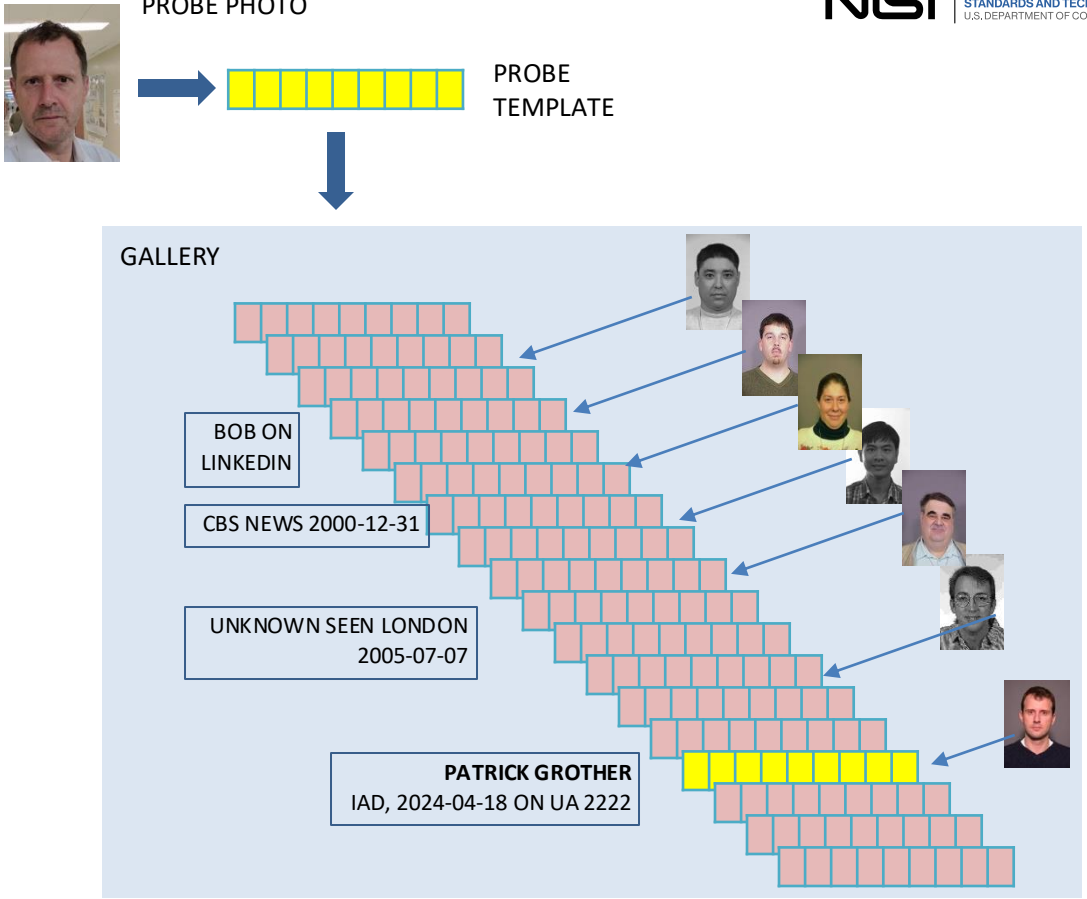
FRTE 1:N

CONTEXT AND EXPLANATION OF THE DEMOGRAPHICS LEADERBOARD

LAST UPDATE 2025-12-22



A GALLERY, OR ENROLLMENT DATABASE, CONSISTS OF **1.** BIOMETRIC TEMPLATES EXTRACTED FROM A COLLECTION OF PHOTOS, and **2.** ASSOCIATED METADATA WHICH MAY INCLUDE IDENTITY INFORMATION OR OTHER CONTEXT ABOUT THE GALLERY PHOTO.

BACKGROUND ON BIOMETRIC ERRORS - PART 1: FALSE NEGATIVES

WHAT IS A FALSE NEGATIVE?

- FAILURE TO ASSOCIATE TWO PHOTOS OF A PERSON

WHEN COMPARING TWO IMAGES OF ONE PERSON, WHAT SHOULD AN AUTOMATED FACE RECOGNITION ALGORITHM (AFR) PRODUCE?

- A HIGH SIMILARITY SCORE

WHAT'S THE STANDARD METRIC?

- IN ONE-TO-ONE COMPARISONS: FALSE NON-MATCH RATE (FNMR)
- IN ONE-TO-MANY SEARCH: FALSE NEGATIVE IDENTIFICATION RATE (FNIR)

1:1 - SO HOW IS FNMR MEASURED?

- CURATE A SET OF N PAIRS OF PHOTOS. EACH PAIR SHOULD BE OF THE SAME PERSON.
- RUN AFR ON THE PAIRS TO PRODUCE N SIMILARITY SCORES
- COMPUTE NUMBER OF SCORES BELOW THESHOLD AND DIVIDE BY N

1:N - SO HOW IS FNIR MEASURED?

- CURATE A GALLERY OF N PHOTOS.
- CURATE A SET OF P PROBES WHICH HAVE MATED ENTRIES IN THE GALLERY
- RUN AFR SEARCHES TO PRODUCE P CANDIDATE LISTS
- COMPUTE NUMBER OF LISTS WHERE THE MATE DOES NOT APPEAR WITH SCORE AT OR ABOVE THRESHOLD, THEN DIVIDE BY P

STANDARD METRICS: ISO/IEC 19795-1:2021 [ANSI, ISO]

BACKGROUND ON BIOMETRIC ERRORS - PART 2: FALSE POSITIVES

WHAT IS A FALSE POSITIVE?

- INCORRECT ASSOCIATION OF PHOTOS OF TWO PEOPLE

WHEN COMPARING TWO IMAGES OF DIFFERENT PEOPLE, WHAT SHOULD AN AUTOMATED FACE RECOGNITION ALGORITHM (AFR) PRODUCE?

- A LOW SIMILARITY SCORE

WHAT'S THE STANDARD METRIC?

- IN ONE-TO-ONE COMPARISONS: FALSE MATCH RATE (FMR)
- IN ONE-TO-MANY SEARCH: FALSE POSITIVE IDENTIFICATION RATE (FPIR)

1:1 - SO HOW IS FMR MEASURED?

- CURATE A SET OF N PAIRS OF PHOTOS OF DIFFERENT PEOPLE - SAME SEX, AGE, ETHNICITY
- RUN AFR ON THE PAIRS TO PRODUCE N SIMILARITY SCORES
- COMPUTE NUMBER OF SCORES AT OR ABOVE THESHOLD AND DIVIDE BY N

1:N - SO HOW IS FPIR MEASURED?

- CURATE A GALLERY OF N PHOTOS.
- CURATE A SET OF P PROBES WHICH DO NOT HAVE A MATE IN THE GALLERY
- RUN AFR SEARCHES TO PRODUCE P CANDIDATE LISTS
- COMPUTE NUMBER OF LISTS WHERE ANY CANDIDATE APPEARS WITH SCORE AT OR ABOVE THRESHOLD, THEN DIVIDE BY P

STANDARD METRICS: ISO/IEC 19795-1:2021 [ANSI, ISO]

BACKGROUND: REVIEW OF ACCURACY METRICS

1:1 COMPARISON

- **FALSE MATCH RATE (FMR)** QUANTIFIES HOW OFTEN COMPARISON OF IMAGES OF TWO PEOPLE PRODUCES A SIMILARITY SCORE AT OR ABOVE A FIXED THRESHOLD T.
- FMR IS ESTIMATED BY EXECUTING MANY COMPARISONS OF IMAGES OF DIFFERENT PEOPLE.
  - IN ANY GIVEN COMPARISON, THE IMAGE PAIR SHOULD BE OF SIMILAR AGE, SEX, ETHNICITY.

- **FALSE NON-MATCH RATE (FNMR)** QUANTIFIES HOW OFTEN COMPARISON OF TWO IMAGES OF ONE PERSON PRODUCES A SIMILARITY SCORE BELOW THE FIXED THRESHOLD T.
- FNMR IS ESTIMATED BY EXECUTING MANY COMPARISONS OF IMAGES OF THE SAME PERSON.

- FALSE REJECT RATE (FFR) AND FALSE ACCEPT RATE (FAR) ARE ANALOGOUS TO FNMR AND FMR BUT ARE RESERVED BY ISO/IEC 19795-1 FOR TRANSACTIONS WHERE SEVERAL MATCH ATTEMPTS MAY BE MADE.

1:N SEARCH

- **FALSE POSITIVE IDENTIFICATION RATE (FPIR)** QUANTIFIES HOW OFTEN SEARCHES OF PERSONS NOT PRESENT IN AN ENROLLED DATABASE YIELD INCORRECT IDENTITIES AT OR ABOVE SOME THRESHOLD T.
- FPIR IS ESTIMATED BY EXECUTING MANY NON-MATED SEARCHES OF A GALLERY OF SIZE N.
- FPIR IS ALSO KNOWN AS “FALSE ALARM RATE”

- **FALSE NEGATIVE IDENTIFICATION RATE (FNIR)** QUANTIFIES HOW OFTEN SEARCHES OF PERSONS PRESENT IN AN ENROLLED DATABASE FAIL TO RETURN THE CORRECT MATED IDENTITY AT OR ABOVE SOME THRESHOLD T.
- FNIR IS ESTIMATED BY EXECUTING MANY MATED SEARCHES OF A GALLERY OF SIZE N.
- FNIR IS ALSO KNOWN AS “MISS RATE”, and 1-FNIR IS “HIT RATE”

- WHY 1:N IS A MORE DIFFICULT TASK THAN 1:1
  - FOR SEARCHES WHERE THERE IS NO MATE, THE ALGORITHM MUST ASSIGN LOW SIMILARITY SCORES TO **ALL N** ENROLLED IDENTITIES.
  - FOR SEARCHES WHERE THERE IS A MATE, THE ALGORITHM MUST ASSIGN LOW SIMLARITY SCORES TO ALL EXCEPT THE MATED ENTRY

