The effectiveness of fusion in face recognition

Carina A. Hahn

Amy N. Yates, P. Jonathon Phillips

National Institute of Standards and Technology

International Face Performance Conference 2020

Background and Goals

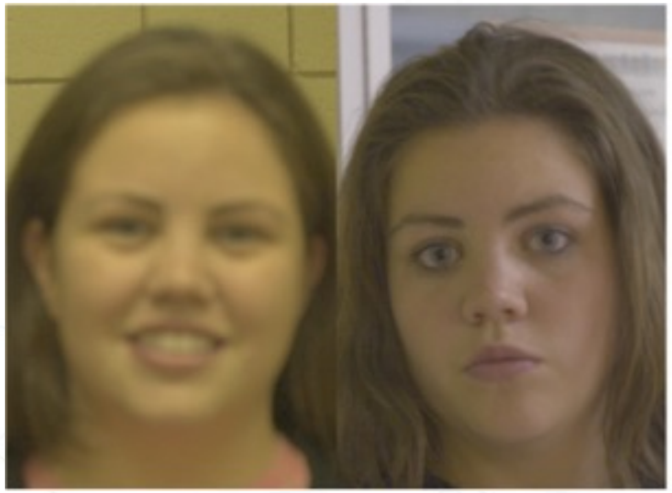
- Fusing: combining judgments across performers (people and/or algorithms)
  - "wisdom of crowds"
- Fusing humans + algorithms highly effective (Phillips et al., 2018)[1]
  - Used simple fusion strategy
  - Every person is fused with algorithm
- Current study: more detailed evaluation of fusing humans and machines
  - When to fuse? When to take only one performer's judgments?
- Goal: improve accuracy of system

[1]Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*

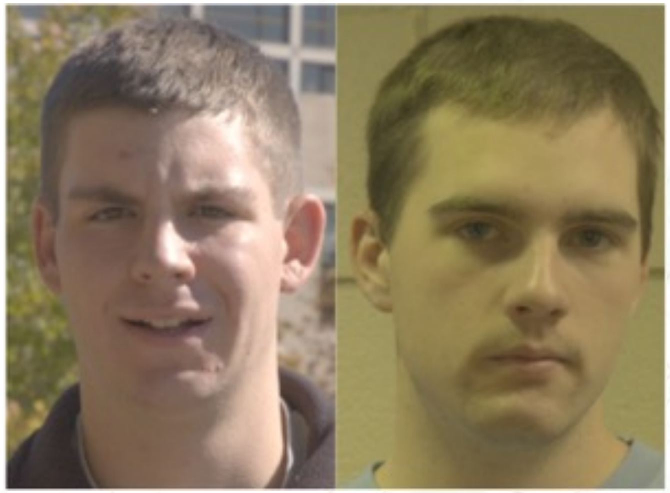
Task: Facial comparisons

- Facial comparisons are conducted for a variety of reasons (AKA face matching or face recognition)
- Task: determine whether images are of same person or of different people

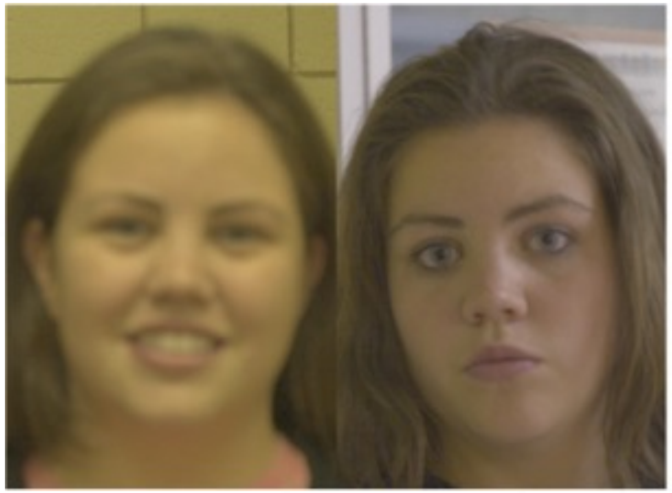
Same-identity pair



Different-identity pair



Background: Simple fusion strategy



+3 The observations strongly support that it is the same person
+2 The observations support that it is the same person
+1 The observations support to some extent that it is the same person
0 The observations support neither that it is the same person nor that it is different persons
-1 The observations support to some extent that it is not the same person
-2 The observations support that it is not the same person
-3 The observations strongly support that it is not the same person

