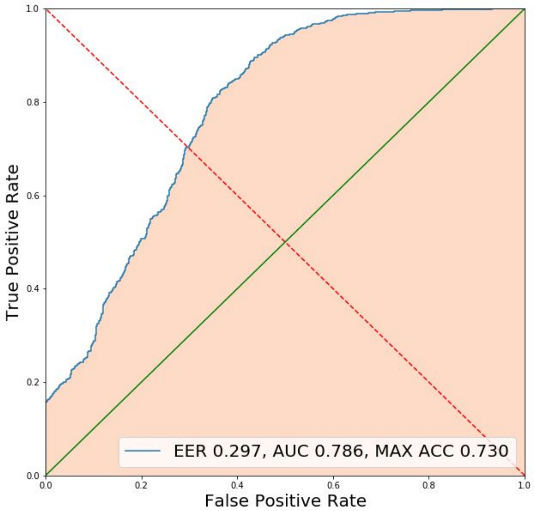
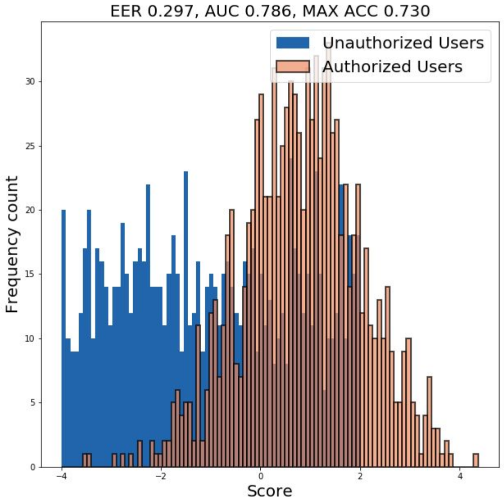
RUTGERS
UNIVERSITY | NEW BRUNSWICK

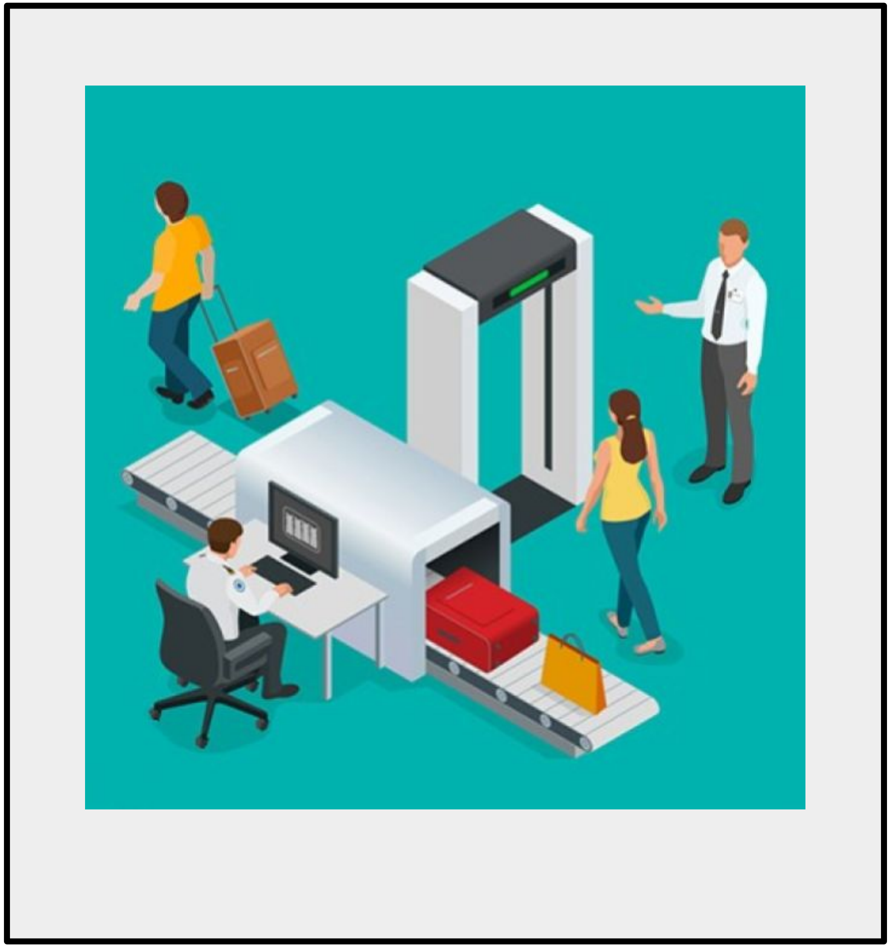
Robust Performance Metrics for Authentication Systems

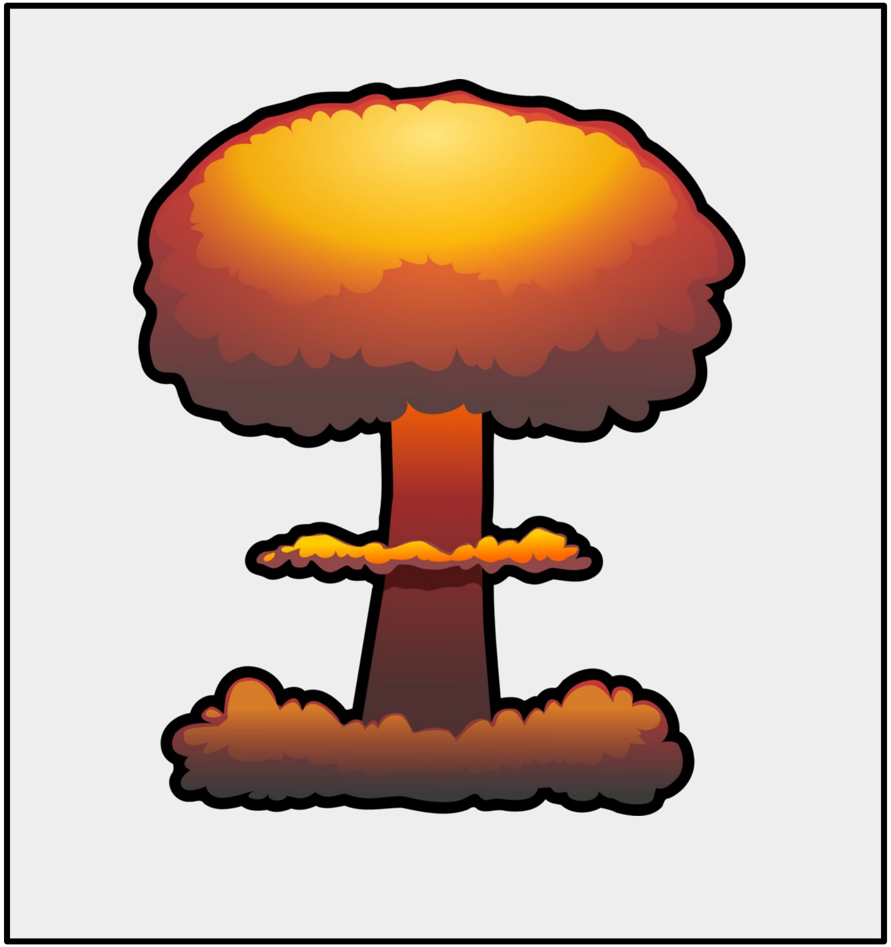
Shridatt Sugrim, Can Liu, Meghan C. McLean, Janne Lindqvist
Rutgers University



