Event Detection in Airport Surveillance

The TRECVid 2008 Evaluation

Jerome Ajot, Jonathan Fiscus, John Garofolo
Martial Michel, Paul Over, Travis Rose, Mehmet Yilmaz
NIST

Heather Simpson, Stephanie Strassel
LDC

Video Analysis
Content Extraction

Outline

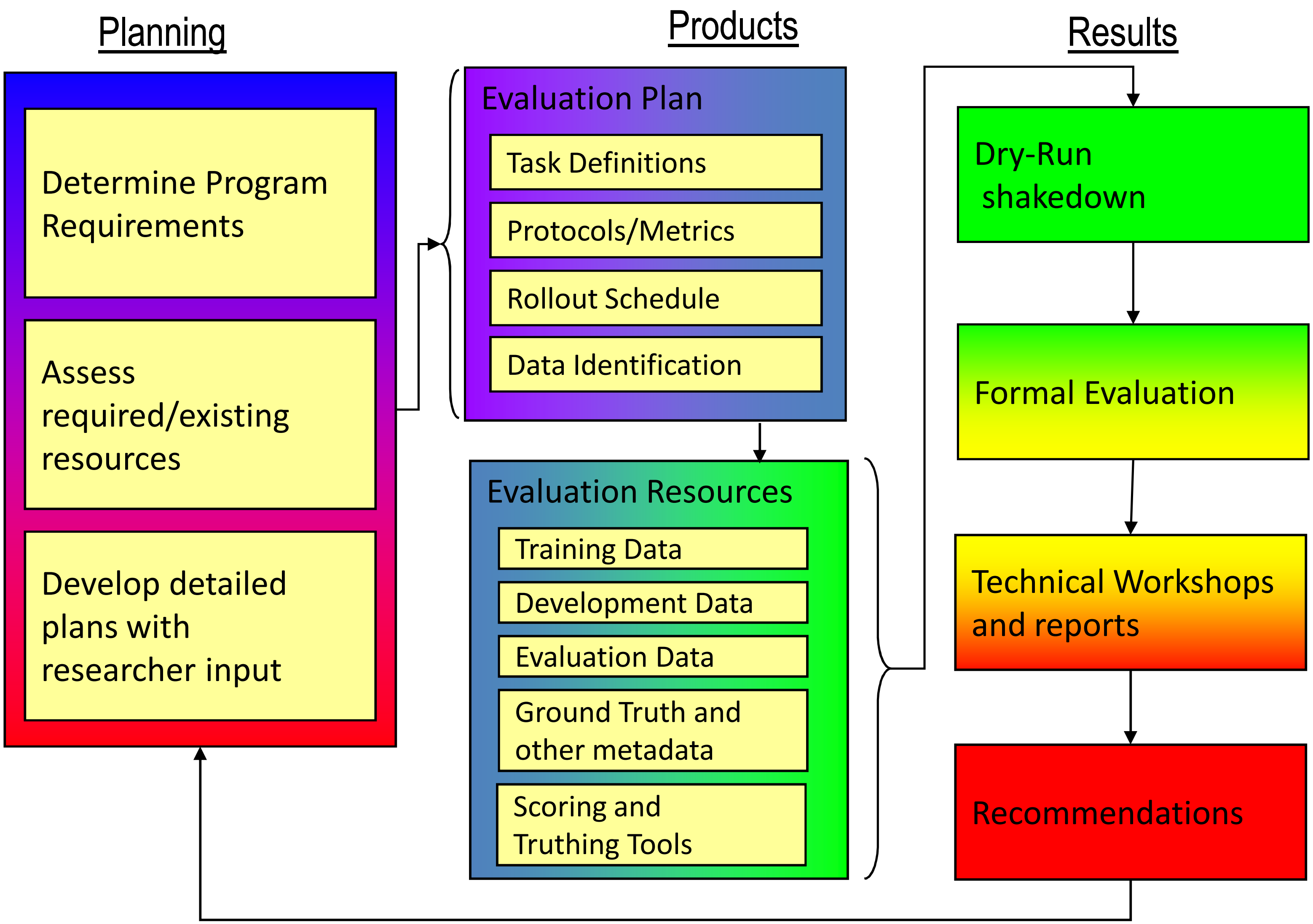
- Motivation
- Evaluation process
- Data
- Task definitions
- Events
- Annotation process
- Scoring
- Adjudication
- Conclusion & Future work

Motivation

- **Problem:** automatic detection of *observable* events 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.

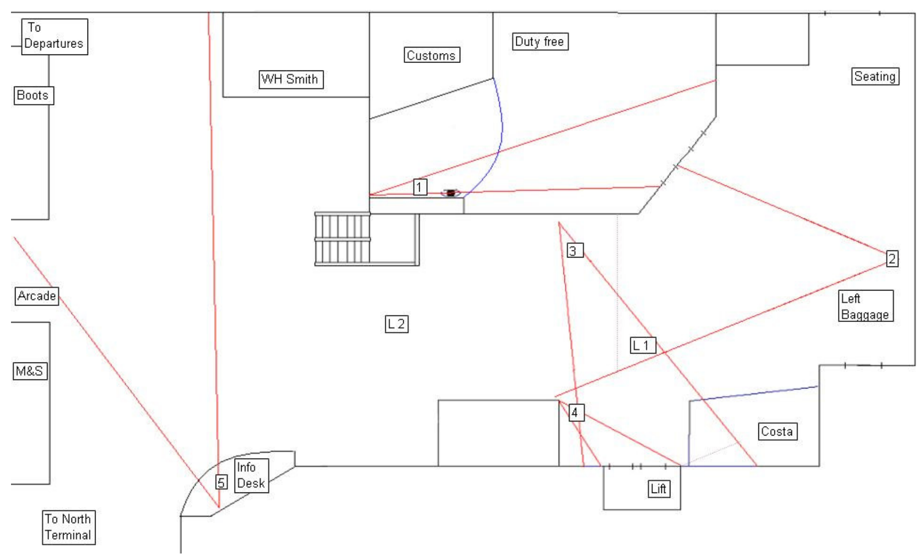
NIST Evaluation Process

Choosing the right task and metric is key



UK Home Office
London Gatwick Airport Data

- Home Office collected two parallel surveillance camera datasets
  - 1 for their multi-camera tracking evaluation
  - 1 for our event detection evaluation
- 100 hour event detection dataset
  - 10 data collection sessions
    * 2 hours per session
    * 5 cameras per session
- Camera views
  - Elevator close-up
  - 4 high traffic areas
  - Camera view features
    - Controlled access door
    - Some overlapping views
    - Areas with low pixels on target



TRECVid
Retrospective Event Detection

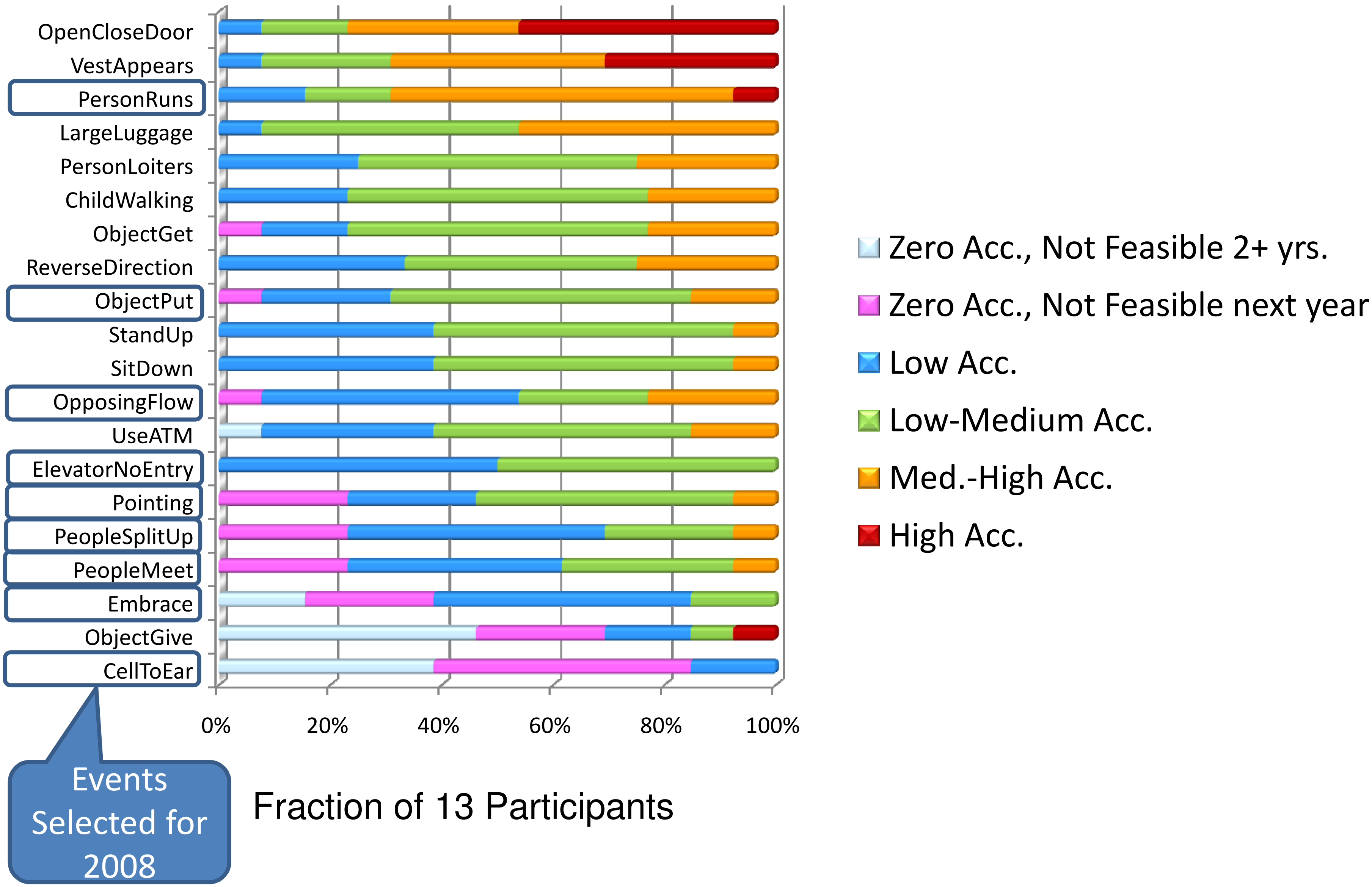
- Task:
  - Given a definition of an ***observable event*** involving humans, detect all occurrences of an event in ***airport surveillance video***
  - Identify each event observation by
    - The ***temporal extent***
    - A detection score indicating the strength of evidence
    - A binary decision on the detection score optimizing performance for a ***surrogate*** application

TRECVid
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

Technology Readiness Discussion Results

Benchmark detection accuracy across a variety of low occurrence events



Fraction of 13 Participants

Event Annotation Guidelines

- Jointly developed by:
  - NIST, Linguistic Data Consortium (LDC), Computer Vision Community
- Rules help users identify event observations
  - Reasonable Interpretation (RI) Rule
    - If according to a reasonable interpretation of the video, the event must have occurred, then it is a taggable event
  - Start/Stop times for occlusion
    - Observations with “occluded start times” begin with the occlusion or frame boundary
    - Observations with “occluded end times” end with the occlusion or frame boundary
    - Frame boundaries are occlusions, but the existence of the event still follows the RI Rule
- Event Definitions left minimal to capture human intuitions
  - Contrast with highly defined annotation tasks such as ACE

Annotator Training

- Training session with lead annotator to introduce task and guidelines
- Complete 1-3 practice files
  - Tool functionality
  - Data and camera views
  - Annotation decisions and rules of thumb
- Regular team meetings for ongoing training
- Annotator mailing list to resolve challenging examples
  - Usually matter of reinforcing basic principles – “How would you describe this event to someone else?”
- Decisions logged to LDC wiki for annotator reference
- NIST input sought on issues that could not be resolved locally

