Video Surveillance Event Detection Track

The TRECVID 2009 Evaluation

Jonathan Fiscus, Martial Michel,
John Garofolo, Paul Over
NIST

Heather Simpson, Stephanie Strassel
LDC

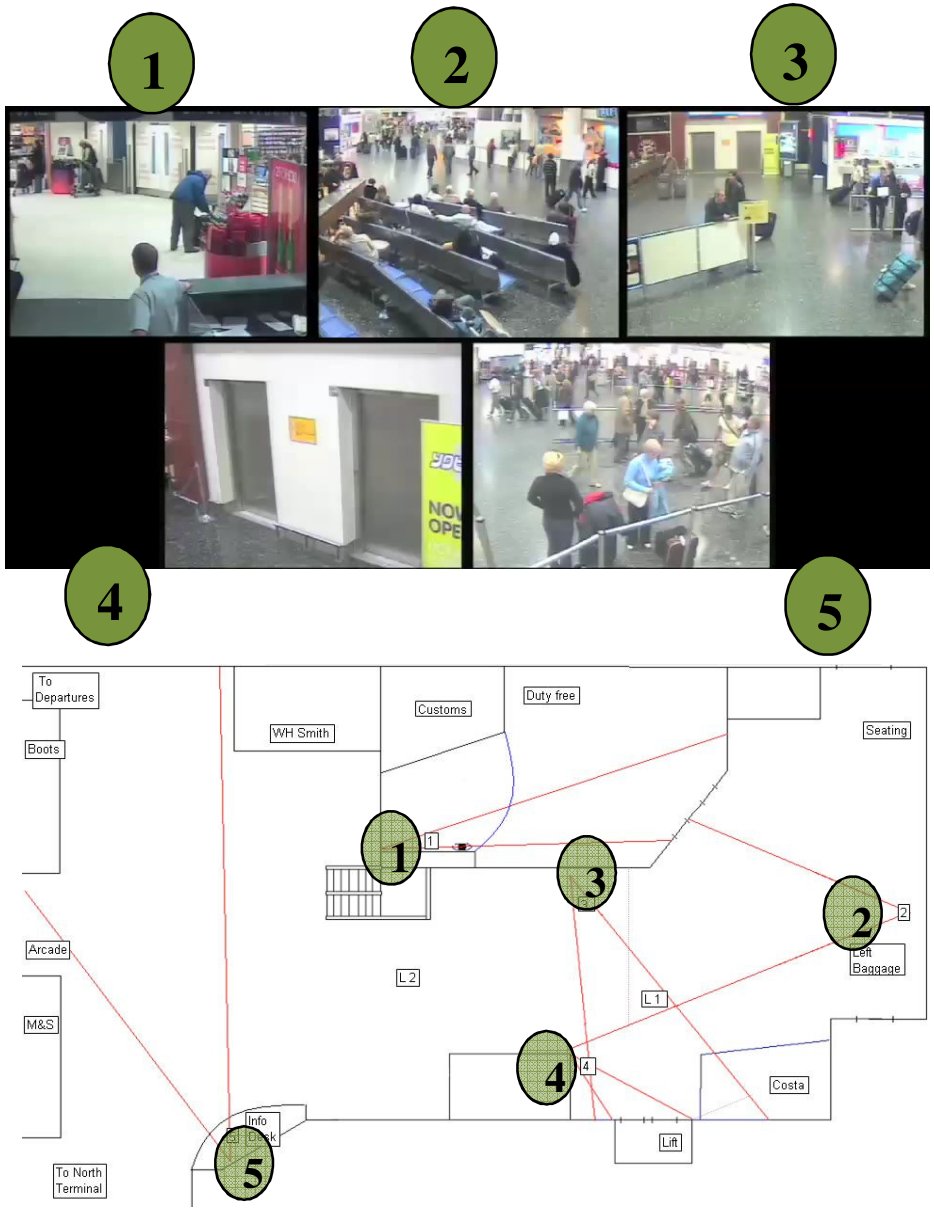
iLIDS® Imagery Library for Intelligent Detection Systems
Home Office Scientific Development Branch
CPNI Centre for the Protection of National Infrastructure
Homeland Security Science and Technology Directorate
VACE Video Analysis Content Extraction
NIST National Institute of Standards and Technology

Motivation

- **Problem:** automatic detection of *observable* events of interest in surveillance video
- **Challenges:**
  - requires application of several Computer Vision techniques
    - segmentation, person detection/tracking, object recognition, feature extraction, etc.
  - involves subtleties that are readily understood by humans, difficult to encode for machine learning approaches
  - can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc.

Evaluation Source Data

- UK Home Office collected CCTV video at a busy airport
  - 5 Camera views: (1) controlled access door, (2) waiting area, (3) debarkation area, (4) elevator close-up, (5) transit area
- Development data resources:
  - 100 camera hours of video from the 2008 VSED Track
    - Complete annotation for 10 events on 100% of the data
- Evaluation data resources:
  - 45 camera hours of video from the iLIDS Multiple Camera Tracking Scenario Training data set
    - Complete annotation for 10 events annotated on 1/3 of the data
    - Also used for the AVSS 2009 Single Person Tracking Evaluation



TRECVID VSED
Retrospective Event Detection

- Task:
  - Given a textual description of an ***observable event of interest*** in the airport surveillance domain, configure a system to detect all occurrences of the event
  - Identify each event observation by:
    - The ***temporal extent***
    - A ***detection score*** indicating the system's confidence that the event occurred
    - A ***binary decision*** on the detection score optimizing performance for the primary metric

TRECVID
VSED Freestyle Analysis

- Goal is to support innovation in ways not anticipated by the retrospective task
- Freestyle task includes:
  - rationale
  - clear definition of the task
  - performance measures
  - reference annotations
  - baseline system implementation

Event Annotation Guidelines

- Jointly developed by NIST, Linguistic Data Consortium (LDC), Computer Vision Community
  - Event Definitions left minimal to capture human intuitions
- Updates from 2008 guidelines :
  - Based on annotation questions from 2008 annotation
  - End Time Rule :
    - If Event End Time = a person exiting the frame boundary, frame for end time should be the earliest frame when their body and any objects they are carrying (e.g. rolling luggage) have passed out of the frame. If luggage remains in the frame not moving, can assume person left the luggage and tag at person leaving the frame.
  - People Meet/Split Up rules:
    - If people leave a group but do not leave the frame, the re-merging of those people do not qualify as PeopleMeet
    - If a group is standing near the edge of the frame, people are briefly occluded by frame boundary but under RI rule have not left the group, that is not PeopleSplitUp
  - Some specific case examples added to Annotator guidelines

Annotation Tool and Data Processing

- No changes from 2008
  - Annotation Tool
    - ViPER GT, developed by UMD (now AMA)
    - http://viper-toolkit.sourceforge.net/
    - NIST and LDC adapted tool for workflow system compatibility
  - Data Pre-processing
    - OS limitations required conversion from MPEG to JPEG
      - 1 JPEG image for each frame
    - For each video clip assigned to annotators
      - Divided JPEGs into framespan directories
      - Created .info file specifying order of JPEGs
      - Created ViPER XML file (XGTF) with pointer to .info file
    - Default ViPER playback rate = about 25 frames (JPEGs)/second

Annotation Workflow Design

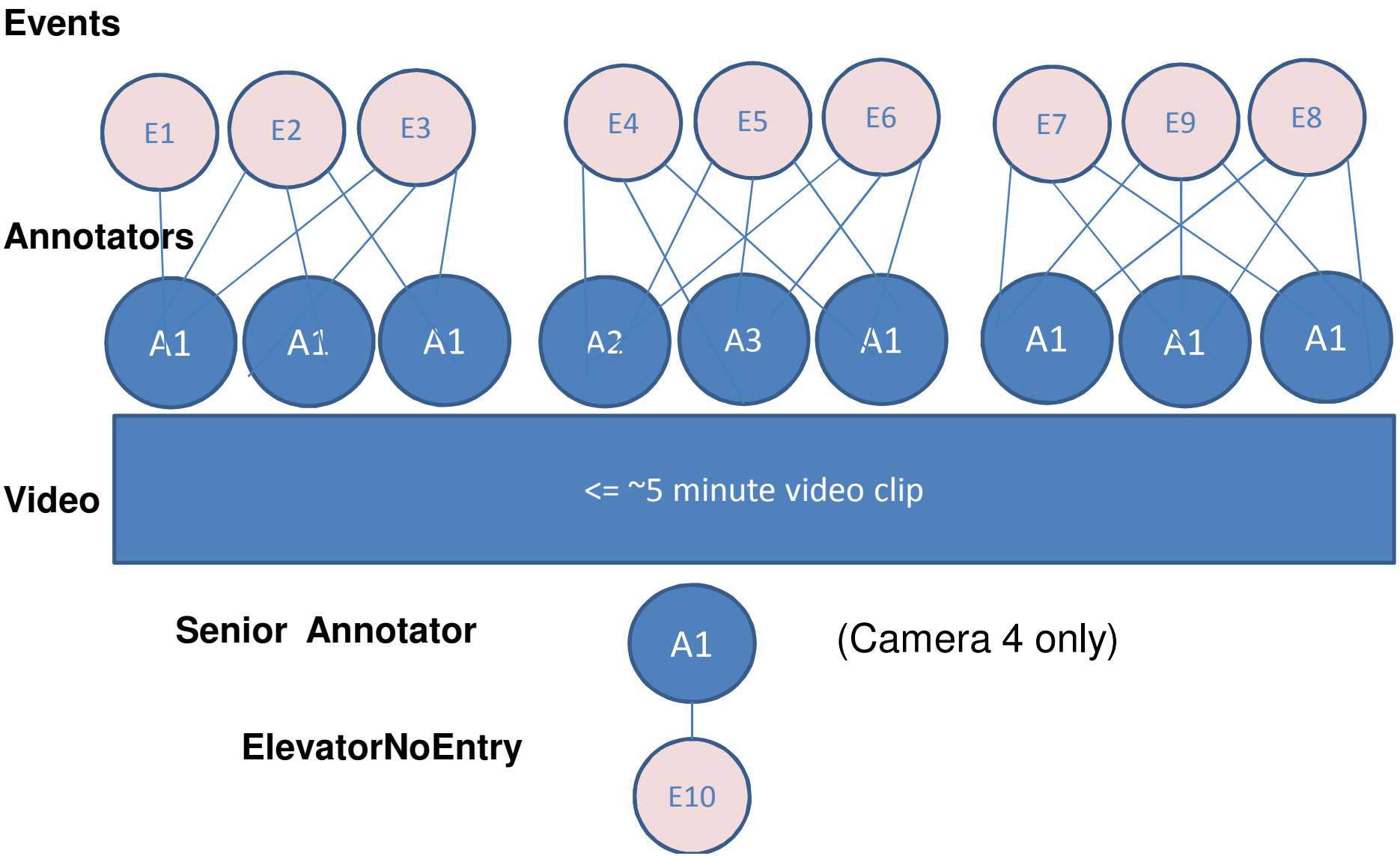
- Clip duration about same or smaller than 2008
- Rest of workflow revised based on 2008 annotations and experiments
  - 3 events per work session for 9 events
  - 1 pass by senior annotator over ElevatorNoEntry for Camera 4 only
    - ElevatorNoEntry very infrequent, only 1 set of elevators which are easy to see in Camera 4 view
    - Camera 4 ElevatorNoEntry annotations automatically matched to corresponding timeframe in other camera views
  - 3 passes over other 9 events for 14 hours of video
    - (2008 – 1 pass over all 10 events for 100 hours of video)
  - Additional 6 passes over 3 hour subset of video
- Adjudication performed on 3x and 9x annotations
  - 2008 Adjudication performed on system + human

Event Sets

- 3 sets of 3 events, ElevatorNoEntry separate set
- Goal to balance sets by event type and frequency

Event Type	Tracking	Object	Gesture
Set 1	OpposingFlow	CellToEar	Pointing
Set 2	PeopleSplitUp	ObjectPut	Embrace
Set 3	PeopleMeet	TakePicture	PersonRuns

Visualization of Annotation Workflow



Annotation Challenges

- Ambiguity of guidelines
  - Loosely defined guidelines tap into human intuition instead of forcing real world into artificial categories
  - But human intuitions often differ on borderline cases
  - Lack of specification can also lead to incorrect interpretation
    - Too broad (e.g. baby as object in ObjectPut)
    - Too strict (e.g. person walking ahead of group as PeopleSplitUp)
- Ambiguity and complexity of data
  - Video quality leads to missed events and ambiguous event instances
    - Gesturing or pointing? ObjectPut or picking up an object? CellToEar or fixing hair?
- Human factors
  - Annotator fatigue a real issue for this task
  - Lower number of events per work session helps
- Technical issues

2009 Participants

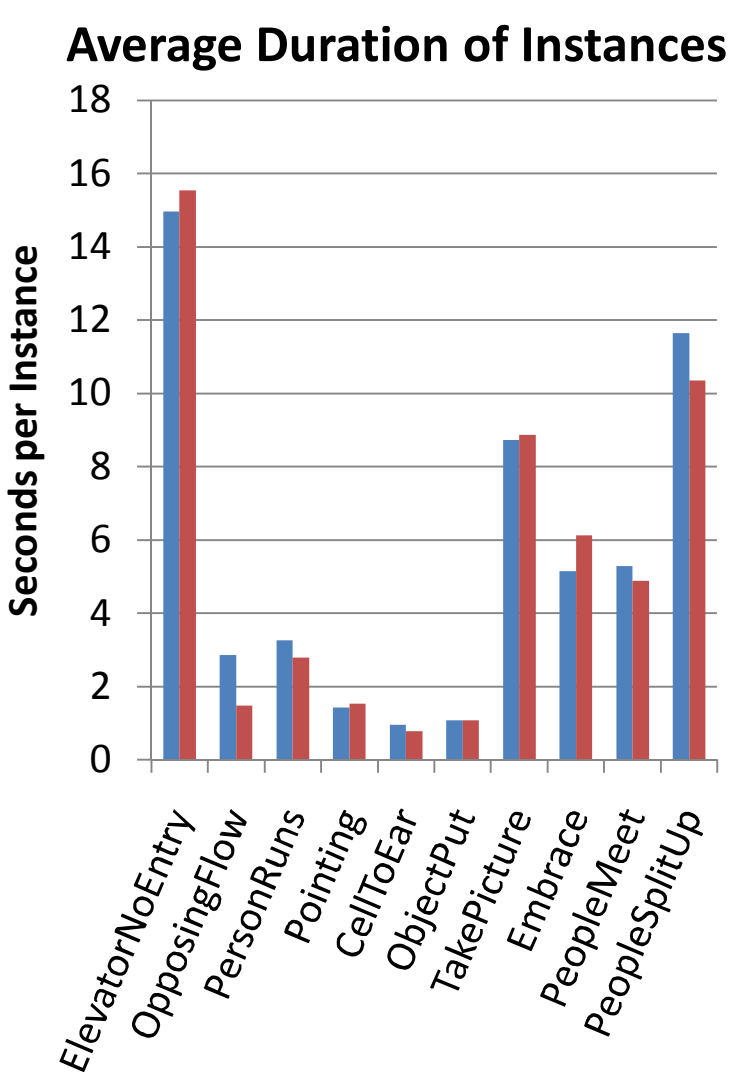
11 Sites (45 registered participants)

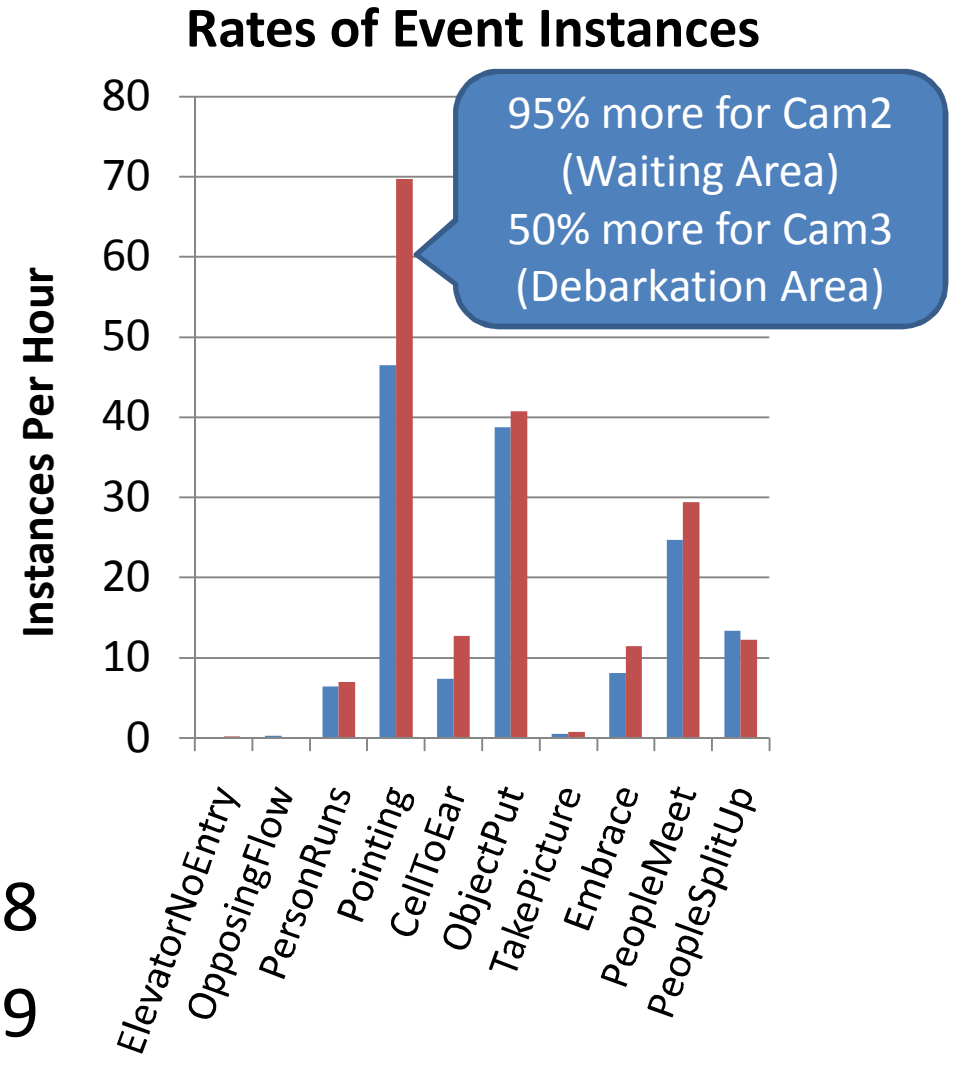
75 Event Runs

			Single Person				Single Person + Object			Multiple People		
			ElevatorNoEntry	OpposingFlow	PersonRuns	Pointing	CellToEar	ObjectPut	TakePicture	Embrace	PeopleMeet	PeopleSplitUp
2008	Shanghai Jiao Tong University	SJTU	x	x	x	x			x	x	x	x
	Universidad Autónoma de Madrid	UAM		x	x			x				
	Carnegie Mellon University	CMU	x	x	x	x	x	x	x	x	x	x
	NEC Corporation/University of Illinois at Urbana-Champaign	NEC-UIUC			x	x	x	x		x		
	NHK Science and Technical Research Laboratories	NHKSTRL		x	x			x			x	
	Beijing University of Posts and Telecommunications (MCPRL)	BUPT-MCPRL	x	x	x	x			x			
	Beijing University of Posts and Telecommunications (PRIS)	BUPT-PRIS	x	x	x							
	Peking University (+ IDM)	PKU-IDM	x		x					x	x	x
New	Simon Fraser University	SFU			x	x				x		
	Tokyo Institute of Technology	TITGT			x						x	x
	Toshiba Corporation	Toshiba	x	x	x							
	Total Participants per Event		6	7	11	5	2	4	3	5	5	4