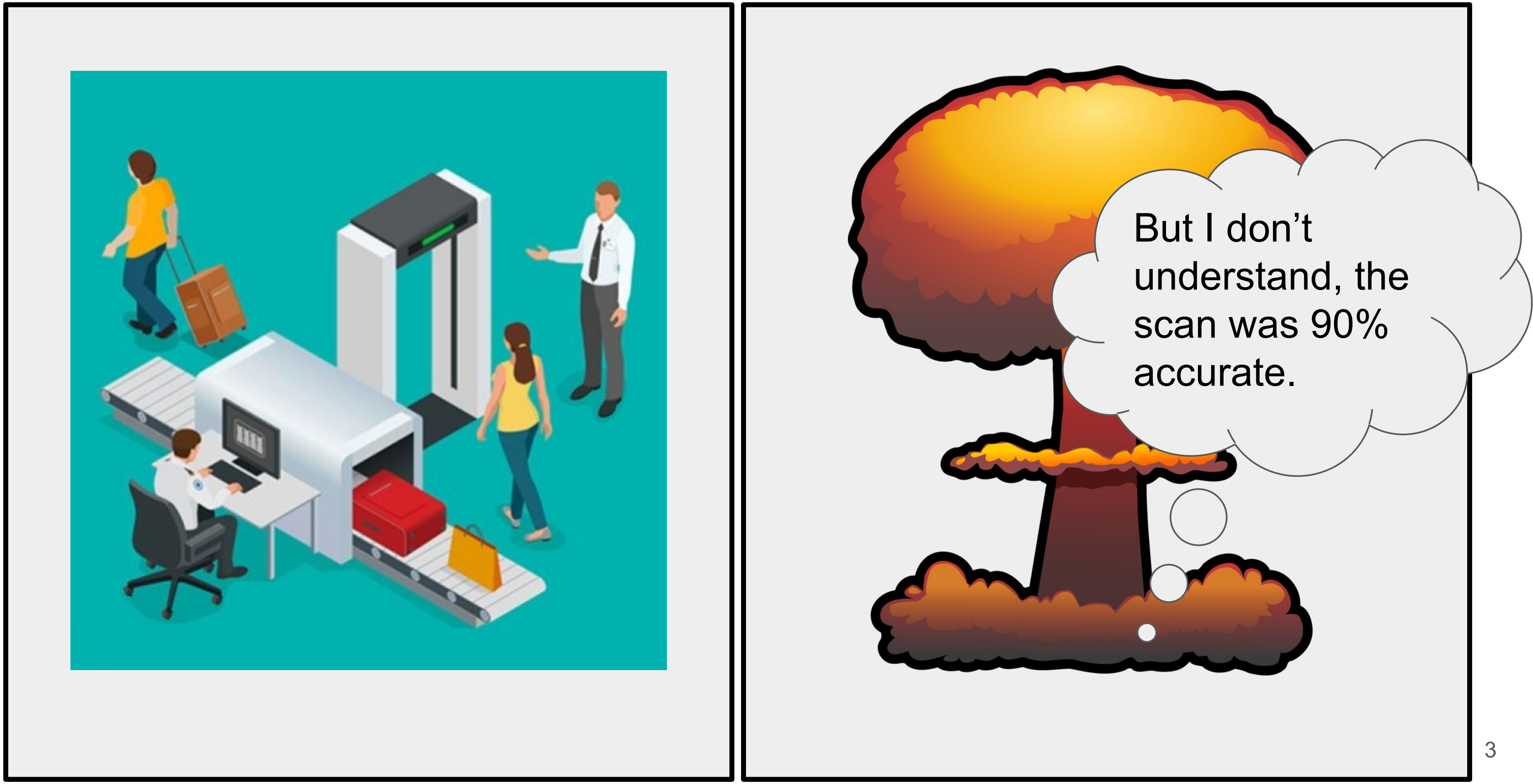


TL; DR:
Performance reporting might surprise you

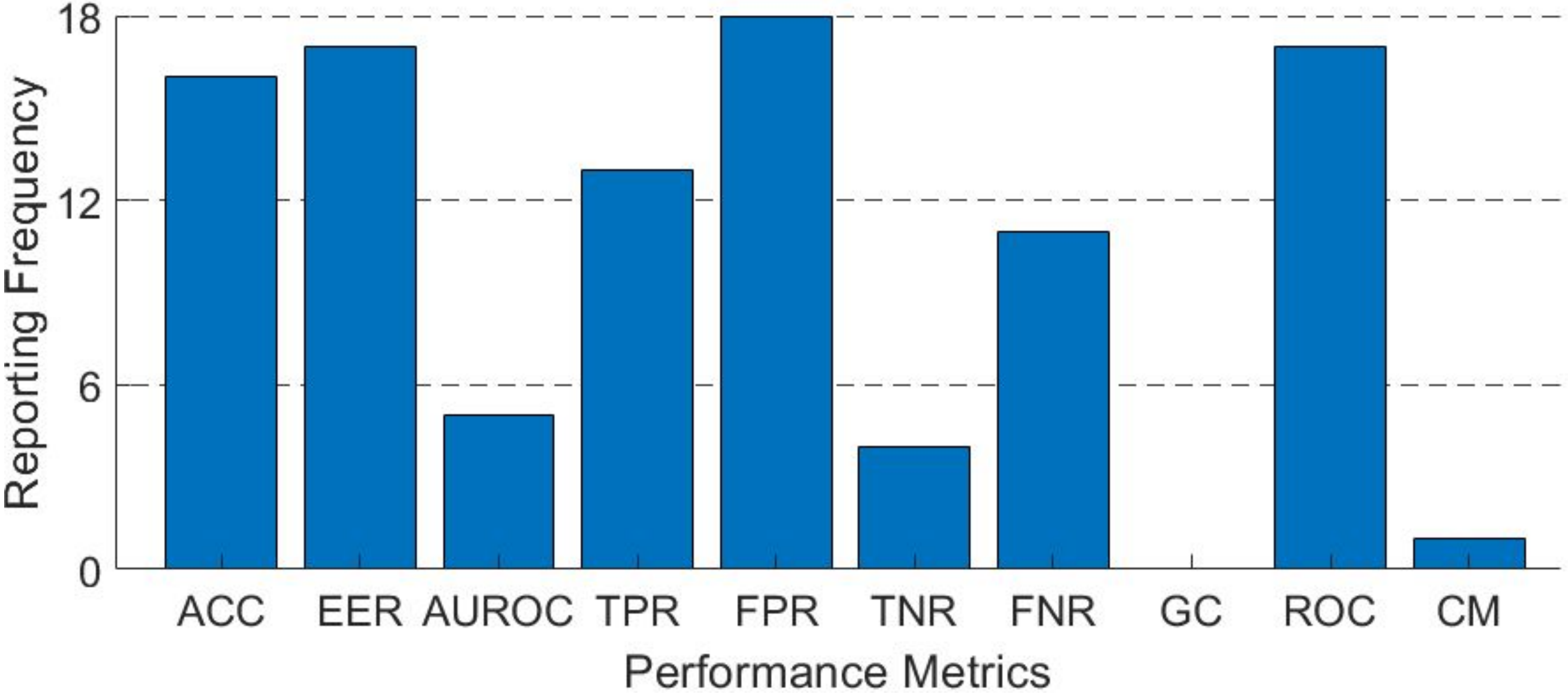




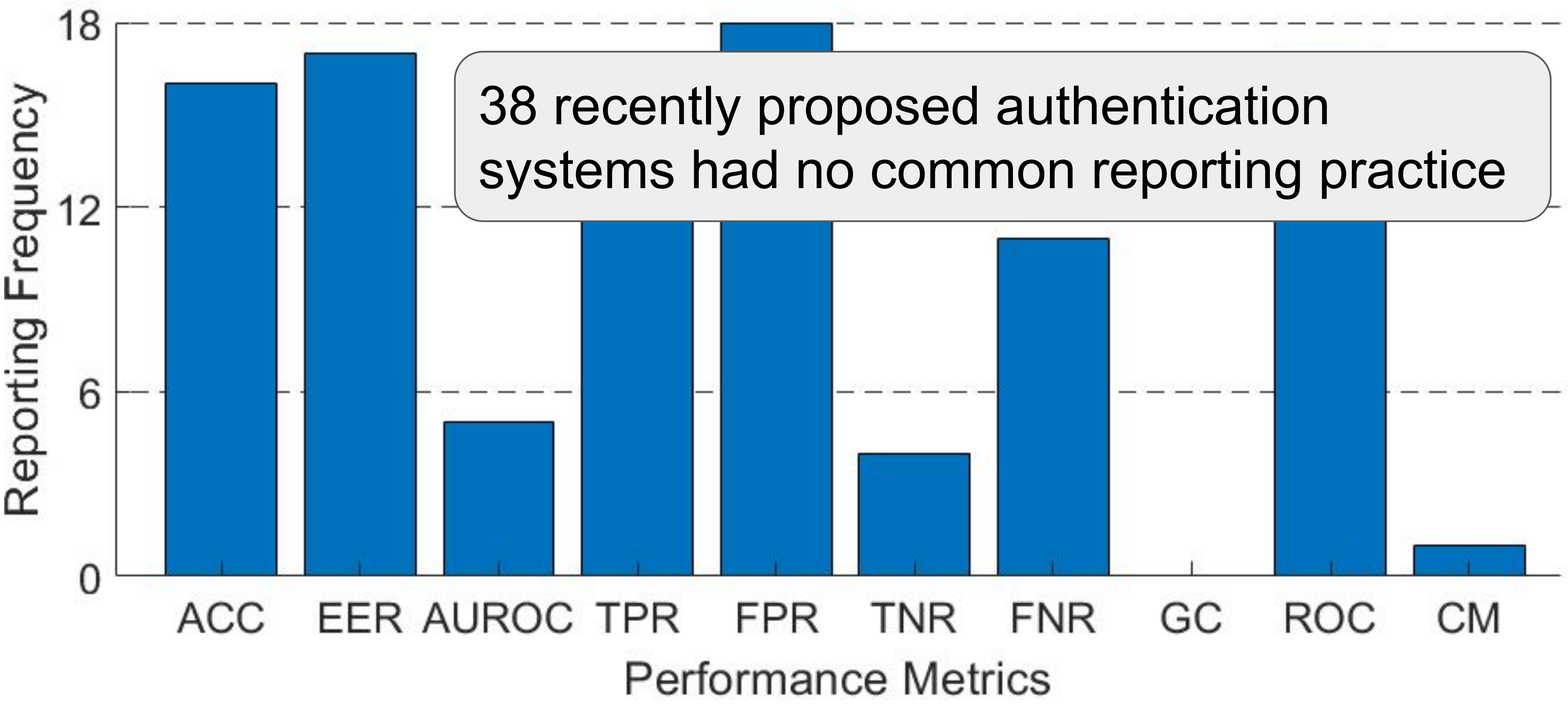
TL; DR:
Performance reporting might surprise you



Why is evaluation of authentication systems hard?

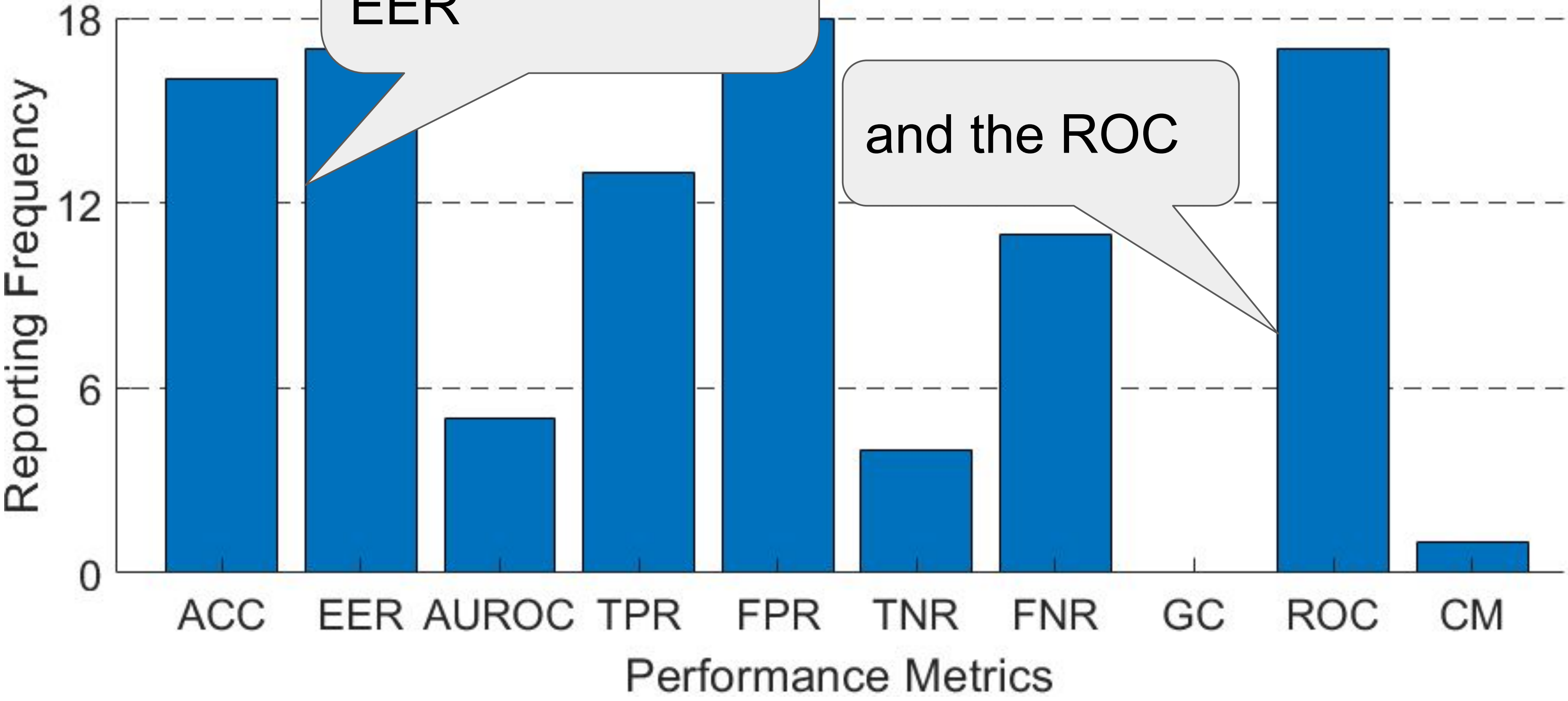


Why is evaluation of authentication systems hard?



38 recently proposed authentication systems had no common reporting practice

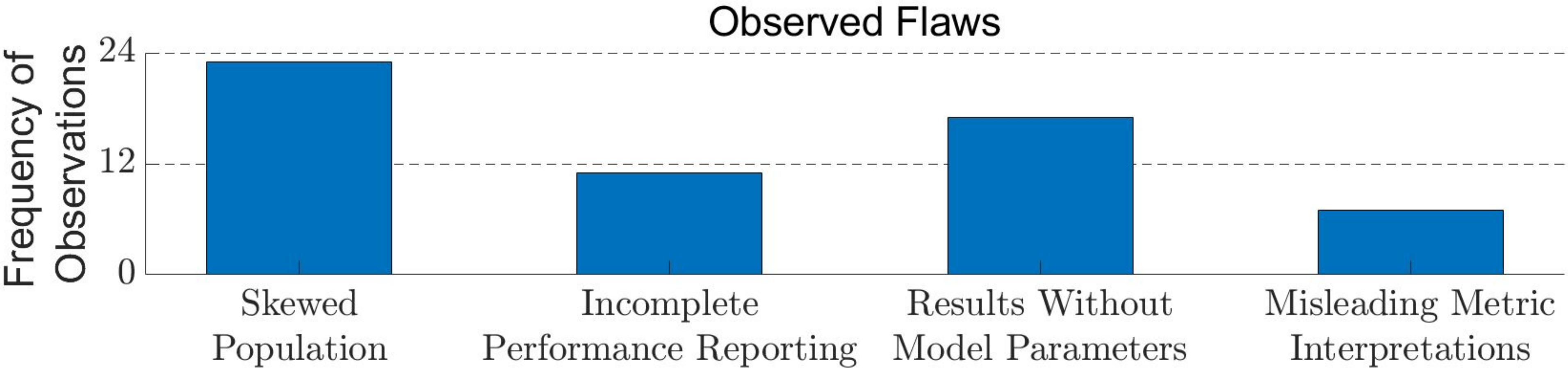
Why is ev tication systems hard?



We are going to focus on ACC and EER

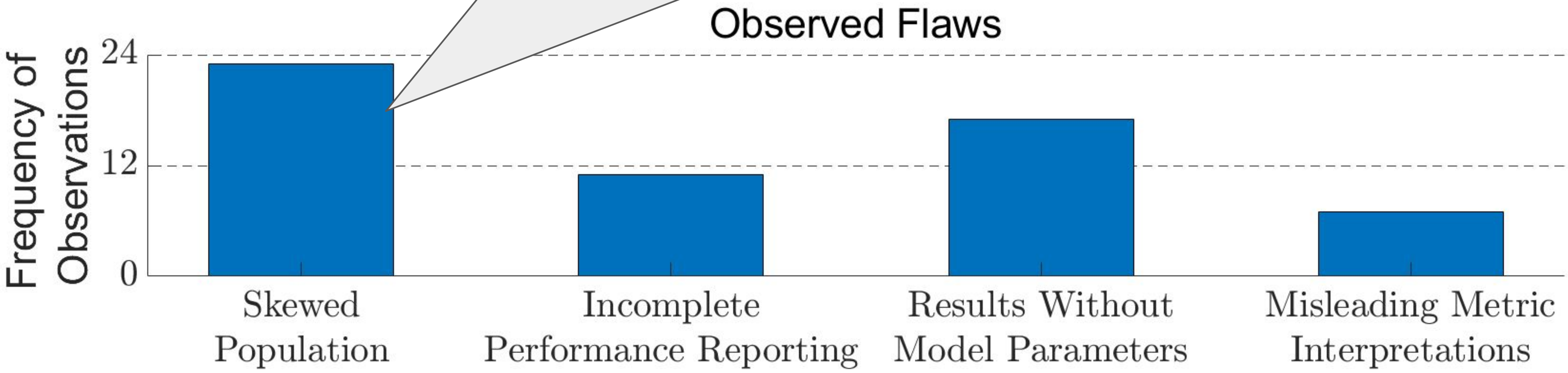
and the ROC

Most (36 of 38) of the performance reporting had some flaw



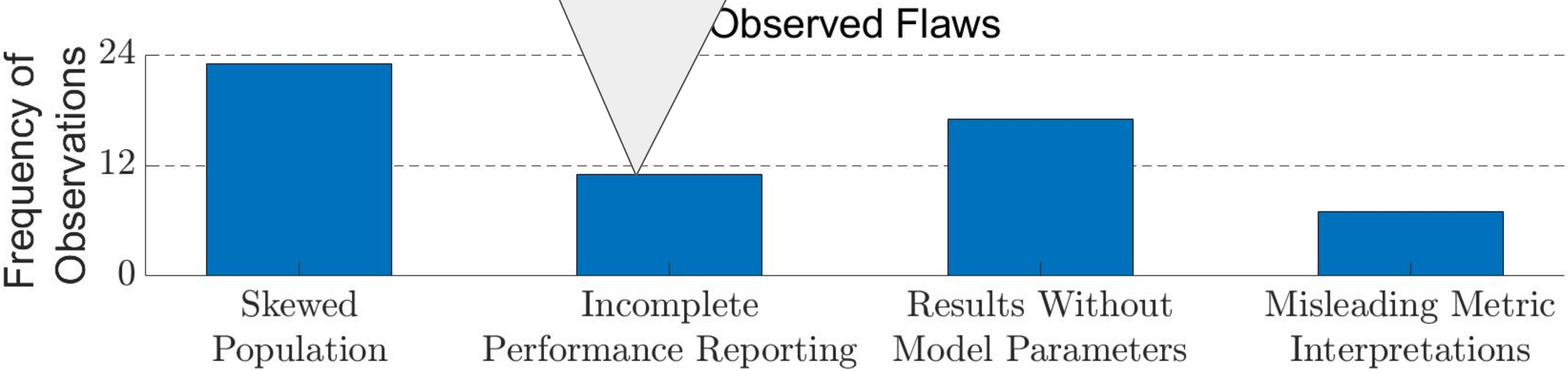
Most (36 of 38) of the performance reporting had some flaw

Inflate performance metric values



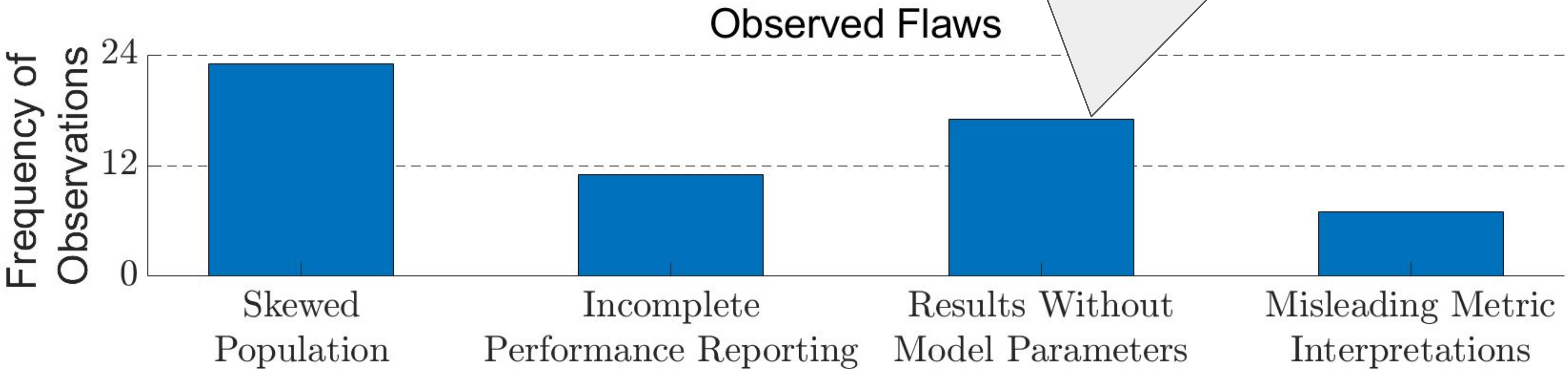
Most (36 of 38) of the performance reporting had some flaw

