DHS SCIENCE AND TECHNOLOGY

Iris Vendor Results From the 2018 Biometric Technology Rally

Iris Experts Group, June 2018

John J. Howard, Maryland Test Facility

Arun Vemury, DHS S&T

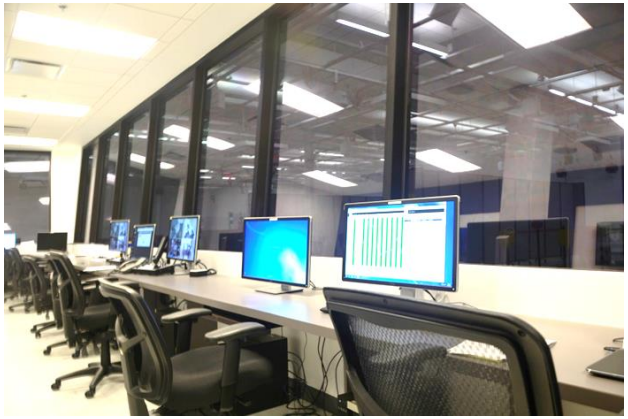
Biometrics Technology Engine
Department of Homeland Security
Science and Technology Directorate

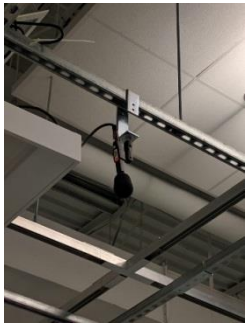
Outline

- The Maryland Test Facility
- The 2018 Biometric Technology Rally
  - Motivation
  - Process
  - Metrics
- Rally Results
  - True identification rates in a high throughput environment
  - Primary error determinants
  - Positive outcomes
- General Discussion
  - High throughput systems
  - High throughput metrics
  - Industry expectations
  - Primary error determinants
  - Operational tradeoffs
  - Demographics

The Maryland Test Facility (MdTF)

- 10,000 square feet of test space, consenting and debriefing areas.
- Designed and constructed to facilitate DHS efforts to incorporate biometrics at border crossings
- Fully instrumented, custom software
- To date over 2500 subjects have progressed through the MdTF
  - Ages 18-81
  - Over 72 countries of origin



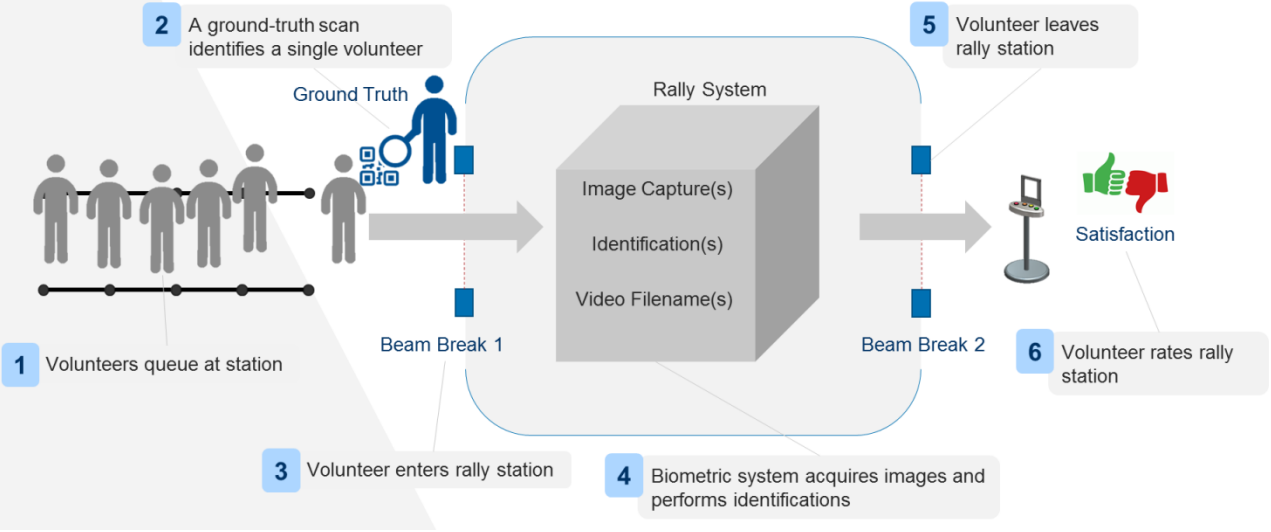


2018 Biometric Technology Rally

- Response to observed high failure to acquire rates in operational deployments
- Goals:
  - Formalize the “high-throughput” use case
  - Fairly access the state of the industry in regards to EES
  - Promote industry innovation and further market maturity
  - Inform DHS and other government acquisition
  - Guide promising technologies, share information via CRADA
- Benefits to the vendors:
  - Data
  - Immediate feedback
  - Showcase systems via VIP day

