Rubin Difference Detectives - Zooniverse project for the ML reliability model for Rubin

Tatiana Acero-Cuellar

Ph.D. Candidate @ University of Delaware

3rd LSST@LATAM

2nd November 2026



H. Stockebrand

VERA C. RUBIN OBSERVATORY

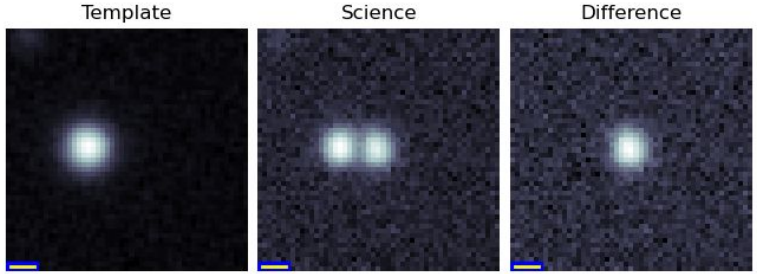
U.S. National Science Foundation

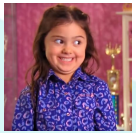
U.S. DEPARTMENT of ENERGY | Office of Science

Difference Image Analysis output: Real and Bogus

Changing objects can be detected as clusters of adjacent pixels deviating from the background.

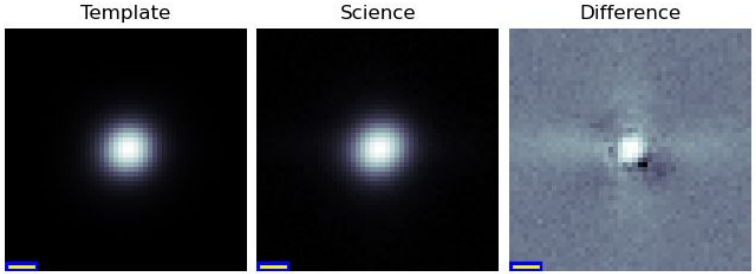
Real detection

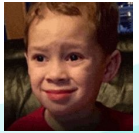




… artifacts are also detected under the same threshold…

Bogus detection





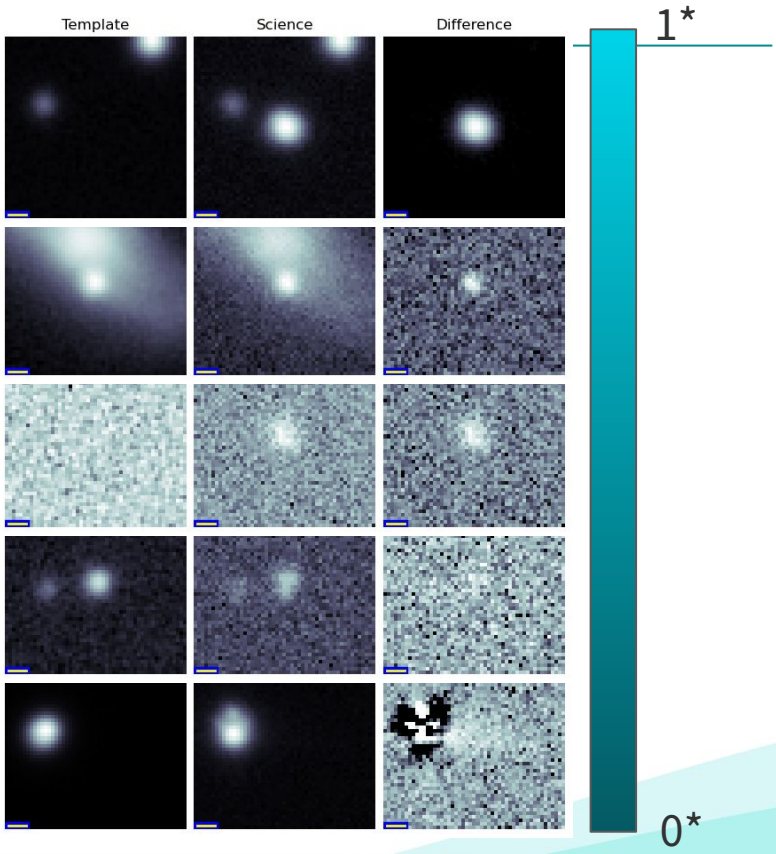
Real/Bogus: Reliability Score

Each source detected on a difference image will have a **reliability** score, values closer to 1 indicates a real object.

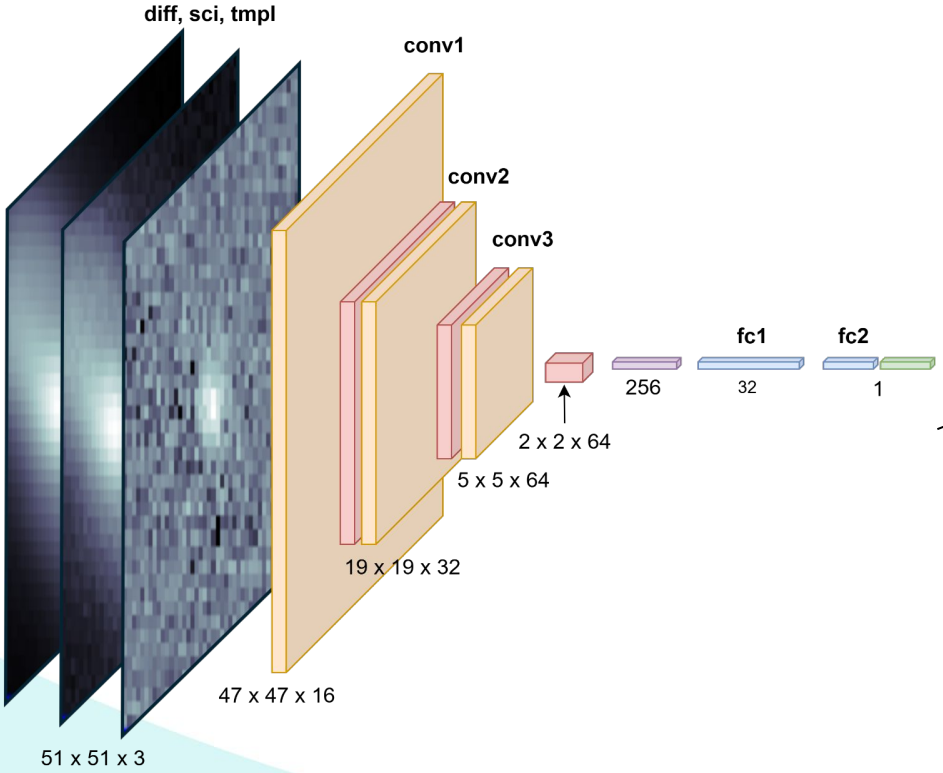
Reliability [as defined in Alert Production Database]:

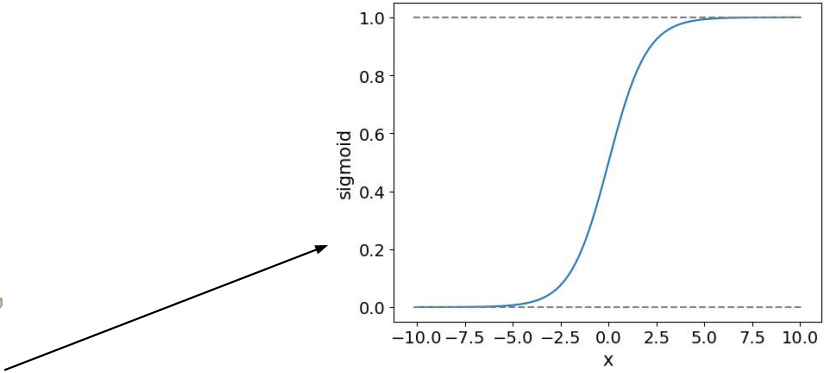
A measure of reliability, computed using information from the source and image characterization, as well as the information on the Telescope and Camera system (e.g., ghost maps, defect maps, etc.).

