U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Feature Vector Clustering – A Step Toward Fixing Broad Homogeneity Effects

Science and Technology

John J. Howard & Yevgeniy B. Sirotin
Identity and Data Sciences Laboratory at
the Maryland Test Facility

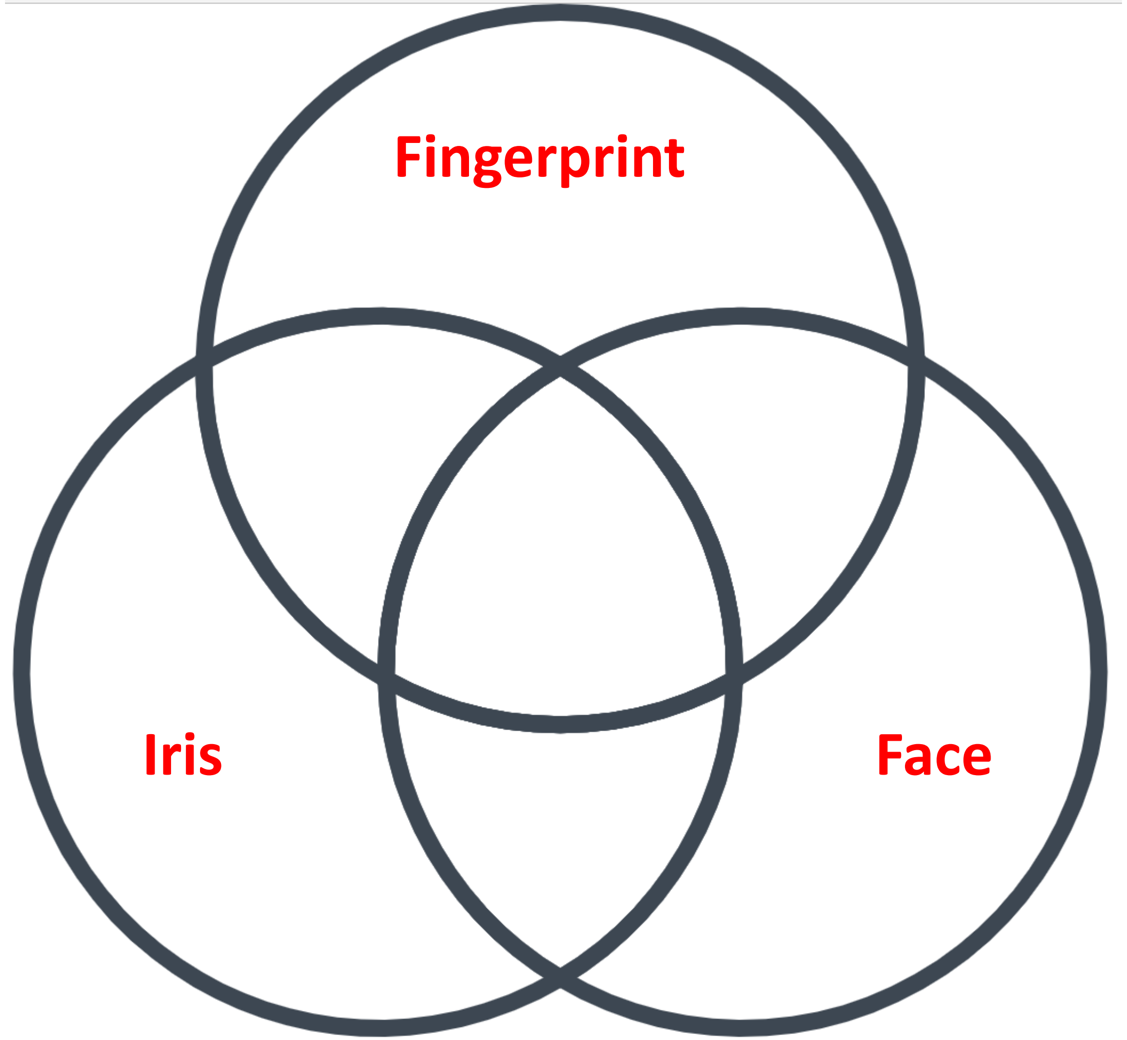
Arun Vemury
Lead
Biometric and Identity Technology Center
DHS Science & Technology Directorate

November 2022

Disclaimer

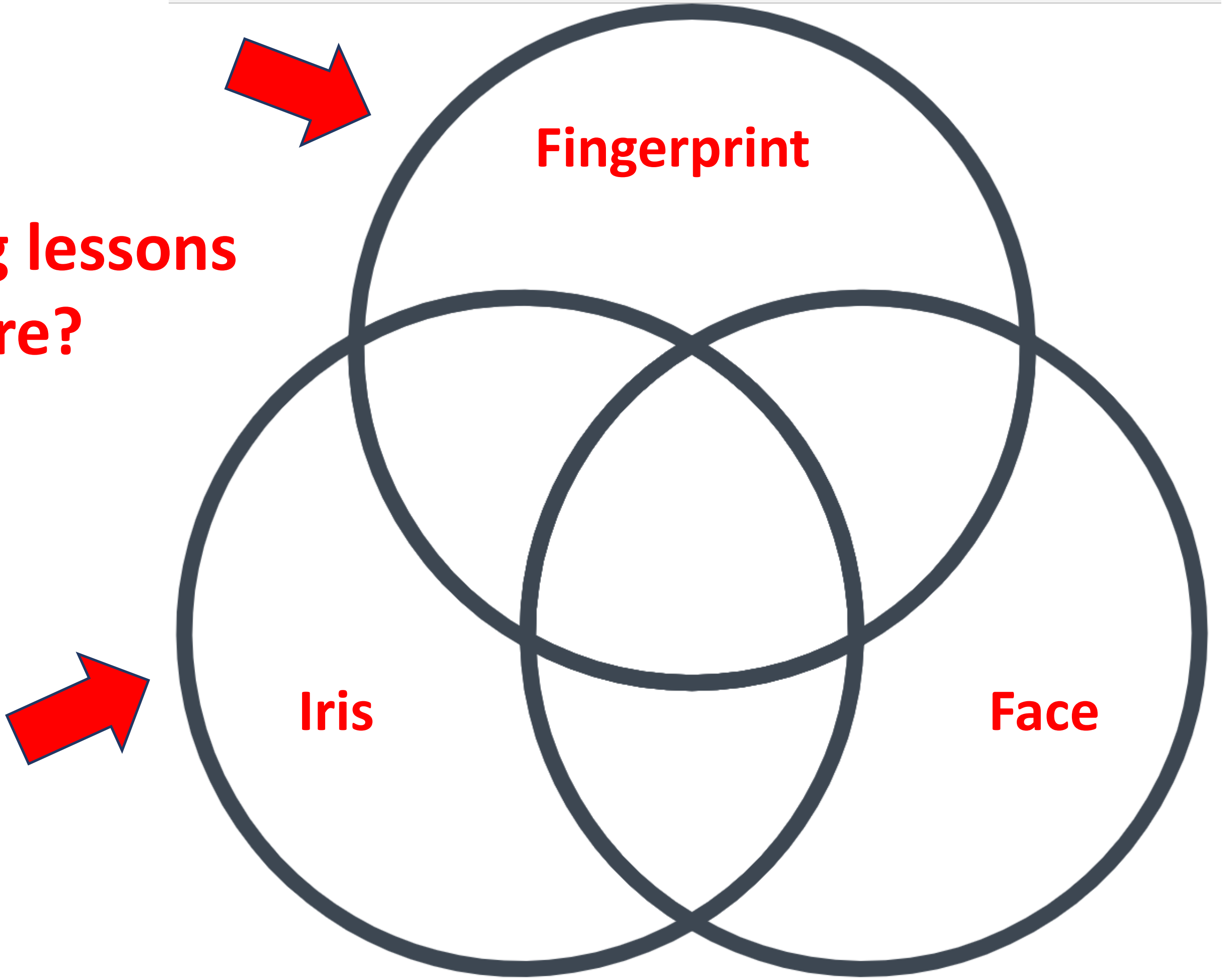
- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.
- This work was performed by the Identity and Data Sciences Laboratory team at the Maryland Test Facility.
- The views presented here are those of the authors 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 or is publicly available non-PII data.

