

# Pitfalls in ROC Analysis when Evaluating Normalized 1:N Matcher Scores

**Brian DeCann, Ph.D**

**Brad Ulery**

**Nat Hall**

**Tim Busse**

**Delia McGarry**

**May 3<sup>rd</sup>, 2016**

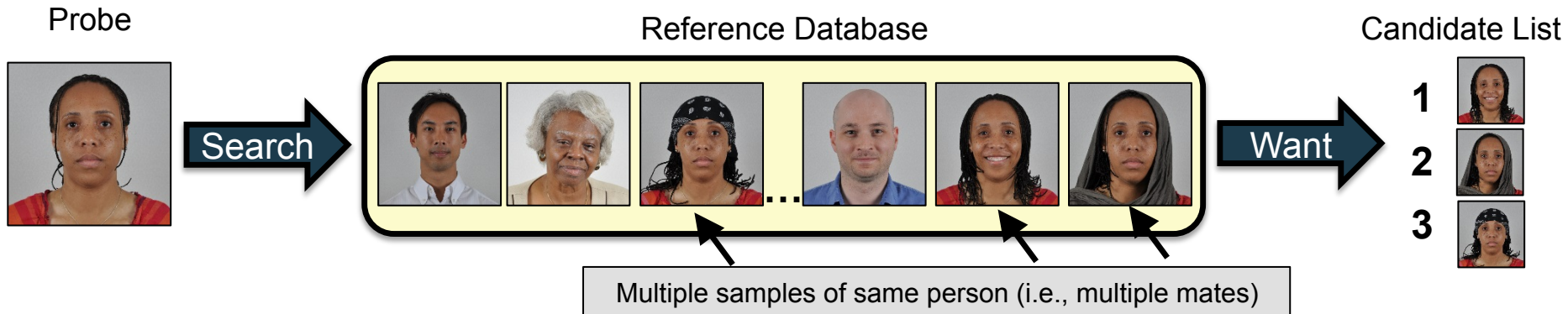
**noblis**<sup>®</sup>

**20**  
YEARS

*celebrating  
excellence*

# Types of 1:N Matching Scenarios

- Find all matching samples for the probe
  - Example: U.S. Department of State Face Recognition System



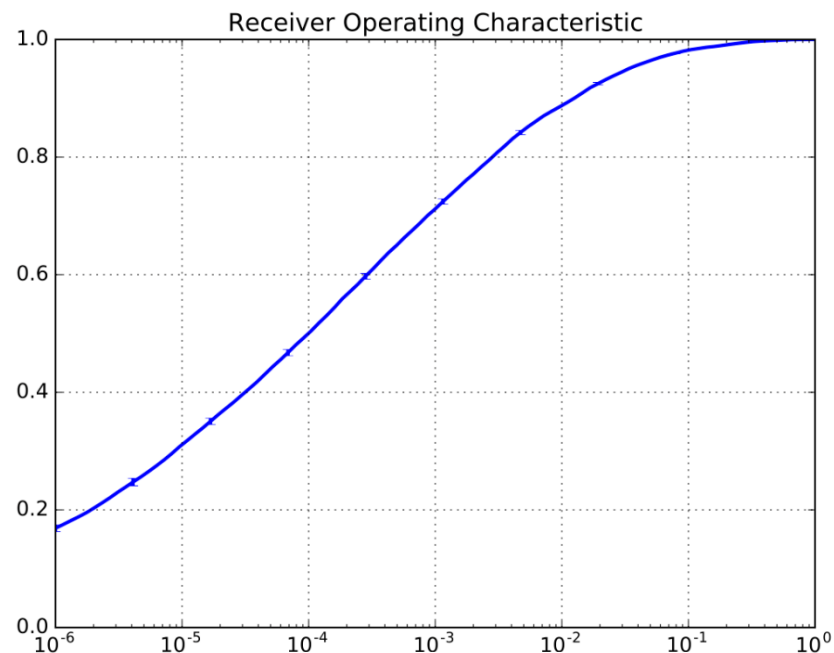
- Find (at least) one matching sample of the probe
  - Example: Access control, watch-list