Report on Operational Performance

2018 Biometric Technology Rally



- Required:
  - Collect 1 Face
  - Fit in a 7x8 ft. space
  - Be unmanned
  - Direct all interaction
  - Take on average 10 s. per person
- Optional:
  - Collect 3 Faces
  - Collect 3 Irises
  - Provide Facial Identifications
  - Collect Video
- Test Process:
  - 3 month development
  - 11 systems, 2 day install
  - 363 subjects
  - Groups of 15 over 5 days
  - General instructions provided
  - Enrollment
  - Counterbalanced

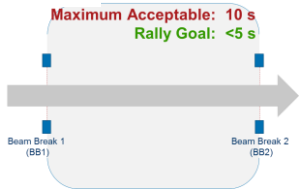
Rally Metrics

- Efficiency
  - Refers to the amount of time required to use each biometric system
  - Quantified as average transaction time (beam-break to beam-break) for Test Volunteers at each Rally System

- Effectiveness
  - Refers to the accuracy and completeness with which users are identified.
  - Measured in two time intervals:
    - By 5 seconds after the entry beam break
    - By 20 seconds after the entry beam break
  - Failure to Acquire Rate (FtAR) for face and iris images
    - Proportion of Test Volunteers for whom no images were captured
  - True Identification Rate (TIR) for face and iris images
    - The proportion of Test Volunteers correctly identified
    - vTIR: Identity of Test Volunteers provided by Rally Systems
    - mTIR: MdTF ability to identify Test Volunteers using images provided

- Satisfaction
  - Refers to Test Volunteers' positive attitudes toward the Rally Systems
  - Measured using a 4-button kiosk from Very Happy to Very Unhappy
  - Quantified as proportion of Happy or Very Happy responses

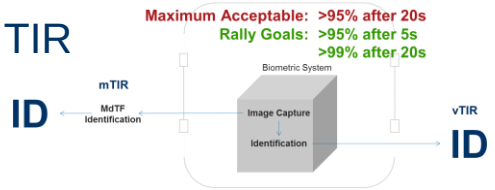
Efficiency



Effectiveness:

FtAR

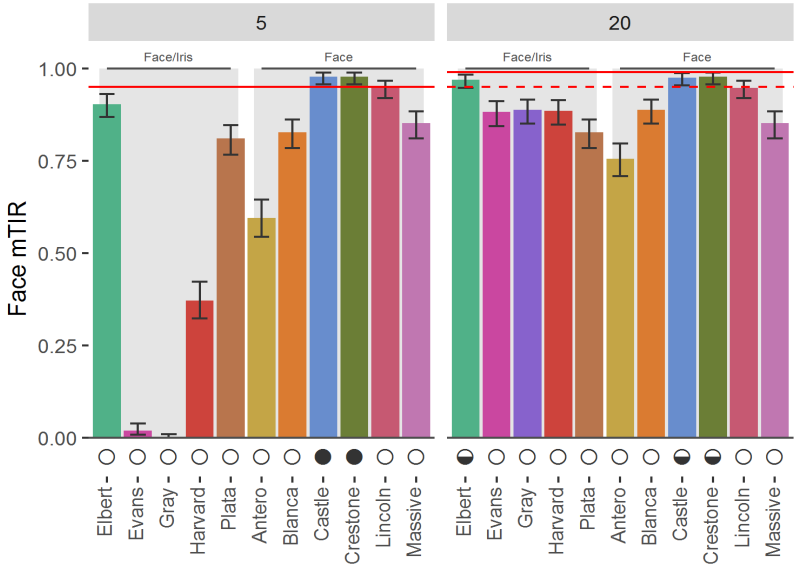
TIR

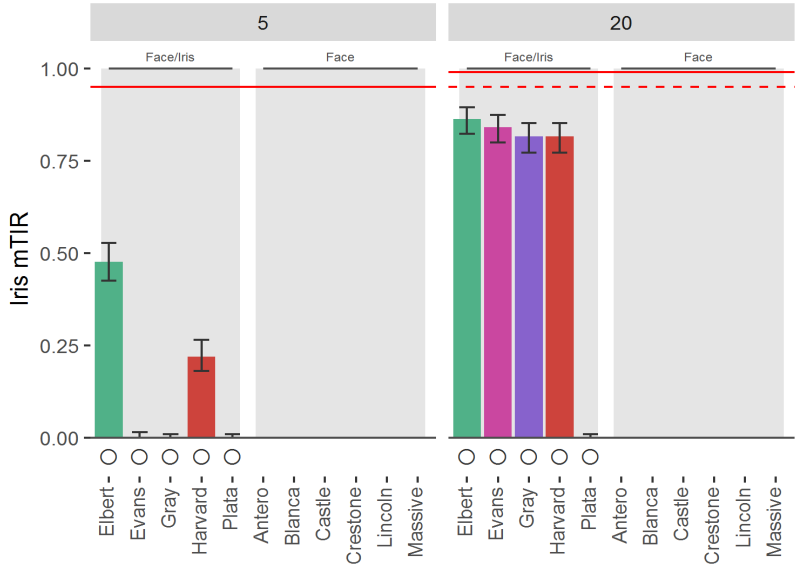


Satisfaction

Rally Results – True Identification Rates (Face & Iris)

- True Identification Rates:



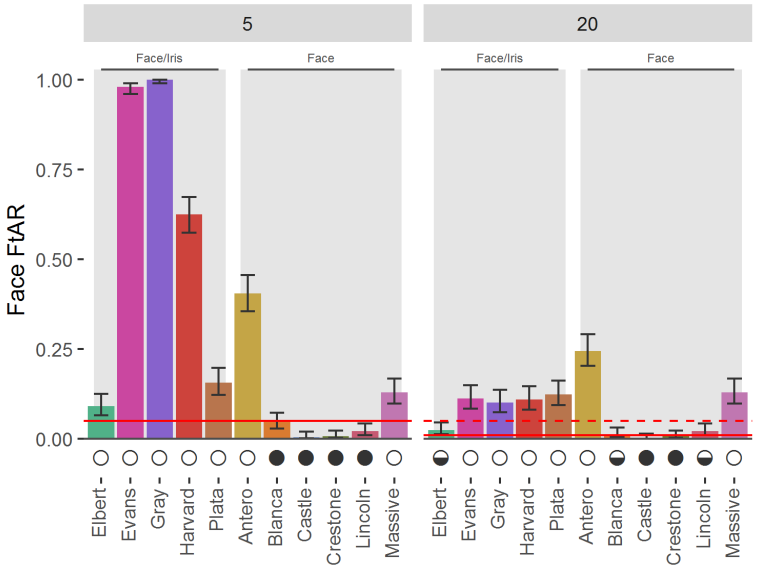


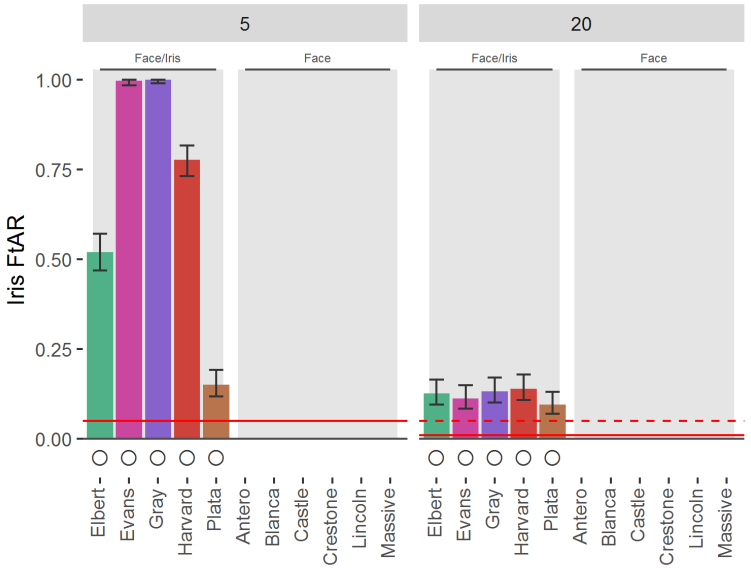
- Three face only stations with high TIR with face samples
  - Castle (97.5), Crestone (97.8), Lincoln (94.8)
- One face/iris system
  - Elbert (97.0)

- No face/iris system achieved a high TIR with iris samples
  - Elbert (86.2), Evans (84), Gray (81.5), Harvard (81.5)

