

# Use of machine learning for anomaly detection in large astronomical databases

Konstantin Malanchev<sup>1,2</sup>, Alina Volnova<sup>3</sup>,  
Matvey Kornilov<sup>1,2</sup>, Maria Pruzhinskaya<sup>1</sup>,  
Emille Ishida<sup>4</sup>, Florian Mondon<sup>4</sup>, Vladimir Korolev<sup>5,6</sup>

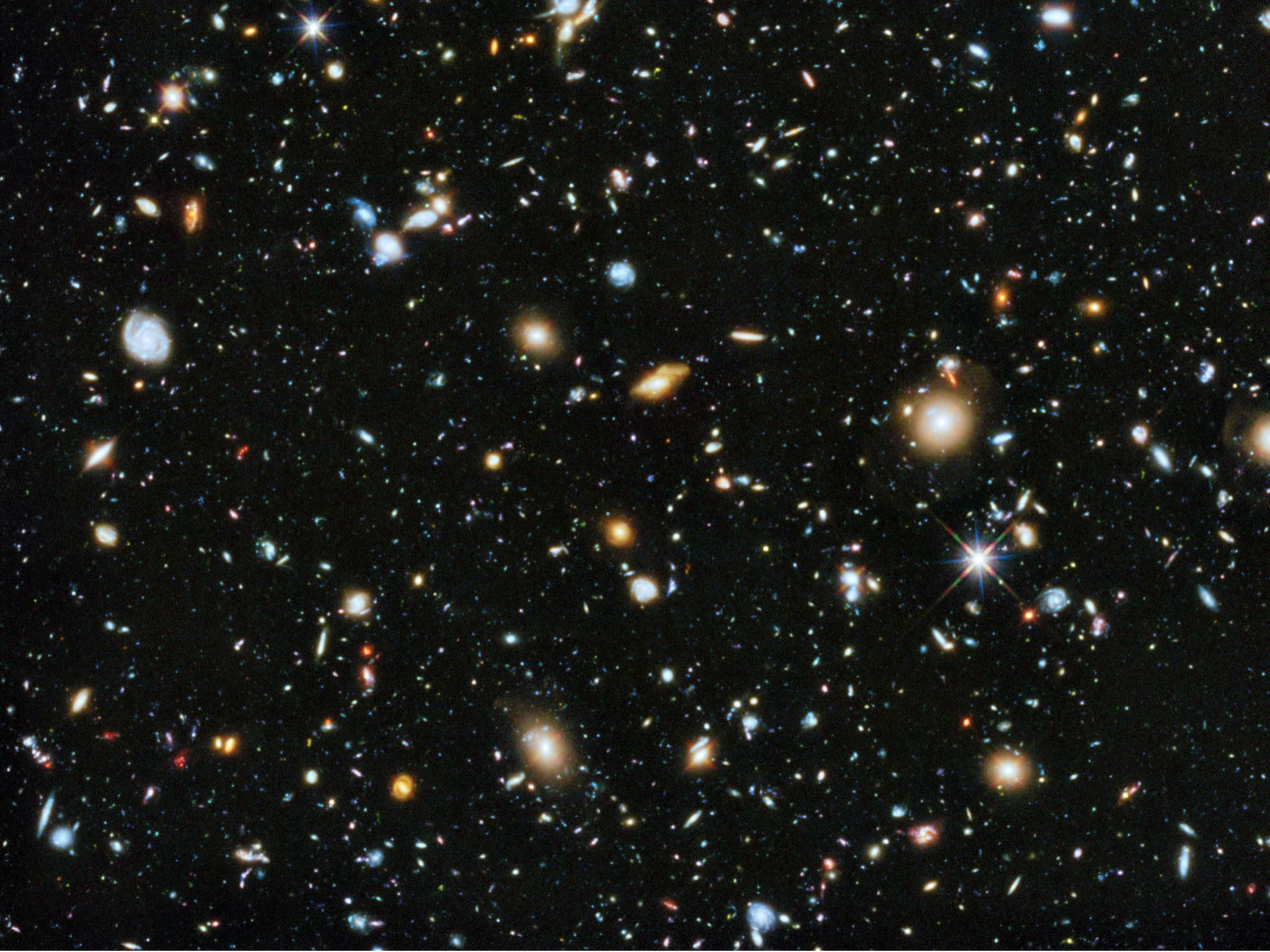
1) Sternberg Astronomical Institute MSU, 2) Higher School of Economics,

3) Space Research Institute (IKI) RAS, 4) Université Clermont Auvergne,

5) Central Aerohydrodynamic Institute, 6) Moscow Institute of Physics and Technology

DAMDID–2019, Kazan, October 17







# Astrophysics as a “Big Data” science

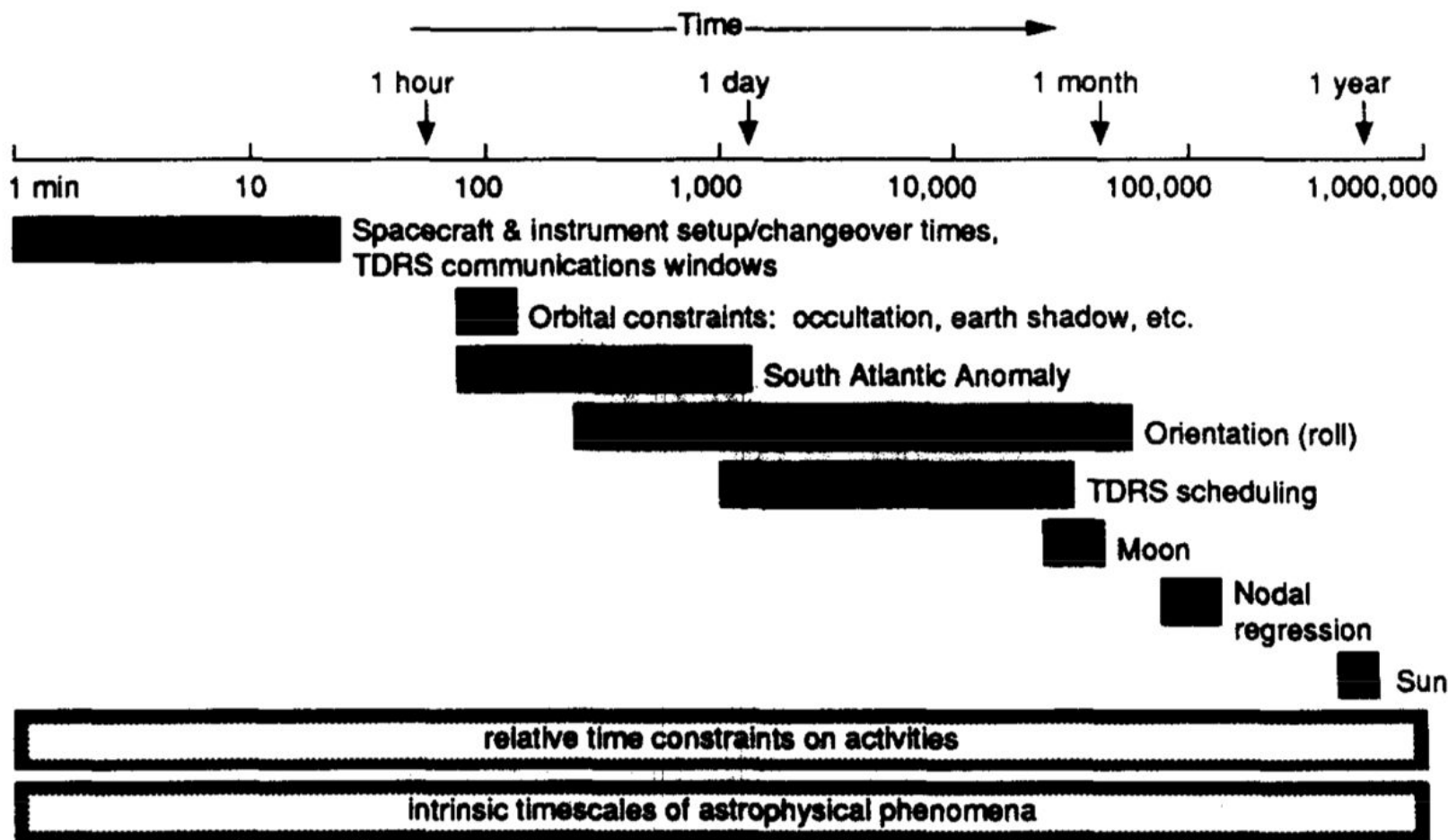
Astronomical catalogs can contain huge amount of open access data.

- Second data release of GAIA space observatory: over billion of positions, and proper motions
- First data release of Zwicky Transient Facility (ZTF): over one hundred million light curves of variable objects with at least 100 observations
- Sloan Digital Sky Survey (SDSS): half billion of objects, four millions of galactic spectra, over 150 TB of data
- Future survey of Large Synoptic Survey Telescope (LSST) will collect several PB of data for ten years

