Understanding and Mitigating Bias in Human and Machine Face Recognition

Identity and Data Sciences Laboratories

John J. Howard
Chief Data Scientist
Identity and Data Sciences Laboratory at the
Maryland Test Facility

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT23CB0000003.

- This work was performed by the SAIC 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 an IRB protocol.

Agenda

- The Maryland Test Facility / Identity and Data Sciences Lab

- Demographic differentials or “bias” in Face Recognition
  - What is it?
  - Where does it come from?
  - Why are they bad?
  - How do we measure it (and why we are currently doing that wrong)?
  - How do we fix it?

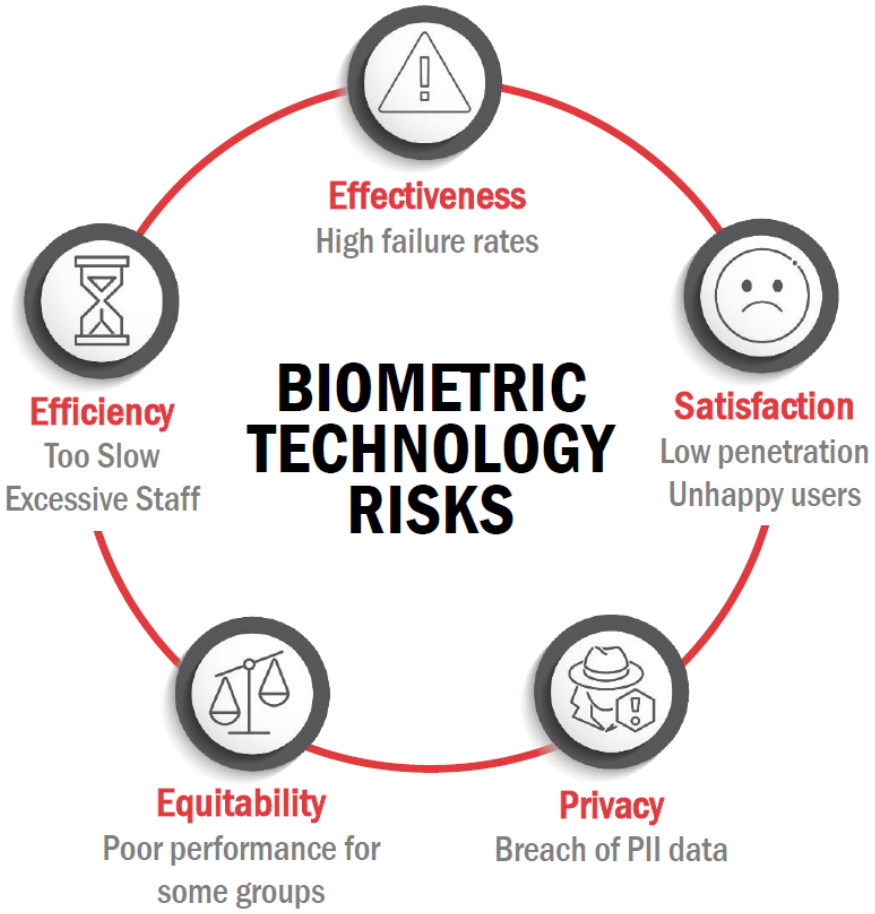
The Identity and Data Sciences Laboratory

- AI testbed specializing in scenario tests of biometric and identity systems
  - Scientists, Engineers, and Biometric SMEs

- Trusted by government and industry stakeholders to perform unbiased assessments

- Biometric and identity systems:
  - Biometric data on ~4000 subjects since 2014.
  - Diverse & ground-truthed collection of gender, race, age, skin-tone, etc.

We work to mitigate risks associated with biometric and identity technologies.

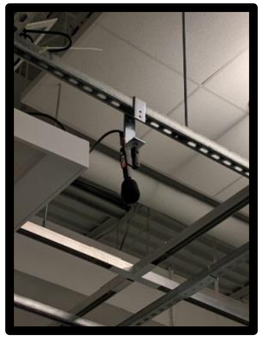


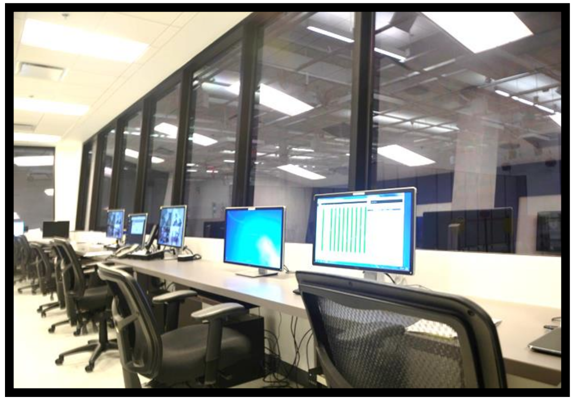
The Maryland Test Facility

- **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.**





What is demographic “bias” in FR

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.

The Alan Turing Institute

Understanding bias in facial recognition technologies

ACLU

NEWS & COMMENTARY

How is Face Recognition Surveillance Technology Racist?

SITN science in the news celebrating 20 years

HARVARD UNIVERSITY
The Graduate School of Arts and Sciences

OCTOBER 24, 2020

BLOG, SCIENCE POLICY, SPECIAL EDITION: SCIENCE POLICY AND SOCIAL JUSTICE

Racial Discrimination in Face Recognition Technology

MIT Schwarzman College of Computing

HE CASES AUTHOR RESOURCES

Winter 2021 ⌄ Published on Feb 05, 2021 DOI 10.21428/2c646de5.62272586

The Bias in the Machine: Facial Recognition Technology and Racial Disparities

IDSL

What is demographic “bias” in FR

- Despite all the attention, the term “bias” is not well defined
- Overloaded term (computer science, statistics, psychology, public discourse)
- Not specific enough (How is it biased? Does it have an impact?)
- Howard, Sirotin, Vemury. *The Effect of Broad and Specific Demographic Homogeneity on the Imposter Distributions and False Match Rates in Face Recognition Algorithm Performance (2019).*

What is demographic “bias” in FR

- **False negative differential** – tendency for a group not to match
- **False positive differential** – tendency for a group to false match

Algorithm: No Match

FND(τ) =

If the rate that this happens

> or <

the rate that this happens

Algorithm: No Match

Algorithm: Match

FPD(τ) =

> or <

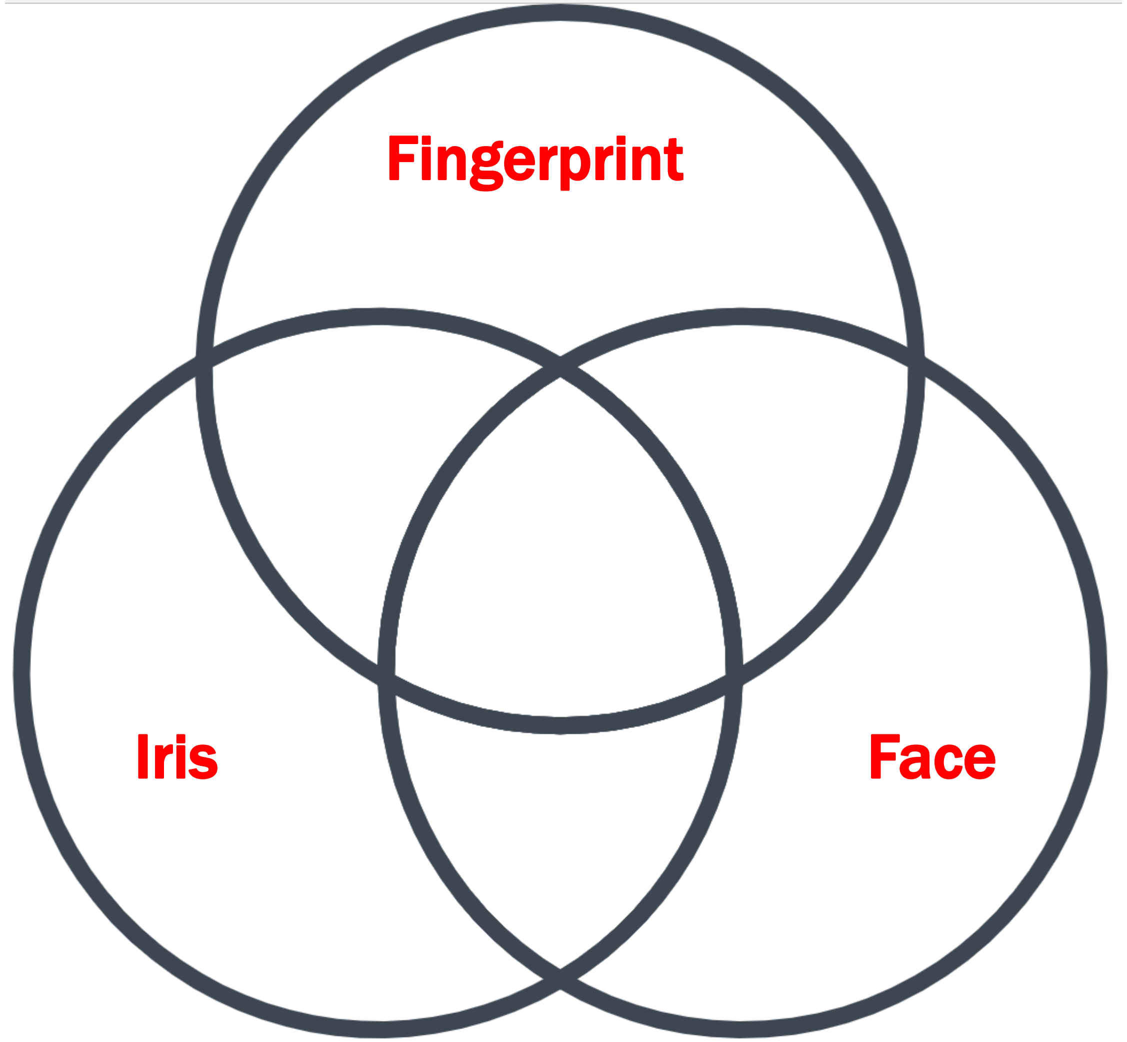
the rate that this happens

Algorithm: Match

Where does “bias” in FR come from?

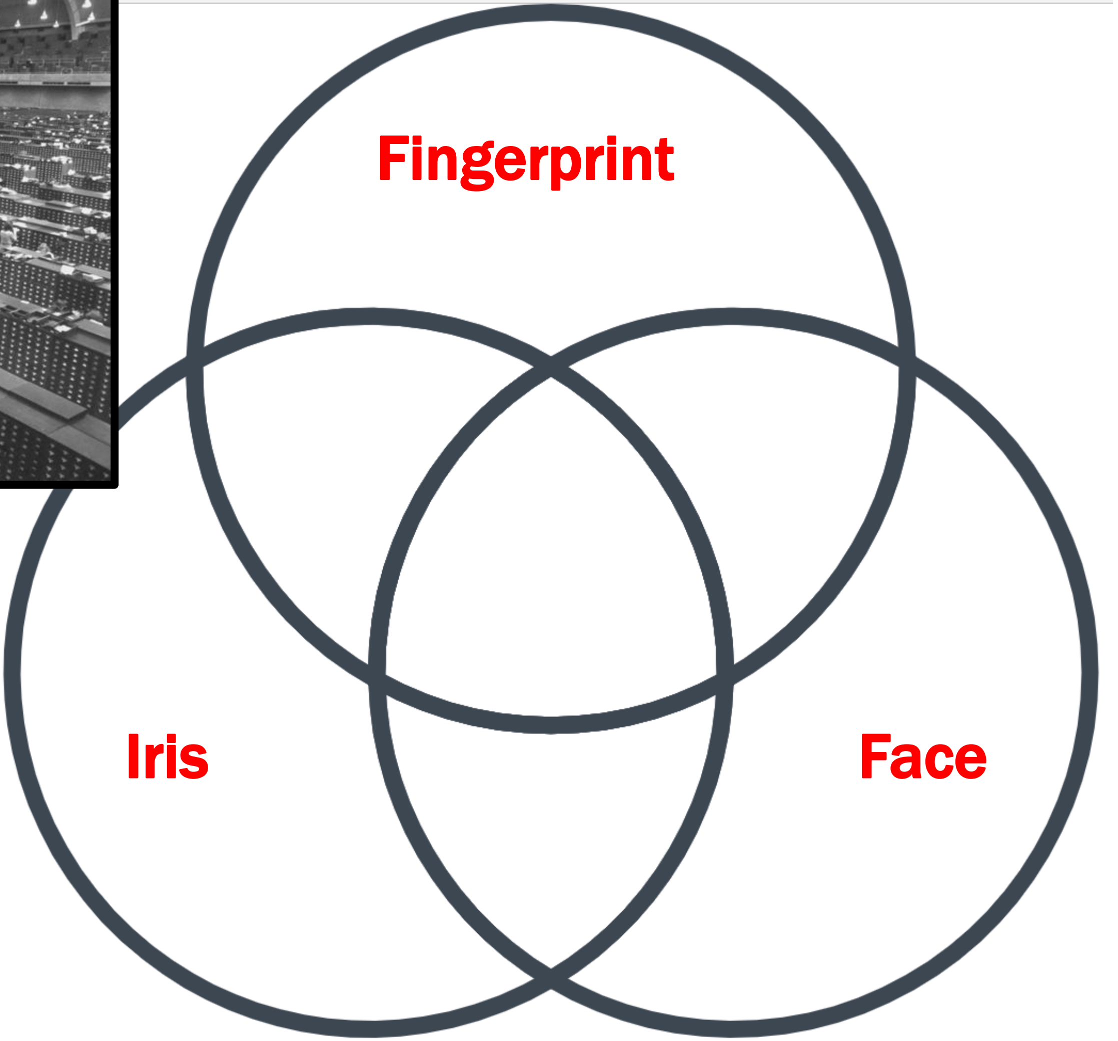
- Many sources:
  - **Most** people (and almost all computer scientists) will say **“the data”**
  - Far fewer people bring up:
    - Loss function
    - **Evaluation bias & historical anchoring**
    - **Our own brains**
      - **Projection bias** (we think machine ought to behave like us)
      - **Confirmation bias** (we like it when the machine confirms our beliefs)
      - **Automation bias** (we do what the machine tells us)