Inadequate for system comparison

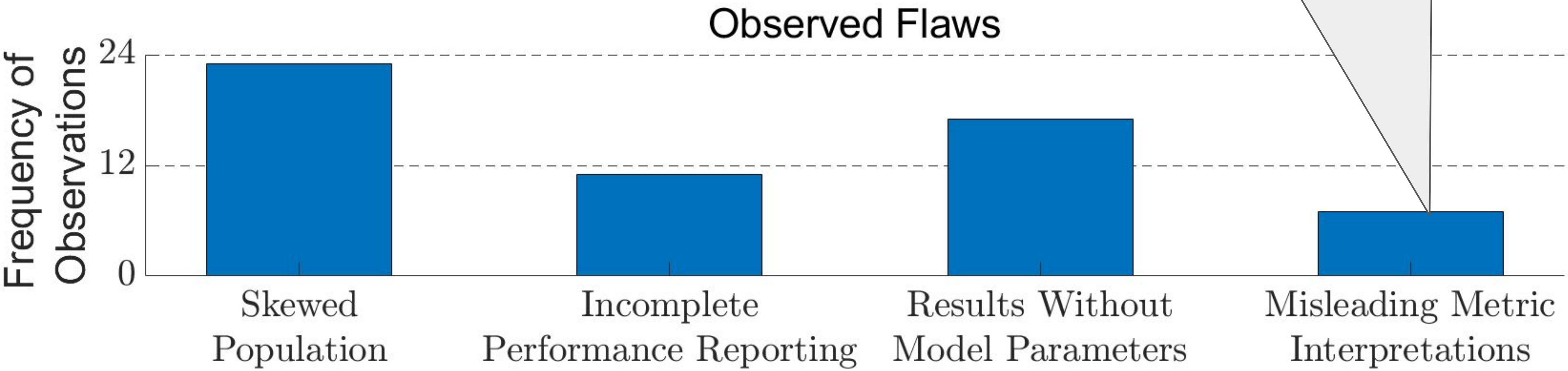


Most (36 of 38) of the performance reporting had some flaw

Prevents reproducing results

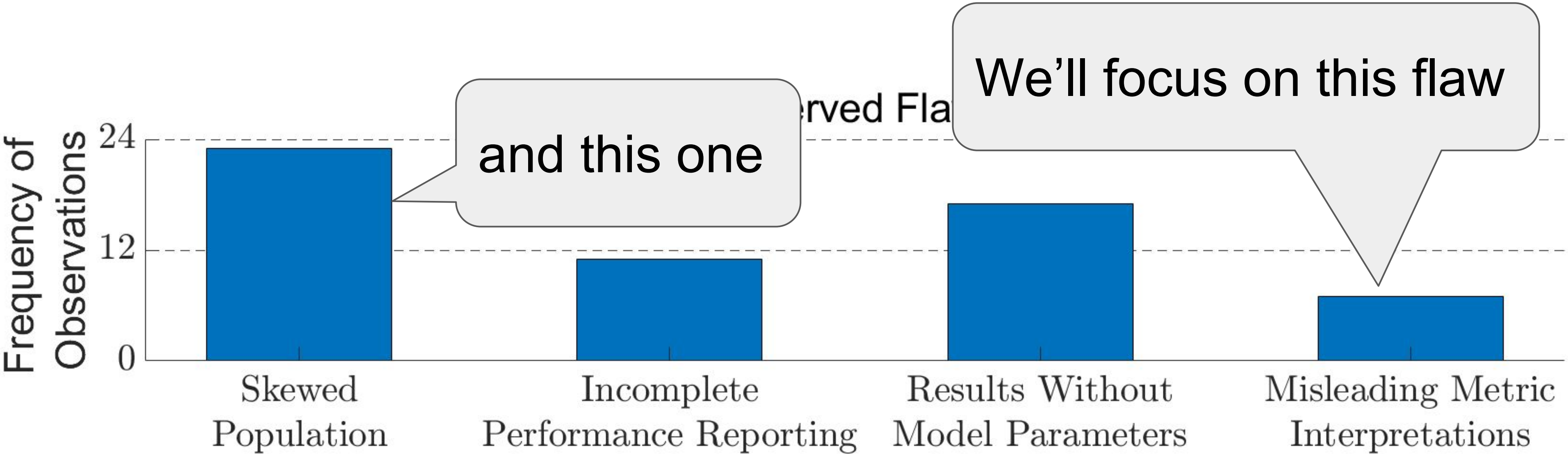


Most (36 of 38) of the performance reporting had some flaw



Unexpected failures

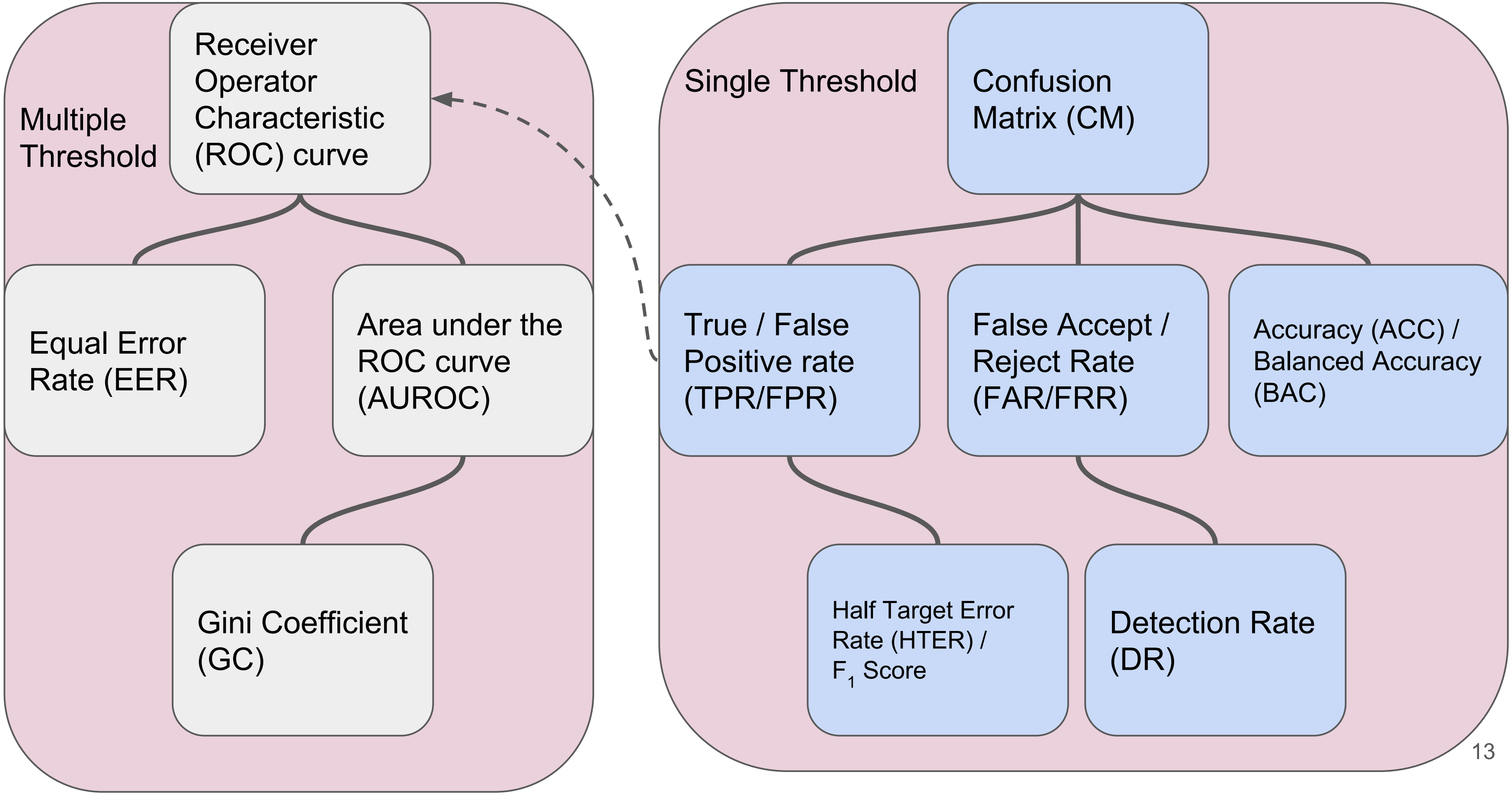
Most (36 of 38) of the performance reporting had some flaw



and this one

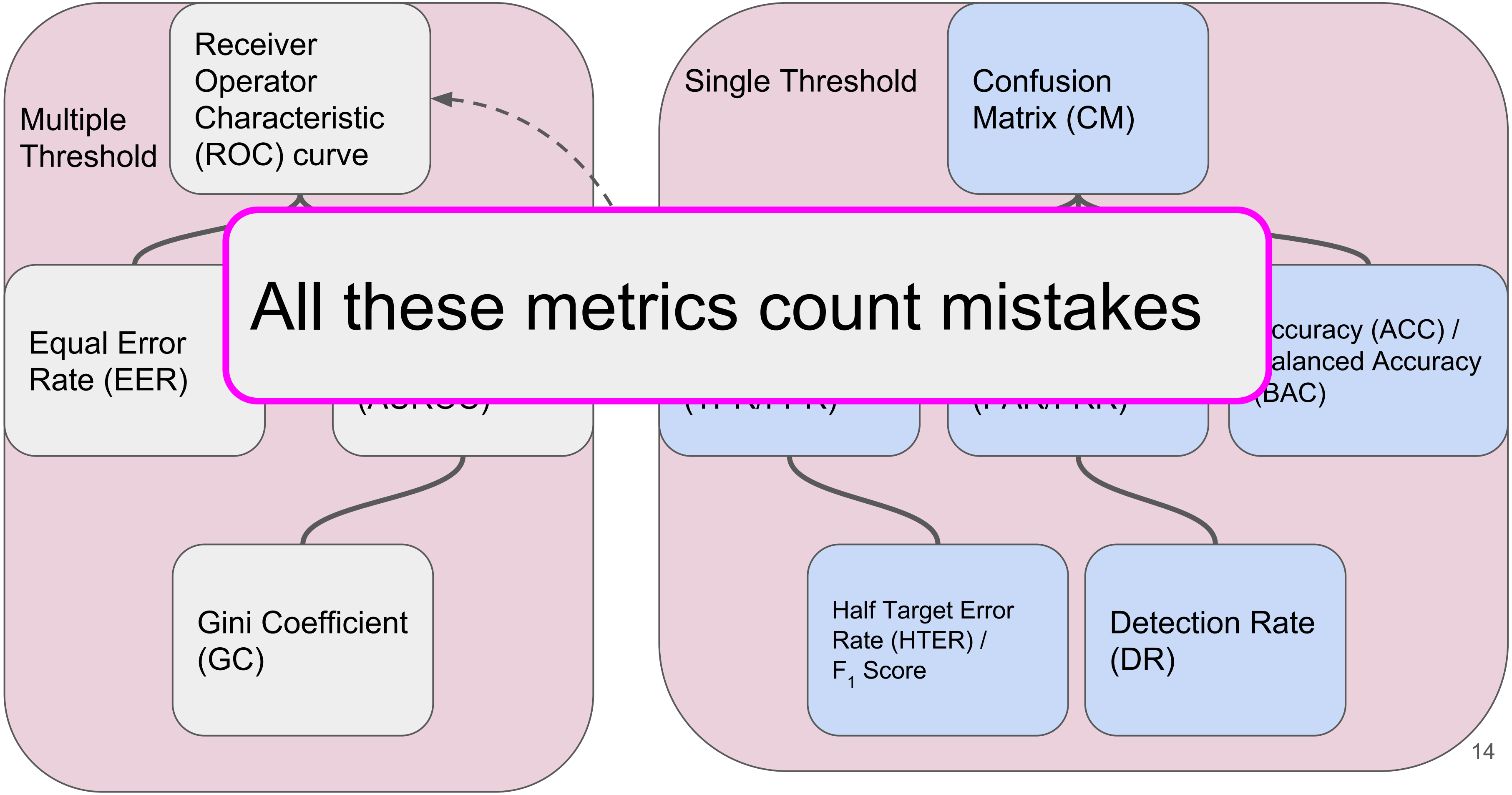
We'll focus on this flaw

Related metrics have similar properties



Multiple Threshold
Receiver Operator Characteristic (ROC) curve
Equal Error Rate (EER)
Area under the ROC curve (AUROC)
Gini Coefficient (GC)
Single Threshold
Confusion Matrix (CM)
True / False Positive rate (TPR/FPR)
False Accept / Reject Rate (FAR/FRR)
Accuracy (ACC) / Balanced Accuracy (BAC)
Half Target Error Rate (HTER) / F1 Score
Detection Rate (DR)

Related metrics have similar properties.



