Secure Identity Verification

Anthony Vetro

Mitsubishi Electric Research Labs
Cambridge, Massachusetts, USA
avetro@merl.com

IST – Lisbon, Portugal
November 15, 2010

Biometrics for Identity Verification

Biometrics is the science and technology of measuring and statistically analyzing biological data.

BIOMETRIC	FINGERPRINT	FACE	HAND GEOMETRY	IRIS	VOICE
					
Barriers to universality	Worn ridges; hand or finger impairment	None	Hand impairment	Visual impairment	Speech impairment
Distinctiveness	High	Low	Medium	High	Low
Permanence	High	Medium	Medium	High	Low
Collectibility	Medium	High	High	Medium	Medium
Performance	High	Low	Medium	High	Low
Acceptability	Medium	High	Medium	Low	High
Potential for circumvention	Low	High	Medium	Low	High

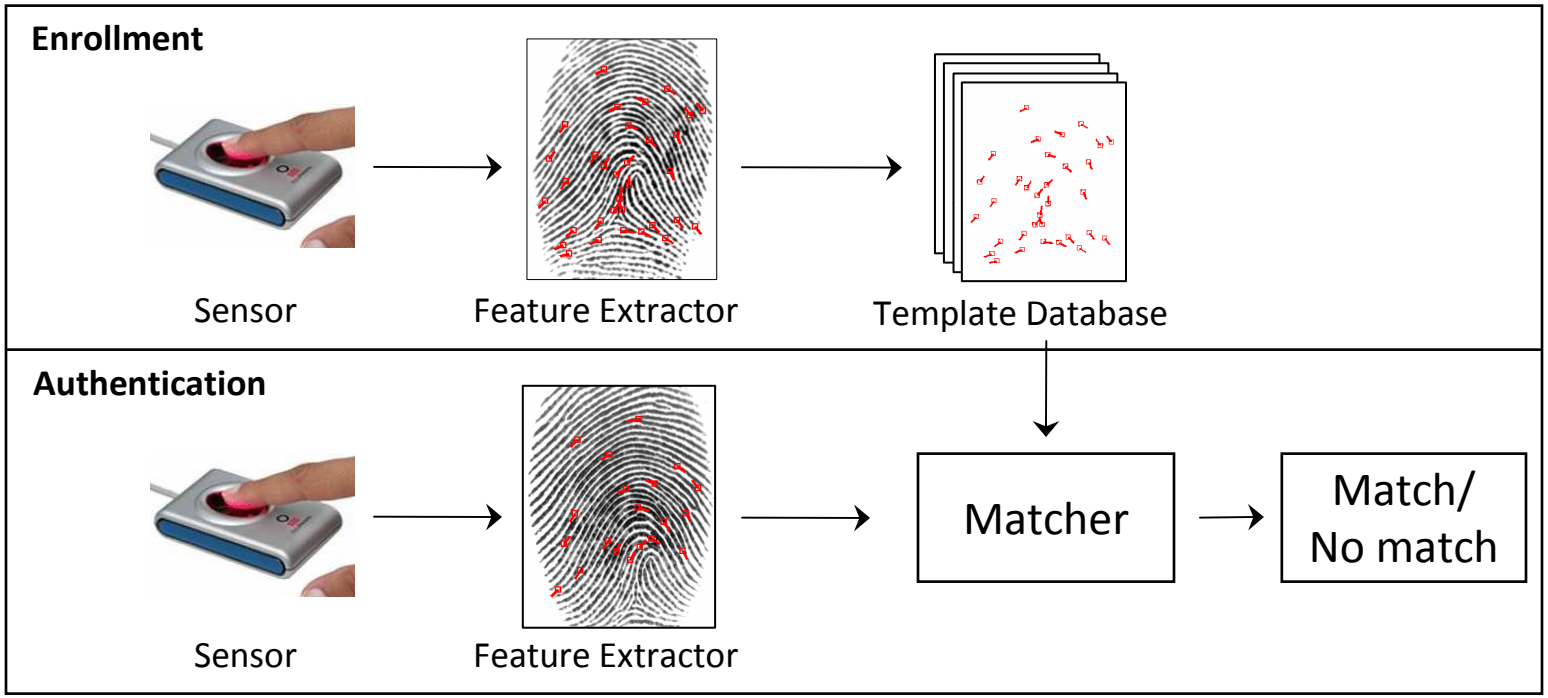
S. Prabhakar, S. Pankanti, and A. K. Jain, "Biometric Recognition: Security and Privacy Concerns", IEEE SECURITY & PRIVACY, 2003.

- Universality : do all people have it ?
- Distinctiveness : can people be distinguished based on an identifier ?
- Permanence : how permanent is the identifier ?
- Collectability : how well can the identifier be captured and quantified ?
- Performance : speed and accuracy
- Acceptability : willingness of the people to use
- Circumvention : foolproof

Biometric Matching System

Four main components:

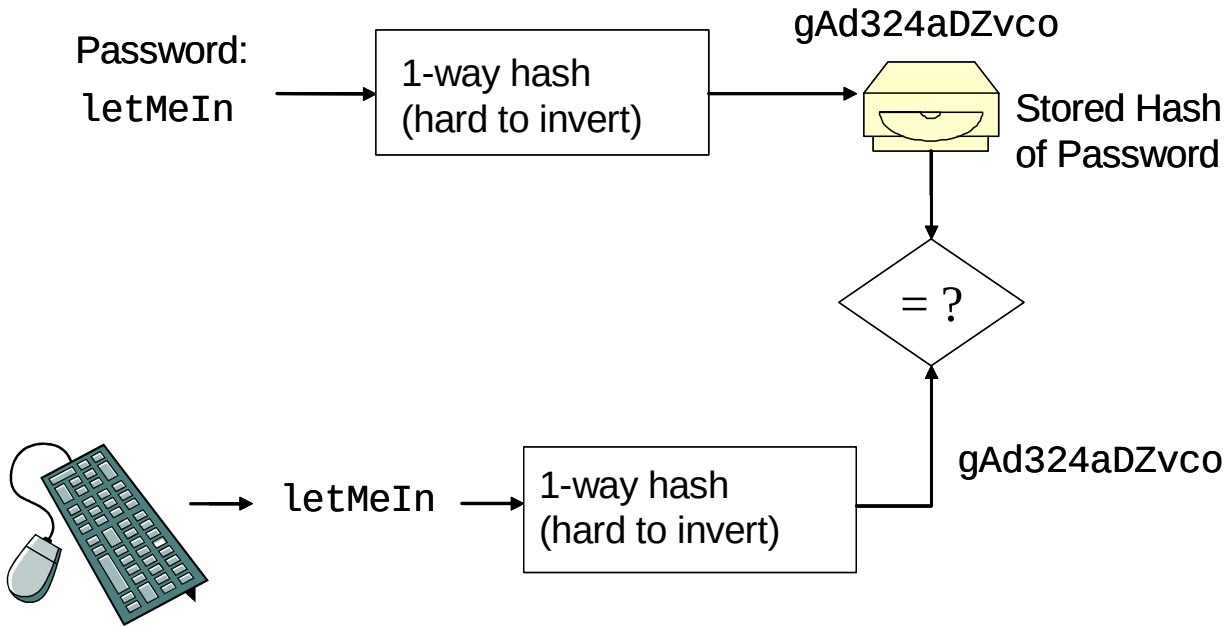
sensor, feature extractor, template database, and matcher



Question: How can we design a secure identity verification system?

Securing Passwords

- Do not store passwords as clear text - store hash of password instead
- If computer stolen / broken into, password remains secure
- Enter identical password to gain access



Challenge for Biometrics

- Biometric data is noisy
  - Each feature extraction results in different but similar data
    Reasons: sensor, feature extraction algorithm, environment
  - Extremely difficult to model both the data and noise
  - Conventional hash functions not applicable

Four impressions
from the same finger

- Traditional encryption schemes won't help much either
  - Clear template is needed for matching
  - Where to store the key?

Two Approaches Considered

ECC-Based Systems

- Extract error correcting information from the biometric (aka helper data)
- Authentication performed by recovery of external key or original biometric
- Difficult to recover biometric from stored data; information-theoretic security analysis is possible

attempt to cope with noise in data

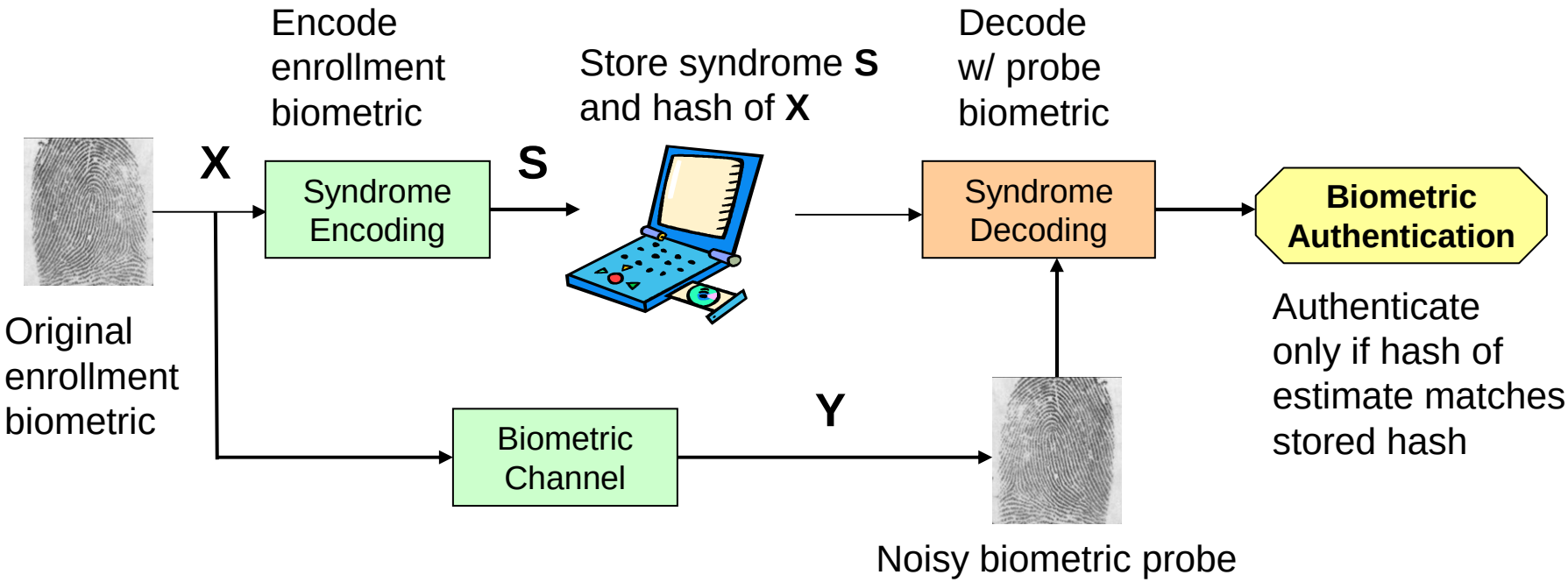
Encryption-Based Systems

- Apply homomorphic encryption to the biometric
- Authentication performed on encrypted data through a protocol that does not reveal user biometric
- Computational security as offered by cryptographic primitives

utilize properties of homomorphic functions to maintain security and data privacy