We cannot deal with  
such data volumes  
without ML

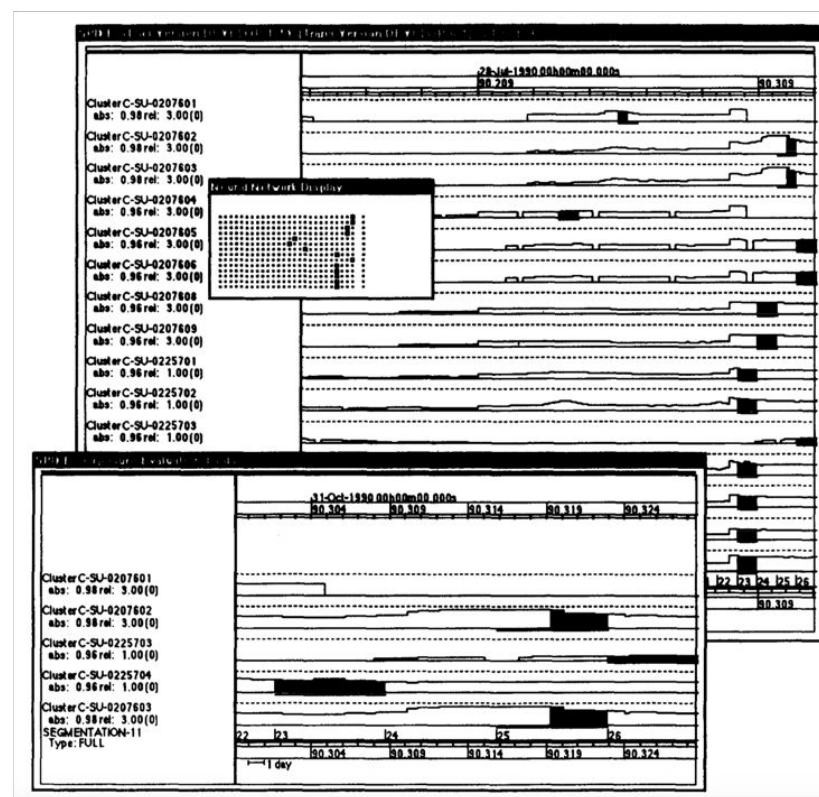
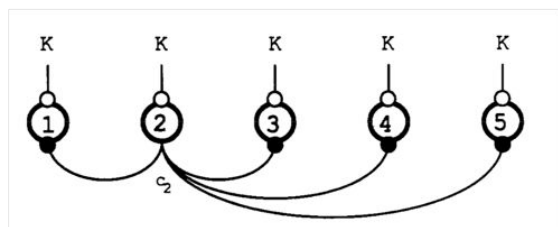
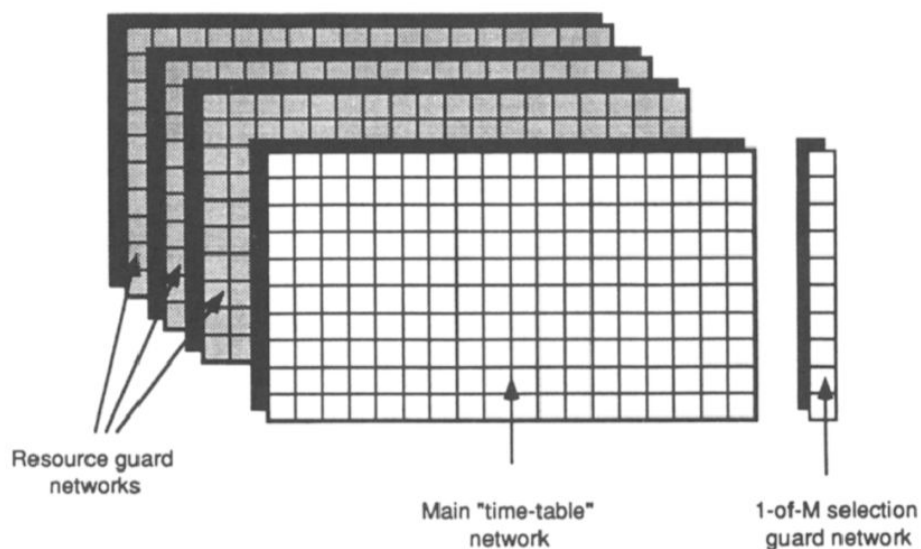
# Machine learning in astrophysics

