DHS SCIENCE AND TECHNOLOGY

Quantifying Race and Gender Effects in Face versus Iris Algorithms

The International Face Performance Conference 2020

John Howard, Yevgeniy Sirotin, & Jerry Tipton
The Maryland Test Facility

Arun Vemury
Director
Biometric and Identity Technology Center
Science and Technology Directorate

Homeland Security
Science and Technology

Disclaimer

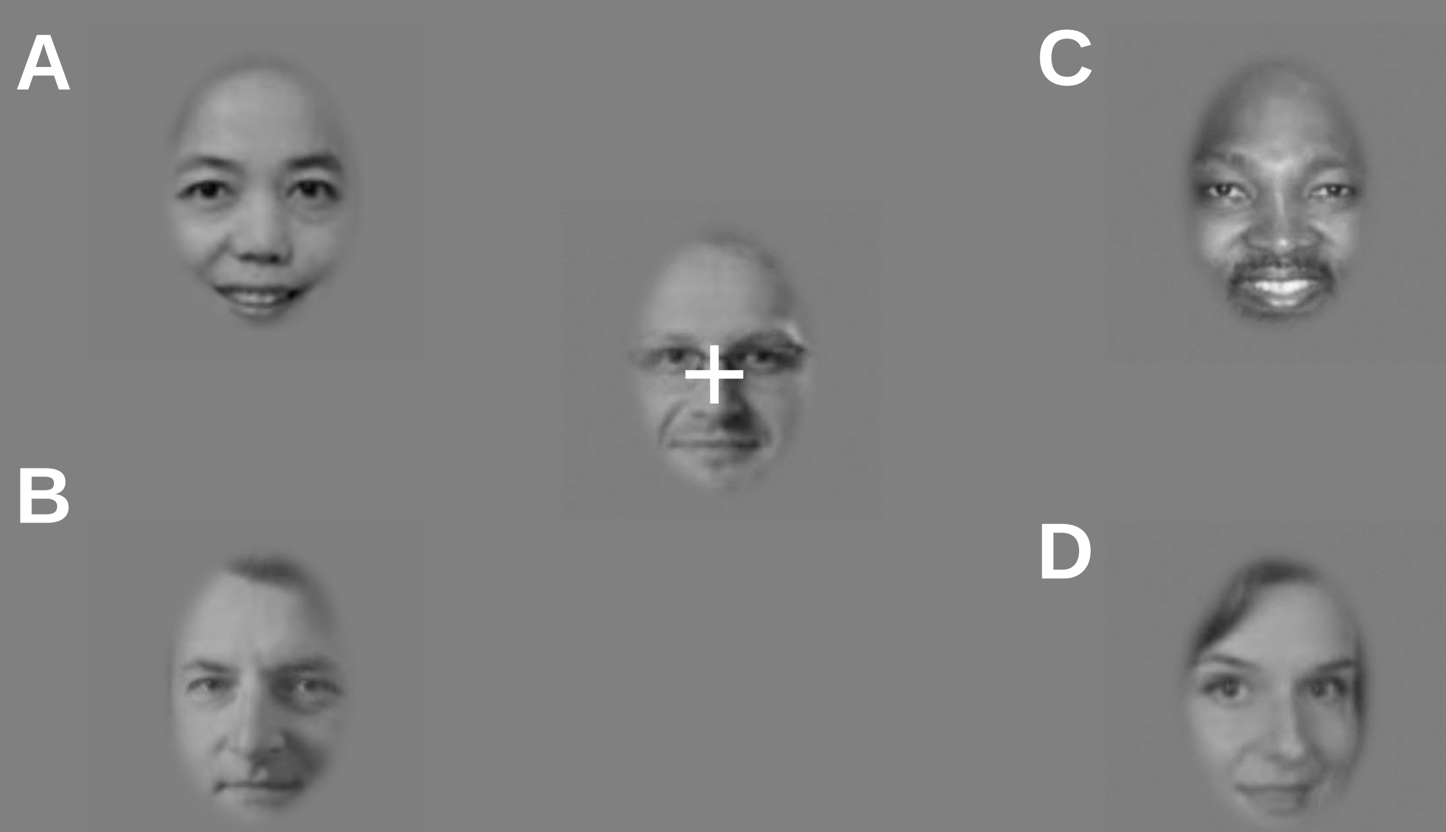
- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.
- This work was performed by a team of researchers at the Maryland Test Facility.
- The views presented here are those of the authors and do not represent those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

