NIST FRVT Quality Assessment: Quality Scalar

Patrick Grother, Austin Hom, **Mei Ngan**, Kayee Hanaoka

National Institute of Standards & Technology
Department of Commerce

Workshop on Face Image Quality
November 18th, 2021

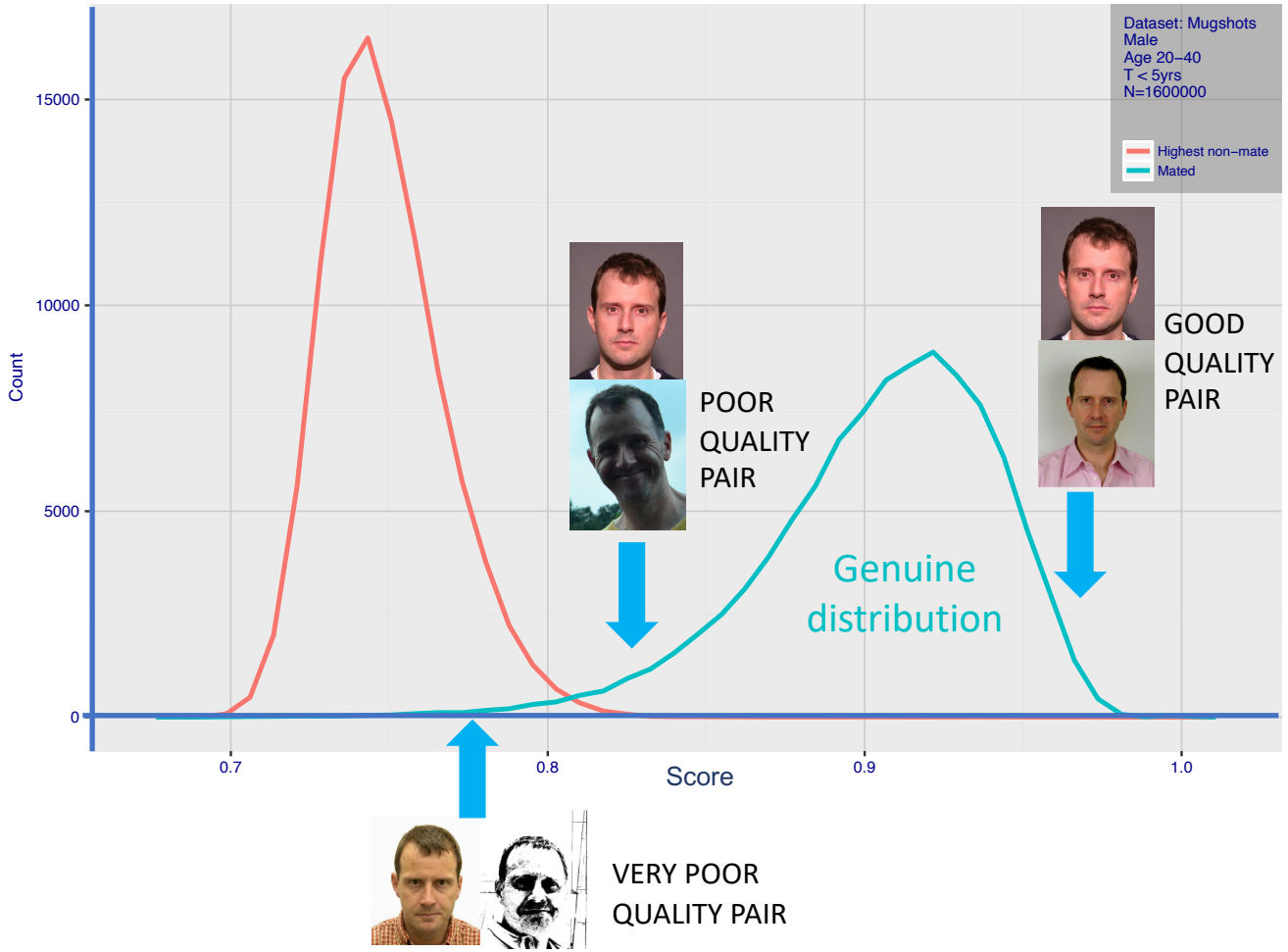
National Institute of Standards and Technology
U.S. Department of Commerce

INFORMATION TECHNOLOGY LABORATORY

Quantitative goals for quality scalar

ISO/IEC 29794-1 delineates three aspects of the umbrella term quality:

- *Character: a*n expression of quality based on the inherent properties of the source from which the biometric sample is derived
- *Fidelity*: reflects the degree of the sample similarity to its source
- *Utility*: an expression of quality based on utility reflects the predicted positive or negative contribution of an individual sample to the overall performance of a biometric system



Quality problems exist in the left tail of the genuine distribution

NIST Quality Assessment Evaluation [2019 – current]

Automated Face Image Quality Assessment

- Independent, sequestered evaluation quality assessment capabilities across large datasets
- “Black-box” testing
- Free of charge
- Ongoing testing + public reporting (report + interactive webpage)

Tracks

- Quality Scalar
- Quality Vector (coming soon…)

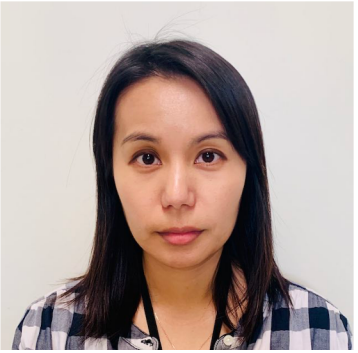
Participation

- China Electronics Import-Export Corp (CN)
- Lomonosov Moscow State University (RU)
- Paravision (US)
- Guangzhou Pixel Solutions (CN)
- Rank One Computing (US) x4
- Universidad Autónoma de Madrid + Joint Research Center (EU) x2
- Neurotechnology (LT)
- Xiamen University (CN)
- Dermalog (DE)
- Tevian (RU)

Many of these developers have also submitted recognition algorithms to FRVT 1:1

FRVT Quality draft report out for public comment (last updated: September 2021)
Ongoing quality assessment submissions accepted! Google: FRVT Quality

Quality Scalar… as predictor of true matching performance



Q = 95

Q = 90

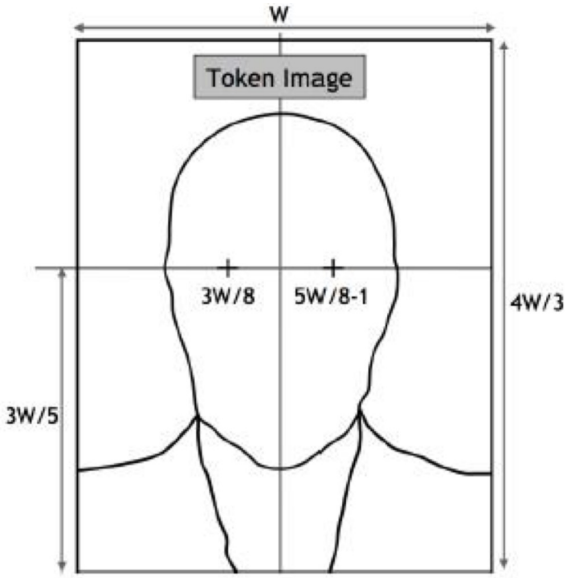
Q = 55

Q = 40

Quality scalar = $F(X_{IMAGE})$

By implicitly predicting verification outcomes of comparing X_{IMAGE} with a canonical portrait image of the same subject

Verification(X_{IMAGE}, $X_{PORTRAIT}$)

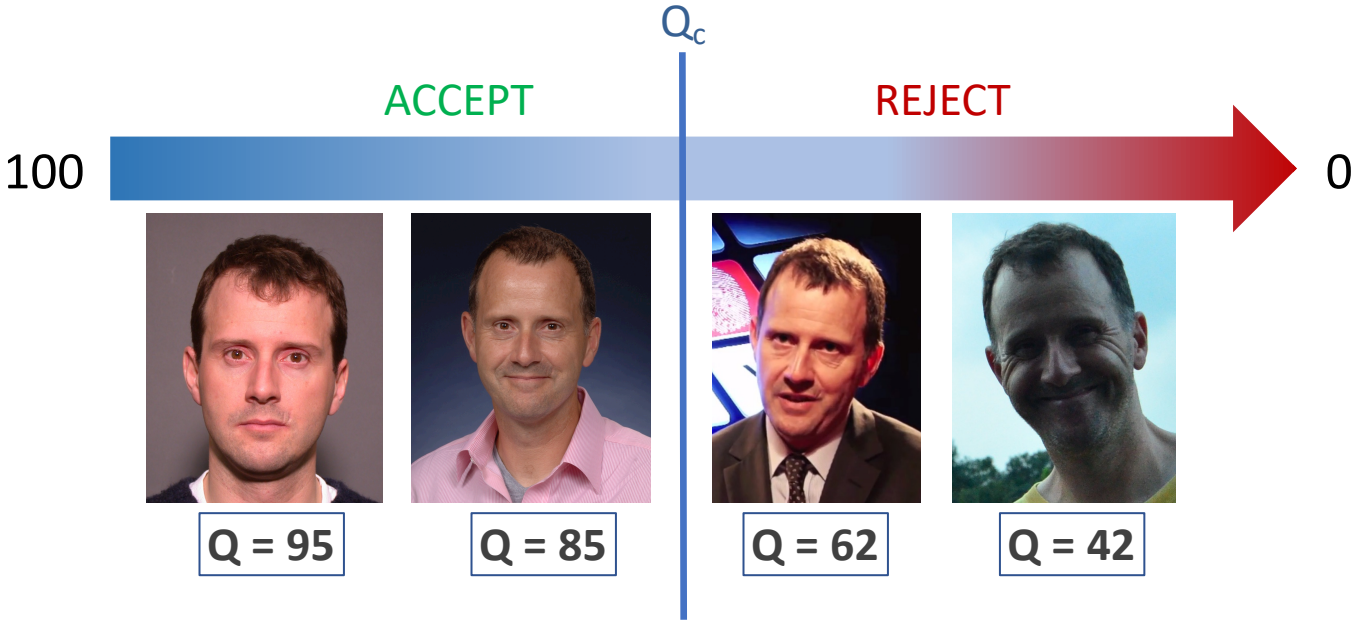


Canonical Portrait Photograph, as standardized in ISO/IEC 19794-5 (now superseded by ISO/IEC 39794-5).

