ASTRONOMICAL DATA TRIAGE FOR RAPID SCIENCE RETURN

Brian D. Bue, Kiri L. Wagstaff, Umaa D. Rebbapragada, David R. Thompson and Benyang Tang*

Jet Propulsion Laboratory, California Institute of Technology

ABSTRACT

The quantity of astronomical observations collected by today's instruments far exceeds the capability of manual inspection by domain experts. Rather than relying on human eyes to examine and analyze all collected data, we employ data triage algorithms shortly after data collection. Automated data triage enables increased science return by prioritizing interesting or anomalous observations for follow-up inspection, while also expediting analysis by filtering out noisy or redundant observations. We describe three specific astronomical investigations that are currently benefiting from upstream data triage techniques in their respective processing pipelines.

Index Terms— data analysis, data triage, astronomy, transient detection, asteroid detection

1. INTRODUCTION

Today's scientific instruments can collect data at increasingly higher resolutions (in temporal, spatial, and spectral dimensions). The increase in resolution leads directly to an increase in data volumes. For scientific campaigns in which rare, unusual, but extremely valuable phenomena are present, the ability to sift quickly through large amounts of data to find observations of interest is vital.

In astronomical applications, the primary bottleneck limiting the amount of data that can be analyzed is the amount of human reviewer time available for examining observations. Automated analysis methods such as classifiers, clustering, and statistical anomaly detection algorithms can assist by triaging data where it is collected and prioritizing data before demanding the attention of a human expert. Rather than relying on human eyes to examine and analyze all collected data, automated data triage allows domain experts to focus their attention on interesting or anomalous observations and reduces the time and effort necessary to manually filter out false detections.

We have applied this strategy to a variety of applications in which the goal is to detect rare but scientifically valuable phenomena. In this abstract, we describe three specific astronomical investigations that have benefited from automated data triage systems. In each case, our automated data triage systems have already been integrated into the relevant data processing and analysis pipelines.

2. OPTICAL TRANSIENT EVENT DETECTION

The intermediate Palomar Transient Factory (iPTF) is a fully automated synoptic sky survey operating since 2013 for the purpose of detecting optical transient events such as supernovae, variable stars, and asteroids [1]. iPTF uses the 48-inch Samuel Oschin telescope to image large swaths of sky at a rapid cadence for transient detection, and the 60-inch Palomar telescope for multi-color follow-up of detected candidates. By subtracting nightly-captured images from reference images of the sky, iPTF generates large quantities of candidate transient sources (≈500K-1M nightly). iPTF succeeds the original Palomar Transient Factory (PTF) that was in operation from 2009 through 2012 [2], and differs only in science goals and data processing. Figure 1 is an example of iPTF images of a nova from the M31 galaxy. On the left is the science image, its associated reference (center), and the differenced image (right) that results from subtracting the reference from the science image.

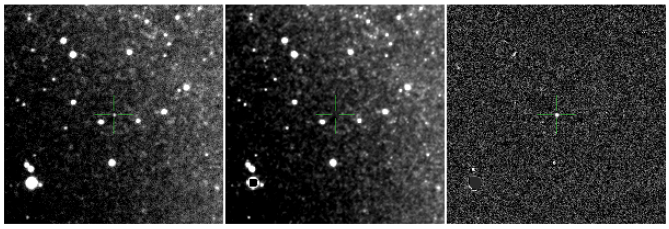

Fig. 1. Examples of a science (left), reference (center) and differenced (left) images from the intermediate Palomar Transient Factory (iPTF). This is a point source detection of a nova in the M31 galaxy that was scored highly by RB4.

Unfortunately, the image differencing process introduces an overwhelming number of "bogus" candidates for every candidate source representing a real astronomical real. These are typically caused by instrument or image processing artifacts. Automated data triage is essential for rapid and effective science return from sky surveys such as iPTF. Human scanners cannot vet the volume of nightly detections that may number into the hundreds of thousands. Instead, iPTF relies on an automated RealBogus system to serve as an initial filtering step and prioritize up to 200 high-quality candidates per night that humans will vet for follow-up analysis at iPTF Consortium follow up assets. It is important that human scanners not waste time vetting detections that are truly bogus in order to effectively use the highly-constrained time at Consortium telescopes. In other words, it is critical for the RealBogus systems to have a low false positive rate, and assign high scores to true astronomical transients.

