U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Assessing variation in human skin tone to inform face recognition system design.

Science and Technology

Yevgeniy B. Sirotin
Technical Director
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.

Science and Technology

Biometric & Identity Technology Center

Vision

- **Drive biometric and identity innovation** at DHS through RDT&E capability
- **Facilitate and accelerate understanding of biometrics and identity technologies** for new DHS use cases
- Follow "**Build once, use widely**" approach

Goals

- **Drive efficiencies** by supporting cross cutting methods, best practices, and solutions across programs
- **Deliver Subject Matter Expertise** across the DHS enterprise
- **Engage Industry** and provide feedback
- **Encourage Innovation** with Industry and Academia

Science and Technology

DHS S&T Scenario Testing of Biometric Technology

- Address technology risks prior to deployment:
  - Centered around a specific use-case (e.g., land border)
  - Gathering new biometric samples
  - Full multi-component biometric system

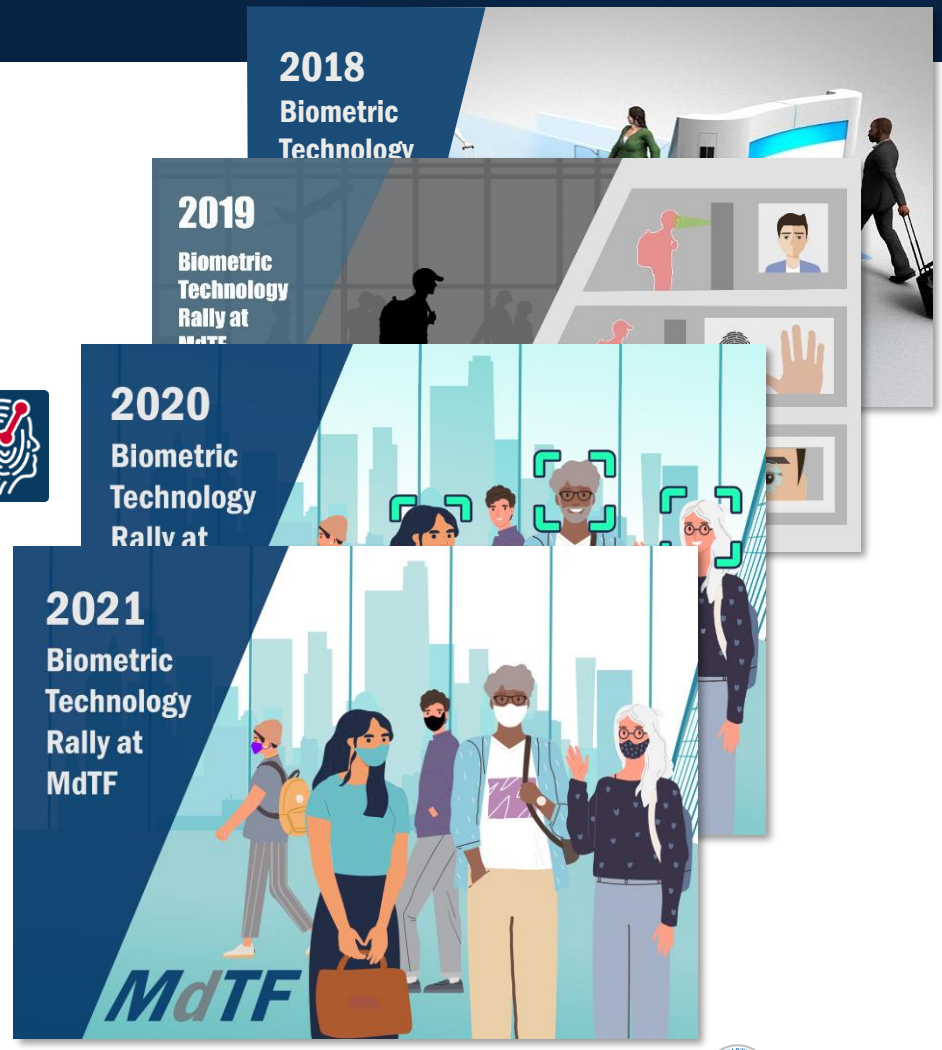
- Outcomes:
  - Understand technology performance in use
  - Inform stakeholders about technology performance
  - Direct engineering resources to the main causes of error

DHS S&T Scenario Testing of Face Recognition Technology

- Since 2018, over 200 commercial systems selected for testing by a panel of experts

- Testing performed by the Identity and Data Sciences Laboratory at the Maryland Test Facility

- Tests provide comprehensive metrics:
  - Efficiency – transaction times
  - Effectiveness – image capture and matching success
  - Satisfaction – user feedback
  - Equitability – performance across demographic groups
  - https://mdtf.org



IDSL

Science and Technology

Key Takeaways from the 2021 Biometric Technology Rally

DO MASKS AFFECT PERFORMANCE?

MASKS REDUCE FACE RECOGNITION PERFORMANCE.*

Without Masks

95% Of All People Succesfully Identified

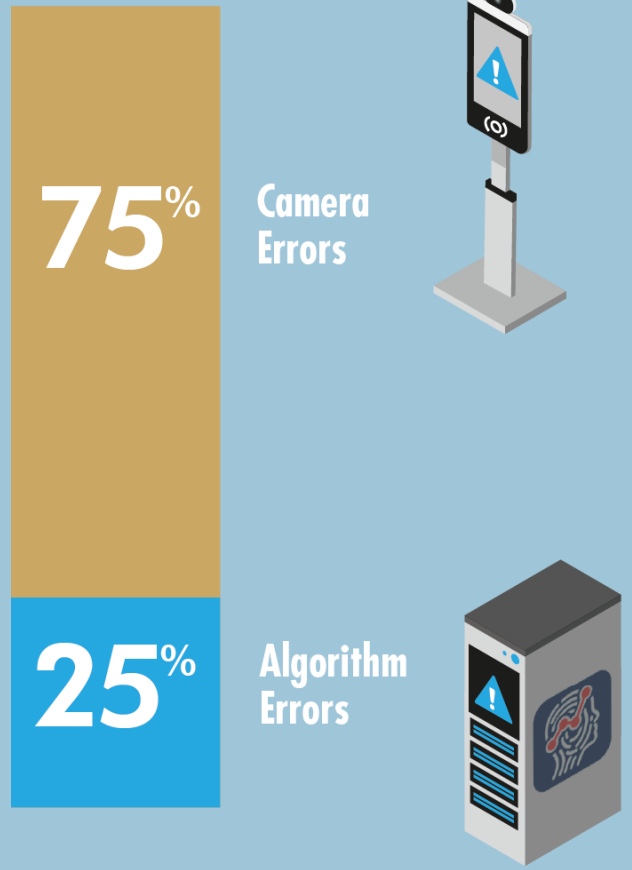
With Masks

86% Of All People Succesfully Identified

Science and Technology

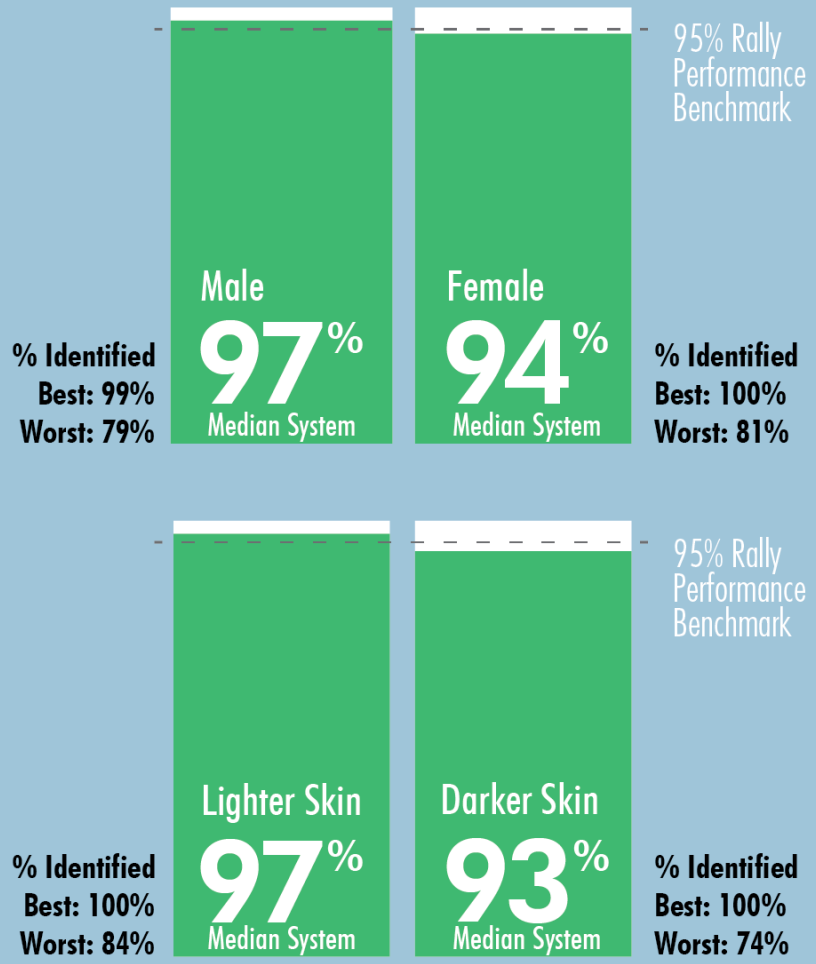
WHAT CAUSES ERRORS?

MOST ERRORS ARE DUE TO PHOTO CAPTURE, NOT MATCHING.**



HOW DOES PERFORMANCE VARY?

PERFORMANCE VARIES BY GENDER AND SKIN TONE.* **



*Numbers are representative of the median system combination (25th best system out of 50 total systems) in each test condition.

**Results from system combinations tested without masks.

Prepared by the Identity and Data Sciences Lab

Key Takeaways from the 2021 Biometric Technology Rally

DO MASKS AFFECT PERFORMANCE?

