Поиск аномалий в Открытом Каталоге Сверхновых звезд методами машинного обучения

М.В. Корнилов[1,2] К.А. Маланчев[1,2] М.В. Пружинская[1] Э. Ишида[3]
Ф. Мондон[3] А.А. Вольнова[4] В.С. Королев[5]

17 декабря 2018

[1]Московский государственный университет им.М.В.Ломоносова
[2]НИУ Высшая школа экономики
[3]Клермонтский университет
[4]Институт космический исследований РАН
[5]Центральный аэрогидродинамический институт им. проф. Н.Е. Жуковского

Introduction

Present state of astronomical data

Too much data is a problem

- LSST claims to produce $\sim 10PB$ per night
- 10^6 new SN per operation year
- Time-domain astronomy

Machine learning

Arthur Samuel (1959, doi: 10.1147/rd.33.0210)

> *ML is a field of computer science that gives computer systems the ability to "learn" with data, without being explicitly programmed.*

Machine learning

Tom Mitchel (1997, ”Machine Learning”, McGraw Hill)

> *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**.

Anomaly detection

Anomalies are data patterns that have different data characteristics from **normal** instances.

Anomaly detection algorithms

- Supervised and unsupervised

SN Anomaly Detection

The Open Supernova catalog

- `https://sne.space`
  - Meta info: 50K
  - Light curves: 12K, only about 2000 are useful for us.
  - Spectra: 6K
- Pros:
  - It is open
  - It is supernova catalog
- Cons:
  - Total mess

Homogenize the data

Machine learning algorithms usually work with homogeneous data

- Every input sample is multidimensional vector
- All the vectors have the same length

Homogenize the data

The Open Supernova catalog

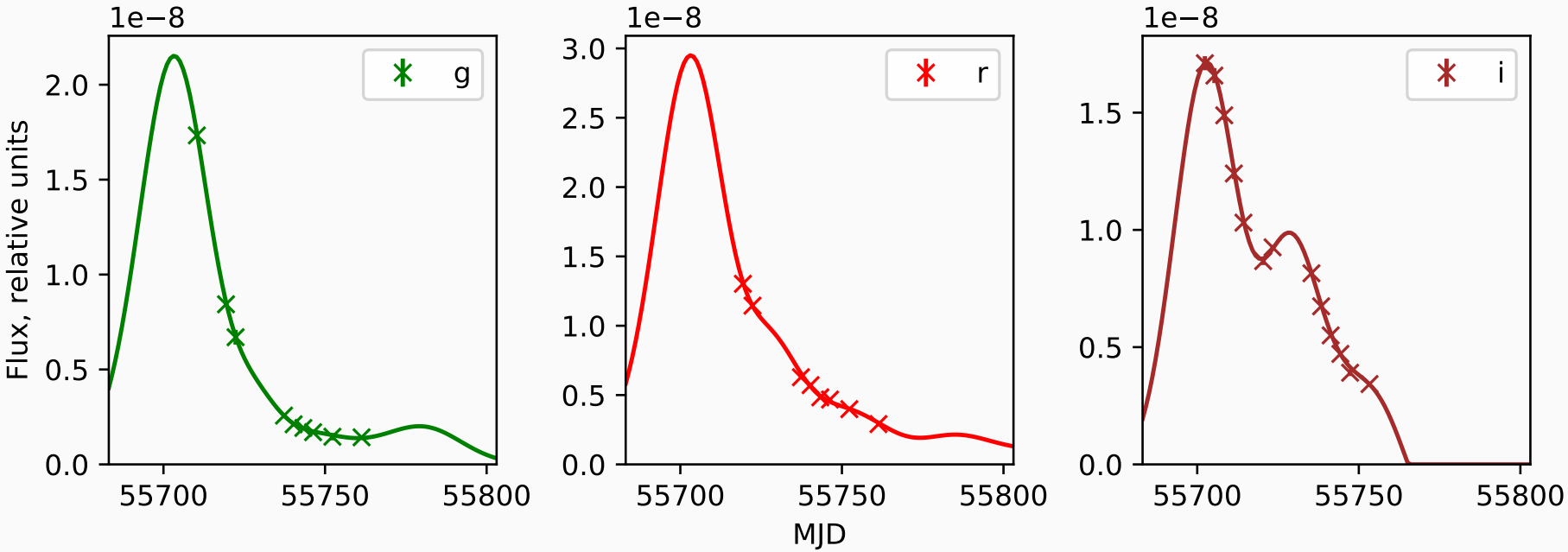
- Unevenly distributed flux measurements
- Only a few passbands usually available per light curve
- For each LC we have different time span before the maximum
- Unreliable measurement accuracy estimations

