U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Science and Technology

Yevgeniy Sirotin
Principal Investigator
The Maryland Test Facility

Arun Vemury
Director
Biometric and Identity Technology Center

November 16, 2021

Disclaimer

- This research was funded by the U.S. Department of Homeland Security (DHS), Science and Technology Directorate (S&T) 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

DHS S&T scenario testing of face recognition technology

- The DHS Biometric Technology Rally is a yearly biometric system evaluation focused on DHS technology use-cases.

- Since 2018, we have tested more than 200 combinations of commercial face acquisition systems and matching algorithms in a high-throughput unattended use case.

- The Rallies provide comprehensive metrics about the tested technologies:
  - Efficiency – transaction times
  - Effectiveness – image acquisition and matching success
  - Satisfaction – user feedback
  - Equitability – technology works well for different groups
  - https://mdtf.org

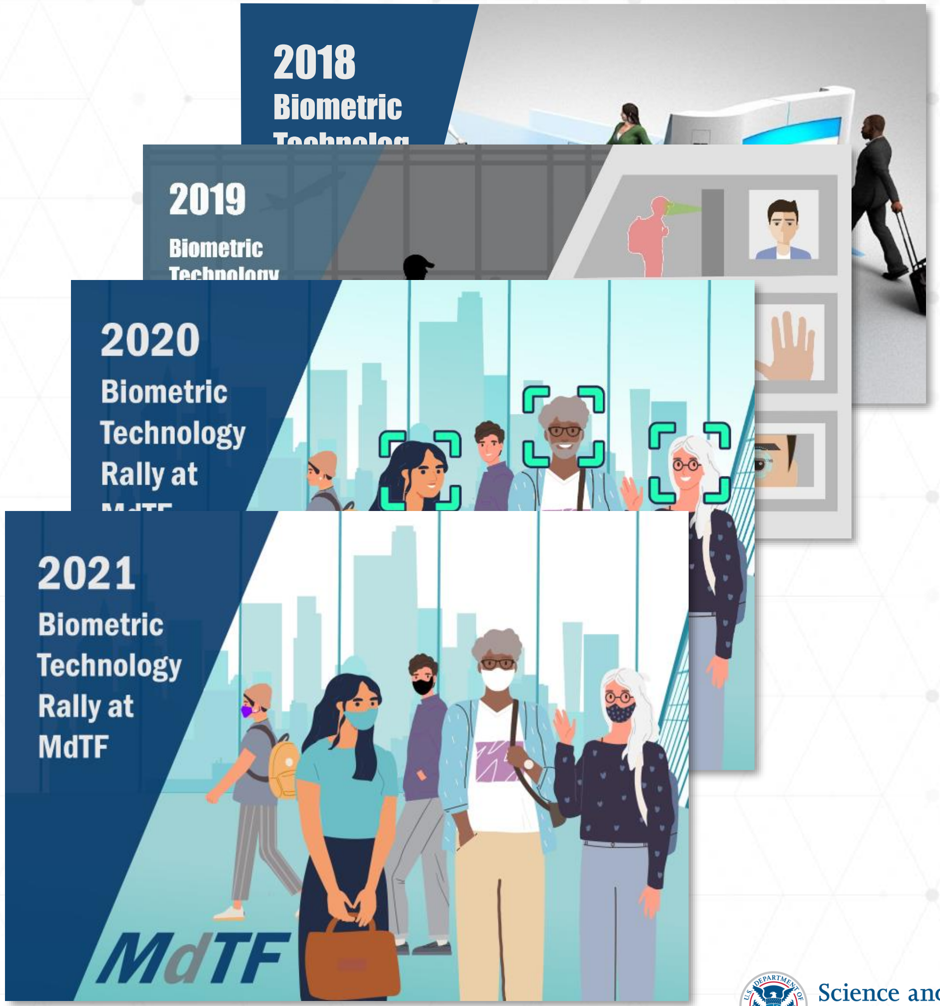
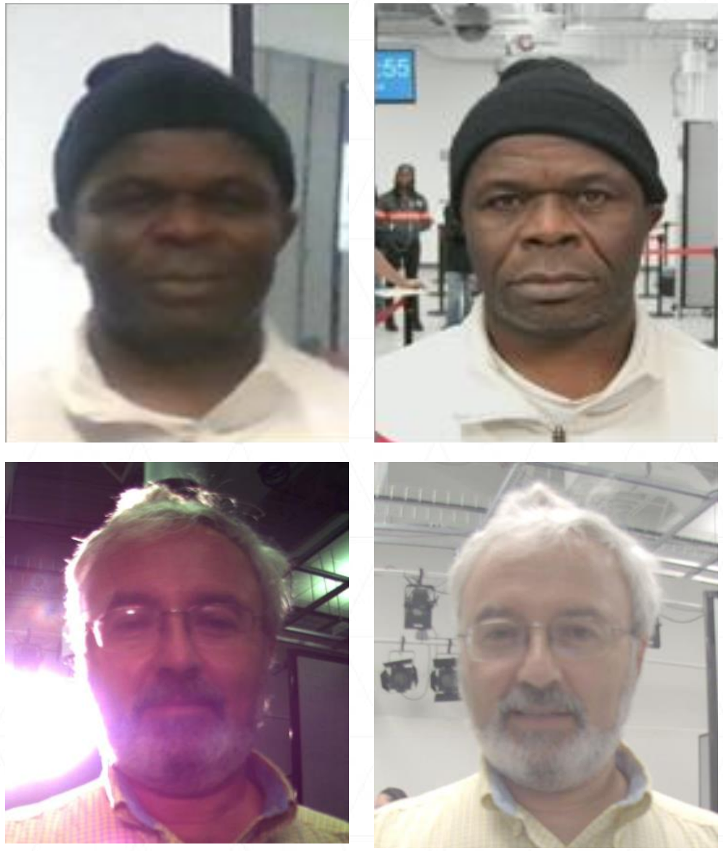


