FACE RECOGNITION: A SOLVED PROBLEM?
IFIP WG10.4

PATRICK GROTHER
NIST, U. S. DEPARTMENT OF COMMERCE
GAITHERSBURG, MARYLAND

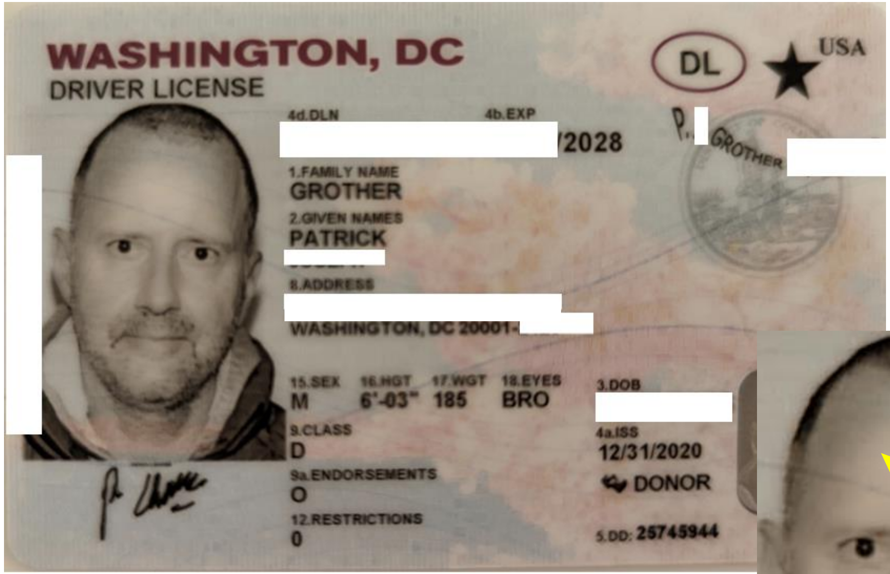
JUNE 26, 2026

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Science and Technology

Office of Biometric Identity Management

USING FR TO BIND LIVE-PERSON TO ID CARD



1. Scan ID credential

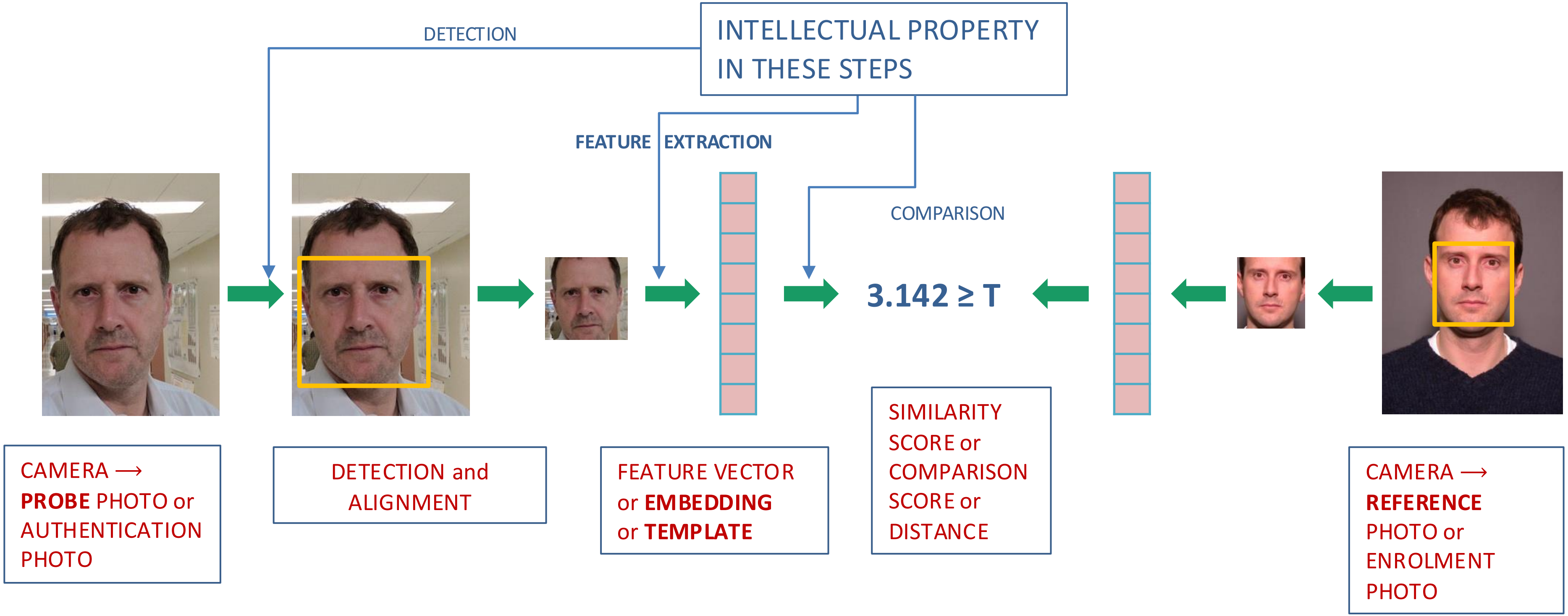
2. Segment

3. Live Image from phone or tablet

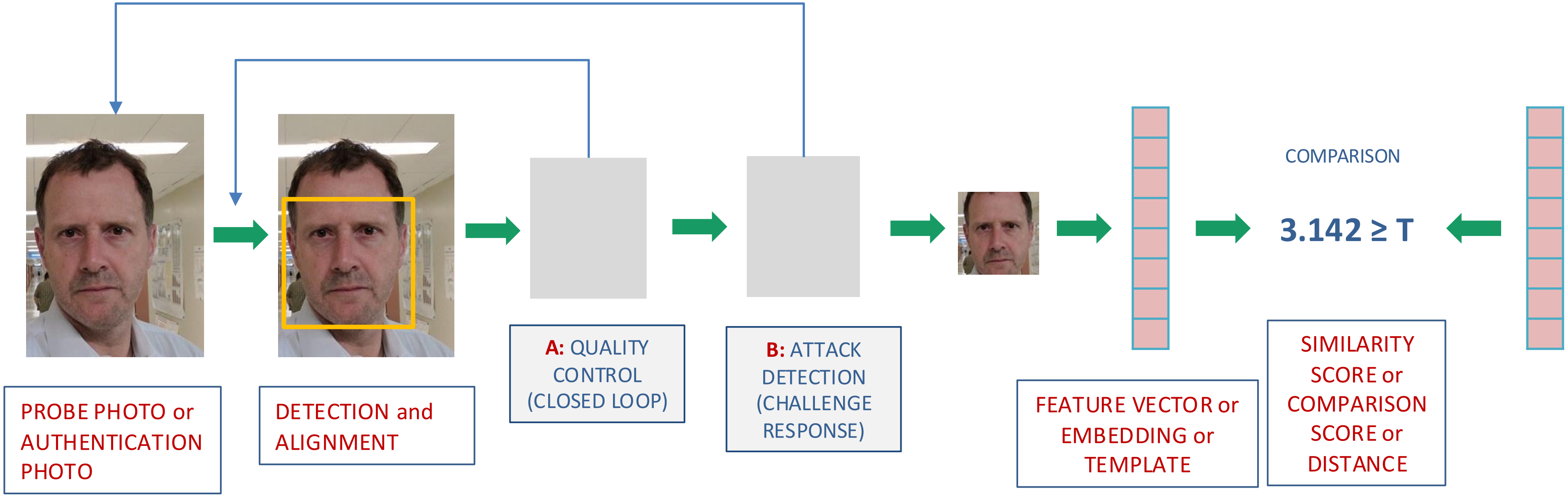
Face Recognition Verification

4. Same person, or not?

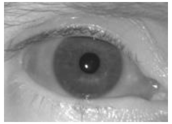
TERMINOLOGY + ARCHITECTURE 1:1: DETECTION, TEMPLATES, COMPARISON



TERMINOLOGY + ARCHITECTURE 1:1: ADDITIONAL COMPONENTS



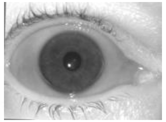
CORE BIOMETRIC OPERATION: COMPARISON



(IDENTITY EXTRACTOR)

TWO IMAGES

(IDENTITY EXTRACTOR)



TEMPLATE

TEMPLATE COMPARATOR

TEMPLATE

SIMILARITY SCORE

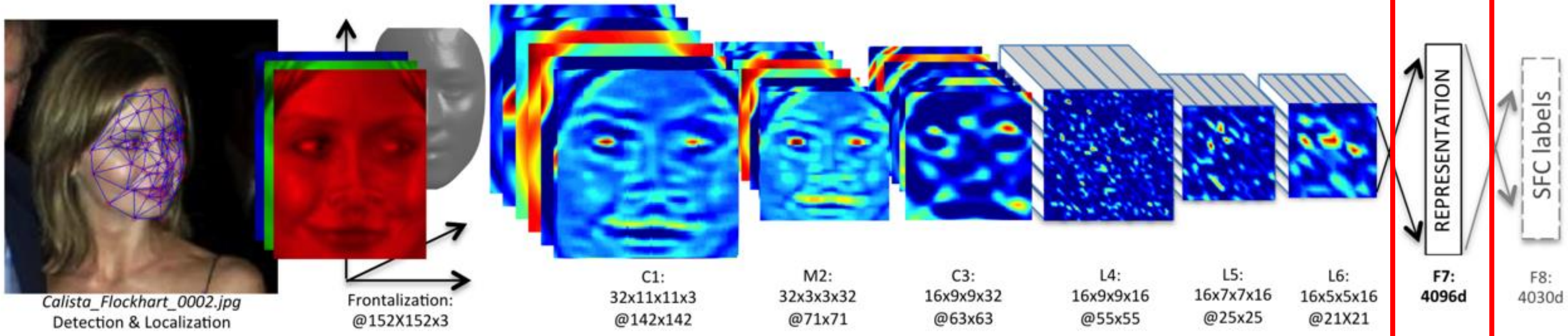
NO, DIFFERENT PERSON*

IS SCORE ABOVE THRESHOLD?

YES, SAME PERSON!

THRESHOLD IS SET BY SYSTEM OWNER TO IMPLEMENT A POLICY

DEEP NEURAL NETWORKS

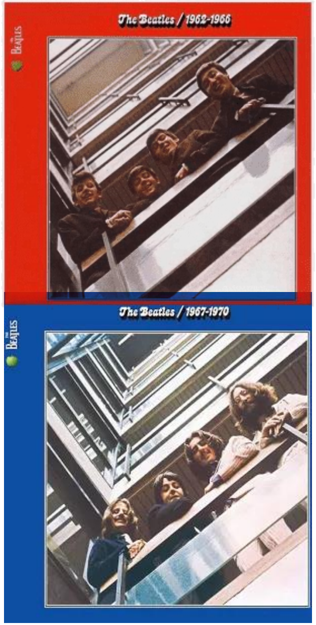


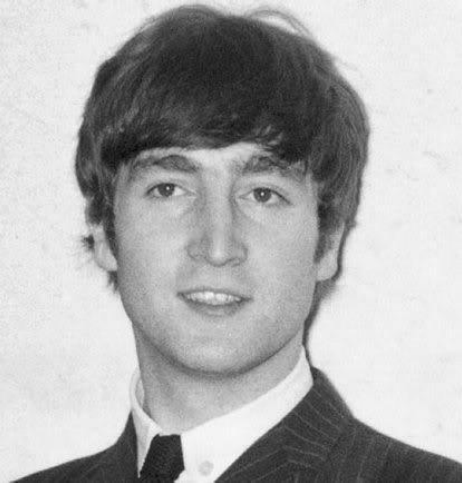
Y. TAIGMAN, M. YANG, M. RANZATO AND L. WOLF, *DEEPFACE: CLOSING THE GAP TO HUMAN-LEVEL PERFORMANCE IN FACE VERIFICATION*, IEEE CVPR, 2014, PP. 1701-1708, DOI: 10.1109/CVPR.2014.220.

EMBEDDING aka
TEMPLATE

STATE OF THE ART FACE RECOGNITION
TECH REFRESH, CAPTURE ENVELOPE, THRESHOLDS

Beatle John Lennon between the release of the Red Album and the Blue Album, ~5 years.





Year	Developer	Algorithm	Score	FMR	Outcome
2024	Idemia	011	6803.72	< 5e-07	**Strong match**
2025	Paravision	020	0.4118	< 5e-07	**Strong match**
2014	Cogent Thales	A20A	2521	0.48	**Failed match**
2014	NEC	E20A	0.562	0.002	**Failed match**

NIST FACE BENCHMARKS

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

FRTE

FACE **RECOGNITION** TECHNOLOGY EVALUATION

RECOGNITION: **WHO** IS IN AN IMAGE

1:1 VERIFICATION	SAME PERSON OR NOT?
1:N SEARCH	WHO? WHERE? WHEN?
FACE IN VIDEO 2024	1:N ON NON-COOP PEOPLE
TWINS DISAMBIGUATION	SAME PERSON, OR TWIN?

FATE

FACE **ANALYSIS** TECHNOLOGY EVALUATION

ANALYSIS: **WHAT** ABOUT AN IMAGE

MORPH DETECTION	TWO PEOPLE IN ONE FACE?
QUALITY + DIAGNOSTICS	HOW BAD IS THIS PHOTO
PAD	SUBVERSIVE PHOTO?
AGE ESTIMATION	HOW OLD? OLD ENOUGH?

Benchmarks are:

- Independent
- Free
- Regular
- Fast
- Repeatable
- Fair
- Black box
- IP-protecting
- Open globally
- Large-scale
- Sequestered datasets
- Statistically robust
- Public
- Transparent
- Extensible
- **ABSOLUTE PERFORM.**
- **RELATIVE PERFORM.**
- **! Certification**
- **! Procurement**

FRTE

FACE **RECOGNITION** TECHNOLOGY EVALUATION

RECOGNITION: **WHO** IS IN AN IMAGE

1:1 VERIFICATION	←	SAME PERSON OR NOT?
1:N SEARCH	←	WHO? WHERE? WHEN?
FACE IN VIDEO 2024	←	1:N ON NON-COOP PEOPLE
TWINS DISAMBIGUATION	←	SAME PERSON, OR TWIN?

FATE

FACE **ANALYSIS** TECHNOLOGY EVALUATION

ANALYSIS: **WHAT** ABOUT AN IMAGE

MORPH DETECTION	←	TWO PEOPLE IN ONE FACE?
QUALITY + DIAGNOSTICS	←	HOW BAD IS THIS PHOTO
PAD	←	SUBVERSIVE PHOTO?
AGE ESTIMATION	←	HOW OLD? OLD ENOUGH?

