# Active Anomaly Detection Tutorial

**Konstantin Malanchev malanchev@cmu.edu**

**LINCC Frameworks / Carnegie Mellon University**

**AISSAI Anomaly Detection Workshop, March 2024, Clermont-Ferrand**

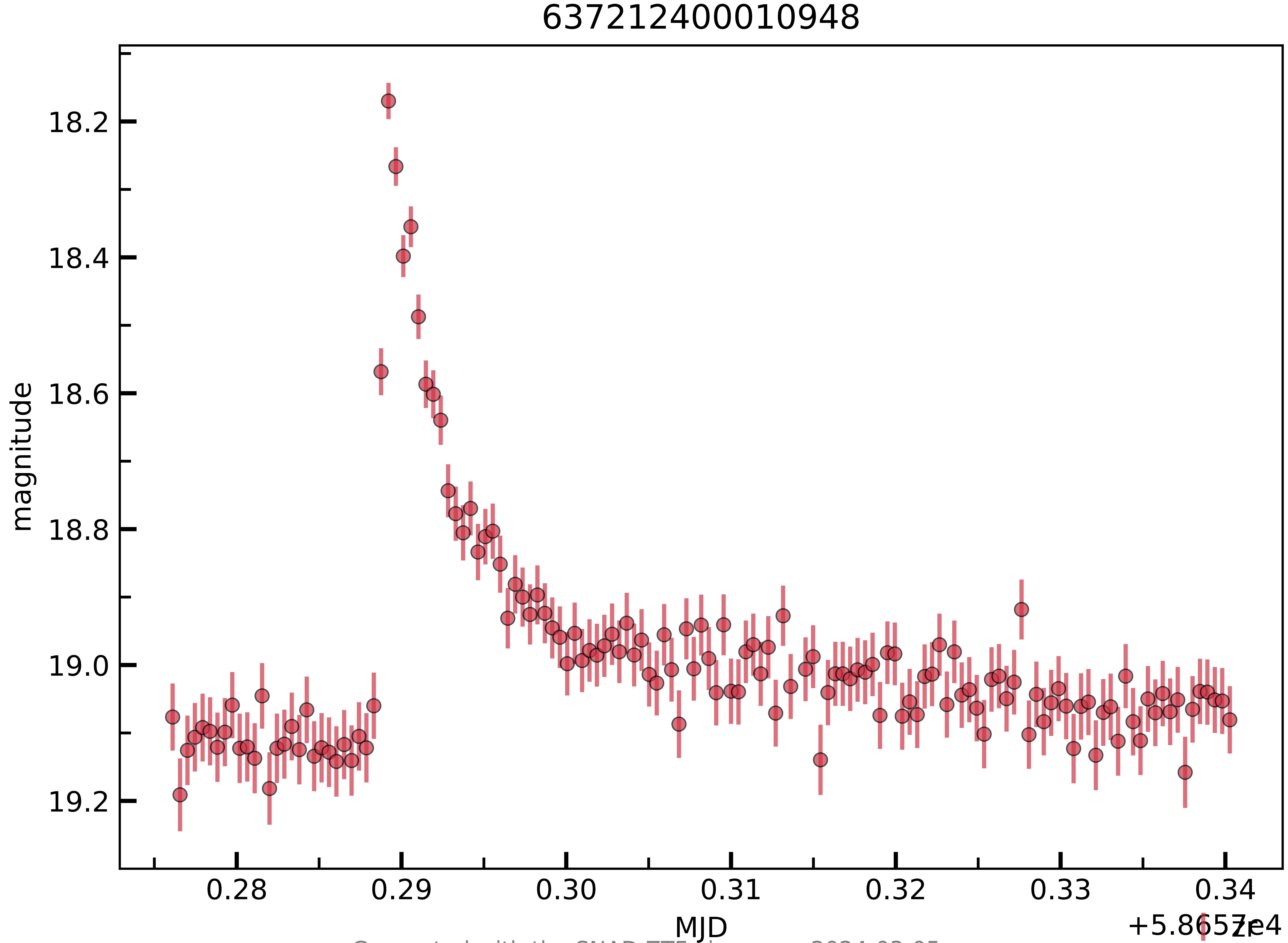
# Before we start

## What we will need

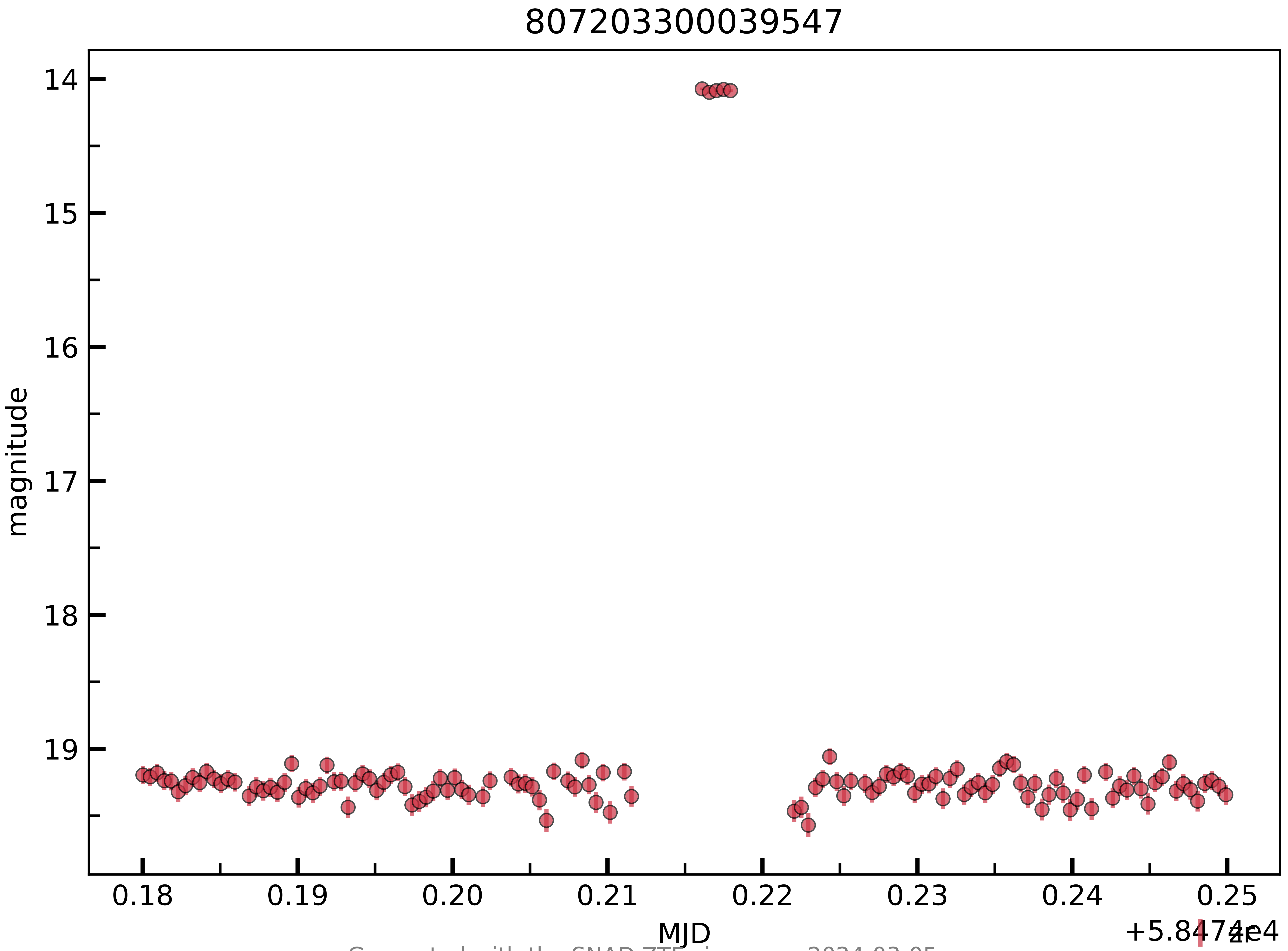
- Google account for Google Colab, OR
- Python 3.9-3.11 and Jupyter with virtual environment