M
RECOG

Without Masks

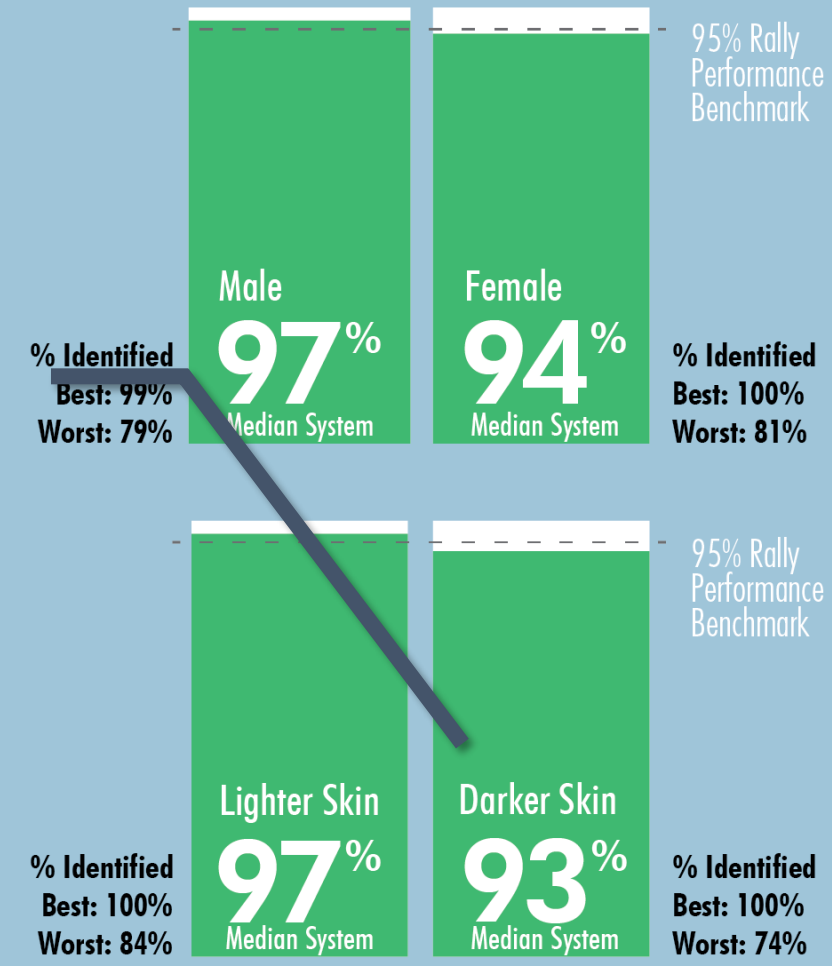
With Masks

WHAT CAUSES ERRORS?

HOW DOES PERFORMANCE VARY?

PERFORMANCE VARIES BY GENDER AND SKIN TONE.* **

- **Top performing systems identified >98% of people in each group**
- **Median system failed to meet Rally threshold for volunteers with darker skin tone**
- **The worst performing system had a 10% difference based on skin tone**



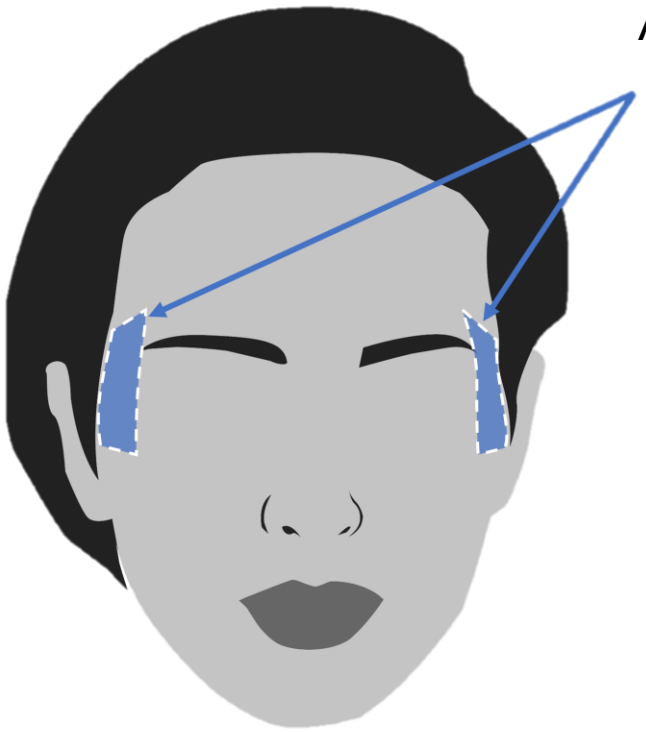
Science and Technology

*Numbers are representative of the median system combination (25th best system out of 50 total systems) in each test condition.

**Results from system combinations tested without masks.

Prepared by the Identity and Data Sciences Lab

Facial Skin Tone – MdTF Sample

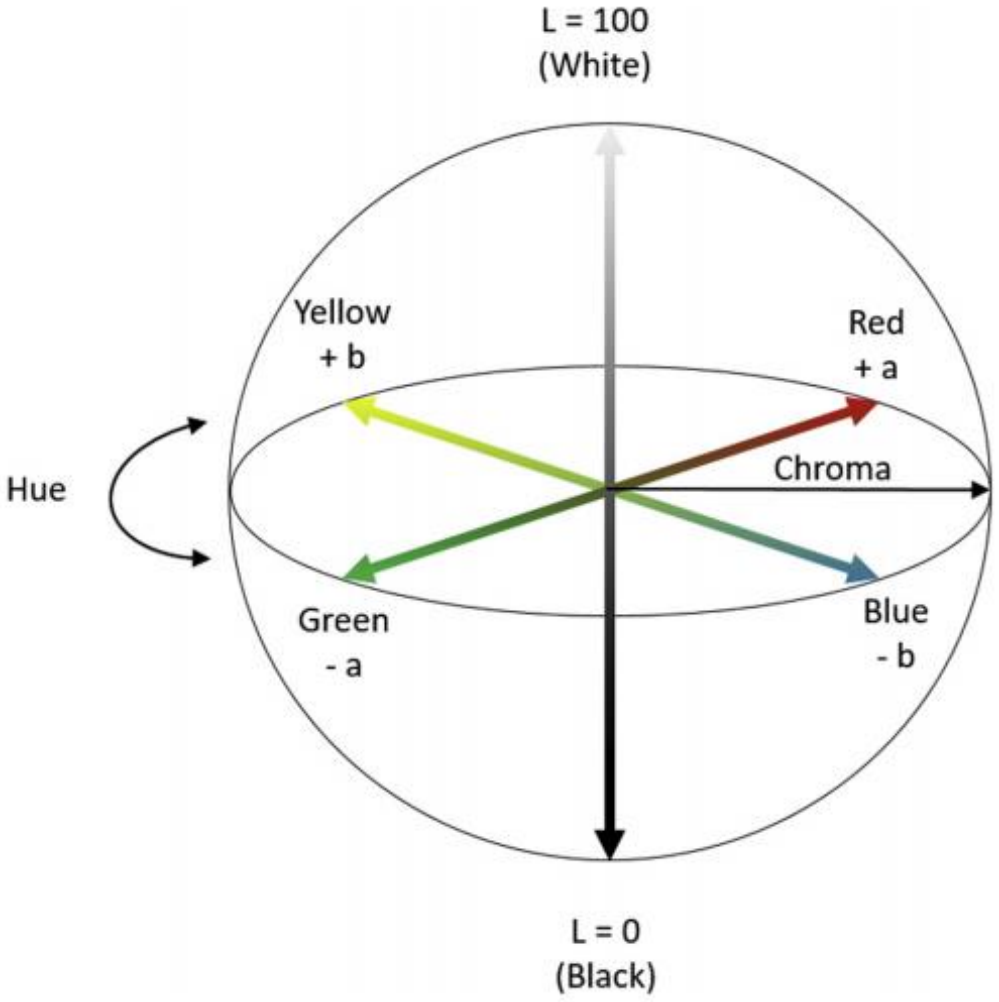


One reading each from the left and the right temple.
Average value computed.

1,000+ unique volunteers.
Diverse race, gender, age.
2,000+ color samples.

DSM III Colormeter
Cortex Technology

CIELAB color space



Distances in CIELAB color space are normalized for human perception.

Lightness: L*
Red/green: a*
Yellow/blue: b*

$$\Delta E = \sqrt{(\Delta L)^2 + (\Delta a)^2 + (\Delta b)^2}$$

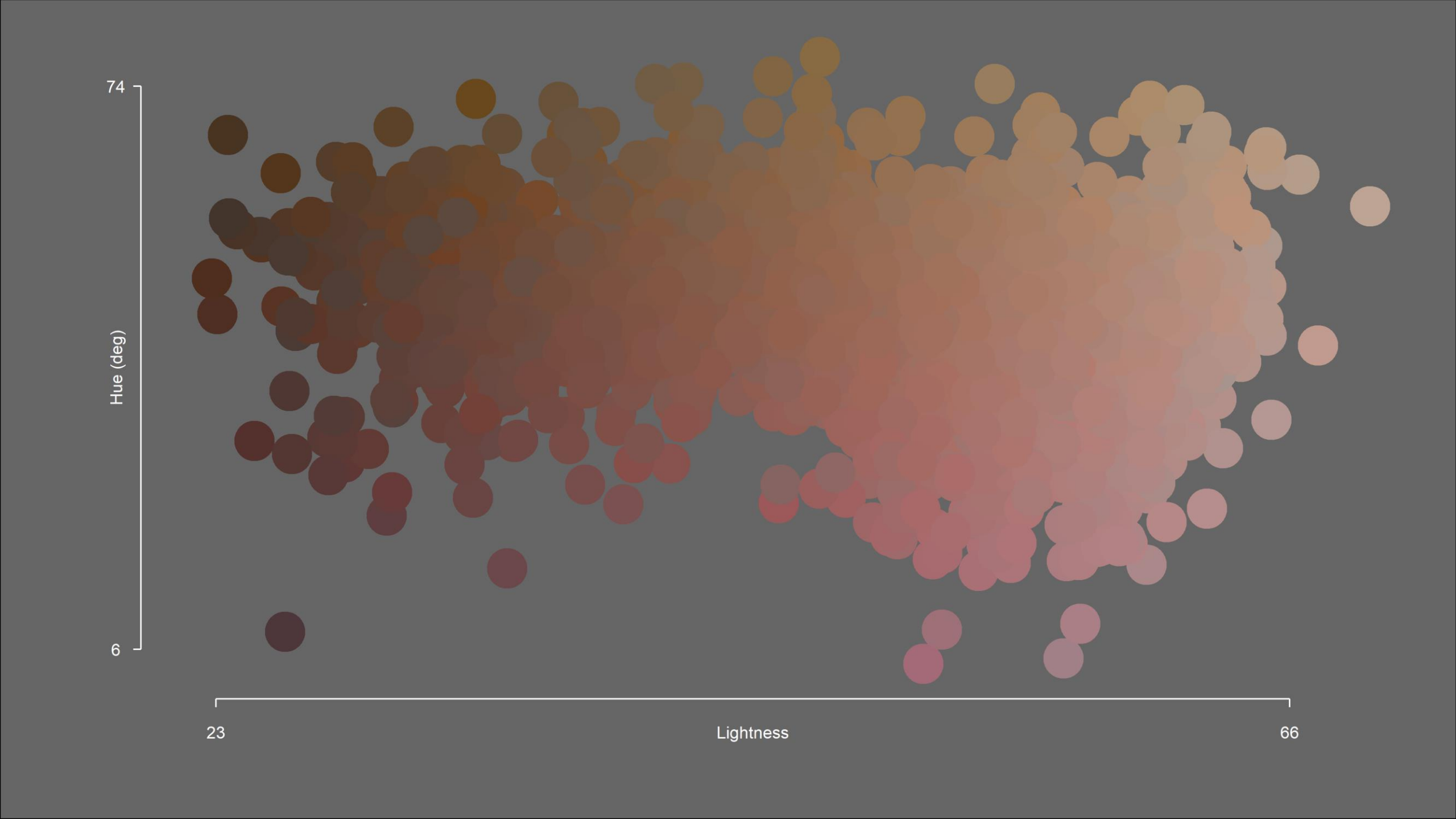
$\Delta E = 2.3$ is a just-noticeable difference in human perception.