Human judgments

Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*

Background: Simple fusion strategy

Algorithm: Similarity score

Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*

Background: Simple fusion strategy

Humans and Algorithms

Accuracy: AUC

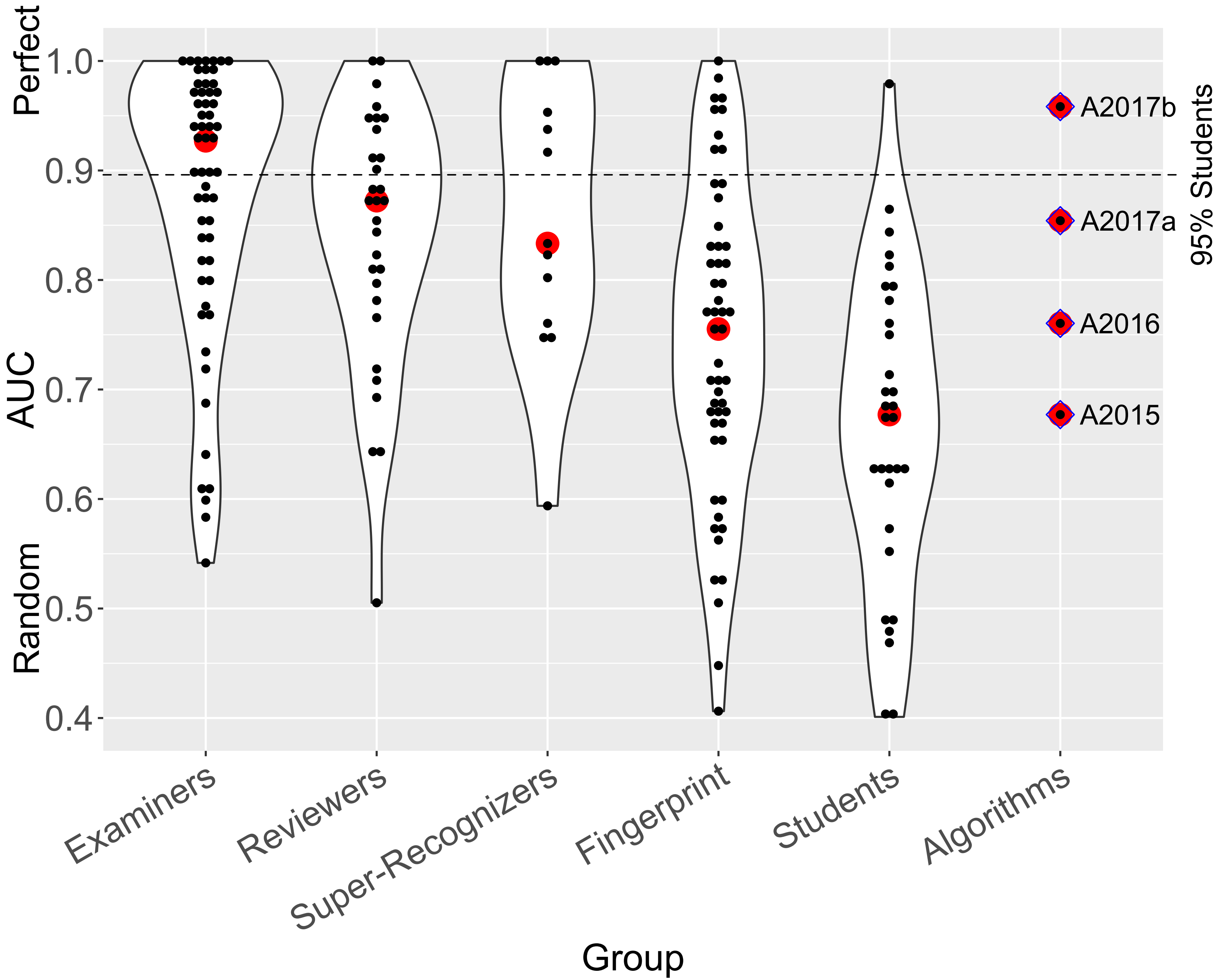
Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*

Five Subject Groups and Algorithms

- Forensic facial professionals (n=87, 5 continents)
  - Examiners (n=57)
  - Reviewers (n=30)
- Super-recognizers (n=13)
- Fingerprint examiners with no face experience (n=53)
- Undergraduate Students (n=30)
- Algorithms
  - VGG-Face (A2015)
  - U. of Maryland (A2016, A2017a, A2017b)

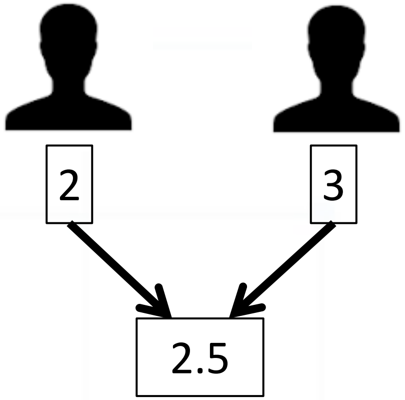
Phillips, P. J. et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*

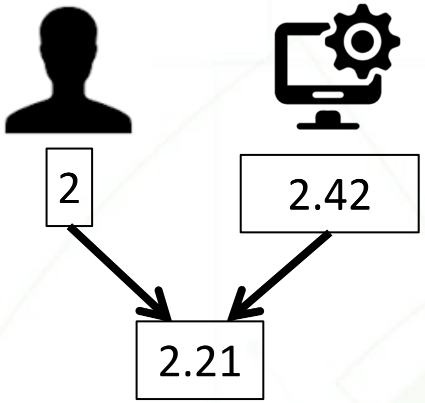
Results: Individual judgments



Phillips, P. J. et al. (2018)

