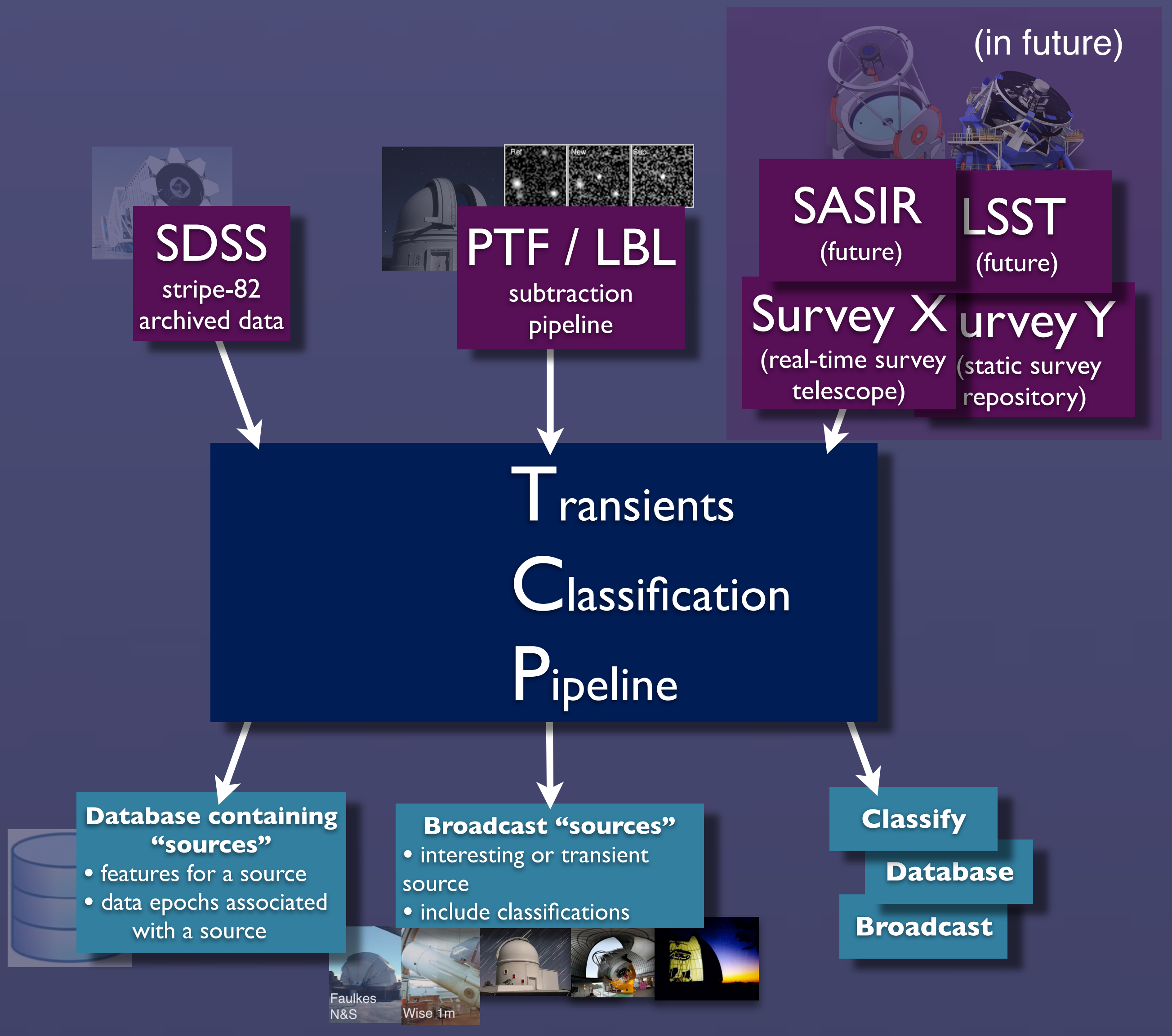
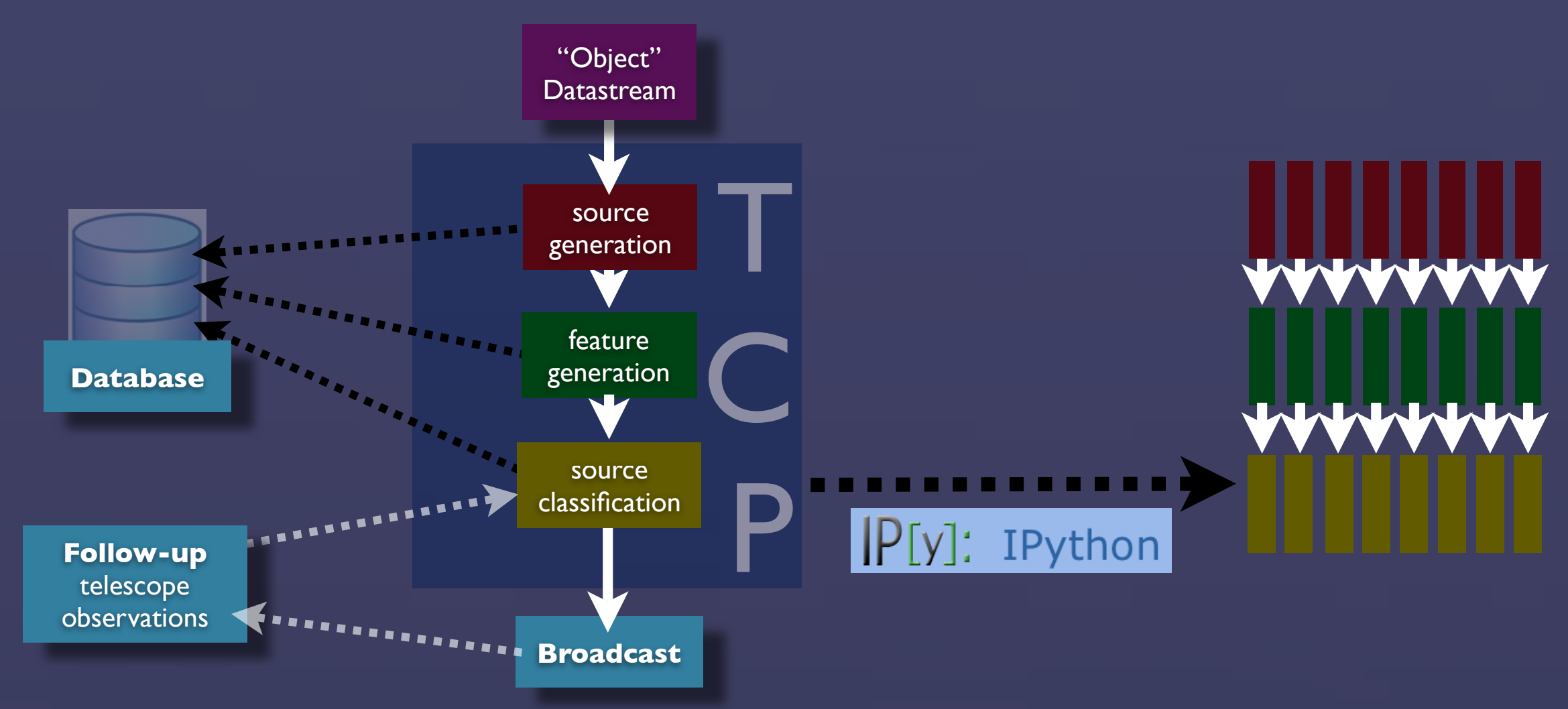
Source Discovery and Time Series Classification Using Berkeley's Transient Classification Pipeline

