Large-Scale Evaluation of Biometric Algorithms

Elaine M. Newton, PhD
Computer Security Division
Information Technology Laboratory

National Institute of Standards and Technology (NIST)
Department of Commerce
17 January 2011

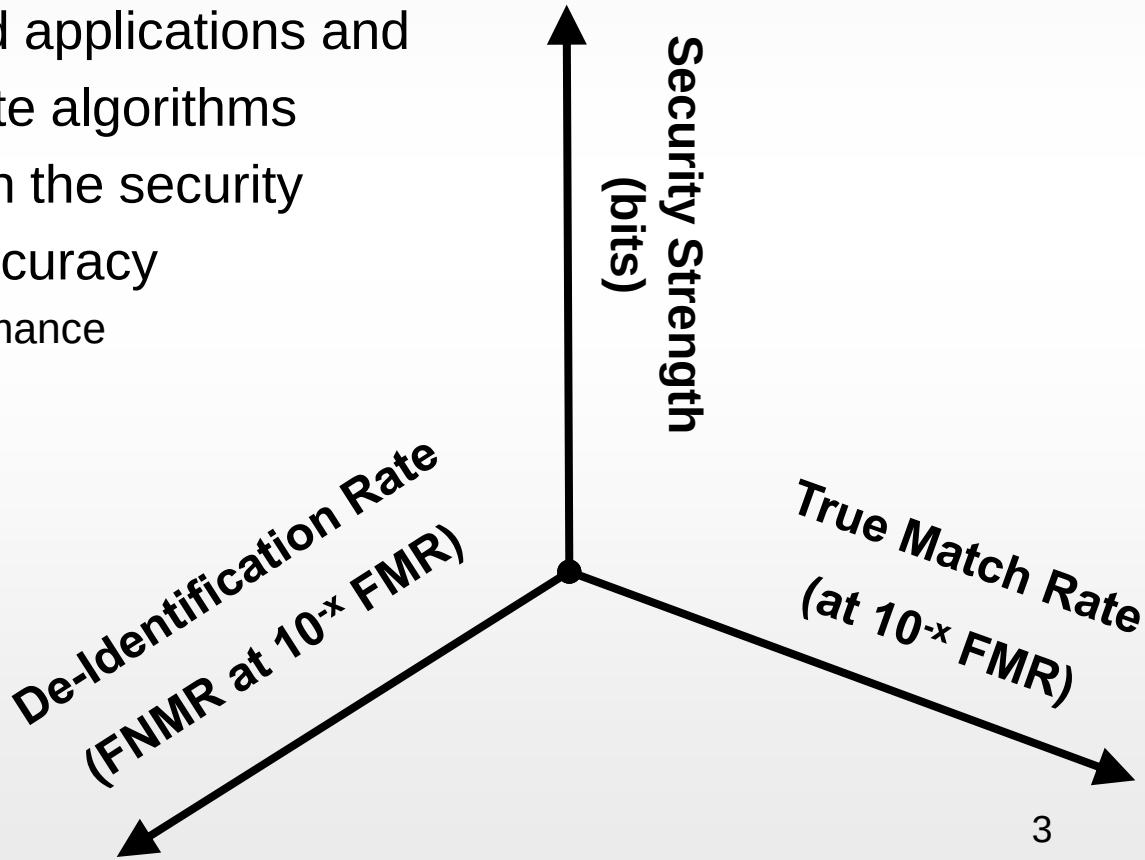
Why do we test?

- To assess the state-of-the-art
- To determine if products are interoperable
- To balance the asymmetry of information in the market place
- To give consumers confidence

➢ Ultimately, to provide a basis for trust in systems relying on the technology

Take Home Message

- Next steps for template protection testing will need to address
  - scale for intended applications and
  - metrics to evaluate algorithms incorporating both the security properties and accuracy
    - Biometric Performance
    - De-Identification
    - Irreversibility
    - Others



Evaluation of Technology

Principles

- Testing organization is an independent party.
- Sequestered data is used for evaluation.
- Results are published by testing organization.
- Testing methods are made publicly available to so that consumers of the results can assess the
  - technical merits of the methods used and
  - applicability to their needs.

