One Year MediFor PI meeting

# MediFor Nimble Challenge Evaluation 2017

June 27, 2017

**Jonathan Fiscus (PI)**
Dr. Haiying Guan
Dr. Yooyoung Lee
Dr. Amy Yates
Andrew Delgado
Daniel Zhou
David Joy
August Pereira

Multimodal Information Group and Image Group
Information Access Division
Information Technology Laboratory
National Institute of Standards and Technology (NIST)

# Thanks to the Test and Evaluation Team!

- DARPA Media Forensic (MediFor) Team – Role: Program administration
- TA3 Team – Role: Data production and curation
  - PAR Government
  - National Center for Media Forensics, University of Colorado Denver
  - RankOne
  - Rochester Institute of Technology
  - Drexel University
  - University of Michigan
- Air Force Research Lab – Role: Contracting
- NIST MediFor Team – Role: Evaluation designed and implementation

# Outline

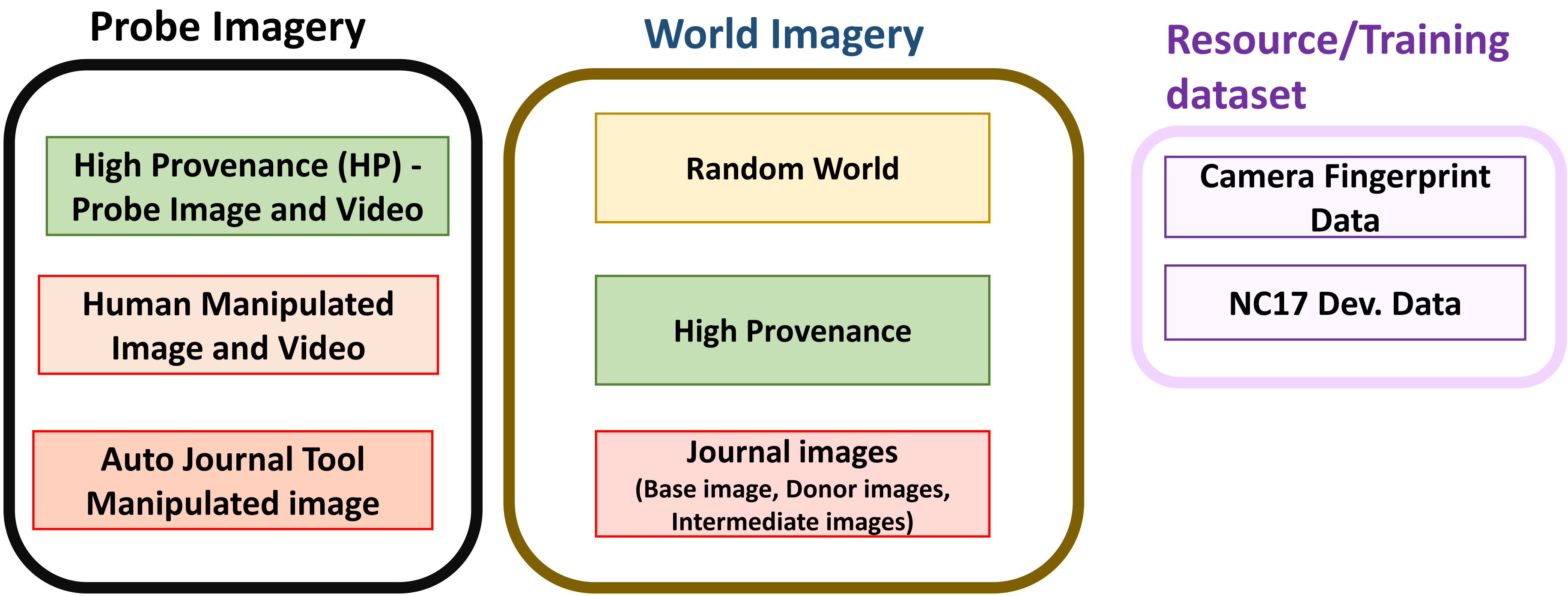
- NIST NC2017 Evaluation Overview
- Detection
  - Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - Splice – paired: Task, Data, Results
  - Video: Data, Result
- Localization
  - Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# NIST NC17 Evaluation Overview

# Nimble Challenge (NC) 2017

- Media Forensics -- “the science and practice of determining the authenticity and establishing the integrity of visual media”
- Evaluation tasks probe the media forensics technology space
  - Is the imagery manipulated?
  - How was the imagery manipulated?
  - Where is the imagery manipulated?
  - What are the sources of manipulated imagery?
- Four technology evaluation tasks for NC2017
  - Manipulation Detection and Localization
    - Images
    - Videos (detection only)
  - Splice Detection and Localization
  - Provenance Filtering
  - Provenance Graph Building

# NC17 Datasets Overview



**Probe Imagery**

- High Provenance (HP) - Probe Image and Video
- Human Manipulated Image and Video
- Auto Journal Tool Manipulated image

**World Imagery**

- Random World
- High Provenance
- Journal images (Base image, Donor images, Intermediate images)

**Resource/Training dataset**

- Camera Fingerprint Data
- NC17 Dev. Data

“Forensic **Probe** – what a forensic analyst would study”

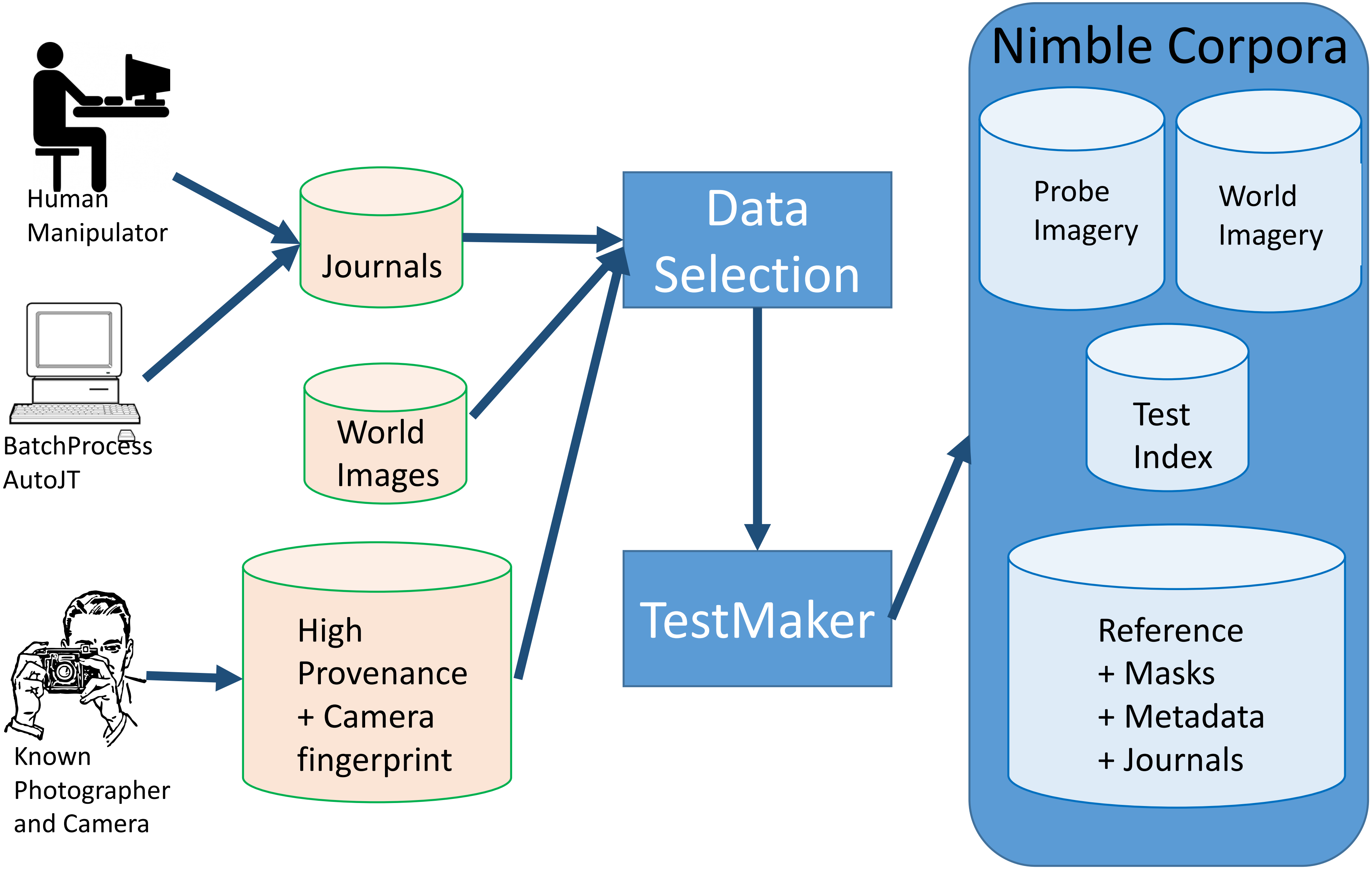
“Image collection potential containing related imagery ”

# List of Datasets

| | Image | | | | | Videos | |
|---|---|---|---|---|---|---|---|
| | Manipulation Journals | MDL Probes | SDL Probes | Provenance Probes | World | Manipulation Journals | MD Probes |
| NC17 Dev | 394-Human | 3563 | 1 M | 2528 | 115K | 25 | 214 |
| NC17 Eval | [redacted] | 10 K | 1 M | 2991 | 1 M | [redacted] | 1083 |
| NC17 EvalPart1 | 132-Auto<br>274-Human | 4077 | 330301 | 992 | 1 M | 45 | 360 |

MDL: Manipulation Detection and Localization
MD: Manipulation Detection
SDL: Splice Detection and Localization
NC17_Dev: the combination of Dev 1, 2, and 3.

# Overview: Data Set Production Data Flow



Human Manipulator

BatchProcess AutoJT

Known Photographer and Camera

Journals

World Images

High Provenance + Camera fingerprint

Data Selection

TestMaker

Nimble Corpora

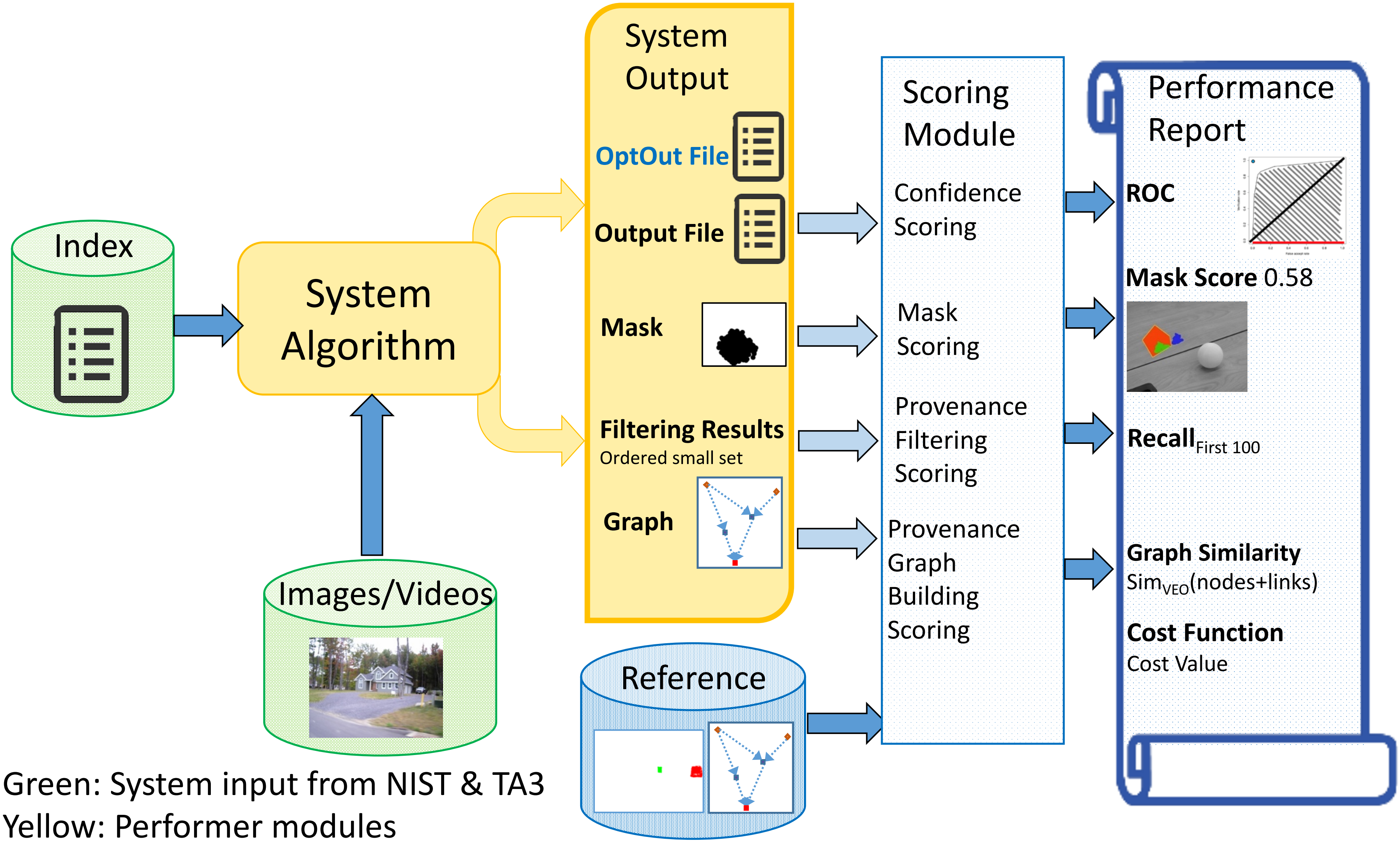
Probe Imagery

World Imagery

Test Index

Reference + Masks + Metadata + Journals

# Overview: Evaluation Modules & Data Flow



Index

System Algorithm

Images/Videos

System Output

OptOut File

Output File

Mask

Filtering Results

Ordered small set

Graph

Reference

Scoring Module

Confidence Scoring

Mask Scoring

Provenance Filtering Scoring

Provenance Graph Building Scoring

Performance Report

ROC

Mask Score 0.58

$\textbf{Recall}_{\text{First 100}}$

Graph Similarity

$\text{Sim}_{\text{VEO}}$(nodes+links)

Cost Function

Cost Value

Green: System input from NIST & TA3

Yellow: Performer modules

Blue: NIST Evaluation modules

# NC17 Participates Overview

| Team Abb. | Organization ID | MDL (image/video) | SDL | PF | PG |
|---|---|---|---|---|---|
| BIN | Binghamton University | 1 | - | - | - |
| FIB | Honeywell ACS Laboratories | 1 | 1 | - | - |
| KIT | Kitware<br>UC Berkeley<br>Dartmouth College<br>University at Albany, SUNY | 4 + 1(video) | - | 1 | - |
| MAY | MAYACHITRA<br>Naval Air Warfare Center, China Lake<br>UC Riverside | 9 | - | - | - |
| PUR | Purdue<br>Politecnico di Milano, Italy<br>University of Siena<br>Univ. of Notre Dame; University of Campinas, Brazil | 5 | - | 5 | 4 |
| SRI-TA2 | SRI International, Princeton (Ajay Divakaran) | 1 | - | - | - |
| SRPPRI | SRI International, Princeton (Jeffrey Lubin) | 1+1(video) | - | - | - |
| UMD | University of Maryland, College Park | 1 | - | - | - |
| UNIFI | University of Florence, FENCE, Prato, Italy | 3 | 2 | - | - |
| USCISI | University of Southern California, ISI | 5 | 1 | 1 | 1 |
| 10 teams | 19 organizations, 49 systems | 31 + 2(video) | 4 | 7 | 5 |

# System OptIn vs. Selective Scoring

- System OptIn Protocol
  - Media Forensics techniques often address a specific manipulation type, sources, etc.
    - This is NOT intended to be generalized ‘shunt’ for failure to read/process
  - The OptIn Protocol allows developer/system provide a response for a probe IFF a response is appropriate given ‘only the imagery and imagery metadata’
  - Score reporting
    - Trail Response Rate – Fraction of probes for which the system responded
    - Performance measures on the subset of trials.
- Selective Scoring
  - Performed by NIST
  - Data analysis technique using metadata to condition analysis, i.e., manipulations of a certain type, etc.

# Submissions Labelling Motivation: Site, Team, Primary and Contrastive

- Blind evaluations can be difficult for both teams and evaluators
  - Teams want to evaluate two kinds of systems
    1. **Primary**: Competitive, bells and whistles, optimized collaborations
    2. **Contrastive**: Diminished systems to test theories, components, etc.
  - Evaluators want to make apples/apples comparisons
    - Compare performance within team
    - Compare performance across team
- NIST Approach –
  - Sites - independent organization/lab, signature authority for licenses (Kitware, Berkeley)
  - Teams - one or more collaborating sites (e.g., Kitware, Kitware-Berkley, Berkley)
  - A Submission is made by a Team and labelled as Primary or Contrastive

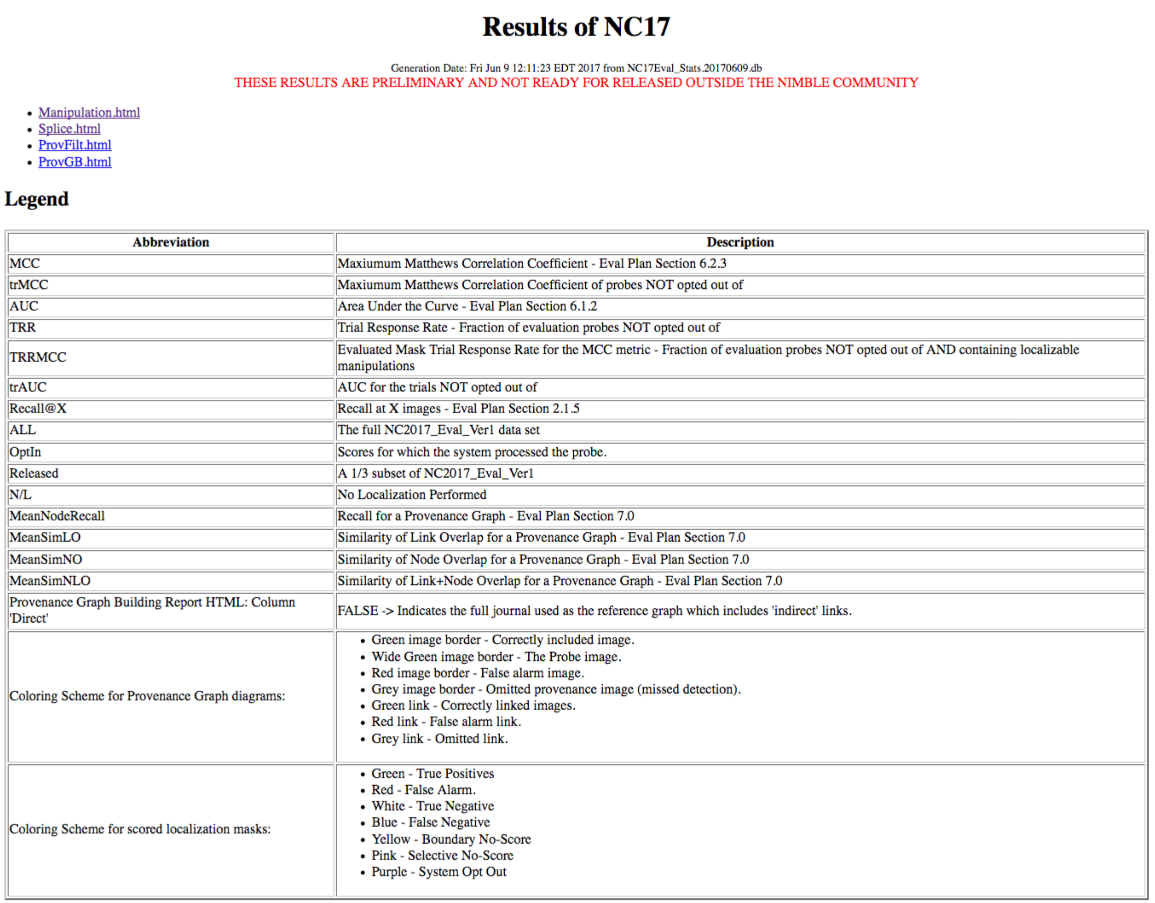
# Nimble Evaluation Rules/Procedures

- Follow the Evaluation Agreement
  - "The site agrees to not publicly compare its results with the results of other participants. Sites are free to do what they wish with their own results, but may neither redistribute nor publish results from another site without that site's explicit permission."
- NC2016 and NC2017 Development data sets are free to use for development and training
- NC2017 Evaluation data is for 'Evaluation' but not training or development
  - References for 1/3 (called NC2017 EvalPart1) have been released for use as an internal test set.

# NIST Results Report:

https://mig.nist.gov/MediforBP/NC17_Participants_LatestResults/

User: mediforBPperf
PW: firstMedi4Data



## Results of NC17

Generation Date: Fri Jun 9 12:11:23 EDT 2017 from NC17Eval_Stats.20170609.db
THESE RESULTS ARE PRELIMINARY AND NOT READY FOR RELEASED OUTSIDE THE NIMBLE COMMUNITY

- Manipulation.html
- Splice.html
- ProvFilt.html
- ProvGB.html

### Legend

| Abbreviation | Description |
|---|---|
| MCC | Maxiumum Matthews Correlation Coefficient - Eval Plan Section 6.2.3 |
| trMCC | Maxiumum Matthews Correlation Coefficient of probes NOT opted out of |
| AUC | Area Under the Curve - Eval Plan Section 6.1.2 |
| TRR | Trial Response Rate - Fraction of evaluation probes NOT opted out of |
| TRRMCC | Evaluated Mask Trial Response Rate for the MCC metric - Fraction of evaluation probes NOT opted out of AND containing localizable manipulations |
| trAUC | AUC for the trials NOT opted out of |
| Recall@X | Recall at X images - Eval Plan Section 2.1.5 |
| ALL | The full NC2017_Eval_Ver1 data set |
| OptIn | Scores for which the system processed the probe. |
| Released | A 1/3 subset of NC2017_Eval_Ver1 |
| N/L | No Localization Performed |
| MeanNodeRecall | Recall for a Provenance Graph - Eval Plan Section 7.0 |
| MeanSimLO | Similarity of Link Overlap for a Provenance Graph - Eval Plan Section 7.0 |
| MeanSimNO | Similarity of Node Overlap for a Provenance Graph - Eval Plan Section 7.0 |
| MeanSimNLO | Similarity of Link+Node Overlap for a Provenance Graph - Eval Plan Section 7.0 |
| Provenance Graph Building Report HTML: Column 'Direct' | FALSE -> Indicates the full journal used as the reference graph which includes 'indirect' links. |
| Coloring Scheme for Provenance Graph diagrams: | • Green image border - Correctly included image.<br>• Wide Green image border - The Probe image.<br>• Red image border - False alarm image.<br>• Grey image border - Omitted provenance image (missed detection).<br>• Green link - Correctly linked images.<br>• Red link - False alarm link.<br>• Grey link - Omitted link. |
| Coloring Scheme for scored localization masks: | • Green - True Positives<br>• Red - False Alarm.<br>• White - True Negative<br>• Blue - False Negative<br>• Yellow - Boundary No-Score<br>• Pink - Selective No-Score<br>• Purple - System Opt Out |

# Manipulation Detection

- Image Manipulation Detection
- Splice Manipulation Detection
- Video Manipulation Detection

# Outline

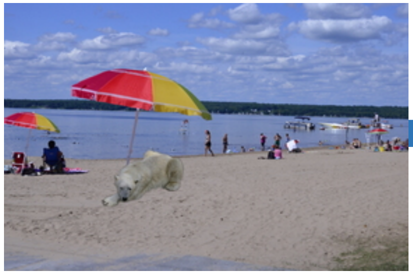
✓NIST NC2017 Evaluation Overview

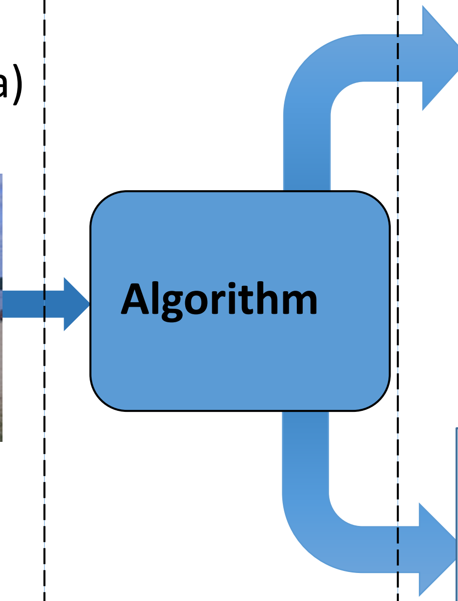
- Detection
  - ➢Image – single: **Task, Data, Metrics, Selective Scoring, Results,** Analysis
  - Splice – paired: Task, Data, Results
  - Video: Data, Result
- Localization
  - Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# Manipulation Detection and Localization Evaluation Model: Expected Output for a **Manipulated Image**

System Input

Image(s) + (metadata)





**Algorithm**

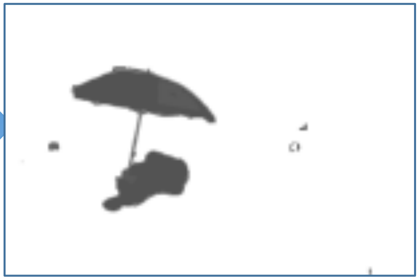
System Output

Detection

Confidence score

27.58

Localization

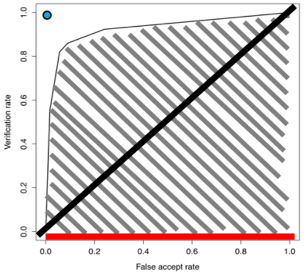


Probe Mask
(If a manipulation)

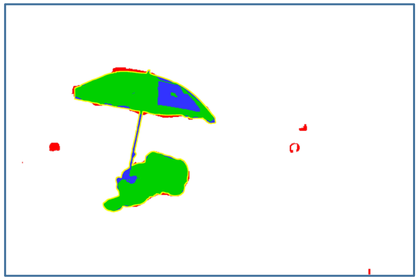
Metrics

**Confidence Score**

Receiver operating characteristic (ROC)



AUC: 0.85

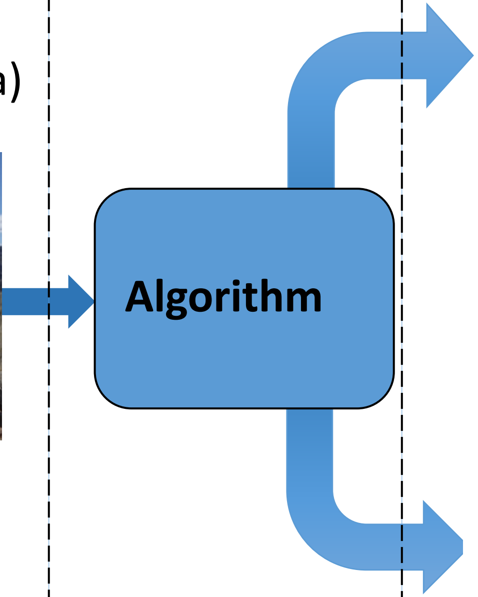


Maximum
Matthews Correlation Coefficient
0.873343237591

# Manipulation Detection and Localization Evaluation Model: Expected Output for a **Non-Manipulated Image**

## System Input

Image(s) + (metadata)



Algorithm

## System Output

### Detection

Confidence score

-17.58

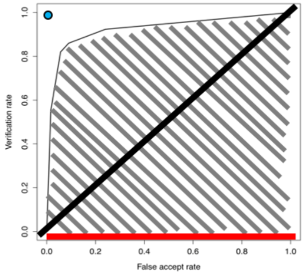
### Localization

NULL

## Metrics

**Confidence Score**

Receiver operating characteristic(ROC)



AUC: 0.85

# NC17 Image Manipulation Datasets

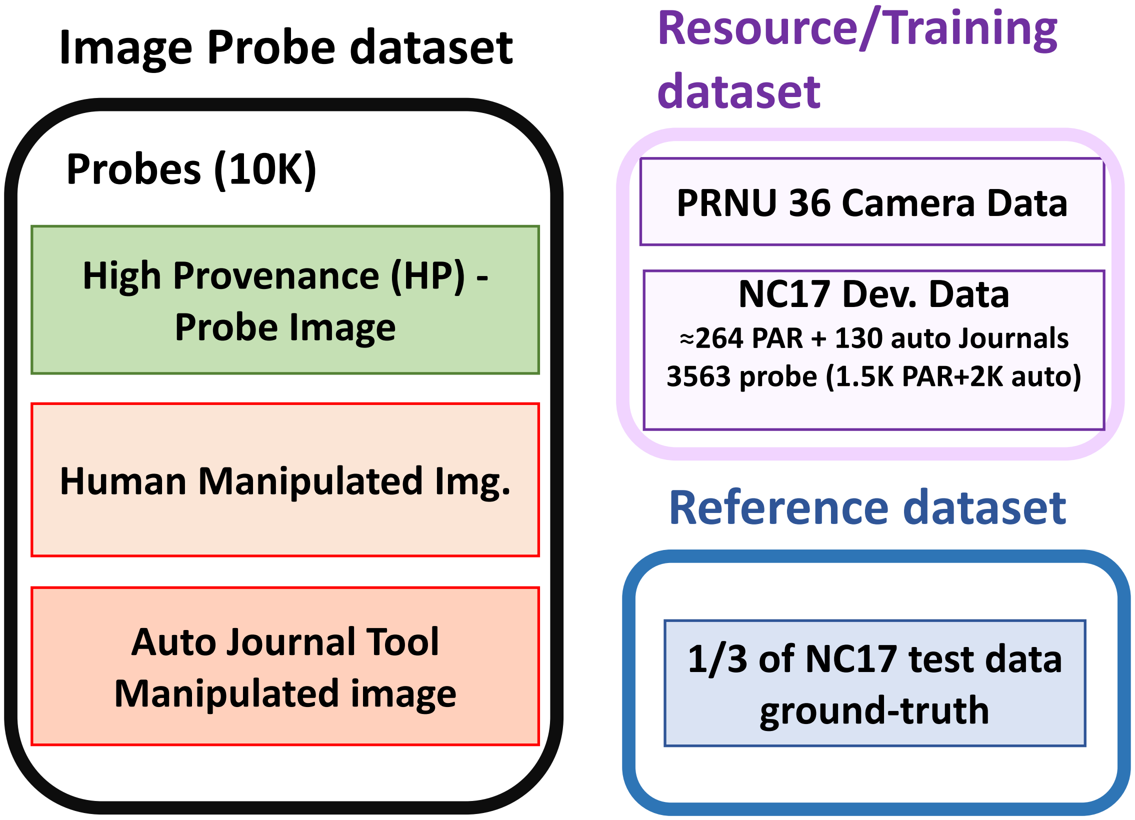


Image Probe dataset

Probes (10K)

High Provenance (HP) - Probe Image

Human Manipulated Img.