Berkeley Astronomy: Dan Starr, Josh Bloom, Justin Higgins, Dovi Poznanski, Maxime Rischard, Nat Butler, Chris Klein, Rachel Kennedy, Adam Morgan
San Francisco State University: John M. Brewer ***Berkeley Statistics***: Noureddine El Karoui, John Rice
Berkeley CS: Massoud Nikravesh, Martin Wainwright ***Lawrence Berkeley Lab***: Peter Nugent, Horst Simon ***Los Alamos Nat. Lab. / UC Santa Cruz***: Damian Eads

The Berkeley Transients Classification Pipeline (TCP) is a parallelized source identification and classification pipeline which ingests several data streams and emits classified science events.

Currently, the TCP ingests and classifies subtracted sources from the Palomar Transient Factory 48" telescope survey. The 7.8 sq-deg, 100 Mpix instrument samples the sky at 120 seconds, producing ~1 Million source detections per night.





Each new observation is first associated with a source in a database, then "feature" attributes are regenerated which characterize that source, and finally science classifications are made using those features.

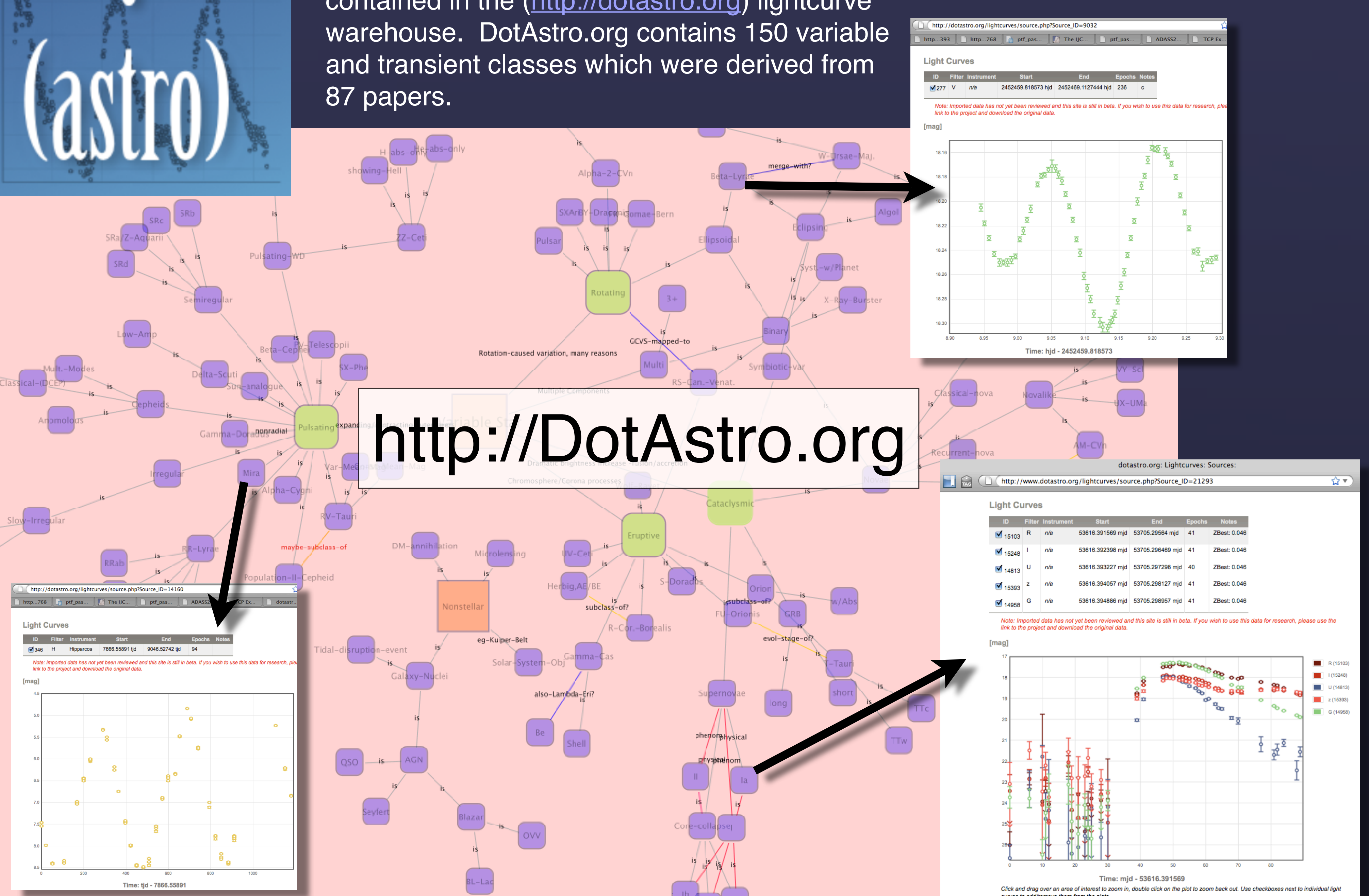
This triplet of steps for a single observation are parallelized over 22-110 CPU cores using IPython.