Check notebook links on the workshop website, QRs are following

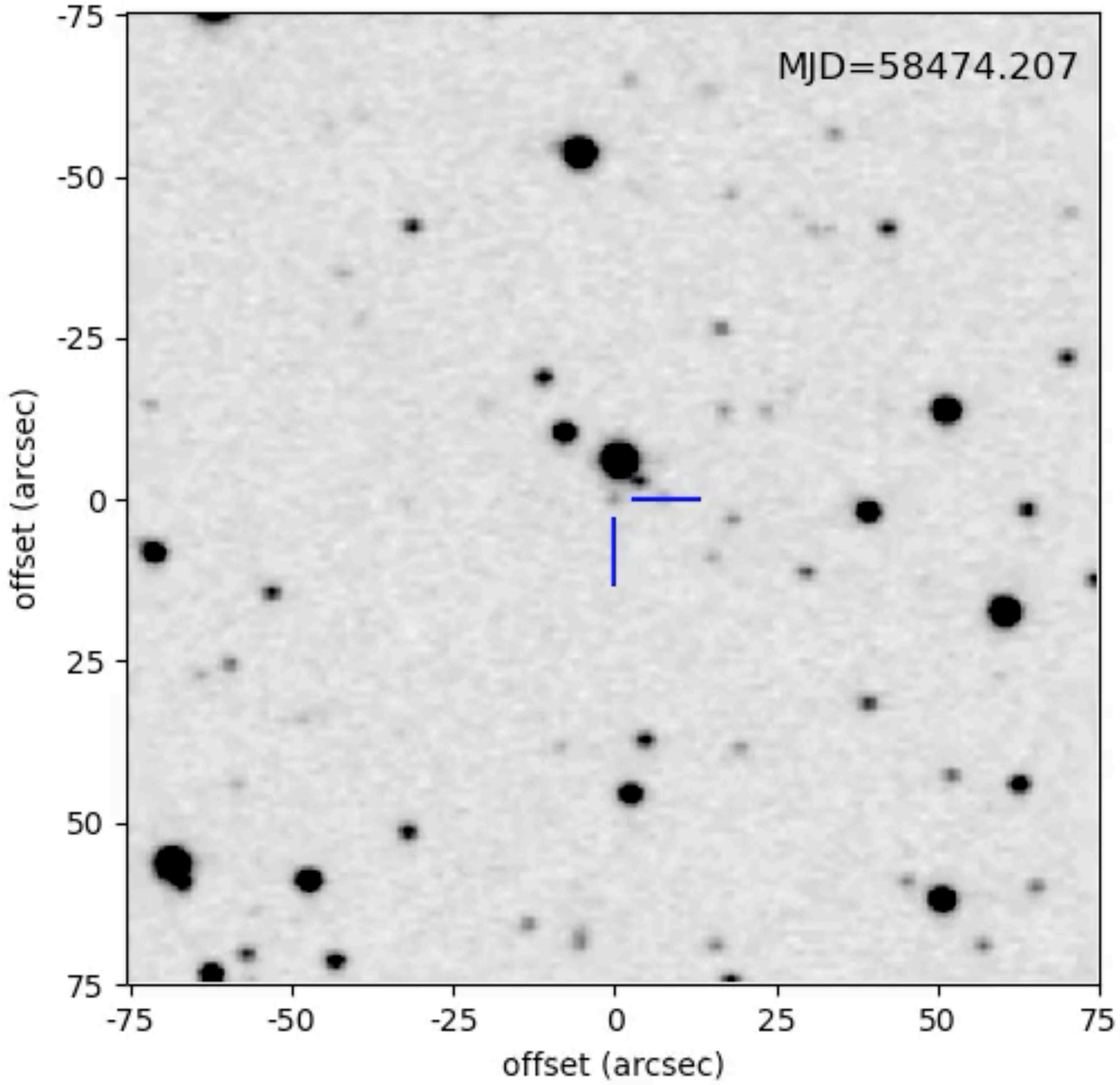
# Why do we want to go active?



