U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Key Considerations when Evaluating the Performance of Facial Recognition Systems: Cameras, Humans, and Demographics

Science and Technology

John J. Howard
Principal Scientist
The Maryland Test Facility

Arun Vemury
Director
Biometric and Identity Technology Center

July 7th, 2022

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.

Science and Technology

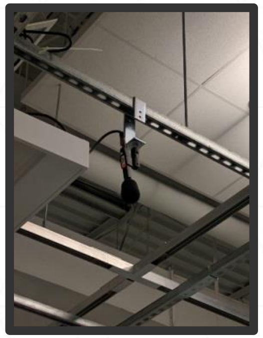
Biometric & Identity Technology Center

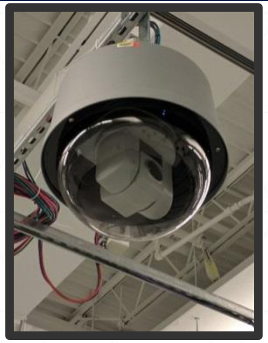
S&T conducts foundational research to ensure advancements in science and technology are harnessed for cutting-edge solutions to new and emerging operational challenges.

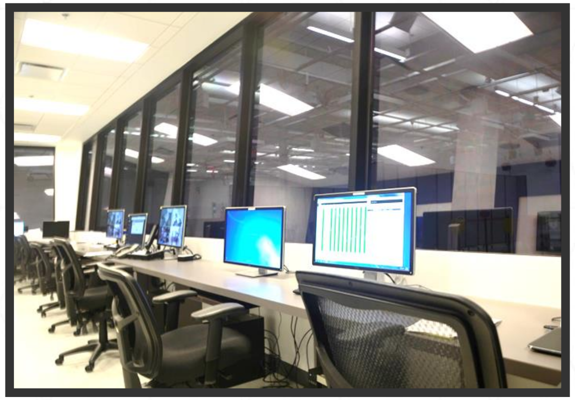
- Drive biometric and identity innovation at DHS through RDT&E capabilities
- Facilitate and accelerate understanding of biometrics and identity technologies for new DHS use cases
- Drive efficiencies by supporting cross cutting methods, best practices, and solutions across programs
- Deliver Subject Matter Expertise across the DHS enterprise
- Engage Industry and provide feedback
- Encourage Innovation with Industry and Academia

The Maryland Test Facility (MdTF)

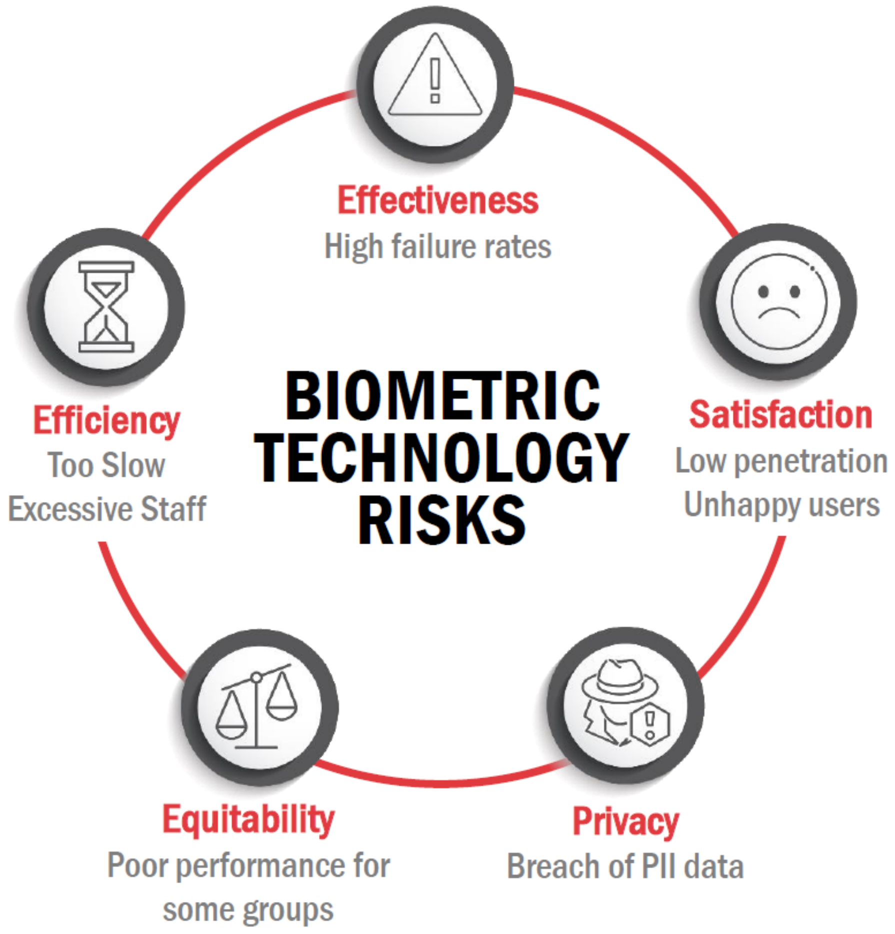
- Founded in 2014 by the Department of Homeland Security, Science and Technology Directorate.
- 20,000 ft^2 of office and reconfigurable laboratory space
- Fully instrumented and designed for human subject testing
  - Data collection infrastructure: Cameras, ambient light, noise, humidity, real time control center and monitoring capability, informed consent collection facilities, etc.
- Since its founding over 2500 subjects have participated in biometric testing at the MdTF
  - Ages 18-72
  - 114 countries of origin

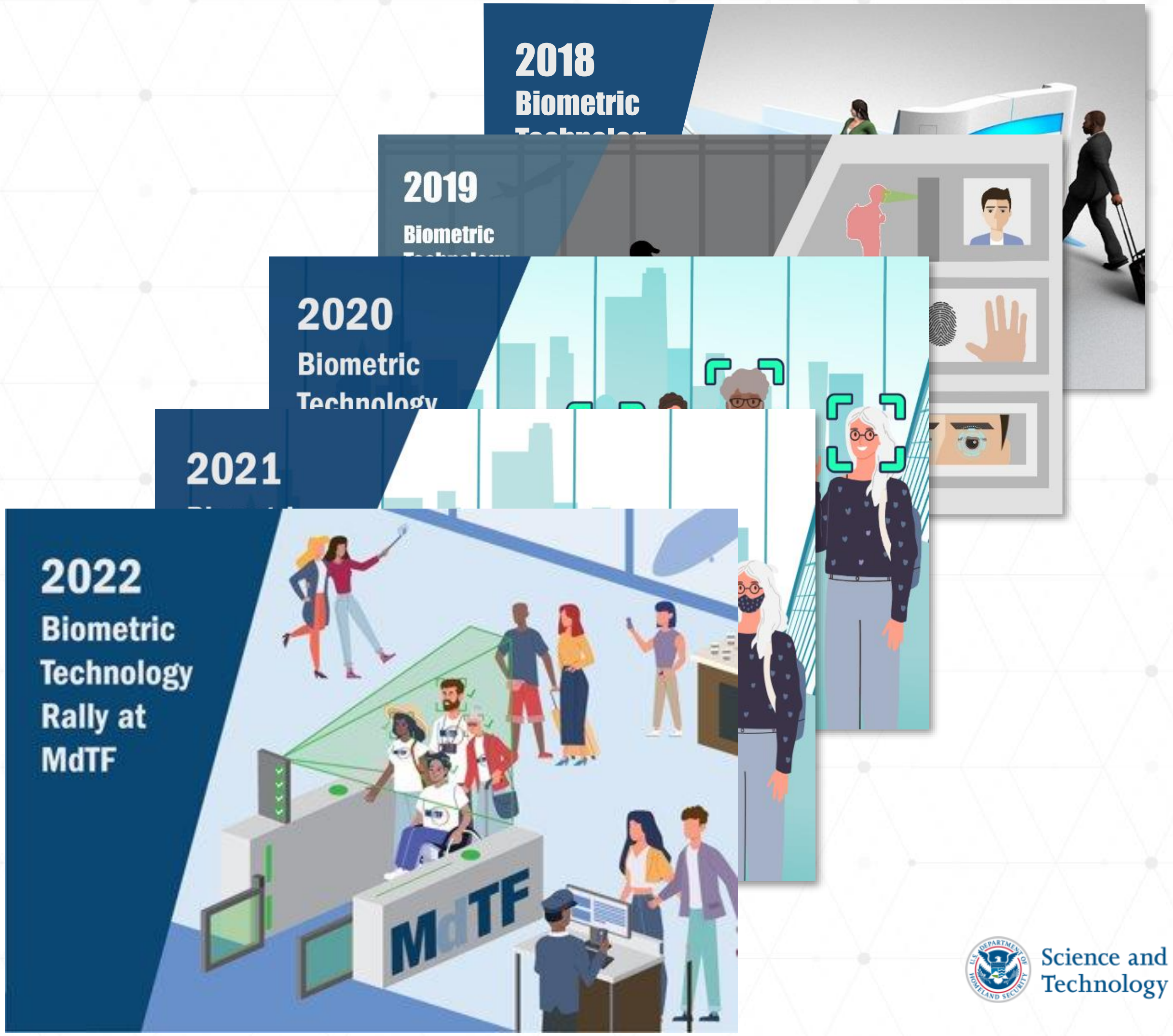






DHS S&T Biometric Technology Rallies





Scenario Testing vs. Technology Testing

- **Scenario Testing:**
  - Centered around a use-case,
  - Full multi-component biometric system,
  - Gathering new biometric samples,
  - Smaller sample size. Important to delineate the effect size you can find

  - 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. Important to delineate the effect size you are looking for.