Evaluation Bias and Historical Anchoring

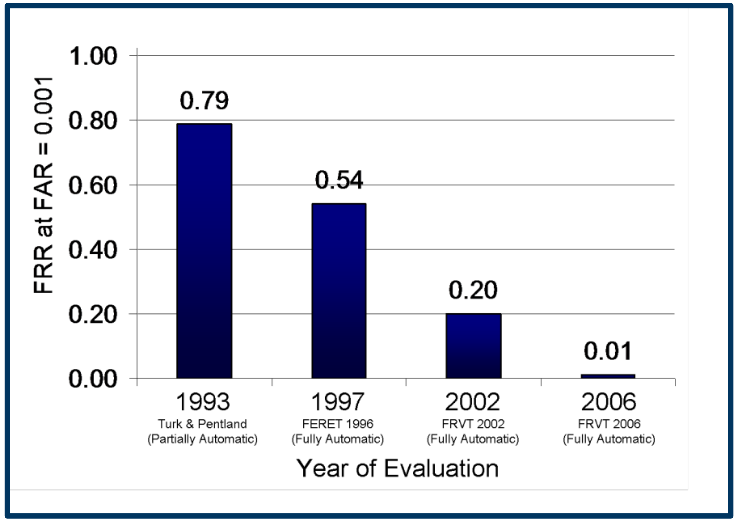


The **means** by which we evaluate fairness **impacts the outcome** of a fairness evaluation

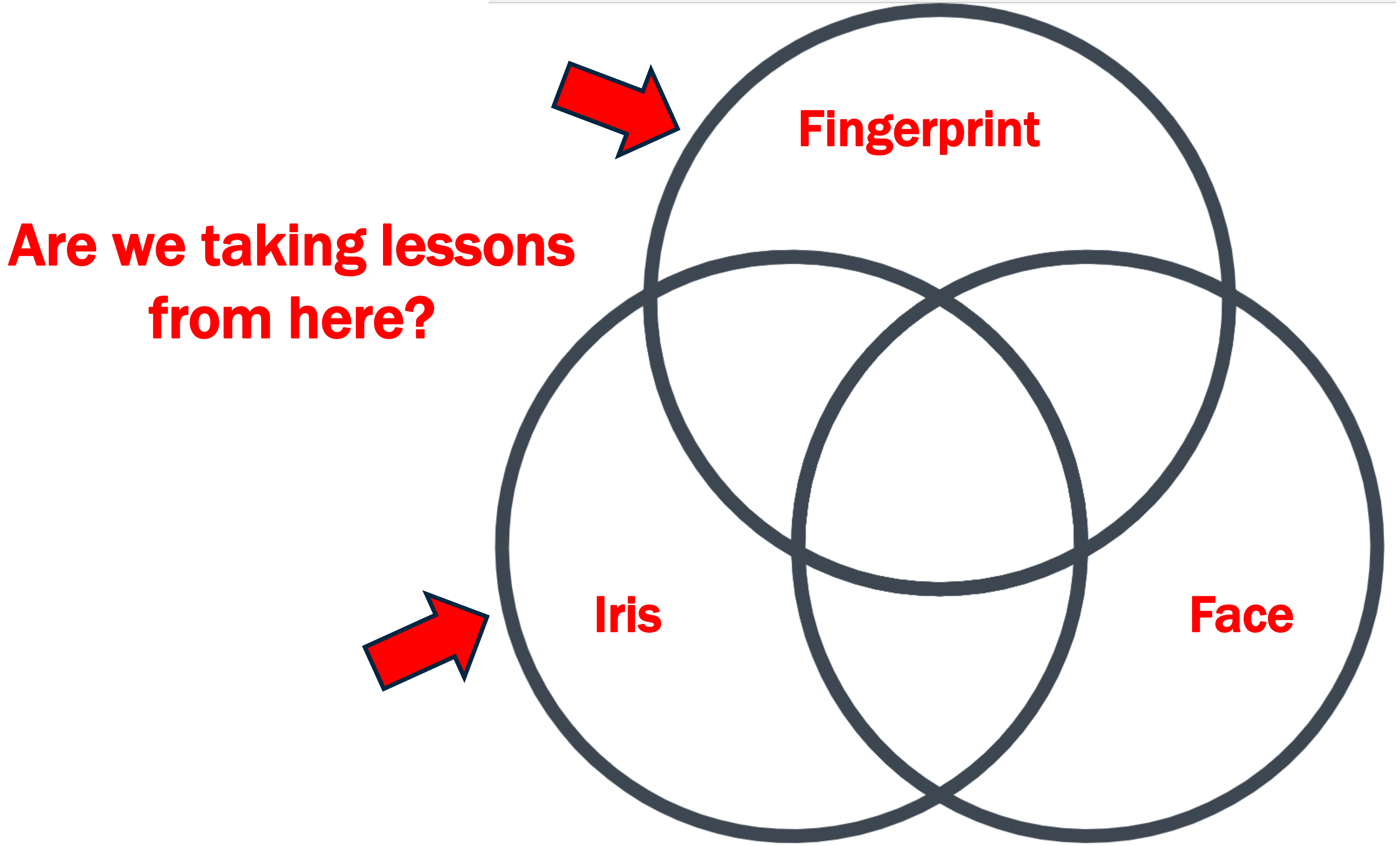
Evaluation Bias and Historical Anchoring



The **means** by which we evaluate fairness **impacts the outcome** of a fairness evaluation

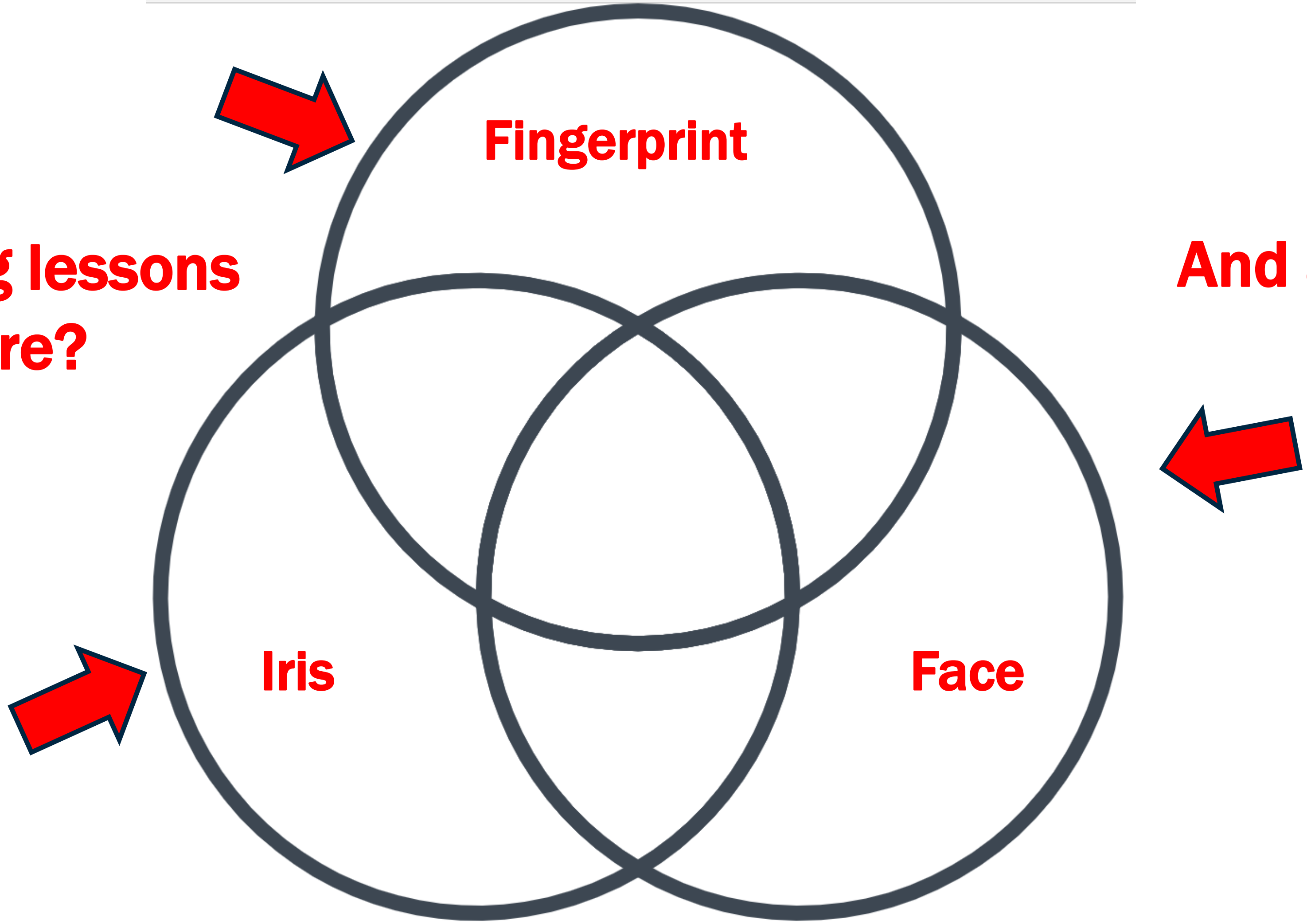


Evaluation Bias and Historical Anchoring



Evaluation Bias and Historical Anchoring

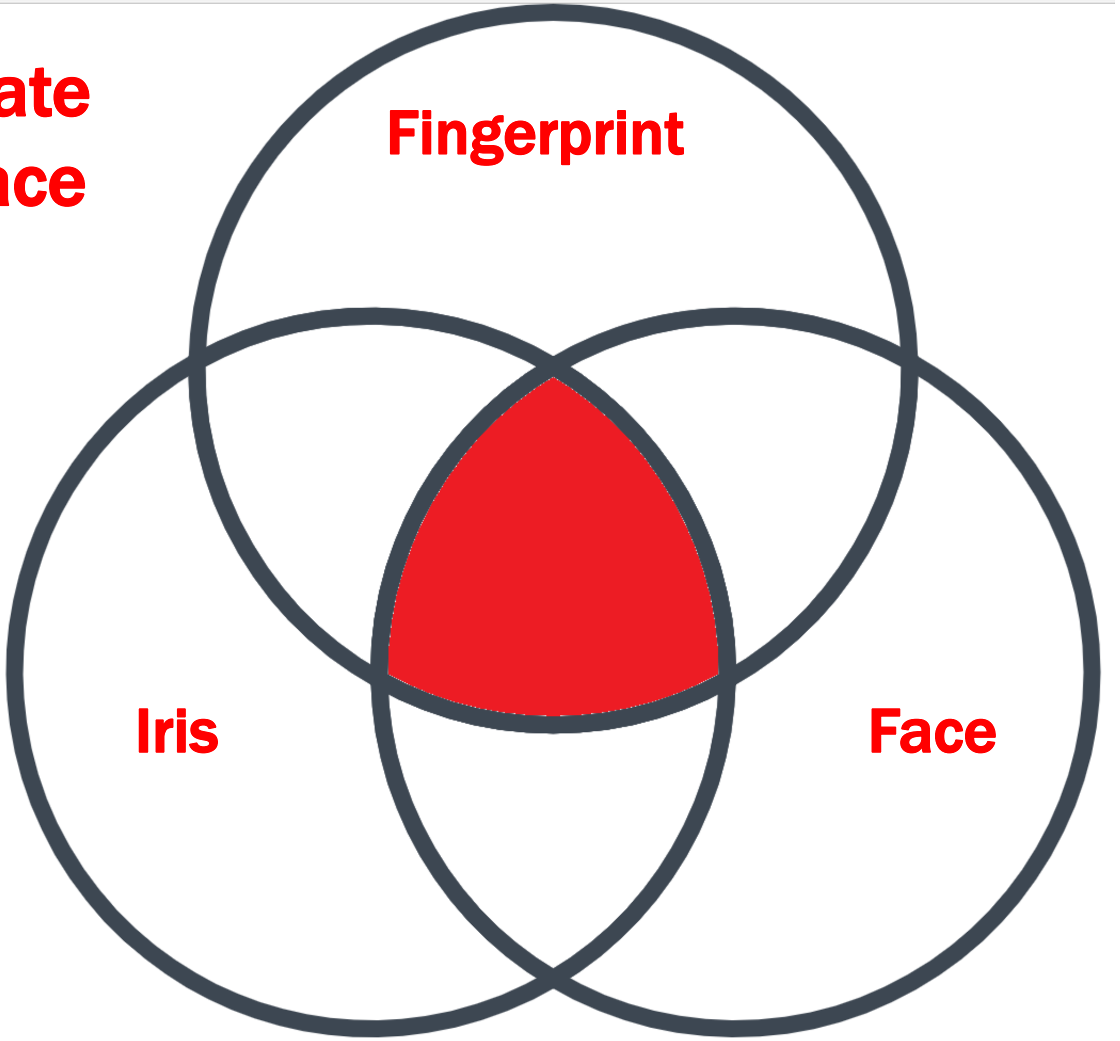
Are we taking lessons from here?



