Machine Learning with ZTF

Ashish Mahabal, Brian Bue, Kevin Burdge, Dave Cook, Mansi Kasliwal, Thomas Kupfer, Frank Masci, Adam Miller, Umaa Rebbapragada, Quan-Zhi Ye

Section 1: Scientific motivation:
As we go to ZTF, there are many changes we will see: An increase in collection area, new CCDs, an extension into parts of the plane that could see more vignetting effects, a larger number of objects (including variables and transients, and a possible shift from focus on individual objects to populations of different sizes). With 700 exposures per night, ~50,000 subtraction images, and 1 million target point source candidates, and 200,000 streak candidates are expected nightly.

The aim of the machine learning (ML) activity will be to automate procedures, and allow incorporation of newer observations, and corresponding feedback in a straightforward manner. At the same time it will be important for astronomers to retain control over tweaking follow-up target selection, and that means an organic interface with one or more Marshals. The ML will be agnostic to cadence (be it the 3-5 day, nightly, or higher cadence), and will heavily make use of past experience (including but not limited to [i]PTF pipelines) as well as newer techniques. The non-dependence on cadence and specific follow-up observations means that this WP will skip Section 2 (Proposed observations) and Section 3 (Supporting observations).

We envisage four main areas where ML can make a real difference:

1. RealBogus for point sources
2. RealBogus for streaks
3. Classification/Characterization of real transients and variables
4. Transfer learning using other datasets (PTF/iPTF), but also CRTS/Gaia/Pan-STARRS/…

From a science perspective #3 is most important. In order to get there #1 is crucial, while #4 is essential to accomplish the RealBogus goals as quickly as possible using previous datasets that share some aspects with ZTF (e.g., area, cadence, filters, etc.). #2 utilizes similar architecture to #1, but involves different characterization of the sources. Figure 1 shows a schematic block diagram covering the data flow and decision boundaries.

All the four tasks above are in progress using iPTF data, and will be improved upon in terms of techniques used. The biggest requirement for ZTF will be matching the rapid generation of new samples with labeled real and bogus sources, and not just with spectroscopic follow-up. Getting up to speed includes a few unknowns as of now: (1) Who will do the labeling (spread over different science teams), and (2) how the labeling will be done (some spectroscopy and a lot of known variable sources). These are connected with the Marshal, and is also with the early commissioning data. While July 1 is the likely date for first light, it is by August 1 that we can expect a few processed data frames (for each chip, and verified by humans) to come through. New chips necessarily mean artifacts we have not seen before. These include ghosts and glints

(from bright stars; bands along ribs may have to be excluded). There will be a filter-dependant effect to take into account. Effort will be made to get pre-first-light samples of artifacts that are available. Known detector issues will be factored in (e.g. those seen from WASP chips).

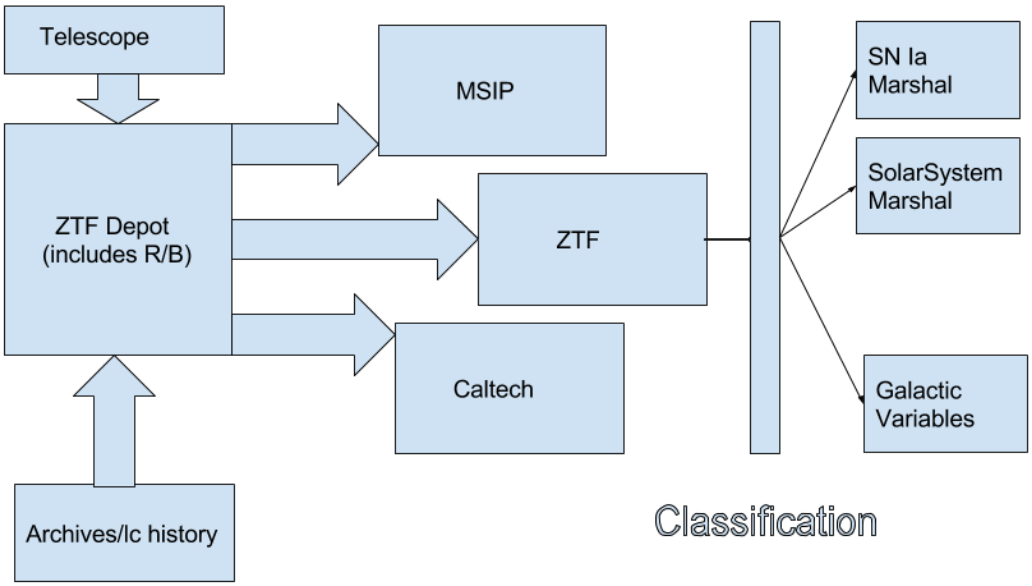

Figure 1. ZTF Depot contains #1, and #2, the RB routines for point sources and streaks. It also has access to archival datasets. It serves the ZTF, MSIP, and the caltech portions of the time. The classification layer comes next. How exactly that interfaces with the marshals required by different science groups is somewhat nebulous at the moment. Many details of the ZTF depot have already been worked out.

The effort will initially be directed towards the first year data when least is known about systematics. During the first year appropriate changes in the approach will be affected for the remaining period. In the longer run it will also be useful to reorient this so that the effort is in line with a LSST broker-like configuration.