  - 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. <

Scenario Testing

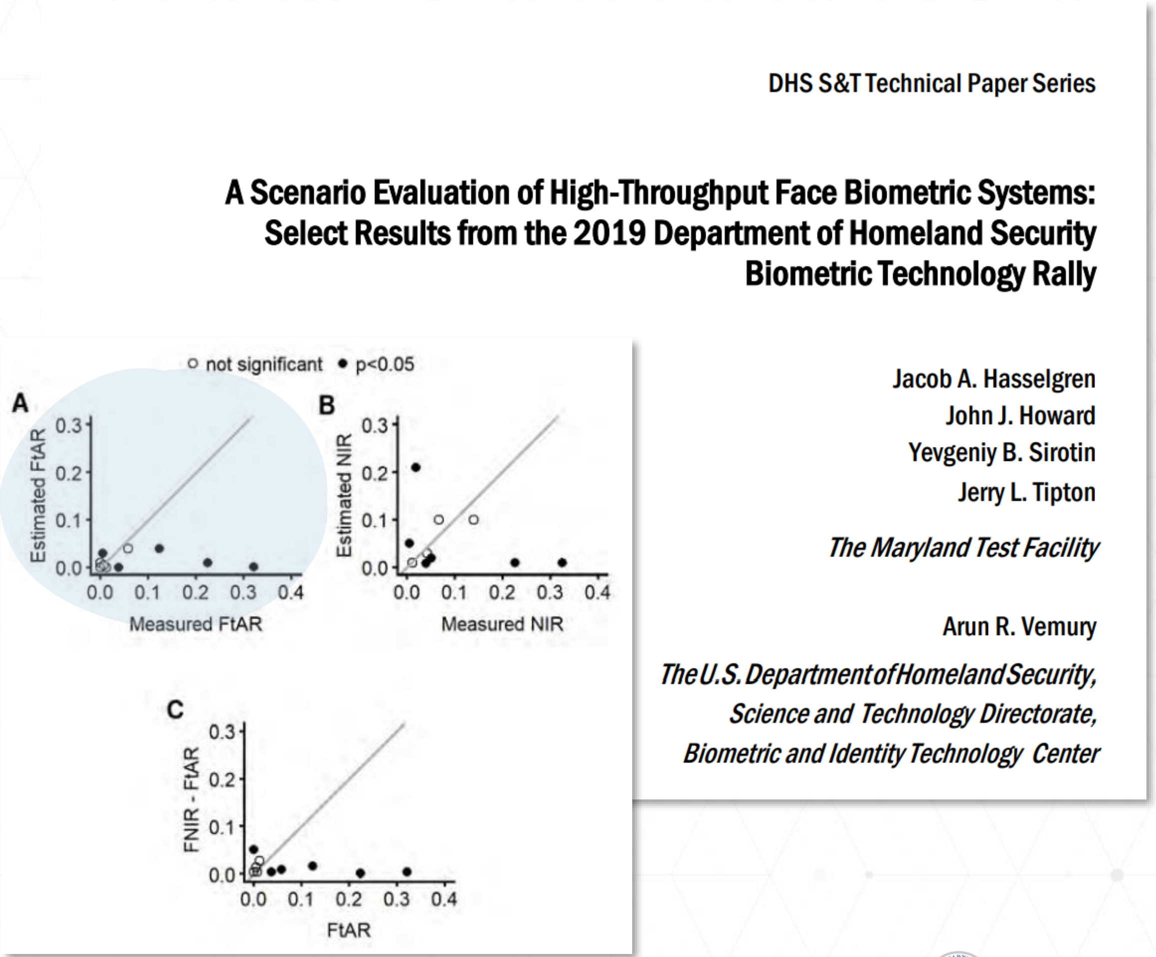
- Answers key questions not addressed by technology testing:
  - What is the performance of the full facial recognition system (camera + human computer interface + matching system).
  - What is the performance in a simulated, real world environment?
  - Are their demographic effects in the full system? What part of the system can those effects be attributes to?

- Is a necessary part of pre-deployment testing of facial recognition systems

Scenario Testing, Lesson 1: Acquisition Errors can Drive Performance.

- In 2019 DHS S&T examined the major source of errors in high-throughput unstaffed biometric systems.

- 2019 Rally compared acquisition error to matching error:
  - **Finding 1: Vendors under-estimate failure to acquire.**
  - Finding 2: Measured acquisition error can be much higher than matching error.



DHS S&T Technical Paper Series

A Scenario Evaluation of High-Throughput Face Biometric Systems: Select Results from the 2019 Department of Homeland Security Biometric Technology Rally

Jacob A. Hasselgren
John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility

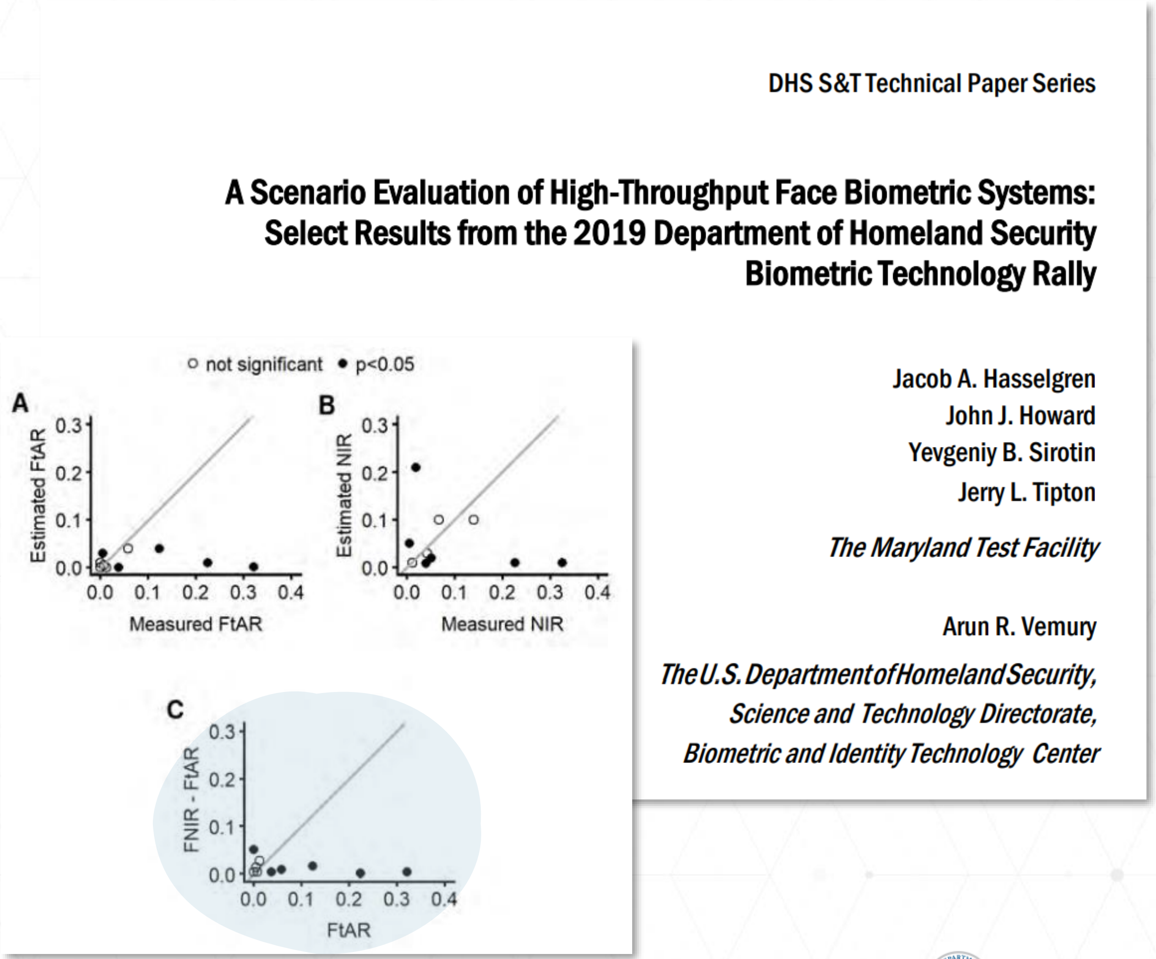
Arun R. Vemury

The U.S. Department of Homeland Security, Science and Technology Directorate, Biometric and Identity Technology Center

Scenario Testing, Lesson 1: Acquisition Errors can Drive Performance.

- In 2019 DHS S&T examined the major source of errors in high-throughput unstaffed biometric systems.

- 2019 Rally compared acquisition error to matching error:
  - Finding 1: Vendors under-estimate failure to acquire.
  - **Finding 2: Measured acquisition error can be much higher than matching error.**


DHS S&T Technical Paper Series

A Scenario Evaluation of High-Throughput Face Biometric Systems: Select Results from the 2019 Department of Homeland Security Biometric Technology Rally

Jacob A. Hasselgren
John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility

Arun R. Vemury

The U.S. Department of Homeland Security, Science and Technology Directorate, Biometric and Identity Technology Center

Scenario Testing, Lesson 1: Acquisition Errors can Drive Performance.

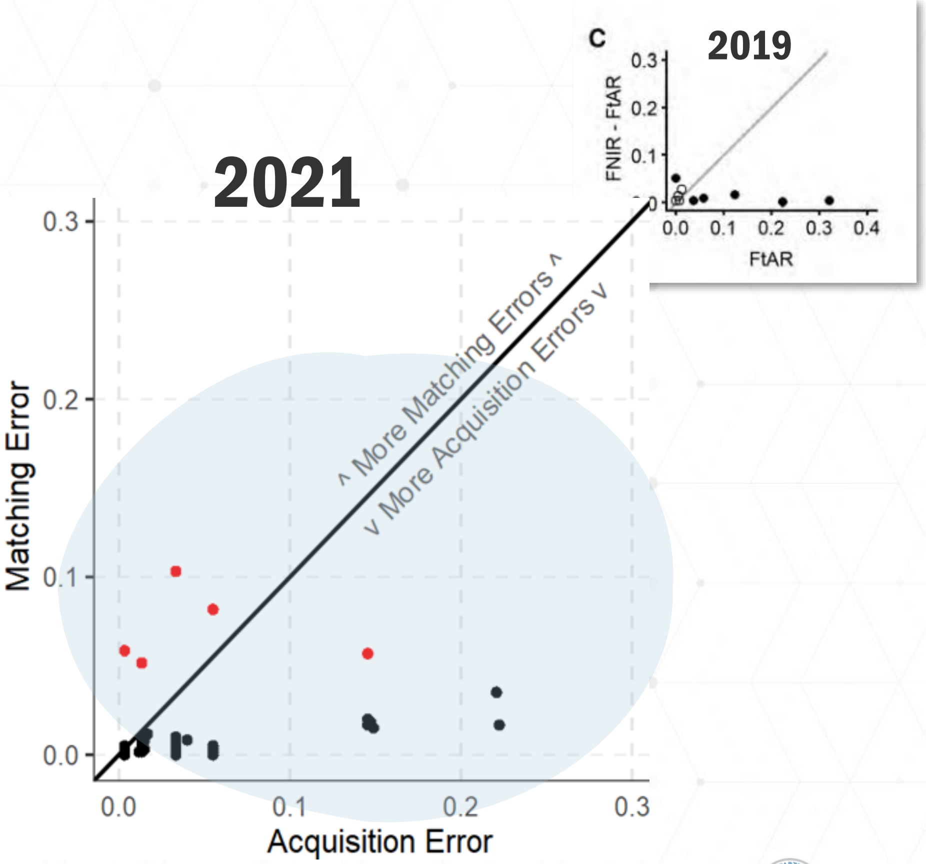
- In the 2021 Rally, DHS S&T again measured the primary source of error in 50 combinations of acquisition and matching systems.

- 75% of system combinations had acquisition errors in excess of matching errors.

