U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Face Recognition Scenario testing, performance, and fairness

Science and Technology

Arun Vemury
Lead
Biometric & Identity Technology Center

Jerry Tipton
Executive Director
The Maryland Test Facility

Yevgeniy Sirotin
Technical Director
The Maryland Test Facility

April, 2023

Technology, Scenario, and Operational Testing

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 Testing:

- Centered around a use-case,
- Full multi-component biometric system,
- Gathering new biometric samples,
- Robust experimental control.

- Answers questions about how technology performs for an intended use.
- Answers questions about the suitability of a system for an intended use.
- Answers questions regarding demographic performance that cannot be answered through operational testing (E.g. performance across race categories or skin tones)

- E.g. How will face recognition perform in a high-throughput unattended scenario?

Operational Testing:

- Centered around a specific environment,
- Specific biometric system implementation,
- New data collected in the course of operational use,
- Little experimental control.

- Answers questions about how technology performs within the specific operational environment and with specific users.
- Answers questions regarding whether the technology meets specific operational performance benchmarks.

- E.g. Is the face recognition system meeting organizational performance objectives?

Past Biometric Technology Rallies

2018 Rally assessed acquisition systems

2019 Rally assessed acquisition systems *and* matching systems

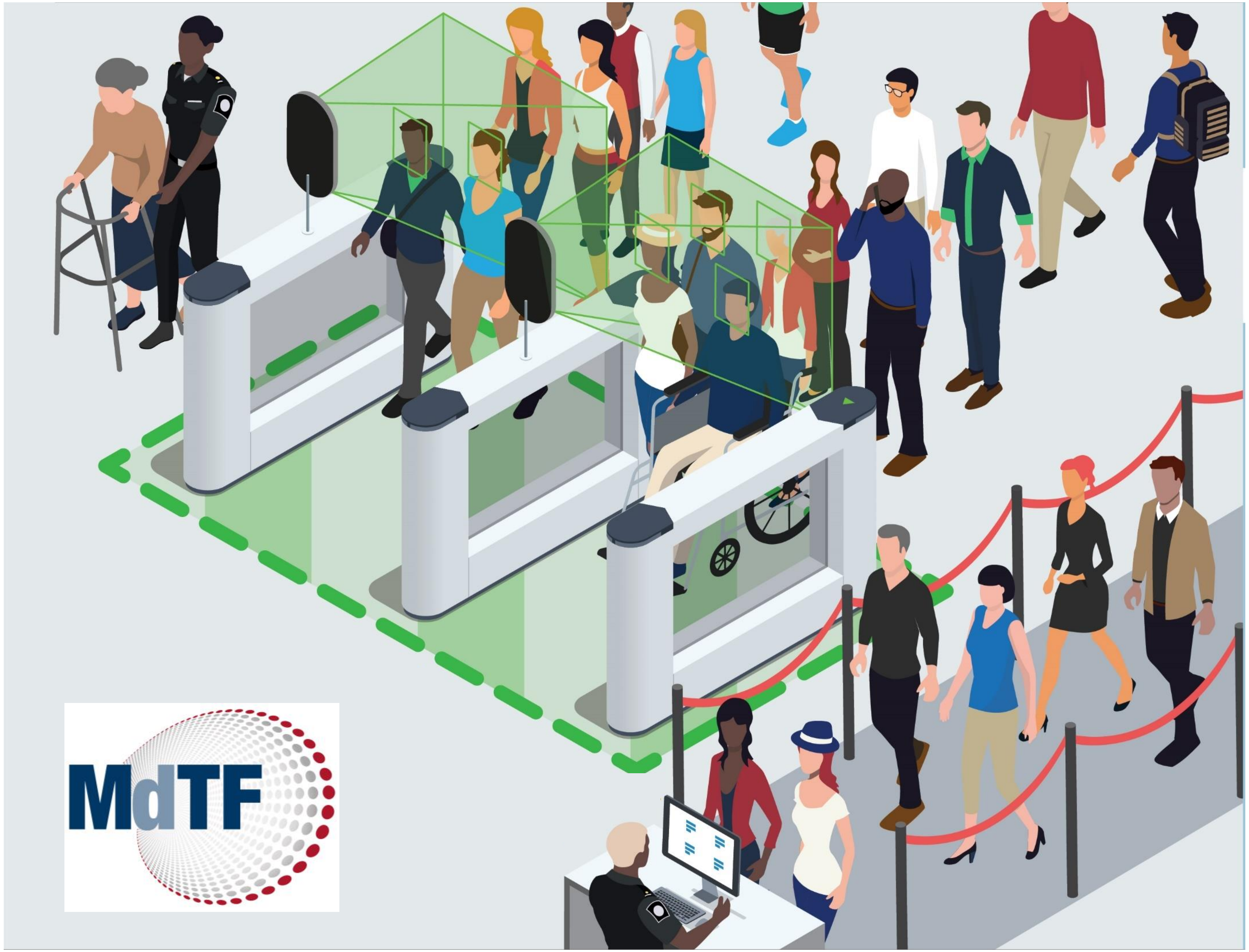
2020 Rally assessed acquisition *and* matching systems *with* face masks

2021 Rally assessed acquisition *and* matching systems *with* face masks *and* system equitability

- Since 2018, the Rallies have demonstrated progress in the performance and maturity of biometric acquisition and matching systems
  - Rally results provide insights into how people interact with biometric systems to improve usability
  - Rally results have been used to inform participating vendors, leading to improved performance of both acquisition and matching systems
  - **There are continuing challenges with respect to reliable image acquisition in the high throughput unattended use-case**

Science and Technology

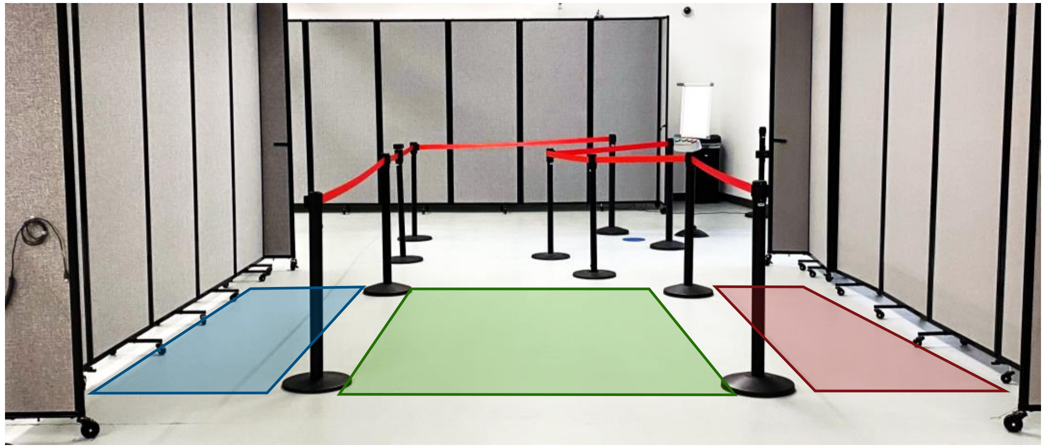
Group Processing at Checkpoints (Concept):