Rally Results – FtA as a primary driver of non-identification

- Why? Failure-to-acquire

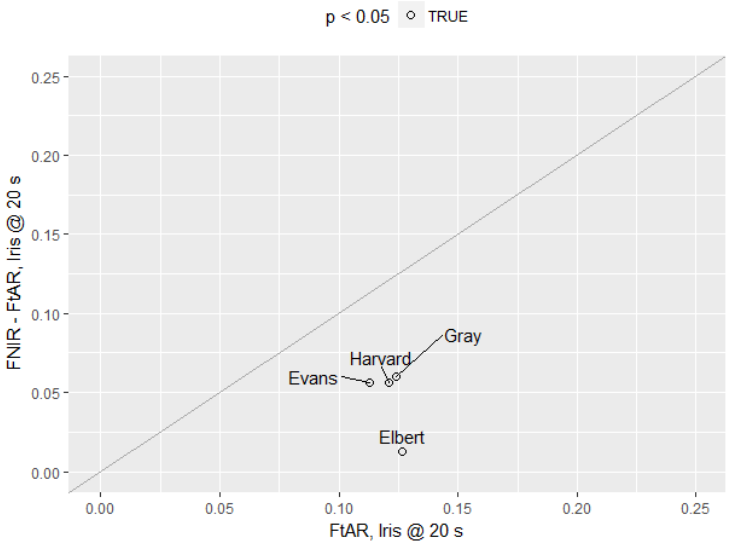


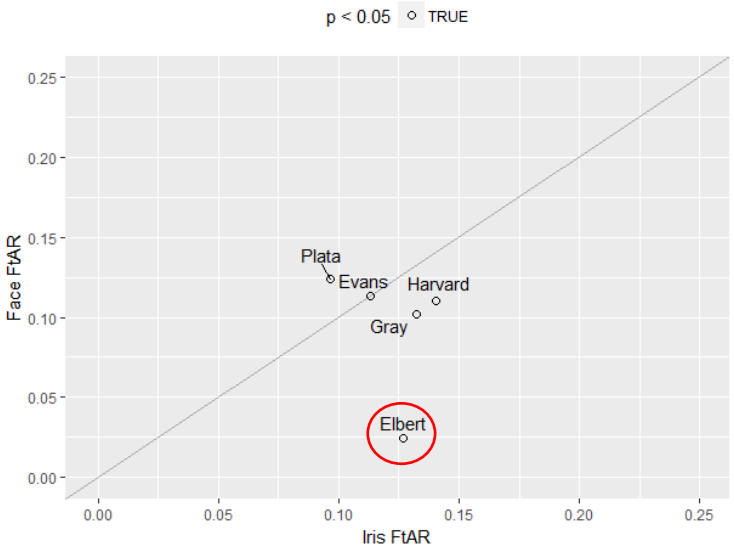


- Face FtAR
  - Elbert (2.5), Evans(11.3), Gray(10.2), Harvard (11.0), Plata (12.4)
- Iris FtAR
  - Elbert (12.7), Evans (11.3), Gray (13.2), Harvard (14.0), Plata (9.6)

Rally Results – FtA as a primary driver of non-identification

- Why? Failure-to-acquire



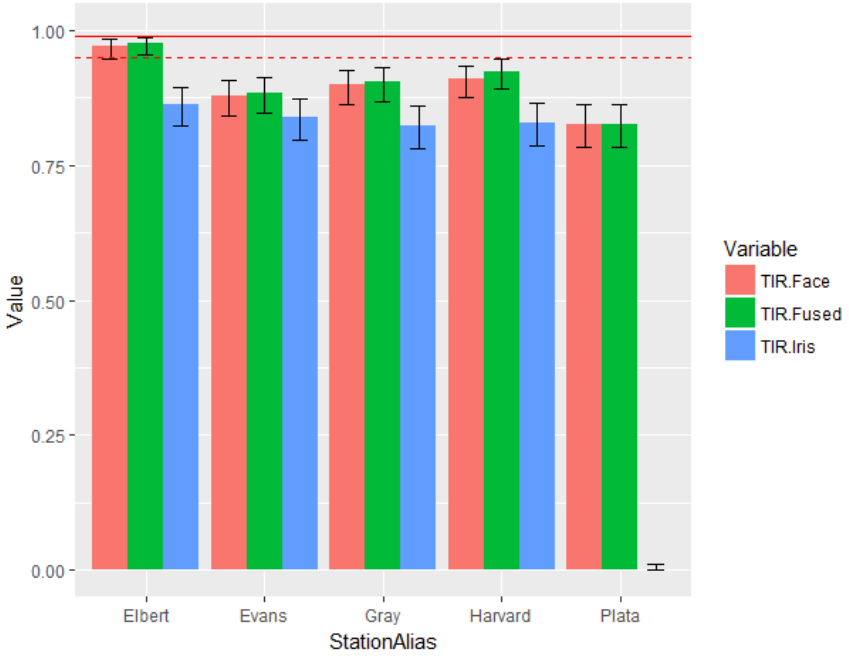


- Failure to acquire was main driver of non-identification for all face/iris systems
- FtA was main driver of non-identification for 3 of 6 face only stations

- Only one station appeared to have independent face + iris image submission capability

Rally Results – Impact of Fusion

- Face + two iris fusion helped, a little



- Fused (match any) TIR results raised two face/iris systems above 90%.