Supported by:

- Testing standards (e.g. ISO/IEC 19795)
- Accreditation programs (e.g. NVLAP)

Large-Scale Biometric Testing

In collaboration

with Dr. Jonathon Phillips

NIST Biometric Testing/Data

- Fingerprint
  - Ongoing Proprietary Fingerprint Test (PFTII) and MINEX (MINutiae EXchange) testing using various databases of 120K+ subjects
  - Software development kit (SDKs) –based testing
- Face
  - Data from grand challenges and vendor tests
  - DOS Database of 37K subjects
  - Algorithm-based testing
- Iris
  - Data from grand challenges and vendor tests
  - Algorithm-based testing

And we can see improvements over time…

Progress in Fingerprint Recognition

FVC

	2000	2002	2006
Top 10% Mean	0.03640	0.00173	0.00074
Top 10% Median	0.03640	0.00170	0.00095
Top 10% Range	0.03640	0.14% to 0.21%	0.021% to 0.121%
Top Quartile Mean	0.04337	0.00599	0.00131
Top Quartile Median	0.04010	0.00605	0.00137
Top Quartile Range	3.64% to 5.36%	0.14% to 1.18%	0.021% to 0.237%

PFT

	POEBVA	DOS	DHS2
2006 Mean	0.0373	0.0437	0.0967
2007 Mean	0.0079	0.0095	0.0429

Progress in Face Recognition

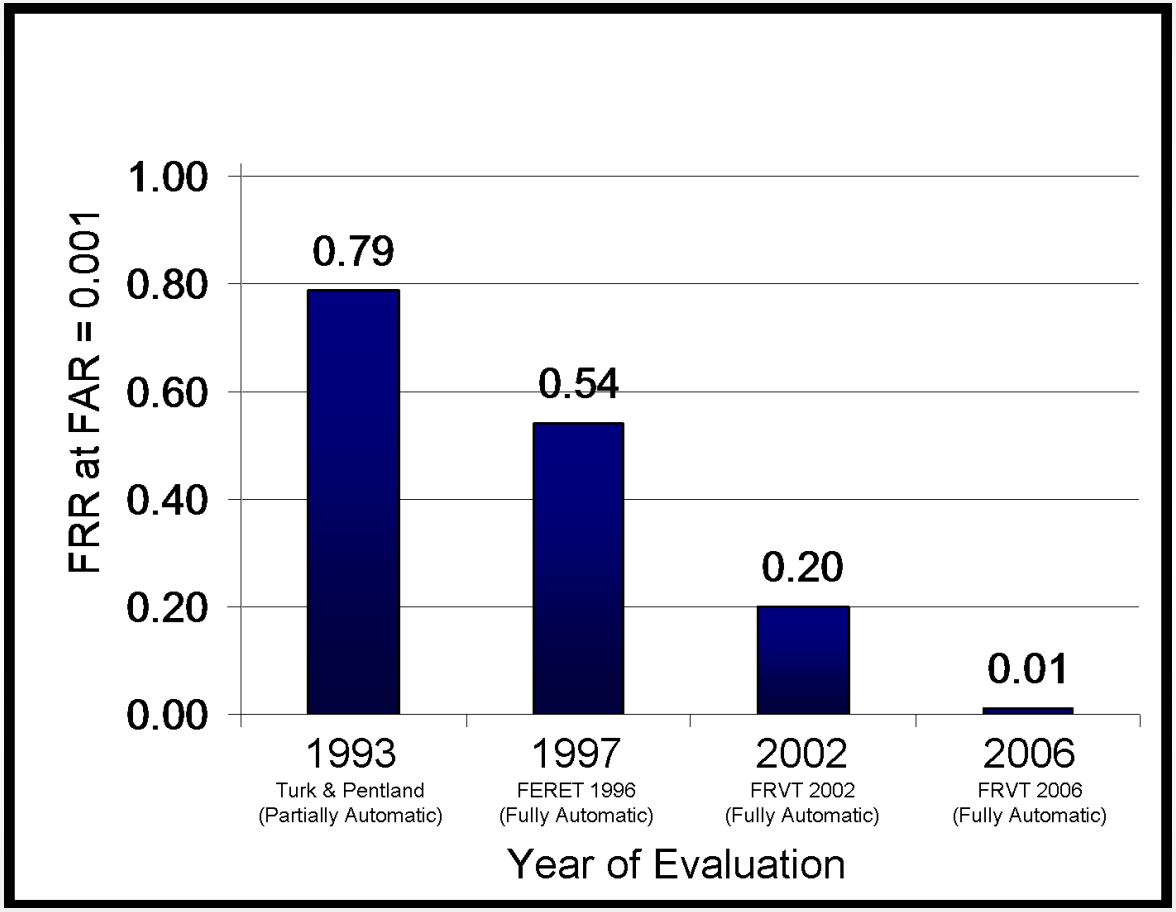