*estimated values



The Real/Bogus-Reliability Score model

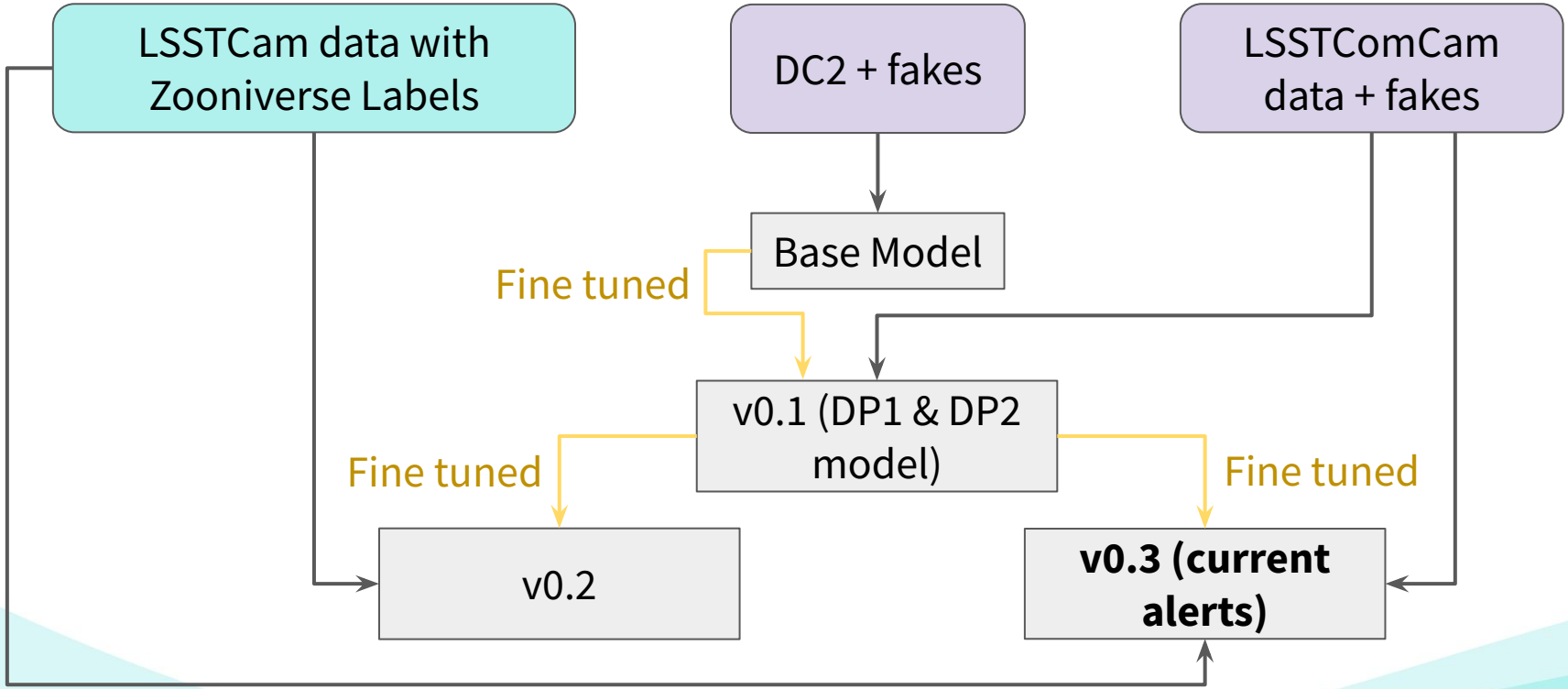




convolutional + ReLU
max pooling
fully connected + ReLU
flatten
sigmoid

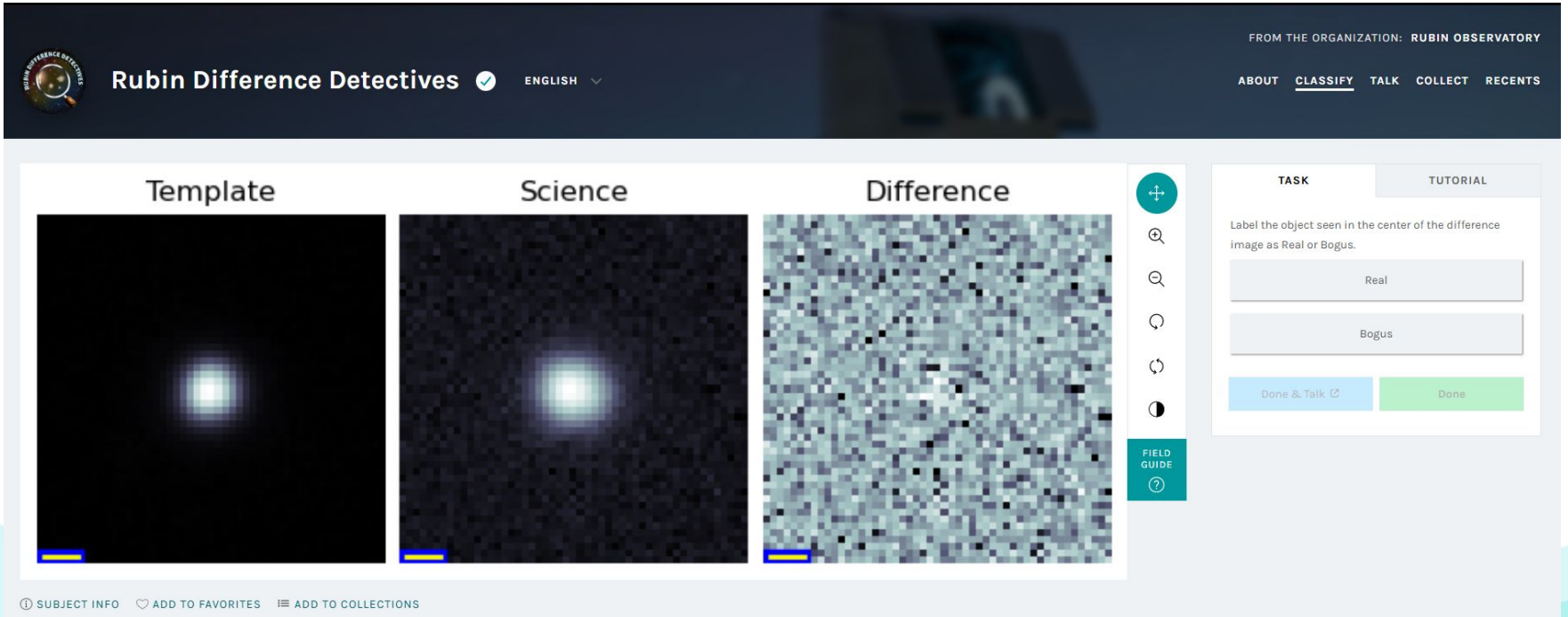
Output can be interpreted as the probability of being real: **the reliability score.**

Model versions and data



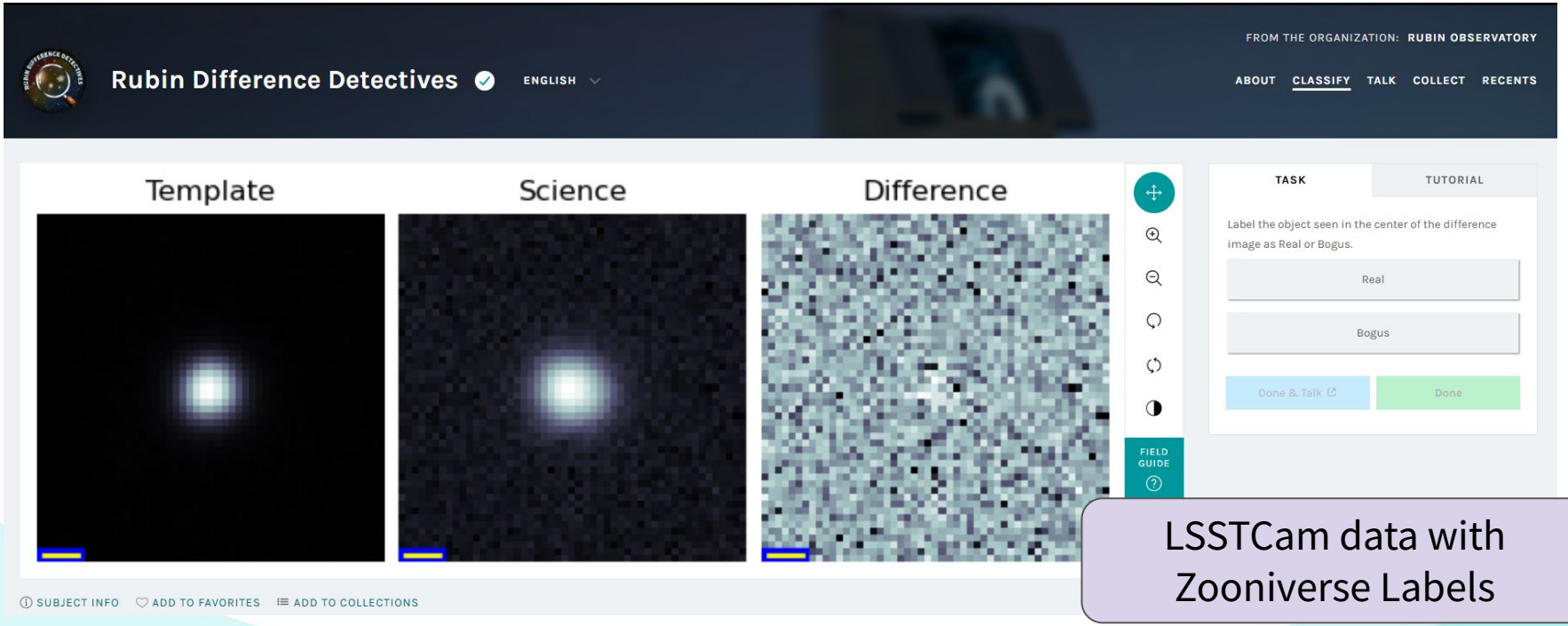
Rubin Difference Detectives: Citizen science project

Launch on November 11, 2025 and reached 100% completion on January 21, 2026



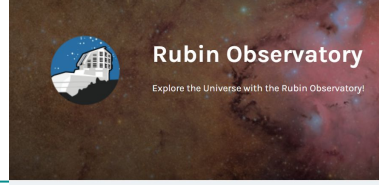
Rubin Difference Detectives: Citizen science project

Launch on November 11, 2025 and reached 100% completion on January 21, 2026



LSSTCam data with Zooniverse Labels

Rubin Difference Detectives



242K unique subjects**

2,075 Volunteers	1,339,732 Classifications
282,196 Subjects	230,236 Completed subjects

Active Projects



RUBIN DIFFERENCE DETECTIVES



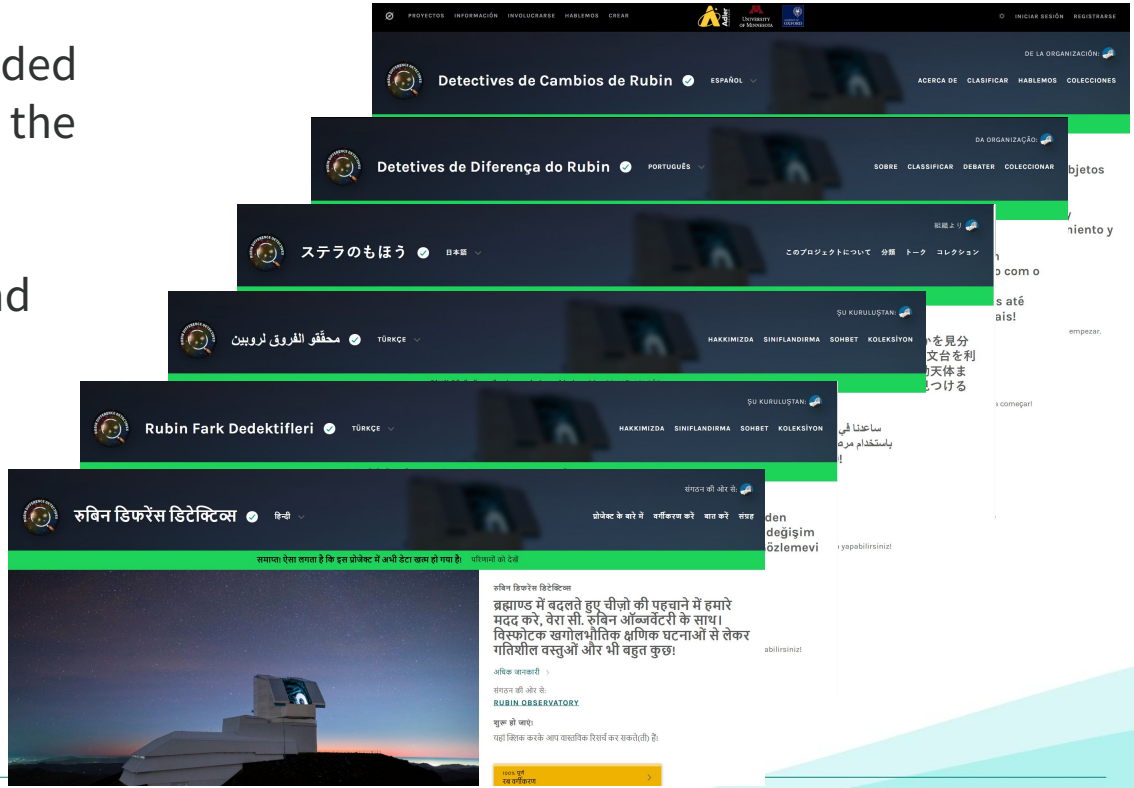
RUBIN COMET CATCHERS

	Origin	Num subjects
Transients	TNS catalog match	4,054
Solar system	Solar System catalog match	88,058
Variables	Gaia catalog match	89,410
	0.5< reliability v0.1 model < 0.9	49,993
	Sources w extendedness = 1	10,932

Rubin Difference Detectives - Available in 8 languages

Two (english and spanish) provided directly by the team developing the Zooniverse Project.