Rally Results – Positive Outcomes for Iris

- Two encouraging outcomes (for iris community)
  - Zero in gallery false positives for iris

Face *t* = 0.54
FMR = between 1/5000 (us) and 1/10,000 (them)
Expected FPIR = 9.0% and 5.1%
Actual FPIR = 19%

Table 1: MdTF Reported False Positive Identifications

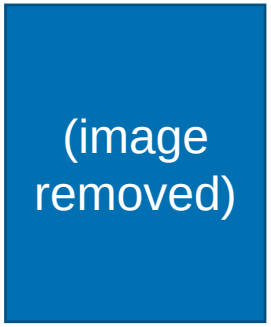
Station Alias	In Gallery False Positives	In Gallery Image Captures	In Gallery FPIR	Out of Gallery False Positives	Out of Gallery Identification Attempts	Out of Gallery FPIR
Elbert	0	318	0%	1	36	2.7%
Evans	0	290	0%	2	32	6.3%
Gray	0	295	0%	1	35	2.9%
Harvard	0	298	0%	1	34	2.9%
Plata[1,2]	5	287	1.7%	3	33	9.1%
Antero	0	258	0%	0	29	0%
Blanca	0	326	0%	0	37	0%
Castle[1]	2	324	0.6%	1	37	2.7%
Crestone	0	324	0%	0	37	0%
Lincoln	0	318	0%	1	37	2.7%
Massive[2]	5	284	1.8%	1	33	3.0%

Iris *t* = 51.6
FMR = 1/10,000
Expected FPIR = 5.2 %
Actual FPIR = TBD

- For out of out-of-gallery subjects (37 in total) the majority of Rally Stations provided a face sample which was found to be a match to the incorrect individual in the gallery. Additionally, the 11 out-of-gallery false positive errors shown came from 7 different subjects, meaning nearly a fifth of the out-of-gallery subjects were incorrectly matched to *someone* on the gallery during the course of the Rally.
- Points to difficulty associated with open set facial identifications
- All iris subjects were “in gallery” also used same day samples – not apples to apples comparison, yet.

Rally Results – Positive Outcomes for Iris

- Two encouraging outcomes (for iris community)
  - Zero in gallery false positives for iris
  - One subject's face sample was out-of-gallery FPIR on images from 4 of 11 systems.
  - "Persistent" false positive error or zero effort imposter
  - Thresholds for face and iris algorithms set at similar FMR
  - Ongoing work to ascertain if there is a similar effect in iris

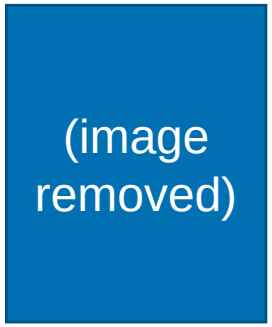


FMR	Genders	Races	Ages
0.0000283460	different	same	similar
0.0000386420	Same	different	similar
0.0001594400	Same	same	different
0.0004339340	Same	same	similar

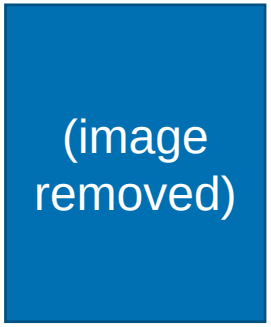
Rally Results – Positive Outcomes for Iris

- Two encouraging outcomes (for iris community)
  - Highest quality facial samples came from iris devices

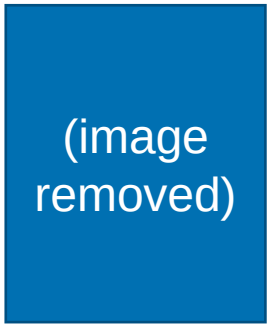
Gray



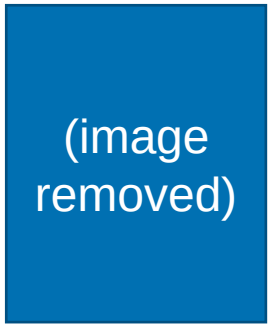
Harvard



Crestone



Lincoln



Blanca

Station	In Gallery ID Rate
Castle	0.9876543
Crestone	0.9785276
Elbert	0.9754601
Lincoln	0.9601227
Blanca	0.9171779
Harvard	0.9110429
Gray	0.8957055
Evans	0.8865031
Massive	0.8650307
Plata	0.8466258
Antero	0.7760736

