



**SNAD**

# **SNAD: anomaly detection for large scale time-domain astronomy**

**Patrick Aleo (UIUC), Emille Ishida (CNRS-LPC), Matwey Kornilov (SAI MSU), Vladimir Korolev, Konstantin Malanchev (UIUC), Maria Pruzhinskaya (SAI MSU), Etienne Russeil (CNRS-LPC), Sreevarsha Sreejith (BNL), Alina Volnova (IKI RAS)**

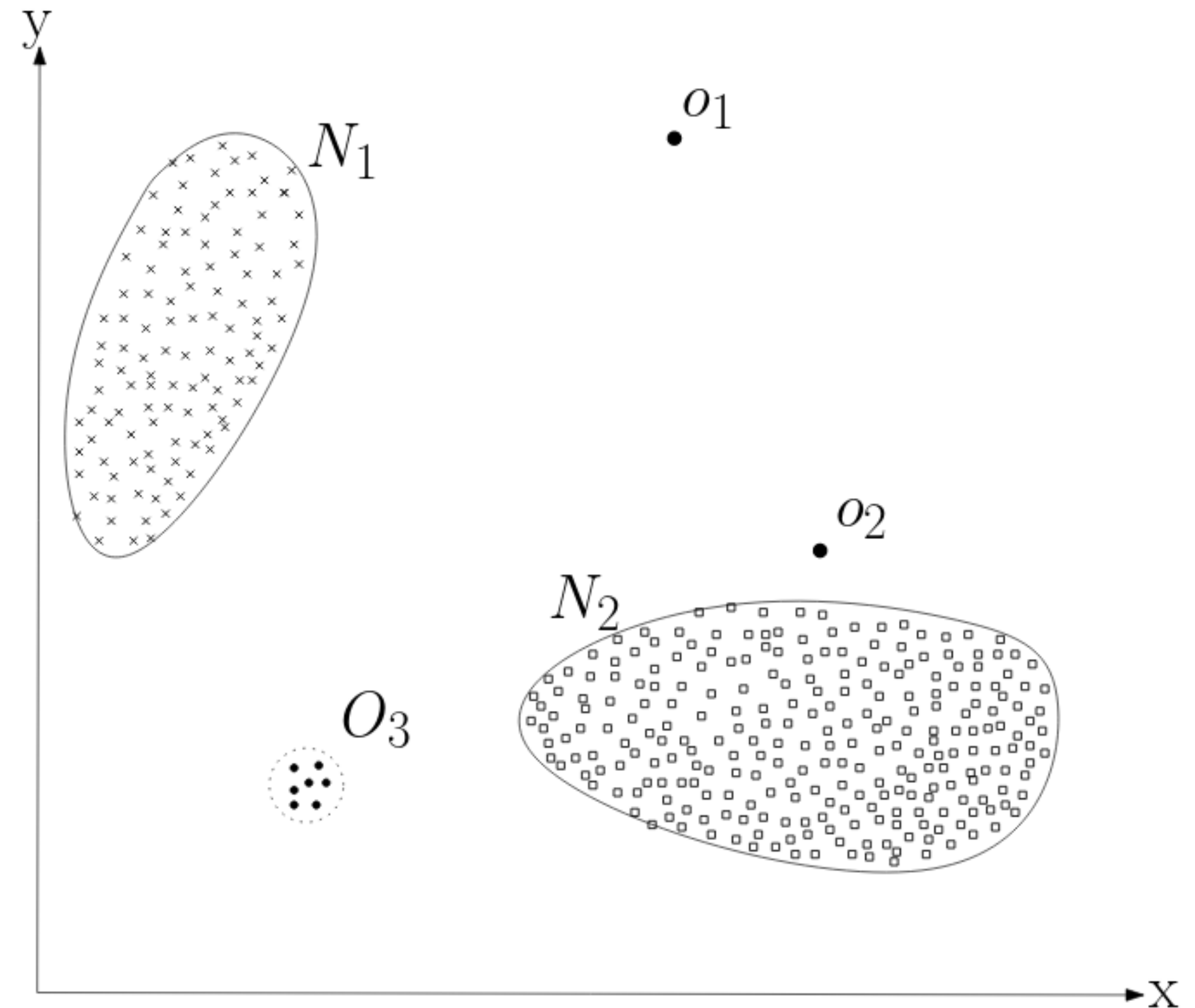
**on the behalf of the SNAD team  
+ all our side collaborators**

**Paris, 2022.06.21**

# Anomaly detection

## We look for anomalies

- Def. *Outlier* is an object located in a sparse region of the feature space
- Def. *Anomaly* is an astrophysical source having unusual properties for its class or a representative of some rare class



Chandola+ 2009



# SNAD

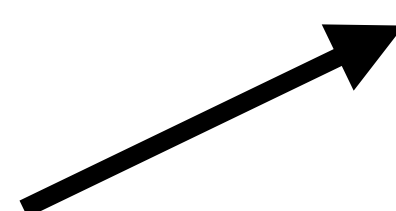
# Discovery

## ML only produces recommendations

Light curves



Preprocessing



Machine learning  
Outlier detector

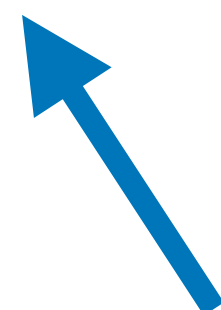


Potentially interesting anomalies:

- Candidate 1
- Candidate 2
- Candidate 3
- ...
- ...
- ...
- ...



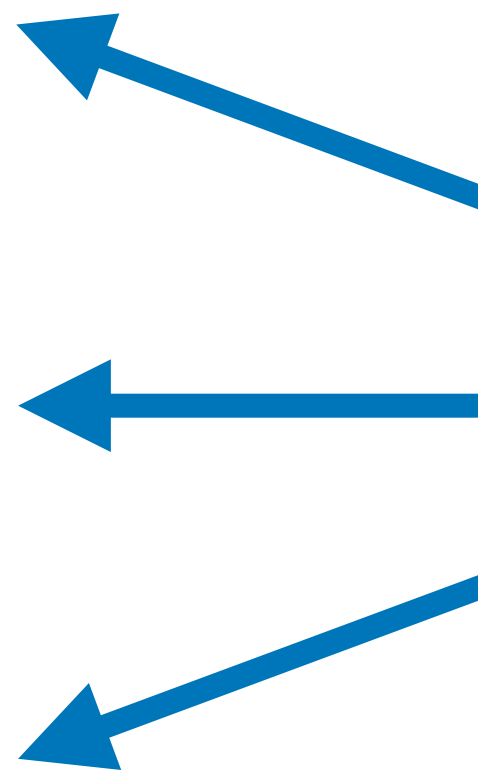
Get more  
data  
and  
Publication



Not interesting

Interesting

Very interesting!



