DHS SCIENCE AND TECHNOLOGY

Evaluation of Rapid Face Capture Devices: Results of the 2018 Biometric Technology Rally

International Face Performance Conference, November 2018

Homeland Security
Science and Technology

MdTF

John J. Howard, Ph.D.
Principal Data Scientist,
SAIC Identity and Data Sciences Laboratory,
Maryland Test Facility

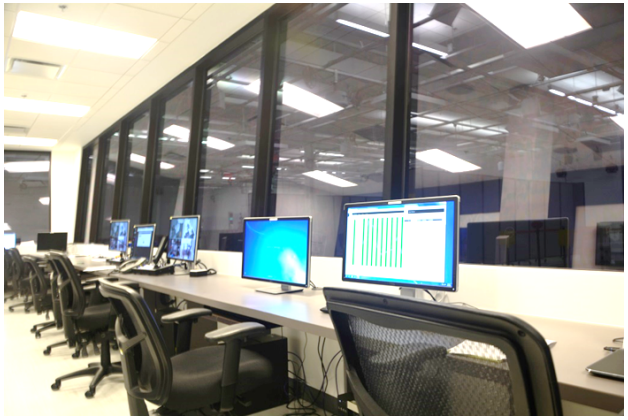
Arun Vemury
Biometrics and Identity Technology Engine,
Department of Homeland Security,
Science and Technology Directorate

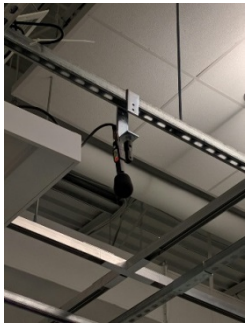
Outline

- The Maryland Test Facility
- The 2018 Biometric Technology Rally
  - Motivation
  - Timeline & Process
  - Metrics
- Rally Results
  - Consumer reports
  - Efficiency, Satisfaction, Effectiveness
- Conclusions
  - Industry expectations
  - Primary error determinants
  - High throughput systems
  - High throughput metrics
  - Acquisition system choice
  - 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

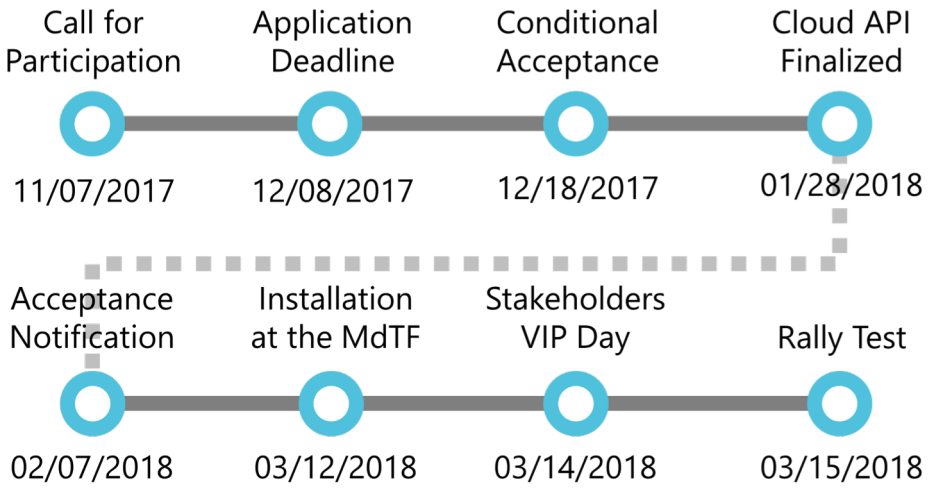




2018 Biometric Technology Rally – Motivation and Goals

- Multiple components within DHS collect and match biometric information during day-to-day operations
  - New biometric technologies and collection methods are being considered for DHS processes, especially in the travel environment
  - Some commercial biometric systems show undesirable rates of failure in the field, in part driven by failure to acquire images
  - Selecting the wrong technology carries significant risk of the system failing to meet performance expectations
- Goals of the 2018 Rally:
  - Formalize the “high-throughput” use case
  - Obtain a fair assessment of the state of the industry in regards to efficiency, effectiveness, and satisfaction
  - 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