Auto Journal Tool Manipulated image

Resource/Training dataset

PRNU 36 Camera Data

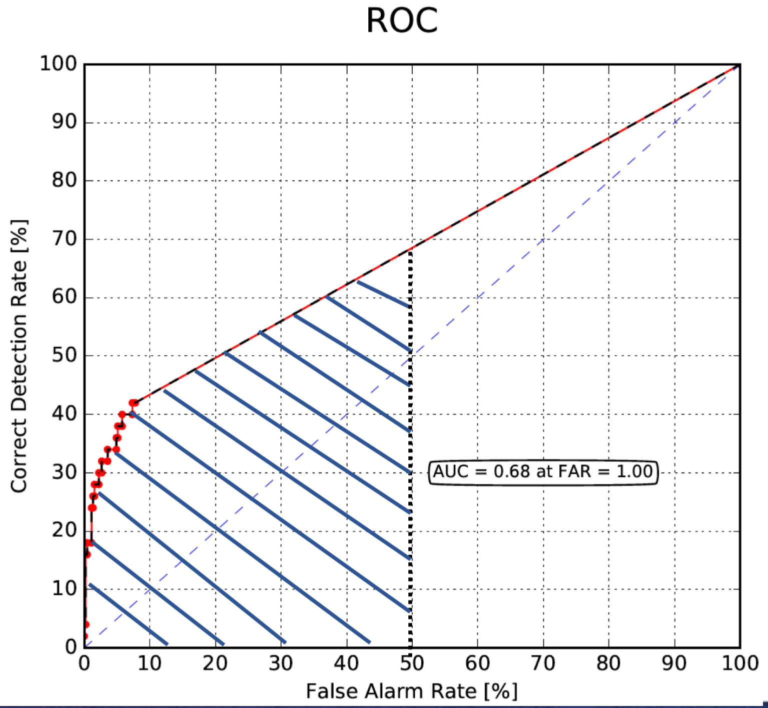
NC17 Dev. Data
≈264 PAR + 130 auto Journals
3563 probe (1.5K PAR+2K auto)

Reference dataset

1/3 of NC17 test data ground-truth

# Detection System Evaluation Metrics

- Detection Scorer
  - ROC (Receiver Operating Characteristic)
  - AUC (Area Under Curve)
- Evaluation Software Package:
  - MediScore



# NC17 Image Manipulation Detection

## - Participate Summary

- Task condition is all 'Image Only'.

| Team | Organizations | Image Mani. Det. | |
|---|---|---|---|
| | | All Probe | OptIn |
| BIN | Binghamton University | - | 1 |
| FIB | Honeywell ACS Laboratories | - | 1 |
| KIT | Kitware<br>UC Berkeley<br>Dartmouth College<br>University at Albany, SUNY | 4 | - |
| MAY | MAYACHITRA<br>Naval Air Warfare Center, China Lake<br>UC Riverside | 7 | 2 |
| PUR | Purdue<br>Politecnico di Milano, Italy<br>University of Siena<br>Univ. of Notre Dame; University of Campinas, Brazil | 3 | 2 |
| SRI-TA2 | SRI International, Princeton (Ajay Divakaran) | 1 | |
| SRPPRI | SRI International, Princeton (Jeffrey Lubin) | - | 1 |
| UMD | University of Maryland, College Park | - | 1 |
| UNIFI | University of Florence, FENCE, Prato, Italy | - | 3 |
| USCISI | University of Southern California, ISI | - | 5 |
| **10 teams** | **16 organizations, 31 systems** | **15** | **16** |

# NC17 Image Manipulation Detection

## - OptIn System Summary

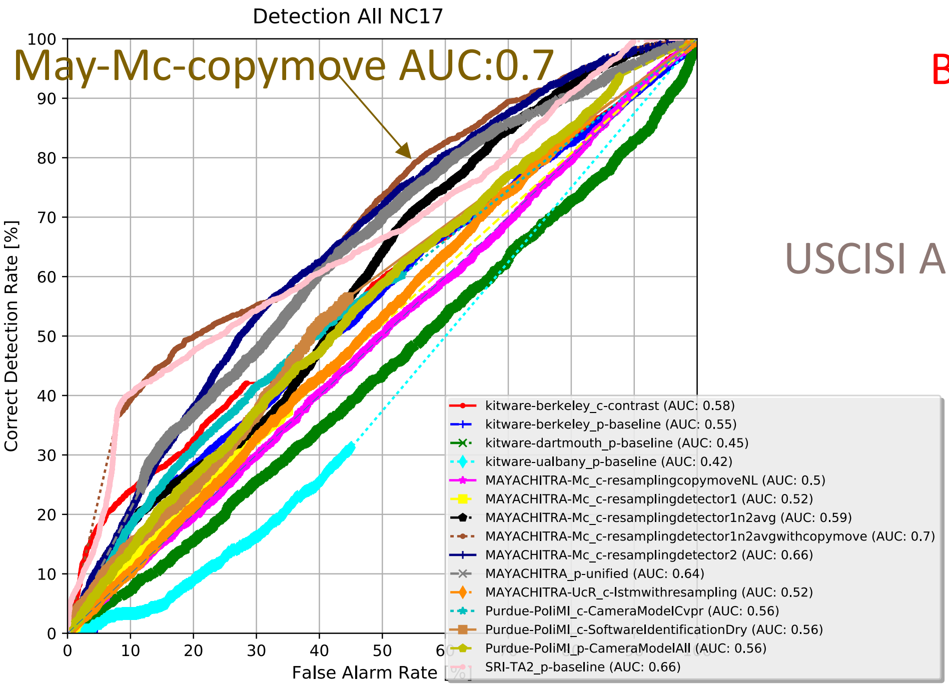
| Team-Org-ID | SystemID | OptIn (TRR*) |
|---|---|---|
| MAYACHITRA-UcR | c-lstmwithoutresampling_2 | ≈ 1 |
| USCISI | c-PMcopymove01a_1 | ≈ 1 |
| USCISI | c-PMinpainting01a_1 | ≈ 1 |
| Purdue | p-MFCN1_1 | 0.96 |
| USCISI | c-Autoencoder01a_1.txt | 0.95 |
| USCISI | p-Splicebuster01a_1 | 0.95 |
| FIBBER | p-FourIGH_1 | 0.93 |
| MAYACHITRA-Cl | c-acontrario_3 | 0.92 |
| Purdue-Unisi | p-baseline_1 | 0.92 |
| UNIFI | p-baselineMOD1_1 | 0.43 |
| SRIPRI | p-baseline_1 | 0.13 |
| USCISI | c-gradbased01a_1 | 0.12 |
| BINGHAMTON | p-prnu_1 | 0.08 |
| UMD | p-facesteganalysis_1.txt | 0.04 |
| UNIFI | c-baselineMOD3_1 | 0.03 |
| UNIFI | c-baselineMOD4_1 | 0.03 |

*Trial Response Rate (TRR) - Fraction of evaluation probes NOT opted out of

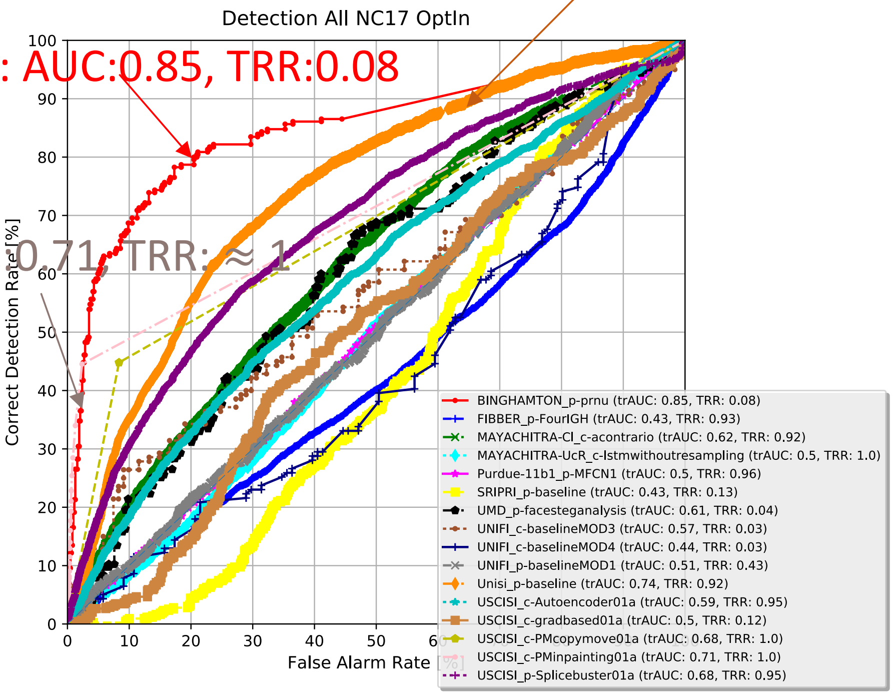
# NC17 Image Manipulation Detection Results

- Observation: when focusing on OptIn, some systems' performance is higher.

UNISI: AUC:0.74, TRR:0.92



May-Mc-copymove AUC:0.7



BIN: AUC:0.85, TRR:0.08

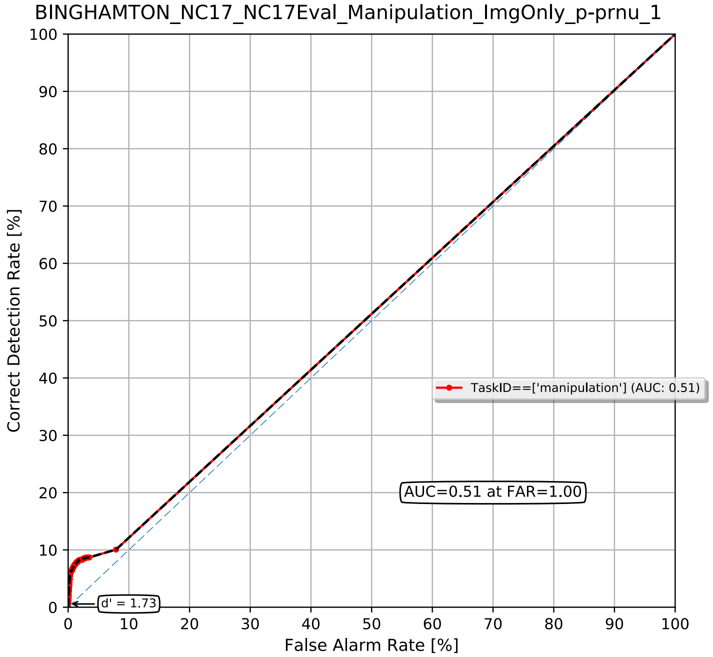
USCISI AUC:0.71, TRR: ≈ 1

All Probes

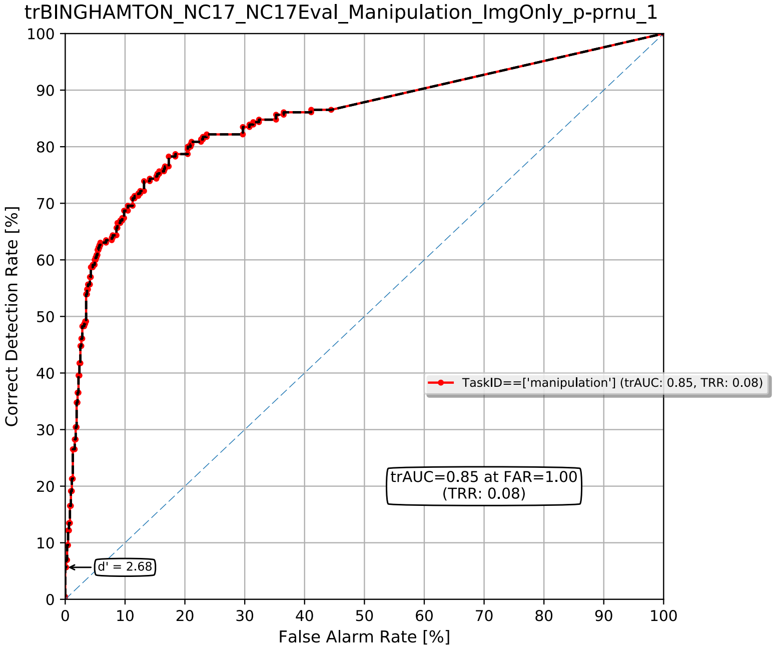
OptIn

# NC17 Image Manipulation Detection Metrics Limitation (1)

- Metrics interpretation: AUC. vs. trAUC (OptIn)
  - AUC: Area Under the Curve - Eval Plan Section 6.1.2; trAUC:  AUC for the trials NOT opted out of
  - ROC: keep shape stretch the curve in X, Y with different scale respectively.



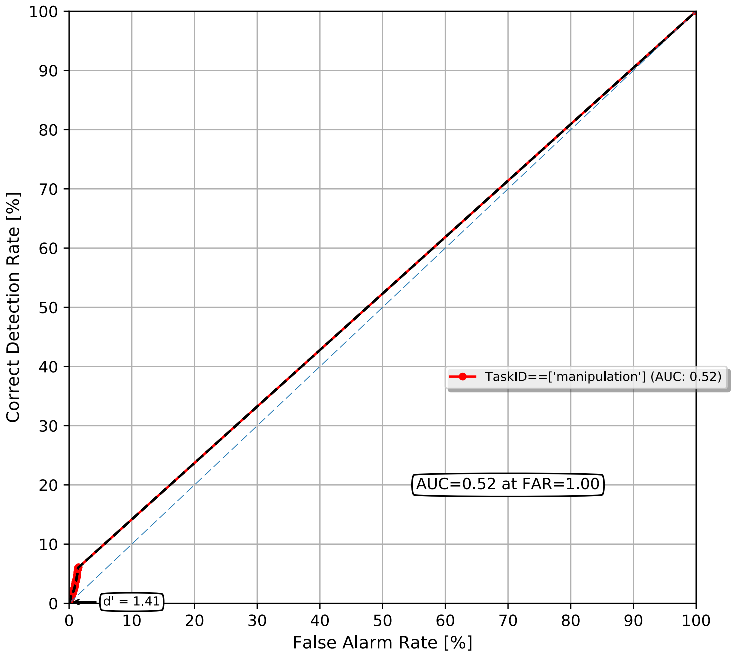
BIN: AUC= 0.513641



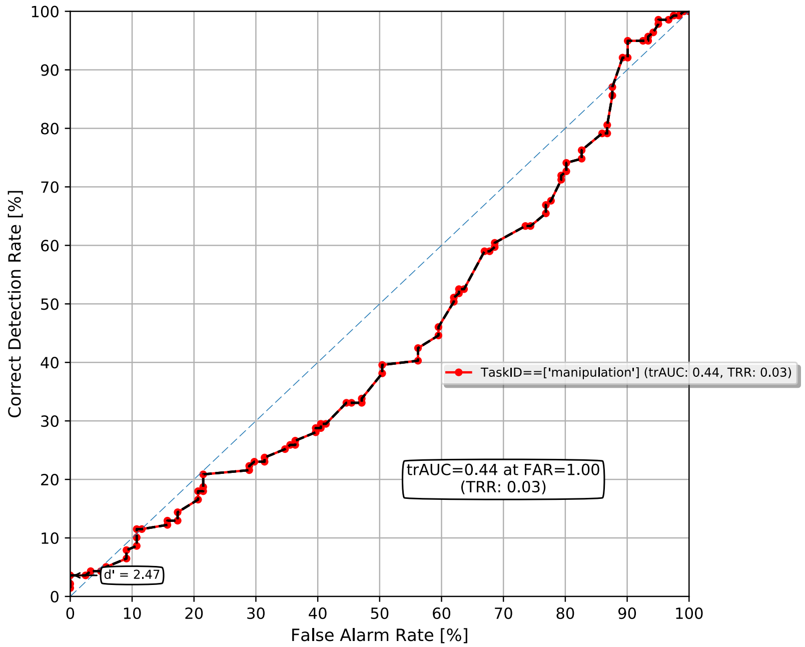
BIN: trAUC = 0.848965

# NC17 Image Manipulation Detection Metrics Limitation (2)

- Metrics interpretation: AUC. vs. trAUC
  - ROC: keep shape stretch the curve in X, Y with different scale respectively.



AUC=0.522761



trAUC=0.435876

# Selective Scoring Evaluation Infrastructure

- What?
  - NIST developed evaluation infrastructure
  - Select a subset of data given a condition to evaluate the system performance
- Why?
  - What kind of manipulation it is?
  - For a certain manipulation, or combinations of a serial of manipulations, which system has the highest performance?
  - Deeply understand detection system performance – factor analysis
- How?
  - Structured data and metadata collection and annotation
    - Manipulation history graph
    - Journal level, probe image level, and operation level
    - Semantic annotations
  - Dynamic reference ground-truth generation
- Benefits:
  - Support evaluation subtask evaluation;
  - Support system analysis in depth;
  - The relationships between three sub-domains:
    - manipulation operation (image processing algorithms, plugin, tools);
    - manipulation semantic meaning (human understandable);
    - detection system performance (MediFor systems).
  - Fully utilize all data collected: high data collection cost.

# NC17 Image Manipulation Selective Scoring SubTasks

| **Selective Scoring** | **qm command** | #Data |
|---|---|---|
| **PasteSplice** | -qm "Operation==['PasteSplice']" | 926 |
| **Remove** | -qm "Operation==['PasteSampled','FillContentAwareFill'] && Purpose==['remove']" | 1455 |
| **ContentAwareRemove** | -qm "Operation==['FillContentAwareFill']" | 1039 |
| **Clone** | -qm "Operation==['PasteSampled'] && Purpose==['clone']" | 1229 |
| **Recapture** | -qm "Operation==['Recapture']" | 57 |
| **Crop** | -qm "Operation==['TransformCrop']" | 245 |
| **PasteSampled** | -qm "Operation==['PasteSampled']" | 1830 |
| **Paste** | -qm "Operation == ['PasteSampled', 'PasteSplice']" | 2756 |
| **FaceManipulation** | -qm "OperationArgument==['face']" | 88 |
| **LocalBlur** | -qm "Operation==['AdditionalEffectFilterBlur', 'AdditionalEffectFilterSmoothing', 'AdditionalEffectFilterMedianSmoothing', 'FilterBlurMotion','FilterBlurNoise'] && Color!=['']" | 1082 |
| **GlobalBlur** | -qm "Operation==['AdditionalEffectFilterBlur', 'AdditionalEffectFilterSmoothing', 'AdditionalEffectFilterMedianSmoothing', 'FilterBlurMotion','FilterBlurNoise'] && Color==['']" | 582 |
| **Blur** | -qm "Operation==['AdditionalEffectFilterBlur','AdditionalEffectFilterSmoothing', 'AdditionalEffectFilterMedianSmoothing','FilterBlurMotion','FilterBlurNoise'] " | 1664 |
| **LocalSharpening** | -qm "Operation==['AdditionalEffectFilterSharpening'] && Color!=['']" | 76 |
| **GlobalSharpening** | -qm "Operation==['AdditionalEffectFilterSharpening'] && Color==['']" | 560 |
| **Sharpening** | -qm "Operation==['AdditionalEffectFilterSharpening'] " | 636 |
| **GlobalSmoothing** | -qm "Operation==['AdditionalEffectFilterSmoothing'] && Color==['']" | 535 |
| **GlobalIntensityNormalization** | -qm "Operation==['IntensityNormalization'] && Color==['']" | 1069 |

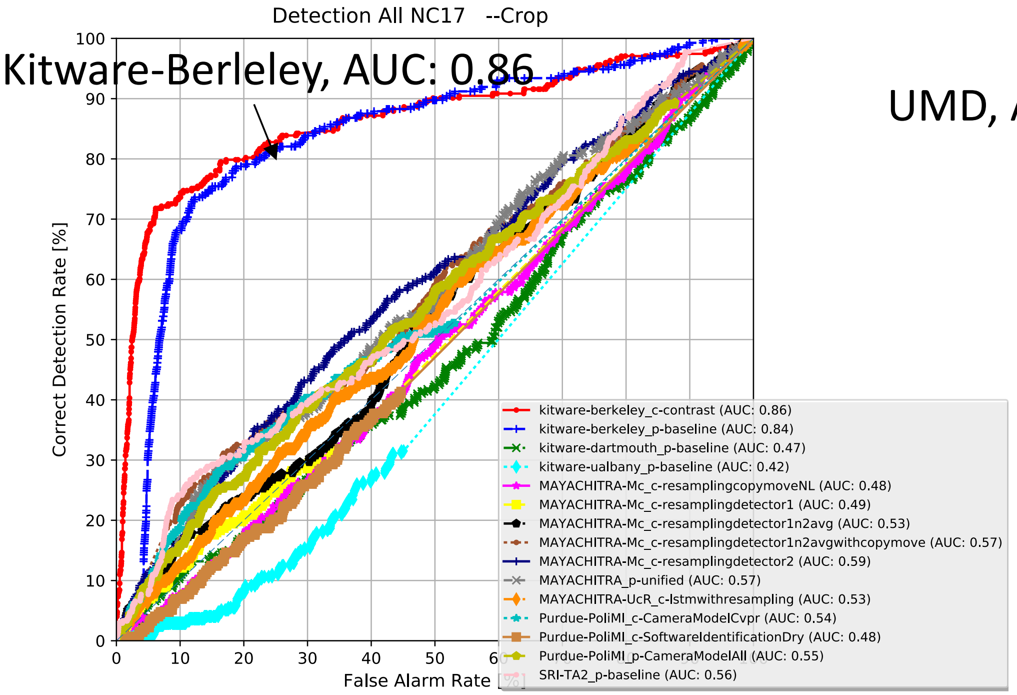
# Selective Scoring Command Line

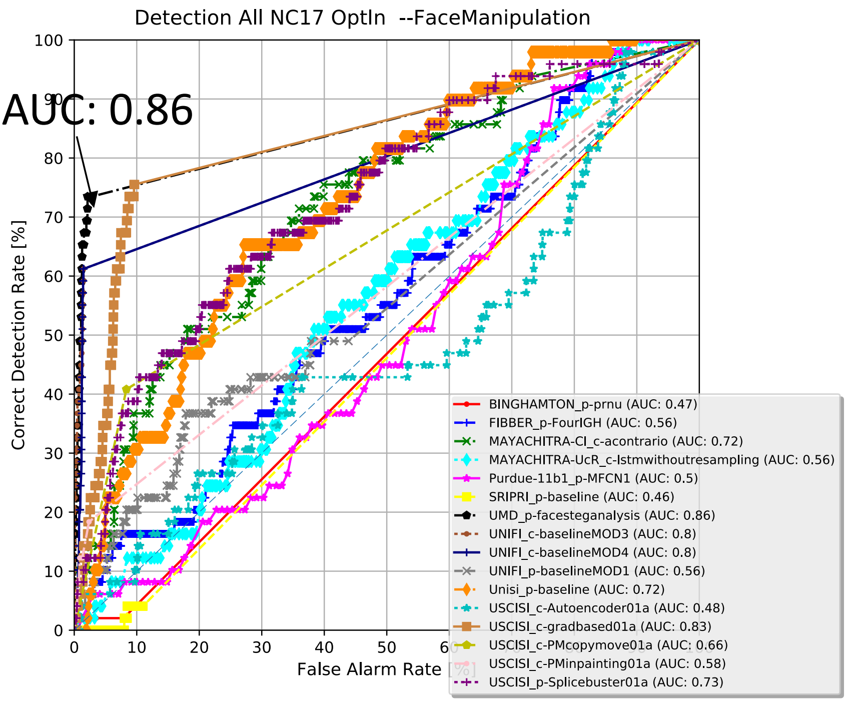
```
python2 /… … /MediScore/tools/MaskScorer/MaskScorer.py
--refDir 00-Reference/NC2017_Eval_Ver1-Part1PAR
--sysDir 10-Submissions/10-EXPIDS/UNIFI_NC17_NC17Eval_Manipulation_ImgOnly_c-baselineM
-s UNIFI_NC17_NC17Eval_Manipulation_ImgOnly_c-baselineMOD4_1.csv
-oR 30-ScorePrelim/UNIFI_NC17_NC17Eval_Manipulation_ImgOnly_c-baselineMOD4_1/Localiz
-html
-v 1
-t manipulation
-qm "OperationArgument==['face']"
-xF
-x indexes/NC2017_Eval-manipulation-image-index.csv
-r reference/manipulation-image/NC2017_Eval-manipulation-image-ref.csv
```

# NC17 Image Manipulation Selective Scoring Results: Operation comparison: single operation vs. all

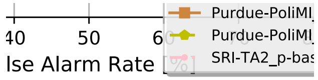
- All Probes; Human+Auto manipulation;
- Observation: performances changed greatly on manipulations.

Detection All NC17





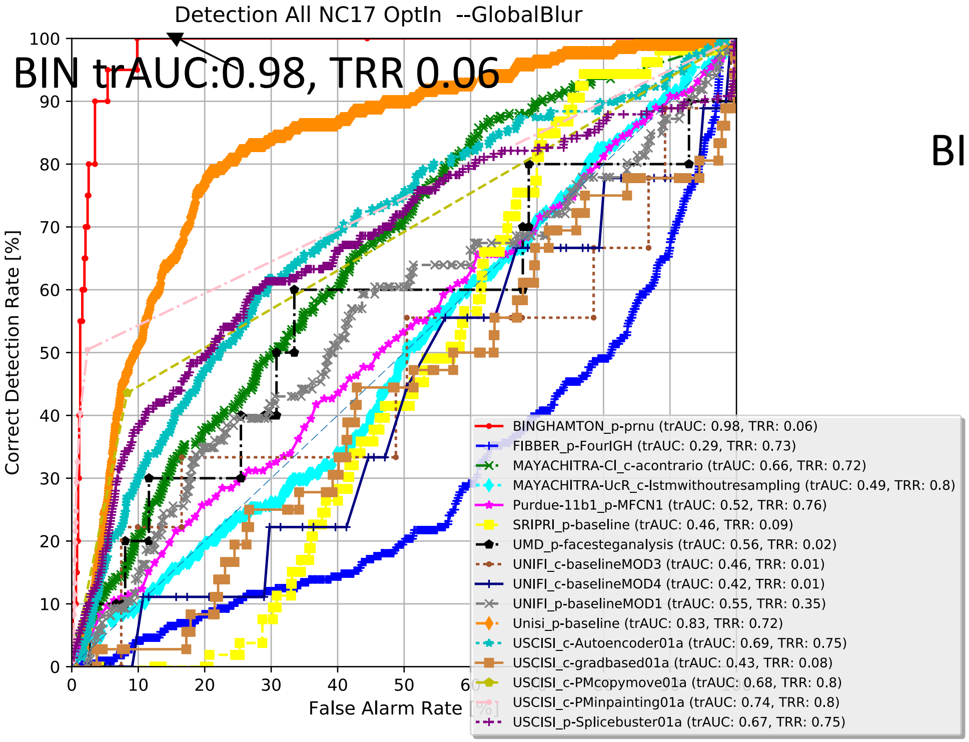
Crop



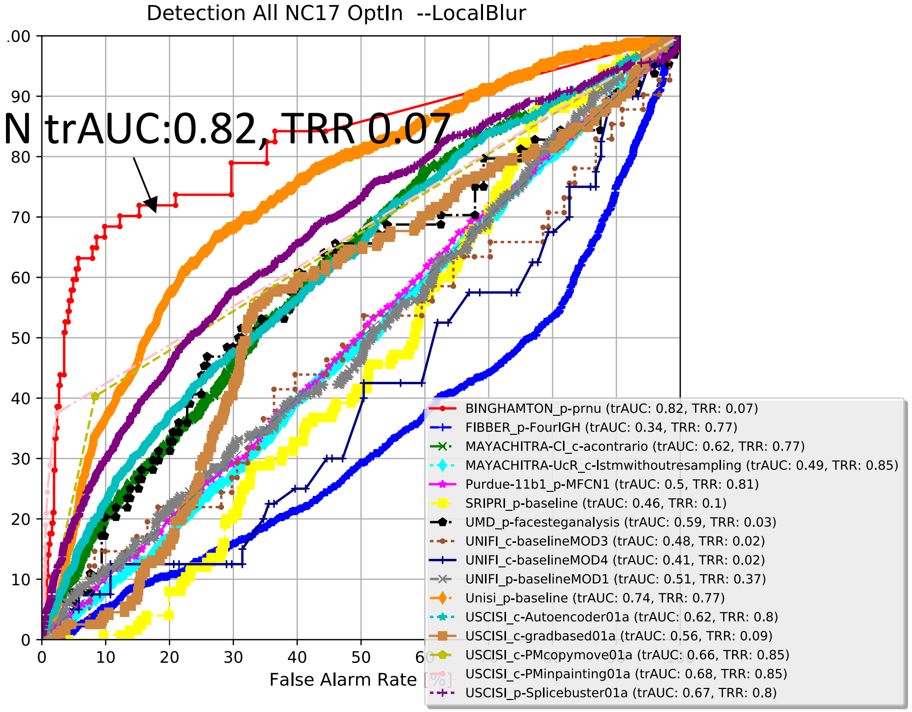
Face

# NC17 Image Manipulation Selective Scoring Results: - global operation vs. local operation

- OptIn; Human+Auto manipulations;
- Observation: even for the same operation, system performs differently given global or local data



Global Blur

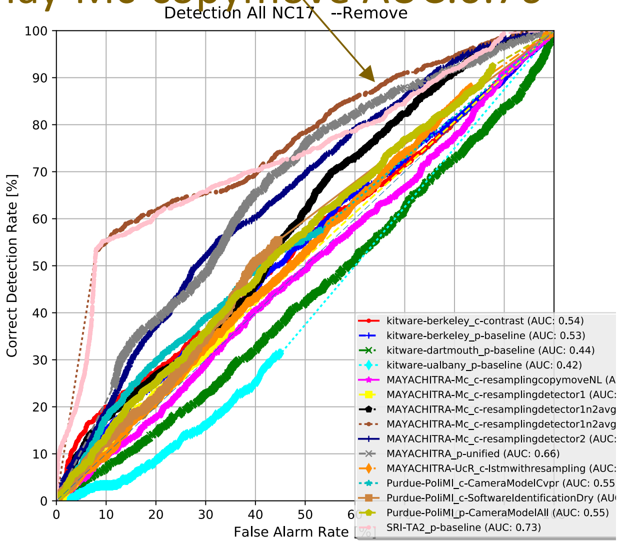


Local Blur

# NC17 Image Manipulation Selective Scoring Results: Semantic comparisons

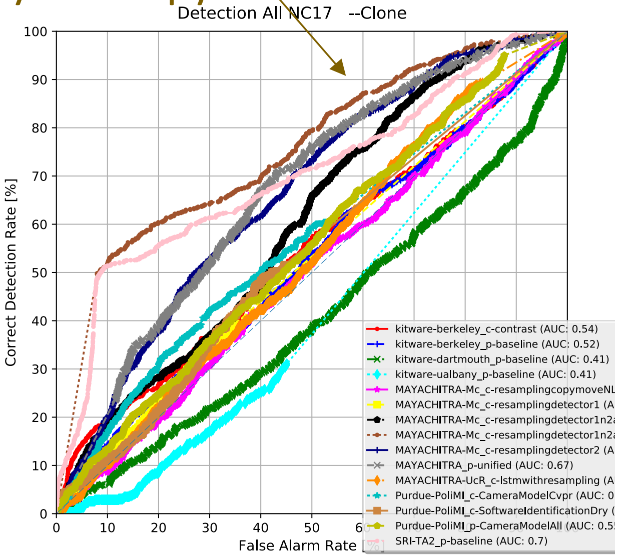
- All probes; Operations:
  - Remove: within image operation, semantic: remove
  - Clone: within image operation, semantic: clone, add object
  - Splice: between image operation, semantic: splice, add object
- System performance: manipulation operation vs. semantic meaning.