Use Case: Photo Acceptance

Image acceptance / rejection decision during enrollment

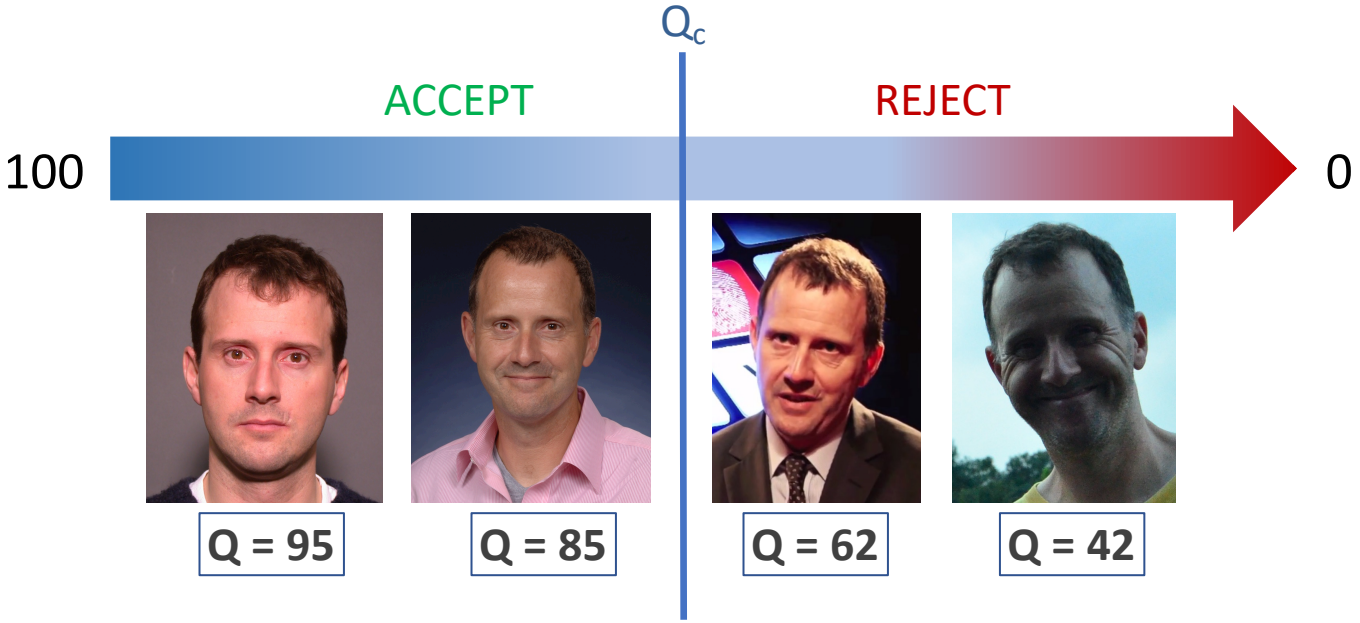
- When only **one** image is available (first encounters) or
- Matching is not possible



Use Case: Photo Acceptance

Image acceptance / rejection decision during enrollment

- When only **one** image is available (first encounters) or
- Matching is not possible



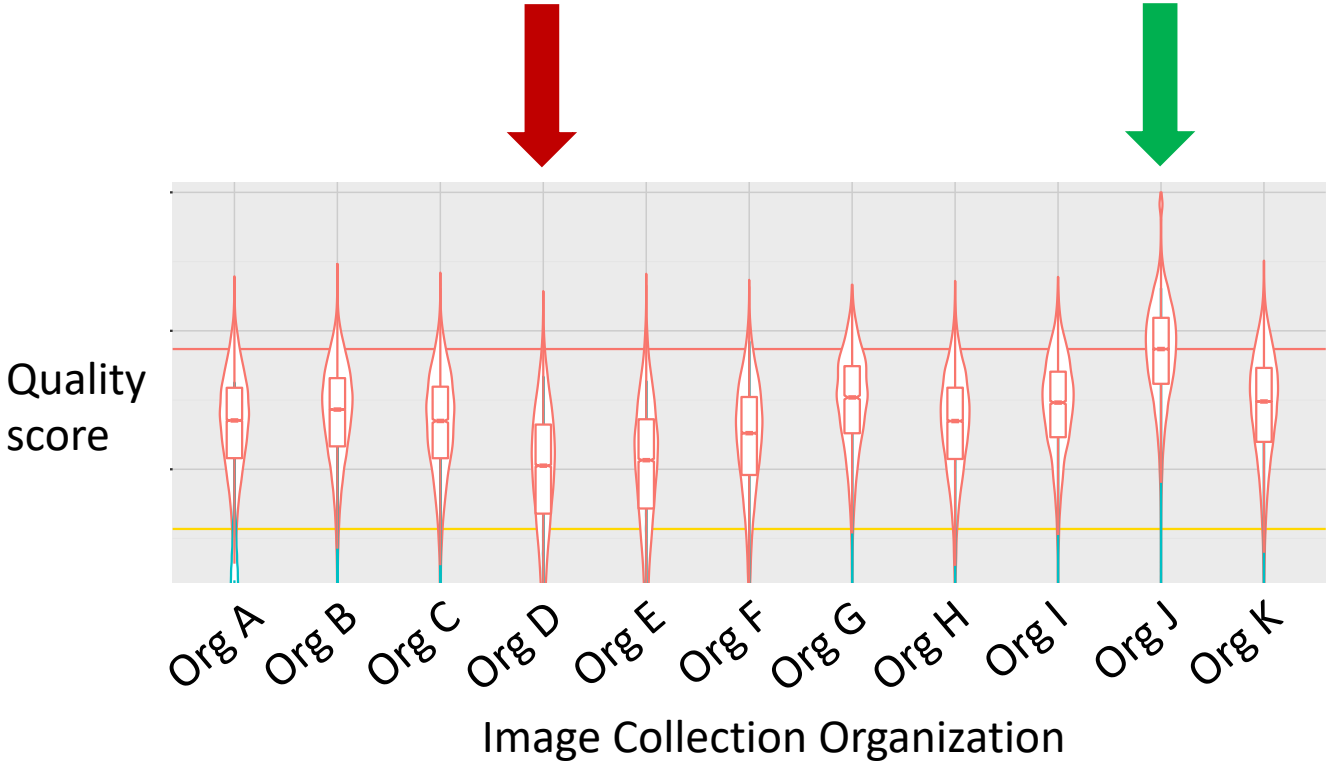
Note: The best indicator of quality is **RECOGNITION ACCURACY**

With two or more images of the person, match it against the claimed reference sample -- a match result is the ultimate quality indicator

Use Case: Quality Summarization

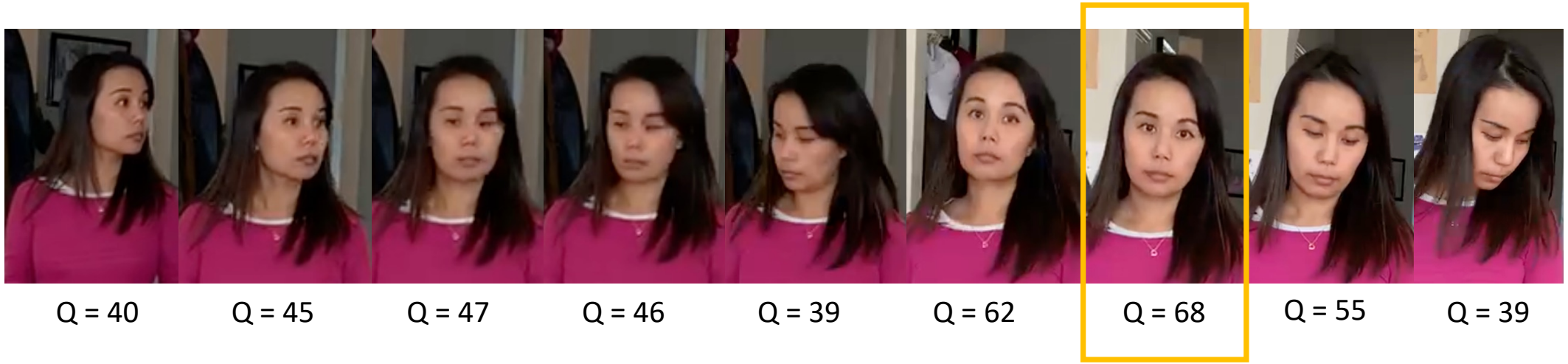
Quality as a management indicator

- Survey over large collections of images collected at certain sites or times
- Monitor a statistic over ongoing operation – time, place, camera, organization, etc.

