Latent Matching: Wisdom of Crowd

Anil Jain, Kai Cao, Sunpreet Arora, Tarang Chugh

Biometrics Research Group
Michigan State University
http://biometrics.cse.msu.edu

May 13, 2016

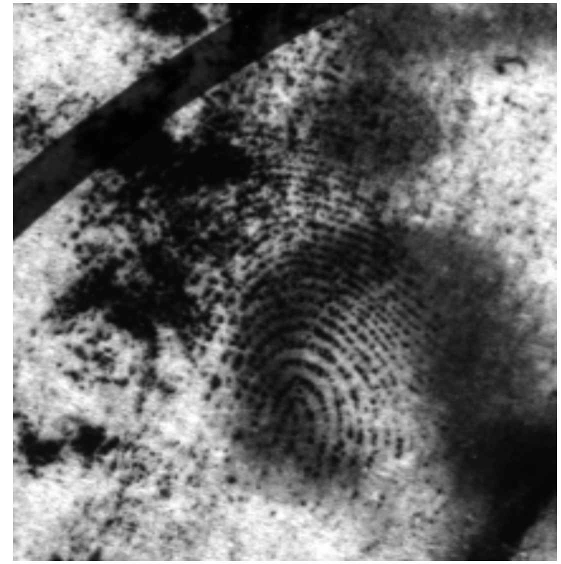

WELL, KEEP SEARCHING LADS, THEY MUST'VE LEFT SOME SORT OF CLUE

Who Left This Print?

AFIS Performance

Rolled

Plain

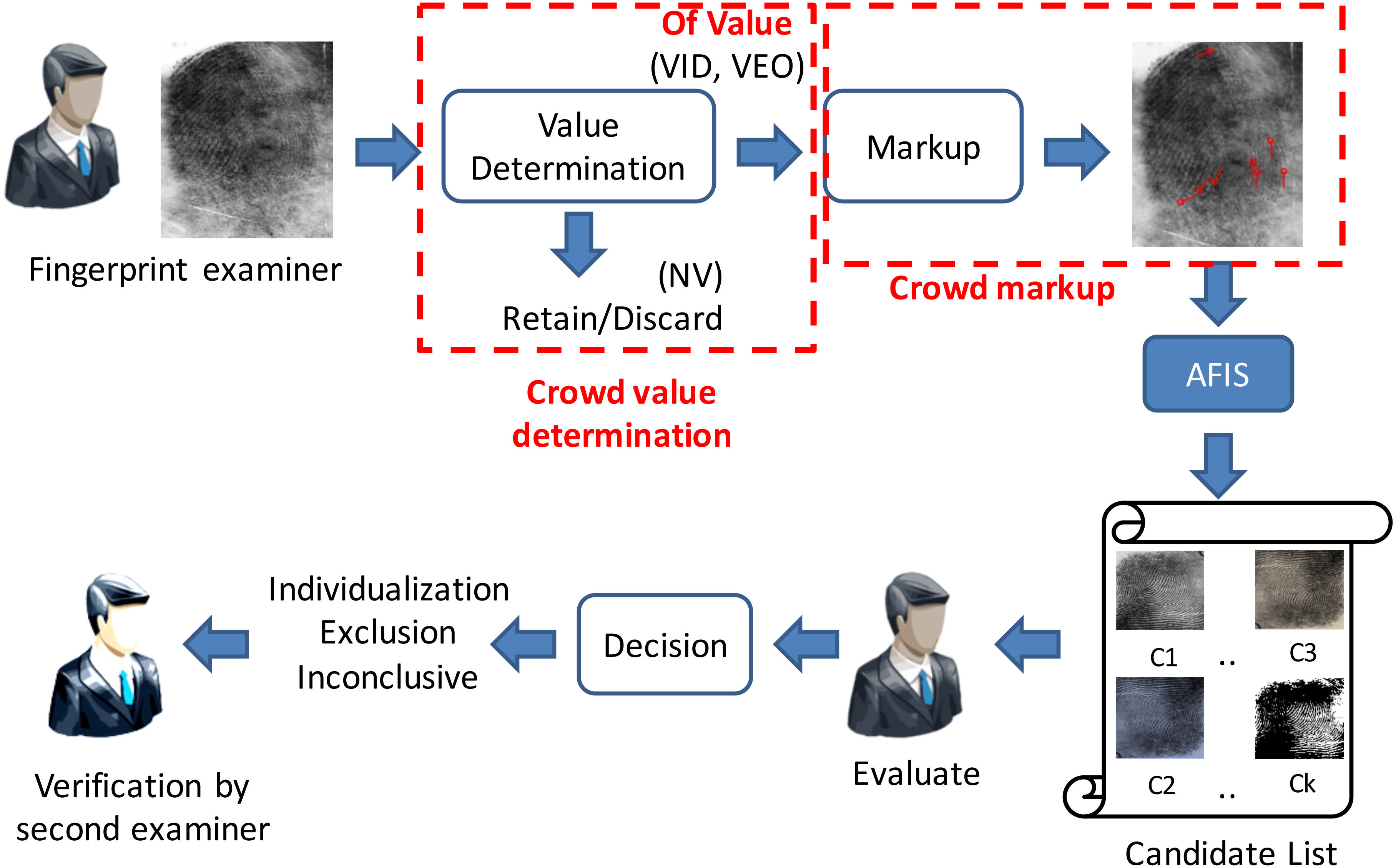


Latent

- Lights-out rank-1 hit rate
  - Plain: 99.3%
  - **Latent: 67.2% (70.2% with image + markup)**

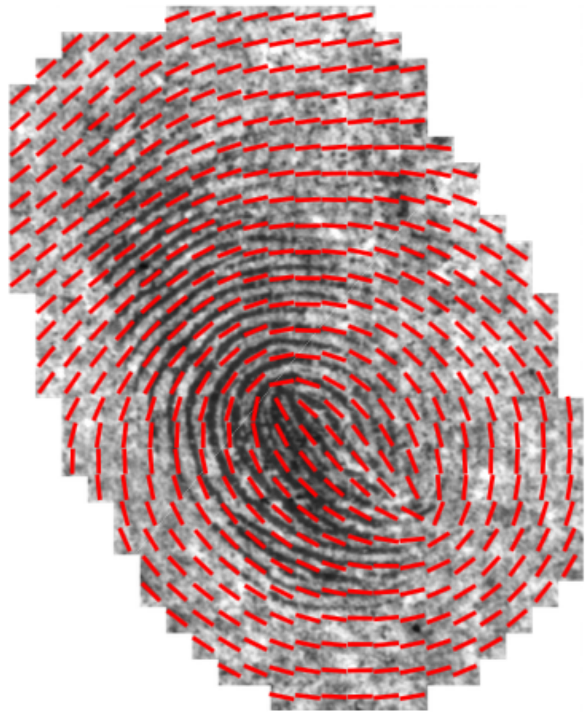
[1] C. Watson, G. Fiumara, E. Tabassi, S. L. Cheng, P. Flanagan, W. Salamon. Fingerprint Vendor Technology Evaluation, NISTIR, 8034, 2012.
[2] M. Indovina, V. Dvornychenko, R. Hicklin, and G. Kiebuzinski. ELFT-EFS Evaluation of Latent Fingerprint Technologies: Extended Feature Sets [Evaluation# 2], NISTIR, 7859, 2012.

Latent Matching: ACE-V Protocol

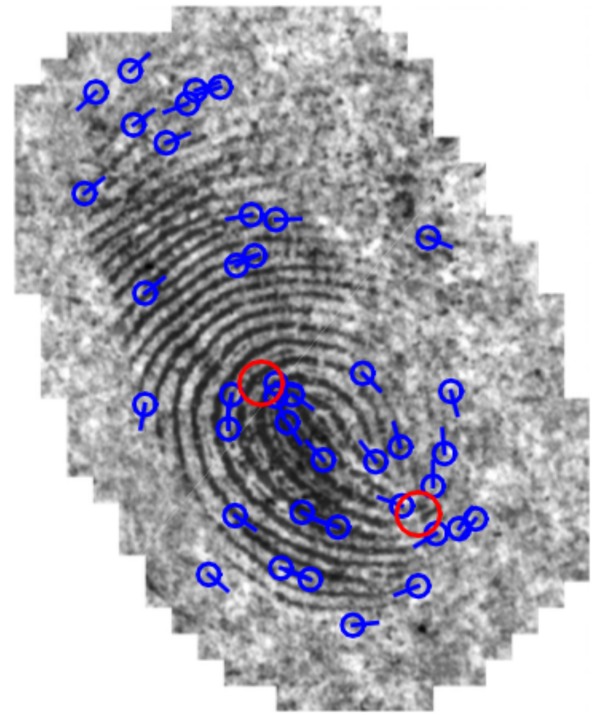


Automatic Latent Processing

Input latent



Ridge flow



Minutiae, core

Crowdsourcing

- *"Leveraging the combined intelligence, knowledge, or experience of a group of people to answer a question, solve a problem, or manage a process"*

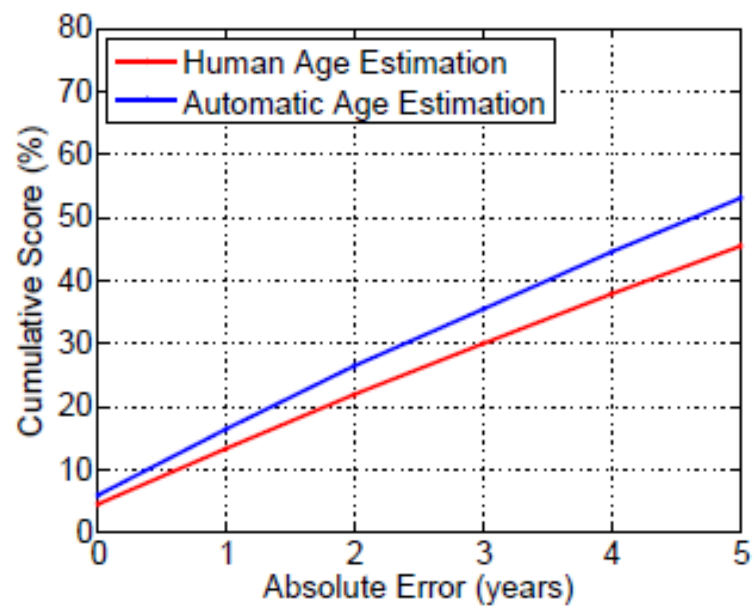
Eric Paul Dennis, Richard Wallace, and Brian Reed. "Crowdsourcing Transportation Systems Data." Technical report, Center for Automotive Research and Parsons Brinkerhoff, 2015.

Image Tagging

Luis Von Ahn & Laura Dabbish. "Labeling images with a computer game." Proc. *SIGCHI conference on Human factors in computing systems*, 2004

Human Age Estimation

- Amazon Mechanical Turk
- 10 estimates (from 10 workers)/image; $0.02/image
- Discard highest/lowest estimates
- Average of response is the human age estimate
- Baseline for automatic age estimation



Crowdsourcing: FR Baseline

Is it the same person in both the videos?

○ I am sure they are the same

○ I think they are the same

○ I cannot tell whether they are the same

○ I do not think they are the same

○ I am sure they are not the same

Play

Familiar? ☐

Name:

Familiar? ☐

Name:

- 5,000 total video pairs
- 40 workers/video; 20 each from US & India
  - ~ 800 total workers
- 6 cents for completing 5 video pairs

	TAR @ 1% FAR	TAR @ 10% FAR
USA workers	80.6	96.7
Indian workers	63.7	92.4
FR COTS	54.4	81.4

L. Best-Rowden, S. Bisht, J. Klontz and A. K. Jain. Unconstrained Face Recognition: Establishing Baseline Human Performance via Crowdsourcing. In *Proc. IJCB*, 2014.

Crowdsourcing: Details

- Human Intelligence Task (HIT): What to ask?
- Non-expert v. **Expert** crowd
- How to incentivize workers?
- How many workers?
- How to consolidate information?
- How do we know a worker can be trusted?