Observation Durations and Event Densities

Comparing 2008 and 2009 Test Sets

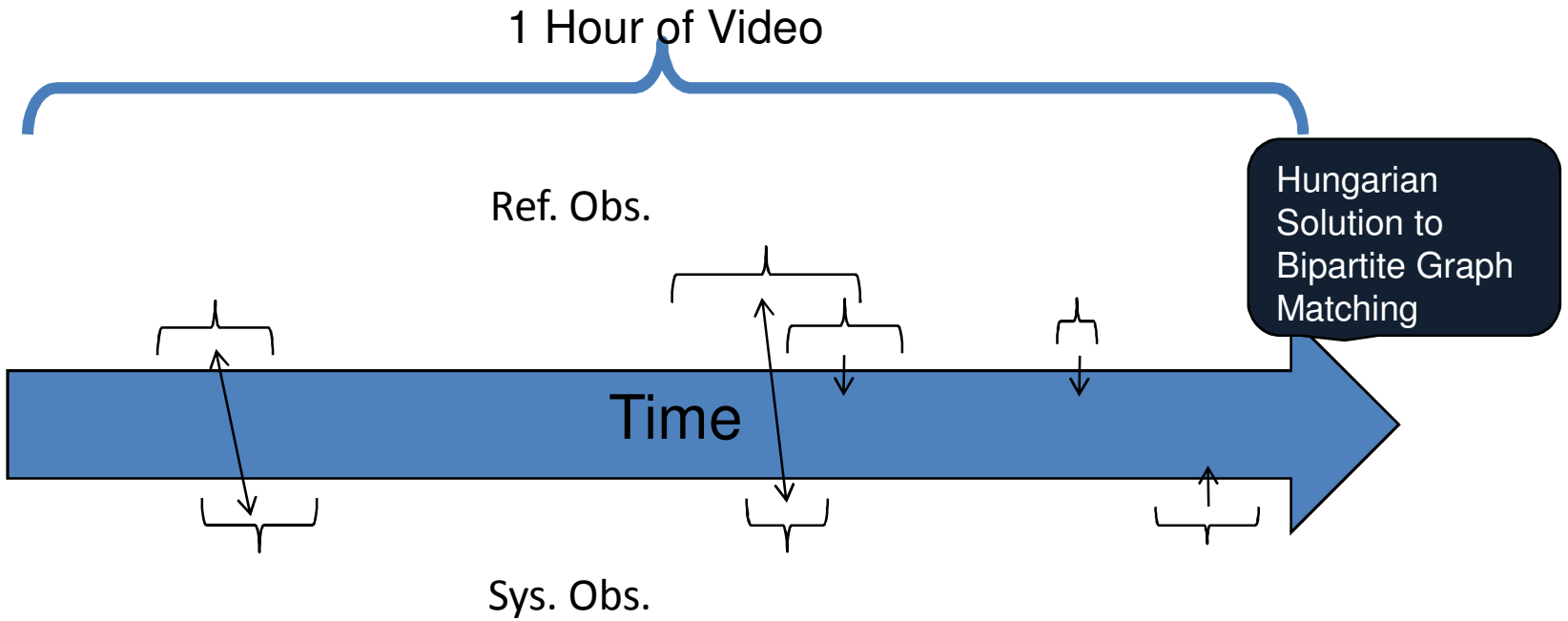




Evaluation Protocol Synopsis

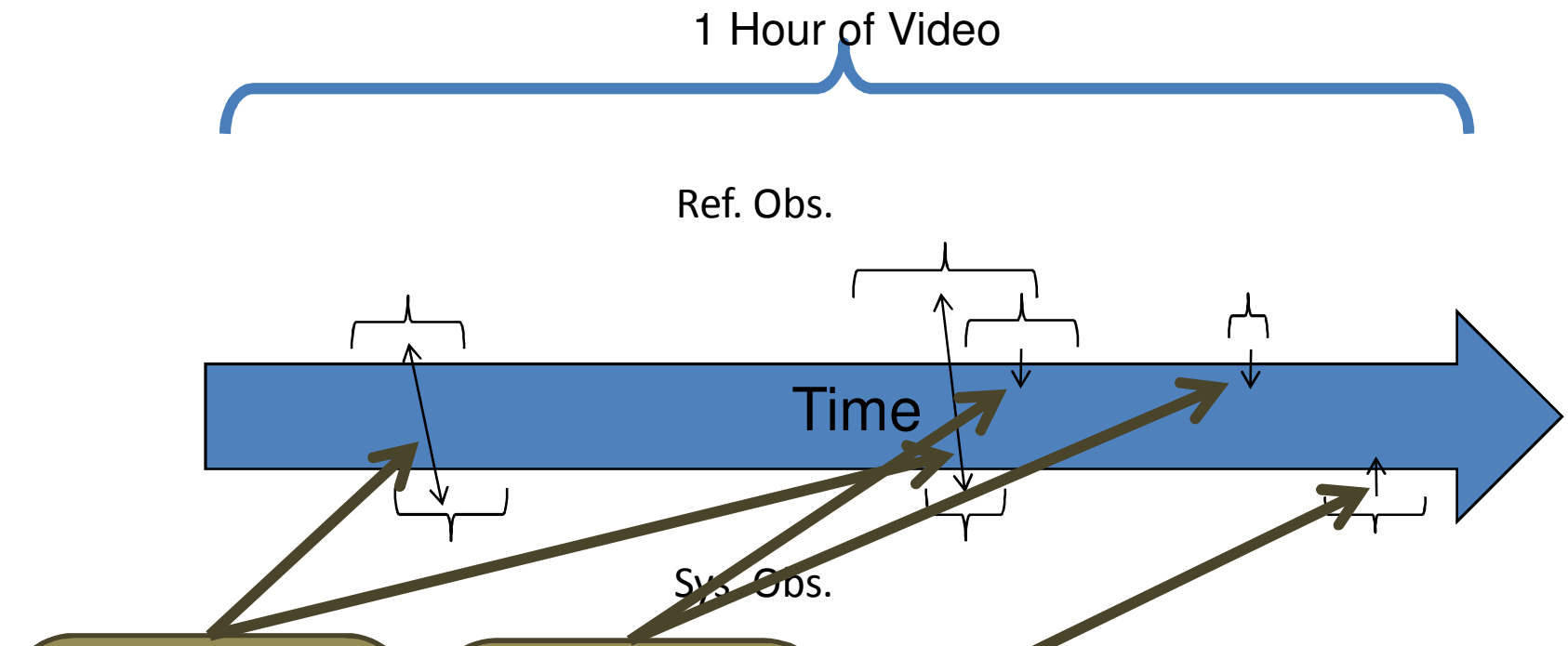
- NIST used the Framework for Detection Evaluation (F4DE) Toolkit
  - Available for download on the VSED Web Site
  - http://www.itl.nist.gov/iad/mig/tools
- Events are scored independently
- Five step evaluation process
  - Segment mapping
  - Segmented scoring
  - Score accumulation
  - Error metric calculation
  - Error visualization

Segment Mapping for Streaming Media



- Mapping kernel function
  - The mid point of the system-generated extent must be within the reference extent extended by 1 sec.
  - Temporal congruence and decision scores give preference to overlapping events

Segment Scoring



Correct Detections

When reference and system observations are mapped

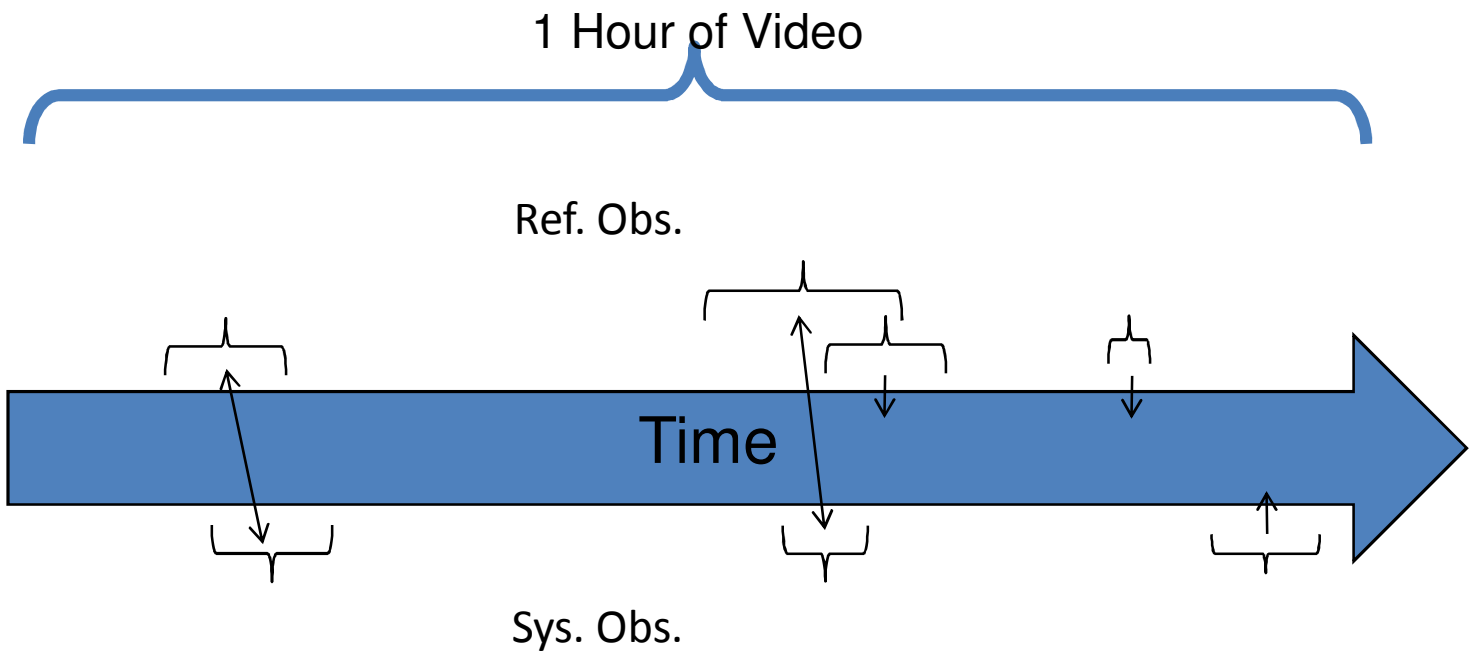
Missed Detections

When a reference observation is NOT mapped

False Alarms

When a system observation is NOT mapped

Compute Normalized Detection Cost



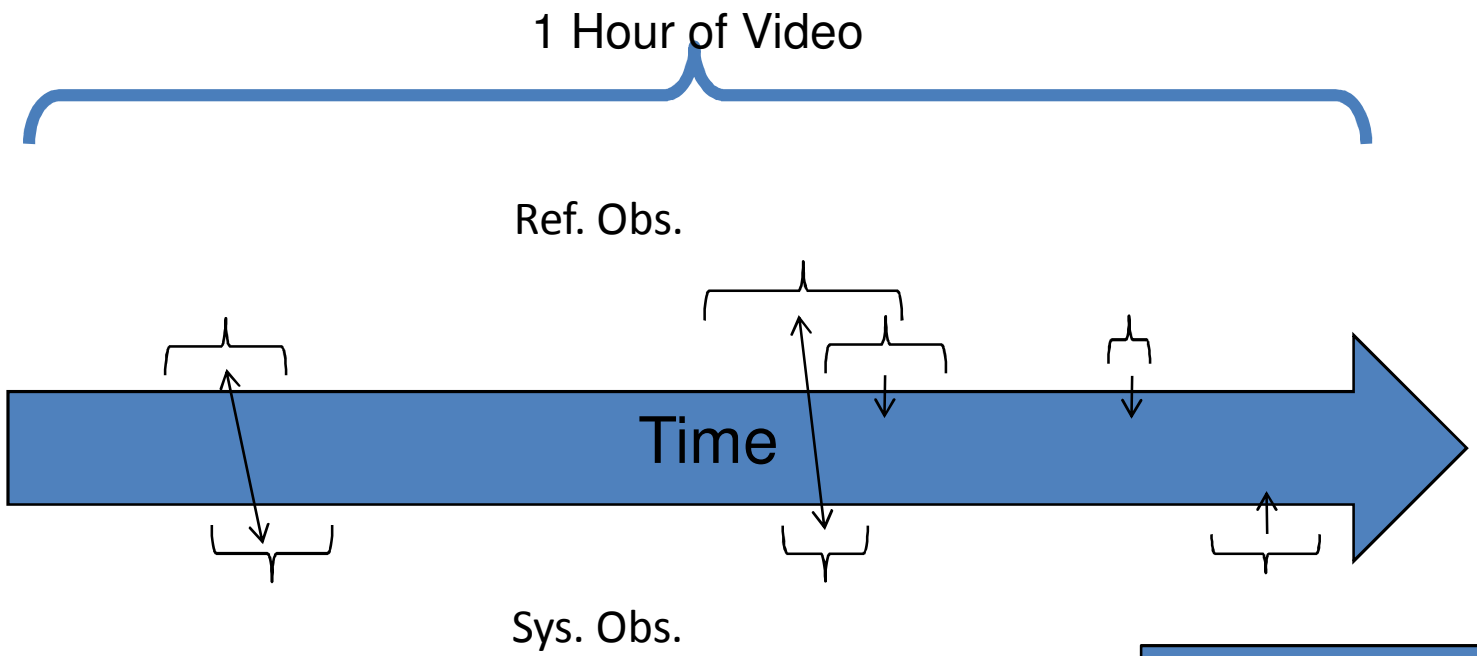
$$P_{Miss}() = \frac{\#MissedObs}{\#TrueObs}$$

$$P_{Miss}() = \frac{2}{4} = .50$$

$$Rate_{FA}() = \frac{\#FalseAlarms}{SignalDuration}$$

$$Rate_{FA}() = \frac{1}{1Hr} = 1FA/Hr$$

Compute Normalized Detection Cost Rate



$$NDCR() = P_{Miss}() + \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} * R_{FA}()$$