Hue and chromaticity are more intuitive means of describing color than *a* and *b*:

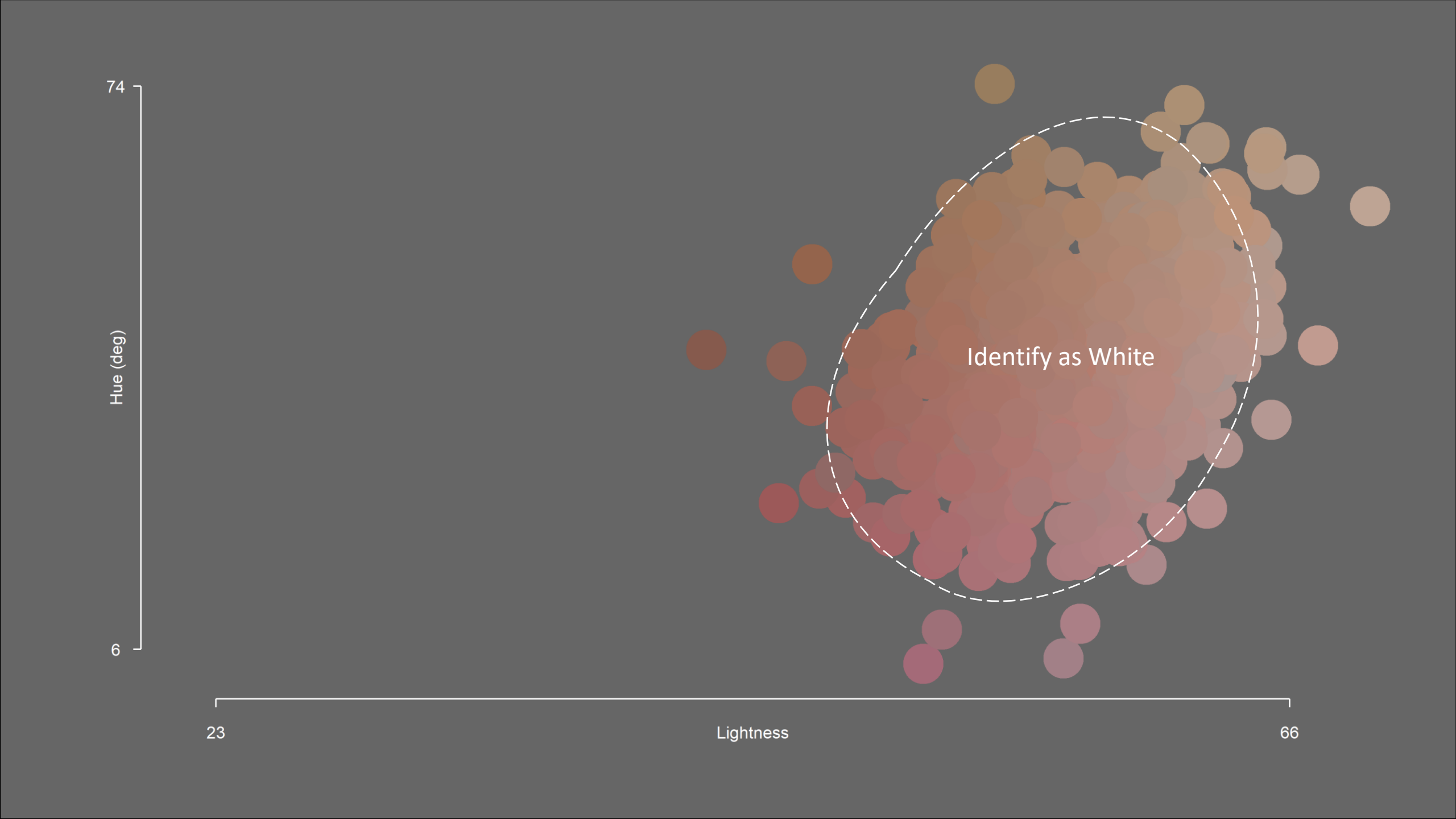
$$\text{Hue} = \frac{180}{\pi} \text{atan}\left(\frac{b}{a}\right)$$

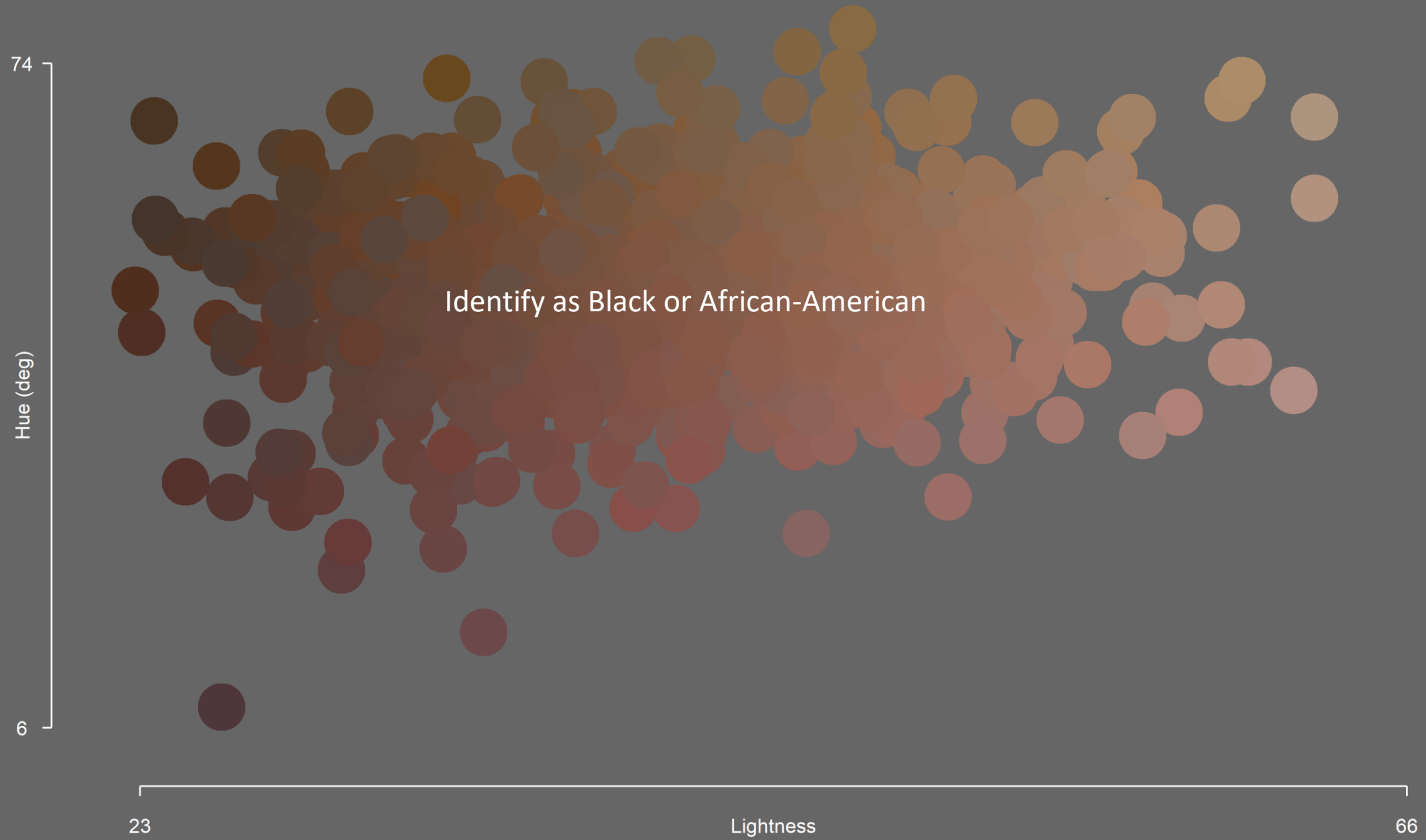
$$\text{Chromaticity} = \sqrt{a^2 + b^2}$$