ERROR RATE TRADEOFF
THRESHOLD SETTING

- ERROR RATES ARE DEFINED IN TERMS OF A SCORE THRESHOLD, T.
  - AT HIGH T: FPIR IS LOW, FNIR IS HIGH
  - AT LOW T: FPIR IS HIGH, FNIR IS LOW
- THE THRESHOLD IS USUALLY SET TO TARGET SOME PARTICULAR FPIR POLICY, WHICH WILL DEPEND ON THE APPLICATION
  - POLICY CAN SE SET TO MEET A SECURITY OBJECTIVE SAY
  - OR MAY BE SET TO LIMIT HOW MANY CANDIDATES ARE REFERRED TO A POOL OF HUMAN REVIEWERS
- THE THRESHOLD VALUE HAS A DEVELOPER-DEFINED RANGE OF VALUES: SOME USE SIMILARITY SCORES ON RANGE [0,1], OTHERS ON [0,100], OR [300,10000].
  - THIS MEANS THRESHOLD VALUES ARE DEVELOPER, ALGORITHM AND (OFTEN) VERSION DEPENDENT
  - SIMILARITY SCORES CANNOT BE UNDERSTOOD AS PROBABILITIES OF MATCH, OR NON-MATCH

- **FNIR AND FPIR CAN BE LARGER IN ONE DEMOGRAPHIC GROUP THAN ANOTHER - NEXT SLIDE**
  - **THIS IS THE HEART OF THE BIAS ISSUE**
  - **THE TERM DEMOGRAPHIC DIFFERENTIAL IS MORE PRECISE THAN "BIAS"** WHICH HAS OVERLOADED MEANINGS IN STATISTICS, SOCIOLOGY, ELECTRICAL ENGINEERING, AI, AND ELSEWHERE

ONE ALGORITHM, TWO DEMOGRAPHIC GROUPS, TWO ERROR RATES

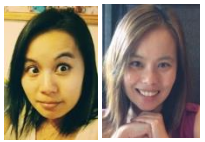
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

FNIR
FALSE NEGATIVE IDENTIFICATION RATE

PROPORTION OF MATED SEARCHES FAILING TO RETURN MATE WITH SCORE AT OR ABOVE THRESHOLD, T.

SEE ISO/IEC 19795-1

EVERY SEARCH IS RUN WITH THE SAME FIXED THRESHOLD - HERE IT IS SET TO T = 3.142 (A TOY VALUE)
A SYSTEM USES ONE FIXED THRESHOLD BECAUSE IT USUALLY HAS NO MEANS TO RE-CONFIGURE ITSELF ACCORDING TO WHO IS IN THE PROBE PHOTO



ALGORITHM X, DEMOGRAPHIC 2
- HIGHER FPIR
- LOWER FNIR

Δ FPIR

Δ FNIR

T = 3.142

T = 3.142

ALGORITHM X, DEMOGRAPHIC 1
- LOWER FPIR
- HIGHER FNIR

TRADEOFF BETWEEN FNIR AND FPIR: IF ONE RATE GOES UP, THE OTHER MUST GO DOWN

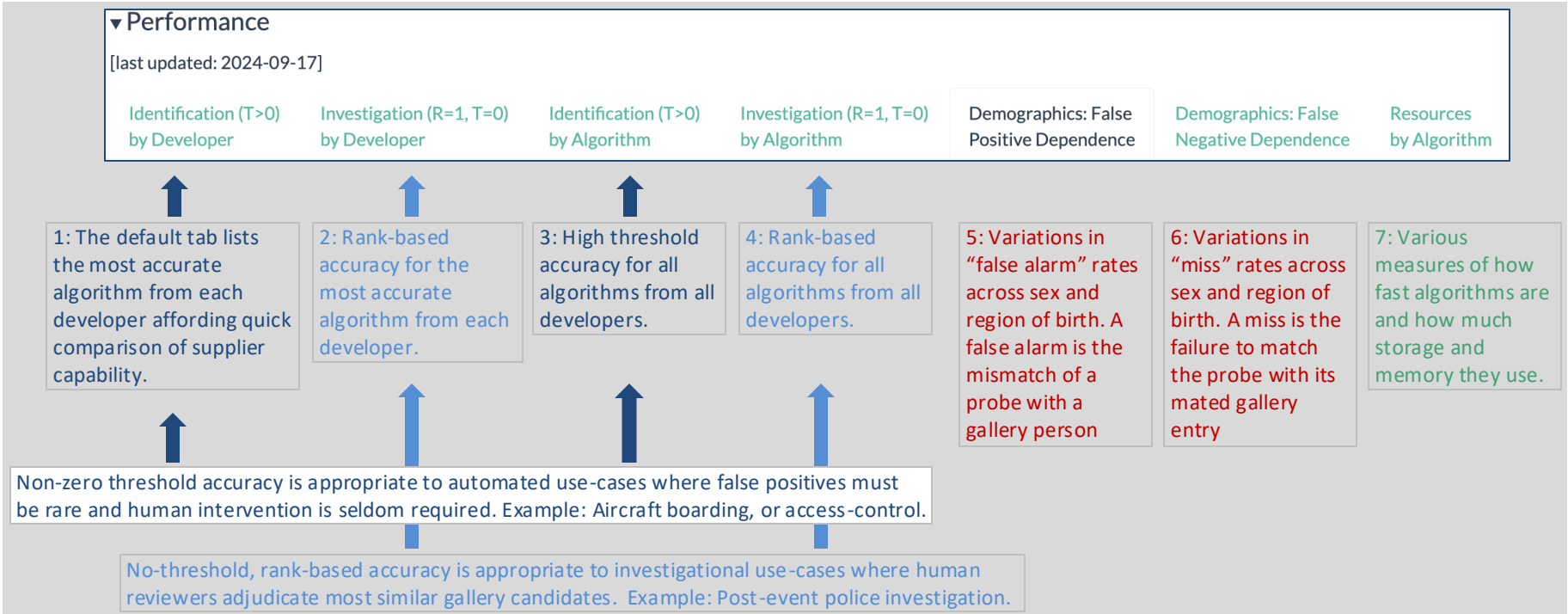
LOW FPIR VALUES ACHIEVED WITH HIGHER, I.E. MORE STRINGENT, THRESHOLDS.

LOG-SCALE IS OFTEN REQUIRED BECAUSE LOW FPIR VALUES ARE OPERATIONALLY RELEVANT.

FPIR FALSE POSITIVE IDENTIFICATION RATE
PROPORTION OF NON-MATE SEARCHES YIELDING ANY CANDIDATES AT OR ABOVE THRESHOLD, T.

NAVIGATING THE 1:N LEADERBOARD

PERFORMANCE DATA APPEARS IN SEVEN TABS, EACH OF WHICH INCLUDES A RESULTS TABLE AND SOME NARRATIVE



DEMOGRAPHICS IN 1:N EXPERIMENTAL SETUP

- WHICH IMAGES
- HOW MANY IMAGES, PEOPLE
- WHICH DEMOGRAPHIC GROUPS
- GALLERY COMPOSITION
- PROBE SET COMPOSITION

CONSIDER SEX AND ETNICITY

SEX:	IS PRESENT IN THE AVAILABLE METADATA
ETHNICITY:	IS APPROXIMATED BY COUNTRY OF BIRTH (NEXT SLIDE)
AGE:	HAS AN EFFECT TOO BUT THIS IS NOT CONSIDERED HERE

COUNTRIES SELECTED (NEXT SLIDE)

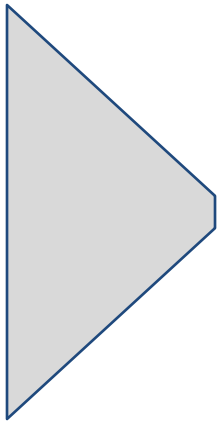
- THAT HAVE LOW LEVELS OF RECENT TRANSCONTINENTAL MIGRATION (UNLIKE FOR EXAMPLE, FRANCE, UK, USA)
- THREE CONTINENTS: AFRICA, EUROPE, AND SOUTH AND EAST ASIA SEPARATED BY THE HIMALAYAS
- THAT HAVE LARGE NUMBERS OF IMAGES IN PARENT DATABASE

WHICH DEMOGRAPHIC GROUPS: COUNTRY OF BIRTH AS PROXY FOR ETHNICITY

NOTIONAL POWER OF SINGLE-WORD REGION LABELS TO DESCRIBE PEOPLE OF ETHNIC ORIGINS

Very High Resolution

Nigeria 2018
Hausa 30%
Yoruba 15.5%
Igbo 15.2%
Ijaw/Izon 1.8%
Ibibio 1.8%
Tiv 2.4%
Fulani 6%
Kanuri 2.4%
Others 24.7%
Source: https://www.cia.gov/the-world-factbook/countries/nigeria/



NIST DOES NOT HAVE SUCH FINE-GRAINED ETHNICITY METADATA

High Resolution

Poland Hungary
Russia Romania
Ukraine Czechia

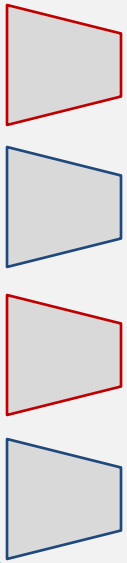
Benin Togo
Ghana Male
Liberia Sierra Leone
Nigeria Senegal

Korea Taiwan
China
Japan

Afghanistan Pakistan
India Bangladesh
Myanmar
Nepal

NIST HAS COUNTRY OF BIRTH METADATA

Medium Resolution



1. E. EUROPE
2. W. AFRICA
3. E. ASIA
4. S. ASIA

NIST GROUPS IMAGES FROM SEVERAL COUNTRIES TO INCREASE SAMPLE SIZE, AND TO EASE INTERPRETATION

Low Resolution

ASIA

SUCH GROUPING IS NOT DONE BECAUSE IT WOULD TOO COARSELY COMBINE PEOPLE EXHIBITING DIFFERENT PHENOTYPES

FRTE 1:N :: GALLERY USED IN DEMOGRAPHICS

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

GALLERY PROPERTIES

- 727975 IMAGES
- 349746 PEOPLE
- ADULTS 18+
- BORDER-CROSSING IMAGES

- VARIABLE NUMBER OF IMAGES PER PERSON
  - UNCONSOLIDATED: THE IMAGES OF A PERSON ARE ENROLLED UNDER DIFFERENT RANDOM IDENTIFIERS - ALGORITHM IS NOT TOLD THAT IMAGES X AND Y ARE FROM THE SAME PERSON.
  - DETAIL: SEE NIST IR 8272 SEC 2.3.