CELEBRATING EST. 1901 NIST 125 YEARS
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

... AND

IREX

IRIS EXCHANGE (INTEROP) TECHNOLOGY EVALUATION

1:N IRIS + NEW DATASET

FRIF

FINGERPRINT TECHNOLOGY EVALUATION

1:N FINGERPRINT

TATT

TATTOO TECHNOLOGY EVALUATION

1:N TATTOO

SRE

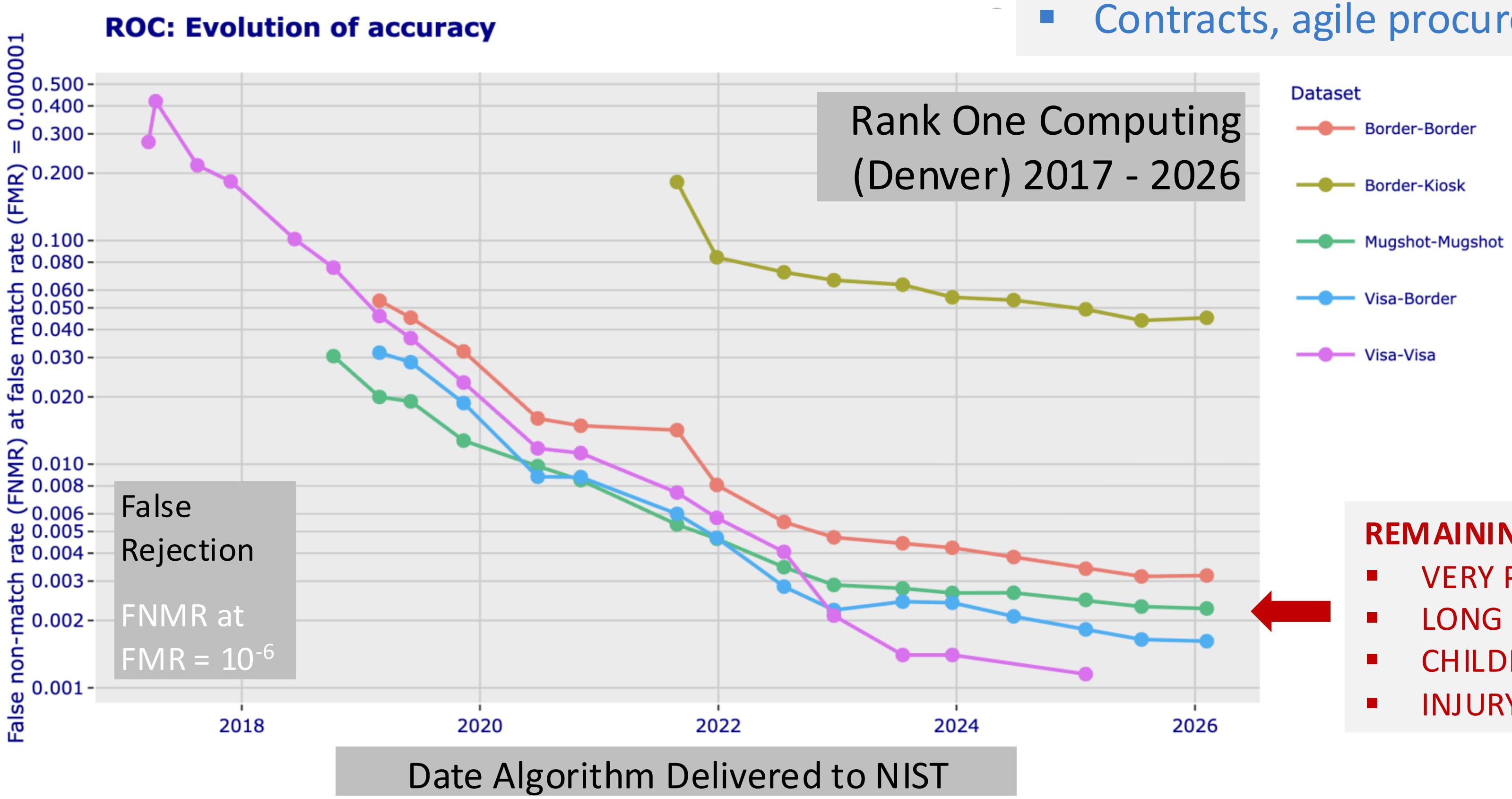
SPEAKER RECOGNITION TECHNOLOGY EVALUATION

1:1 VOICE

BROAD DURABLE GAINS
RISK: USE OF LEGACY ALGORITHM

Management + Mitigation

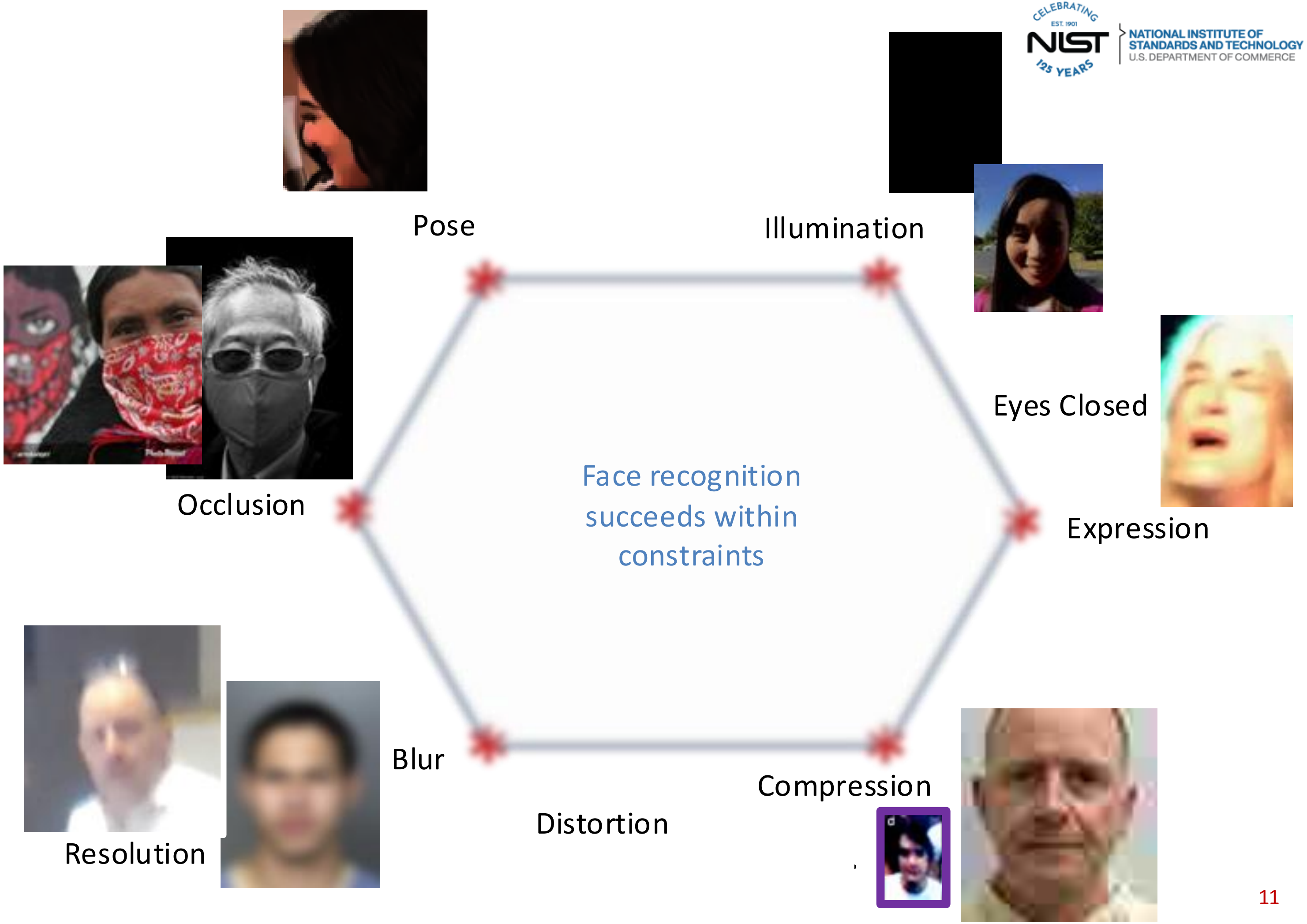
- Algorithms improve regularly
- Do tech refresh!
- Contracts, agile procurement



REMAINING ERRORS

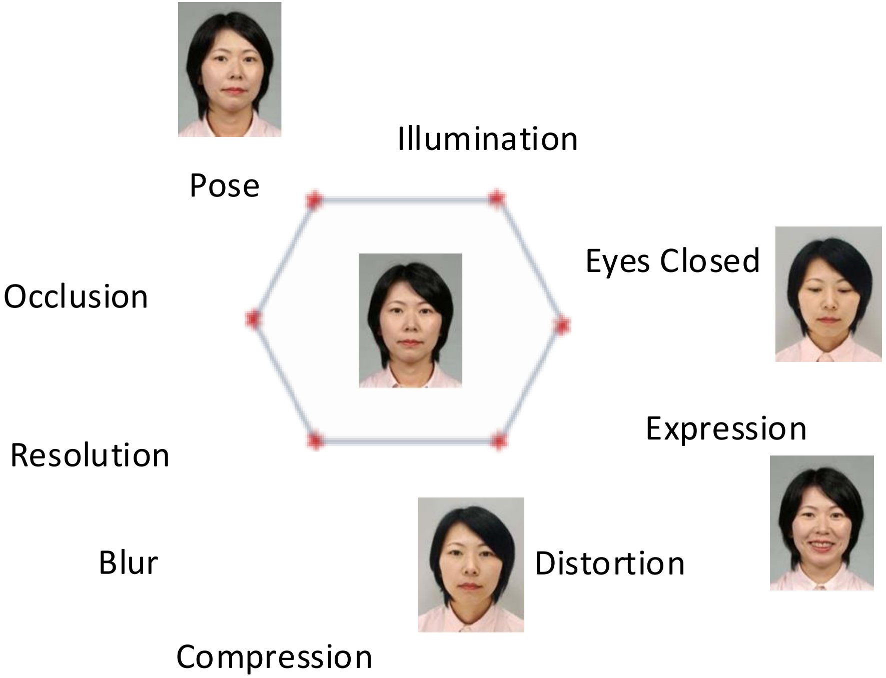
- VERY POOR QUALITY
- LONG TERM AGEING
- CHILDREN
- INJURY

FACE RECOGNITION EXPECTED TO WORK WITHIN A "CAPTURE ENVELOPE"

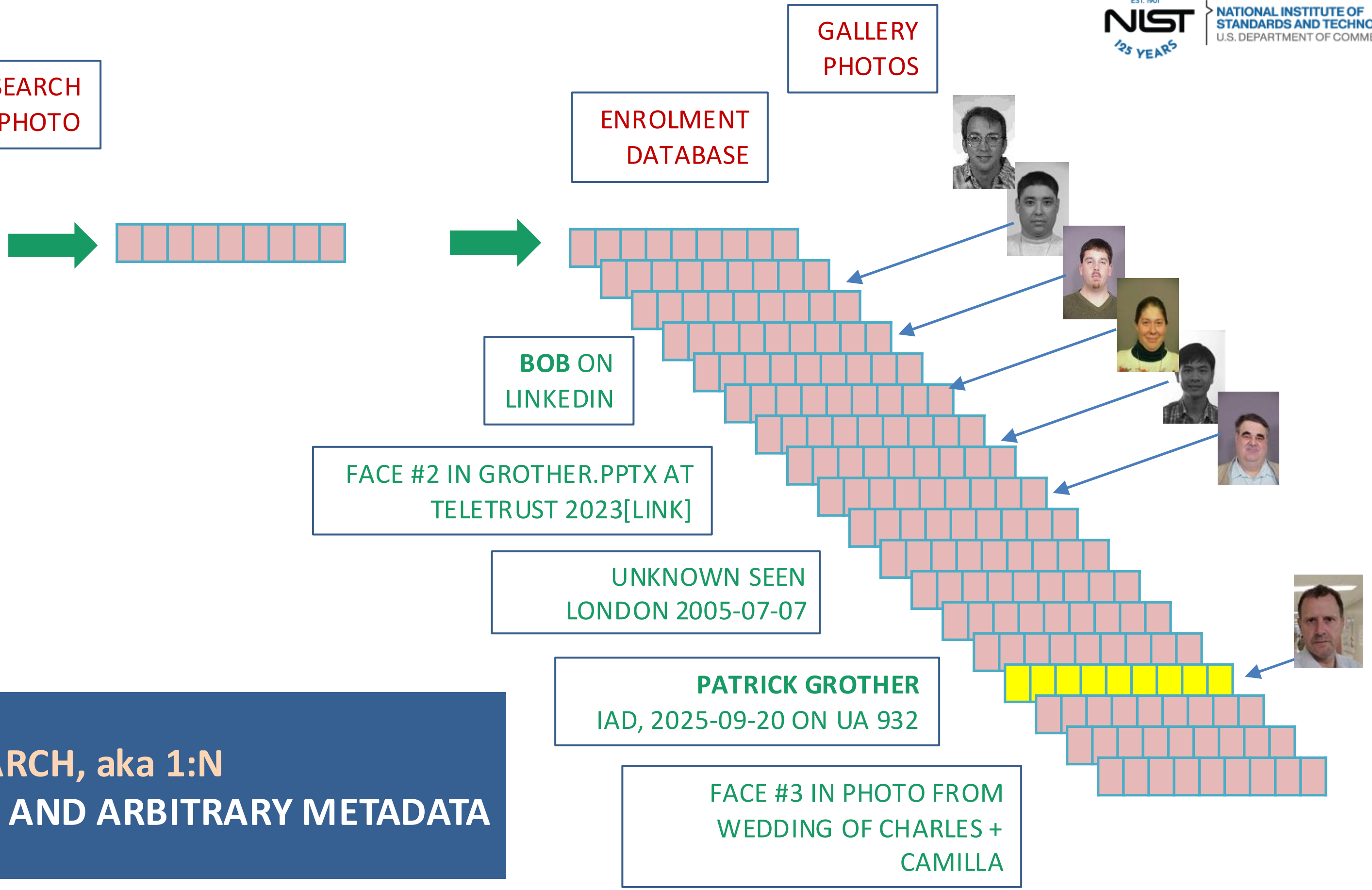


FACE RECOGNITION EXPECTED TO WORK WITHIN A "CAPTURE ENVELOPE"

Role +/- 8°
Yaw +/- 5°
Pitch +/- 5°



PROBE or QUERY or SEARCH PHOTO

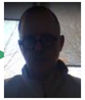


BASICS: GALLERY SEARCH, aka 1:N
ASSOCIATING IMAGES AND ARBITRARY METADATA

INVESTIGATION
T = 0

vs.

(LIGHTS OUT) IDENTIFICATION
T > 0



Search image

Score	Rank
3.142	1
2.998	2
1.626	3
0.707	4
0.330	5
0.198	6
0.074	7
0.016	8

MATE

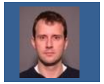
Policy A: Review up to 8 candidates

Source: https://www.google.com/doodles/jim-hensons-75th-birthday

"Investigation"
Returns candidates for human review

"Lights Out"
System concludes this the correct result

Search image



Score	Rank
3.142	1

T

Policy B: Review only candidates with score above threshold T = 3.0

1:N LEADERBOARD SNAPSHOT

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Identification (T>0) by Developer | Investigation (R=1, T=0) by Developer | Identification (T>0) by Algorithm | Investigation (R=1, T=0) by Algorithm | Demographics: False Positive Dependence | Demographics: False Negative Dependence | Resources by Algorithm

Show 25 entries

Search: ex. 2023|insight

Algorithm	Gallery / Probe / Date	Mugshot / Mugshot / N = 12000000	Mugshot / Mugshot / N = 1600000	Mugshot / Webcam / N = 1600000	Mugshot / Profile 90° / N = 1600000	Visa / Border / N = 1600000	Visa / Kiosk / N = 1600000	Border / Border ΔT ≥ 10 YRS / N = 1600000	Mugshot / Mugshot ΔT ≥ 12 YRS / N = 3000000
qazsmartvisionai_002	2024-12-03	0.0007(2)	0.0005(1)	0.0057(1)	0.0460(1)	0.0013(1)	0.0443(1)	0.0211(6)	0.0031(3)
nec_010	2025-01-24	0.0007(1)	0.0006(2)	0.0057(2)	0.0462(2)	0.0015(2)	0.0444(2)	0.0052(1)	0.0019(1)
cloudwalk_mt_002	2023-02-24	0.0017(16)	0.0015(25)	0.0113(22)	0.0484(3)	0.0016(3)	0.0478(4)	0.0157(2)	0.0030(2)
stcon_004	2026-01-26	0.0013(7)	0.0007(7)	0.0069(6)	0.0701(7)	0.0018(4)	0.0469(3)	0.0181(4)	0.0047(8)
megvii_004	2023-10-18	0.0011(5)	0.0009(15)	0.0081(12)	0.0599(6)	0.0019(5)	0.0565(10)	0.0167(3)	0.0053(9)
innovatrics_015	2025-09-29	0.0012(6)	0.0008(11)	0.0065(3)	0.0815(12)	0.0021(6)	0.0564(9)	0.0252(9)	0.0040(5)
sensetime_011	2026-03-12	0.0009(3)	0.0006(3)	0.0065(4)	0.0717(8)	0.0021(7)	0.0741(35)	0.0229(7)	0.0043(7)

2004

THE POWER OF AFR

Use 15 FRTE 1:N algorithms, c. 2023 to

1. enrol my 2004 portrait into a gallery with N = 12 million other people
2. search images below:

SEARCH PHOTOS GIVING HIGH-SCORE MATE AT RANK 1

2002 | 2018 | 2007 50º YAW | 2019 | 2008 SUNGLASSES | 2014 60º YAW | 2009 SHADOW

THE POWER OF AFR: 2026 EDITION

1. Enroll 2004 portrait in a gallery with N = 12 million other people
2. Search with 10 algorithms

SEARCH PHOTOS GIVING HIGH-SCORE MATE AT RANK 1

2003
SHADOW

2009
POSE

2007
55º PITCH

2021
DL

2019
PITCH

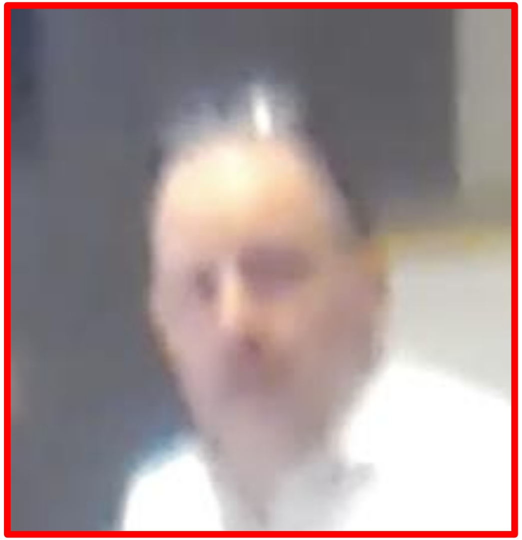
2014
90º YAW

2014
40º PITCH

AFR IN 2026 :: WHAT'S LEFT

Enroll 2004 portrait into a gallery with
N = 12 million other people

RANK 1
WEAK HIT



MISSED
OUTSIDE 50
RANKS

HIT BY ONE
CHINESE
ALGORITHM
RANK 15

QUESTION :: HOW ACCURATE IS FACE RECOGNITION?