Multiple Threshold

Receiver Operator Characteristic (ROC) curve

Equal Error Rate (EER)

(AUROC)

Gini Coefficient (GC)

Single Threshold

Confusion Matrix (CM)

(TPR/FPR)

(FAR/FRR)

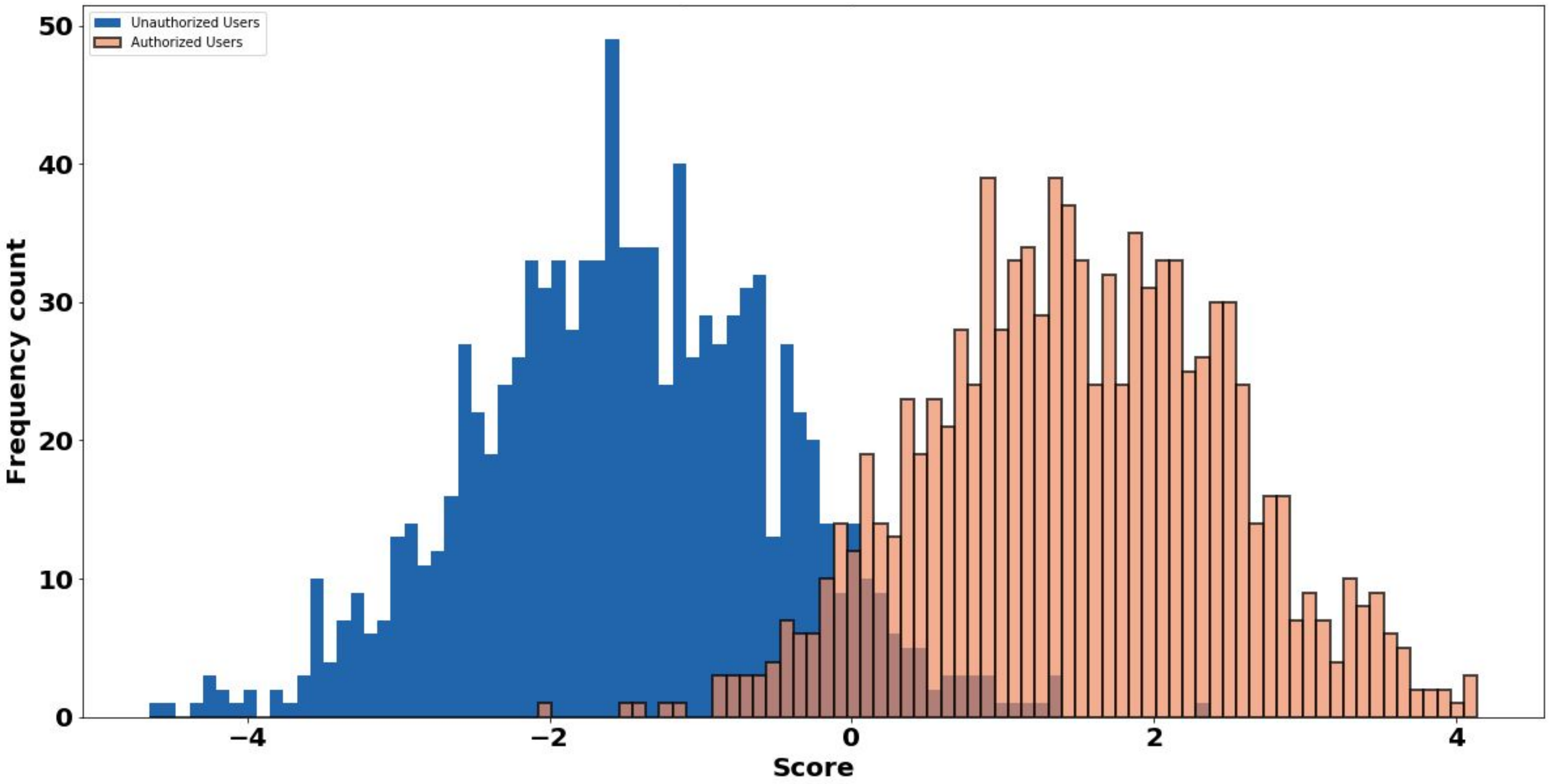
ccuracy (ACC) / alanced Accuracy (BAC)

Half Target Error Rate (HTER) / F_1 Score

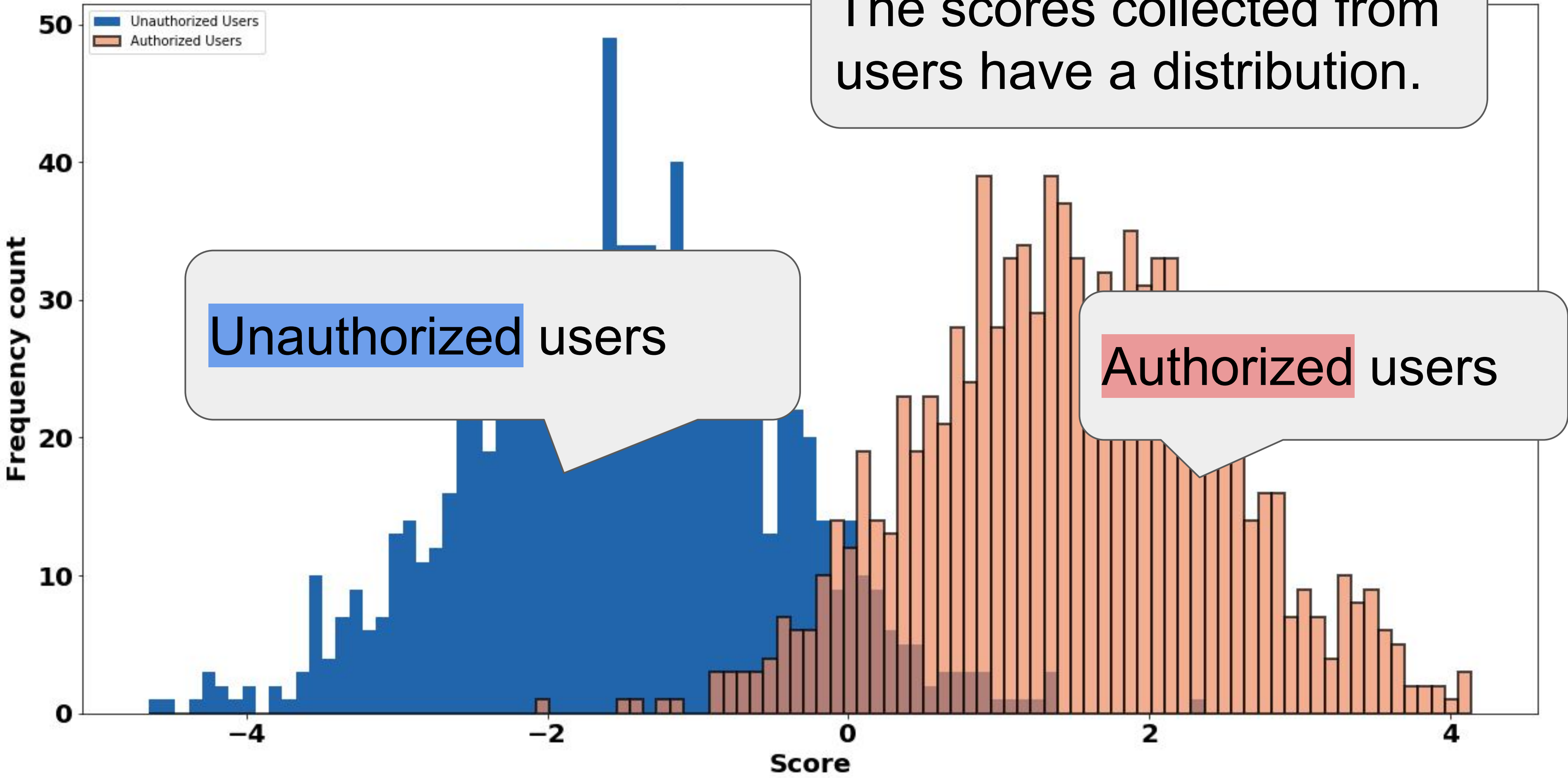
Detection Rate (DR)

All these metrics count mistakes

How authentication systems work



How authentication systems work

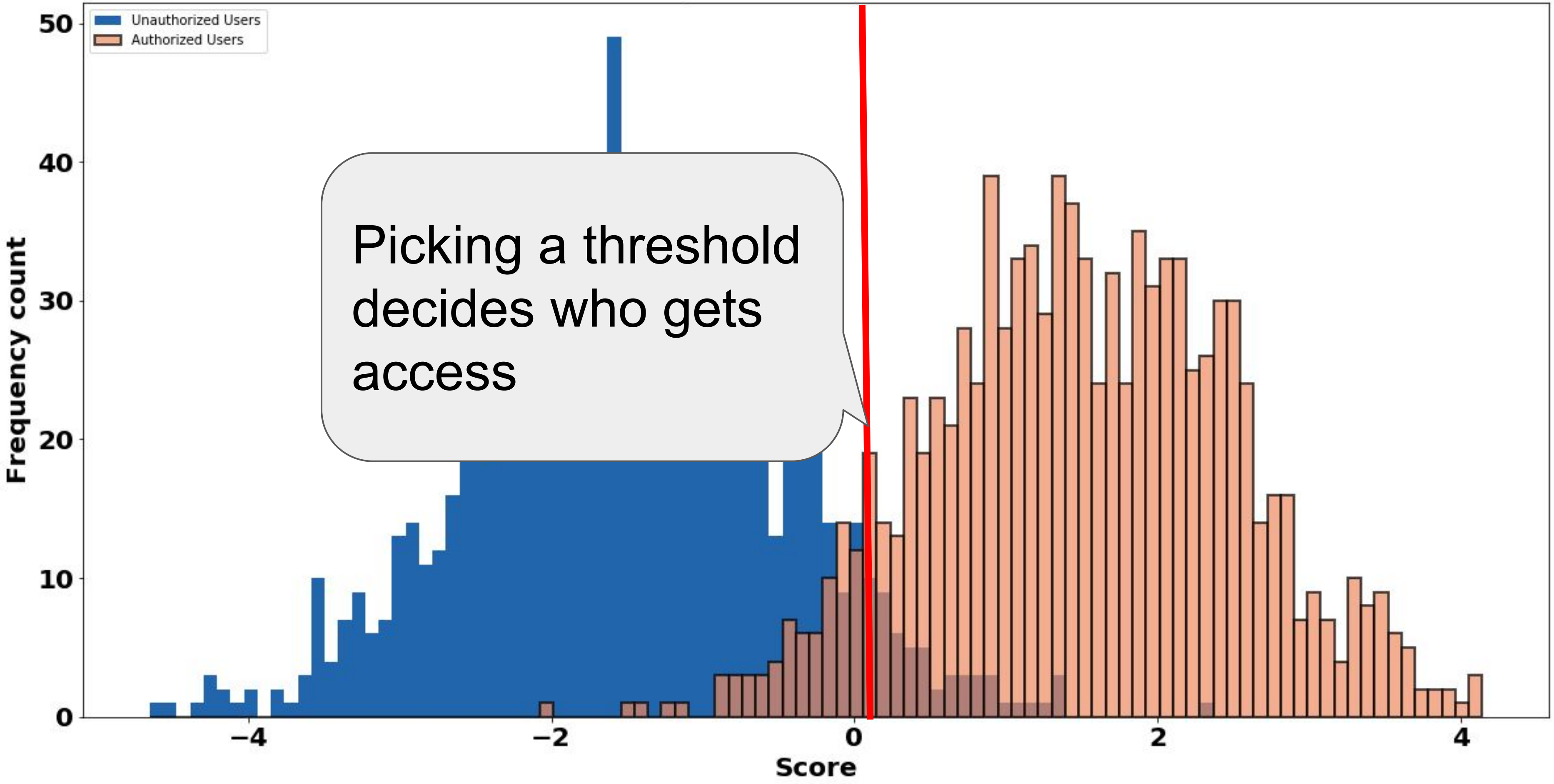


The scores collected from users have a distribution.