Station	In Gallery ID Rate 2
Castle	0.9444444
Elbert	0.9079755
Gray	0.8865031
Harvard	0.8773006
Evans	0.8680982
Crestone	0.8619632
Massive	0.8312883
Plata	0.8006135
Lincoln	0.7791411
Antero	0.7361963
Blanca	0.601227

General Conclusions – High Throughput Systems

- High-throughput systems need further definition, system and human factors engineering, and overall maturity

What makes H.T. Biometrics Different[1]:

1) Hundreds to thousands of users in a short time frame
2) Because of these volumes, these systems must emphasize speed
3) Also because of these volumes, even sub percentage error rates equate to dozens of exception cases
4) In order to scale, H.T. systems must be optionally manned or purposefully understaffed. Need to be intuitive to naïve user without human intervention

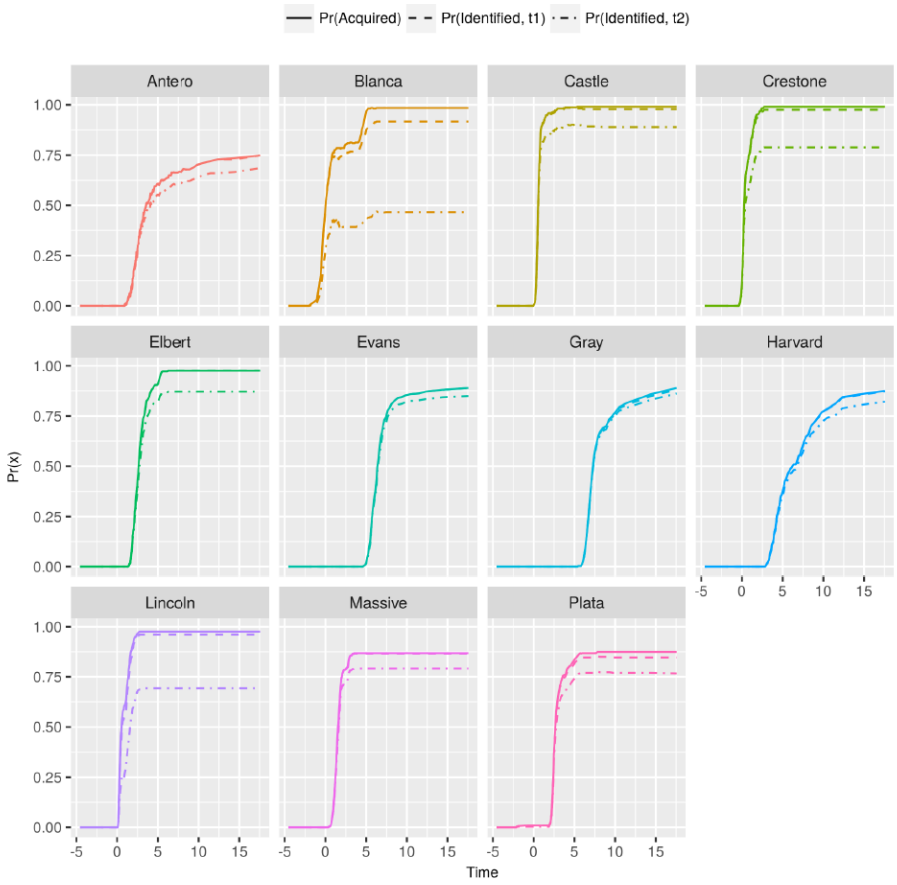
H.T. Systems need unique workflows[1]:

1) To achieve shortened processing times, high-throughput systems should have a strategy for acquiring a sample of "good-enough" quality quickly and to recognize when that condition has been achieved.
2) To maintain high biometric accuracy, high-throughput systems should adjust when good-enough quality samples are not being acquired.
3) To allow for scalability, high-throughput systems should perform collections with minimal operator intervention and need to be intuitive to the untrained user.

[1] Howard, et al. *On Efficiency and Effectiveness Tradeoffs in High-Throughput Facial Biometric Recognition Systems.* BTAS 2018.

General Conclusion – High Throughput Metrics

- High-throughput systems may need different kinds of metrics for proper evaluation
  - **ISO 19795-1, 8.2.2.3 "**The failure-to-acquire rate will depend on thresholds for sample quality, as well as the allowed duration for sample acquisition or allowed number of presentations. These settings shall be reported along with the observed failure-to-acquire rate"
  - How do you do that for 11 different "black box" biometric systems as in the Rally?
  - Time based performance curves[1]



[1] Howard, et al. *On Efficiency and Effectiveness Tradeoffs in High-Throughput Facial Biometric Recognition Systems.* BTAS 2018.