$$NDCR() = 0.5 + \frac{1}{10 * 20} * 1 = .505$$

Range of NDCR() is [0:∞)
NDCR() = 1.0 is a system that outputs nothing

Event Detection Constants

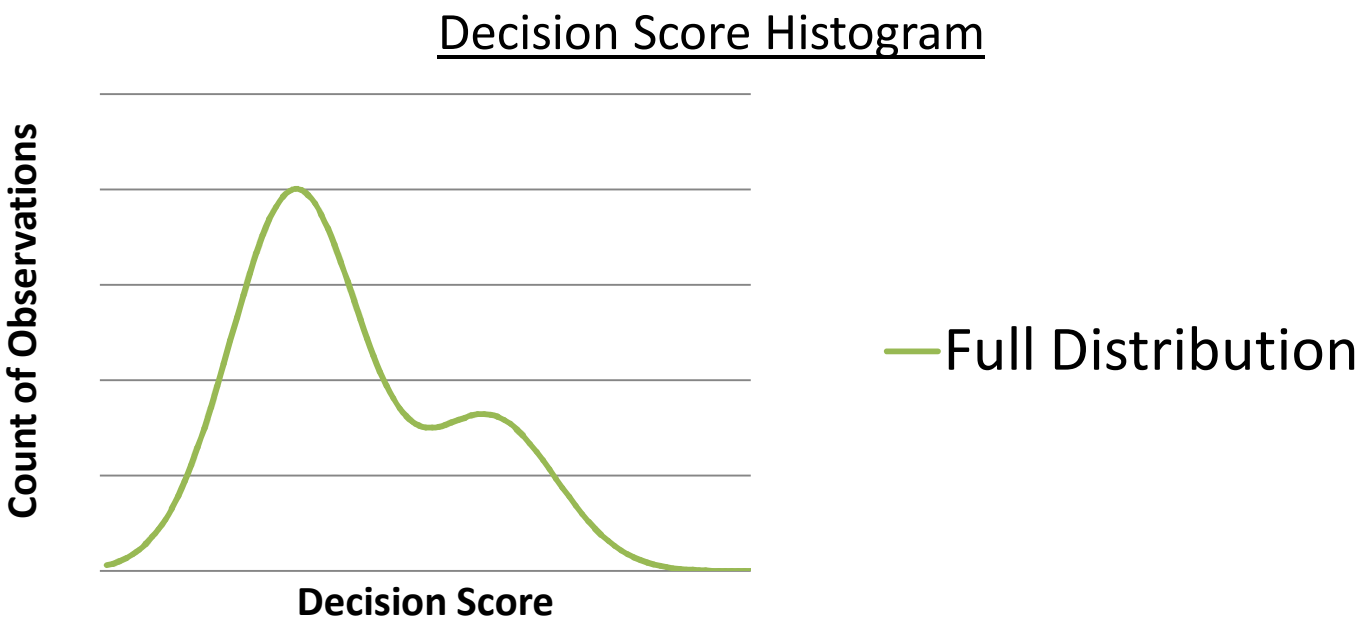
$Cost_{Miss} = 10$

$Cost_{FA} = 1$

$R_{Target} = 20$

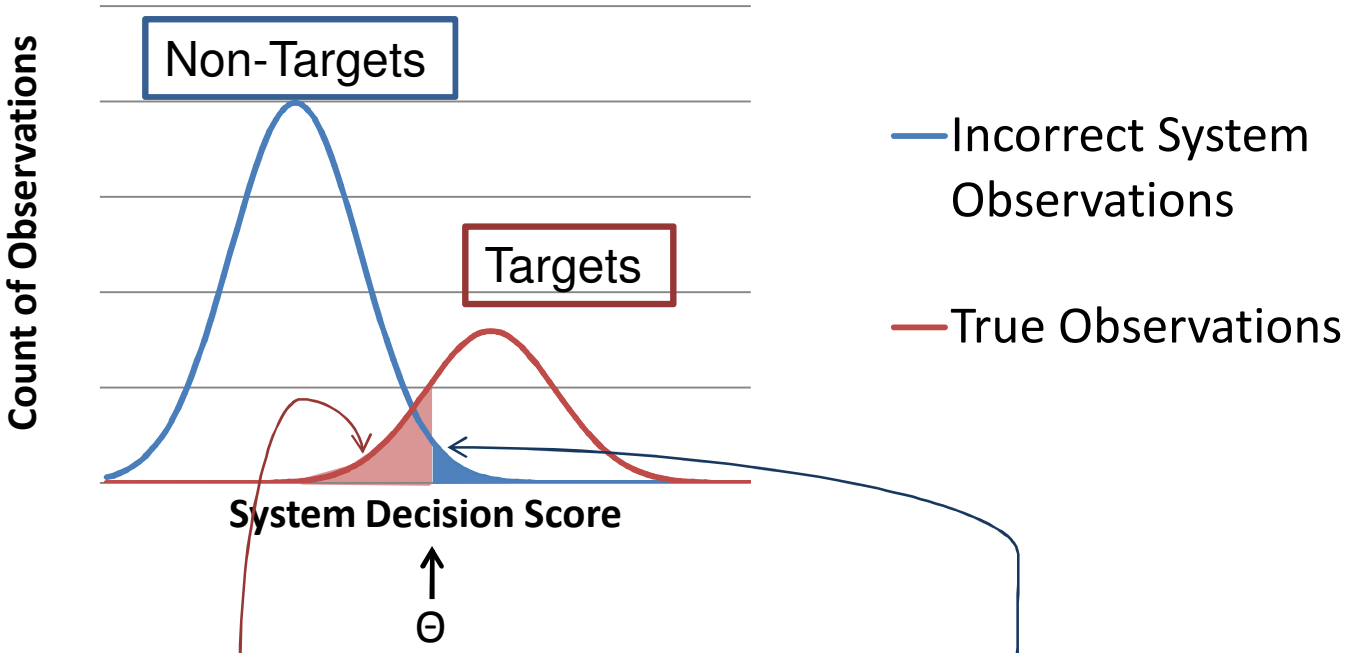
Decision Error Tradeoff Curves

$Prob_{Miss}$ vs. $Rate_{FA}$



Decision Error Tradeoff Curves
$Prob_{Miss}$ vs. $Rate_{FA}$

Decision Score Histogram Separated wrt. Reference Annotation s



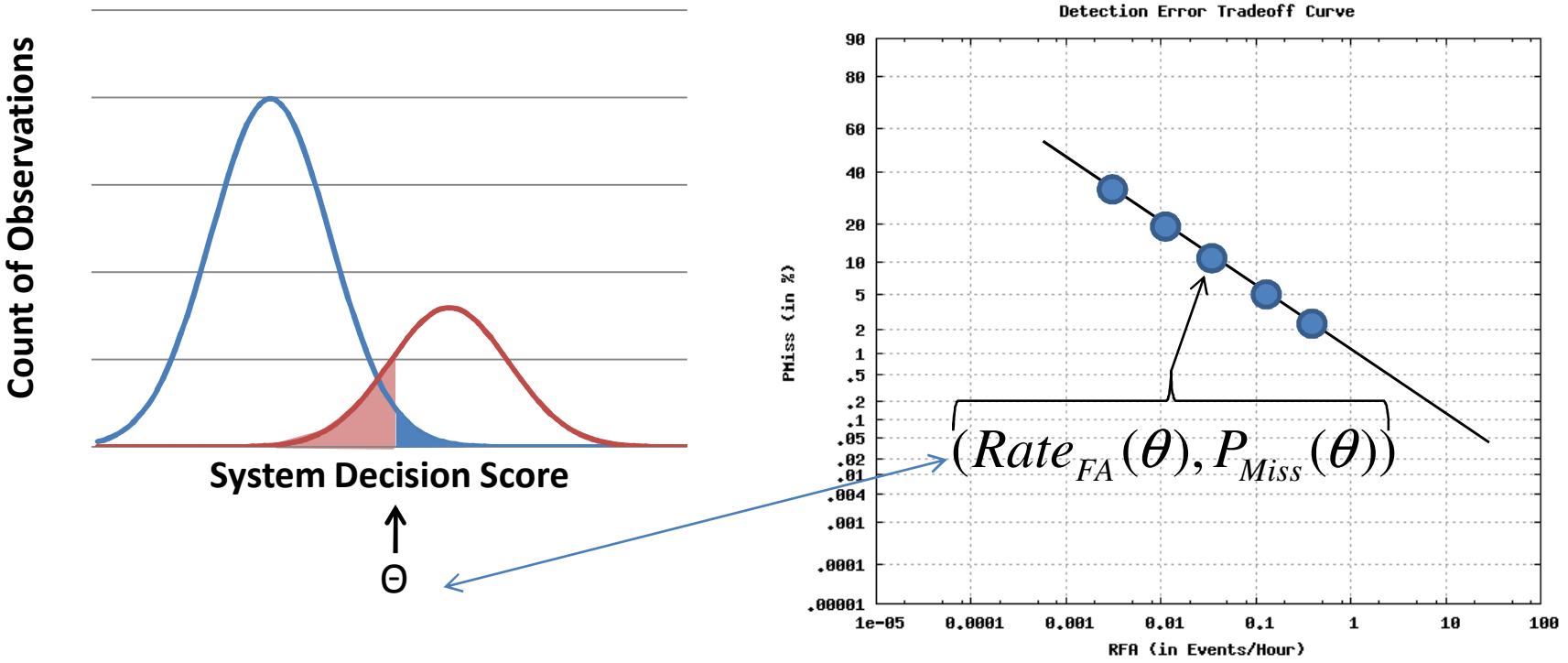
$$P_{Miss}(\theta) = \frac{\#MissedObs}{\#TrueObs}$$

$$Rate_{FA}(\theta) = \frac{\#FalseAlarms}{SignalDuration}$$

Normalizing by # of Non-Observations is impossible for Streaming Detection Evaluations

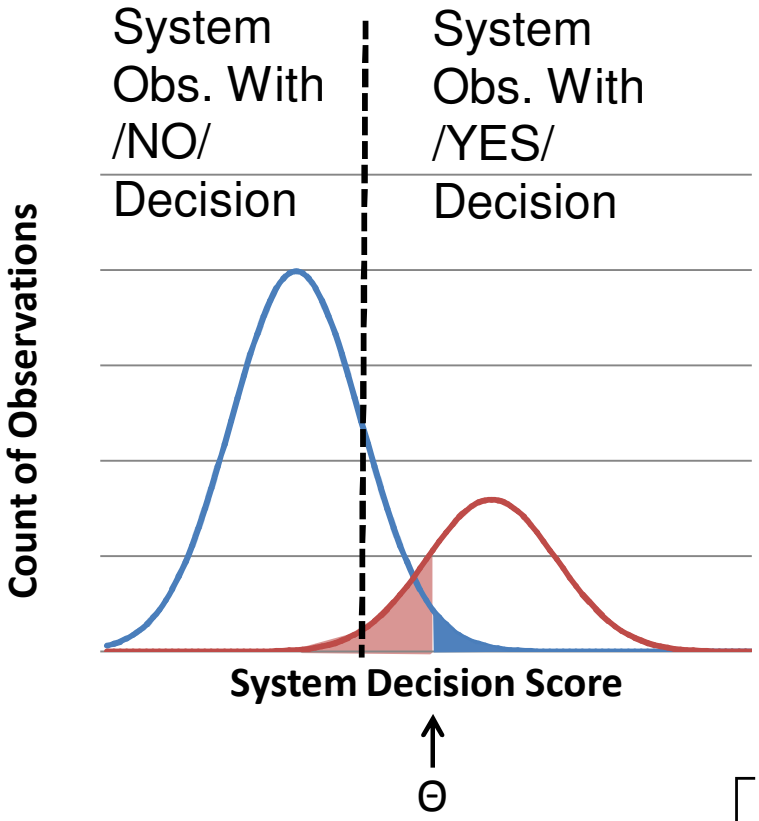
Decision Error Tradeoff Curves
$Prob_{Miss}$ vs. $Rate_{FA}$

Compute $Rate_{FA}$ and P_{Miss} for all Θ



$$MinimumNDCR(\theta) = \arg\min_{\theta}\left[P_{Miss}(\theta) + \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} * R_{FA}(\theta)\right]$$

Decision Error Tradeoff Curves
Actual vs. Minimum NDCR

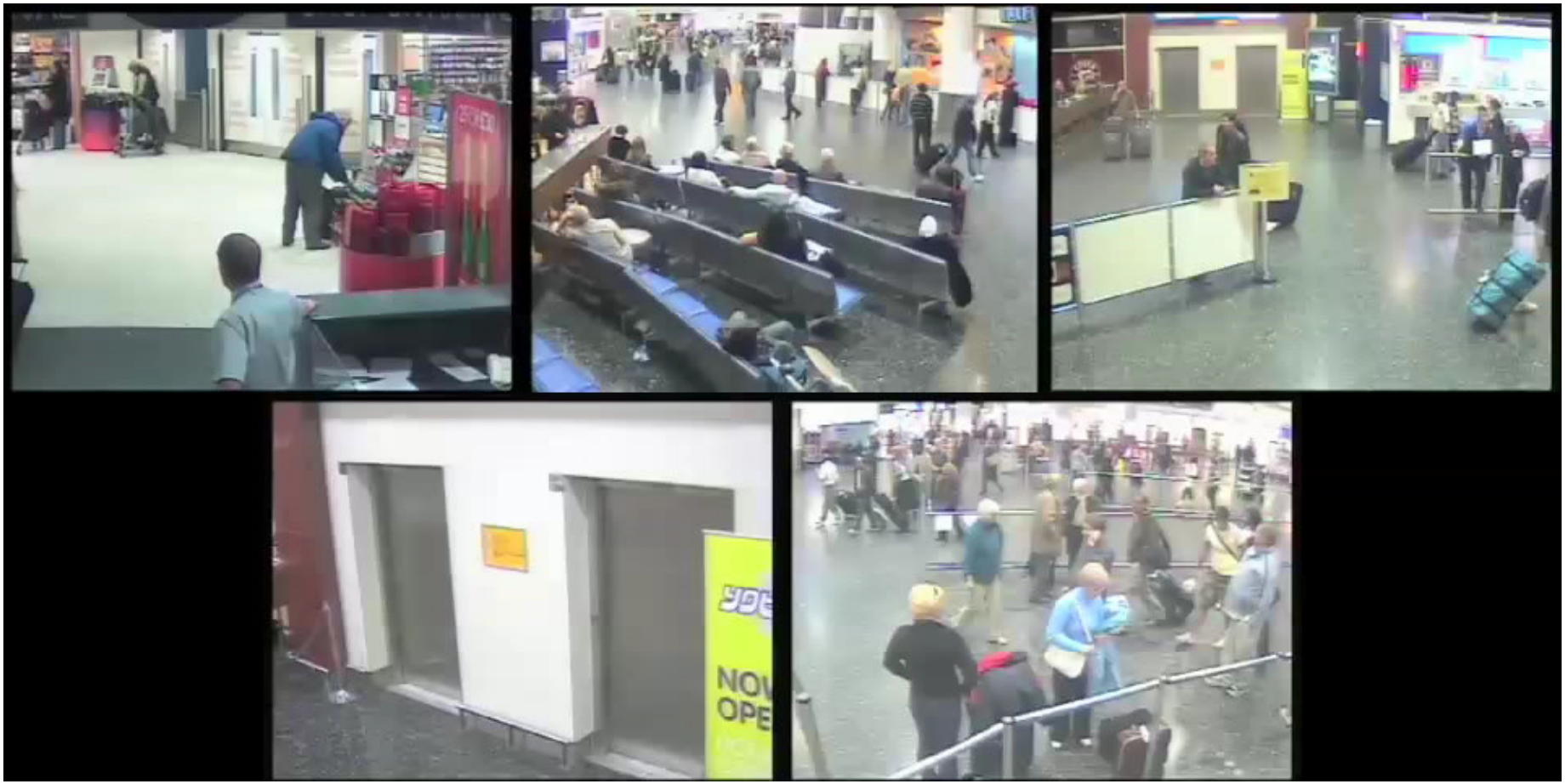


Event Detection Constants

$$Cost_{Miss} = 10$$
$$Cost_{FA} = 1$$
$$R_{Target} = 20$$

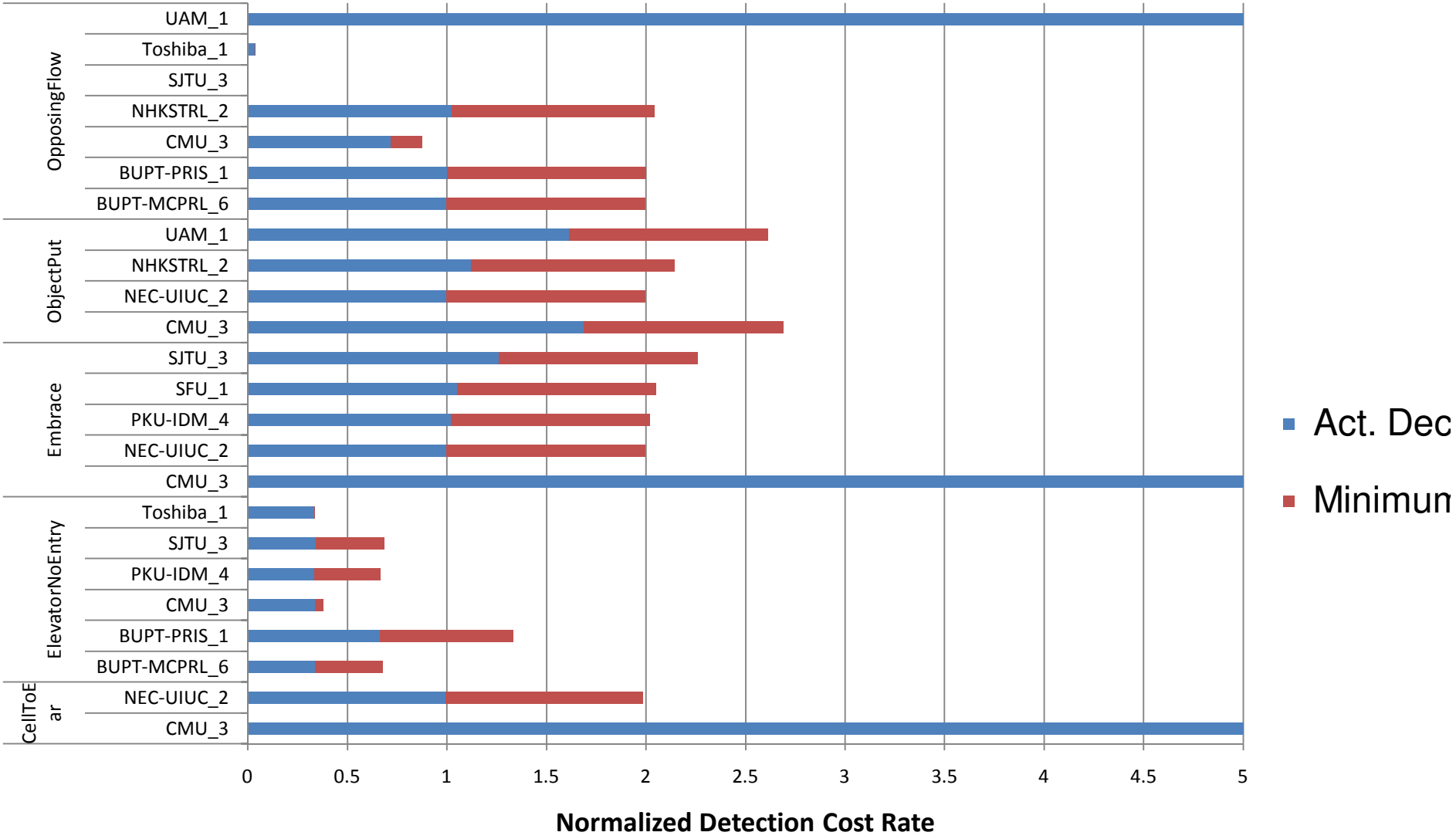
$$MinimumNDCR(\theta) = \arg\min_{\theta}\left[P_{Miss}(\theta) + \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} * R_{FA}(\theta)\right]$$

$$ActualNDCR(Act.Dec.) = P_{Miss}(Act.Dec.) + \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} * R_{FA}(Act.Dec.)$$