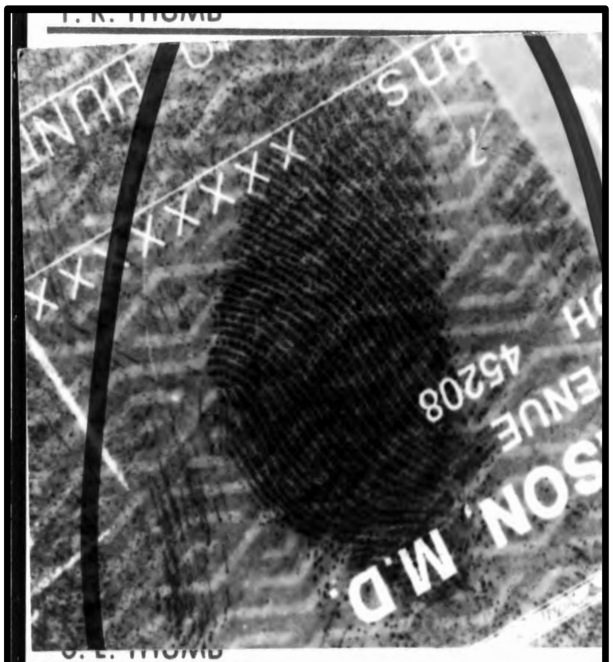
Latent Value Determination: Learning From Crowd

Goals

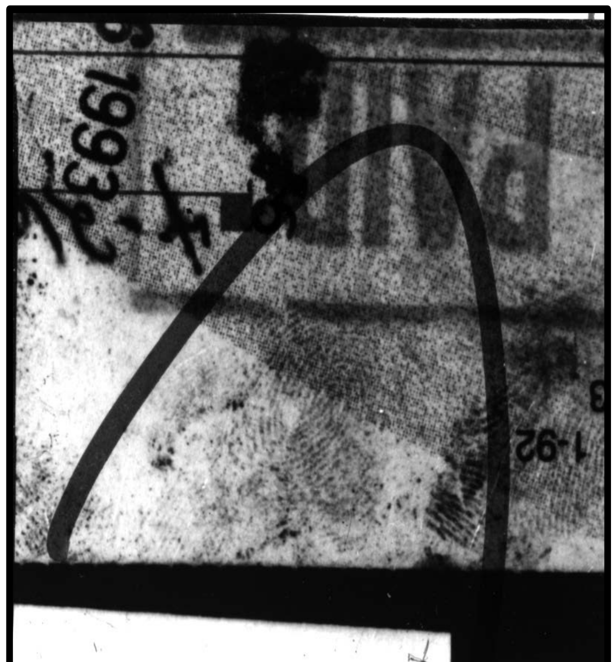
- Learn a model for **quantitative value** assignment to a latent
- How do fingerprint examiners assign **subjective value**?
- Understand the relationship between subjective value assignment and **latent image features**

Value Determination

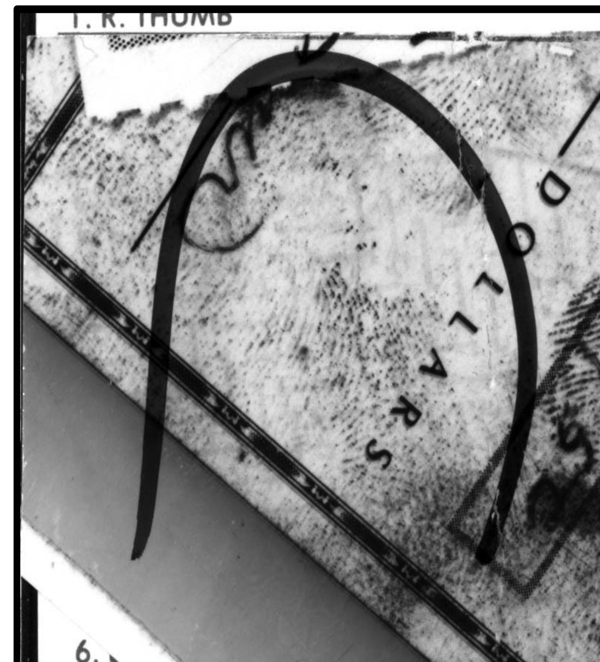
- Presumably based on two factors:
  - Image quality (e.g., clarity of friction ridges)
  - Information content (e.g., no. of minutiae)



VID
(Value for Identification)

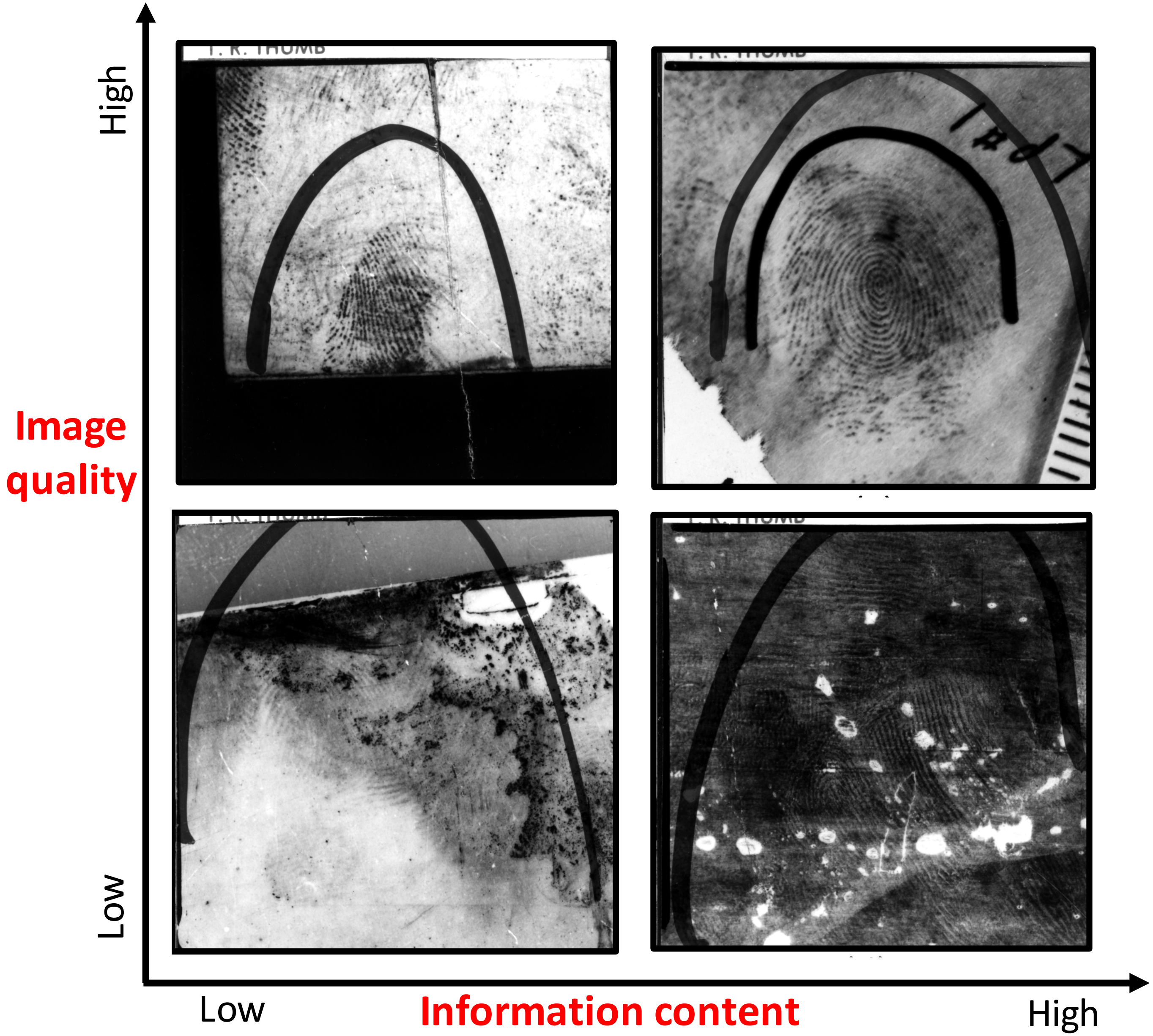


VEO
(Value for Exclusion Only)



NV
(No Value)

Quality v. Information

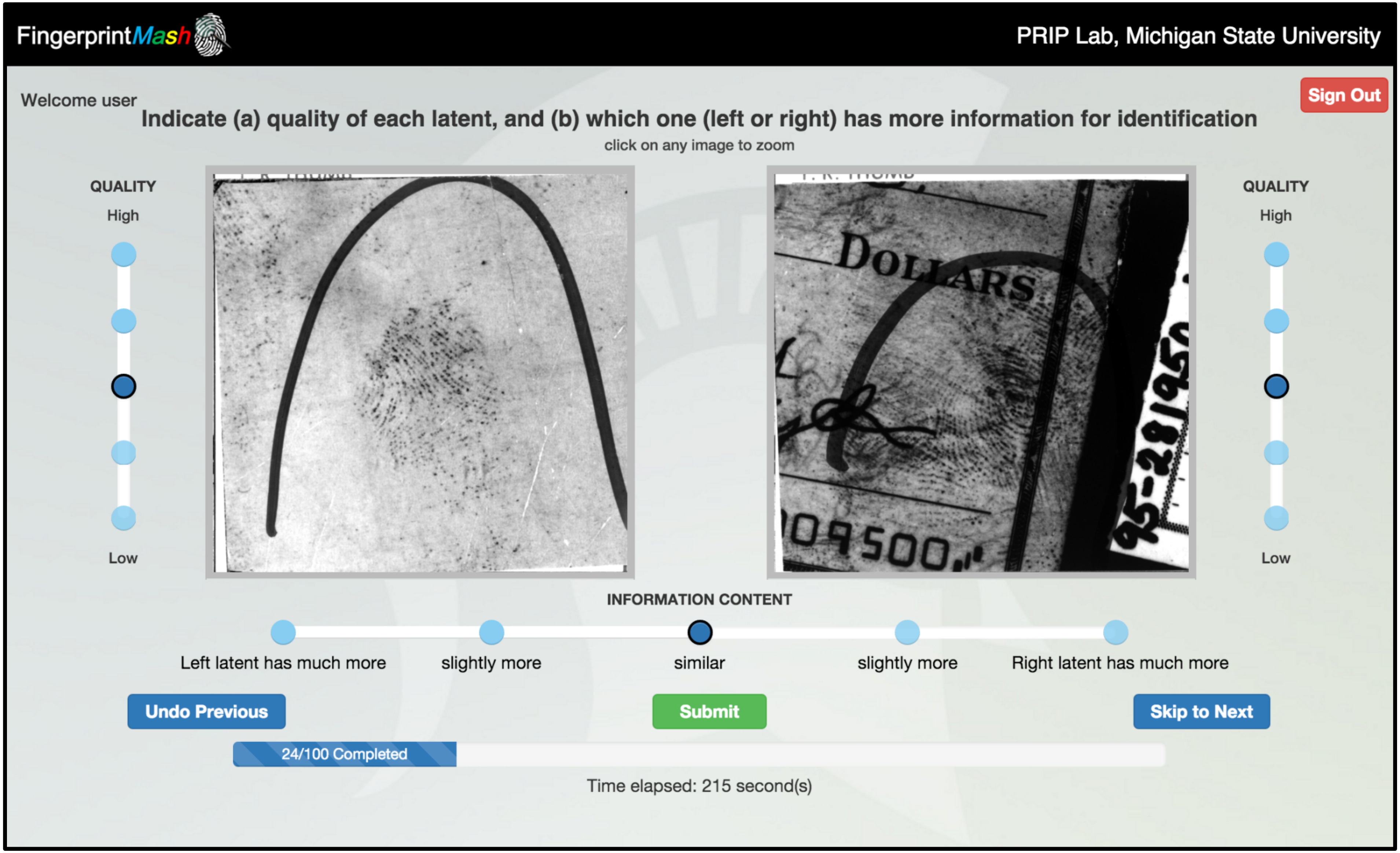


Why Value Determination?