Background

- Recent reports have shown that biometric performance can vary for people based on demographic group membership
- This has been most notable in commercial face recognition algorithms
  - NIST's FRVT showed **some face algorithms can have 100-fold difference in FMR across groups**
  - However, there are also **"broad homogeneity"** effects in face algorithms whereby comparisons between **individuals similar in race, age, and gender produce higher scores**
  - This does not appear to occur in iris recognition

Differential Performance

Broad Homogeneity



Images taken from public sources, posted under fair use doctrine per 17 U.S. Code § 107

Broad Homogeneity

A

Unknown Person:

C

B

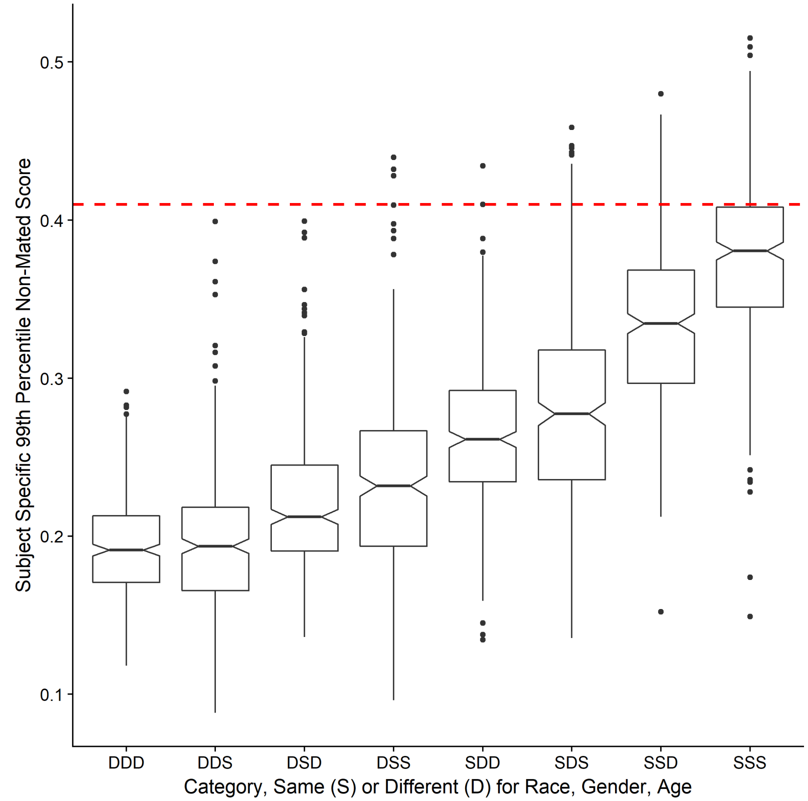
D

Images taken from public sources, posted under fair use doctrine per 17 U.S. Code § 107

Broad Homogeneity

- In face recognition, **you are more likely to match to someone who shares your demographic characteristics**

- We showed this was true in one commercial face recognition algorithm in 2019 [1]



[1]: Howard, Sirotin, Vemury. *The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance*. BTAS 2019. Copy available: https://mdtf.org/publications/broad-and-specific-homogeneity.pdf

Broad Homogeneity

- Evaluated five other commercial face algorithms in 2019/2020. The "broad homogeneity" effect was observed in each algorithm [1].

- We observe broad homogeneity is a general property of current commercial face recognition systems.

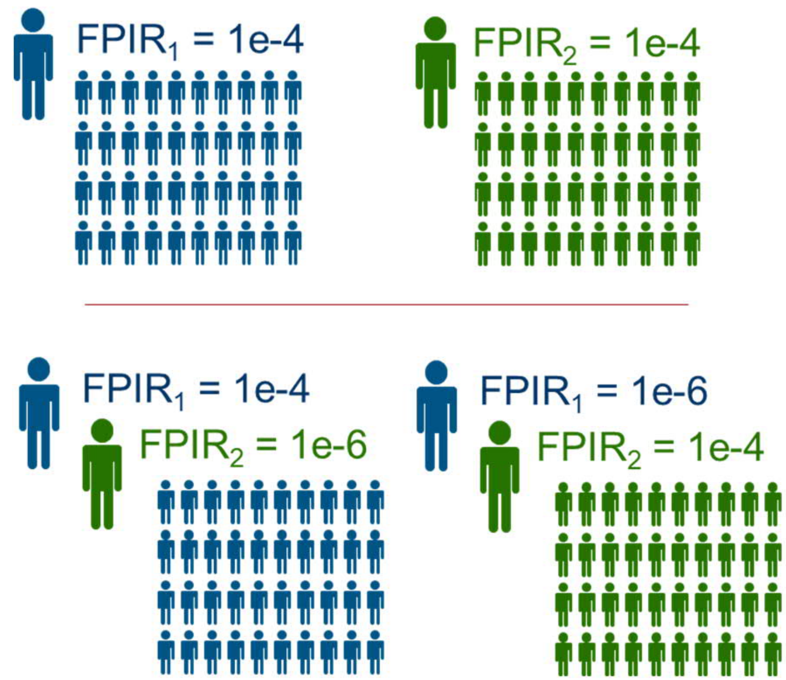
- While intuitive, this property of face algorithms **can create undesirable behavior in many identification scenarios.**