Homogenize the data

Gauss process interpolation

- Commonly used technique to interpolate unevenly distributed data onto uniform grid.
- Have to account for multiple passbands
- Have to deal with extrapolation

Improved Gauss process interpolation (PTF11dec)



- Interpolate unevenly multiband distributed data
- `https://github.com/matwey/gp-multistate-kernel`

Homogenize the data

Coping with the photometric systems

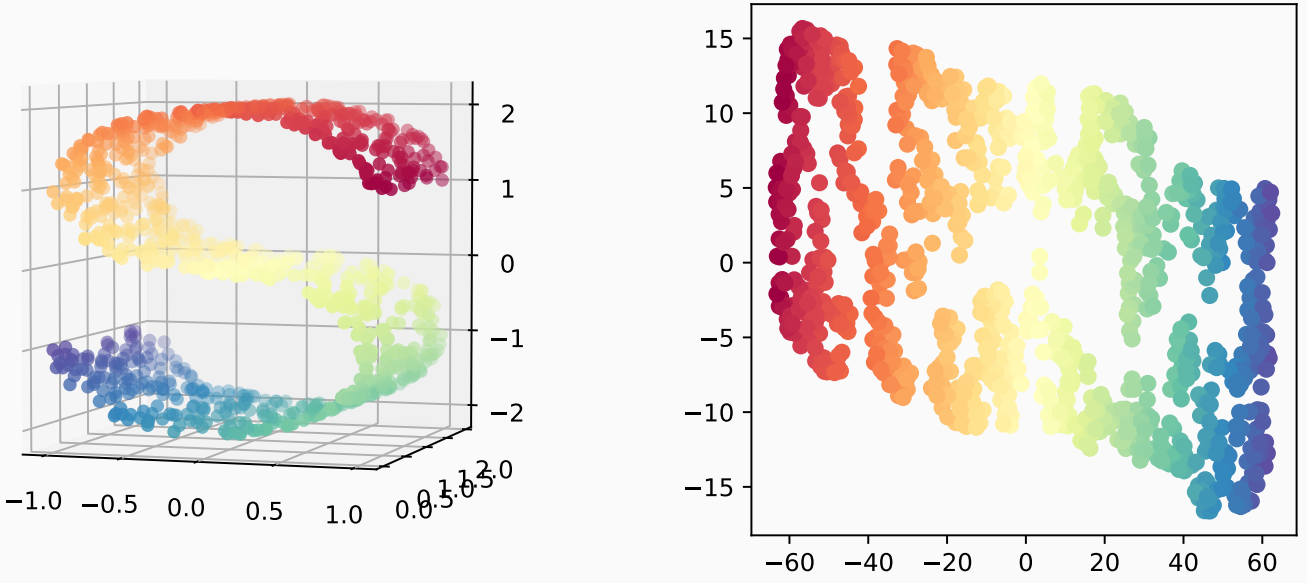
- We use *gri* passbands (most of the LCs have measurements there)
- $g'r'i'$ considered to be the same as *gri*
- Interpolated *BRI* light curves converted to *gri* by Lupton ad-hoc equations (2005).

