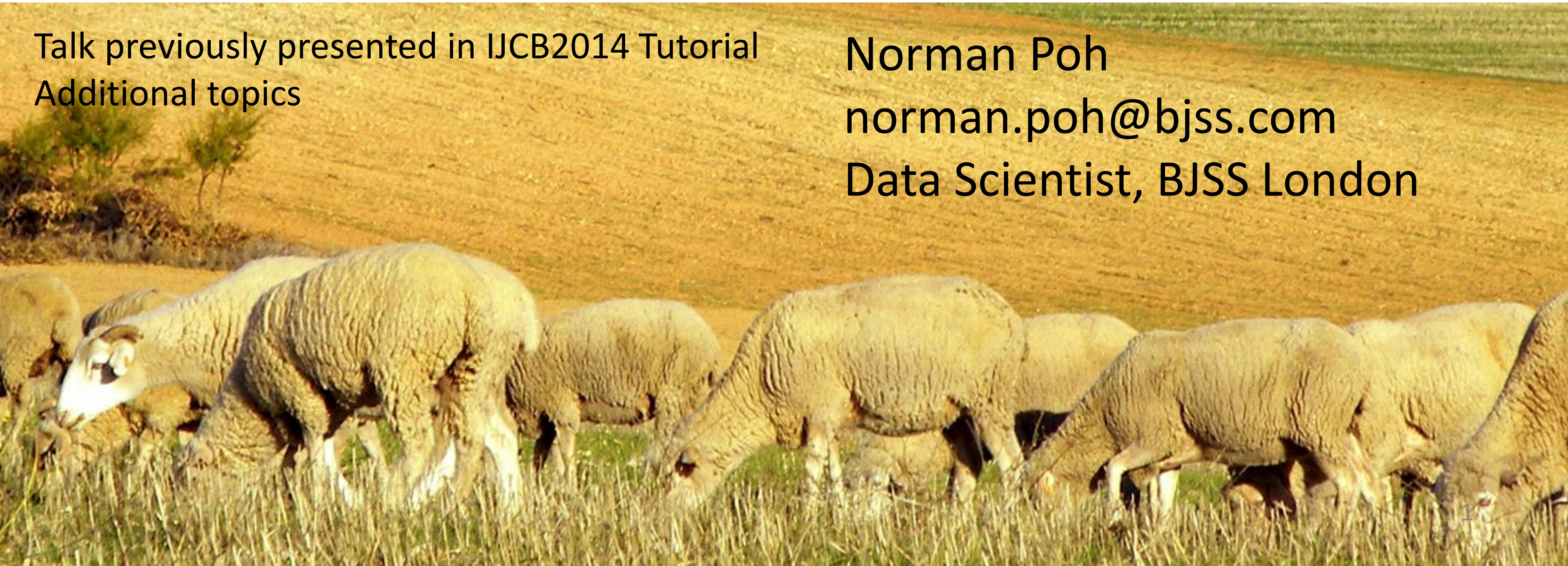
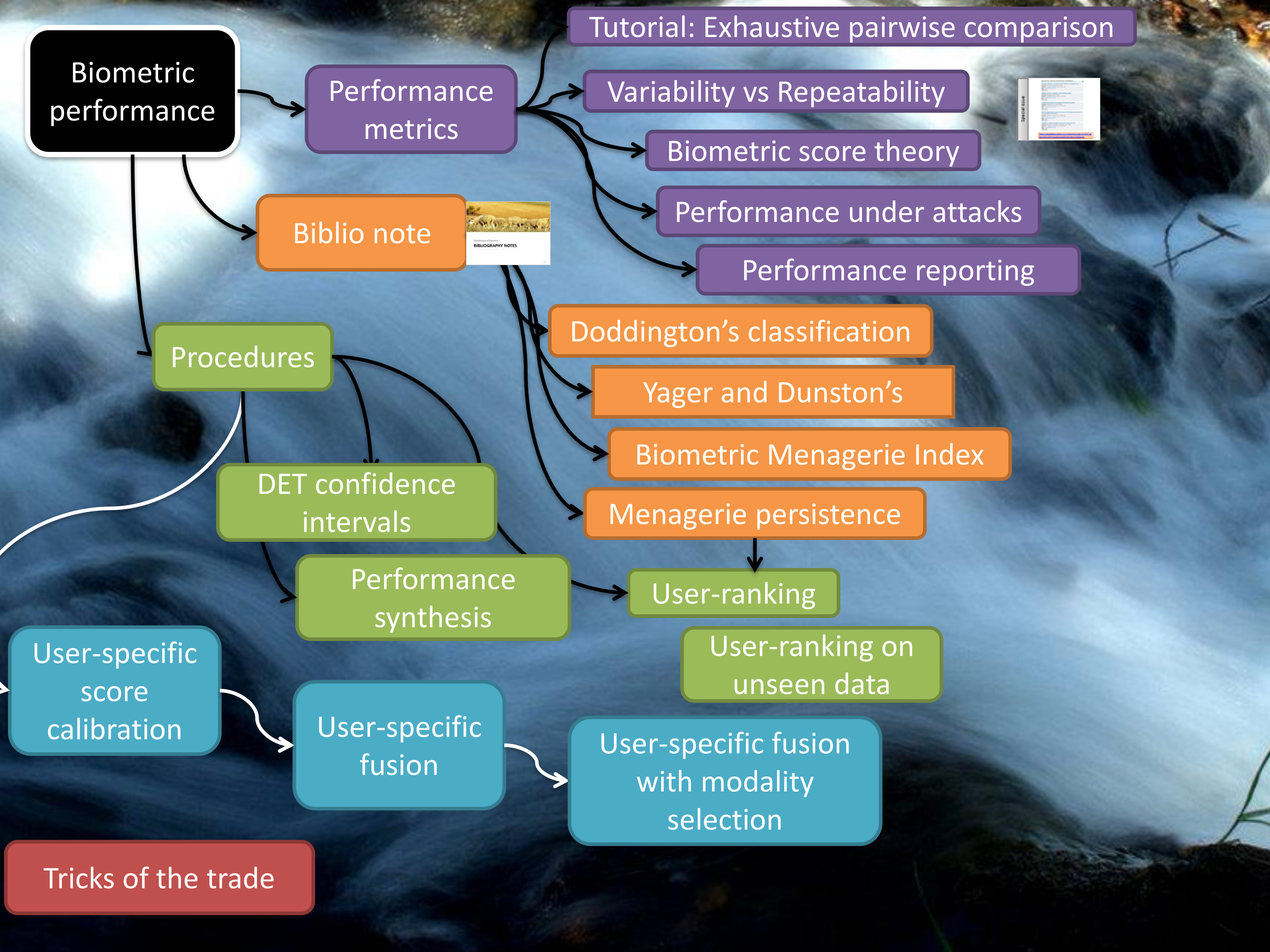
Biometric Performance and Its Optimal Calibration



Talk previously presented in IJCB2014 Tutorial
Additional topics

Norman Poh
norman.poh@bjss.com
Data Scientist, BJSS London

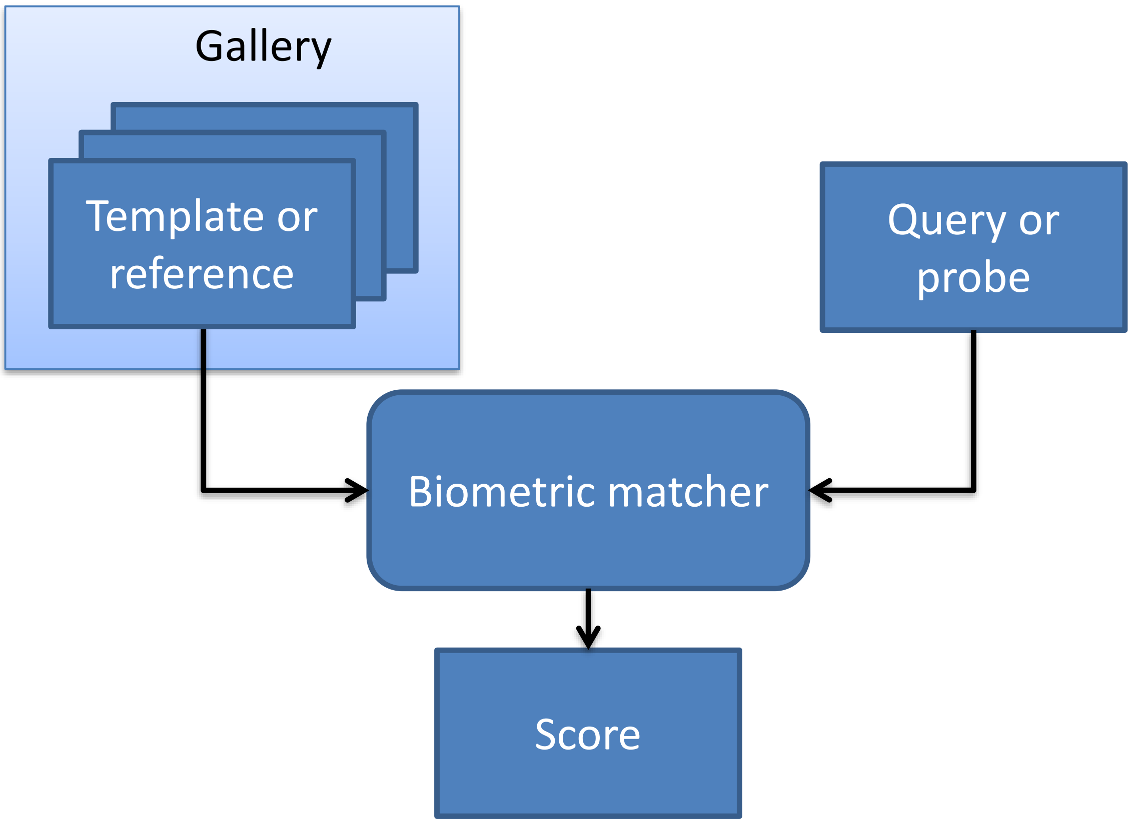

Biometric performance
Performance metrics
Tutorial: Exhaustive pairwise comparison
Variability vs Repeatability
Biometric score theory
Performance under attacks
Performance reporting
Biblio note
Doddington's classification
Yager and Dunston's
Biometric Menagerie Index
Menagerie persistence
Procedures
DET confidence intervals
Performance synthesis
User-ranking
User-ranking on unseen data
User-specific score calibration
User-specific fusion
User-specific fusion with modality selection
Tricks of the trade

Thank you

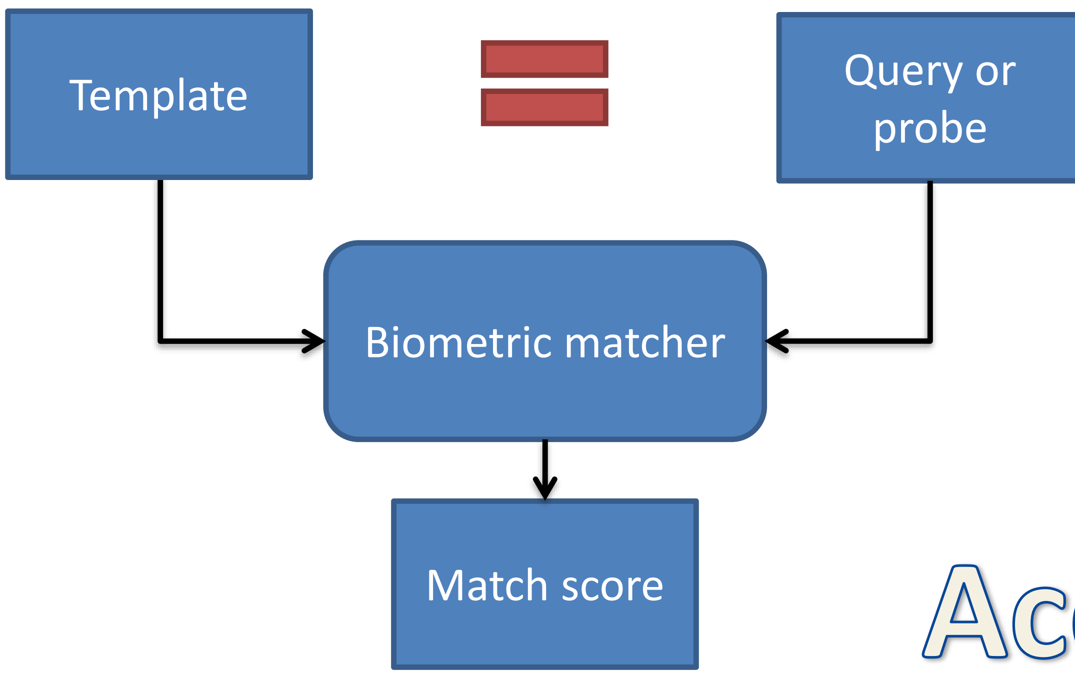
n.poh@surey.ac.uk

TERMS AND PERFORMANCE METRICS

Some basic terms



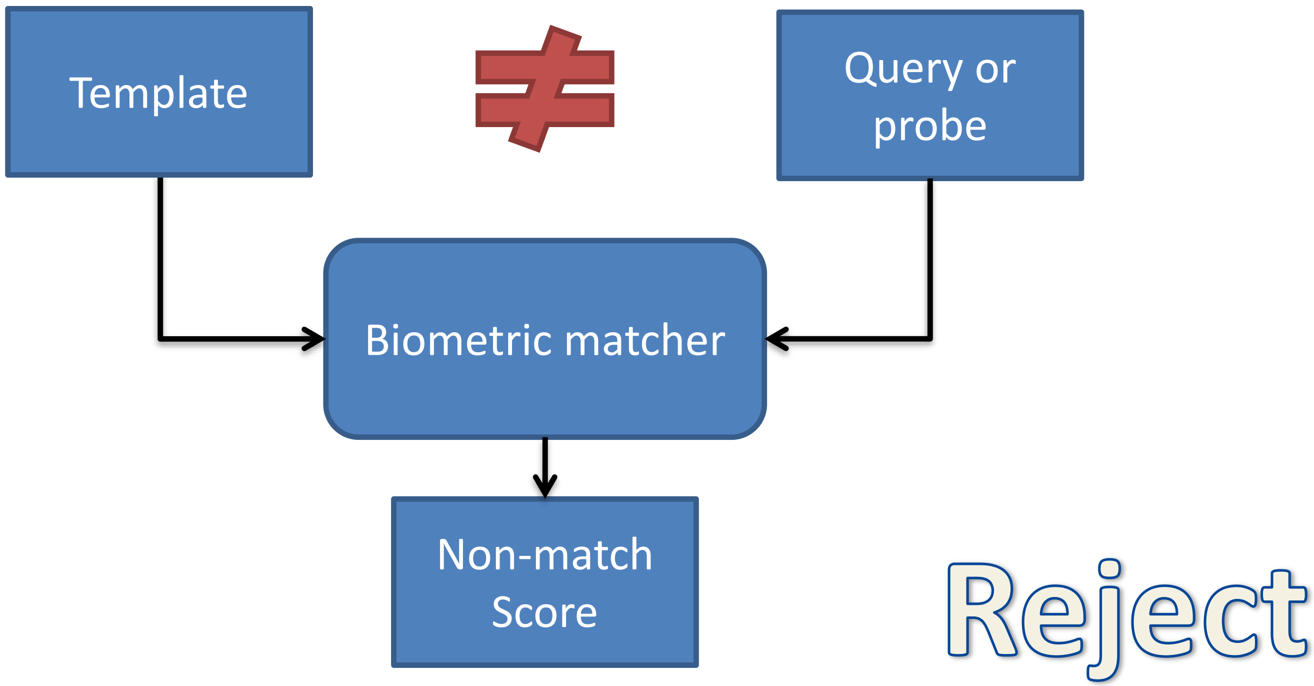
Genuine score



Accept

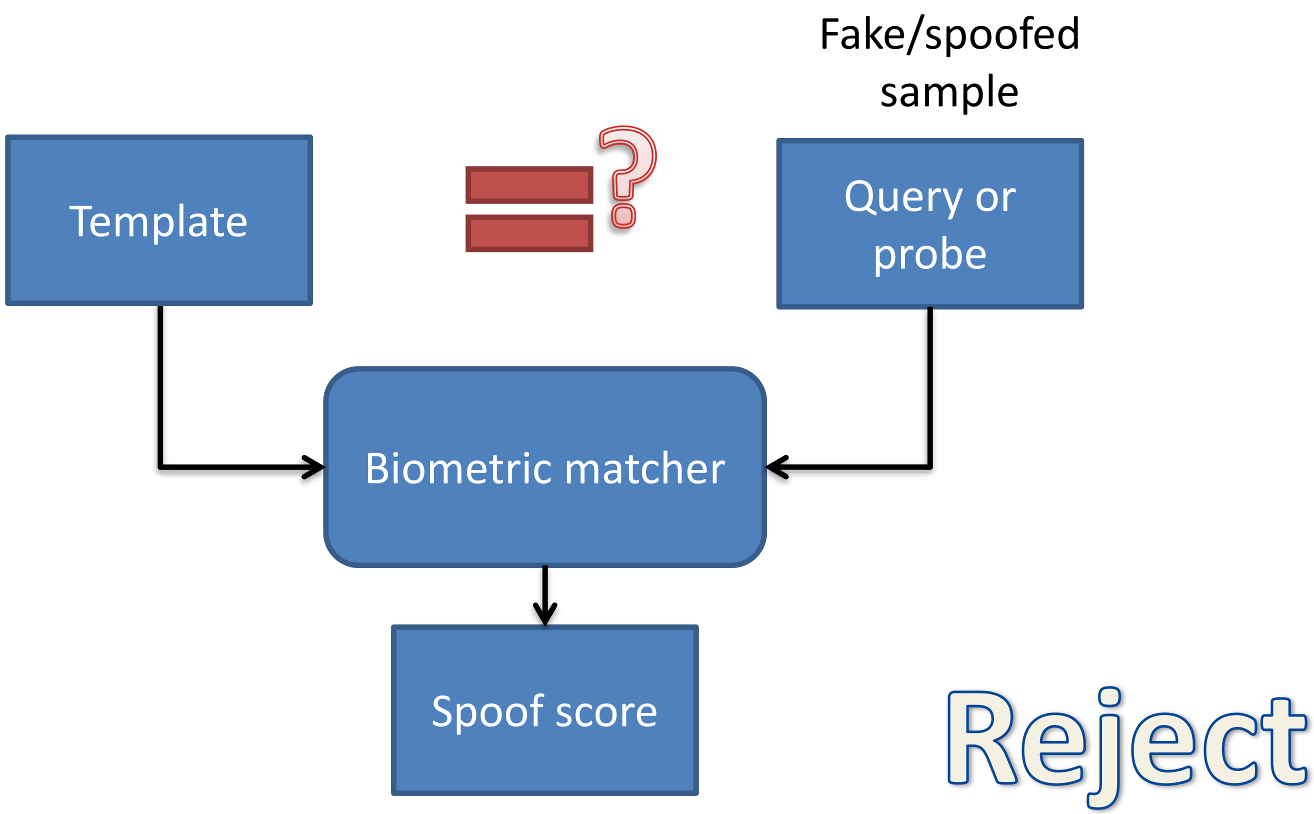
Genuine score
(Mated score)

Zero-effort impostor score



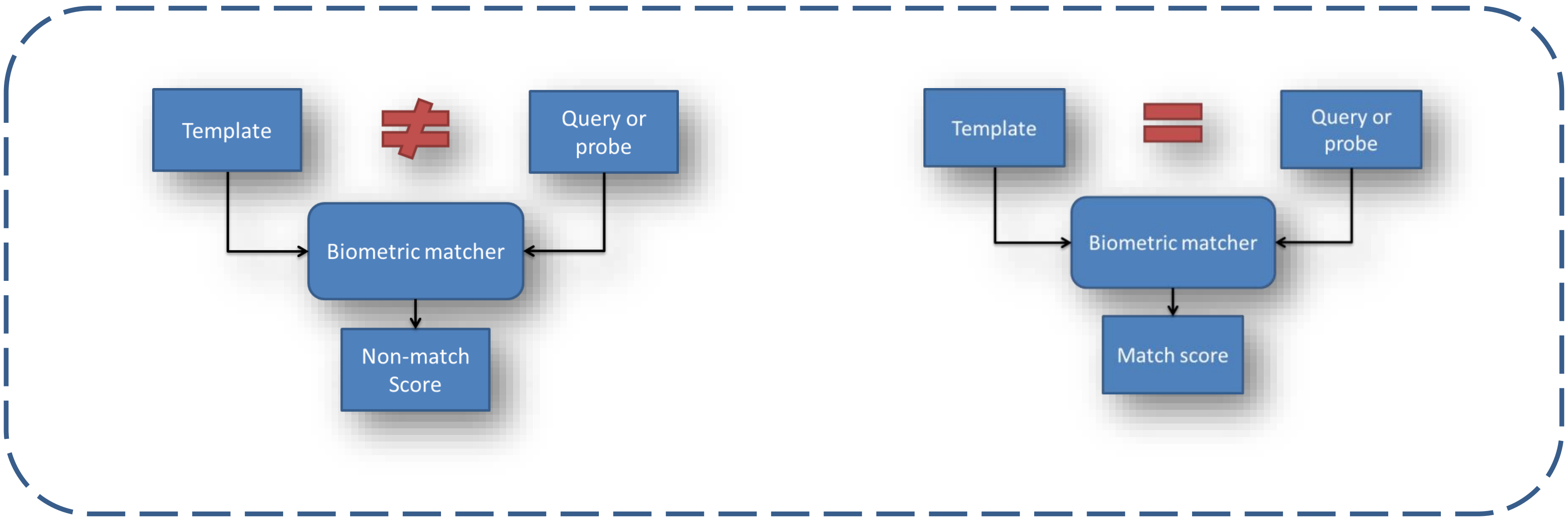
Zero-effort impostor score
(Non-mated score)

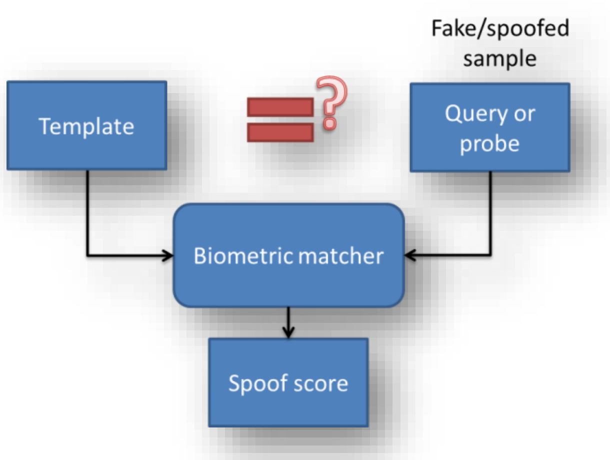
Nonzero-effort impostor score



Nonzero-effort impostor score

Biometric Menagerie – scope



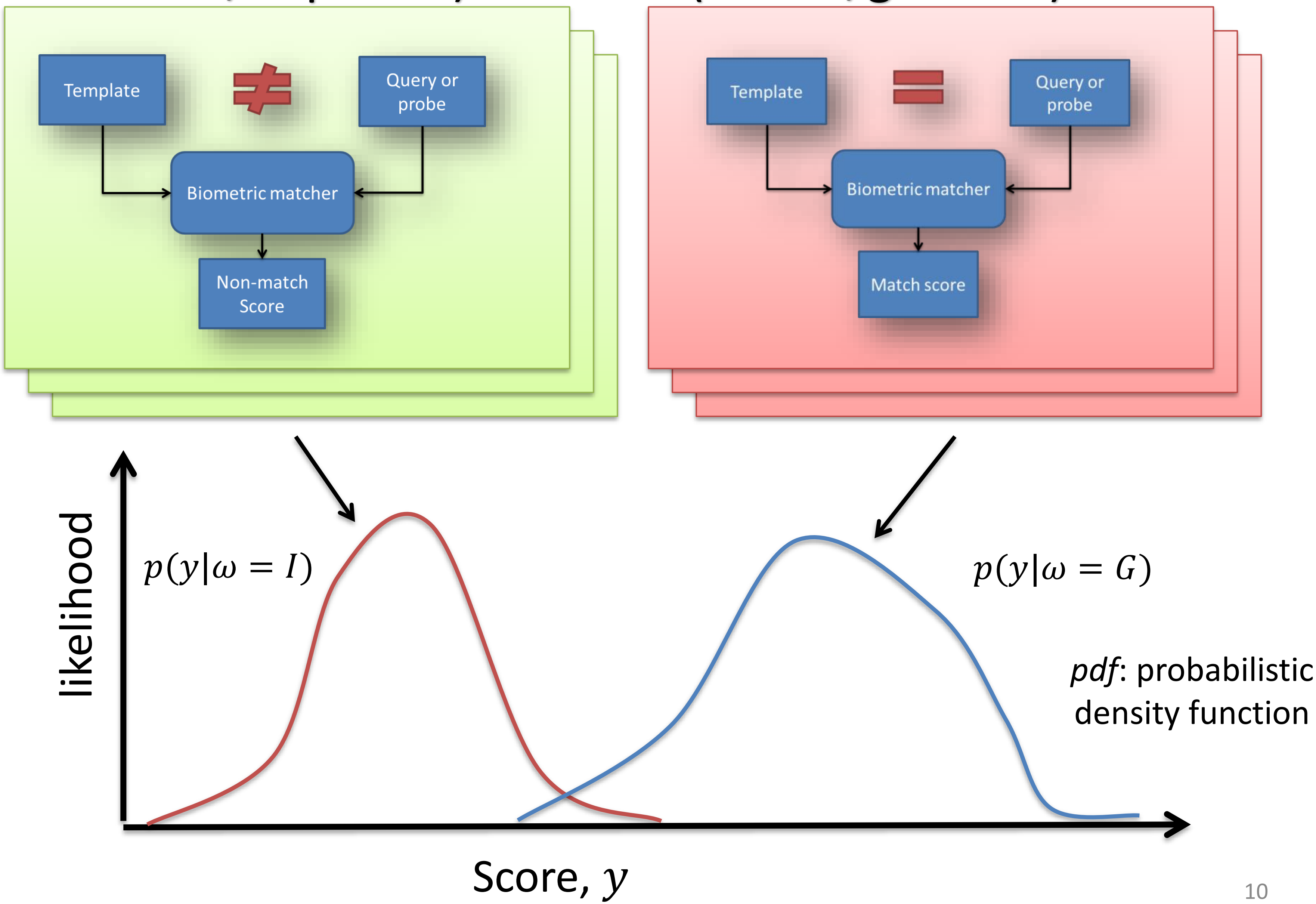


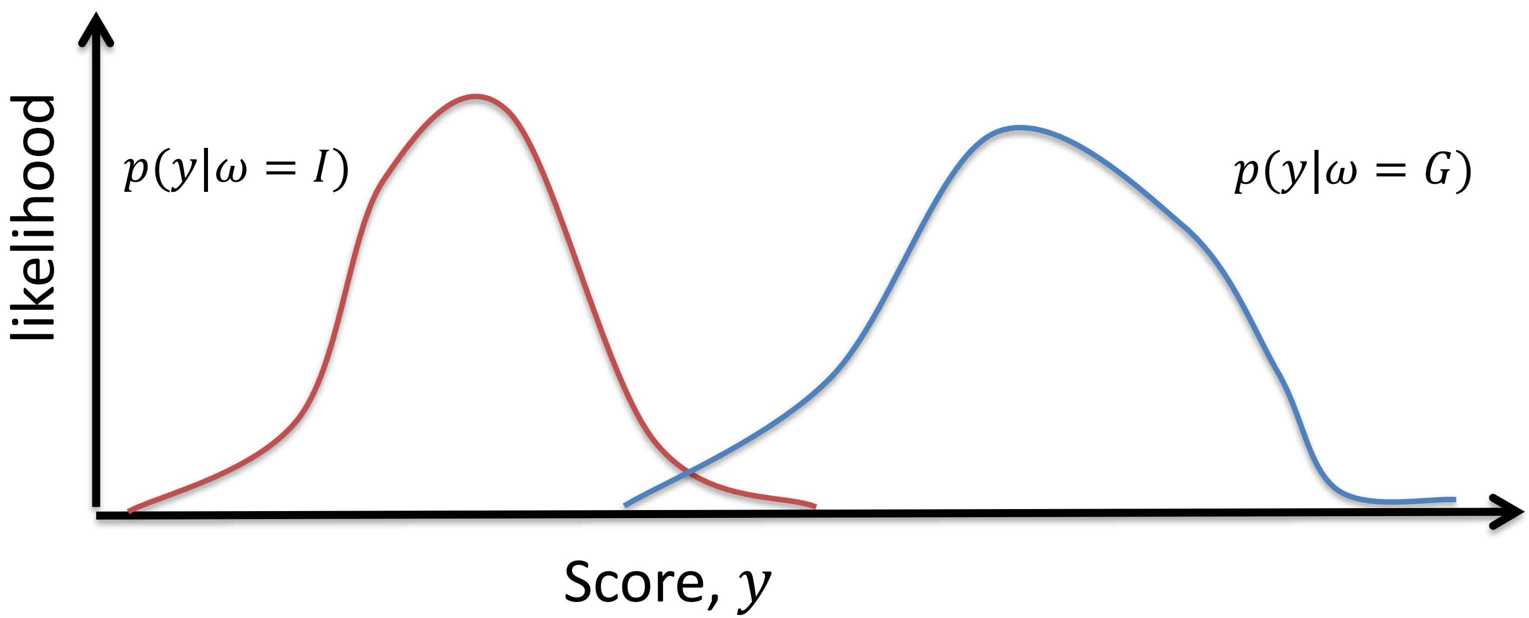
Future research

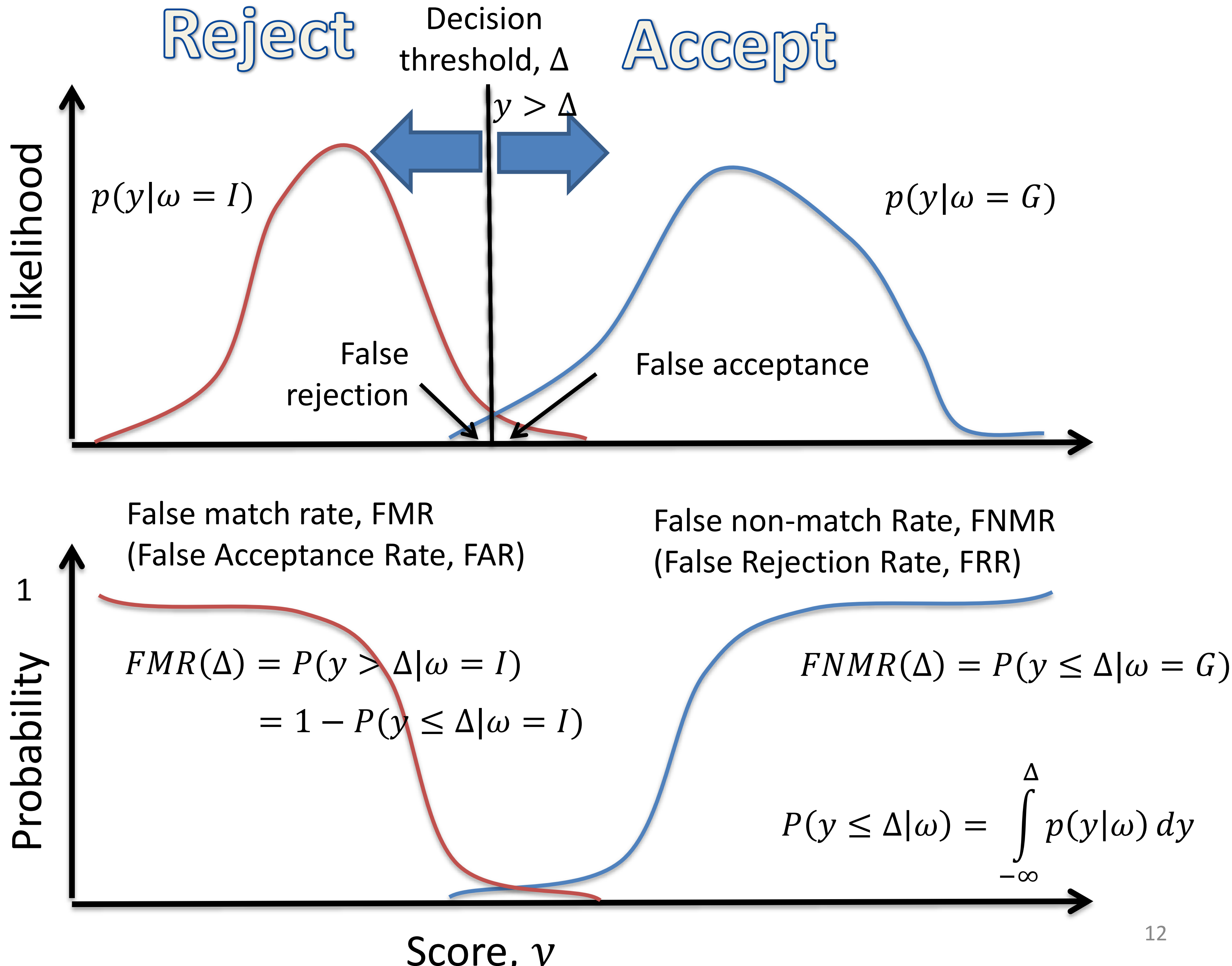
- Biometric antispoofing
- biometric presentation attack and countermeasures

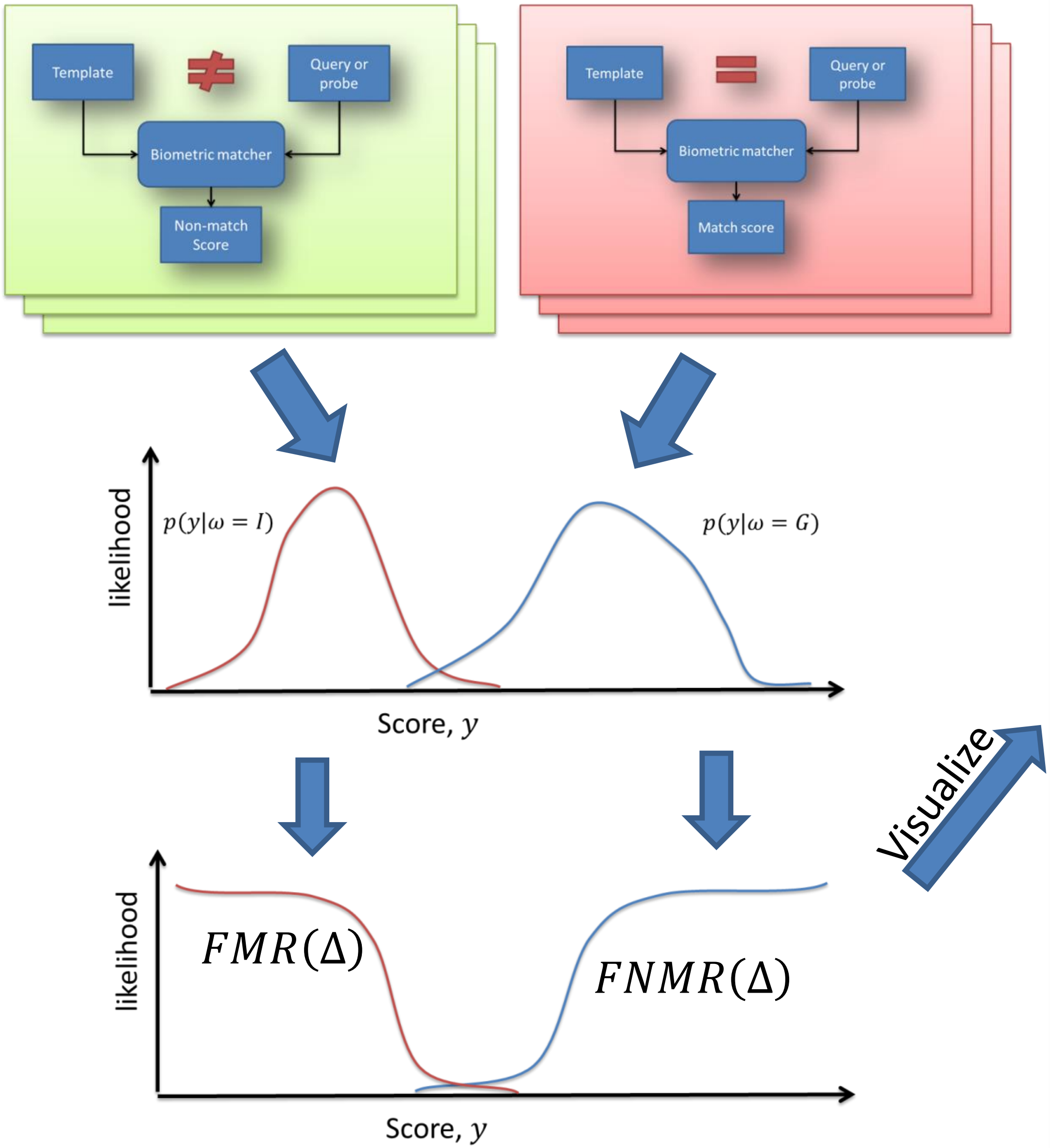
Different-pair (nonmatch/impostor) scores

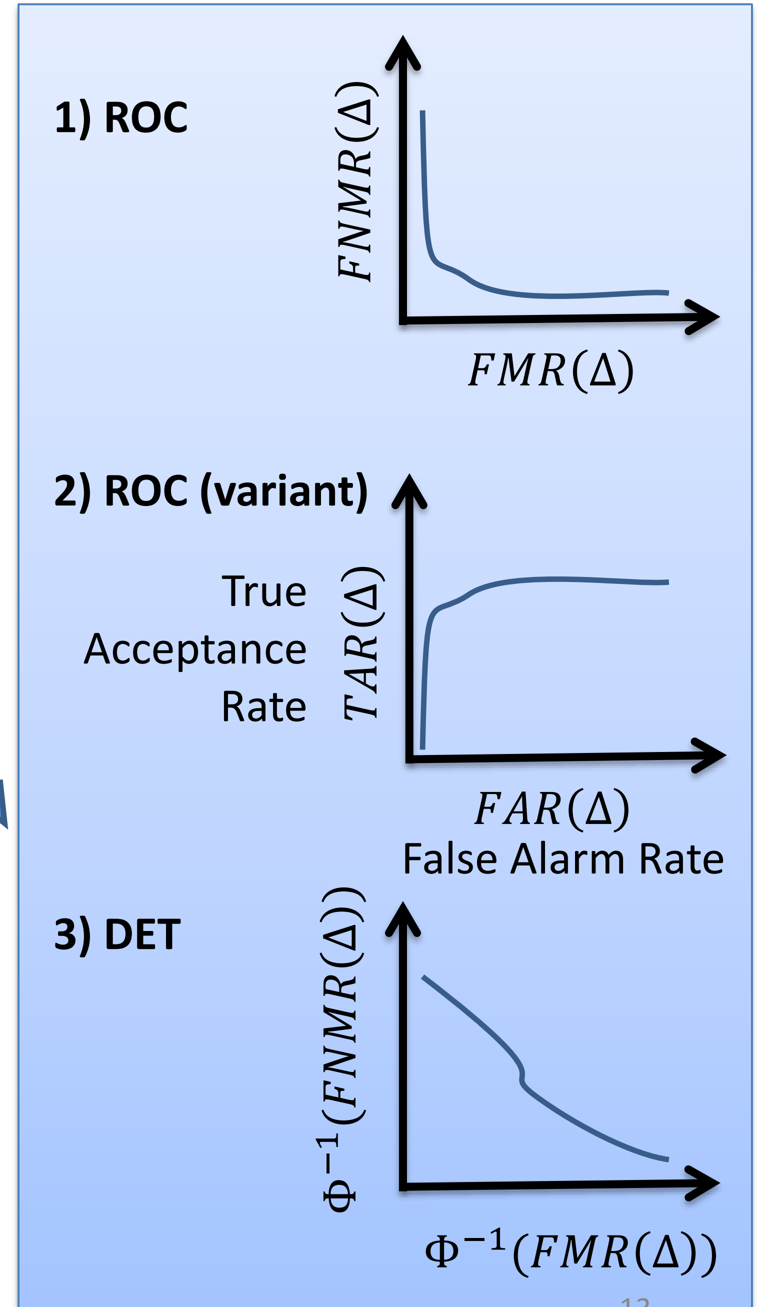
Same-pair (match/genuine) scores




likelihood
$p(y|\omega = I)$
$p(y|\omega = G)$
Score, y


Reject
Decision threshold, Δ
Accept
$y > \Delta$
likelihood
$p(y|\omega = I)$
$p(y|\omega = G)$
False rejection
False acceptance
False match rate, FMR
(False Acceptance Rate, FAR)
False non-match Rate, FNMR
(False Rejection Rate, FRR)
1
Probability
$FMR(\Delta) = P(y > \Delta|\omega = I)$
$= 1 - P(y \leq \Delta|\omega = I)$
$FNMR(\Delta) = P(y \leq \Delta|\omega = G)$
$P(y \leq \Delta|\omega) = \int_{-\infty}^{\Delta} p(y|\omega)\, dy$
Score, y

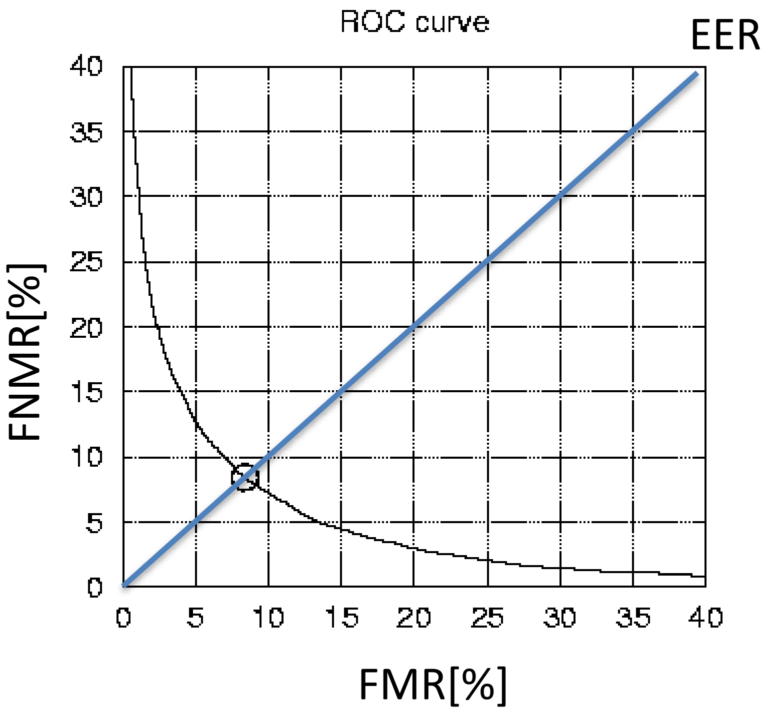




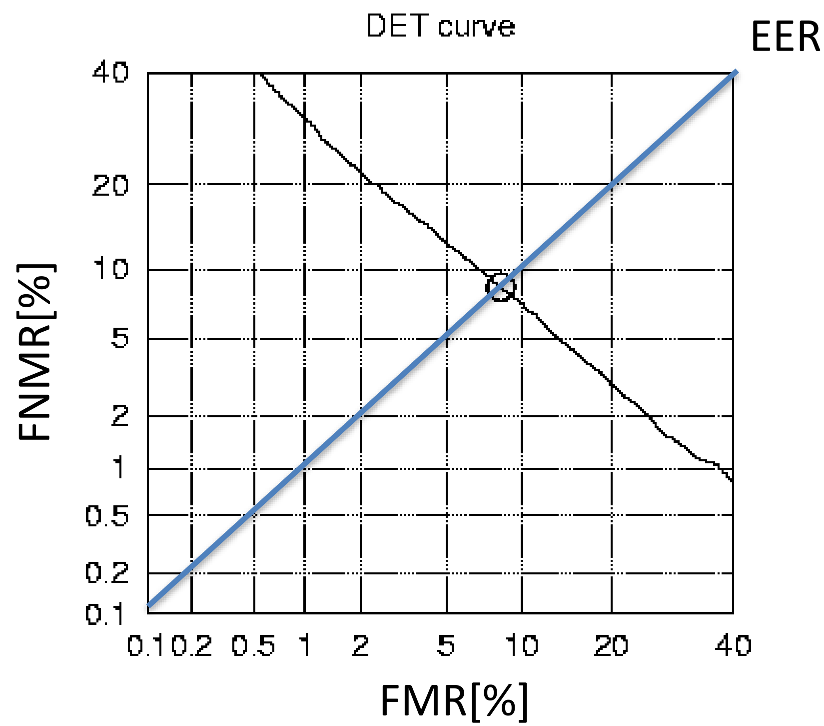
Note: Φ is the PDF of a normal distribution; $\Phi^{-1}$is its inverse

ROC versus DET

Receiver's Operating Characteristic (ROC) curve



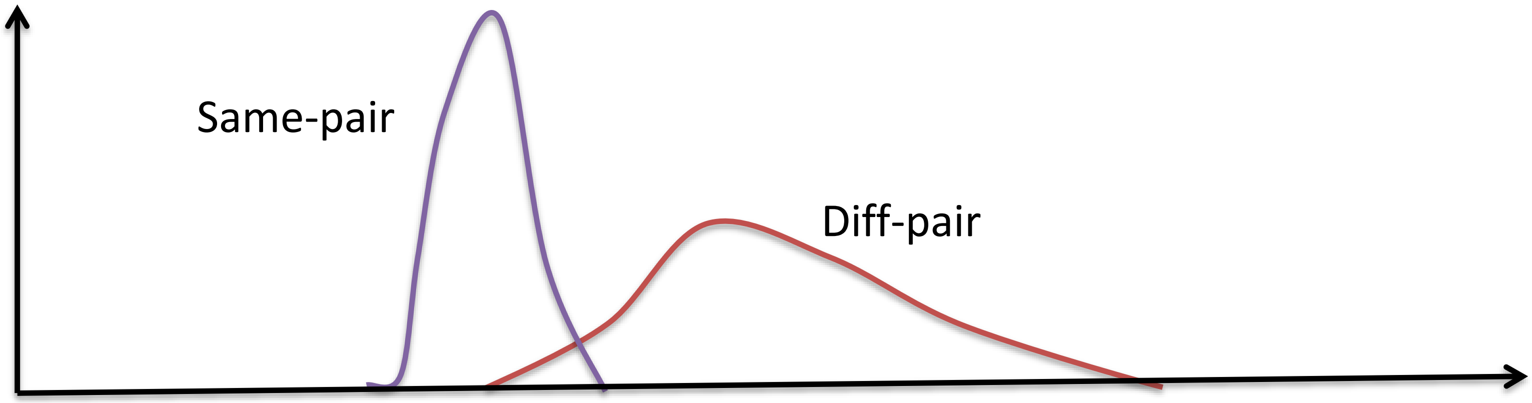
Detection Error Trade-off (DET) curve



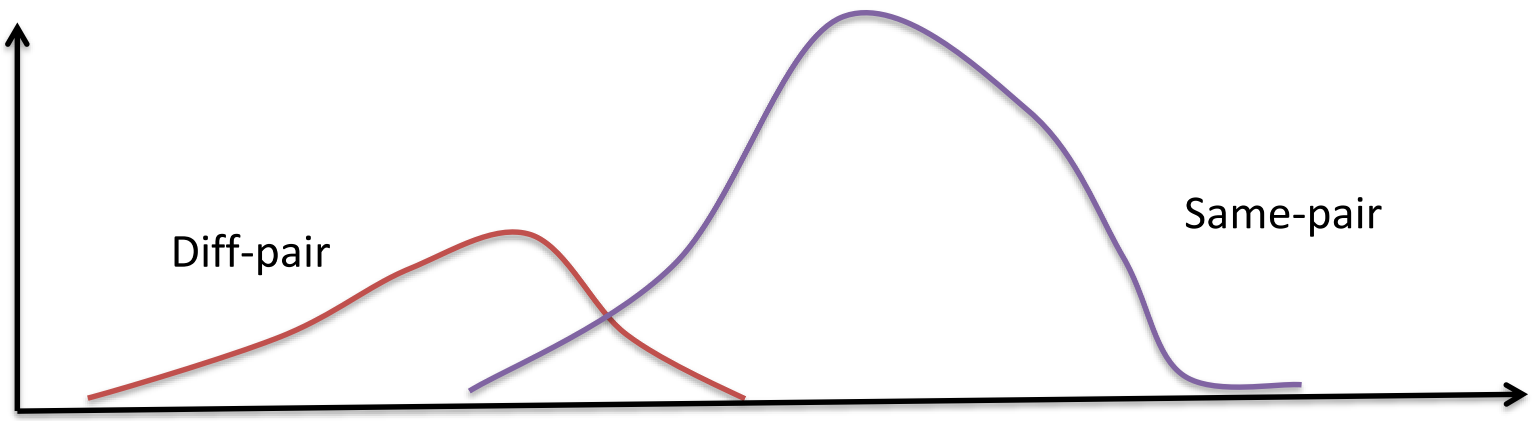
EER: Equal Error Rate or Cross-over error rate

Further reading: Biometric Testing and Statistics
http://www.biometrics.gov/Documents/BioTestingAndStats.pdf

Side note: Similarity vs dissimilarity scores



Distance (dissimilarity) scores

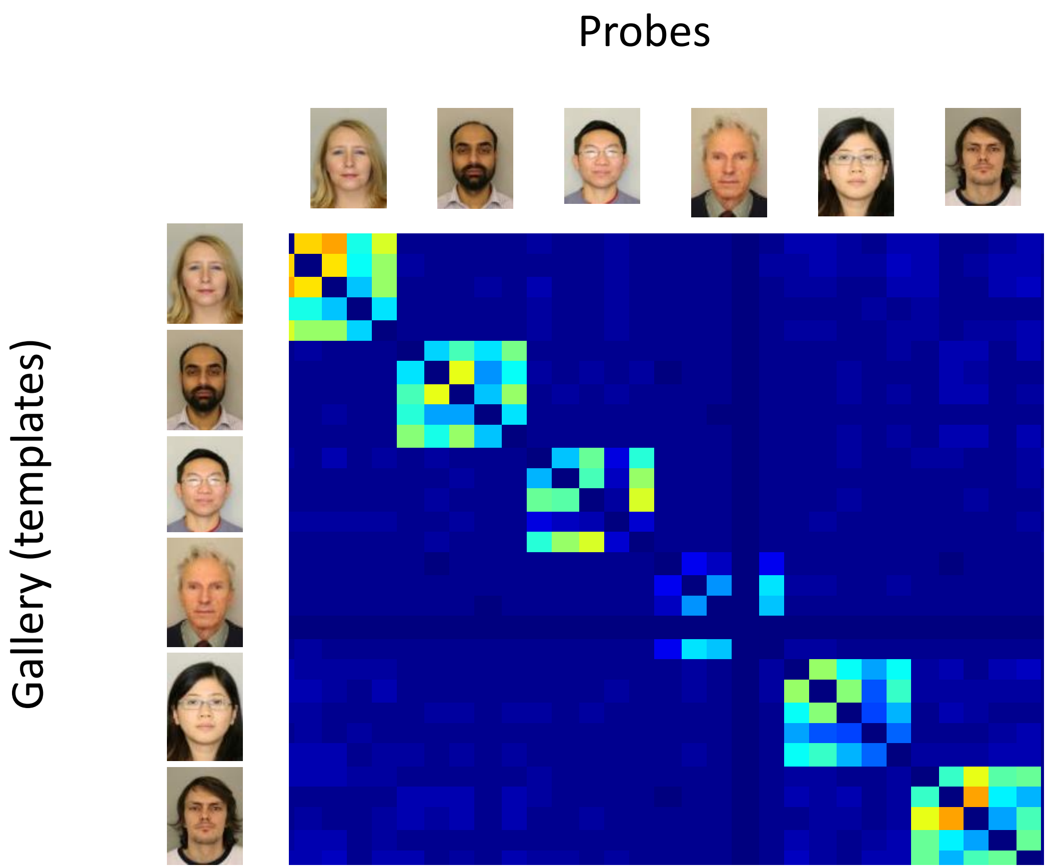


Similarity scores

Menu

TUTORIAL: EXHAUSTIVE PAIRWISE COMPARISON EXPERIMENT

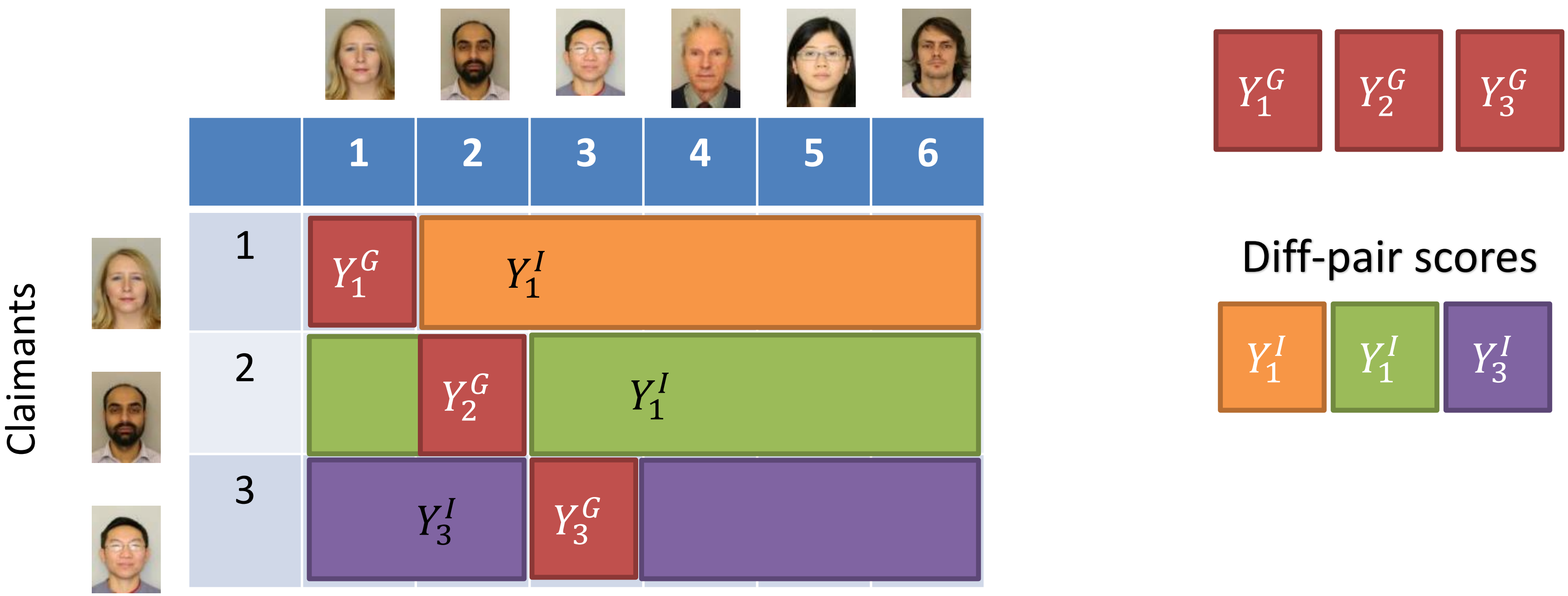
Generating an exhaustive pairwise comparison experiments



What's the difference between a wolf and a lamb???

