DHS SCIENCE AND TECHNOLOGY

Demographic variation in the performance of biometric systems: insights gained from large-scale scenario testing

Homeland Security
Science and Technology

3/30/2021

Yevgeniy Sirotin
Maryland Test Facility
Principal Investigator

Arun Vemury
Biometric and Identity Technology Center
Director

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034
- This work was performed by a team of researchers at the Maryland Test Facility
- The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers
- The data used in this research was acquired under IRB protocol

Scenario Testing vs. Technology Testing

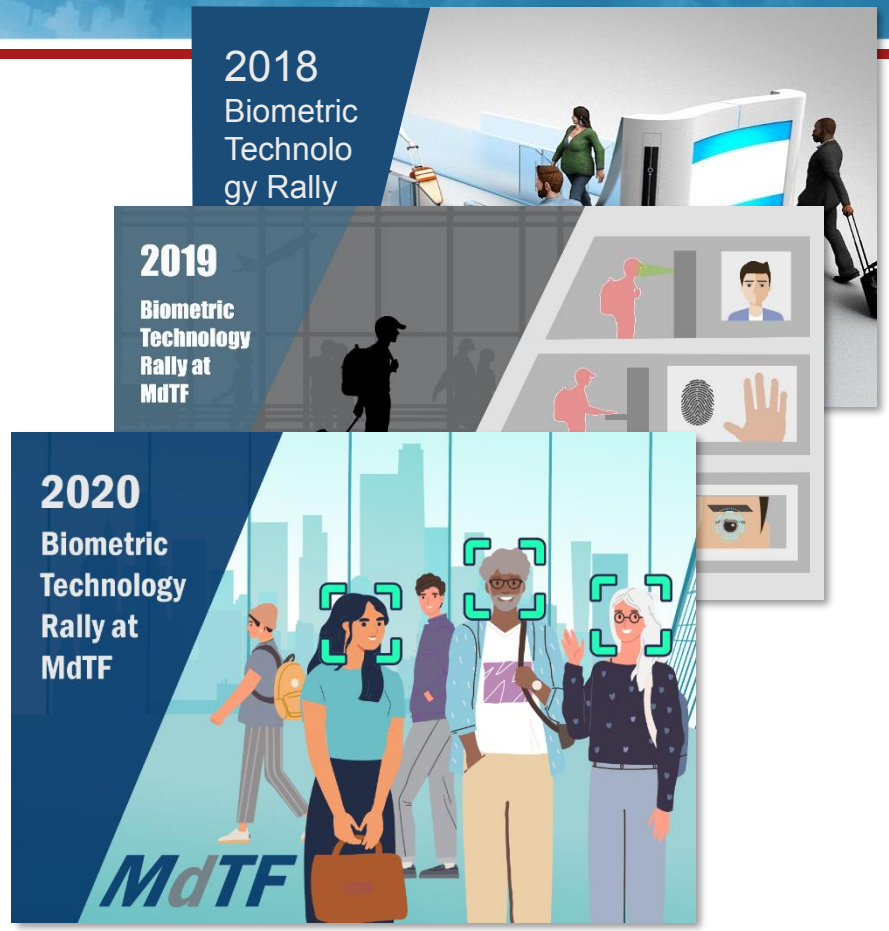
- **Scenario Testing:**
  - Centered around a use-case
  - Full multi-component biometric system
  - Gathering new biometric samples
  - Smaller sample size
  - Answers questions about how technology performs for an intended use
  - Answers questions about the suitability of a system for an intended use
    - e.g., How will face recognition perform in a high-throughput unattended scenario?

- **Technology Testing:**
  - Centered around a technology
  - Focused on a specific system component
  - Re-use of biometric datasets
  - Larger sample size
  - Answers questions about how technologies advance or perform relative to each other
  - Answers questions about the limits of a technology's performance
    - e.g., What is the minimum false match rate achievable by face recognition technology?

> Scenario test thinking can help frame questions of technology fairness during use <

DHS S&T Scenario Testing of Face Recognition Technology

- The DHS Biometric Technology Rally is a yearly biometric system evaluation focused on DHS technology use-cases
- Since 2018, we have tested 151 combinations of commercial face acquisition systems and matching algorithms in a high-throughput unattended use case
- The Rallies provide comprehensive metrics about the tested technologies:
  - Efficiency – transaction times
  - Effectiveness – image acquisition and matching success
  - Satisfaction – user feedback
  - https://mdtf.org



Homeland Security
Science and Technology

Face Recognition Technology Fairness

- Scientific analyses of data collected in the Rallies have addressed demographic effects in face recognition technologies:
  - The role of image acquisition in shaping demographic differences in a face recognition system
  - Establishing the influence of race, gender, and age on false match rate (FMR) estimates of a face recognition system
  - Quantification and comparison of race and gender differences in commercial face and iris recognition systems
  - Cognitive biases introduced by face recognition algorithm outcomes on human reviewers

- While some systems test well with diverse demographic groups, some demographic performance differentials persist in both acquisition and matching components of biometric systems and require careful evaluation

Homeland Security
Science and Technology

2020 Biometric Technology Rally
Unmasking Demographic Differentials

- Completed during COVID-19, this Rally tested face recognition systems under two conditions:

Without Masks

volunteers removed their masks prior to using the system

With Masks

volunteers kept their masks on while using the system

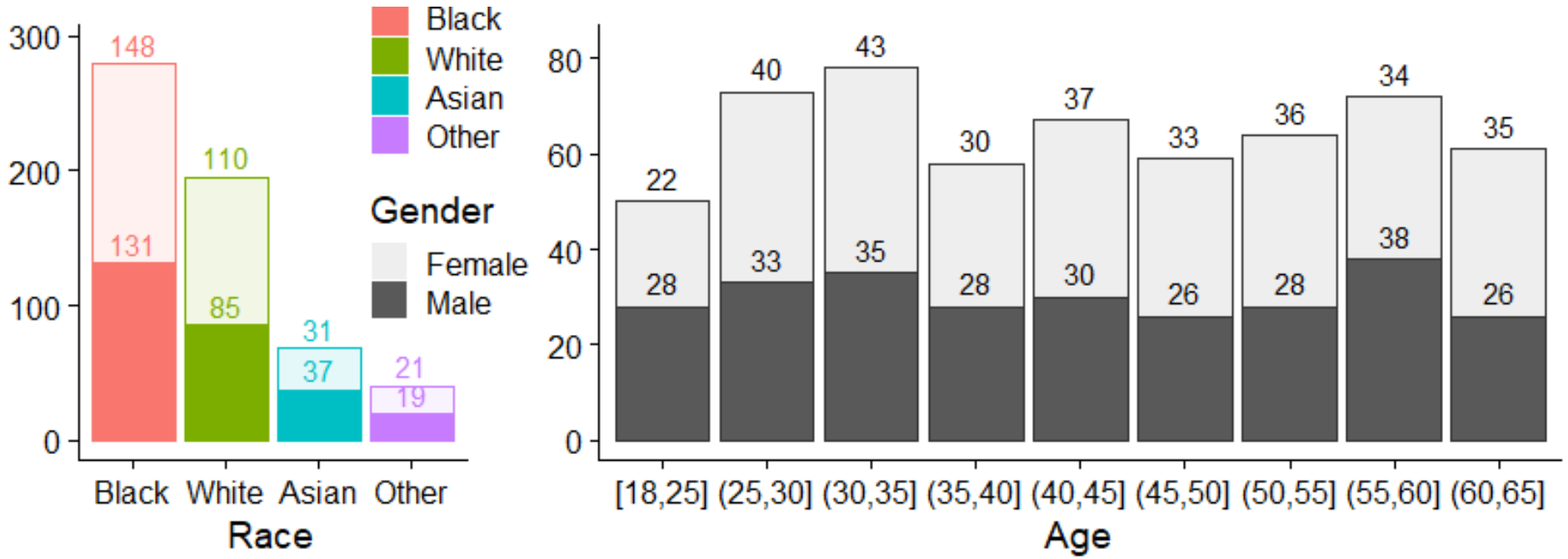
> Unattended high-throughput scenario similar to aircraft boarding <