- If an identification gallery, such as a most wanted list, skews predominantly male, then men who are not in the gallery are more likely to be mis-identified when searched against that gallery than women, solely on the basis of their male facial features.

[1]: Howard, Sirotin, Tipton, & Vemury. *Quantifying the Extent to Which Race and Gender Features Determine Identity in Commercial Face Recognition Algorithms*. Pre-print available: https://arxiv.org/abs/2010.07979.

Why are broad homogeneity effects problematic?

- Suppose two algorithms are evaluated separately on two groups (group 1 and group 2)
  - With equal FPIR against their peers
- However, members of the two groups can still have different FPIR against homogeneous galleries
  - Differential performance even if algorithm performs equally well for each group
- This may lead to **differential impact** in a law enforcement context reflecting pre-existing gallery demographic composition



Broad Homogeneity

- This was discussed in the Georgetown Perpetual Lineup paper in 2016 [1]

- As a scientific community, we don't have a metric to measure this

- FMR's per specific group (i.e. white females vs. black females) are measures of "specific homogeneity"
  - NIST FRVT revealed 100x difference in FMR across demographic groups
  - Measure of how often the event (false match) occurs, per group

- Currently little formal reporting on the effect of cross group "sameness"

- Here we will present an approach to understanding and **quantifying broad homogeneity effects** so that they can be compared **across algorithms**
  - We will discuss implications of these results for face and iris recognition

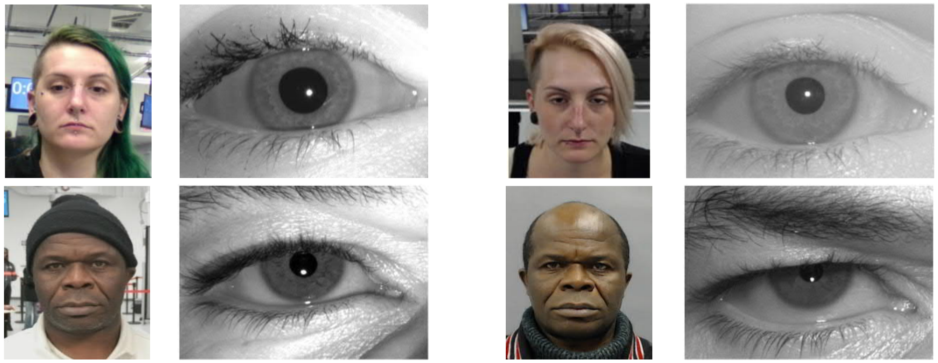
[1]:Garvie, Clare; Bedoya, Alvaro M.; Frankle, Jonathan (2016): The Perpetual Line-Up. Unregulated Police Face Recognition In America. Georgetown Law Center on Privacy & Technology. Available online at www.perpetuallineup.org

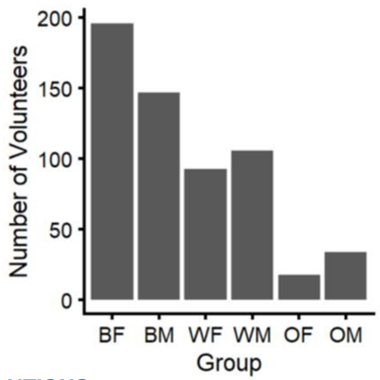
Dataset

- All images were **acquired under IRB protection** and used here with explicit subject consent
- A total of 333 volunteers were used in this analysis
  - 1,205 face images and 1,083 left iris images were gathered from the same volunteers over a five year period from 2012-2018

- Unstaffed high-throughput acquisition environment
- All acquisition and matching systems were **commercial biometric technologies**