Image quality and face recognition

- Quality **is a property of a face image**
- Quality is **not a property of the person** in the image
- Quality should be predictive of biometric performance:
  - Algorithms: Lower quality → lower mated scores
  - Humans: Lower quality → poorer decisions

different devices
same conditions

same devices
different conditions

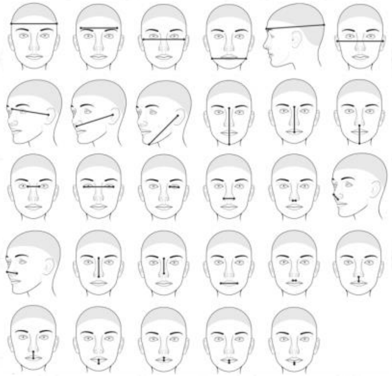


Demographics and face recognition

- Demographics **are properties of a person** that are related to their face
  - Race, gender, age
  - Skin tone, face structure
  - Self-styling behaviors
  - Apparel (e.g. hats, glasses)

- Demographics may influence face image capture through an interaction between face properties and the biometric sensor
  - Differential performance
    - (i.e. differences in biometric match outcomes)
  - Latent differentials
    - (e.g. differences in match scores)

- Broad public deployment of face recognition has raised concerns about differential performance for protected demographic groups

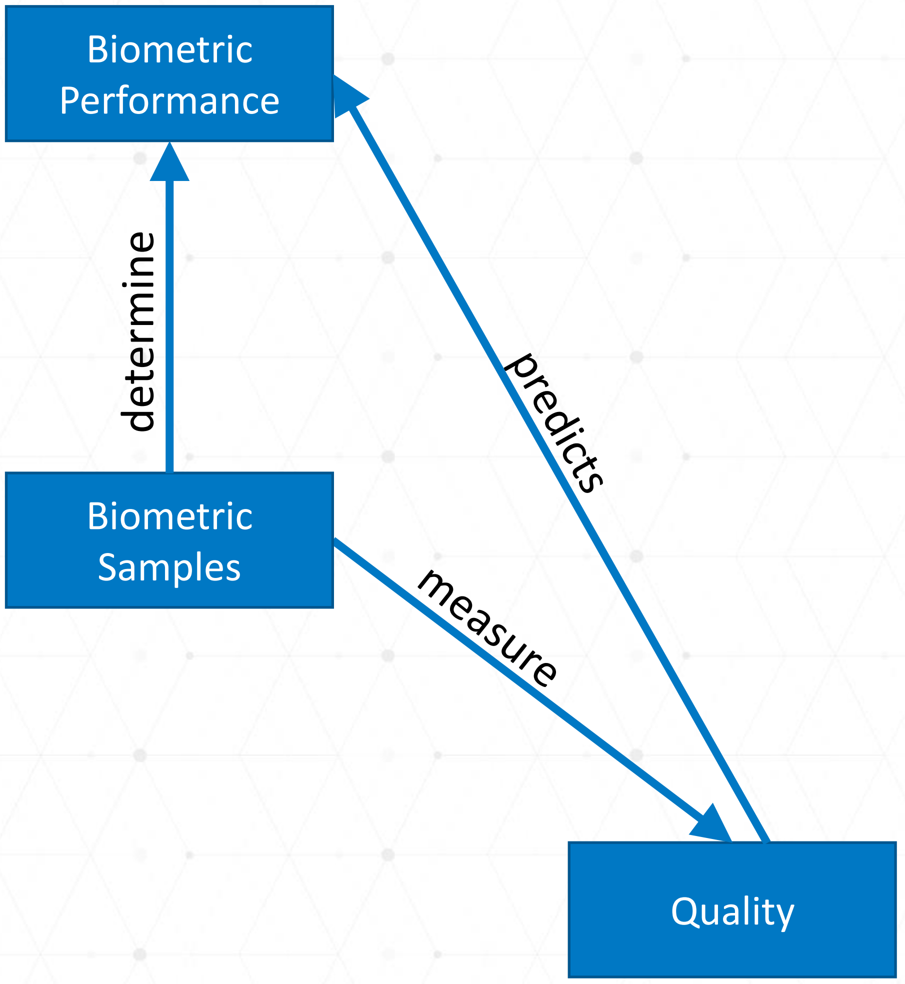


Kesterke et al. Biology of Sex Differences (2016) 7:23

Dantcheva, Antitza, C. Chen, and A. Ross. "Makeup challenges automated face recognition systems." *SPIE Newsroom* (2013): 1-4.

