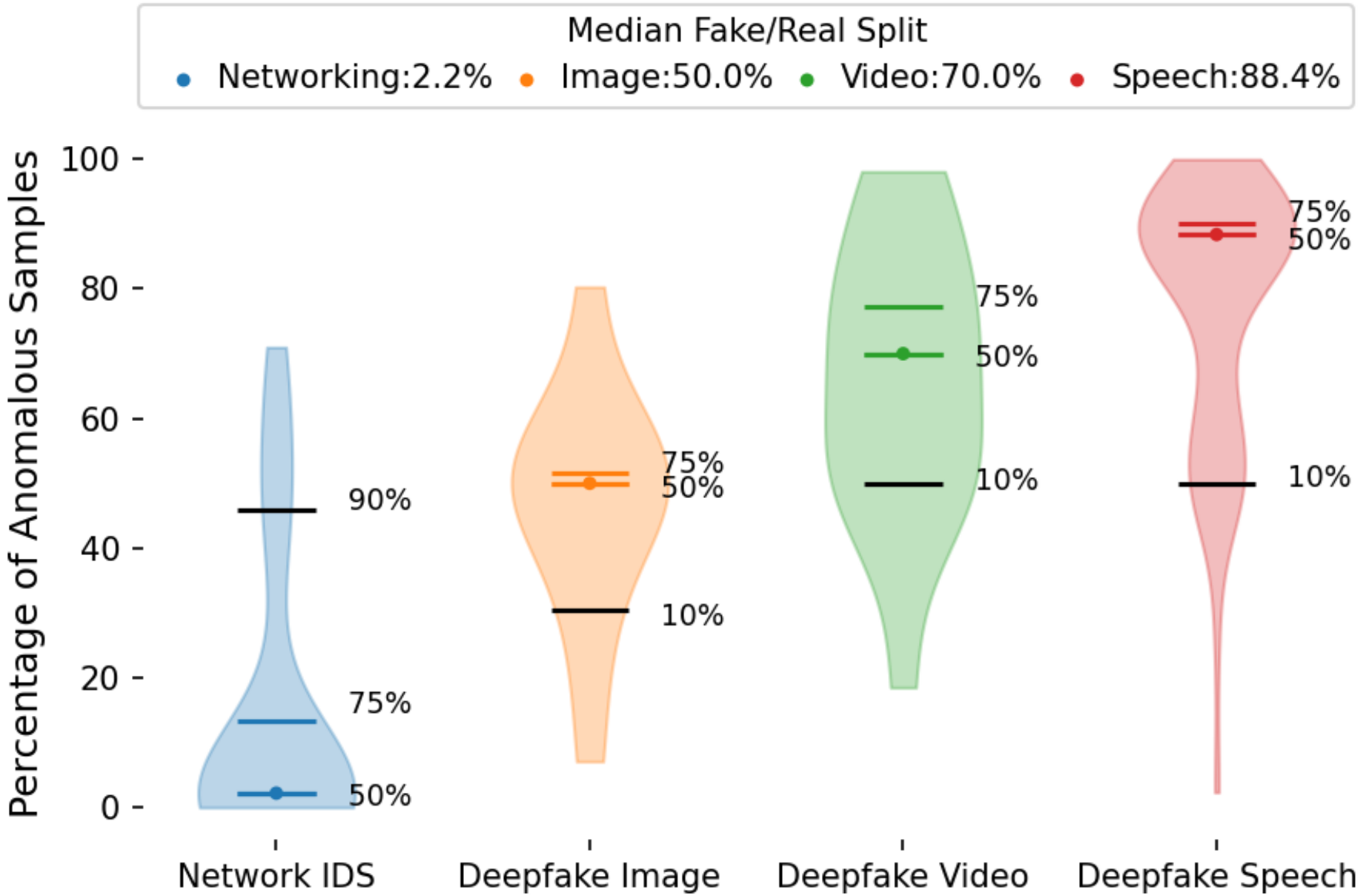
Class Distribution and Bias

Metric Usage Impacts

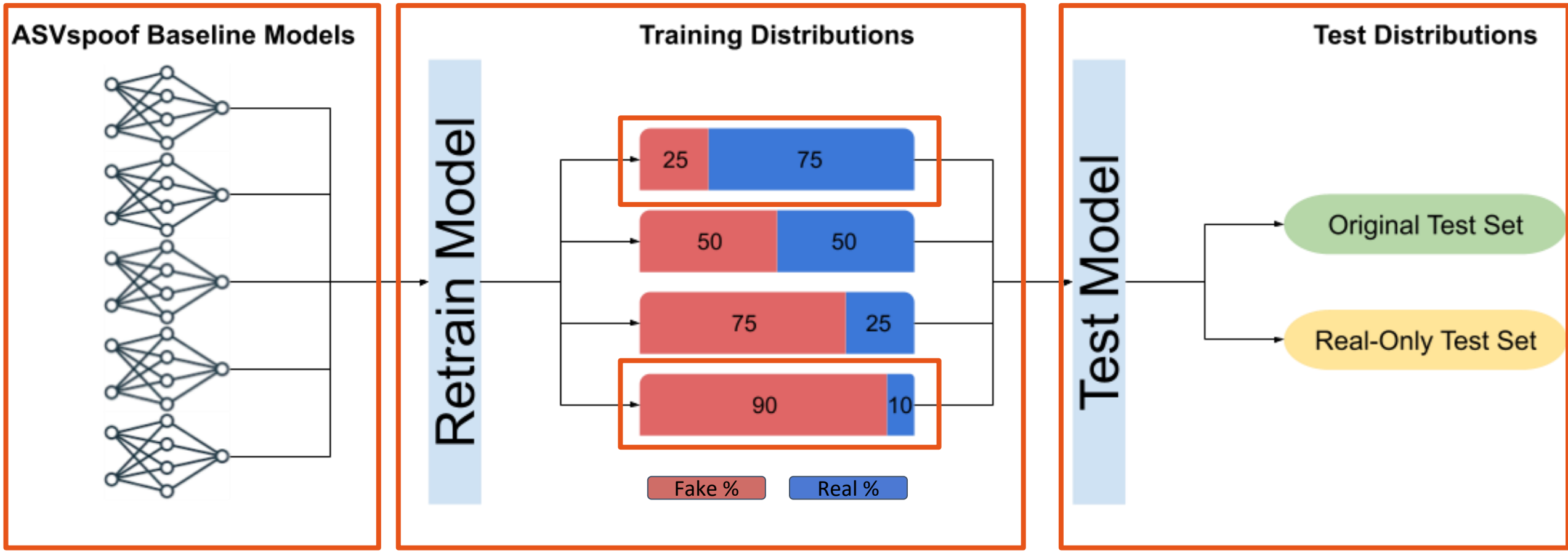
Base-Rate Contextualization

Suggestions and Greater Impacts

