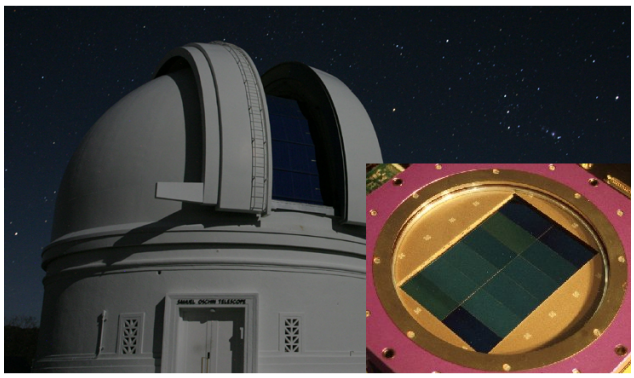
Realtime Astronomical Time-series Classification and Broadcast Pipeline

Dan Starr (dstarr@astro.berkeley.edu) – *UC Berkeley,* USA
Josh Bloom (jbloom@astro.berkeley.edu) – *UC Berkeley,* USA
John Brewer (bizard@propellerheads.com) – *UC Berkeley,* USA

The Transients Classification Pipeline (TCP) is a Berkeley-led, Python based project which federates data streams from multiple surveys and observatories, classifies with machine learning and astronomer defined science priors, and broadcasts sources of interest to various science clients. This multi-node pipeline uses Python wrapped classification algorithms, some of which will be generated by training machine learning software using astronomer classified time-series data. Dozens of context and time-series based features are generated in real time for astronomical sources using a variety of Python packages and remote services.



Palomar 48 inch telescope, PTF survey detector on an older instrument (inset).

Project Overview

Astronomy is entering an era of large aperture, high throughput surveys with ever increasing observation cadences and data rates. Because of the increased time resolution, science with fast variability or of an explosive, "transient" nature is becoming a focus for several up and coming surveys. The Palomar Transients Factory (PTF) is one such project [PTF]. The PTF is a consortium of astronomers interested in high variability, "transient" science, and which has a dedicated survey telescope as well as several follow-up telescope resources.

Berkeley's Transients Classification Pipeline (TCP) is a Python based project that will provide real-time identification of transient science for the Palomar Transients Factory's survey telescope, which goes online in November 2008.

The PTF's survey instrument is a ~100 megapixel, 7.8 square-degree detector attached to Palomar Observatory's Samuel Oschin 48 inch diameter telescope. Taking 60 second exposures, this instrument produces up to 100 Gigabytes of raw data per night. Immediately following an observation's exposure and read-out from the detector, the data is uploaded to Lawrence Berkeley Laboratory for calibration and reduction, which results in tables of positions and flux measurements for objects found in each image. The TCP takes this resulting data and after several steps, identifies astronomical "sources" with interesting science classes, which it broadcasts to follow-up scheduling software. PTF affiliated telescopes are then scheduled to make follow-up observations for these sources.

The TCP is designed to incorporate not only the Palomar 48 inch instrument's data stream but other telescope data streams and static surveys as well. The software development and testing of the TCP makes use of SLOAN Digital Sky Survey's "stripe 82" dataset [SDSS] and the near-infrared PAIRITEL telescope's real-time data stream [PAIRITEL]. Prior to the commissioning of the Palomar instrument, the TCP will be tested using a historically derived data stream from the preceding Palomar Quest survey on the Palomar 48 inch telescope.