Homogenize the data

Data normalization

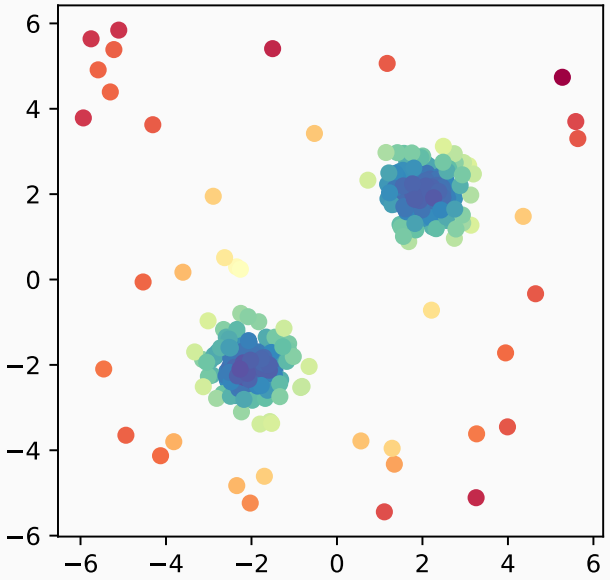
- Each light curve is 360 points (3 passbands, 120 days)
- Interpolation parameters (about 10 values)
- Normalize light curve (another column)
- Input dimension is close to 400

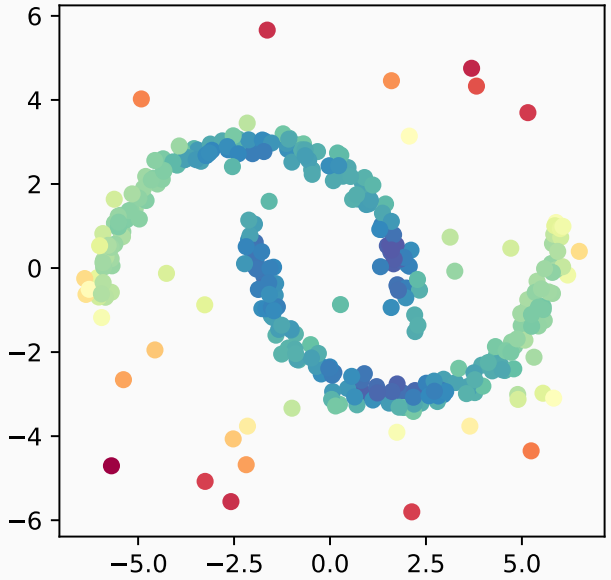
Dimensionality reduction: t-SNE



- Nonlinear dimensionality reduction technique.
- van der Maaten, L. and Hinton, G. "Visualizing High-Dimensional Data Using t-SNE" Journal of Machine Learning Research (2008)
- Let $p(\mathbf{x}_i|\mathbf{x}_j) \sim \exp\left(-\frac{1}{2}||\mathbf{x}_i - \mathbf{x}_j||^2\right)$ in high-dimensional space and find $\mathbf{y}_i$ in low-dimensional space to keep distributions "close".
- Cons: $O(N^2)$ complexity

