SEARCHING FOR DIFFERENT AGN POPULATIONS IN MASSIVE DATASETS WITH MACHINE LEARNING

Paula Sánchez Sáez (on behalf of the ALeRCE team)

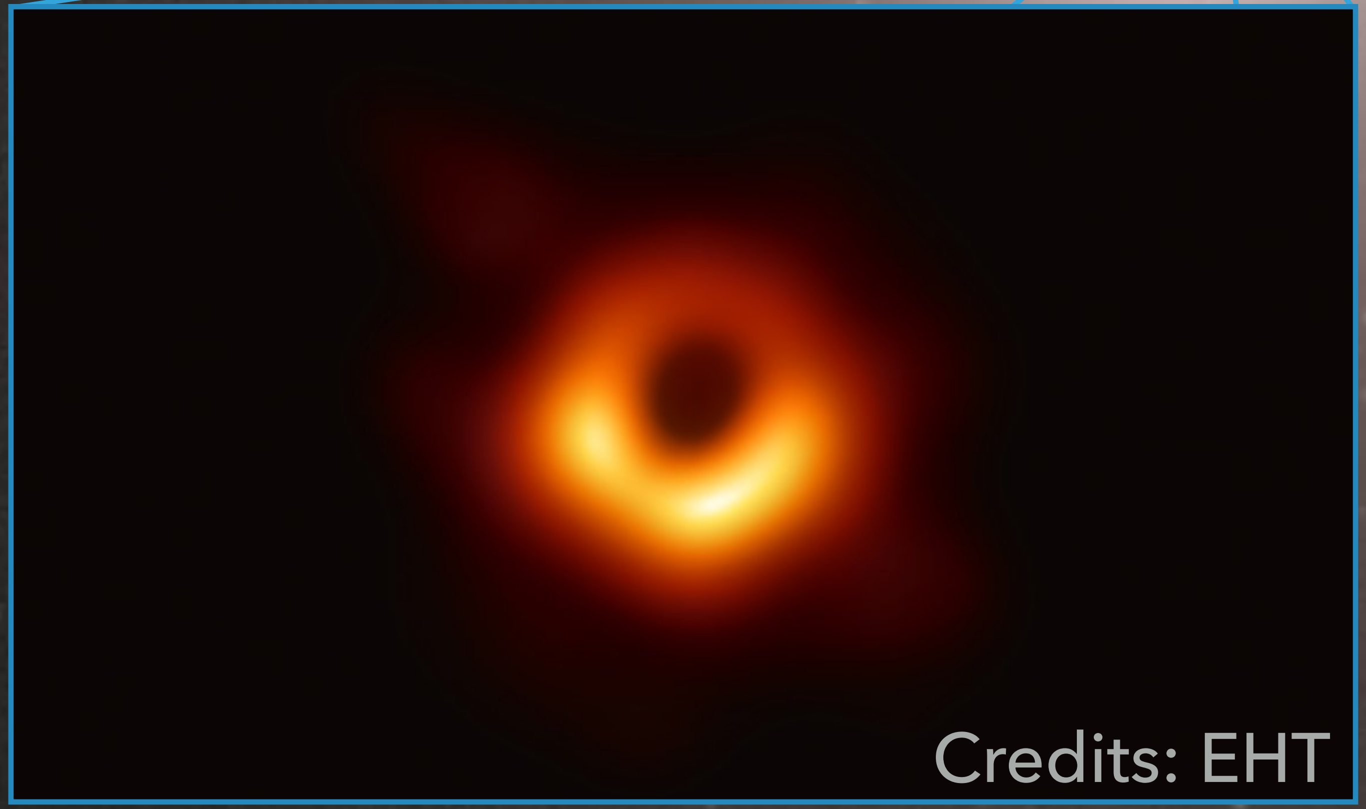
ESO Fellow (Garching)

ALeRCE
Automatic Learning for the Rapid Classification of Events

INSTITUTO MILENIO DE ASTROFÍSICA

Active Galactic Nuclei (AGN)

AGNs are powered by the release of gravitational energy related with the accretion of material onto a supermassive black hole (SMBH), with masses larger than $10^6\,M_\odot$

