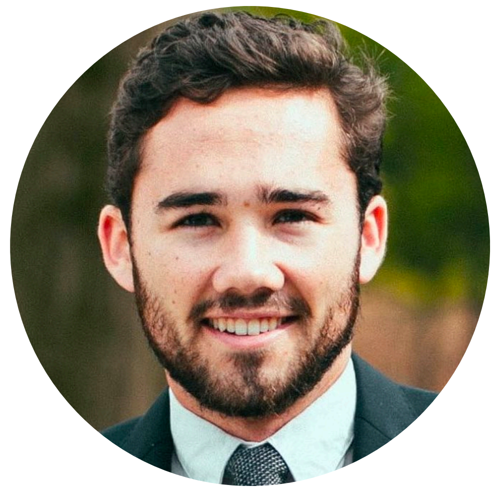
@a_a_cabrera

IEEE VIS 2019

FairVis

Visual Analytics for Discovering Intersectional Bias in Machine Learning

Alex
Cabrera
Carnegie Mellon



Will
Epperson
Georgia Tech

Fred
Hohman
Georgia Tech

Minsuk
Kahng
Oregon State

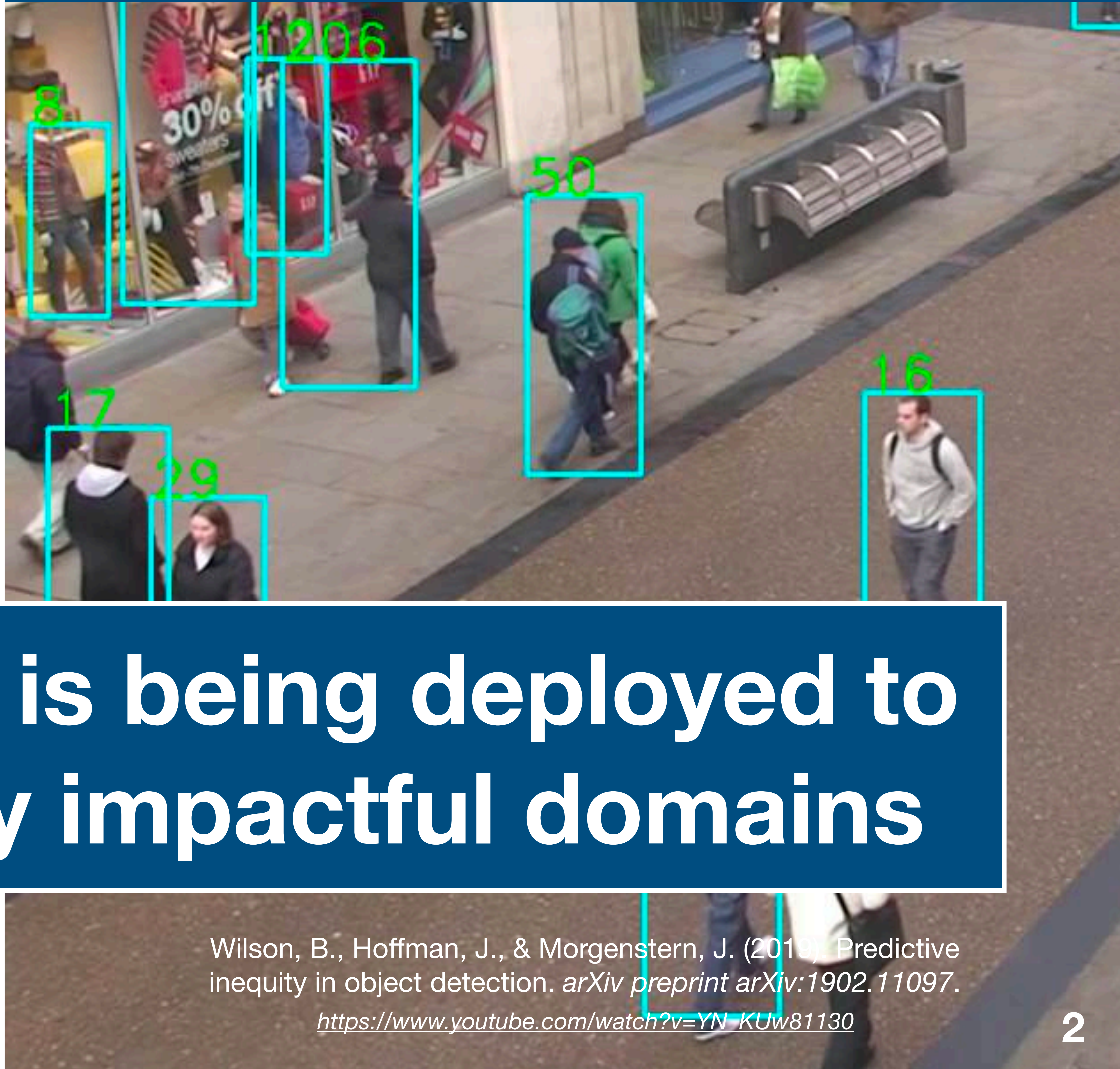
Jamie
Morgenstern
Univ. of Washington

Polo
Chau
Georgia Tech

Carnegie Mellon University
Georgia Tech
Oregon State University
UNIVERSITY of WASHINGTON

Recidivism Prediction

Self-Driving Cars



Machine learning is being deployed to various societally impactful domains

Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. www.propublica.org
https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*.
https://www.youtube.com/watch?v=YN-KUw81130