Generated with the SNAD ZTF viewer on 2024-03-05



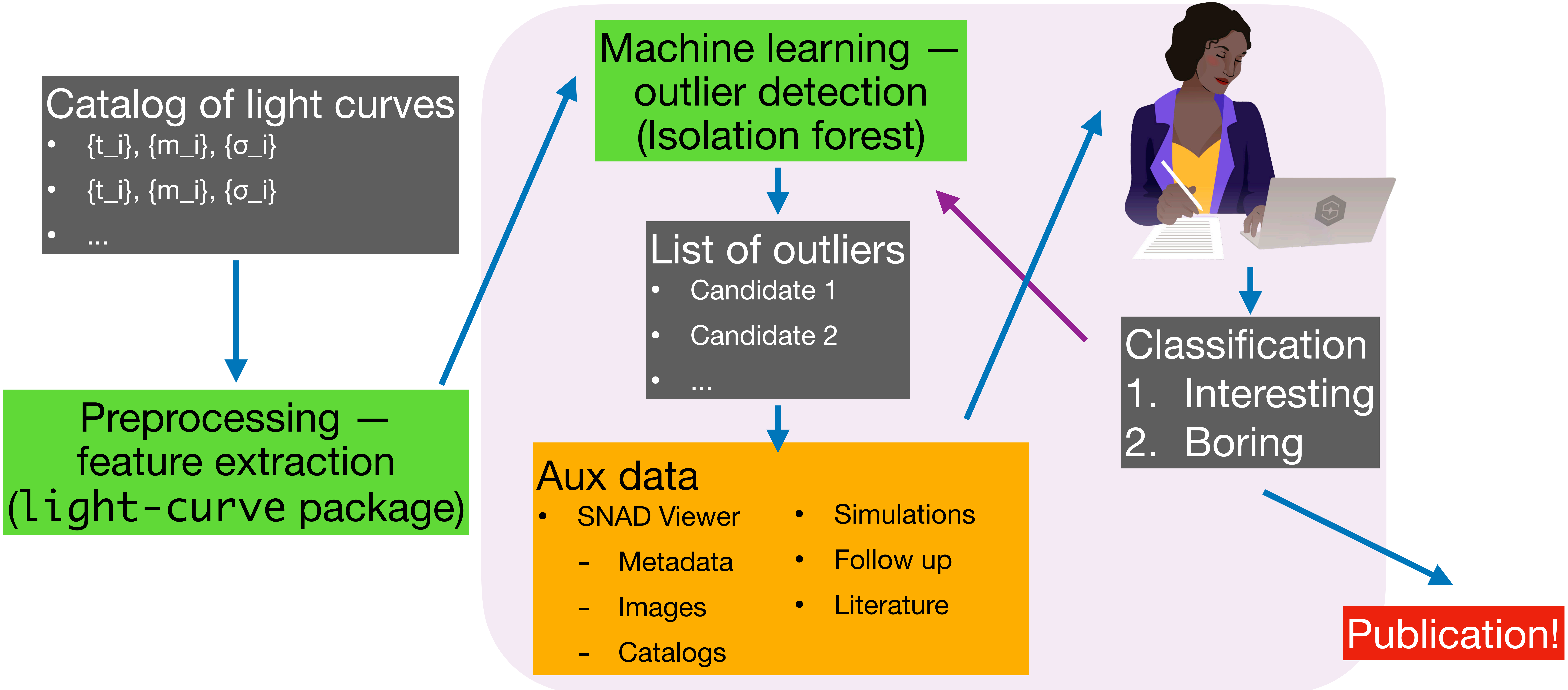
Generated with the SNAD ZTF viewer on 2024-03-05


MJD=58474.207
offset (arcsec)
-75
-50
-25
0
25
50
75
offset (arcsec)
-75
-50
-25
0
25
50
75

# Pipeline

# Anomaly Detection for Light Curves

## SNAD papers: Pruzhinskaya+19,22; Ishida+19; KM+21; Aleo+22.



Catalog of light curves

- {t_i}, {m_i}, {σ_i}
- {t_i}, {m_i}, {σ_i}
- …

Preprocessing — feature extraction (`light-curve` package)

Machine learning — outlier detection (Isolation forest)

List of outliers

- Candidate 1
- Candidate 2
- …

Aux data

- SNAD Viewer
  - Metadata
  - Images
  - Catalogs
- Simulations
- Follow up
- Literature

Classification

1. Interesting
2. Boring

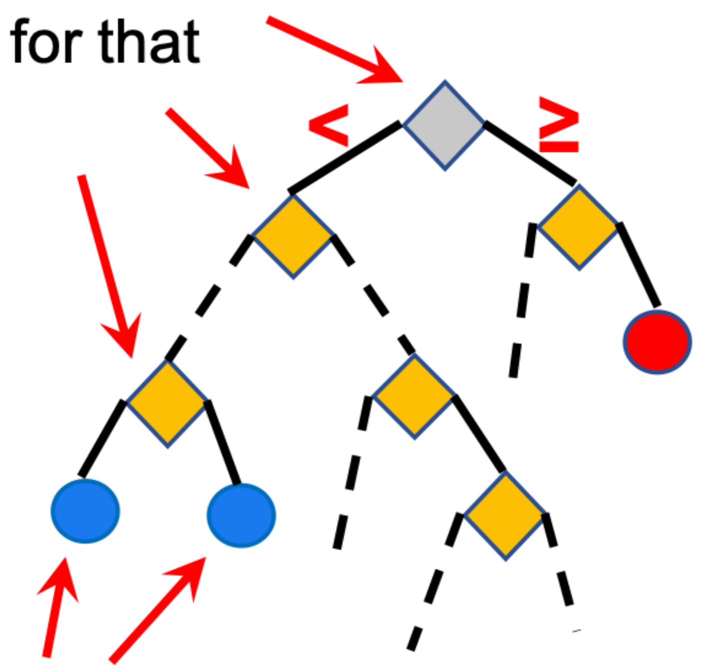
Publication!

# Isolation Forest and Pineforest

# Isolation Forest

## iTree

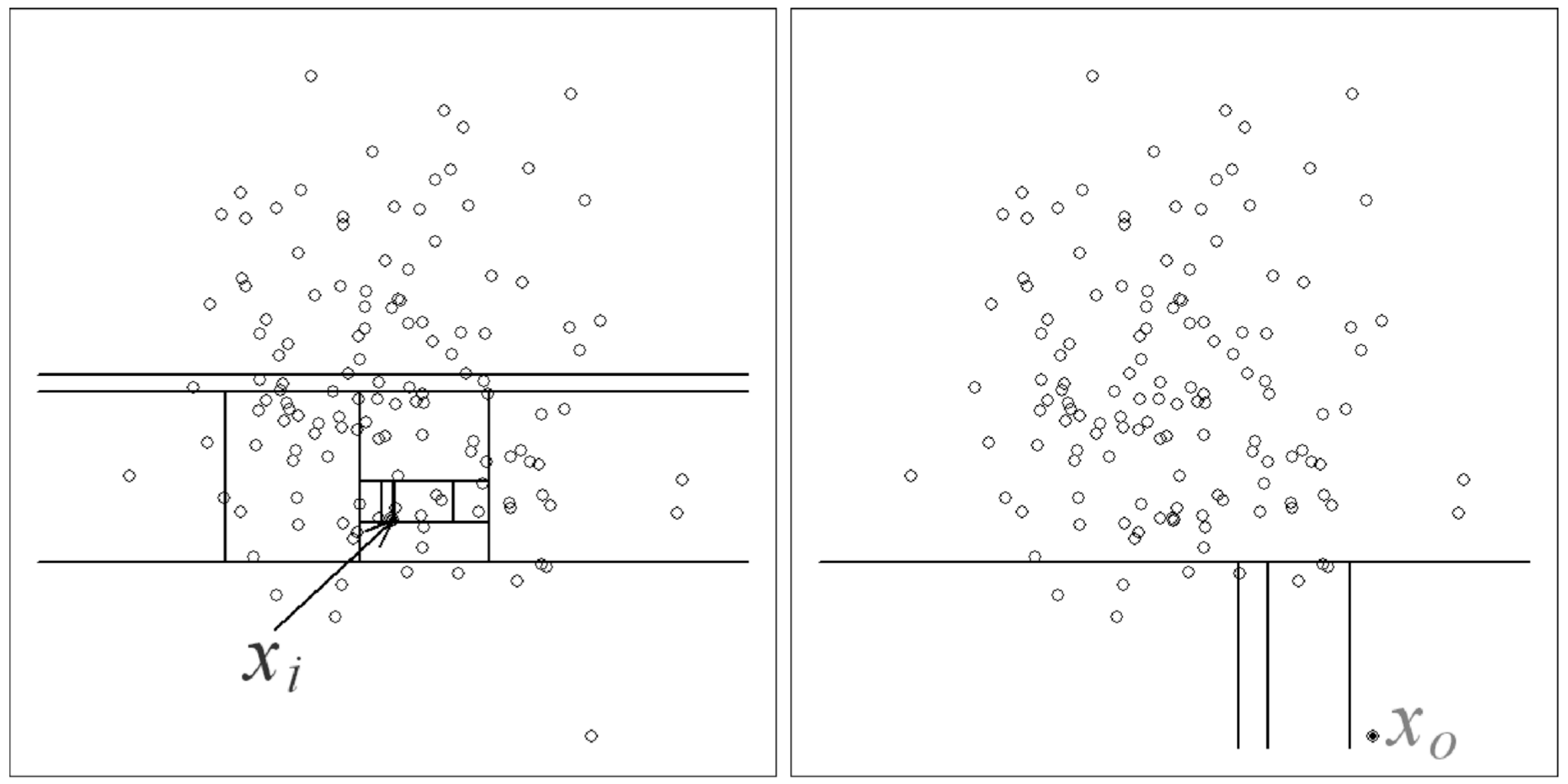
Select a random feature at each node, and a random split point for that feature