Six provided by volunteers, and (some) review by our team (japanese and italian).



Rubin Difference Detectives - Forum

Rubin Difference Detectives Talk

Search or enter a #tag

Notes

General comment threads about individual subjects

145 Participants
1484 Discussions
1919 Comments

jcash_scsu Subject 113677538 *3 months ago*

General

For discussing anything that doesn't seem to belong to one of the existing boards.

28 Participants
22 Discussions
83 Comments

Barbalbero TRANSLATOR Paper about Rubin *6 months ago*

Troubleshooting

A place to bring up any challenges encountered during classification for this project.

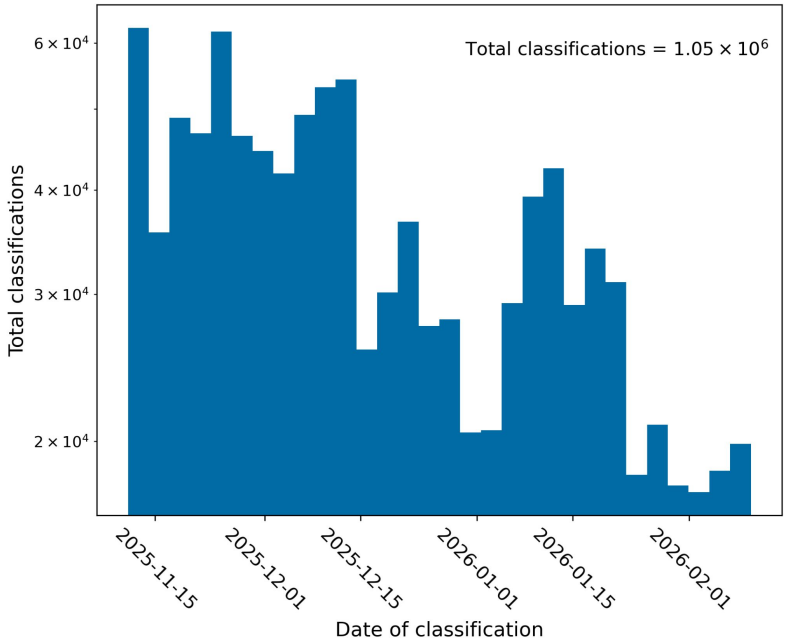
14 Participants
8 Discussions
24 Comments

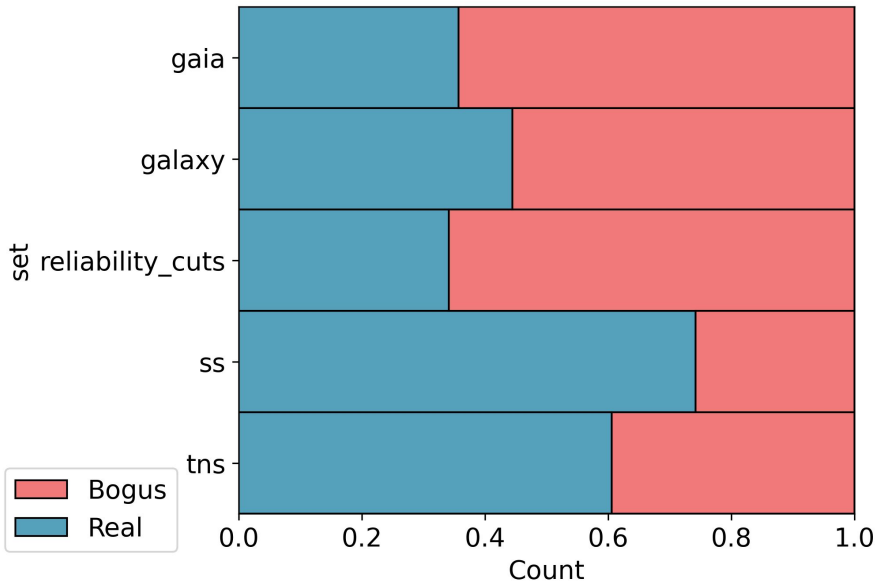
CChaser Any option to go back and correct classification? *7 months ago*

Popular Tags:

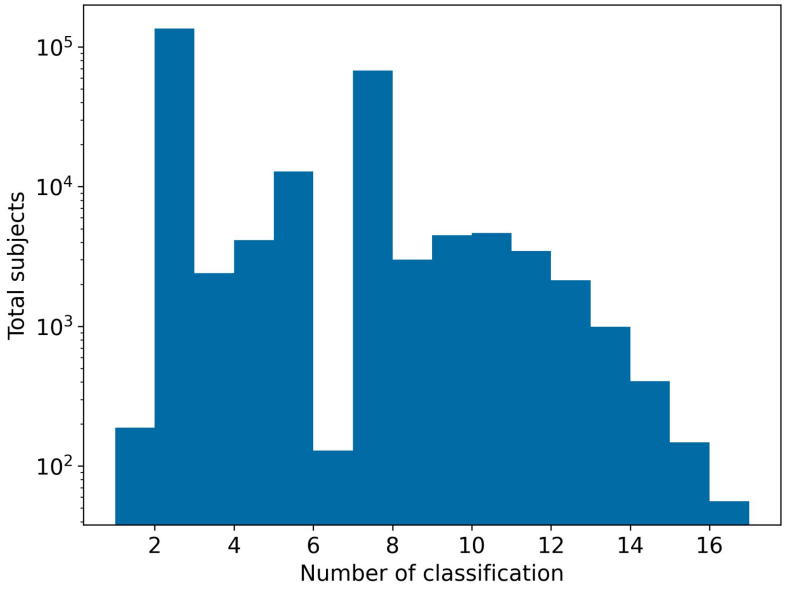
faint
noise
real
bogus
supernova
unsure
focus
asteroid
transient
bad_template
cosmic-ray
diffraction-spikes
cosmicray
comet
doubtful
too-noisy
trail
two-stars
bad_image
cosmic