Homeland Security
Science and Technology

Unattended High-throughput Scenario

- The face recognition system has limited time to operate
- The face recognition system acquires one image per individual
- The identification gallery is small (500 people)
- Most people being matched are in the identification gallery
- Impact of errors of those being matched is dominated by false negative identifications
  - **Example Impact:** Delay or denial of access to an aircraft

Diverse Test Volunteers – Commercial Face Recognition Systems



- Volunteers used their own personal face masks during testing

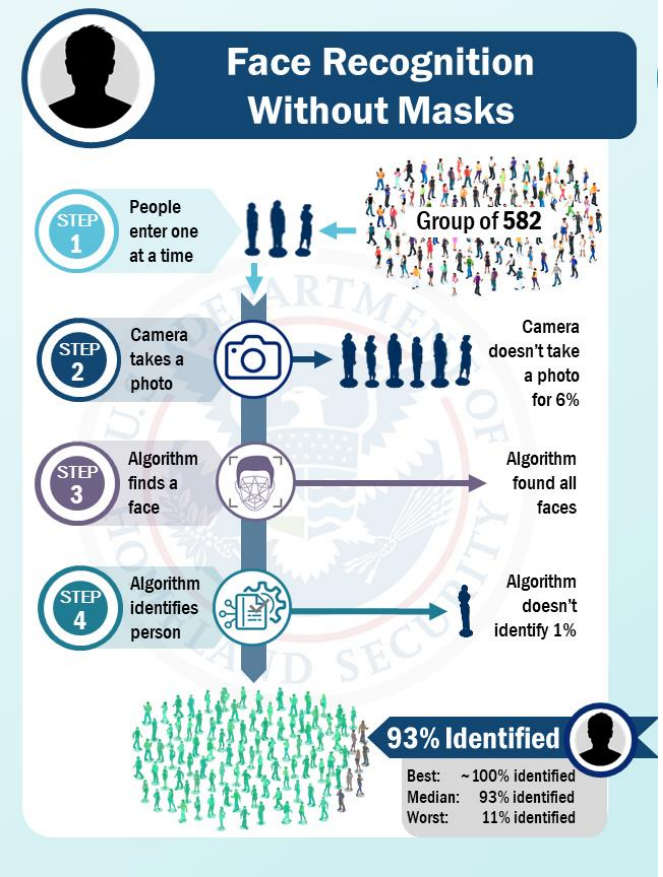
- 6 Commercial Acquisition Systems
- 10 Commercial Matching Systems

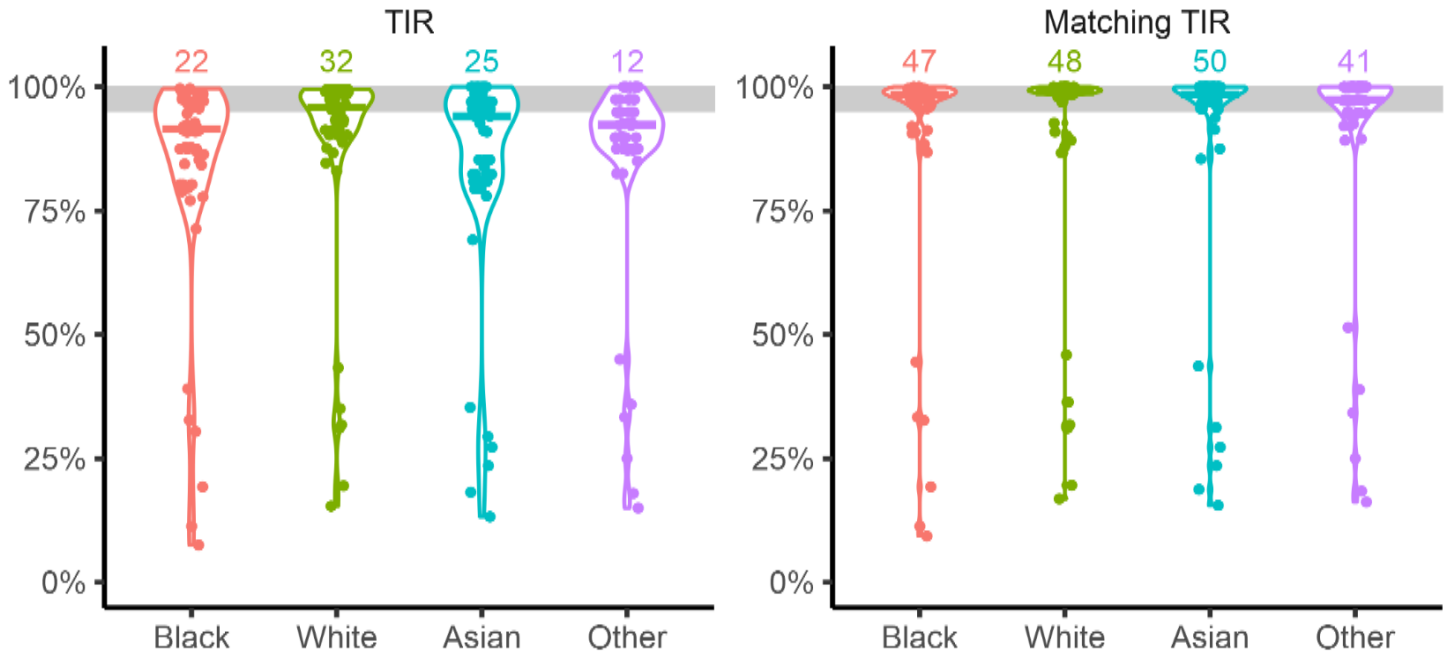
> 60 System Combinations <

- Systems had to:
  - Acquire face images from each volunteer
  - Identify each volunteer against a gallery
    - 1479 images of 500 individuals