2.1. Detecting Transients and Variable Stars

The original PTF triage system, developed by Bloom et al. [3], and its later upgrade, "RB2," deployed by Brink et al. [4], have both demonstrated real-time discovery of transients and variable stars. However, upgrades to the image subtraction pipeline for iPTF altered the features extracted from newly-captured imagery, in contrast to the PTF-based features used to train the RB2 classifier. We demonstrate this difference in Figure 3, which shows the correlation between the distributions of 31 features for PTF vs. iPTF real transients. Training a classifier using the PTF features exhibiting small or negative correlation coefficients will potentially degrade classification accuracy, as such features are not informative for the iPTF imagery.

*Corresponding author. 4800 Oak Grove Dr., Pasadena CA, 91001.
E-mail address: bbue@jpl.nasa.gov (B. Bue).

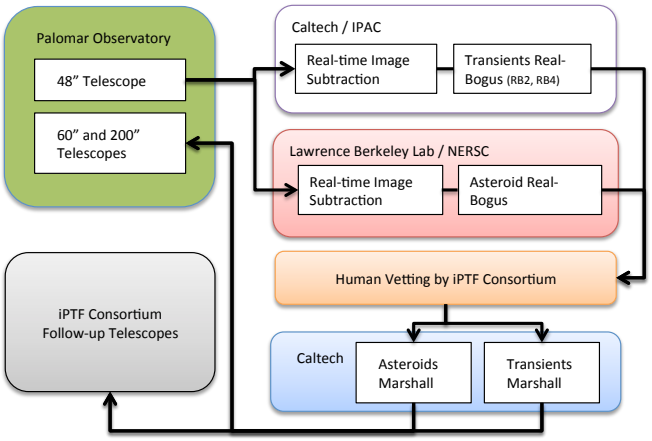

Fig. 2. Data flow (simplified and updated from original diagram published by Law et al. [2]) from the 48" telescope at Palomar Observatory to the real-time image differencing pipelines at both Lawrence Berkeley Laboratory (LBL) and Caltech. Two different "RealBogus" systems are in operation, one at LBL that is trained to look for point-source astronomical transients, specifically those in extragalactic fields. The second focuses on vetting "streaks" found on differenced images as either real near-Earth object (NEO) detections, or bogus objects such as image artifacts, cosmic rays or fast-moving satellites.

In July 2014, we deployed a new RealBogus classifier, "RB4," designed specifically for iPTF imagery. We constructed a new training set consisting of 18.5K spectroscopically confirmed iPTF transients, along with an equal number of bogus candidates. Following the methodology of Brink et al., we selected bogus candidates via uniform random sampling from the set of unconfirmed iPTF candidates. Using our new training set, we trained a random-forest classifier to distinguish between real and bogus iPTF candidates. Figure 4 compares the predicted scores produced by the RB2 classifier in comparison to our RB4 classifier scores on an independent test set. Scores near zero indicate bogus predictions, and scores near one indicate real predictions. As Figure 4 shows, the PTF-trained RB2 classifier produces scores that are more ambiguous in comparison to the iPTF-trained RB4 predictions that are more skewed toward a score of one. We also discovered cases where RB2 completely missed an interesting transient. An M31 nova (pictured in Figure 1) was missed by RB2 in July 2014. RB2 returned an average score of 0.0925 over sixteen candidate detections. RB4 predicted an average score of 0.8775 on those detections, and would have caught the nova had it been in operation (RB4 was deployed at iPTF on August 1, 2014).

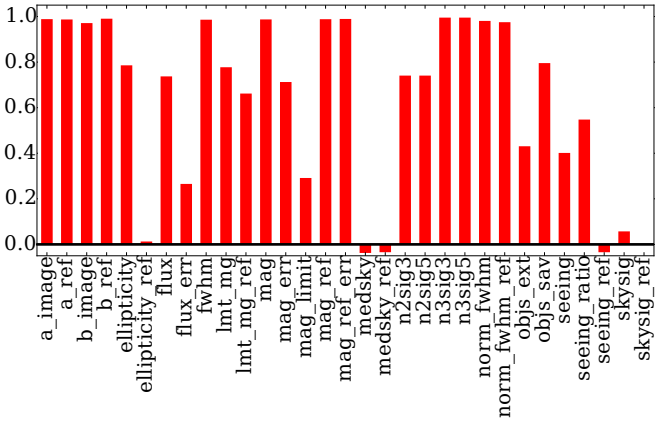

Fig. 3. Correlation between distributions of PTF and iPTF features for real transient sources.

