Поиск аномалий в фотометрических каталогах переменных звезд

ГАИШ: Константин Маланчев, Мария Пружинская, Матвей Корнилов
Université Clermont Auvergne: Emille Ishida, Florian Mondon, Sreevarsha Sreejith
ИКИ РАН: Алина Вольнова
ЦАГИ: Владимир Королёв
Cinimex: Анастасия Маланчева
Washington State University: Shubhomoy Das

Микросеминар отдела релятивистской астрофизики ГАИШ, 22 октрябя 2019

SNAD Team

Sternberg Astronomical Institute MSU, 2018

Laboratoire de Physique de Clermont, 2019

Machine learning is a study of algorithms that computers use to build a model from input data

"A computer program is said to **learn** from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by P, improves with experience *E*"

— Tom M. Mitchell, «Machine Learning», 1997

Why Machine Learning

- LSST: photometry for ~$5 \cdot 10^8$ objects, ~$4 \cdot 10^6$ spectra
- GAIA DR2: coordinates, proper motions and photometry for ~10^9 stars
- ZTF DR1: ~$10^{11.5}$ individual photometry measurements for ~10^9 transient objects for less than a year of observations
- LSST (prediction): ~10^{10} galaxies, ~10^{10} stars, ~10^7 supernova, ~100 PB of data for ten years of observations

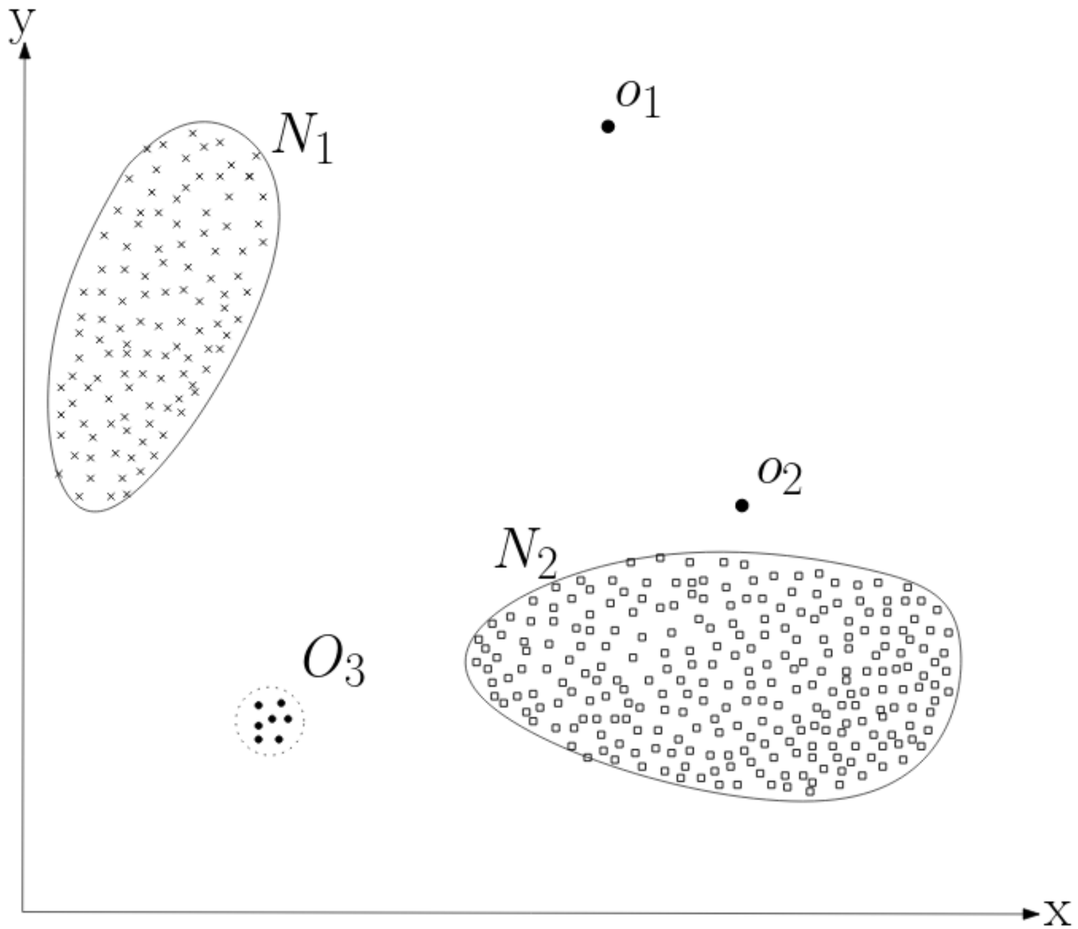
Anomalies

Definition of an anomaly depends on a problem. In astrophysics it could be:

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

Outliers

Machine gives an outlier, expert makes it an anomaly

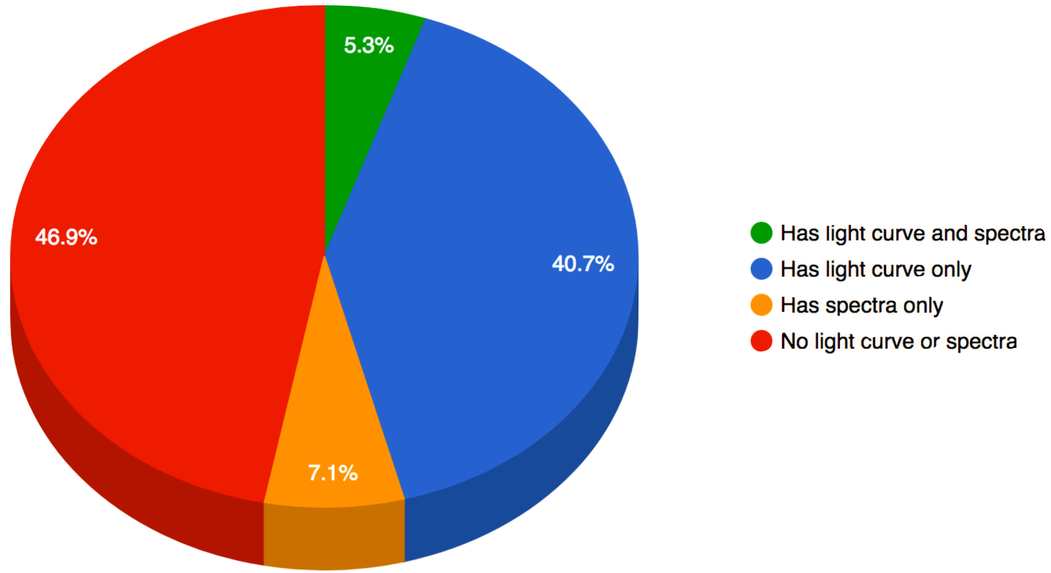


Chandola V., et al. 2009

We search anomalies in light curve data bases

Open Supernova Catalog

- https://sne.space (Guillochon et al. 2017)
- ~55 000 of supernovae and candidates
- ~600 000 of photometric measurements
- ~20 000 спектров



OSC: Data

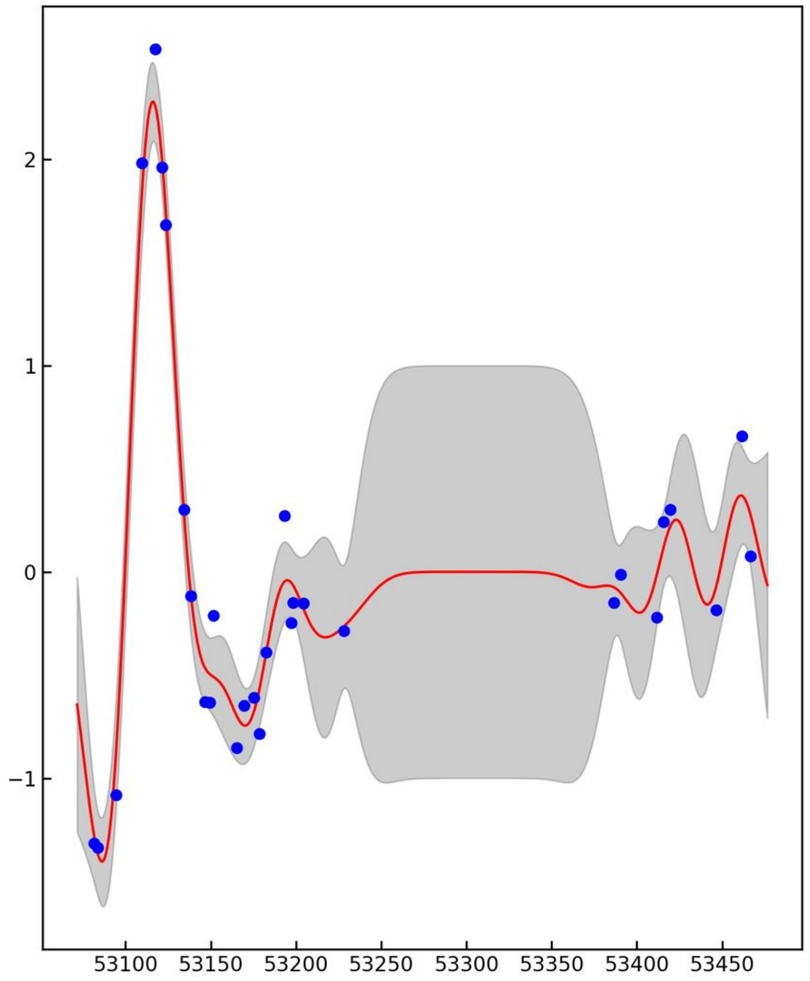
OSC:

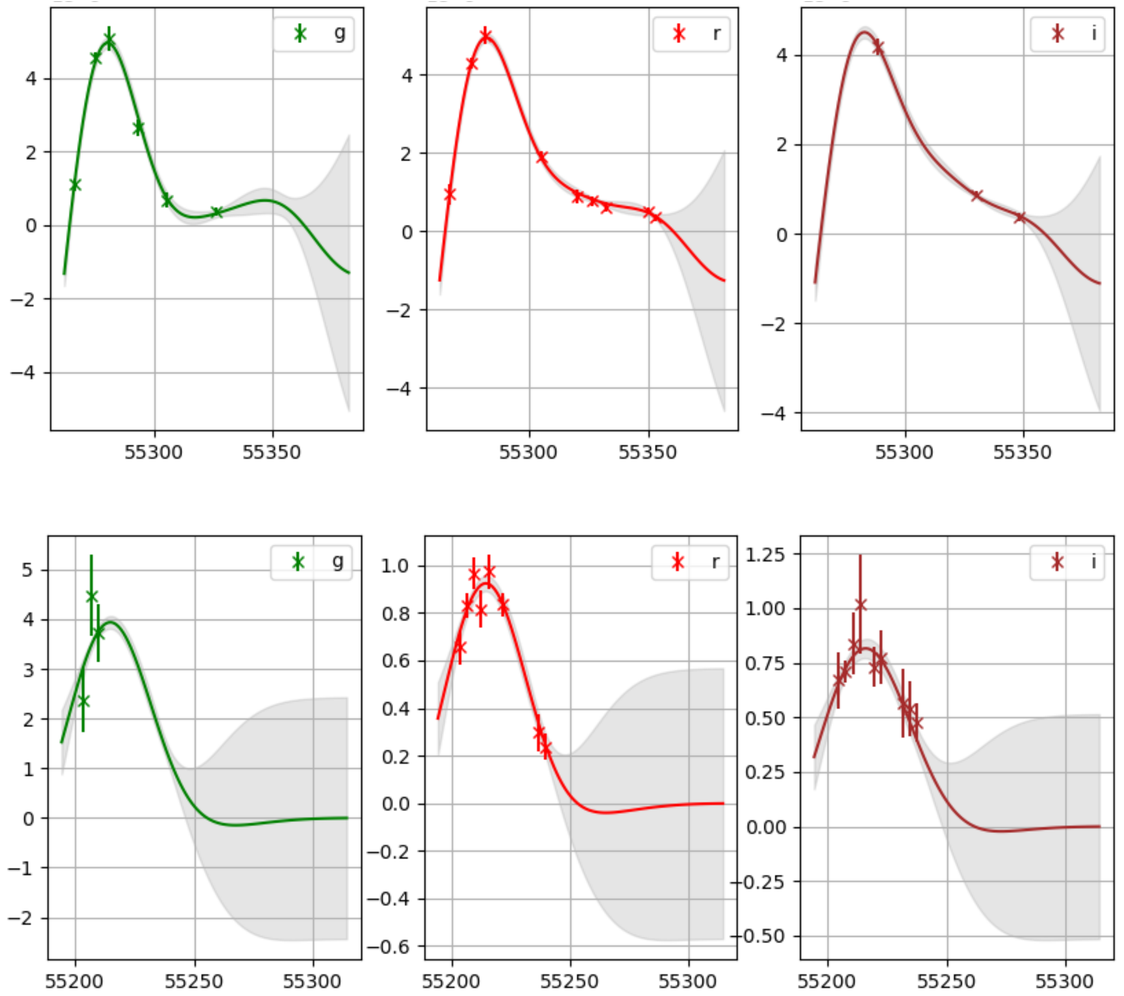
- ~55000 of objects
- Heterogeneous data
- Different passbands
- Observations without errors
- Upper limits
- Light curves are unevenly time series

Our data sample:

- 1999 of objects
- *gri*, *g'r'i'* and *BRI* → *gri*
- Min 3 points per passband
- Only observations with errors
- Upper limits are used
- Approximation via Gaussian processes

