# The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Data set

We briefly describe the PLAsTiCC data set that will be tested by the Kaggle team. This note will be updated for the full release of the data to the community.

## 1. Introduction

While we think of the night sky as static, it is filled with sources of light that vary in brightness on timescales from seconds and minutes to months and years.

Some of these events are classified as *transients*, and can be the observational consequences of a large variety of astronomical phenomena. For example, the cataclismic event that occurs when a supernova explodes generates a bright signal that fades with time, but does not repeat.

Other events are classified as *variables*, since they can vary their brightness in a periodic (or aperiodic) fashion, and originate from physical process governing high density regions of the Universe such as emission from the active galactic nuclei (AGN) at the hearts of galaxies, or our position in relation to a set of observed objects (e.g. eclipsing binaries).

These relatively rapid changing objects can provide important clues about themselves and their environement - as well as the evolution of the universe as a whole (e.g. type Ia supernovae provided the first evidence of the current accelerated expansion of the Universe caused by dark energy). Crucially, the different types of transients and variables, given *different* pieces of information, and therefore, the proper classification of transients is a crucial task in observational astronomy - specially in the light of large data volumes expected for the next generation of astronomical surveys.

The two modes for characterising the light from the objects, are called spectroscopy and photometry.

Spectroscopy is the modern equivalent of using a prism to separate a beam of white light in its composite (e.g. rainbow) colours. It is a high resolution measurement which allows us to identify any emission/absorption features indicative of specific chemical elements present in the object - and are also the main source of information enabling classification. Despite being paramount for the classification task, spectroscopy is an extremely time consuming process - with integration times hanging from 20 minutes to a few hours depending on the telescope and brightness of the source.

Traditionally, classification is completely based on spectroscopic measurements - however, given the volume of data expected from the upcoming large scale sky surveys, this strategy is not sustainable. An alternative approach, based on cheaper/lower resolution measurements must be found. The spectra are described as flux per wavelength interval.

Different physical processes (and chemical elements) are found at different wavelengths of the spectrum, and so different parts are useful to distinguish between objects.

Photometry is gives the overall information of how bright the source is (how much energy the object is emitting) at a particular moment. It can be described as the low resolution counterpart of spectroscopy - or the integral of all the information stored in the spectra. The resultant photometric magnitudes $m$ are related to the flux in a given (average) wavelength through

$$F = 10^{-0.4m+11} + noise \tag{1}$$

These magnitudes are precisely the integral of the spectrum over the filter bandpasses of the instrument. A sequence of photometric observations made at different moments in time is called a light curve. It holds the information of how the energy of the source evolves with time and can also be used to characterize different types of supernovae (although the distinction between classes is much more discrete than we would hope it to be). As a consequence, for each object we will have a number of light curves in each filter (or band). For LSST, the filters are defined as the SDSS $ugrizy$ filter set. Not all bands have the same noise properties, and therefore the bands will vary in usefulness for different objects. In addition, the bands are sensitive to different physical processes in the same way as the spectrum, and so will be more or less useful to distinguish between objects.

2

In addition to being more cost effective, it is possible to obtain photometry of objects to greater distances (and volumes) in the universe. For an object of given brightness (luminosity), the flux received on earth decreases with the distance to the object as

$$F = \frac{L}{4\pi d_L^2} \tag{2}$$

Hence if the intrinsic luminosity of an object is known, then its flux can be used to determine the distance to the object. This distance is an important tool in classification, since different objects are more likely to be found within or outside of our own galaxy.

When working in magnitudes, we connect the distance to the magnitudes via the distance modulus $\mu$:

$$\mu = m - M = 5 \log d_L + 5 \log \left( \frac{c}{H_0} \right) + 25, \tag{3}$$

where $M$ is the (unknown) intrinsic brightness of the object, $d_L$ is measured in mega-parsec (Mpc), $c$ is the speed of light and $H_0 \simeq 70$ km/s/Mpc is the Hubble constant. Programs exist to fit for the distance modulus $\mu$ given the light curve points and models for the type of object.

New instruments, like the Large Synoptic Survey Telescope (LSST) - scheduled to begin observations in 2022 - will produce an unprecedented number of light curves. Once these transients are detected, we rely on agreements with other telescopes in order to acquire a small number of spectroscopic observations.

The final connection to a model of the cosmology of the universe, and the distances to objects comes through the redshift, $z$. Redshift is an emperical quantity that is defined by measuring the difference in the observed wavelength $\lambda_o$ of a given feature (e.g. in the spectrum described above) compared to the emitted wavelength $\lambda_e$, or

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e}. \tag{4}$$

It is the light analogy of the Doppler effect that we are used to in sound. It is clear then that using a spectrum to determine the redshift of an object gives the most precise result (with the smallest error $\sigma$). However photometry can also be used to determine a redshift for an object, the so-called photometric redshift, with a larger uncertainty. Given that photometric error relies on prominent features moving between the (large) bandpasses, it is also possible to insert 'catastrophic' redshift error, where the redshift of the object is misassigned. These errors are rare (roughly $2\%$ of the total number of objects), however can pose serious problems for classification of objects.

The distance modulus is connected to the cosmology as a function of redshift $z$ and the components of the universe. We do not expect participants of this challenge to compute these functions, and will provide the distane modulus and error on the modulus as if the data had been run through a light-curve model fit.

We will describe the data in the following sections, and discuss the metrics used to classify objects in a separate note.


## 2. The data

The lightcurve data is made of non-homogeneously sampled, non-periodic time series with correlated errors obtained in several wavelength filters.

A database over all objects will be provided. The following data are provided:


- objectID

- right ascension and declination (the sky co ordinates of the object)

- MJD (Mean Julian Date) when a trigger was generated

- the value of the Milky Way extinction at that position on the sky

- redshift

- redshift error (this will be small for the spectroscopic training set and large for the photometric test set)

- distance modulus

- error on the distance modulus

Each object will also have a list of

- MJD (Mean Julian Date)
- flux in each of the LSST bands
- flux error in each of the LSST bands
- signal-to-noise value for the flux in each of the LSST bands
- photometry flag

The training set will also contain the simulated model name, while the test data will not: it is the task of those participating to determine the classification flag of the test set.

## 2.1. Training data

The training data will be a subset of $\sim 5000$ objects taken from the larger test data set. The relative rates of the data will be adjusted slightly in the training sample. In addition, the test data will contain some rare objects that are not seen in the training data. The final classification classes will therefore contain an 'other' class to be used on any object that the classifiers have not seen before.

# 3. Conclusion

## Acknowledgments