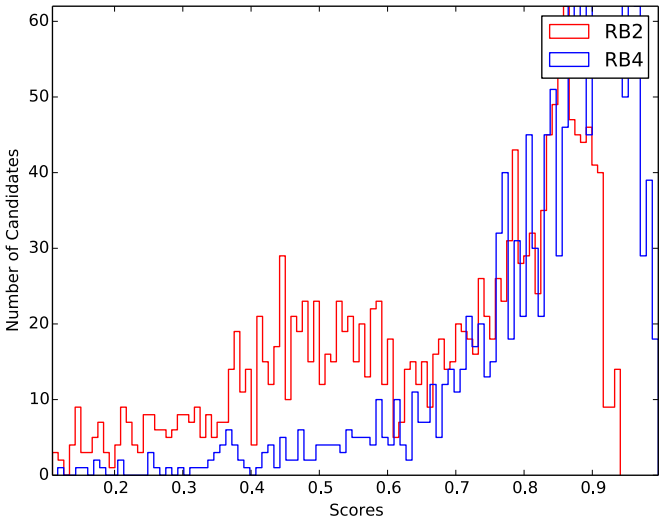

Fig. 4. Distribution of RB2 (red) vs. RB4 (blue) scores on an independent test set of real, spectroscopically-confirmed iPTF sources. The iPTF-trained RB4 classifier produces fewer ambiguous predictions than the PTF-trained RB2 classifier, which scores large numbers of these real detections below 0.7.

2.2. Near-Earth Asteroid Detection

We also developed a RealBogus classifier to detect streaking Near-Earth Objects (NEOs). This project focuses on asteroids ranging from 3 to 300 meters in diameter passing between 0.1 and 10 lunar distances from Earth, typically moving between 10 and 100 arcsec/minute.[1] Such NEOs are a poorly understood population of solar system bodies. They are of scientific interest for the hazards they pose, and resources they may contain. They are also key destinations in NASA's future plans for human spaceflight [5].

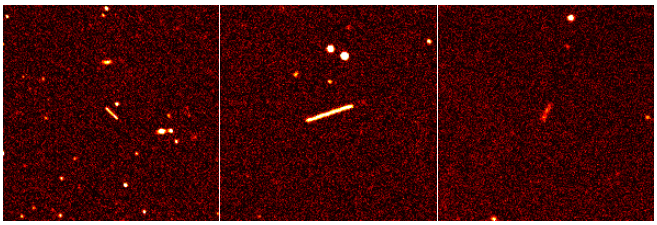

Fig. 5. Example streaking NEO candidates detected with the streak detector (figures courtesy Adam Waszczak, Caltech).

As with the iPTF transient detection pipeline, the number of detected streak candidates far exceeds manual inspection capabilities, and many are bogus candidates representing imaging artifacts, radiation hits, or periodic changes in brightness caused by fast-moving satellites. However, unlike the iPTF pipeline, very few real candidates ($\approx$ 250) are available for training. To ameliorate this issue, (1)

[1] http://ptf.caltech.edu/marshals/asteroids/

supplement the training set with roughly 1500 synthetic streaks generated within the span of the real streaks' features; and (2) augment the feature space with morphological features for each candidate. We then train a random-forest classifier using the real and synthetic streaks, along with 20K randomly-selected bogus candidates. The streaks classifier was deployed in the real-time asteroid processing pipeline in April 2014, and it continues to detect several follow-up worthy candidates each week, including nine confirmed streaking NEOs at this time of this writing, while reducing the number of non-streak detections for manual inspection by a factor of 100. Figure 5 shows three streaking NEO candidates detected by the streak detector.

3. RADIO TRANSIENT EVENT DETECTION

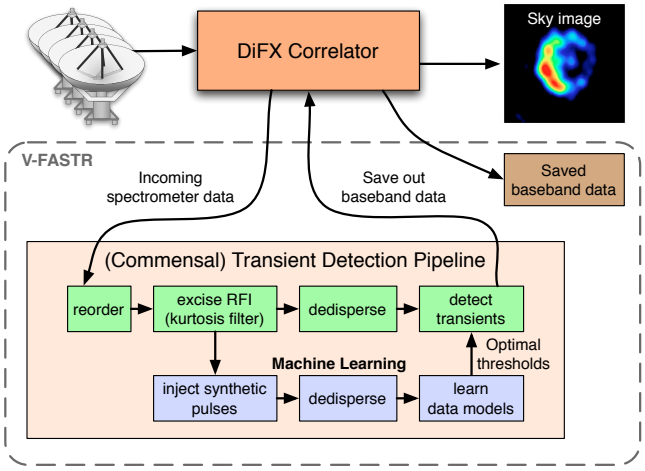

Fig. 6. V-FASTR system diagram for commensal radio transient detection at the VLBA. Data flows from the telescopes through the correlator and is simultaneously analyzed by a machine learning detection pipeline. Full baseband data is saved for interesting candidates; the rest is deleted to make room for new observations.