metadata  
images  
simulations  
catalogs



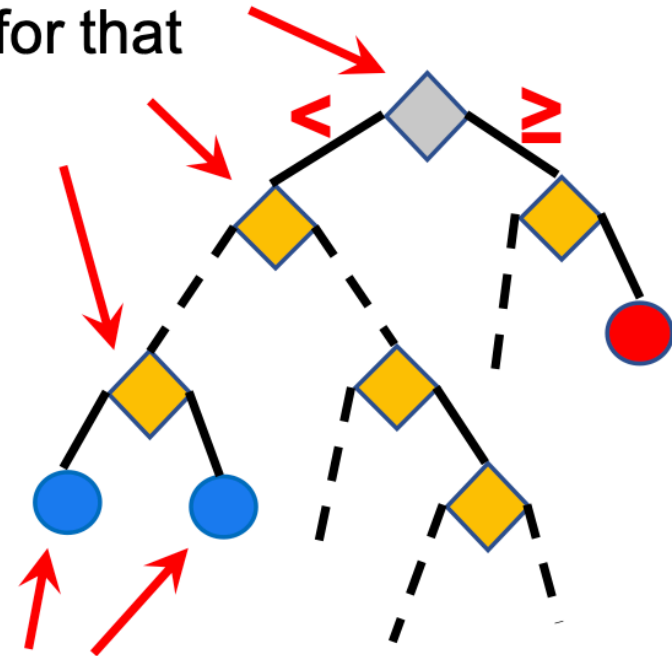




# Outlier detection: Isolation Forest

## iTree

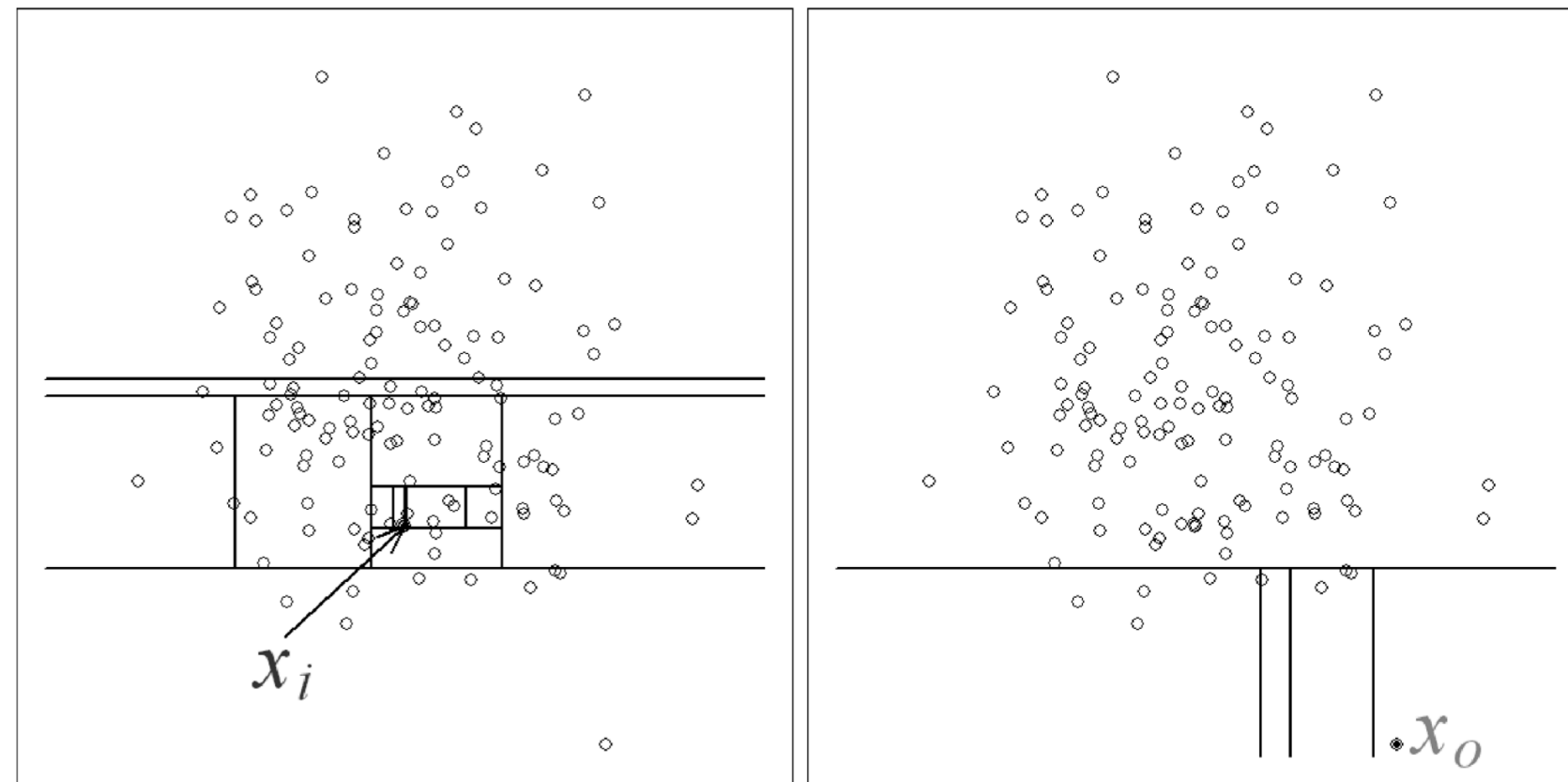
Select a random feature  
at each node, and a  
random split point for that  
feature



Shallower leaf nodes  
have higher anomaly  
scores, whereas, deeper  
leaf nodes have lower  
anomaly scores.