Science and Technology

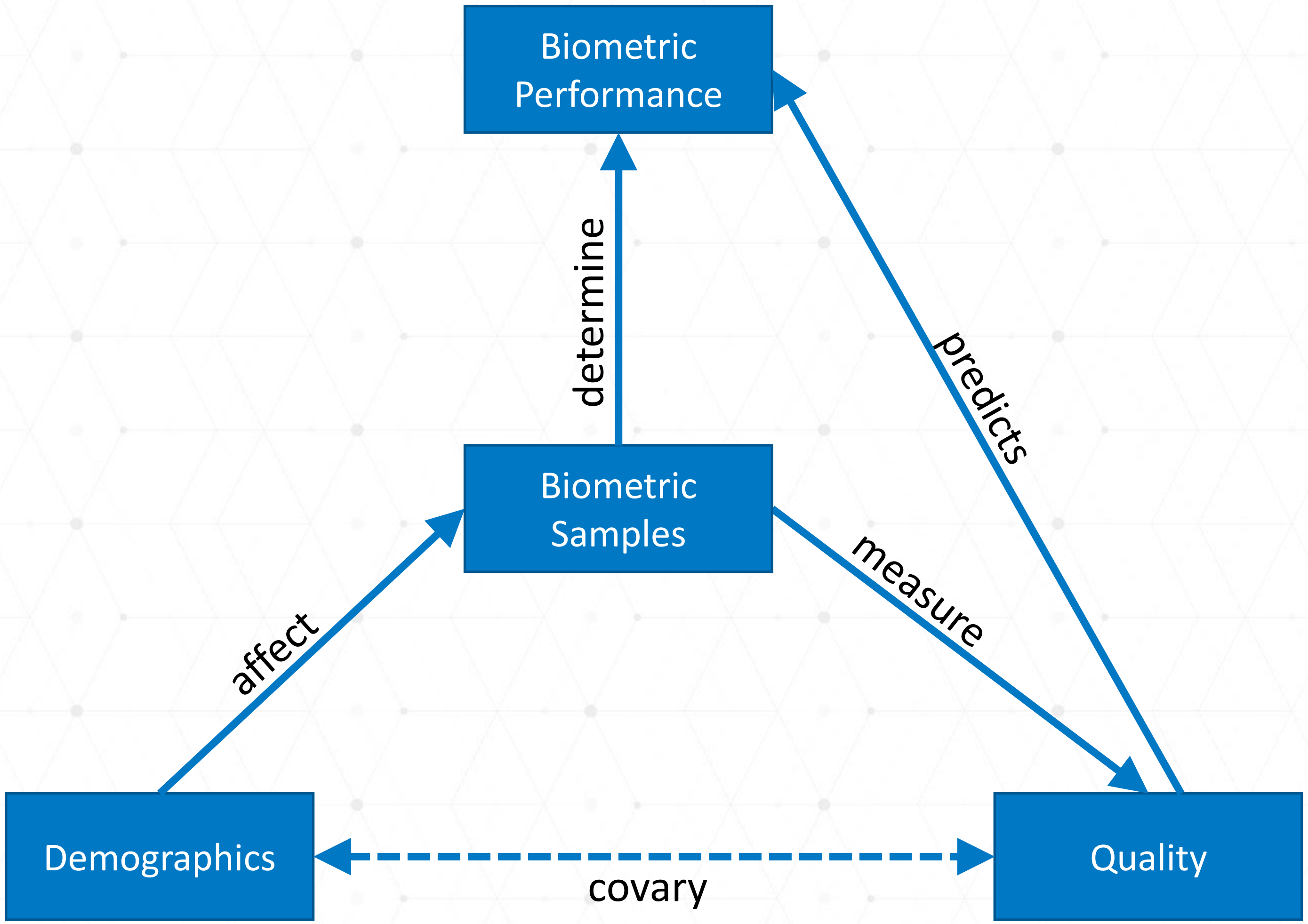
Quality predicts performance



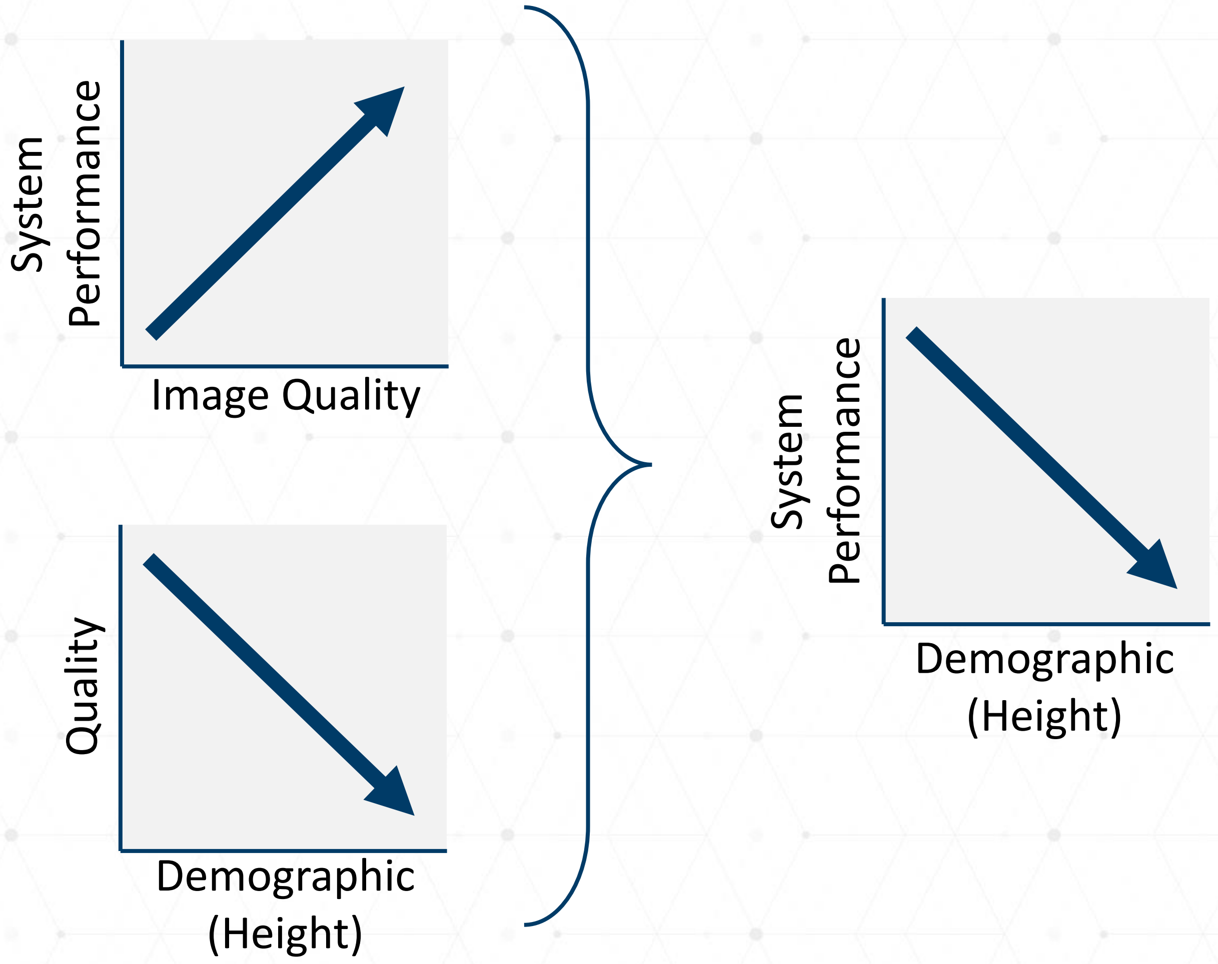
Quality may covary with demographics

Which demographic factors generally covary with face recognition system performance?