Class Distribution

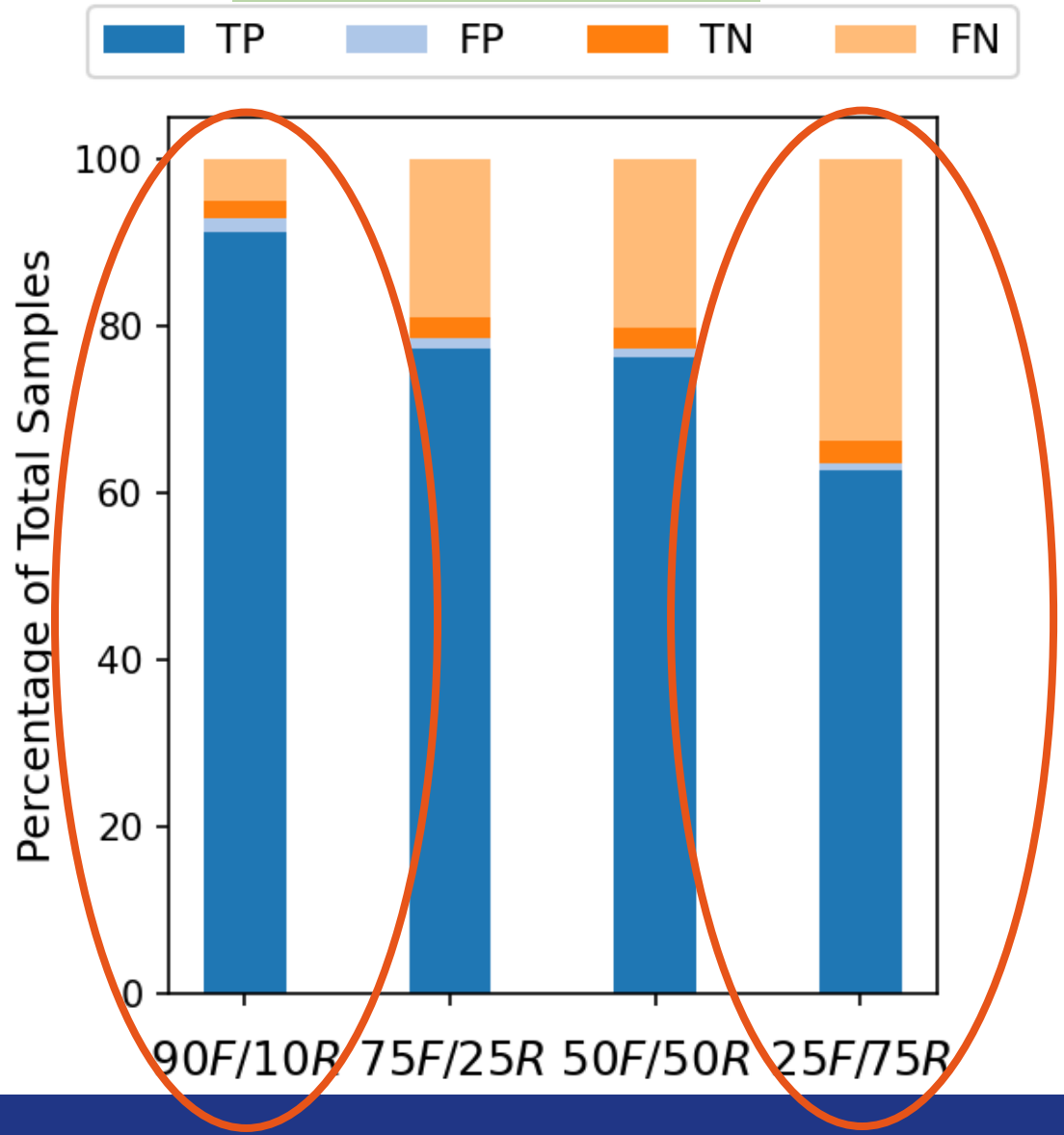


Impact of Class Distribution

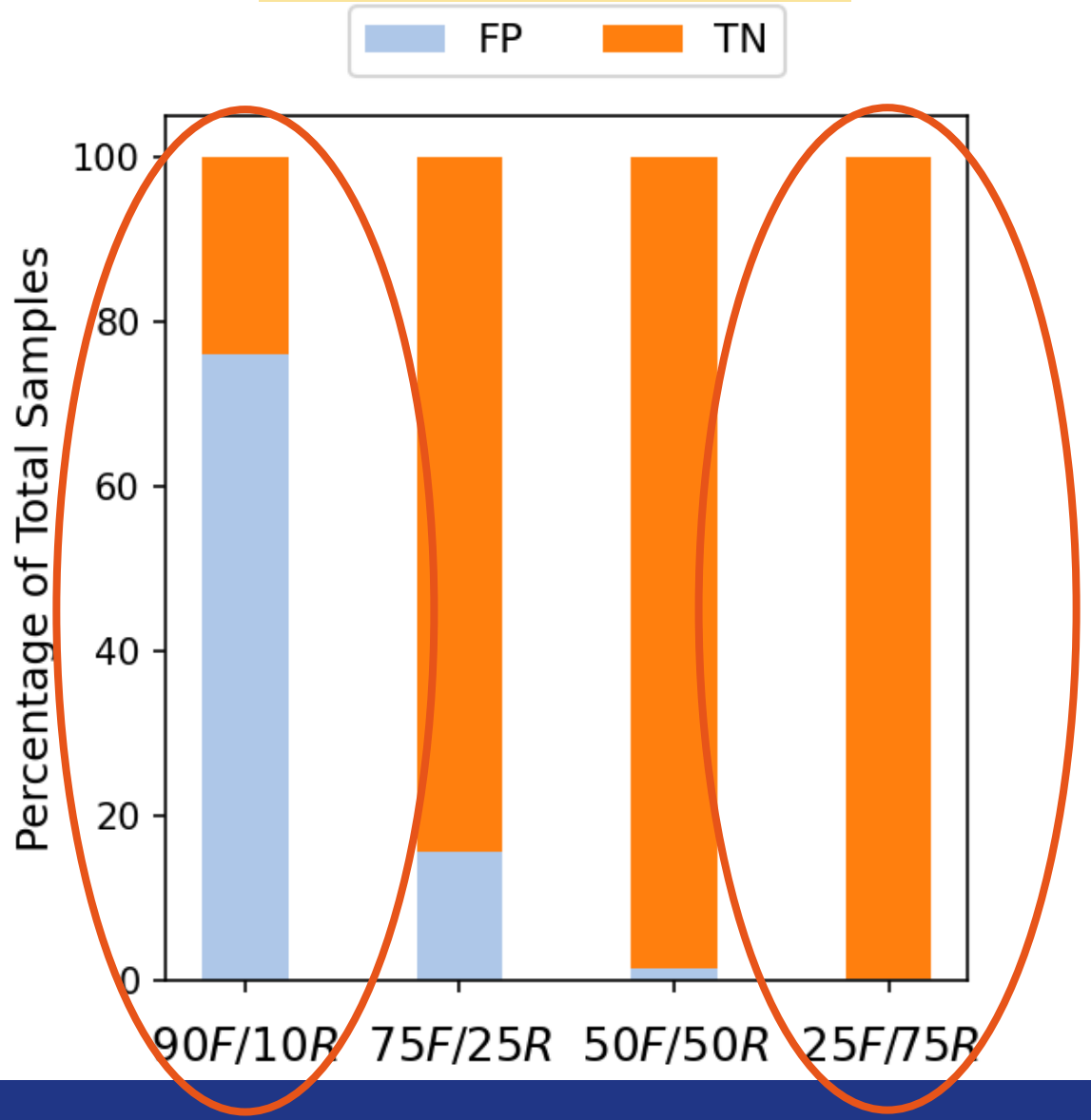


Impact of Class Distribution

Original Test Set