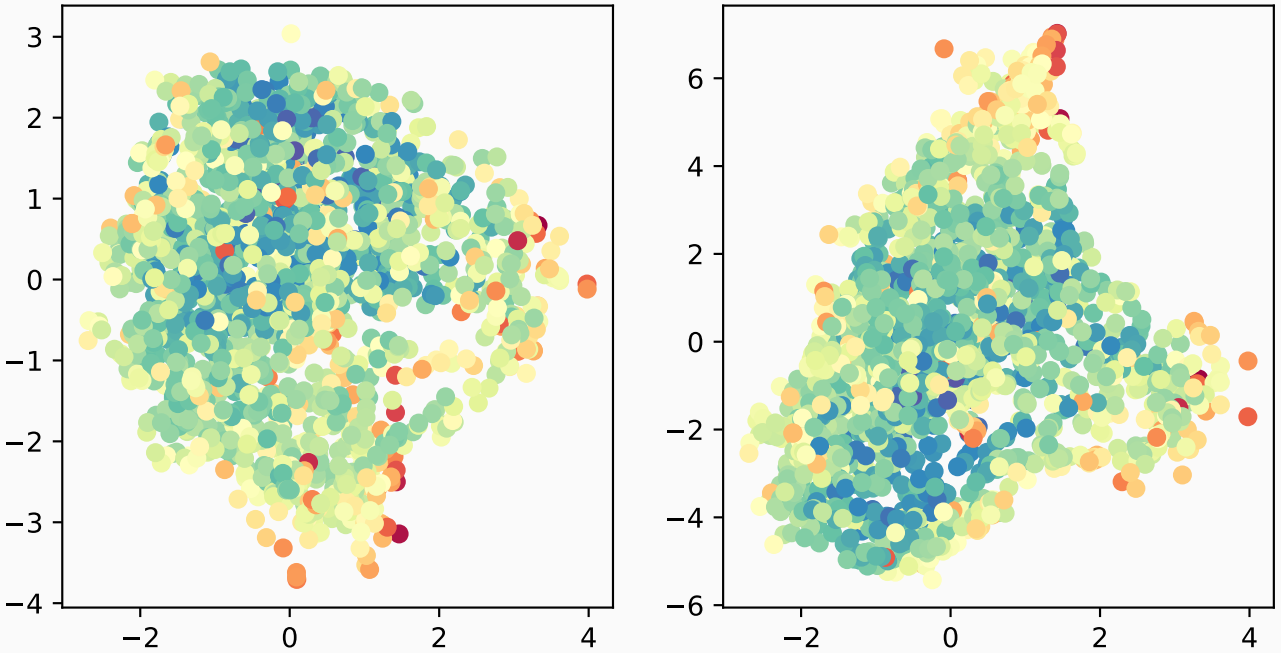
Isolation forest algorithm





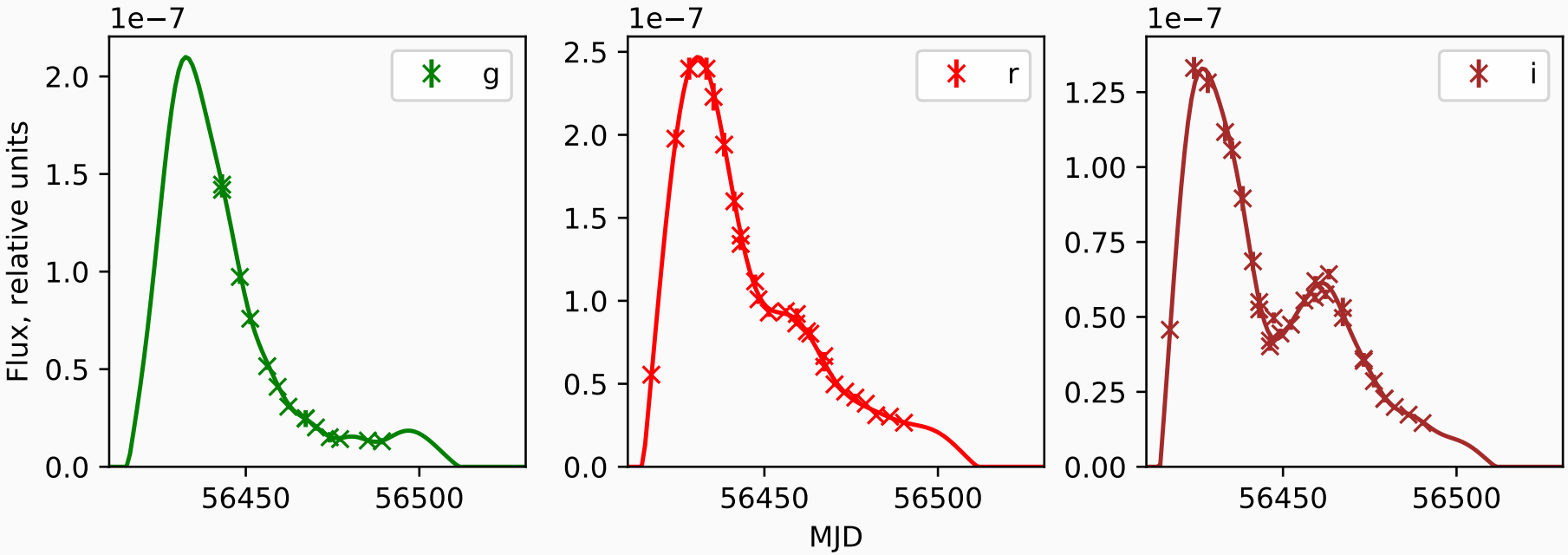
- Anomaly isolation algorithm
- Liu, F. et al. ”Isolation-based anomaly detection” ACM TKDD
- Do **not** rely on ”normal data” distribution