May-Mc-copymove AUC:0.76



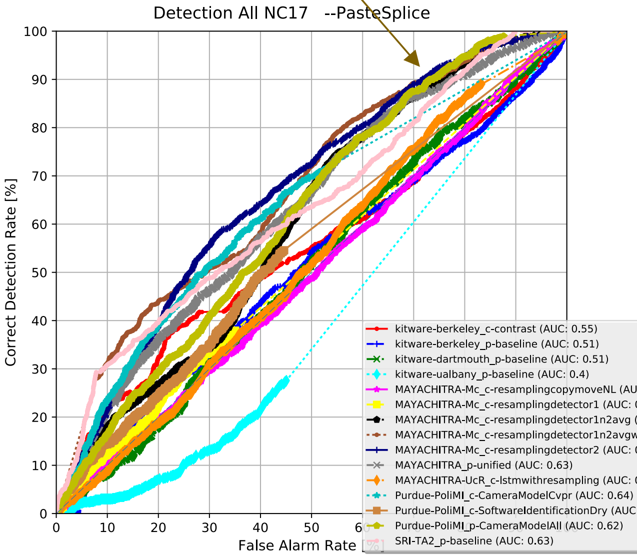
Remove

May-Mc-copymove AUC:0.75



Clone

May-Mc-copymove AUC:0.67



Splice

# Outline

✓NIST NC2017 Evaluation Overview

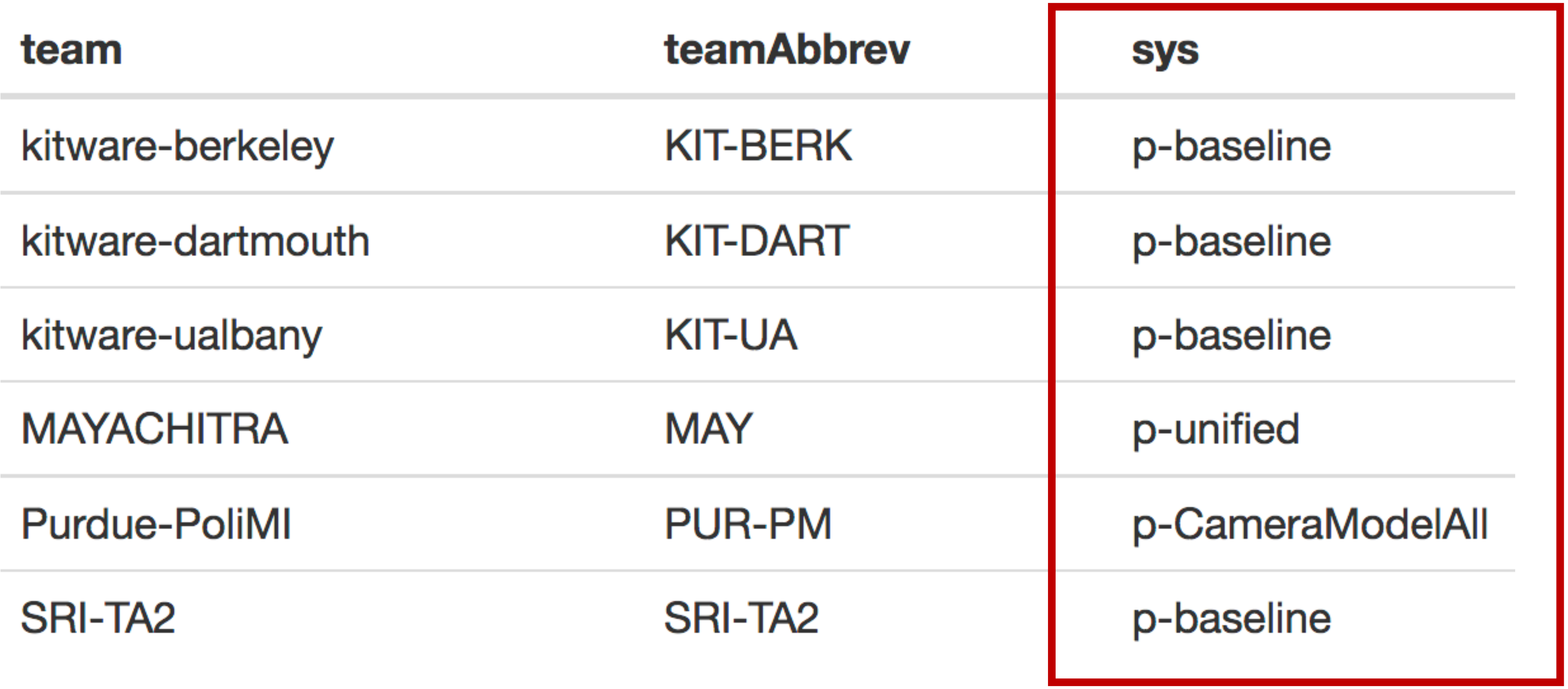
- Detection
  - ➢Image – single: Task, Data, Metrics, Selective Scoring, Results, **Analysis**
  - Splice – paired: Task, Data, Results
  - Video: Data, Result
- Localization
  - Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# Purpose

- Evaluating the accuracy of a system (primary-only)
  - All data: 6 teams
  - OptIn data: 8 teams
- Characterizing a system performance
- Sensitivity analysis: understanding key factors that affect the quality and performance of a system
- Comparative analysis
- Conclusions robust (vs. interactions)
- Providing a research direction for system improvements and optimizations

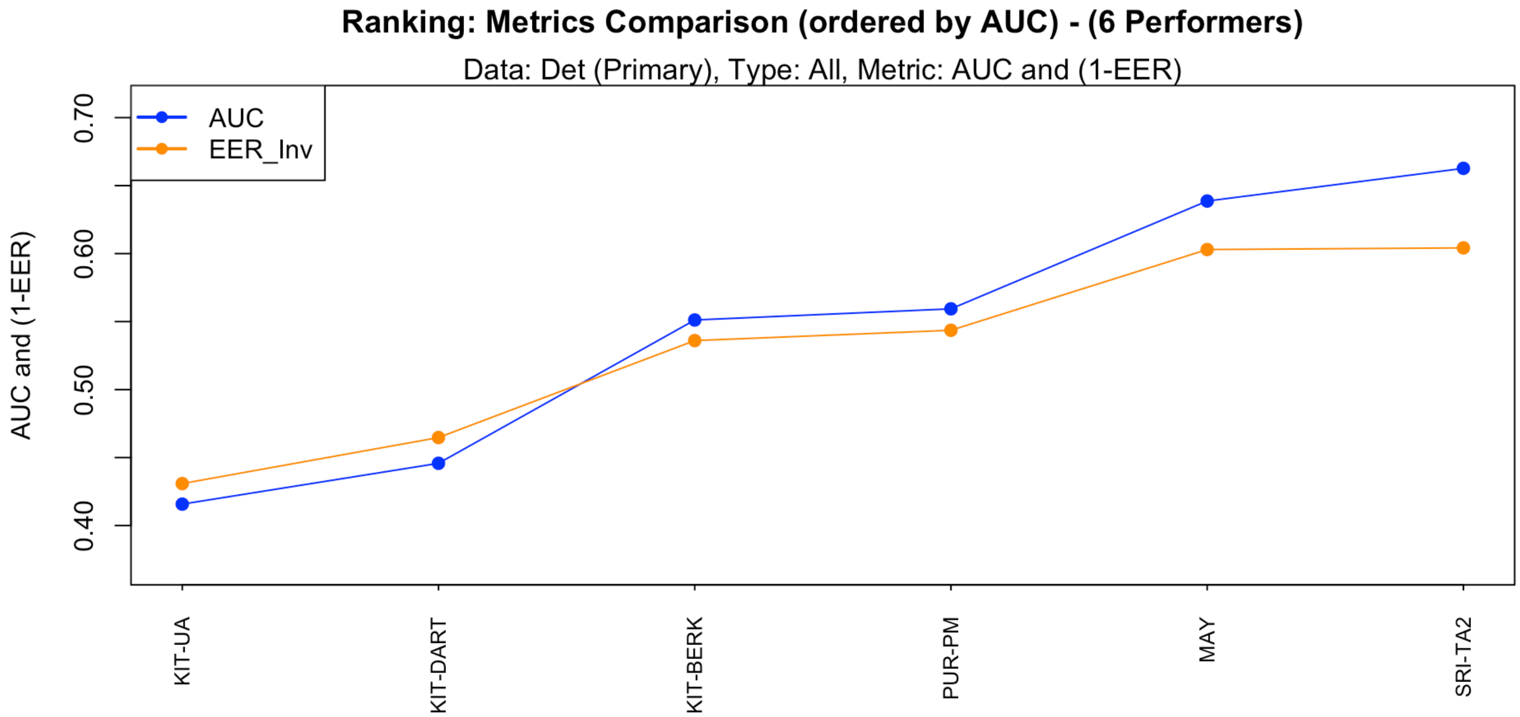
# All Data: 6 Teams

Primary-only



| team | teamAbbrev | sys |
|---|---|---|
| kitware-berkeley | KIT-BERK | p-baseline |
| kitware-dartmouth | KIT-DART | p-baseline |
| kitware-ualbany | KIT-UA | p-baseline |
| MAYACHITRA | MAY | p-unified |
| Purdue-PoliMI | PUR-PM | p-CameraModelAll |
| SRI-TA2 | SRI-TA2 | p-baseline |

# Q: What is the ranking on system performance? (All Data)



# Sensitivity Analysis

**Cautionary Note:**

The experiment design for this data is not orthogonal; therefore, the interpretation may require caution

# Chosen Factors (17 out of 25)

- 25 factors (in total) from the reference and journal information
- Examined the relationship among these factors and removed factors that were:
  - highly correlated
  - less than 50 target trials

**X0: Team (system)**

**1) Target objects and Semantic changes:**

X1: People,
X2: Natural,
X3: SemanticRepurposing,
X4: SemanticRestaging,

**2) Manipulation process:**

X5: JournalSource,
X6: ManipulationCategory,
X7: CompositePixelSize,
X8: BrowserUnit,
X9: OperationArgument,
X10: Purpose,
X11: Operation,

**3) Post-processing:**

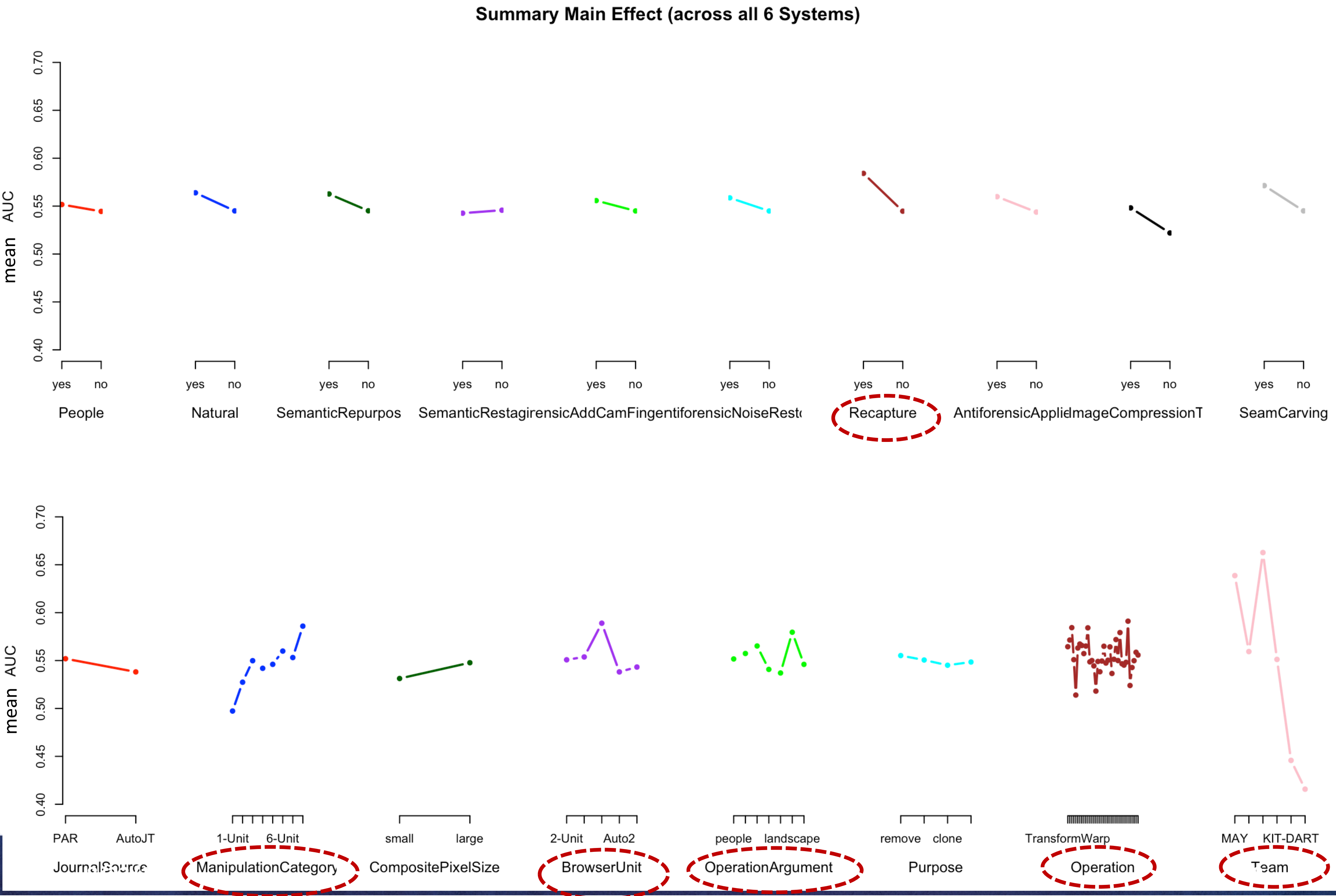
X12: AntiforensicAddCamFingerprintPRNU,
X13: AntiforensicNoiseRestoration,
X14: Recapture,
X15: AntiforensicApplied

**4) Others:**

X16: ImageCompressionTable,
X17: SeamCarving

The removed factors are:
1) FaceManipulations, 2) ReflectionManipulations,
3) ShadowManipulations, 4) SemanticRefabrication,
5) Sequence, 6) Color,
7) LaunderingMedianFiltering, 8) ImageReformat

# Q: What are the important factors affecting system performance?



# Q: What are the important factors affecting system performance?

## 17 factors

- X0: Team (6)
- 1) Target objects and semantic changes:
  - X1: People (2)
  - X2: Natural (2)
  - X3: SemanticRepurposing (2)
  - X4: SemanticRestaging (2)
- 2) Manipulation process:
  - X5: JournalSource (2)
  - X6: ManipulationCategory (8 out of 11)
  - X7: CompositePixelSize (2)
  - X8: BrowserUnit (6)
  - X9: OperationArgument (7)
  - X10: Purpose (2)
  - X11: Operation (34 out of 79)
- 3) Post-processing:
  - X12: AntiforensicAddCamFingerprintPRNU (2)
  - X13: AntiforensicNoiseRestoration (2)
  - X14: Recapture (2)
  - X15: AntiforensicApplied (2)
- 4) Others:
  - X16: ImageCompressionTable (2)
  - X17: SeamCarving (2)

## Ranking List (Top 5)

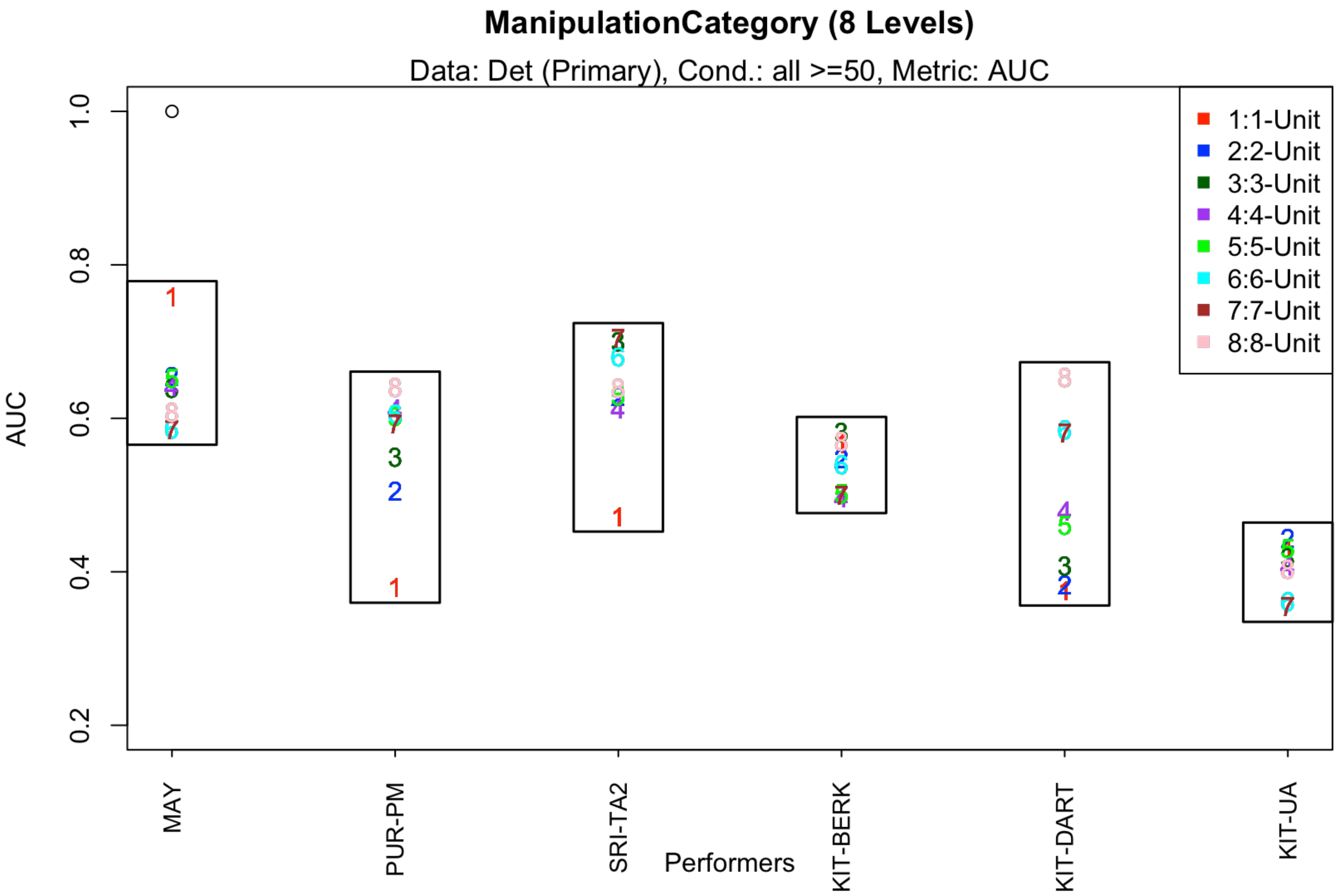
| Ranking | Factors | Effect |
|---|---|---|
| **1** | **ManipulationCategory** | **0.089** |
| **2** | **Operation** | **0.077** |
| **3** | **BrowserUnit** | **0.050** |
| **4** | **OperationArgument** | **0.043** |
| **5** | **Recapture** | **0.039** |
| 6 | SeamCarving | 0.026 |
| 7 | ImageCompressionTable | 0.026 |
| 8 | Natural | 0.019 |
| 9 | SemanticRepurposing | 0.018 |
| 10 | CompositePixelSize | 0.016 |
| 11 | AntiforensicApplied | 0.016 |
| 12 | JournalSource | 0.014 |
| 13 | AntiforensicNoiseRestoration | 0.014 |
| 14 | AntiforensicAddCamFingerprintPRNU | 0.011 |
| 15 | Purpose | 0.010 |
| 16 | People | 0.007 |
| 17 | SemanticRestaging | 0.003 |

* Note that we haven't yet calculated a statistical significance for main effect

# Top 5 Important Factor Definitions

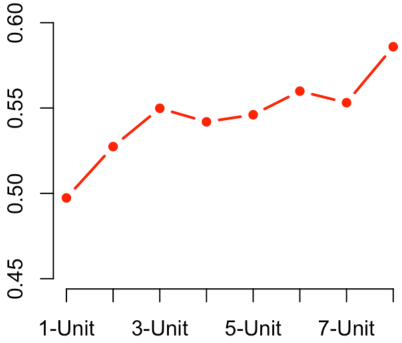
- ManipulationCategory (8)
  - The number of manipulations applied to the image that changed the image in some manor
- Operation (11)
  - Manipulation operation techniques
- BrowserUnit (5)
  - The complexity control for manipulation assignments as specified in the experimental design
- OperationArgument (7)
  - Object type (e.g., landscape) that were manipulated
- Recapture (2)
  - The recaptured image after the manipulation is done

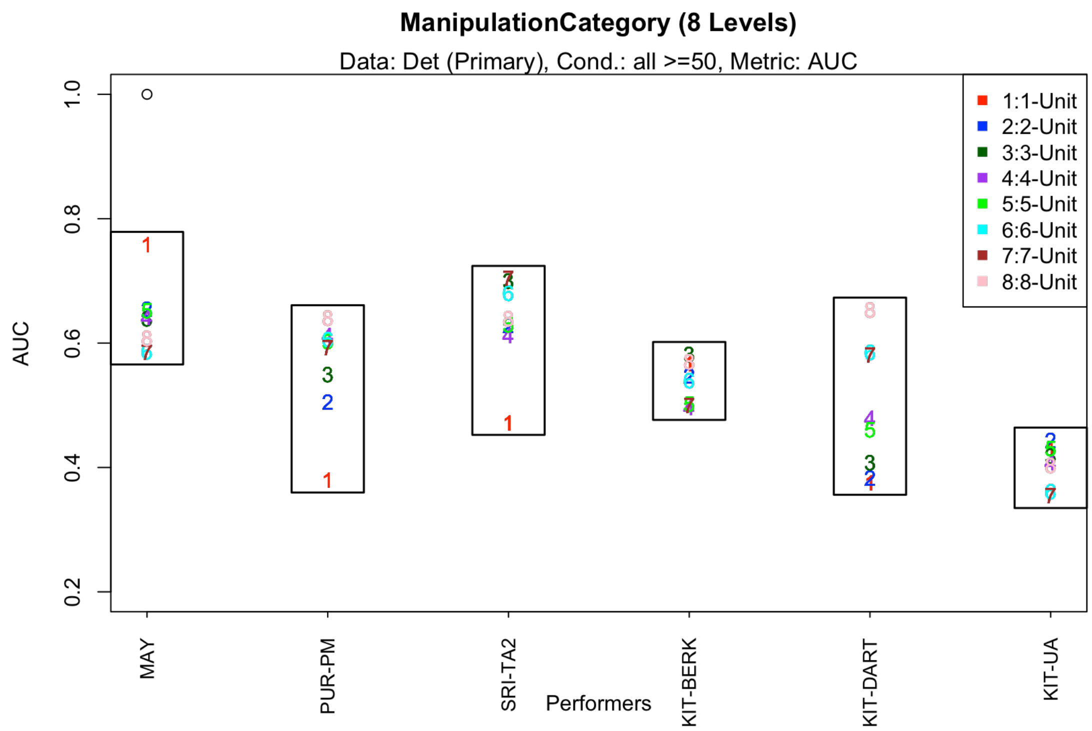
# Q: How does ManipulationCategory behave per system? Robustness? Effect comparison?

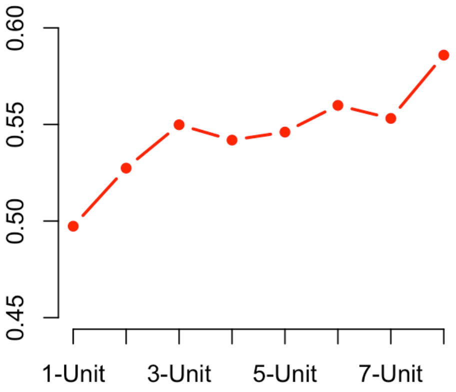


ManipulationCategory (8 Levels)

Data: Det (Primary), Cond.: all >=50, Metric: AUC

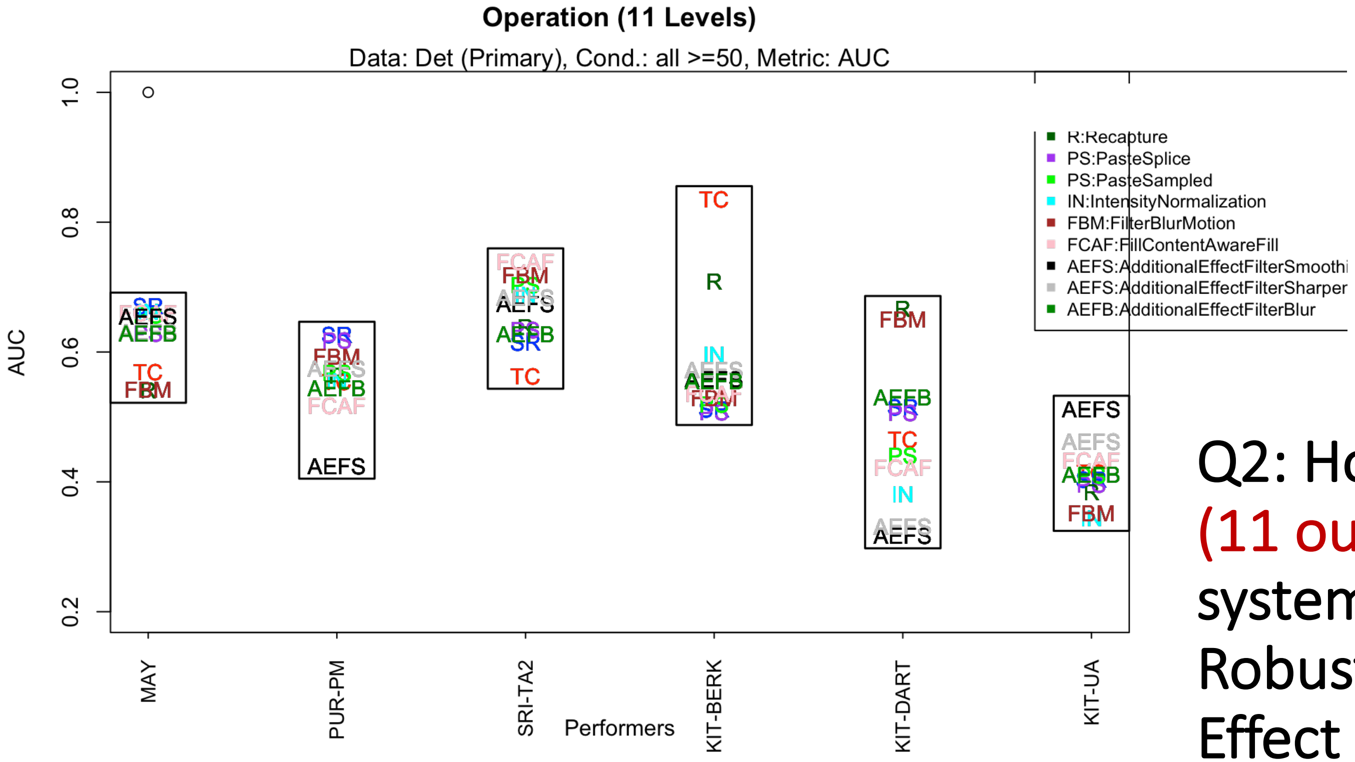


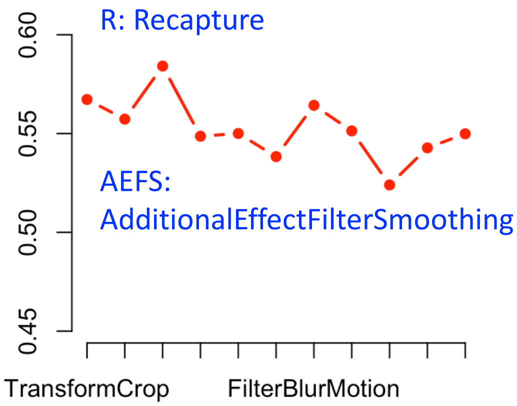




## Conclusion

1. **How does the manipulation units behave per system?**
   - For PUR-PM, SRI-TA2, and KIT-DART, the smaller units have a lower performance compared to higher units (more number of units are easier to detect)
2. **Robustness?**
   - The ranked list of manipulation units are not consistent across the 6 systems– therefore, the best/worst level depends on the team (interaction)
3. **Effect comparison?**
   - KIT-DART, PUR-PM, SRI-TA2, and MAY have larger effect on the manipulation category compared to the rest systems.

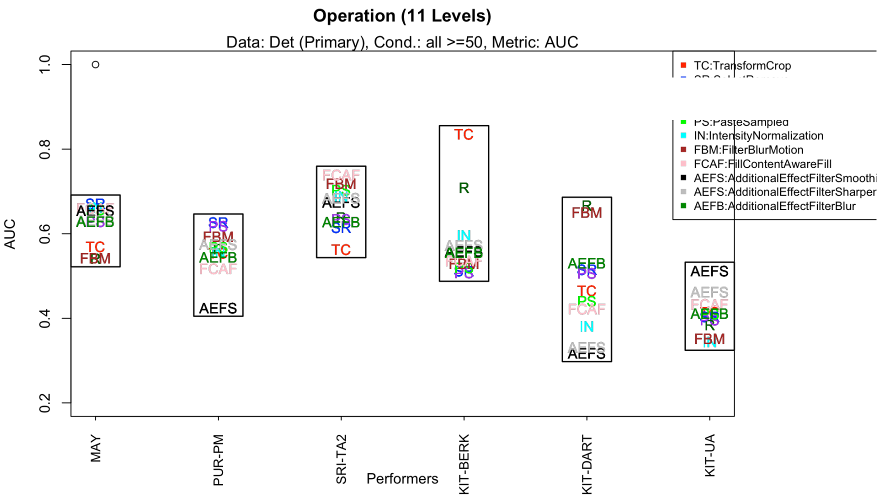


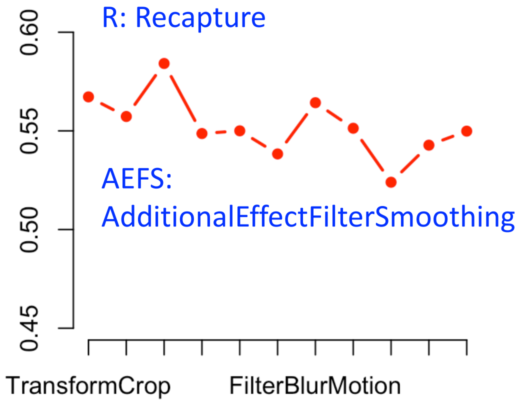


Q2: How does Operation (11 out of 79) behave per system?

Robustness?

Effect comparison?





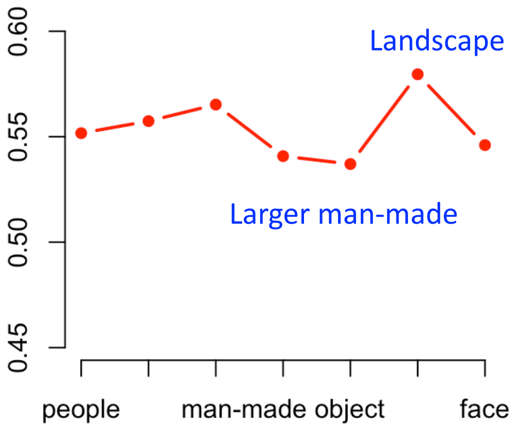
R: Recapture

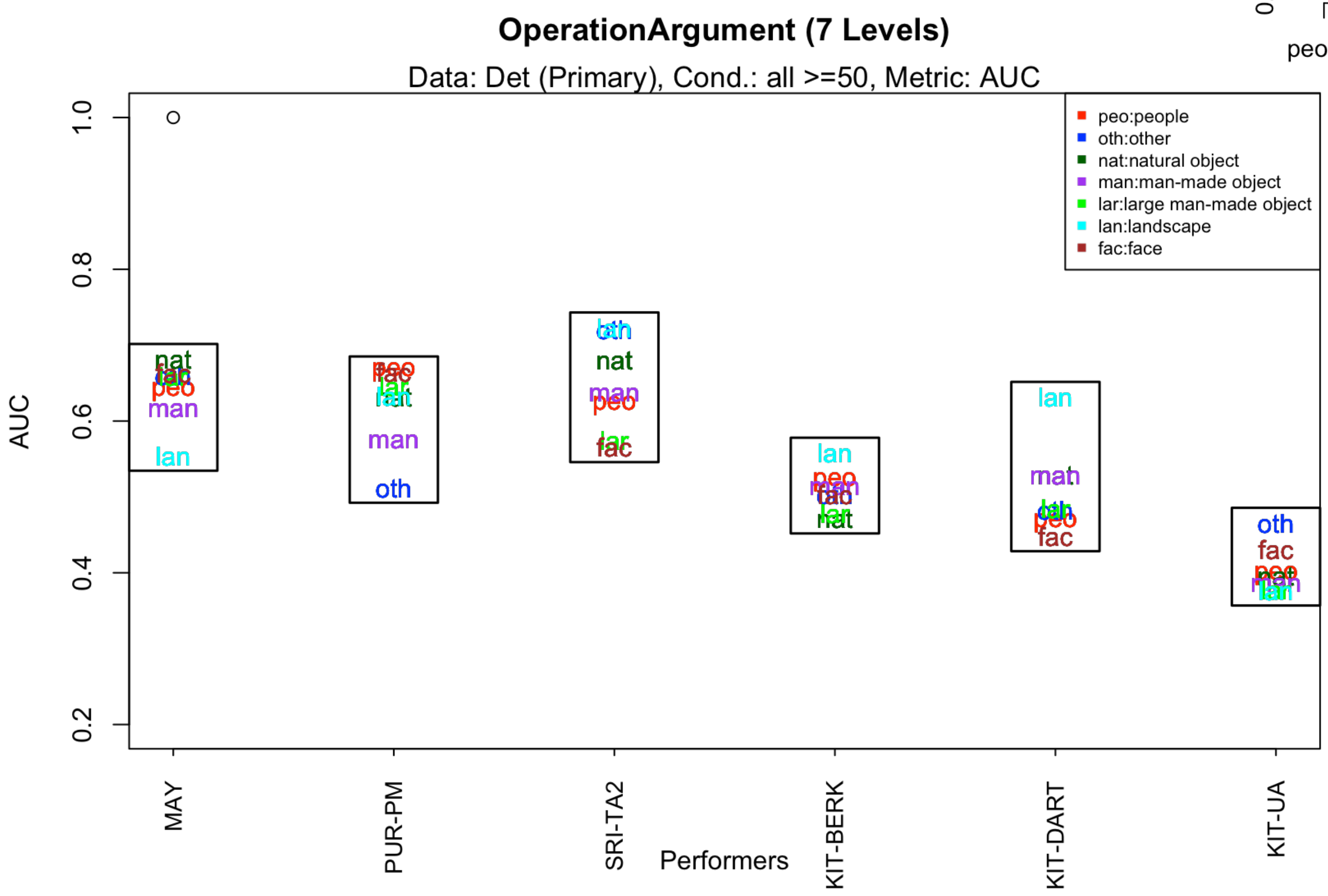
AEFS: AdditionalEffectFilterSmoothing

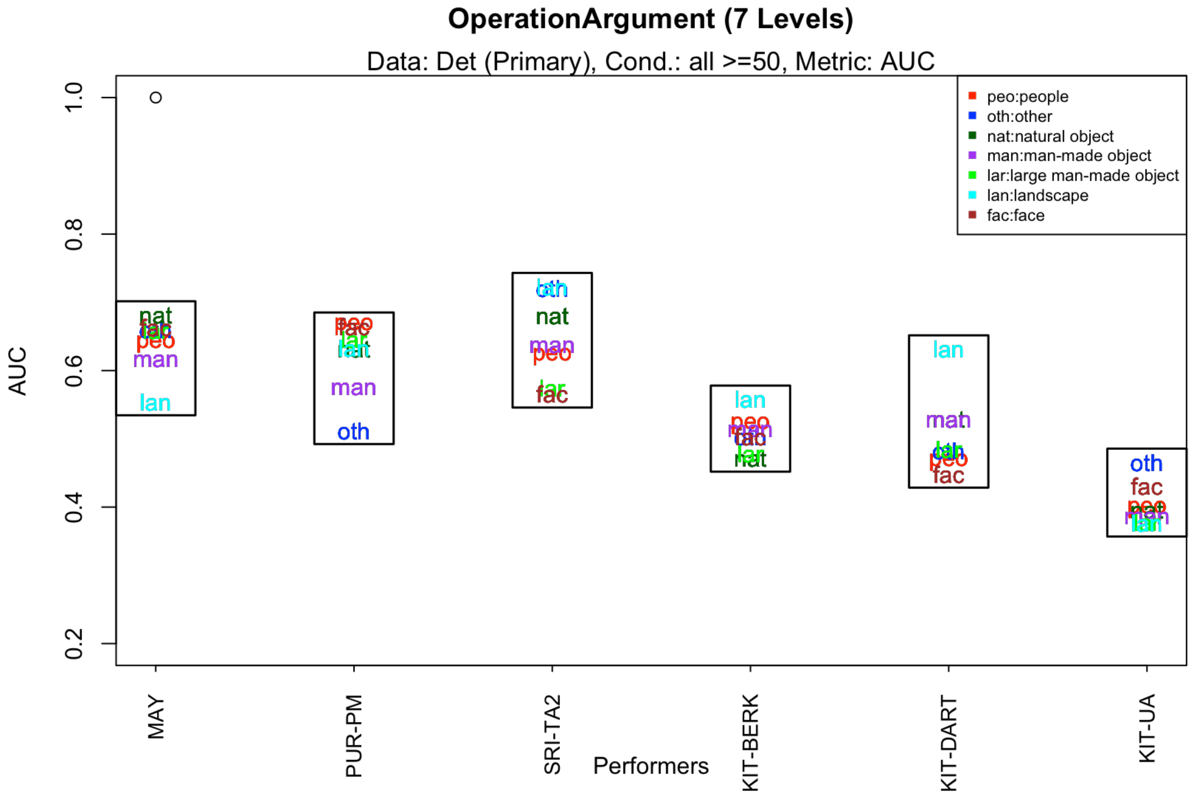
# Conclusion

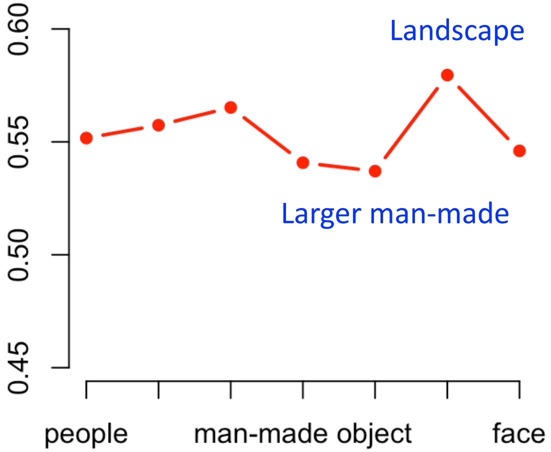
1. **How does the operation behave per system?**
   - Overall, FillContentAwareFill has the highest AUC while FilterMedianSmooth has the lowest AUC across the 6 systems
2. **Robustness?**
   - The ranked list of operation are not consistent across the 6 systems – (interaction)
3. **Effect comparison?**
   - KIT-DART and KIT-BERK have larger effect on the operation