The Third Wave of Biometrics



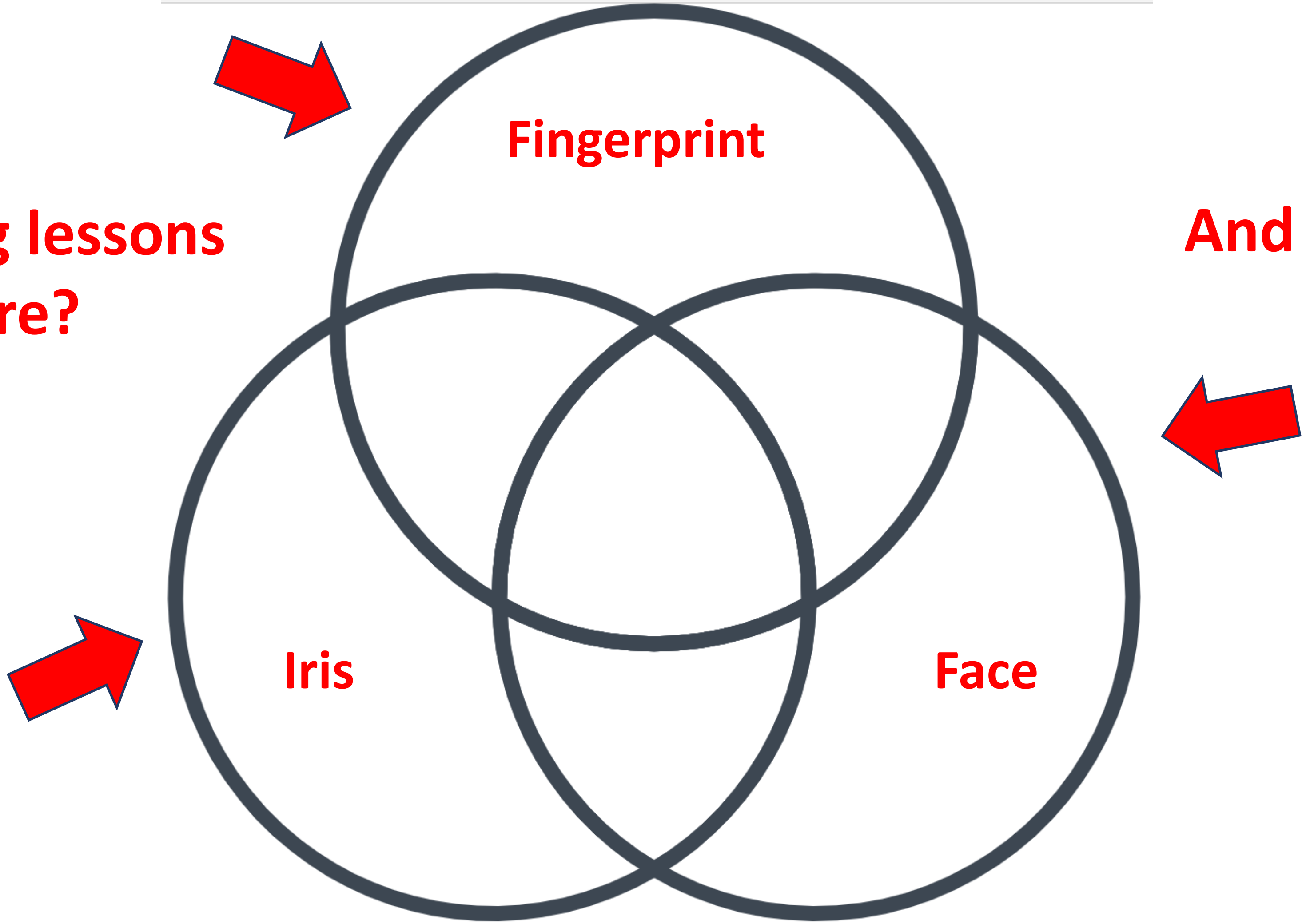
Lessons Learned

Are we taking lessons from here?



Lessons Learned

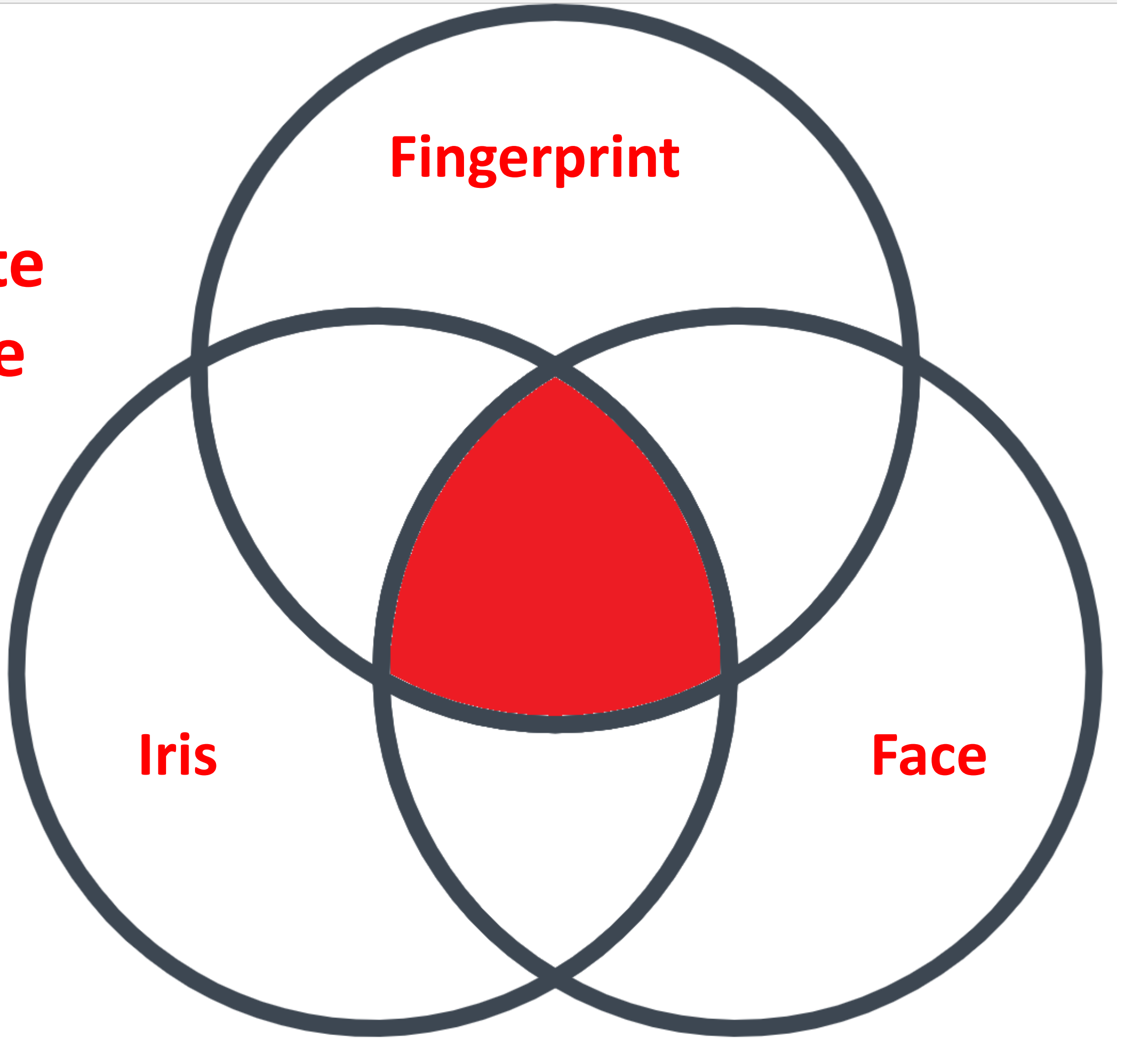
Are we taking lessons from here?