- EIGHT DEMOGRAPHIC GROUPS
  - **2 SEXES**
  - **4 REGIONS** OF BIRTH: EAST ASIA, SOUTH ASIA, WEST AFRICA, EAST EUROPE.
  - **IMBALANCED**: UNEQUAL NUMBERS OF EACH GROUP, AS IS TYPICAL IN OPERATIONS.
  - NUMBERS OF GALLERY ENTRIES CAN AFFECT FPIR
    - "MORE CHANCES" TO MAKE A FALSE MATCH

NUMBERS OF PEOPLE IN 8 GALLERY GROUPS = 2 SEXES x 4 REGIONS OF BIRTH

27994 F EAST EUROPE
17188 M EAST EUROPE
63412 F SOUTH ASIA
64094 M SOUTH ASIA
10675 F WEST AFRICA
12163 M WEST AFRICA
94115 F EAST ASIA
61147 M EAST ASIA

350788: TOTAL

TOTAL DIFFERS FROM 349746 BECAUSE SOME (0.2%) INDIVIDUALS HAVE TWO OR MORE GALLERY PHOTOS WITH DIFFERENT SEX METADATA - THIS COULD HAVE OCCURRED DUE TO A CHANGE OF SELF-REPORTED SEX OR TYPOGRAPHICAL ERROR.

LARGEST GROUP IS NEARLY 9 TIMES MORE PEOPLE THAN SMALLEST

BORDER IMAGES:

- EXAMPLES HERE FROM MEDS DATASET
- MEDIUM QUALITY
  - HEAD POSE
  - CROPPING
  - EXPOSURE

- ARE USED IN FRTE 1:N ACCURACY LEADERBOARD AS THE WEBCAM PART OF THE MUGSHOT-WEBCAM DATASET

FRTE 1:N :: PROBE SET USED IN DEMOGRAPHICS

PROBE SET PROPERTIES

- 858098 IMAGES
- 654431 PEOPLE, SOME WITH GALLERY MATE, SOME WITHOUT
- ADULTS 18+
- APPLICATION OFFICE IMAGES

- RANDOMLY SEQUENCED SEARCHES
  - MATED SEARCHES ⟶ FNIR ESTIMATES
  - NON-MATED SEARCHES ⟶ FPIR ESTIMATES

- SAME EIGHT DEMOGRAPHIC GROUPS AS GALLERY
  - IMBALANCED - NOTE THAT, UNLIKE THE GALLERY, IMBALANCE ONLY MATTERS TO UNCERTAINTY IN FNIR AND FPIR ESTIMATES.

APPLICATION IMAGES:

- EXAMPLES ARE FROM NIST STAFF
- HIGH QUALITY
  - "VISA-LIKE"
  - DEDICATED CAMERA
  - ATTENDED CAPTURE
  - UNIFORM POSE
  - NO GLASSES
  - GOOD EXPOSURE

- ARE USED IN FRTE 1:N ACCURACY LEADERBOARD AS THE VISA PART OF THE VISA-BORDER DATASET

DEMOGRAPHICS #1
THE FALSE NEGATIVE ASPECT

WHO	SEARCHES OF PERSON WITH A MATE IN THE DATABASE
ERROR	THEY ARE NOT RETURNED IN A SEARCH
METRIC	FALSE NEGATIVE IDENTIFICATION RATE (FNIR)
DIFFERENTIAL	FNIR FOR GROUP 1 > FNIR FOR GROUP 2

IMPACT OF A FALSE NEGATIVE IS APPLICATION DEPENDENT:

- **EXAMPLE 1:** ACCESS CONTROL TO A FACILITY: IF A PERSON IS NOT MATCHED TO A GALLERY ENTRY, THEY ARE INCONVENIENCED BY HAVING TO RETRY OR SEEK AN ALTERNATIVE WAY TO AUTHENTICATE.
- **EXAMPLE 2:** A PERSON APPLIES FOR A VISA UNDER A DIFFERENT NAME, HAVING BEEN DENIED PREVIOUSLY. IF A 1:N SEARCH FAILS, THE PERSON BENEFITS FROM GETTING THE VISA; THE COUNTRY IS DISADVANTAGED DEPENDING ON SUBJECT'S INTENT.
- **EXAMPLE 3:** CRIMINAL INVESTIGATION: IF A PHOTO TAKEN AT A CRIME SCENE IS NOT-MATCHED, THE PERSON (WHETHER INVOLVED IN THE CRIME OR NOT) IS ADVANTAGED BY POSSIBLY NOT BEING FURTHER INVESTIGATED. SOCIETY, ON THE OTHER HAND, IS DISADVANTAGED BY A FAILURE TO MATCH IF THAT PERSON WAS INDEED GUILTY OF THE CRIME.

DEMOGRAPHICS: THE FALSE NEGATIVE TAB

The tab shows false negative identification rates (FNIR), the proportion of searches of given demographic group for which the correct mated identity is not returned by the algorithm above a threshold. Th threshold is set for each algorithm to give FPIR = 0.002 on women born in E. Europe. The threshold is is used across all demographic groups - this is an operational necessity.

SHOW MORE HERE

Show 8 entries

CLICK ARROWS TO SORT

Search: ex. 2023|eyematic

SELECT BY NAME OR DATE HERE

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
qnap-006	2024_08_09	0.32	2.32	0.0096(70)	0.0077(72)	0.0106(71)	0.0155(71)	0.0032(65)	0.0037(69)	0.0046(67)	0.0060(65)
hisign-001	2024_08_09	0.23	1.87	0.0049(61)	0.0059(67)	0.0068(61)	0.0096(65)	0.0028(63)	0.0037(69)	0.0039(61)	0.0063(66)
psl-002	2024_07_26	0.21	1.67	0.0010(1)	0.0008(2)	0.0013(2)	0.0016(5)	0.0007(6)	0.0007(14)	0.0006(3)	0.0012(3)
optiexacta-000	2024_07_23	0.31	2.07	0.0012(6)	0.0009(5)	0.0021(14)	0.0014(4)	0.0006(2)	0.0004(4)	0.0006(4)	0.0017(16)
omnigarde-001	2024_06_25	0.21	1.64	0.0016(18)	0.0017(27)	0.0020(11)	0.0024(20)	0.0010(28)	0.0008(16)	0.0011(27)	0.0017(16)
roc-017	2024_06_24	0.26	1.77	0.0021(30)	0.0020(36)	0.0034(44)	0.0038(42)	0.0013(40)	0.0012(34)	0.0017(48)	0.0034(53)
clearviewai-002	2024_06_18	0.26	1.91	0.0020(25)	0.0016(25)	0.0029(32)	0.0023(18)	0.0009(22)	0.0008(20)	0.0010(20)	0.0017(19)

1: THE NAME OF THE PROTOTYPE ALGORITHM SUBMITTED TO NIST.

2: THE DATE OF SUBMISSION TO NIST

3: SUMMARY INDICATORS OF INEQUITY, QUANTIFYING VARIATION ACROSS NEXT EIGHT COLUMNS.

- SMALLER VALUES ARE BETTER
- **MXOG** EXPRESSES THE HIGHEST FNIR AS A MULTIPLE OF THE GEOMETRIC MEAN
- **GINI** MEASURES SPREAD.
- SEE NEXT SLIDE, NIST INTERAGENCY REPORT 8429, AND ISO/IEC 19795-10:2024 [ANSI, ISO]

4: FALSE NEGATIVE MISS RATES BY DEMOGRAPHIC GROUP

- HIGHER VALUES ARE WORSE, IDEAL VALUE IS 0.
- VARIATION ACROSS ROWS INDICATES POTENTIAL DIFFERENTIAL IMPACT

TWO DEMOGRAPHIC SUMMARY MEASURES

GINI COEFFICIENT

- Given n demographic groups d_i and estimates of an error rate for that group such as FMR_i

$$\mathrm{GINI}(\tau) = \frac{\sum_i \sum_j |\mathrm{FMR}_{d_i}(\tau) - \mathrm{FMR}_{d_j}(\tau)|}{2n(n-1)\,\mathrm{FMR}^{\diamond}} \quad (1)$$

- GINI is on the range [0,1] with smaller values indicating more uniform error rates across demographic groups.
- GINI has been used in economics for a century to quantify wealth or income disparity.

MAXIMUM OVER GEOMETRIC MEAN

- Given n demographic groups d_i and estimates of an error rate for that group such as FMR_i

$$\mathrm{MXOG}(\tau) = \frac{\max_{d_i} \mathrm{FMR}_{d_i}(\tau)}{\mathrm{FMR}^{\dagger}} \quad (2)$$

- MXOG simply states how many times larger the worst-case error rate is above the geometric mean of all error rates.
- Lower values imply more uniform error rates

BOTH SUMMARY MEASURES ARE:

1. MANDATED IN ISO/IEC 19795-10:2024
2. DETAILED IN NIST Interagency Report 8429

DEMOGRAPHICS #2
THE FALSE POSITIVE ASPECT

WHO: PEOPLE NOT IN THE DATABASE
HAZARD: THEY MATCH SOMEONE WHO IS

IMPACT: EITHER OR BOTH PEOPLE

IMPACT OF A FALSE POSITIVE IS APPLICATION DEPENDENT:

- **EXAMPLE 1:** CASINO USE TO DETECT "HIGH-ROLLERS": IF A PERSON WAS INCORRECTED MATCHED TO A GALLERY ENTRY, THEY COULD BE ADVANTAGED BY, SAY, THE CASINO OFFERING FREE HOSPITALITY. THE CASINO COULD BE DISADVANTAGED BY LESS-THAN-EXPECTED REVENUE.

- **EXAMPLE 2:** AIRCRAFT BOARDING:
  - 1. IF A STOWAWAY INCORRECTLY MATCHES ANY GALLERY ENTRY, THEY COULD BE ADVANTAGED BY TRAVELING WITHOUT A TICKET.
  - 2. IF A TRAVELER GOES TO THE WRONG GATE, MATCHES A GALLERY ENTRY, THEY COULD GO TO THE WRONG DESTINATION.
  - THE AIRLINE COULD BE DISADVANTAGED IN TERMS OF, SAY, DELAYED BOARDING, OR A FINANCIAL PENALTY FOR SENDING PERSON TO A COUNTRY FOR WHICH THEY ARE INADMISSABLE

- **EXAMPLE 3:** LIVE SURVEILLANCE: IF A FACE IN REAL TIME VIDEO IS MATCHED TO A GALLERY ENTRY, AND THAT PERSON COULD BE EVICTED FROM SAY A SPORTS ARENA VENUE, PHARMACY (NEXT SLIDE), OR EVEN DETAINED INCORRECTLY. THE OPERATOR COULD, FOR EXAMPLE, BE DISADVANTAGED BY SUBSEQUENT LITIGATION OVER LIABILITIES.