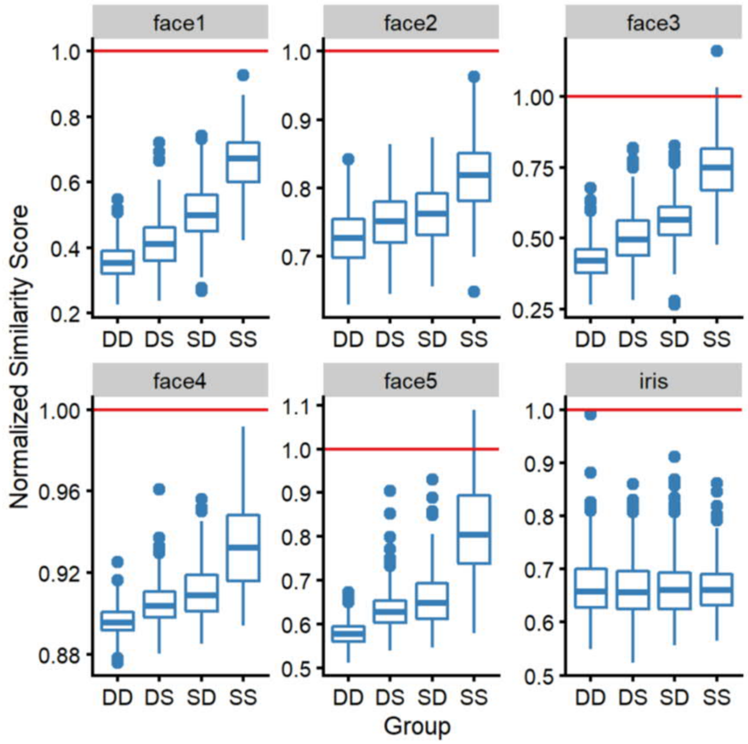
Sample Images





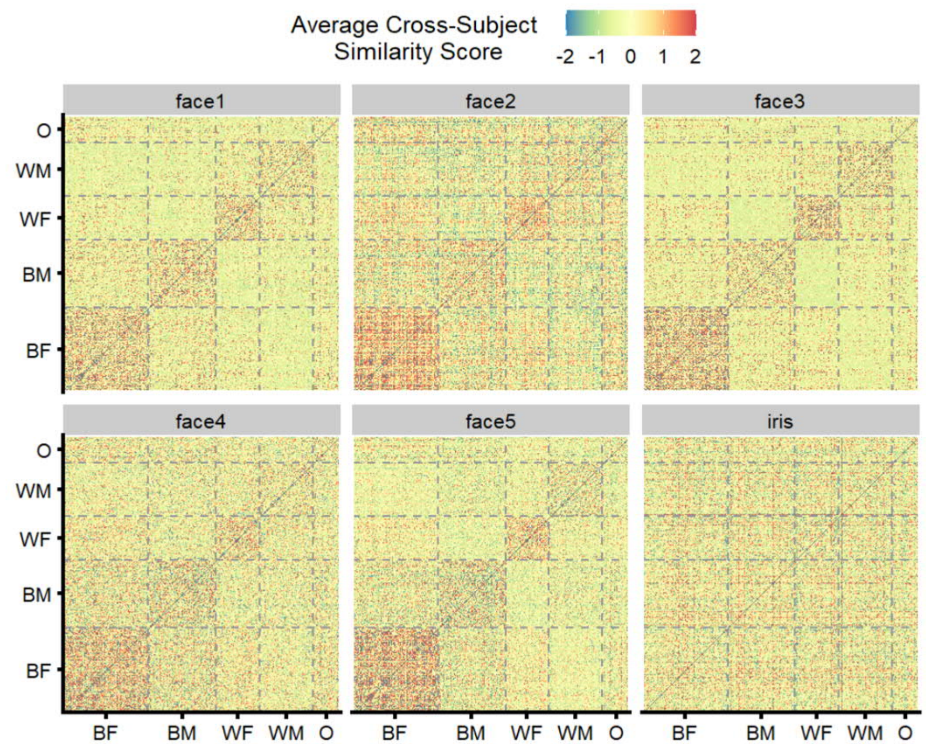
Broad Homogeneity

- All 5 **commercial face algorithms show broad homogeneity effects**
  - Non-mated similarity scores increased with increasing demographic similarity
  - Figure plots 99th percentile non-mated score for each of 333 subjects
  - **DD**: different gender and race
  - **DS**: different gender, same race
  - **SD**: same gender, different race
  - **SS**: same gender and race
- The reference **commercial iris recognition algorithm does not show broad homogeneity effects**
  - This is a classic "Daugman" algorithm