General Conclusions – Industry Expectations

- FtAR and TIR results were not well anticipated by industry[2]:
  - Six of the eleven Rally Participants elected not to provide FtA estimates, indicating this metric may be poorly understood or documented from an industry perspective
  - Measured FtAR was uniformly higher than those anticipated by the Rally Participants

Table 2. 2018 Biometric Technology Rally Anticipated Metrics

System Alias	Anticipated Face Failure to Acquire Rate	Anticipated Face True Identification Rate
Antero	NA	0.950
Blanca	NA	0.990
Castle	NA	0.950
Crestone	0.0003	0.991
Elbert	0.0150	0.980
Evans	0.0000	1.000
Gray	NA	NA
Harvard	NA	NA
Lincoln	NA	0.780
Massive	0.0000	0.970
Plata	0.0000	1.000

Table 4. 2018 Biometric Technology Rally Matching Results at 20 seconds

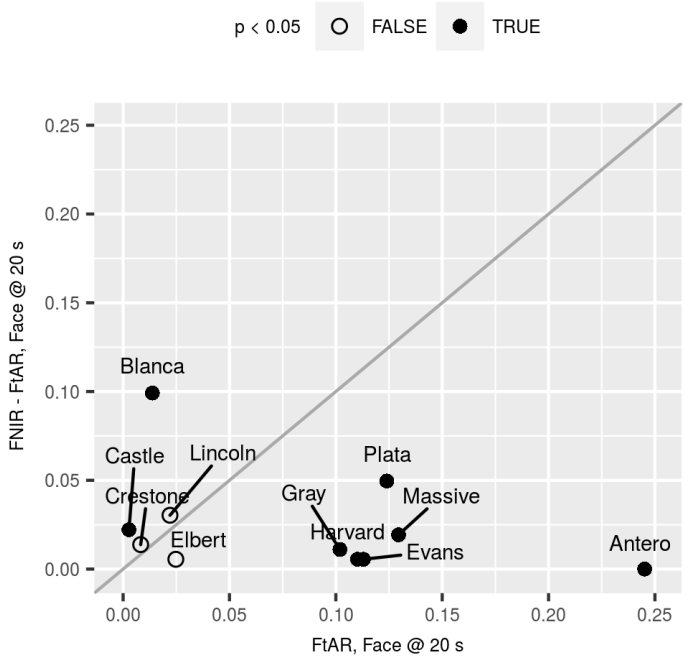
System Alias	Face FtAR	Iris FtAR	Face mTIR	Face vTIR	Iris mTIR
Antero	0.245	NA	0.755	0.457	NA
Blanca	0.014	NA	0.887	0.645	NA
Castle	0.003	NA	0.975	0.989	NA
Crestone	0.008	NA	0.978	0.970	NA
Elbert	0.025	0.127	0.970	0.763	0.862
Evans	0.113	0.113	0.882	0.879	0.840
Gray	0.102	0.132	0.887	NA	0.815
Harvard	0.110	0.140	0.884	NA	0.815
Lincoln	0.022	NA	0.948	0.915	NA
Massive	0.129	NA	0.851	0.813	NA
Plata	0.124	NA	0.826	NA	NA

- Two of nine measured TIR exceeded anticipated TIR (Castle & Lincoln)
- Had these vendor-provided, anticipated error rates been used to plan the details of an operational deployment, such as expected throughput, staffing requirements, etc., costly redesigns would have likely been required
- Our population was compliant, cooperative, undistracted, unencumbered, and paid for their efforts.

[2] Howard, et al. *An Investigation of High-Throughput Biometric Systems: Results of the 2018 Biometric Technology Rally.* BTAS 2018.

General Conclusions - FtA as a primary driver of non-identification

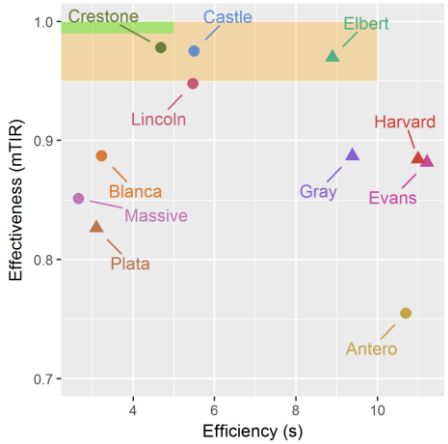
- Failure to acquire is a primary driver of error but is currently understudied by the community[2]:
  - Dominant source of error in 7 of 11 Rally Systems
  - Rally CONOP was fully transparent, well-defined, and communicated months in advance
  - Demonstrates the difficulty of the biometrics in environment defined by [1].
  - Have copious bodies of knowledge & datasets on algorithm performance (IREX, FpVTE, FRVT, FIVE, etc.)
  - Little work on system level testing
    - Moving, installing, maintaining systems is a challenge
  - Supports continued “Rally-like” efforts