2022 Rally Process

Group Processing at Checkpoints (Testing):

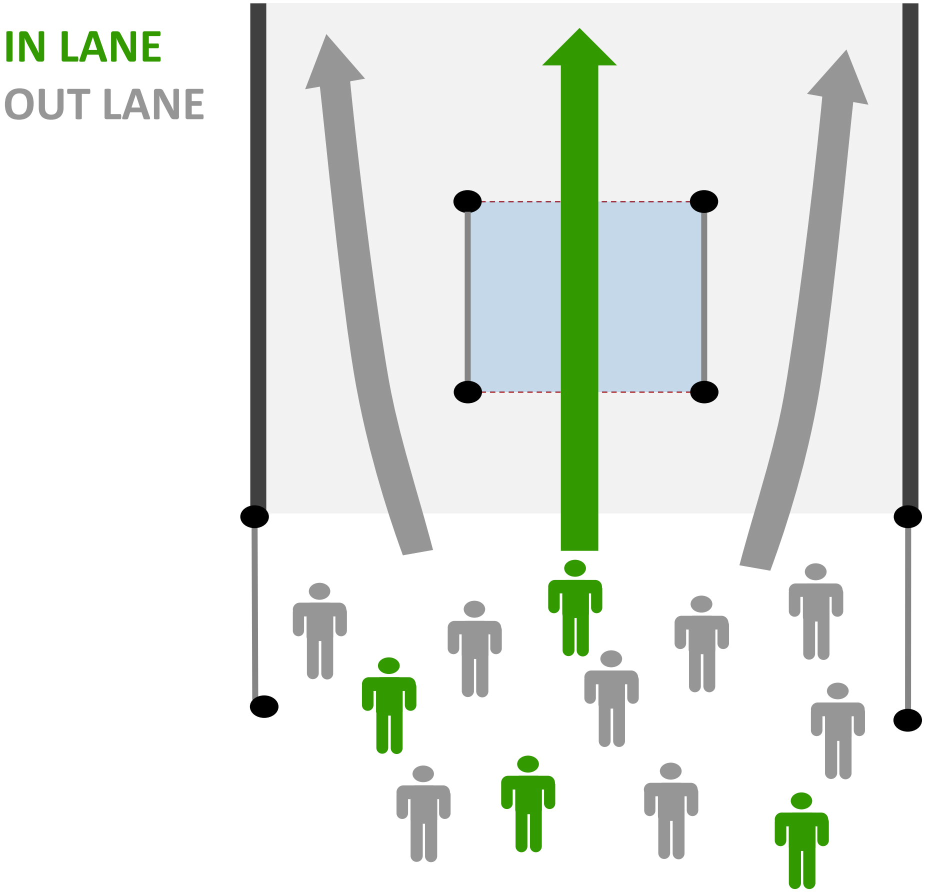
2022 Rally Station Configuration



Left Lane

Center Lane

Right Lane



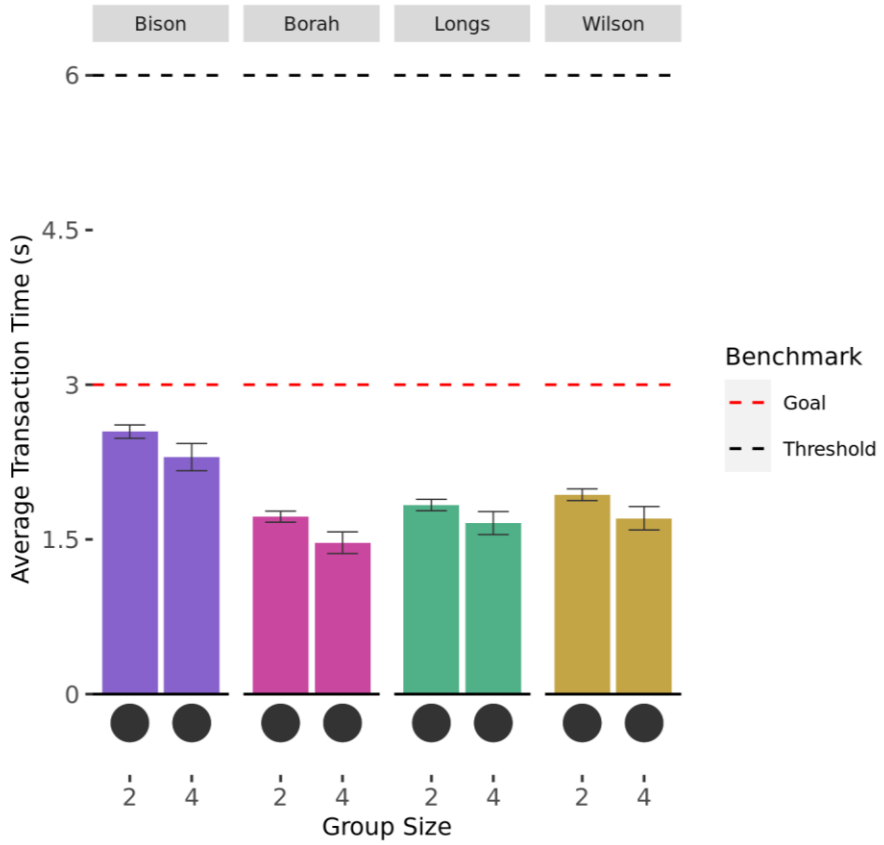
Efficiency

- All acquisition systems met the goal of 3 seconds or less and had faster per person transaction times for larger groups
- Quantified as average transaction time per group size per volunteer at each Rally Station

