Accuracy Comparison Across Face Recognition Algorithms: Where Are We On Measuring Race Bias?

Jacqueline G. Cavazos[1], P. Jonathon Phillips[2], Carlos D. Castillo[3], Alice J. O'Toole[1]
The University of Texas at Dallas (UTD)[1]
National Institute of Standards and Technology[2]
*Johns Hopkins University[3]

International Face Performance Conference (IFPC) - 2020

*correction: original presentation stated: The University of Maryland

OVERVIEW

- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks (Cavazos, Phillips, Castillo, O'Toole, 2020)
- Final thoughts/considerations on race accuracy variation

MYTHS ABOUT RACE PERFORMANCE VARIATION

- Myth #1: There would be no race performance variation in face identification if we eliminated machines.

- Myth #2: Face recognition systems used to be "fair" before 2015 and the emergence of deep convolutional neural networks (DCNNs).

- Myth #3: Race is categorical. And we know what these categories are.

OVERVIEW

- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks (Cavazos, Phillips, Castillo, O'Toole, 2020)
- Final thoughts/considerations on race accuracy variation

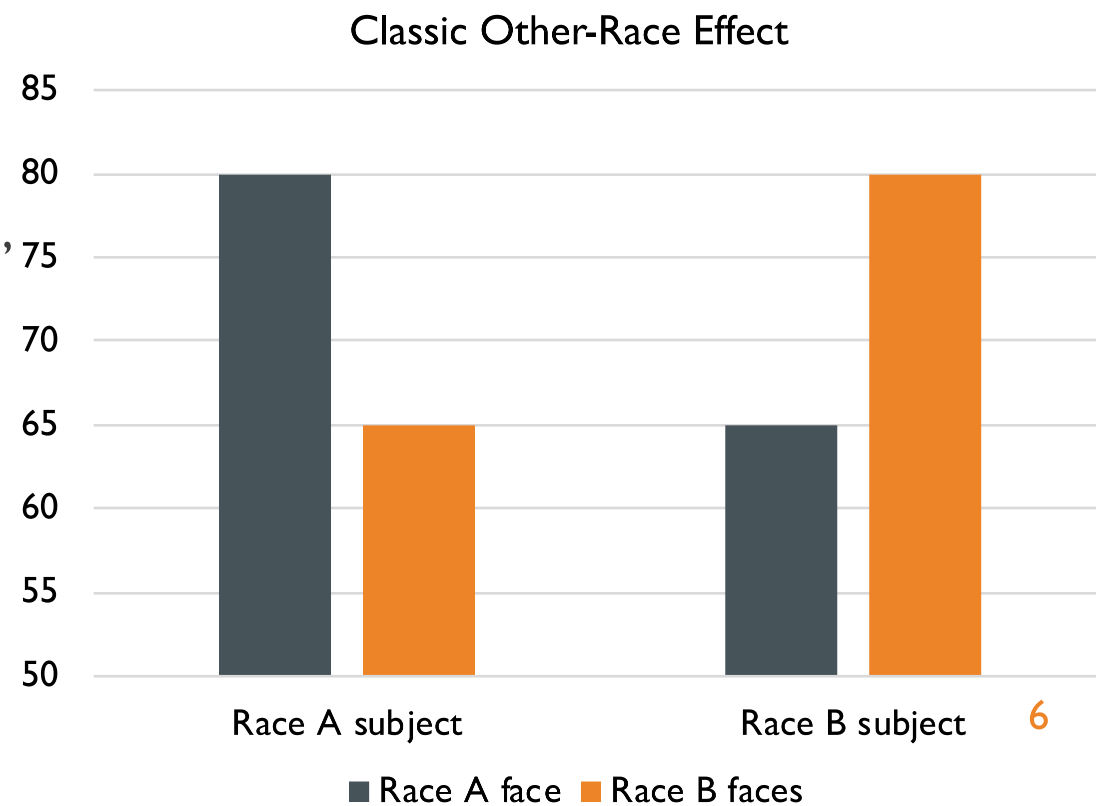
THE OTHER-RACE EFFECT FOR HUMANS

- Greater identification accuracy for own-race faces compared to other-race faces. (Malpass & Kravtiz, 1969; Meissner & Brigham, 2001)

- Multiple racial/ethnic groups (Meissner & Brigham, 2001)
- Methodological paradigms (Meissner & Brigham, 2001; Sporer et al., 2001)
- Age groups (Sangrigoli and De Schonen, 2004; Kelly et al., 2005; Pezdek et al., 2003; Anzures et al., 2014; Tham et al., 2017)

OTHER - RACE EFFECT VS RACE PERFORMANCE VARIATION

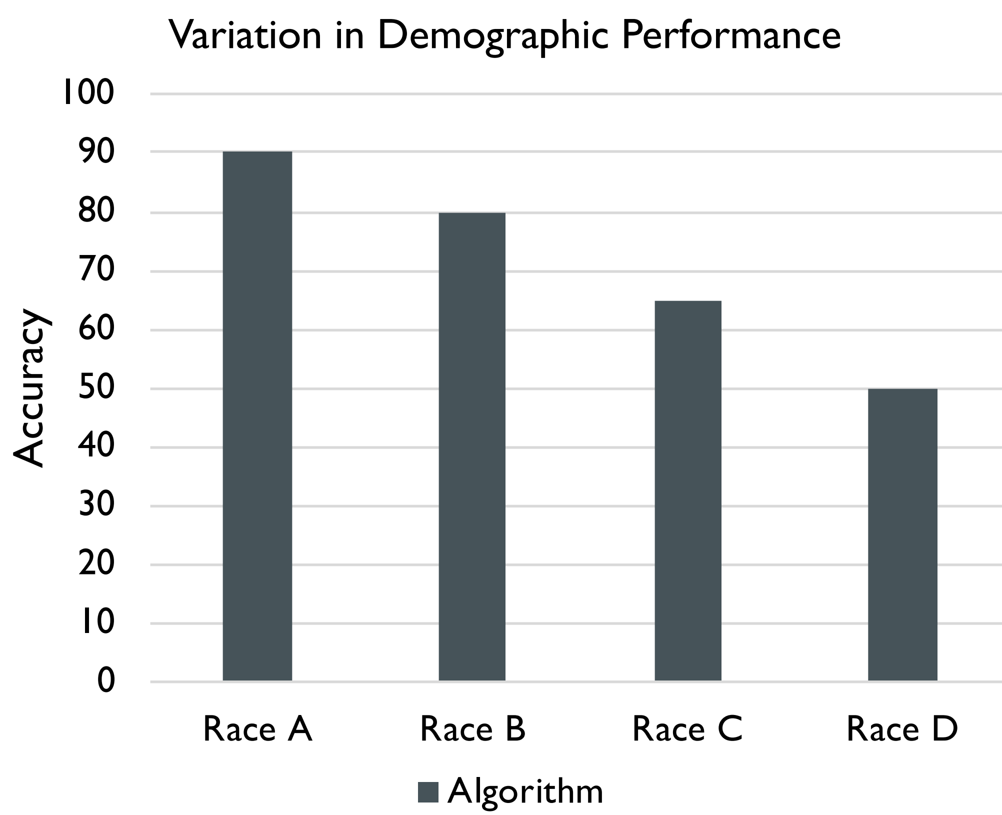
- <u>Other-race effect for humans</u>
  - interaction between the race of "subject" and the race of the "face"



OTHER - RACE EFFECT VS RACE PERFORMANCE VARIATION

- Other-race effect for humans
  - interaction between the race of "subject" and the race of the "face"

- Race performance variation
  - machine more accurate for race A vs. race B



EVIDENCE OF RACE DEMOGRAPHIC VARIATION

Pre-DCNNs

- Asian and Caucasian (Furl et al., 2002)
- East Asian and Caucasian- "Other-race effect"(Phillips et al., 2011)
- Black, White, Hispanic (multiple demographics: gender, race, age)(Klare et al., 2012)

DCNNs

- Black and White (multi-class demographics) (El Khiyari et al., 2016)
- African American and Caucasian (Krishnapriya et al., 2019; 2020)
- NIST report on demographic effects (Grother et al., 2019)

OVERVIEW

- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks (Cavazos, Phillips, Castillo, O'Toole, 2020)
- Final thoughts/considerations on race accuracy variation

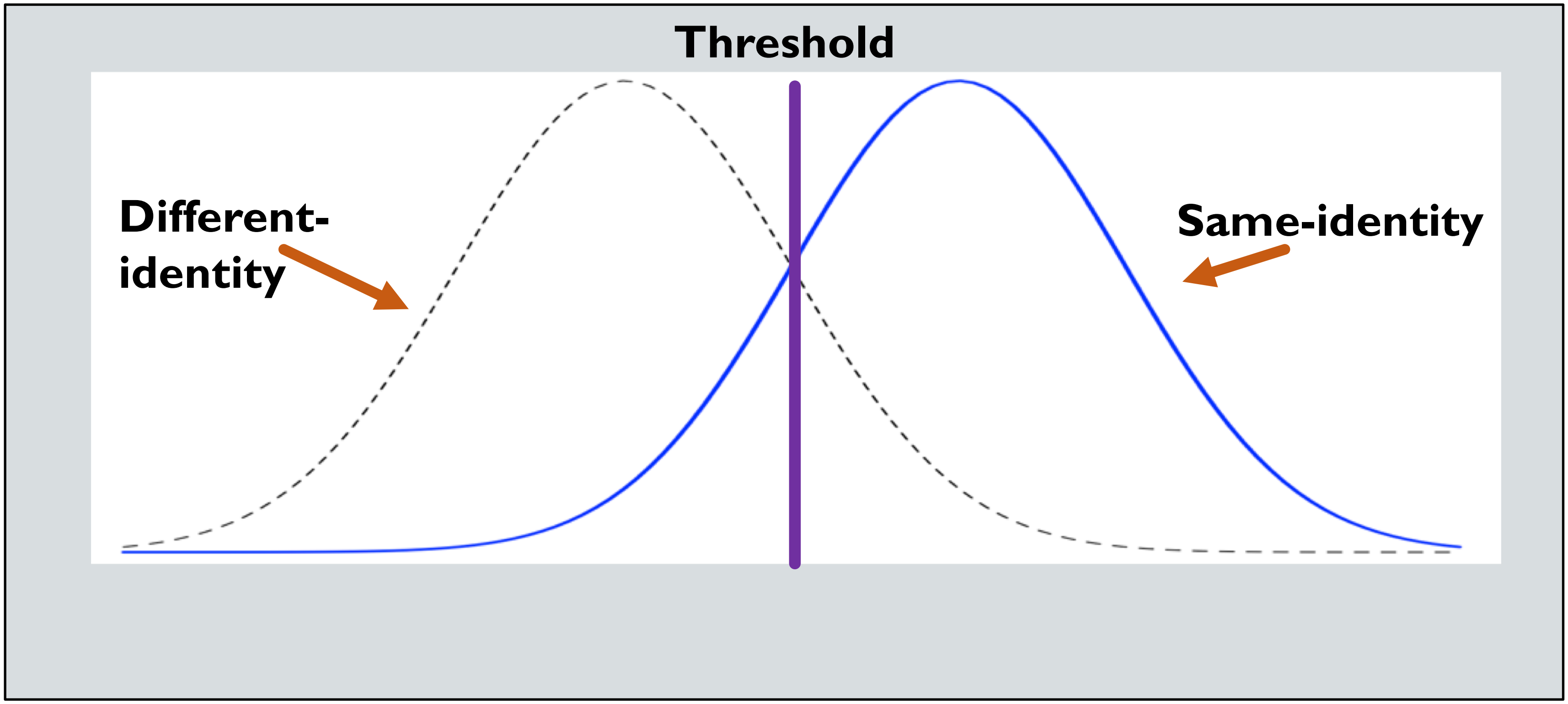
HUMAN TASK

Are these images of the same person or two different people?