Approach: Simple fusion strategy

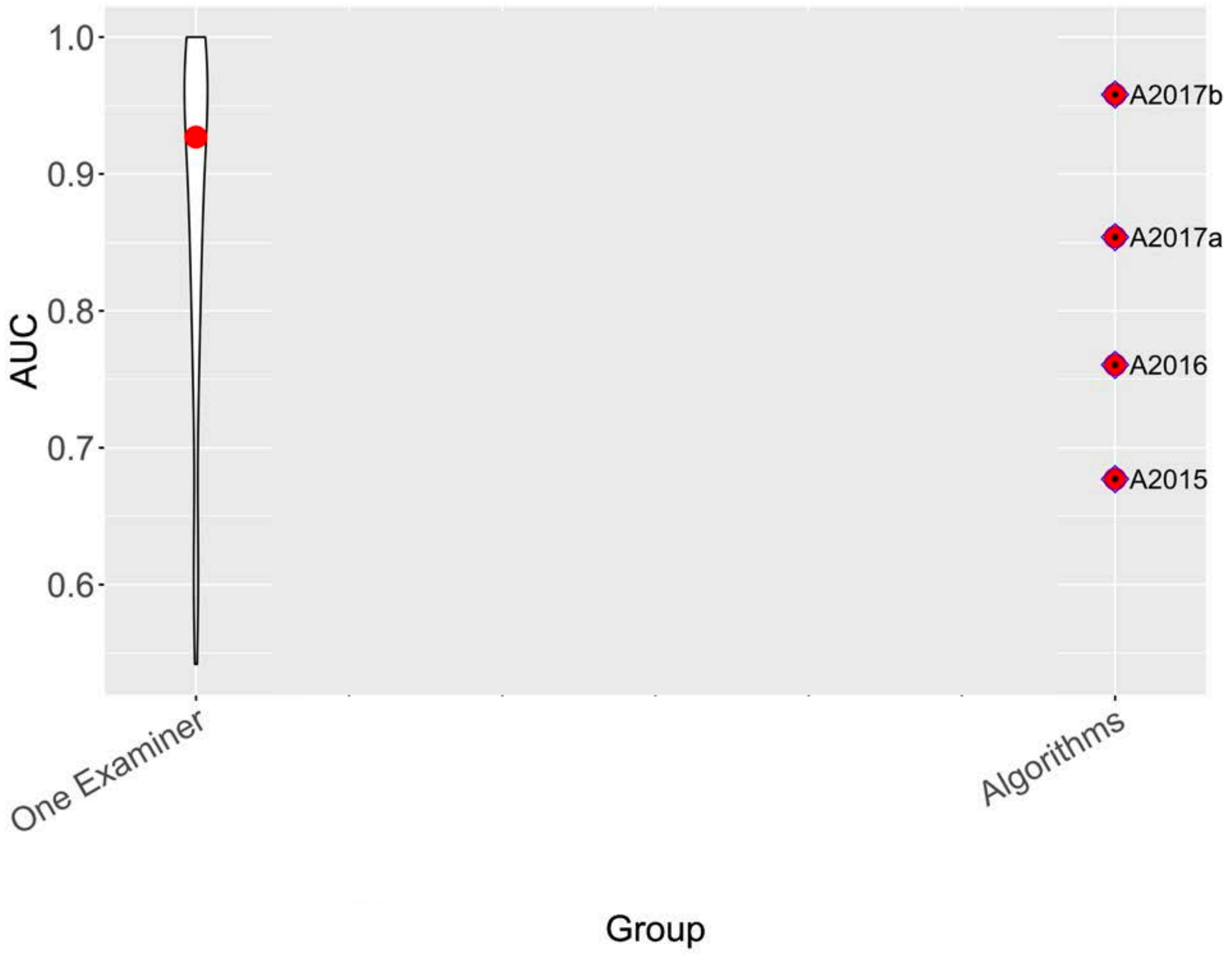




Algorithm:
Rescaled to human judgments

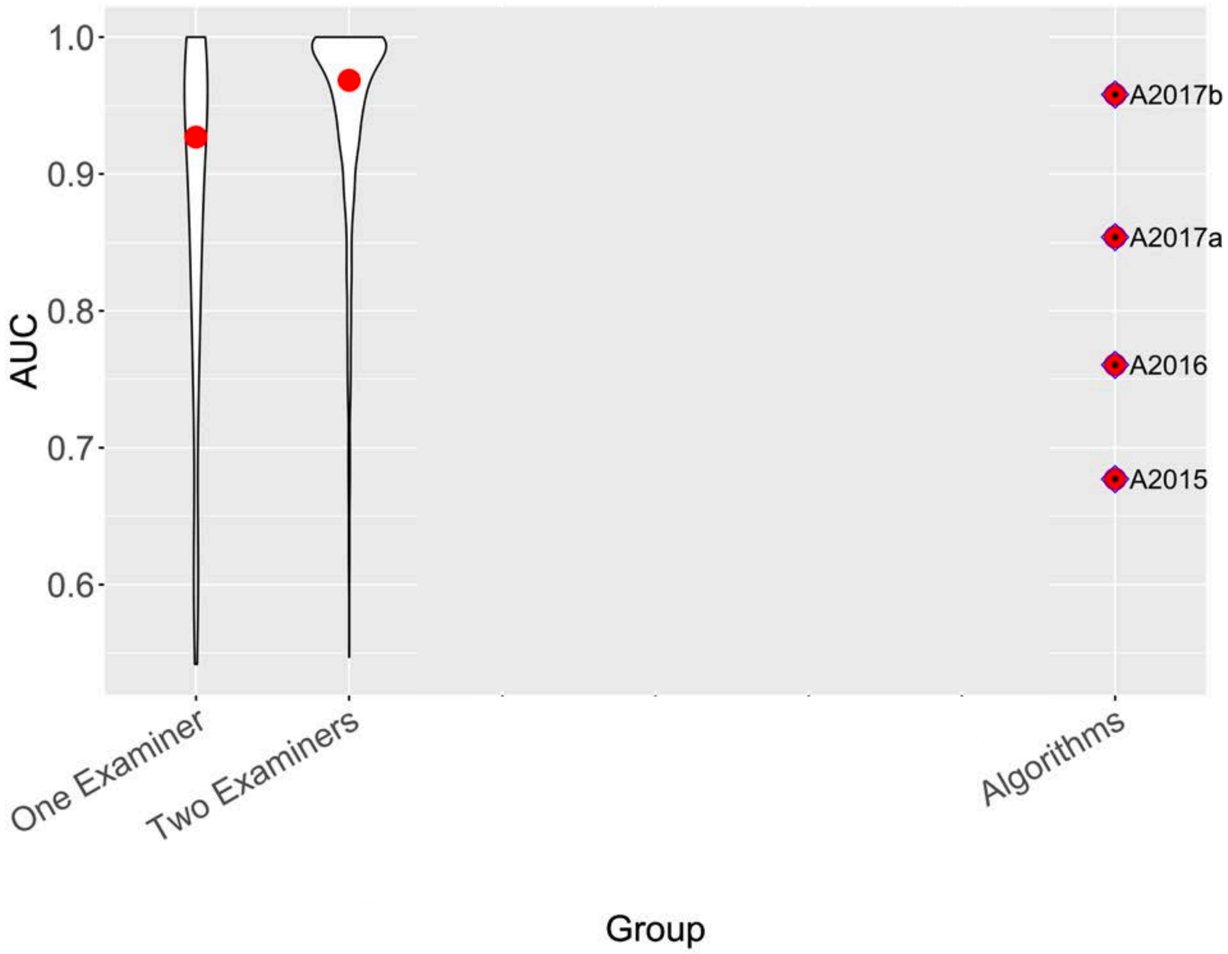
Phillips, P. J. et al. (2018)