ECC-Based Systems

Modeled as a Slepian-Wolf system

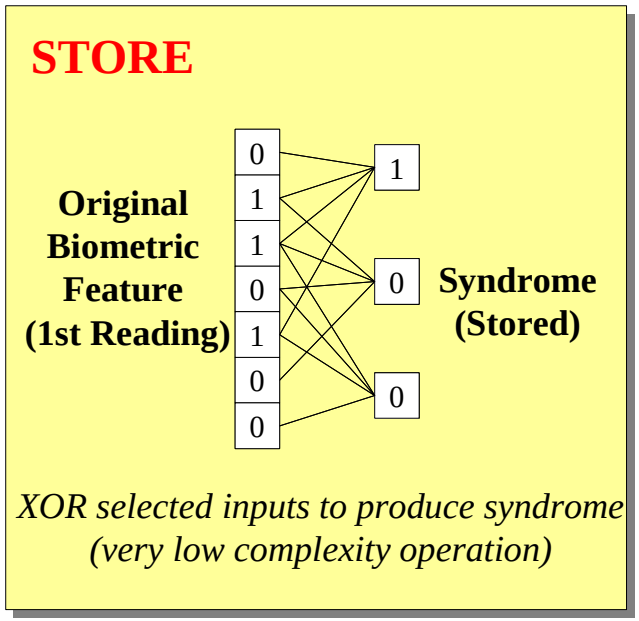


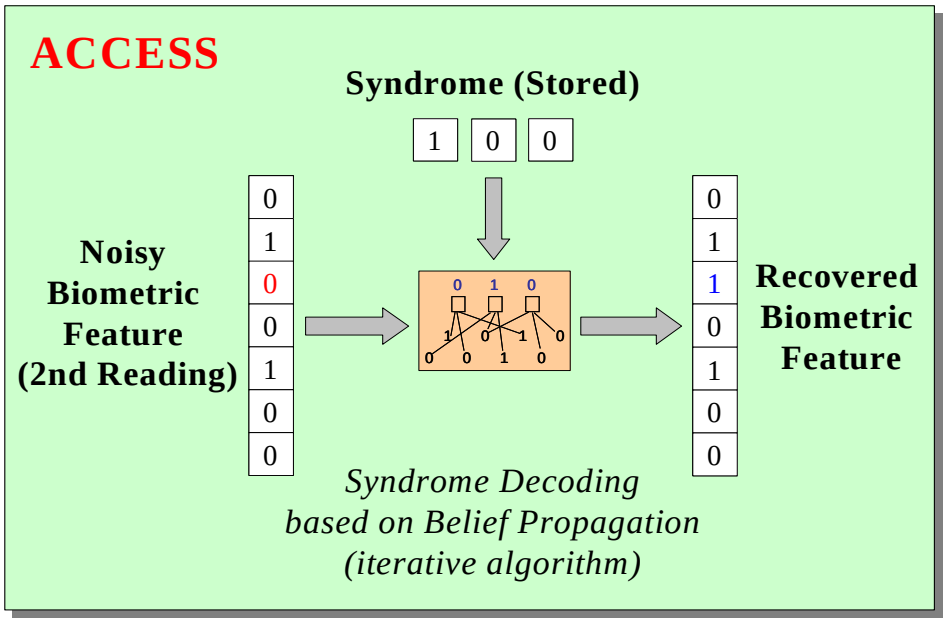
Encode into syndrome S

- **S** cannot be uncompressed by itself & is therefore secure
- In combination with a noisy second reading **Y** the original **X** can be recovered using a Slepian-Wolf decoder
- Compare hash of estimate with stored hash to permit access

[Martinian, et al., Allerton 2005] [Draper, et al., ICASSP 2007]

Overview: Syndrome encoding / SW decoding



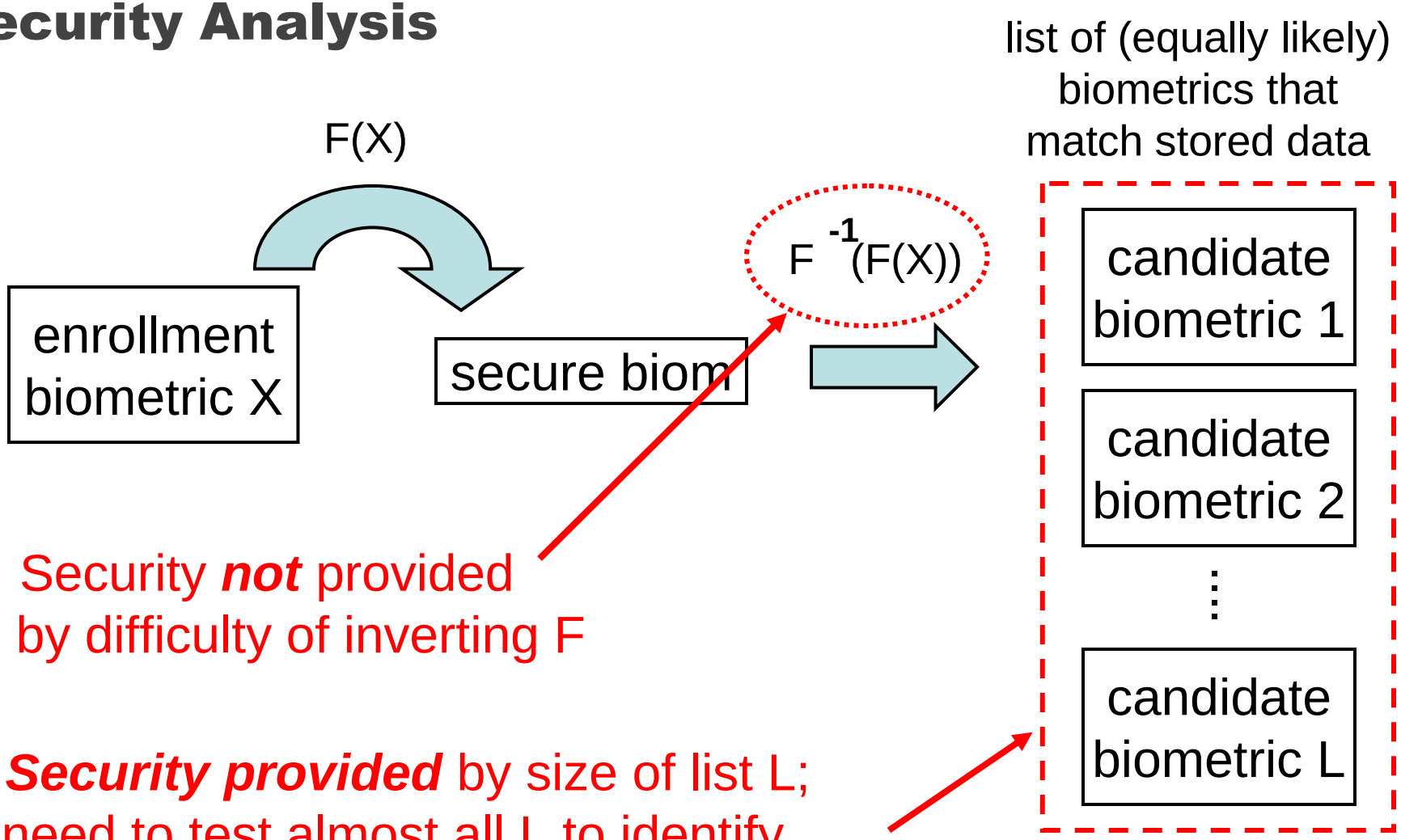


Security = number "missing" bits
= original bits – syndrome bits
Translates into number guesses
to identify original biometric w.h.p.

Robustness = false-rejection rate
Robustness to variations in biometric readings
achieved by syndrome decoding process
(syndrome + noisy biometric => original biometric)

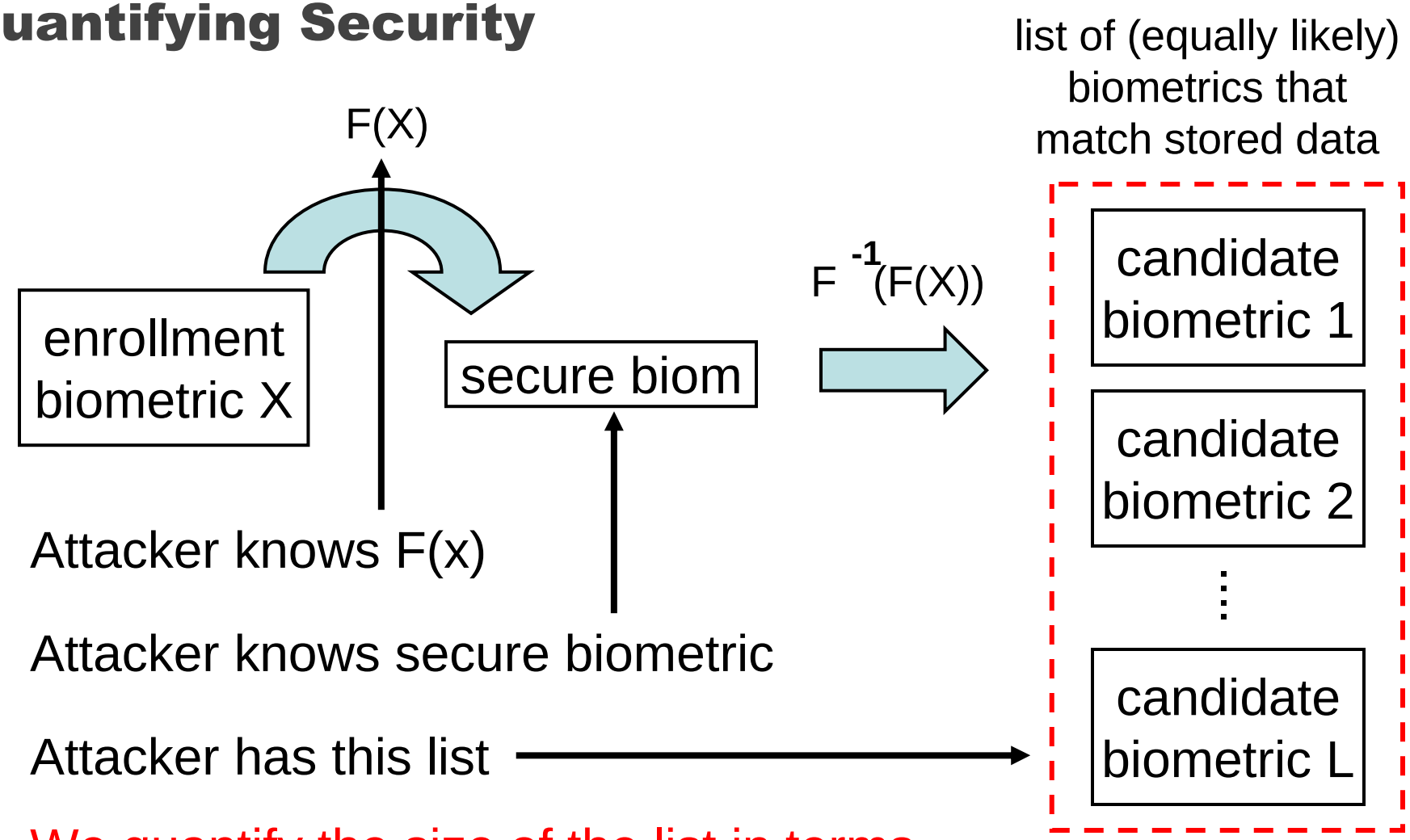
Fewer syndrome bits = greater security, but less robustness