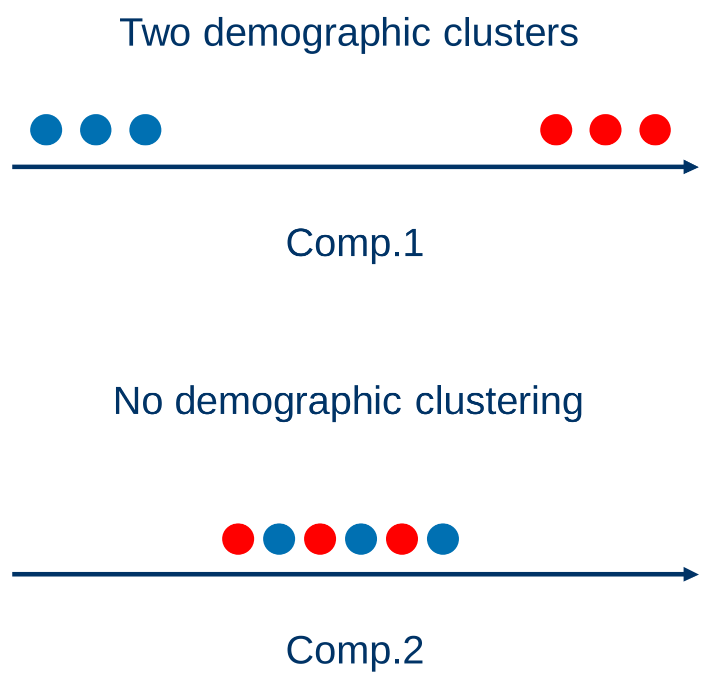
Visualizing Broad Homogeneity

- We measured **average cross-subject similarity** scores and arranged these into **score matrices**
- These matrices were **sorted by demographic group**
- Face algorithms showed **clear block structure** with respect to demographic group membership
- The iris algorithm did not show obvious patterns



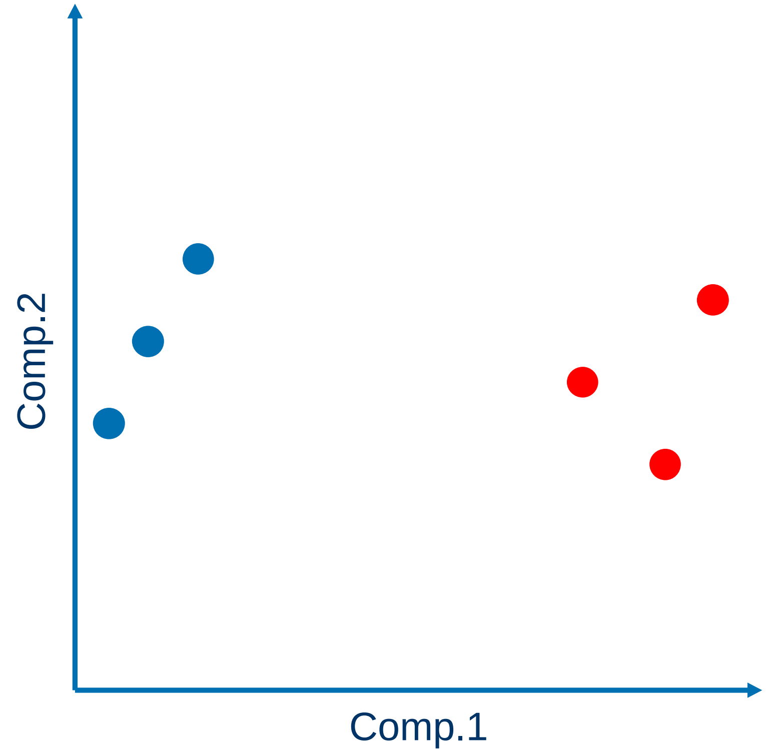
Score Matrix PCA

- Principal components analysis (PCA) is a **linear matrix decomposition technique**
  - It can be used to transform high dimensional data into a series of principal components
  - Each component explains a portion of the total variance in the data
  - The highest level of variance is found on the first component, Comp 1
  - Each subsequent component is orthogonal to the preceding and explains less variance
- Each component corresponds to **a pattern across subjects**
  - We can examine how subjects are arranged along each component