Results: Simple fusion strategy



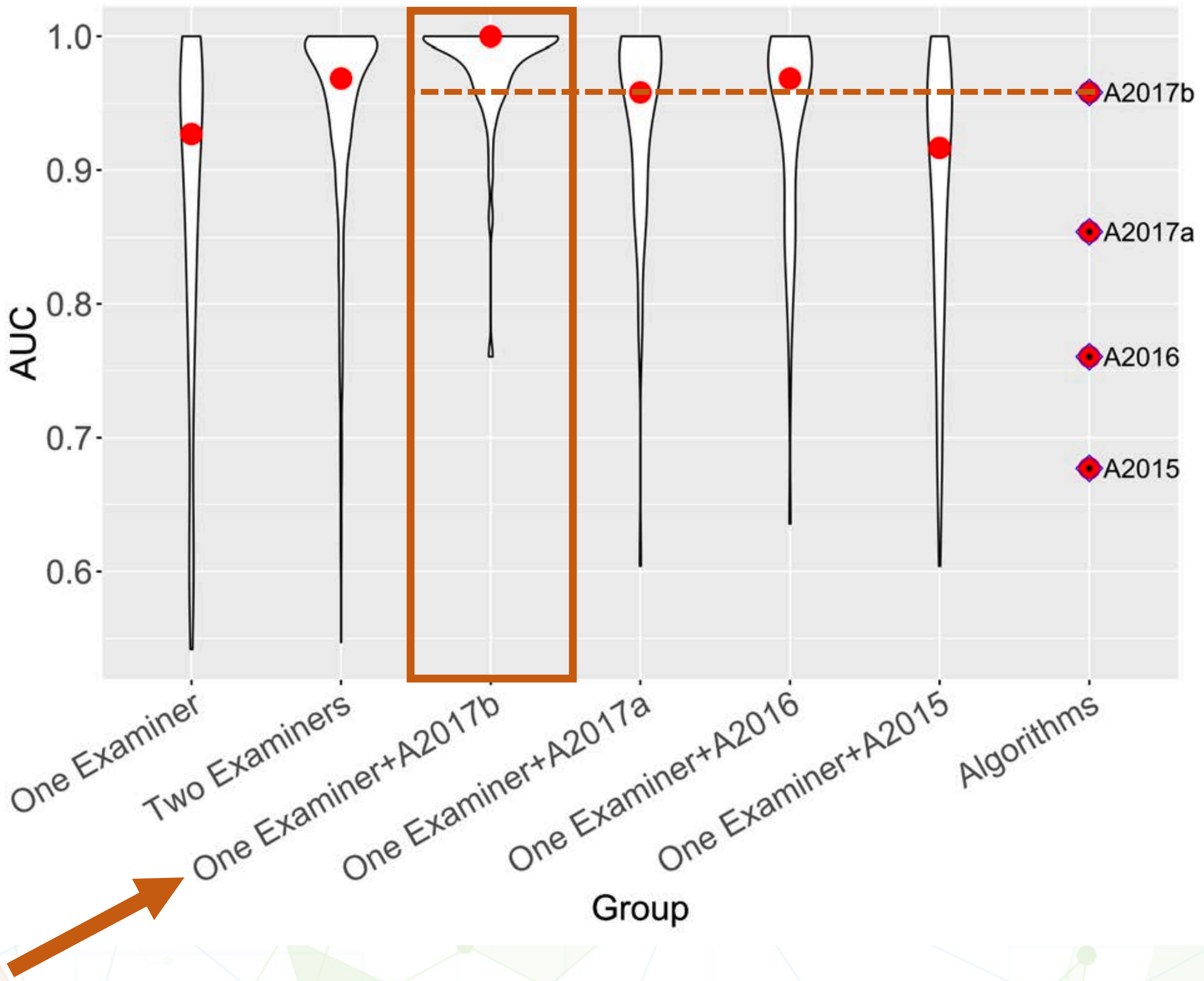
Phillips, P. J. et al. (2018)

Results: Simple fusion strategy



Phillips, P. J. et al. (2018)

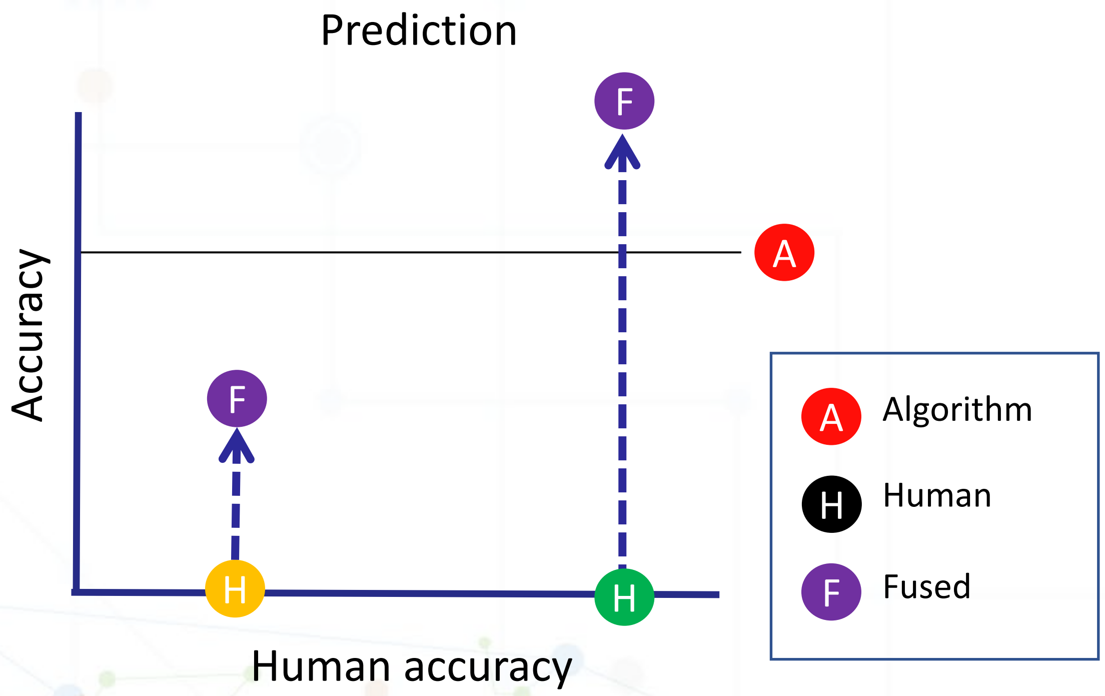
Results: Simple fusion strategy



- Some humans contribute to increase
- Some humans decrease
- Threshold to determine who to fuse
- **How to find that threshold?**

Phillips, P. J. et al. (2018)

Who to fuse?