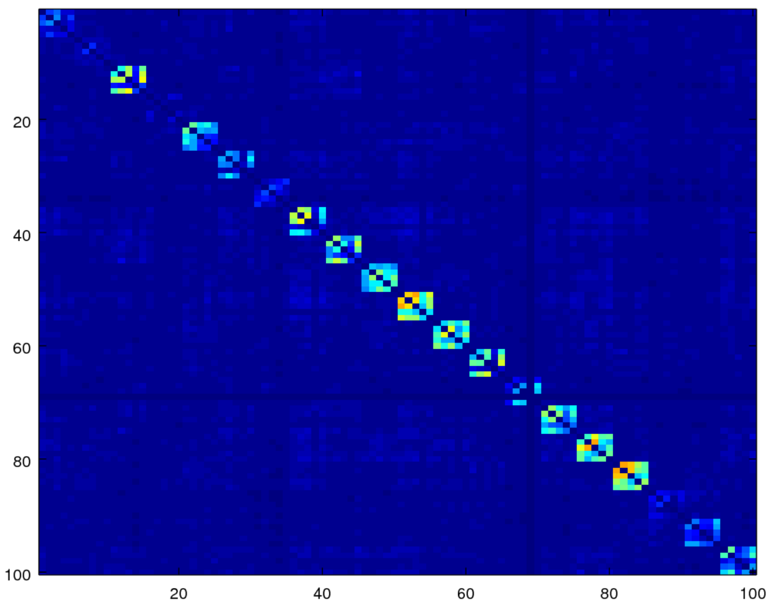
Each subject may have multiple query images

Same-pair scores

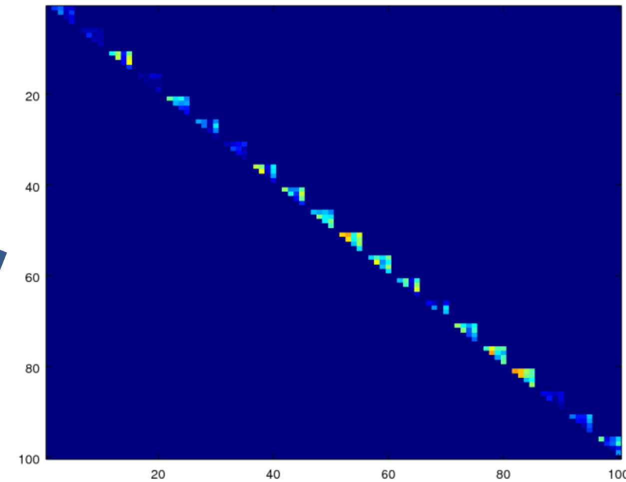


(1 template each)

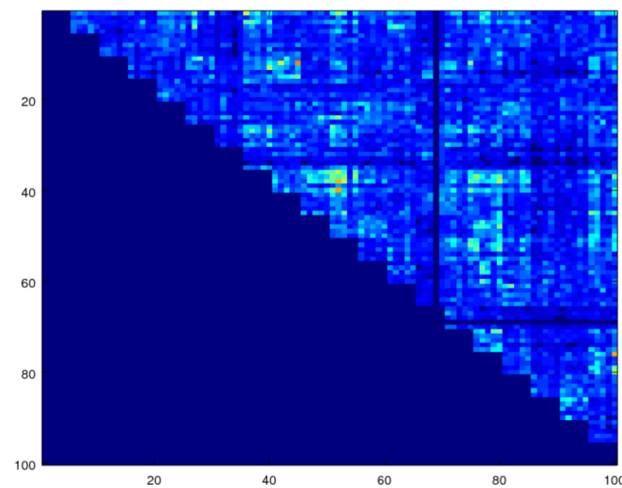
Exhaustive pair-wise comparison



Same-pair (match) scores

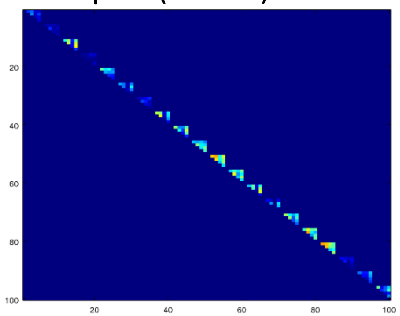


Different-pair (nonmatch) scores

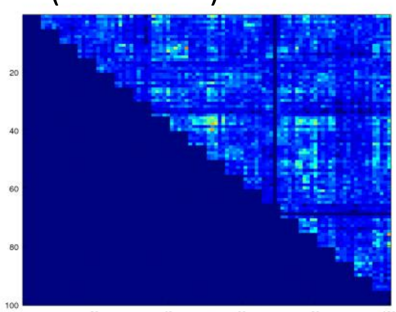


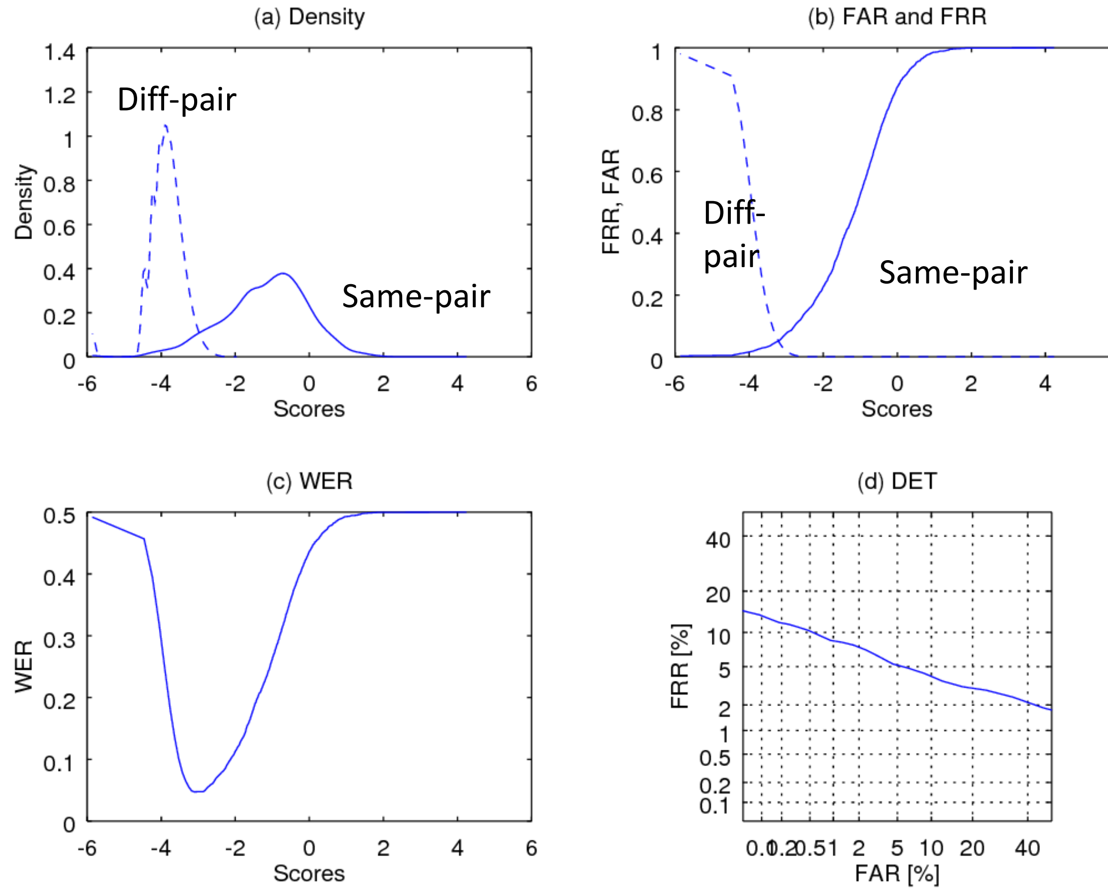
https://normanpoh.github.io/blog/2017/12/29/generate-pairwise-fprint-scores.html

Same-pair (match) scores

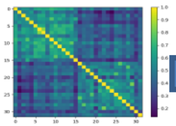


Different-pair (nonmatch) scores

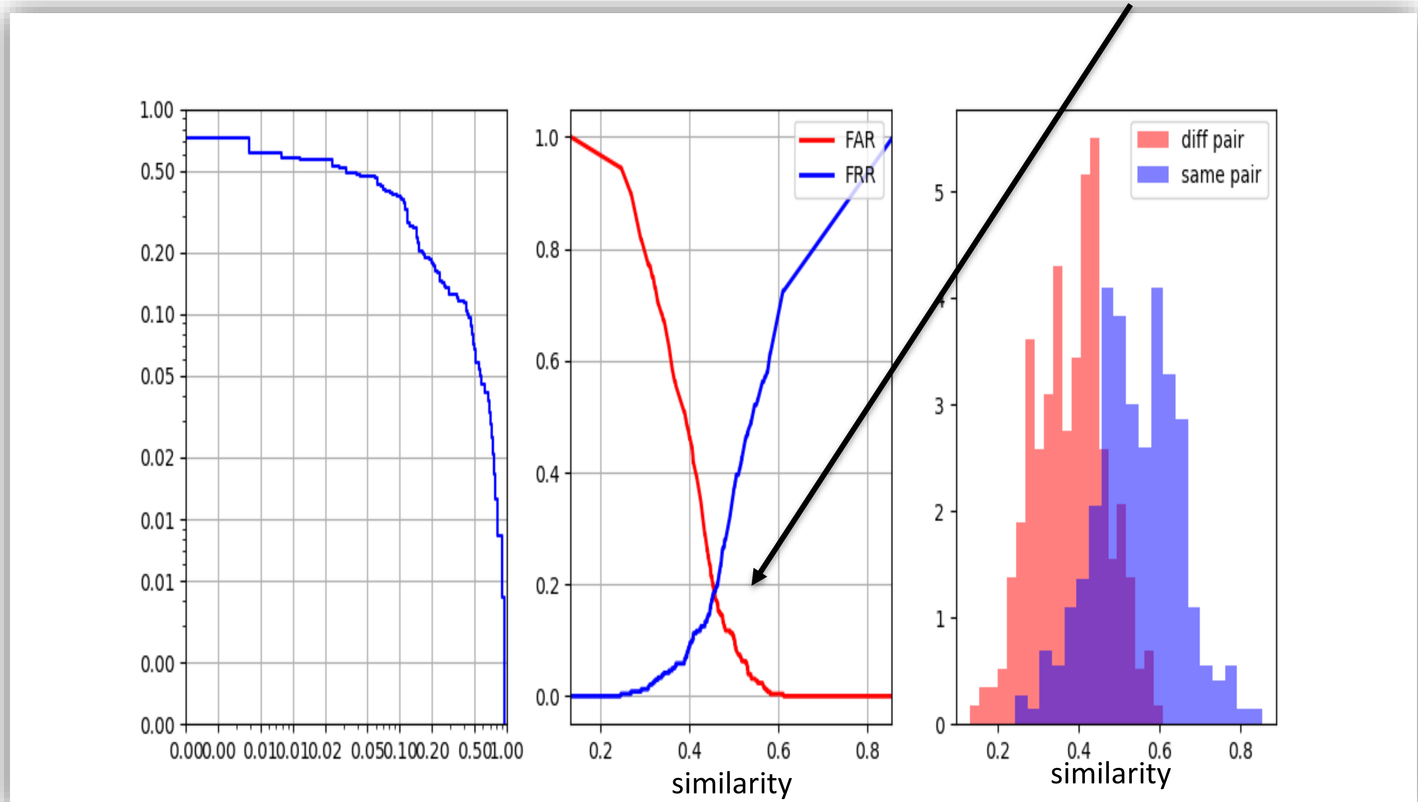




Binary classification



Cross-over operating point
Equal error rate (EER)



Menu

PERFORMANCE REPORTING

Performance Reporting

Security applications:

Fix FAR to 1 in a million, report performance on FRR (security application)

Convenient applications:

Fix FRR to 1 in 100K, report performance on FAR

Equal:

FAR and FRR

$$HTER(\Delta) = \frac{1}{2}FAR(\Delta) + \frac{1}{2}FRR(\Delta)$$

Weighted:

$$WER(\Delta) = \beta FAR(\Delta) + (1-\beta)FRR(\Delta)$$

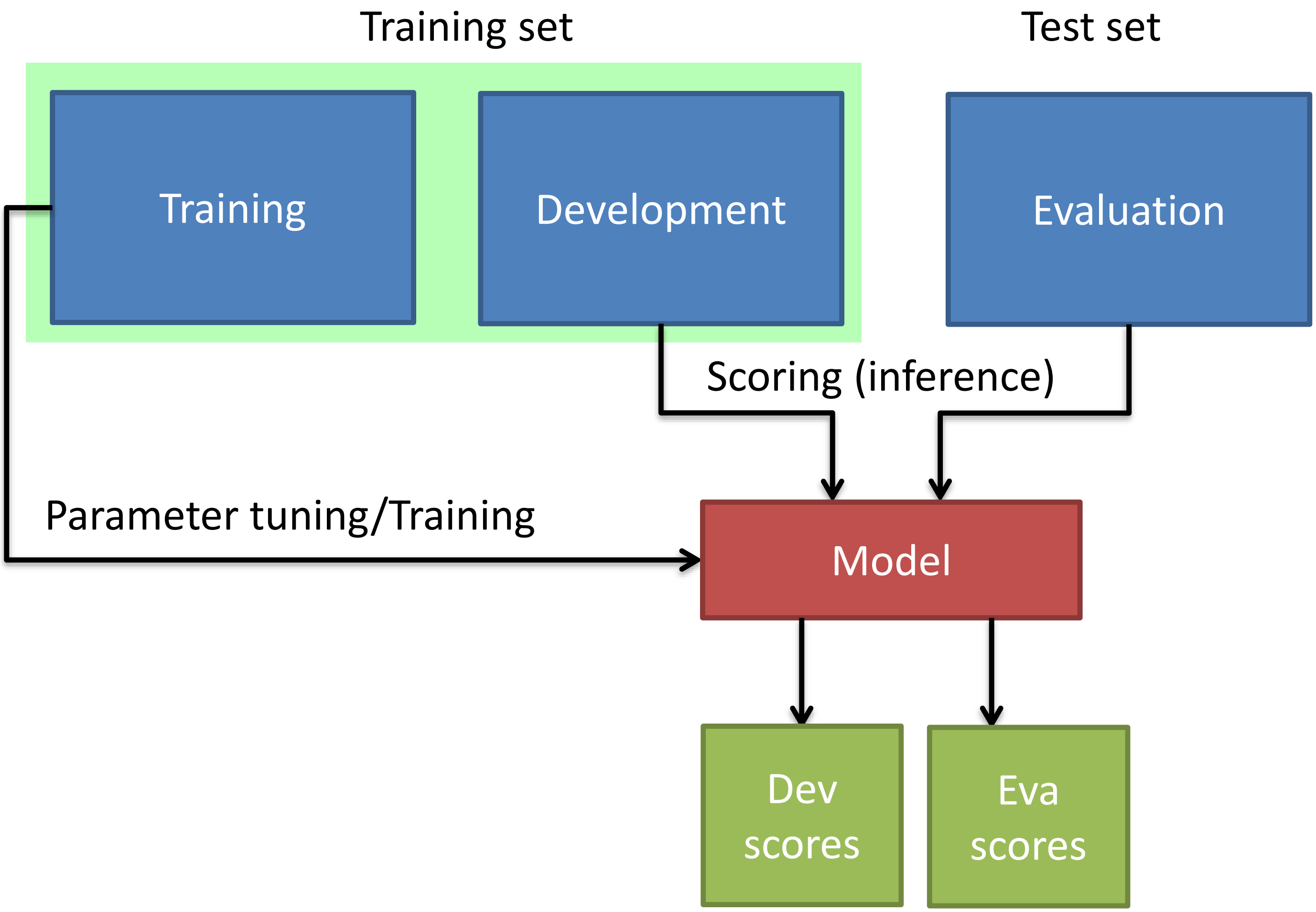
Security applications

FAR	FRR
1 in 10^4	
1 in 10^5	
1 in 10^6	

Convenient applications

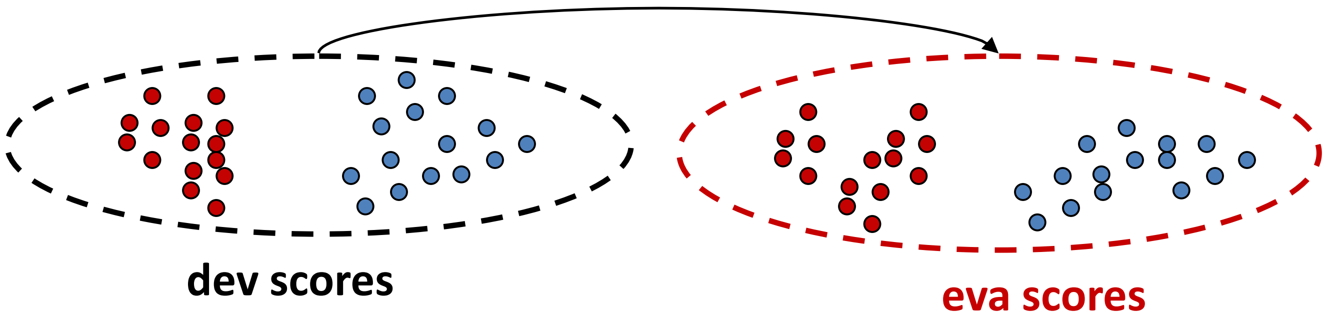
FRR	FRR
	1 in 10^4
	1 in 10^5
	1 in 10^6

Organising your data to get the scores



Reporting performance

$$\Delta_\beta = \arg\min_{\Delta} \mathrm{cost}_\beta(\Delta|dev)$$



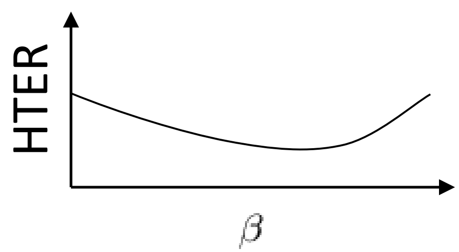
Where Δ is tuned *a priori* on a **dev set** according to the criterion:

$$\mathrm{cost}_\beta^{wer}(\Delta|dev) = \beta \mathrm{FAR}(\Delta|dev) + (1-\beta)\mathrm{FRR}(\Delta|dev)$$

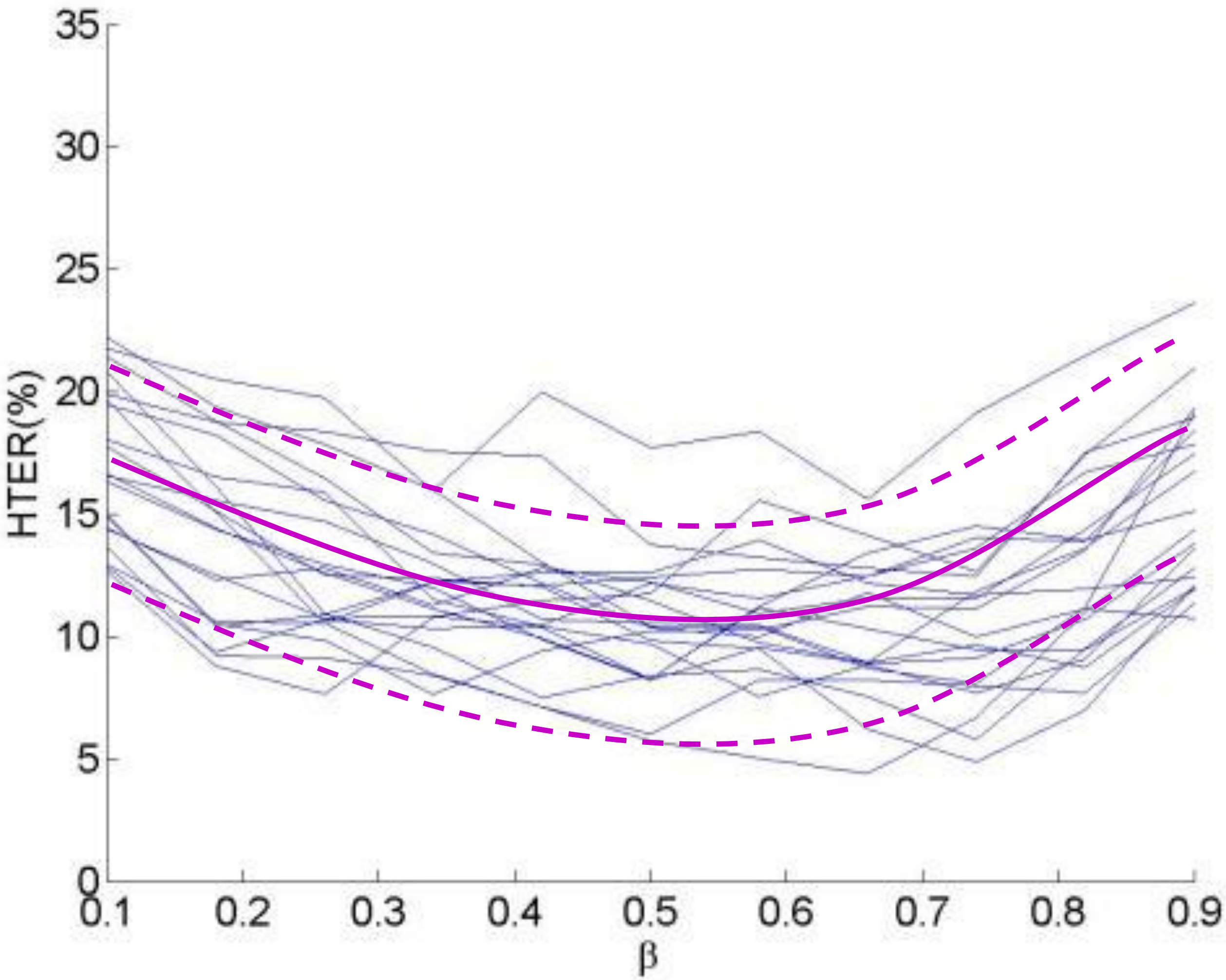
$$\beta \in [0, 1]$$

Final performance on **eva set** :

$$\mathrm{HTER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2}$$



Expected Performance Curve (EPC)

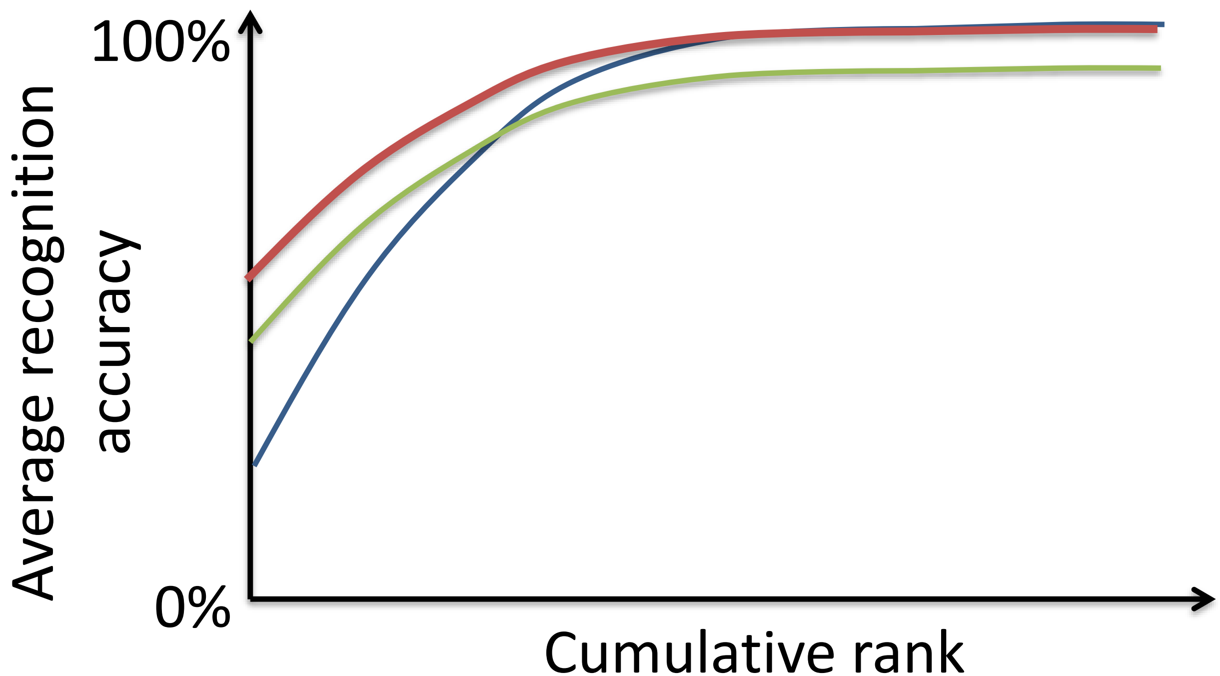


Low cost of FAR

High cost of FAR

Cumulative Match Characteristic (CMC) curve – Closed set identification

- Rank-k accuracy: Classification accuracy when the correct candidate is found within top k

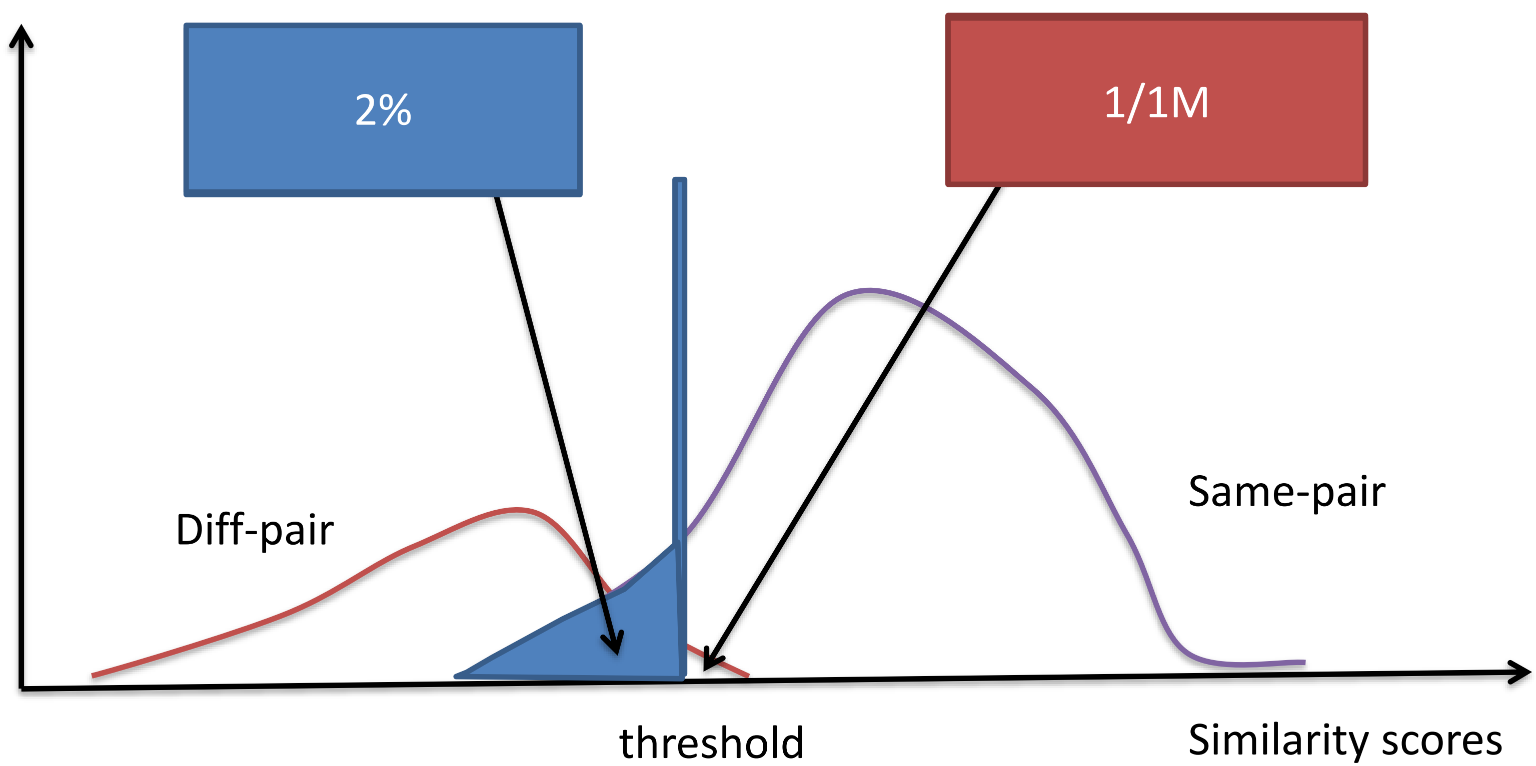


Menu

VARIABILITY AND REPEATABILITY OF BIOMETRIC PERFORMANCE

Ask the right questions

Vendor: “Our false rejection is 2% when operating at a false acceptance rate of 1 in a million”



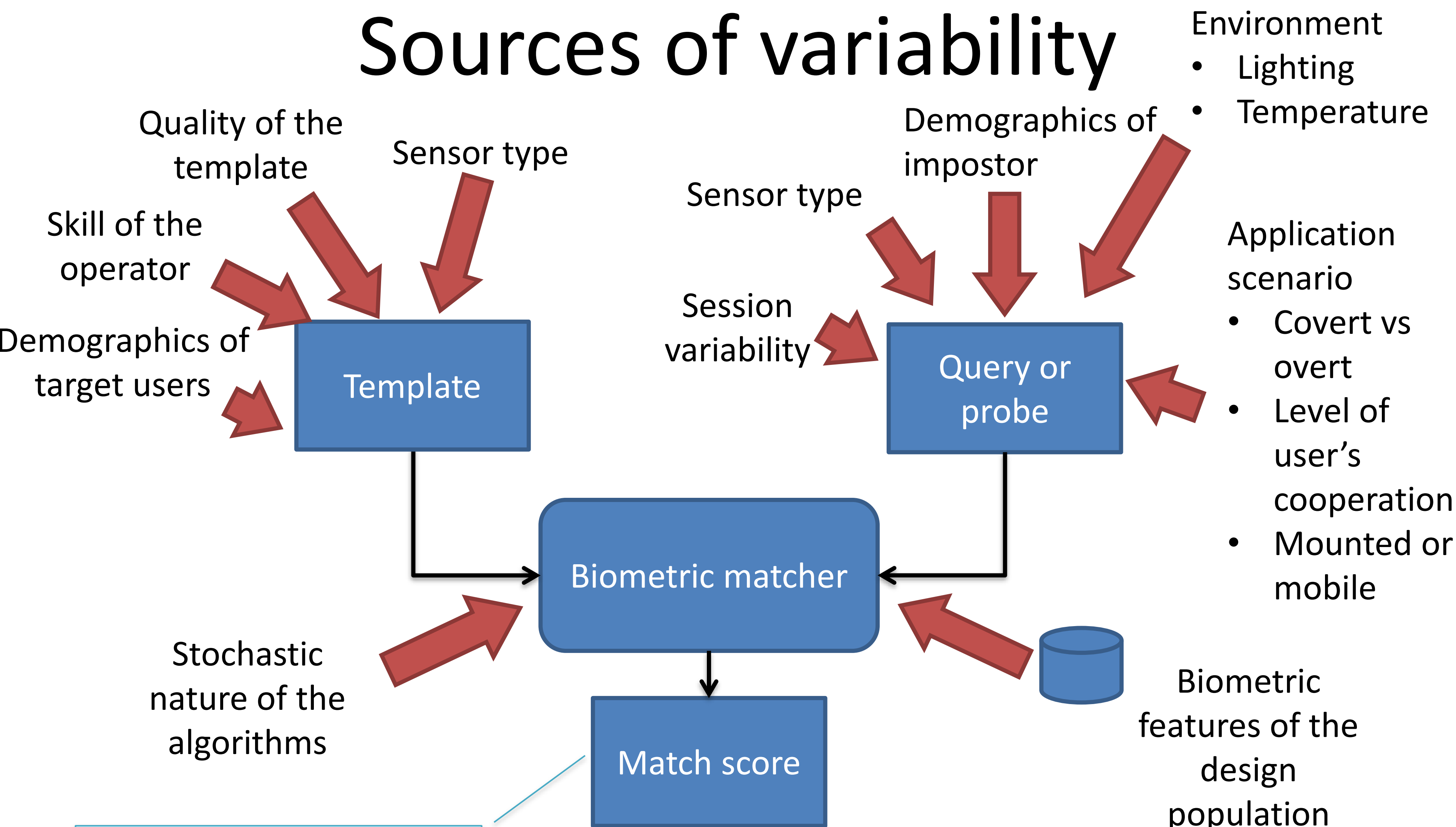
What questions should you ask?

1/50 FRR

1/1M FAR

Questions?

Sources of variability



We are trying to collect the scores and measure the system performance

Wayman, James L., Antonio Possolo, and Anthony J. Mansfield. "Modern statistical and philosophical framework for uncertainty assessment in biometric performance testing." *IET biometrics* 2.3 (2013): 85-96.

Three types of biometric tests

Technology test

- Use a (sequestered) database to test/compare algorithms
- **Repeatable** because many factors are ‘fixed’ once data is collected

Scenario testing

- Measure the overall system (sensor + algorithm) performance in a typical real-world application domain
- It is not repeatable because another field test will not give exactly the same result

Operational testing

- Model the performance in a specific application using a specific biometric system
- Not repeatable

Source: Philips, P.J., Martin, A., Wilson, C., Przybocki, M.:‘An introduction to the evaluation of biometric systems’, IEEE Comput., 2000, 33, (2), pp. 56–63

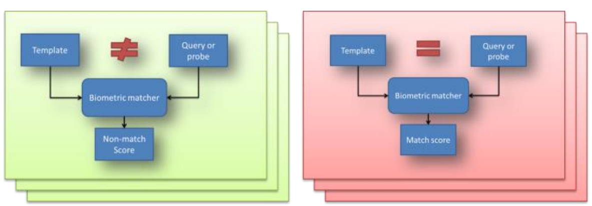
Comments: Wayman, James L., Antonio Possolo, and Anthony J. Mansfield. "Modern statistical and philosophical framework for uncertainty assessment in biometric performance testing." *IET biometrics* 2.3 (2013): 85-96.

Repeatability

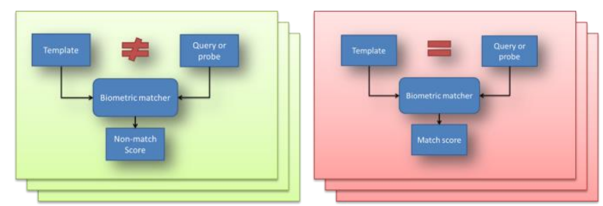
Fix the conditions

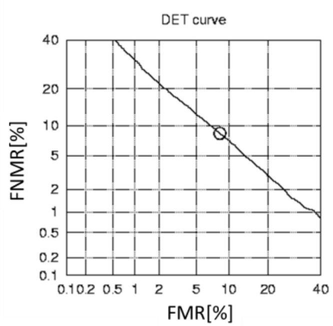
- Same SOP for enrolment, same operators
- Same biometric system & sensor
- Same application scenario
- Similar (controlled) environment
- Comparable user size but different subjects

Installation site A



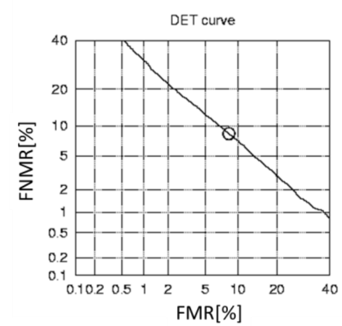
Installation site B



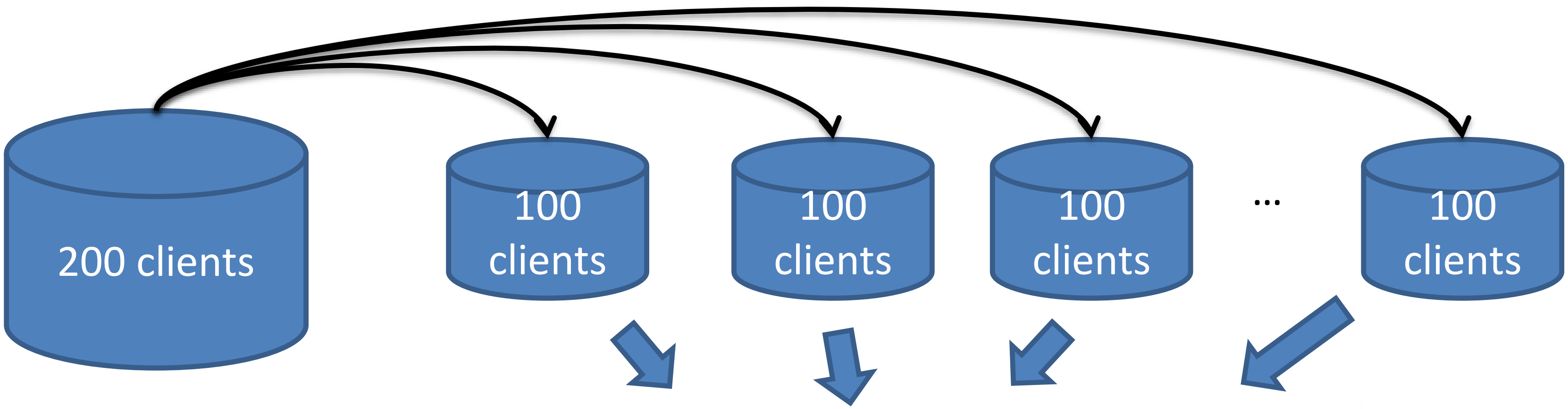


Will the two curves be the same?

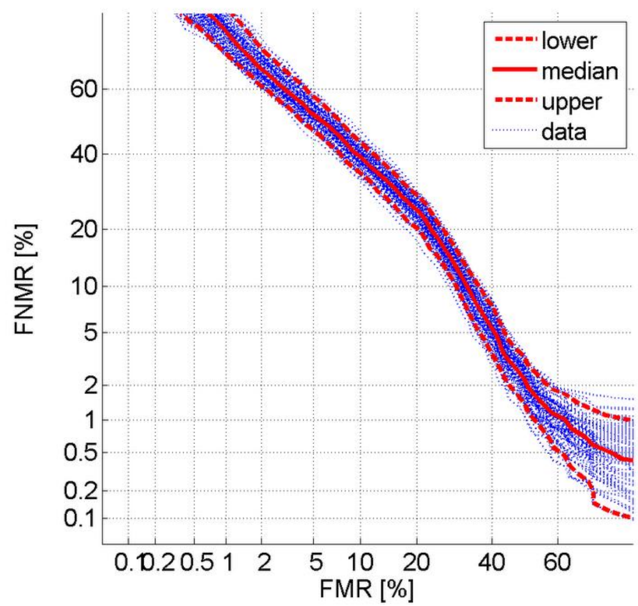
How much do they vary?



Clients-based bootstrapping



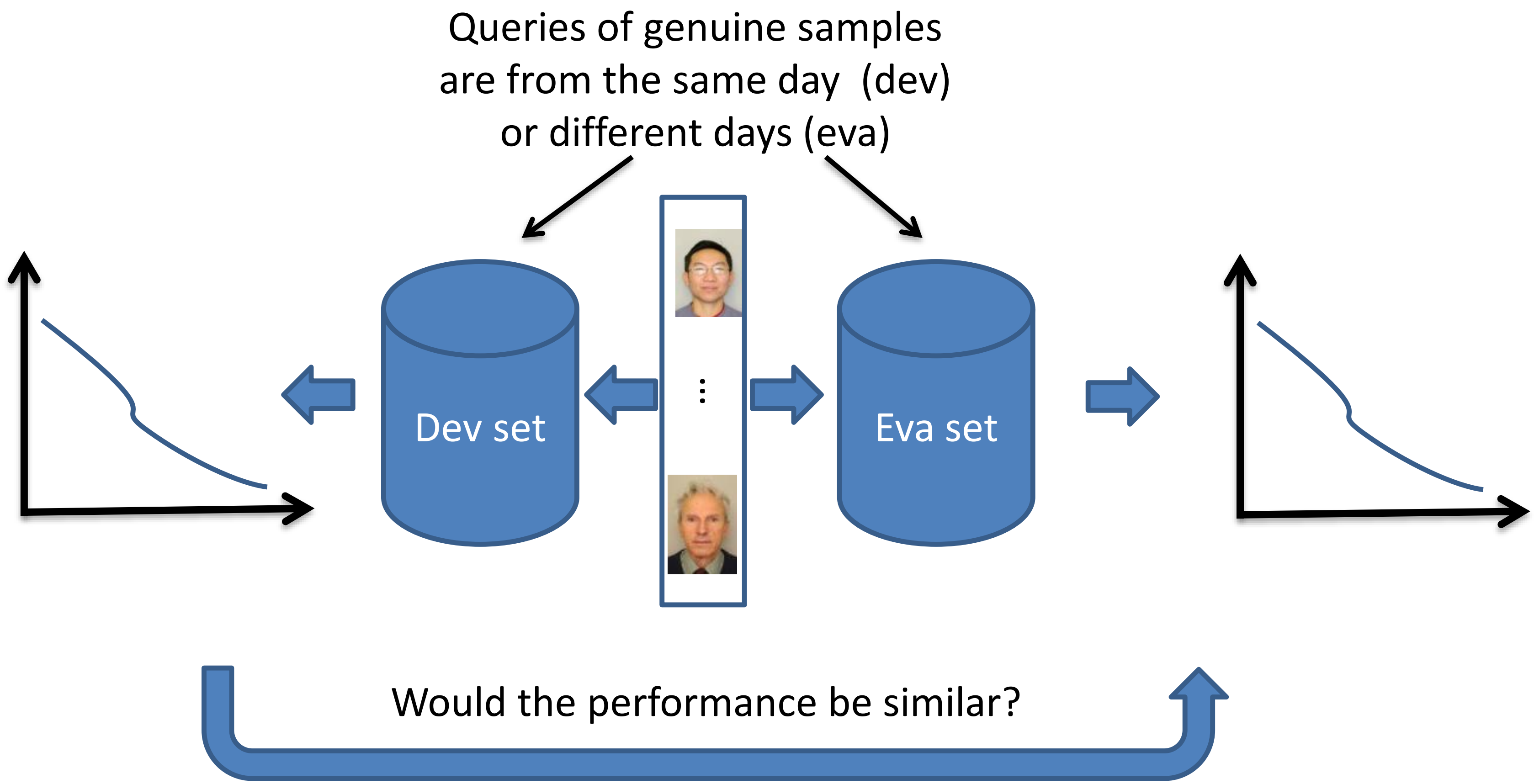
Variability due to the user composition is observed under <u>technology test</u>



Questions:

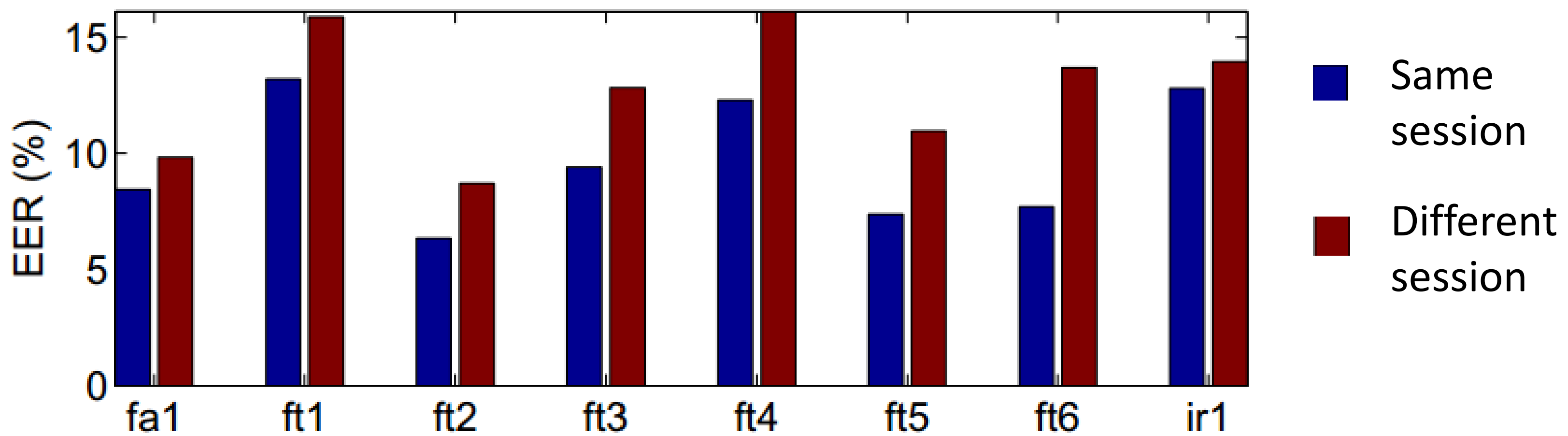
- How much do they vary?
- Why do they vary?
- Can we extrapolate (generalize) to another scenario?

Session variability – why it matters?



With vs without session variability

Label	template ID {n}	Modality	Sensor
fa	1	Still Face	web cam
ft	1–6	Fingerprints	Thermal
ir	1	Left iris image	LG



Biosecure DS2 score database

http://personal.ee.surrey.ac.uk/Personal/Norman.Poh

Revisit the questions

1/50 FRR

1/1M FAR

Your answers…

Menu

Quiz

Should we consider session variability when testing a biometric system?

A: No, we should not because we want to fix all parameters in favour of experimental repeatability

B: Yes, we must do so in every biometric testing

Need a hint?

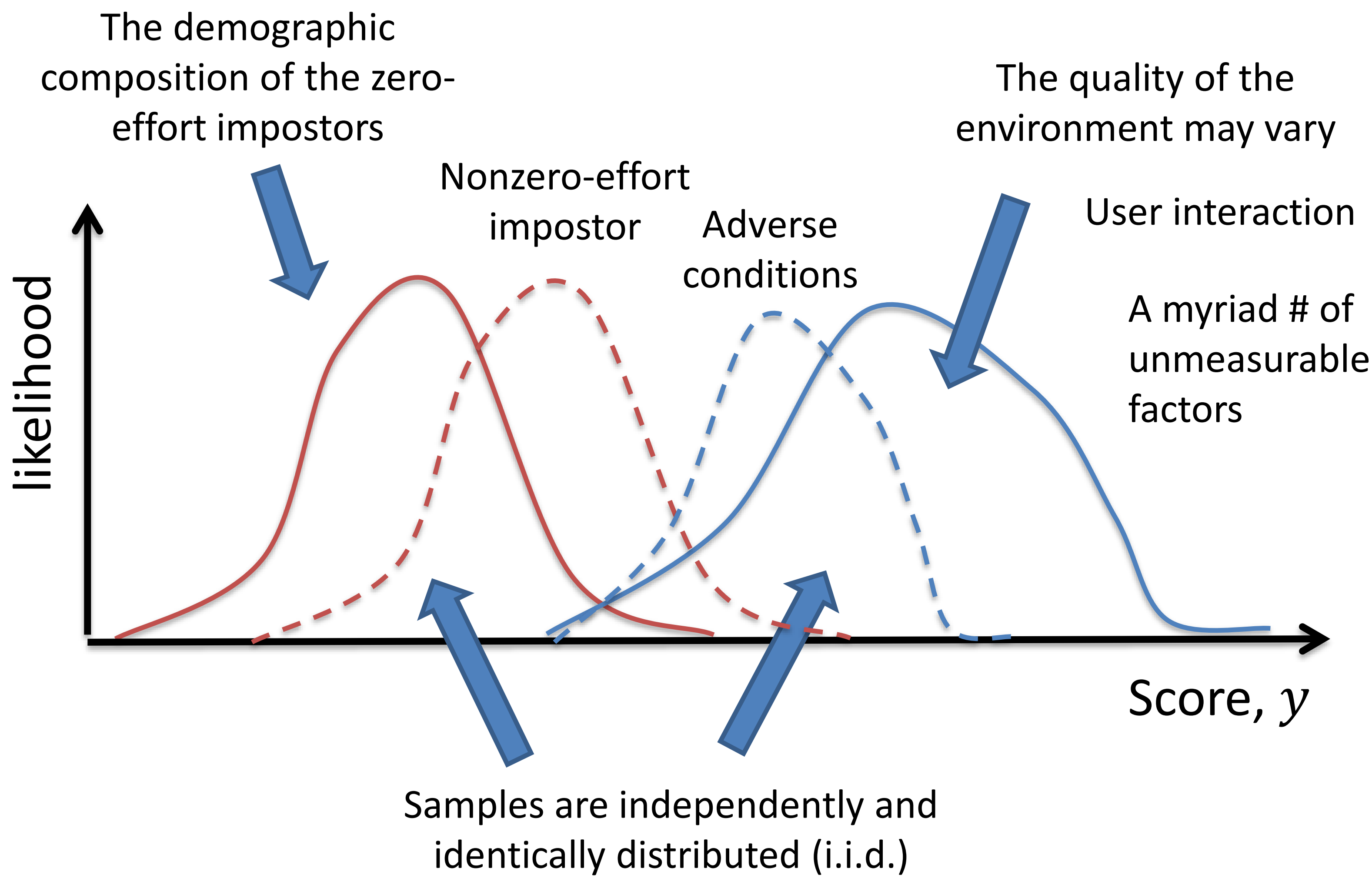
Do you know of any biometric system that is designed to operate only on the same day as enrolment?

Explanation

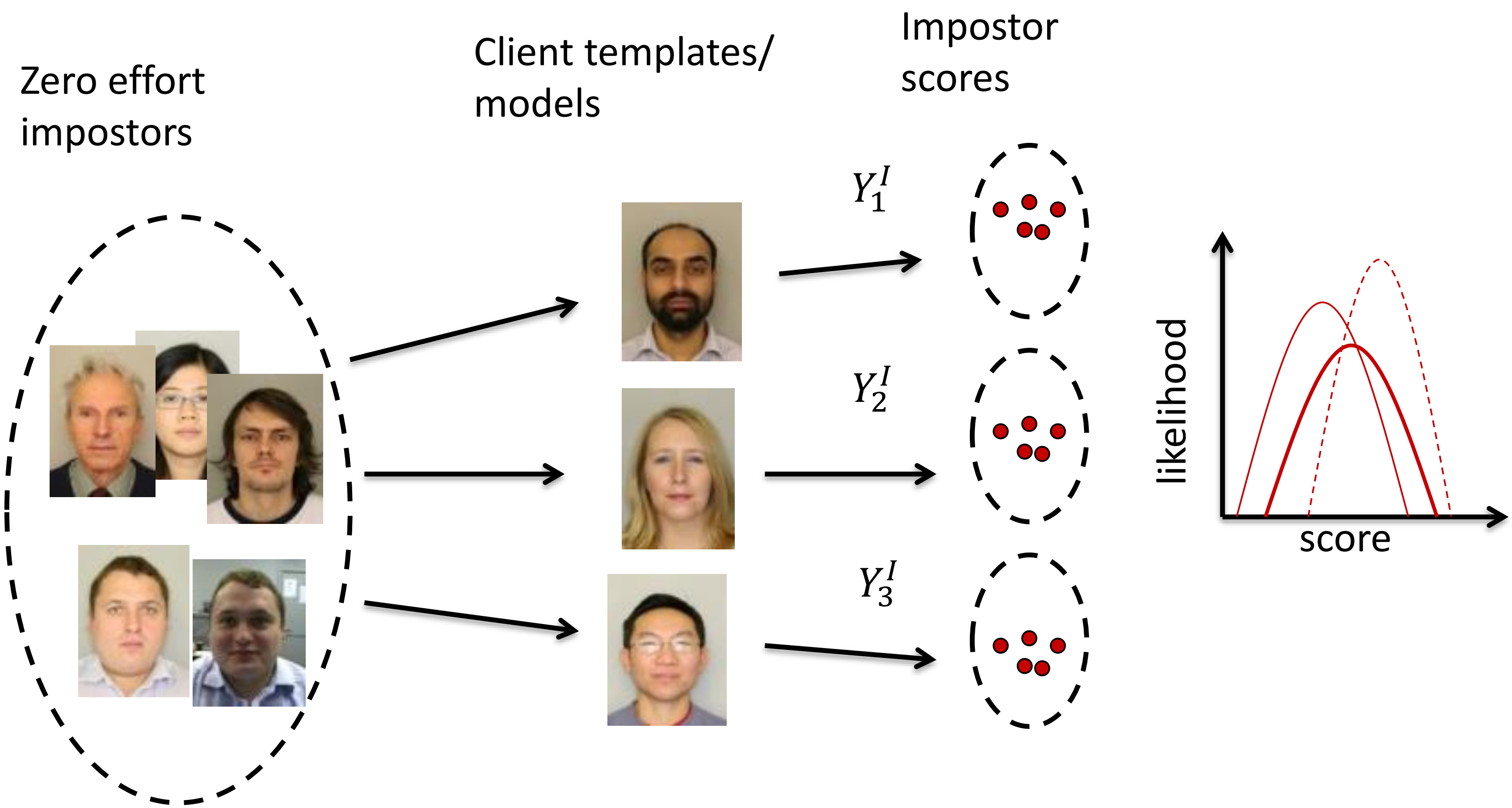
We must acknowledge that we can never fix all parameters and fixing all is not always desirable

BIOMETRIC SCORE THEORY

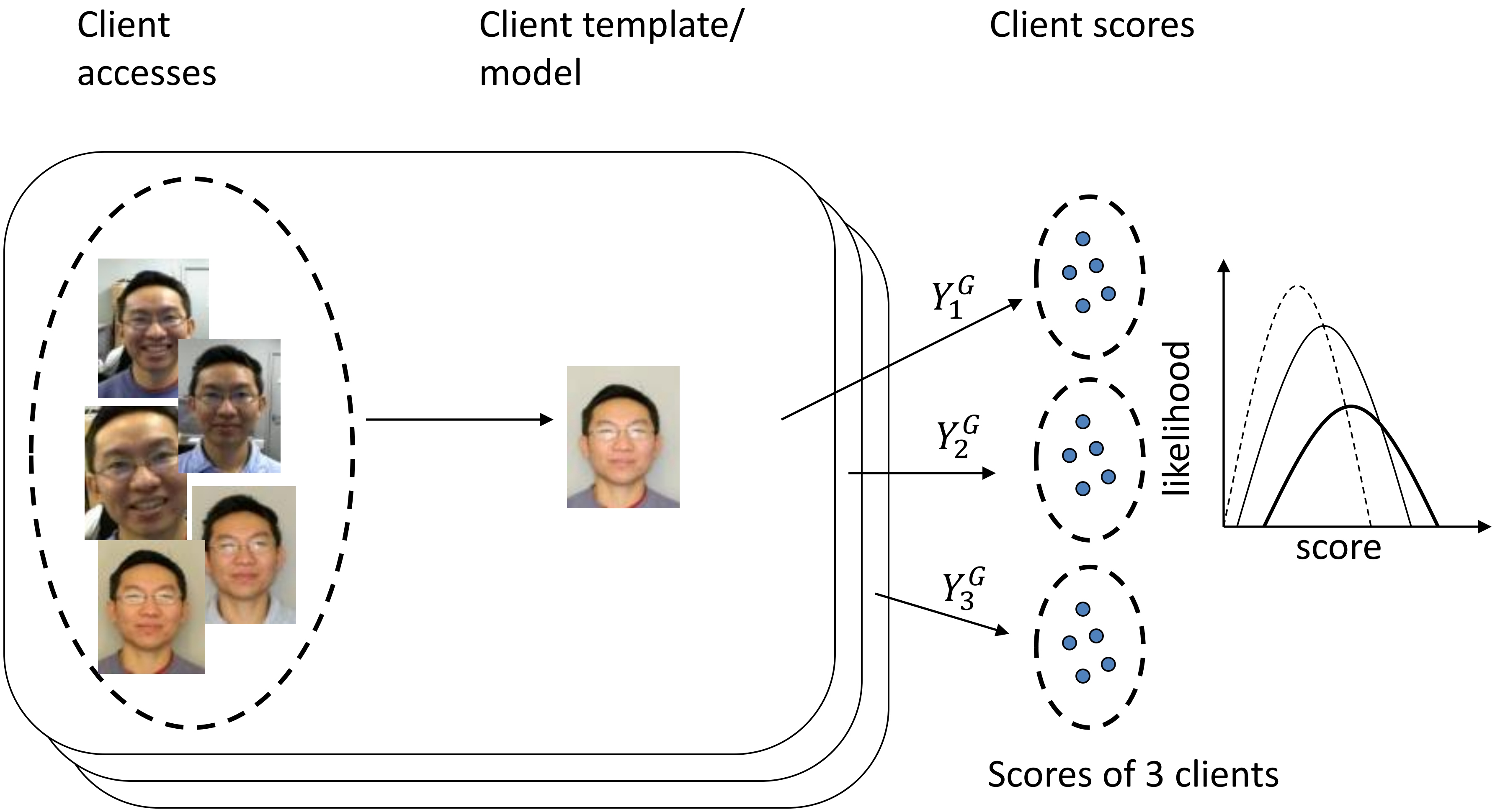
What assumptions?



Formation of nonmatch scores



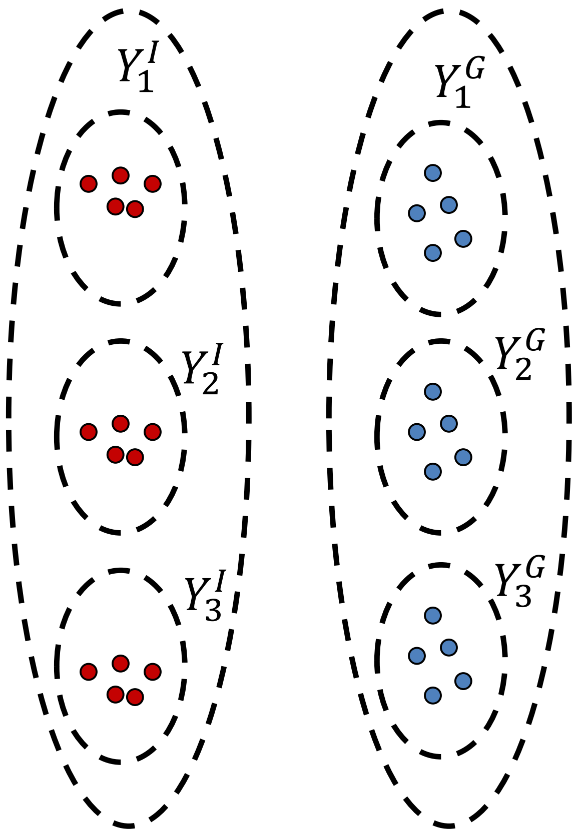
Formation of match scores



An Observed Experimental Outcome

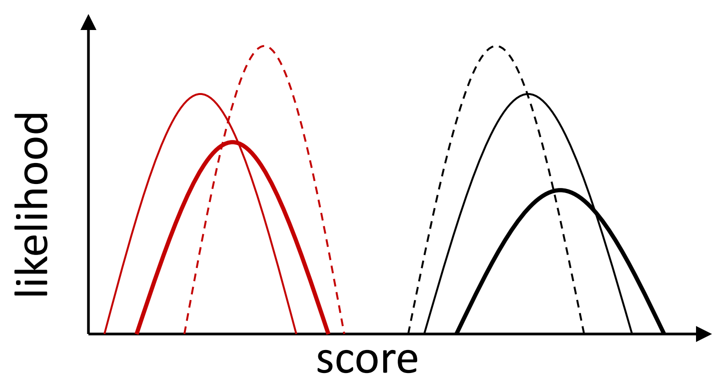
Impostor scores

Client scores



Samples are independently and identically distributed (i.i.d.) **given the claimant ID**

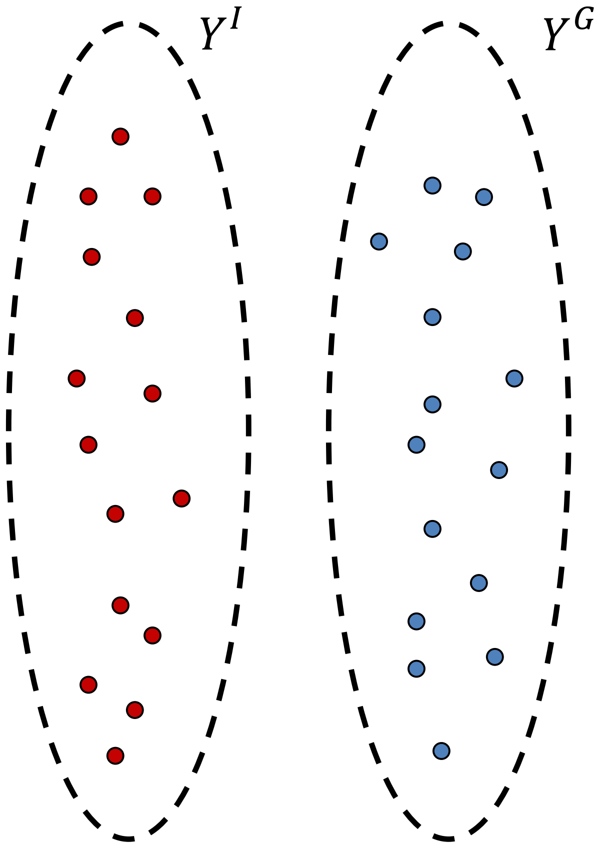
Strong correlation exists among scores generated by the common client model



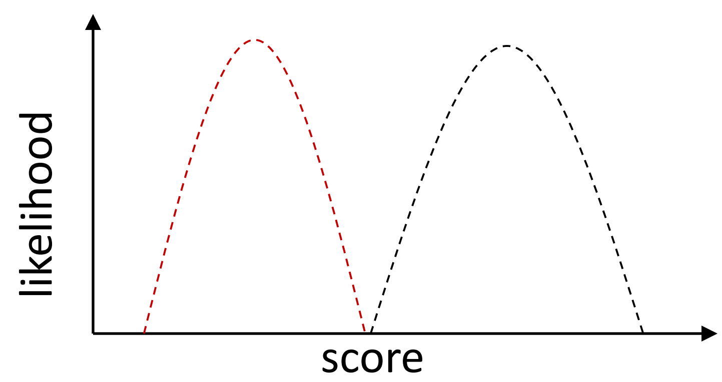
How would i.i.d. samples look like?