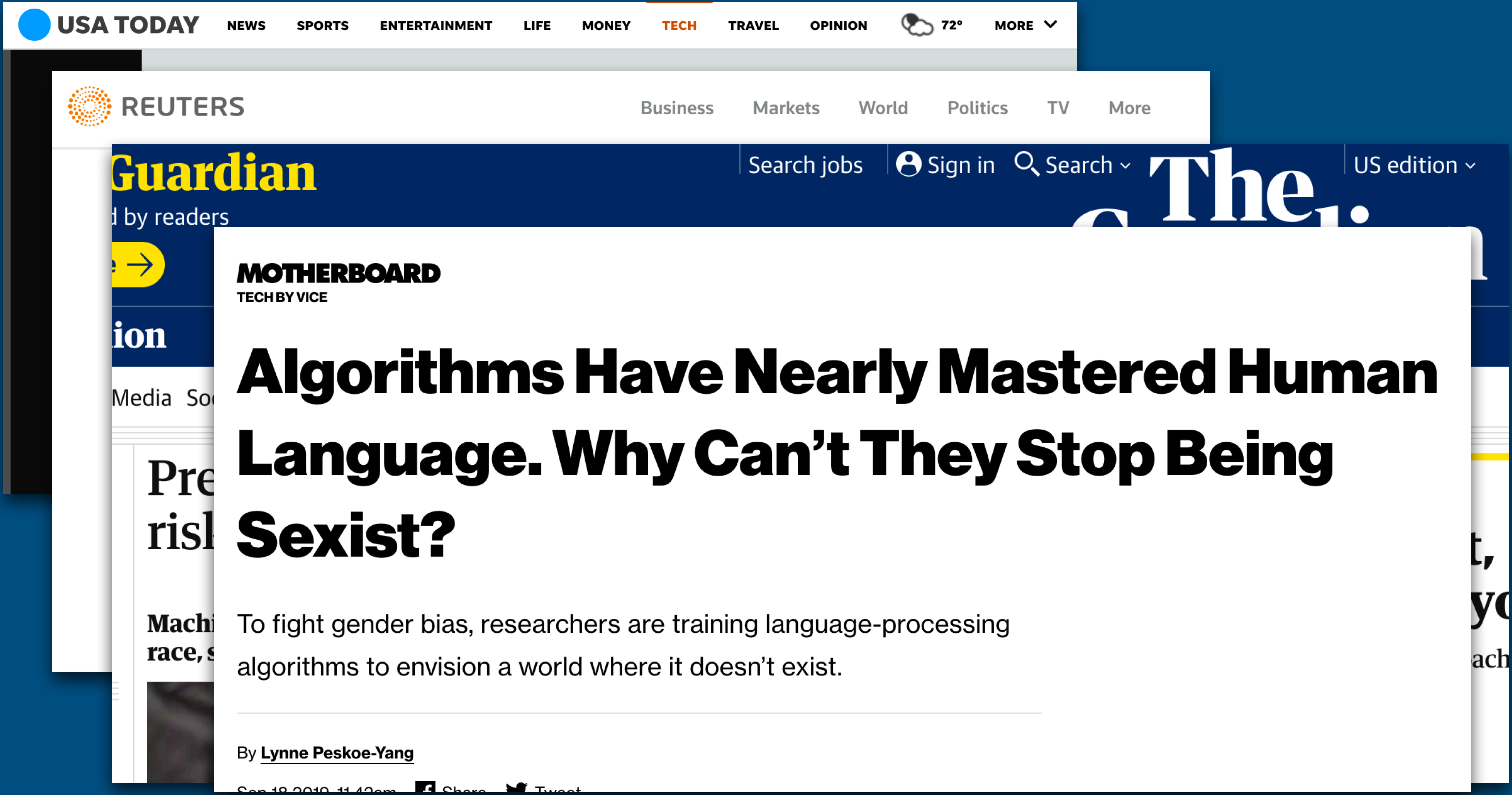
Recidivism Prediction

Self-Driving Cars

Unfortunately, these systems can perpetuate and worsen societal biases

Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. www.propublica.org
https://www.wired.com/story/crime-predicting-algorithms-may-not-outperform-untrained-humans/

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097.*
https://www.youtube.com/watch?v=YN-KUw81130


USA TODAY
NEWS
SPORTS
ENTERTAINMENT
LIFE
MONEY
TECH
TRAVEL
OPINION
72°
MORE
REUTERS
Business
Markets
World
Politics
TV
More
Guardian
by readers
Search jobs
Sign in
Search
US edition
Media
MOTHERBOARD
TECH BY VICE
Algorithms Have Nearly Mastered Human Language. Why Can't They Stop Being Sexist?
To fight gender bias, researchers are training language-processing algorithms to envision a world where it doesn't exist.
By Lynne Peskoe-Yang

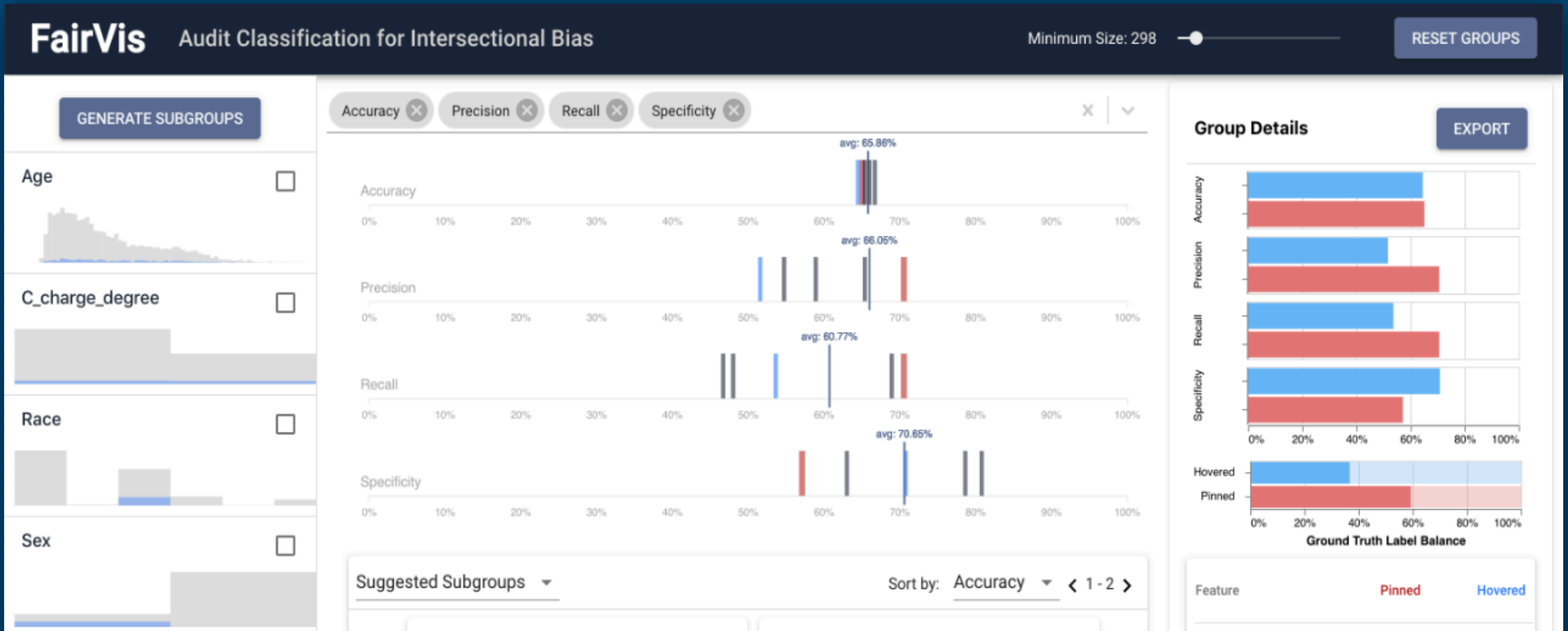
Fairness is a
wicked problem

Issues so complex and dependent on so many factors that it is hard to grasp what exactly the problem is, or how to tackle it.