- Triage for Large Caseload
  - FBI's I*AFIS conducted 14,311 latent feature searches and 2,370 latent image searches (Dec. 2015)*
- **Of Value** [VID, VEO] latent classified as **No Value** (NV) latent is a missed opportunity to identify the suspect
- **No Value** latent determined as **Of Value** leads to ineffective use of examiner's effort in feature markup & verification

Need for automatic value assignment

FingerprintMash: A Crowdsourcing Tool

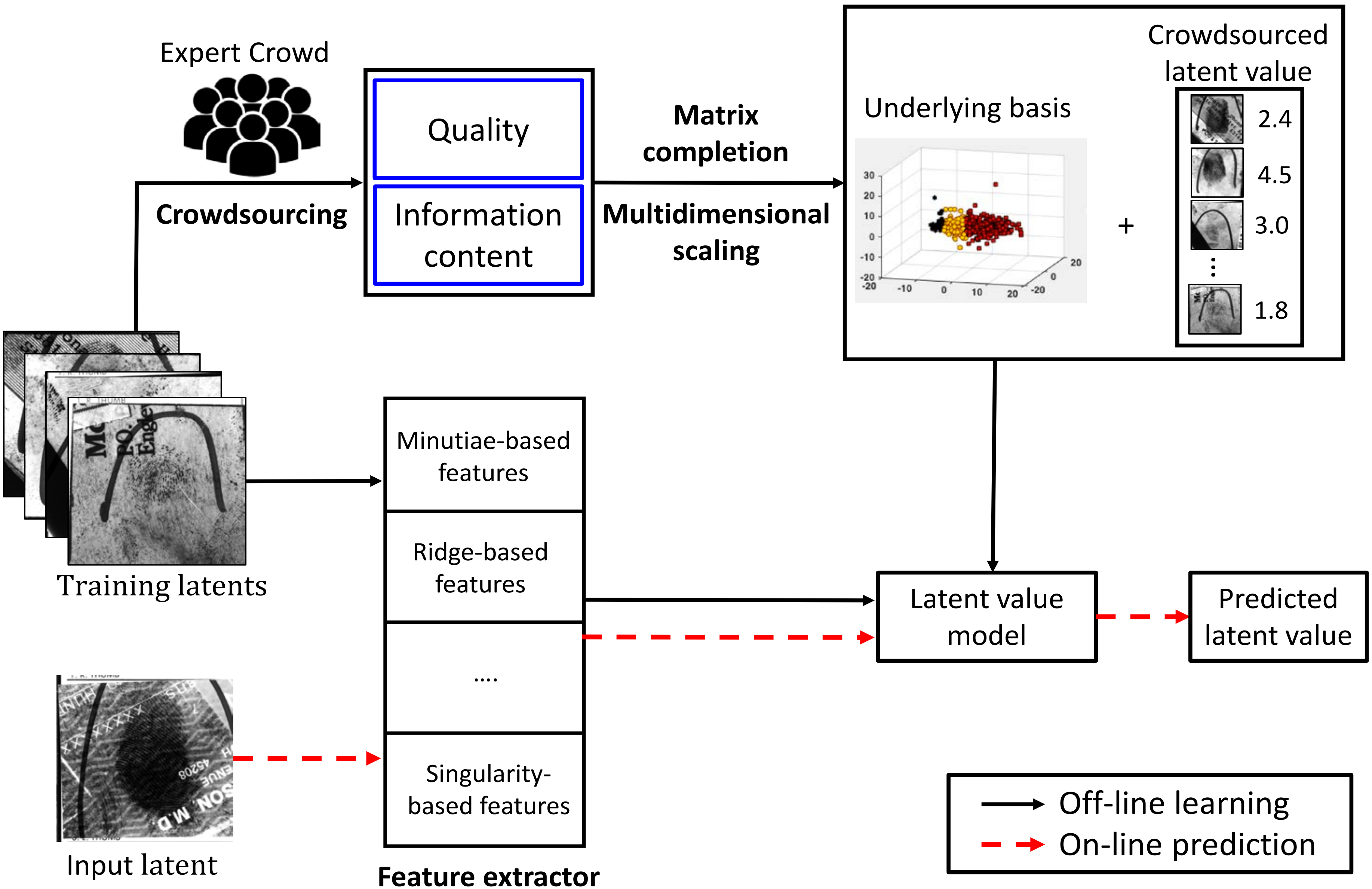


http://fingerprintmash.org

Crowdsourcing Details

- **Expert Crowd**
  - 31 experts (latent examiners and researchers)
- **Dataset**
  - 258 latents from NIST SD27
  - 258 latents from the Michigan State Police (MSP)
- **Protocol**
  - 5 levels for numerical quality rating; 1: low, 5: high
  - 5 relative levels for information content; 1: left is better
  - 100 randomly selected pairs presented to each expert
  - 1 pair is repeated at every 5th comparison for validity

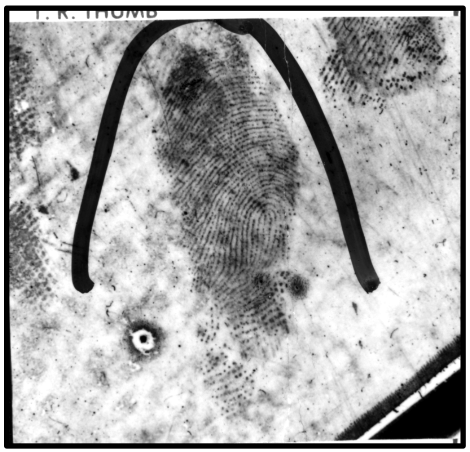
Automatic Latent Value Assignment



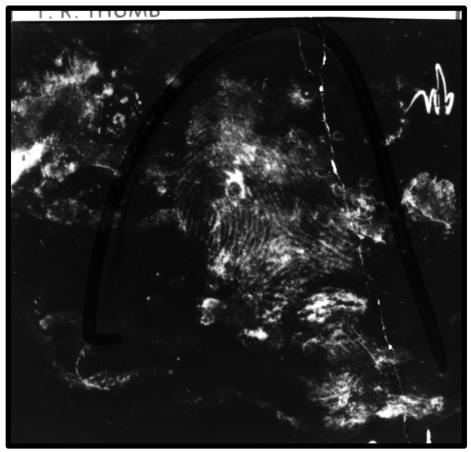
Intra-Expert Variance

Numerical Rating

Mean = 0.35; Std. Dev. = 0.16



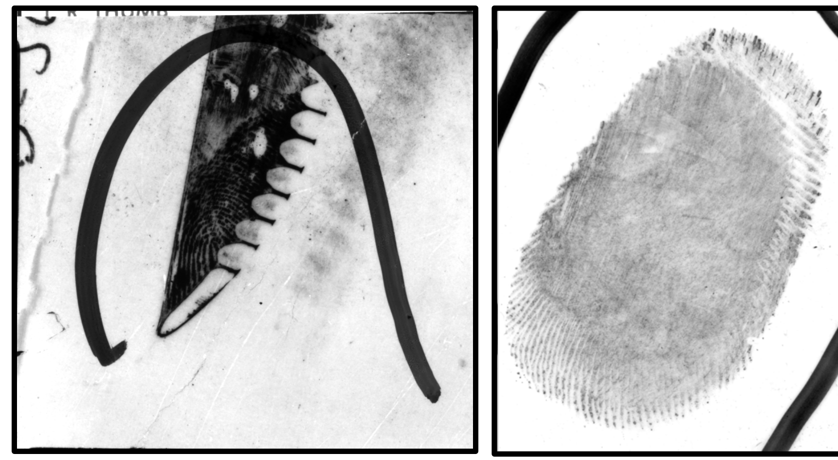
Low variance
Values = [4,4,4,4]



High variance
Values = [1,2,2,4]

Pairwise Comparison

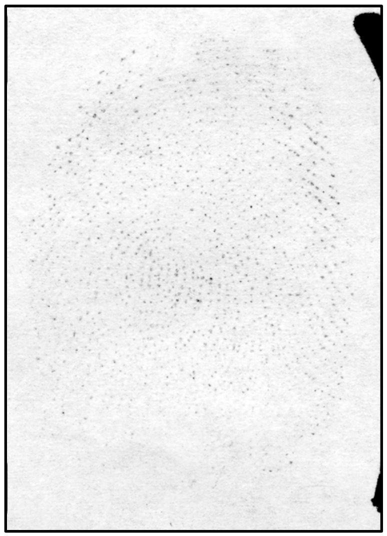
Mean = 0.24; Std. Dev. = 0.15



High variance in comparison
Value ranged from 1 to 4

Inter-Expert Variance

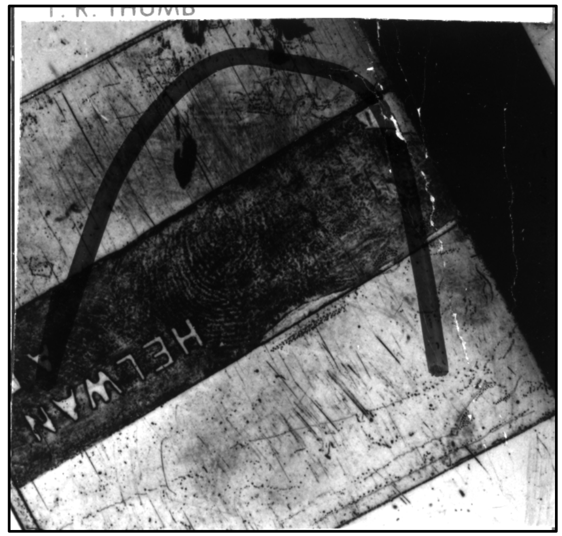
Examples of strong agreement among experts



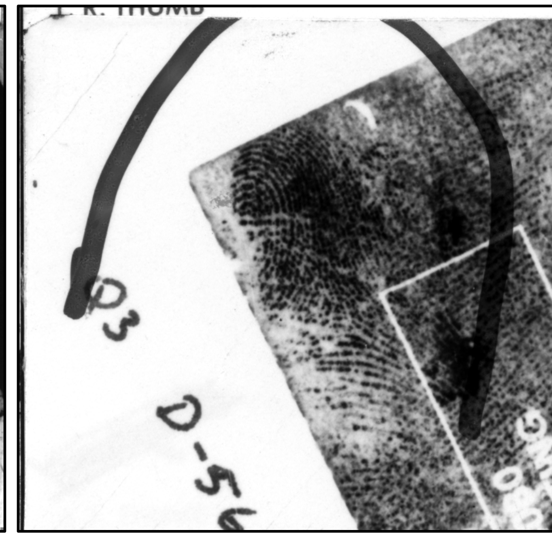
Low Value (1)

High Value (5)

Examples of strong disagreement among experts



Value Range (2-5)



Value Range (1-5)

Expert Crowd v. Value Determination

- NIST SD27: 210 VID, 41 VEO, and 7 NV from [5]

Rank-1 hit rates from AFIS; 250K reference ptints

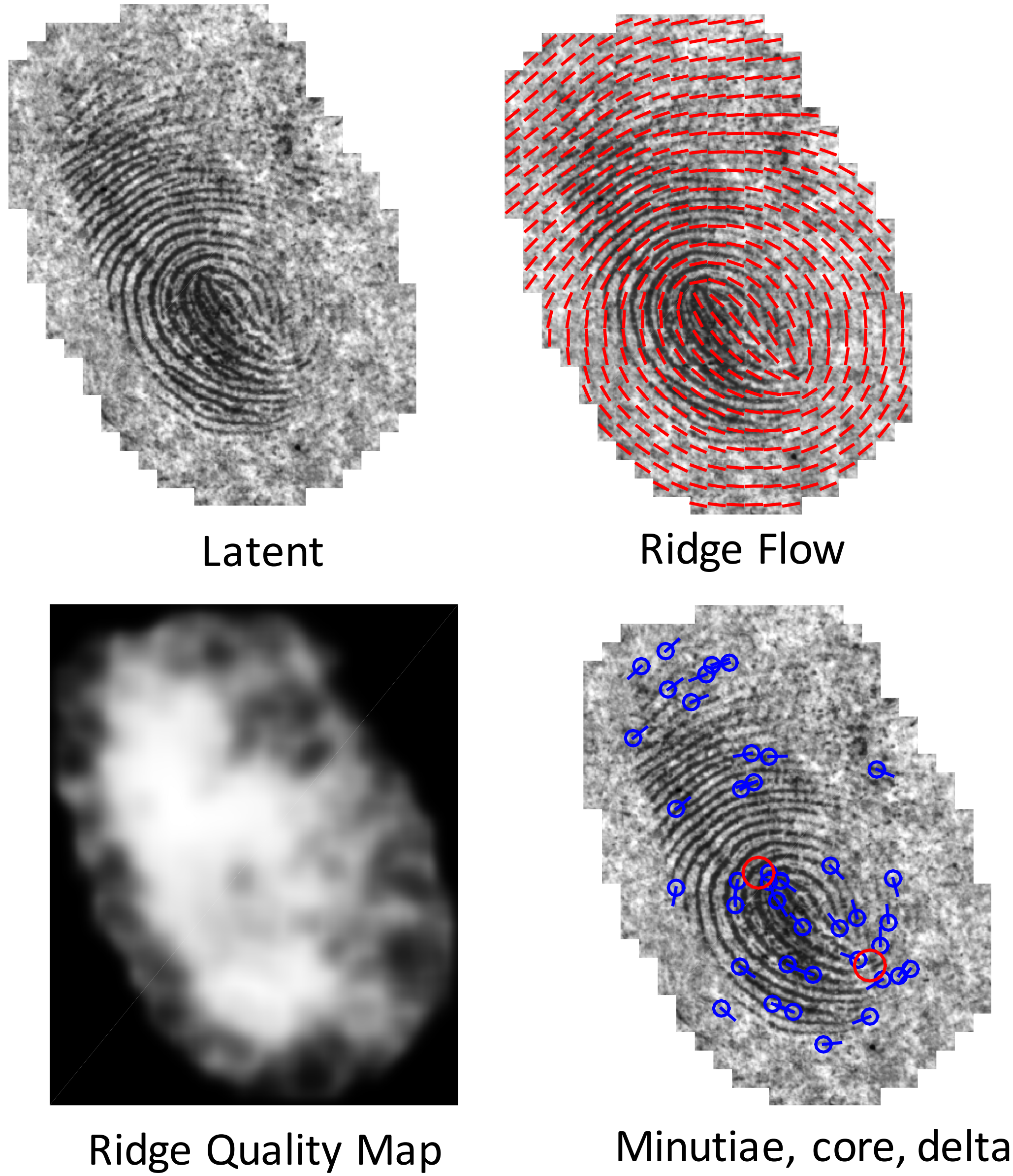
Value Determination	VID	VEO	NV
	163/210	13/41	0/7
Expert Crowd	High value	Medium value	Low value
	169/210	**7/41**	**0/7**