# Q: How does OperationArgument behave per system? Robustness? Effect comparison?





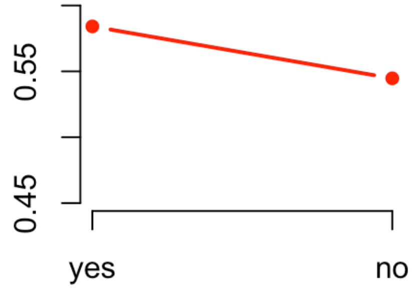


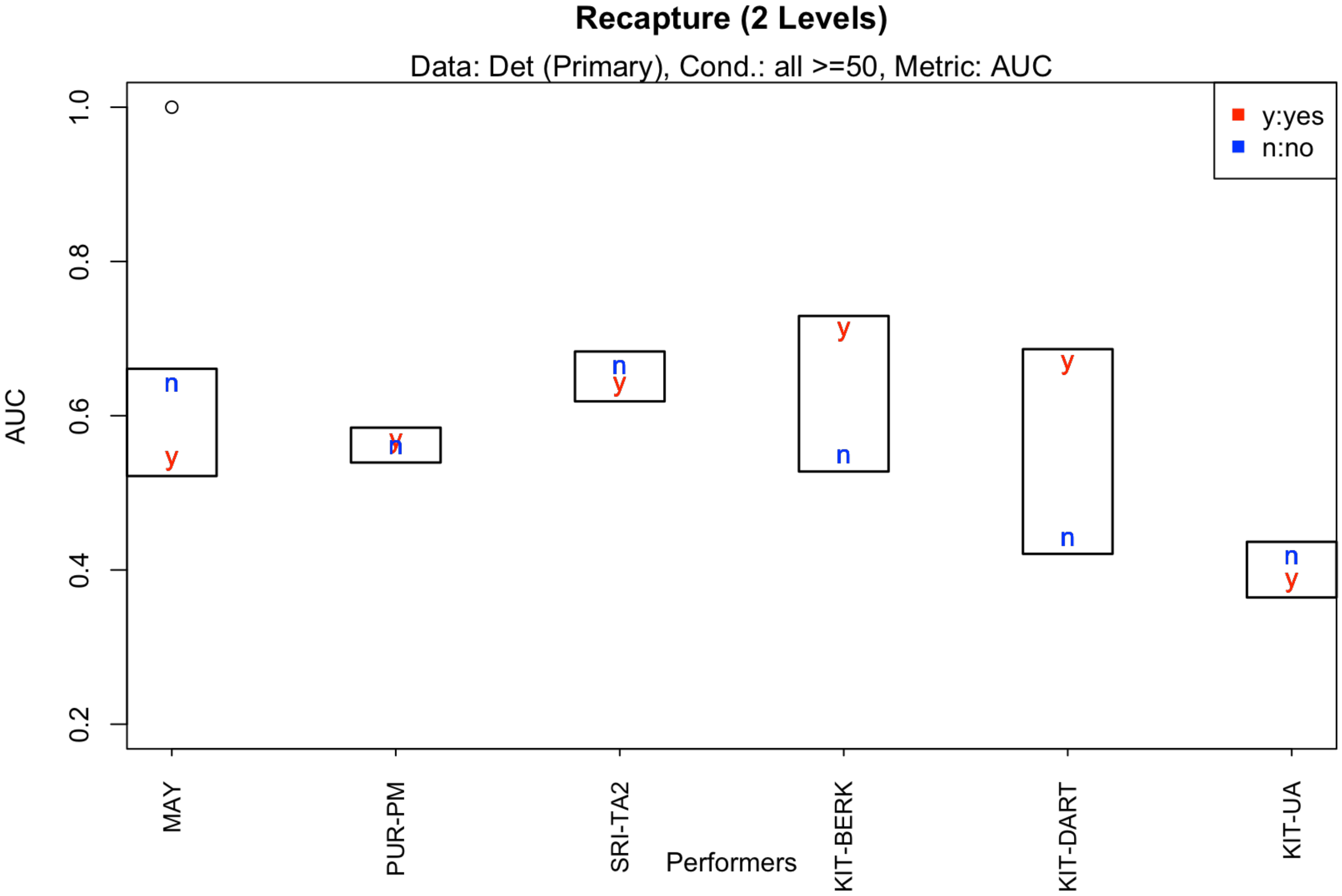


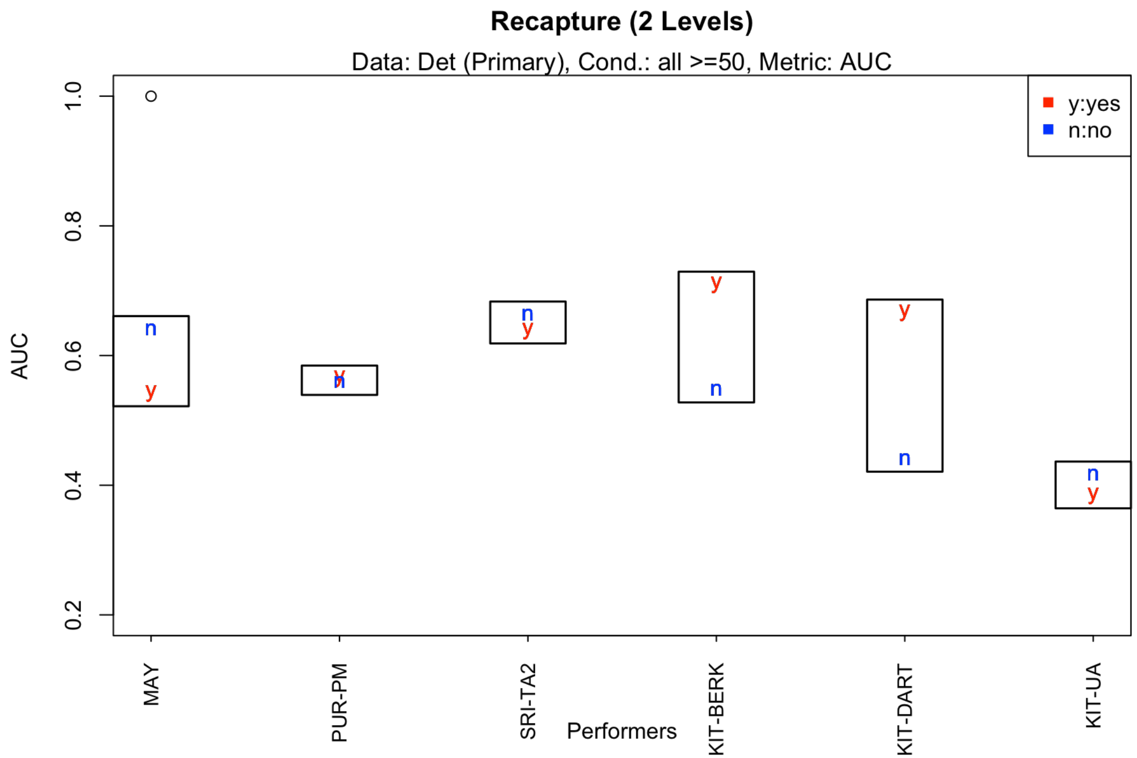
## Conclusion

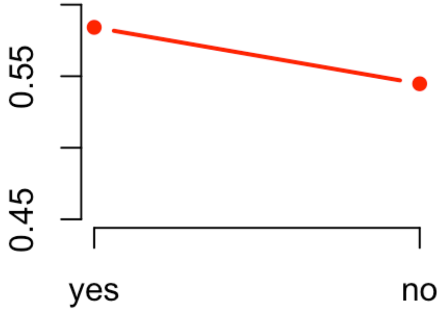
1. **How does the OperationArgument behave per system?**
   - Overall, the landscape is easier for detection across the 6 systems
2. **Robustness?**
   - The ranked list of operation argument are not consistent across the 6 systems (interaction)
3. **Effect comparison?**
   - KIT-DART, SRI-TA2, and PUR-PM have larger effect on the operation argument.

# Q: How does Recapture behave per system? Robustness? Effect comparison?







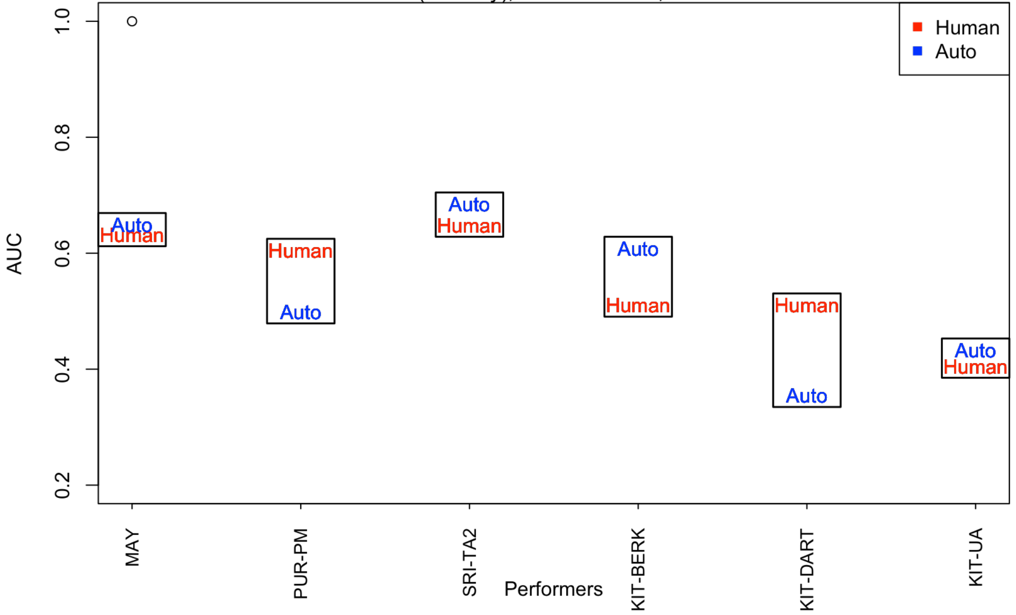


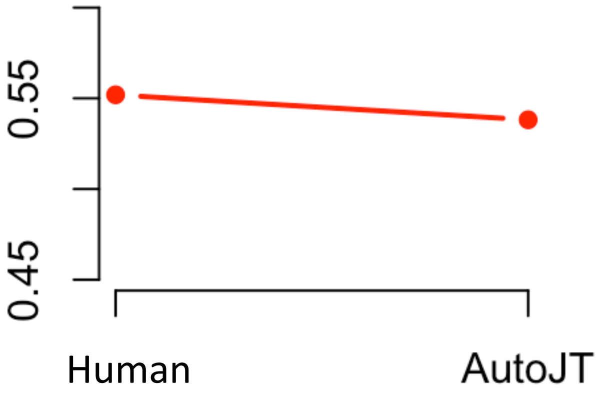
## Conclusion

1. **How does the Recapture behave per system?**
   - The recaptured images are easier to detect for especially KIT-DART, KIT-BERK, and PUR-PM.
2. **Robustness?**
   - The ranked list of recapture are not consistent across the 6 systems (interaction)
3. **Effect comparison?**
   - KIT-DART and KIT-BERK have larger effect on Recapture

**JournalSource (2 Levels)**

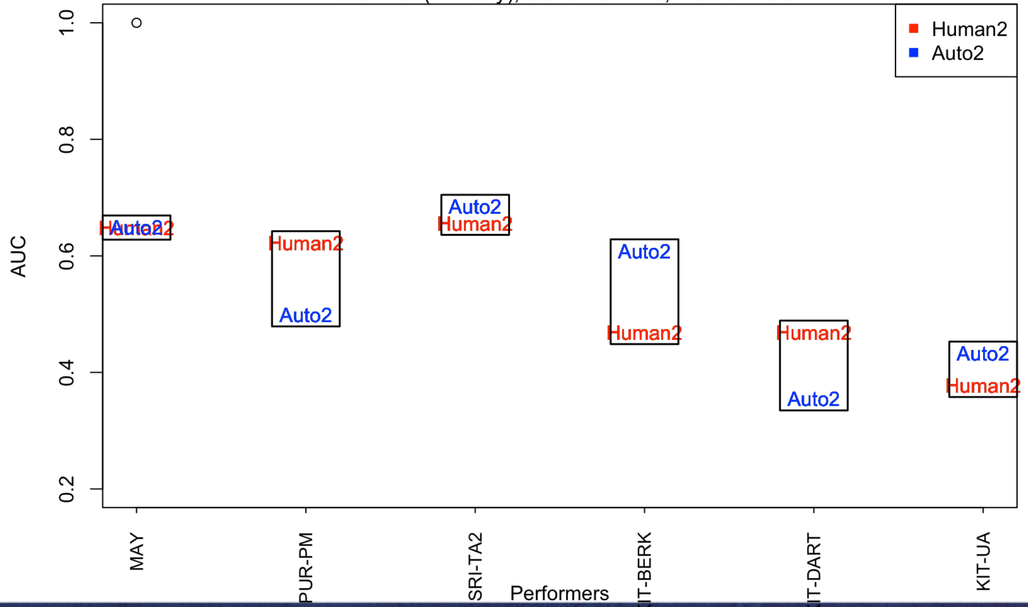
Data: Det (Primary), Cond.: all >=0, Metric: AUC



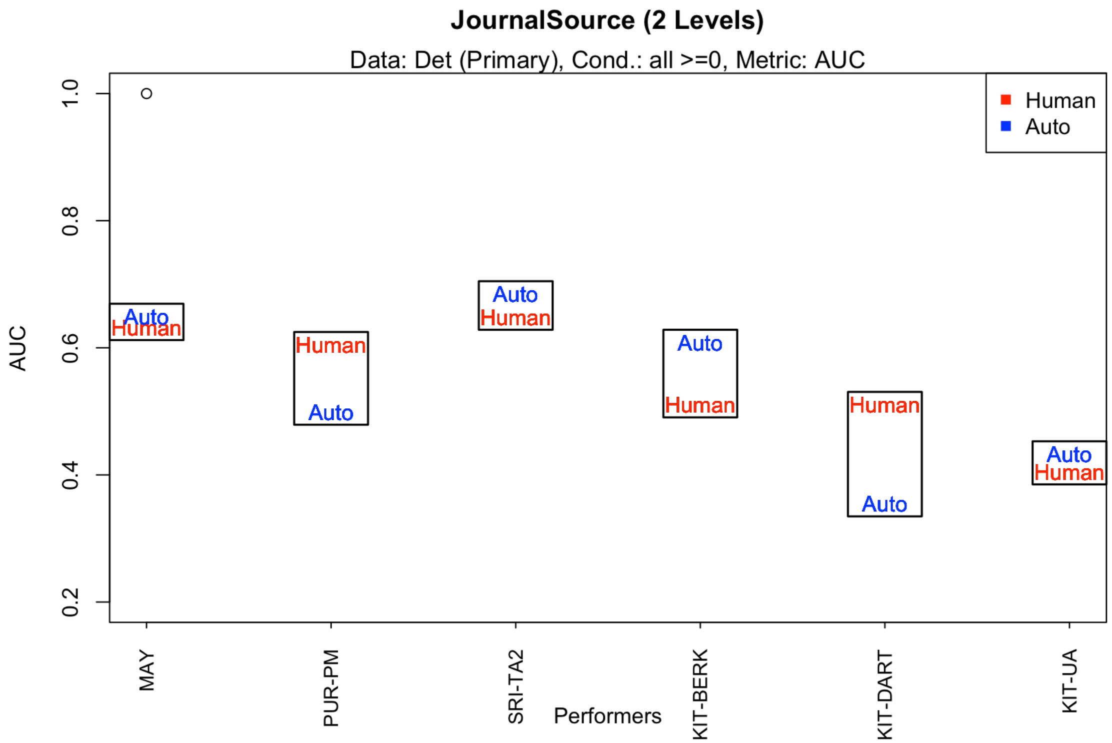


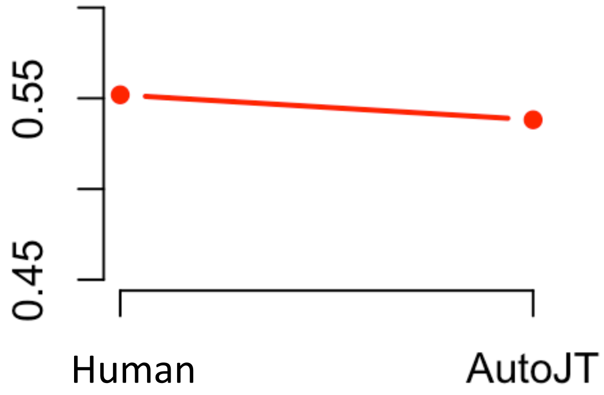
**JournalSource_2Unit (2 Levels)**

Data: Det (Primary), Cond.: all >=0, Metric: AUC



Q: How does the system perform on JournalSource (Human vs Auto)?
Robustness?
Effect comparison?





## Conclusion

1. **Comparison of Human vs Auto manipulations?**

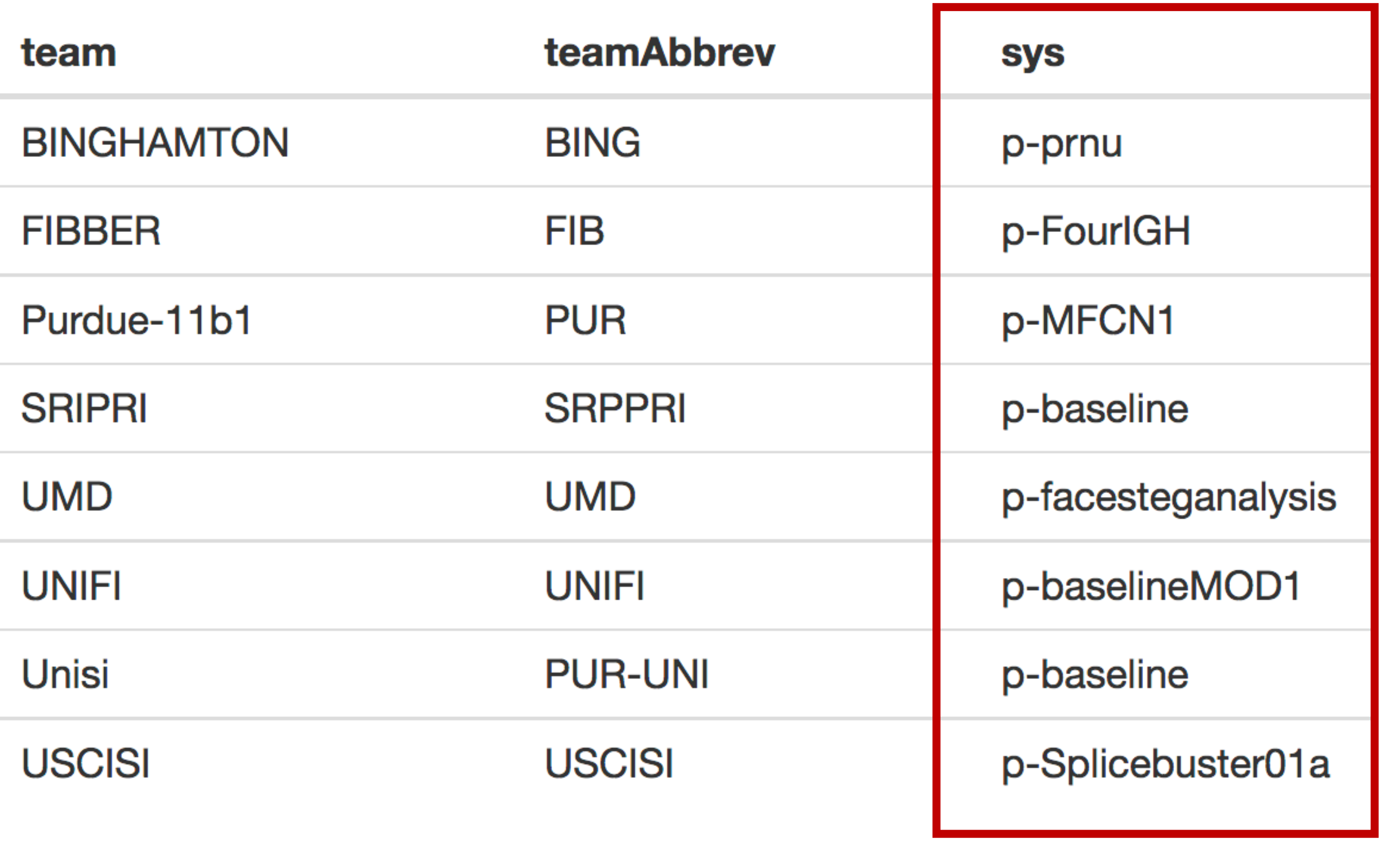
- The Human-manipulations are easier to detect for SRI-TA2 KIT-BERK, MAY, and KIT-UA while the Auto-manipulations are easier to detect to for PUR-PM and KIT-DART.

2. **Robustness?**
   - The ranked list of journal source are not consistent across the 6 systems (interaction)

3. **Effect comparison?**
   - KIT-DART, KIT-BERK, and PUR-PM have larger effect on the Journal source.

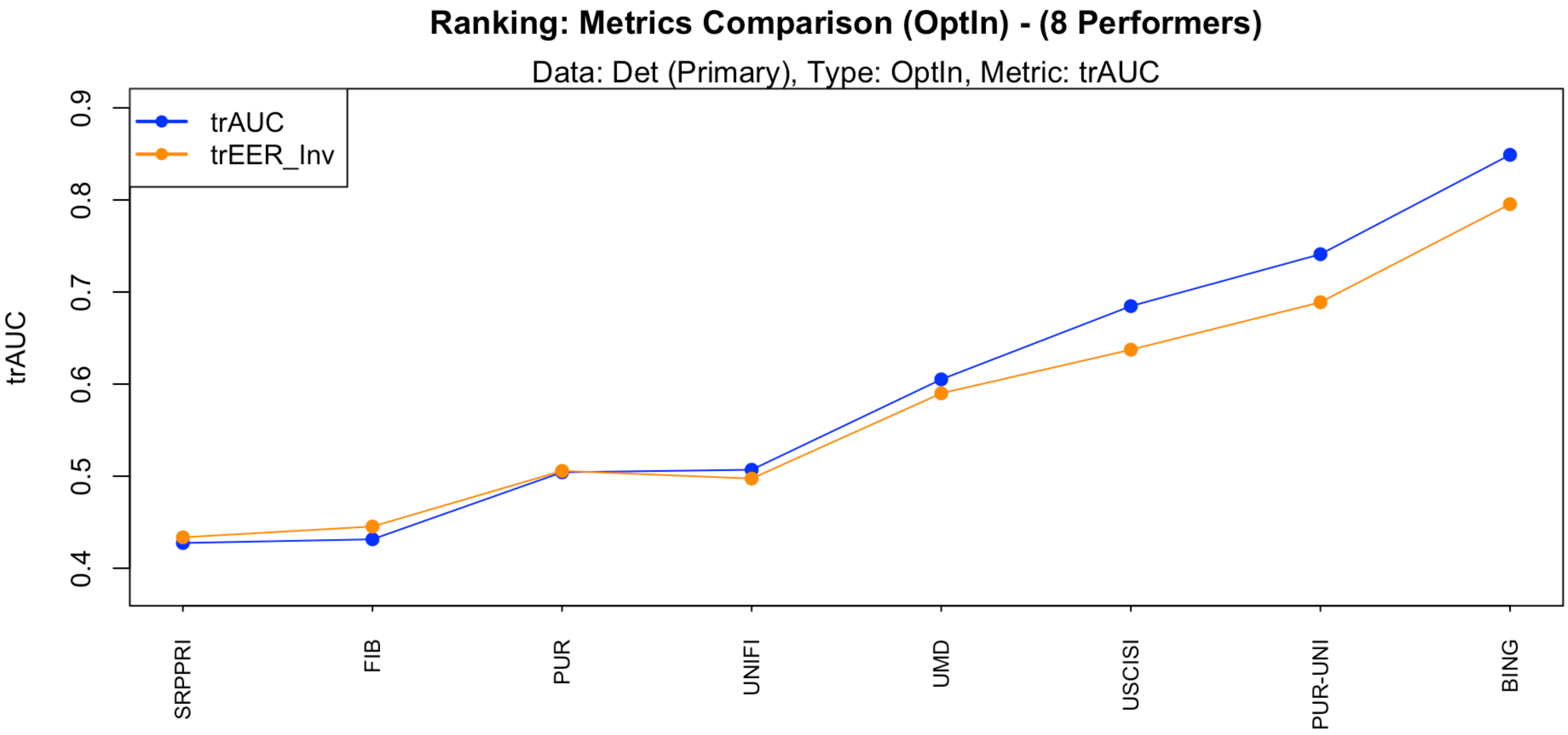
# OptIn Data: 8 Teams

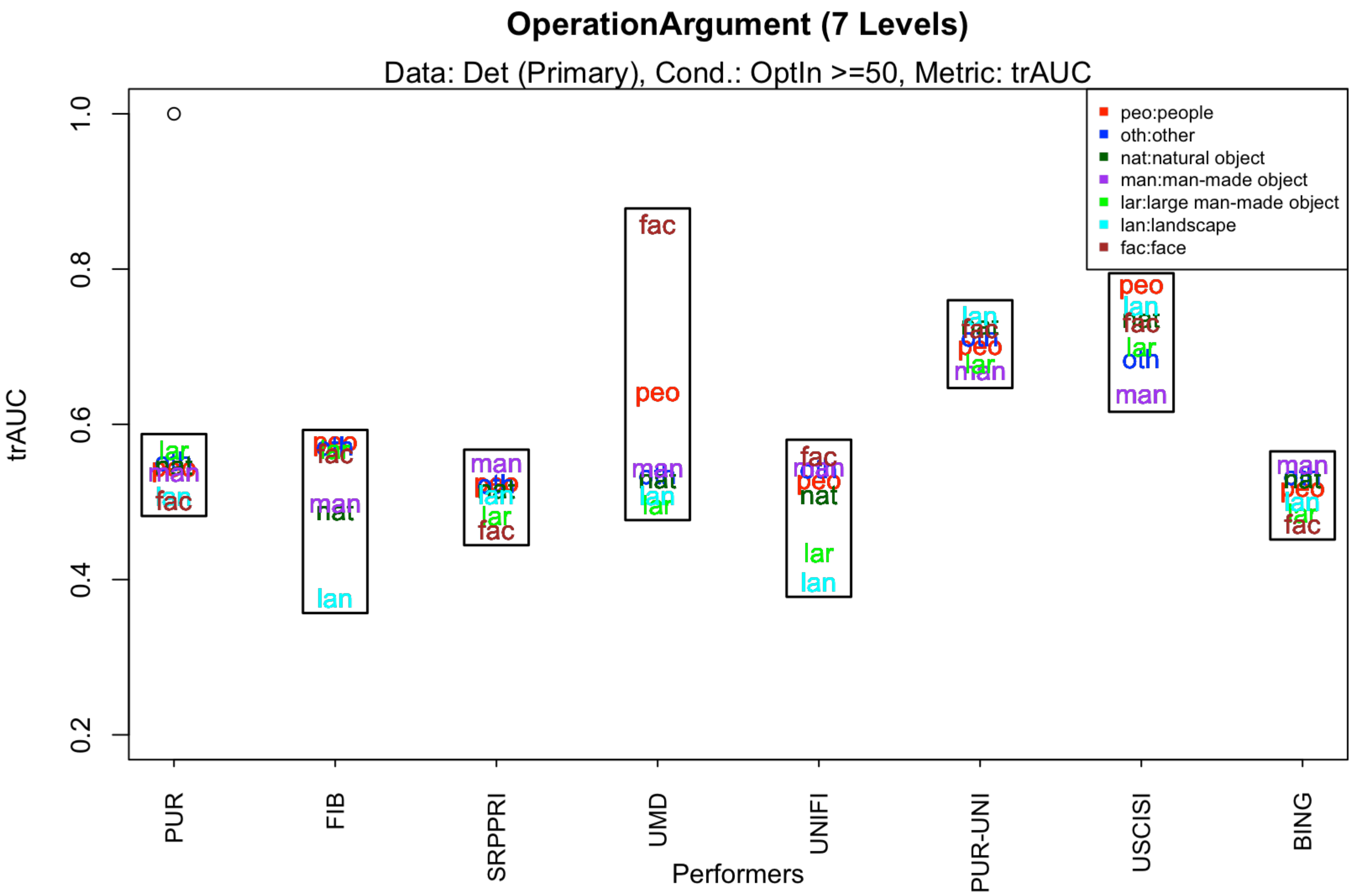
Primary-only

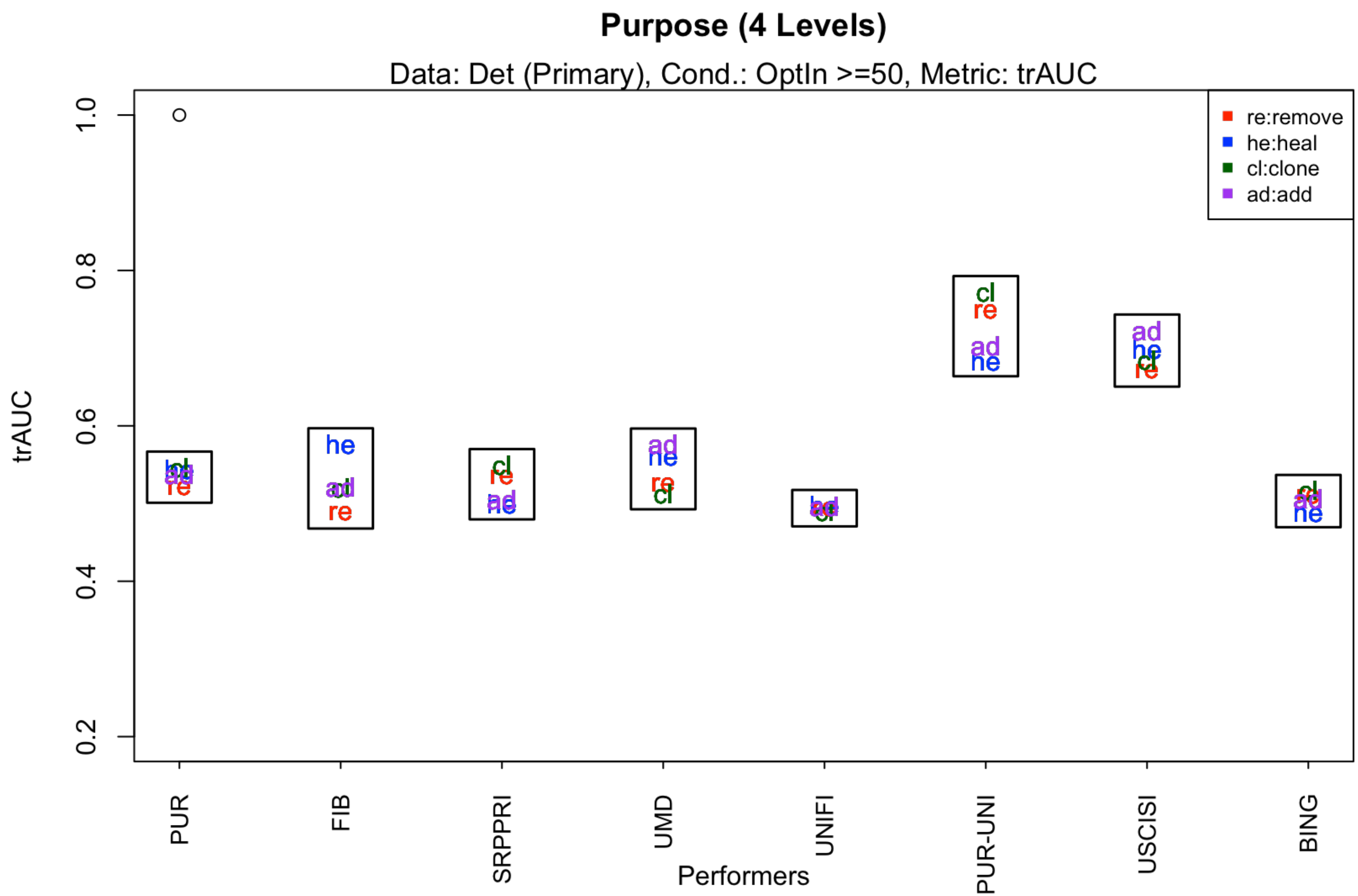


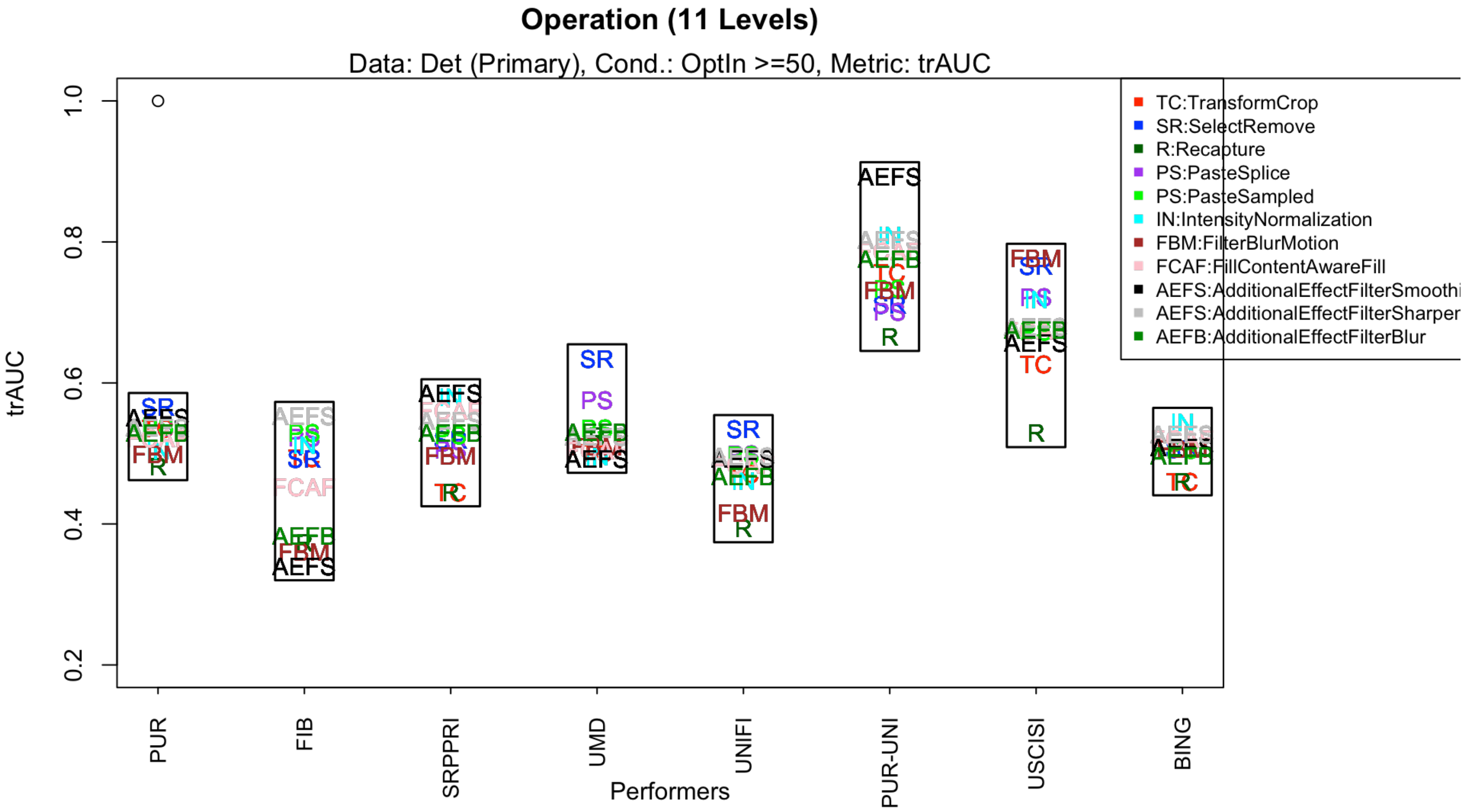
| team | teamAbbrev | sys | TRR |
|---|---|---|---|
| BINGHAMTON | BING | p-prnu | 0.08 |
| FIBBER | FIB | p-FourIGH | 0.93 |
| Purdue-11b1 | PUR | p-MFCN1 | 0.96 |
| SRIPRI | SRPPRI | p-baseline | 0.13 |
| UMD | UMD | p-facesteganalysis | 0.04 |
| UNIFI | UNIFI | p-baselineMOD1 | 0.43 |
| Unisi | PUR-UNI | p-baseline | 0.92 |
| USCISI | USCISI | p-Splicebuster01a | 0.95 |