Homeland Security
Science and Technology

Face Recognition Can Work Well Across Demographic Groups

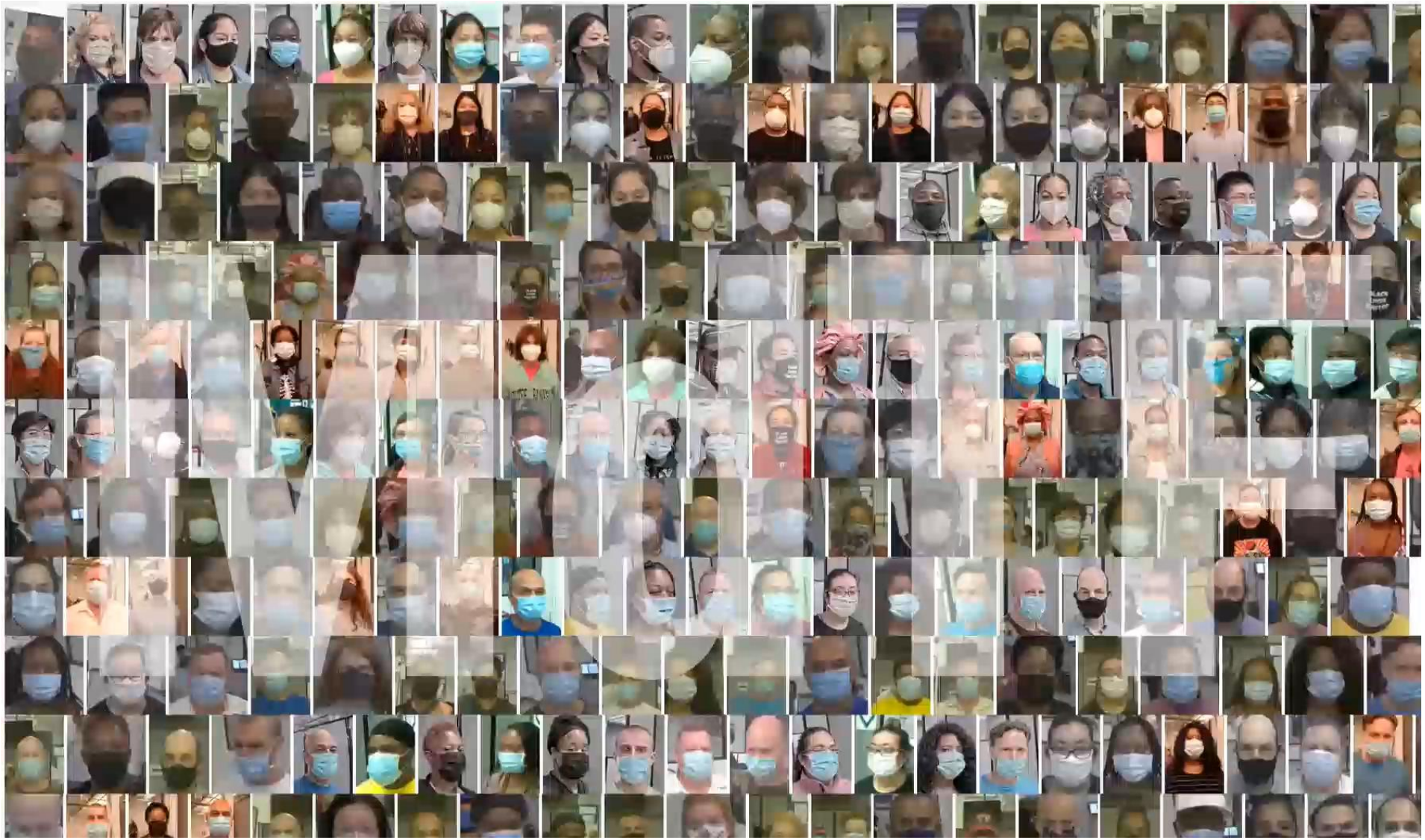




Each point in the graph represents the true identification rate (TIR) of a combination of an acquisition and matching system (n = 60) across our sample of 582 volunteers.

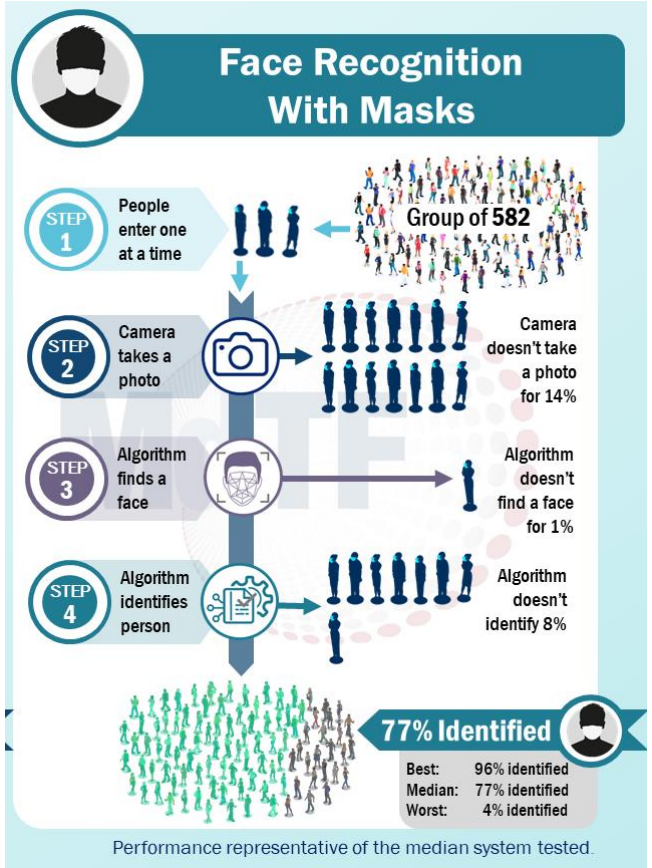
TIR includes failure of acquisition systems to submit images.
Matching TIR discounts any failure of acquisition systems to submit images.

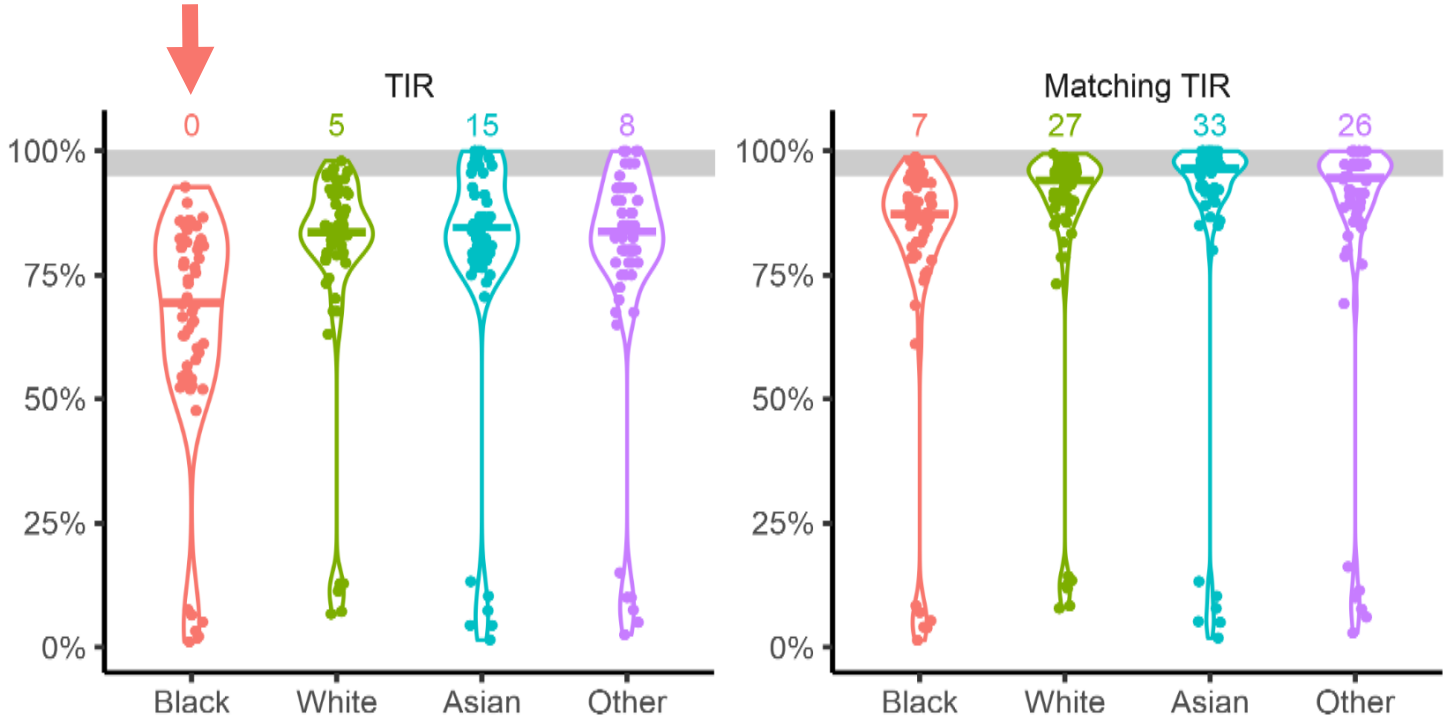
Homeland Security
Science and Technology