The Very Long Baseline Array (VLBA) is a distributed set of ten 25-m radio telescopes spread across 8600 km. The array is controlled by the VLBA Array Operations Center in Socorro, NM, where data collected by all of the telescopes is combined using a software correlator [6]. Because the correlator is implemented in software, instead of hardware like older telescope arrays, we are able to modify its operation without altering the physical hardware. Automated analysis is essential for this system given the volume of data collected.

3.1. The V-FASTR Transient Detection System

In 2010, an international collaboration of researchers from the Jet Propulsion Laboratory, Curtin University, ASTRON, and the National Radio Astronomy Observatory (NRAO), implemented V-FASTR, the VLBA Fast Transient detection system. V-FASTR employs machine learning methods to quickly analyze radio data collected at a 1-ms resolution across more than 100 channelized frequency bands [7, 8]. The V-FASTR system architecture is shown in Figure 6. Data from multiple telescopes is processed by the DiFX software correlator to generate spectrometer data that reports signal intensity at a range of frequencies for each time-step. The commensal pipeline first reorders the spectrometer packets by time-step, then uses a statistical (kurtosis-based) method to remove frequency bands that are corrupted by radio frequency interference (RFI). Each telescope's data is separately de-dispersed, a process by which the frequency-dependent delay induced by the interstellar medium is reversed. The signals from all telescopes are combined and candidate transient events are detected. Combining information from multiple telescopes increases the signal to noise ratio and reduces the number of spurious detections for local events visible only to a single telescope.

Candidate detections are archived along with derived features describing the event such as its signal strength and dispersion measure (DM), which indicates the amount of dispersion that was removed in the de-dispersion step. The DM provides information about how far the signal has traveled, enabling us to separate local versus astronomical events.

V-FASTR also employs a parallel pipeline that uses adaptive excision to selectively ignore signals from individual telescopes that may contain RFI that was not caught by the kurtosis filter. The system periodically injects pseudo-transients with known signal strengths into a parallel copy of the incoming data and determines which telescopes' data to include in the data combination and transient detection step by maximizing detection accuracy on these known transients. The number of telescopes that are masked by this process varies from zero to six; depending upon the current noise conditions. Such adaptive excision further improves the precision of the detections by preventing a large burst of RFI at a single station from dominating the combined sum across all stations.

V-FASTR is supported by a team of human reviewers that examine the candidate detections that are found; these amount to tens to hundreds of candidates per day. The reviewers browse the detections using a web portal that reports the latest detections and allows them to take action on individual detections. Reviewers can discard spurious detections, save interesting detections for further analysis, and add content-based tags to classify detections by type. To date, the system has accumulated more than 150,000 detections, of which 9143 have human-assigned tags.

3.2. Transient Classification in V-FASTR

V-FASTR uses a machine learning classifier to further sort detections by type in an automated fashion. The goal is to reduce the burden of human review time by allowing reviewers to focus on the most promising candidates first. Radio transients with an astrophysical origin are rare, and without additional filtering, reviewers may spend a lot of time rejecting subtle RFI signals that would ideally be classified as such in an automated fashion.

Each candidate is detected in a detection time window, along with two margin time windows before and after the detection window. Sixteen features are calculated from these three windows, reflecting various properties of the candidate. We trained a random forest classifier using 7130 labeled candidates and evaluated it on a separate collection of 2008 labeled candidates. The possible classes were "pulsar" and three types of spurious detections: "aligned RFI", "single antenna detection" (SAD), and "system state switch" (SSS). In this setting, very high reliability of the output is essential, since the classification may ultimately be used to discard detections without further review. Therefore, we only publish the top 10% most confident pulsar classifications and the top 20% of the other three output classes. This also allows the system to abstain from classifying novel detections that do not fit into these four classes, such as

Table 1. V-FASTR transient classification performance, showing the number of detections assigned to each class by the random forest classifier (columns) versus the class assigned by human reviewers (rows). Pulsars have no false or missed detections. The three spurious classes show minor confusion. Overall accuracy is 93.5%.