1:N SEARCH :: ONE EXAMPLE OF 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
By Eduardo Medina, Dec. 21, 2023

FTC REPORTS THAT "THE SYSTEM GENERATED THOUSANDS OF FALSE-POSITIVE MATCHES"

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

DEMOGRAPHICS: THE FALSE POSITIVE TAB

- False positive dependence on demographics is shown in below. It is derivedfrom searches of high quality visa-like application photo searches into a gallery of medium quality border crossing photos. The gallery is unconsolidated and imbalanced.
- The **FPIR** values are direct measurements of false positive identification rate. These rates depend, in part, on:
  1. The number of people of each demographic group in the gallery – see prior slide.
  2. The ability of the algorithm to distinguish two individuals from the particular demographic group.
- Point (1) above means that the FPIR table does not provide generalizable insight into the many other galleries that face recognition would be used with. The table is useful for comparing algorithms*on this specific gallery*. It useful also for revealing how any one algorithm does not produce equal FPIR values.

⚠ Technical: Later slides show how FPIR can depend on the gallery composition and the underlying one-to-one false match rates.

They also show how – for some algorithms – one might infer those underlying false match rates. A more immediate way to find false match rates is to look at the FRTE one-to-one demographic results which apply if the same underlying model was submitted to this 1:N test

The table shows false positive identification rates (FPIR), the fraction of searches of a given demographic that incorrectly return any non-mated gallery entry above a threshold. The threshold is set for each algorithm to give a FPIR of 0.002 (1 in 500) or less on searches of women born in Eastern Europe. In columns three and four are inequity summaries required by ISO/IEC 19795-10:2024: The Gini coefficient is on the range [0,1]; MXOG is the maximum of eight FPIR values divided by their geometric mean. For both measures lower values are better.

Show 8 entries Search: ex. 2023|eyematic

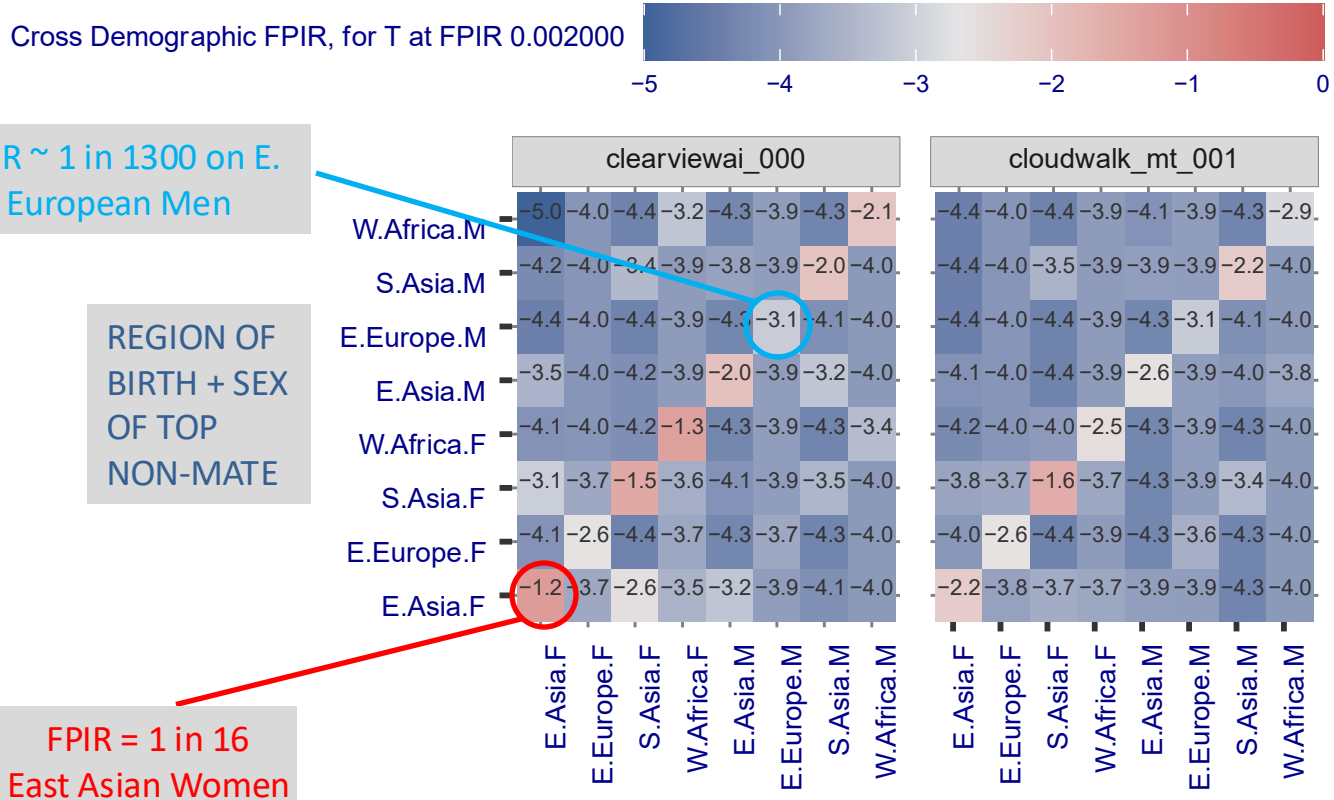
Algorithm	Date	GINI	MXOG	E-Asia F	E-Europe F	S-Asia F	W-Africa F	E-Asia M	E-Europe M	S-Asia M	W-Africa M
optiexacta-000	2024_07_23	0.60	9.68	0.4336(74)	0.0020(42)	0.2727(74)	0.1038(74)	0.2200(72)	0.0007(64)	0.0749(74)	0.0592(74)
visionbox-001	2024_03_19	0.68	12.70	0.1992(71)	0.0020(42)	0.0828(60)	0.0787(72)	0.0366(71)	0.0003(1)	0.0157(61)	0.0088(63)
s1-005	2023_07_03	0.65	8.84	0.1471(70)	0.0020(42)	0.1340(72)	0.0779(71)	0.0187(63)	0.0004(26)	0.0216(70)	0.0108(68)
clearviewai-001	2024_02_16	0.59	6.87	0.0860(39)	0.0019(4)	0.0648(36)	0.0564(68)	0.0144(44)	0.0004(7)	0.0143(52)	0.0132(71)
psl-002	2024_07_26	0.74	12.65	0.0861(40)	0.0020(42)	0.0221(20)	0.0372(39)	0.0098(31)	0.0004(26)	0.0043(12)	0.0018(12)
nec-3	2018_10_30	0.47	4.08	0.0093(11)	0.0020(42)	0.0192(18)	0.0268(21)	0.0066(25)	0.0005(42)	0.0105(26)	0.0102(67)
cognitec-007	2023_12_01	0.53	4.50	0.0034(7)	0.0020(42)	0.0077(7)	0.0184(19)	0.0019(6)	0.0005(42)	0.0116(32)	0.0071(57)
idemia-011	2024_03_05	0.33	2.75	0.0026(6)	0.0020(42)	0.0024(3)	0.0060(10)	0.0017(5)	0.0006(60)	0.0033(6)	0.0020(16)

FPIR in E. Asian women is about 200 times higher than in European women.

FPIR in W. African women is about 19 times higher than in European women.

FPIR no more than 3 times E. Euro women

THE FALSE MATCHES :: WHICH DEMOGRAPHIC GROUPS?



FPIR ~ 1 in 1300 on E. European Men

REGION OF BIRTH + SEX OF TOP NON-MATE

FPIR = 1 in 16 East Asian Women

THE FIGURE SHOWS HOW OFTEN A SEARCH OF A PERSON OF DEMOGRAPHIC GROUP X RETURNS A PERSON OF GROUP Y ABOVE THE THRESHOLD FOR THAT ALGORITHM

- THE VALUES ARE BASE-10 LOGS SO A VALUE OF -2.0 MEANS $10^{-2} = 0.01$, ONE IN ONE HUNDRED

KEY POINTS:

1. THE ON-DIAGONAL VALUES VARY BY 1.9 ($_{3.1 - 1.2}$) AND 1.5 ($_{3.1 - 1.6}$) ORDERS OF MAGNITUDE i.e., FACTORS OF 80 and 30.
2. THE OFF-DIAGONAL ELEMENTS ARE GENERALLY A FACTOR OF TEN OR MORE SMALLER.
   - SO GROUPS X and Y ARE PERHAPS 10-100 TIMES LESS LIKELY TO PRODUCE A FALSE POSITIVE THAN GROUPS X and X.

THE REMAINING SLIDES ARE BACKGROUND MATERIAL NOT REQUIRED FOR INTERPRETATION OF THE 1:N WEB PAGE

BACKGROUND:
WHAT'S THE CONNECTION BETWEEN 1:1 FALSE MATCH RATES and 1:N FALSE POSITIVE IDENTIFICATION OUTCOMES

1. MANY 1:N ALGORITHMS IMPLEMENT 1:N SEARCH
   A. BY COMPARING THE PROBE TEMPLATE WITH N ENROLLED ENTRIES
   B. THEN SORTING THE N SIMILARITIES TO RANK THE MOST SIMILAR

2. CAUTION: SOME SYSTEMS DON'T DO JUST THAT
3. CAUTION: SOME SYSTEMS DON'T DO THAT AT ALL

4. FOR THOSE THAT DO, WHAT'S THE RELATIONSHIP BETWEEN FPIR AND FMR?
5. CAN WE RUN 1:1 TESTS AND PREDICT 1:N OUTCOMES?

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

$$FPIR = 1 - (1 - FMR)^N$$

GALLERY SIZE, N

FALSE MATCH RATE FOR A 1:1 COMPARISON, FMR

ESTIMATED PROBABILITY THAT A SEARCH YIELDS *ANY* FALSE POSITIVES, FPIR

NONE OF THE N COMPARISONS MUST GIVE A FALSE POSITIVE

i.e. ALL N COMPARISONS MUST NOT MATCH

APPROXIMATION WHEN $N \cdot FMR \ll 1$

$$FPIR = N\,FMR$$

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

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

TERMINOLOGY STANDARDIZED IN ISO/IEC 19795-1:2021

FALSE POSITIVE :: CASINO EXAMPLE

- N = 500 PEOPLE (PEOPLE WHO CHEATED IN CASINOS)
- ALGORITHM CONFIGURED FOR FMR = 1 IN 1 MILLION

FALSE MATCH RATE EXPECTED FROM A 1:1 COMPARISON, FMR

FPIR = N FMR

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

MODELED PROBABILITY THAT A SEARCH YIELDS ANY FALSE POSITIVES

BUT SUPPOSE ALSO P = 10000 PEOPLE VISIT THE CASINO EACH DAY

EXPECTED NUMBER OF FALSE POSITIVES PER DAY =
P . FPIR = $10000 \times 5\ 10^{-4} = 5$

TO DETERMINE IF A CANDIDATE MATCH IS AN ACTUAL TRUE POSITIVE (A CHEAT) OR A FALSE POSITIVE, THE CASINO STAFF WOULD INVESTIGATE: POSSIBLY BY TAKING FURTHER PHOTOS, PERFORMING HUMAN FACE COMPARISON, LOOKING AT CARS, COMPANIONS, CONTEXT etc.

