Science of Security: Authentication and Predictive Logical Models

Janne Lindqvist
Human-Computer Interaction and Security Engineering Lab

October 4, 2019
http://lindqvistlab.org

How Does Science Get Published?

How Does Science Get Published?

- Peer review

Big Picture

PLOS | MEDICINE

BROWSE PUBLISH

OPEN ACCESS

ESSAY

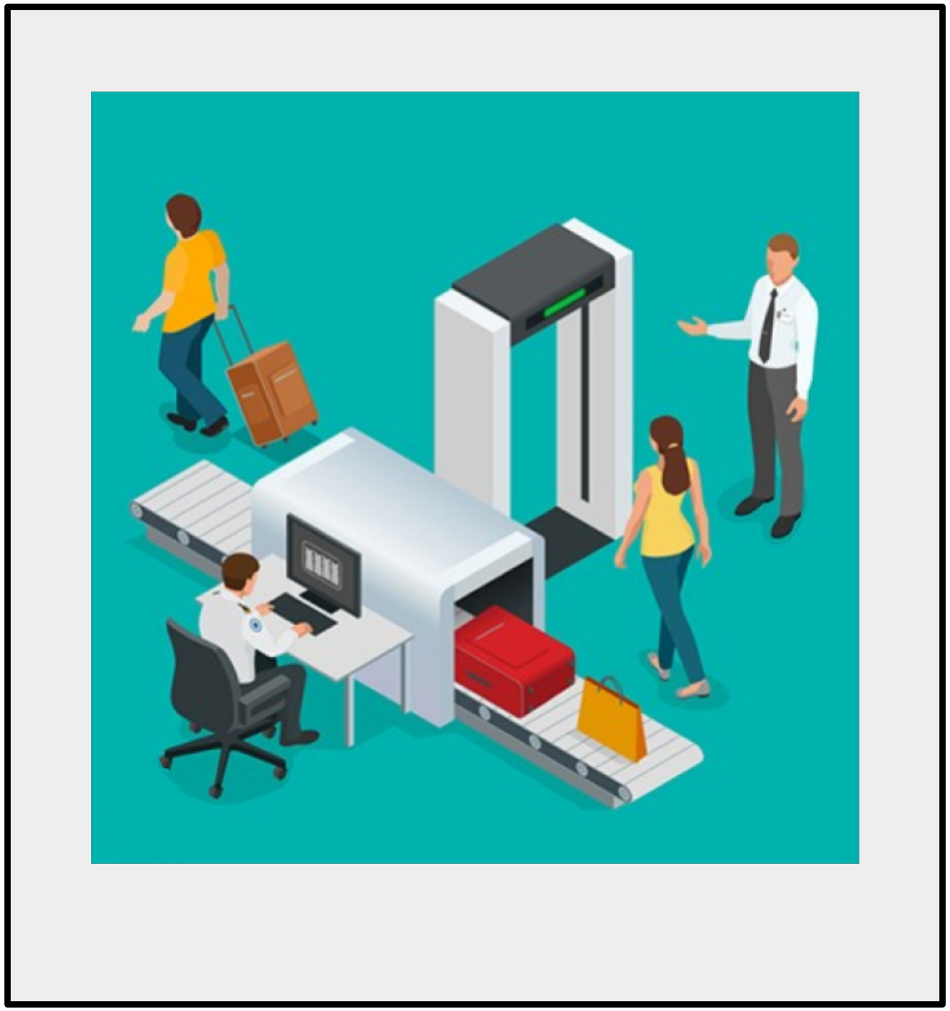
Why Most Published Research Findings Are False

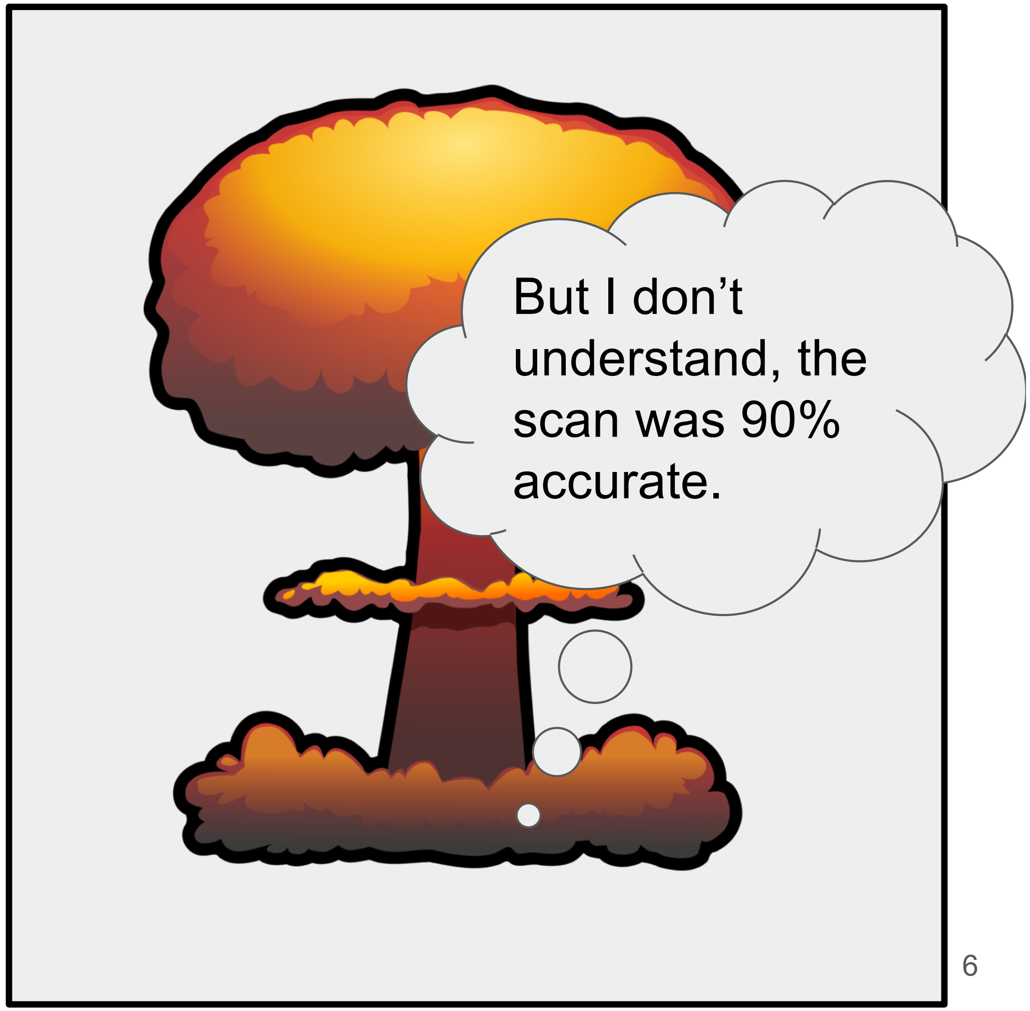
John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

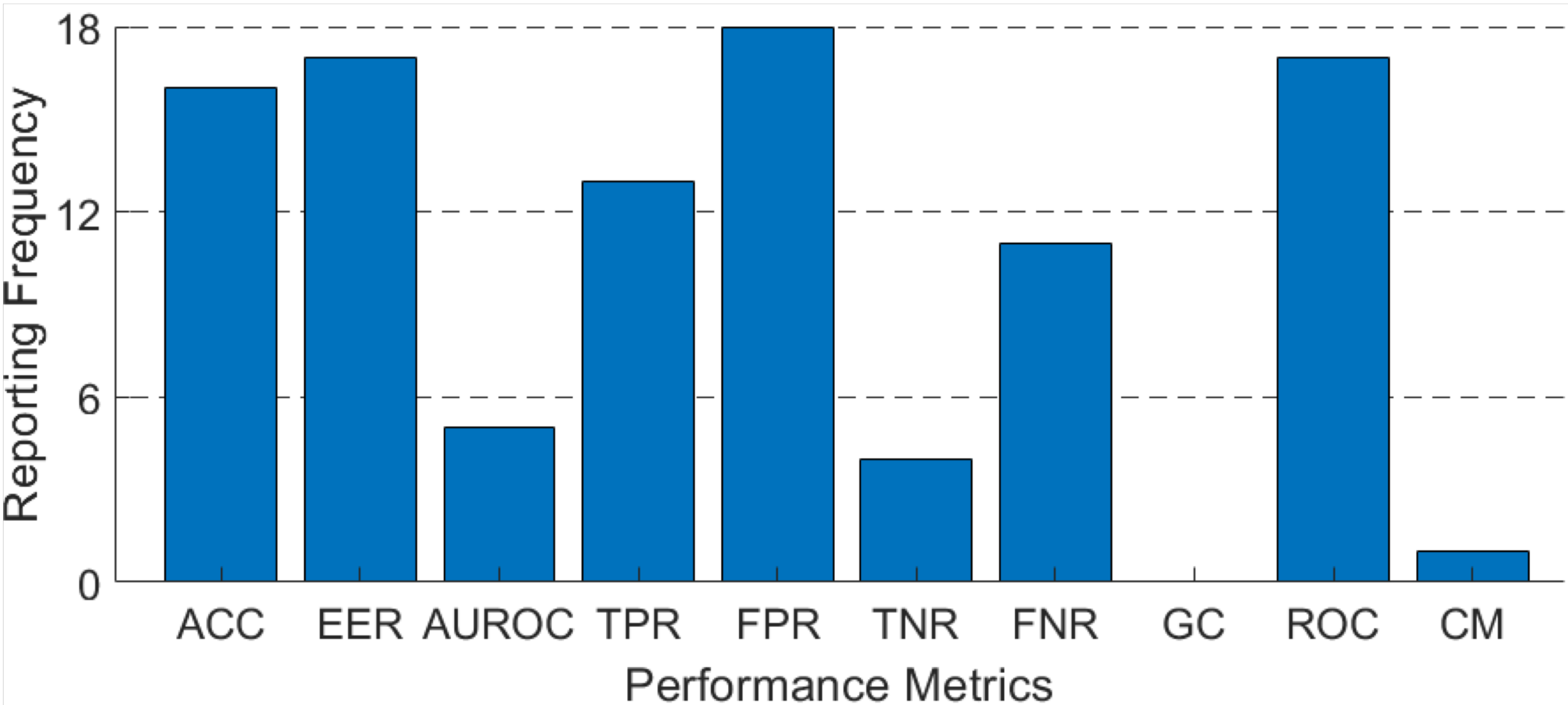
TL; DR:
Performance reporting might surprise you

TL; DR:
Performance reporting might surprise you

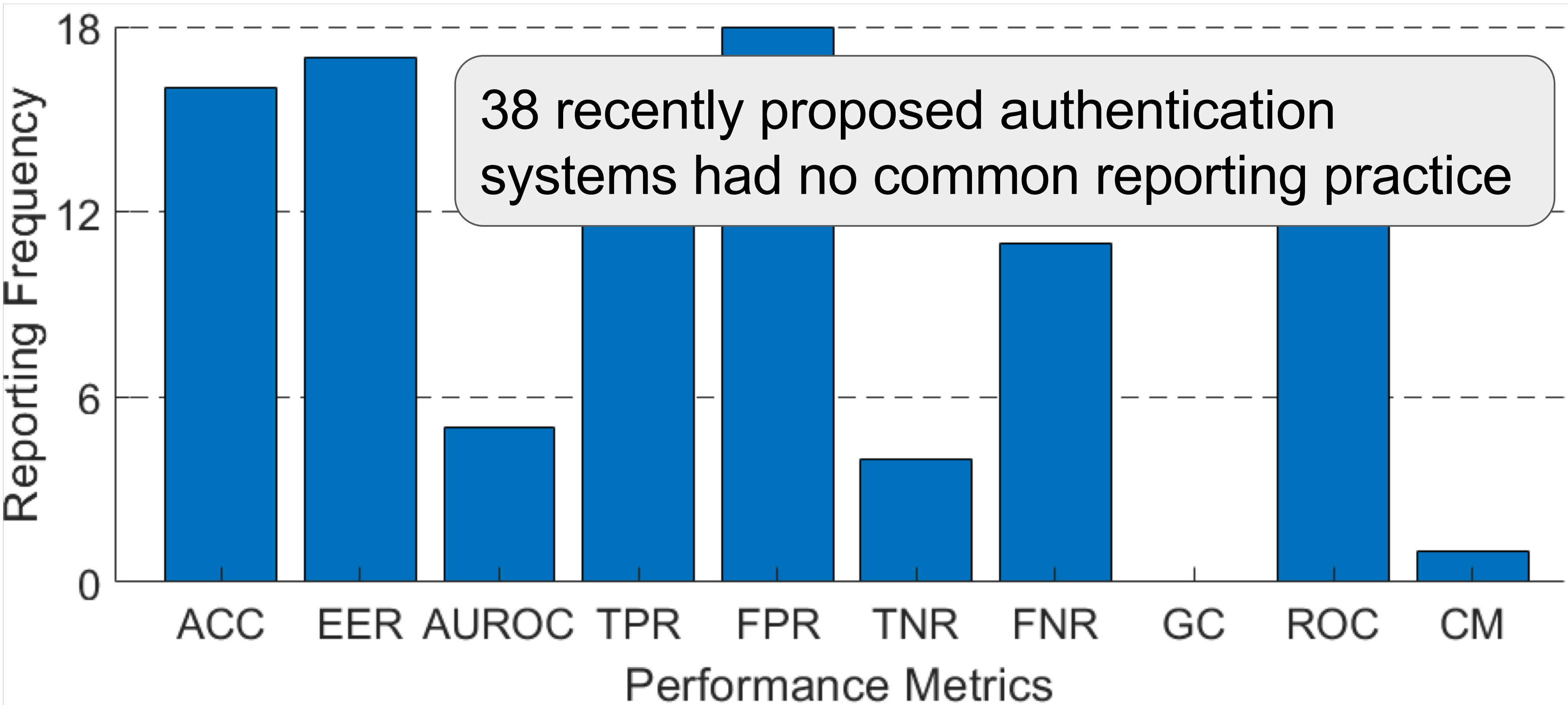




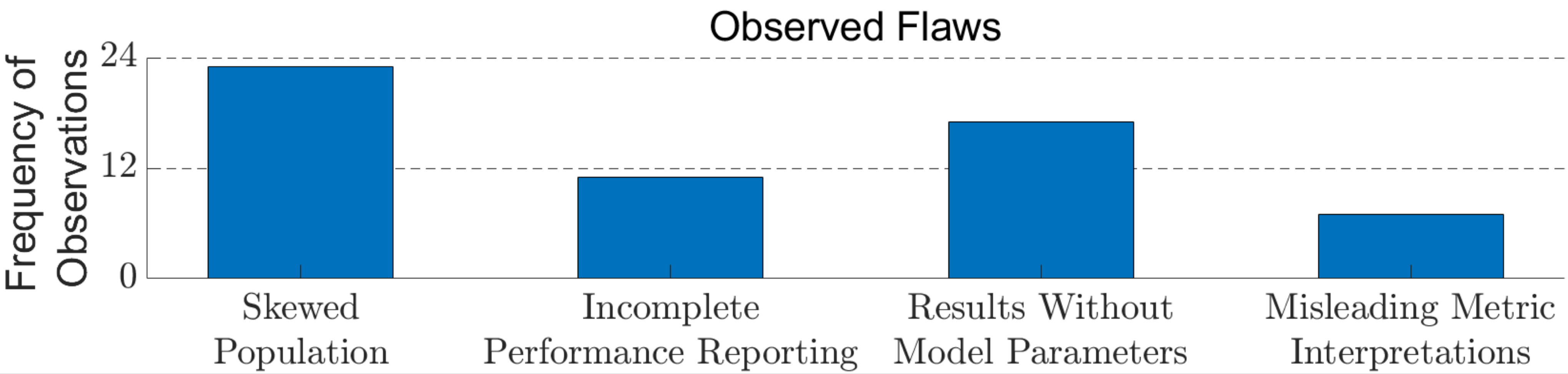
Example: Why Evaluation is Hard?



Example: Why Evaluation is Hard?