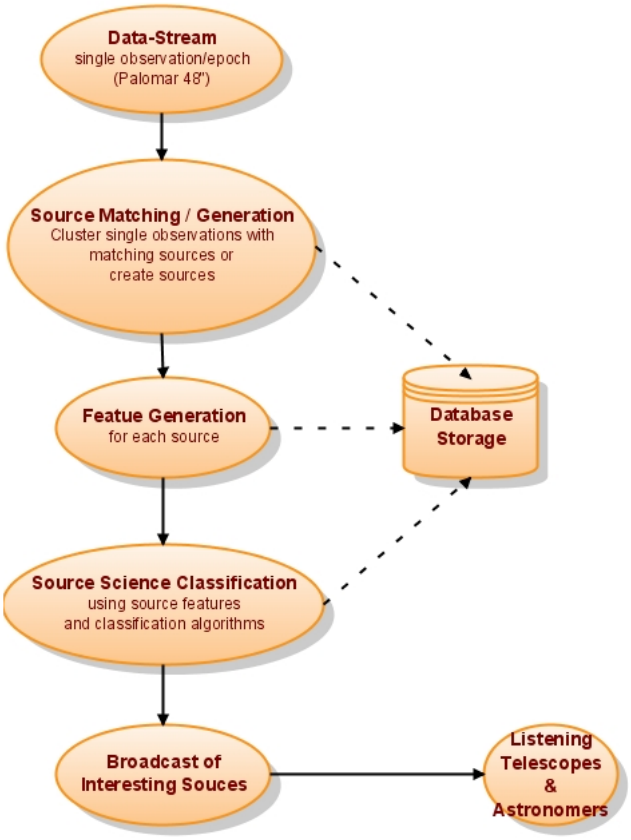
A long term goal of the Transients Classification Pipeline is to produce a scalable solution to much larger, next generation surveys such as LSST. Being a Python based project, the TCP has so far been relatively easy to implement and scale to current processing needs using the parallel nature of "IPython". Although the TCP's processing tasks are easily parallelized, care and code optimization will be needed when scaling to several orders of magnitude larger next generation survey data streams.

Two of PTF's several follow-up telescopes: Palomar 60 inch, PAIRITEL (inset).

TCP Data Flow

The TCP incorporates several data models in its design. Initially, the telescope data streams feeding into the TCP contain "object" data. Each "object" is a single flux measurement of an astronomical object at a particular time and position on the sky. The pipeline clusters these objects with existing known "sources" in the TCP's database. The clustering algorithm uses Bayesian based methods [Bloom07] and sigma cuts to make its associations. Each source then contains objects which belong to a single light source at a specific sky position, but are sampled over time. Objects not spatially associated with any sources in the TCP database are then used to define new sources.

Once a source has been created or updated with additional object data points, the TCP then generates a set of real-number "features" or properties for that source. These features can describe context information, intrinsic information such as color, or properties characterizing the source's time-series "light-curve" data. The source's features are then used by classification algorithms to determine the most probable science classes a source may belong to.



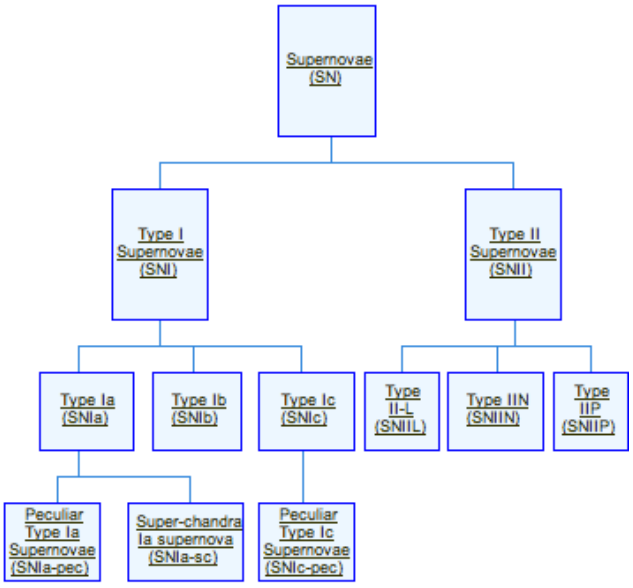
Data flow for the TCP.

Sources which match with a high probability science classes that interest PTF collaborators, will be broadcast to the PTF's "Followup Marshal" software. The Followup Marshal delegates and schedules follow-up observations on various telescopes.

One important design point of the Transients Classification Pipeline is that it allows the addition of new data streams, feature generators, and science classification algorithms while retaining its existing populated database. To meet this constraint, the TCP will use autonomous source re-evaluation software, which traverses the existing source database and re-generates features and science classifications for stale sources.