## Example: Observation scheduling



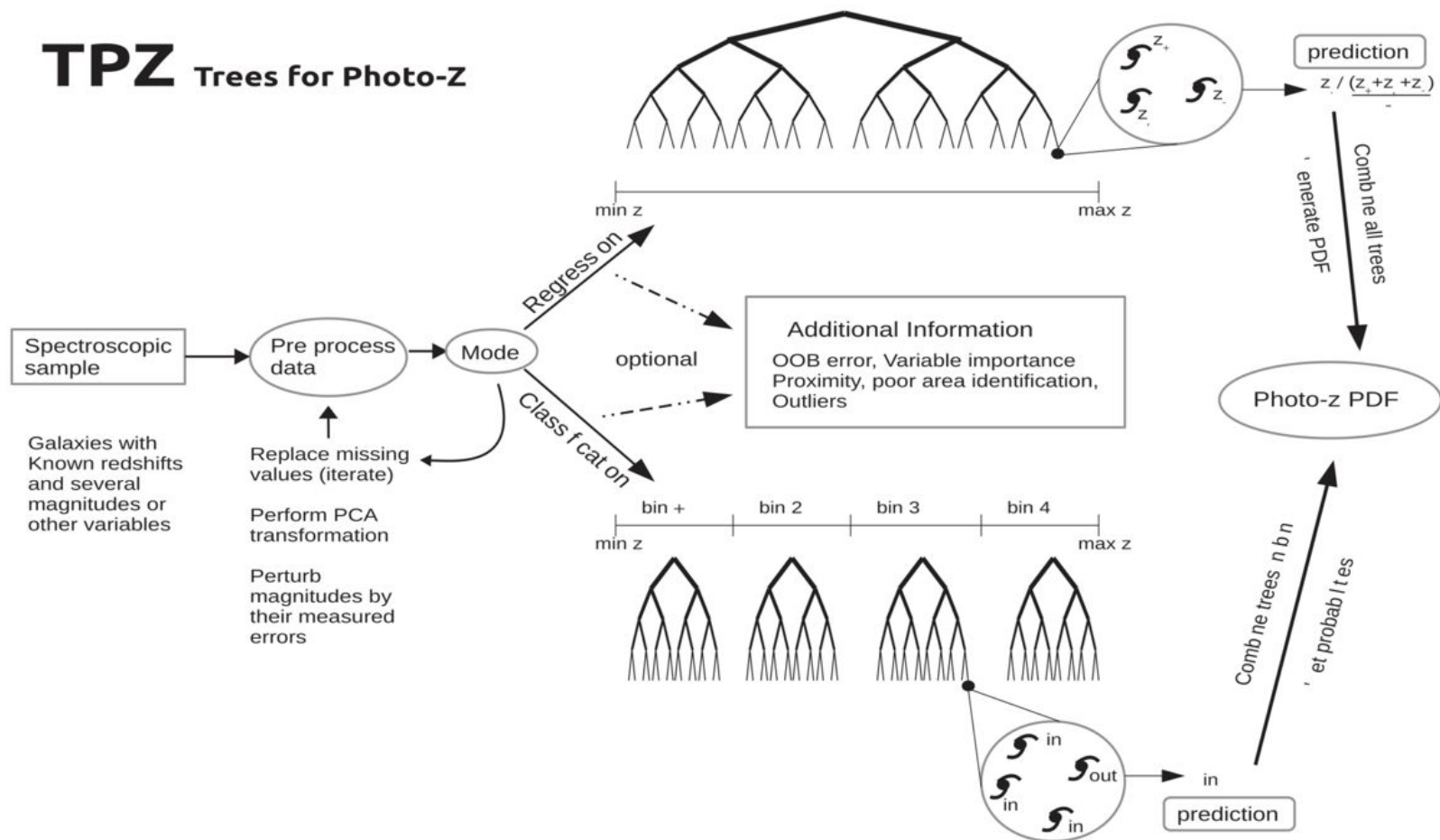
# Machine learning in astrophysics

## Example: Observation scheduling



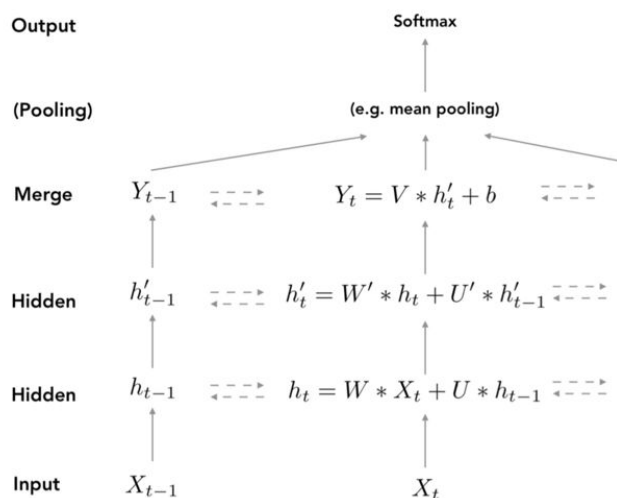
# Machine learning in astrophysics

## Example: Photometry distance estimation

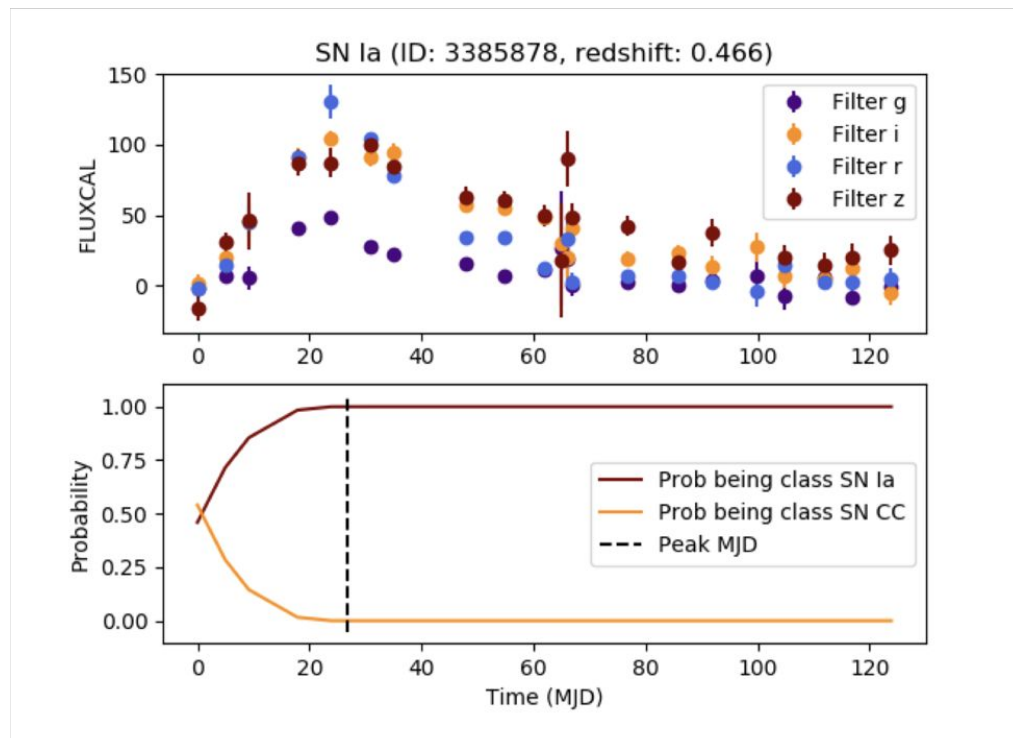


# Machine learning in astrophysics

## Example: Photometric classification of Supernovae



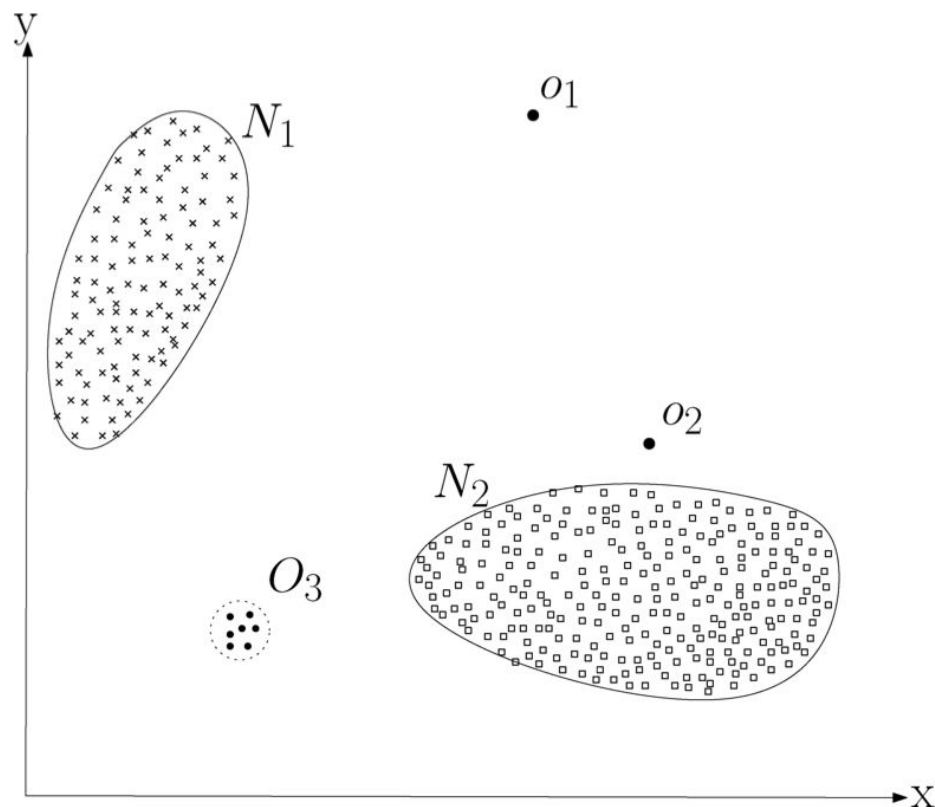
**Figure 1.** Partial schema of a vanilla bidirectional RNN with mean pooling. Layers are indicated in the first left column. Inputs ( $X_t$ ) are given to the network. Arrows indicate transmission of temporal information. The network extends to the right by the number of  $t \in [1, T]$  available inputs.





# Machine learning for anomaly detection

## Outliers

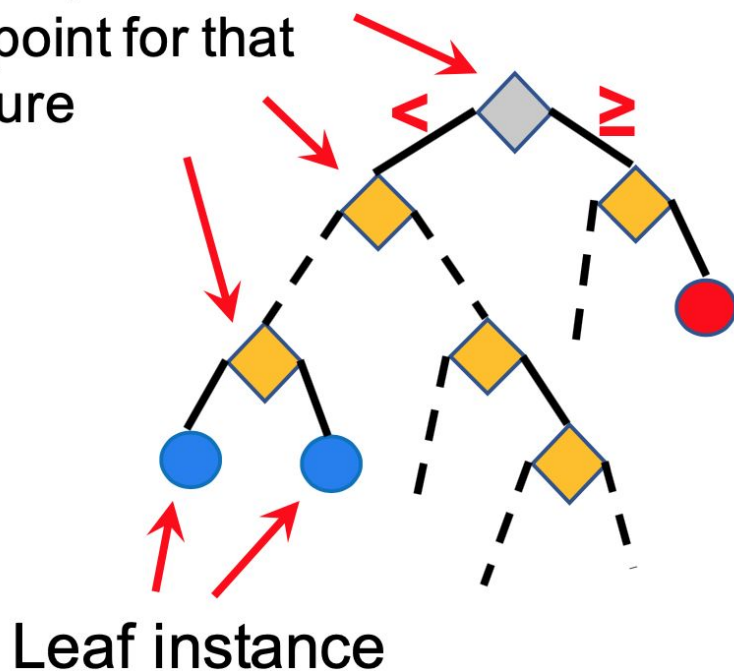


**Fig. 1.** A simple example of anomalies in a two-dimensional data set.

# Machine learning for anomaly detection

## Isolation Tree

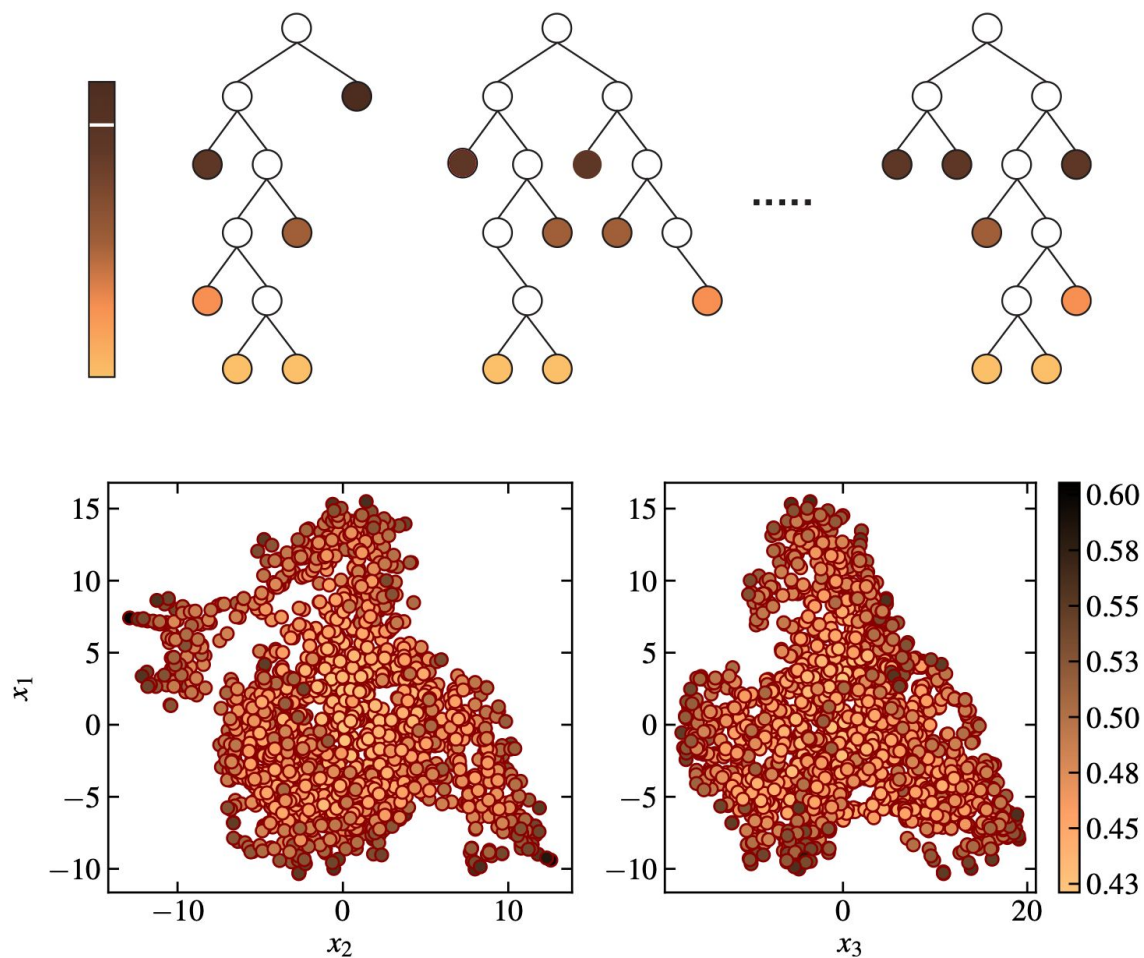
Select a random feature  
at each node, and a  
random split point for that  
feature



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

# Machine learning for anomaly detection

## Isolation Forest



# What is anomaly?

Definition of “anomaly” depends on a problem.

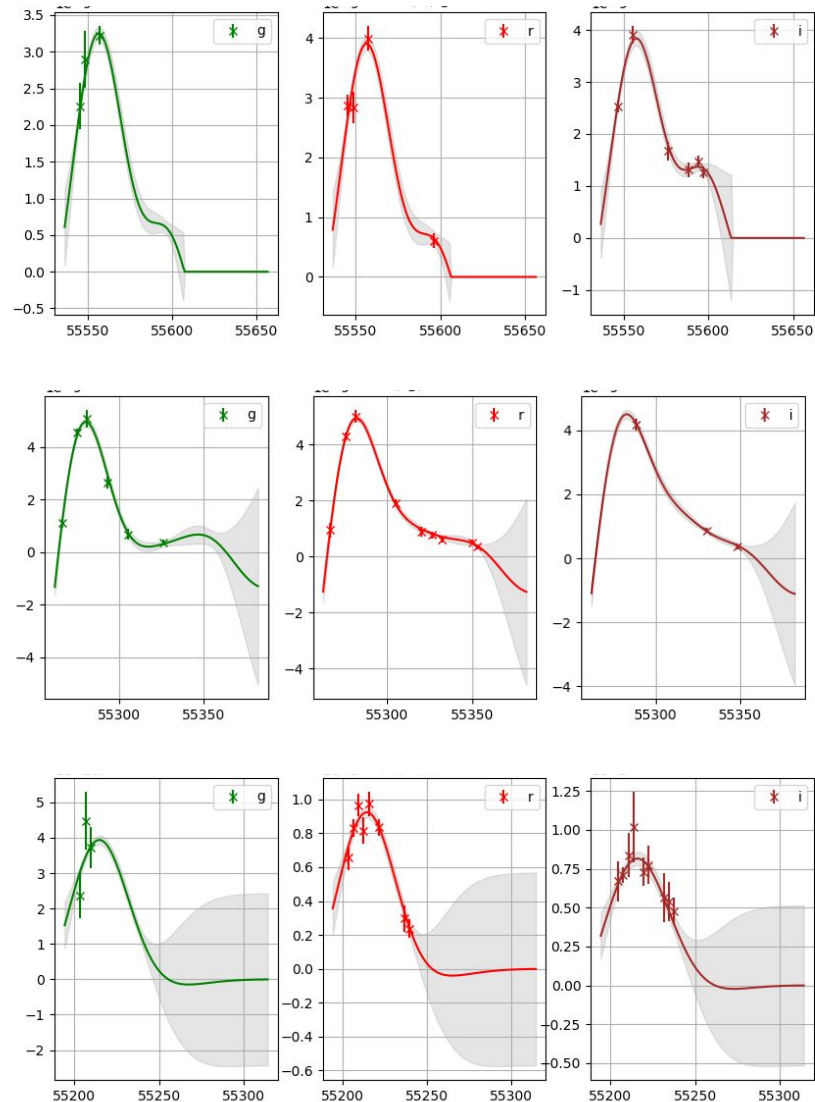
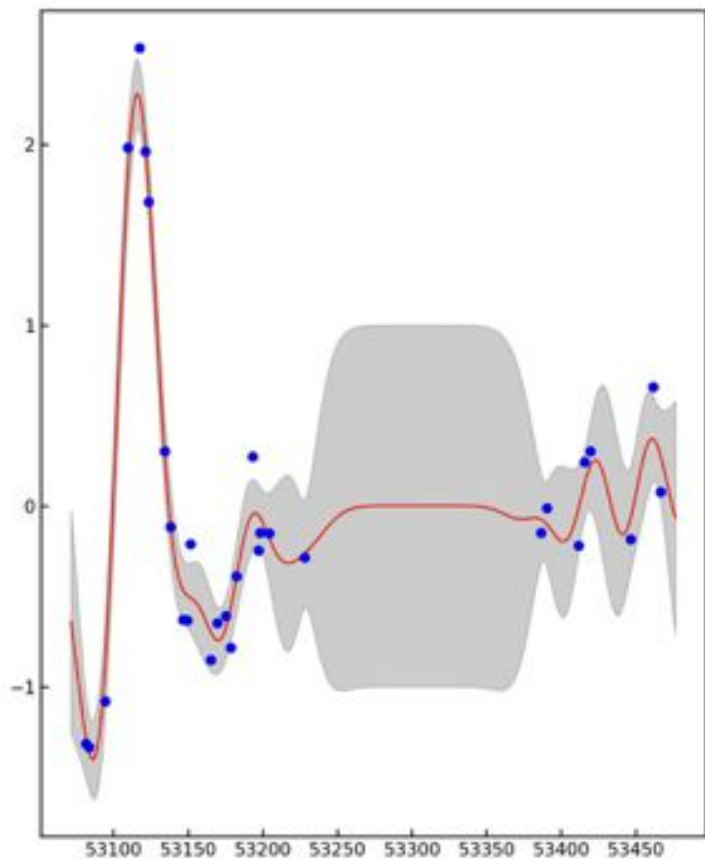
In astrophysics it could be:

- Observation or data reduction artifacts
- Misclassified objects, i.e. active galaxy nuclei in supernova catalog
- Rare class of objects, i.e. microquasar in variable star catalog or gamma ray burst in supernova catalog.
- New physics

Unsupervised machine learning selects outliers, expert analysis of outliers provides anomalies

# Light curve features

## Gaussian processes



<https://gp-multistate-kernel.readthedocs.io>



# Light curve features

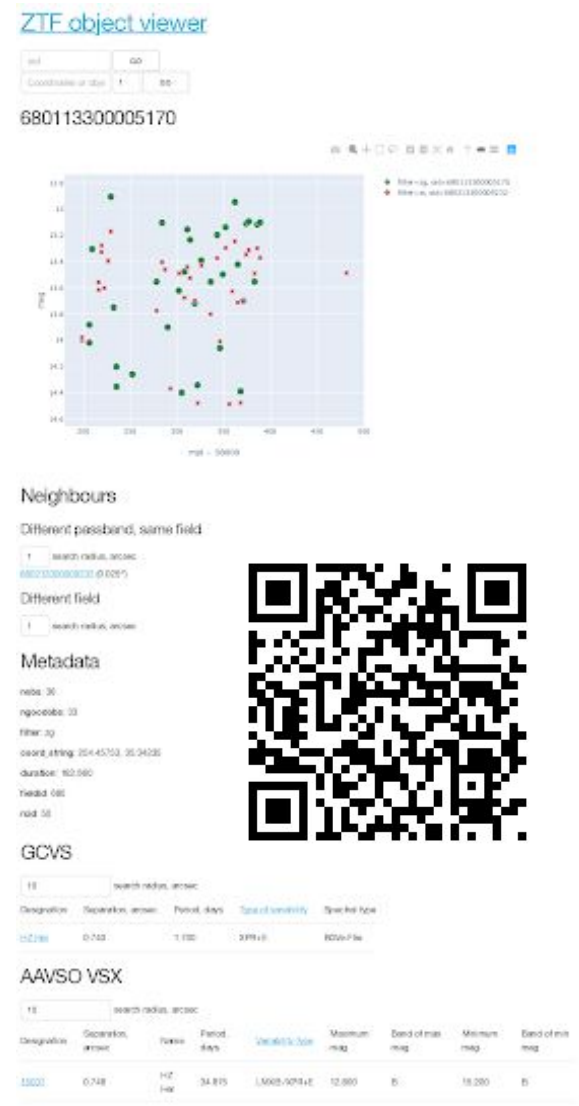
## Variable star approach

We implemented dozens of light curve features based on ML papers on variable star classification

- Magnitude distribution features: amplitude, sample moments, Cusum (Kim et al. 2014), Stetson (1996)  $K$ , ...
- Light curve shape features: maximum slope, linear trend, linear least square fit, ...
- Periodogram based features: peak period, peak significance, shape based features

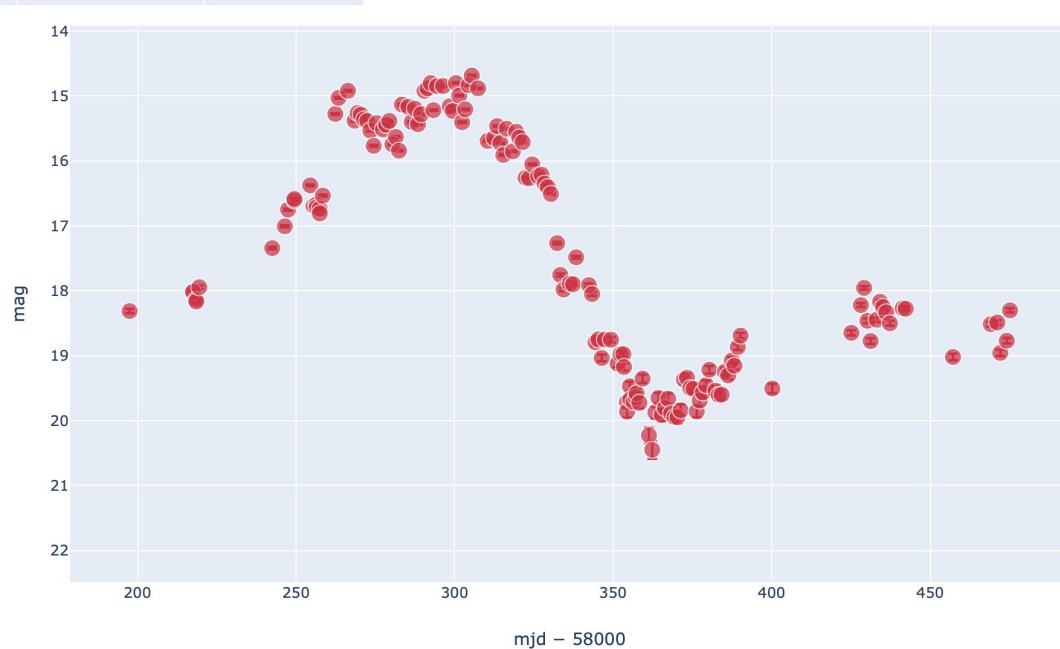
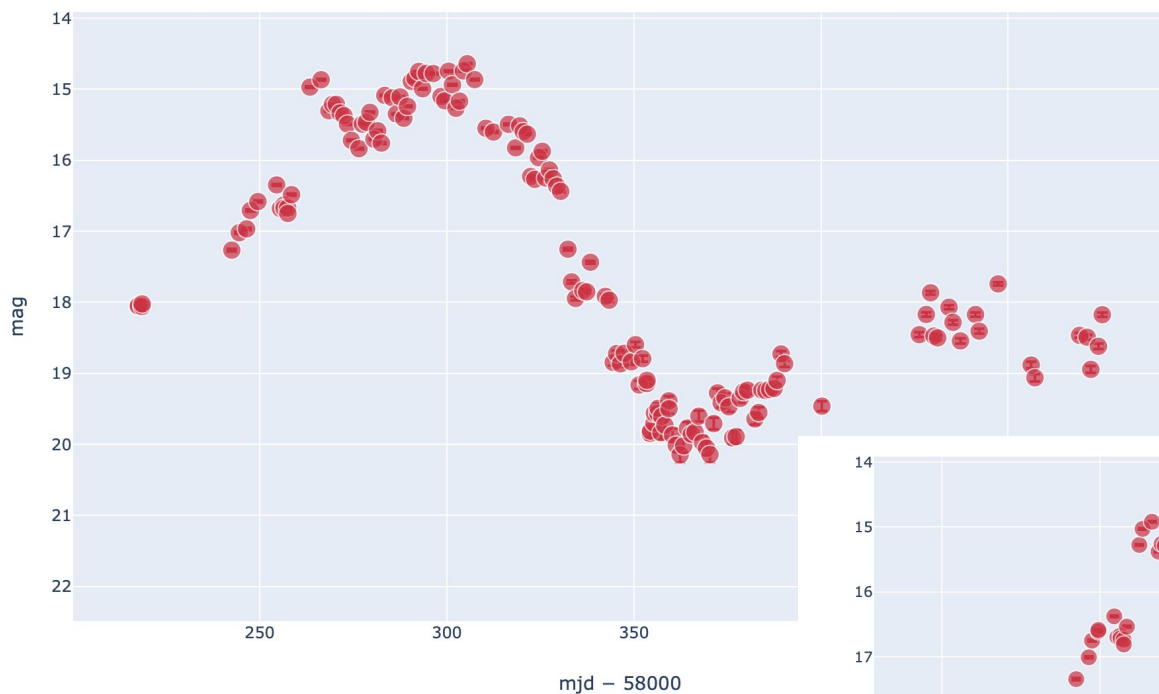
# Anomaly detection in Zwicky Transient Facility DR1 Dataset

- Full catalog contains  $\sim 1.6 \cdot 10^9$  of “objects” in  $g$  &  $r$ , collected in 284 days.
- Raw data are  $\sim 2$  TB, our PostgreSQL database has  $\sim 5 \cdot 10^{11}$  rows and occupies  $\sim 4$  TB
- At least 100 observations per light curve, each covers at least 200 days.  
 $\sim 8 \cdot 10^7$  of light curves in  $r$ .
- 38 features per light curve
- <http://ztf.snad.space> object viewer for expert analysis



