Demographic issues related to face image quality

Identity and Data Sciences Laboratories

Yevgeniy Sirotin
Technical Director
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.

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:
  - Document validation, presentation attack detection
  - Face, fingerprint, iris
  - Comprehensive holdings of responsibly acquired images suitable for evaluating biometric systems

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

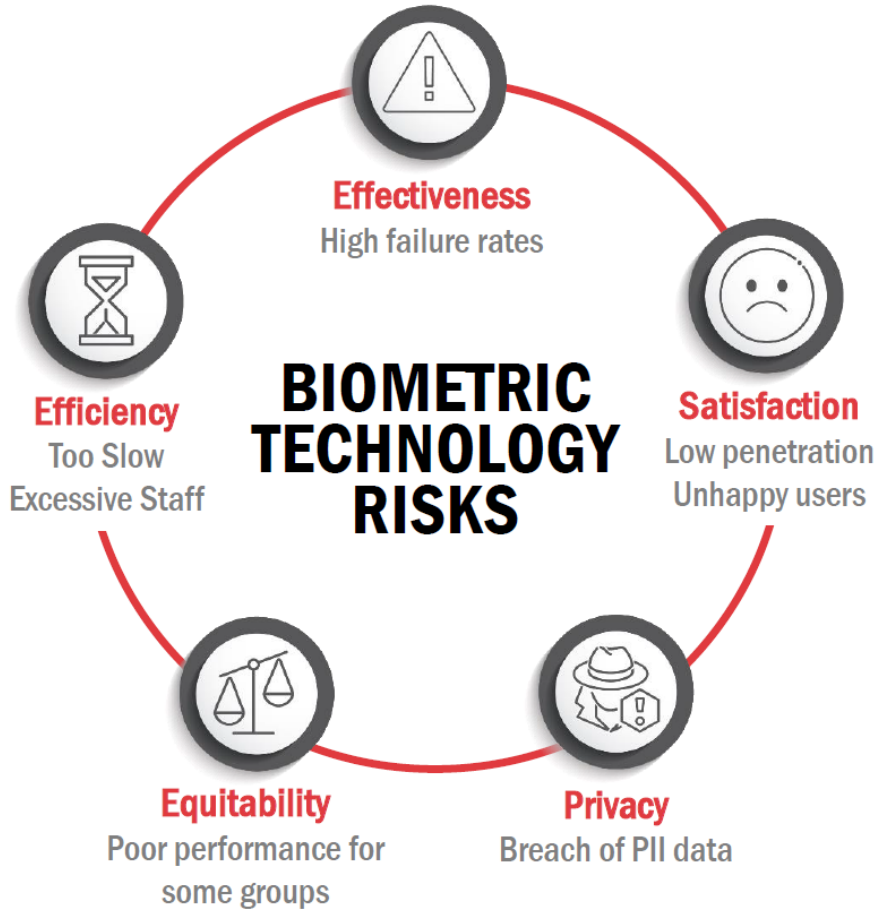
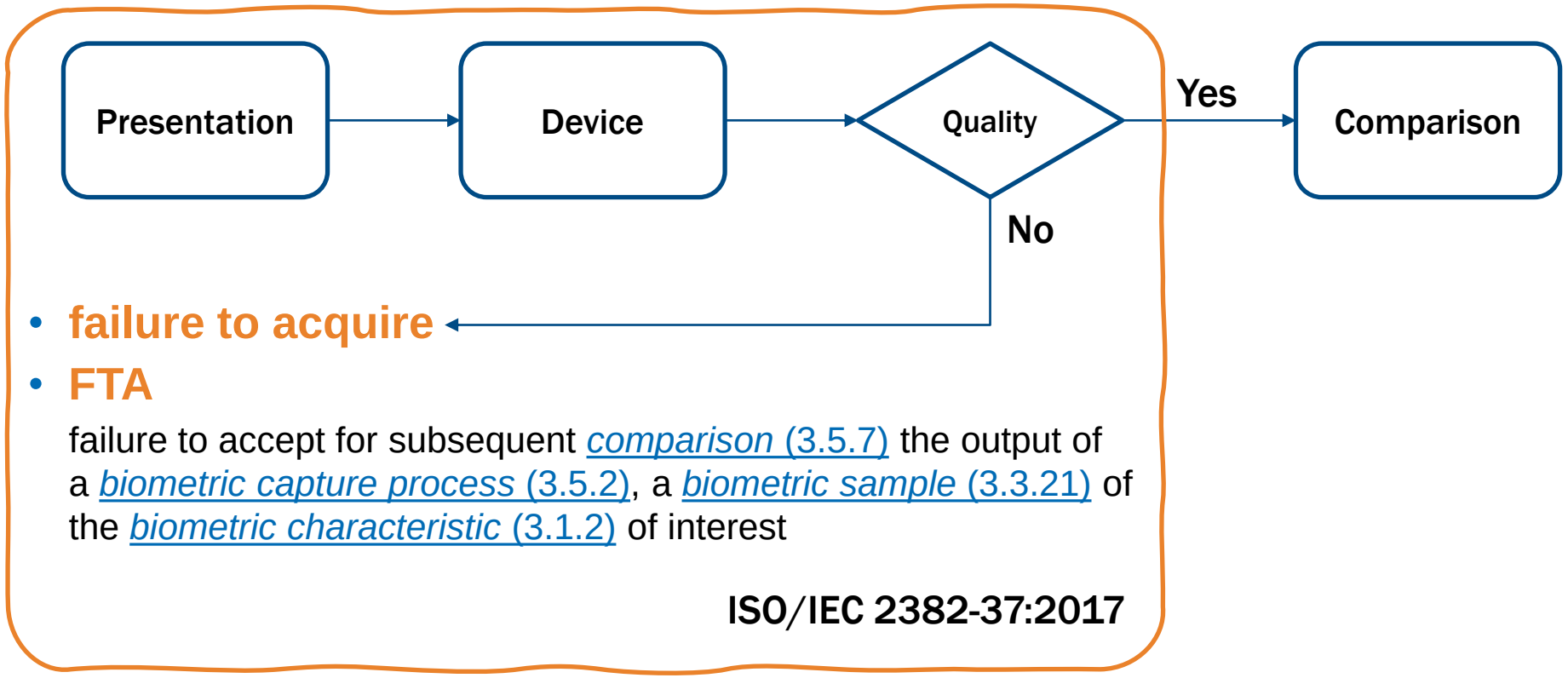


Image quality dictates what is compared



- **failure to acquire**
- **FTA**

  failure to accept for subsequent *comparison* (3.5.7) the output of a *biometric capture process* (3.5.2), a *biometric sample* (3.3.21) of the *biometric characteristic* (3.1.2) of interest

ISO/IEC 2382-37:2017

Overview of the 2022 Biometric Technology Rally

1 INFORMED CONSENT:
575 diverse people from the local area were briefed about the Rally and consented to participate.

2 GROUND TRUTH:
Volunteers self-report their gender and race. Staff measures their skin tone.

3 FORMING GROUPS:
Volunteers form groups of two and groups of four to use each system.

4 IMAGE ACQUISITION:
Acquisition systems had to select one best photo of each volunteer in the group to submit for matching.