Unauthorized users

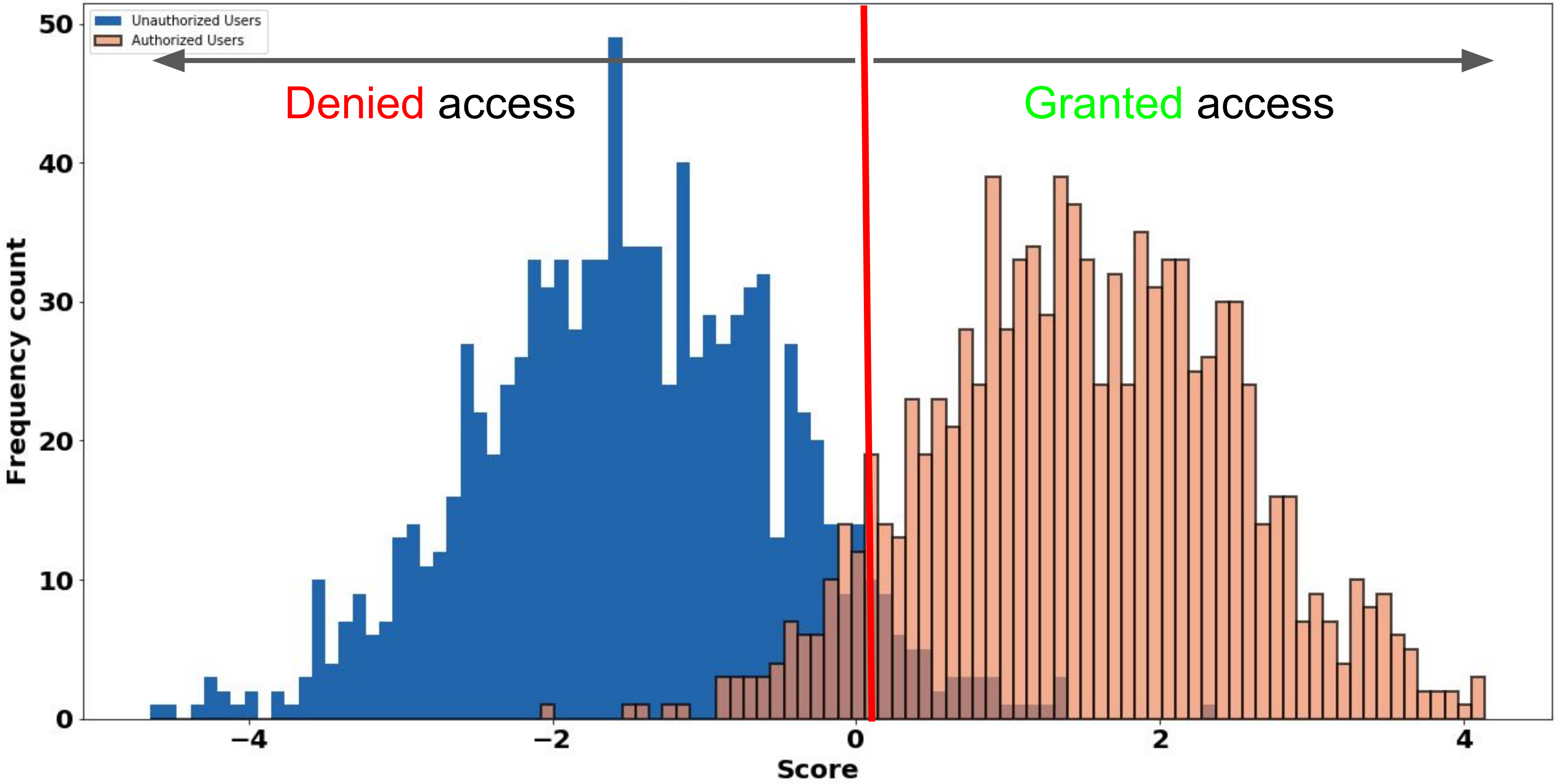
Authorized users

How authentication systems work

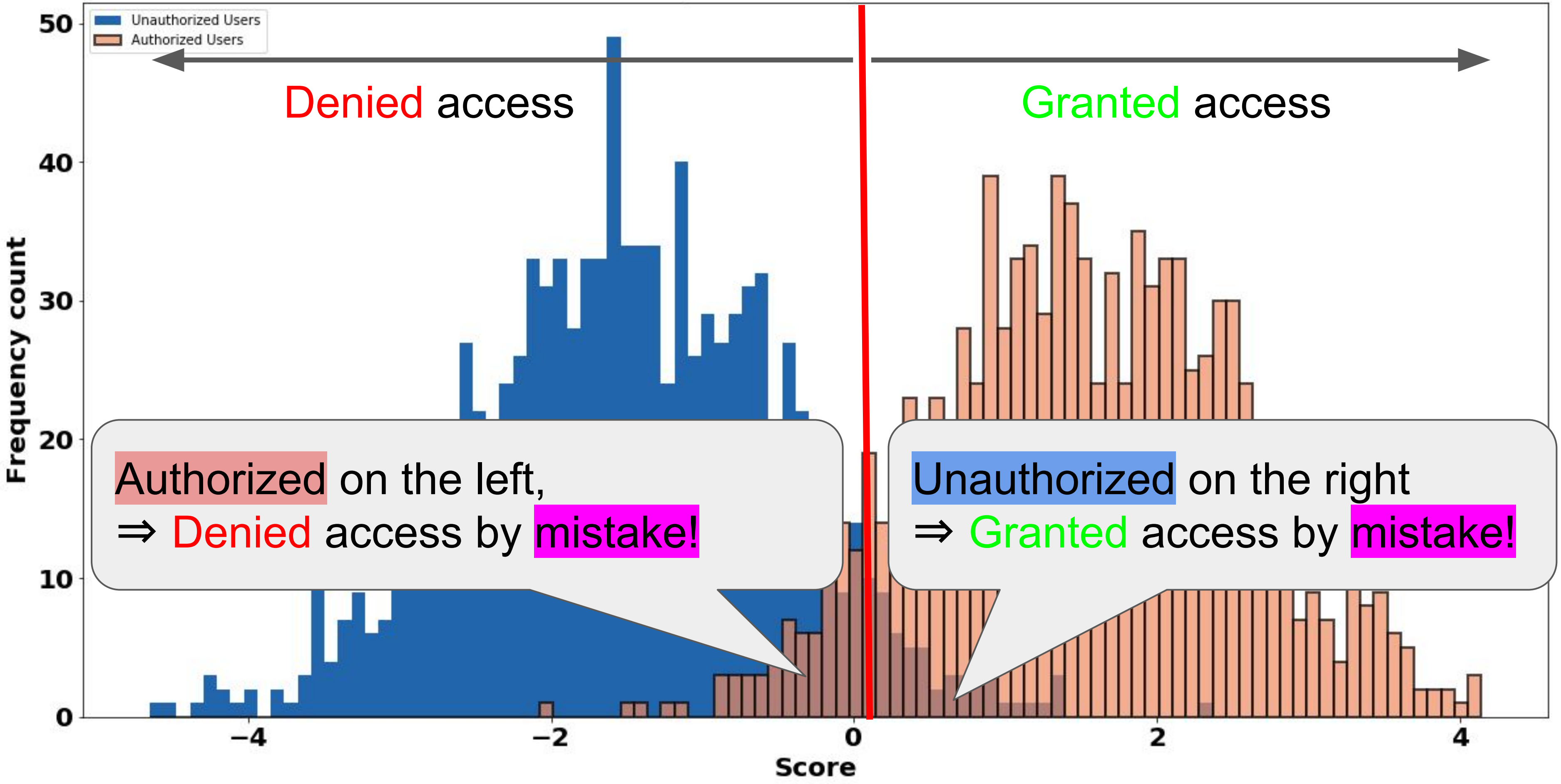


Picking a threshold decides who gets access

How authentication systems work



Where do authentication mistakes come from?



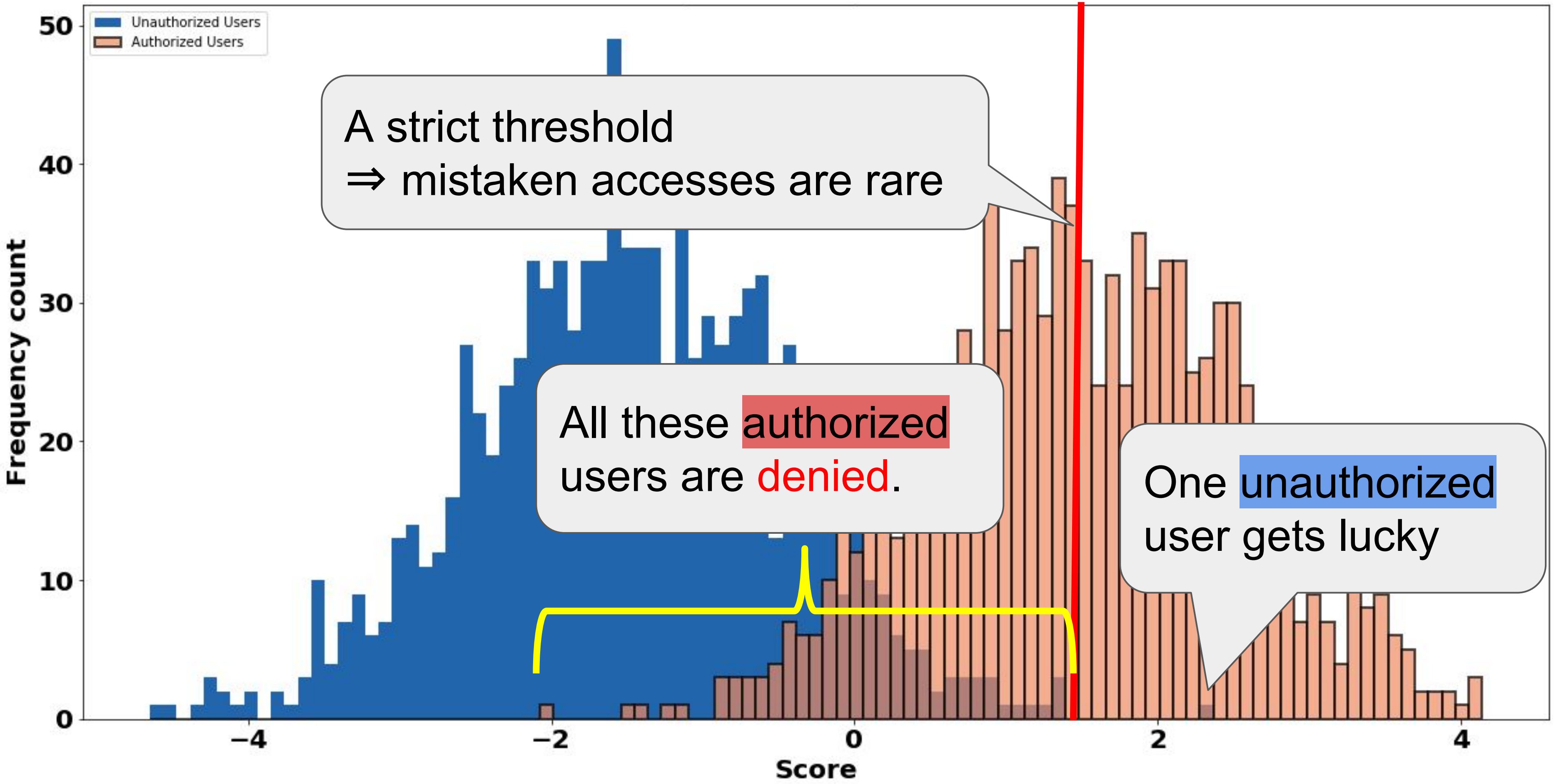
Denied access

Granted access

Authorized on the left,
⇒ Denied access by mistake!

Unauthorized on the right
⇒ Granted access by mistake!

Thresholds matter

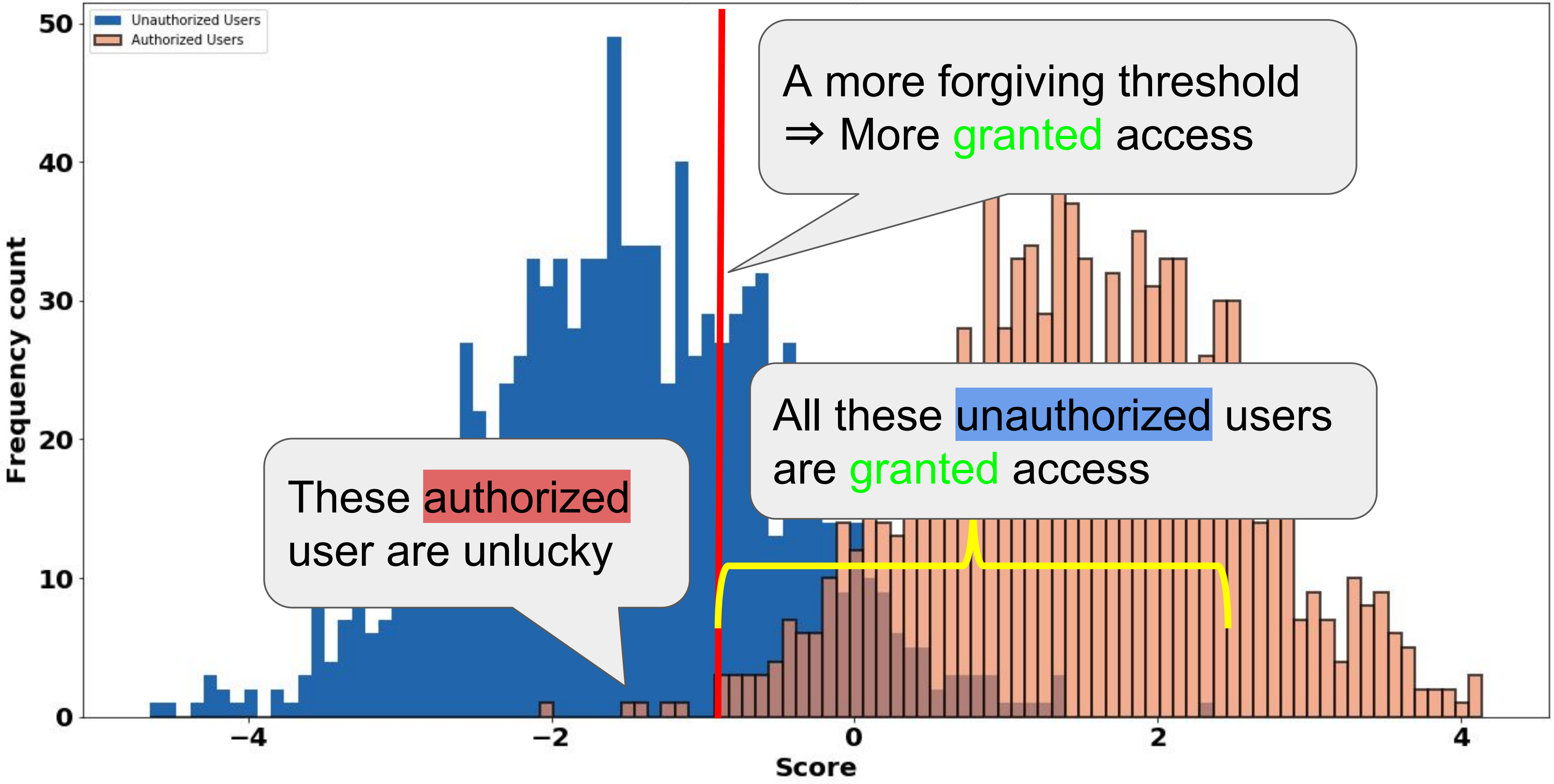


A strict threshold
⇒ mistaken accesses are rare

All these authorized users are denied.