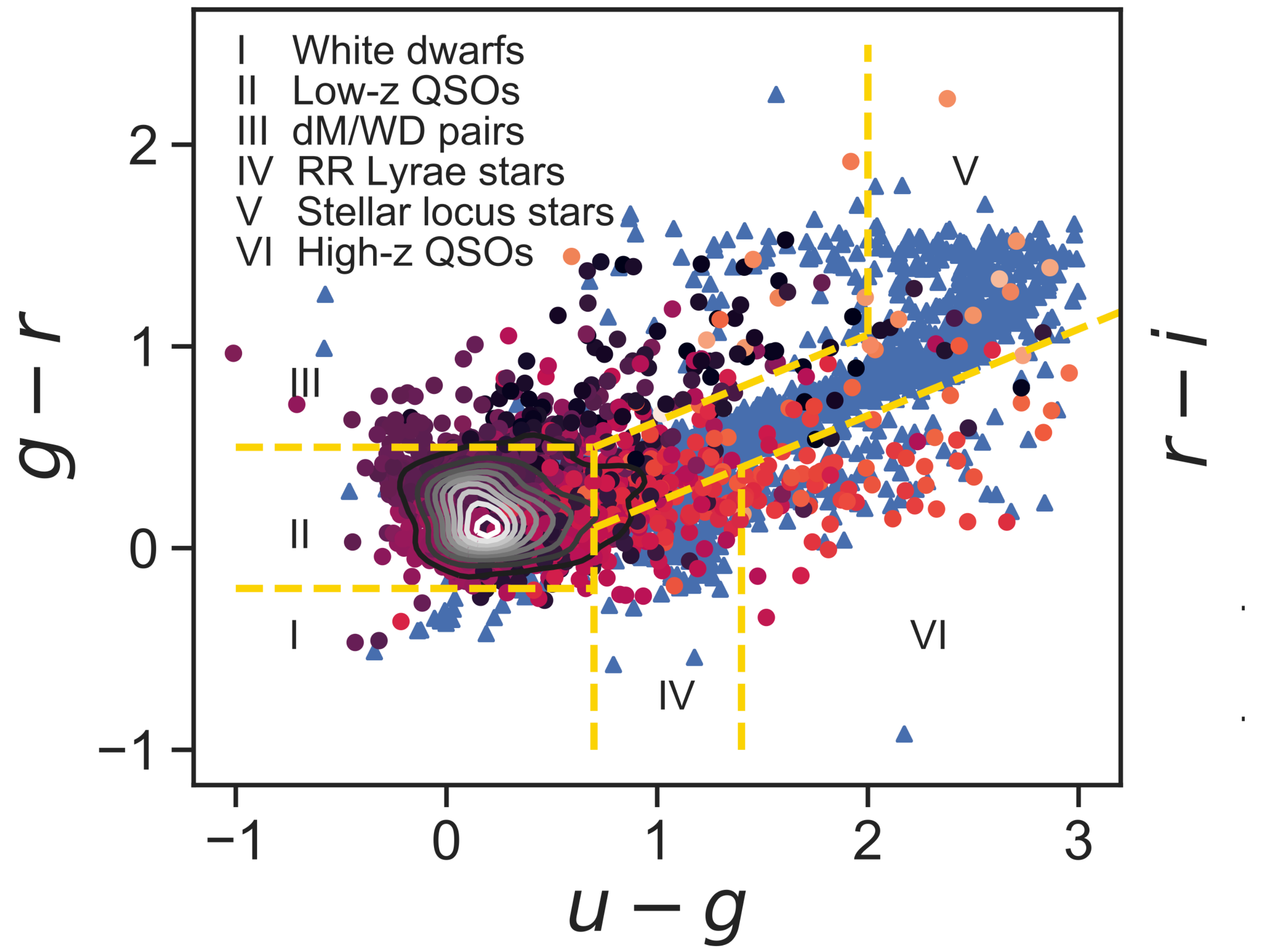


Credits: EHT

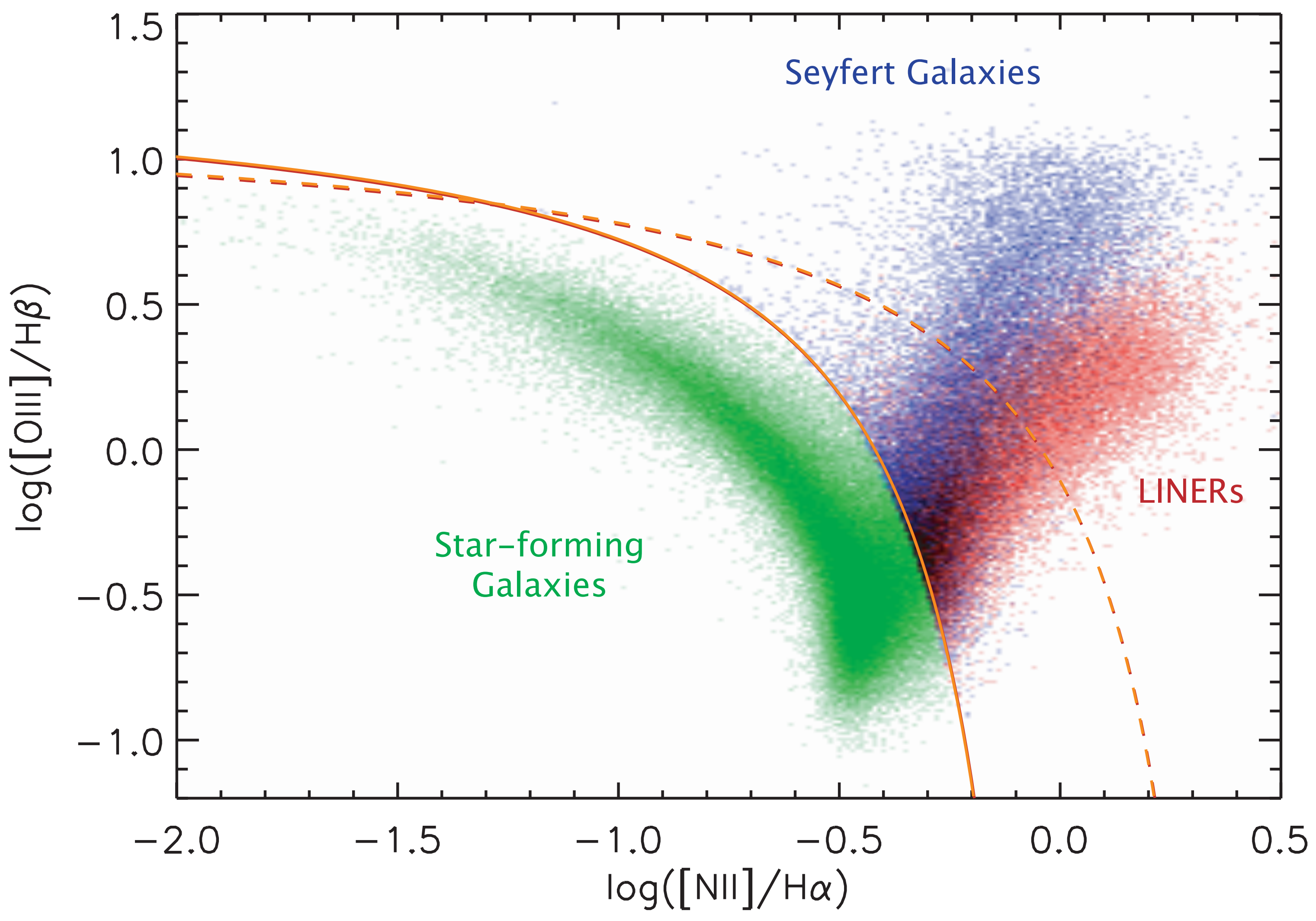
M87, Credits: ESO

Traditional selection of AGN candidates

Color selection

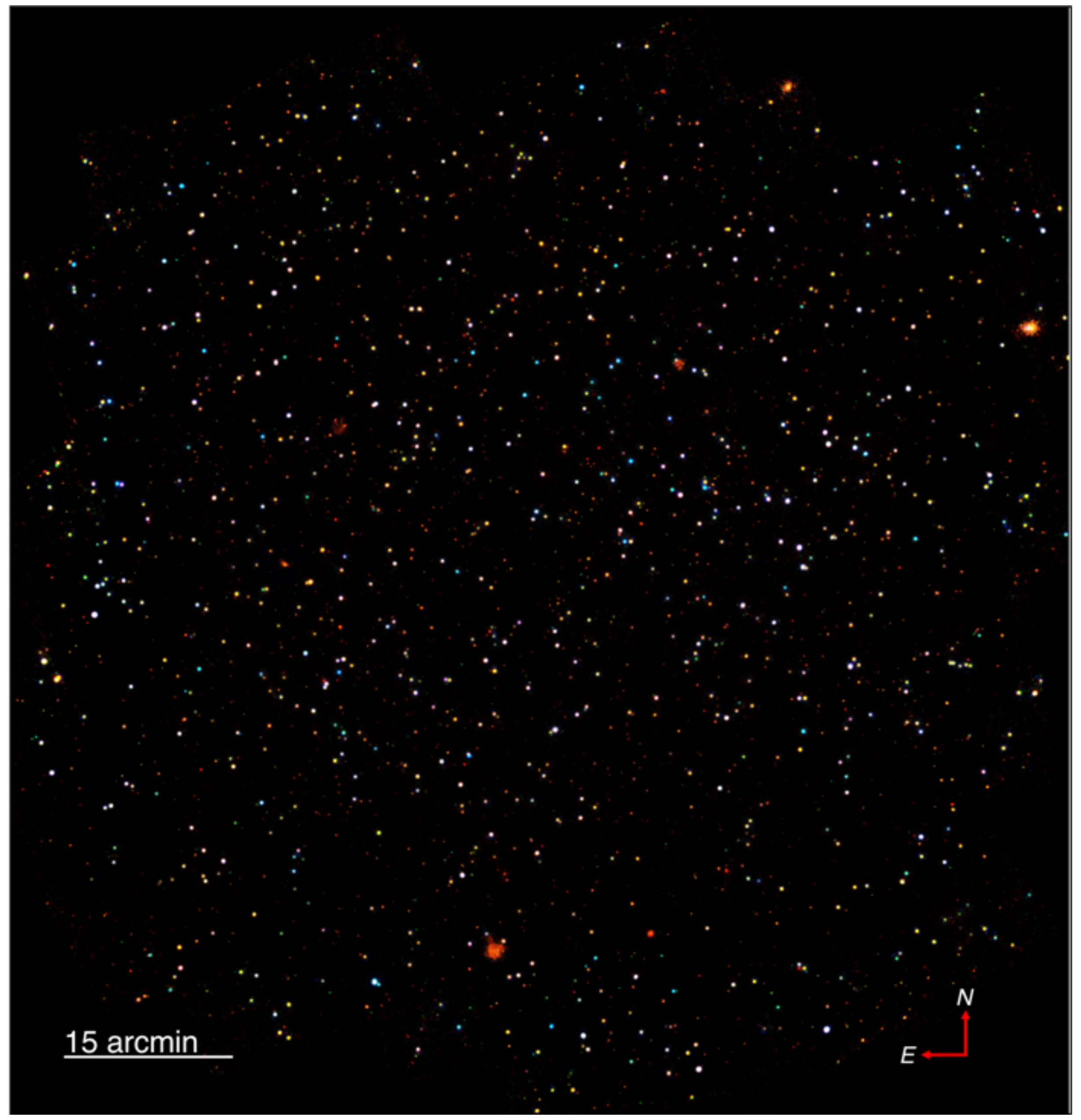


Line-ratio diagnostic (BPT) diagram



Fosbury+2007

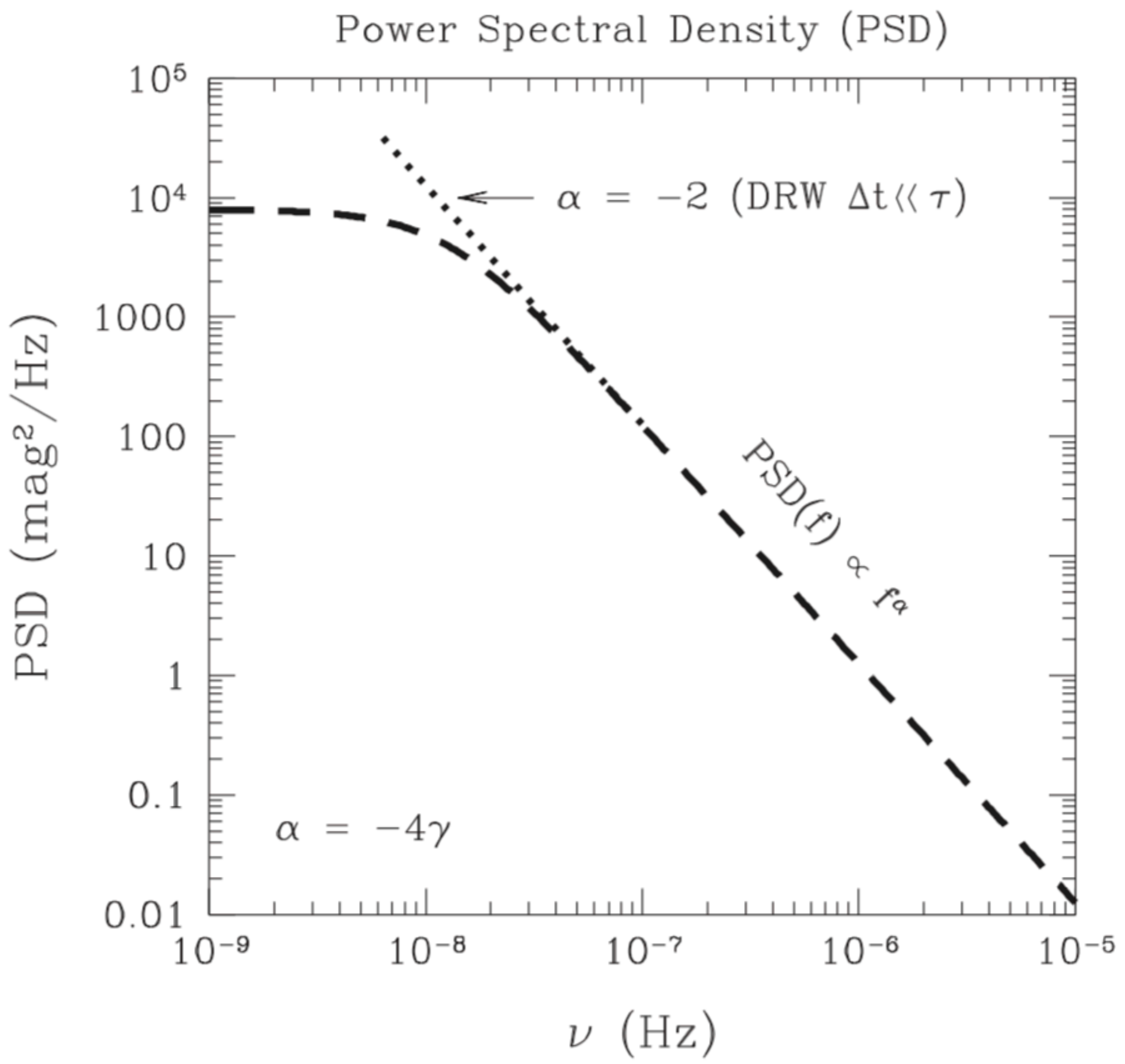
X-ray selection



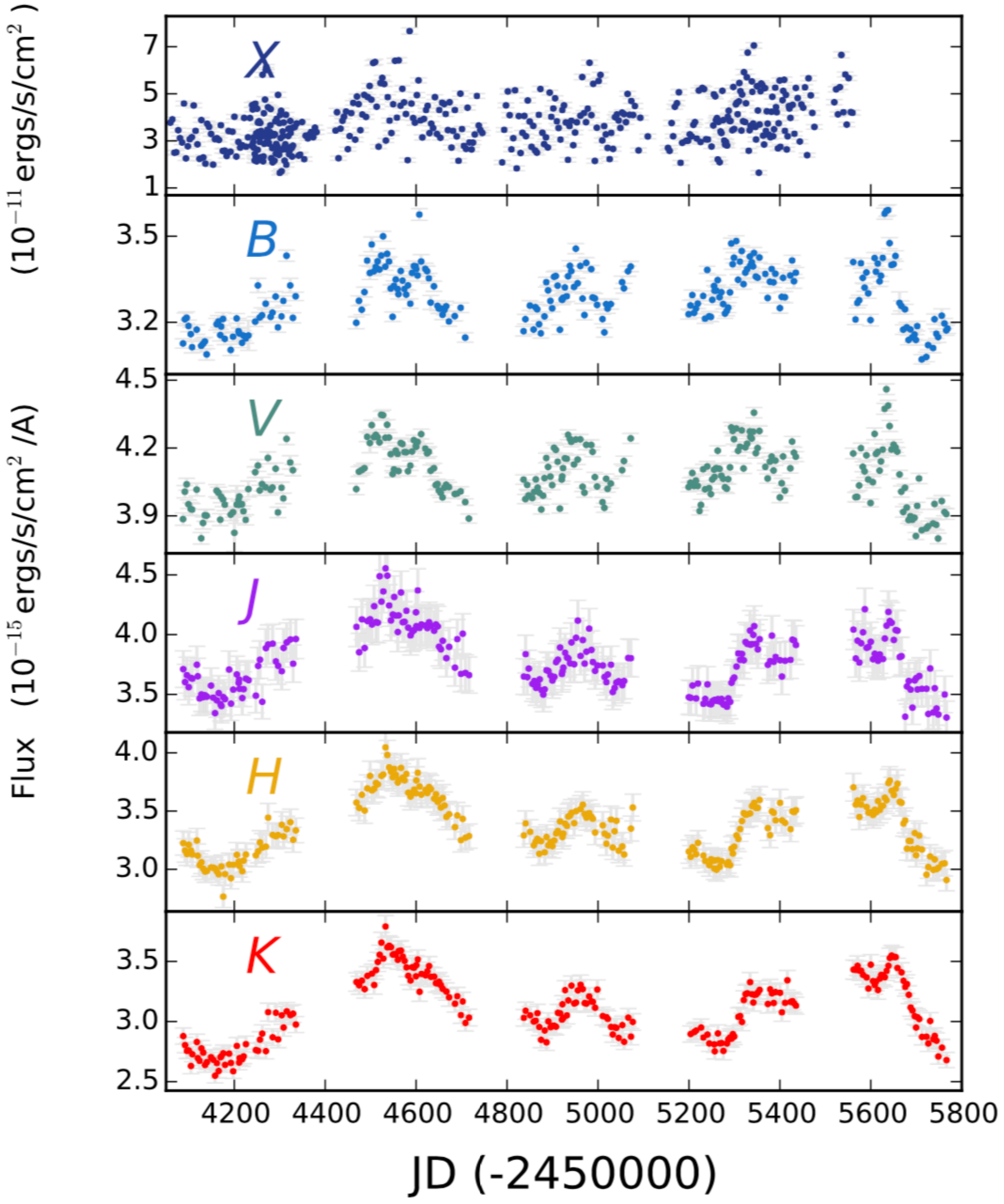
Civano+2016

AGN are variable

- AGN variability seems to be well described as a stochastic process.
- The characteristic time-scales of the variability range from hours to years, with the shortest time-scales being associated with shorter emission wavelengths.



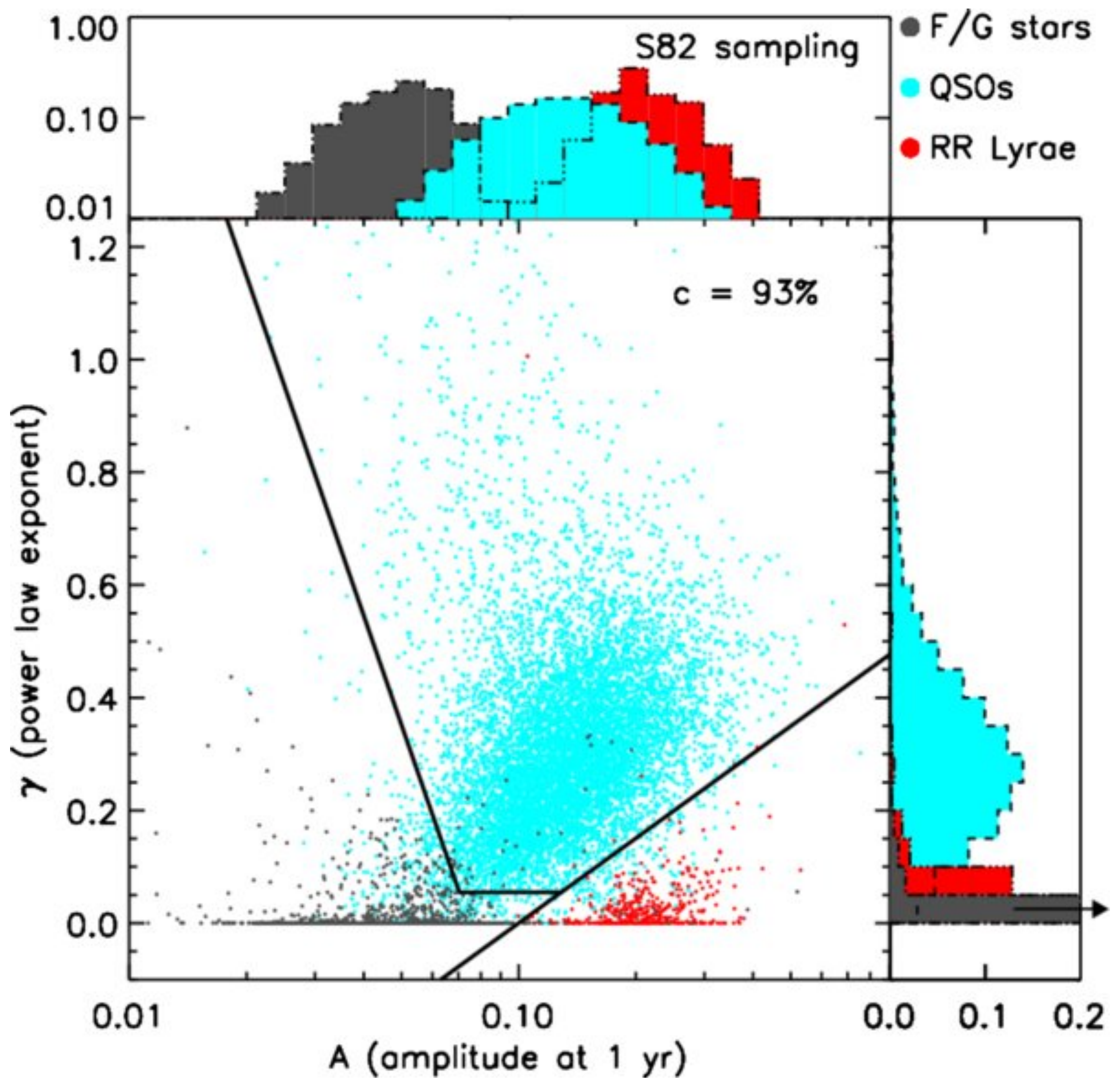
Kozlowski 2016



MCG-6-30-15 from Lira et al. 2015

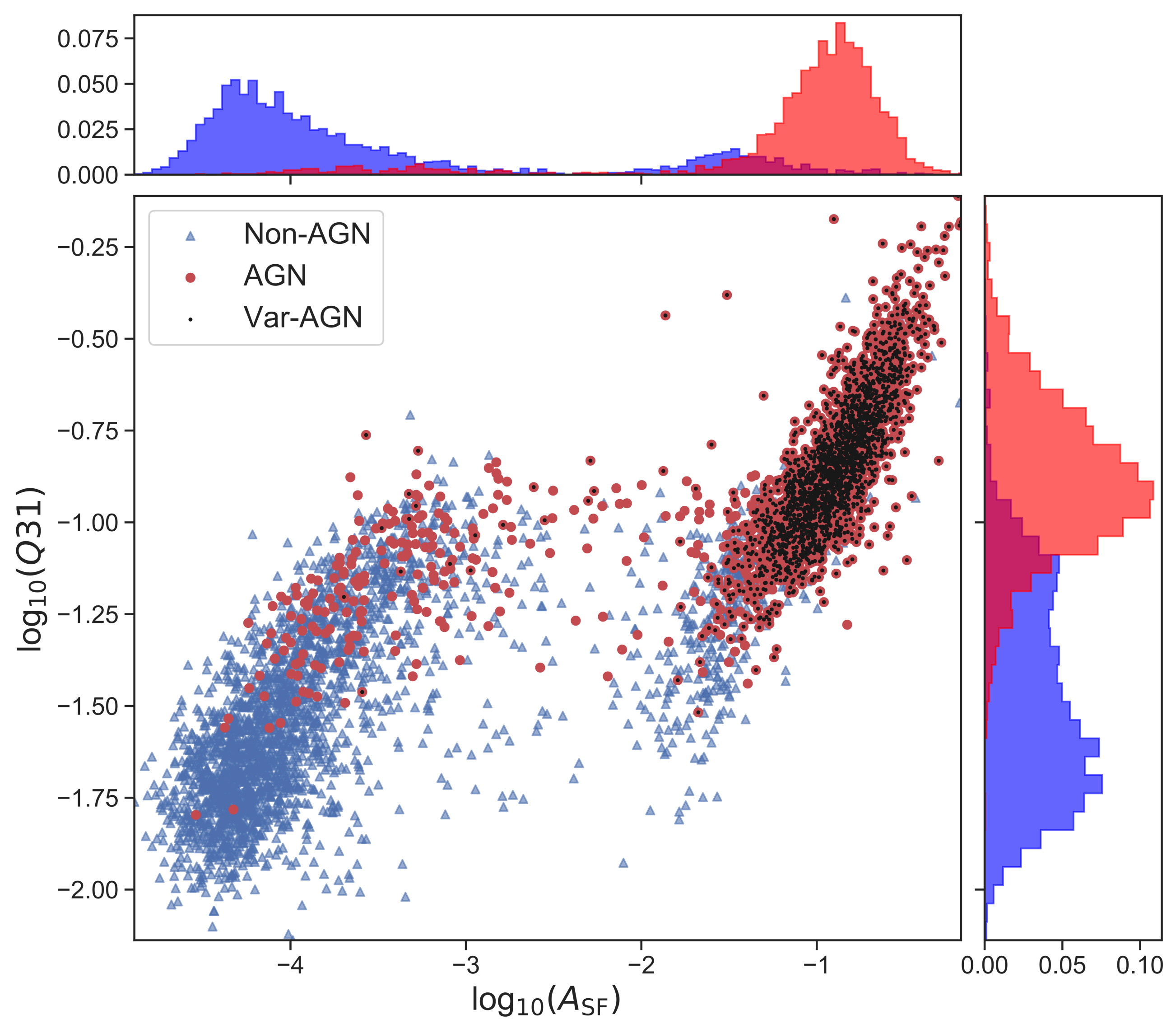
Selection of AGN candidates with variability-based methods

Selection with pre-defined cuts

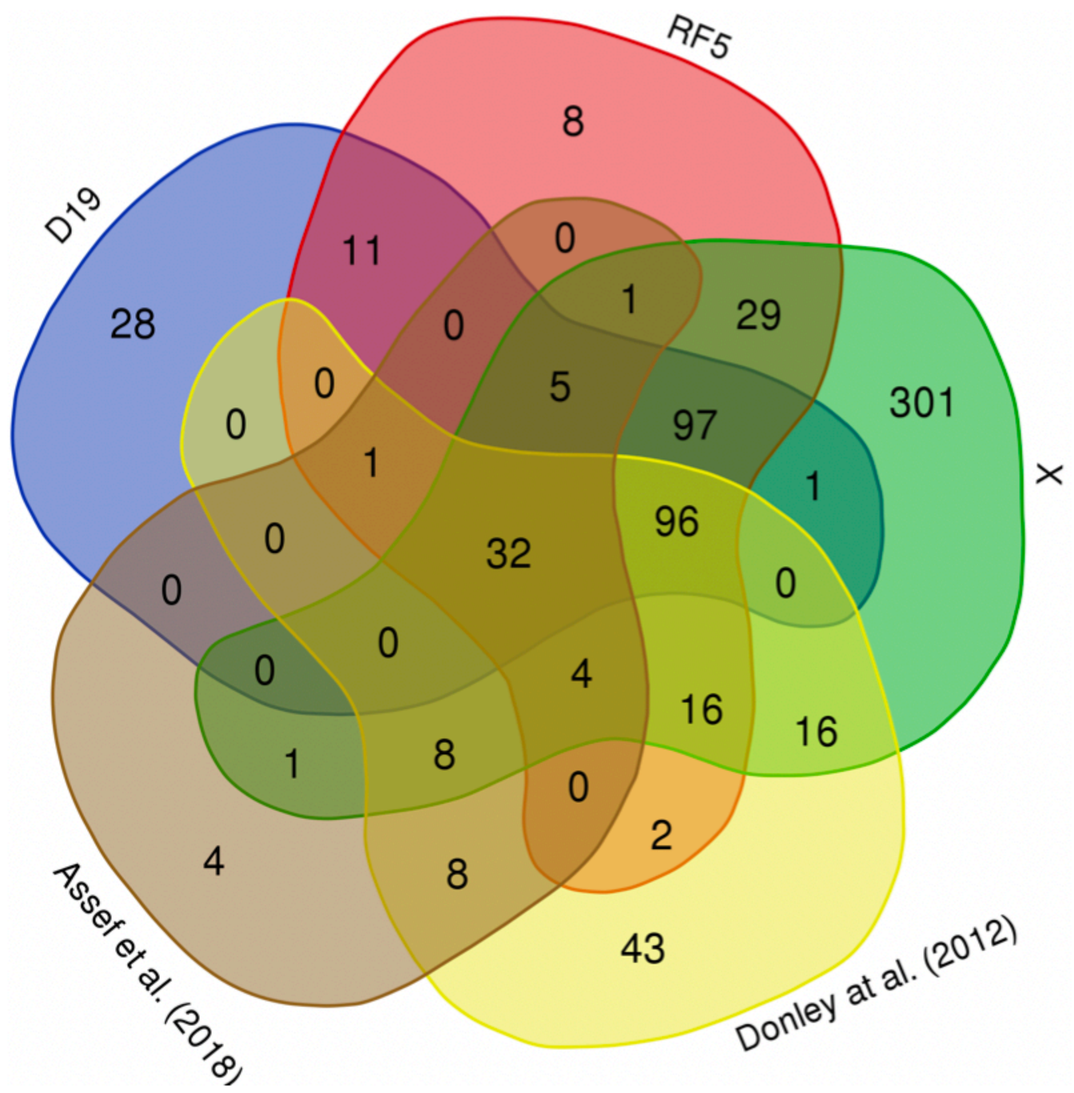


e.g., Schmidt+2010

Selection using ML techniques



Sánchez-Sáez+2019



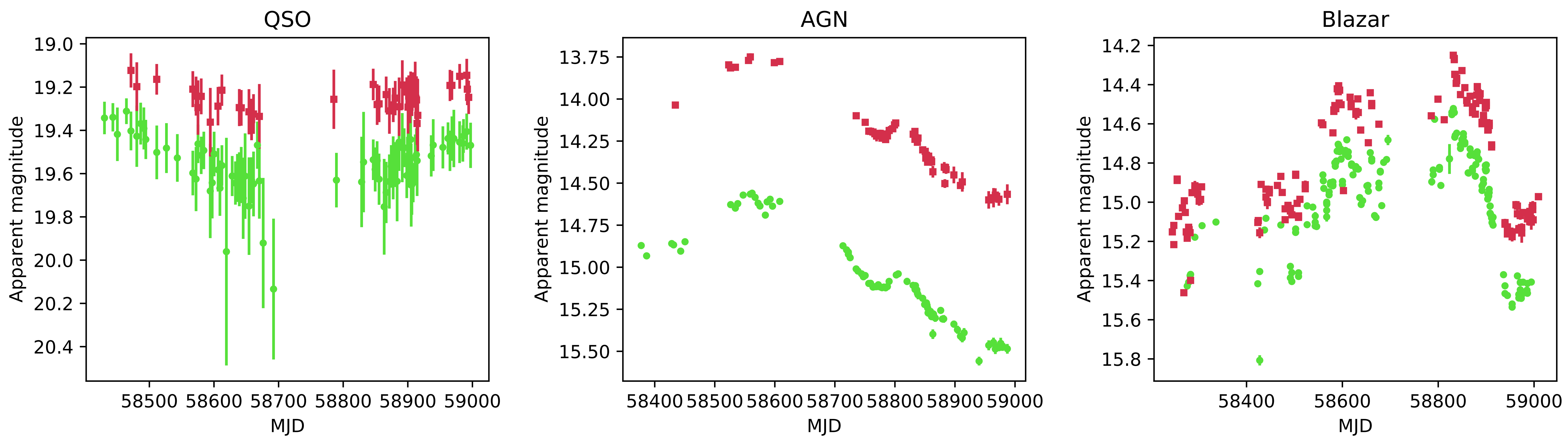
De Cicco+2021

Selection of different AGN populations with variability-based methods

QSO: Bright-Blue and core-dominated

AGN: Red-faint and host-dominated

Blazar: extreme variations and jet-dominated

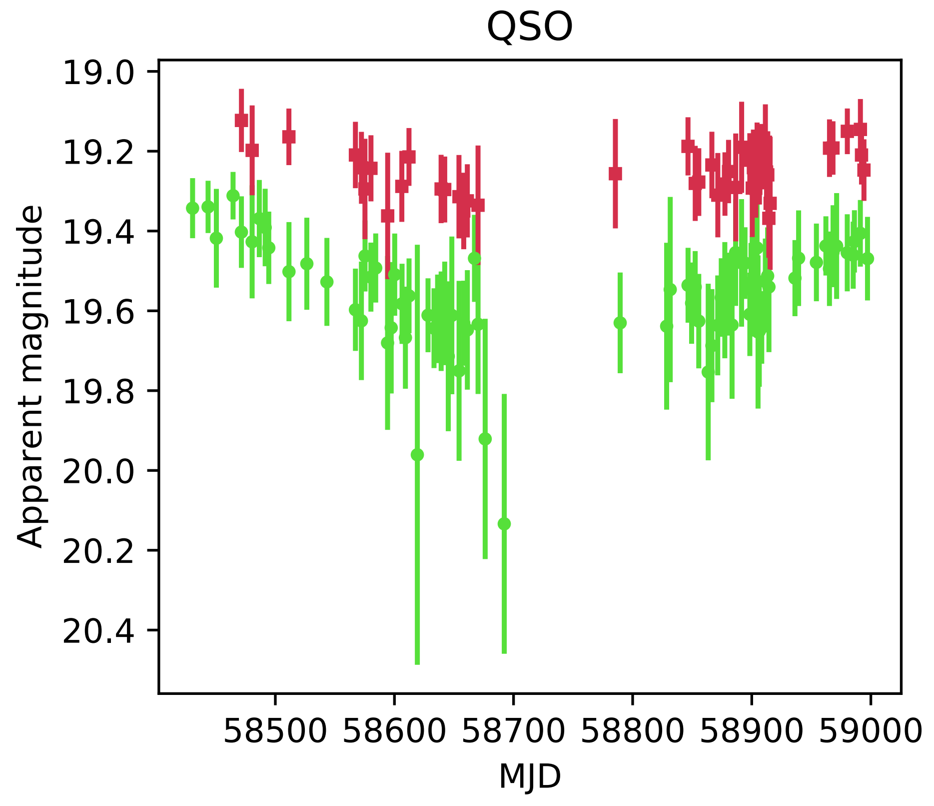


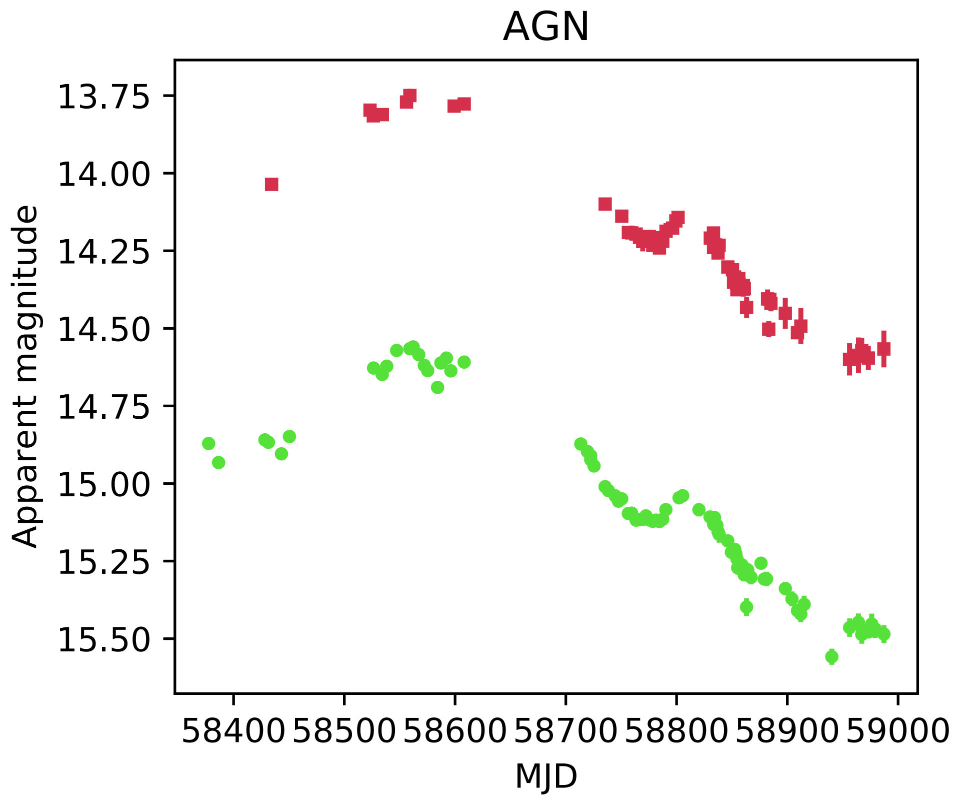
Selection of different AGN populations with variability-based methods