Leaf instance

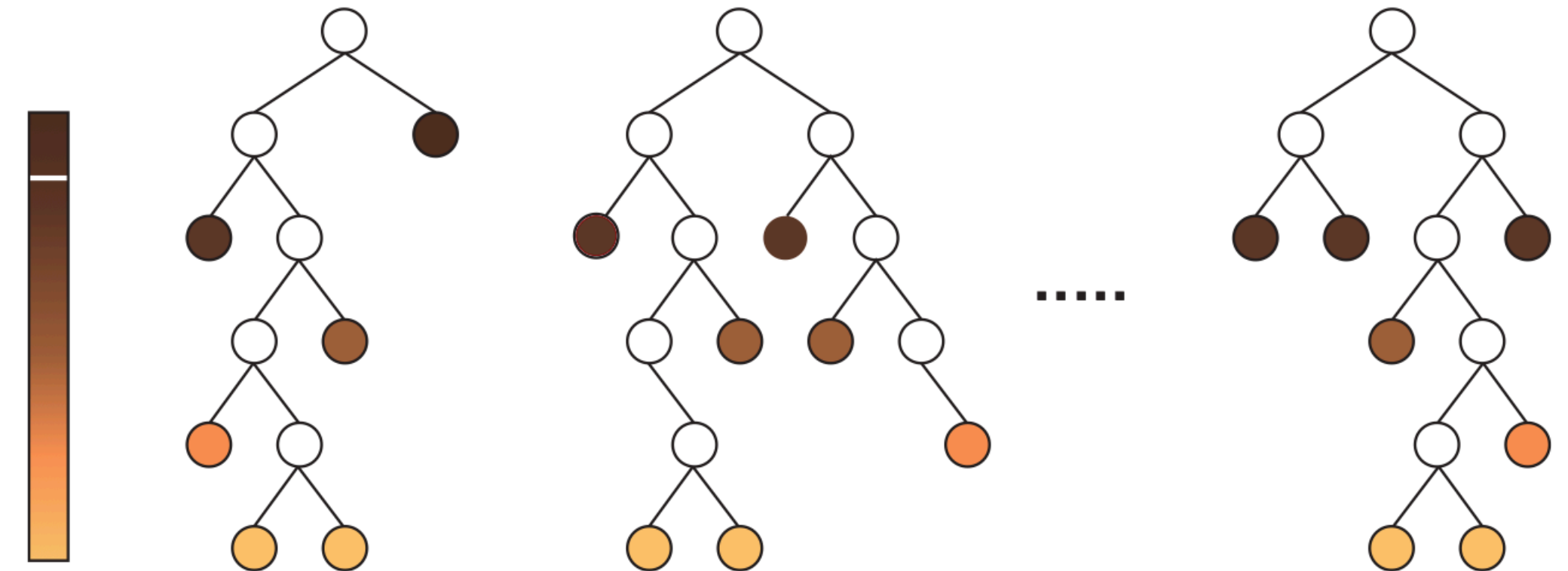
arXiv:1708.0944



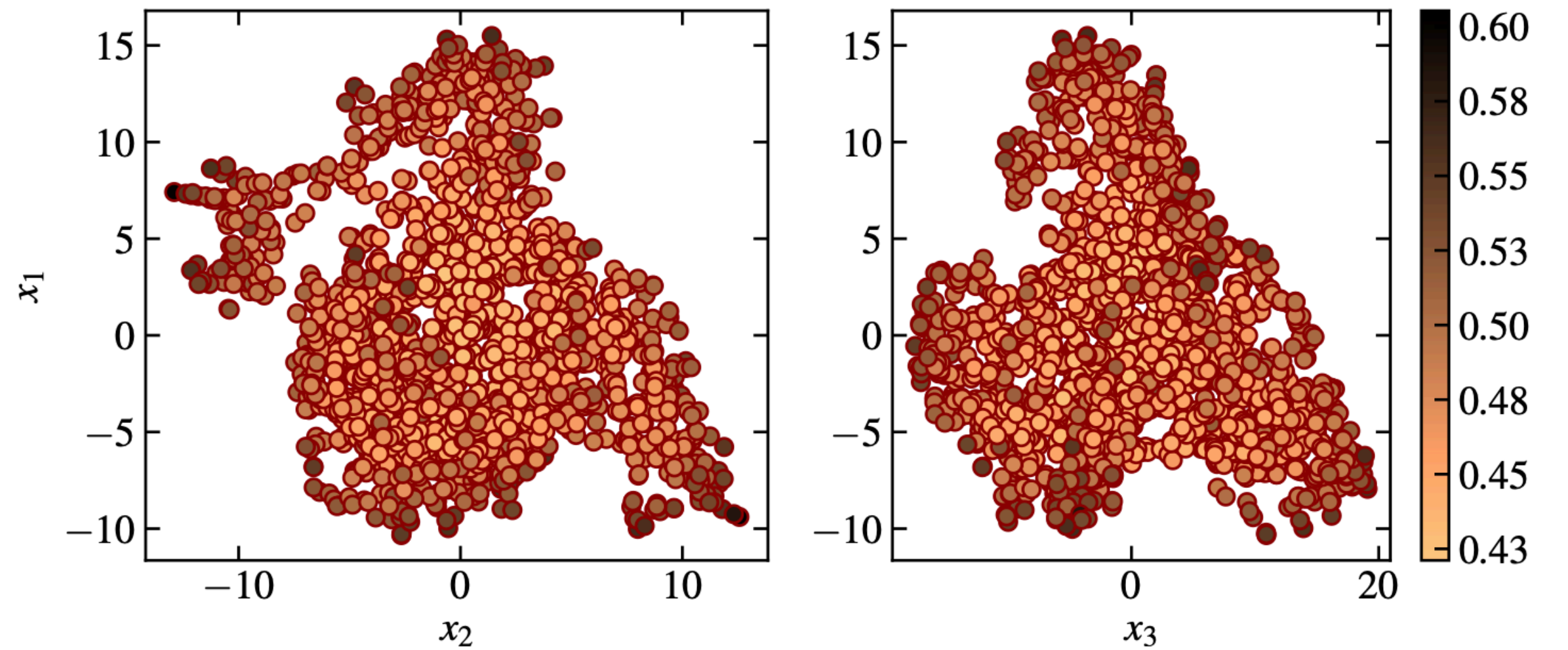
(a) Isolating  $x_i$

(b) Isolating  $x_o$

Liu+ 2008, Liu+ 2012



Darker is more anomalous



arXiv:1905.11516

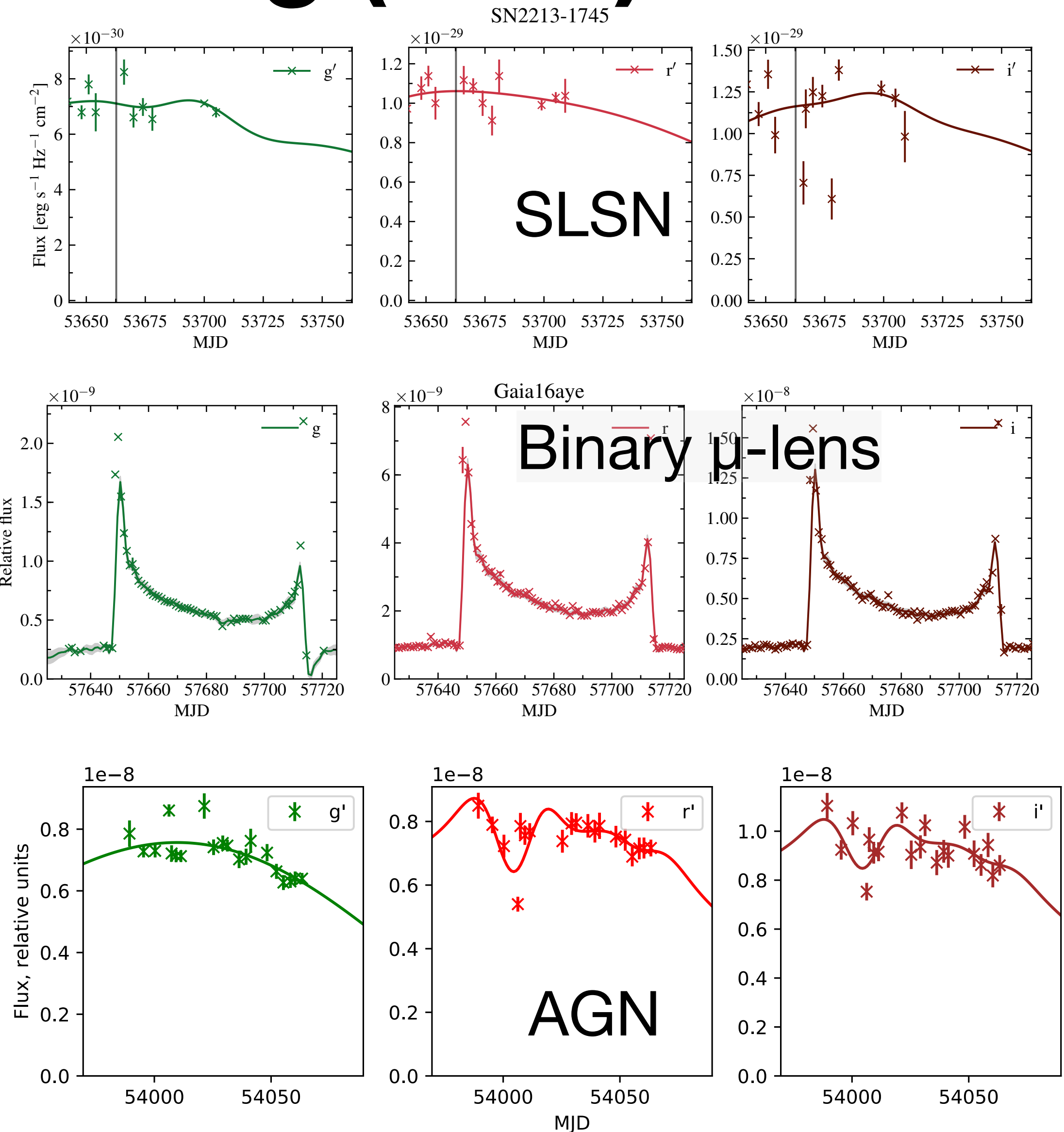




# Case: Open Supernova Catalog (OSC)

arXiv:1905.11516

- 1999 SNe in *gri*, *g'r'i'* & *BRI* taken from the OSC (Guillochon+ 2017)
- Multivariate Gaussian process approximation (Semenikhin+ in prep.) & t-SNE
- 30/100 anomaly candidates
  - Two known SLSNe
  - Several known peculiar SNe
  - Several known cases of misclassification, including binary  $\mu$ -lens
  - **16 previously unknown cases of misclassification** (10 stars and 6 AGNs), including SN 2006kg suggested as a "template" SN II (Okumara+ 2014)