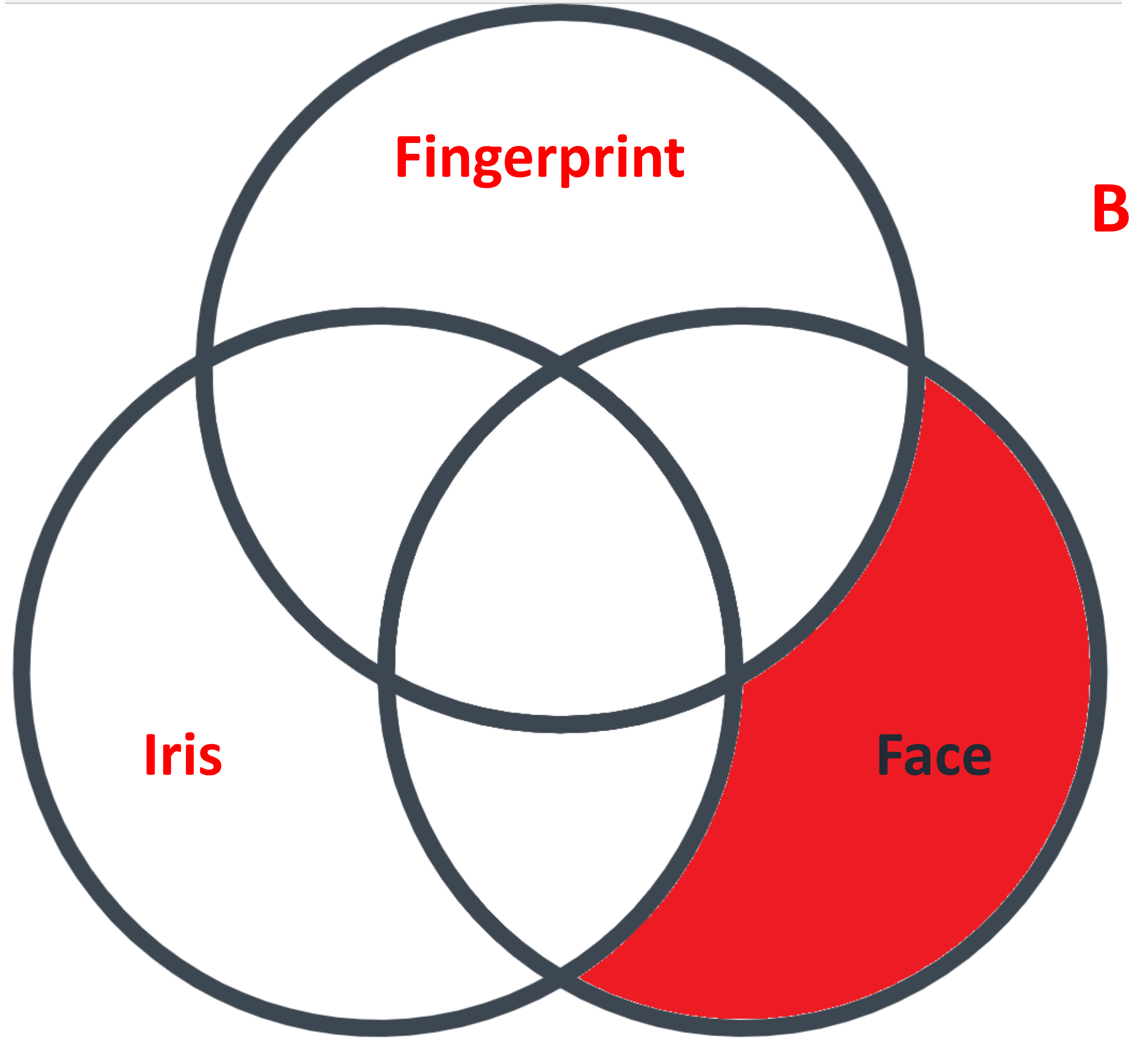
And applying them here?

Lessons Learned

May be appropriate because this space exists



Lessons Learned

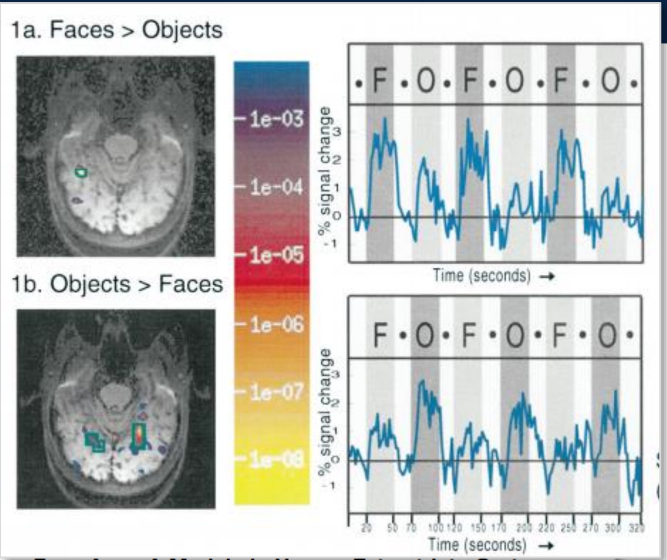


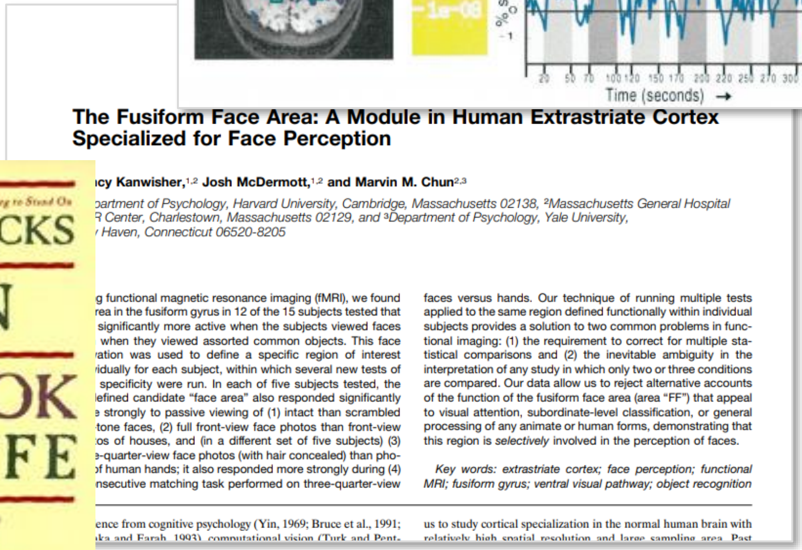
But we need to keep in mind that this space exists as well