Impostor scores

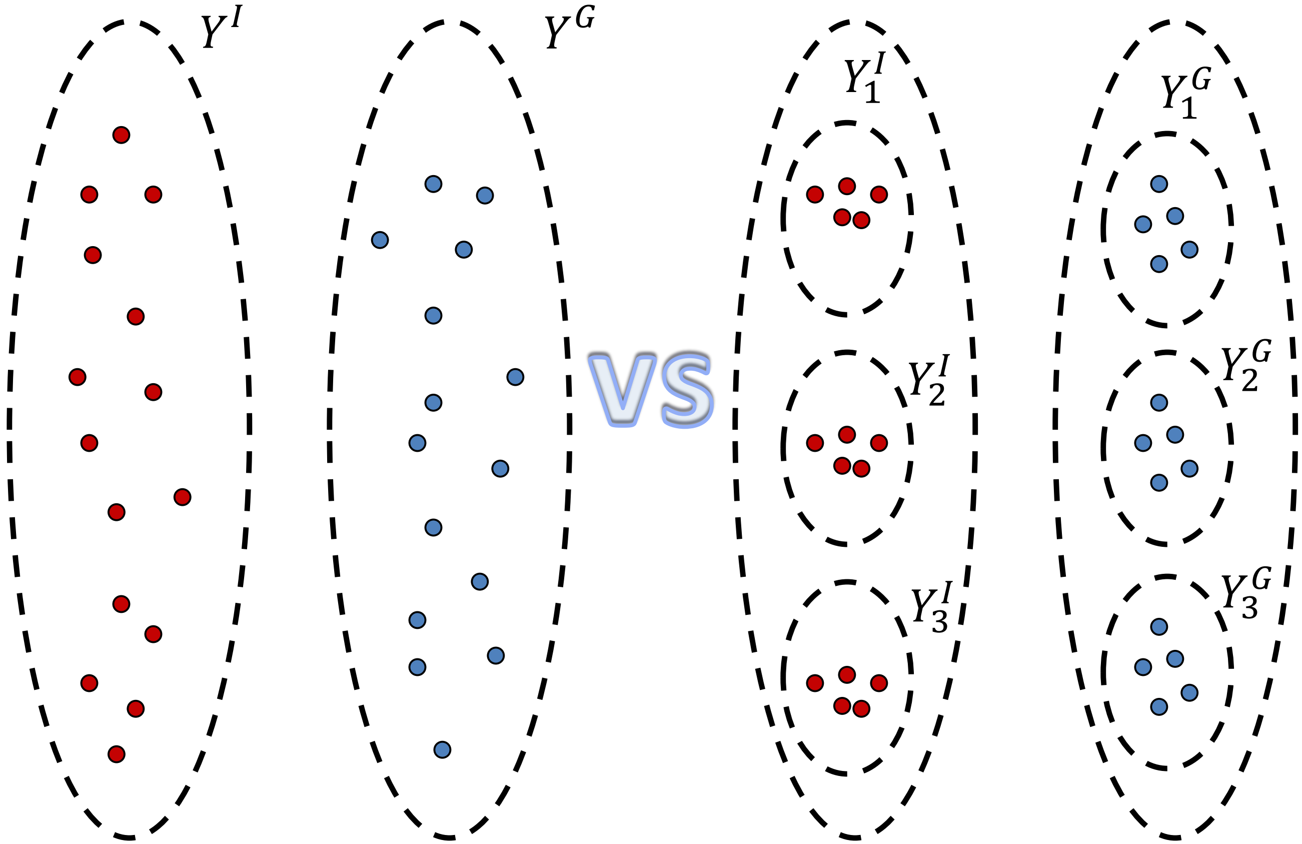
Client scores



Samples are independently and identically distributed (i.i.d.)

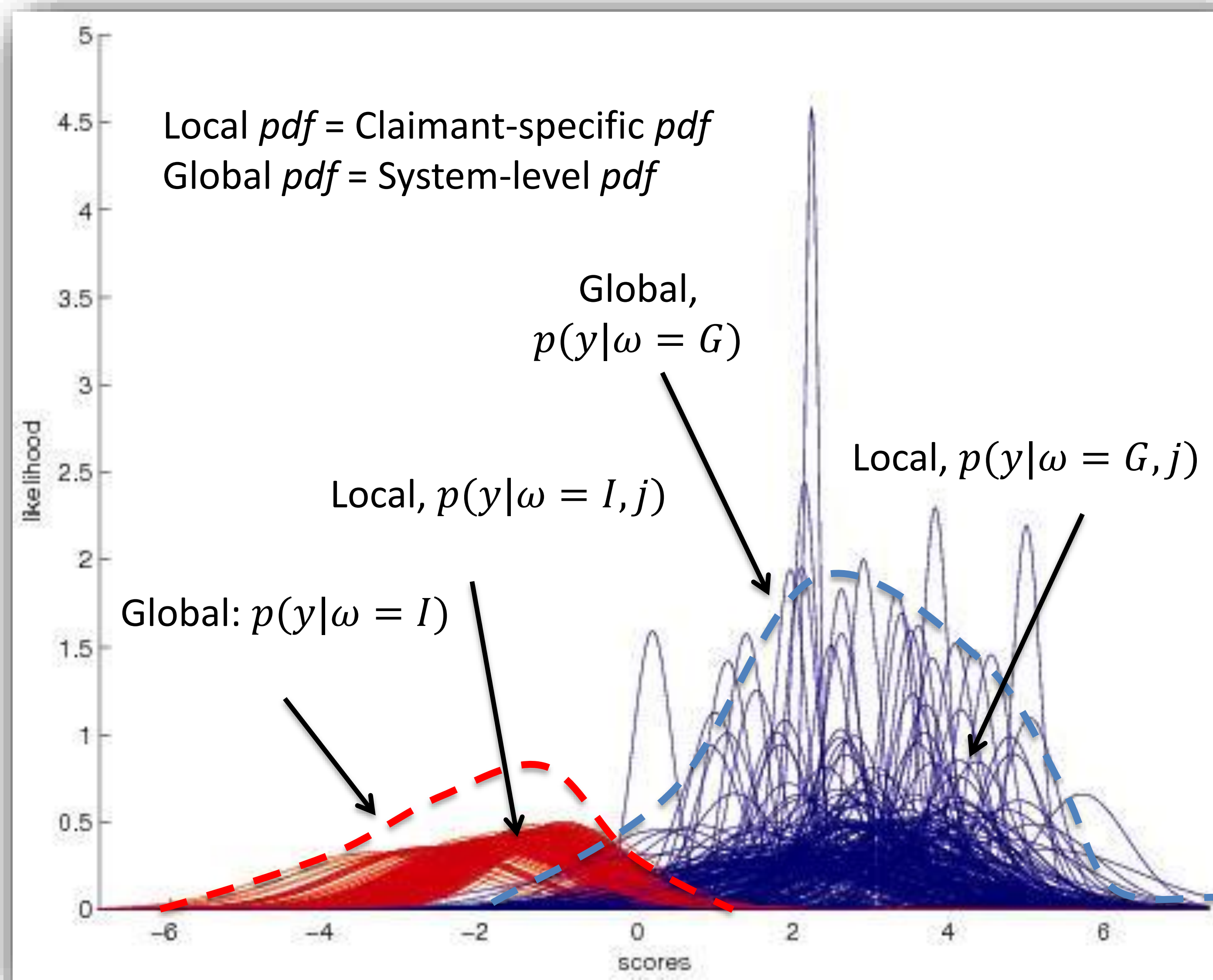


Which is correct?



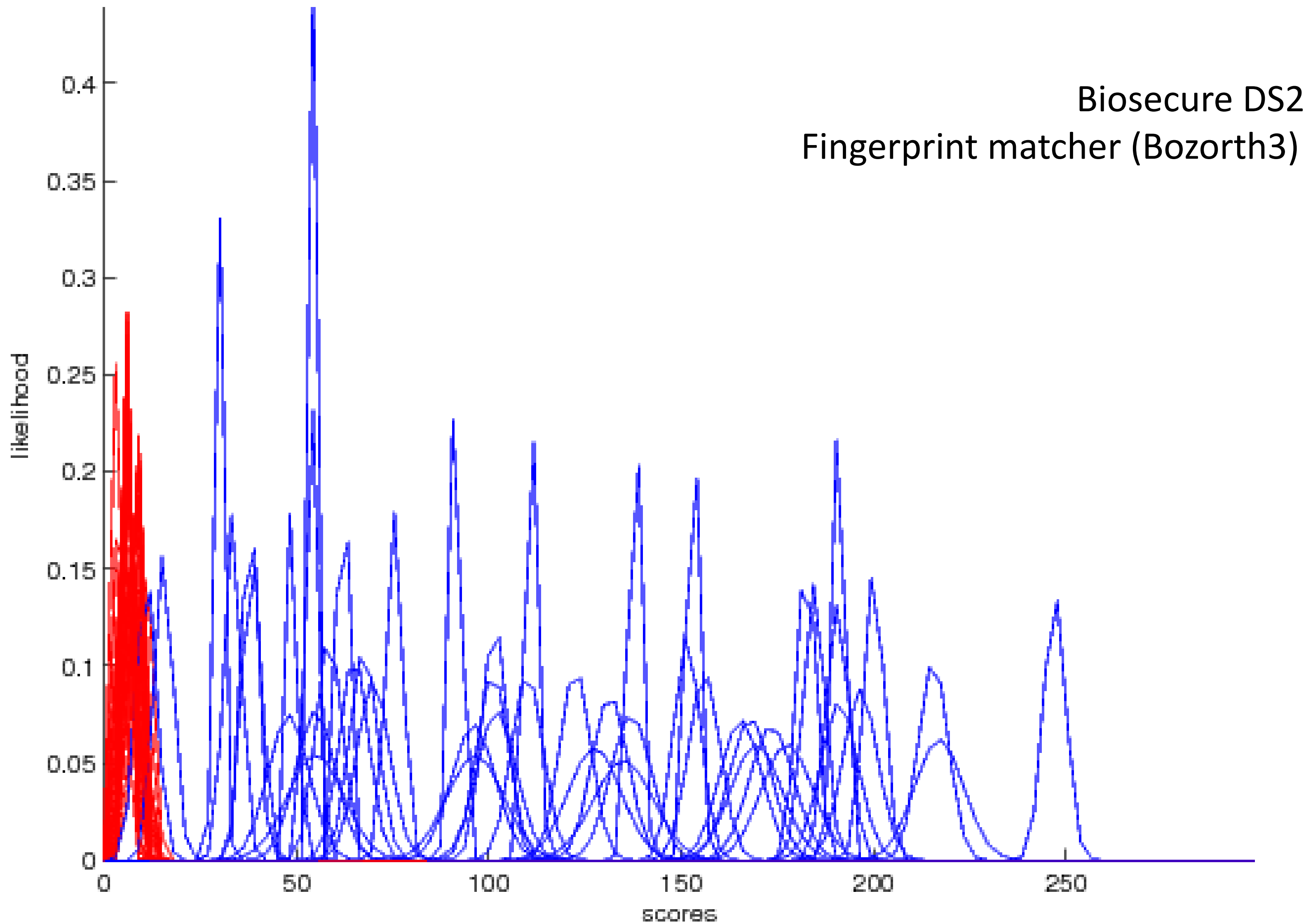
Samples are i.i.d.

Samples are i.i.d. given the claimed ID



- XM2VTS face system (DCTmod2,GMM)
- 200 users/clients
- 3 genuine scores per user (blue curve)
- 400 impostor scores per user (red curve)

User-specific fingerprint score distribution



How are the local and global models related?

$$p(y|\omega) = \sum_{j=1}^{J} p(y|\omega, j) P(j|\omega) \qquad \text{for } \omega \in \{G, I\}$$

The class-conditional pdf associated to claimant j

The prior probability of claimant j given the class label ω

There is no reason why one claimant is more important than another; so, $P(j|\omega)$ should be uniform: $P(j|\omega) = \frac{1}{J}$

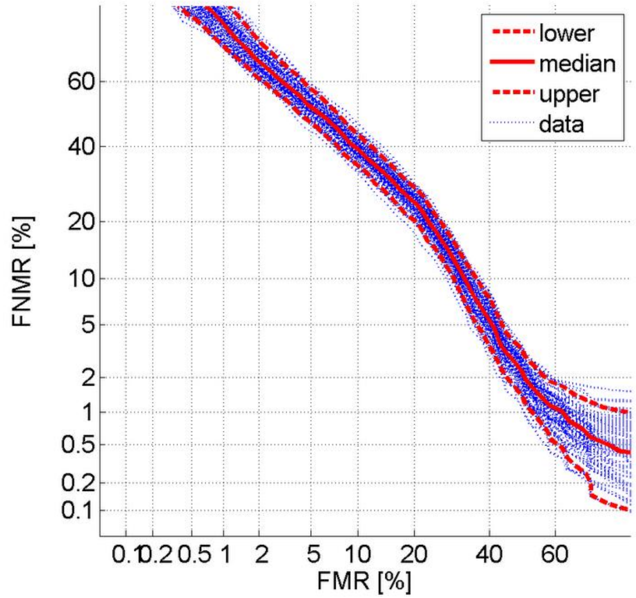
$$p(y|\omega) = \frac{1}{J} \sum_{j=1}^{J} p(y|\omega, j)$$

The system-level class-conditional *pdf* score is simply an average of the *pdf* of the claimant-specific *pdf*.

Consequence

The user composition has direct influence on the system performance even when all experimental conditions are fixed

$$P(y \leq \Delta | \omega = G) = \frac{1}{J} \sum_{j=1}^{J} P(y \leq \Delta | \omega = G, j)$$

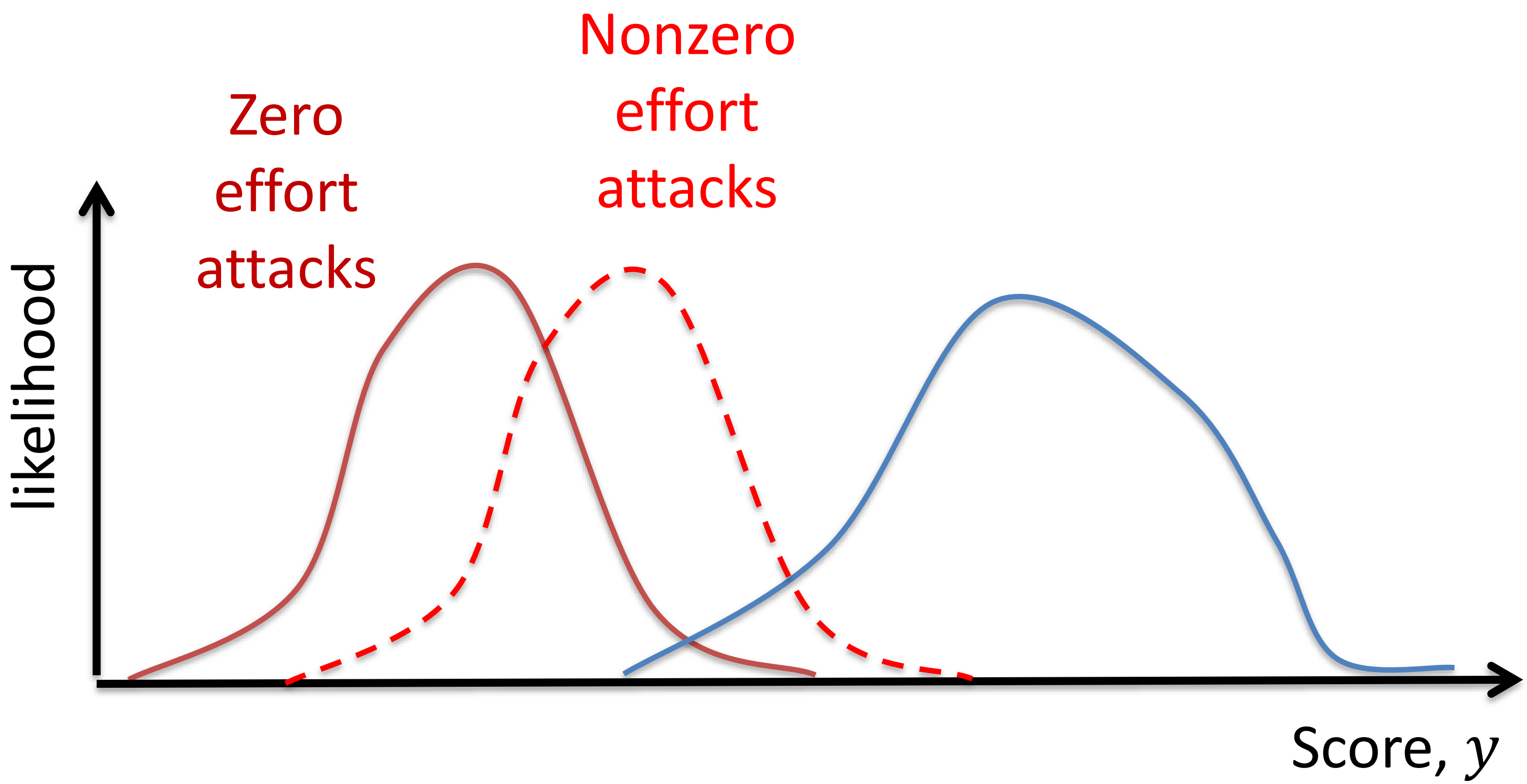


$$P(y > \Delta | \omega = I) = \frac{1}{J} \sum_{j=1}^{J} P(y > \Delta | \omega = I, j)$$

Menu

PERFORMANCE UNDER PRESENTATION ATTACKS

Nonzero effort attacks



ISO/IEC 30107-3 defined metrics for assessing the performance of the PAD methods

Attack Presentation Classification Error Rate (APCER)

N_{PAI} number of PAI

$$APCER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - Res_i)$$

FAR

$APCER_{PAI}$ is computed once for each PAI

Bona Fide Presentation Classification Error Rate (BPCER)

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} Res_i}{N_{BF}}$$

FRR

$Res_i = 1$, declare presentation attack for sample i; 0 otherwise

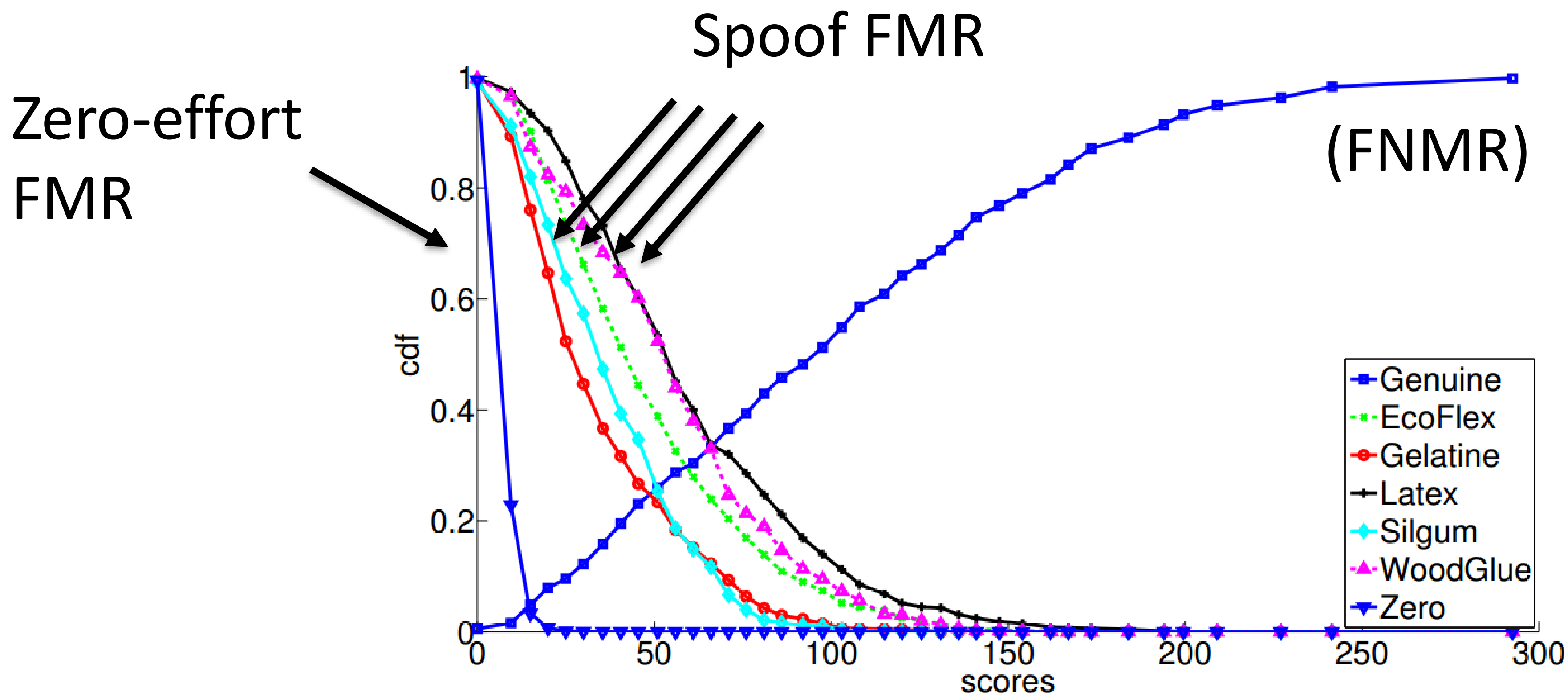
A single metric summarising all PAIs

- Average Classification Error Rate (ACER)

$$ACER = \frac{\max\limits_{PAI=1\ldots S}(APCER_{PAI}) + BPCER}{2}$$

$APCER_{PAI}$ is computed once for each PAI

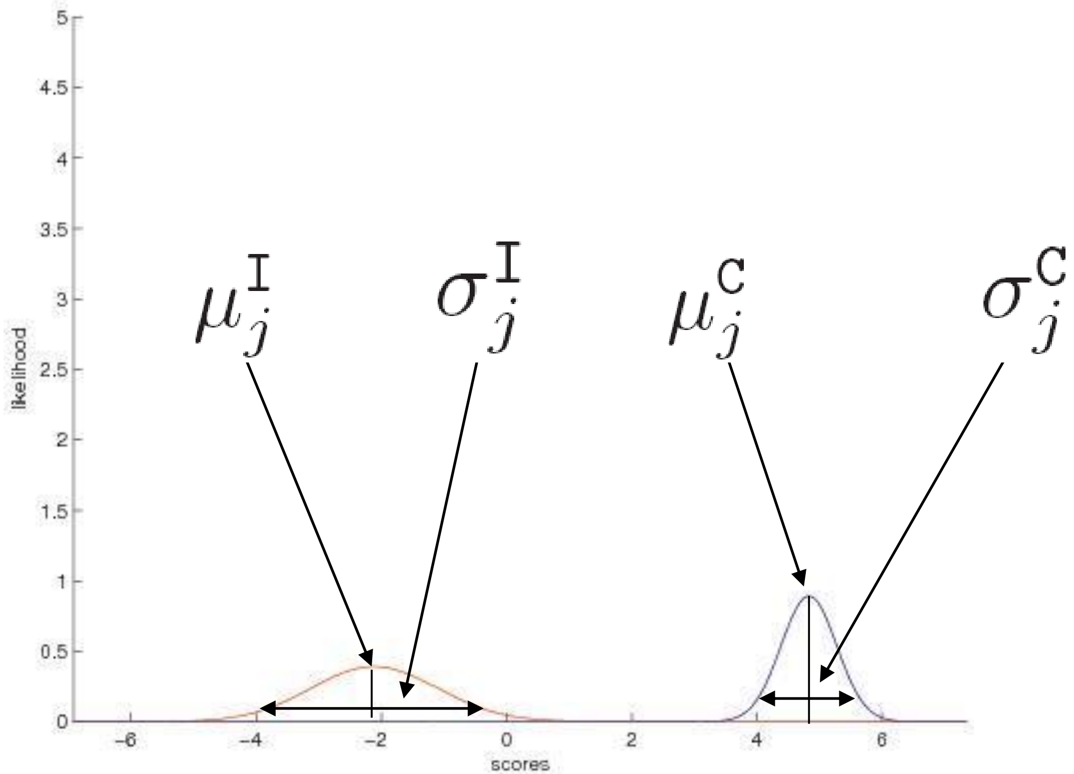
Performance under attack



Menu

MODEL-BASED PERFORMANCE EVALUATION

Model-based assessment



Other Parameteric Point-based Assessment

$$\text{EER} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right)$$

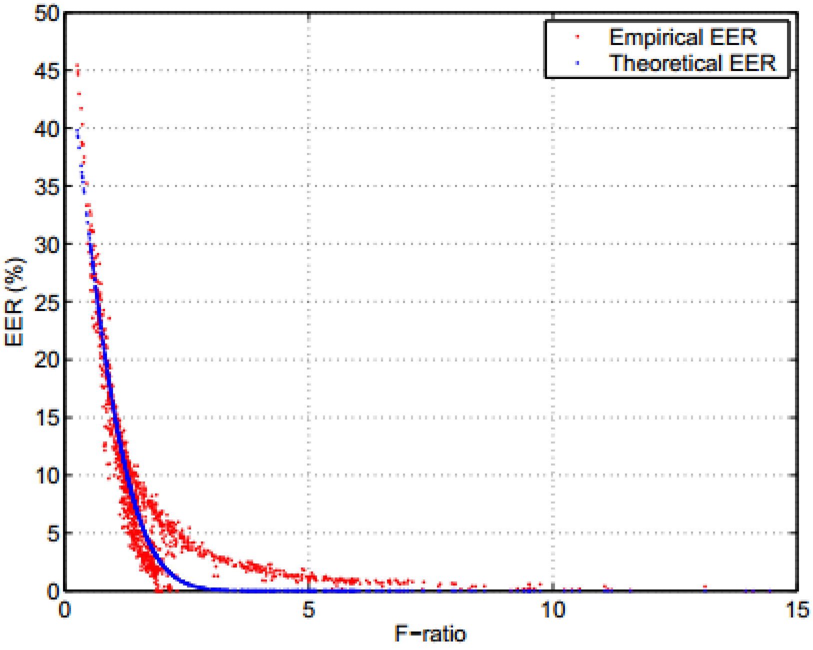
where $\text{F-ratio} = \dfrac{\mu^{\mathrm{C}} - \mu^{\mathrm{I}}}{\sigma^{\mathrm{C}} + \sigma^{\mathrm{I}}}$

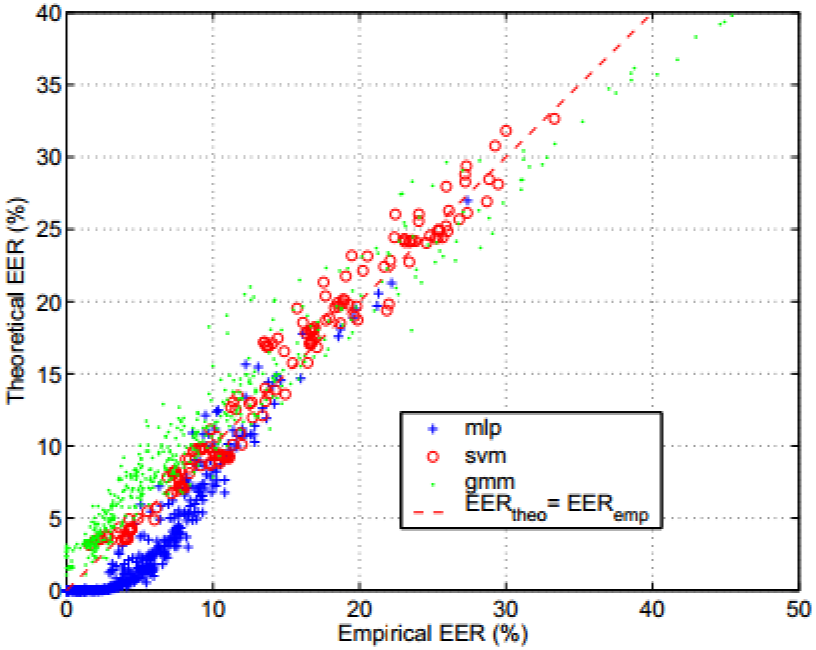
and $\text{erf}(z) = \dfrac{2}{\sqrt{\pi}} \displaystyle\int_0^z \exp\left[-x^2\right] dx$

$$\text{Fisher-ratio} = \frac{(\mu^{\mathrm{C}} - \mu^{\mathrm{I}})^2}{(\sigma^{\mathrm{C}})^2 + (\sigma^{\mathrm{I}})^2}$$

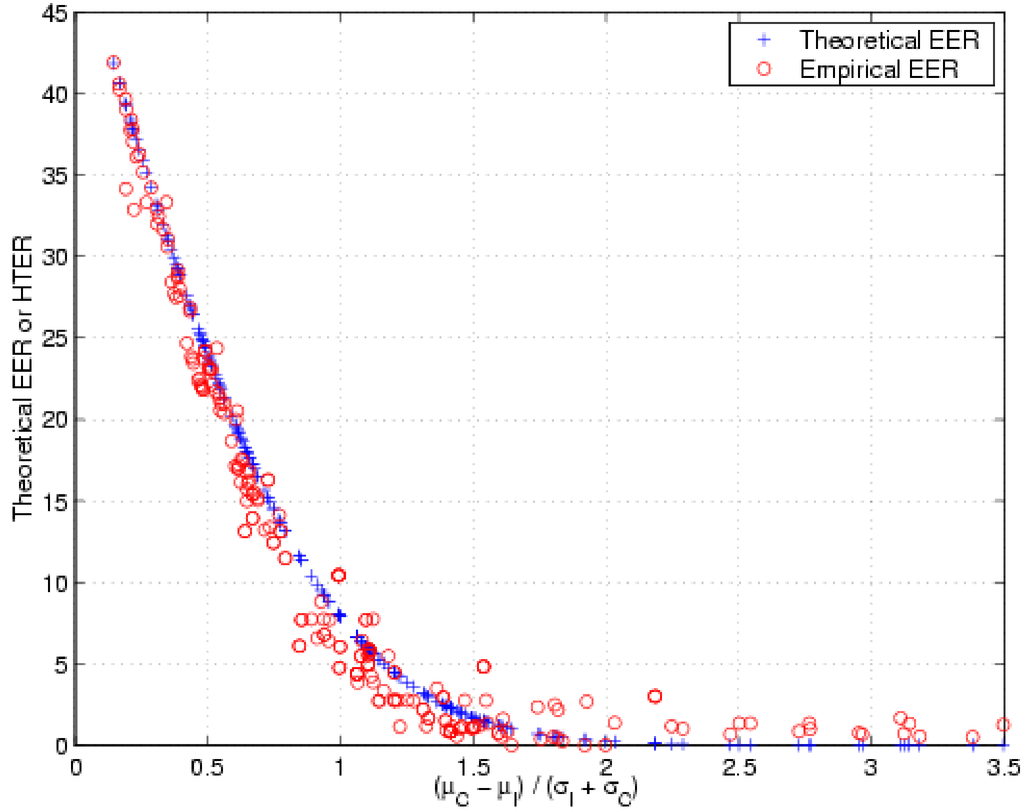
$$d' = \frac{|\mu^{\mathrm{C}} - \mu^{\mathrm{I}}|}{\sqrt{\frac{1}{2}(\sigma^{\mathrm{C}})^2 + \frac{1}{2}(\sigma^{\mathrm{I}})^2}}$$

F-ratio and EER





F-ratio and EER



180 experiments from XM2VTS, NIST2001 and BANCA

Application of parametric analyses

User-specific performance analysis

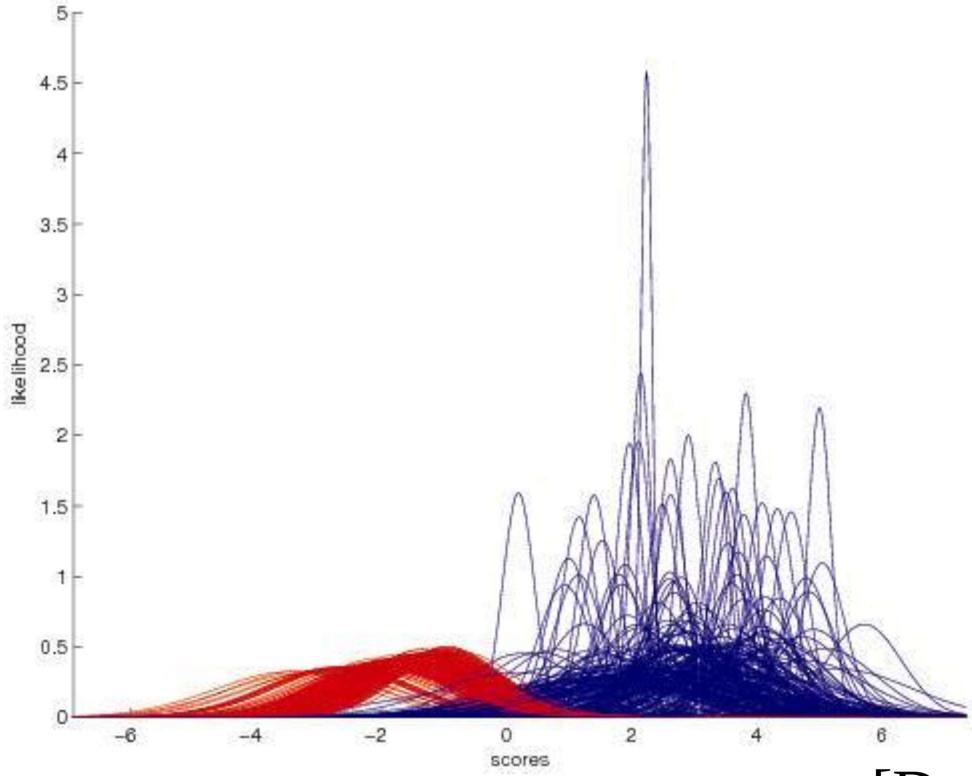
Performance trend estimation over time

When no empirical error is observed

Multimodal fusion diagnosis

- How correlation affect the system performance?
- Which combination of biometric traits are optimal?

User-specific performance analysis



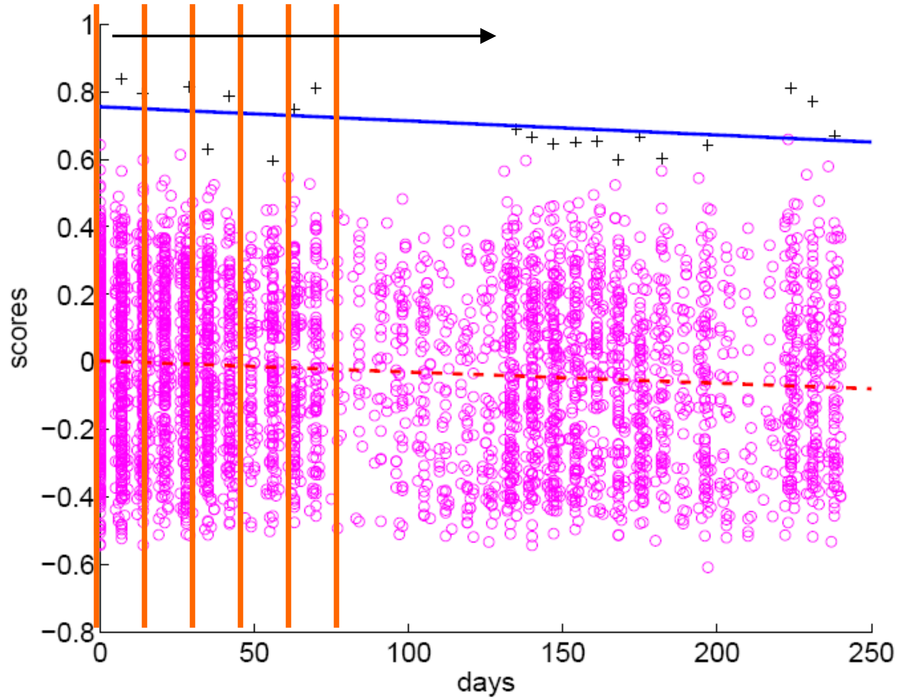
- XM2VTS face system (DCTmod2,GMM)
- 200 users/clients
- 3 genuine scores per user (blue curve)
- 400 impostor scores per user (red curve)

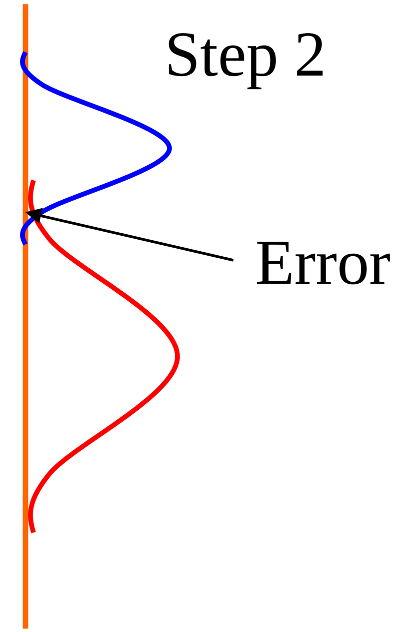
[Doddington et al, 1998]

Error estimation over time

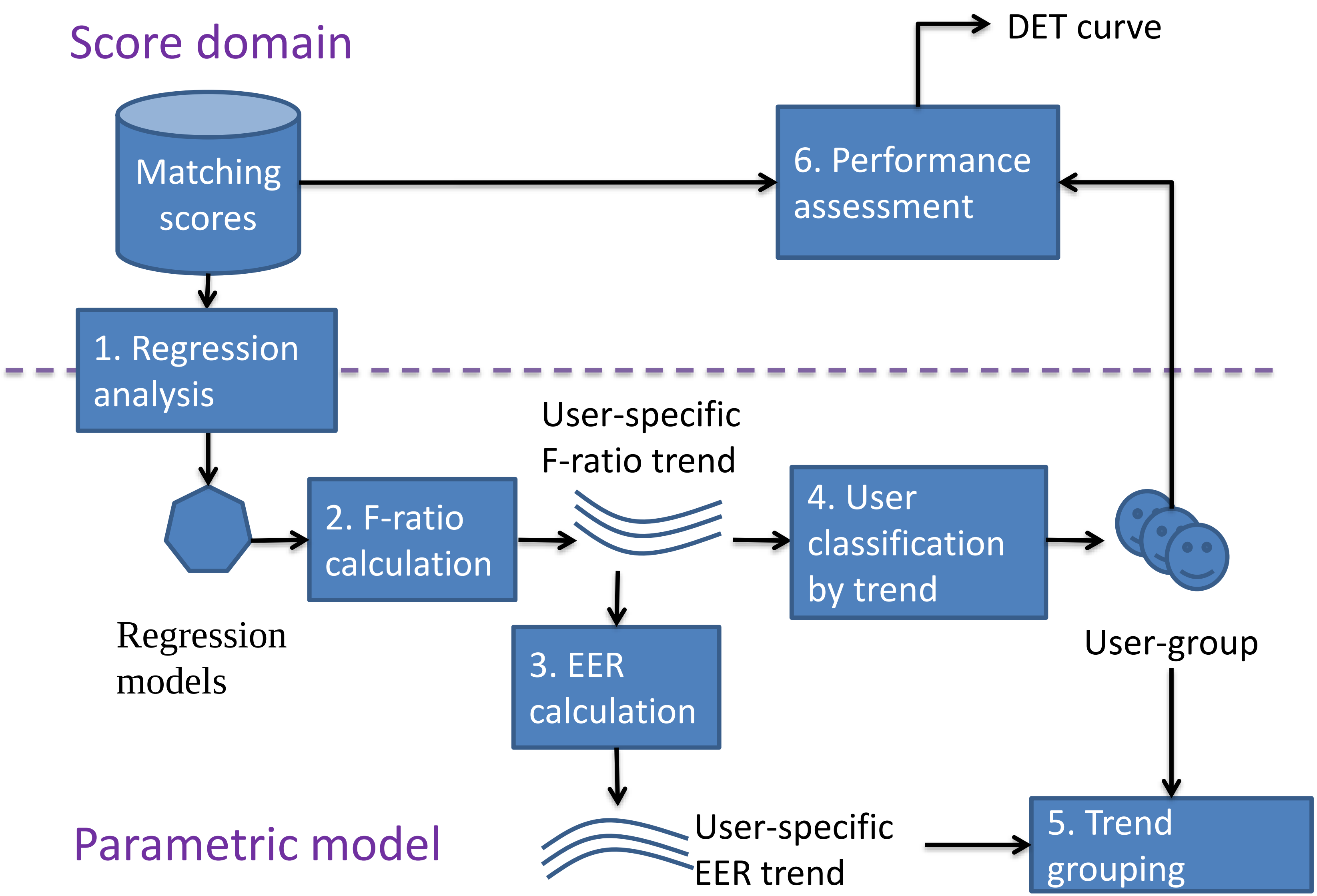
1. Estimate score density over time
2. Estimate the performance over time

Step 1

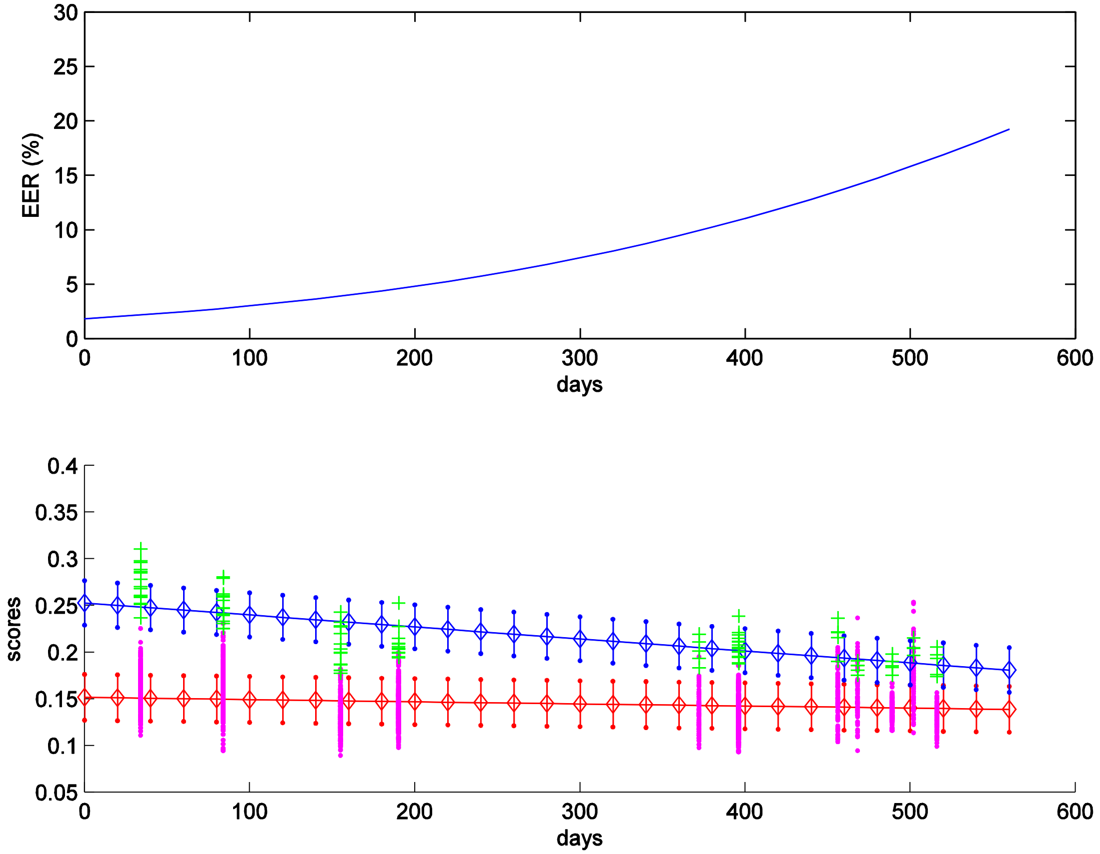




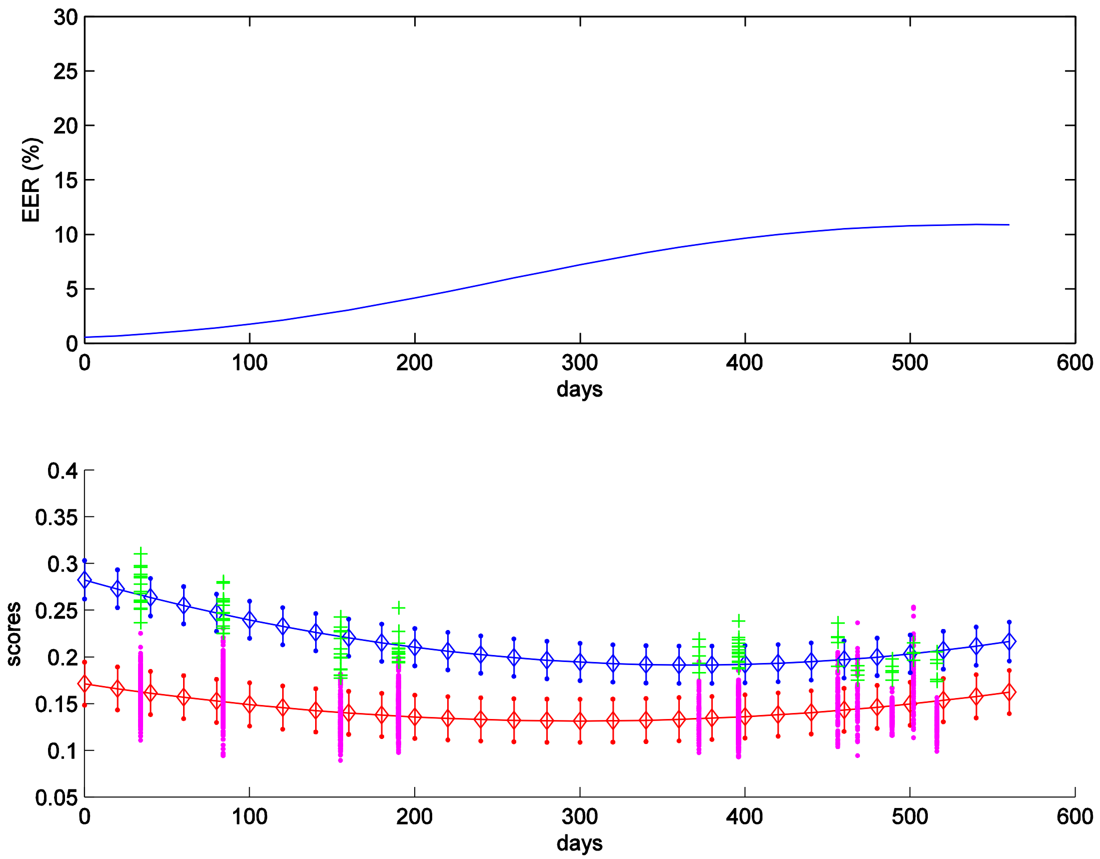
The homomorphic framework



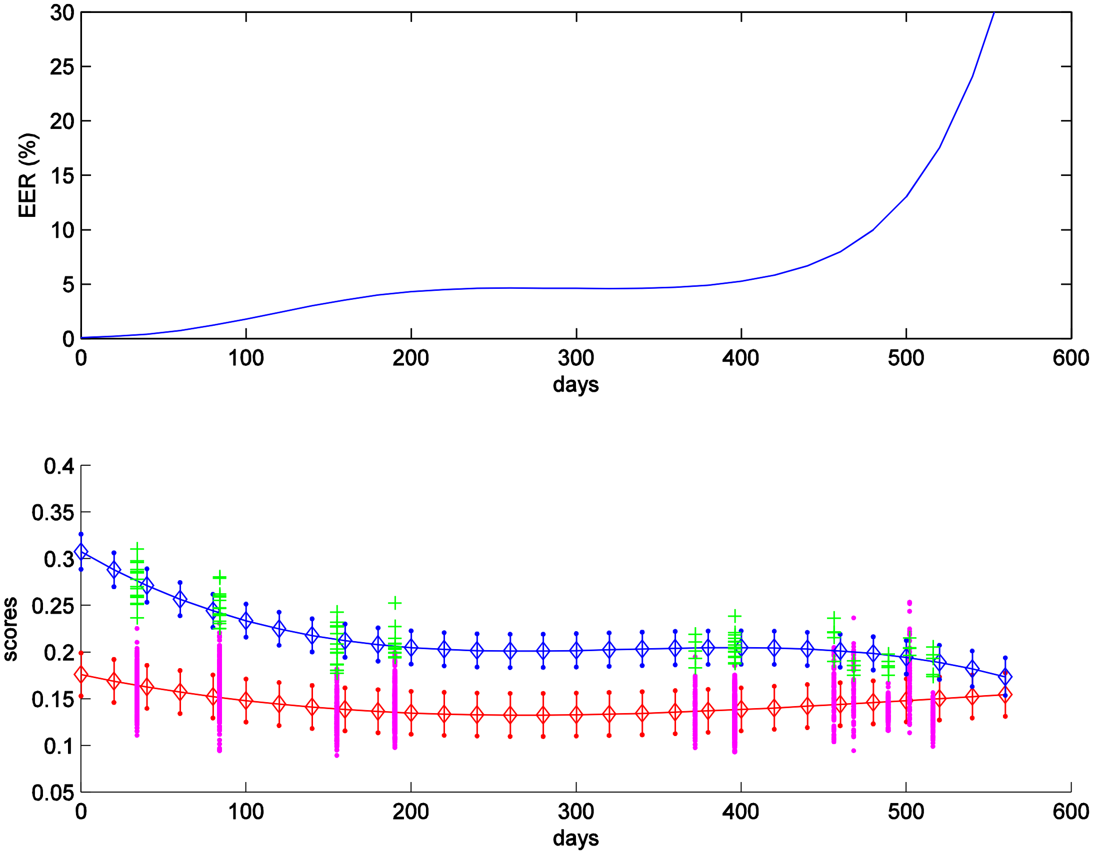
Degree of polynomial (1)



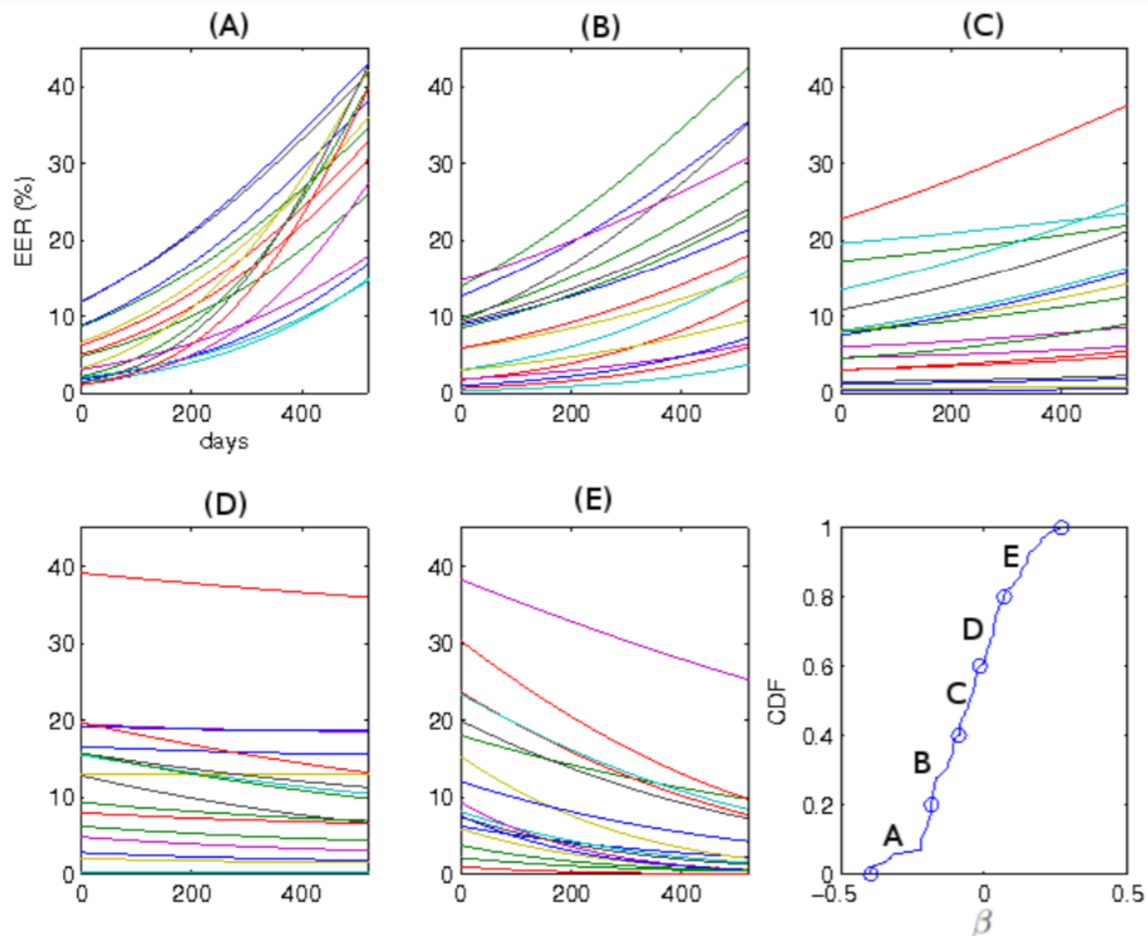
Degree of polynomial (2)



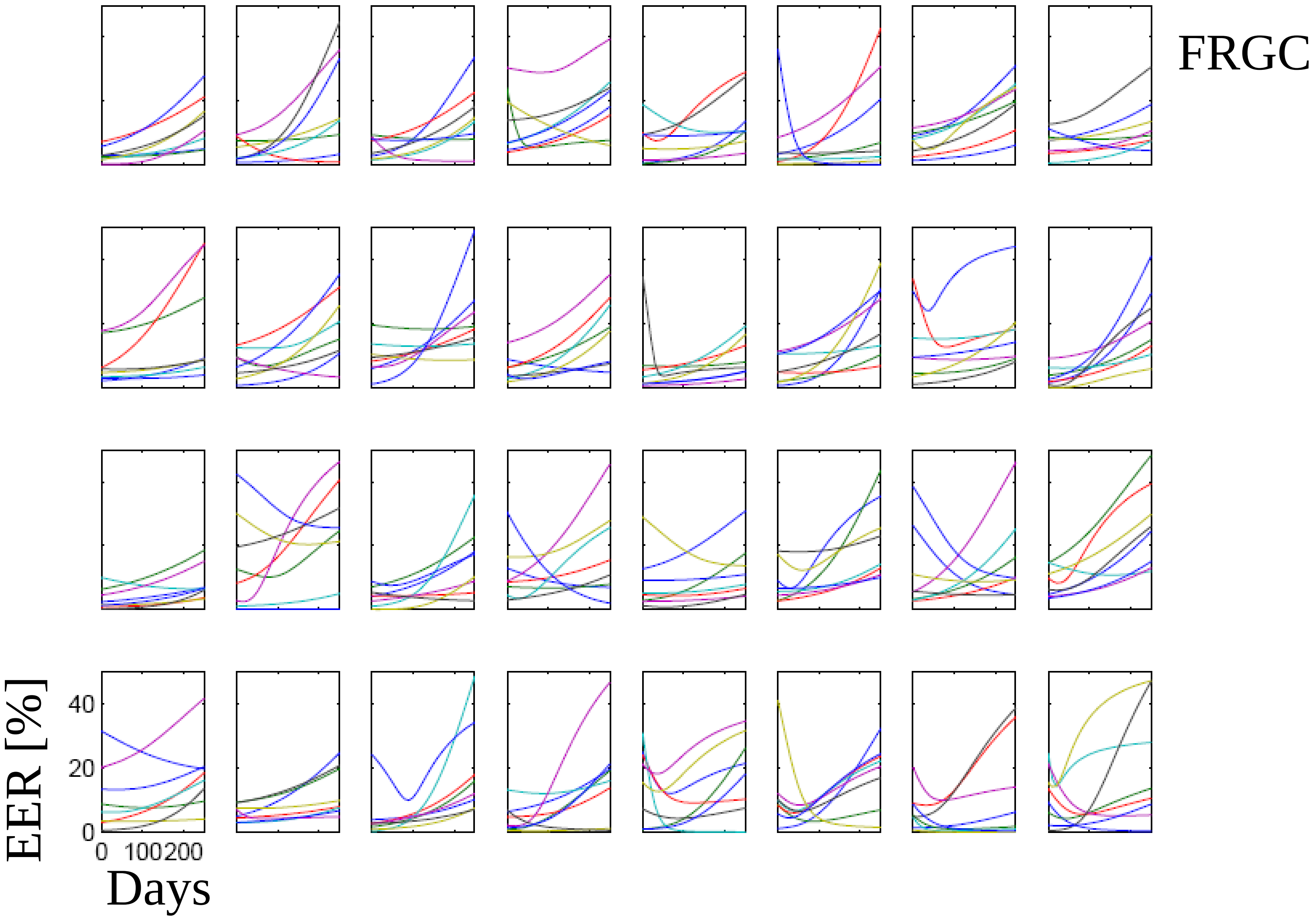
Degree of polynomial (3)



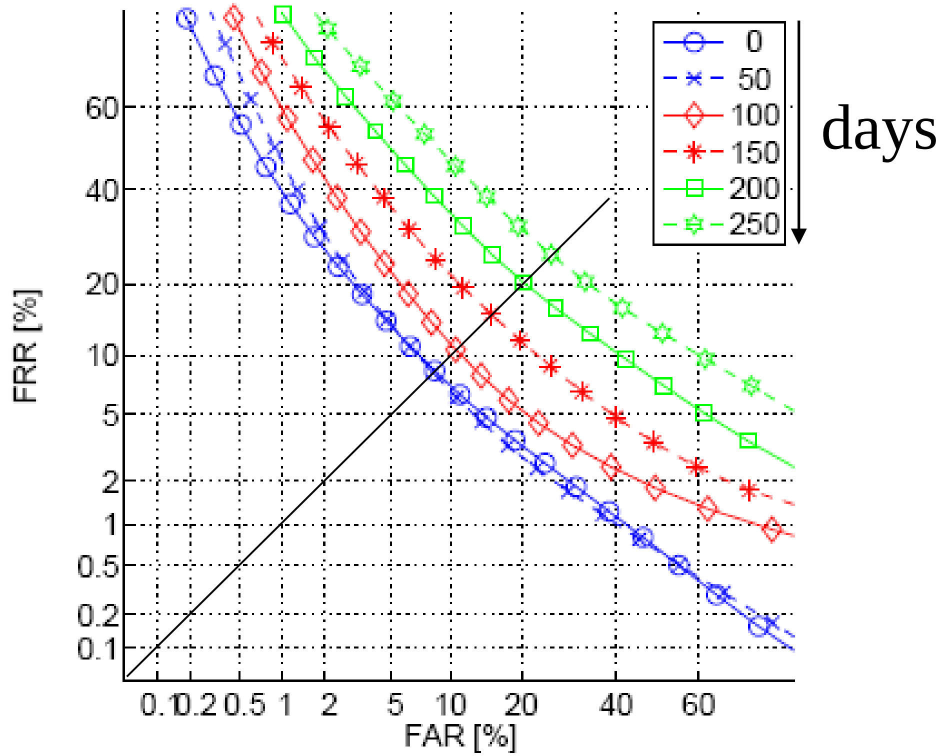
Group into 5 groups



EER of 256 Users



DET curve evolution



http://epubs.surrey.ac.uk/812522/1/norman_eer_trend_mmbio_journal_v1.pdf

Menu

Special issue

Special Issue: Biometric Performance and Statistics

Biometrics statistics: a foreword and introduction to the special issue
Author(s): Norman Poh and Michael Schuckers
Source: IET Biometrics, Volume 4, Issue 4, p. 206 –208
DOI: 10.1049/iet-bmt.2015.0100
Type: Article
Show details

Modelling errors in a biometric re-identification system
Author(s): Brian DeCann and Arun Ross
Source: IET Biometrics, Volume 4, Issue 4, p. 209 –219
DOI: 10.1049/iet-bmt.2015.0061
Type: Article
Show details

Performance evaluation of continuous authentication systems
Author(s): Patrick Bours and Soumik Mondal
Source: IET Biometrics, Volume 4, Issue 4, p. 220 –226
DOI: 10.1049/iet-bmt.2014.0070
Type: Article
Show details

Impact of (segmentation) quality on long vs. short-timespan assessments in iris recognition performance
Author(s): Peter Wild ; James Ferryman ; Andreas Uhl
Source: IET Biometrics, Volume 4, Issue 4, p. 227 –235
DOI: 10.1049/iet-bmt.2014.0073
Type: Article
Show details