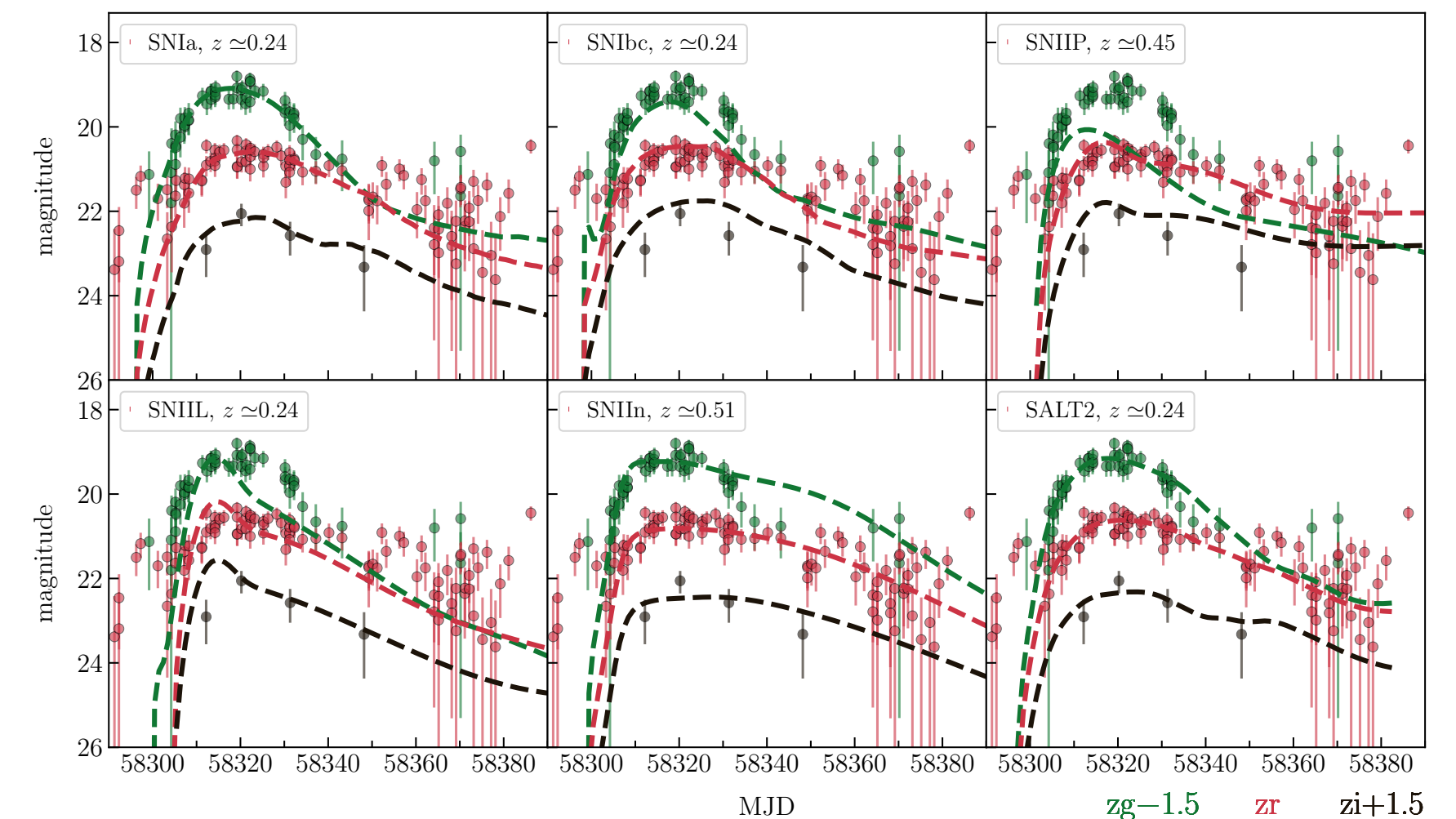
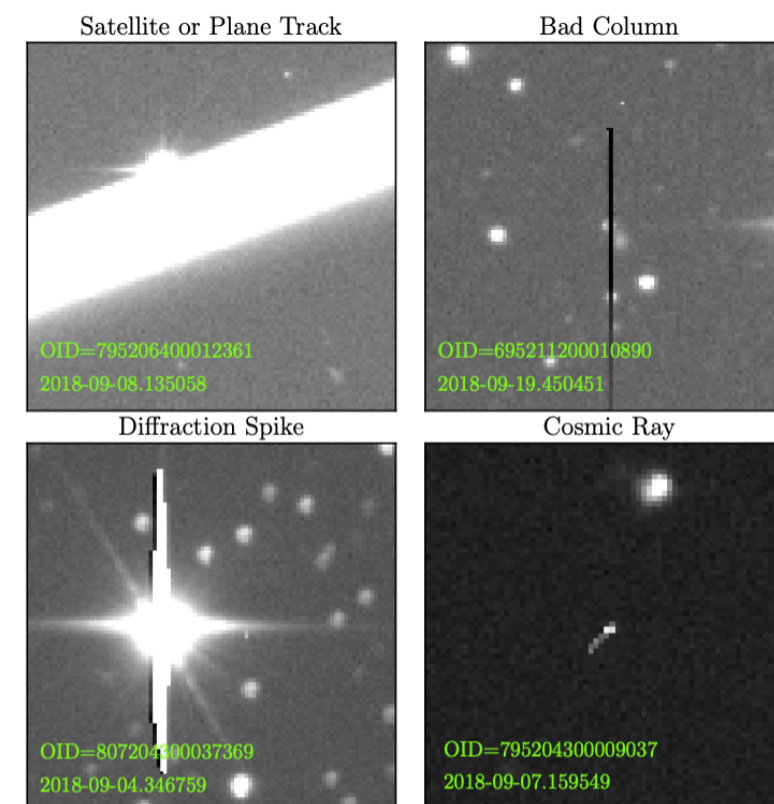
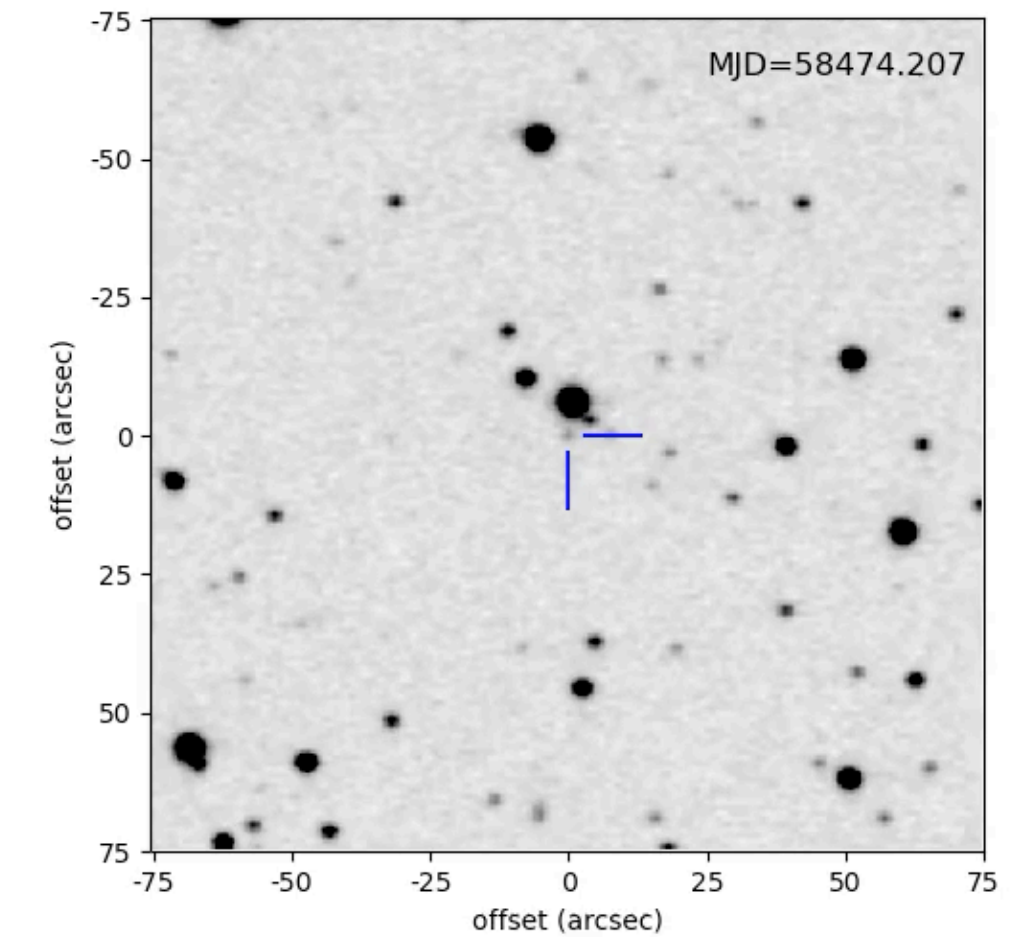
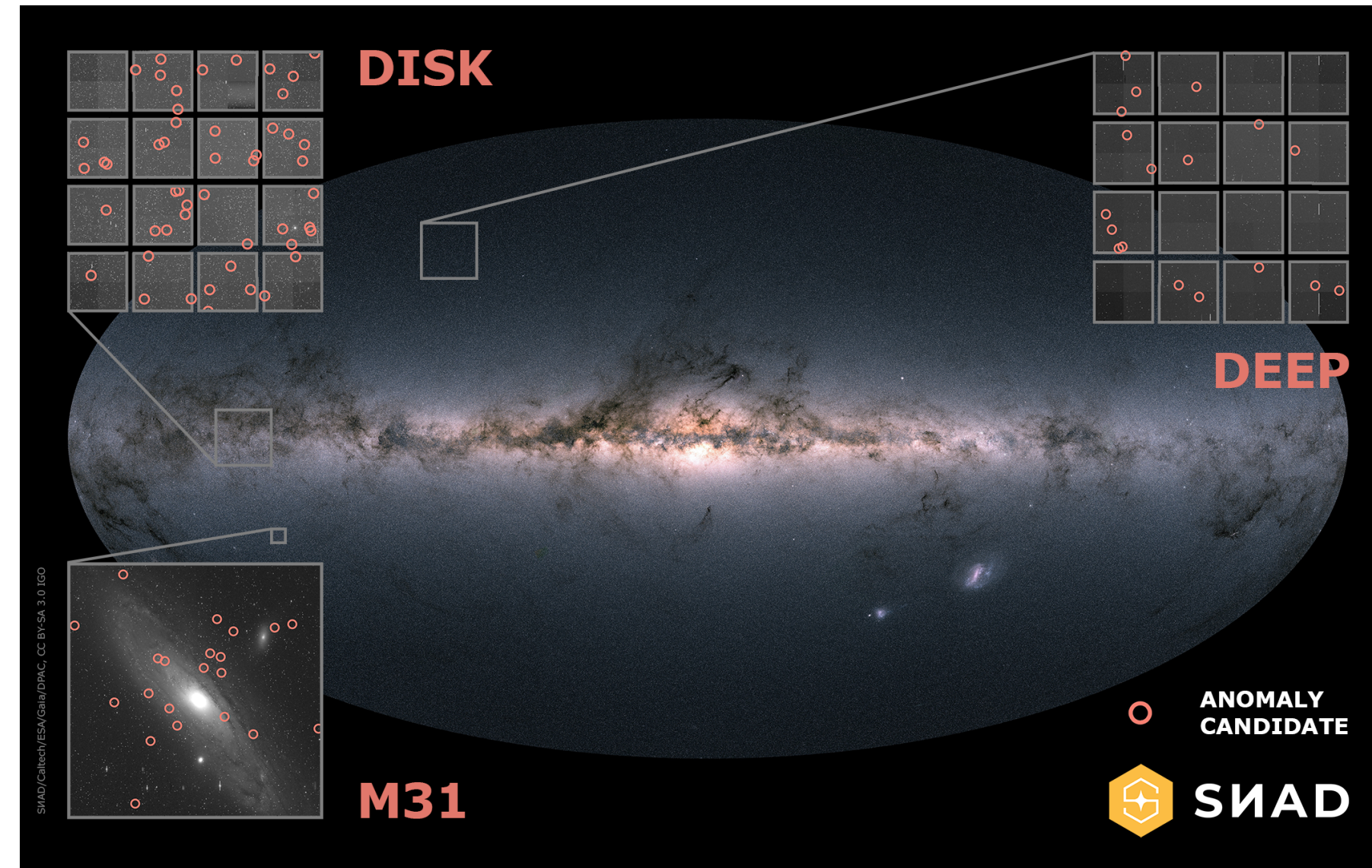




# Case: Zwicky Transient Facility DR3

arXiv:2012.01419

- Three fields of ZTF DR3
- $\sim 2 \times 10^6$  objects total
- Four outlier detection algorithms
- 89/227 anomaly candidates
  - Six (5/6 are new!) SN Ia candidates
    - RS CVn (confirmed by our spectra)
    - Mira binary candidate
- 188/277 bogus light curves
  - Double star defocusing
  - Bright Mira "echos"
  - Asteroid overlap
  - Bad columns, satellites, spikes, ghosts, etc







# Applying expert bias to anomaly detection

## From outlier to anomaly detection algorithm

- How to discriminate annoying non-anomalies sources and bogus light curves?
  - We can ask an expert interactively about each new outlier
  - If it is not an anomaly, set lower probability to objects like this
  - Retrain, ask the expert again
- We can do the opposite: highlight interesting class of objects for classification of rare objects. Listen Emille Ishida's talk about this

Data

Train initial model

Machine

The best outlier  
up date

Update model  
with the  
outlier label



Inspect outliers  
using external  
data



# Active anomaly detection (AAD)

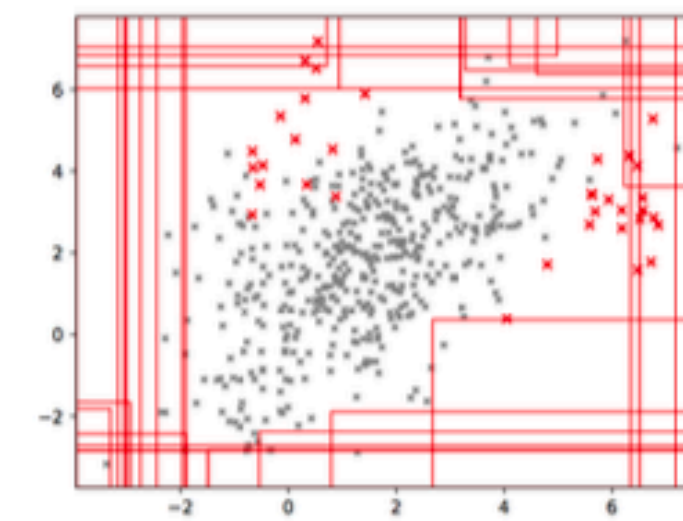


SNAD

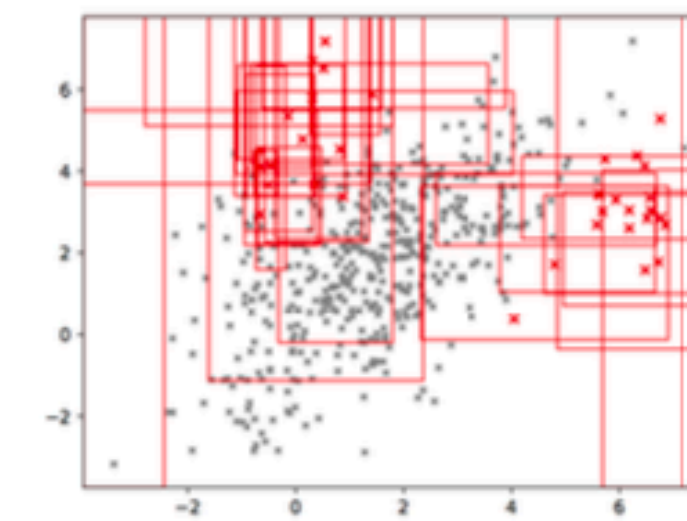
## Implementation of the machine–expert loop, Das+2018