Faces are 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

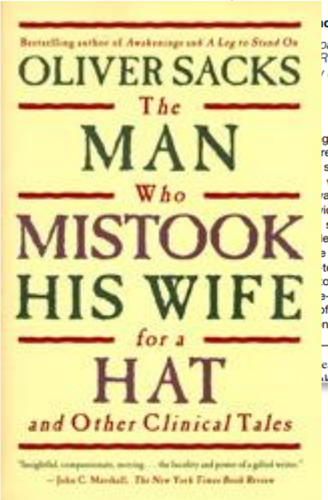
ncy Kanwisher,[1,2] Josh McDermott,[1,2] and Marvin M. Chun[2,3]

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 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

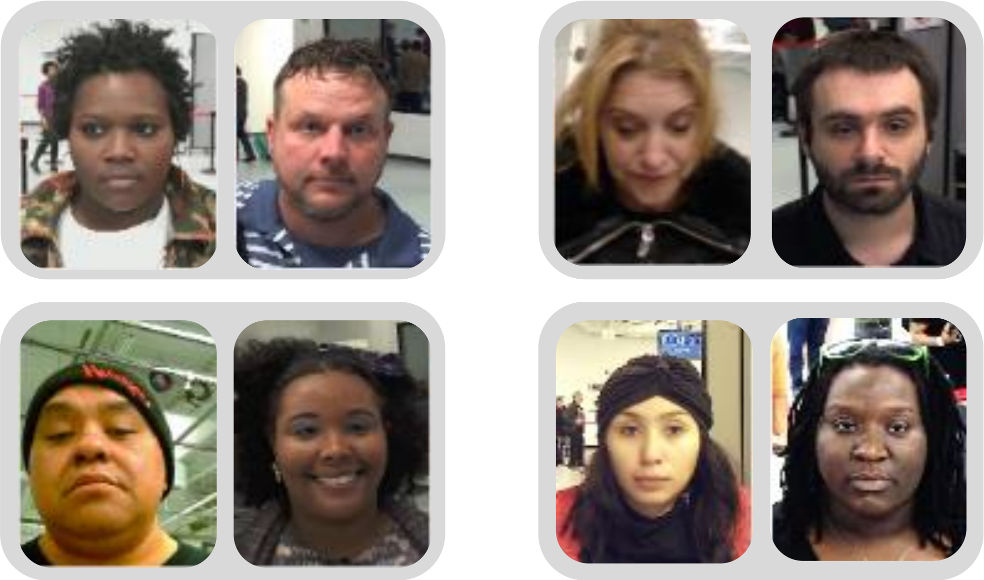
ence from cognitive psychology (Yin, 1969; Bruce et al., 1991; us to study cortical specialization in the normal human brain with



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

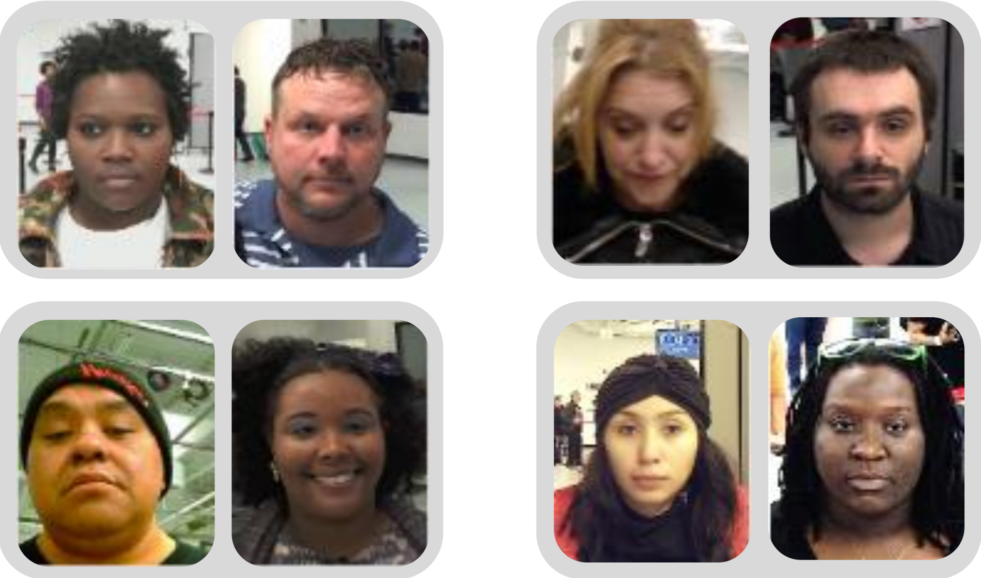
80% of face recognition false positives were between people of the same race and gender

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



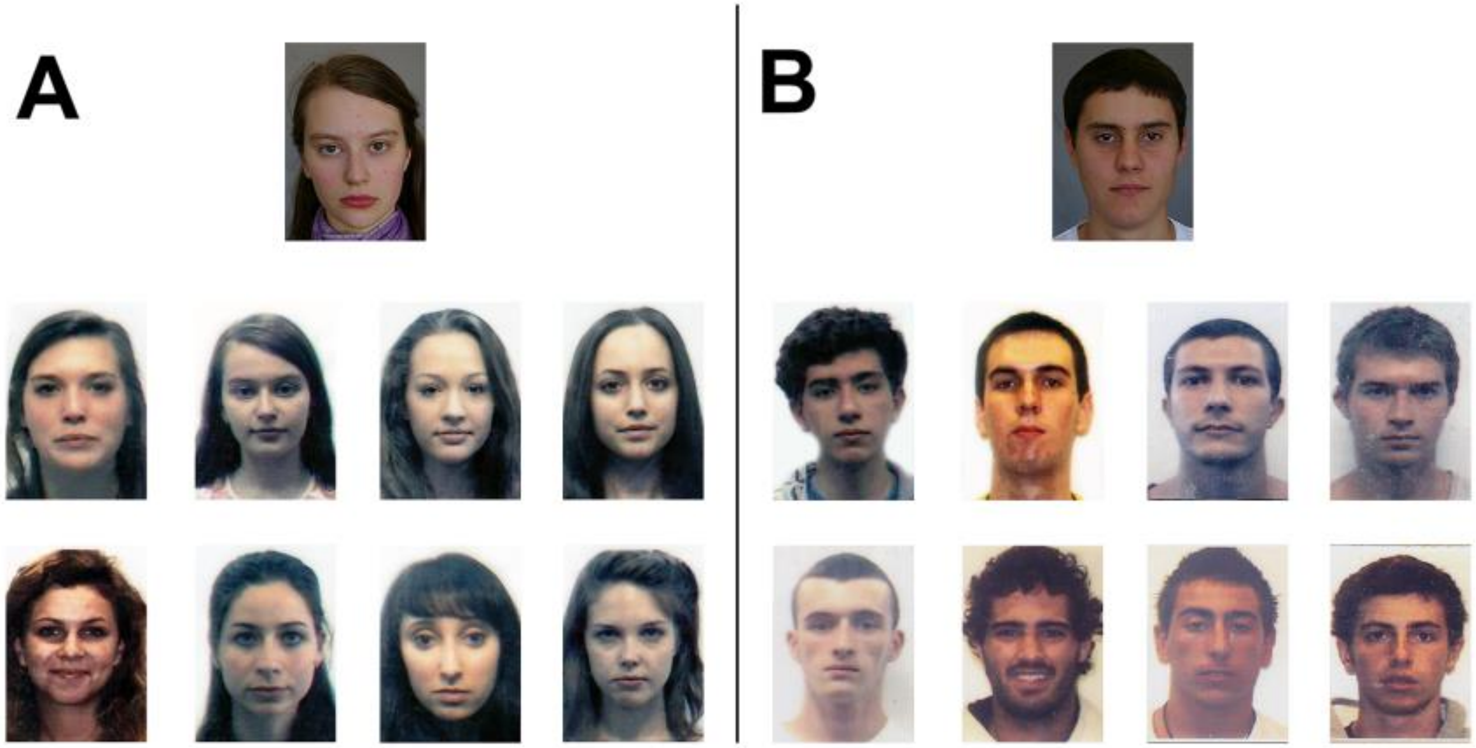
Iris recognition false positives were random relative to race and gender