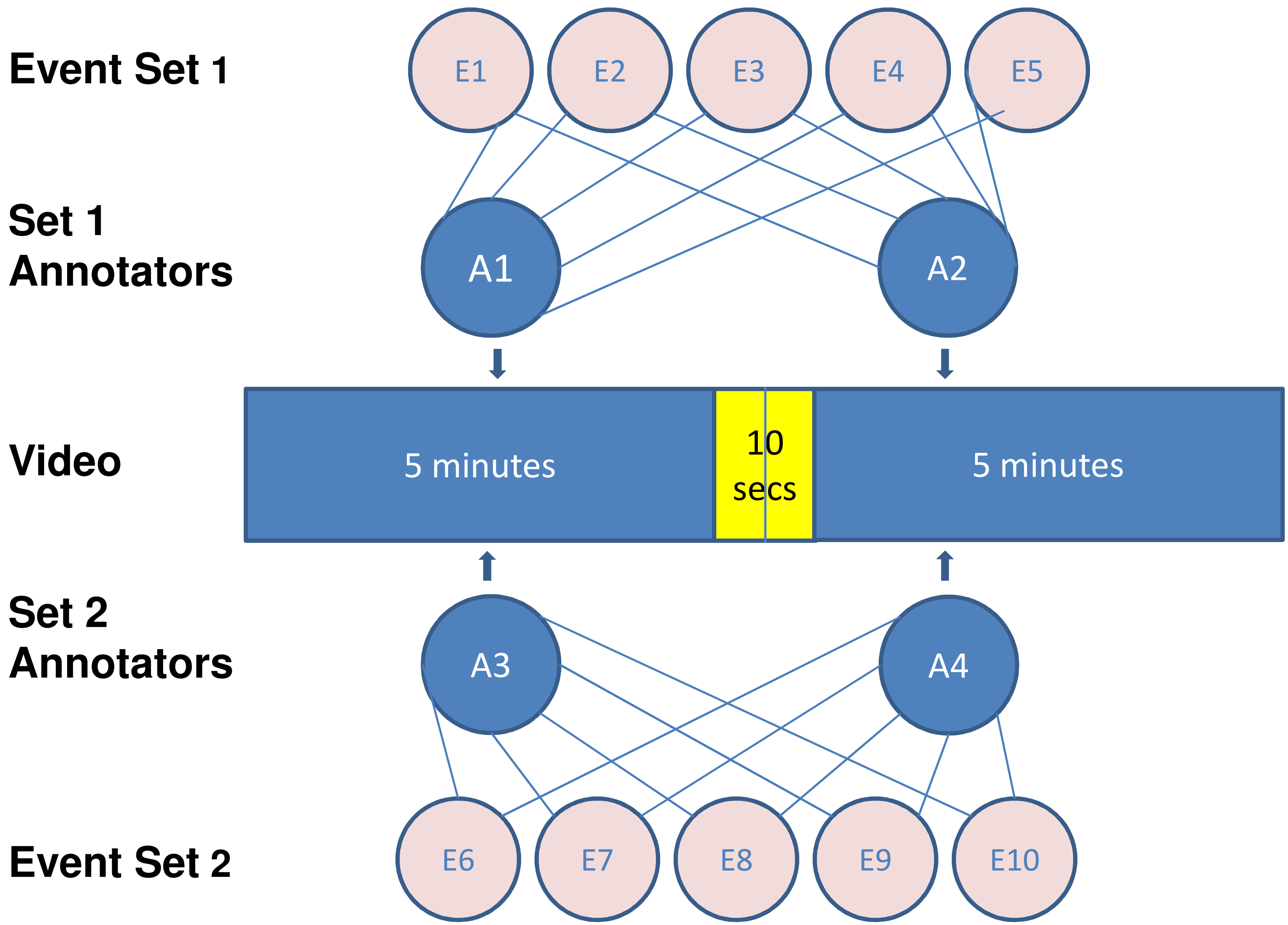
Annotation Tool and Data Processing

- 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

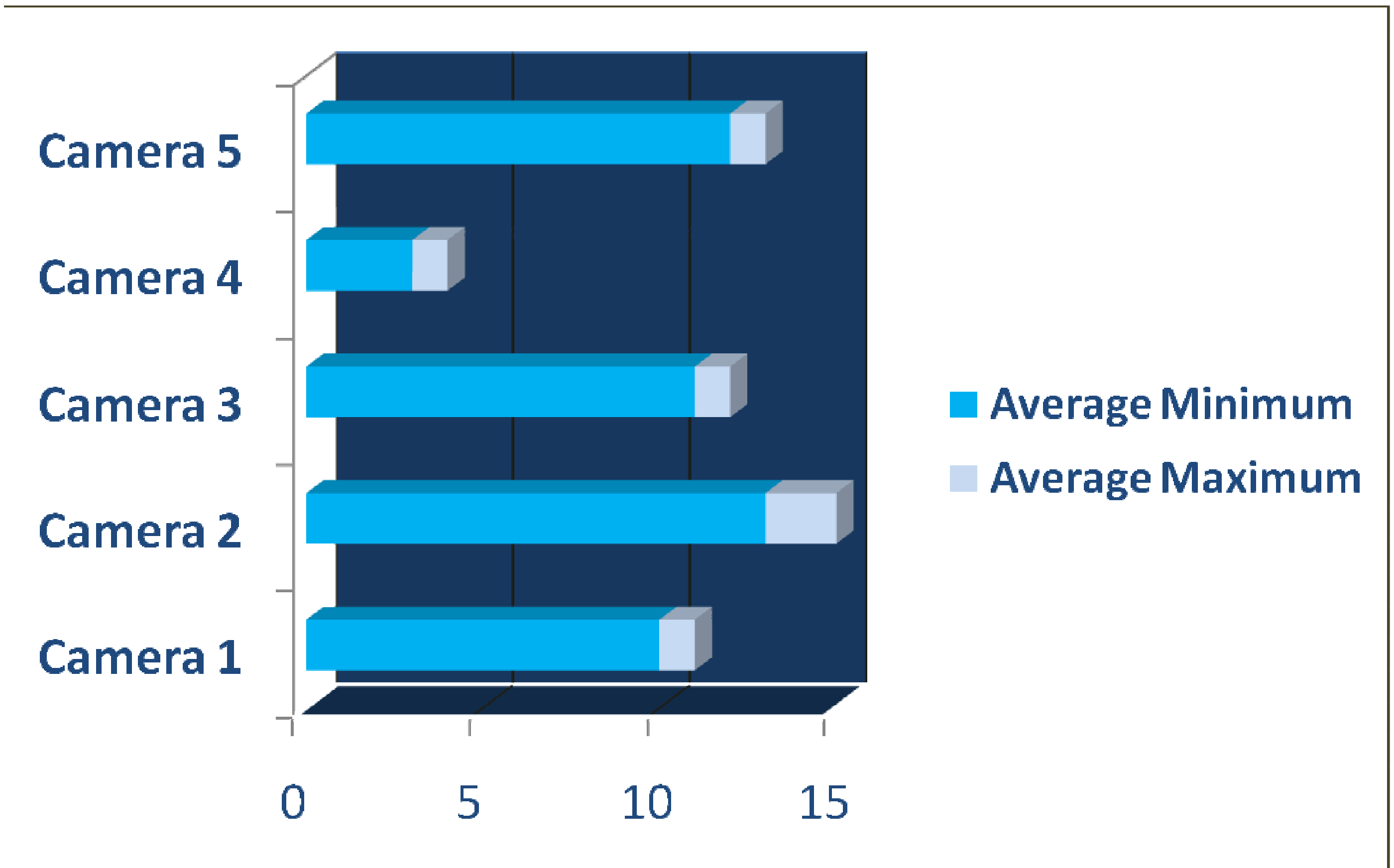
- Pilot study to determine optimal balance of clip duration and number of events per work session
- Source data divided into 5m 10s clips
  - 10s = 5s of overlap with the preceding and following clips
- Events divided into 2 sets of 5
  - Set 1: PersonRun, CellToEar, ObjectPut, Pointing, ElevatorNoEntry
  - Set 2: PeopleMeet, PeopleSplitUp, Embrace, OpposingFlow, TakePicture
- For each assigned clip + event set, detect any event occurrence and label its temporal extent
- 5% of devtest set dually annotated (double-blind) to establish baseline IAA and permit consistency analysis

Visualization of Annotation Workflow



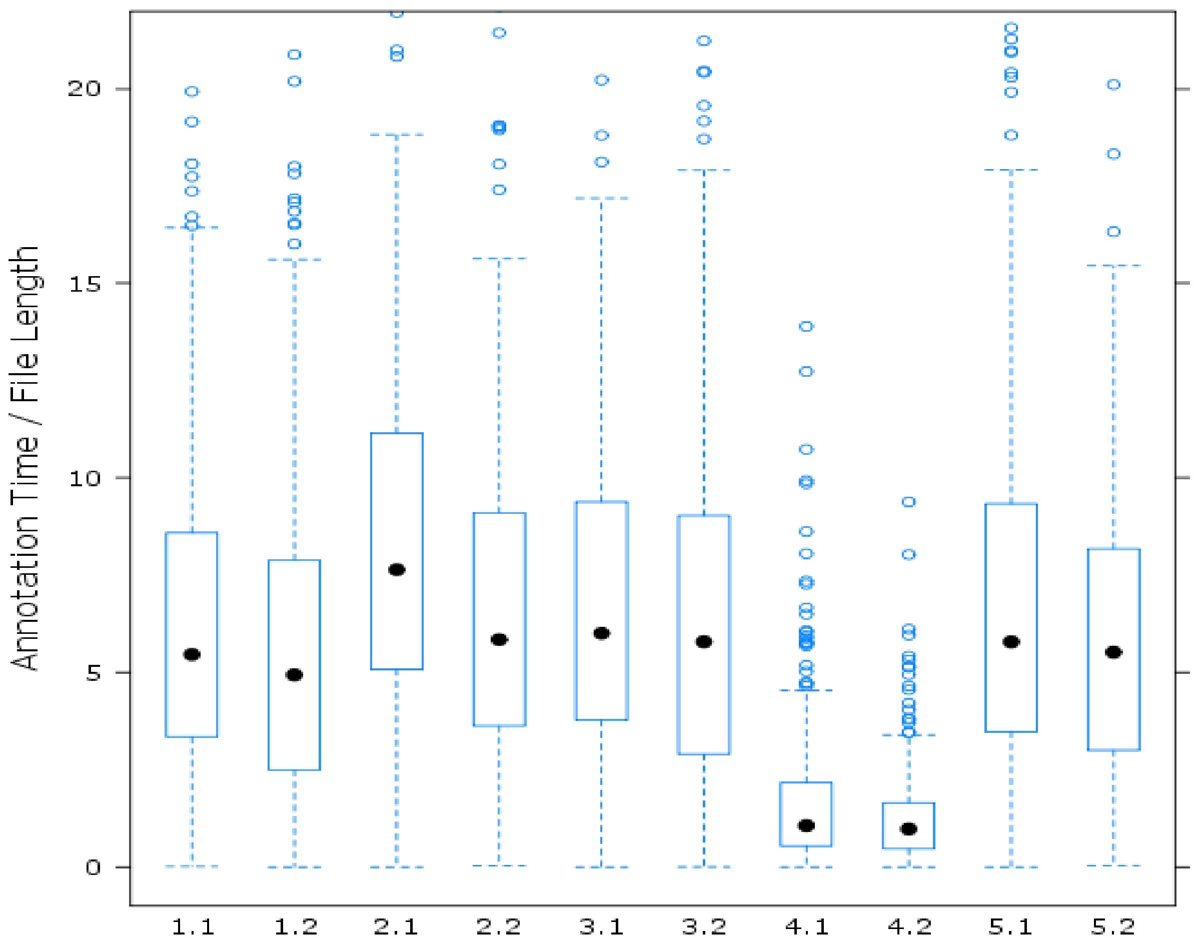
Annotation Rates

- Average 10-15 x Real Time
  - i.e. 50-75 mins per 5m clip, with 5 events under consideration per clip
- Annotation rates heavily conditioned by camera view



Annotation Rates

- Average 6-9 x Real Time (10x-15x Real Time including upper outliers)
  - i.e. 31-46.5 mins per 5m clip, with 5 events under consideration per clip
- Annotation rates heavily conditioned by camera view



Annotation Challenges

- Ambiguity of guidelines
  - Loosely defined guidelines tap into human intuition instead of forcing real world data 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
- Technical issues

Example Observations

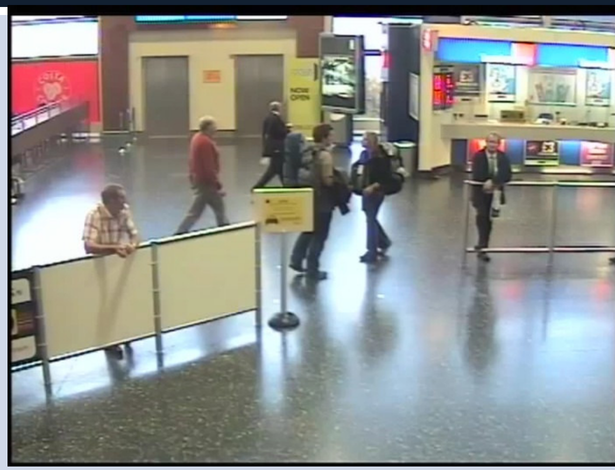
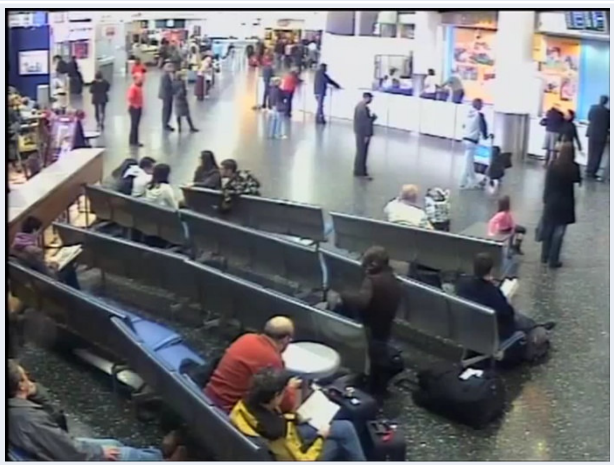
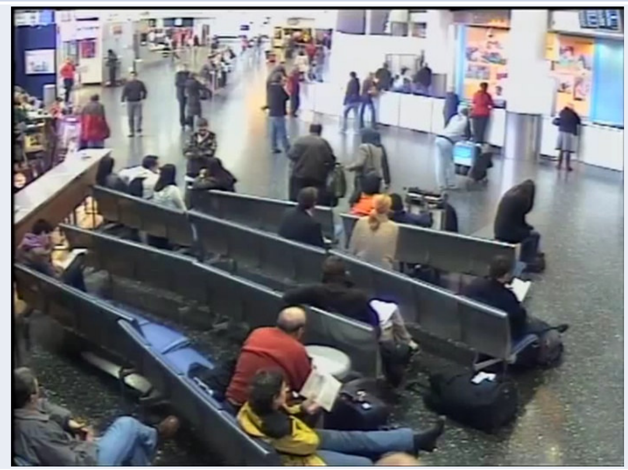
	Easy to Find Example	Hard to Find Example
Pointing		
Embrace		