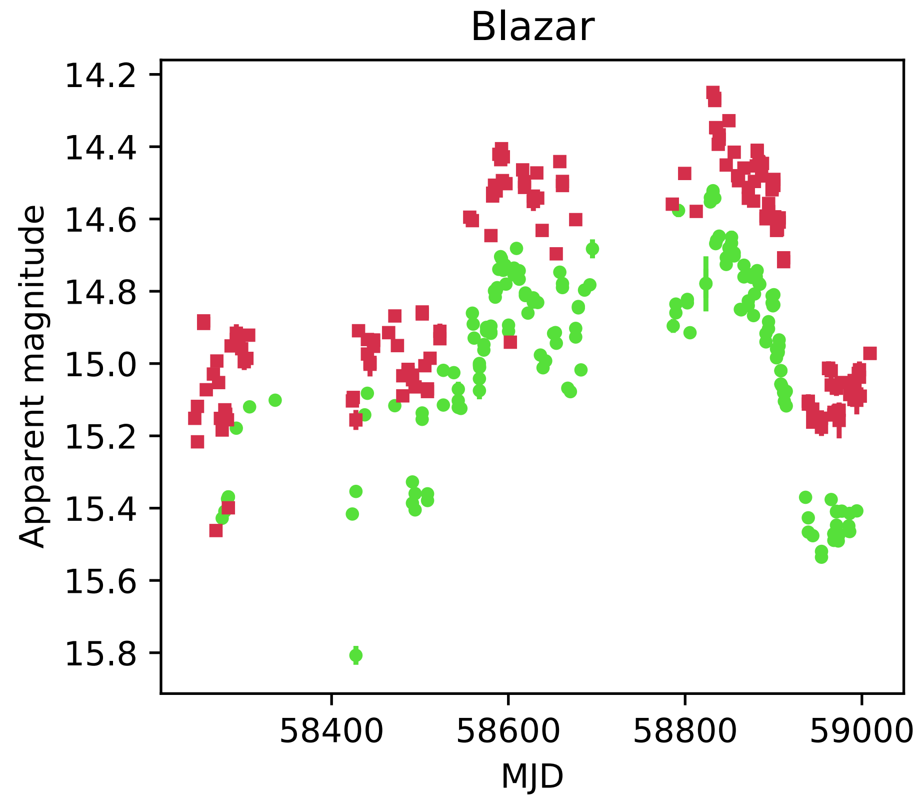
QSO: Bright-Blue and core-dominated

AGN: Red-faint and host-dominated

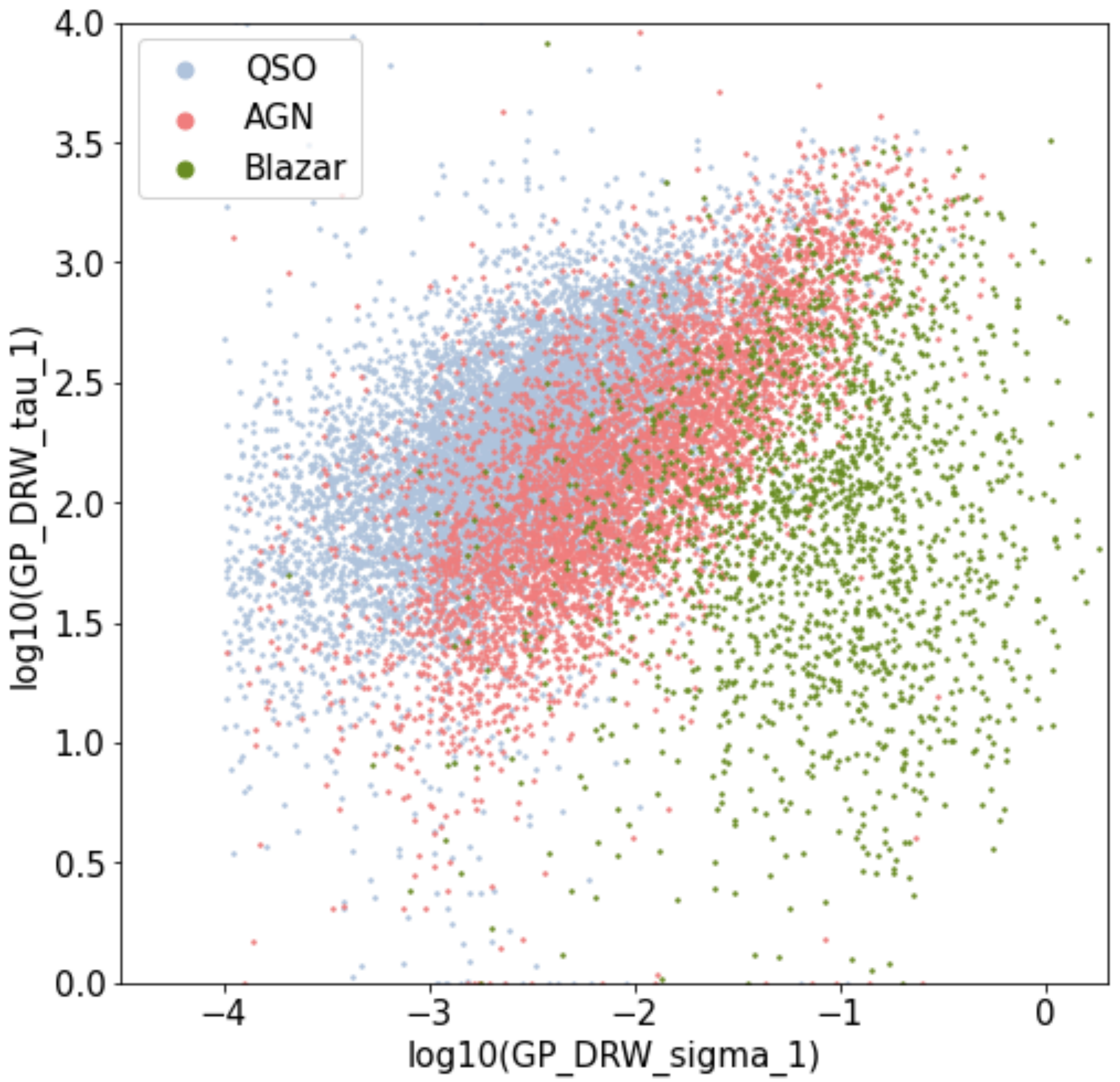
Blazar: extreme variations and jet-dominated







Time scale of the variability



Variability amplitude

1. THE ALERCE BROKER LIGHT CURVE CLASSIFIER

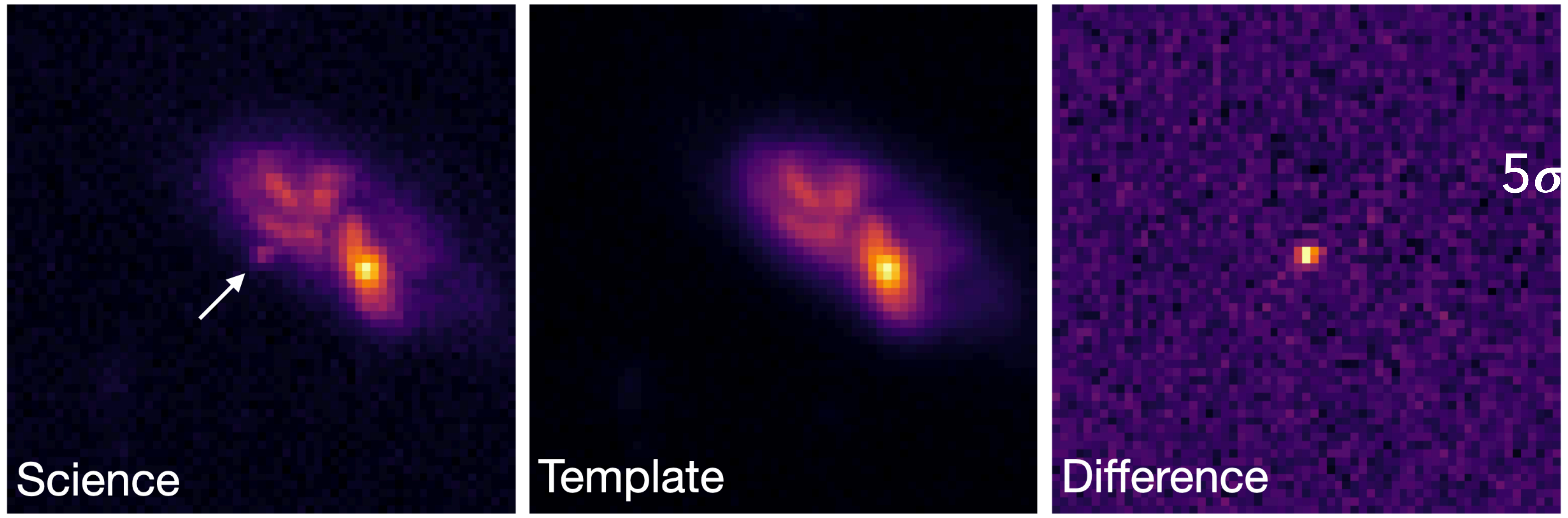
2. SEARCHING FOR CSAGNS WITH ANOMALY DETECTION

1. THE ALERCE BROKER LIGHT CURVE CLASSIFIER

The ALeRCE broker

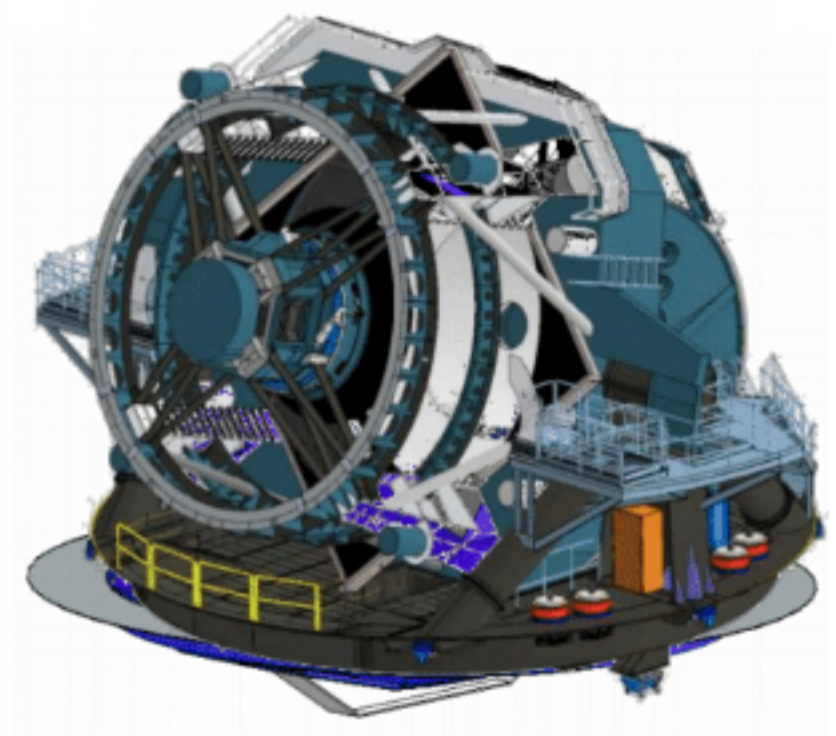
ALeRCE
Automatic Learning for the Rapid Classification of Events

Brokers are astronomical alert processing systems.



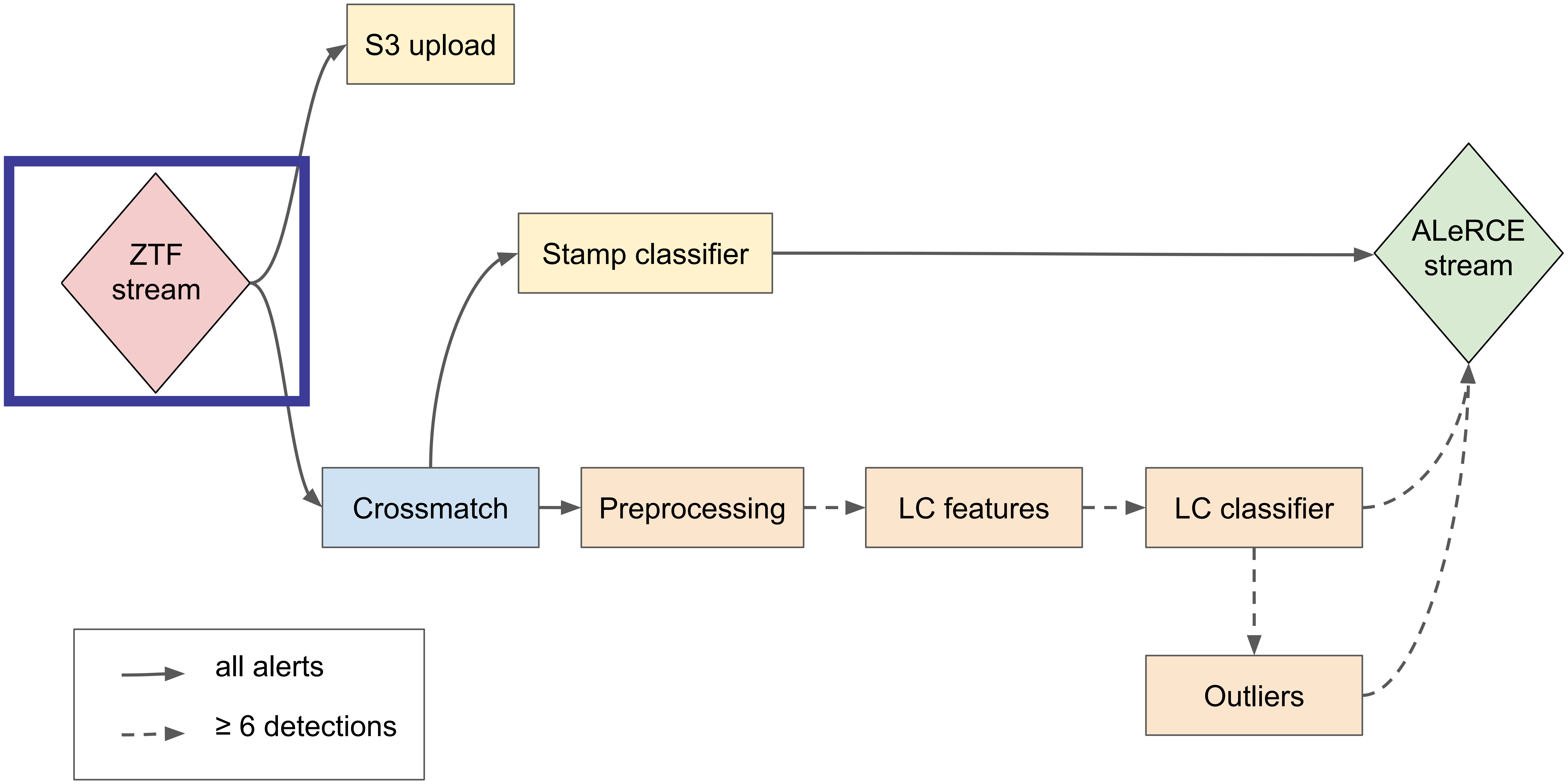
ZTF

VERA C. RUBIN OBSERVATORY



ZTF		LSST
2018-2023	→	2022-2032
1.4 TB per night	**10x**	**15 TB per night**
~1 billion objects		~37 billion objects
~1 trillion measurements		~7 trillion measurements
~1 million alerts per night		**~10 million alerts per night**