Security Analysis



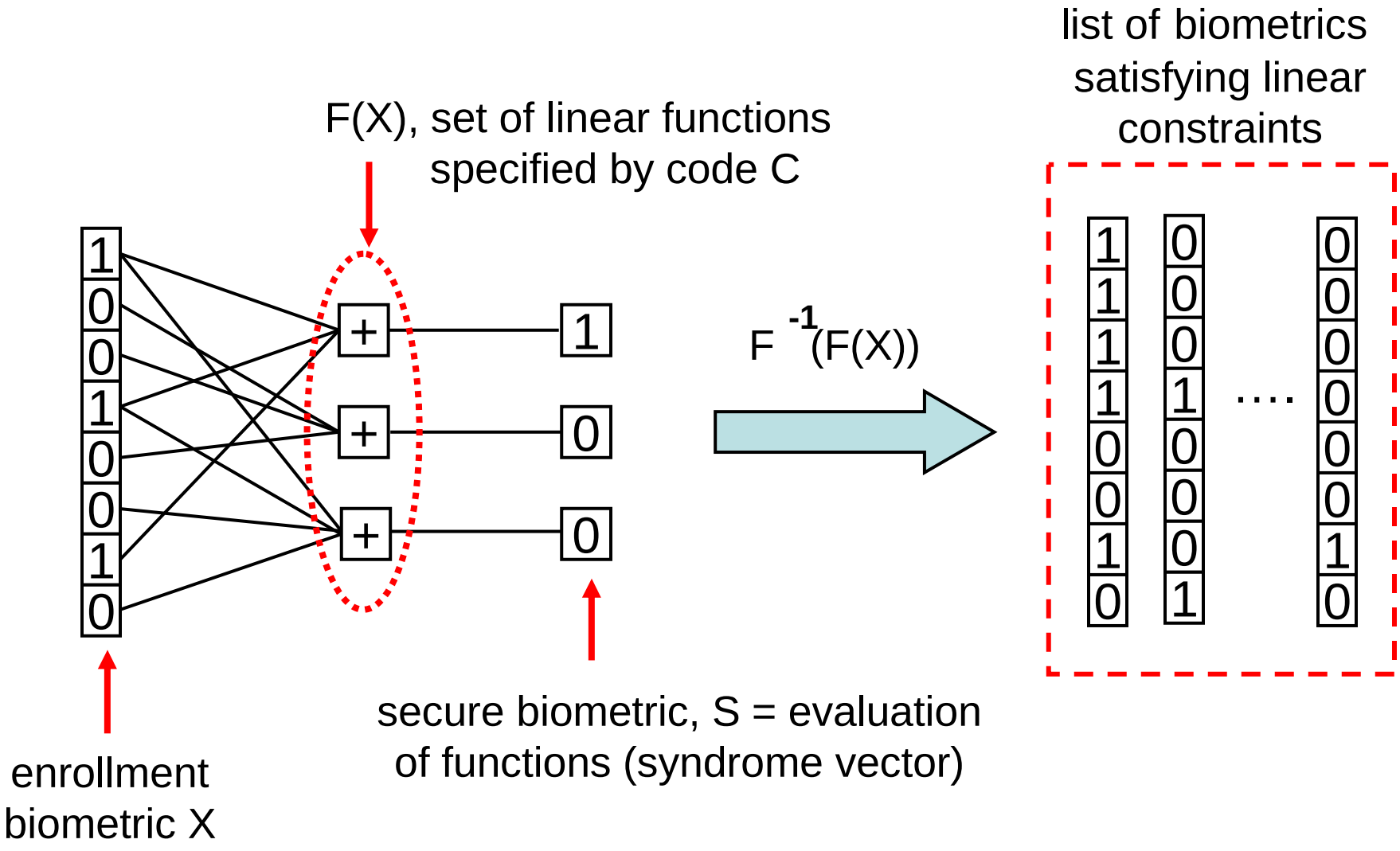
- Security ***not*** provided by difficulty of inverting F
- ***Security provided*** by size of list L; need to test almost all L to identify enrollment biometric

Quantifying Security



- Attacker knows F(x)
- Attacker knows secure biometric
- Attacker has this list
- We quantify the size of the list in terms of measurable characteristics of F

Security of Syndrome-Based System



Security/Robustness evaluation: information-theoretic analysis

X = biometric feature (length n binary vector)
S = syndrome (length nR_{SW} binary vector, R_{SW} is compression rate)
Y = biometric probe (length n binary vector)

Security corresponds to number of missing bits

- Guess from typical sequences in bin
- $2^{H(X|S)}$ guesses required for successful attack w.h.p.
- $R_{sec} = H(X|S) = H(X,S) – H(S) = H(X) – H(S) = H(X) – nR_{SW}$
- Lower values of R_{SW} → higher security

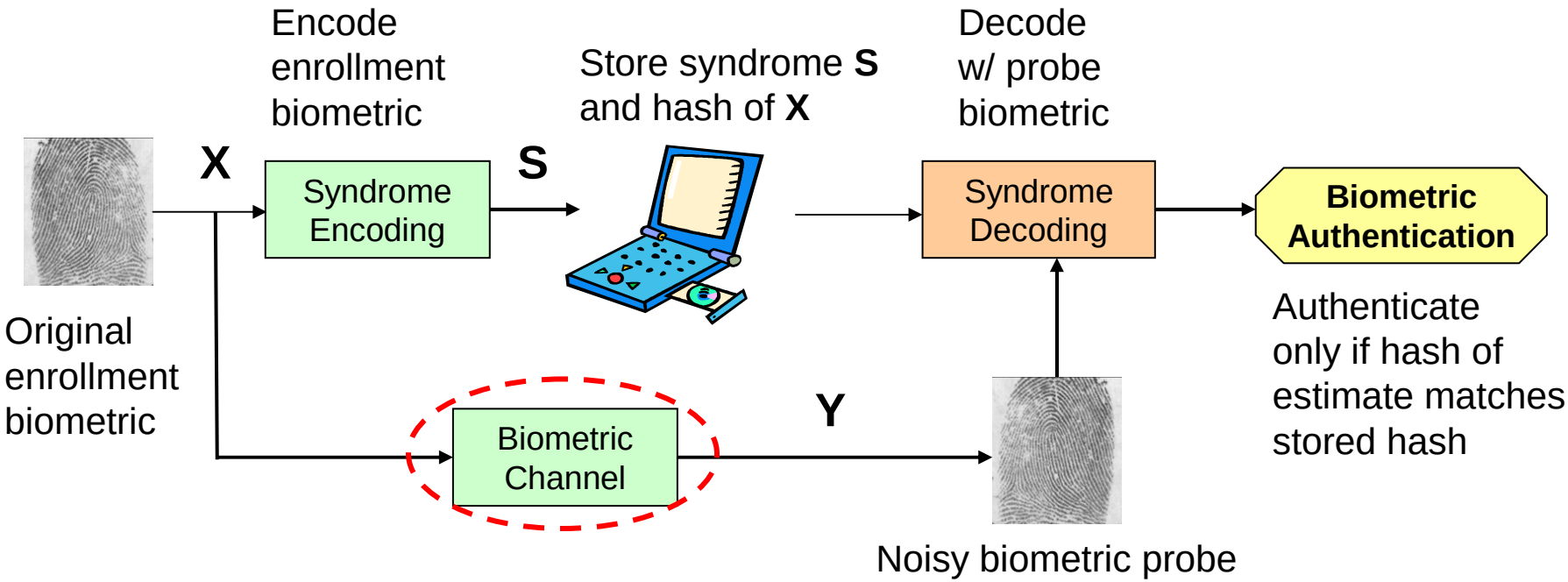
Robustness determined by Slepian-Wolf error exponent

- Pr[false rejection] = $\exp\{ -n E_{SW}(R_{SW}) \}$
- Lower values of R_{SW} → higher false-rejection-rate

Security/Robustness range

- $R_{SW} < (1/n) H(X)$ needed for positive information security
- $R_{SW} > (1/n) H(X|Y)$ needed for positive error exponent

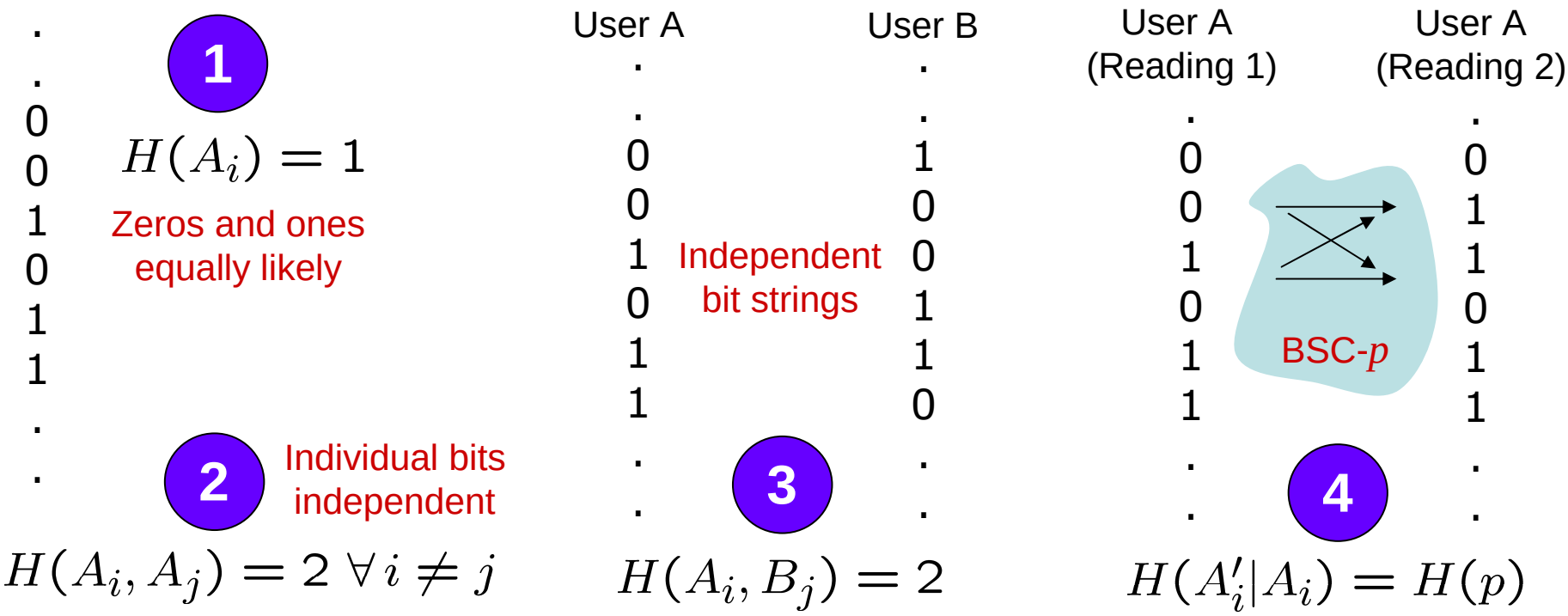
System Design



- Key issue: what does the biometric channel look like?
  - Depends heavily on the input X

- Our approach: transform the input to a binary feature vector so that the biometrics channel looks like a BSC

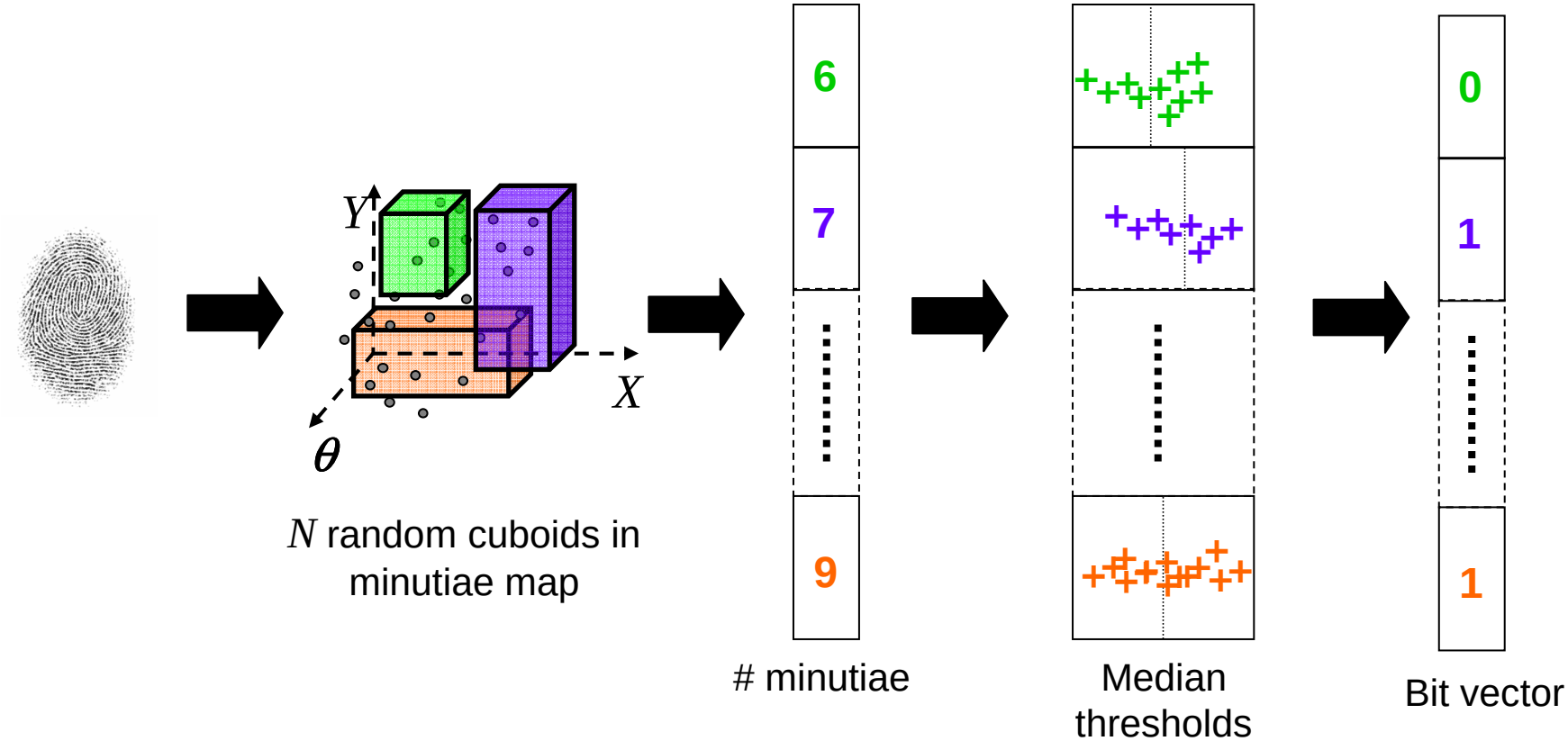
Desired Properties of Extracted Binary Features