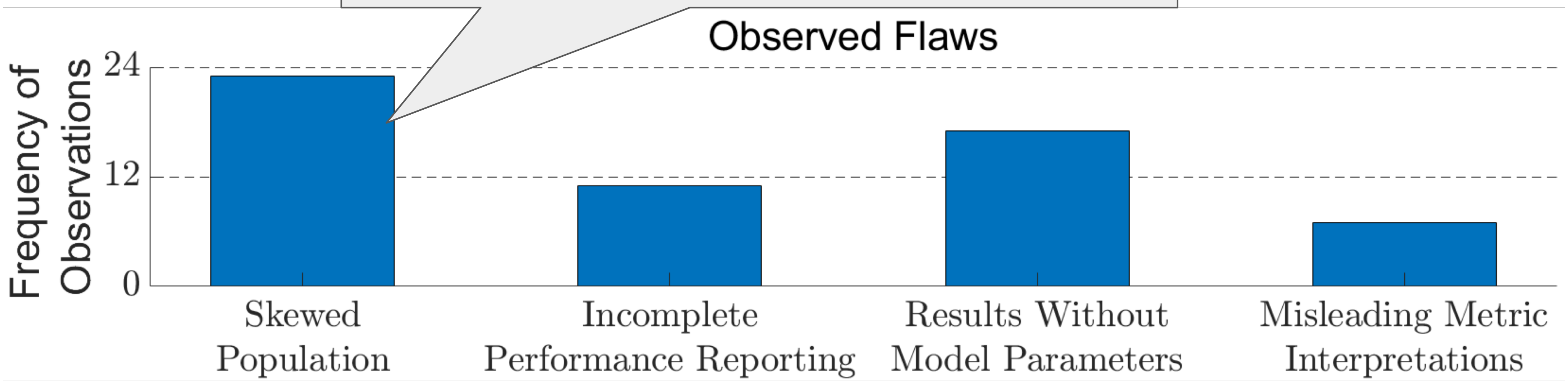


Most (36 of 38) Reporting Had Some Flaw

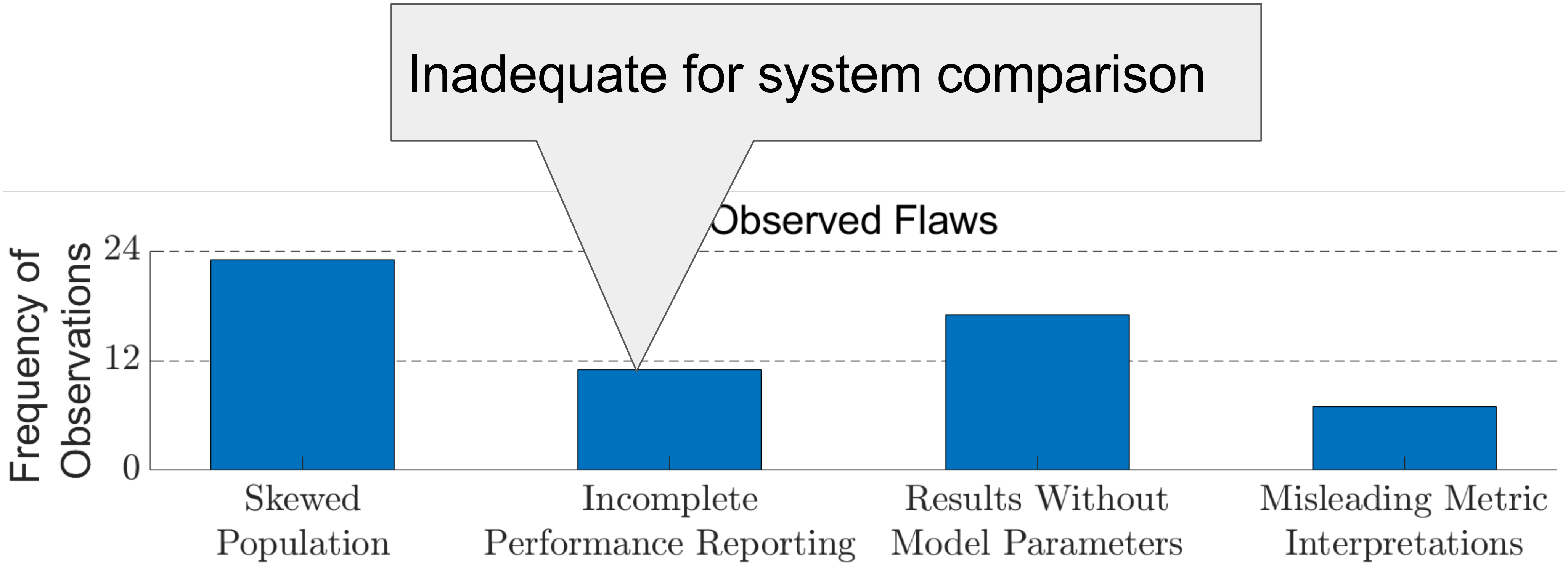


Most (36 of 38) Reporting Had Some Flaw

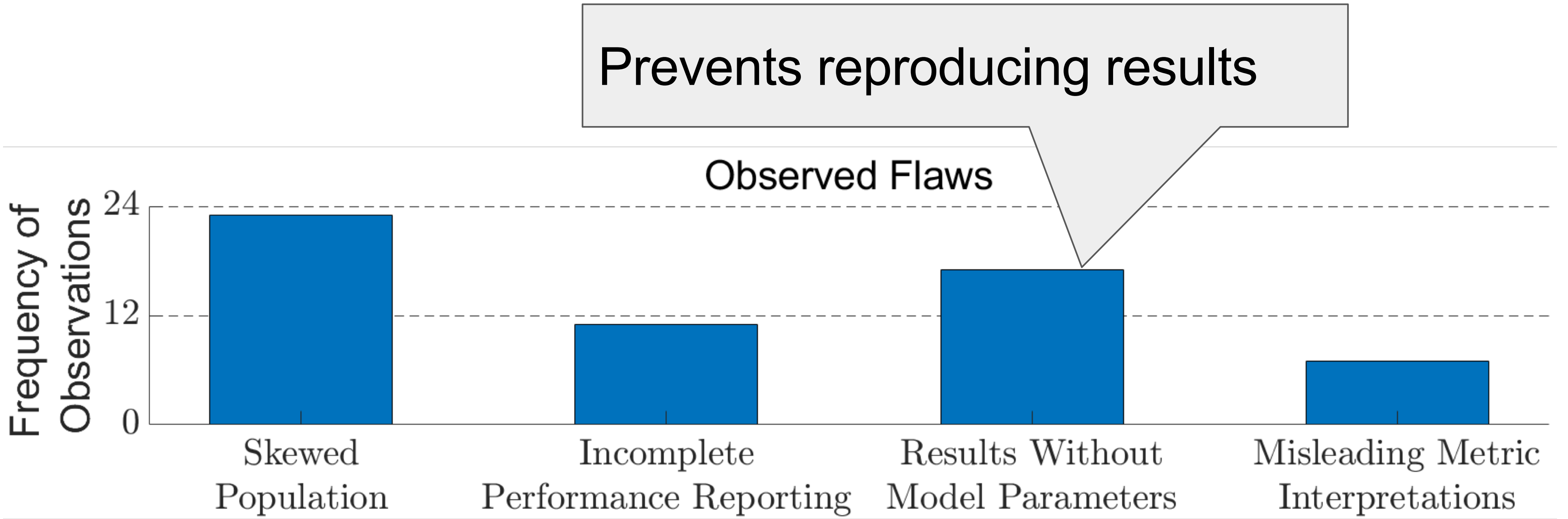
Inflate performance metric values



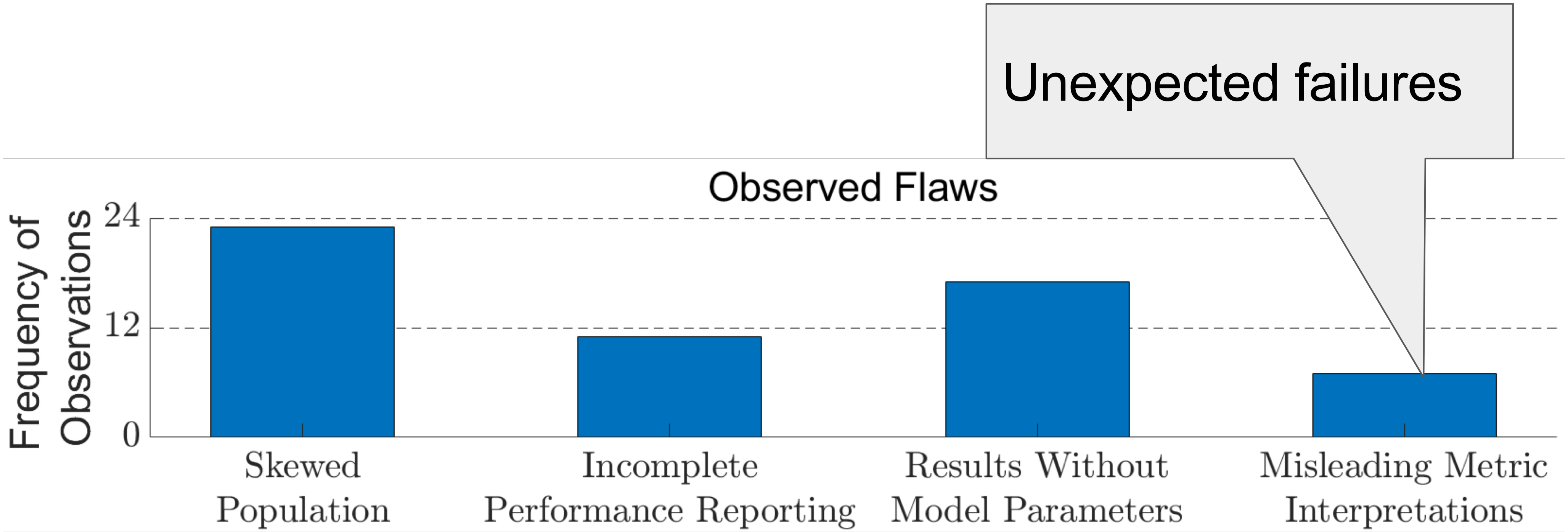
Most (36 of 38) Reporting Had Some Flaw



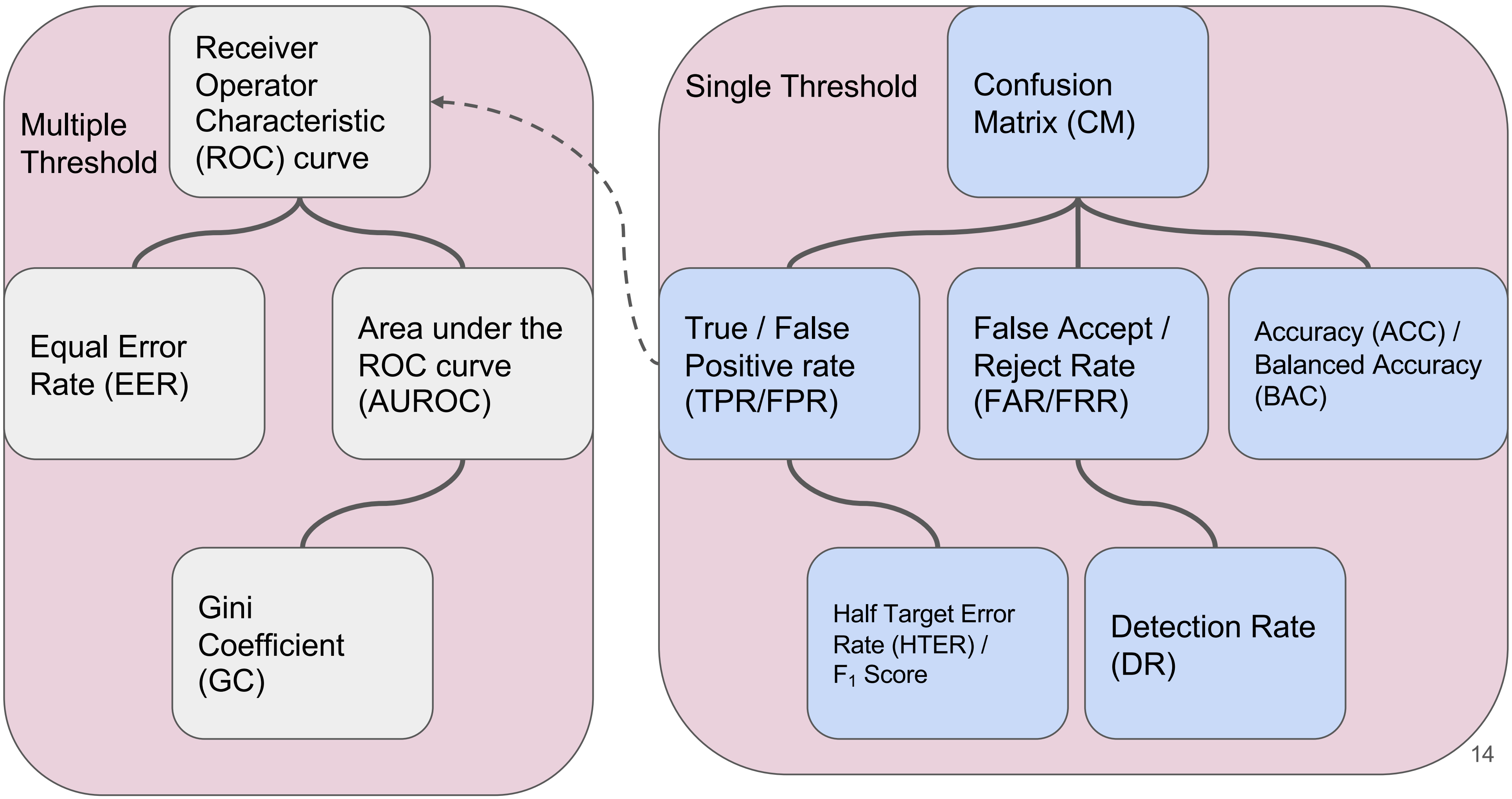
Most (36 of 38) Reporting Had Some Flaw



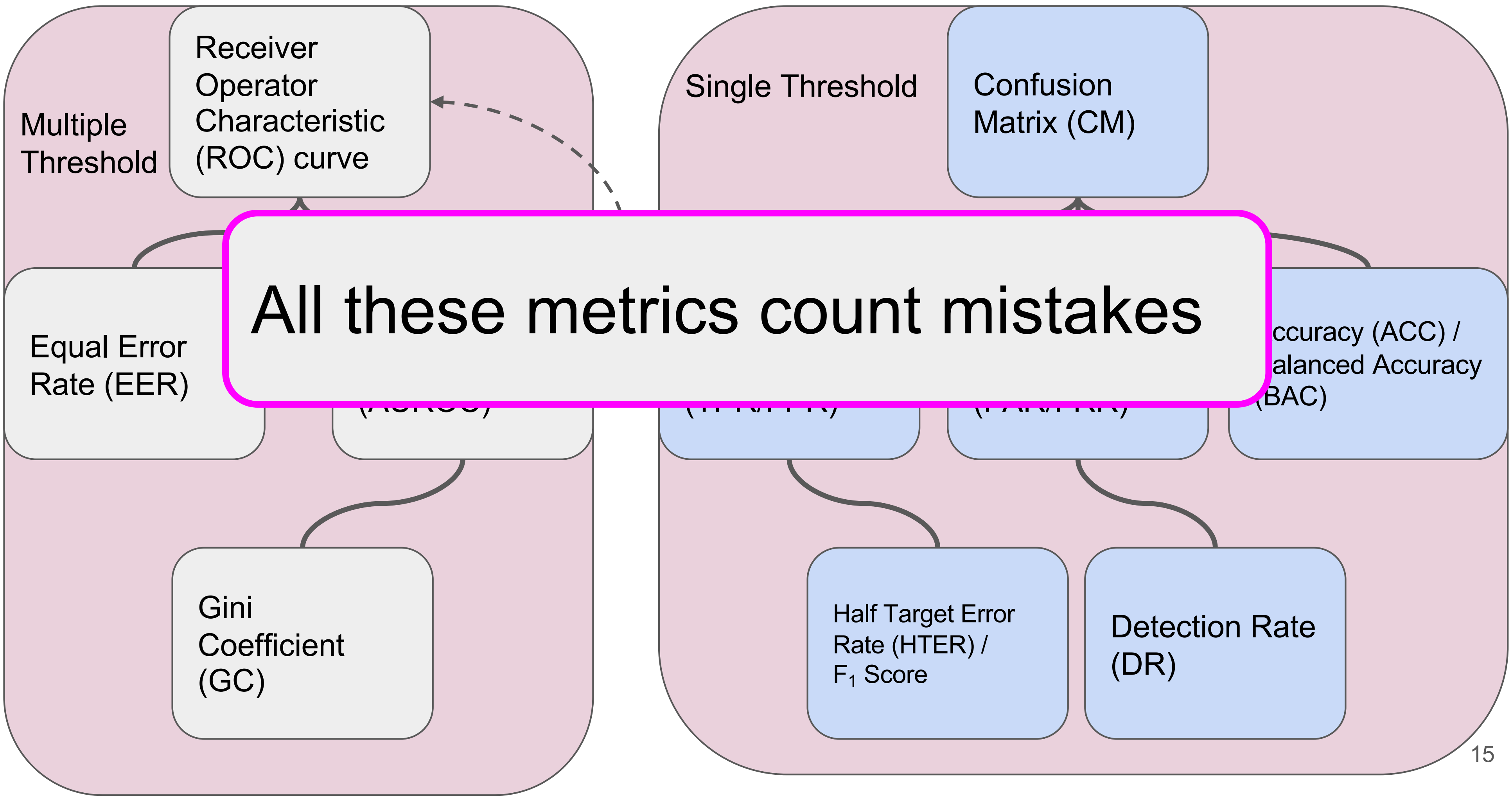
Most (36 of 38) Reporting Had Some Flaw



Metrics Are Related

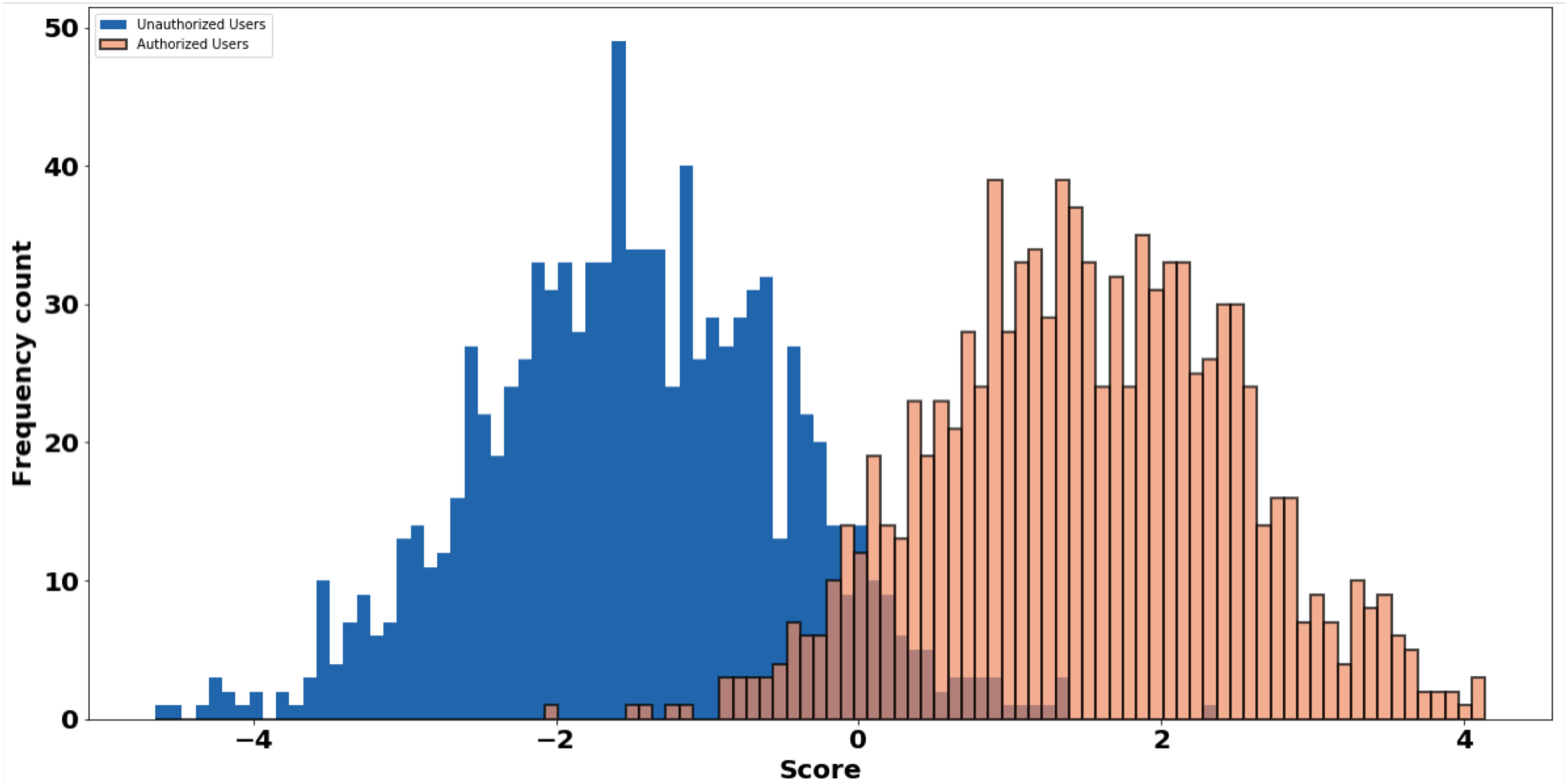


Metrics Are Related



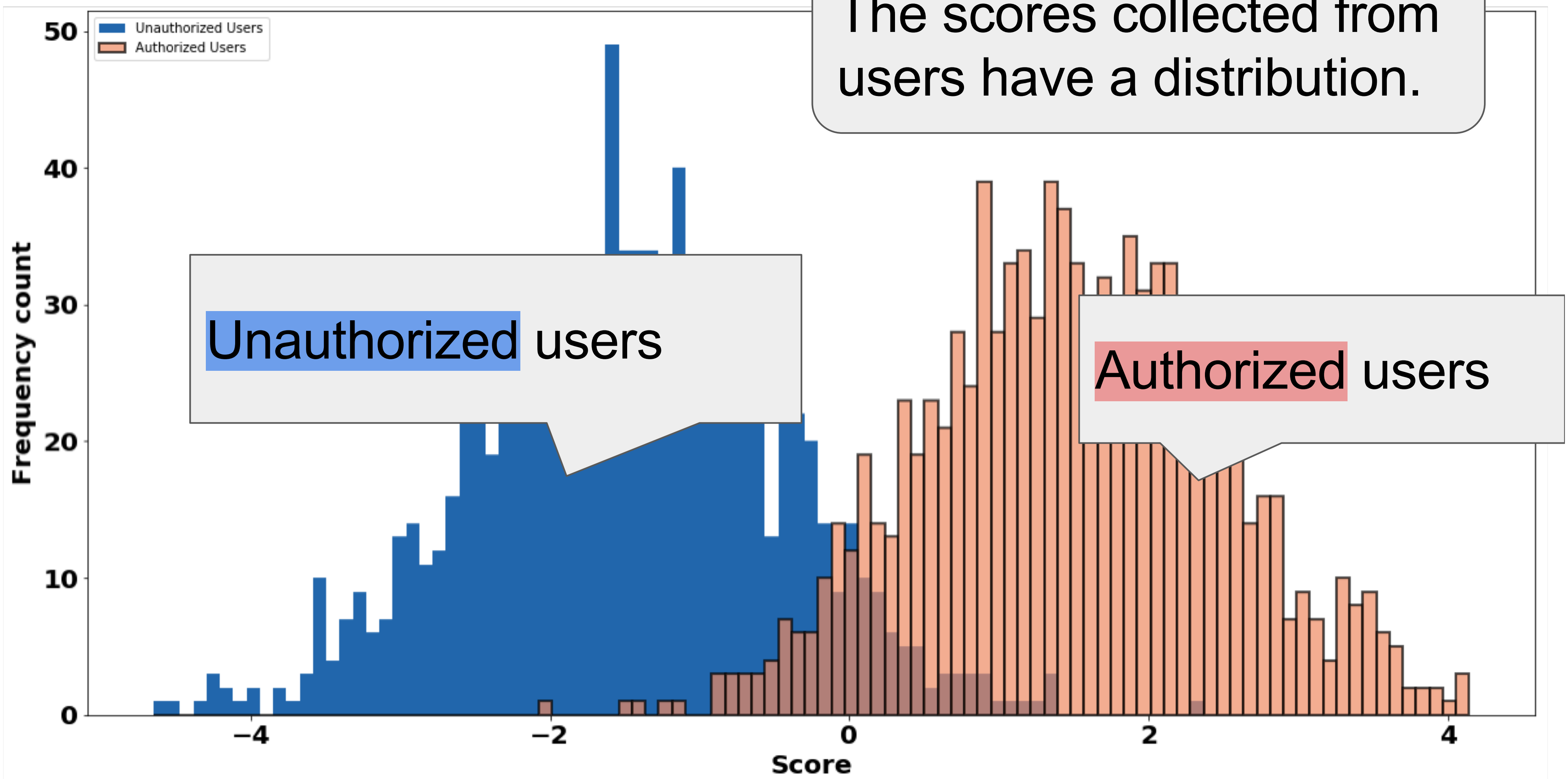
All models are wrong, but some are useful

How Authentication Systems Work

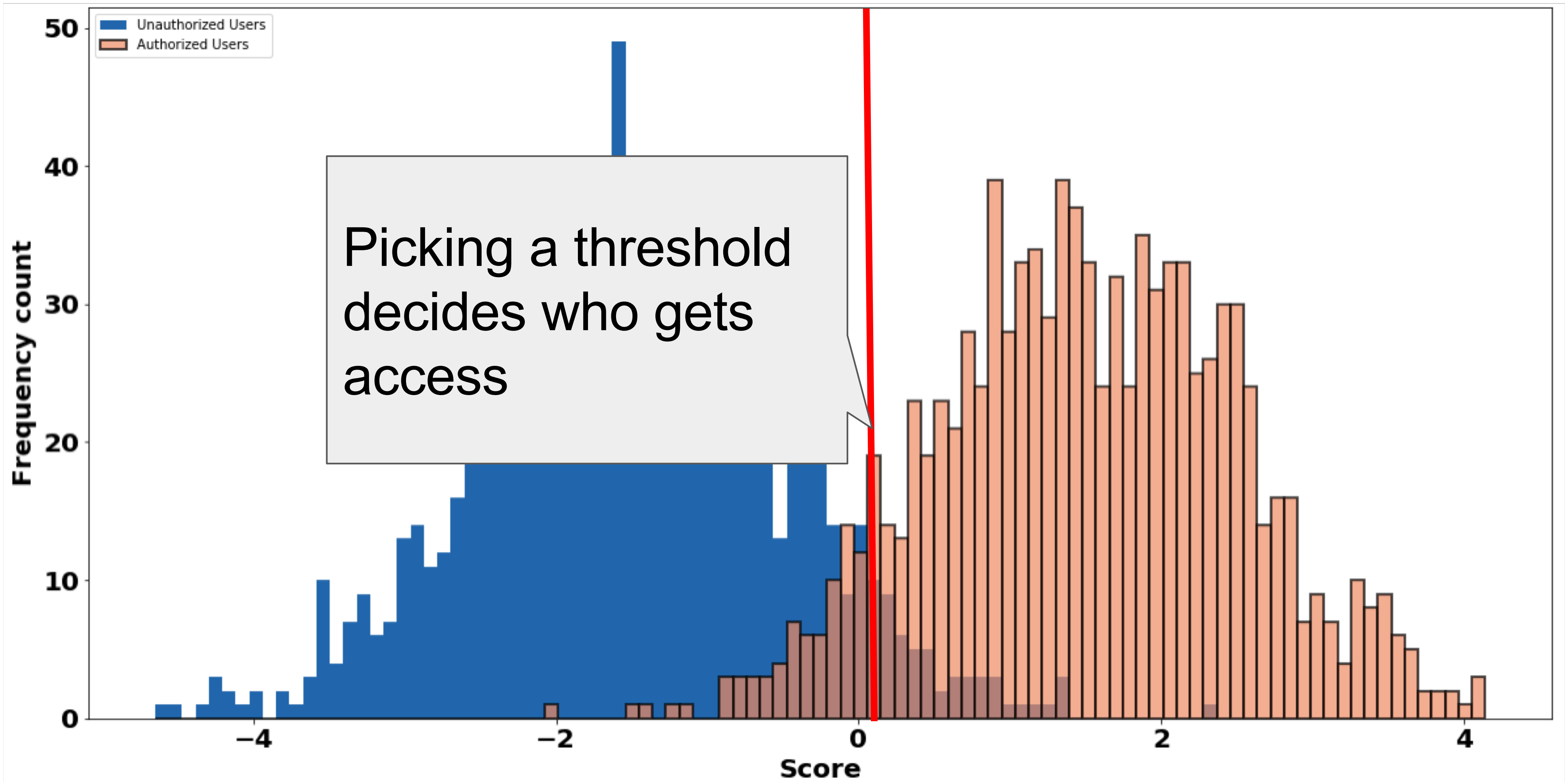


How Authentication Systems Work

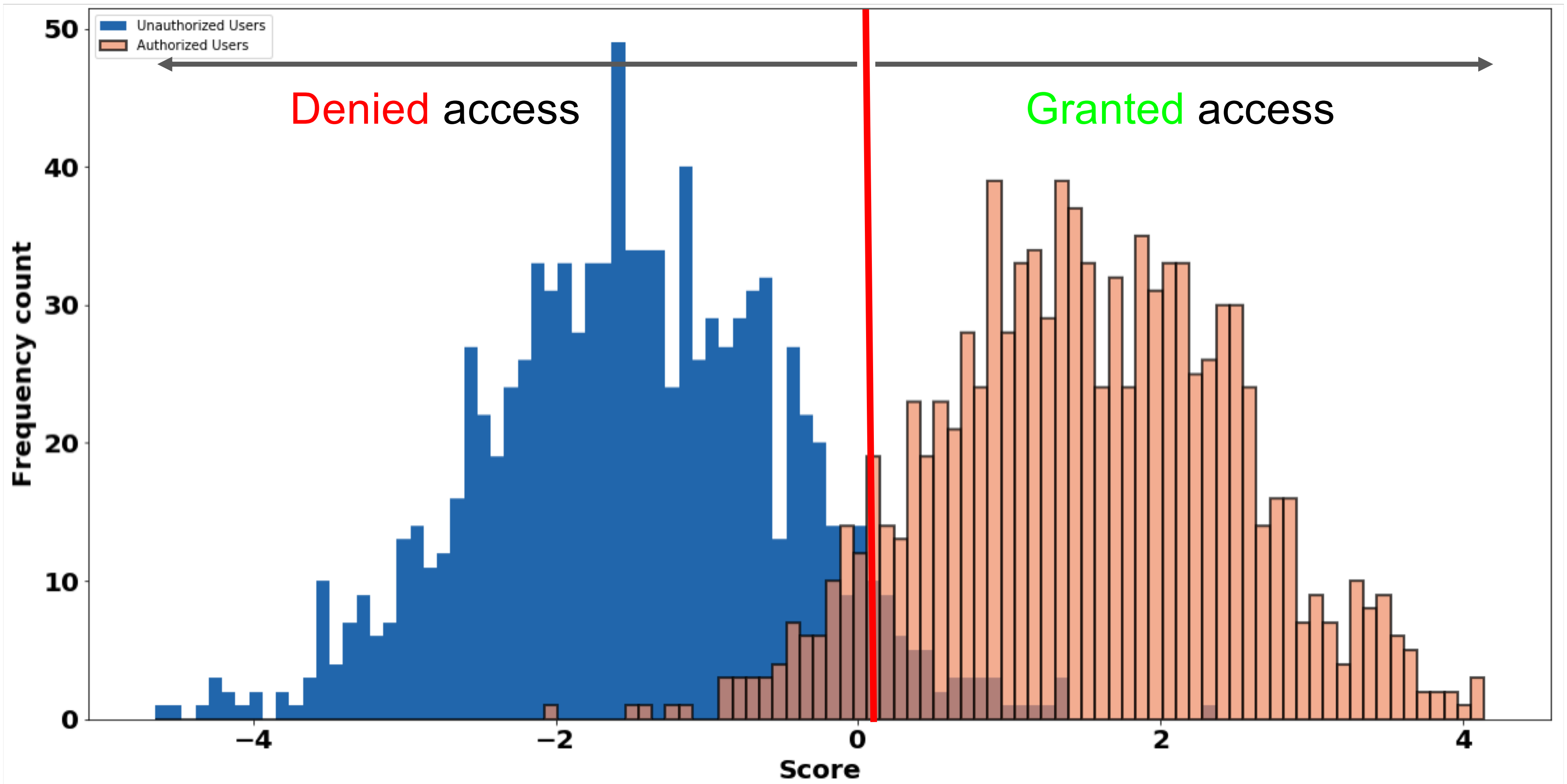
The scores collected from users have a distribution.