Common to use ROC Analysis to evaluate matchers for these scenarios

# Matcher for Study

- Noblis Research Algorithm<sup>1</sup>
  - Deep learning approach
  - Template: 1280 bytes
  - Search 1M templates ~10s
  - C++ w/o licensing restrictions
  - Available for transition to Government
- Performance
  - TMR @ FMR = 0.1%: 70%



Recognition Performance on the Benchmark of Large-scale Unconstrained Face Recognition (BLUFR) dataset.<sup>2</sup>

## Contact

Dr. Mark Burge: [mark.burge@noblis.org](mailto:mark.burge@noblis.org)  
Jordan Cheney: [jordan.cheney@noblis.org](mailto:jordan.cheney@noblis.org)

<sup>1</sup> Sponsored by Noblis Internal Research (NSR)

<sup>2</sup> <http://www.cbsr.ia.ac.cn/users/scliao/projects/blufr/>

# 1:1 Verification

Genuine Comparison

Variable: How many comparisons?



Probe



Reference



0.87

To genuine (mate) scores  
(Compute TAR)



ROC: TAR vs. FAR



0.24

To impostor (non-mate) scores  
(Compute FAR)



Probe



Reference



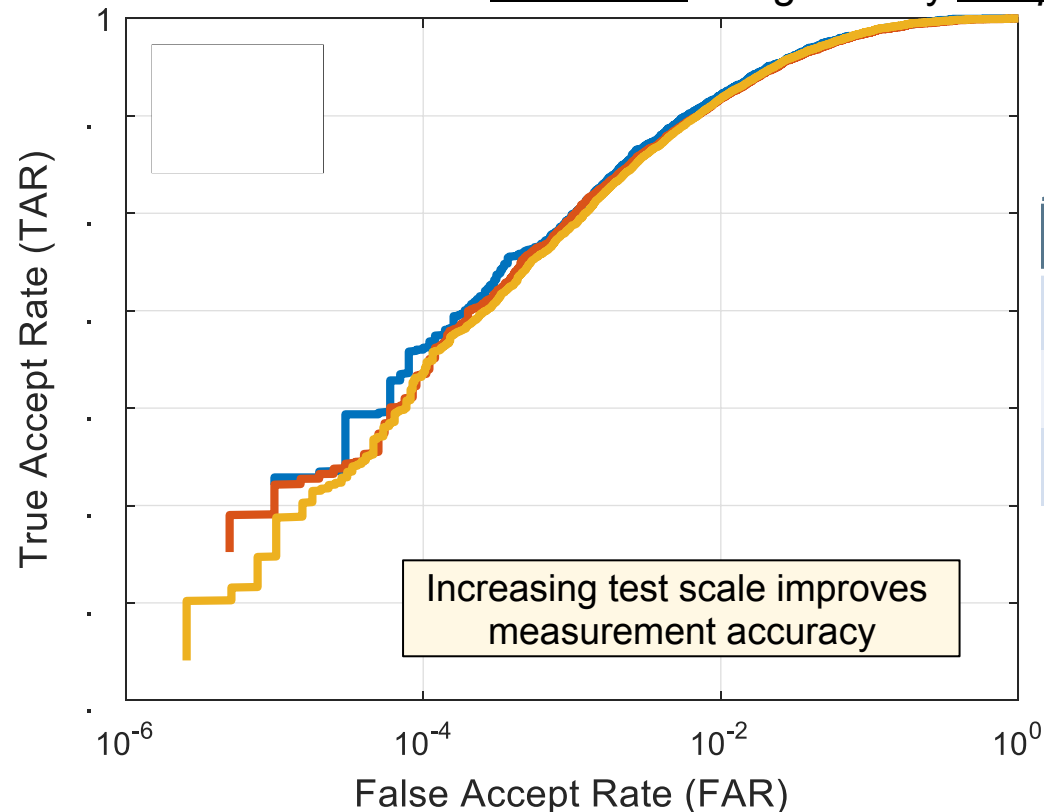
Variable: How many comparisons?