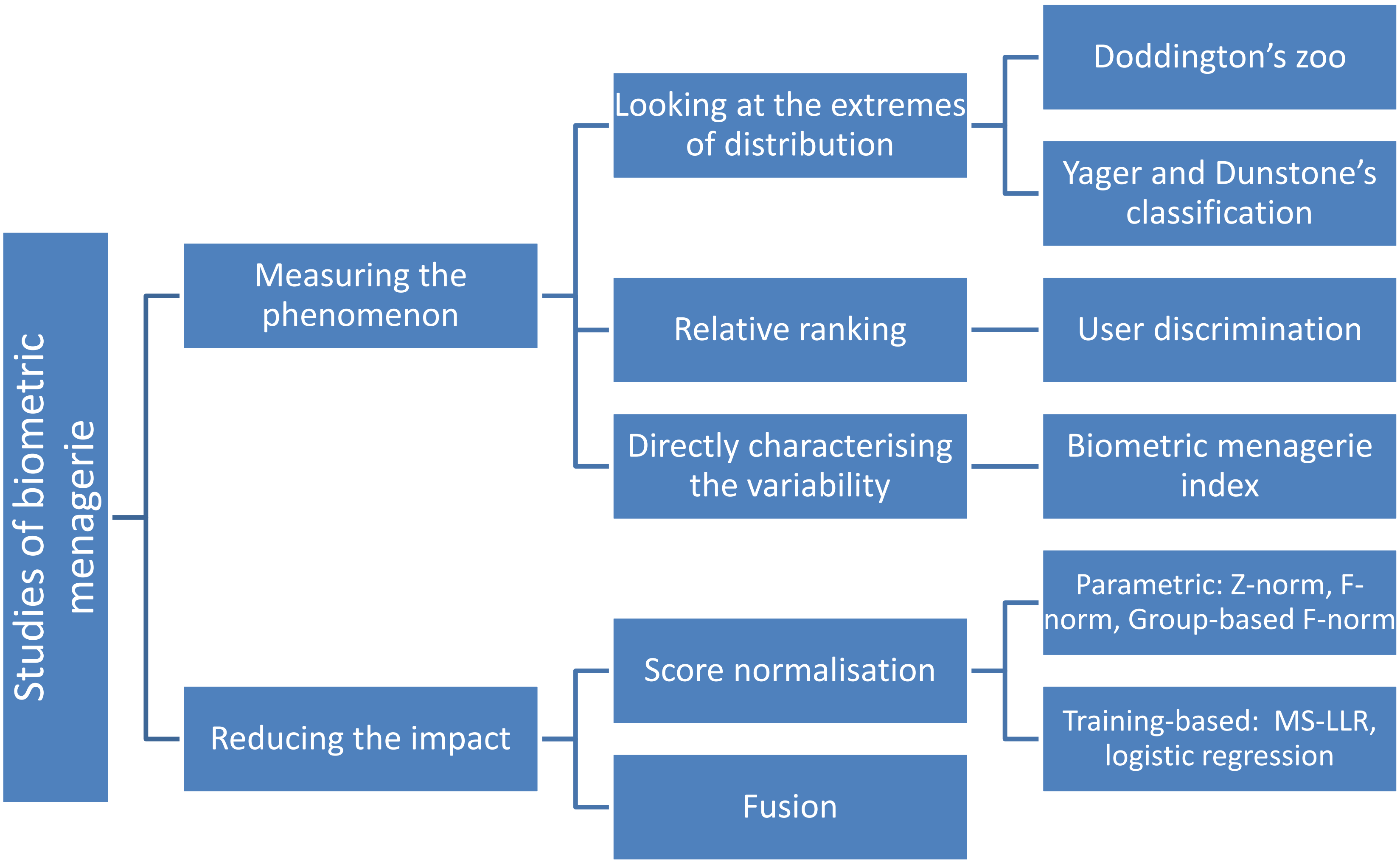
Algorithm to estimate biometric performance change over time
Author(s): Norman Poh ; Josef Kittler ; Chi-Ho Chan ; Medha Pandit
Source: IET Biometrics, Volume 4, Issue 4, p. 236 –245
DOI: 10.1049/iet-bmt.2014.0107
Type: Article
Show details

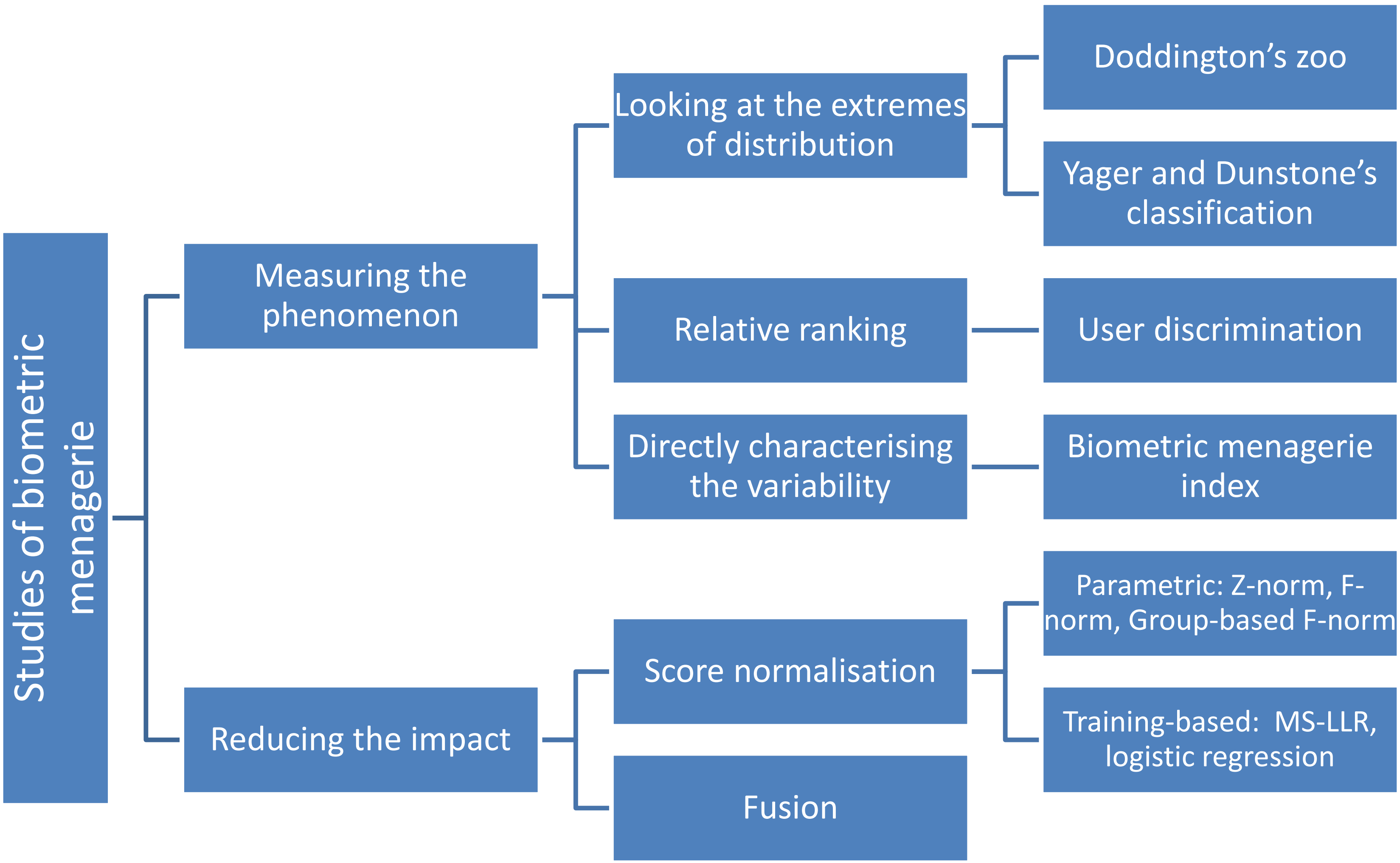
https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2014.0107;jsessionid=xe60qfewntgh.x-iet-live-01

Classification of literature

BIBLIOGRAPHY NOTES

Classification of research in Biometric Menagerie




Studies of biometric menagerie
Measuring the phenomenon
Reducing the impact
Looking at the extremes of distribution
Relative ranking
Directly characterising the variability
Score normalisation
Fusion
Doddington's zoo
Yager and Dunstone's classification
User discrimination
Biometric menagerie index
Parametric: Z-norm, F-norm, Group-based F-norm
Training-based: MS-LLR, logistic regression

Some references

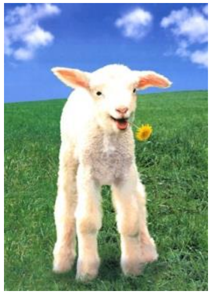
Doddington's zoo	Doddington, George, et al. *Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation*. NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD, 1998.
Yager and Dunstone's classification	Yager, Neil, and Ted Dunstone. "The biometric menagerie." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.2 (2010): 220-230.
User discrimination	* N. Poh and J. Kittler, A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion, in 6th Biometrics Symposium, pp. 17–22, 2008. * N. Poh, A. Ross, W. Li, and J. Kittler, A User-Specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects, Pattern Recognition 46(12): 3341-57, 2013.
Biometric menagerie index	N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Modelspecific Variation, in Int'l Conf. on Biometrics (ICB'09), 2009.
Parametric: Z-norm, F-norm, Group-based F-norm	* N. Poh, A. Rattani, M. Tistarelli and J. Kittler, Group-specific Score Normalization for Biometric Systems, in IEEE Computer Society Workshop on Biometrics (CVPR), pages 38-45, 2010. * N. Poh and S. Bengio, An Investigation of F-ratio Client-Dependent Normalisation on Biometric Authentication Tasks, in IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP), volume I, pages 721-724, 2005
Training-based: MS-LLR, logistic regression	* N. Poh and M. Tistarelli, Customizing Biometric Authentication Systems via Discriminative Score Calibration, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012 * N. Poh and J. Kittler, Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems, IEEE Trans. Audio, Speech and Language Processing, 16(3):594-606, 2008.

Menu

DODDINGTON’S ZOO

Doddington's Zoo

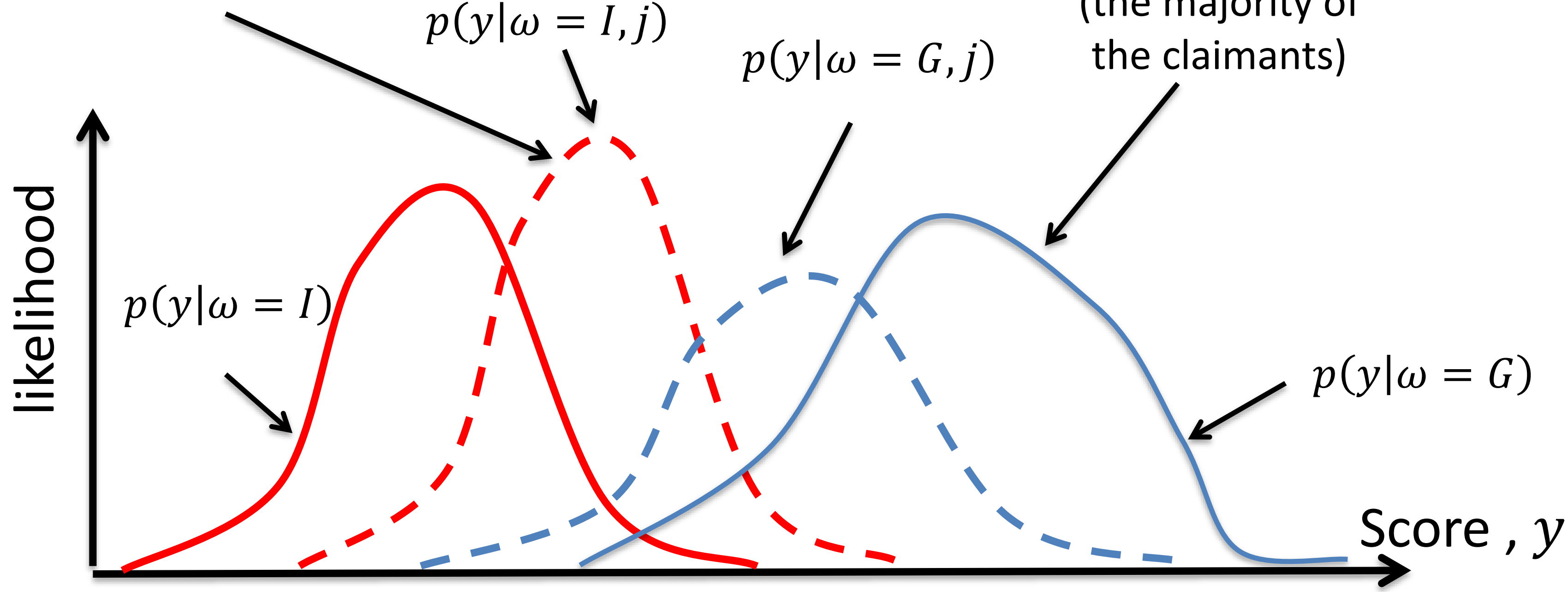
Strong impostor



High FAR

High FRR

Good clients (the majority of the claimants)



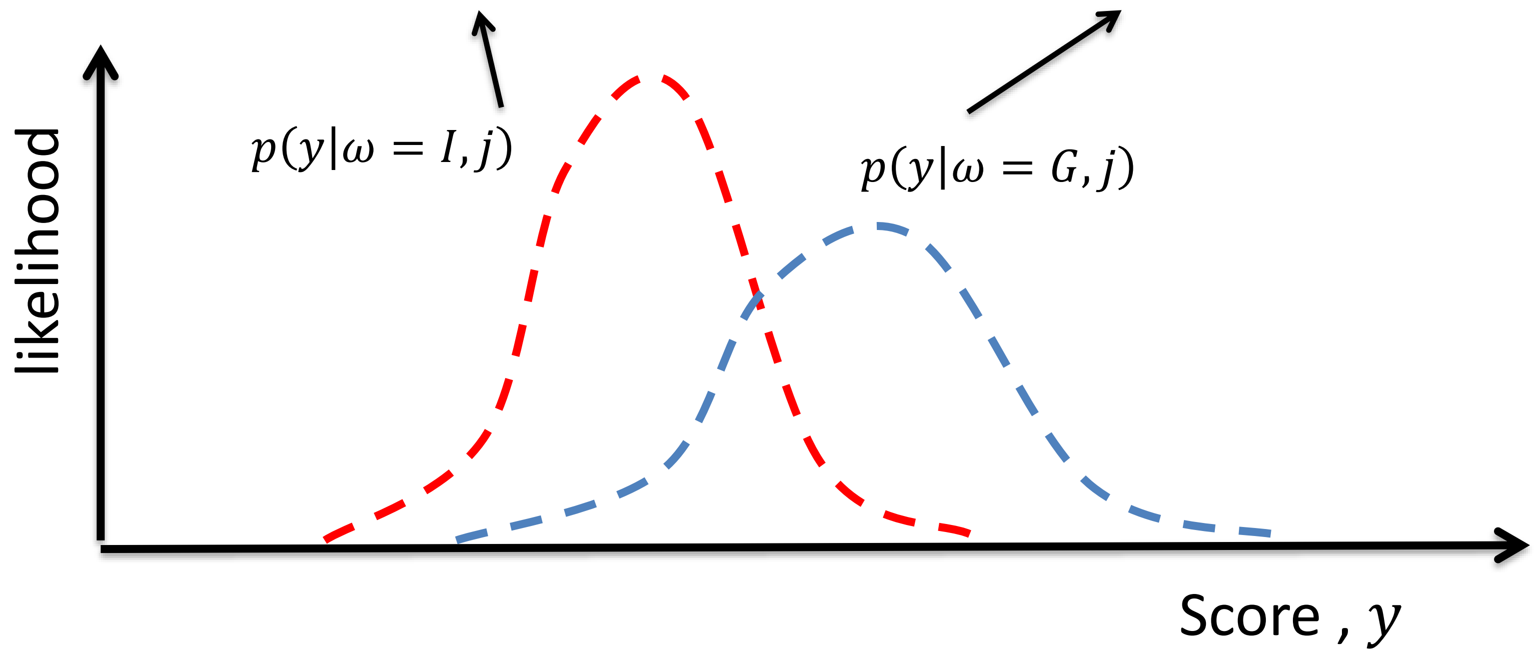
Doddington, George, et al. *Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation*. NATIONAL INST OF STANDARDS AND TECHNOLOGY GAITHERSBURG MD, 1998.

Characterising claimant-specific *pdf*

Need more and more data

- First order: μ_j^I
- Second order: $\left(\sigma_j^I\right)^2$
- Skewness: ${m_3}_j^I$
- Kurtosis: ${m_4}_j^I$

- First order: μ_j^G
- Second order: $\left(\sigma_j^G\right)^2$
- Skewness: ${m_3}_j^G$
- Kurtosis: ${m_4}_j^G$

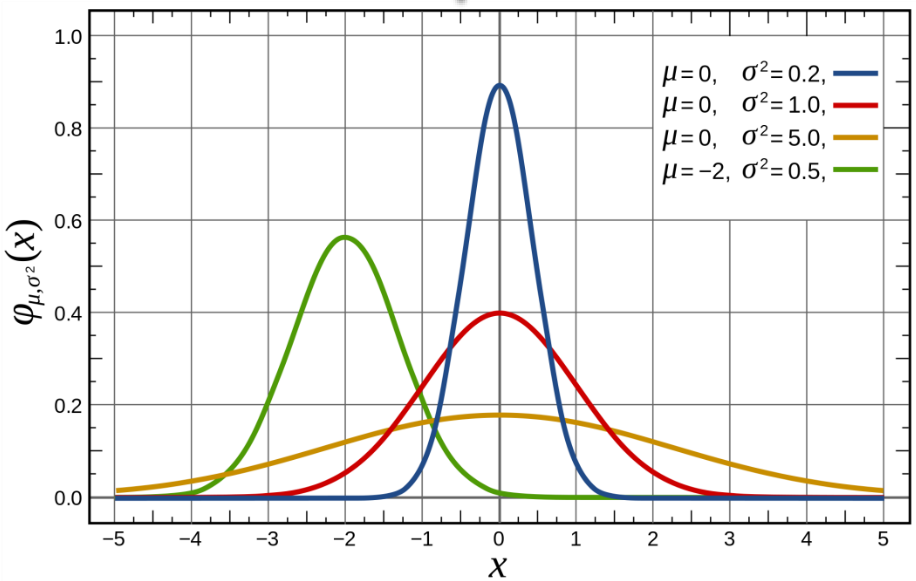


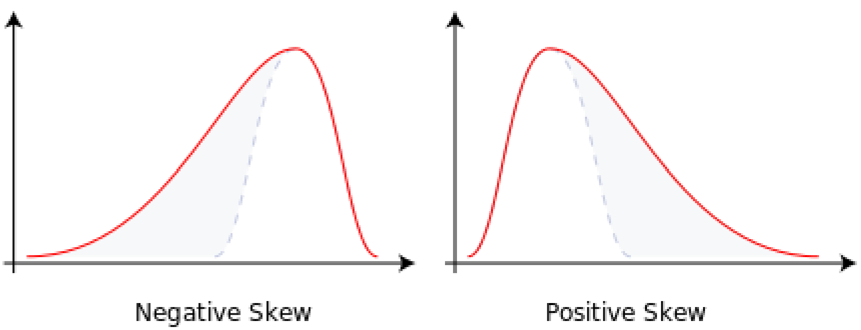
Mean
$\mu = E_y[y]$

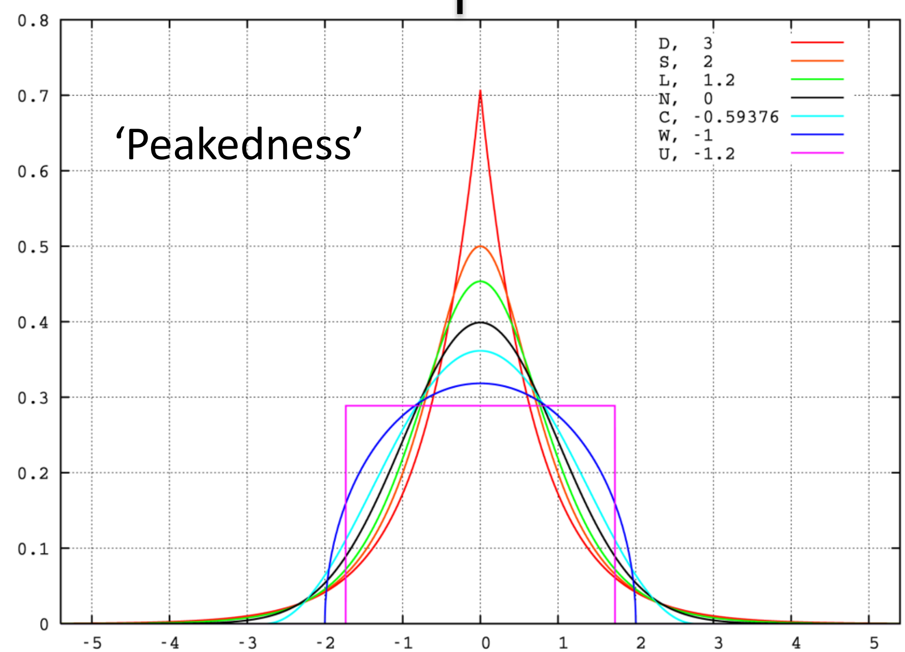
Variance
$\sigma^2 = E_y[(\mu - y)^2]$

Skewness
$\frac{E_y[y-\mu]^3}{\sigma^3}$

Kurtosis
$\frac{E_y[y-\mu]^4}{\sigma^4}$



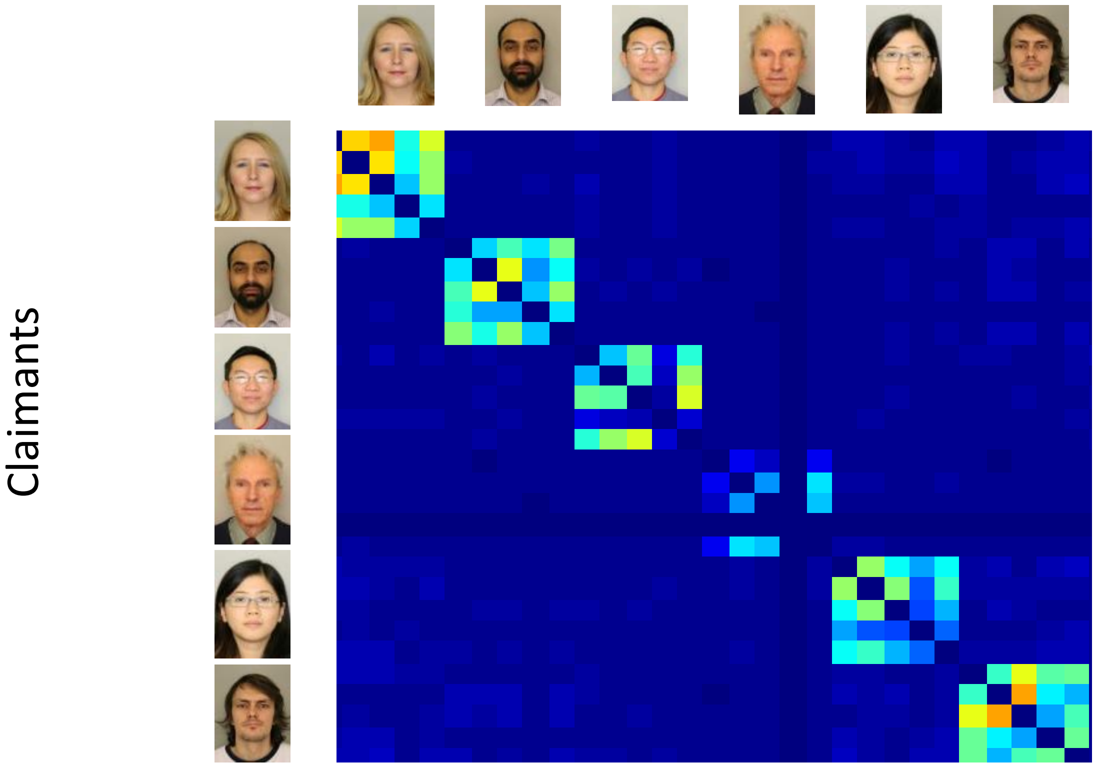




Source: Wikipedia entries: mean, variance, skewness ,and kurtosis

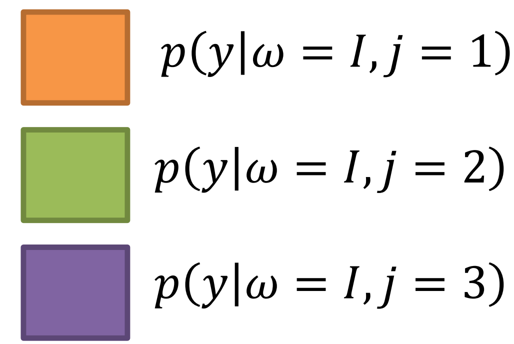
What's the difference between a wolf and a lamb???

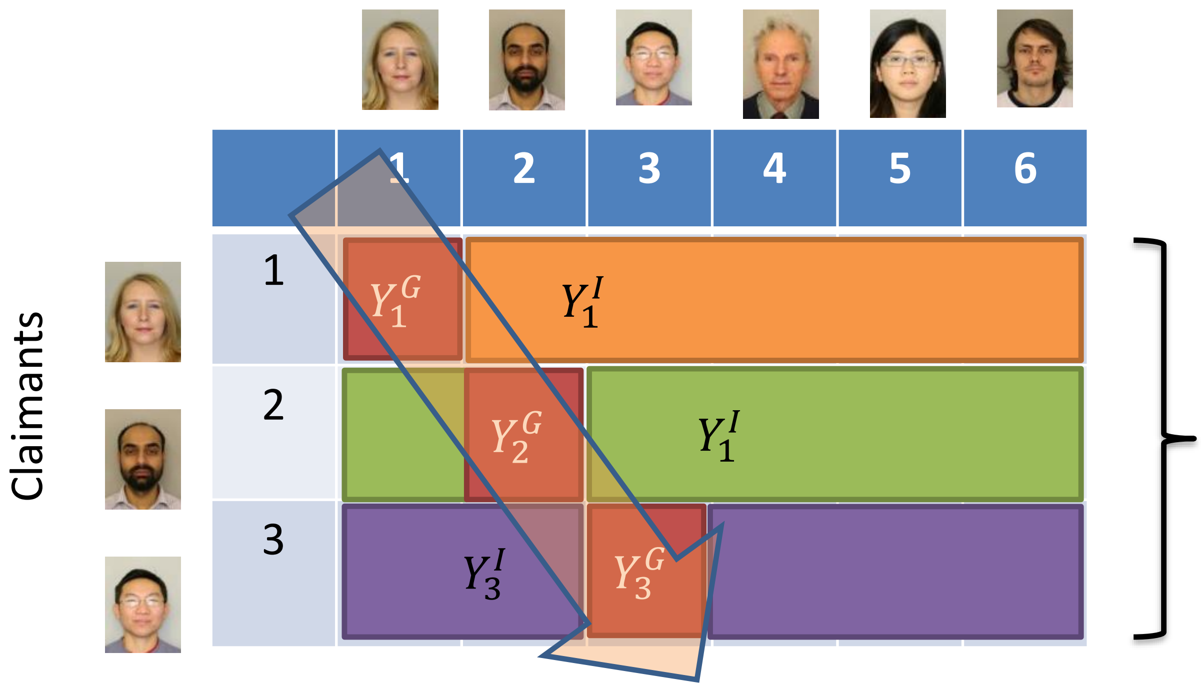
Each subject may have multiple query images



What’s the difference between a wolf and a lamb???

Each subject may have multiple query images



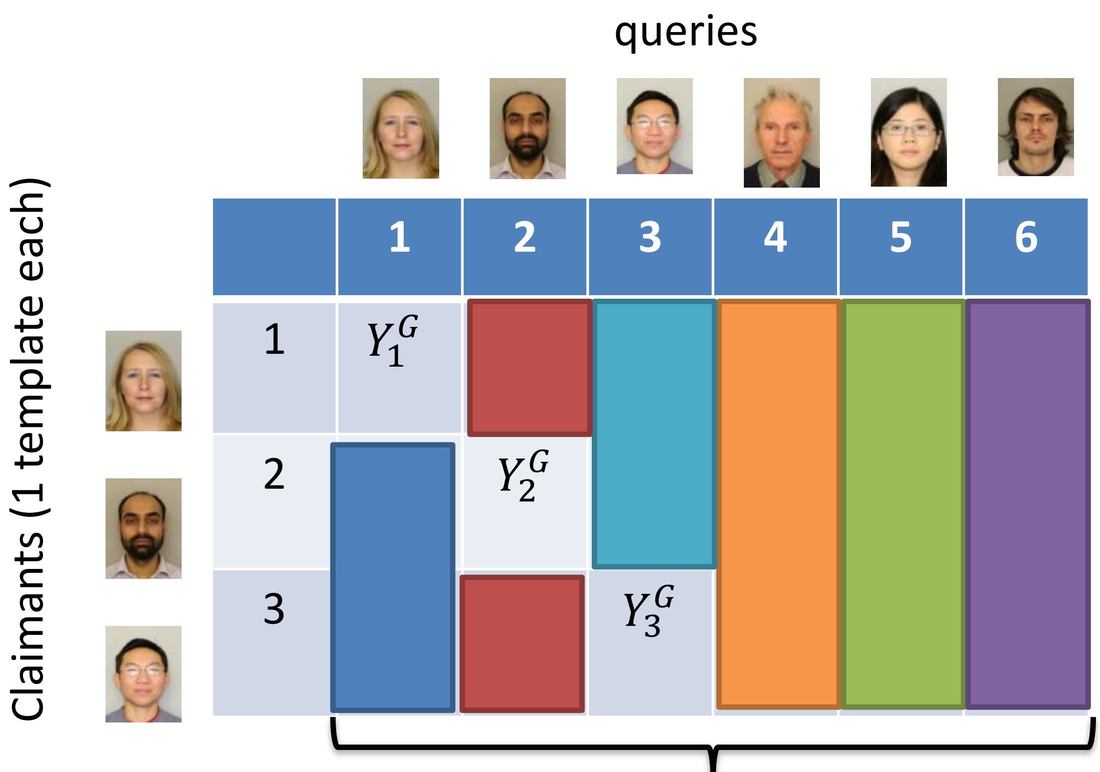


(1 template each)

Lambs are claimants whose zero-effort impostor scores are **relatively high**

Goats are claimants whose genuine scores are **relatively low**

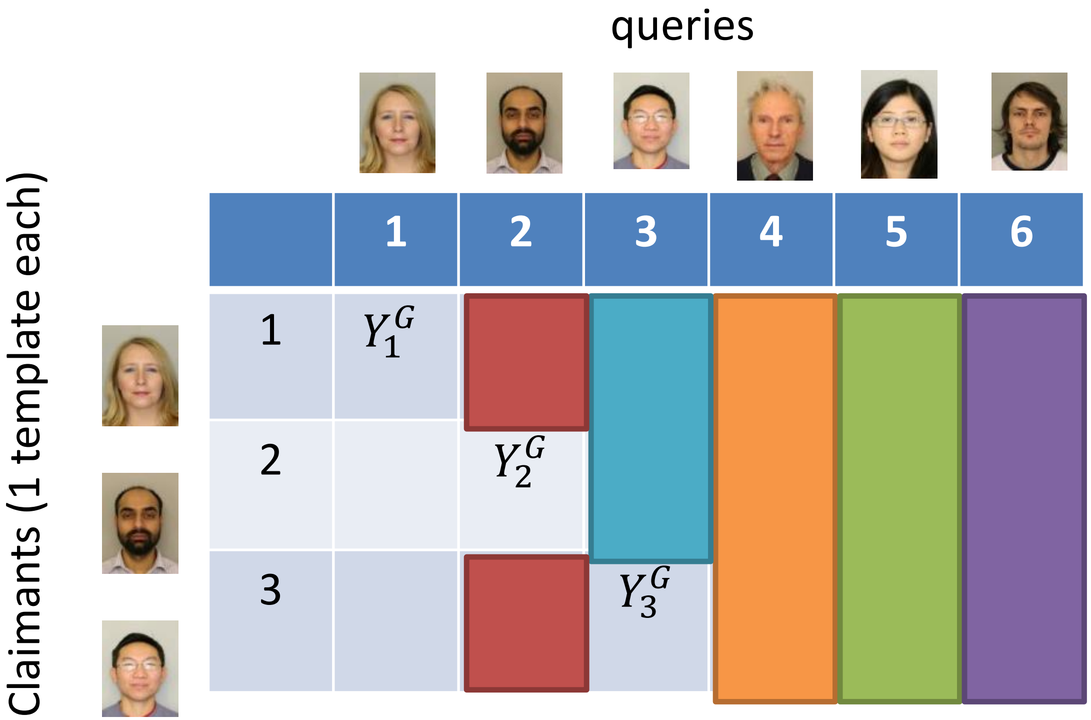
What's the difference between a wolf and a lamb???



Wolves are those who can generate very high ZE impostor scores

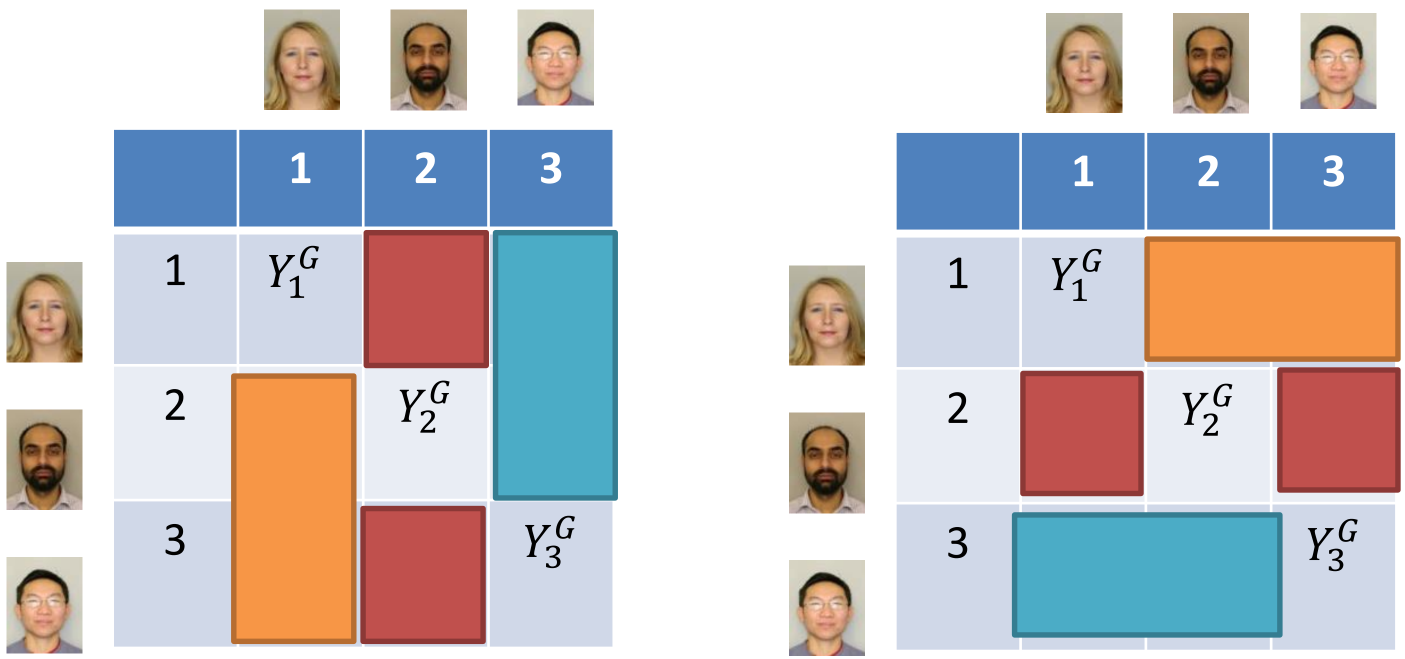
Claimants can be wolves, too!

Open-set experiment



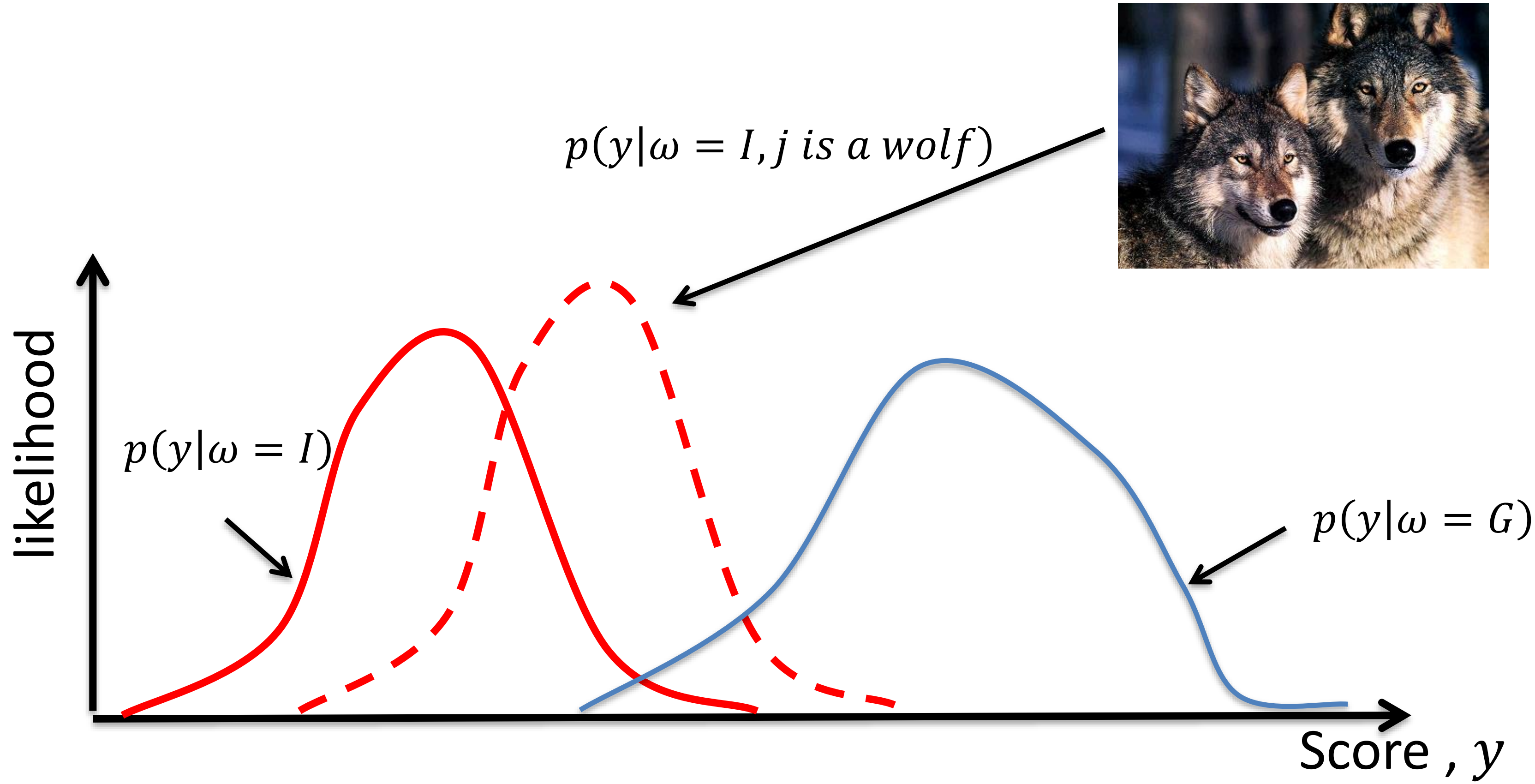
Wolves are lambs and lambs are wolves when …

We use a symmetric matcher and consider a closed set



A symmetric matcher is one that satisfies: M(a,b) = M(b,a)

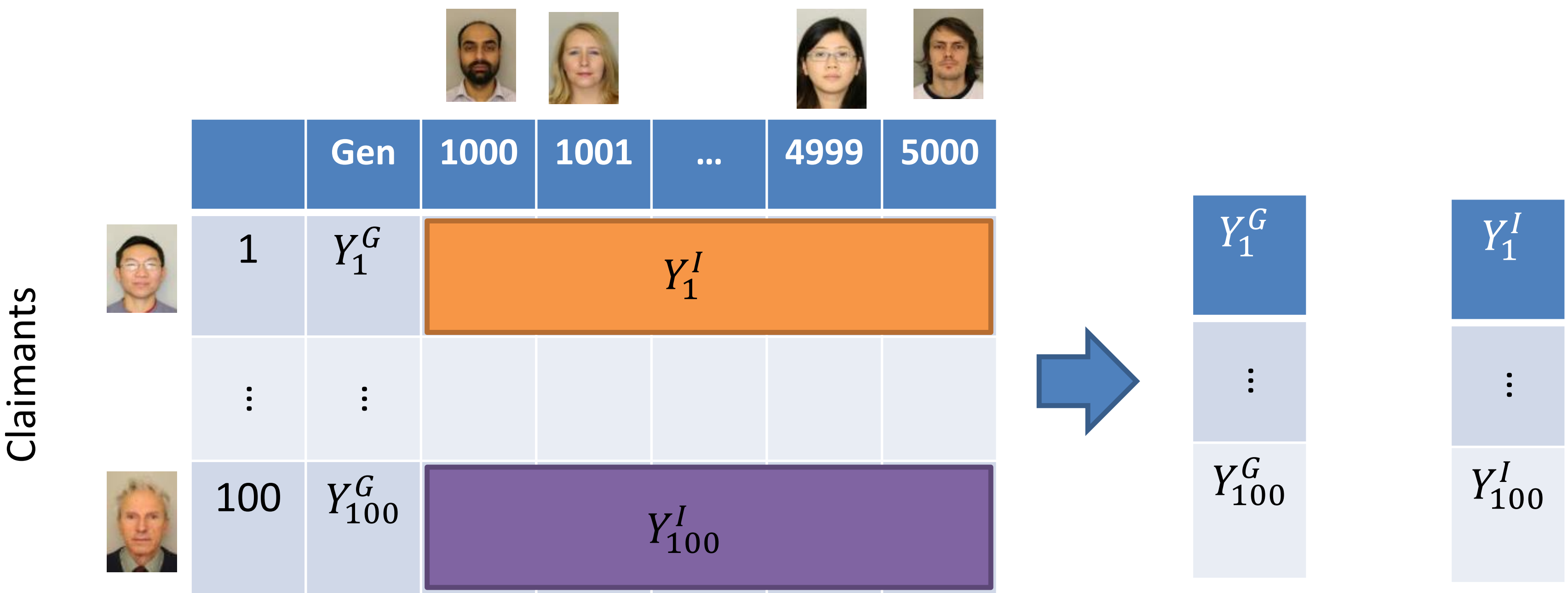
Why do wolves matter?



Murakami, Takao, Kenta Takahashi, and Kanta Matsuura. "Towards optimal countermeasures against wolves and lambs in biometrics." *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*. IEEE, 2012.

Summary: Goats and lamb detection

Each subject may have multiple query images

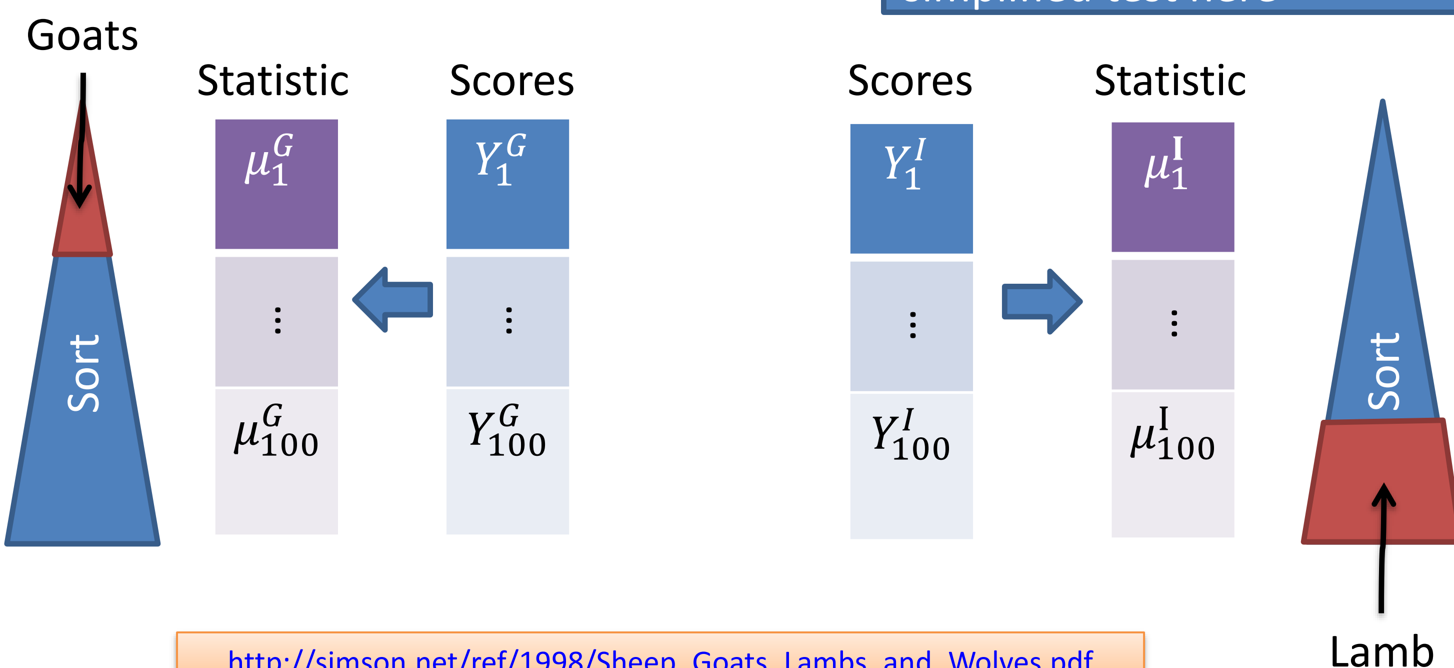


Summary: Goats and lamb detection

Y_1^G	Y_1^I
$\vdots$	$\vdots$
Y_{100}^G	Y_{100}^I

Summary: Goats and lamb detection

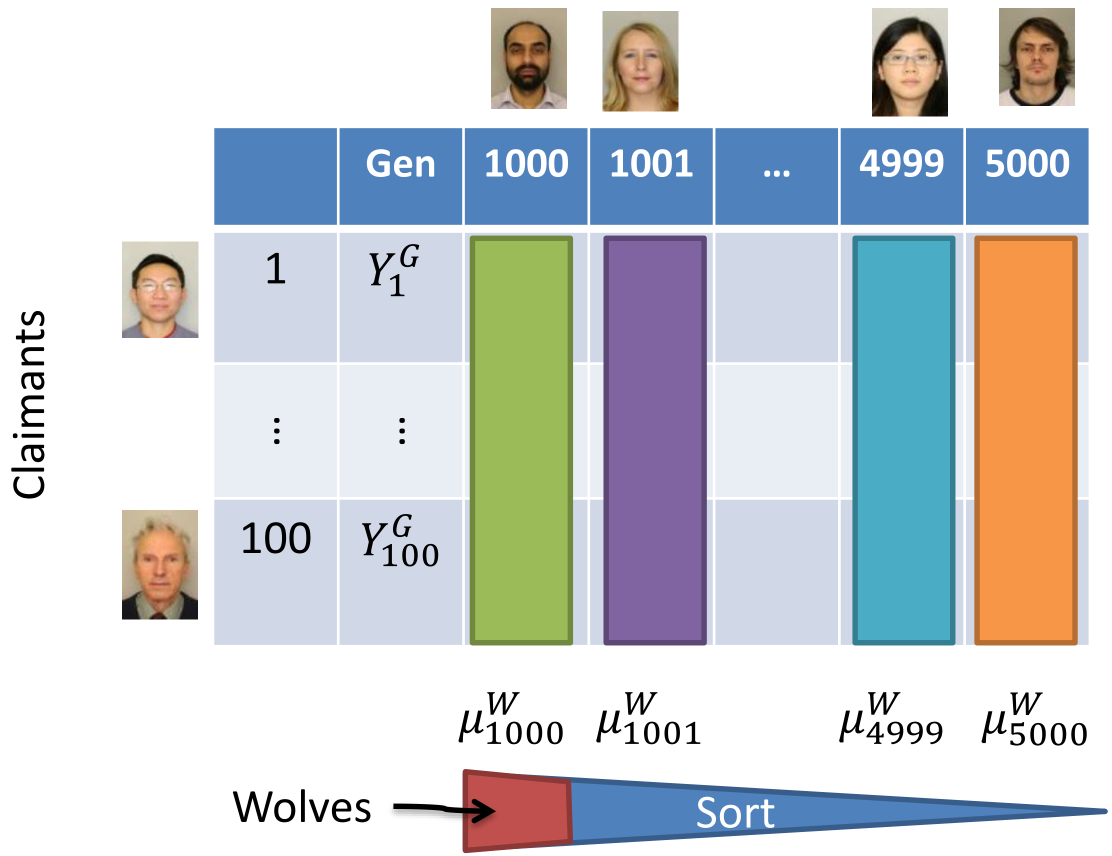
The original Goat test was implemented using **nonparametric Kruskal-Wallis rank sum test**; we use a simplified test here



http://simson.net/ref/1998/Sheep_Goats_Lambs_and_Wolves.pdf

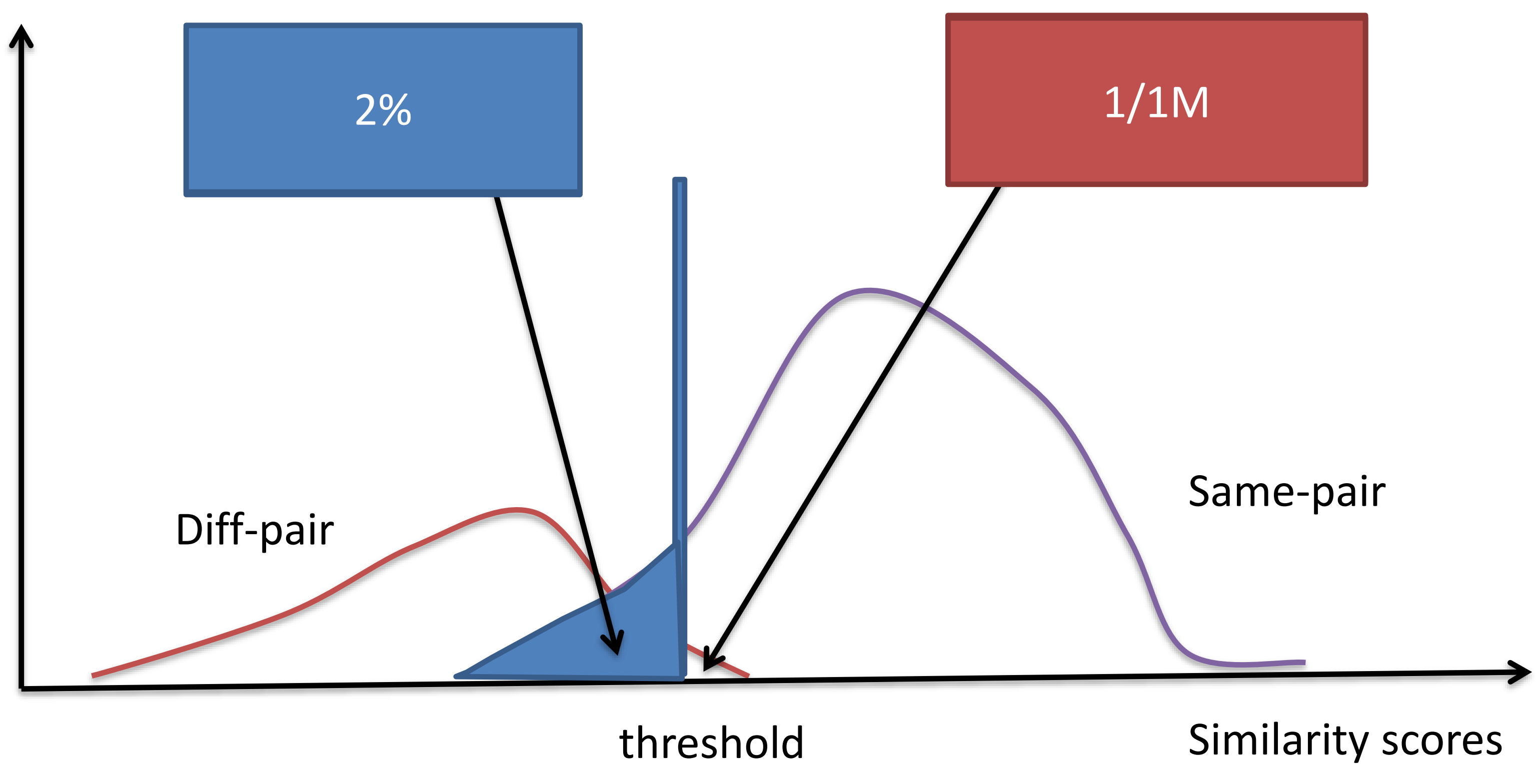
Summary: Wolves detection

Each subject may have multiple query images



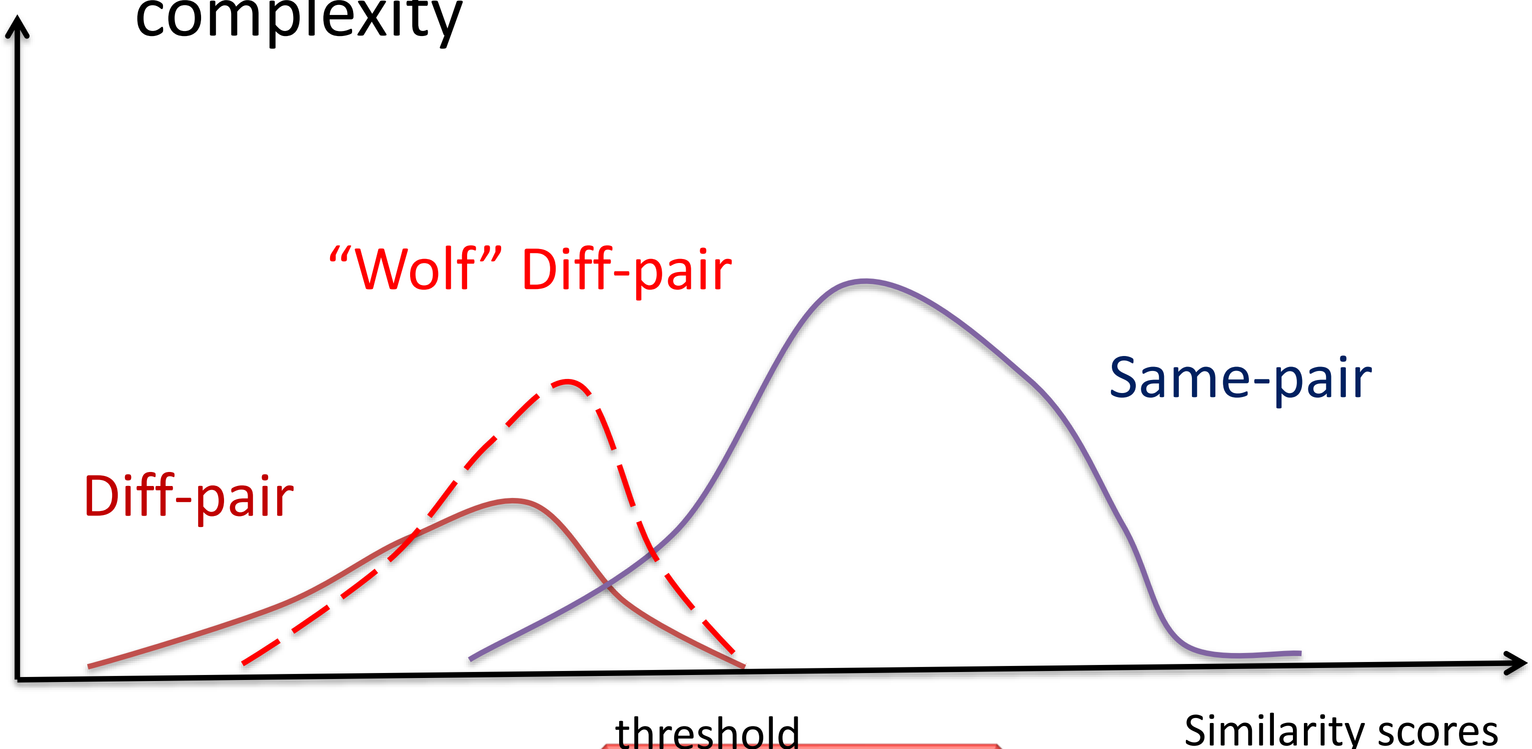
Discussion: How were the 1M nonmatch scores chosen?