For a fair comparison, we identify, from crowdsourced value data, the top 210 valued latents as high value, next 41 as medium value and the remaining 7 as low value.

R. A. Hicklin, et al, “Latent Fingerprint Quality: A Survey of Examiners,” Journal of Forensic Identification, vol. 61, no. 4, pp. 385–419, 2011.

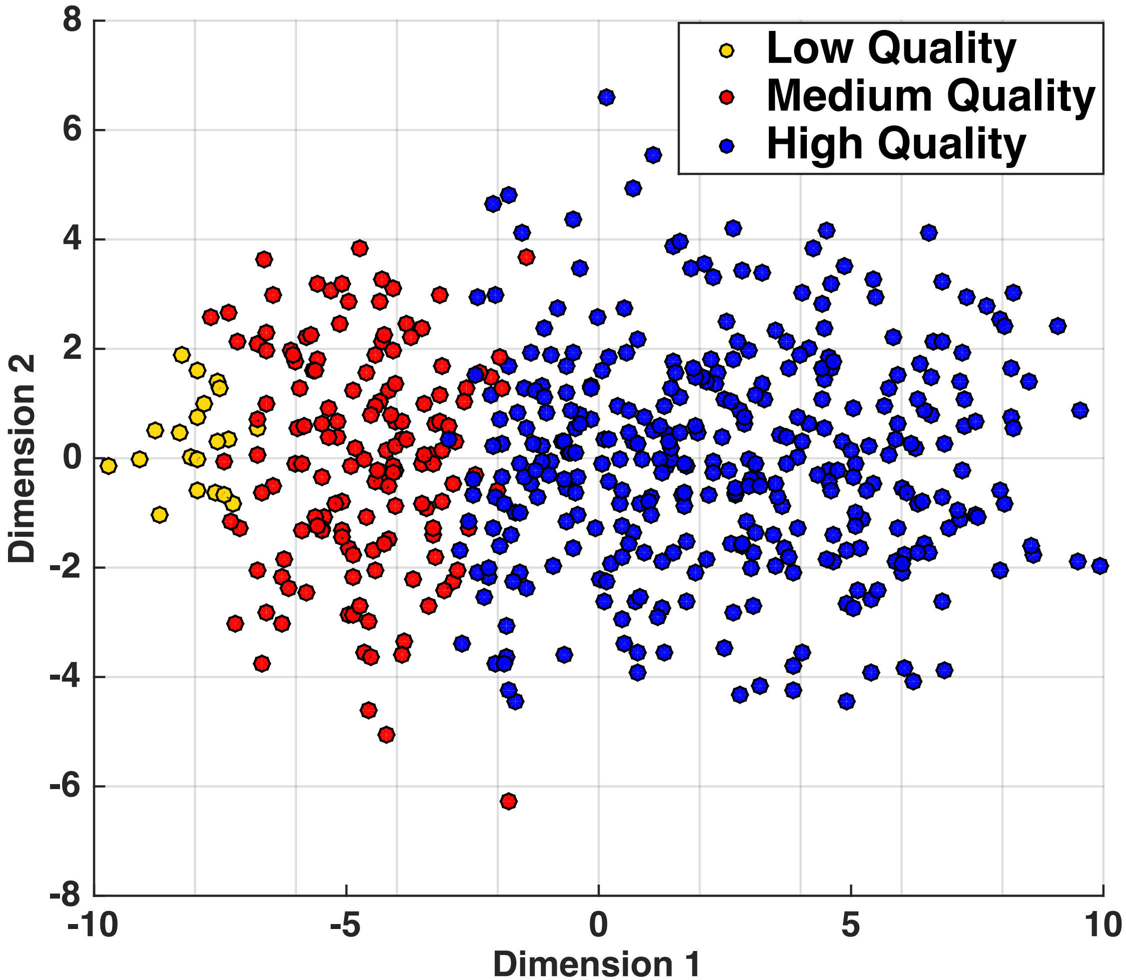
Automatic Feature Extraction

- #Minutiae
- Minutiae reliability
- Friction ridge area
- Ridge quality
- #singular points
- S.d. of ridge flow



K. Cao, T. Chugh, J. Zhou, E. Tabassi, and A. K. Jain, "Automatic latent value determination," in ICB, 2016, pp. 1–6.

Multi-Dimensional Scaling



Inter-latent similarity explained by two dimensions (bases)

Interpretation of the Two Bases

- Each basis is a weighted sum of the latent features

Weight Vector* for Dimension #	#Minutiae	Minutiae reliability	Friction ridge area	Ridge quality	#Singular points	S.d. of ridge flow
1	0.62	0.35	0.02	0.24	0.59	0.29
2	0.15	0.11	0.95	0	0.24	0.07

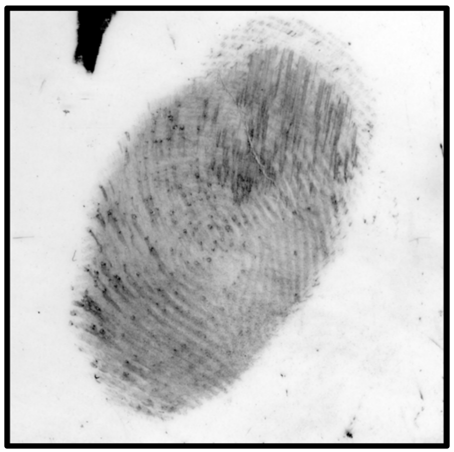
*weight vectors are normalized to make them unit length

Value Prediction

- Learned a predictor for value assignment
- Average MSE for predicted value = 0.24