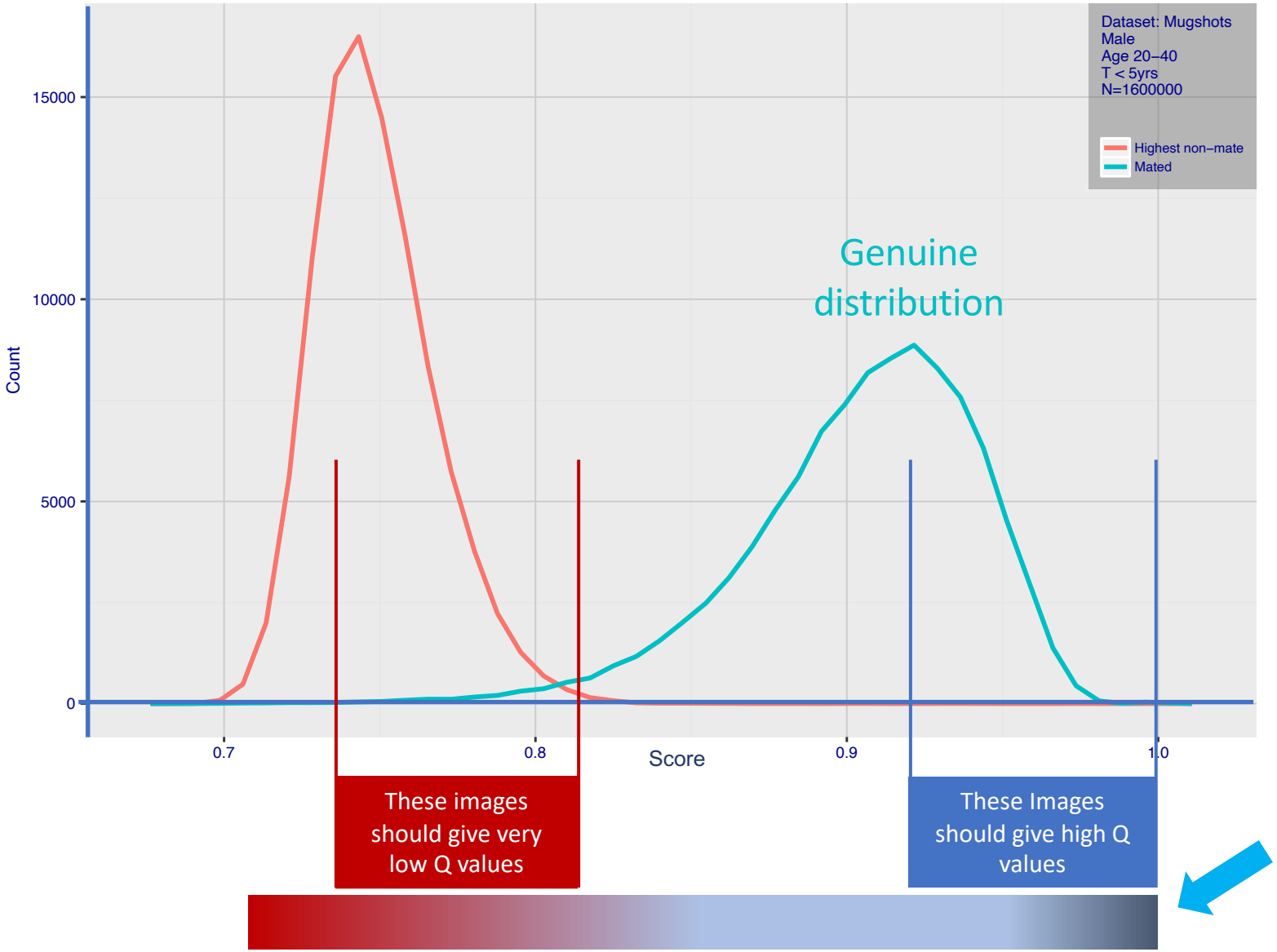


Use Case: Photo Selection from Capture Stream

Given K > 1 images of a person (e.g., from a capture stream), compute their quality values and select the best



Building a quality algorithm test set



Assign target quality scores that are continuous monotonic function of similarity scores

Quality values as predictors of FR outcomes

1. Mediocre Images

2. Quality Values from algorithm

42	57	22	48	20	39	48	55

Quality values as predictors of FR outcomes

1. Mediocre Images

2. Quality Values from algorithm

42	57	22	48	20	39	48	55

3. Pristine reference images/canonical portraits

4. Mate match scores

2.49	2.27	2.32	1.38	1.90	1.89	2.78	3.00

Quality values as predictors of FR outcomes

1. Mediocre Images								
2. Quality Values from algorithm	42	57	22	48	20	39	48	55
3. Pristine reference images								
4. Mate match scores	2.49	2.27	2.32	1.38	1.90	1.89	2.78	3.00
5. Match? Match score threshold = 2.0	Y	Y	Y	N	N	N	Y	Y

Can Q **predict** score?

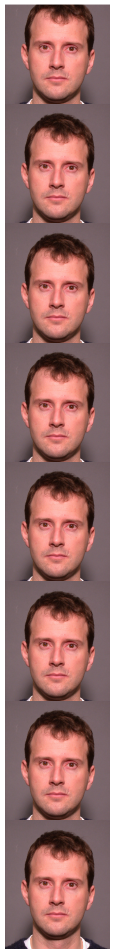
Quality values as predictors of FR outcomes

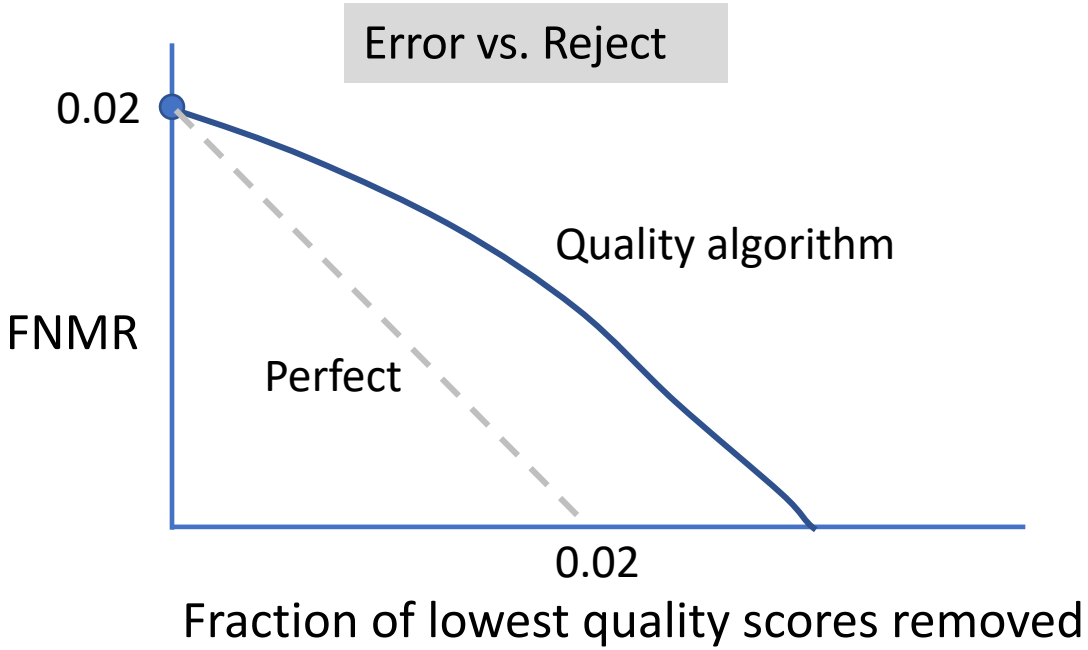
1. Mediocre Images								
2. Quality Values from algorithm	42	57	22	48	20	39	48	55
3. Pristine reference images								
4. Mate match scores	2.49	2.27	2.32	1.38	1.90	1.89	2.78	3.00
5. Match? Match score threshold = 2.0	Y	Y	Y	N	N	N	Y	Y
			Low Q but high score	High Q but low score	Low Q + low score			

Can Q **predict** score?

6. Compute "prediction failures" for many Q values

Quality values as predictors of genuine scores – Error vs. Reject

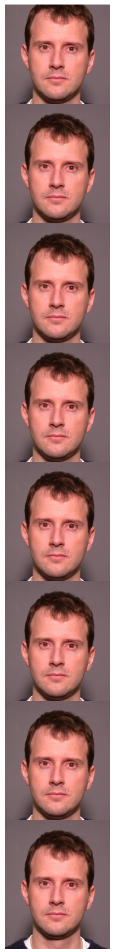
	Quality algorithm		Recognition algorithm	
Ref	Probe	Quality	MATCH SCORE	MATCH YES?
		97	0.97	TRUE
		94	0.91	TRUE
		74	0.60	FALSE
		68	0.85	TRUE
		57	0.81	TRUE
		32	0.72	FALSE
		29	0.90	TRUE
		27	0.57	FALSE



The matching threshold is set to give, for example, FNMR = 0.02 i.e. lowest 2 percent of mate scores

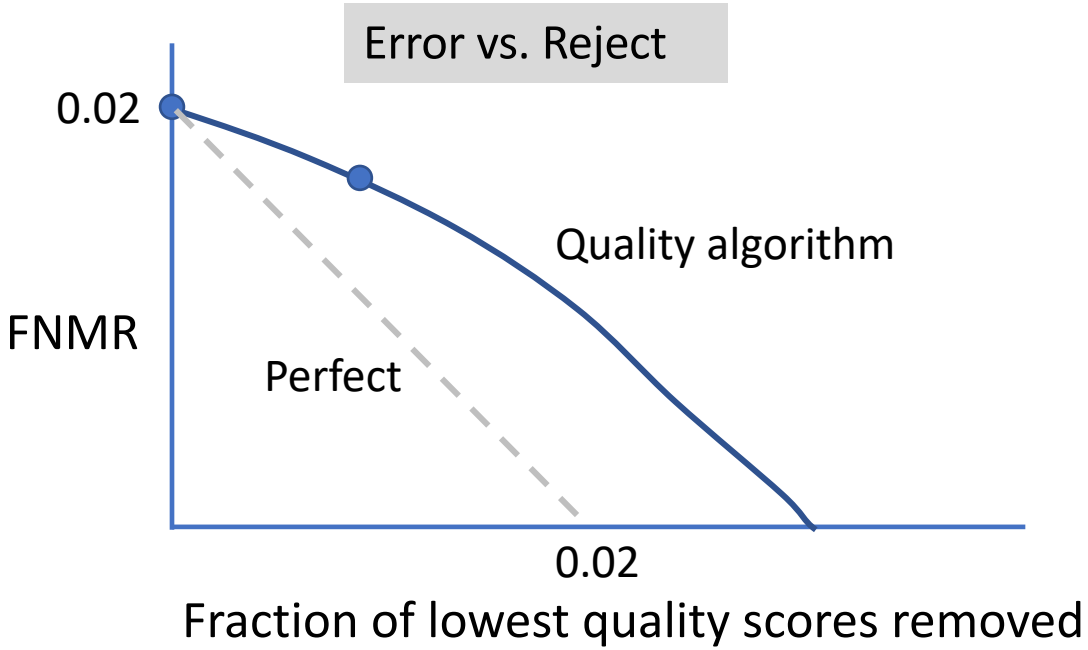
Quality values as predictors of genuine scores – Error vs. Reject