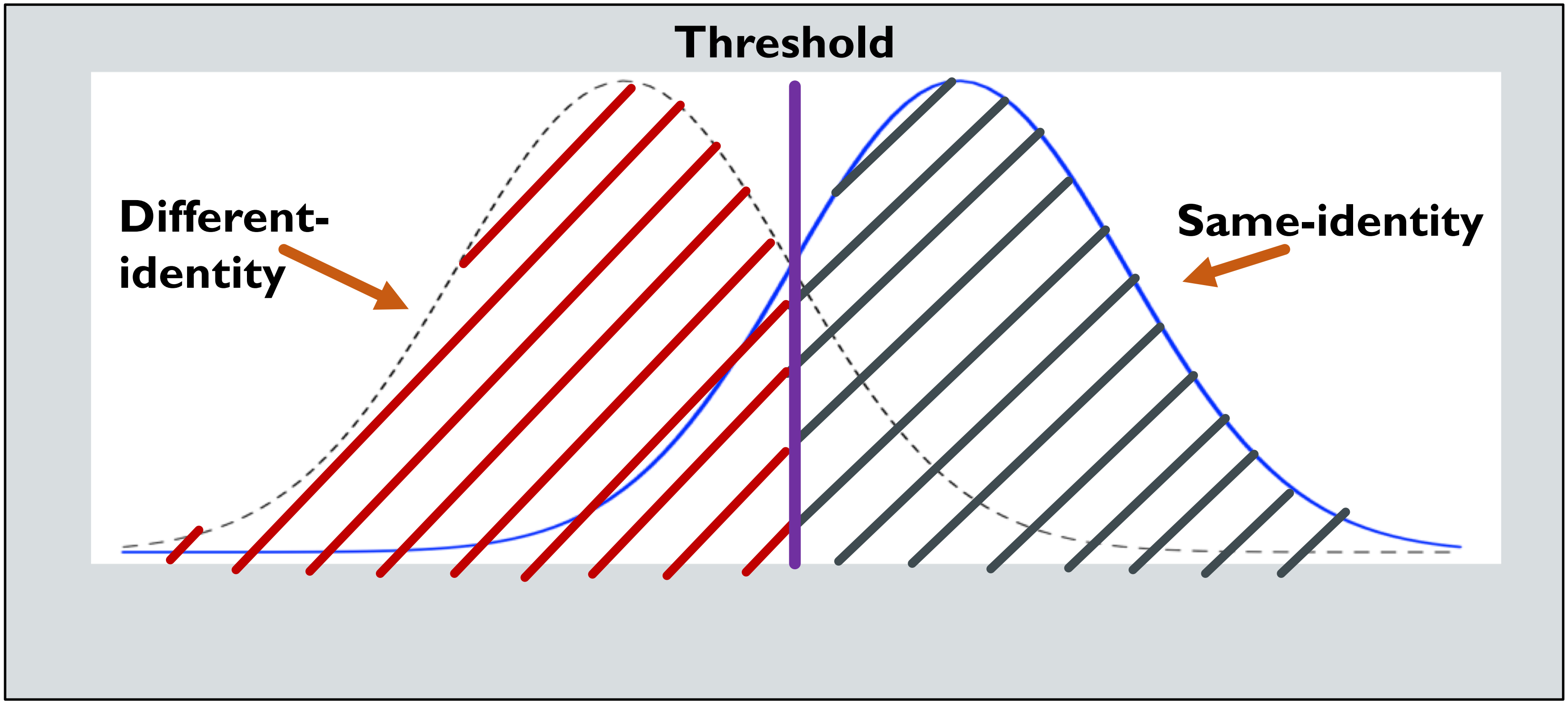
Response Options

1: Sure they are the same
2: Think they are the same
3: Do not know
4: Think they are different
5: Sure they are different

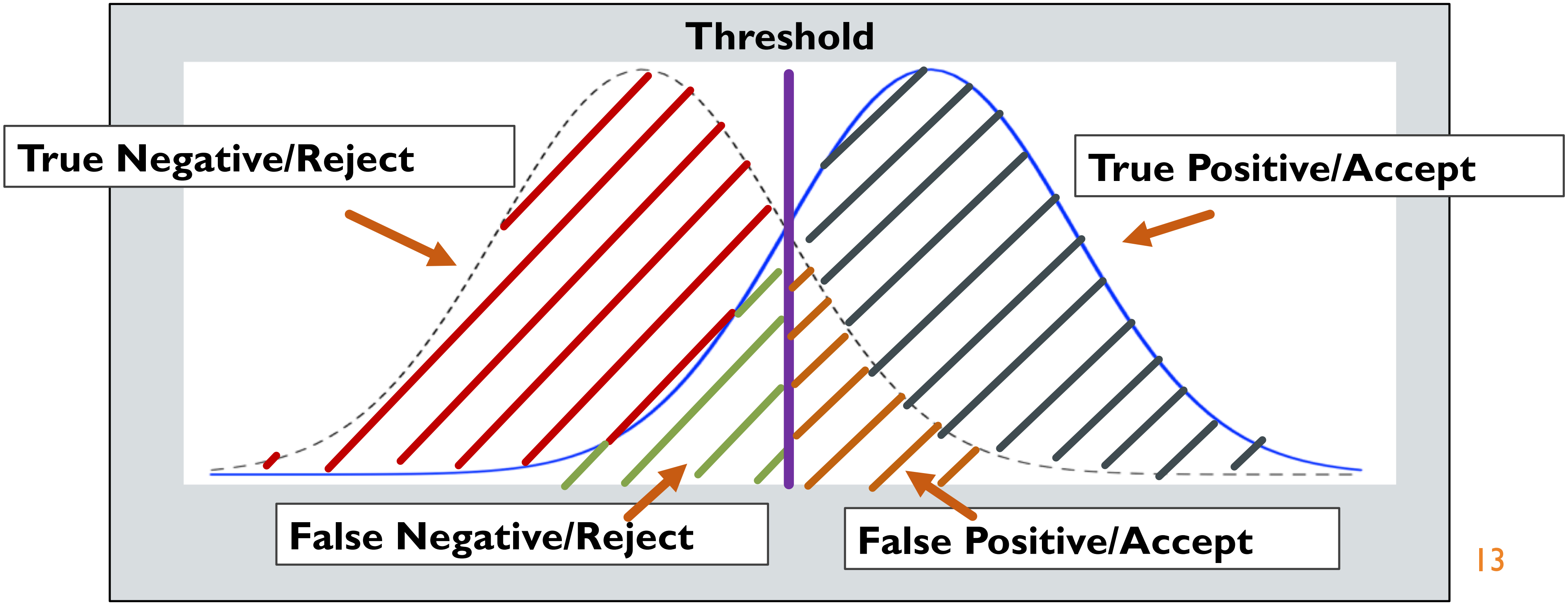
MEASURING HUMAN PERFORMANCE



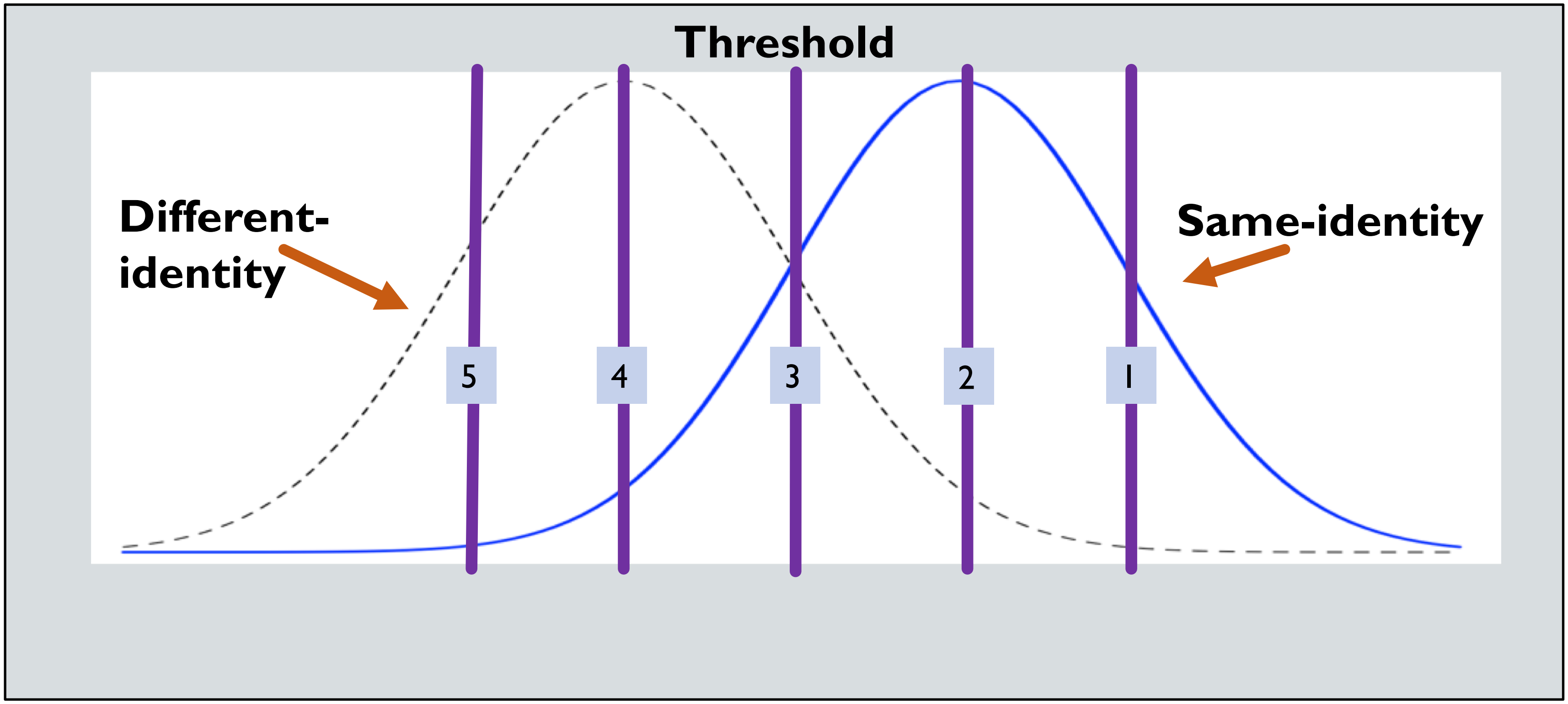
MEASURING HUMAN PERFORMANCE



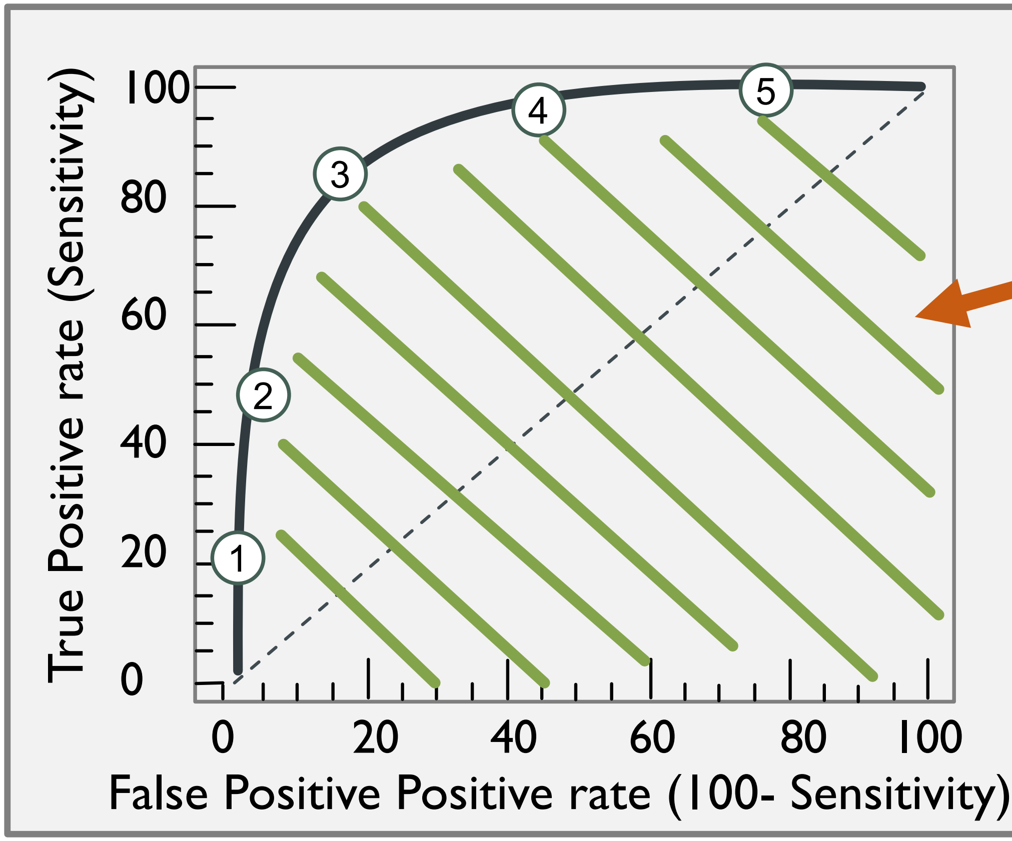
MEASURING HUMAN PERFORMANCE



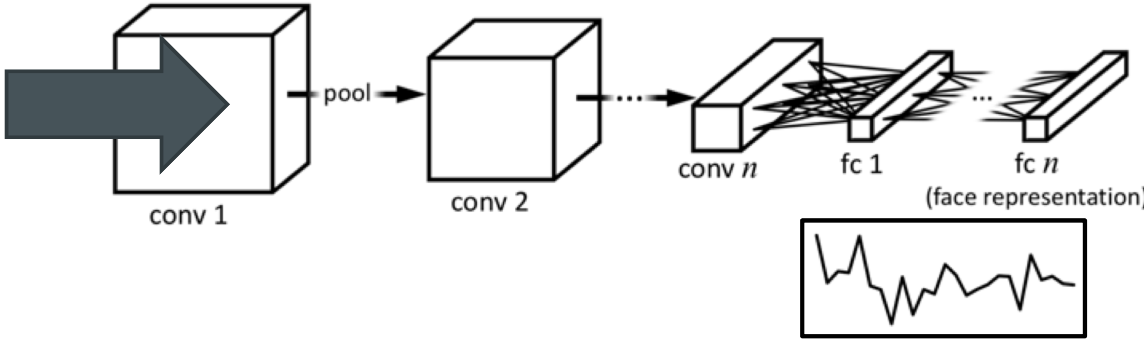
MEASURING HUMAN PERFORMANCE



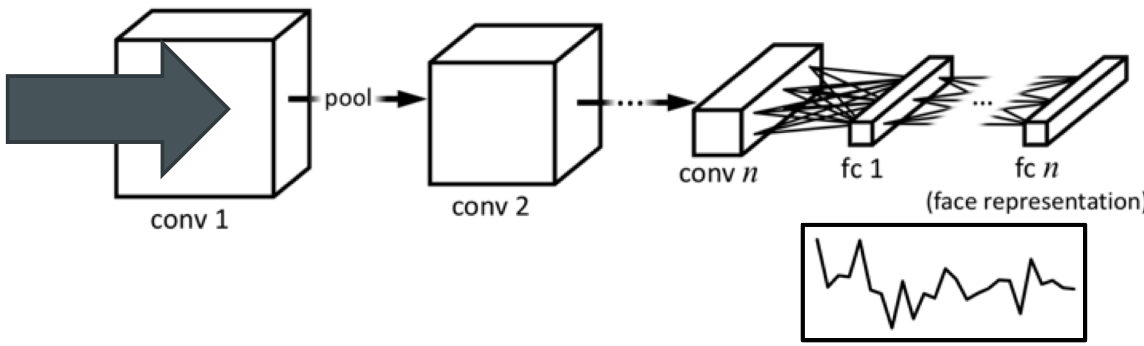
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