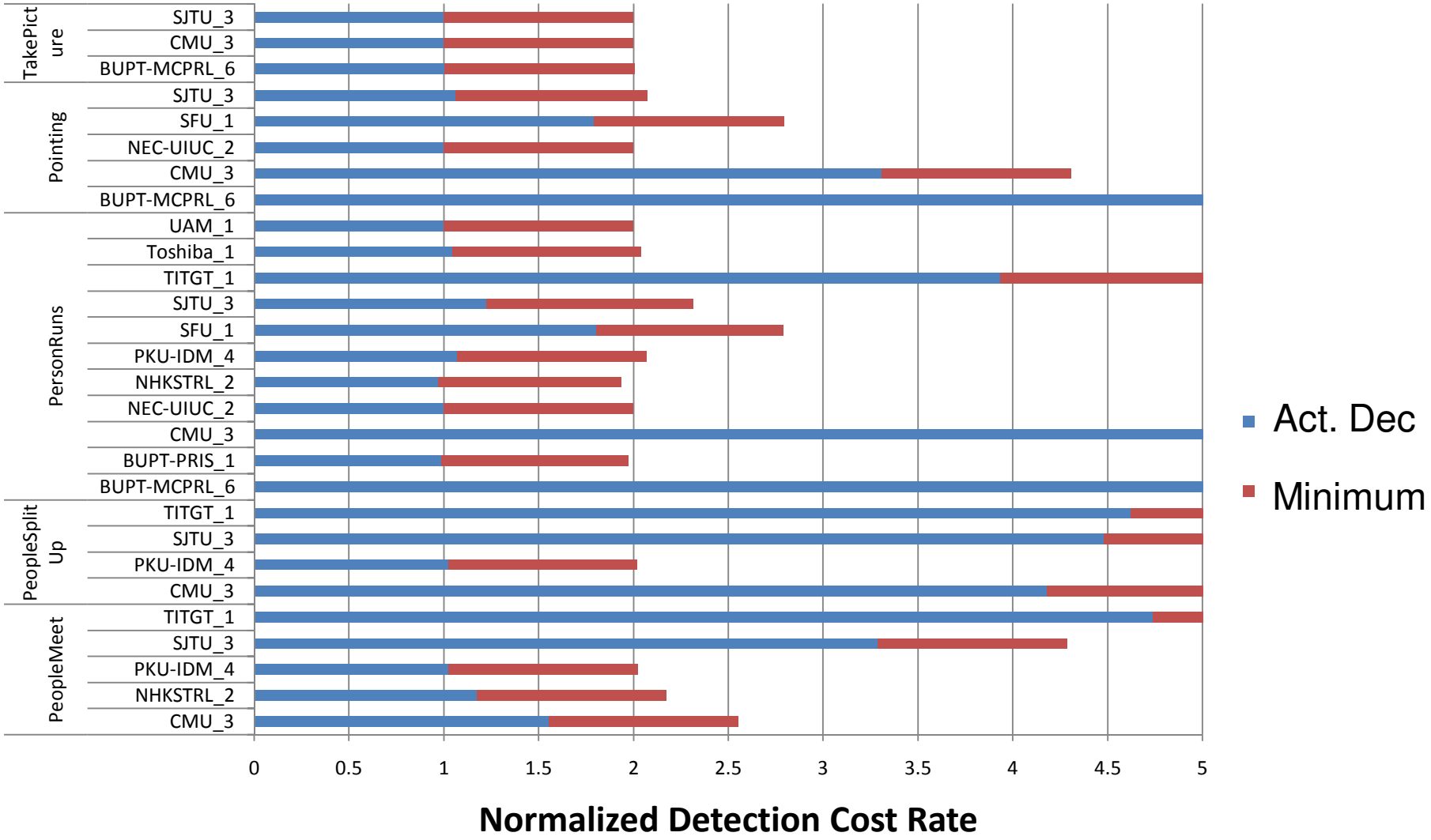


2009 Event Detection Results

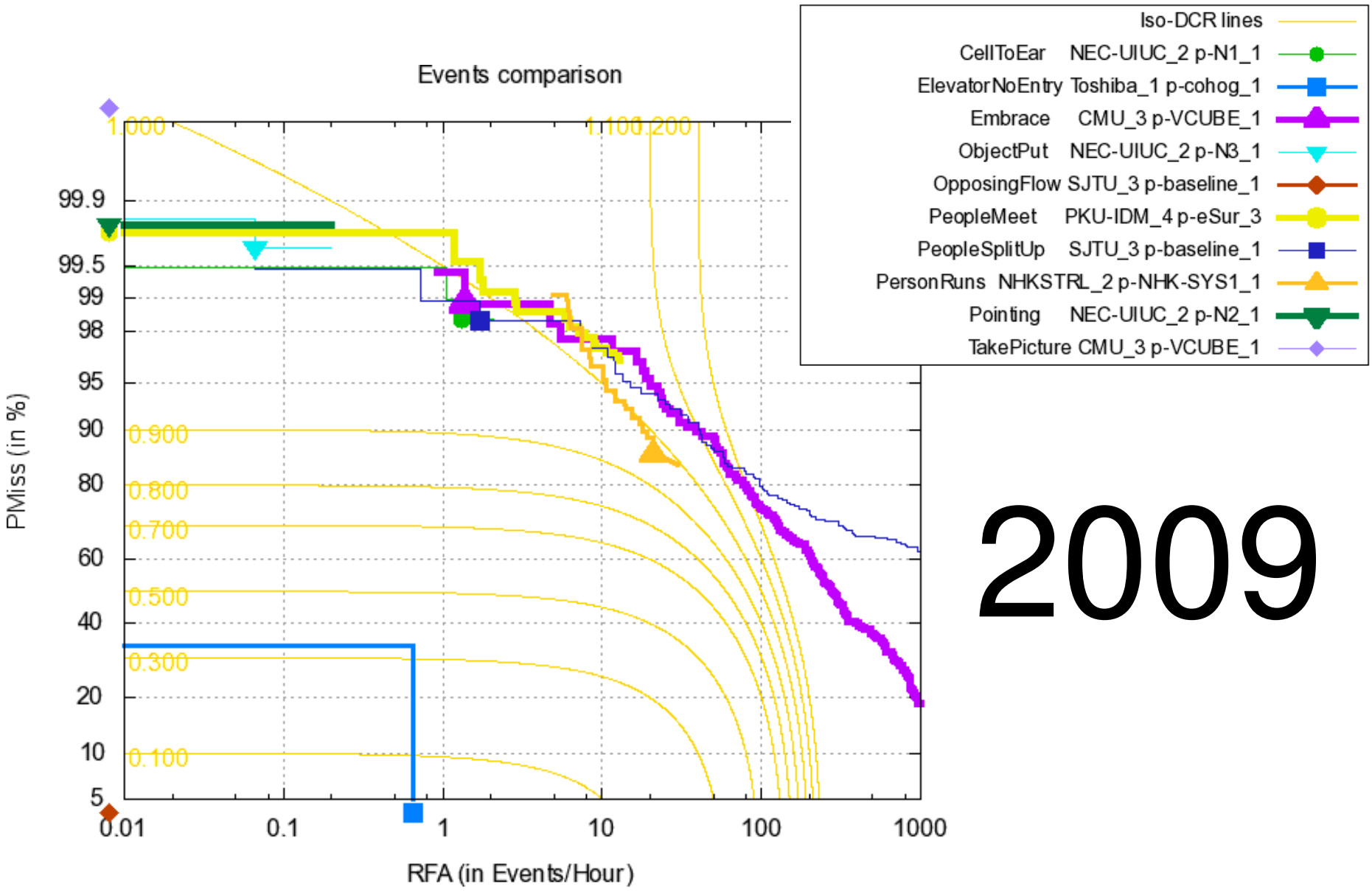
2009 Minimum and Actual NDCRs (Set 1)



2009 Minimum and Actual NDCRs (Set 2)

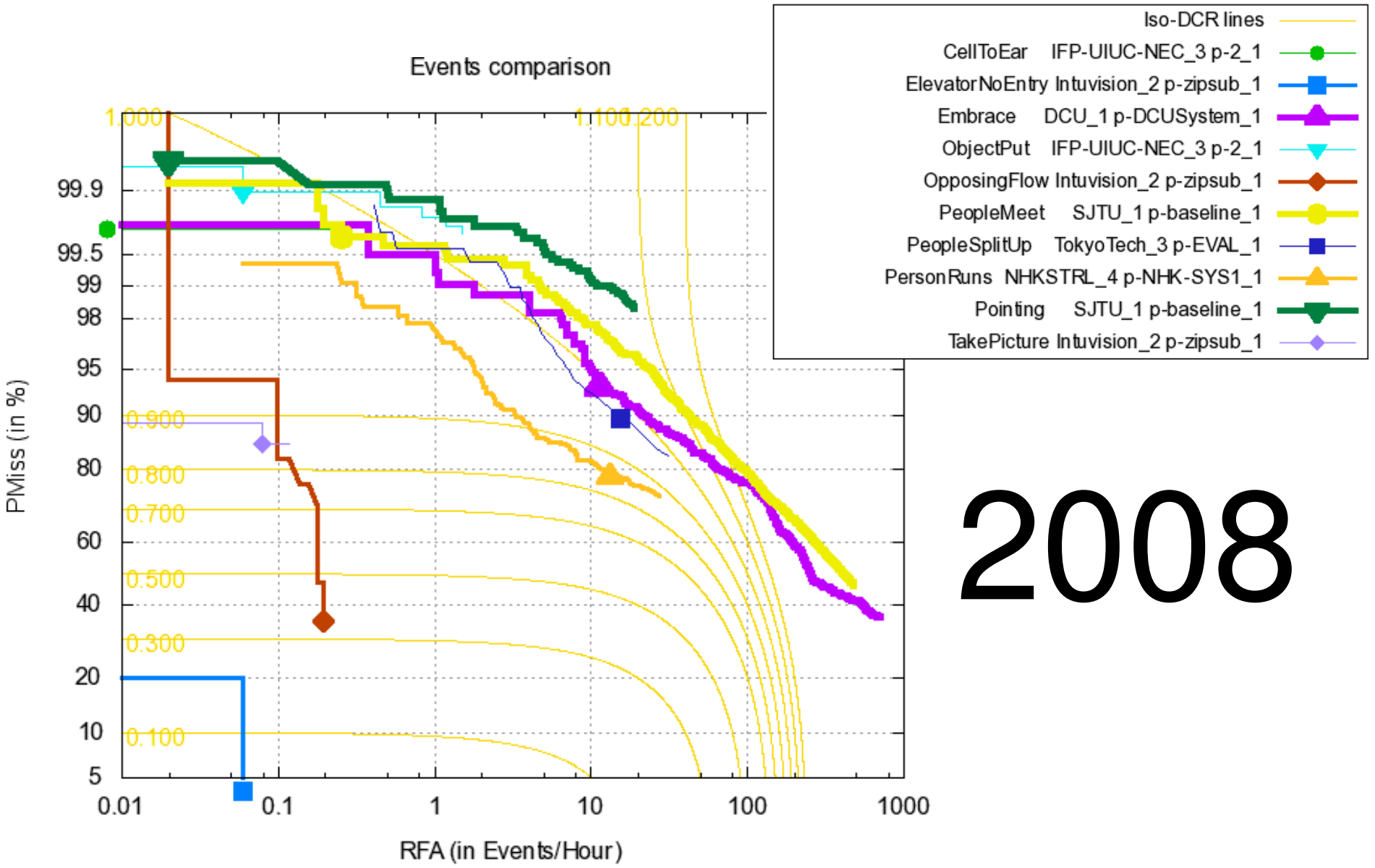


2009 Best DET Curves for Events



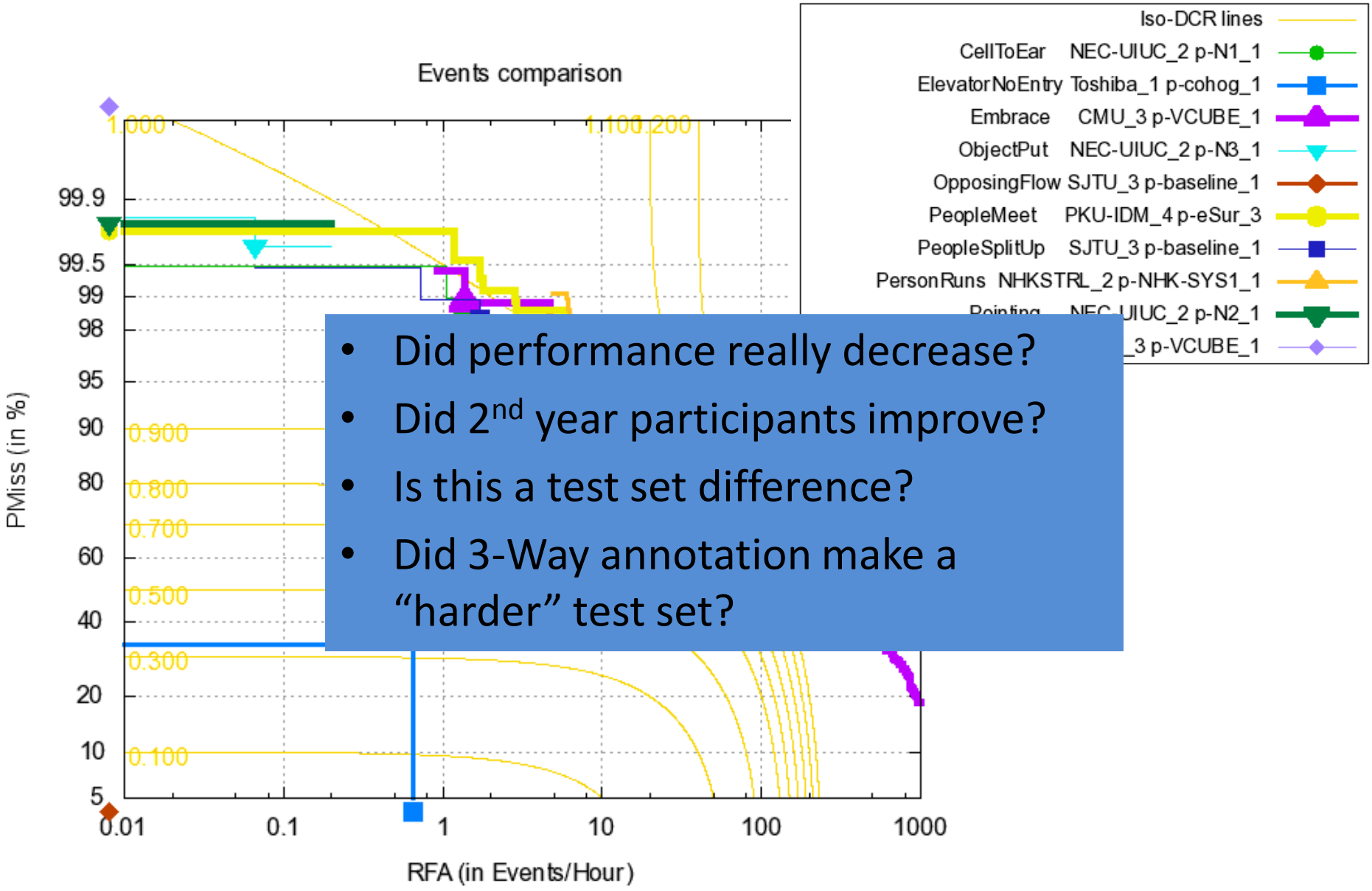
2009

2008 Best DET Curves for Events



2008

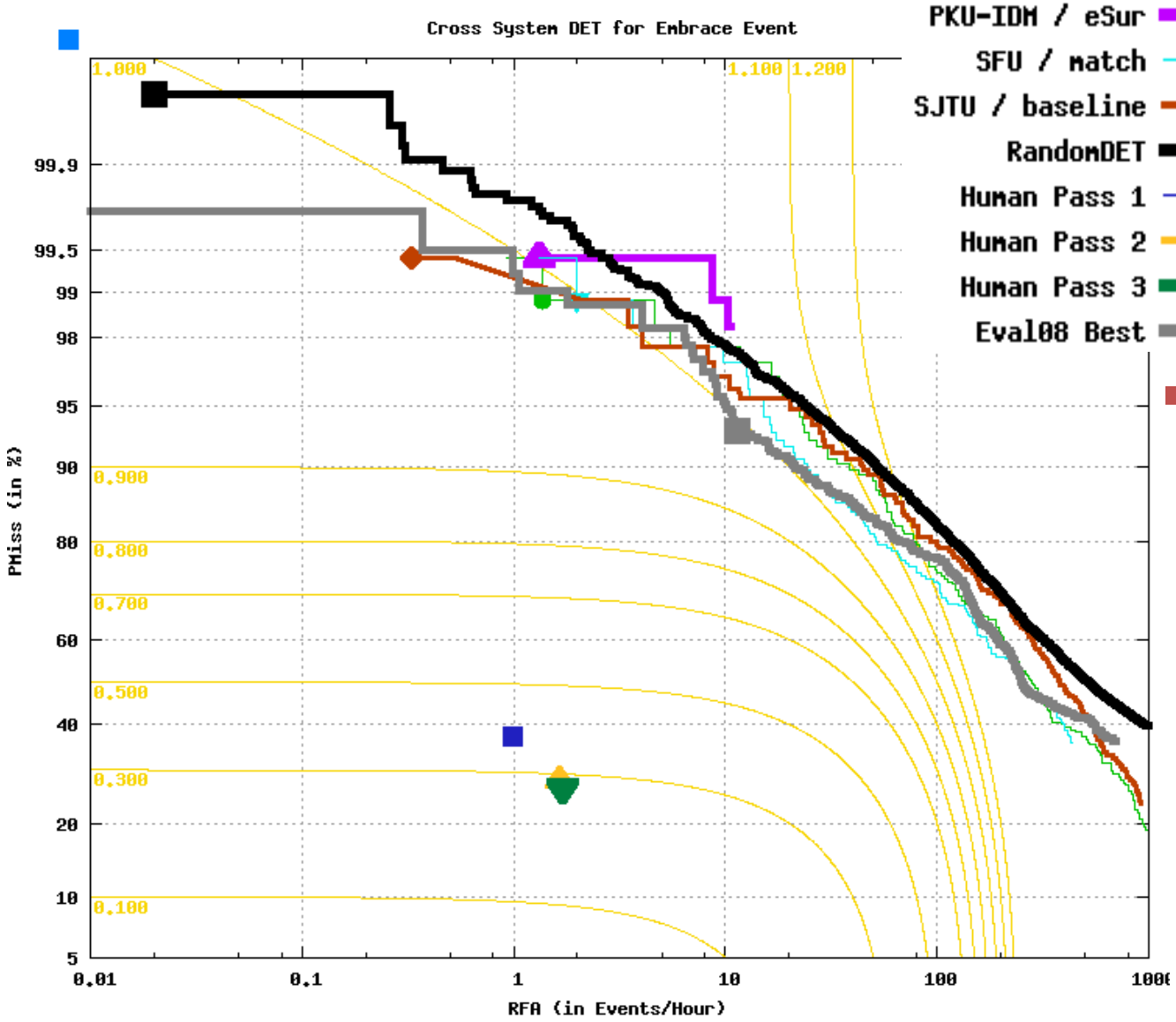
2009 Best DET Curves for Events

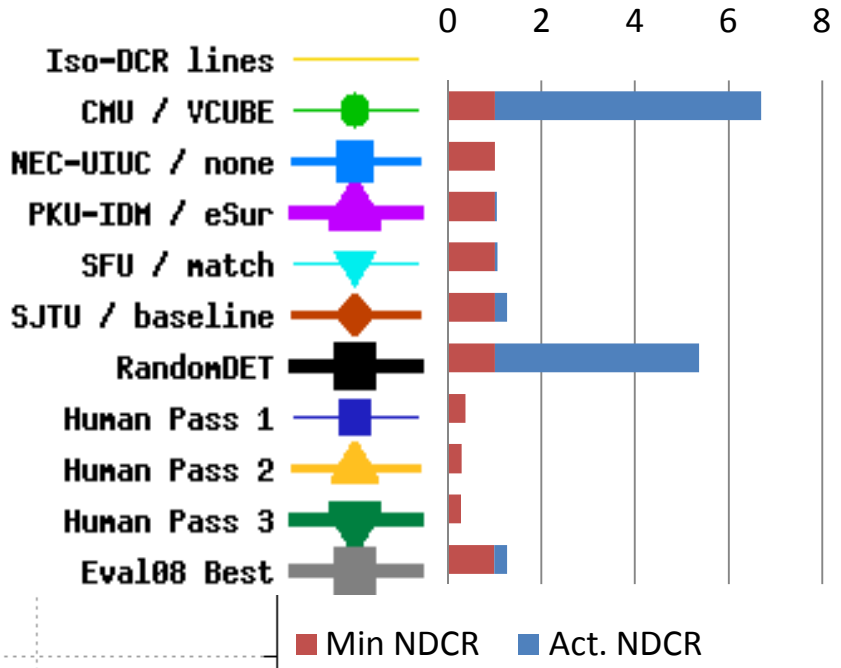


- Did performance really decrease?
- Did 2nd year participants improve?
- Is this a test set difference?
- Did 3-Way annotation make a "harder" test set?