Impostor Comparison

# Receiver Operating Characteristic (ROC) Analysis (1:1)

## ■ 1:1 Verification

- Measured error rates are generally *independent* of scale of operations



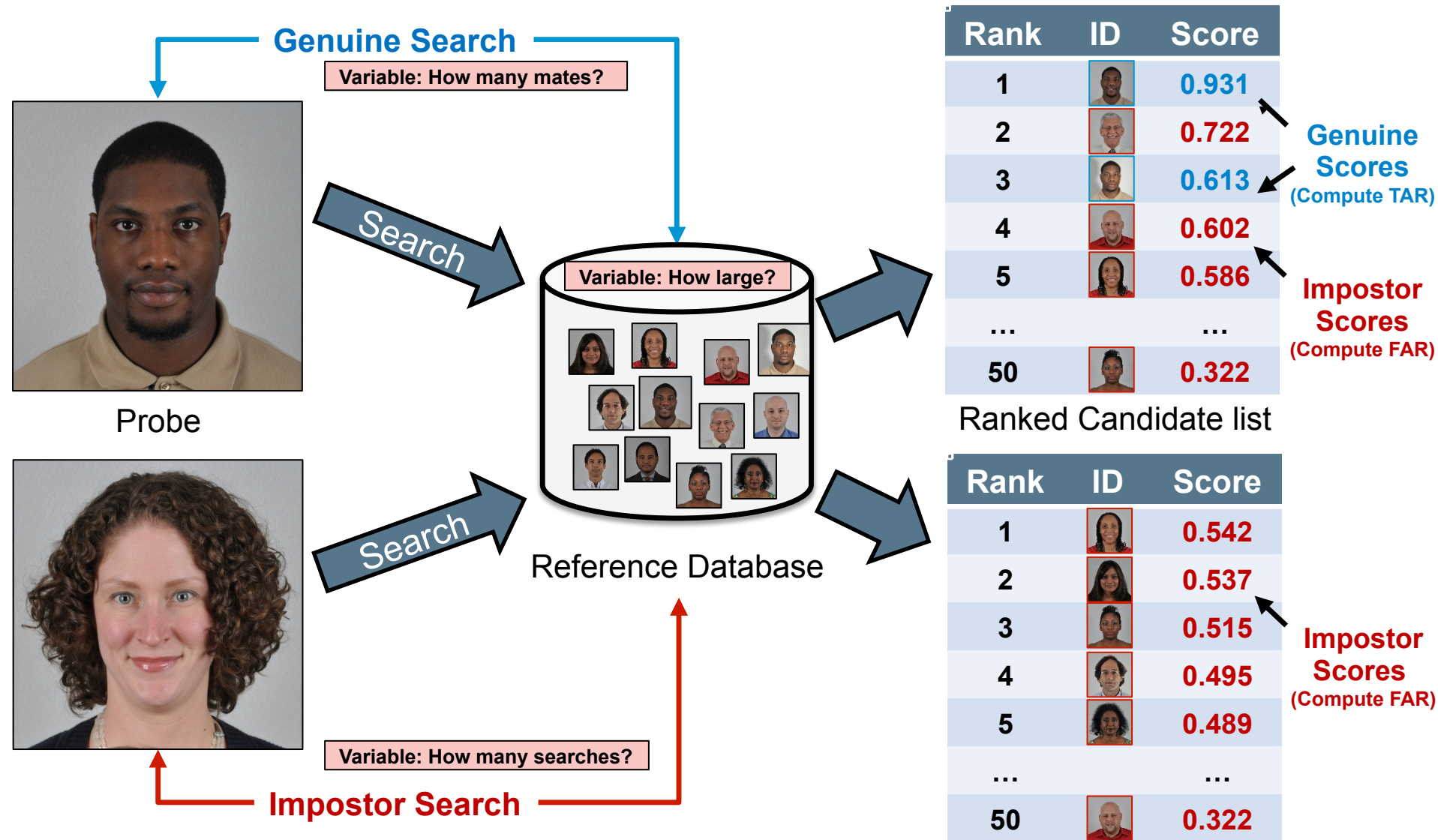
Test	#Genuines	#Impostors
A	2,500	100,000
	10,000	200,000
	15,000	400,000

Match scores obtained from Noblis research FR algorithm on a frontal face dataset