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

Leaf instance

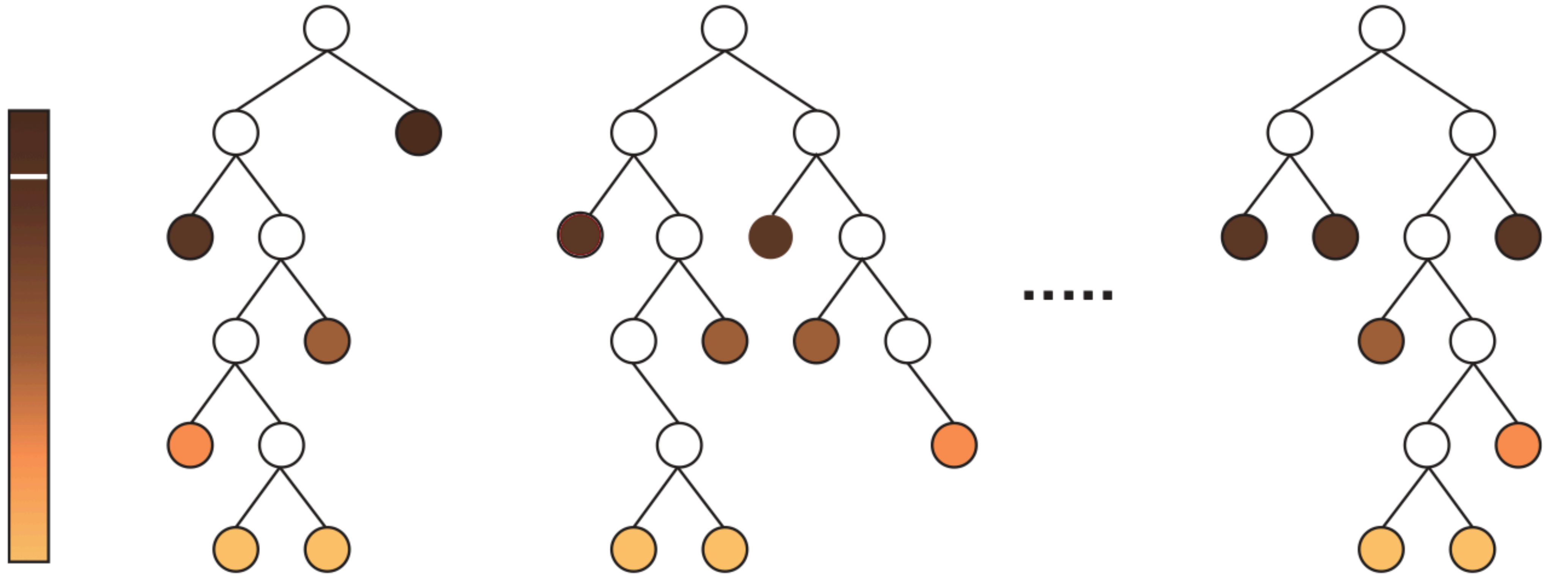
arXiv:1708.0944



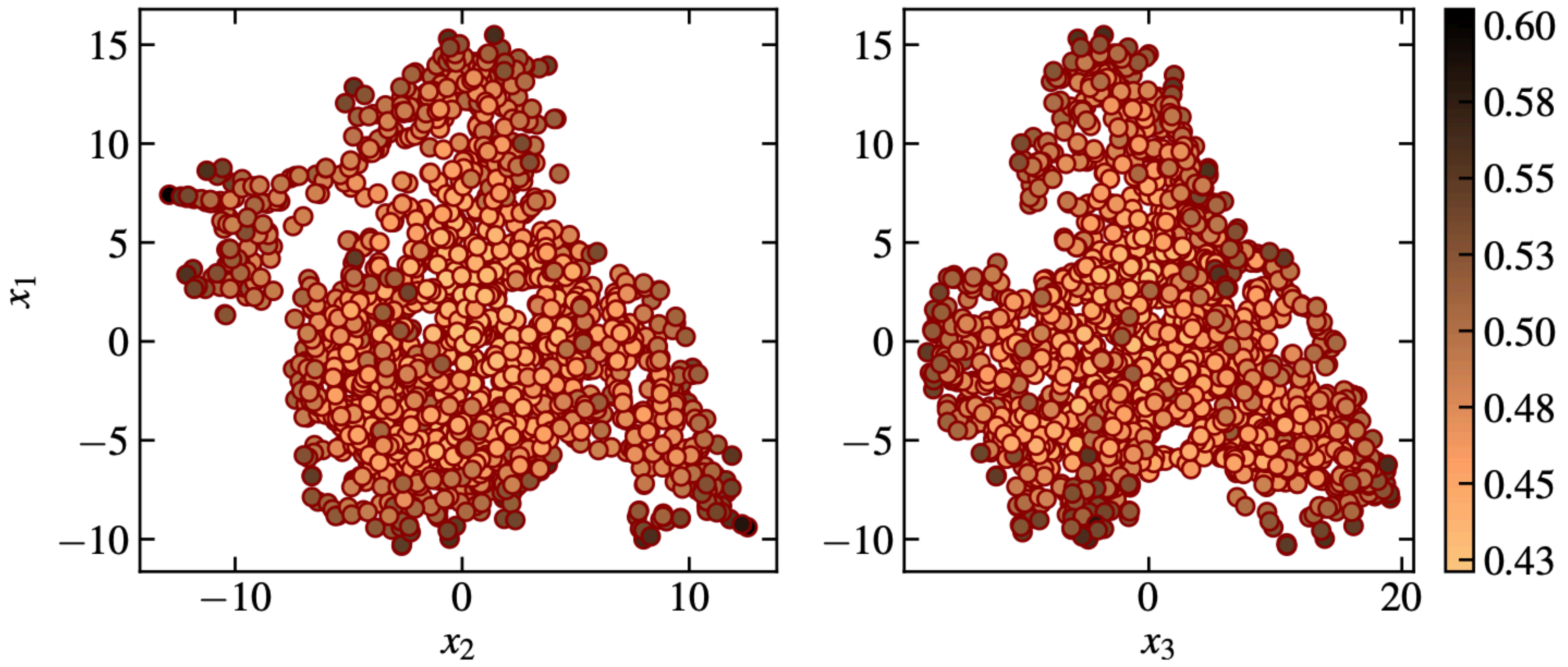
(a) Isolating $x_i$

(b) Isolating $x_o$

Liu+ 2008, Liu+ 2012

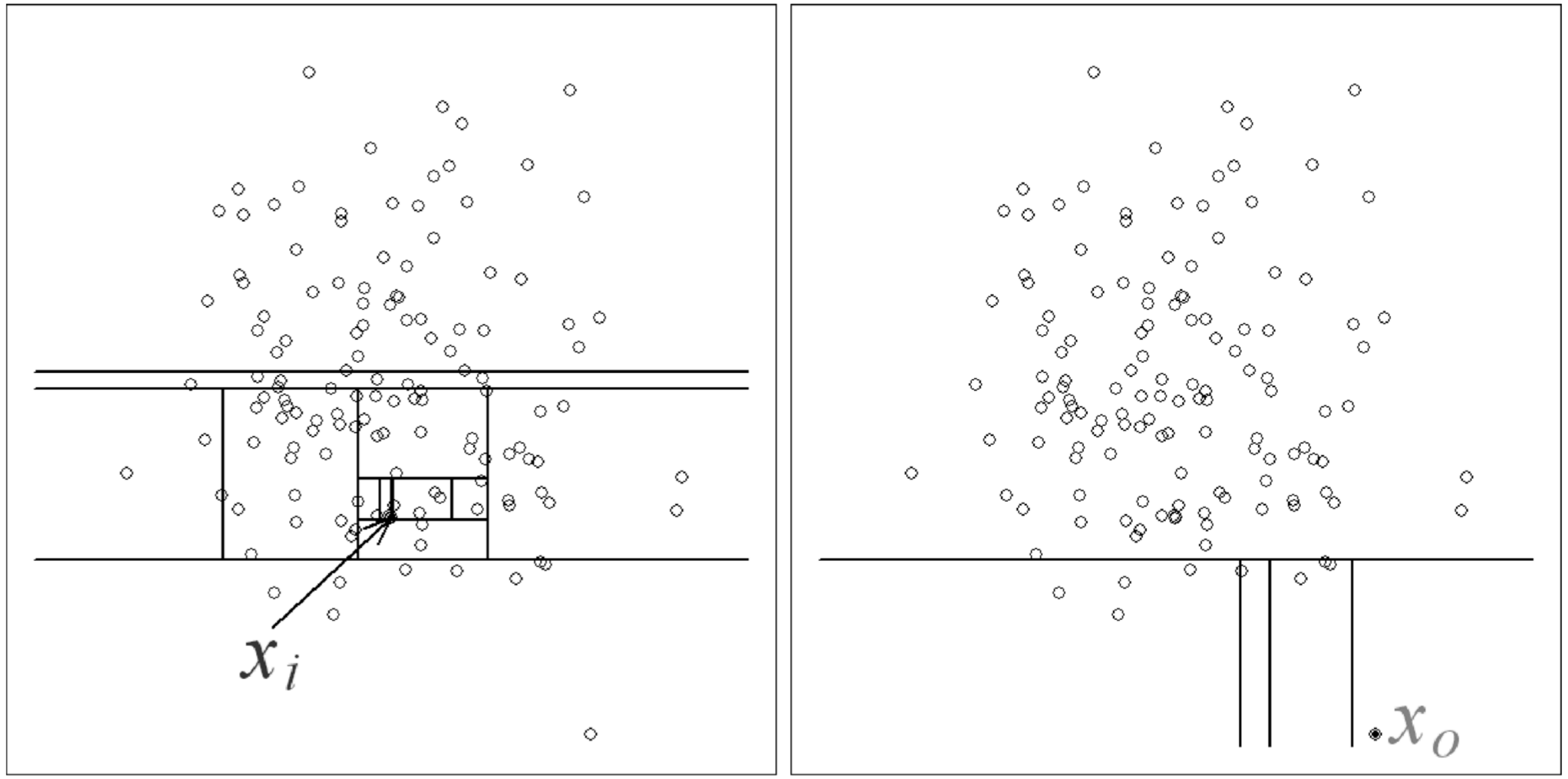


Darker is more anomalous



arXiv:1905.11516

# Isolation Tree



(a) Isolating $x_i$ (b) Isolating $x_o$

$$c(\psi) = \begin{cases} 2H(\psi - 1) - 2(\psi - 1)/\psi & \text{for } \psi > 2, \\ 1 & \text{for } \psi = 2, \\ 0 & \text{otherwise,} \end{cases}$$