http://theconversation.com/wicked-problems-and-how-to-solve-them-100047

FairVis

Visual analytics for
discovering biases
in machine learning models

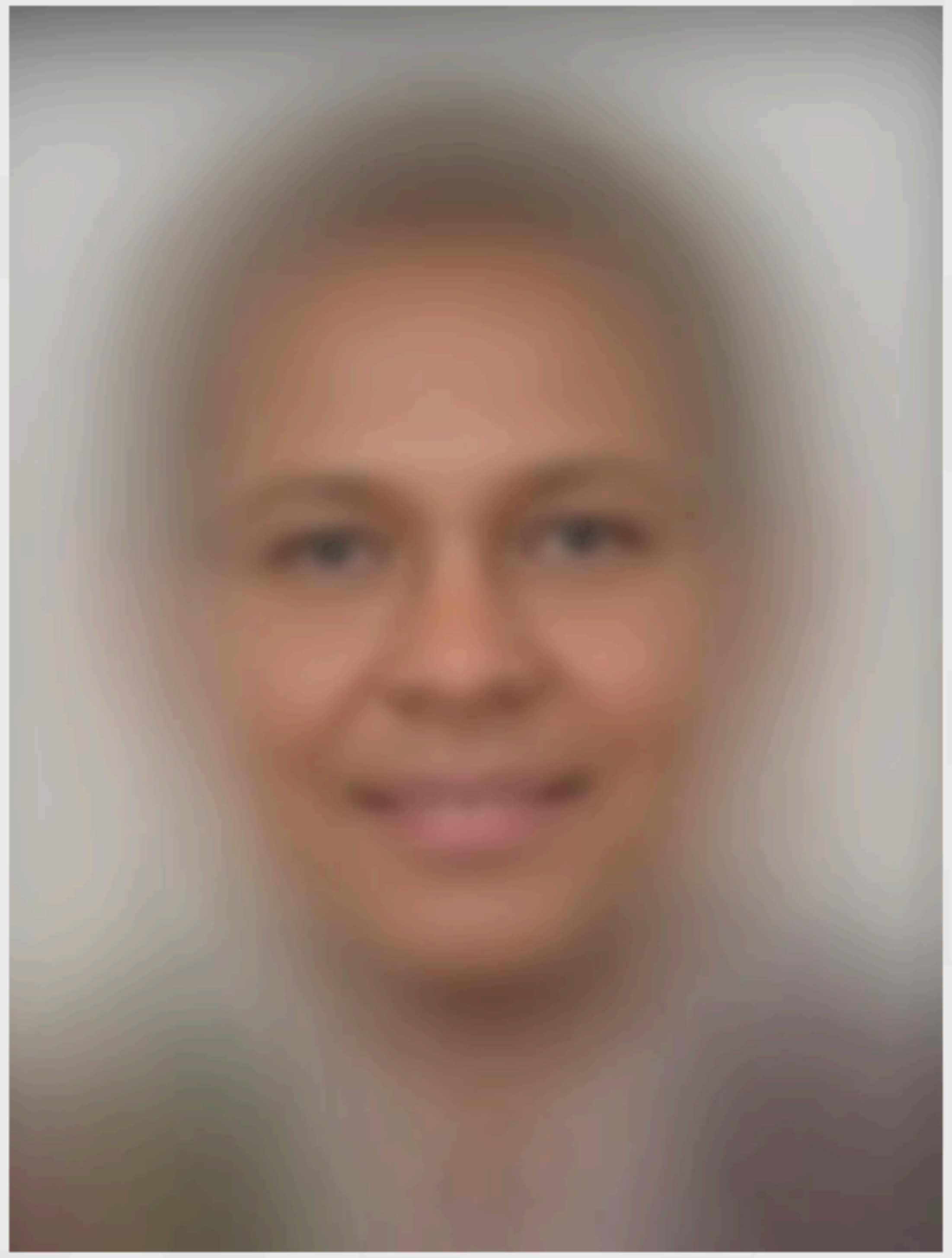


Challenges for Discovering Bias

1

Intersectional bias

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

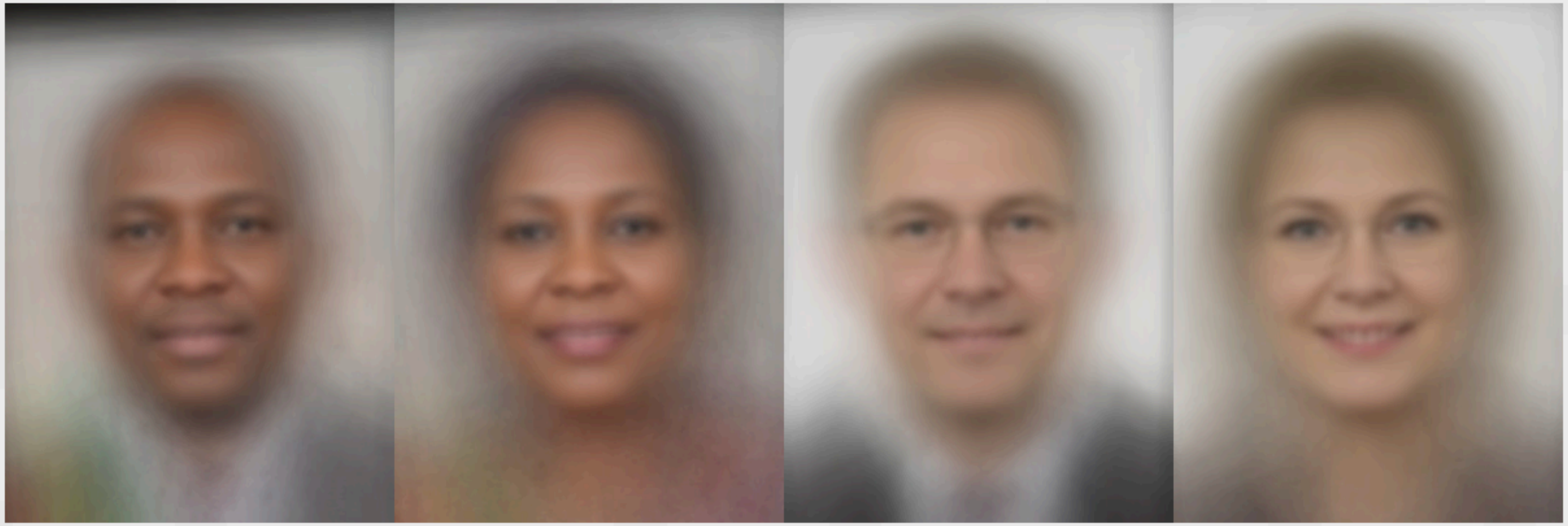


Disparities in Gender Classification

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91).

Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

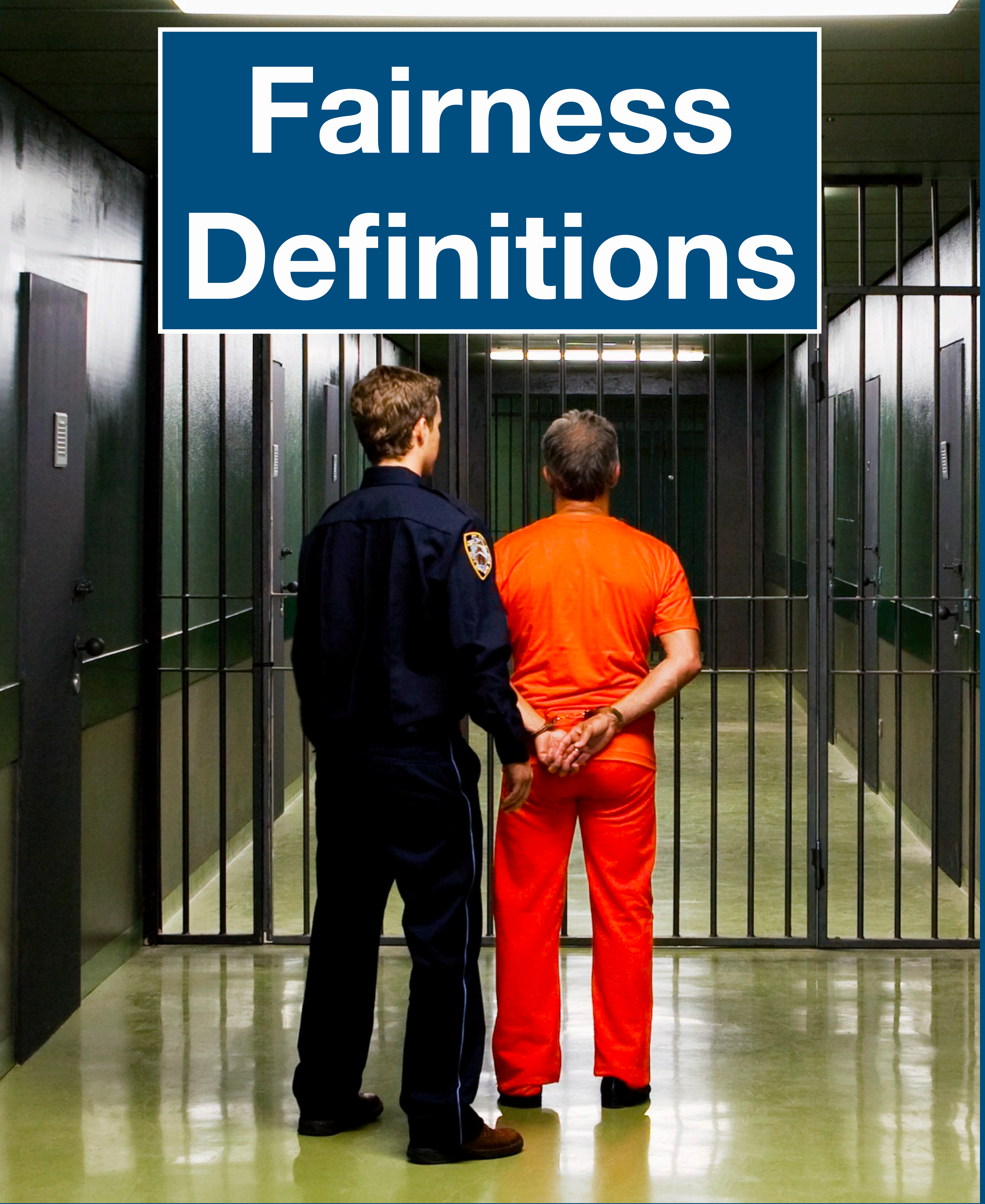
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



2

Defining Fairness

Fairness Definitions



Accuracy? Recall?