> Acquisition continues to be the main source of error in high throughput, unstaffed face-recognition systems. <

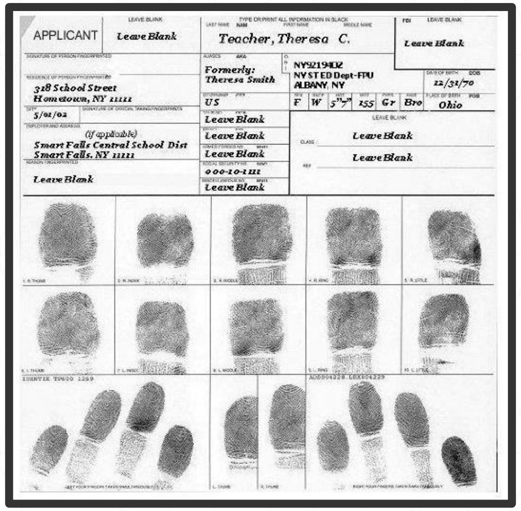
> Vendors are often unaware of this. <

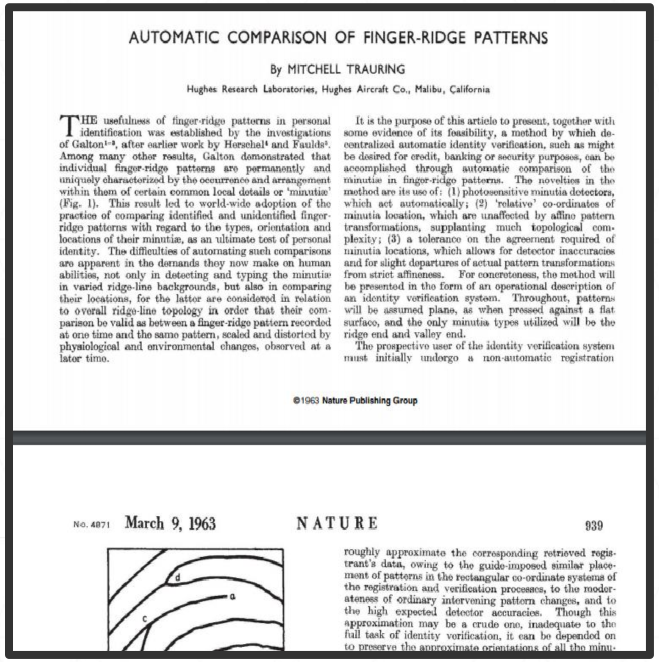
>This can be discovered in scenario testing. Difficult to ascertain in technology/operational <



Scenario Testing, Lesson 2: Demographic Effects Exist, Our Understanding of Them May be Clouded.

- A brief biometric history:
  - Fingerprint Recognition:
    - Oldest, non-innate biometric modality, dating to the 1800s
    - U.S. Fingerprint repository began at FBI in 1924
    - Estimated over 200 million cards processed from 1924 – 1999
    - First automated in 1963 by Trauring
    - 2008 – 63,000 fingerprint receipts daily





AUTOMATIC COMPARISON OF FINGER-RIDGE PATTERNS

By MITCHELL TRAURING

Hughes Research Laboratories, Hughes Aircraft Co., Malibu, California

THE usefulness of finger-ridge patterns in personal identification was established by the investigations of Galton[1-3], after earlier work by Herschel[4] and Faulds[5]. Among many other results, Galton demonstrated that individual finger-ridge patterns are permanently and uniquely characterized by the occurrence and arrangement within them of certain common local details or 'minutiæ' (Fig. 1). This result led to world-wide adoption of the practice of comparing identified and unidentified finger-ridge patterns with regard to the types, orientation and locations of their minutiæ, as an ultimate test of personal identity. The difficulties of automating such comparisons are apparent in the demands they now make on human abilities, not only in detecting and typing the minutiæ in varied ridge-line backgrounds, but also in comparing their locations, for the latter are considered in relation to overall ridge-line topology in order that their comparison be valid as between a finger-ridge pattern recorded at one time and the same pattern, scaled and distorted by physiological and environmental changes, observed at a later time.

It is the purpose of this article to present, together with some evidence of its feasibility, a method by which decentralized automatic identity verification, such as might be desired for credit, banking or security purposes, can be accomplished through automatic comparison of the minutiæ in finger-ridge patterns. The novelties in the method are its use of: (1) photosensitive minutia detectors, which act automatically; (2) 'relative' co-ordinates of minutia location, which are unaffected by affine pattern transformations, supplanting much topological complexity; (3) a tolerance on the agreement required of minutia locations, which allows for detector inaccuracies and for slight departures of actual pattern transformations from strict affineness. For concreteness, the method will be presented in the form of an operational description of an identity verification system. Throughout, patterns will be assumed plane, as when pressed against a flat surface, and the only minutia types utilized will be the ridge end and valley end.

The prospective user of the identity verification system must initially undergo a non-automatic registration

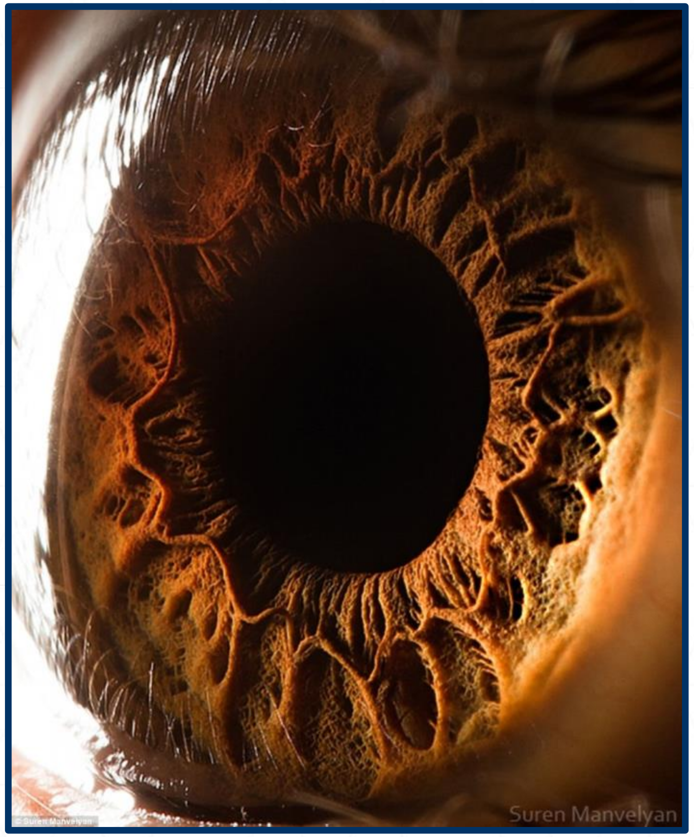
No. 4871 March 9, 1963 NATURE 939

roughly approximate the corresponding retrieved registrant's data, owing to the guide-imposed similar placement of patterns in the rectangular co-ordinate systems of the registration and verification processes, to the moderateness of ordinary intervening pattern changes, and to the high expected detector accuracies. Though this approximation may be a crude one, inadequate to the full task of identity verification, it can be depended on to preserve the approximate orientations of all the minu-

Science and Technology

Scenario Testing, Lesson 2: Demographic Effects Exist, Our Understanding of Them May be Clouded.

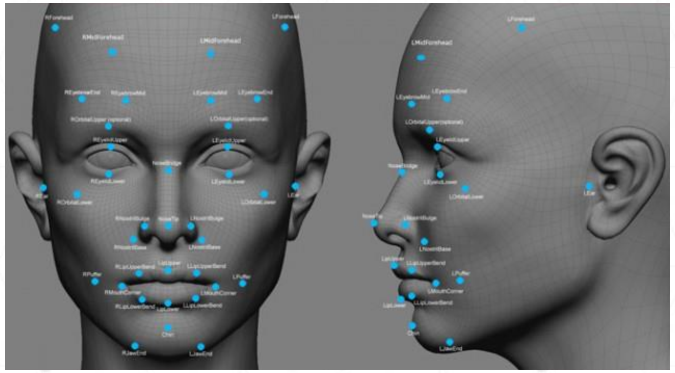
- A brief biometric history:
  - Iris Recognition:
    - 1985 Safir and Flom patent – “Methods and apparatus are disclosed for identifying an eye, especially a human eye, on the basis of the visible features of the iris and pupil”
    - 1991 – John Daugman formalized & automated the process
    - 2004 – Method released publicly
    - Limited adoption in the U.S. Border and travel adoptions here and abroad throughout the 2000s.
      - US Canada Nexus Program (2000)
      - UAE Border (2001)
      - India UIDAI (2009)



Science and Technology

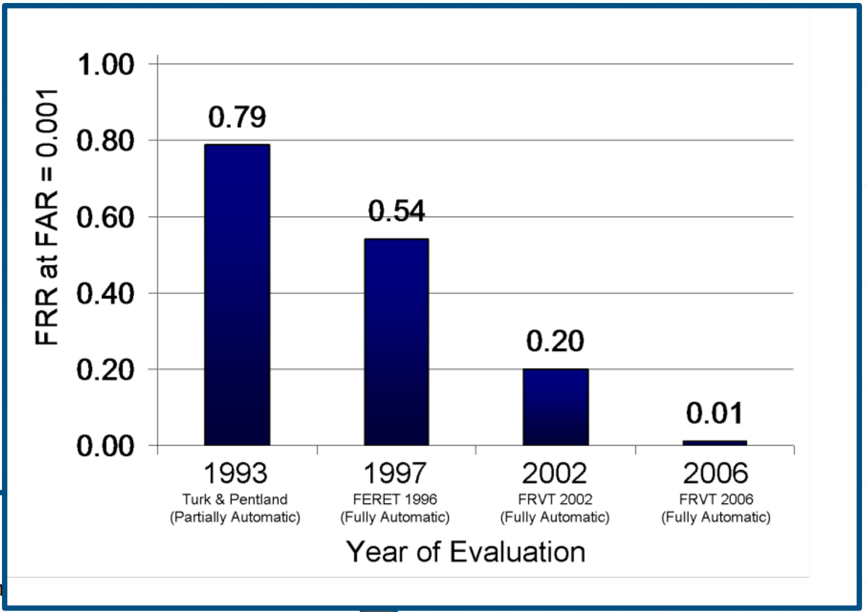
Scenario Testing, Lesson 2: Demographic Effects Exist, Our Understanding of Them May be Clouded.

- A brief biometric history:
  - Face Recognition:
    - Early approaches date to around the same time automated fingerprints 1960s - based on distances & ratios between facial points
    - Eigenfaces, fundamental face vectors, in the 1990s was major improvement.

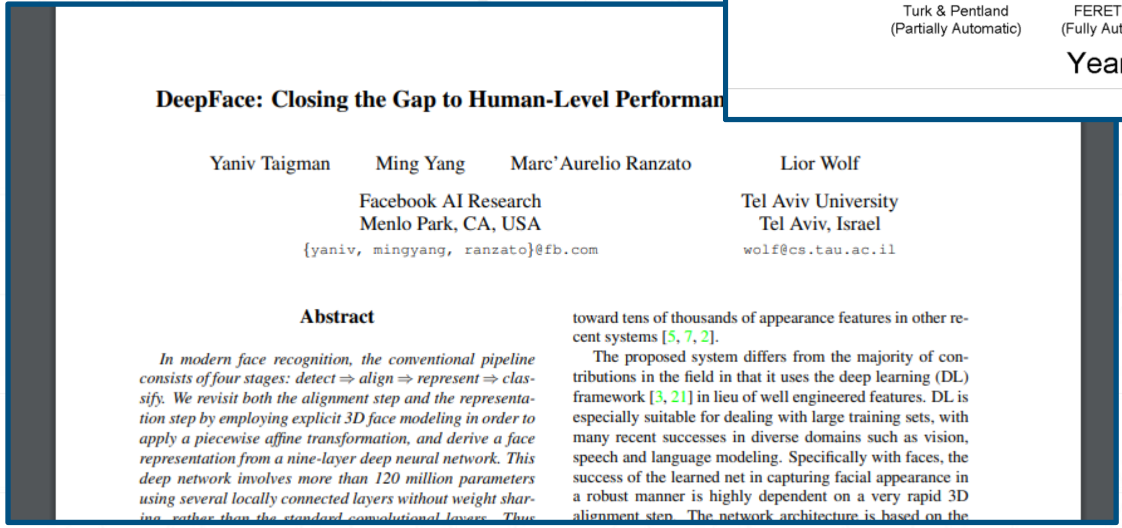


Scenario Testing, Lesson 2: Demographic Effects Exist, Our Understanding of Them May be Clouded.