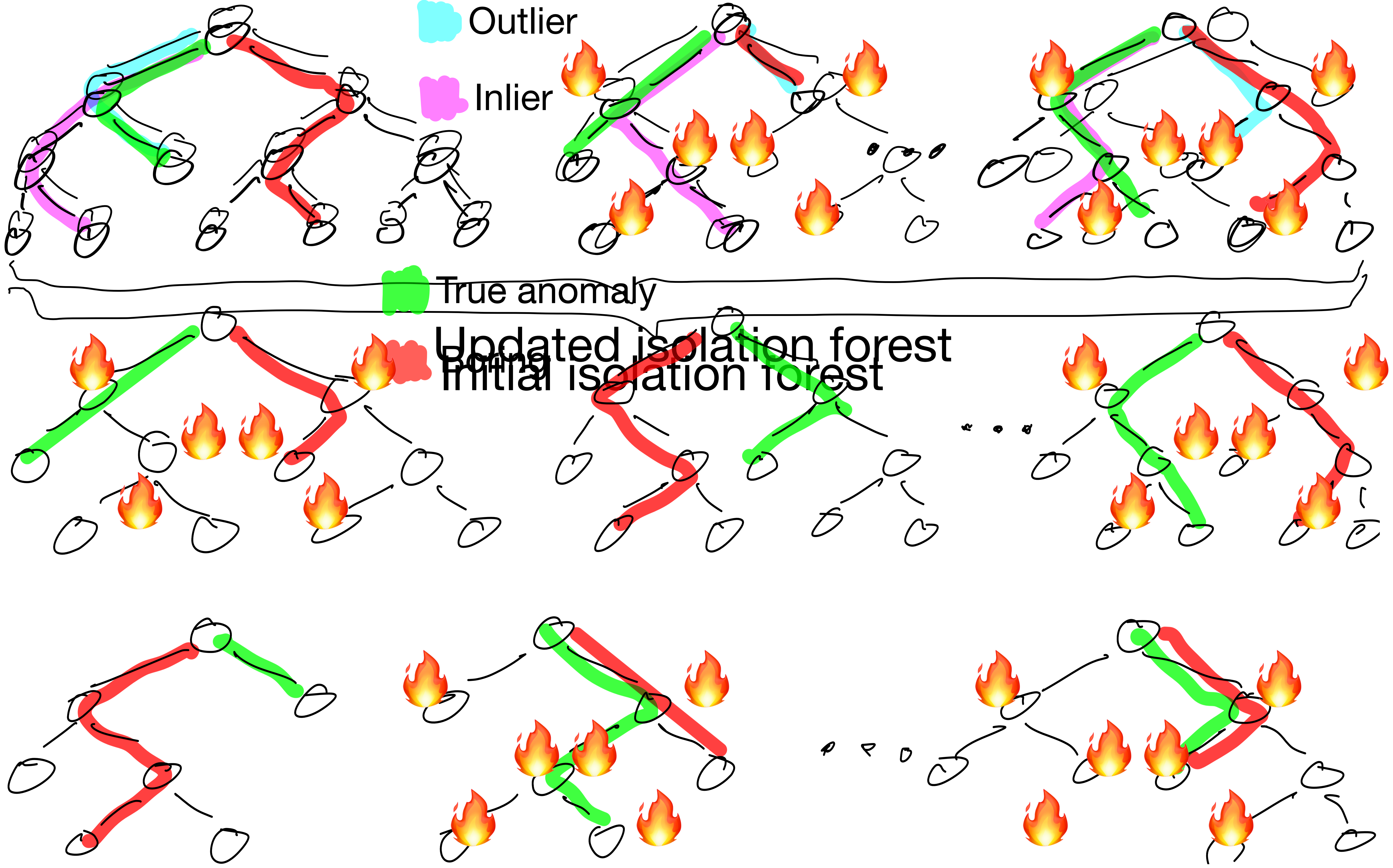
$$s(x, \psi) = 2^{-\frac{E(h(x))}{c(\psi)}},$$

**Liu et al 2008, Liu et al 2012**

# Active Anomaly Detection with Pineforest

## Based on Isolation Forest (Liu+08) and inspired by AAD (Das+17)

1. Build an isolation forest
2. Select the best outlier from the unlabeled data
3. Ask the expert to classify
4. Build more trees
5. Rank the trees with labeled data
6. Select the best trees and prune the rest
7. Go to 2.



# Tools we are going to use

# Coniferest Package

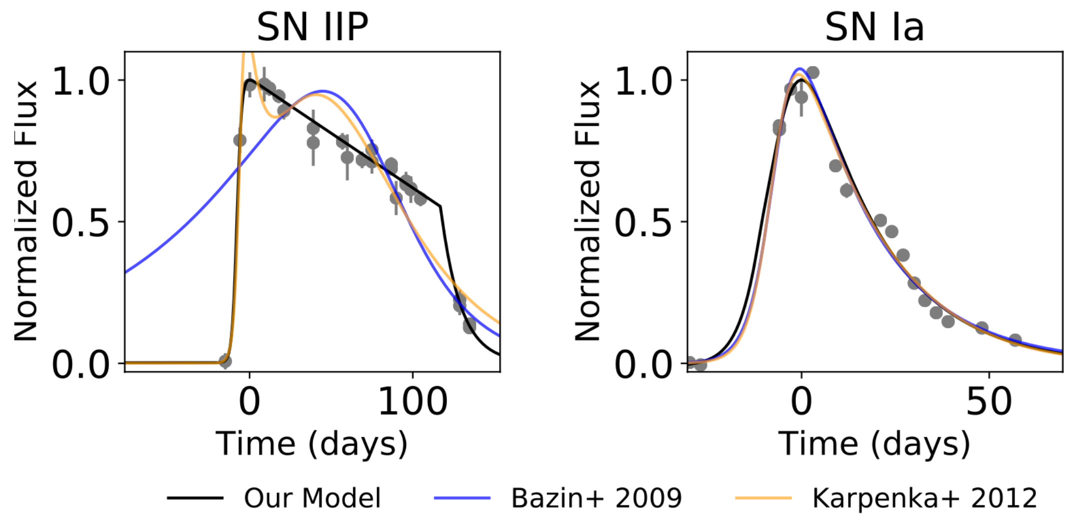
**Docs: https://coniferest.snad.space**
**Code: https://github.com/snad-space/coniferest**

- Performant re-implementation of scikit-learn's IsolationForest
- Two "active" algorithms atop of it: AAD (Das+2017) and Pineforest
- `Session` class which handles interactive pipeline

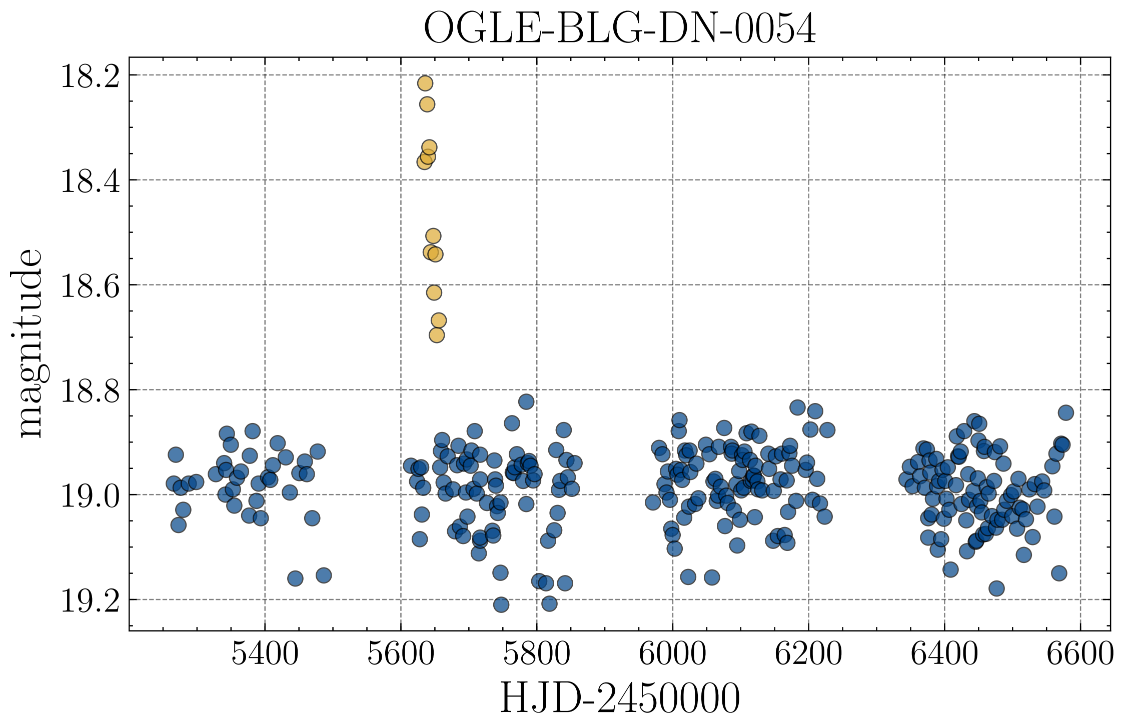
# Astronomical data: time-series features