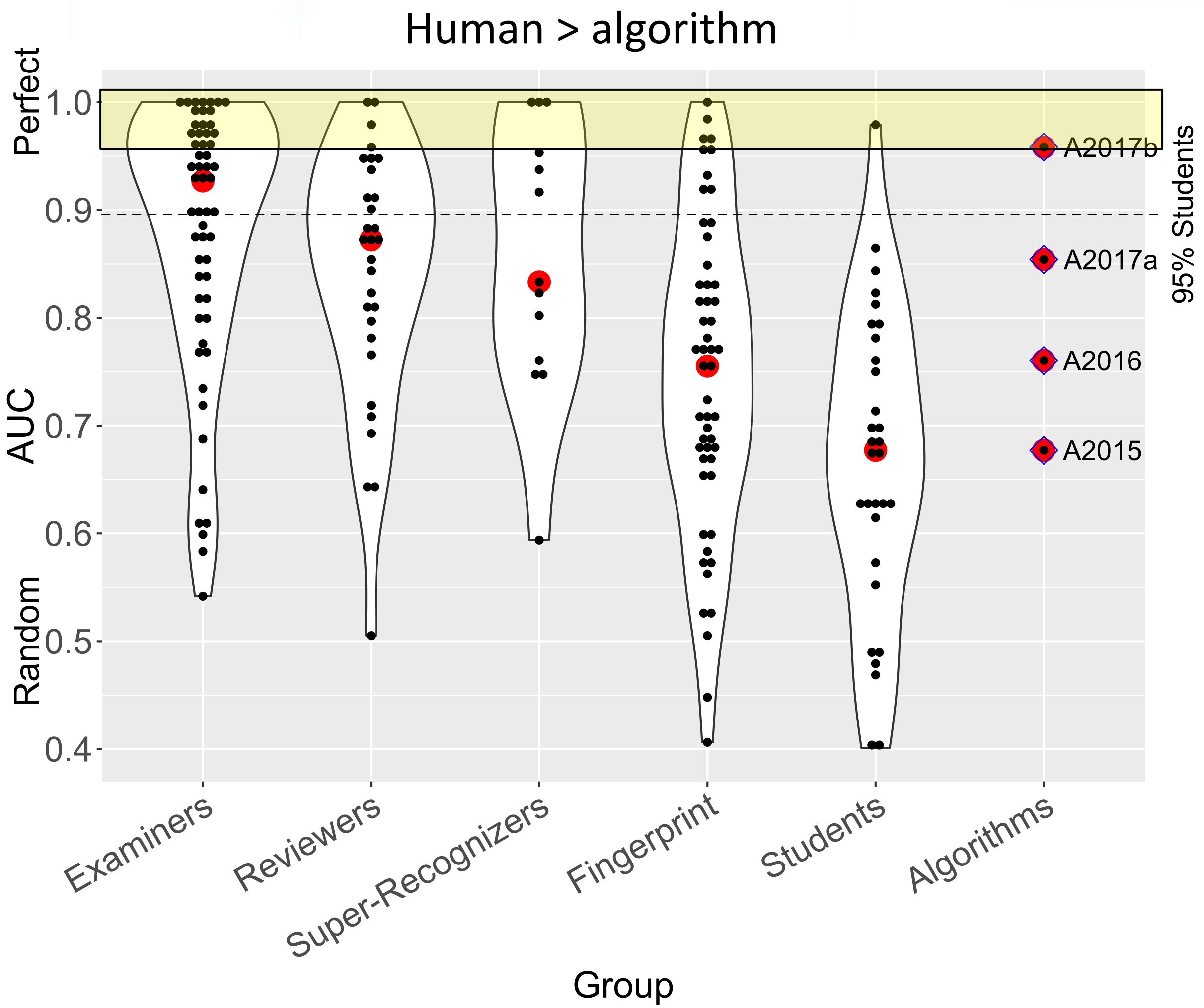


Prediction for similar accuracies: Fusing judgments increases accuracy

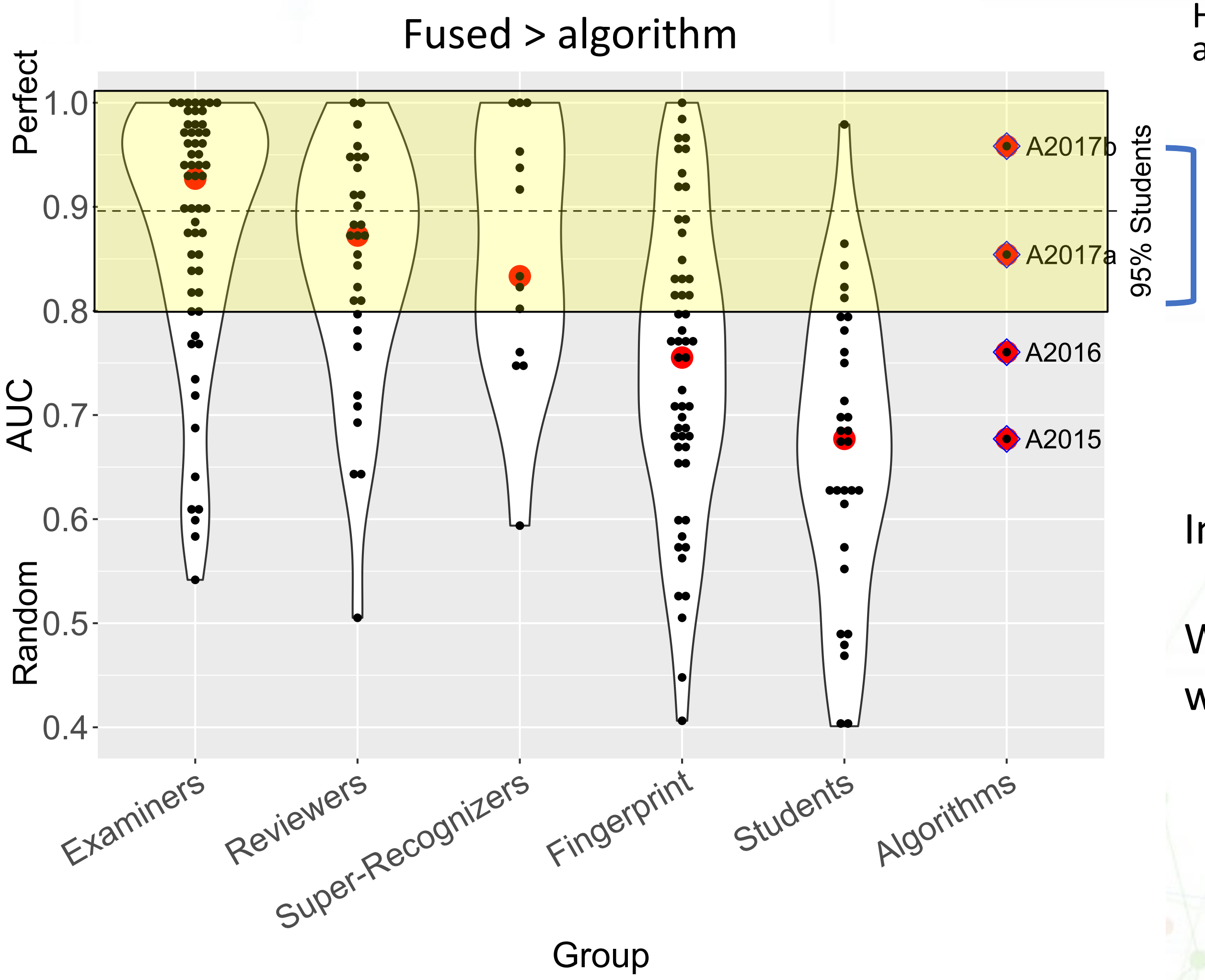
Prediction for large differences: Fusing judgments decreases accuracy

- Only *some* people should be fused
- Judgments from more accurate performer should be used otherwise