Isolation forest algorithm



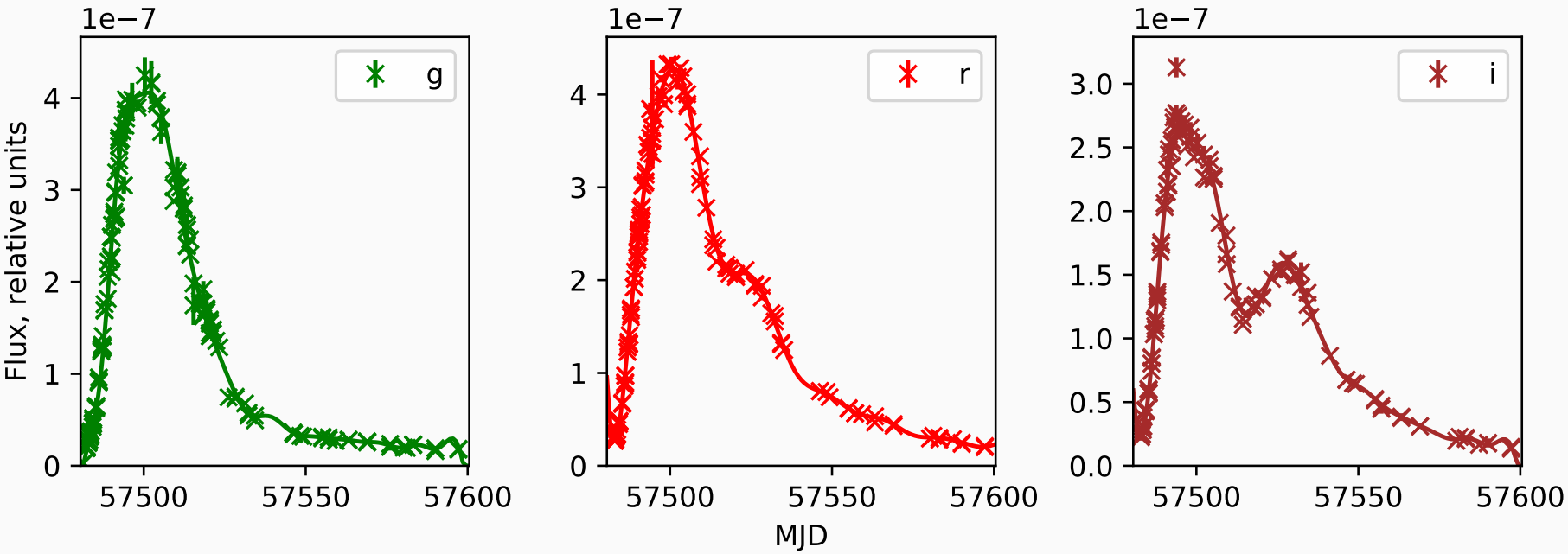
Results

SN 91-T: SN 2013cv



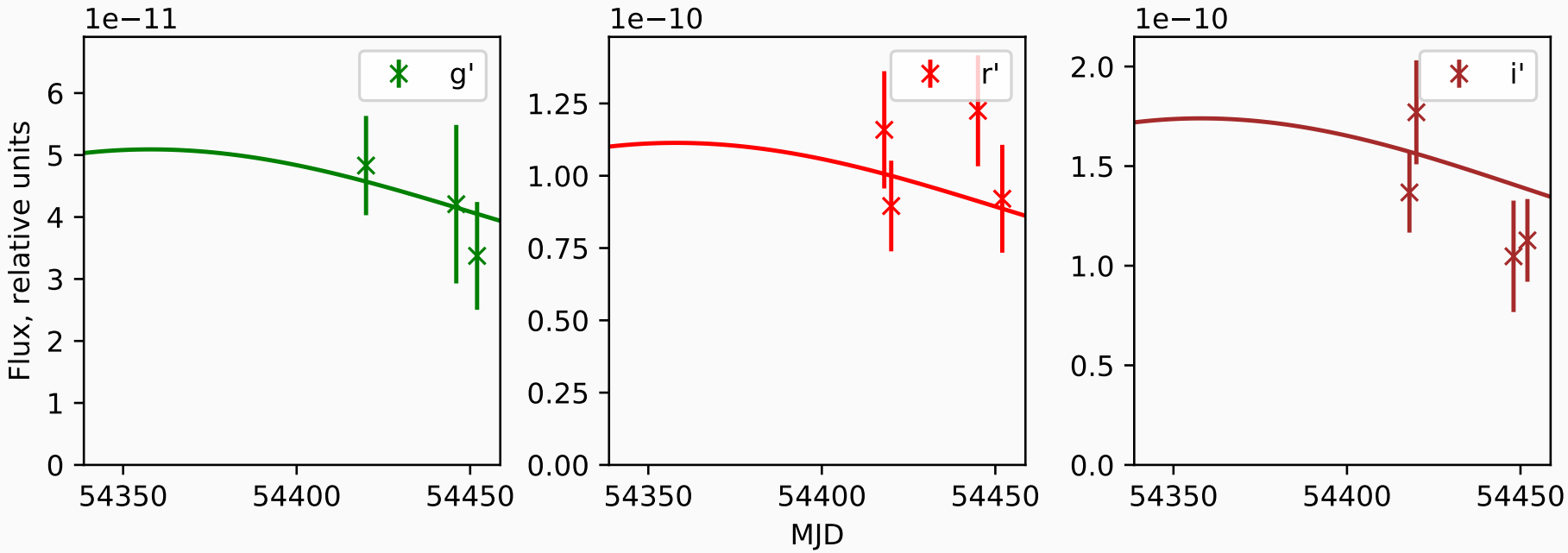
- SN 91-T looks similar to Ia. The issue that it is brighter.