**For 1:1 verification, the ROC enables operational performance estimates from representative test data.**



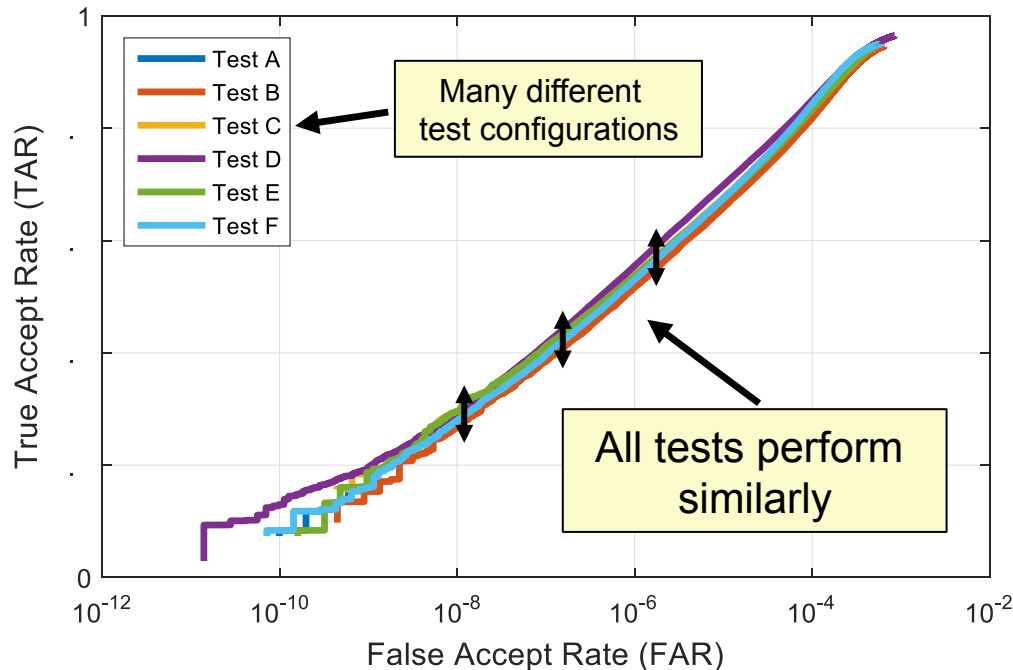
# 1:N Identification



# Receiver Operating Characteristic (ROC) Analysis (1:N)

## ■ 1:N Identification

- For **matcher scores that are strictly dependent** on the probe and reference sample, measured error rates generally independent of test configurations.
  - e.g.,  $FAR \downarrow N = 1 - (1 - FAR) \uparrow N \cong N \cdot FAR^1$