Vendor: “Our false rejection is 2% when operating at a false acceptance rate of 1 in a million”



Wolf diff-pair

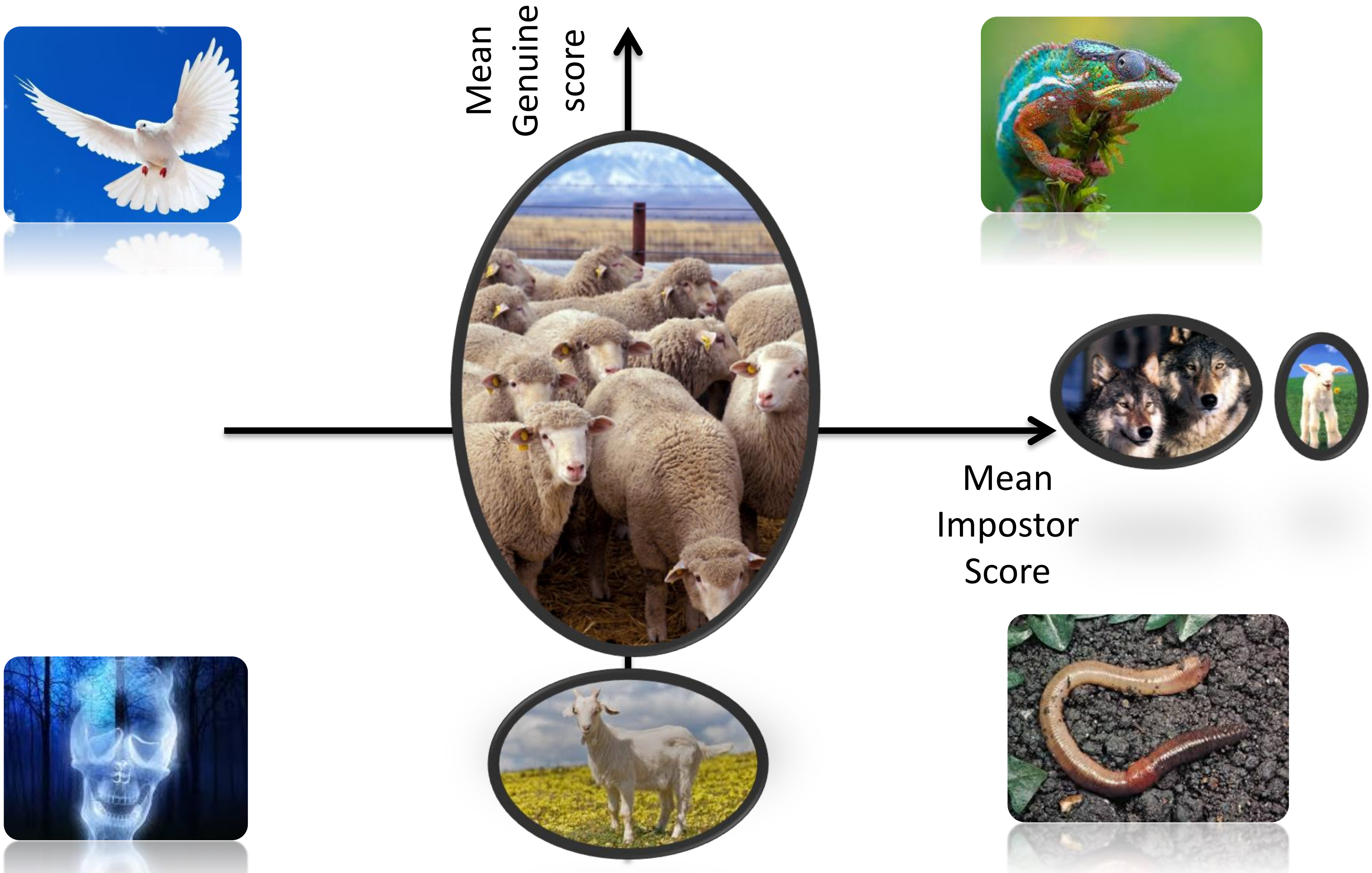
- How will the performance look like using difficult samples chosen based on the wolf?
- Why would you do that? Computational complexity



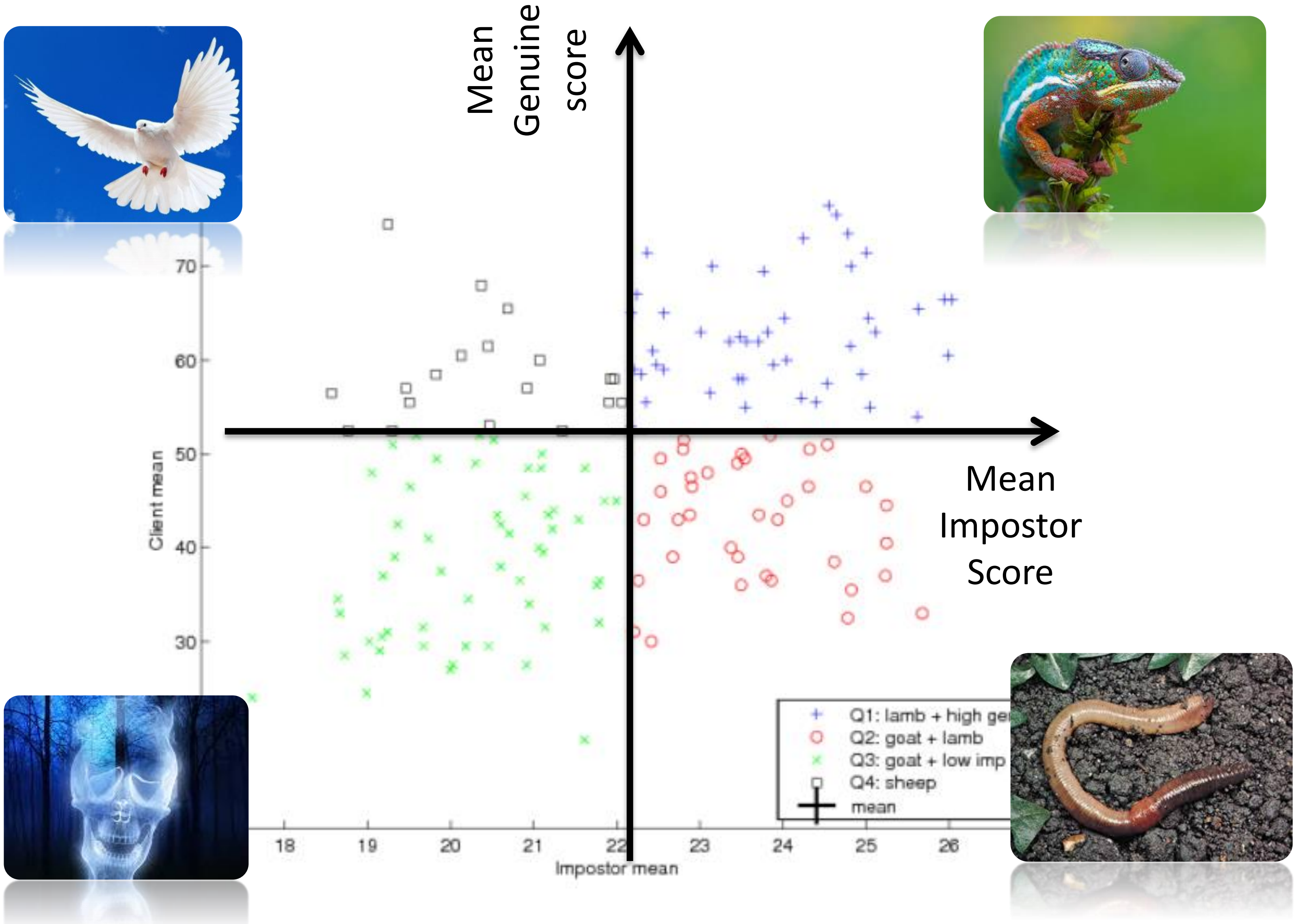
Menu

YAGER AND DUNSTONE'S CLASSIFICATION

Biometric Menagerie

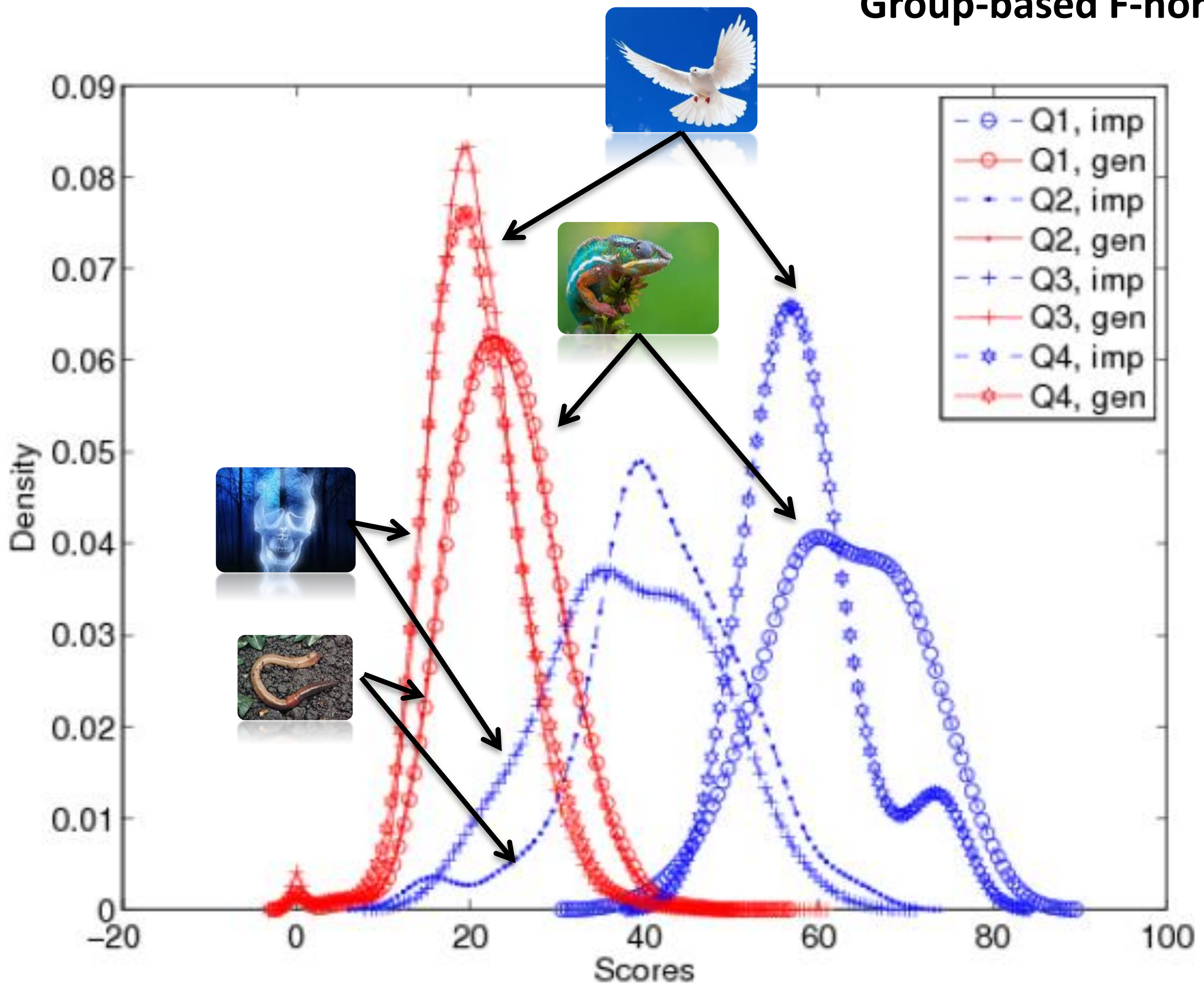


Yager, Neil, and Ted Dunstone. "The biometric menagerie." Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.2 (2010): 220-230.



N. Poh, A. Rattani, M. Tistarelli and J. Kittler, **Group-specific Score Normalization for Biometric Systems**, in IEEE Computer Society Workshop on Biometrics (CVPR), pages 38-45, 2010.

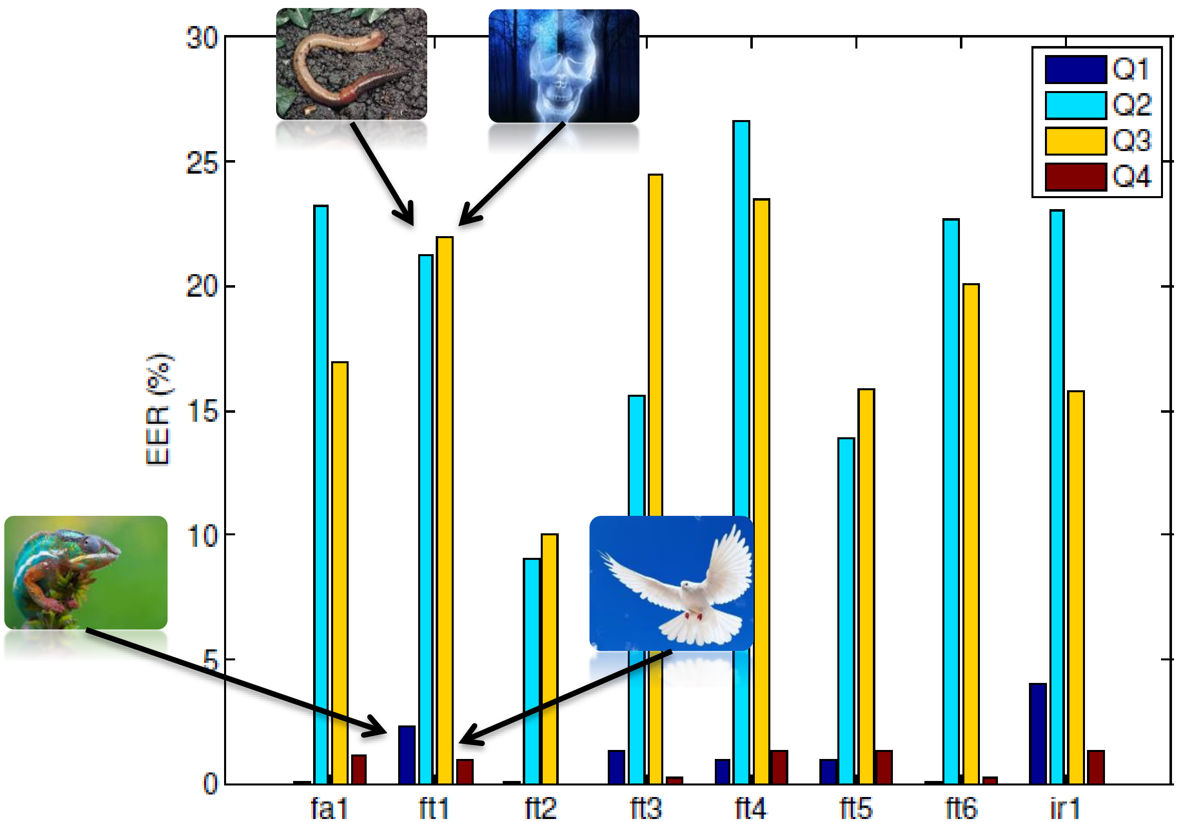
Group-based F-norm



N. Poh, A. Rattani, M. Tistarelli and J. Kittler, **Group-specific Score Normalization for Biometric Systems**, in IEEE Computer Society Workshop on Biometrics (CVPR), pages 38-45, 2010.

Menu

Error rates



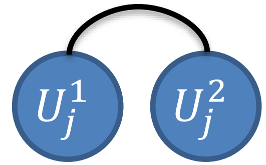
ON THE PERSISTENCE OF BIOMETRIC MENAGERIE

Key issues

- Are subjects inherently “hard to match”?
  - Match score analysis: This question depends on how well we understand the environment in which biometric operates, thus requiring some generalization capability to *unseen* data
  - Nonmatch score analysis: This can be answered!
- Why biometric menagerie: a conjecture

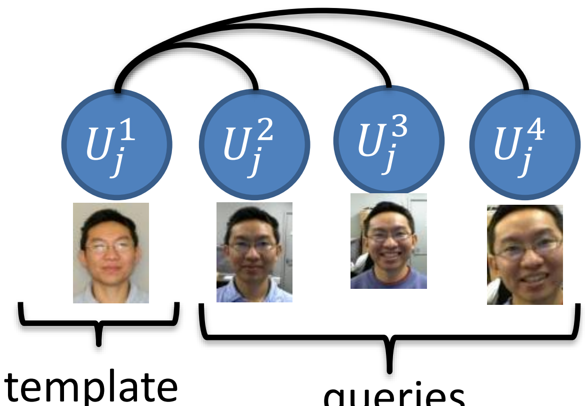
Are subjects inherently hard to match? Same-pair score analysis

$\{U_j^1, U_j^2, U_j^3, U_j^4\}$



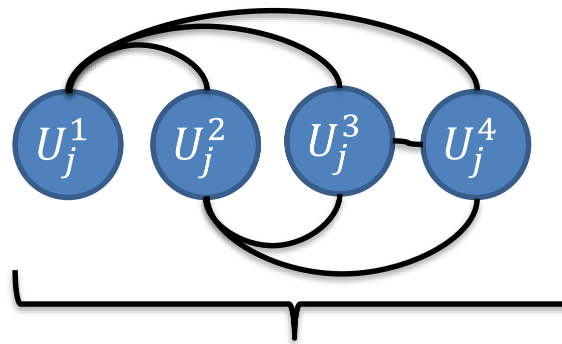
$M(U_j^1, U_j^2)$

Scenario



n-1 scores

VS



$\binom{n}{2} = \frac{n(n-1)}{2}$ scores

Application

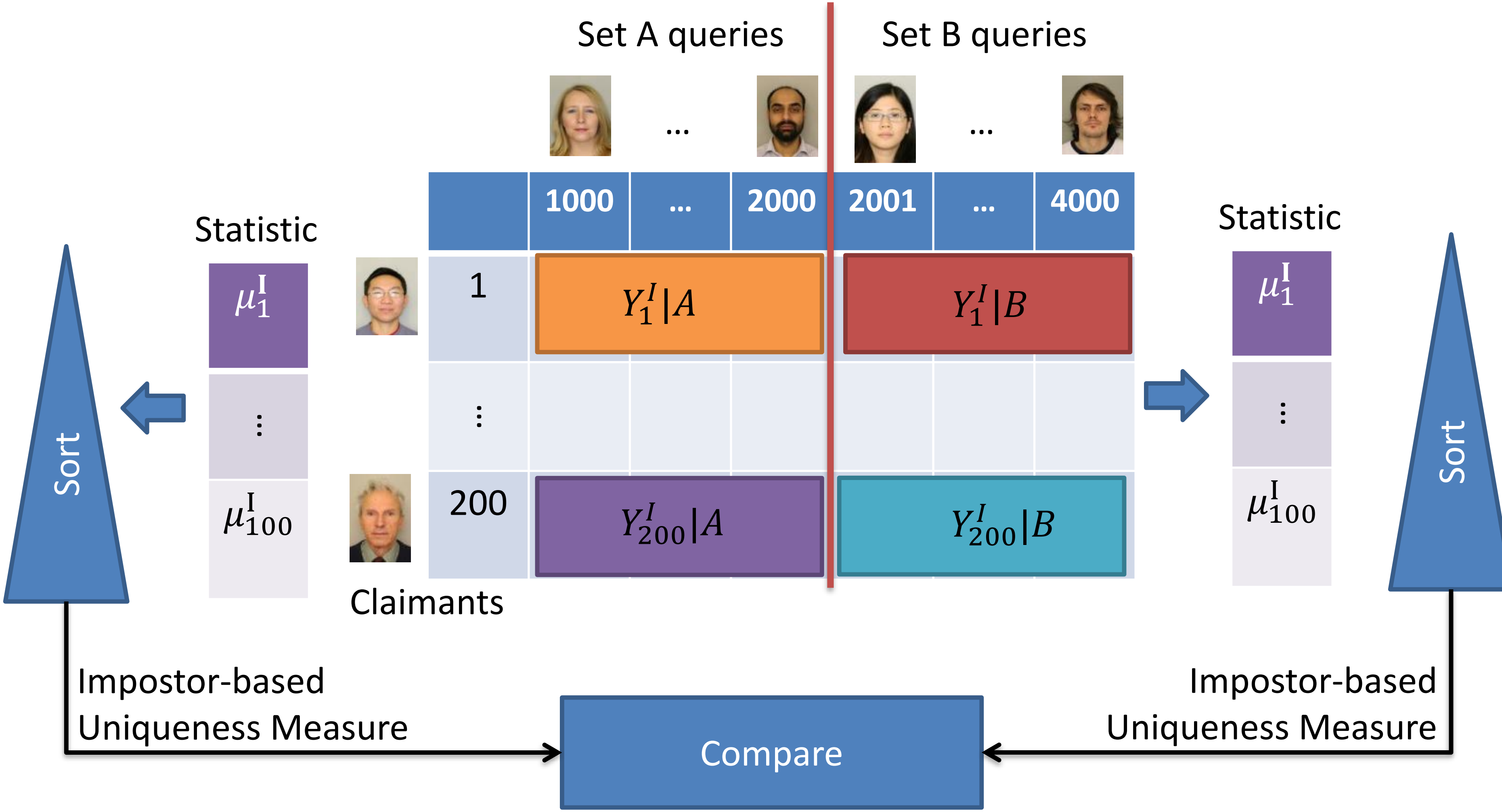
User-specific score calibration: to reduce the menagerie effect

Poh, Norman, and Josef Kittler. "Incorporating variation of model-specific score distribution in speaker verification systems." *IEEE Transactions on Audio, Speech and Language Processing* 16.3 (2008): 594-606.

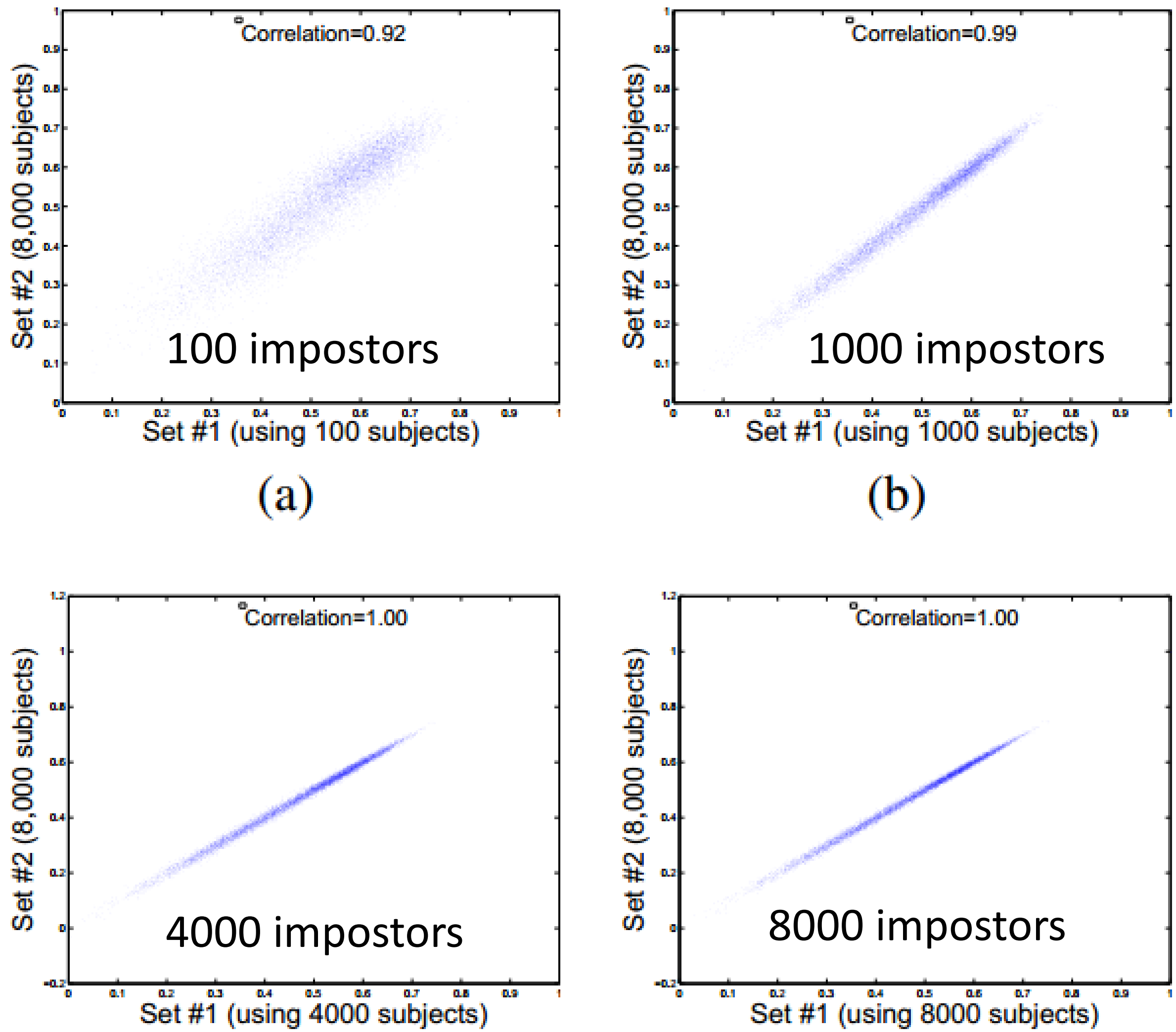
Template selection: to find a 'representative' template

Jain, Anil, Umut Uludag, and Arun Ross. "Biometric template selection: a case study in fingerprints." *Audio-and Video-Based Biometric Person Authentication*. Springer Berlin Heidelberg, 2003.

Are subjects inherently hard to match?
Diff-pair score analysis

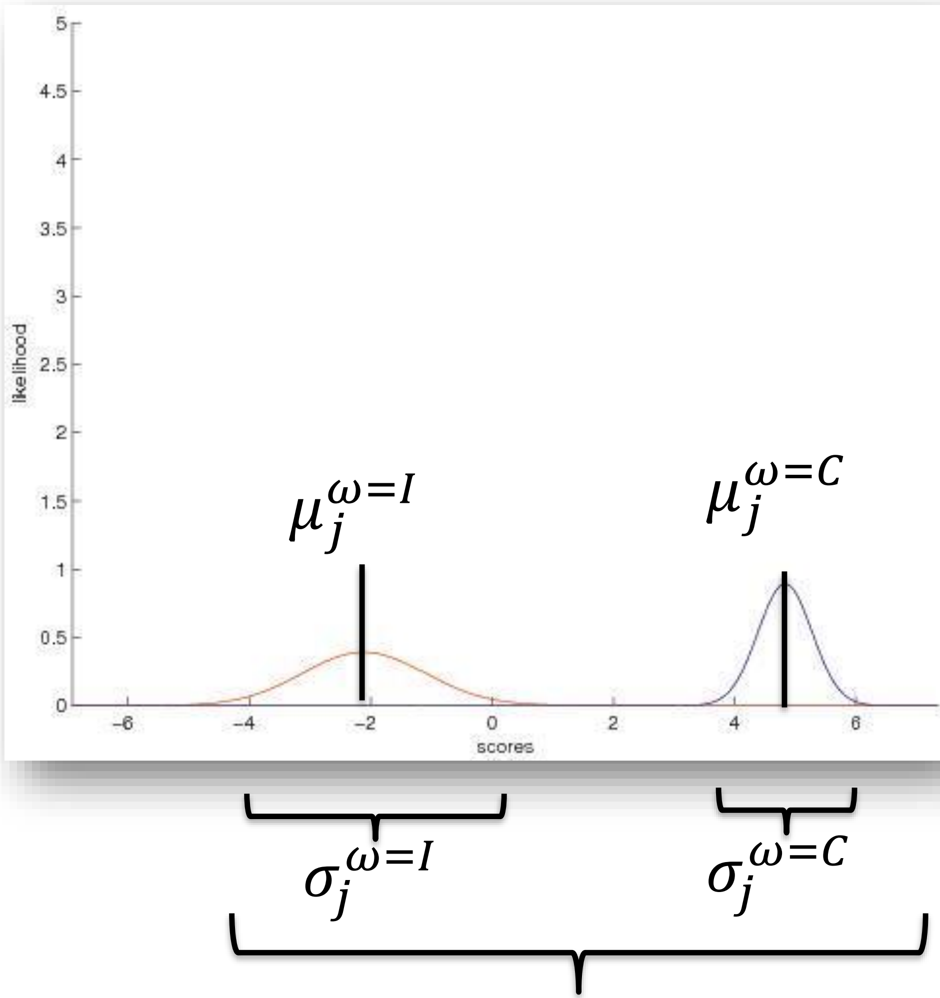


Non-match scores are very stable



Klare, Brendan F., and Anil K. Jain. "Face recognition: Impostor-based measures of uniqueness and quality." *Biometrics: Theory, Applications and Systems (BTAS), 2012 IEEE Fifth International Conference on*. IEEE, 2012.

XM2VTS score-level benchmark database



One data set

$\times$ 200 *claimants*
$\times$ 13 *experiments*

Download the database here: https://gitlab.com/normanpoh/xm2vts_fusion

Norman Poh and Samy Bengio, Database, Protocol and Tools for Evaluating Score-Level Fusion Algorithms in Biometric Authentication, Pattern Recognition, Volume 39, Issue 2, Pages 223-233, 2006.

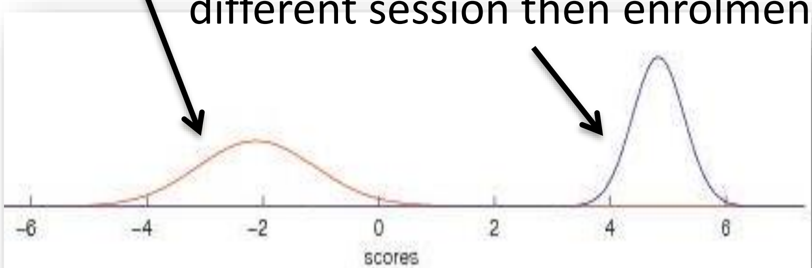
Generalisation ability of biometric menagerie

~ 1 month apart

Dev set ⟷ **Eva set**

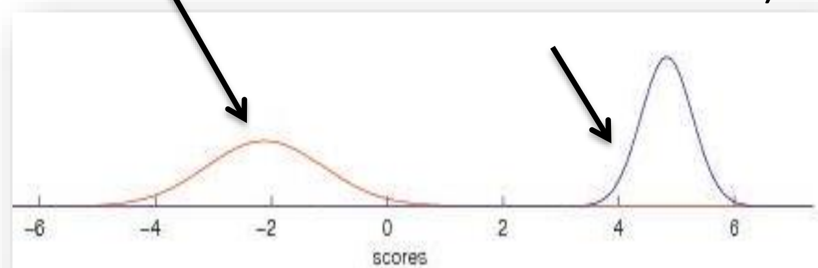
25 impostors × 8 samples

2 genuine samples from a different session then enrolment)



70 impostors × 8 samples

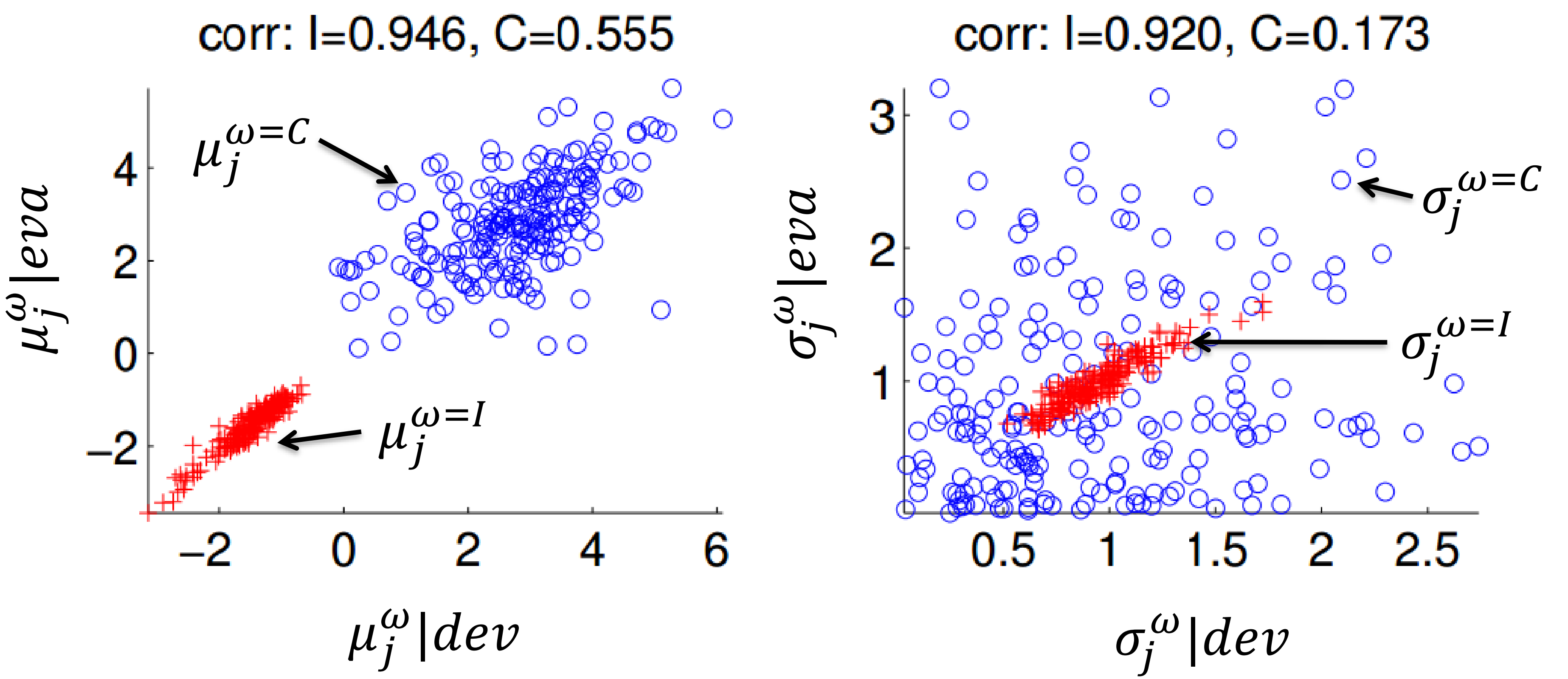
3 genuine samples (from a different session then enrolment)



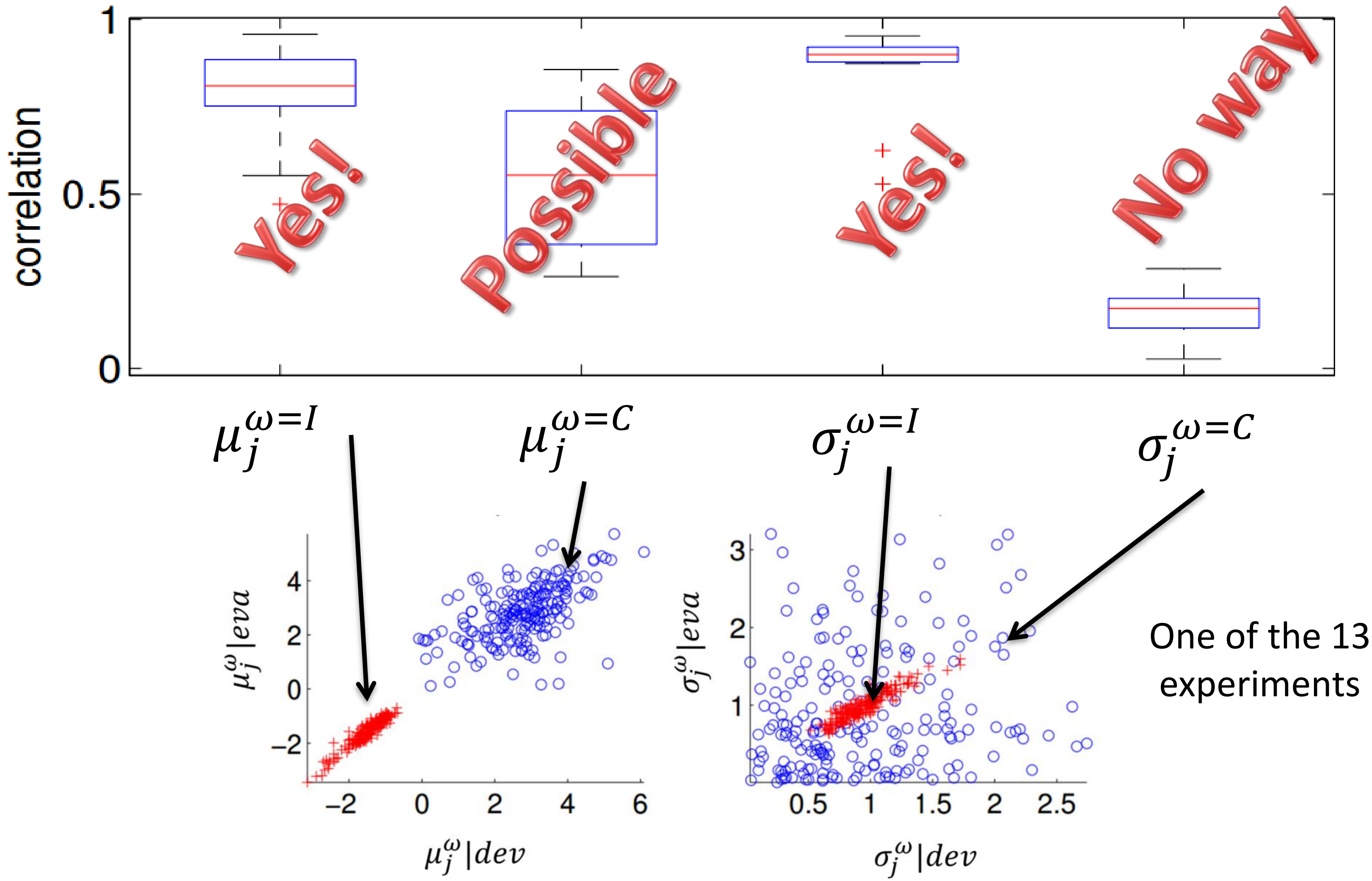
Can we predict the statistics?

N. Poh, A. Ross, W. Li, and J. Kittler, A User-Specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects, Pattern Recognition 46(12): 3341-57, 2013.

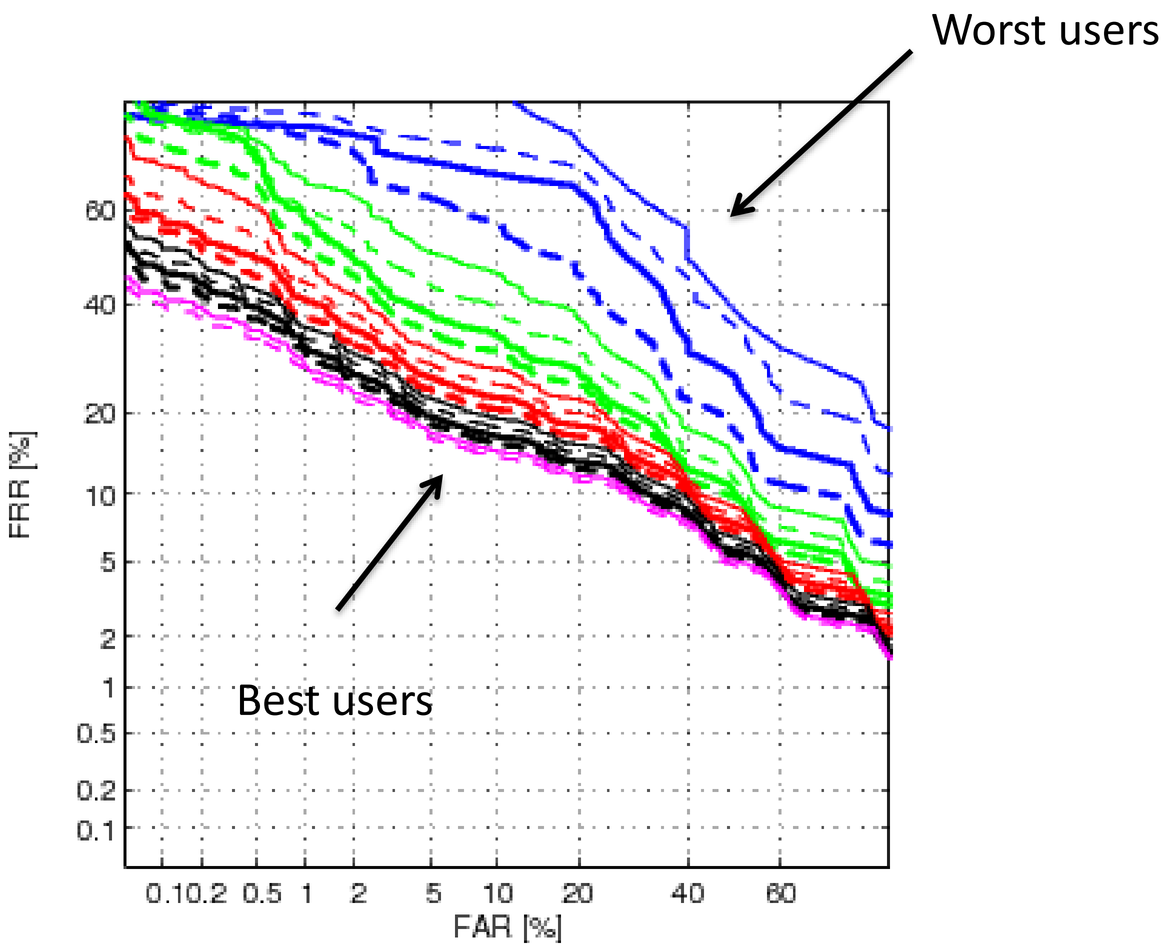
One of the 13 experiments



How predictable are the statistics?



Grouping users by their performance



Why is the subject bias so persistent???

L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

Conjectures

Biometric menagerie is system-dependent

- Feature space

Teli, Mohammad Nayeem, et al. "Biometric zoos: Theory and experimental evidence." Biometrics (IJCB), 2011 International Joint Conference on. IEEE, 2011.

It is related to subject representativeness

- Poor quality enrolment

Even with well-controlled enrolment, the menagerie still exists

Wittman et al, Empirical Studies of the Existence of the Biometric Menagerie in the FRGC 2.0 Color Image Corpus, CVPRW 2006

User-specific schemes can reduce the effect

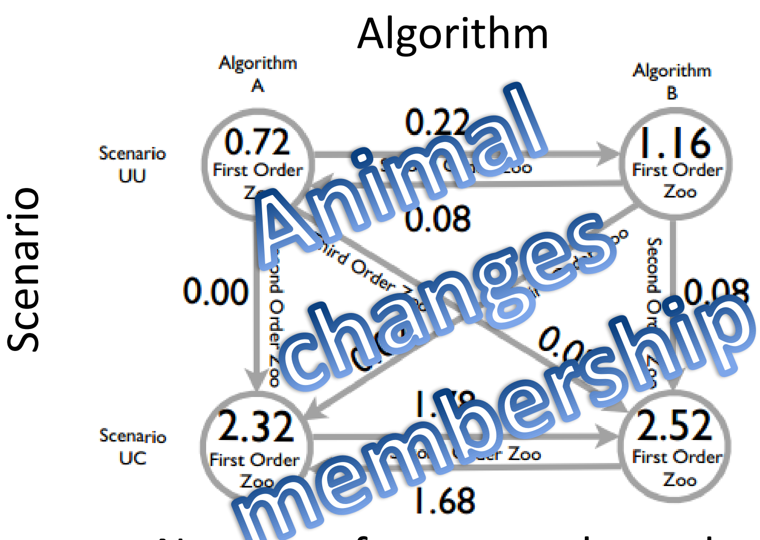
* Paone, Liberating the Biometric Menagerie Through Score Normalization, PhD thesis, 2013
* N. Poh and J. Kittler, Incorporating Variation of Model-specific Score Distribution in Speaker Verification Systems, IEEE Trans. Audio, Speech and Language Processing, 16(3):594-606, 2008.

Changing menagerie membership

Popescu-Bodorin, Nicolaie, Valentina Emilia Balas, and Iulia Maria Motoc. "The Biometric Menagerie–A Fuzzy and Inconsistent Concept." *Soft Computing Applications*. Springer Berlin Heidelberg, 2013. 27-43.

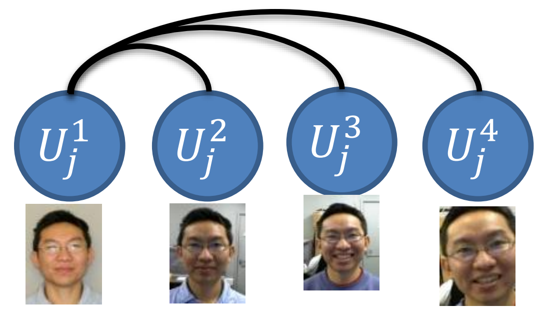
Should the menagerie classify the users or their templates?

Classify users	Classify templates



Number of goats overlapped out of 50 trials

Teli, Mohammad Nayeem, et al. "Biometric zoos: Theory and experimental evidence." Biometrics (IJCB), 2011 International Joint Conference on. IEEE, 2011.

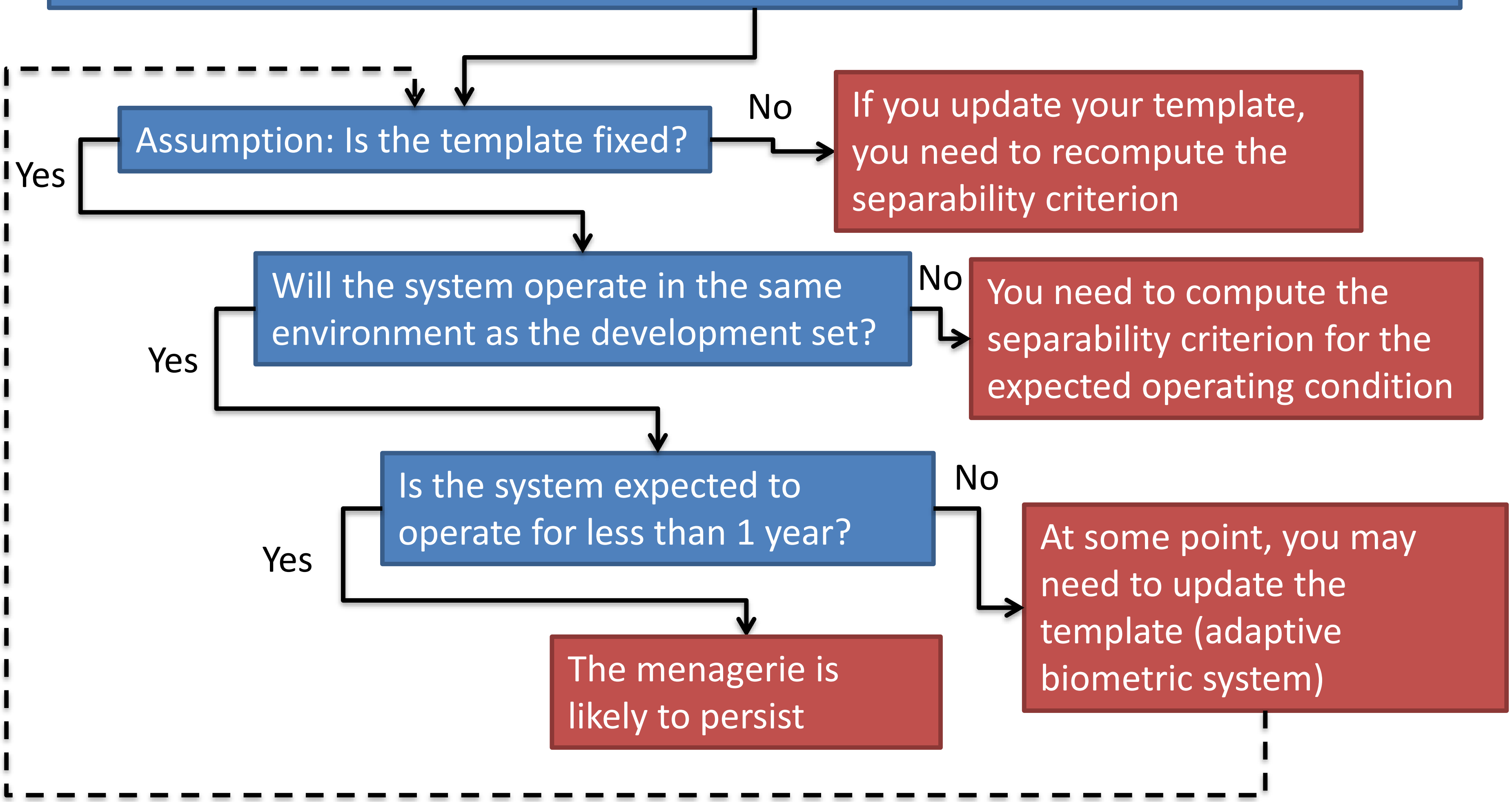


Users' score characteristics persist!

N. Poh, A. Ross, W. Li, and J. Kittler, A User-Specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects, Pattern Recognition 46(12): 3341-57, 2013.

Popescu-Bodorin, Nicolaie, Valentina Emilia Balas, and Iulia Maria Motoc. "The Biometric Menagerie–A Fuzzy and Inconsistent Concept." *Soft Computing Applications*. Springer Berlin Heidelberg, 2013. 27-43.

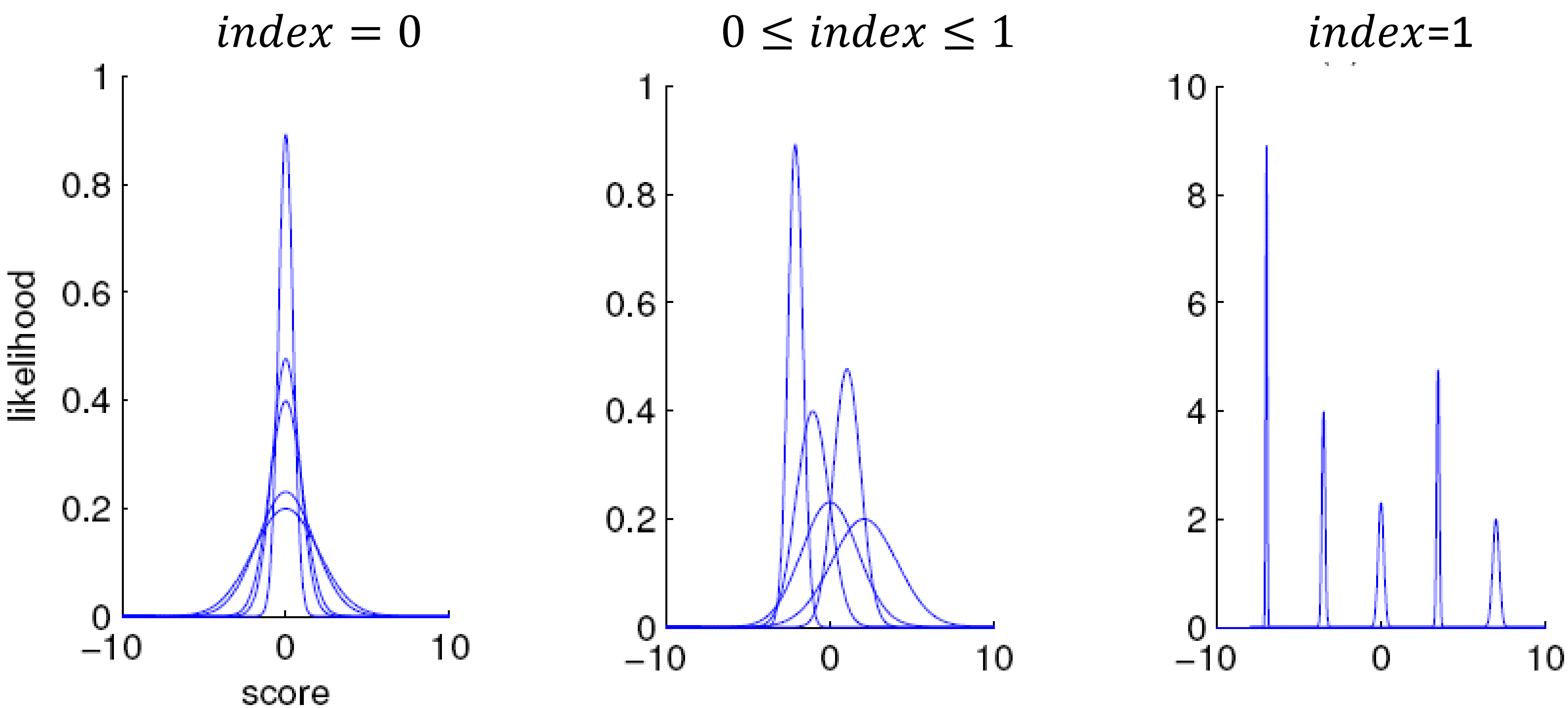
Is biometric menagerie persistent?



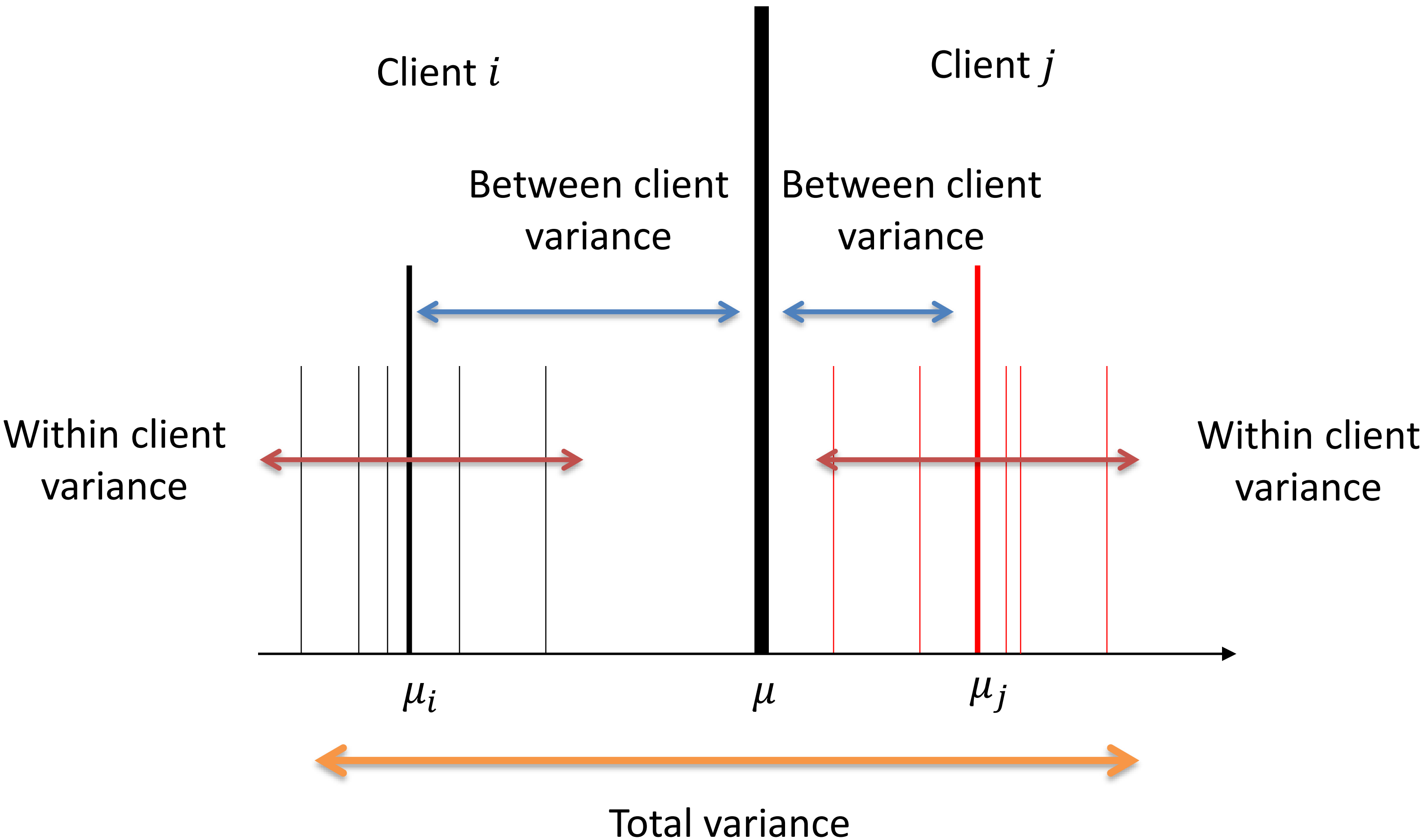
Menu

BIOMETRIC MENAGERIE INDEX

Desirable properties of a menagerie index



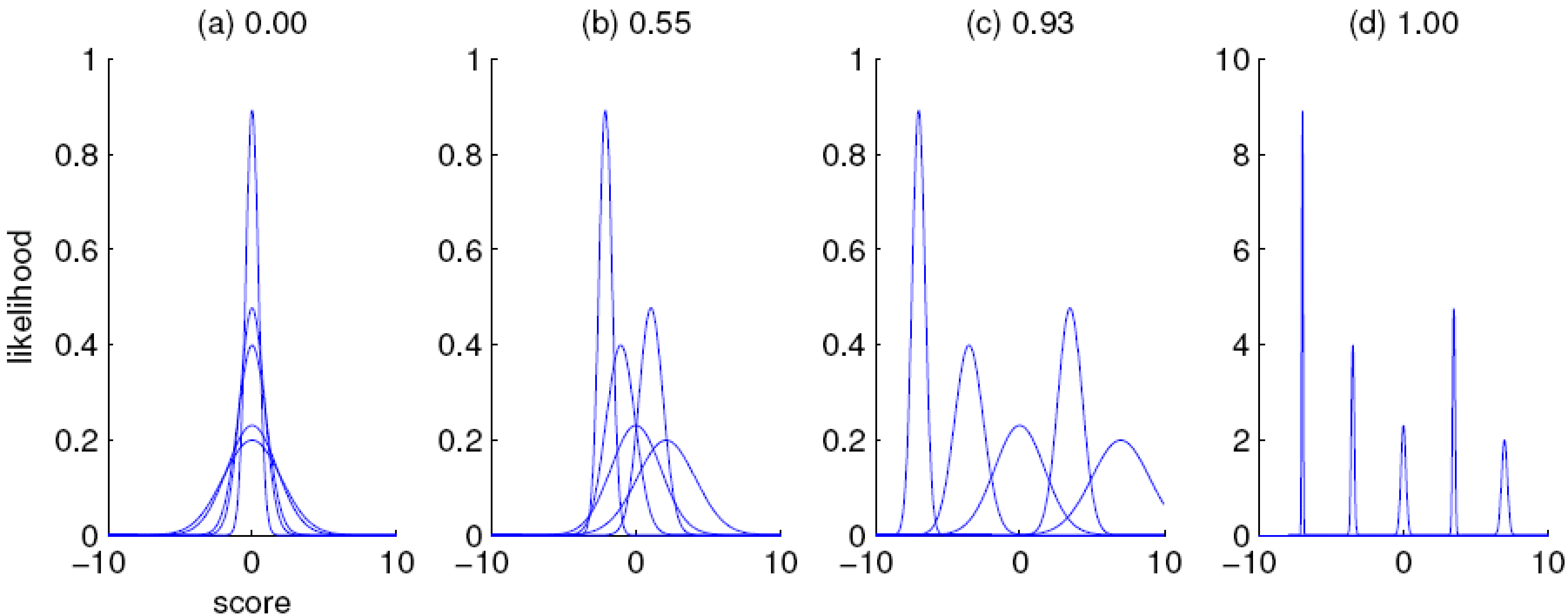
Biometric Menagerie Index



N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Modelspecific Variation, in Int'l Conf. on Biometrics (ICB'09), 2009.

$$\mathrm{BMI} = \frac{E\left[\qquad\qquad\right]}{\underbrace{\qquad\qquad\qquad}_{\text{Total variance}}} = \frac{\text{Mean of between client variance}}{\text{Total variance}}$$

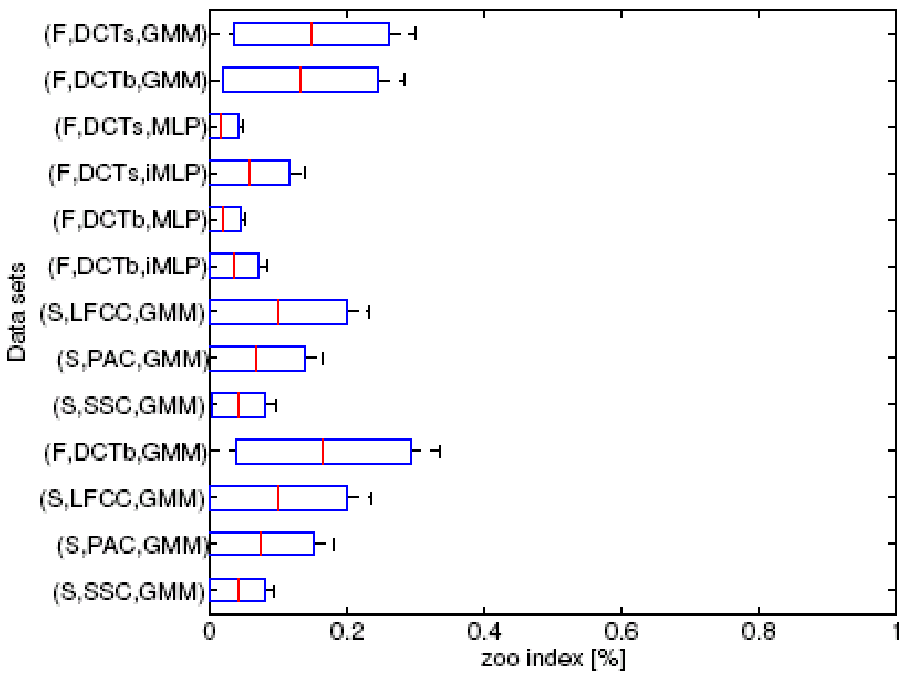
Global mean =
client mean



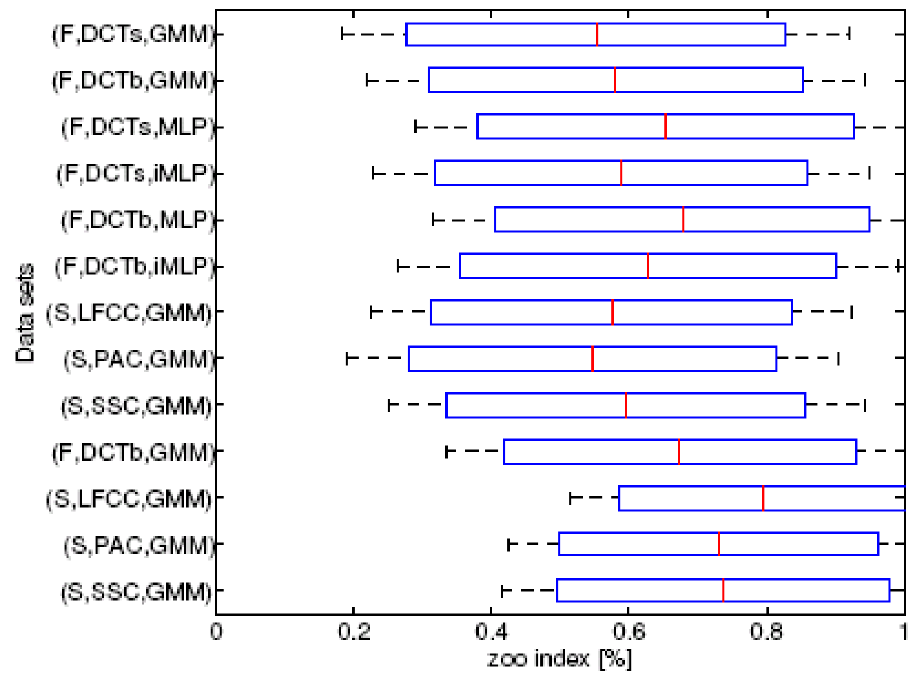
Within-client variance dominates

Between client variance dominates

XM2VTS database



Nonmatch comparisons



Match comparisons

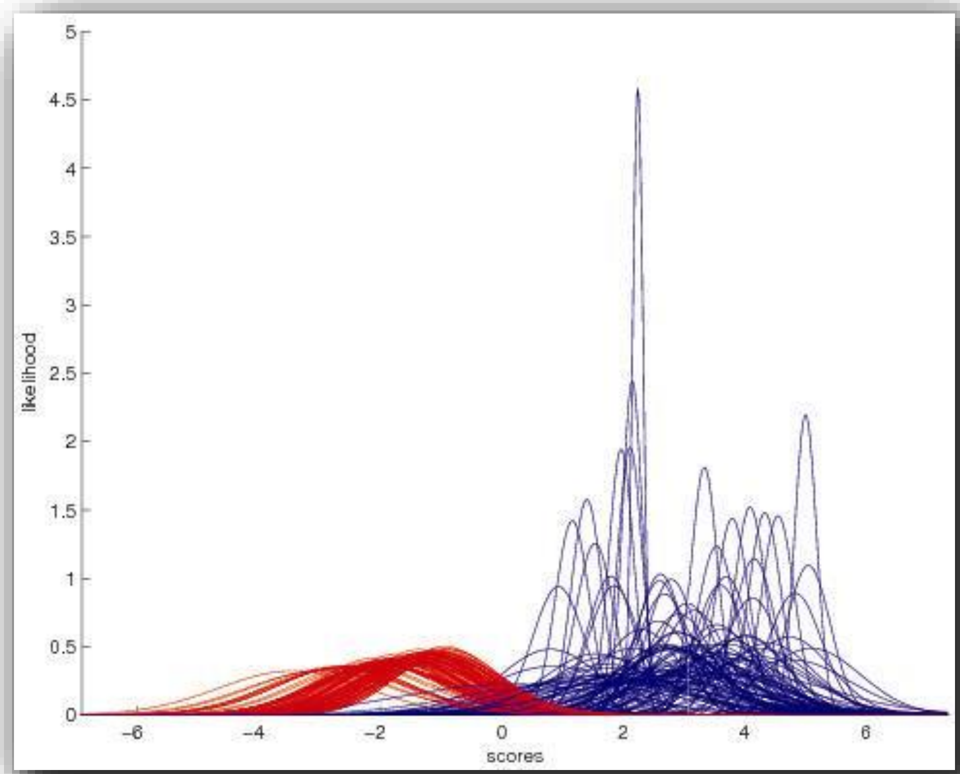
- N. Poh and J. Kittler, A Biometric Menagerie Index for Characterising Template/Modelspecific Variation, in Int'l Conf. on Biometrics (ICB'09), 2009.
- Get the data set: http://goo.gl/CdXw9Z

Findings

Impostor BMI is insensitive to the choice of (zero-effort) impostor

client BMI > Impostor BMI

High client BMI values suggest that client-specific threshold or normalization is essential



Menu

USER RANKING

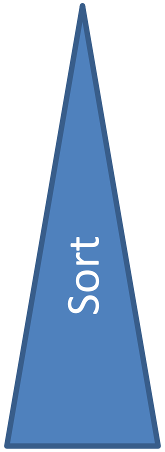
User ranking

	Genuine	ZE Impostor
1	Y_1^G	Y_1^I
⋮	⋮	
100	Y_{100}^G	Y_{100}^I

Calculate the statistic

μ_1^{G}	$\mu_1^{\mathrm{I}}, \sigma_1^{\mathrm{I}}$
⋮	⋮
μ_{100}^{G}	$\mu_{100}^{\mathrm{I}}, \sigma_{100}^{\mathrm{I}}$

Rank the users



Why rank the subjects?