Rally Timeline and Systems



7 Months to design and execute the rally from start to finish (9/17 – 3/18)

Developed 11 comprehensive performance metrics

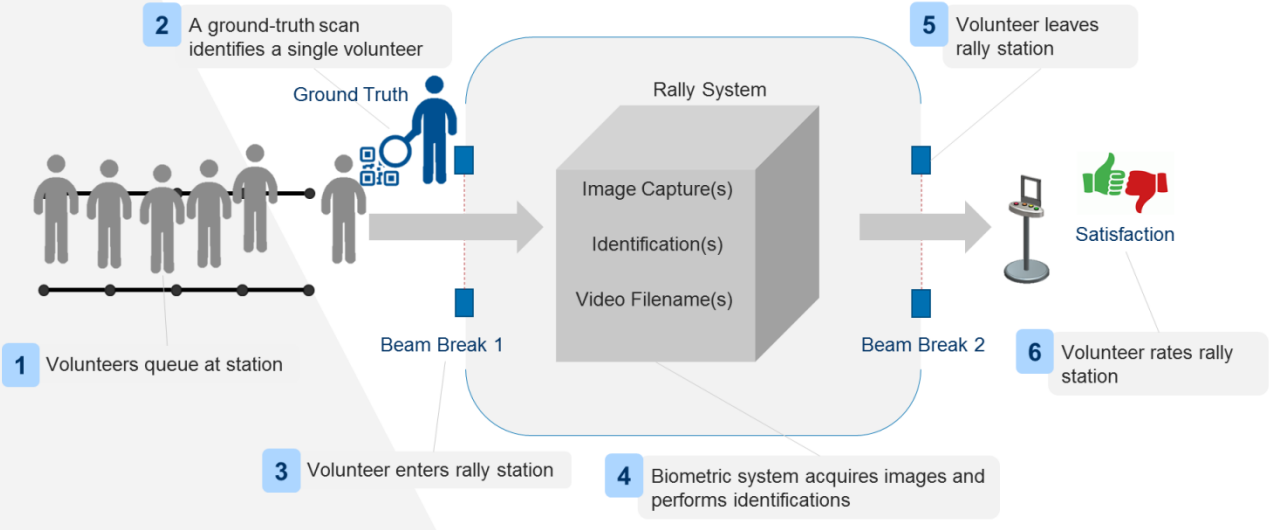
Developed cloud API so participants could start integrating remotely

Rally Systems:

- Required:
  - Collect 1 Face
  - Fit in a 7x8 ft. space
  - Be unmanned
  - Direct all interaction
  - Take on average 10 seconds per person
- Optional:
  - Collect 3 Faces
  - Collect 3 Irises
  - Provide Facial Identifications
  - Collect Video

- Timeline:
  - Announced in November 2017
  - One month for applications.
  - 19 applications, 12 selected
  - Cloud hosted API in January
  - Test in March
  - Results live in May

2018 Biometric Technology Rally Test Process

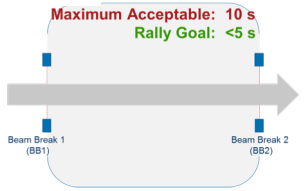


- Eleven systems, two day install
- Rally Gallery of 525 unique people, 1848 images total – to support onboard identifications
- 363 diverse subjects, groups of 15, over 5 day period
- General instructions were provided
- Enrollment by a trained operator
- All subjects interacted with all systems in a counterbalanced manner

Rally Metrics

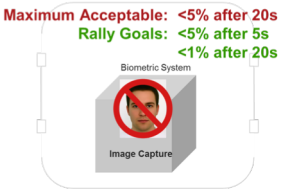
- Efficiency
  - Refers to the amount of time required to use each biometric system
  - Quantified as average transaction time (beam-break to beam-break) for Test Volunteers at each Rally System
- Effectiveness
  - Refers to the accuracy and completeness with which users are identified.
  - Measured 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

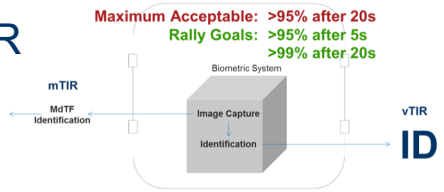


Effectiveness:

FtAR



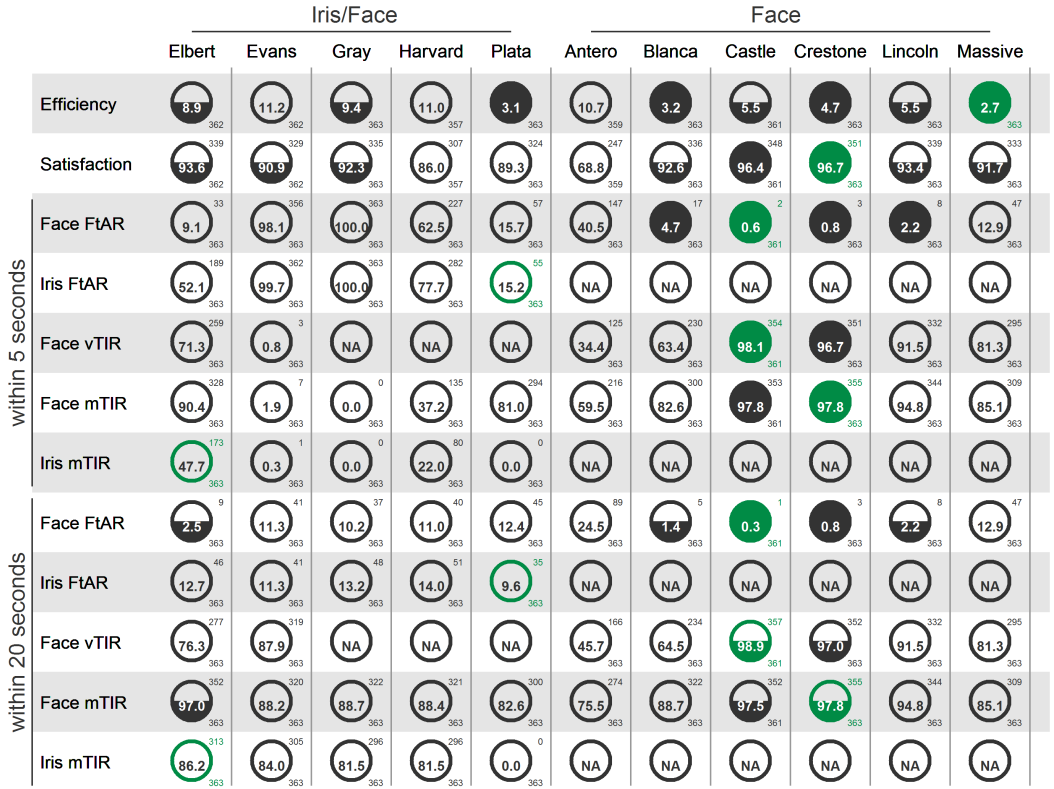
TIR



Satisfaction

Rally Results – Consumer Report

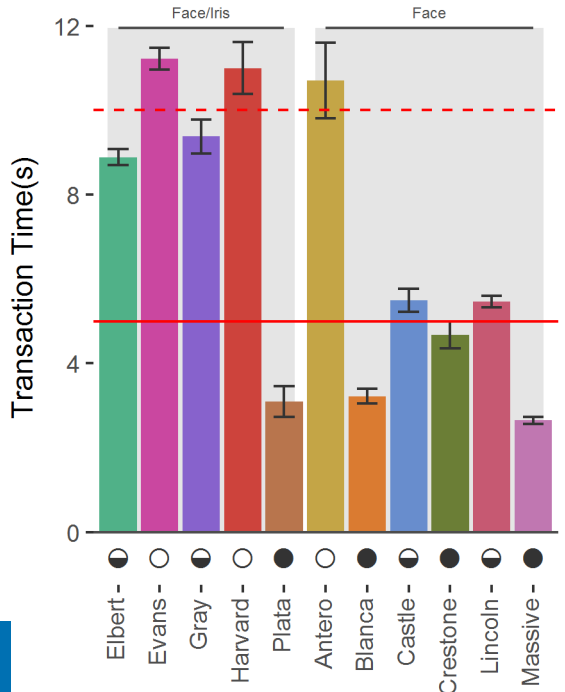
http://mdtf.org



This figure presents a summary of the performance of the participating face and face/iris systems, plotting the code name for each Rally Participant as column headers and each rally metric as the row headers. Circles show the value for each metric. The units are seconds for efficiency and are proportions for all other metrics. Circles are coded are as follows: ○ - below rally threshold; ◒ - below rally goal; ● - meets or exceeds rally goal. The number to the lower right of each circle is the denominator and the number on the top right of each circle is the numerator for the proportion.