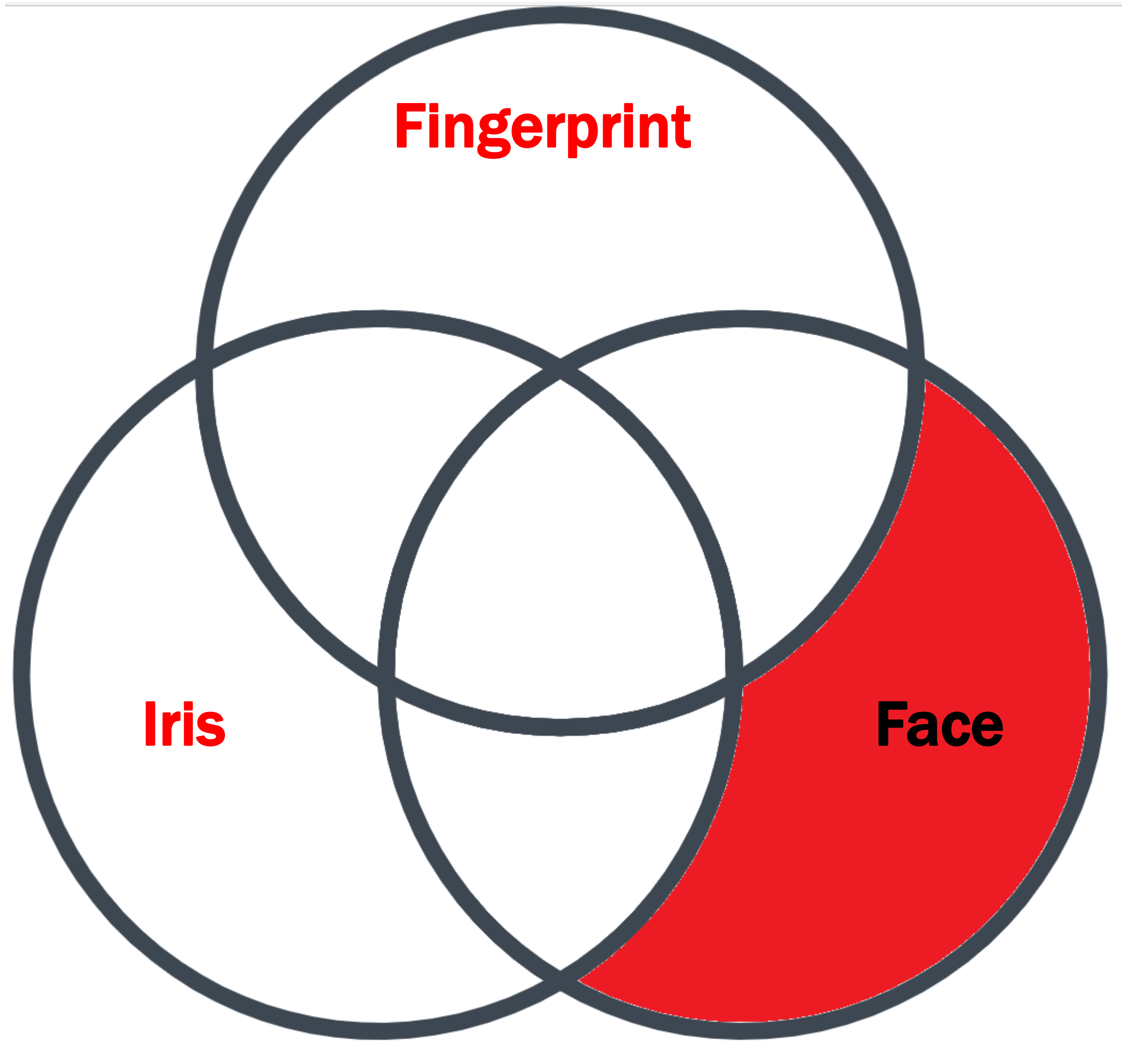
And applying them here?

Evaluation Bias and Historical Anchoring

May be appropriate because this space exists



Evaluation Bias and Historical Anchoring

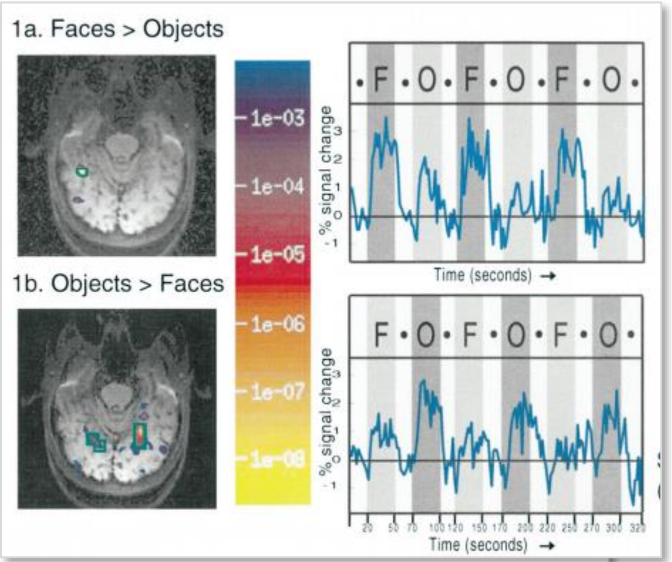


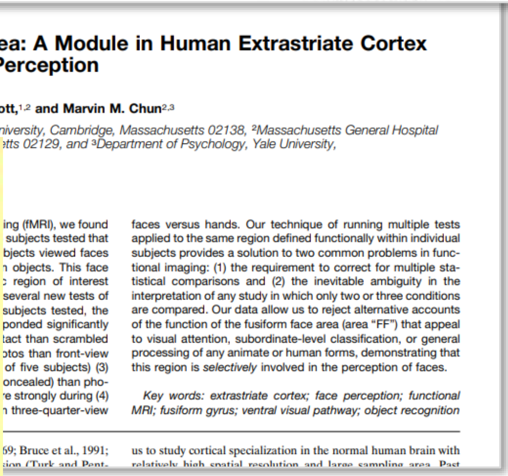
But we need to keep in mind that this space exists as well

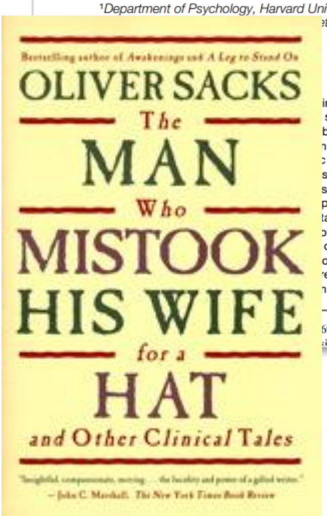
Faces are different for (at least) two reasons

- Faces are **genetic**, iris and fingerprint characteristics are determined during development.
  - Face are more alike for siblings, those with common ancestry, and those of the same sex

- 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”



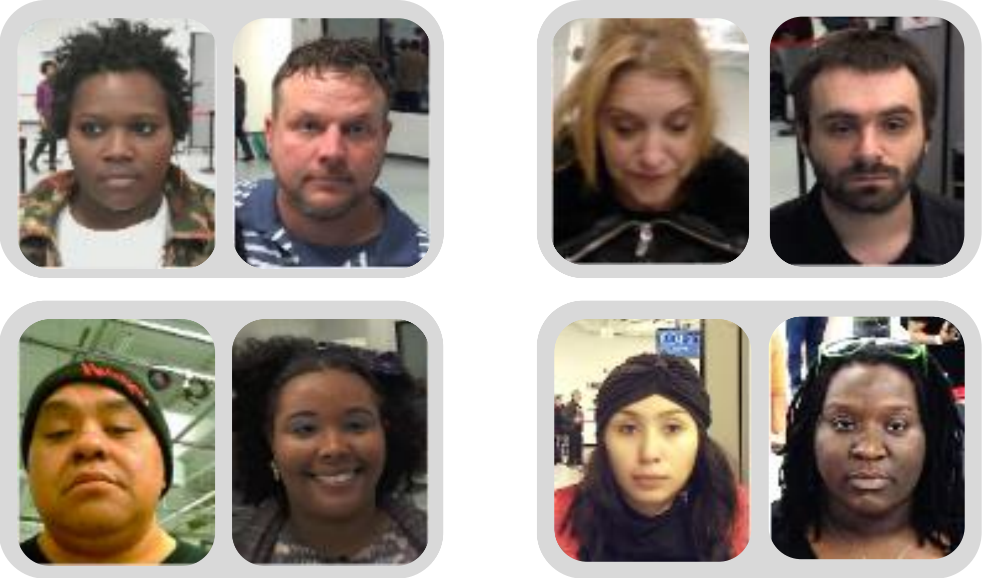




Demographic effects exist, our understanding of them may be clouded

> It may seem natural to us that FR “clusters” people based on race and gender (projection bias) <

Iris recognition



Iris recognition false positives were random relative to race and gender

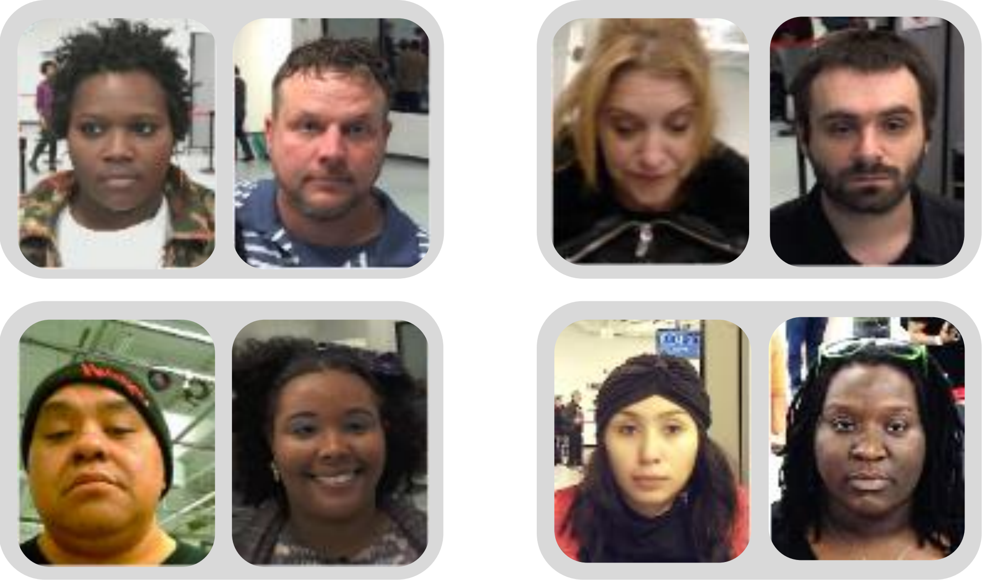
Face recognition

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

Demographic effects exist, our understanding of them may be clouded

> 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

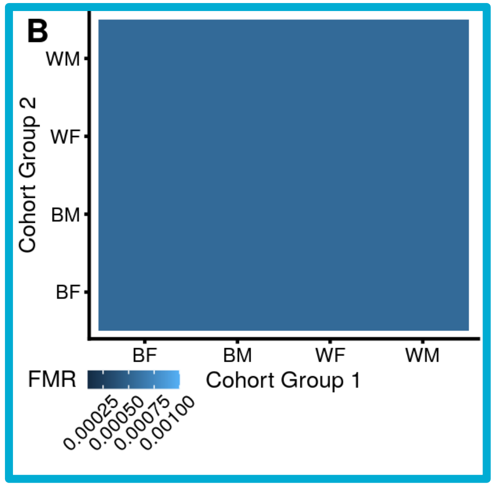
80% of face recognition false positives were between people of the same race and gender