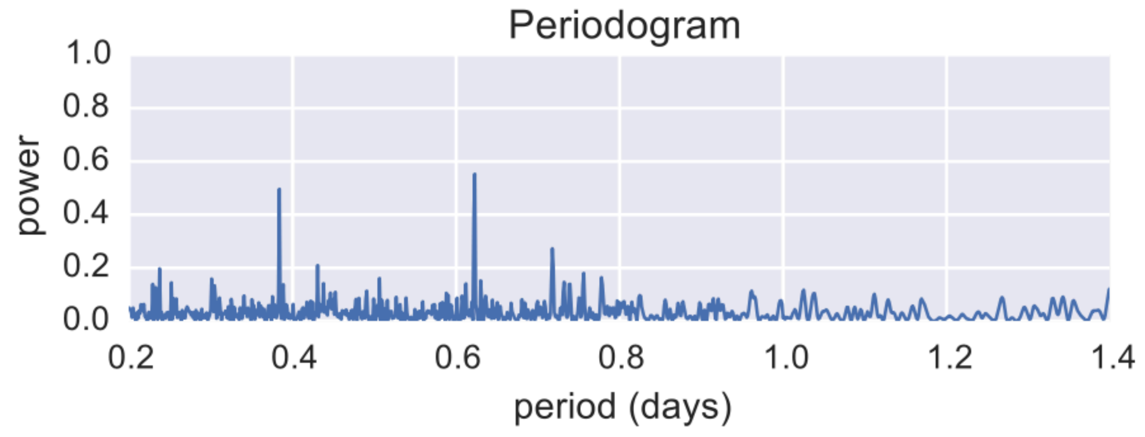
`light-curve` package in Python and Rust, `https://github.com/light-curve`

- Rich feature set
  - Magnitude statistics: mean-, median-, momentum- quartile-bas
  - 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 Rainbow approach Russeil+2024**
  - New Otsu-split extractor: powerful features to classify recurren outbursts, eclipsing binaries, etc (Lavrukhina+2023)
- Hundreds of unit tests, pre-built wheels for Linux and macOS
- Serves **three ZTF/LSST brokers**: AMPEL, ANTARES, Fink
- `python3 -m pip install light-curve`