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
- EXCEPTIONS FOR SOME ALGORITHMS FROM ASIA
- EXCEPTIONS FOR SOME ONE-TO-MANY ALGORITHMS

WITH SUBSTANTIAL VARIATIONS ACROSS ALGORITHMS PROMPTING THE NISTIR 8280 RECOMMENDATION TO "KNOW YOUR ALGORITHM"

EXAMPLES ON NEXT TWO SLIDES

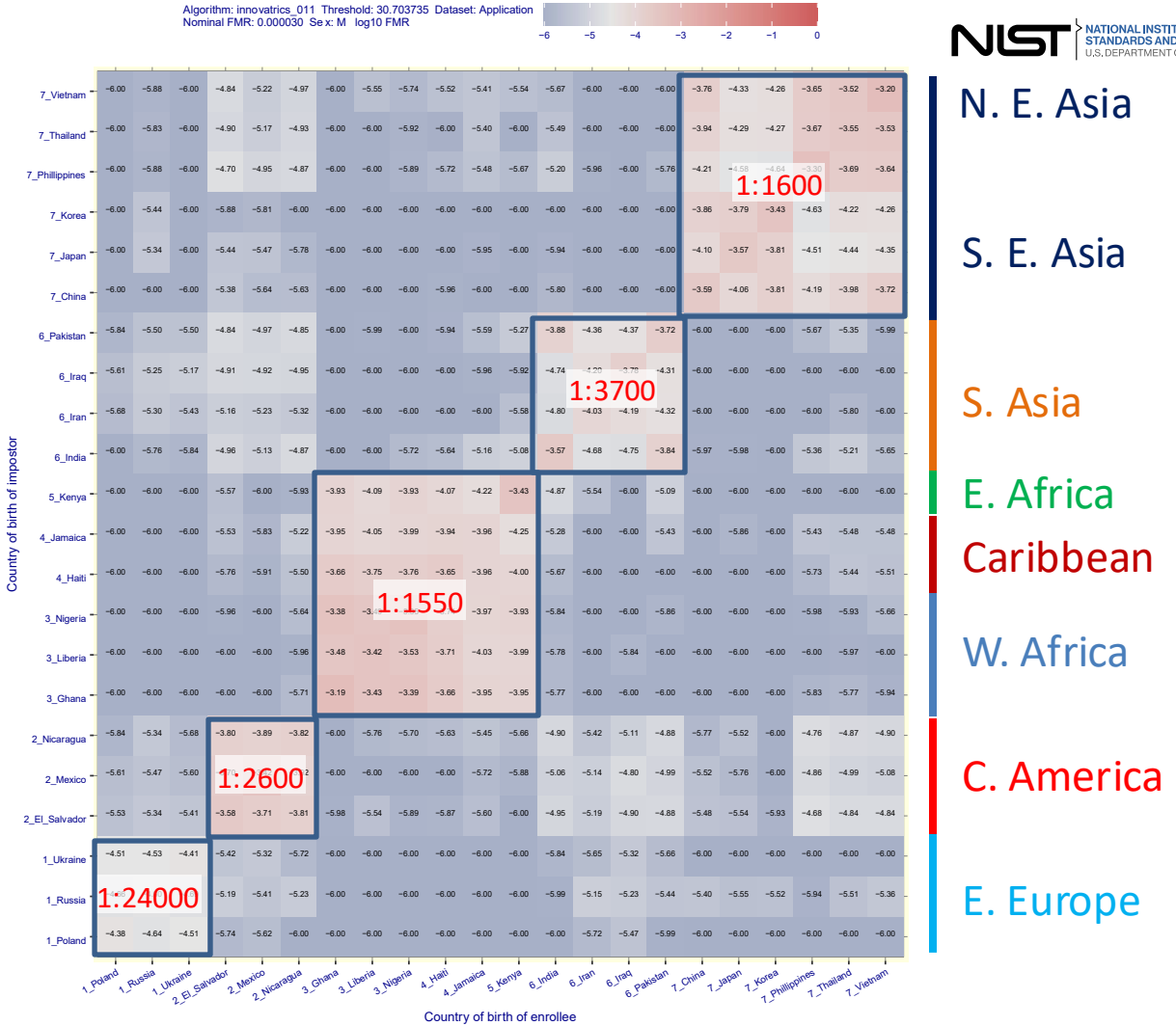
NIST ONGOING 1:1 FRTE (SINCE 2019) GIVES DEMOGRAPHIC DEPENDENCE OF FMR FOR MANY MORE RECENTLY SUBMITTED ALGORITHMS

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

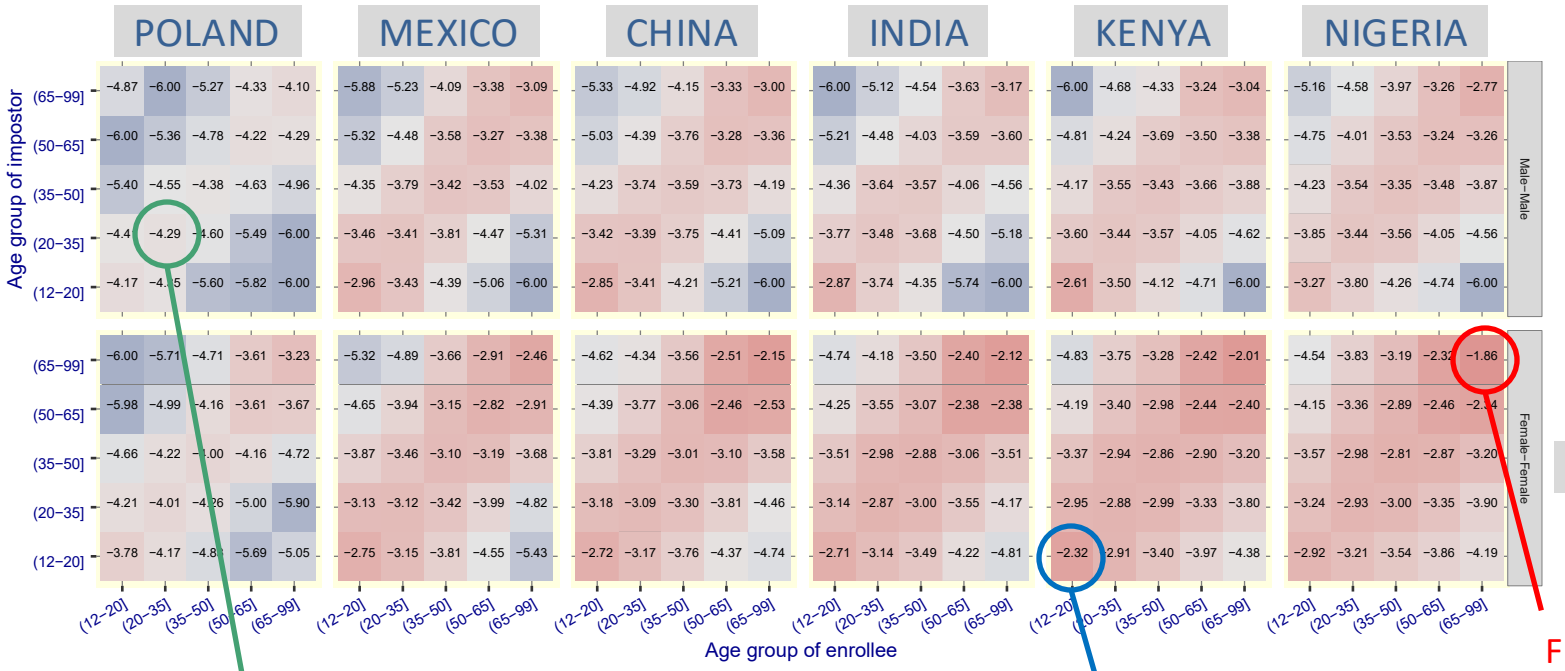
Caribbean

W. Africa

C. America

E. Europe

NOW ALSO CONSIDER WOMEN AND AGE



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

SO HOW TO HANDLE VARIABLE FALSE MATCH RATES IN PREDICTION?