Homeland Security
Science and Technology

Un-equal Impact of Masks on Performance



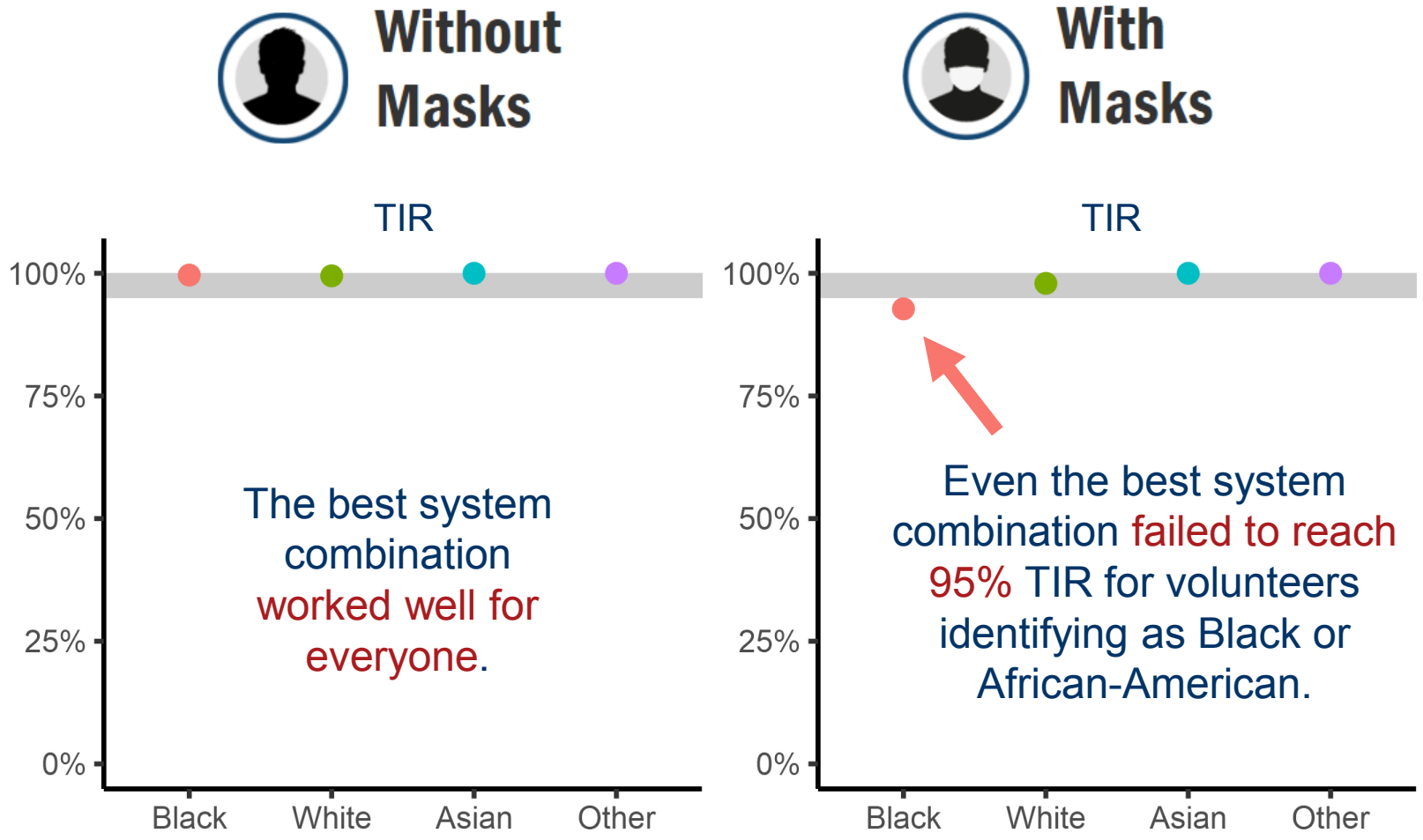


Each point in the graph represents the true identification rate (TIR) of a combination of an acquisition and matching system (n = 60) across our sample of 582 volunteers.

TIR includes failure of acquisition systems to submit images.
Matching TIR discounts any failure of acquisition systems to submit images.

Homeland Security
Science and Technology

Best-performing Acquisition and Matching System Combination



Homeland Security
Science and Technology

Unattended High-throughput Scenario Summary

- Face recognition technology can work well across demographic groups without face masks
  - Similar to findings from past Rally scenario tests

- However, acquisition and matching errors do not increase equally when the system is perturbed by the addition of face masks
  - Performance declines for some demographic groups more than for others
  - Both acquisition and matching performance is affected; it is not just the matching algorithm
  - Future research will investigate the differential performance of the technology that underlies these differential outcomes

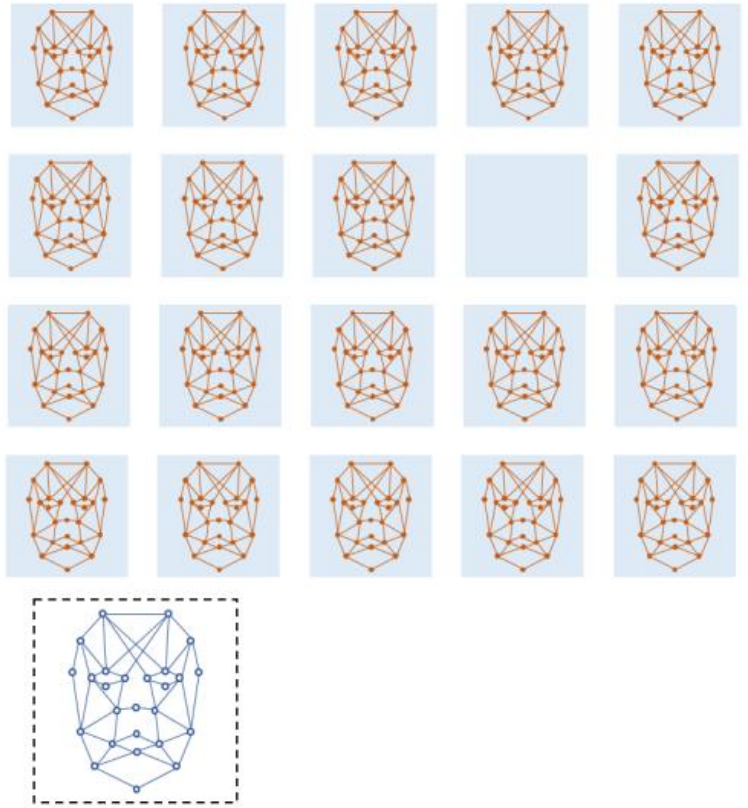
> Takeaway <

A fair system under one set of operational conditions, may become unfair when conditions change.