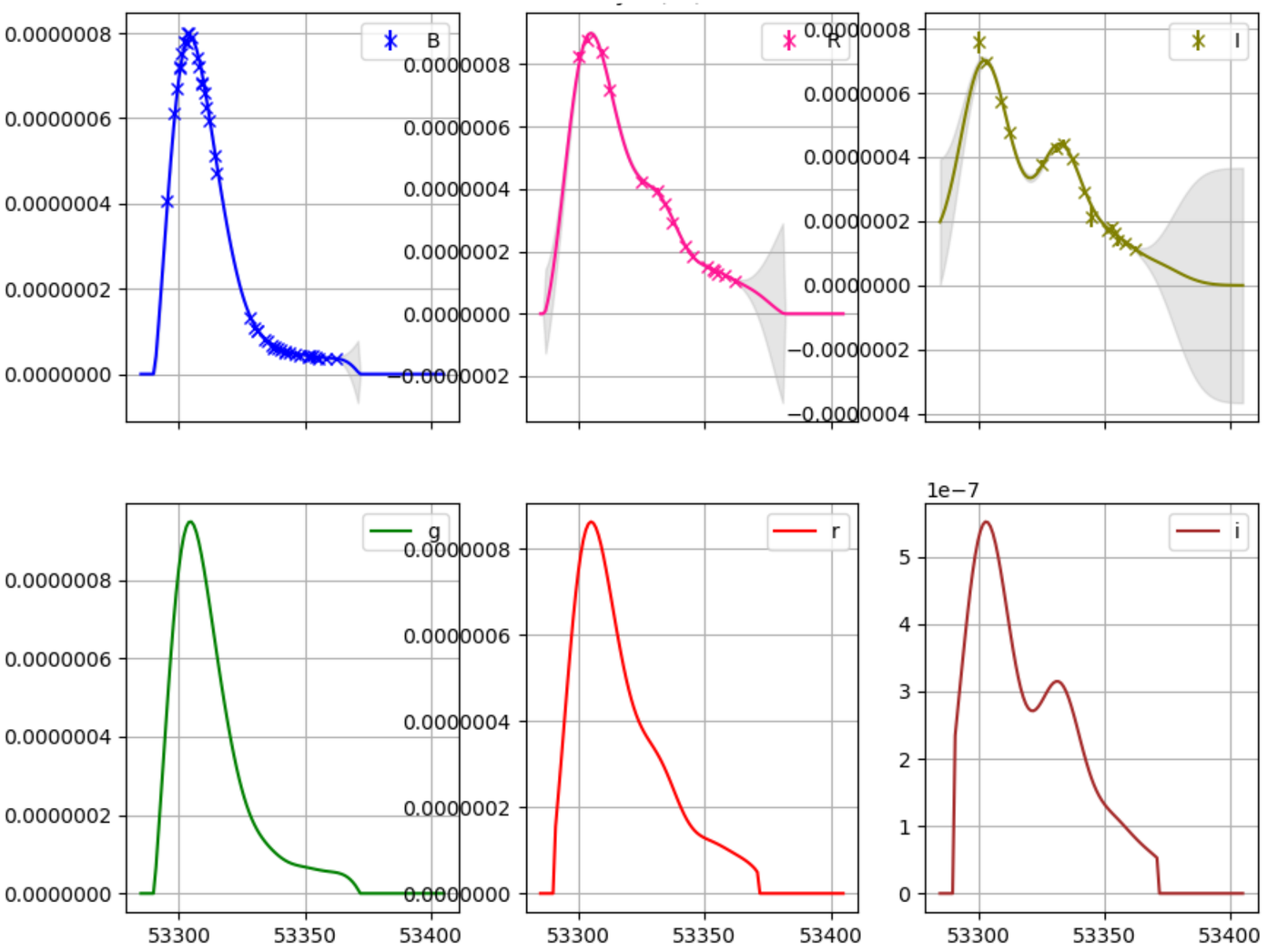
Gaussian Processes



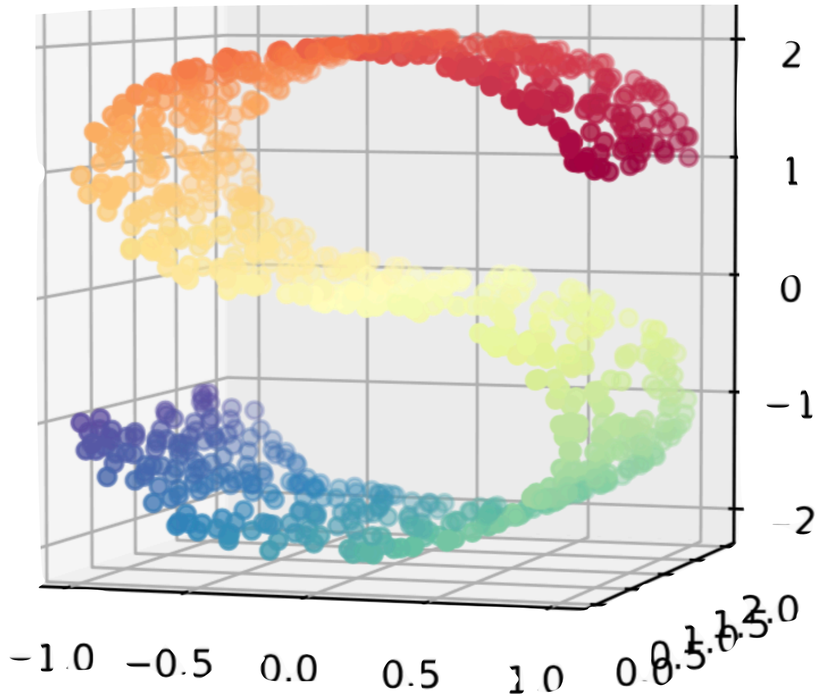


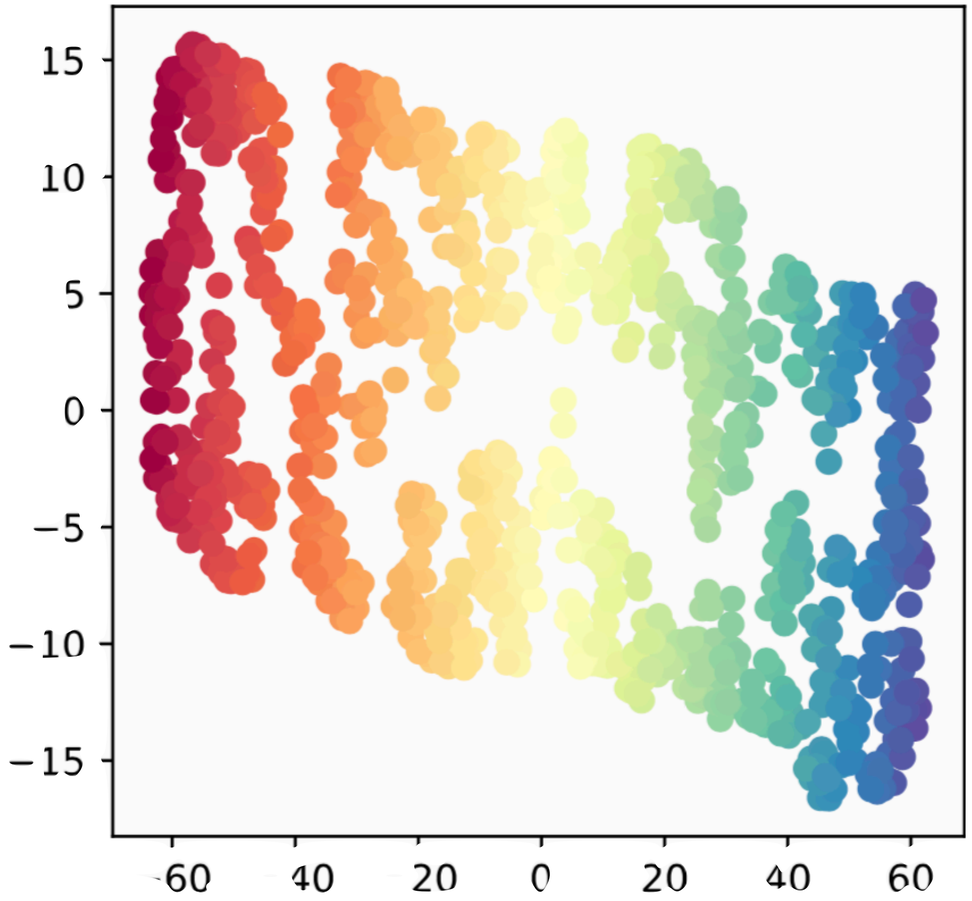
http://gp.snad.space

BRI → gri



Dimensionality Reduction: t-SNE





http://scikit-learn.org

OSC: 10 Feature Sets

1. 364 Gaussian processes approximated points: 3 passbands ✖ 121 points $\in [-20; 100]$ days after peak in r normalised to peak, peak flux

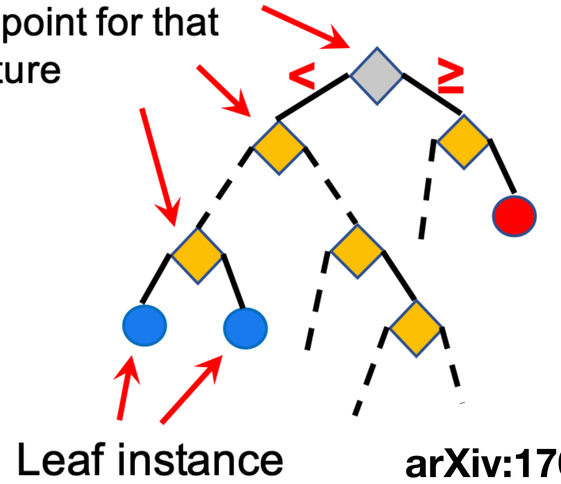
2. 10 parameters of Gaussian process fit: 6 values of correlation matrix, 3 lengths of kernels, likelihood

3. Nine datasets obtained by reducing 374 Gaussian process features to 2–9 t-SNE dimensions

Isolation Forest

Isolation Tree

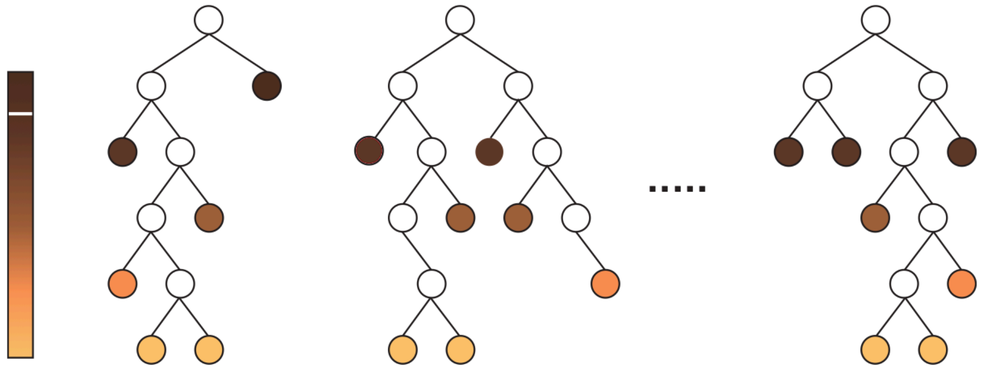
Select a random feature at each node, and a random split point for that feature