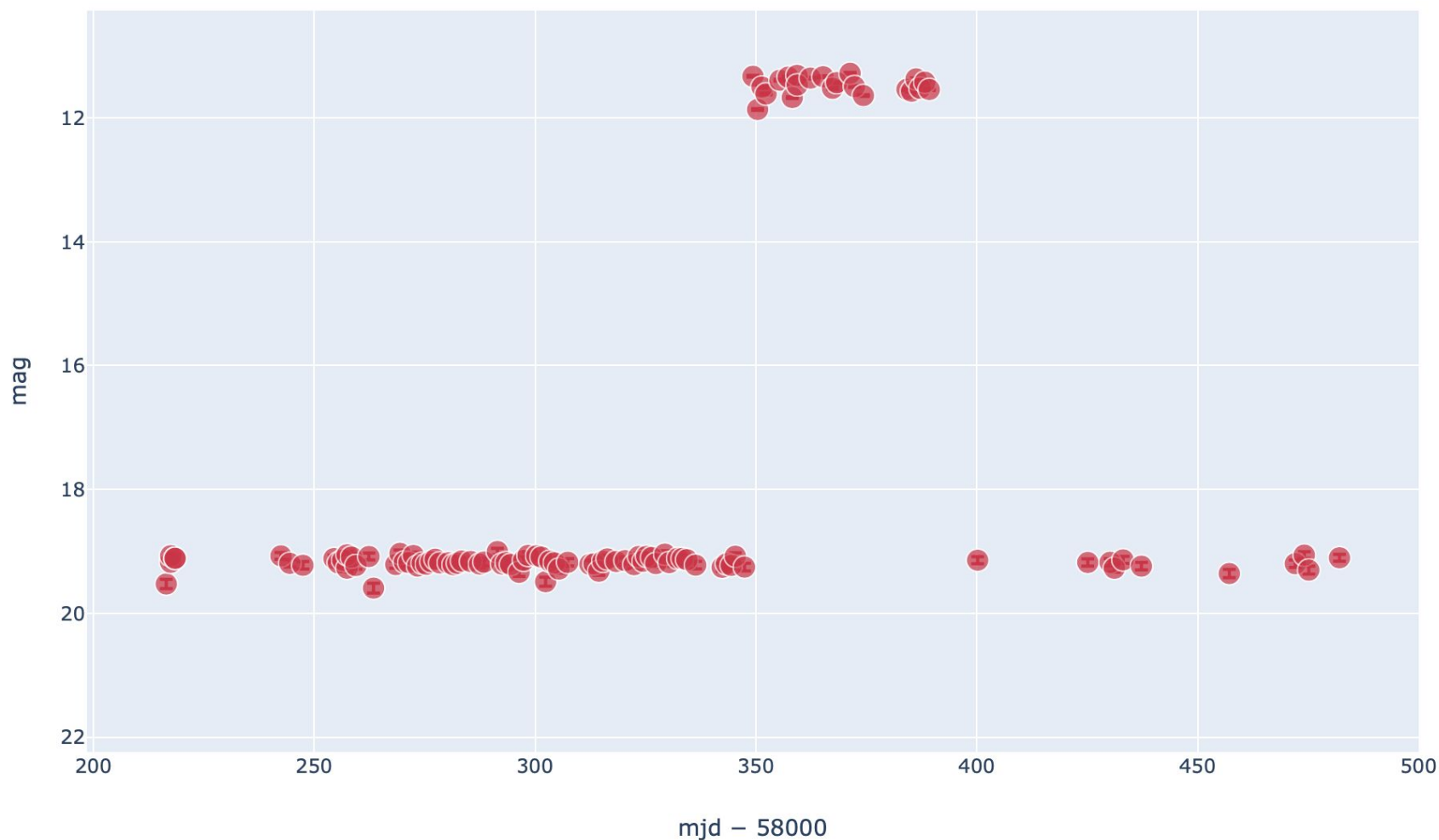
# Anomaly detection in Zwicky Transient Facility DR1

## Preliminary results



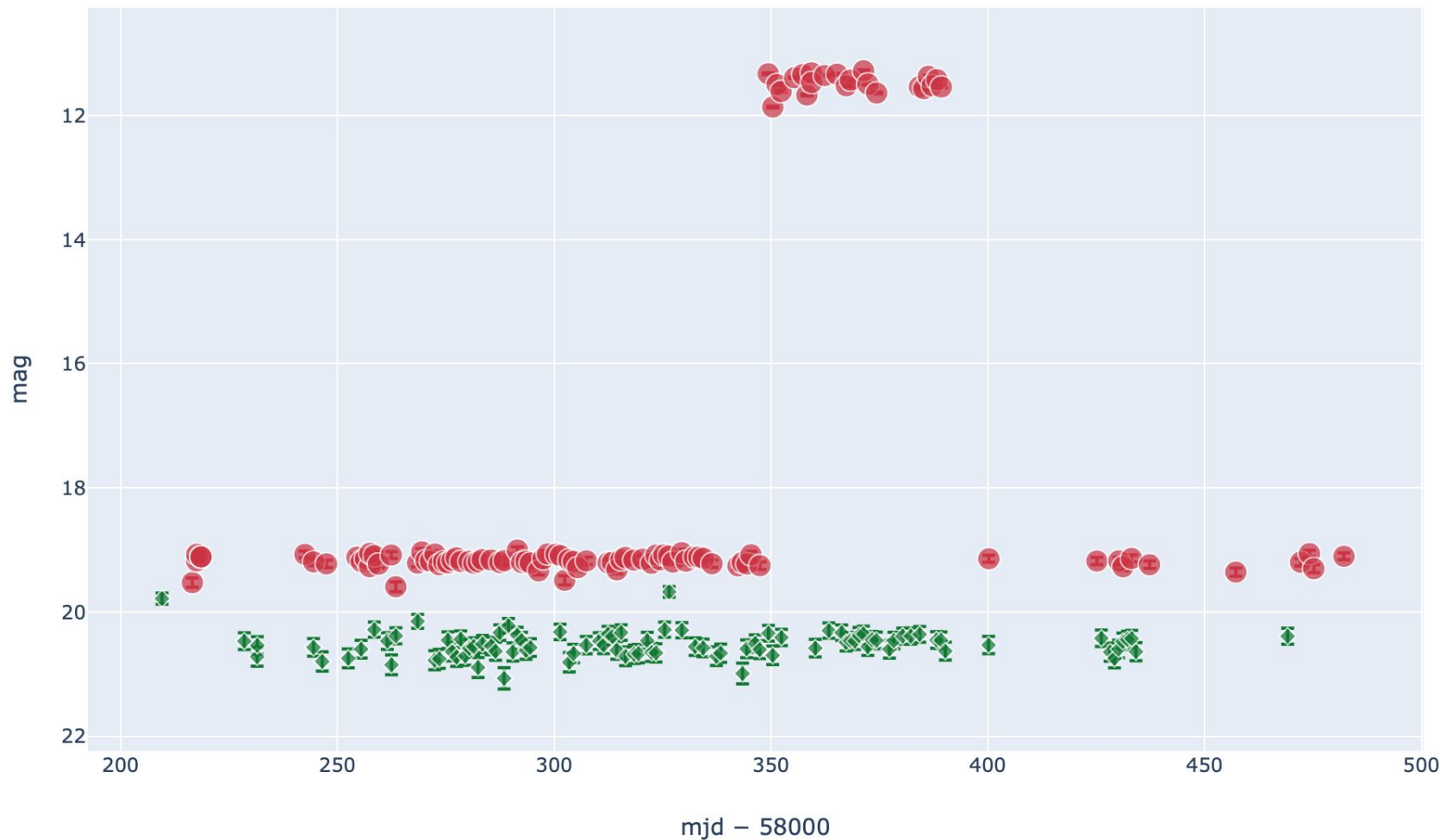
# Anomaly detection in Zwicky Transient Facility DR1