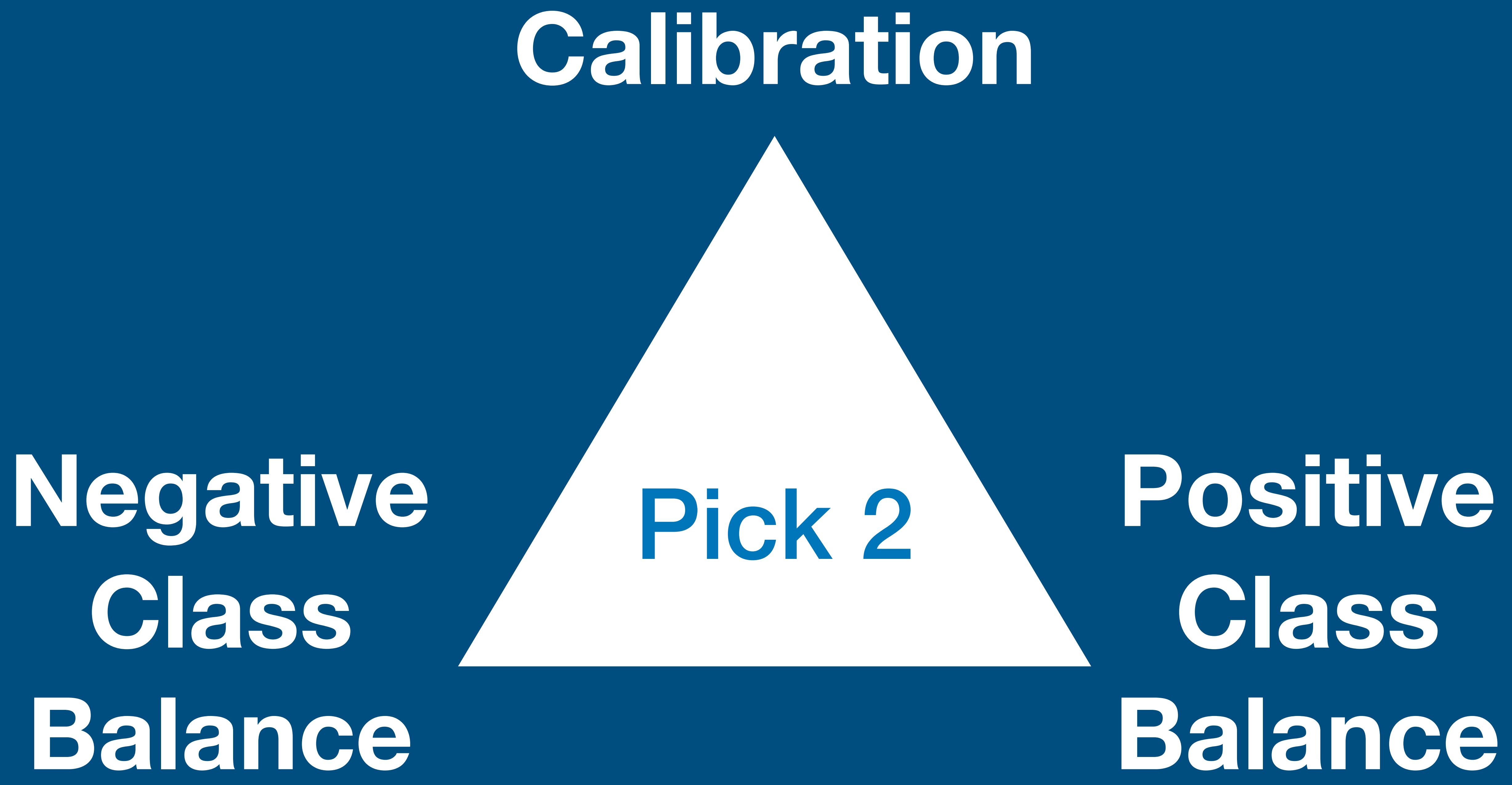
False Positive Rate?

F1 Score? Predictive Power?

Over 20 different measures of fairness are found in the ML fairness literature

Verma, Sahil, and Julia Rubin. "Fairness definitions explained." 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 2018.

Impossibility of Fairness



Some measures of fairness are mutually exclusive, have to pick between them

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

Challenges

1

Auditing the performance of hundreds or thousands of intersectional subgroups

2

Balancing dozens of incompatible definitions of fairness

Race	Accuracy
African-American	73
Asian	77
Caucasian	79
Hispanic	91
Native American	88
Other	67

Race, Sex	Accuracy
African-American, Male	60
Asian, Male	86
Caucasian, Male	96
Hispanic, Male	91
Native American, Male	75
Other, Male	81
African-American, Female	97
Asian, Female	66
Caucasian, Female	73
Hispanic, Female	91
Native American, Female	92
Other, Female	84

Race, Sex	Accuracy	FPR	FNR	F1	Precision	...
African-American, Male	87	74	61	68	95	86
Asian, Male	83	93	77	74	88	84
Caucasian, Male	80	82	93	71	72	88
Hispanic, Male	96	86	85	92	81	63
Native American, Male	89	85	76	85	93	97
Other, Male	78	69	90	76	68	62
African-American, Female	72	72	99	67	75	61
Asian, Female	84	68	65	91	71	71
Caucasian, Female	88	100	91	63	87	95
Hispanic, Female	76	94	99	71	77	64
Native American, Female	82	65	65	98	81	78
Other, Female	86	98	72	83	72	69


FairVis
Audit Classification for Intersectional Bias
Minimum Size: 0
RESET GROUPS
GENERATE SUBGROUPS
Age
C_charge_degree
Race
Sex
Priors_count
Days_b_screening_arrest
Accuracy
Precision
Recall
avg: 65.86%
avg: 66.05%
avg: 60.77%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Suggested Subgroups
Sort by:
Accuracy
1 - 2
Group 1
4 Instances
C_charge_degree
Misdemeanor
Felony
Race
Native American
African-American
Asian
Caucasian
Sex
Male
Female
Group 2
2 Instances
C_charge_degree
Felony
Misdemeanor
Race
Asian
African-American
Caucasian
Hispanic
Sex
Female
Male
Group Details
EXPORT
Hovered
Pinned
Ground Truth Label Balance
Feature
Pinned
Hovered
Size
2626
29
race
African-American
Asian
sex
Male
Male