Approach: Finding threshold



Approach: Finding threshold



0.96

Threshold AUC: 0.16

0.80

Humans with accuracy < algorithm contributed to boosting overall accuracy

In this case: We have all the answers

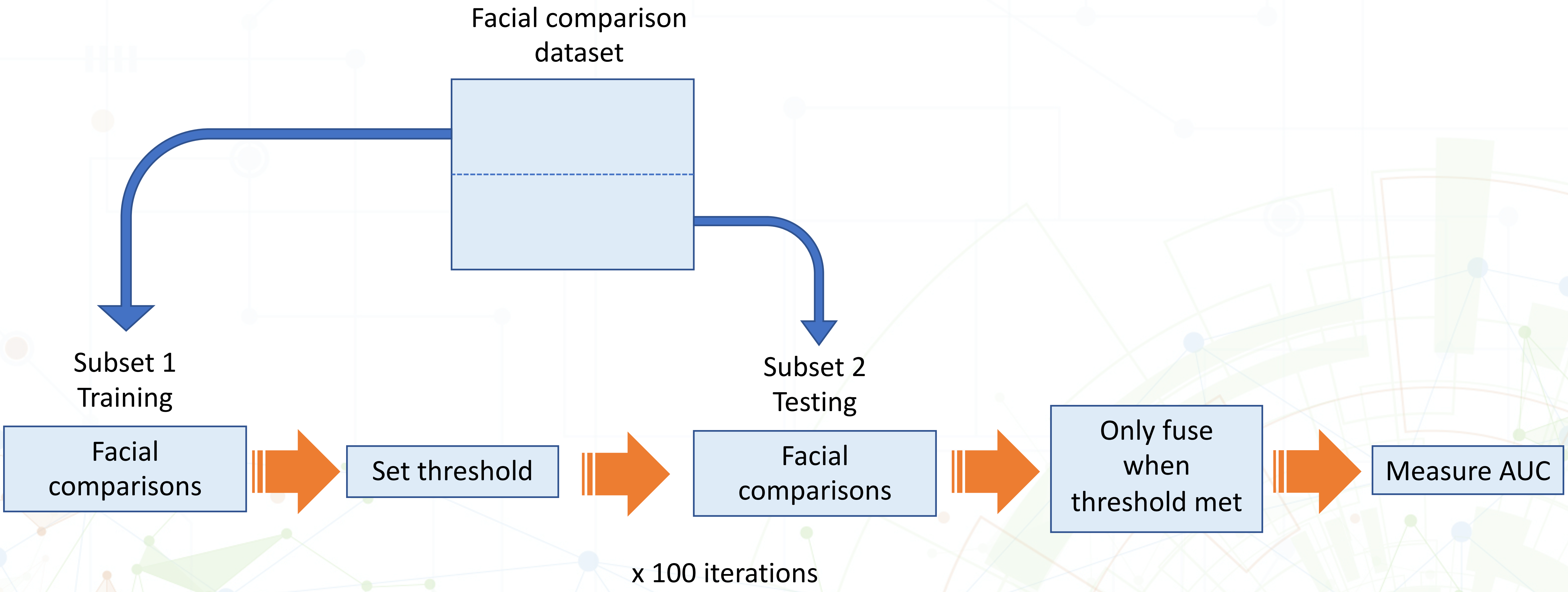
What happens when threshold is determined and used with new images and new people?