Ongoing testing is recommended to track performance, including fairness, as conditions change.

Large Watch-list Identification Scenario

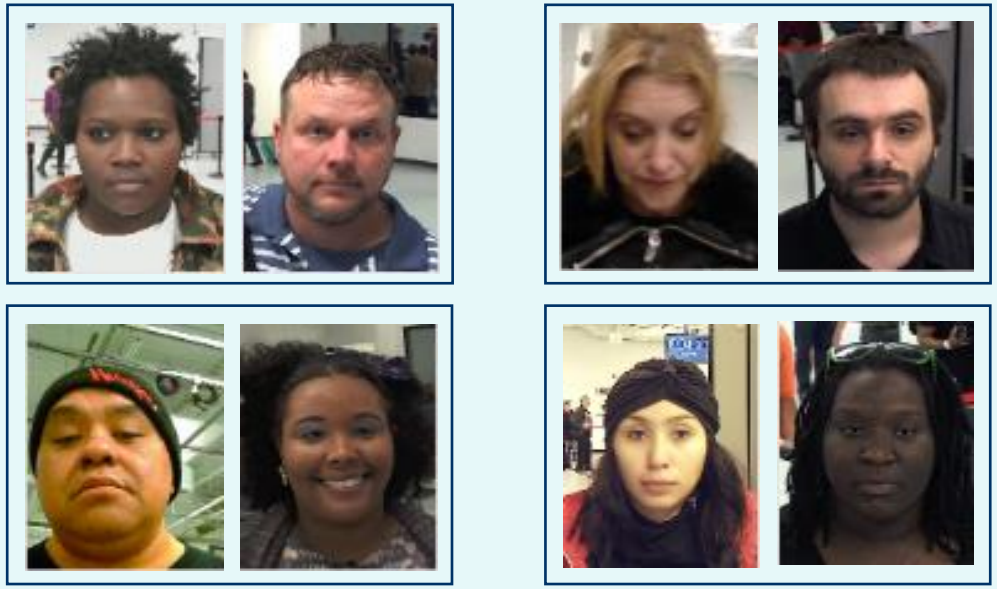
- The face recognition system receives many images for matching (e.g., from various sources)
- The identification gallery is large (1000+ people)
- Most people being matched are not in the identification gallery
- Impact of errors of those being matched is dominated by false positive identifications
  - **Example Impact:** Investigation by authorities



Homeland Security
Science and Technology

Face Recognition False Positives and Demographics

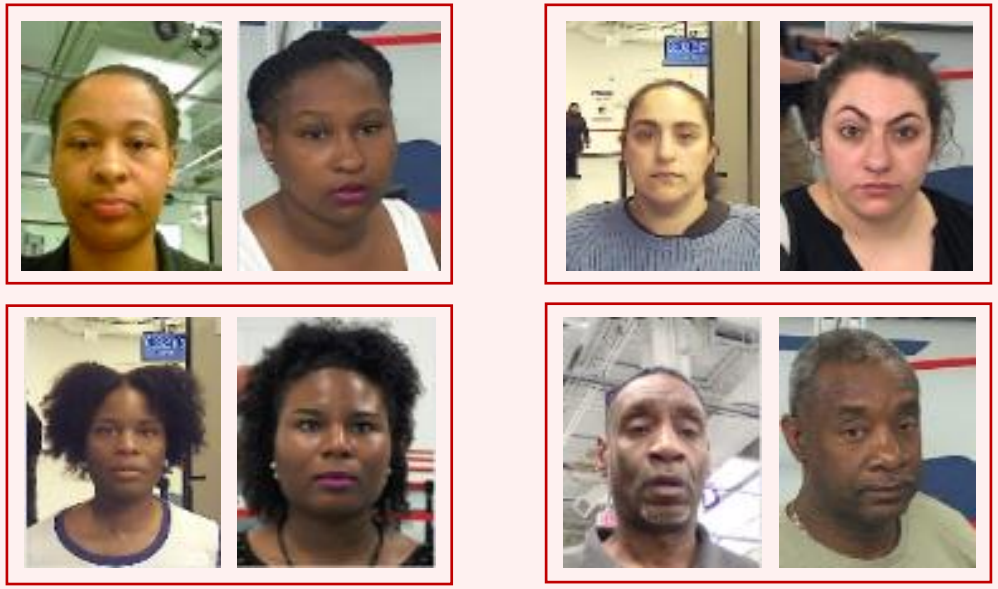
Iris False-Matches



~1 in 4 iris false matches are of the same Race and Gender (FMR = 1e-5)

> As expected from random assortment <

Face False-Matches



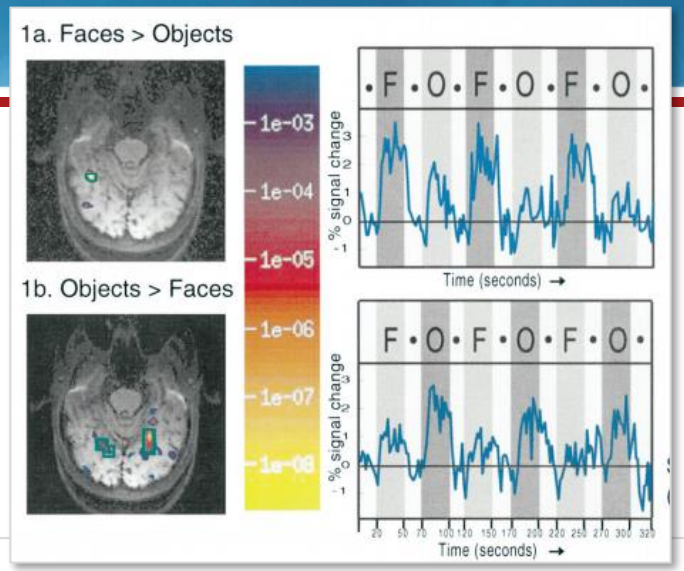
~4 in 5 face false matches are of same Race and Gender (FMR = 1e-5)

> Intuitive? Why? <

Homeland Security
Science and Technology

Our Unique Neurobiology Biases Our Intuition for Face Recognition

- Humans have a dedicated perceptual face recognition capability
  - This architecture is shared with other primates (e.g., Macaques)
  - Evolved to recognize familiar individuals within small social groups
- A landmark neurophysiological study in 1997 identified areas of the human brain dedicated to face processing
  - No dedicated areas exist for fingerprint or iris processing
- Our own face processing capability biases our judgement about how face recognition should work:
  - Our perceptual system is bad at distinguishing unfamiliar individuals
  - It is easier to distinguish unfamiliar individuals based on race or gender
  - We assume face recognition algorithms must work this way because this is intuitive
  - But this is not an assumption we make about fingerprint or iris biometrics because we don't have the right neurobiology



The Journal of Neuroscience, June 1, 1997, 17(11):4302–4311

The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception

Nancy Kanwisher,[1,2] **Josh McDermott,**[1,2] **and Marvin M. Chun**[2,3]

[1]*Department of Psychology, Harvard University, Cambridge, Massachusetts 02138,* [2]*Massachusetts General Hospital NMR Center, Charlestown, Massachusetts 02129, and* [3]*Department of Psychology, Yale University, New Haven, Connecticut 06520-8205*

Using functional magnetic resonance imaging (fMRI), we found an area in the fusiform gyrus in 12 of the 15 subjects tested that was significantly more active when the subjects viewed faces than when they viewed assorted common objects. This face activation was used to define a specific region of interest individually for each subject, within which several new tests of face specificity were run. In each of five subjects tested, the predefined candidate "face area" also responded significantly more strongly to passive viewing of (1) intact than scrambled two-tone faces, (2) full front-view face photos than front-view photos of houses, and (in a different set of five subjects) (3) three-quarter-view face photos (with hair concealed) than photos of human hands; it also responded more strongly during (4) a consecutive matching task performed on three-quarter-view faces versus hands. Our technique of running multiple tests applied to the same region defined functionally within individual subjects provides a solution to two common problems in functional imaging: (1) the requirement to correct for multiple statistical comparisons and (2) the inevitable ambiguity in the interpretation of any study in which only two or three conditions are compared. Our data allow us to reject alternative accounts of the function of the fusiform face area (area "FF") that appeal to visual attention, subordinate-level classification, or general processing of any animate or human forms, demonstrating that this region is *selectively* involved in the perception of faces.