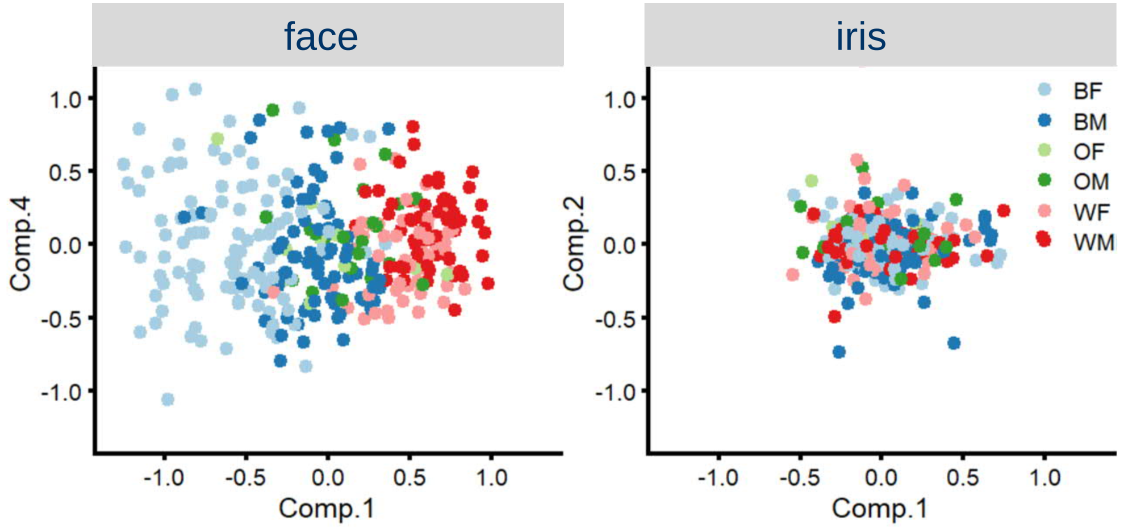
Score Matrix PCA

- Principal components analysis (PCA) is a **linear matrix decomposition technique**
  - It can be used to transform high dimensional data into a series of principal components
  - Each component explains a portion of the total variance in the data
  - The highest level of variance is found on the first component, Comp 1
  - Each subsequent component is orthogonal to the preceding and explains less variance
- Each component corresponds to **a pattern across subjects**
  - We can examine how subjects are arranged along each component



Demographic Clustering in PC Space

- The figure shows two example components for one representative face algorithm and the iris algorithm
- Face algorithm component 1 shows strong demographic clustering
- Face algorithm component 4 does not show clustering
- No iris algorithm components show clustering



Homeland Security
Science and Technology

Quantifying Demographic Clustering

- Each PCA component explains a certain proportion of score variance

- We quantified demographic clustering across demographic groups (D) within each component as:

$$C_k = 1 - \frac{\sum_D \sum_{i \in D} (x_i - \bar{x_D})^2}{\sum_i (x_i - \bar{x})^2}$$

Sum of component variances within each group

Component's total variance

- And total clustering for the algorithm as the sum of clustering for each component weighted by the amount of variance it explains:

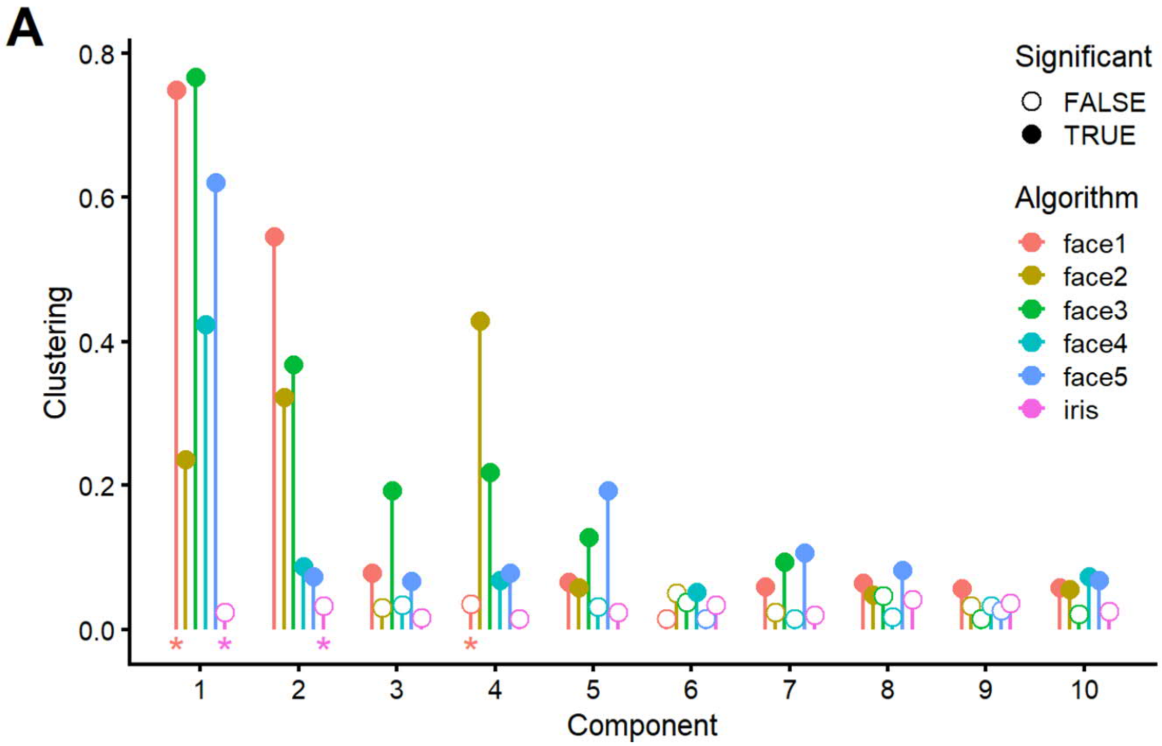
$$C_{tot} = \frac{1}{\sigma_{tot}^2} \sum_k \sigma_k^2 C_k$$