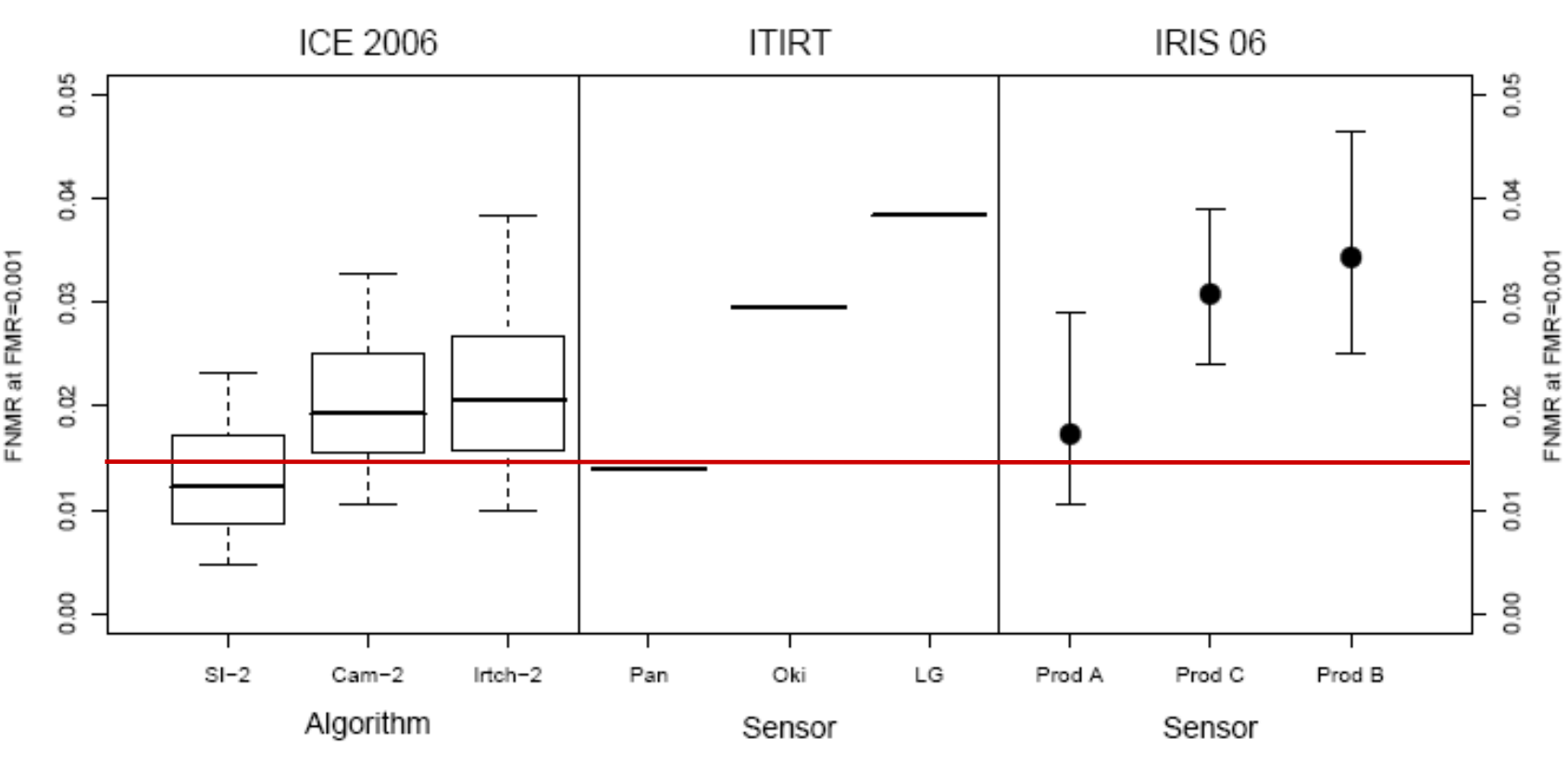
Figure is from the FRVT 2006 Report.

Three Recent Iris Evaluations

- ICE 2006
  - NIST
  - Results released in 2007
- ITIRT
  - Independent Test of Iris Recognition Technology
  - International Biometric Group
  - Results released in 2005
- IRIS06
  - Authenti-Corp
  - Results released in 2007

Analysis of Best Performers

- **Average FNMR = 0.0146**
- **Average Absolute Difference = 0.004**



Relevant use cases

Scenario Name:	**Personal Security**		**Forensics**		**Watchlist**		**Large-Scale ID**	
Operational FMR:	0.01		0.0001		1.E-07		1.E-07	
Min|Max E[Subjects]	1	10	1000	1E+09	10	1.E+05	1.E+07	1E+09

Gap in Testing Large-Scale Applications

- Large-scale (150K and greater) verification and/or identification performance, especially at very low error rates.
  - What is currently understood about the accuracy of biometric systems through empirical study?
  - Is testing answering questions relevant to anticipated uses of biometrics, such as personal security, forensic ID, watchlists, and large-scale ID?
  - Can current testing methods be leveraged to fill the gaps between current testing and anticipated needs?