- A brief biometric history:
  - Face Recognition:
    - Led to first national testing program (NIST FERET) in 1993
    - Results improved slowly through the 2000s
    - Then came the application of AI in 2014



- Ongoing NIST FRVT 1:1 Challenge (February 9, 2021):
  - 271 algorithms from over 200 different companies
  - 1:1 – now have a 0.2 % non match rate at a false match rate of 1 in a million
- Allowed us to start thinking about doing *identification* operations with face



DeepFace: Closing the Gap to Human-Level Performan

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Research
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il

Abstract

In modern face recognition, the conventional pipeline consists of four stages: detect ⇒ align ⇒ represent ⇒ classify. We revisit both the alignment step and the representation step by employing explicit 3D face modeling in order to apply a piecewise affine transformation, and derive a face representation from a nine-layer deep neural network. This deep network involves more than 120 million parameters using several locally connected layers without weight shar-

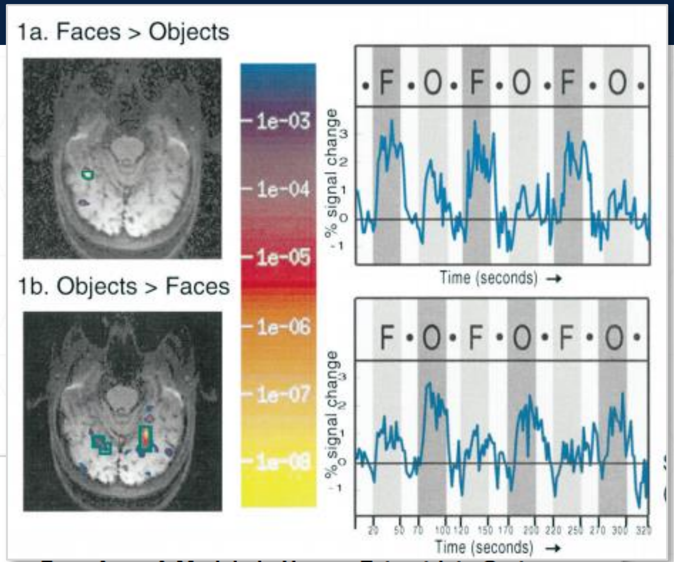
toward tens of thousands of appearance features in other recent systems [5, 7, 2].

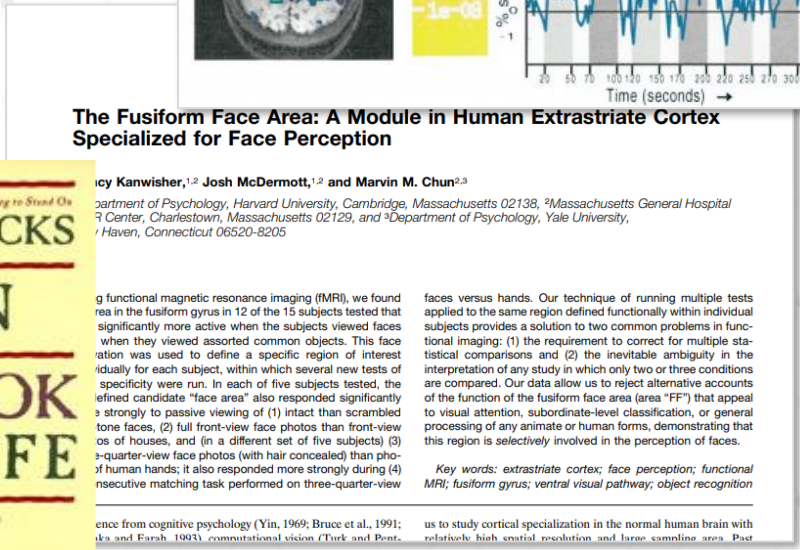
The proposed system differs from the majority of contributions in the field in that it uses the deep learning (DL) framework [3, 21] in lieu of well engineered features. DL is especially suitable for dealing with large training sets, with many recent successes in diverse domains such as vision, speech and language modeling. Specifically with faces, the success of the learned net in capturing facial appearance in a robust manner is highly dependent on a very rapid 3D

But Faces are Fundamentally Different for (at least) Two Reasons

- Faces are genetic, iris and fingerprint characteristics are determined during development.
  - To us, individuals look more like their parents, siblings, and those that share racial and gender categories.

- Humans have an innate ability to perform face recognition tasks, not so with iris and fingerprints.
  - Humans have dedicated brain areas that process faces quickly
  - This was an important function for human evolution
    - Mates, Friends, Foes, Family members
    - Other primates have a similar capability
  - Intuitively perceive same-gender and same-race faces as more similar
  - We even know the exact part of the human brain dedicated to face processing.
    - Evolved to recognize familiar individuals within small social groups (25-100)
  - Prosopagnosia – “face blindness”



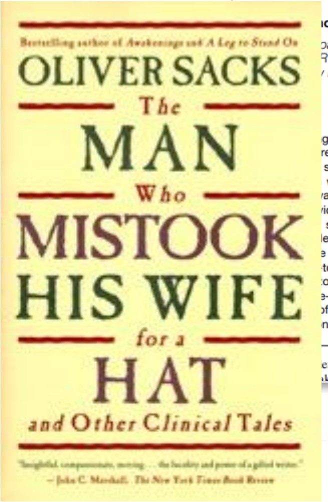

The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception

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

...partment of Psychology, Harvard University, Cambridge, Massachusetts 02138, [2]Massachusetts General Hospital ...R Center, Charlestown, Massachusetts 02129, and [3]Department of Psychology, Yale University, ...Haven, Connecticut 06520-8205

...g functional magnetic resonance imaging (fMRI), we found ...rea in the fusiform gyrus in 12 of the 15 subjects tested that ...significantly more active when the subjects viewed faces ...when they viewed assorted common objects. This face ...ation was used to define a specific region of interest ...idually for each subject, within which several new tests of ...specificity were run. In each of five subjects tested, the ...efined candidate “face area” also responded significantly ...strongly to passive viewing of (1) intact than scrambled ...tone faces, (2) full front-view face photos than front-view ...os of houses, and (in a different set of five subjects) (3) ...-quarter-view face photos (with hair concealed) than pho... ...f human hands; it also responded more strongly during (4) ...nsecutive 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

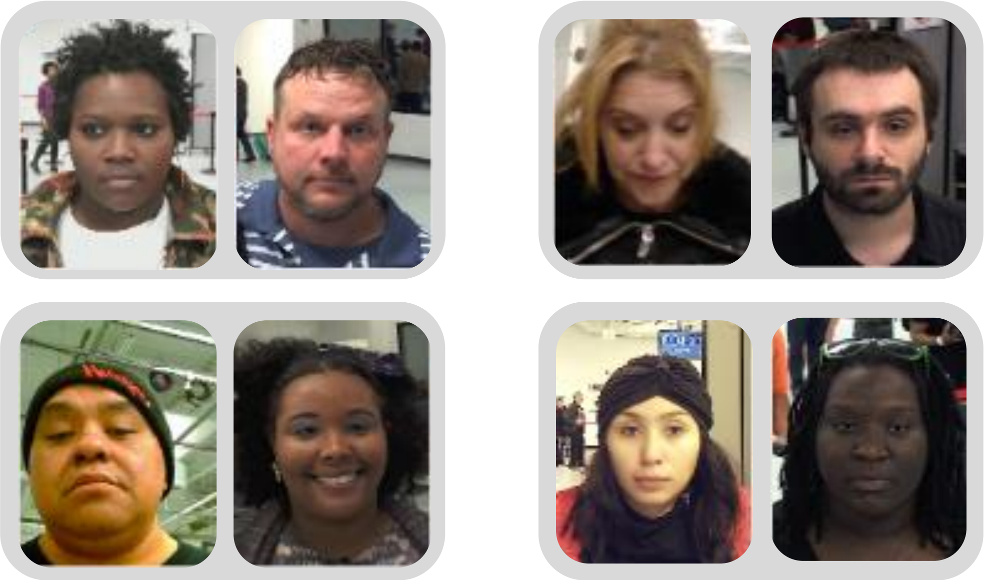


Science and Technology

Scenario Testing, Lesson 2: Demographic Effects Exist, Our Understanding of Them may be Clouded.

> It may seem natural to us that face recognition “clusters” people based on race and gender <

Iris recognition



Iris recognition false positives were random relative to race and gender

Face recognition

80% of face recognition false positives were between people of the same race and gender

Subjects consent for use of their image in publications was obtained

ience and Technology

This "clustering" is often referred to negatively

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › news feature › article

NEWS FEATURE | 18 November 2020

Is facial recognition too biased to be let loose?

The technology is improving – but the bigger issue is how it's used.

MIT Technology Review

Featured Topics Newsletters Events Podcasts

Artificial intelligence Dec 20

A US government study confirms most face recognition systems are racist

Bloomberg

Technology

EU Data Watchdogs Call for Ban on Facial Recognition Through AI

By Stephanie Bodoni +Follow

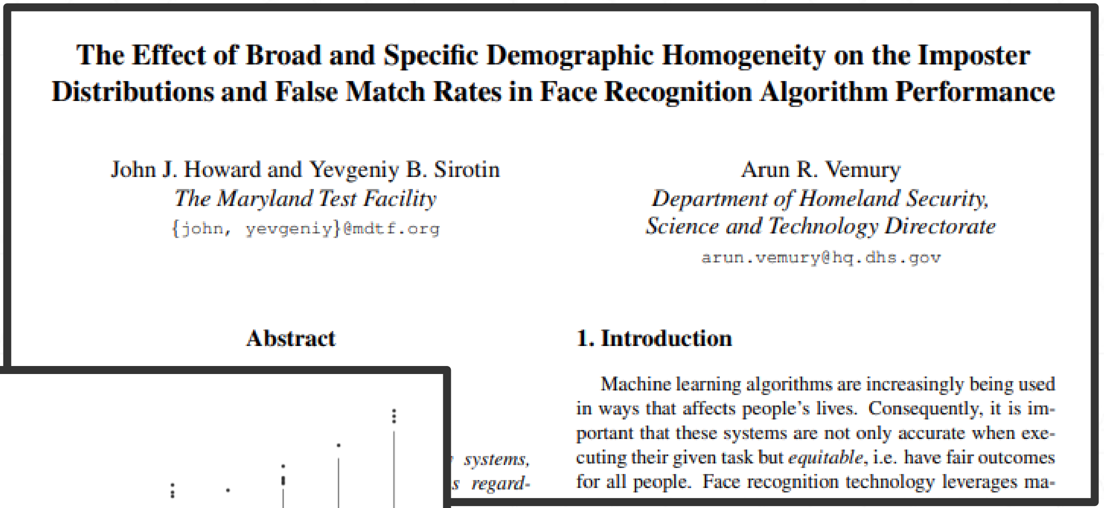
June 21, 2021, 7:48 AM EDT

The two bodies charged with overseeing compliance with the bloc's strict data protection rules called for the ban "on any use of AI." The embargo should cover remote biometric identification of people in public and the use of technology "to categorize individuals into clusters based on ethnicity, gender, political or sexual orientation," which could lead to discrimination.

Science and Technology

It is also (likely) (currently) a Universal Feature of Face Recognition

- We first highlighted this in 2019 using one commercial algorithm
- NIST subsequently confirmed this exists in 138 algorithms
  - NIST FRVT Part 3: Demographics – Annex 5.