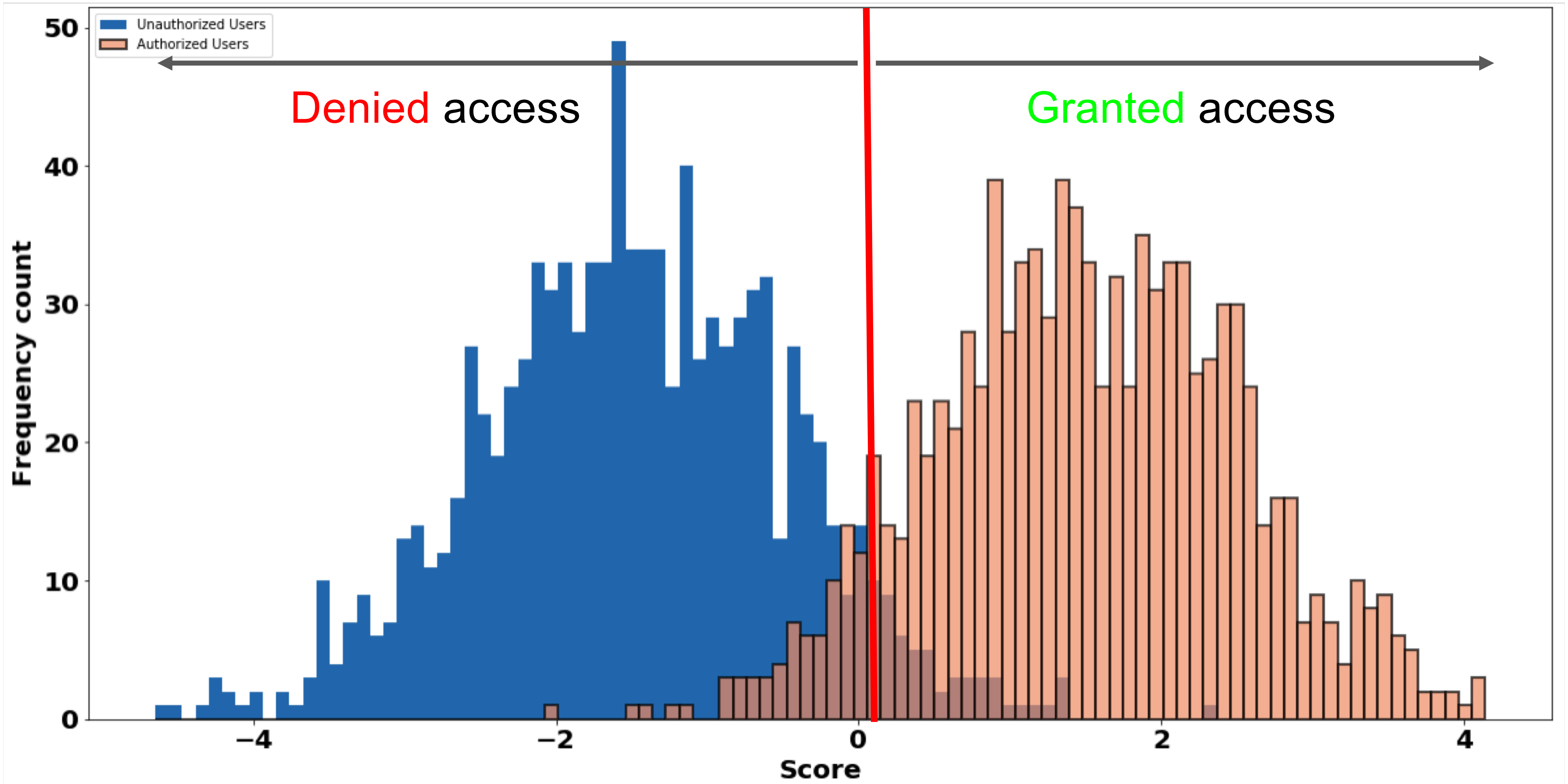
How Authentication Systems Work



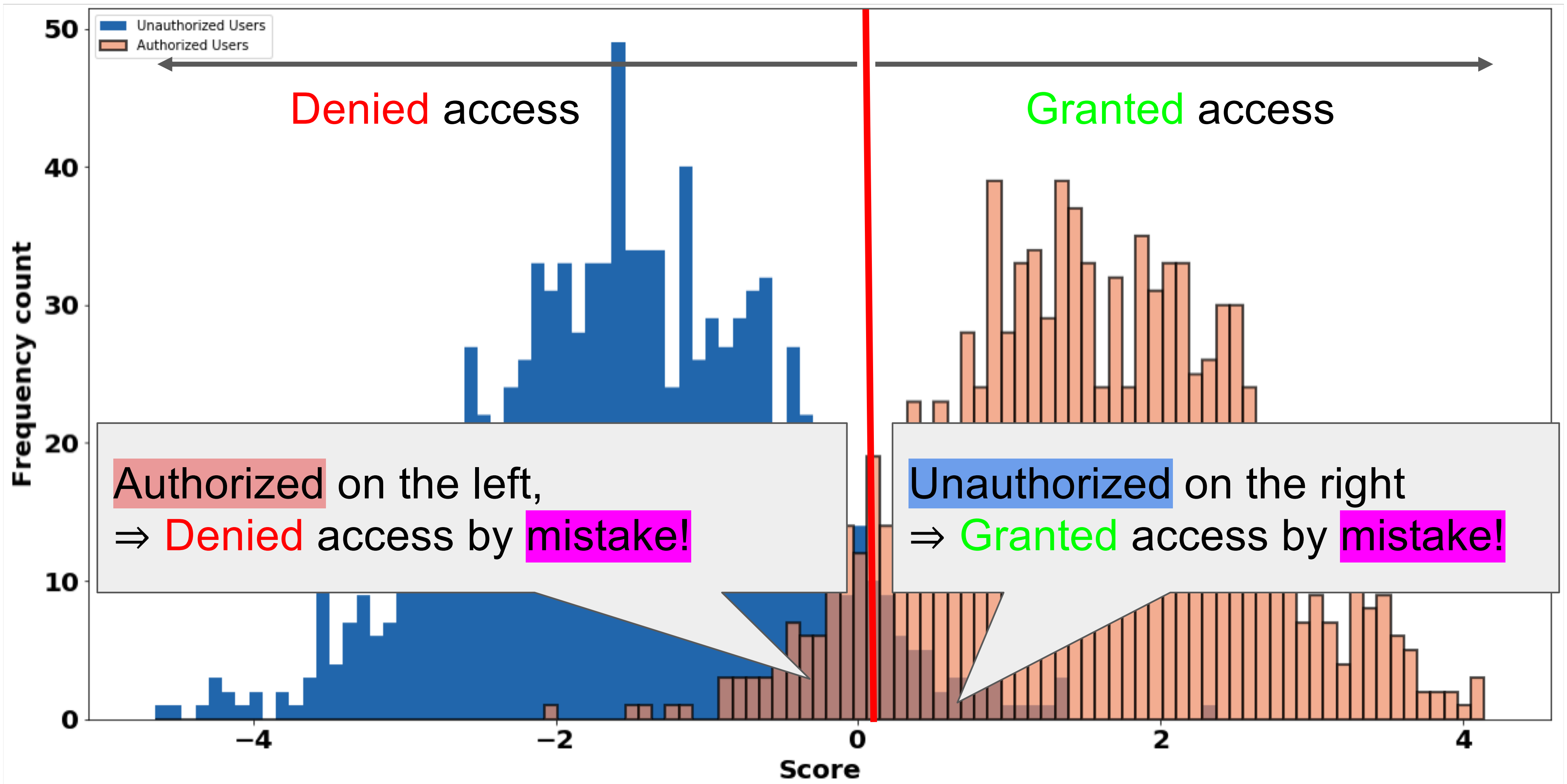
How Authentication Systems Work



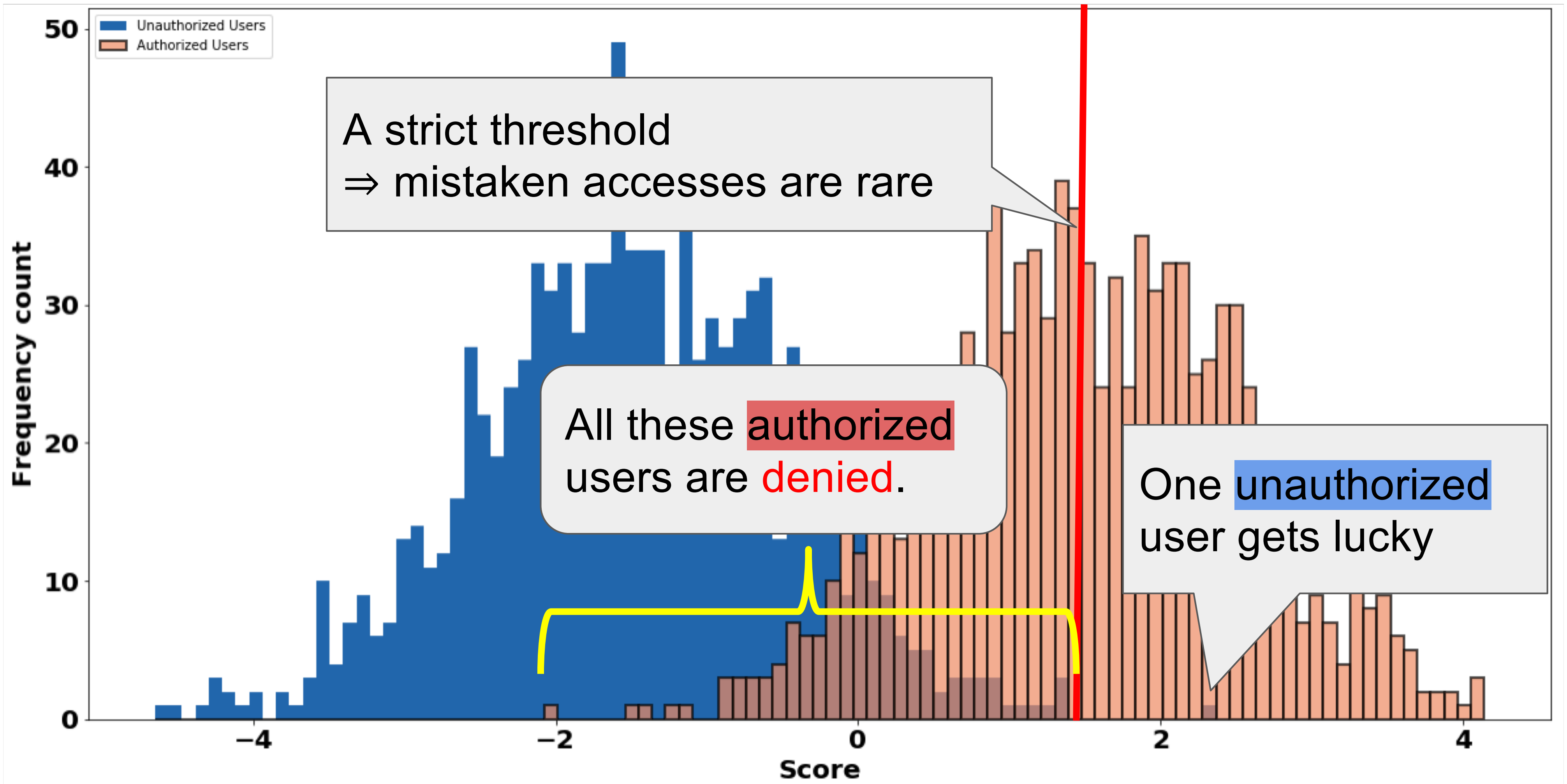
Where are the Mistakes?



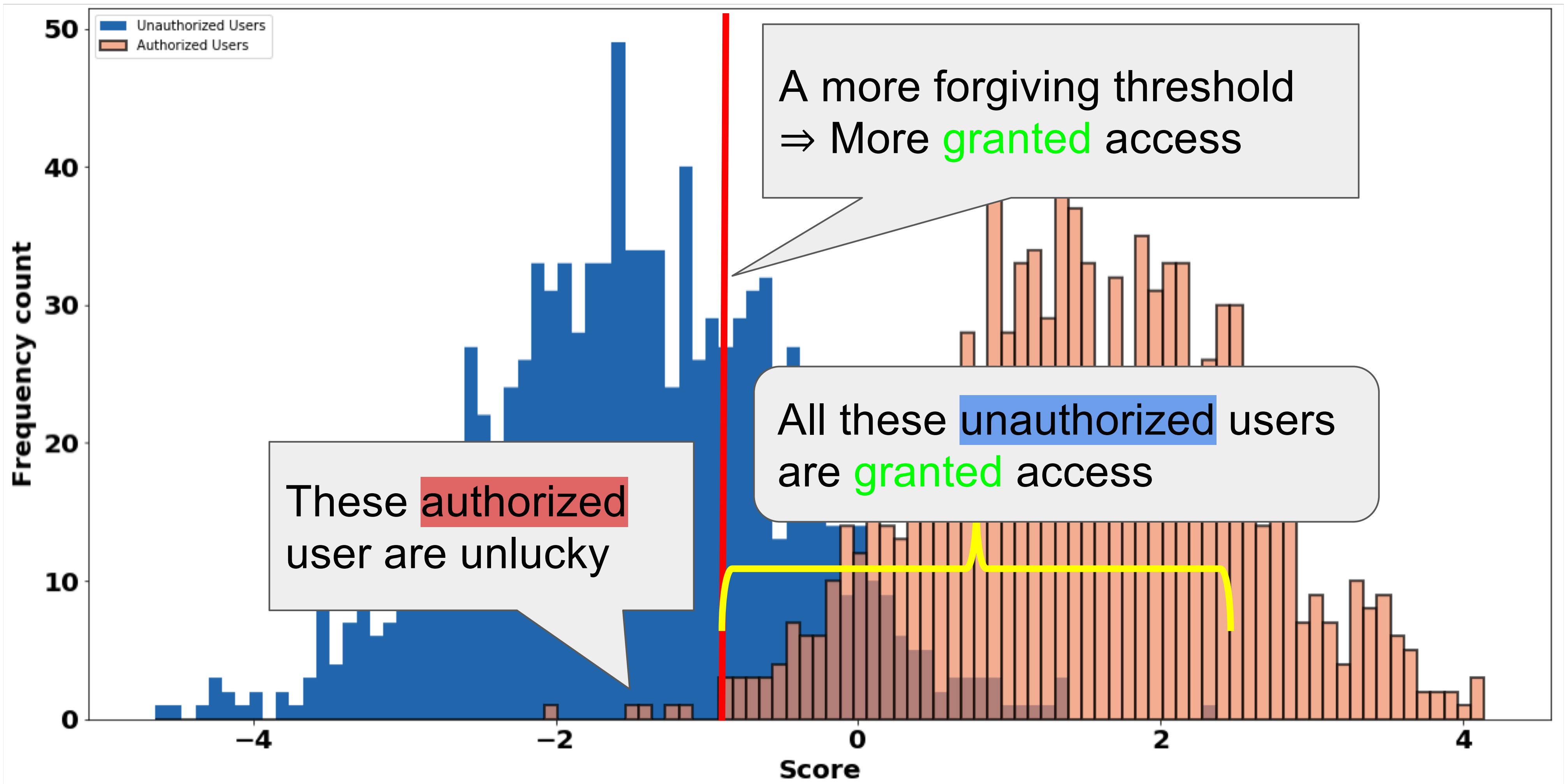
Mistakes


Unauthorized Users
Authorized Users
Denied access
Granted access
Authorized on the left,
⇒ Denied access by mistake!
Unauthorized on the right
⇒ Granted access by mistake!
50
40
30
20
10
0
Frequency count
−4
−2
0
2
4
Score

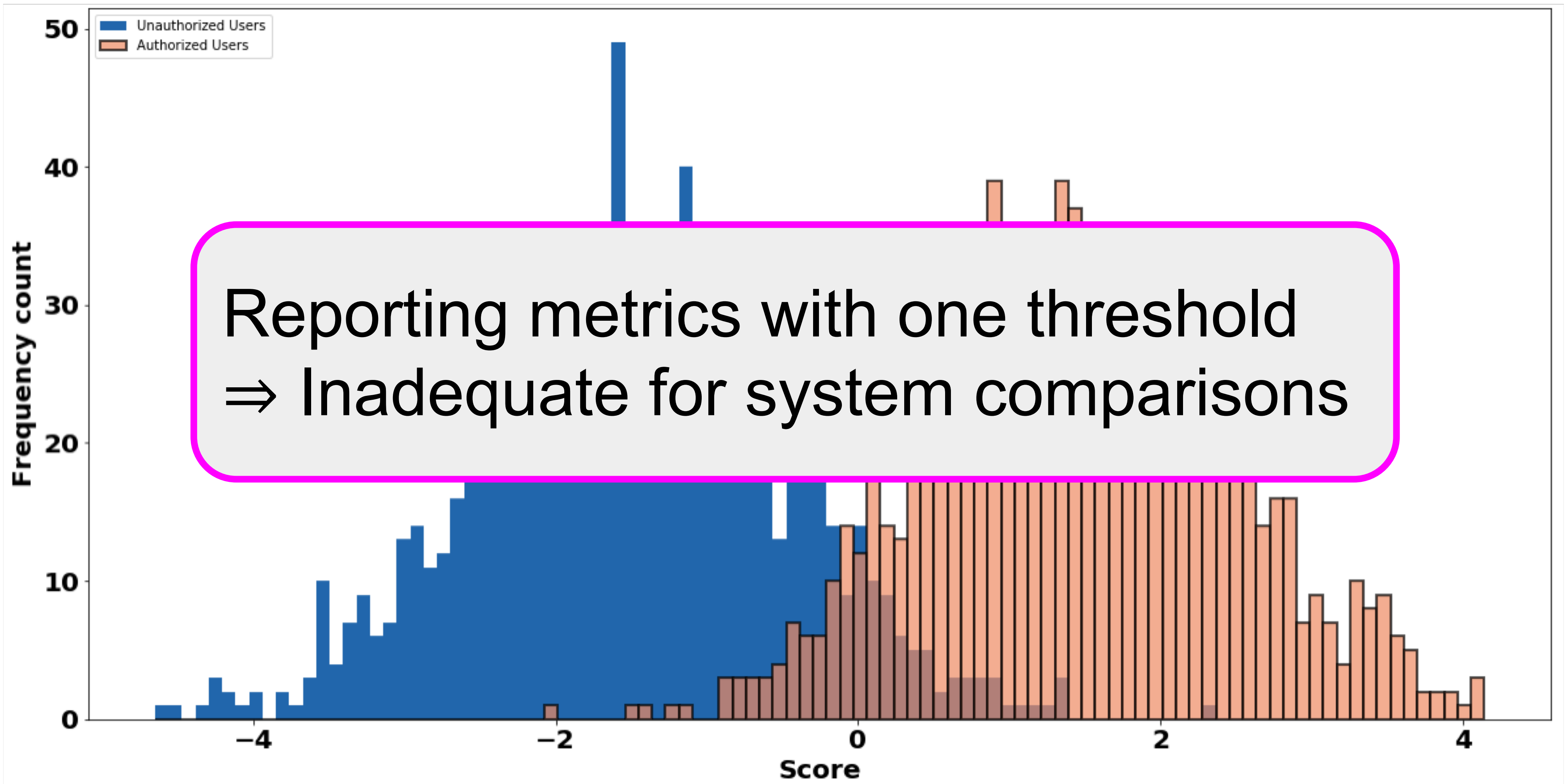
Thresholds Matter



Thresholds Matter



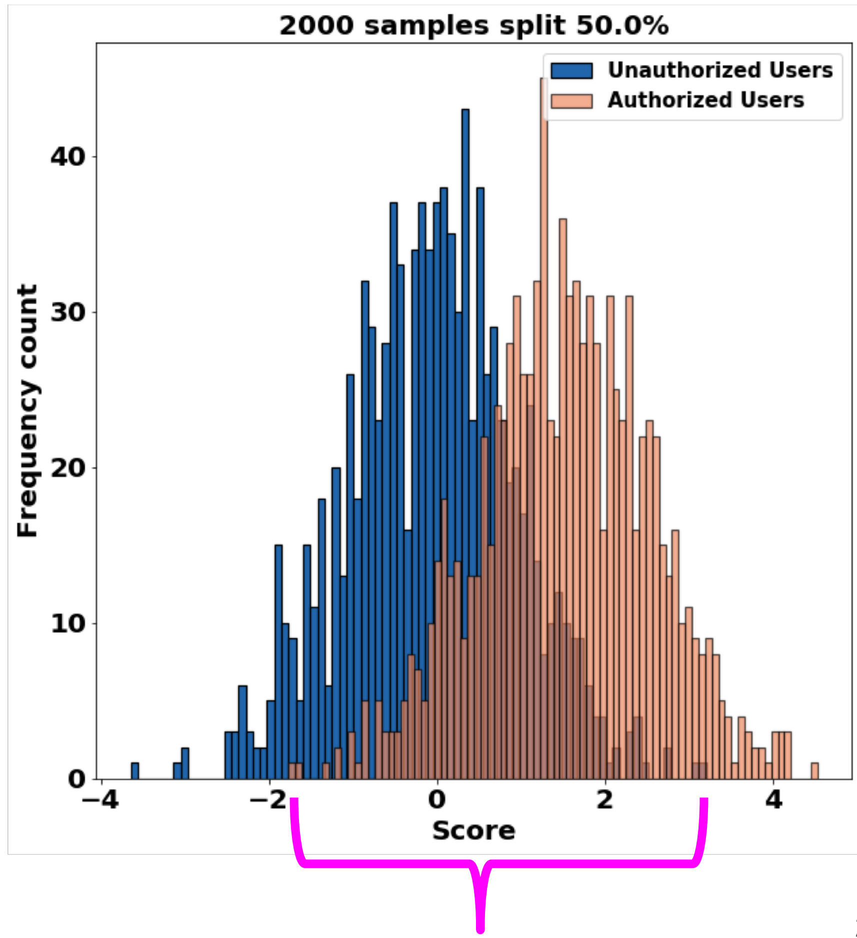
Thresholds Matter



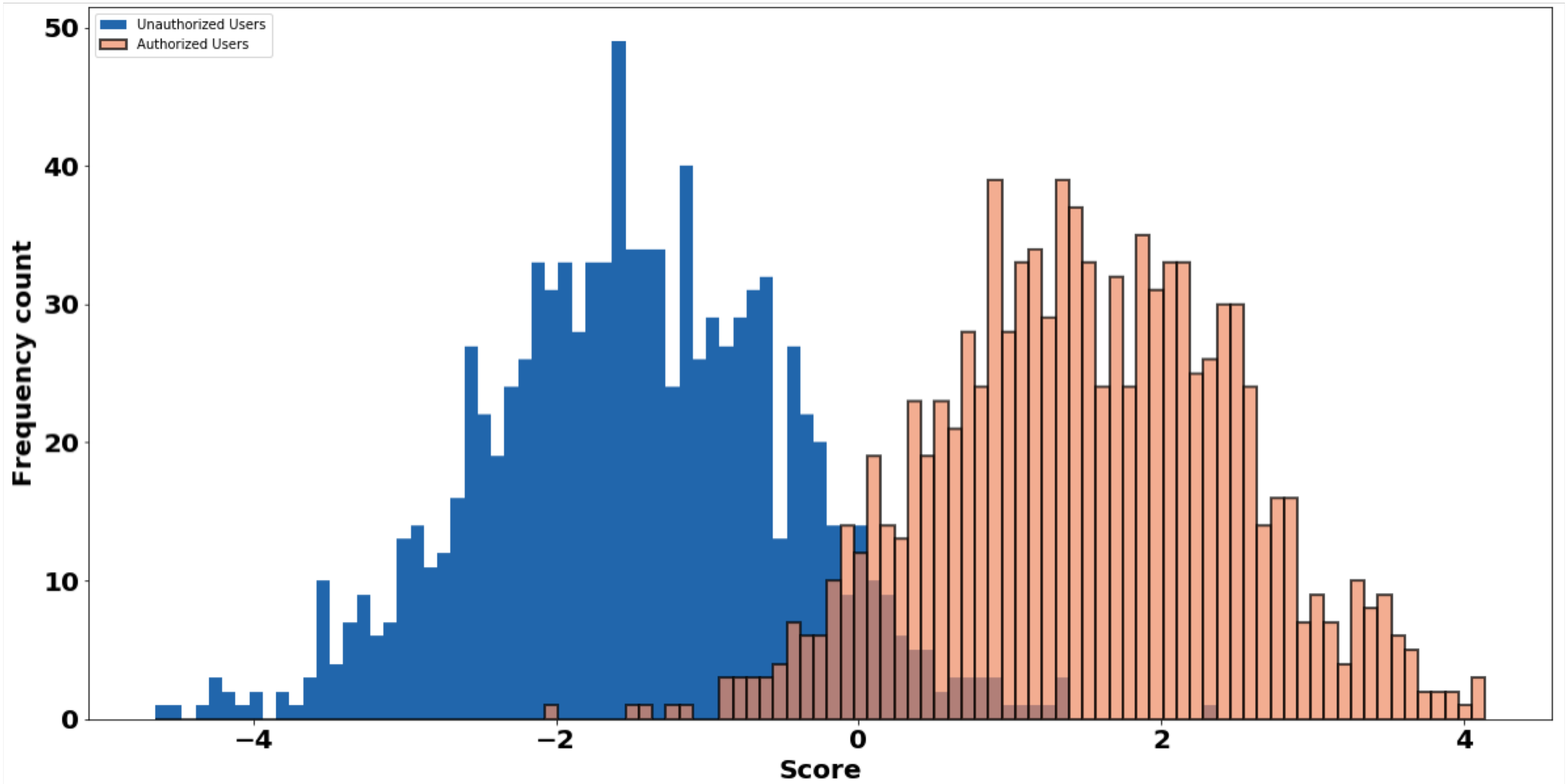
Reporting metrics with one threshold
⇒ Inadequate for system comparisons

We Propose: Frequency Count of Scores (FCS) can help

- The distribution of scores plays an important role in the system performance
- The potential for error is directly proportional to the width of the score overlap
- The FCS can be used to identify problems with scoring



Wait? Where Does All This User Data Come From?