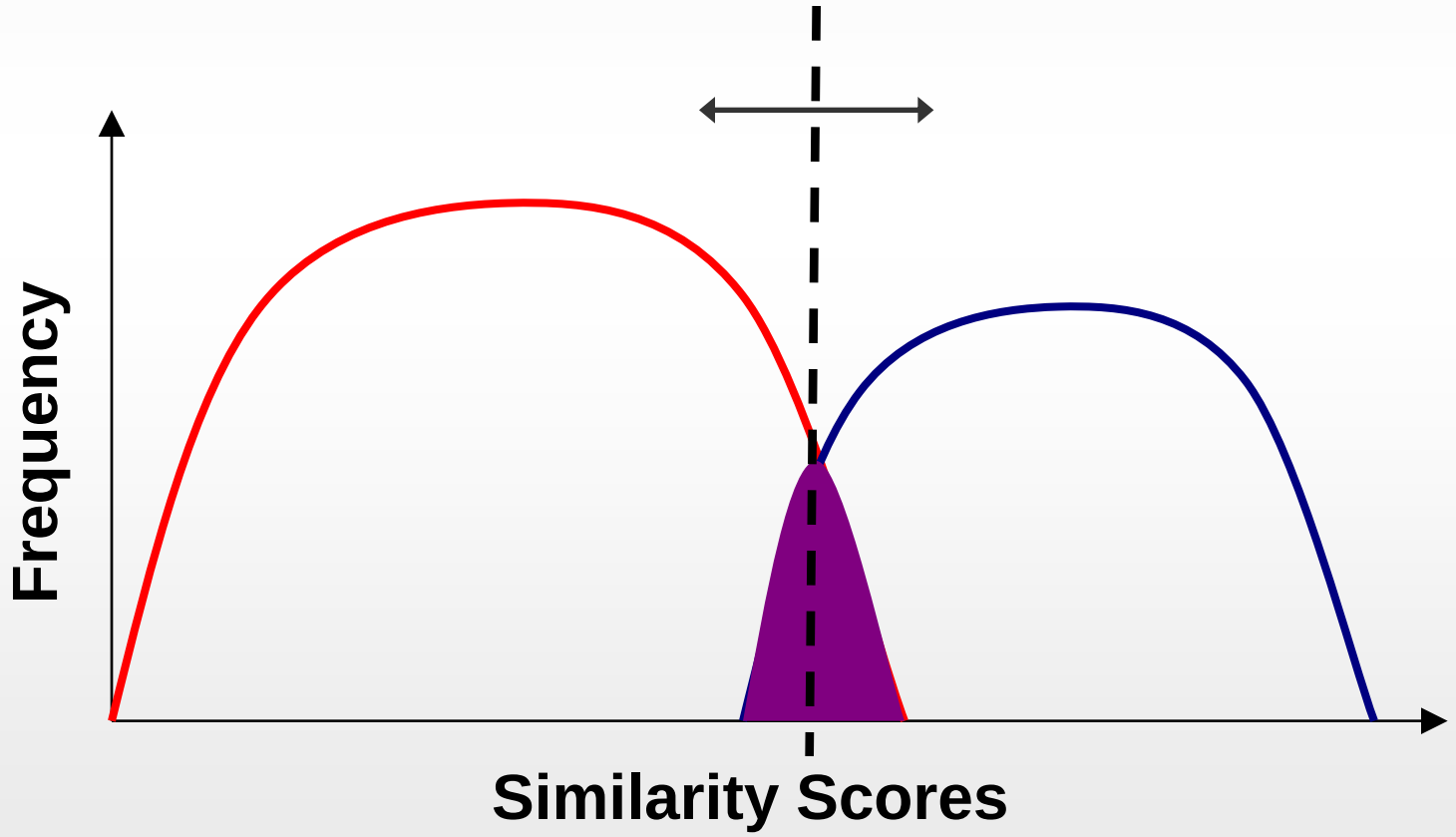
i→I Experiment (1)