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

Problem #1 - This Makes Adjudicator Jobs Harder & Slower



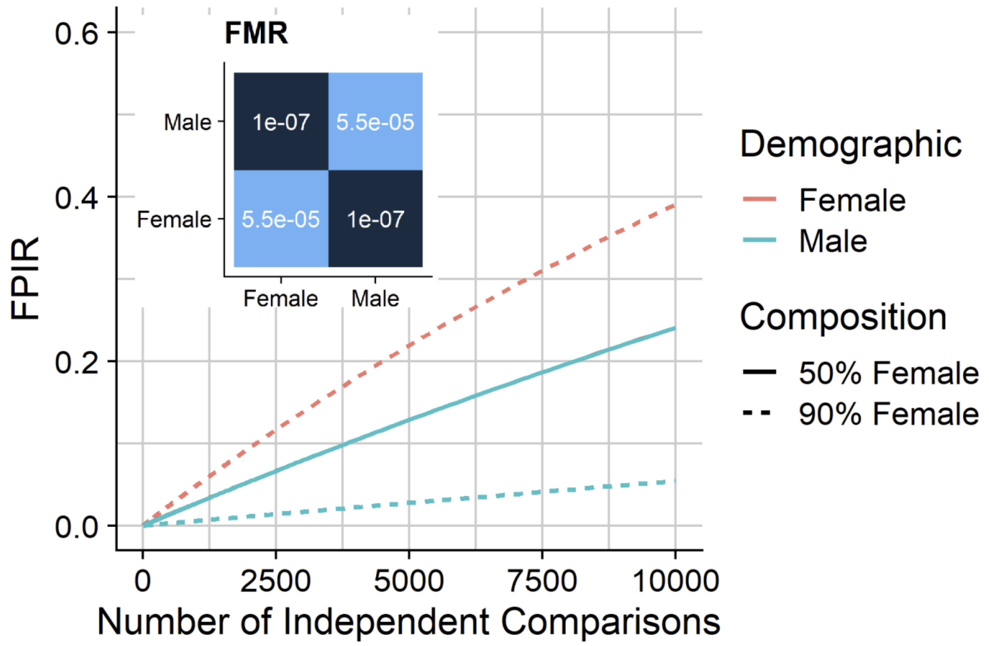
- White et. al “Error Rates in Users of Automatic Face Recognition Software”
- **50% - 60%** errors rates
- If ability of the human to correct the error is the distinguishing factor, **within group false match is not the same as an out group false match**

Problem #2: This Can Impact “Fairness”

- The “watchlist imbalance effect”
  - Howard et. al (2021)
  - Drodowski et. al (2021)

- In the presence of “broad homogeneity”, if you have a watch-list gallery of majority white males:
  - An innocent white male has a higher likelihood of a false positive..
  - .. than a similarly innocent member of a different demographic group

- If impact on 1:N fairness is the distinguishing factor, **within group false match is not the same as an out group false match**

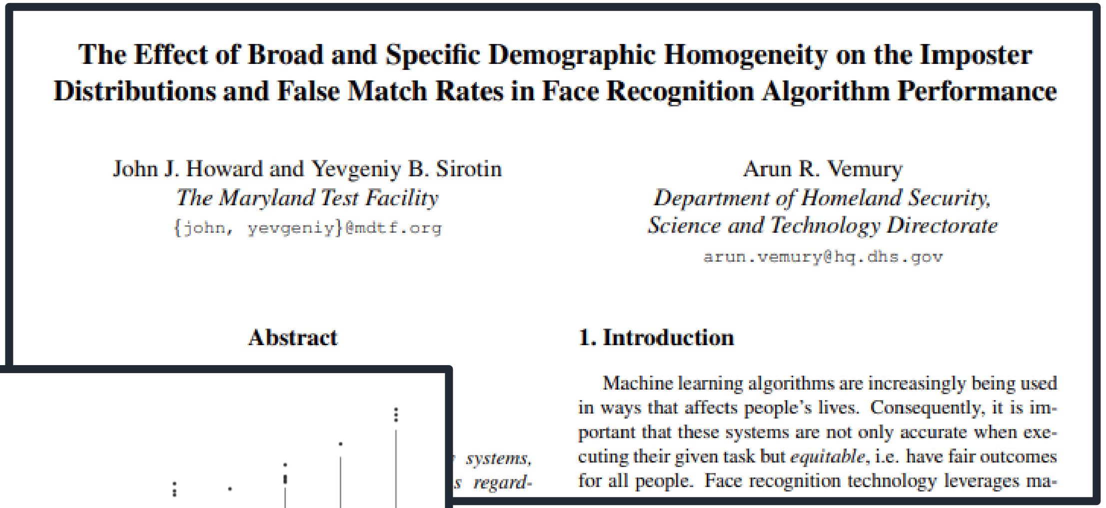


Problem #3 – Overly Optimistic Security

- Imagine a system that matches people to their driver's license photo
- The system designer sets a FMR threshold so that the odds of someone stealing someone else's driver's license and getting away with it are 1 in 1,000 (global FMR)
- But people wouldn't try to assume a random face
- The within group FMR is much lower, two orders of magnitude by some estimates
- What you thought was a 1 in 1,000 FMR, may be more like 1 in 10
- Mismatch between what computer scientists think is "zero-effort" (all faces) and what an imposter thinks is "zero-effort" (finding faces of a similar gender, race, and age).
- If estimating real world error rates is the objective, **within group false match is not the same as an out group false match**

Broad Homogeneity – A Note on Prevalence

- We coined the term “broad homogeneity” to describe this “sameness” effect 2019
- We showed this effect exists in **one** commercial face recognition algorithm