4 Group Acquisition Systems.

5 MATCHING:
Matching systems identify the face in each photo by comparing it to photos of known people.

10 Algorithms.

6 REPORTING:
Performance was measured for each of the 40 possible combinations of acquisition and matching systems.

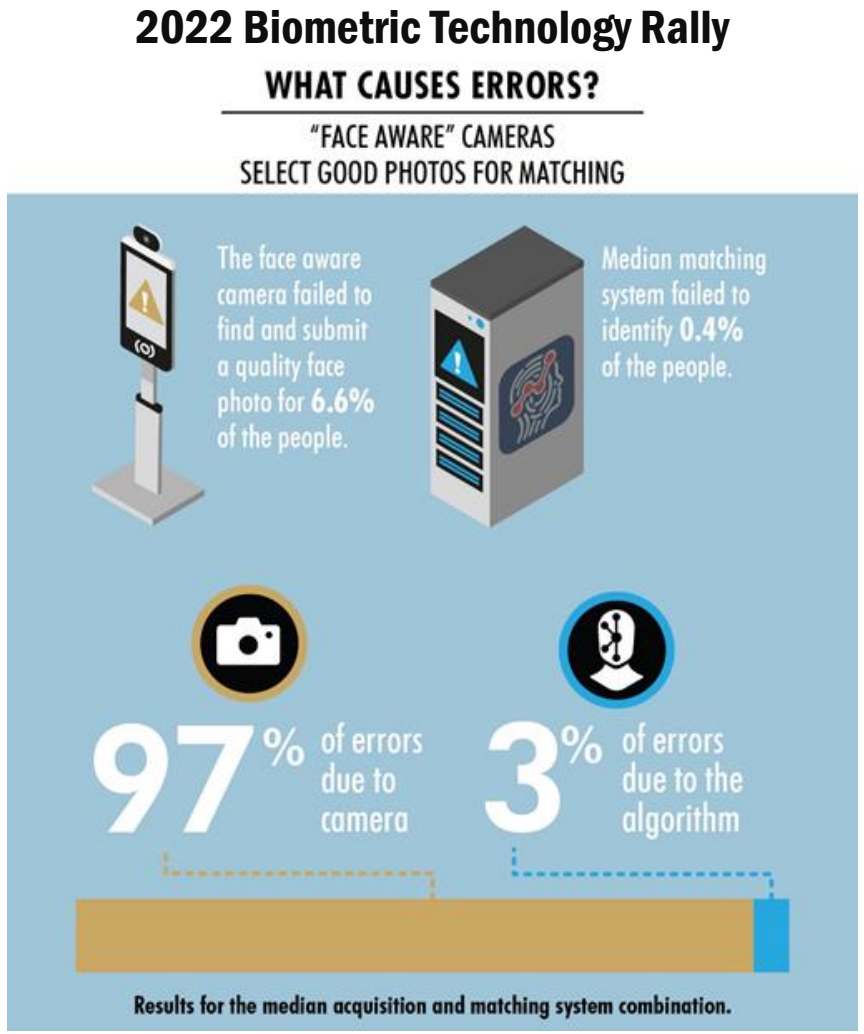
40 System Combinations Tested.

Science and Technology

Prepared by the Identity and Data Sciences Lab

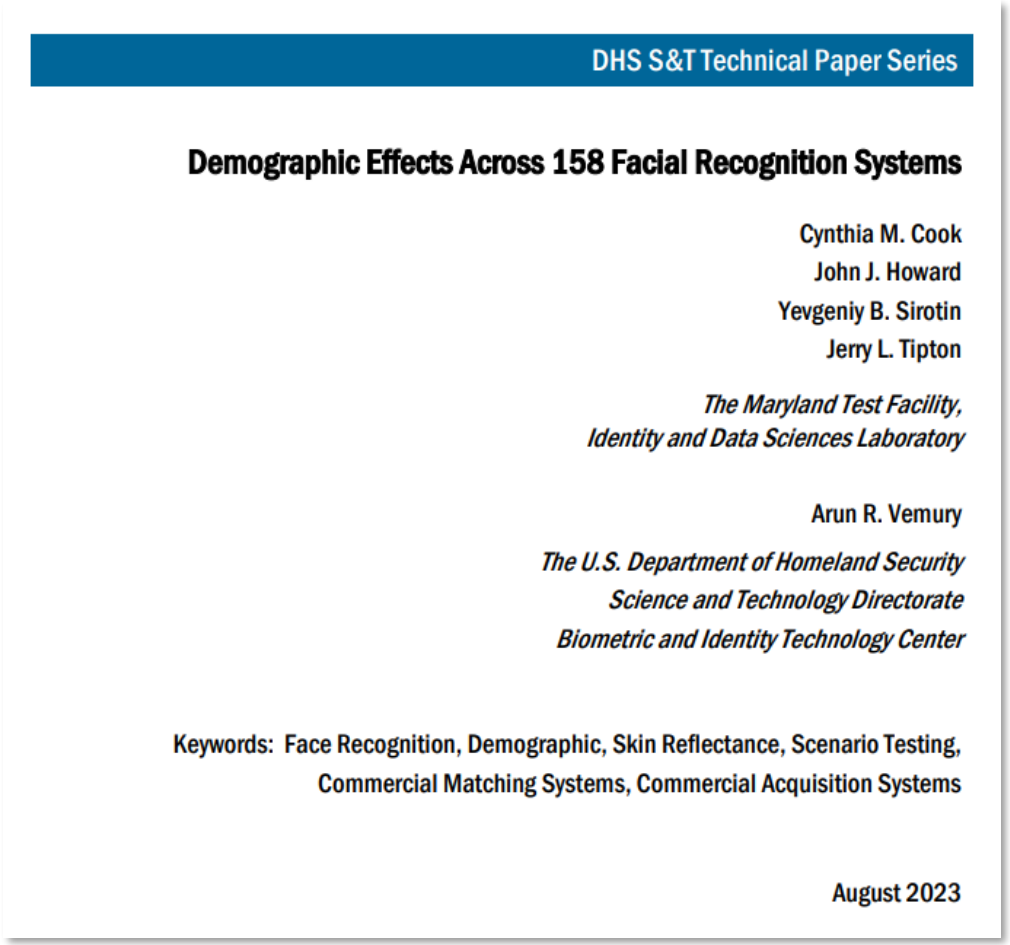
Failure to acquire can be the largest source of error in face recognition

- ISO/IEC 29794-5:2023(CD2)
  - *"This standard is needed because without significant modernization of capture procedures, recognition errors will become more prevalent as volumes increase."*
- Since 2017, we have tested 100+ systems in high throughput unattended applications
- We find, consistently, that failure to acquire (FTA) is the single largest source of error in such systems.
- Quality filters reduce matching error but increases FTA errors



Face recognition is sensitive to demographics

- 158 face recognition systems
  - 2019 to 2021
  - Combinations of acquisition and matching systems
  - Examined rank one mated similarity scores using linear modeling
- Mated similarity scores:
  - Lower for people wearing eyewear (96% of models)
  - Lower for women than men (74% of models)
    - No gender effect when matching same day face images
  - Lower for people with darker skin tone (57% of models)
- Skin lightness is a better predictor of average mated similarity scores than self-reported race



DHS S&T Technical Paper Series

Demographic Effects Across 158 Facial Recognition Systems

Cynthia M. Cook
John J. Howard
Yevgeniy B. Sirotin
Jerry L. Tipton

The Maryland Test Facility,
Identity and Data Sciences Laboratory

Arun R. Vemury

The U.S. Department of Homeland Security
Science and Technology Directorate
Biometric and Identity Technology Center

Keywords: Face Recognition, Demographic, Skin Reflectance, Scenario Testing, Commercial Matching Systems, Commercial Acquisition Systems

August 2023

Demographic differentials in commercial systems...

True Identification Rate Darker Skin

Matching System \ Acquisition System	Bison	Longs	Wilson	Borah
Kenai	96.8	92.6	88.8	69.8
Miami	96.8	93.1	88.8	69.8
Tioga	96.8	93.1	88.8	69.8
Mill	96.8	93.1	88.8	69.8
Bronx	96.2	92.1	88.8	68.8
Grant	96.8	91.0	88.8	69.8
Hop	95.7	92.1	88.8	69.3
Entiat	96.2	92.1	87.8	69.8
Flag	96.8	90.5	88.3	68.3
Row	79.6	81.0	74.5	57.7

Table legend

Max	>99%	>95%	>90%	≤90%

9 of 40 met the TIR threshold

True Identification Rate Lighter Skin

Matching System \ Acquisition System	Bison	Longs	Wilson	Borah
Kenai	96.3	96.3	94.1	72.5
Miami	96.3	96.3	94.1	72.5
Tioga	96.3	96.3	94.1	72.5
Mill	96.3	96.3	94.1	72.0
Bronx	96.3	96.3	94.1	72.5
Grant	95.7	96.3	94.1	71.4
Hop	96.3	96.3	94.1	72.5
Entiat	95.7	95.8	93.6	72.0
Flag	96.3	94.7	93.6	70.9
Row	80.7	85.2	81.9	59.3