AUC: Area under the Curve

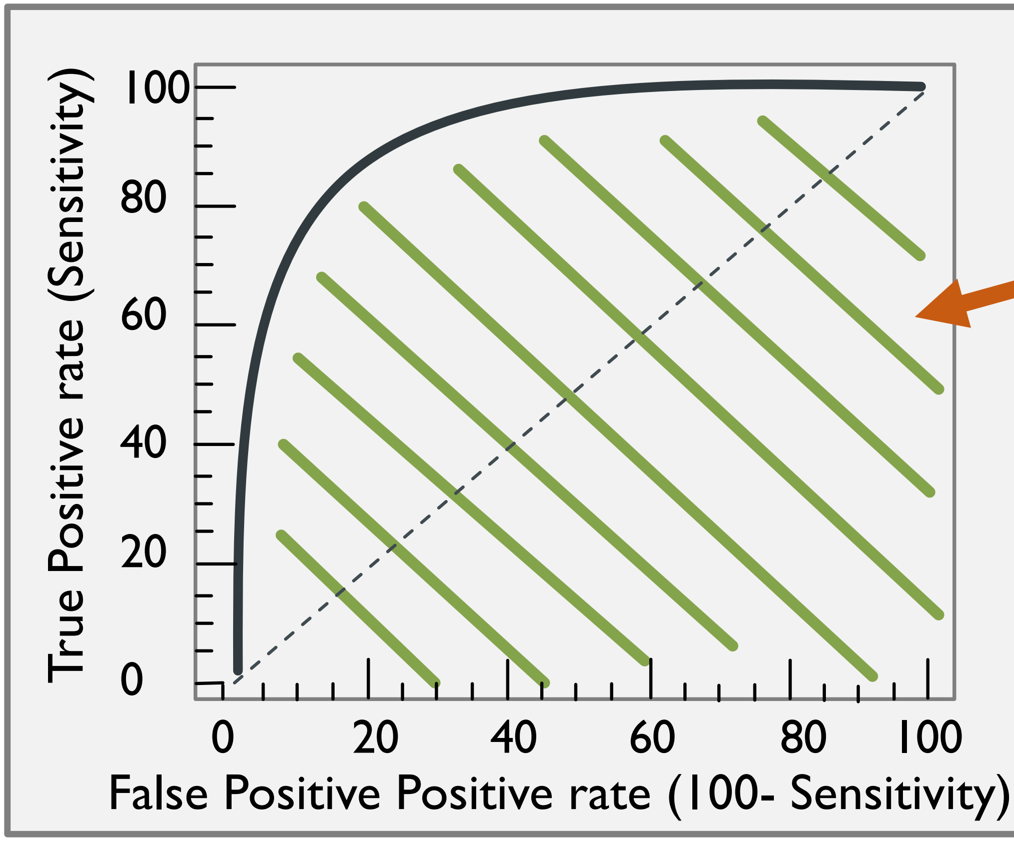


pool
conv 1
conv 2
conv n
fc 1
fc n
(face representation)



pool
conv 1
conv 2
conv n
fc 1
fc n
(face representation)

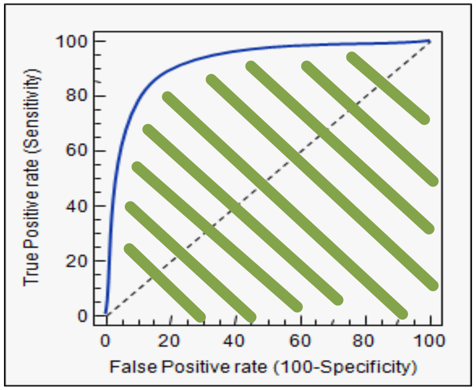
RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



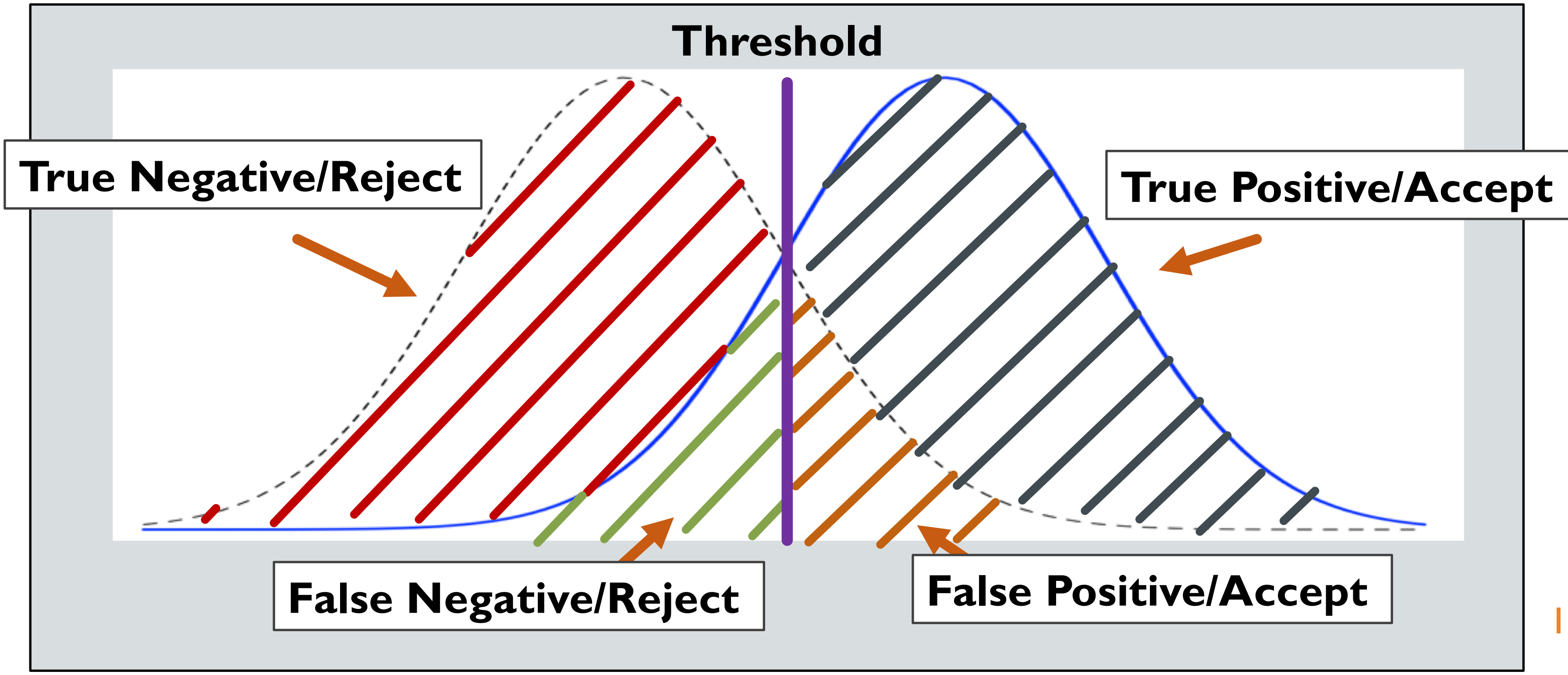
AUC: Area under the Curve

MEASURING ALGORITHM PERFORMANCE

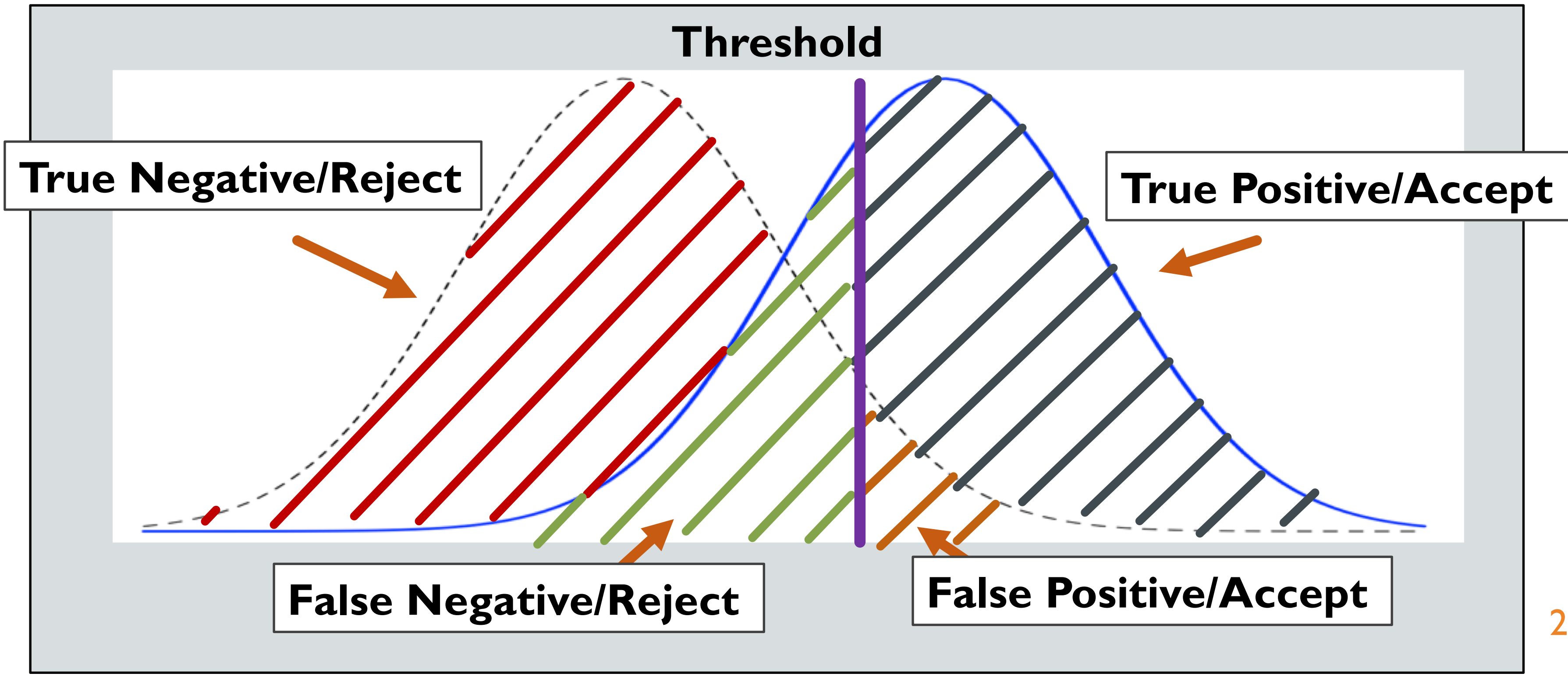
- Performance measures
  - **threshold independent**: characterize "system" as a whole
    - Area under the ROC curve (AUC, aROC)
  - **threshold dependent**: operational measure



THRESHOLD DEPENDENT MEASURE

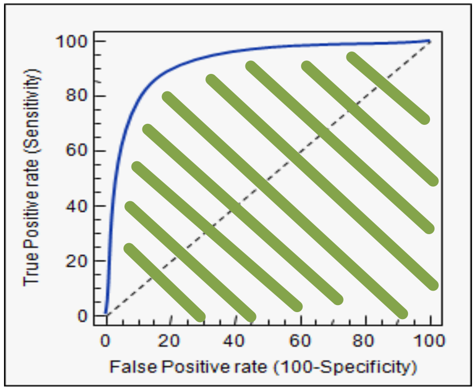


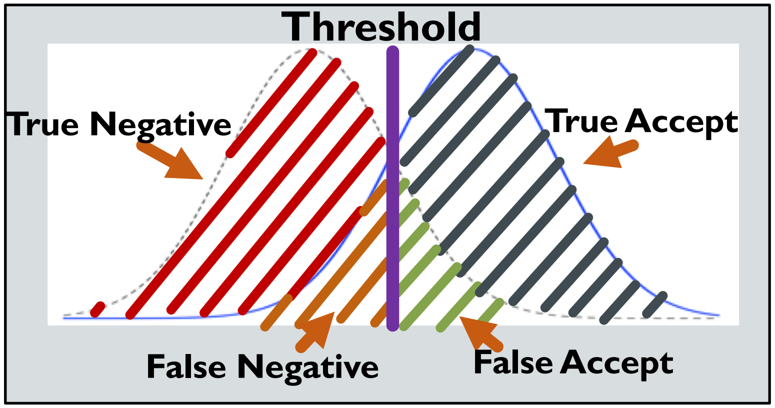
THRESHOLD DEPENDENT MEASURE



MEASURING ALGORITHM PERFORMANCE

- Performance measures
  - **threshold independent**: characterize "system" as a whole
    - Area under the ROC curve (AUC, aROC)
  - **threshold dependent**: operational measure
    - measure true accept rate (TAR) @ a pre-set FAR
      - FAR usually very low FAR: 10^{-3}, 10^{-4}, 10^{-5}