Component's total variance

Total score variance

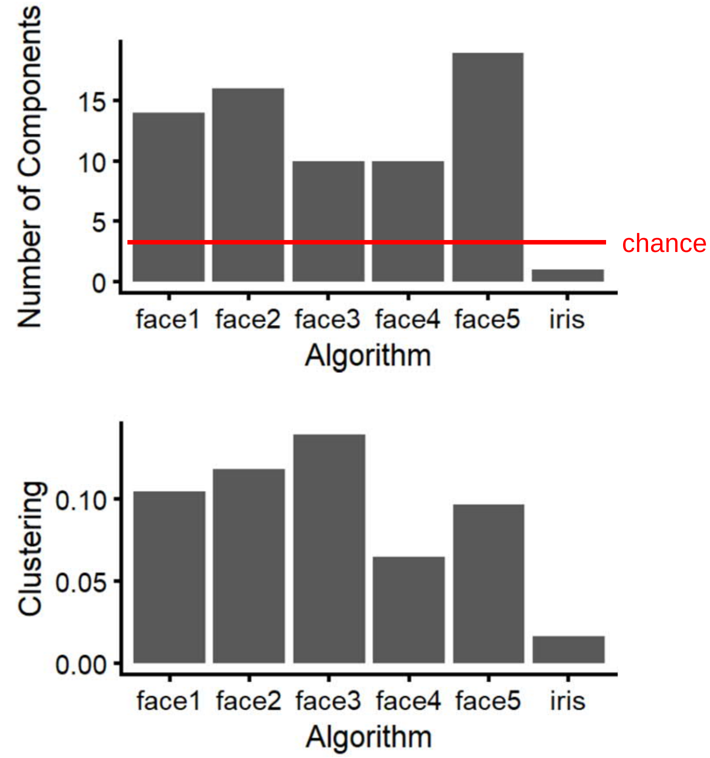
Demographic Clustering in Each Component

- Many, but not all, of the top 10 face algorithm components showed high levels clustering
- Statistical significance of clustering was assessed using bootstrap resampling with randomized demographic labels
- No significant clustering for the iris algorithm



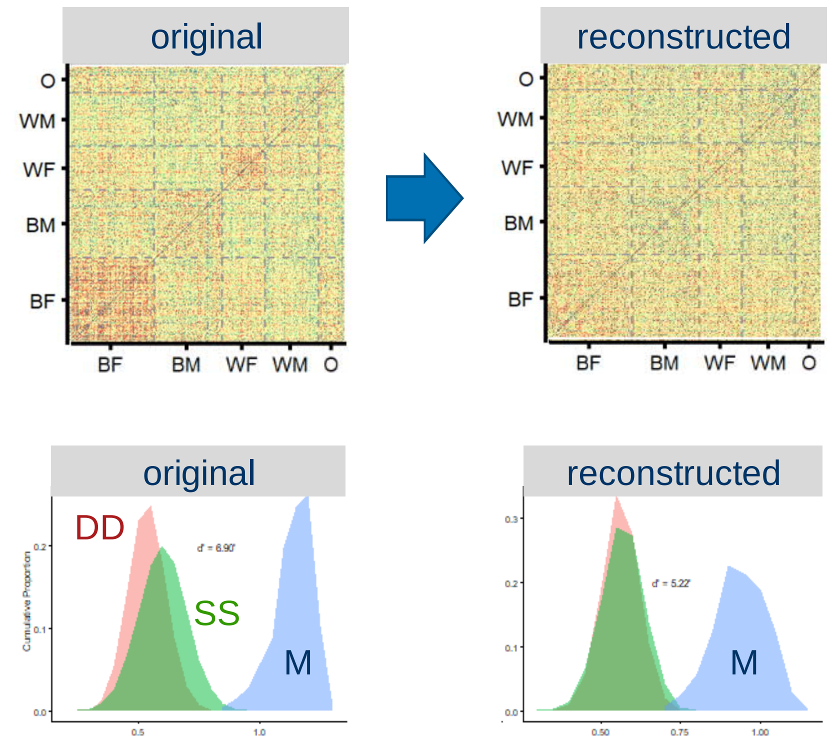
Comparing Clustering across Algorithm

- On average **~10 components** showed significant demographic clustering for face algorithms
- Clustering accounted for **10% of total score variance** in face algorithm scores
- The **iris algorithm had no clustering** in excess of what would be expected by chance
- This quantification is independent of match threshold and can be computed even in the absence of any overlap between the mated and non-mated distributions (ROC = 1)