Correct prediction

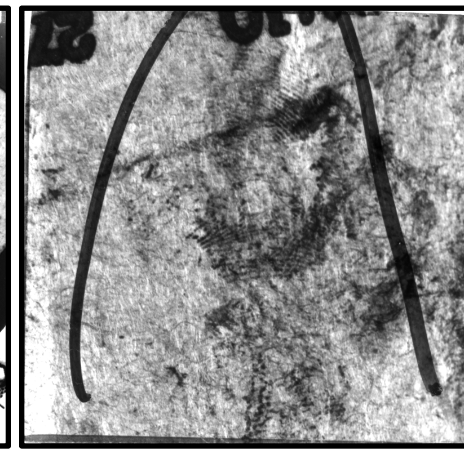
Crowd : 4.67
Predicted : 4.68



Crowd : 2.08
Predicted : 2.04

Incorrect prediction

Crowd : 2.99
Predicted : 4.02

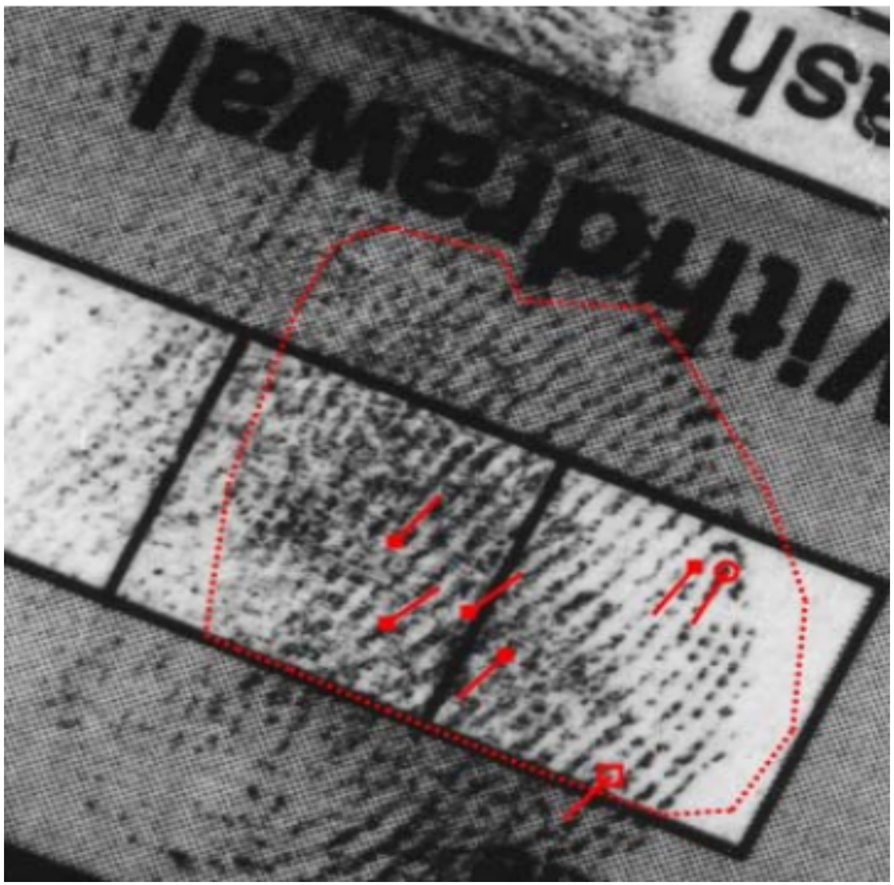


Crowd : 2.96
Predicted : 2.02

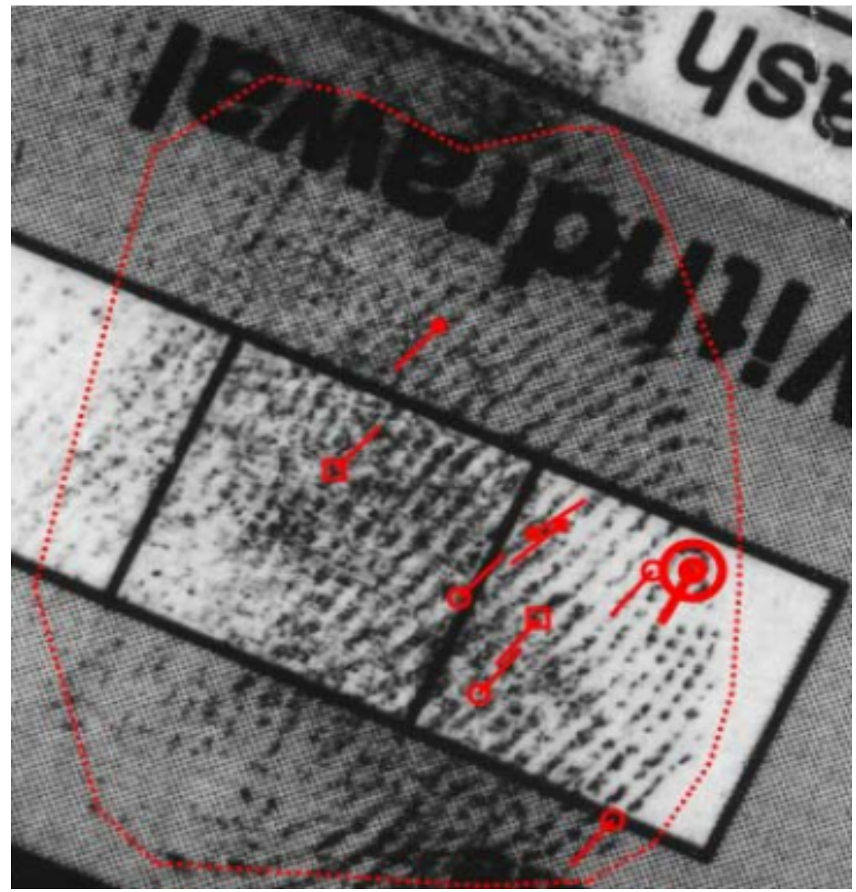
Crowdsourcing Markup

Latent Mark up

- Different examiners provide different mark ups



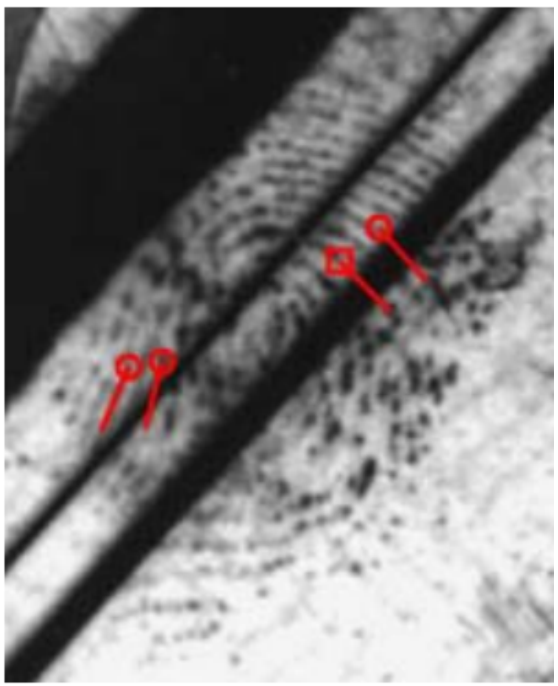
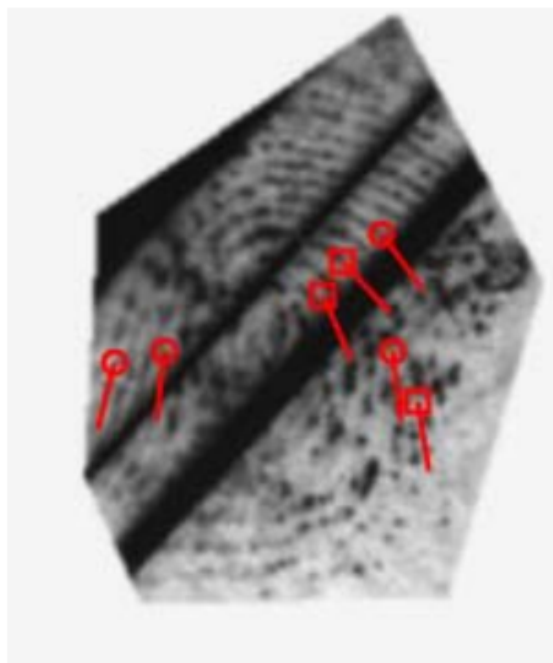
Markup 1: Hit at **rank-1**



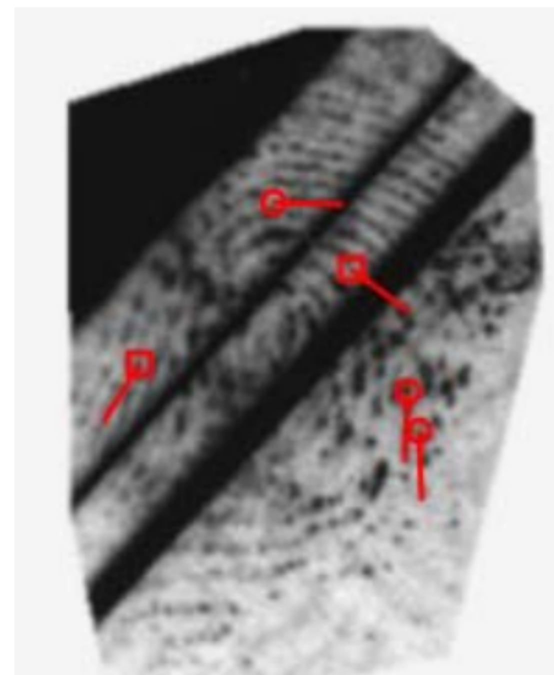
Markup 2: Hit at **rank-129**

Approach

- Use collective wisdom of multiple examiners
- Use a team of latent examiners (expert crowd) for mark up as needed

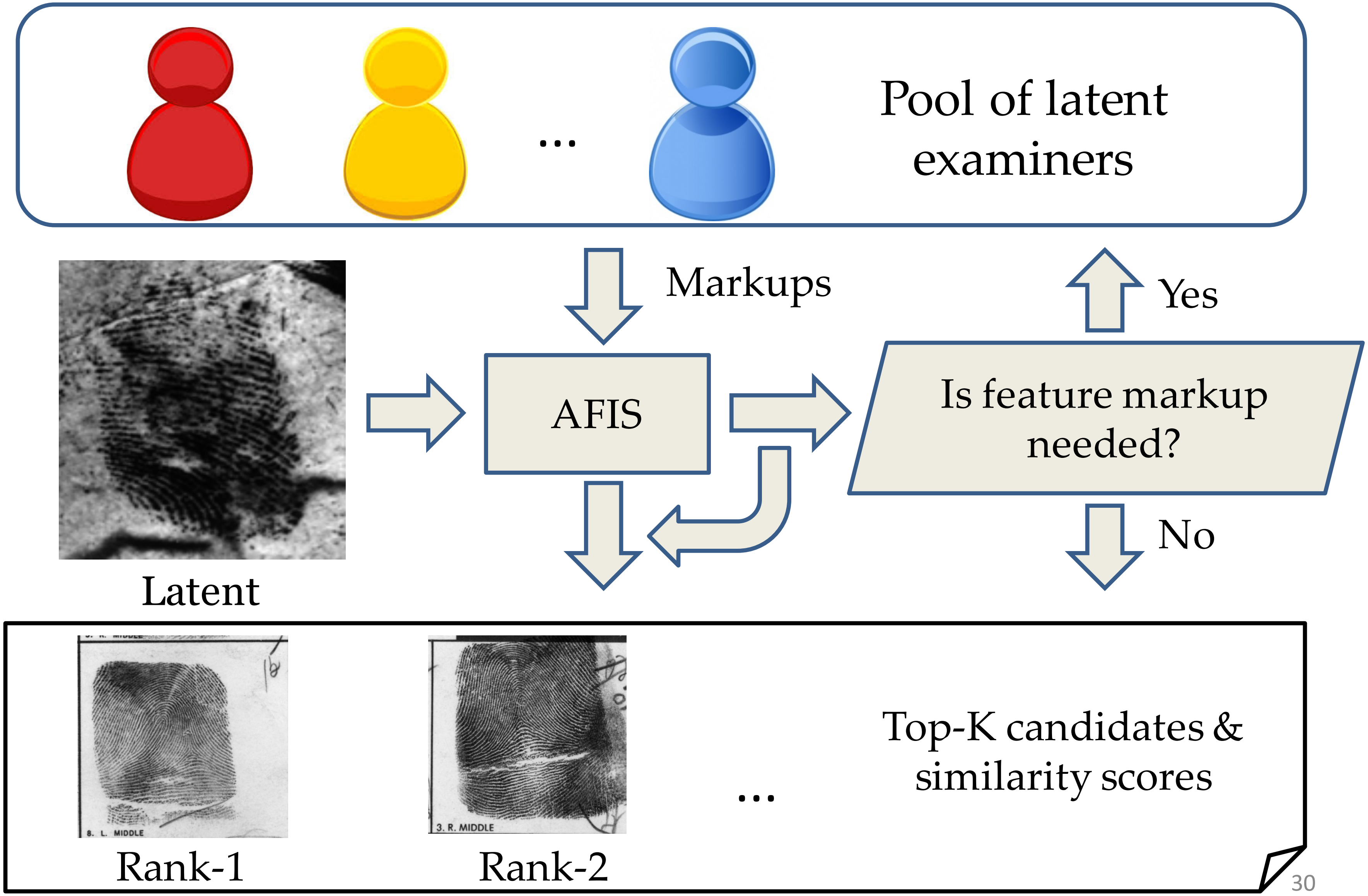



...



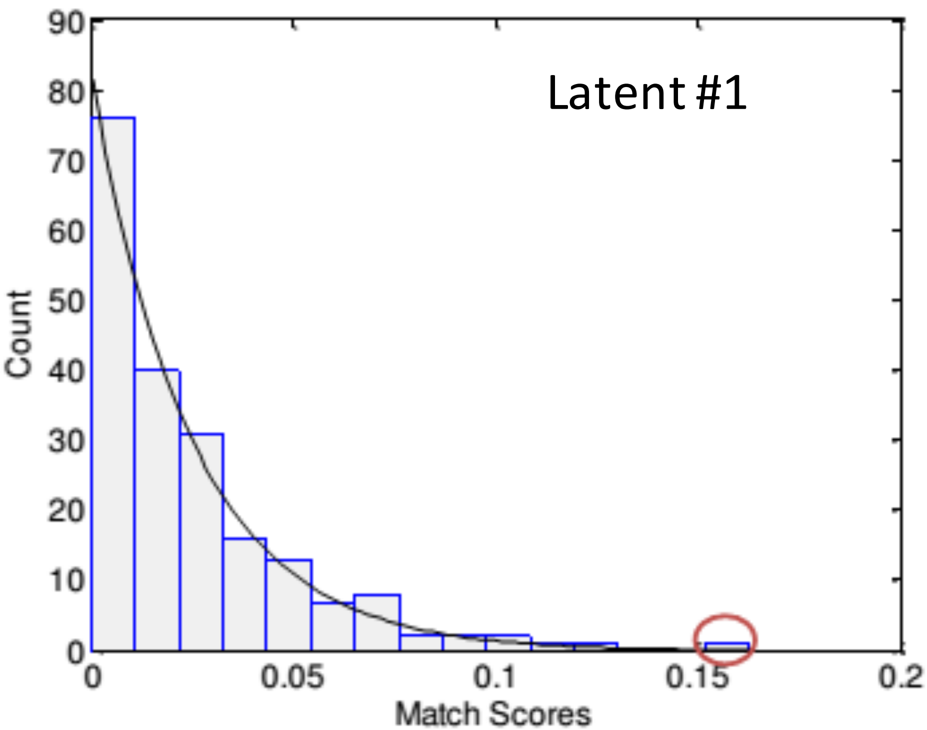
D. Retelny, et. al. Expert crowdsourcing with flash teams. In ACM Symposium on UIST, 2014.

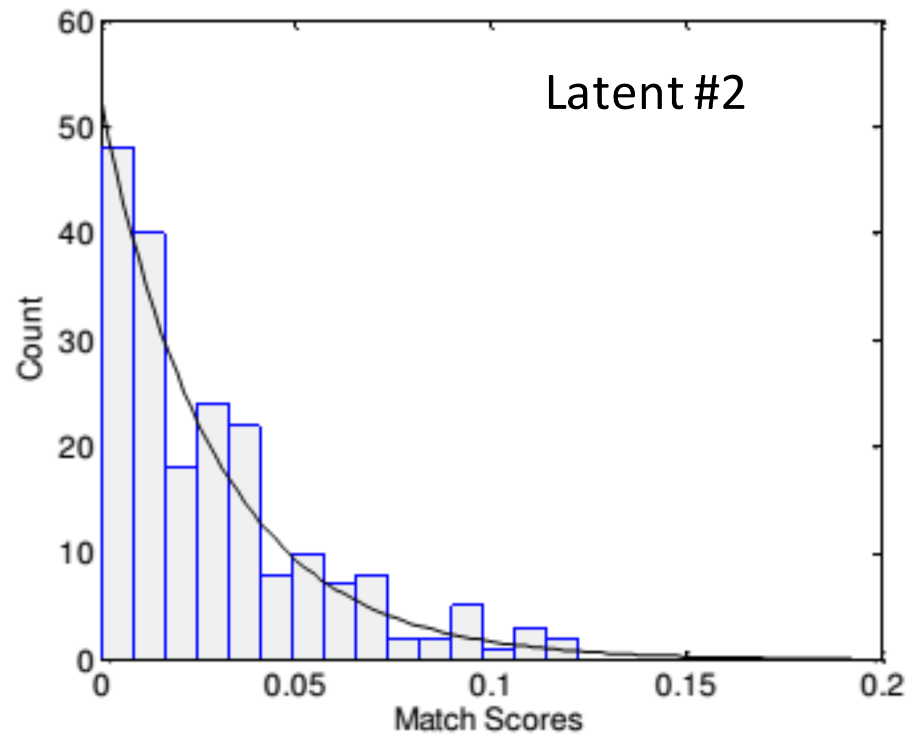
Expert Crowdsourcing Framework



When to Crowdsource?

