# Anomaly Detection in the Open Supernova Catalog

**F. Mondon (CNRS-LPC, florian.mondon@clermont.in2p3.fr), A. A. Volnova (IKI RAS, alinusss@gmail.com), K. L. Malanchev (SAI MSU, HSE), M. V. Pruzhinskaya (SAI MSU), M. V. Kornilov (SAI MSU, HSE), E. E. O. Ishida (CNRS-LPC), and V. S. Korolev (MIPT)**

## Introduction:

The next generation of large astronomical surveys will discover millions of transients making available a combined data set of unprecedented volume. The use of machine learning (ML) method become essential to process such large data volumes. Moreover, rare or completely new transients are expected and the task of finding them can be framed as an anomaly detection problem. In this analysis we turn to the automatic search for anomalies in the real photometric data using the Open Supernova Catalog (OSC, [1]), which serves as a proof of concept for the applicability of these methods to future large scale surveys.

## Pre-processing:

Machine Learning algorithms generally need a homogeneous input data matrix. For this purpose, the pre-processing procedure allows to extract features from OSC light curves like illustrated by Fig 1. First, we prepared the photometric data extracted from the OSC; we transformed the magnitudes to the flux units, converted the upper limits, and implemented 1-day time-binning. After that we approximate the light curves with Gaussian processes [2]. In general each light curve are approximated by GP independently but in this study for each object we used Multivariate Gaussian Process approximation that takes into account the correlation between light curves in different bands, approximating the data by GP in all filters in a one global fit (for details see Kornilov et al. 2019, in prep.). Finally to reduce the e dimensionality of the data we applied t-SNE [3], a variation of the stochastic neighbour embedding method [4],

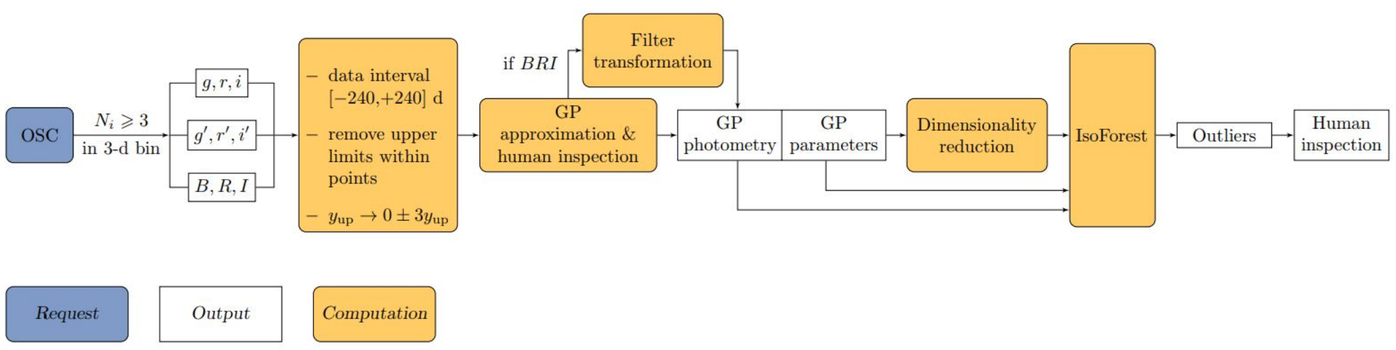


**Figure 1.** Workflow for the analysis. Ni denotes the number of observations in i'th band. GP photometry includes 364 features: 121 × 3 normalized fluxes and the LC flux maximum; GP parameters are 9 fitted parameters of the Gaussian process kernel and the log-likelihood of the fit.

## Isolation forest:

To find the outliers we use the isolation forest algorithm (Liu et al. 2008, 2012) and describe Fig 3. We run the isolation forest algorithm on 10 data sets obtained using the same photometric data (Fig. 1):

A. data set of 364 photometric characteristics
B. data set of 10 parameters of the Gaussian process
C. 8 data sets obtained by reducing 374 features to 2–9 t-SNE dimensions

For each data set we obtained a list of outliers. The distribution of objects in each of 10 data sets by anomaly score is presented in Fig. 8. An example of the isolation forest algorithm applied to the three-dimensional reduced data set is shown in Fig. 9.