This method provides positive information theoretic secrecy

[Sutcu, et al, ISIT 2008]

Feature Extraction (based on fingerprints)

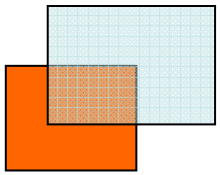


Each cuboid contributes a 0 or 1 bit to the feature vector, if it contains less or more minutia points than the median

[Sutcu, et al, CVPR 2008]

Performance Improvements

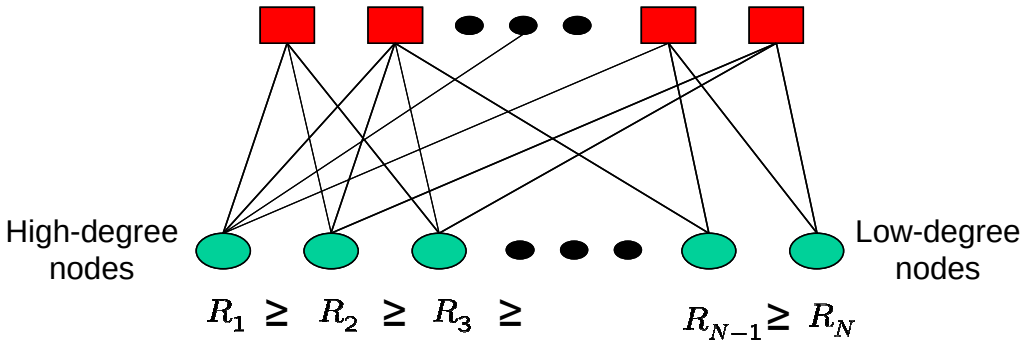
- Minimize Cuboids Overlap
  - Large overlap → similar bits
  - → easy for attacker to guess



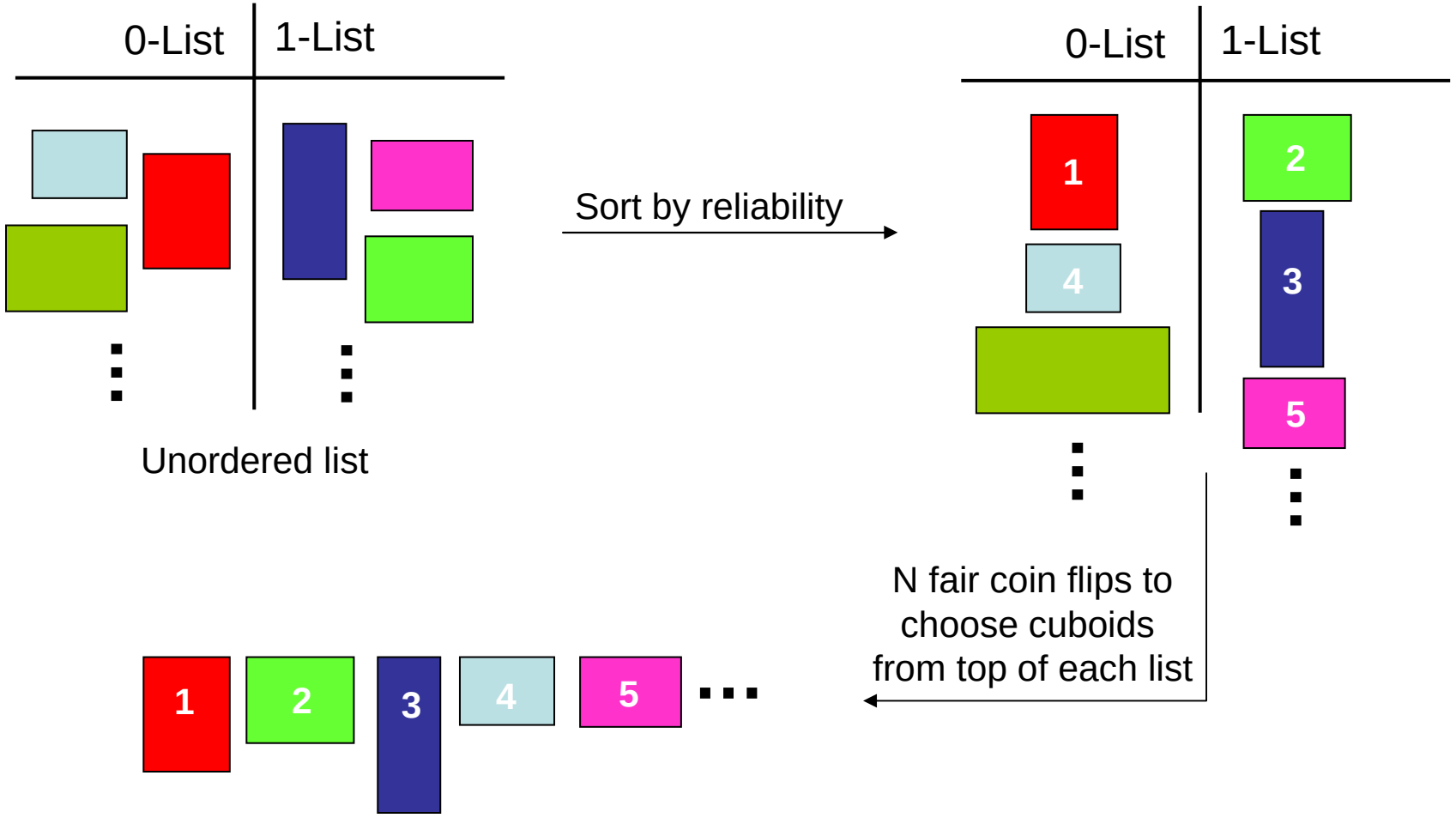
$$O_{i,j} = \frac{V_i \bigcap j}{V_i \bigcup j}$$

- Leverage Bit Reliability
  - Differ depending on where the biometric bits are derived from
  - Reliabilities could even be user-specific
  - Possible to leverage reliabilities in [Wang, et al, WIFS 2009]
    - Initialization of LDPC decoding
    - Degree distribution for irregular LDPC

$$R_i = \left| \log\left(\frac{1-p_i}{p_i} \right) \right| = |\mathrm{LLR}_i|$$

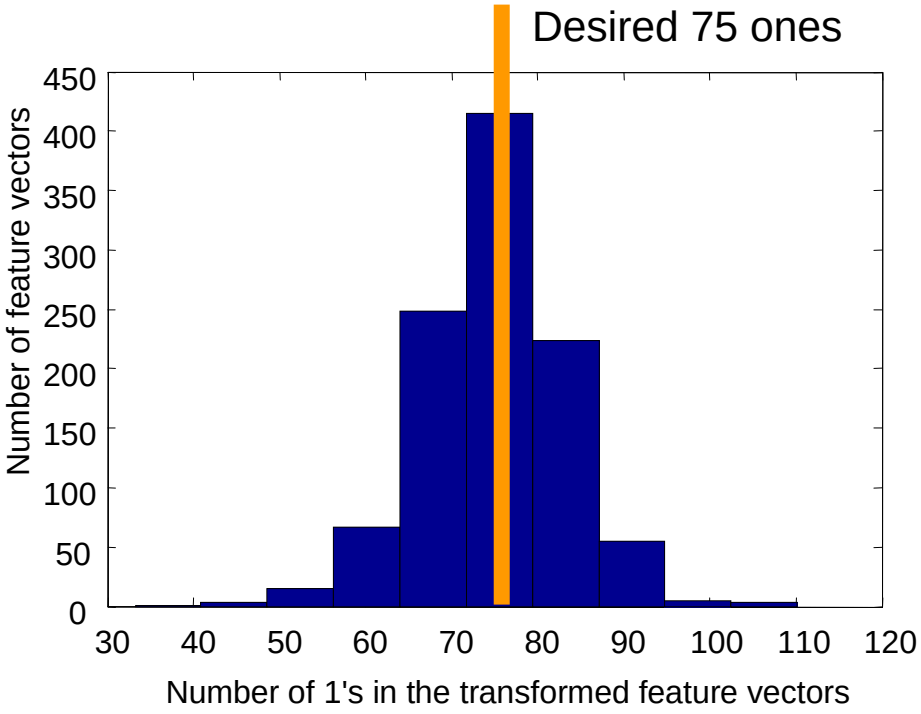


User-Specific Reliable Cuboids

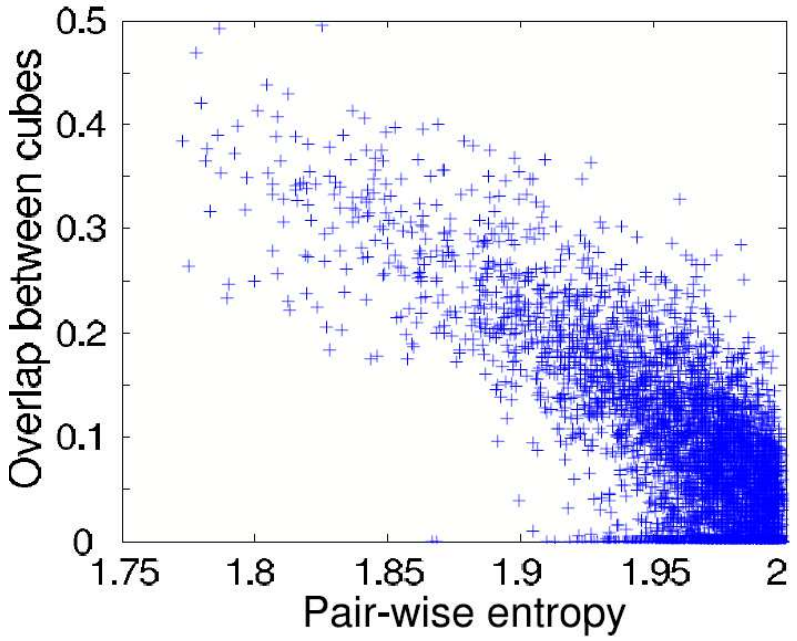


To what extent are the 4 desired properties are satisfied ?