Embrace Event

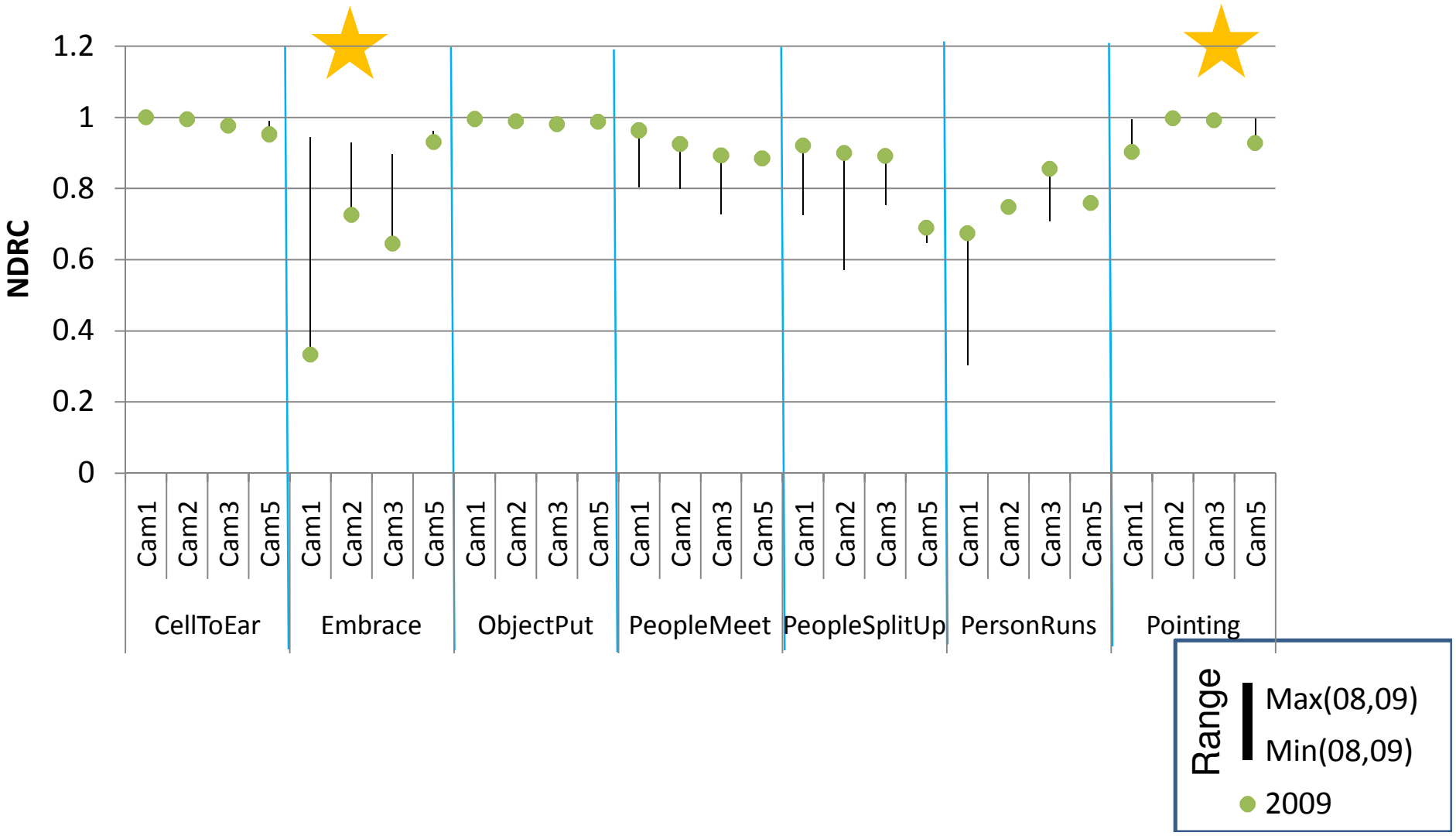
Best Submission per Site





2008 vs. 2009 Minimum NDCRs

Conditioned by Selected Events and Cameras

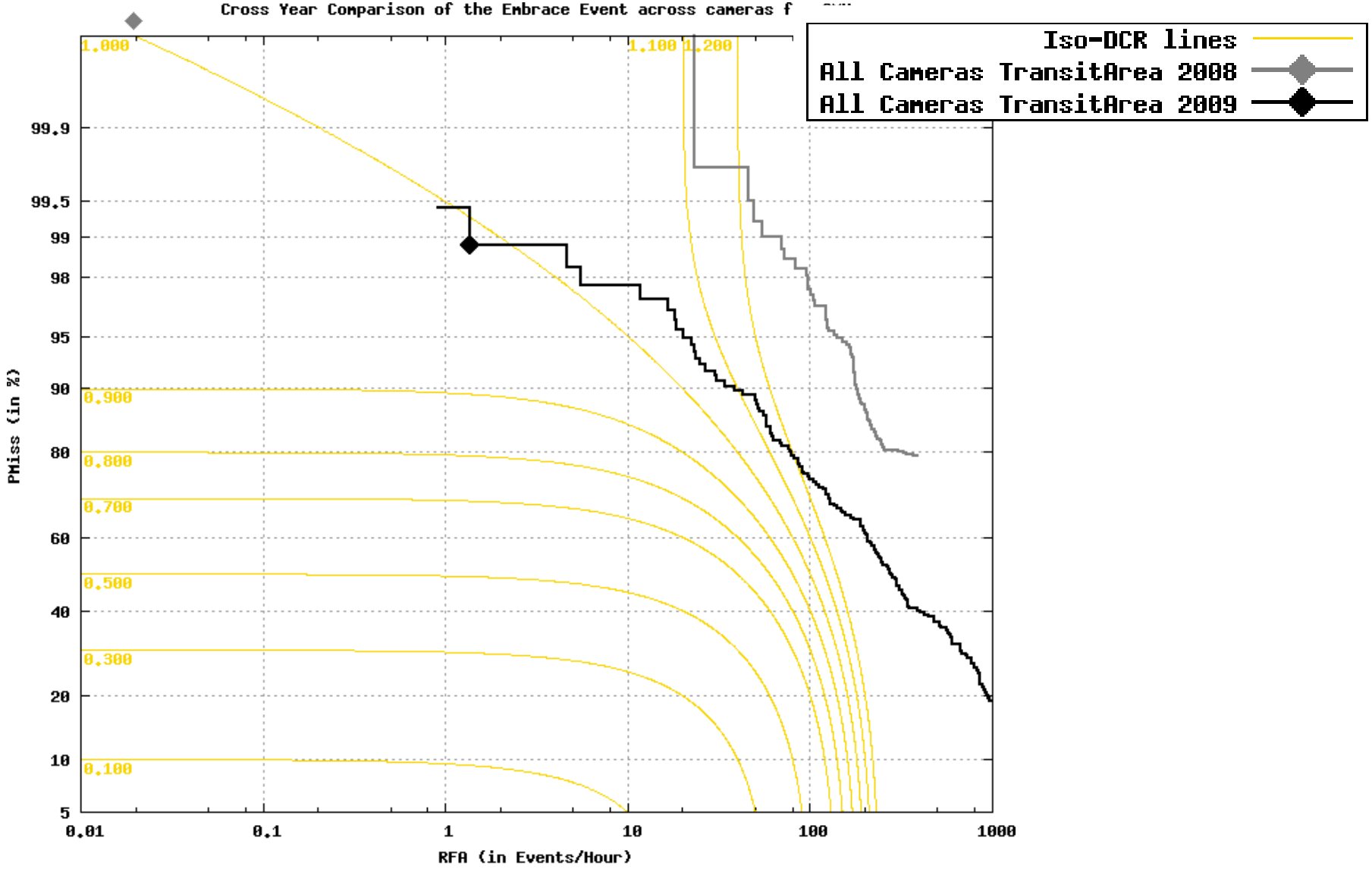


2008 vs. 2009 Minimum NDCRs

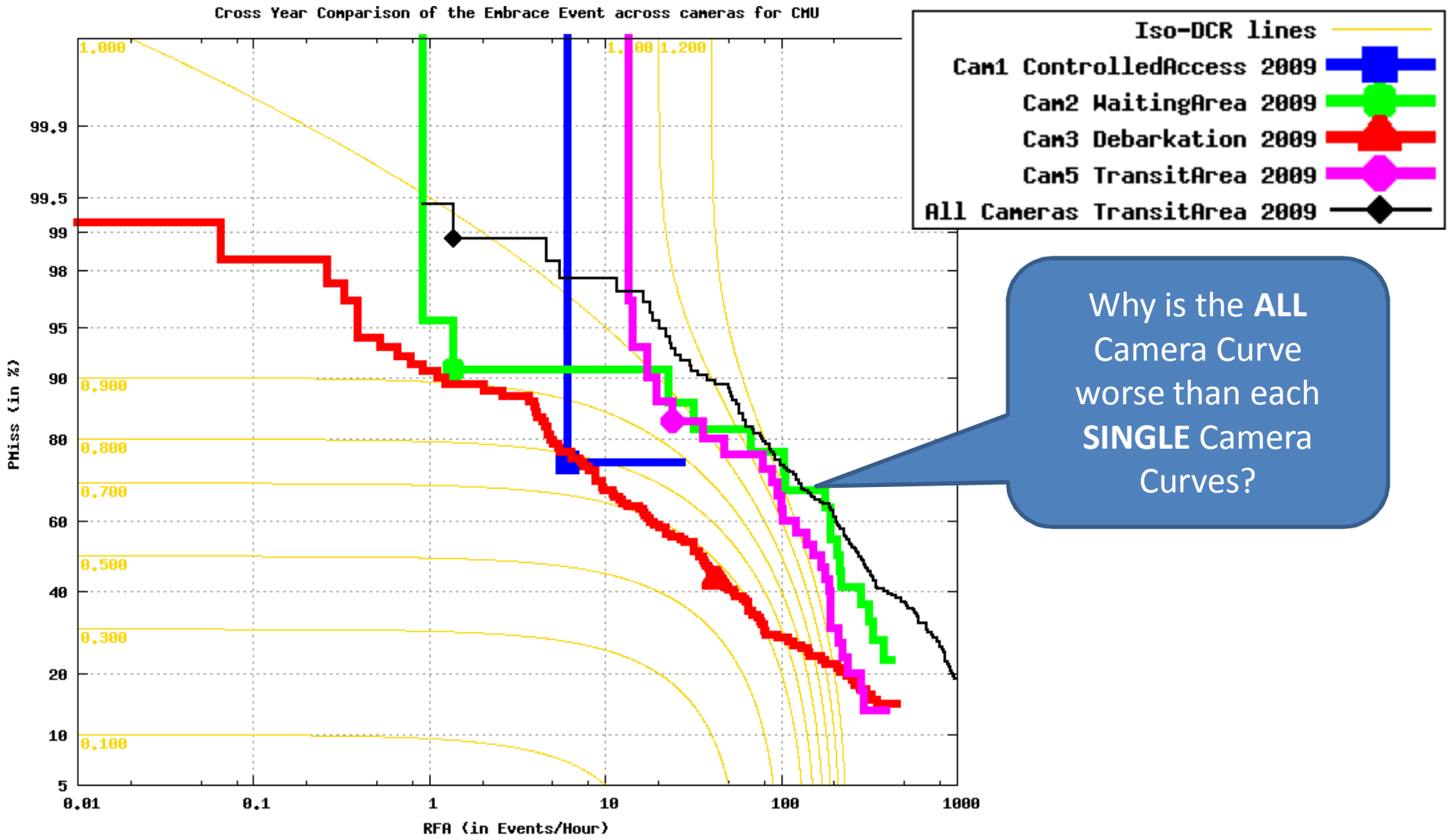
Limited to 2nd Year Participants

Conditioned by Selected Events and Cameras

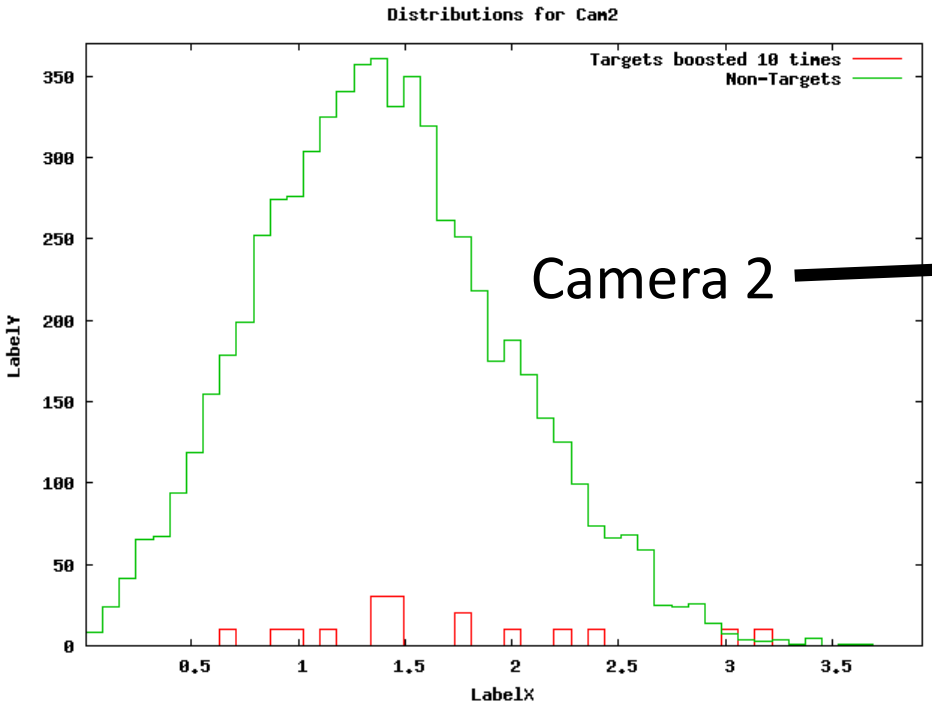
CMU 2008 and 2009 Embrace Event Submissions



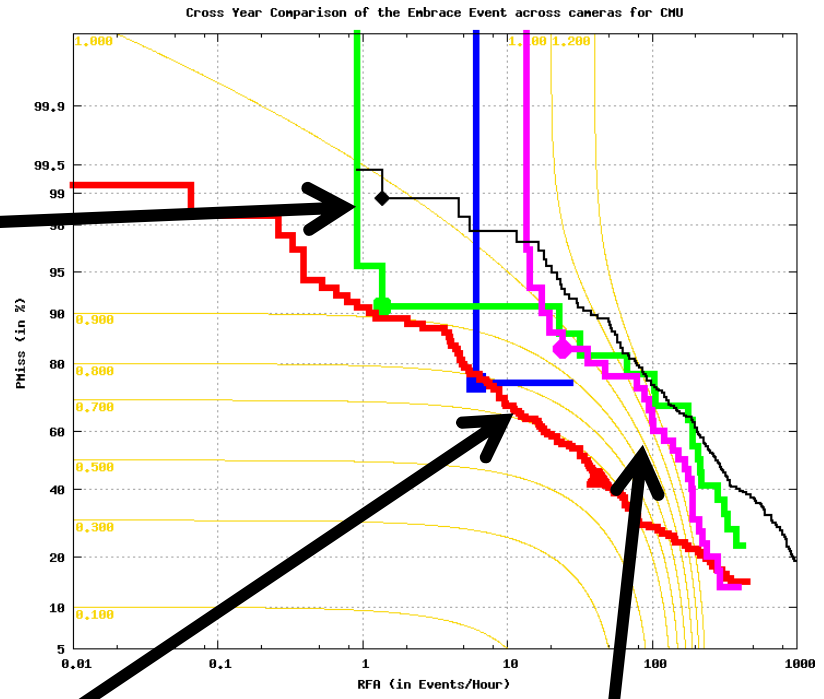
CMU 2008 and 2009 Embrace Event Submissions Split By Cameras

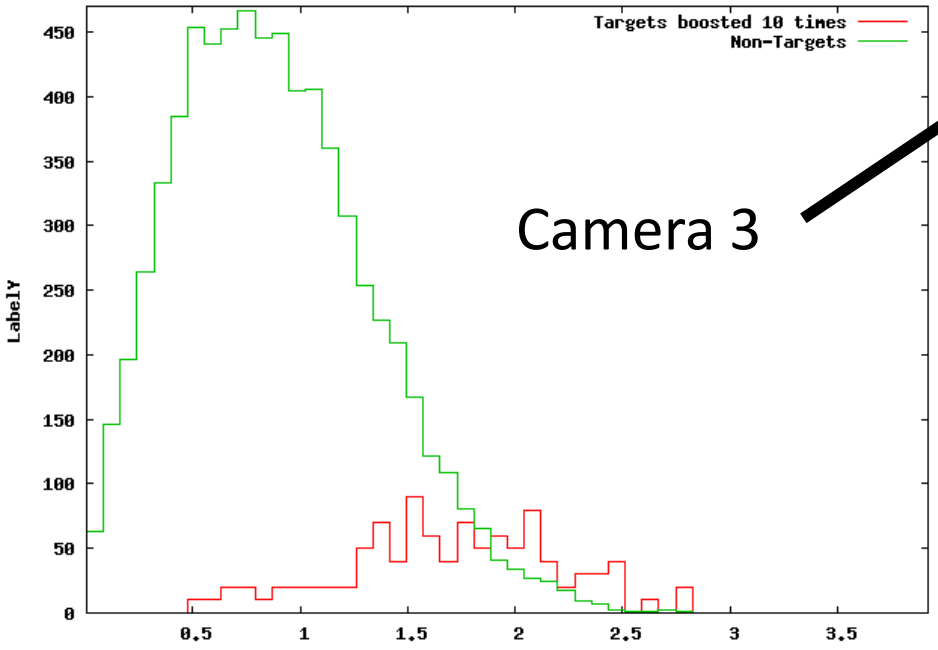


Why is the **ALL** Camera Curve worse than each **SINGLE** Camera Curves?