Rally Results - Efficiency

- Average Transaction Time
- Most efficient:
  - **Massive – 2.65 seconds on average**
- Met the Goal (4):
  - Transaction time < 5 seconds
  - Massive, Plata, Blanca, and Crestone
- Met the Threshold (4):
  - Transaction time < 10 seconds
  - Lincoln, Castle, Elbert, and Gray



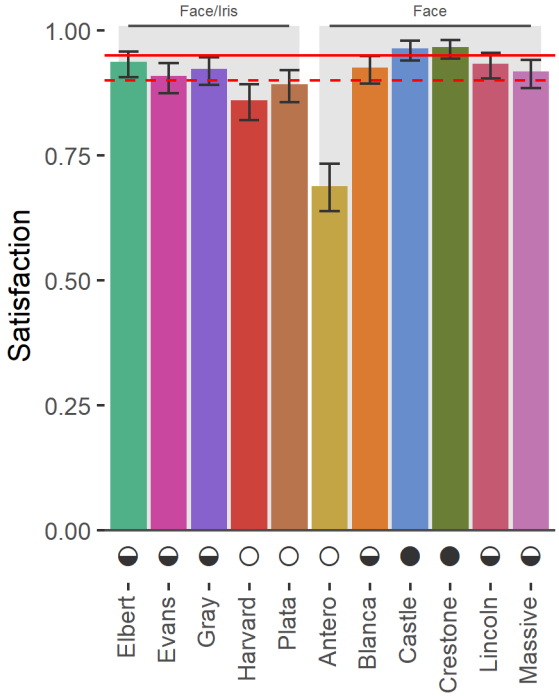
Most Rally systems were fast

Face - 10 seconds is enough, 5 seconds is possible

Iris – 10 seconds is possible

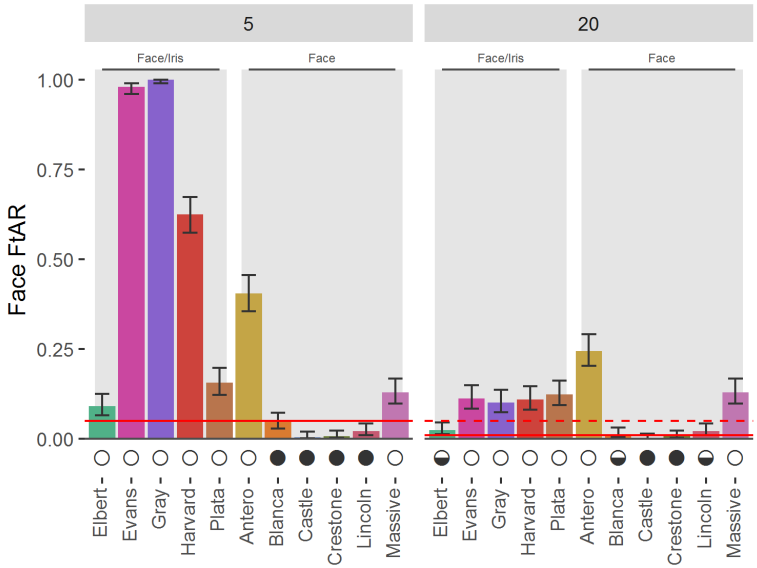
Rally Results - Satisfaction

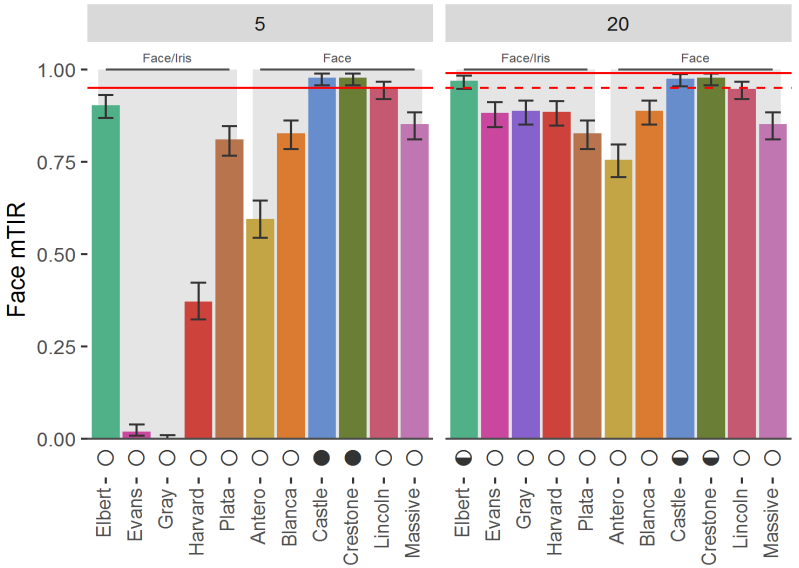
- Proportion of positive responses (“Happy” or “Very Happy”)
- Most Satisfying:
  - **Crestone – 96.7% Happy or Very Happy**
- Met the Goal (2):
  - Satisfaction > 95%
  - Castle and Crestone
- Met the Threshold (6):
  - Satisfaction > 90%
  - Elbert, Evans, Gray, Blanca, Lincoln, and Massive