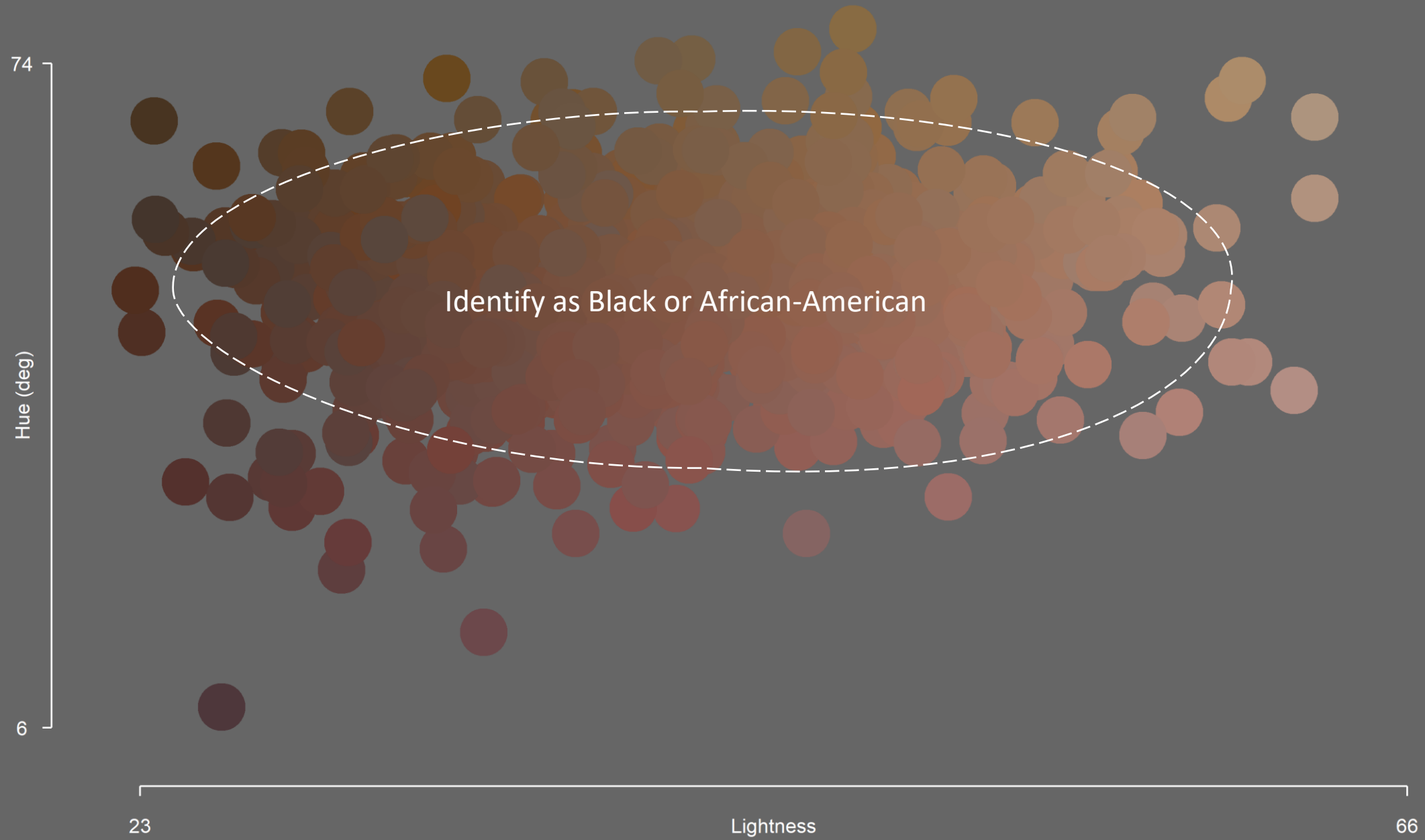
Science and Technology

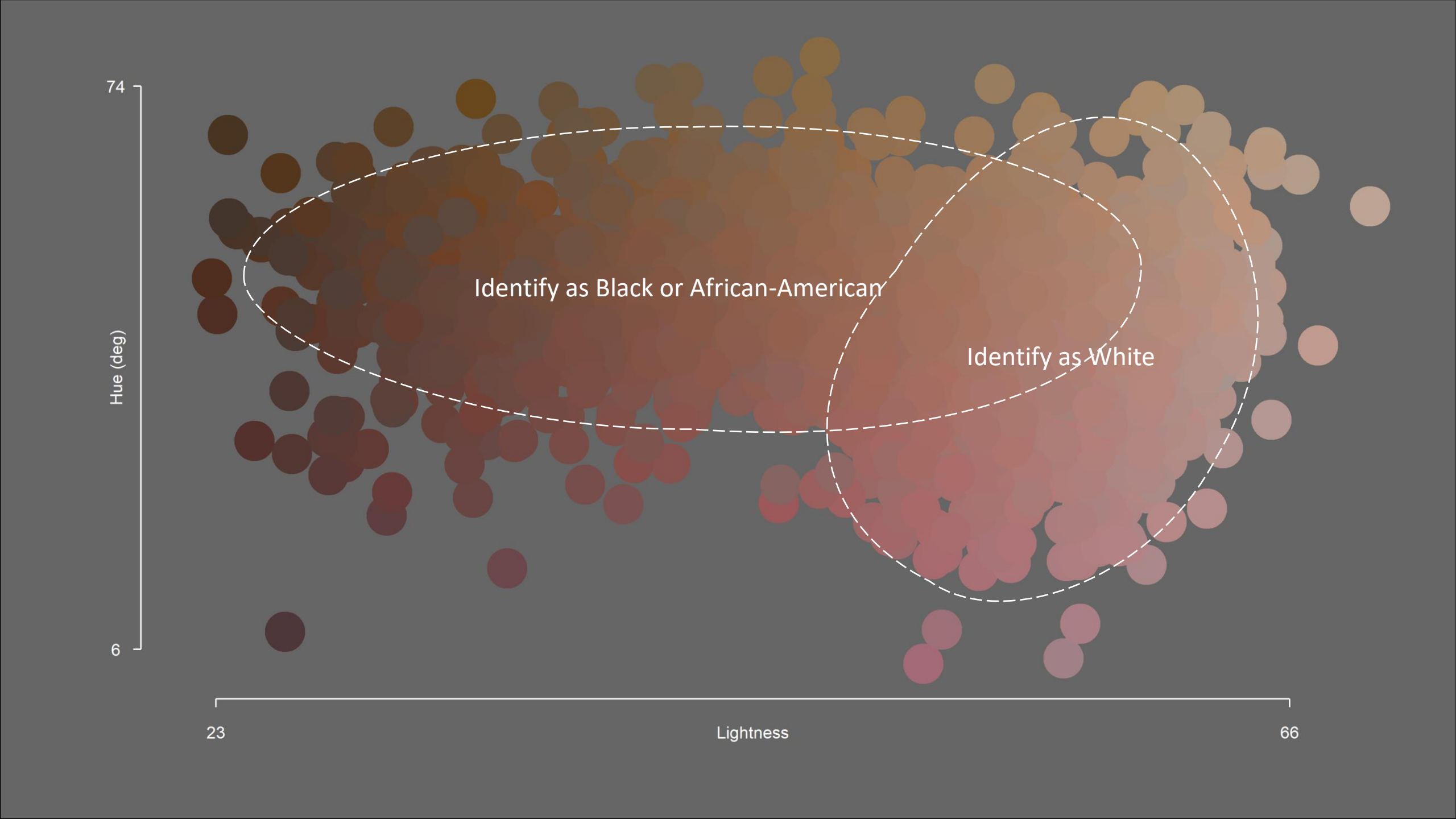

74
Hue (deg)
6
23
Lightness
66


74
Hue (deg)
6
Identify as White
23
Lightness
66


74
Hue (deg)
6
Identify as White
23
Lightness
66


Identify as Black or African-American
74
Hue (deg)
6
23
Lightness
66


74
Hue (deg)
6
Identify as Black or African-American
23
Lightness
66

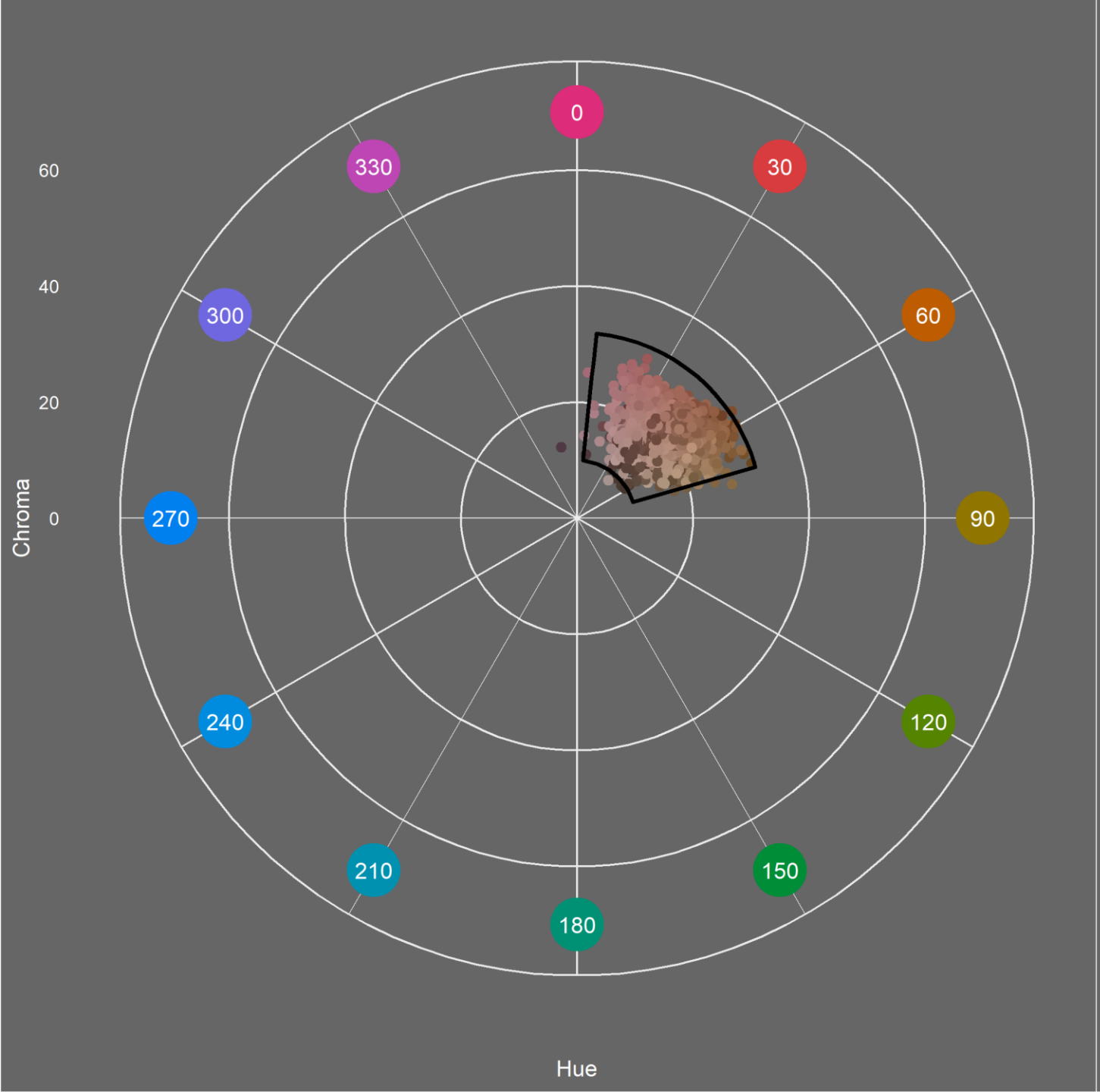

74
Hue (deg)
6
Identify as Black or African-American
Identify as White
23
Lightness
66

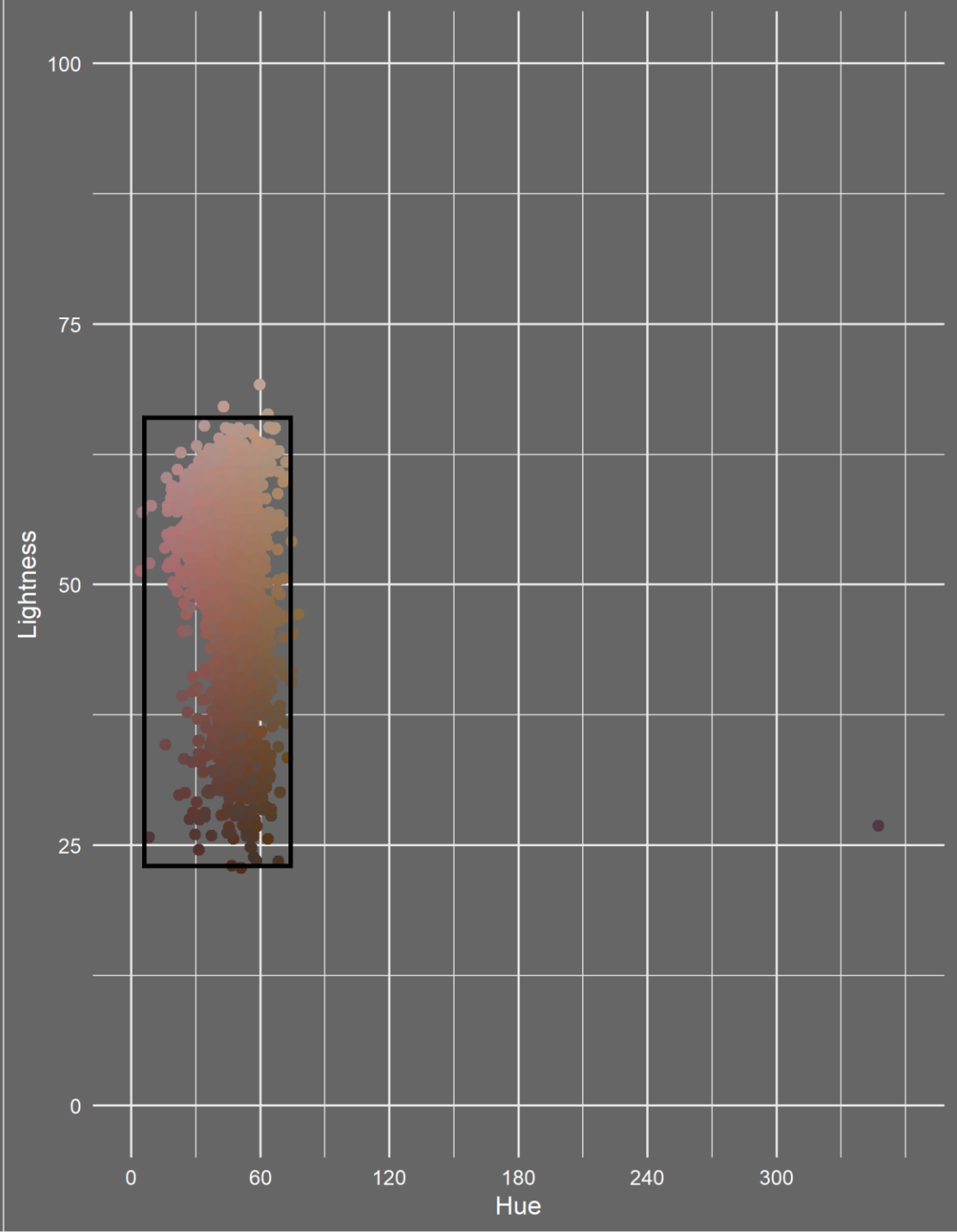
Face skin tone: measured “natural” range

- Face image quality standards specify that face images have no “unnatural color”
  - ISO/IEC 19794-5 Biometric data interchange formats - Part 5: Face image data requires appropriate white balance
  - ISO/IEC 39794-5 Extensible biometric data interchange formats - Part 5: Face image data discusses “unnatural skin tone” in CIELAB space
  - ISO/IEC WD 29794-5 Biometric sample quality — Part 5: Face image data: describes a measure of the degree of face color “unnaturalness”

- MdTF samples suggest a “natural” skin tone range can be defined
  - Range of values in MdTF sample:

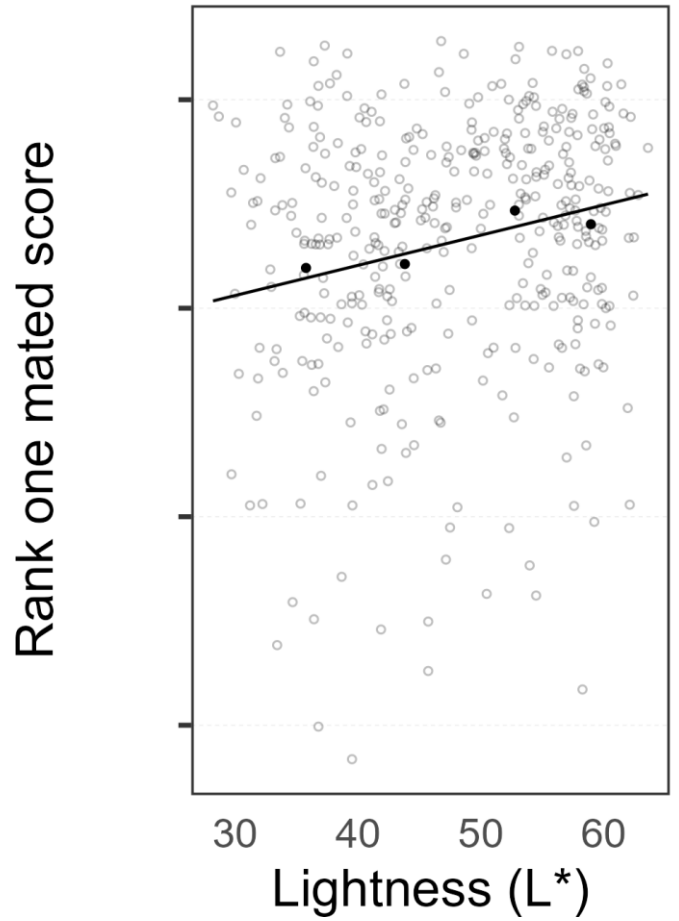
Lightness:	23 – 66
a*:	5 – 26 (avg = 15)
b*:	2 – 28 (avg = 16)
Hue:	6 – 74 degrees
Chromaticity:	10 – 32


0
30
60
90
120