ANSWER: Face search will succeed 100% of the time if

a. You're using a recent leading FR algorithm AND
b. There is a mate in the database AND
c. The mate is not more than X years old AND
d. The image is not manipulated AND
e. The image has limited quality problems – within the "capture envelope"

but there are caveats: Twins, occlusion, attacks, demographics, application details

GOOD → BAD → UGLY

FACE SEARCH IN ACTION

Hot Money podcast World + Add to myFT

How western governments took down Europe's most powerful crime group

Our second series investigates the rise and fall of a Dubai-based super cartel

Miles Johnson

Published DEC 4 2023 | Updated DEC 26 2023, 06:16

55

Pimeyes.com ?
Eyematch.ai
Lenso.ai
…

https://www.ft.com/content/40406849-e120-40d4-8c21-f41b5e5f8a9c

ERROR TRADEOFF FOR 1:N FACE

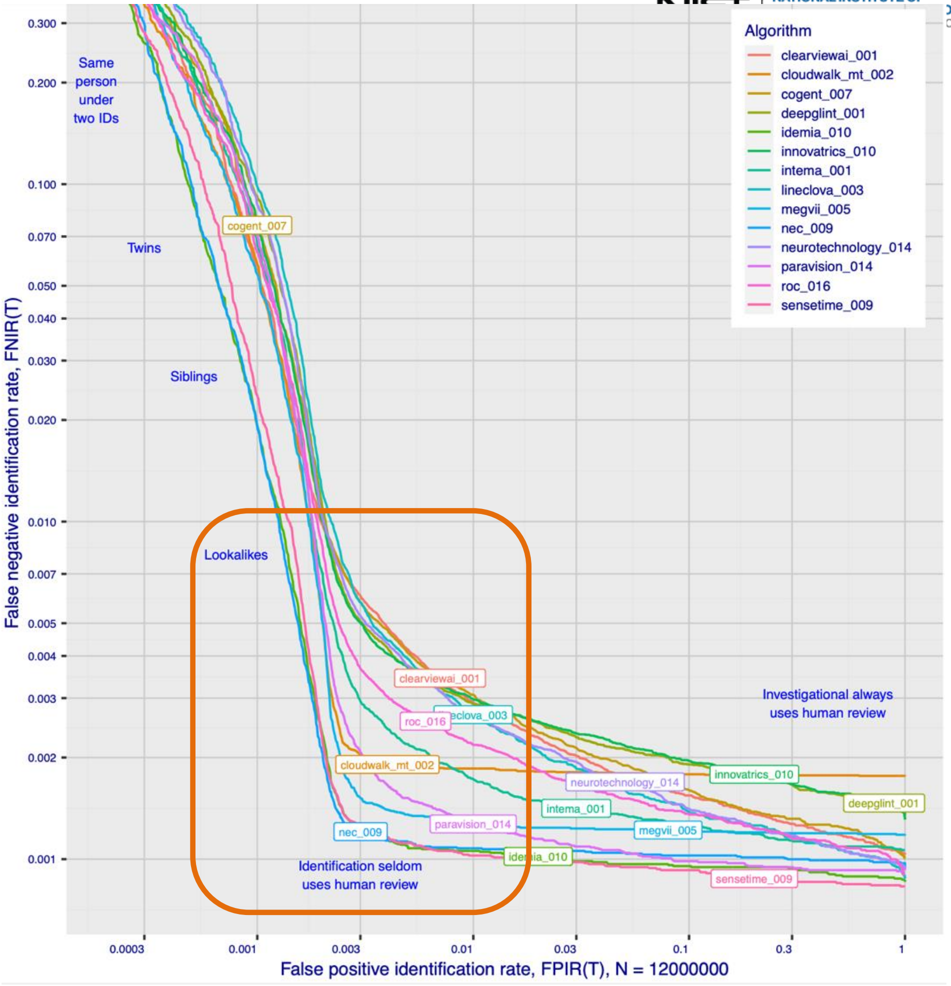
N = 12 000 000 MUGSHOTS

SYMPTOM

- High non-mate scores mixed in with mate scores

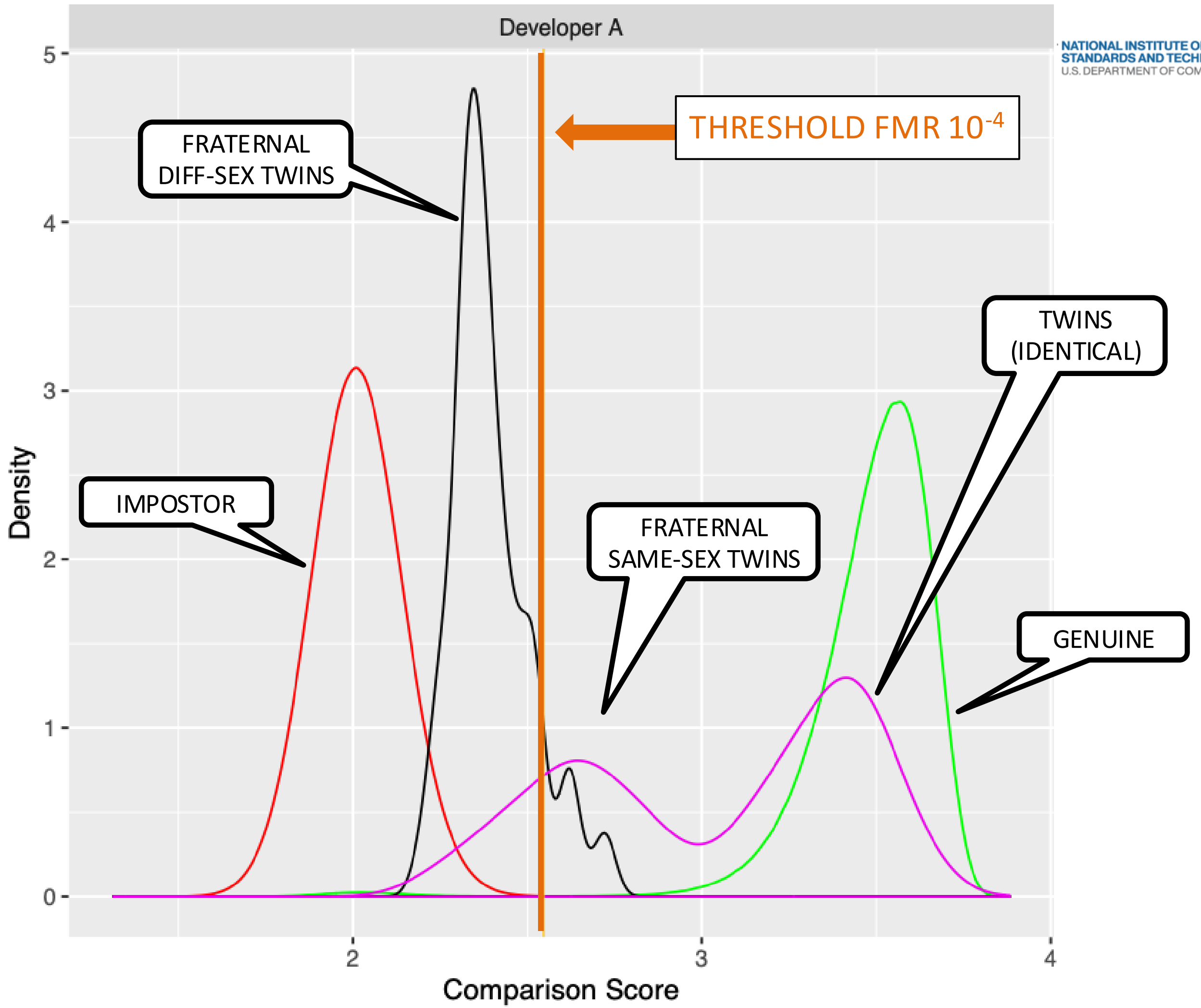
CAUSES

- ID label errors: Instances of one person under two IDs
- In Face Recognition: Twins, Siblings.



SAME PERSON? OR NOT?

AUTOMATED FACE RECOGNITION: TWINS ARE SIMILAR



TWINS ARE COMMON

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Source: Notre Dame's Twins Day Collection

	Identical	Fraternal
How	Monozygotic	Dizygotic
USA proportion that are a twin	0.7%	2.4%
West Africa	0.5%	2.8%
East Asia	0.3%	0.9%
Same-sex	100%	50% in theory 58% actually
Twinning rate	x1.5 since 1980	x1.9 since 1980
Demographics	~ constant with age, geography	varies with mothers age, order, geography

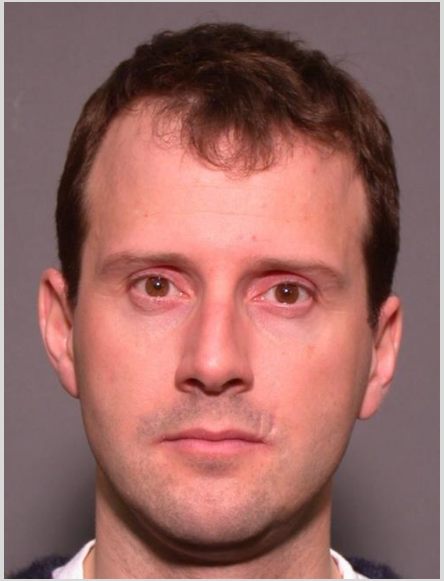
Twins, triplets … constituted 117,000 out of 3.7M live births in US 2022 = 3.2%

https://www.cdc.gov/nchs/data/nvsr/nvsr73/nvsr73-02.pdf

DETOUR:
USE OF ALGORITHM "OUT OF ENVELOPE"

LOW QUALITY: A CAUSE OF FALSE POSITIVES?

OUT-OF-ENVELOPE RISK
KEEP ANOMALOUS INPUTS OUT OF DNNS

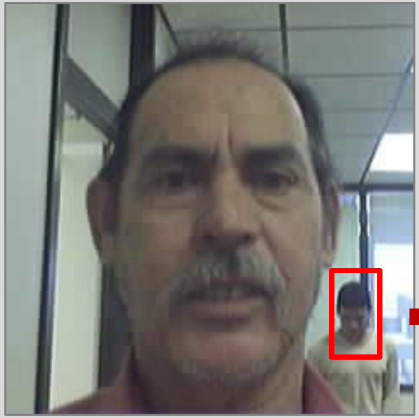


3000 x 2000

A: SLR Camera c. 2010 6 megapixel image
Standard Mugshot (SAP 50)