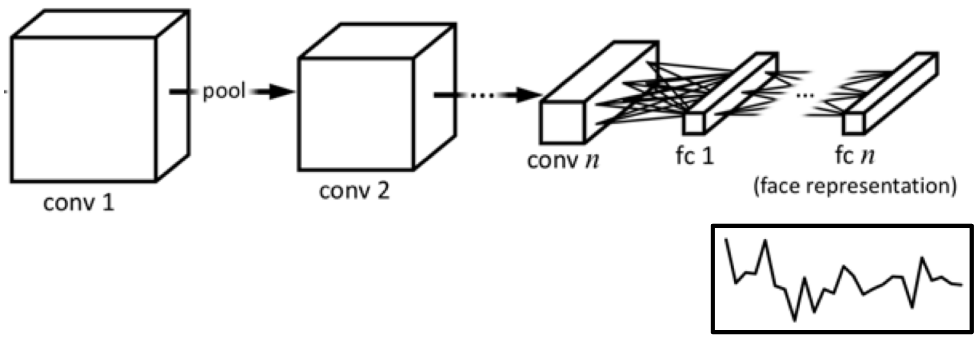
OVERVIEW

- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks (Cavazos, Phillips, Castillo, O'Toole, 2020)
- Final thoughts/considerations on race accuracy variation

DATA-DRIVEN FACTORS

same- identity and different-identity distributions differ across demographics

Quality of algorithms' representation



Faces representativeness

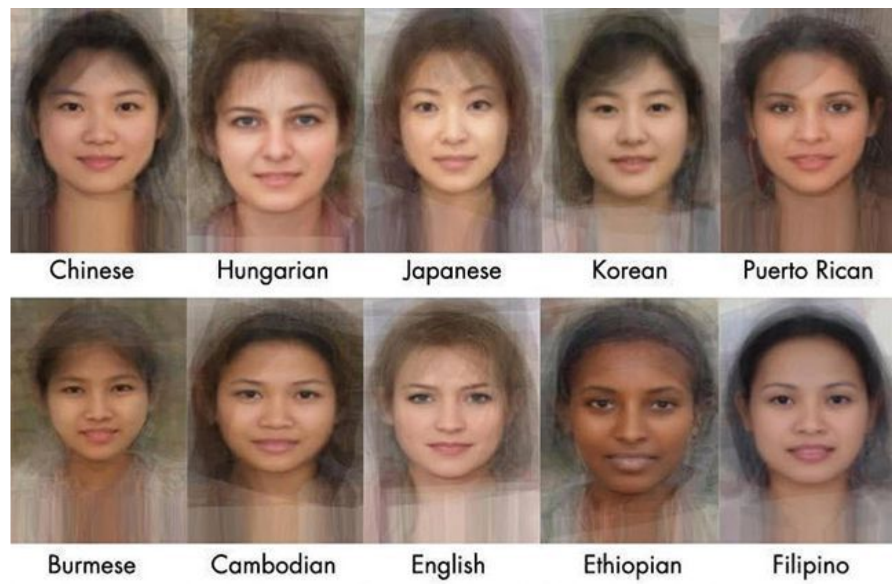


Photo credit : IJB-B/IJB-C datasets

DATA-DRIVEN FACTORS

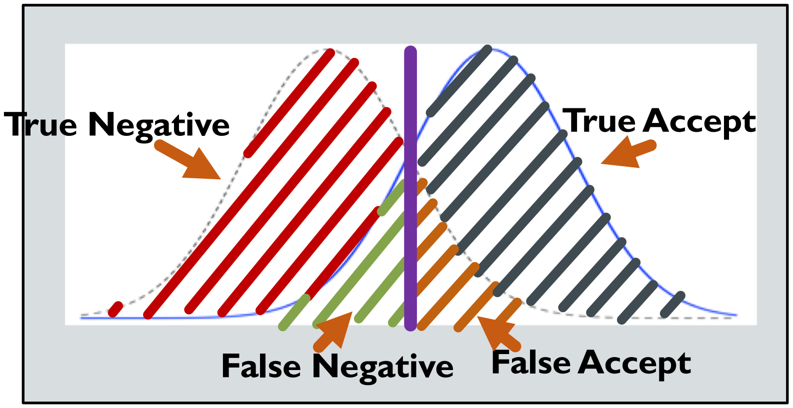
same- identity and different-identity distributions differ across demographics

Quality of photographs

Photo credit : IJB-B/IJB-C datasets

OPERATIONAL FACTORS

Thresholds



(O'Toole et al., 2012; Krishnapriya et al., 2019; NIST; Bowyer, 2019; Cavazos et al., 2020)

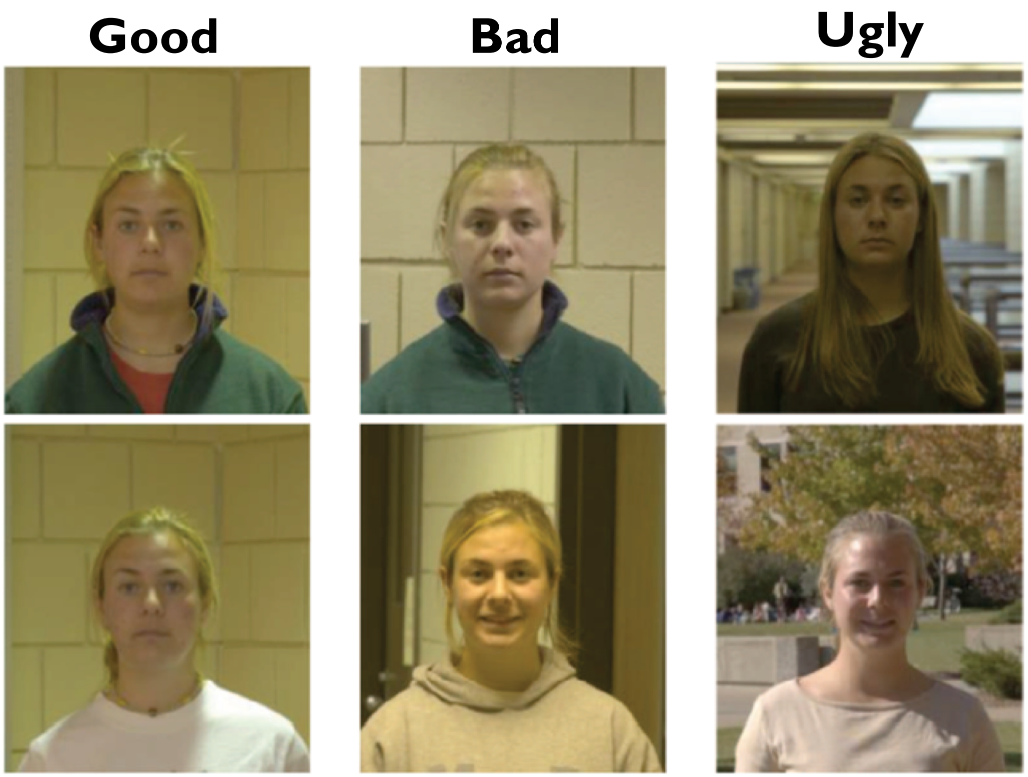
"Yoking"- different-identity distribution

(O'Toole et al., 2012; Cavazos et al., 2020)

OVERVIEW

- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks (Cavazos, Phillips, Castillo, O'Toole, 2020)
- Final thoughts/considerations on race accuracy variation

DATA SET



- NIST Good, Bad, Ugly Challenge (Phillips et al., 2011)
  - stimulus difficulty levels stratified with previous generation algorithm
- East Asian and Caucasian faces

ALGORITHMS

A2011

- Pre-DCNN
- Fused algorithm of top three algorithms in FRVT 2006

(Phillips et al., 2011)

A2015

- Early DCNN
- Training: VGG Face, 982,803 images

(Parkhi, Vedaldi, & Zisserman, 2015)

A2019

- Recent DCNN
- Training: 993,153 images

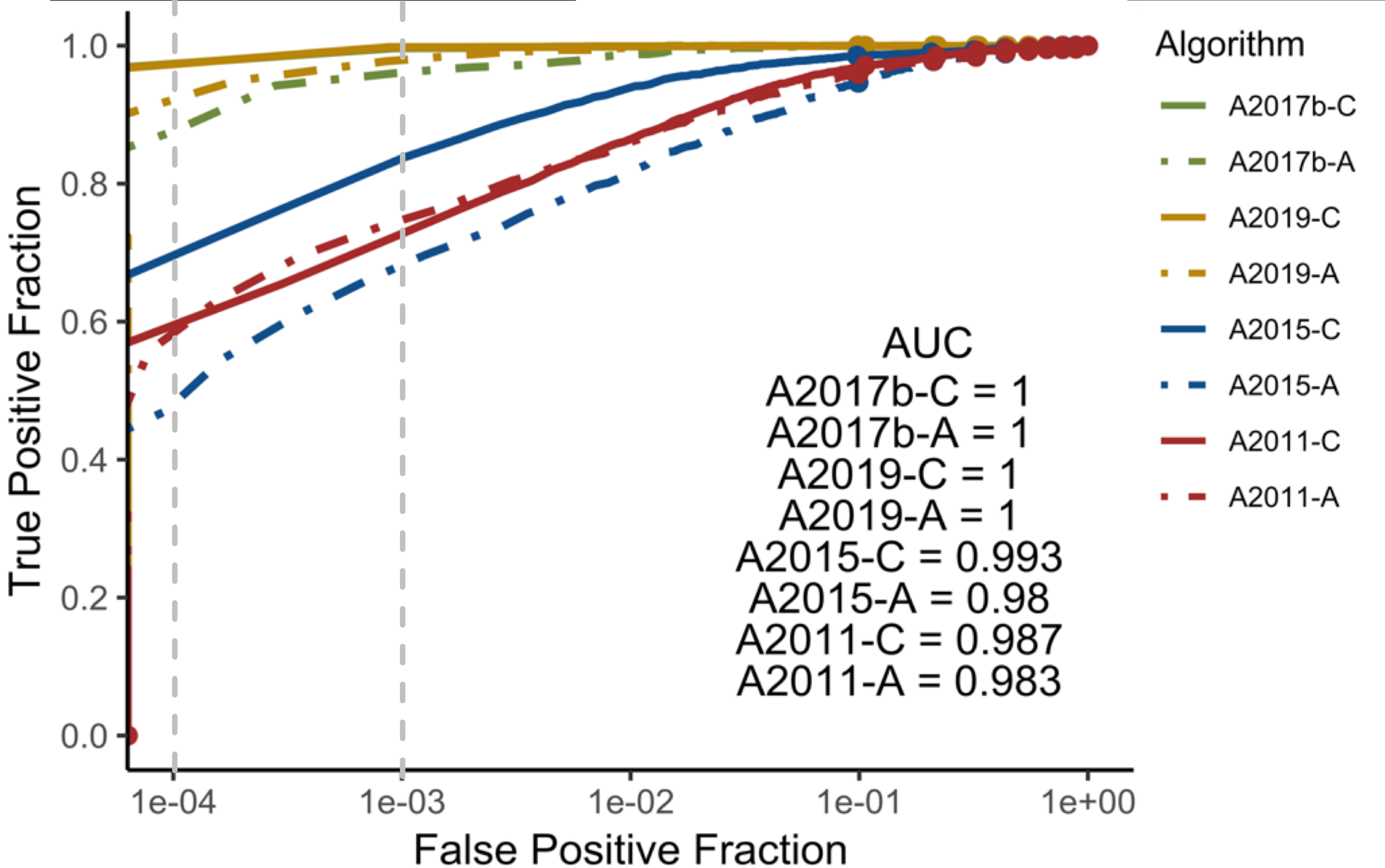
(Ranjan et al., 2019)