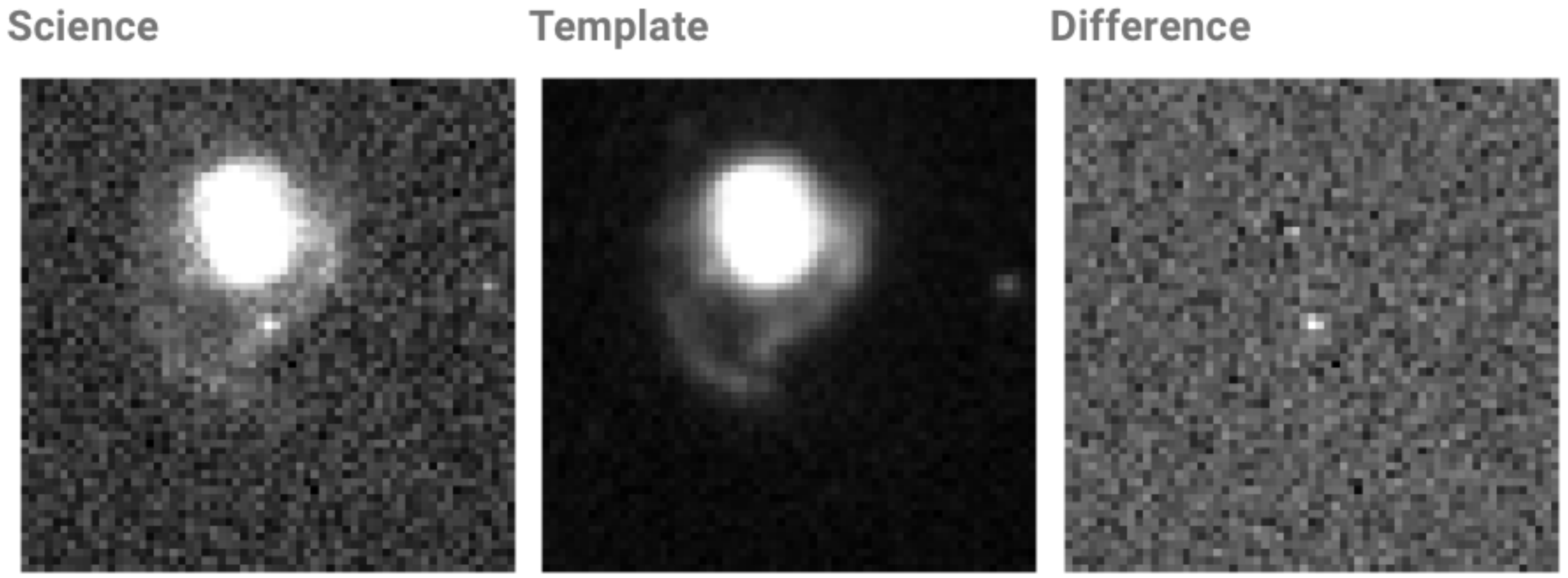
The ALeRCE broker pipeline

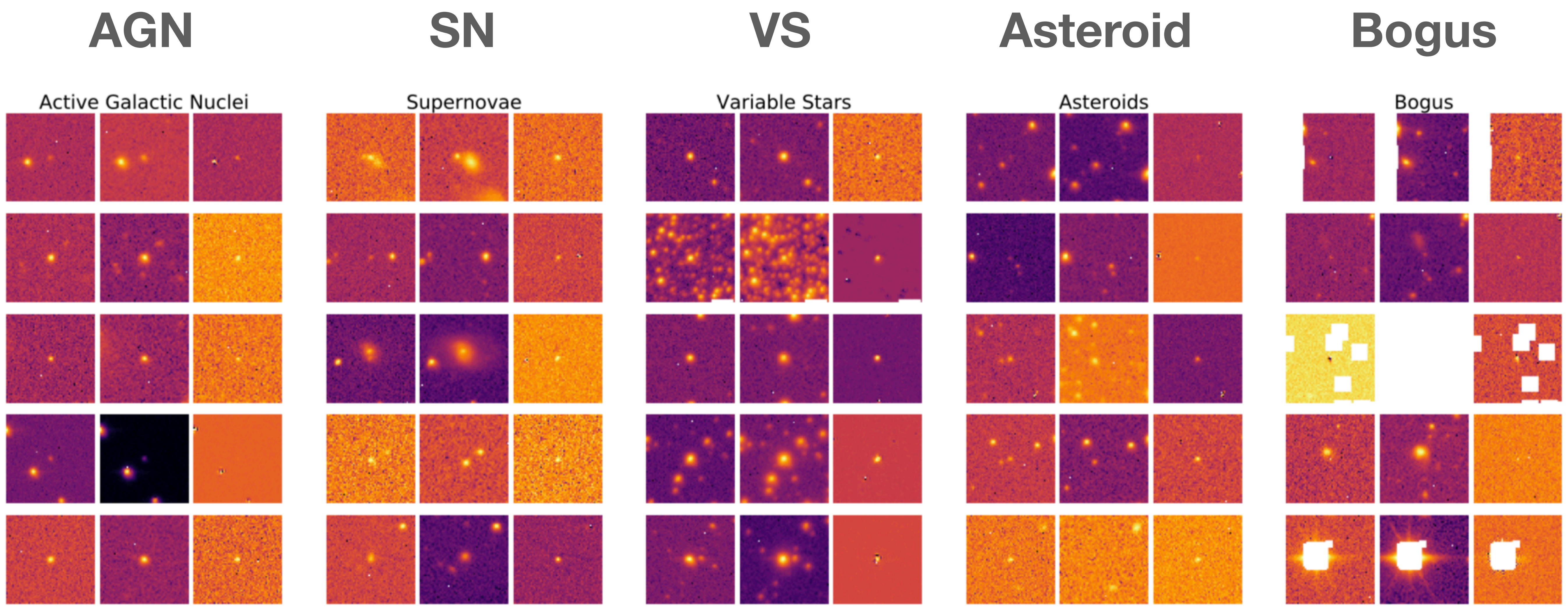


The ALeRCE broker stamp classifier

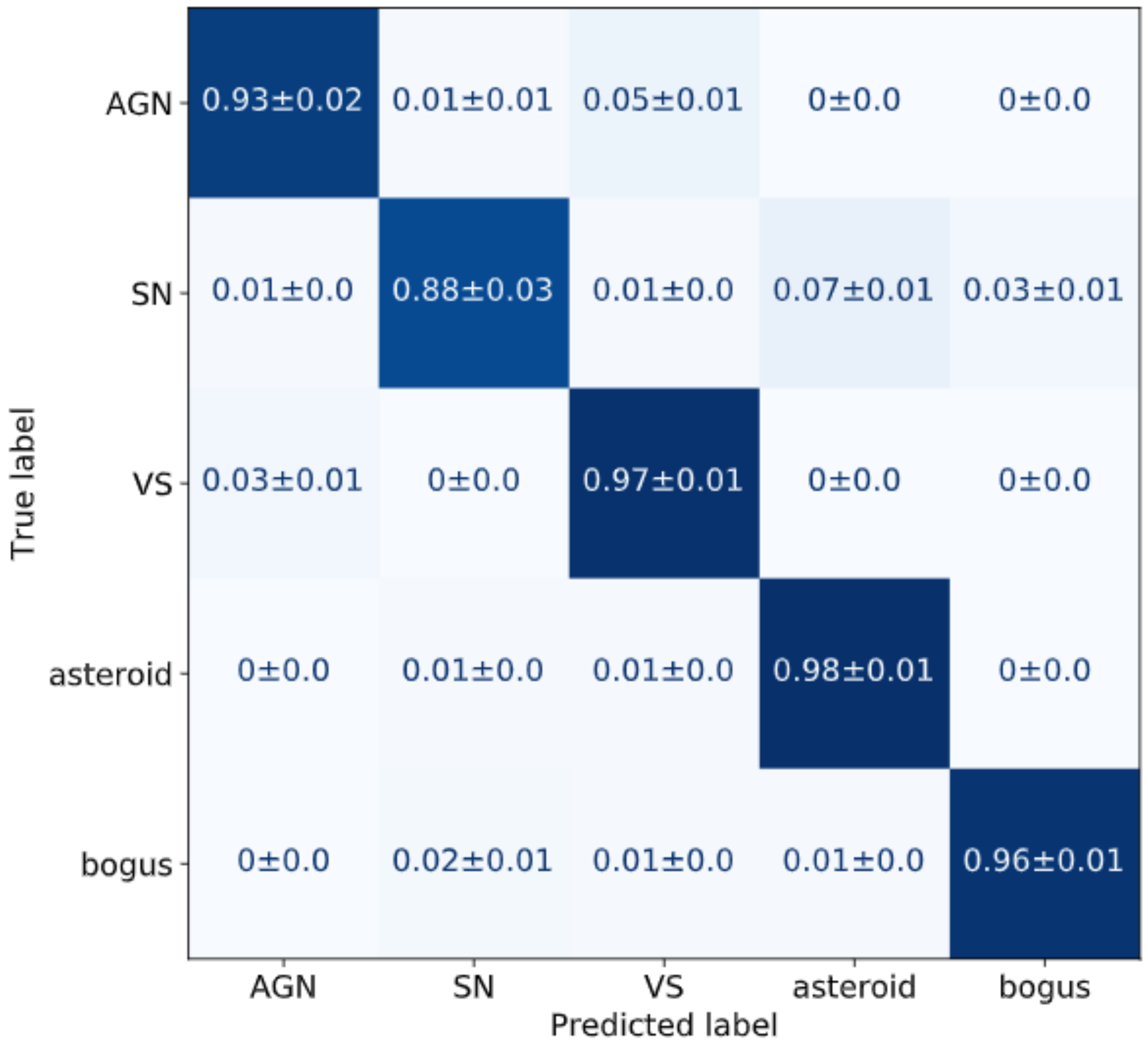
Carrasco-Davis et al. 2021, AJ, 162, 231

Convolutional Neural Network
(using 1st detection stamp)