640 x 480

B: e-Passport
2004…2016
0.3 megapixels

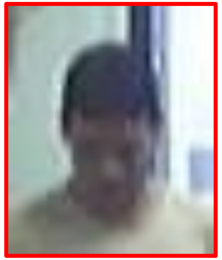


240 pix

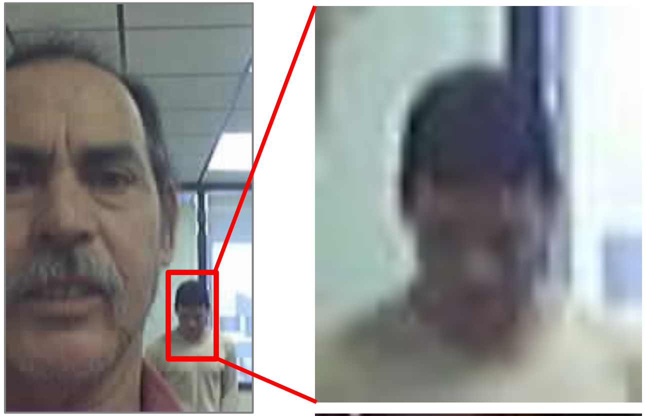
240 pix

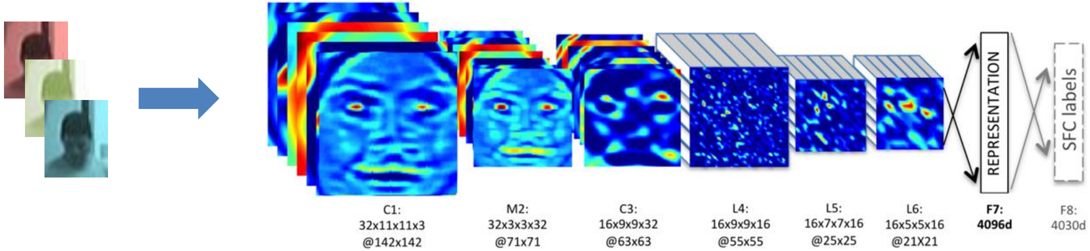
C: Common image / video

32 x 24
NOT TO SCALE



LOW RESOLUTION COMPARISON YIELDS FALSE POSITIVES

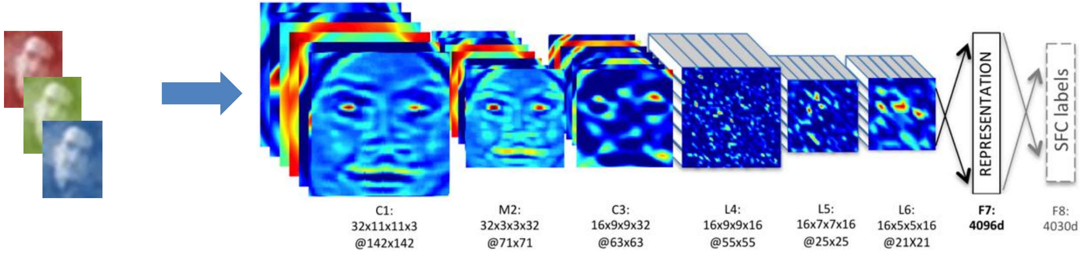




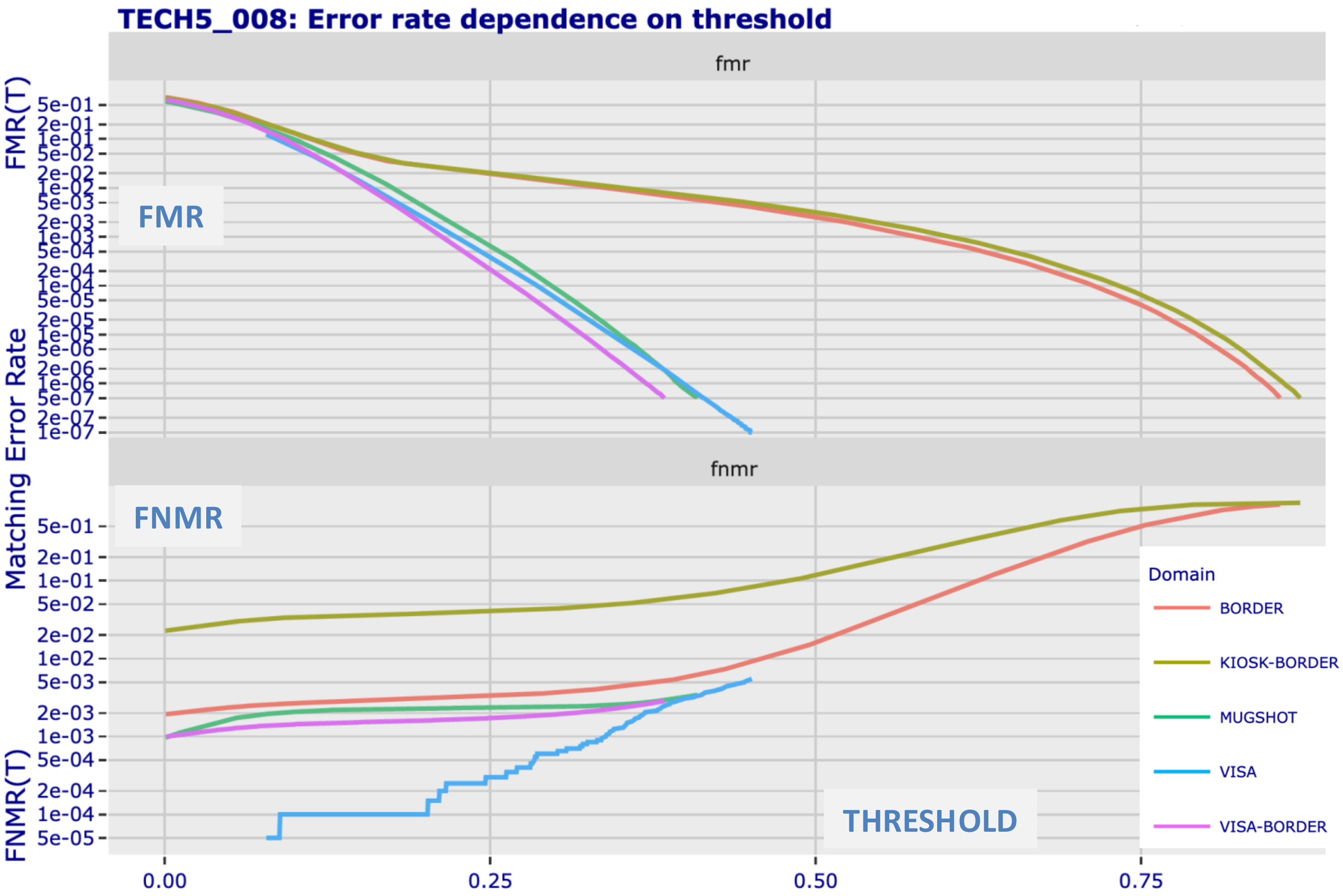
32 x 24

142x142

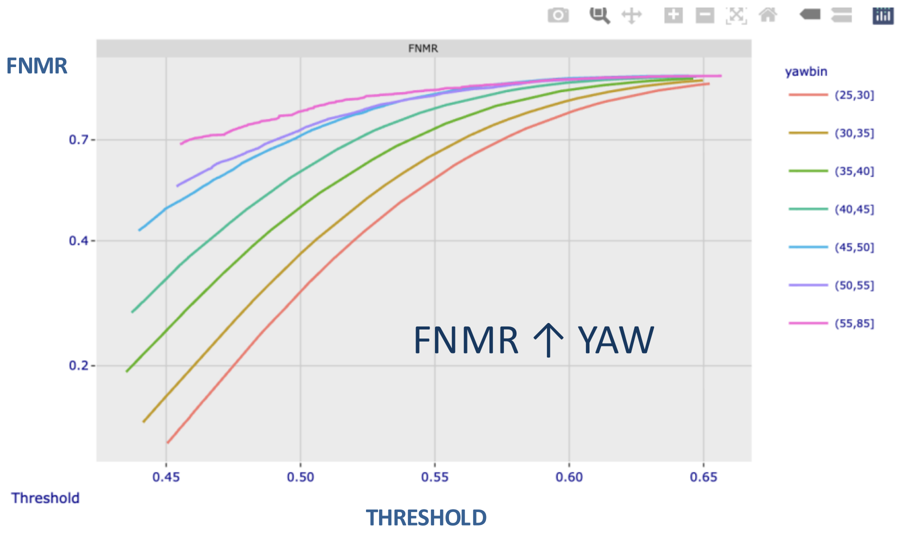
COMPARISION GIVES HIGH NON-MATE SCORE



ANOMALOUS BEHAVIOUR: ELEVATED FMR(T)



TWO ERROR RATES vs. YAW ANGLE



ATTACK OVERVIEW

ANALOG

DIGITAL

LEGITIMATE PRESENTATION OR PRESENTATION ATTACK ?

TRUSTED CAMERA OR HACKED INJECTOR?

Source: https://www.biometricupdate.com/202406/vote-begins-on-biometric-injection-attack-standard

AFR IS OPEN TO PRESENTATION ATTACK

1. Spoofing:

- Goal: Impersonate someone else
- How: Maximize similarity score

Enrollment

Verify

Verify

2. Evasion:

- Goal: Do not match your prior enrollment, to impede 1:N detection
- Minimize similarity score

Database

Search

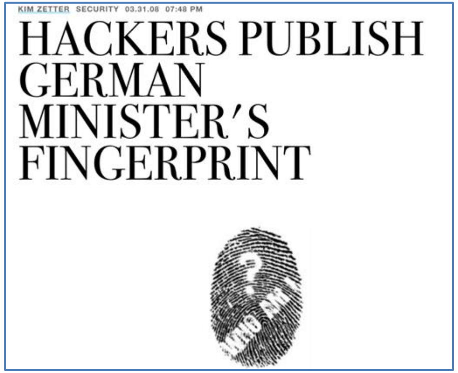
STEALING SAMPLES TO ATTEMPT IMPERSONATION

FACE

- **Digital availability:** Easy
  - Social media
  - Weddings
  - Conferences
  - Corporate websites
- **Physical availability:** Easy
  - Take a photo in public
  - From an ID card
- Replay: Easy
- Recognition:

FINGER

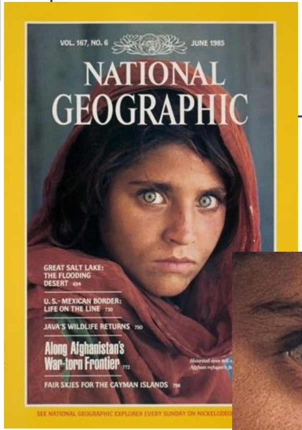
- **Digital availability:** Difficult
  - Some immigration and most criminal databases
  - Some ePassports DG3
- **Physical availability:** Difficult
  - Latent mark on surface
  - Close-range high resolution photograph


KIM ZETTER SECURITY 03.31.08 07:48 PM

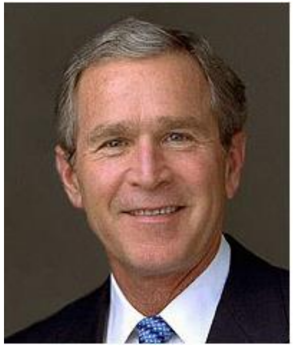
HACKERS PUBLISH GERMAN MINISTER'S FINGERPRINT

IRIS

- **Digital availability:** Difficult
  - Some Immigration databases
  - Rarely ePassports DG4
- **Physical availability:** Difficult
  - Photography is optically difficult
    - Reflections, noise, resolution
    - Brown eyes need near infrared



Face morphing: single image of two people



+

George W. Bush
(43rd US president)

Barack Obama
(44th US
president)

Face morphing generates an image that resembles both contributing subjects

MORPH OR NOT: ANY GUESSES?

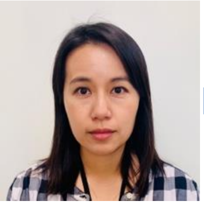
ONE DOCUMENT, TWO FACES: FOUNDATIONAL ID?

PASSPORT APPLICATION

Accomplice (passport owner)

+

Attacker (other identity)



Morph = A + B

Printed on 2in x 2in photo paper and mailed to passport office

Passport office scans printed photo, conducts human examination, …

Image Source: NIST

- **Some passport applications are susceptible to manipulation of user-submitted photos.**
- **Risk #1: Countries accepting user-submitted ID photos enable morphing attacks.**
- **Risk #2: Countries that process such ID documents acquire immigration risk.**

Morphing: two detection opportunities

Morph

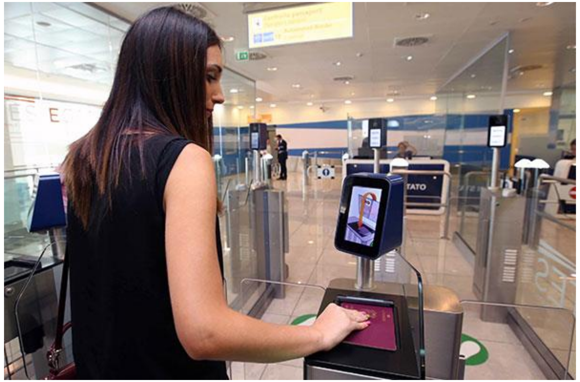
Image Source: NIST

DOCUMENT ISSUANCE: Suspect image in isolation (Single-image morph detection)

Opportunity: Can be effective on morphs from a particular morph generator, if they have been trained on examples from that generator

Challenge: Does not generalize across not-previously-seen morphing methods

Image Source: http://www.futuretravelexperience.com/2016/01/automated-border-control-e-gates-go-live-at-naples-airport/



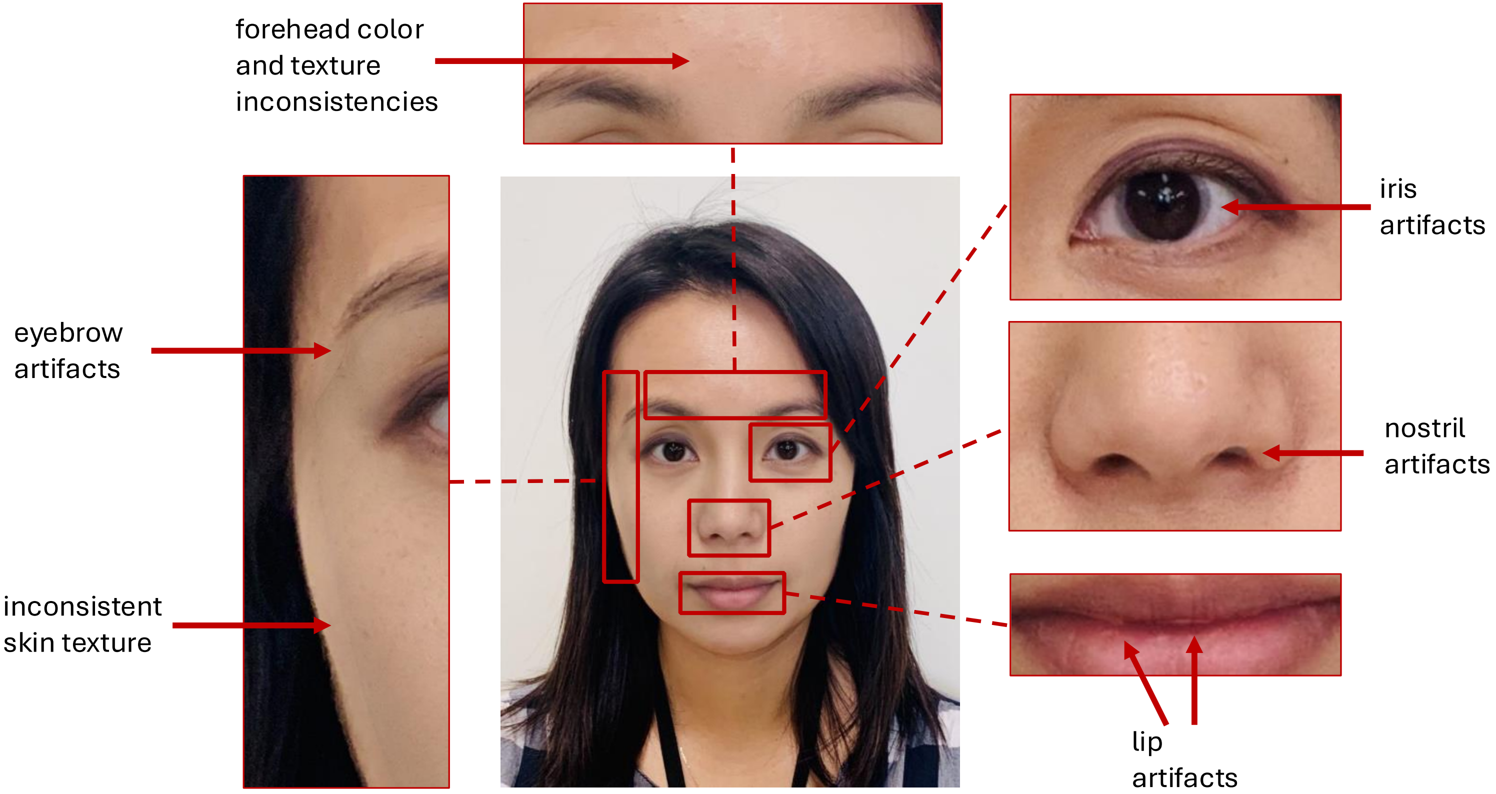
BORDER CROSSING: Suspect image + live image (Differential morph detection)

Opportunity: Morph detection is more generalizable, because identity info can be analyzed (instead of specific image artifacts)

Considerations: Requires an additional live capture constituent photo to be available

INVESTIGATION PROCESS: VISUAL INSPECTION

Inspection for presence of artifacts



DEMOGRAPHICS

DO SOME GROUPS HAVE HIGHER FALSE POSITIVE RATES?


https://freephotos.cc/

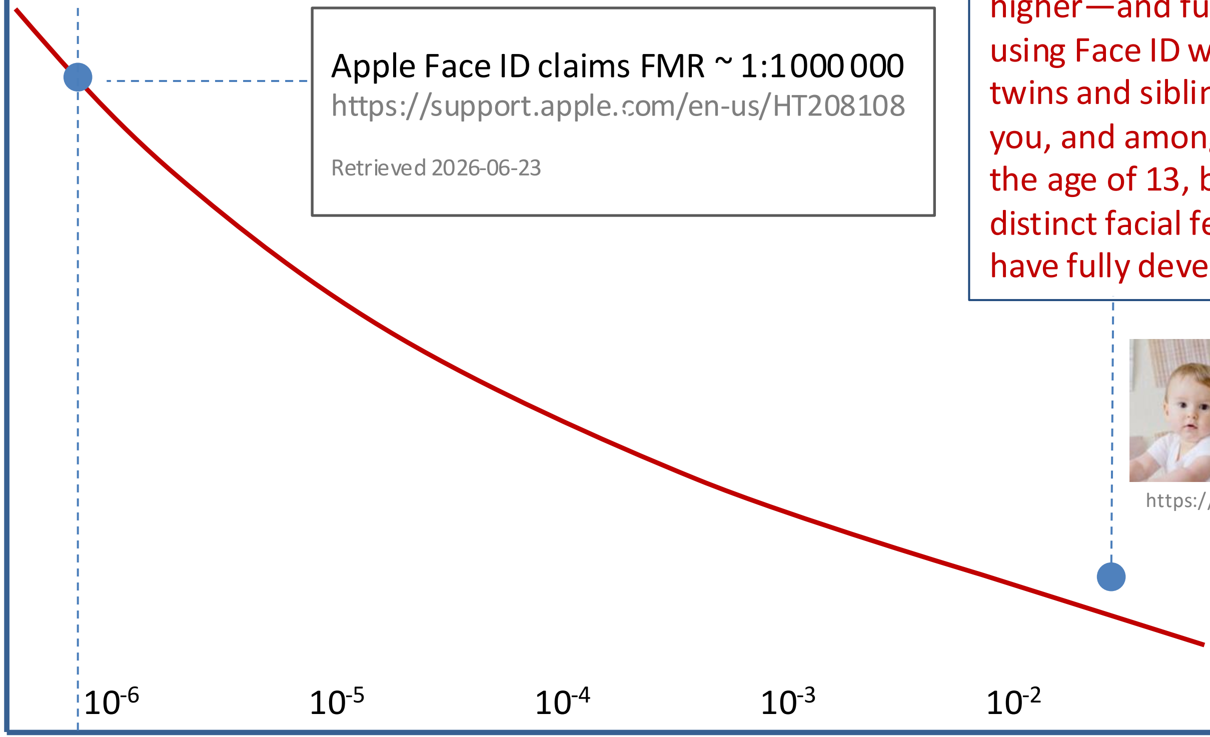
ERROR TRADEOFF CHARACTERISTIC
FMR AS INSTRUMENT OF SECURITY POLICY

FNMR
False non-match rate

Proportion of genuine comparisons producing score below threshold, T.