Zeros & Ones Equally Likely

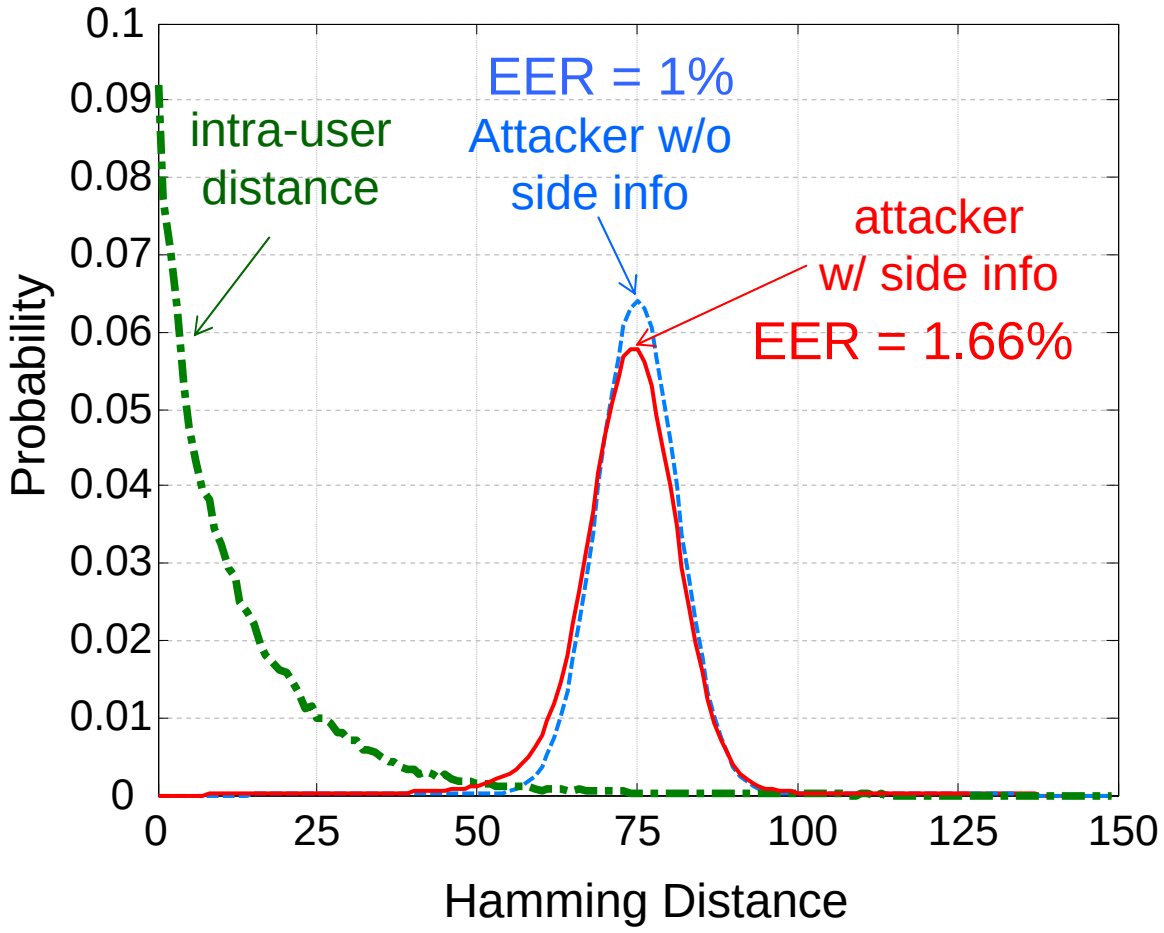


Individual Bits Independent



Proprietary database of 1035 users, 15 pre-aligned samples per user, 150 cuboids

Intra-user & Inter-user Distance



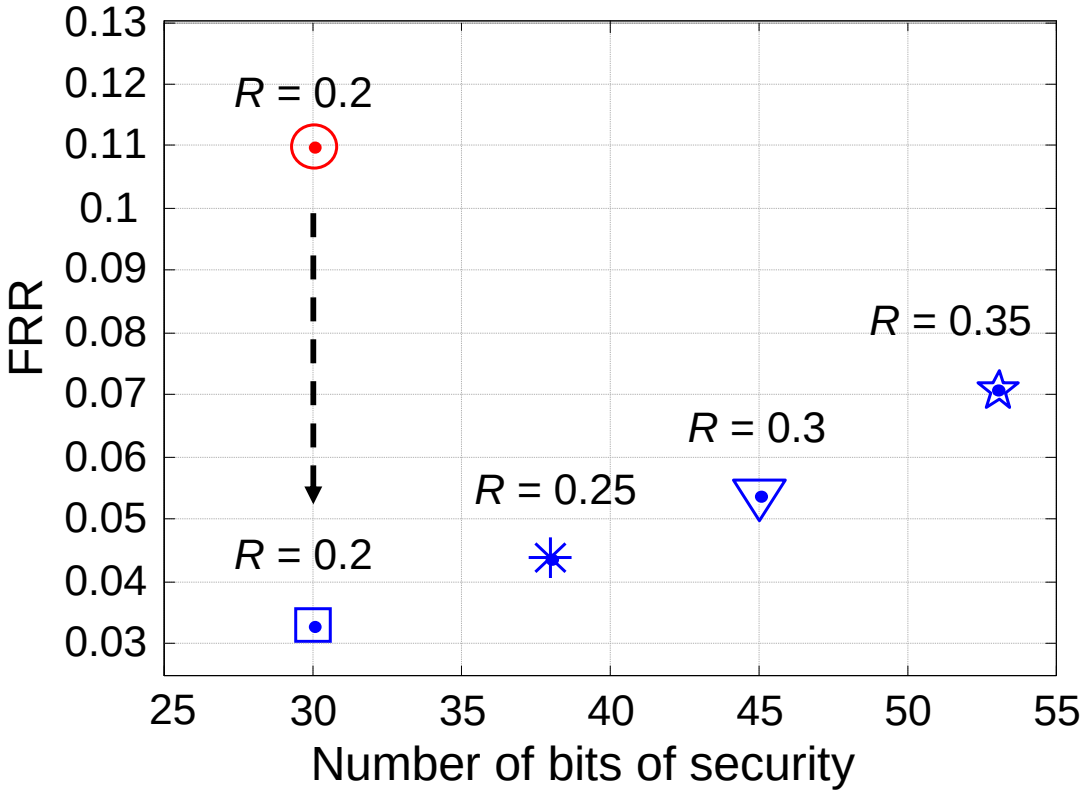
* EER: equal error rate [false accept = false reject]

Overall Security & Robustness (Syndrome Code Rate = 0.2)

Scheme	FRR	FAR	SAR*
Unordered Bits Equal LLR	11%	0.0003%	0.012%
Unordered Bits Unequal LLR	9.9%	0.0002%	0.044%
Reordered Bits Unequal LLR	3.7%	0.0001%	0.043%
Reordered Bits Unequal LLR Shuffled BP	3.3%	0.00016%	0.050%

* Successful Attack Rate (SAR) = Pr{ Successful imposter login with side-info } ≥ FAR

Bits of Security



- # bits of security = # bits the attacker must guess
  ≈ # feature bits - # syndrome bits
- Can trade off FRR for # bits of security

Beyond Minutiae Counts

- Expanded feature set could enable better accuracy and increased security

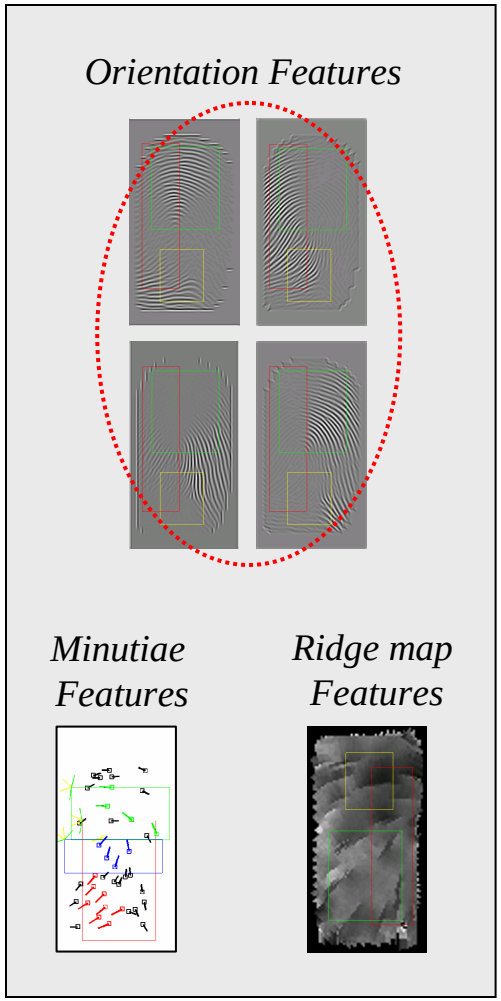
- Need uncorrelated and discriminable features
  - Correlated features lead to redundancy; loss in security so must eliminate pair-wise correlations
  - Discriminability of ith bit corresponding to the jth user is given as

$$d_i^j = I_i^j - G_i^j$$

I_i^j : Impostor bit-flip probability

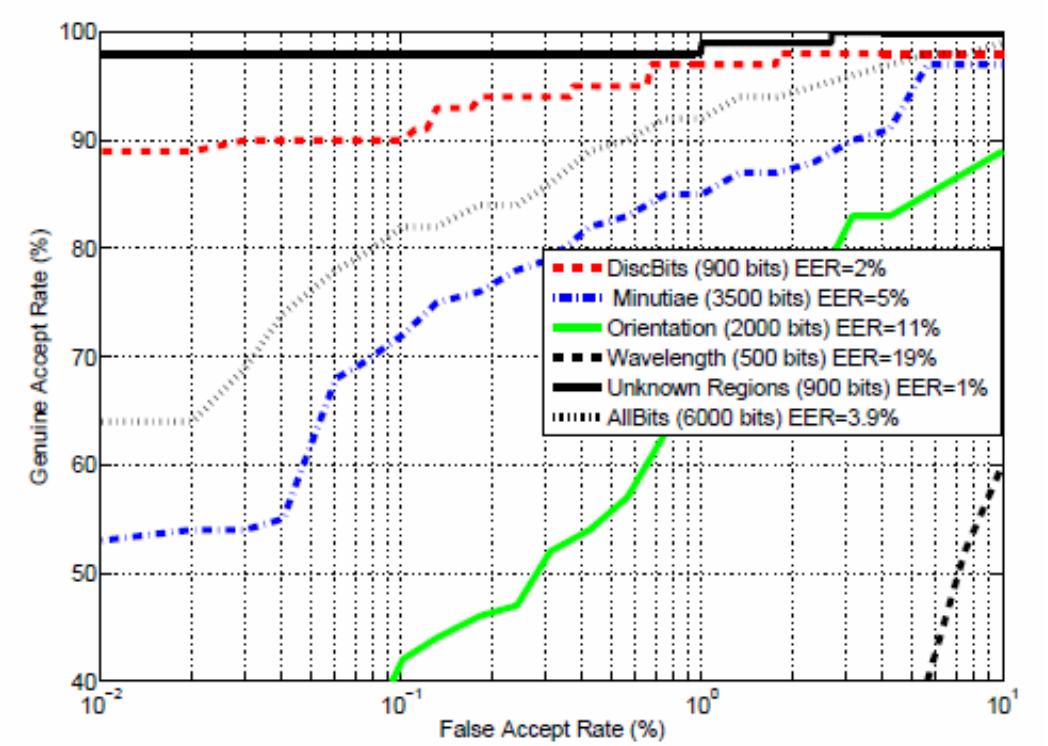
G_i^j : Genuine bit-flip probability

Bits having highest discriminability are selected as final features



[Nagar, et al, SPIE 2010]

Results with Expanded Fingerprint Features



- FVC2002 Database-2
  - 100 fingers, 8 impressions per finger
  - One impression is enrolled, six used for training and one for testing
- Consider seven minutiae features, four ridge orientation features and ridge wavelength