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-

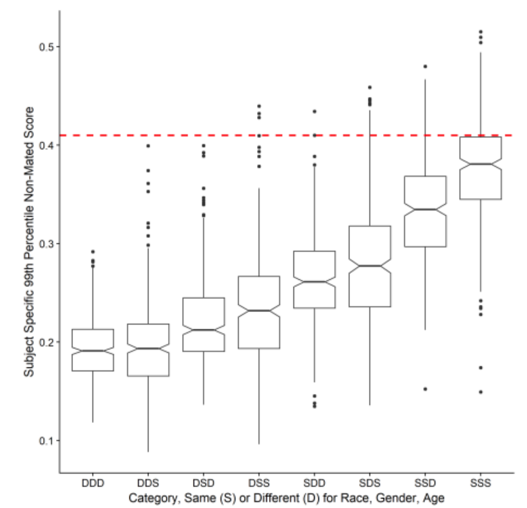
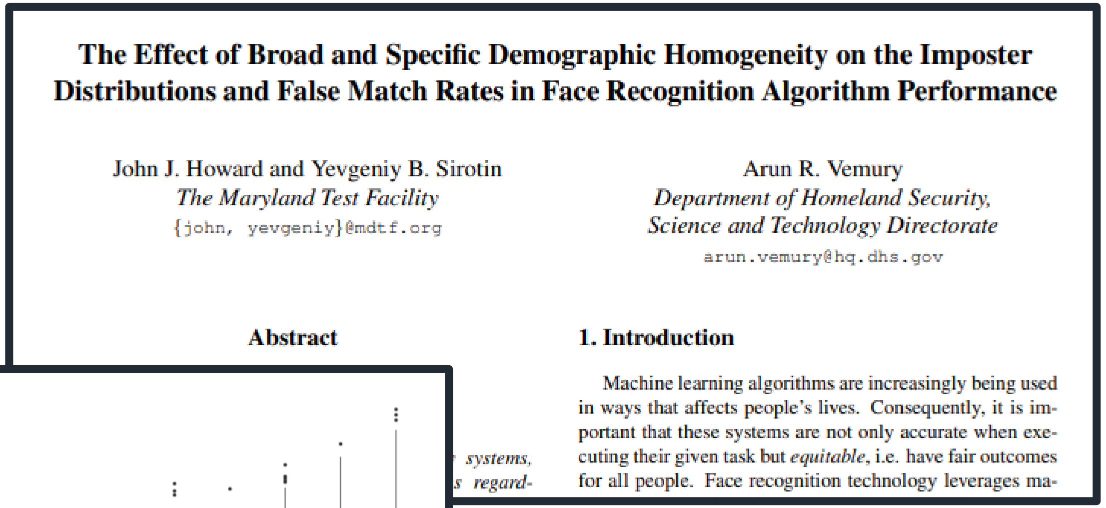


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

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

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-

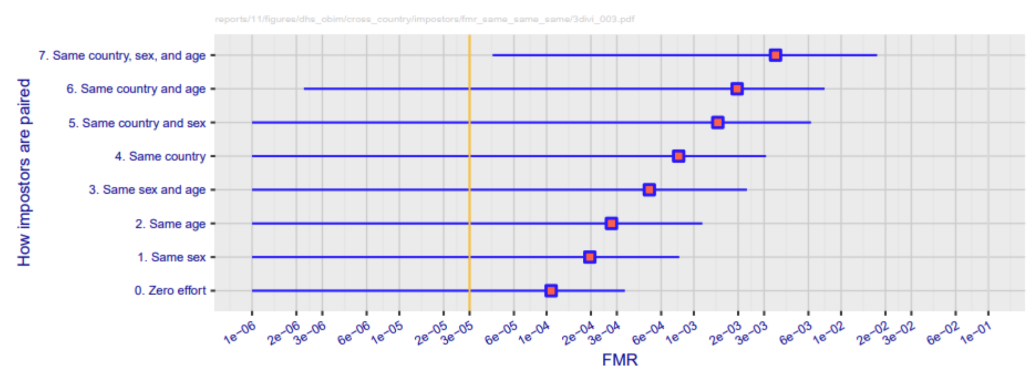


Figure 1: FMR for increasing matched covariates, 3divi-003

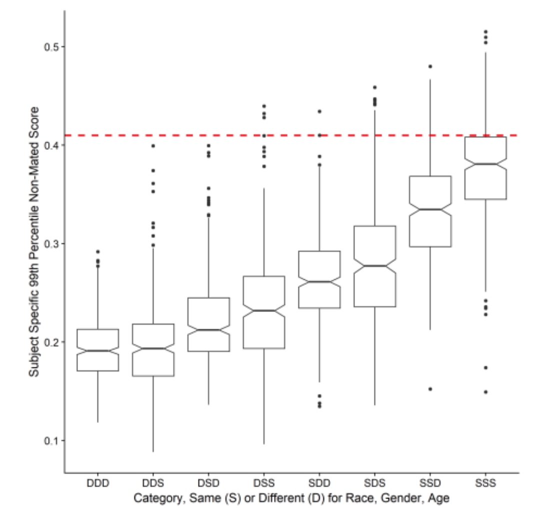


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

Science and Technology

But There May be Solutions

- **IF** we recognize this as a problem..
- We may be able to address it
- Estimated **6 – 14%** of face information content clustered by race and gender (2021).

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

But There May be Solutions

- **IF** we recognize this as a problem..
- We may be able to address it
- Estimated **6 – 14%** of face information content clustered by race and gender (2021).
- Showed a method to **remove this clustering** improved "fairness" across five different fairness measures (2022).

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

epartment of Homeland Security

Appeared in 26th International Conference on Pattern Recognition (ICPR 2022), Fairness in Biometrics Workshop, Montreal, Quebec, August 2022.

Disparate Impact in Facial Recognition Stems from the Broad Homogeneity Effect: A Case Study and Method to Resolve

John J. Howard*[1], Eli J. Laird*†[1], and Yevgeniy B. Sirotin*[1]

The Identity and Data Sciences Lab at The Maryland Test Facility, Maryland, USA
{elaird, jhoward, ysirotin}@idslabs.org

Abstract. Automated face recognition algorithms generate encodings of face images that are compared to other encodings to compute a similarity score between the two originating face images. These face encodings, also known as feature vectors, contain representations of various facial features. Some of these facial features, but not all, have been shown to resemble each other across different subjects that happen to share a de-

What data did we use?

- Data
  - Three of face samples collected from the 2018-200 Biometric Technology Rallies:
    - S1 – demographically balanced training set
    - S2 – disjoint test set
    - S3 – mated pairs to subjects in S1

Dataset	Subjects (Samples)			
	Black Female	**Black Male**	**White Female**	**White Male**
S1	150 (150)	150 (150)	150 (150)	150 (150)
S2	50 (50)	50 (50)	49 (49)	43 (43)
S3	106 (300)	117 (339)	126 (321)	117 (278)

  - Two algorithms
    - ArcFace pre-trained on MS-Celeb-1M
    - ArcFace pre-trained on Glint 360k
  - Requirement for white box template structures

What did we do?

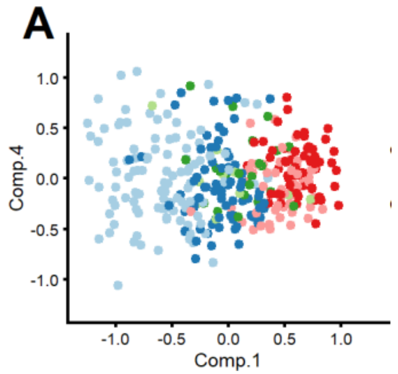
- **Goal:** Given a matrix V of face recognition **feature vectors**, identify components of those vectors that exhibit demographic clustering.
- **Process:**
  - SVD on normalized feature vector matrix, creates subject specific space (U) and a feature space (W^T)

$$\hat{V} = U\Sigma W^T, \text{where } U \in \mathbb{R}^{n\times n}, \Sigma \in \mathbb{R}^{n\times p}, W^T \in \mathbb{R}^{p\times p}$$

  - Calculate clustering index (C_k)

$$C_k = 1 - \frac{\sum_D \sum_{i\in D}(u_i - \bar{u}_D)^2}{\sum_i (u_i - \bar{u})^2}, \quad k, i \in \{1, \ldots, n\}$$

  - Identify components in U with $C_k > 99^{th}\ percentile$ of the bootstrapped C_k distribution



Science and Technology

What did we do?