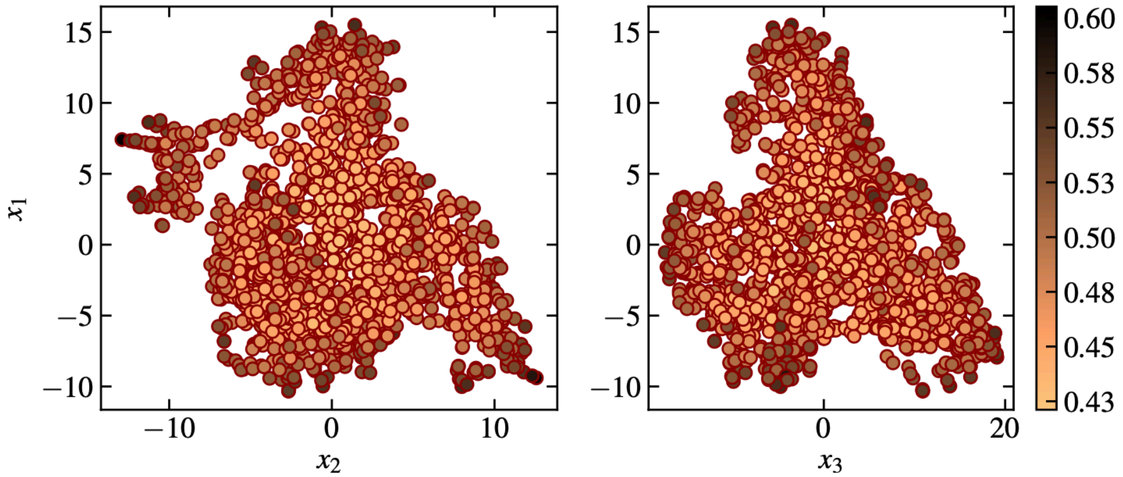


Leaf instance

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

arXiv:1708.09441

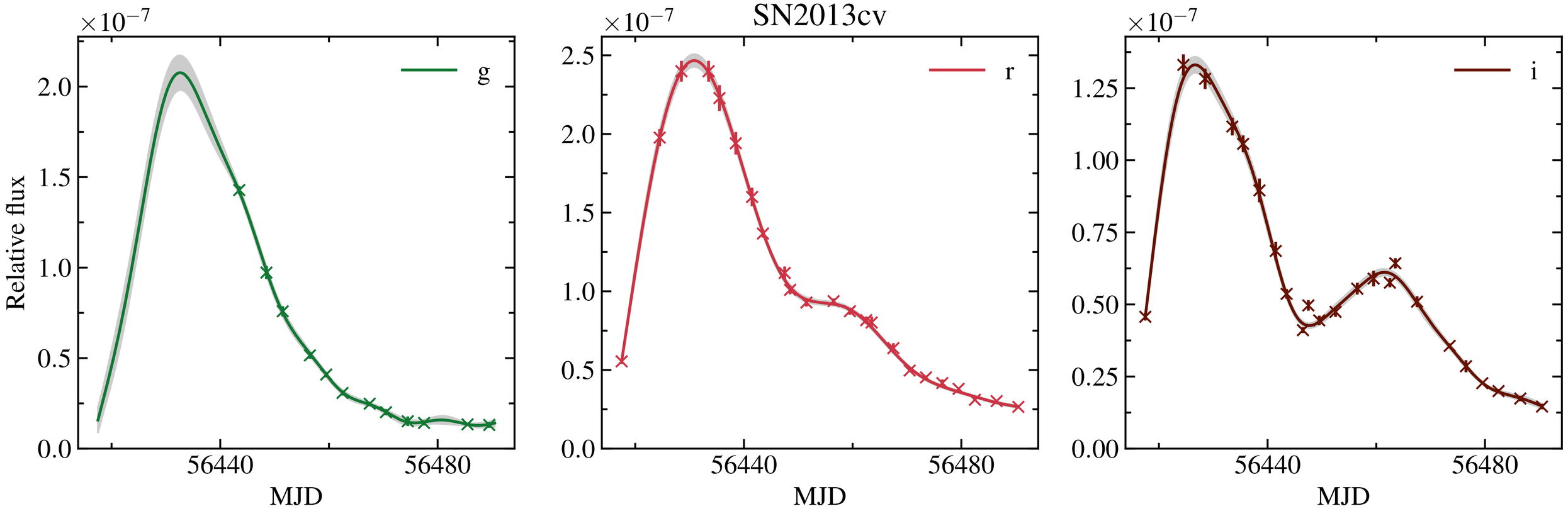




arXiv:1905.11516

OSC: Results

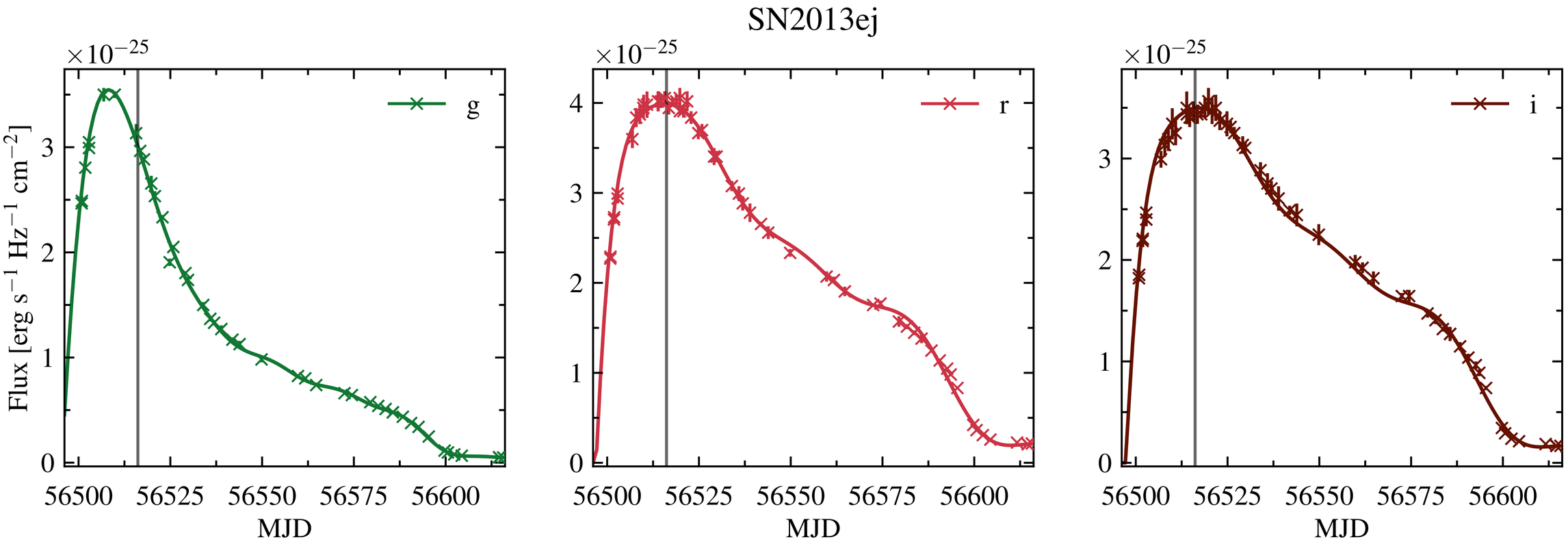
Ia 91T-like (pecular supernovae)