# Q: What is the ranking on system performance? (OptIn)




OperationArgument (7 Levels)
Data: Det (Primary), Cond.: OptIn >=50, Metric: trAUC
trAUC
0.2
0.4
0.6
0.8
1.0
Performers
PUR
FIB
SRPPRI
UMD
UNIFI
PUR-UNI
USCISI
BING
peo:people
oth:other
nat:natural object
man:man-made object
lar:large man-made object
lan:landscape
fac:face
fac
peo
man
nat
lan
lar
oth


Purpose (4 Levels)
Data: Det (Primary), Cond.: OptIn >=50, Metric: trAUC
trAUC
1.0
0.8
0.6
0.4
0.2
re:remove
he:heal
cl:clone
ad:add
PUR
FIB
SRPPRI
UMD
UNIFI
PUR-UNI
USCISI
BING
Performers



Optout target:

PSL&PSM

All except
TCP, INT

PSL&PSM

PRNU

# Detection Analysis Summary

- The number of manipulation units, the type of operations, and the type of object manipulated matter to system performance
  - The manipulation process group is important for affecting system performance
- Although the average effect (6 systems) of Human vs Auto-manipulation is small, it is larger for some systems
- Recaptured images are easier to detect for some (but not all) systems
- For OptIn case, the systems mostly met their design-purpose with a few exceptions

# Future Analysis Plan

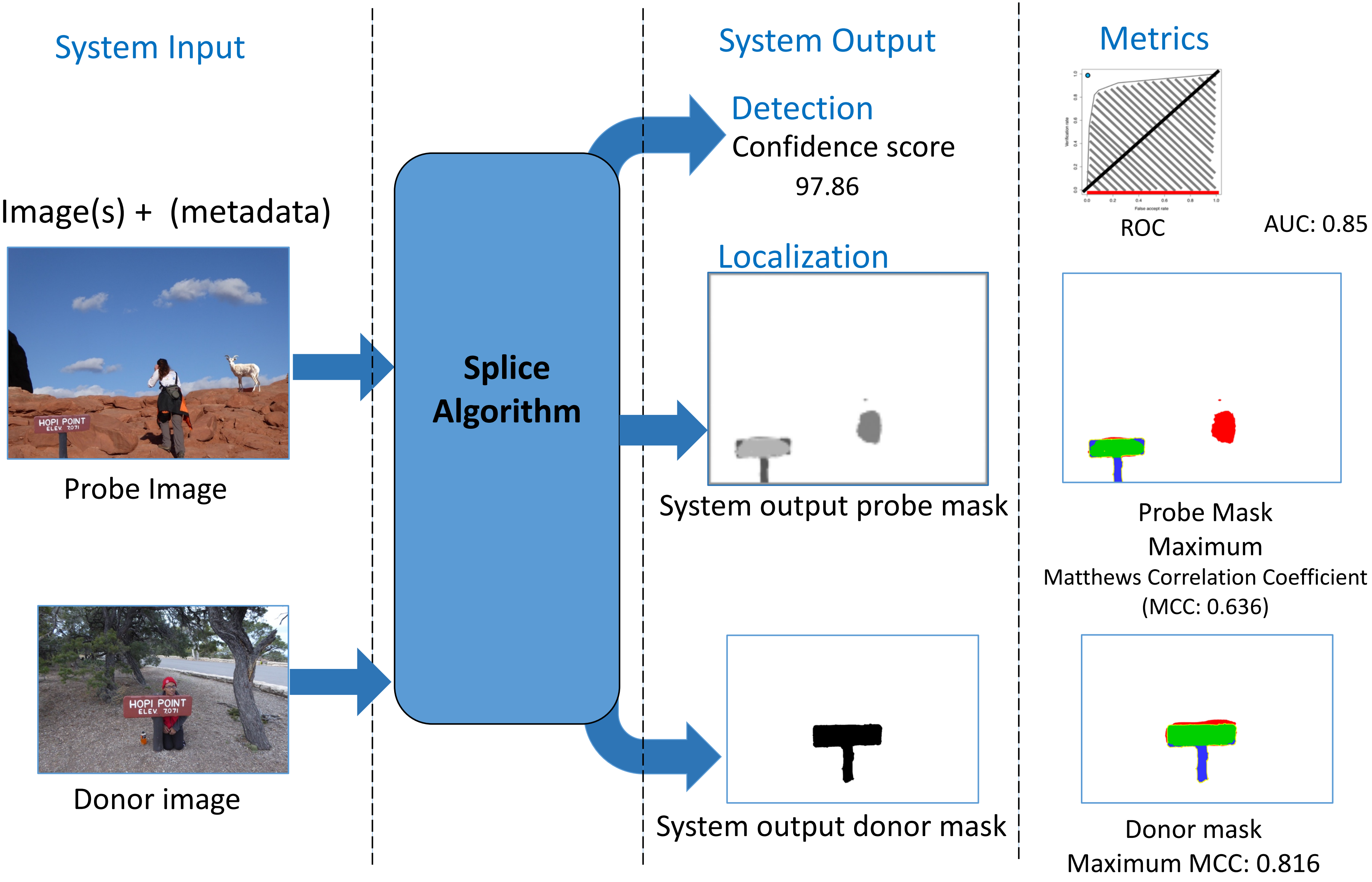
- Calculate statistical significance
- Estimate interaction effect among factors
- Improve experiment design (at least balanced data)
- Apply additional methodologies to characterize the system performance better
- TA2 infrastructure to perform robust system testing and its analysis

# Outline

✓ NIST NC2017 Evaluation Overview

- Detection
  - ✓ Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ➢ Splice – paired: Task, Data, Results
  - Video: Task, Data, Result
- Localization
  - Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# Splice Detection and Localization Evaluation Model



System Input

Image(s) + (metadata)

Probe Image

Donor image

Splice Algorithm

System Output

Detection

Confidence score

97.86

Localization

System output probe mask

System output donor mask

Metrics

ROC

AUC: 0.85

Probe Mask

Maximum

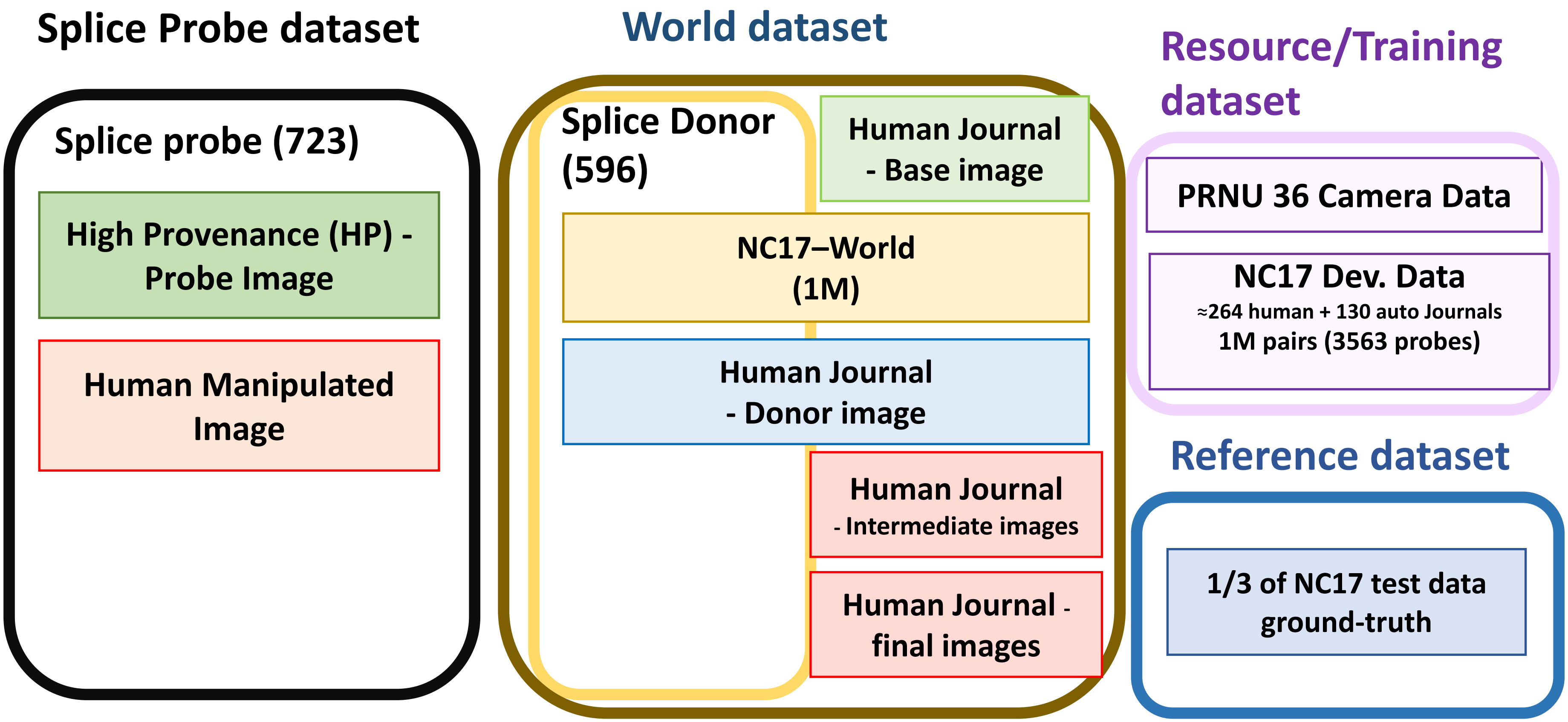
Matthews Correlation Coefficient

(MCC: 0.636)

Donor mask

Maximum MCC: 0.816

# NC17 Manipulation Splice Datasets



Splice Probe dataset

Splice probe (723)

High Provenance (HP) - Probe Image

Human Manipulated Image

World dataset

Splice Donor (596)

Human Journal - Base image

NC17–World (1M)

Human Journal - Donor image

Human Journal - Intermediate images

Human Journal - final images

Resource/Training dataset

PRNU 36 Camera Data

NC17 Dev. Data

≈264 human + 130 auto Journals

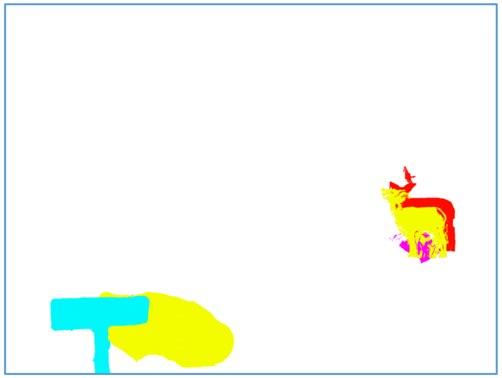
1M pairs (3563 probes)

Reference dataset

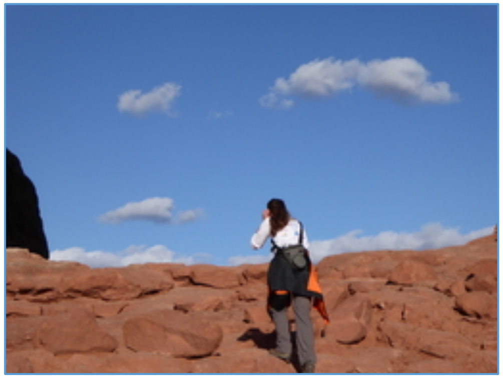
1/3 of NC17 test data ground-truth

# Splice Detection and Localization Example

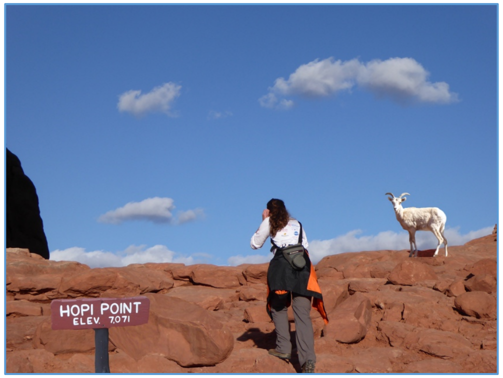
Color Composite Mask



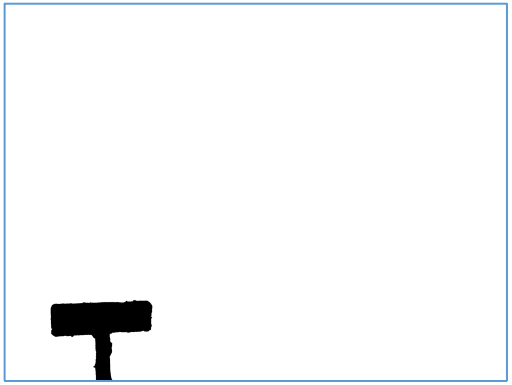
Base Image



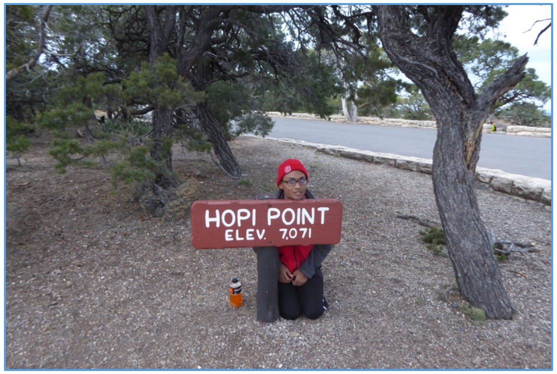
Probe Image



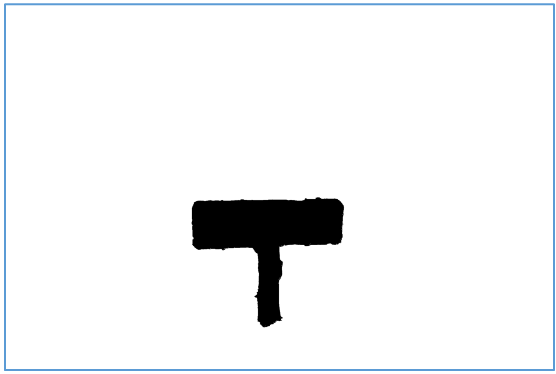
Reference Probe Mask Given the Donor



Donor Image

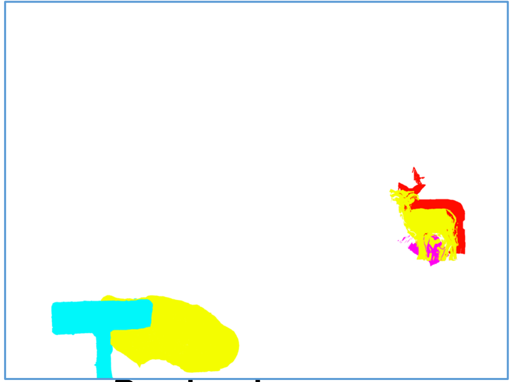


Donor Mask

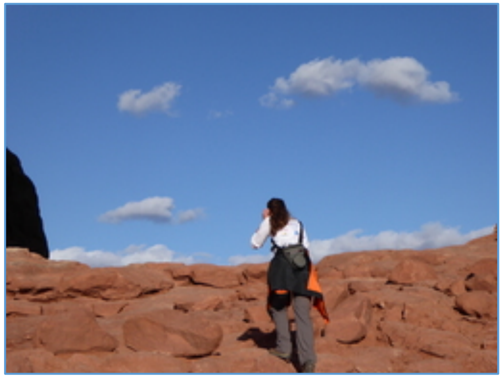


# Splice Detection and Localization Example

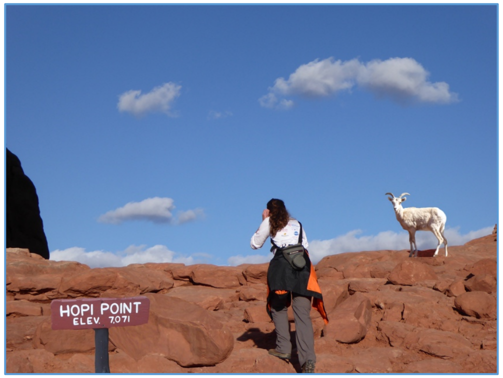
Color Composite Mask



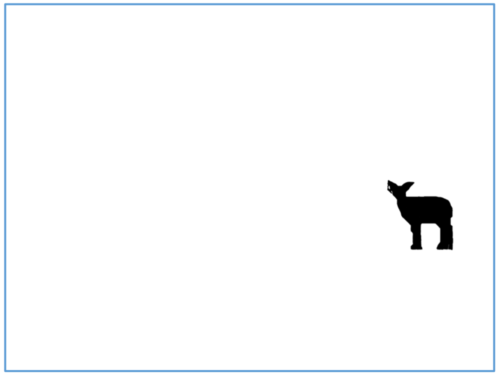
Base Image



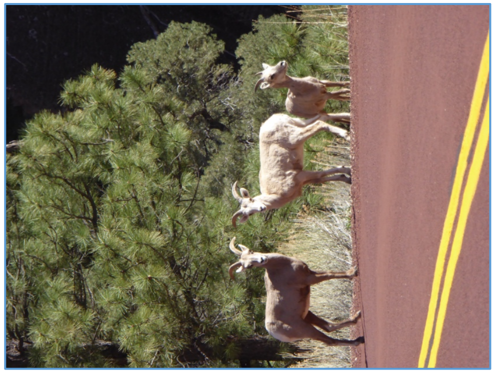
Probe Image



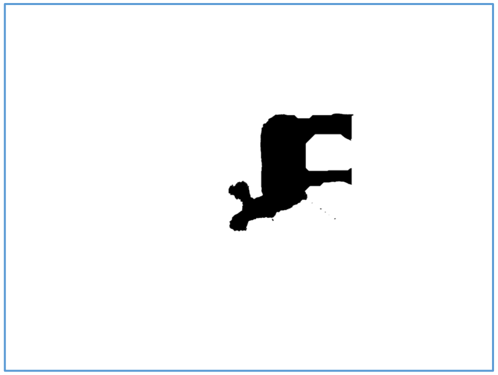
Reference Probe Mask
Given the Donor



Donor Image



Donor Mask



# NC17 Splice Manipulation Eval. Results
## - Detection AUC

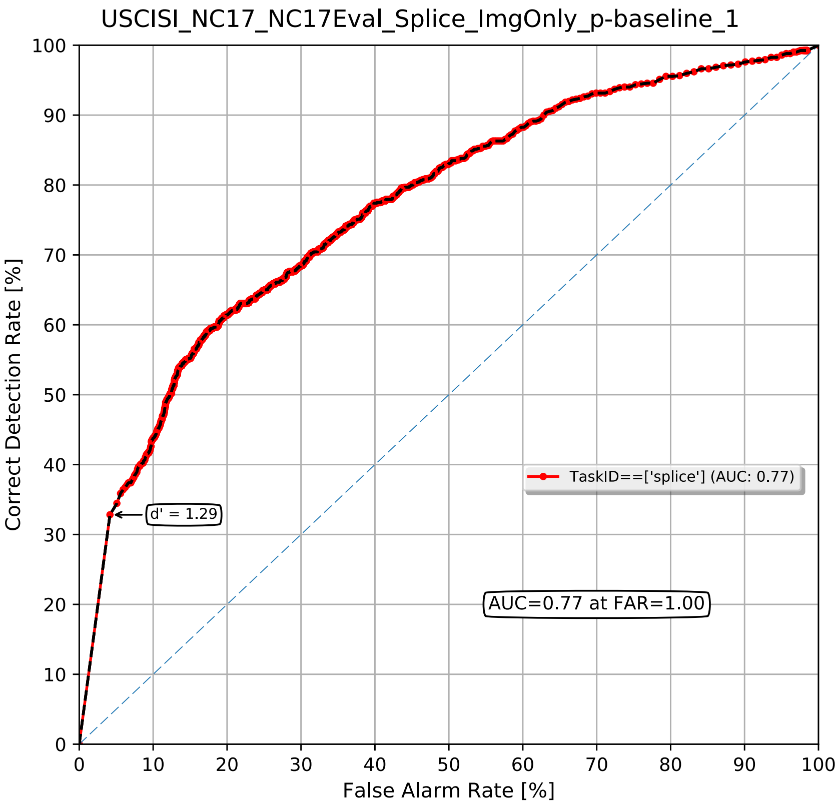
- Condition: Image Only
- **Data: All human manipulated**

| TeamID | SystemID | All_AUC | OptIn_AUC | OptIn_TRR | OptIn_trAUC |
|---|---|---|---|---|---|
| FIBBER | p-FourIGH_2 | | 0.528087 | 0.37 | 0.583492 |
| UNIFI | c-baselineMOD4_1 | | 0.526053 | 0.04 | 0.45844 |
| | p-baselineMOD3_1 | | 0.526492 | 0.04 | 0.459056 |
| USCISI | p-baseline_1 | 0.767365 | | | |

| | |
|---|---|
| AUC | Area Under the Curve - Eval Plan Section 6.1.2 |
| OptIn | Scores for which the system processed the probe. |
| trAUC | AUC for the trials NOT opted out of |
| TRR | Trial Response Rate - Fraction of evaluation probes NOT opted out of |

# NC17 Splice Manipulation Eval. Results
# - Detection ROC (All probes)

- Condition: Image Only; All probes; All human manip.
- USCISI



Splice Task, AUC: 0.767365

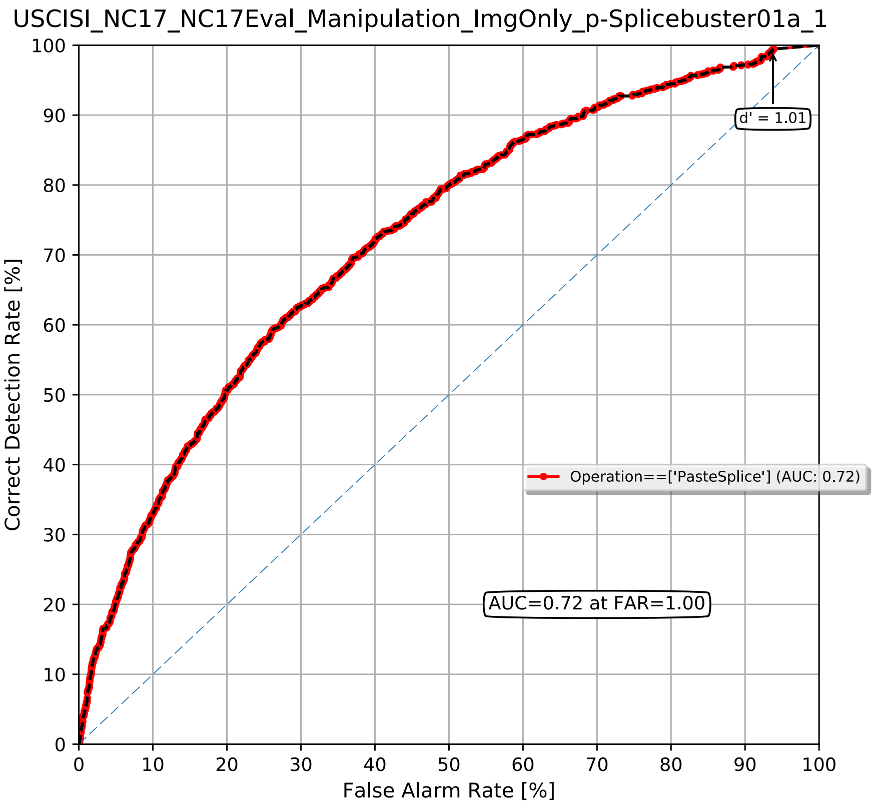


Image Manipulation Task, AUC: 0.722241

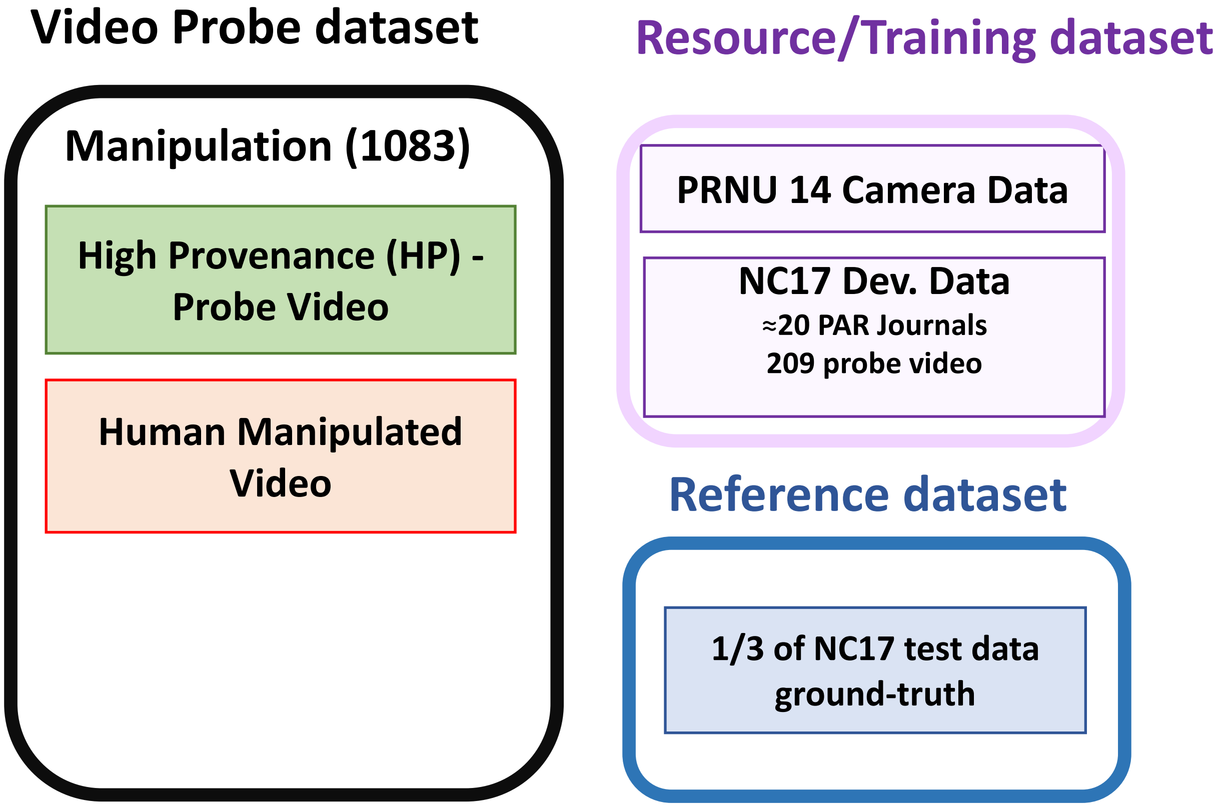
Selective Scoring – PasteSplice

# Outline

✓ NIST NC2017 Evaluation Overview

- Detection
  - ✓ Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ✓ Splice – paired: Task, Data, Results
  - Video: Data, Result
  - Discussion and Future Opportunities
- Localization
  - Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
  - Discussion and Future Opportunities
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
  - Discussion and Future Opportunities
- Summary

# NC17 Manipulation Video Datasets



Video Probe dataset

Manipulation (1083)

High Provenance (HP) - Probe Video

Human Manipulated Video

Resource/Training dataset

PRNU 14 Camera Data

NC17 Dev. Data
≈20 PAR Journals
209 probe video

Reference dataset

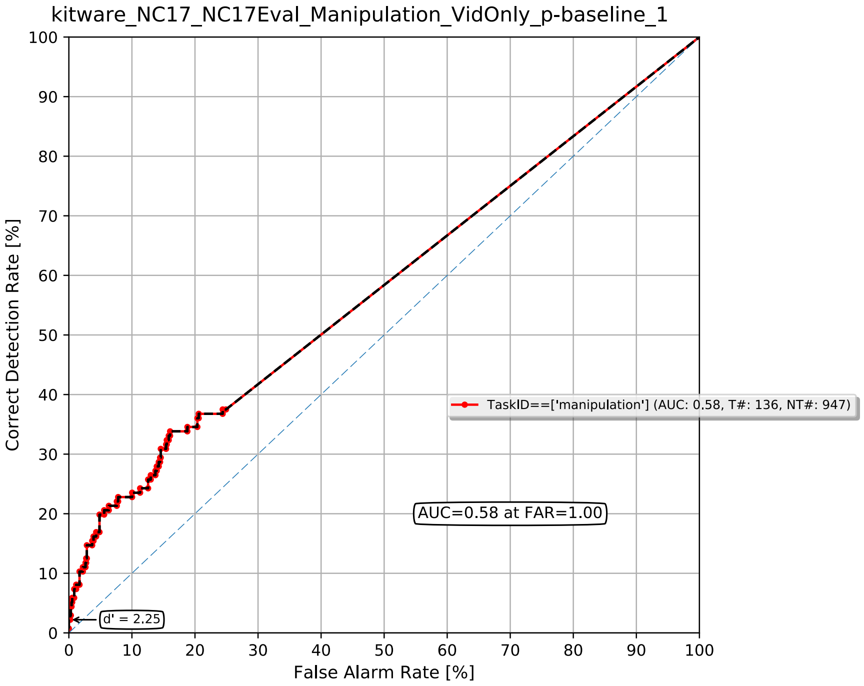
1/3 of NC17 test data ground-truth

# NC17 Video Manipulation Eval. Results

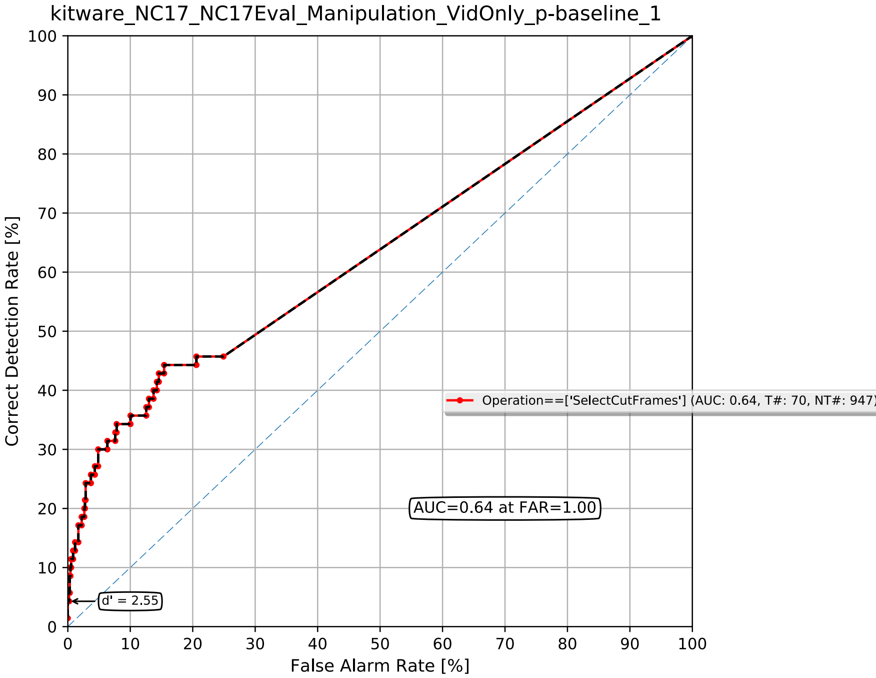
| Video | TeamID | All_AUC | OptIn_AUC | OptIn_TRR | OptIn_trAUC |
|---|---|---|---|---|---|
| VidMeta | SRIPRI | | 0.493715 | 0.32 | 0.492333 |
| VidOnly | kitware | 0.580436 | | | |