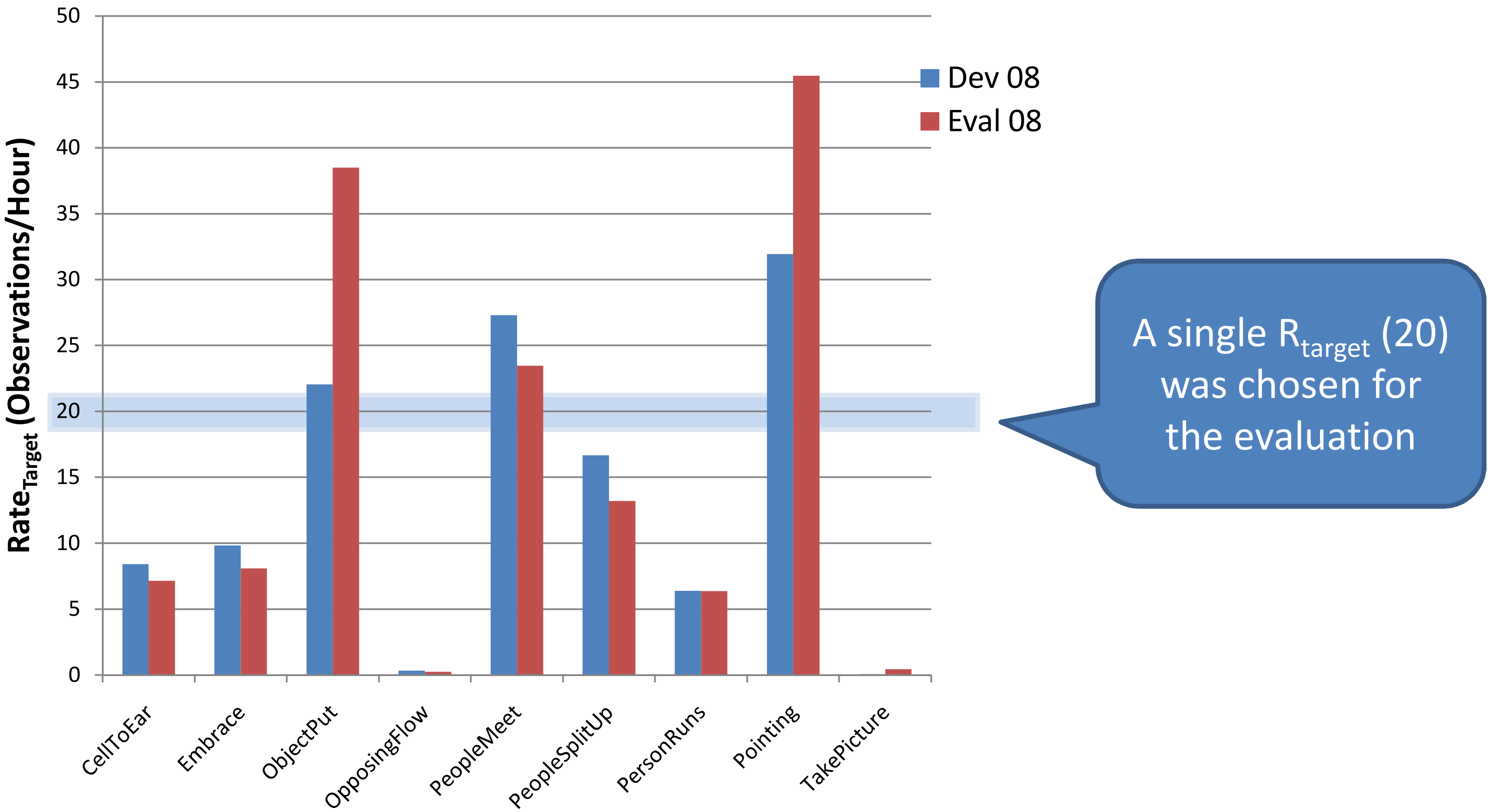
Table of Participants Vs Events

	Cell To Ear	Elevator NoEntry	Embrace	ObjectPut	Opposing Flow	People Meet	People Split Up	Person Runs	Pointing	Take Picture
AIT		x			x			x		
BUT		x		x	x			x		
CMU	x	x	x	x	x	x	x	x	x	x
DCU		x	x		x	x		x		
FD					x			x		x
IFP-UIUC-NEC	x	x	x	x	x	x	x	x	x	x
Intuvision		x			x					x
MCG-ICT-CAS		x	x		x	x	x	x		x
NHKSTRL		x			x			x		
QMUL-ACTIVA		x			x			x		
SJTU		x			x	x		x	x	
THU-MNL	x				x			x		
TokyoTech						x	x	x		
Toshiba		x			x			x		
UAM				x	x			x		
UCF				x	x			x		x
Total	**3**	**11**	**4**	**5**	**15**	**6**	**4**	**15**	**3**	**6**

- 16 Sites
- 72 Event Runs

Rates of Event Observations

Development vs. Evaluation data

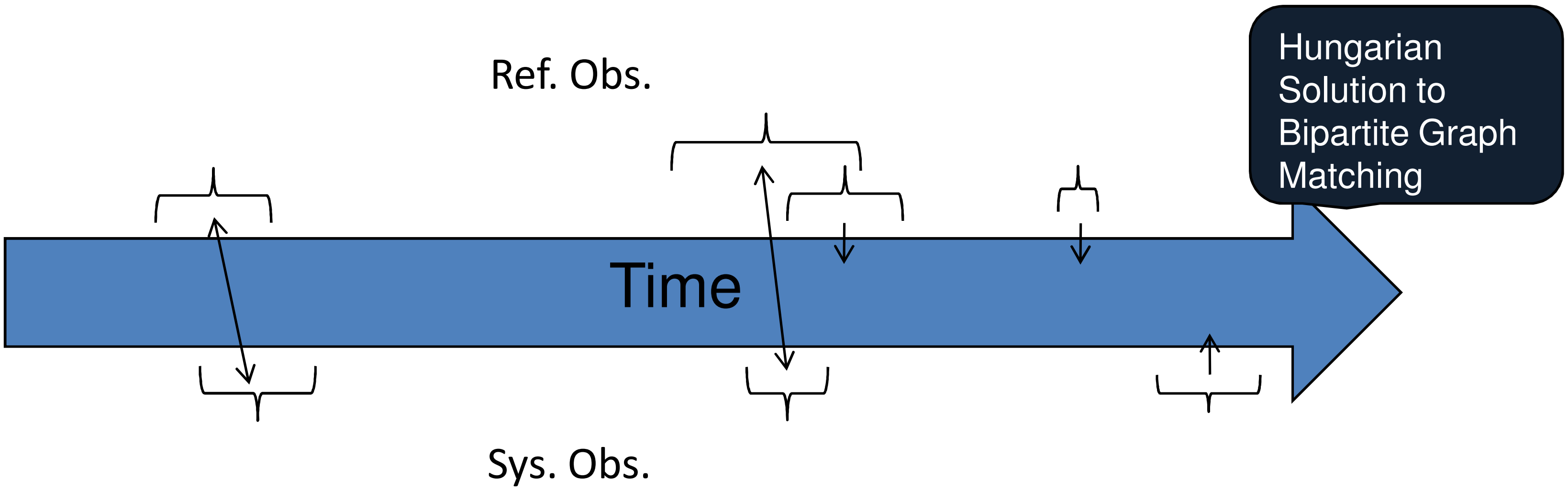


Evaluation Protocol Synopsis

- NIST used the Framework for Detection Evaluation (F4DE) Toolkit
  - Available for download on the Event Detection Web Site
- Events are independent for eval. purposes
- Two step evaluation process
  - System observations are “aligned” to reference observations
  - Detection performance is a tradeoff between missed detections and false alarms

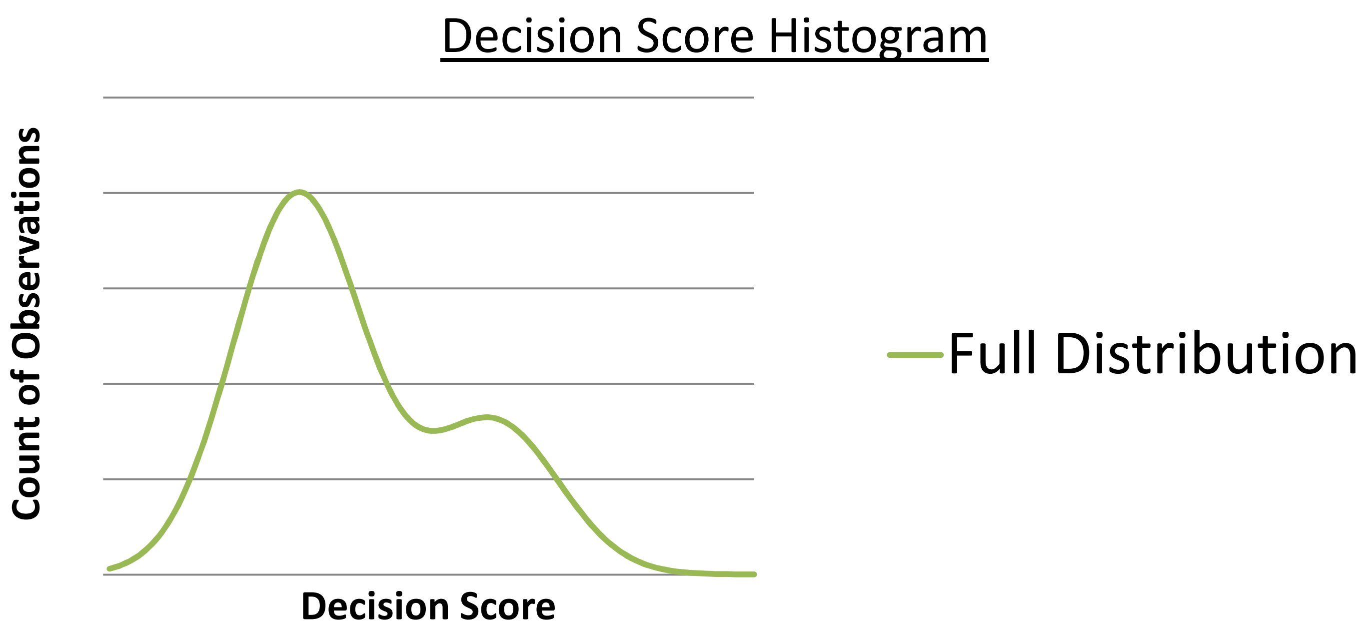
- Two methods of evaluating performance
  - Decision Error Tradeoff curves graphically depict performance
  - A “Surrogate Application”: Normalized Detection Cost Rate
    - *A priori* application requirements unknown
    - Optimization to be achieved using a “System Value Function”

Temporal Alignment for Detection in Streaming Media



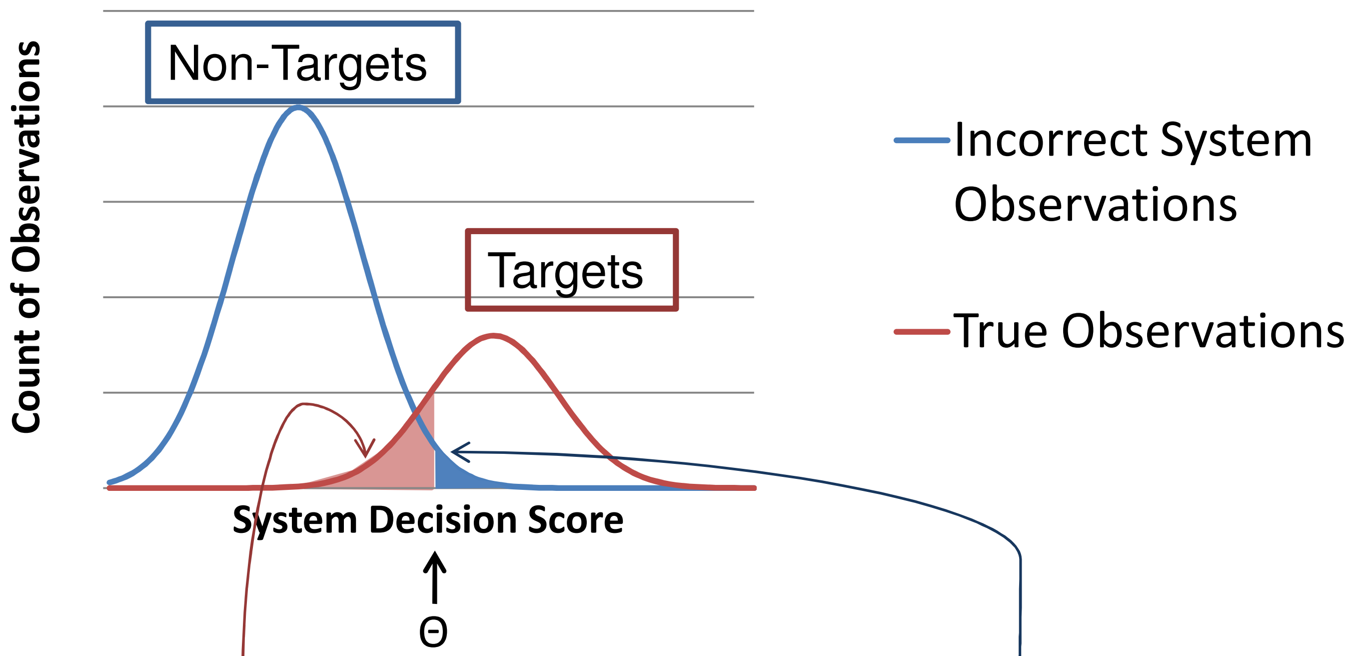
- Mapping Alignment Rules
  - Mid point of system with Δ*t* of reference extent
  - Temporal congruence and decision scores give preference to overlapping events