Quality algorithm

Ref	Probe	Quality
		97
		94
		74
		68
		57
		32
		29
		27

Recognition algorithm

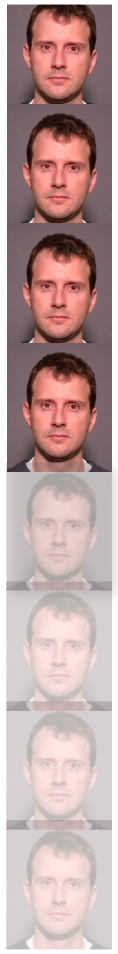
MATCH SCORE	MATCH YES?
0.97	TRUE
0.91	TRUE
0.60	FALSE
0.85	TRUE
0.81	TRUE
0.72	FALSE
0.90	TRUE
0.57	FALSE



Discard n-% of lowest quality probes

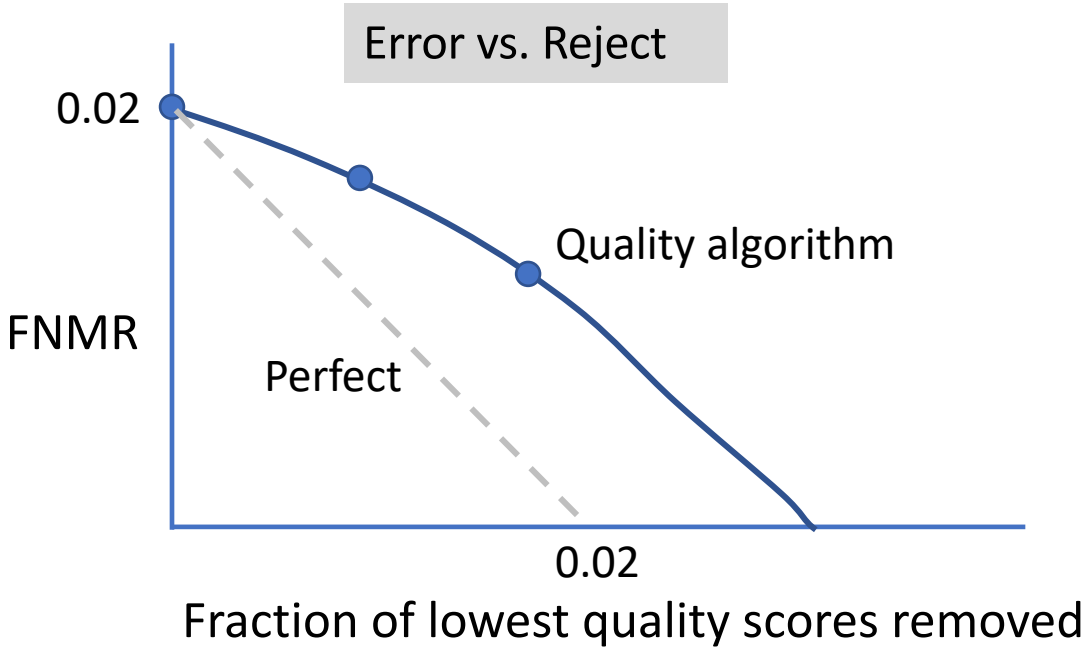
Quality values as predictors of genuine scores – Error vs. Reject

Quality algorithm

Ref	Probe	Quality
		97
		94
		74
		68
		57
		32
		29
		27

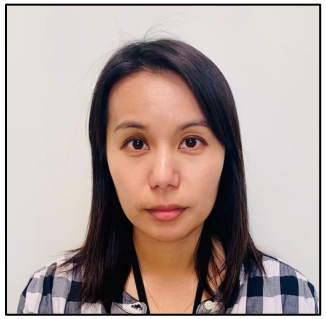
Recognition algorithm

MATCH SCORE	MATCH YES?
0.97	TRUE
0.91	TRUE
0.60	FALSE
0.85	TRUE
0.81	TRUE
0.72	FALSE
0.90	TRUE
0.57	FALSE



FNMR is ideally reduced as quality algorithm is used to discard low quality probes

Error vs. Reject - quality algorithm performance against target FR matchers



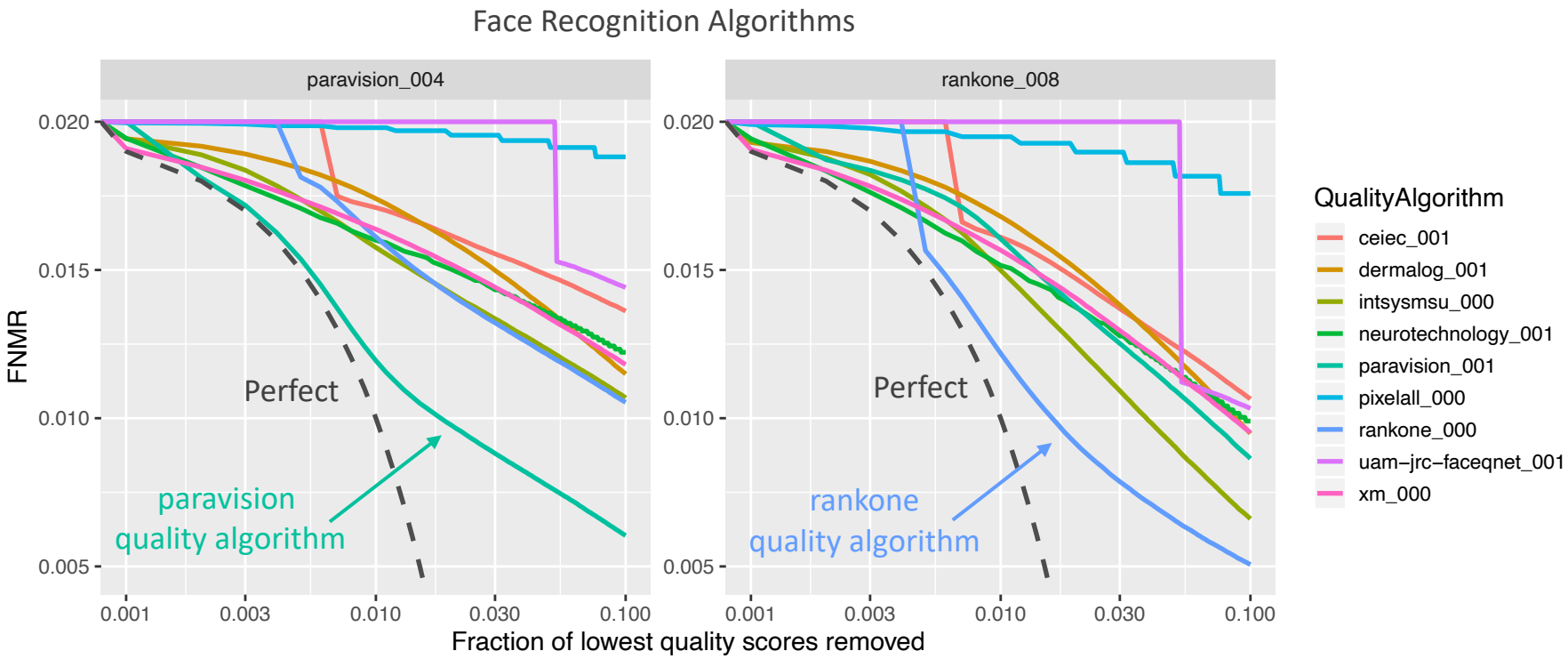
visa-like reference photo

webcam probe

Mate scores are from comparison of high quality visa–like application photos with medium quality webcam photos (3 225 633 genuine scores)

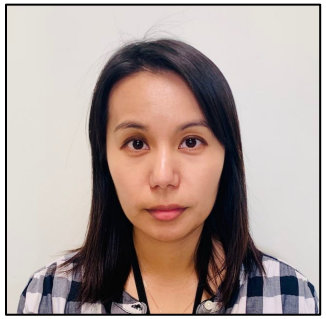
Matching threshold set to give FNMR = 0.02 i.e. lowest 2% of mate scores

Quality is computed on the webcam photos (5 225 633 images)



Some developers can predict false negative decisions produced by their respective face recognition algorithms