- Given we found r components in the U matrix with statistically significant clustering

- Remove r columns from W which correspond to the r clustered components in U,
  - *Leaving* $\widehat{W} \in \mathbb{R}^{p \times m}$*, where* $m = p - r$

- Define **de-clustering transform** $\widehat{\mathrm{W}}\widehat{\mathrm{W}}^{\mathrm{T}}$

What did we do?

- Can apply $\widehat{\mathrm{W}}\widehat{\mathrm{W}}^{\mathrm{T}}$ to the set of feature vectors it was learned on
  - $\dot{V} = V\widehat{W}\widehat{W}^T$
  - **Q1**: How demographically “fair” are comparison scores generated from $\dot{V}$ versus V ?

- Can apply $\widehat{\mathrm{W}}\widehat{\mathrm{W}}^{\mathrm{T}}$ to any arbitrary face feature vector v (from the same algorithm)
  - $\dot{v} = v\widehat{W}\widehat{W}^T$
  - **Q2**: If we learn features that exhibit demographic clustering on one set of subjects, do those same featured cluster on other subjects?

What did we do?

- **Experiment 1** - De-clustering Learned and Applied to the Same Dataset (S1)
  - Performed *n x n* comparisons for S1 (360,000 comparisons)
  - Learned & Applied de-clustering transform to S1 feature vectors
  - Evaluated false match rate (FMR) differentials pre- and post-applying transformation

- **Experiment 2** - De-clustering Learned on One Dataset and Applied to a Disjoint Dataset (S2)
  - Performed *n x n* comparisons for S2 (36,864 comparisons)
  - Applied de-clustering transform learned on S1 to S2 feature vectors
  - Evaluated false match rate differentials (FMR) pre- and post-applying transformation

Dataset	Subjects (Samples)			
	Black Female	**Black Male**	**White Female**	**White Male**
S1	150 (150)	150 (150)	150 (150)	150 (150)
S2	50 (50)	50 (50)	49 (49)	43 (43)
S3	106 (300)	117 (339)	126 (321)	117 (278)

How did we measure success?

- Five face recognition fairness measures:
  - Net Clustering [1]
  - Gini Aggregation Rate for Biometric Equitability (GARBE) [2]
  - Fairness Discrepancy Rate (FDR) [3]
  - NIST Inequity Ratio* – all ratios
  - NIST Inequity Ratio [4] – along the diagonal

- Investigated these measures at a threshold that gives a global FMR of 1e-3

- Broad homogeneity is a non-mated effect (alpha = 1, Beta = 0)

[1] Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms (2020)
[2] Howard, J., Laird, E., Sirotin, Y., Rubin, R., Tipton, J., and Vemury, A.. (2022). Evaluating Proposed Fairness Models for Face Recognition Algorithms.
[3] Pereira, T.d.F., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Transactions on Biometrics, Behavior, and Identity Science pp. 11 (2021). https://doi.org/10.1109/TBIOM.2021.3102862
[4] Grother, P.: Face recognition vendor test (frvt) part 8: Summarizing demographic differentials (2022)

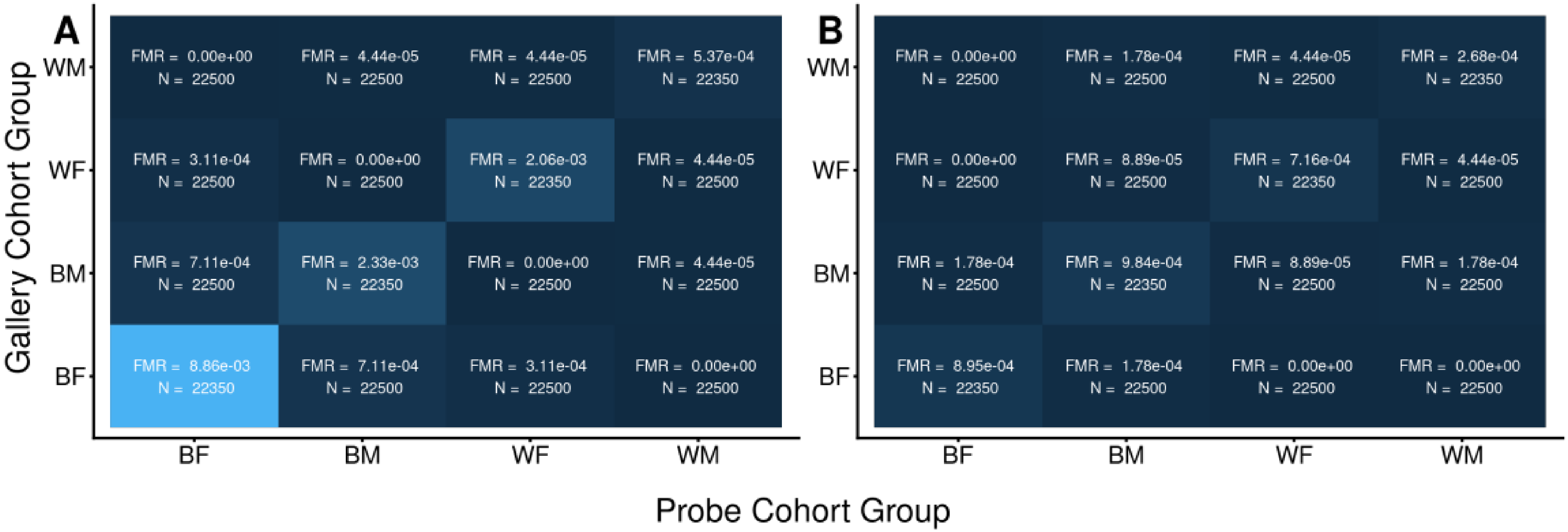
What we found

- Most "fair" values are in **bold** (higher for FDR, lower for all others)
- Applying this demographic de-clustering **universally improved "fairness"**
- Across **two face recognition algorithms**
- Even when applied to an **"unknown" set of subjects** (S2)

Algorithm	Fairness Metric	Experiment 1		Experiment 2	
		S1 Original	S1 Transformed	S2 Original	S2 Transformed
ArcFace-MS1MV2	Net Clustering	0.0163	**0.00549**	0.0252	**0.0207**
	GARBE	0.8540	**0.65000**	0.922	**0.909**
	FDR	0.9900	**0.99900**	0.991	**0.993**
	INEQ	219.00	**30.2000**	22.00	**18.00**
	INEQ*	15.58	**3.74**	10.56	**6.62**
ArcFace-Glint360k	Net Clustering	0.0150	**0.00497**	0.0250	**0.0197**
	GARBE	0.8350	**0.67100**	0.955	**0.881**
	FDR	0.9910	**0.99900**	0.990	**0.996**
	INEQ	199.00	**22.1000**	12.5	**10.20**
	INEQ*	16.23	**3.67**	12.47	**3.68**

What does this do to false match cohort matrices?