Wait? Where Does All This User Data Come From?

Big Picture: What Is Common Between Medicine, Interventional Behavioral Sciences and Security?

PLOS | MEDICINE

BROWSE PUBLISH

OPEN ACCESS

ESSAY

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Big Picture: What Is Common Between Medicine, Interventional Behavioral Sciences and Security?

PLOS | MEDICINE BROWSE PUBLISH

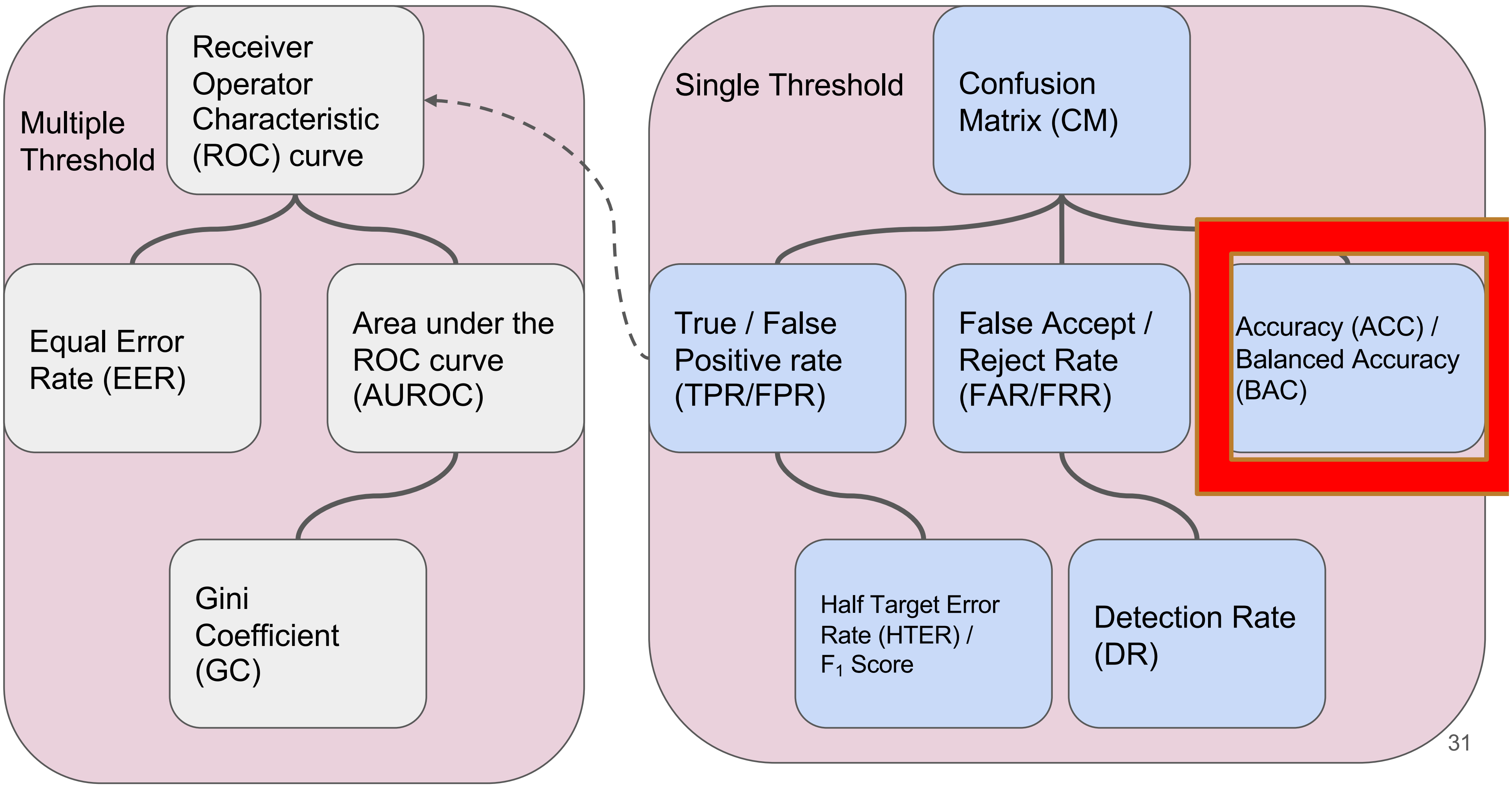
Common: Is your cure, treatment or system effective (and better than before)?

Why Most Published Research Findings Are False

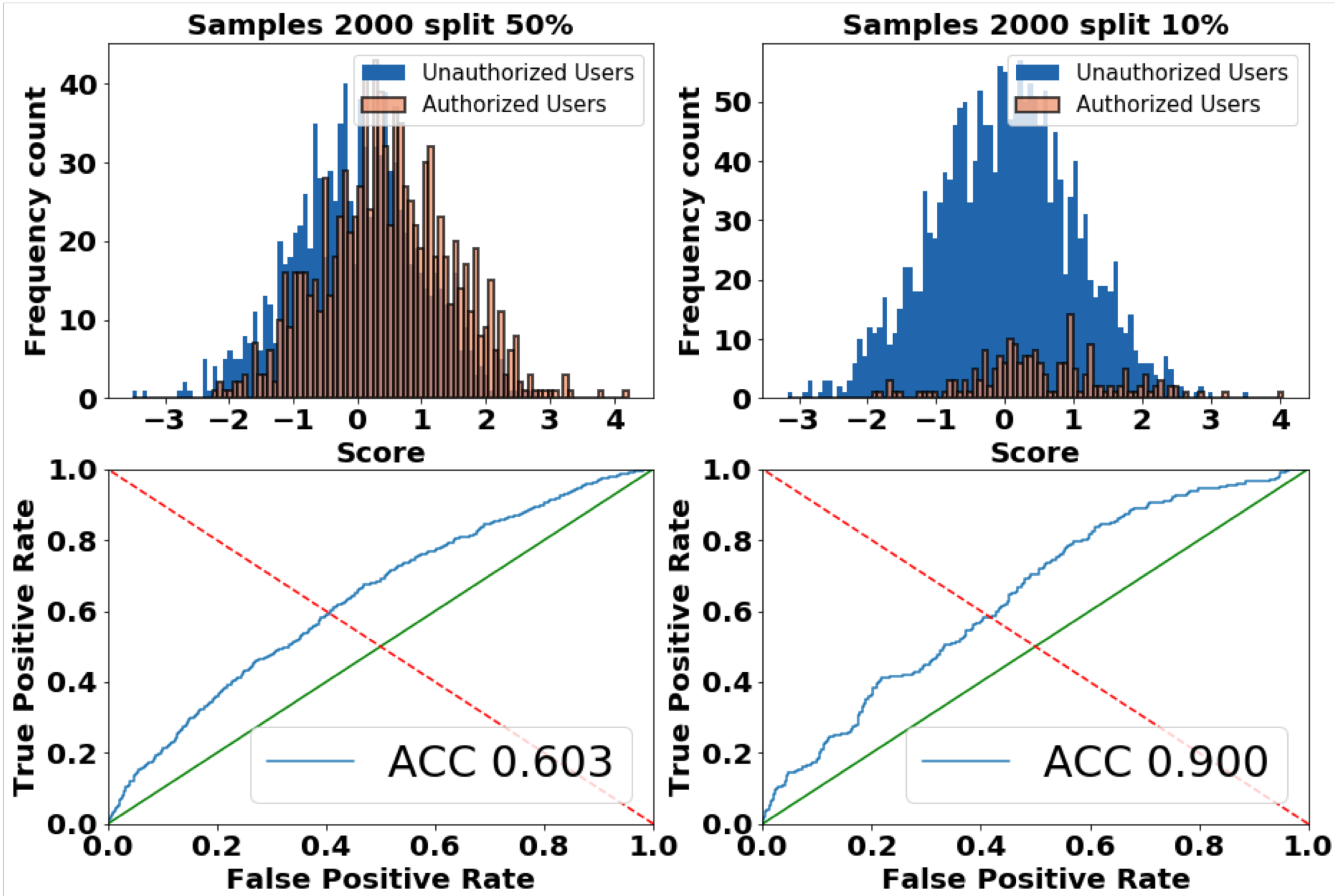
John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

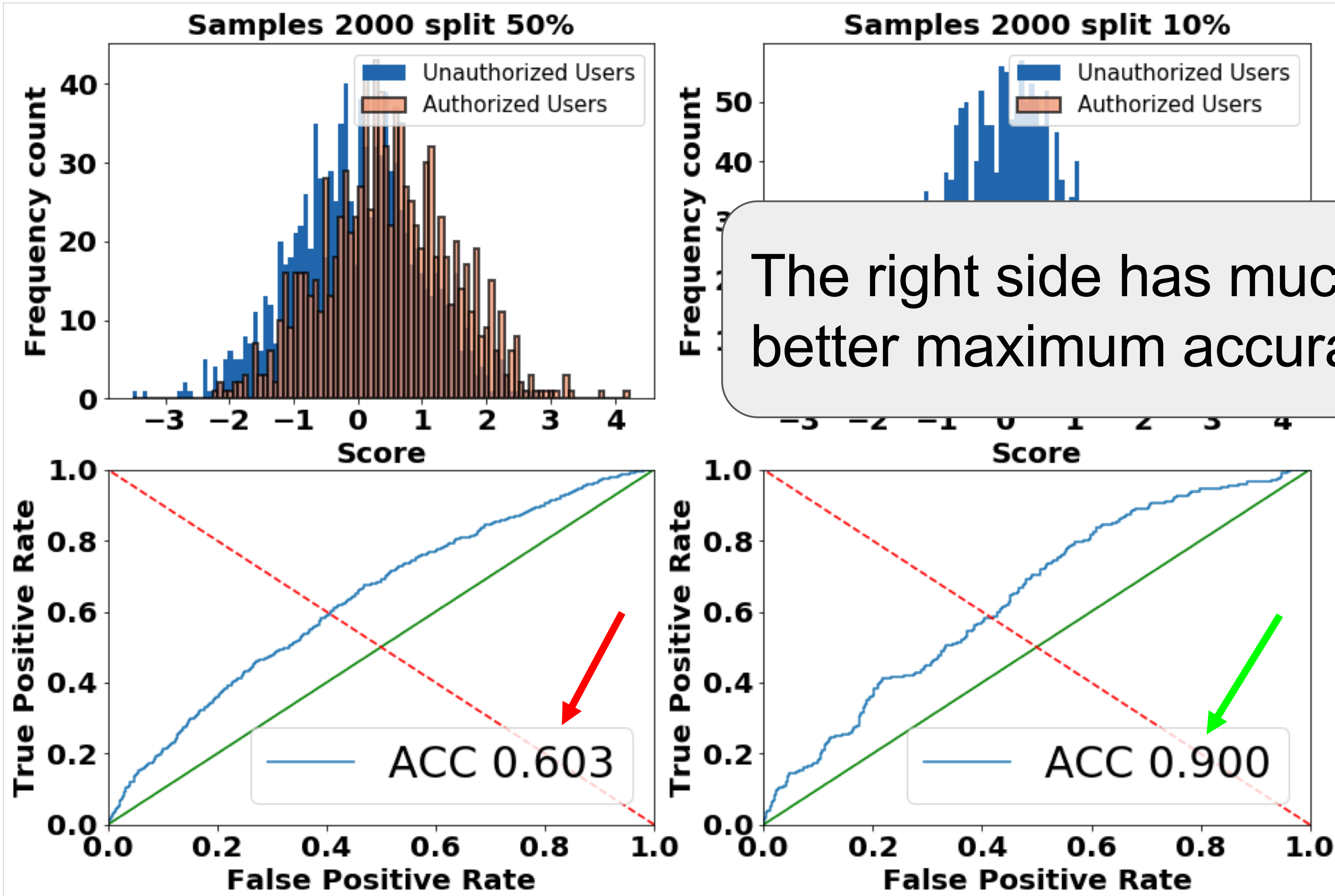
Effective: We Got High Accuracy



Sample bias -> Accuracy unreliable

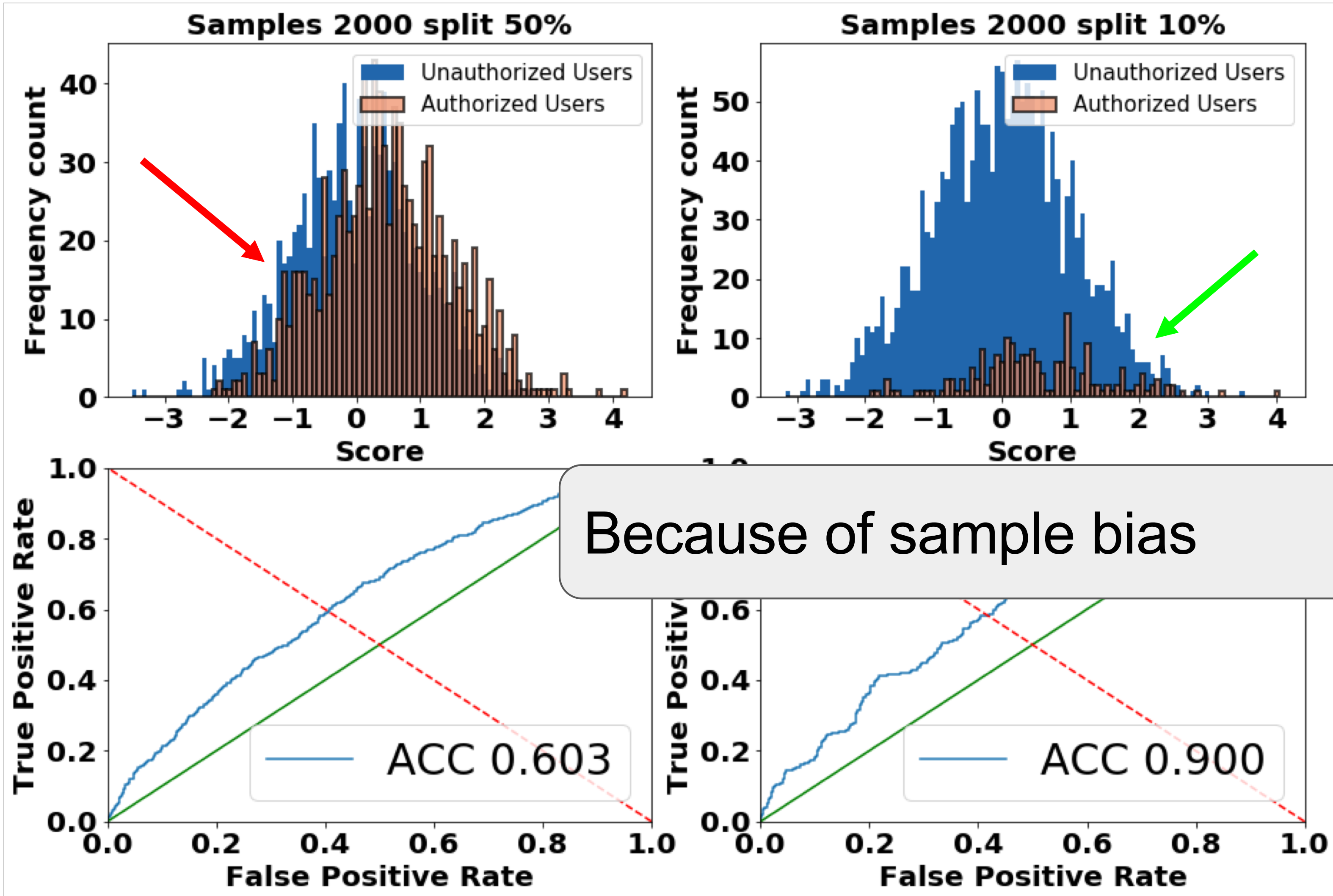


Sample bias -> Accuracy unreliable



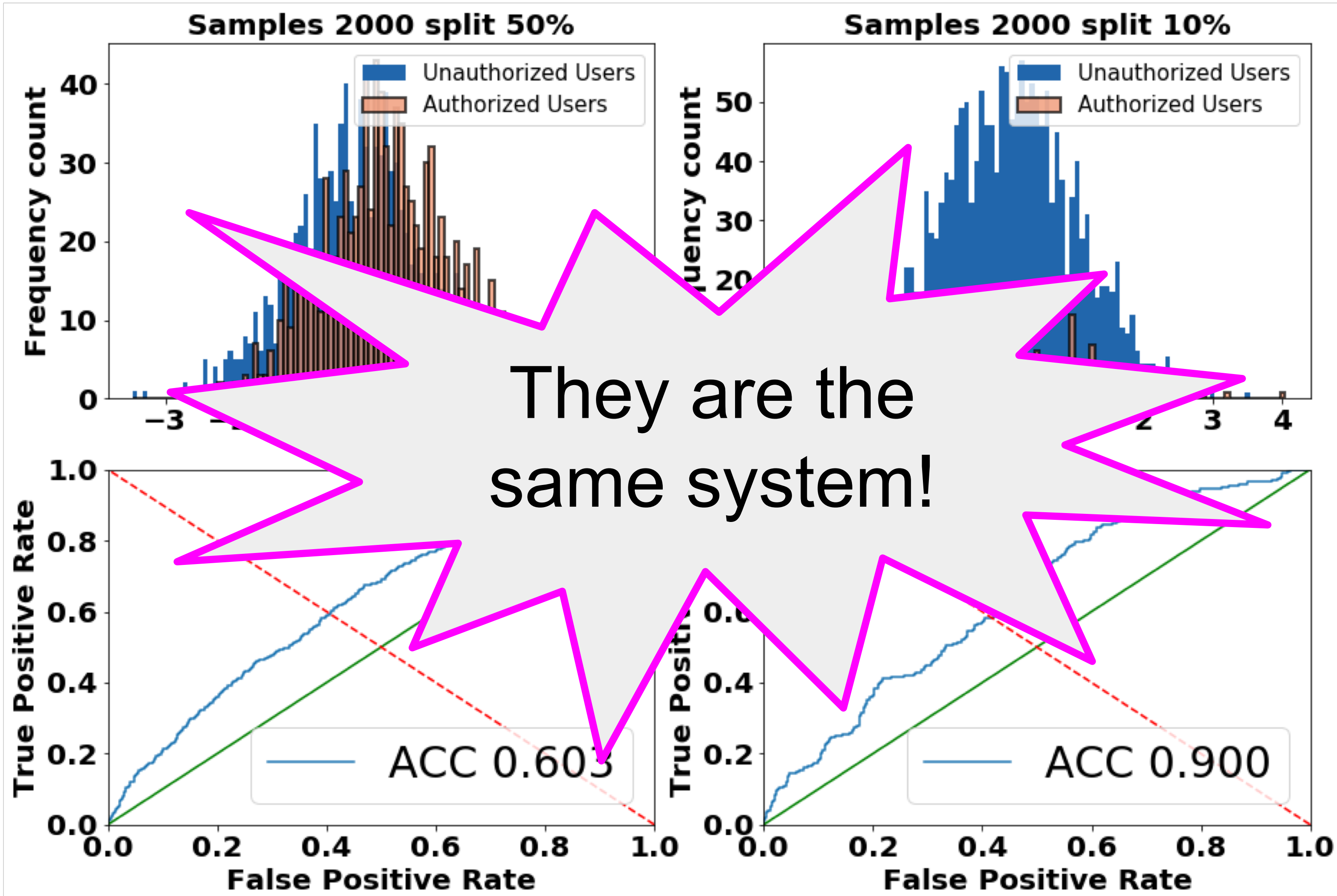
The right side has much better maximum accuracy!

Sample bias -> Accuracy unreliable



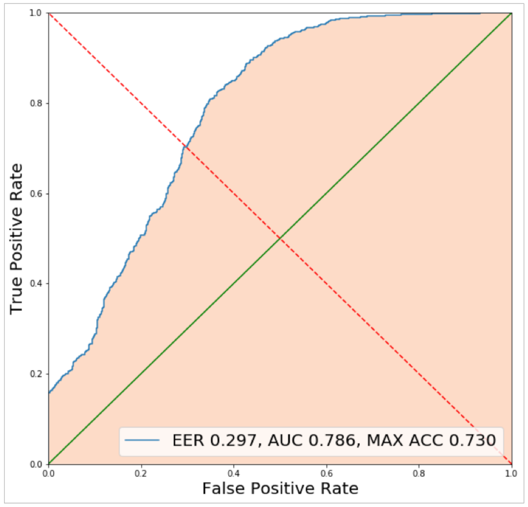
Because of sample bias

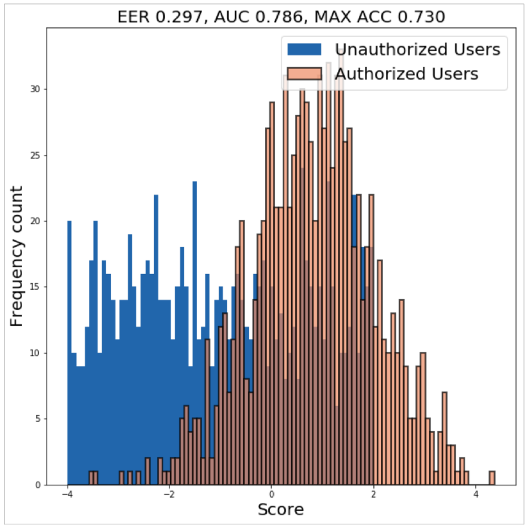
Sample bias -> Accuracy unreliable



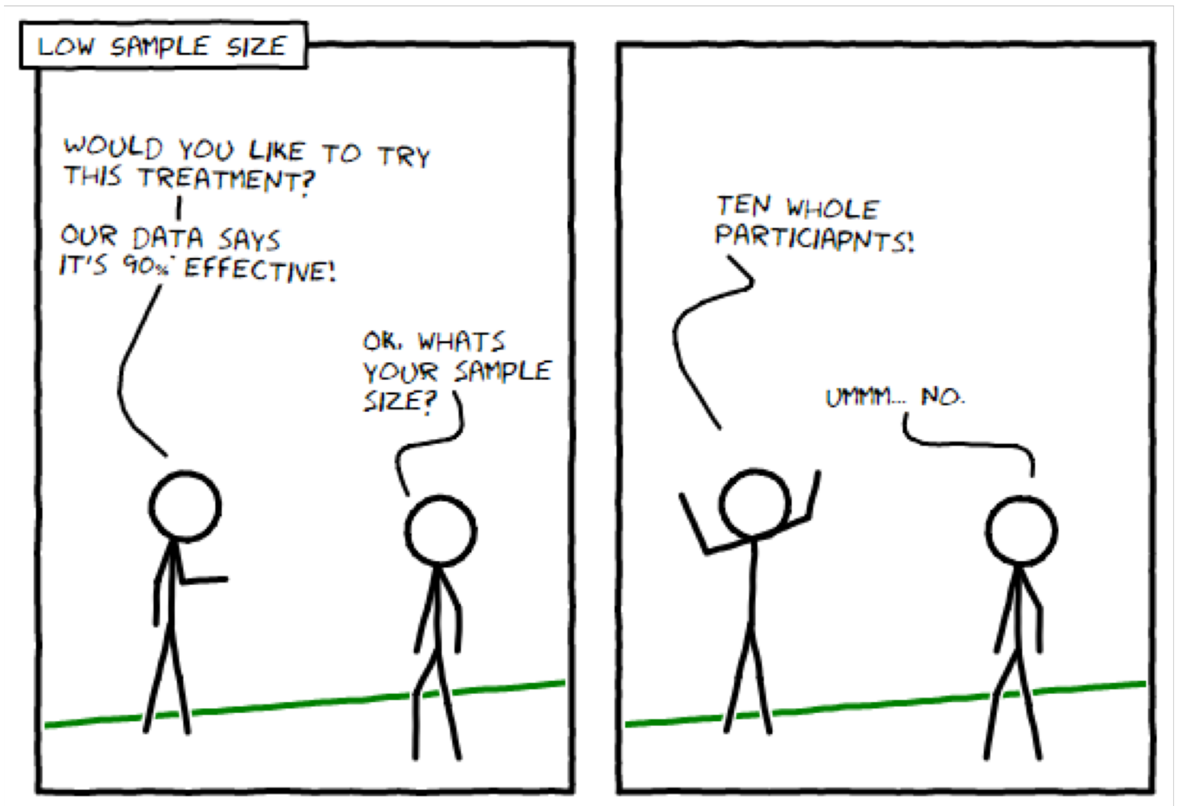
Summary so far

- We propose reporting ROC and FCS to increase transparency
- No common reporting practice across surveyed systems
  - 36 out of 38 proposed systems had flaws in reporting
- Poor performance reporting impedes system comparison and replication
- Common metrics (e.g. accuracy, EER) can be misleading and hide performance tradeoffs