## Preliminary results



# Anomaly detection in Zwicky Transient Facility DR1

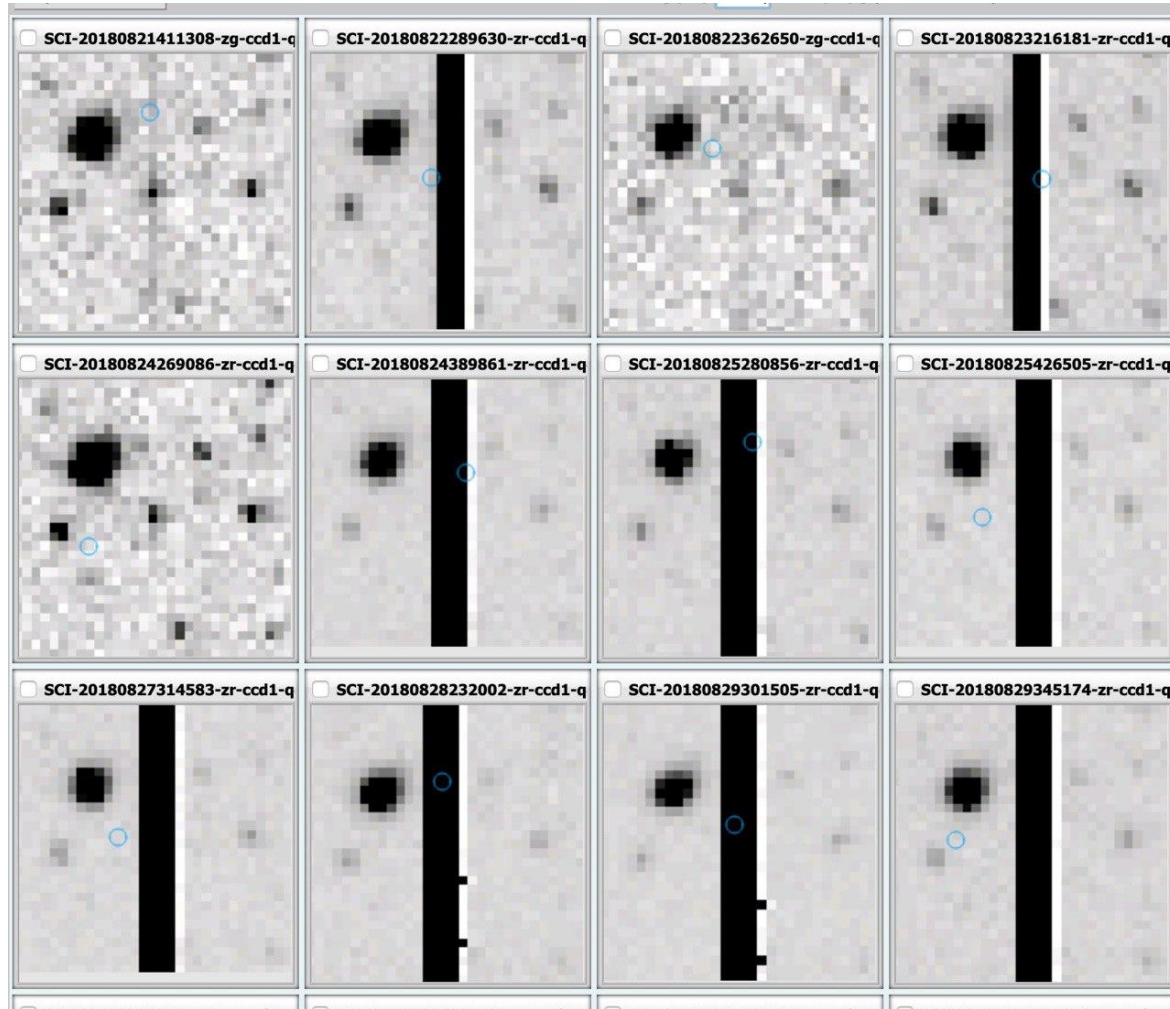
## Preliminary results





# Anomaly detection in Zwicky Transient Facility DR1

## Preliminary results



# SNAD - SuperNova Anomaly Detection

Non formal group from **Sternberg astronomical institute**, **Laboratoire de Physique de Clermont**, **Space Research Institute** and **Central Aerohydrodynamic Institute** joined together to solve the problem of detecting unusual objects in supernova datasets with machine learning methods.



<https://sne.space>




# OSC - Open Supernova Catalog (sne.space)

## The Open Supernova Catalog



[Catalog](#) [About](#) [Contribute](#) [Derivations](#) [Statistics](#) [Download](#) [API](#) [Bibliography](#) [Links](#)

Welcome to the Open Supernova Catalog! The goal of this catalog is to act as a centralized, open repository for supernova metadata, light curves, and spectra. The list of supernovae available here is scraped from various data repositories and individual publications. If you use this data, please [reference the cited sources of that data](#). We'd also appreciate if you referenced [the paper describing this catalog](#). Thanks!

The table below is generated from JSON files produced by [AstroCats](#), which agglomerates the data for each supernova as a series of JSON files that all adhere to the same [schema](#). The entirety of the data available for any supernova can be downloaded by clicking the  icon in the Data column. Advanced data retrieval options are available via the [OACAPI](#). If you would like to contribute data yourself, please visit our [contribute](#) page. If you spot any errors, [please create a new issue on our GitHub issue tracking page](#), or [contact us](#) via e-mail.

Most requested objects for week of October 10, 2019:

- SN2002er, Ia supernova  
<https://t.co/VoStn6c7PZ>
- SN2008ax, IIb supernova  
<https://t.co/miuFsPRIPn>
- SN2000F, Ic supernova  
<https://t.co/WDT0bn36pu>
- SN2004gk, Ic supernova  
<https://t.co/MiEbkv39NT>

[Astro Catalogs](#)   2

6 days ago

[Select all](#) [Deselect all](#) [Column visibility](#) [Export selected: CSV](#) [Export selected: JSON](#) [Permalink](#) [+ Add Supernova](#) Search:

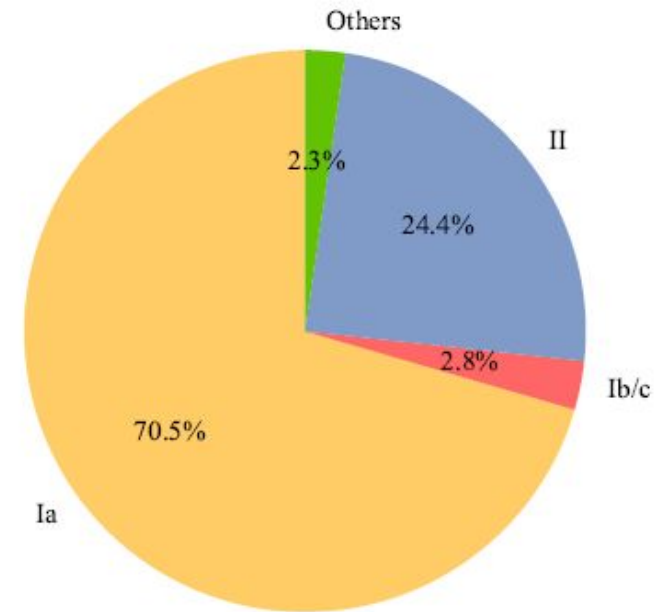
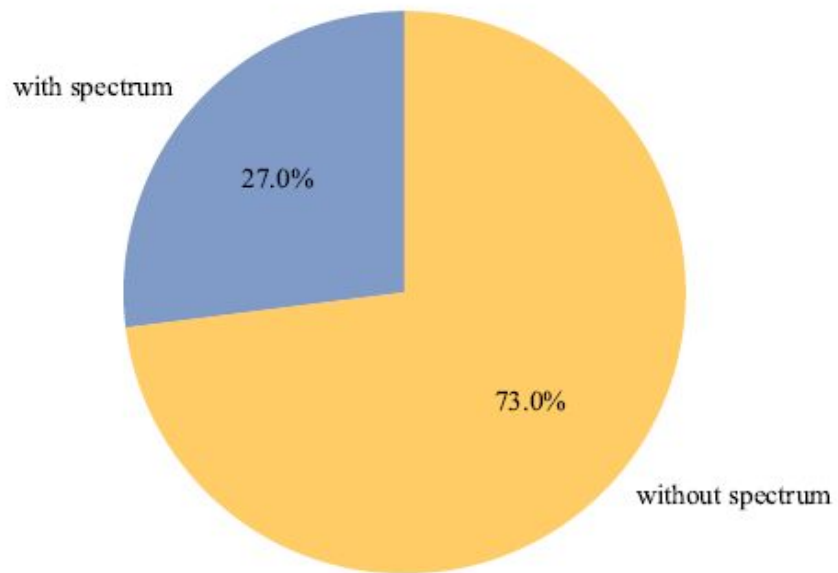
Show  entries [Previous](#) [1](#) [2](#) [3](#) [4](#) [5](#) ... [1086](#) [Next](#)

	Name	Disc. Date	$m_{\text{max}}$	Host Name	R.A.	Dec.	Type	Phot.	Spec.	Radio	X-ray	Data
<input type="checkbox"/>	SN2011fe	2011/08/24	9.48	 M101	14:03:05.711	+54:16:25.22	Ia	 3371	 80	 0	 0	  
<input type="checkbox"/>	SN1987A	1987/02/24	1.9	 LMC	05:35:28.020	-69:16:11.07	II Pec	 3332	 36		 105	  
<input type="checkbox"/>	SN2003dh	2003/03/31	12.62	 A104450+2131	10:44:50.030	+21:31:18.15	Ic BL	 2781	 16			  
<input type="checkbox"/>	SN2013dy	2013/07/10	12.8	 NGC 7250	22:18:17.60	+40:34:09.6	Ia	 2275	 85			  

# OSC composition (Guillochon et al. 2017)

Asiago Supernova Catalog	<b>55 000</b> SNe and candidates
Gaia Photometric Science Alerts	<b>600 000</b> photometrical data points
Nearby Supernova Factory	<b>20 000</b> spectra
Pan-STARRS	
SDSS Supernova Survey	
Sternberg Astronomical Institute Supernova Light Curve Catalogue	
Supernova Legacy Survey	
MASTER Global Robotic Net	
All-Sky Automated Survey for Supernovae	
Palomar Transient Factory	

# OSC Statistics





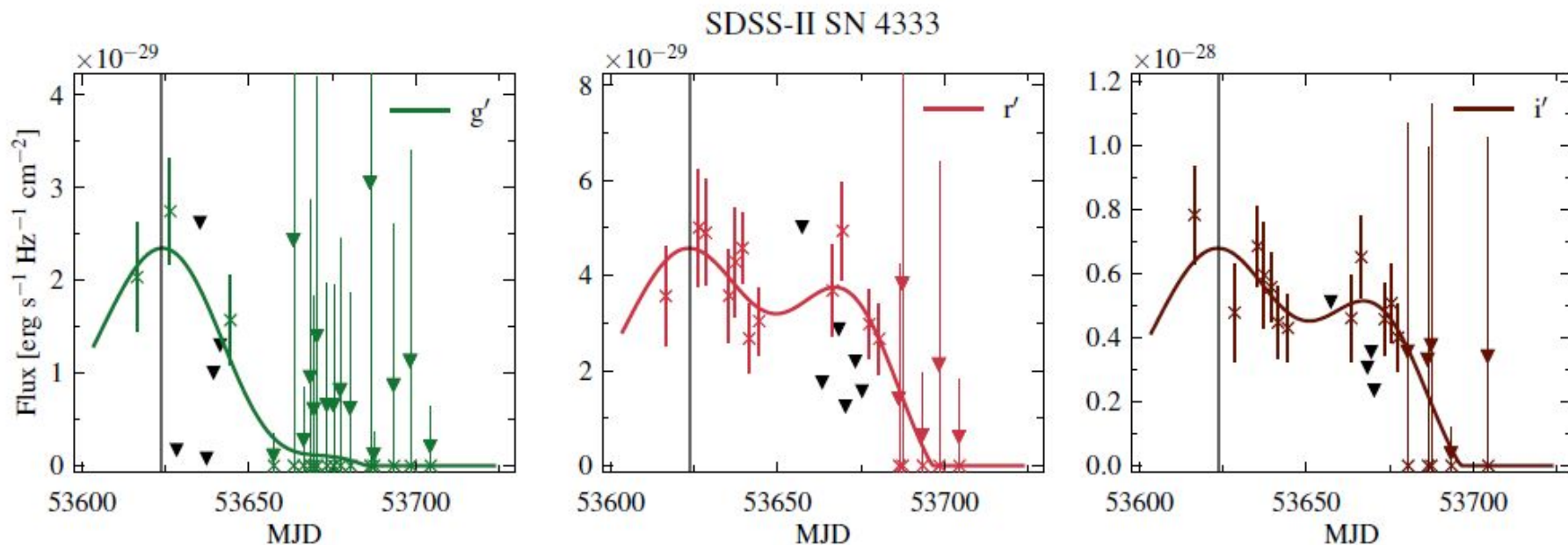
# OSC problems

- Unevenly distributed flux measurements
- Only a few passbands usually available per light curve
- For each LC we have different time span before the maximum
- Unreliable measurement accuracy estimations

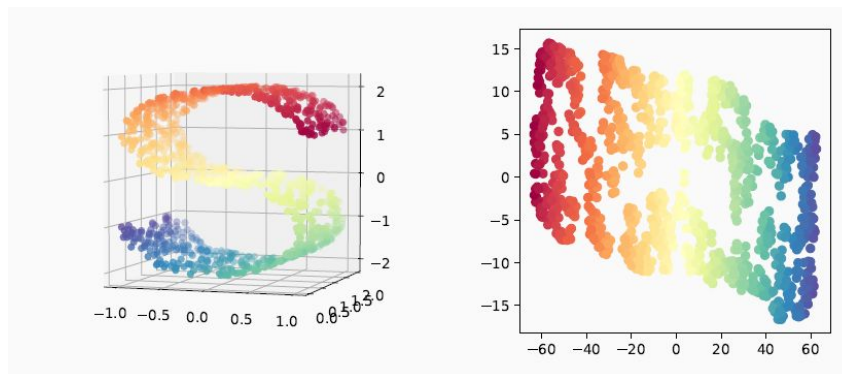
# Choosing data

- Filters gri or g'r'i' or BRI
- At least 3 data points in each filter
- All data converted to gri using known photometrical equations
- Fitting with Gaussian Processes in the range of [-20, +100] (arbitrary zero point)

=> 1999 objects



# Dimensionality reduction



Nonlinear dimensionality reduction technique t-SNE (Maaten & Hinton 2008).