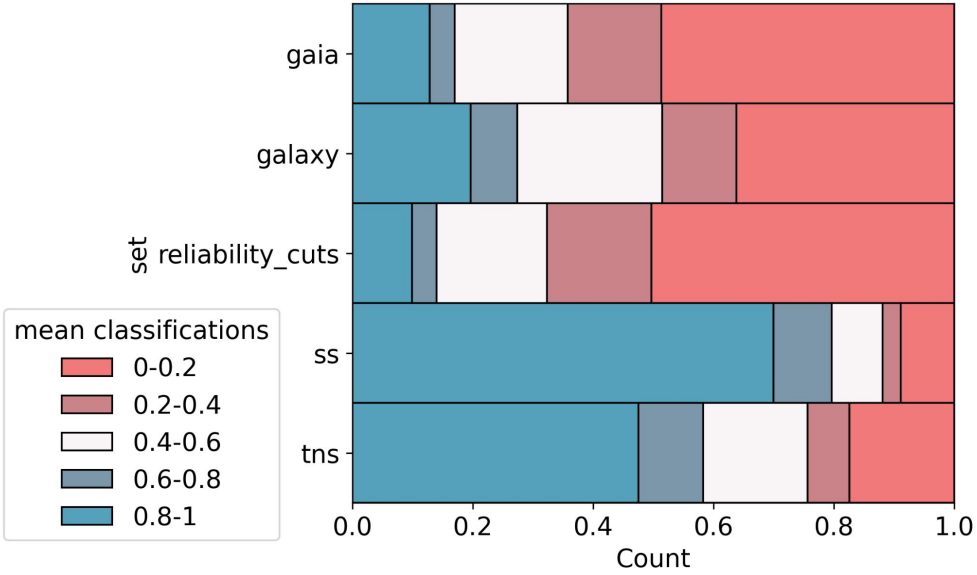
General statistics



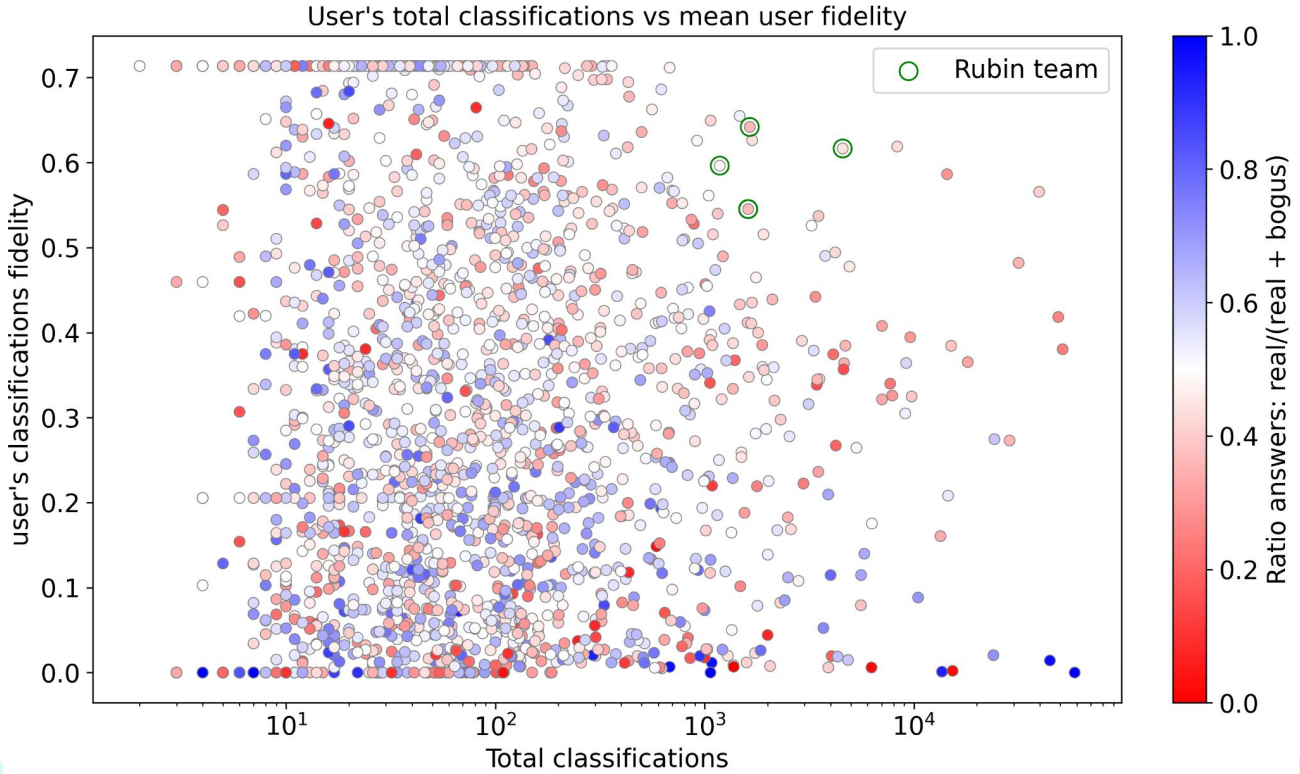


General statistics by subject

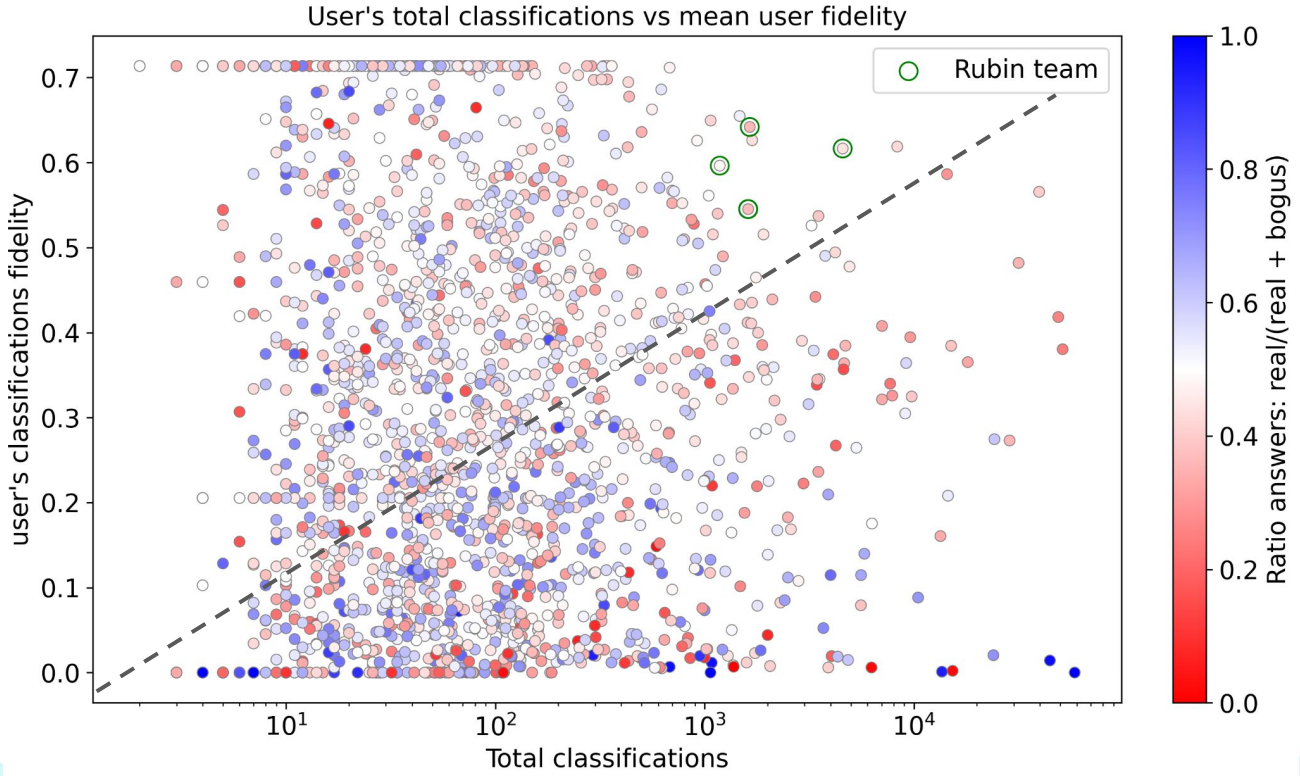




More classifications != reliable classifications



More classifications != reliable classifications



Analysis of user's classifications

SPACE WARPS: I. Crowd-sourcing the Discovery of Gravitational Lenses
Marshall, Phill et al (2016): https://doi.org/10.1093/mnras/stv2009

$$P(\text{Real}|C) = \frac{P(C|\text{Real}) \cdot P(\text{Real})}{[P(C|\text{Real}) \cdot P(\text{Real}) + P(C|\text{Bogus}) \cdot P(\text{Bogus})]}$$

Each classification have a different weight that depends on how well they did on a **reference data**. (sources with know ground truth)

Analysis of user's classifications

SPACE WARPS: I. Crowd-sourcing the Discovery of Gravitational Lenses
Marshall, Phill et al (2016): https://doi.org/10.1093/mnras/stv2009

$$P(\text{Real}|C) = \frac{P(C|\text{Real}) \cdot P(\text{Real})}{[P(C|\text{Real}) \cdot P(\text{Real}) + P(C|\text{Bogus}) \cdot P(\text{Bogus})]}$$

For each subject, the 'prior' is update with the previous posterior.

Each classification have a different weight that depends on how well they did on a **reference data**. (sources with know ground truth)

Analysis of user's classifications

SPACE WARPS: I. Crowd-sourcing the Discovery of Gravitational Lenses
Marshall, Phill et al (2016): https://doi.org/10.1093/mnras/stv2009

For each subject, the 'prior' is update with the previous posterior.

$$P(\text{Real}|C) = \frac{P(C|\text{Real}) \cdot P(\text{Real})}{[P(C|\text{Real}) \cdot P(\text{Real}) + P(C|\text{Bogus}) \cdot P(\text{Bogus})]}$$

$$P(\text{Real}|\text{'Real'}) = \frac{TP \cdot P(\text{Real})}{[TP \cdot P(\text{Real}) + FP \cdot P(\text{Bogus})]}$$

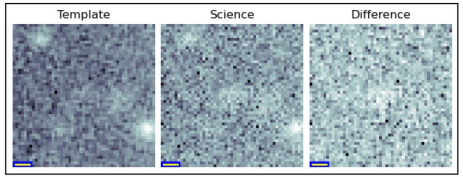
Each classification have a different weight that depends on how well they did on a **reference data**. (sources with know ground truth)

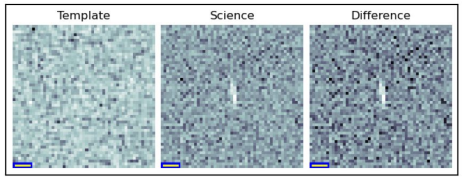
The user information:

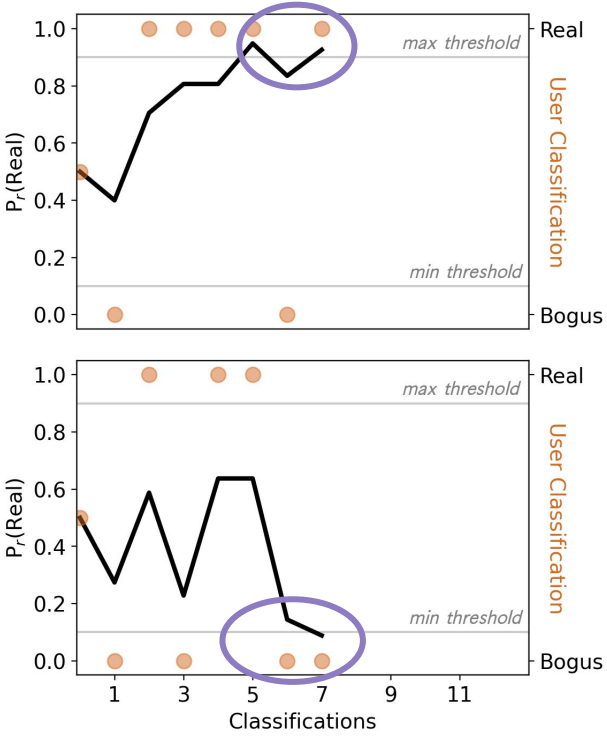
FP = 1-TN	**TP = correct R / total R**
TN = correct B / total B	**FN = 1 - TP**