Algorithm:

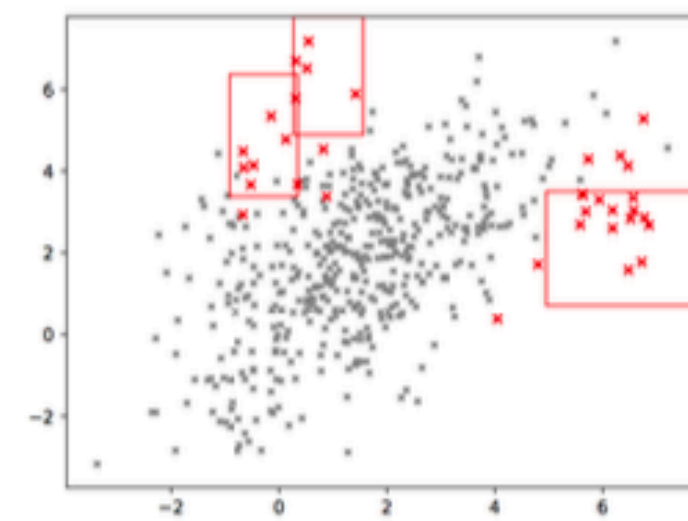
1. Initialize isolation forest, set equal  $w_i$  to each iTree
2. Ask the forest for the outlier with the largest score
3. Ask an expert to classify the object as normal or anomaly
4. If anomaly, go to step 2 and ask next outlier
5. If normal, update  $\{w_i\}$  to give lower influence to wrong detectors, go to step 2



(a) Baseline

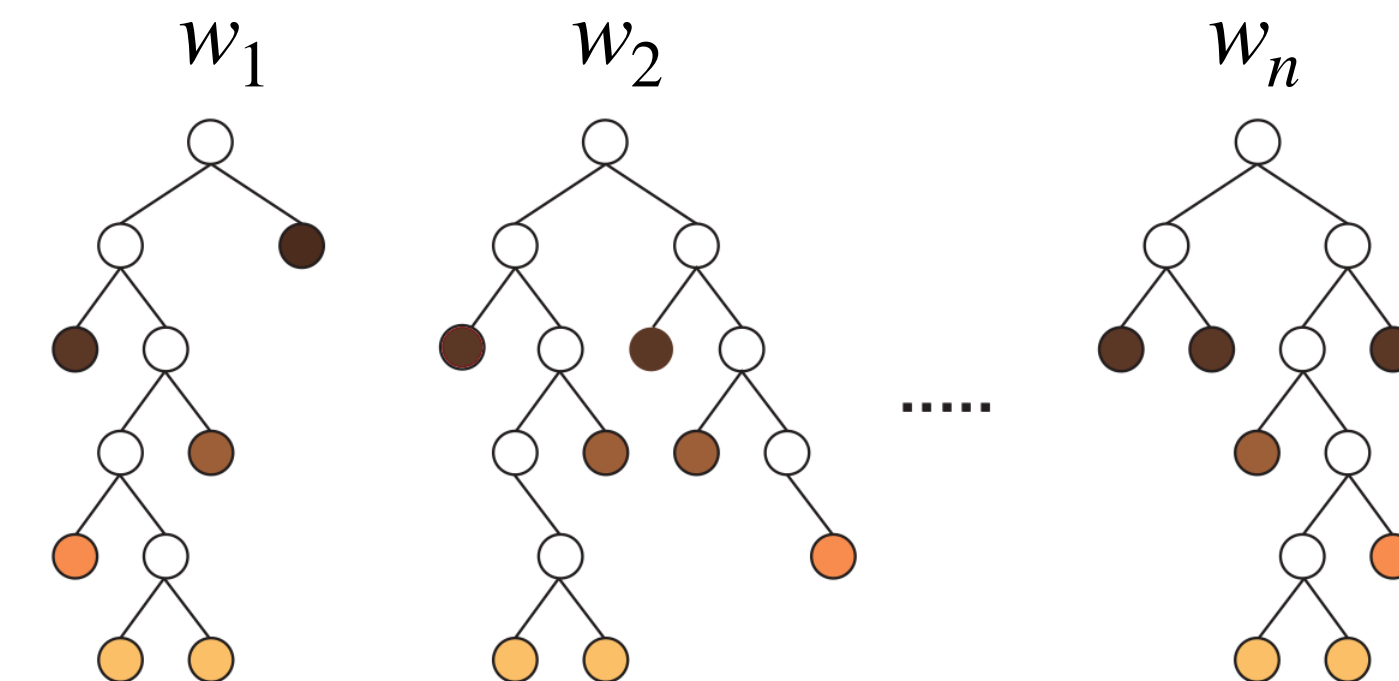


(b) AAD



(c) Descriptions

arXiv:1809.06477



There are other algorithms to solve this problem, we are developing a (better) alternative (Korolev+, in prep)

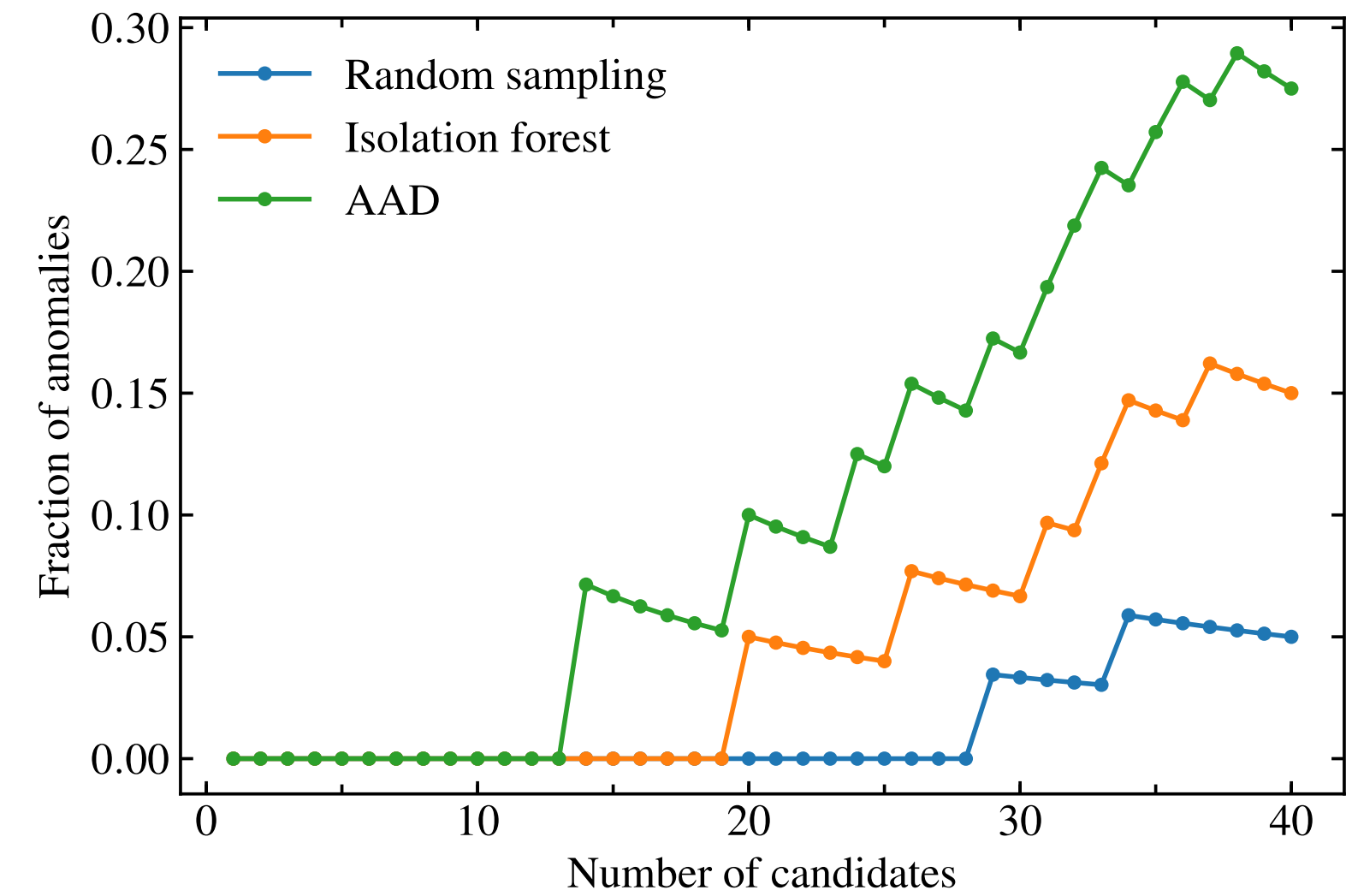
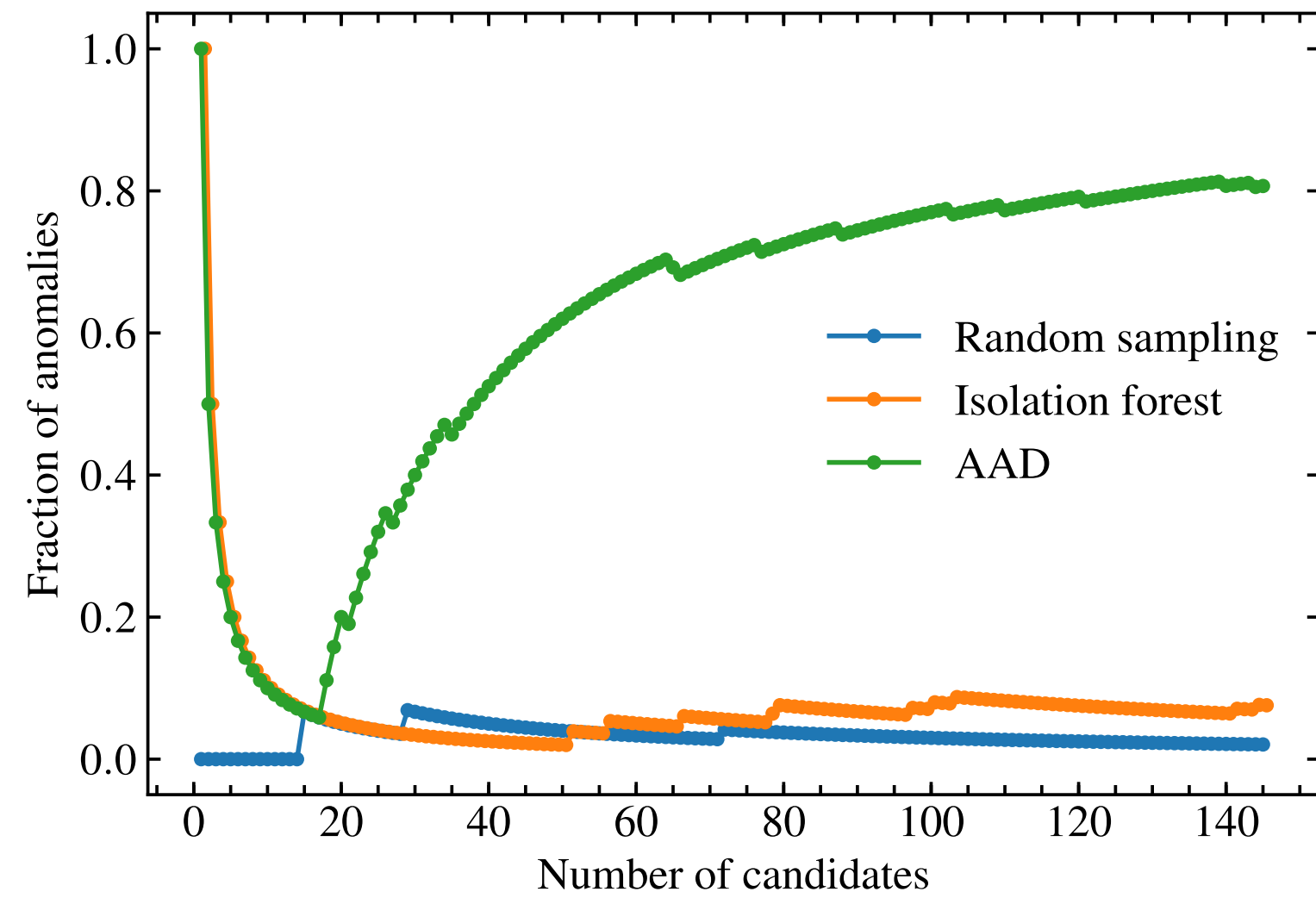
# Case: PLAsTiCC & OSC