arXiv:1905.11516

OSC: Results

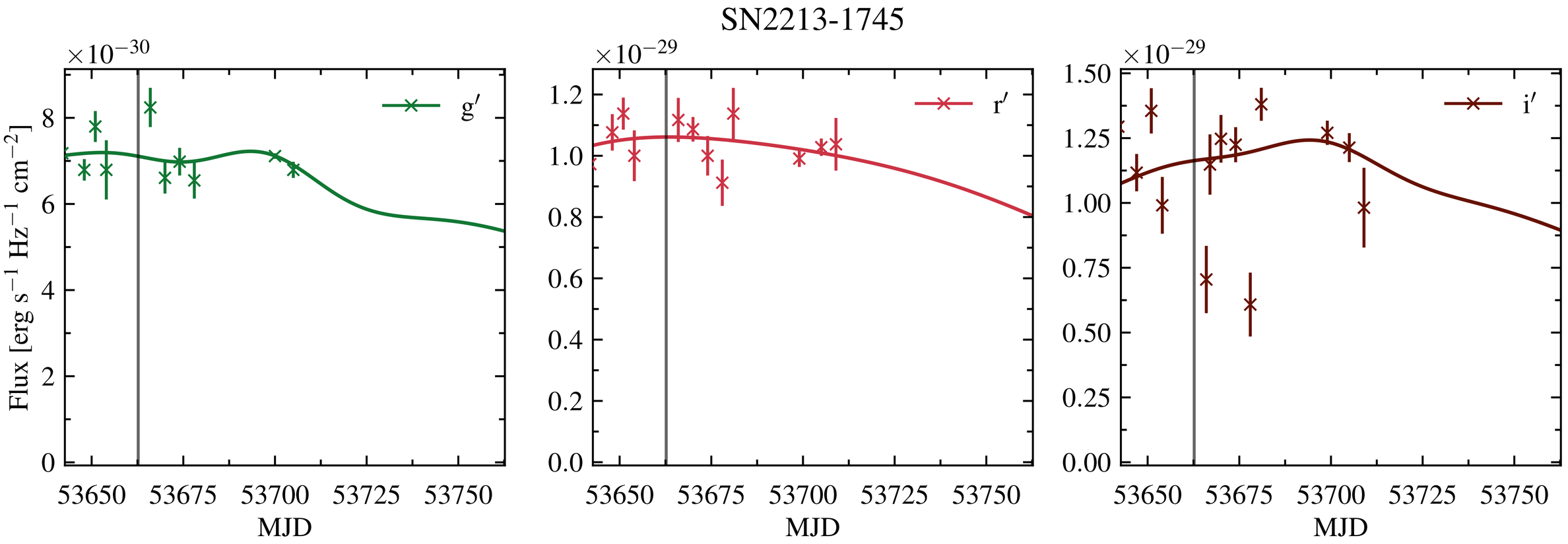
Pecular type II supernovae



arXiv:1905.11516

OSC: Results

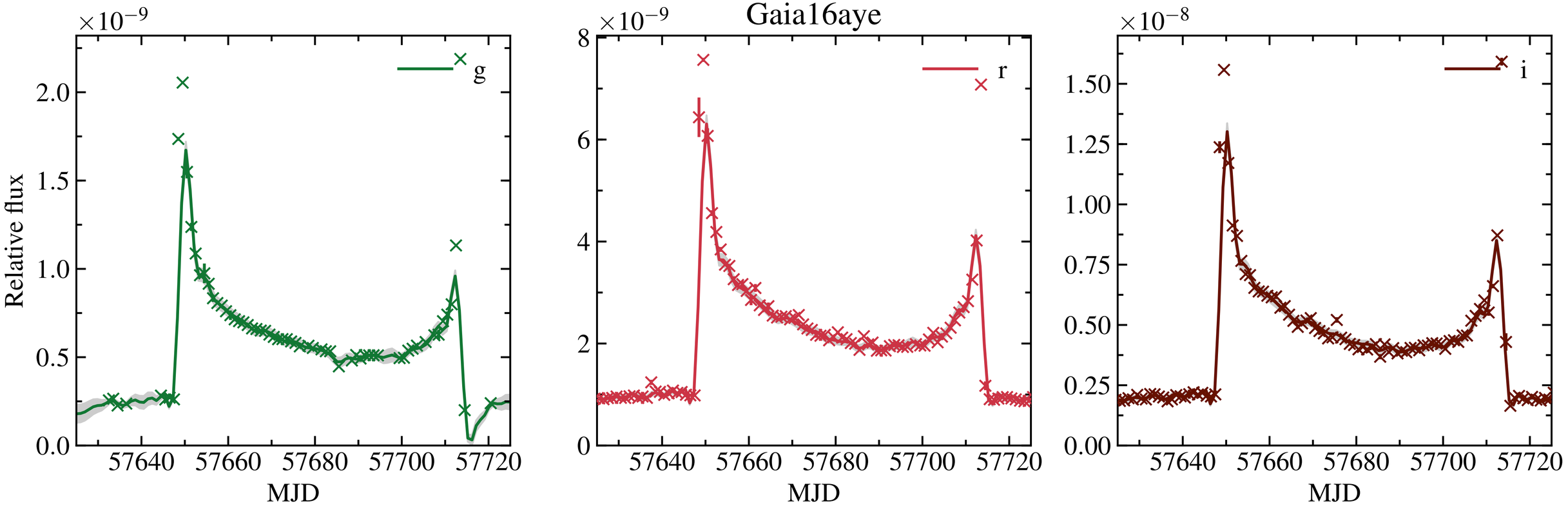
Super-luminous supernovae



arXiv:1905.11516

OSC: Results

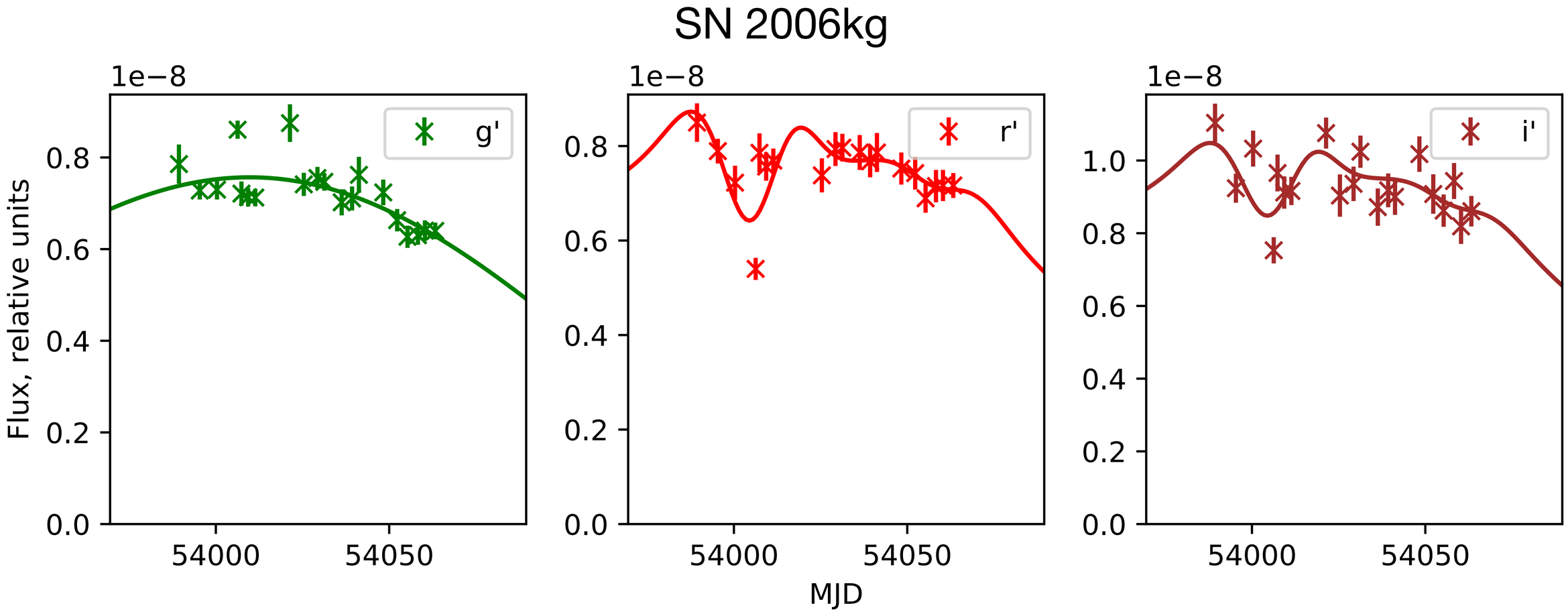
Binary microlensing



arXiv:1905.11516

OSC: Results

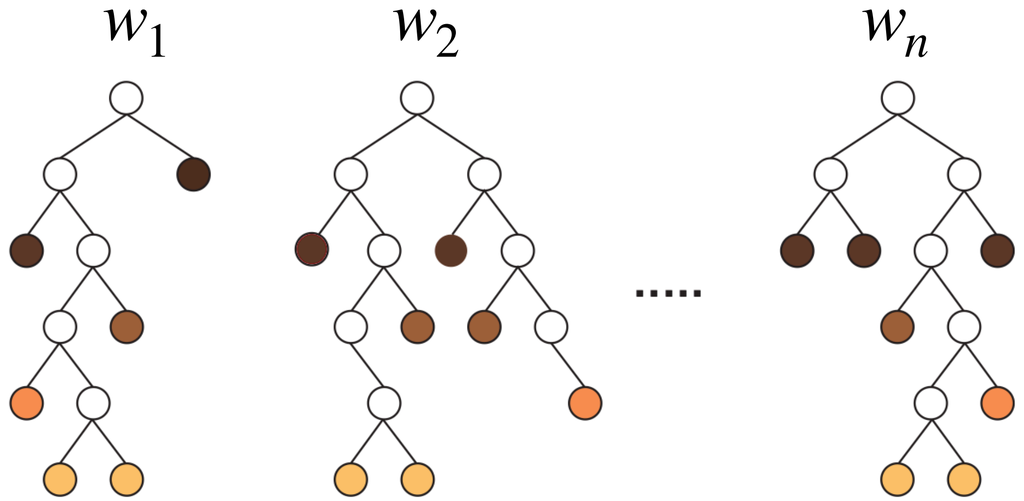
Misclassification of SDSS objects: 10 stars, 6 AGNs



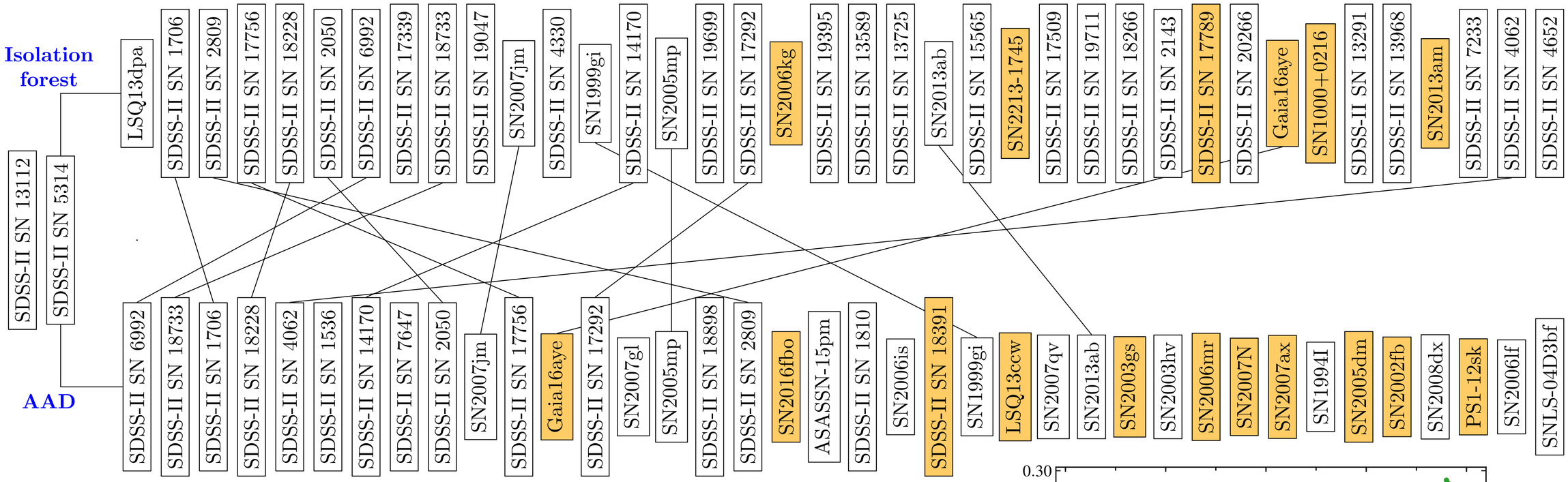
arXiv:1905.11516

Active Anomaly Detection

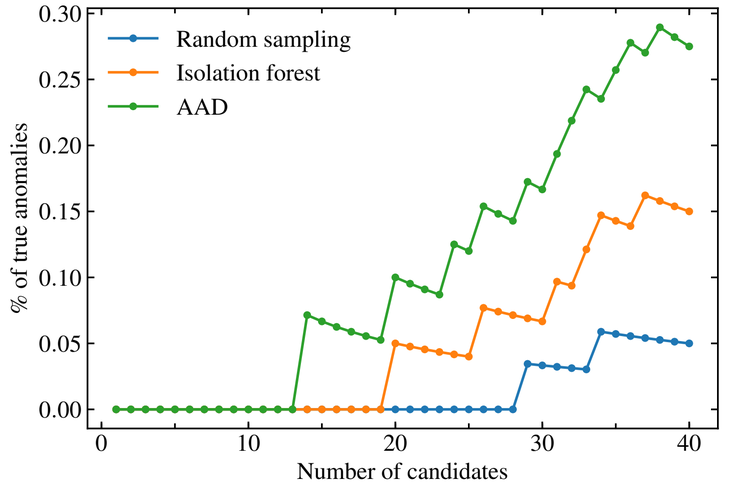
1. Initialize isolation forest or other ensemble of anomaly detectors, set equal w_i to each detector
2. Ask the ensemble for the outlier with the largest score
3. Ask an expert to classify the object as normal or anomaly
4. If anomaly, go to step 2 and ask next outlier