Most people are happy using these biometric systems

Rally Results – Acquisition and Matching





- 4 systems met 5 second goal (<5%)
- 2 systems met the 20 second goal (<1%)
- Lowest Face FtAR:
  - **Castle – 0.6% by 5 sec, 0.3% by 20 sec**

- 3 systems met the 5 second goal (> 95%)
- 0 systems met the 20 second goal (> 99%)
- Highest face mTIR:
  - **Crestone – 97.8% by 5 sec, 97.8% by 20 sec**

General Conclusions – High-throughput biometric systems[1]

- "Defined" a new high-throughput (HT) biometric use case:
  - 1000s of users
  - Short time frames
  - Unmanned
- Tested 11 commercial biometric systems
- All failed to meet the goal of a 99% TIR
- Context of 1% failure rate in the HT environment - dozens of exception cases
- Modern IT systems that handle similar volumes measure reliability in the far fraction of a percent (99.99..% uptime).
- Minority in the 95% range, majority in the 70-80 % range
- Points to the challenge of HT environment
- Need for improved system design and the ability to handle non-optimal user behavior.

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

General Conclusions – Industry Expectations

- FtAR and TIR results were not well anticipated by industry[1]:
  - 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

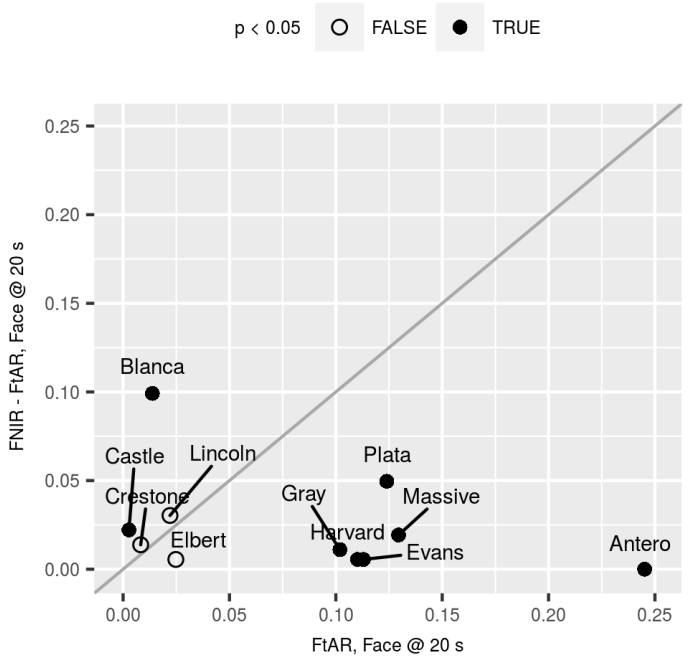
System Alias	Face FtAR	Iris FtAR	Face mTIR	Face vTIR	Iris mTIR
Antero	0.245	NA	0.755	0.457	NA
Blanca	0.014	NA	0.887	0.645	NA
Castle	0.003	NA	0.975	0.989	NA
Crestone	0.008	NA	0.978	0.970	NA
Elbert	0.025	0.127	0.970	0.763	0.862
Evans	0.113	0.113	0.882	0.879	0.840
Gray	0.102	0.132	0.887	NA	0.815
Harvard	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.

[1] 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[1]:
  - 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



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

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[2]:

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

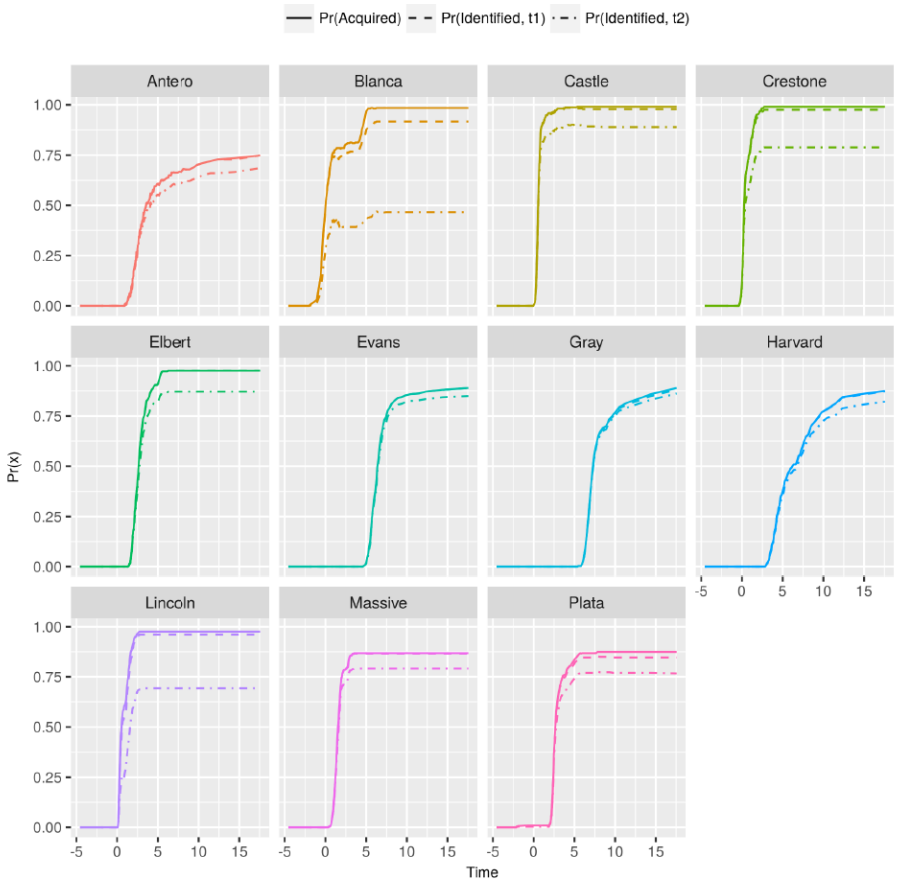
H.T. Systems need unique workflows[2]:

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.

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

General Conclusion – High Throughput Metrics

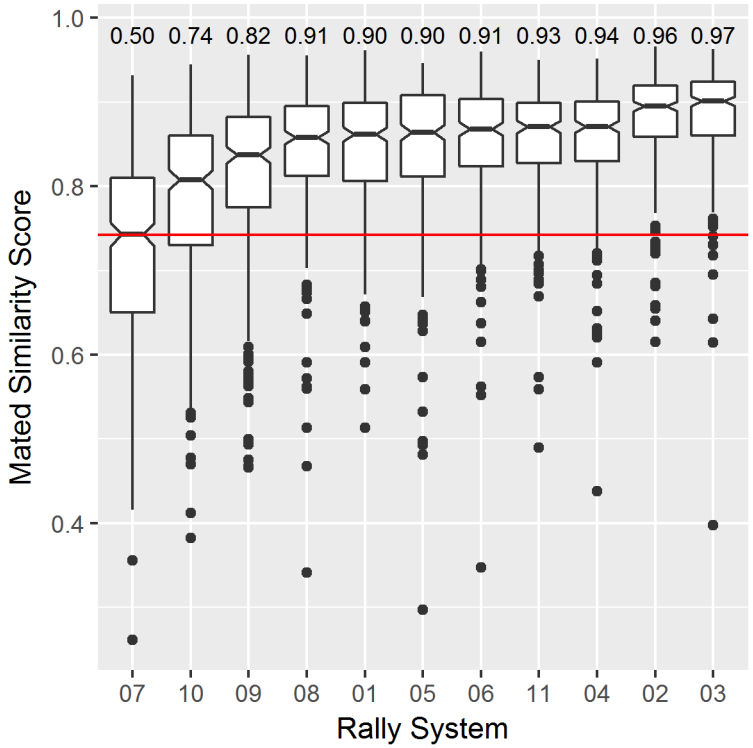
- High-throughput systems may need different kinds of metrics for proper evaluation
  - **ISO 19795-1, 8.2.2.3 "**The failure-to-acquire rate will depend on thresholds for sample quality, as well as the allowed duration for sample acquisition or allowed number of presentations. These settings shall be 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[2]