17 of 40 met the TIR threshold

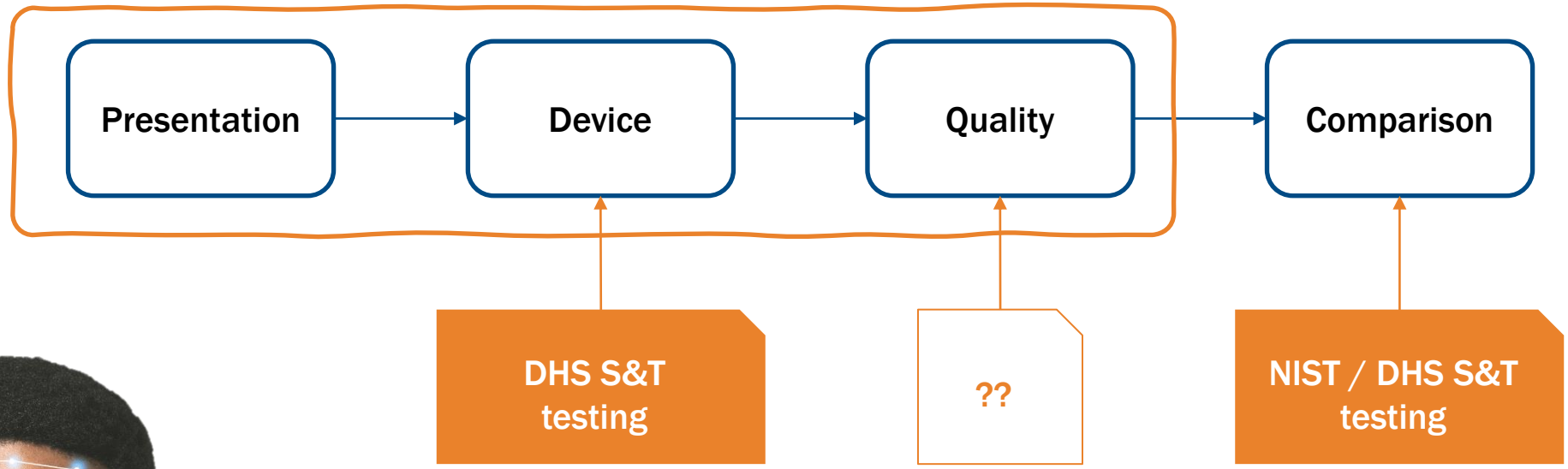
36 of 40 met 95% TIR requirements when discounting FTA regardless of skin tone.

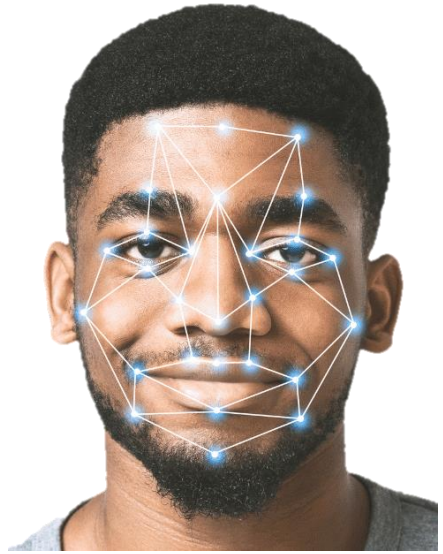
Face image quality measures

- Background uniformity
- Illumination uniformity
- Luminance mean
- Luminance variance
- Skewed
- Abnormal kurtosis illumination prevention
- Underexposure prevention
- Overexposure prevention
- Dynamic range
- Sharpness
- Motion blur prevention
- Compression ratio
- Natural colour
- Single face present
- Eyes visible
- Eyes open
- Mouth occlusion prevention
- Mouth closed
- Face occlusion prevention
- Inter-eye distance
- Head size
- Leftward crop of face in image
- Rightward crop of face in image
- Downward crop of face in image
- Upward crop of face in image
- Pose angle yaw frontal alignment
- Pose angle pitch frontal alignment
- Pose angle roll frontal alignment
- Shoulder presentation
- Expression neutrality
- No head covering
- Radial distortion
- Pixel aspect ratio
- Camera subject distance

ISO/IEC 29794-5:2023(CD2)

Face image quality dictates what is compared





https://eab.org/images/banners/2023-11-07-09_EAB-FIQWS.png

Are we ensuring that image quality measures perform equitably across demographic groups?

Lack of data is a known problem in equitability assessment

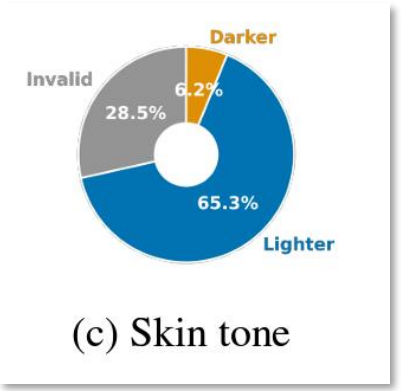
- Sony AI examined 20+ datasets for body pose estimation
  - Only 1 annotated with skin tone and gender (using Amazon Mechanical Turk)
  - Image annotations resulted in unreliable, poorly distributed demographic labels
- New regulations will likely require more testing (e.g., the October 31 Executive Order on AI)
  - We need a better data and methods to test image analyses like quality measures for demographic effects
- What about face pose? – e.g., CMU Multi-PIE:
  - "The subjects were predominantly men (235 or 69.7% vs. 102 or 30.3%). 60% of subjects were European-Americans, 35% Asian, 3% African-American and 2% others. The average age of the subjects was 27.9 years."
  - 3% ~ 10 people

The AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI (R²HCAI)

A Case Study in Fairness Evaluation:
Current Limitations and Challenges for Human Pose Estimation

Julienne LaChance* **William Thong*** **Shruti Nagpal** **Alice Xiang**
Sony AI
julienne.lachance@sony.com, william.thong@sony.com, shruti.nagpal@sony.com, alice.xiang@sony.com

Dataset	Demographic annotations
COCO (2014)	Augmented with skin tone and gender in (2021); *R.P.O.*
MPII Human Pose (2014)	None
Human3.6M (2014)	Gender
Frames Labeled in Cinema Plus (2014)	None
HumanEVA (2010)	None
DensePose (2018)	Inherits gender and skin tone from COCO; *R.P.O.*
Leeds Sports Pose (2010)	None
JHMDB (2013)	None
CMU Panoptic Studio Dataset (2015)	None
Frames Labeled in Cinema (2013)	None
Unite the People (2017)	None
CrowdPose (2018)	None
PoseTrack (2018)	None
UPenn Action (2013)	None
ITOP Dataset (2016)	None
VGG Human Pose Estimation (2016)	None
OCHuman (2019b)	None
FashionPose (2014)	None
Mannequin RGB and IRS in-bed (2017)	None
UAV Human (2021)	None



(c) Skin tone

Sequestered multiangle dataset

- Collected as part of the 2022 Biometric Technology Rally
- Controlled imagery of 613 subjects at 4 yaw angles and 3 pitch angles