One unauthorized user gets lucky

Thresholds matter

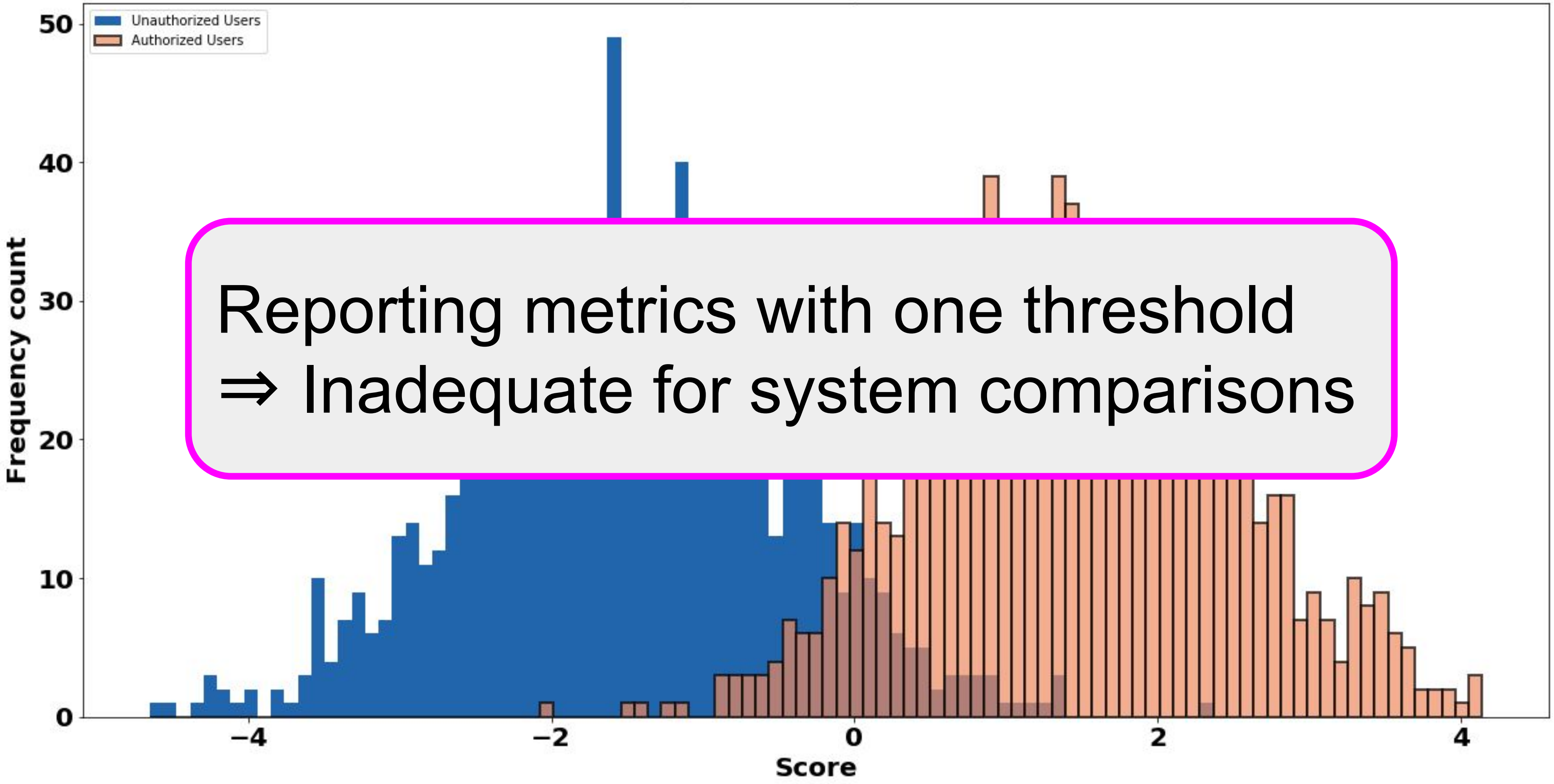


A more forgiving threshold
⇒ More granted access

All these unauthorized users are granted access

These authorized user are unlucky

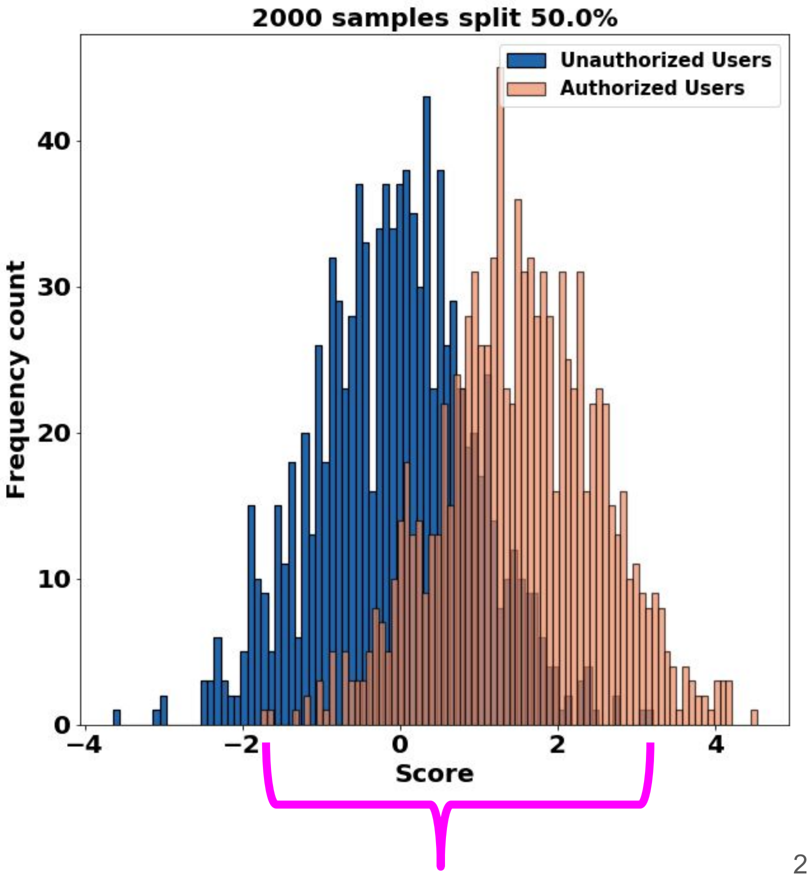
Thresholds matter



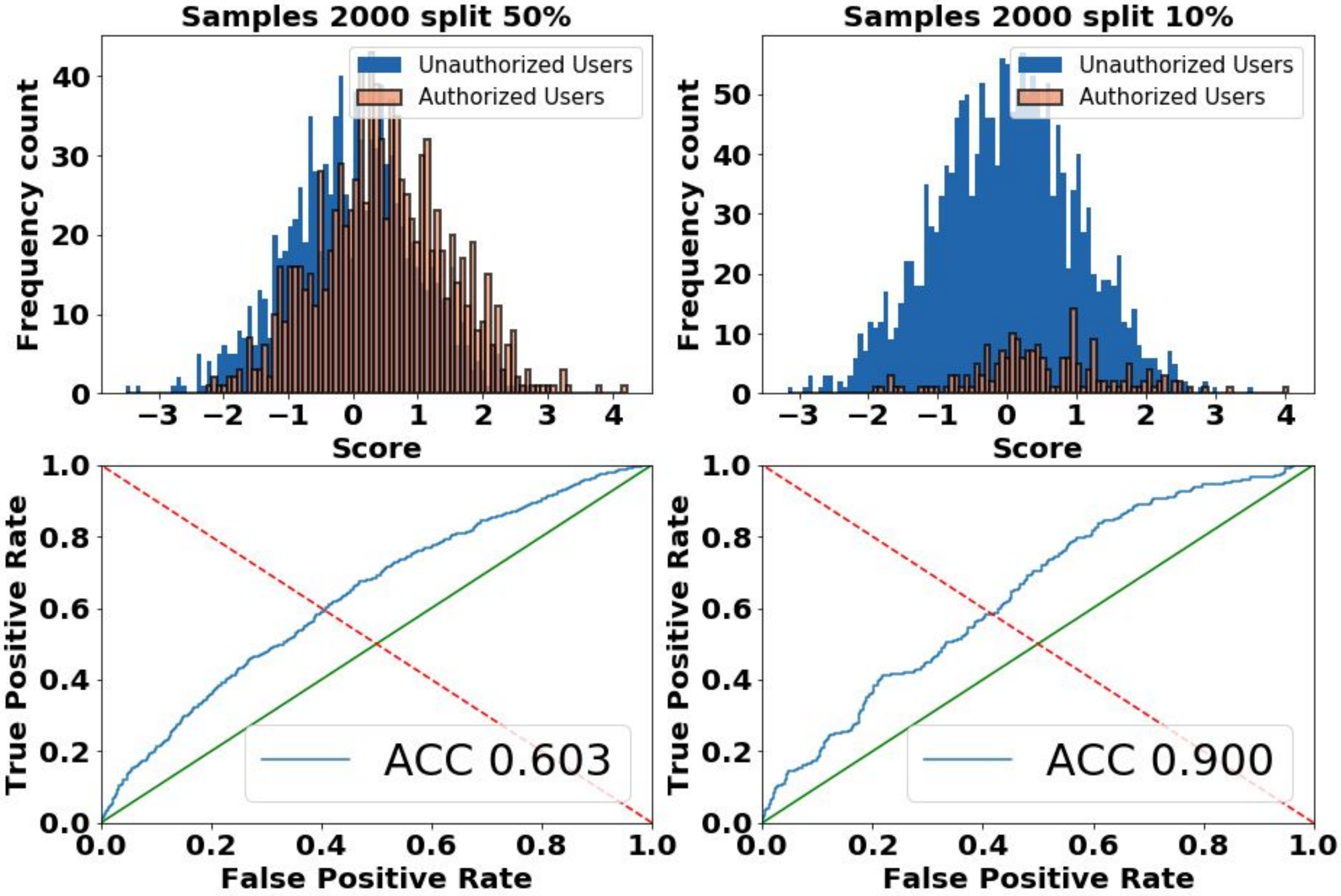
Reporting metrics with one threshold
⇒ Inadequate for system comparisons

Frequency Count of Scores (FCS) helps visualize problems

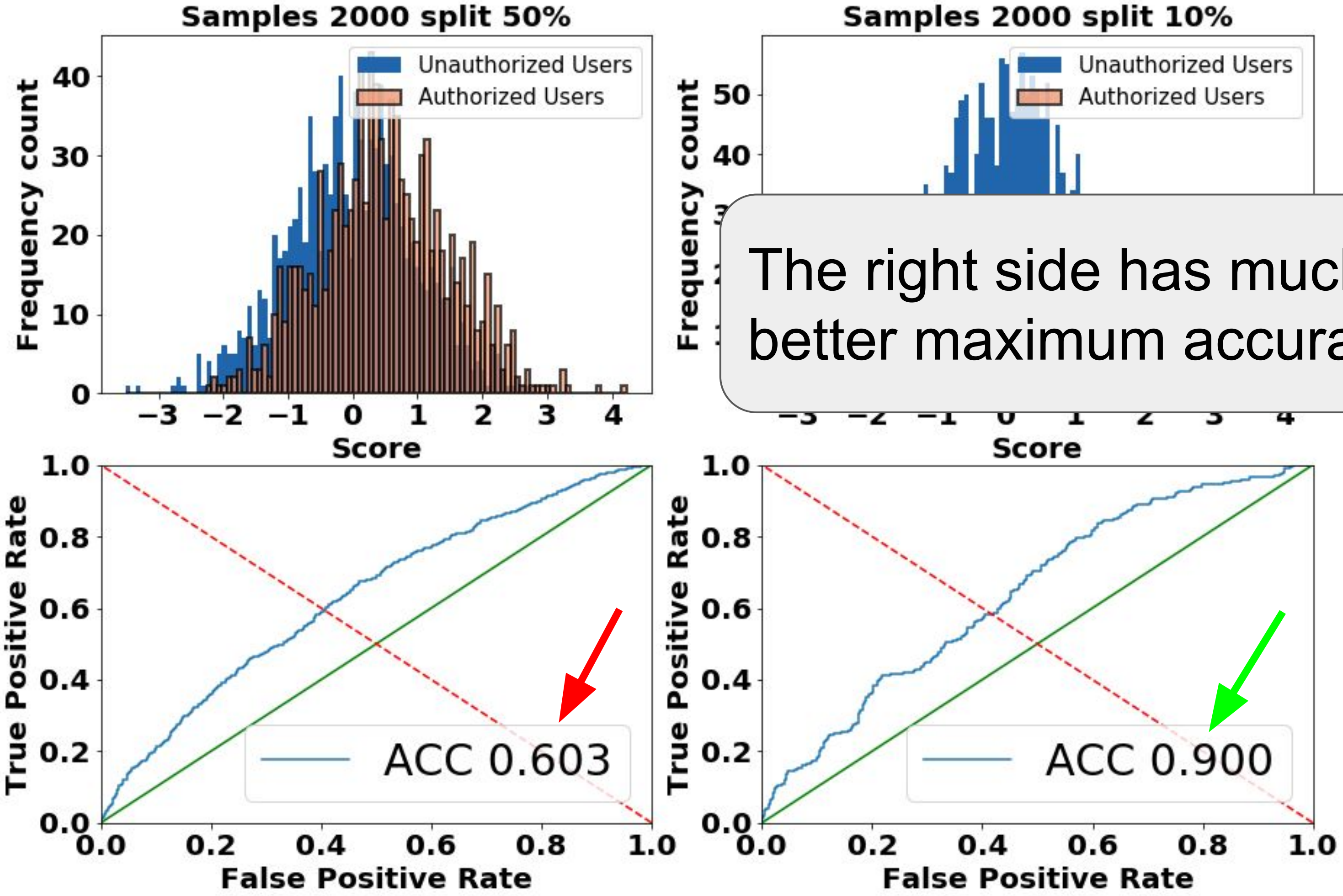
- The distribution of scores plays an important role in the system performance
- The potential for error is directly proportional to the width of the score overlap
- The FCS can be used to identify problems with scoring