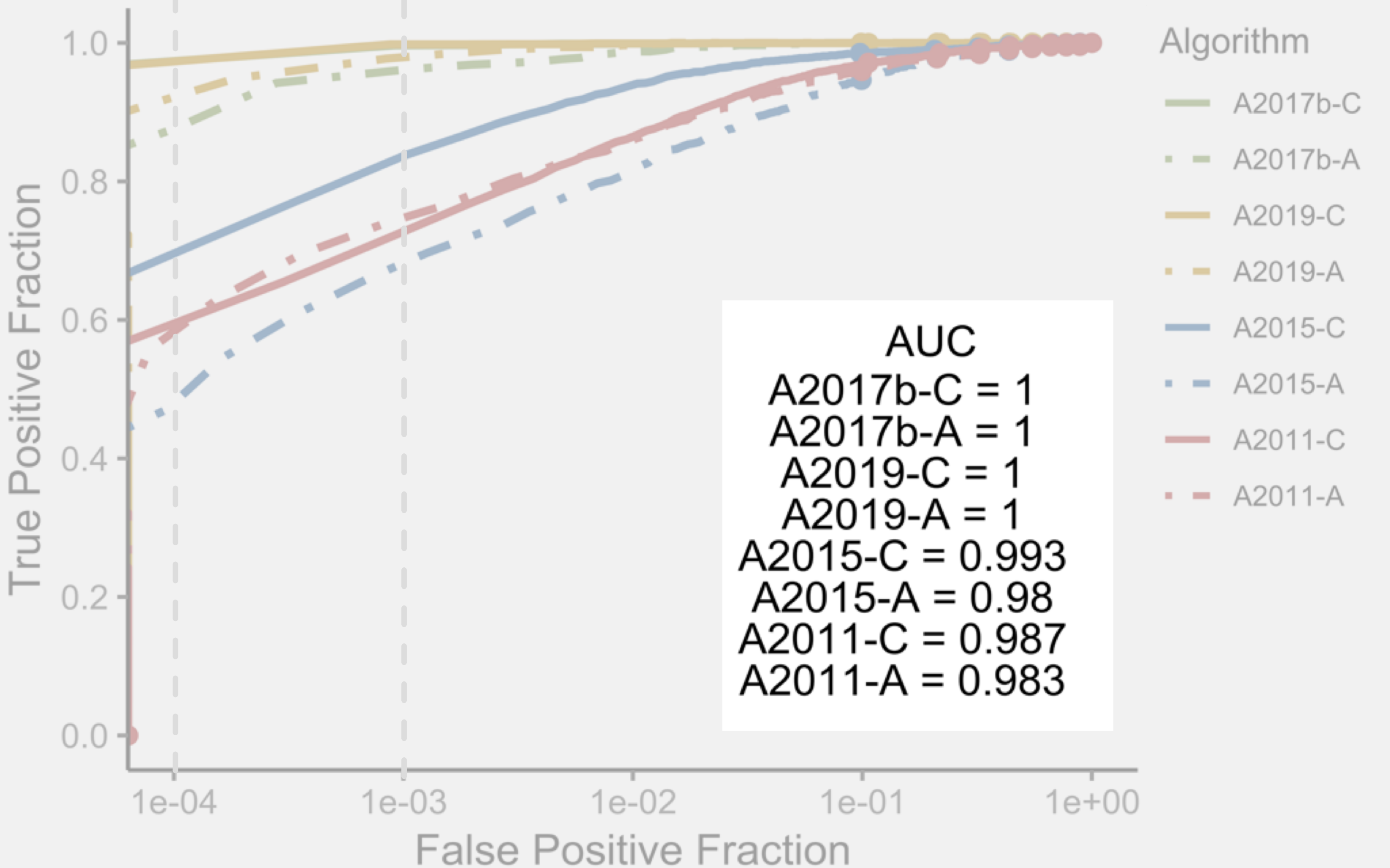
A2017b

- Recent DCNN
- Training: 5,714,444 images
- Accuracy = forensic face examiners (Phillips et al., 2018)

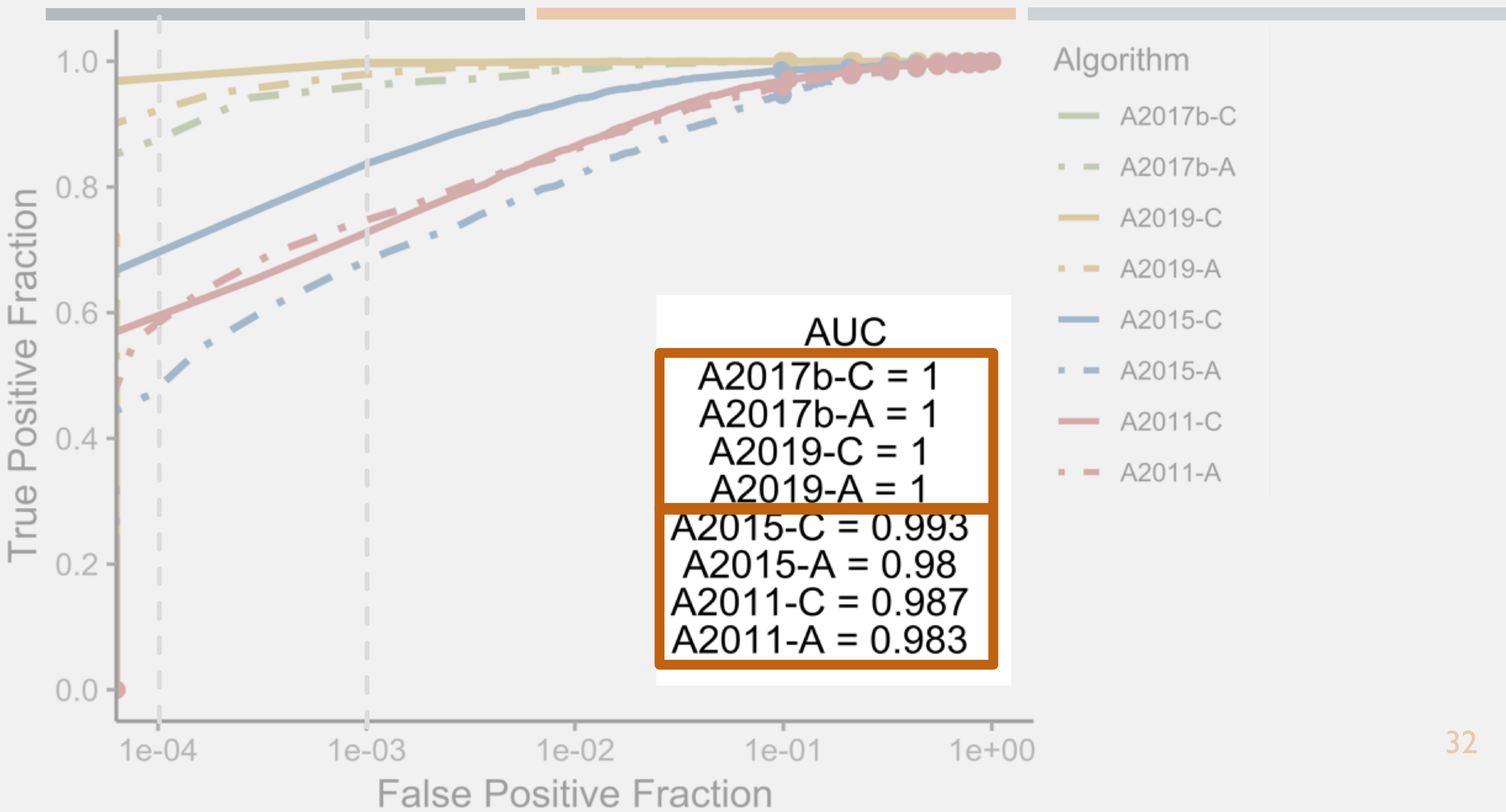
(Ranjan, Castillo & Chellappa, 2017)

RACE RESULTS

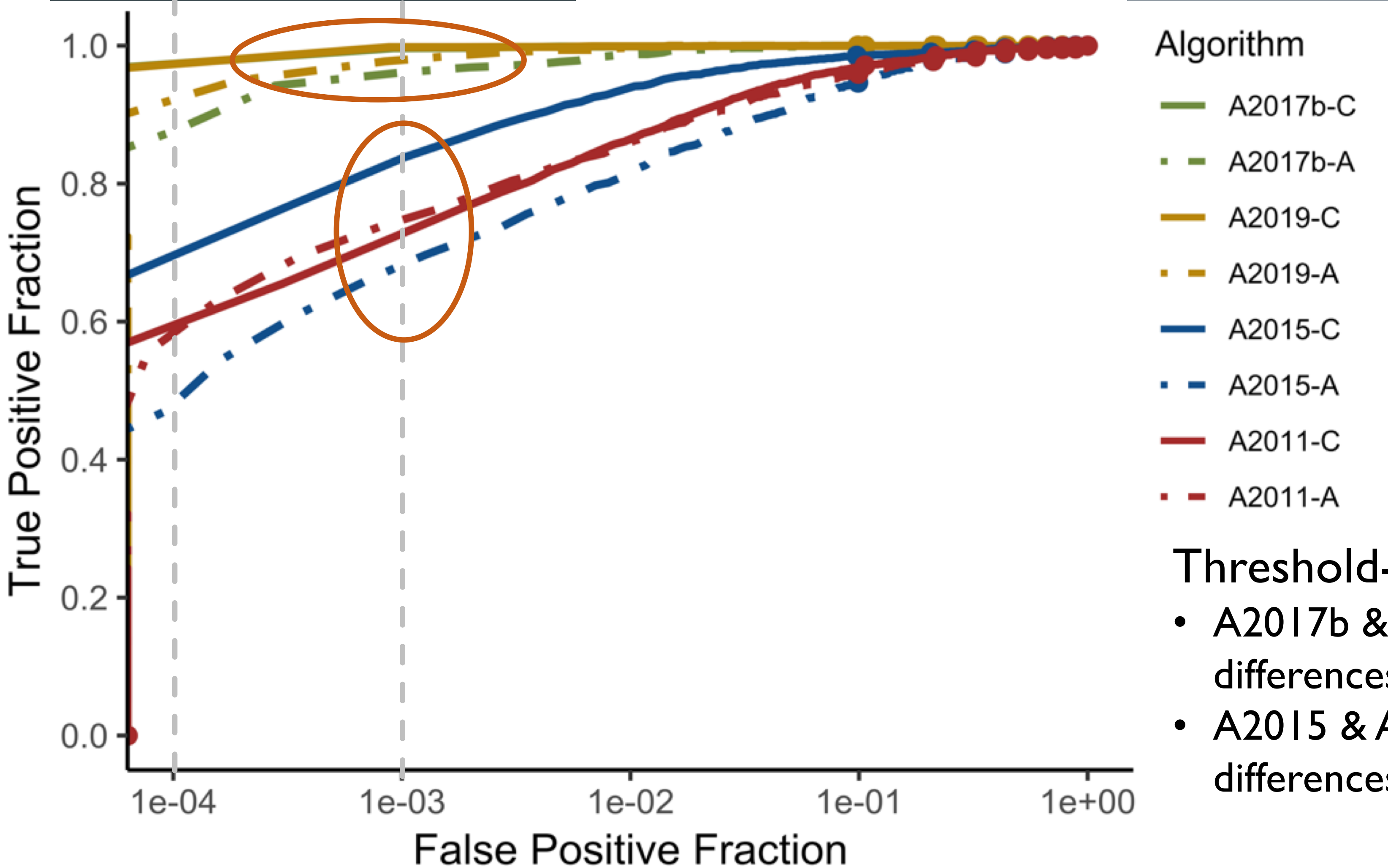

True Positive Fraction
1.0
0.8
0.6
0.4
0.2
0.0
1e-04
1e-03
1e-02
1e-01
1e+00
False Positive Fraction
AUC
A2017b-C = 1
A2017b-A = 1
A2019-C = 1
A2019-A = 1
A2015-C = 0.993
A2015-A = 0.98
A2011-C = 0.987
A2011-A = 0.983
Algorithm
A2017b-C
A2017b-A
A2019-C
A2019-A
A2015-C
A2015-A
A2011-C
A2011-A



True Positive Fraction
1.0
0.8
0.6
0.4
0.2
0.0
1e-04
1e-03
1e-02
1e-01
1e+00
False Positive Fraction
AUC
A2017b-C = 1
A2017b-A = 1
A2019-C = 1
A2019-A = 1
A2015-C = 0.993
A2015-A = 0.98
A2011-C = 0.987
A2011-A = 0.983
Algorithm
A2017b-C
A2017b-A
A2019-C
A2019-A
A2015-C
A2015-A
A2011-C
A2011-A