○ - below Rally threshold

◒ - between Rally threshold and Rally goal

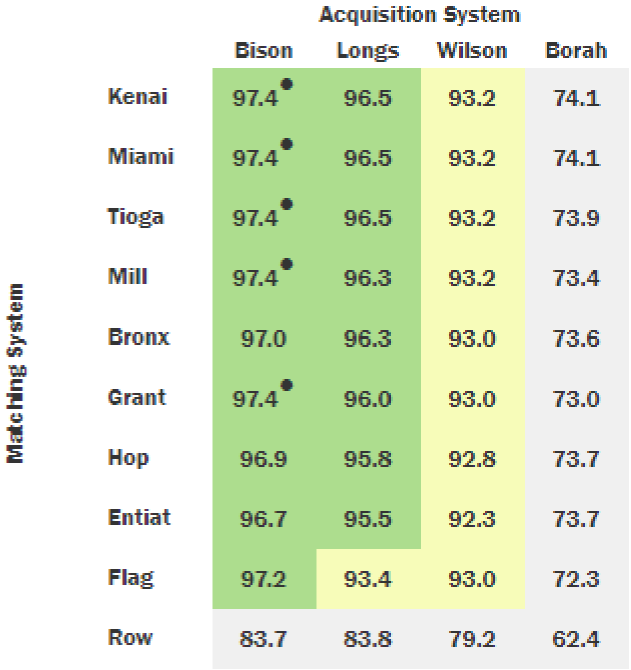
● - meets or exceeds Rally goal



- Most efficient:
  - **Borah – 1.72 seconds per person for groups of 2, 1.47 seconds per person for groups of 4**

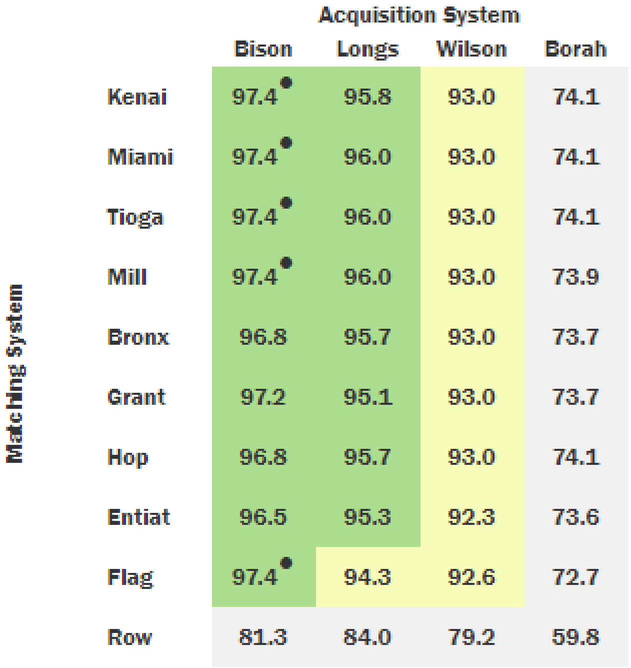
Effectiveness – Operational Focus

- TIR: True Identification Rate: quantified as the **percentage of users** who were correctly identified
- (Correct Identifications / Total People)



		Acquisition System			
		Bison	Longs	Wilson	Borah
Matching System	Kenai	97.4●	96.5	93.2	74.1
	Miami	97.4●	96.5	93.2	74.1
	Tioga	97.4●	96.5	93.2	73.9
	Mill	97.4●	96.3	93.2	73.4
	Bronx	97.0	96.3	93.0	73.6
	Grant	97.4●	96.0	93.0	73.0
	Hop	96.9	95.8	92.8	73.7
	Entiat	96.7	95.5	92.3	73.7
	Flag	97.2	93.4	93.0	72.3
	Row	83.7	83.8	79.2	62.4

Groups of 2



		Acquisition System			
		Bison	Longs	Wilson	Borah
Matching System	Kenai	97.4●	95.8	93.0	74.1
	Miami	97.4●	96.0	93.0	74.1
	Tioga	97.4●	96.0	93.0	74.1
	Mill	97.4●	96.0	93.0	73.9
	Bronx	96.8	95.7	93.0	73.7
	Grant	97.2	95.1	93.0	73.7
	Hop	96.8	95.7	93.0	74.1
	Entiat	96.5	95.3	92.3	73.6
	Flag	97.4●	94.3	92.6	72.7
	Row	81.3	84.0	79.2	59.8

Groups of 4

- Seventeen (17) system combinations **met the TIR threshold of 95%** for groups of 2 and 4
- Same system combinations **across groups of 2 and 4**
- **No system** combinations met the TIR goal of 99%

Effectiveness – Demographics

- TIR performance was disaggregated into eight demographic groups

- Gender (self-reported)
  - Male, Female

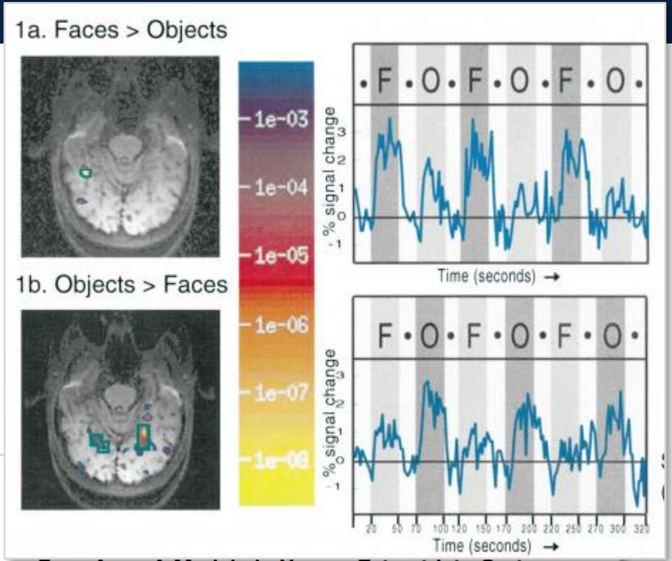
- Race (self-reported)
  - Asian, Black, White