Skewed populations make accuracy unreliable

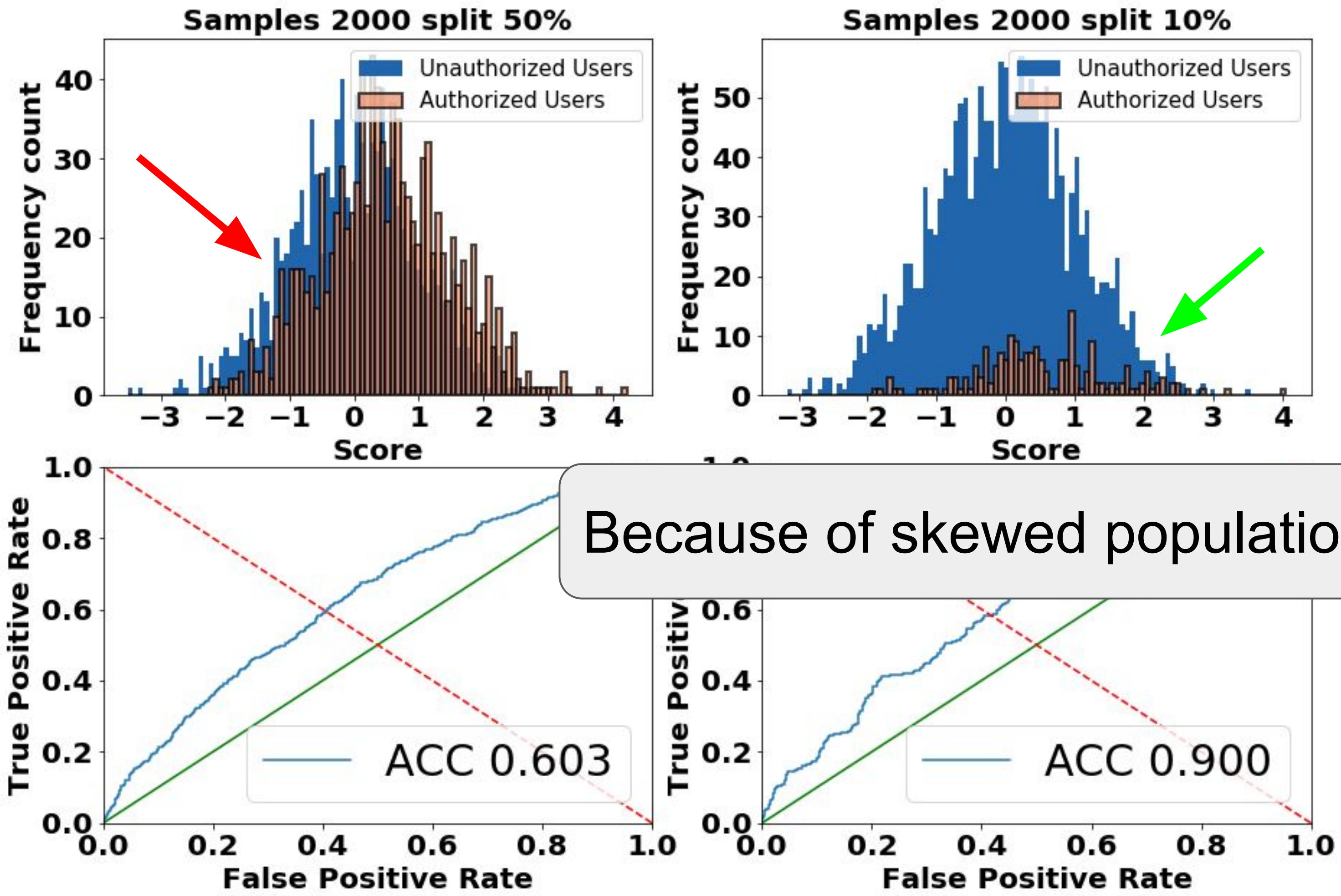


Skewed populations make accuracy unreliable



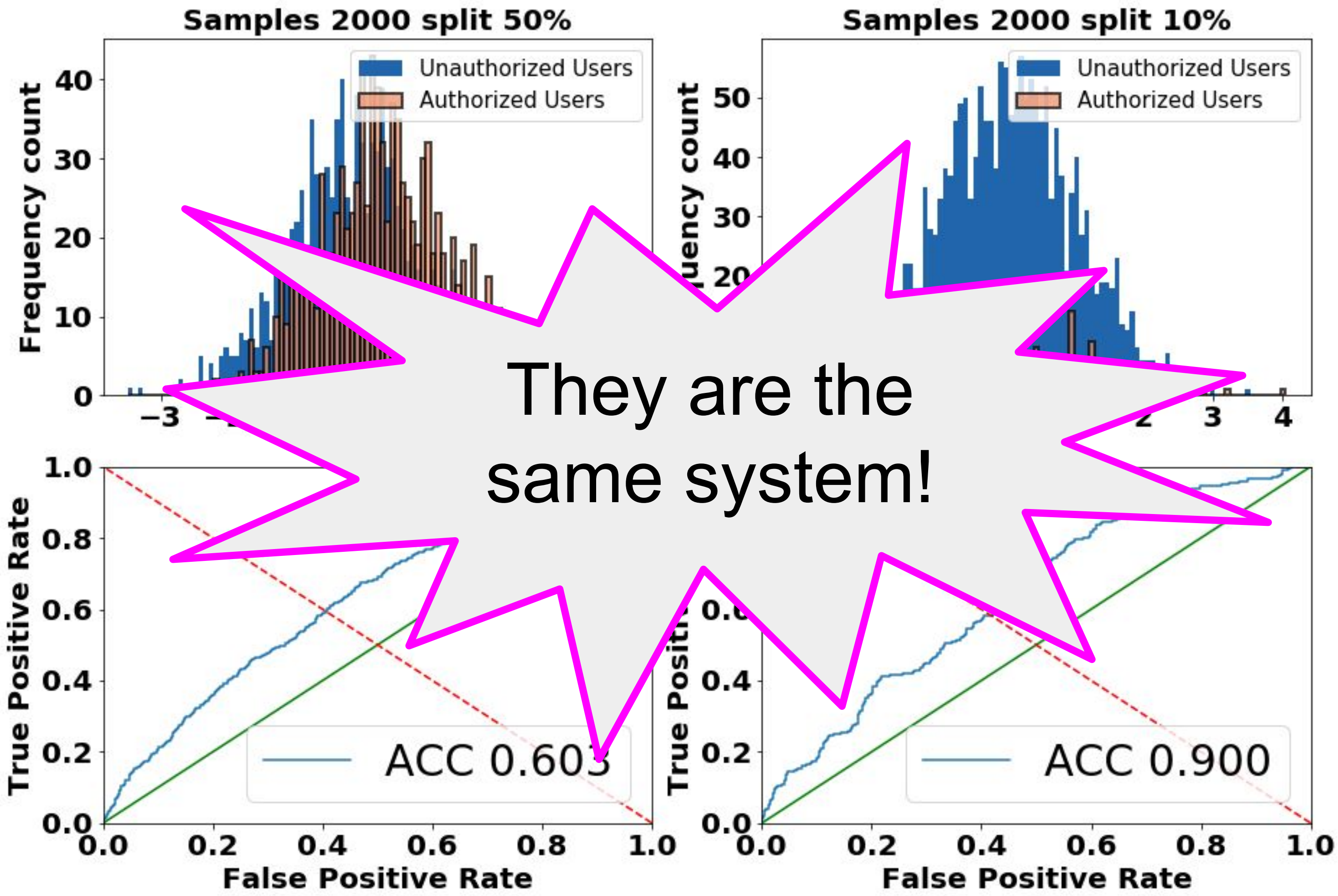
The right side has much better maximum accuracy!

Skewed populations make accuracy unreliable



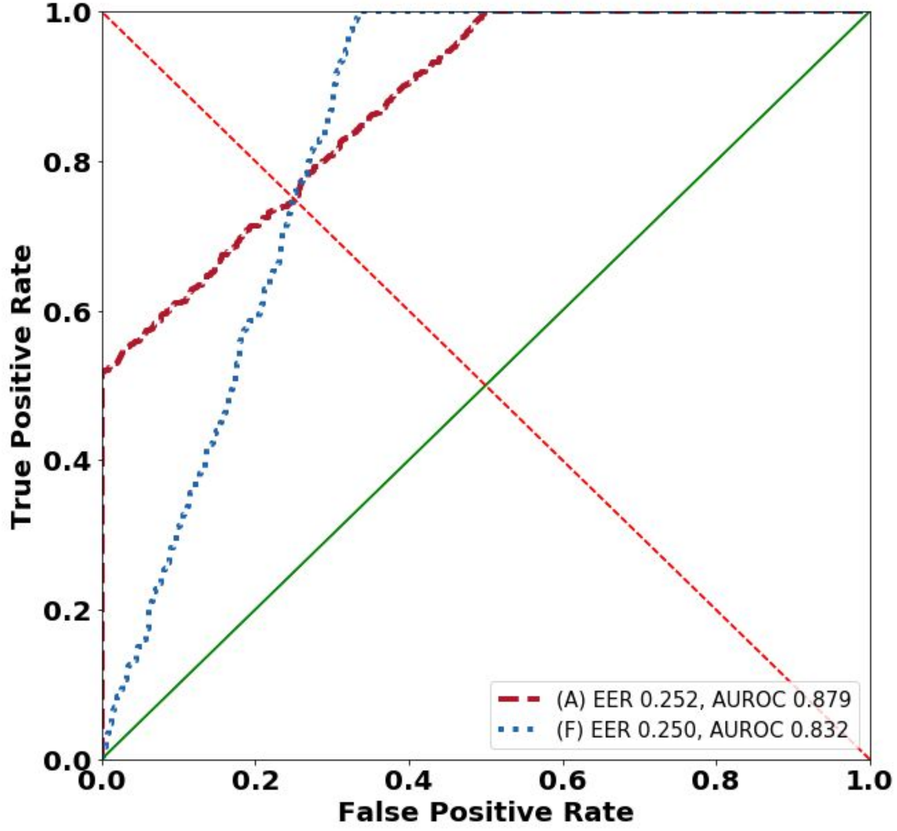
Because of skewed population

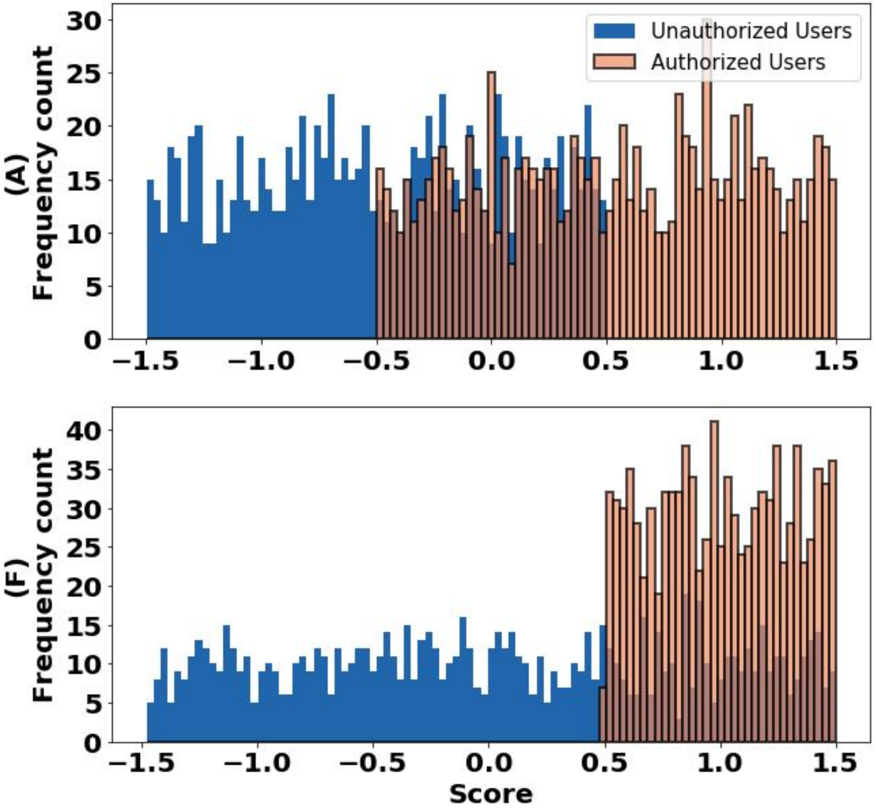
Skewed populations make accuracy unreliable



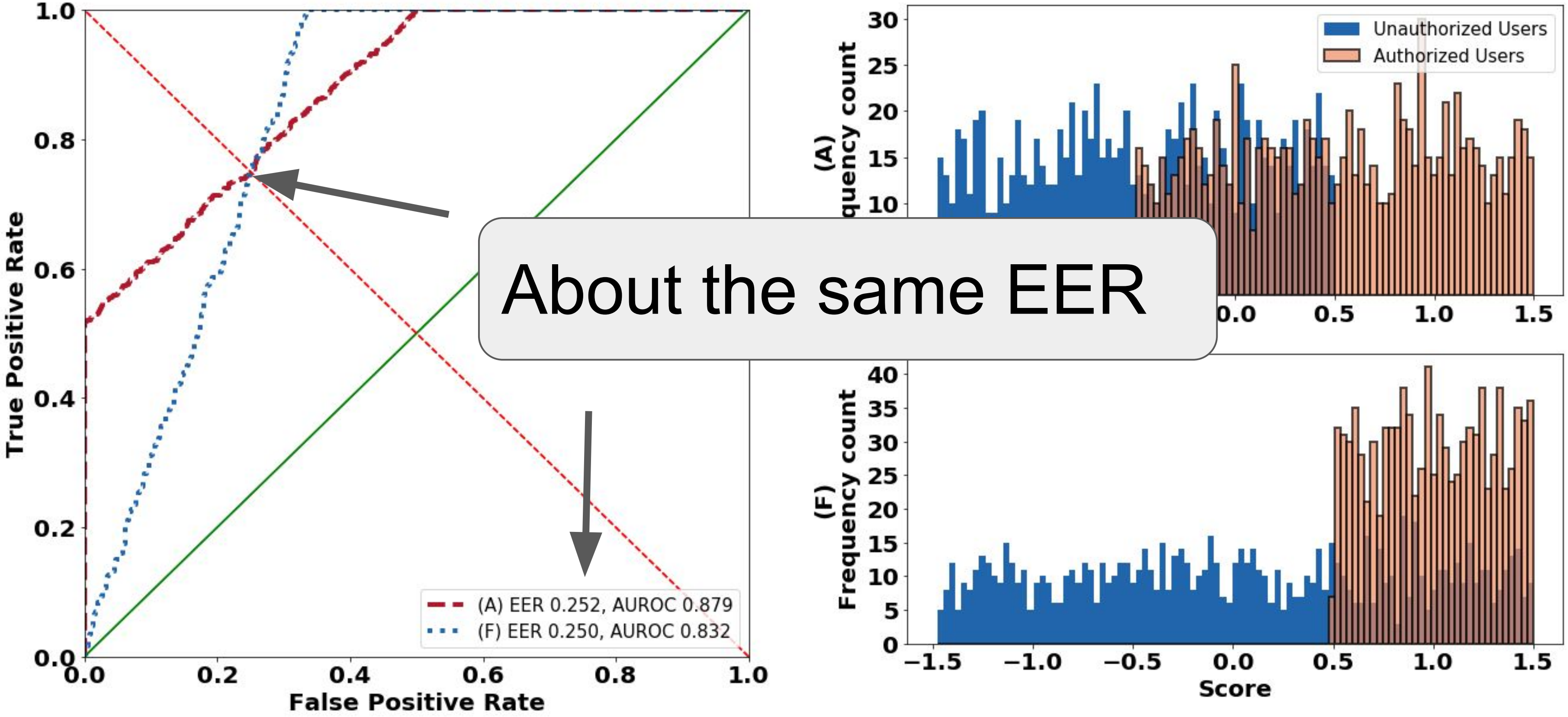
They are the same system!