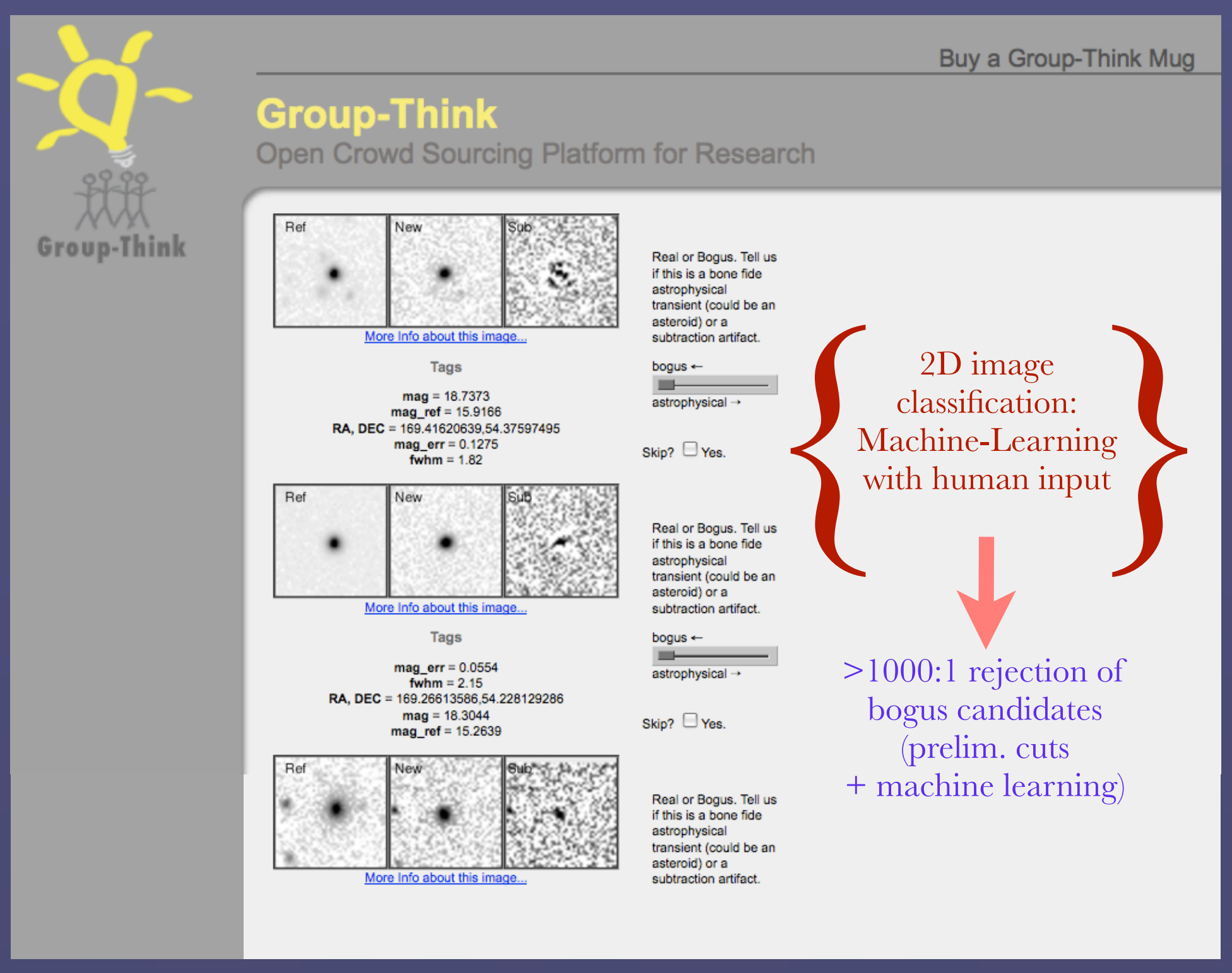
The real-time TCP is generally distributed over 6 cluster computers, although a 96 core beowulf cluster is occasionally included for re-evaluation purposes.

Additionally, we've recently started reclassification and exploration of classifier and attribute parameter spaces using Hadoop on Yahoo!'s 4000 core M45 cluster. This was made possible once we ported WEKA to the Hadoop environment.