FairVis

Auditing the COMPAS Model

Risk scoring for recidivism prediction

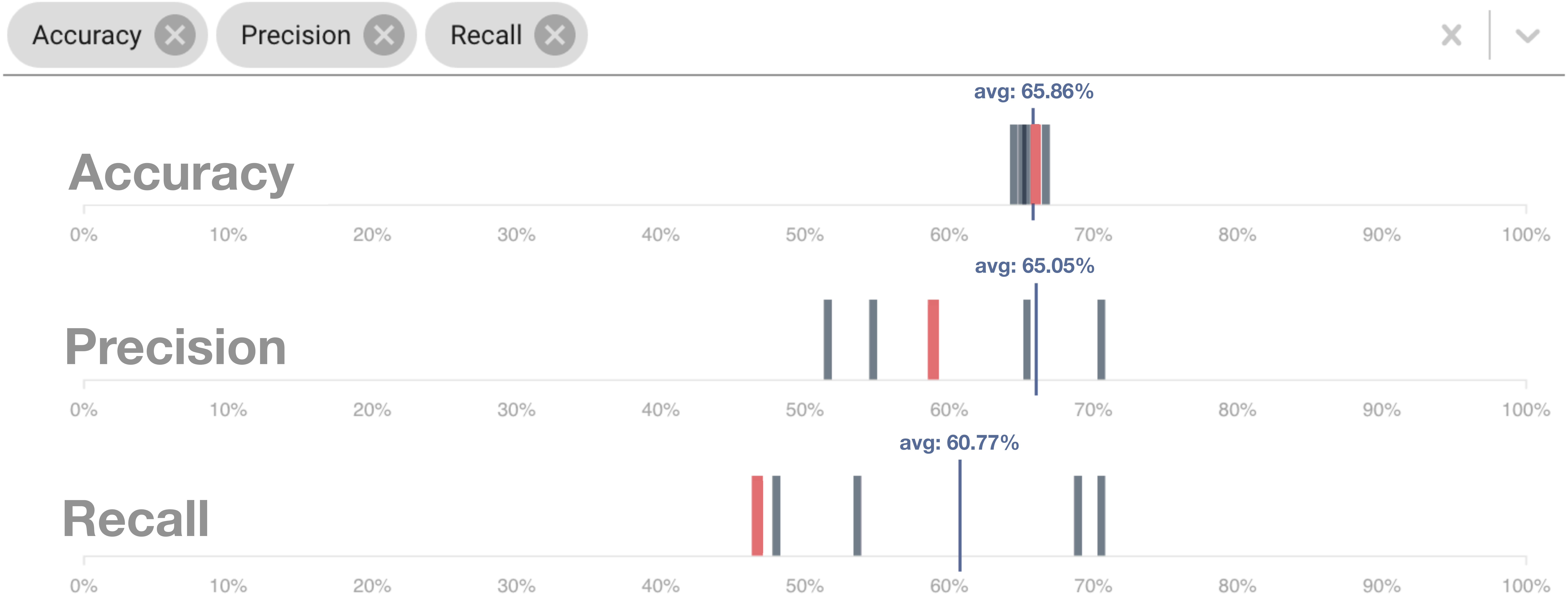
Use Case 1

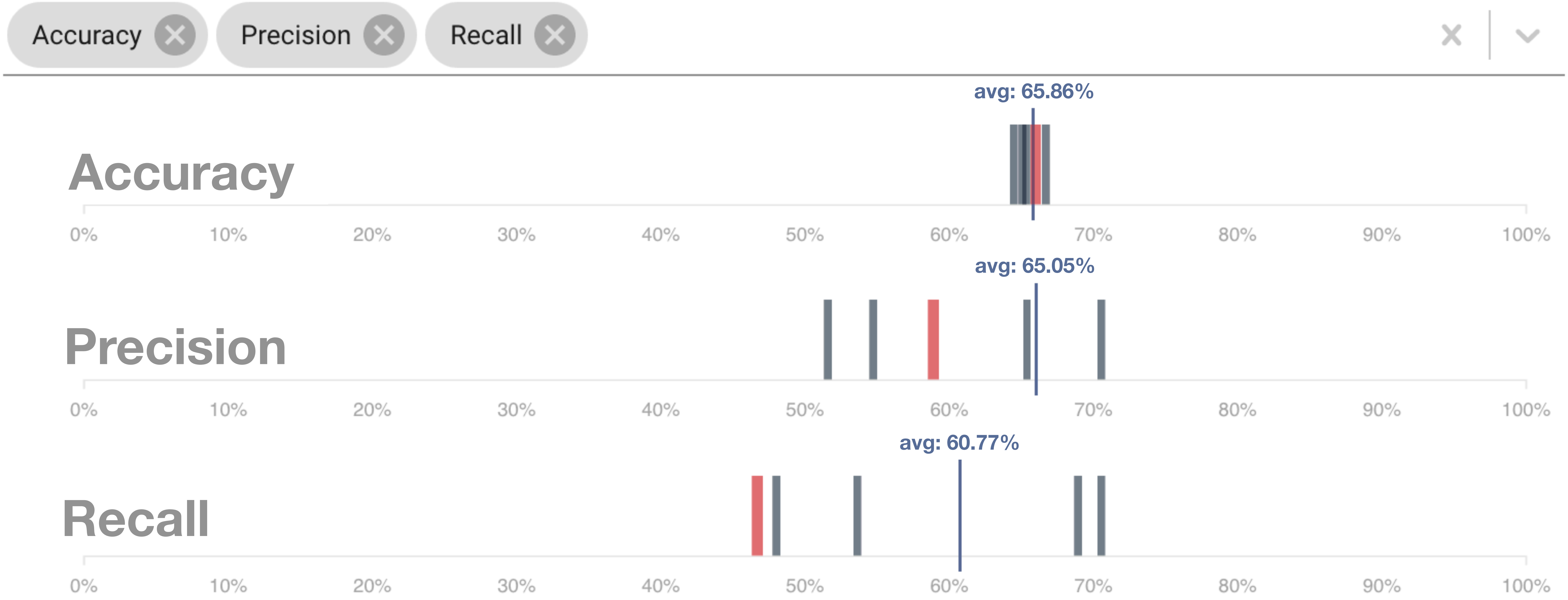
Auditing for Suspected Bias

Visualize specific subgroups

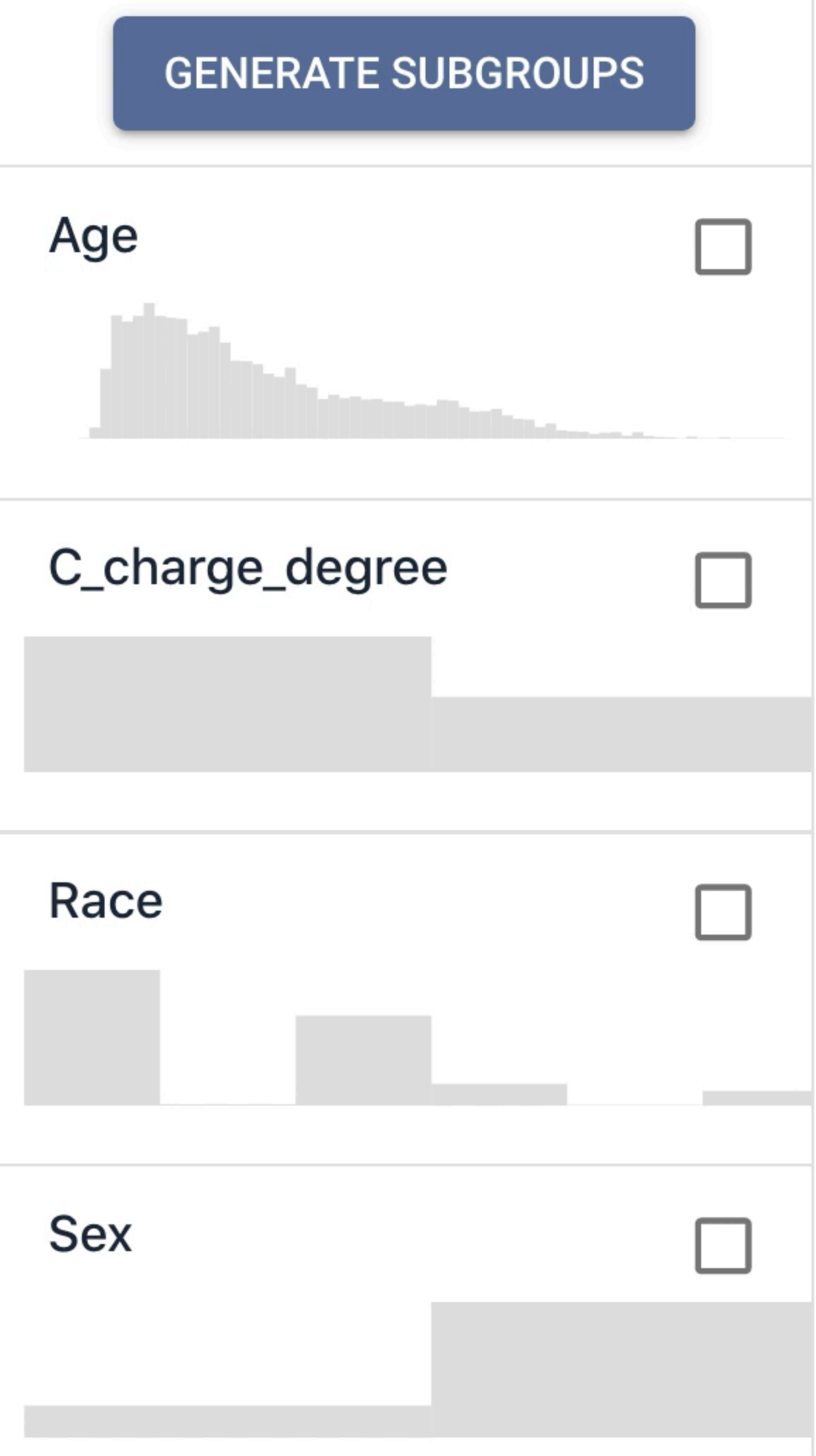
Performance of the

African-American Male subgroup


Accuracy
Precision
Recall
avg: 65.86%
Accuracy
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
avg: 65.05%
Precision
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
avg: 60.77%
Recall
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%



= Subgroup of African-American Males



Visualize all the combinations of subgroups for selected features

African-American Male, Caucasian Male, African-American Female, etc.


FairVis
Audit Classification for Intersectional Bias
Minimum Size: 0
RESET GROUPS
GENERATE SUBGROUPS
Age
C_charge_degree
Race
Sex
Priors_count
Accuracy
Precision
Recall
avg: 65.86%
avg: 66.05%
avg: 60.77%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Group Details
EXPORT
Ground Truth Label Balance
Feature
Pinned
Hovered
Size
Suggested Subgroups
Sort by:
Accuracy
1 - 2
Group 1
4 Instances
Group 2
2 Instances
Filter for significantly large subgroups


FairVis
Audit Classification for Intersectional Bias
Minimum Size: 368
RESET GROUPS
GENERATE SUBGROUPS
Age
C_charge_degree
Race
Sex
Priors_count
Accuracy
Precision
Recall
avg: 65.86%
avg: 66.05%
avg: 60.77%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Group Details
EXPORT
0% 20% 40% 60% 80% 100%
0%
Ground Truth Label Balance
Feature
Pinned
Hovered
Size
Select preferred metrics,
in this case the
false positive rate


FairVis
Audit Classification for Intersectional Bias
Minimum Size: 368
RESET GROUPS
GENERATE SUBGROUPS
Age
C_charge_degree
Race
Sex
Priors_count
Accuracy
False Positive Rate
avg: 65.86%
avg: 29.35%
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Group Details
EXPORT
0% 20% 40% 60% 80% 100%
Ground Truth Label Balance
Feature
Pinned
Hovered
Size
Compare the subgroups
with the highest and lowest
false positive rate