R: real
B: bogus

Final Posterior and Stability criteria





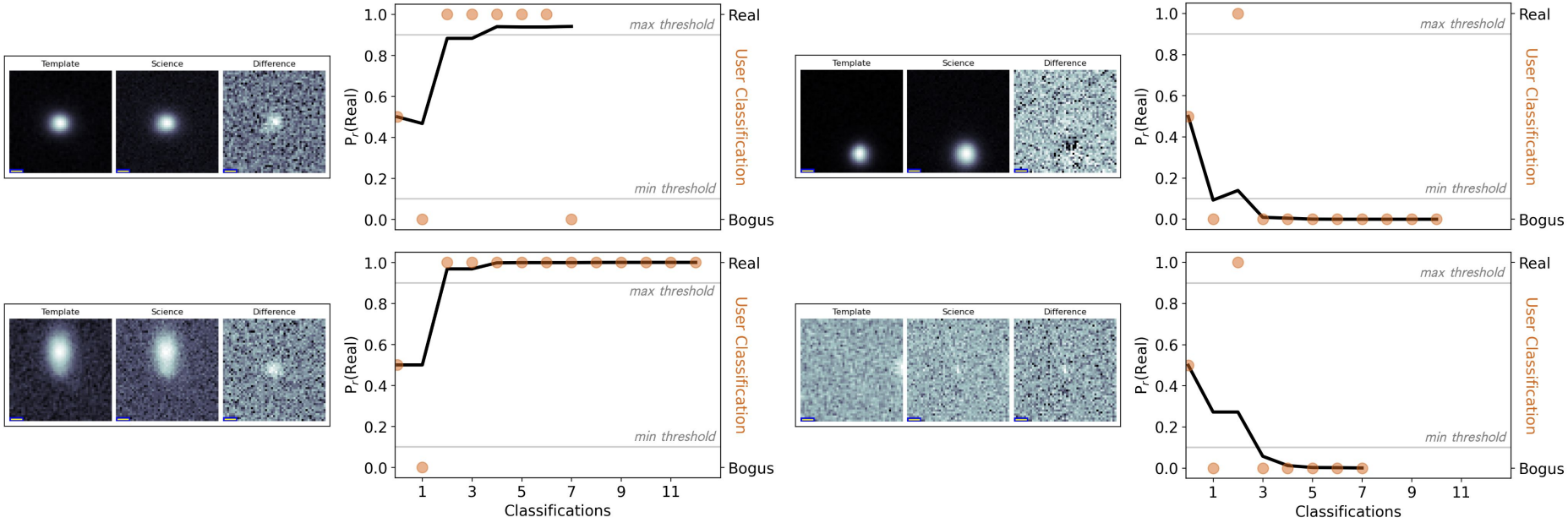


Real: Final posterior > 0.9
Bogus: Final posterior < 0.1

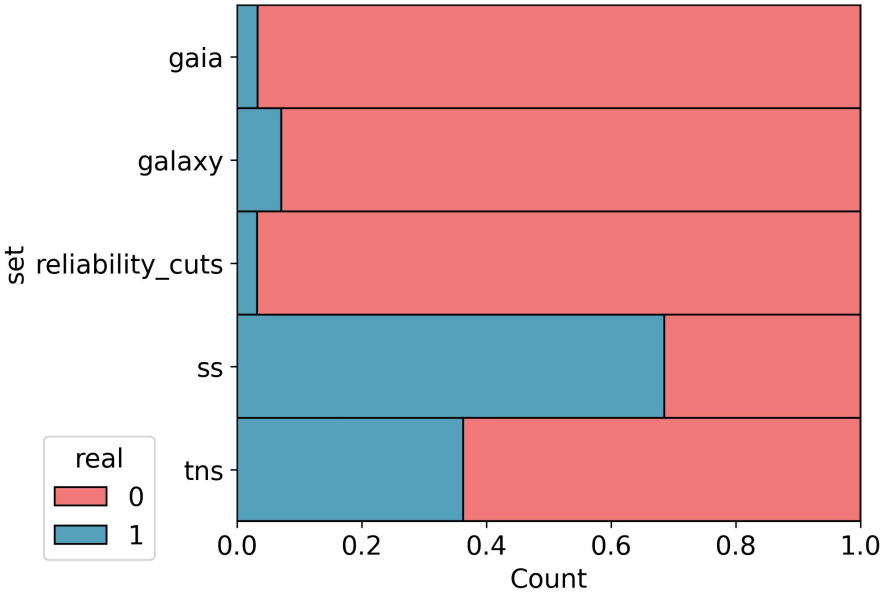
Posterior for the last three classifications has to be above/below the real/bogus threshold.

Examples that did not pass stability criteria

High confidence example (stable)

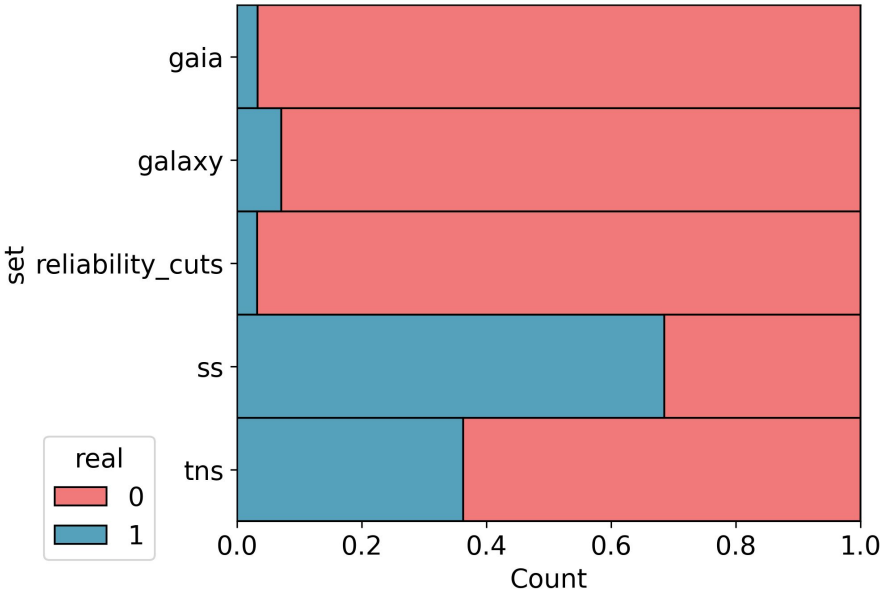


Data distribution



Scenario	Total Real	Total Bogus	Reference Data
stable	8236	40,402	Rubin
unstable	25,218	53,024	Rubin
stable	*8025*	*46,124*	*Oxford & Rubin*
unstable	*23,889*	*57,570*	*Oxford & Rubin*

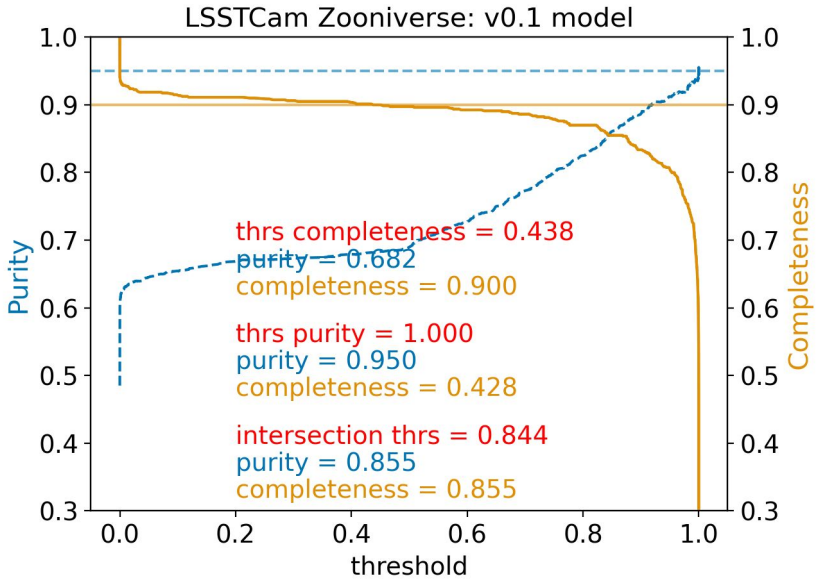
Data distribution



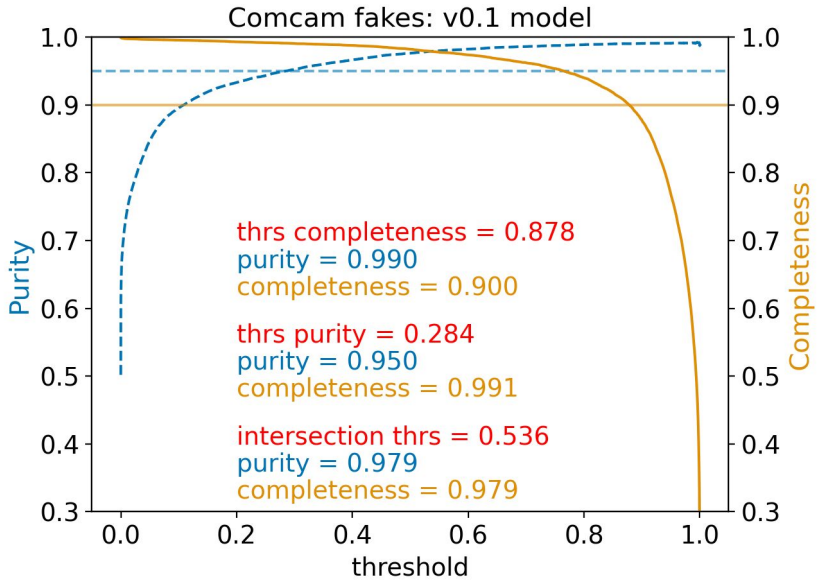
Scenario	Total Real	Total Bogus	Reference Data
stable	8236	40,402	Rubin
unstable	25,218	53,024	Rubin
stable	*8025*	*46,124*	*Oxford & Rubin*
unstable	*23,889*	*57,570*	*Oxford & Rubin*