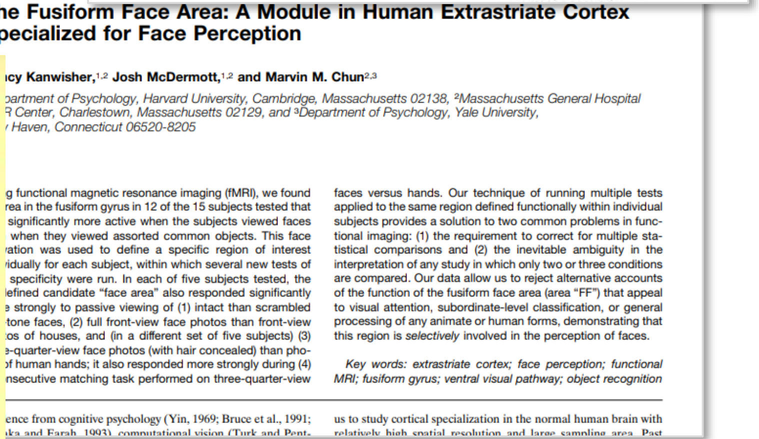
- Skin-Tone (measured)
  - Lighter, Medium, Darker

Faces are different from other biometric modalities 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

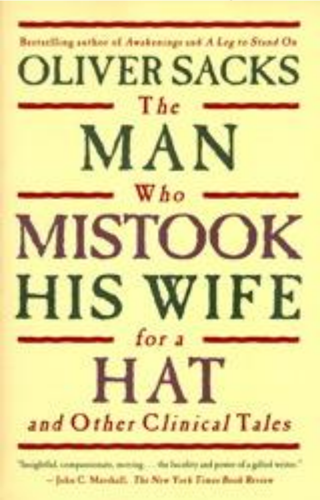
cy Kanwisher,[1,2] Josh McDermott,[1,2] and Marvin M. Chun[2,3]

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 e 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) e-quarter-view face photos (with hair concealed) than photos 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

ence from cognitive psychology (Yin, 1969; Bruce et al., 1991; ka 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


Bestselling author of *Awakenings* and *A Leg to Stand On*

OLIVER SACKS

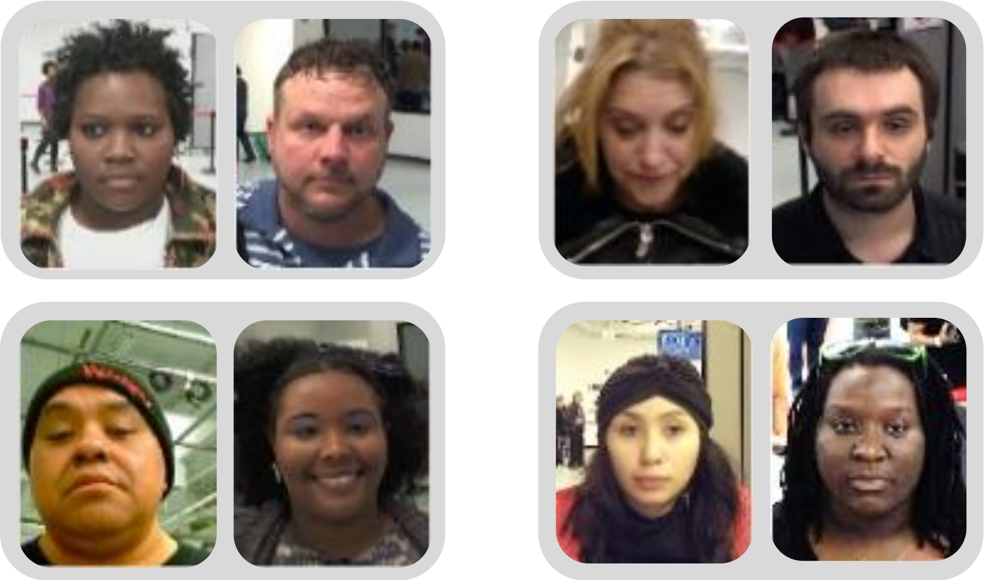
The MAN Who MISTOOK HIS WIFE for a HAT and Other Clinical Tales

"Insightful, compassionate, moving . . . the lucidity and power of a gifted writer." —John C. Marshall, *The New York Times Book Review*

Demographic Effects Exist, Our Understanding of Them may be Clouded.

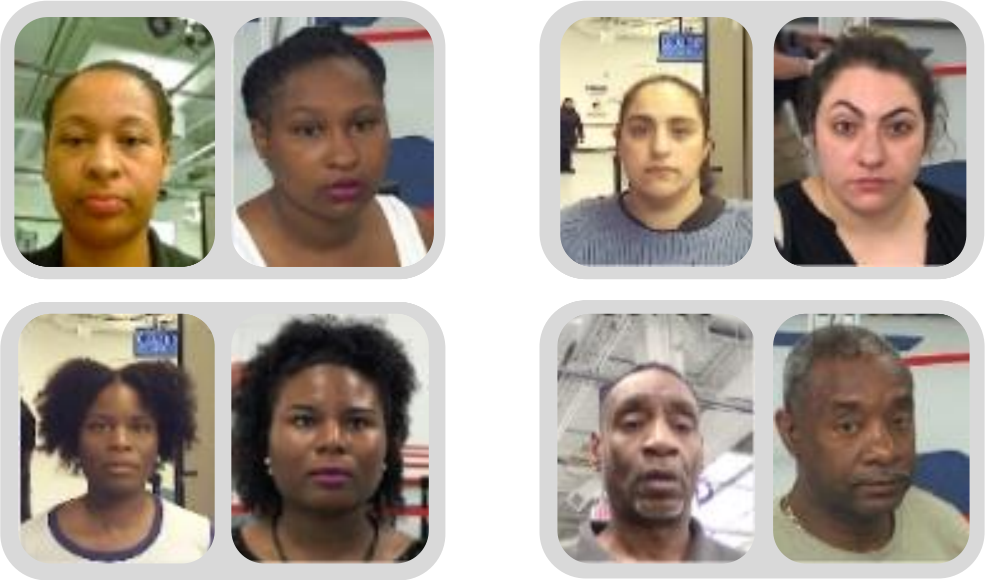
> It may seem natural to us that face recognition "clusters" people based on race and gender <