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



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

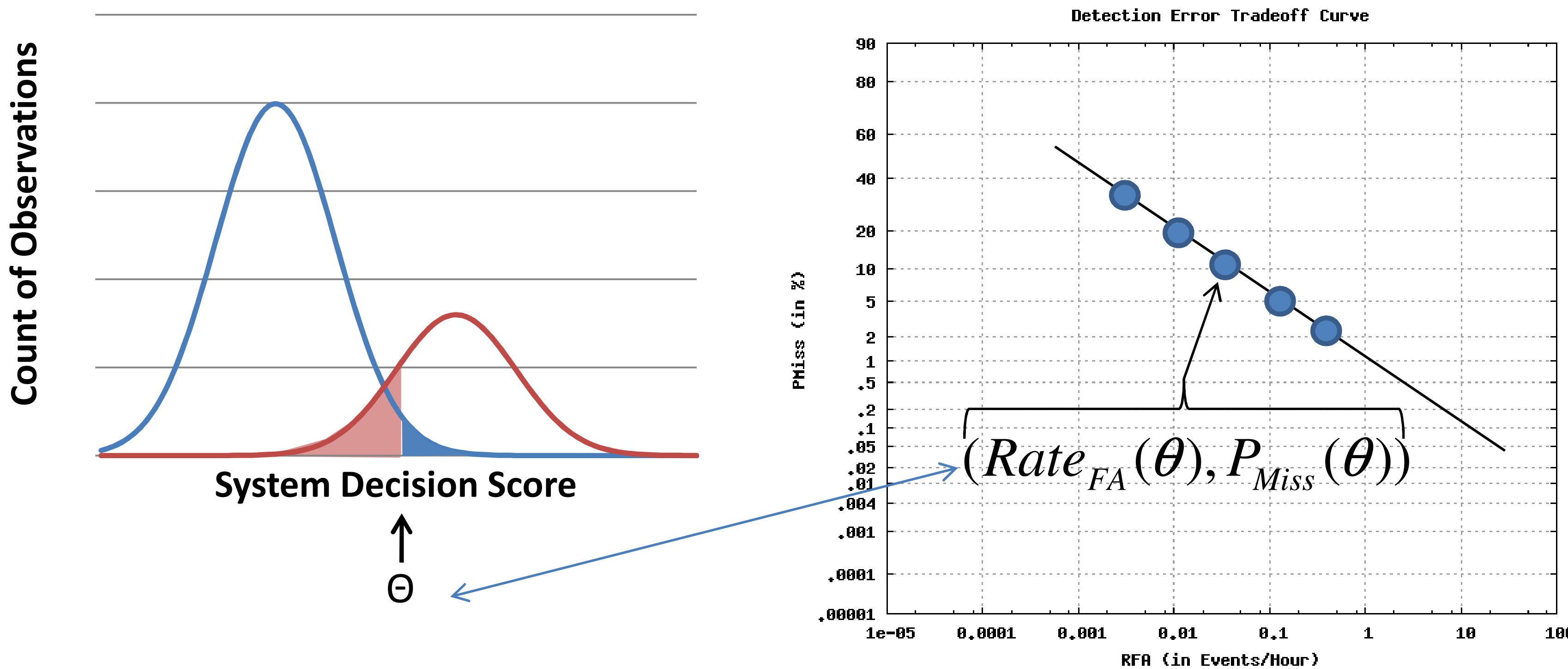
$$Rate_{FA}(\theta) = \frac{\#FalseAlarms}{SignalDuration}$$

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

Decision Error Tradeoff Curves

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

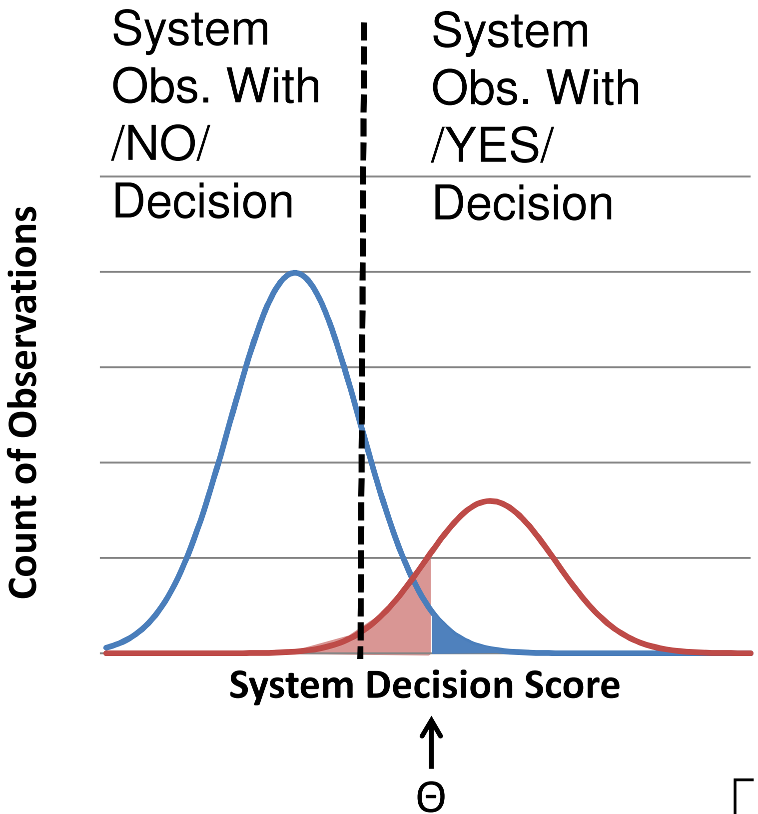
Compute $Rate_{FA}$ and P_{Miss} for all Θ



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

Decision Error Tradeoff Curves

Actual vs. Minimum NDCR



Event Detection Constants

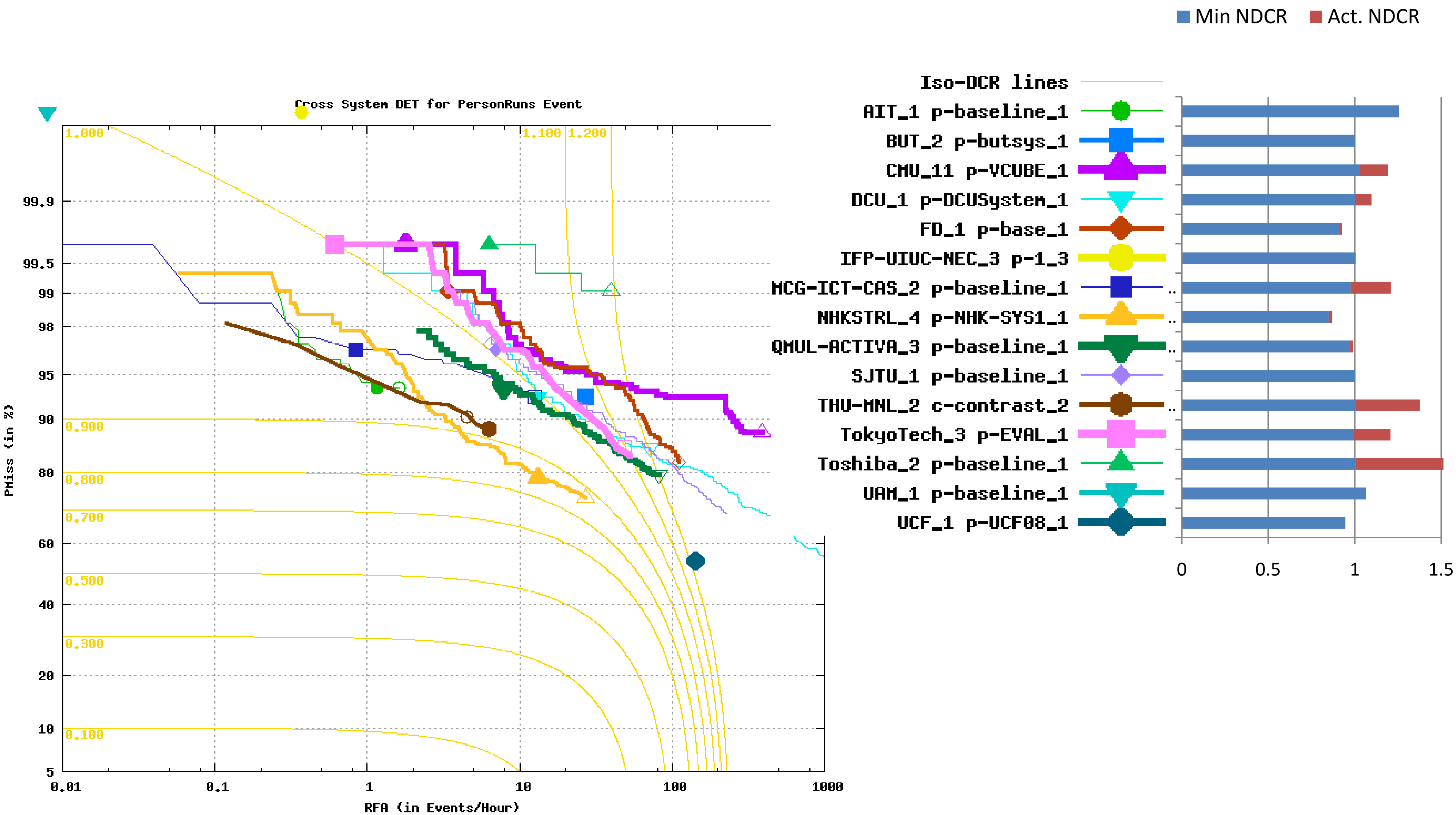
$$Cost_{Miss} = 10$$
$$Cost_{FA} = 1$$
$$R_{Target} = 20$$

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

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

PersonRuns Event

Best Submission per Site



Estimating Human Error Rates:

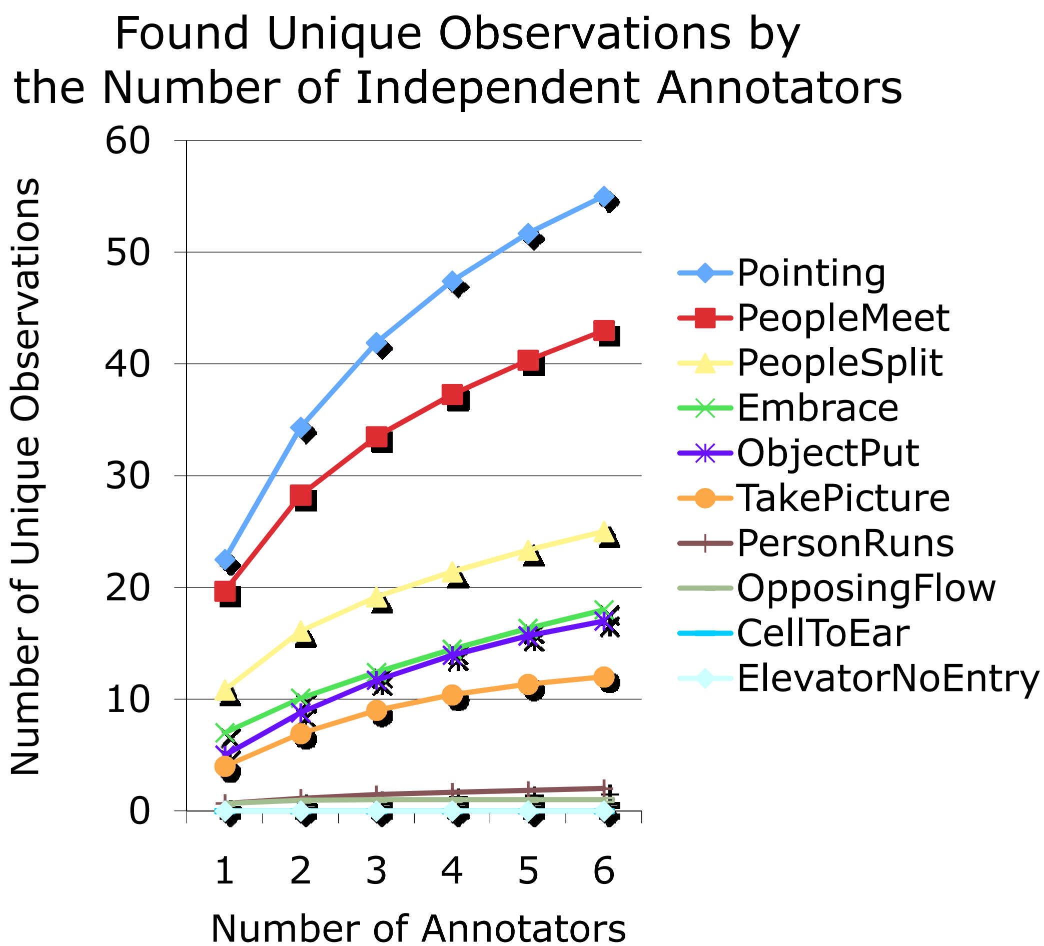
6-Way Annotation Study

- LDC create 6 independent annotations for each excerpt

Caveats of the experiment

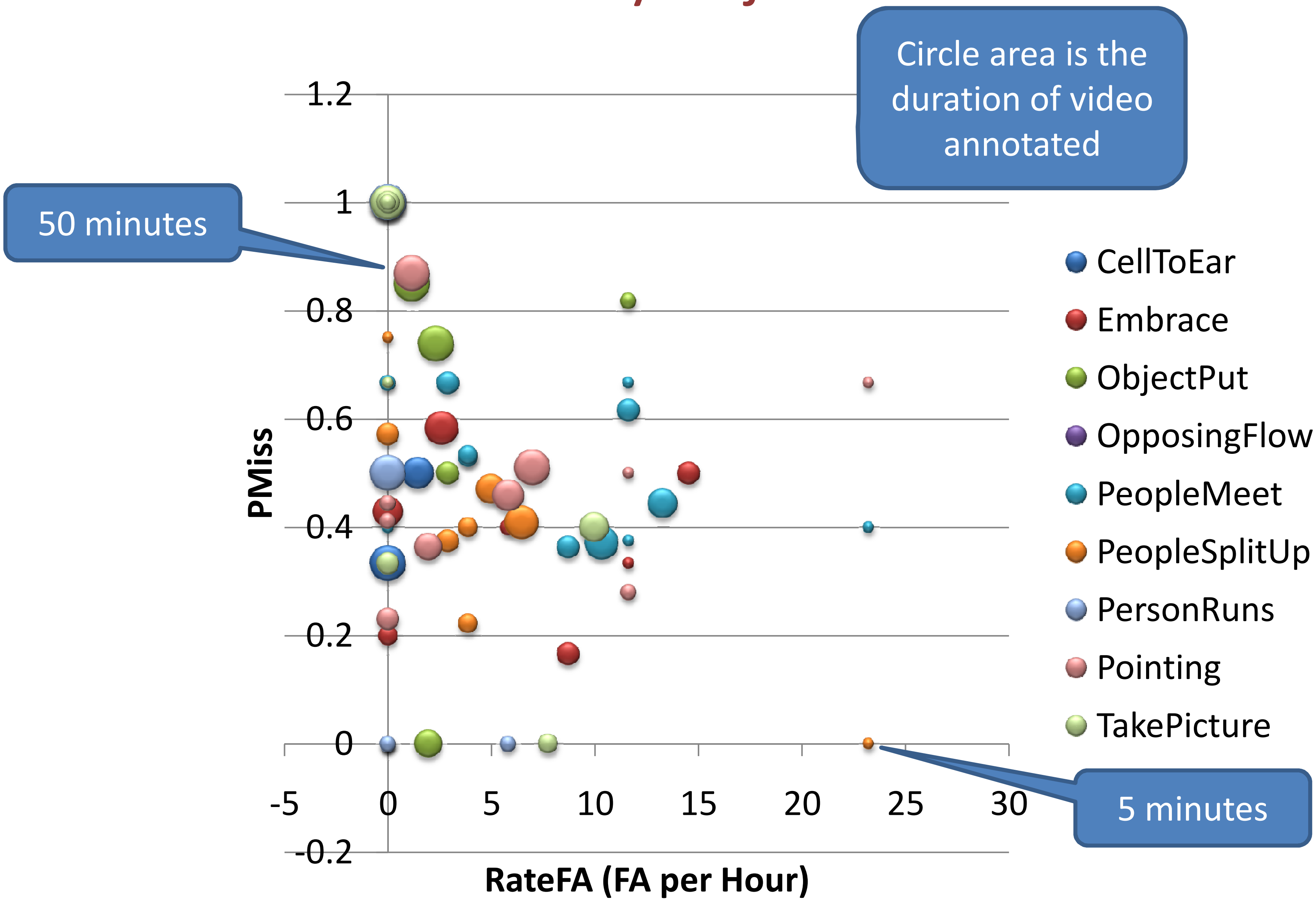
  - Not balanced by events
  - Not balanced by annotators