5. If normal, reweight detectors to give lower influence to wrong detectors, go to step 2



AAD: OSC Case



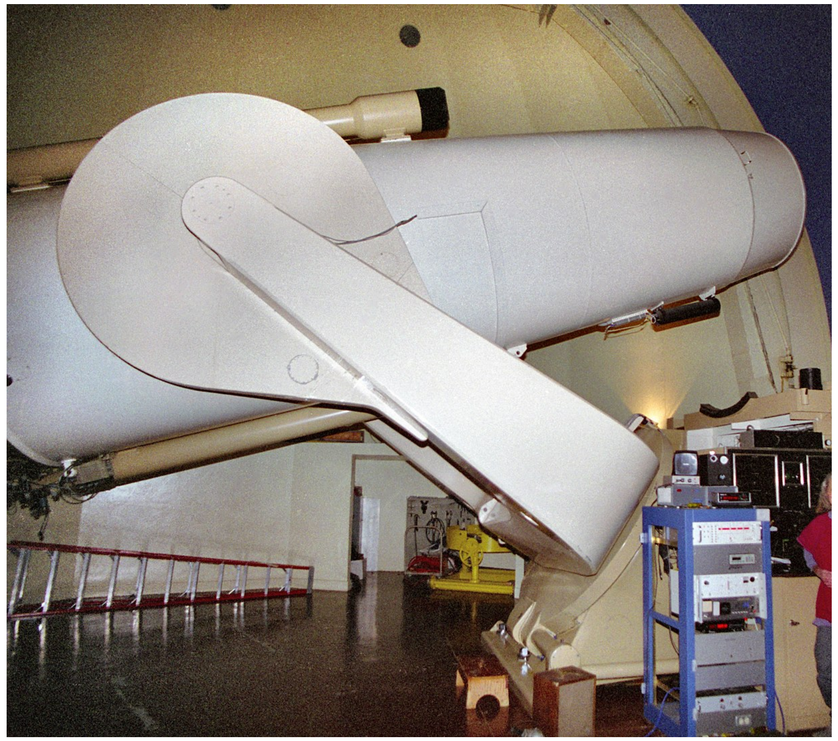
Anomaly



arXiv:1909.13260

Zwicky Transient Facility DR1

- Full light curve catalog contains ~$1.6 \cdot 10^9$ of "objects" in *g* & *r* collected in 284 days
- Mostly Galactic objects, detected in near-real-time extragalactic transients are excluded
- Raw data are 2 TB, our PostgreSQL database has ~$5 \cdot 10^{11}$ rows and occupies 4 TB
- We use objects observed in *r* with at least 100 points covering at least 200 days. Totally ~$8 \cdot 10^7$ light curves



Samuel Oschin 48-inch Schmidt telescope

ZTF DR1: Feature Extraction

Totally three dozens of features are used

- 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, periodogram shape based features