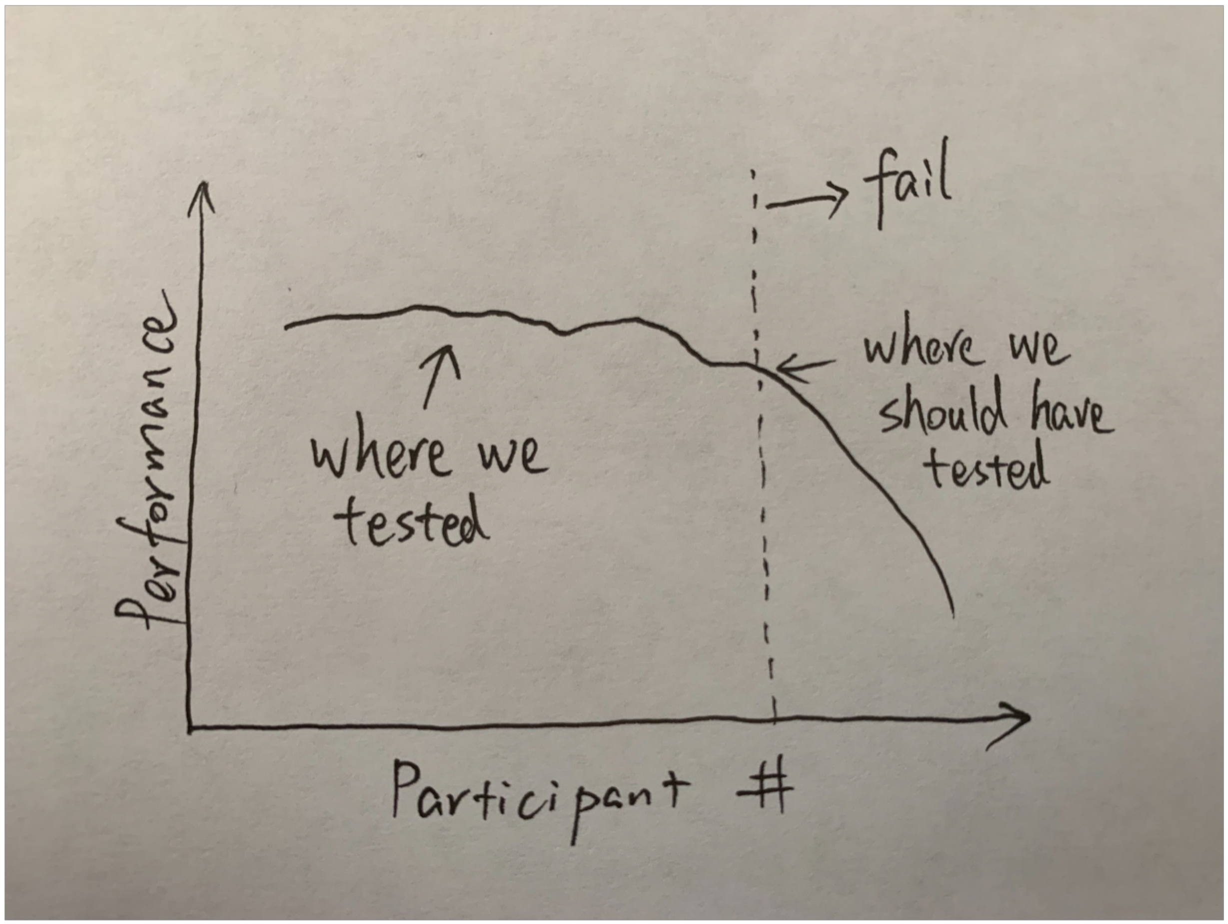




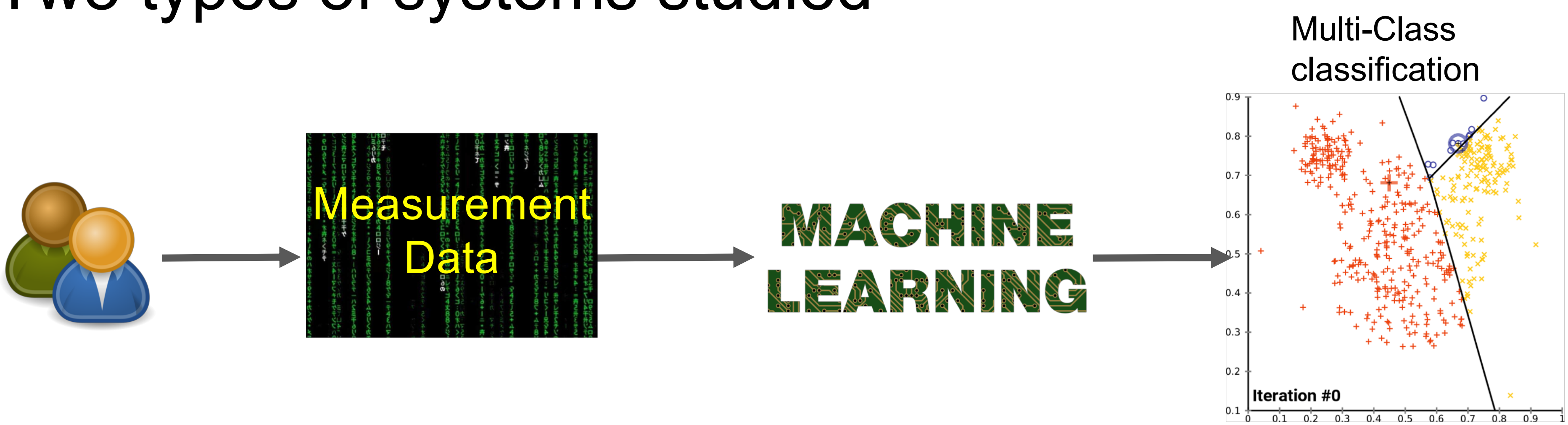
TL;DR
Testing with low participant counts does not identify the limits of system performance

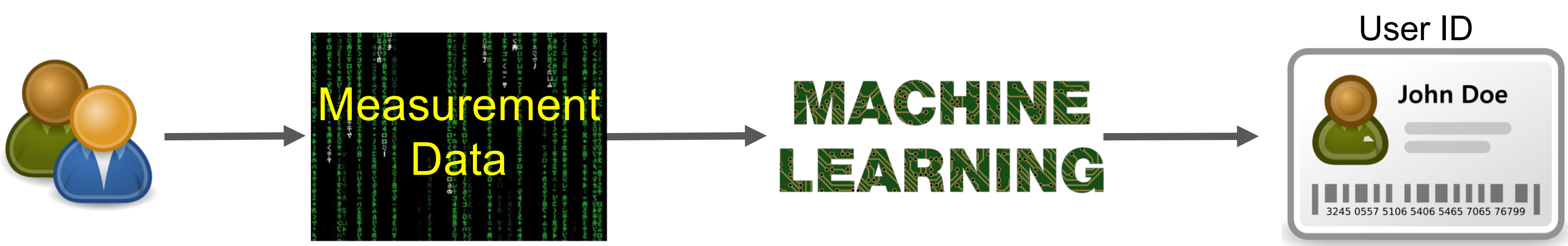


TL;DR
Testing with low participant counts does not identify the limits of system performance



Two types of systems studied



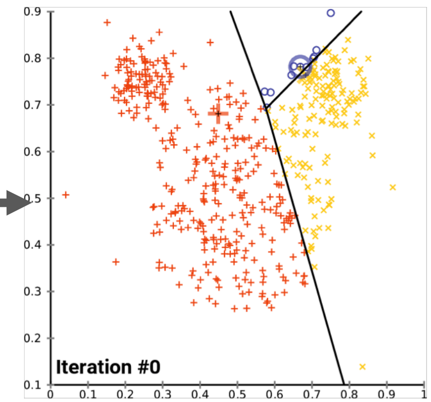


Not Really

Two types of systems studied

MACHINE LEARNING

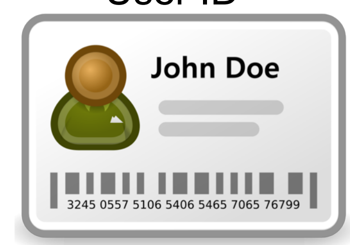
Multi-Class classification



User identification is a special case of multi-class classification where there each class corresponds to a specific user

MACHINE LEARNING

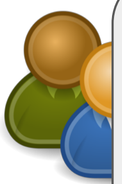
User ID

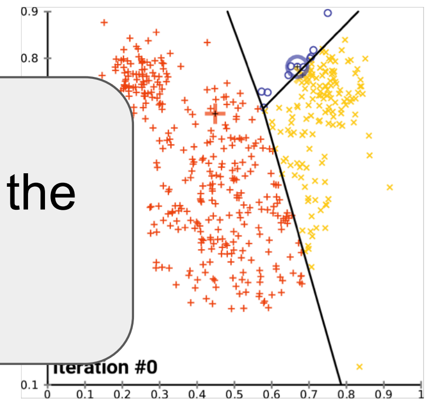


Not Really

Two types of systems studied

Multi-Class classification

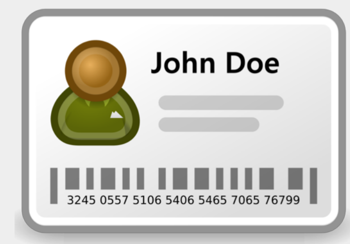




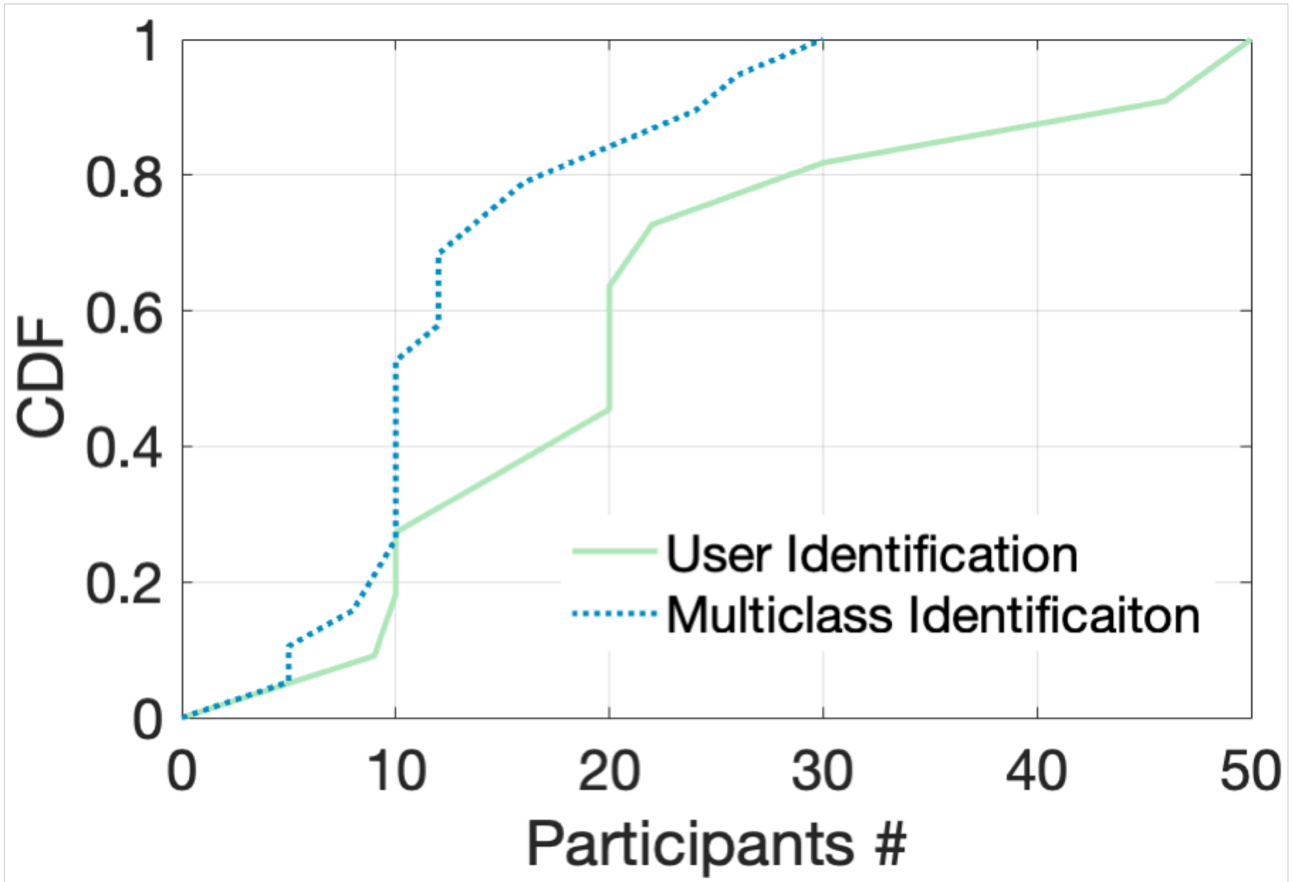
We will focus on user identification systems, but some of the analysis holds for the broader class of problems.

MACHINE LEARNING

User ID

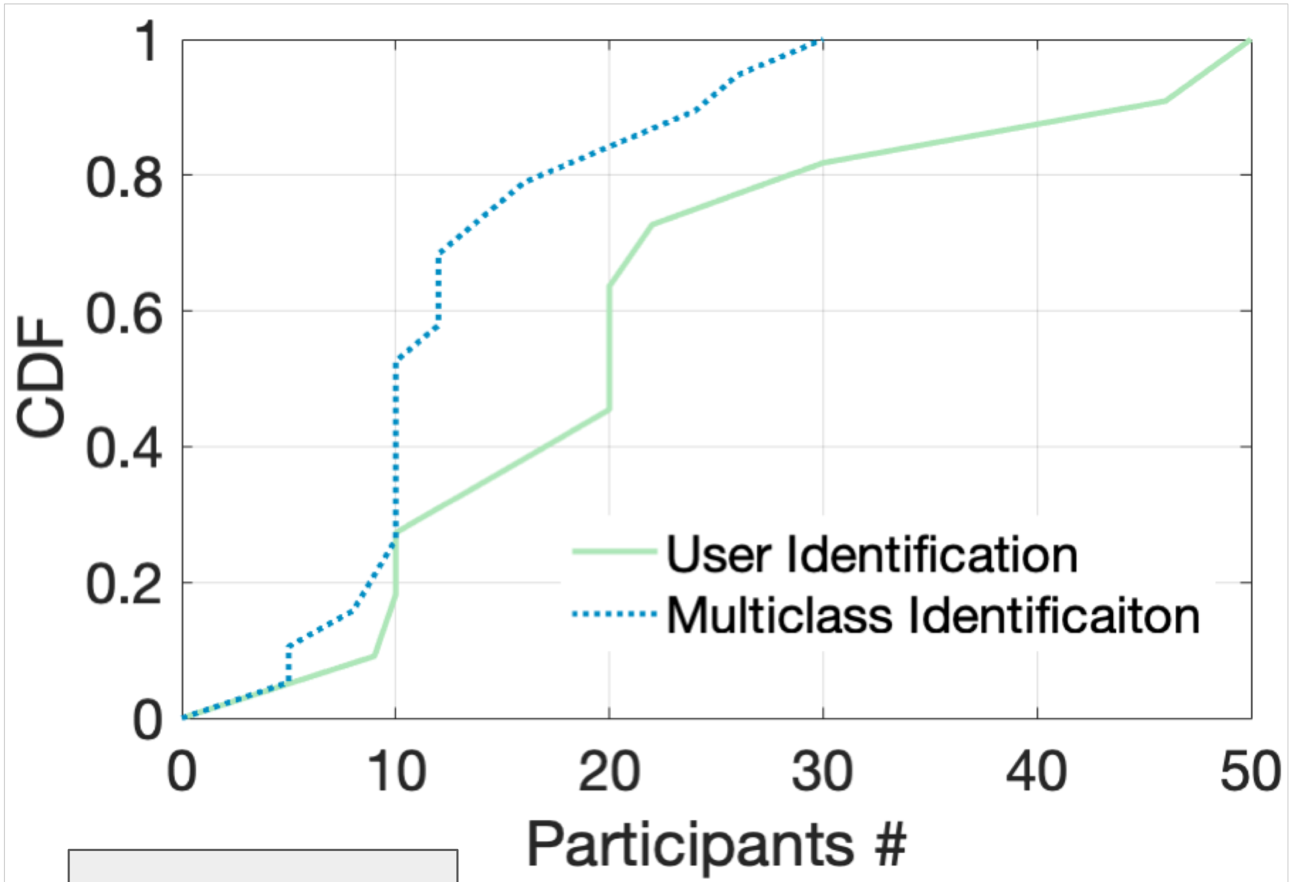


Problem: Testing with too few participants



Venue	Case 1: User Identification	Case 2: Multiclass Identification
IMWUT/UBICOMP	[55], [49], [67], [26], [61], [62], [13]	[31], [52], [63], [69], [57], [65], [43]
CHI	[46], [38], [45]	[66], [39], [64], [47], [34]
UIST		[5], [22], [33]
OTHERS	[24]	[28], [56], [60], [59]

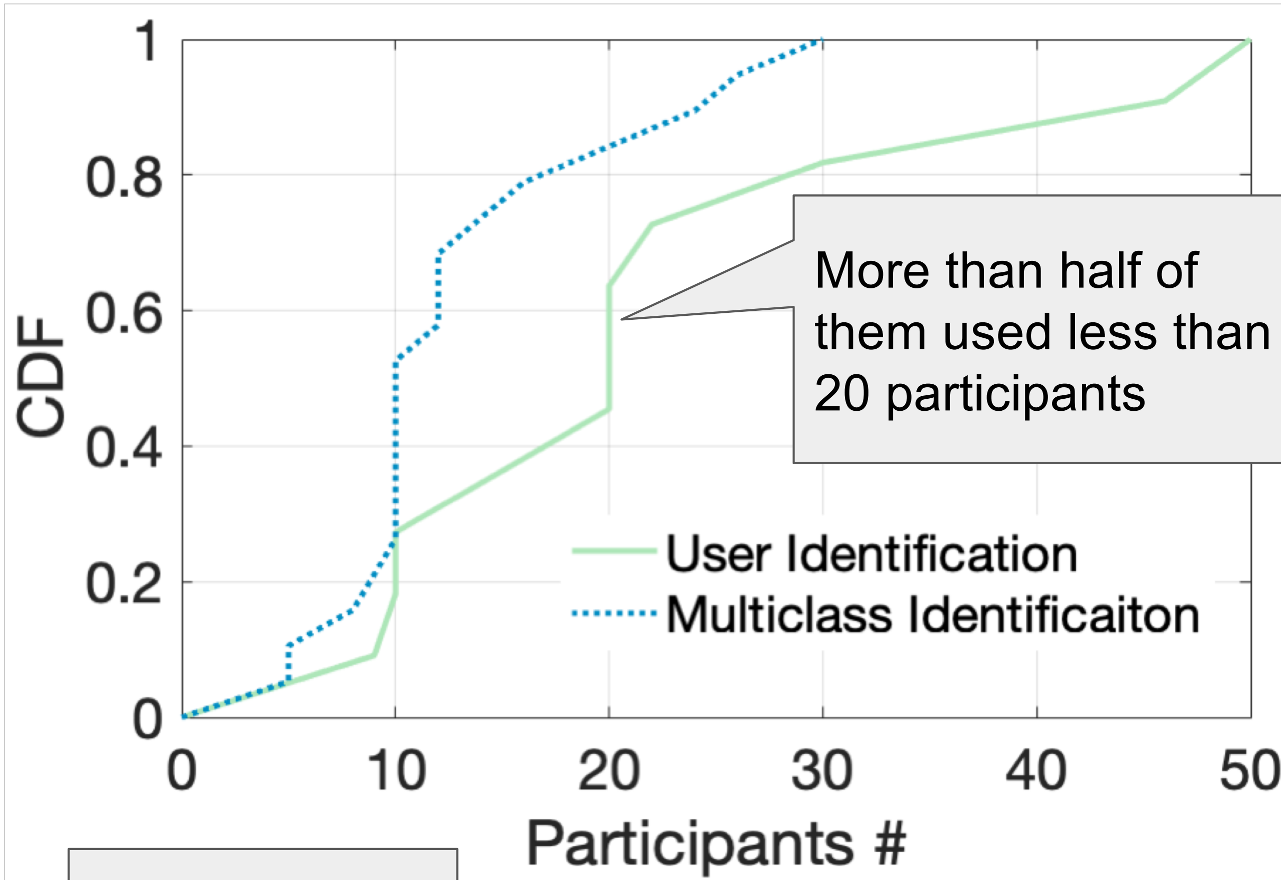
Problem: Testing with too few participants



We surveyed 30 proposed systems

Venue	Case 1: User Identification	Case 2: Multiclass Identification
IMWUT/UBICOMP	[55], [49], [67], [26], [61], [62], [13]	[31], [52], [63], [69], [57], [65], [43]
CHI	[46], [38], [45]	[66], [39], [64], [47], [34]
UIST		[5], [22], [33]
OTHERS	[24]	[28], [56], [60], [59]