- Blindly merge all annotations
  - Use evaluation code to iteratively merge annotations
  - Commonly detected observations counted once
- Analysis:
  - Curves follow published studies on finding software bugs*
  - Curves suggest more annotation is needed for some events but False Alarms haven't been accounted for
  - LDC reviewed all observed events (100% Adjudication)



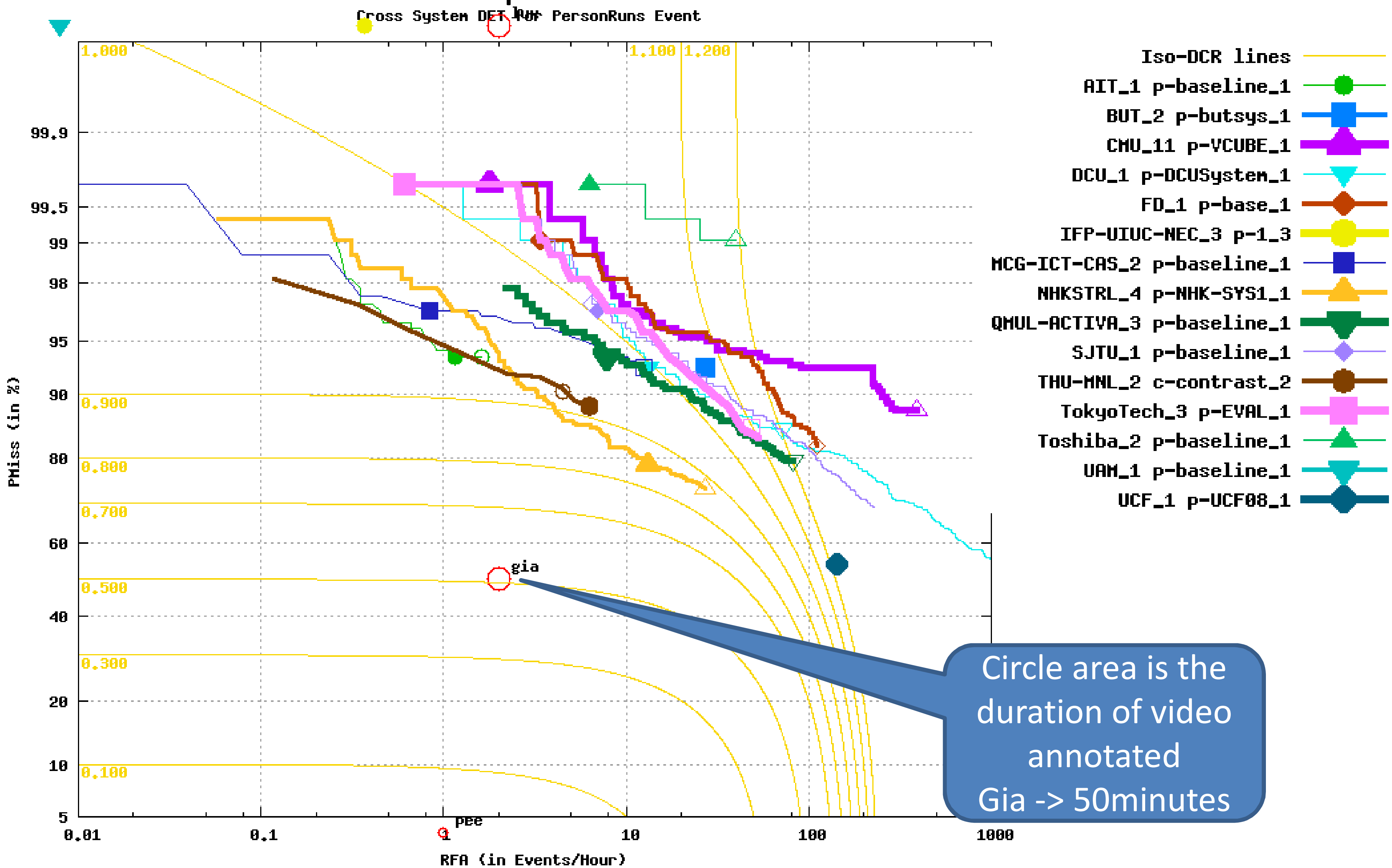
* Nielsen and Landauer: "A Mathematical Model of Finding Usability Problems"

Estimating Human Error Rates:
Humans vs. 6-Way Adjudicated References



PersonRuns Event

Best Submission per Site with Human Error Estimates

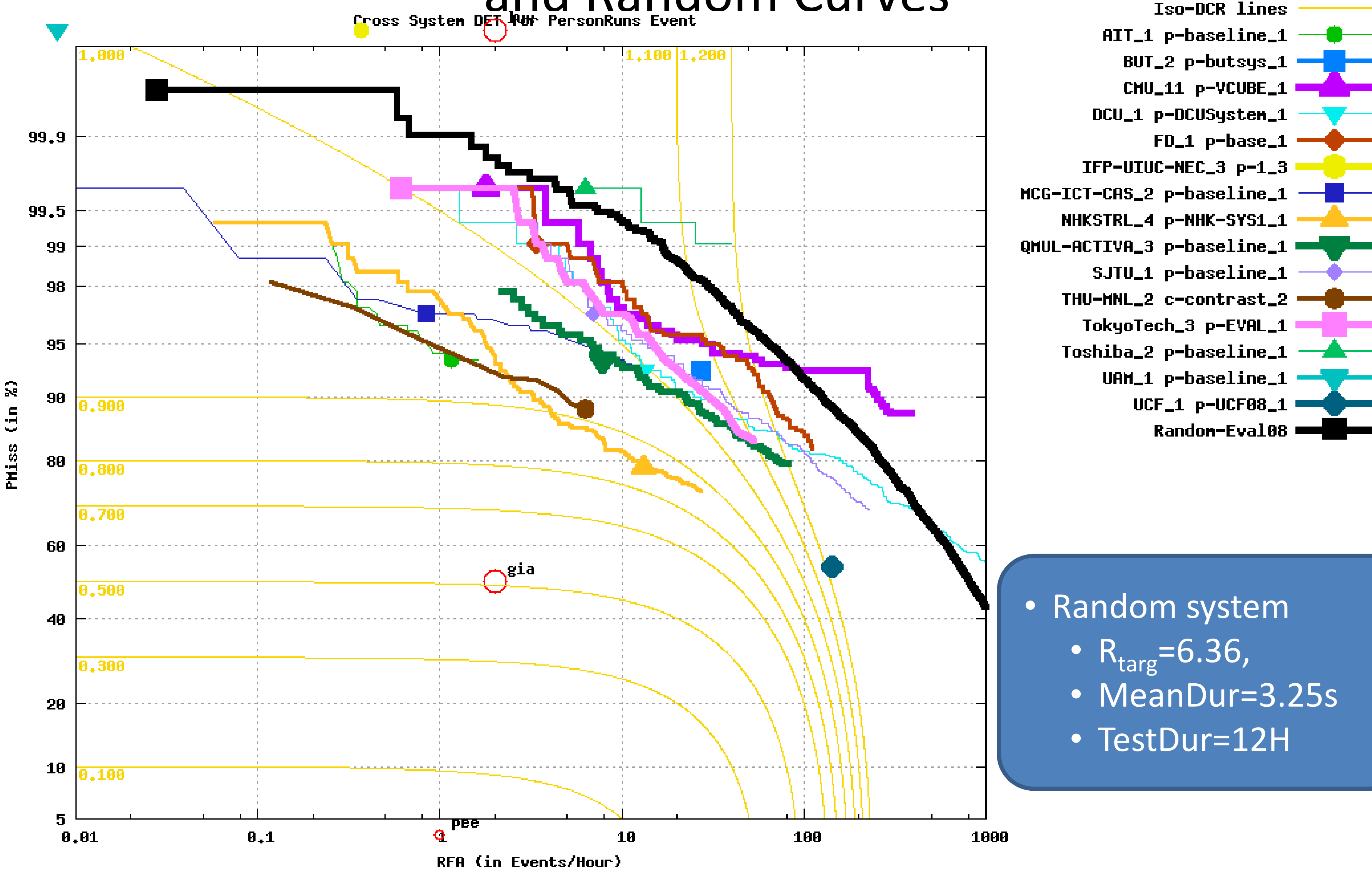


Random DET Curves for Streaming Detection Evaluations

- Parametric random curves are not possible
  - Due to un-countable non-target trials
  - Monte Carlo simulation is a feasible method
- Monte Carlo Random DET Curves
  - Two factors influence a random system
    - R_{Target} -- Primary effect
    - Observation duration statistics -- Secondary effect
      - Distribution measurements: Mean, Standard Deviation, etc.
  - Test set size computation (Rule of 30 @ 40% P_{miss})
    - #Hours = 30 errs / .4 (Pmiss) / R_{Target}
  - Our procedure:
    1. Measure R_{target} and Mean Duration of observations in the eval set
    2. Construct 50 pairs of a random test set and system output with decision scores from a uniform random distribution, 1000 system obs./hour
    3. Compute an ref/sys pair-averaged, DET Curve

PersonRuns Event

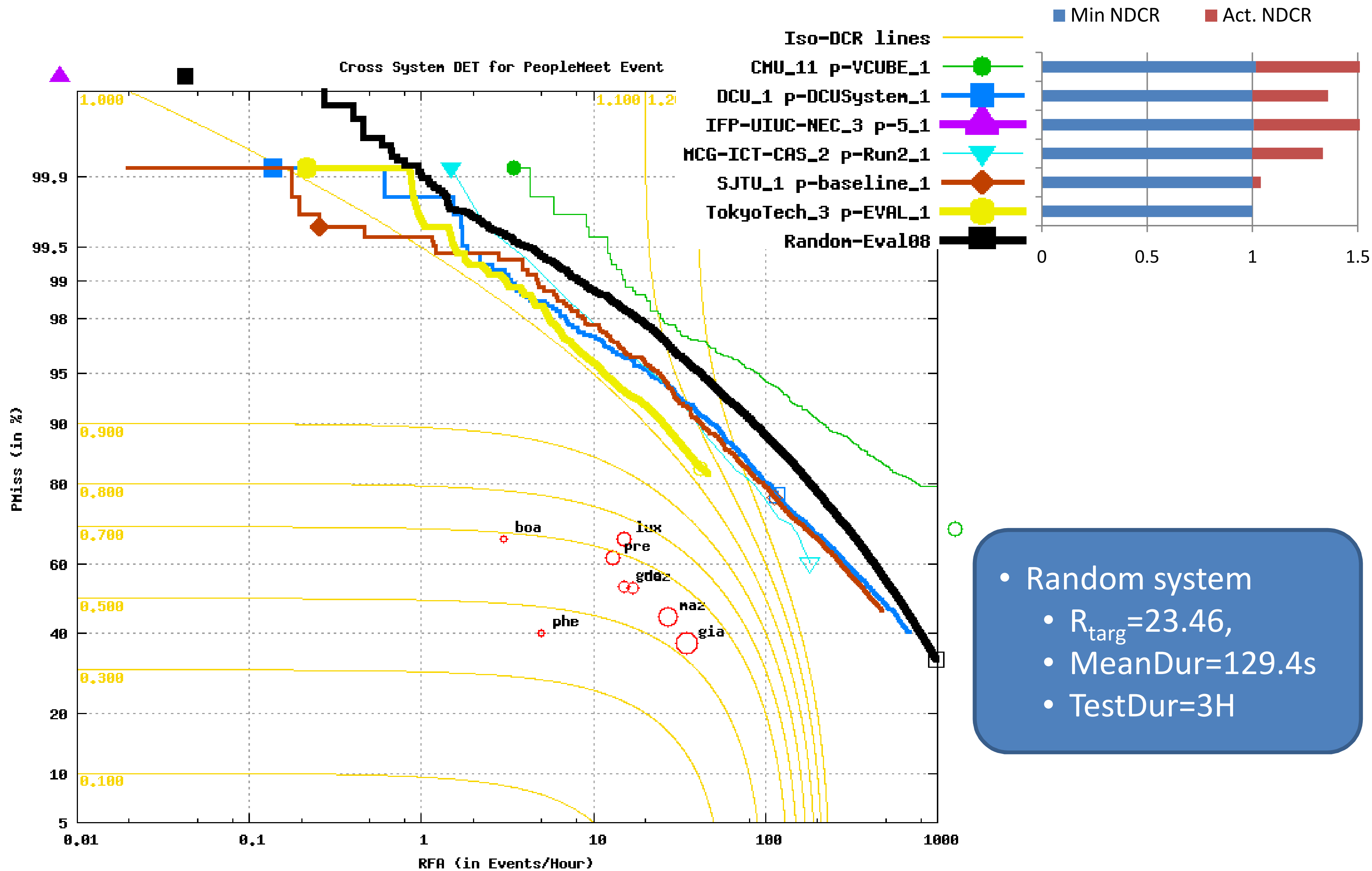
Best Submission per Site with Human Error Estimates and Random Curves