hadoop + WEKA The University of Waikato YAHOO!

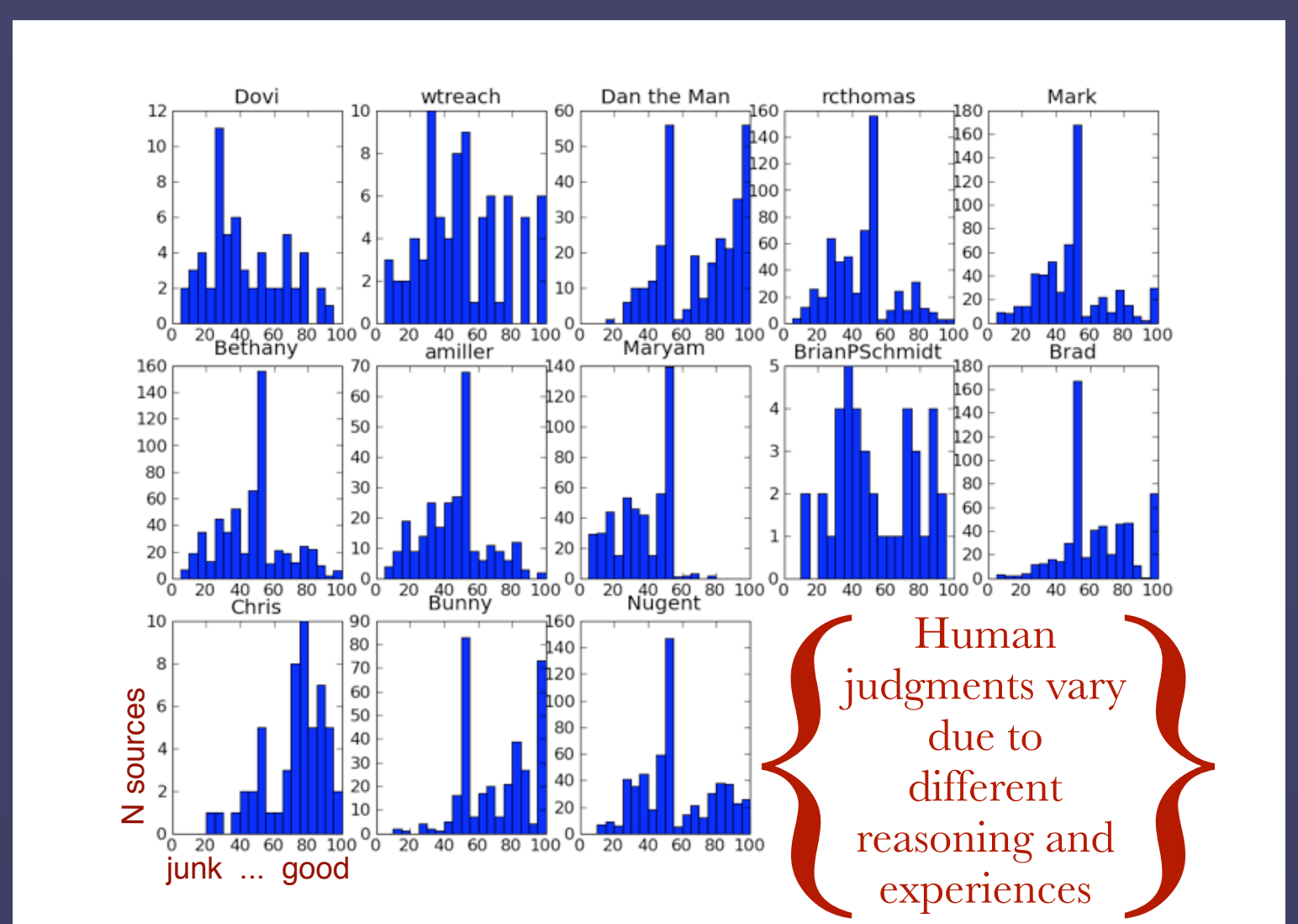
To train the TCP's science classifiers, the TCP references ~20,000 lightcurves which are contained in the (http://dotastro.org) lightcurve warehouse. DotAstro.org contains 150 variable and transient classes which were derived from 87 papers.