N. Poh and J. Kittler, A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion, in 6th Biometrics Symposium, pp. 17–22, 2008.

Quality control during enrollment

- If the newly acquired template is not representative of the person, acquire a *better* (more representative) sample

A tool to assess the worst-scenario DET curve

- How bad could a system perform for a group of weak (the weakest) users?
- The experience of *each* user in interacting with a biometric device matters!

A modality selection criterion in fusion

- Use only the more representative biometric modality instead.

Novel group specific decision

- An optimal decision threshold for each group of users

N. Poh, A. Ross, W. Li, and J. Kittler, A User-Specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects, Pattern Recognition 46(12): 3341-57, 2013.

N. Poh, A. Rattani, M. Tistarelli and J. Kittler, Group-specific Score Normalization for Biometric Systems, in IEEE Computer Society Workshop on Biometrics (CVPR), pages 38-45, 2010.

Criteria to rank the users

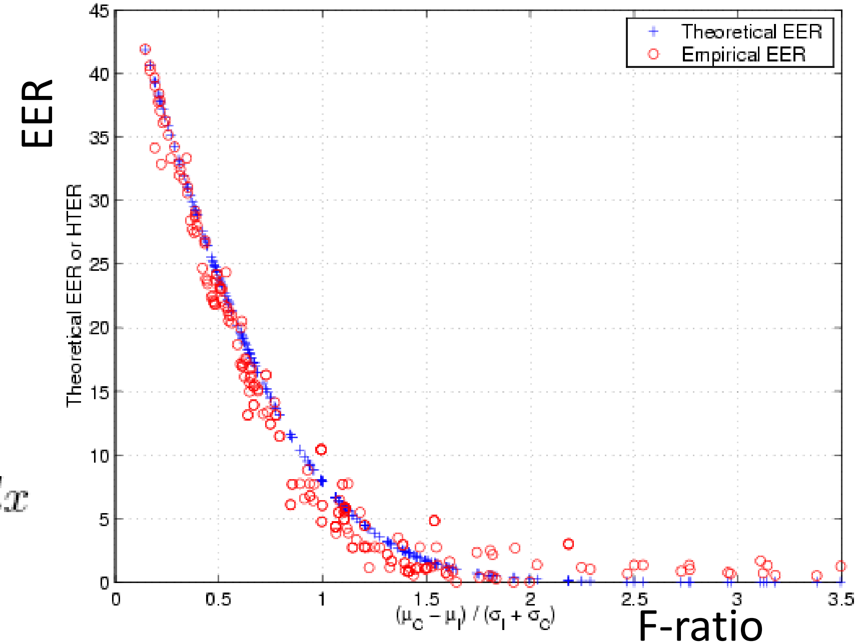
- Calculate the performance of each user template and rank them somehow!

$$\mathrm{EER}_j = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\text{F-ratio}}{\sqrt{2}}\right)$$

$$\text{F-ratio} = \frac{\mu^{\mathrm{C}} - \mu^{\mathrm{I}}}{\sigma^{\mathrm{C}} + \sigma^{\mathrm{I}}}$$

$$\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp\left[-x^2\right] dx$$

For each user/template!



[IEEE Trans. SP, 2006]

Many parametric discrimination criteria available

$$\text{F-ratio} = \frac{\mu^{\mathrm{C}} - \mu^{\mathrm{I}}}{\sigma^{\mathrm{C}} + \sigma^{\mathrm{I}}}$$

$$d' = \frac{|\mu^{c} - \mu^{I}|}{\sqrt{\frac{1}{2}(\sigma^{c})^2 + \frac{1}{2}(\sigma^{I})^2}}$$

$$J_1 = \frac{\mu^{\mathrm{C}}}{\mu^{\mathrm{I}}}, J_2 = \frac{(\mu^{\mathrm{C}} - \mu^{\mathrm{I}})^2}{\mu^{\mathrm{C}}\mu^{\mathrm{I}}}$$

$$J_3 = \frac{(\mu^{c} - \mu^{I})^2}{(\sigma^{c})^2 + (\sigma^{I})^2}$$

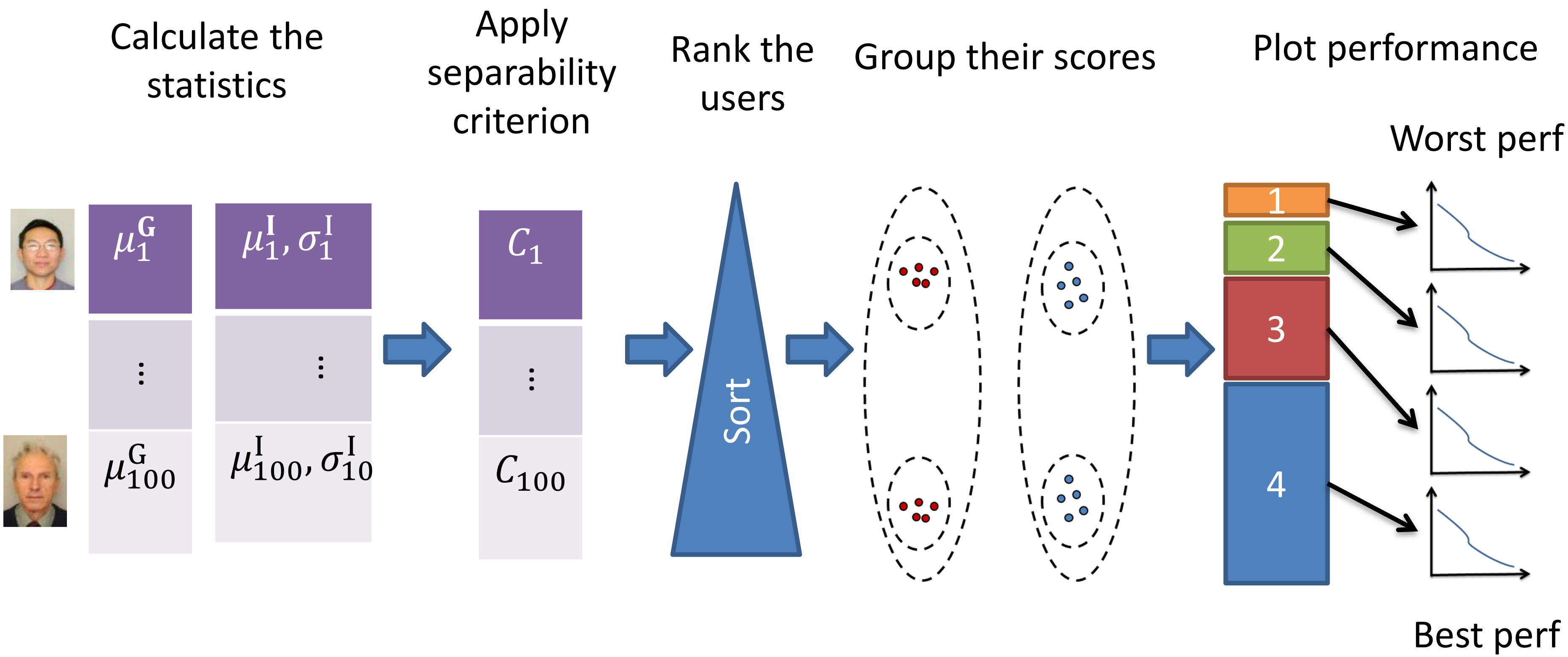
Multivariate statistics

Bhattacharyya distance

Chernoff Bound

$$\text{Fisher-ratio} = \frac{(\mu^{c} - \mu^{I})^2}{(\sigma^{c})^2 + (\sigma^{I})^2}$$

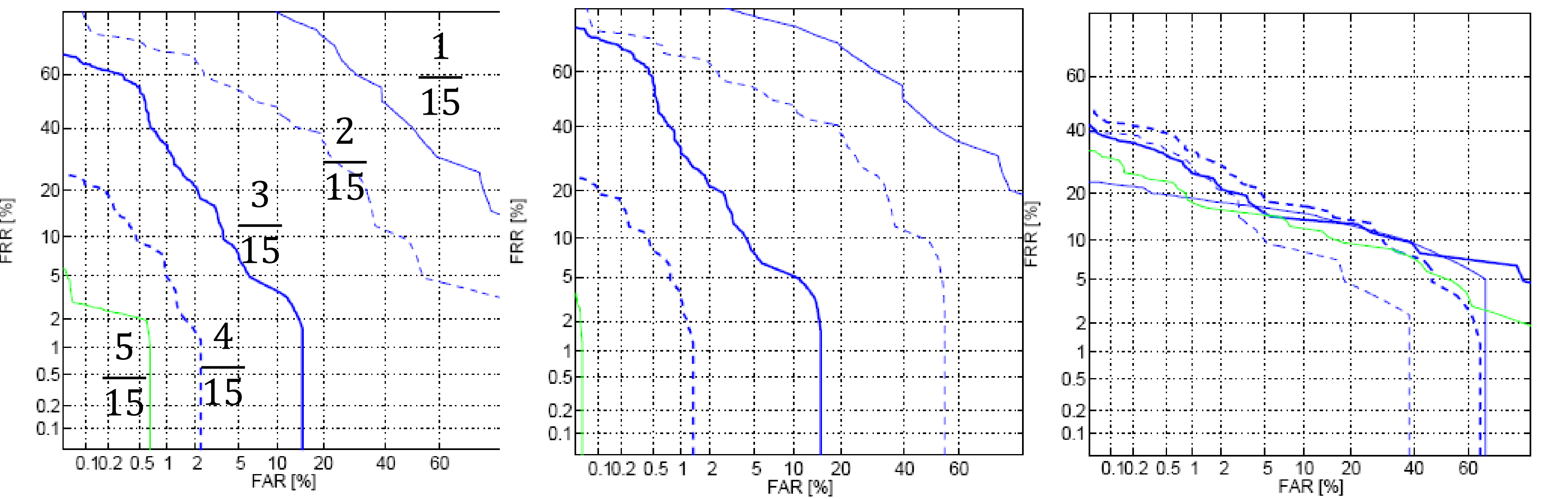
User-ranking algorithm



N. Poh and J. Kittler, A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion, in 6th Biometrics Symposium, pp. 17–22, 2008.

Results on Face, Fingerprints & Iris

Biosecure database – a face example



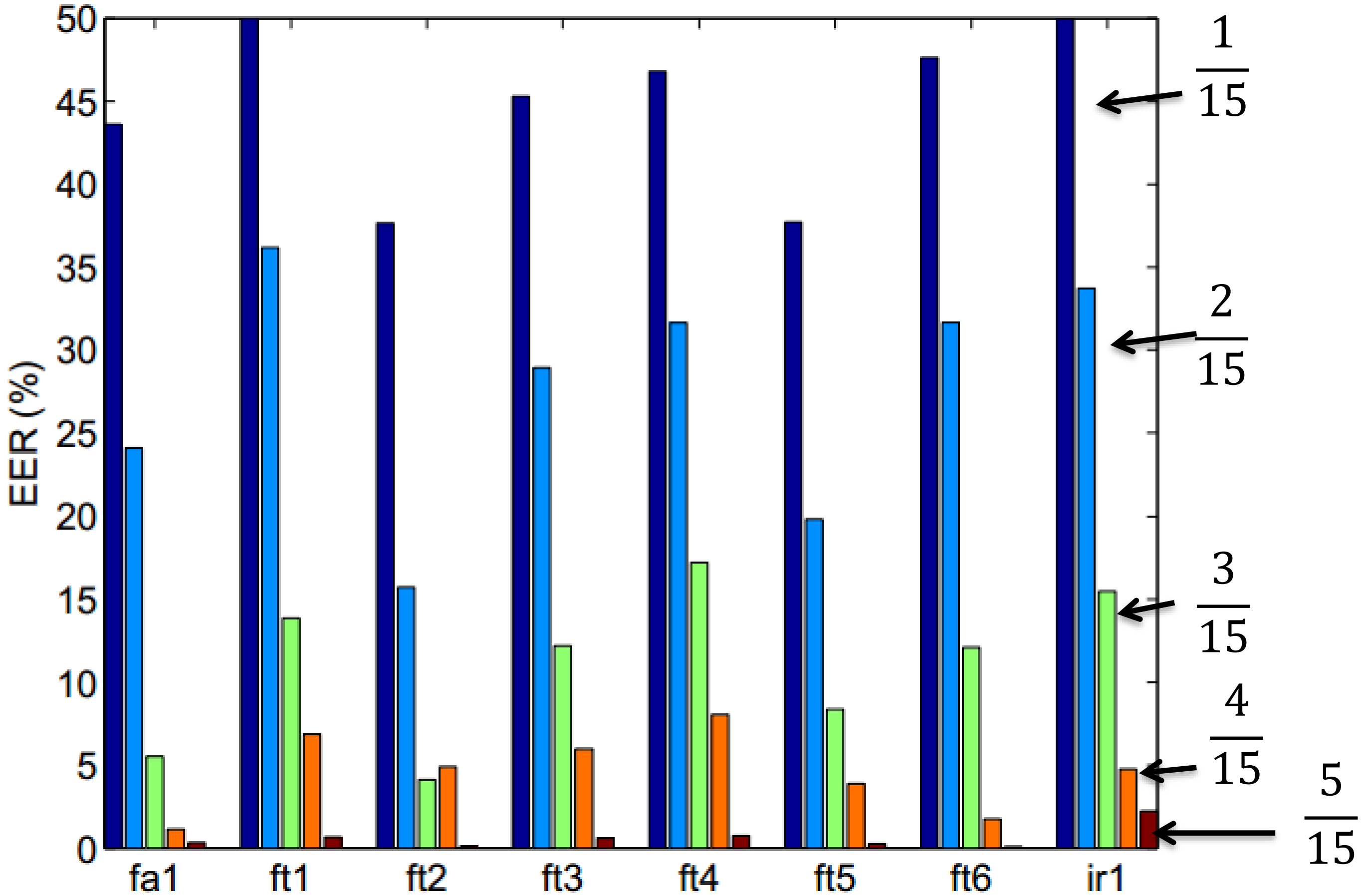
F-ratio d-prime random

$$\text{F-ratio} = \frac{\mu^{\mathrm{C}} - \mu^{\mathrm{I}}}{\sigma^{\mathrm{C}} + \sigma^{\mathrm{I}}}$$

$$d' = \frac{|\mu^{\mathrm{C}} - \mu^{\mathrm{I}}|}{\sqrt{\frac{1}{2}(\sigma^{\mathrm{C}})^2 + \frac{1}{2}(\sigma^{\mathrm{I}})^2}}$$

N. Poh and J. Kittler, A Methodology for Separating Sheep from Goats for Controlled Enrollment and Multimodal Fusion, in 6th Biometrics Symposium, pp. 17–22, 2008.

Performance disparity across users



Label	template ID {n}	Modality	Sensor
fa	1	Still Face	web cam
ft	1–6	Fingerprints	Thermal
ir	1	Left iris image	LG

Findings

Very few weak users; but they dominate the errors

F-ratio and d-prime can be used to rank the users

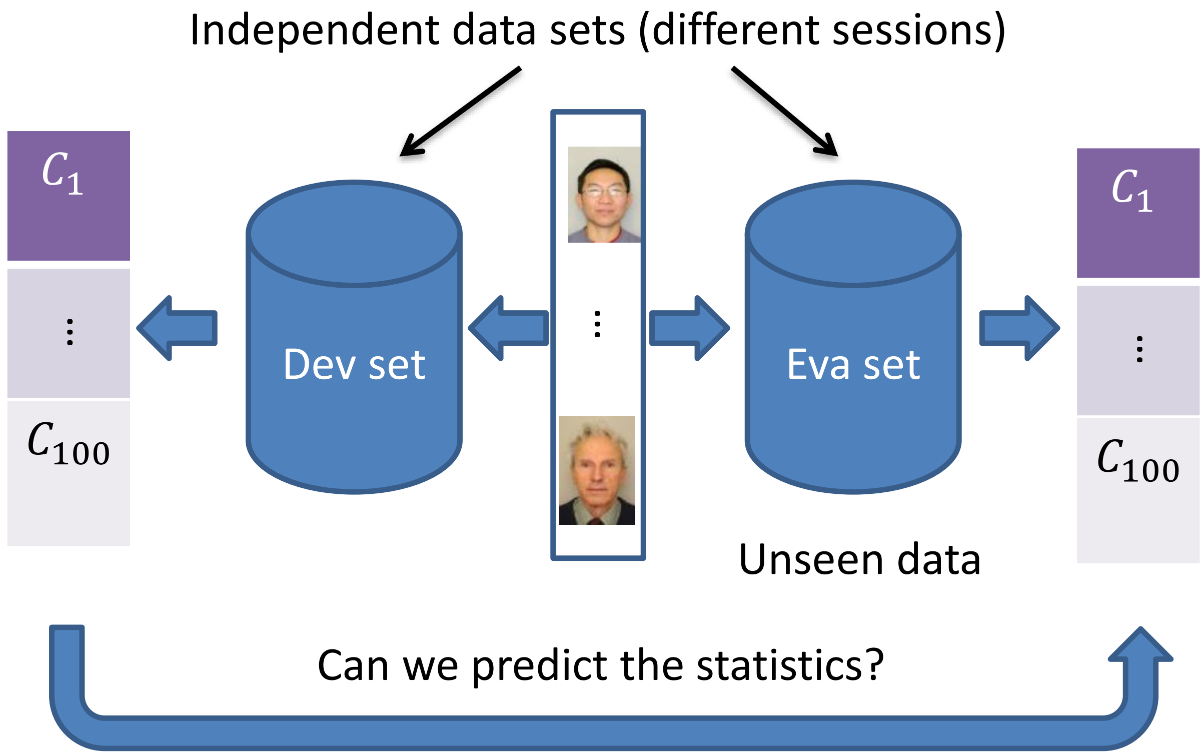
Can we rank the subject with *maximal stability (on unseen data)*?

- “B-ratio” (PRJ 2013) ranks users on unseen data

Menu

USER RANKING ON UNSEEN DATA

Rank the users on unseen data!



We designed 6 criteria and searched the best one

N. Poh, A. Ross, W. Li, and J. Kittler, A User-Specific and Selective Multimodal Biometric Fusion Strategy by Ranking Subjects, Pattern Recognition 46(12): 3341-57, 2013.

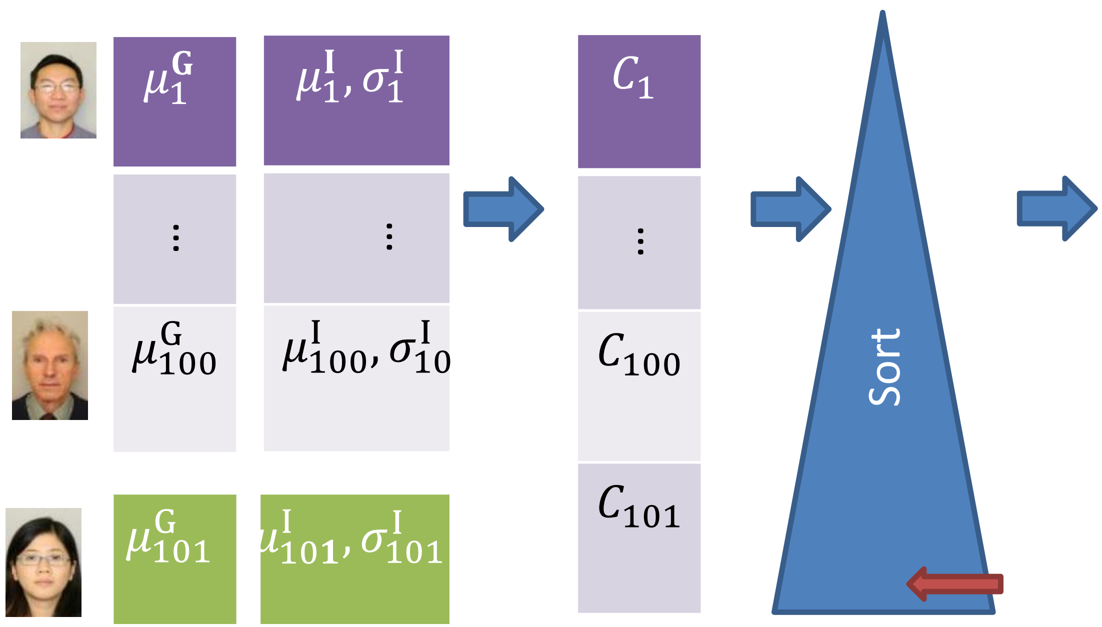
Approach

Calculate the statistics

Apply separability criterion

Rank the users

We should be able to predict problematic users before they use the technology

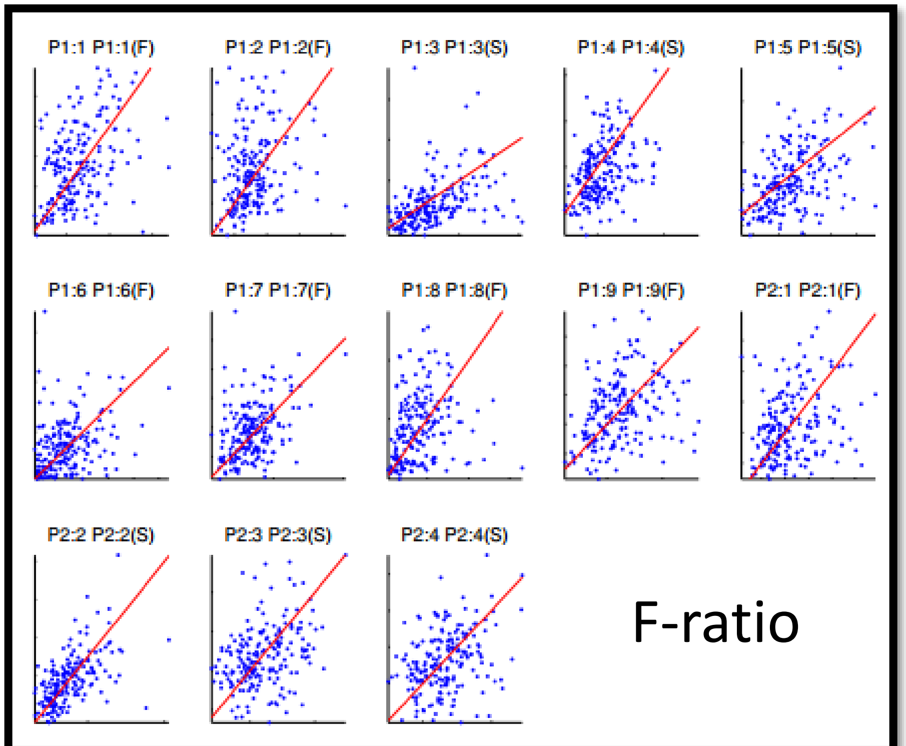


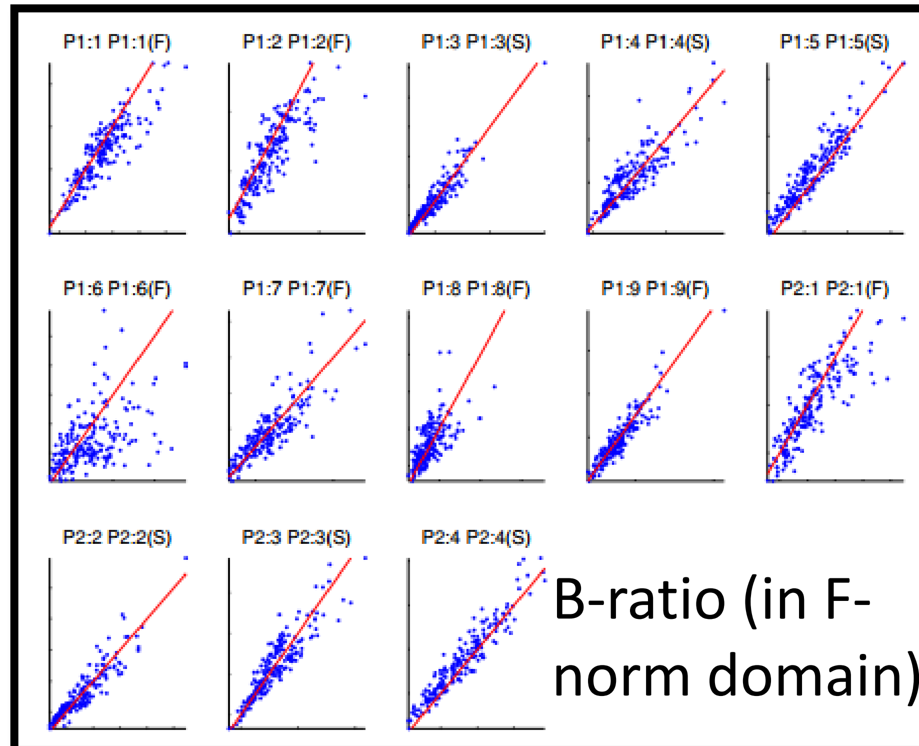
Re-enrol with new samples

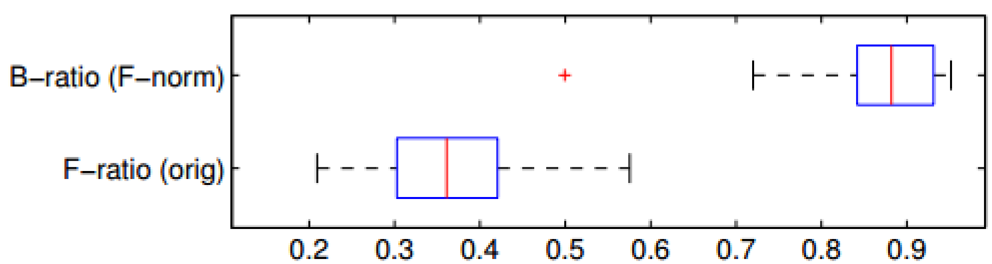
Use a different biometric modality

Recommend another authentication mechanism

Predictability of user ranking on unseen data







XM2VTS 13 systems

Menu

BIOMETRIC MENAGERIE “PROCEDURE”

Biometric Menagerie “Procedures”

Definition 1 – Performance finetuning/calibration: Interventions that allow us to exploit Biometric Menagerie to improve the system’s future operation

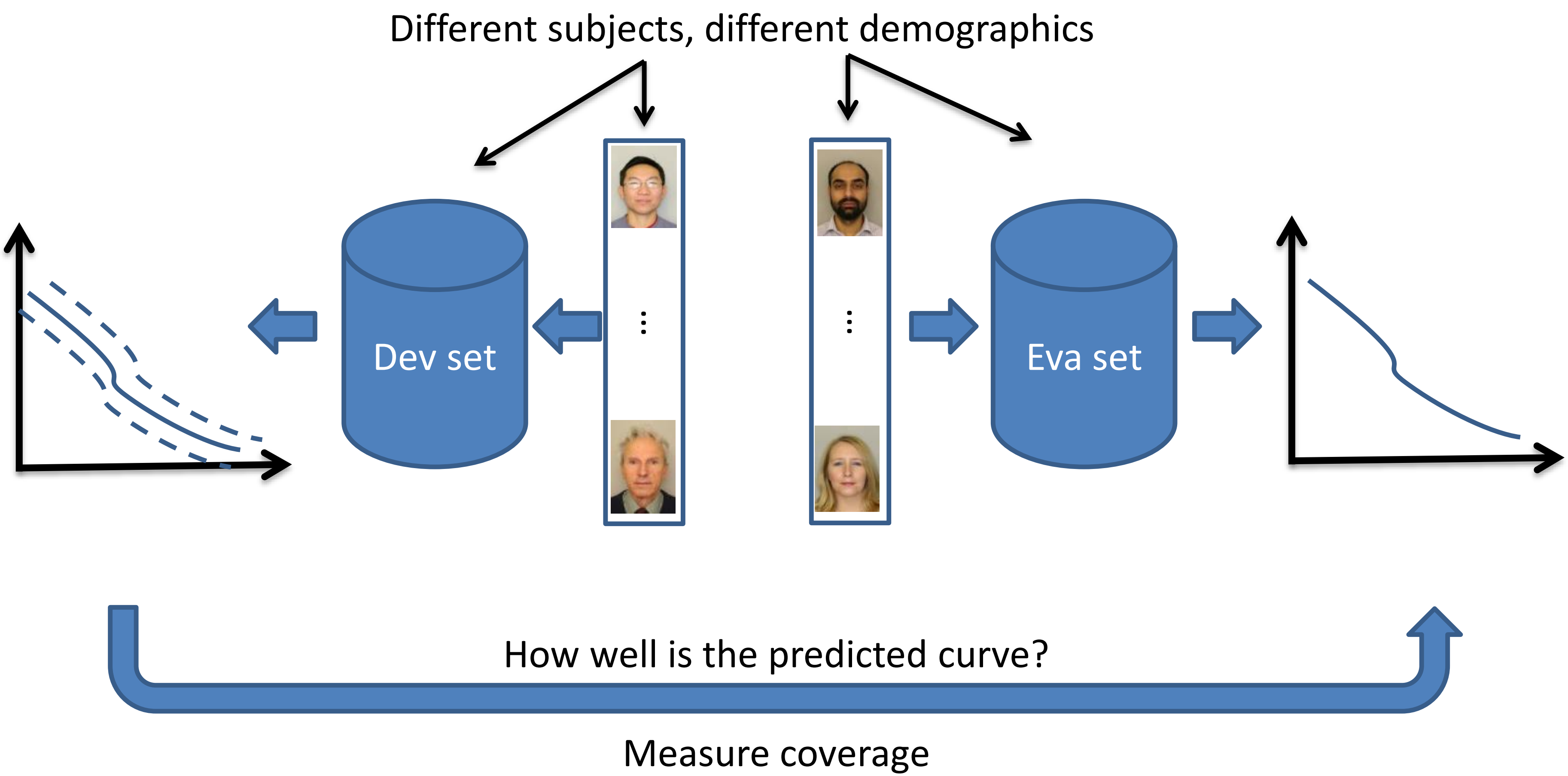
User ranking	user tailoring

Definition 2 – Performance estimation / generalization / prediction: Algorithms that exploit Biometric Menagerie to estimate the system’s future operational performance

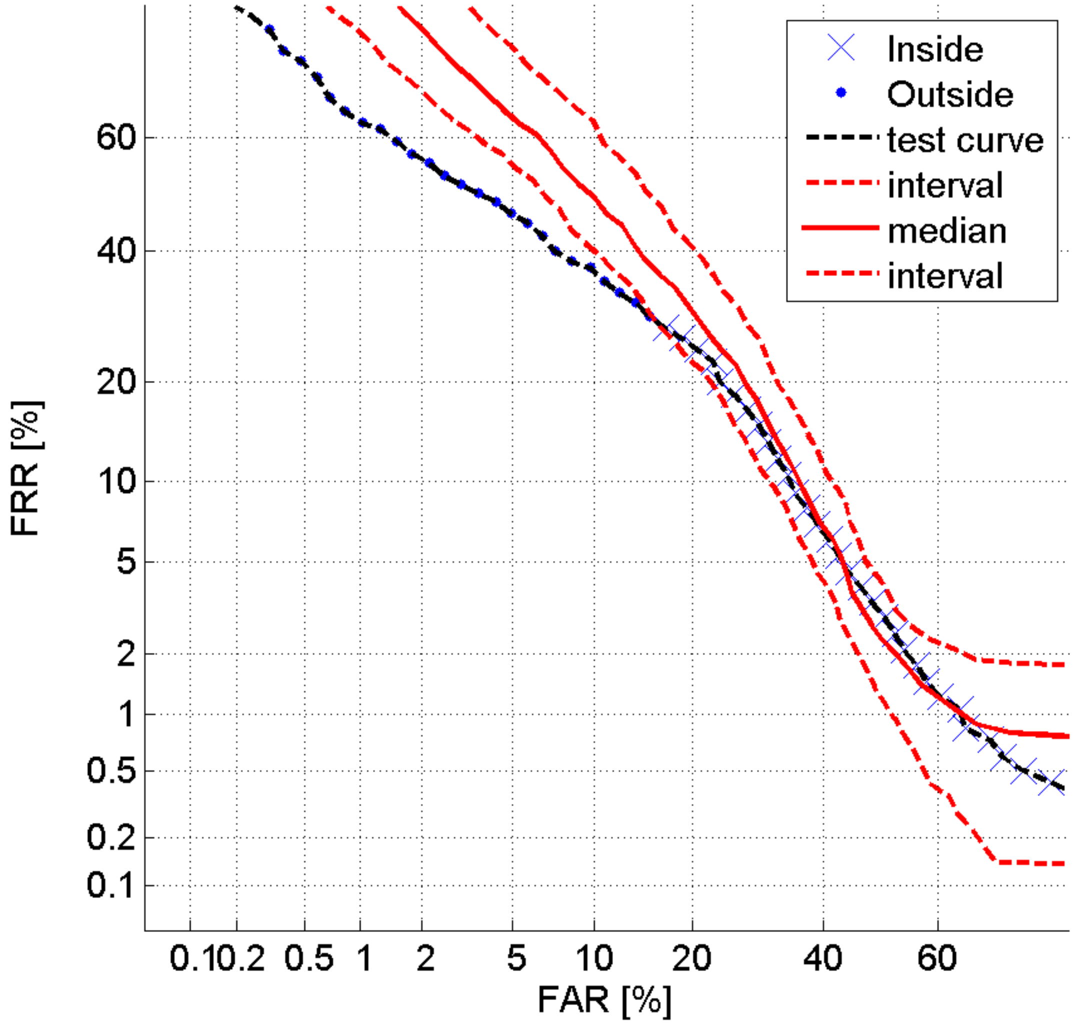
Performance intervals estimation	Performance synthesis

PERFORMANCE CONFIDENCE ESTIMATION

DET confidence intervals

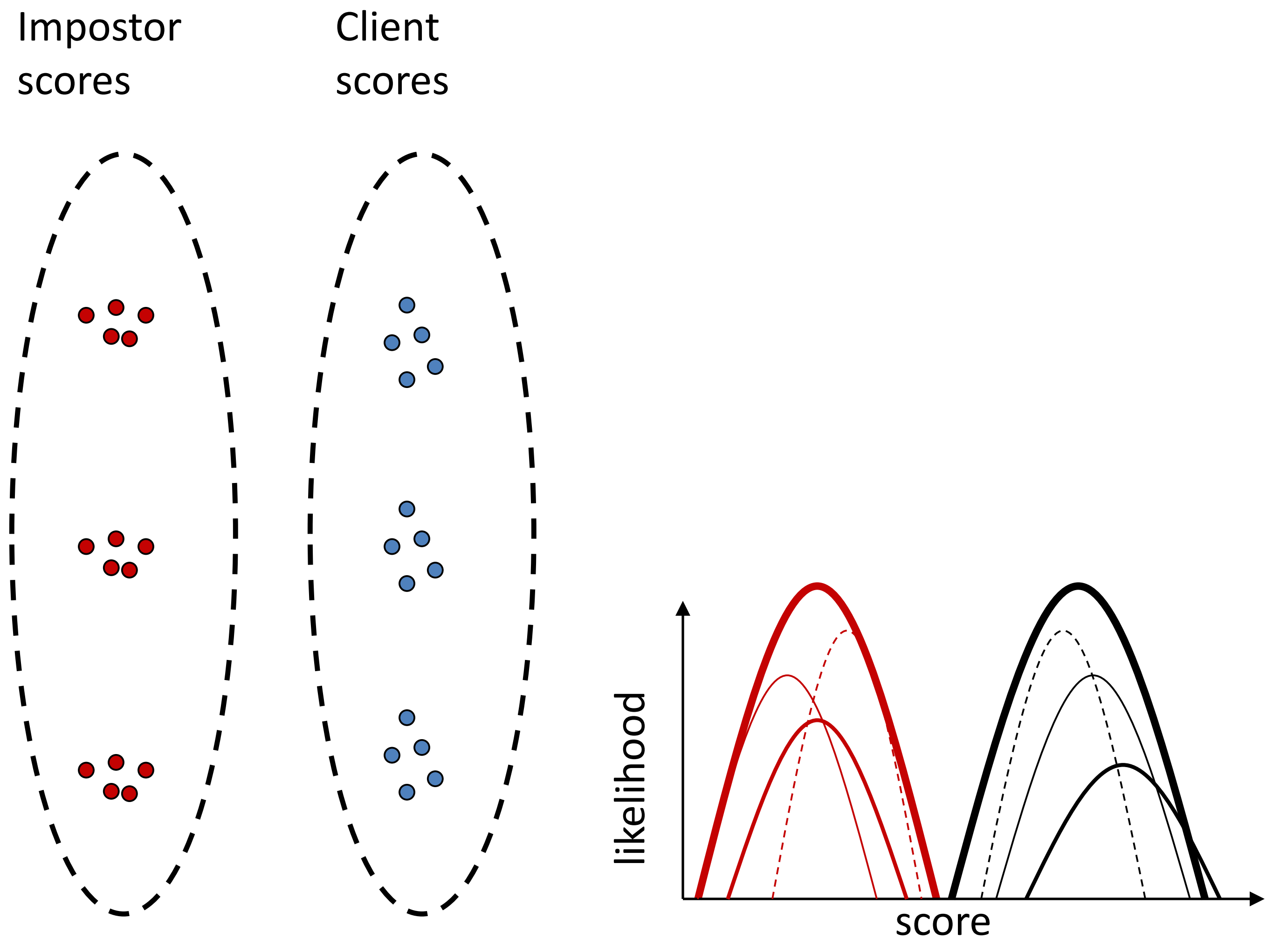


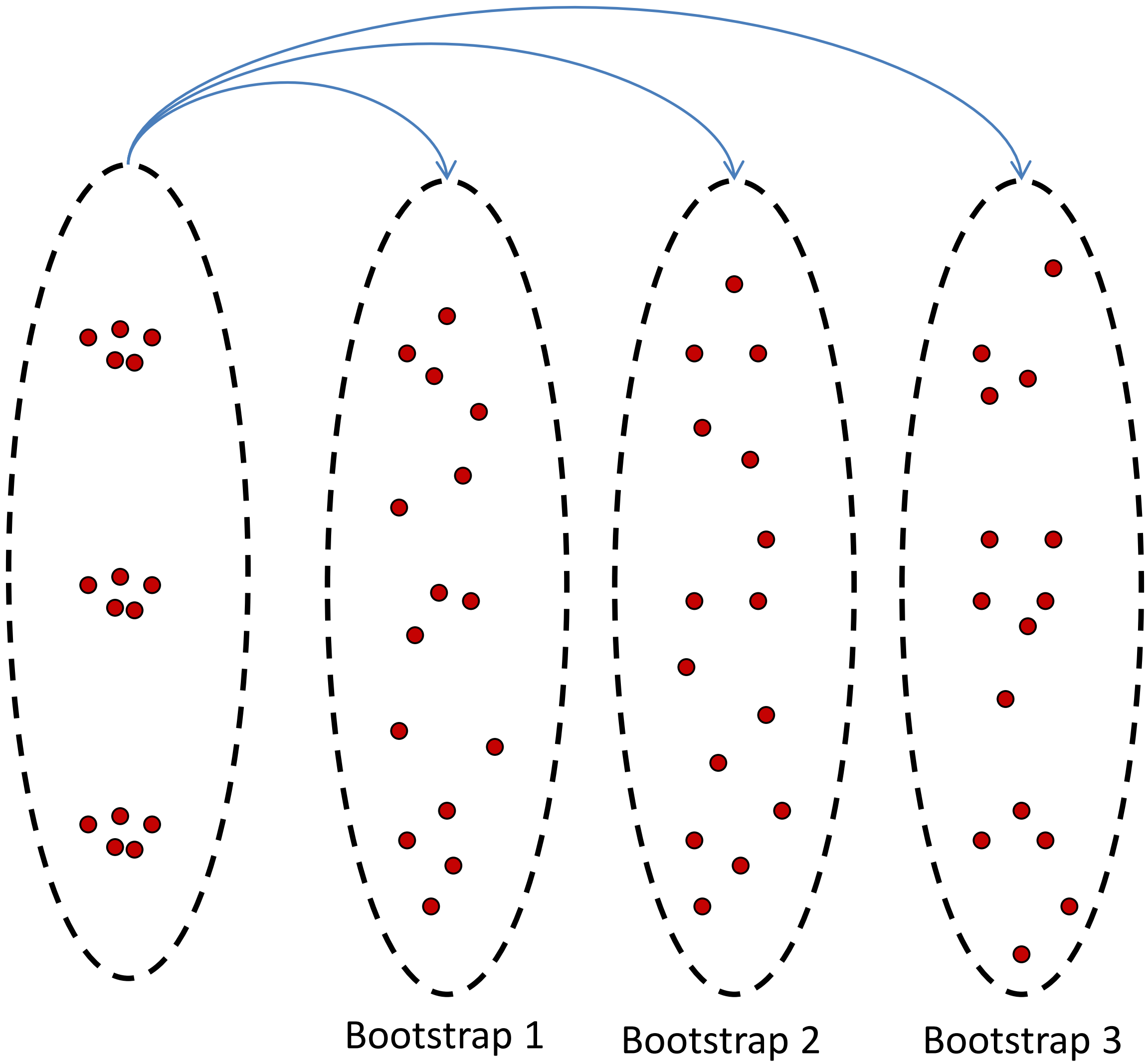
Coverage



Inside
Outside
test curve
interval
median
interval
FRR [%]
60
40
20
10
5
2
1
0.5
0.2
0.1
0.1
0.2
0.5
1
2
5
10
20
40
60
FAR [%]

Conventional bootstrap

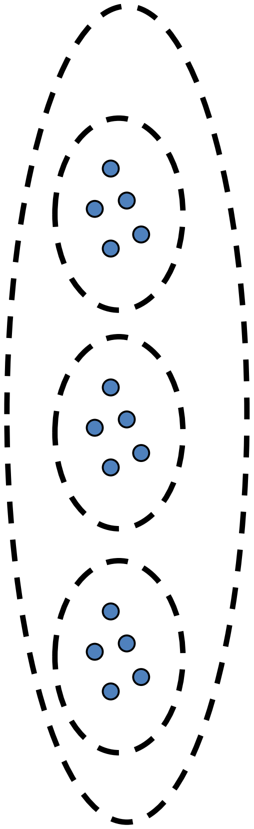



Bootstrap 1
Bootstrap 2
Bootstrap 3

A two-level bootstrap

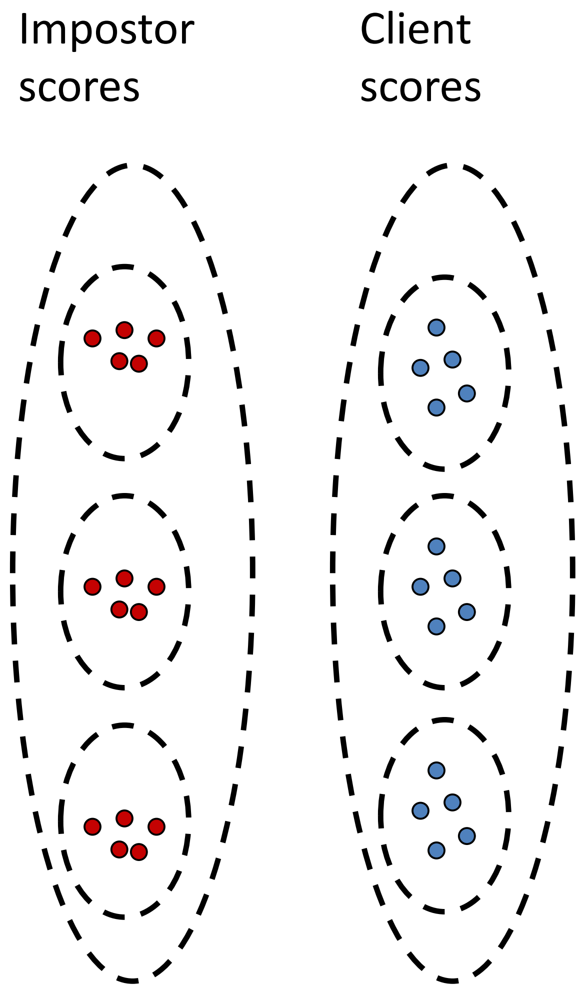
Handle **between model variance** by boostrapping the enrolled identities

Handle **within model variance** by bootstrapping the within-model scores

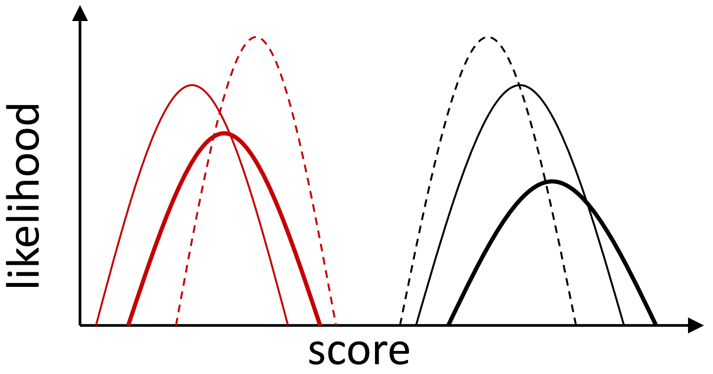


N. Poh and S. Bengio, Performance Generalization in Biometric Authentication Using Joint User-specific and Sample Bootstraps , IEEE Trans. on Pattern Analysis and Machine Intelligence, 29(3):492-498, March 2007.

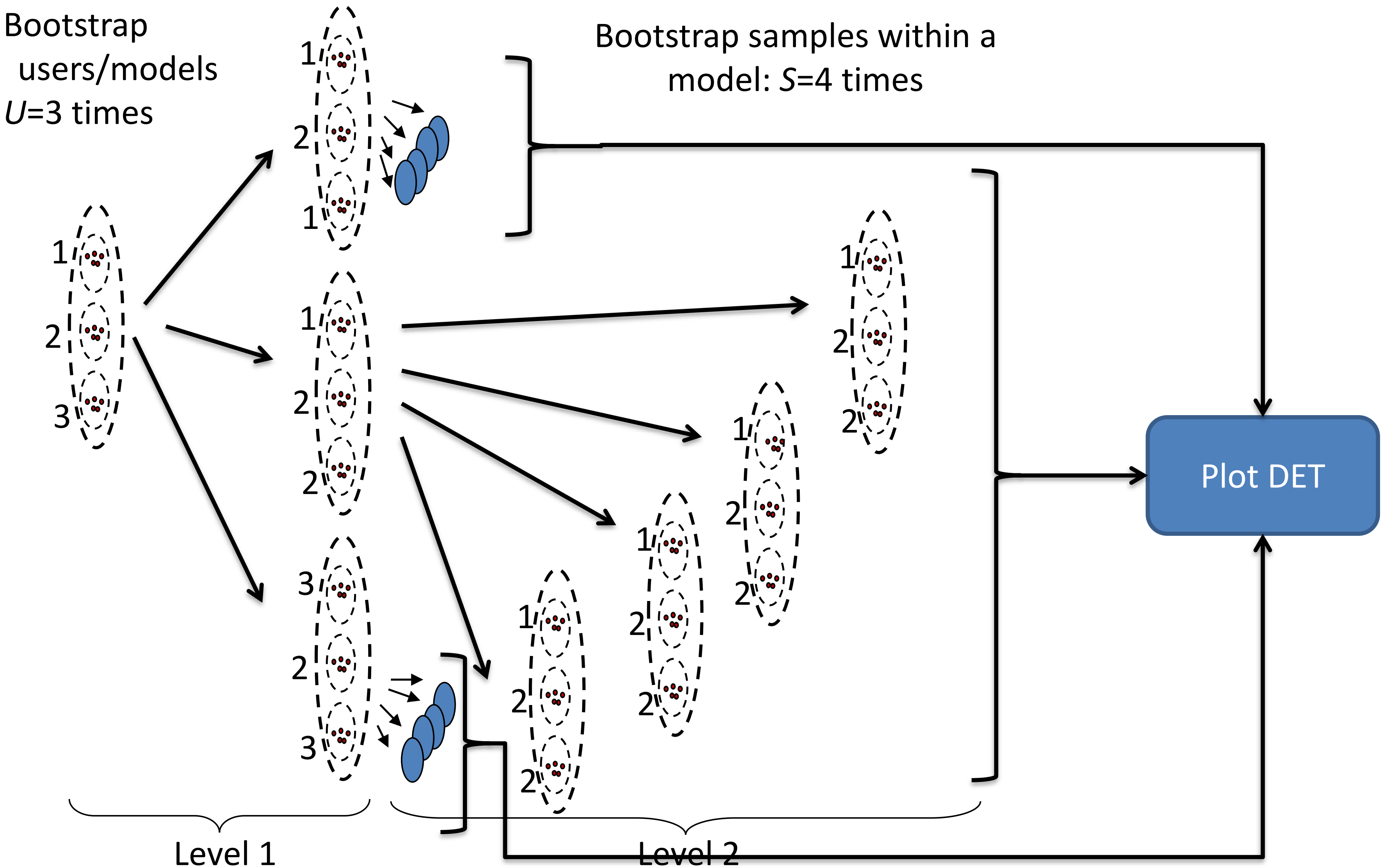
Subset Bootstrap



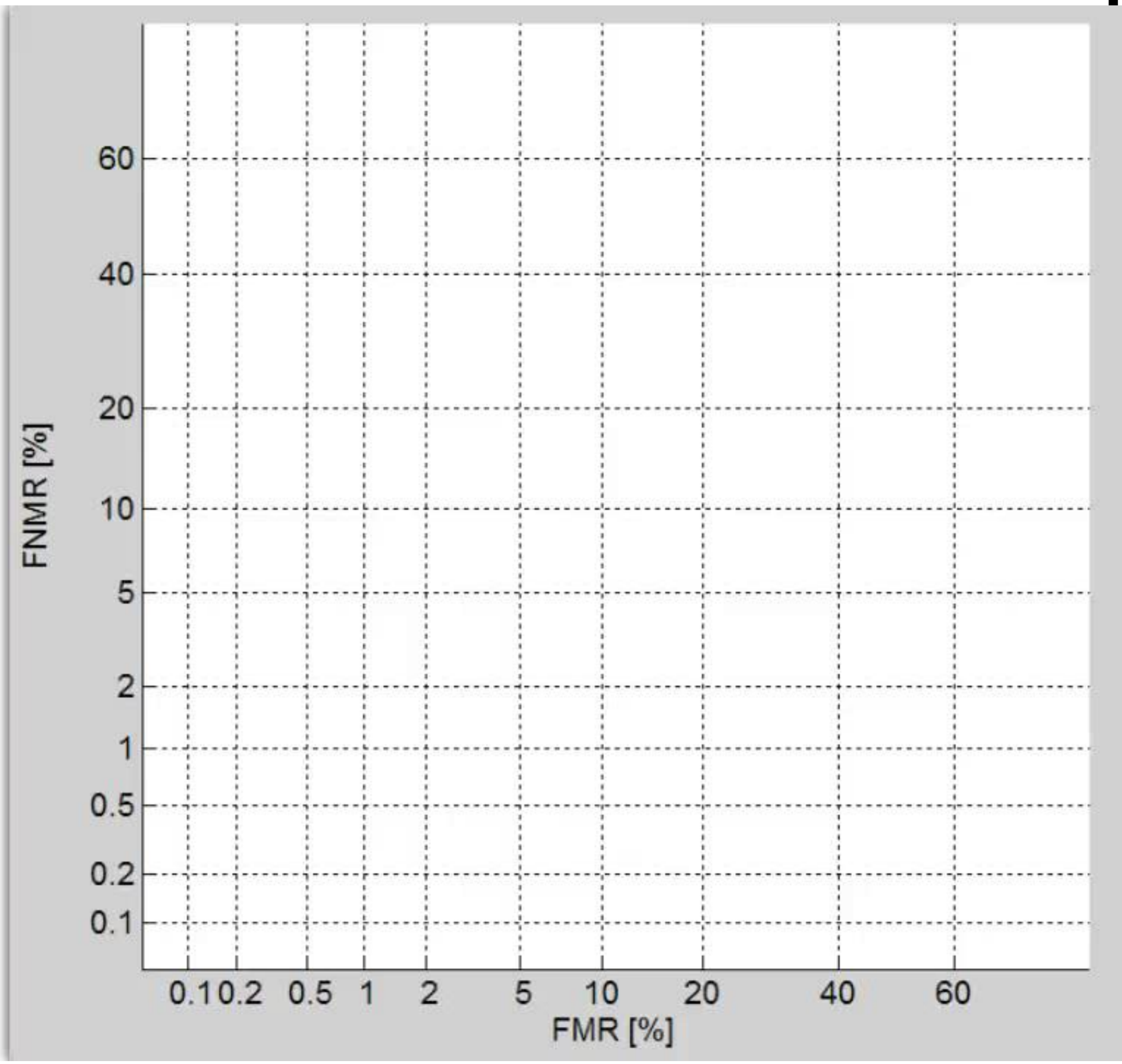
Bolle, Ratha & Pakanti, Error analysis of pattern recognition systems: the subsets bootstrap, CVIU 2014