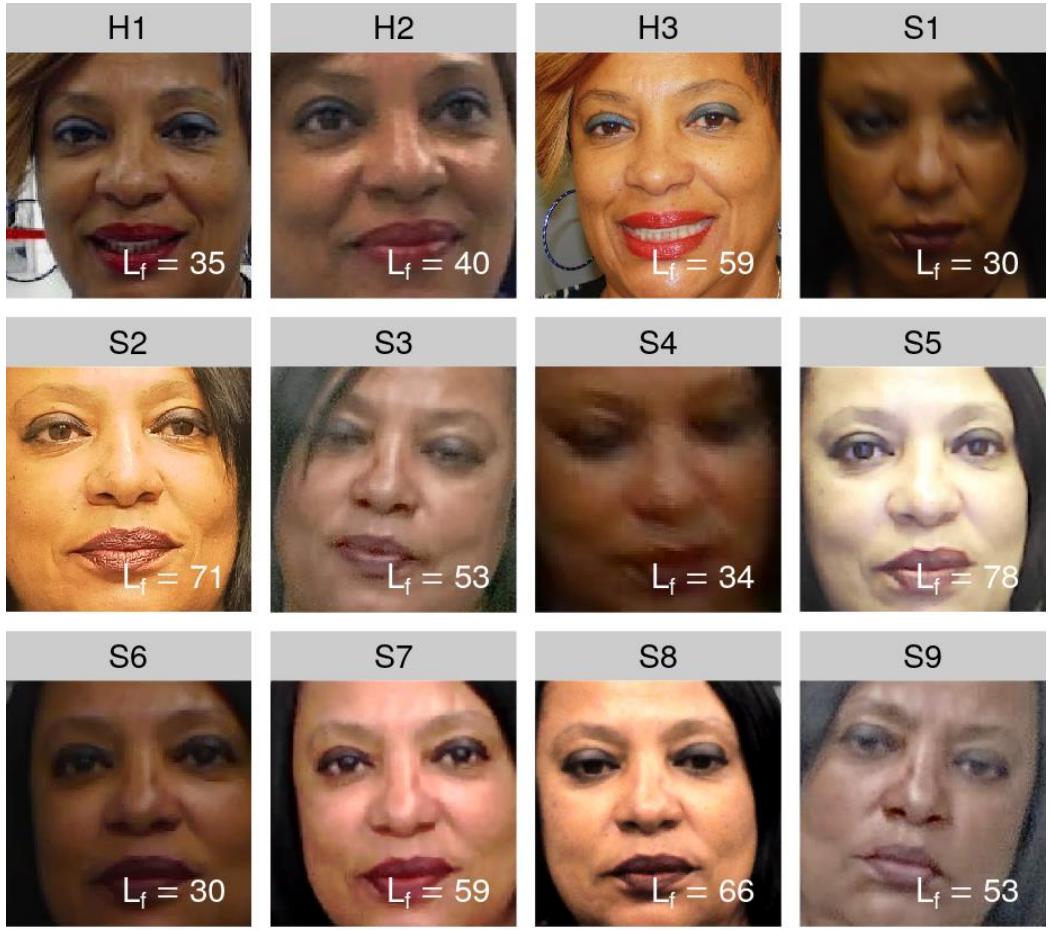
Sequestered multiangle dataset

- Neutral grey background.
- 50% African-American, 35% White, 15% Asian + Other

- 53% Female, 46% Male
- Skin tone measurements using a controlled instrument

Biometric cameras do not reproduce skin tone accurately

- H1-H3: Images taken on different days by different cameras
- S1-S9: Images taken on the same day by different cameras
- All images taken within the same lighting environment
- Which image reflects her “natural color”?



Skin tone and mistaken identity?

- Trevor Noah is a South African comedian of mixed ancestry
- In his book “Born a Crime” he recounts his childhood growing up in South Africa
- He tells a story where his identity is mistaken, to his benefit, based on poor color reproduction when he and a friend are caught shoplifting on camera
- “*The camera chose white*” for his skin tone says Trevor, but black for his friend
- The police never suspected Noah though they showed him the video and asked who the white kid in it was

Trevor Noah, Wikipedia

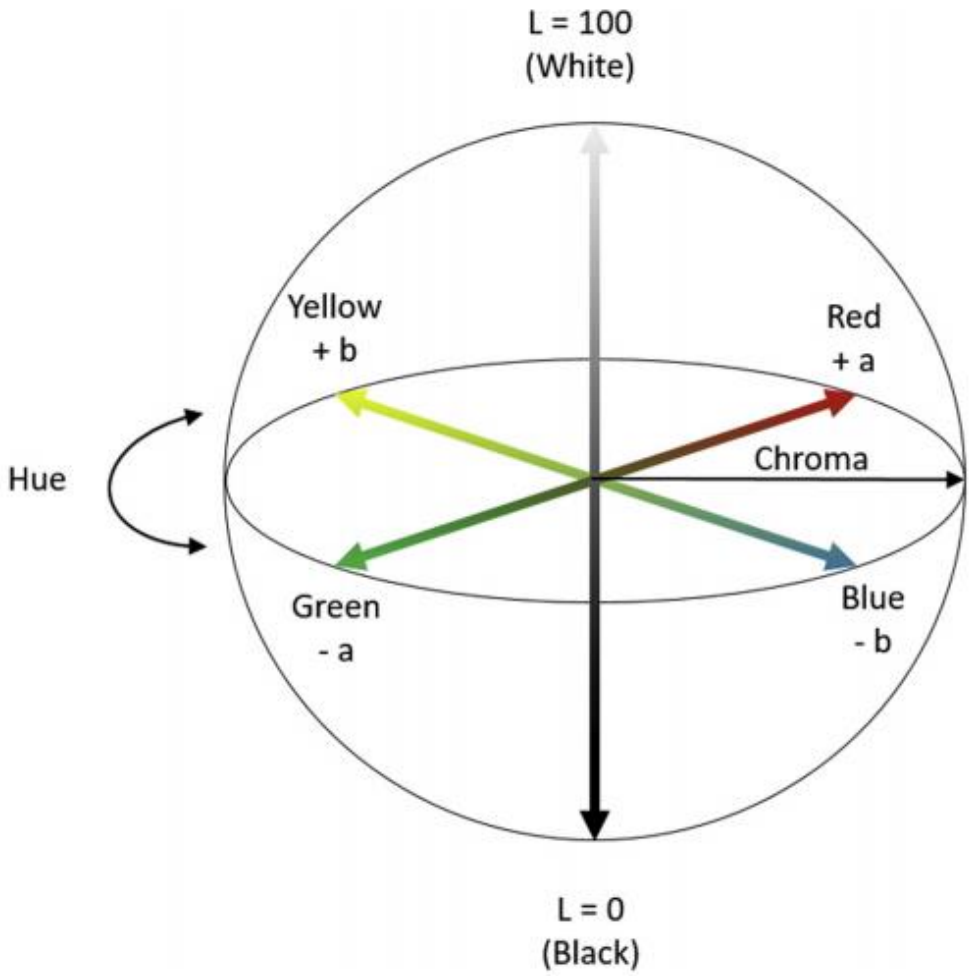
Measuring color

- ISO/IEC 29794-5:2023(CD2)
- CIELAB color space

- Perceptually calibrated

- L* - lightness
- a* - red/green
- b* - yellow/blue

- Hue – $(180/\pi)\tan^{-1}(b^*/a^*)$
- Chromaticity – $\sqrt{a^{*2} + b^{*2}}$



Natural color

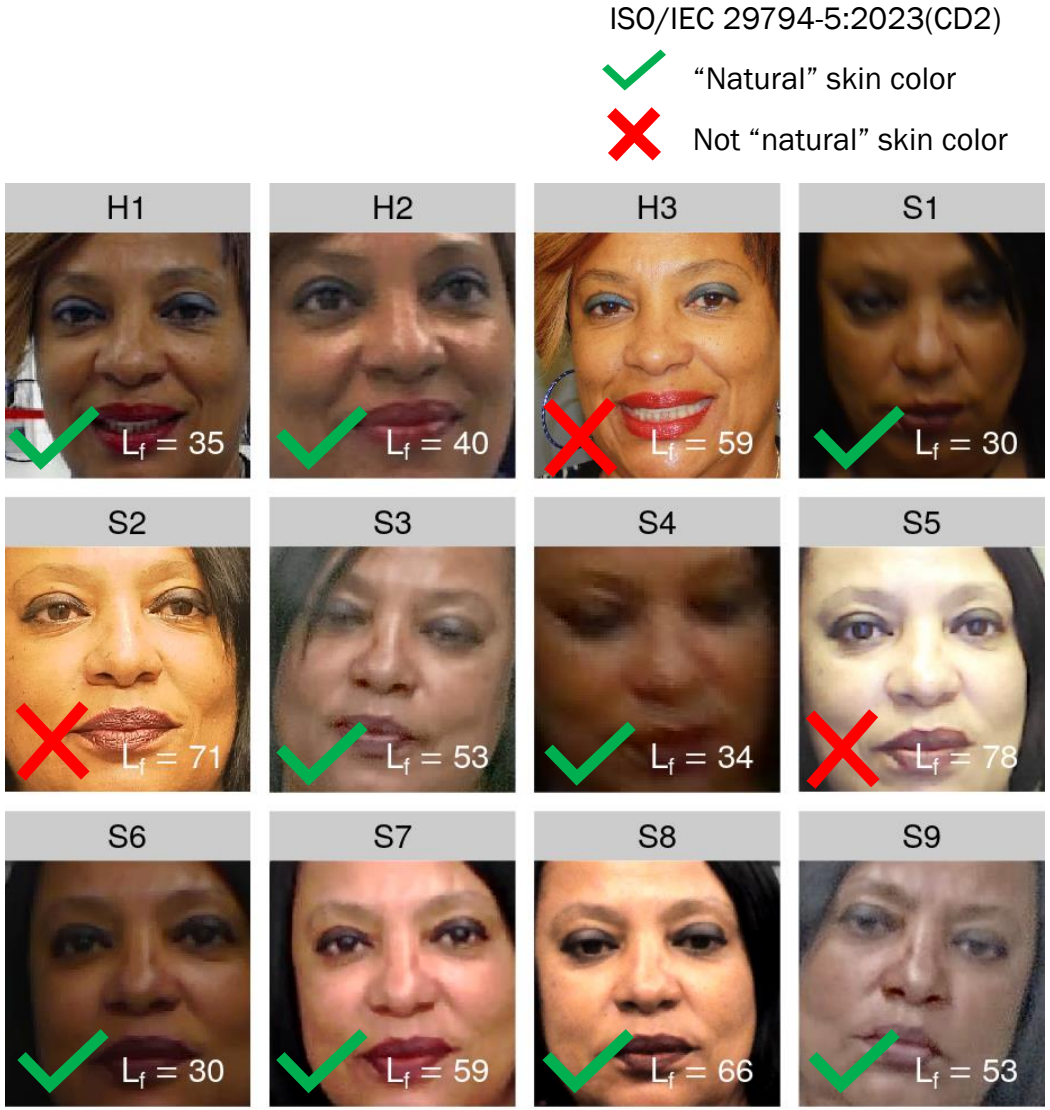
- ISO/IEC 29794-5:2023(CD2)
- CIELAB color space