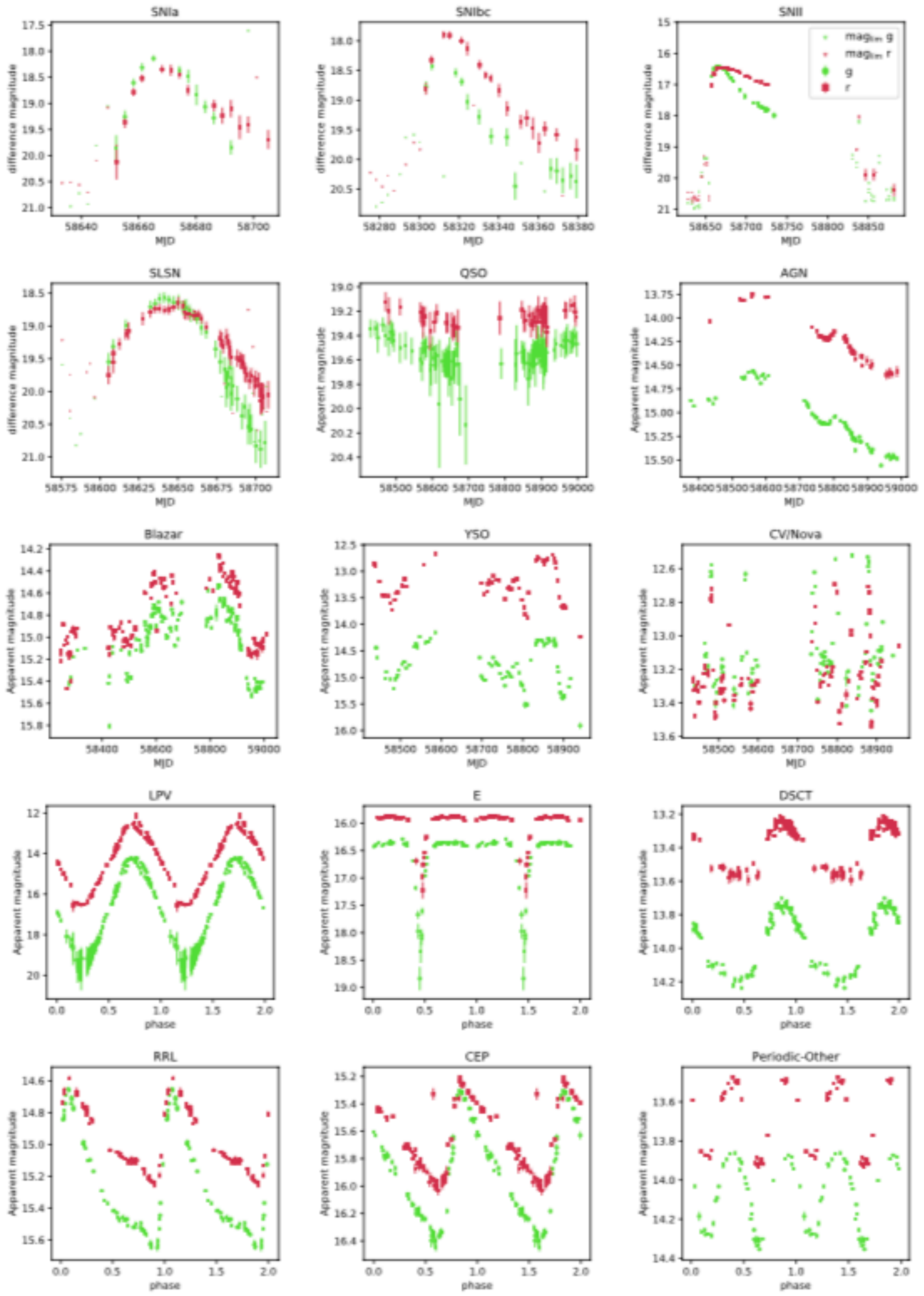
accuracy = 0.94

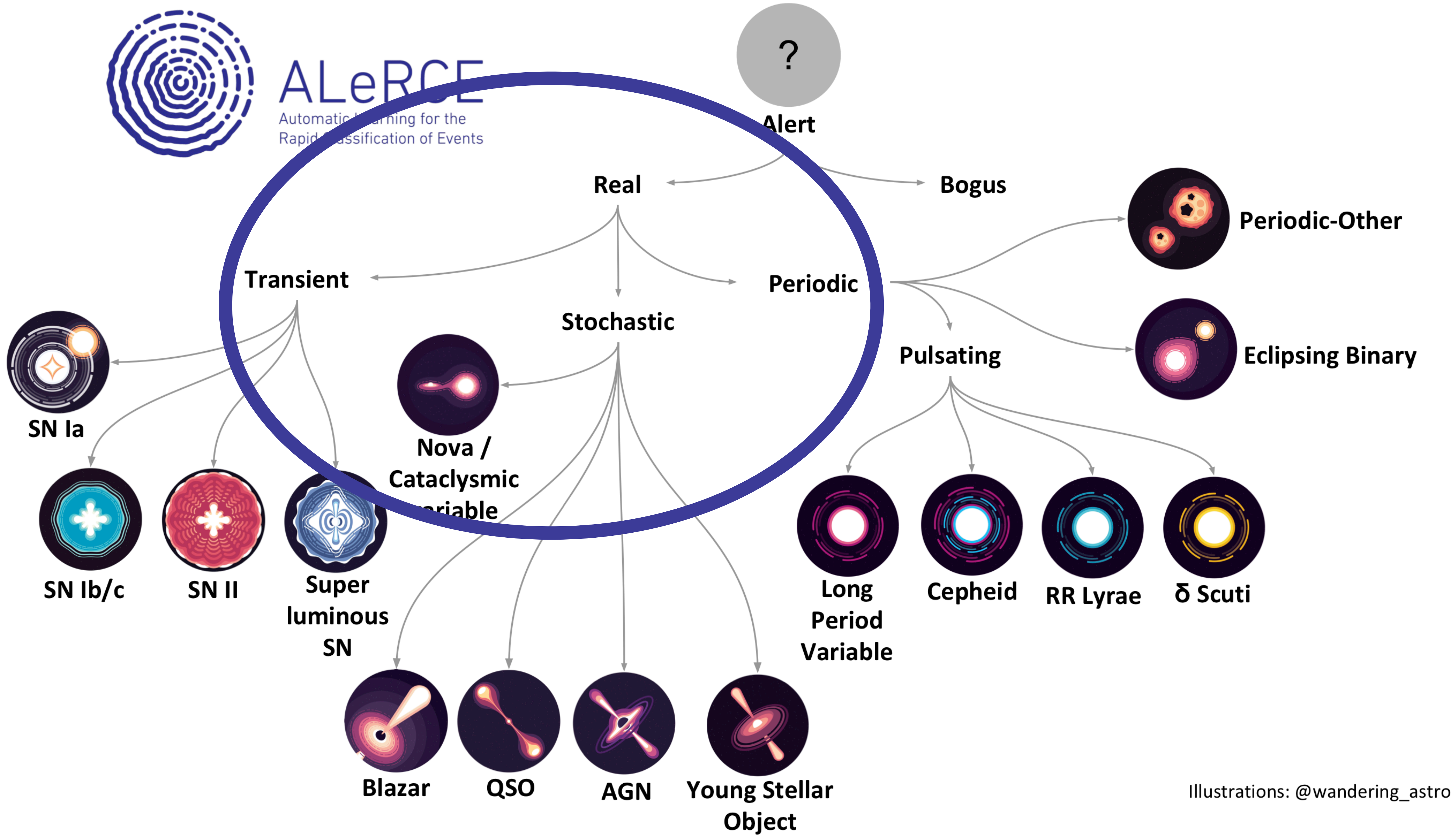


The ALeRCE broker light curve classifier

Sánchez-Sáez et al. 2021, AJ, 161,141

Balanced and Hierarchical Random Forest Model

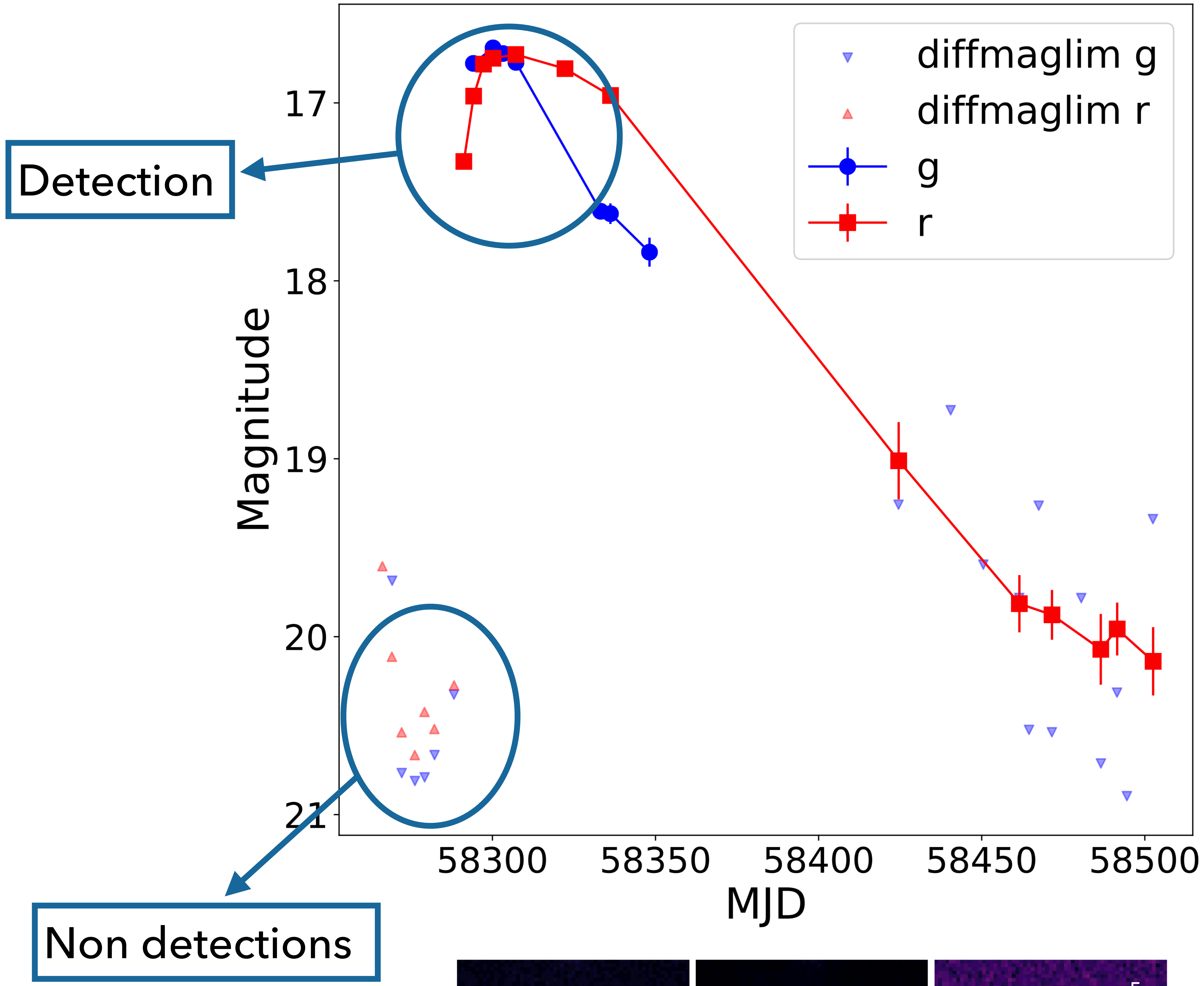


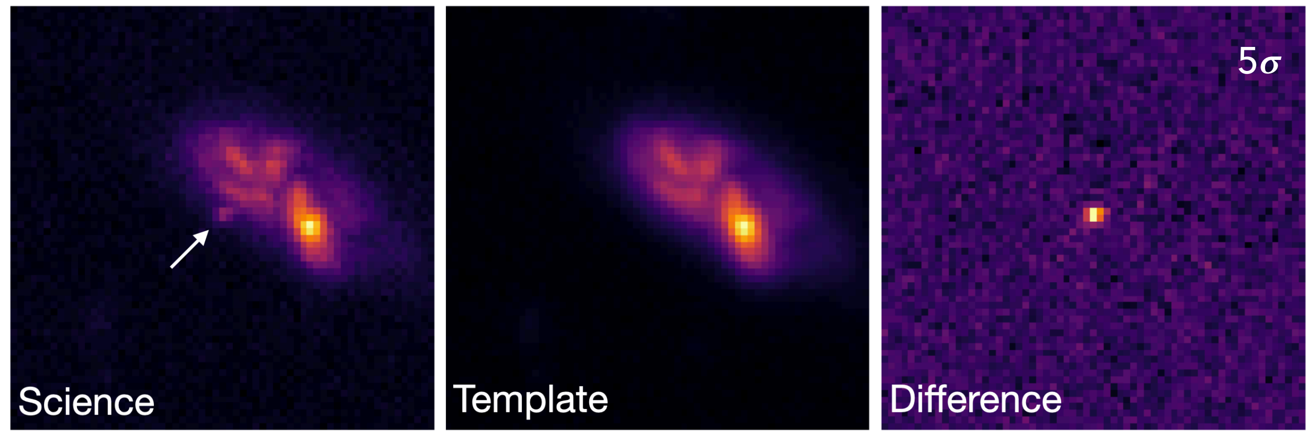


*With imbalanced-learn package

The ALeRCE broker light curve classifier

Sánchez-Sáez et al. 2021, AJ, 161,141





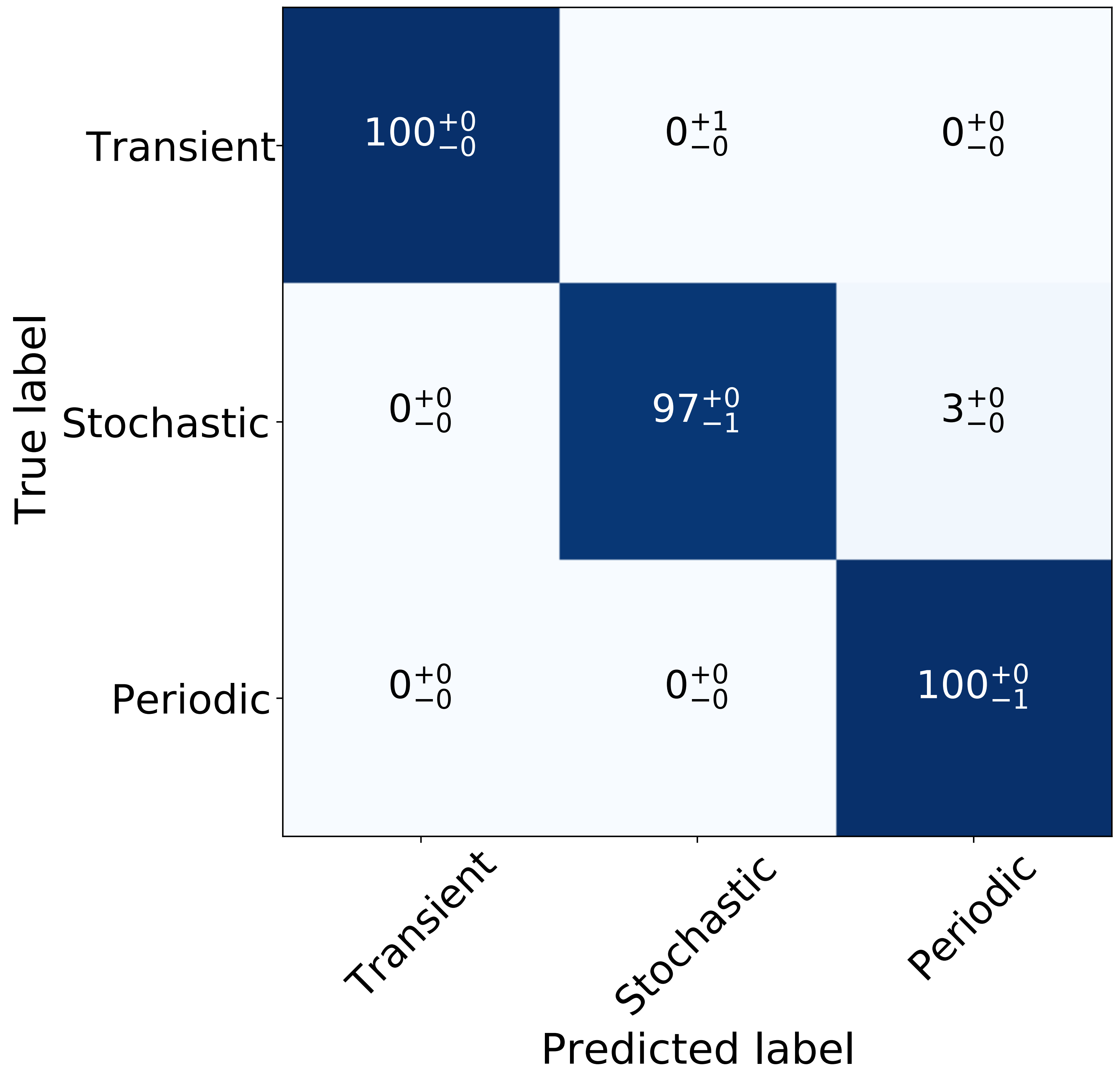
152 features in total:

1) Colors from AllWISE and ZTF (8 in total)
2) Detection features (for *g* and *r* ZTF bands, 124 in total):
   - Supernova parametric model (SPM; adapted from Villar et al. 2019b)
   - Multiband period (adapted from Mondrik et al. 2015)
   - **Irregular autoregressive model (IAR; Eyheramendy et al. 2018)**
   - **Mexican Hat Power Spectrum (MHPS; adapted from Arévalo et al. 2012)**
3) Non-detection features (for *g* and *r* ZTF bands, 18 in total)
4) Features from ZTF metadata (galactic coordinates, sgscore1 from PanSTARRS1, and real-bogus)

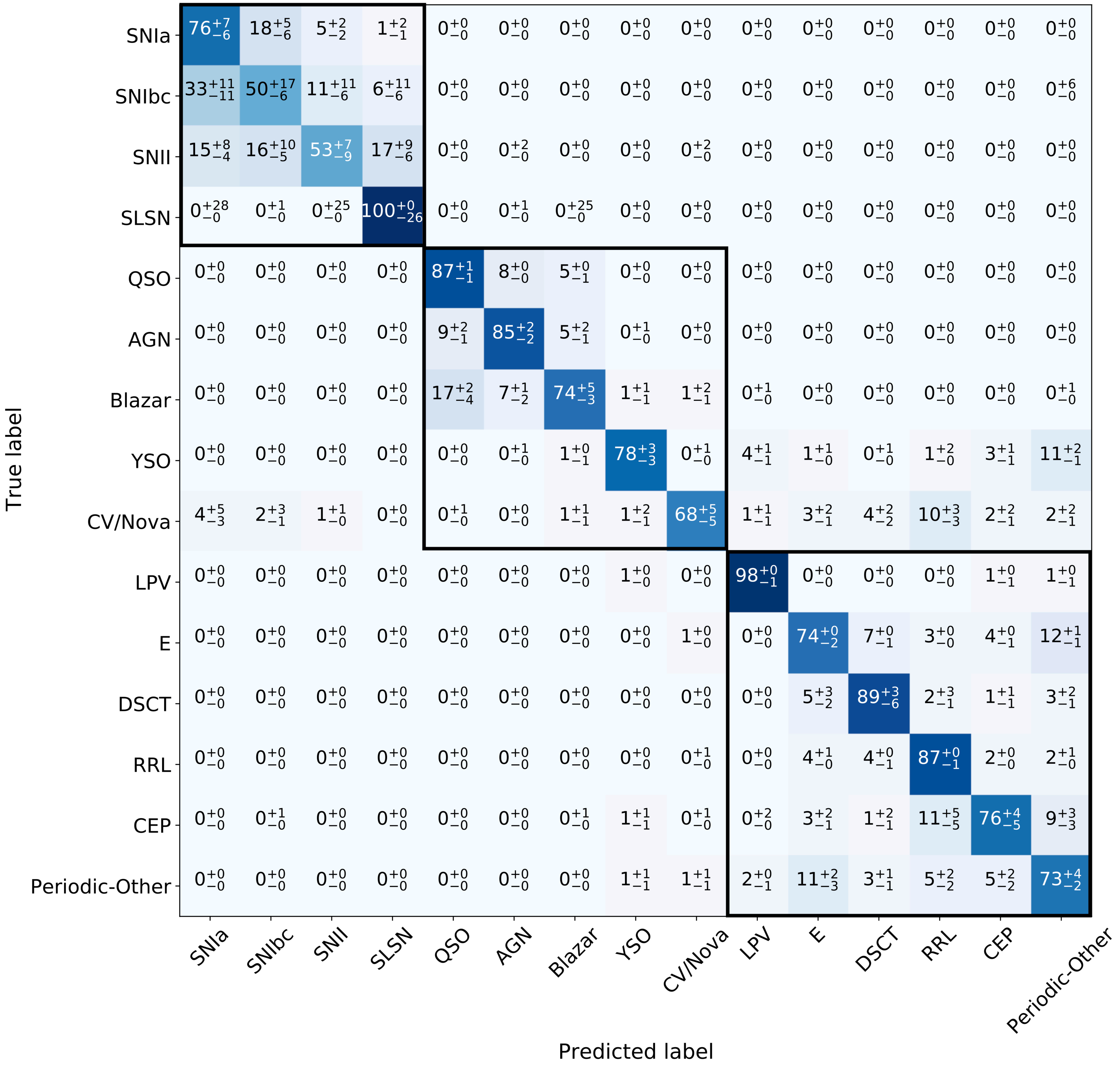
ZTF → LSST

The ALeRCE broker light curve classifier

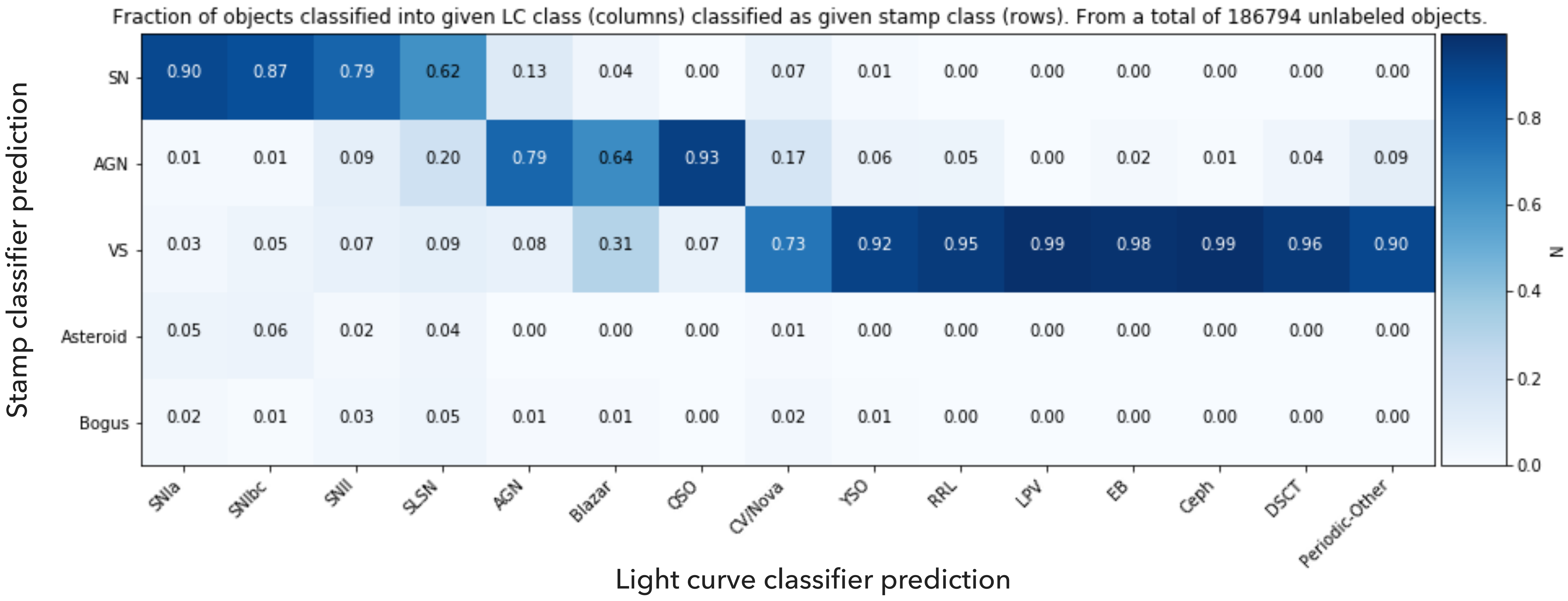
Sánchez-Sáez et al. 2021, AJ, 161,141



Classifier	Precision	Recall	F1-score
BHRF - top	0.96 ± 0.01	0.99 ± 0.01	0.97 ± 0.01
BHRF - bottom	0.57 ± 0.01	0.76 ± 0.02	0.59 ± 0.01

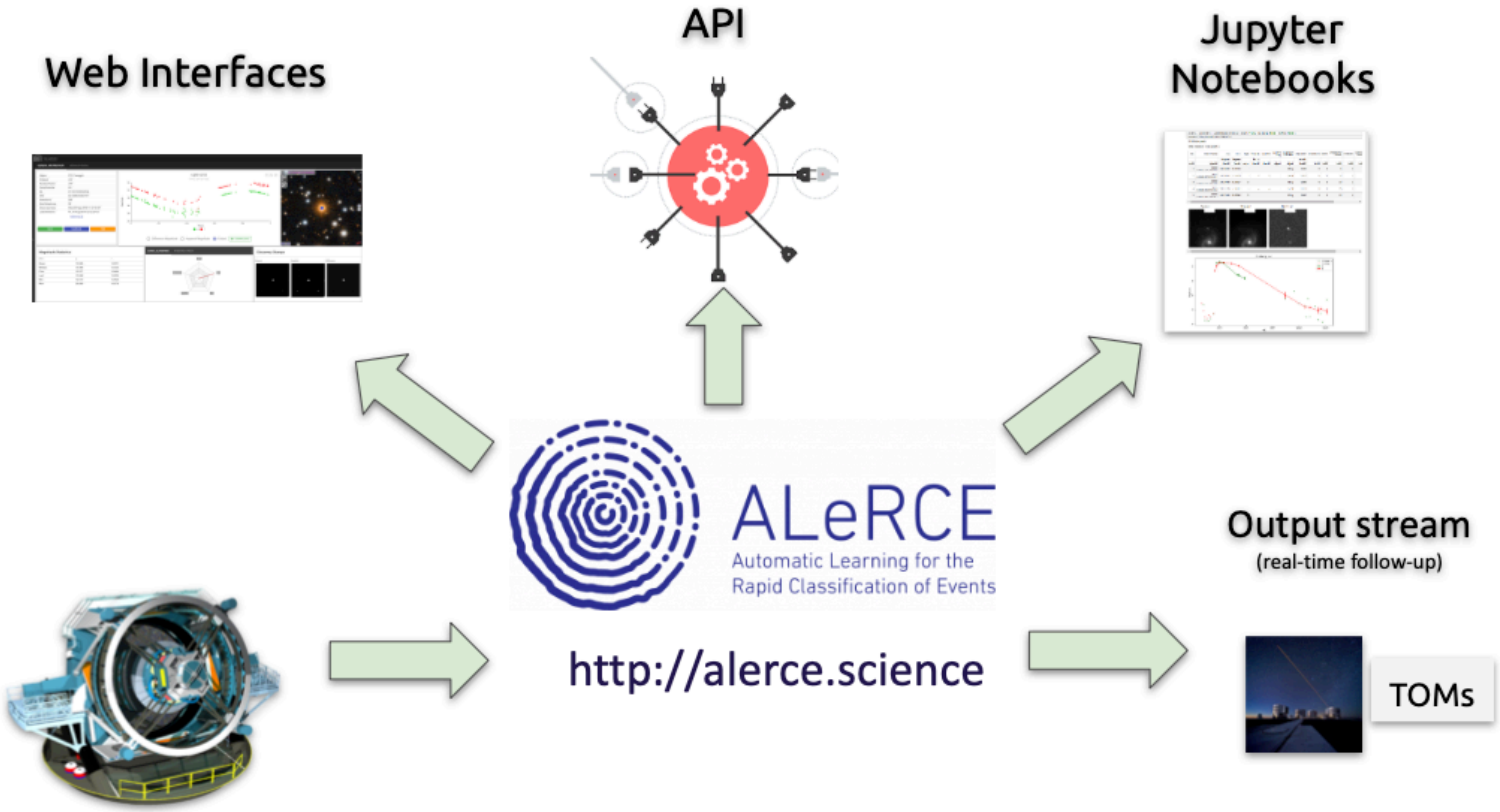


Synergy between the stamp and light curve classifiers



Förster et al. 2021. 2021, AJ, 161, 242

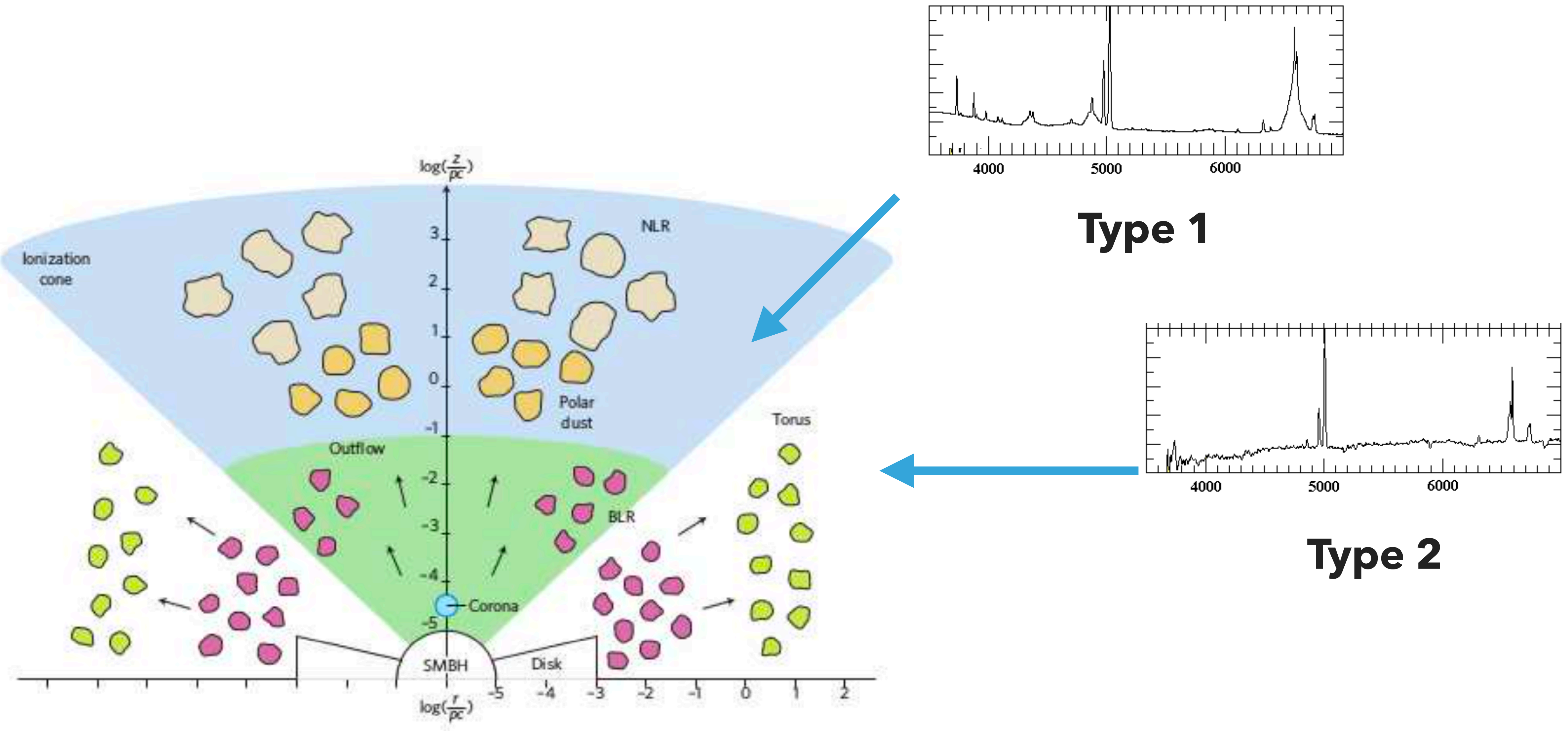
Accessing the outputs of the stamp and light curve classifiers