Use Case 2

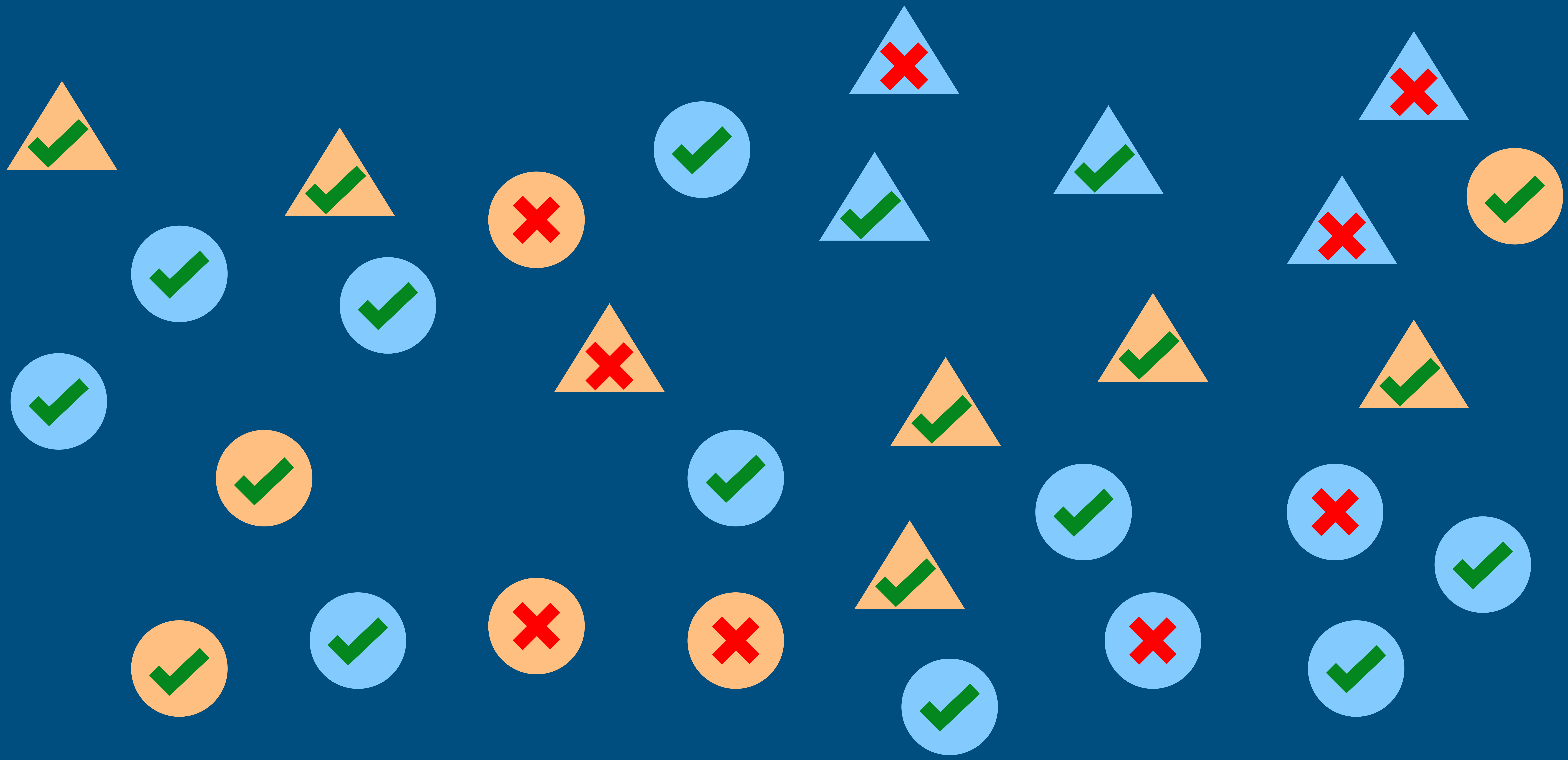
Discovering Unknown Biases

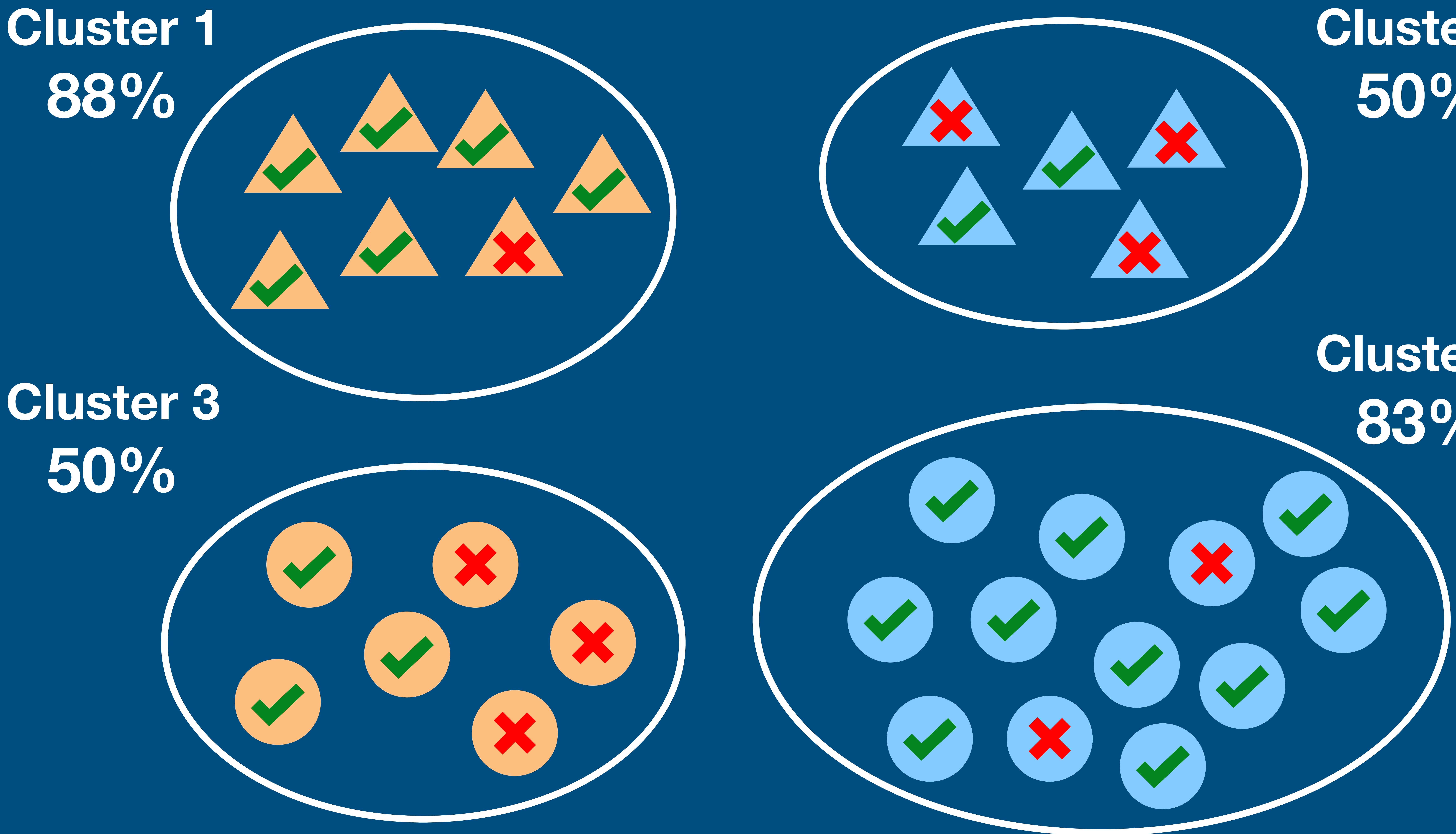
A

Suggested Subgroups

Shape Classification

70% Accuracy

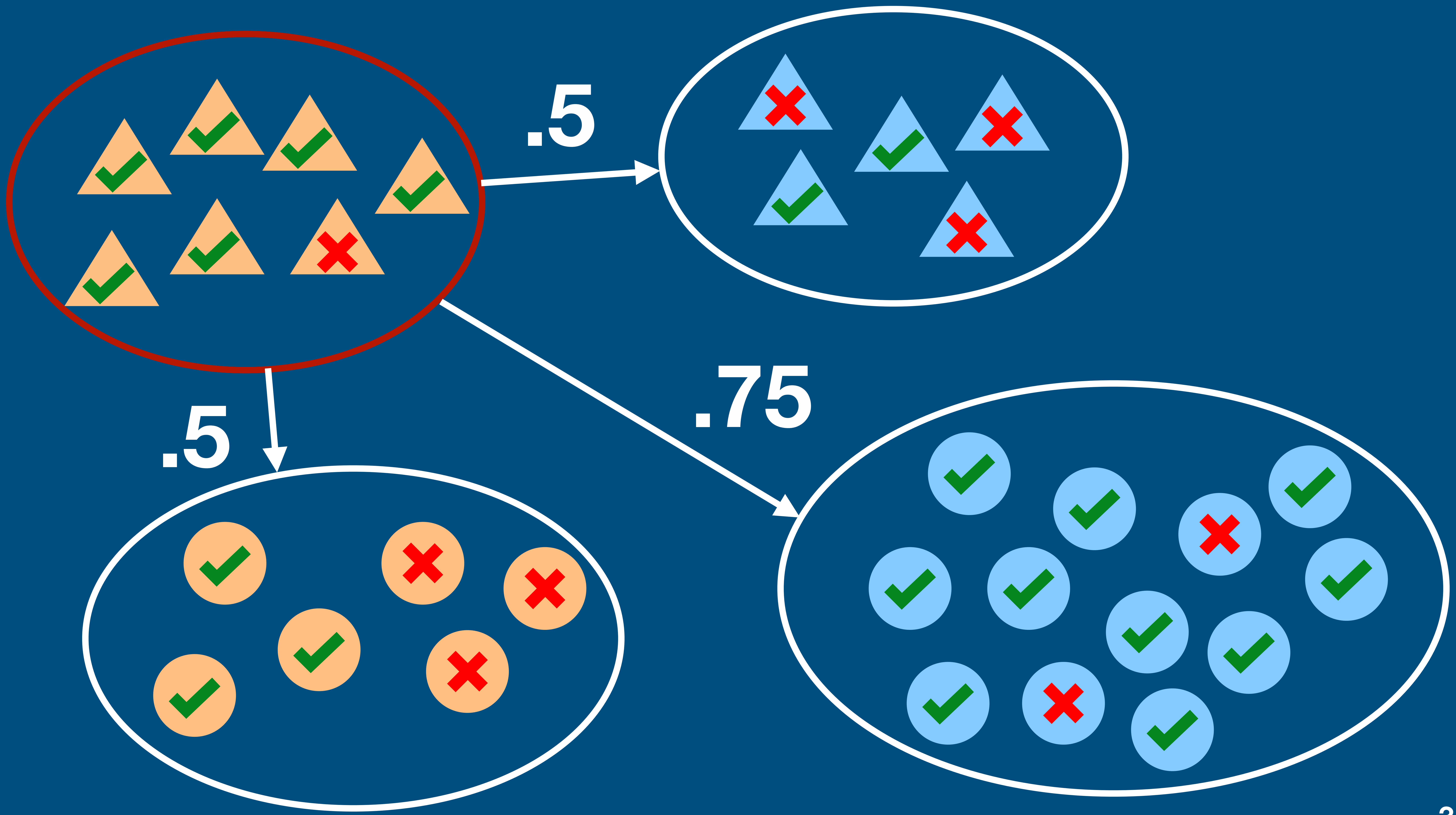



Cluster 1
88%
Cluster 2
50%
Cluster 3
50%
Cluster 4
83%


Accuracy
False Positive Rate
avg: 65.86%
Accuracy
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
avg: 29.35%
False Positive Rate
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Group Details
EXPORT
0% 20% 40% 60% 80% 100%
0%
Ground Truth Label Balance
Feature
Pinned
Hovered
Size
Suggested Subgroups
Sort by:
False Positive Rate
5 - 6
Group 5
425 Instances
C_charge_degree
Misdemeanor
Misdemeanor
Felony
0% 50% 100%
Race
African-American
African-American
Asian
Caucasian
Hispanic
0% 50% 100%
Sex
Male
Male
Female
0% 50% 100%
Group 6
249 Instances
C_charge_degree
Felony
Felony
Misdemeanor
0% 50% 100%
Race
African-American
African-American
Asian
Caucasian
Hispanic
0% 50% 100%
Sex
Female
Female
Male
0% 50% 100%