Real-Only Test Set



Background

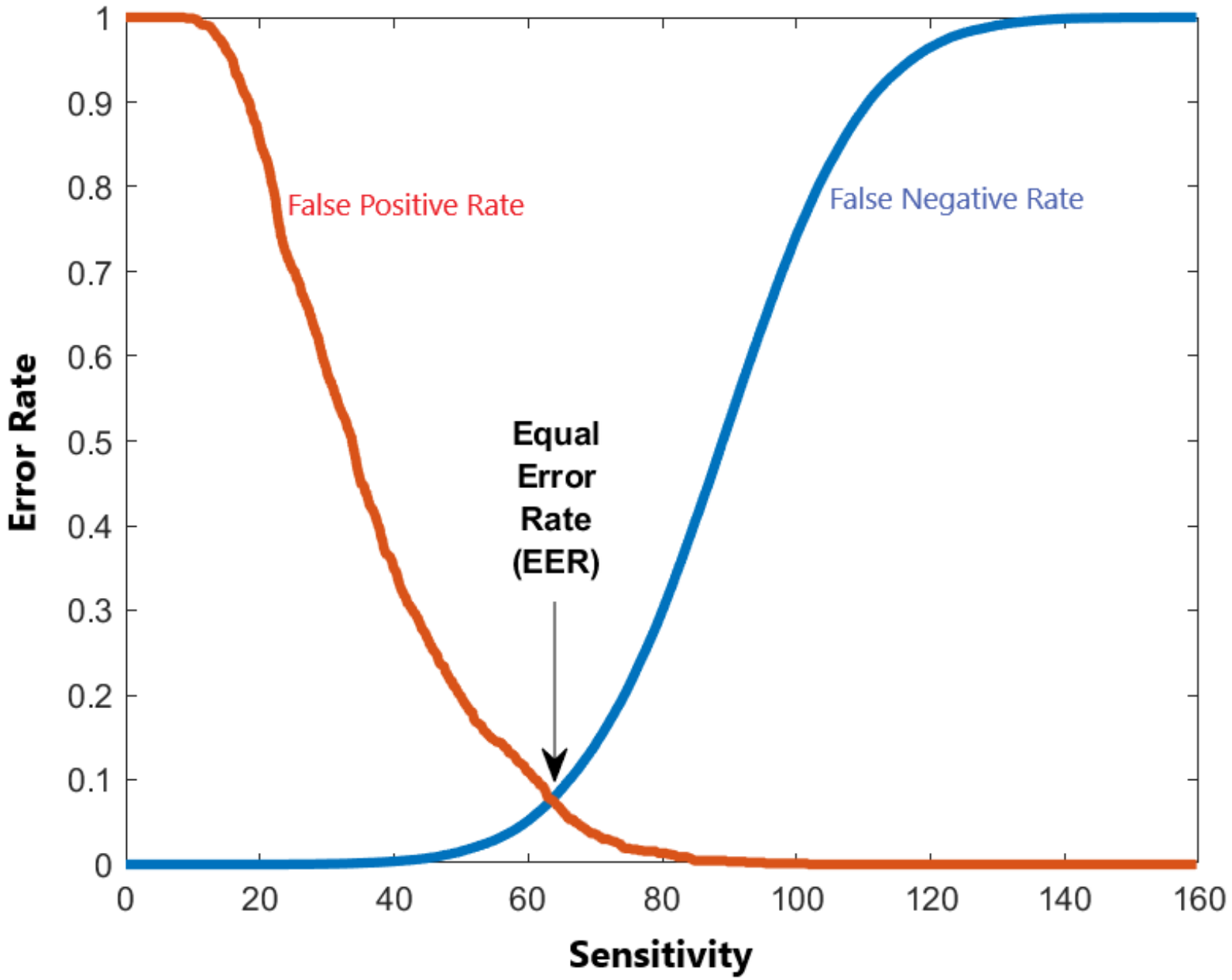
Class Distribution and Bias

Metric Usage Impacts

Base-Rate Contextualization

Suggestions and Greater Impacts

Metrics


False Positive Rate
False Negative Rate
Equal
Error
Rate
(EER)
Error Rate
Sensitivity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160

Metrics

Current Deepfake Detectors claim EERs < 1%

Does that mean this is a “solved” space?

Impact of EER

	EER	TPR	FPR
LFCC-GMM	25.5%	44.9%	8.80%
LFCC-LCNN	22.9%	94.7%	41.7%
Absolute Difference	2.6%	49.8%	32.9%