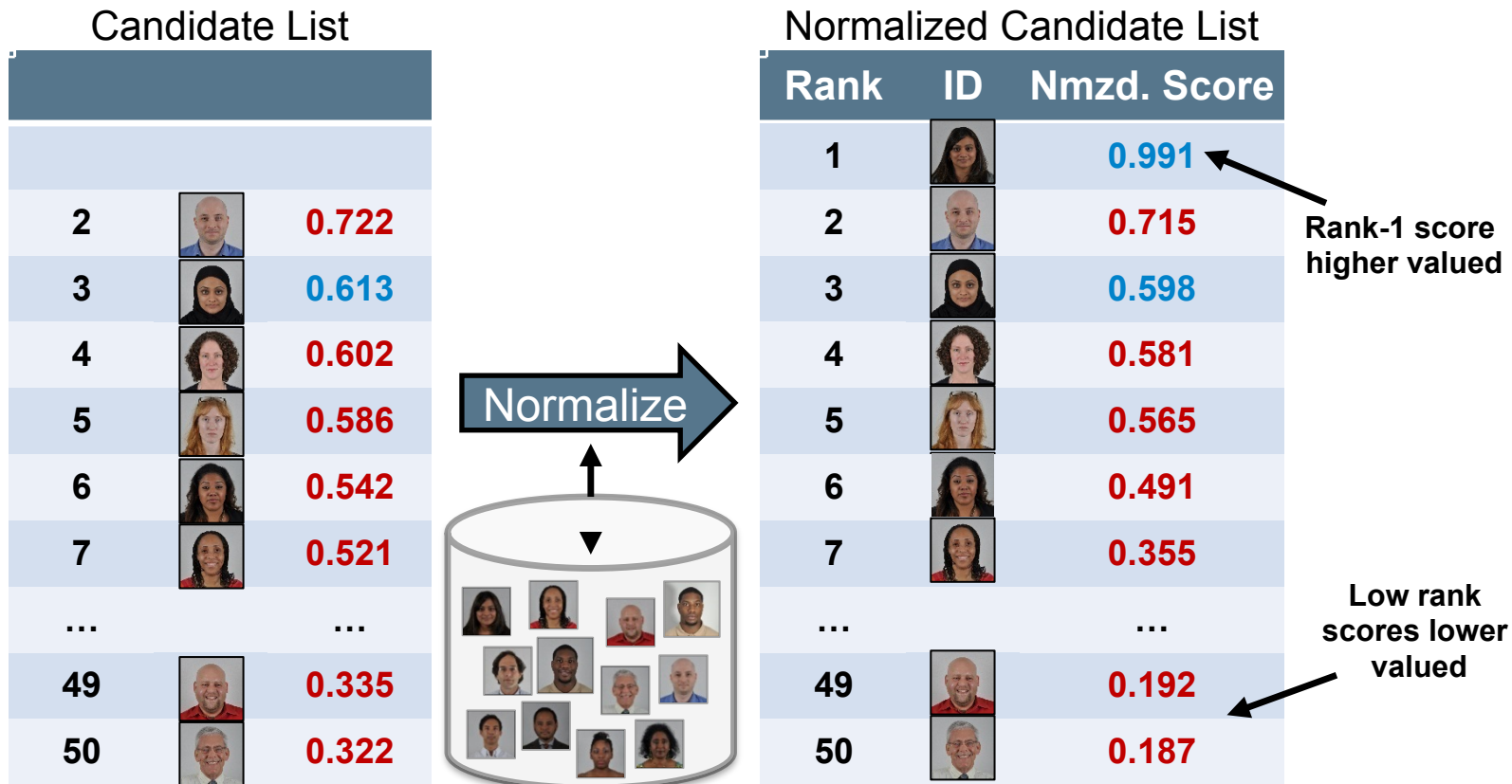
Test	Test Description
A	Gallery: <u>0, 1, 2, ... mates</u>
B	Gallery: <u>0 or 1 mates</u>
C	(A) <u>with additional mates</u>
D	(A) <u>with larger gallery</u>
E	(A) <u>without impostor searches</u>
F	(A) <u>with additional impostor searches</u>

Match scores obtained from Noblis research FR algorithm on a frontal face dataset

1 Jain, A., Ross, A., and Prabhakar, S., "An Introduction to Biometric Recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, 2014