- Random system
  - R_{targ}=6.36,
  - MeanDur=3.25s
  - TestDur=12H

PeopleMeet Event

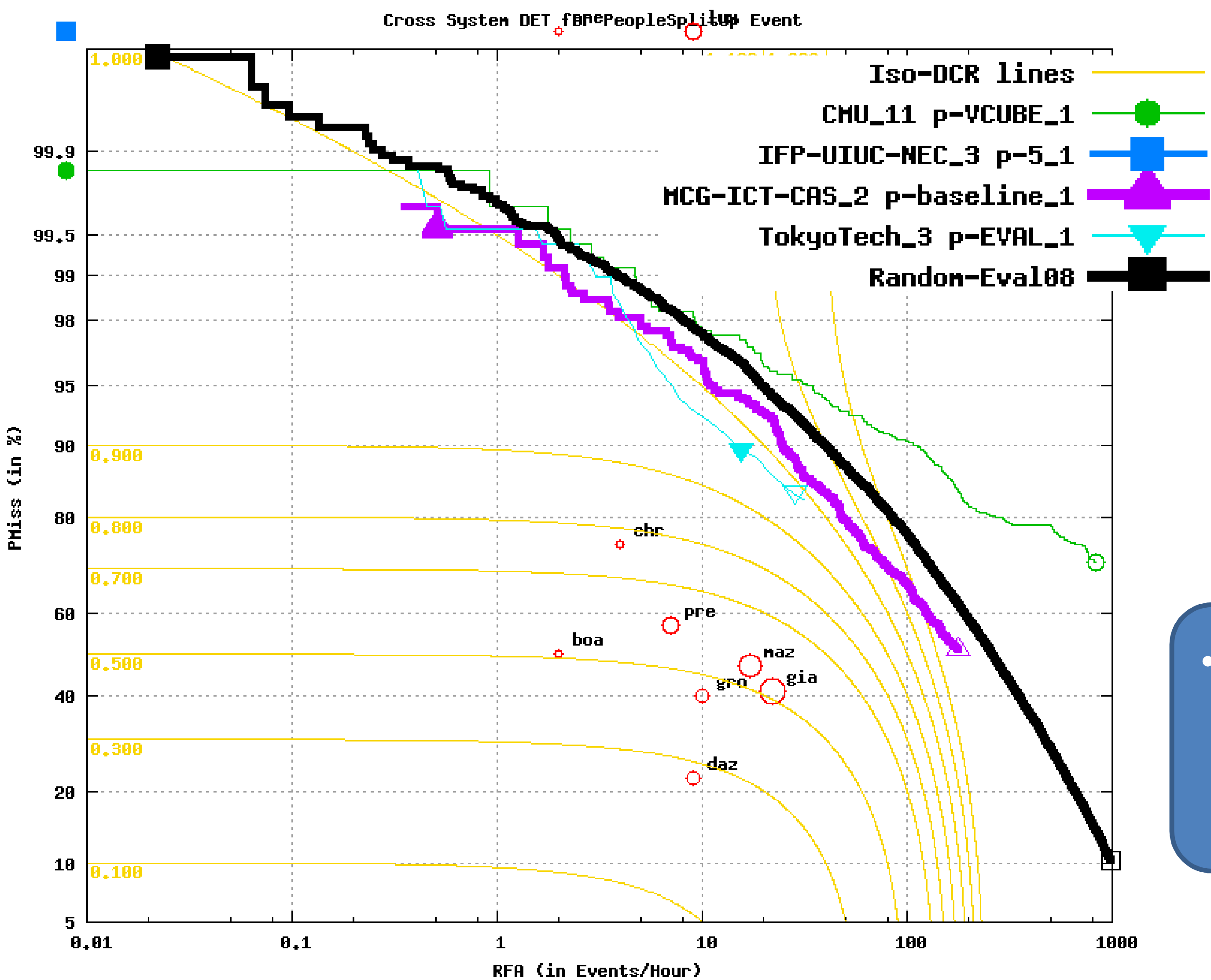
Best Submission per Site

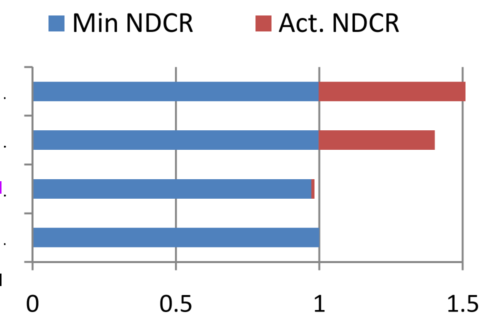


- Random system
  - R_{targ}=23.46,
  - MeanDur=129.4s
  - TestDur=3H

PeopleSplitUp Event

Best Submission per Site

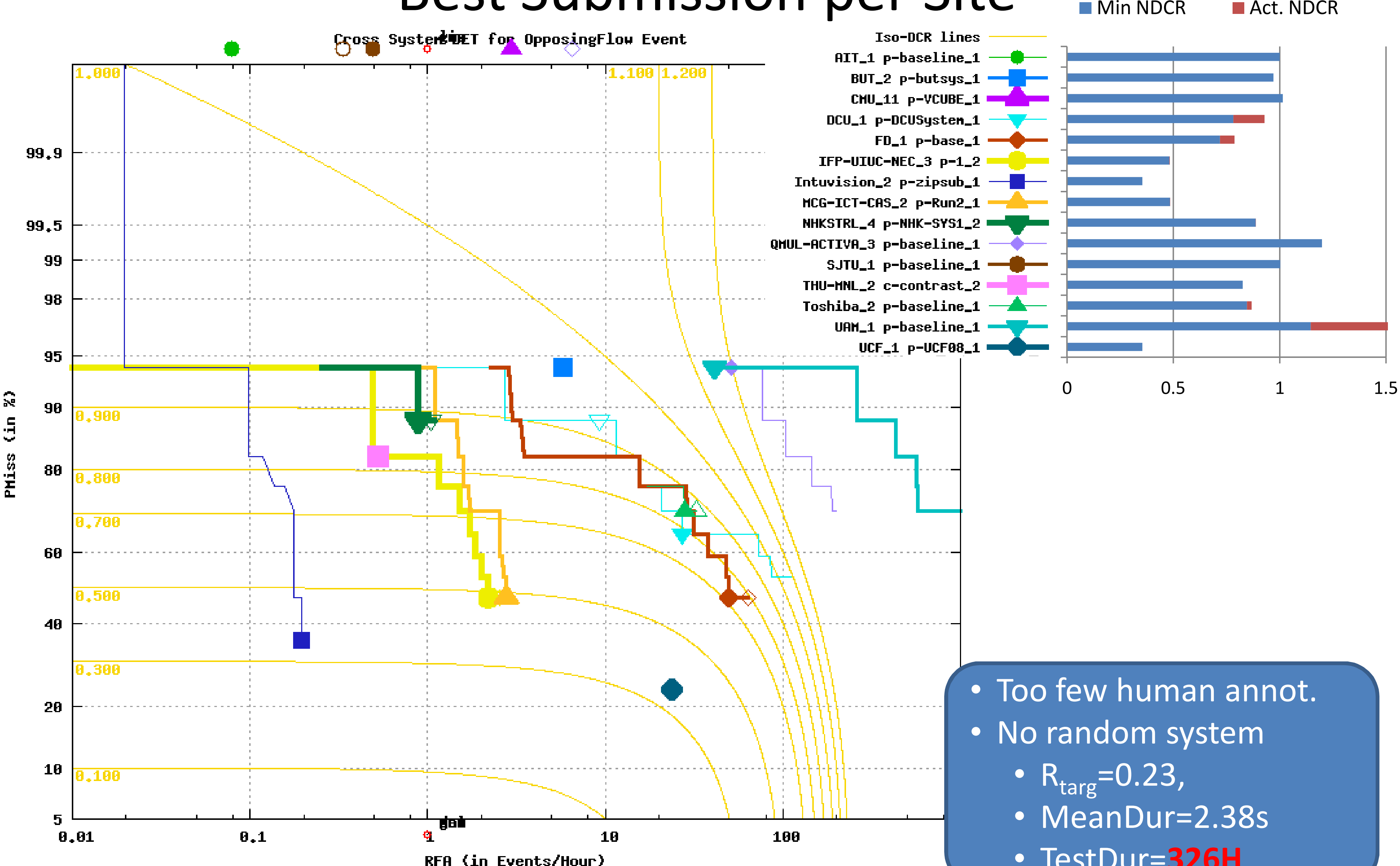




- Random system
  - R_{targ}=13.2,
  - MeanDur=11.66s
  - TestDur=6H

Opposing Flow Event

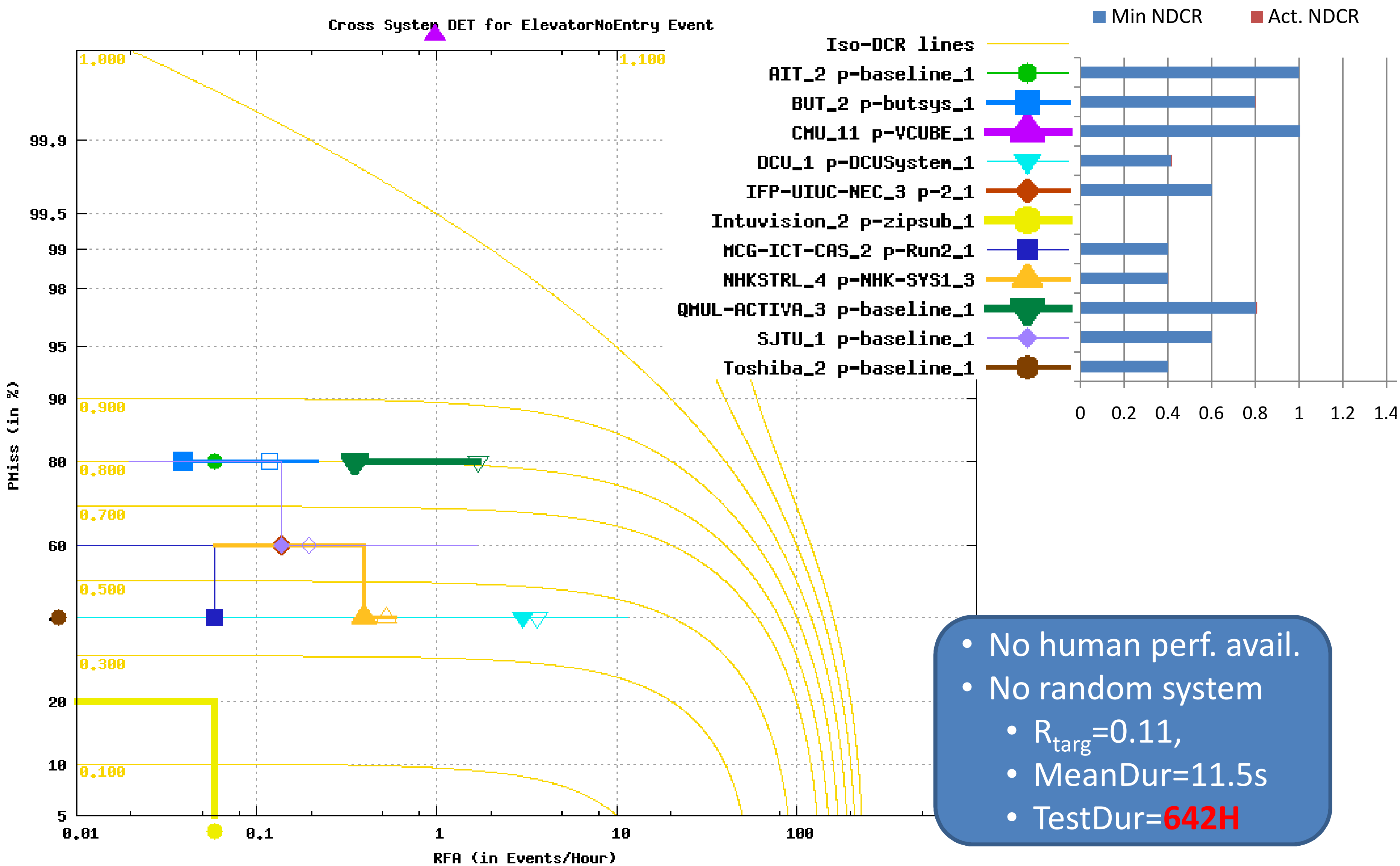
Best Submission per Site



- Too few human annot.
- No random system
  - R_{targ}=0.23,
  - MeanDur=2.38s
  - TestDur=326H