EER Inherently Obfuscates Results

Background

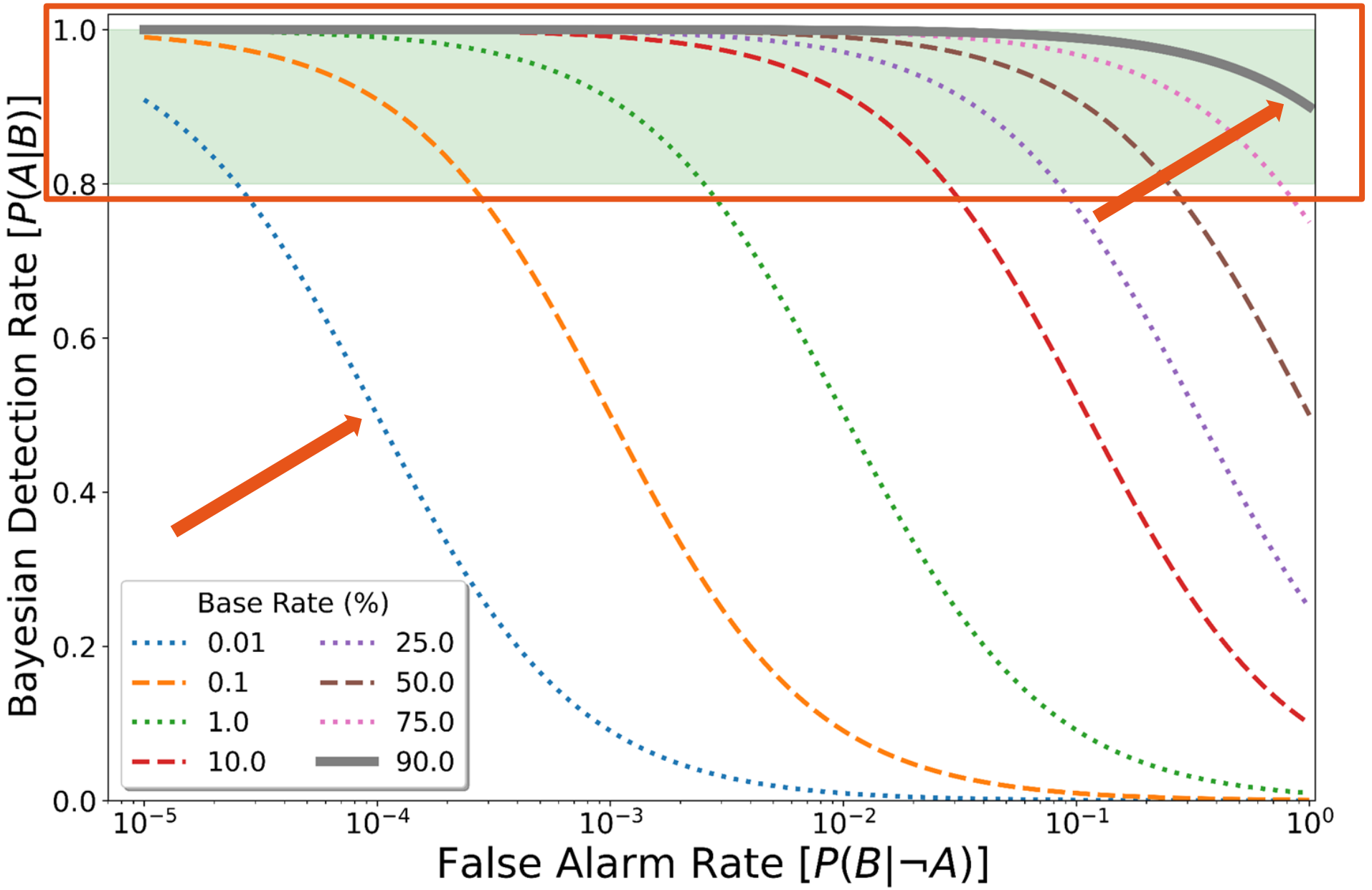
Class Distribution and Bias

Metric Usage Impacts

Base-Rate Contextualization

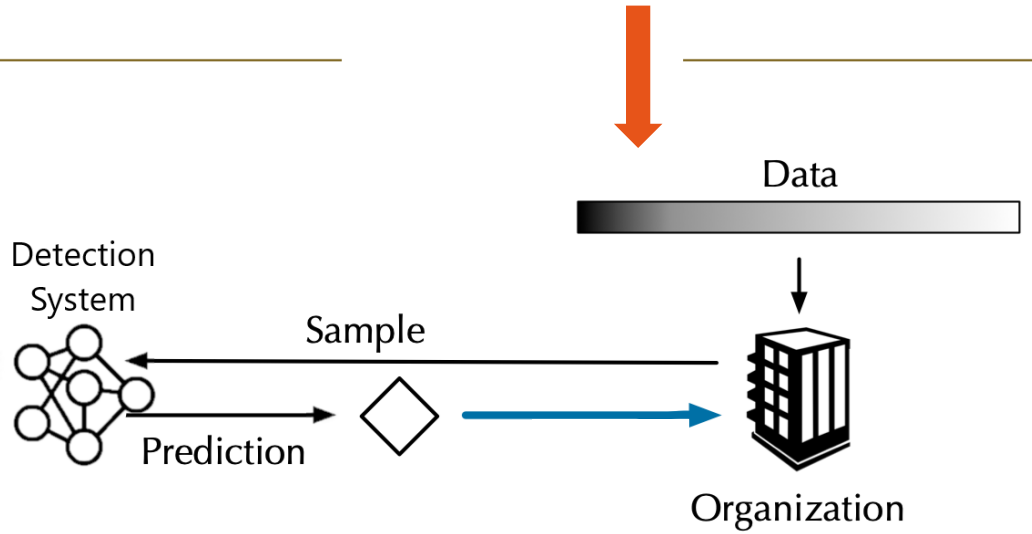
Suggestions and Greater Impacts

Characterizing Model Efficacy



Call Center Scenario

1/1074 Incoming Calls are Deepfakes



~4,400 Calls Monthly

~4 Deepfake Calls Monthly

	EER	True Positive	False Positive
M_{LG}	25.5%	2	387
M_{SW}	4.14%	4	1182

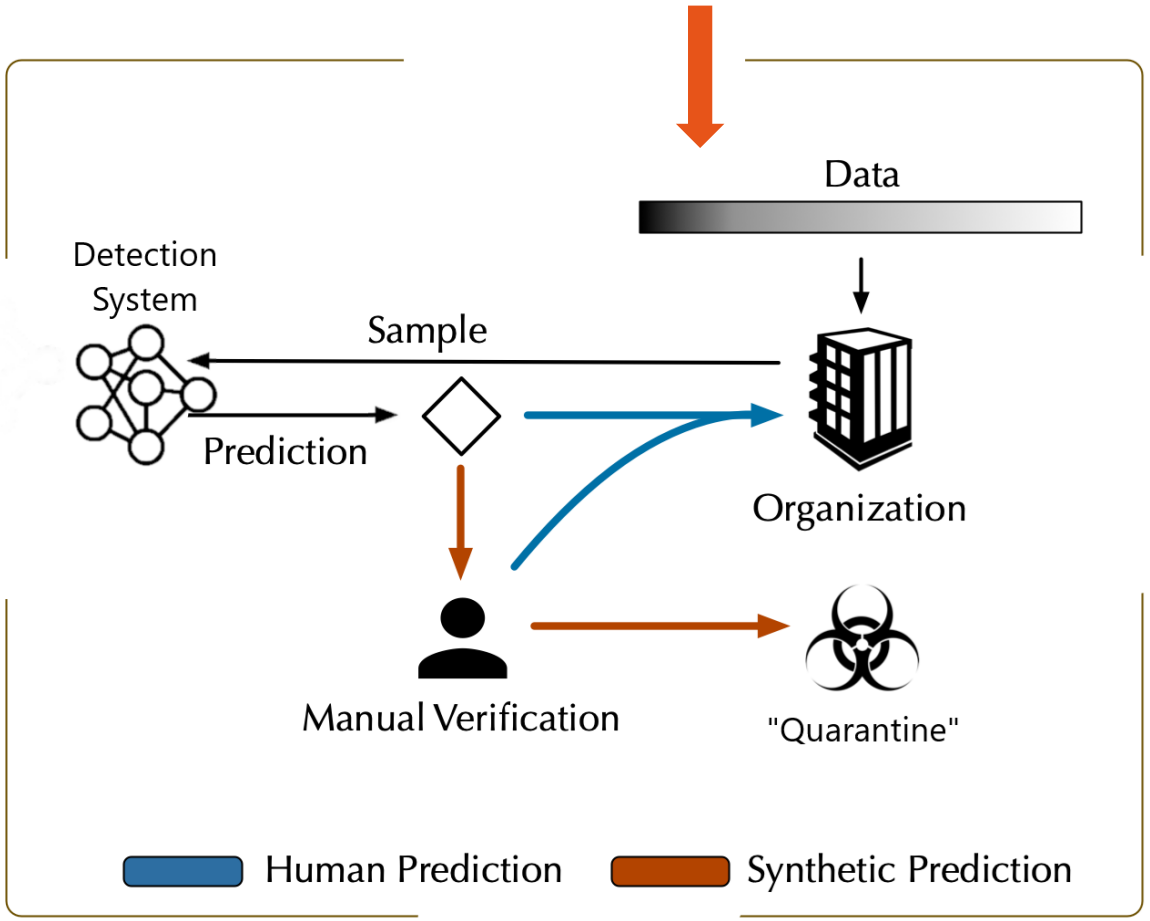