Face Algorithms Do Not Need Race/Gender Features To Be Viable

- PCA can be used to reconstruct data using select components
- Removed components with significant clustering and **reconstructed the score matrices**
  1. Better overlap for non-mated distributions for comparisons between volunteers of the same gender and race (SS) and those between volunteers of different gender and race (DD)
  2. Reduced separation between the mated (M) and non-mated distributions (DD, SS)
- The reduction in separation was **not "catastrophic" to performance**:
  - d' for the best face algorithm after reconstruction was better than all other face algorithms before reconstruction



Why is this Important?

- "Broad homogeneity" is an undesirable characteristic, particularly if you want to do large identifications
- Exists in (likely all) currently available commercial FR systems
- Being talked about in civil liberties / privacy law circles
  - They are aware of this because its intuitive that face recognition algorithms behave in this way
- We are working to develop a scientific measure for this effect
  - Few researchers have formulated this as a problem
  - Not clear commercial vendors are aware this is a problem

Why is this Important?

- Broad homogeneity based on race and gender **doesn't currently exist in commercial iris recognition** algorithms (we think)
  - Many current commercial iris algorithms use the "Daugman" algorithm
  - Demonstrated to provide unique iris codes with independent features generally not linked to demographics
- However, race/gender-linked information is plainly available in periocular images
  - E.g. makeup and eye shape
  - Research documenting gender prediction results from iris images
- Face algorithms have experienced significant performance improvements from the use of DCNNs
- Use of **DCNNs for iris recognition may** inadvertently **introduce race and gender features into iris** performance

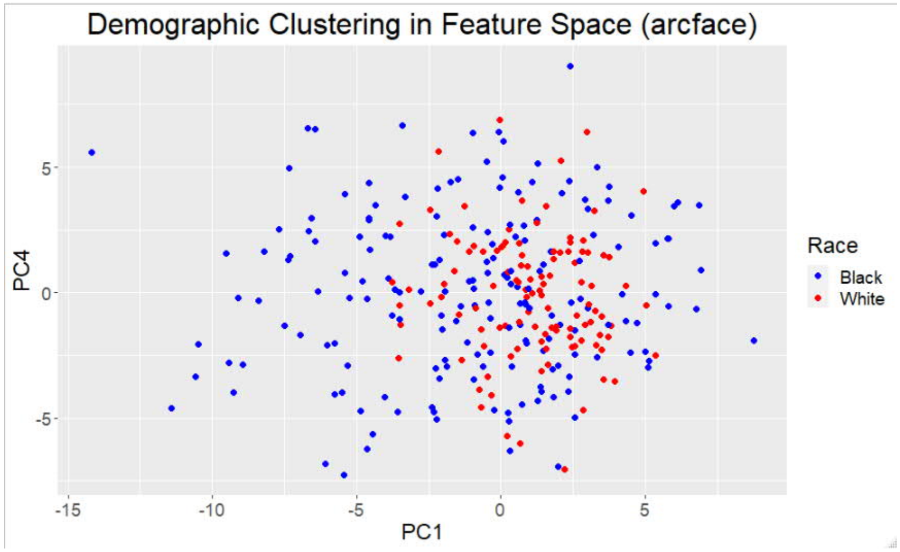
Where do we go from here?

- Methods to **ensure** iris recognition remains independent of demographics should be considered.

- Methods to **remove** face recognition reliance on features that are consistent within demographic categories should be considered.

- We quantified this effect in the **score space** because we were working with black box commercial algorithms (no insight into the template)

- To remove this effect, we need to identify and discard components in the **feature space** that are consistent within demographic group (currently working on this)



Questions?

- This work was performed by a team of researchers at the Maryland Test Facility.
  - Detailed paper at: https://arxiv.org/ftp/arxiv/papers/2010/2010.07979.pdf

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

- john@mdtf.org
- yevgeniy@mdtf.org

- jerry@mdtf.org
- **arun.vemury@hq.dhs.gov**

Homeland Security
Science and Technology