Which of these demographic factors are likely to affect images such that they impact quality measures?



Covariation of quality and demographics is a problem



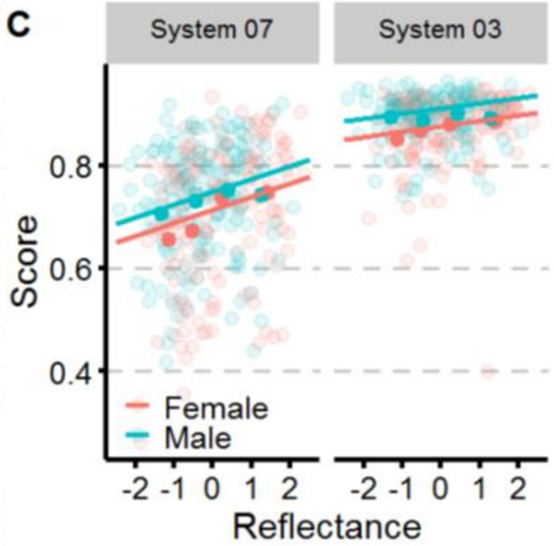
Algorithm bias or image quality difference?

Prior work from our group

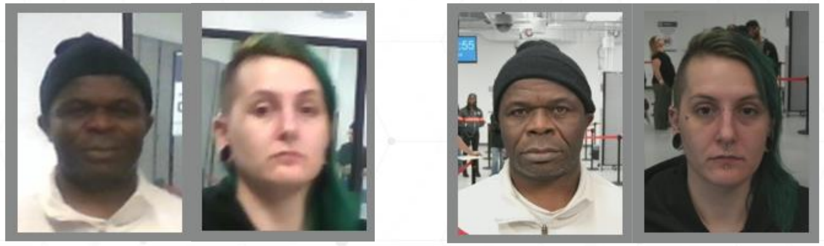
- Tested **one** matching system with 11 acquisition systems
- Used linear modeling to identify demographic factors influencing scores

- Mated scores increased with face area lightness (FAL) of the subject
- Influence of FAL depended on the acquisition system

- FAL was a better predictor of mated score than Race
- Mated scores were higher for men relative to women when matching different-day, but not same day images



System 07 System 03

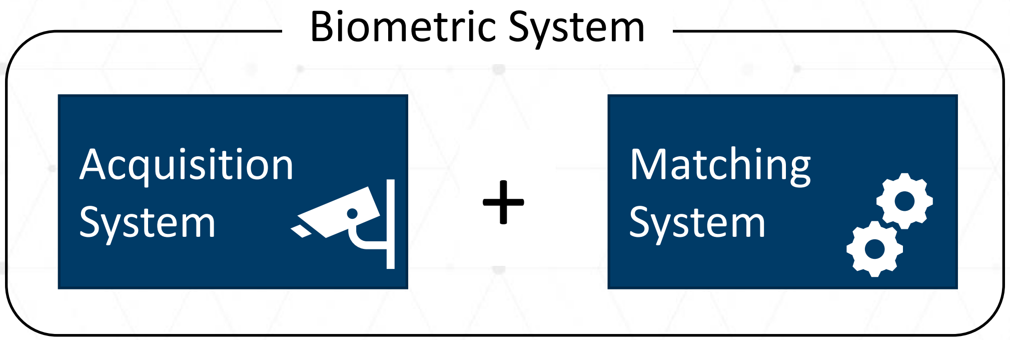


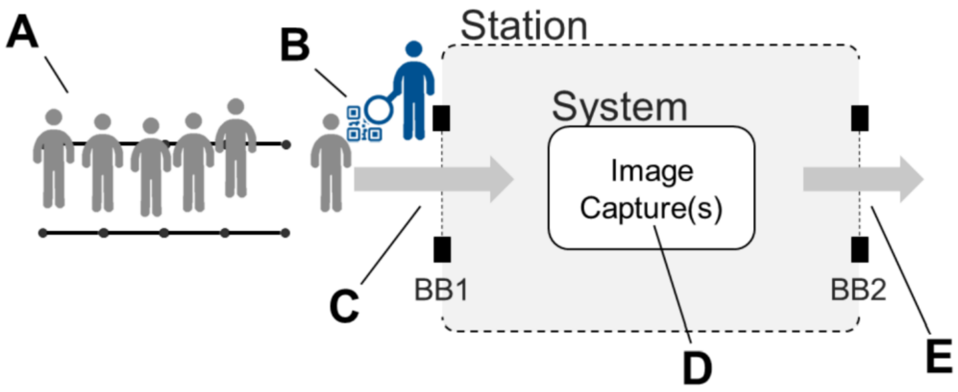
0.658 0.802 0.930 0.882

Cook, Howard, Sirotin, Tipton, and Vemury. Fixed and Varying Effects of Demographic Factors on the Performance of Eleven Commercial Facial Recognition Systems. T-BIOM.