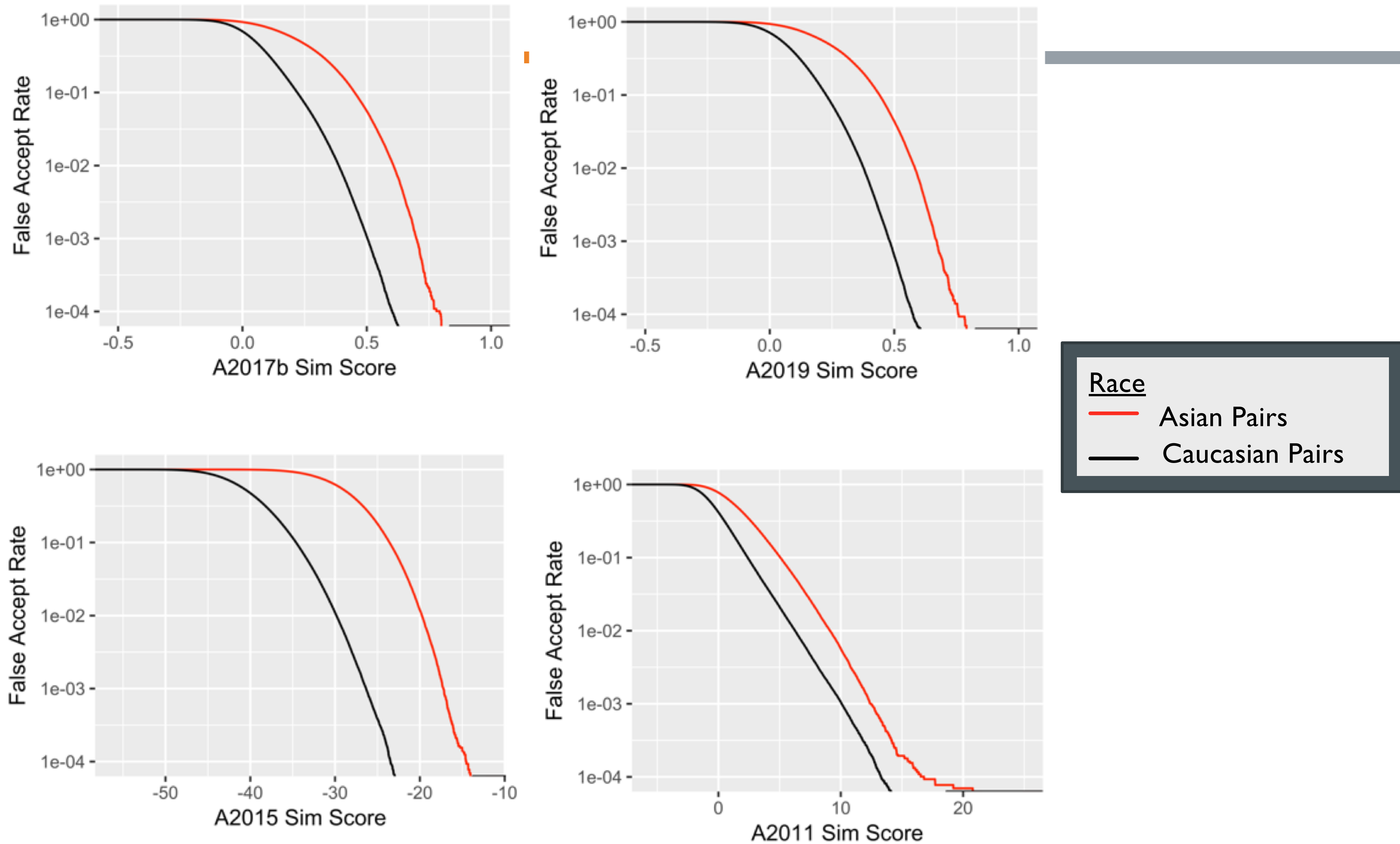


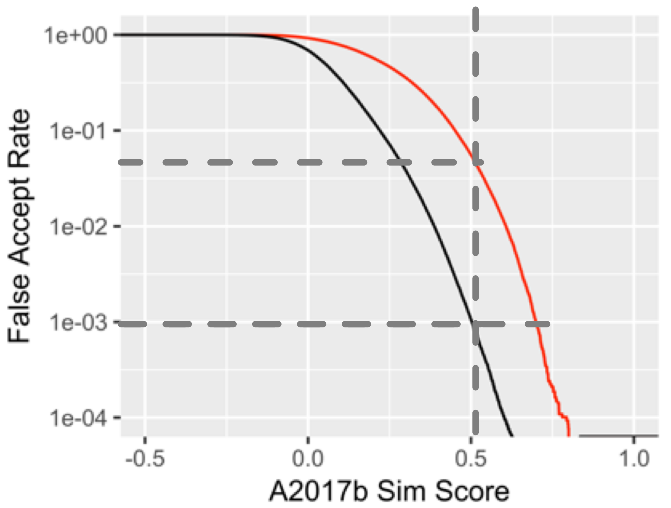
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Fraction
1e-04
1e-03
1e-02
1e-01
1e+00
False Positive Fraction
AUC
A2017b-C = 1
A2017b-A = 1
A2019-C = 1
A2019-A = 1
A2015-C = 0.993
A2015-A = 0.98
A2011-C = 0.987
A2011-A = 0.983
Algorithm
A2017b-C
A2017b-A
A2019-C
A2019-A
A2015-C
A2015-A
A2011-C
A2011-A

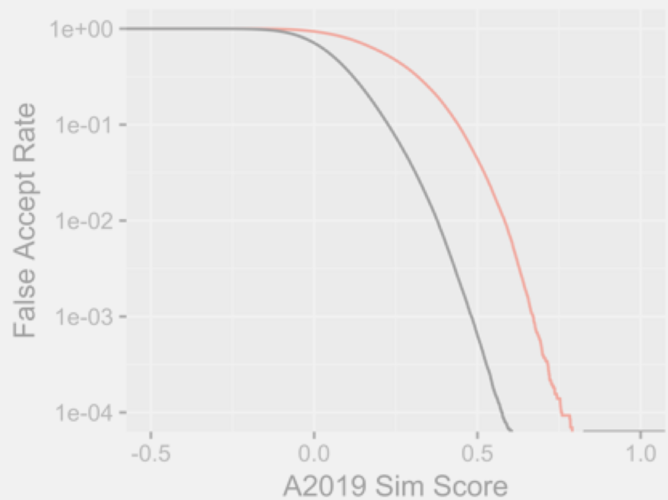


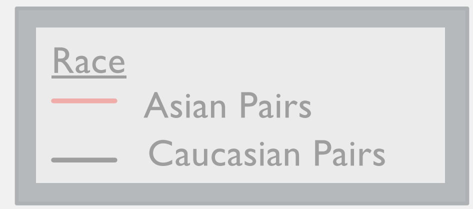
Threshold-independent:

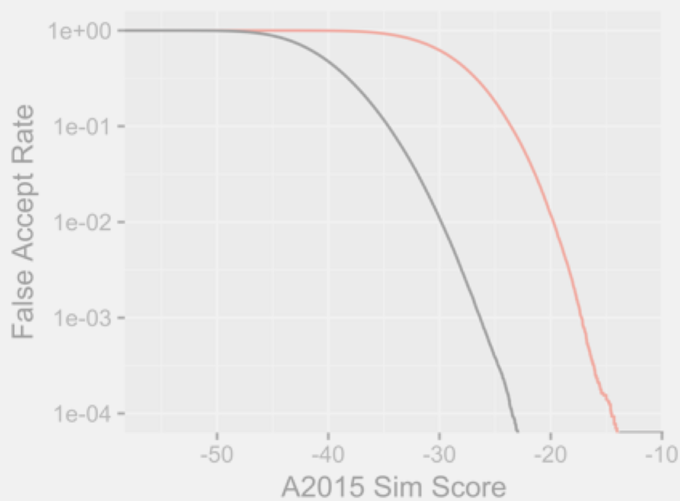
- A2017b & A2019 = **no** differences across race
- A2015 & A2011 = **minimal** differences across race

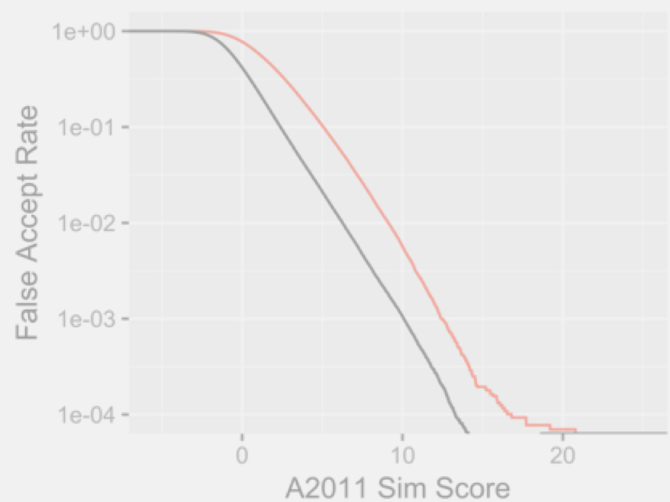

False Accept Rate
1e+00
1e-01
1e-02
1e-03
1e-04
-0.5
0.0
0.5
1.0
A2017b Sim Score
A2019 Sim Score
-50
-40
-30
-20
-10
A2015 Sim Score
0
10
20
A2011 Sim Score
Race
Asian Pairs
Caucasian Pairs

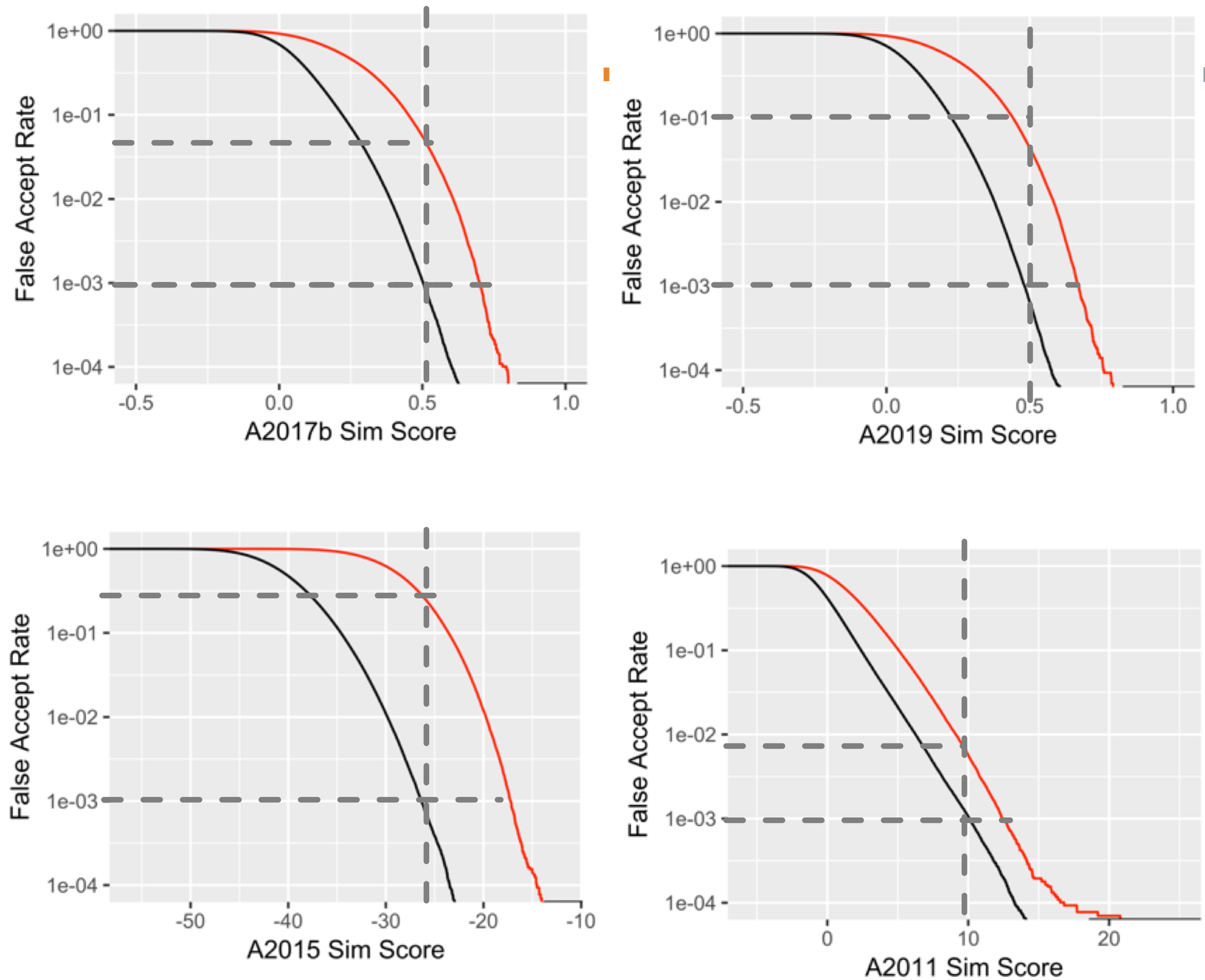

1e+00
1e-01
1e-02
1e-03
1e-04
False Accept Rate
-0.5
0.0
0.5
1.0
A2017b Sim Score


1e+00
1e-01
1e-02
1e-03
1e-04
False Accept Rate
-0.5
0.0
0.5
1.0
A2019 Sim Score


Race
Asian Pairs
Caucasian Pairs


1e+00
1e-01
1e-02
1e-03
1e-04
False Accept Rate
-50
-40
-30
-20
-10
A2015 Sim Score