M_{SW}	4.14%

M_{LG} = 1 DF per 200 Alarms

M_{SW} = 1 DF per 333 Alarms

Call Center Scenario

1/11 Incoming Calls are Deepfakes



~4,400 Calls Monthly

~400 Deepfake Calls Monthly

	EER	True Positive	False Positive
M_{LG}	25.5%	180	352
M_{SW}	4.14%	400	1076

M_{LG} = ~ 1 DF per 2 Alarms

M_{SW} = ~ 2 DF per 5 Alarms

Background

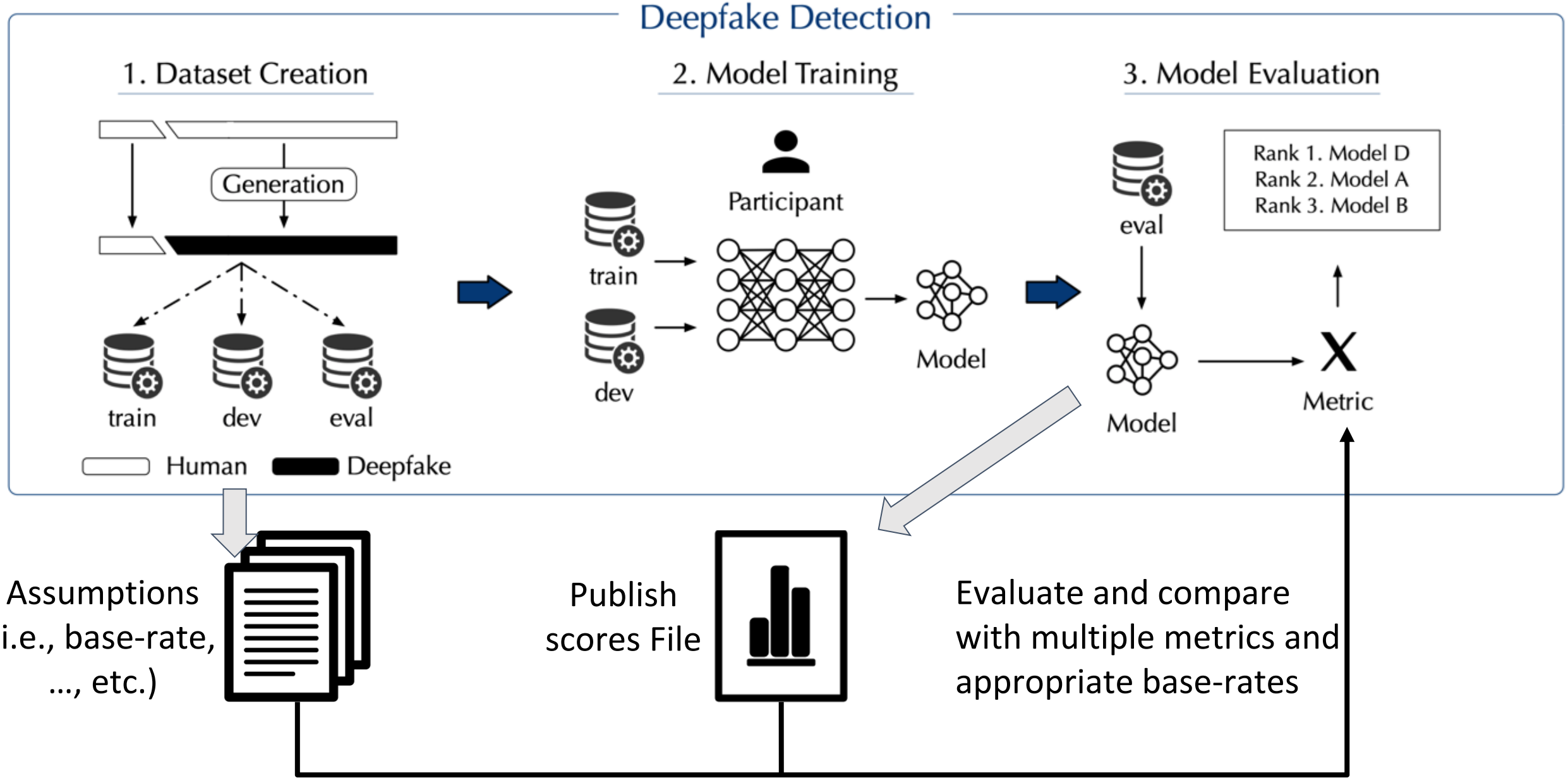
Class Distribution and Bias

Metric Usage Impacts

Base-Rate Contextualization

Suggestions and Greater Impacts

Improving Deepfake Detection



Broader Impacts

1. This applies to more than deepfakes
2. Honestly characterize model performance
3. Facilitate more research
4. Facilitate more meaningful research

In Conclusion

1. Currently used metrics obfuscate results
2. The class distribution in datasets can impose bias on results
3. Contextualization of a dataset is important
4. Current standards of reproducibility and comparability are lacking

sethlayton@ufl.edu