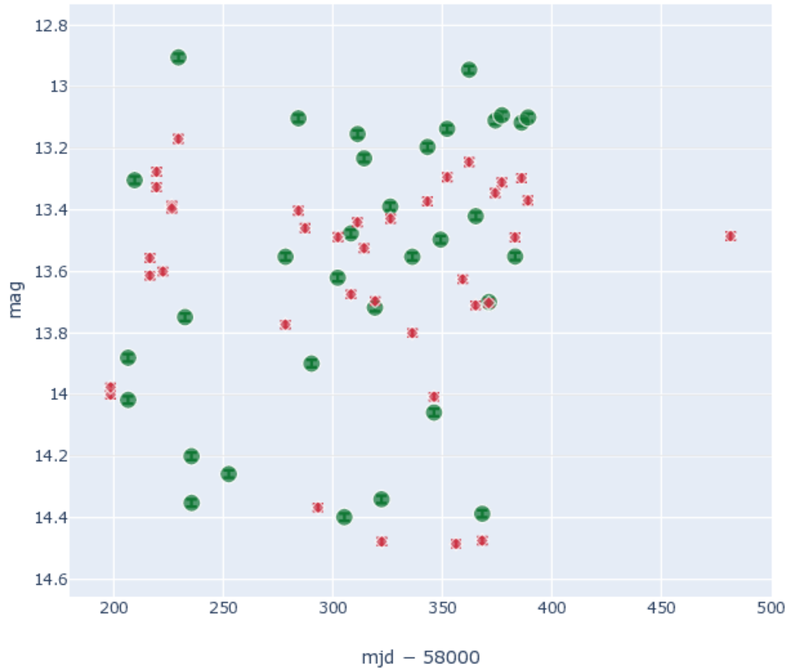
ZTF DR1: Object Viewer

ZTF object viewer

oid GO

Coordinates or obje 1 GO

680113300005170



Metadata

nobs: 36

ngoodobs: 33

filter: zg

coord_string: 254.45753, 35.34235

duration: 182.660

fieldid: 680

rcid: 50

GCVS

10 search radius, arcsec

Designation	Separation, arcsec	Period, days	Type of variability	Spectral type
HZ Her	0.740	1.700	XPR+E	B0Ve-F5e

AAVSO VSX

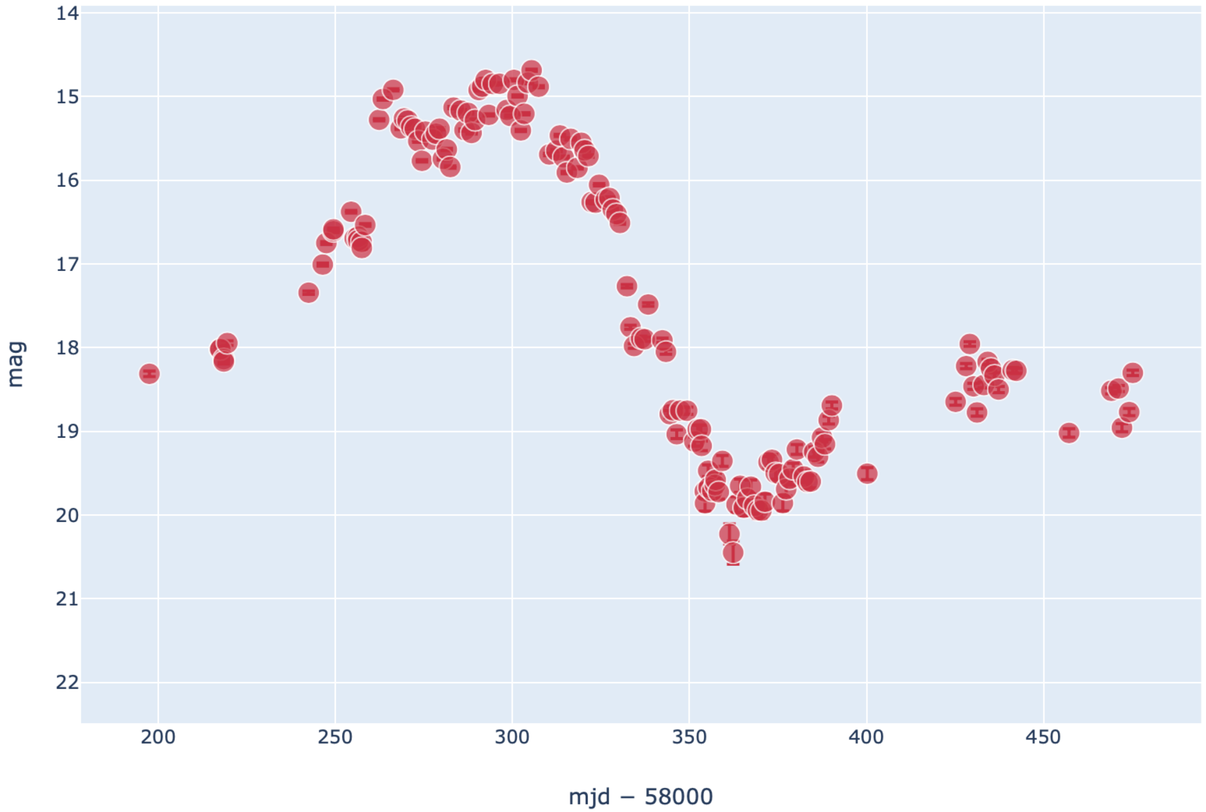
10 search radius, arcsec

Designation	Separation, arcsec	Name	Period, days	Variability type	Maximum mag	Band of max mag	Minimum mag	Band of min mag
15037	0.748	HZ Her	34.875	LMXB:/XPR+E	12.800	B	15.200	B