- Department of State database with 36,000 subject's faces

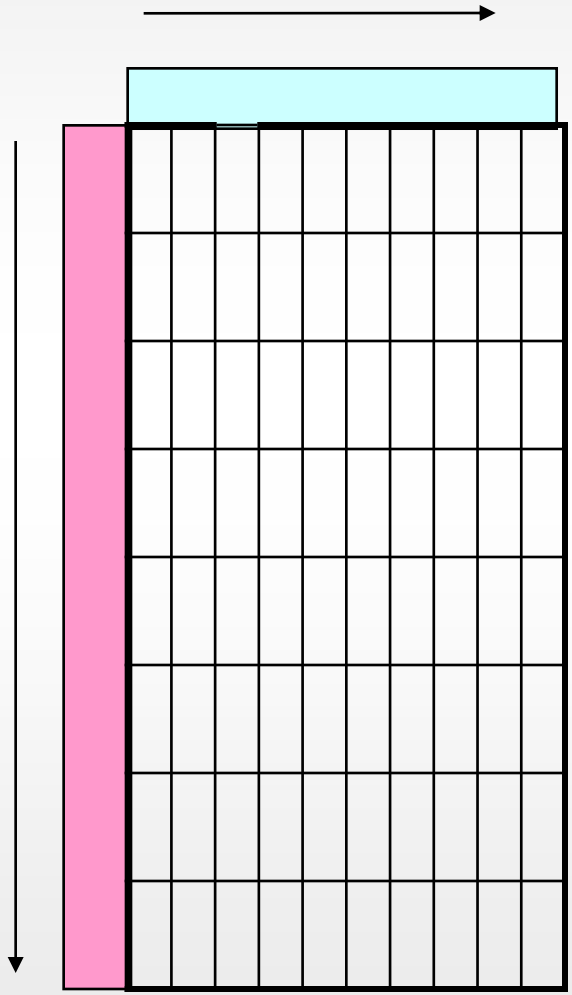
- Ran open-set identification tests to observe the effect of small scale testing on large-scale scenarios.

i→I Experiment (2)

- Tested two FMR levels: FMR = 0.01 and FMR = 0.001. Observing change in threshold.



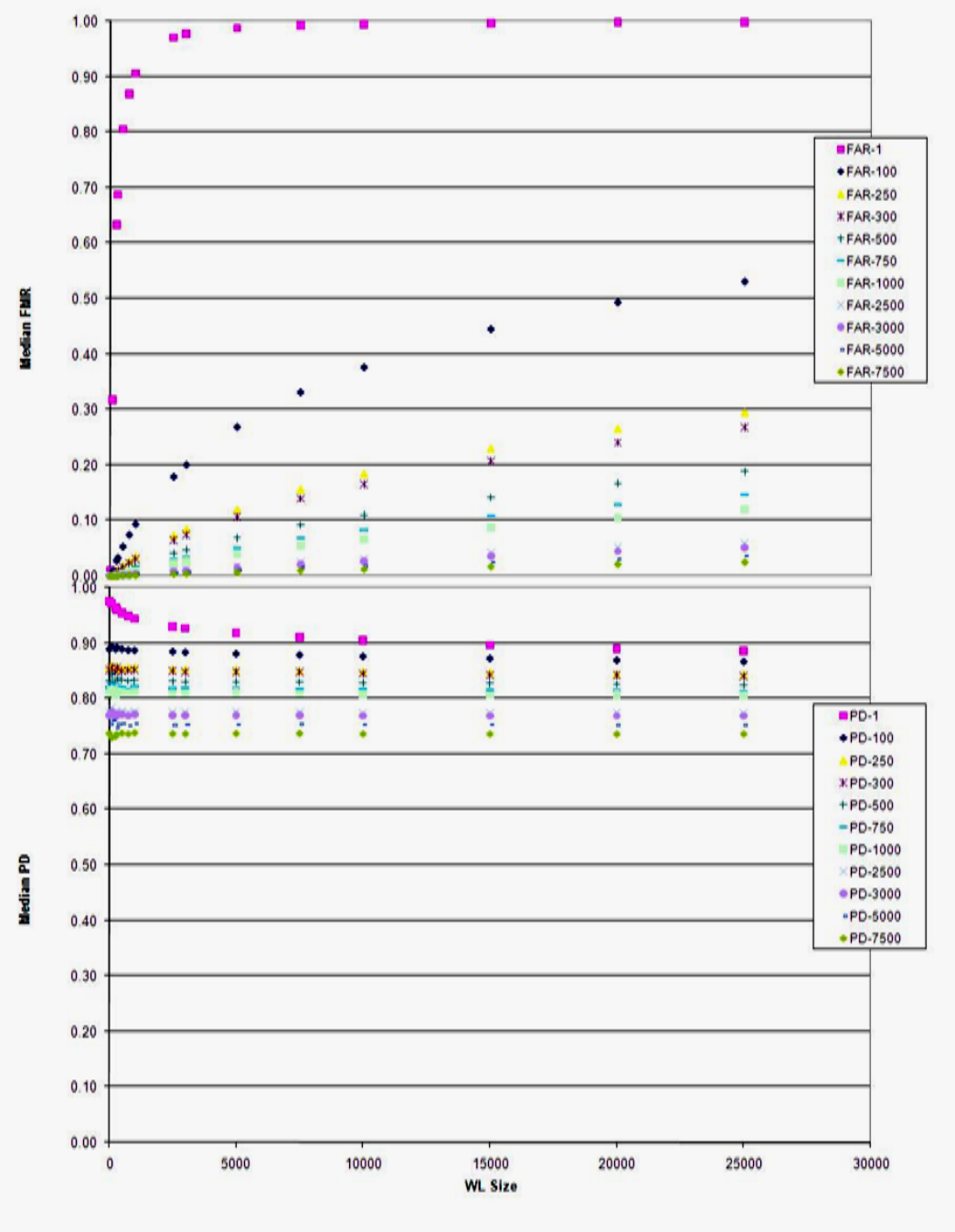
i→I Experiment



- Ran trials with 10,000 imposters (rose) against different size galleries (blue)
  - Varied gallery size
  - Top match retrieved from similarity scores
  - Repeated 100 times