- Cenko, S. et al., ATel 8909 (2016)

SN 91-T: SN 2016bln



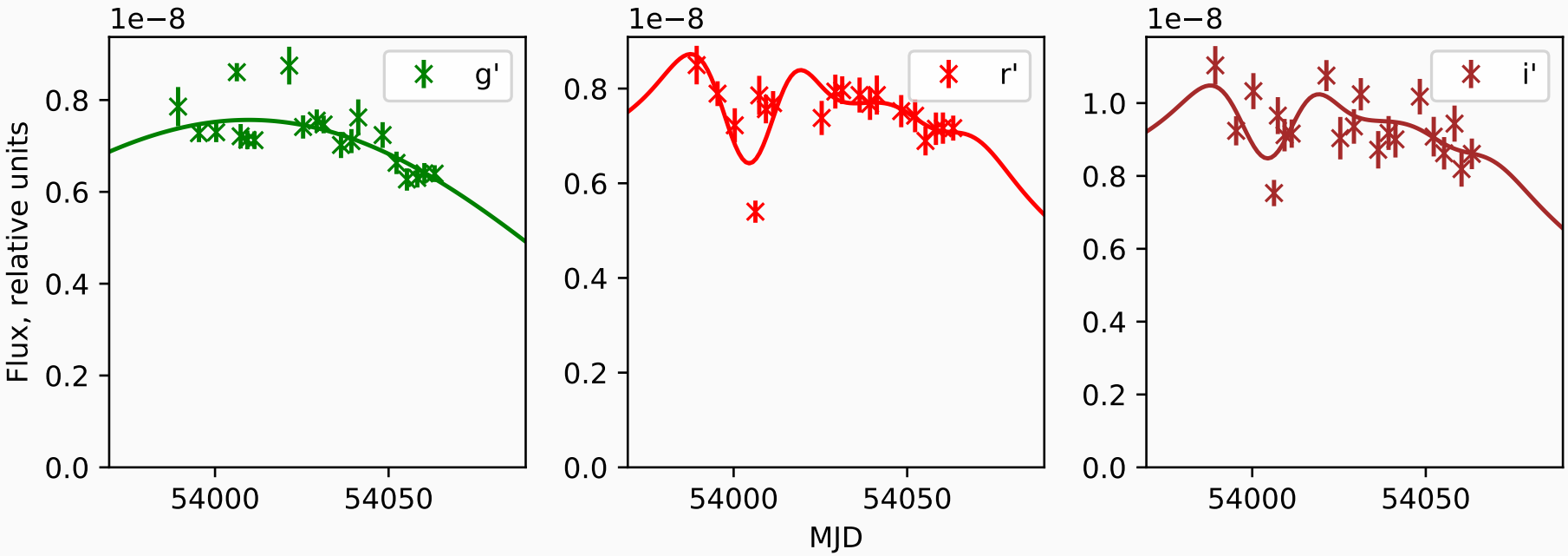
- Cao, Yi et al., AJ, Vol. 823, Issue 2, 147, 13 pp. (2016) doi:10.3847/0004-637X/823/2/147