# Video Manipulation Detection without Metadata

- Kitware, All Probe Data



All operations; AUC = 0.58



Selective Scoring
Drop Frame; AUC = 0.64

# Outline

✓NIST NC2017 Evaluation Overview

✓Detection

- ✓Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
- ✓Splice – paired: Task, Data, Results
- ✓Video: Data, Result

- Localization
  - ➢Image – single: Metrics, Results, Analysis
  - Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# Localization (Mask) Scorer

- Evaluate the accuracy of a system output mask to a reference mask (localization)
- Evaluate on **ALL** target trials only (manipulations)
  - If the system output mask for a trial was not provided, the scorer uses and empty system mask for that trial
  - Not accounting for detection confidence scores
- Evaluation metrics (for both binary and grayscale)
  - ***Maximum MCC (Matthews Correlation Coefficient)***
  - NMM (Nimble Mask Metric)
  - WL1 (Weighted L1 Loss)
    - Binary: BWL1
    - Grayscale: GWL1
- Code speedup –
  - Using "USCISI_NC17_NC17Eval_Manipulation_ImgOnly_c-Autoencoder01a_1"
  - 20170526 release – 64 hours
  - 20170607 release – 9.9 hours (Single Thread, '—speedup' option)
    – 1.4 hours (10 Threads)

# NC17 Image Manipulation Localization

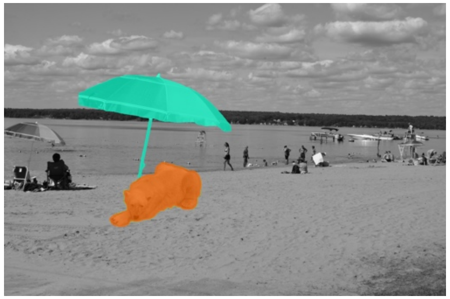
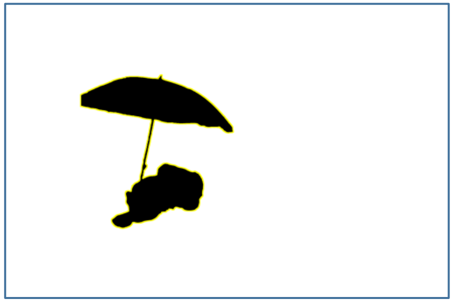
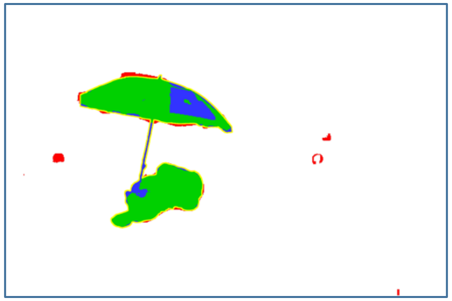
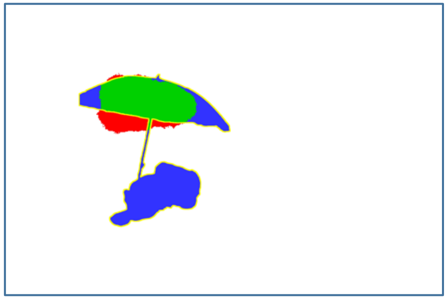
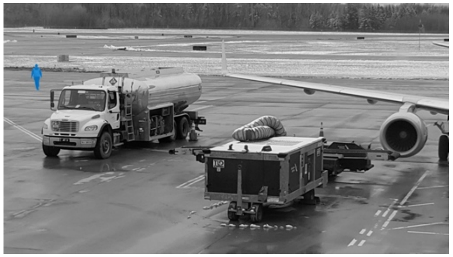
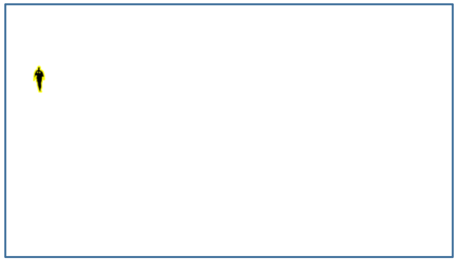
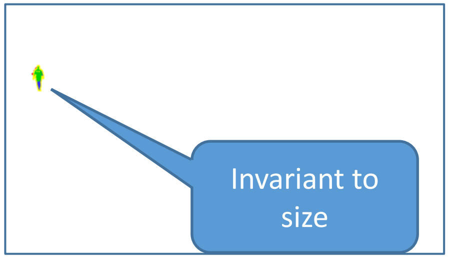
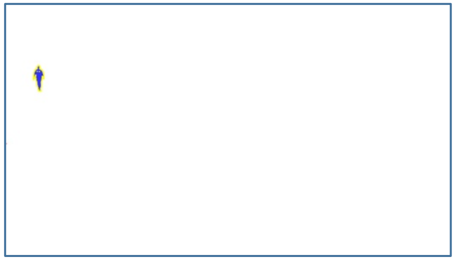
- Participate Summary

| Team | | Image Mani. Loc. | |
|---|---|---|---|
| | | All Probe | OptIn |
| BIN | Binghamton University | | 1 |
| FIB | Honeywell ACS Laboratories | | 1 |
| KIT | Kitware<br>UC Berkeley<br>Dartmouth College<br>University at Albany, SUNY | 4 | |
| MAY | MAYACHITRA<br>Naval Air Warfare Center, China Lake<br>UC Riverside | 1 | 2 |
| PUR | Purdue<br>Politecnico di Milano, Italy<br>University of Siena<br>Univ. of Notre Dame; University of Campinas, Brazil | 3 | 1 |
| SRI-TA2 | SRI International, Princeton (Ajay Divakaran) | 1 | |
| SRPPRI | SRI International, Princeton (Jeffrey Lubin) | - | 1 |
| UMD | University of Maryland, College Park | - | 1 |
| UNIFI | University of Florence, FENCE, Prato, Italy | - | 2 |
| USCISI | University of Southern California, ISI | - | 5 |
| 10 teams | 16 organizations, 23 systems | 9 | 14 |

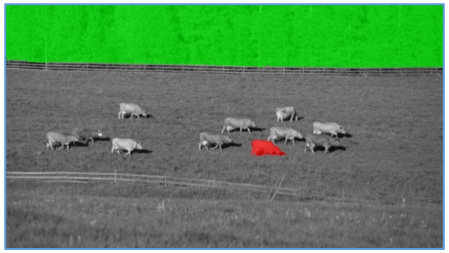
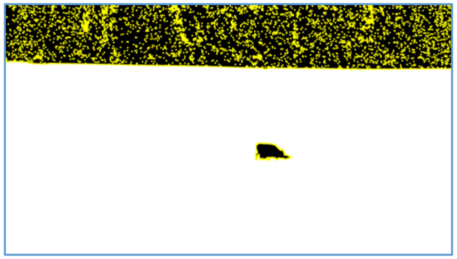
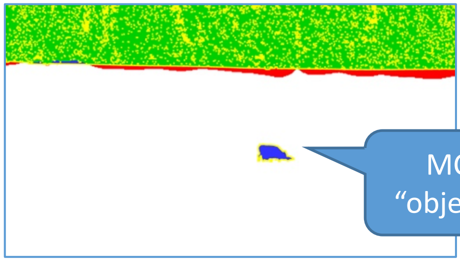
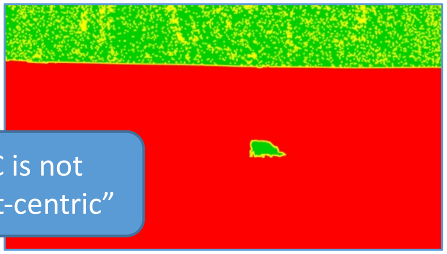
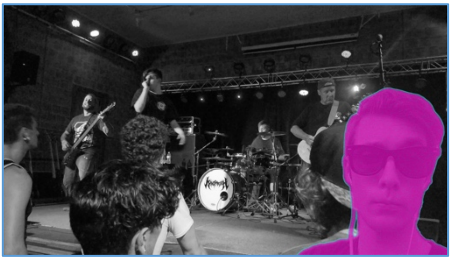
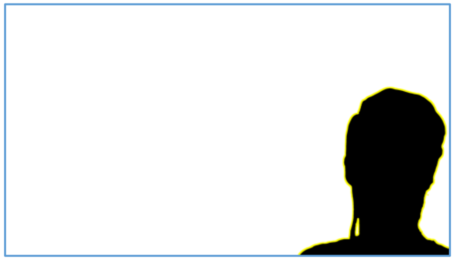
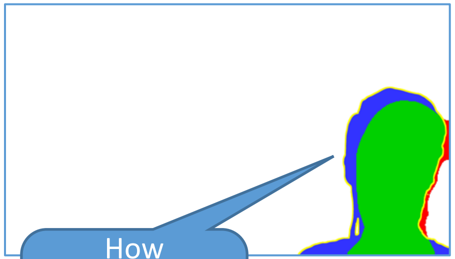
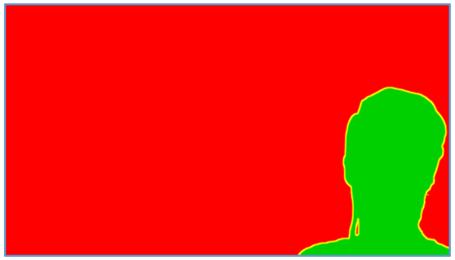
# NC17 Image Manipulation Localization Results (11 teams, 16 systems)

| Team | System | All-MCC | tr-MCC | Trial Response Rate |
|---|---|---|---|---|
| BINGHAMTON | p-prnu_1 | | 0.1853 | 0.1000 |
| FIBBER | p-FourIGH_1 | | 0.0365 | 0.9886 |
| MAYACHITRA-Cl | c-acontrario_3 | | 0.0345 | 0.9945 |
| MAYACHITRA-Mc | c-resamplingdetector1_3 | 0.0202 | | |
| MAYACHITRA-UcR | c-lstmwithoutresampling_2 | | 0.0035 | 0.9975 |
| Purdue-11b1 | p-MFCN1_1 | | 0.0596 | 0.9980 |
| SRI-TA2 | p-baseline_1 | 0.0887 | | |
| SRIPRI | p-baseline_1 | | 0.0831 | 0.1870 |
| UMD | p-facesteganalysis_1 | | 0.1876 | 0.1054 |
| UNIFI | c-baselineMOD3_1 | | 0.2241 | 0.0686 |
| | c-baselineMOD4_1 | | 0.2237 | 0.0681 |
| USCISI | c-Autoencoder01a_1 | | 0.1893 | 0.9727 |
| | c-PMcopymove01a_1 | | 0.1317 | 0.9995 |
| | c-PMinpainting01a_1 | | 0.1209 | 0.9995 |
| | c-gradbased01a_1 | | 0.1957 | 0.2337 |
| | p-Splicebuster01a_1 | | 0.1991 | 0.9727 |

# NC17 Image Manipulation Localization - All operation example (1)

| | • Black – Manipulation<br>• Yellow - No-Score | • Green - True Positives<br>• Red - False Alarm.<br>• White - True Negative<br>• Blue - False Negative | |
|---|---|---|---|
| Composite | Binarized Reference | SystemID1 | SystemID2 |
| 21a1b6501b9c0d84fa46ad6eddf8bbe4 | | MCC: 0.87 | MCC: 0.57 |
|  |  |  |  |
| fb8785800546e9602ef35c7ee0cee8b7 | | MCC: 0.84 | MCC: 0 |
|  |  |  Invariant to size |  |

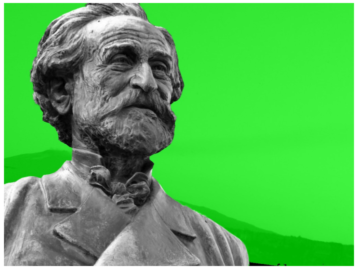
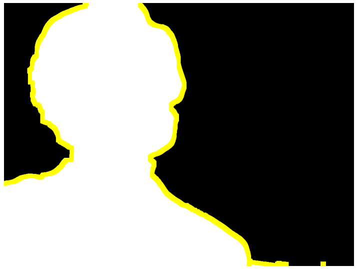
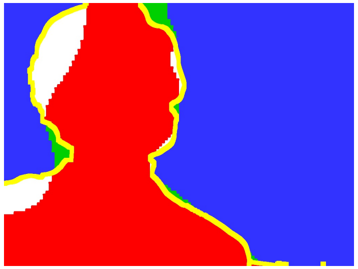
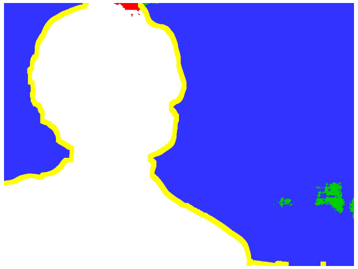
# NC17 Image Manipulation Localization
# - All operation example (2)

| | • Black – Manipulation<br>• Yellow - No-Score | • Green - True Positives<br>• Red - False Alarm.<br>• White - True Negative<br>• Blue - False Negative | |
|---|---|---|---|
| Composite | Binarized | SystemID1 | SystemID2 |
| ebaaa9df1cbbb21a2bfdc99da637fd1b | | MCC: 0.94 | MCC: 0.0079 |
|  |  |  |  |
| 5130e1013704be24a7f1c7ac8d3d67c9 | | MCC: 0.85 | MCC: 0.0 |
|  |  |  |  |

MCC is not “object-centric”

How “important” is the spatial mismatch?

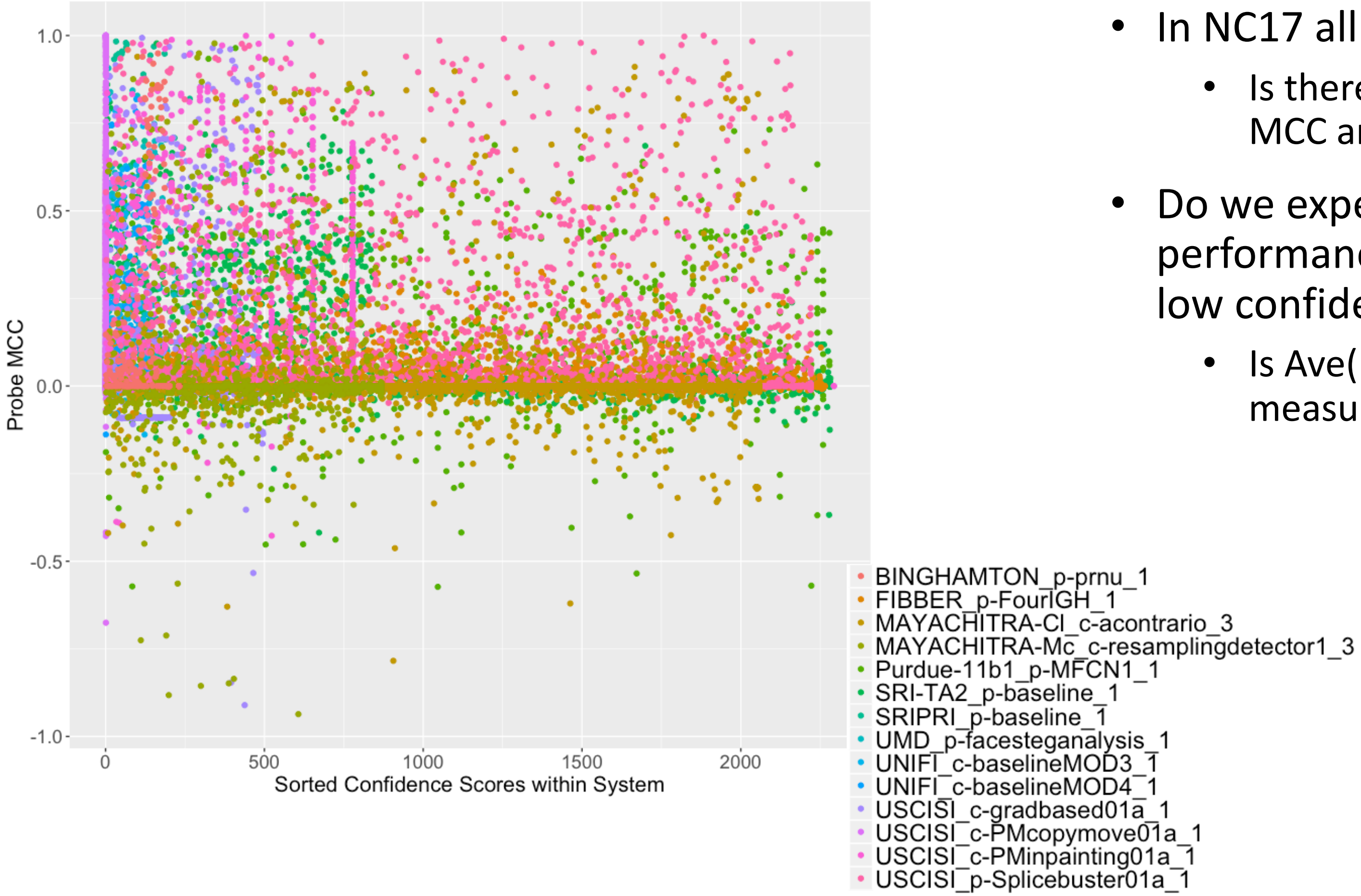
# NC17 Image Manipulation Localization
# - All operation example (3)

| | - Black – Manipulation<br>- Yellow - No-Score | - Green - True Positives<br>- Red - False Alarm.<br>- White - True Negative<br>- Blue - False Negative | |
|---|---|---|---|
| Composite | Binarized | SystemID1 | SystemID2 |
| **903ced75f1755d6819a857da2a475121** | | MCC: -0.882 | MCC: 0.046 |
|  |  |  |  |

<u>Description</u>

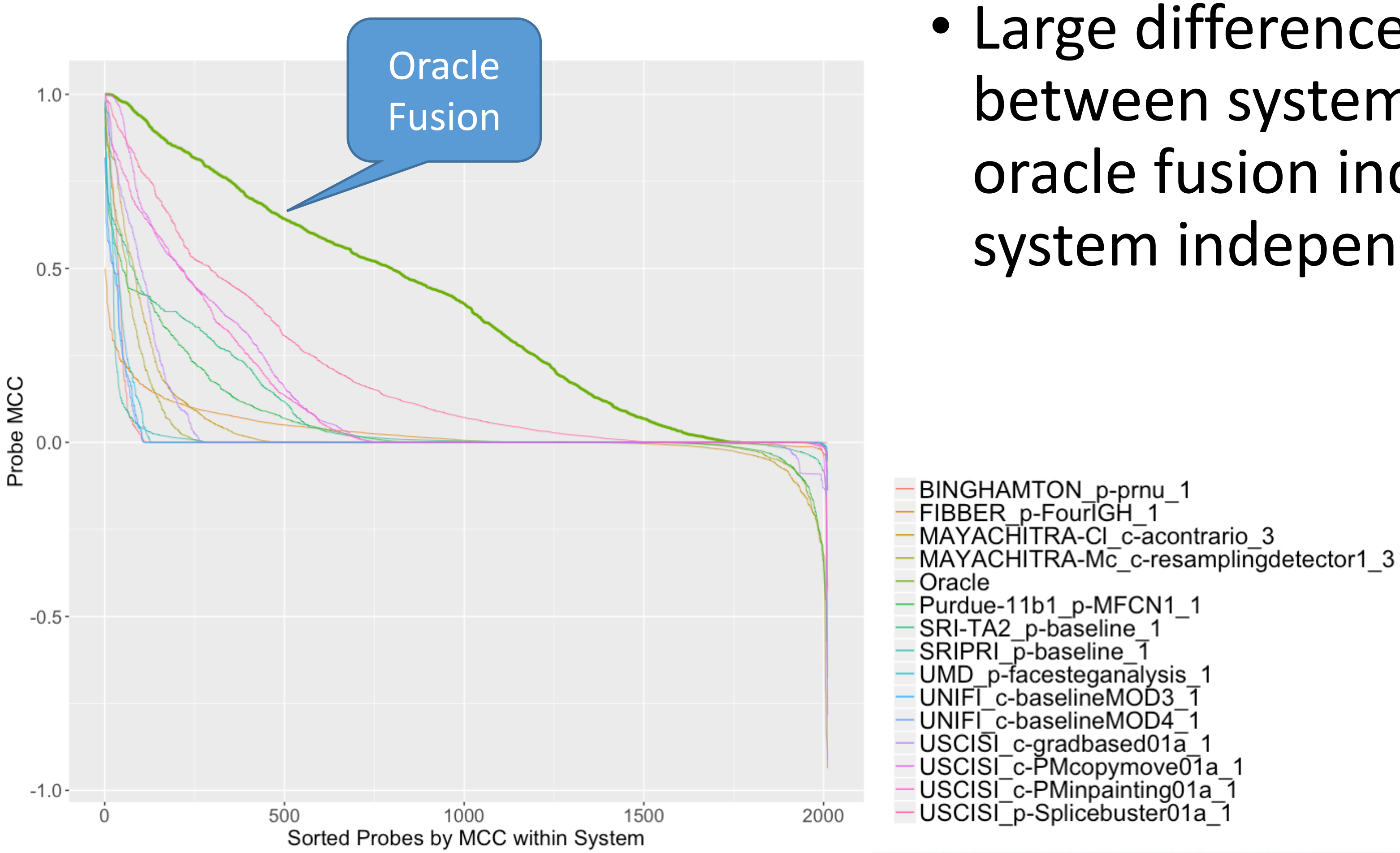
"spliced landscape of mountain at dusk into background of statue"

# Understanding Localization Performance MCC vs. Rank(Confidence)



- In NC17 all targets scored
  - Is there a correlation between MCC and detection rank?
- Do we expect localization performance to be needed for low confidence images?
  - Is Ave(max MCC) the right measure?

# Understanding Localization Performance
## MCC Score Distribution Across Manipulations



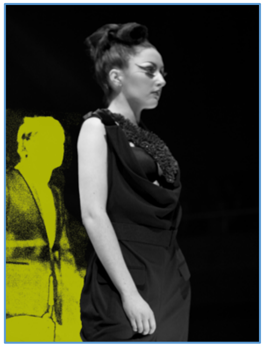
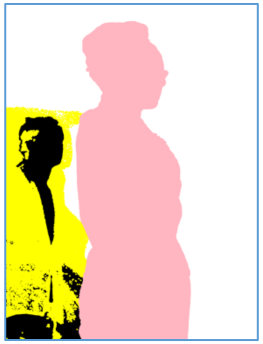
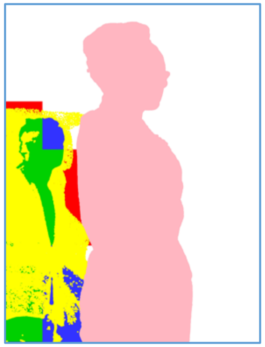
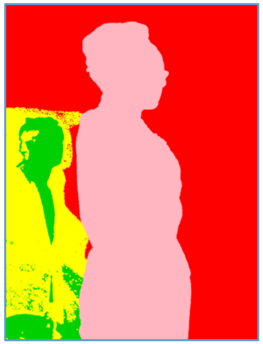
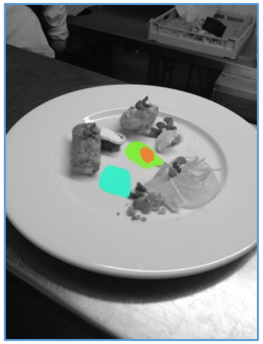
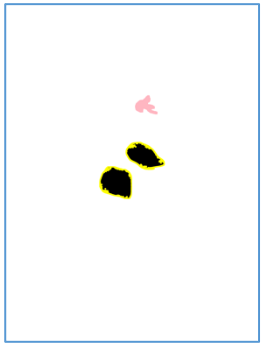
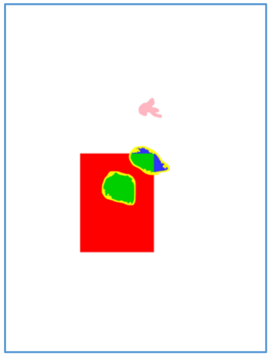
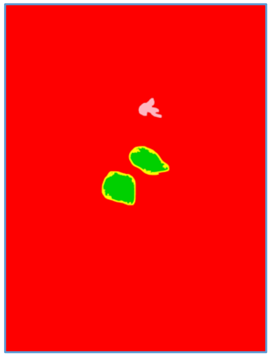
Oracle Fusion

- Large difference between systems and oracle fusion indicates system independence

# NC17 Image Manipulation Localization - Selective Scoring Example (1)

- Green - True Positives
- Red - False Alarm.
- White - True Negative
- Blue - False Negative
- Yellow - Boundary No-Score
- Pink - Selective No-Score

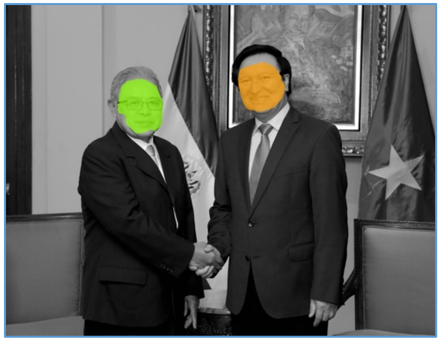
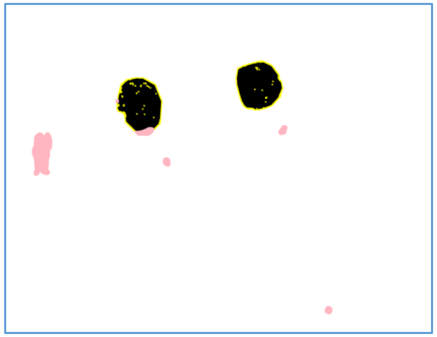
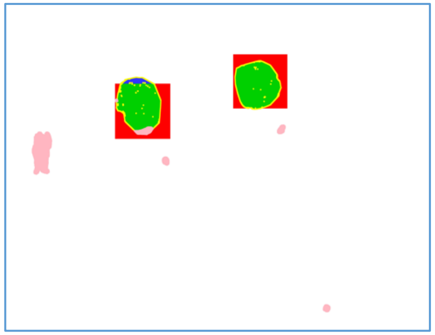
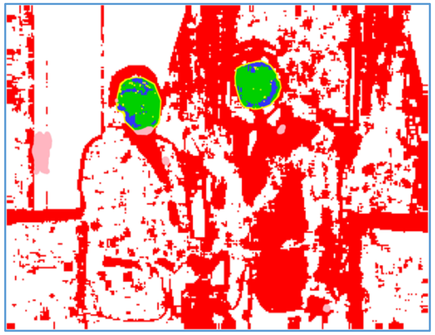
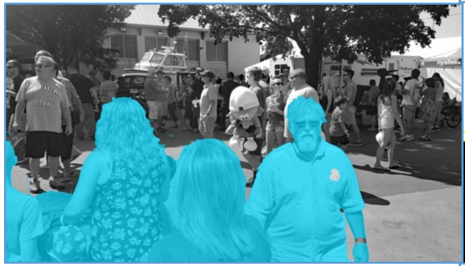
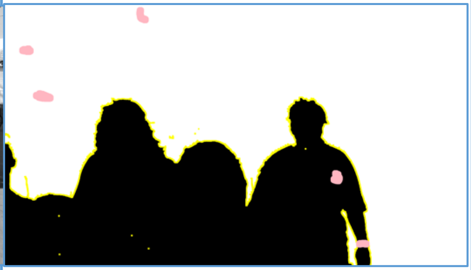
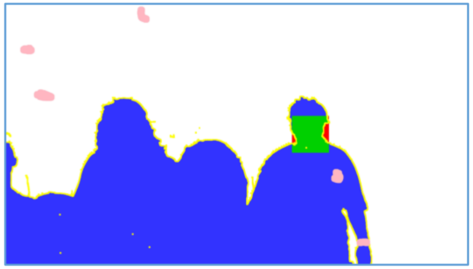
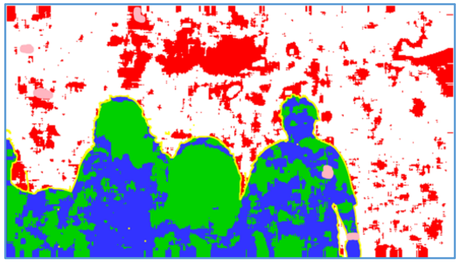
- Content Aware Remove: -qm "Operation==['FillContentAwareFill']"

| Composite | Binarized | SystemID1 | SystemID2 |
|---|---|---|---|
| 3ab10f081a83602cef7d4907e35c94b7 | | MCC: 0.662457150011 | MCC: 0.0 |
|  |  |  |  |
| ad8b348b5d4e82f261633bc979ac98eb | | MCC: 0.334995460552 | MCC: 0.0 |
|  |  |  |  |

# NC17 Image Manipulation Localization - Selective Scoring Example (1)

- Face: -qm "OperationArgument==['face']”
- 11 probes: need more data; (OptIn: partial)

- Green - True Positives
- Red - False Alarm.
- White - True Negative
- Blue - False Negative
- Yellow - Boundary No-Score
- Pink - Selective No-Score

| Composite | Binarized | SystemID1 | SystemID2 |
|---|---|---|---|
| 728a0b1ba50a962b21fafb7ef372bf75 | | MCC: 0.74087844111 | MCC: 0.120222463722 |
|  |  |  |  |
| 52d568ca915ac608d12c0bbafbea3bb8 | | MCC: 0.128545171054 | MCC: 0.305342054323 |
|  |  |  |  |

# Outline

✓NIST NC2017 Evaluation Overview

✓Detection
  - ✓Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ✓Splice – paired: Task, Data, Results
  - ✓Video: Data, Result

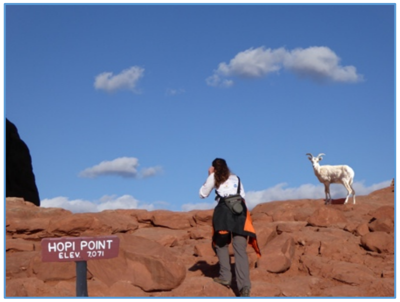
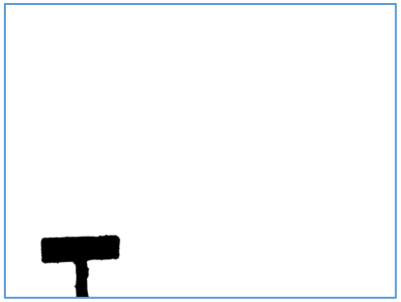
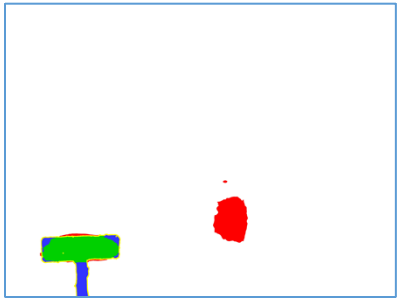
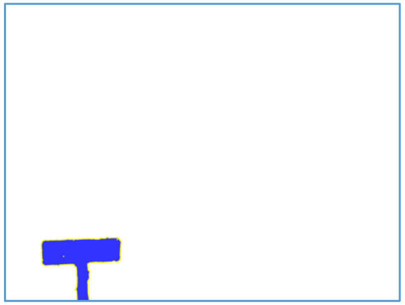
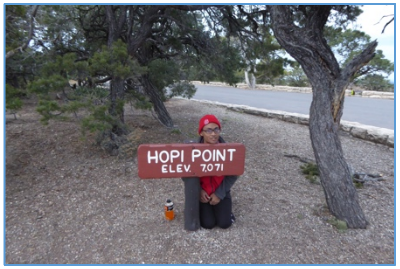
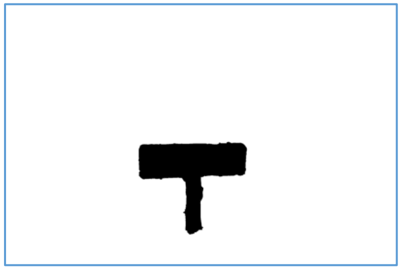
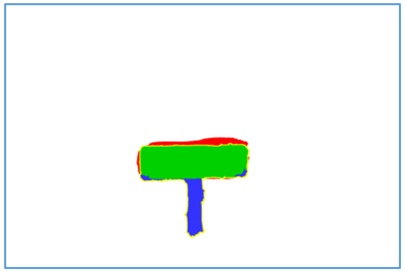
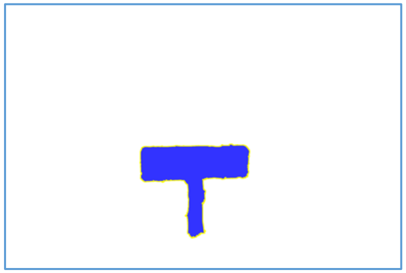
- Localization
  - ✓Image – single: Metrics, Results, Analysis
  - ➢Splice – paired: Results
- Provenance
  - Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary

# NC17 Splice Manipulation Eval. Results
## - Localization MCC , Image Only

| | | Donor TRR | Donor trMCC | Probe TRR | Probe trMCC |
|---|---|---|---|---|---|
| UNIFI | c-baselineMOD4_1 | 0.0907 | 0.1010 | 0.0910 | 0.1940 |
| | p-baselineMOD3_1 | 0.0918 | 0.0998 | 0.0921 | 0.1916 |
| USCISI | p-baseline_1 | 1.000 | 0.1862 | 1.0000 | 0.1740 |

OptIn

# NC17 Splice Manipulation Localization - Example (1)

| | Input images | Binarized Mask | SystemID1 | SystemID2 |
|---|---|---|---|---|
| Probe | | | MCC: 0.636 | MCC: 0.0 |
| |  |  |  |  |
| Donor | | | MCC: 0.816 | |
| |  |  |  |  |

# Provenance

- Provenance Filtering
- Provenance Graph Building

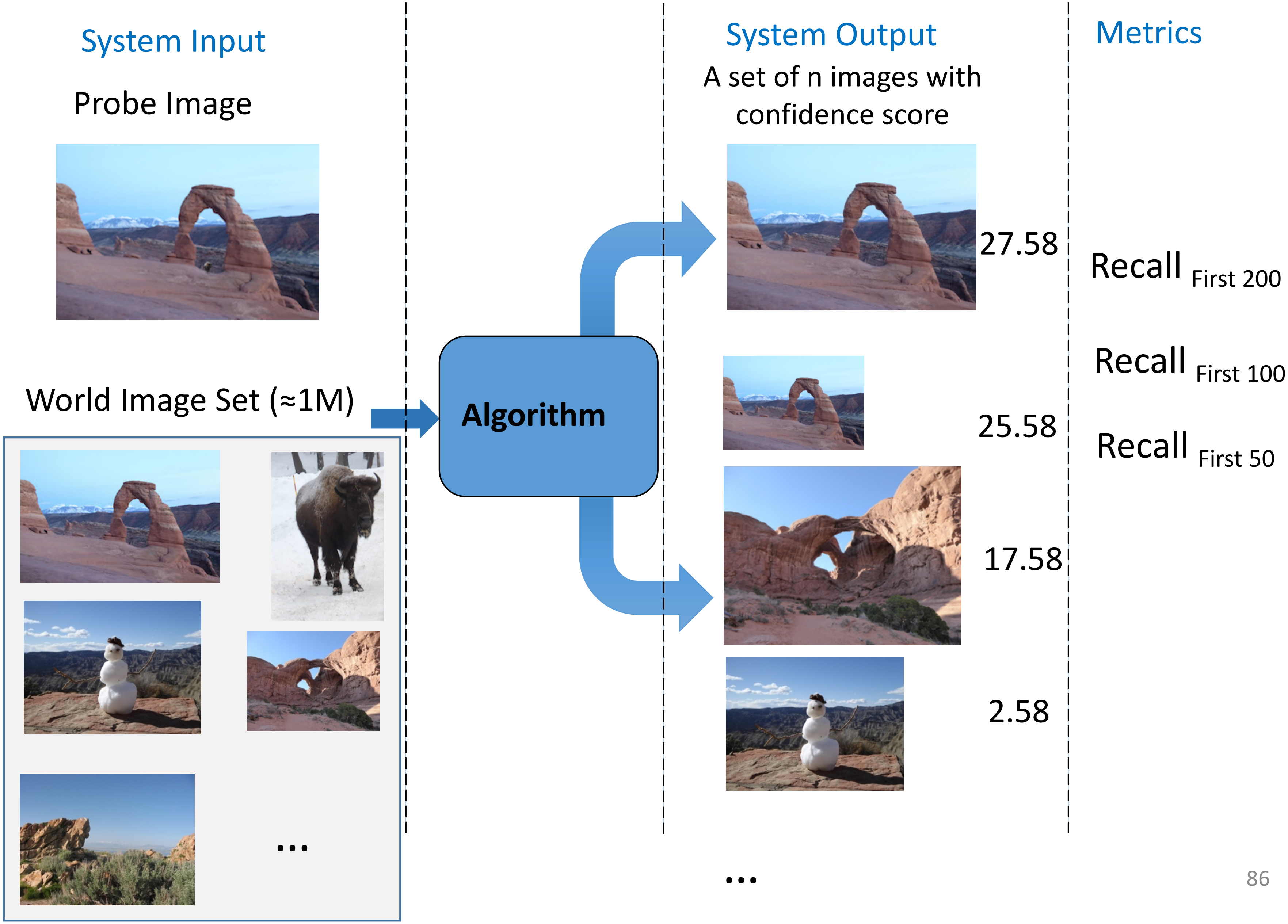
# Outline

- ✓NIST NC2017 Evaluation Overview
- ✓Detection
  - ✓Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ✓Splice – paired: Task, Data, Results
  - ✓Video: Data, Result
- ✓Localization
  - ✓Image – single: Metrics, Results, Analysis
  - ✓Splice – paired: Results
- Provenance
  - ➢Filtering: Task, Data, Metrics, Result, Analysis
  - Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

# Provenance Filtering Task

- Task description
  - Given a probe image, return all images (its ancestors and descendants in the world dataset) in its genealogy graph.
- Task inputs
  - A probe image
  - A world dataset
- System input conditions:
  - Image Only (no header or metadata)
  - Image + Metadata
- Task outputs
  - For each probe, a set of $n$ images as potential candidates with their confidence scores.