Testing a fusion strategy

- Find threshold (*selective fusion*)
  - Generalize to new facial comparisons
  - Generalize to new people
- Options:
  - If within threshold: Fuse
  - If outside of threshold: Take more accurate
    - Human alone
    - Algorithm alone
- Tested with data from White et al. 2015[2]
  - More facial comparisons to separate into training and test
- Algorithm: VGG-Face on White et al. 2015 facial comparisons

[2] White, D., Phillips, P. J., Hahn, C. A., Hill, M., & O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*

First case: Generalize to new images



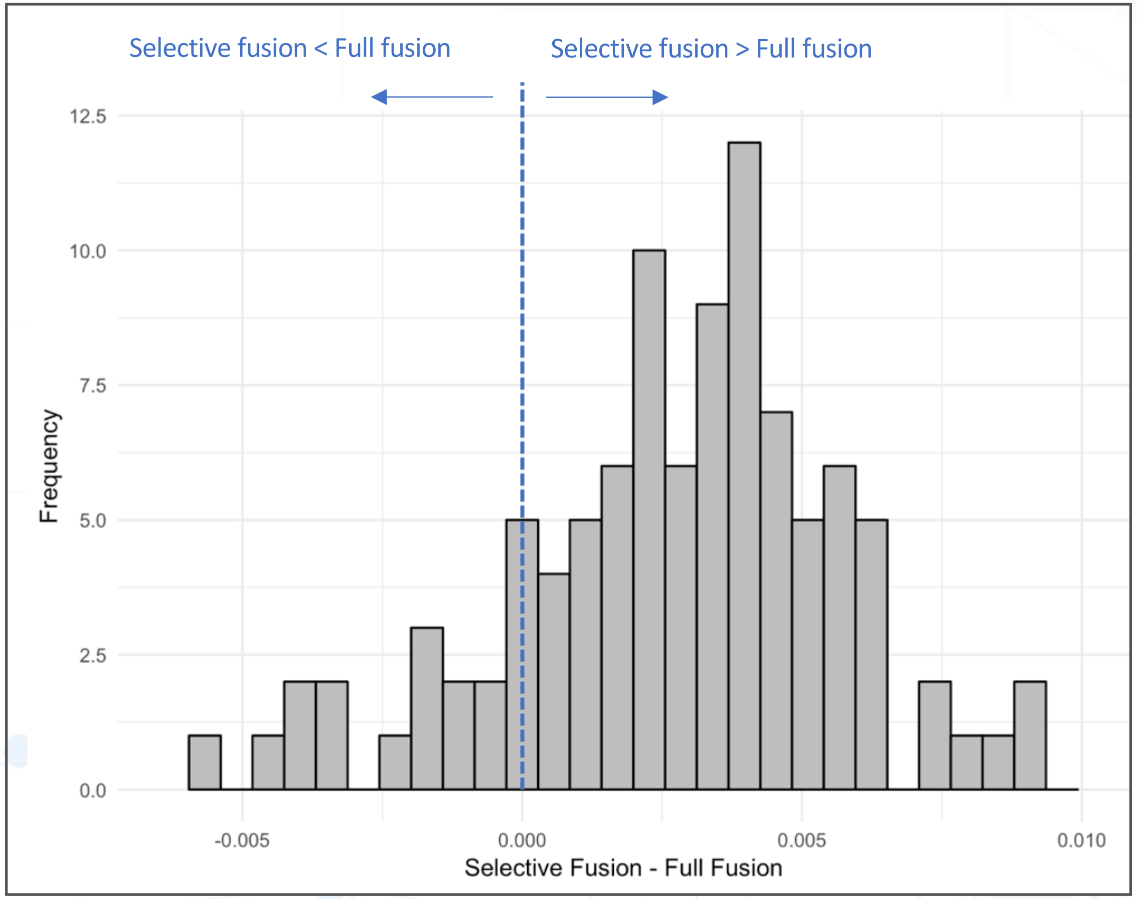
Question: Higher accuracy than fusing with everyone? *(full fusion)*

Measuring success

- % of cases where selective fusion (threshold based) > full fusion (everyone is fused with algorithm)