SLSN-II: SN1000+0216

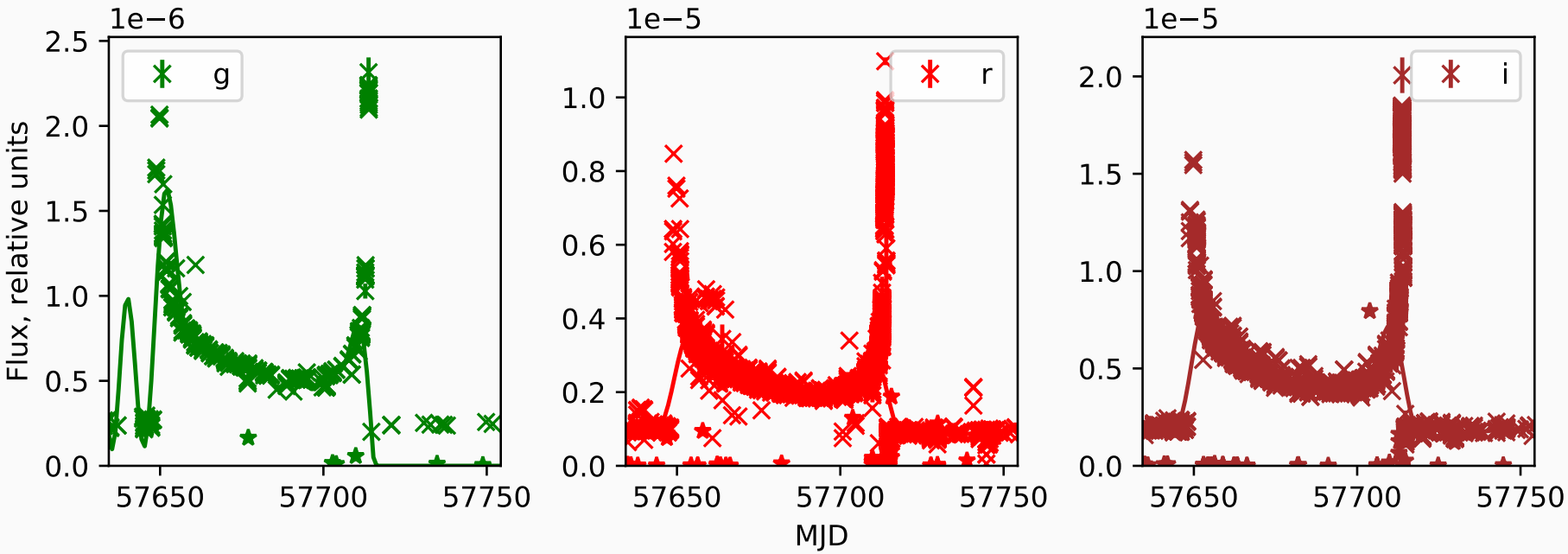


- Cooke, J et al. Nature, Volume 491, Issue 7423, pp. 228-231 (2012) doi:10.1038/nature11521

AGN: SN2006kg



Binary microlensing event: Gaia16aye



- Wyrzykowski, L. et al. ATel 9507 (2016)

Questions?