Science Classification and Follow-up

Time-variable astronomical science can be described as a hierarchy of science classes. At the top level, there are major branches such as "eruptive variables" or "eclipsing binary systems", each of which contain further refined sub-classes and child branches. Each science class has an associated set of constraints or algorithms which characterize that class ("science priors"). Determining the science classes a source belongs to can be thought of as a traverse along the class hierarchy tree; following the path where a source's characteristics ("features") fall within a science class's constraints. These science priors / classification algorithms can be either astronomer defined or generated by machine learning software which is trained using existing classified source datasets.



12 of the ~150 science classes in the current TUTOR light-curve repository.

In the case of science types which are of particular interest to the PTF group, science priors can be explicitly defined by experts in each field. This is important since PTF's primary focus is in newly appearing, sparsely sampled sources which are tricky to identify using only a couple data points. As the consortium develops a better understanding of the science coming from the Palomar 48-inch data stream, these priors can be refined. The TCP will also allow different astronomers or groups to define slightly different science priors, in order to match their specific science definitions and constraints.

Besides using astronomer crafted science priors, a primary function of the TCP is its ability to automatically

generate algorithms using an archive of classified light-curves. These algorithms are then able to distinguish science classes for sources. Also, as follow-up observations are made, or as more representative light-curves are added to the internal light-curve archive, the TCP can further refine its classification algorithms.

In a future component of the TCP, a summary of the priors / algorithms which identify a source's science class may be added to an XML representation for that source. The XML is then broadcast for follow-up observations of that source. The source's priors in this XML may be useful for follow-up in cases where the TCP was unable to decide between several science classes for that source. Here, software could parse the priors and identify any features which, if better sampled, would be useful in discerning between these unresolved science classes for that source.

The PTF's "Followup Marshal" will receive TCP's broadcasted source XMLs and then schedule follow-up observations on available telescopes. In the case of a source with unresolved science classes, the Followup Marshal may be responsible for choosing a follow-up telescope based upon the priors mentioned in the source XML.