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

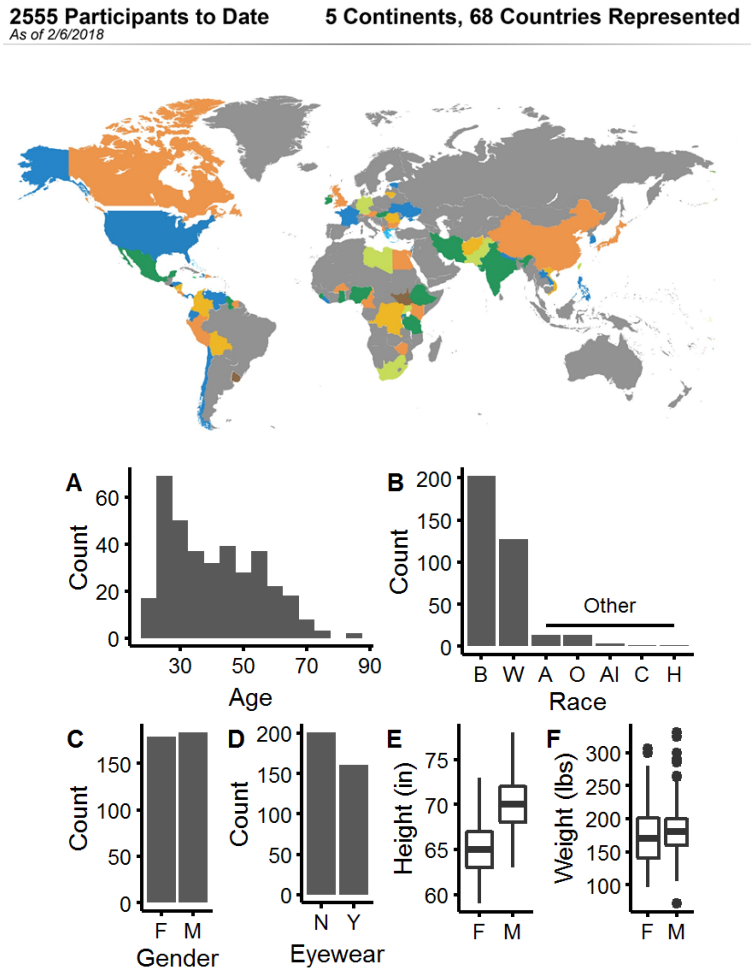
General Conclusions – Acquisition Camera Matters

- mTIR computed using common algorithm for all camera systems
- Mated similarity score distributions varied significantly across the 11 rally systems
- Proportion of mated similarity scores above a fixed threshold (0.74, red line) varied from **50% to 97%** depending on camera
- Choice of camera system will significantly affect biometric performance independent of algorithm

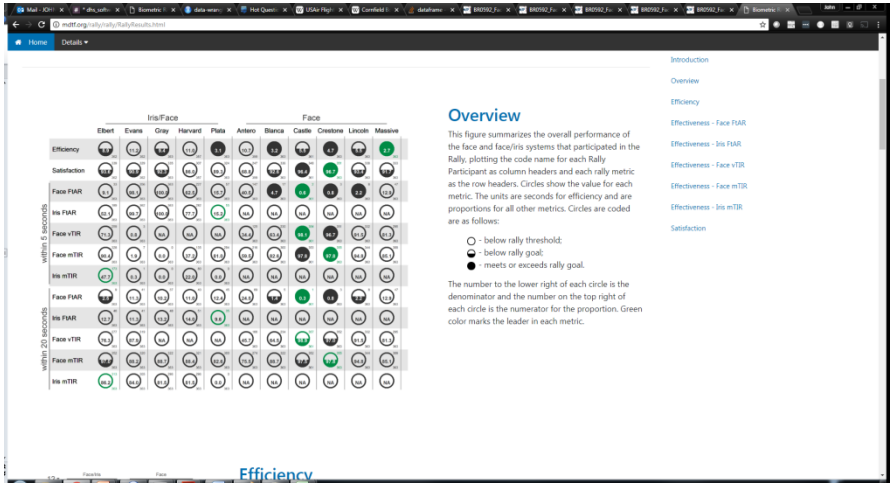


General Conclusions – Demographic Effects

- MdTF test populations are designed to mimic travelling public
- Have been collecting ~5 years
- Collect controlled, manned, enrollment images and self reported demographics
- Allows for investigations of these covariates on:
  - Capture speed
  - Match performance
  - Longitudinal analysis



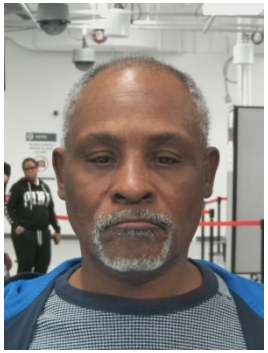
- We're hiring!
- 2019 Rally announced November 11, 2018
- For more information:
  - john@mdtf.org
  - arun.vemury@hq.dhs.gov
- http://mdtf.org
  - All 2018 Biometric Technology Rally Results
  - Information on 2019 Rally, applications due 11/30
- Thank you
- Questions?