Biometric systems and data

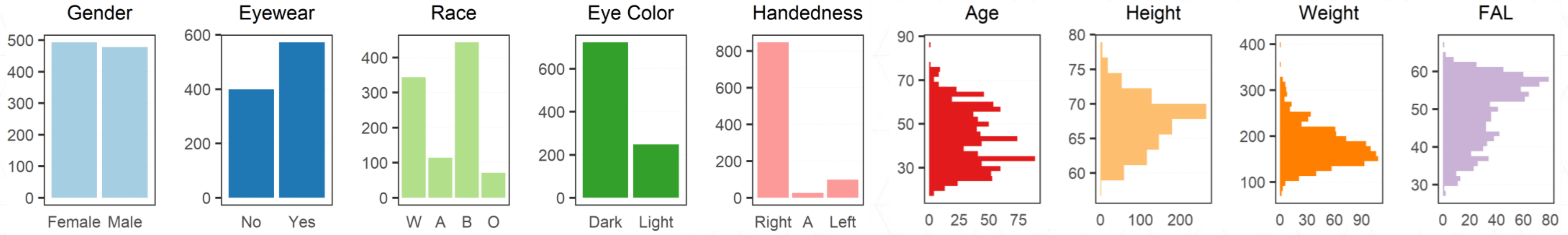
- 2019 and 2020 Biometric Technology Rallies
- **148** algorithm-camera combinations
  - Treated as different biometric systems
- Fit models to explain rank-1 mated score variation across sample of diverse participants:
  - 422 from 2019 Rally
  - 560 from 2020 Rally





Demographic factors and full mated score model

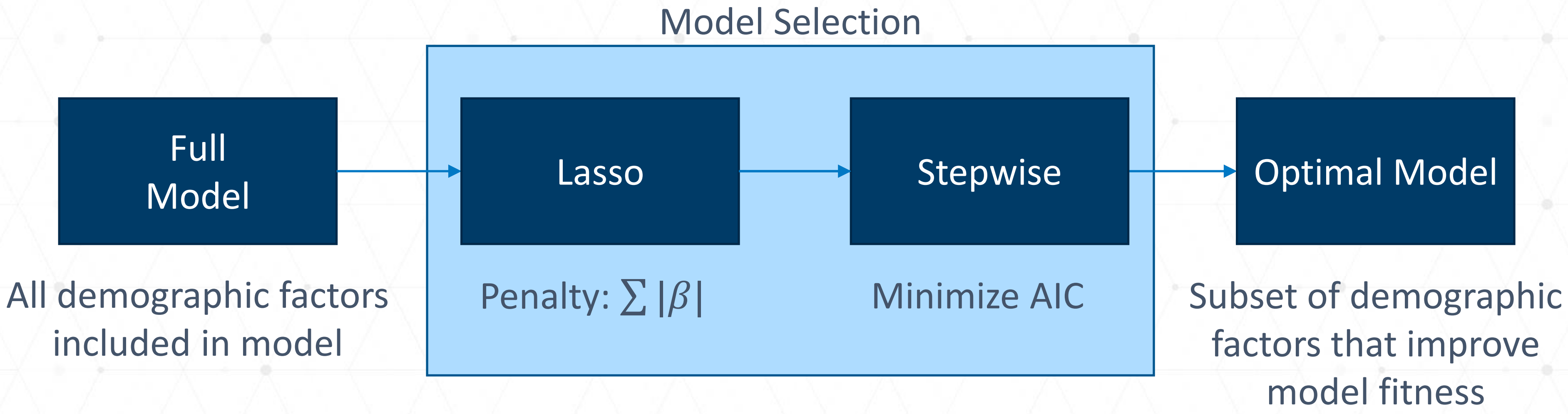
Categorical | **Continuous**



self-reported | measured

$$Score \sim \beta_0 + \beta_1 Gender + \beta_2 Eyewear + \beta_3 Race + \beta_4 EyeColor + \beta_5 Handedness + \beta_6 Age + \beta_7 Height + \beta_8 FAL$$