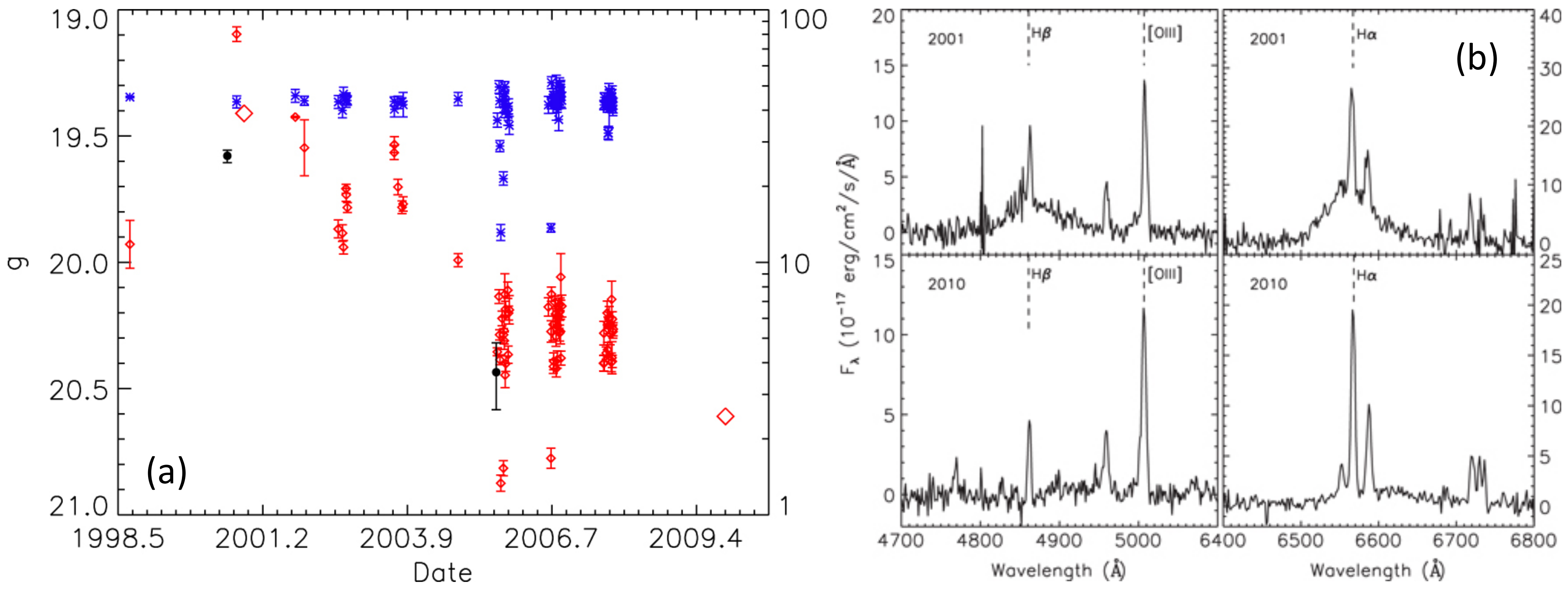
2. SEARCHING FOR CSAGNS WITH ANOMALY DETECTION

Changing-state AGNs

Changing-state AGNs (CSAGNs) in the optical range correspond to sources that change their classification as type 1 or type 2 AGN, as well as to sources that present large changes in the flux of their broad emission lines, within a timescale of months or years. **This transition phase is accompanied by a drastic change in the AGN continuum flux.**



Ramos Almeida & Ricci (2017)

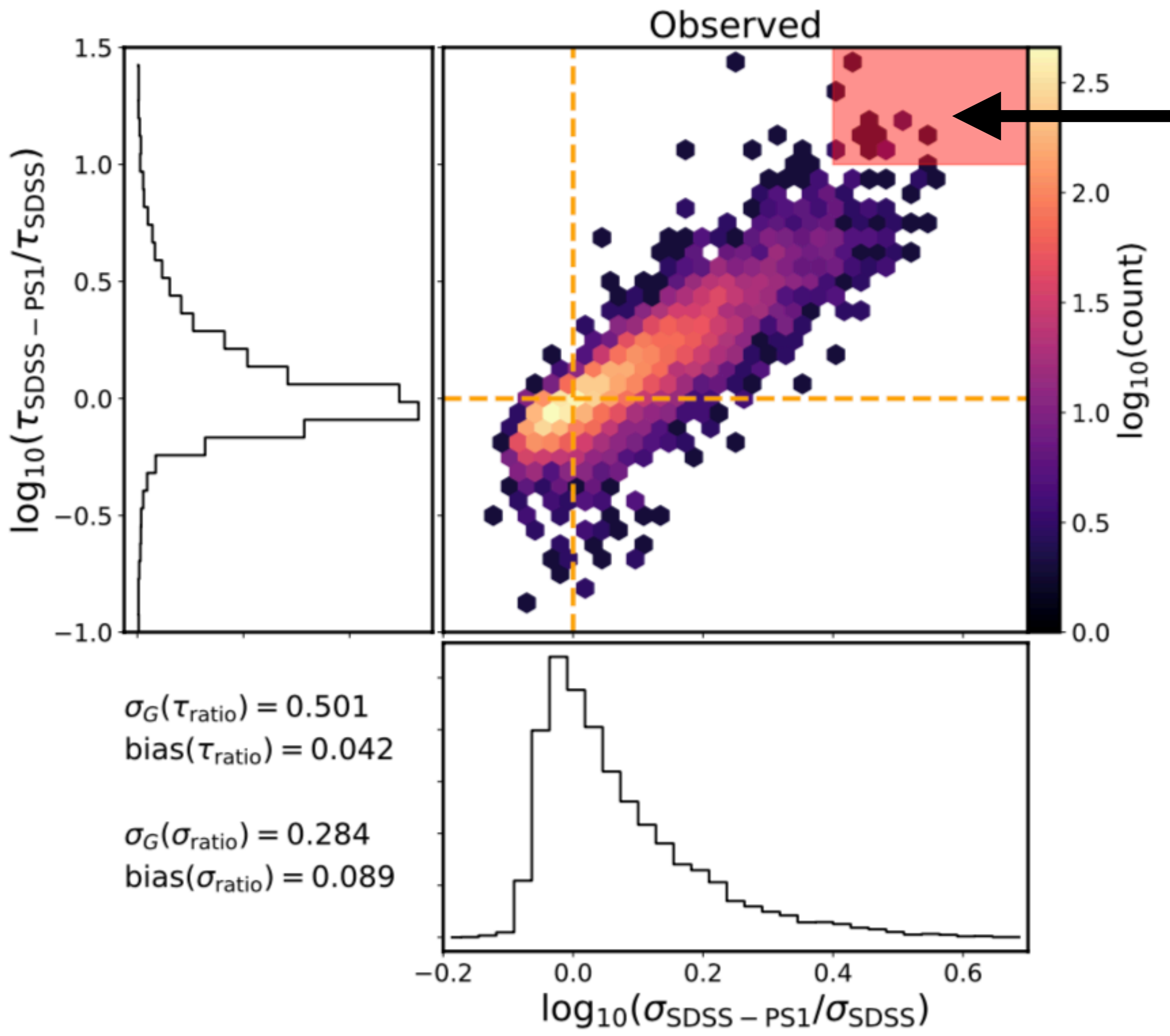


LaMassa et al. 2015

Detecting CSAGN events in massive datasets

The goal of this work is to create a method to search for CSAGN candidates in massive data sets, using anomaly detection techniques.

Currently, we use data from the Zwicky Transient Facility data releases, and in the future we will apply this to Vera Rubin / LSST data.



CSAGN candidates

Suberlak et al. 2021

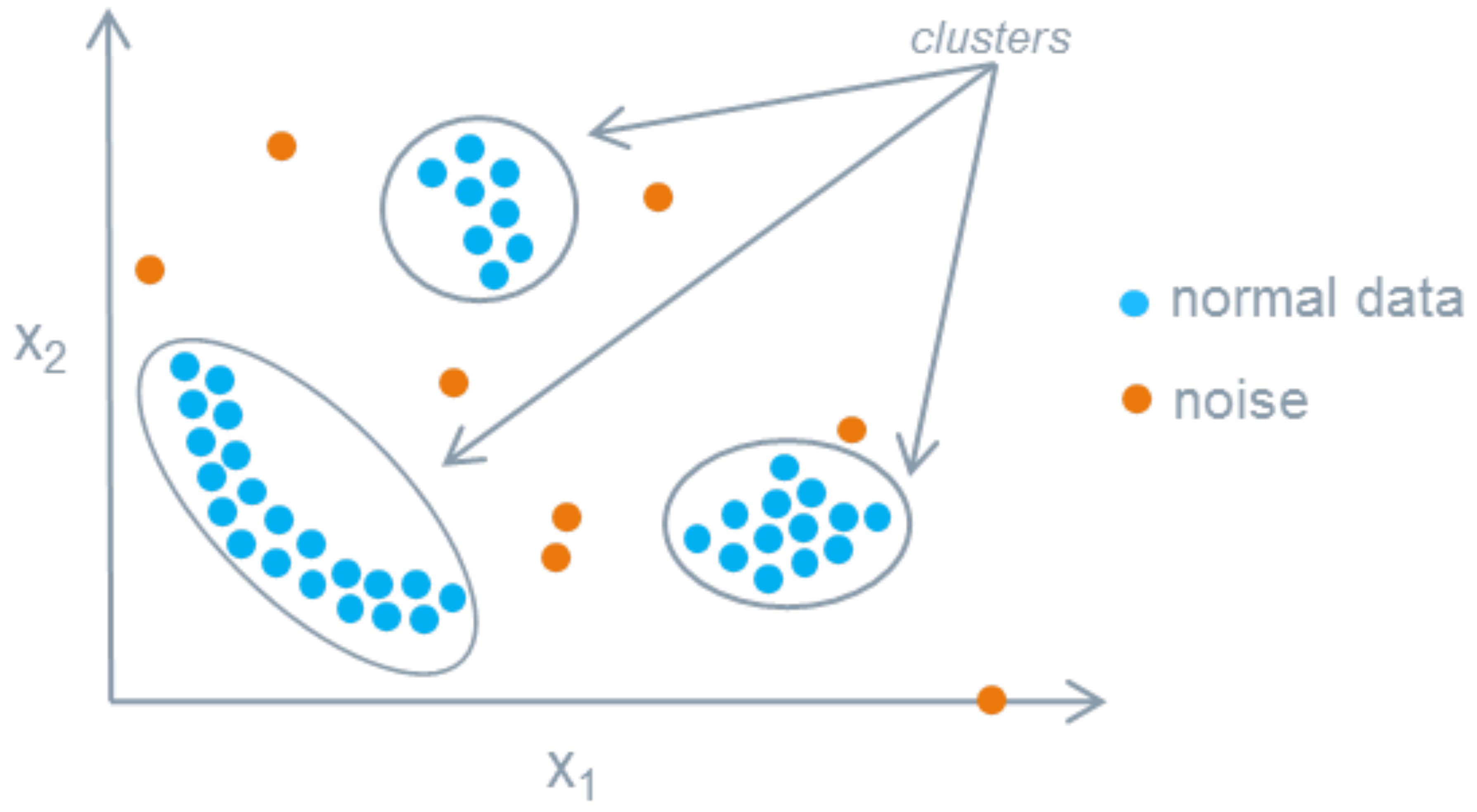
ZTF → VERA C. RUBIN OBSERVATORY

ANOMALY DETECTION TECHNIQUES

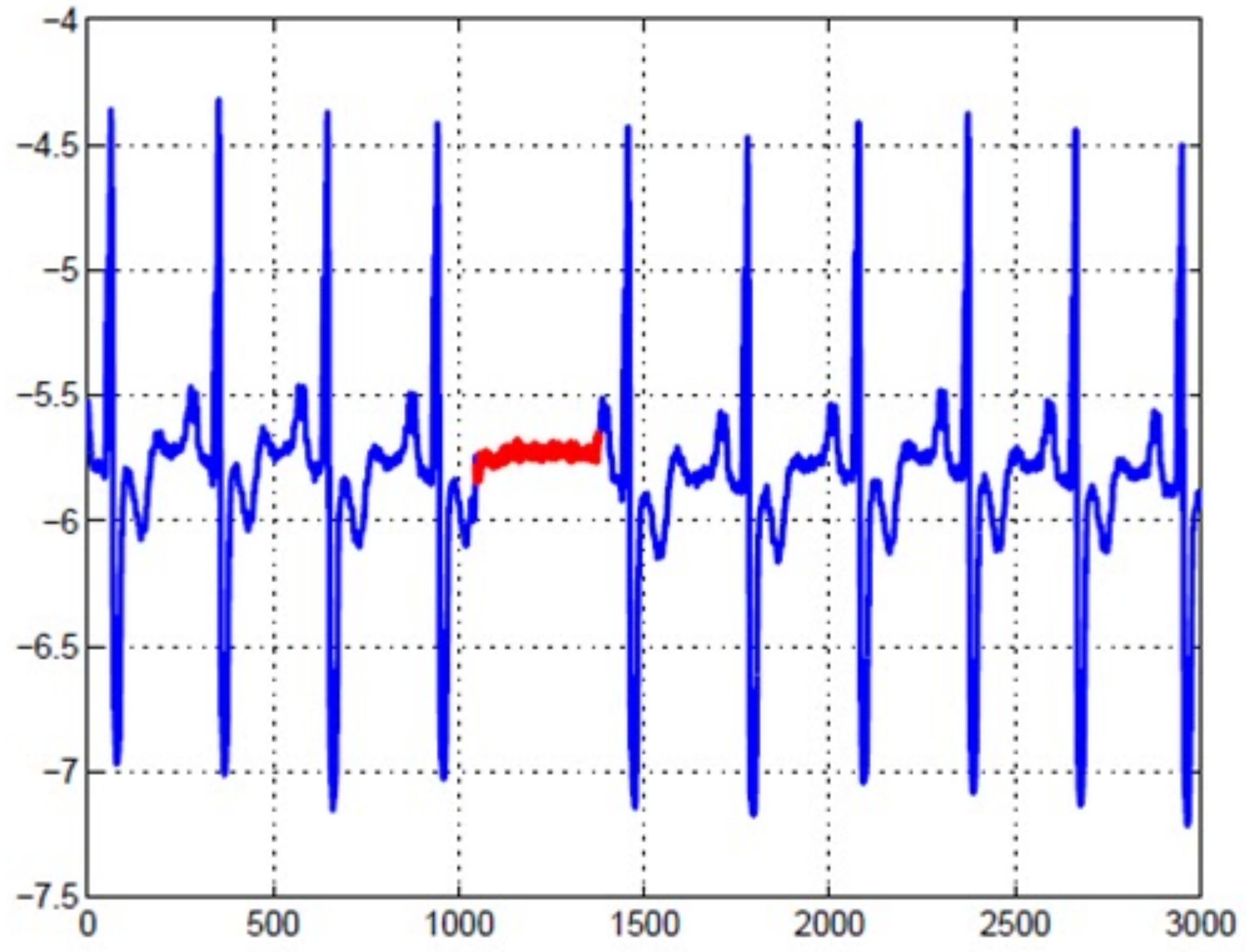
Anomaly detection (AD)

AD correspond to the identification of rare events or observations that differ significantly from the majority of the data.

Out of distribution anomaly: searching for unusual objects within datasets.

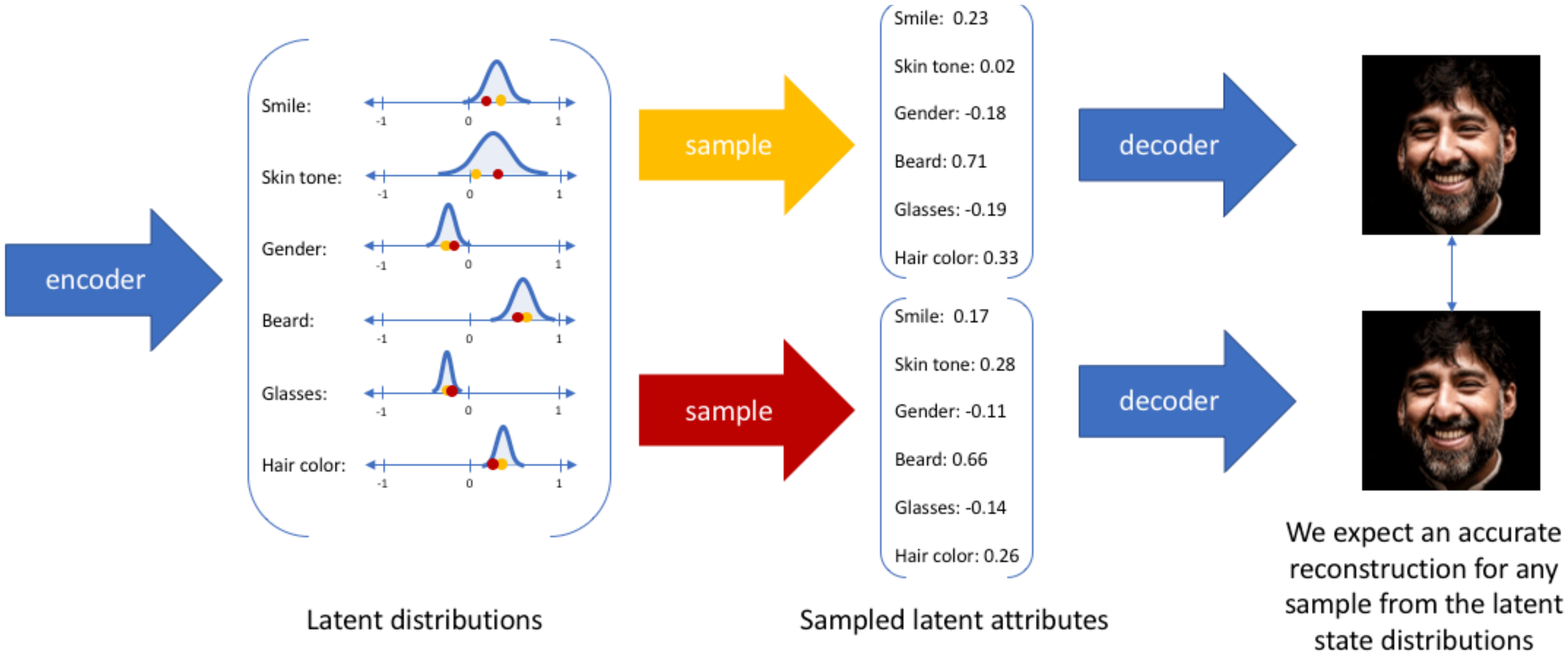


Contextual anomaly: searching for objects that suddenly start presenting unusual behaviors.