See ISO/IEC 19795-1

Log-scale is typical to show small numbers.



Log-scale is often required because low FMR values are operationally relevant.

"The statistical probability is higher—and further increased if using Face ID with a mask—for twins and siblings that look like you, and among children under the age of 13, because their distinct facial features might not have fully developed."

https://freephotos.cc/

NIST staff + sister, with permission

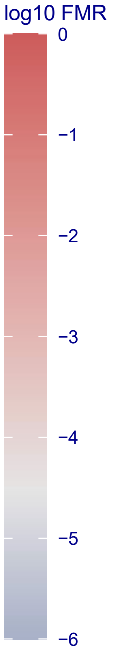
FMR False match rate
Proportion of impostor comparisons searches yielding any candidates at or above threshold, T.

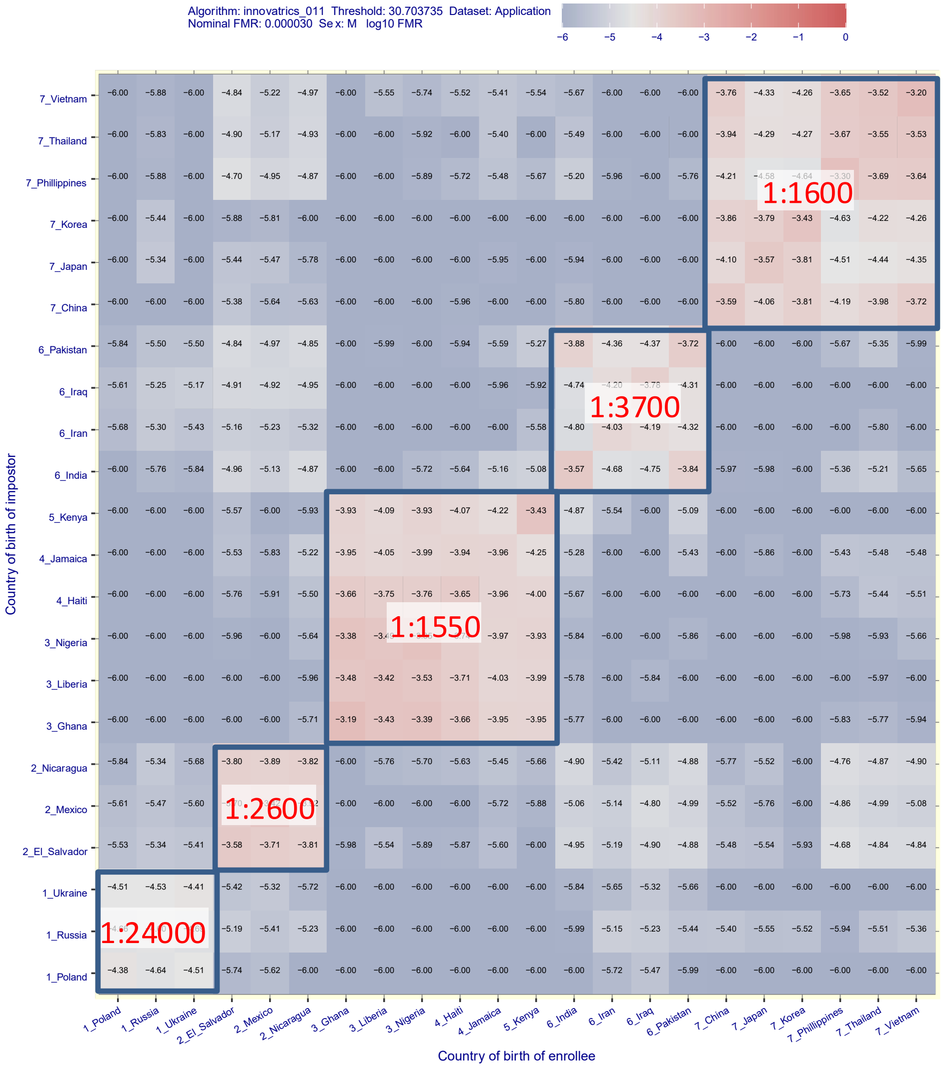
GLOBAL CROSS-COUNTRY FMR
MEN 20-35 ONLY

THIS FIGURE SHOWS FMR VARIATIONS FOR MEN 20-35 FOR ONE 1:1 COMPARISON ALGORITHM AT A GLOBALLY FIXED THRESHOLD

FOR ALL AGE GROUPS, BOTH SEXES, SEE FIGURES HYPERLINKED FROM [LEADERBOARD] E.G.:

[PDF][PDF][PDF][PDF]





N. E. Asia

S. E. Asia

S. Asia

E. Africa

Caribbean

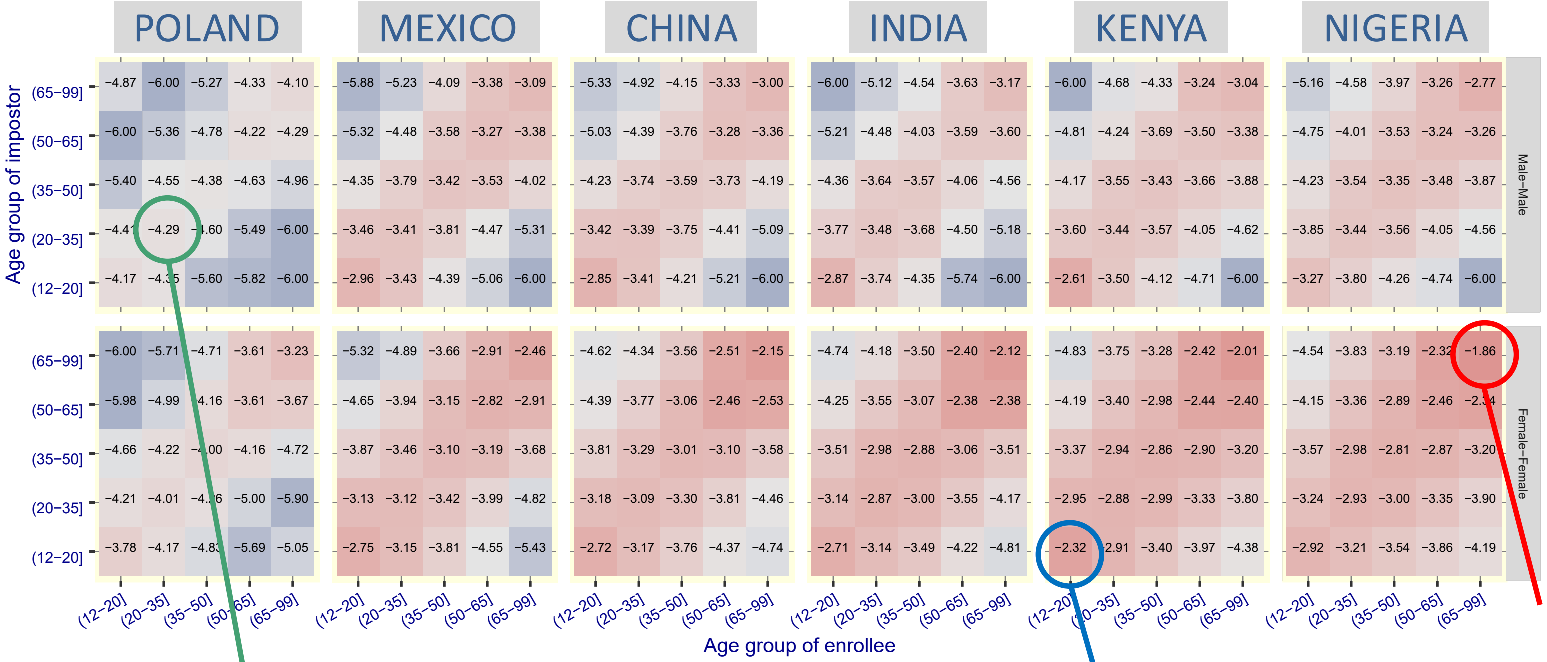
W. Africa

C. America

E. Europe

NOW ALSO CONSIDER WOMEN AND AGE

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE



MEN

WOMEN

FMR ~ 1 IN 16000 ON POLISH MEN 20-35

FMR ~ 1 IN 200 ON KENYAN WOMEN <= 20

FMR ~ 1 IN 72 ON NIGERIA WOMEN > 65

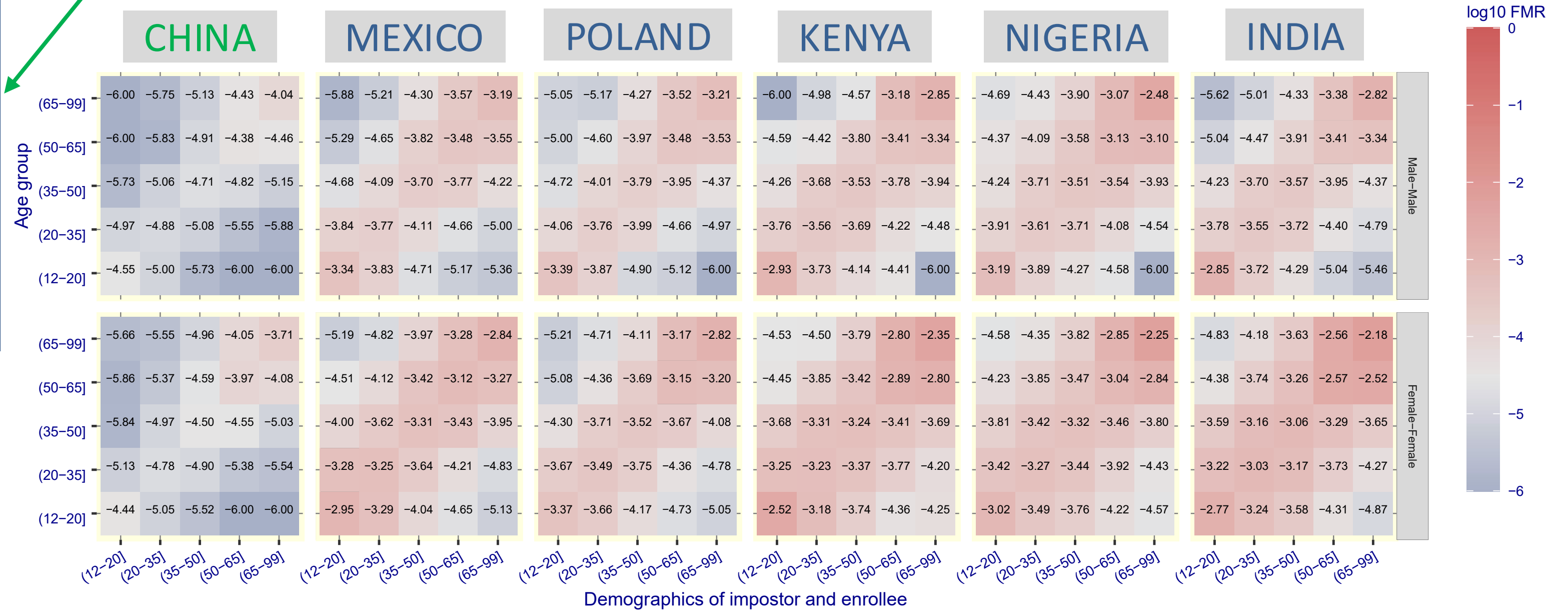
IS THE MOST ACCURATE ALGORITHM UNBIASED?

CLOUDWALK (CHINA)

DECEMBER 2020

SOME CHINESE ALGORITHMS GIVE LOW FMR ON CHINESE FACES

Algorithm: mt_003
Threshold: 228.280239



RMF REQUIREMENT

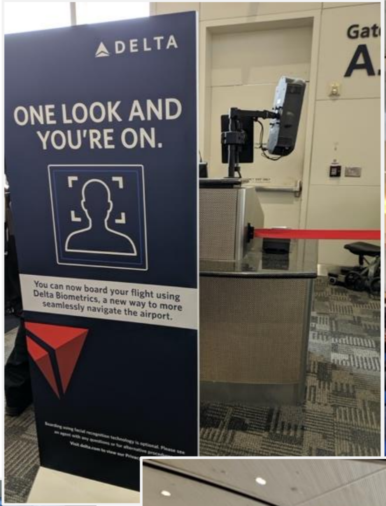
ORGANIZATIONS SHALL DOCUMENT IMPACTS OF ALL POSSIBLE ERRORS

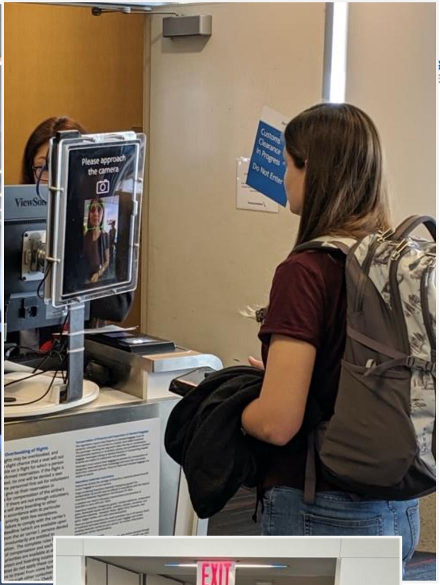
DEPARTING A COUNTRY

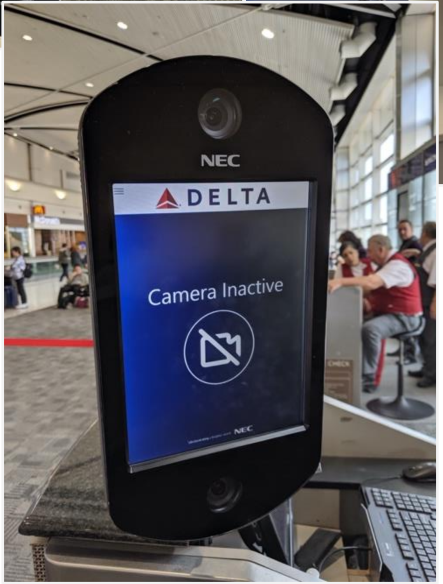
1. POSITIVE ACCESS CONTROL
2. IMMIGRATION EXIT FACILITATION

HOW

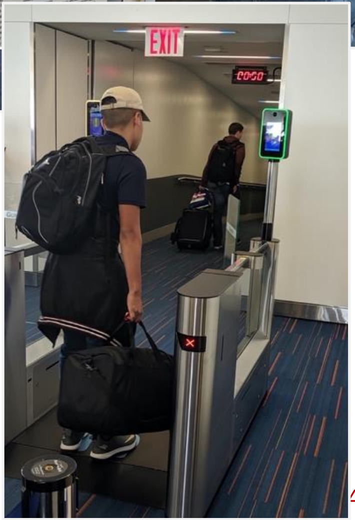
1. FACE RECOGNITION 1:N
2. PAPERLESS BOARDING







Diverse hardware, common matcher (TVS)



FALSE POSITIVES IN REAL-TIME IDENTIFICATION

WHO: PEOPLE NOT IN THE DATABASE
RISK: **THEY MATCH SOMEONE WHO IS**
IMPACT: TO EITHER OR BOTH PEOPLE

TWO TECHNICAL COMPONENTS

1. DEMOGRAPHIC EFFECTS
2. LOW RESOLUTION

1:N SEARCH BIAS
FALSE POSITIVES IN OPERATIONS

Rite Aid's A.I. Facial Recognition Wrongly Tagged People of Color as Shoplifters

Under the terms of a settlement with the Federal Trade Commission, the pharmacy chain will be barred from using the technology as a surveillance tool for five years.

https://www.nytimes.com/2023/12/21/business/rite-aid-ai-facial-recognition.html Eduardo Medina, 2023-12-21

FTC [REPORTS](https://www.ftc.gov) THAT *"THE SYSTEM GENERATED THOUSANDS OF FALSE-POSITIVE MATCHES"*

https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without

https://www.engadget.com/ftc-bans-rite-aid-from-using-facial-surveillance-systems-for-five-years-053134856.html

BIOMETRICS 101: NON-MATED ONE-TO-MANY SEARCHES

GALLERY SIZE, N

$$\text{FPIR} = 1 - (1 - \text{FMR})^N$$

ESTIMATED PROBABILITY THAT A SEARCH YIELDS ANY FALSE POSITIVES, FPIR

FALSE MATCH RATE FROM A 1:1 COMPARISON, FMR

NONE OF THE N COMPARISONS MUST GIVE A FALSE POSITIVE

Approximation when
$N \cdot \text{FMR} << 1$

$$\text{FPIR} = N\,\text{FMR}$$

THIS SAYS FPIR SCALES LINEARLY WITH NUMBER OF PEOPLE IN GALLERY, SO 10x MORE PEOPLE ⟶ 10x INCREASE IN LIKELIHOOD OF FALSE MATH. TYPICALY REMEDY: INCREASE THE THRESHOLD TO MAINTAIN FPIR.