v0.1 model performance - Only fakes

‘real’ data with zooniverse labels



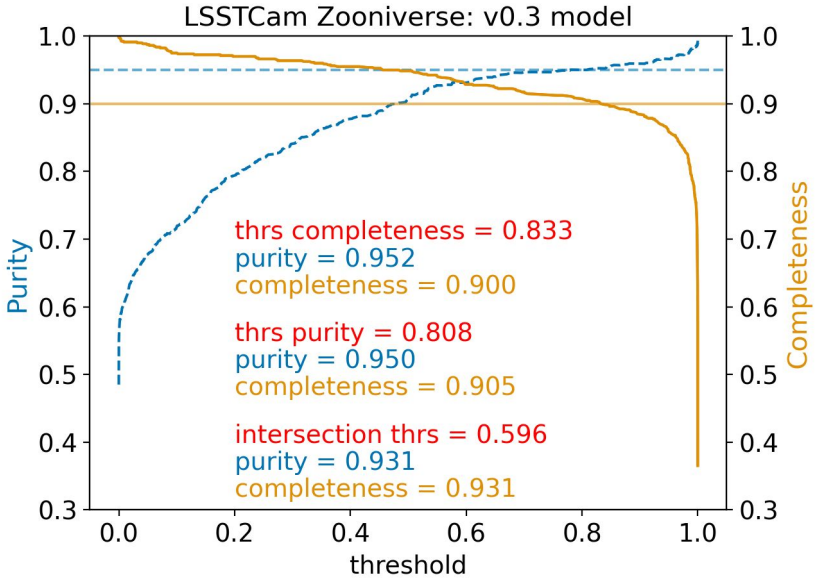
LSSTComcam Fakes



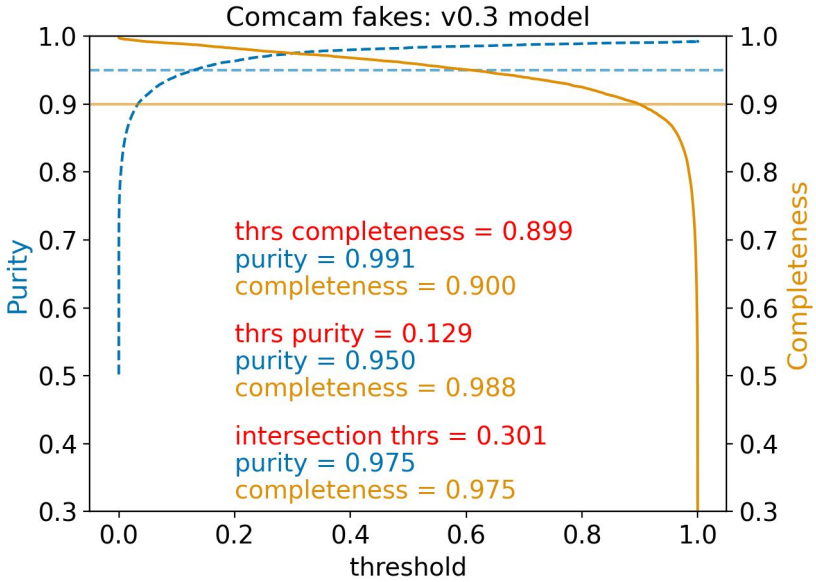
Note that performance is calculated on a test set and may not be fully representative of on-sky performance.

v0.3 model performance - Including citizen science

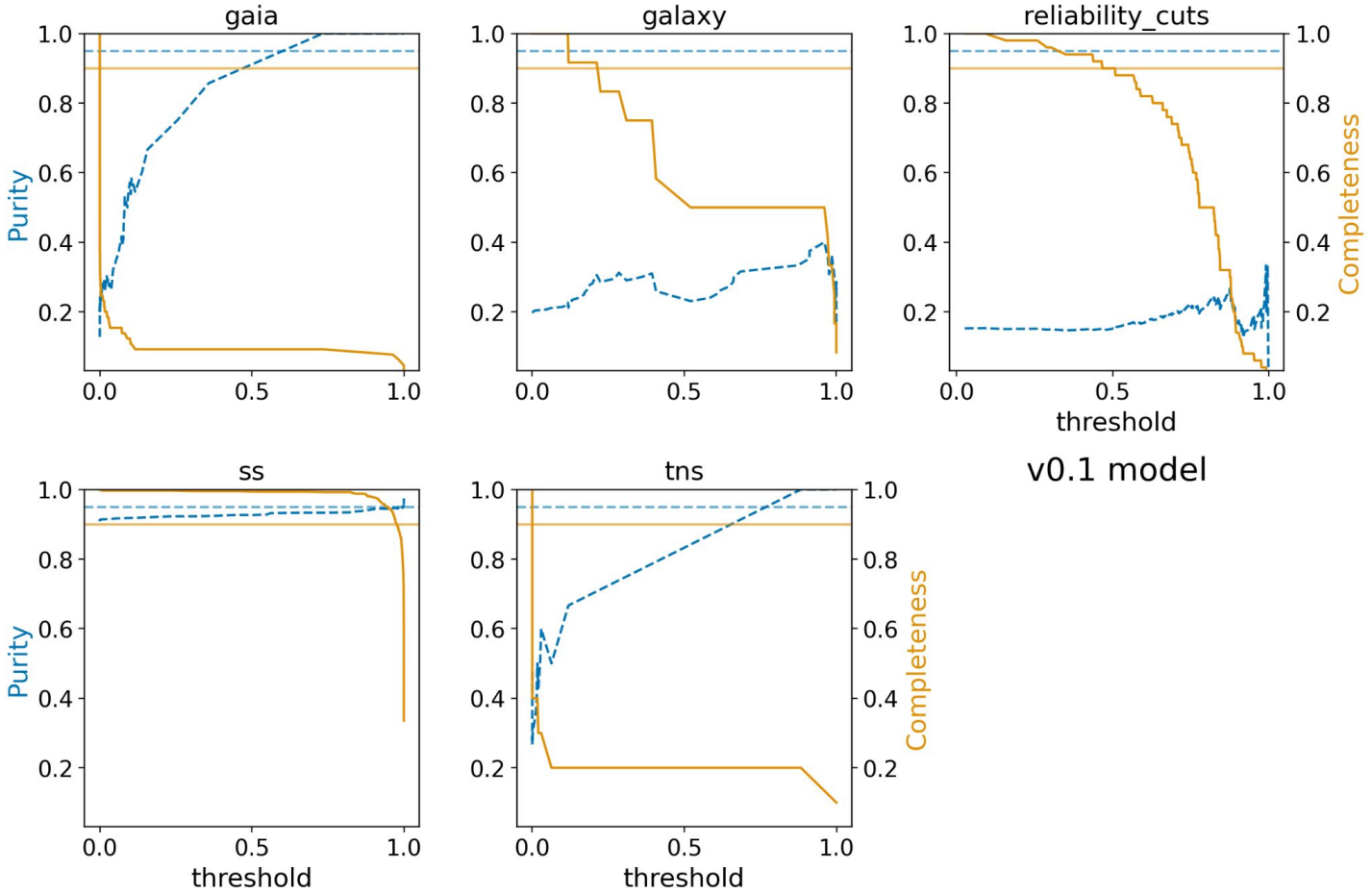
‘real’ data with zooniverse labels

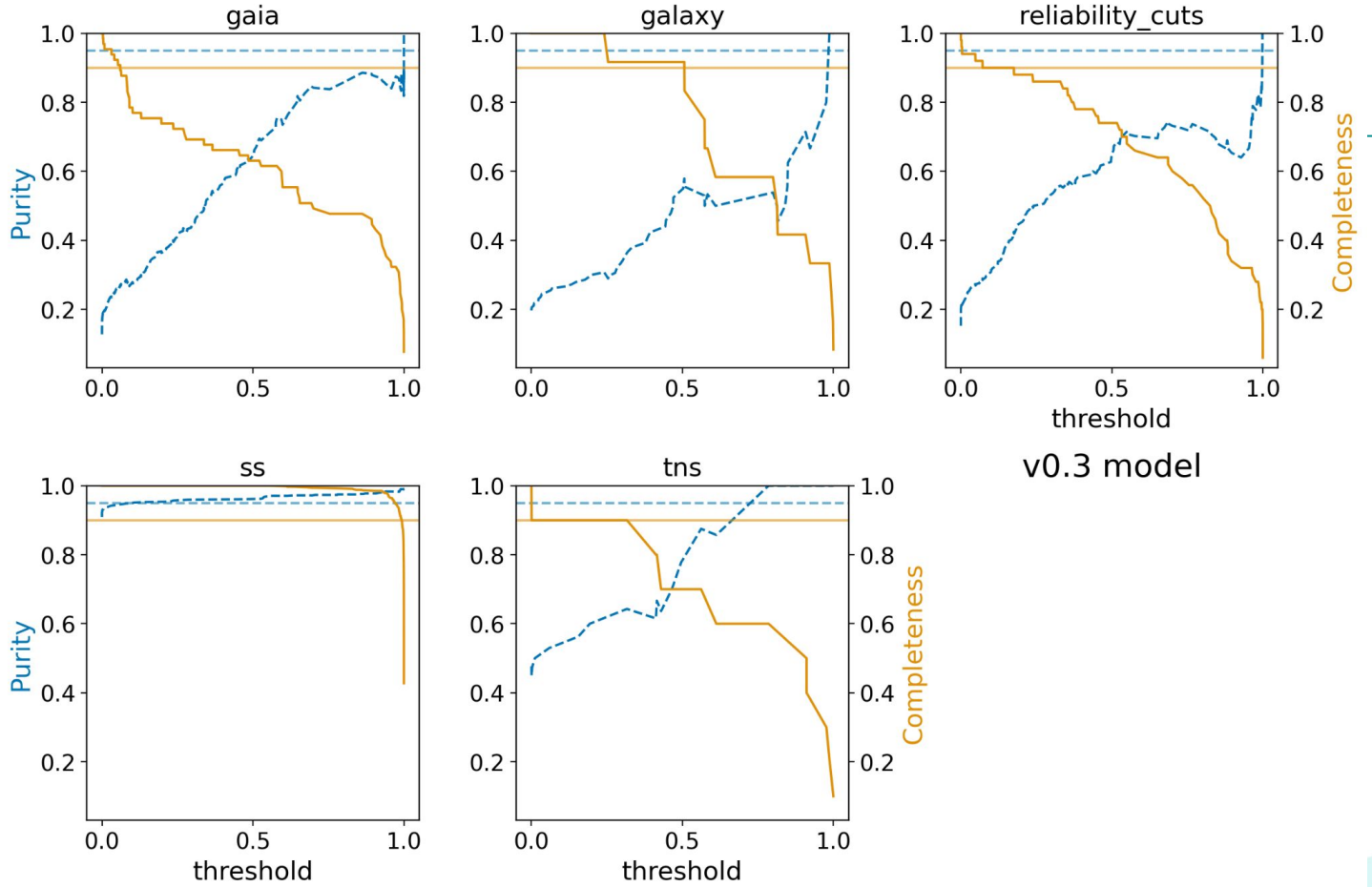


LSSTComcam Fakes



Note that performance is calculated on a test set and may not be fully representative of on-sky performance.