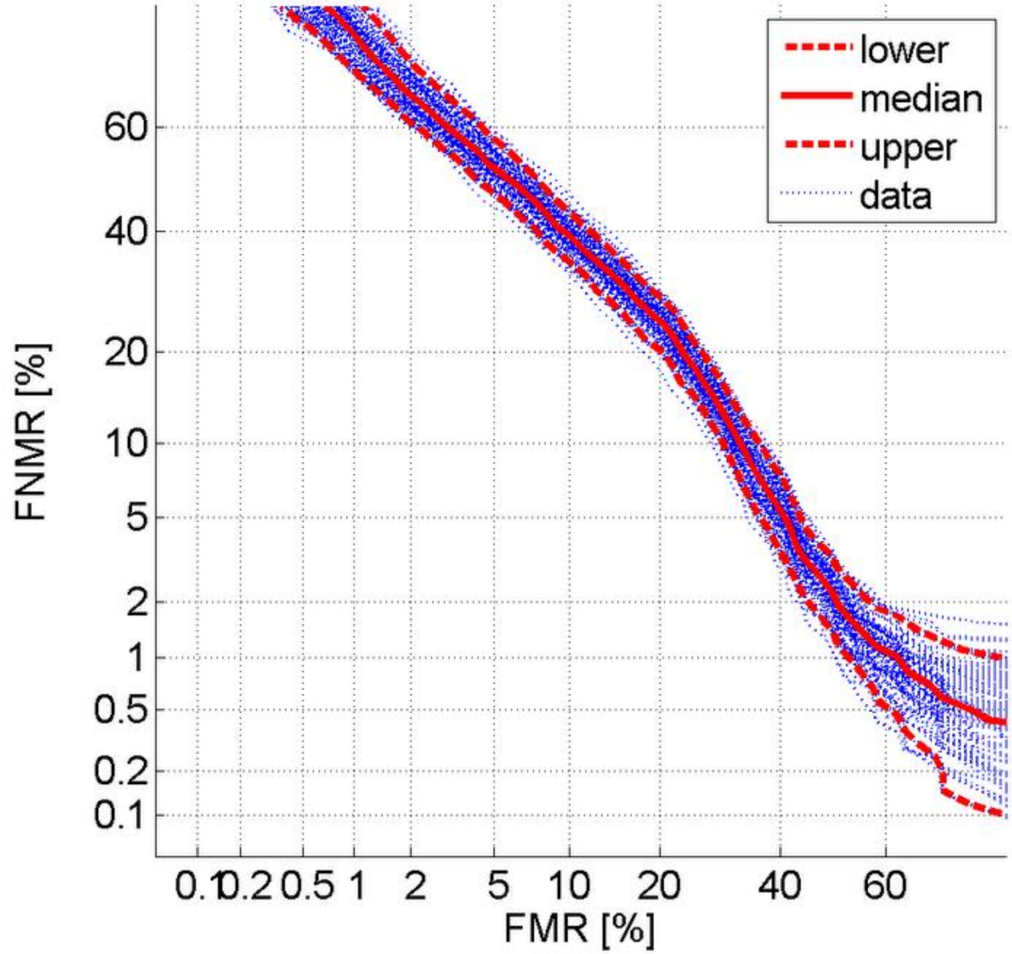


Two-level Bootstrap



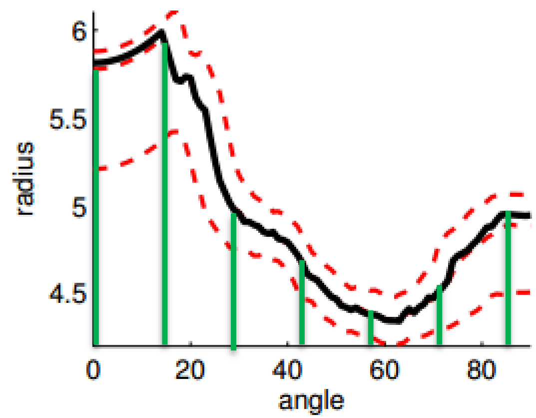
DET confidence via bootstrapping

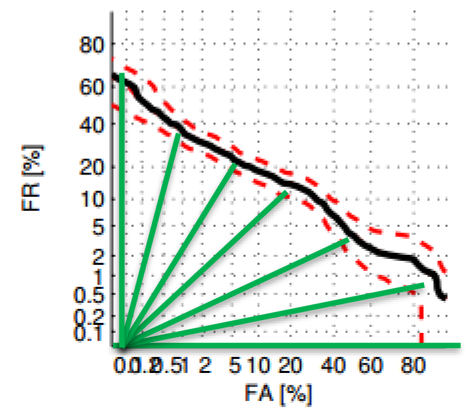



lower
median
upper
data
60
40
20
10
5
2
1
0.5
0.2
0.1
FNMR [%]
0.1
0.2
0.5
1
2
5
10
20
40
60
FMR [%]

DET angle

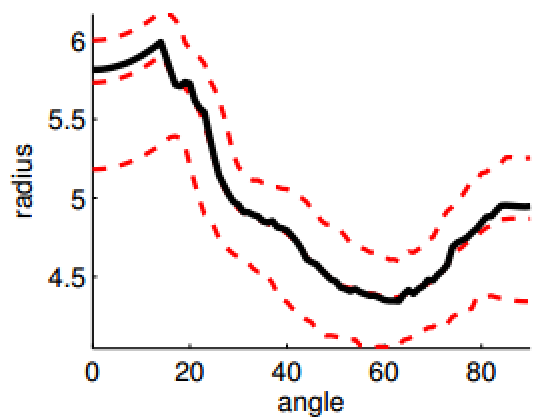
Sample bootstrap

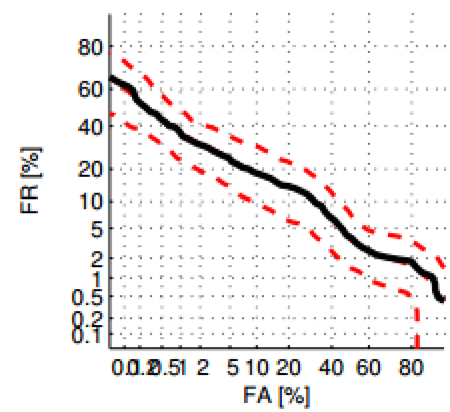




Fix the users (claimants), bootstrap only the samples

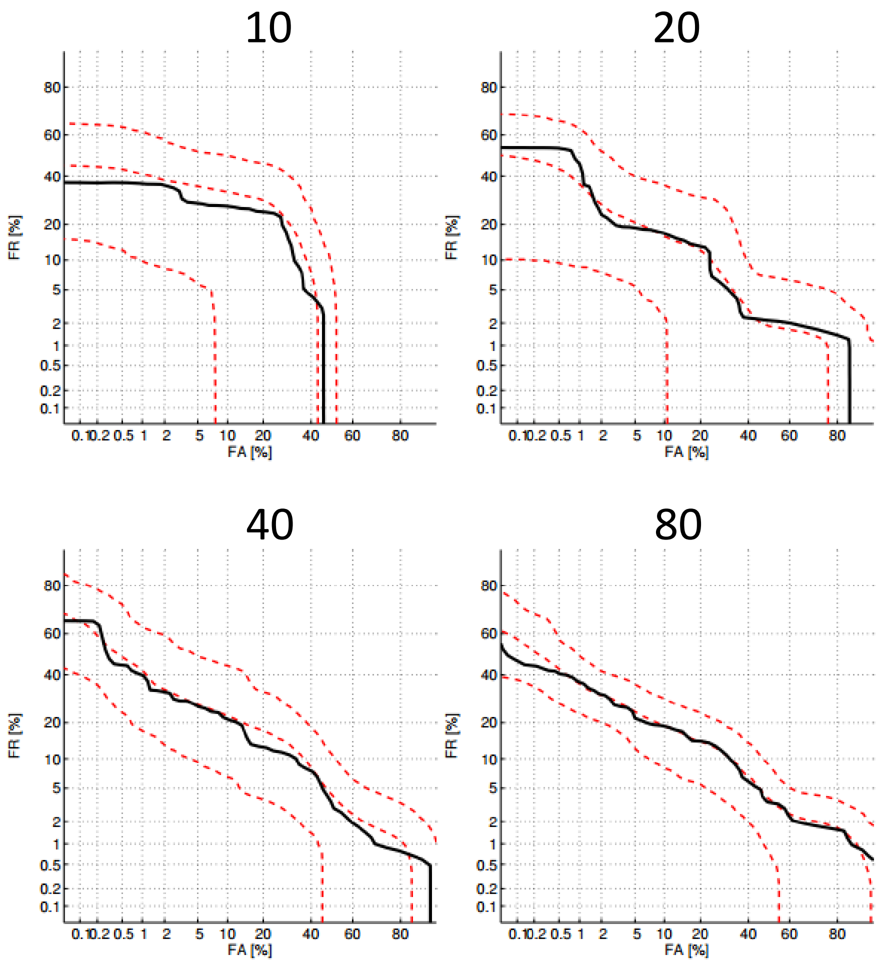
User-specific bootstrap





Bootstrap the claimants, and keep all the scores due to the chosen claimants

Effect of subject samples



Increased subject

Better precision

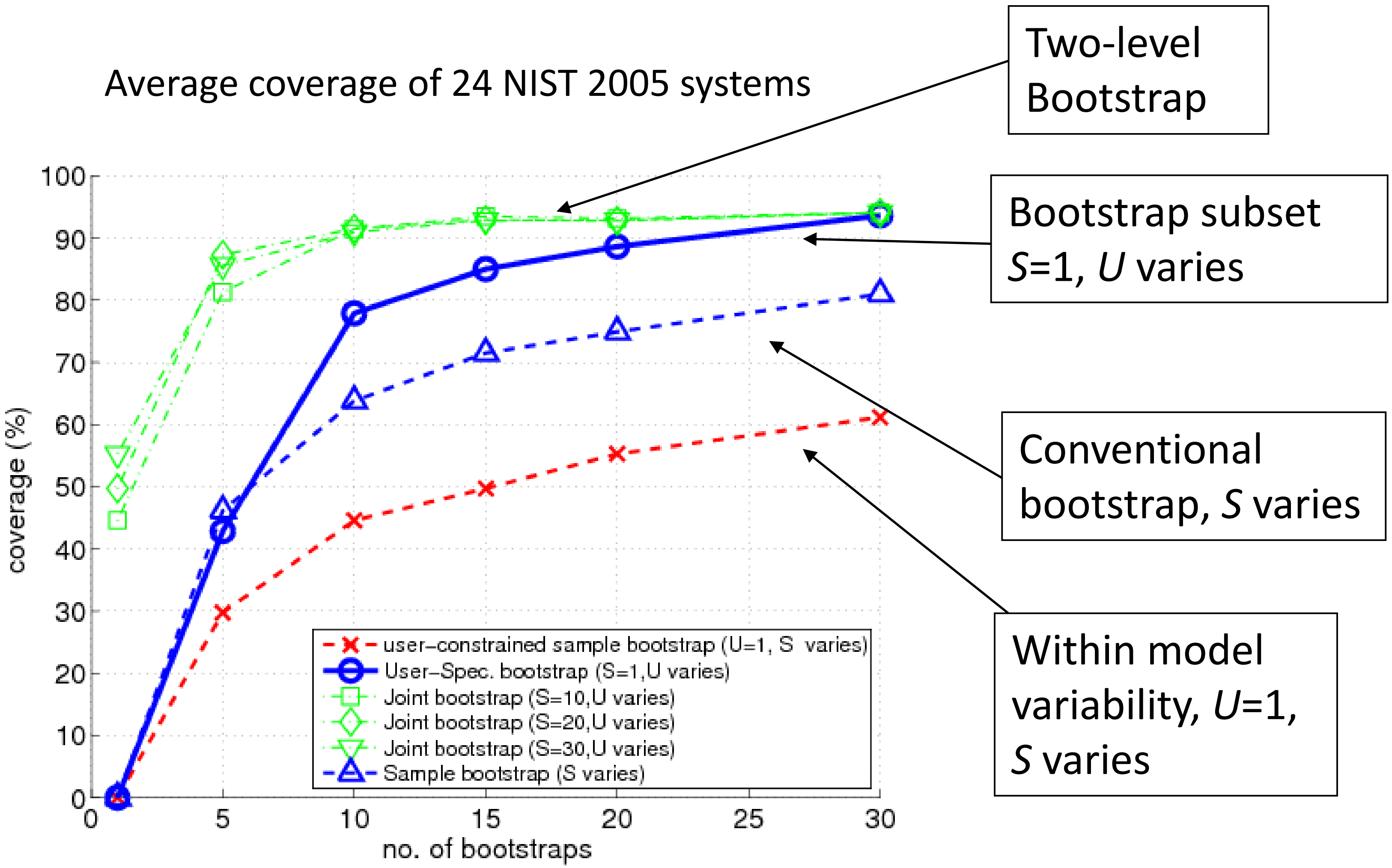
Findings

User-induced variability is more important than the sample variability

Increasing the subject population reduces the uncertainty

Coverage achieved 70-90% on NIST 2005 Speaker Evaluation data set

Average coverage of 24 NIST 2005 systems



Configuration: Trained on 31 users; tested on another 62 users

Quiz

How to best generate 10,000 genuine score samples in order to conduct a biometric test?

A

Recruit 10 subjects each contribute 1000+1 samples

B

Recruit 100 subjects each contribute 100+1 samples

C

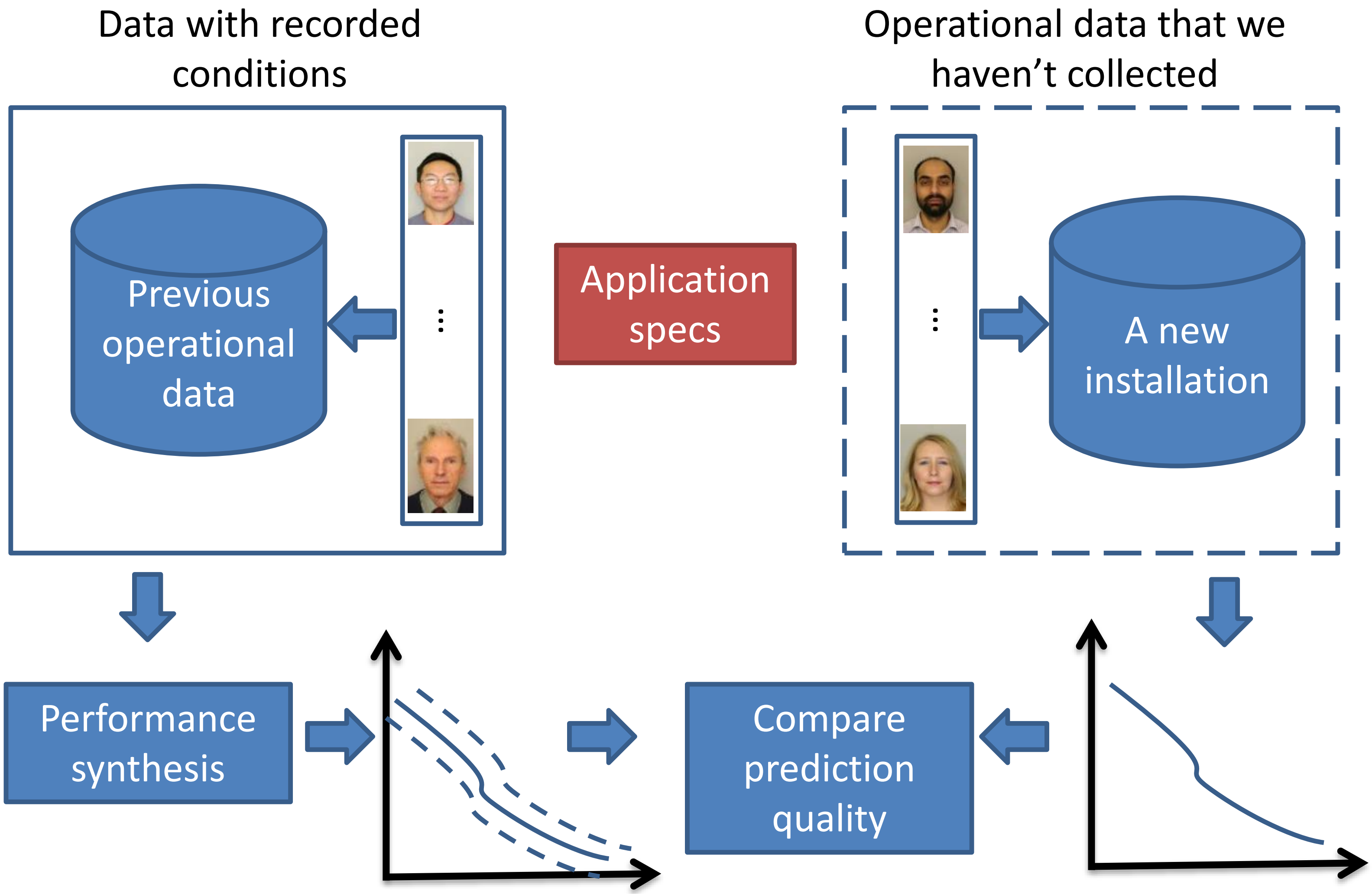
Recruit 1000 subjects each contribute 10+1 samples

Menu

On-going research

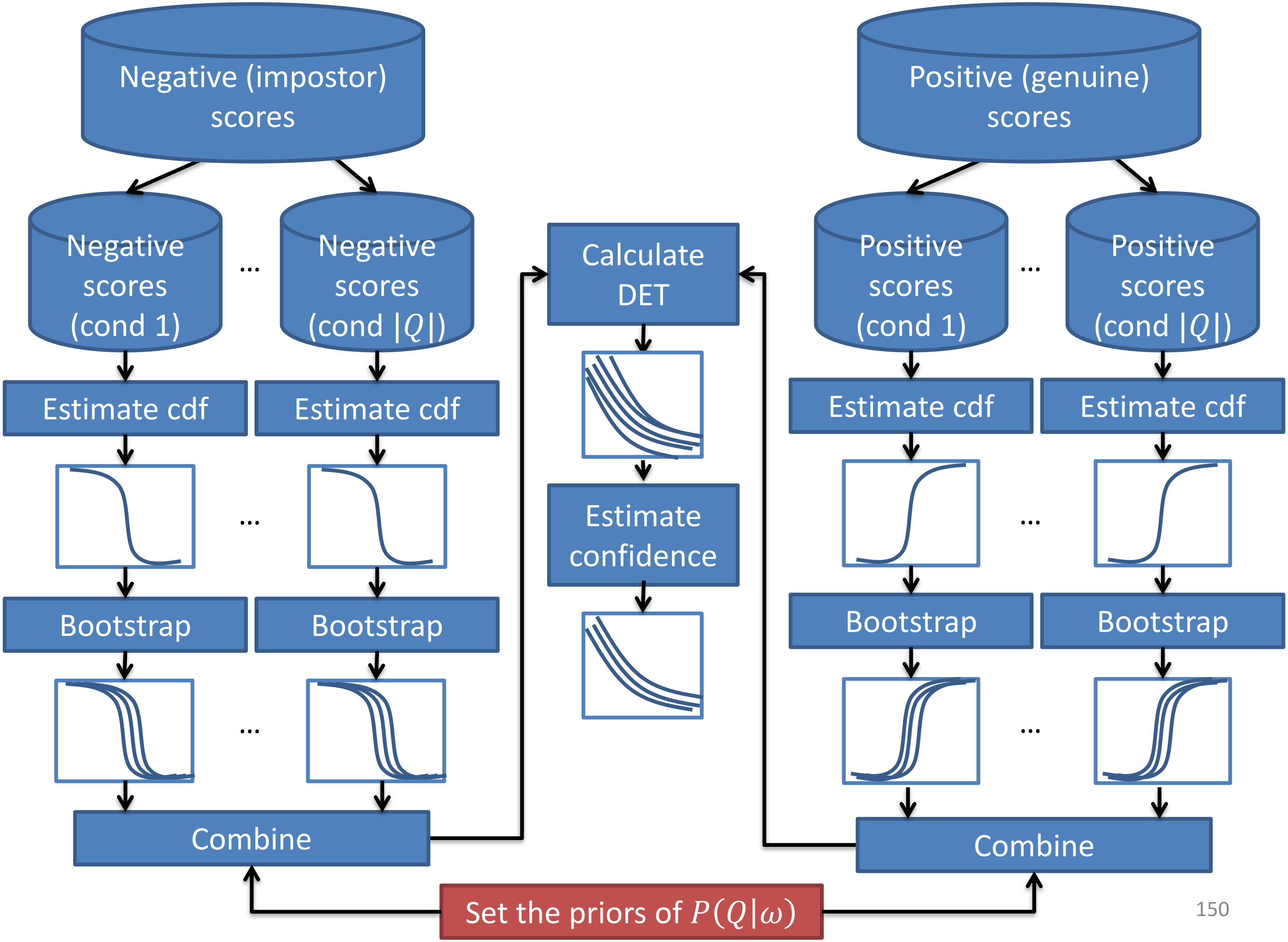
PERFORMANCE SYNTHESIS

Synthesizing DET curves to other operating conditions



Benefits of performance synthesis

- Liberates performance curves from its assessment data set
- Provides a framework for test reporting that promotes documentation and measurement of experimental conditions
- Potentially reduces cost in testing


Negative (impostor) scores
Positive (genuine) scores
Negative scores (cond 1)
...
Negative scores (cond $|Q|$)
Calculate DET
Positive scores (cond 1)
...
Positive scores (cond $|Q|$)
Estimate cdf
Estimate cdf
Estimate cdf
Estimate cdf
...
Estimate confidence
...
Bootstrap
Bootstrap
Bootstrap
Bootstrap
...
...
Combine
Combine
Set the priors of $P(Q|\omega)$

The underpinning theory

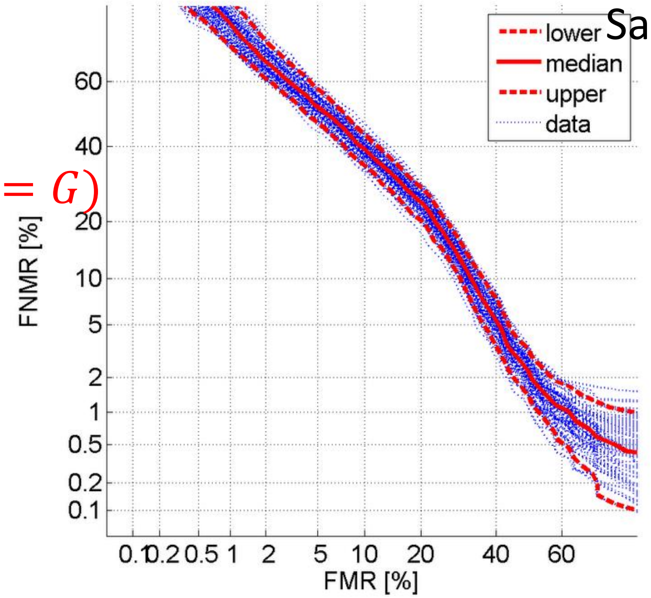
$$p(y|\omega) = \sum_Q p(y|\omega, Q) P(Q|\omega)$$

The system-level (class-conditional) score *pdf* is a mixture of factor-dependent *pdfs*

$$P(y \leq \Delta|\omega) = \sum_Q P(y \leq \Delta|\omega, Q) P(Q|\omega)$$

For performance estimating, we don't need to estimate the pdfs, but only their *cdf*s, which is monotonic function.

$$P(y \leq \Delta|\omega = G)$$
$$= \sum_{Q} P(y \leq \Delta|\omega = G, Q) P(Q|\omega = G)$$



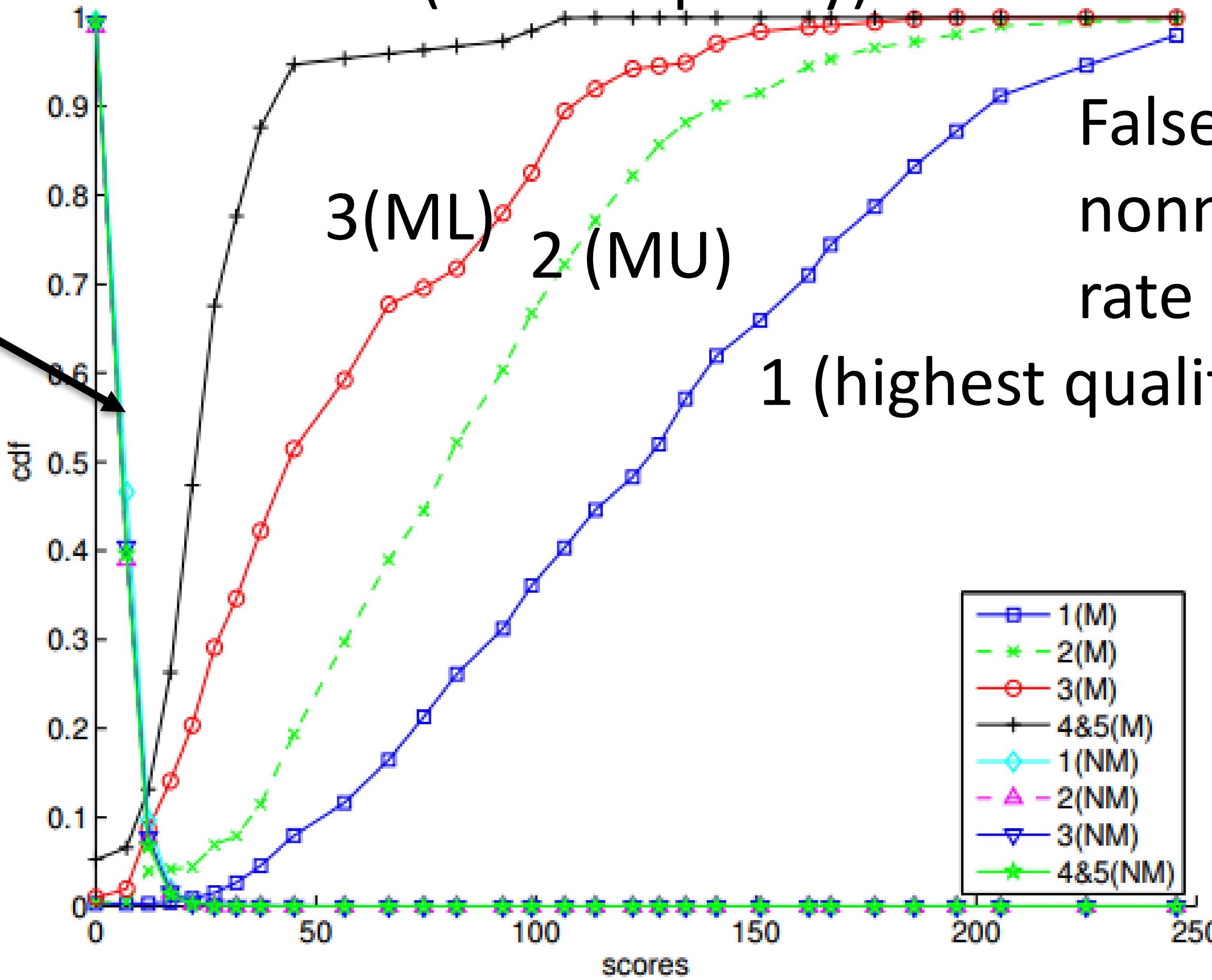
Same-pair

$$P(y > \Delta|\omega = I) = \sum_{Q} P(y > \Delta|\omega = I, Q) P(Q|\omega = I)$$

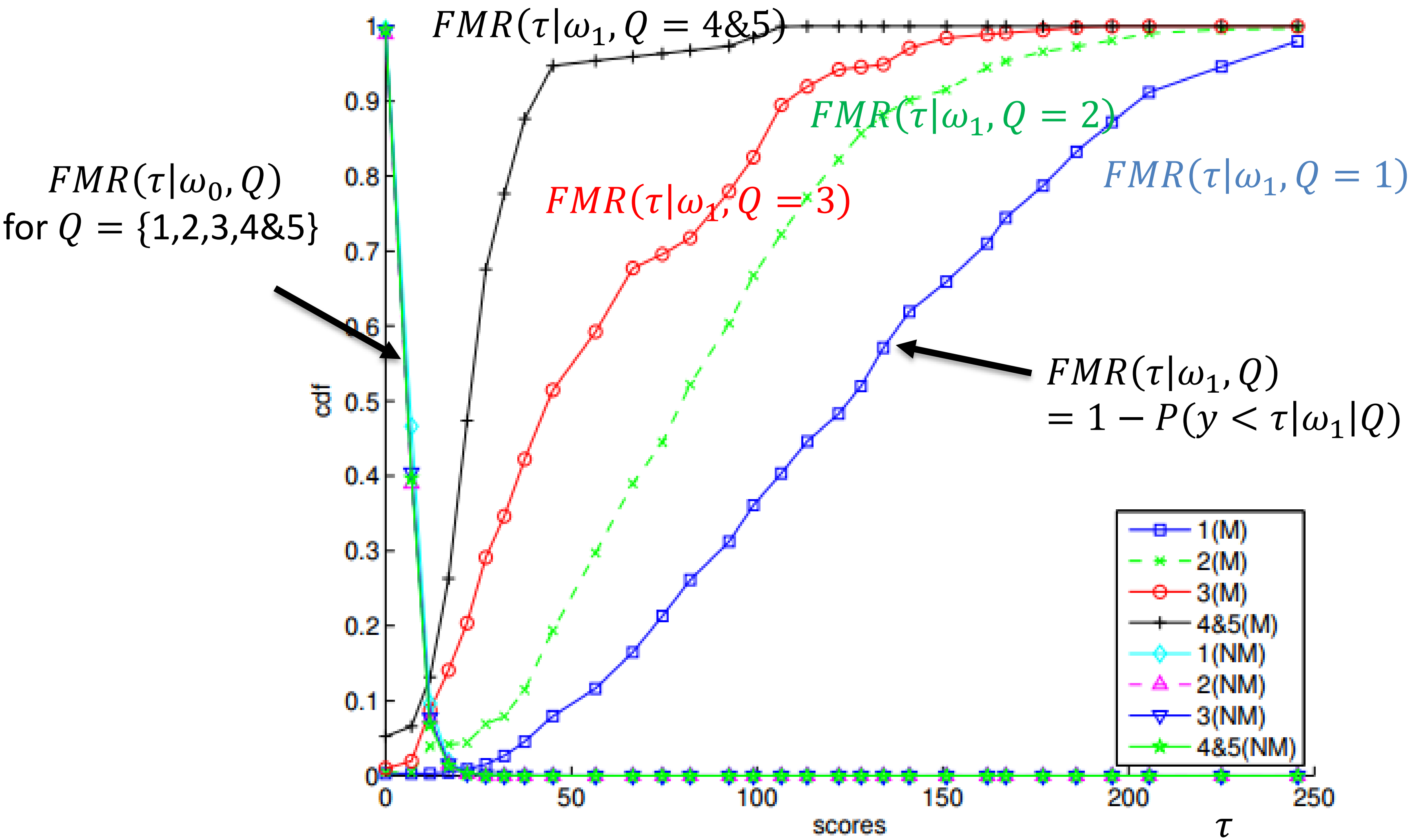
Fingerprint with NFIQ

4&5 (lowest quality)

Zero-effort impostor scores (False match Rate)

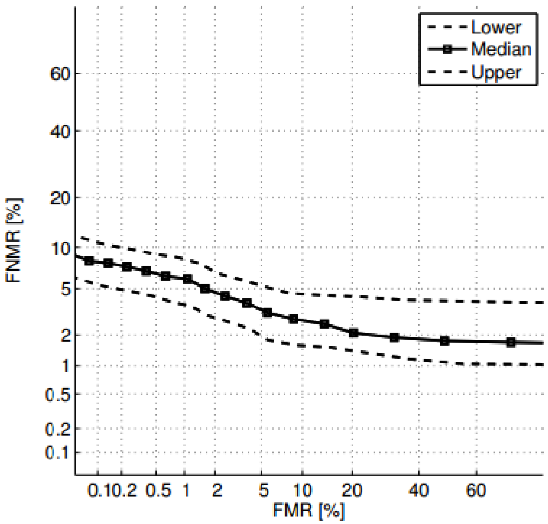


False nonmatch rate

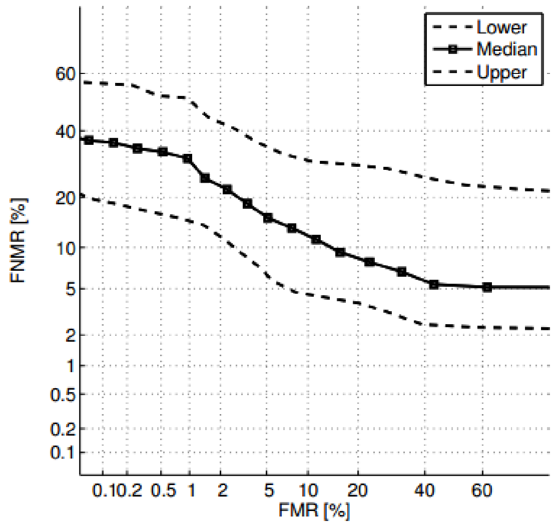


$FMR(\tau|\omega_1, Q = 4\&5)$
$FMR(\tau|\omega_1, Q = 2)$
$FMR(\tau|\omega_1, Q = 1)$
$FMR(\tau|\omega_0, Q)$
for $Q = \{1,2,3,4\&5\}$
$FMR(\tau|\omega_1, Q = 3)$
$FMR(\tau|\omega_1, Q)$
$= 1 - P(y < \tau|\omega_1|Q)$
cdf
scores
τ
1(M)
2(M)
3(M)
4&5(M)
1(NM)
2(NM)
3(NM)
4&5(NM)

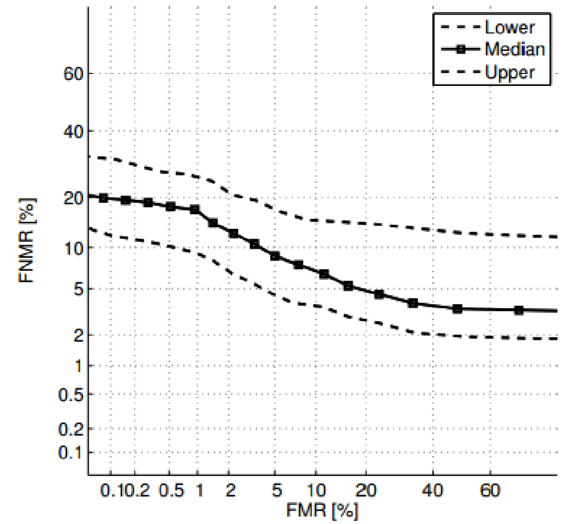
Simulated DET curves of varying quality



High quality tendency
[8 4 2 1]



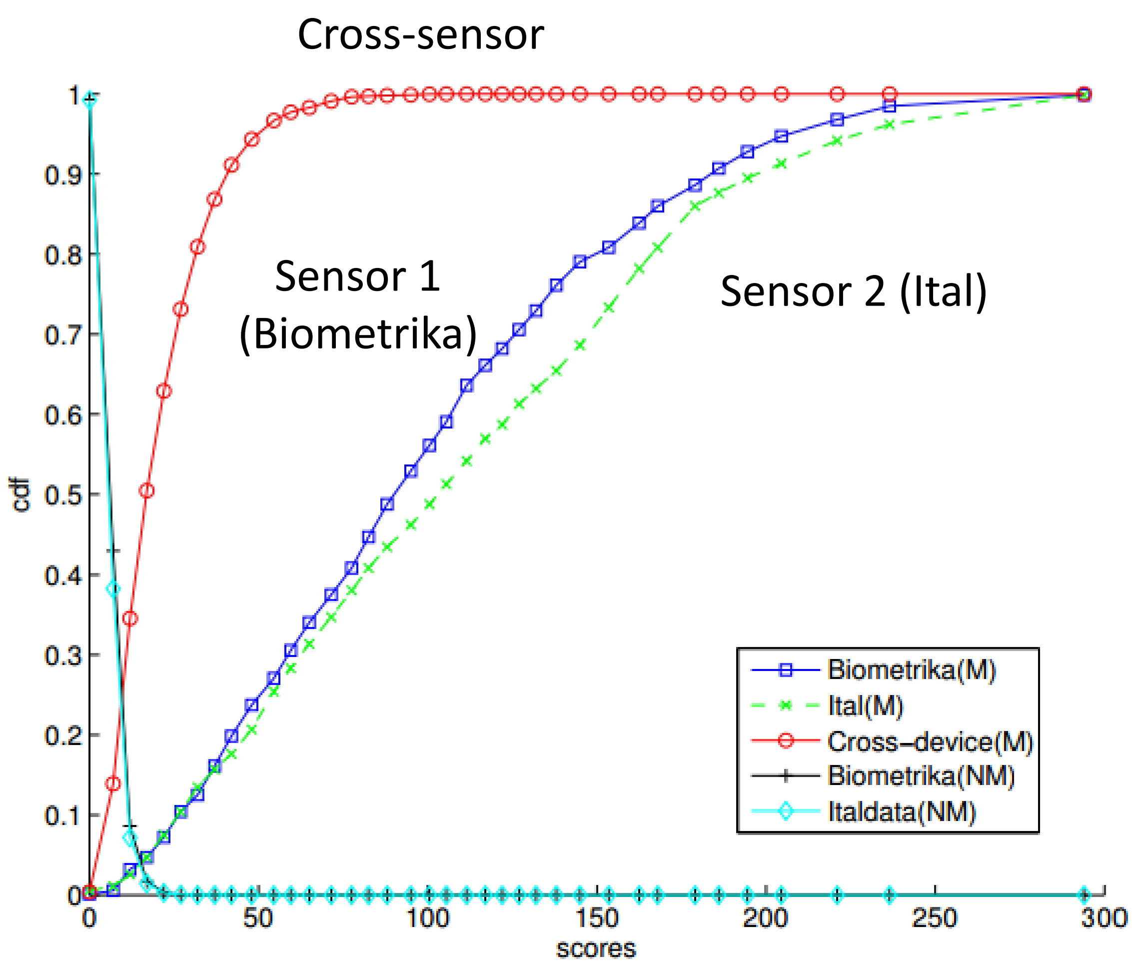
Low quality tendency
[1 2 4 8]



Equal prior quality
[1 1 1 1]

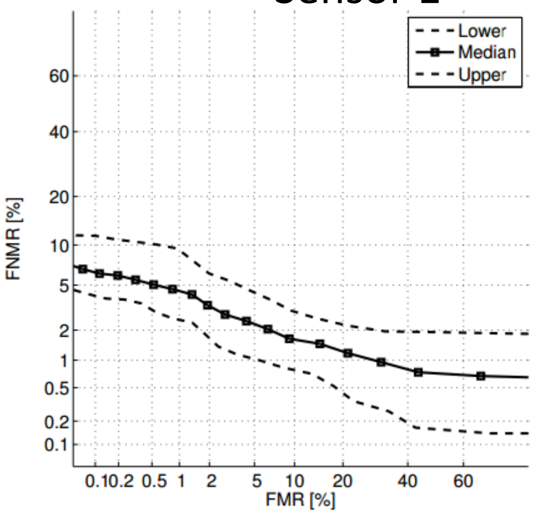
Relative weights of the 4 curves:
[Q1 Q2 Q3 Q4&5] or [H MU ML L]

Study II: Cross-sensor performance



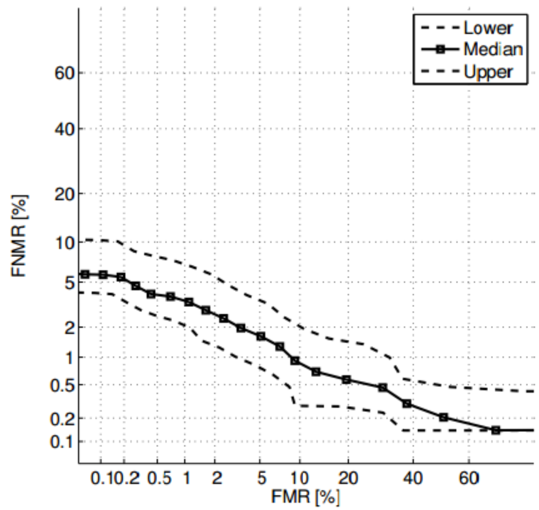
Simulated DET curves with mixture of data

Sensor 1



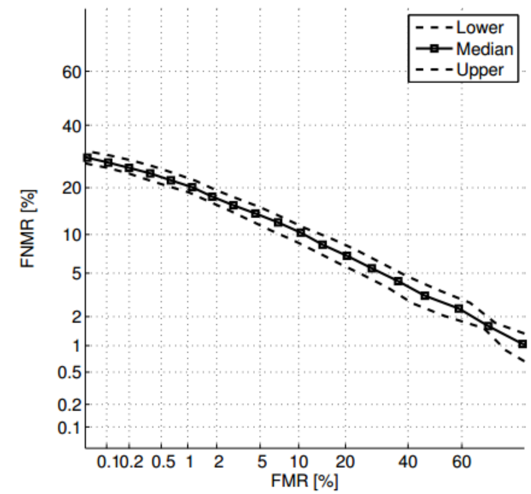
[1 0 0]

Sensor 2



[0 1 0]

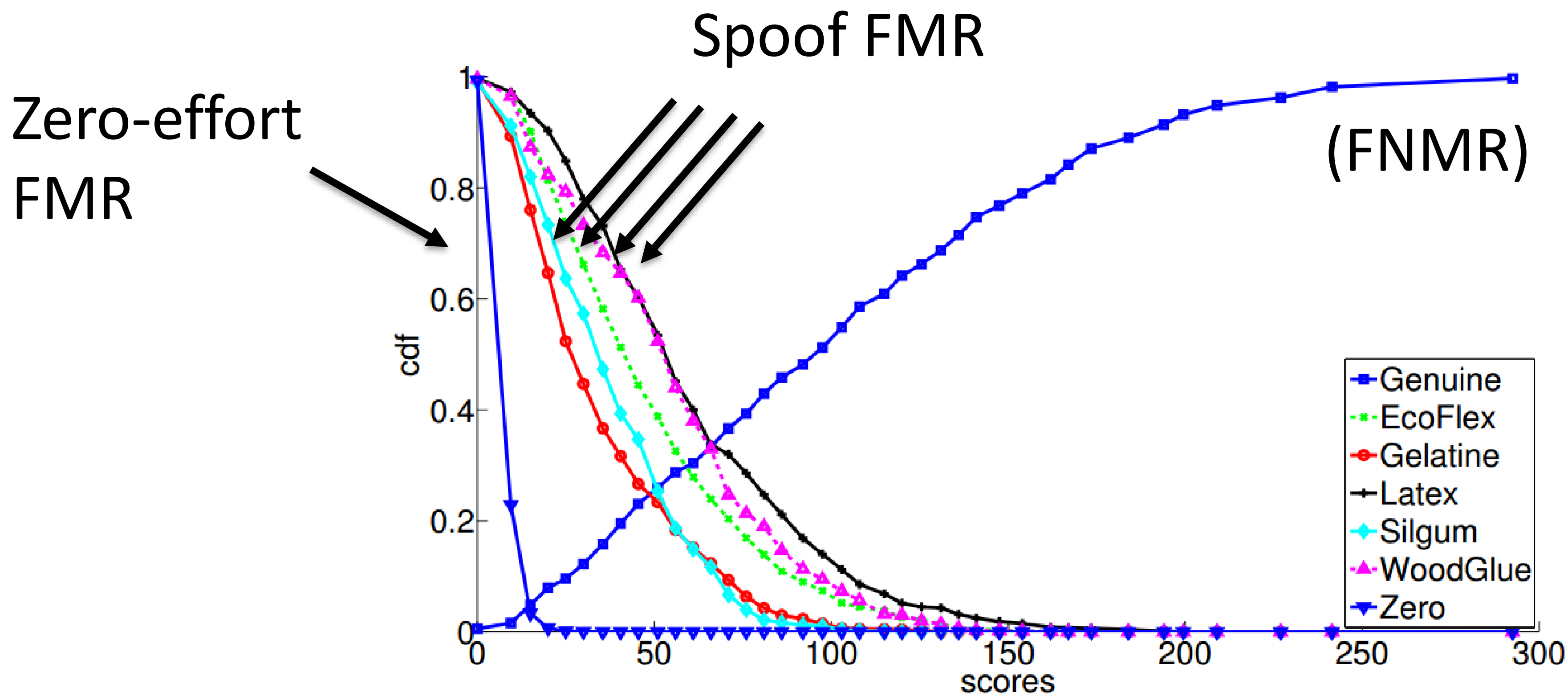
Sensor 1 + Senso 2
+ Cross-sensor



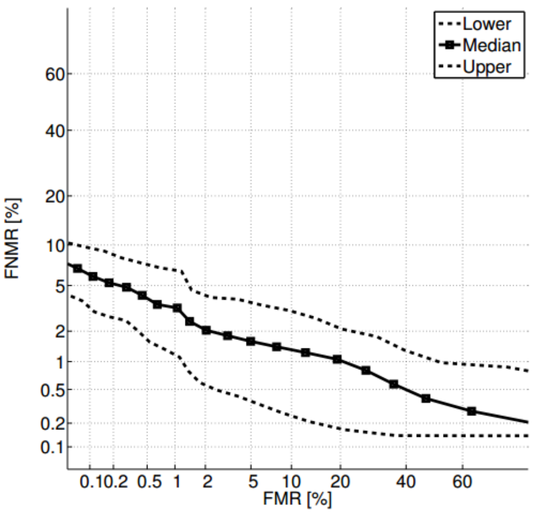
[1 1 1]

The ratio shows [S1 S2 cross-sensor]

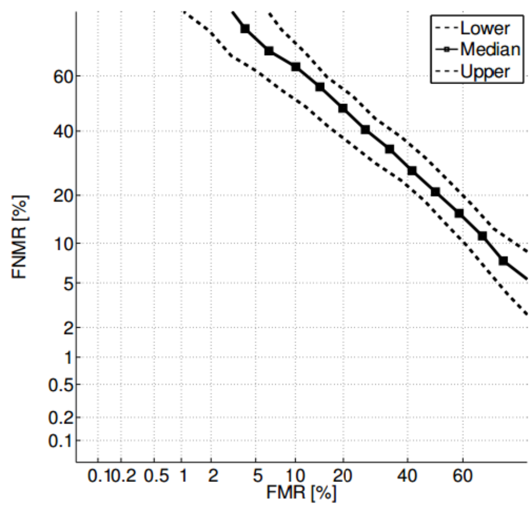
Study III: Spoof attack



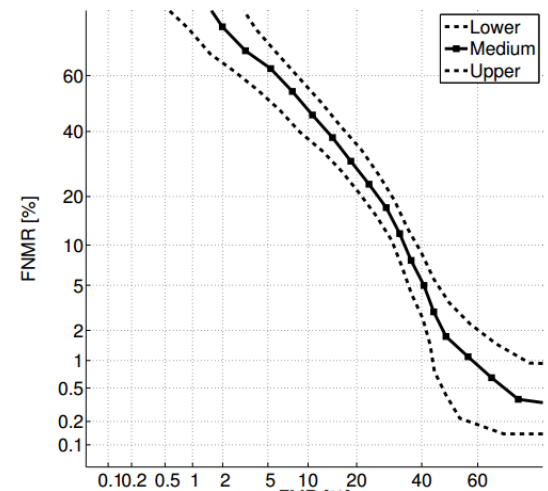
Performance under various attacks



Zero-effort attack



Spoof attack

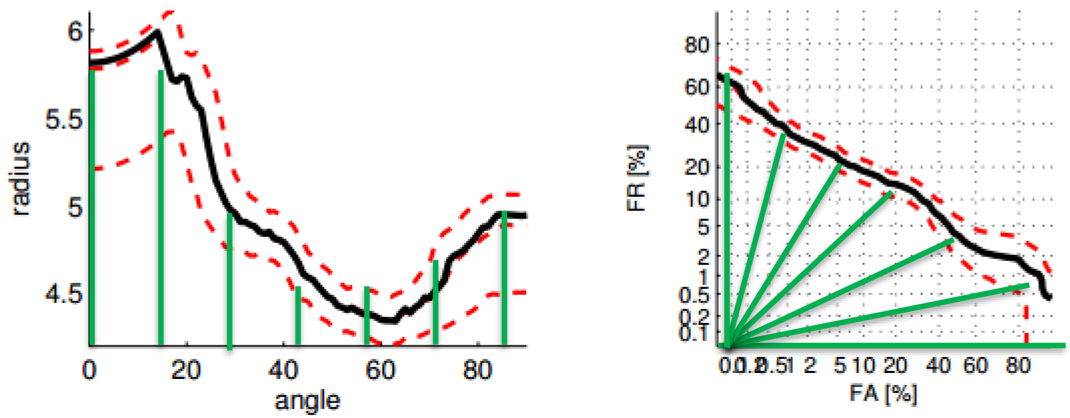


50% zero-effort
+ 50% spoof attack

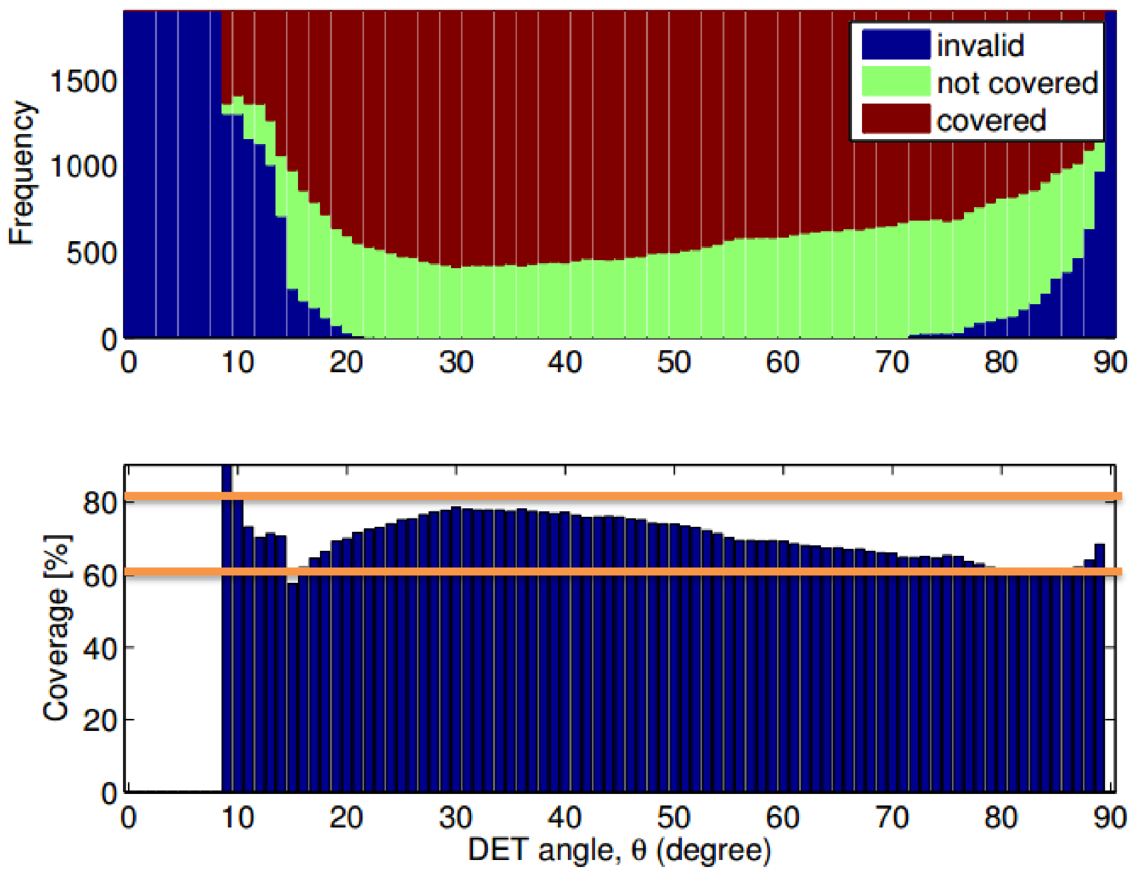
Poh, Norman, and Chi Ho Chan. "Generalizing DET curves across application scenarios." *Information Forensics and Security, IEEE Transactions on* 10.10 (2015): 2171-2181.

Menu

Assessment with DET angles



Achievable coverage



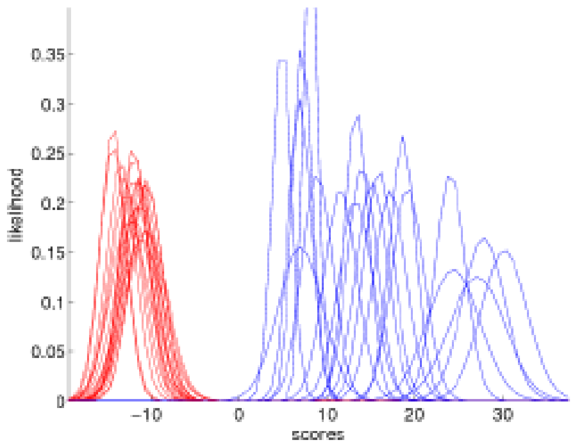
Coverage achieved between 60% and 80%

USER-SPECIFIC SCORE NORMALISATION

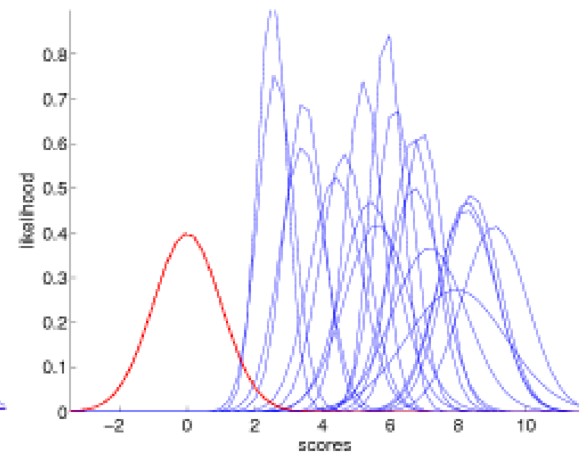
Score normalisation

Original matching scores

$$y$$



(a) baseline



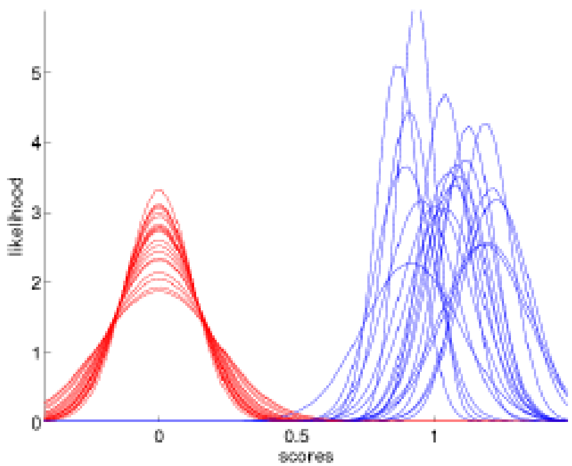
(b) Z-norm

Z-norm

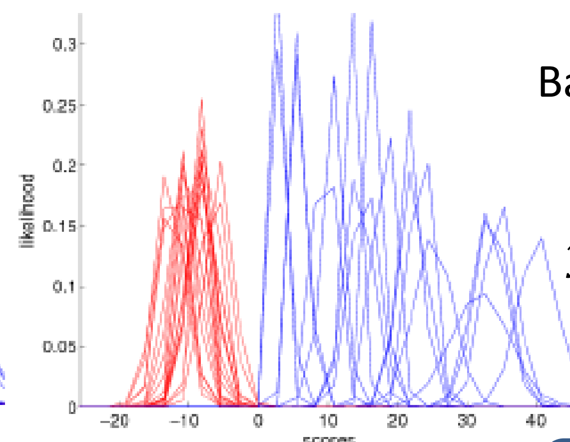
$$y^Z = \frac{y - \mu_j^I}{\sigma_j^I}$$

F-norm

$$y^F = \frac{y - \mu_j^I}{{\sim}^{G}_{j} - \mu_j^I}$$



(c) F-norm



(d) likelihood ratio norm

Bayesian classifier

$$y^B = \log \frac{p(y|I,j)}{p(y|G,j)}$$

[IEEE TASLP'08]

Procedures	Formulas	Properties
Z-norm	$y_j^Z = \frac{y-\mu_j^{\mathtt{I}}}{\sigma_j^{\mathtt{I}}}$	$E_j[y_j^Z\vert\mathtt{I}] = 0$ and $var_j[y_j^Z\vert\mathtt{I}] = 1$
F-norm	$y_j^F = \frac{y-\mu_j^{\mathtt{I}}}{\gamma\mu_j^{\mathtt{C}}+(1-\gamma)\mu^{\mathtt{C}}-\mu_j^{\mathtt{I}}}$	$E_j[y_j^F\vert\mathtt{I}] = 0$ and $E_j[y_j^F\vert\mathtt{C}] = 1$
EER-norm	$y^{EER} = y - \Delta_j$	$y_j^{EER} > 0$ is an optimal decision function (at EER) for all j
MS-LLR norm	$y^{llr} = \log\frac{p(y\vert\mathtt{C},j)}{p(y\vert\mathtt{I},j)}$	$y_j^{llr} > 0$ is an optimal decision function (at EER) for all j

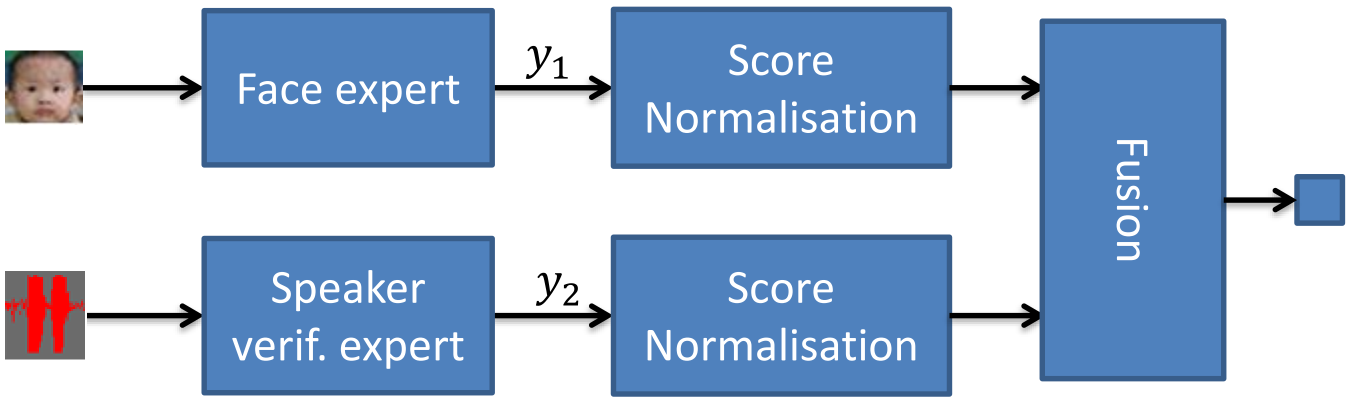
$$\mu_j^{F,\mathtt{C}} \equiv E[y_j^F|\mathtt{C}] = \frac{E[y_j^{\mathtt{C}}] - \mu_j^{\mathtt{I}}}{\mu_j^{\mathtt{C}} - \mu_j^{\mathtt{I}}} = 1, \text{ for all } j$$

$$\mu_j^{F,\mathtt{I}} \equiv E[y_j^F|\mathtt{I}] = \frac{E[y_j^{\mathtt{I}}] - \mu_j^{\mathtt{I}}}{\mu_j^{\mathtt{C}} - \mu_j^{\mathtt{I}}} = 0, \text{ for all } j$$