Summary

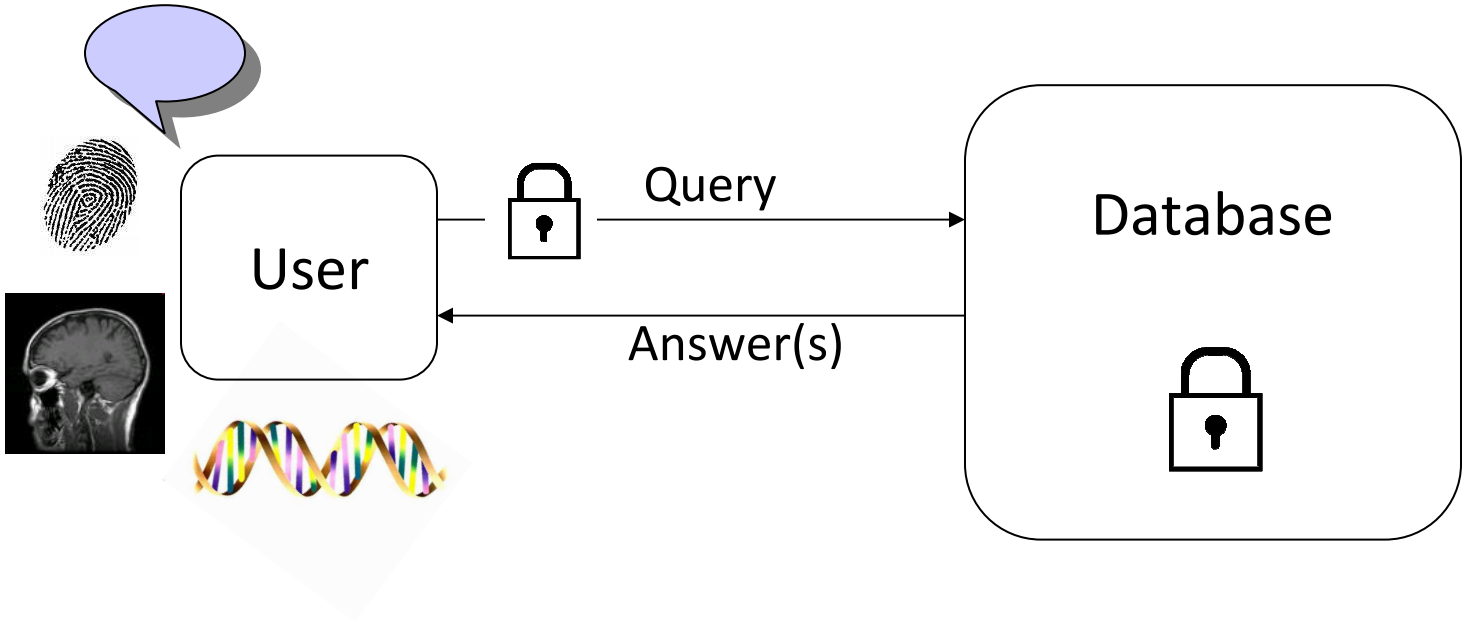
- ECC techniques can be utilized to cope with noise in secure verification of biometric data

- Important points to note
  - Biometric feature vectors should be designed according to the ECC to achieve a good security-robustness tradeoff
  - Possible to leverage reliability of extracted feature bits in code design and decoding process
  - Extraction of bits from ridge orientation and ridge wavelength in addition to minutiae improves matching performance

- Drawback: attacker can eavesdrop on reconstructed biometric and verification result

Encryption-Based Systems

Private Information Retrieval



- Keyword search on encrypted documents
- Privacy-preserving medical analysis
- Private biometric authentication

Oblivious Transfer (OT)

- Input: Bob has $\mathbf{z} = z_1, z_2, \ldots, z_N$
- Output: Alice gets z_k
- Requirements
  - Alice will know nothing about Bob's other elements
  - Bob will not know k
- Example:
  - Alice has $x = 5$, Bob has $y = 7$
  - Alice wants to compute $(x - y)^2$ where $1 \leq x, y \leq 10$
  - Bob keeps a list of $(x - 7)^2$ i.e., $z = [36,25,16,9,4,1,0,1,4,9]$
  - Alice wants z_5 w/o Bob's knowledge

1 out of 10 Oblivious Transfer

- Alice ← 10 public keys $K_1, K_2, \ldots, K_{10}$ — Bob

- Alice → $K_5(E)$ → Bob

- Bob tries to decrypt $K_5(E)$ using all 10 decryption keys to obtain $D_1[K_5(E)]$, … $D_2[K_5(E)]$, …, $D_{10}[K_5(E)]$. The 5th entry is Alice's key, others are garbage. $G_1, G_2, \ldots, G_5 = E, \ldots, G_{10}$

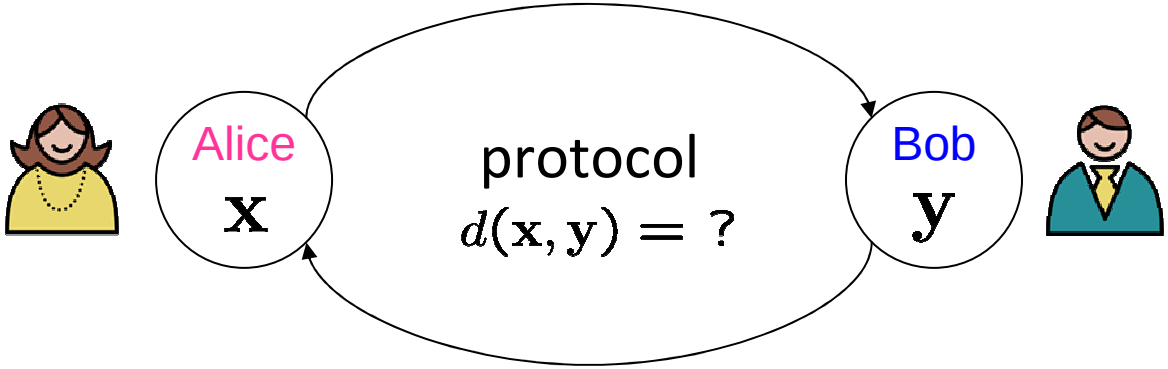
- Alice ← $G_1(z_1), G_2(z_2), \ldots, E(z_5), \ldots, G_{10}(z_{10})$ — Bob

- Alice decrypts the 5th entry. She can't decrypt anything else.

Practical Issues with OT

- Generality is good, but protocol overhead becomes heavy even for very simple circuits (esp. with large values and long vectors)
  - O(N) encrypted transmissions
  - For naïve OT, # decryptions required = O(N)
- With homomorphic encryption, possible to reduce encrypted transmissions and decryptions drastically
- Traditional uses of homomorphic encryption
  - Secure voting [Adida, Rivest, ‘06]
  - Secure auctions and bidding [Damgard, ‘09]

Secure Distance Computation



$$\mathbf{x} = x_1, x_2, \ldots, x_n$$
$$\mathbf{y} = y_1, y_2, \cdots, y_n$$

- Alice and Bob want to evaluate $d(\mathbf{x},\mathbf{y})$ without sharing $\mathbf{x}$ and $\mathbf{y}$
- Need protocols with low transmission and computation overhead
- Focus of this talk: consider $d(\mathbf{x},\mathbf{y})$ as Hamming, L2 or L1 distance

Additively Homomorphic Functions

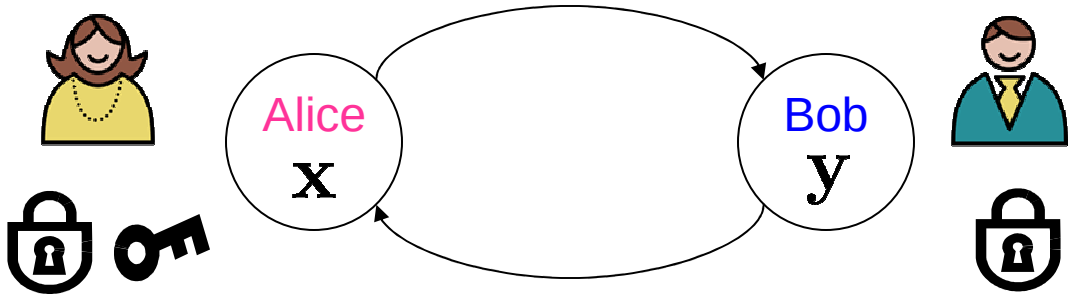
$$\xi(m_1 + m_2) = \xi(m_1)\xi(m_2)$$

$$\xi(km_1) = \xi(m_1)^k$$

Additively homomorphic schemes in the literature:
[Paillier,`99; Benaloh,`86; Damgard–Jurik,`01]

(Our protocol will work with any of them)

Squared Distance Protocol (Setup)



- $s(\mathbf{x},\mathbf{y}) = \sum (x_k - y_k)^2 = \sum x_k^2 + y_k^2 - 2x_k y_k = A + B + C$
- $A = \sum x_k^2, B = \sum y_k^2, \; C = -2\sum x_k y_k$
- Alice knows A, Bob knows B

Protocol

[Rane, et al., ICIP 2009]

1. Alice $\xrightarrow{\xi(x_k) \text{ for all } k}$ Bob

2. Bob: $[\xi(x_k)]^{-2y_k} = \xi(-2x_k y_k)$ for all k

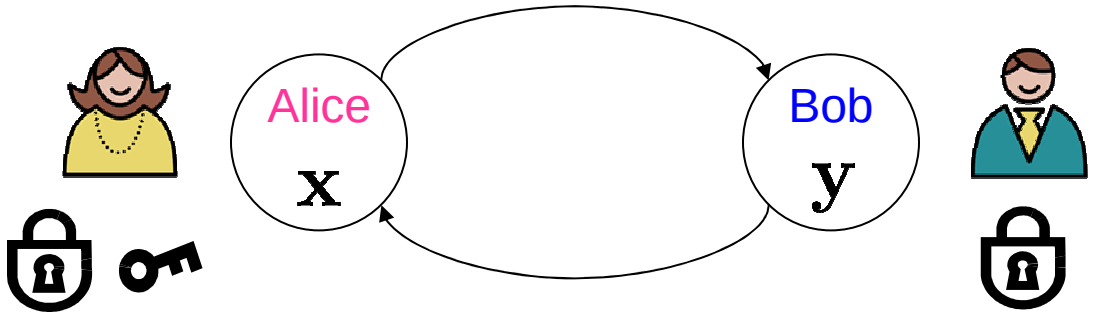
3. Bob: $\prod_k \xi(-2x_k y_k) = \xi(-2\sum_k x_k y_k) = \xi(C)$

4. Bob: $B = \sum y_k^2,\ \xi(B)\xi(C) = \xi(B+C)$

5. Alice $\xleftarrow{\xi(B+C)}$ Bob

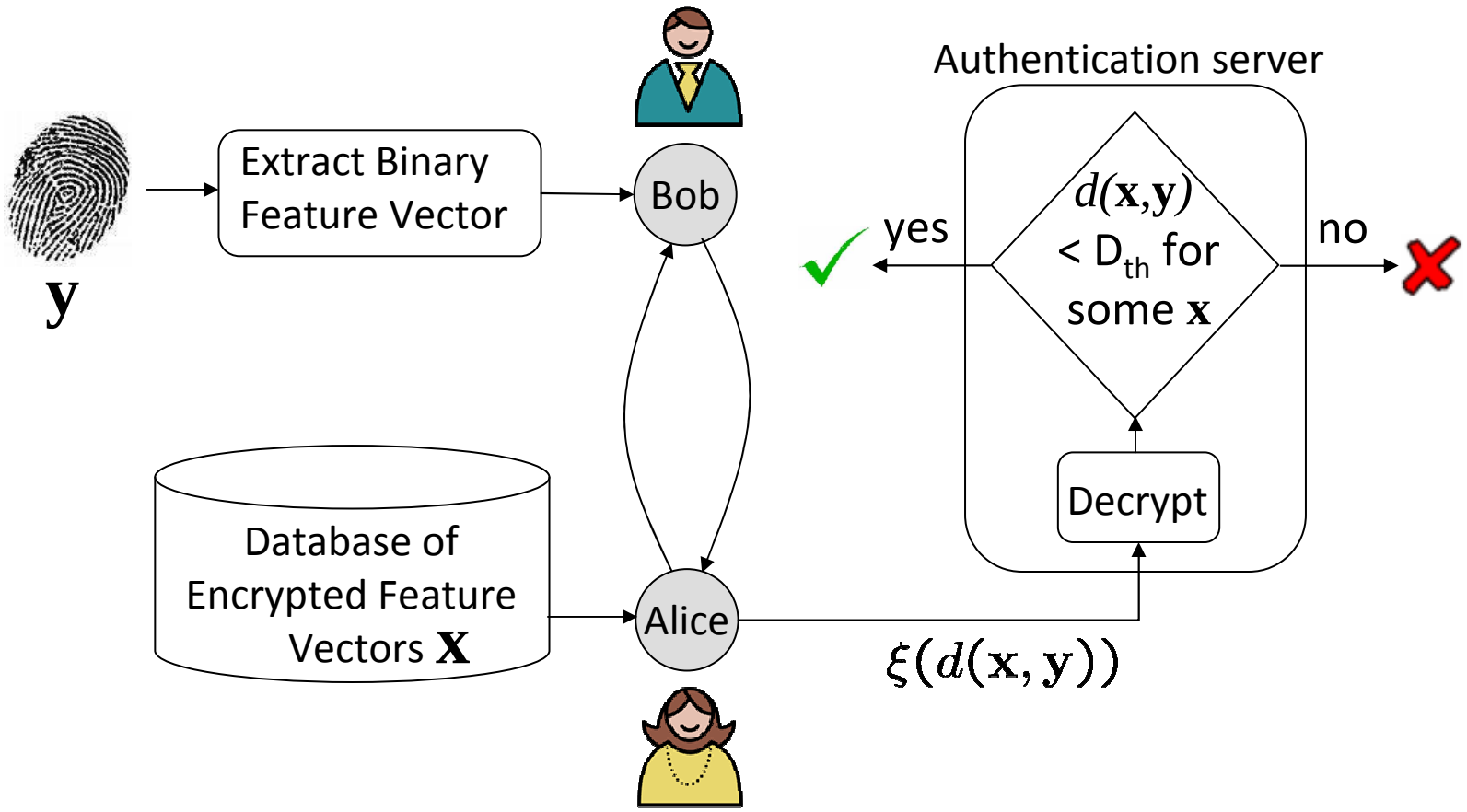
6. Alice: $A = \sum x_k^2,\ \xi(A)\xi(B+C) = \xi(A+B+C)$
$= \xi(s(\mathbf{x}, \mathbf{y}))$