150
180
210
240
270
300
330
60
40
20
0
Chroma
Hue


100
75
50
25
0
Lightness
0
60
120
180
240
300
Hue

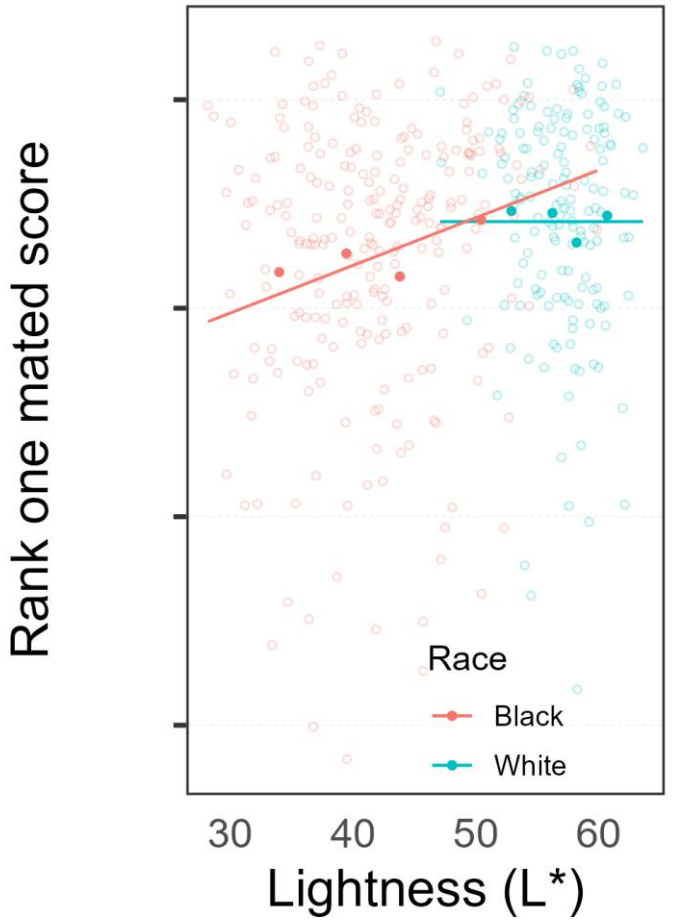
Relation to Biometric System Performance

- Acquisition:
  - Failure to Acquire is greater for volunteers with darker skin tone.
- Matching:
  - Rank one mated scores are higher for those with lighter skin tone (Cook et al., TBIOM 2018)
  - Relation of scores with skin tone is stronger than with Race
  - Skin tone effects found for >50% of acquisition-matching system combinations tested in DHS S&T Rallies (85 of 158)



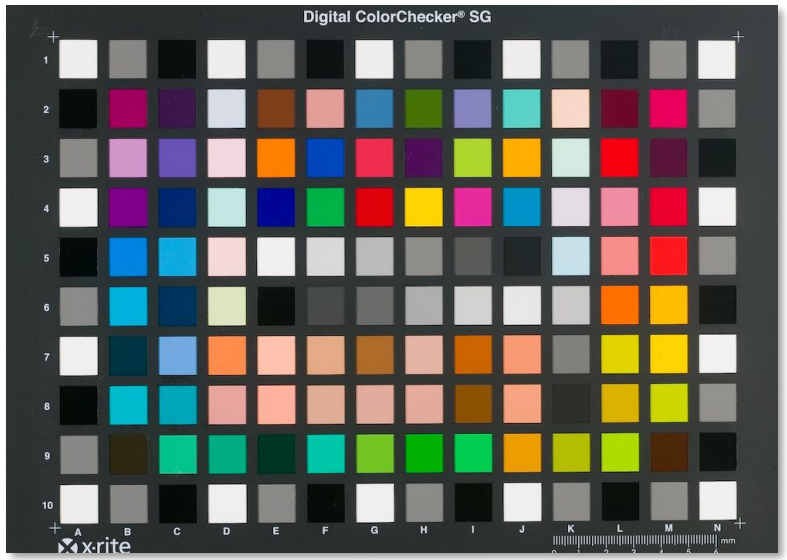
Relation to Biometric System Performance

- Acquisition:
  - Failure to Acquire is greater for volunteers with darker skin tone.
- Matching:
  - Rank one mated scores are higher for those with lighter skin tone (Cook et al., TBIOM 2018)
  - Relation of scores with skin tone is stronger than with Race
  - Skin tone effects found for >50% of acquisition-matching system combinations tested in DHS S&T Rallies (85 of 158)
  - Relation of scores with skin tone exists for volunteers identifying as Black or African-American, but not for those that identify as White.
  - Relationship between skin tone and mated score can vary across acquisition systems.
- Some of these effects may be due to poor quality of acquired imagery.