Variational Autoencoders (VAEs)

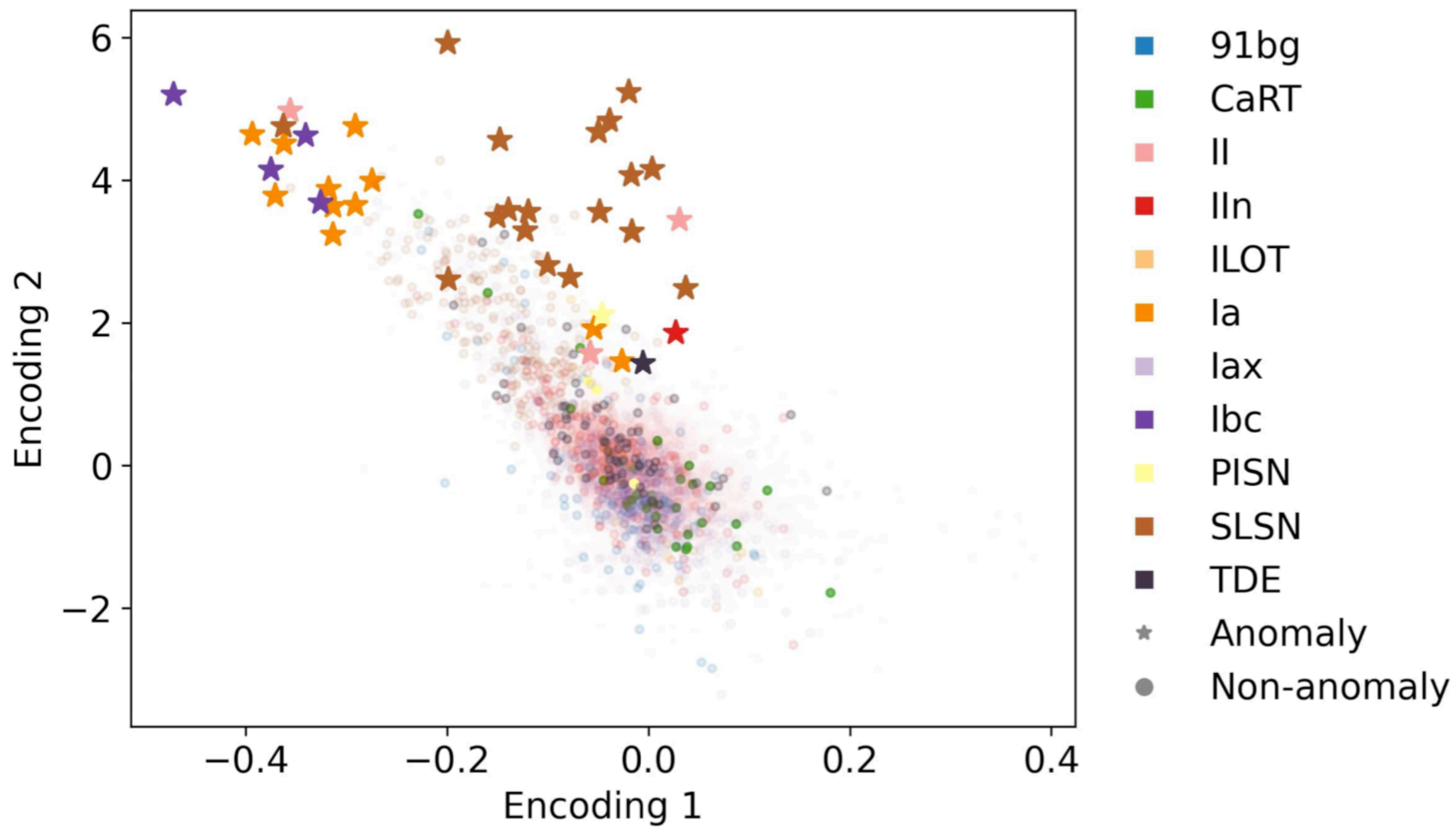
VAEs correspond to a modification of the more classical Autoencoder (AE) architectures. In this case, the latent representations are described by multivariate normal distributions, where each attribute or feature in the latent space is described by a latent mean (μ) and a latent variance (σ^2), which can be used to randomly sample a set of attributes.



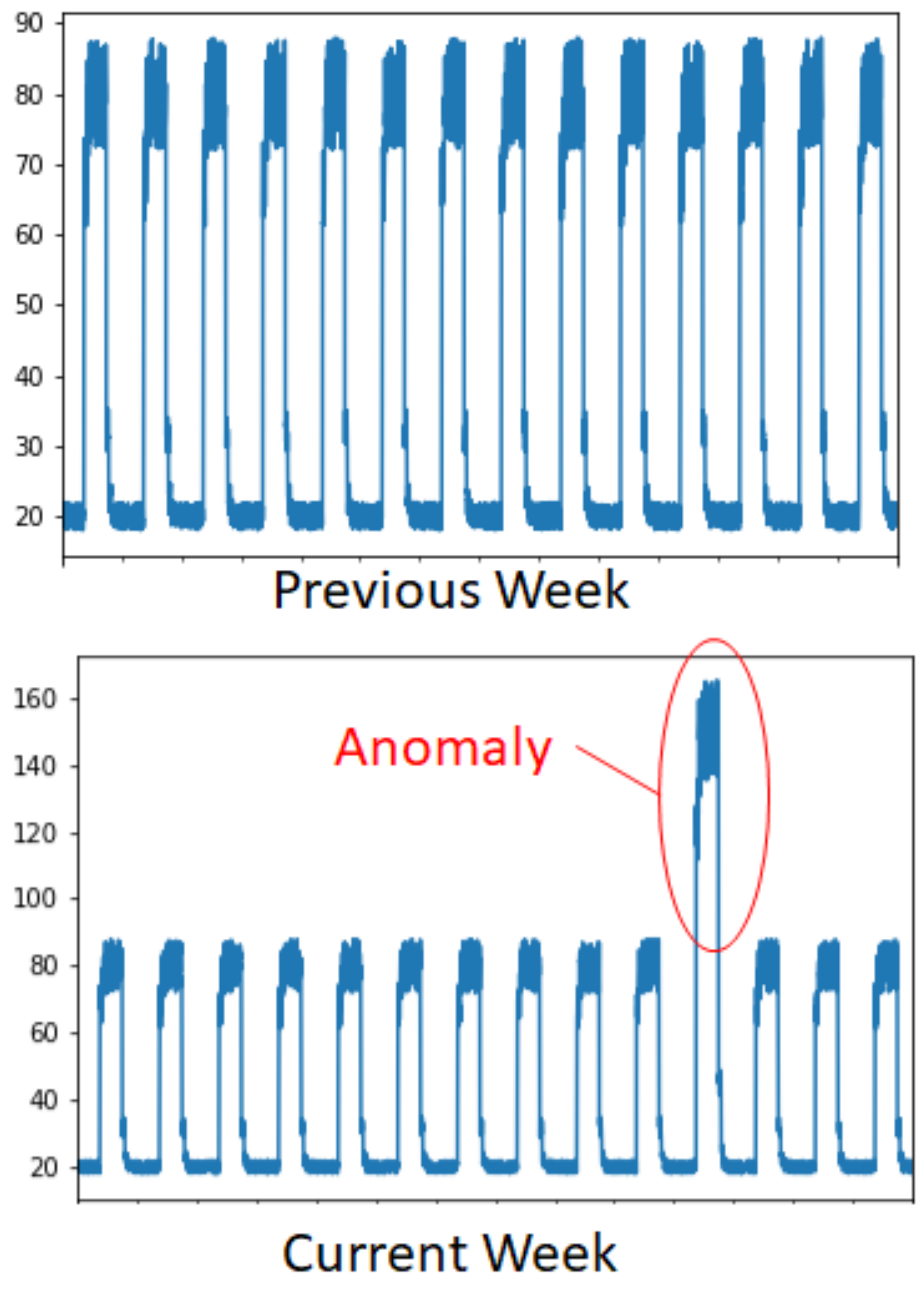
Credits: https://www.jeremyjordan.me/variational-autoencoders/

VRAEs for time series anomaly detection

Out of AD: using the latent space to define outliers that are in atypical locations of latent space (e.g., Villar+2021)



Contextual AD: using the reconstruction error of the VRAE as an anomaly score



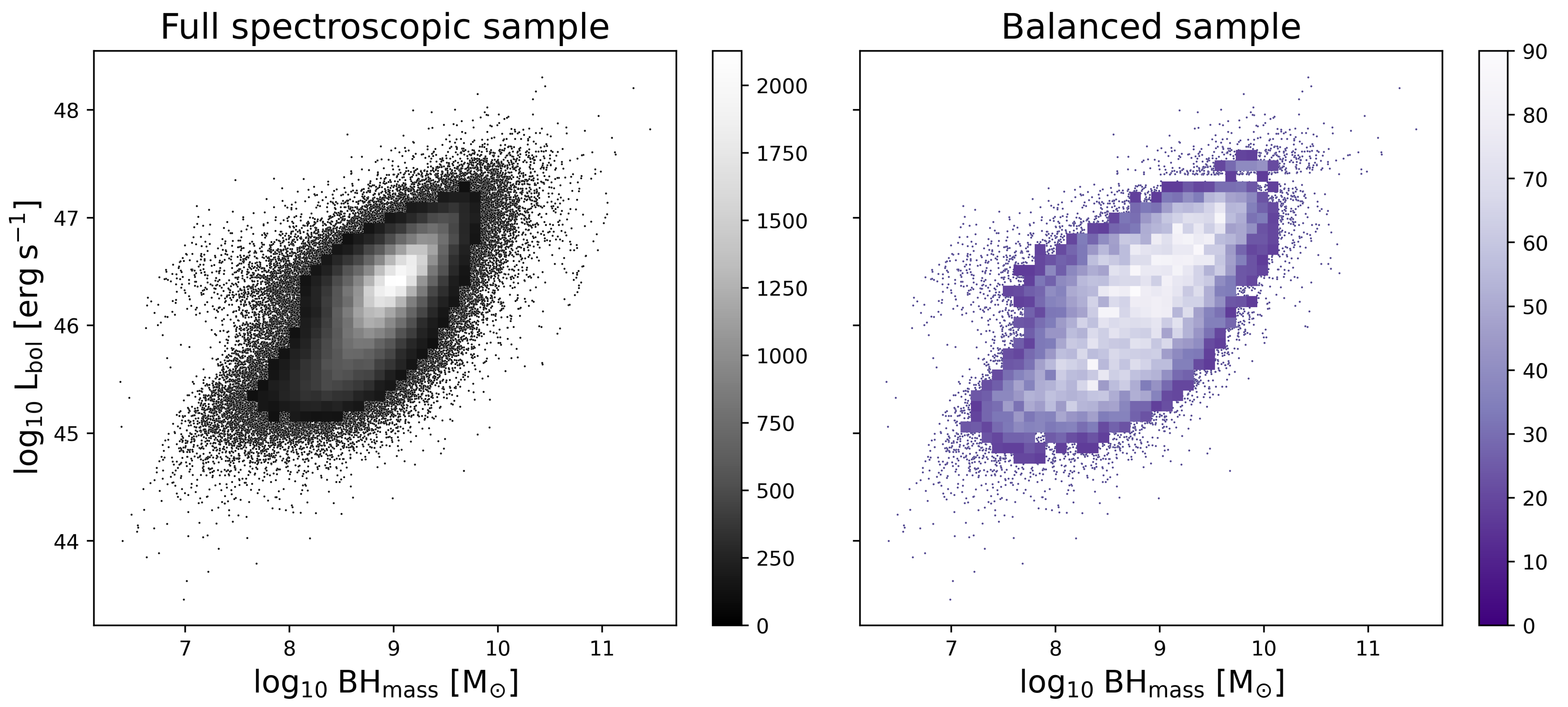
SEARCHING FOR ANOMALOUS AGN VARIABILITY WITH ANOMALY DETECTION

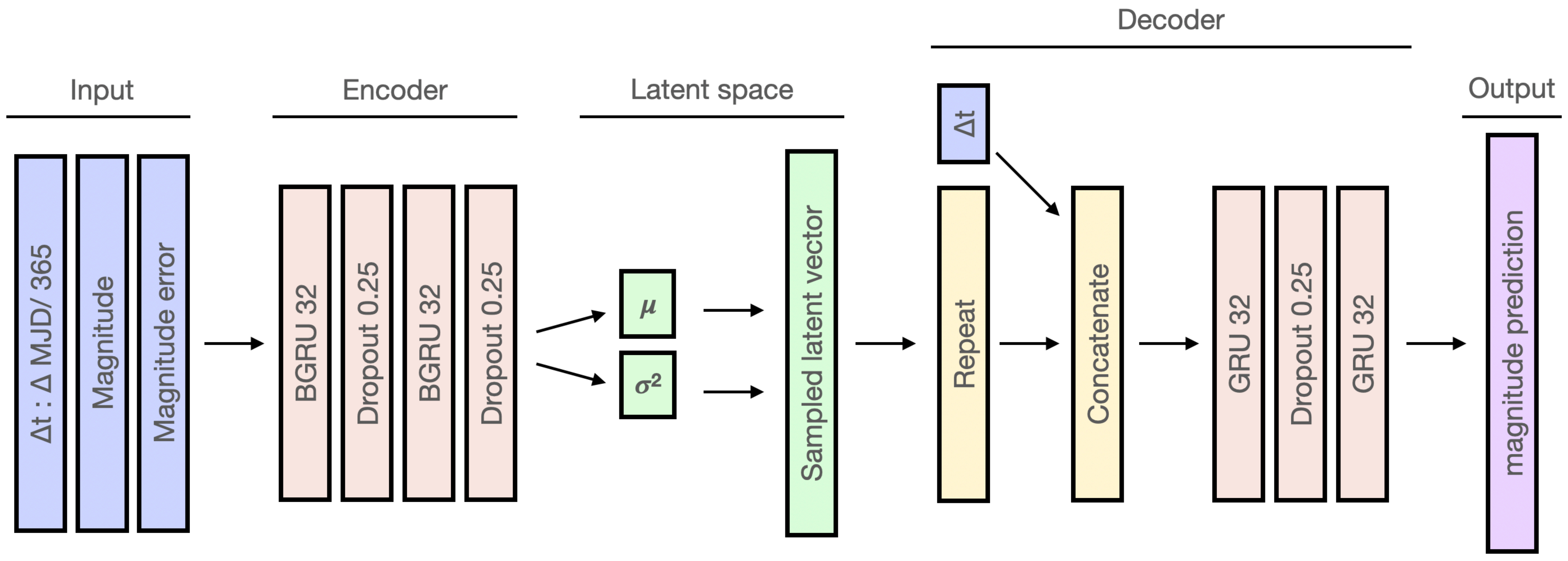
VRAEs to model AGN variability

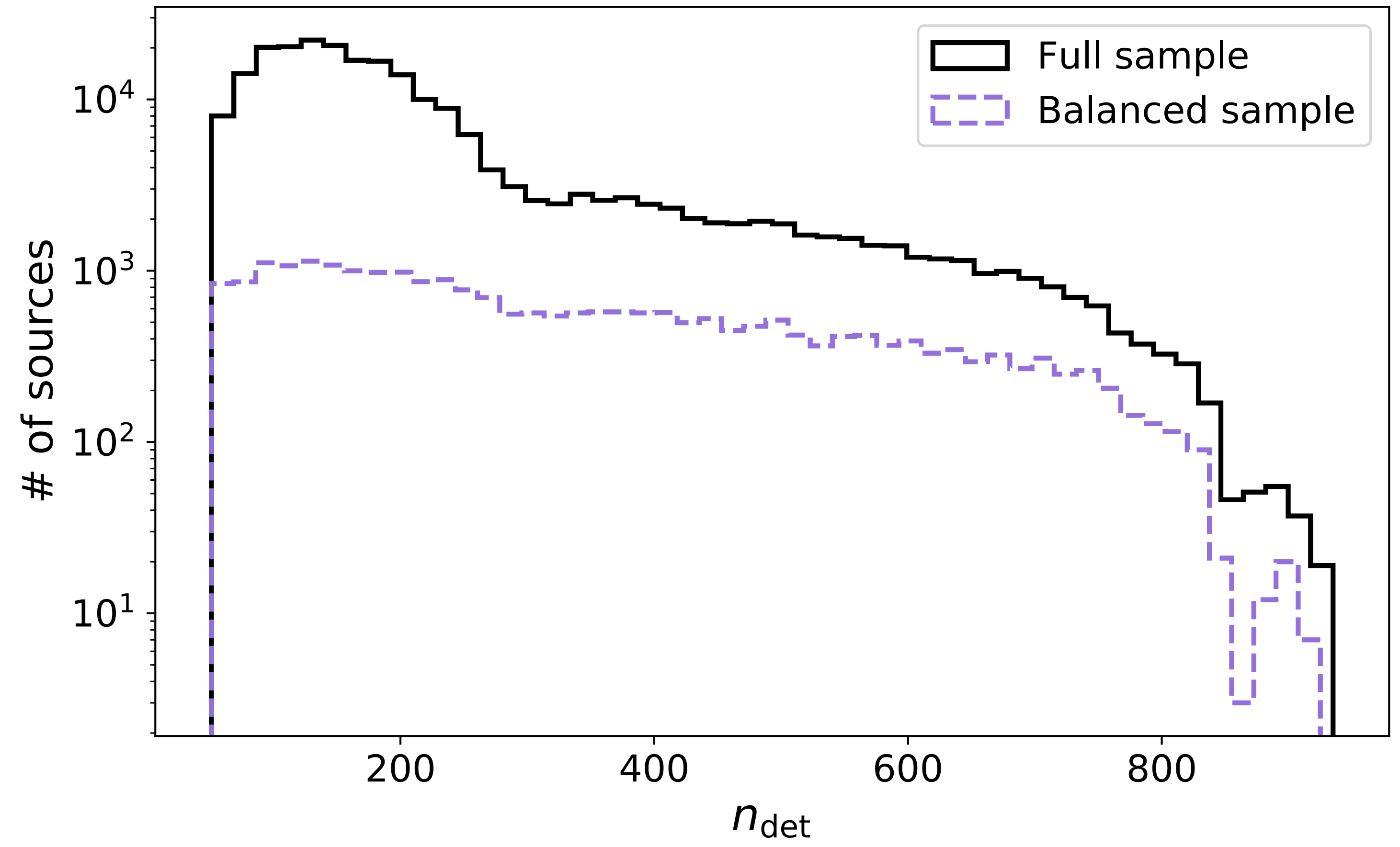
Sánchez-Sáez et al. 2021, AJ, 162, 206

230,451 AGN light curves from ZTF DR5 (including different classes from the MILLIQUAS and ROMABZCAT catalogs)

- VRAE architecture (inspired by Tachibana+2020's model)
- Trained with a dataset balanced by means of their physical properties and number of epochs per light curve.