Privacy & Cost

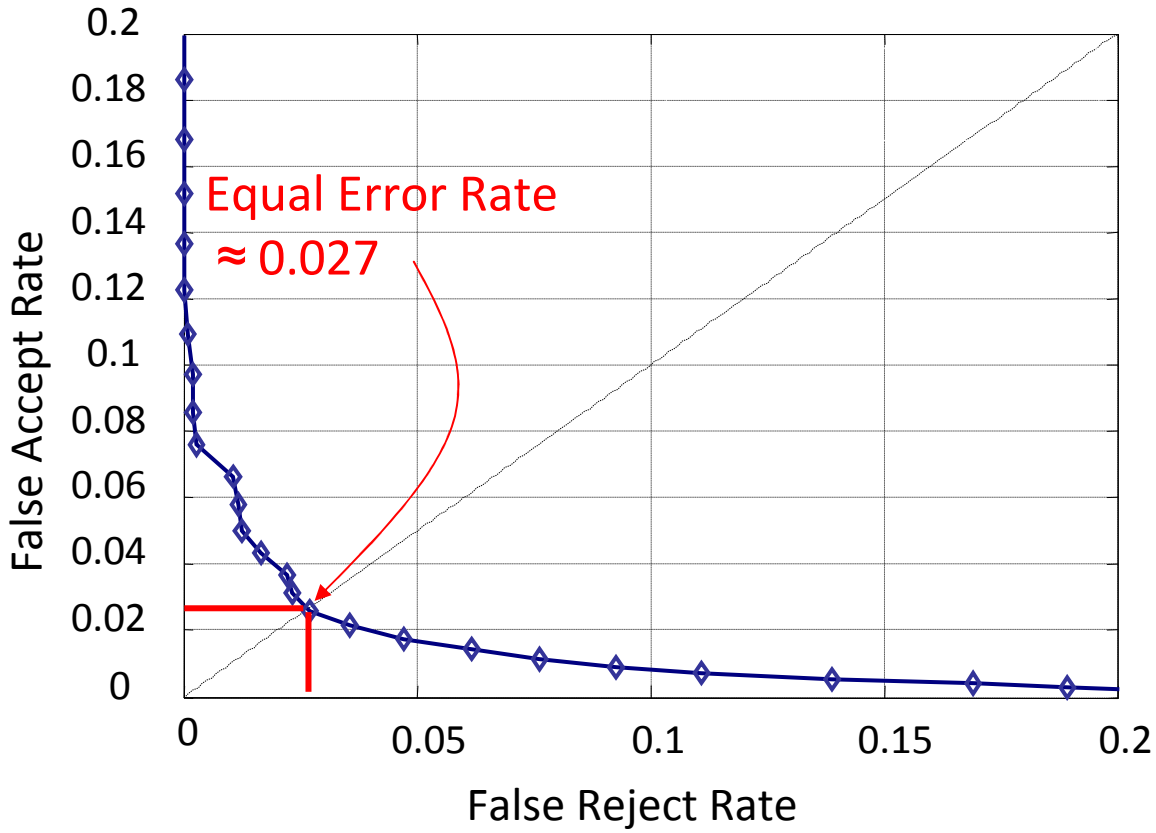


- Bob operates only on encrypted $x_1, x_2, ..., x_N$
- Alice can decrypt $d(\mathbf{x},\mathbf{y})$ and try to obtain $y_1, y_2, ... y_N$
  - No privacy for $N = 1$
  - Privacy for $N \geq 2$
- Alice: O(N) encryptions, 1 multiplication
- Bob: 1 encryption, O(N) exponentiations, O(N) multiplications in encrypted domain

Anonymous Fingerprint Biometrics



Validation of Operating Characteristics



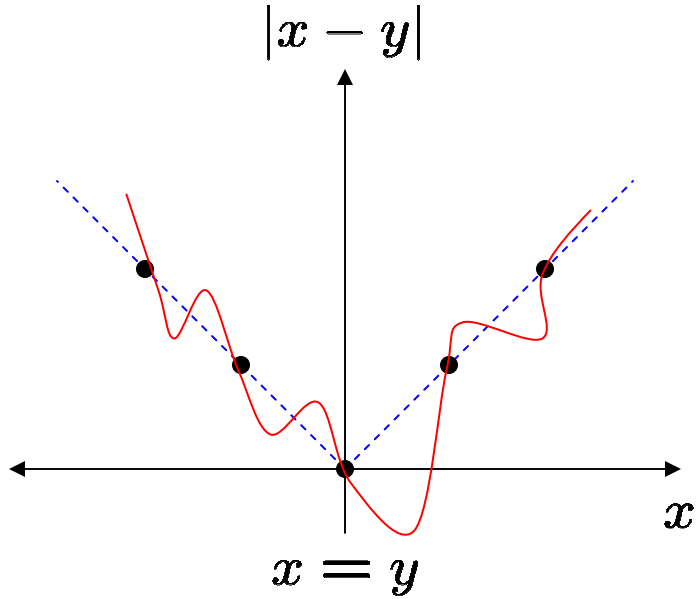
1000 fingers, 15 samples per finger

Similar protocol not possible for L1 distance

$$(x-y)^2 = x^2 + y^2 - 2xy$$

$$x \oplus y = x + y - 2xy$$

$$|x-y| = \ ?$$



- Can express integer L1 distance function as a polynomial in a large finite field
- However, tremendously large degree → high protocol overhead

Convert L1 to L2

- Alice and Bob can binarize their inputs as follows:

  Let alphabet size = 5

  2 ≡ [11000] , 4 ≡ [11110]

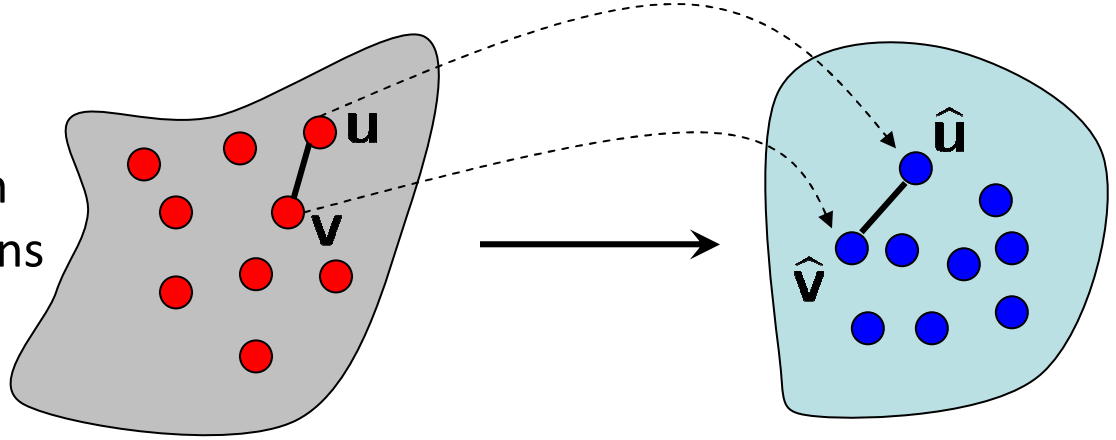
  Then u = [2,4] → ũ = [1100011110]

- Then, $\| \mathbf{x} - \mathbf{y} \|_1 = \| \tilde{\mathbf{x}} - \tilde{\mathbf{y}} \|_1 = \| \tilde{\mathbf{x}} - \tilde{\mathbf{y}} \|_2^2$

- Possible to use squared distance protocol, but this is impractical because we have made our vectors so long

- For vectors of length n, and alphabet size M, size increases to Mn

Reduce dimensionality of new L2 problem



S points in D dimensions

S points in $O(\frac{1}{\epsilon^2} \log S)$ dimensions

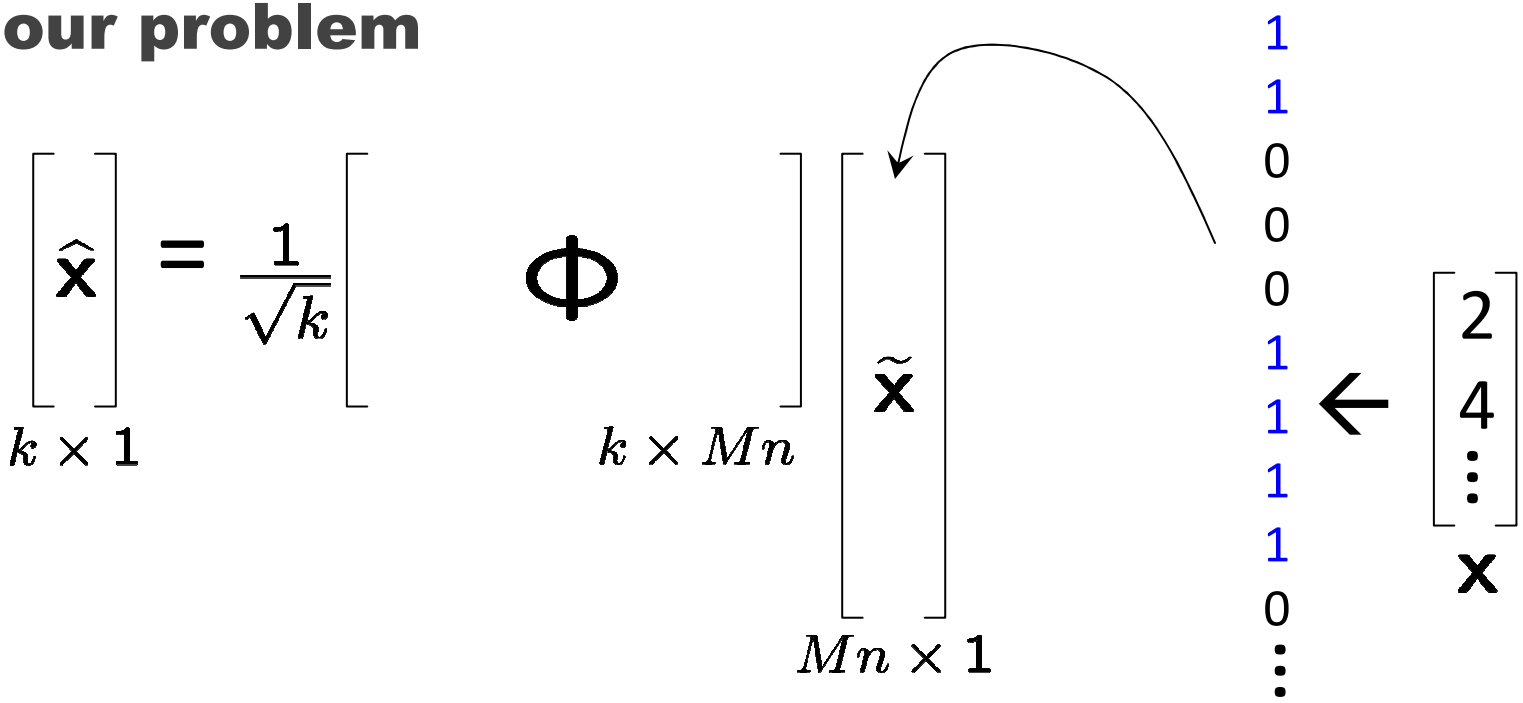
$$\begin{bmatrix} \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \Phi \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \\ \\ \end{bmatrix}_{D \times 1}$$