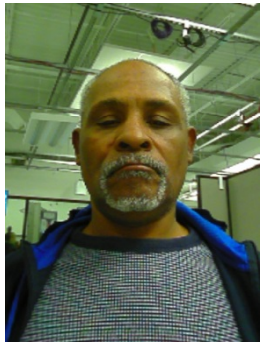
General Conclusions – Iris Systems as a Face Capture Device

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

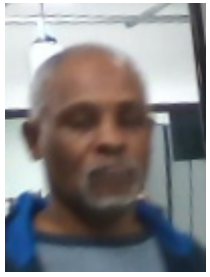
Gray



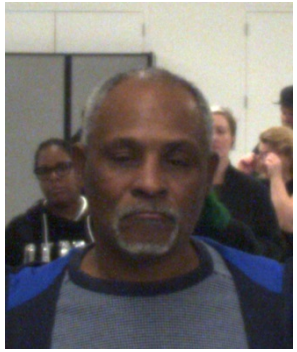
Harvard



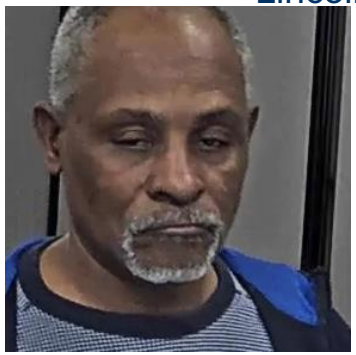
Blanca



Crestone



Lincoln

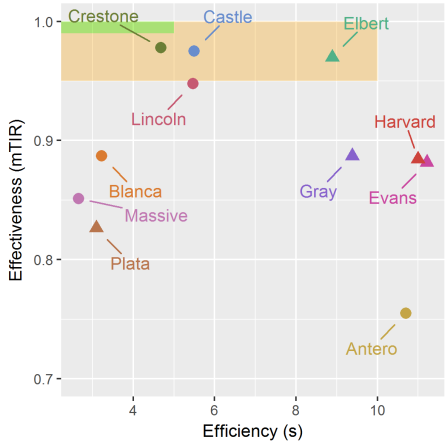


Station	In Gallery ID Rate
Castle	0.9876543
Crestone	0.9785276
Elbert	0.9754601
Lincoln	0.9601227
Blanca	0.9171779
Harvard	0.9110429
Gray	0.8957055
Evans	0.8865031
Massive	0.8650307
Plata	0.8466258
Antero	0.7760736

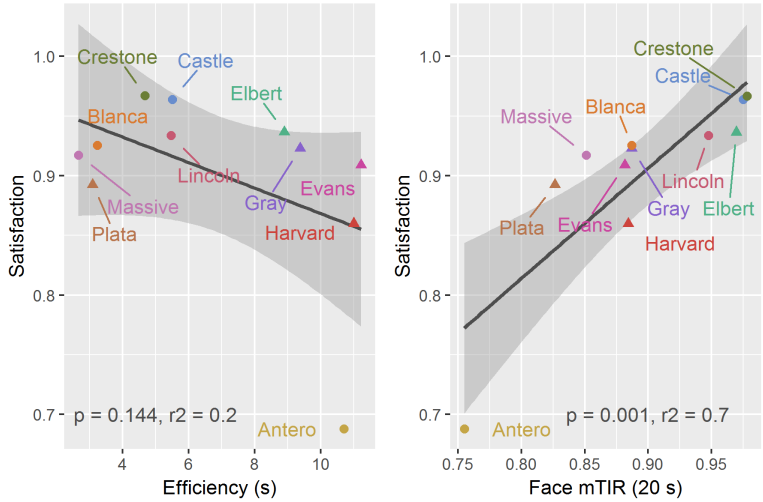
Station	In Gallery ID Rate 2
Castle	0.9444444
Elbert	0.9079755
Gray	0.8865031
Harvard	0.8773006
Evans	0.8680982
Crestone	0.8619632
Massive	0.8312883
Plata	0.8006135
Lincoln	0.7791411
Antero	0.7361963
Blanca	0.601227

General Conclusions – 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).