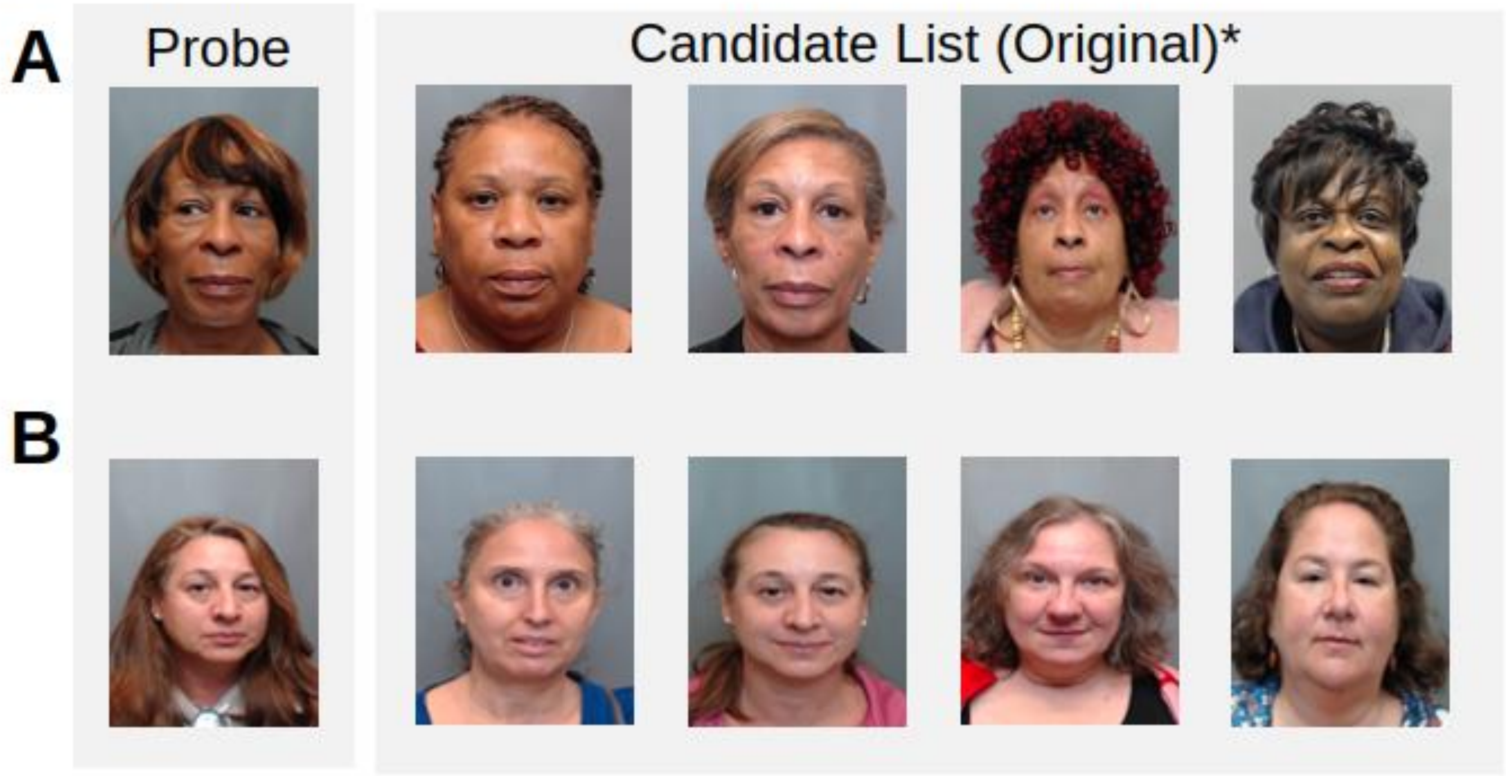
- One example (Glint 360k S1->S1 dataset):



Science and Technology

What does this do to human review?

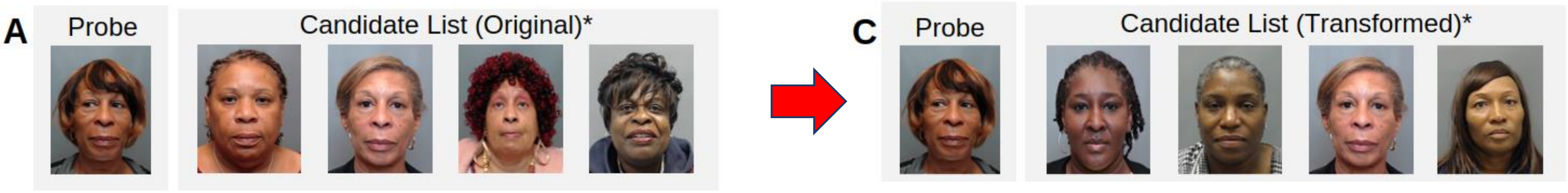
- Pulled two rank 4 probe and candidate lists:



What does this do to human review?

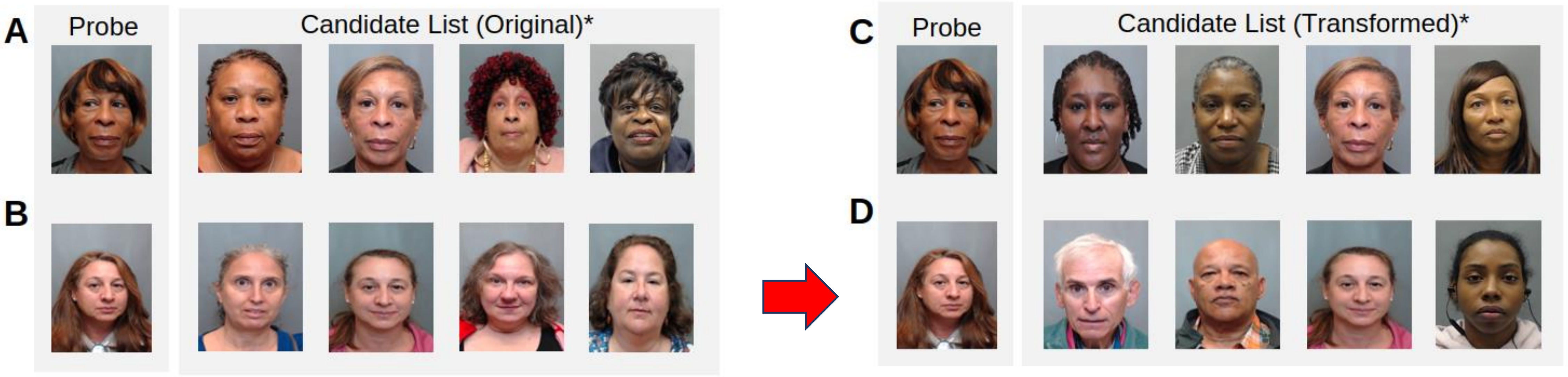
Before De-clustering

After De-clustering



For some subjects, one broadly homogenous candidate set was replaced with another

What does this do to human review?



But for others, a homogenous set was replaced with a non-homogenous one

Current literature on face matching in humans work suggest these are much easier for humans to review

Future Work

- There are **numerous** additional questions to answer in this area.
- What is the best means to identify and remove “clustering” in feature vector space?
- What is the best metric for results? Need something beyond false match rate.
- How stable are these transforms across and within demographic group? Can they be made more stable?
- What is the best algorithm for a human to work with? Might not be “the best algorithm”

Questions & Answers

- Contact information
  - arun.vemury@hq.dhs.gov
  - jhoward@idslabs.org
  - ysirotin@idslabs.org
  - peoplescreening@hq.dhs.gov

- Visit our websites for additional information
  - To see additional work DHS S&T supports, visit www.dhs.gov/science-and-technology
  - Detailed application instructions will be available in a separate document on https://mdtf.org
  - To view additional information about this year and prior Rallies, visit https://mdtf.org