Marshals will need to have visual interfaces, fault monitoring, and database design appropriate for labeling, feedback, and follow-up, and well integrated with the rest of the system.

Details on the work done on these aspects for [i]PTF are included below.

Rapid Deployment of RB Classifiers: the Case for Domain Adaptation: Effective automated transient classification requires a representative set of annotated data to be available to train a classifier to distinguish between real transients versus bogus detections (e.g., image artifacts). PTF and iPTF addressed these challenges with RealBogus (RB), statistical classifiers that scored candidate transients and streaks from zero (bogus) to one (real).

The success of RB is predicated on training data that is well sampled with respect to the true distributions of real and bogus candidates on any given night. Over time, the PTF and iPTF Real-Bogus systems took advantage of spectroscopically-confirmed detections (e.g., supernovae, variable stars, gap transients, cataclysmic variables andor novae) and a large collection of image subtraction artifacts.

Well-sampled RB training sets will not be available during the early stages of ZTF. Furthermore, updates to the ZTF image processing pipelines will impact any RB models, as the statistical distributions of the new observations can, as a result, significantly differ from the existing training data. The lack of training data, and (anticipated future) major pipeline adjustments will result in inaccurate predictions.

There are, effectively, two choices to deal with the lack of initial training data, and the anticipated changes in the pipeline software: (1) Automated vetting following commissioning observations. This can be done in conjunction with a machine learning technique called active learning, which aims to efficiently utilize human labeling resources to build high-quality training data. (2) Use domain adaptation to transform training data of real astrophysical iPTF sources into ZTF data characteristics.

Domain adaptation techniques – which seek to reconcile differences between data captured under similar (but not identical) measurement regimes – provide a potential solution to these issues. For example, we examined an instance of “data shift” caused by a data pipeline upgrade when PTF was upgraded to iPTF in Jan 2013. We demonstrated that domain adaptation techniques substantially improved transient classification accuracy across the PTF and iPTF measurement regime. These results suggest that a similar approach would be beneficial to bootstrap transient classification systems for ZTF. The combination of both manual vetting, in the early stages of ZTF, and domain adaptation will enable the rapid deployment of the RB systems necessary to achieve the ZTF science goals. Following the initial deployment of RB, future efforts will incorporate data from Pan-STARRS, CRTS, Gaia, and SDSS and utilize deep learning techniques to improve the overall performance of the new classifiers.

Real-time Detection of Small Near-Earth Asteroids in ZTF: In addition to the RB point-source classifier, a separate RB model was developed to identify streaks from near-Earth asteroids in iPTF. This approach leverages a nearly identical feature extraction methodology with minor modifications to capture asteroid-specific morphological features.

Future ML Projects for ZTF: Once the RB ML models have been incorporated into the ZTF pipelines, we will develop new methods to identify and eliminate training set contamination. We have built an active learning prototype that identifies candidates that are likely mislabeled, and presented those candidates to the science teams to be re-labeled. We have also developed a new way to randomly sample against the PTFIDE database to ensure our bogus sample is not overly biased towards certain types of artifacts and well represents the full distribution of observing conditions. Finally, we have used ML to analyze the frequency of certain types of

bogus artifacts produced by image subtraction in order to aid understanding of the software and facilitate improvements.

Another area of potential improvement is the implementation of software capable of detecting large deviations between nightly data characteristics and training data characteristics. This information will alert both the science and software pipeline teams that the Real-Bogus system may be performing sub-optimally. If such deviations persist over time (indicating a permanent shift in data characteristics), then the domain adaptation techniques discussed above could be used to bootstrap the initial model, or trigger a process to retrain and validate the existing model with the latest observations. .

Section 4: Expertise to undertake project

iPTF ML included members from Caltech, IPAC, JPL, NWU to carry out the tasks outlined above. Continuing with the team will be useful to avoid transfer learning at human level.

Tools: Fast databases, GPUs for processing will be required. Disks for the main project can be reused.

Section 5: Manpower and time-line

All co-authors will be involved in various aspects of the ML development. Additionally various science team members will be required to help with early labeling of real and bogus sources.

Table 1: Task dates, and FTEs. The estimates on FTEs are not concrete yet as the interfacing with marshals etc. is not clear.

Task	**Start Date**	**End Date**	**FTE**
Extract lightcurve features from PTF+iPTF	03/01/2017	10/31/2017	0.40
Create variability priors from PTF+iPTF	03/01/2017	10/31/2017	0.20
Collect real ZTF data (phase 0)	08/01/2017	11/01/2017	0.10
Label ZTF data for RB (phase 0)	08/02/2017	10/01/2017	0.25
Train version 0.0 RB for point sources	10/02/2017	10/20/2017	0.40
Integrate version 0.0 RB for point sources into pipeline and test	10/21/2017	11/01/2017	0.10
Collect real streaks labels from ZTF (phase 0)		1/1/2018	0.25
First phase of transfer learning		2/1/2018	0.50
First phase of real-time classification		3/1/2018	0.50
First phase of streaks RB		11/1/2018	0.40
Full versions of RB, transfer learning, classification		11/1/2018	1.00