Science and Technology

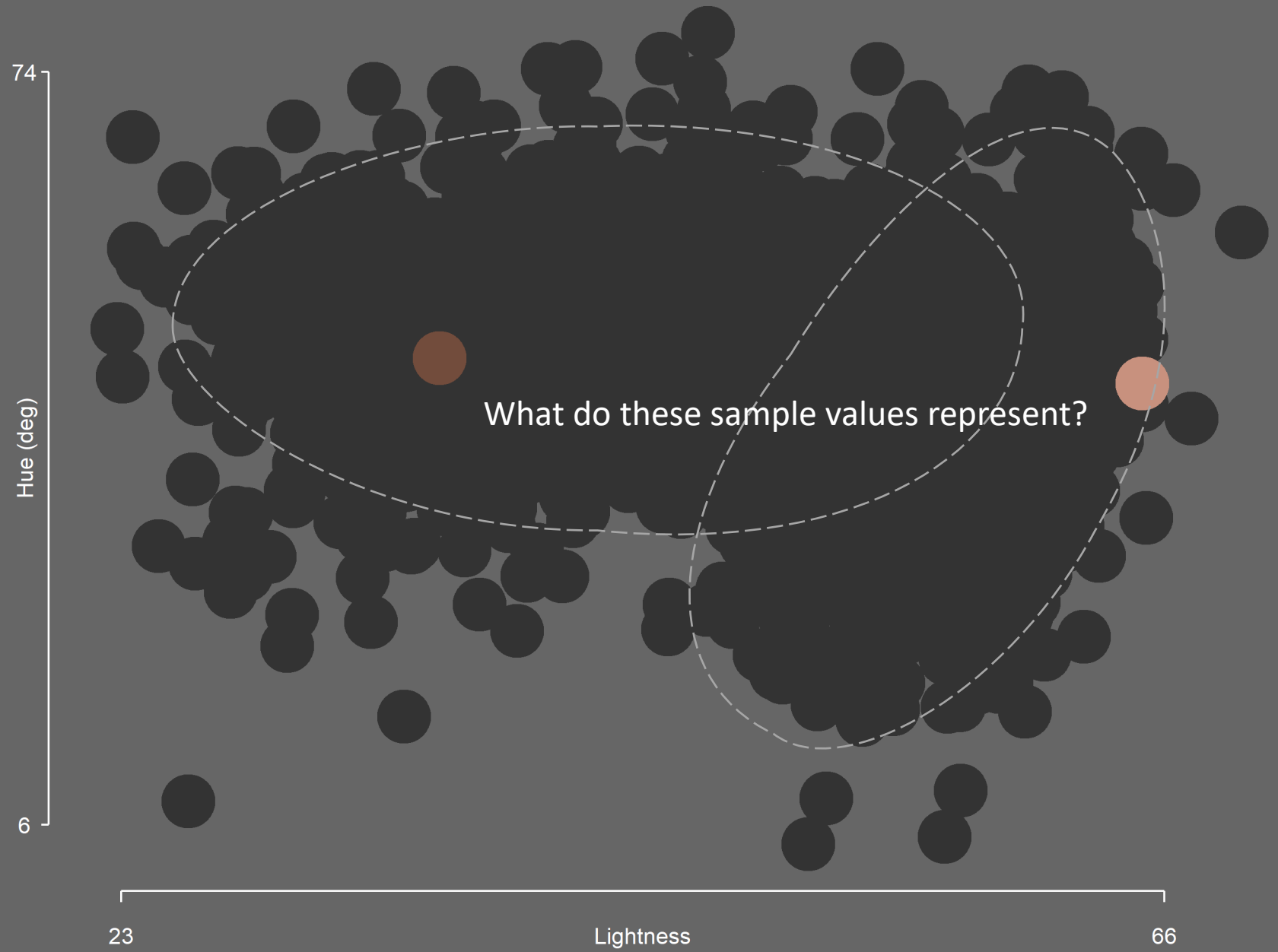
Color Calibrating Cameras



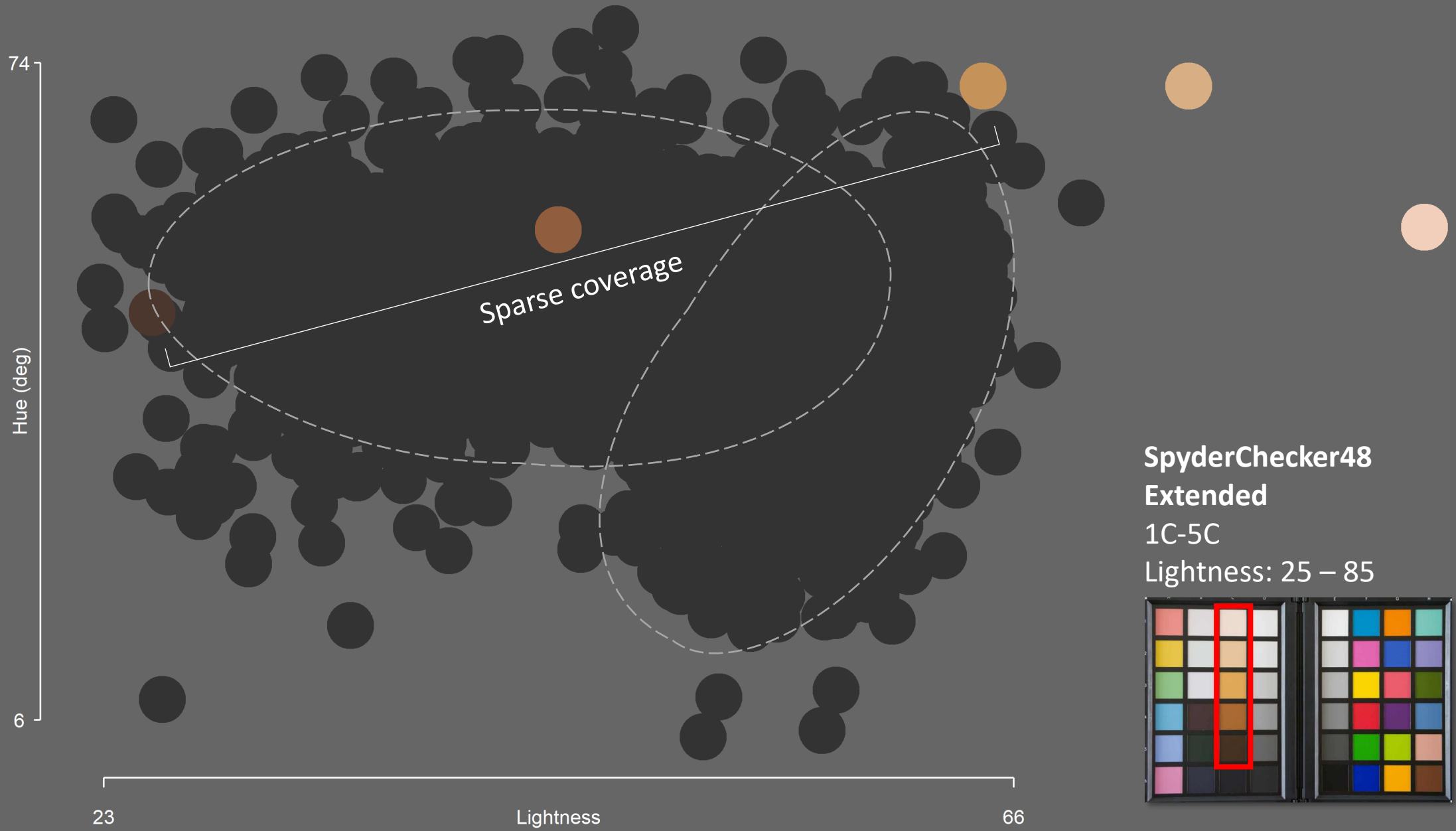
CIELAB values are provided for each color element in these color checkers.

Goal is to minimize color error across color squares (i):

$$\sum_i \sqrt{(\Delta L_i)^2 + (\Delta a_i)^2 + (\Delta b_i)^2}$$



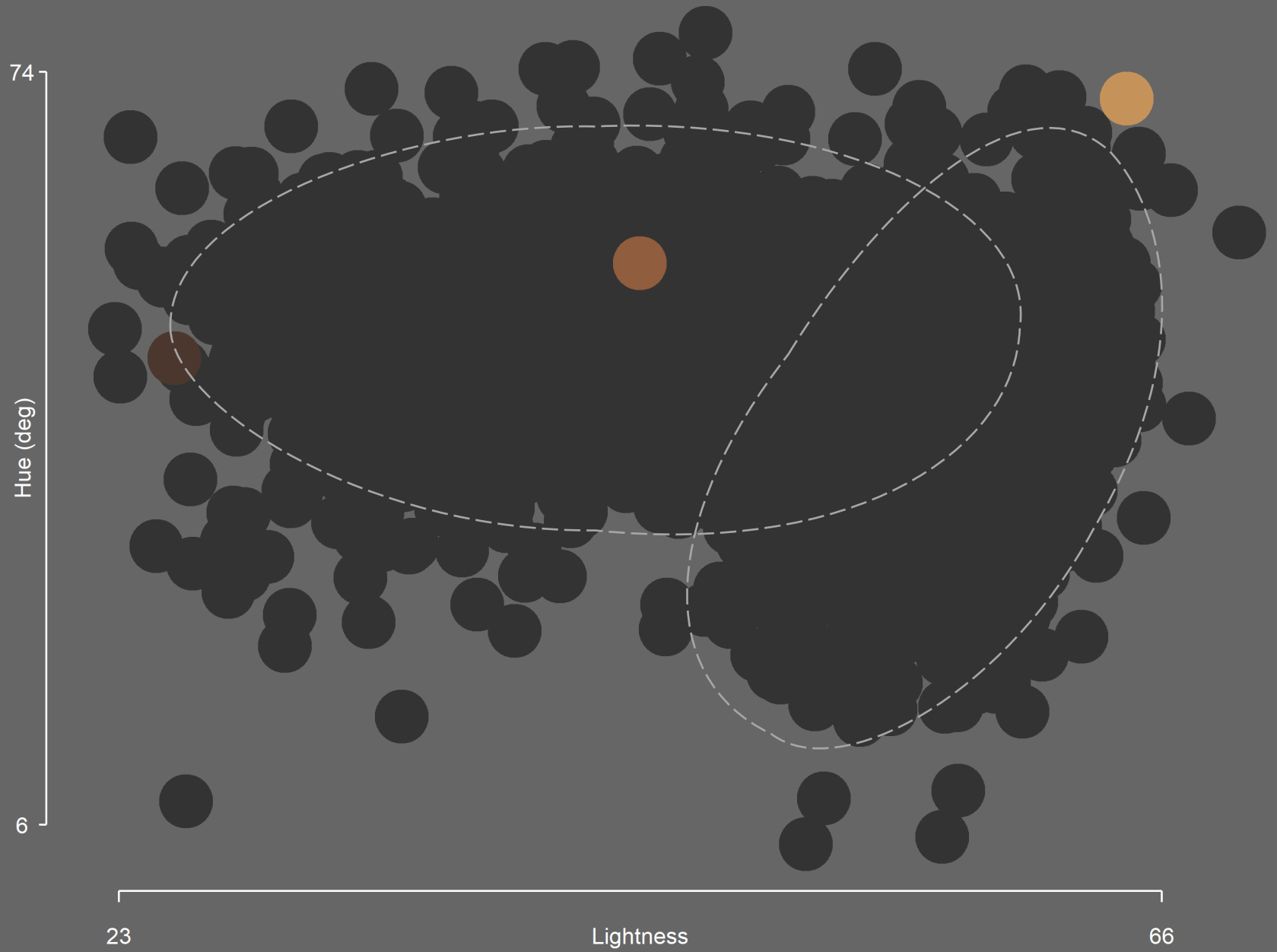
SpyderCheckr48
Classic Macbeth Chart
5H-6H
Lightness: 36 – 65