Histogram of top-K scores between latent and reference prints



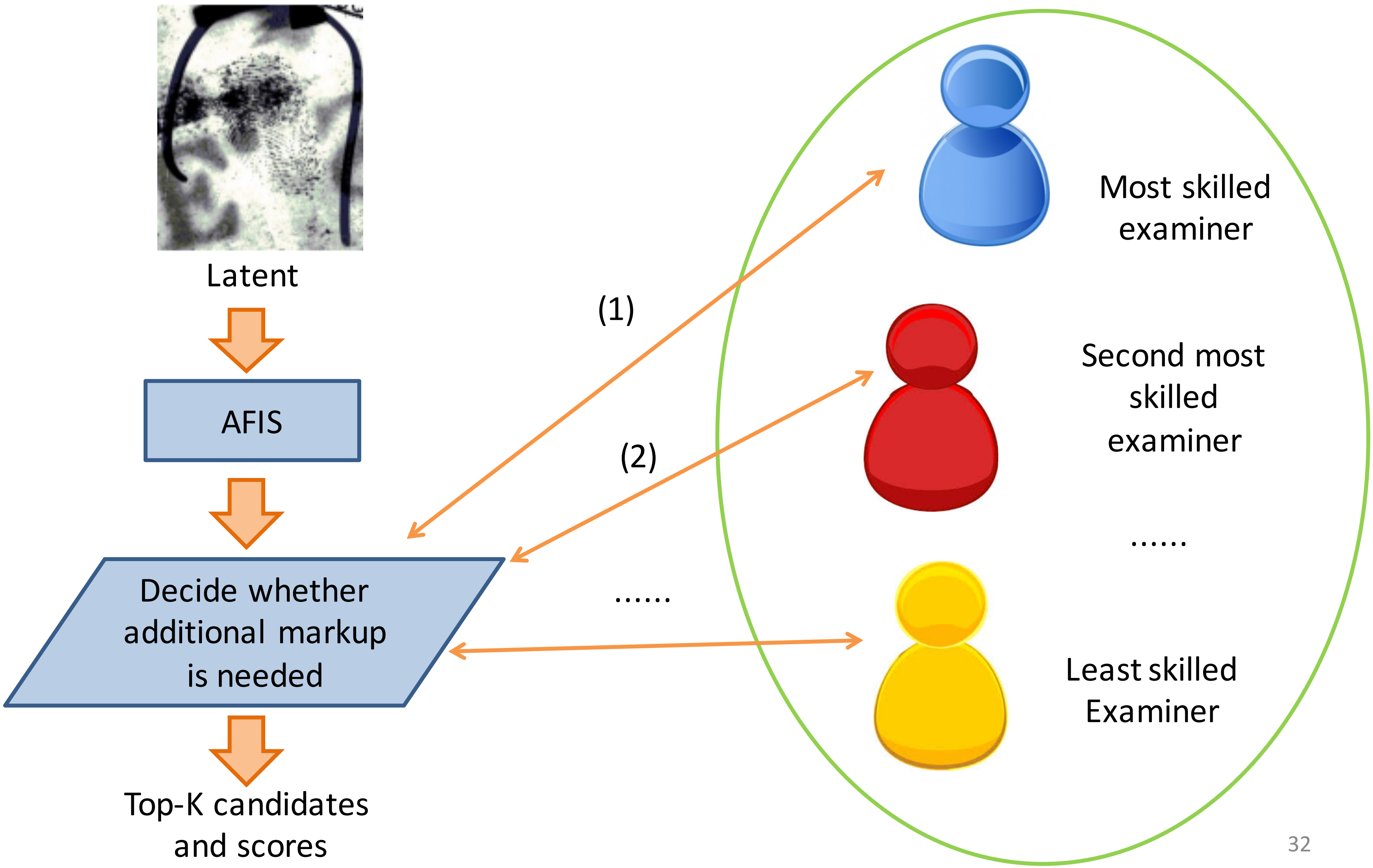


Upper outlier detected
=> Lights-out AFIS adequate

No upper outlier detected
=> Obtain examiner markups

S. S. Arora, E. Liu, K. Cao and A. K. Jain, "Latent Fingerprint Matching: Performance Gain via Feedback from Exemplar Prints", IEEE TPAMI, 2014.

How Many Experts are Enough?



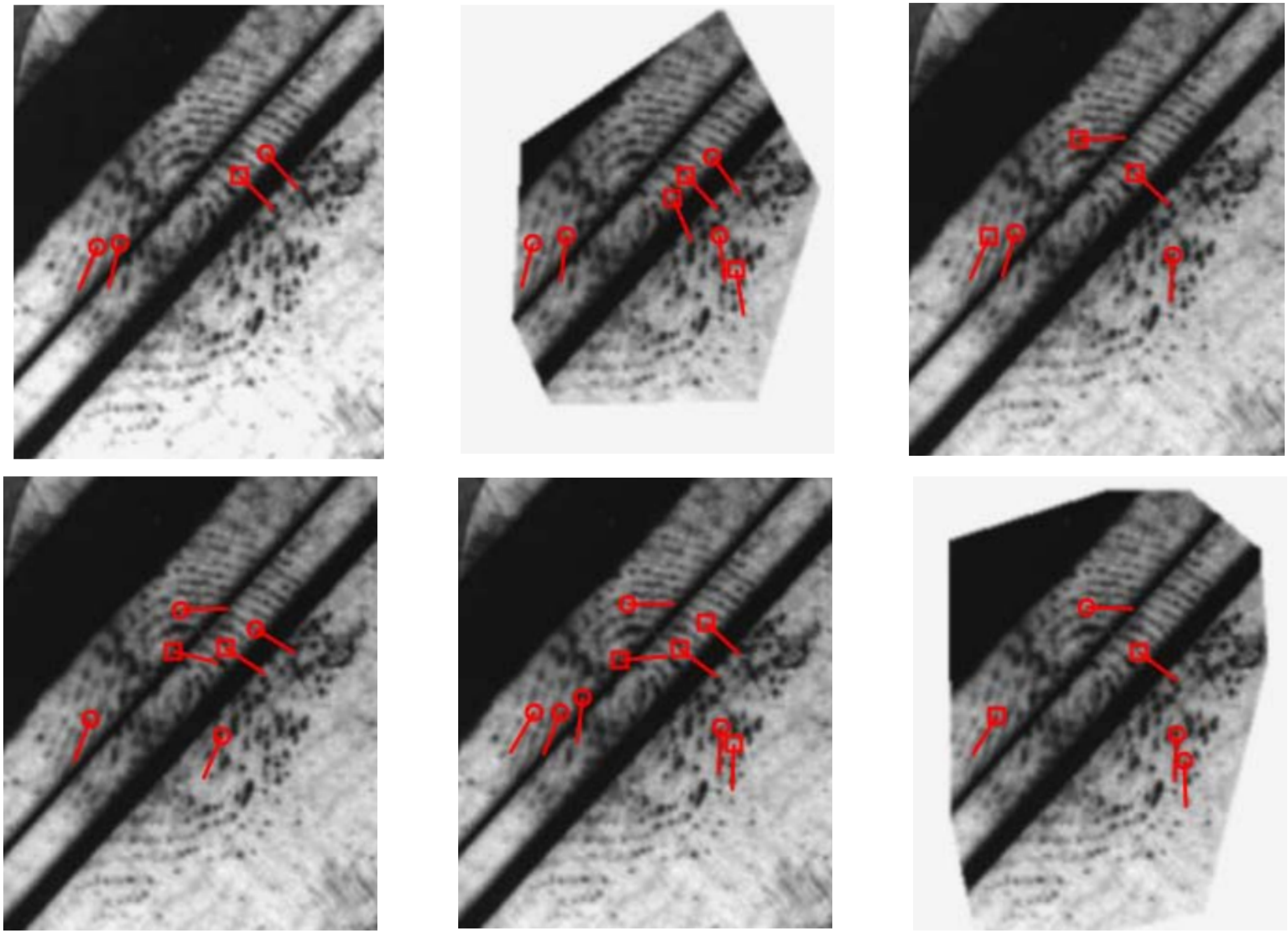
Experiments

- Latent Databases

Database	#Latents	Resolution	Latent type	#Examiner Markups
NIST SD27	258	500	operational	6
ELFT EFS	255	1000	operational	2
RS&A	200	1000	collected in lab	1

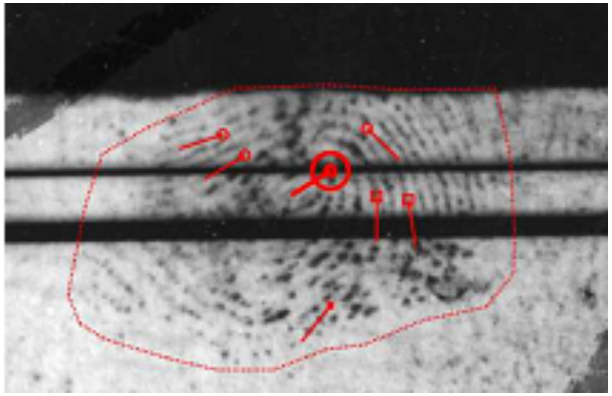
- Reference Database
  - 250K rolled prints (true mates, MSP, NIST)
- Latent AFIS
  - Top performer in NIST ELFT-EFS 2

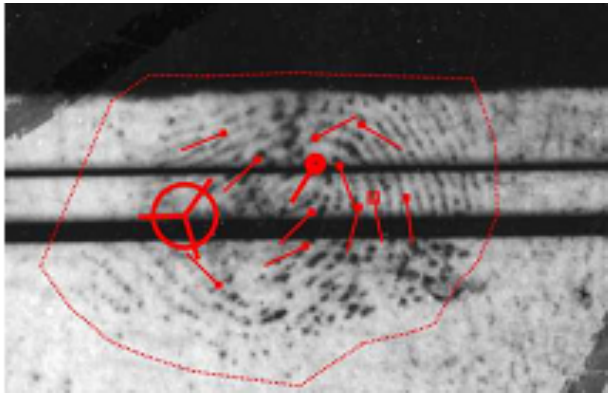
Sample Markups: NIST SD27



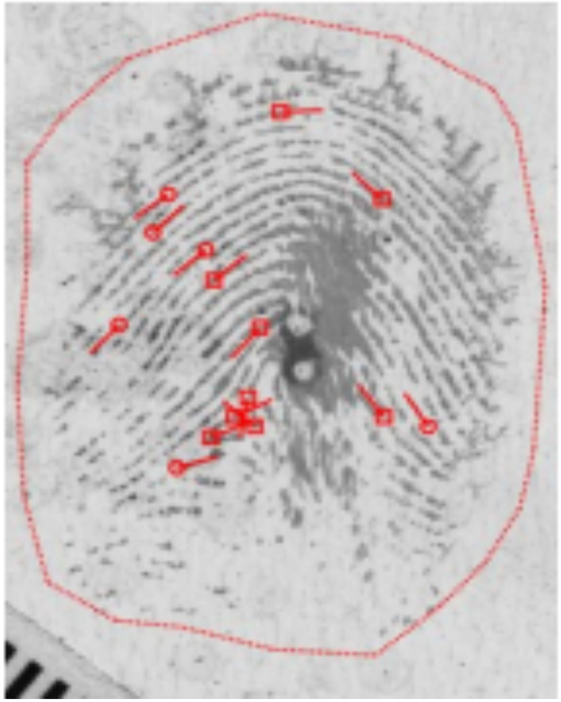
Markups by 6 examiners for a latent in NIST SD27