WITH HOMOGENOUS FALSE MATCH RATES:

$$\mathrm{FPIR}(\tau) = 1 - (1 - \mathrm{FMR}(\tau))^N \approx N\,\mathrm{FMR}(\tau)$$

BUT WITH HETEROGENOUS FALSE MATCH RATES, THE FPIR FOR GROUP i IS:

$$\mathrm{FPIR}_i(\tau) = 1 - \prod_j (1 - \mathrm{FMR}_{ij}(\tau))^{n_j} \approx \sum_j \mathrm{FMR}_{ij}(\tau)\, n_j$$

WHICH HAS A CONCISE MATRIX FORM:

$$\mathbf{FPIR}(\tau) = \mathbf{FMR}(\tau)\ \mathbf{n}$$

TO PREDICT NUMBERS OF FALSE POSITIVES, INCLUDE WHO GETS SEARCHED

$$\mathrm{NFP}(\tau) = \mathbf{p}^T\ \mathbf{FMR}(\tau)\ \mathbf{n}$$

NEXT SLIDE EXPLAINS THIS EQUATION

TERMS:
FPIR = False Positive Identification Rate
FMR = False Match Rate
τ = Threshold
i = index of probe demographic group
j = index of gallery demographic group
n_j = number of gallery entries for group j
p_i = number of probes for group i
N = total number of gallery entries
FMR_{ij} = rate at which groups i and j false match

CAUTION: AGAIN THIS IS BINOMIAL THEORY WHICH ASSUMES INDEPENDENT COMPARISONS: WE KNOW SOME ALGORITHMS DON'T DO THAT, SO THIS MODEL IS WRONG FOR THEM

PREDICTING 1:N FALSE POSITIVES FROM 1:1 RESULTS

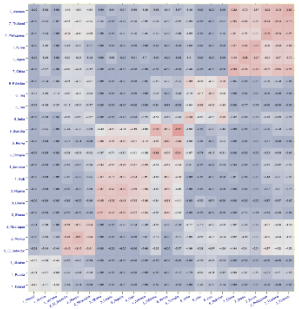
NIST NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY U.S. DEPARTMENT OF COMMERCE

SOCIO-TECHNICAL CONTEXT

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

$$NFP = p^T \; FMR \; g$$

DEMOGRAPHIC COMPOSITION OF PROBE SET



DEMOGRAPHIC COMPOSITION OF GALLERY

WHO FALSE MATCHES WHO: HOW FREQUENTLY

CROSS-DEMOGRAPHIC FALSE MATCH RATE MATRIX

FALSE POSITIVE IDENTIFICATION RATE (FPIR) WHEN GALLERY HAS MIXED DEMOGRAPHICS: A TOY EXAMPLE

FALSE POSITIVE IDENTIFICATION RATE (FPIR) IS THE PROPORTION OF NON-MATED SEARCHES YIELDING ONE OR MORE CANDIDATES ABOVE THRESHOLD

WHO FALSE MATCHES WHO: HOW FREQUENTLY

WHO GETS PUT IN GALLERY, E.G., 20000 WOMEN and 80000 MEN

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

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

CROSS-DEMOGRAPHIC FALSE MATCH RATES (WHICH ARE A PROPERTY OF THE ALGORITHM)

DEMOGRAPHIC COMPOSITION OF GALLERY

NOTE 1: Even though women are a minority in the gallery, they have highest false positive outcomes

NOTE 2: The error inherent in the N.FMR approximation to the binomial is relatively small FPIR = 0.18 vs. 0.20.

FP RATE (FPIR)

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

FPIR = FMR g

0.2	=	1×10^{-5}	0	20000
0.008		0	1×10^{-7}	80000

FP NUMBER (NFP)

NUMBER OF FALSE POSITIVES | WHO GETS SEARCHED | x WHO FALSE MATCHES WHO: HOW FREQUENTLY x | WHO GETS PUT IN GALLERY

NFP = p^T FMR g

200+8	=	1000	1000	1×10^{-5}	0	20000
				0	1×10^{-7}	80000

DEMOGRAPHIC COMPOSITION OF PROBE SET

CROSS-DEMOGRAPHIC FALSE MATCH RATES

DEMOGRAPHIC COMPOSITION OF GALLERY

BUT GIVEN EMPIRICAL ESTIMATES OF FPIR CAN WE RECOVER FMR ?

TERMS:

τ = Threshold
i = index of probe demographic group
j = index of gallery demographic group
n_j = number of gallery entries (group j)
N = total number of gallery entries
FMR_{ij} = rate at which groups i and j false match
FPIRij = rate at which group I search produced group j non-mate

CAN WE INVERT THE PREDICTION FORMULA

$$\mathrm{FPIR}_i(\tau) = 1 - \prod_j (1 - \mathrm{FMR}_{ij}(\tau))^{n_j}$$

REWRITE FPIR IN TERMS OF FPIR, USING APPROXIMATION

$$\mathrm{FPIR}_i(\tau) = 1 - \prod_j (1 - \mathrm{FPIR}_{ij}(\tau))$$

RE-ARRANGE AND INCLUDE FMR AGAIN

$$\mathrm{FPIR}_{ij}(\tau) = (1 - \mathrm{FMR}_{ij}(\tau))^{n_j}$$

APPROXIMATE AGAIN AND ASSUME FMR_{ij} is ZERO WHEN i ≠ j
INVERT TO PRODUCE *IMPLIED* FMR

$$\mathrm{IFMR}_i(\tau) = n_i^{-1}\,\mathrm{FPIR}_i$$

THE TERM *IMPLIED* IS NECESSARY BECAUSE SOME ALGORITHMS DO NOT IMPLEMENT 1:N BY EXECUTING N 1:1 COMPARISONS - SEE LATER SLIDE

THE 1:N FPIR VALUES BY DEMOGRAPHICS GROUP CAN BE USED TO PRODUCE AN "IMPLIED FMR"

$$\mathrm{IFMR}_i(\tau) = n_i^{-1}\,\mathrm{FPIR}_i$$

THIS IS THE FALSE MATCH RATE IMPLIED BY AN EMPIRICAL MEASUREMENT OF FALSE POSITIVE IDENTIFICATION RATE FROM A SET OF 1:N SEARCHES.

- **FPIR_i** = FRACTION OF GROUP i SEARCHES THAT YIELD ANY NON-MATED GALLERY ENTRIES ABOVE THRESHOLD τ
- **IFMRi** = IMPLIED FALSE MATCH RATE FOR COMPARING PHOTOS OF TWO GROUPiI PEOPLE.
- **n_i** IS THE NUMBER OF GROUP i PEOPLE IN THE GALLERY.

⚠ THIS EQUATION IS A MODEL THAT IS LIKELY TO BE WRONG FOR THOSE ALGORITHMS THAT DO OBEY BINOMAL ASSUMPTIONS – SEE SLIDE 36.

DEMOGRAPHICS: THE FALSE POSITIVE TAB 2/2

1. Searches of non-mated probe images of persons in group X gives an empirical measurement $FPIR_X$.
2. If there are N_X gallery entries for that group then the false match rate implied by that measurement is: **$ImpliedFMR_X = FPIR_X / N_X$**
3. This formula is an approximation because it ignores that some of the false positives may be against gallery elements of demographic groups Y, Z etc.
4. The incidence of cross-group false positives is empirically much lower than within-group false positives. This is true because the experimental design only includes persons from geographically separate regions.
5. The tabulated values are -log10($ImpliedFMR_X$) so that a value of 5 means ImpliedFMR = 0.00001.
6. High values in the table are inferior.
7. The threshold is set for PFIR = 0.002 in East. European women. Implied FMR for this group is 6.3 10^{-8} so the tabulated value is 7.2.
8. Tabulated values higher than 7.2 indicate higher FMR. A value of 5.2 indicates FMR is 100 times higher.

FPIR | Implied FMR

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
megvii-004	2023_10_18	0.52	4.07	7.6(4)	7.2(42)	7.4(2)	6.6(4)	7.5(6)	7.0(73)	7.4(3)	6.7(18)
canon-003	2023_09_05	0.7	9.76	7.6(5)	7.2(42)	7.3(5)	6.1(12)	7.4(10)	7.2(71)	7.3(5)	6.5(20)
veridas-004	2023_02_03	0.69	10.17	5.9(55)	7.2(42)	5.9(59)	5.4(53)	6.5(61)	7.4(60)	6.6(65)	6.3(49)
idemia-011	2024_03_05	0.66	8.52	7.6(6)	7.2(42)	7.4(3)	6.2(10)	7.6(5)	7.4(60)	7.3(6)	6.8(16)
hyperverge-002	2022_04_13	0.78	15.55	6.1(33)	7.2(42)	6.6(13)	5.4(48)	6.7(39)	7.5(42)	7.0(17)	6.3(50)
corsight-000	2023_07_13	0.72	11.8	6.0(38)	7.2(42)	6.0(42)	5.4(49)	6.7(36)	7.6(26)	6.7(34)	6.2(56)
clearviewai-002	2024_06_18	0.72	14.22	6.0(41)	7.2(42)	6.0(38)	5.3(69)	6.6(47)	7.7(3)	6.6(56)	6.0(70)

Implied FMR in W. African women is about 4 times higher than in European women.

Implied FMR in W. African women is about 60 times higher than in European women, and 125 times higher than in European men.

SO WHAT?
CONSEQUENCES OF UNKNOWN FALSE POSITIVE RATE VARIATIONS

1. THRESHOLDS ARE OFTEN SET BASED ON A DEVELOPER OR SUPPLIER RECOMMENDATION.
    - THE INTENT OF THE THRESHOLD CALIBRATION PROCEDURE IS TO TABULATE THE EXPECTED FPIR FOR EACH THRESHOLD T AND DATABASE SIZE N.

1. THE DEVELOPER WILL OFTEN CALIBRATE FMR BASED ON INTERNAL TRIALS THAT COULD, IN PRINCIPLE, HAVE DIFFERENT DEMOGRAPHIC REPRESENTATION RELATIVE TO SOME FUTURE OPERATIONAL USES.

2. HOWEVER, IF
    A: FPIR IN A NEW DEMOGRAPHIC GROUP Y IS HIGHER, AND
    B: PEOPLE OF GROUP Y ARE ENROLLED AND SEARCHED
    THEN FALSE POSITIVES WILL BE MORE COMMON THAN EXPECTED

3. MITIGATION:
    1. CONSIDER EXTERNAL (PRE-DEPLOYMENT) TESTS
    2. RUN TESTS OF THE OPERATIONAL SYSTEM INCLUDING IN AN OFFLINE BULK-SEARCH MODE TO GENERATE SUFFICIENT TRANSACTIONS
    3. SET THRESHOLD BASED ON A NEW CALIBRATION FOR WHICHEVER GROUP GIVES THE HIGHEST FPIR.

BUT ONE REACTION TO NIST IR 8280 WAS "SO WHAT?"

Q1: ALL THOSE FMR VALUES ARE SO TINY -- 1 IN 10000 OR 1 IN 100, SAY. SO WHY CARE?

A1: BECAUSE IF THE 1:1 ALGORITHM IS USED TO IMPLEMENT 1:N SEARCH THEN THESE FMR VALUES ARE EXPECTED TO SCALE UP WITH THE GALLERY SIZE N.

- **THAT IS A 1:N ALGORITHM MUST CORRECTLY REJECT ALL N NON-MATES**

THESE FMR VARIATIONS ARE OBSERVED WITH HIGH QUALITY IMAGES

IN THE PRECEDING FIGURES THE IMAGES ARE ALL WELL CONTROLLED FRONTAL PORTRAITS.