Iris recognition



Iris recognition false positives were random relative to race and gender

Face recognition



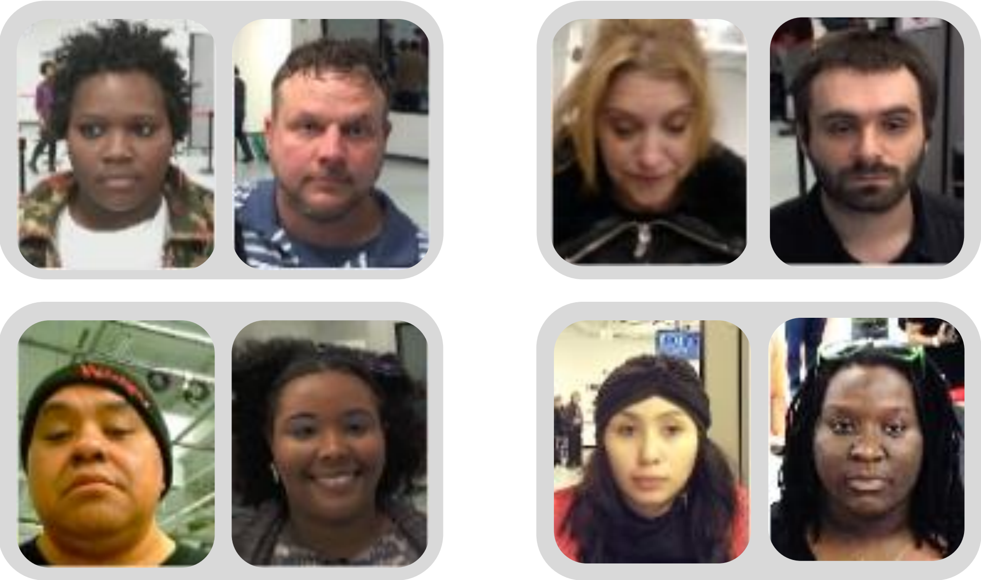
80% of face recognition false positives were between people of the same race and gender

Subjects consent for use of their image in publications was obtained

Apples and Apples or Apples and Oranges?

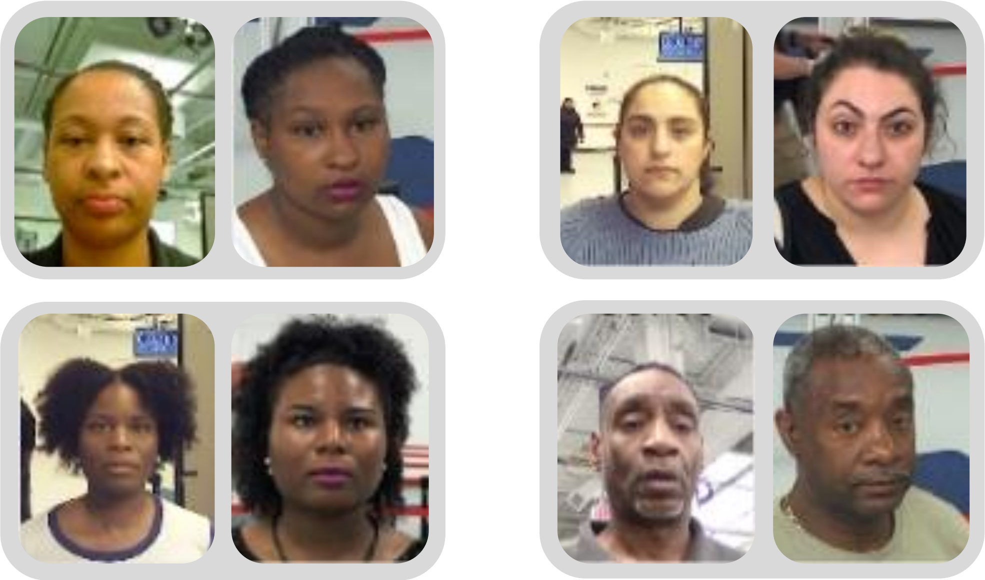
> All of these “errors” are called “false matches”, but those on the right are different than those on the left <

Iris recognition



Iris recognition false positives were random relative to race and gender

Face recognition



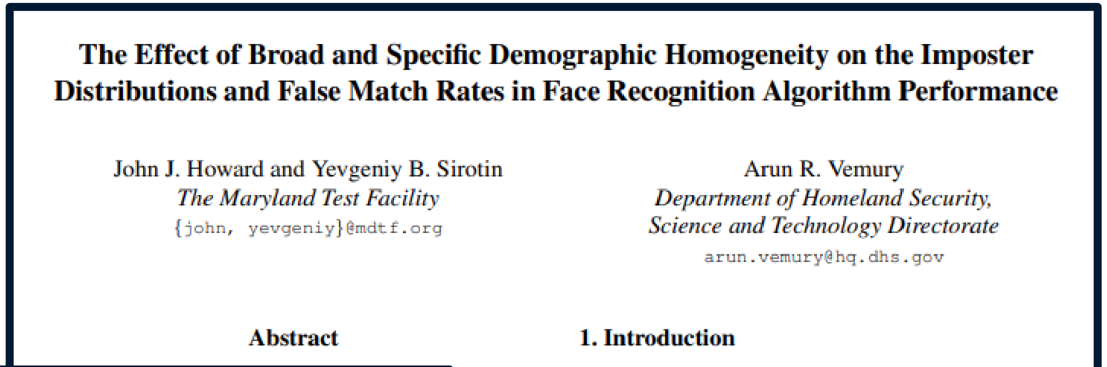
80% of face recognition false positives were between people of the same race and gender

Subjects consent for use of their image in publications was obtained

This is (likely) (currently) a Universal Feature of Face Recognition

- We first highlighted this in 2019 using one commercial algorithm

- NIST subsequently confirmed this exists in **all 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