VRAEs for AGN variability anomaly detection

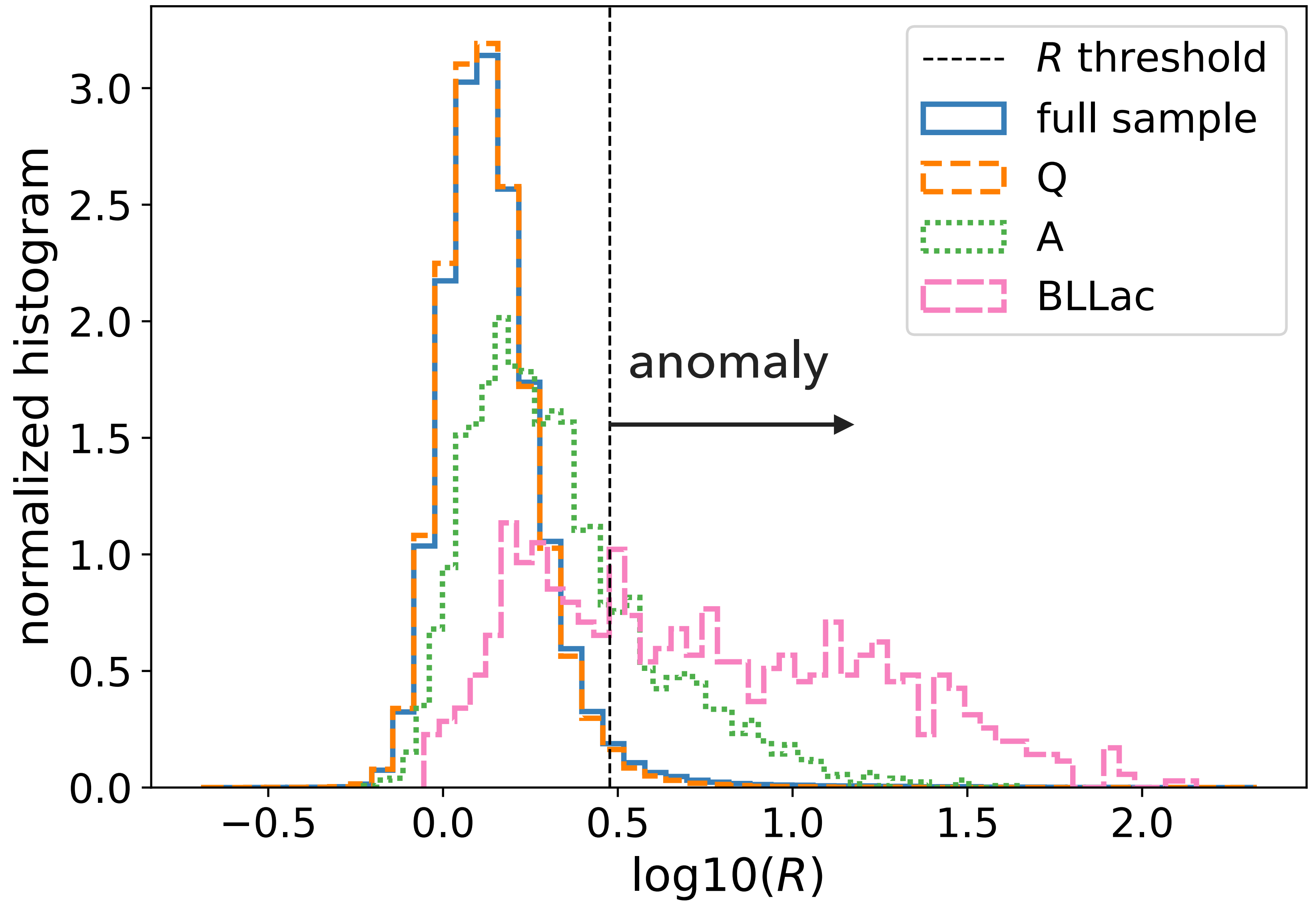
Sánchez-Sáez et al. 2021, AJ, 162, 206

Selection of outlier candidates

Using the reconstruction error of the VRAE as an anomaly score:

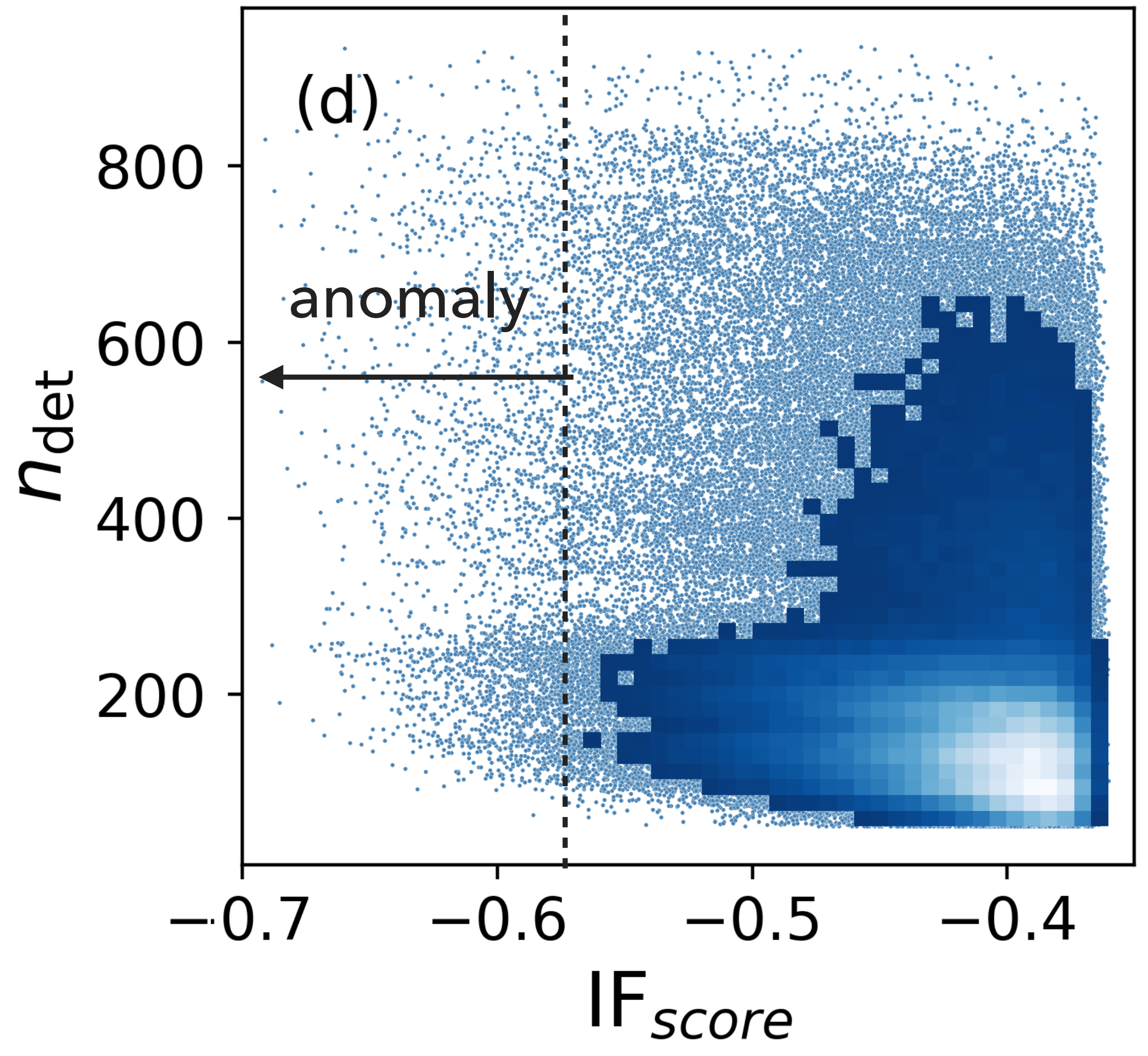
$R > 3$

$$R = \frac{1}{N_T} \sum_{i}^{N_T} \frac{(m_i - m_{pred_i})^2}{\mathrm{err}_i^2 + \mathrm{err}_{pred_i}{}^2}.$$



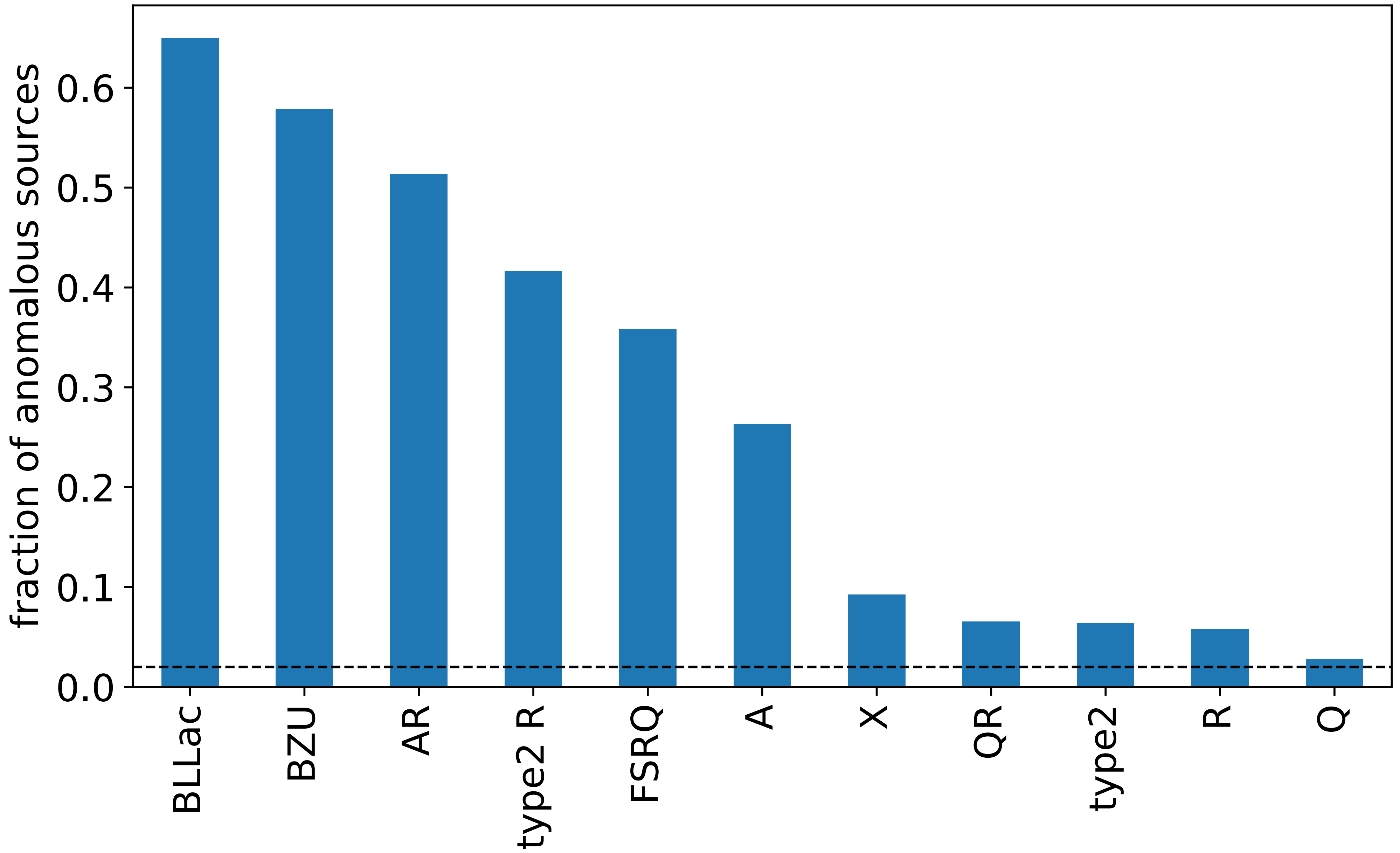
Using the latent space attributes with an Isolation Forest algorithm (IF):

IF_score < IF threshold 2% contaminants (-0.57633)



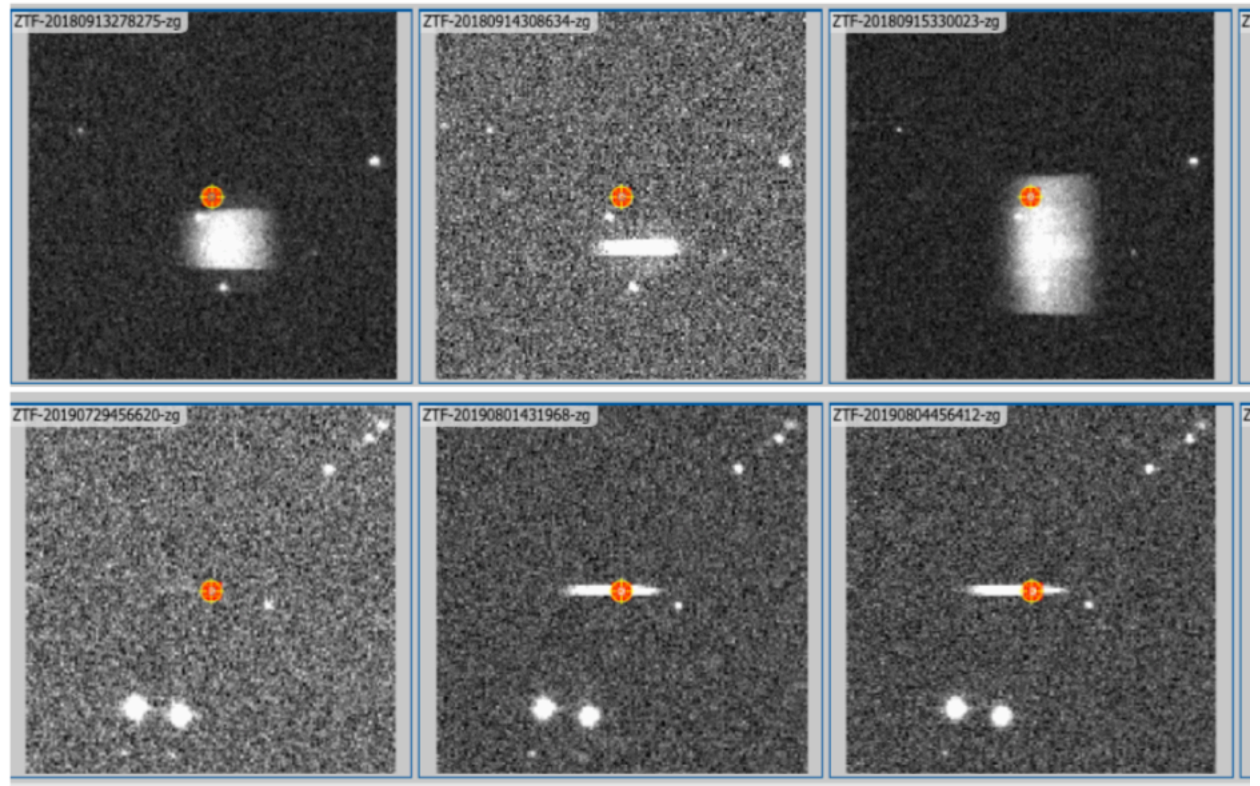
VRAEs for AGN variability anomaly detection: results

We selected 8,809 anomalies.

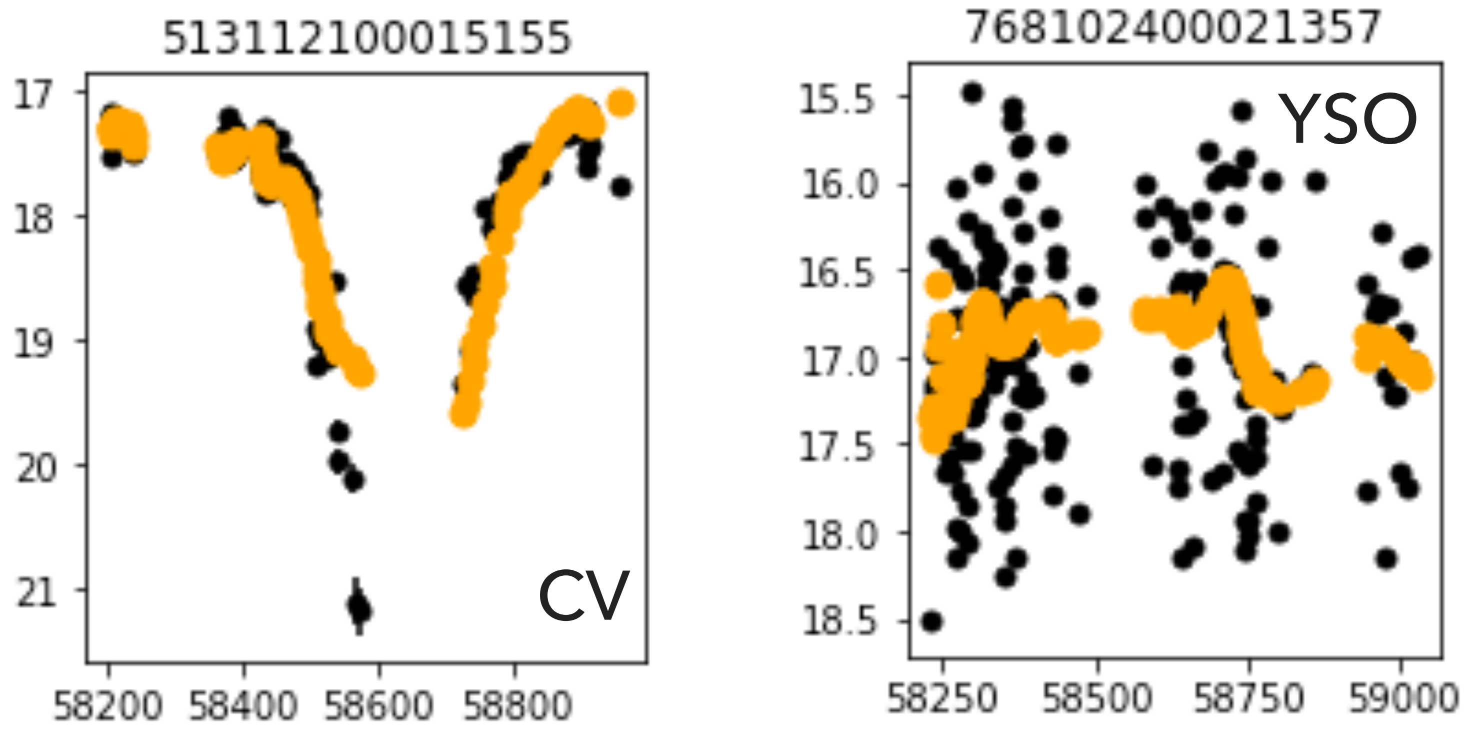


Sánchez-Sáez et al. 2021, AJ, 162, 206

Dominated by photometric issues



And miss-classified sources

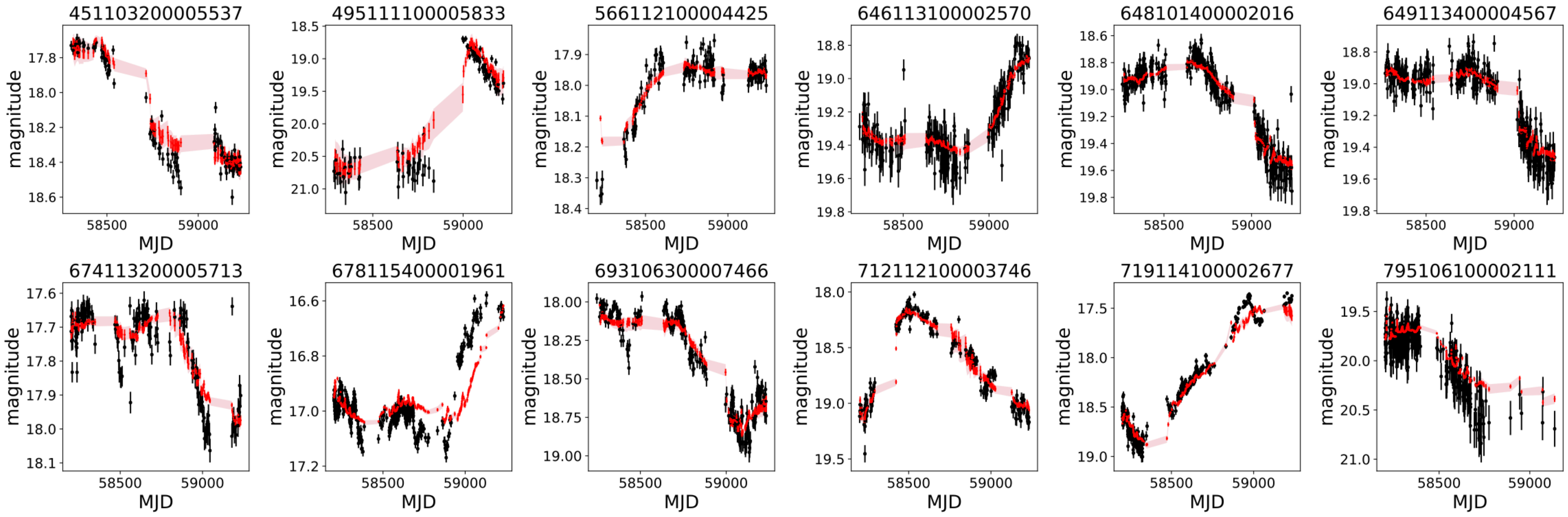


CSAGN candidates

Sánchez-Sáez et al. 2021, AJ, 162, 206

We visually inspected the list of candidates and selected as promising CSAGN candidates those anomalies that present evidence of flares, and/or abrupt increment or decrement in the luminosity. **We identified 75 CSAGN candidates** (65% are regular QSOs).

Further spectroscopic follow-up is required to confirm the nature of our candidates. Although 4 are known CSAGN candidates (Graham+2020), 2 have been spectroscopically confirmed (M. Graham, private communication), and 28 are candidates using other techniques (Graham+ in prep).



Summary

- Variability-ML-based classifiers can help us to select AGN populations that can me missed by more traditional selection techniques.

- The ALeRCE light curve classifier corresponds to the first attempt to classify multiple classes of stochastic variables (including nucleus- and host-dominated active galaxies, blazars, young stellar objects, and cataclysmic variables) in addition to different classes of periodic and transient sources, using real data.

- Detection of CSAGN events in massive data sets is crucial to understand these events and to improve our knowledge of the physical mechanisms behind AGN variability.

- We used a Variational Recurrent Autoencoder (VRAE) architecture to model 230,451 AGN light curves from the ZTF DR5. We used reconstruction error and the latent space attributes to search for anomalous AGN light curves.

- We found 8,809 anomalies. These anomalies are dominated by bogus candidates (photometric issues, miss-classified sources in the original catalogs), but we were able to identify 75 promising CSAGN candidates.

pasanchezsaez@gmail.com
psanchez@eso.org