> Because the errors on the left are unique to FR, FR has unique problems <

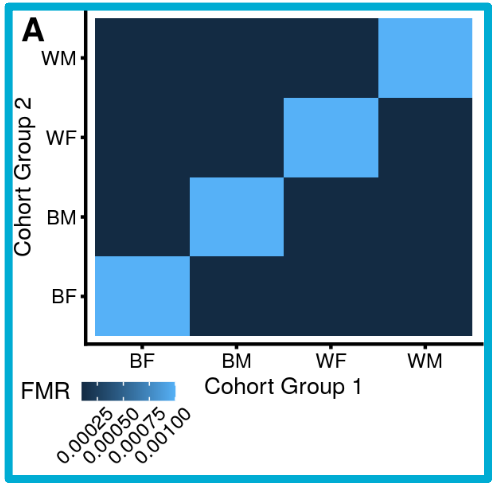
IDSL

Problem 1 – This can impact fairness in identification scenarios

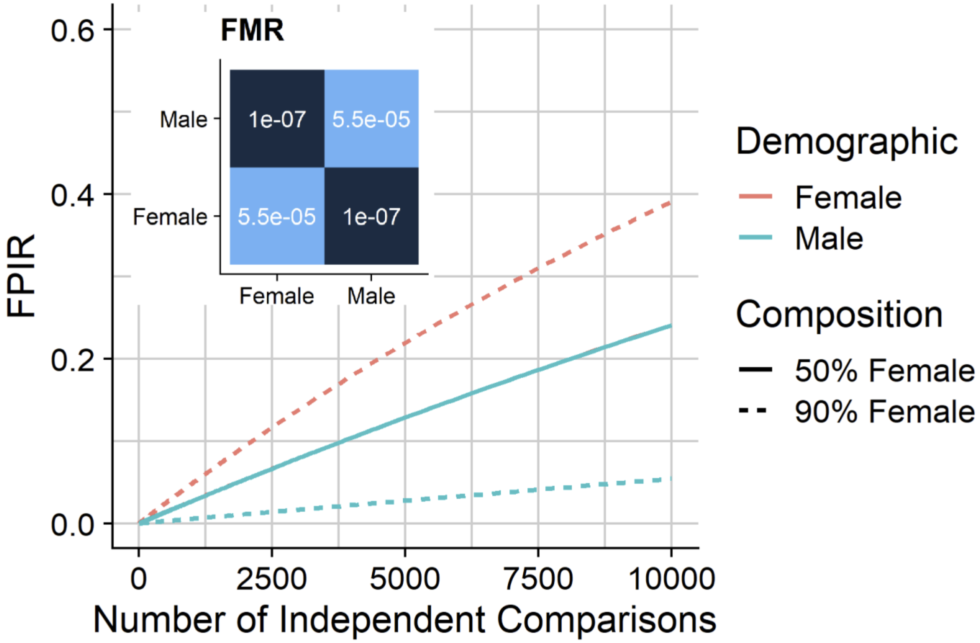
- The "watchlist imbalance effect"
  - Howard et. al (2021)
  - Drodowski et. al (2021)



False match cohort matrix for finger, iris, etc.



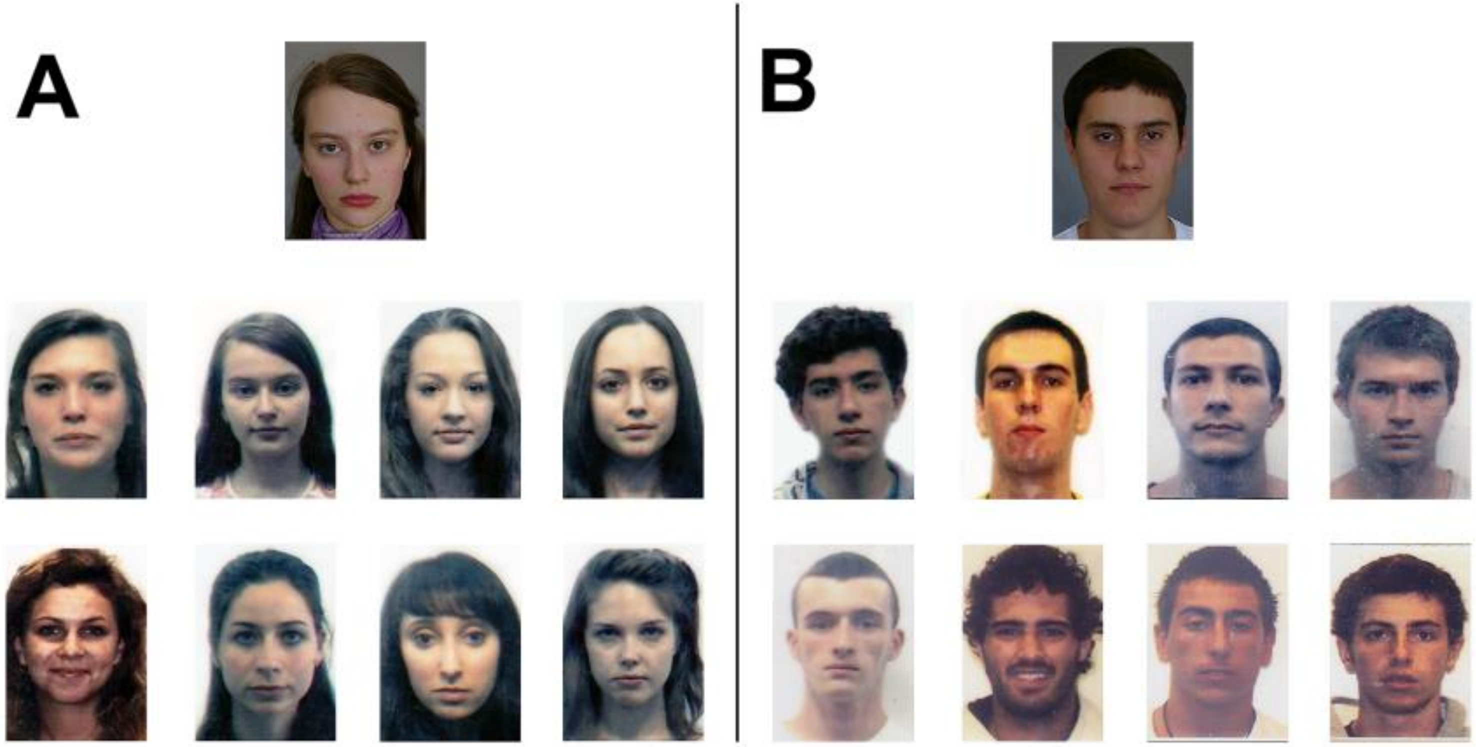
False match cohort matrix for face



- "broad homogeneity": if you have a watch-list gallery of majority female:
  - An innocent white female 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 2 – Errors like this make the human's job harder and slower



- White et. al "Error Rates in Users of Automatic Face Recognition Software" (2015)
- **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**

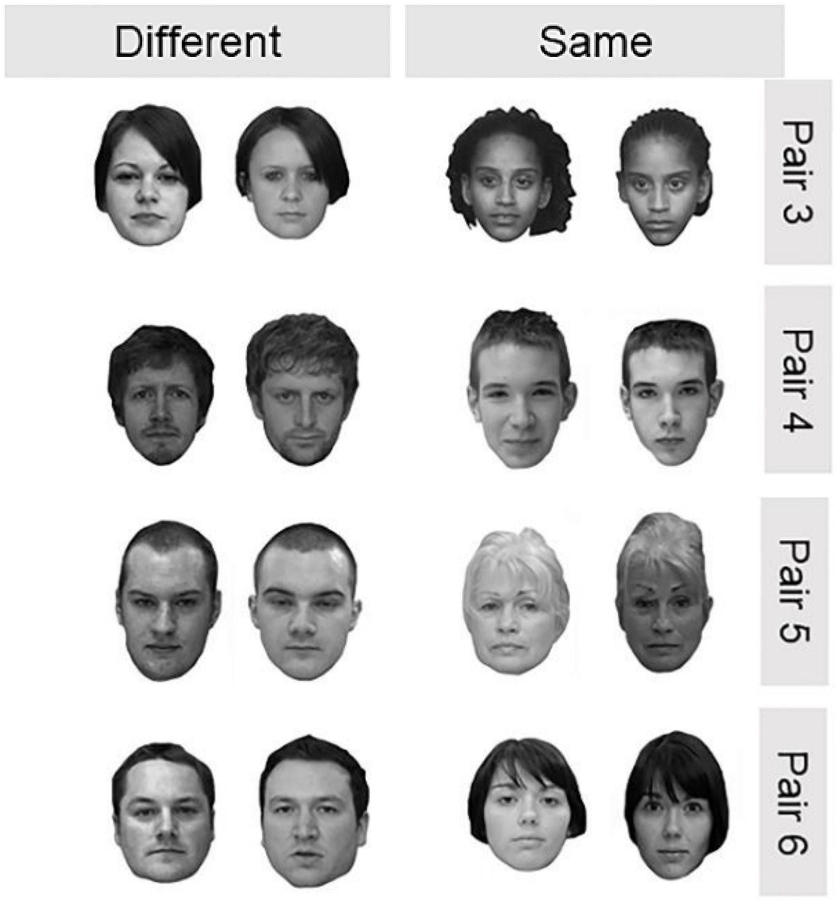
Problem 3 – Errors like this make us more susceptible to automation bias

- Howard, Rabbitt, Sirotin, *Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making.* PLoS 2020

- **343 volunteers performed face matching task (12 face pairs)**
  - Glasglow Face Matching Test (8 pairs)
  - **Select stimuli from MEDS for diversity in pairs (4 face pairs)**

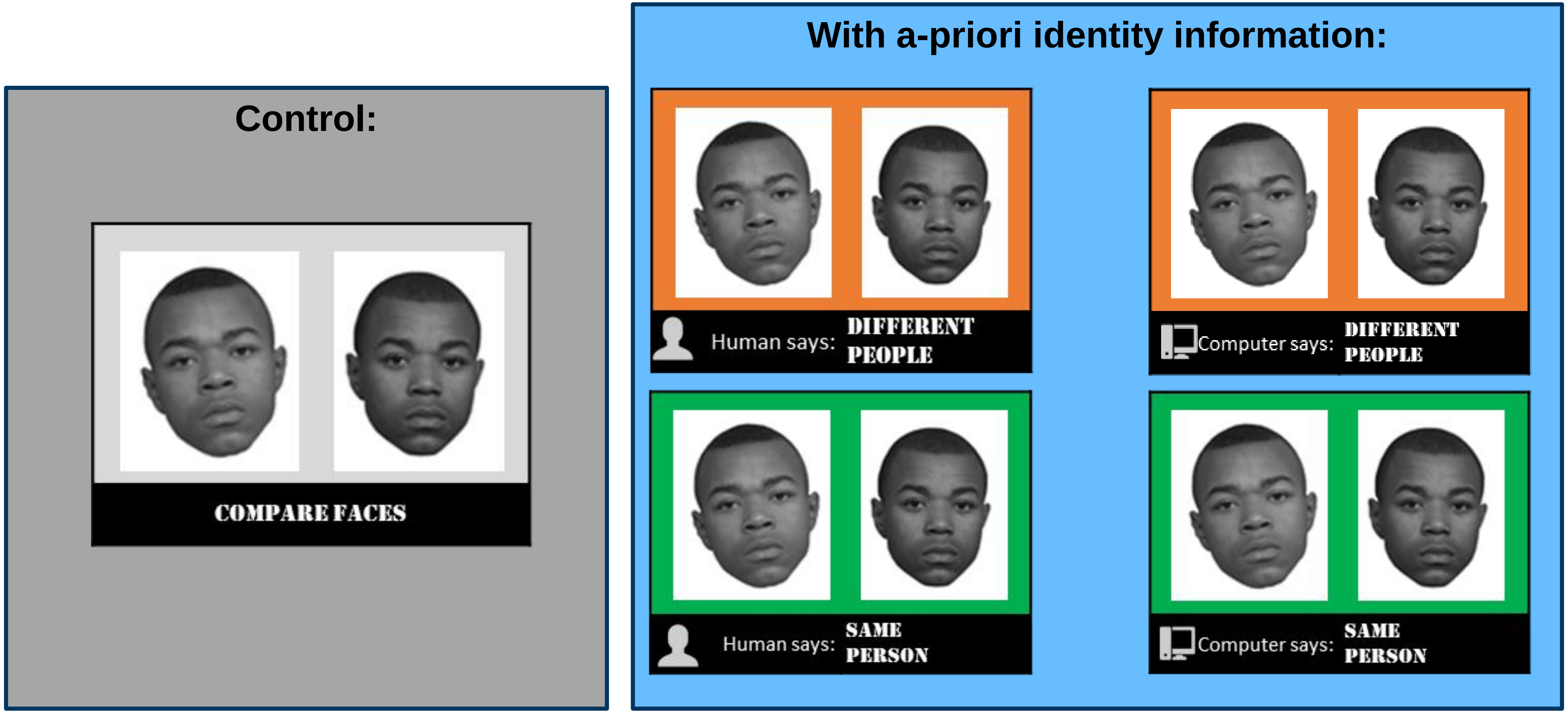
- Asked to rate similarity on a 7-point scale:

-3	**I am absolutely certain these are different people**
-2	**I am mostly certain these are different people**
-1	**I am somewhat certain this is the different person**
0	**I am not sure**
1	**I am somewhat certain these are same people**
2	**I am mostly certain this is the same person**
3	**I am absolutely certain this is the same person**



Automation Bias in FR

- Subjects were given face pairs under two conditions



Automation Bias in FR

- At a threshold of 0.5

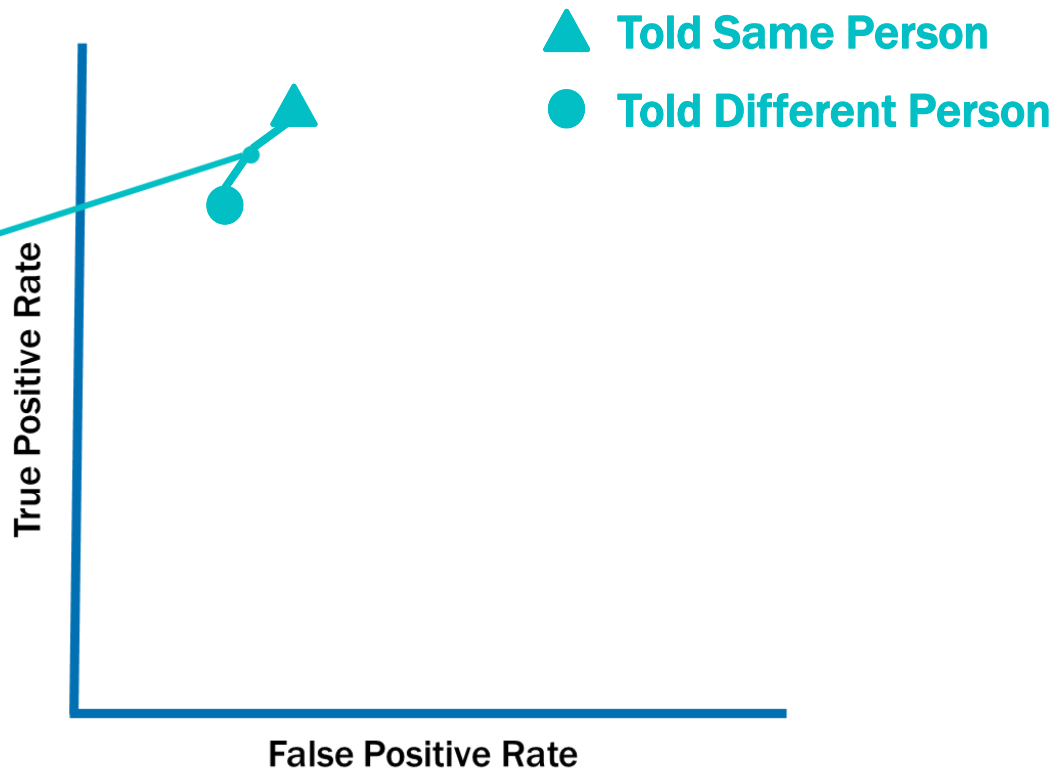
No Match	-3	I am absolutely certain these are different people
	-2	I am mostly certain these are different people
	-1	I am somewhat certain this is the different person
	0	I am not sure
Match	1	I am somewhat certain these are same people
	2	I am mostly certain this is the same person
	3	I am absolutely certain this is the same person

Source	N	Accuracy	FPR	TPR
Control	120	0.75	0.19	0.70
Same	223	0.73	0.25	0.72
Different	223	0.75	0.17	0.66

Automation Bias in FR

- At the threshold of 0.5:

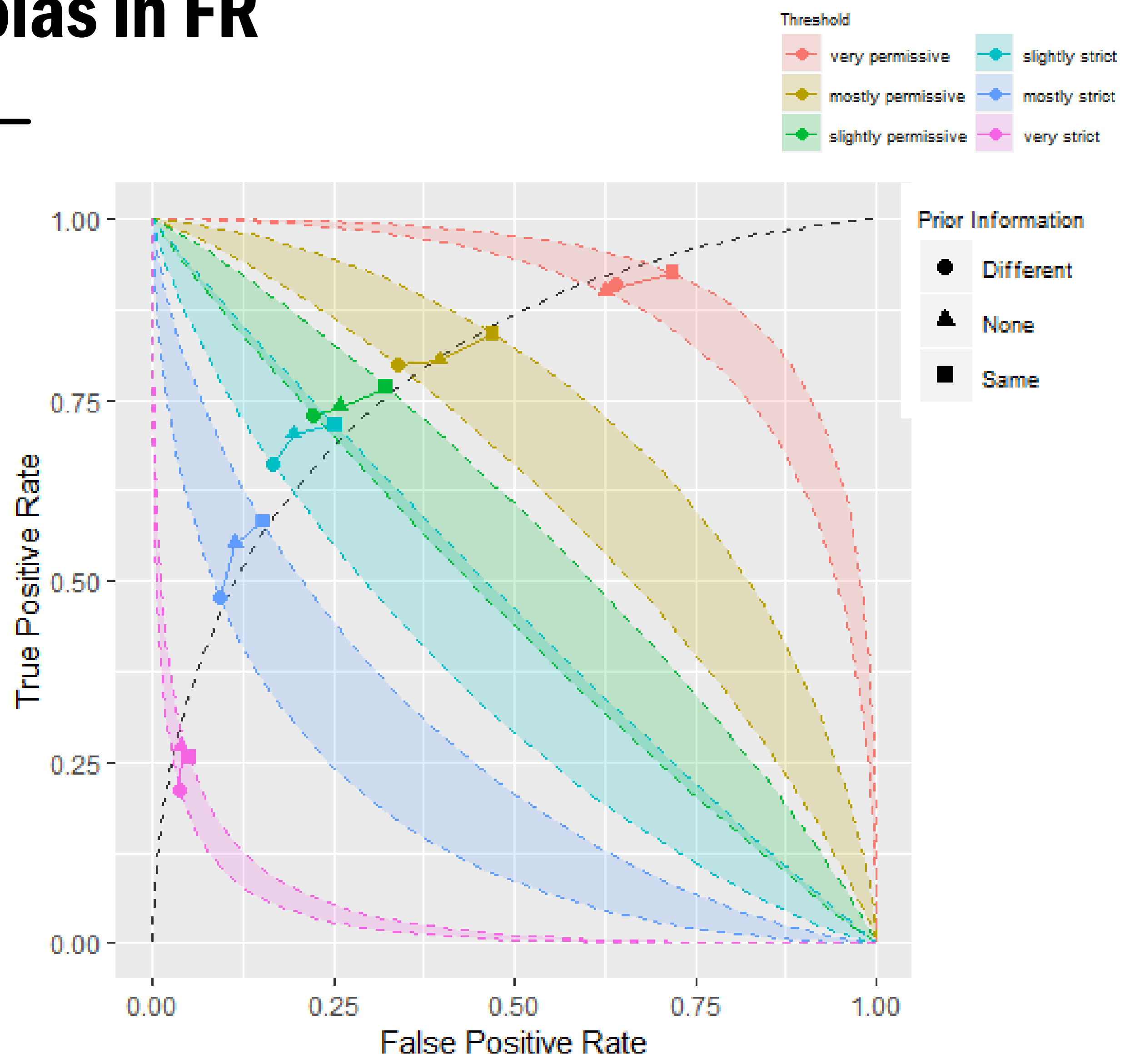
-3	I am absolutely certain these are different people
-2	I am mostly certain these are different people
-1	I am somewhat certain this is the different person
0	I am not sure
1	I am somewhat certain these are same people
2	I am mostly certain this is the same person
3	I am absolutely certain this is the same person



Source	FPR	TPR
Control	0.19	0.70
Same	0.25	0.72
Different	0.17	0.66

Automation bias in FR

- Across thresholds:
- The overlap in middling threshold indicates **prior identity information** can shift responses by a whole step
  - I am not sure → I am somewhat sure
- But only for **challenging** face pairs (I am not sure)
- Prior identity information effect was present but modest
- Humans mostly trusted their own abilities (**under ideal conditions**)



But what about when FR is hard?

- Barragan, Howard, Rabbitt, Sirotin. *COVID-19 Masks Increase The Influence of Face Recognition Algorithm Decisions on Human Decisions in Unfamiliar Face Matching.* PLoS 2022

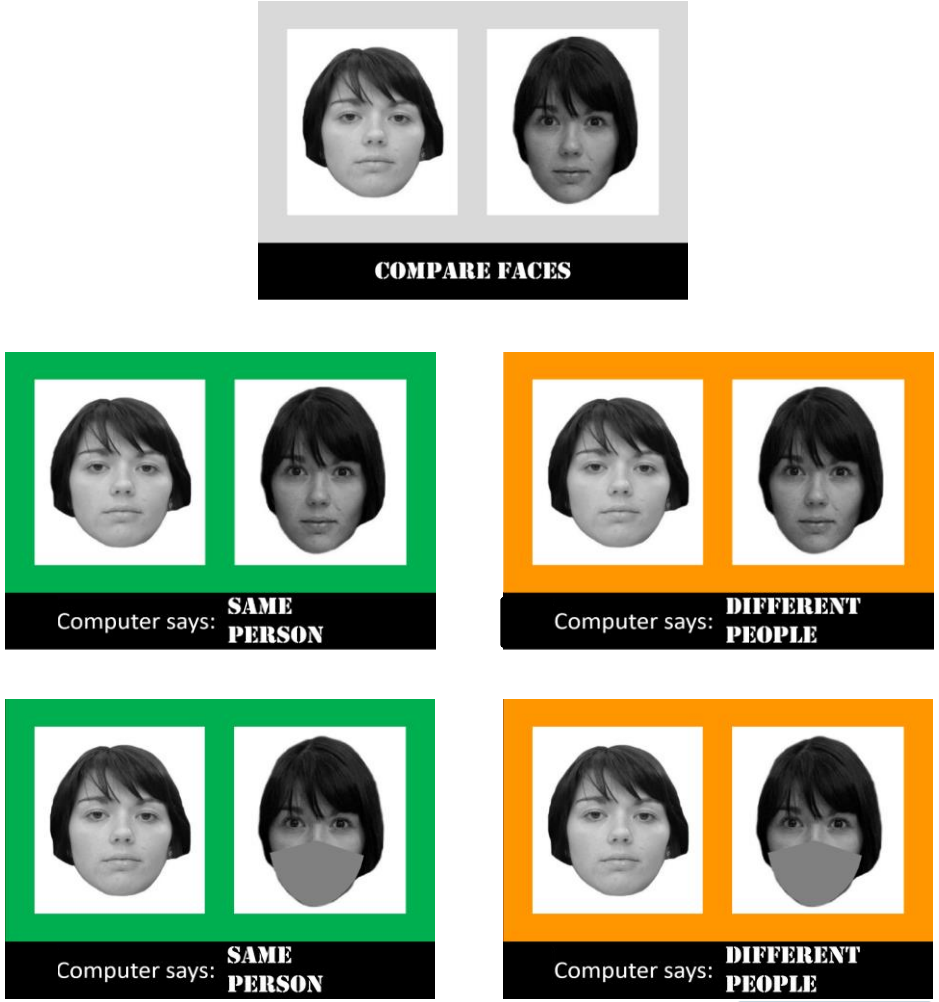
But what about when FR is hard?