Problem: Testing with too few participants

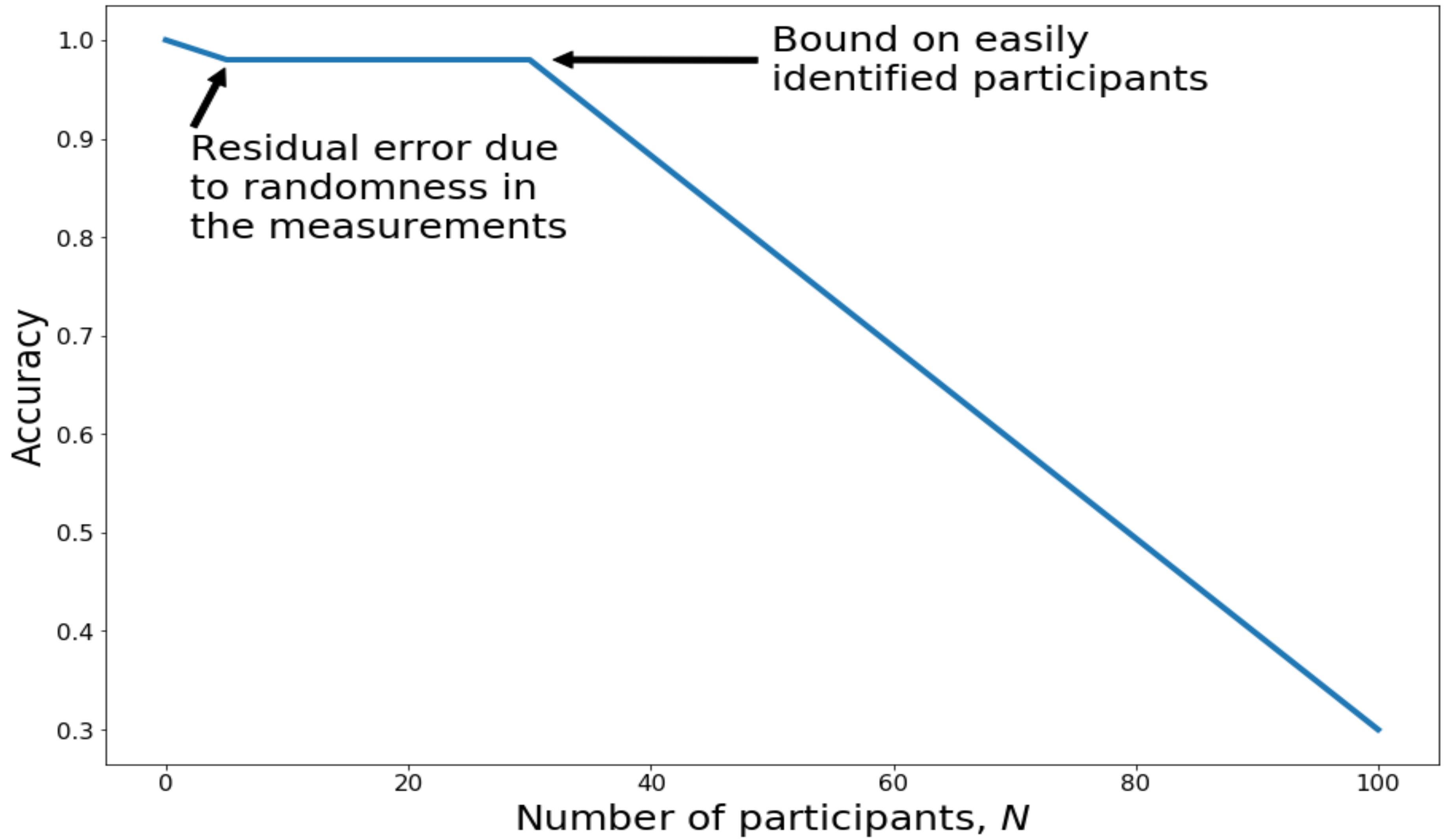


More than half of them used less than 20 participants

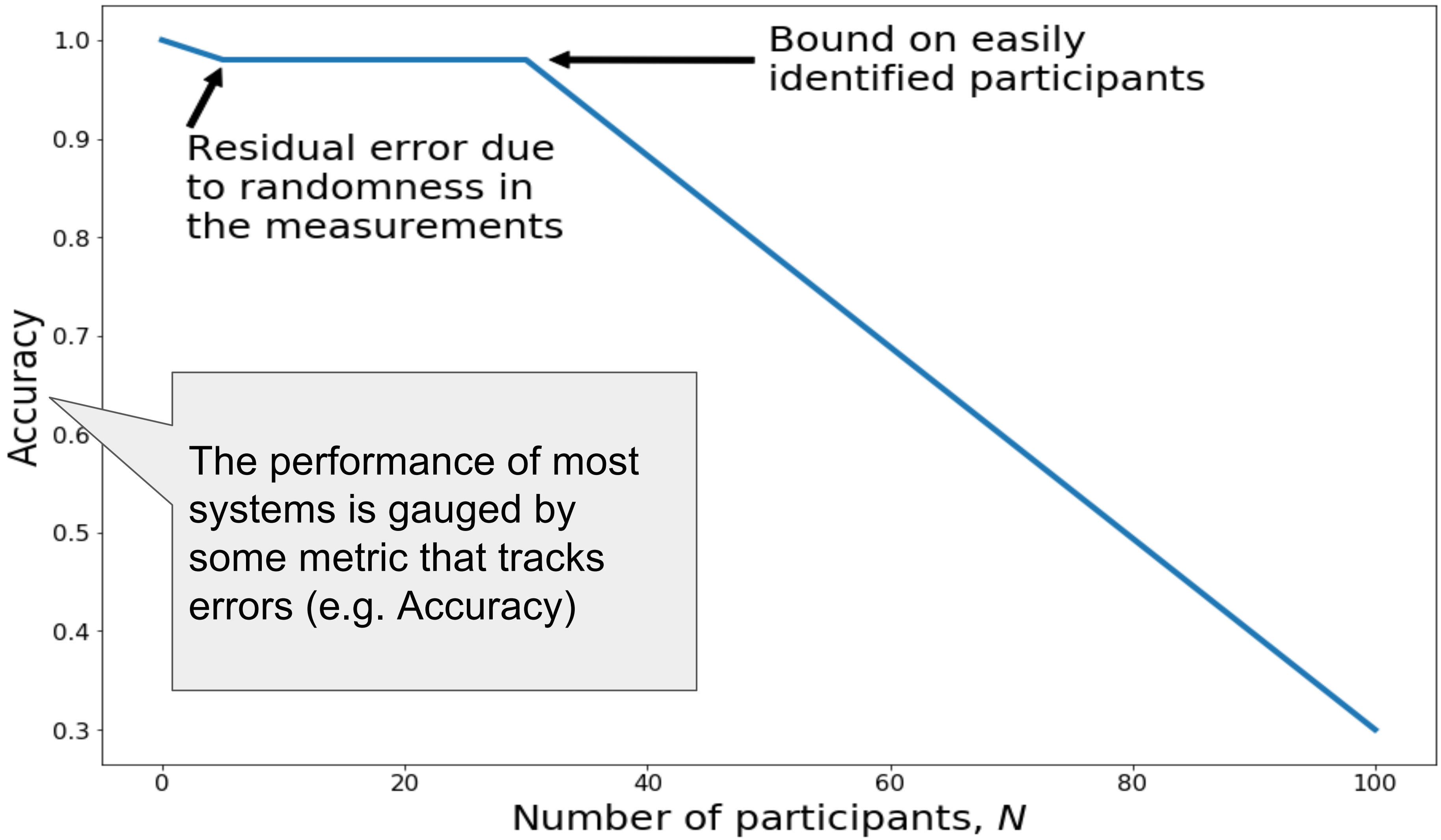
We surveyed 30 proposed systems

Venue	Case 1: User Identification	Case 2: Multiclass Identification
IMWUT/UBICOMP	[55], [49], [67], [26], [61], [62], [13]	[31], [52], [63], [69], [57], [65], [43]
CHI	[46], [38], [45]	[66], [39], [64], [47], [34]
UIST		[5], [22], [33]
OTHERS	[24]	[28], [56], [60], [59]

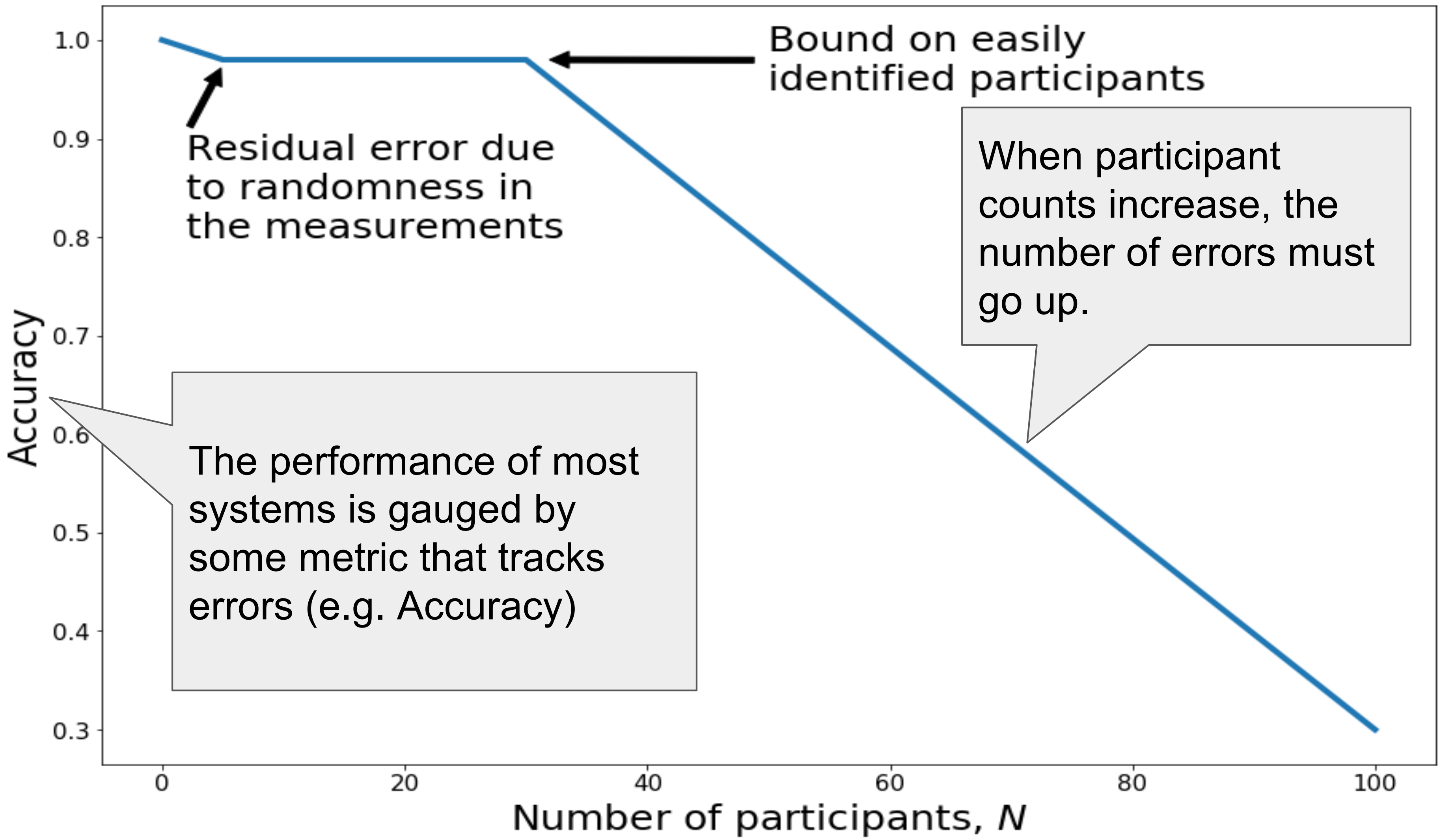
What Is Going On? (simplified)



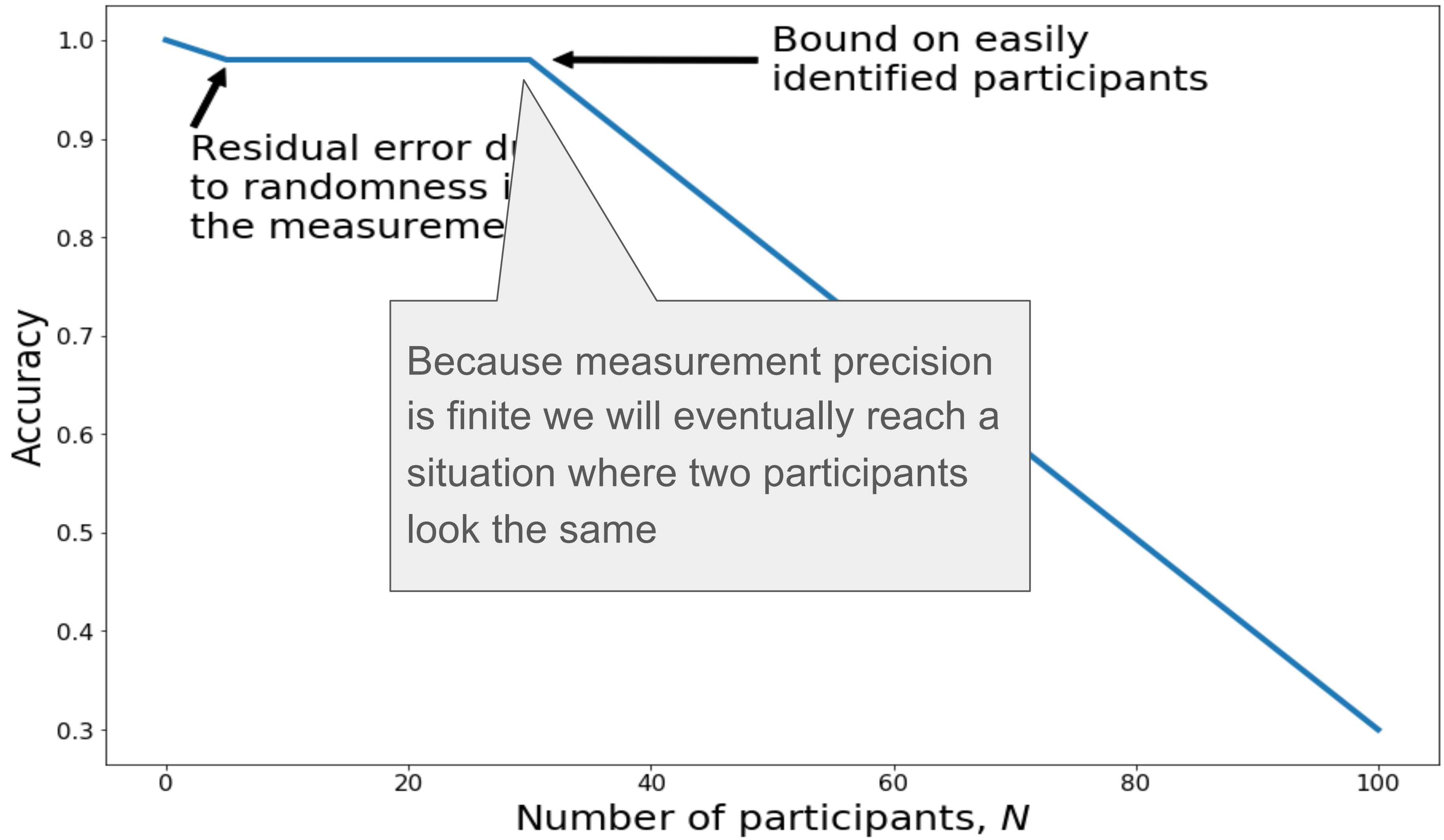
What Is Going On? (simplified)



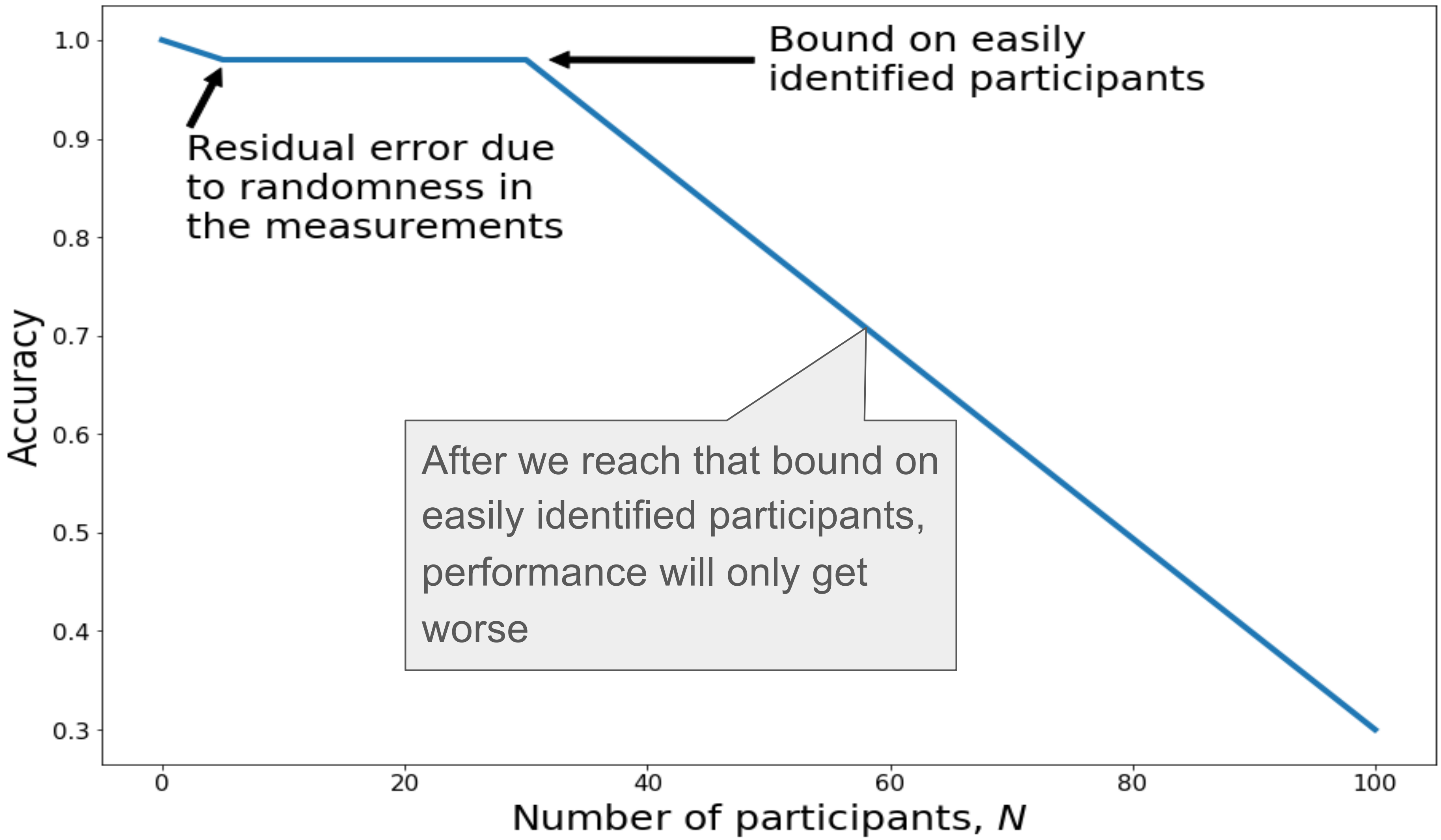
What Is Going On? (simplified)



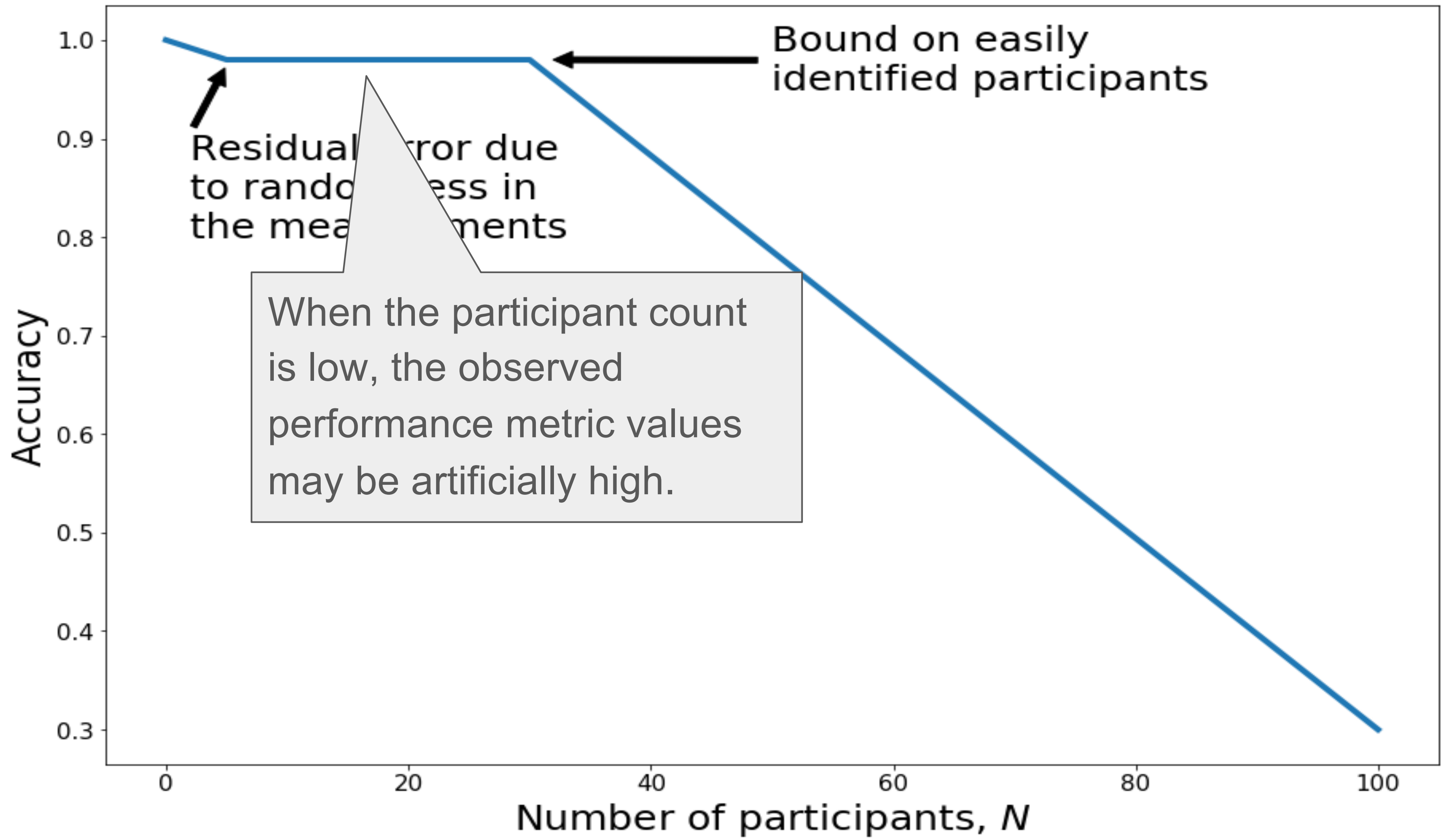
What Is Going On? (simplified)



What Is Going On? (simplified)



What Is Going On? (simplified)



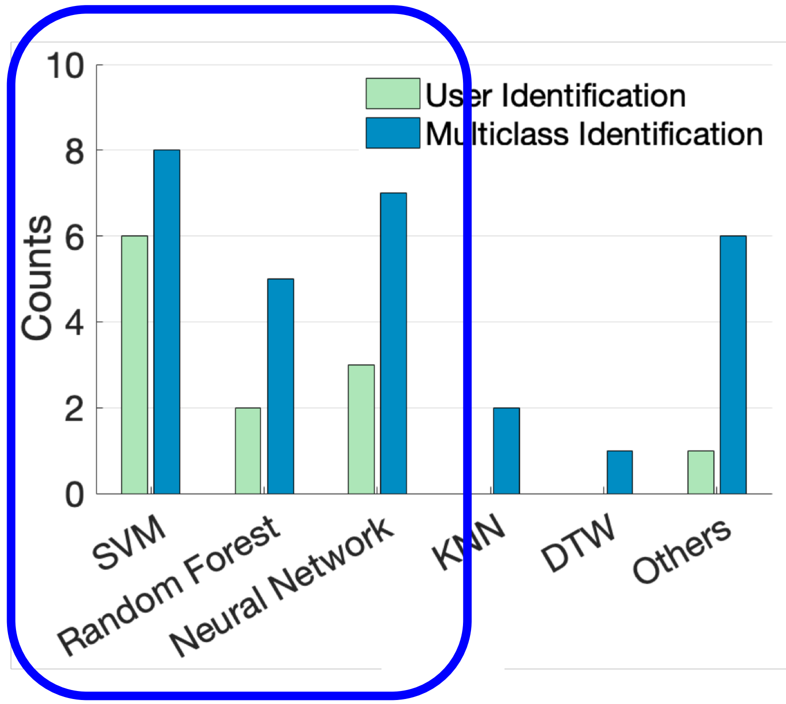
Hammer Time

- What if…
- … we just took some arbitrary datasets of humans and used **artificial intelligence** on them?