gaia
galaxy
reliability_cuts
ss
tns
Purity
Completeness
threshold
v0.1 model

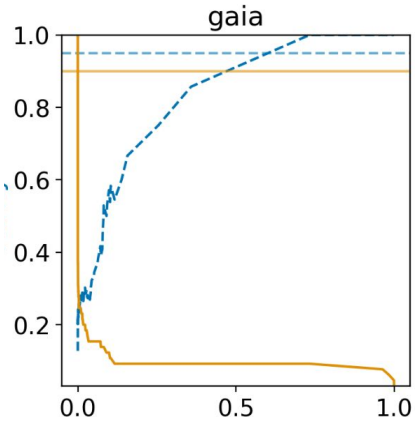

gaia
galaxy
reliability_cuts
ss
tns
Purity
Completeness
threshold
v0.3 model

real data is great because we can take …

Purity
Completeness

real data is great because we can take …

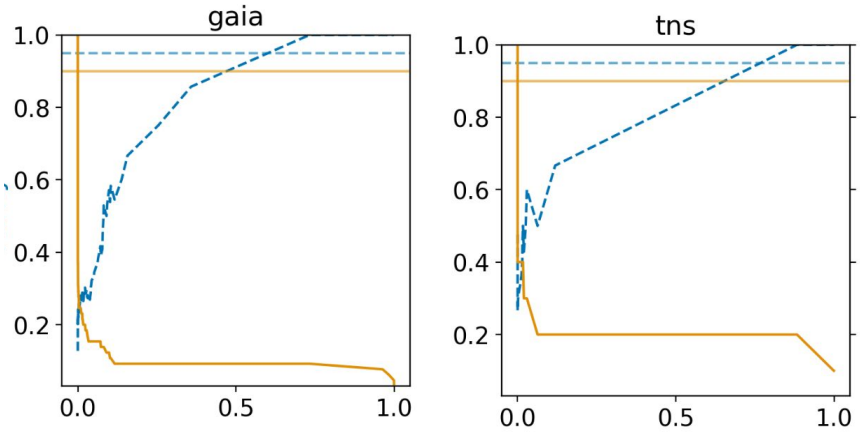
Purity
Completeness



This and this

real data is great because we can take …

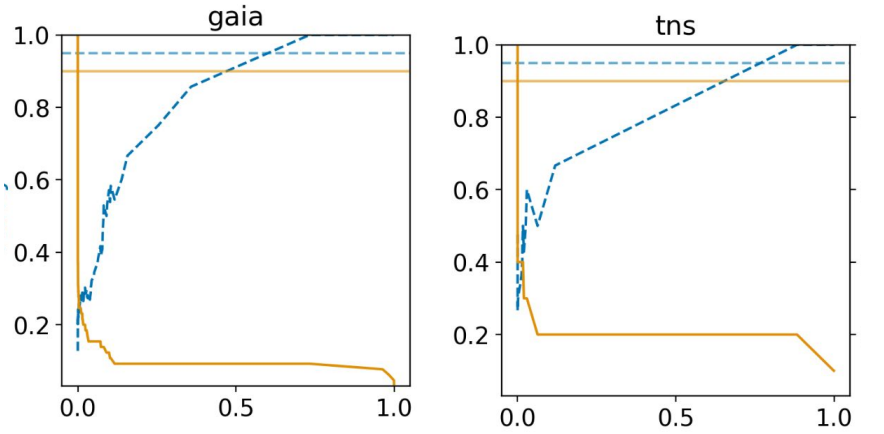
Purity
Completeness



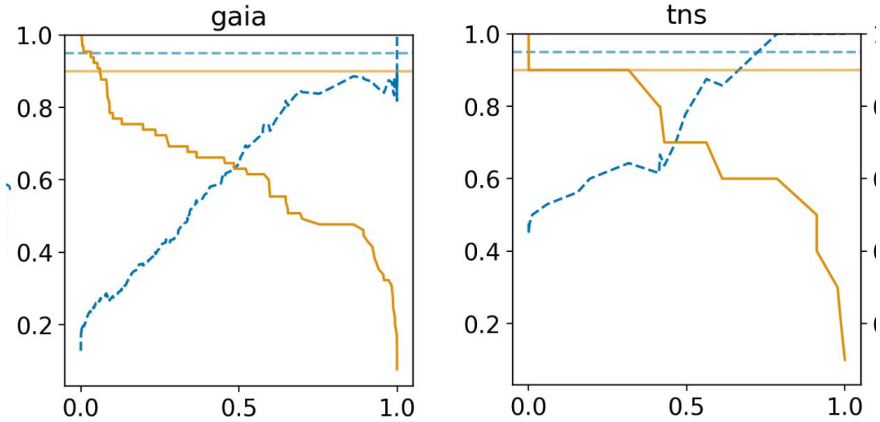
This and this

real data is great because we can take …

Purity
Completeness



This and this

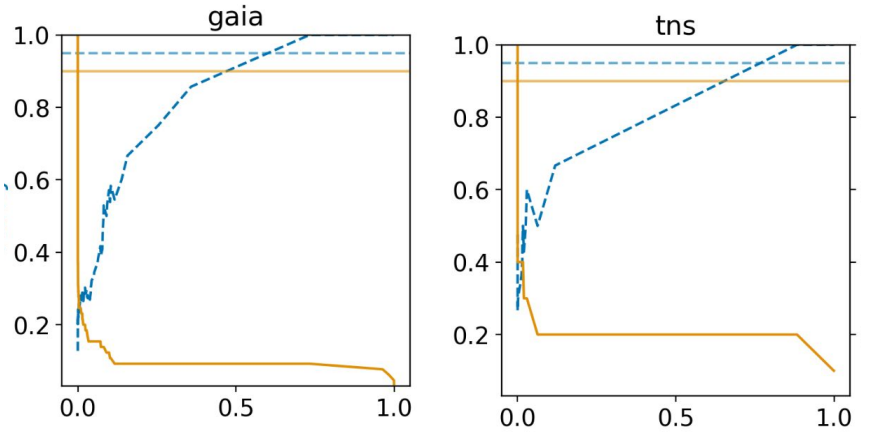


And give you

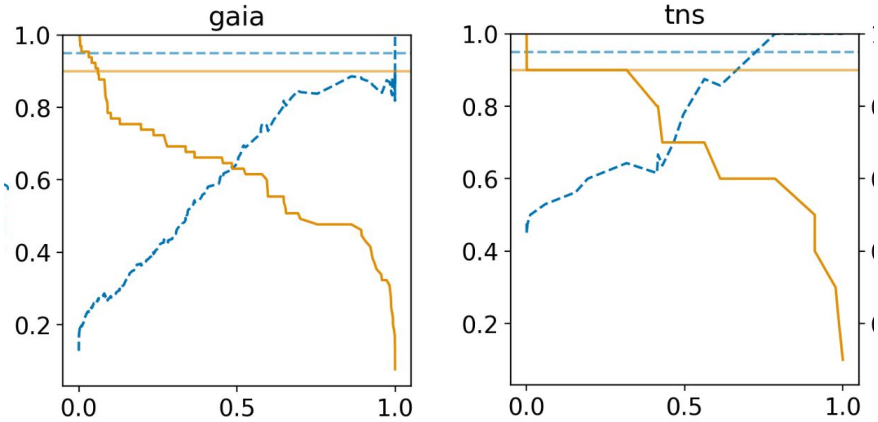


real data is great because we can take …

Purity
Completeness



This and this

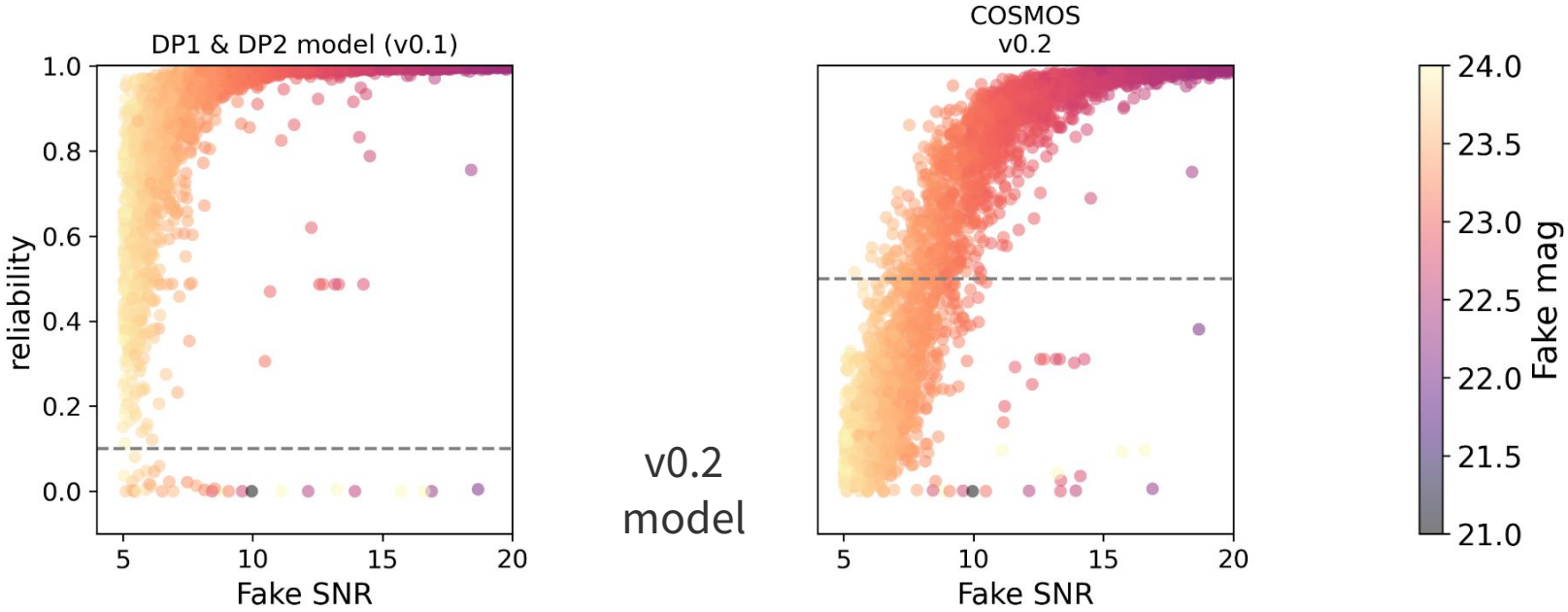


And give you



By just given a few examples to the model!