- **Barragan, Howard, Rabbitt, Sirotin. *COVID-19 Masks Increase The Influence of Face Recognition Algorithm Decisions on Human Decisions in Unfamiliar Face Matching.* PLoS 2022**

Control

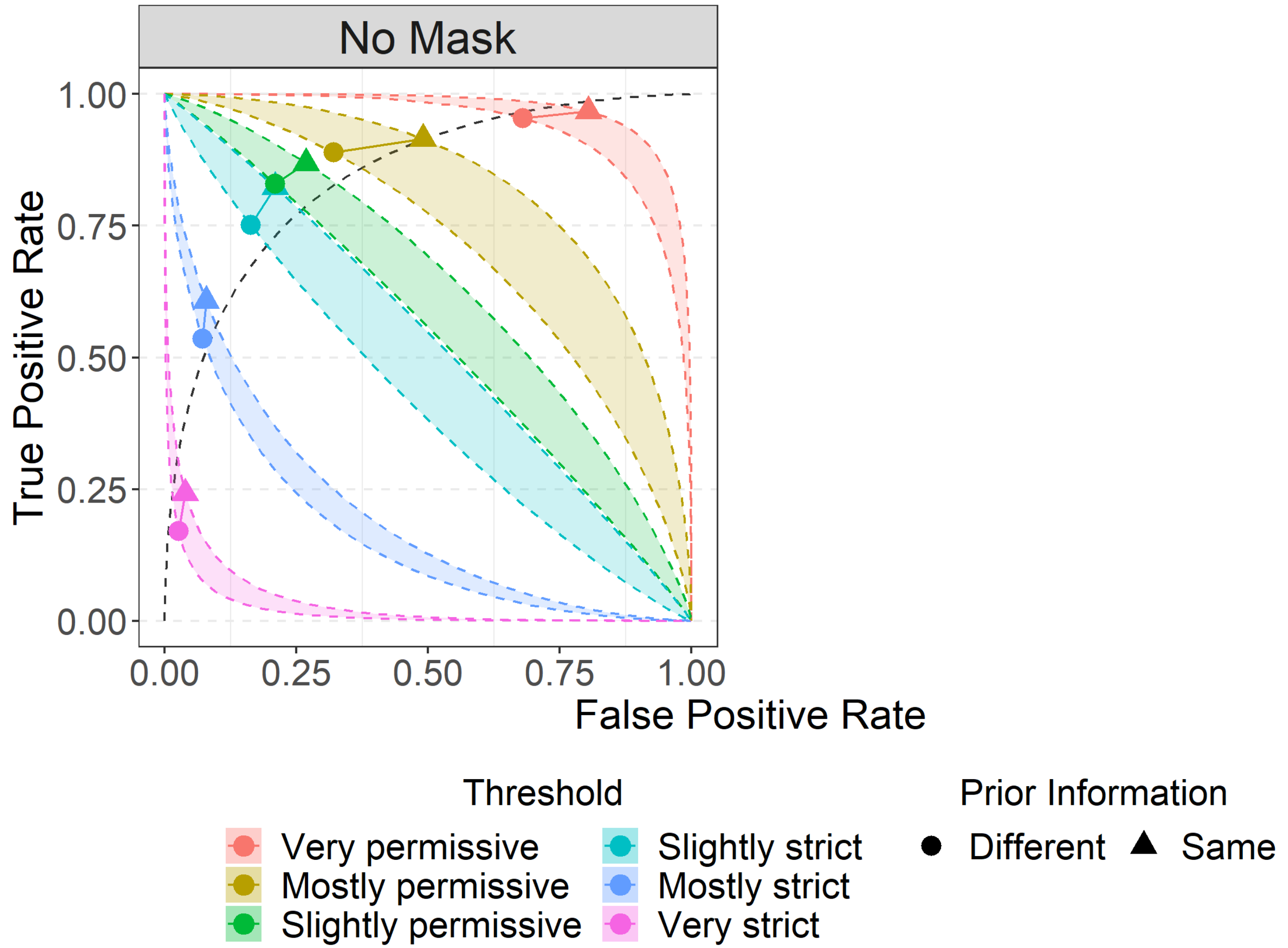
Computer-No Mask

Computer-Mask



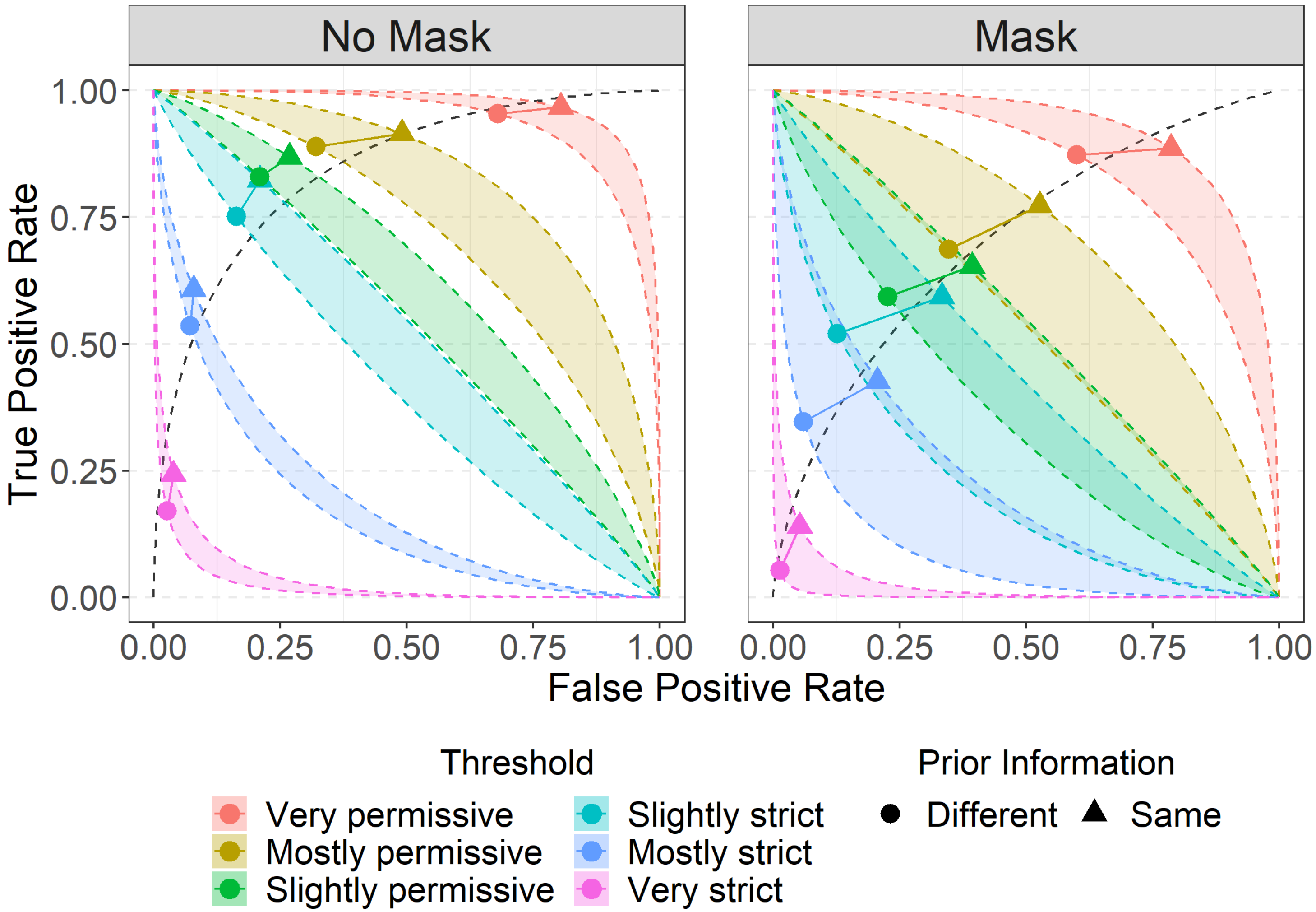
Automation Bias in FR (when it's hard)

- 150 test subjects
- Largely replicated 2020 "No Mask" study



Automation Bias in FR (when it's hard)

- **150 test subjects**
- **Largely replicated 2020 "No Mask" study**
- **However, the presence of masks greatly increased the influence of the prior algorithm information**
- **It also reduced accuracy 10-20%**



Automation Bias in FR (when it's hard)

- Our results showed that masks increased human reliance on algorithm determinations (if presented)

- Its likely (in our minds) that this is true for many factors that <u>increase difficulty</u> in face recognition tasks:
  - True across many categories of socio-technical systems (Google maps effect)
  - Lack of information in the image due to pose, blur, lighting etc.
  - Human perceived similarity **demographic homogeneity**

Agenda

- ~~The Maryland Test Facility~~

- Demographic differentials or “bias” in Face Recognition:
  - ~~What is it?~~
  - ~~Where does it come from?~~
  - ~~Why are they bad?~~
  - How do we measure it?
  - How do we fix it?

How do we Measure Demographic Differentials

- Remember, these two things are both called a "false match error" in biometric parlance

Two people who share a similar **iris pattern** (according to an algorithm)

Two people who share a similar **face pattern** (according to an algorithm)

- But the **homogenous** pair is **more severe** because:
  - It can impact **fairness** in large identification scenarios
  - Its harder for a **human to adjudicate**
  - It makes humans more susceptible to **automation bias**

IDSL

Broad Homogeneity – A Note on Prevalence

- We coined the term “broad homogeneity” to describe this sameness effect in face recognition in 2019

Different Demographics

Same Demographics

More similar

Algorithm score

Less similar

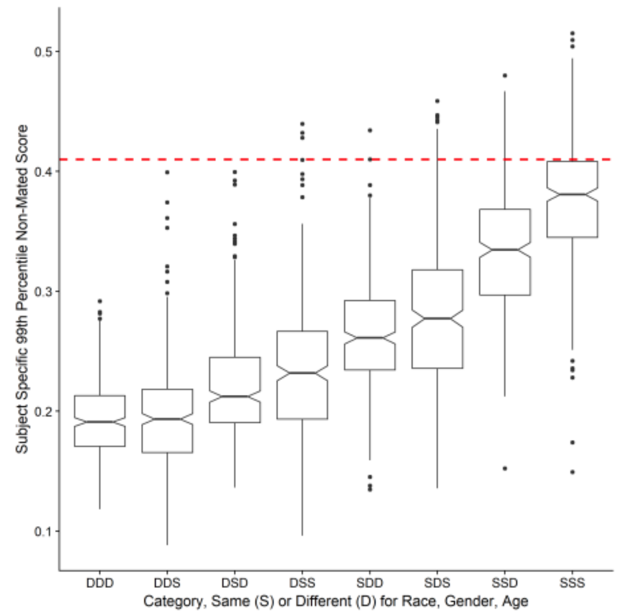
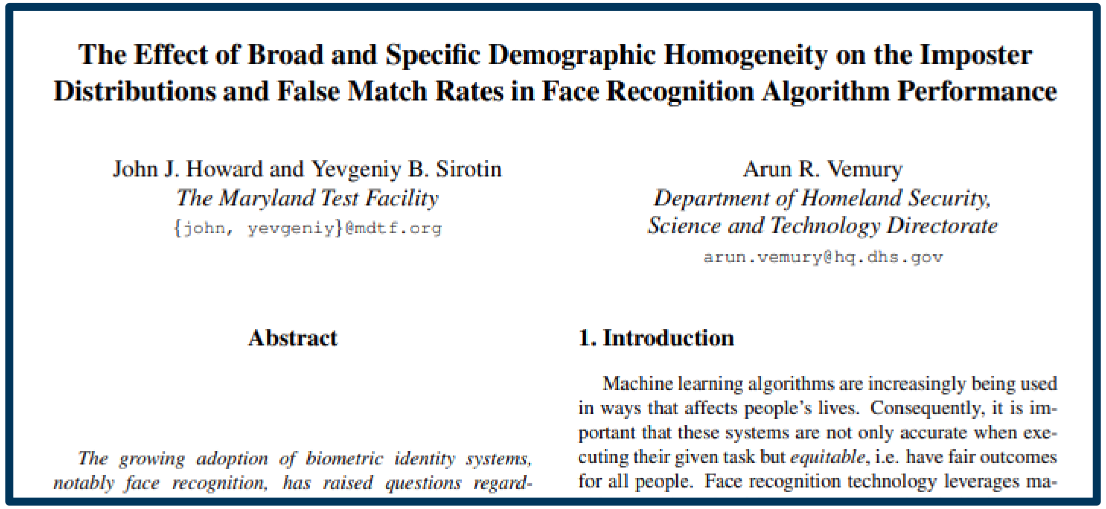


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.



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

The growing adoption of biometric identity systems, notably face recognition, has raised questions 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-

- We show this effect existed in one commercial face recognition algorithm

- Not present in iris or fingerprint biometrics

This is (Likely) (Currently) a Universal Feature of Face Recognition

- NIST subsequently confirmed this exists in **all 138 algorithms** submitted to FRVT in 2019.
  - NIST FRVT Part 3: Demographics – Annex 5.

Higher (non-mate) similarity score

More similar demographics

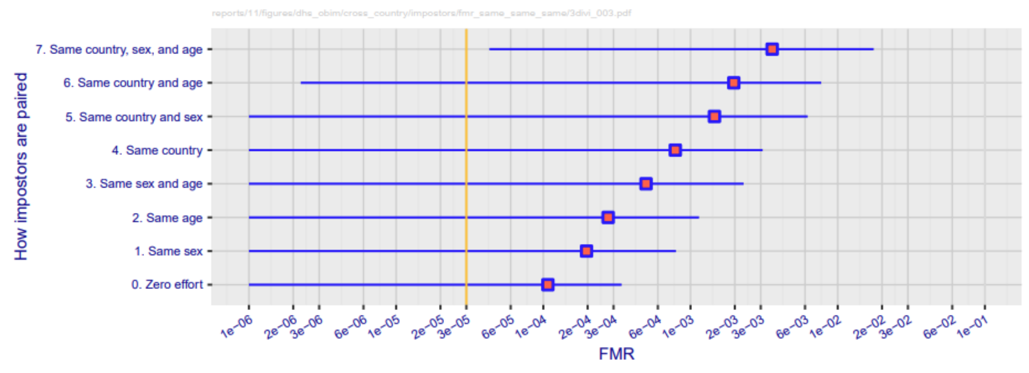


Figure 1: FMR for increasing matched covariates, 3divi-003

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

uction

learning algorithms are increasingly being used
affects people's lives. Consequently, it is im-
these systems are not only accurate when exe-
given task but *equitable*, i.e. have fair outcomes
le. Face recognition technology leverages ma-

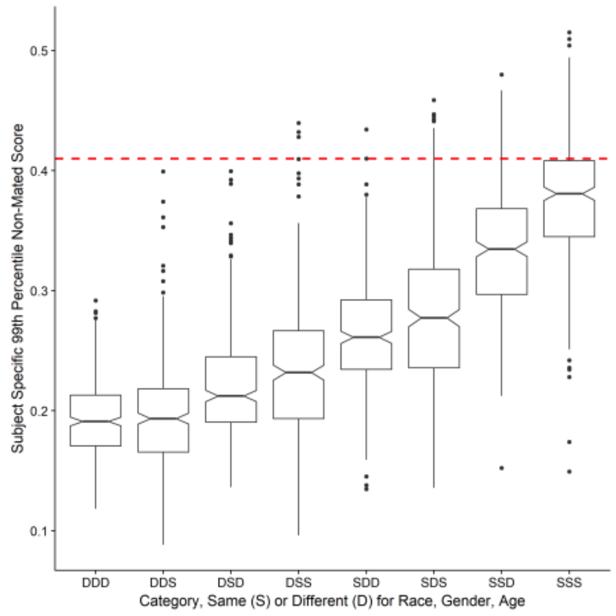


Figure 4. Distributions of the 99th percentile subject-specific non-mated scores across broad homogeneous versus heterogeneous race, gender, and age categories.