Building 5 example user identification systems

- 3 criteria to selecting datasets:
  - Have a unique identifier for each participant
  - Have at least 20 participants
  - Have more than one measurement per participant

- 3 most popular classification algorithms
  - Support Vector Machine
  - Random Forest
  - Neural Network
- 10 iterations with randomly selected participants
  - We did the minimal amount of tuning necessary to generate output



Building 5 example user identification systems

- We performed the minimum tuning possible for each system
- We report two metrics
  - confusion matrix (not shown)
  - accuracy (ACC)

	EEG	NBA Stat.	Act. Recogn.	Walking Act.	CT Scan
Neural Network	0.5122	0.9452	0.8153	0.5901	0.9996
Random Forest	0.5176	0.9583	0.9246	0.7119	0.9992
SVM	0.3297	0.7945	0.7875	0.5679	1.0000

Building 5 example user identifi[illegible]ation systems

We performed the mini[illegible]m tuning pos[illegible]r each system

- W[illegible]t two metrics
    - c[illegible]
    - max[illegible]

[illegible]hich al[illegible]

- th[illegible]

[illegible]e

With minimal tuning 3 of the systems achieved accuracy > 90%

	EE[illegible]	NBA Stat.	R[illegible]gn.	Walking Act.	CT Scan
Neural Network	0.51[illegible]	0.9452	0.8[illegible]53	0.5901	0.9996
Random Forest	0.517[illegible]	0.9583	0.9246	0.7119	0.9992
SVM	0.3297	0.7945	0.7875	0.5679	1.0000

Building 5 example user identification systems

A favorable combination of algorithm and dataset can inflate the performance values significantly, making the classification artificially easy

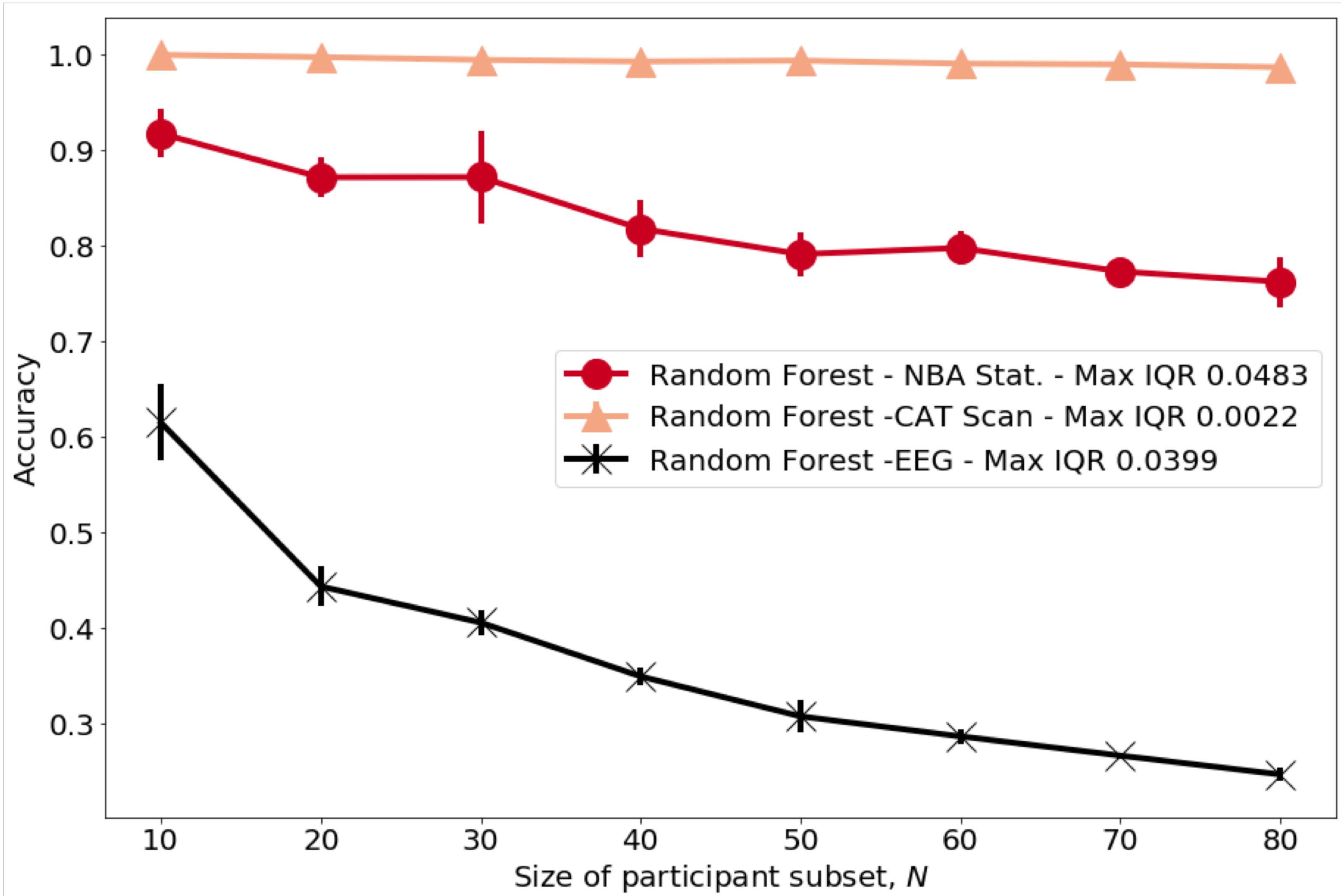
96%

92%

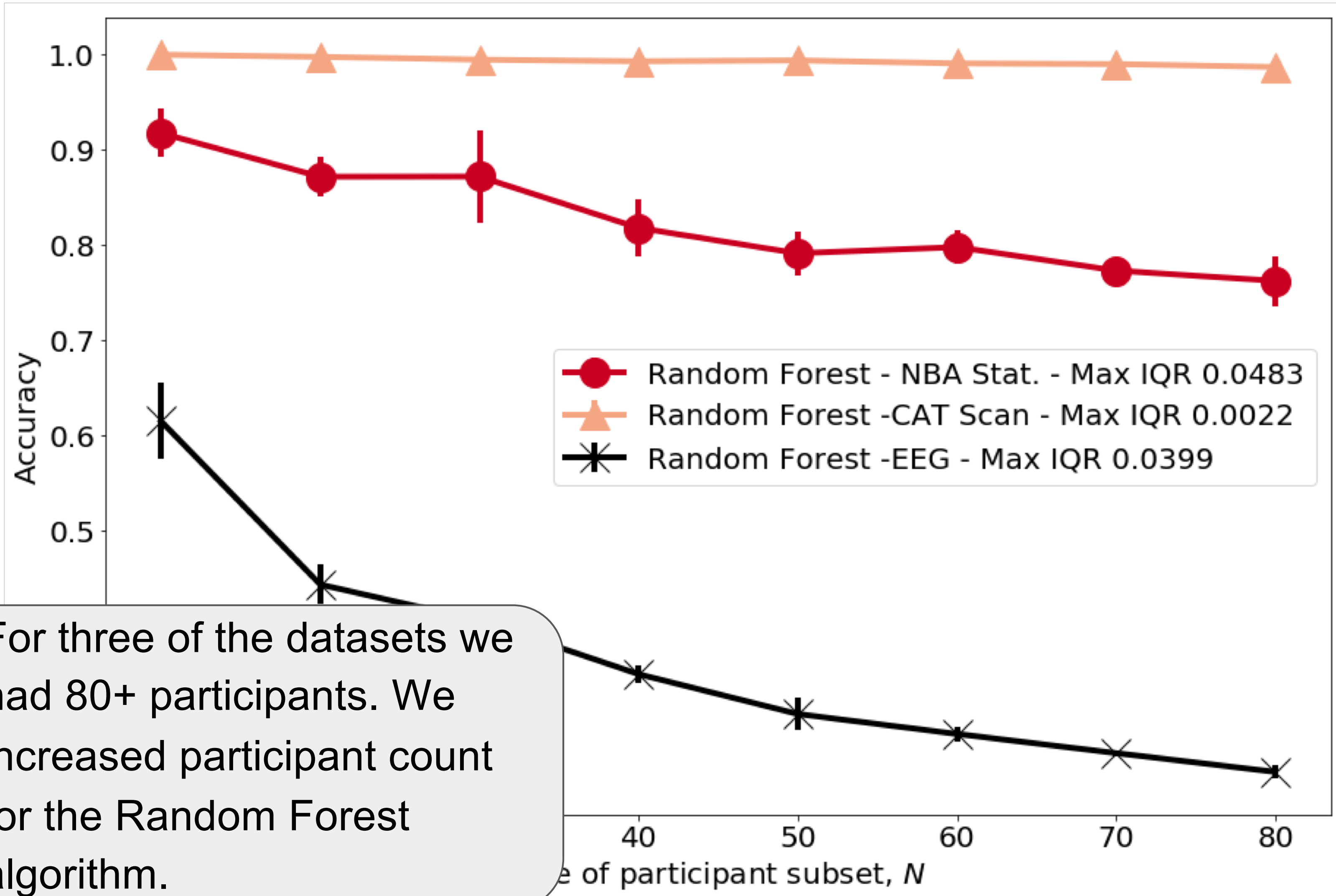
100%

	EEG	NBA Stat.	Act. Recogn.	Walking Act.	CT Scan
Neural Network	0.5122	0.9452	0.8153	0.5901	0.9996
Random Forest	0.5176	0.9583	0.9246	0.7119	0.9992
SVM	0.3297	0.7945	0.7875	0.5679	1.0000

Each system fails at a different point

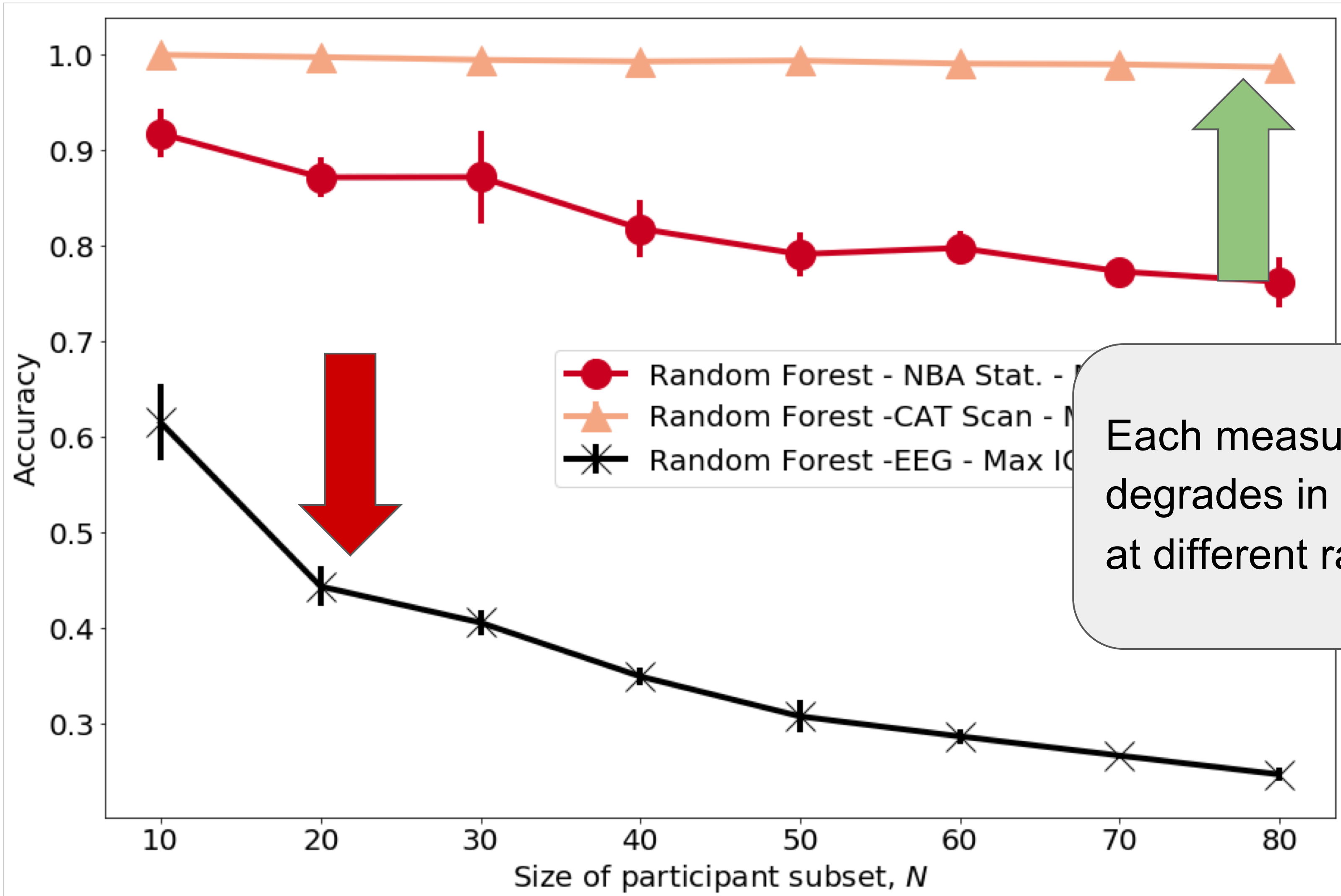


Each system fails at a different point



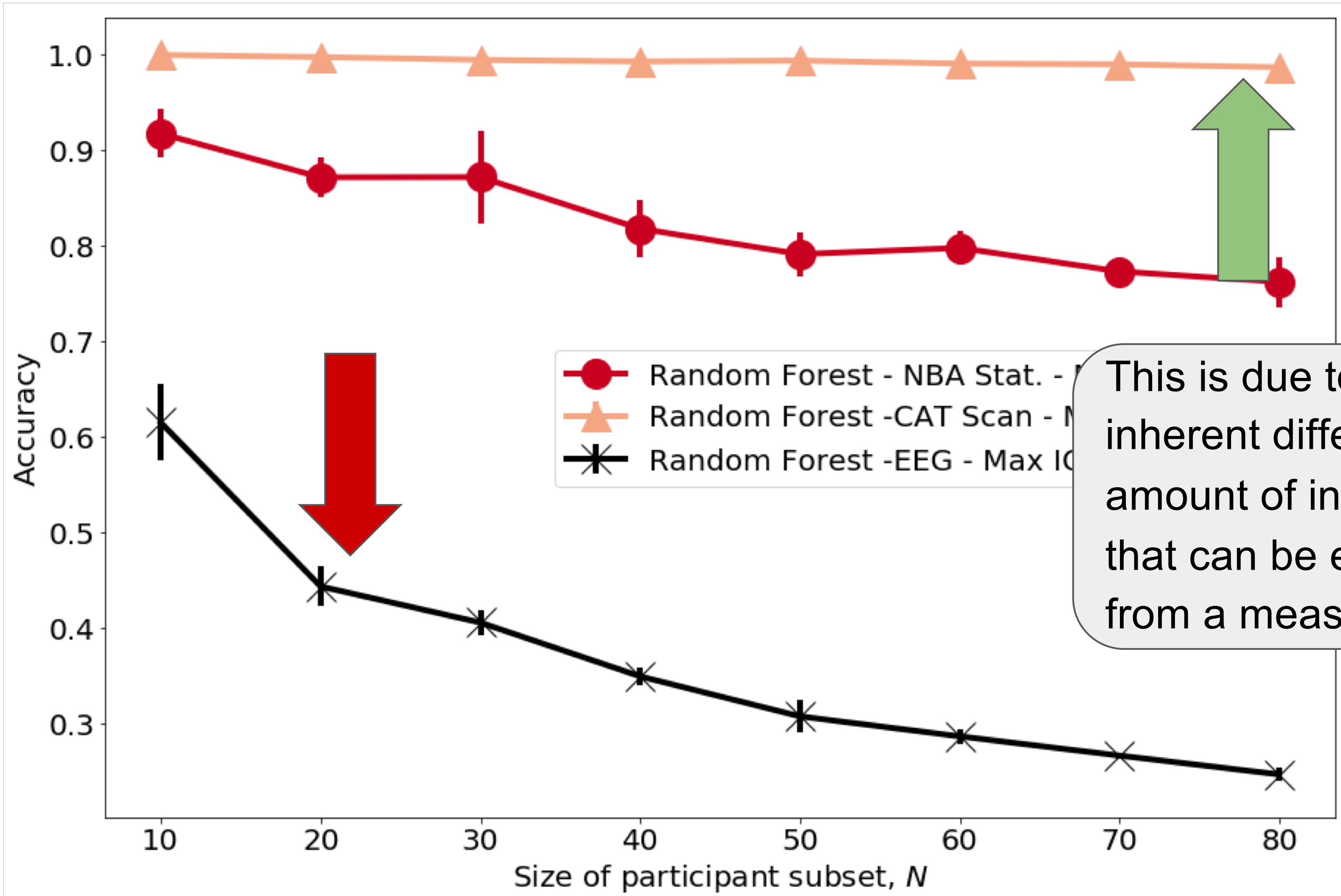
For three of the datasets we had 80+ participants. We increased participant count for the Random Forest algorithm.

Each system fails at a different point



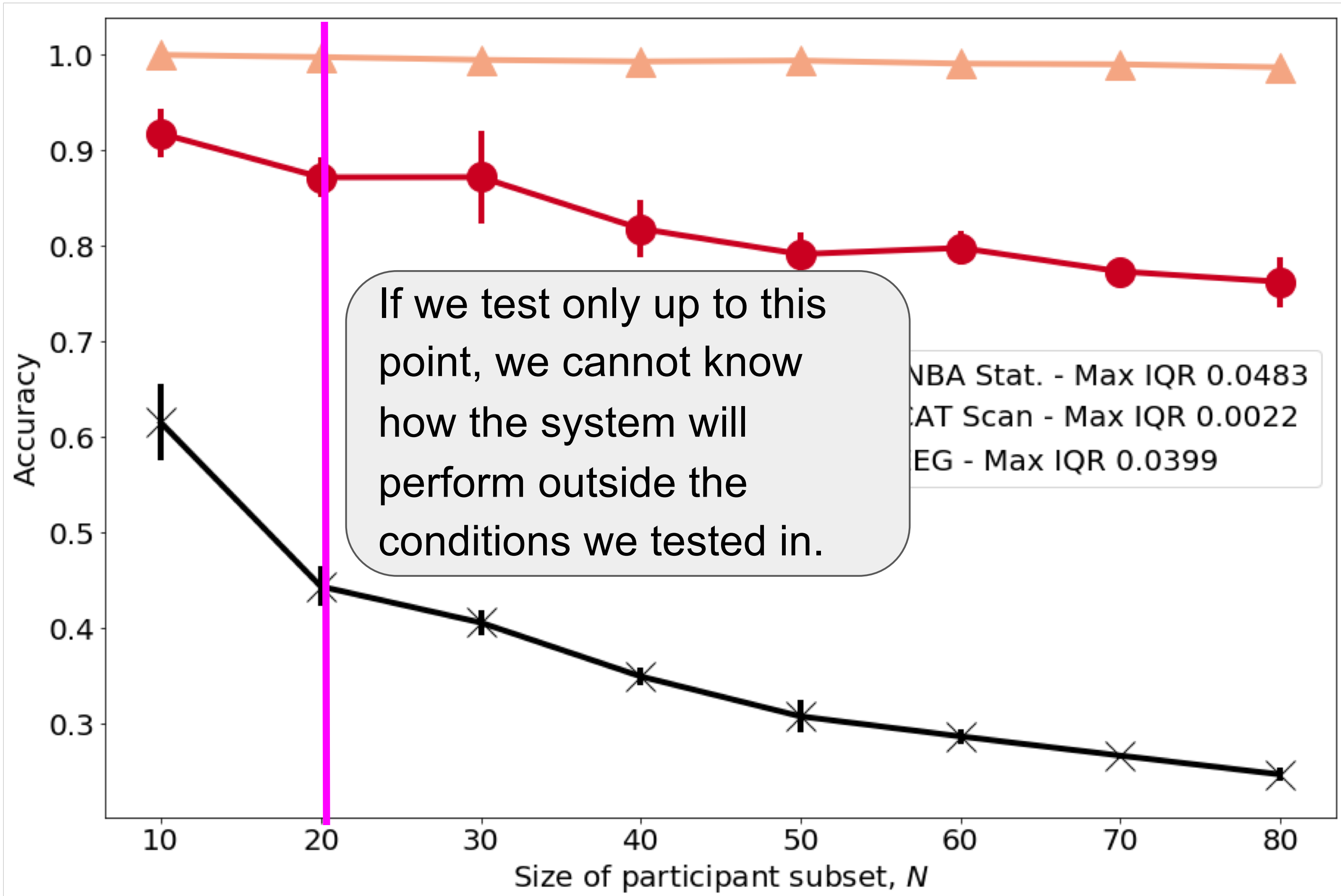
Each measurement degrades in performance at different rates.

Each system fails at a different point



This is due to the inherent difference in the amount of information that can be extracted from a measurement

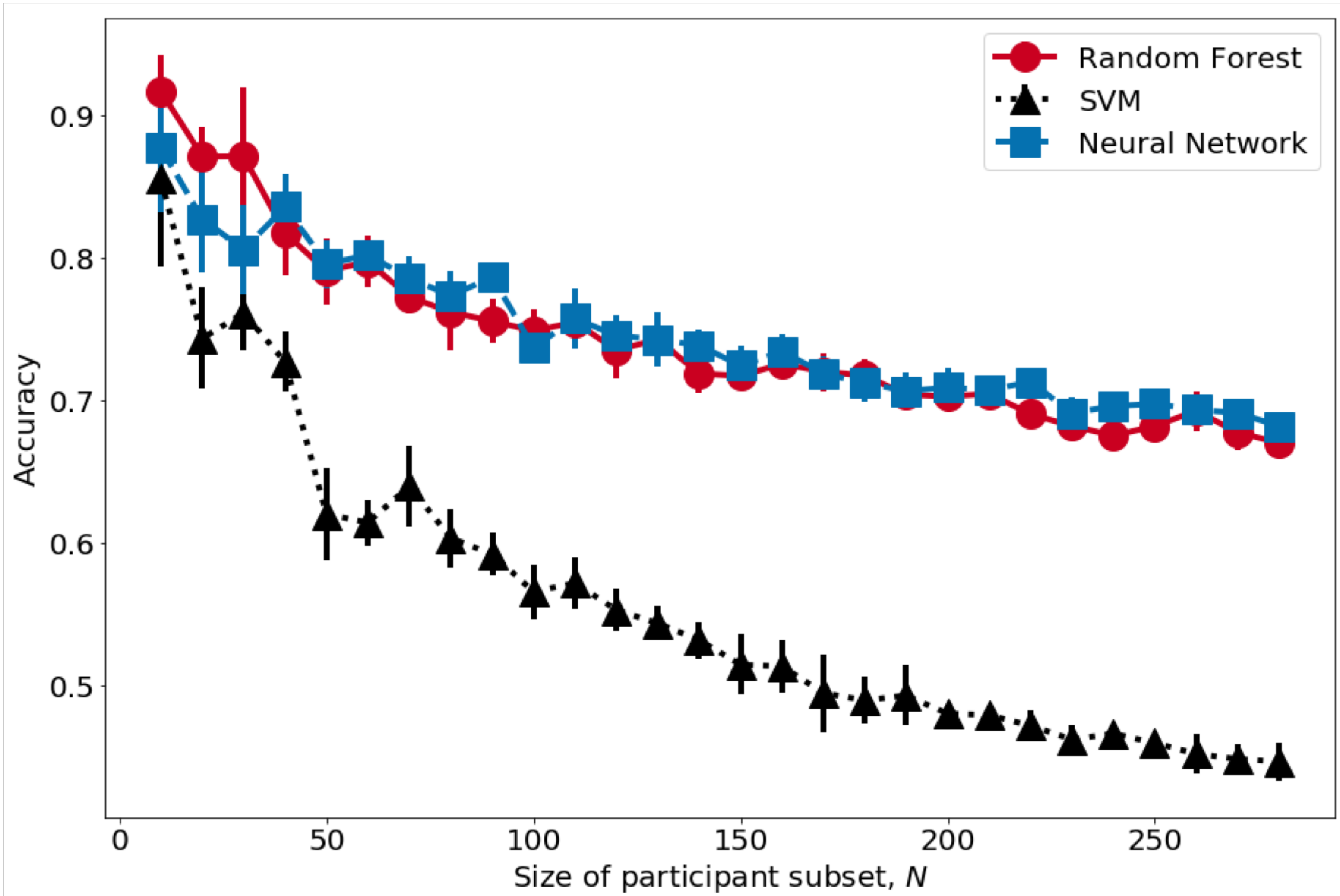
Each system fails at a different point



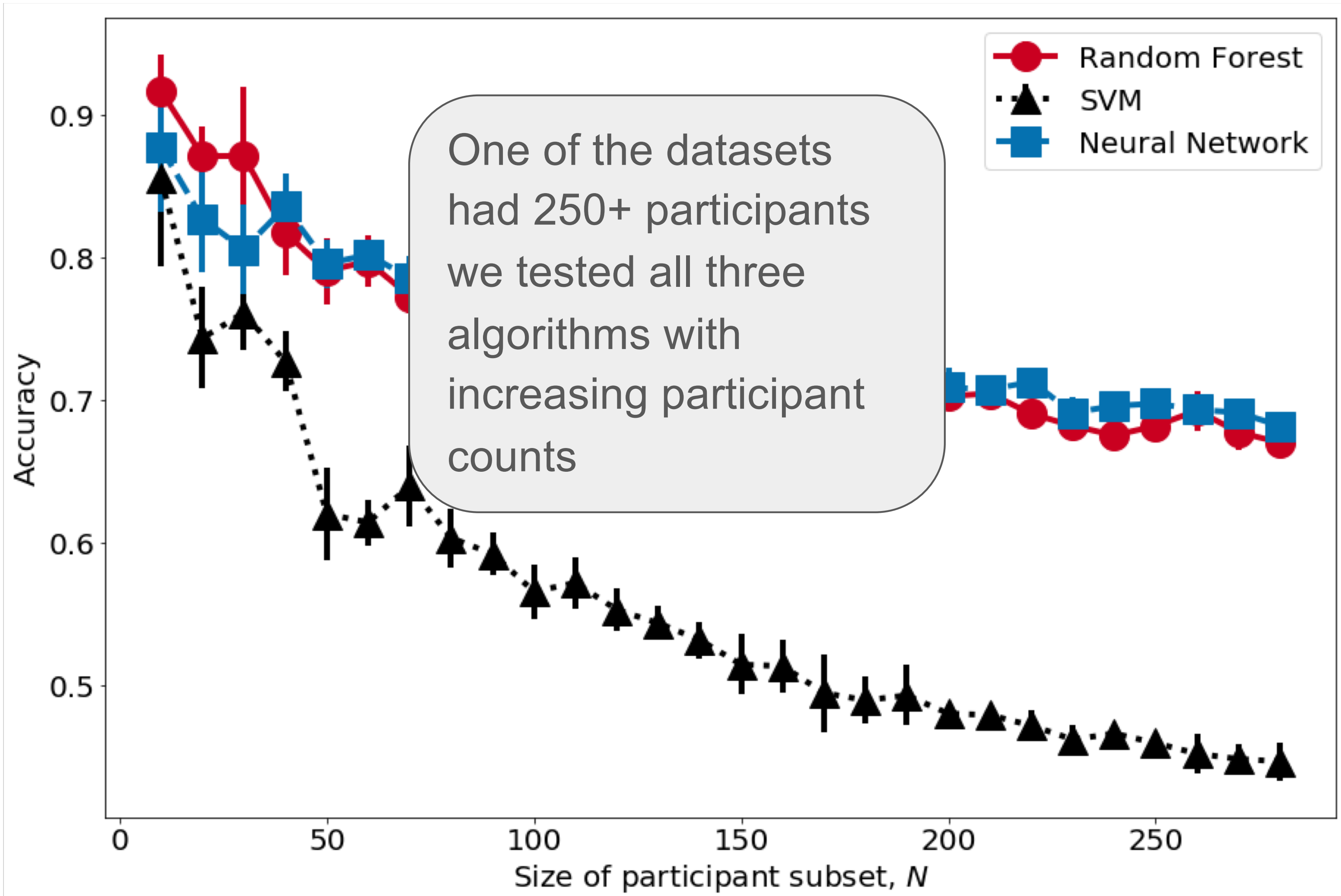
What To Do?