i→I Experiment

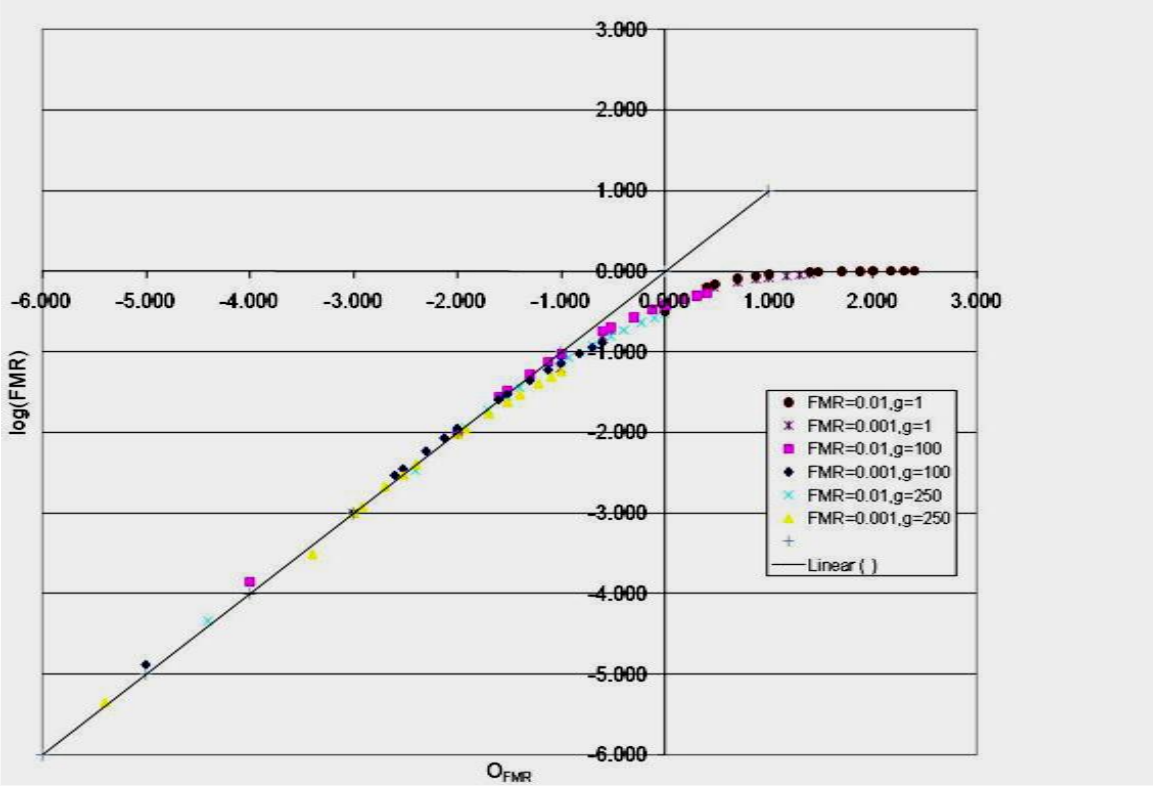
- For each gallery size *g*, there are 100 similarity scores corresponding to the tested FMR level FMR_{Λ} (0.01 or 0.001)
  - Median similarity score assigned to $\tilde{T}_g^{FAR_{\Lambda}}$
- For every *g* for FMR_{Λ} , $\tilde{T}_g^{FAR_{\Lambda}}$ was reapplied to each trial matrix to determine the FMR and P_D
  - Varied scenario watchlist size from 1 to 25,000


Median FMR
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
FAR-1
FAR-100
FAR-250
FAR-300
FAR-500
FAR-750
FAR-1000
FAR-2500
FAR-3000
FAR-5000
FAR-7500
Median PD
PD-1
PD-100
PD-250
PD-300
PD-500
PD-750
PD-1000
PD-2500
PD-3000
PD-5000
PD-7500
0
5000
10000
15000
20000
25000
30000
WL Size

i→I Results

- When applying the median threshold for FMR-level 0.01 from experiments of WL=100 to WL=10,000, the observed-median-FMR is 0.38.
  - 38 times greater FMR for change in gallery size of 2 orders of magnitude.
  - For WL = 25,000, applying the median threshold when g = 100, FMR=0.53.
  - For WL = 25,000, applying the median threshold when g = 1, FMR is nearly 100%.
- When applying the median threshold for FMR-level 0.001 from experiments of WL=100 to WL=10,000, the observed-median-FMR is 0.07.
  - 70 times greater FMR for change in gallery size of 2 orders of magnitude.
  - For WL=25,000, applying the median threshold when g = 100, FMR=0.13.
  - For WL = 25,000, applying the median threshold when g = 1, FMR=0.91.