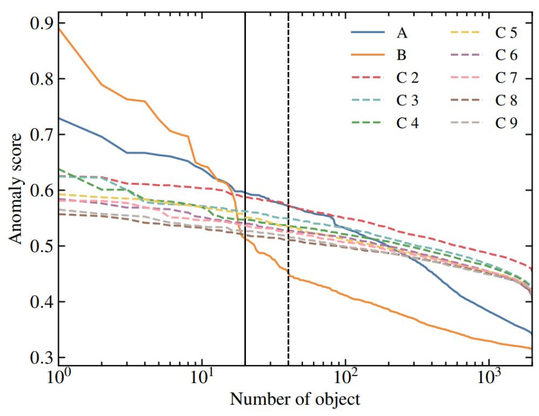


**Figure 4.** Distribution of objects by anomaly score in 10 data sets, C 2 – C 9 denote C data sets with 2–9 t-SNE dimensions. In each data set objects are ordered by score. Black solid and dashed lines denote 1% and 2% contamination level of outliers, respectively

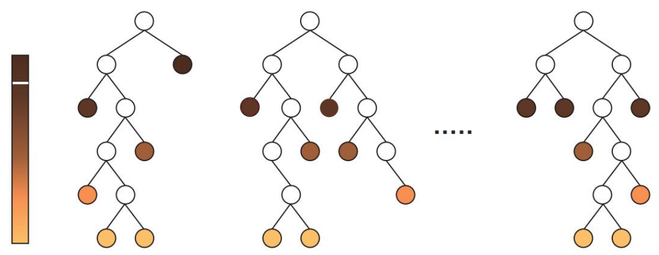

**Figure 3.** Isolation forest structure. Forest consists of the independent decision trees. To build a branching in a tree a random feature and a random splitting are selected. The tree is built until each object of a sample is isolated in a separate leaf — the shorter path corresponds to a higher anomaly score which is also illustrated by the colour. For each object, the measure of its normality is a function of the depths of the leaves into which it is isolated.

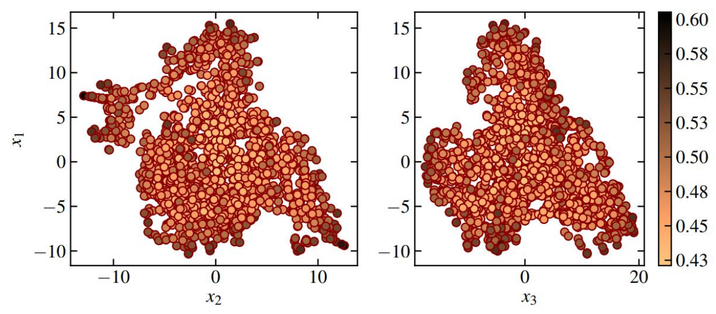


**Figure 5.** Three-dimensional t-SNE reduced data after application of the isolation forest algorithm. Each point represents a supernova light curve from the data set projected into the three-dimensional space with the coordinates (x1, x2, x3). The intensity of the colour indicates the anomaly score for each object as estimated by the isolation forest algorithm. A darker color corresponds to the objects with higher anomaly scores.

## Results:

Applying the unsupervised learning to the photometric data extracted from the Open Supernova Catalog we found ~100 outliers among a total of 1999 objects (Fig. 1). However, not all of them are necessary anomalies. Among the detected outliers there are few known cases of miss-classifications, representatives of rare classes of SNe (e.g., superluminous supernovae, 91T-like SNe Ia) and highly reddened objects. We also found that 16 anomalies classified as supernovae in [5], are likely to be quasars or stars. Light curves with GP approximation for all 1999 objects can be found at http://snad.space/osc/ and those who considered anomalous according to the criteria described in the previous section are listed in Table A1 in [6]. In our list of anomalies we found three peculiar SNe Ia (SN 2002bj, SN 2013cv, SN 2016bln - Fig. 7); two peculiar SNe type II (SN 2013ej, SN 2016ija - Fig. 6); two superluminous SNe (SN2213-1745, PTF10aagc); and two known misclassifications: SN 2006kg which is an active galactic nucleus, and binary microlensing event Gaia16aye (Fig. 8).

In summary, the isolation forest analysis identified 81 potentially interesting objects, from which 27 (33%) where confirmed to be non-SN events or representatives of the rare SN classes. Found anomalies correspond to 1.4% of the original data set of ~2000 objects which was identified demanding significantly less resources than a manual search would entail. Among these objects, we report for the first time the 16 star/quasar-like objects misclassified as SNe.

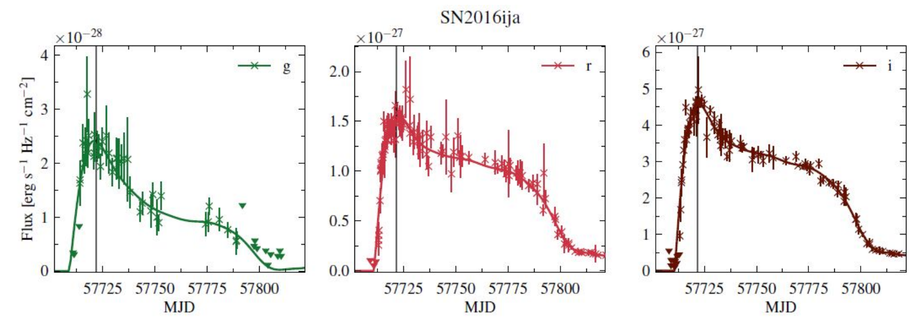


**Figure 6.** Light curves in *gri* filters of peculiar SN II 2016ija (Tartaglia et al. 2018). Solid lines are the results of our approximation by Multivariate Gaussian Process. The vertical line denotes the moment of maximum in *r* filter.