SpyderChecker48 Extended

1C-5C

Lightness: 25 – 85

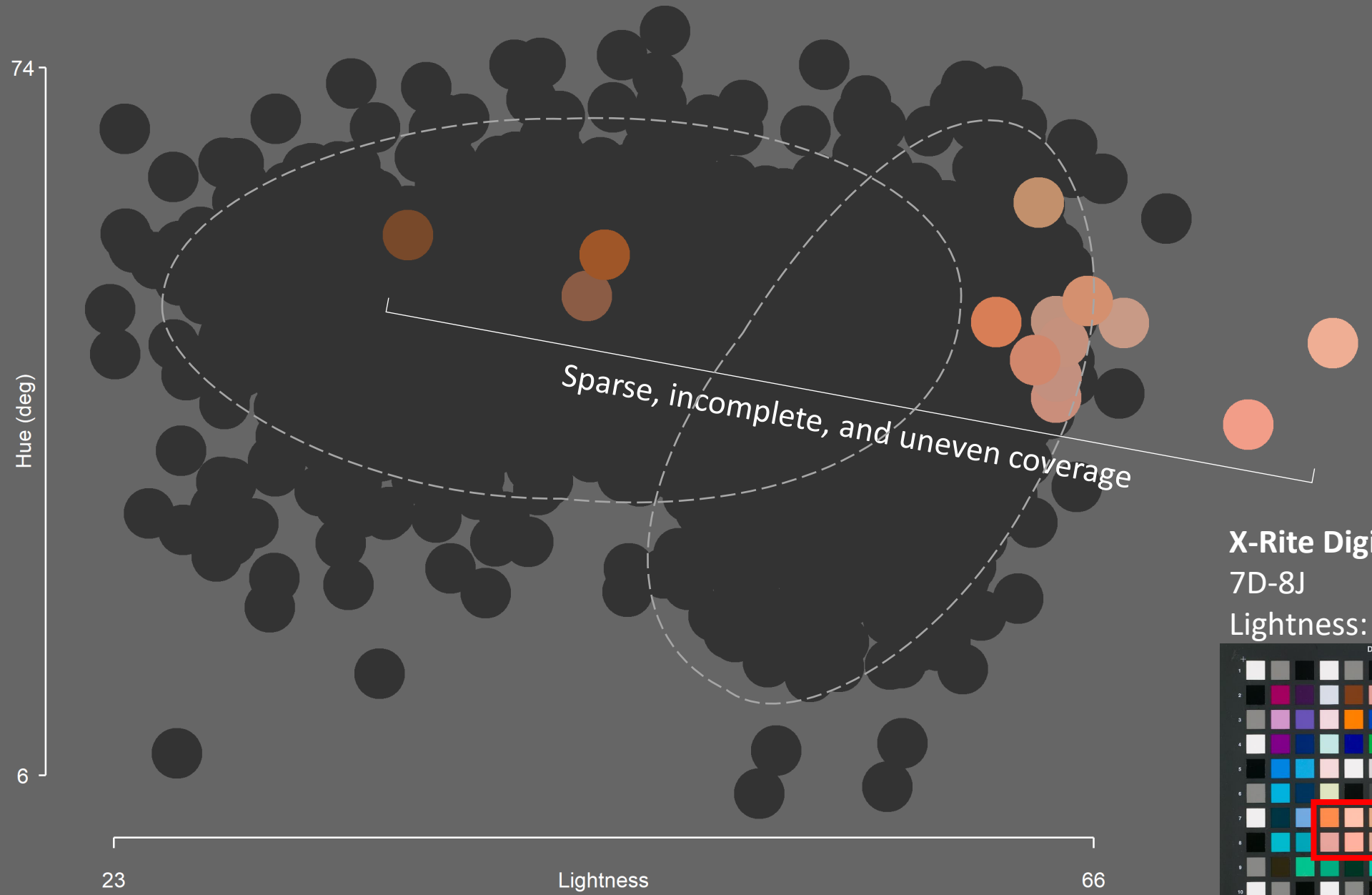


Too light for face skin

SpyderChecker48 Extended

1C-5C

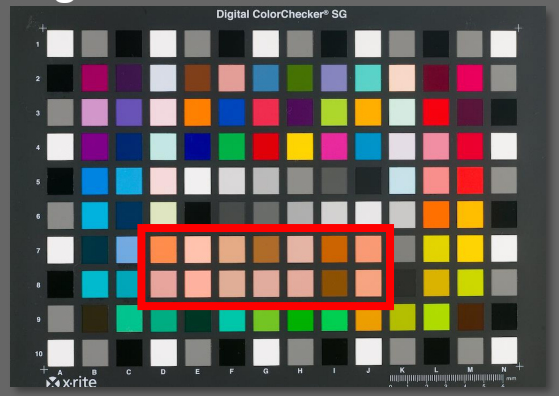
Lightness: 25 – 85

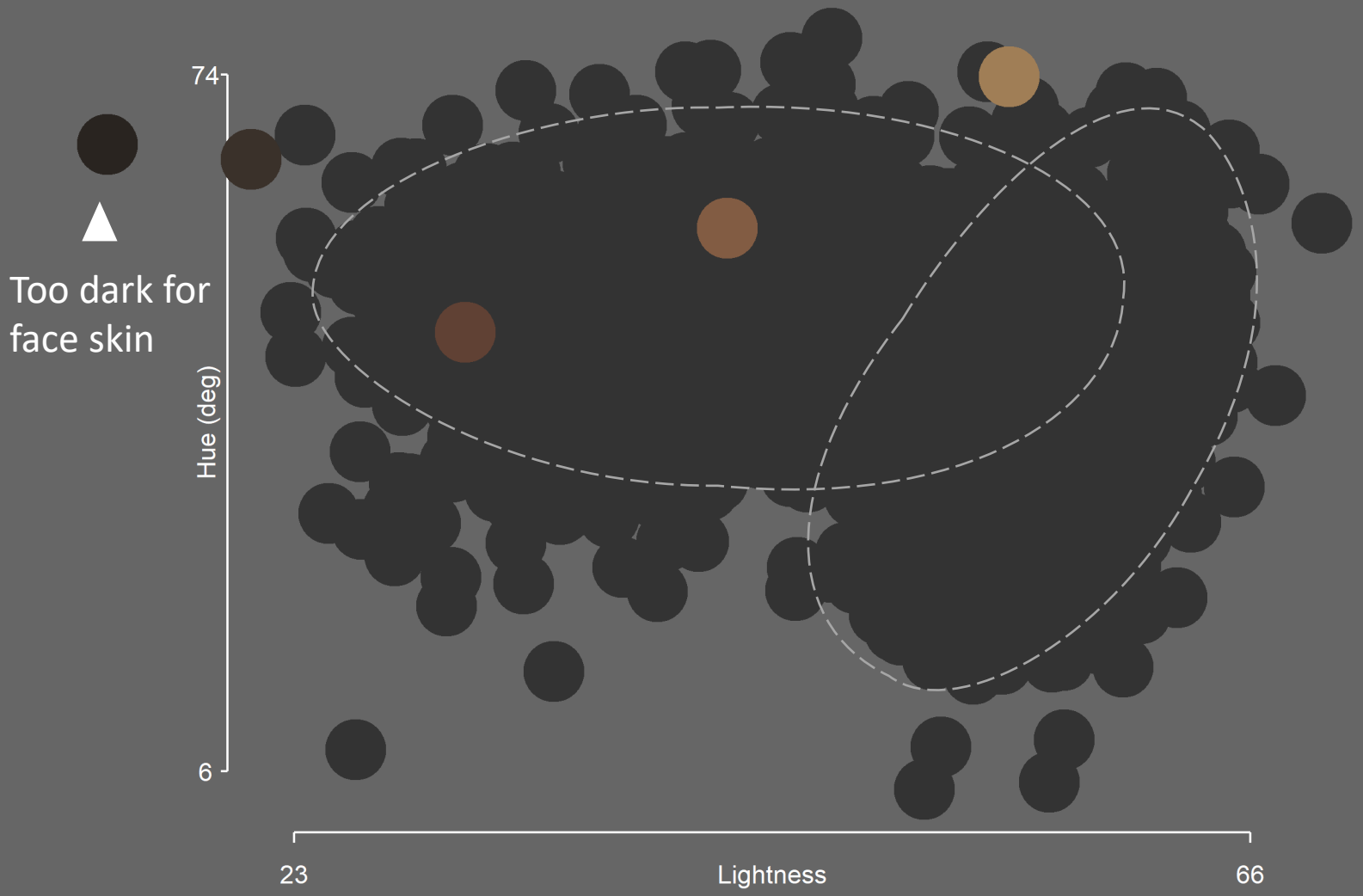


X-Rite Digital ColorChecker SG

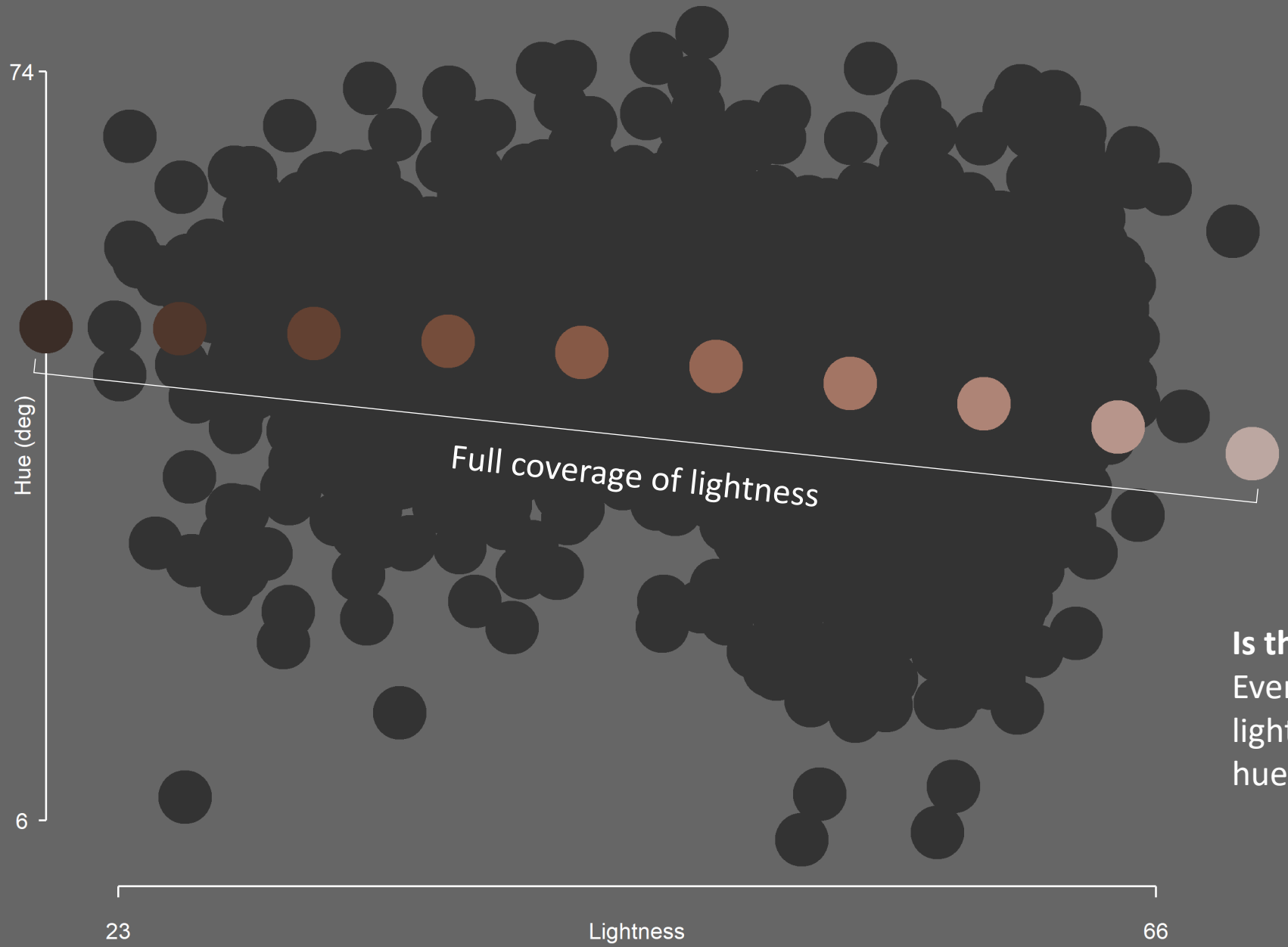
7D-8J

Lightness: 36 – 77





Google Monk Scale
Developed for Labeling Images in Studies of AI Fairness
Lightness: 15 – 94