systems, s regard-

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-

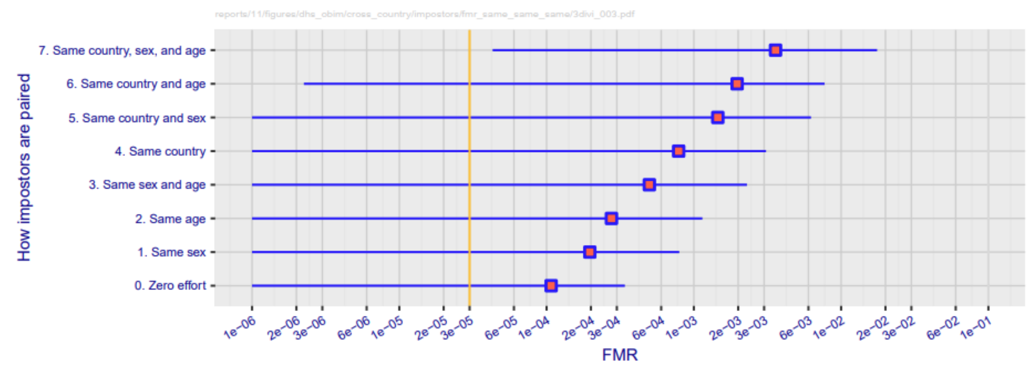


Figure 1: FMR for increasing matched covariates, 3divi-003

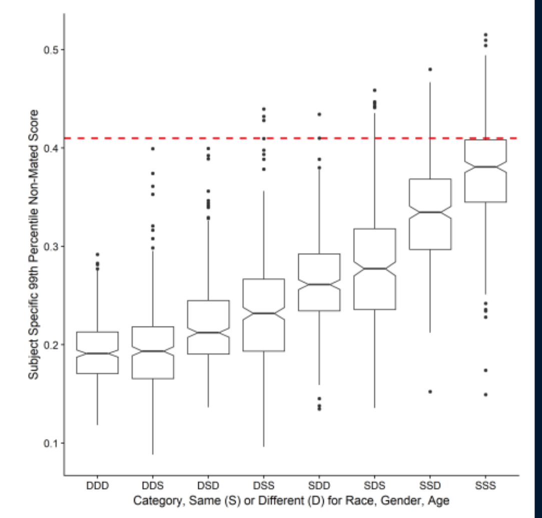
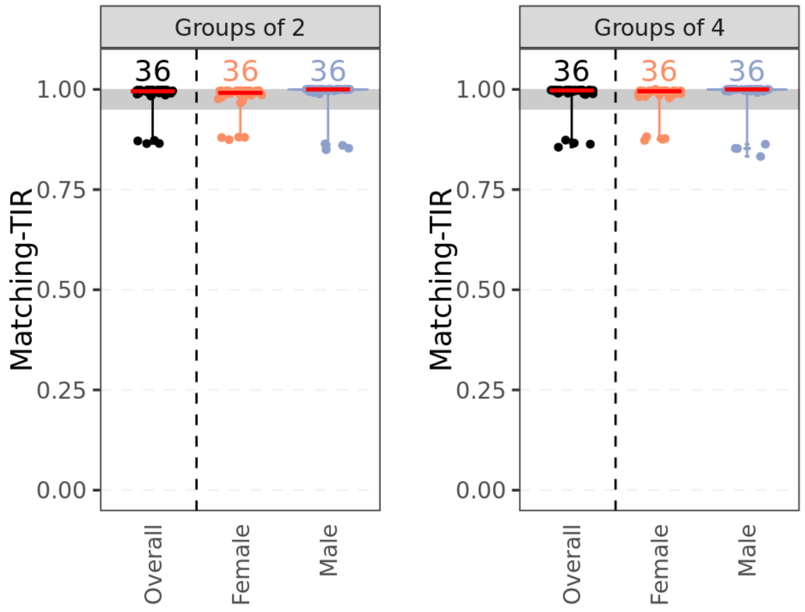
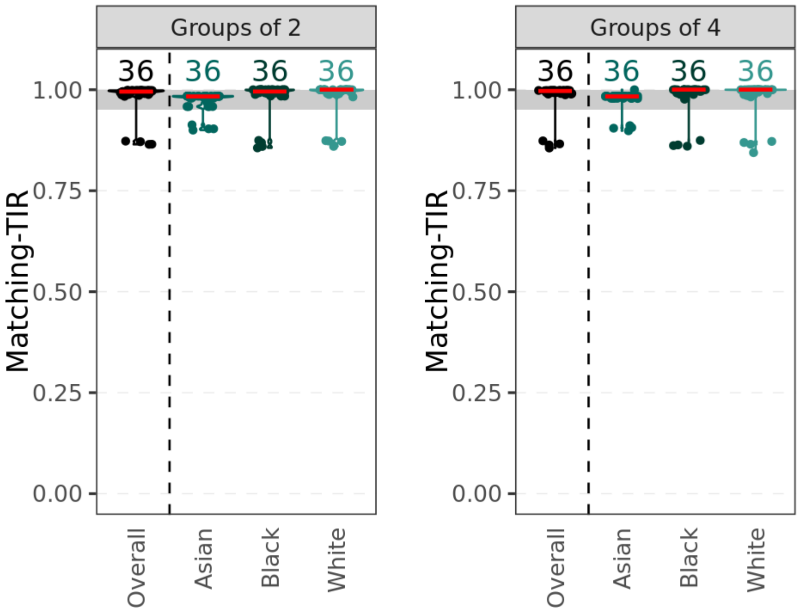


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

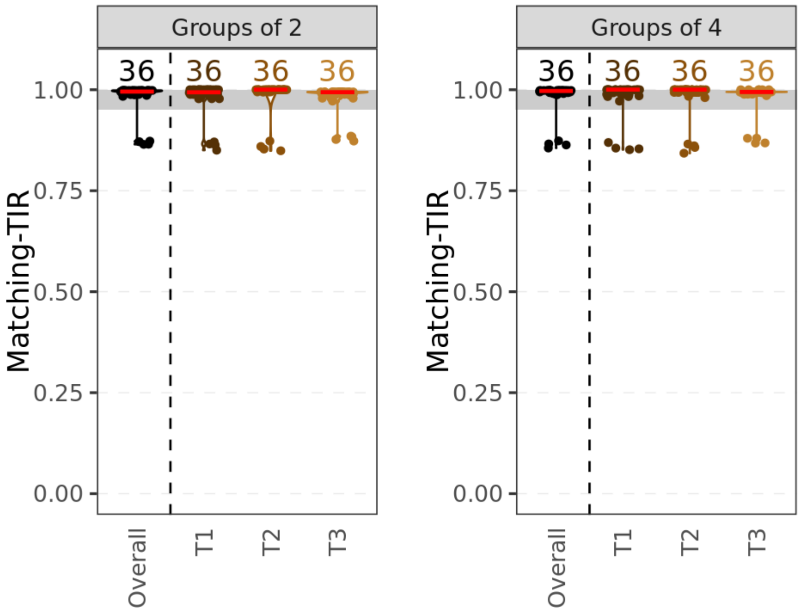
Matching Focus Demographic Differentials



Gender



Race



Skin Tone

- When discounting failures to submit images of suitable quality, **most system combinations** were able to meet the 95% Rally matching-TIR threshold