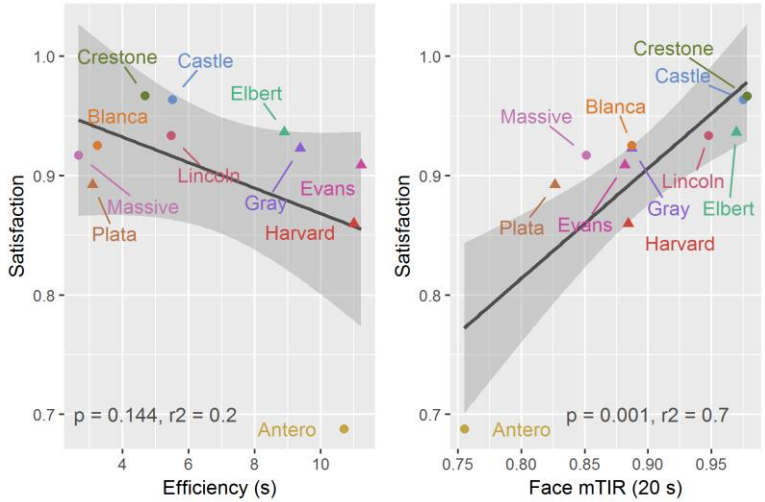
[2] Howard, et al. *An Investigation of High-Throughput Biometric Systems: Results of the 2018 Biometric Technology Rally.* BTAS 2018.

General Conclusions – Operational Tradeoffs

- There is more than one way to evaluate a given system
- System designers need to consider *relationships* in evaluation criteria[3]



- Effectiveness of systems with mid range efficiency is higher than extremes
- Capturing too quickly can lead to reduced image quality
- Linking face capture to iris capture significantly increases time (iris)



- Satisfaction strongly related to *perceived* effectiveness, not as much to efficiency (iris)
- No true effectiveness feedback in our test
- Systems that compromise effectiveness for efficiency may be less satisfying to use

[3] Hasselgren, Howard, Sirotin. *Operational Tradeoffs in the 2018 Biometric Technology Rally.* IEEE-HST 2018 (pending).

General Conclusions – Demographic Effects

- Demographic effects exist at the acquisition level as well as the matcher level:
  - For five systems, increasing skin reflectance one σ decreased transaction time by 5.5% on average [4]
  - For nine systems, increasing age one σ increased transaction time by 7.3% on average [4]
  - Impacts at the matcher level may be visible after careful statistical analysis
  - Impacts at the acquisition level may be visible in real time
  - Much more work in this area, currently only relates to face matching [4]

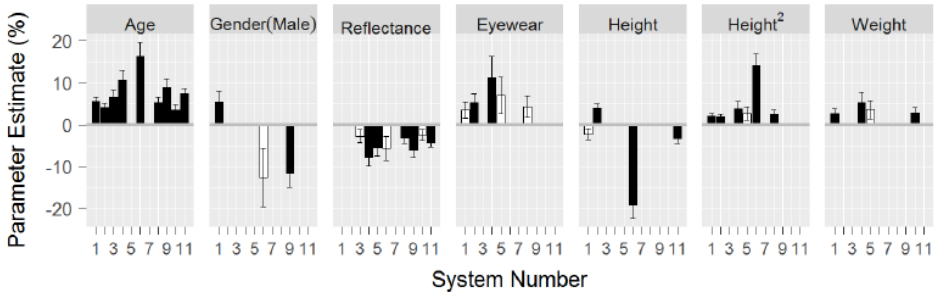


Fig. 6: Transaction Time Model Parameter Estimates for 11 Biometric Systems

[4] Cook, Howard, Sirotin. *Effects of User Demographics on the Performance of Eleven Commercial Biometric Systems. IEEE BIOSIG* 2018 (pending).

- Questions?
- Thank you
- For more information:
  - arun.vemury@hq.dhs.gov
  - peoplescreening@hq.dhs.gov
  - jerry.l.tipton@saic.com
  - john@mdtf.org
- http://mdtf.org
  - All 2018 Biometric Technology Rally Results
  - Material from all webinars, briefings, outreach
  - Announcements about future Rally's