Error vs. Reject - quality algorithm performance against target FR matchers



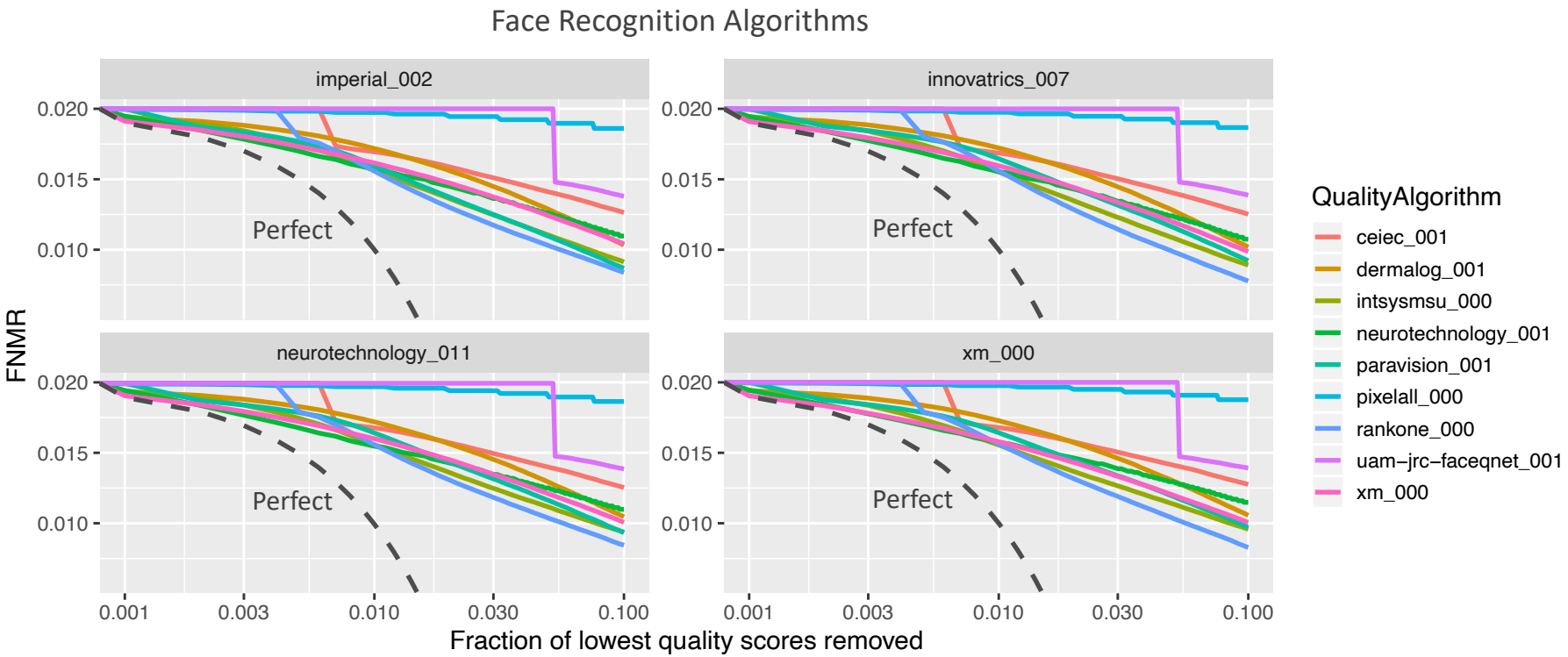
visa-like reference photo

webcam probe

Mate scores are from comparison of high quality visa–like application photos with medium quality webcam photos (3 225 633 genuine scores)

Matching threshold set to give FNMR = 0.02 i.e. lowest 2% of mate scores

Quality is computed on the webcam photos (5 225 633 images)



Current quality algorithms are **not** effective at predicting false negative decisions across **different** developer face recognition algorithms

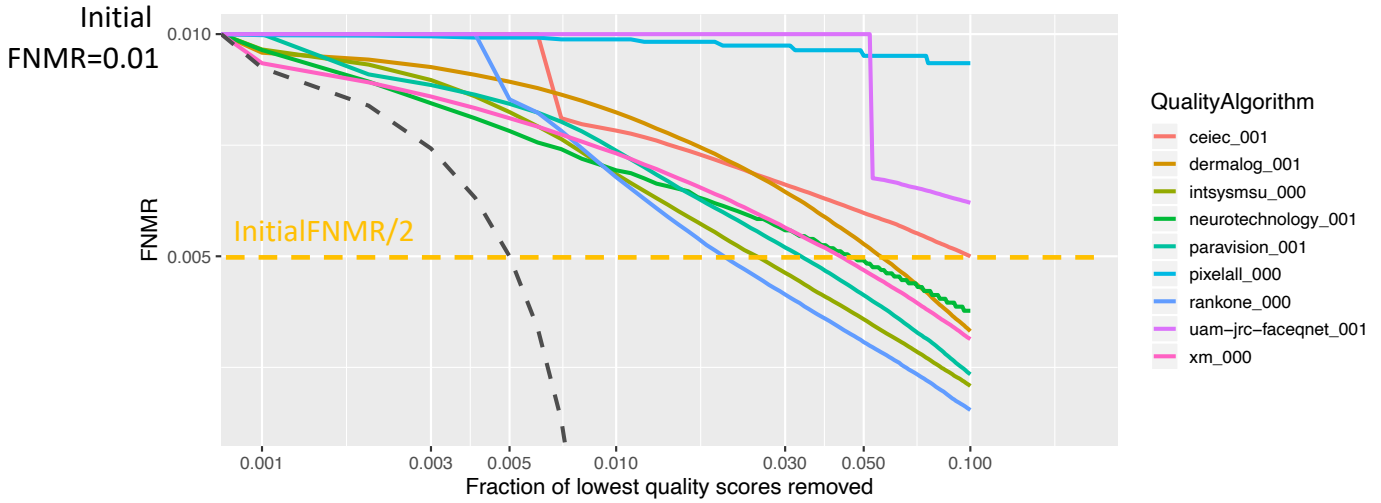
Error vs. Reject - quality performance against consensus across multiple FR matchers

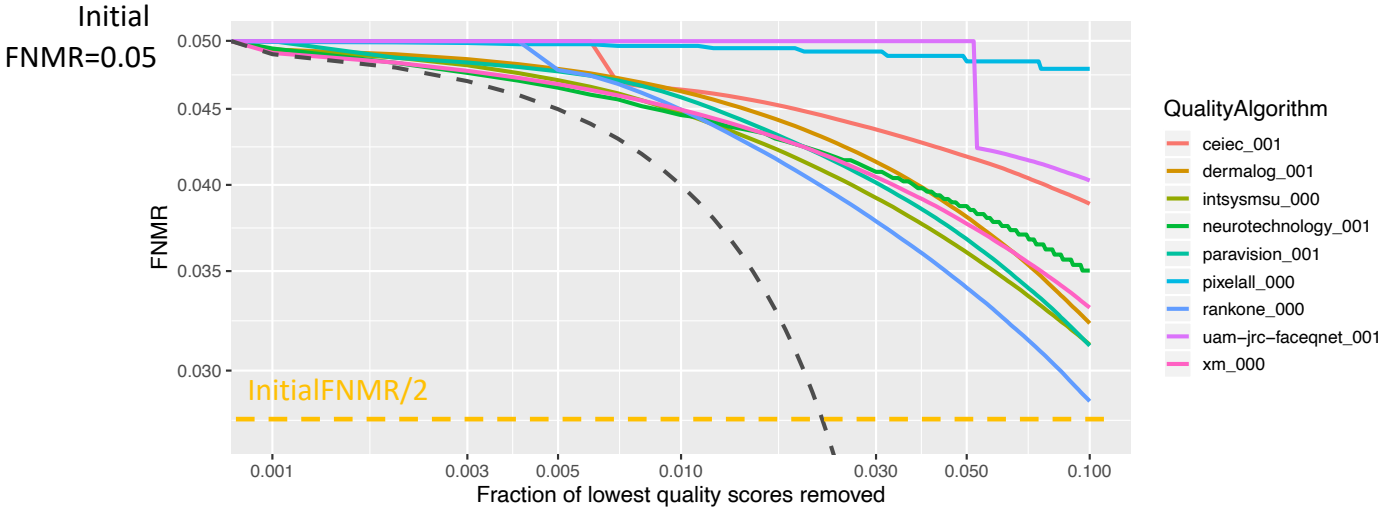
Ground truth for quality is set as false negatives from 33 face recognition algorithms

Consensus approach gives more weight to the problematic images generally

Algorithms are more effective when detecting the least recognizable images

Consensus across 33 face recognition algorithms



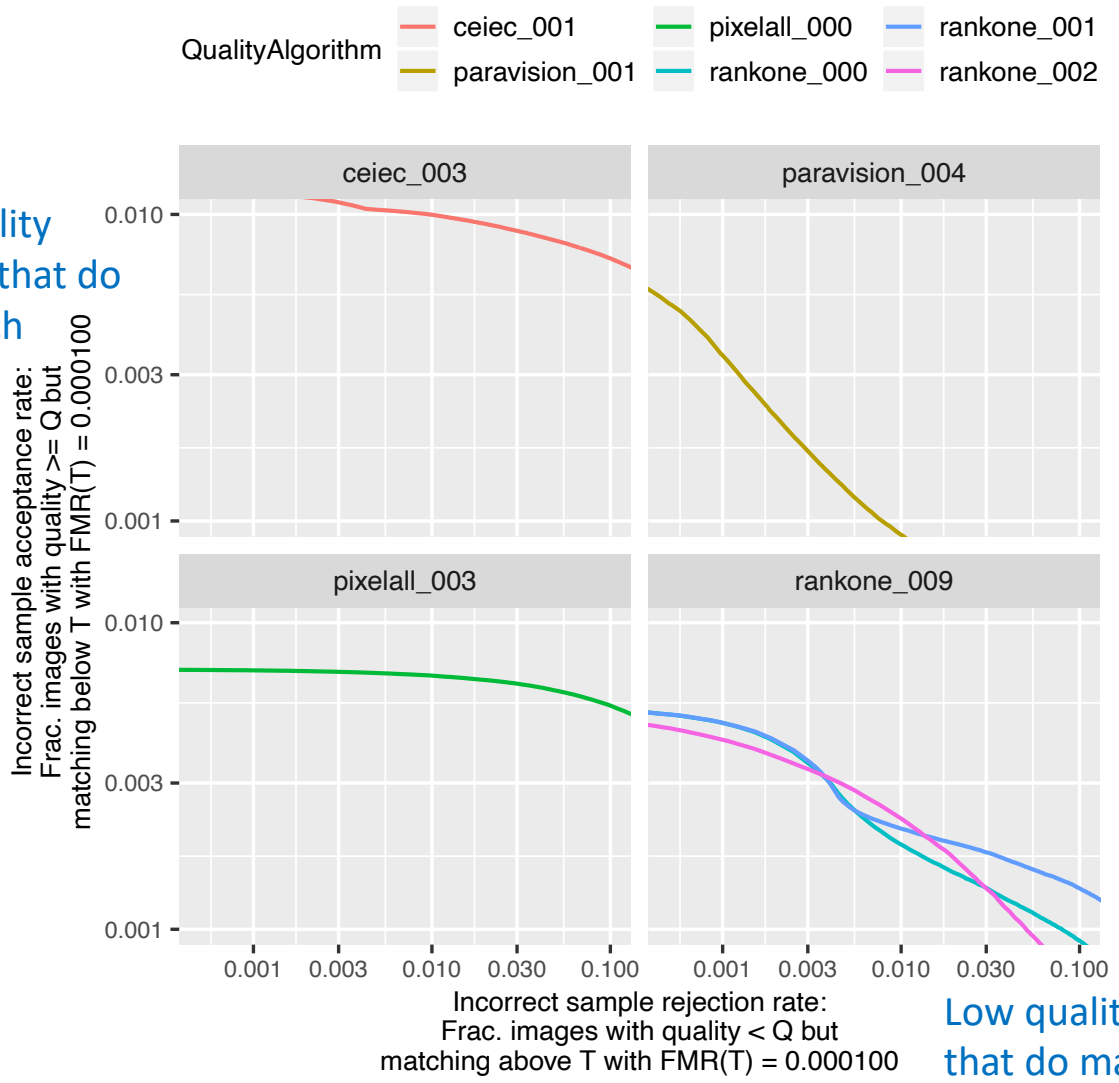


Quality measurement for use in photo acceptance - Sample acceptance error tradeoff (ISAR vs. ISRR)