# Provenance Filtering Evaluation Model



System Input

Probe Image

World Image Set (≈1M)

...

Algorithm

System Output

A set of n images with confidence score

27.58

25.58

17.58

2.58

...

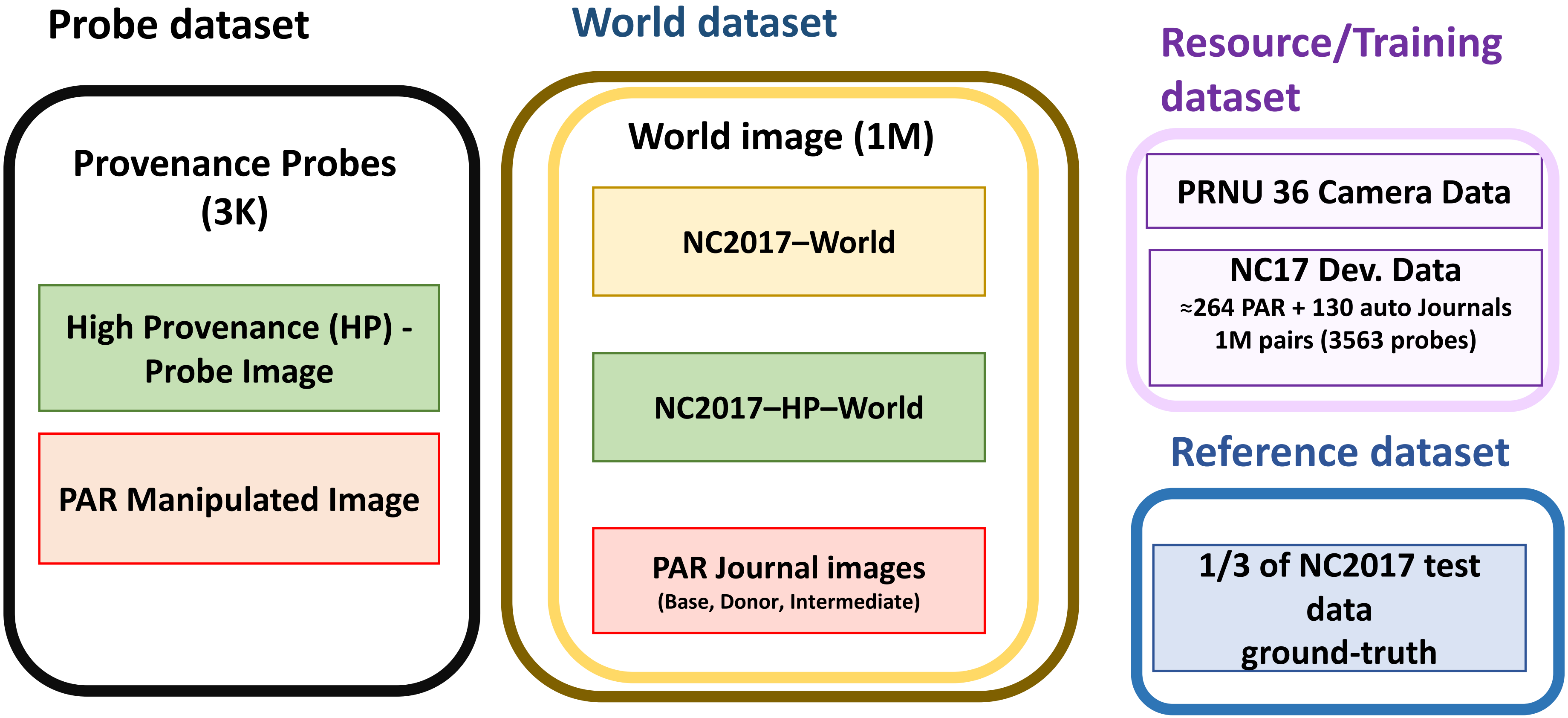
Metrics

Recall$_{\text{First 200}}$

Recall$_{\text{First 100}}$

Recall$_{\text{First 50}}$

# NC2017 Provenance Datasets



## Probe dataset

Provenance Probes (3K)

High Provenance (HP) - Probe Image

PAR Manipulated Image

## World dataset

World image (1M)

NC2017–World

NC2017–HP–World

PAR Journal images
(Base, Donor, Intermediate)

## Resource/Training dataset

PRNU 36 Camera Data

NC17 Dev. Data
≈264 PAR + 130 auto Journals
1M pairs (3563 probes)

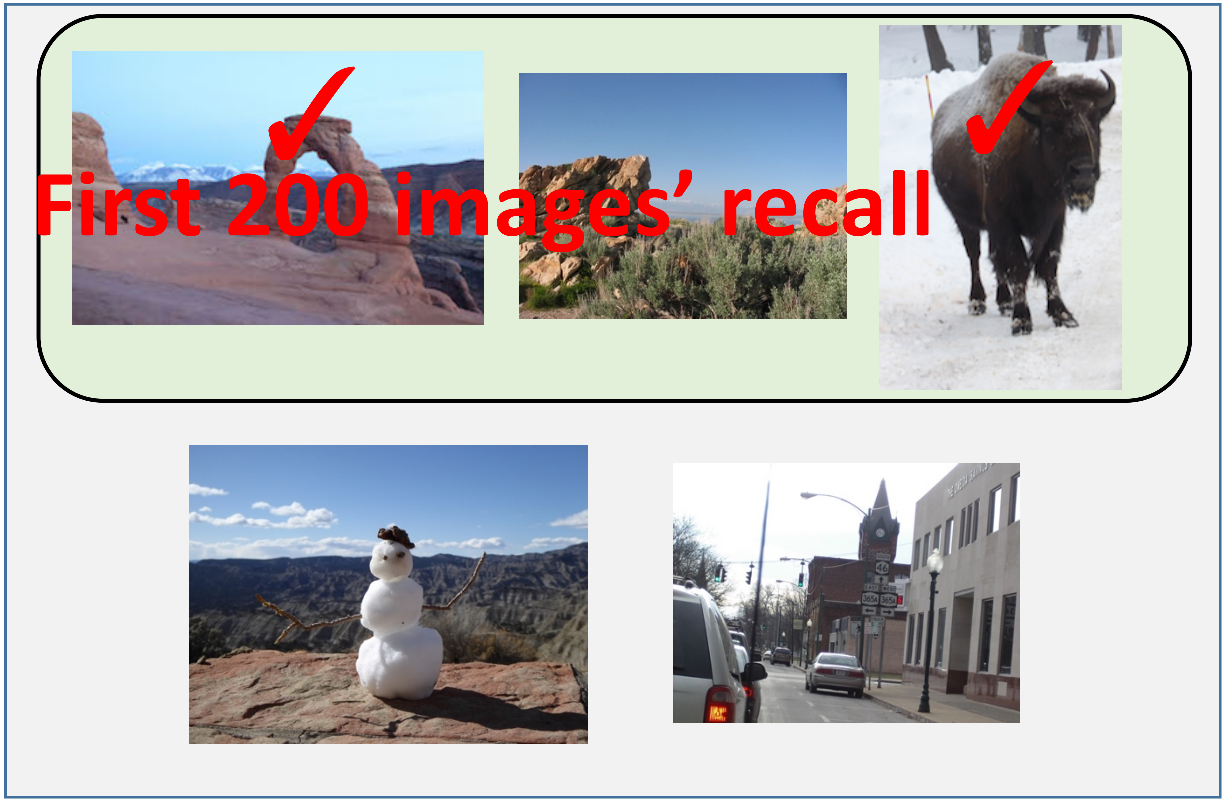
## Reference dataset

1/3 of NC2017 test data ground-truth

# Provenance Filtering Example

Probe Image

NC2017 Evaluation World Set (≈1M)



First 200 images' recall

# Provenance Filtering Metrics

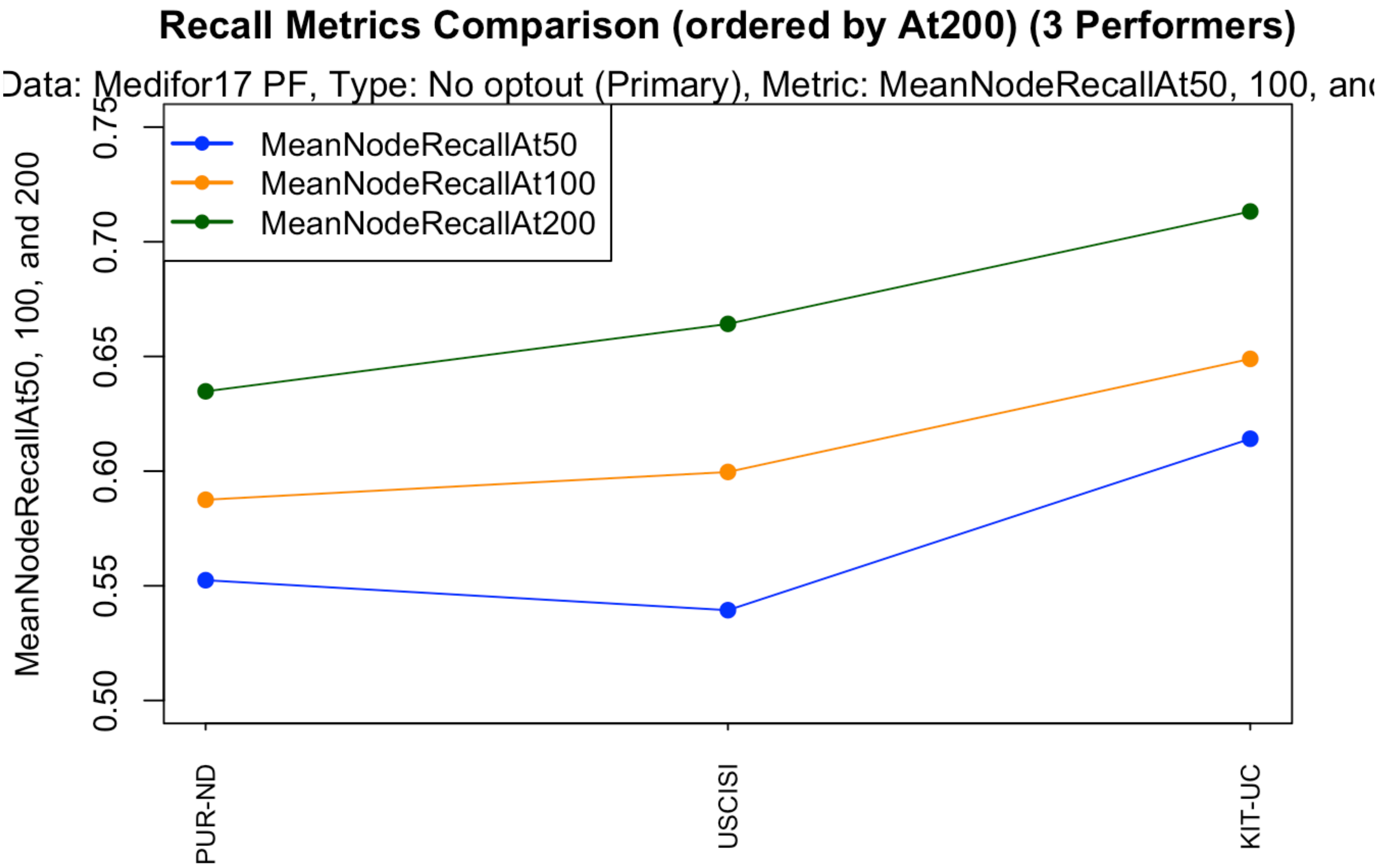
$$recall = \frac{|\{relevant\} \cap \{retrieved\}|}{|\{relevant\}|}$$

- The recall of first $n$ images from the world dataset (≈1M) sorted by ‘confidence score’
- Evaluated only true manipulated probes whose contributors are in the world data set
- Variations:
  - The depth of retrieval will be varied, e.g., recall@50, recall@100, recall@200

# NC2017 Provenance Filtering Results

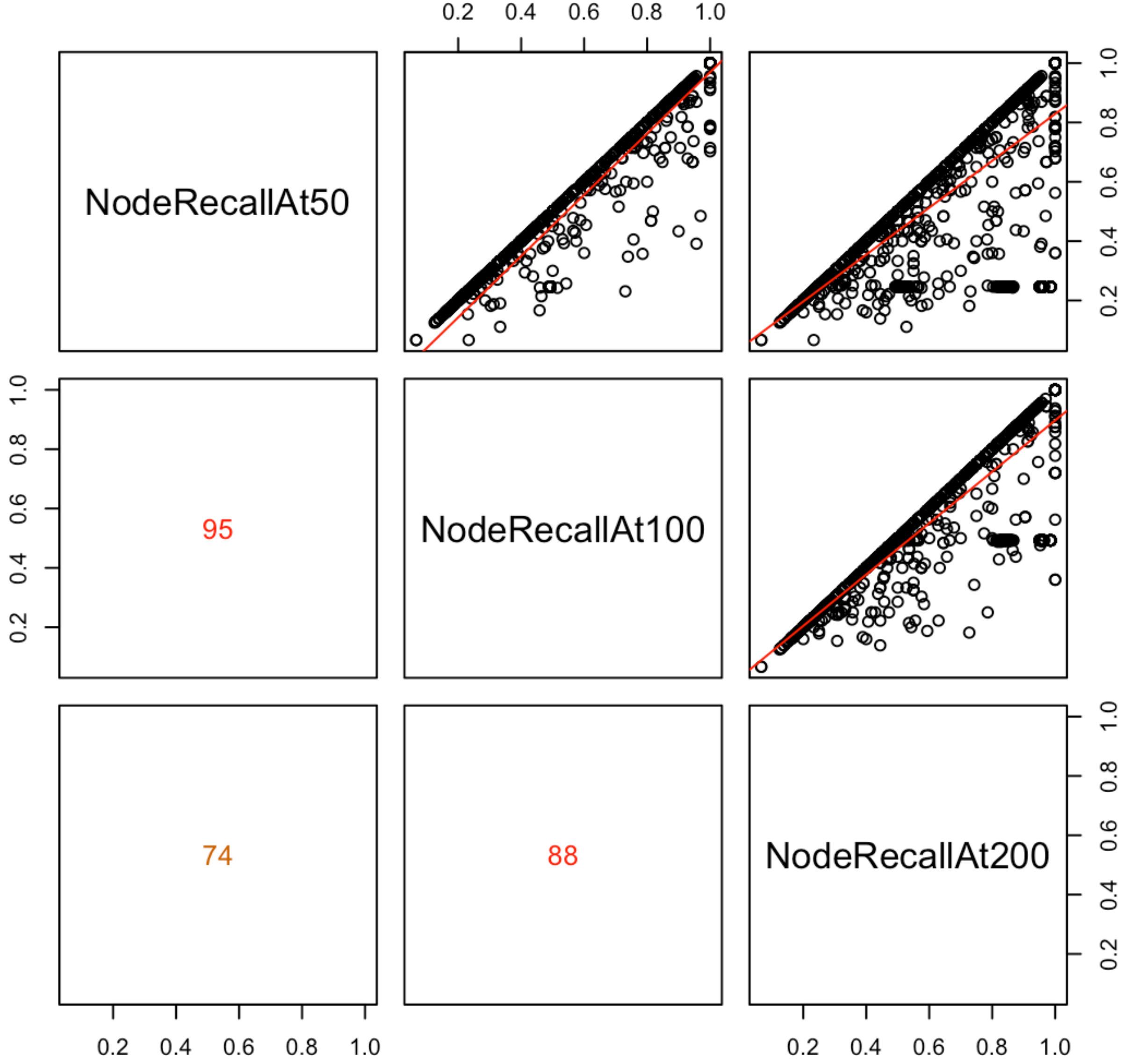
- 3 teams/organizations, 7 systems

| | | | Recall@050 | Recall@100 | Recall@200 |
|---|---|---|---|---|---|
| NDPURDUE | p | p-baseline_1 | 0.5524 | 0.5875 | 0.6348 |
| USCISI | p | p-baseline_1 | 0.5394 | 0.5996 | 0.6642 |
| kitware-ucolumbia | p | p-baseline_1 | 0.6141 | 0.6489 | 0.7132 |


Recall Metrics Comparison (ordered by At200) (3 Performers)
Data: Medifor17 PF, Type: No optout (Primary), Metric: MeanNodeRecallAt50, 100, anc
MeanNodeRecallAt50
MeanNodeRecallAt100
MeanNodeRecallAt200
MeanNodeRecallAt50, 100, and 200
0.50
0.55
0.60
0.65
0.70
0.75
PUR-ND
USCISI
KIT-UC

# Correlation of Node Recall Metrics

## Primary Systems, All Provenance Probes



# Outline

- ✓NIST NC2017 Evaluation Overview
- ✓Detection
  - ✓Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ✓Splice – paired: Task, Data, Results
  - ✓Video: Data, Result
- ✓Localization
  - ✓Image – single: Metrics, Results, Analysis
  - ✓Splice – paired: Results
- Provenance
  - ✓Filtering: Task, Data, Metrics, Result, Analysis
  - ➢Graph Building: Task, Data, Metrics, Results, Analysis
- Summary and Future Opportunities

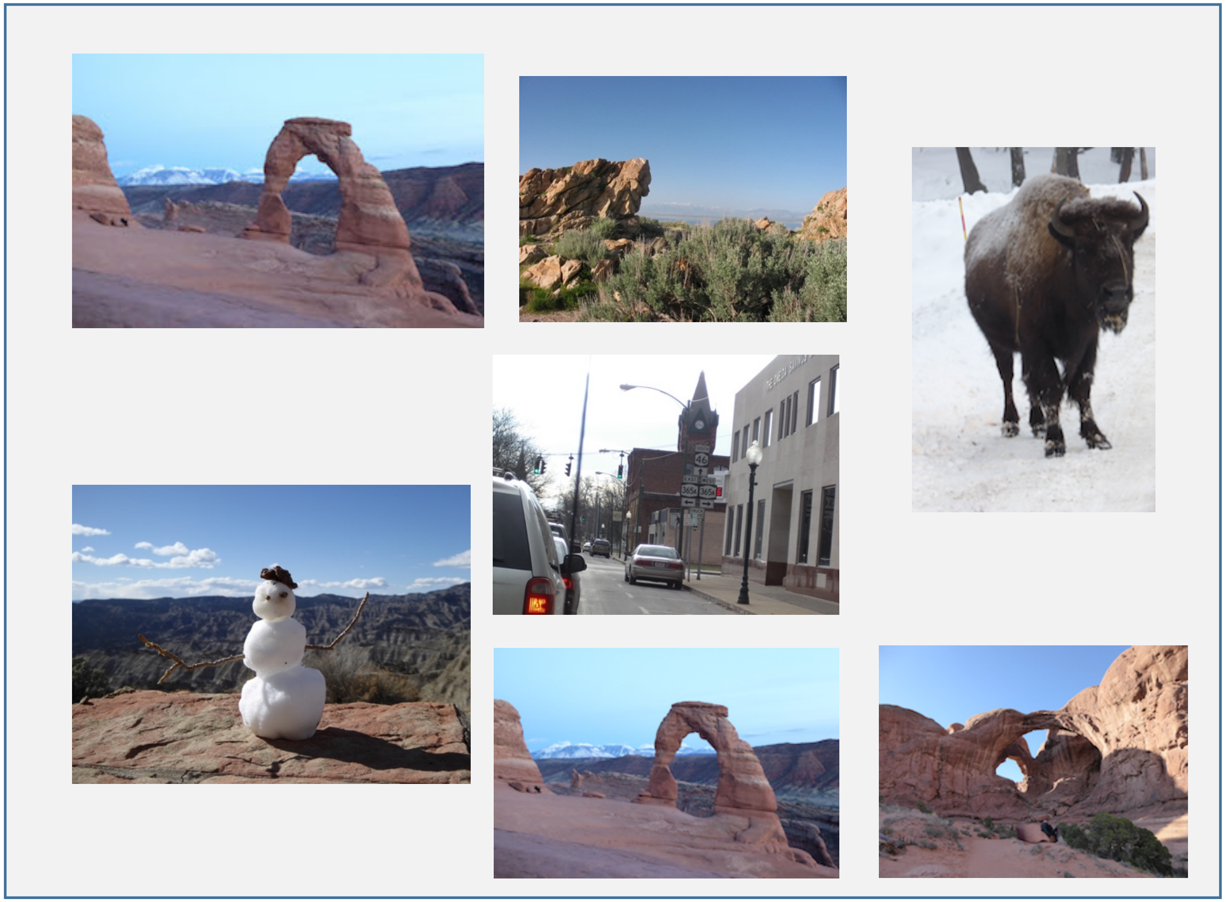
# Provenance Graph Building

- Task Description
  - Given a probe image, construct and label the manipulation provenance graph that includes all its ancestors and descendants in the world dataset.
- Task Inputs
  - End-to-End Provenance:  a probe image, a large world set (1M images)
  - Oracle Provenance: a probe image, a small world set (≈200 images, all contributor images with some distractor world images)
- System input conditions:
  - Image Only
  - Image + Metadata
- Task outputs:
  - a provenance graph

# Provenance Graph Building System Input

Probe Image

NC2017 Evaluation World Set / Oracle Set



# Evaluation Options:
## Direct Path Limited vs. Full Graph

- probe: node with circle;
- world: all other nodes in concise graph

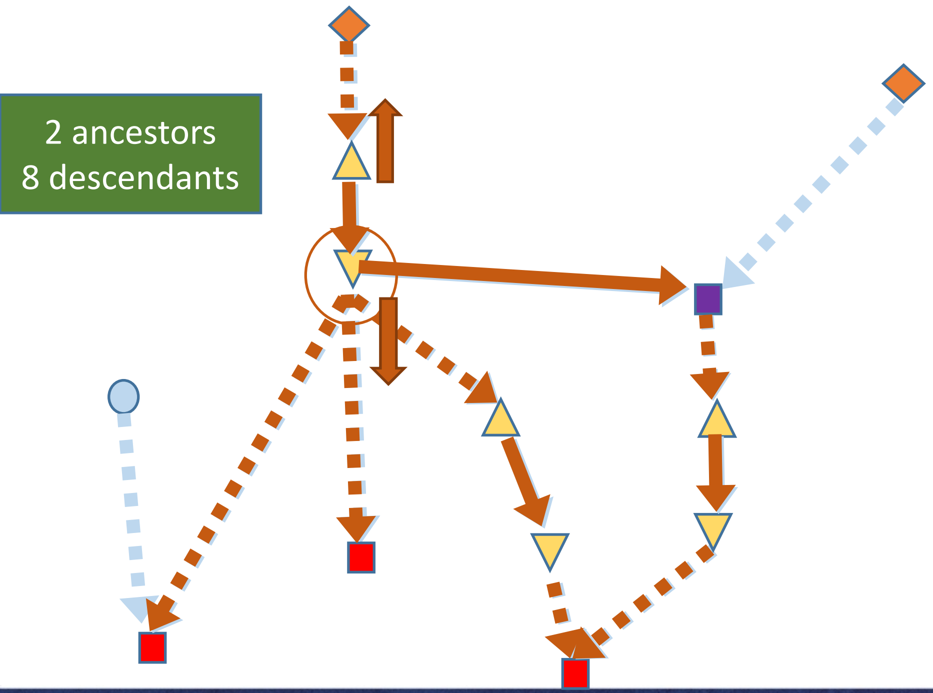
**Direct Path Limited**

All direct ancestors and descendants of given probe

**Full Graph**

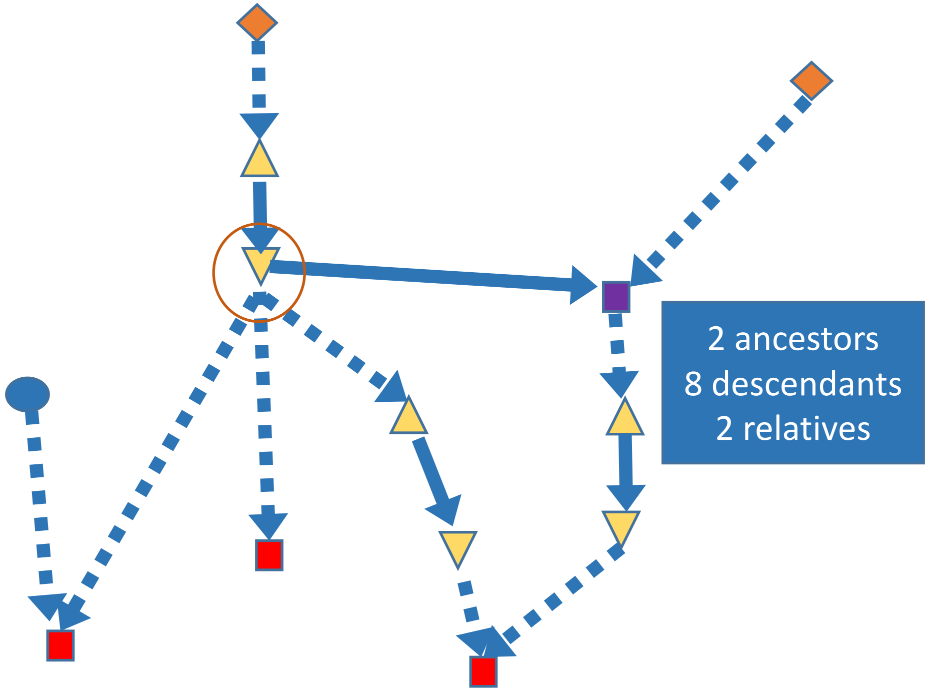
Recursively include all direct paths connected to the probe and all ancestors and descendants of given probe

**Direct Path Reference Graph**



2 ancestors
8 descendants

**Full Reference Graph**



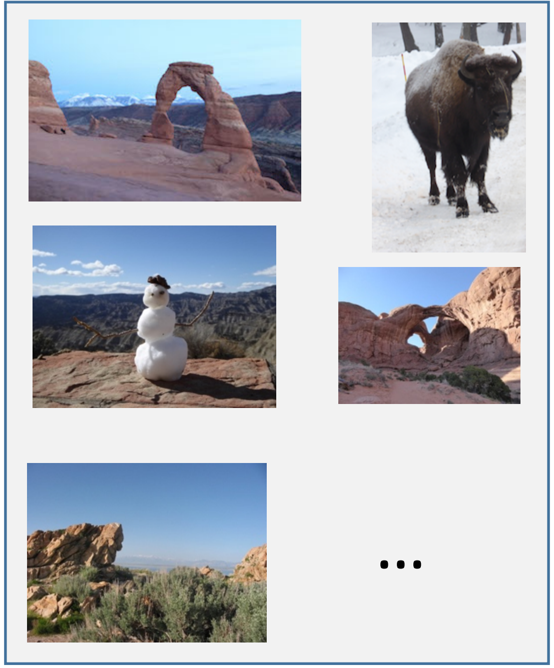
2 ancestors
8 descendants
2 relatives

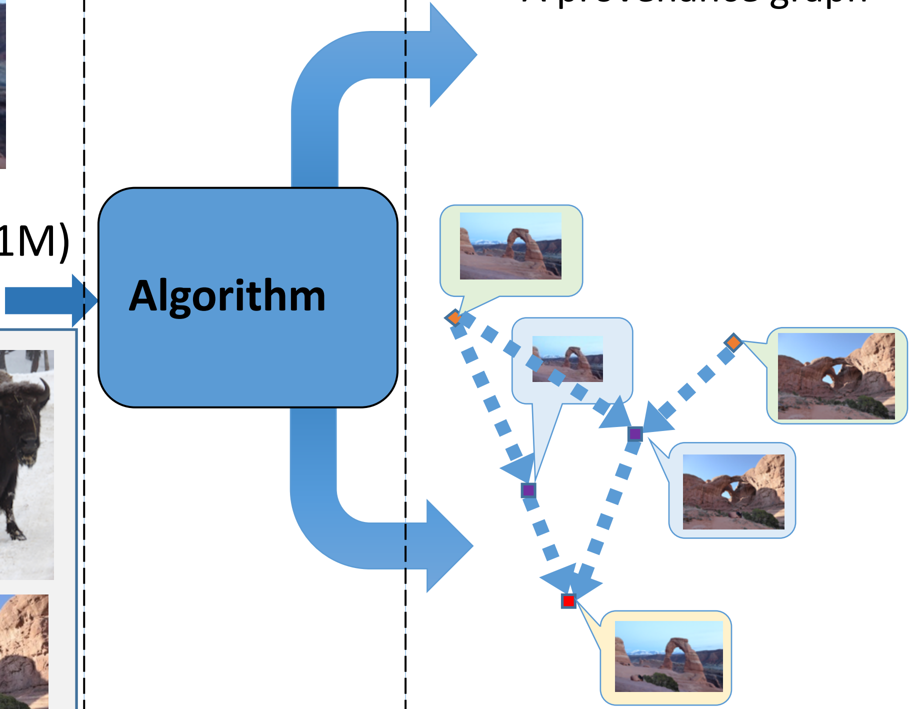
# Provenance Graph Building Evaluation Model

System Input

Probe Image

(1) World Image Set (≈1M)
(2) Oracle Set (≈200)