BUT .. SOME ALGORITHMS DON'T BEHAVE LIKE THIS: FPIR IS. ROUGHLY CONSTANT WITH CHANGE IN N.

TERMINOLOGY STANDARDIZED IN ISO/IEC 19795-1:2019

FALSE POSITIVES: A CASINO EXAMPLE

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

- N = 500 people ("The Griffin Book" people who cheated in casinos)
- Algorithm configured for FMR = 1 in 1 million

False match rate from a 1:1 comparison, FMR

$$\text{FPIR} = N \;\; \text{FMR}$$

$$\text{FPIR} = 500 \times 10^{-6} = 5 \times 10^{-4} = 1 \text{ in } 2000$$

Probability that a search yields any false positives

Suppose also

- P = 10000 people visit the casino today

Number of false positives = P . FPIR = $10000 \times 5\ 10^{-4}$ = 5 per day

BUT … FMR IS NOT A SINGLE FIXED VALUE FOR ALL SUBJECTS …

NIST INTERAGENCY REPORT 8280 (2019) REPORTED

- HIGHER FMR IN WOMEN
- HIGHER FMR IN E. ASIANS and AFRICANS
- HIGHER FMR IN THE OLD AND VERY YOUNG

WITH SUBSTANTIAL VARIATIONS ACROSS ALGORITHMS SUSTAINING THE NISTIR 8280 RECOMMENDATION TO “KNOW YOUR ALGORITHM”

EXAMPLES ON NEXT TWO SLIDES

NIST ONGOING 1:1 FRTE (SINCE 2019) REPORTS DEMOGRAPHIC DEPENDENCE OF FMR FOR MOST SUBMITTED ALGORITHMS

PREDICTING 1:N FALSE POSITIVES FROM 1:1 RESULTS

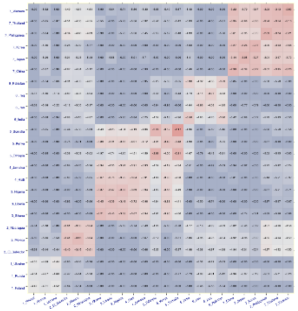
NUMBER OF FALSE POSITIVES = WHO GETS SEARCHED x CROSS-DEMOGRAPHIC FALSE MATCH RATES x WHO GETS PUT IN GALLERY

$$\mathbf{NFP} = \mathbf{p}^T \; \mathbf{FMR} \; \mathbf{g}$$

SOCIO-TECHNICAL CONTEXT FOR FR USE:

- WHO IS IN GALLERY
- WHO FALSE MATCHES WHO
- WHO GETS SEARCHED

DEMOGRAPHIC COMPOSITION OF PROBE SET



DEMOGRAPHIC COMPOSITION OF GALLERY

WHO FALSE MATCHES WHO: HOW FREQUENTLY

CROSS-DEMOGRAPHIC FALSE MATCH RATE MATRIX

FP RATE (FPIR)

WHO FALSE MATCHES WHO: HOW FREQUENTLY × WHO GETS PUT IN GALLERY

$$\mathbf{FPIR} = \mathbf{FMR} \; \mathbf{g}$$

$$\begin{bmatrix} 0.2 \\ 0.008 \end{bmatrix} = \begin{bmatrix} 1 \times 10^{-5} & 0 \\ 0 & 1 \times 10^{-7} \end{bmatrix} \begin{bmatrix} 20000 \\ 80000 \end{bmatrix}$$

FP NUMBER (NFP)

NUMBER OF FALSE POSITIVES = WHO GETS SEARCHED × WHO FALSE MATCHES WHO: HOW FREQUENTLY × WHO GETS PUT IN GALLERY

$$\mathbf{NFP} = \mathbf{p}^T \; \mathbf{FMR} \; \mathbf{g}$$

$$\begin{bmatrix} 200+8 \end{bmatrix} = \begin{bmatrix} 1000 & 1000 \end{bmatrix} \begin{bmatrix} 1 \times 10^{-5} & 0 \\ 0 & 1 \times 10^{-7} \end{bmatrix} \begin{bmatrix} 20000 \\ 80000 \end{bmatrix}$$

DEMOGRAPHIC COMPOSITION OF PROBE SET

CROSS-DEMOGRAPHIC FALSE MATCH RATES

DEMOGRAPHIC COMPOSITION OF GALLERY

1:N FPIR IN 8-GROUP GLOBAL SEARCH
BORDER-CROSSING vs. VISA ENROLLMENT

Algorithm	Date	GINI	MXOG	E-Asia F	E-Europe F	S-Asia F	W-Africa F	E-Asia M	E-Europe M	S-Asia M	W-Africa M
toshiba-003	2026_01_09	0.41(15)	3.29(18)	0.0015(6)	0.0020(82)	0.0018(3)	0.0003(1)	0.0008(4)	0.0019(131)	0.0039(22)	0.0004(1)
idemia-011	2024_03_05	0.33(8)	2.76(12)	0.0026(12)	0.0020(82)	0.0024(9)	0.0061(19)	0.0017(10)	0.0006(98)	0.0033(14)	0.0020(28)
cognitec-009	2024_08_29	0.48(23)	3.80(22)	0.0061(20)	0.0020(82)	0.0116(26)	0.0199(47)	0.0026(24)	0.0007(105)	0.0131(84)	0.0090(115)
nec-011	2025_10_16	0.45(17)	2.96(14)	0.0130(26)	0.0020(82)	0.0100(21)	0.0113(34)	0.0047(39)	0.0007(112)	0.0045(29)	0.0030(37)
incode-007	2025_07_30	0.72(122)	11.37(118)	0.0280(39)	0.0020(82)	0.0075(18)	0.0100(30)	0.0019(16)	0.0002(4)	0.0010(1)	0.0007(8)
cogent-007	2023_01_30	0.63(41)	7.59(45)	0.0402(49)	0.0020(82)	0.0645(75)	0.0378(82)	0.0067(52)	0.0003(12)	0.0118(72)	0.0055(76)
roc-020	2026_02_06	0.67(90)	9.23(92)	0.0424(51)	0.0020(82)	0.0073(17)	0.0077(21)	0.0127(75)	0.0004(30)	0.0037(20)	0.0022(30)
paravision-021	2026_03_12	0.74(126)	13.11(126)	0.0442(54)	0.0020(82)	0.0116(27)	0.0045(13)	0.0051(42)	0.0003(8)	0.0037(20)	0.0007(7)
clearviewai-003	2025_11_24	0.62(37)	7.66(46)	0.0880(76)	0.0020(82)	0.0688(81)	0.0505(118)	0.0145(84)	0.0002(4)	0.0143(95)	0.0100(120)
intema-001	2023_02_22	0.67(85)	9.71(104)	0.1316(118)	0.0020(82)	0.0808(101)	0.0550(122)	0.0174(103)	0.0004(48)	0.0145(99)	0.0088(113)
onfido-000	2026_03_03	0.83(134)	27.41(134)	0.1450(124)	0.0019(17)	0.0200(42)	0.0040(10)	0.0232(122)	0.0002(4)	0.0075(50)	0.0006(6)
viante-002	2025_12_16	0.73(123)	12.23(123)	0.2087(128)	0.0020(82)	0.0363(53)	0.0278(53)	0.1322(128)	0.0010(118)	0.0109(65)	0.0118(123)

TWO DEMOGRAPHIC DIFFERENTIALS SUMMARY MEASURES

ONE THRESHOLD PER ALGORITHM
SET TO VALUE THAT GIVES FPIR ~ 0.002 ON E. EUROPEAN WOMEN.

FP VARIATION: MANAGEMENT + MITIGATION

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Algorithm	E-Asia F	E-Europe F	S-Asia F	W-Africa F	E-Asia M	E-Europe M	S-Asia M	W-Africa M
toshiba-00								
idemia-01								
cognitec-0								
nec-011								
incode-00								
cogent-00								
roc-020								
paravision-								
clearviewai-								
intema-00								
onfido-00								
viante-00								

1. Understand that FP rates may vary
   - Collect data on FP rates, consider public test results (below)
   - Look out for FP rate variation in operations
   - Find out who matches who
   - **Raise threshold to control FP**

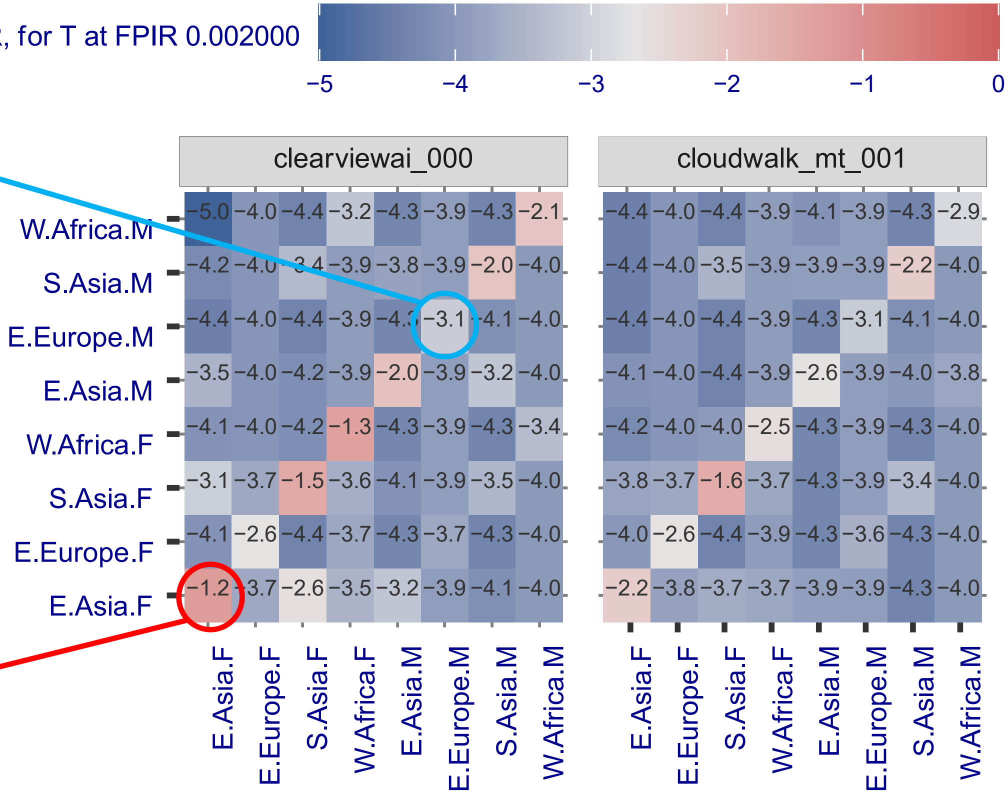
2. Avoid algorithms with FP rate variations across demographics
   - Engage algorithm developer to seek remediation

FALSE MATCHES: WHICH DEMOGRAPHIC GROUPS?

Cross Demographic FPIR, for T at FPIR 0.002000

FMR ~ 1 in 1300 on E. European Men

REGION OF BIRTH + SEX OF TOP NON-MATE

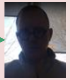


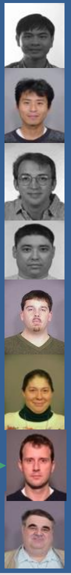
FPIR = 1 in 16 East Asian Women

REGION OF BIRTH + SEX OF PROBE

HUMAN REVIEW (+ APPLIED CASE INFO) AS MITIGATION OF FALSE POSITIVES IN 1:N IDENTIFICATION

SEARCH IMAGE





SCORE	RANK
3.142	1
2.998	2
1.626	3
0.707	4
0.330	5
0.198	6
0.074	7
0.016	8

MODE A: FR SYSTEM IS CONFIGURED TO RETURN A FIXED NUMBER OF CANDIDATES (HERE 8) REGARDLESS OF SCORE. THIS IS LIKE MOST TEXT SEARCHES WHERE WE REVIEW CANDIDATE DOCS.

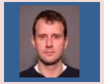
ANALOGY TO TEXT SEARCH

Source: https://www.google.com/doodles/jim-hensons-75th-birthday

"Investigation"
Returns candidates for human review

"Lights Out"
System concludes this the correct result

SEARCH IMAGE



SCORE	RANK
3.142	1

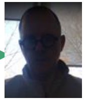
T

MODE B: FR SYSTEM CONFIGURED TO RETURN ONLY THOSE CANDIDATES WITH SCORE ABOVE THRESHOLD E.G. T = 3.0. THIS IS SIMILAR TO THE "I'M FEELING LUCKY" or "LIGHTS OUT" TEXT SEARCHES WITHOUT REVIEW.

HUMAN ROLE

OPERATOR LED INVESTIGATIONS

INVESTIGATION VS. (LIGHTS OUT) IDENTIFICATION



Search image

Score	Rank
3.142	1
2.998	2
1.626	3
0.707	4
0.330	5
0.198	6
0.074	7
0.016	8

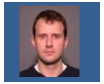
MATE

Policy A: Review upto 8 candidates

"Investigation"
Returns candidates for human review

"Lights Out"
System concludes this the correct result

Search image



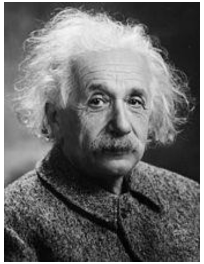
Score	Rank
3.142	1

T

Policy B: Review only candidates with score above threshold T = 3.0

FALSE POSITIVES AND SYSTEM BOUNDARIES

NIST — NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, U.S. DEPARTMENT OF COMMERCE — CELEBRATING 125 YEARS, EST. 1901



NON-MATE

1. Scene

ENVIRONMENTAL DESIGN
+
UNCONTROLLED VARIATION

2. Camera

OPTICAL DESIGN OPPORTUNITY

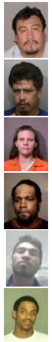
Enrollment Database

FR algorithm has False Alarm Rate = 100%

3. Face Recognition Algorithm

VAST ALGORITHM RESEARCH HERE

ACCURACY VARIES: BUYER BEWARE



4. Hypothesized candidate(s)

Human has False Alarm Rate > 0%

5. Human review including comparisons

CASE INFO, OTHER PHOTOS, BIOGRAPHICS,

PROCESS, DISPLAY, WORKFLOW, TIME ALLOWED, TRAINING, APTITUDE, GROUP WORK, FATIGUE

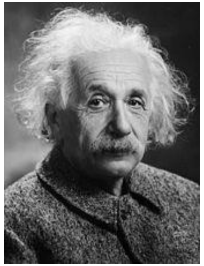
Possible outcomes:

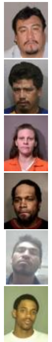
1. **TRUE NEGATIVE (not present)**
2. **FALSE POSITIVE (bad press, money)**

LAW ENFORCEMENT INVESTIGATIONS: SYSTEMS + BOUNDARIES

NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

Image Database





1. Scene

2. Camera

3. Face Recognition Algorithm

4. Fixed number of hypothesized candidate(s)

5. Human review including comparisons

SYSTEM = ALGORITHM + HUMAN

FALSE ALARM RATE = F($\text{FPIR}_{\text{ALGORITHM}}$, $\text{FPIR}_{\text{HUMAN}}$)

Possible outcomes:

1. TRUE POSITIVE (hit)
2. NO DECISION (caution)
3. FALSE POSITIVE (bad press, money)

FPIR = False Positive Identification Rate, synonym for False Alarm Rate

INCORRECT ARRESTS: MICHIGAN AND NEW JERSEY

PROBE

INCORRECT PERSON
ROBERT WILLIAMS

https://www.cbsnews.com/news/facial-recognition-60-minutes-2021-05-16/

INCORRECT PERSON
NIJEER PARKS

PROBE

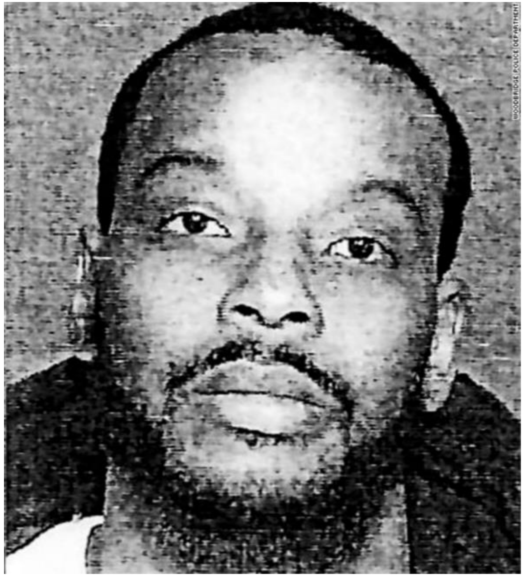
GALLERY RETRIEVED

NEWS PHOTO

https://www.cnn.com/2021/04/29/tech/nijeer-parks-facial-recognition-police-arrest/index.html

https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html

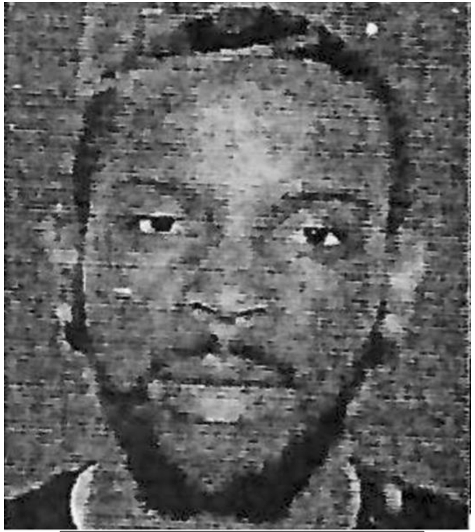
NIJEER PARKS: PHOTO QUALITY



NIJEER PARKS
IN GALLERY, N = 12M.

NIJEER PARKS
POST EXONERATION

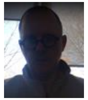
- 10 ALGS FIND GALLERY MATCH AT RANK 1, HIGH SCORE



SUSPECT PHOTO
NOT NIJEER PARKS

- 5 ALGS FIND GALLERY MATCH AT RANK 1, WEAK SCORE
- 1 ALG FINDS GALLERY MATCH AT RANK 8, WEAK SCORE

THE CLEARVIEW AI VERSION: T > 0



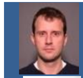
Search image

	Score	Rank
	43.1	1
	42.8	2
	16.26	3

STEP 1: T = 45

OUTPUT: NOTHING RETURNED

Search image

	Score	Rank
	43.1	1
	42.8	2
	16.26	3

STEP 2: (OPTIONAL) T = 42

OUTPUT: TWO LEADS

Search image | Reference image

STEP 3: A "LEAD" FOR HUMAN COMPARISON

STEP 4: INVESTIGATE WITH ANY CONTEXT INFO

STEP 5: A SECOND "LEAD" FOR HUMAN COMPARISON

STEP 6: INVESTIGATE WITH ANY CONTEXT INFO

SAME PERSON OF NOT? ONE-TO-ONE COMPARISON

- ❑ **ACTUAL: SAME IDENTITY**
- ❑ **ERROR: FALSE NEGATIVE**

- ❑ **ACTUAL: DIFFERENT-IDENTITY**
- ❑ **ERROR: FALSE POSITIVE**

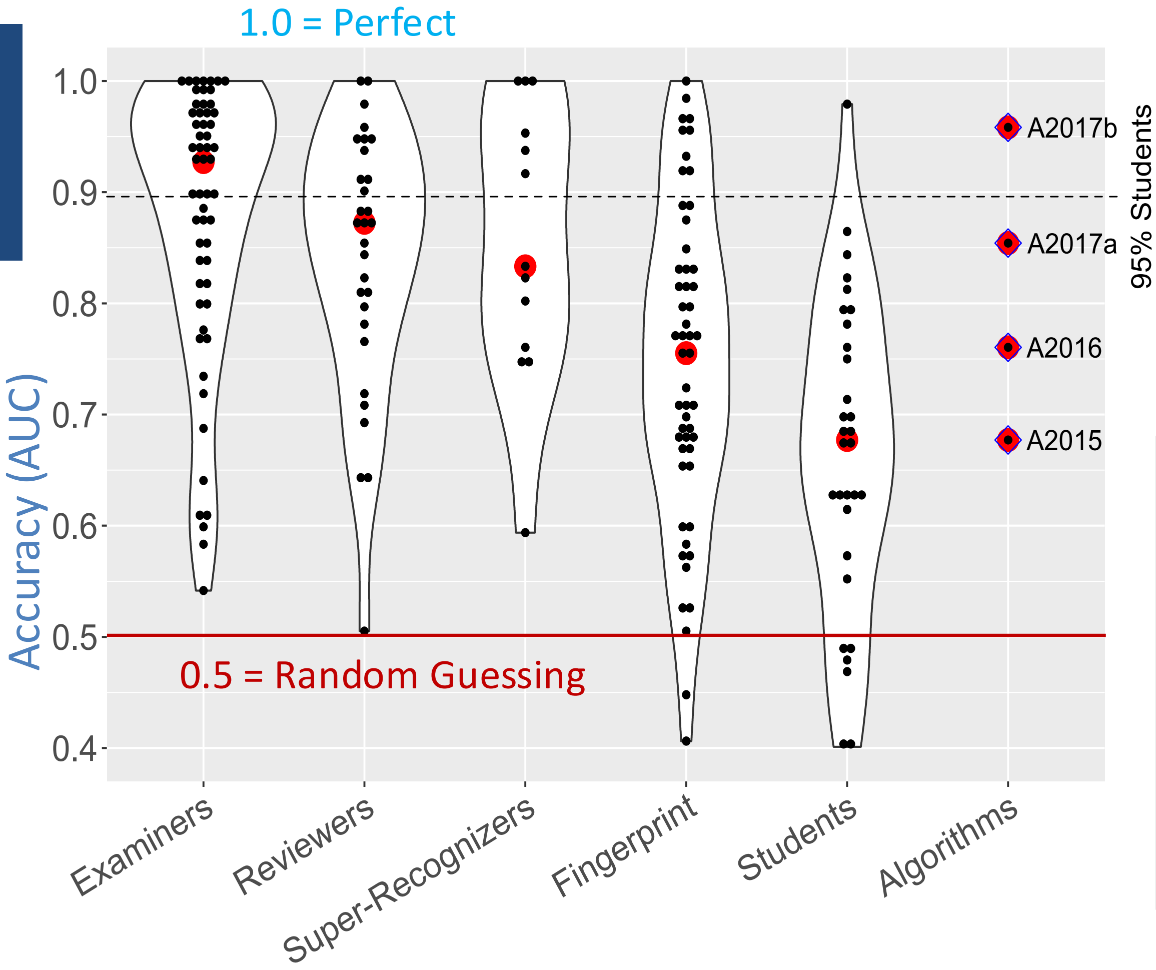
Image source:
Notre Dame via PNAS Phillips18

Images available:
https://www.pnas.org/content/115/24/6171/tab-figures-data

Takeaways:

- HUMAN COMPARISON OF PHOTOS IS HARD!
- TRY IT! https://facetest.psy.unsw.edu.au/

ALGORITHMS >> HUMANS



- Examiners, reviewers, super-recognizers statistically the same
- Each subject group has large range of accuracy
- Four groups had subjects with AUC=1

P. J. Philipps et al. *Face Recognition Accuracy of Forensic Examiners, Super-recognizers, and Machines*
Proceedings of the National Academy of Sciences, 2018

Human capability vs. automated FR

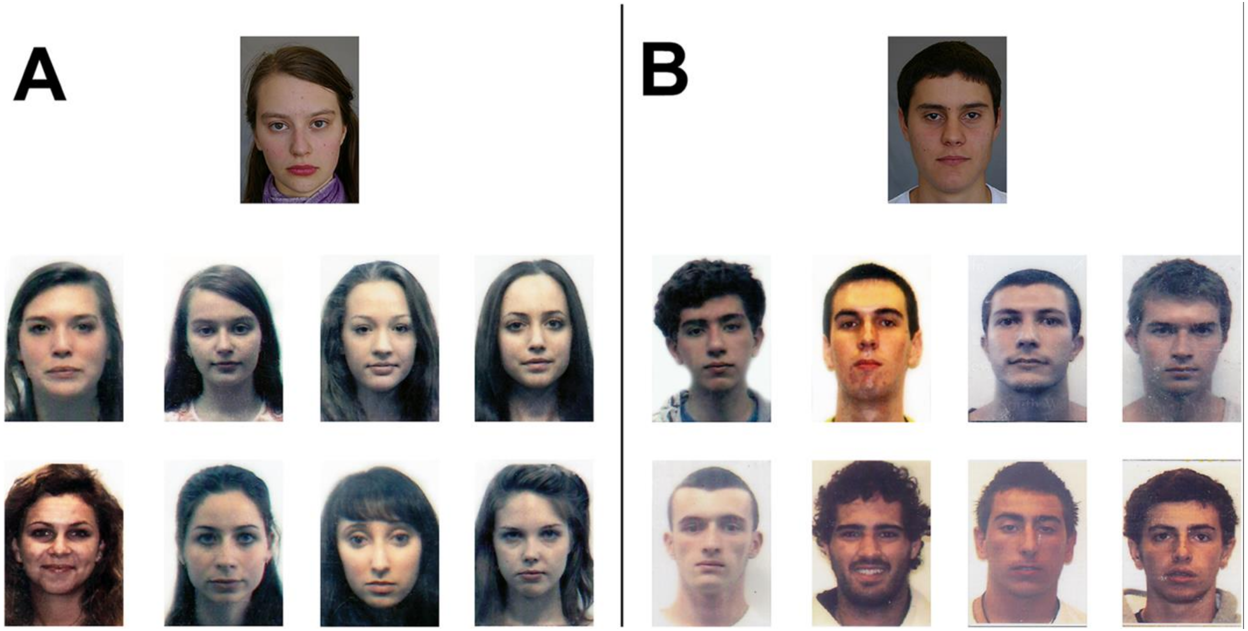
AFR is far superior to humans…

- In images of people we don't know
- especially when human is
  - operating under time pressures (e.g. border)
  - fatigued
  - untrained
  - looking at a novel demographic
- In searching large galleries
  - Human capacity ~ 5000 (variable)
  - AFR scales to arbitrary sizes – just dependent on computer availability

When humans may be superior to AFR

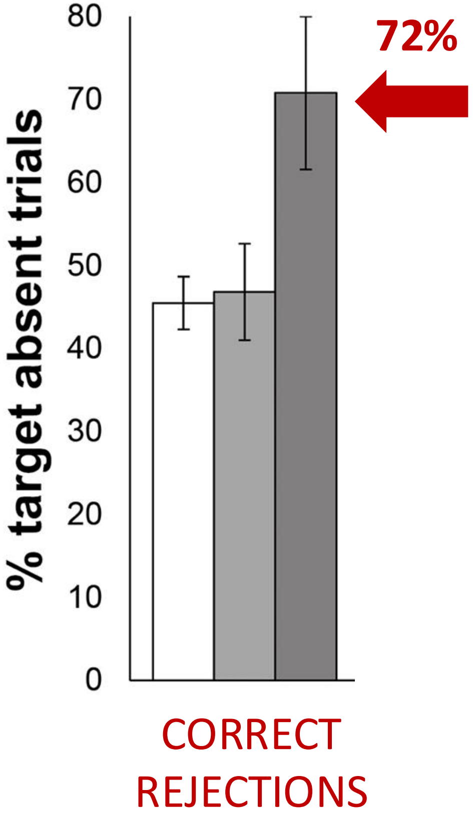
- In images/videos of people we know well
- In definitively **exculpating / excluding** someone saying "no match" e.g., by finding specific minor feature absent in one image but present in another
- On twins, siblings
  - (given adequate image quality)

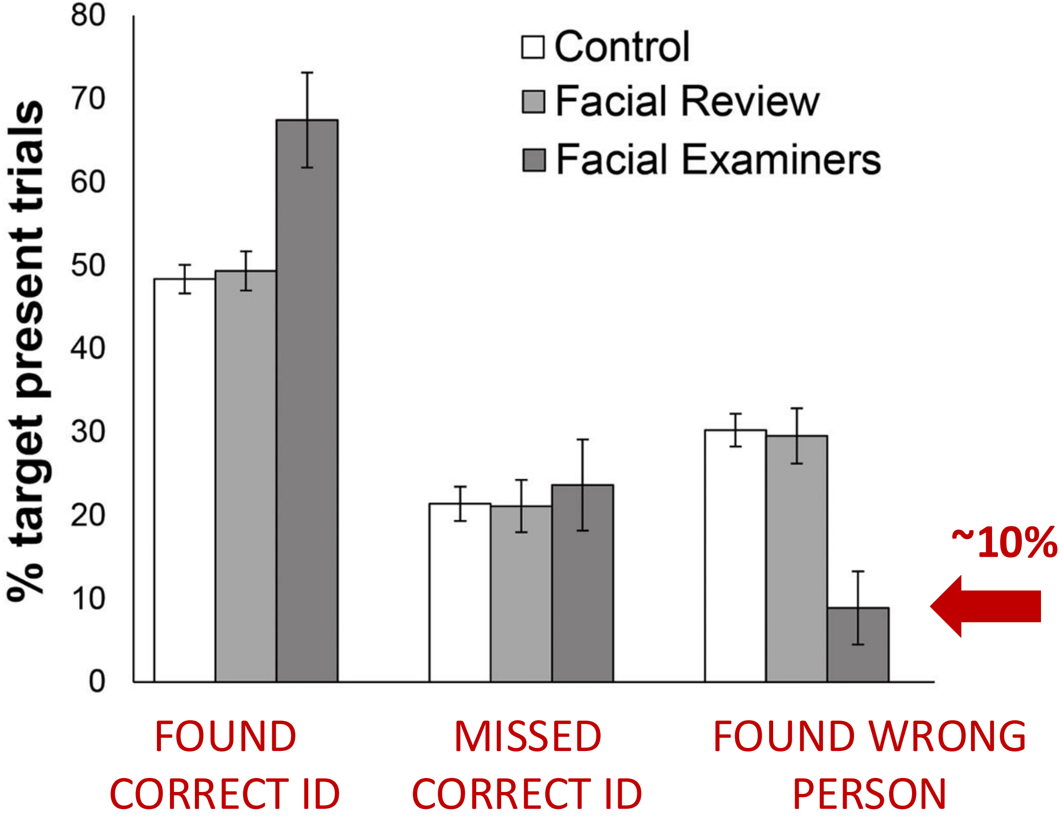
HUMAN ADJUDICATION OF UNFAMILIAR FACES IS HARD!



David White, James D. Dunn, Alexandra C. Schmid, and Richard I. Kemp.
Error rates in users of automatic face recognition software. PLoS ONE, October 2015.

HUMAN PERFORMANCE ON CANDIDATE LISTS, L = 8





SAME PERSON? OR NOT?