Is there a better approach?
Evenly sample extended lightness range at appropriate hue and chromaticity.

Implications for Acquisition Systems

- Face recognition systems should be able to maintain performance for face targets within “natural” CIELAB color range:
  - Lightness: 23 – 66
  - Hue: 6 – 74 degrees
  - Chromaticity: 10 - 32

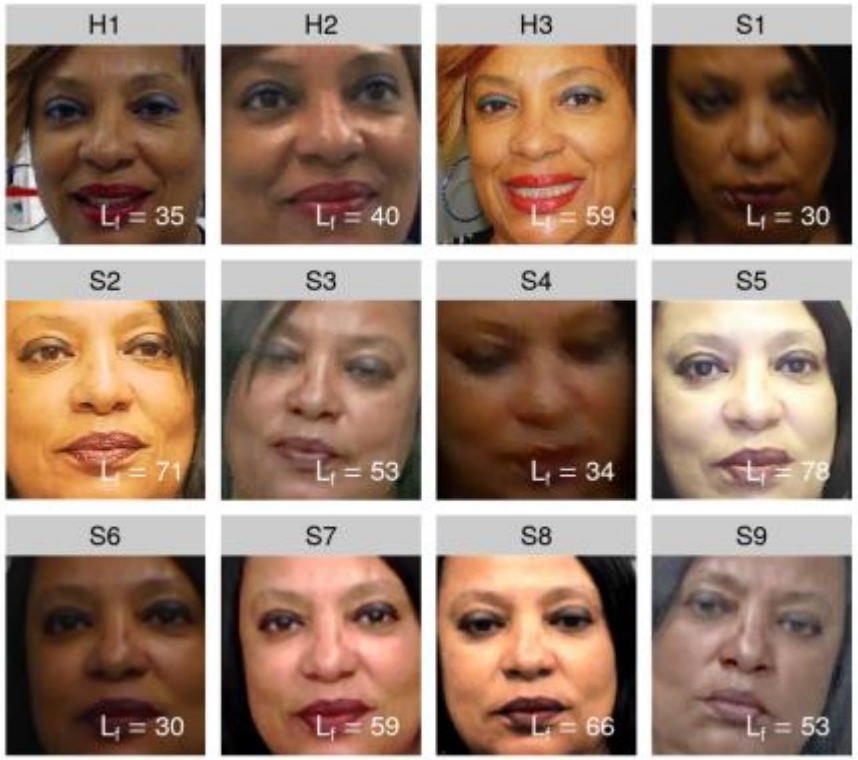
- Can measure reproduction of CIELAB color within defined error from target
  - $\Delta E = \sqrt{(\Delta L)^2 + (\Delta a)^2 + (\Delta b)^2}$

- Optimize for diffuse – not specular reflection
  - Bahmani et al., IWBF 2021

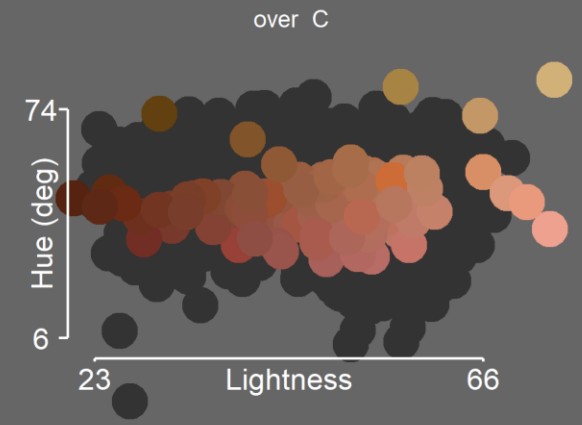
Implications for Face Image Datasets

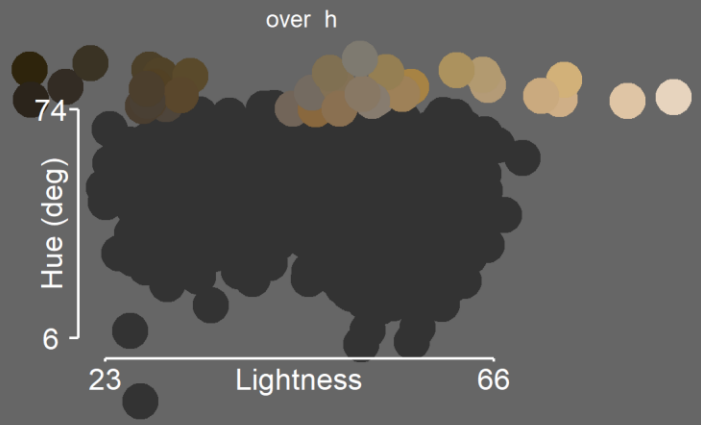
- What are we using to train and evaluate face recognition?
- Skin tone in images can vary – some datasets may have face images with values outside the natural range.
- This indicates dataset images are improperly color calibrated.
- A measure of face dataset “color health”:

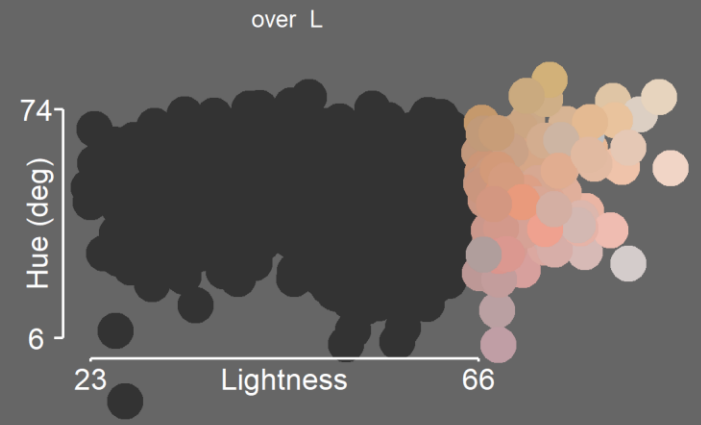
$$CH_{dataset} = \frac{1}{N}\sum_{i=1}^{N} c_i \in C$$



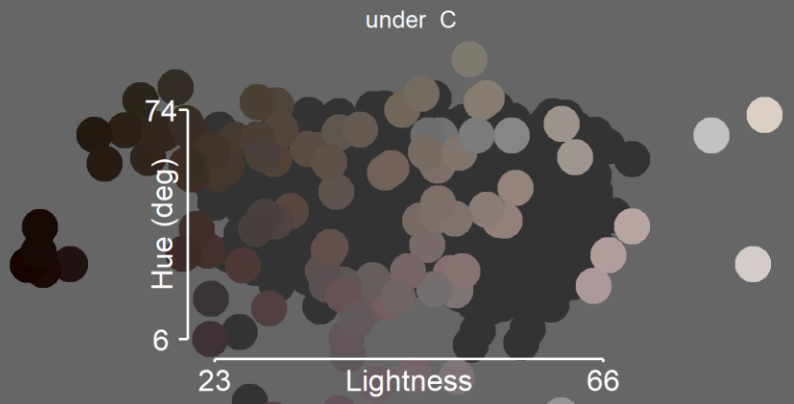
Howard et al., TBIOM 2021

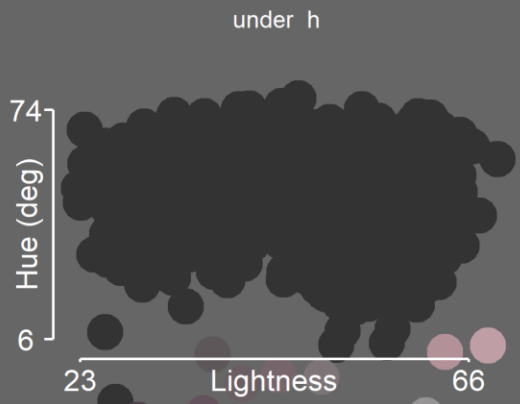


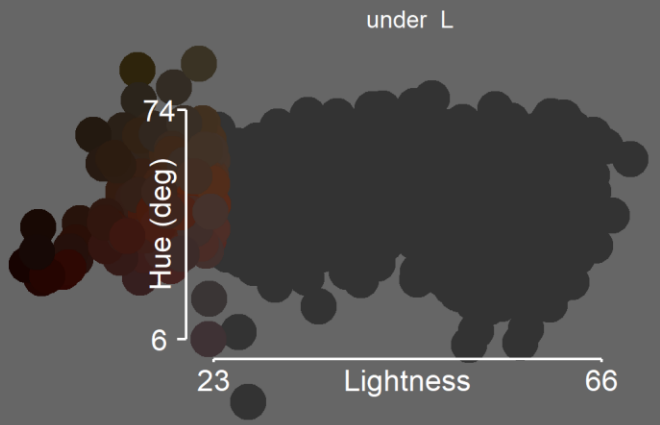




$$CH_{MEDSII} \sim 60\%$$







Takeaways

- Face recognition system performance varies as a function of skin tone
  - Reduced performance for people with darker skin
  - Skin tone is linked with performance independent of self-reported race

- Skin tone is not reliably represented in images returned by commercial biometric systems
  - Error in color reproduction
  - Over and under-exposure

- Color calibration targets used to calibrate digital cameras do not provide even and complete coverage of measured skin tone values

- Face recognition systems should be engineered to take quality images for individuals within the full range of measured skin tone values
  - Color targets better representing skin tone variation may help

Questions & Answers

- Contact information
  - 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