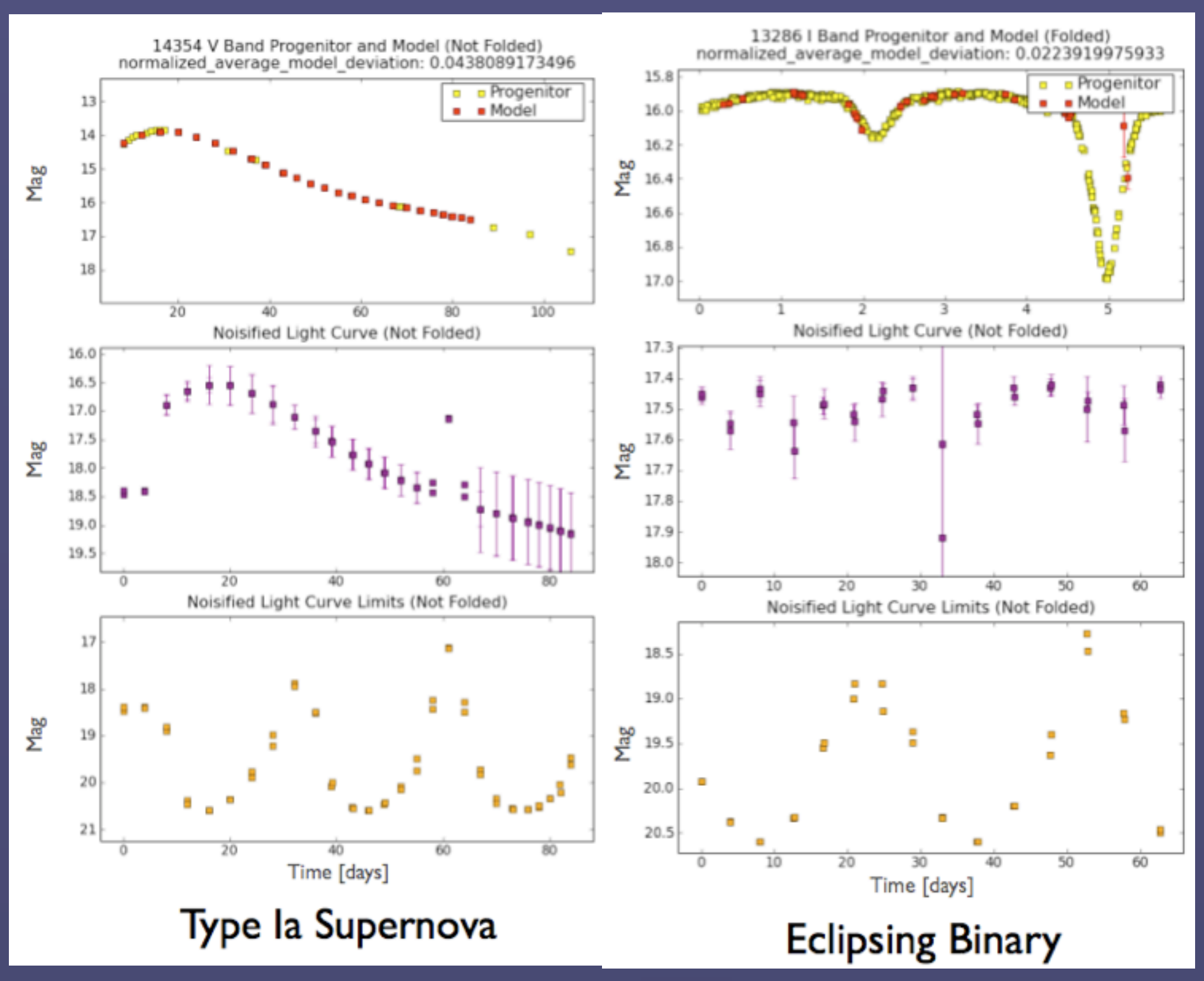



Group-Think
Open Crowd Sourcing Platform for Research
Buy a Group-Think Mug
Ref
New
Sub
More info about this image...
Real or Bogus. Tell us if this is a bone fide astrophysical transient (could be an asteroid) or a subtraction artifact.
bogus
astrophysical
Skip? □ Yes.
2D image classification: Machine-Learning with human input
>1000:1 rejection of bogus candidates (prelim. cuts + machine learning)