CANMETIN ET AL. HAVE FAULTED NIST'S TESTS FOR USING "*BENCHMARK IMAGES [THAT] ARE OVERLY IDEAL COMPARED TO REAL WORLD CONDITIONS*" WITHOUT THE "*COMPLEXITY OF IMAGES CAPTURED IN REAL-WORLD DEPLOYMENTS*". WHILE THAT IS INEVITABLY TRUE – NIST DOES NOT HOLD IMAGES EXHIBITING ALL POSSIBLE QUALITY DEGRADATIONS – THE CRITICISM ASSUMES THAT POOR QUALITY IS NEEDED TO ELICIT THE HIGH ERROR RATES - IT IS NOT - THE CLAIM DEPENDS ON THE PARTICULAR KIND OF ERROR:

- **FALSE POSITIVES:** THIS SLIDE DECK SHOWS ADVSERSE FALSE MATCHES OCCUR EVEN WITH GOOD QUALITY - FALSE POSITIVES VARIATIONS OCCUR BECAUSE ALGORITHMS ARE a) NOT TRAINED ON DIVERSE DATA AND b) NOT TRAINED TO EQUALIZE FALSE POSITIVE RATES ACROSS DEMOGRAPHICS.
  - POOR QUALITY PHOTOS MAY EXACERBATE THE EFFECT, OR MAY REDUCE THE EFFECT, DEPENDING ON THE ALGORITHM - THIS REMAINS TO BE FULLY CHARACTERIZED

- **FALSE NEGATIVES:** POOR QUALITY PHOTOS DO ELEVATE FALSE NEGATIVE RATES, AND THIS CAN BE COUPLED TO DEMOGRAPHICS, FOR EXAMPLE:
  - PHOTOGRAPHY OF SKIN THAT REFLECTS LESS LIGHT CAN GIVE UNDEREXPOSURE ⟶ POTENTIALLY INCREASING FNIR
  - PHOTOGRAPHY OF TALL PEOPLE CAN GIVE NON-ZERO HEAD PITCH ANGLE ⟶ POTENTIALLY INCREASING FNIR

  BUT, DEPENDING ON THE APPLICATION, HIGH FALSE NEGATIVE RATES CAN BE ADVANTAGEOUS OR NOT
  - IN COOPERATIVE ACCESS CONTROL, HIGHER FALSE NEGATIVE RATES IMPLY INCONVENIENCE FOR THE USER.
  - IN CRIMINAL LAW ENFORCMENT, HIGHER FALSE NEGATIVE RATES ARE ADVANTAGEOUS TO THE CRIMINAL, DISADVANTAGEOUS TO THE INVESTIGATOR

* NIST'S DATASETS ARE VARIOUS QUALITIES

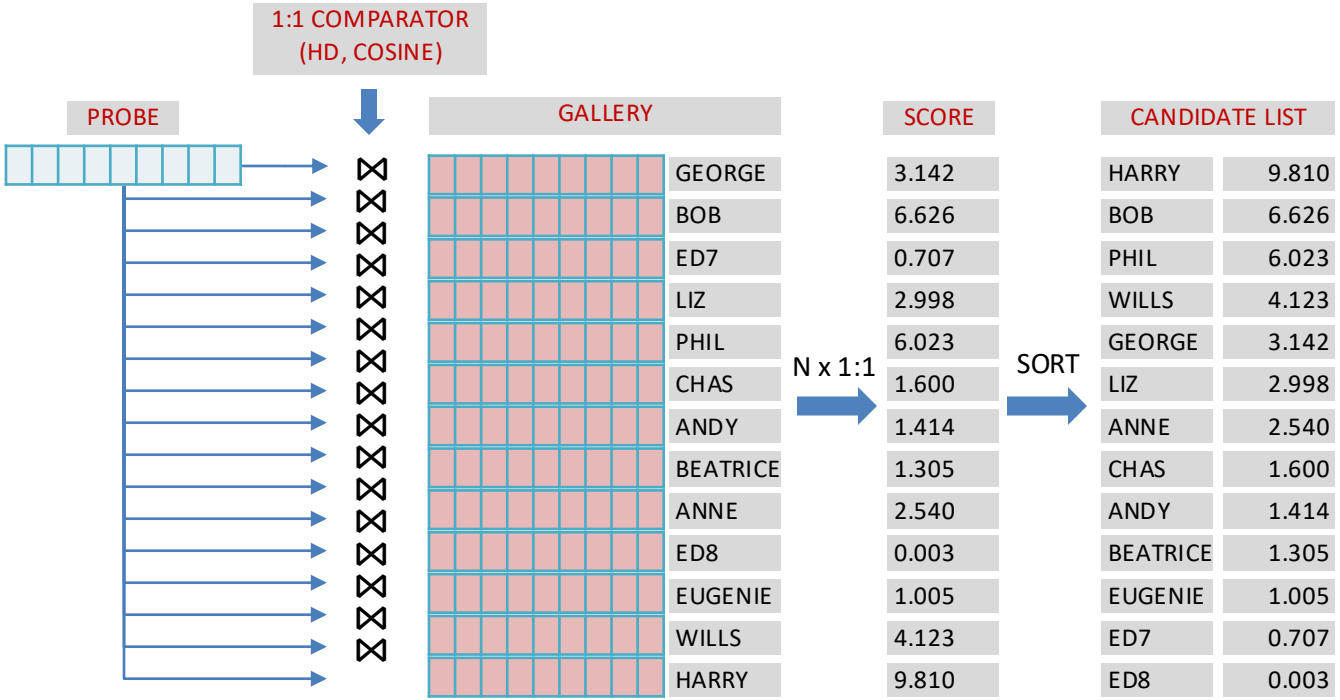
SOME ALGORITHMS IMPLEMENT 1:N AS N 1:1 COMPARISONS.

+ OTHERS DO NOT…
+ OTHERS DO MORE …

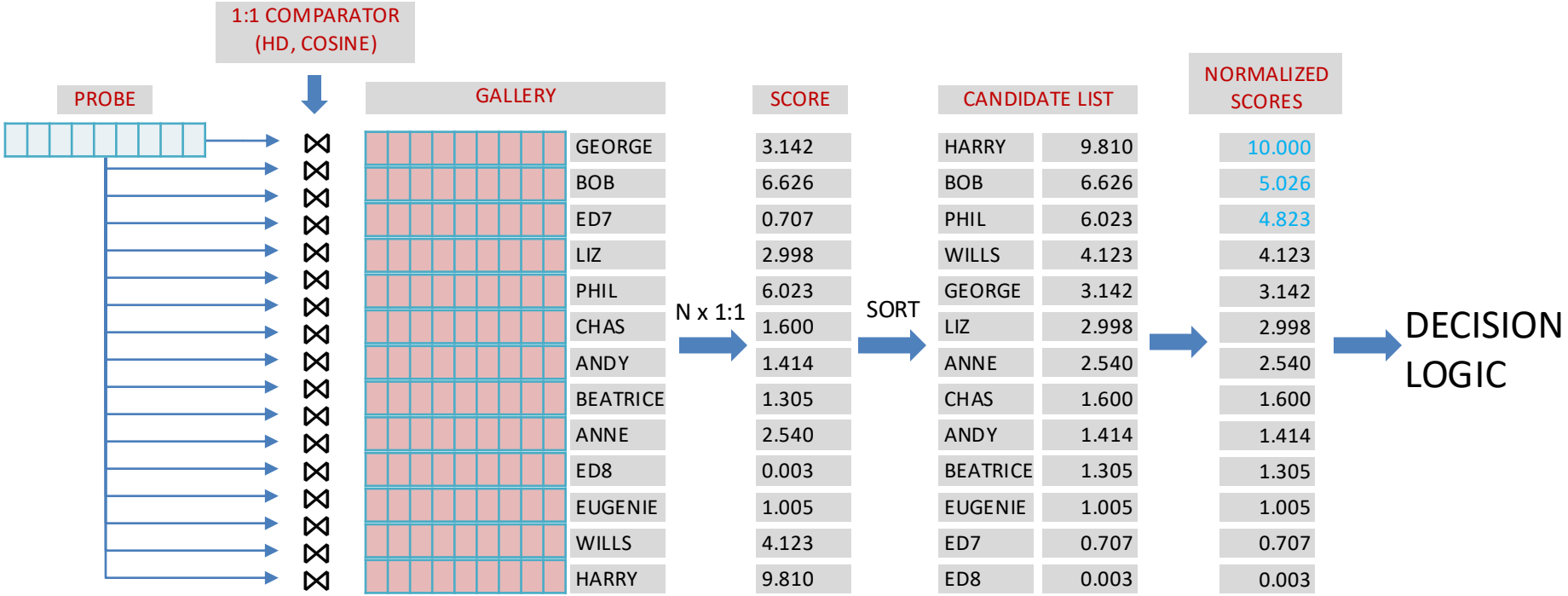
EXAMPLES FOLLOW…

THE BINOMIAL MODEL OF THE LAST FEW SLIDES WILL NOT WORK CORRECTLY FOR SUCH SEARCH ALGORITHMS

DEFAULT IMPLEMENTATION OF 1:N:
EXHAUSTIVE SEARCH = N 1:1 COMPARISONS + SORT



BUT: SOME DEVELOPERS NORMALIZE SCORES INTRODUCES DEPENDENCE ON GALLERY

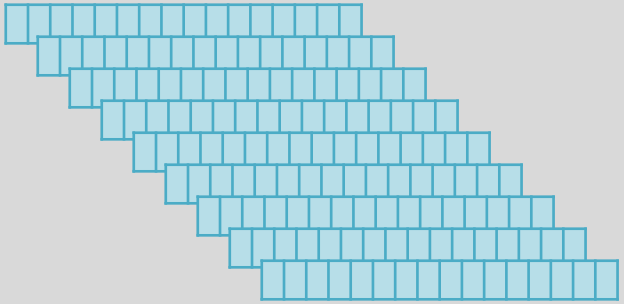


- Jens Peter Hube *Using Biometric Verification To Estimate Identification Performance* Identix Corporate Research, Biometrics Symposium 2006 https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4341620
- Ross J. Micheals, Walter Scheirer, Terrance Boult, Anderson Rocha Robust Fusion: Extreme Value Theory for Recognition Score Normalization, ECCV 2010
- Ethan M. Rudd , Lalit P. Jain, Walter J. Scheirer, and Terrance E. Boult The Extreme Value Machine, IEEE PAMI, March 2018

OTHER 1:N ALGORITHMS DO NOT COMPUTE N COMPARISONS

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

SET OF N TEMPLATES, PRODUCED INDEPENDENTLY



...

FINALIZE()

CONVERT LINEAR SET OF TEMPLATES INTO A SPECIALIZED GRAPH STRUCTURE.