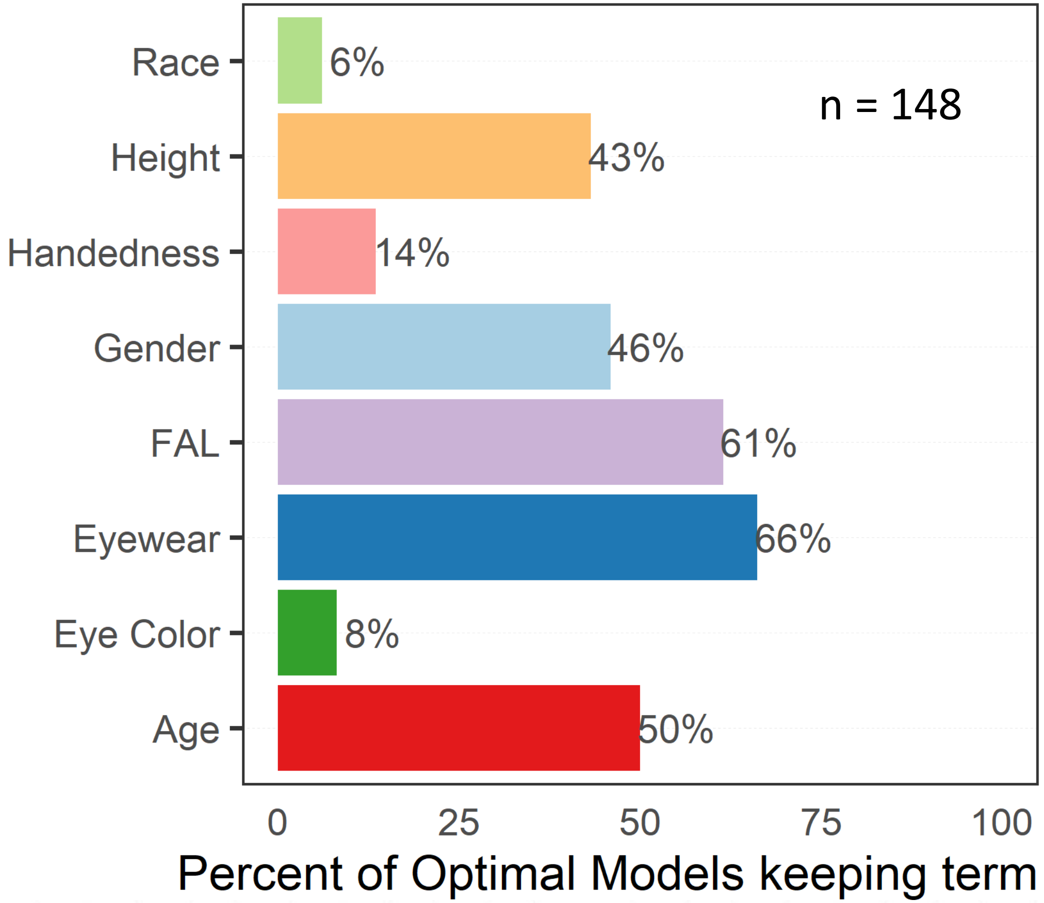
Optimal model selection



One optimal model per biometric system.

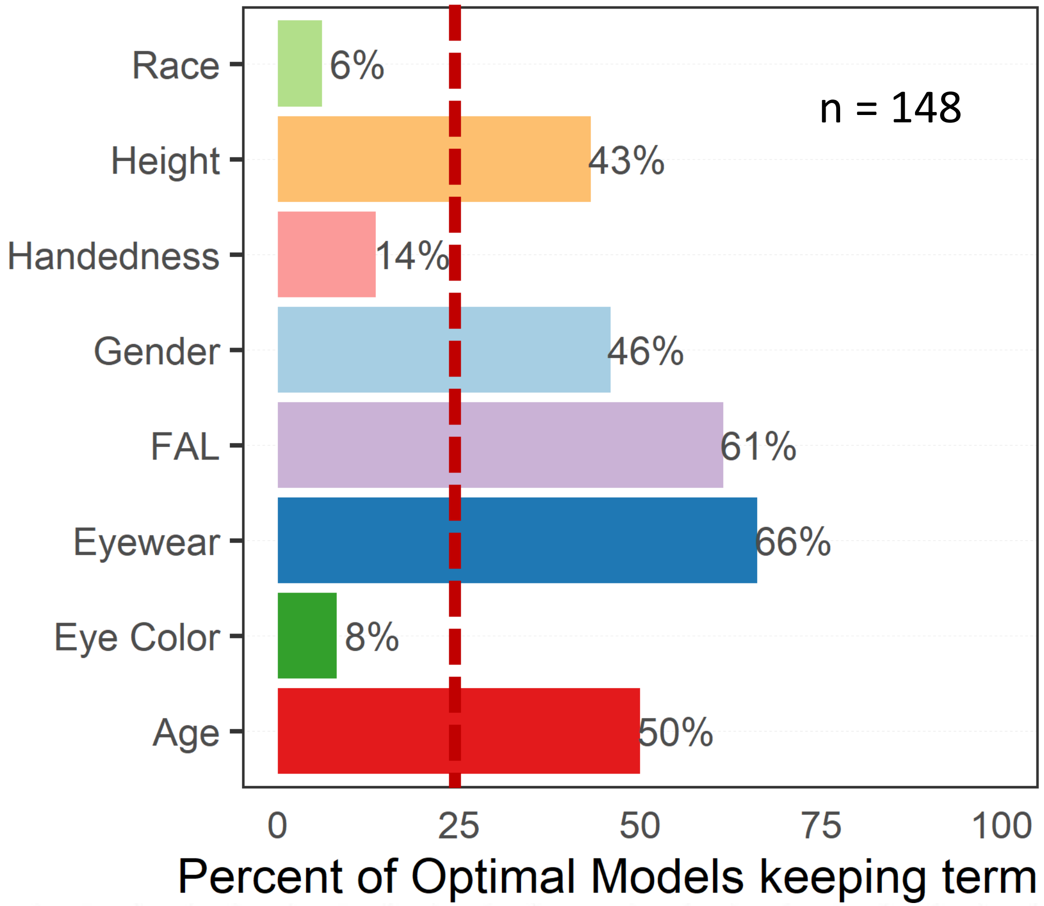
Optimal models

$$Score \sim \beta_0 + \beta_1 Gender + \beta_2 Eyewear + \beta_3 Race + \beta_4 EyeColor + \beta_5 Handedness + \beta_6 Age + \beta_7 Height + \beta_8 FAL$$



Optimal models

$$Score \sim \beta_0 + \beta_1 Gender + \beta_2 Eyewear + \cancel{\beta_3 Race} + \cancel{\beta_4 EyeColor} + \cancel{\beta_5 Handedness} + \beta_6 Age + \beta_7 Height + \beta_8 FAL$$