The GroupThink web-app displays thumbnails of astrophysical as well as defective and poorly subtracted sources from the PTF pipeline and allows a dozen users to judge whether the sources are astrophysical or not.

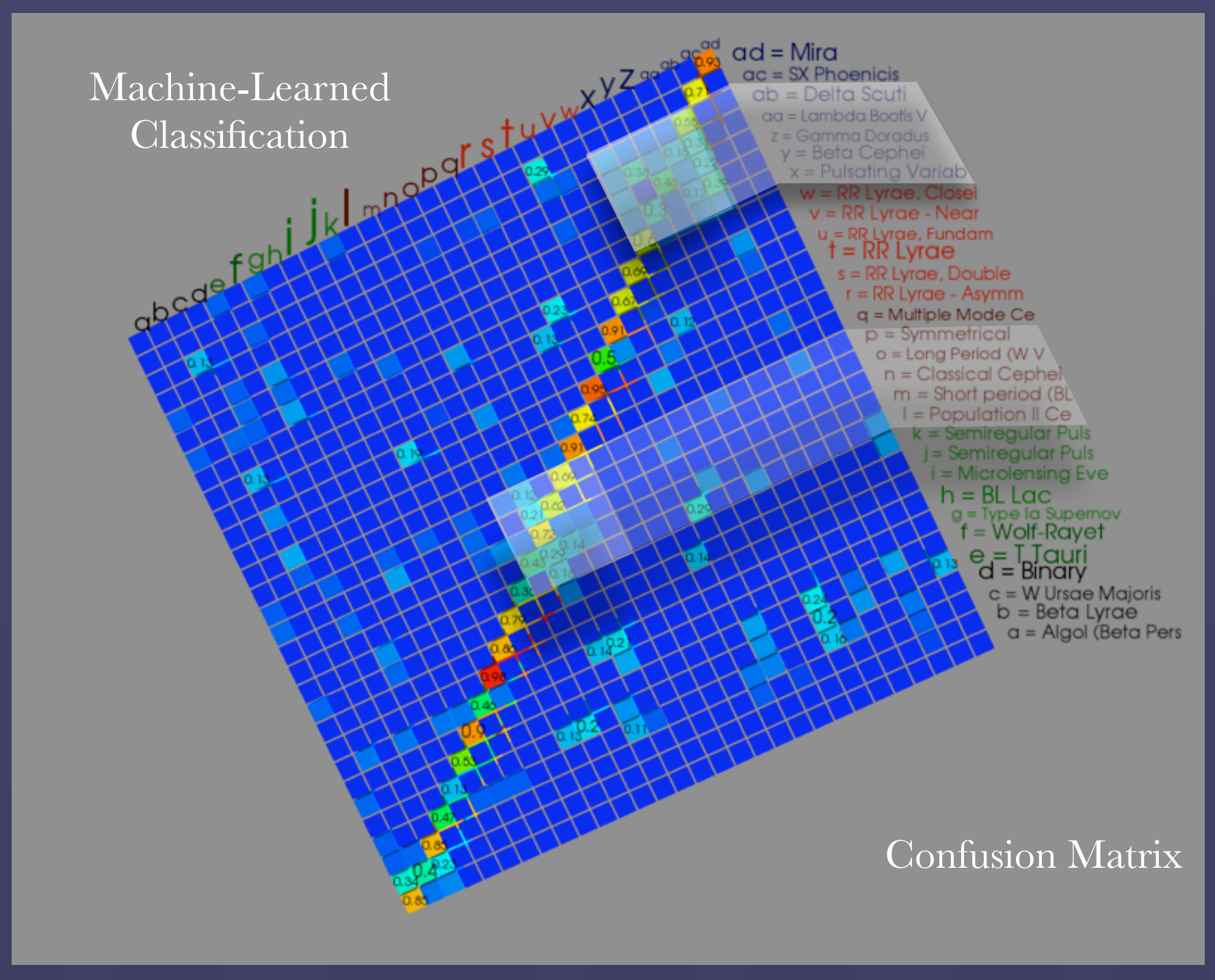
The resulting dataset from GroupThink is used to train a machine-learning based filter. This filter and it's associated metrics effectively model the user judgments, allowing the TCP to automatically filter out most non-astrophysical detections without the help of a human.