- a* and b* components are considered and checked against a “natural” range

- L*, or lightness is not included in determining natural color

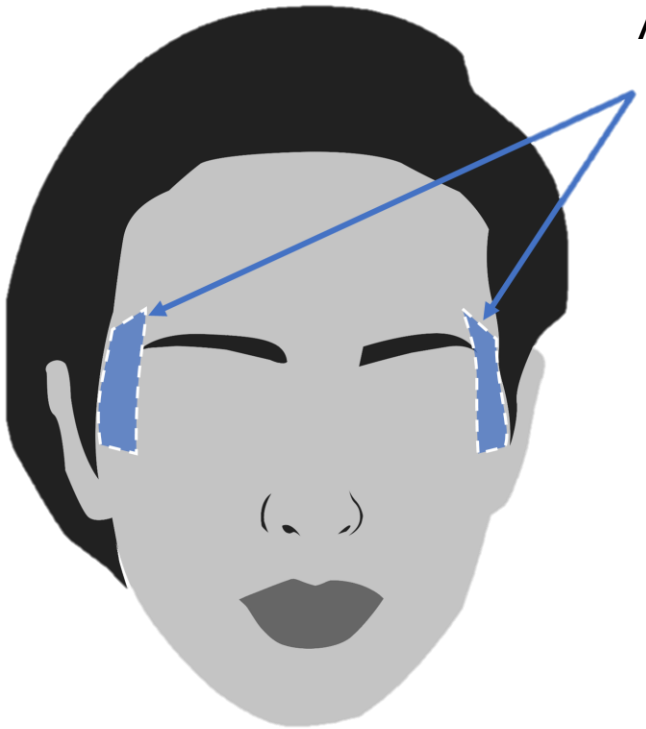
- All but three images of this woman are considered “natural”

- Should we attempt to do better?



Facial skin tone – IDSL Sample

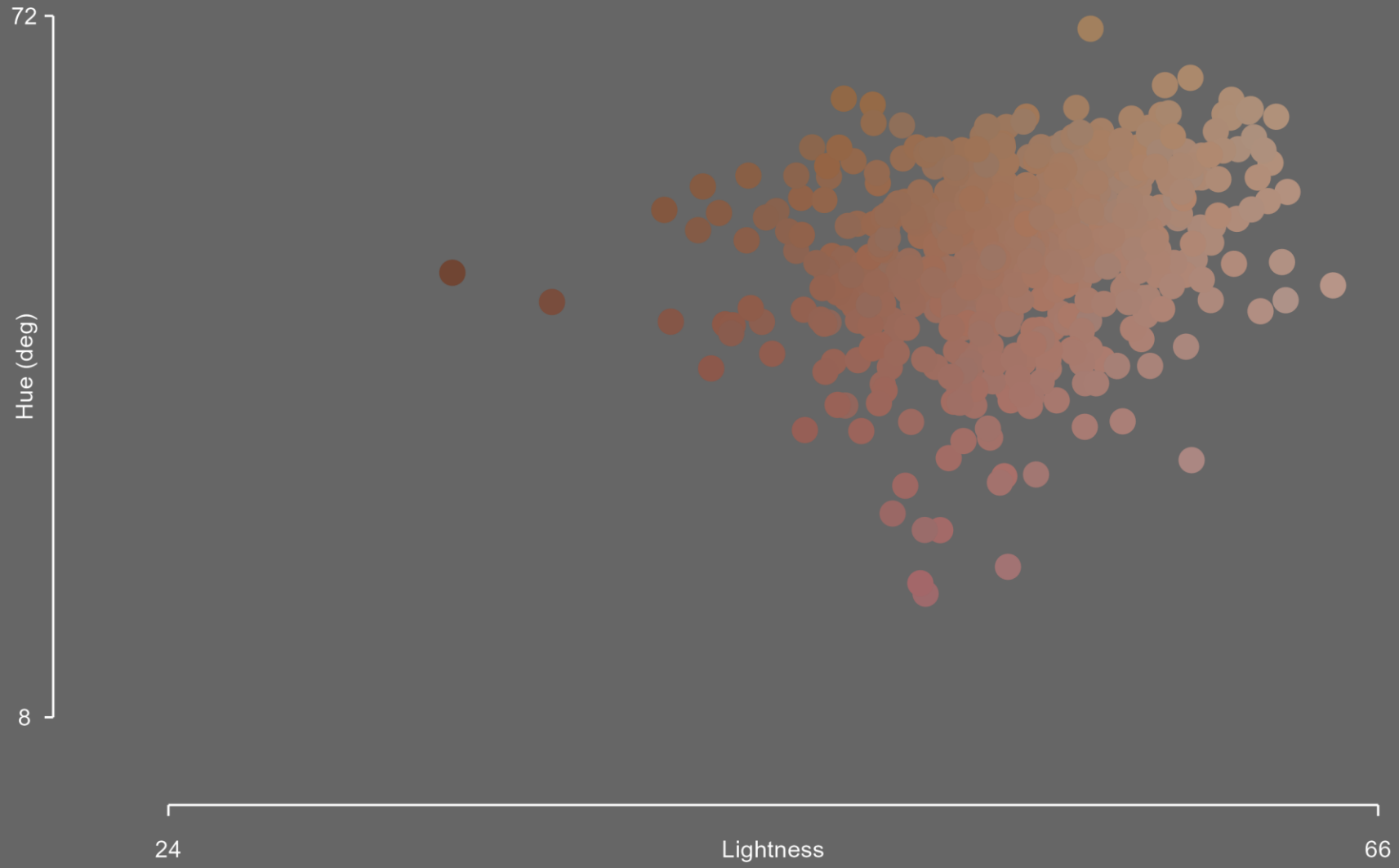
One reading each from the left and the right temple.
Average value computed.



2,500+ unique volunteers.
Diverse race, gender, age.
3,500+ facial color readings.