The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance

John J. Howard and Yevgeniy B. Sirotin
The Maryland Test Facility
{john, yevgeniy}@mdtf.org

Arun R. Vemury
Department of Homeland Security, Science and Technology Directorate
arun.vemury@hq.dhs.gov

Abstract

1. Introduction

Machine learning algorithms are increasingly being used in ways that affects people's lives. Consequently, it is important that these systems are not only accurate when executing their given task but *equitable*, i.e. have fair outcomes for all people. Face recognition technology leverages ma-

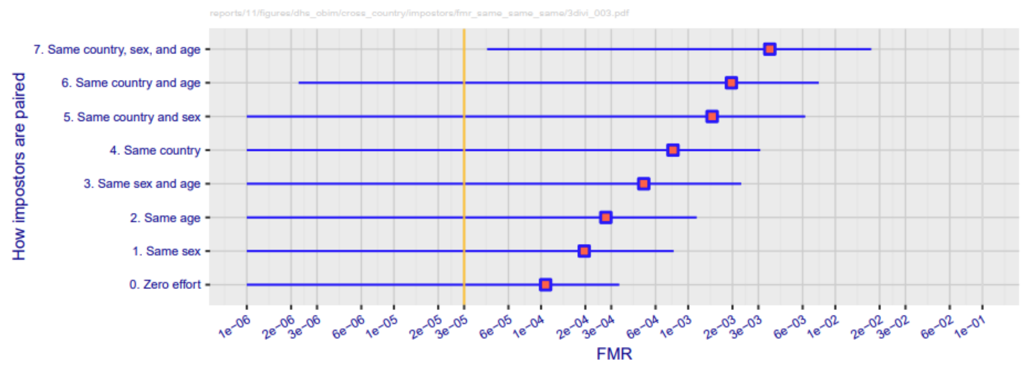


Figure 1: FMR for increasing matched covariates, 3divi-003

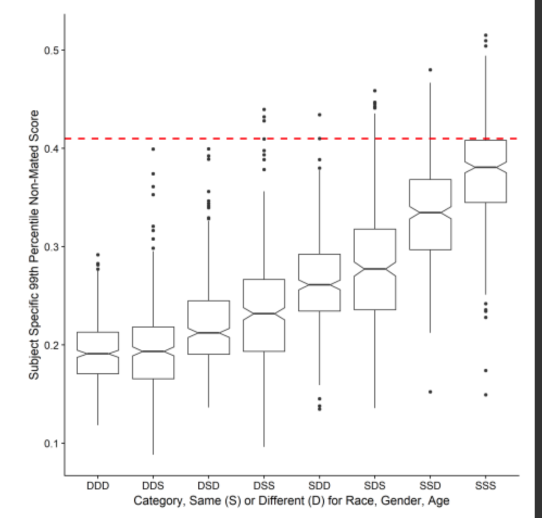
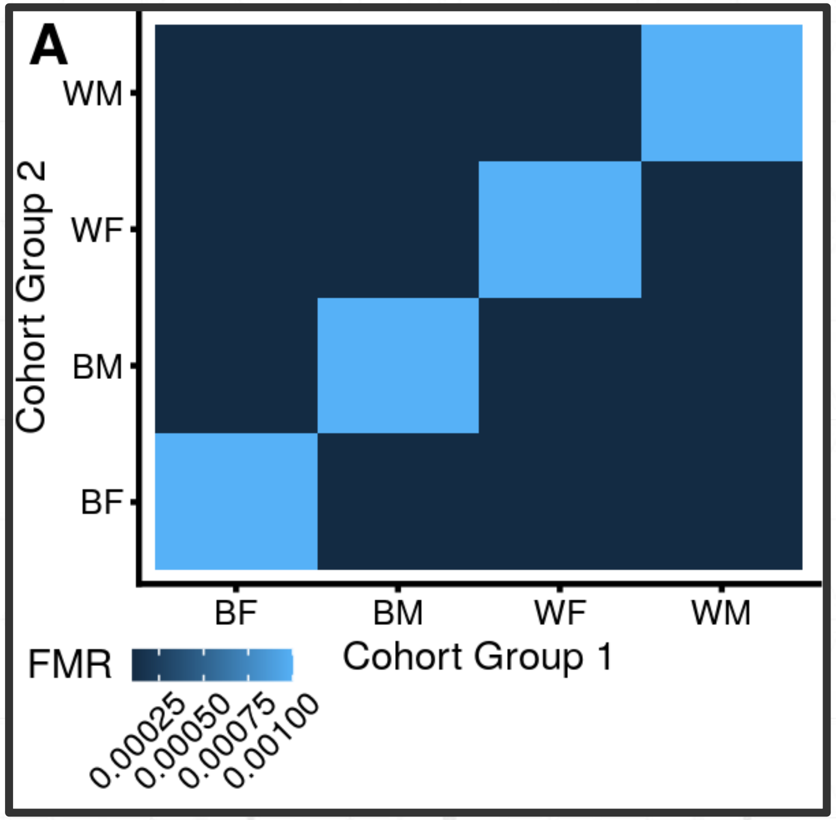


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

Science and Technology

But must it be so?

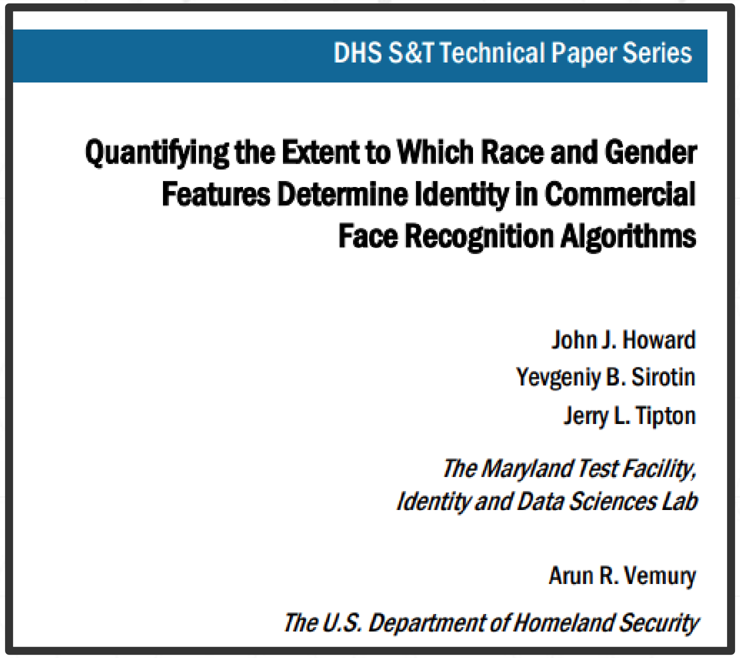
- We need to overcome our human intuition to evaluate face recognition artificial intelligence (AI) objectively.
- Is this the goal?



But must it be so?

- We need to overcome our human intuition to evaluate face recognition artificial intelligence (AI) objectively.

- Or is this the goal?



DHS S&T Technical Paper Series

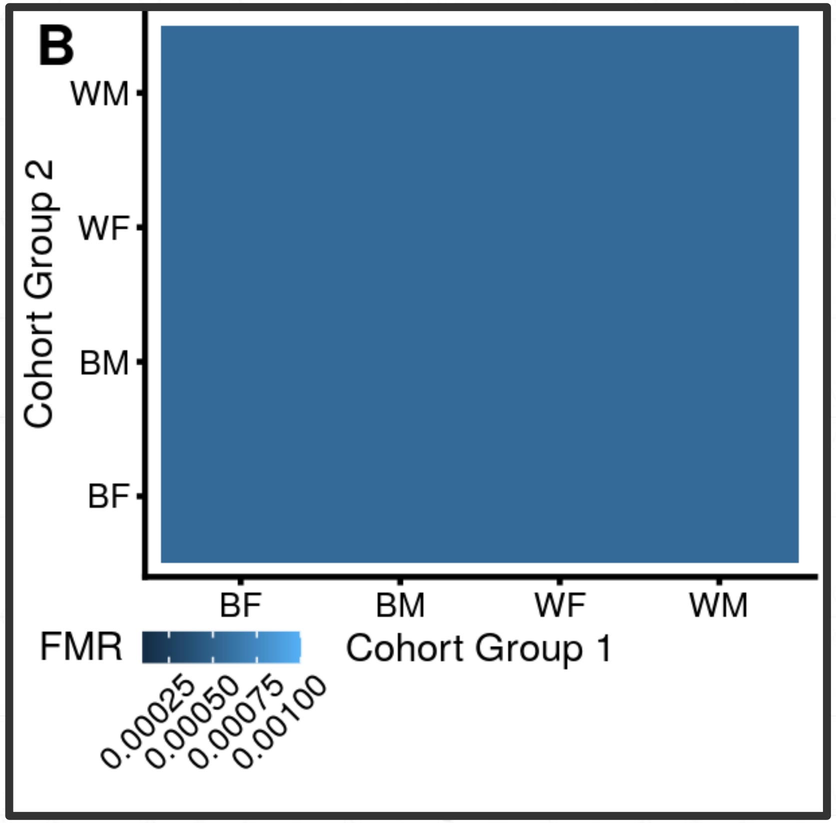
Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms

John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility, Identity and Data Sciences Lab

Arun R. Vemury

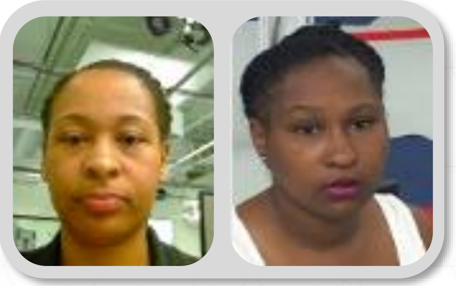
The U.S. Department of Homeland Security