1e+00
1e-01
1e-02
1e-03
1e-04
False Accept Rate
0
10
20
A2011 Sim Score

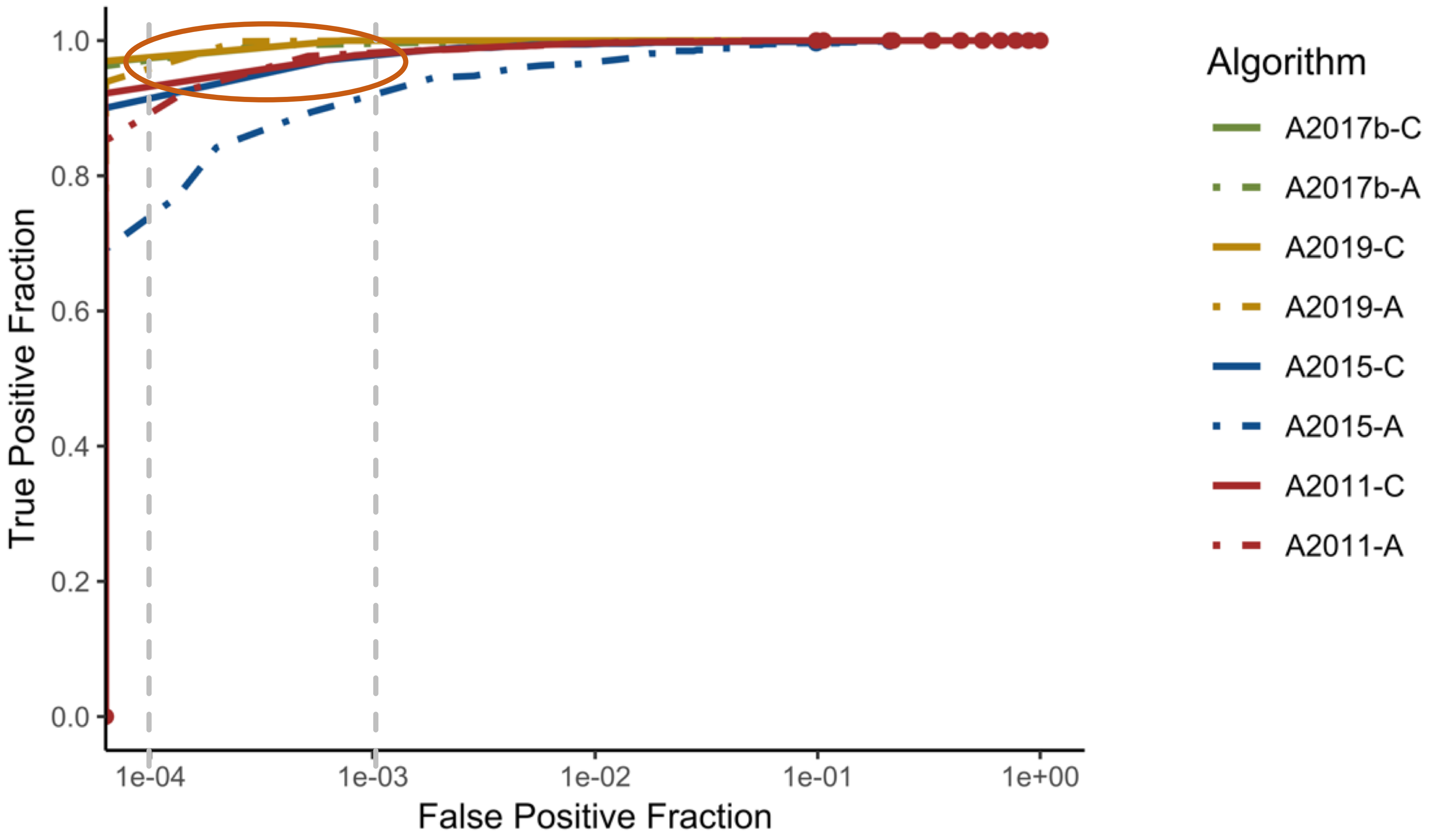


Threshold dependent:

- **All four** algorithms need greater Asian threshold when setting False Accept Rates.

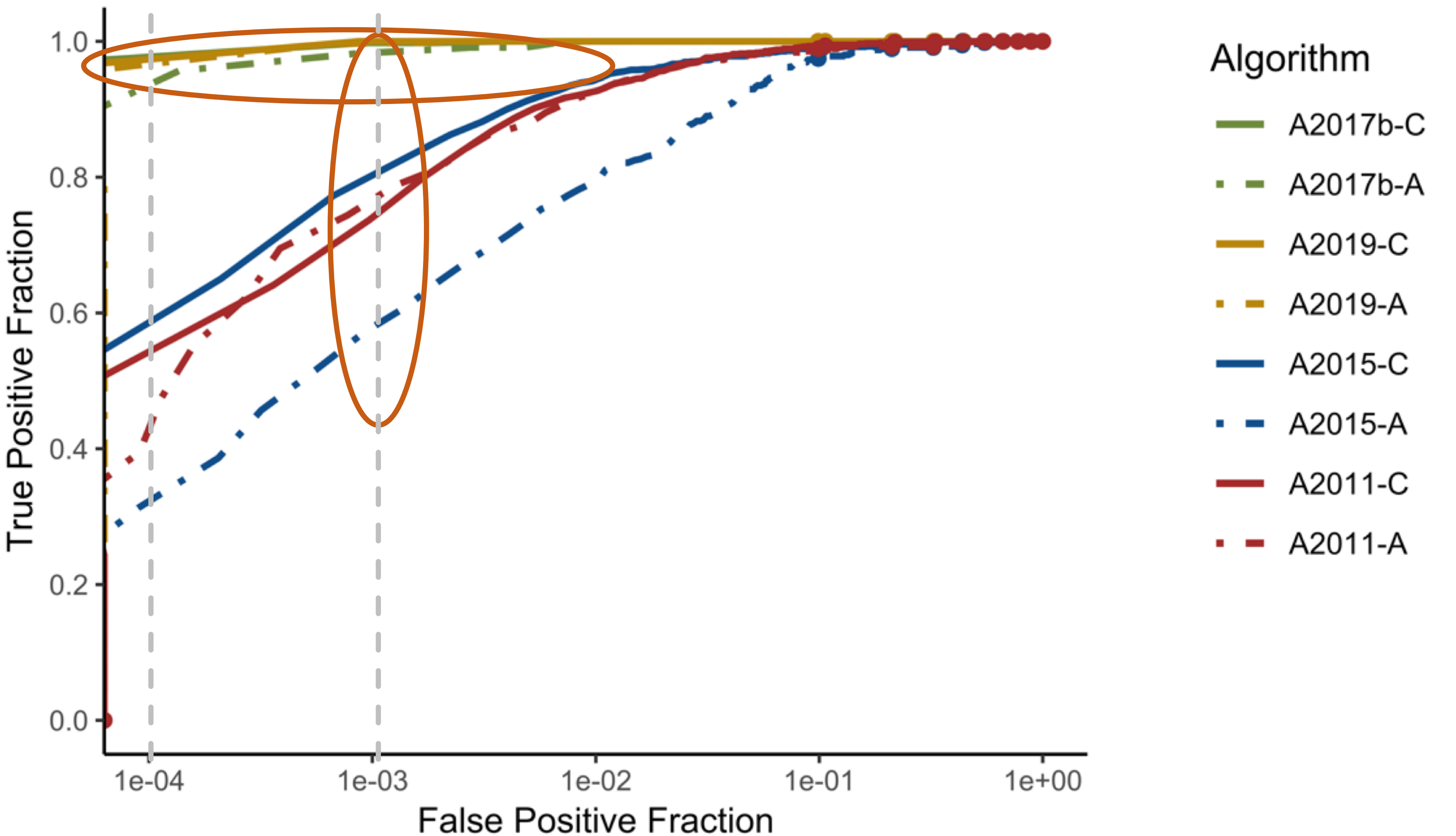
IMAGE DIFFICULTY RESULTS

Good



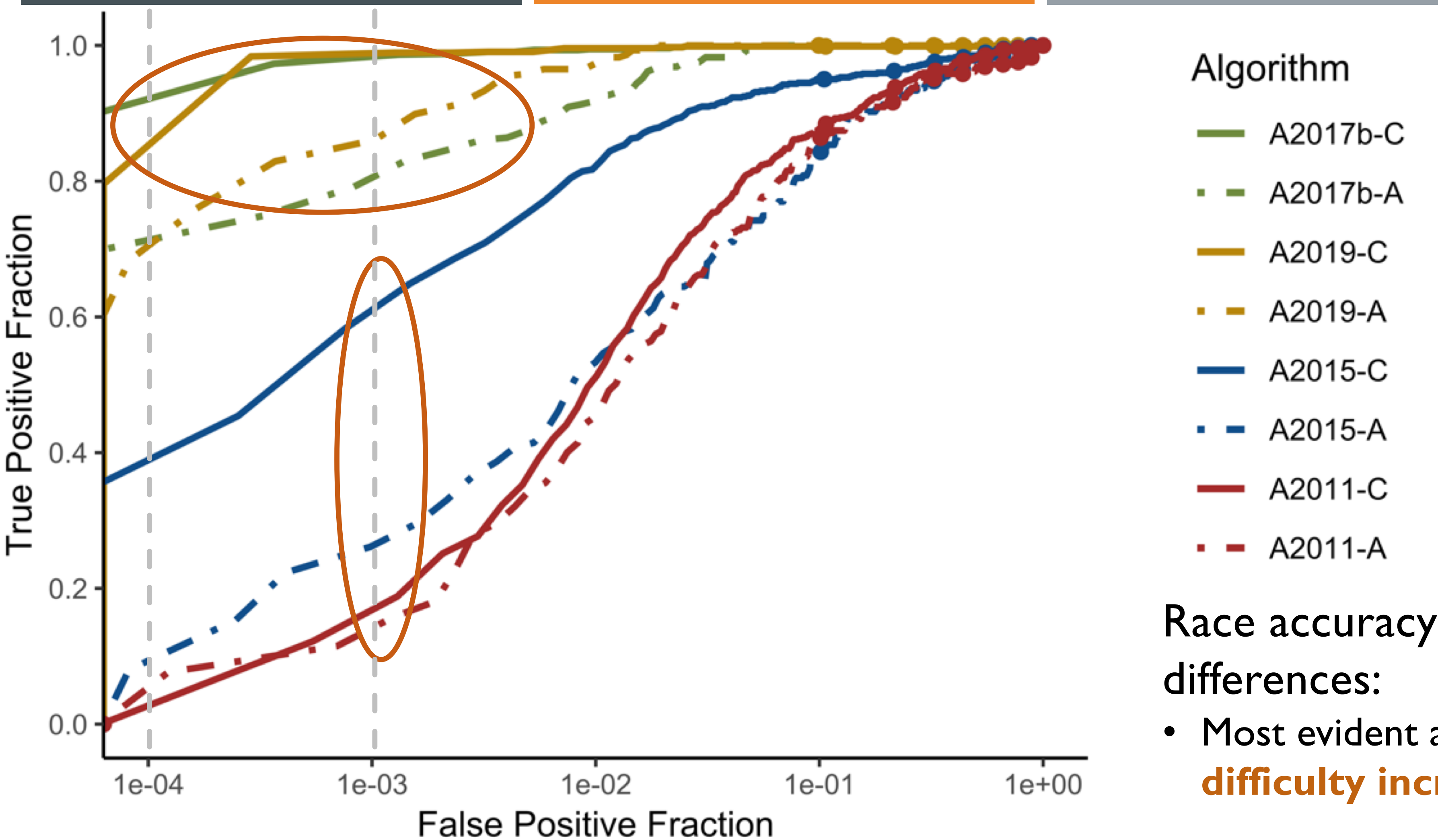
1.0
0.8
0.6
0.4
0.2
0.0
True Positive Fraction
1e-04
1e-03
1e-02
1e-01
1e+00
False Positive Fraction
Algorithm
A2017b-C
A2017b-A
A2019-C
A2019-A
A2015-C
A2015-A
A2011-C
A2011-A

Bad



1.0
0.8
0.6
0.4
0.2
0.0
True Positive Fraction
1e-04
1e-03
1e-02
1e-01
1e+00
False Positive Fraction
Algorithm
A2017b-C
A2017b-A
A2019-C
A2019-A
A2015-C
A2015-A
A2011-C
A2011-A

Ugly



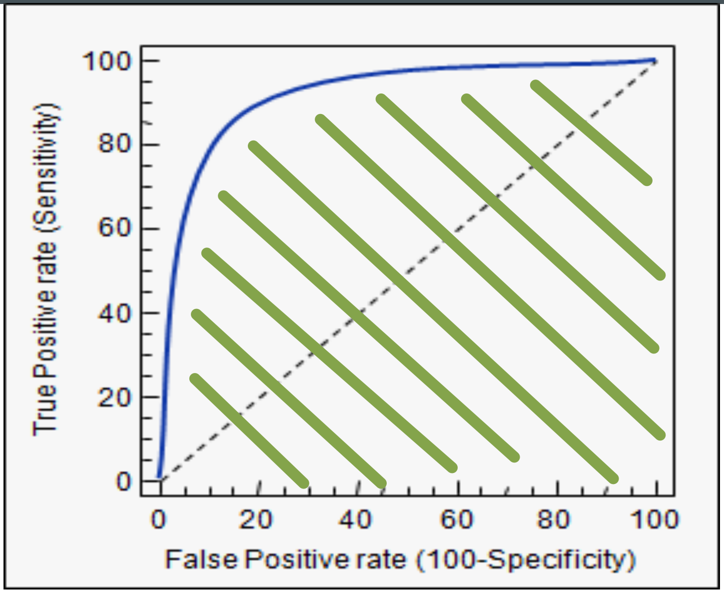
Race accuracy differences:

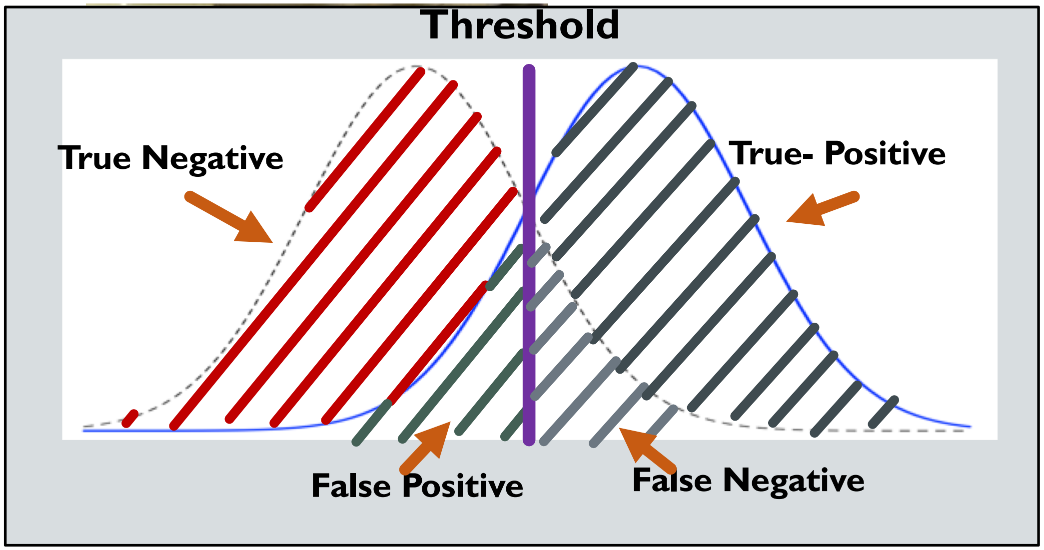
- Most evident as image **difficulty increases.**