Science and Technology

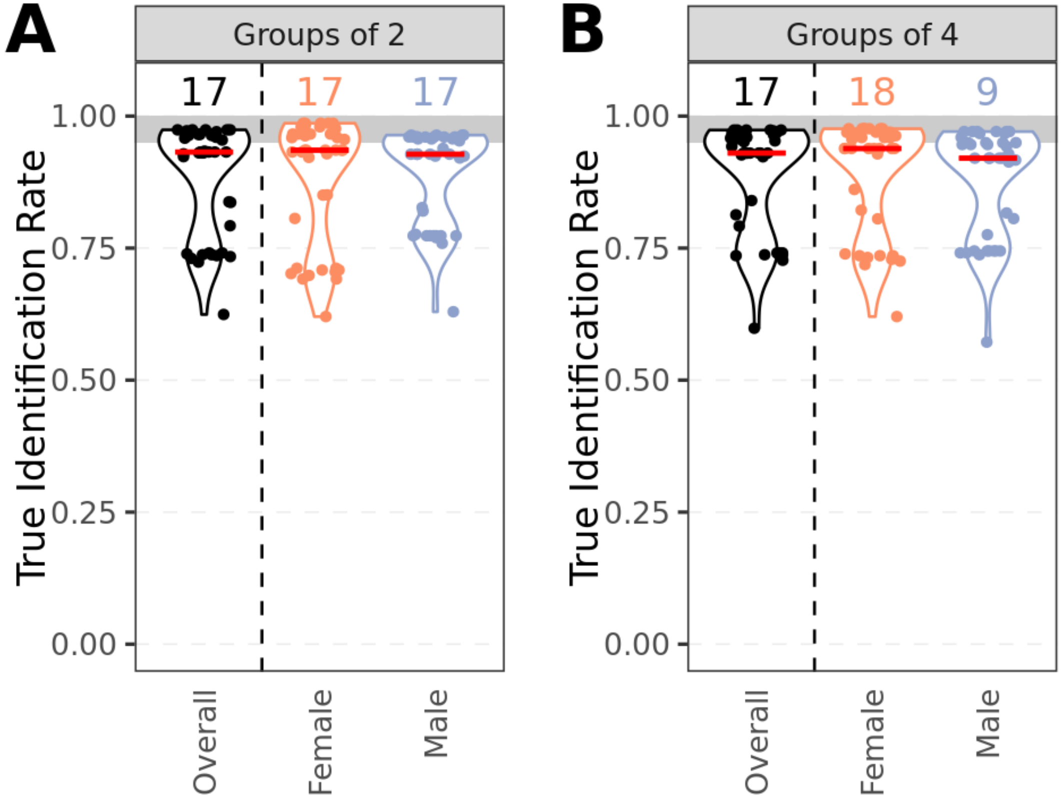
Operational Focus Demographic Differentials

- Some system combinations were able to meet the 95% Rally TIR threshold for all demographic group

- However, considering acquisition some demographic differentials remained

- Median system performance was:
  - Lower for “Male” relative to “Female” volunteers **(gender differential)**

Group Size	Female		Male
2	93.5%	↘	92.8%
4	93.9%	↘	92.0%



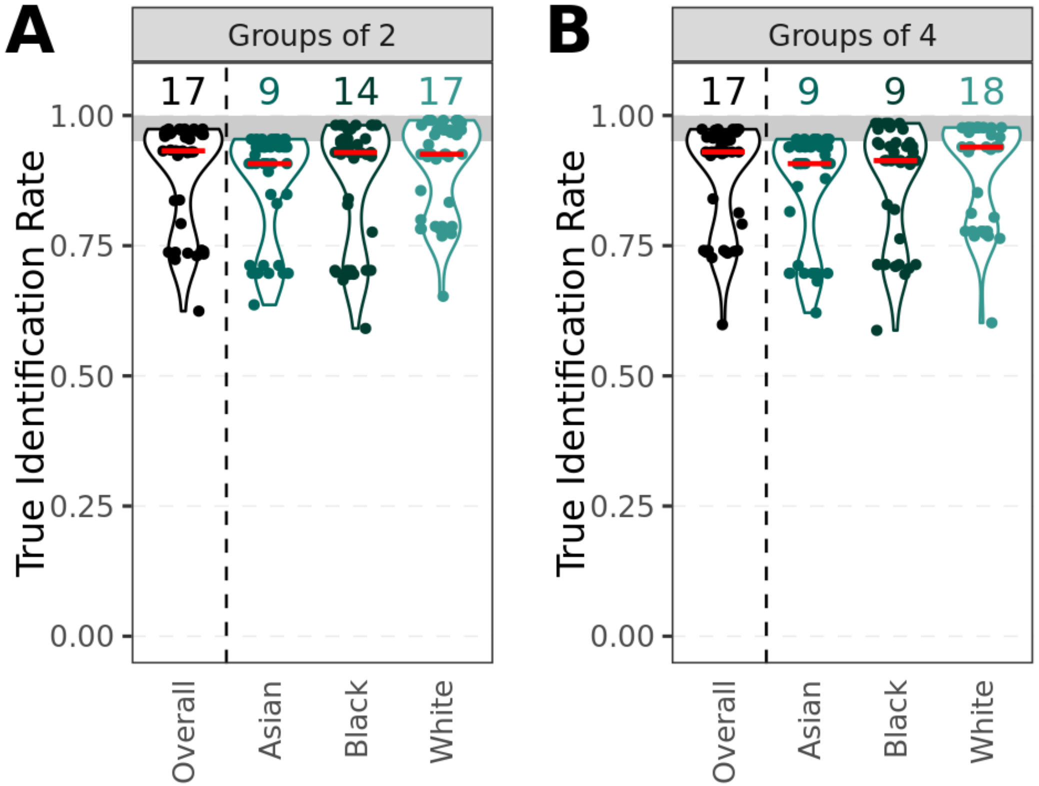
Operational Focus Demographic Differentials

- Some system combinations were able to meet the 95% Rally TIR threshold for all demographic groups

- However, considering acquisition some demographic differentials remained

- Median system performance was:
  - Lower for volunteers that self-identified as “Asian” **(race differential)**