Sample Markups: ELFT EFS & RS&A



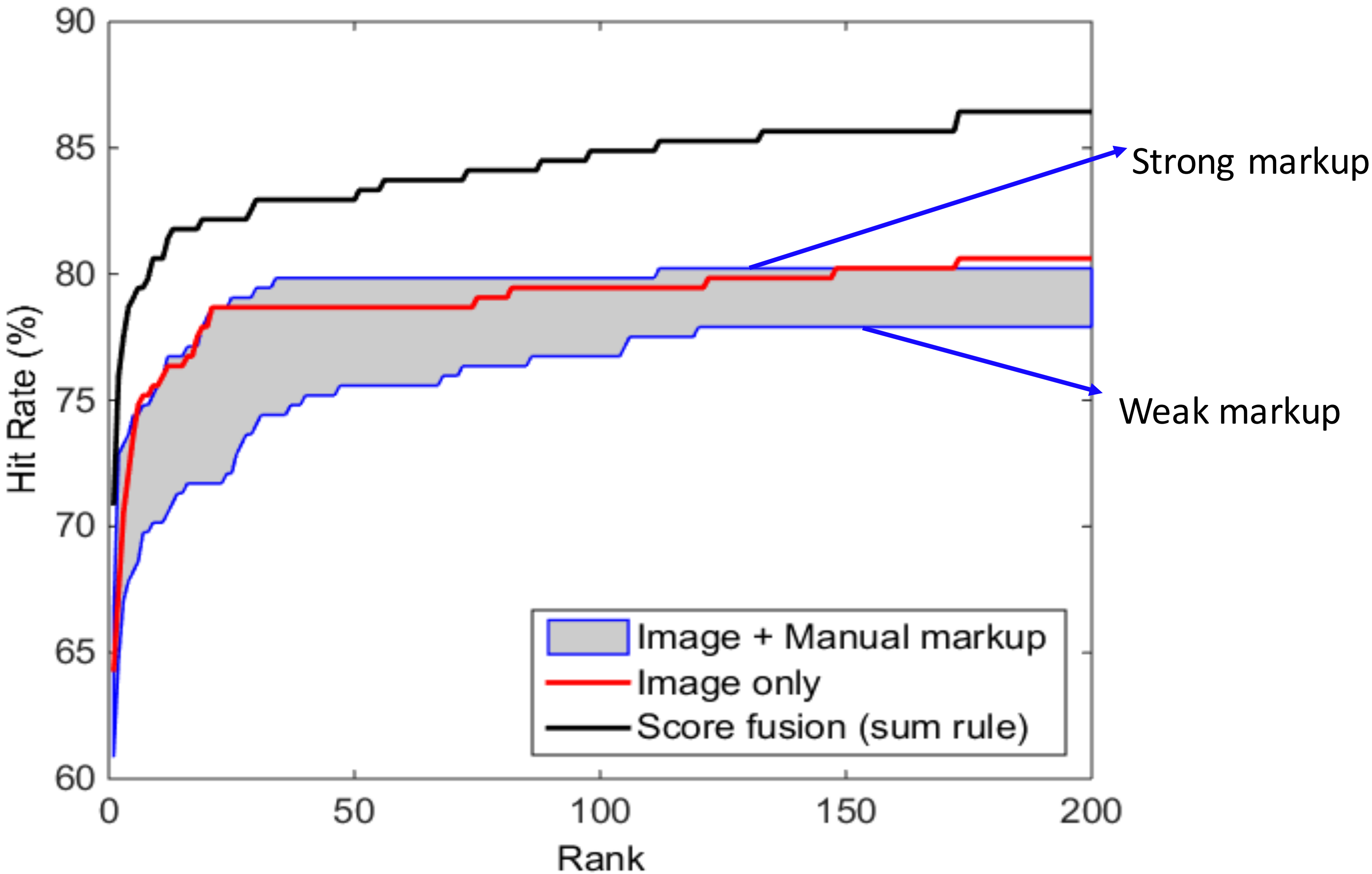


Markups by two examiners
for a latent in ELFT EFS



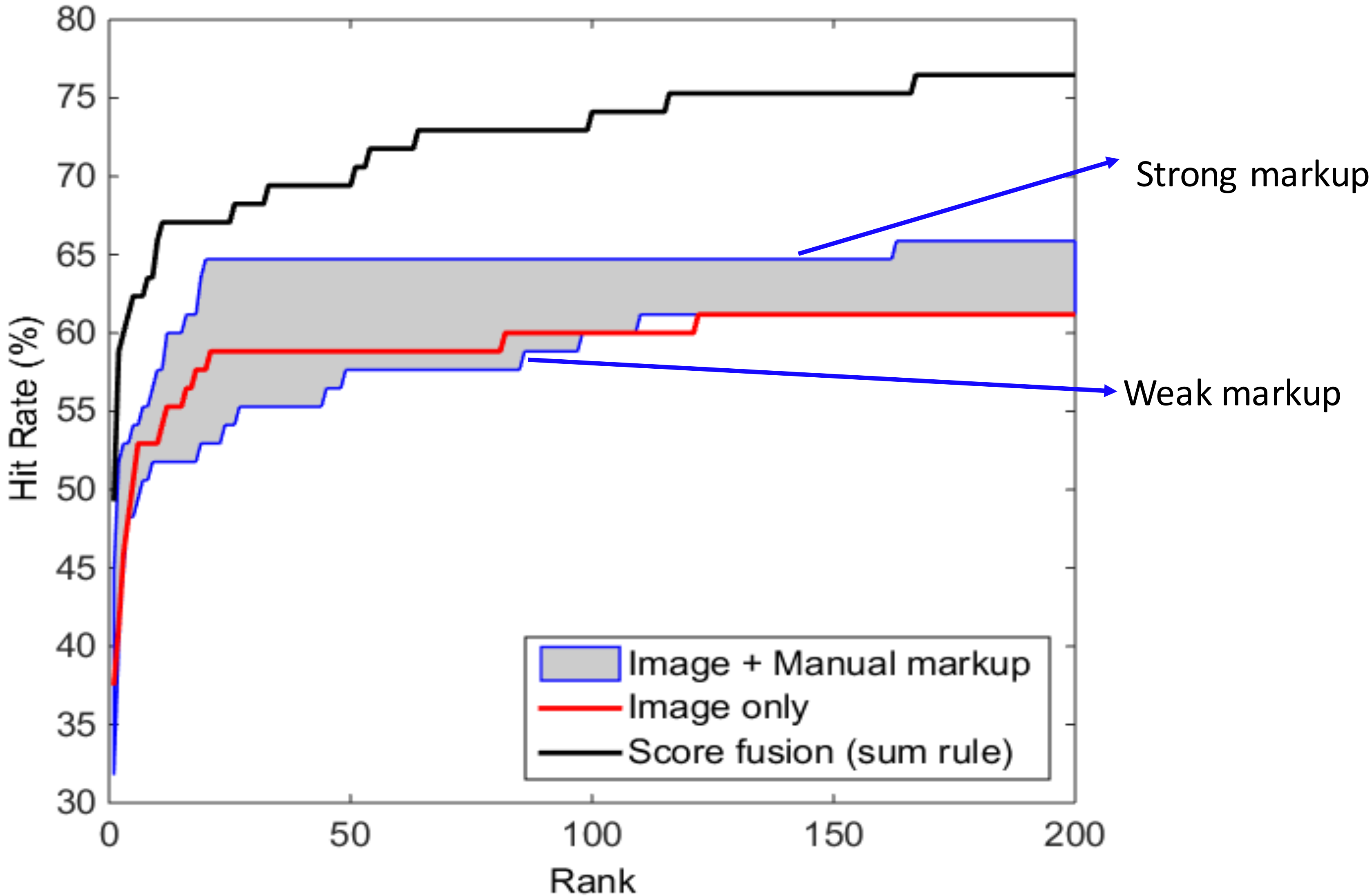
Only a single markup
available for latents in RS&A