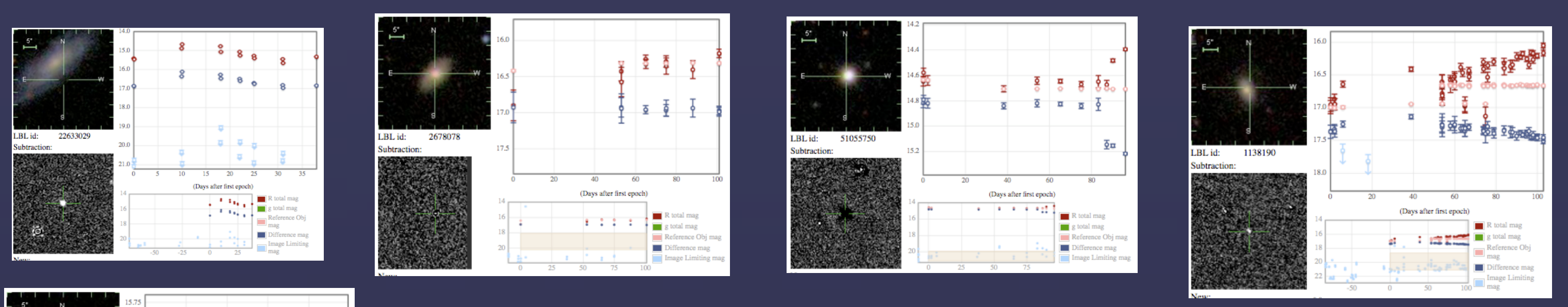
Before training the TCP's classifiers, we re-sample the well sampled lightcurves found in DotAstro.org to better match the PTF survey's observing cadence.

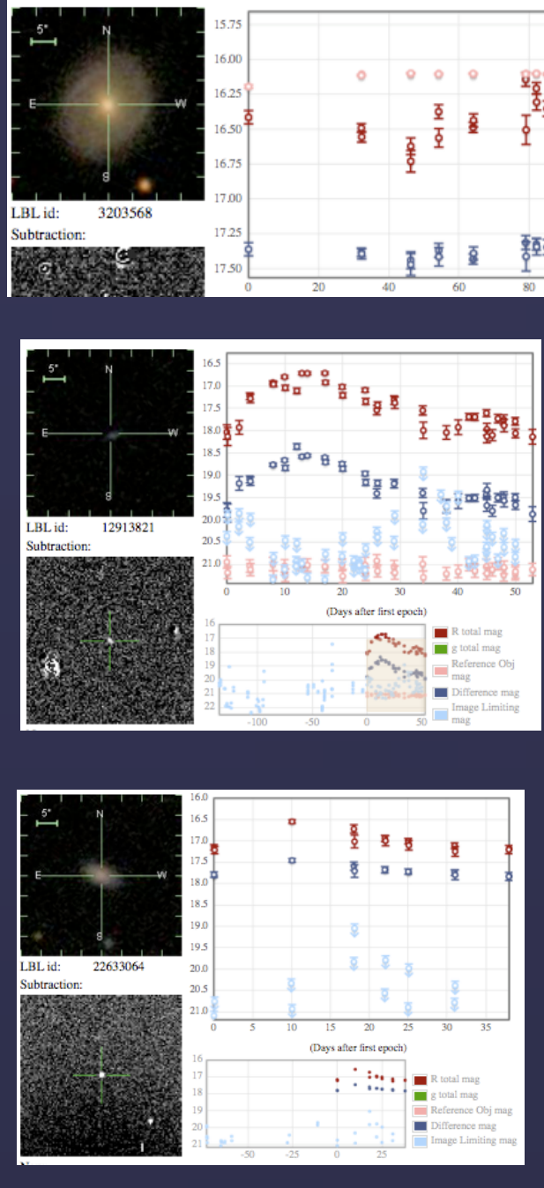
This "Noisification" software also adds simulated instrument and limiting magnitude offsets to better represent real PTF data.