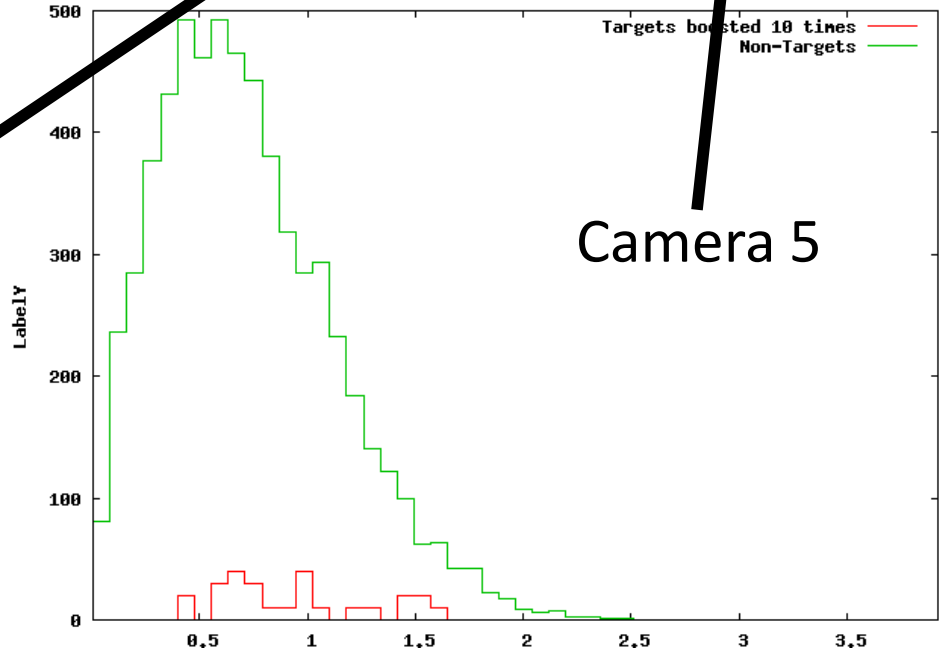


Camera 2





Camera 3



Camera 5

Conclusions and Lessons Learned

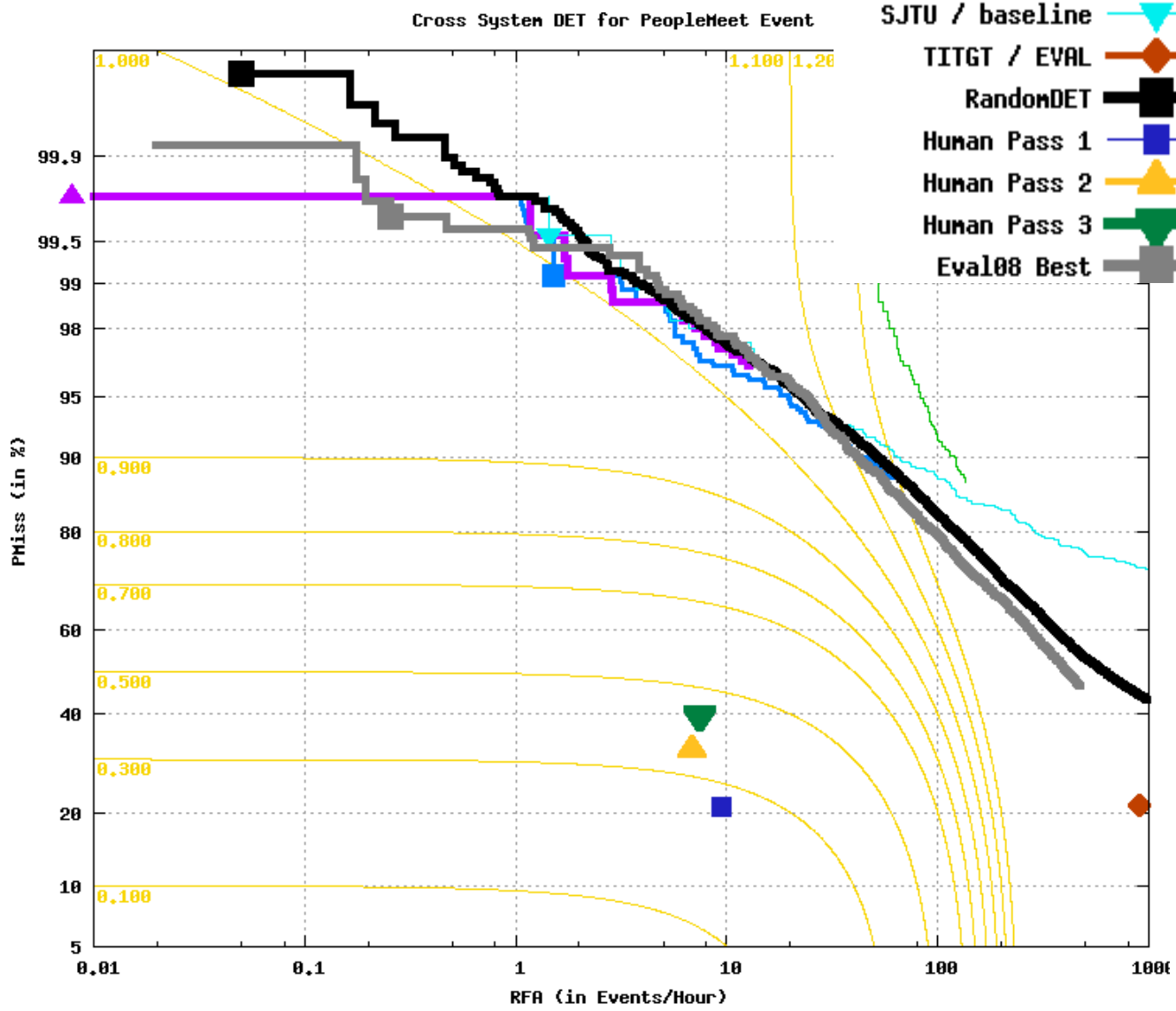
- Improvement can be seen in 2 of the events on specific cameras
- Multiple-year participants have shown improvement on 3 events
  - Decision score normalization is important
  - Non-optimal normalization obscures performance gains
- The change in annotation scheme improved the number of found event instances
  - We will be studying the effect on scoring
- Next year's evaluation should re-use this year's test set but in what manner