OVERVIEW

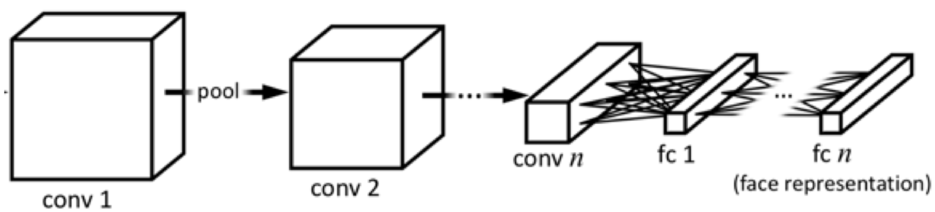
- Background on the other-race effect and race demographic variation
  - Humans and machines
- Measuring human and machine performance
- What factors impact accuracy differences across race groups in algorithms?
- Considerations for measuring these differences?
  - A walk through sample data: demographic variation in deep networks Cavazos, Phillips, Castillo, O'Toole (2019)
- Final thoughts/considerations on race accuracy variation

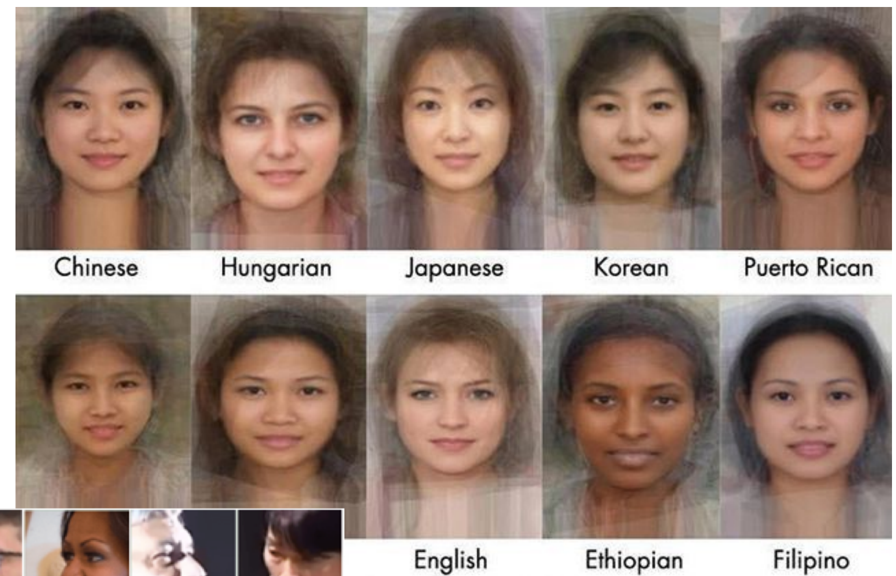
FINAL THOUGHTS

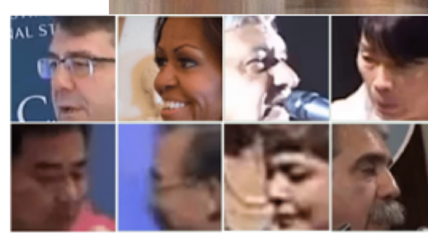




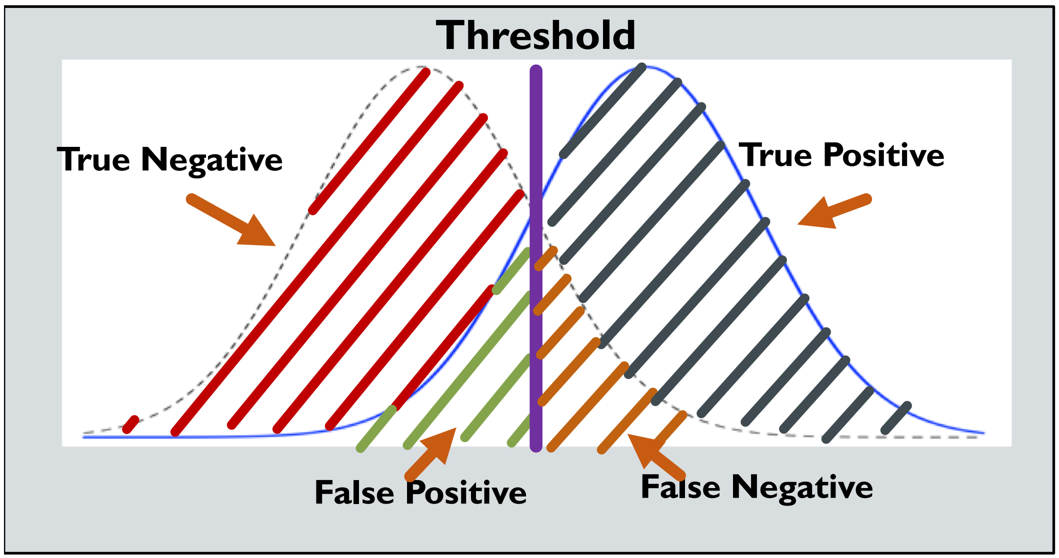
FINAL THOUGHTS

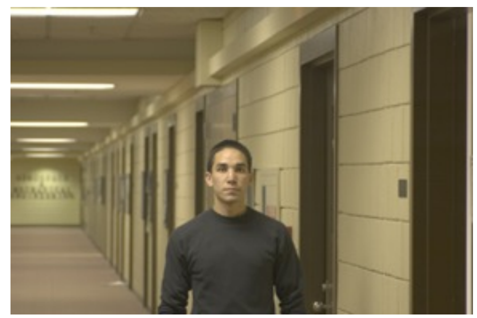


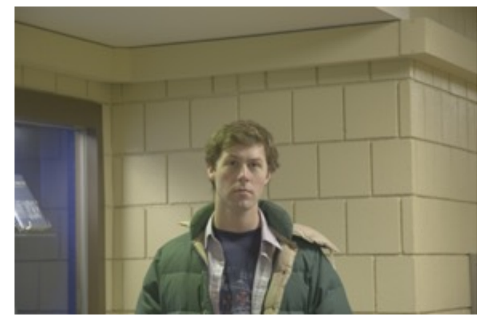


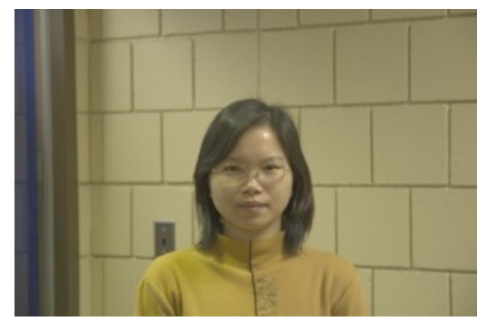


FINAL THOUGHTS



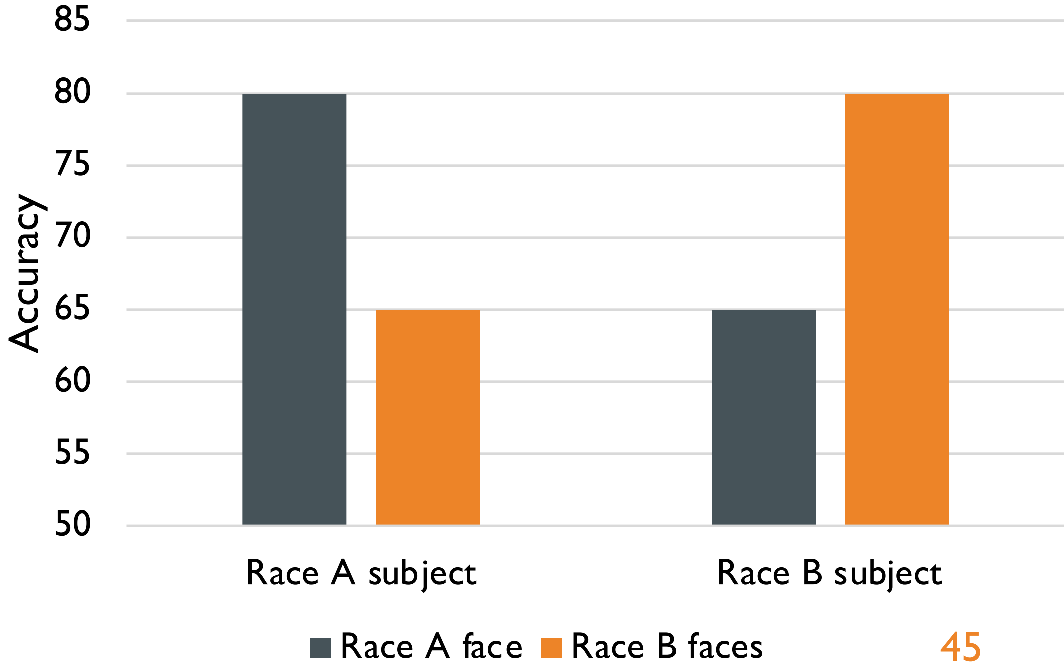






MYTHS ABOUT DEMOGRAPHIC VARIATION

- Myth #1: There would be no race performance variation in face identification if we eliminated machines.



MYTHS ABOUT DEMOGRAPHIC VARIATION

- Myth #2: Face recognition systems used to be fair before 2015 and the emergence of deep convolutional neural networks

A. J. O'Toole, K. Deffenbacher, H. Abdi, and J. C. Bartlett, "Simulating the 'other-race effect' as a problem in perceptual learning," *Connection Science*, vol. 3, no. 2, pp. 163–178, 1991.

N. Furl, P. J. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science*, vol. 26, no. 6, pp. 797–815, 2002.

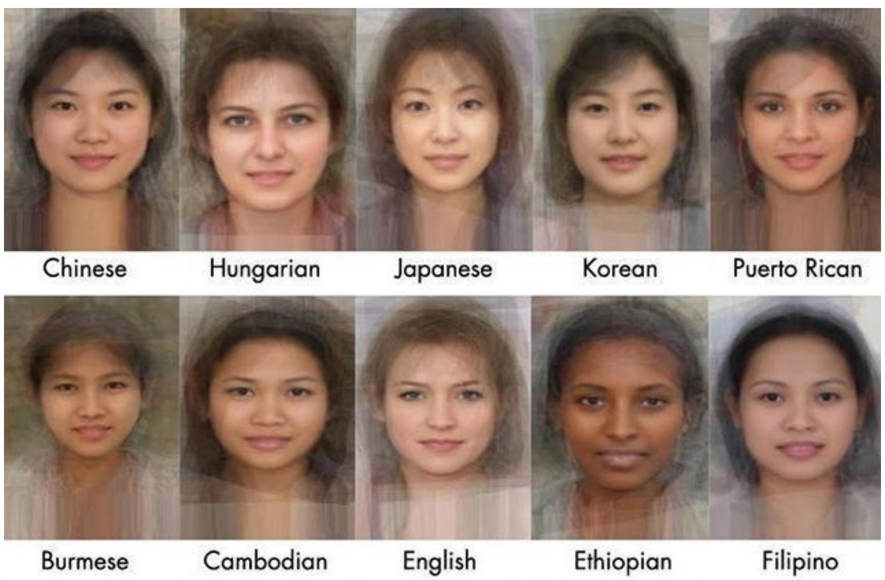
A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop, "Demographic effects on estimates of automatic face recognition performance," *Image and Vision Computing*, vol. 30, no. 3, pp. 169–176, 2012.

P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Transactions on Applied Perception (TAP)*, vol. 8, no. 2, p. 14, 2011.

B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 6, pp. 1789–1801, 2012.

MYTHS ABOUT DEMOGRAPHIC VARIATION

- Myth #3: Race is categorical. And we know what these categories are.



Acknowledgements

- Face Perception Lab
  - **Dr. Alice O'Toole**
  - Asal Barachizadeh
  - Matthew Q. Hill
  - Ying Hu
  - Gerie Jeckeln
  - Connor J. Parde
  - Parisa Jesudasen
  - Victoria Huang
  - Snipta Mallick
- NIST
  - **Dr. P. Jonathon Phillips**
- *Johns Hopkins University
  - **Dr. Carlos Castillo**
  - Dr. Rama Chellappa

Cavazos, J. G., Phillips, P. J., Castillo, C. D., & O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

*correction: original presentation stated: The University of Maryland

Acknowledgements

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Research funded by **National Eye Institute** of the National Institutes of Health RO1 EY 029692-01 to A.O.T, **National Institute of Justice, IARPA JANUS Program**

THANK YOU, QUESTIONS?