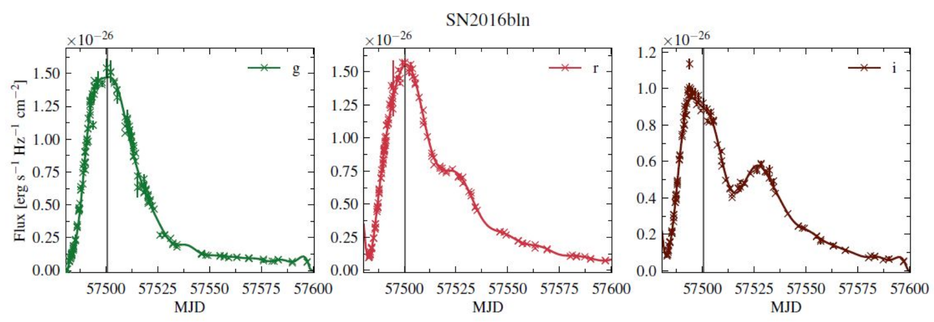


**Figure 7.** Light curves in *gri* filters of 91T-like SN Ia 2016bln (Miller et al. 2018). Solid lines are the results of our approximation by Multivariate Gaussian Process. The vertical line denotes the moment of maximum in *r* filter.

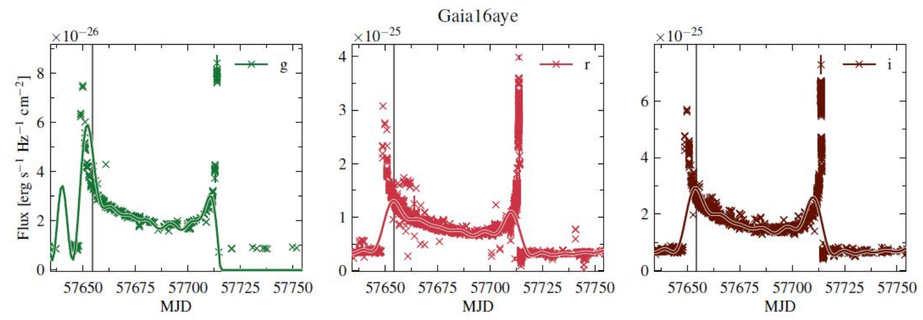


**Figure 8.** Light curves in *gri* filters of binary microlensing event Gaia16aye (http://gsaweb.ast.cam.ac.uk/alerts/alert/Gaia16aye/followup). Solid lines are the results of our approximation by Multivariate Gaussian Process. The vertical line denotes the moment of maximum in *r* filter.

## Acknowledgements:

M. Pruzhinskaya and M. Kornilov are supported by RFBR grant according to the research project 18-32-00426 for outlier analysis and LCs approximation. K. Malanchev is supported by RBFR grant 18-32-00553 for preparing the OSC data. E. E. O. Ishida acknowledges support from CNRS 2017 MOMENTUM grant. A. Volnova acknowledges support from RSF grant 18-12-00522 for analysis of interpolated LCs. We used the equipment funded by the Lomonosov Moscow State University Program of Development. The authors acknowledge the support from the Program of Development of M.V. Lomonosov Moscow State University (Leading Scientific School 'Physics of stars, relativistic objects and galaxies'). This research has made use of NASA's Astrophysics Data System Bibliographic Services and following PYTHON software packages: NUMPY (van der Walt, Colbert & Varoquaux 2011), MATPLOTLIB (Hunter 2007), SCIPY (Jones et al. 2001), PANDAS (McKinney 2010), and SCIKIT-LEARN (Pedregosa et al. 2011).

## Bibliography:

[1] Guillochon J., Parrent J., Kelley L. Z., Margutti R., 2017, ApJ, 835, 64
[2] Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press
[3] Maaten L. v. d., Hinton G., 2008, Journal of machine learning research, 9, 2579
[4] Hinton G. E., Roweis S. T., 2003, in Advances in neural information processing systems. pp 857–864
[5] Sako M., et al., 2018, PASP, 130, 064002
[6] Pruzhinskaya, M. V., Malanchev, K. L., Kornilov, M. V. et al., 2019, MNRAS, 489, 3591