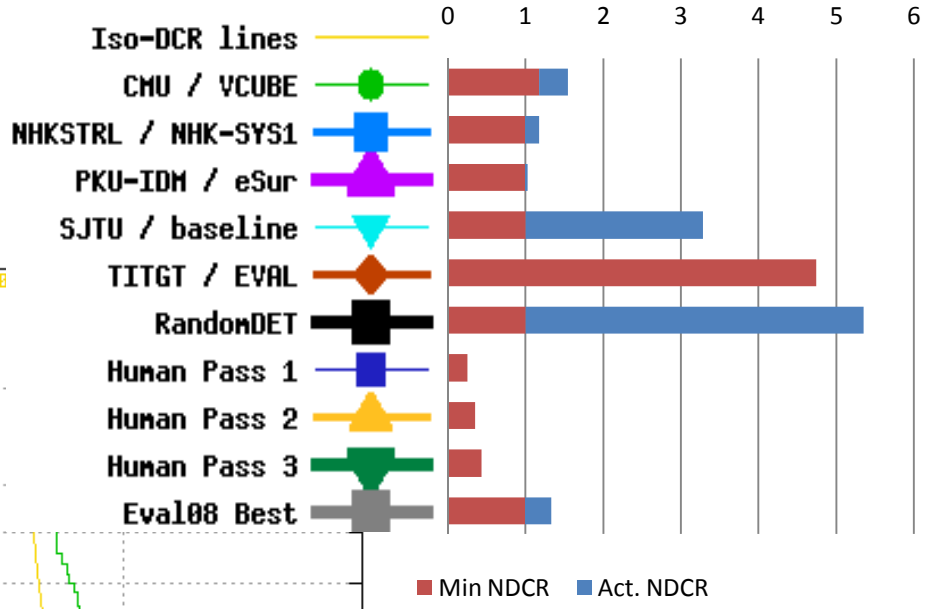
End of Talk

Back up slides to follow

PeopleMeet Event

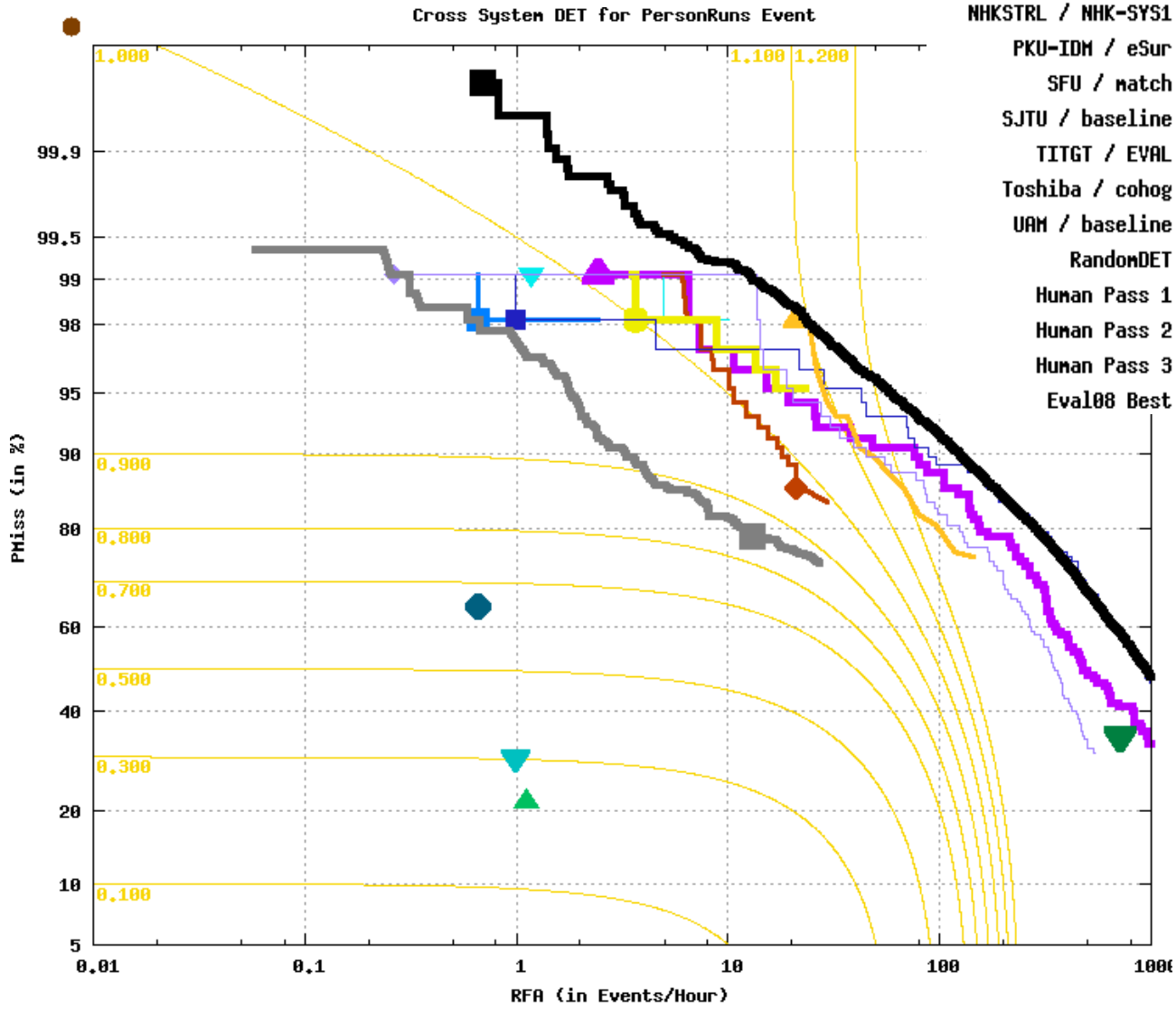
Best Submission per Site

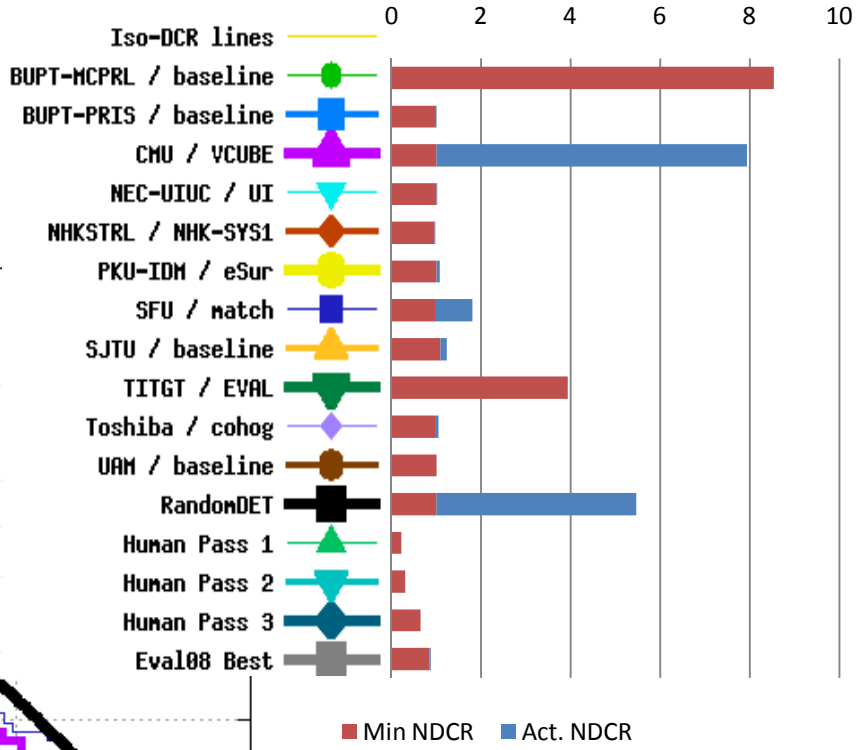




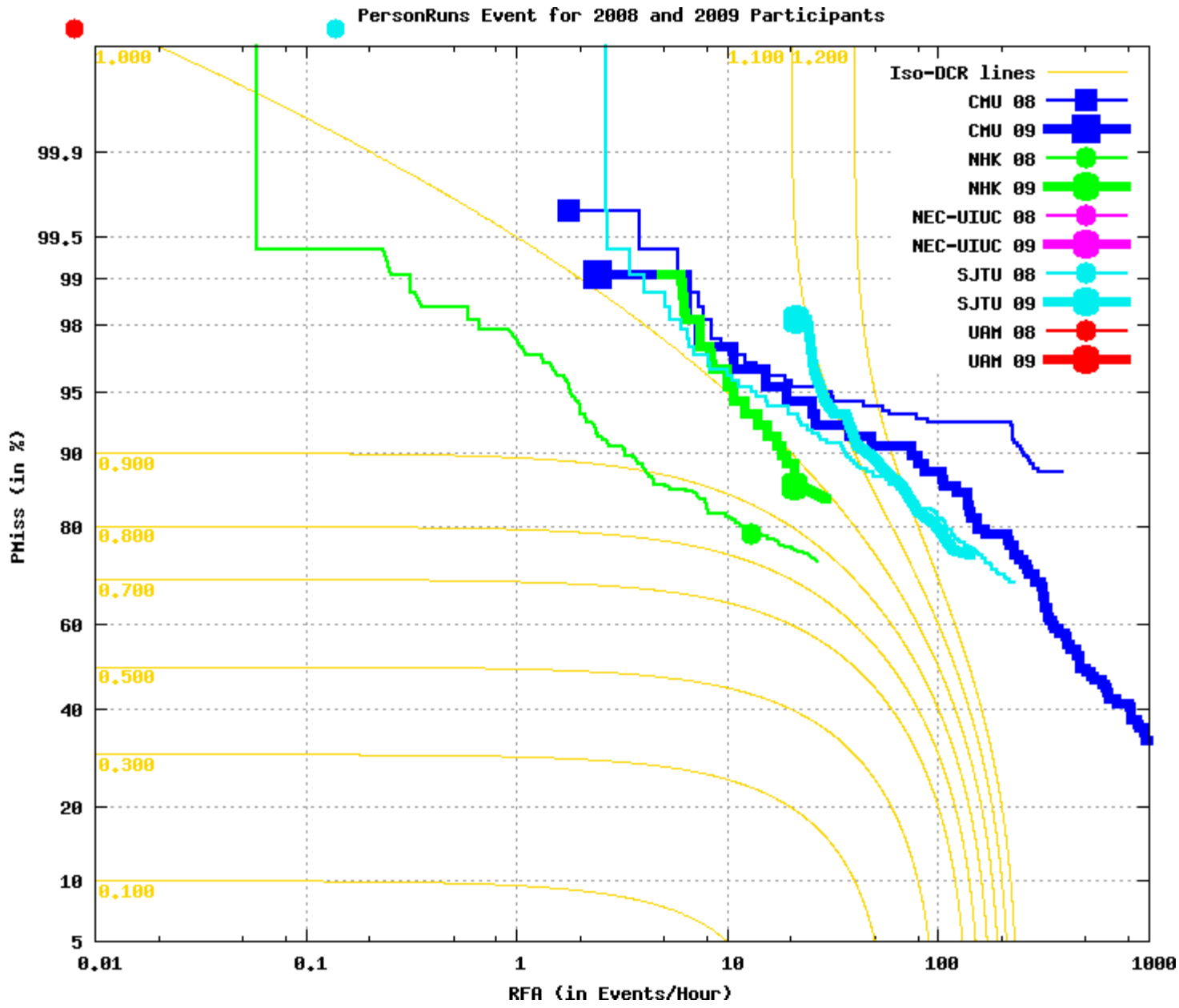
PersonRuns Event

Best Submission per Site



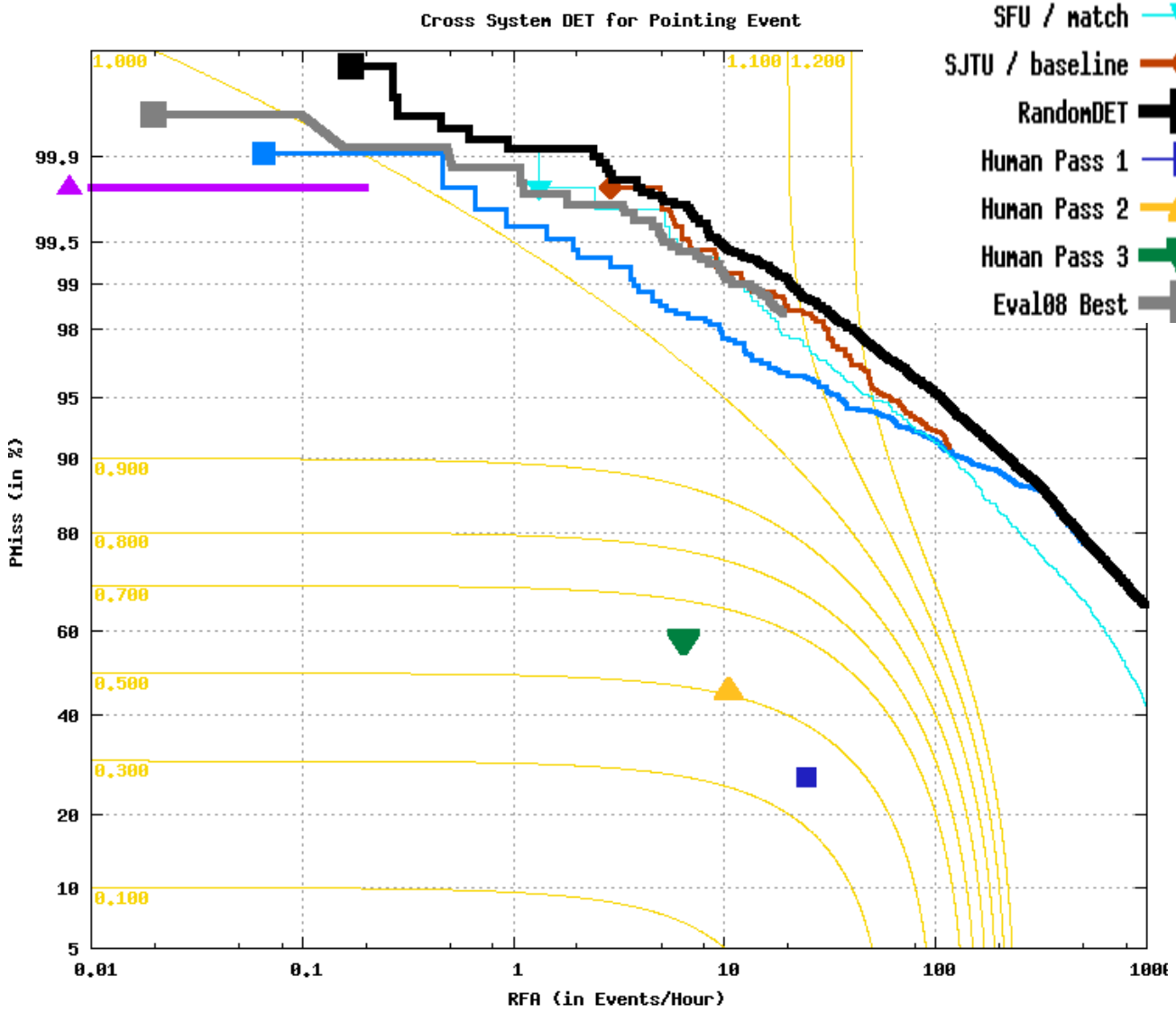


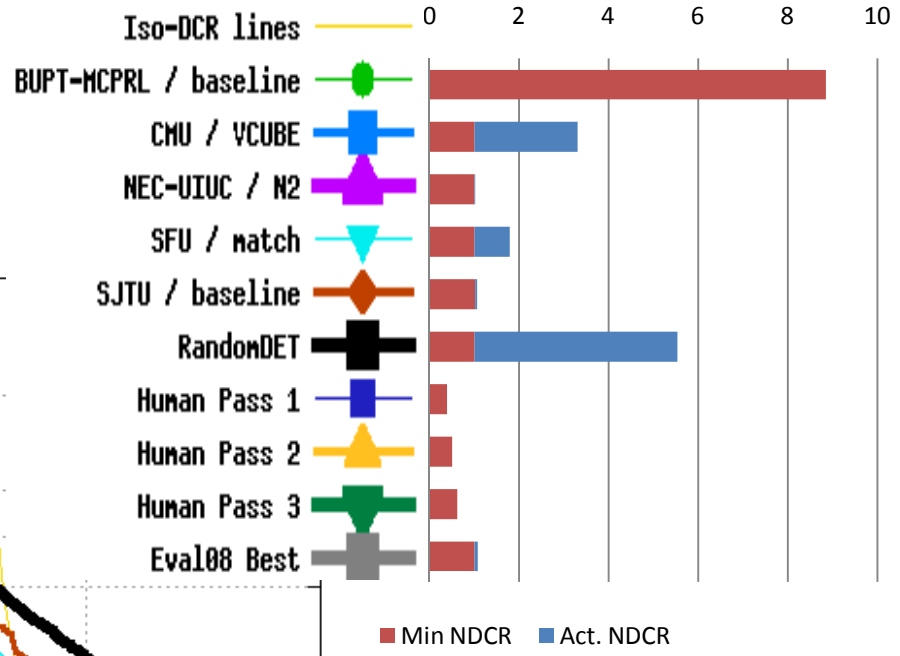
PersonRuns Limited to Participants



Pointing Event

Best Submission per Site

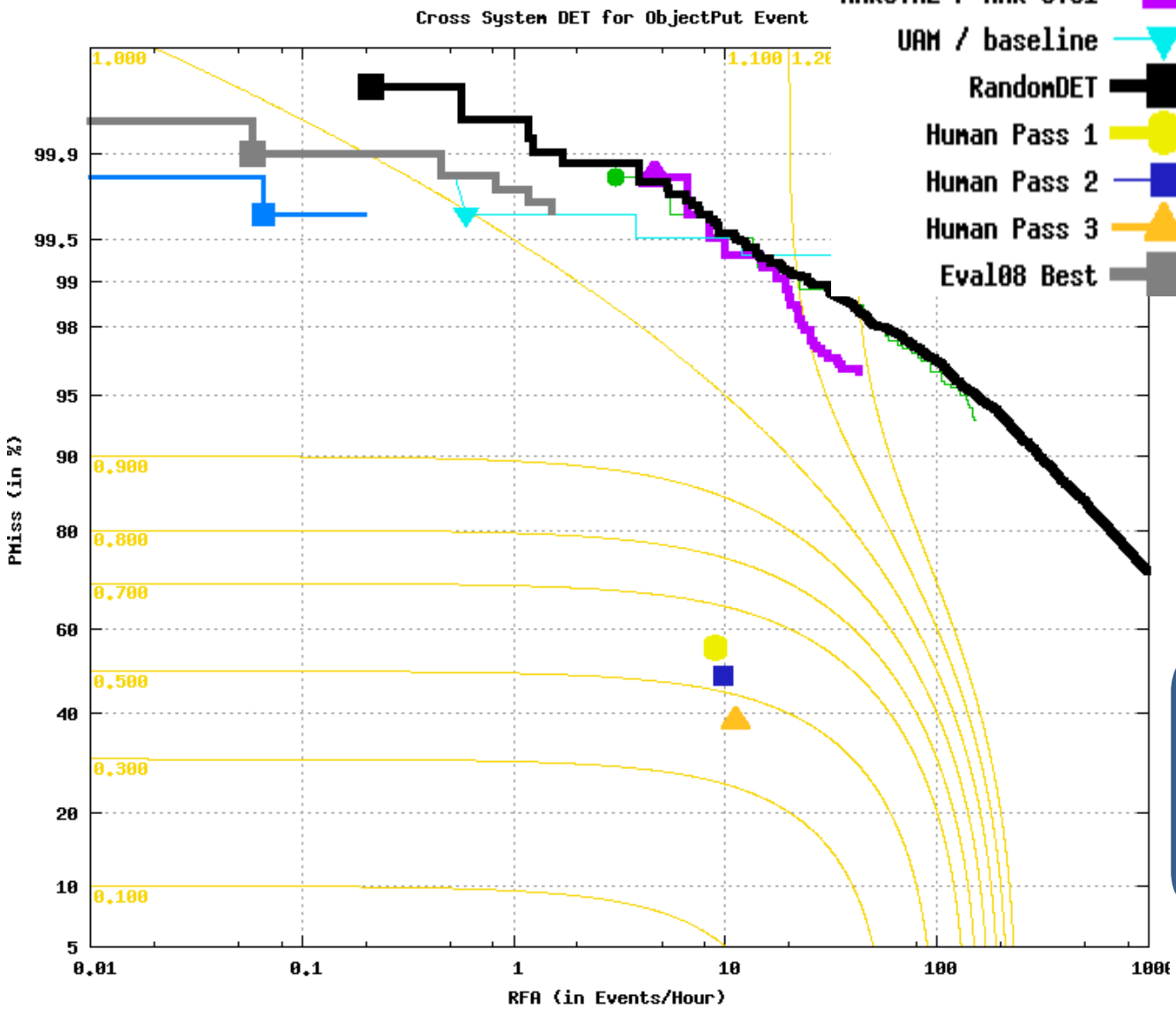


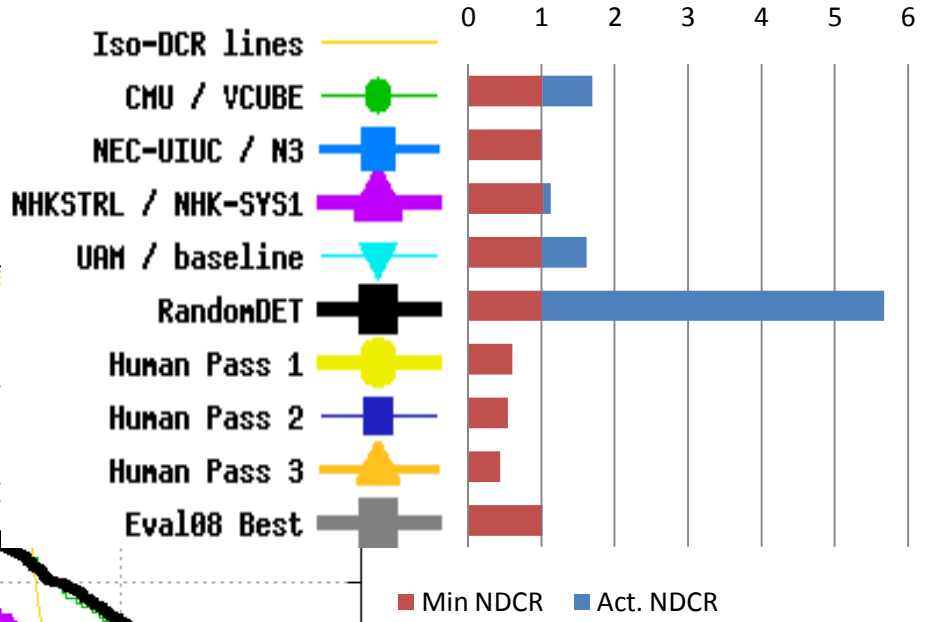


- Random system
  - R_{targ}=XXX,
  - MeanDur=XXs
  - TestDur=XXH

ObjectPut Event

Best Submission per Site

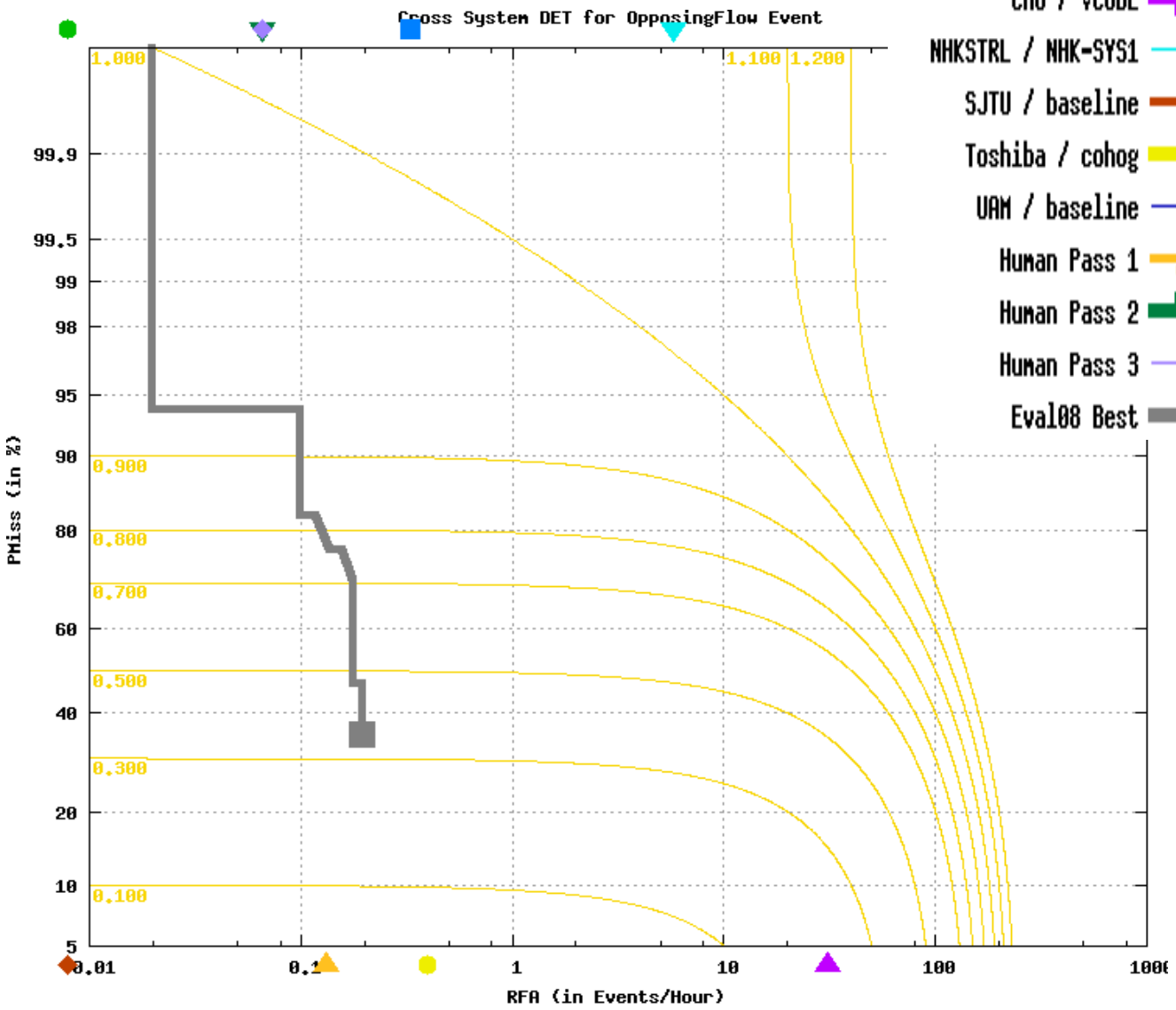