Φ consists of i.i.d. Gaussian or Bernoulli-0.5 entries

$$(1-\epsilon)||\mathbf{u}-\mathbf{v}||_2^2 \leq ||\hat{\mathbf{u}}-\hat{\mathbf{v}}||_2^2 \leq (1+\epsilon)||\mathbf{u}-\mathbf{v}||_2^2$$

[Johnson, Lindenstrauss, 1984] [Achlioptas, 2001]

JL for our problem



$$k = \alpha \log M^n = \alpha n \log M$$

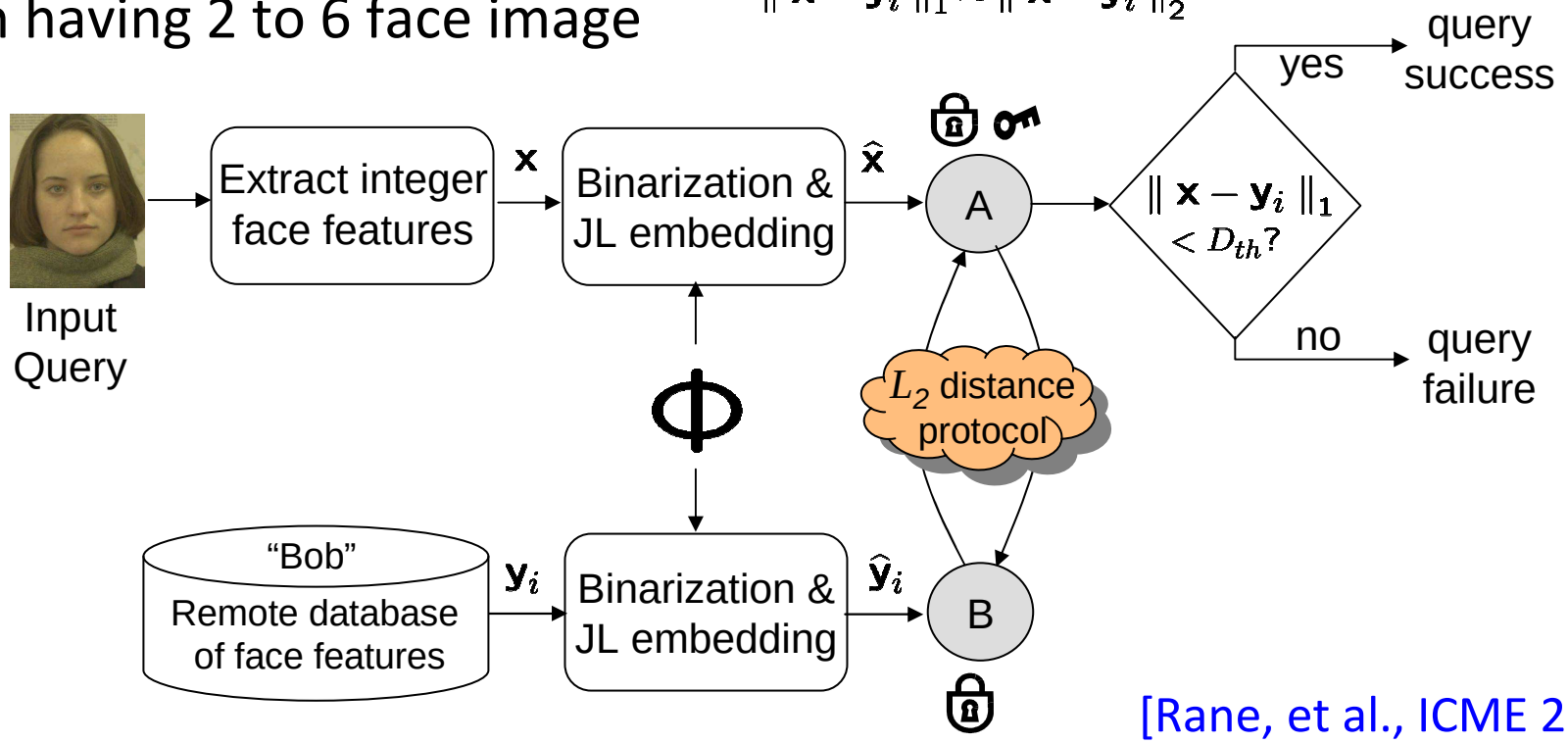
After JL embedding, $||\hat{\mathbf{x}} - \hat{\mathbf{y}}||_2^2 \approx ||\tilde{\mathbf{x}} - \tilde{\mathbf{y}}||_2^2 = ||\mathbf{x} - \mathbf{y}||_1$

Thus, can apply squared distance protocol to JL projections to obtain approximate absolute distance between $\mathbf{x}$ and $\mathbf{y}$

Application: Private Face Image Retrieval

MBGC database, 100 persons,
each having 2 to 6 face image

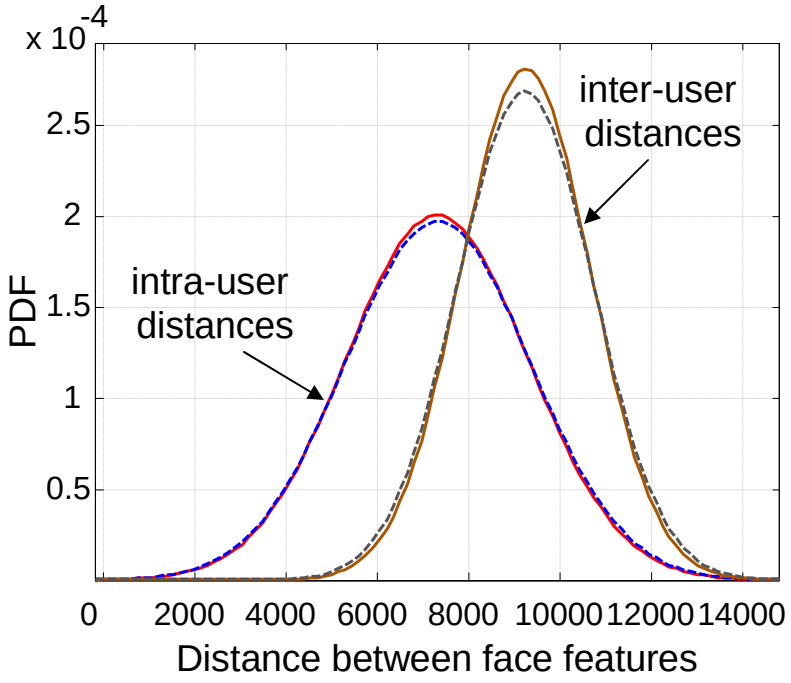
$$\| \mathbf{x} - \mathbf{y}_i \|_1 \approx \| \hat{\mathbf{x}} - \hat{\mathbf{y}}_i \|_2^2$$

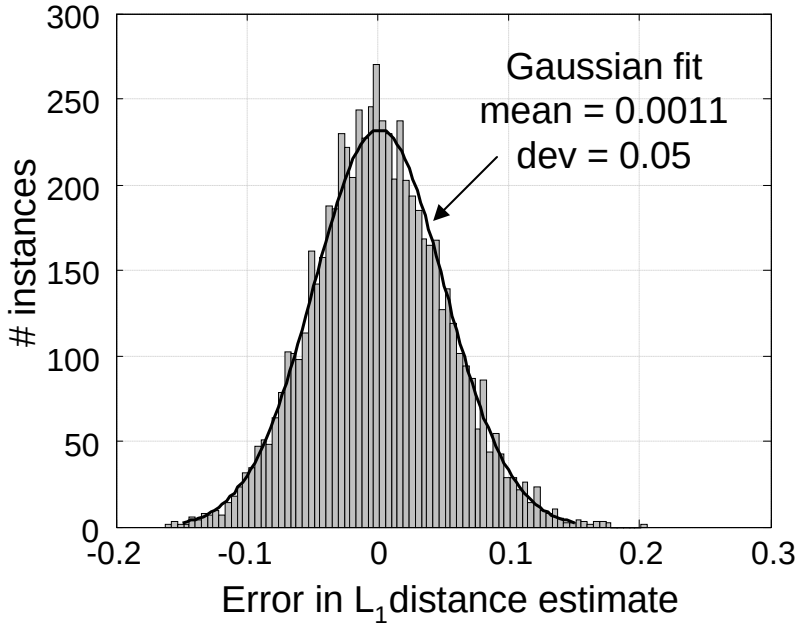


[Rane, et al., ICME 2010]

Feature Vector: 900-length, 8-bit (229.5K after binarization)
JL embedding reduces dimensionality to 7.2K

Accuracy of L1 approximation





6000 pairs of feature vectors chosen at random

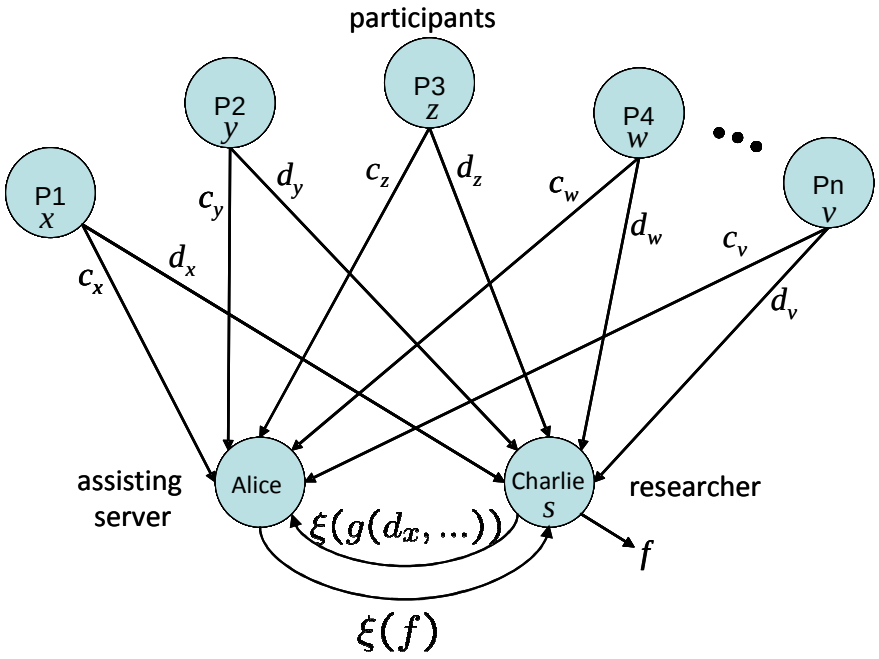
Many other interesting lines of research…

		F	I	R	S	T
	0	1	2	3	4	5
F	1	0	1	2	3	4
A	2	1	1	2	3	4
S	3	2	2	2	2	3
T	4	3	3	3	3	2

Secure Edit Distance

[Rane, et al., WIFS 2010]
(to appear)

Polynomial Evaluation: n parties



[Rane, et al., Allerton 2009]

Summary

- Protocols to evaluate distance between private inputs held by untrusting parties
  - Hamming distance
  - L2 distance
  - L1 distance
  - Edit distance (for some useful substitution costs)
- Use additive homomorphism as a cryptographic primitive to reduce protocol overhead
- Applied to anonymous biometric authentication, but also relevant to many other applications
  - E.g., private image retrieval, comparing DNA sequences, keyword spotting, speaker verification, etc.

Concluding Remarks

- Presented two approaches for secure identity verification
  - ECC-based scheme: Slepian-Wolf setup to cope with noisy data
  - Encryption-based scheme: secure distance calculation

- Various pros and cons for each; best solution depends on application requirements

Thanks for your attention!

Web: http://www.merl.com

Email: avetro@merl.com

Acknowledgments

- Shantanu Rane (MERL)
- Jonathan S. Yedidia (MERL)
- Yige Wang (MERL)
- Wei Sun (MERL)
- Stark C. Draper (U. Wisconsin)
- Yagiz Sutcu (Polytechnic U.)
- Ashish Nagar (MSU)
- Emin Martinian (Bain Capital)
- Ashish Khisti (MIT)