Model for FMR v. FMR Results



$$E\left[\mathrm{E}_{FMR}\right] \approx \frac{WL_O}{g_\Lambda} FMR_\Lambda \quad \text{when } E[\mathrm{E}_{FMR}] \leq 0.10$$

FMR levels to test to determine operating points for scenarios

Scenario Name:	**Personal Security**		**Forensics**		**Watchlist**		**Large-Scale ID**	
Operational FMR:	0.01		0.0001		1.E-07		1.E-07	
Min|Max E[Subjects]	1	10	1000	1E+09	10	1.E+05	1.E+07	1E+09
g (Subjects)								
200								
400								
36,000								
60,000								

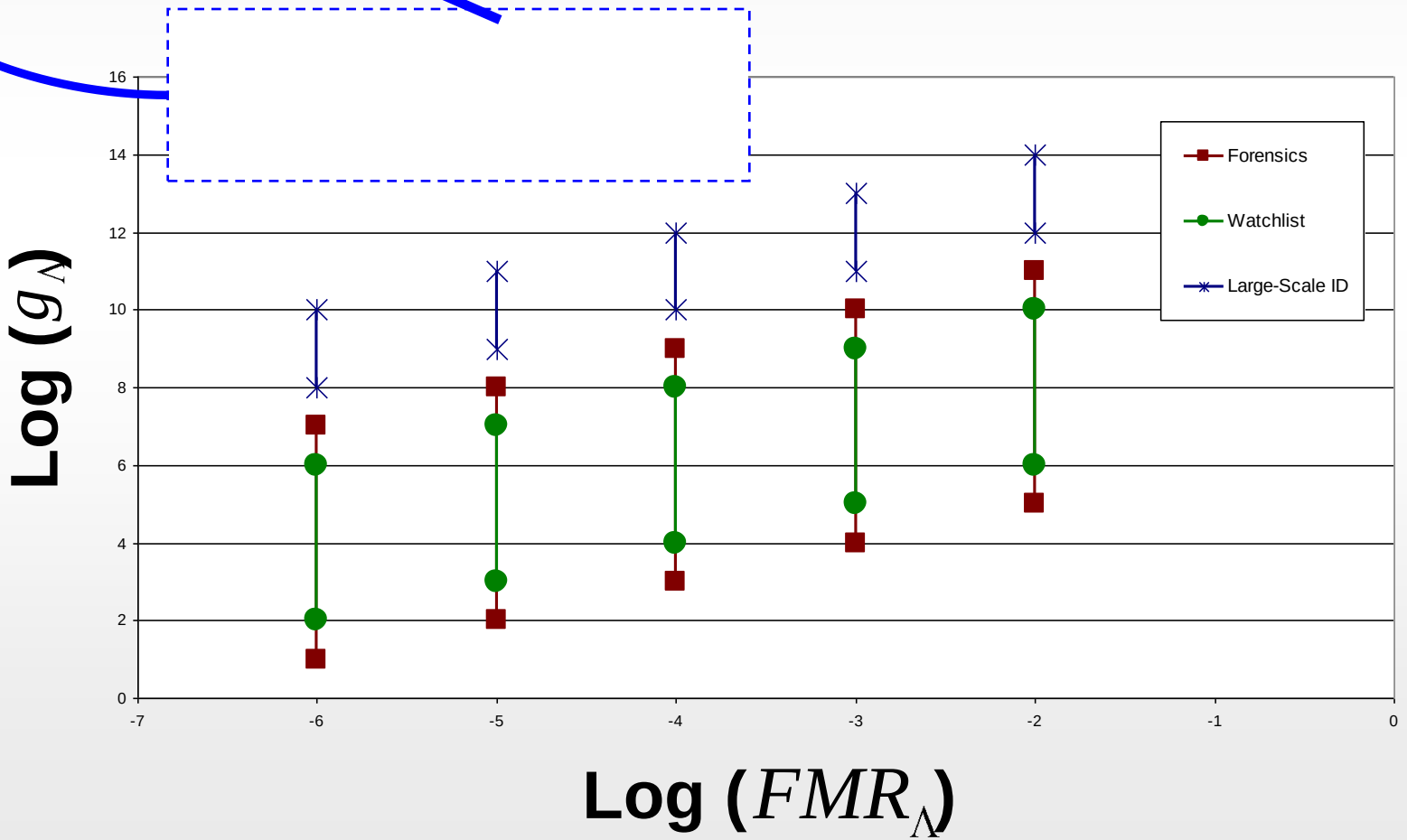
Red = FMR-level can not be tested for this image set size.