Similar EER does not mean similar performance



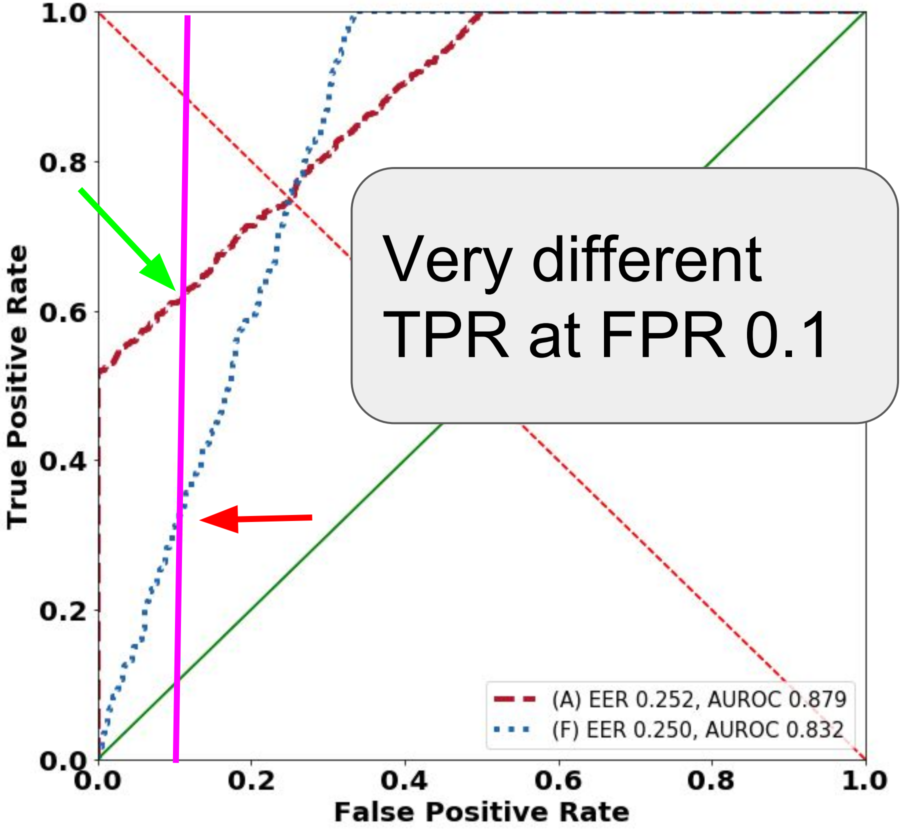


Similar EER does not mean similar performance

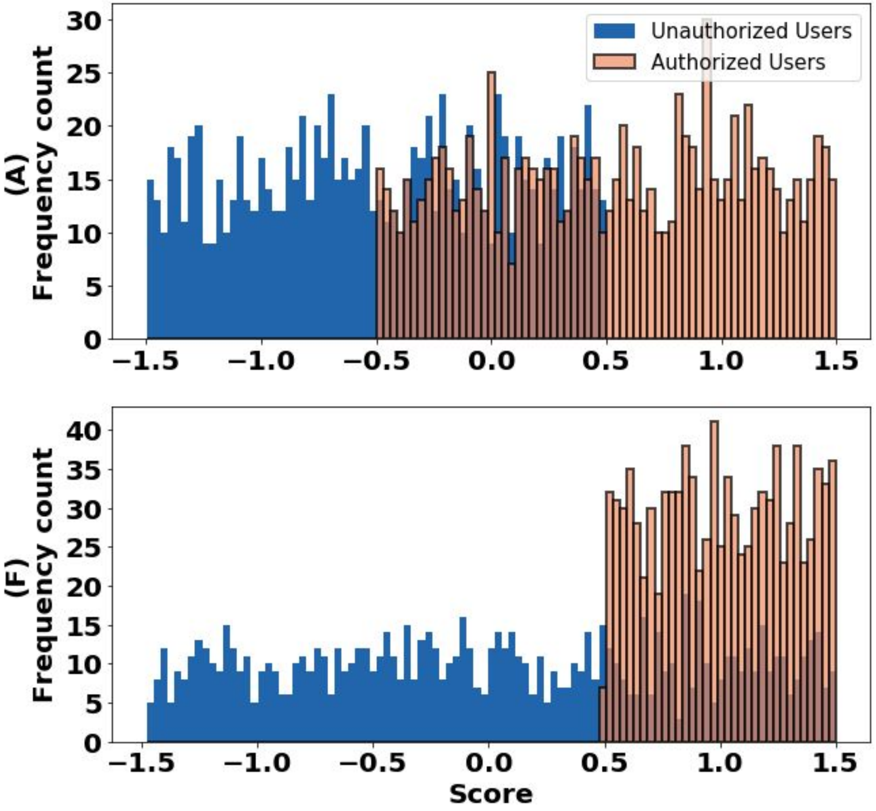


About the same EER

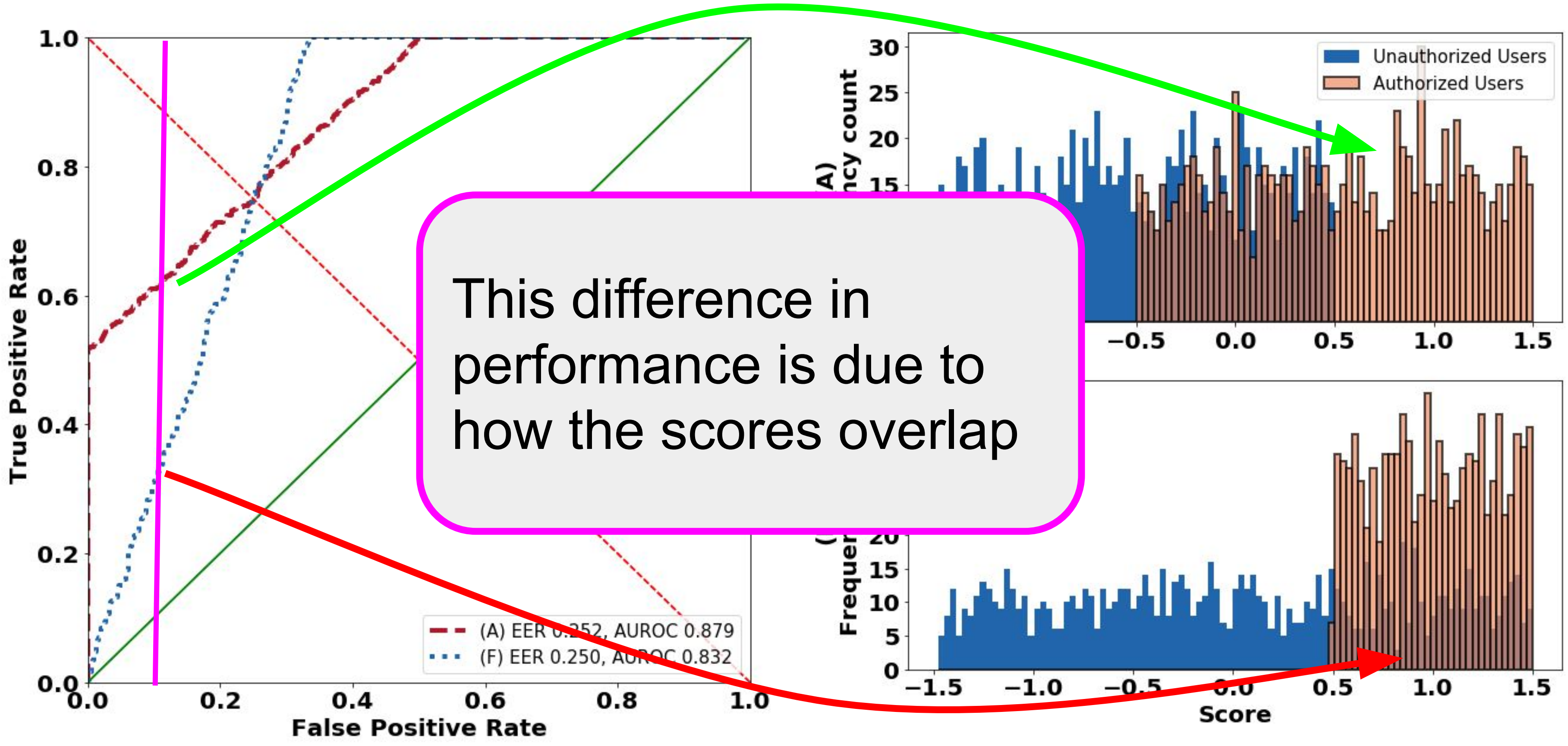
Similar EER does not mean similar performance



Very different TPR at FPR 0.1



Similar EER does not mean similar performance

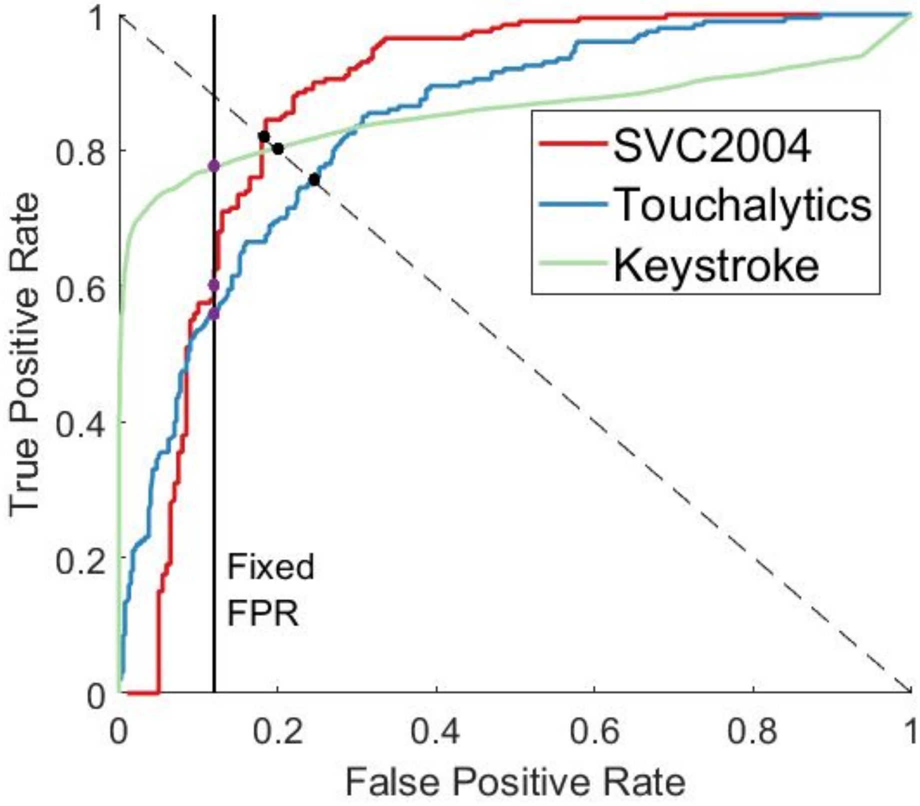


This difference in performance is due to how the scores overlap

EERs hides performance tradeoffs

- Compared 3 systems: SVC2004, Keystroke, Touchalytics

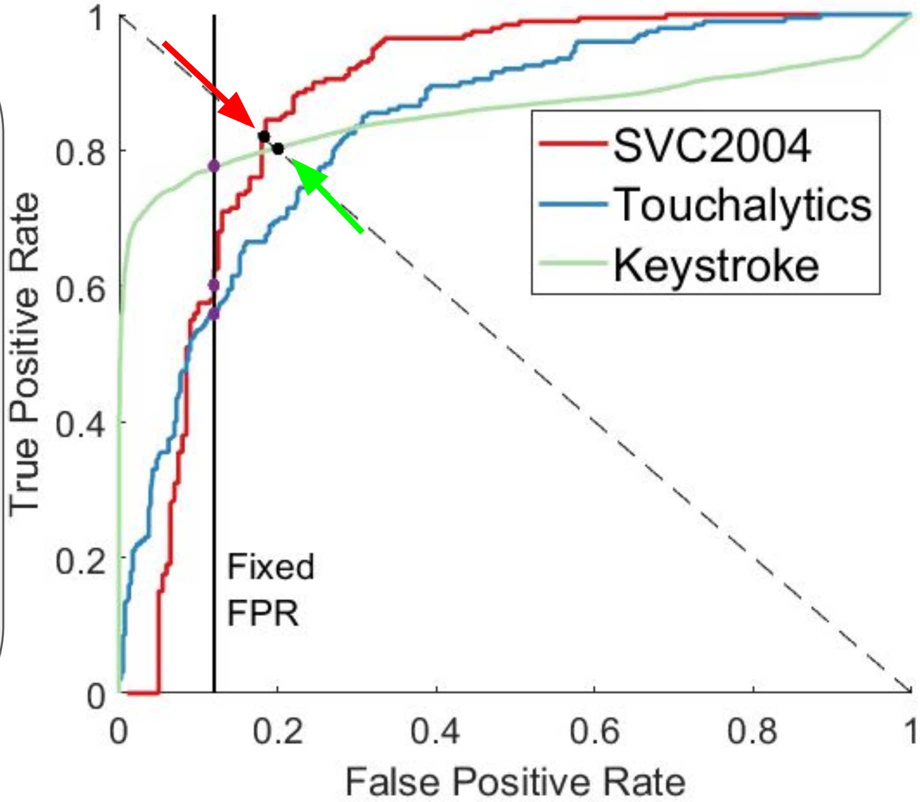
- EER is misleading: SVC2004 and Keystroke have similar EERs – Keystroke has FPR of 0.1



EERs hides performance tradeoffs

SVC2004 has a lower EER making it look better than Keystroke

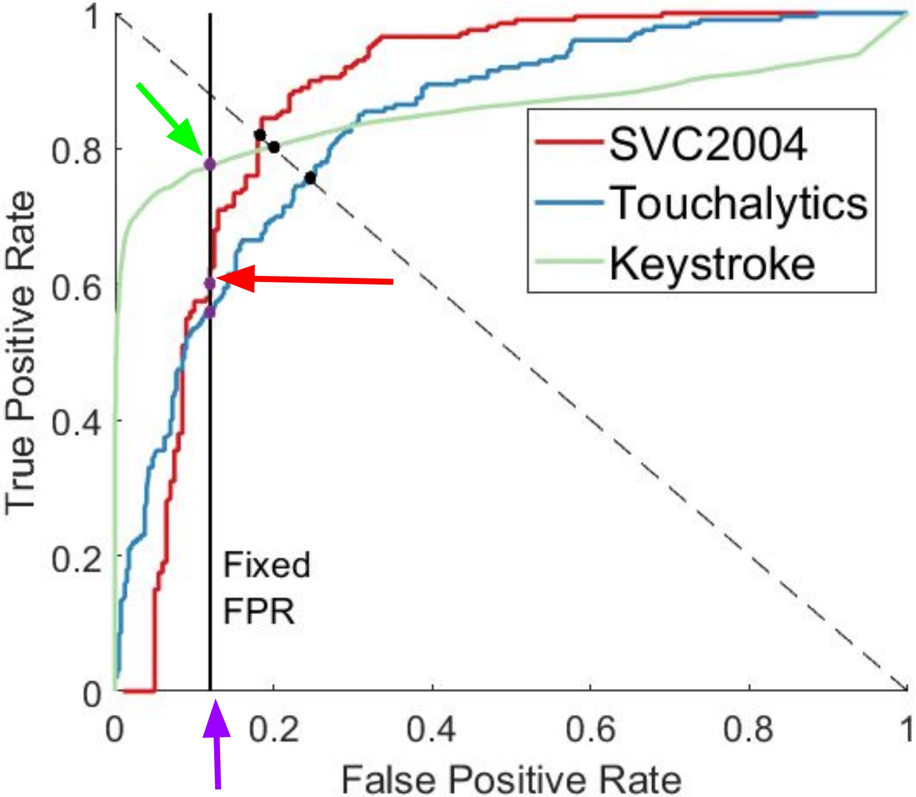
SVC2004 EER = 0.185
Keystroke EER = 0.198



EERs hides performance tradeoffs

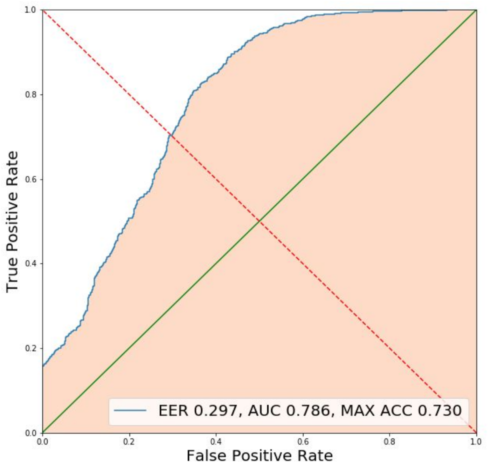
However,
Keystroke performs better at FPR is 0.1:

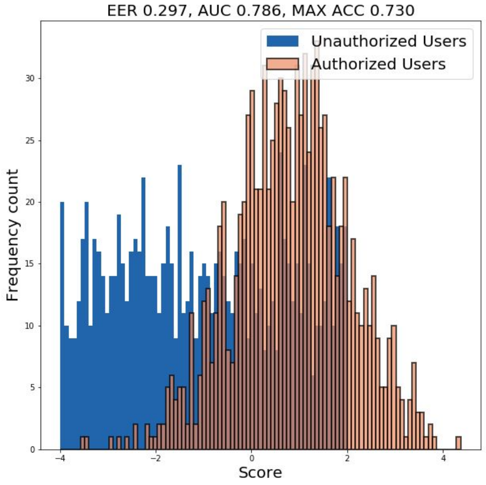
SVC2004 TPR = 0.600
Keystroke TPR = 0.776



Summary

- 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





Questions?

Please visit our websites for more details:

lindqvistlab.org
scienceofsecurity.science