THIS OPERATION MAY BE EXPENSIVE, BUT MAY AFFORD GAINS SUCH AS LOW SEARCH DURATION

FAST SEARCH DATA STRUCTURE

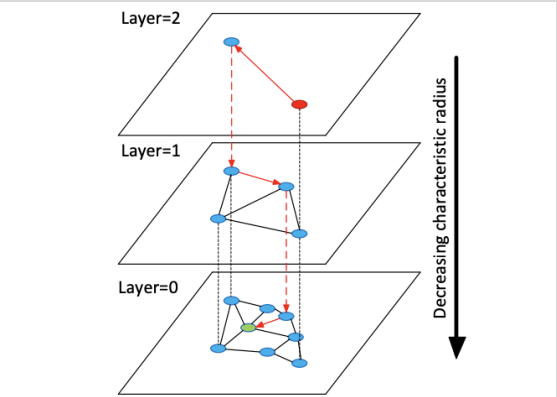


Fig. 1. Illustration of the Hierarchical NSW idea. The search starts from an element from the top layer (shown red). Red arrows show direction of the greedy algorithm from the entry point to the query (shown green).

Yu A. Malkov, D. A. Yashunin, **Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs.** IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 42 No. 4 April 2020 pp. 824–836 https://doi.org/10.1109/TPAMI.2018.2889473

1:N FAST DATA STRUCTURE SUMMARY

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

SEARCHABLE DATA STRUCTURE

- **LINEAR:** Many developers implement 1:N search as N 1:1 comparisons aka exhaustive search
- **CONSTRUCTED:** Others build indexes, graphs, trees, or a dictionary, or other exotic data structure
- Some developers field both types of algorithms

- **PROS**
  - **Speed**
  - **Storage (memory, cloud)**
  - False positive rates ?
  - False positive rates grow as N^0 i.e. flat
  - Demographic dependencies

- **CONS**
  - Cost of
    - Constructor()
    - Insert()
    - Delete()
    - Deleting somebody from a database may not be a simple operation
  - Score interpretation is complicated
    - score is (often) not f(x,y)
    - instead f(x, GALLERY)

- **TESTS**
  - Should not assume 1:N is implemented naively
  - Evaluate separately!

TWO MODES OF OPERATION: INVESTIGATION VS. IDENTIFICATION

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

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.

INVESTIGATIONAL USE OF FACE RECOGNITION AND DEMOGRAPHICS

T = 0 INVESTIGATIONS

- HISTORICALLY LAW ENFORCEMENT SEARCHES USED A SYSTEM CONFIGURED TO RETURN A FIXED NUMBER OF CANDIDATES
  - NO THRESHOLD APPLIED OR, EQUIVALENTLY, THE THRESHOLD IS 0
- IF THE PROBE-PHOTO **SUBJECT IS IN THE GALLERY**, THE SYSTEM RETURNS THE FIXED NUMBER OF MOST SIMILAR CANDIDATES.
  - THE CANDIDATES WILL OFTEN INCLUDE THE CORRECT MATED ENTRY OR ENTRIES
  - THE FALSE NEGATIVE IDENTIFICATION RATE (FNIR) aka “MISS RATE” IS THE FRACTION OF MATED SEARCHES THAT DO NOT INCLUDE THE CORRECT MATE AMONG THE CANDIDATES
  - EXCEPT WITH POOR QUALITY PHOTOS OR WHERE AGEING HAS OCCURRED, FNIR WILL BE CLOSE TO ZERO.
- BUT IF THE PROBE PHOTO **SUBJECT IS NOT IN THE GALLERY**, THE SYSTEM AGAIN RETURNS THE SAME FIXED NUMBER OF MOST SIMILAR CANDIDATES
  - ALL OF THESE ARE FALSE POSITIVES, SO THE FALSE POSITIVE IDENTIFICATION RATE (FPIR) IS 1.
  - A HUMAN INVESTIGATION – WHICH CAN INCLUDE HUMAN COMPARISON OF PHOTOGRAPHS AND CONSIDERATION OF NON-BIOMETRIC INFORMATION – IS NEEDED TO EXONERATE SUCH CANDIDATES

T > 0

- MANY APPLICATIONS OF FACE RECOGNITION IMPOSE A THRESHOLD, T: ONLY THOSE CANDIDATES ABOVE THRESHOLD ARE RETURNED.
- A T > 0 POLICY DEMANDS RETURN OF ONLY SUFFICIENTLY SIMILAR FACES
- WHEN T IS HIGH, AND THE PROBE PHOTO **HAS A MATE IN THE GALLERY**, THE SYSTEM WILL USUALLY RETURN THE CORRECT CANDIDATE, SO FNIR WILL BE SMALL. THIS WILL NOT BE TRUE:
  - FOR POOR QUALITY PHOTOS, SUCH AS THOSE WITH BLUR OR POOR EXPOSURE.
  - WHEN THE GALLERY PHOTO WAS COLLECTED MANY YEARS BEFORE THE PROBE AND THE FACE HAS CHANGED APPEARANCE.
  - LIKEWISE IF THERE IS ANY SUBSTANTIAL CHANGE IN APPEARANCE, A FALSE NEGATIVE CAN OCCUR.
- WHEN T IS HIGH, AND A PROBE PHOTO **HAS NO MATE IN THE GALLERY**, THE SYSTEM WILL USUALLY RETURN NO CANDIDATES, SO FPIR IS SMALL
  - FALSE POSITIVES CAN OCCUR WHEN THE PROBE SUBJECT HAS AN IDENTICAL TWIN IN THE GALLERY AND THE TWIN IS RETURNED.
  - AN ALGORITHM WITH LARGE DEMOGRAPHIC DIFFERENTIALS COULD MORE OFTEN YIELD SIMILARITY SCORES ABOVE THRESHOLD FOR SOME GROUPS.
  - **FALSE POSITIVES CAN ALWAYS BE SURPRESSED BY ELEVATING T**

THINKING THROUGH CONSEQUENCES: THREE EXAMPLE APPLICATIONS

1. DISPENSING DRUGS

- NON-REPUDIATION
- 1:1
- VOLUME: 100S PER DAY
- TRANSACTIONS ARE ALMOST ALWAYS MATED
  - PROB(IMPOSTOR) IS LOW
- FALSE NEGATIVE ⟶ INCONVENIENCE
- FALSE POSITIVE ⟶ PRESCRIPTION DRUG FRAUD
- WHO IS HARMED BY DEMOGRAPHIC DIFFERENTIAL IN FP?
  - SOME PHARMACISTS

2. PAPERLESS BOARDING

- FACILITATION OF RECORDING IMMIGRATION EXIT *AND* ACCESS CONTROL TO AIRCRAFT
- 1:N
- VOLUME: 100S PER FLIGHT
- TRANSACTIONS ARE ALMOST ALWAYS MATED
  - PROB (IMPOSTOR) IS LOW
- FALSE NEGATIVE ⟶
  - PAPER BOARDING WITH AIRLINE STAFF
  - UN-RECORDED EXIT FOR VISA-HOLDER
- FALSE POSITIVE
  - ⟶ STOWAWAY
  - ⟶ MISMATCHES BETWEEN TRAVELERS WHO ARE PERMITTED TO BOARD
- WHO IS HARMED BY FP DIFFERENTIAL?
  - AIRLINE
  - IMMIGRATION DATABASE

3. WATCHLIST

- SOCCER STADIUM. COUNTER-TERRORISM. COMPULSIVE GAMBLERS
- 1:N
- VOLUME: 10S OF THOUSANDS PER DAY
- TRANSACTIONS ARE ALMOST ALWAYS NON-MATED
  - PROB (GENUINE) IS LOW
- FALSE NEGATIVE ⟶ UNDETECTED “BAD GUY”
- FALSE POSITIVE ⟶ INCORRECT ENFORCEMENT ACTION … CIVIL LIBERTIES
- WHO IS HARMED BY DEMOGRAPHIC DIFFERENTIALS IN FP?
  - BYSTANDERS

DEMOGRAPHICS SUMMARY

» LEADING CONTEMPORARY ALGORITHMS
- ARE VERY ACCURATE
- INCREASINGLY TOLERATE POOR IMAGE QUALITY
- GENERALLY DISTRIBUTE ERRORS UNEVENLY ACROSS DEMOGRAPHICS

» FALSE POSITIVE DIFFERENTIALS MUCH LARGER THAN FALSE NEGATIVE DIFFERENTIALS
- MORE FALSE POSITIVES IN ASIAN AND AFRICAN FACES
- MORE FALSE POSITIVES IN WOMEN
- MORE FALSE POSITIVES IN THE OLD AND VERY YOUNG
- WITH EXCEPTIONS TO THIS!

» ONE-TO-MANY ALGORITHMS DON'T NECESSARILY BEHAVE LIKE ONE-TO-ONE
- MANY DO
- BUT SOME ONE-TO-MANY STABILIZE FALSE ALARM RATES

❑ FALSE NEGATIVES FROM CHANGE OF APPEARANCE - OFTEN POOR PHOTOGRAPHY

❑ FALSE POSITIVES FROM ALGORITHMS APPLIED TO "UNKNOWN" DEMOGRAPHIC GROUPS
- ❑ EVEN WITH HIGH QUALITY IMAGES

» ALGORITHM MATTERS
- SOME MORE ACCURATE THAN OTHERS
- SOME DEMOGRAPHICALLY INSENSITIVE
- "KNOW-YOUR-ALGORITHM" (KYA)
- SET THRESHOLD TO LIMIT FPIR FOR THE WORST-CASE DEMOGRAPHIC

» APPLICATION MATTERS
- ERROR IMPACTS RANGE FROM INCONSEQUENTIAL TO GRAVE

» INCOMPLETE REPORTING IN COVERAGE
- CONFUSION OF FACE "ANALYSIS" WITH "RECOGNITION"
  - DETECTION IS NOT RECOGNITION
  - AGE ESTIMATION IS NOT RECOGNITION
- FAILURE TO IDENTIFY WHICH COMPONENT IS AT FAULT
  - DETECTOR? CAMERA? ALGORITHM?
- DIFFERENTIATE FALSE POSITIVES FROM FALSE NEGATIVES
  - MISSING REPORTS ON FALSE POSITIVES

» SINCE 2019
- SOME DEVELOPERS HAVE ADDRESSED DIFFERENTIALS.
- WE HAVE SUMMARY "FITNESS" INDICATORS
- ACADEMIC RESEARCH

» CONSULT SUMMARY "BIAS" MEASURES