**Not all 1:N matchers function this way!**

# 1:N Identification with Gallery Normalization

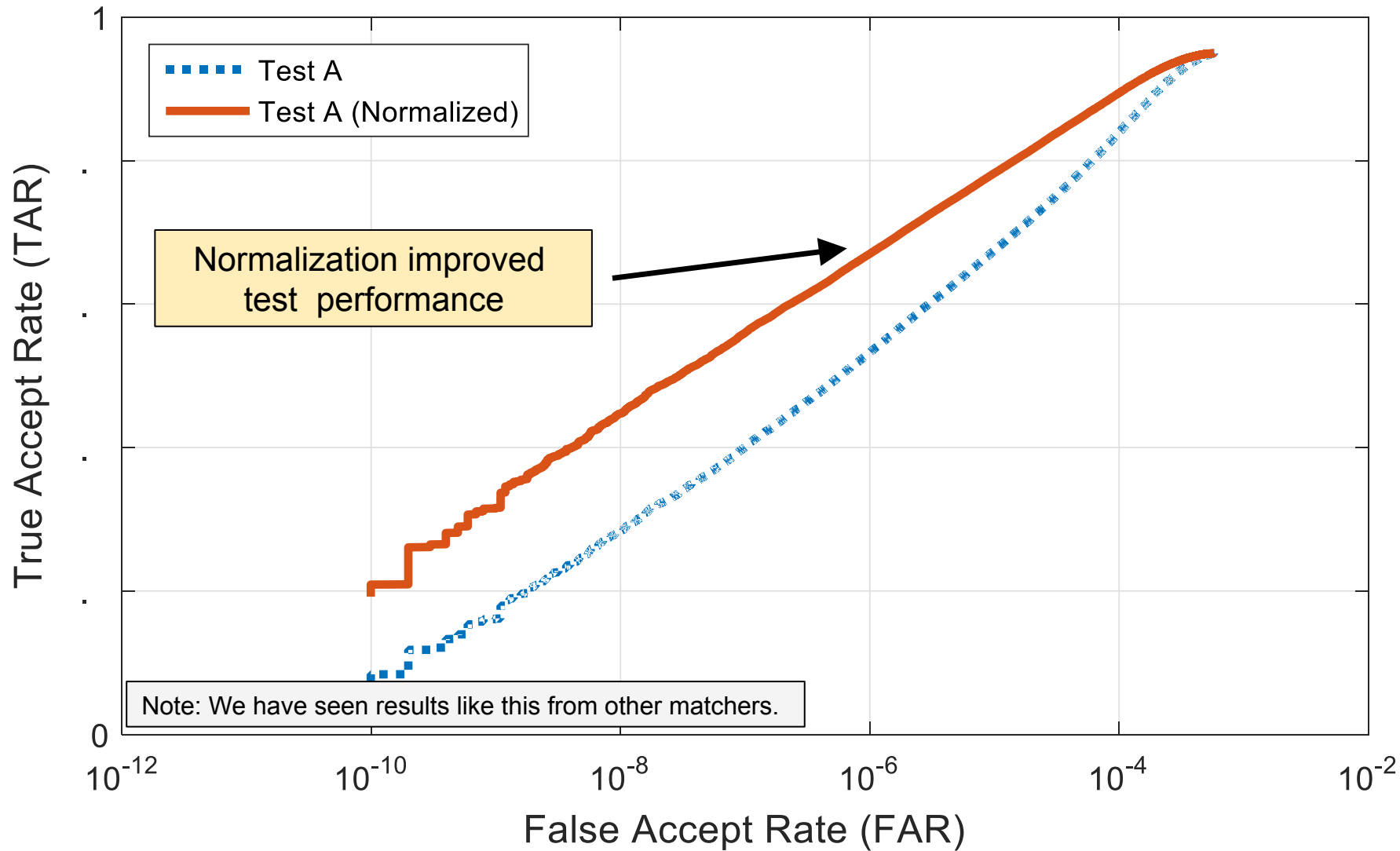


**Why do this?**

A 1:N matcher with gallery normalization may “boost” high scores and “suppress” low scores based on rank position. Note in our example we simply boosted the rank-1 score and suppressed the others.








# Normalization Can Improve ROC Performance



# Potential Pitfalls







Algorithm A (Normalized)

Genuine Search

Rank	ID	Nmzd. Score
1		0.991
2		0.815
3		0.568
4		0.541
5		0.515
6		

Algorithm returns mates at top ranks in candidate list.  
(desirable for identification, not captured by ROC)