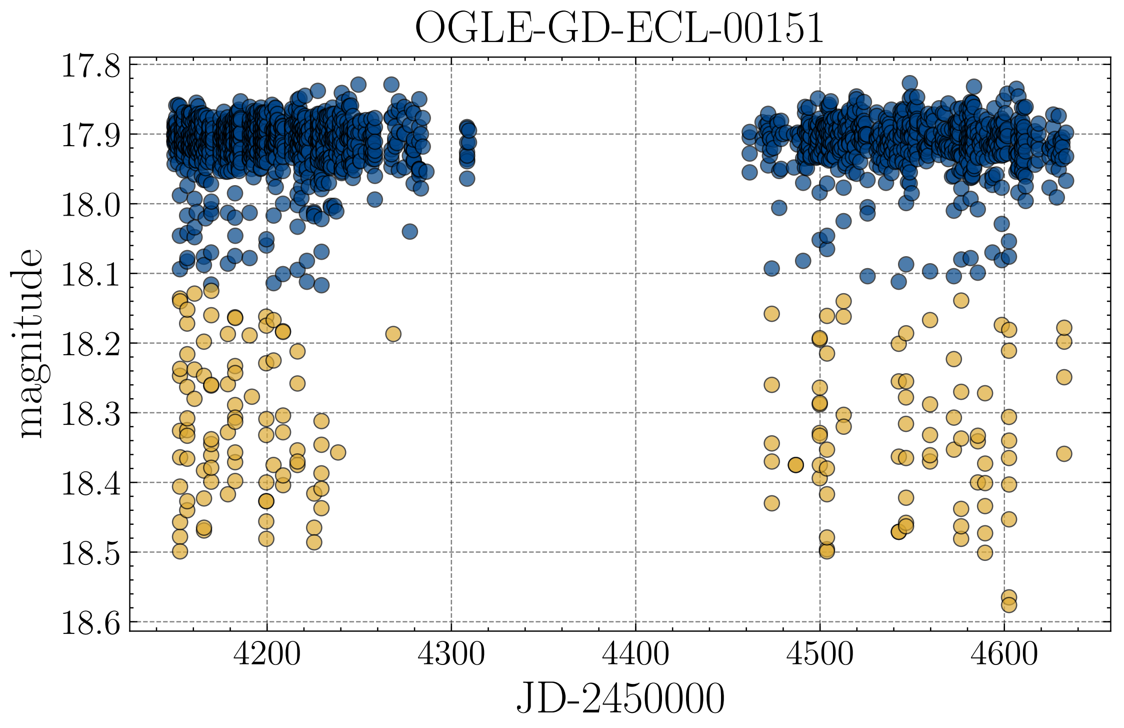


Villar+19, arXiv:1905.07422





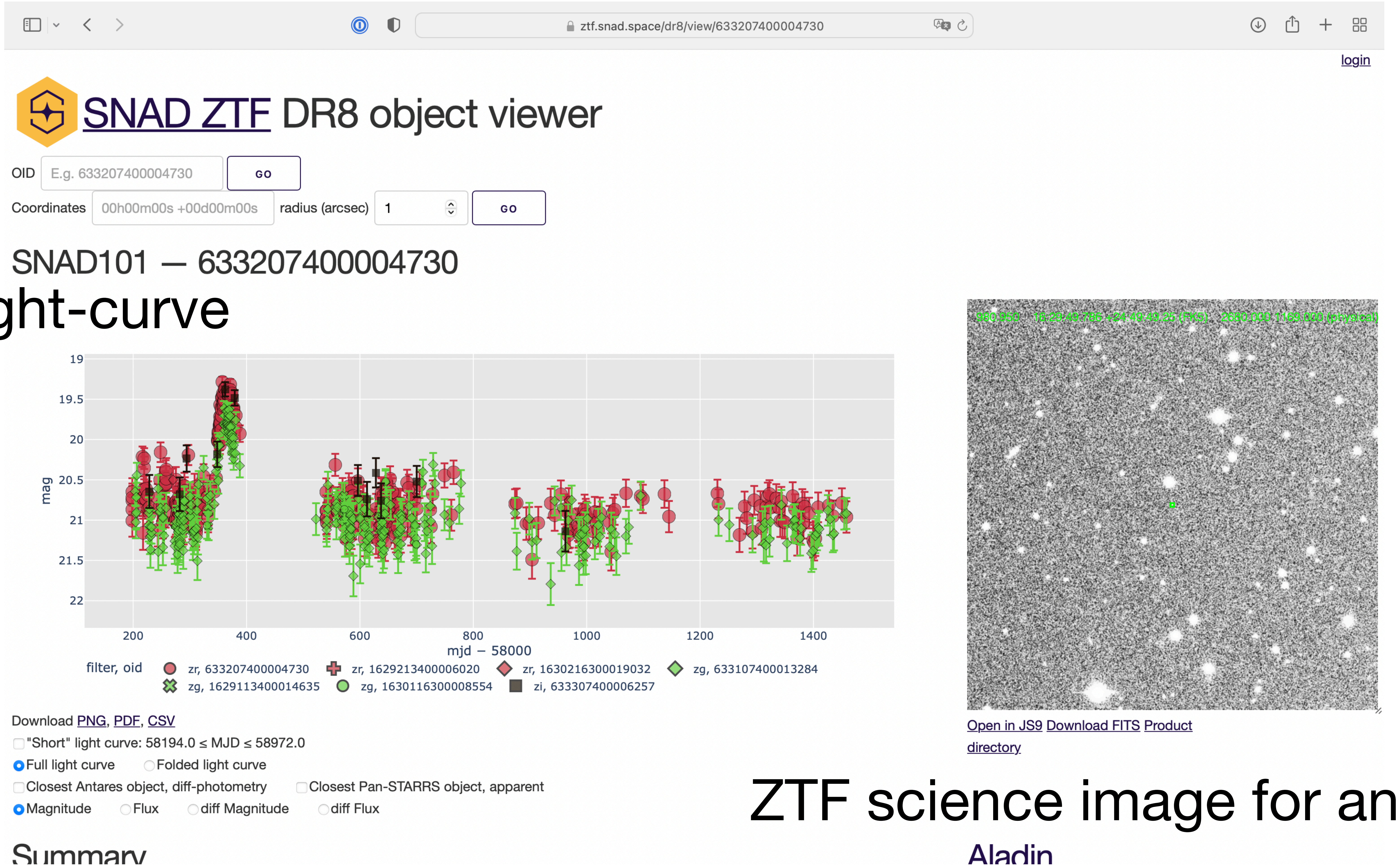
VanderPlas & Ivezić 15, arXiv:1502.01344



# Astronomical data: Expert Portal

## SNAD VIewer, https://ztf.snad.space

Self-matched ZTF light-curve



ztf.snad.space/dr8/view/633207400004730

login

SNAD ZTF DR8 object viewer

OID E.g. 633207400004730 GO

Coordinates 00h00m00s +00d00m00s radius (arcsec) 1 GO

SNAD101 — 633207400004730

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

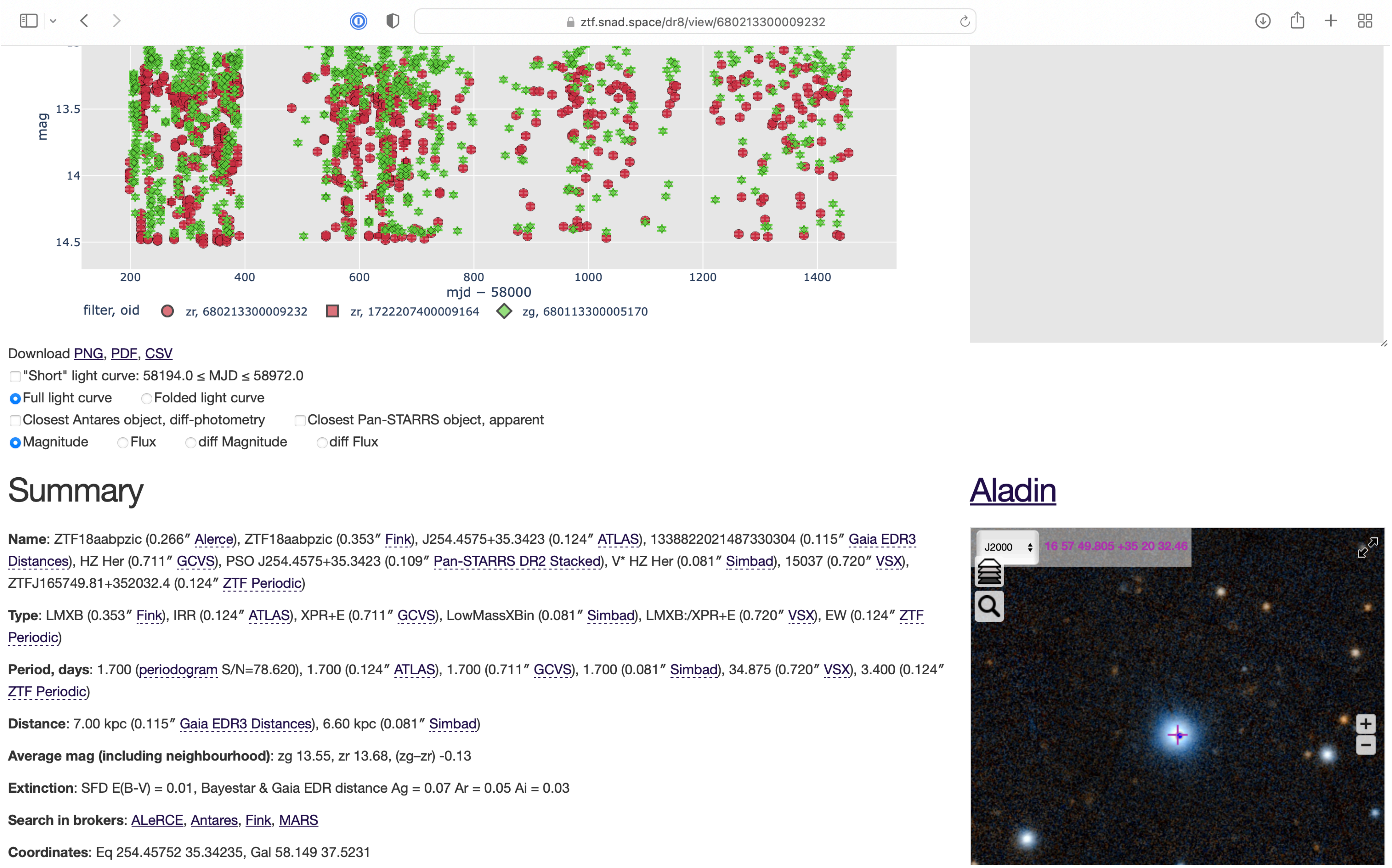
Open in JS9 Download FITS Product directory

Aladin

ZTF science image for any detection

# Astronomical data: Expert Portal

## SNAD VIewer, htts://ztf.snad.space



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

**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 Ag = 0.07 Ar = 0.05 Ai = 0.03

**Search in brokers**: ALeRCE, Antares, Fink, MARS

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

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

# Tutorial Notebooks

Basic tutorial

- "Static" anomaly detection
- Toy data
- Light-curve features

US names time series

MNIST digit images