System Output

A provenance graph

Metrics

Graph Similarity

Generalized F-measure:

- Sim(nodes)
- Sim (links)
- Sim(nodes+links)

# Provenance Graph Building Task Evaluation Metrics

- Graph Similarity and Generalized F-measure
  - Overlap of nodes: $\mathrm{sim}_{\mathrm{NO}}(G_r, G_s) = 2\frac{|V_r \cap V_s|}{|V_r| + |V_s|}$
  - Overlap of links: $\mathrm{sim}_{\mathrm{LO}}(G_r, G_s) = 2\frac{|E_r \cap E_s|}{|E_r| + |E_s|}$
  - Overlap of node and links: $\mathrm{sim}_{\mathrm{NLO}}(G_r, G_s) = 2\frac{|V_r \cap V_s| + |E_r \cap E_s|}{|V_r| + |V_s| + |E_r| + |E_s|}$

| MeanNodeRecall | From Provenance Filtering |
|---|---|
| MeanSimNO | Similarity of Node Overlap for a Provenance Graph - Eval Plan Section 7.0 |
| MeanSimLO | Similarity of Link Overlap for a Provenance Graph - Eval Plan Section 7.0 |
| MeanSimNLO | Similarity of Link+Node Overlap for a Provenance Graph - Eval Plan Section 7.0 |

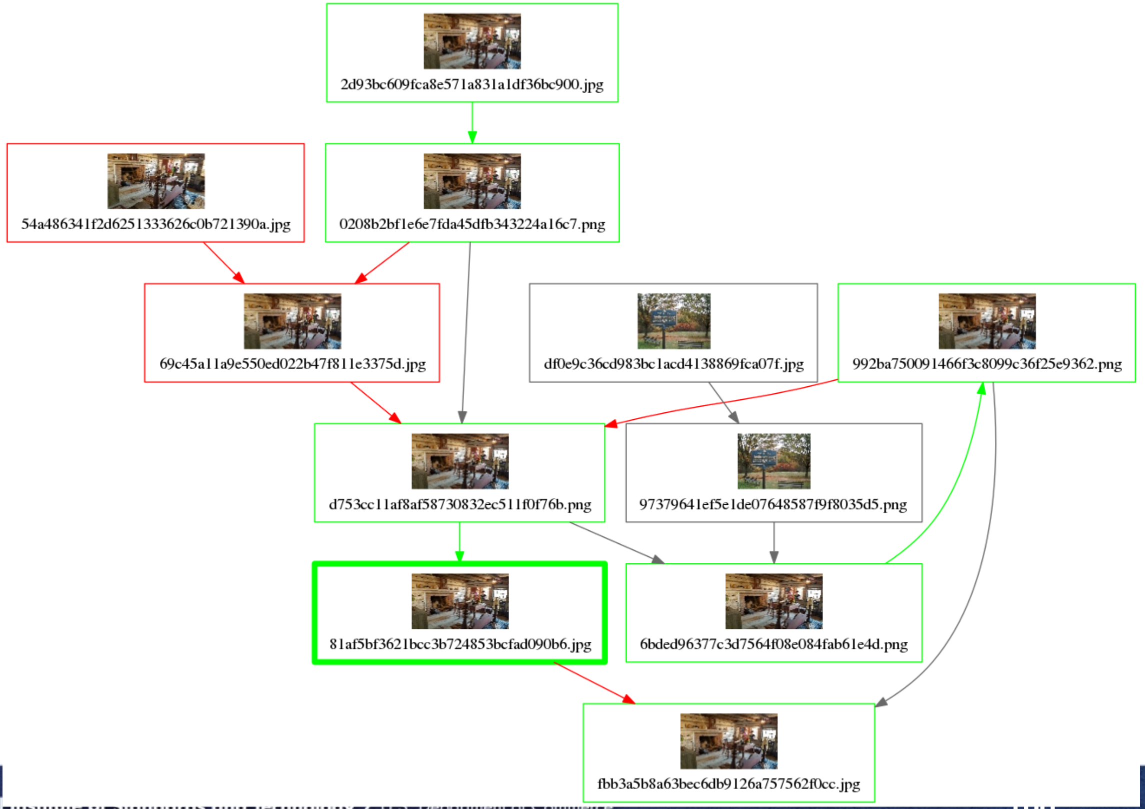
# NC2017 Provenance Graph Building Eval. Results

- 2 teams/organizations, 5 systems (end-to-end)

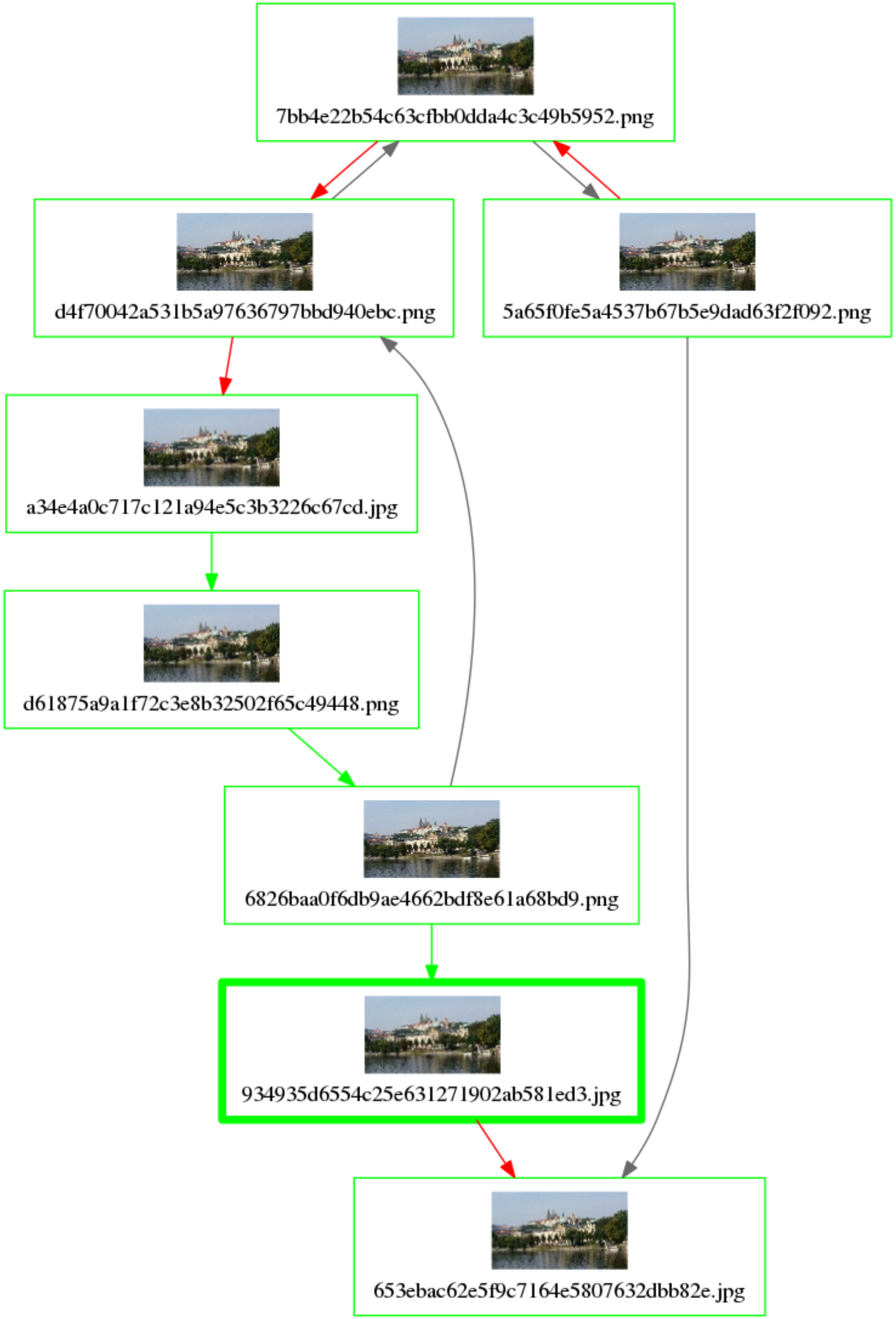
| | | | MeanNodeRecall | MeanSimNO | MeanSimLO | MeanSimNLO |
|---|---|---|---|---|---|---|
| NDPURDUE | c | c-contrast1_1 | 0.5249 | 0.5913 | 0.1812 | 0.3875 |
| | | c-contrast2_1 | 0.5228 | 0.6124 | 0.2189 | 0.4170 |
| | | c-contrast3_1 | 0.5246 | 0.5909 | 0.1809 | 0.3872 |
| | p | p-baseline_1 | 0.5230 | 0.6127 | 0.2085 | 0.4124 |
| USCISI | p | p-baseline_1 | 0.4786 | 0.4146 | 0.0776 | 0.2674 |

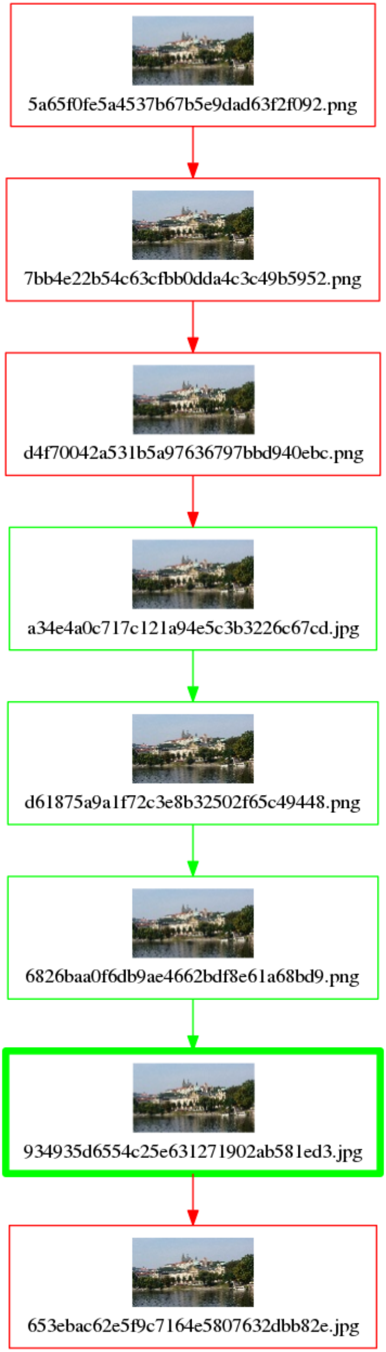
# Graph Evaluation: An Example

- Green image border - Correctly included image.
- **Wide Green** image border - The Probe image.
- Red image border - False alarm image.
- Grey image border - Omitted provenance image (missed detection).
- Green link - Correctly linked images.
- Red link - False alarm link.
- Grey link - Omitted link.



2d93bc609fca8e571a831a1df36bc900.jpg
54a486341f2d6251333626c0b721390a.jpg
0208b2bf1e6e7fda45dfb343224a16c7.png
69c45a11a9e550ed022b47f811e3375d.jpg
df0e9c36cd983bc1acd4138869fca07f.jpg
992ba750091466f3c8099c36f25e9362.png
d753cc11af8af58730832ec511f0f76b.png
97379641ef5e1de07648587f9f8035d5.png
81af5bf3621bcc3b724853bcfad090b6.jpg
6bded96377c3d7564f08e084fab61e4d.png
fbb3a5b8a63bec6db9126a757562f0cc.jpg


7bb4e22b54c63cfbb0dda4c3c49b5952.png
d4f70042a531b5a97636797bbd940ebc.png
5a65f0fe5a4537b67b5e9dad63f2f092.png
a34e4a0c717c121a94e5c3b3226c67cd.jpg
d61875a9a1f72c3e8b32502f65c49448.png
6826baa0f6db9ae4662bdf8e61a68bd9.png
934935d6554c25e631271902ab581ed3.jpg
653ebac62e5f9c7164e5807632dbb82e.jpg


5a65f0fe5a4537b67b5e9dad63f2f092.png
7bb4e22b54c63cfbb0dda4c3c49b5952.png
d4f70042a531b5a97636797bbd940ebc.png
a34e4a0c717c121a94e5c3b3226c67cd.jpg
d61875a9a1f72c3e8b32502f65c49448.png
6826baa0f6db9ae4662bdf8e61a68bd9.png
934935d6554c25e631271902ab581ed3.jpg
653ebac62e5f9c7164e5807632dbb82e.jpg

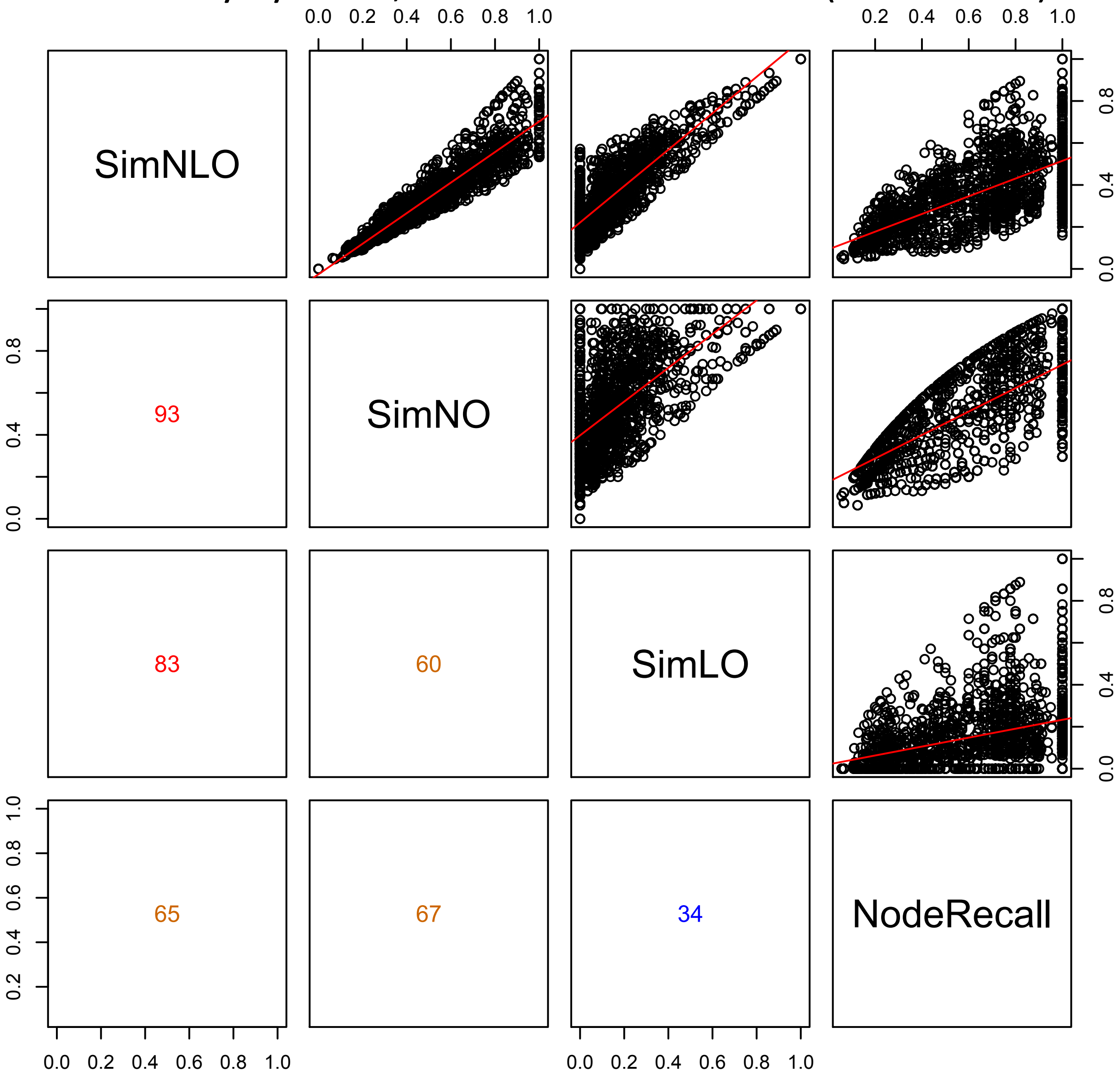
# NC2017 Provenance Graph Building Eval. Results

- 2 teams/organizations, 5 systems (end-to-end)

| | | | MeanNodeRecall | MeanSimNO | MeanSimLO | MeanSimNLO |
|---|---|---|---|---|---|---|
| NDPURDUE | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] | [redacted] |
| | p | p-baseline_1 | 0.5230 | 0.6127 | 0.2085 | 0.4124 |
| USCISI | p | p-baseline_1 | 0.4786 | 0.4146 | 0.0776 | 0.2674 |

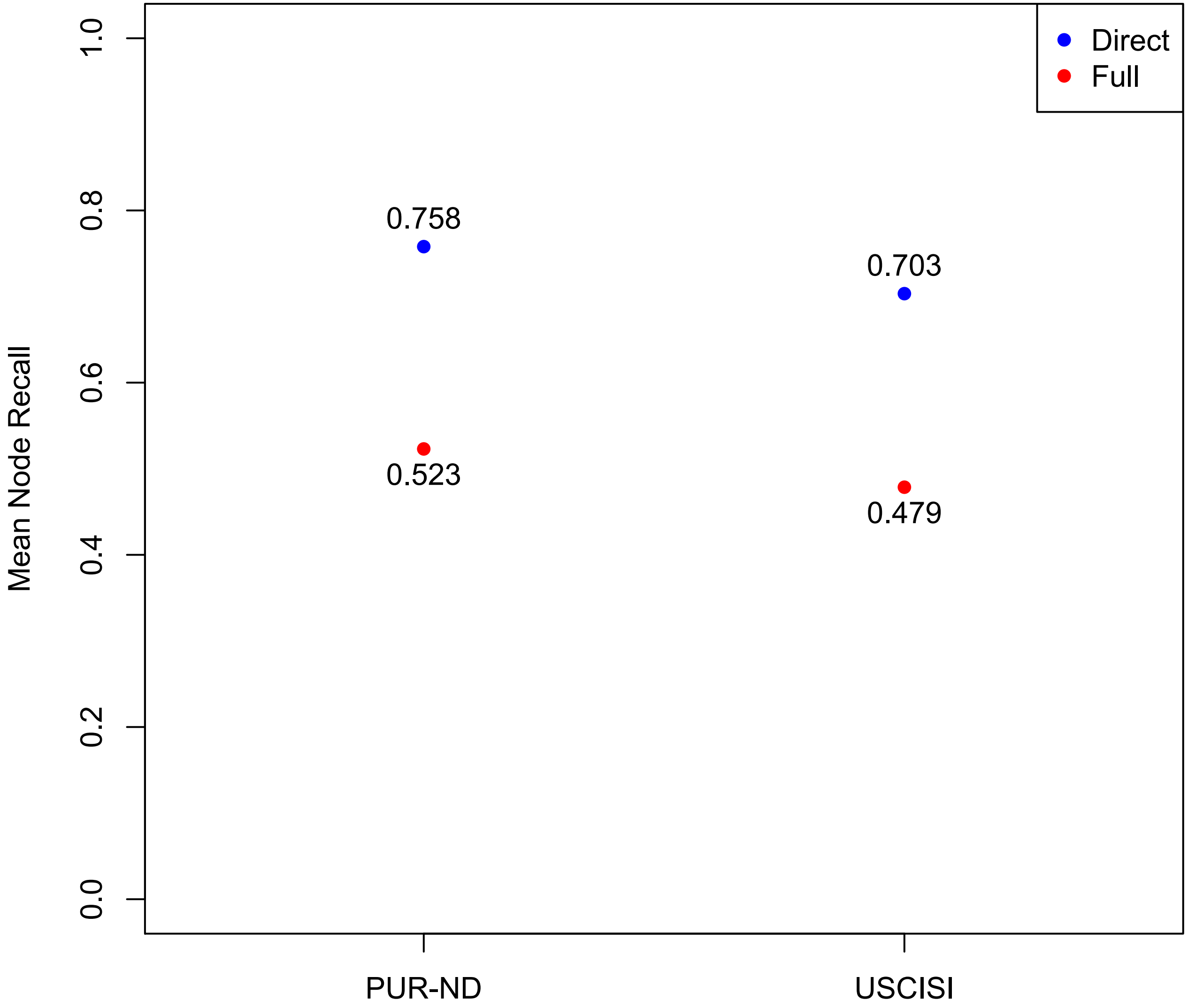
# Correlation of Graph Building Metrics

## Primary Systems, All Provenance Probes (End-to-End)



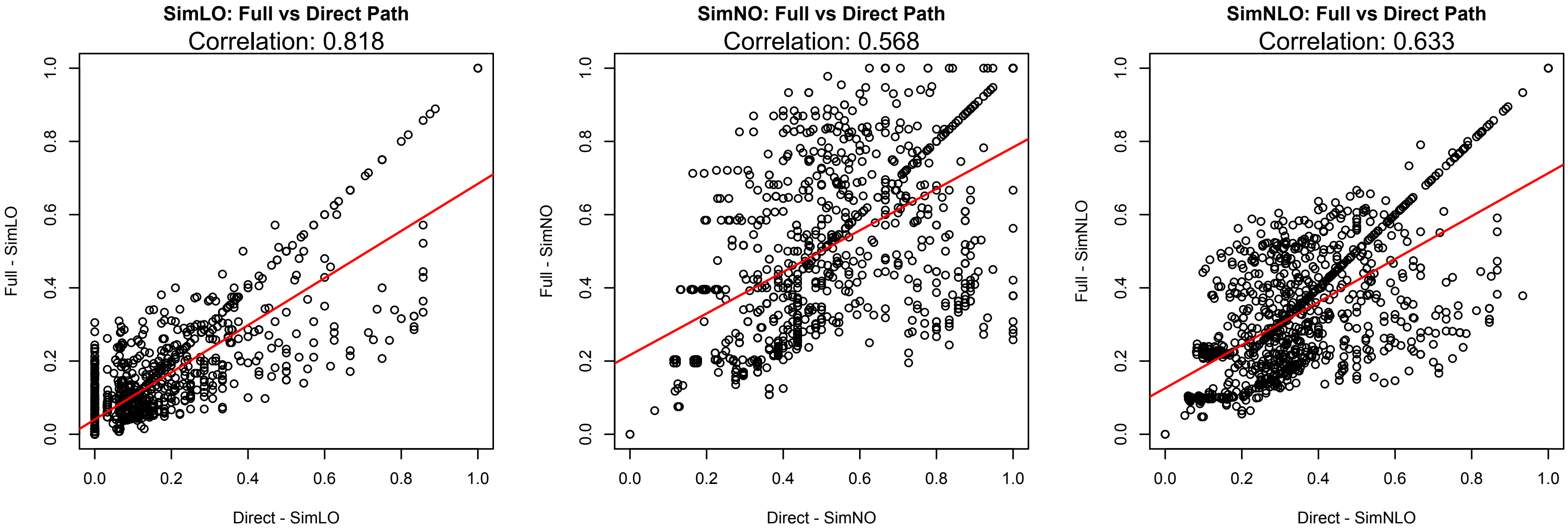
# Mean Node Recall over Graph Conditions

Primary Systems, All Provenance Probes (End-to-End)



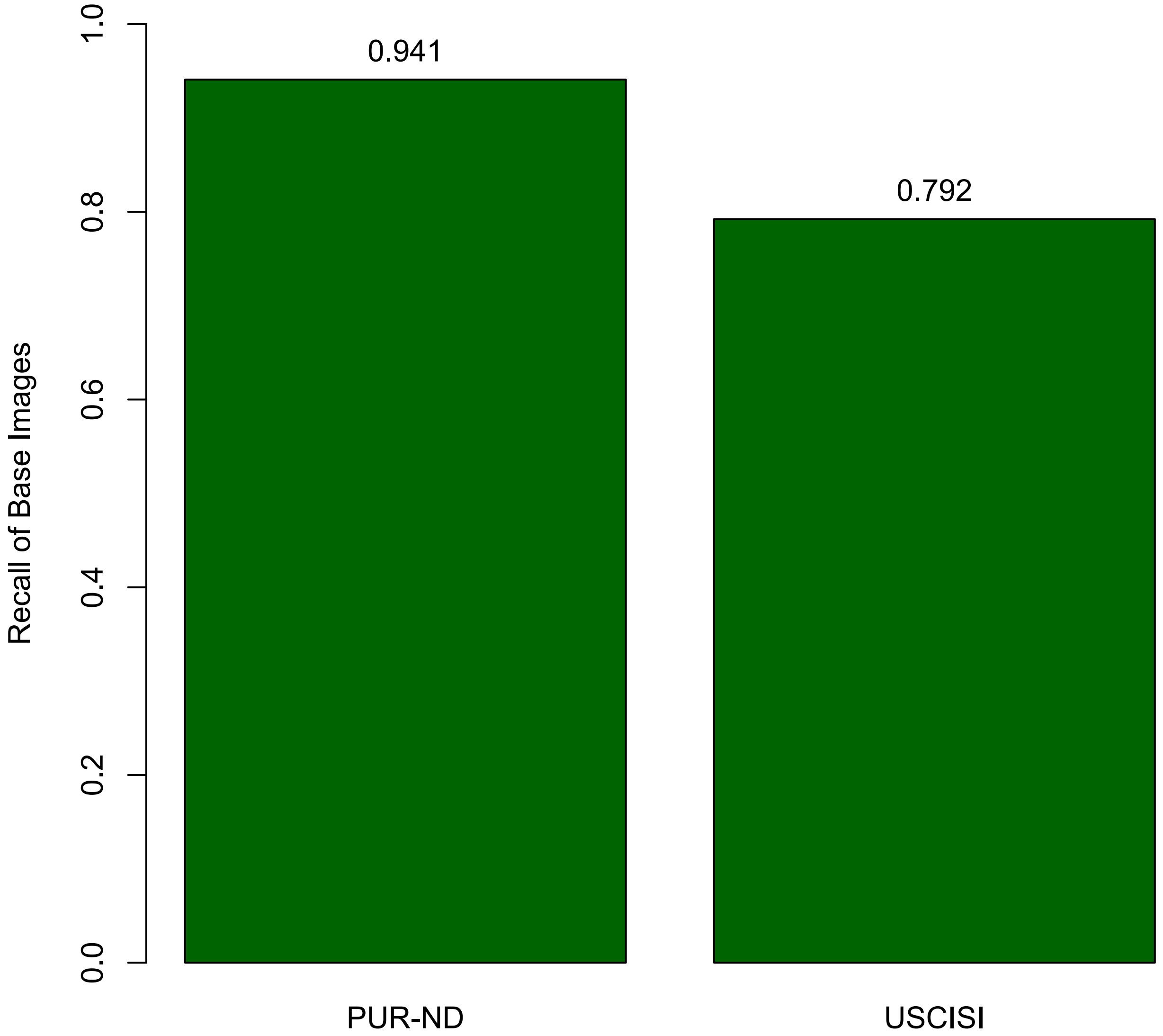
# Correlation of Graph Building Metrics: Full vs. Direct Graph Condition

## Primary Systems, All Provenance Probes (End-to-End)



# Recall of Base Images

## Primary Systems, All Provenance Probes (End-to-End)



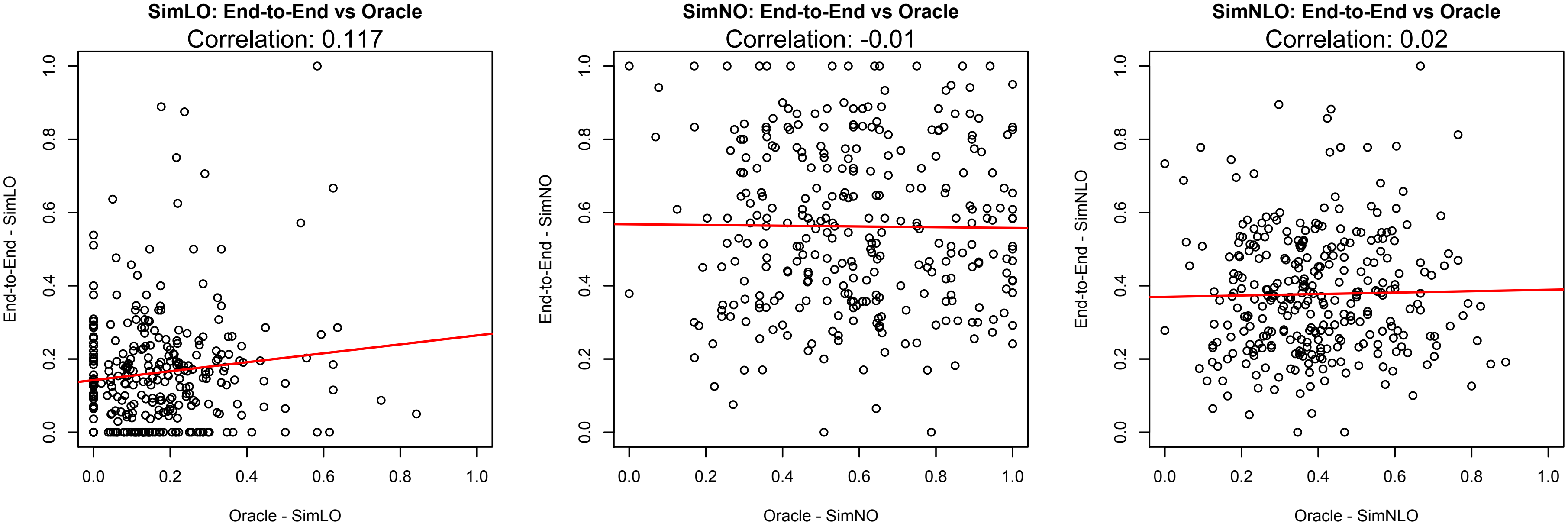
# NC2017 Provenance Graph Building Eval. Results

- 2 teams/organizations, 9 systems (oracle – Part1PAR)

| | | | MeanNodeRecall | MeanSimNO | MeanSimLO | MeanSimNLO |
|---|---|---|---|---|---|---|
| NDPURDUE | p | p-baseline_1 | 0.5919 | 0.6596 | 0.2393 | 0.4530 |
| | | p-oracle_2 | 0.7405 | 0.7172 | 0.2525 | 0.4889 |
| USCISI | p | p-baselineOracle_1 | 0.5349 | 0.4645 | 0.0923 | 0.2998 |
| | | p-baseline_1 | 0.5349 | 0.4644 | 0.0896 | 0.3016 |

# Correlation of Graph Building Metrics: End-to-End vs. Oracle

## Primary Systems, Part 1 PAR Provenance Probes



# Future Work

- How well were donors found?
- How well were other final images found?
- Provenance probes being intermediate or base/donor images

# Outline

- ✓NIST NC2017 Evaluation Overview
- ✓Detection
  - ✓Image – single: Task, Data, Metrics, Selective Scoring, Results, Analysis
  - ✓Splice – paired: Task, Data, Results
  - ✓Video: Data, Result
- ✓Localization
  - ✓Image – single: Metrics, Results, Analysis
  - ✓Splice – paired: Results
- ✓Provenance
  - ✓Filtering: Task, Data, Metrics, Result, Analysis
  - ✓Graph Building: Task, Data, Metrics, Results, Analysis
- ➢Summary and Future Opportunities

# Summary of Test and Evaluation Team Accomplishments

- Built the Nimble Challenge Evaluation Data Set
    - Image data - ~10,000 Forensic Probes, 1M Image world data set
    - Video Data - 1000 Forensic Probes
- Conducted the Nimble 17 Baseline evaluation
    - Baseline benchmark performance on four evaluation tasks exploring the space forensic systems
    - Developed the novel 'OptIn' System Evaluation Protocol
- Developed the initial evaluation infrastructure
    - Detailed annotation of manipulation actions for both human and automatic
    - Data set creation
    - Evaluation code
    - Actionable data analysis

# NIST Evaluation Infrastructure Products for Performer Use

- Data creation infrastructure
  - Data selection tools/methods
  - TestMaker translates annotations to evaluation corpora
  - Automatic Journal builder supporting full factorial design via plugins
- New python-based evaluation tools
  - Mask Scorer
  - Detection Scorer
  - Provenance Scorer (both filtering and graph building)

# Potential Infrastructure Enhancements

- Data pipeline/production capabilities
  - TestMaker: use any node as a forensic probe
  - Produce link masks in addition to the colorized masks
  - Identify data gaps - creation to support specific team needs
  - Understanding test set size requirements
  - Data bug tracking
  - Evaluation set version control
- Automatic manipulation tools:
  - Additional plug in/functionalities;
  - Build synthetic journals using manual journals as the base – support variant processing steps for consistent major steps
  - Leverage the TA2 system integration for exhaustive testing
- Scoring tool enhancements
  - Instead of filtering journal content for analysis, use the journals as a metadata source
  - Improve selective localization scoring masks –
    - current colorized masks occlude over-layed operations - dynamic generation of masks
  - Data analysis integrated into the development cycle
- Opt In Localization support
- Support localization scoring by region/object rather than the whole image

# NC2018 Changes

- Evaluation Task Changes
  - Provenance task - Add link type?
- New data resources
  - 2-3 additional development releases
  - Bigger evaluation collection yet same 2-week evaluation period
    - 50,000 image probes, 5,000 video probes, 5 Million world images
- Metric changes
  - Localization – Thresholded MCC vs Grey Scale WL1; AUC for localization,
  - Object/operation/sub-unit/region level scoring
  - Detection metrics focused on low false alarm – Correct Detection @ X False Alarm
  - Direct Path Provenance Filtering scoring
  - Video temporal/spatial localization scoring
- Scoring Server
  - Internal/External teams
  - Leaderboard vs. blind evaluation
  - Developer-controlled selective scoring
  - Statistical system comparisons
  - System output submissions vs. Docker modules
- Association Evaluations

# Thank You for Your Attention!

NIST MediFor Team: medifor-nist@nist.gov

**Disclaimer**

Any mention of commercial products or reference to commercial organizations in this report is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.