- Clearly, 20 is not a good participant count although often used.
- Unfortunately, there is no correct number.
- Power analysis does not work.

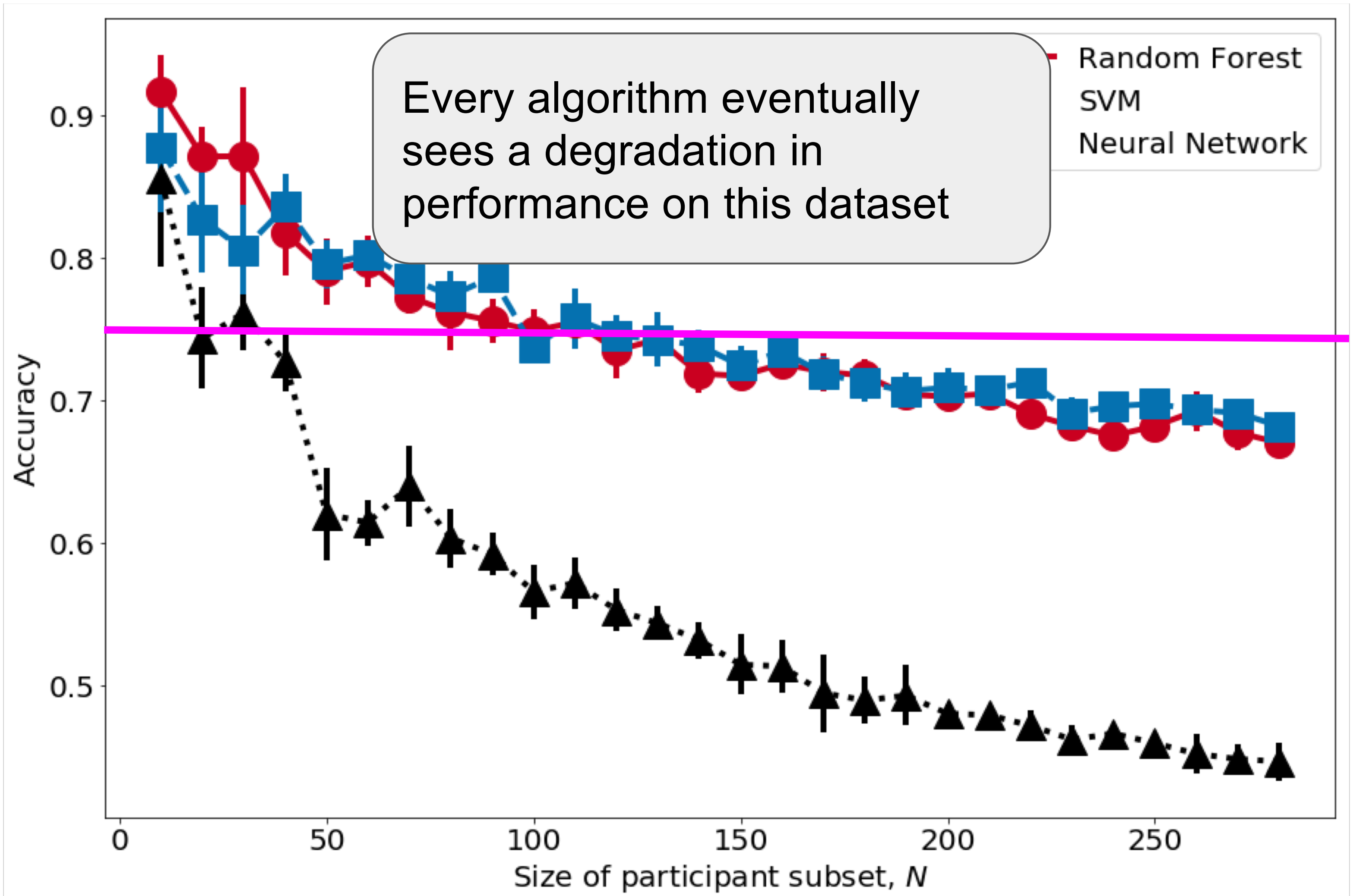
We Propose: Recruit Until It Fails



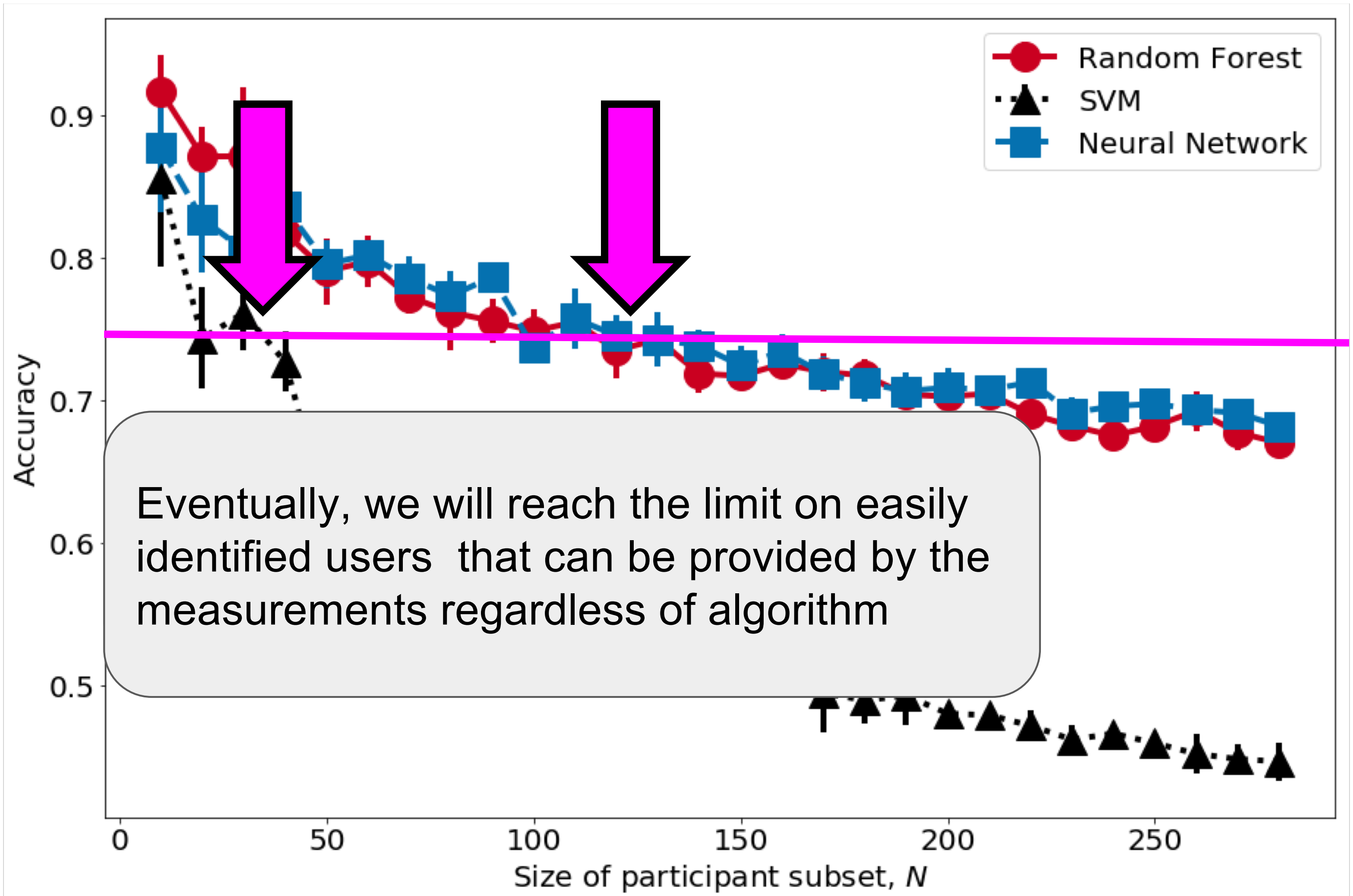
We Propose: Recruit Until It Fails



Recruit until it fails

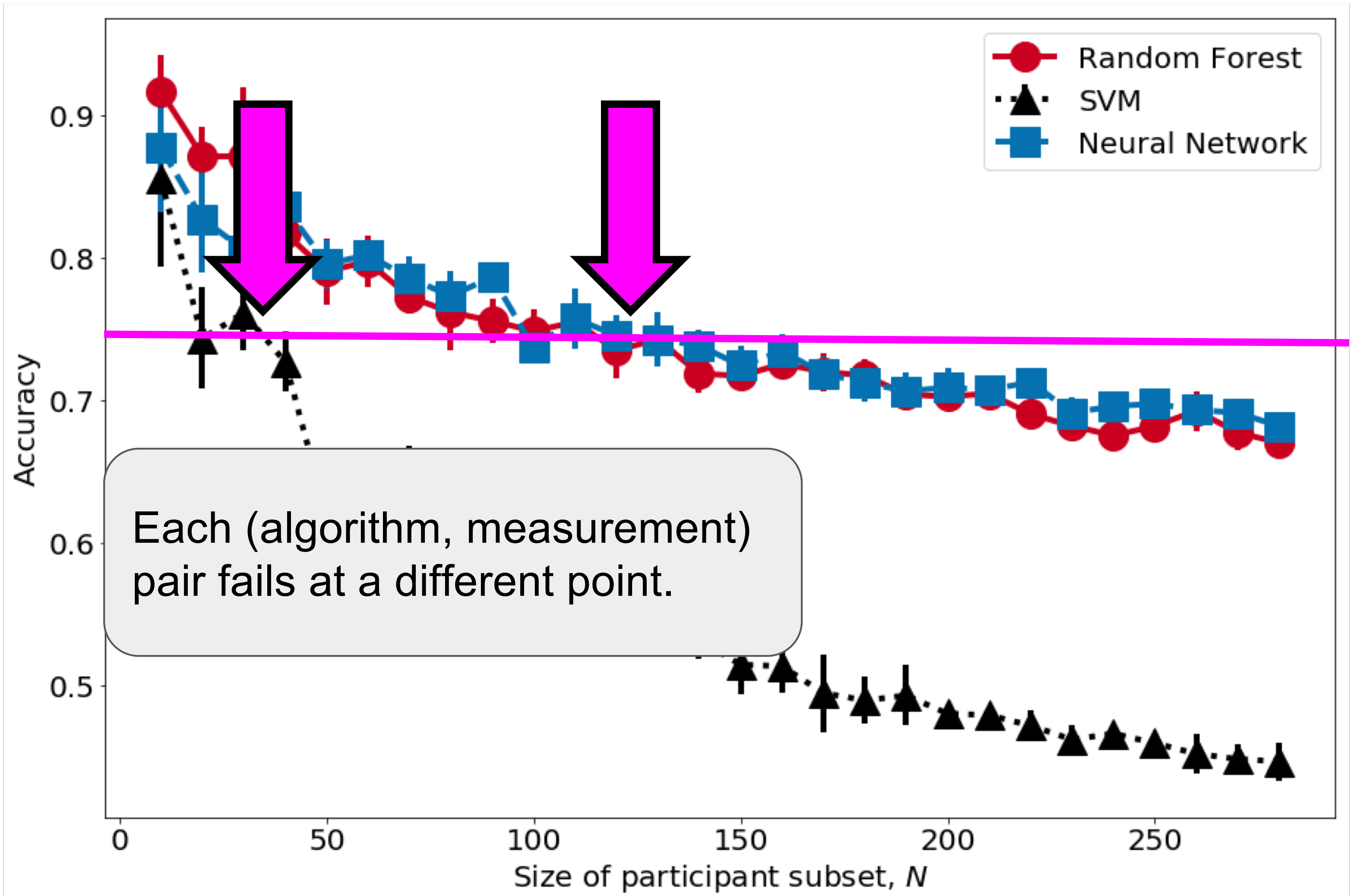


We Propose: Recruit Until It Fails



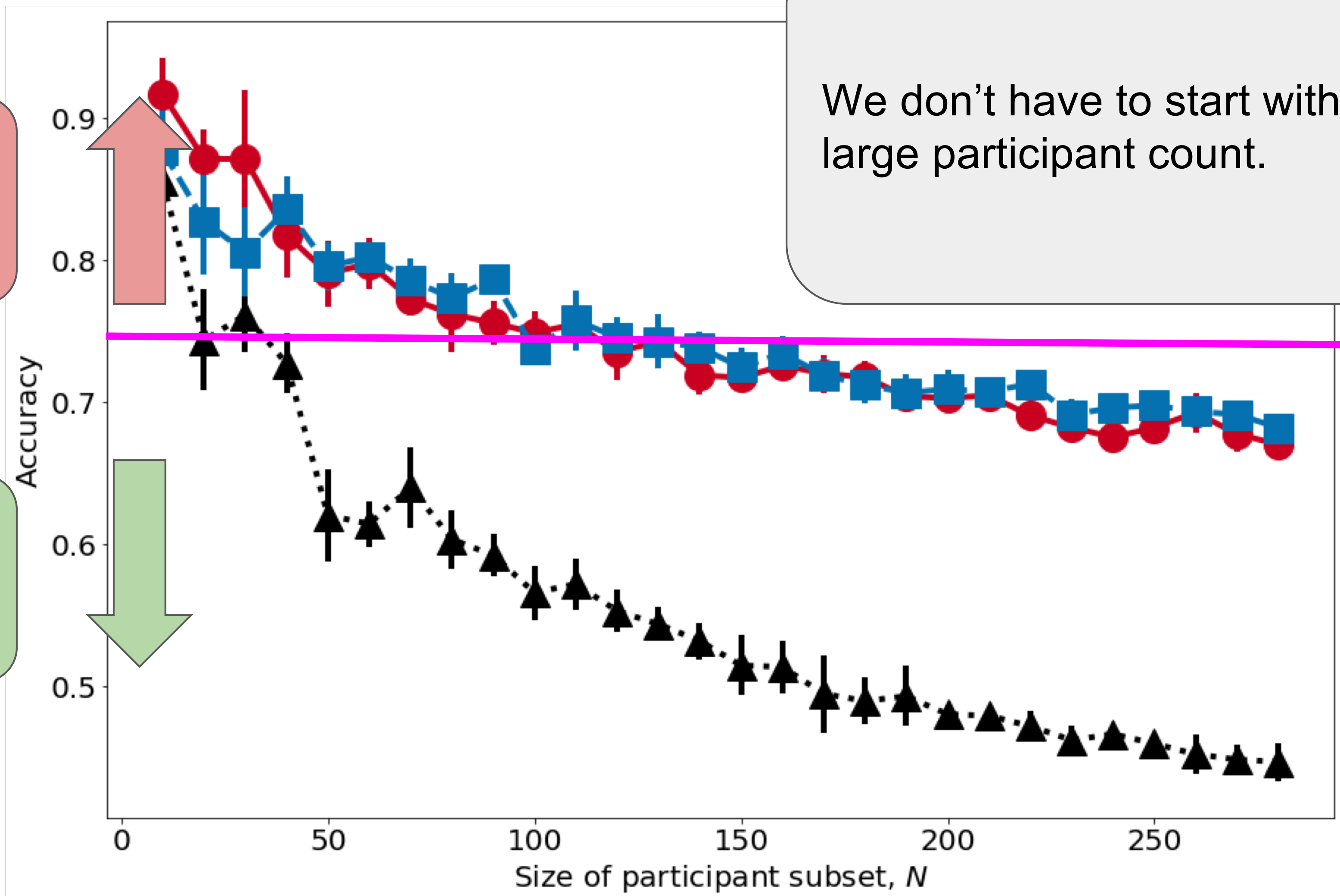
Eventually, we will reach the limit on easily identified users that can be provided by the measurements regardless of algorithm

We Propose: Recruit Until It Fails



We Propose: Recruit Until It Fails

Keep Testing



Good enough

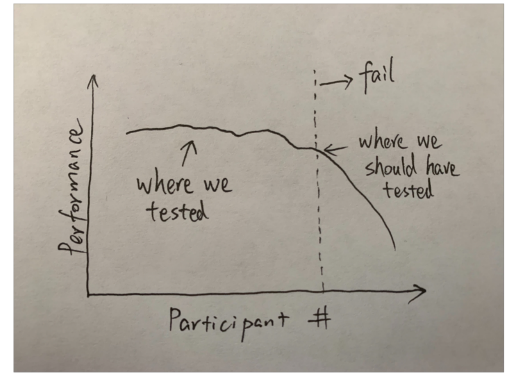
We don’t have to start with a large participant count.

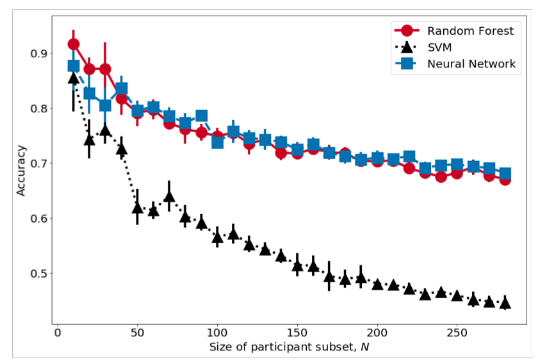
We Propose: Recruit Until It Fails

www.lindqvistlab.org

Summary: Recruit until it fails!

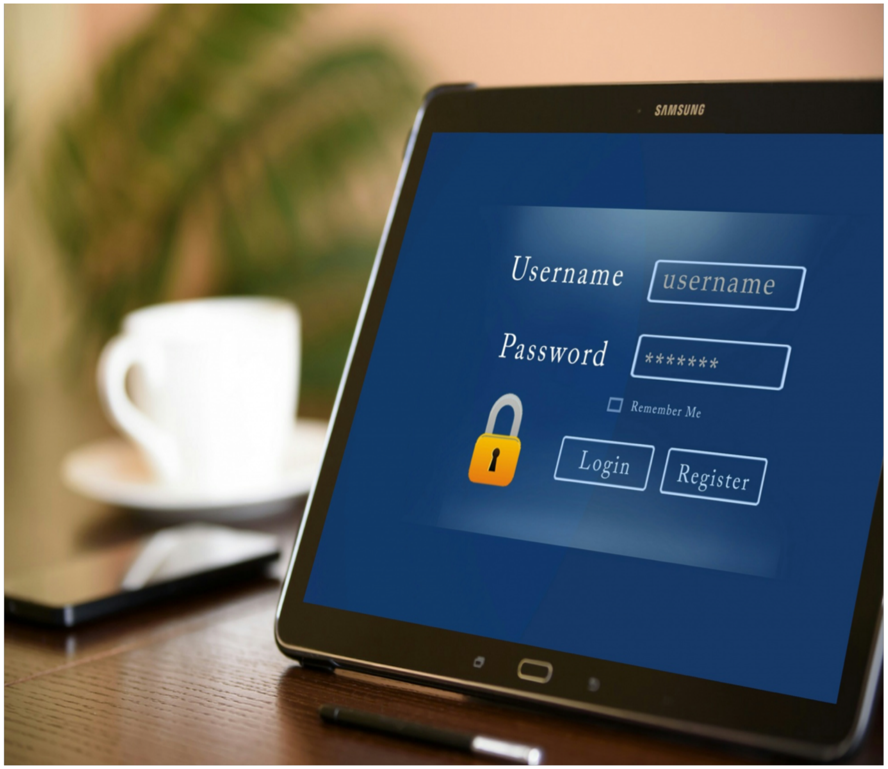
- We show with 5 identification systems
  - Why small participant pools are inadequate
  - Upper limits on easily identified participants
- **New approach to participant recruitment: recruit until it fails**



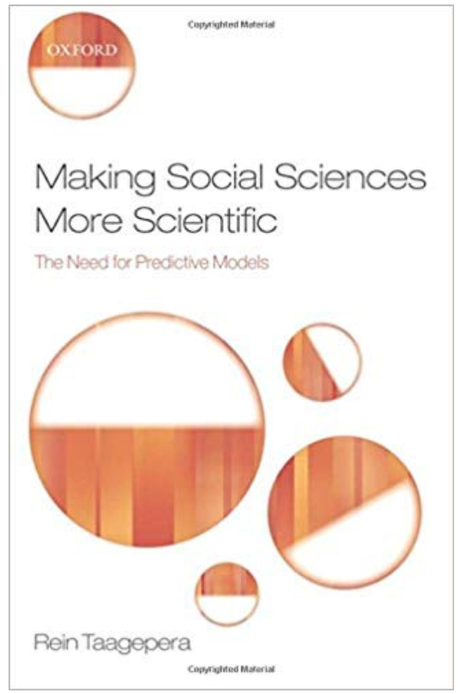


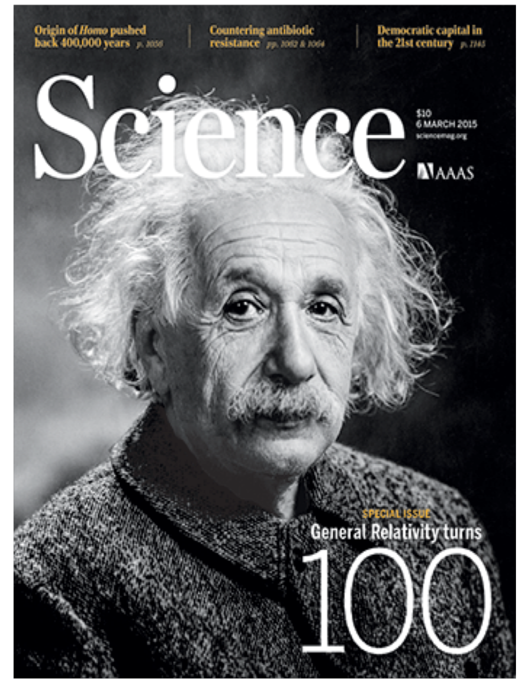

And now
for something
completely different...

“Systems Security”



Not A New Idea: We Need Predictive Models





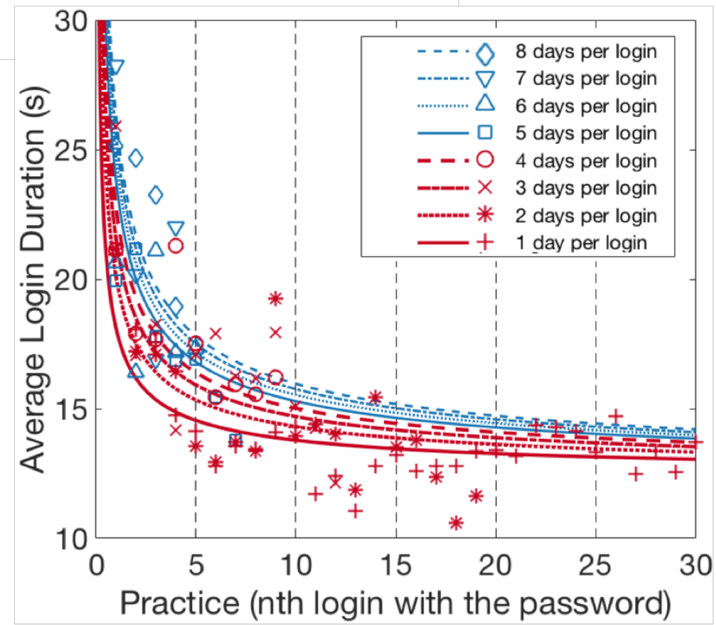
You Are Welcome

Forgetting of Passwords: Ecological Theory and Data

Xianyi Gao[†], Yulong Yang[†], Can Liu[†], Christos Mitropoulos[†], Janne Lindqvist[†], Antti Oulasvirta[*]
[†]*Rutgers University,* [*]*Aalto University*

$$E[Time_{login}] \approx \frac{Kf^{d}(1-d)}{n^{1-d}} + E[Time_{act}]$$

$$R_o \approx e^{-\tau/s+C/s}(1-d)^{-1/s}f^{-d/s}n^{(1-d)/s}$$



Big Picture: What Is Common Between Medicine, Interventional Behavioral Sciences and Security?

PLOS | MEDICINE   BROWSE   PUBLISH

Common: Is your cure, treatment or system effective (and better than before)

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • https://doi.org/10.1371/journal.pmed.0020124

Thank You!

This material is based upon work supported by the National Science Foundation under Grant Number 1750987. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Please visit our websites for more details:

lindqvistlab.org
scienceofsecurity.science