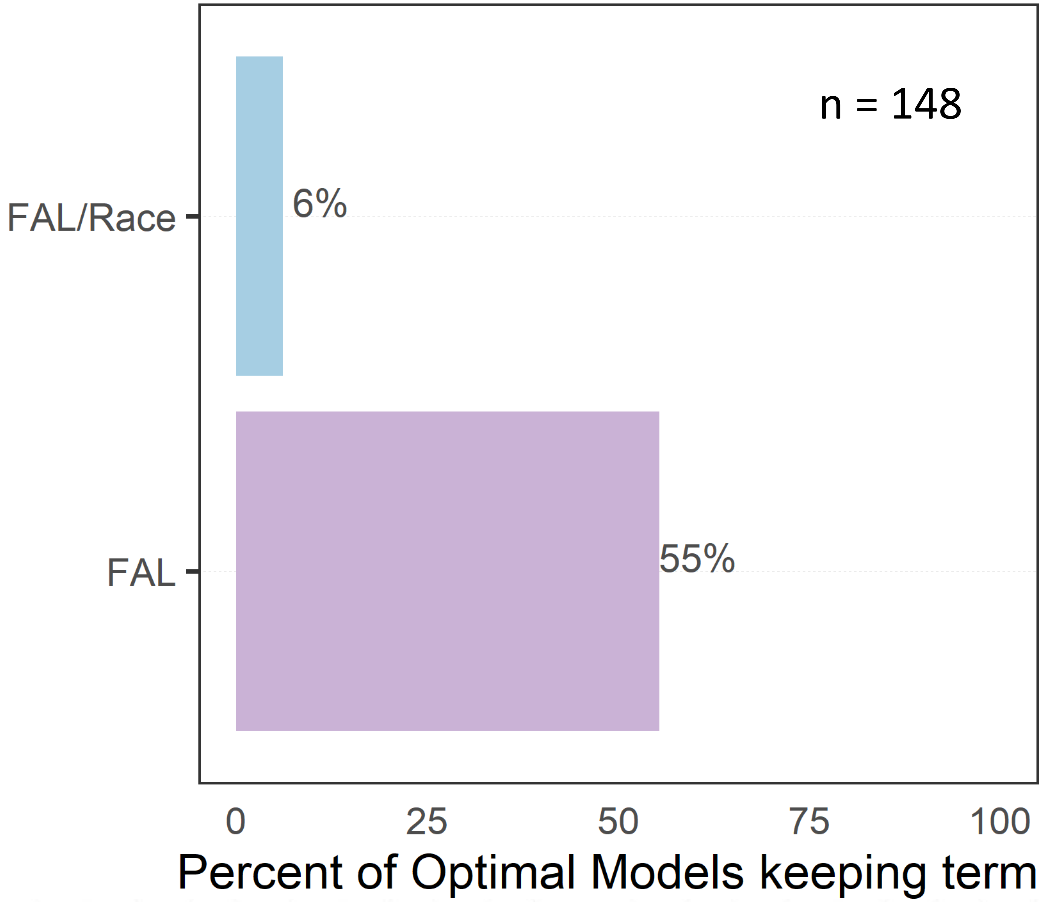


Race is a poorer predictor of mated scores than face area lightness

- Only 6% of optimal models included race as a factor as compared with 61% of optimal models that included FAL

- Each model that included race also included FAL

- FAL appears to be a better predictor of mated scores than race across our sample of face recognition systems

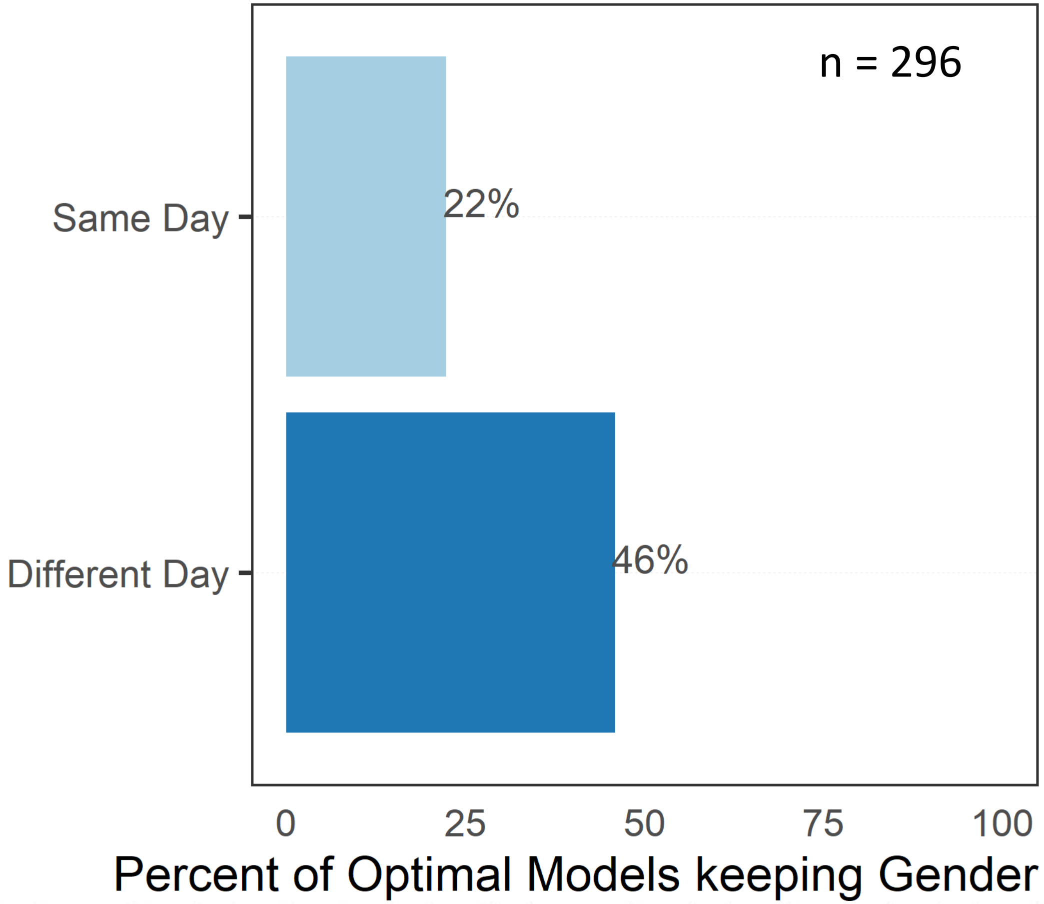


Gender effects reduced for same-day reference images

- Gender had a consistent influence on mated scores
  - Scores for men were higher than scores for women
  - Gender effects present in 46% of models
  - Scores computed using gallery images collected on prior days