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

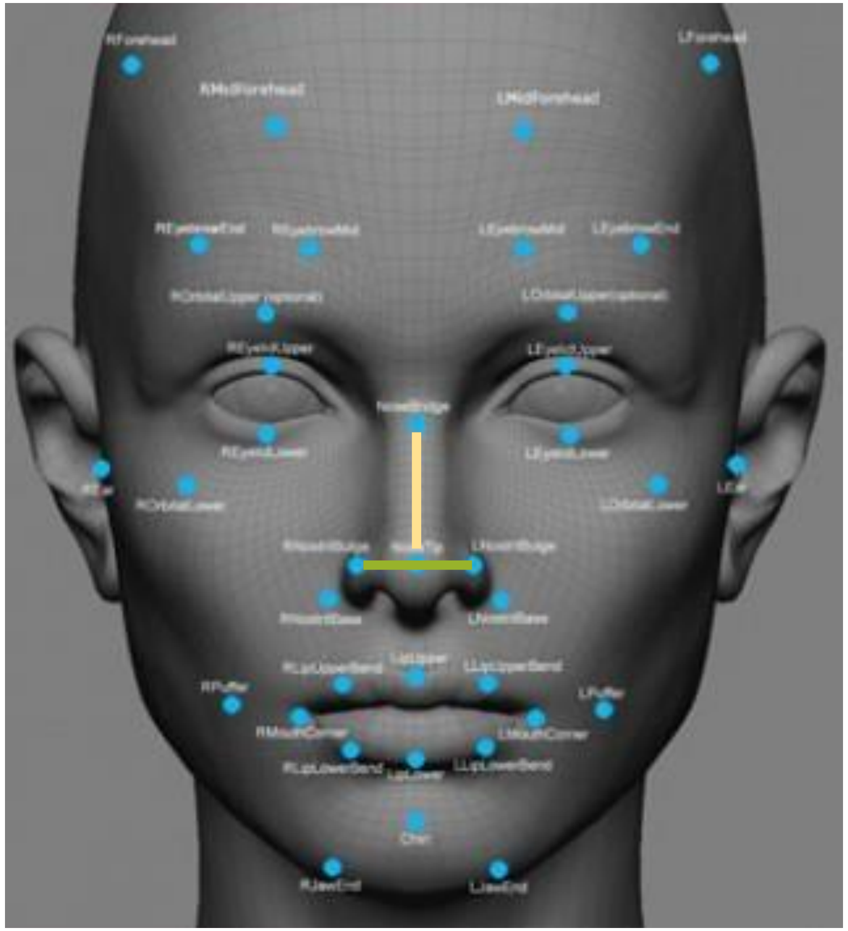
The Maryland Test Facility, Identity and Data Sciences Lab

Arun R. Vemury

The U.S. Department of Homeland Security

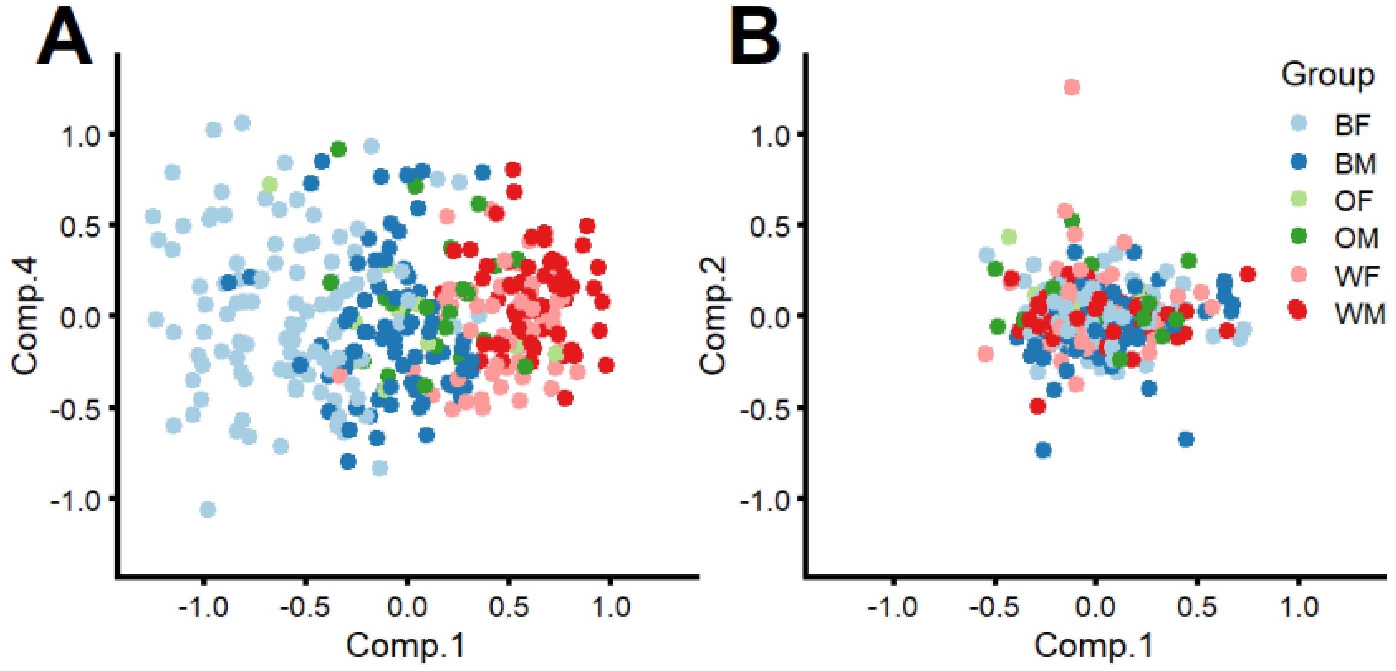
Face Information Content?

- There are many detectable **points** on the human face
- The distances, shapes, and contours formed by those points make up some of the **face information** used by face recognition algorithms
- Some of that information content (but not all) **can cluster** people by ancestry, gender, etc.
- For example, male noses are on average **shorter** and **broader** than female noses



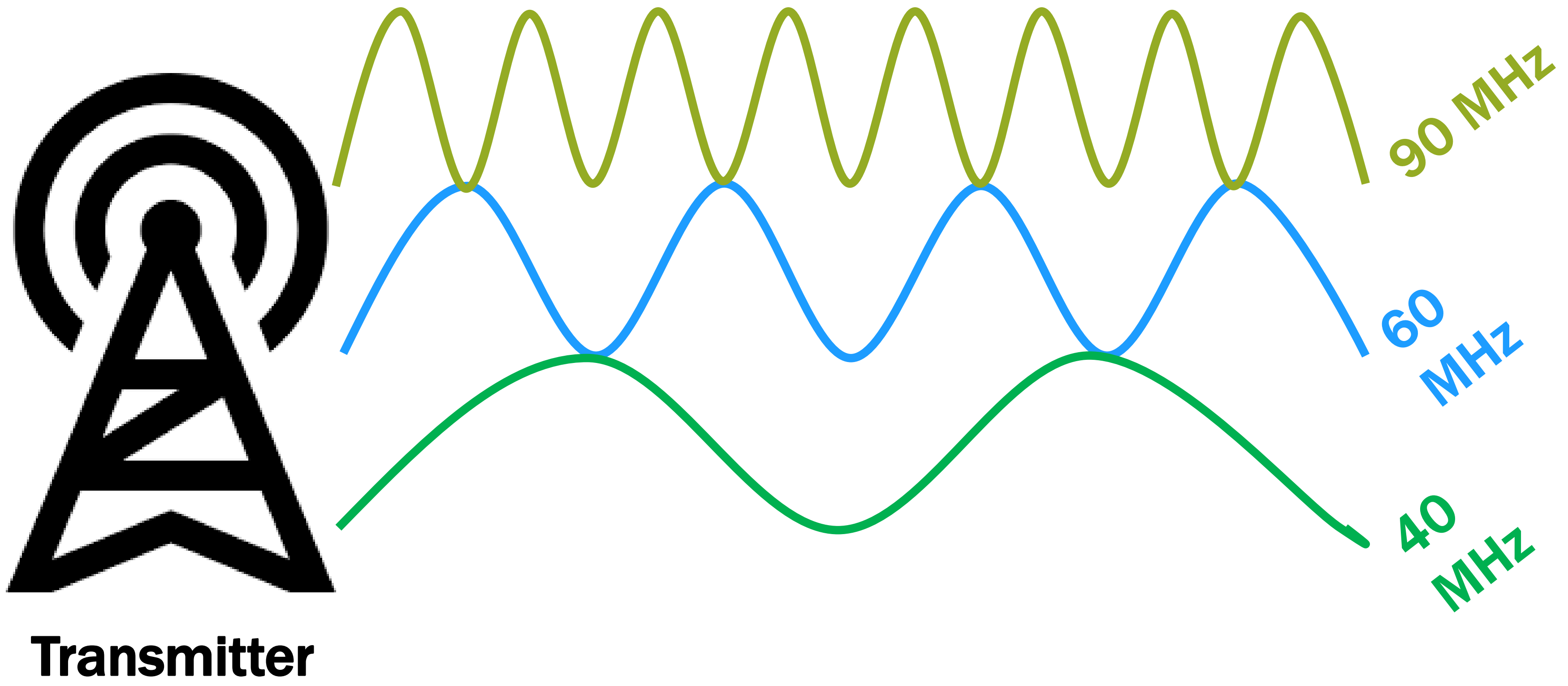
Face Information Content?

- We can visualize this clustering
- And measure it across many types of face information
- To find components that cluster (Comp.1, plot A)*
- And those that don't (Comp.3, plot B)*

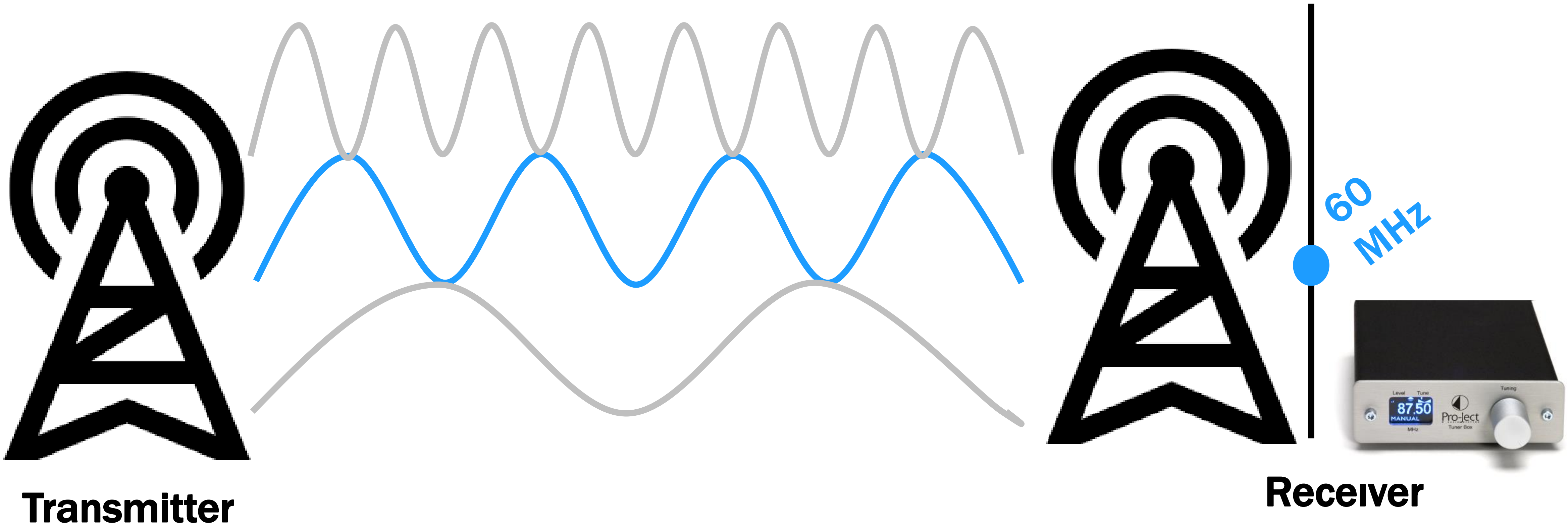


* Howard, Sirotin, Tipton, Vemury. *Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms*. DHS Technical Paper Series 2020.