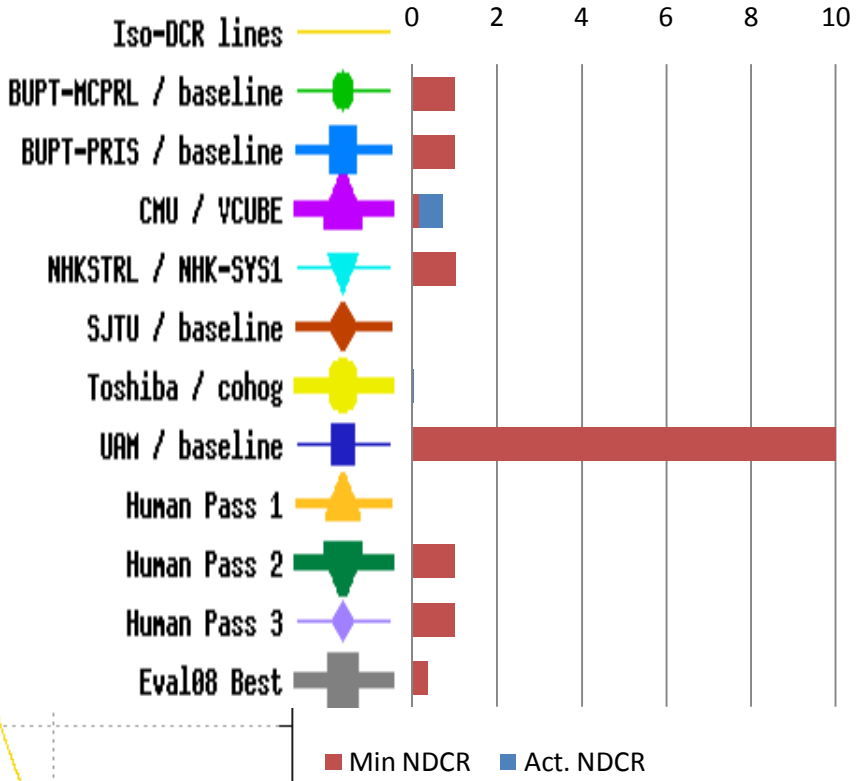


- Random system
  - R_{targ}=XXX,
  - MeanDur=XXs
  - TestDur=XXH

OpposingFlow Event

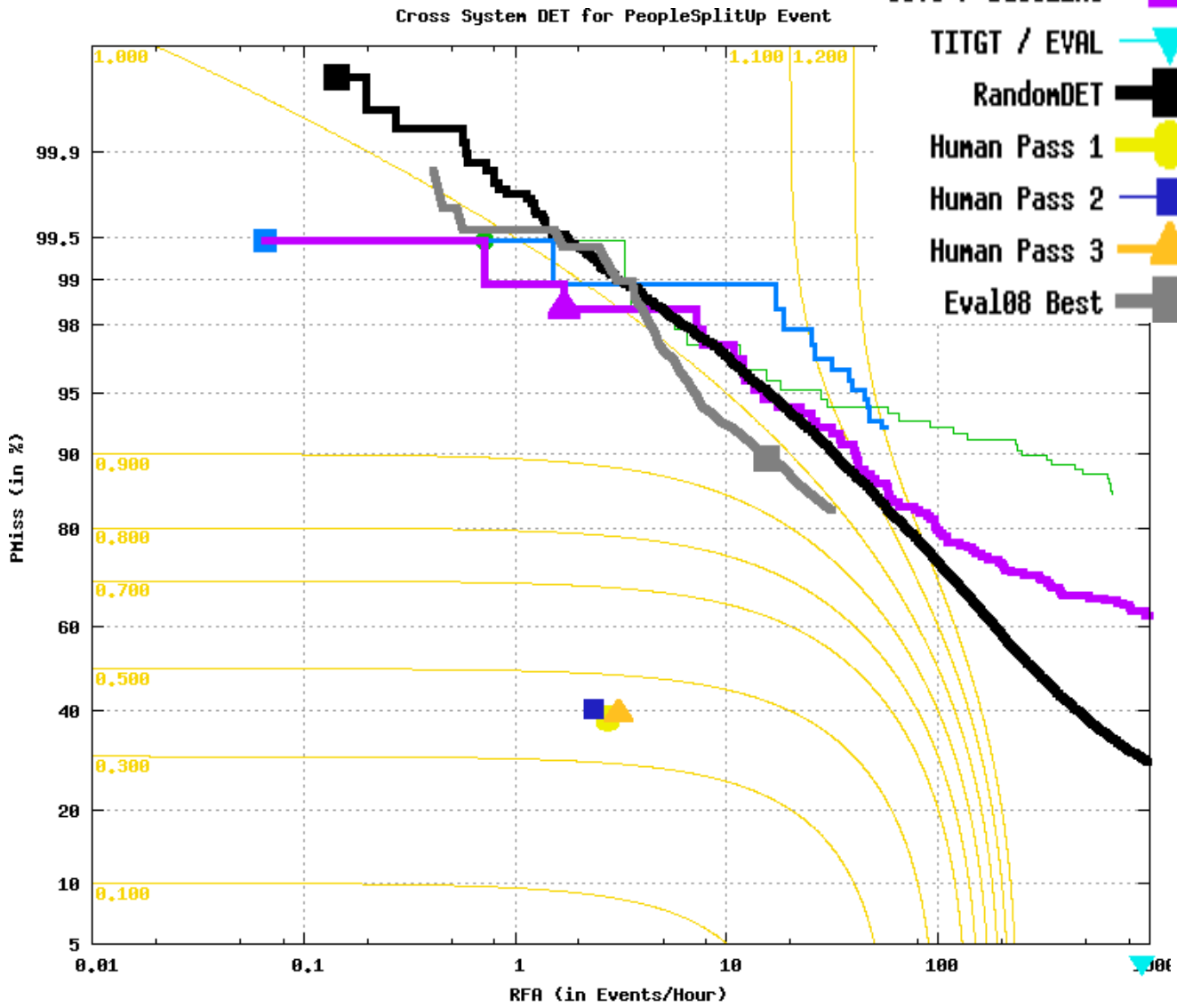
Best Submission per Site

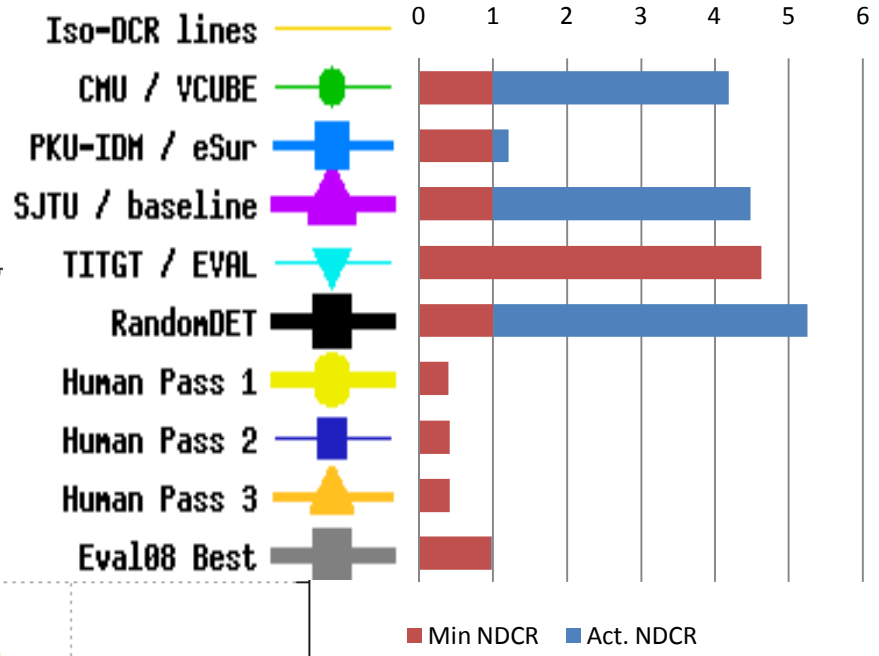




PeopleSplitUp Event

Best Submission per Site

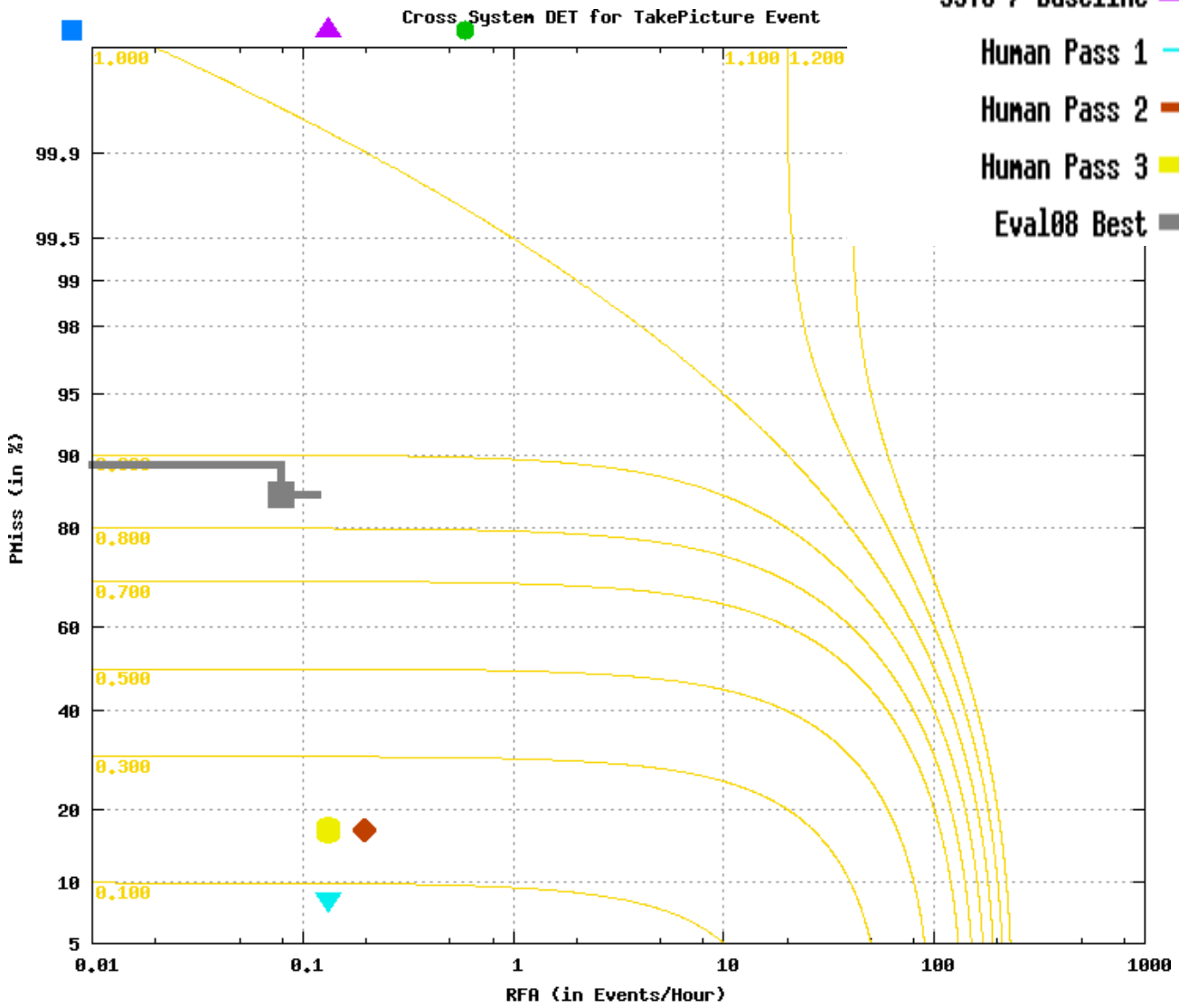


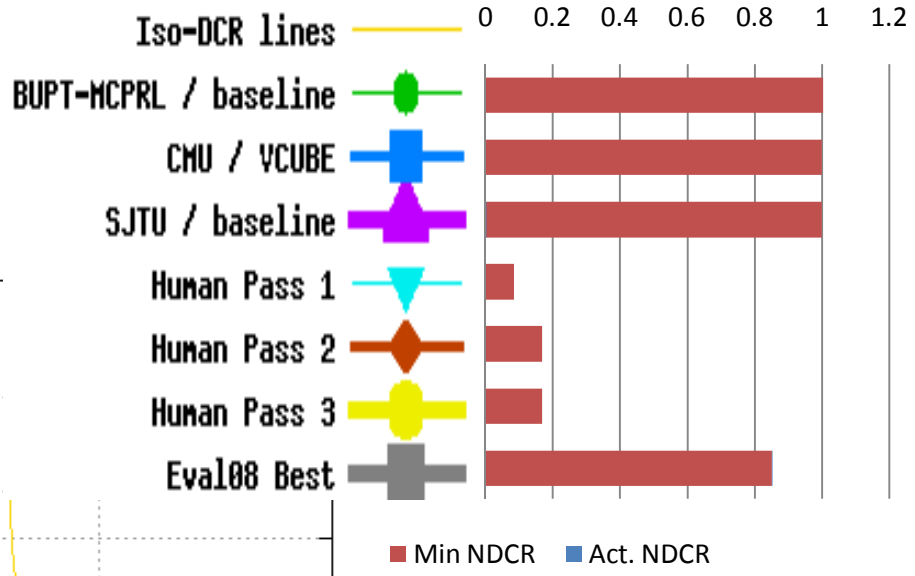


- Random system
  - R_{targ}=XXX,
  - MeanDur=XXs
  - TestDur=XXH

TakePicture Event

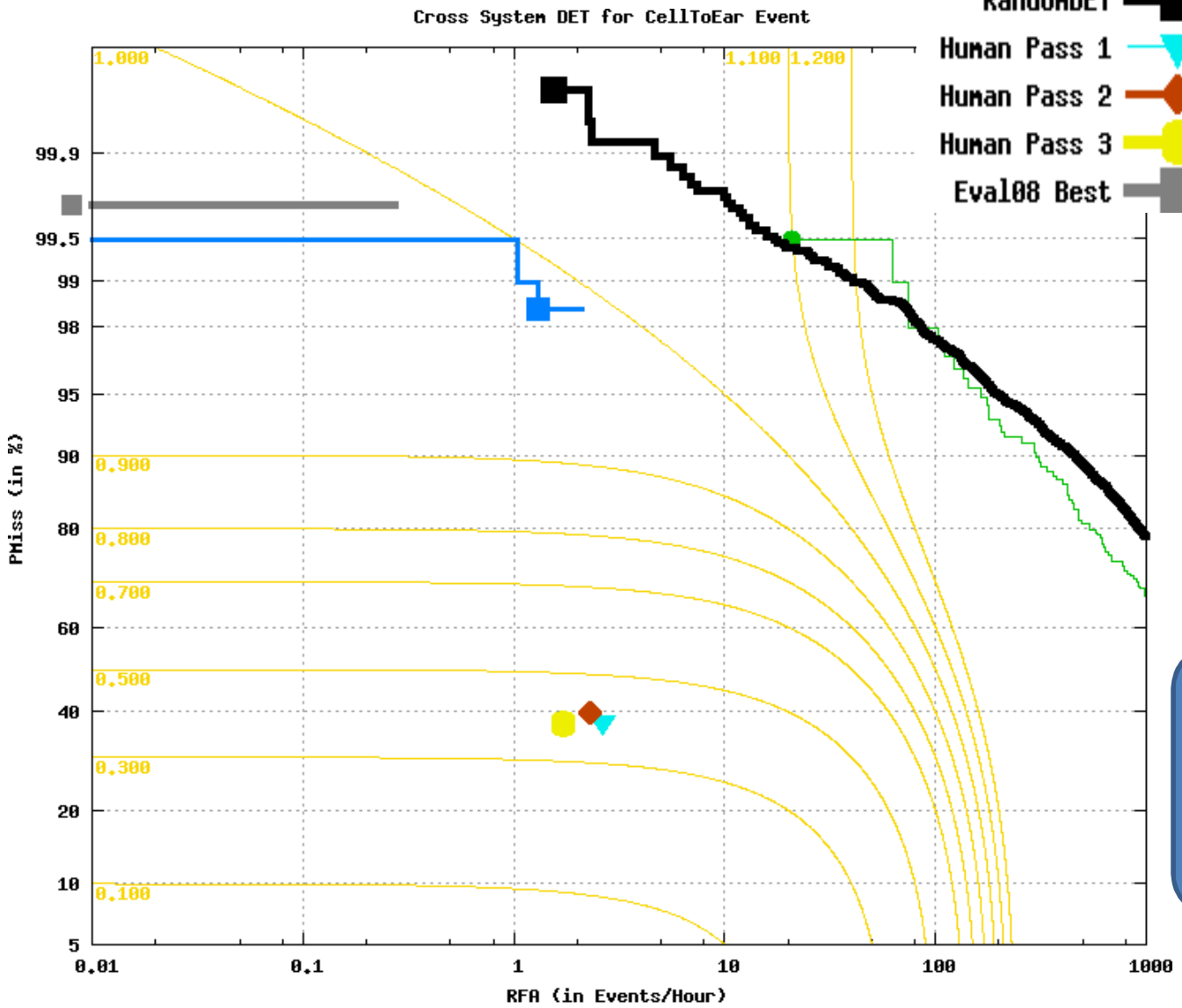
Best Submission per Site

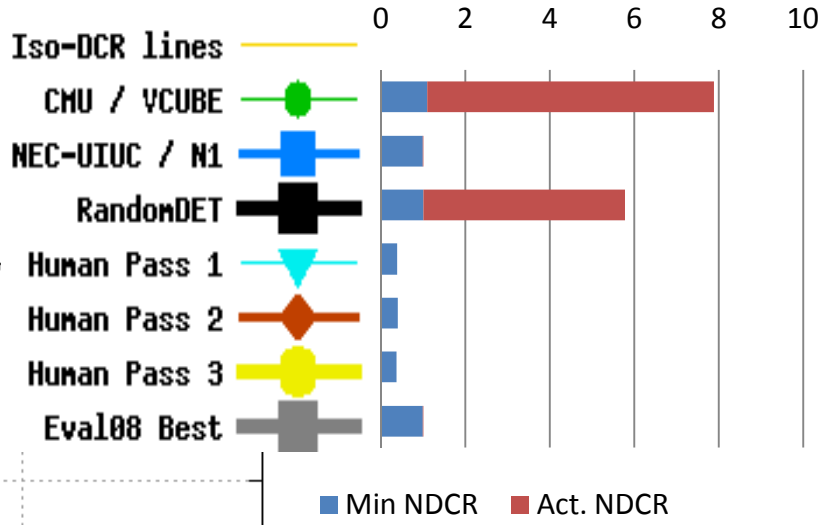




CellToEar Event

Best Submission per Site





- Random system
  - R_{targ}=XXX,
  - MeanDur=XXs
  - TestDur=XXH

ElevatorNoEntry Event

All Submissions