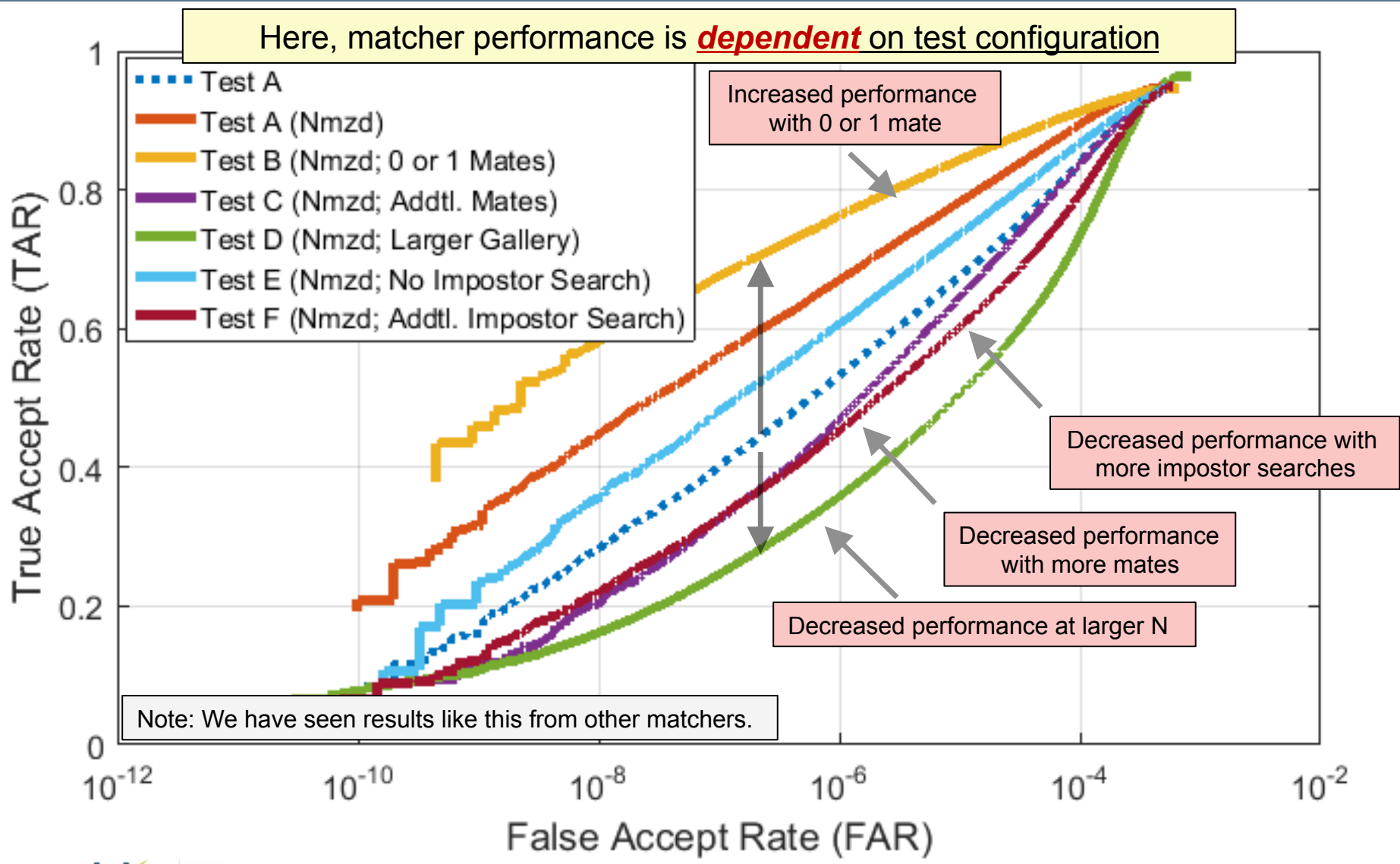
Impostor Search

Rank	ID	Nmzd. Score
1		0.788
2		0.575
3		0.559
4		0.552
5		0.538
6		0.512

But, **lower rank genuine scores** suppressed compared to impostor scores.  
(decreases TAR, ROC performance)

Boosting of high rank impostor scores increases FAR.  
More **impostor searches** → Lower performance  
Fewer **impostor searches** → Higher performance