AGEING (UK) vs. AGING (US)

APPEARANCE CHANGES ⟶ REDUCED SIMILARITY

Images from presenter

ONE-TO-MANY MISS RATES BY AGE AND AGEING

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

Slow increase between 10-15 years

Higher error rates in teenagers:
Rapid ageing

AGE AT ENROLLMENT / TIME LAPSE BETWEEN SEARCH AND ENROLLMENT

Female

Age at enrollment	10	11	12	13	14	15
(60:99]	0.014	0.017	0.022	0.022	0.026	0.024
(45:60]	0.013	0.013	0.015	0.019	0.022	0.021
(30:45]	0.014	0.015	0.018	0.022	0.025	0.025
(21:30]	0.018	0.019	0.022	0.027	0.029	0.029
(18:21]	0.034	0.034	0.038	0.046	0.044	0.049
(15:18]	0.054	0.049	0.057	0.064	0.070	0.066
(12:15]	0.080	0.070	0.076	0.089	0.106	0.109

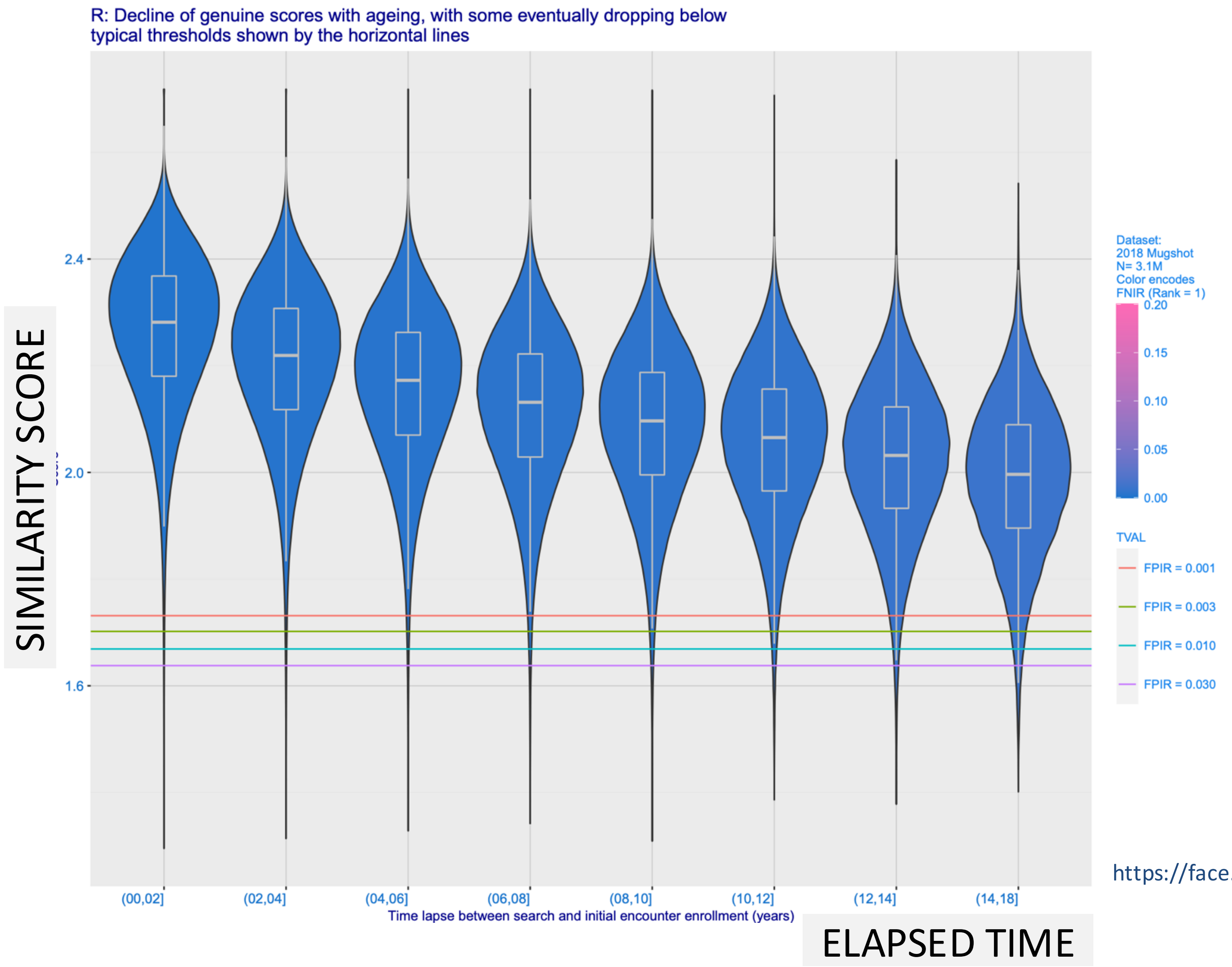
Male

Age at enrollment	10	11	12	13	14	15
(60:99]	0.008	0.008	0.009	0.012	0.009	0.010
(45:60]	0.006	0.007	0.008	0.010	0.011	0.012
(30:45]	0.007	0.006	0.008	0.010	0.012	0.013
(21:30]	0.010	0.010	0.011	0.014	0.017	0.019
(18:21]	0.030	0.031	0.036	0.041	0.053	0.046
(15:18]	0.090	0.091	0.101	0.132	0.145	0.150
(12:15]	0.174	0.160	0.206	0.237	0.252	0.246

Algorithm: Panasonic 2023_08

Gallery: AIR-ENTRY, N = 1.6 million, balanced by sex and by specific age-groups

Probes: AIR-ENTRY, 3.8 million searches, balanced by sex, age-group and time-lapse (years)

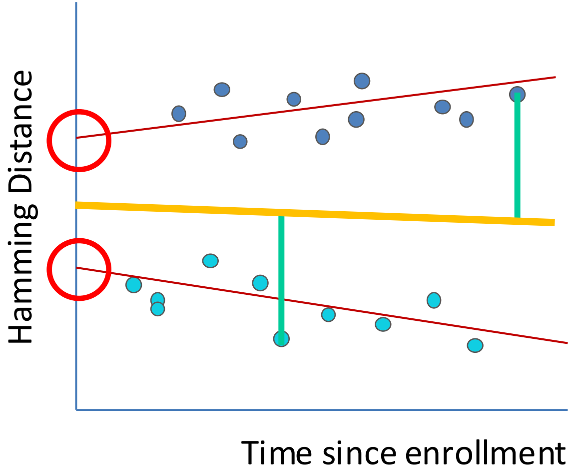


CELEBRATING 125 YEARS EST. 1901 NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

MUGSHOT AGEING EFFECT TO 18 YRS

https://face.nist.gov/frte/reportcards/1N/ntechlab_009.pdf

QUANTIFYING PERMANENCE: MIXED-EFFECTS REGRESSION



Model for the j-th score from the i-th eye

$$HD_{ij} = \pi_{0i} + \pi_{1i}T_{ij} + \epsilon_{ij}$$

Intercept is sum of population average term, the *fixed effect*, and an eye-specific *random effect*

$$\pi_{0i} = \gamma_{00} + \psi_{0i}$$

Slope is sum of population average term, the *fixed effect*, and an eye-specific *random effect*

$$\pi_{1i} = \underline{\gamma_{10}} + \psi_{1i}$$

Permanence stated by the population wide rate at which Hamming Distances are increasing.

Subject to assumptions:

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$$

$$\begin{bmatrix} \psi_{0i} \\ \psi_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_1^2 \end{bmatrix}\right)$$

MIXED EFFECTS MODEL RESPECT IDENTITY INFORMATION. SIMPLE LINEAR REGRESSION, IN YELLOW, DOES NOT AND HAS OTHER PROBLEMS

ONE-TO-MANY MISS RATES BY AGE GROUP

TAKEAWAYS

- 10 YEARS IS A LONG TIME FOR FACE RECOGNITION OF A TEEN
- BEST RECOGNITION ACCURACY IN AGES 30 TO 60

15.5% miss rate for recognition of females enrolled as a young teen, after 10 years of time lapse.

Algorithm: Clearview AI 2024-02

Gallery: AIR-ENTRY, N = 1.6 million, balanced by sex and by specific age-groups

Probes: AIR-ENTRY, 3.8 million searches, balanced by sex, age-group and time-lapse (years)

AGEING: SEARCH MISS RATES

PROBE TAKEN 10 YRS LATER

AGE AT ENROLLMENT	FEMALE	MALE
(60:99]	0.019	0.009
(45:60]	0.018	0.005
(30:45]	0.021	0.006
(21:30]	0.030	0.011
(18:21]	0.064	0.037
(15:18]	0.105	0.119
(12:15]	0.155	0.223

TAKEAWAYS:

1. 10 YEARS IS A LONG TIME FOR A TEEN
2. MOST ACCURATE AGES 30 TO 60
3. MEN EASIER TO RECOGNIZE
   - EXCEPT IN TEEN YEARS
4. SOME ALGORITHMS MUCH BETTER

MISS RATES:
15.5% FEMALE VS. 22.3% MALE

Algorithm: Clearview AI 2024-02

Gallery: AIR-ENTRY, N = 1.6 million, balanced by sex and by specific age-groups

Probes: AIR-ENTRY, 3.8 million searches, balanced by sex, age-group and time-lapse (years)

SEARCH MISS RATES BY ALGORITHM

AGE AT ENROLLMENT

PROBE TAKEN 10 YRS LATER

NEC-2023-12

AGE AT ENROLLMENT	FEMALE	MALE
(60:99]	0.006	0.003
(45:60]	0.003	0.002
(30:45]	0.004	0.002
(21:30]	0.004	0.003
(18:21]	0.006	0.003
(15:18]	0.009	0.009
(12:15]	0.014	0.014

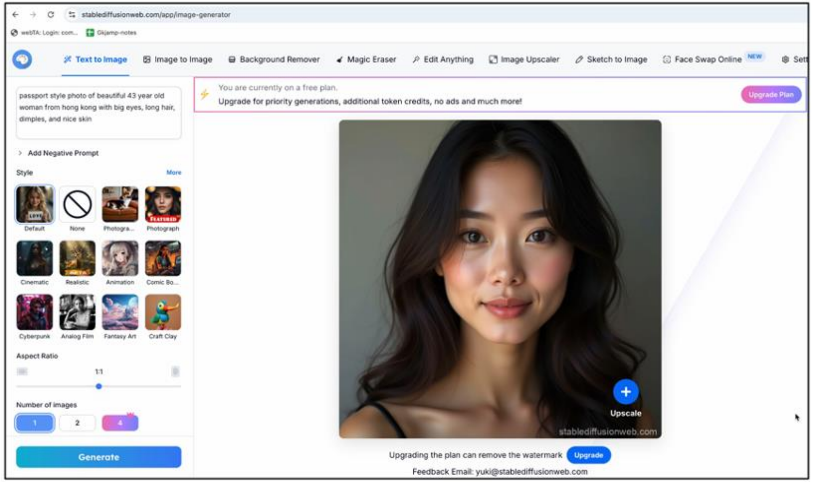
IDEMIA-2024-03

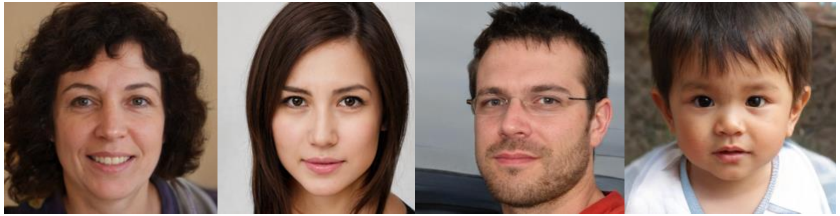
AGE AT ENROLLMENT	FEMALE	MALE
(60:99]	0.018	0.006
(45:60]	0.017	0.004
(30:45]	0.019	0.005
(21:30]	0.027	0.008
(18:21]	0.052	0.024
(15:18]	0.082	0.076
(12:15]	0.111	0.156

CLEARVIEW AI -2024-02

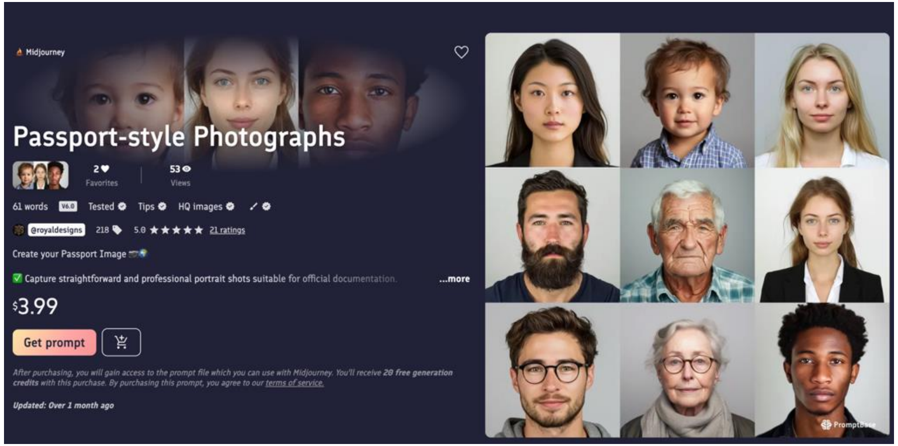
AGE AT ENROLLMENT	FEMALE	MALE
(60:99]	0.019	0.009
(45:60]	0.018	0.005
(30:45]	0.021	0.006
(21:30]	0.030	0.011
(18:21]	0.064	0.037
(15:18]	0.105	0.119
(12:15]	0.155	0.223

SYNTHETIC IMAGERY A.K.A. "DEEPFAKES"





https://thispersondoesnotexist.com/



https://promptbase.com/prompt/passportstyle-photography

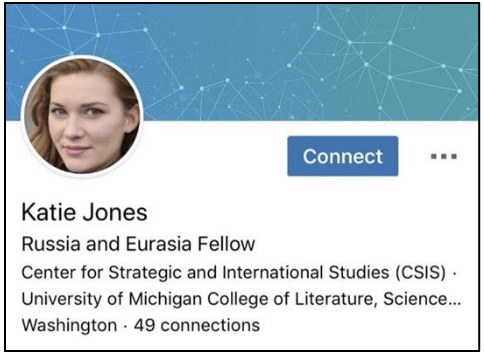
- Deepfakes ⟶ generation of fake identities
  - Many public fake generators
  - High quality

DEEPFAKES: A FAKE IDENTITY RISK

Checklist for determining if an image or video is fake:

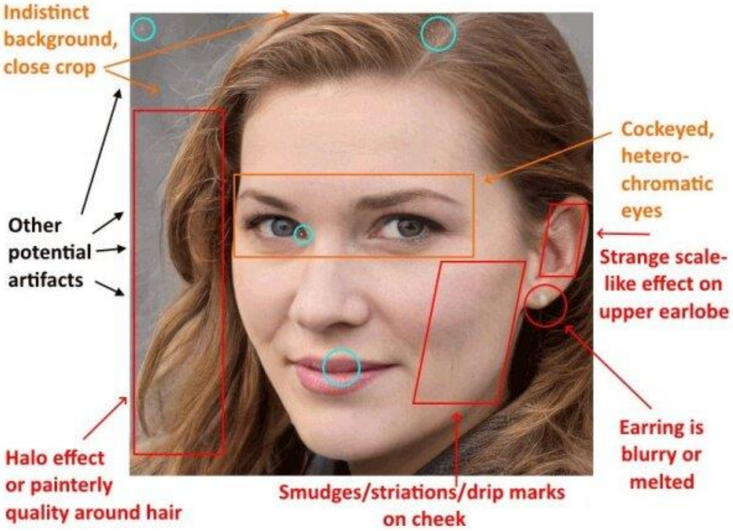
- Blurring evident in the face but not elsewhere (or vice-versa)
- Change of skin tone near the edge
- Double chins, double eyebrows, or double edges to the face
- Face is blurry when it is partially obscured by a hand
- Lower-quality sections in one video
- Box-like shapes, cropped effects near mouth, eyes, and neck
- Un-natural blinking (or lack thereof), movements
- Changes in the background and/or lighting
- Contextual clues – Background inconsistent with foreground and subject?

Source: U.S. Department of Homeland Security Report "[Increasing Threat of Deepfake Identities]"



NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

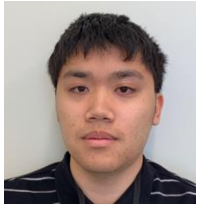
Source: MIT Tech Review



Source: AP

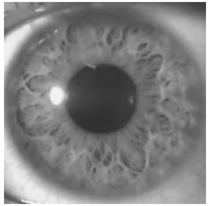
Patrick
Grother

Kayee Hanaoka
(+ Mei Ngan)



Austin
Hom

(not)
Mei Ngan



Jim
Matey

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

THANK YOU!

PATRICK.GROTHER@NIST.GOV
MEI@NIST.GOV