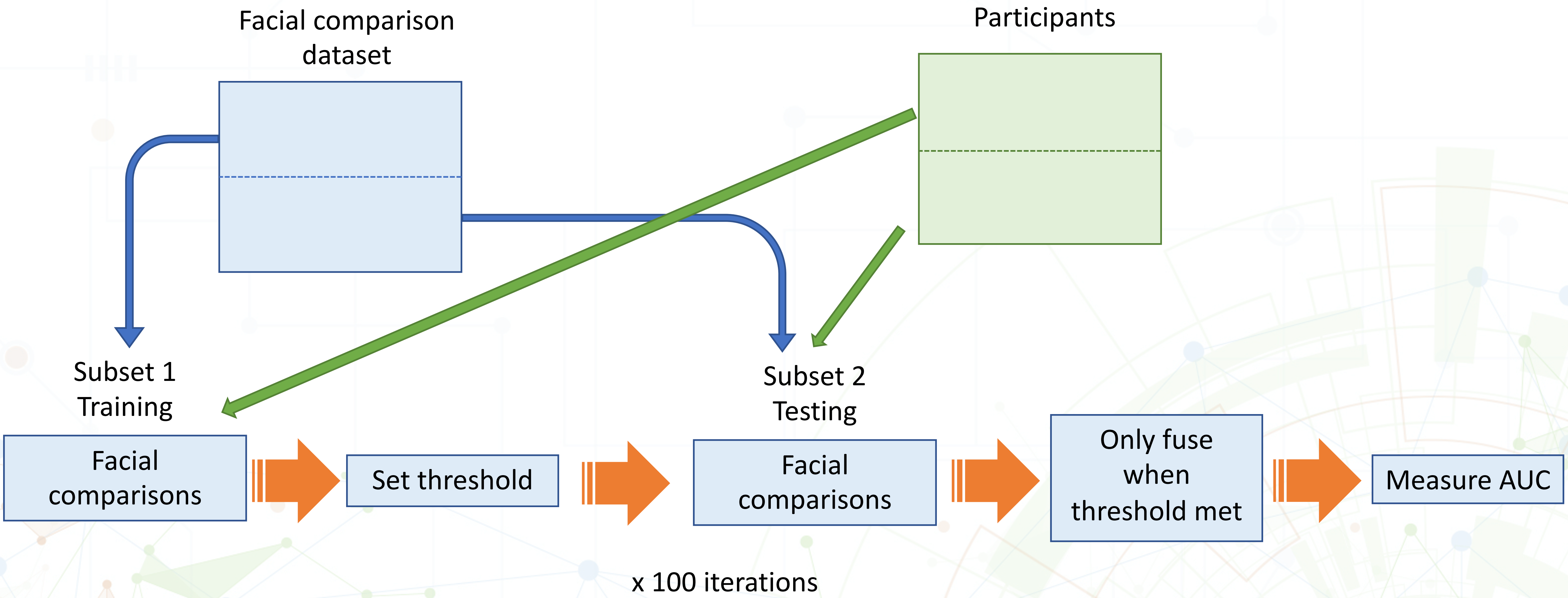
Results

Train and test across images



86% of the time, threshold-based fusion was helpful or neutral
14% of the time, threshold-based fusion does not improve

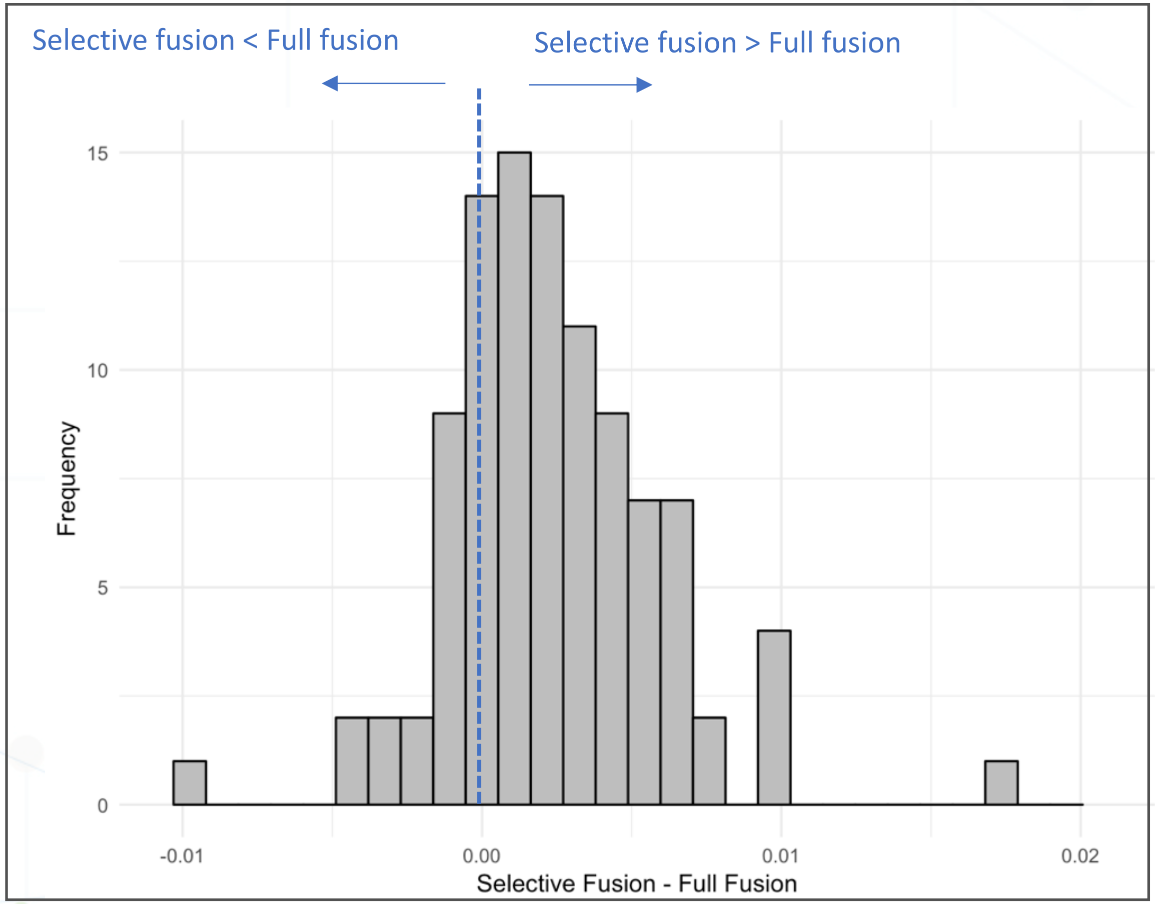
Second case: Generalize to new images & people



Question: Higher accuracy than fusing with everyone?

Results

Train and test across images & participants



79% of the time, threshold-based fusion was helpful or neutral
21% of the time, threshold-based fusion does not improve

Summary

- Selective fusion benefits small but reliable
  - Across new images
  - Across new images and people

Conclusions

- Fusion is effective
  - Not many humans outperform best algorithm
  - Humans with accuracy < algorithm contributed to boosting overall accuracy via fusion
  - Suggests humans and algorithms use different strategy
  - Differences exploited via fusion for benefit
- *Threshold-based, selective fusion* strategy can be applied to improve overall accuracy
  - Benefit of threshold-based fusion generalizes
    - When person's ability on new set of facial comparisons is unknown
    - When new people are added to the system

Conclusions

- Future directions
  - More research: generalization only on one test
  - How translate to other domains?
  - Different threshold types (e.g., weighted relative to AUC distance; asymmetrical)
- Which strategy for highest accuracy?
  - Humans alone
  - Algorithm alone
  - Fusing all humans + algorithm
  - Fusing humans + algorithm, based on *threshold* ←

QUESTIONS?

Contact:
carina.hahn@nist.gov