DSM III Colormeter
Cortex Technology

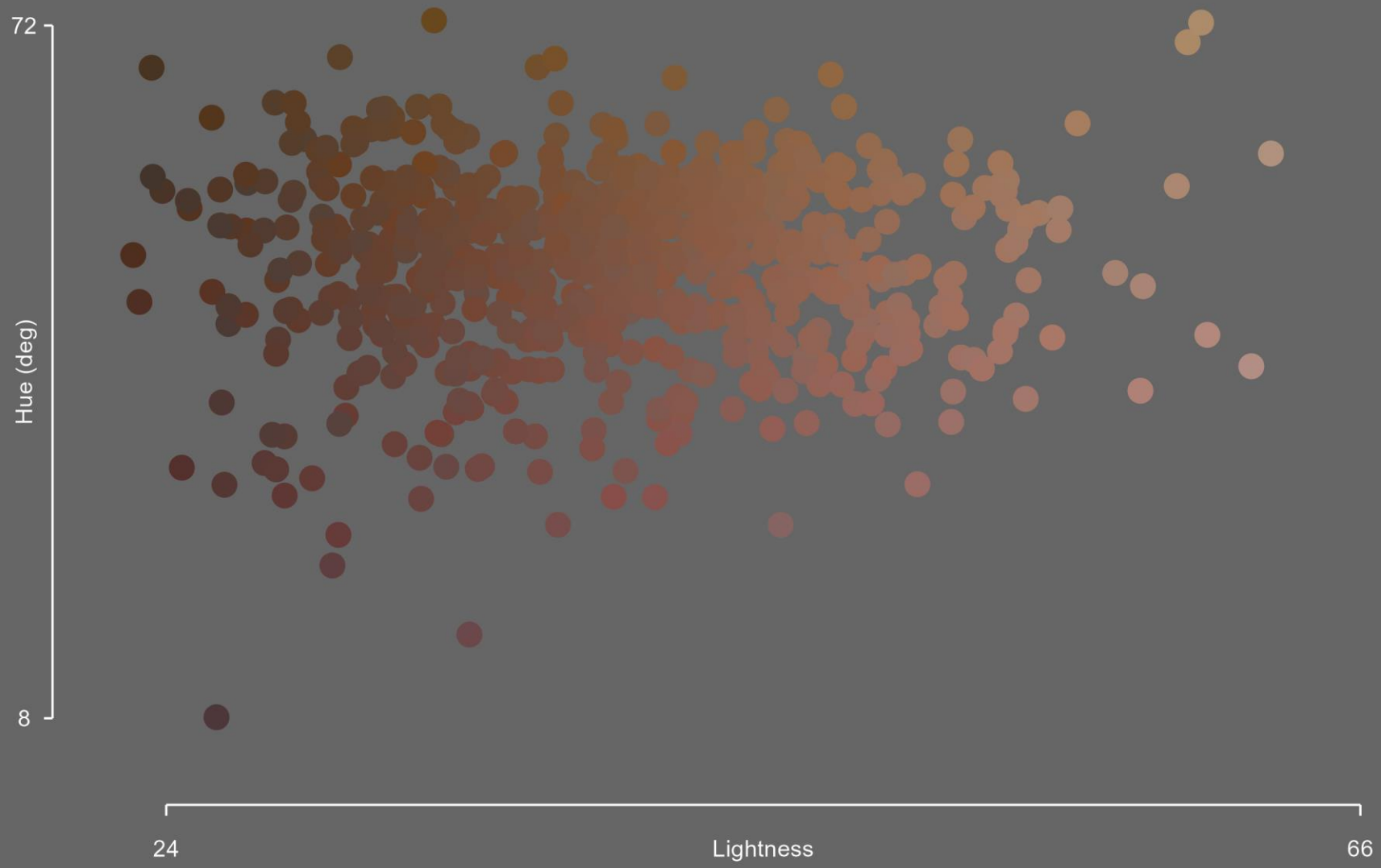
Self report as Asian



Gamut (99%):

L*:	41 – 63
a*:	8 – 21
b*:	8 – 25
hue:	25 – 65
chroma:	16 – 29

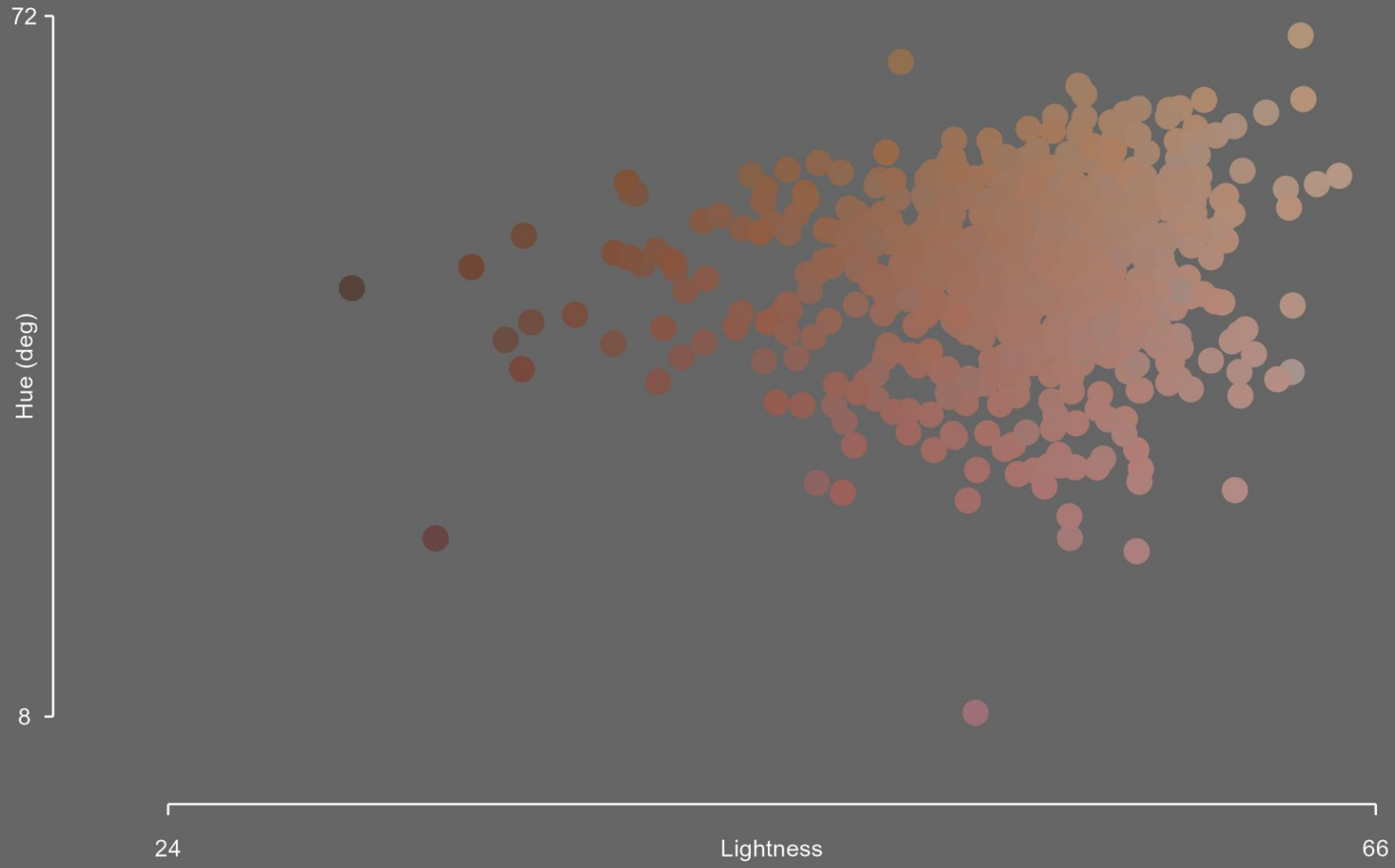
Self report as Black or African American



Gamut (99%):