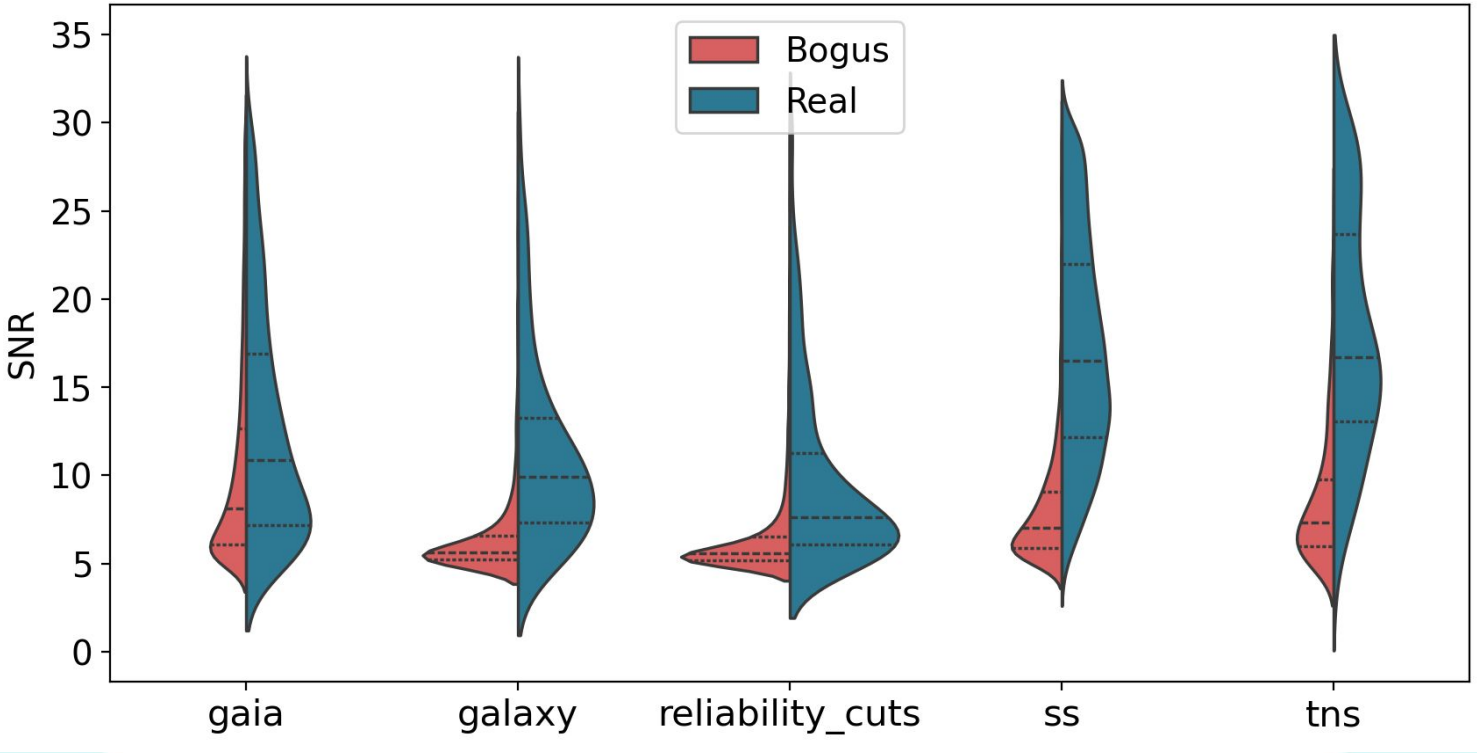
But deep learning models forget so easy…



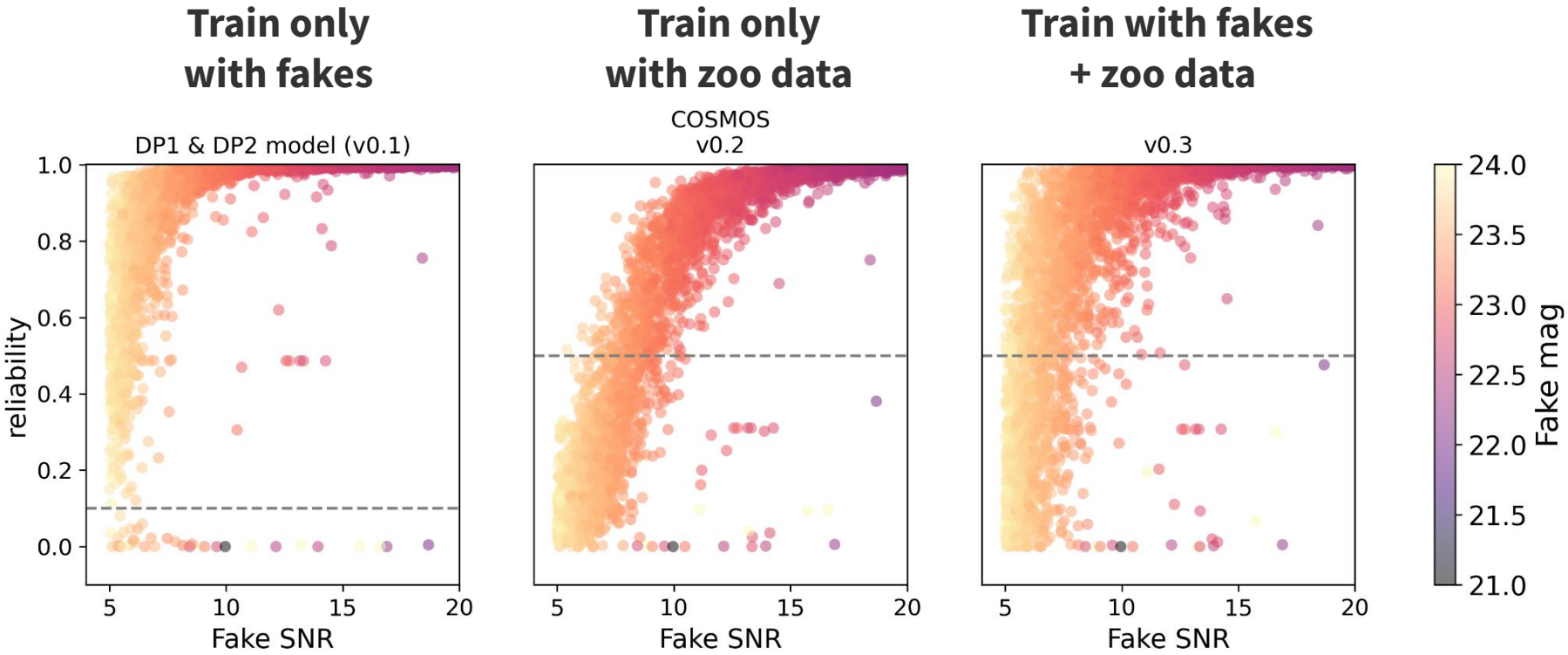
Only including fakes

Only using zoo data

Faint sources are ambiguous



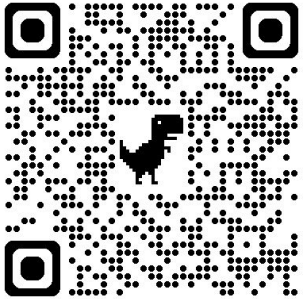
We need to keep fakes to prevent biasing the model against faint sources

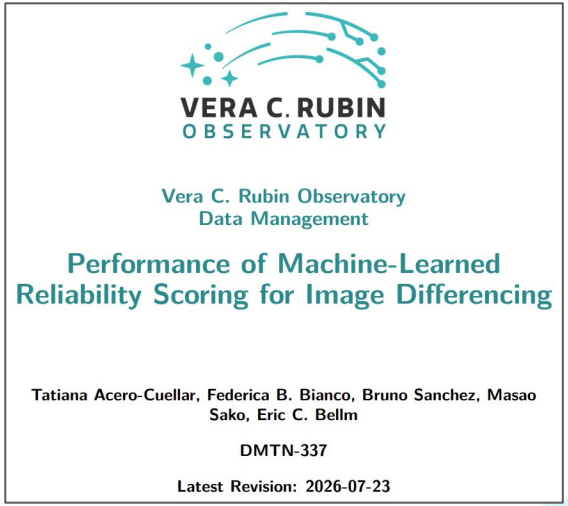


Documentation

More information about the most reliability score for alerts: https://prompt-products.lsst.io/processing/filtering/index.html#filtering-and-reliability-scoring

More detail information in DMTN-337





Current work

The low performance for variables has not been solved yet -> the solution is ‘simply’ that we need more data, but collecting data is not the problem, collecting labels is the real challenge.

Rubin Difference Detectives reached completion and is on pause at the moment while we decide how to best proceed. (collect new data? focus on specific sources?)

Use LSSTCam fakes. (use them as the reference data?)

A lot of work for just one person!


Tatiana Acero-Cuellar


Federica Bianco


Eric Bellm


Masao Sako


Bruno Sánchez


Harshit Rai


Fernanda Urrutia


Clare Higgs


Eric Rosas


Jake Kurlander


Nima Sedaghat


Chris Lintott

& many more!

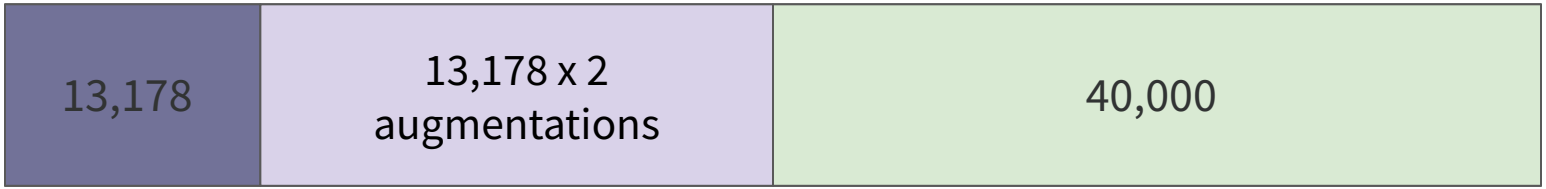
Thank you!

Summary usage of model

Model version	AP deployment date	AP reliability cutoff	DRP usage
0.1	September 2025	0.1	DP1 & DP2 model
0.2	18 February 2026	0.1 (before Feb 24, 2026) 0.5 (after Feb 24, 2026)	
0.3	9 July 2026	0.5	

Starting with v0.3, the model version is now visible: “reliabilityVersion” field in DIASource table.

Data used to fine-tuned the current model (v0.3)



LSSTCam data with Zooniverse Labels

Each set with 50% Reals and 50% Bogus

- Zooniverse labels: high-confidence labels from zooniverse
- Vertical and Horizontal flip of sources with zooniverse labels (high confidence)
- LSSTComcam Fakes

Reliability cutoff for Alerts