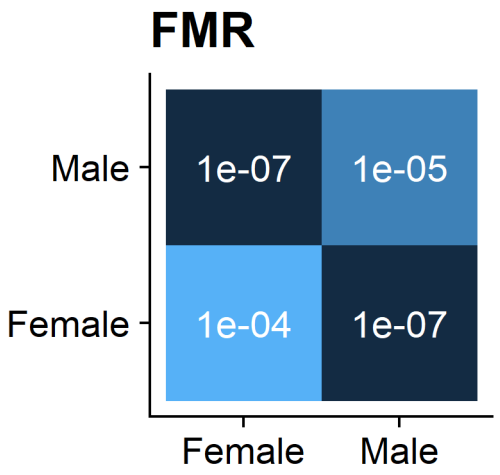
Key words: extrastriate cortex; face perception; functional MRI; fusiform gyrus; ventral visual pathway; object recognition

Evidence from cognitive psychology (Yin, 1969; Bruce et al., 1991; Tanaka and Farah, 1993), computational vision (Turk and Pent-

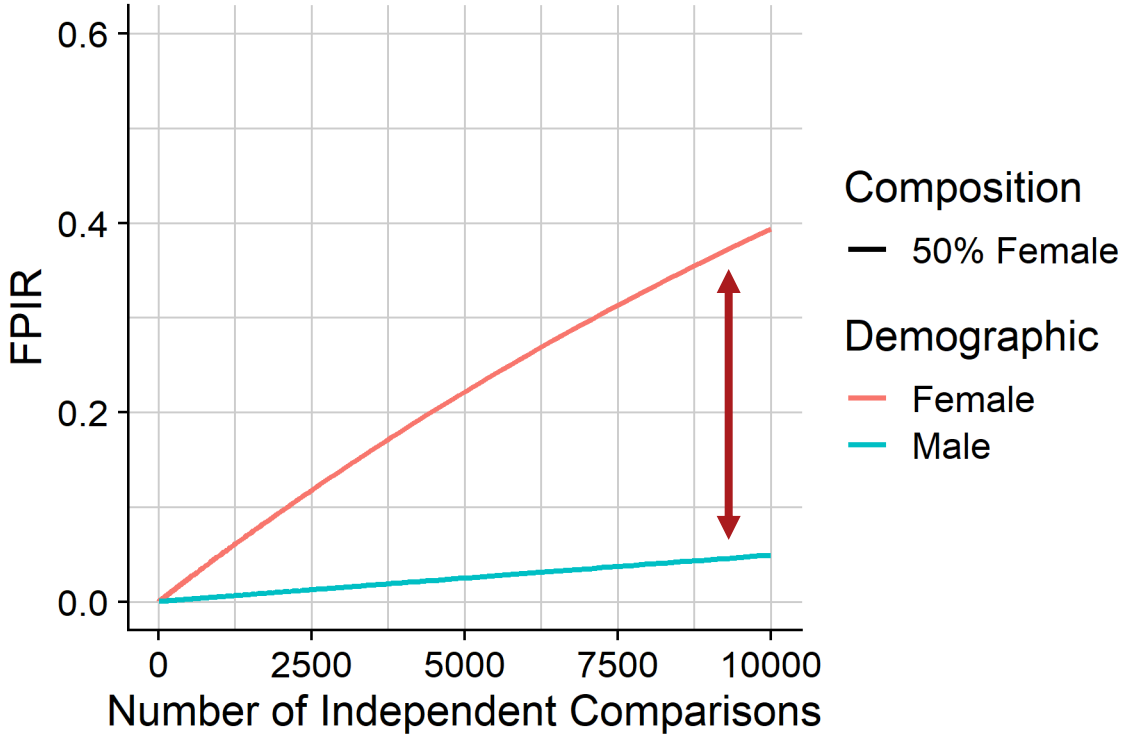
us to study cortical specialization in the normal human brain with relatively high spatial resolution and large sampling area. Past

Differentials in within-group FMR may result in Big FPIR differentials

- Face recognition algorithm False Match Rates (FMR) can vary within demographic groups:
  - NIST FRVT found up to 100-fold disparity in FMR within different demographic groups (typically 10-fold)
  - This is broadly recognized as a problem in the biometrics community
  - With simple assumptions, small FMR differences lead to large differentials in False Positive Identification Rates (FPIR) over many comparisons