- **Incorrect sample acceptance rate (ISAR):** assignment of high quality to photo when it ultimately gives a false negative in recognition
- **Incorrection sample rejection rate (ISRR):** assignment of low quality when the image would be matched by an FR engine correctly

- Good for understanding operational deployment benefits

High quality samples that do **not** match



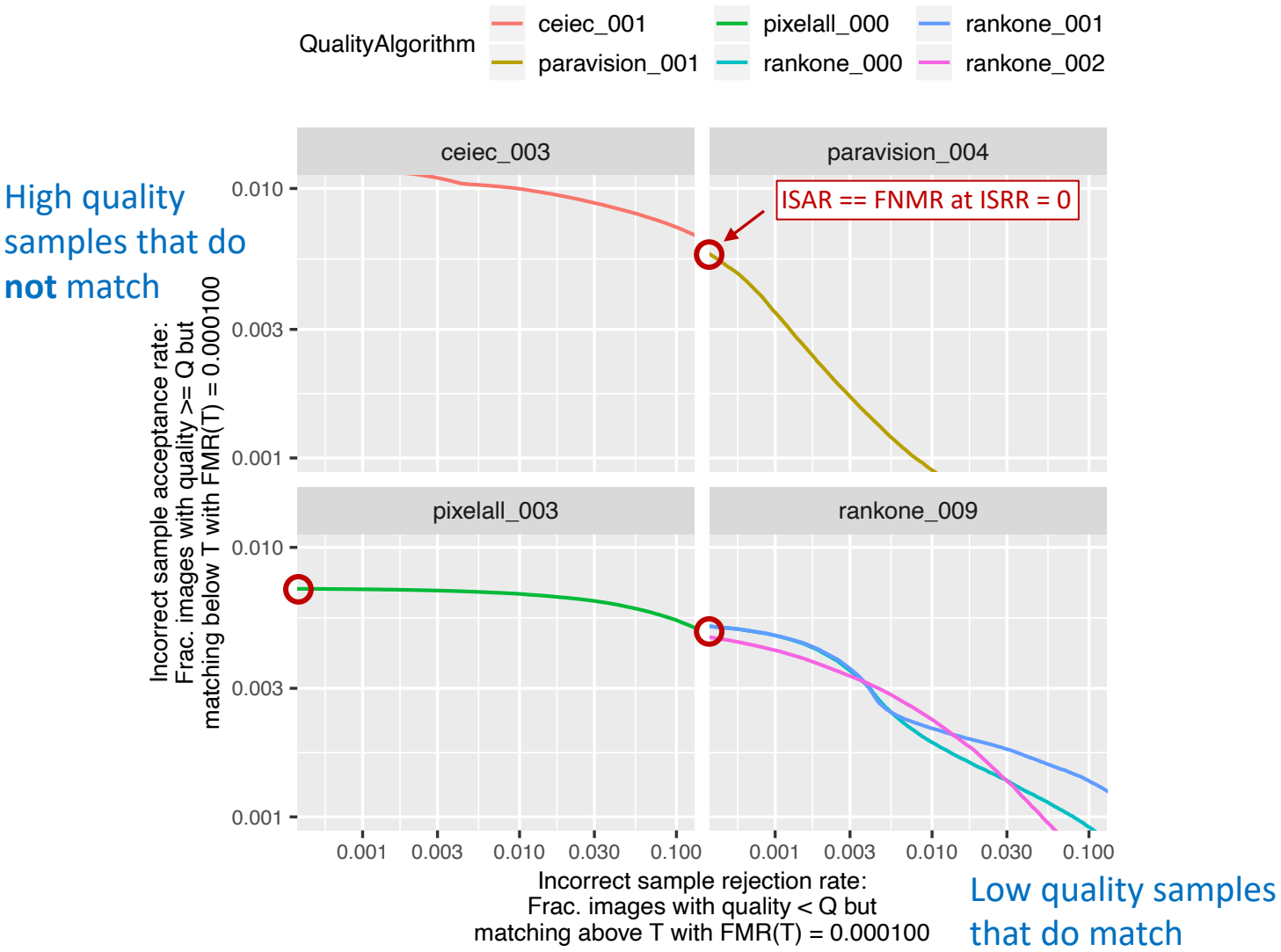
Low quality samples that do match

Quality measurement for use in photo acceptance - Sample acceptance error tradeoff (ISAR vs. ISRR)

- **Incorrect sample acceptance rate (ISAR):** assignment of high quality to photo when it ultimately gives a false negative in recognition
- **Incorrection sample rejection rate (ISRR):** assignment of low quality when the image would be matched by an FR engine correctly

- Good for understanding operational deployment benefits

  - ISAR at ISRR = 0 is essentially your operational FNMR **without** deployment of a quality algorithm

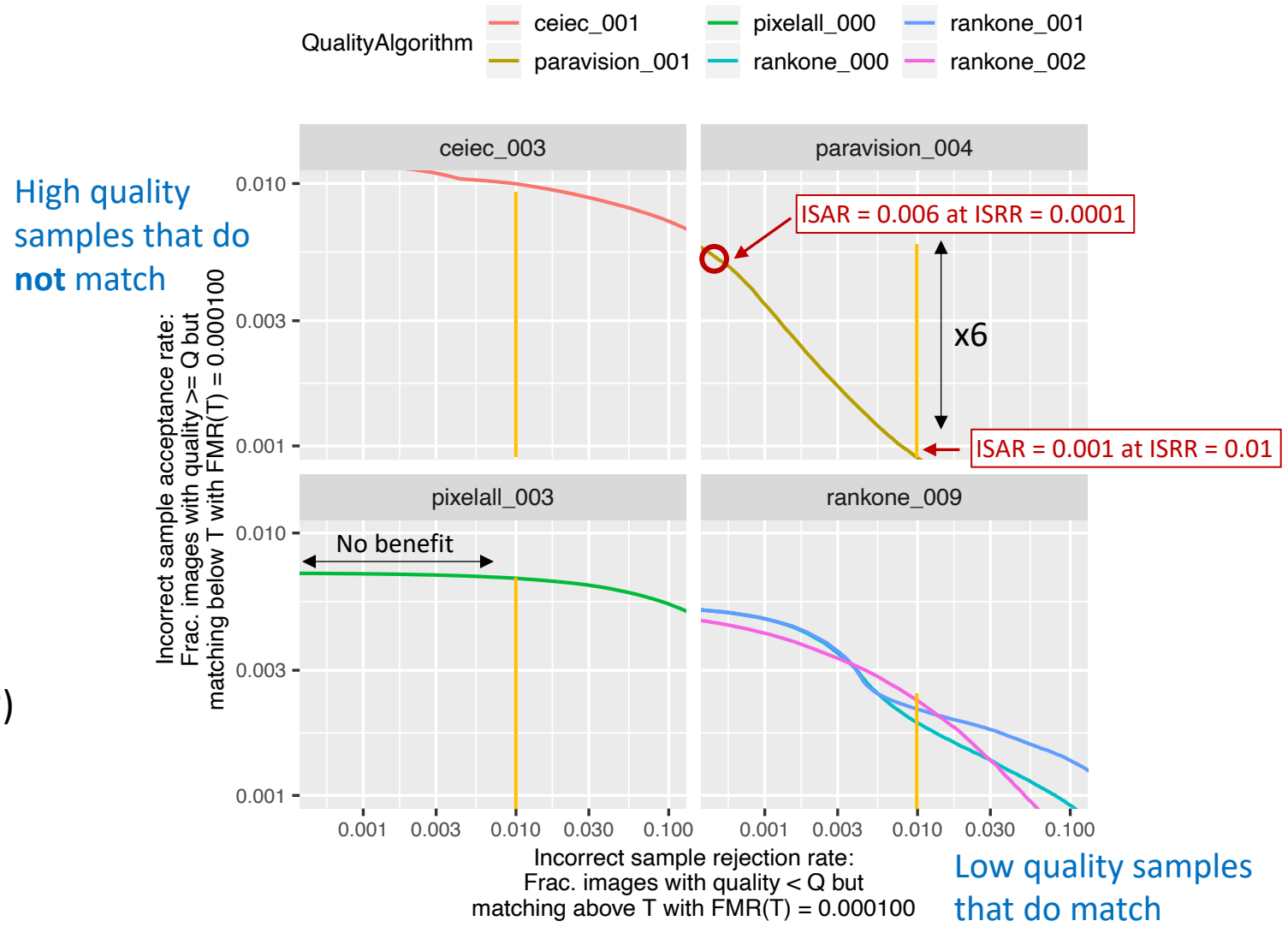


Quality measurement for use in photo acceptance - Sample acceptance error tradeoff (ISAR vs. ISRR)