L*:	24 – 60
a*:	6 – 21
b*:	6 – 27
hue:	24 – 69
chroma:	10 – 31

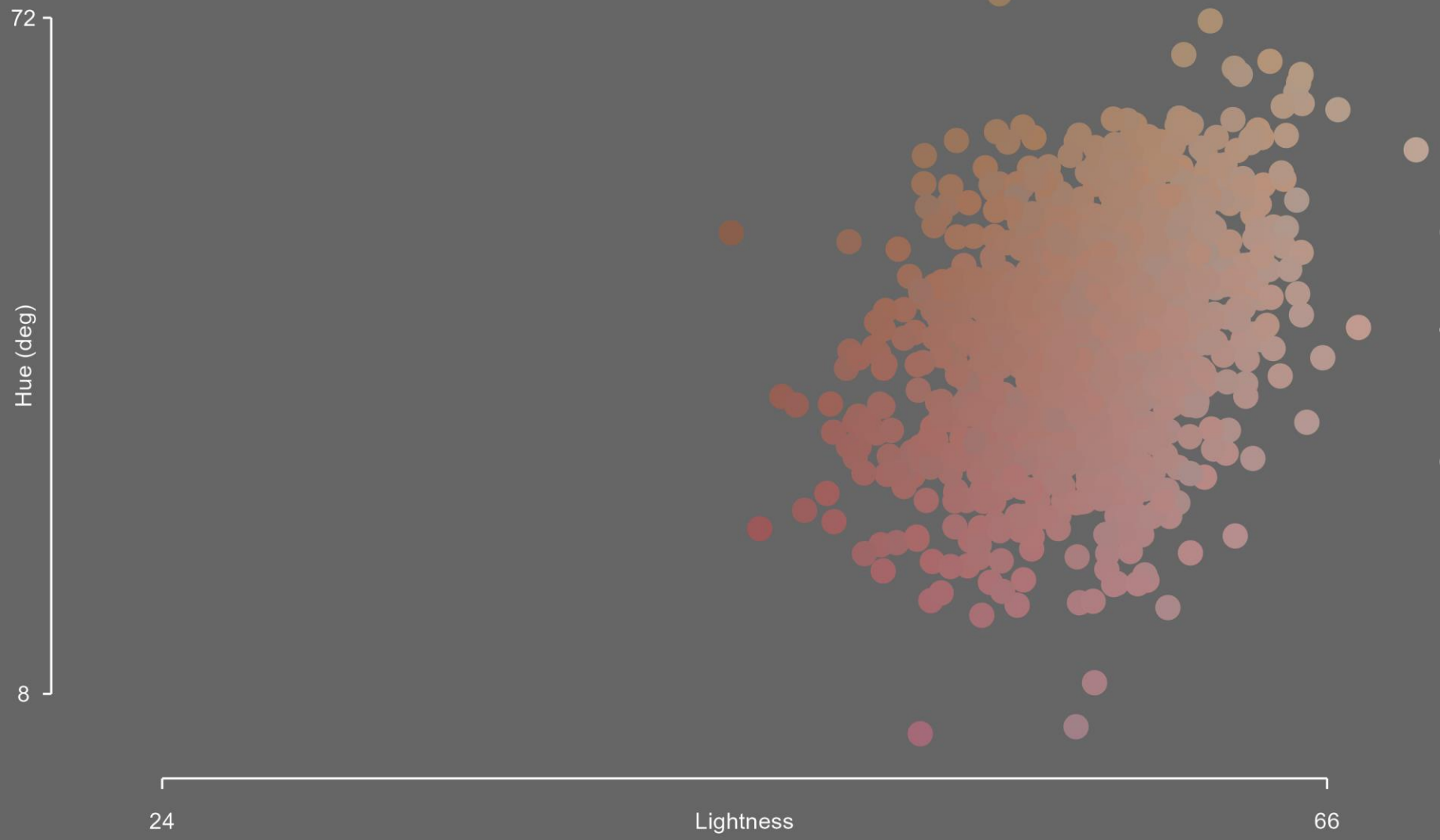
Self report as Hispanic



Gamut (99%):

L*:	36 – 63
a*:	7 – 21
b*:	7 – 24
hue:	25 – 65
chroma:	15 – 29

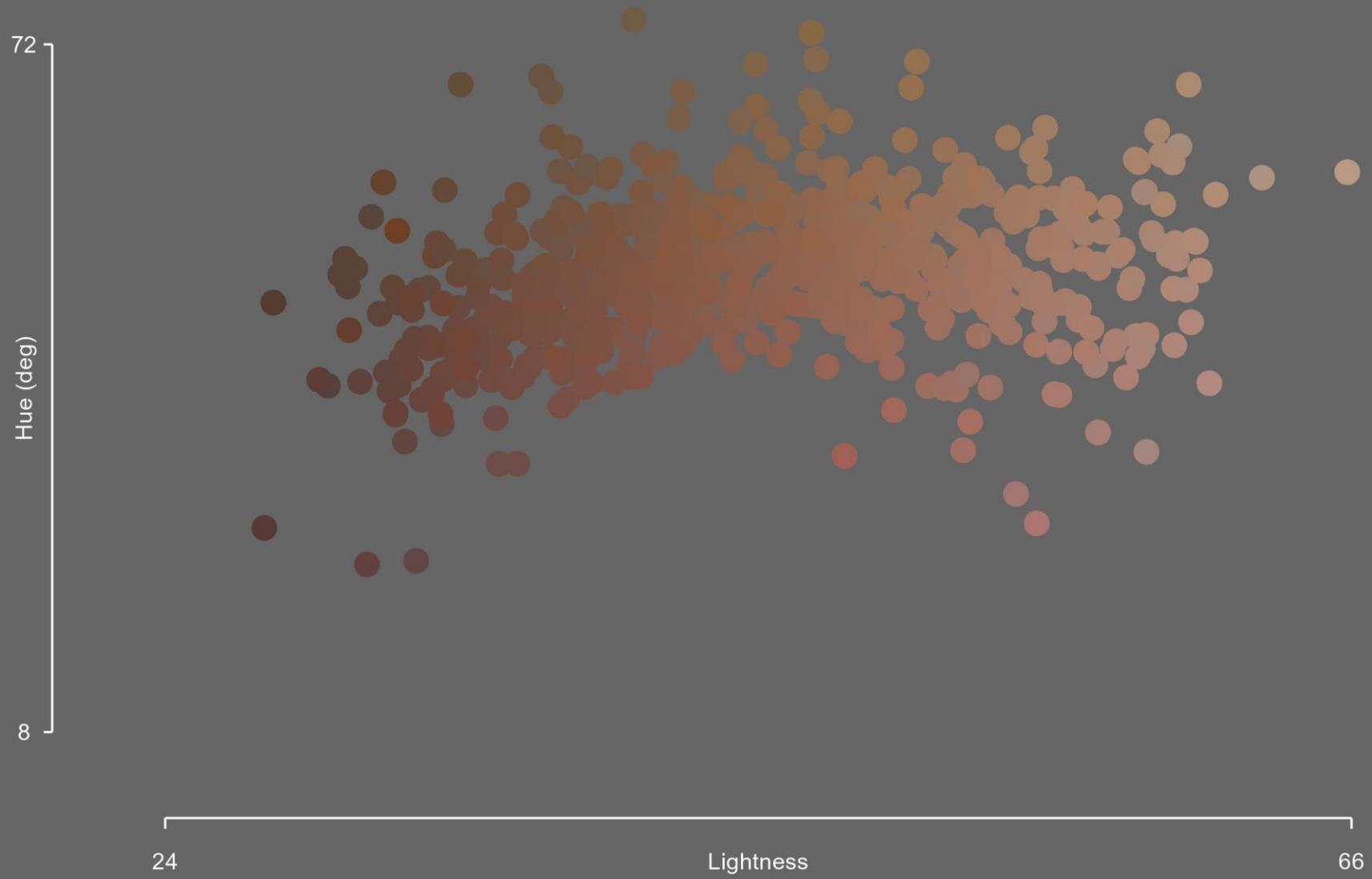
Self report as White



Gamut (99%):