arXiv:1909.13260



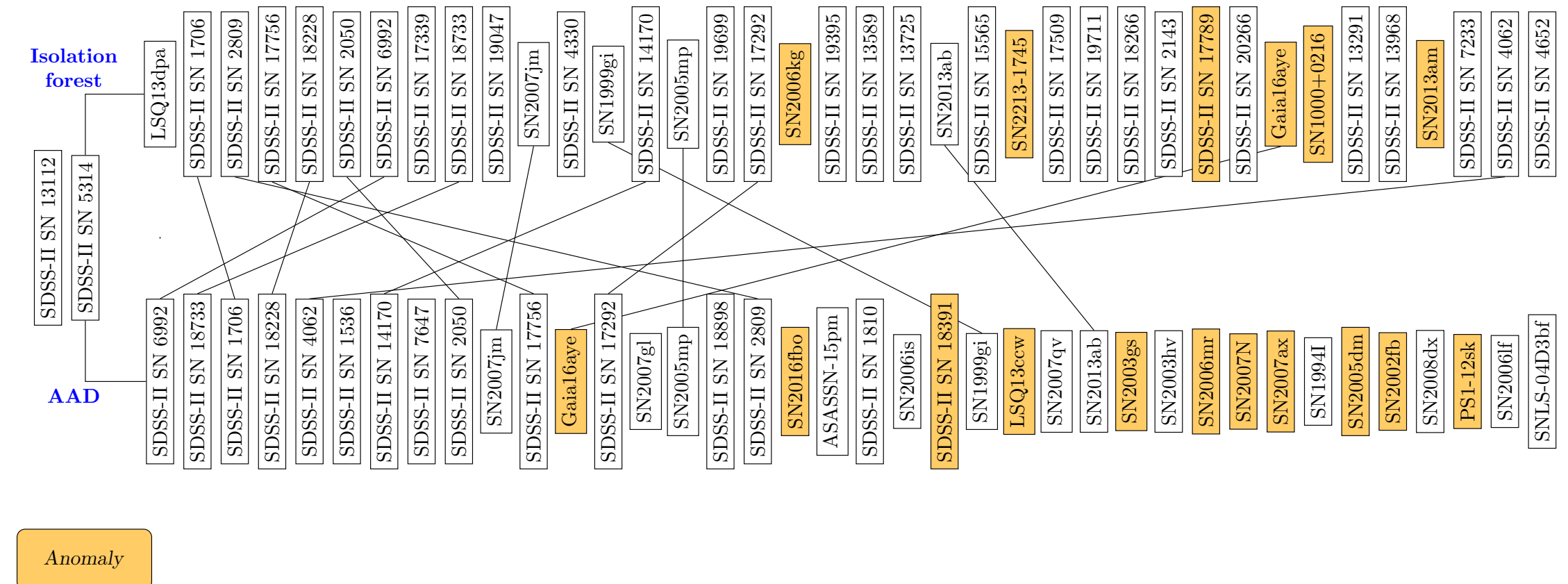
OSC

## PLAsTiCC (LSST sims)



Lessons learnt:

- AAD works
- Anomaly definition (expert bias) matters
- Real data is much harder!