Elevator No Entry Event

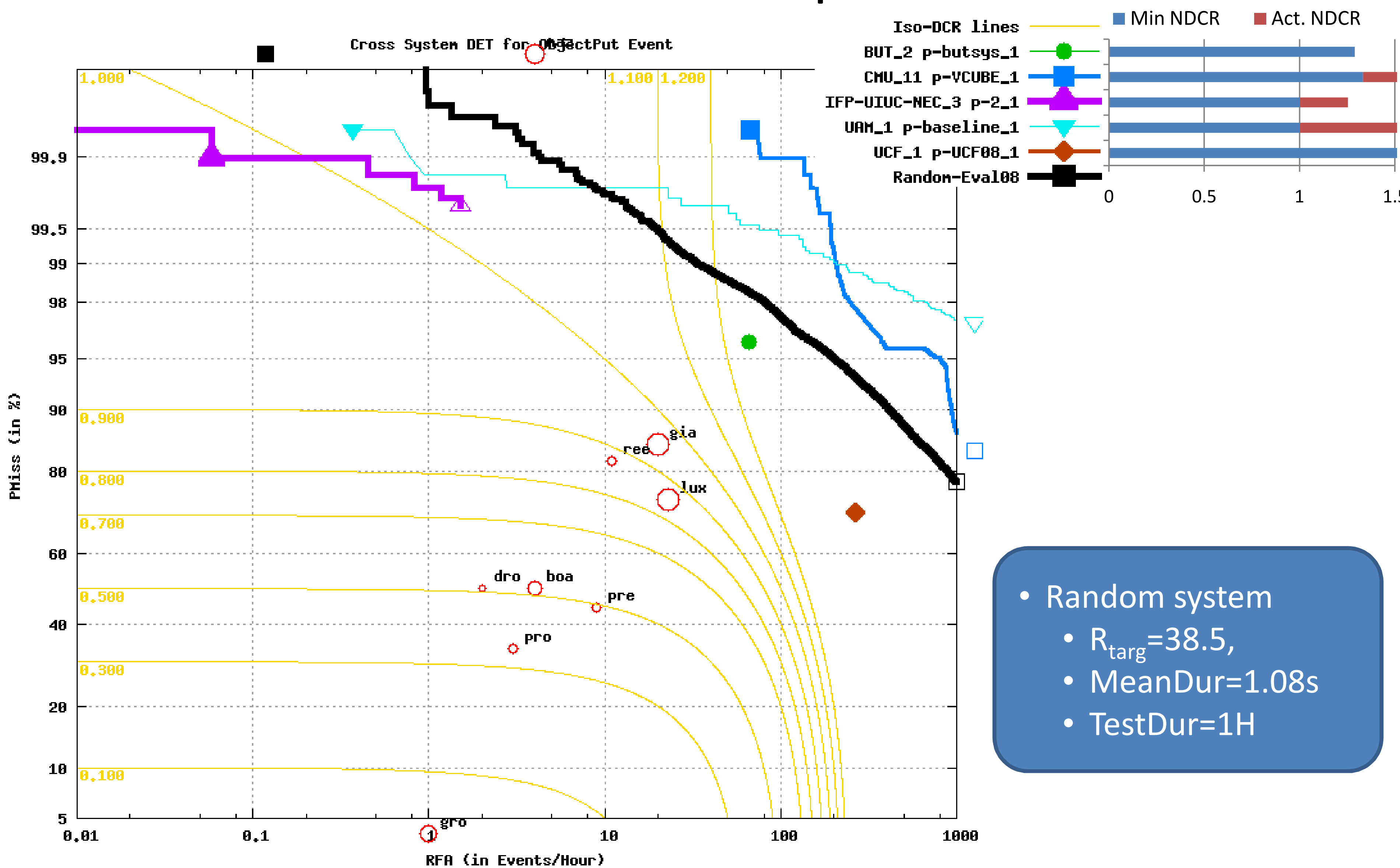
Best Submission per Site



- No human perf. avail.
- No random system
  - R_{targ}=0.11,
  - MeanDur=11.5s
  - TestDur=**642H**