- Fitting models to scores obtained when using high quality gallery images collected on the same day
  - Gender effect prevalence in models more than halved

- Gender effects appear to be related to differences in facial appearance over time



Direction of demographic effects

Lower scores for people of different height than average

Lower scores for taller people

Higher scores for men

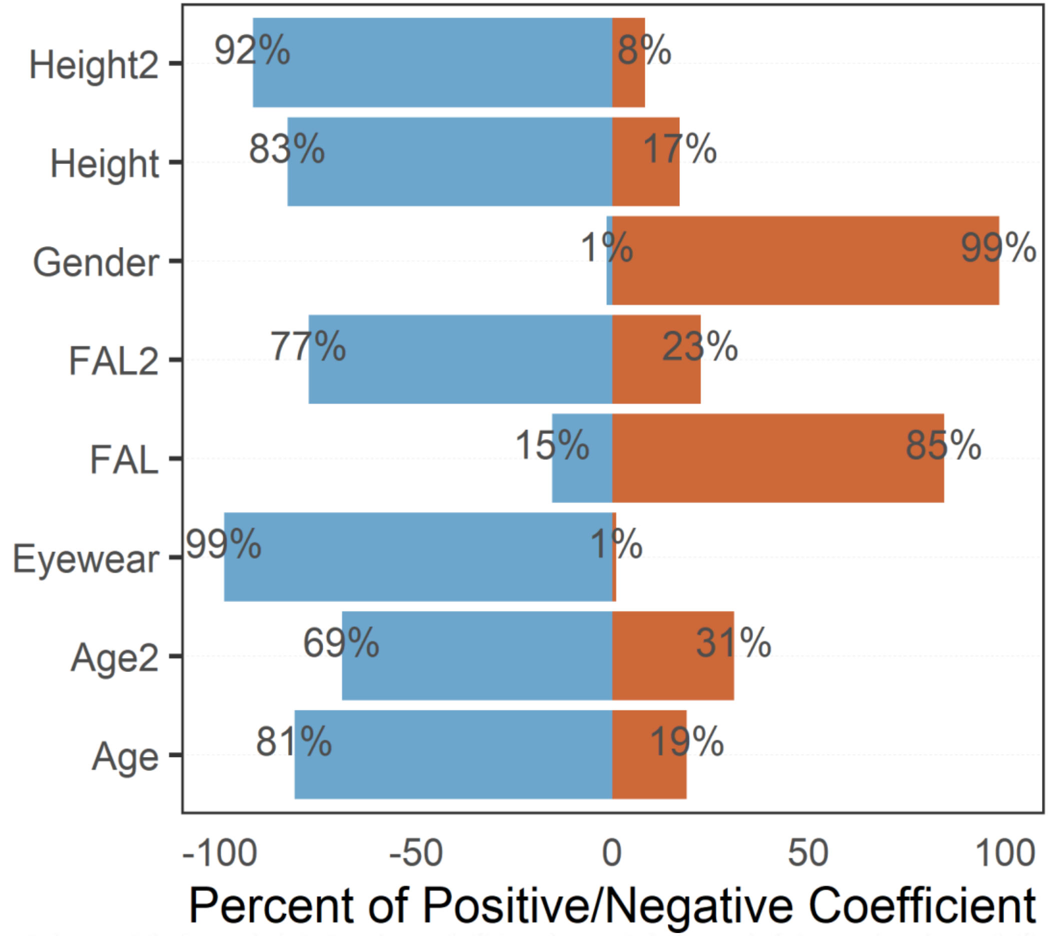
Lower scores for people of different FAL than average

Higher scores for greater face area lightness

Lower scores for people with eyewear

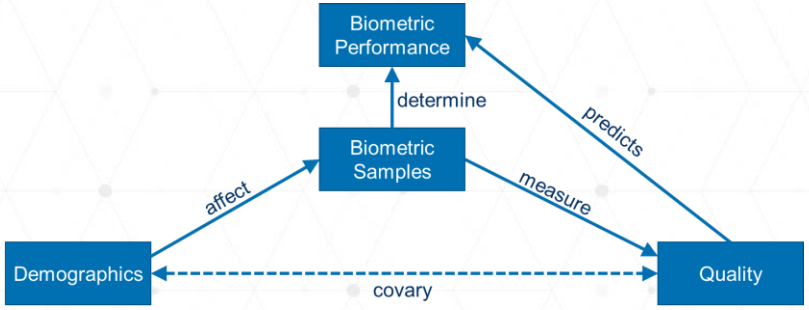
Lower scores for people of different age than average

Lower scores for older people



Summary

- Demographic correlates of mated scores:
    - Present in at least 43% of the 148 tested commercial face recognition systems
    - **Face area lightness** – interacts with camera sensor
    - **Height** – likely interacts with camera height
    - **Eyewear** – occludes part of the face
    - **Age** – algorithm training set?
    - **Gender** – difference appearance over time?
- Face quality may mitigate capture-related differentials:
    - Adjust lighting environment
    - Adjust camera position
    - Ask to remove glasses/apparel
- Open questions:
    - How will proposed face quality measures covary with demographics in practice?
    - How will quality affect non-mated scores?
    - How will this affect datasets used for biometric evaluations?



Demographic Factor	Relevant Quality Measure
Face area lightness	Camera dynamic-range Color balance Illumination
Height	Camera-subject distance Camera lensing Pose Face location
Eyewear	Eyes visible
Gender	--
Age	--

Questions?

yevgeniy@mdtf.org

arun.vemury@hq.dhs.gov

cynthia@mdtf.org

john@mdtf.org

This work was performed by a dedicated team of researchers at The Maryland Test Facility.

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

Science and Technology