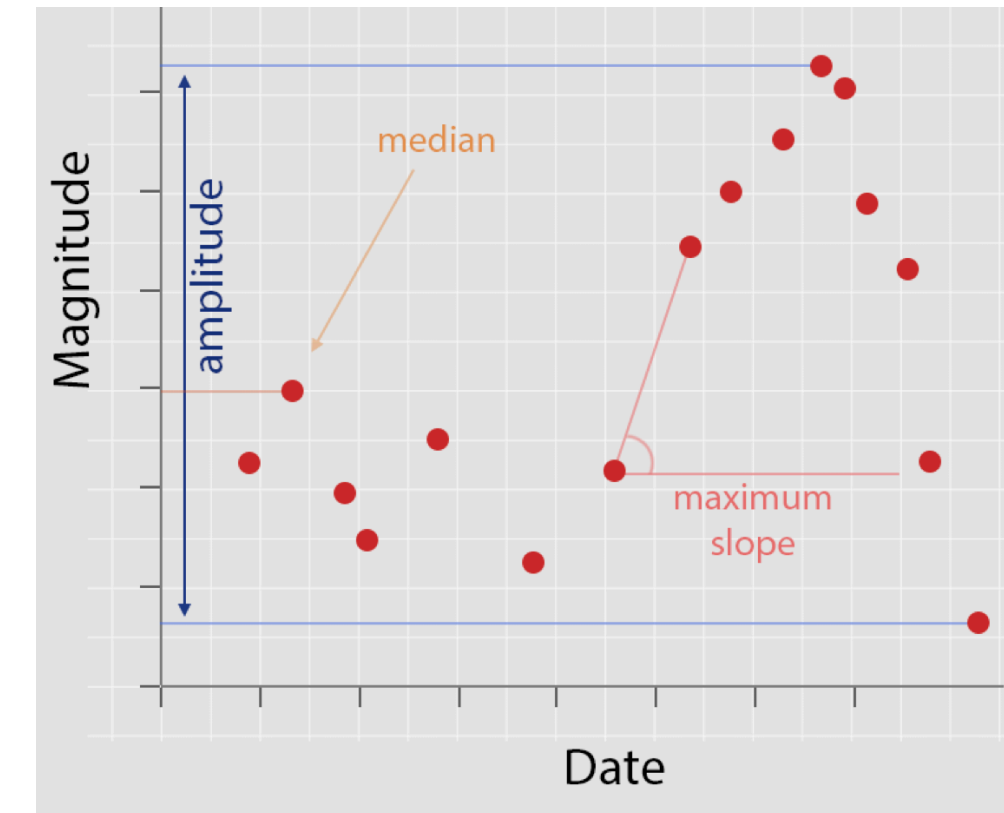




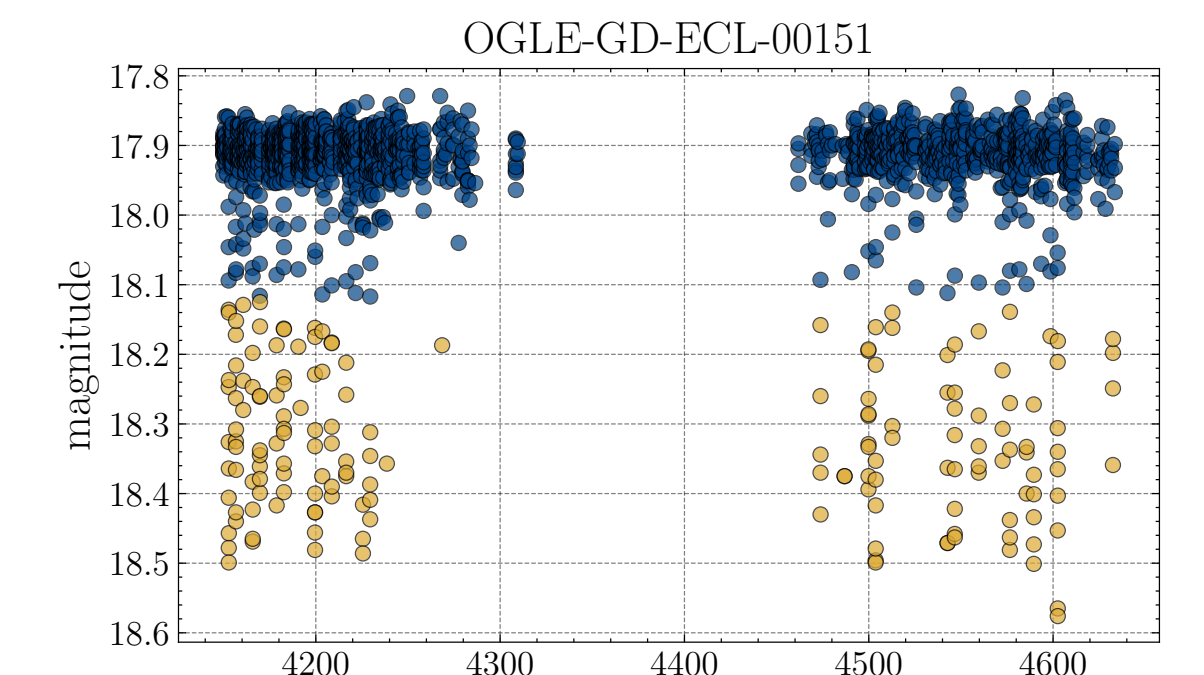
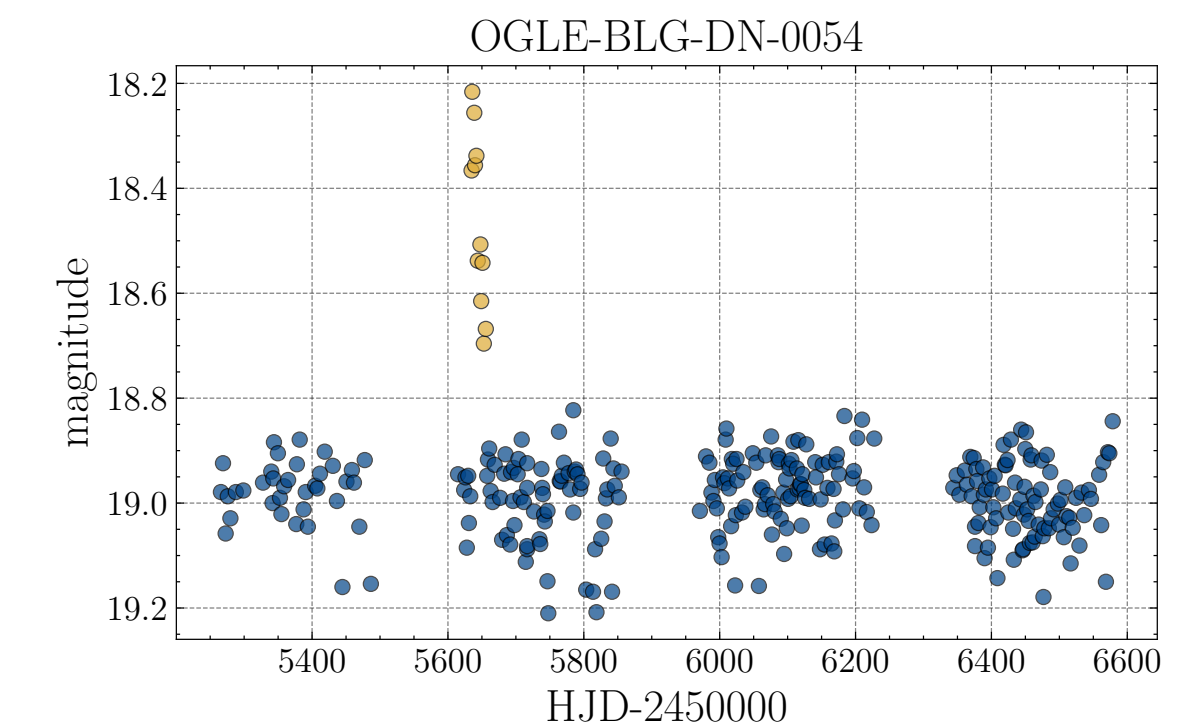
# By-product: light-curve feature extractor

<https://github.com/light-curve>

- **Performant** Rust/Python code: processing of  $\sim 10^6$  light-curves, Nobs  $\geq 100$ , takes few CPU hours
- Rich feature set
  - Magnitude statistics: mean-, median-, momentum- quartile-based
  - Shape-based: Stetson (1996) K,  $\eta^e$  (Kim+ 2014)
  - "Fast" Lomb-Scargle periodogram peaks and other derivatives
  - Parametric fits: linear, SN-like functions (Bazin+ 2009, Villar+ 2019)
  - **New Otsu-split extractor**: powerful features to classify recurrent outbursts, eclipsing binaries, etc (Lavrukhina & Malanchev in prep.)
- Hundreds of unit tests, packages for Linux and Intel Macs
- Serves **three ZTF/LSST brokers**: Ampel, Antares, Fink
- `python3 -m pip install light-curve`



Anastasia Lavrukhina





# By-product: SNAD ZTF viewer



<https://ztf.snad.space>



ztf.snad.space/dr8/view/633207400004730

login

## SNAD ZTF DR8 object viewer

OID

Coordinates  radius (arcsec)

SNAD101 — 633207400004730

mag

mjd - 58000

filter, oid

- zr, 633207400004730
- zr, 1629213400006020
- zr, 1630216300019032
- zg, 633107400013284
- zg, 1629113400014635
- zg, 1630116300008554
- zi, 633307400006257

Download [PNG](#), [PDF](#), [CSV](#)

"Short" light curve: 58194.0 ≤ MJD ≤ 58972.0

Full light curve  Folded light curve

Closest Antares object, diff-photometry  Closest Pan-STARRS object, apparent

Magnitude  Flux  diff Magnitude  diff Flux

Summary

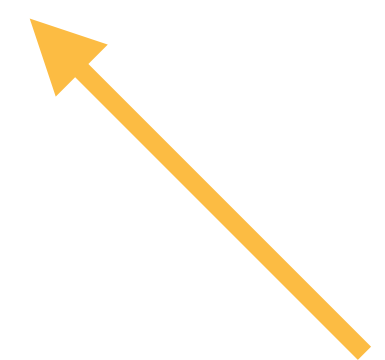
Open in JS9 [Download FITS Product directory](#)

Self-matched ZTF light-curve



ZTF science image for any detection

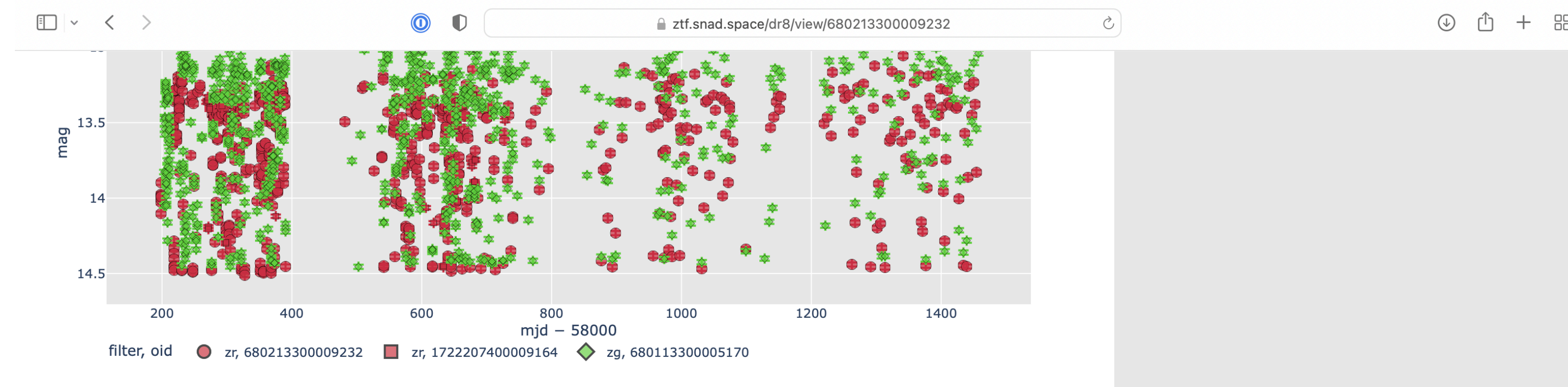
Aladin





# By-product: SNAD ZTF viewer

<https://ztf.snad.space>



Download [PNG](#), [PDF](#), [CSV](#)

- "Short" light curve:  $58194.0 \leq \text{MJD} \leq 58972.0$
- Full light curve  Folded light curve
- Closest Antares object, diff-photometry  Closest Pan-STARRS object, apparent
- Magnitude  Flux  diff Magnitude  diff Flux

## Summary

**Name:** ZTF18aabpzic (0.266" [ALerCe](#)), ZTF18aabpzic (0.353" [Fink](#)), J254.4575+35.3423 (0.124" [ATLAS](#)), 1338822021487330304 (0.115" [Gaia EDR3 Distances](#)), HZ Her (0.711" [GCVS](#)), PSO J254.4575+35.3423 (0.109" [Pan-STARRS DR2 Stacked](#)), V\* HZ Her (0.081" [Simbad](#)), 15037 (0.720" [VSX](#)), ZTFJ165749.81+352032.4 (0.124" [ZTF Periodic](#))