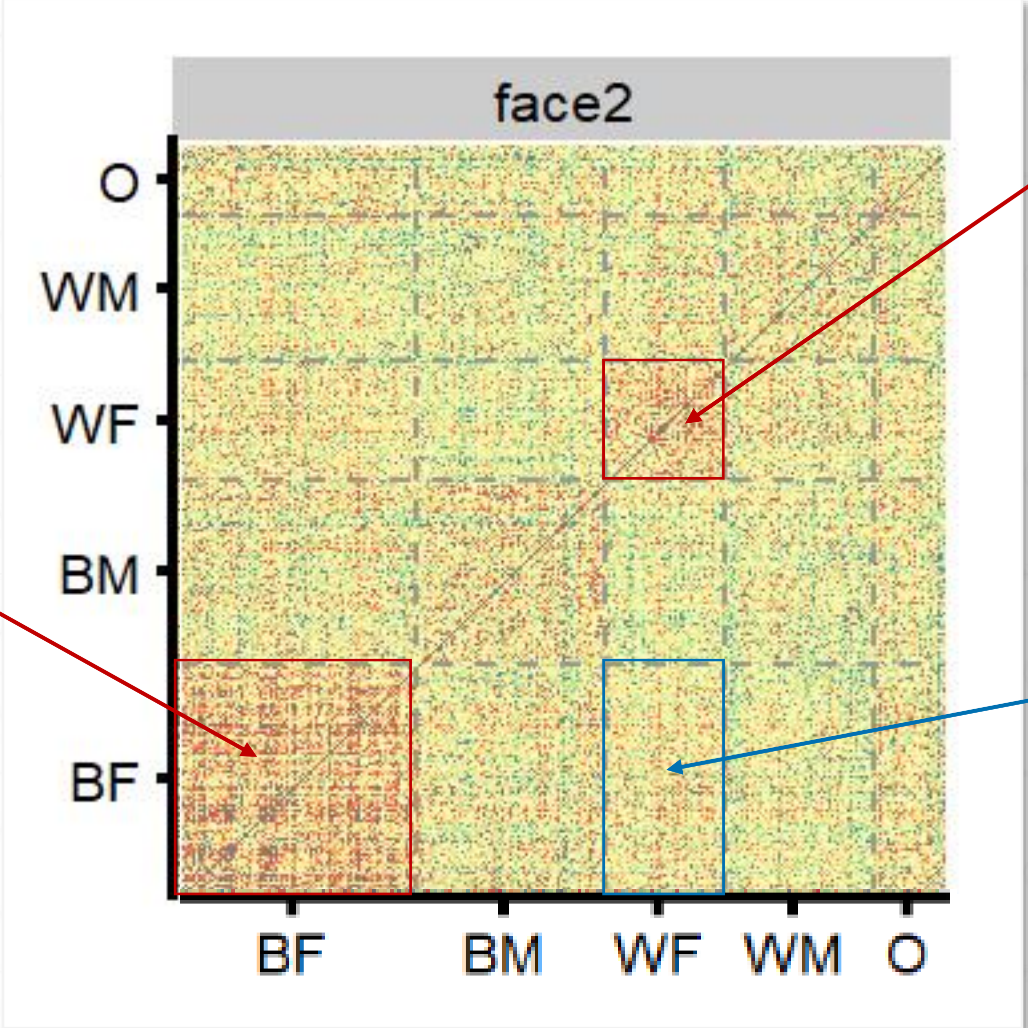


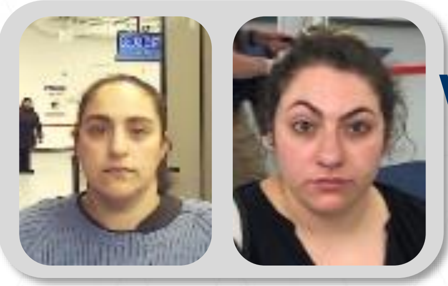
Science and Technology

A new way to think about face similarity

BF to BF

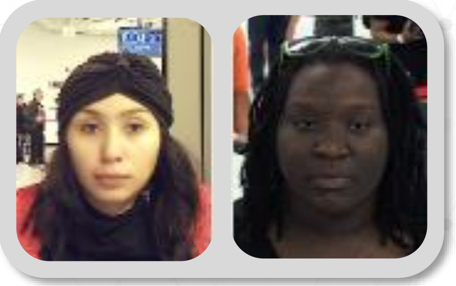






WF to WF

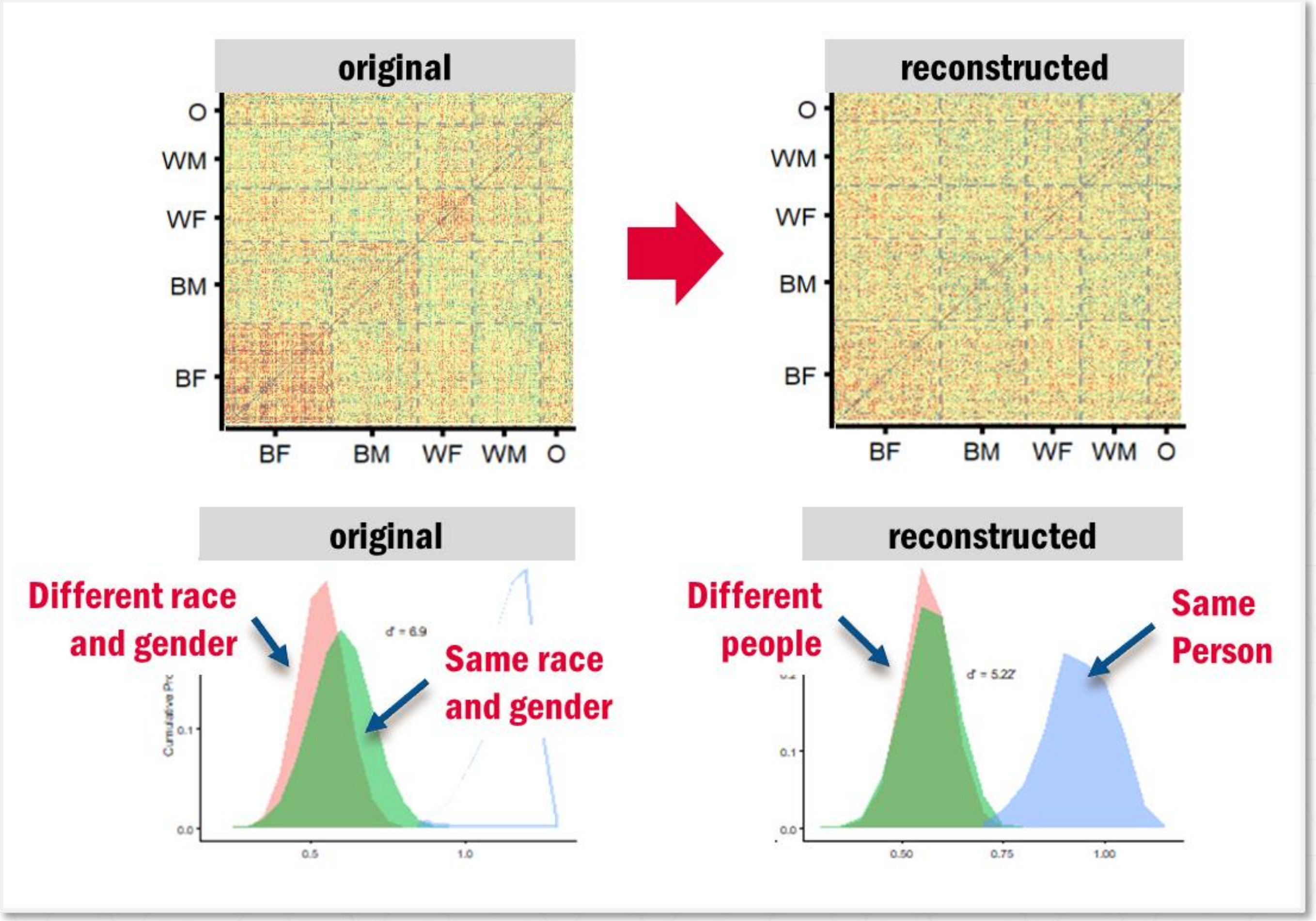
WF to BF



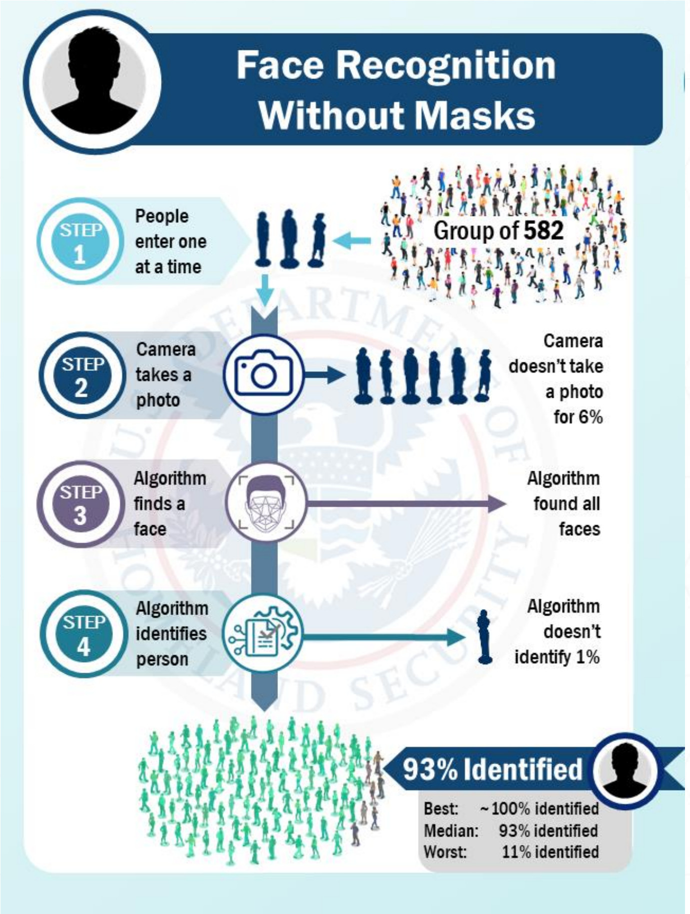
Science and Technology

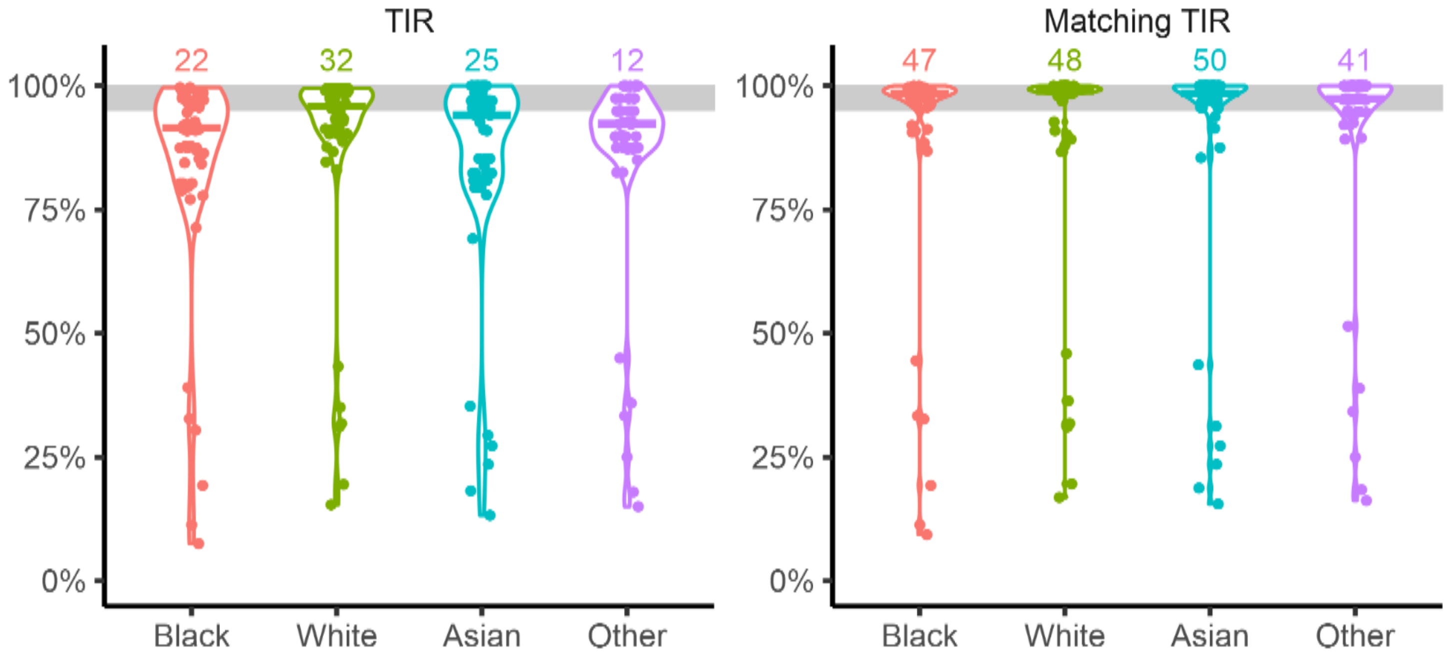
Can face recognition work without relying on race and gender?