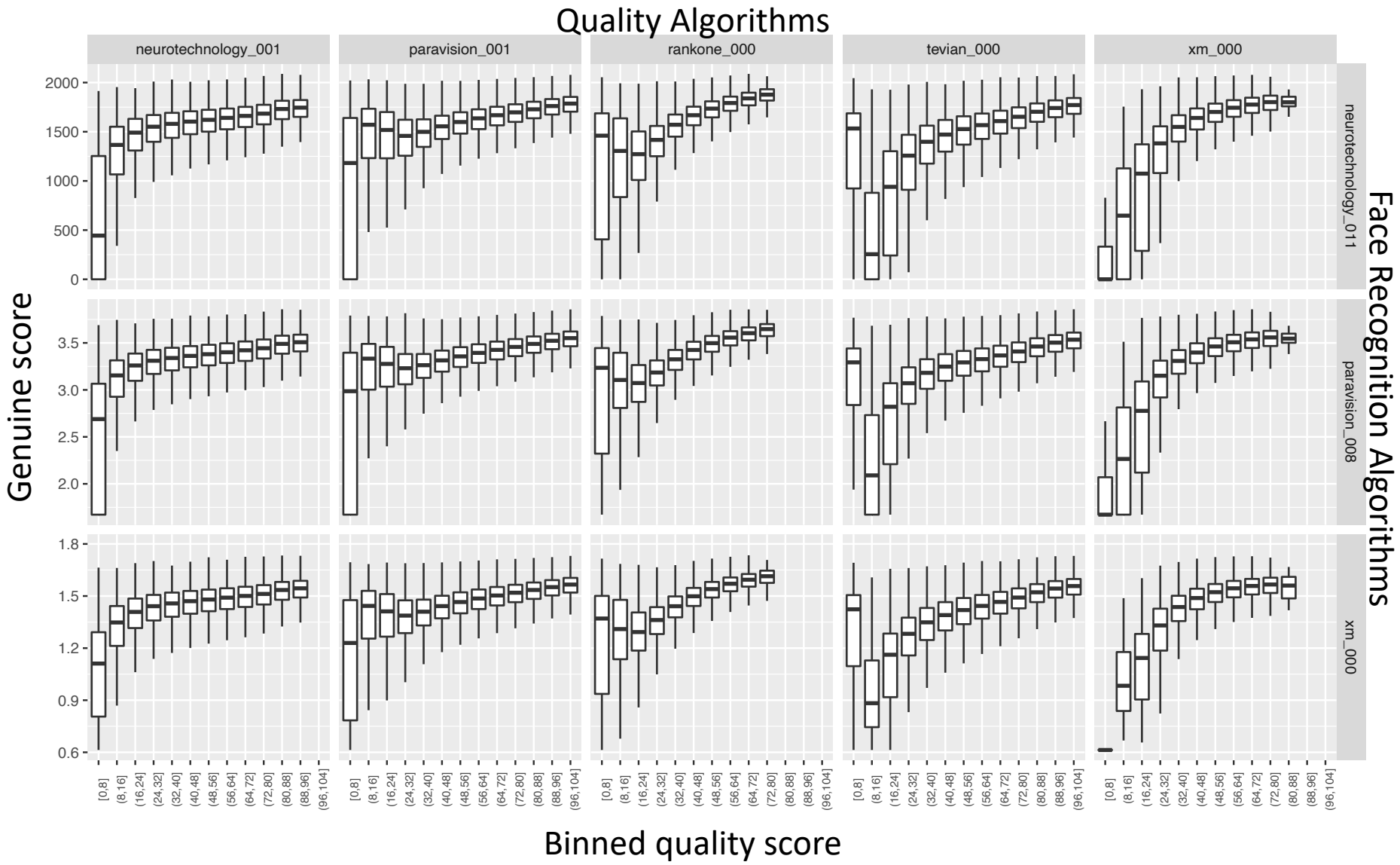
- **Incorrect sample acceptance rate (ISAR):** assignment of high quality to photo when it ultimately gives a false negative in recognition
- **Incorrection sample rejection rate (ISRR):** assignment of low quality when the image would be matched by an FR engine correctly

- Good for understanding operational deployment benefits

ISRR = 0.01 - is it an operationally usable value (?)

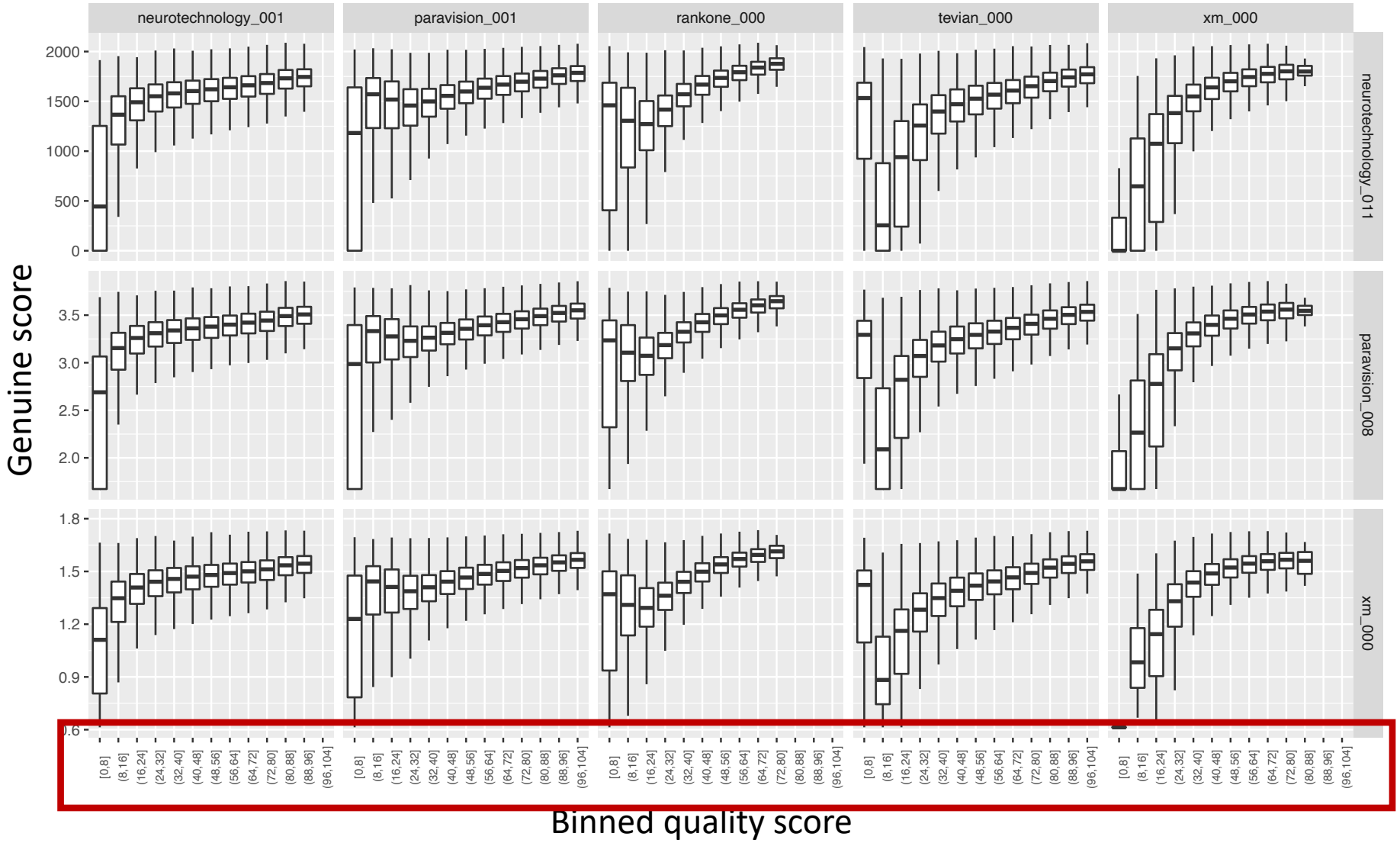


Quality measurement for use as “summary indicator”



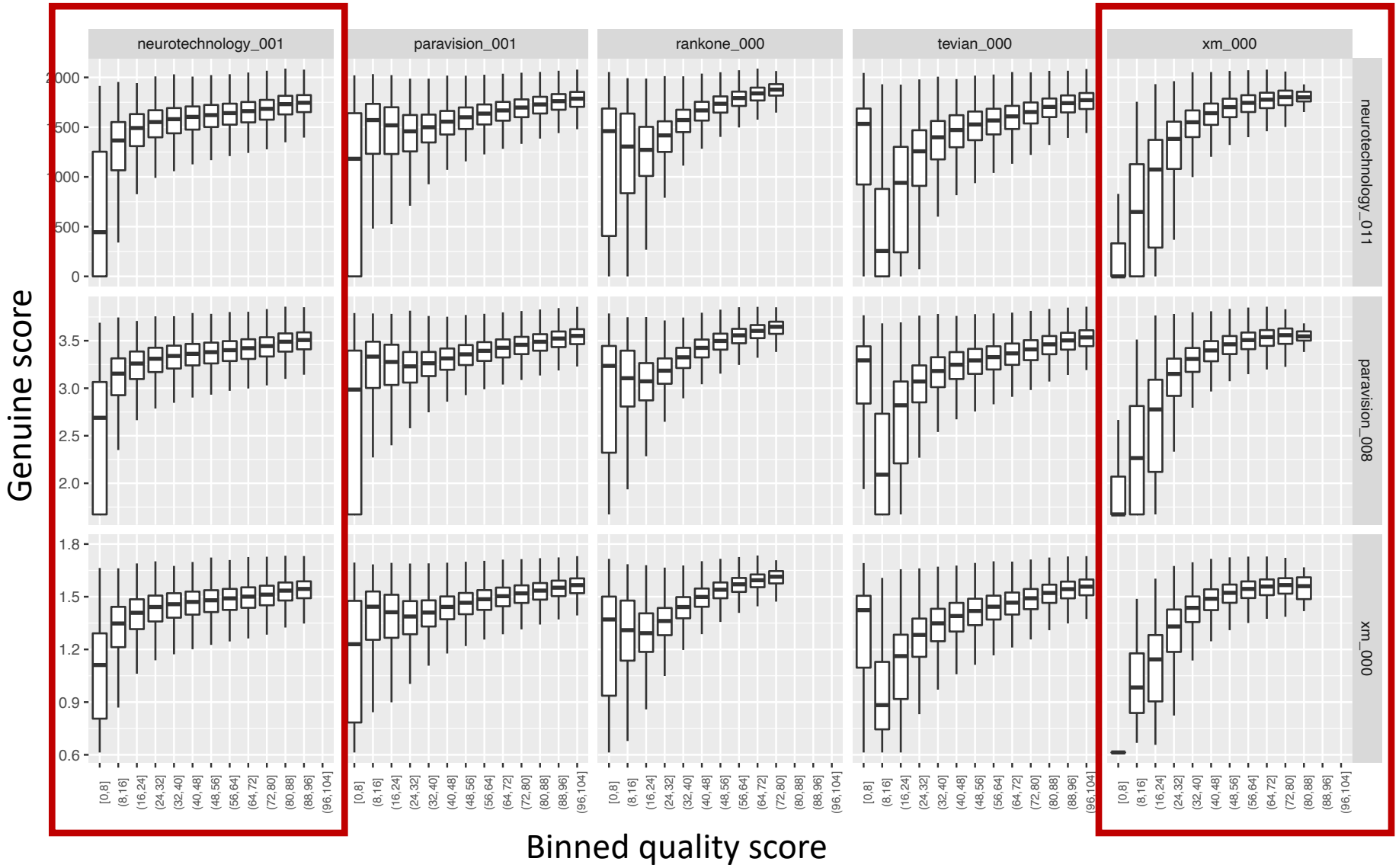
- Quality score binned to 13 levels
- Monotonic medians
- Variance is often high
- Within- vs. cross-developer