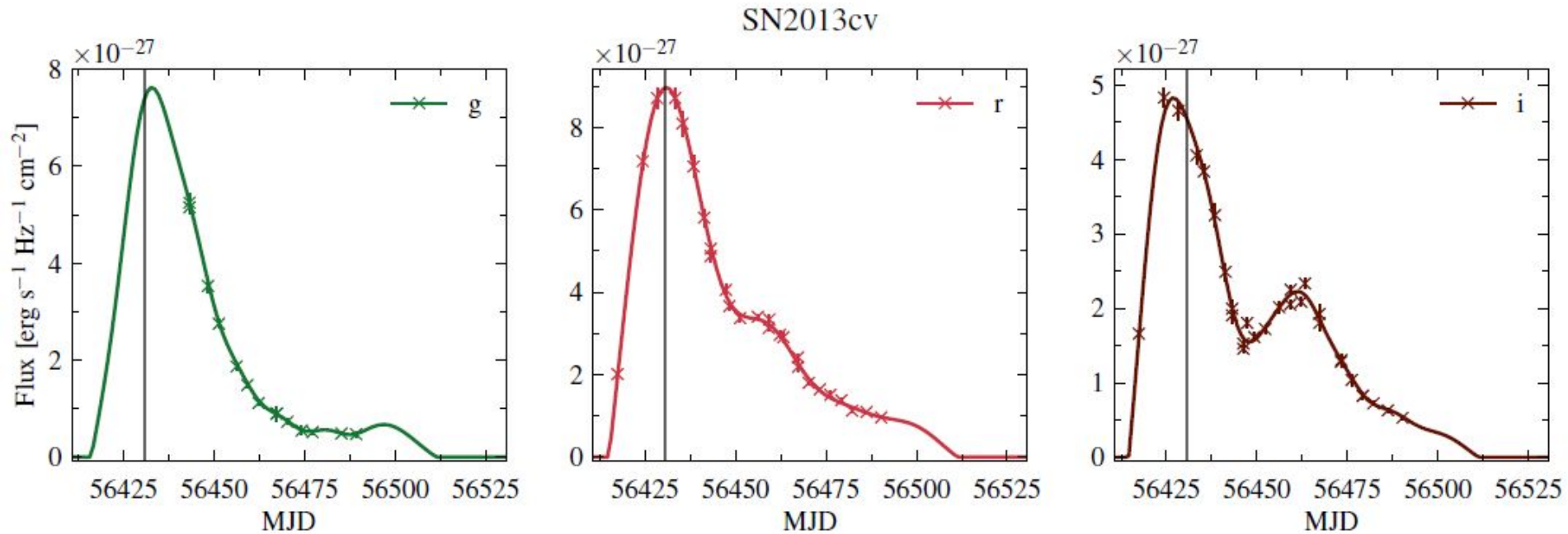
After the approximation procedure, each object has **374** features:

**1213** normalized fluxes, the LC flux maximum, **9** fitted parameters of the Gaussian process kernel, and the log-likelihood of the fit.

OUTPUT: **8** separate reduced data sets corresponding to **2 to 9** t-SNE features (dimensions).

## Results of Isolation forest algorithm

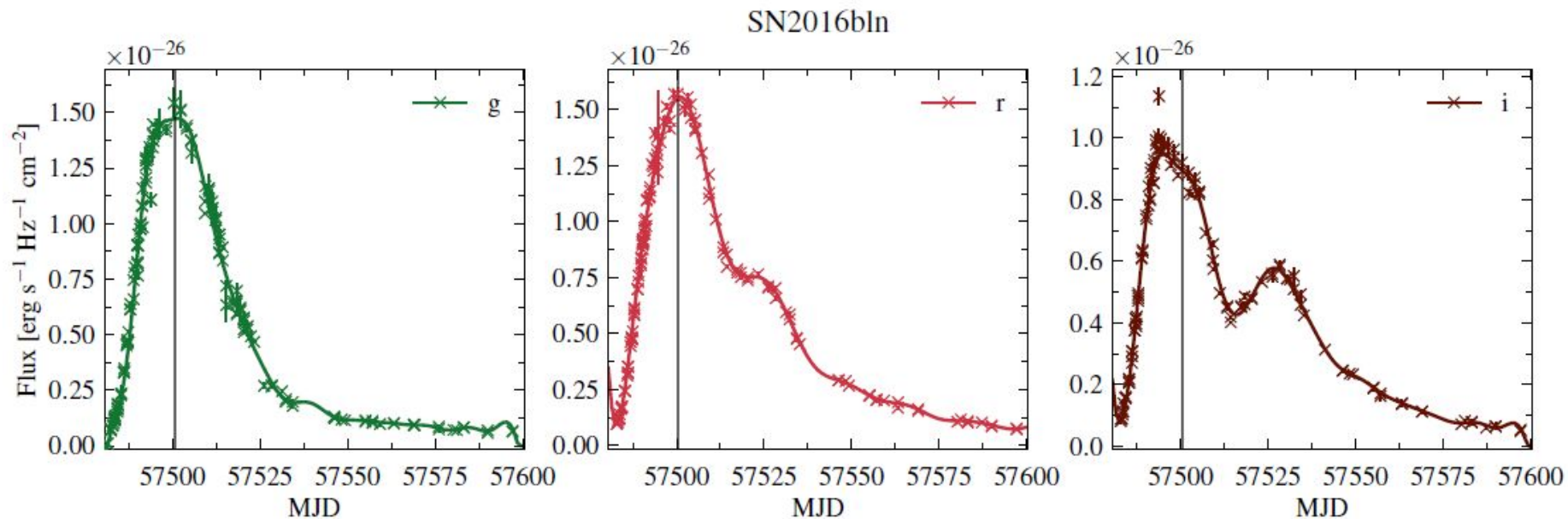
# SN 91-T: SN 2013cv



- SN 91-T looks similar to Ia. The issue that it is brighter.
- Cenko, S. et al., ATel 8909 (2016)