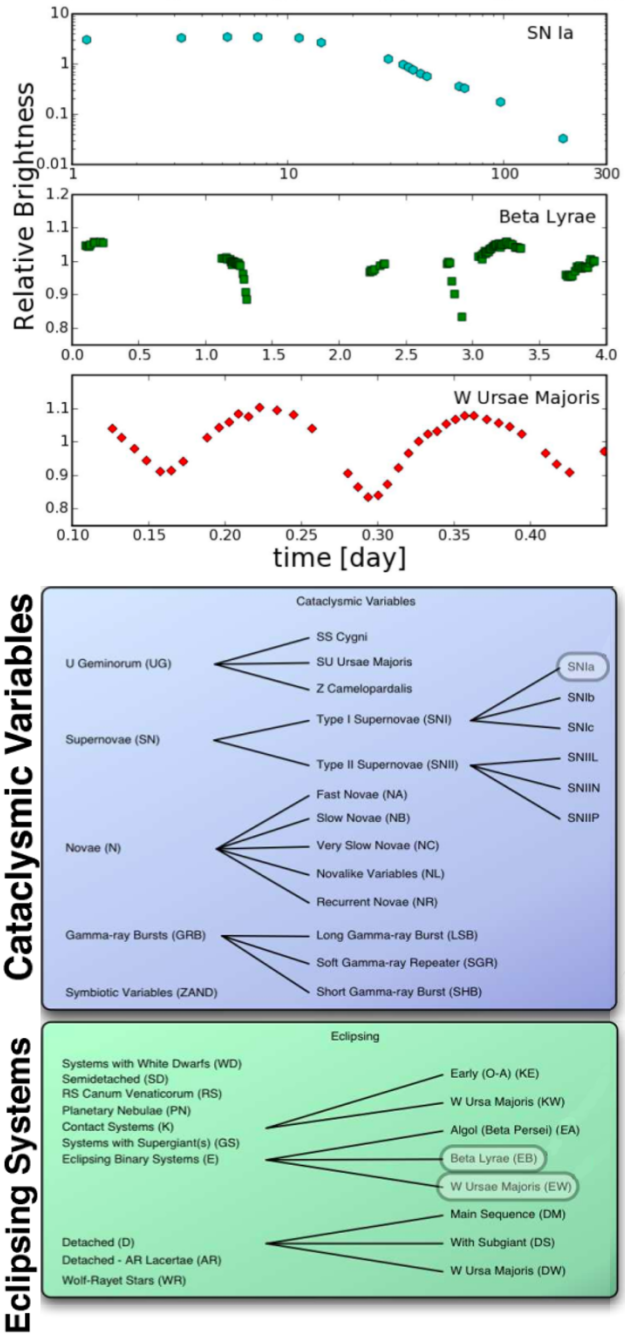
Source Features

To distinguish between a variety of different science classes, many different features need to be generated for a source. Some features represent context information, while others represent intrinsic, non-time-series stellar properties. A third feature set is derived from a source's time-series light-curve.

Context features contain source properties such as the distance of the galaxy nearest to a source, which could represent whether the source resides within a galaxy. In the case that the source closely neighbors a galaxy, the line-of-sight distance to that galaxy would also be a useful context feature of that source. The galactic latitude of a source is another context feature which roughly correlates to whether or not that source is within our galaxy. Some of these features require retrieval of information from external repositories.

Intrinsic features are neither context related or derived from time varying light-curves. The color of a stellar source is one property which astronomers tend to consider as static and unvarying over observable time.

As for time-series derived features, the TCP uses re-sampling software to make better use of locally archived example light-curves in order to generate classification algorithms applicable to the data stream instruments.



Time-series light-curves (above) and their corresponding science classes (below).

Light-Curve Resampling

Currently the TCP has an internal repository of light-curves which we've named TUTOR. As of August 2008, TUTOR contains 15000 light-curves representing 150 science classes, and derived from 87 papers and surveys. Since different instruments and telescopes were used to build the light-curves in this repository, one aspect of the TCP is to re-sample this data to better represent the observation cadences and instrument capabilities of TCP's incoming data streams.
Features generated from these re-sampled light curves are then used as training sets for machine learning software. The machine learning software then produces

science classification algorithms that are more applicable to the incoming data streams and which can be incorporated with existing science identification priors.

Python Use in the TCP

The TCP is developed in Python for several reasons. First, the TCP's primary task, generating features and determining the science classification of a source, is easily parallelized using Python. We've found the parallel aspect of IPython (formerly in the "IPython1" branch) performs well in current tests and should be applicable to the PTF data stream in November. Also, IPython's straightforward one-time importing of modules and calling of methods on client nodes made migration from TCP's original single-node pipeline trivial.

The TCP incorporates astronomer written algorithms in its feature generation software. Python makes for an easy language which programmers with differing backgrounds can code. The TCP makes use of classification code written in other languages, such as "R", but which are easily wrapped in Python. In the future, as we converge on standard science classification algorithms, this code may be re-implemented using more efficient numpy based algorithms.

An example where Python enabled code re-use, was the incorporation of PyEphem into TCP's minor planet correlation code. PyEphem wraps the C libraries of popular XEphem, which is an ephemeris calculating package.

Finally, Python has allowed TCP to make use of several storage and data transport methods. We make use of MySQL for our relational database storage, and have used the Python "dbxml" package's interface to a BerkeleyDB XML database for storage of XML and structured data. The TCP makes use of XMLRPC and socket communication, and smtplib may be used to broadcast interesting follow-up sources to astronomers.

Acknowledgments

Thanks to Las Cumbres Observatory, which partially funds the hardware and software development for the Transients Classification Project. Also thanks to UC Berkeley Astronomy students: Maxime Rischard, Chris Klein, Rachel Kennedy.

References

[PTF] http://www.mpa-garching.mpg.de/~grb07/Presentations/Kulkarni.pdf

[SDSS] Ivezi, Z., Smith, J. A., Miknaitis, G., et al., SDSS Standard Star Catalog for Stripe 82: Optical Photometry, AJ, 2007, 134, pp. 973.

[PAIRITEL] Bloom, J. S. and Starr, D. L., in ASP Conf. Ser. 351, ADASS XV, ed. C. Gabriel, C. Arviset, D. Ponz, E. Solano, 2005, pp. 156.

[Bloom07] Bloom, J.˜S., in HTU Proceedings 2007, Astron Nachrichten 329 No 3, 2008, pp. 284-287.