- Mathematically removed similarity score variation related to race and gender
- Race and gender clustering was removed but individual distinction remained
- Face recognition will likely be useful even without using race and gender



Scenario Testing: Lesson 3, Changes on the Ground Can Reveal Demographic Effects.

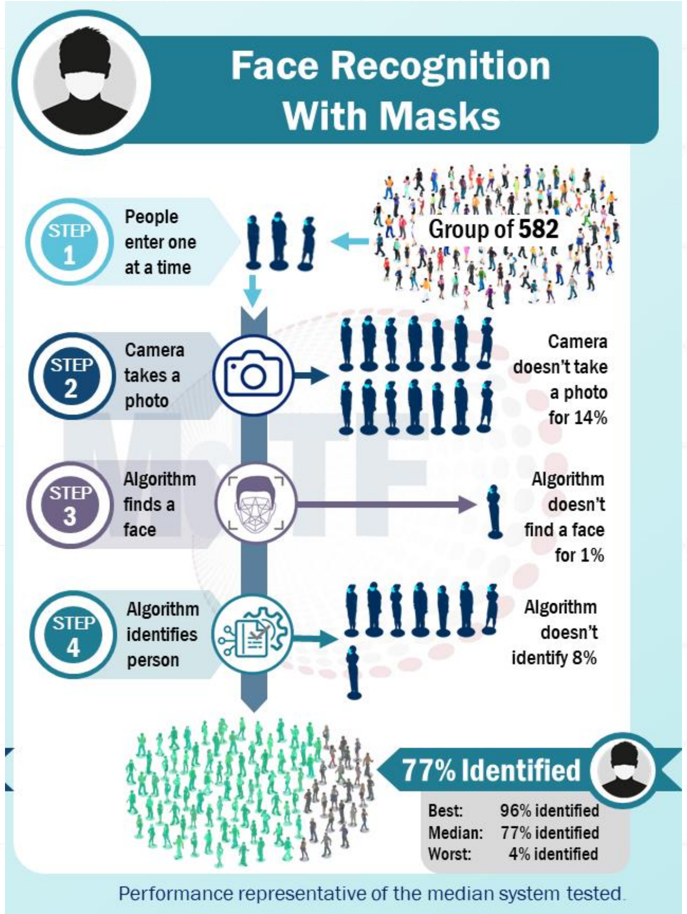


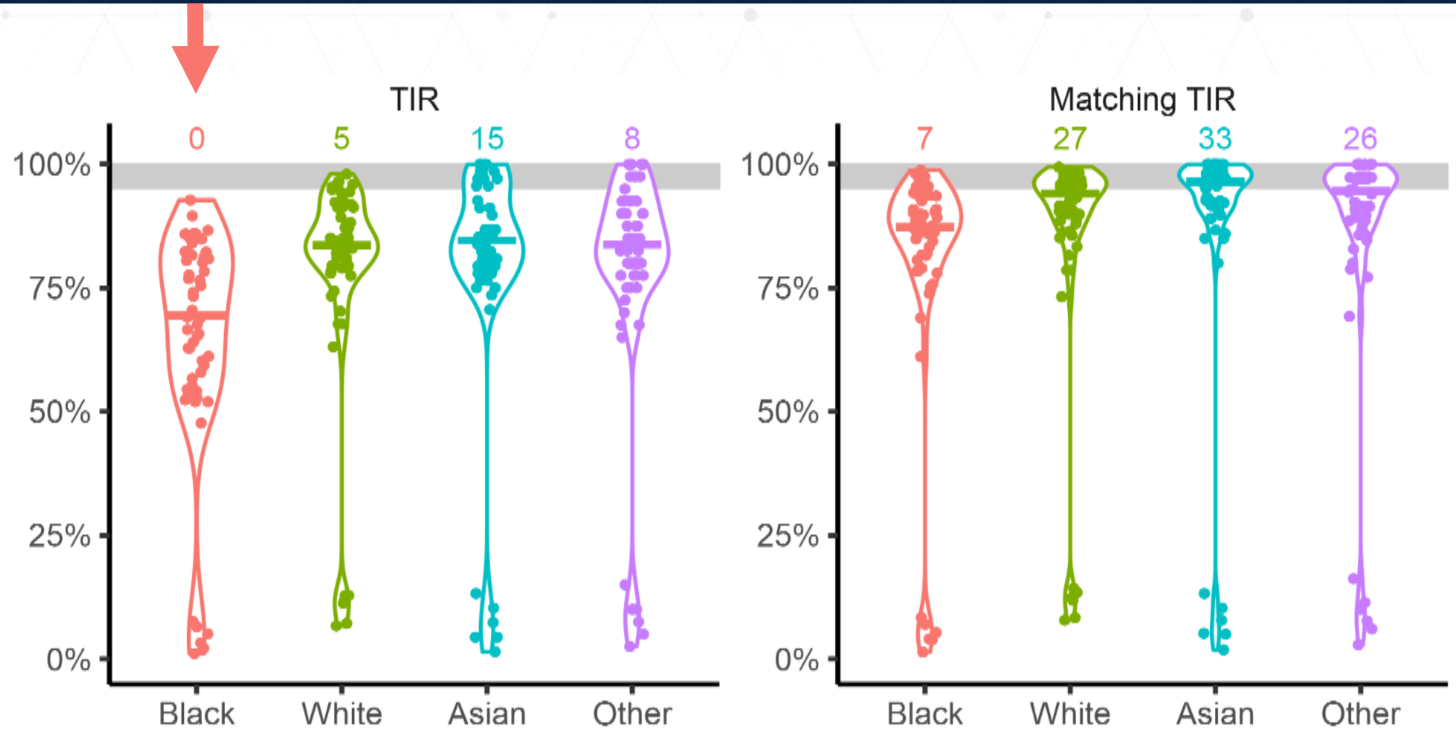


Each point in the graph represents the true identification rate (TIR) of a combination of an acquisition and matching system (n = 60) across our sample of 582 volunteers.

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

Scenario Testing: Lesson 3, Changes on the Ground Can Reveal Demographic Effects.





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.

Scenario Testing, Lesson 4: Humans in the Loop are Susceptible to Influence

Control

Computer-No Mask

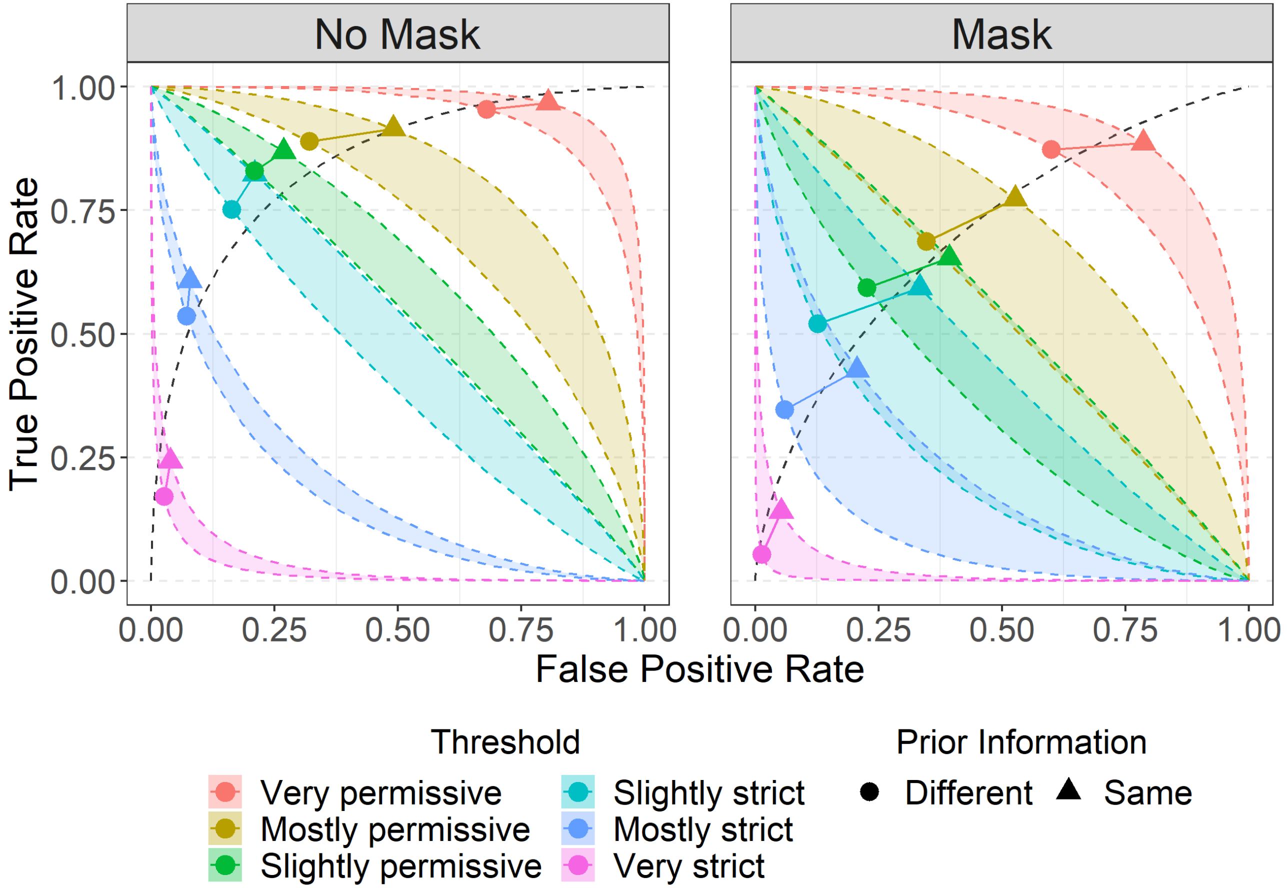
Computer-Mask

374 Untrained Human Subjects:

Similarity-Confidence Scale (Value)
I am absolutely certain this is the same person (3)
I am mostly certain this is the same person (2)
I am somewhat certain this is the same person (1)
I am not sure (0)
I am somewhat certain these are different people (-1)
I am mostly certain these are different people (-2)
I am absolutely certain these are different people (-3)