L*:	48 – 65
a*:	6 – 26
b*:	6 – 21
hue:	17 – 67
chroma:	12 – 27

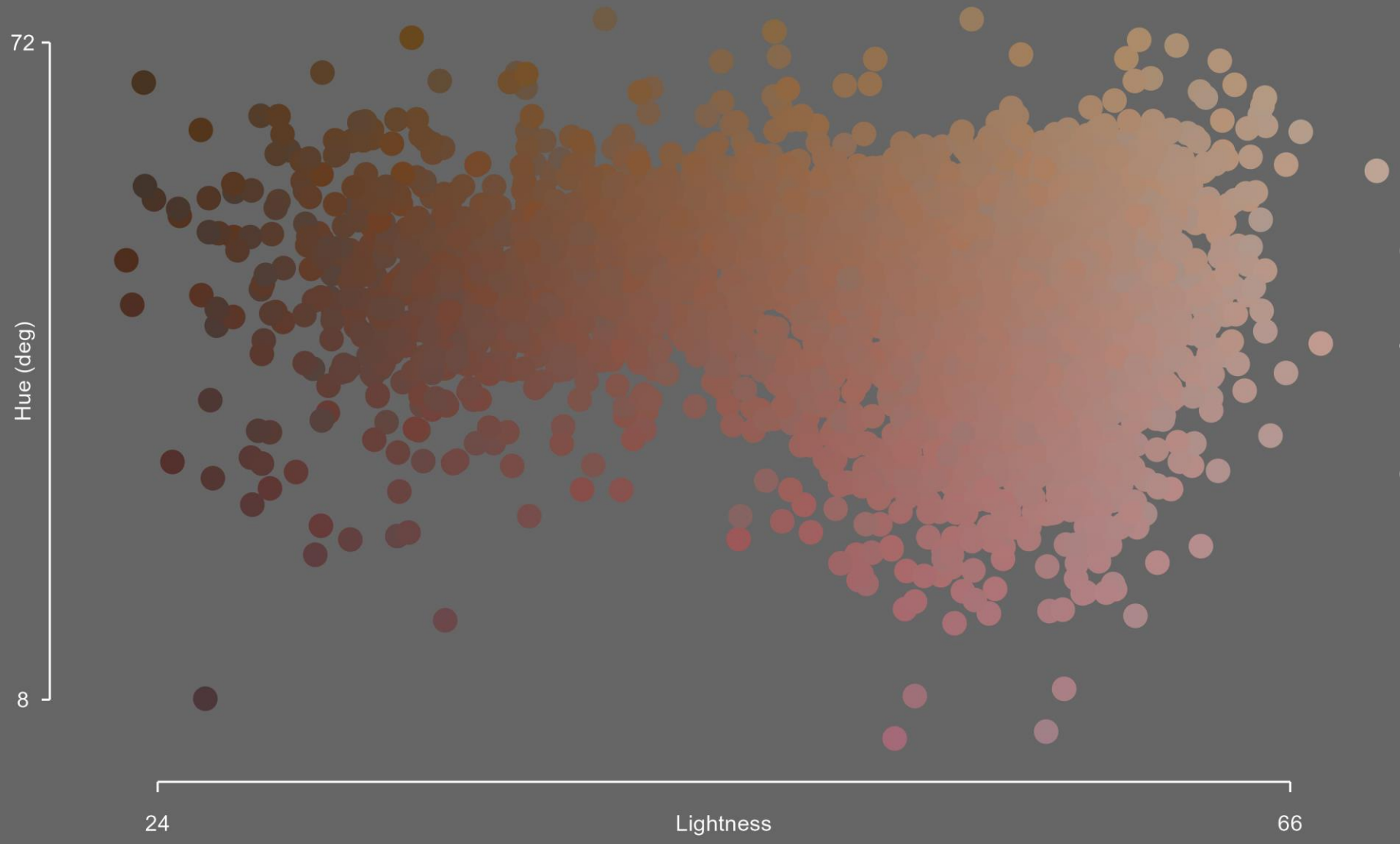
Self report as Multi-Racial or Some Other Race



Gamut (99%):

L*:	30 – 61
a*:	6 – 20
b*:	7 – 25
hue:	28 – 70
chroma:	12 – 29

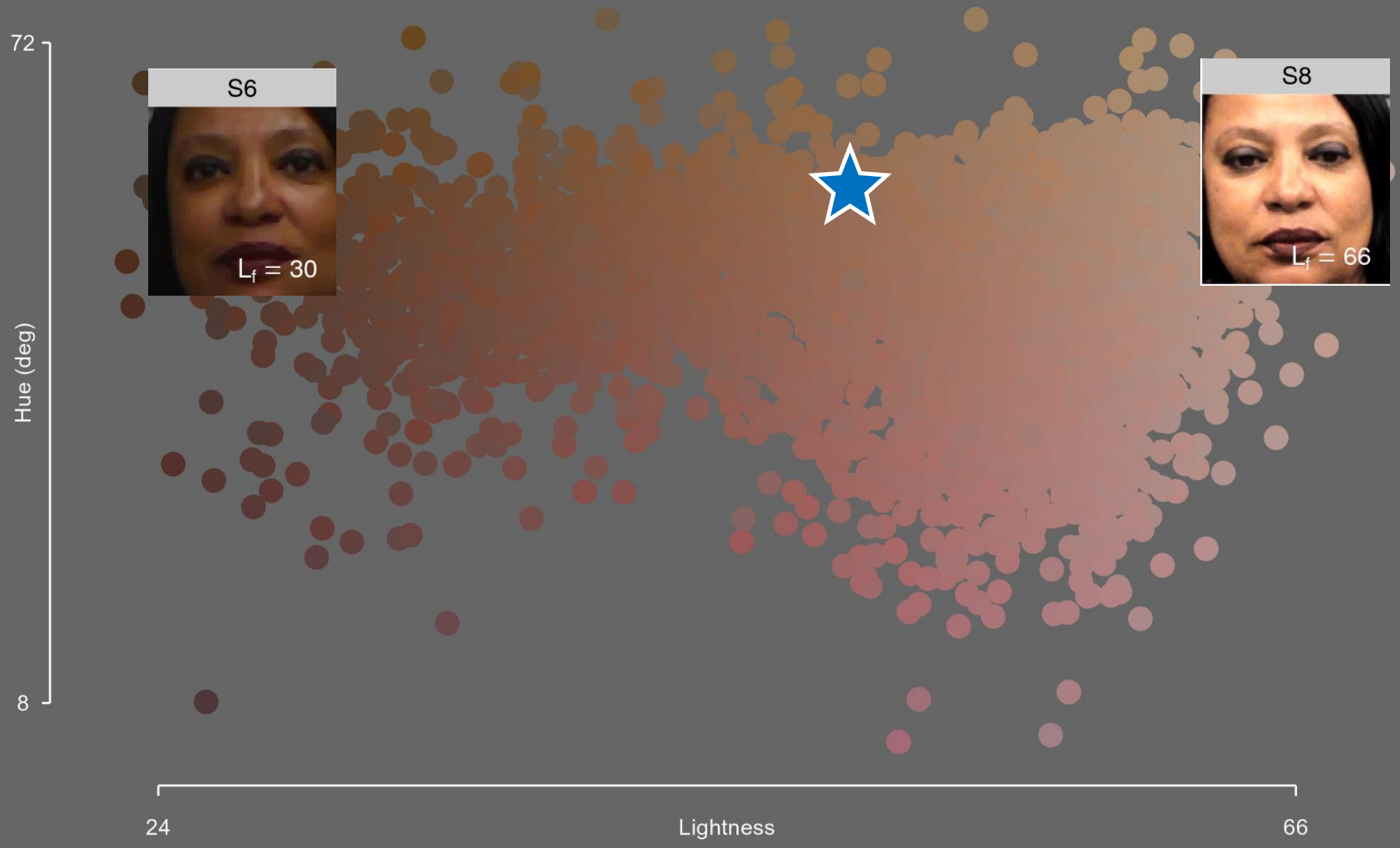
Full gamut of human skin tone (IDSL sample)



Gamut (99.8%):

L*:	24 – 66
a*:	6 – 26
b*:	3 – 27
hue:	8 – 72
chroma:	10 – 31

Full gamut of human skin tone (IDSL sample)



Gamut (99.8%):

L*:	24 – 66
a*:	6 – 26
b*:	3 – 27
hue:	8 – 72
chroma:	10 – 31

Standard reference for human skin tone

- A Standard Reference Material (SRM) that captures the diversity of human face skin tone for use in calibrating digital imaging systems
  - Physical calibration target like a SpyderChecker, but specific to human face skin
  - Multiple versions of the target may be developed for different use-cases

Calibrate to full color gamut

Calibrate to skin tone color gamut

Skin tone SRM use cases

- Camera calibration:
  - Test whether the camera is acquiring quality samples across the full human face skin tone gamut
  - Apply color correction by computing transformation of linear RGB to the SRM target color space

- Image skin-tone labeling:
  - Estimate skin tone of a person from the image with the target present
  - Inform color scales used for labeling images when no target is present

Google/Monk Scale:

Gamut-based Scale:

ISO/IEC 29794-5:2023(CD2)

✓ "Natural" skin color

✗ Not "natural" skin color

Summary

- Failure to acquire (FTA) a sample of sufficient quality is already the main source of error in some face recognition use cases (high-throughput for example)
  - Demographic effects, including those based on skin-tone are already observed here
- New/different quality filters need to be tested for impact to FTA across skin tone, but data is lacking

- We are collecting and sequestering data for testing:
  - Facial skin tone color gamut
  - Face image datasets annotated with calibrated skin tone (e.g., a new pose dataset)

- Sequestered data can be used to:
  - Evaluate performance of quality measures (like pose estimation) across skin tone
  - Develop a standard reference for human facial skin tone
  - This may help us to beyond requiring “natural color” to requiring “accurate color”

Questions: ysirotin@idslabs.org

More Information @ https://mdtf.org/Research/Publications