**Type:** LMXB (0.353" [Fink](#)), IRR (0.124" [ATLAS](#)), XPR+E (0.711" [GCVS](#)), LowMassXBin (0.081" [Simbad](#)), LMXB/XPR+E (0.720" [VSX](#)), EW (0.124" [ZTF Periodic](#))

**Period, days:** 1.700 ([periodogram S/N=78.620](#)), 1.700 (0.124" [ATLAS](#)), 1.700 (0.711" [GCVS](#)), 1.700 (0.081" [Simbad](#)), 34.875 (0.720" [VSX](#)), 3.400 (0.124" [ZTF Periodic](#))

**Distance:** 7.00 kpc (0.115" [Gaia EDR3 Distances](#)), 6.60 kpc (0.081" [Simbad](#))

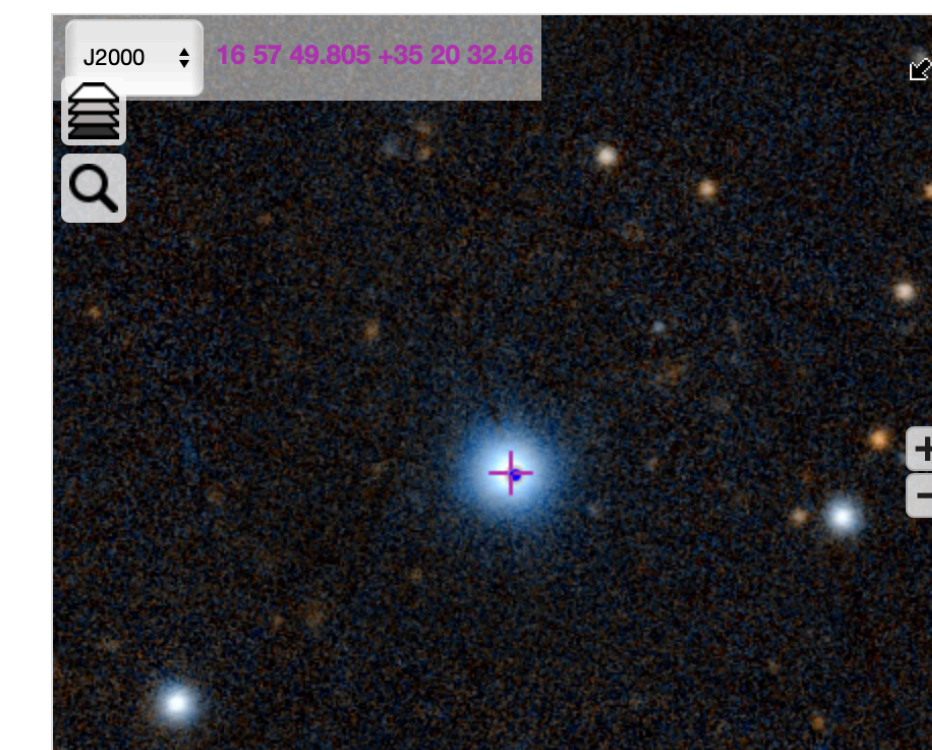
**Average mag (including neighbourhood):** zg 13.55, zr 13.68, (zg-zr) -0.13

**Extinction:** SFD  $E(B-V) = 0.01$ , Bayestar & Gaia EDR distance  $A_g = 0.07$   $A_r = 0.05$   $A_i = 0.03$

**Search in brokers:** [ALeRCE](#), [Antares](#), [Fink](#), [MARS](#)

**Coordinates:** Eq 254.45752 35.34235, Gal 58.149 37.5231

## Aladin



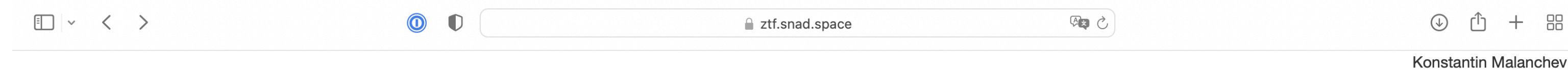
Name, type, period, distance & extension from other catalogs and our periodogram

# By-product: SNAD ZTF viewer



<https://ztf.snad.space>

Home page



 **SNAD ZTF DR8 object viewer**

OID    
Coordinates  radius (sec)

For example see the page for [SNAD101](#)

Welcome to SNAD ZTF object viewer!

This is a tool developed by the [SNAD team](#) in order to enable quick expert investigation of objects within the public [Zwicky Transient Facility \(ZTF\)](#) data releases.

It was developed as part of the [3rd SNAD Workshop](#), held remotely in July, 2020.

The viewer allows visualization of raw and folded light curves and metadata, as well as cross-match information with the [the General Catalog of Variable Stars](#), [the International Variable Stars Index](#), [the ATLAS Catalog of Variable Stars](#), [the ZTF Catalog of Periodic Variable Stars](#), [the Transient Name Server](#), [the Open Astronomy Catalogs](#), [the OGLE III Catalog of Variable Stars](#), [the Simbad Astronomical Data Base](#), [Gaia EDR3 distances \(Bailer-Jones+, 2021\)](#), [Vizier](#).

The viewer is also available for [ZTF DR2](#), [ZTF DR3](#), [ZTF DR4](#)

© 2022 [SNAD](#). Version [2022.5.2](#). Developed by Konstantin Malanchev, based on [the ZTF Caltech data](#). See the source code [on GitHub](#).  
If you use this web-site in your research, please cite [this paper](#) as well as all relevant data source papers.

ZTF object ID / SNAD ID

Eq coordinates / common name





# By-product: SNAD ZTF viewer

<https://ztf.snad.space>



ztf.snad.space/dr8/search/hz%20her/1

Konstantin Malanchev

## SNAD ZTF DR8 object viewer

OID

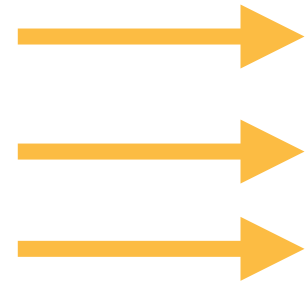
Coordinates  radius (arcsec)

Objects inside cone (254.45755 deg, 35.34236 deg),  $r = 1.0''$

OID	separation, arcsec	filter	Number of "good" observations	Duration, days
<a href="#">1722207400009164</a>	0.069	zr	43	459.824
<a href="#">680113300005170</a>	0.078	zg	734	1253.823
<a href="#">680213300009232</a>	0.082	zr	686	1255.902

© 2022 SNAD. Version 2022.5.2. Developed by Konstantin Malanchev, based on [the ZTF Caltech data](#). See the source code [on GitHub](#).  
If you use this web-site in your research, please cite [this paper](#) as well as all relevant data source papers.

Same source,  
different OIDs



# By-product: SNAD ZTF viewer

<https://ztf.snad.space>



SNAD experts  
tagged  
>2000 objects,  
>70 are submitted  
to the TNS!

Browser address bar: ztf.snad.space/akb

Konstantin Malanchev

### SNAD ZTF DR8 object viewer

OID:

Coordinates:  radius (arcsec)

### Anomaly knowledge base

OID	Tags	Description	Changed by	Changed at
<a href="#">633207400004730</a>	SN, uncertain	SNAD101	maria	2021-08-02T07:46:53.429000+00:00
<a href="#">633216300024691</a>	SN, uncertain	SNAD102	maria	2021-08-02T07:47:54.227000+00:00
<a href="#">634108100006647</a>	AGN, SN, uncertain	SNAD158	maria	2021-10-21T22:22:11.362000+00:00
<a href="#">643105300009229</a>	AGN, SN, uncertain	SNAD153	maria	2021-10-21T21:39:27.291000+00:00
<a href="#">676212400013135</a>	SN, uncertain	SNAD122	maria	2021-08-02T07:52:47.557000+00:00
<a href="#">679108100003227</a>	SN Ia, uncertain, non-catalogued	photo-z of host: 0.303 +/- 0.116... Possible absolute mag between -20.6 and -22.6. SLSN? Too bright for SN Ia... SNAD150	patrick	2021-10-21T14:19:32.575000+00:00
<a href="#">680109100003419</a>	SN	SNAD168, PCA+ k-D tree	maria	2021-11-12T20:32:32.823000+00:00
<a href="#">682102200004200</a>	SN, uncertain	SNAD176	maria	2022-03-03T14:49:49.398000+00:00
<a href="#">682209200018910</a>	SN, uncertain	SNAD143	maria	2021-08-02T09:48:05.990000+00:00
<a href="#">684215200016923</a>	SN, uncertain, non-catalogued	SNAD157	maria	2021-10-21T22:17:58.390000+00:00
<a href="#">692106300027877</a>	SN, uncertain	SNAD174	novinskaya	2022-02-28T15:02:40.405000+00:00
<a href="#">718205300006523</a>	AGN, uncertain, non-catalogued	SNAD155	maria	2021-10-22T09:30:51.131000+00:00
<a href="#">719202100004008</a>	AGN, SN, uncertain	SNAD154	maria	2021-10-21T22:06:23.389000+00:00
<a href="#">720209400014960</a>	SN, uncertain	SNAD123	maria	2021-08-02T10:06:24.720000+00:00
<a href="#">721210100012349</a>	SN, uncertain	SNAD129	maria	2021-08-02T10:47:48.325000+00:00



# Conclusion



- Using real data from the very beginning
- Astronomical experts are queens: their opinion matters from the start of the algorithm construction to the last stage
- Developing new tools — and sharing them with the community
- Recent and ongoing projects:
  - Developing new active anomaly detection algorithm for new features, better computation and detection performance (Korolev+ in prep., ask me about it!)
  - Using AAD for classification, listen **talk by Emille Ishida** about SNe
  - Mining transients with k-D tree, see **poster by Patrick Aleo** (presented by me)