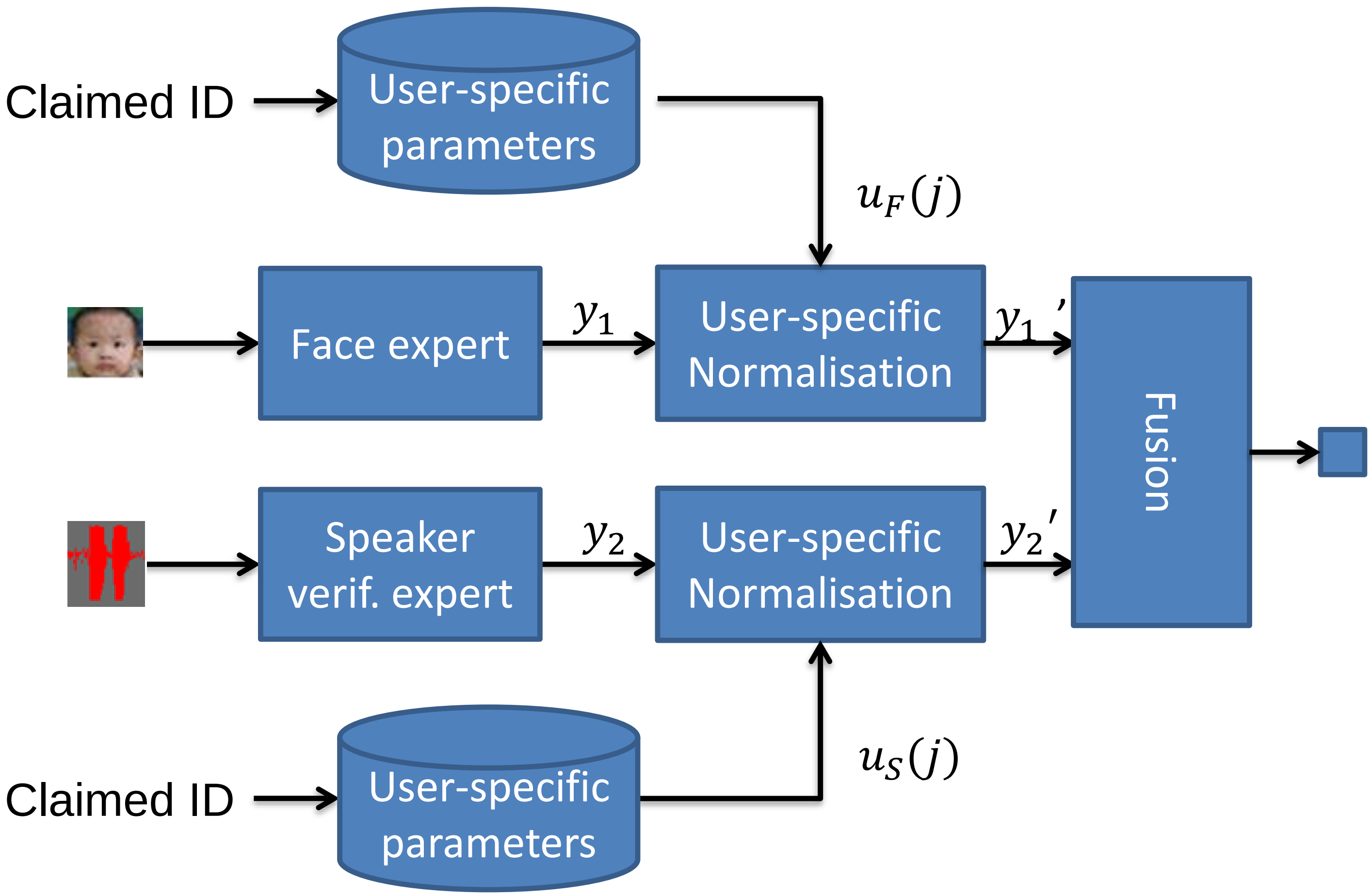
Menu

USER-SPECIFIC FUSION

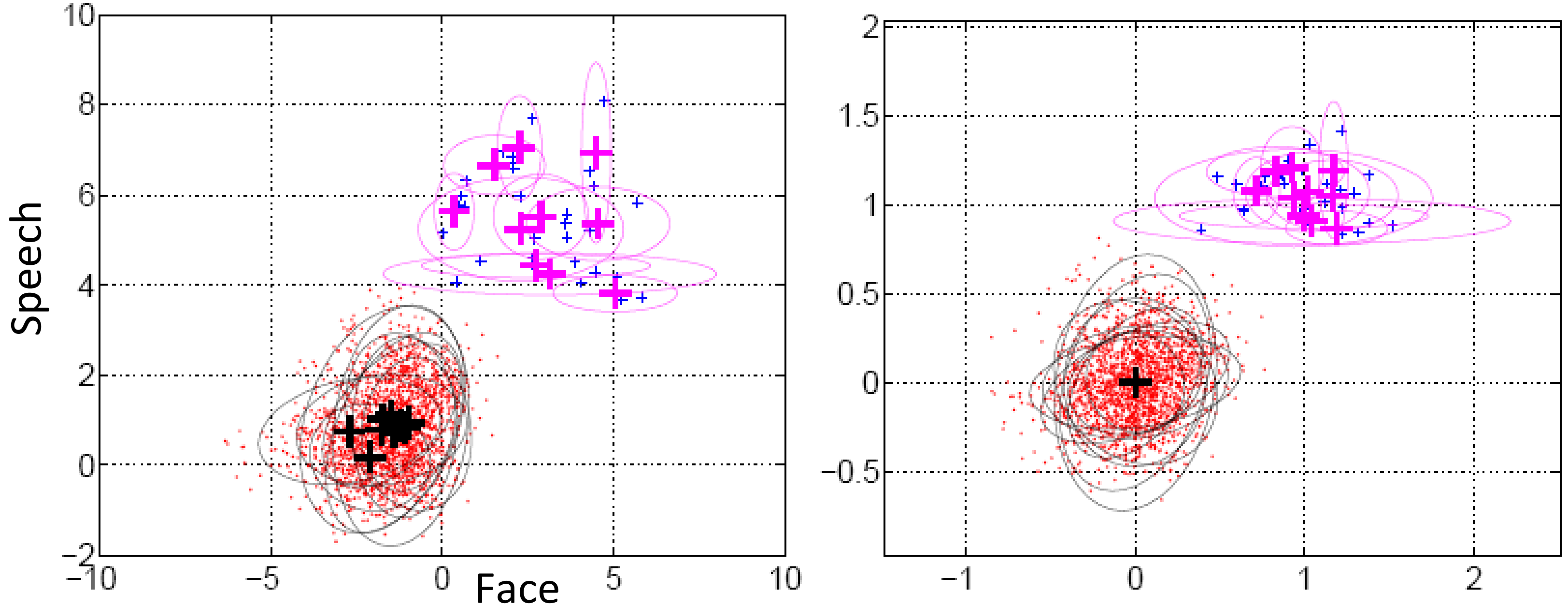
Conventional multimodal fusion



User-specific multimodal fusion



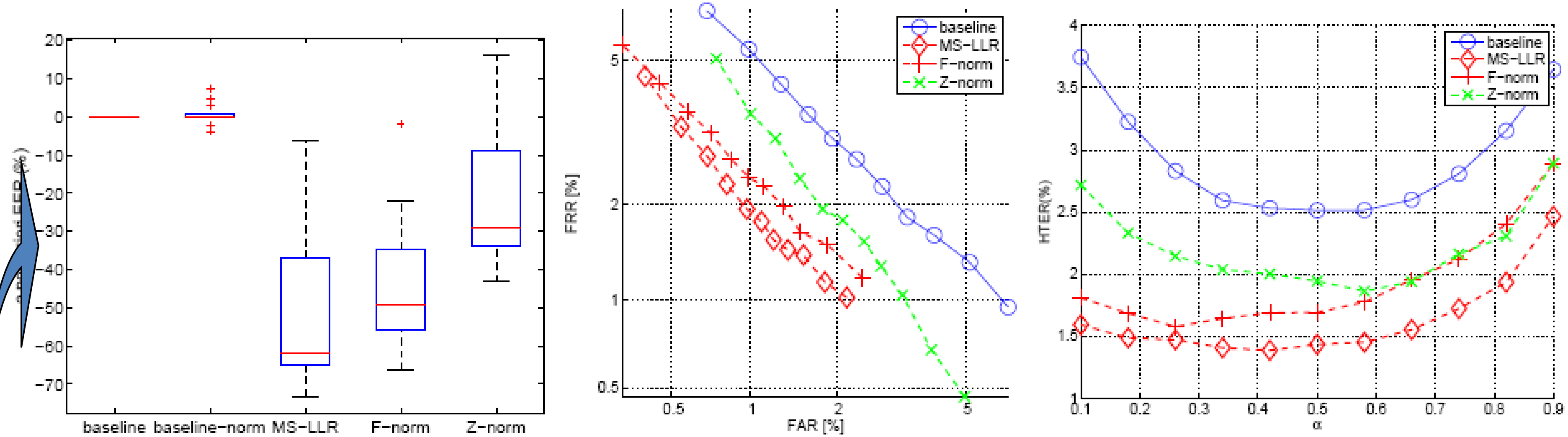
Client-specific fusion



Original fusion problem

Fusion problem after applying F-norm

Some results



(a) relative change of EER (b) DET (c) EPC

$$\boxed{\frac{EER_{\text{norm}} - EER_{\text{bline}}}{EER_{\text{bline}}}}$$

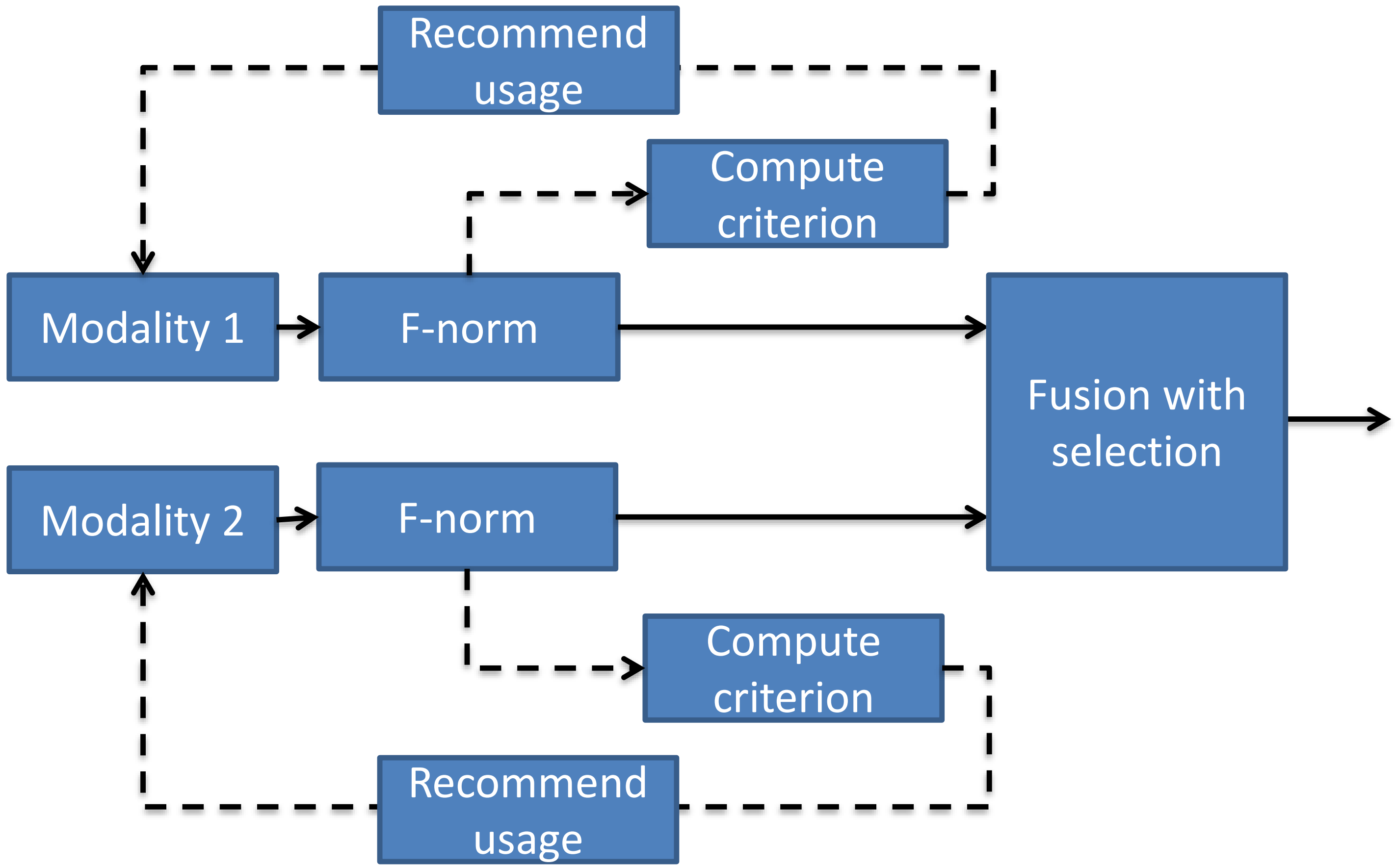
Menu

[IEEE TASLP'08]

USER-SPECIFIC & SELECTIVE FUSION

System architecture

Criterion = “B-norm of Fratio”

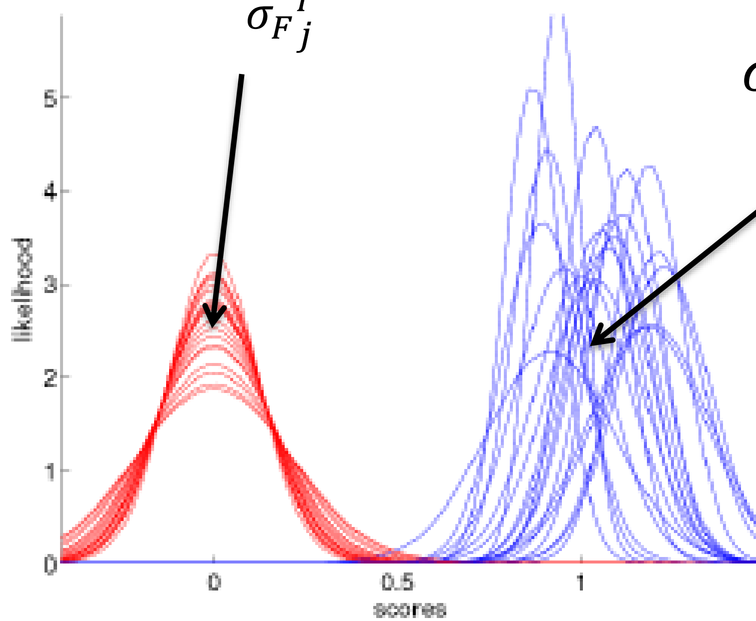


Rationale: The B-ratio of F-norm

$$\text{F-ratio} = \frac{\mu_j^G - \mu_j^I}{\sigma_j^G + \sigma_j^I}$$

$$\text{B-ratio} = \frac{\mu_j^G - \mu_j^I}{\sigma_j^I}$$

The B-ratio of F-norm $= \dfrac{\mu_F{}_j^G - \mu_F{}_j^I}{\sigma_F{}_j^I} = \dfrac{1}{\sigma_F{}_j^I}$

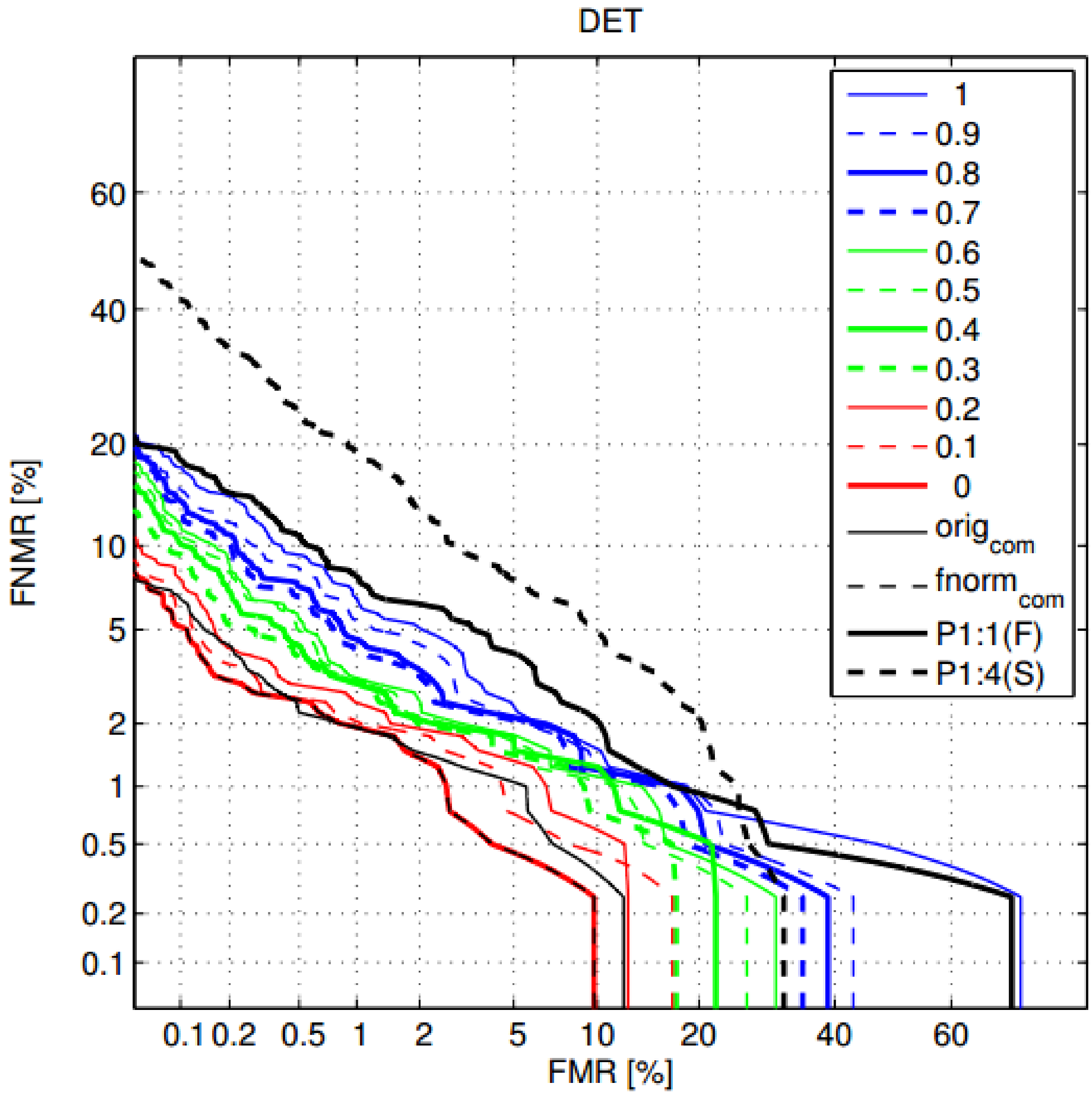


Cost savings

$$\text{computational saving} = 1 - \frac{\sum_{j \in \mathcal{J}} \sum_{i=1}^{N} I(\text{system}_{i,j})}{N \times J}$$

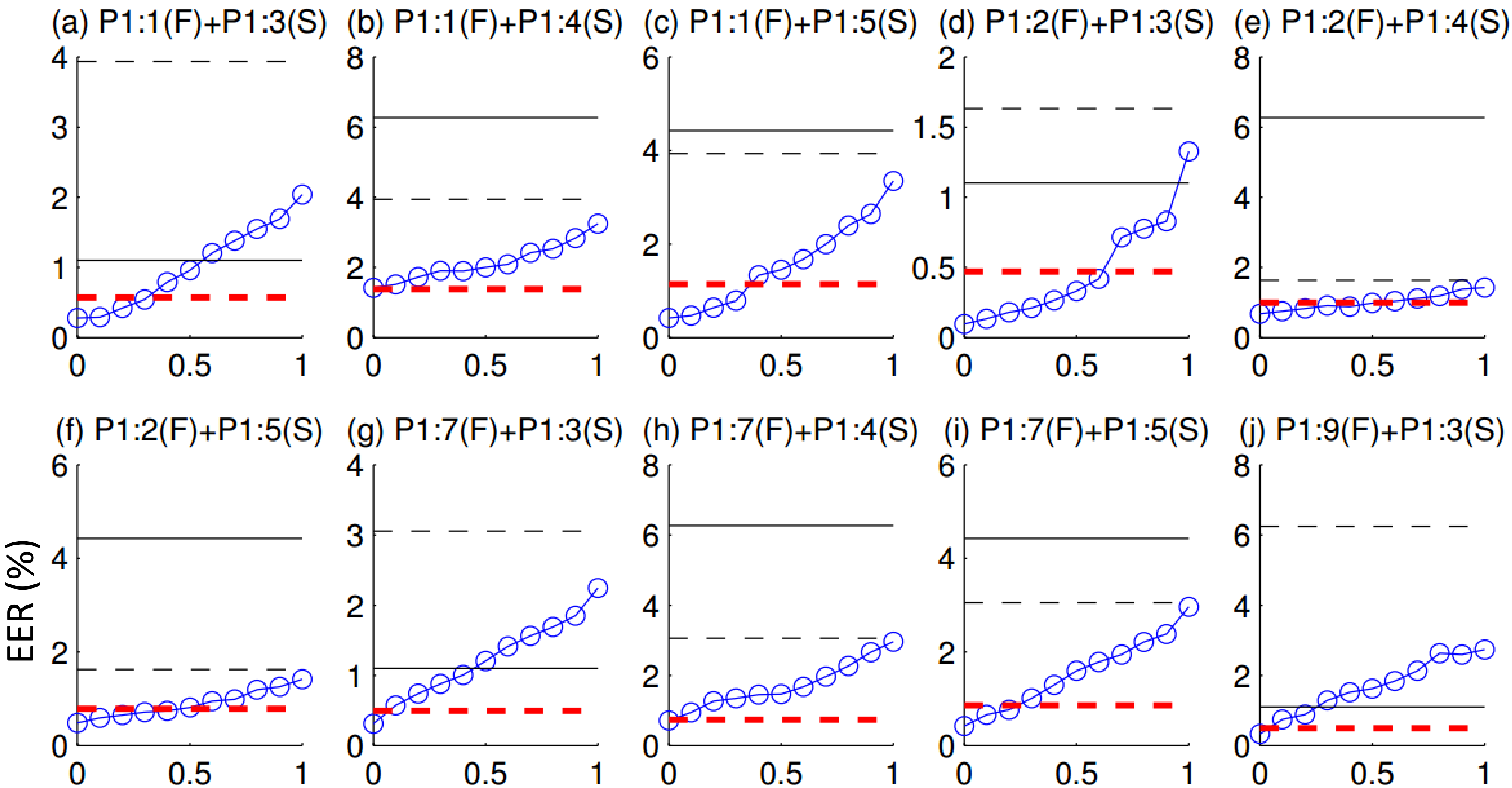
$$= \frac{r}{N} \times 100\%$$

XM2VTS database.
# of users J = 200
# of system = 2



A user-specific and selective multimodal biometric fusion strategy by ranking subjects, PRJ 2013

Performance on test set

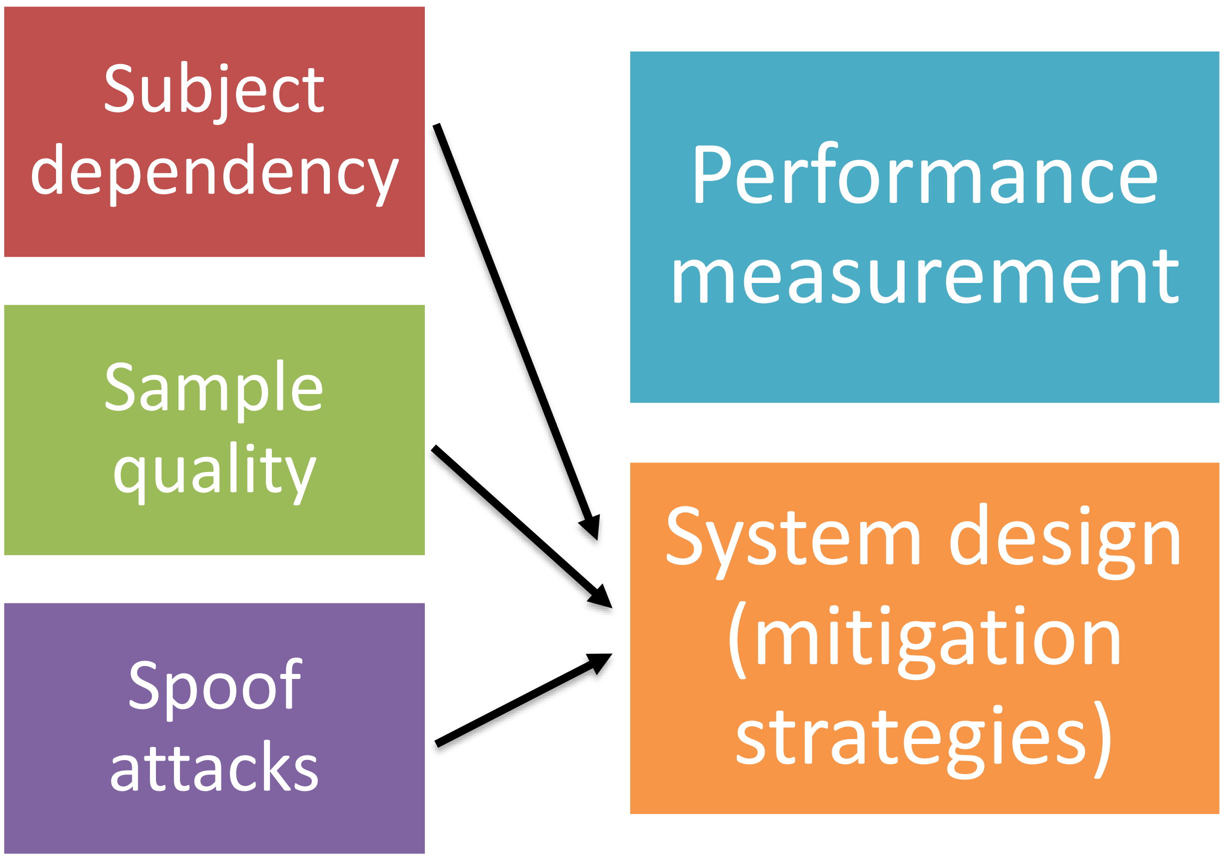


Computational savings

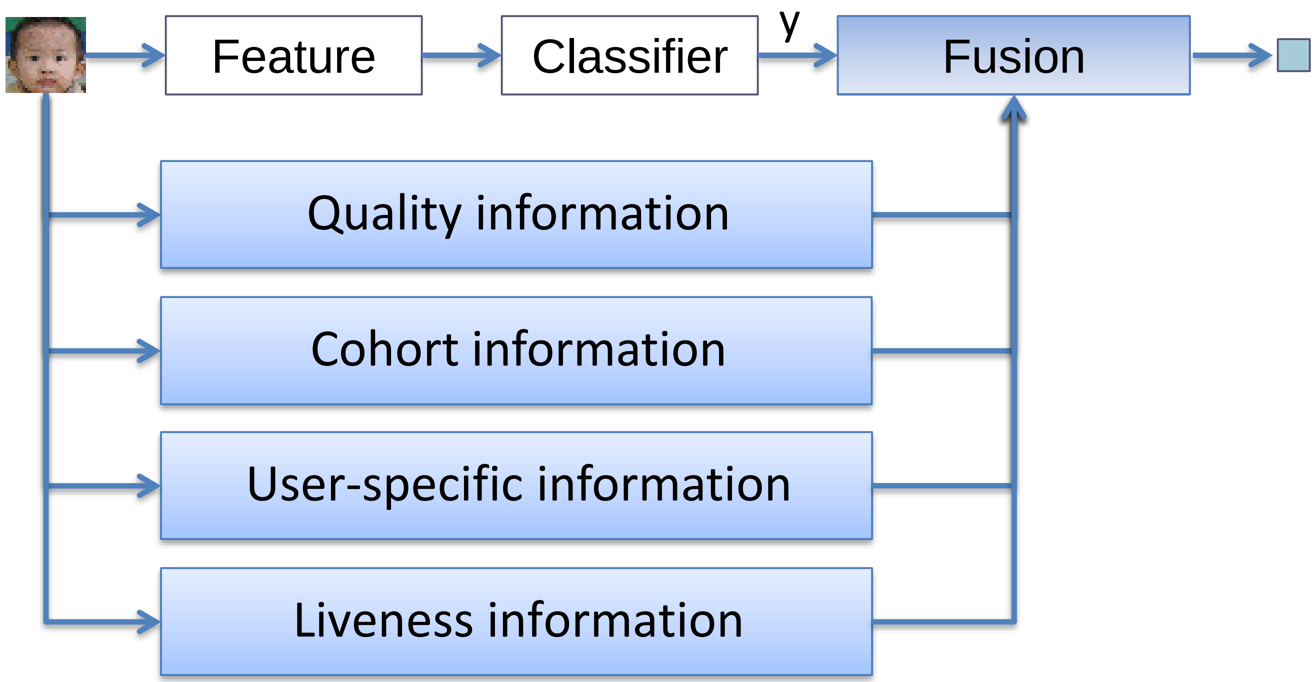
10 of 15 fusion experiments shown here

Menu

HETEROGENOUS INFORMATION FUSION

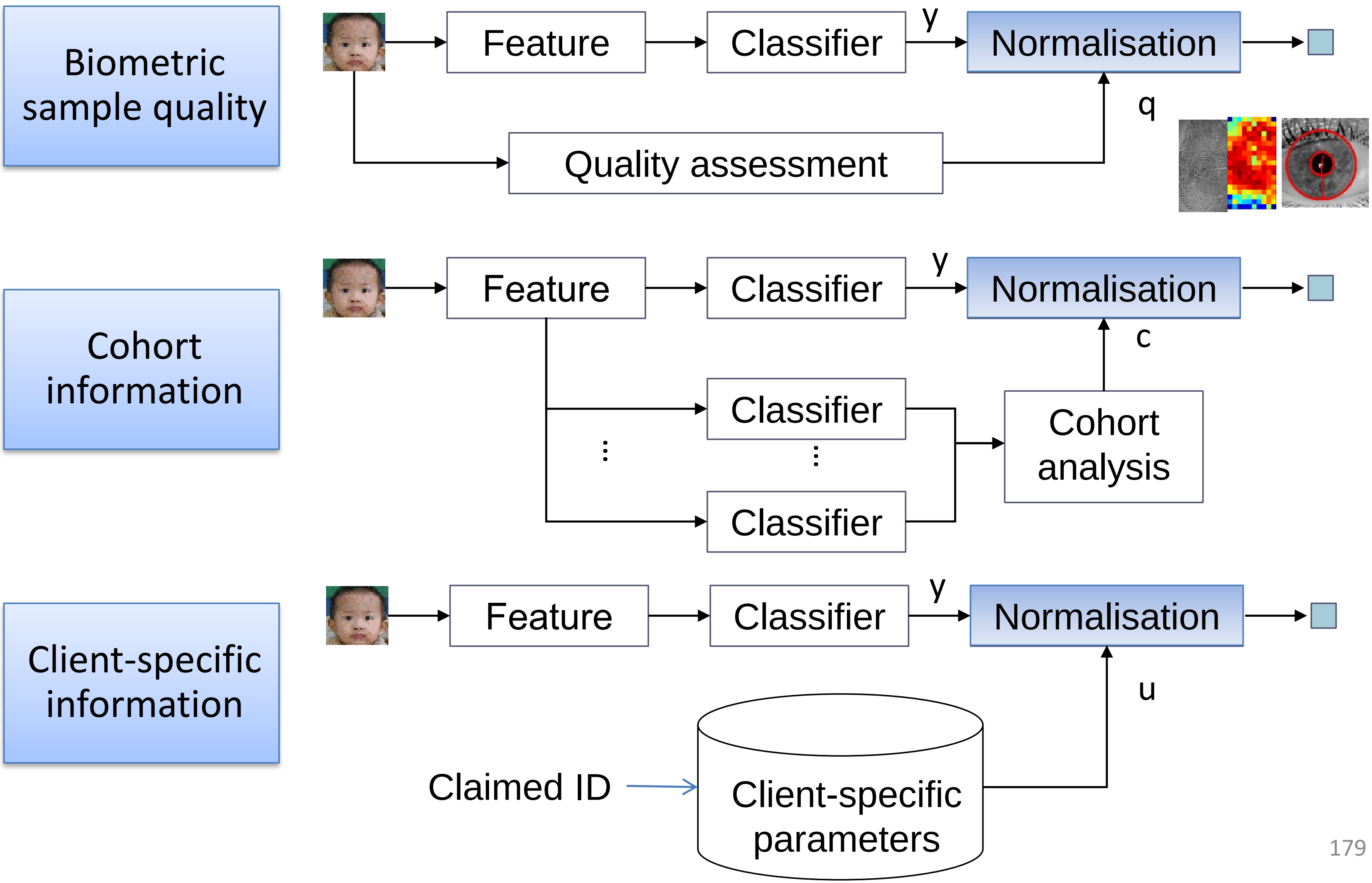

Subject dependency
Sample quality
Spoof attacks
Performance measurement
System design (mitigation strategies)

Heterogeneous information fusion

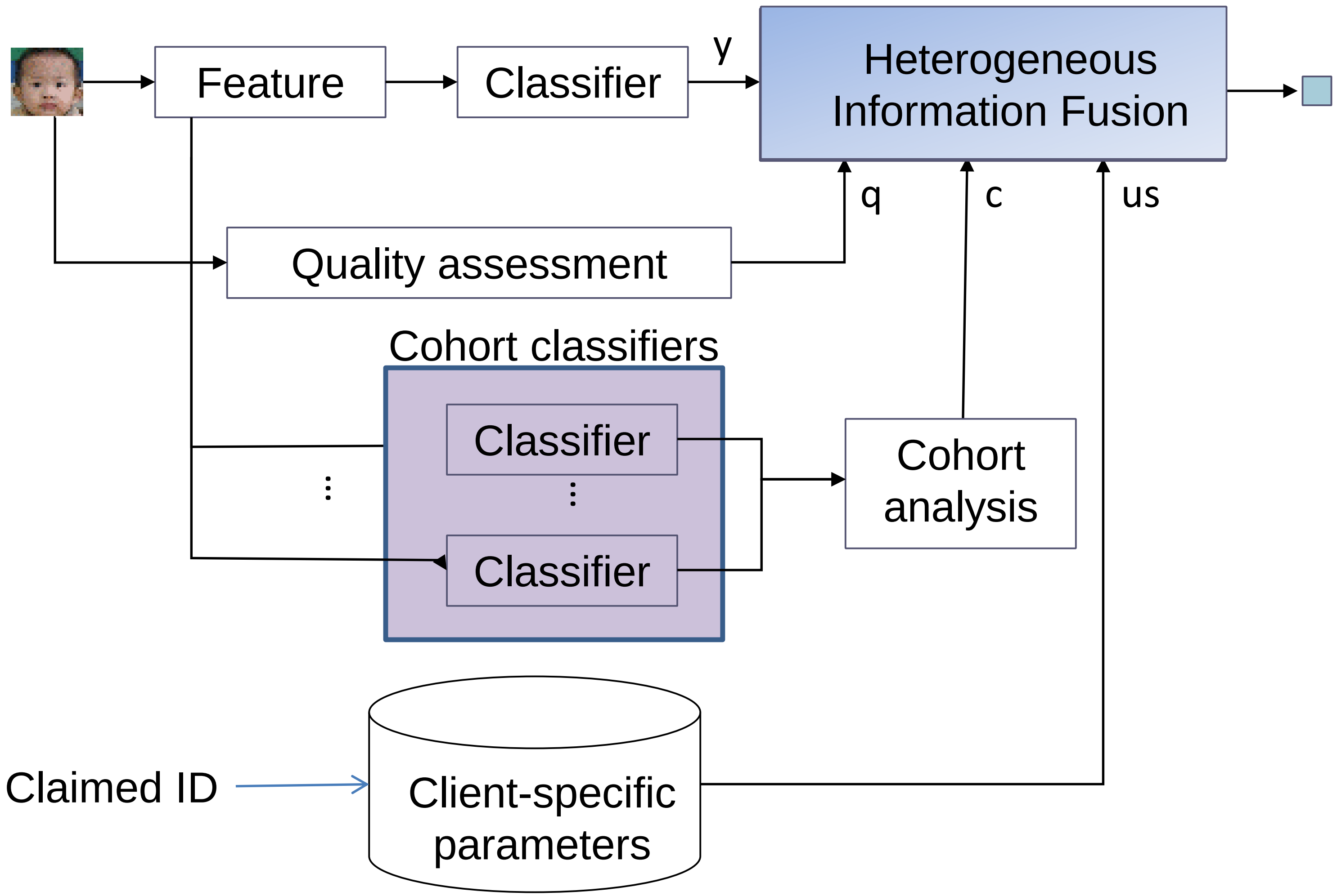


Poh, Norman, Amin Merati, and Josef Kittler. "Heterogeneous information fusion: A novel fusion paradigm for biometric systems." *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 2011.

Three sources of information



One system



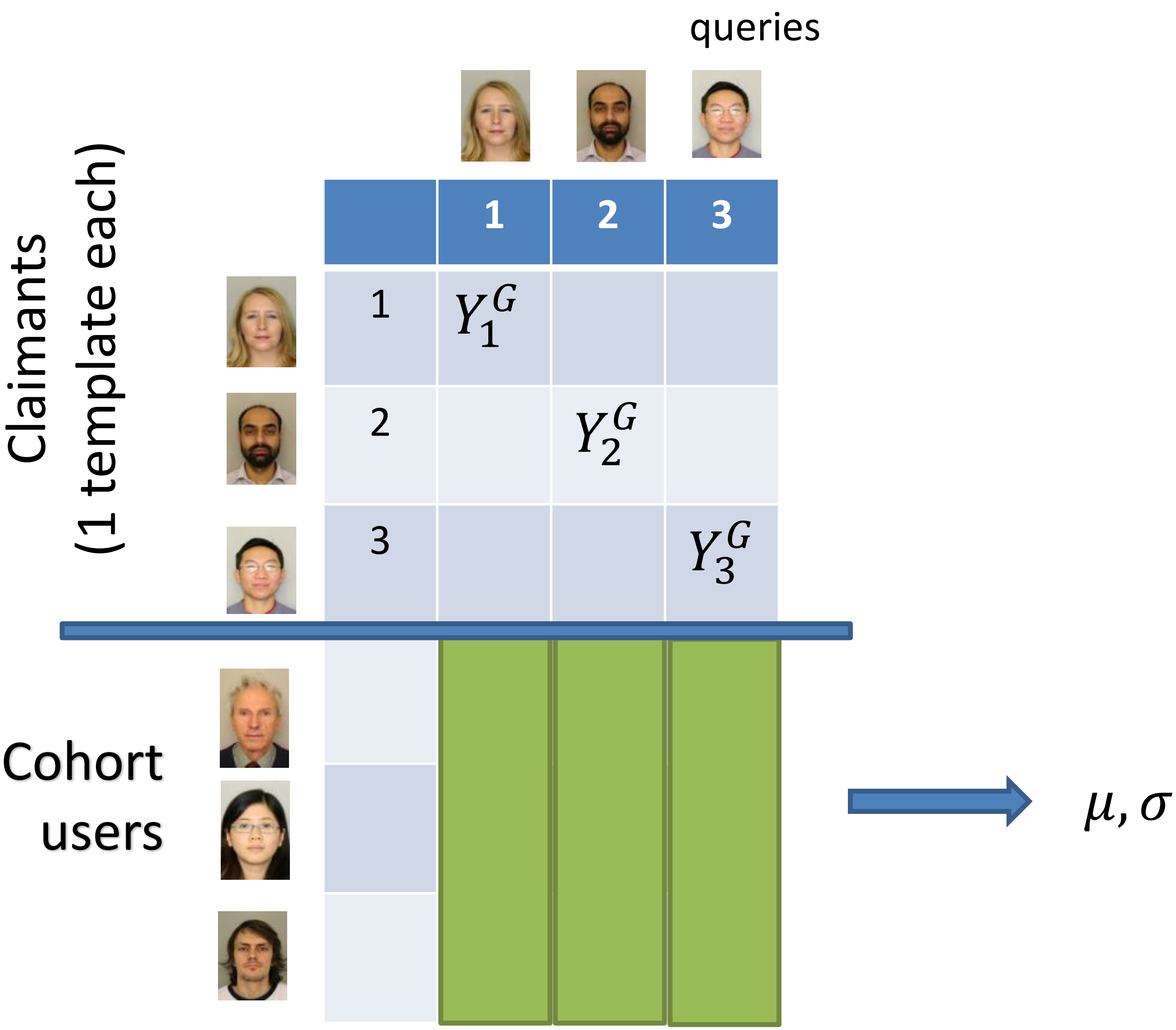
Feature
Classifier
y
Heterogeneous Information Fusion
q
c
us
Quality assessment
Cohort classifiers
Classifier
⋮
Classifier
Cohort analysis
Claimed ID
Client-specific parameters

Representing the information sources

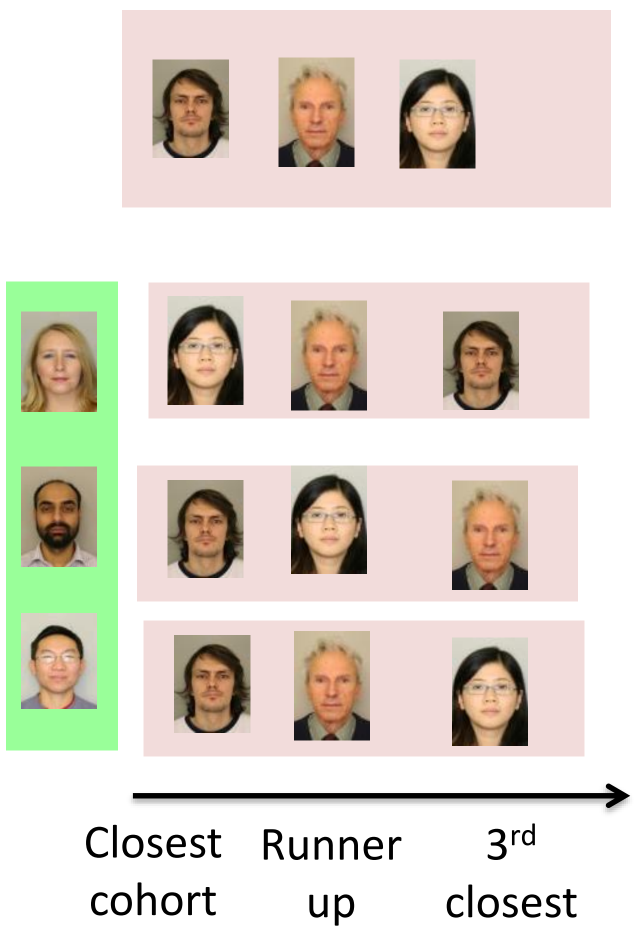
Biometric sample quality	Cohort information	Client-specific information
$s_{\mathtt{Q}} \equiv [q^{\mathtt{tmplt}}, q^{\mathtt{qry}}]$	$s_{\mathtt{C}} \equiv [\mu^{\mathtt{C}}, \sigma^{\mathtt{C}}]$	$s_{\mathtt{US}} \equiv [\mu_1^{\mathtt{US}}, \mu_0^{\mathtt{US}}, \sigma_0^{\mathtt{US}}]$
Quality of template Quality of probe	Mean of cohort scores Std. dev. of cohort scores	Mean of US impostor scores Std. dev. of US impostor scores Mean of US genuine scores

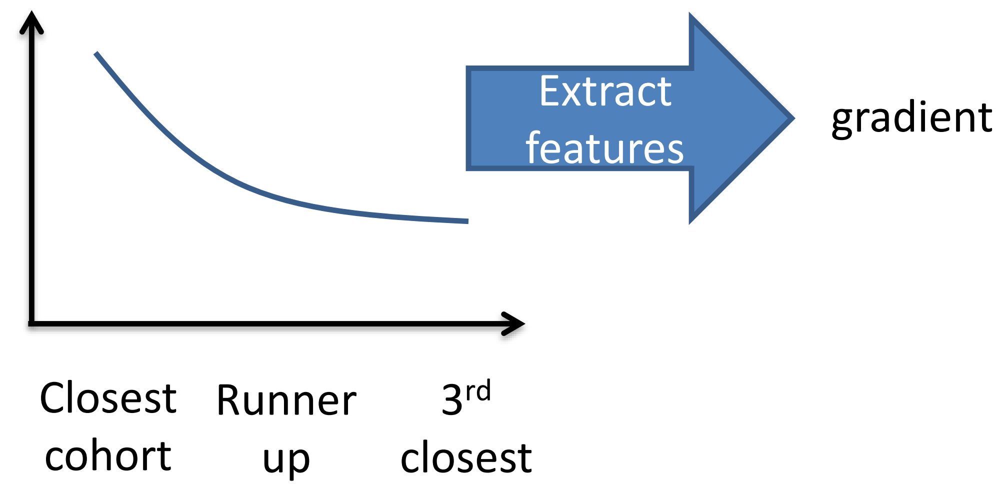
$s_{\mathtt{y}} = [y]$ Matching score

Cohort-based approach



Using cohort scores for normalisation





Representing the information sources

Biometric sample quality

$$s_{\mathtt{Q}} \equiv [q^{\mathtt{tmplt}}, q^{\mathtt{qry}}]$$

Quality of template
Quality of probe

Cohort information

$$s_{\mathtt{C}} \equiv [\mu^{\mathtt{C}}, \sigma^{\mathtt{C}}]$$

Mean of cohort scores
Std. dev. of cohort scores

Client-specific information

$$s_{\mathtt{US}} \equiv [\mu_1^{\mathtt{US}}, \mu_0^{\mathtt{US}}, \sigma_0^{\mathtt{US}}]$$

Mean of US impostor scores
Std. dev. of US impostor scores
Mean of US genuine scores

$s_{\mathtt{y}} = [y]$ Matching score

Experimental setting

- Biosecure DS2 database
- Fingerprint systems:
  - NIST fingerprint matcher”Bozorth3”
  - Quality assessment module “NFIQ”
- Face systems:
  - Omniperception SDK for face recognition and quality assessment
- Each subject provides 4 impressions per device and per finger
- Total of 4 impressions$\times$6 fingers $\times$ 2 devices=48 impressions per subject
- 333 subjects are available

Results

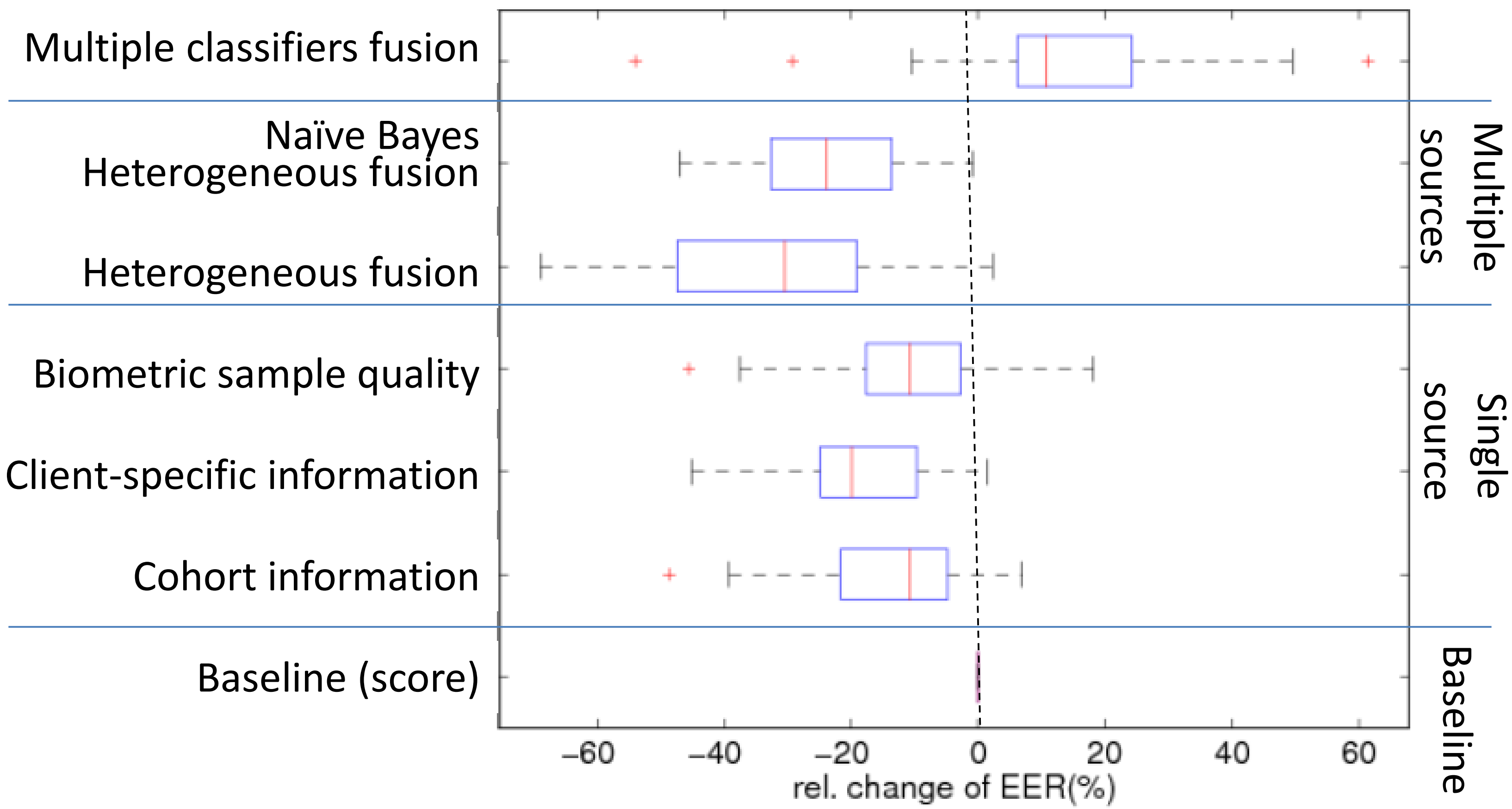
There are 30 experiments.
Do not try to read this.

Summarize the results, instead:

$$\text{RPG} = \frac{\text{metric}_{fusion} - \text{metric}_{baseline}}{\text{metric}_{baseline}} \times 100\%$$

Expe	base-line	individual sources: cohort	individual sources: us	individual sources: quality	hetero. fusion: joint	hetero. fusion: NB	homo. fusion
p1,fo1	3.06	1.85	2.45	2.31	**1.60**	* 1.62	2.16
p1,fo2	1.76	* 0.90	1.32	1.76	**0.81**	1.09	2.84
p1,fo3	4.33	3.34	2.72	3.96	* 2.46	**2.39**	5.70
p1,fo4	4.34	3.11	3.05	3.58	**2.03**	* 2.84	5.36
p1,fo5	3.39	2.93	2.53	2.42	**1.66**	* 2.24	3.74
p1,fo6	2.61	2.27	2.21	2.36	**1.45**	* 1.91	3.36
p2,fo1	2.52	2.26	1.90	1.37	**0.78**	1.51	* 1.16
p2,fo2	1.84	1.44	* 1.42	2.18	**0.97**	1.44	2.67
p2,fo3	1.89	1.30	1.34	1.94	**0.98**	* 1.27	2.35
p2,fo4	3.04	2.35	2.07	1.90	**1.55**	* 1.61	2.96
p2,fo5	3.43	3.26	**1.88**	2.26	* 1.92	2.32	4.21
p2,fo6	3.48	3.02	2.70	2.76	**2.13**	* 2.36	3.80
p1,ft1	10.58	9.51	9.56	10.44	**8.57**	* 9.31	12.20
p1,ft2	6.14	5.05	4.94	5.05	**4.23**	* 4.37	7.65
p1,ft3	8.87	* 7.58	8.50	8.68	**6.90**	7.83	9.98
p1,ft4	12.78	12.42	* 11.72	12.91	**11.14**	11.74	14.05
p1,ft5	6.71	6.13	5.64	5.72	**4.58**	* 5.09	6.43
p1,ft6	8.64	7.69	6.67	8.37	**5.87**	* 6.33	9.45
p2,ft1	8.63	9.23	**7.94**	8.67	8.84	* 8.56	9.15
p2,ft2	4.69	3.37	3.66	4.02	* 3.33	**3.30**	5.75
p2,ft3	8.68	8.51	**6.86**	8.23	7.30	* 7.09	9.85
p2,ft4	12.44	11.47	11.69	12.10	**10.11**	* 10.58	12.85
p2,ft5	7.06	5.59	6.41	6.14	**5.22**	* 5.37	6.32
p2,ft6	8.16	7.51	6.57	7.39	**6.42**	* 6.53	8.77
p1,fa	8.18	7.38	* 6.55	7.35	**6.53**	6.77	8.77
p2,fa	6.97	7.07	**6.17**	7.37	* 6.20	6.64	7.74
p1,fwf	3.73	3.70	* 3.10	3.30	**2.78**	3.22	5.01
p2,fwf	5.55	5.35	5.63	**4.90**	* 5.04	5.07	6.17
p1,fnf	2.57	2.55	2.35	**2.16**	2.35	* 2.34	3.85
p2,fnf	3.59	3.28	2.87	* 2.72	**2.52**	2.97	3.91

Preliminary results

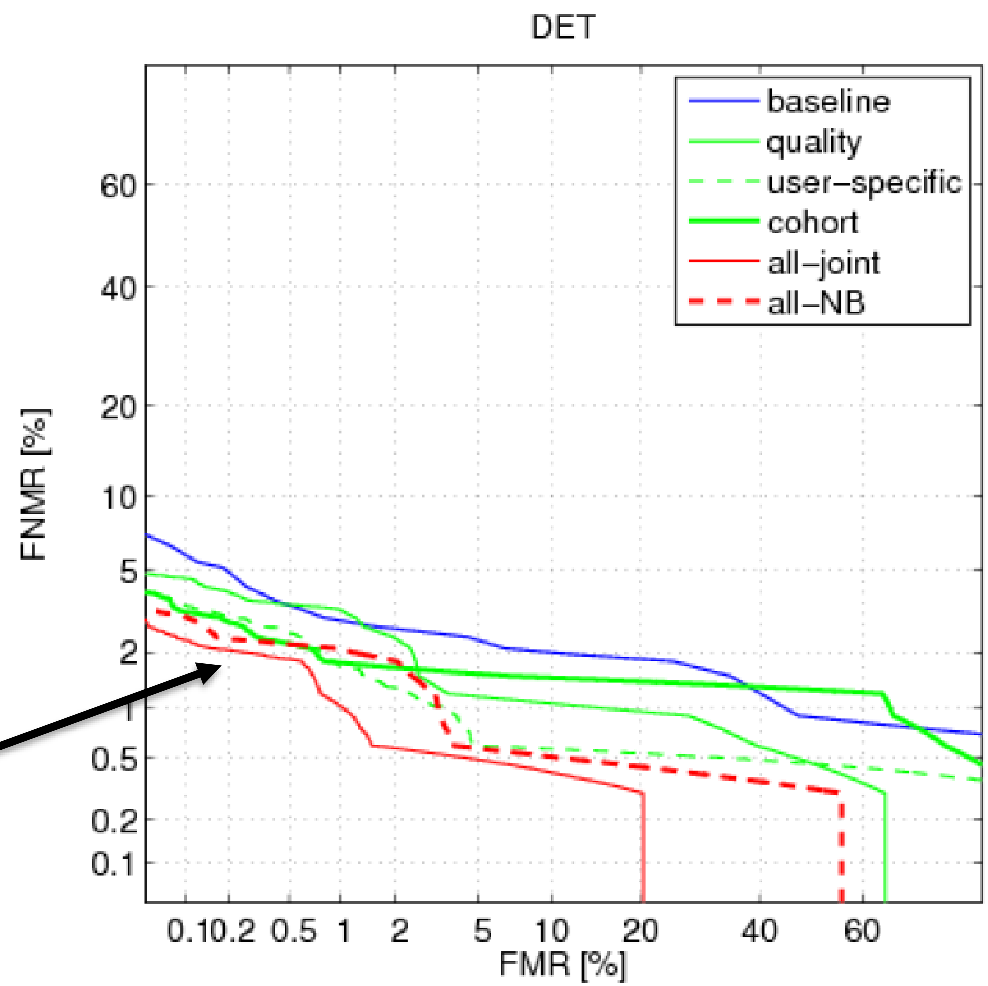


Setting

Biosecure database

Face modality

Zero-effort attacks only



Heterogeneous information fusion

Poh, Norman, Amin Merati, and Josef Kittler. "Heterogeneous information fusion: A novel fusion paradigm for biometric systems." *Biometrics (IJCB), 2011 International Joint Conference on*. IEEE, 2011.

Menu

TRICKS OF THE TRADE

Generalized logit transform

Biometric matcher outputs are not normally distributed

$$\boxed{y' = \log \frac{y-a}{b-y}}$$

Example:

y is the posterior probability of a genuine user given feature x, i.e., $y = P(G|x)$

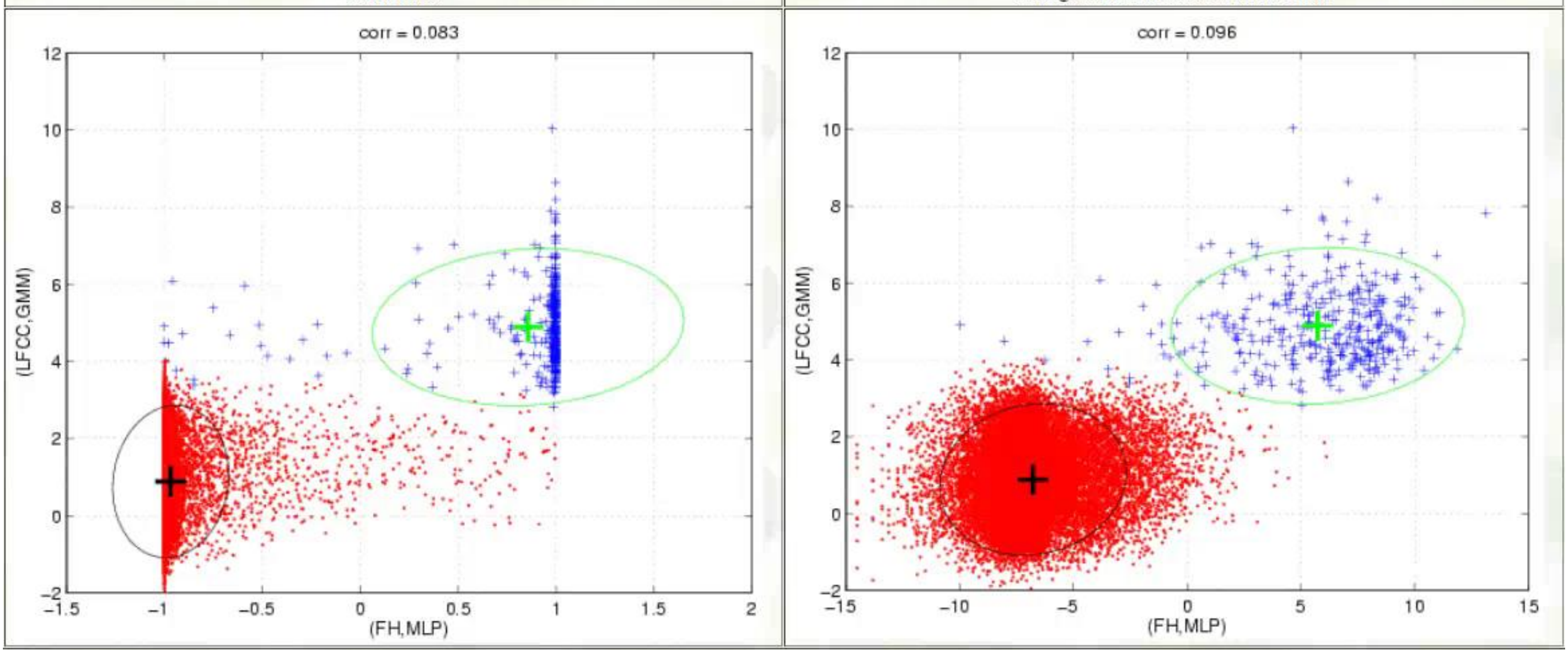
$$y = [a, b]$$

$$\begin{aligned} y^{llr} &= \log\left(\frac{y}{1-y}\right) = \log\left(\frac{P(\mathtt{G}|x)}{P(\mathtt{I}|x)}\right) \\ &= \log\left(\frac{p(x|\mathtt{G})}{p(x|\mathtt{I})}\right) + \log\left(\frac{P(\mathtt{G})}{P(\mathtt{I})}\right) \\ &= \underbrace{\log\left(\frac{p(x|\mathtt{G})}{p(x|I)}\right)} + const, \end{aligned}$$

The effect of logit transform

Original scores
Because some classifiers are MLPs with tanh function, these scores are not normally distributed.

Probabilistic-inversed scores
Contrary to the left, these scores fit better using a single Gaussian distribution. Note the average correlation values in the title.



Menu