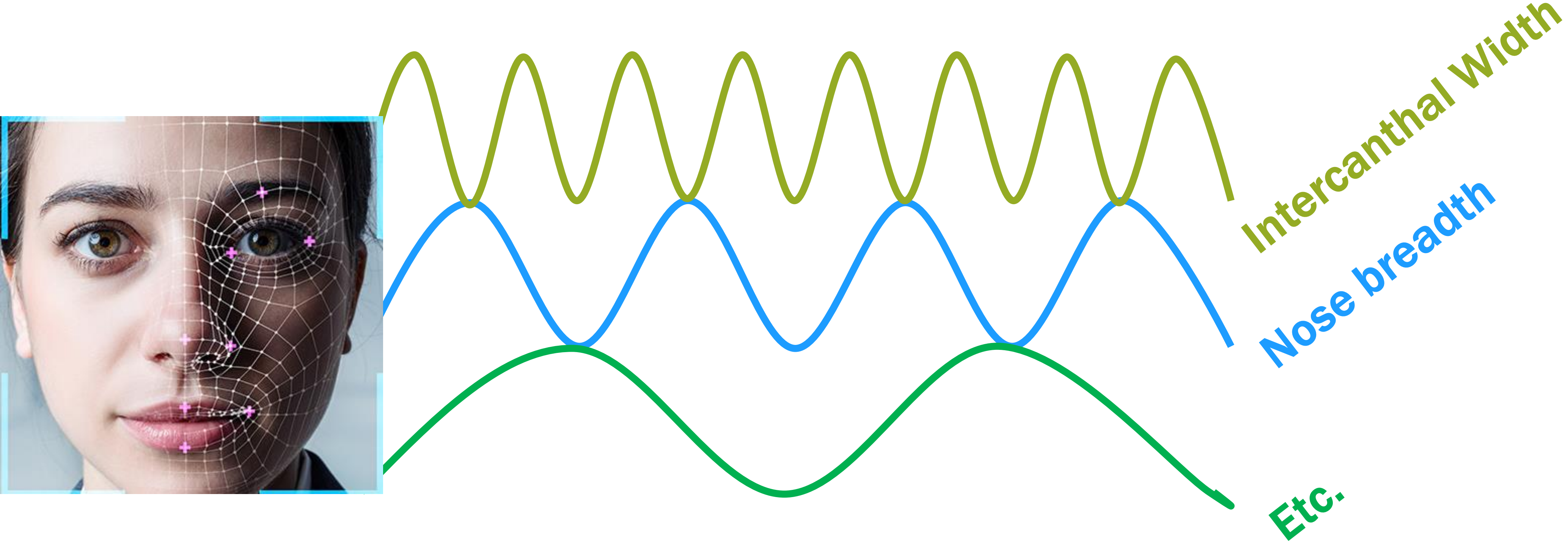
Selecting Face Information Content



Selecting Face Information Content

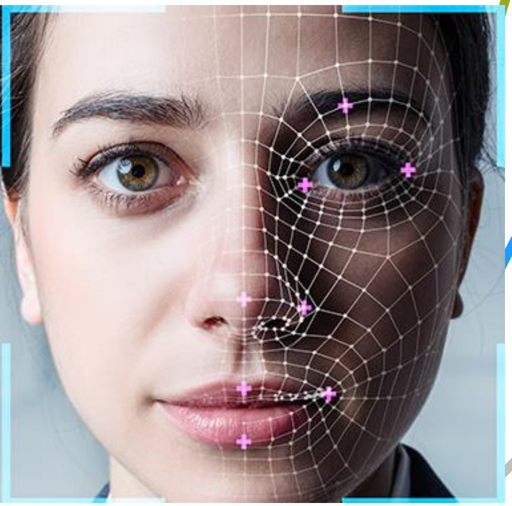


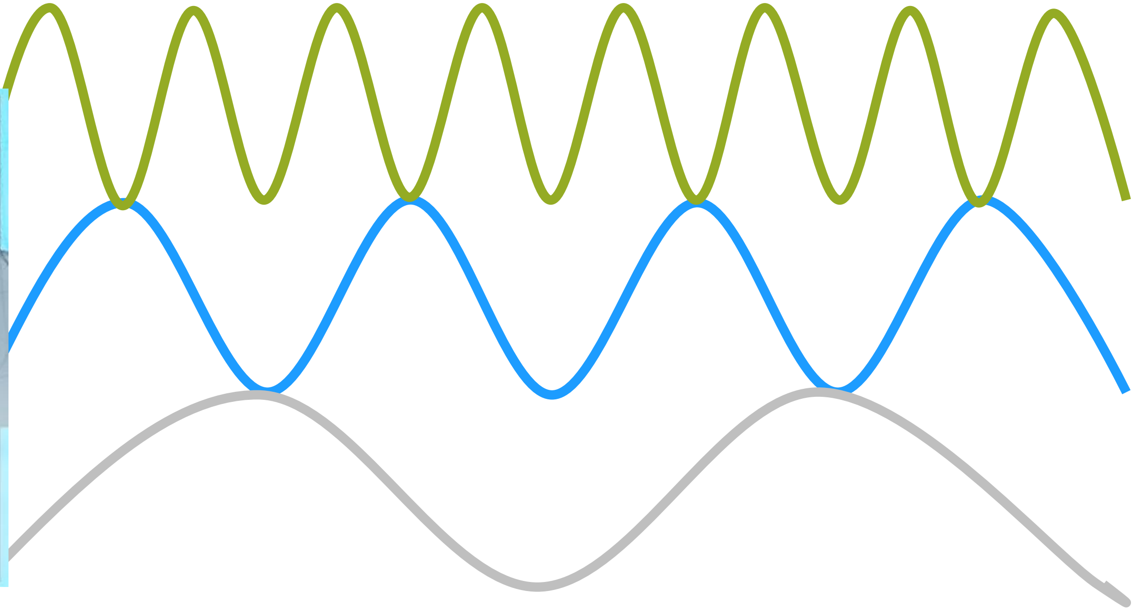
Selecting Face Information Content

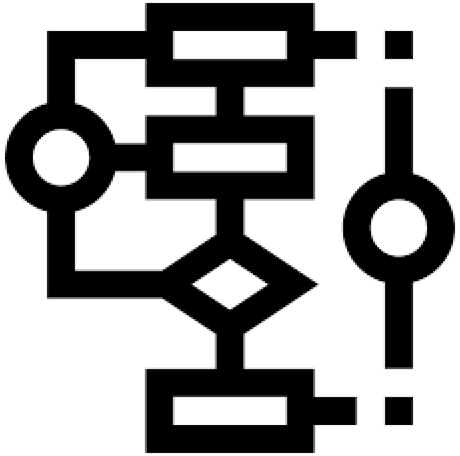


Human Face

Selecting Face Information Content







Non-clustering Face Features

Human Face

Face Recognition Algorithm

But There May Be Solutions

- 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

- 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
Department 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-

IDSL

What data did we use?

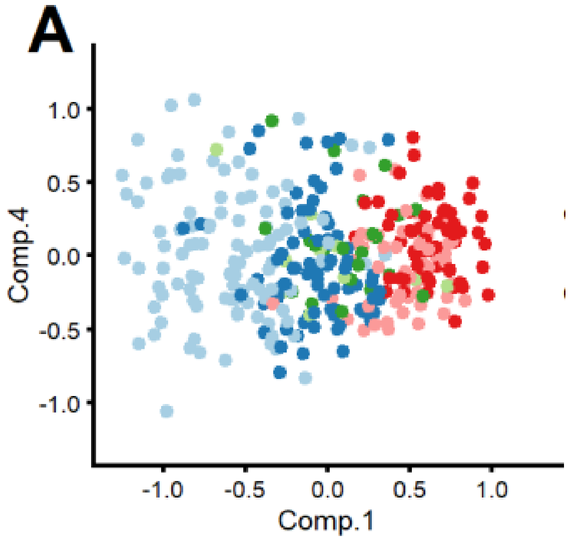
- Data
  - Three of face samples collected from the 2018-200 Biometric Technology Rallies:
    - S1 – demographically balanced training set
    - S2 – disjoint test set
    - S3 – mated pairs to subjects in S1
  - Two algorithms
    - ArcFace pre-trained on MS-Celeb-1M
    - ArcFace pre-trained on Glint 360k
  - Requirement for white box template structures

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)

What did we do?

- **Goal:** Given a matrix V of face recognition **feature vectors**, identify components of those vectors that exhibit demographic clustering.

- Process (high level, details in the paper):
  - SVD on normalized features
  - Calculate clustering index
  - Identify components with significant clutering
  - Remove via a **de-clustering transform** $\widehat{W}\widehat{W}^T$



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

What did we do?

- Experiment 1: apply $\widehat{\mathrm{W}}\widehat{\mathrm{W}}^{\mathrm{T}}$ to the **same feature vectors** it was learned on
  - $\dot{V} = V\widehat{W}\widehat{W}^T$
  - Learned and applied de-clustering transform on S1
  - **Q1**: How demographically "fair" are comparison scores generated from $\dot{V}$ versus V ?

- Experiment 2: $\widehat{\mathrm{W}}\widehat{\mathrm{W}}^{\mathrm{T}}$ to the **arbitrary feature vectors** (from the same algorithm)
  - $\dot{v} = v\widehat{W}\widehat{W}^T$
  - Learned declustering transform on S1 and applied to S2
  - **Q2**: If we learn features that exhibit demographic clustering on one set of subjects, do those same featured cluster on other subjects?

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)

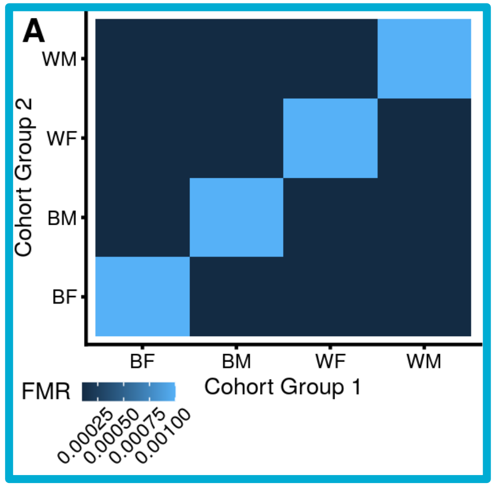
Algorithm	Fairness Metric	Experiment 1		Experiment 2	
		S1 Original	S1 Transformed	S2 Original	S2 Transformed
ArcFace-MS1MV2	Net Clustering	0.0163	**0.00549**	0.0252	**0.0207**
	GARBE	0.8540	**0.65000**	0.922	**0.909**
	FDR	0.9900	**0.99900**	0.991	**0.993**
	INEQ	219.00	**30.2000**	22.00	**18.00**
	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**

Why it matters

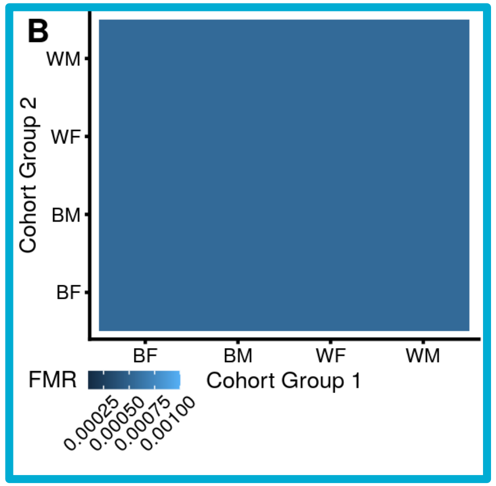
- Why should a male have a higher false positive identification rate when searched against a gallery of all males?

- This doesn't happen with other biometrics, but we've accepted it with face recognition

- But through some fairly simple matrix multiplications, we can make face behave more like iris and fingerprint. This would be a good thing, not just for fairness (human adjudication, automation bias, etc.)



False match cohort matrix
for **face**



False match cohort matrix
for **finger, iris, etc.**

Future Work

- What is the best metric for results? Need something beyond false match rate.
- What is the best means to identify and remove “clustering” in feature vector space?
- 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”

In Summary

- Testing face recognition algorithms for demographic effects is important
- The way we understand and measure these effects continues to evolve (because we are testing)
- “Bias” is multifaceted – comes from data, algorithmic decisions, interactions of humans with technical systems
- Better understanding will lead to better technical solutions

Questions & Thank you

- Thank you

- Contact information
  - jhoward@idslabs.org

- We are hiring! ^^

- Visit our websites for additional information
  - To see additional work DHS S&T supports, visit www.dhs.gov/science-and-technology
  - All papers, lots of slides, video, etc. https://mdtf.org

- Questions?