Science and Technology

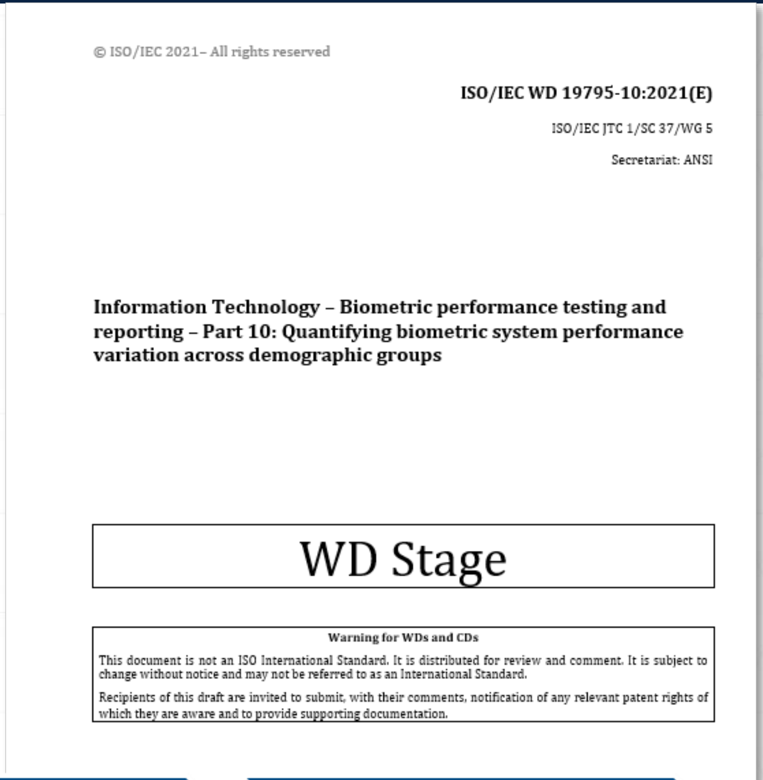
Scenario Testing, Lesson 4: Humans in the Loop are Susceptible to Influence



- Telling a human "same or different" influenced their thinking
- Masks increased this influence
- Sensitivity (d ´) lower in mask condition – more difficulty distinguishing face pairs in presence of mask
- Criterion (c) higher in mask condition – face masks increase cognitive bias and the impact of algorithms on face matching

Scenario Testing, Lesson 5: Need to Standardize How We Measure and Talk About Equitability

- Quantifying biometric system performance across demographic groups
- New work item, approved in 2020
- First draft summer 2021
- Anticipated publication in 2023 - 2024


© ISO/IEC 2021– All rights reserved

ISO/IEC WD 19795-10:2021(E)

ISO/IEC JTC 1/SC 37/WG 5

Secretariat: ANSI

Information Technology – Biometric performance testing and reporting – Part 10: Quantifying biometric system performance variation across demographic groups

WD Stage

Warning for WDs and CDs

This document is not an ISO International Standard. It is distributed for review and comment. It is subject to change without notice and may not be referred to as an International Standard.

Recipients of this draft are invited to submit, with their comments, notification of any relevant patent rights of which they are aware and to provide supporting documentation.

Scenario Testing, Lesson 5: Need to Standardize How We Measure and Talk About Equitability

- Definitions:
  - False positive differential performance – “difference in false positive error rates calculated within multiple demographic groups”
    - If Group A’s false match rate is 1%, and Group B’s false match rate is 3%
- Metrics:
  - Variation from the Mean:

$$A(\tau) = \frac{max_{d_i}(FMR_{d_i}(\tau))}{\overline{FMR}(\tau)} \; \forall d_i \in D¶$$

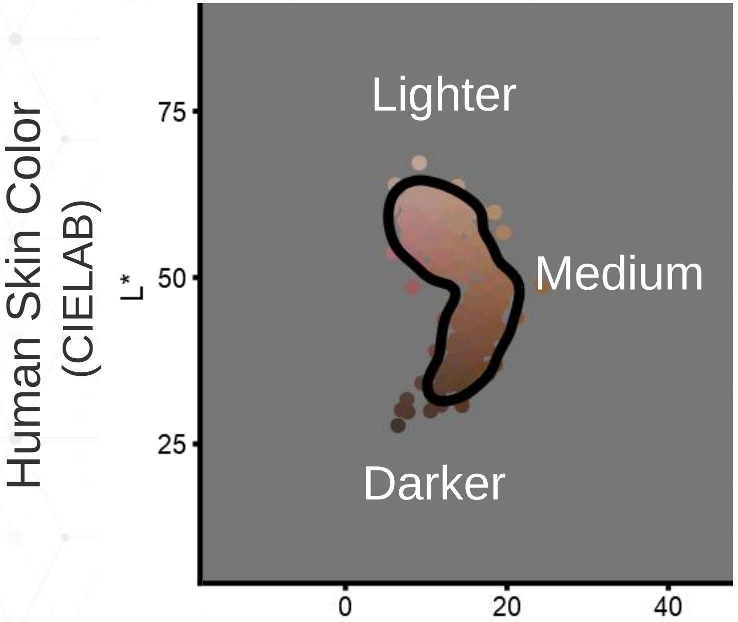
$$B(\tau) = \frac{max_{d_i}(FNMR_{d_i}(\tau))}{\overline{FNMR}(\tau)} \; \forall d_i \in D¶$$

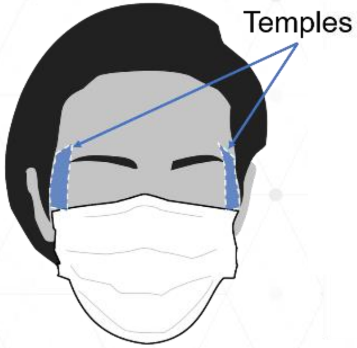
  - Gini Coefficient:

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\bar{x}}\right) \forall d_i, d_j \in D \quad (7)$$

Scenario Testing, Lesson 5: Need to Standardize How We Measure and Talk About Equitability

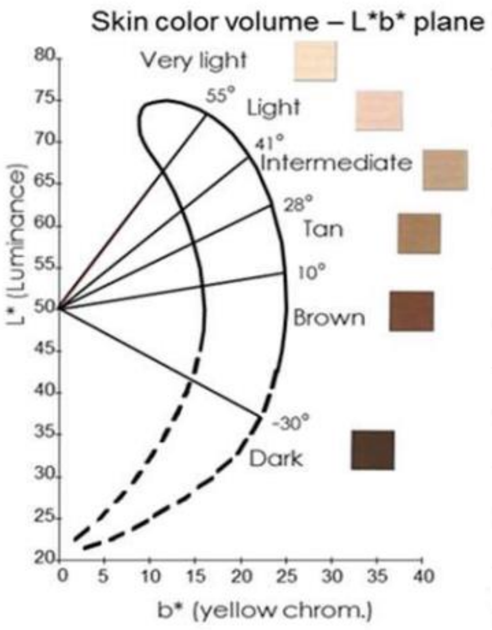
- Protocols:
  - How to collect demographics:
    - Self report – trying to infer demographic variables from the same samples used to perform biometric processing can be problematic
    - Phenotypes
      - Skin tone an important corollary for demographic performance in face recognition
      - Likely explains performance variation better than self reported race
      - Collecting this data is challenging in lab and operational environments

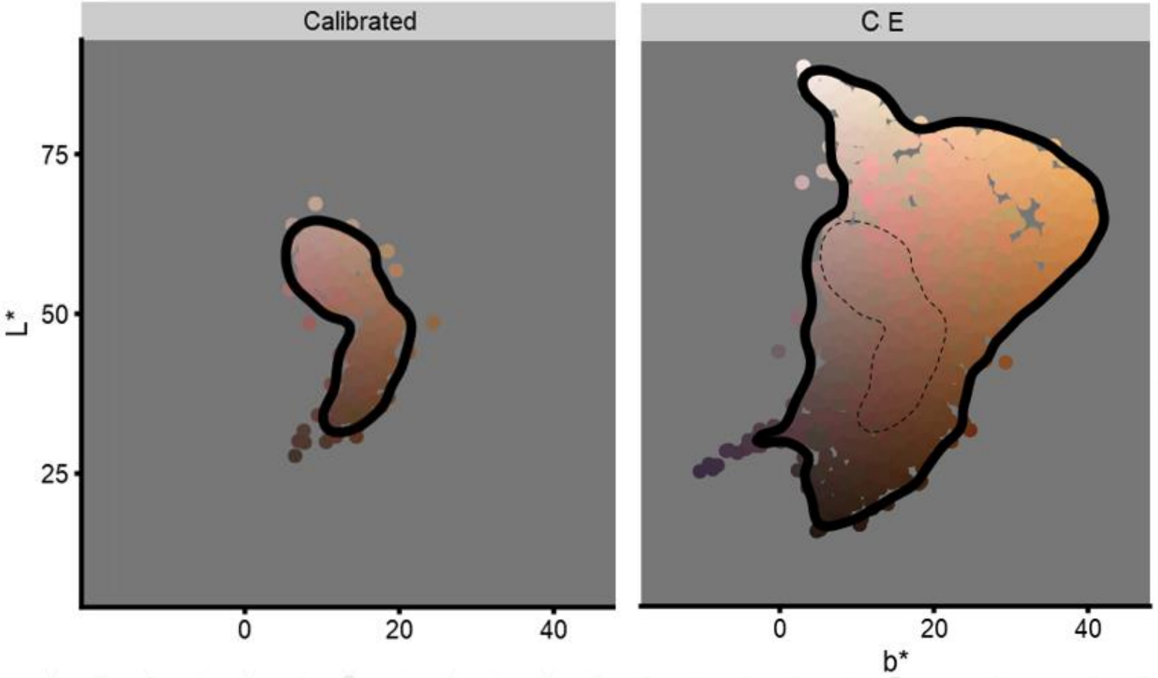




Senario Testing, Lesson 5: Need to Standardize How We Measure and Talk About Equitability

- Protocols





78-30 = 48

Science and Technology

Reporting Differences in Disaggregated Metrics

Median System:
97% TIR for Males
94% TIR for Females

Metric	Value	Pros	Cons
Difference	97%-94% = **3%**	Simple to compute and compare	Easy to mis-interpret as a percent difference
Ratio of Success Rates	94%/97% = **0.97x**	Similar to measure used by EEOC (4/5th rule)	Confusable with another success rate
Ratio of Error Rates	6%/3% = **2x**	Highlights disparities in number of individuals experiencing errors	Neglects high proportion of successful individuals in both groups
Comparison with Benchmarks	94% < 95% - **does not meet threshold** 97% > 95% - **meets threshold**	Easy to understand and trace to requirements	Does not capture magnitude of the difference

Reporting Differences in Disaggregated Metrics

Median System:
97% TIR for Males
94% TIR for Females

How much difference is too much?

Metric	Value	Pros	Cons
Difference	97%-94% = **3%**	Simple to compute and compare	Easy to mis-interpret as a percent difference
Ratio of Success Rates	94%/97% = **0.97x**	Similar to measure used by EEOC (4/5th rule)	Confusable with another success rate
Ratio of Error Rates	6%/3% = **2x**	Highlights disparities in number of individuals experiencing errors	Neglects high proportion of successful individuals in both groups
Comparison with Benchmarks	94% < 95% - **does not meet threshold** 97% > 95% - **meets threshold**	Easy to understand and trace to requirements	Does not capture magnitude of the difference

Biometric Testing and Demographics: A Key Element of Public Trust

- Growing numbers of deployments (law enforcement, border control, private)

- Increased public awareness and concerns

- Concern amongst policy-makers:
  - USS.3284 – Ethical Use of Facial Recognition Act
  - USS.4084 - Facial Recognition and Biometric Technology Moratorium Act of 2020
  - Australian Identity Matching Services Bill 2019
  - European Commission Ethics Guidelines for Trustworthy AI
  - Bridges v. South Wales Police

More information:

- This work was performed by the Identity and Data Sciences Lab, a multi-disciplinary & dedicated team of researchers at the Maryland Test Facility.

- Find out more about the DHS Biometric Technology Rallies:
  - Results at https://mdtf.org/
  - Questions: peoplescreening@hq.dhs.gov

- jhoward@idslabs.org
- arun.vemury@hq.dhs.gov