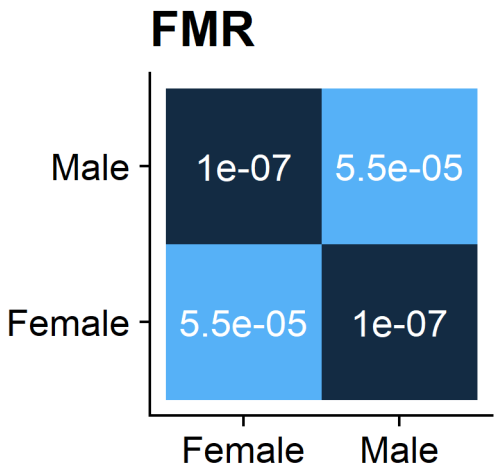
$$FPIR = 1 - (1 - FMR)^{N}$$



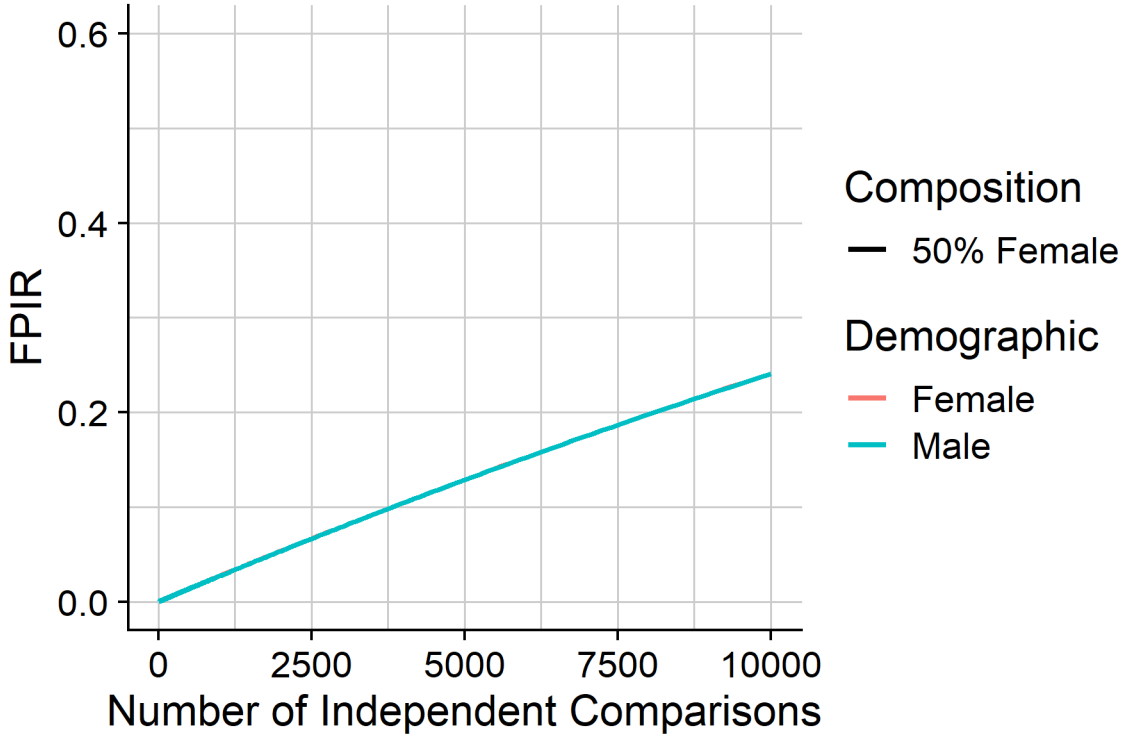
Homeland Security
Science and Technology

Equal within-group FMR + Balanced Gallery = No FPIR differentials

- With equal false match rates within each group:
  - FPIR differentials for a balanced gallery (50:50 male and female) are eliminated
  - This is broadly recognized as the desired ***end state*** in the biometric community and makes intuitive sense
  - **But what if the gallery composition is not balanced?**



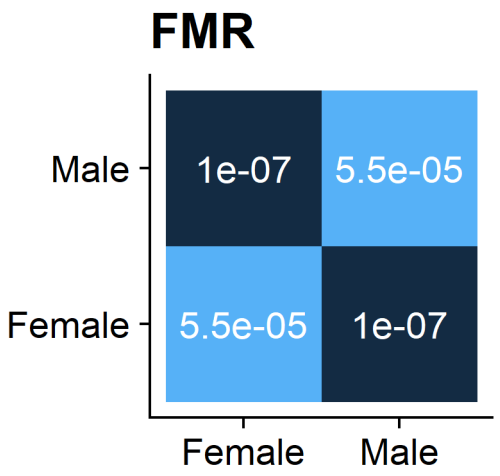
$$FPIR = 1 - (1 - FMR)^{N}$$



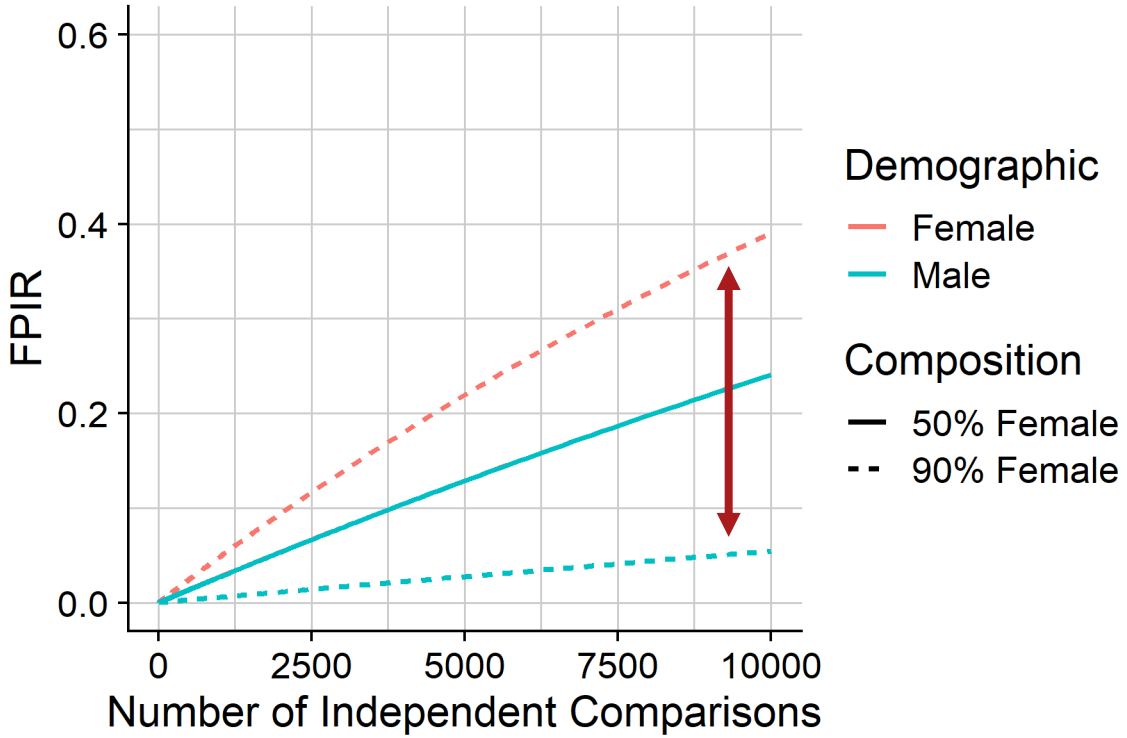
Homeland Security
Science and Technology

Equal within-group FMR + Imbalanced Gallery may result in Big FPIR differentials

- Despite equal false match rates within each group:
  - FPIR differentials for an imbalanced gallery (10% male to 90% female) are again observed!
  - Size of the FPIR differential depends on amount of imbalance in the gallery
  - **How can these differentials be mitigated?**