Yellow = FMR-level may testable with image set size, but may still be weak with respect to statistical significance.

Green = FMR-level can be tested with this image set size.

Gallery size to test at tested FMR level, for scenarios operating points

Scenario Name:	**Forensics**		**Watchlist**		**Large-Scale ID**	
Operational FMR:	0.0001		1.E-07		1.E-07	
Min|Max E[Subjects]	1000	1E+09	10	1.E+05	1.E+07	1E+09



Conclusions (1 of 3)

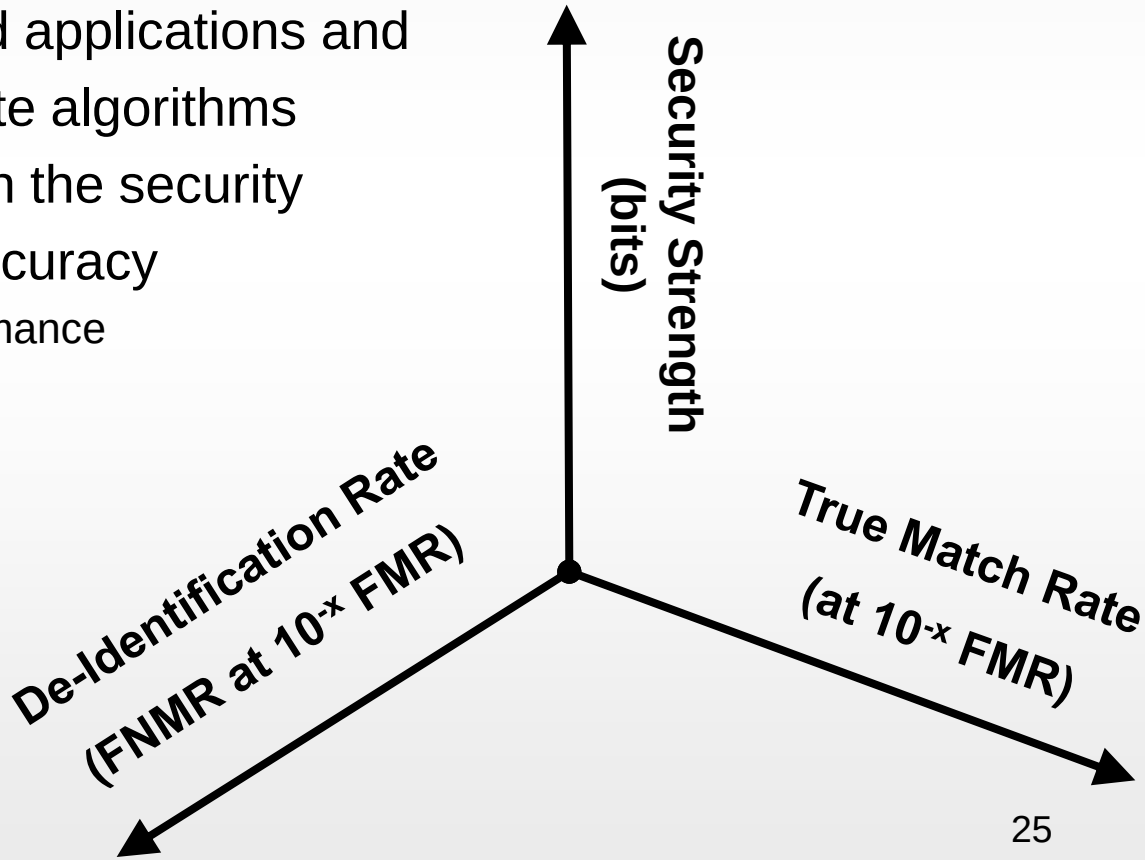
- Major gap in testing biometric systems for larger scale applications
  - largest tests performed on <200,000. Most in verification mode.
- Model for testing shows that
  - Gaps can lead to massive number of false matches.
  - Testing needs to be performed on size relevant to scenario for a reasonable prediction on FMR.
  - FNMR is relatively stable, regardless of gallery/watchlist size.

Conclusions (2 of 3)

- The FMR levels that need to be tested for several common applications can not be supported by the size of tests today.
- To adequately test these, the number of subjects needed exceeds current database sizes, and it would be prohibitively costly to collect for R&D.
- Use of tests for comparison of alternatives on smaller test sets can still be useful.

Conclusions (3 of 3)

- Next steps for template protection testing will need to address
  - scale for intended applications and
  - metrics to evaluate algorithms incorporating both the security properties and accuracy
    - Biometric Performance
    - De-Identification
    - Irreversibility
    - Others



Thank you