B

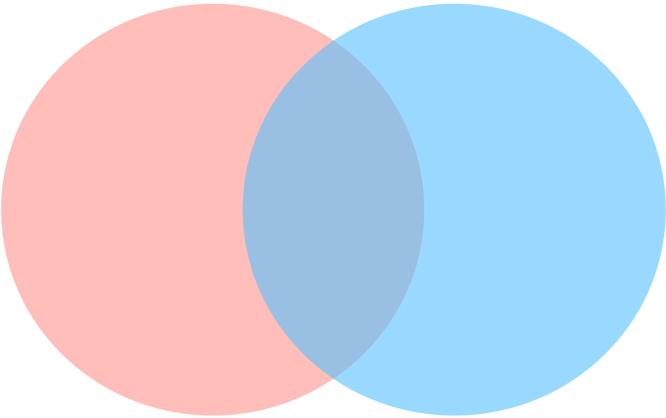
Similar Subgroups


.5
.75
.5

Compare the African-American Male subgroup to a similar subgroup of Other Male

By tackling

Intersectional Bias



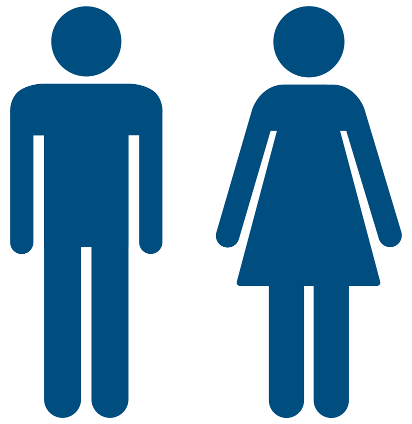
Multiple Definitions of Fairness

FairVis

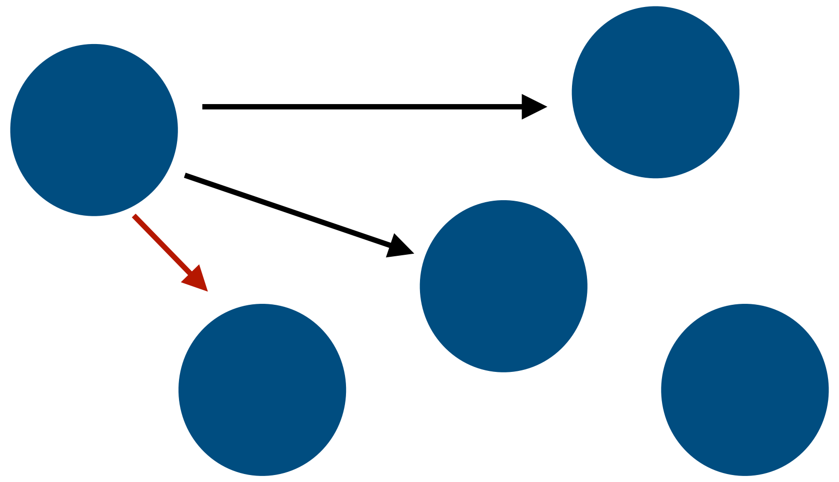
Enables users to find biases in their models

Allowing users to

Audit for Known Biases



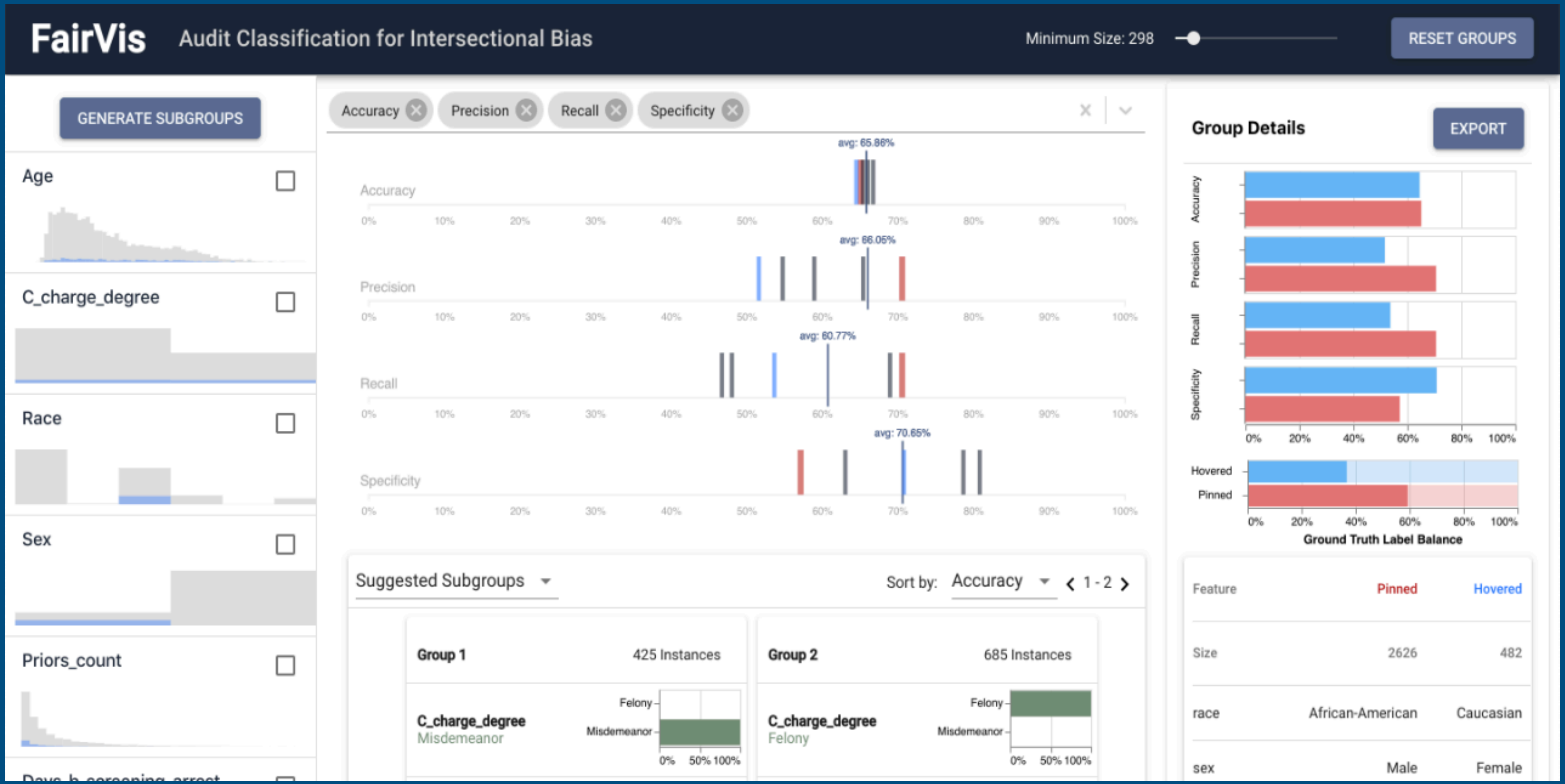
Explore Suggested & Similar Subgroups



FAIRVIS

Learn more at **bit.ly/fairvis**

Visual Analytics for Discovering Intersectional Bias in Machine Learning



Alex Cabrera
Carnegie Mellon

Will Epperson
Georgia Tech

Fred Hohman
Georgia Tech

Minsuk Kahng
Oregon State

Jamie Morgenstern
University of Washington

Polo Chau
Georgia Tech