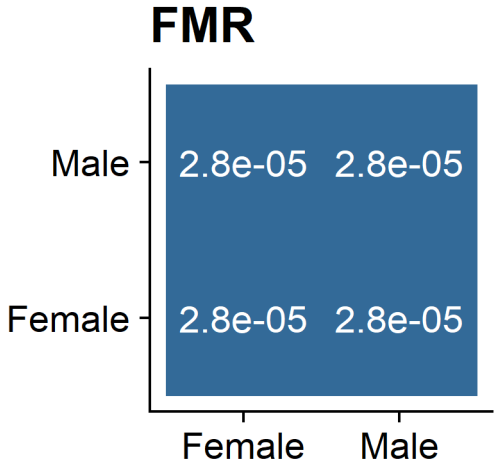
$$FPIR = 1 - (1 - FMR)^{N}$$



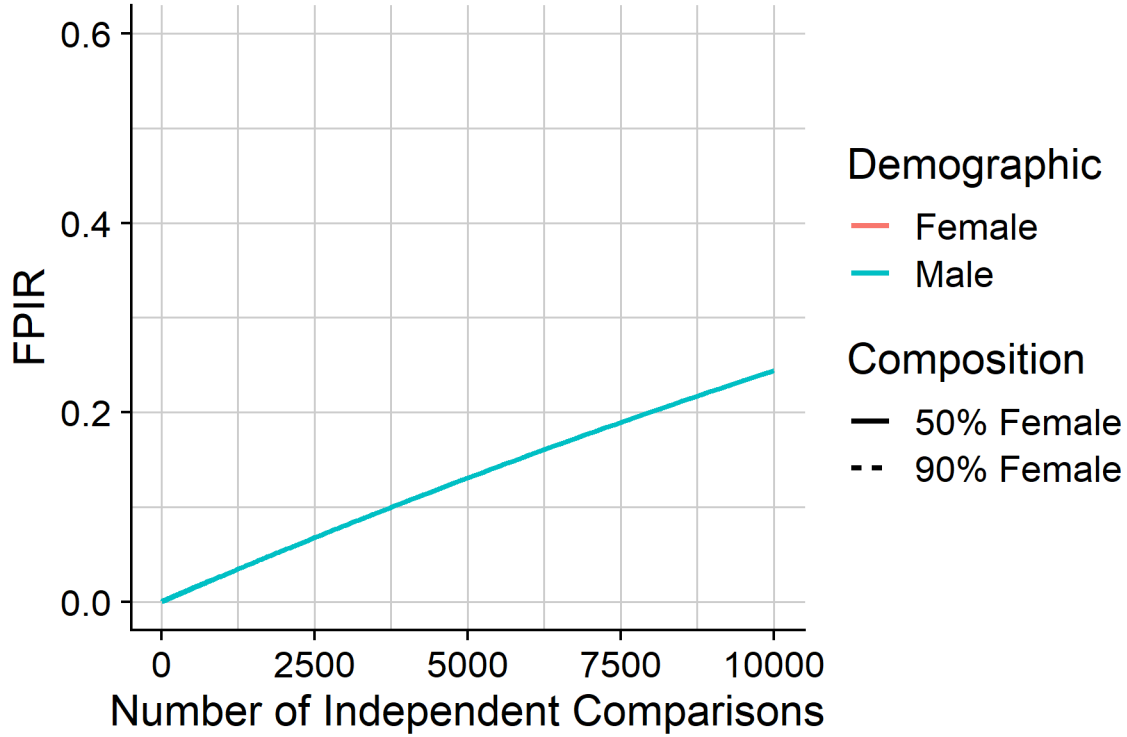
Homeland Security
Science and Technology

Homogeneous FMR = No FPIR differentials Independent of Gallery

- Despite equal false match rates within each group:
  - If face recognition behaved like iris recognition, false match errors would be random with respect to demographic groups
  - We call such false match rates broadly homogeneous, but it goes against human face recognition intuition
  - Homogeneous FMR is not broadly recognized as a desirable ***end-state*** for face recognition

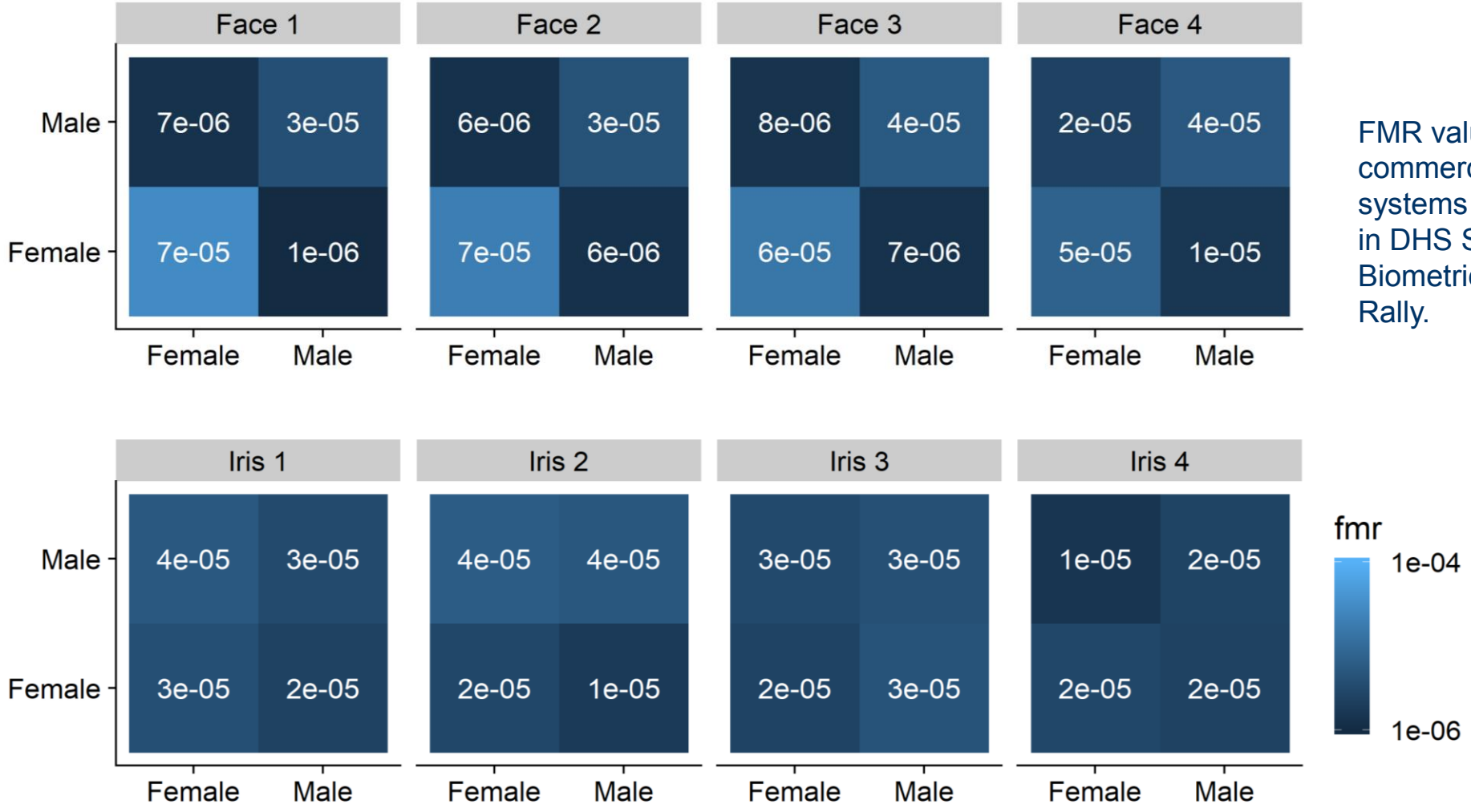


$$FPIR = 1 - (1 - FMR)^{N}$$



Homeland Security
Science and Technology

False Positive Differentials in Commercial Face and Iris Recognition



FMR values based on commercial matching systems participating in DHS S&T's 2019 Biometric Technology Rally.

Homeland Security
Science and Technology

Large Watch-list Identification Scenario Summary

- Differences in FMR linked with race or gender can create an unequal hazard of false positive identification against a watch-list for people based on demographic traits outside their control

- Current focus in face-recognition is to achieve equal within-group error rates
  - This goal is biased by intuition that derives from our own unique neurobiology
  - But, equalizing within-group FMR will create equal FPIR for each group ONLY when gallery composition is exactly balanced
  - Some watch-list galleries may not be balanced for race, gender, or other protected groups

- False Match Rates for iris recognition can be homogeneous:
  - i.e., independent of race or gender both within-groups and between-groups

> Takeaway <

Homogeneous FMRs maintain equal FPIR independent of demographic group membership and gallery composition. To be fair, face identification systems should have homogeneous FMRs.

Questions?

- This work was performed by a dedicated team of researchers at the Maryland Test Facility

- Find out more at https://mdtf.org/

- yevgeniy@mdtf.org
- arun.vemury@hq.dhs.gov

- john@mdtf.org
- jacob@mdtf.org

Homeland Security
Science and Technology