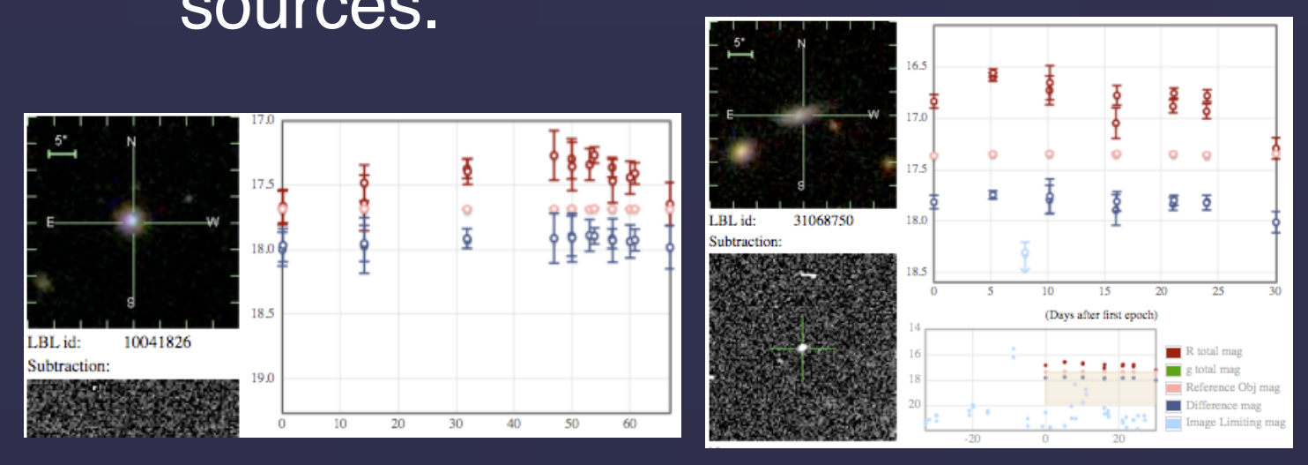


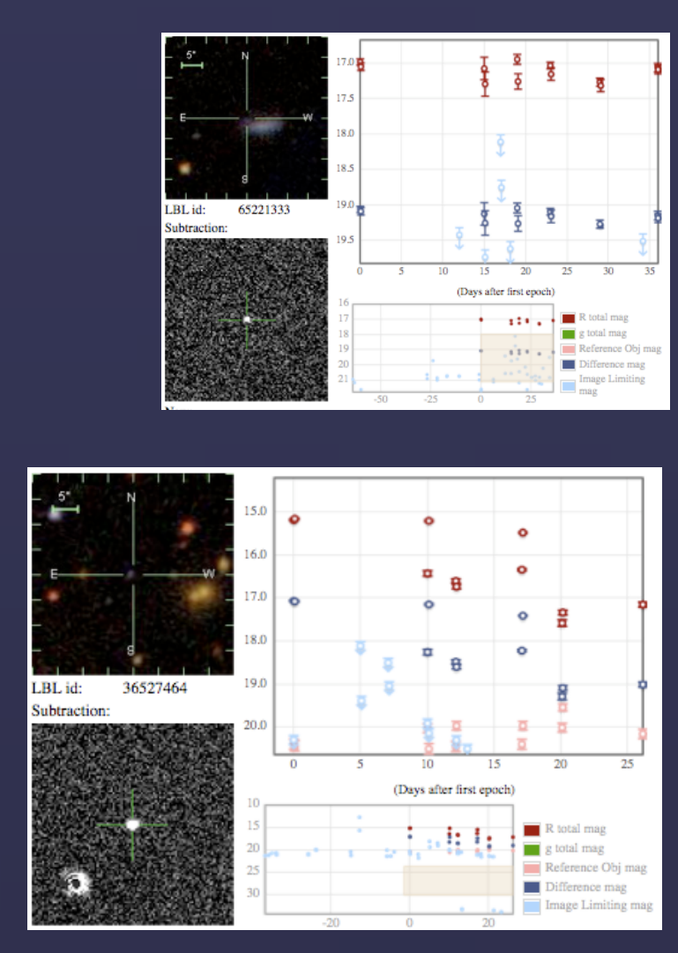
This confusion matrix shows how well a single machine-learned classifier distinguishes between 30 different science classes. A perfect classifier would contain 100% along it's diagonal.

Here, the classifier's effectiveness was evaluated using cross-validation with the training-set data. This means the heat-map colors represent successful classification percentages when given ideal lightcurves. Even when training and testing with the same cross-validated data, we can gain insight on how effective this classifier is with a set of lightcurve attributes.









These are a subset of PTF sources identified and classified by the TCP as potential SN/AGN during a test run. This preliminary test identified 84% of the already known and spectroscopically confirmed Super Novae as an SN/AGN science class. Mis-classifications were primarily faint and nearly-ambiguous sources.

Contacts:

Josh Bloom (PI)
jbloom@astro.berkeley.edu
Dan Starr (primary developer)
dstarr@astro.berkeley.edu

Dan Starr

Links:

DotAstro.org	http://dotastro.org
(TCP links)	http://is.gd/5mA4 http://tinyurl.com/tcp123

EC2 amazon web services SDSC SAN DIEGO SUPERCOMPUTER CENTER NSF LCOGT.net THE UNIVERSITY OF CALIFORNIA 1868