Quality measurement for use as “summary indicator”



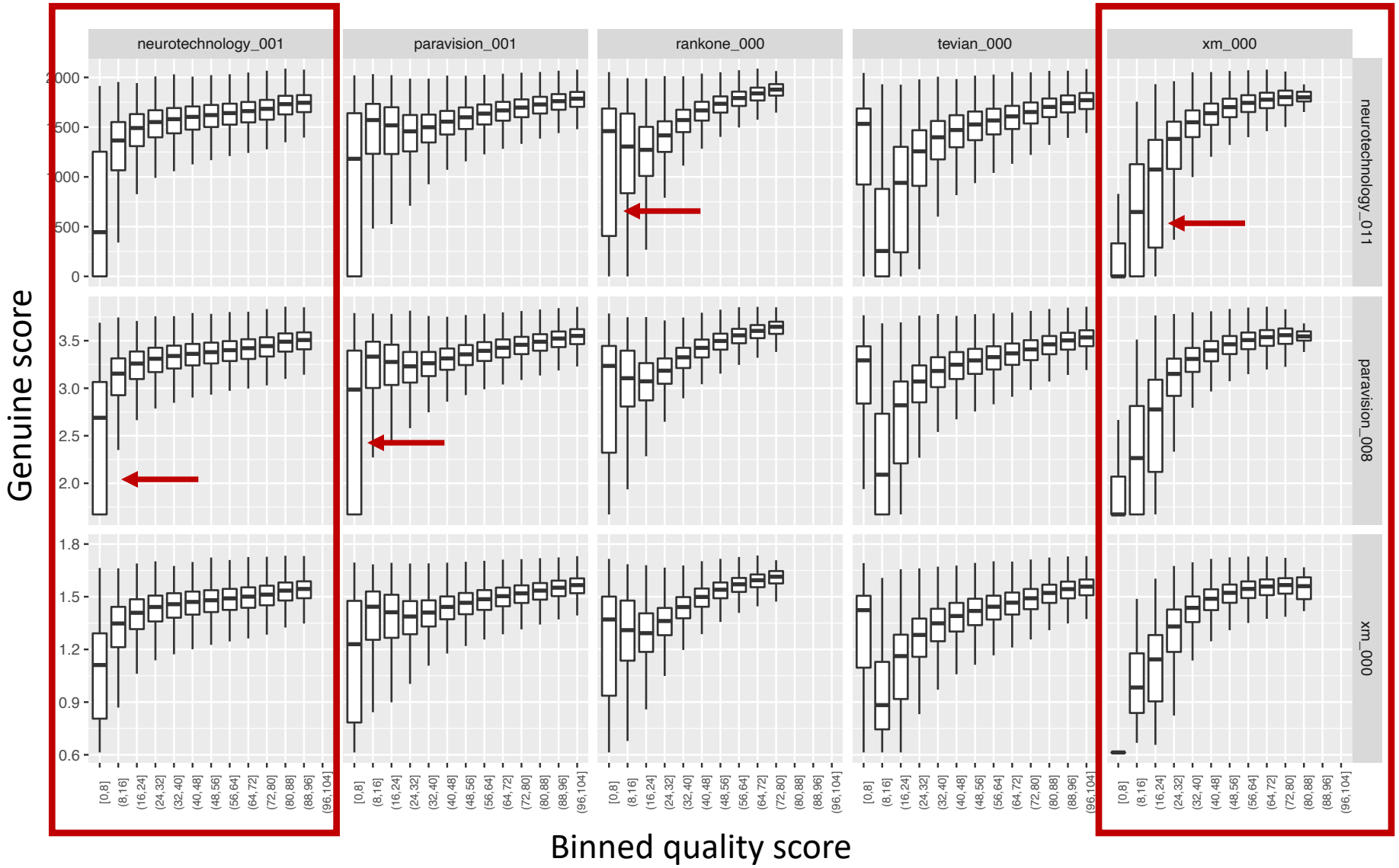
- Quality score binned to 13 levels
- Monotonic medians
- Variance is often high
- Within- vs. cross-developer

Quality measurement for use as “summary indicator”



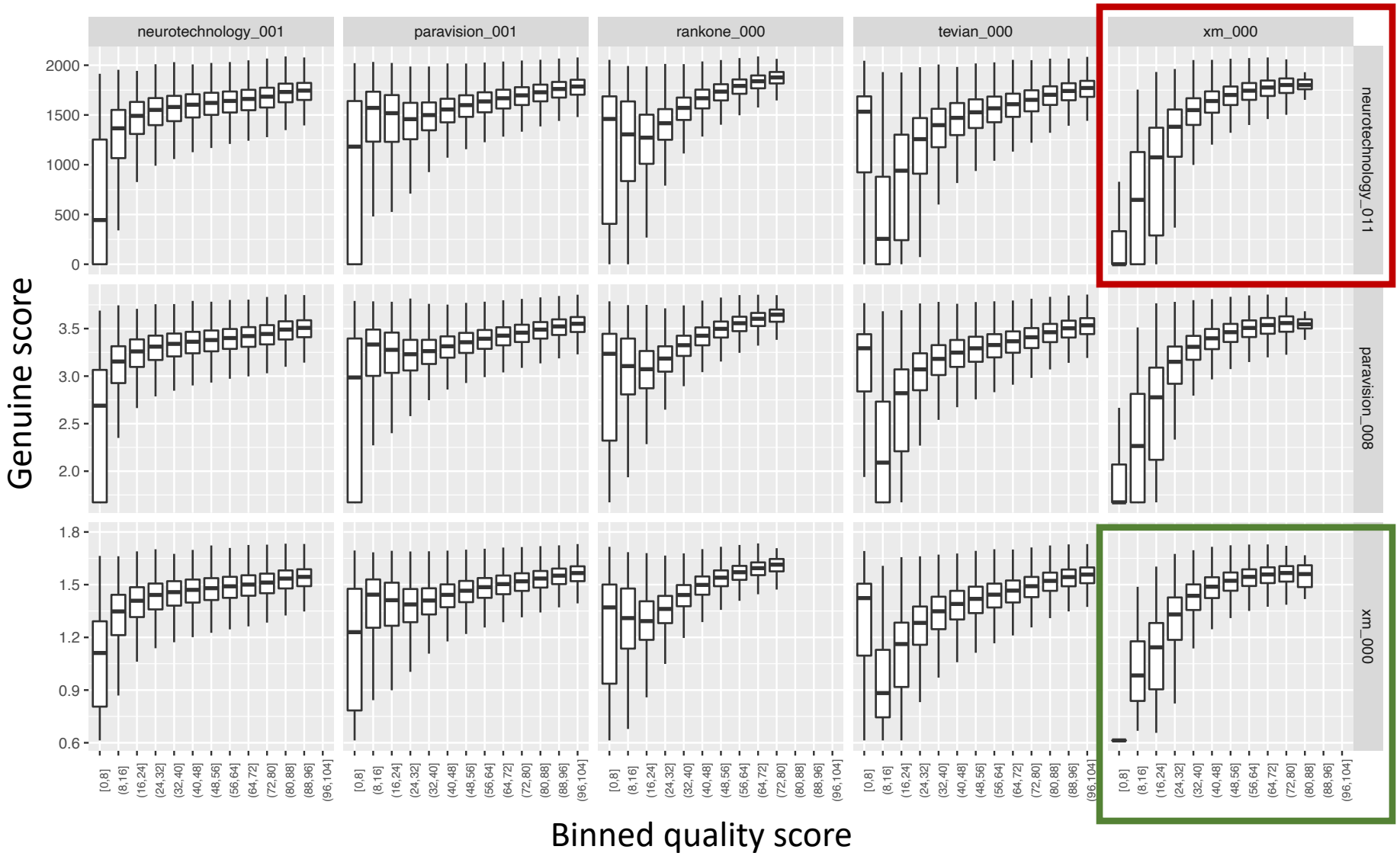
- Quality score binned to 13 levels
- Monotonic medians
- Variance is often high
- Within- vs. cross-developer

Quality measurement for use as “summary indicator”



- Quality score binned to 13 levels
- Monotonic medians
- Variance is often high
- Within- vs. cross-developer

Quality measurement for use as “summary indicator”



- Quality score binned to 13 levels
- Monotonic medians
- Variance is often high
- Within- vs. cross-developer

Looking ahead…

- FRVT Quality Assessment Track
  - Quality summarization (scalar value)
  - Ongoing and will continue

- FRVT Quality Vector Track NEW
  - Starts Q1 2022
  - Specific image defect detection (vector of values)

Thank you!

frvt@nist.gov
Google: NIST FRVT Quality

National Institute of Standards and Technology
U.S. Department of Commerce

FRVT 1:1 Verification: https://pages.nist.gov/frvt/html/frvt11.html
FRVT 1:N Identification: https://pages.nist.gov/frvt/html/frvt1N.html
FRVT MORPH: https://pages.nist.gov/frvt/html/frvt_morph.html
FRVT Quality Assessment: https://pages.nist.gov/frvt/html/frvt_quality.html
FRVT Face Masks: https://pages.nist.gov/frvt/html/frvt_facemask.html
FRVT Paperless Travel: https://pages.nist.gov/frvt/html/frvt_paperless_travel.html