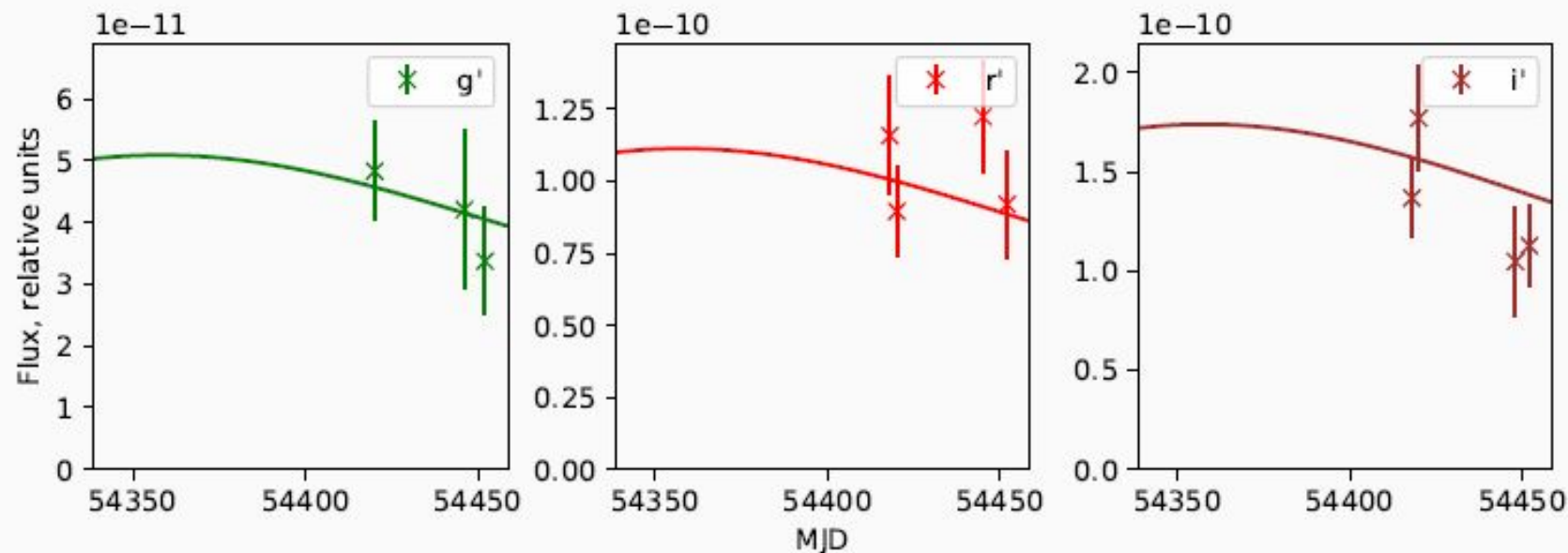


# SN 91-T: SN 2016bln



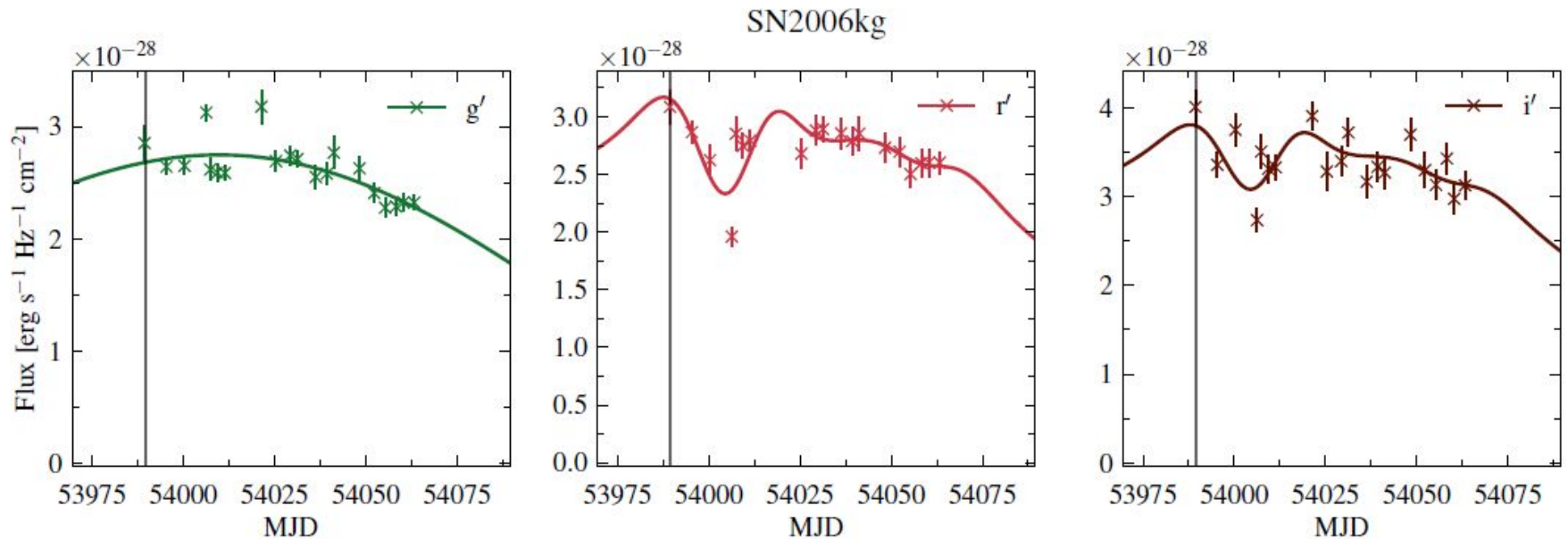
- Cao, Yi et al., AJ, Vol. 823, Issue 2, 147, 13 pp. (2016)

# SLSN-II: SN1000+0216



- Cooke, J et al. Nature, Volume 491, Issue 7423, pp. 228-231 (2012)

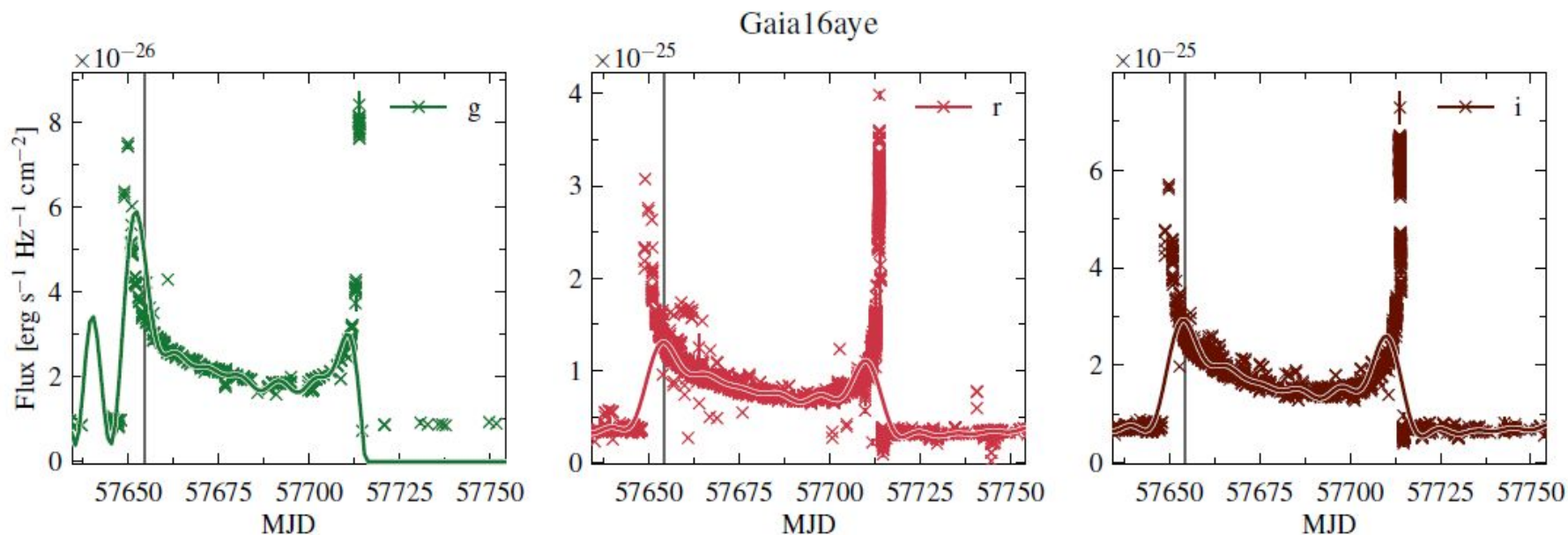
# AGN: SN2006kg



was first classified as a possible Type II SN

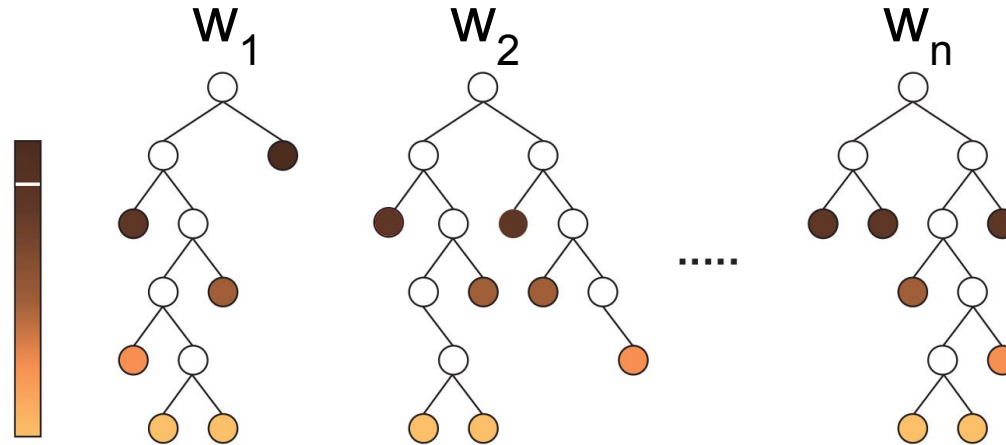
further analysis of 3.6-m New Technology Telescope spectrum revealed that SN2006kg is an active galactic nucleus (Östman et al. 2011; Sako et al. 2018).

# Binary microlensing event: Gaia16aye



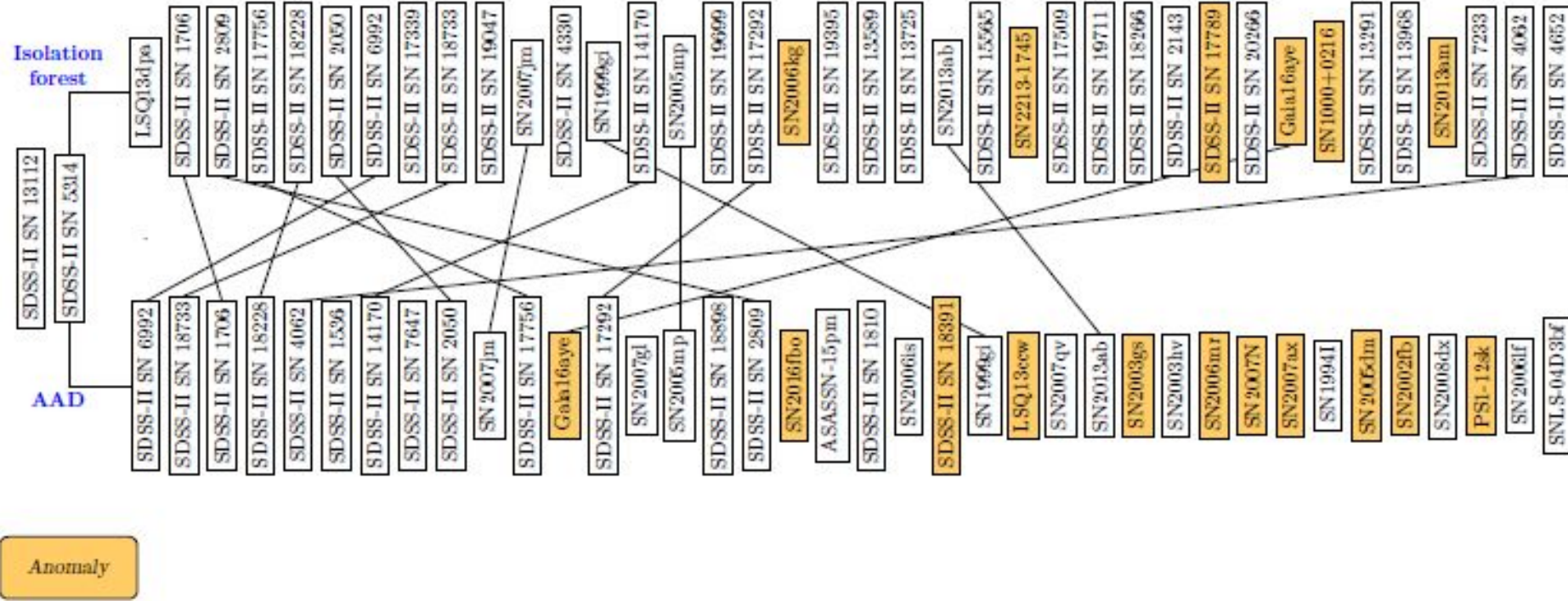
In Wyrzykowski et al. (2016) it was reported that Gaia16aye is a binary microlensing event — gravitational microlensing of binary systems — the first ever discovered towards the Galactic Plane.

# Active Anomaly Detection



1. Initialize isolation forest or other ensemble of anomaly detectors, set equal weight to each detector
2. Ask the ensemble 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, reweight detectors to set lower weight to detectors that give higher score for the object, go to step 2

# Active Anomaly Detection





Thank you for your attention