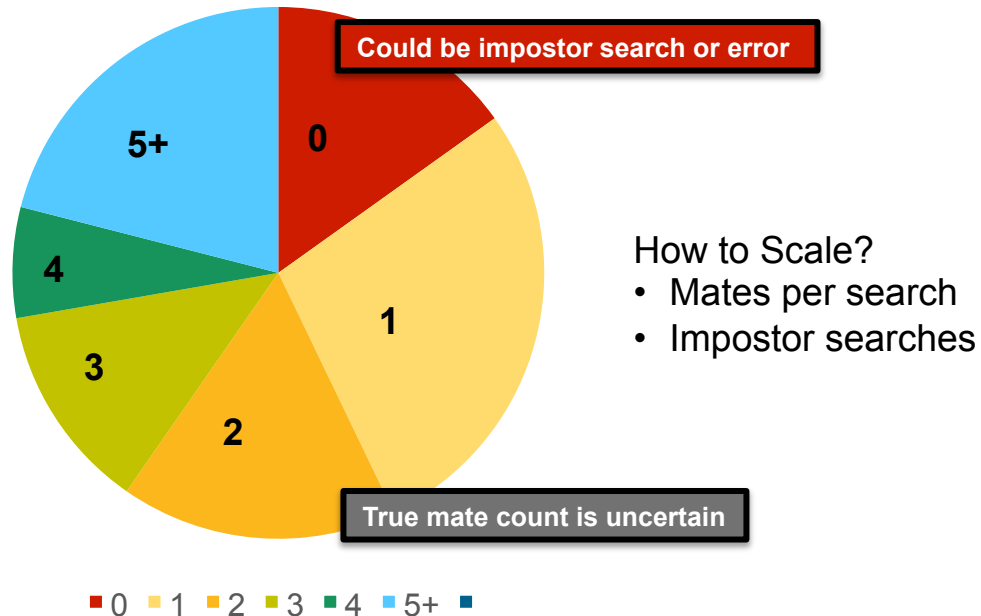
# Matcher Performance (with Normalization) may Depend on Test Configuration



# Challenge: Developing a Test Gallery

- How to appropriately model the distribution of mates per probe?
- How to appropriately model the proportion of genuine / impostor searches?

Mates Returned in Operational Open-set 1:N System



What can be created for testing

≠

Information from the system

# What does this mean?

- Dependent Results
  - Impact: extrapolating performance
  - Impact: comparing multiple matching algorithms
  
- Modeling Issues
  - Size of test database
  - Distribution of mates for genuine searches (how to scale from operations?)
  - Proportion of genuine and impostor searches (how to measure from operations?)
  - Interaction-effects (e.g., demographics)
  
- Best Practices for 1:N Testing
  - (Current): Requires execution of searches with and without mates<sup>1,2</sup>
  - (*Not Present*): Guideline regarding the proportion of mated searches
  - (*Not Present*): Guideline regarding proportion of mates in test database

---

<sup>1</sup> Grother, P., Ngan, M., "Face Recognition Vendor Test (FRVT), Performance of Face Identification Algorithms", NIST Interagency Report 8009, May 2014

<sup>2</sup> Grother, P., Quinn, G., and Phillips, P., "Report on the Evaluation of 2D Still-image Face Recognition Algorithms", NIST Interagency Report 7709, 2010

# Is ROC Analysis Appropriate?

## Common Metrics for Evaluation

	ROC Analysis	FPIR / FNIR / CMC <sup>1,2</sup>
Target Scenario (examples)	Find all mates (e.g., fraud detection)	Find any mate (e.g., watch-list)
Properties	Per-comparison credit Based on match scores	Per-search credit Based on rank and match scores
Weaknesses	Sensitivity to normalization May be dependent on N	Sensitivity to normalization Dependent on N

<sup>1</sup> Grother, P., Ngan, M., "Face Recognition Vendor Test (FRVT), Performance of Face Identification Algorithms", NIST Interagency Report 8009, May 2014

<sup>2</sup> Grother, P., Quinn, G., and Phillips, P., "Report on the Evaluation of 2D Still-image Face Recognition Algorithms", NIST Interagency Report 7709, 2010



# Recommendations

- For Developers / Vendors
  - Keep normalizing!
  - Be cognizant of customer needs
  
- For Operators (and Evaluators)
  - Communicate system specifications and evaluation criteria with developers
  - Identify objectives
    - Value (cost) of finding one vs. some vs. all mates
    - Operating point; Error trade-off
  
- For Evaluators Estimating Operational Performance from Test Data
  - Compose test sets to mimic application specific characteristics
  - Test on full-scale system when possible
  
- For Evaluators Comparing Matching Algorithms
  - Perform sensitivity analysis (varying test configurations)

# Questions?

## Acknowledgements

This work was sponsored by the United States Department of State, Bureau of Consular Affairs, Consular Systems and Technology.

Noblis Visual Analytics Lab