Object Put Event

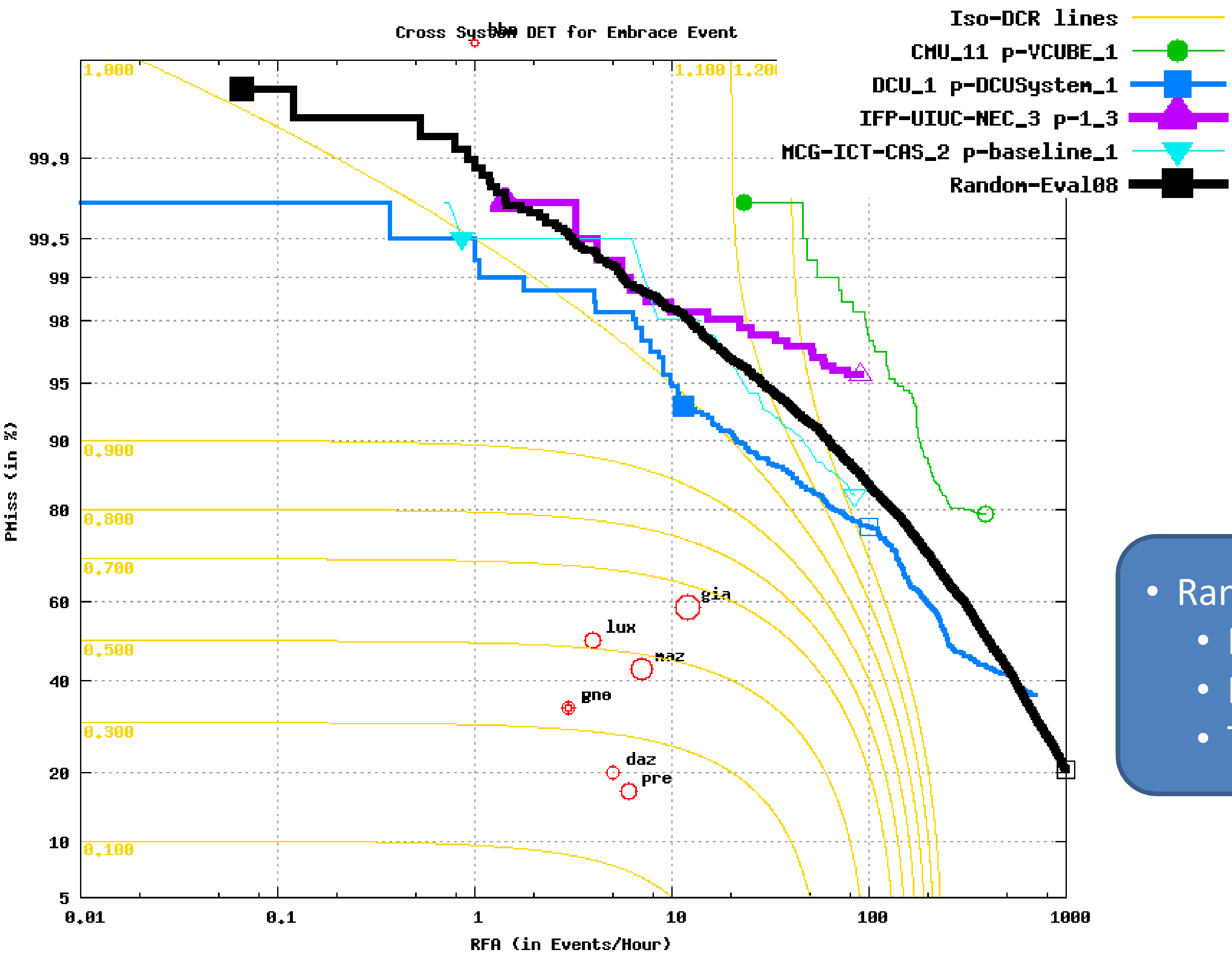
Best Submission per Site

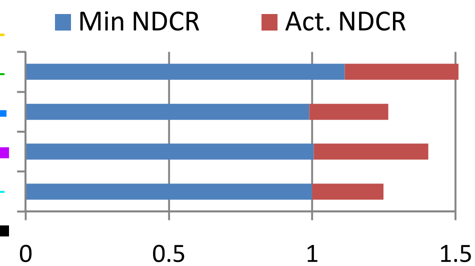


- Random system
  - R_{targ}=38.5,
  - MeanDur=1.08s
  - TestDur=1H

Embrace Event

Best Submission per Site

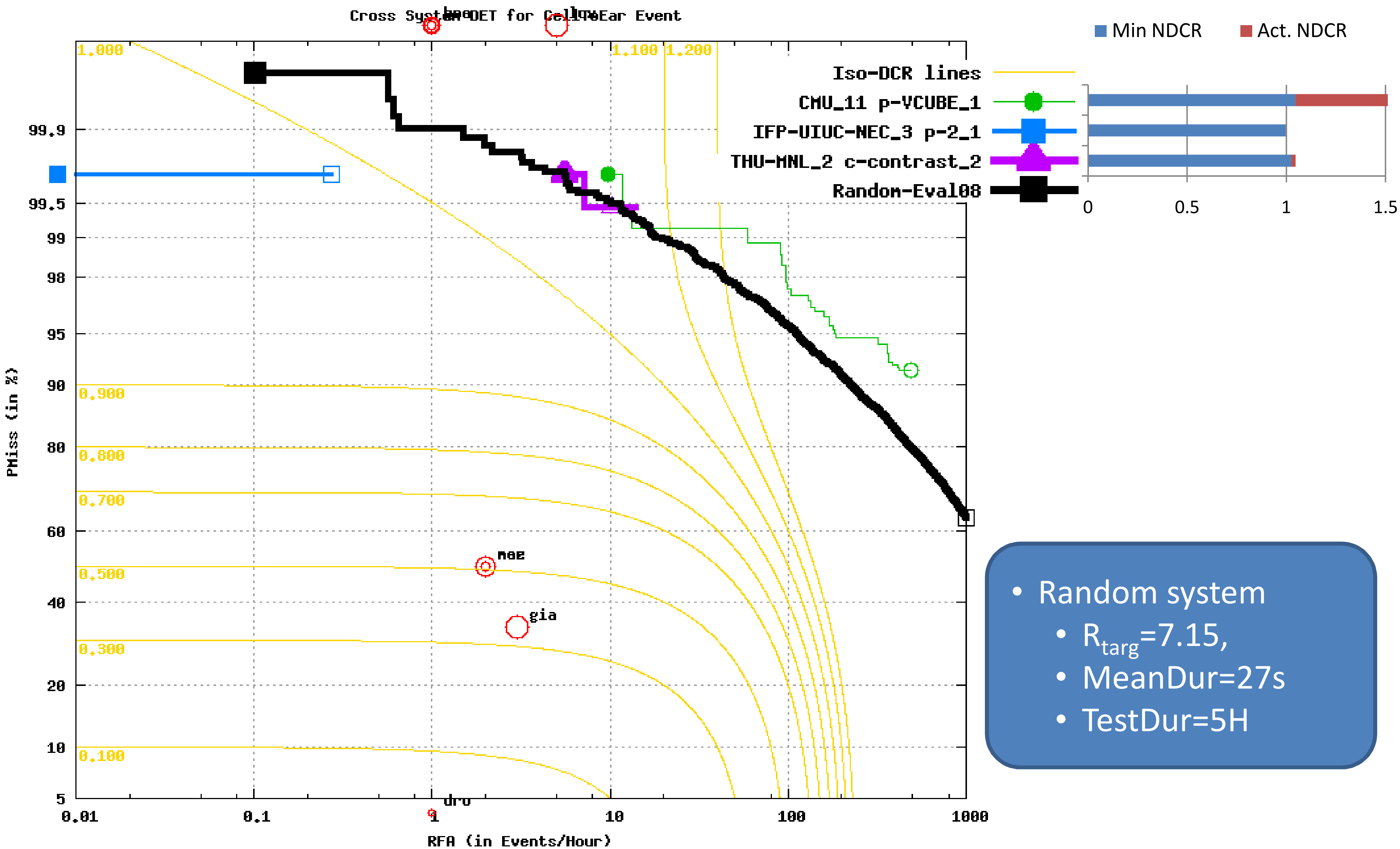




- Random system
  - R_{targ}=8.09,
  - MeanDur=5.2s
  - TestDur=9H

CellToEar Event

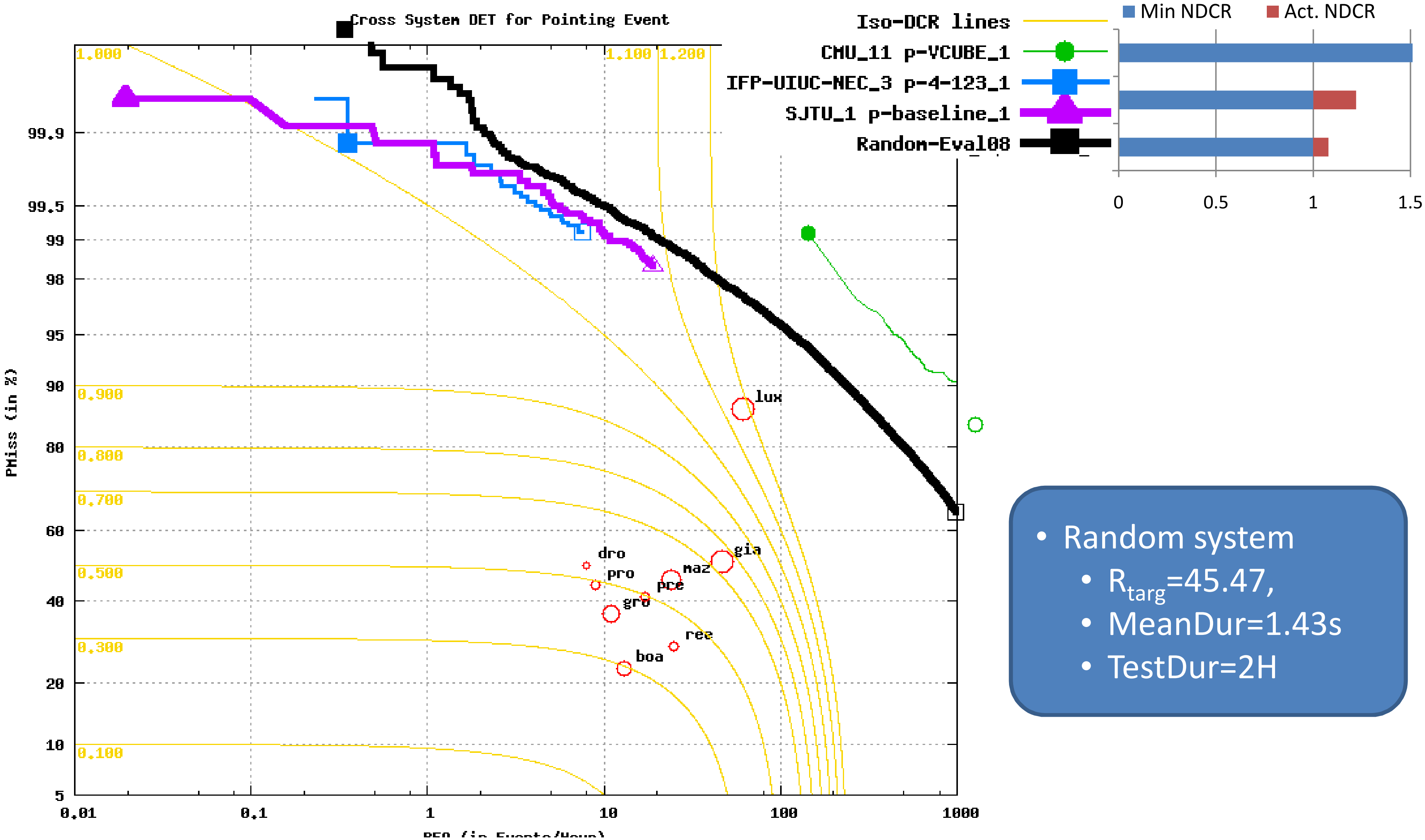
Best Submission per Site



- Random system
  - R_{targ}=7.15,
  - MeanDur=27s
  - TestDur=5H

Pointing Event

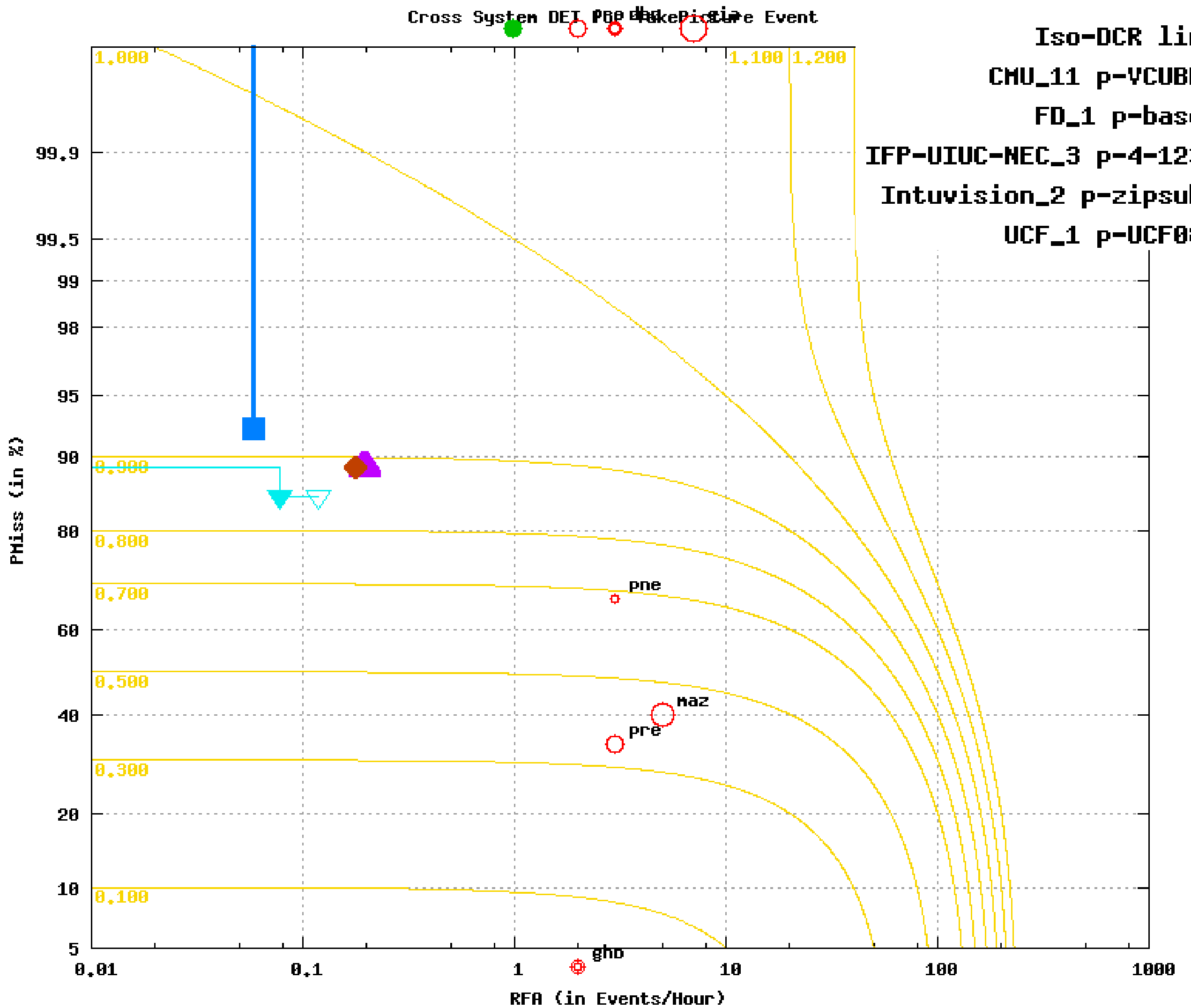
Best Submission per Site

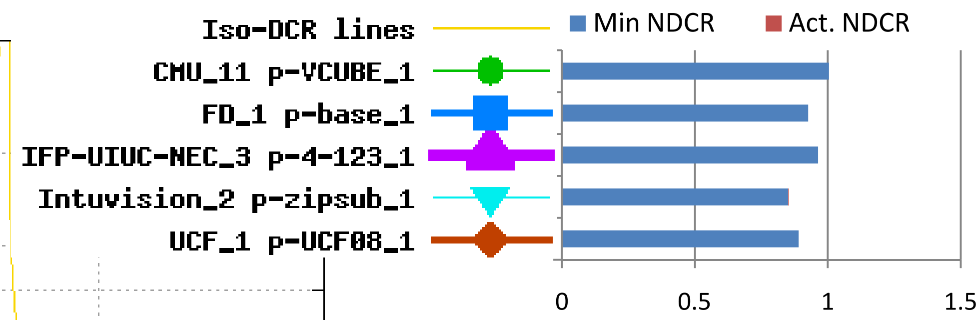


- Random system
  - R_{targ}=45.47,
  - MeanDur=1.43s
  - TestDur=2H

TakePicture Event

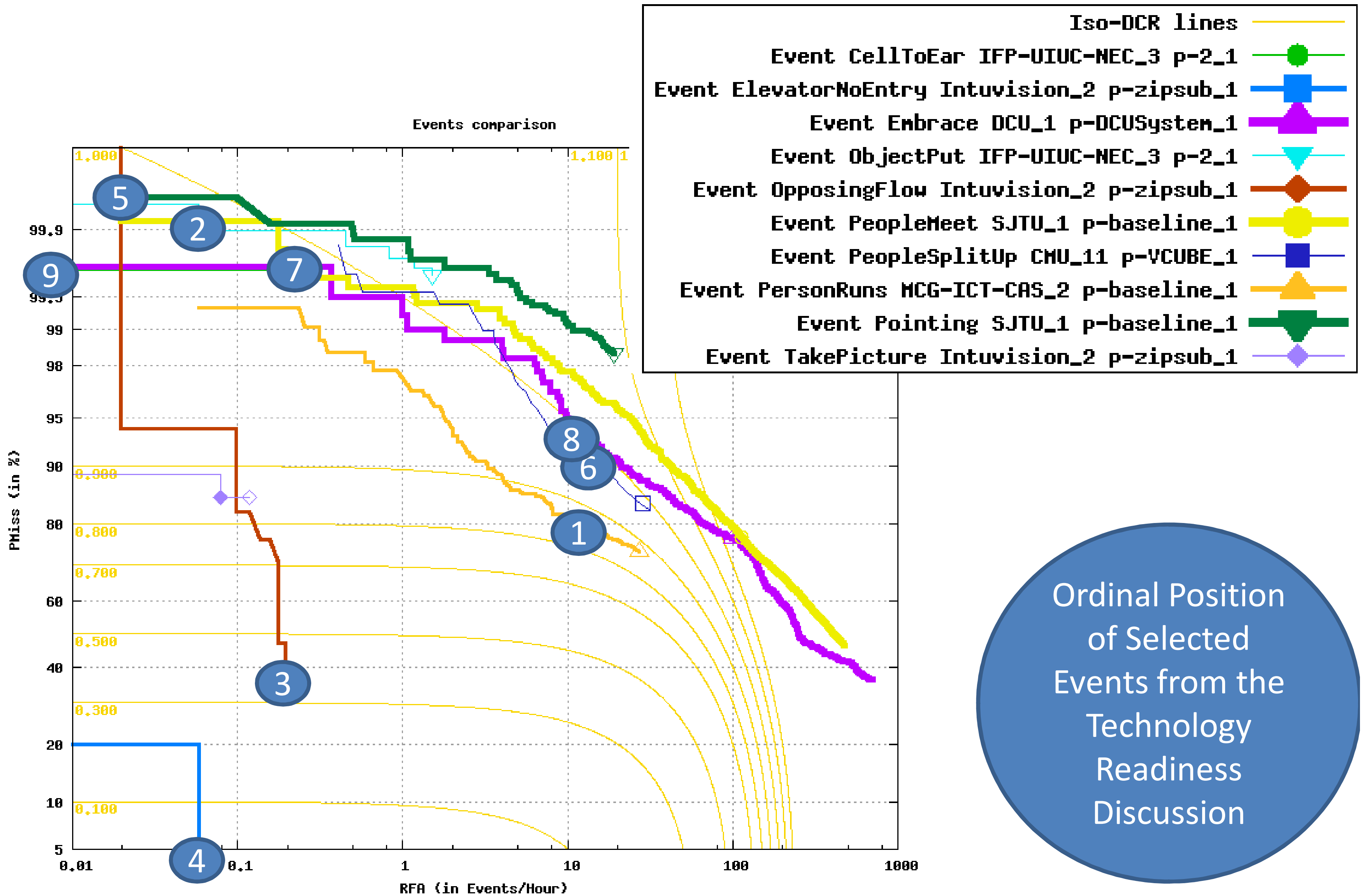
Best Submission per Site





- Random system
  - R_{targ}=.44,
  - MeanDur=9.34s
  - TestDur=170H

Best Run: All Events



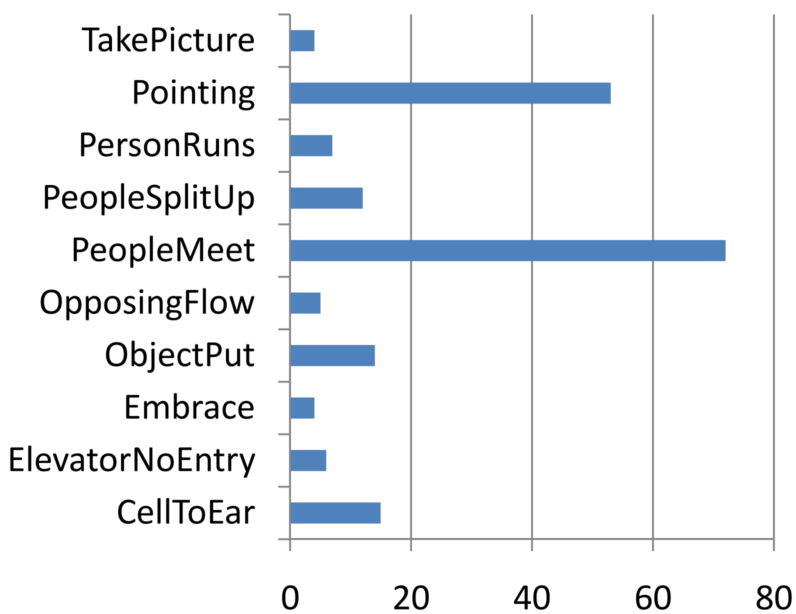
Ordinal Position of Selected Events from the Technology Readiness Discussion

Adjudication Summary

- Dual annotation studies indicated a low recall rate for humans
  - NIST and LDC designed an system-mediated adjudication framework **focused on improving recall**
- Adjudication process for streaming detection
  - Merge system false alarms to develop a prioritized list of excerpts to review:
    - Take into account existing annotations
    - Take into account temporally overlapping annotations
  - Review top 100 false alarm excerpts sorted by
    - Inter-system agreement
    - Average decisions score

Effect of Adjudication

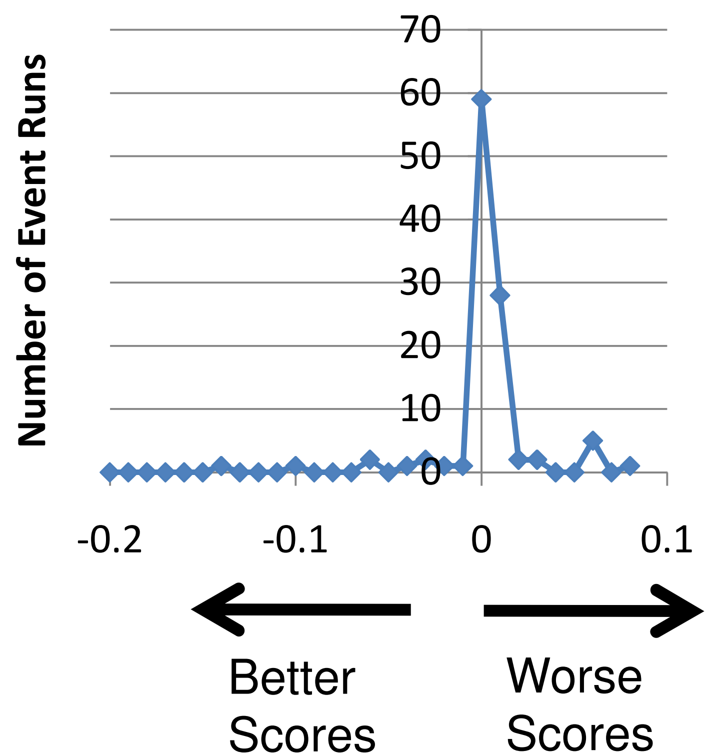
On Annotations



Number of New Event Observations After Reviewing 100 Excerpts

On System Scores

$MinNDCR_{PostAdjud} - MinNDCR_{PreAdjud}$



Conclusions

- Detecting events in high volumes of found data is feasible
  - 16 sites completed the evaluation
  - Human annotation performance indicates the task has a high degree of difficulty
  - 50 Hr. test set insufficient for low frequency events, but 12 Hrs. is sufficient for most events