Crowdsourcing Performance: NIST SD27



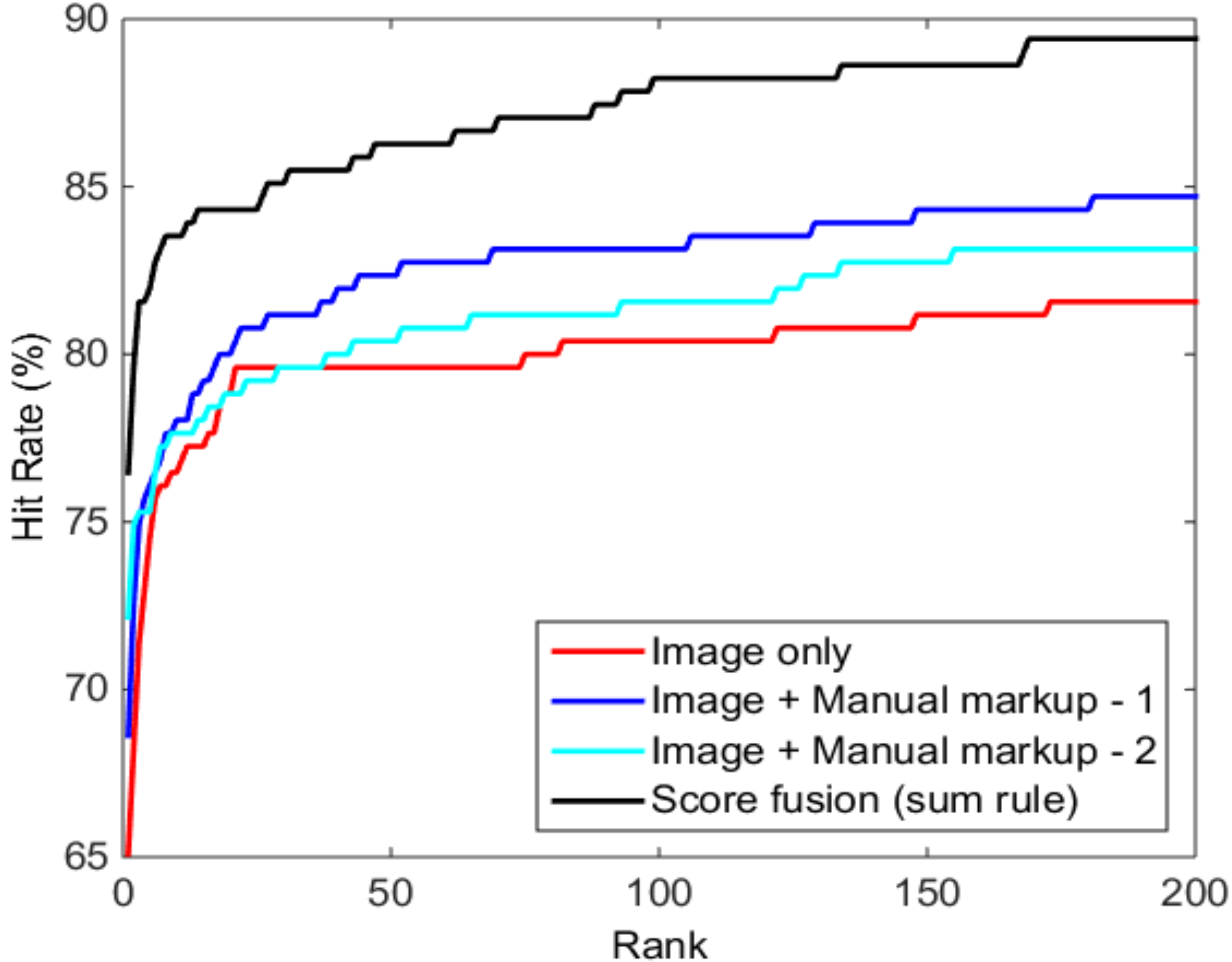
Rank-1 hit rate improves by ~7.75%

Crowdsourcing Performance: SD27 Ugly Latents



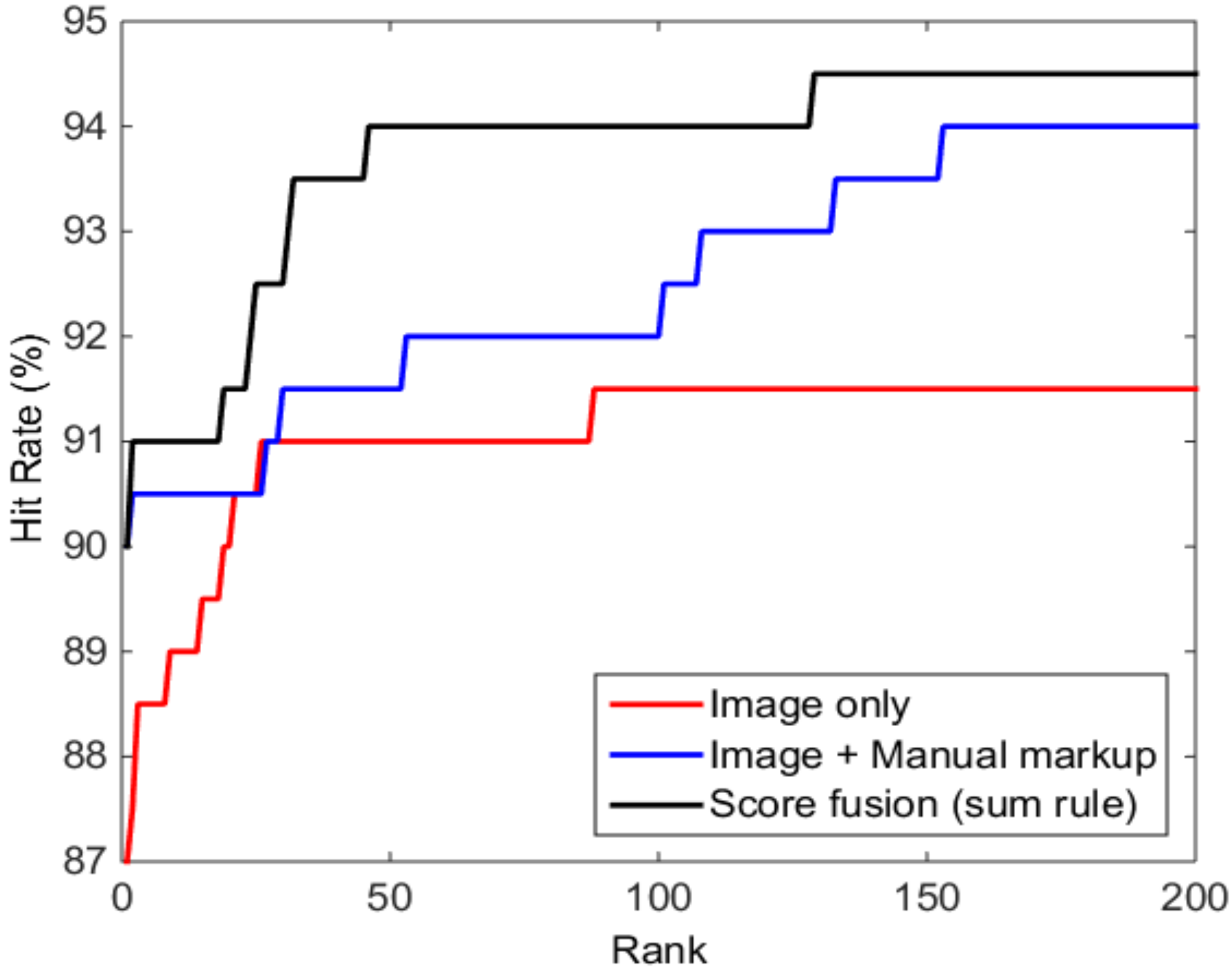
Rank-1 hit rate improves by ~12%

Crowdsourcing Performance: ELFT-EFS



Rank-1 accuracy improves by ~11.5%

Crowdsourcing Performance: RS&A



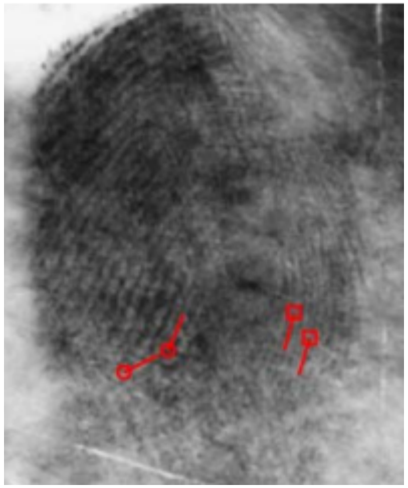
Rank-1 accuracy improves by ~2.5%

How Big a Crowd?

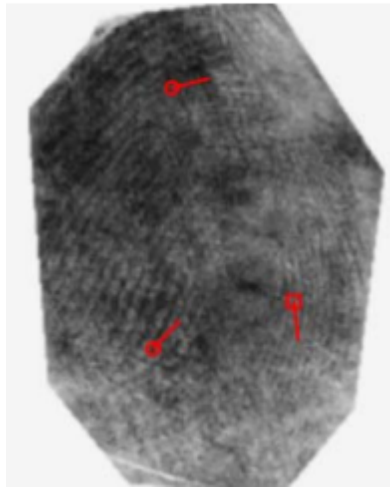
Combination	Rank-1	Rank-50	Rank-100
One examiner	63.11	77.13	78.23
Two examiners	68.04	80.88	81.96
Three examiners	69.42	82.15	83.29
Four examiners	**70.00**	**82.71**	**83.98**
Five examiners	70.80	83.14	84.56
All six examiners	70.93	82.95	84.88

Hit rates using different subsets of latent examiners

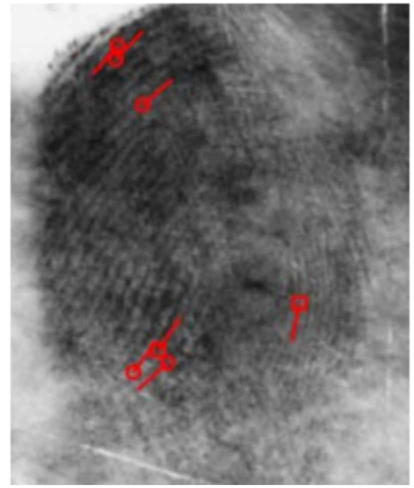
Improved Hit Rate by Crowdsourcing



Markup 1
(Rank 80)

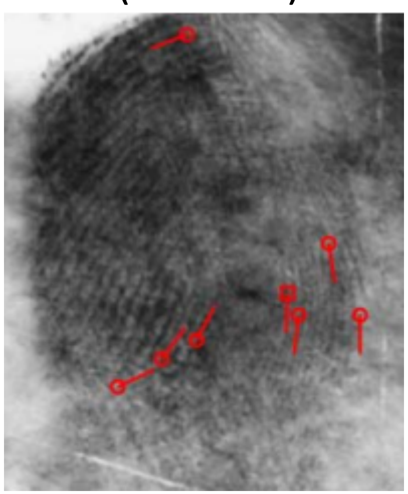


Markup 2
(Failed to match)

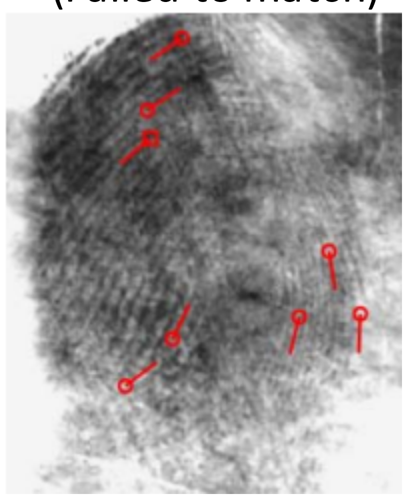


Markup 3
(Rank 45)

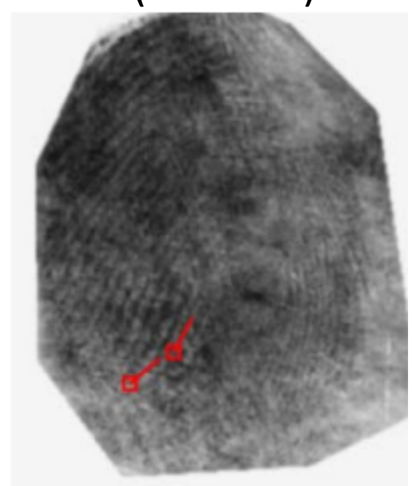
Mated Exemplar



Markup 4
(Rank 7)



Markup 5
(Rank 57)



Markup 6
(Rank 12,971)

Lights-out: Failed to match

Fusion rank: 2

NIST SD27
(Latent 236)

Improved Hit Rate by Crowdsourcing

- NIST SD27 (Latent 83)

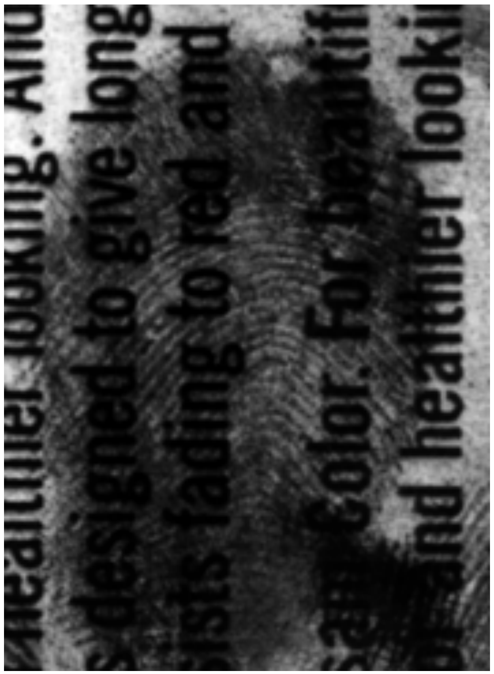


Image only

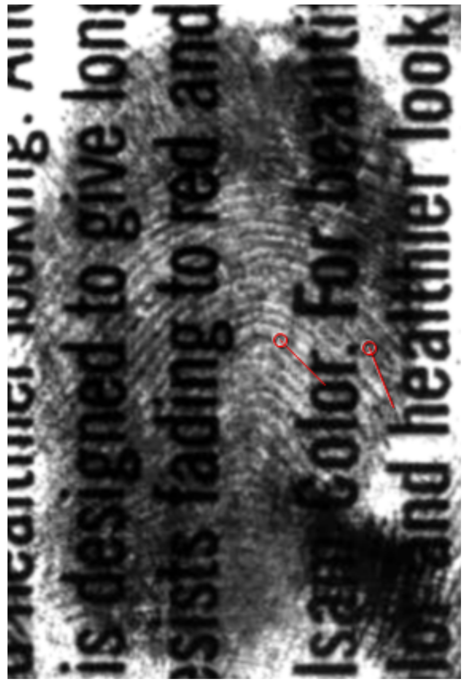


Image + Markup-E1

Mated Exemplar

	Image only	Image + Markup-E1	Fusion (All 6)
Rank	**Failed to match**	**Failed to match**	**2 (score: 226)**

Performance Decrease Example

- NIST SD27 (Latent 206)

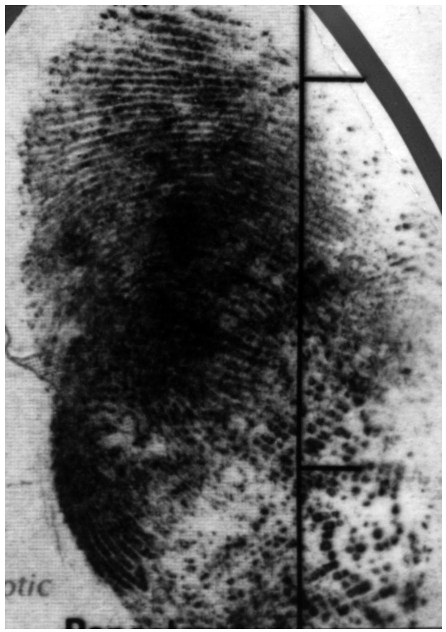


Image only

Mated Exemplar

	Image only	Fusion (All 6)
Rank	82 (score: 97)	116 (score: 411)

Summary

- Significant progress in latent AFIS accuracy; lights out capability is not yet ready
- Examiner-AFIS interaction is crucial for improved performance
- **Wisdom of crowd** leads to better decision making over a single examiner, especially for difficult latents

'COURSE MY FINGERPRINTS WERE EVERYWHERE —WODGER THINK THESE ARE — SAUSAGES?