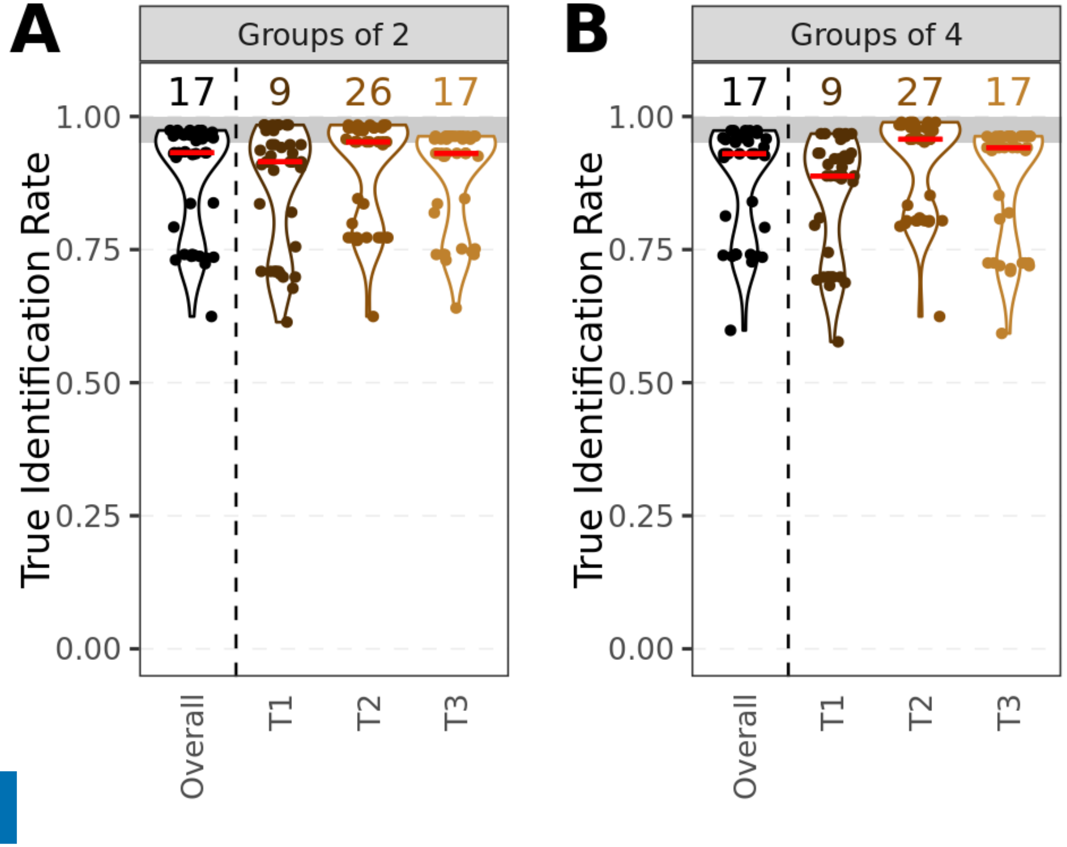
Group Size	Black	White		Asian
2	92.9%	92.5%	↘	90.8%
4	91.3%	93.9%	↘	90.8%



Operational Focus Demographic Differentials

- Some system combinations were able to meet the 95% Rally TIR threshold for all demographic groups
- However, considering acquisition some demographic differentials remained
- Median system performance was:
  - Lower for volunteers with very dark skin tone and very light skin tone **(skin tone differential)**

Group Size	Light Skin Tone		Dark Skin Tone
2	93.1%	➘	91.4%
4	94.1%	➘	88.8%



Demographic Summary

- When discounting failures to submit images of suitable quality, **most system combinations** were able to meet the 99% Rally match-TIR goal for all demographic groups

- Including failure to capture, **some system combinations** were able to meet the 95% Rally TIR threshold for all demographic groups

- Including failure to capture, demographic differentials in the number of systems able to achieve the 95% Rally TIR threshold were observed:
  - Lower for “Male” relative to “Female” volunteers
  - Lower for volunteers that self-identified as “Asian”
  - Lower for volunteers with darker skin tone

Interactive Results Available at mdtf.org

- The data presented today is available for review and exploration at https://mdtf.org

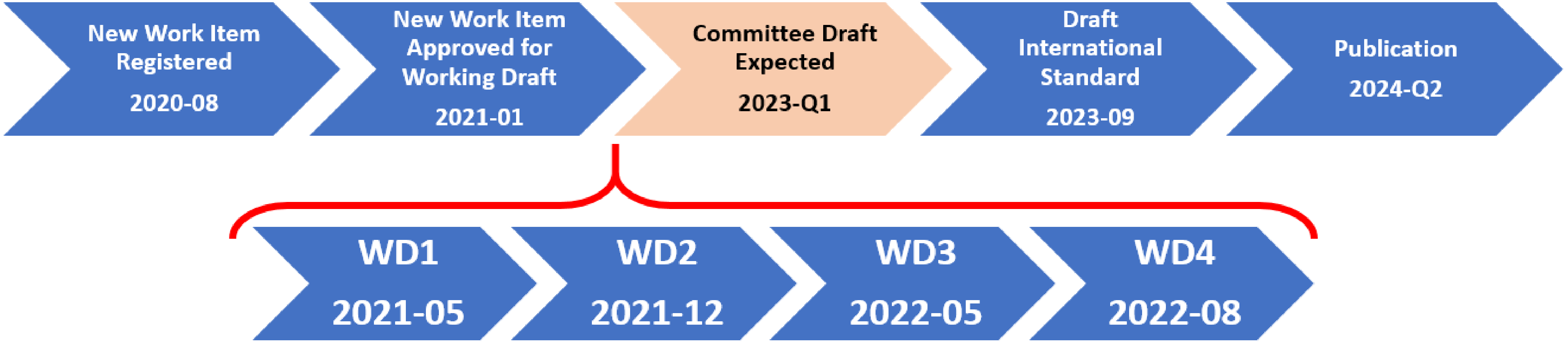
- Interactive visualization of demographically disaggregated performance

- Downloadable PDF report with detailed performance metrics for each tested system

PLACEHOLDER: Video showing interactions with website infographics

ISO/IEC 19795-10: Demographic Differentials

- DHS S&T is supporting development of standard methods of measuring demographic differentials:
  - ISO/IEC 19795-10 WD4 – Biometric performance across demographic groups
  - How to define demographic groups, including skin-tone
  - How to plan and perform an assessment of demographic differentials
  - How to calculate & report error rates across groups



New Work Item Registered

2020-08

New Work Item Approved for Working Draft

2021-01

Committee Draft Expected

2023-Q1

Draft International Standard

2023-09

Publication

2024-Q2

WD1

2021-05

WD2

2021-12

WD3

2022-05

WD4

2022-08

Questions & Answers

- Contact information
  - peoplescreening@hq.dhs.gov
  - rally@mdtf.org

- Visit our websites for additional information
  - To see additional work DHS S&T supports, visit www.dhs.gov/science-and-technology
  - To view additional information about this year and prior Rallies, visit https://mdtf.org

Science and Technology