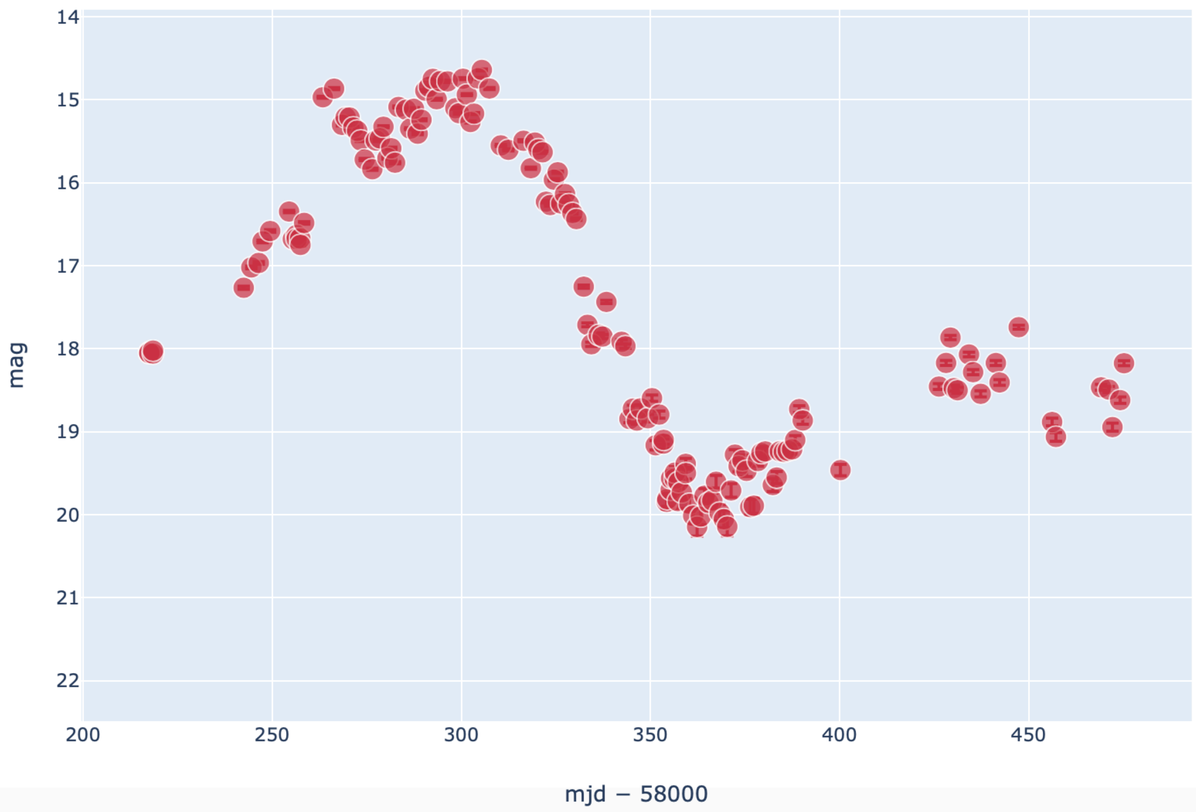
http://ztf.snad.space

ZTF DR1: Preliminary Results

830213400008915



831216400016457

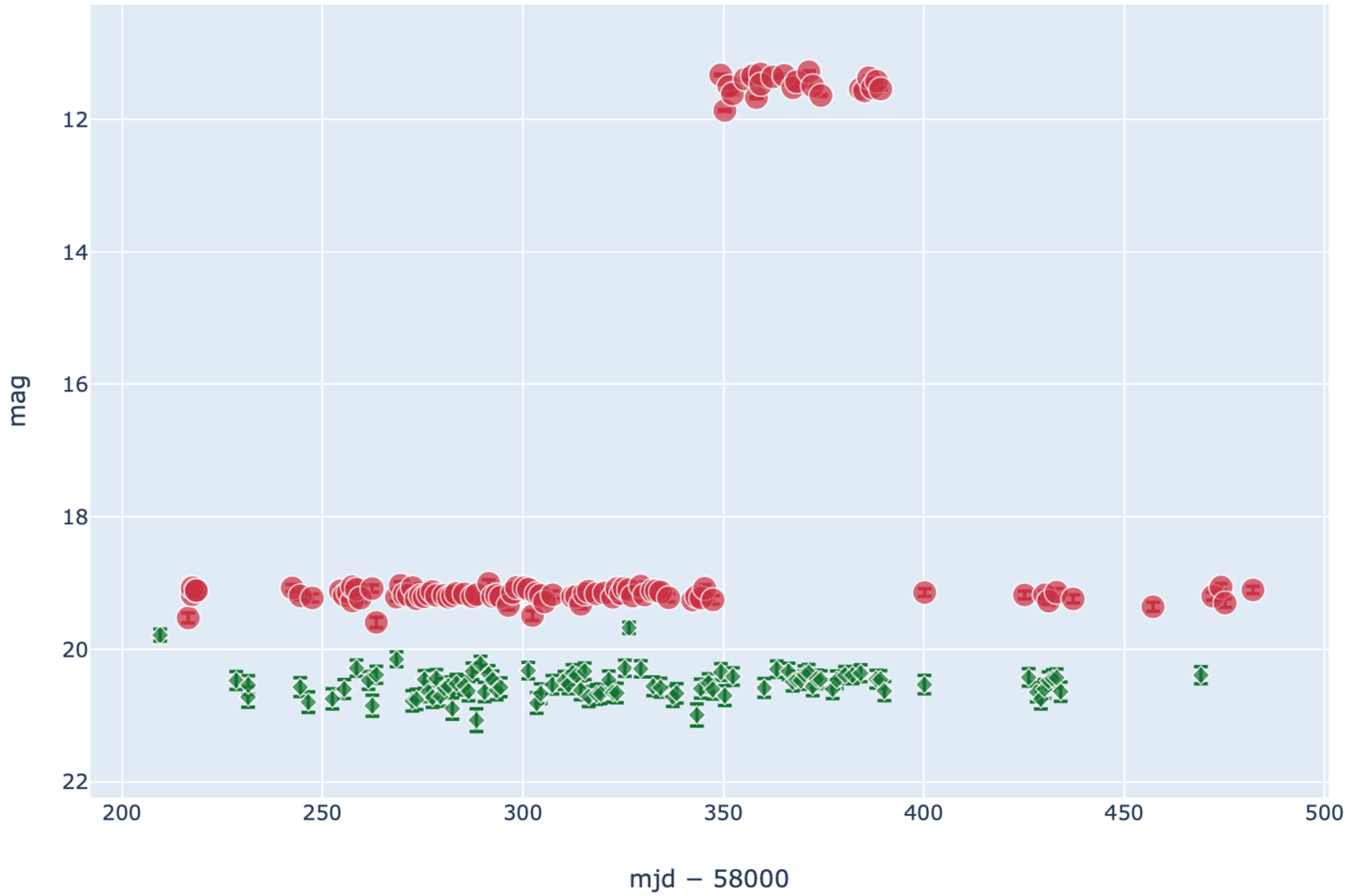


ZTF DR1: Preliminary Results

768201400047639

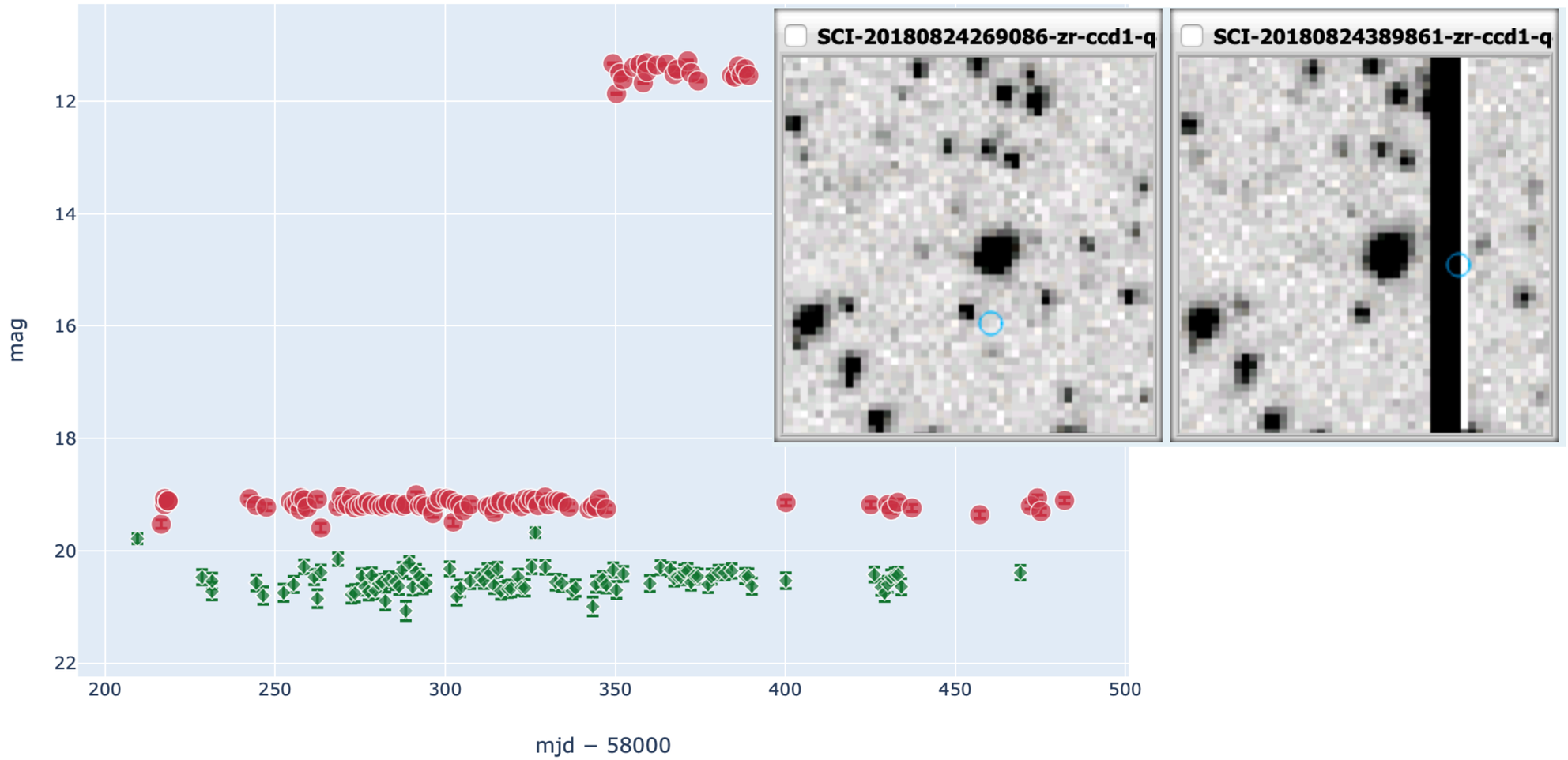
ZTF DR1: Preliminary Results

768201400047639



ZTF DR1: Preliminary Results

768201400047639



Thank you for your attention