Predicted class:	Pulsar	Aligned RFI	SAD	SSS
Pulsar	128	0	0	0
Aligned RFI	0	39	5	6
Single Antenna Det.	0	1	47	1
System State Switch	0	3	2	51

true astrophysical transients. (We do not have enough examples of these rare events to train a classifier directly on them.)

Table 1 compares the classes predicted by the classifier (columns) to the true classes assigned by reviewers (rows) on the held-out test set. The classifier output predictions for only 260 of the 2008 items and abstained on the rest. The overall classification accuracy was 93.5%. The predictions are of very high reliability: there were no false or missed detections for the pulsar class, and only minor confusion among the three spurious classes. This means that if we discard the candidates predicted to belong to these three classes, we are unlikely to miss real pulsar events.

To date, V-FASTR has found thousands pulses from known pulsars and continues to operate commensally with all VLBA observing campaigns, including those with scientific goals other than transient detection [9]. The system continues to detect and classify new events on a daily basis. Reviewers also contribute new tags each day, and have the option of correcting any erroneous predictions inside the web portal. Consequently, we re-train the classifier every day using the latest set of tags and generate new output for all events in the archive. The V-FASTR system thereby has the ability to improve its performance on an ongoing basis using a steadily growing collection of data and tags.

4. CONCLUSIONS AND FUTURE WORK

The high data volumes generated by modern astronomical surveys demand automated data triage techniques to help astronomers identify scientifically interesting or anomalous observations. We have demonstrated that automated data triage with statistical machine learning techniques enables rapid science return in several optical and radio astronomy projects. In each case, the cooperative interaction between astronomers and machine learning practitioners has yielded mutual benefits, both in terms of discoveries and development of advanced detection techniques. Such interaction facilitates refinement of training data and features as discoveries are made, and has played a crucial role in the success of each project. Ongoing efforts include developing domain adaptation techniques to leverage measurements captured by different instruments, and evaluating deep learning methods to learn informative features directly from instrument data.

Acknowledgments: We thank Mansi Kasliwal and Yi Cao (Caltech) and Przemek Wozniak (LANL) for their contributions to the transient detection project;Adam Waszczak, Tom Prince, and Russ Laher (Caltech) for their contributions to the asteroid detection project; and Walter Brisken (NRAO), Adam Deller (ASTRON), Steven Tingay, Divya Palaniswamy, Randall Wayth (Curtin), and Sarah Burke-Spolaor (Caltech) for their contributions to the V-FASTR project. A portion of this research was performed at the Jet Propulsion Laboratory, California Institute of Technology. Copyright 2014. All Rights Reserved. US Government Support Acknowledged.

5. REFERENCES

[1] S R Kulkarni, "The intermediate Palomar Transient Factory (iPTF) begins," *The Astronomer's Telegram*, vol. 4807, pp. 1, 2013.

[2] Nicholas M. Law, Shrinivas R. Kulkarni, Richard G. Dekany, Eran O. Ofek, Robert M. Quimby, Peter E. Nugent, Jason Surace, Carl C. Grillmair, Joshua S. Bloom, Mansi M. Kasliwal, Lars Bildsten, Tim Brown, S. Bradley Cenko, David Ciardi, Ernest Croner, S. George Djorgovski, Julian van Eyken, Alexei V. Filippenko, Derek B. Fox, Avishay Gal-Yam, David Hale, Nouhad Hamam, George Helou, John Henning, D. Andrew Howell, Janet Jacobsen, Russ Laher, Sean Mattingly, Dan McKenna, Andrew Pickles, Dovi Poznanski, Gustavo Rahmer, Arne Rau, Wayne Rosing, Michael Shara, Roger Smith, Dan Starr, Mark Sullivan, Viswa Velur, Richard Walters, and Jeff Zolkower, "The palomar transient factory: System overview, performance, and first results," *Publications of the Astronomical Society of the Pacific*, vol. 121, no. 886, pp. pp. 1395–1408, 2009.

[3] J S Bloom, J W Richards, P E Nugent, R M Quimby, M M Kasliwal, D L Starr, D Poznanski, E O Ofek, S B Cenko, N R Butler, S R Kulkarni, A Gal-Yam, and N Law, "Automating discovery and classification of transients and variable stars in the synoptic survey era," *Publications of the Astronomical Society of the Pacific*, pp. 1175–1196, June 2011.

[4] Henrik Brink, Joseph W Richards, Dovi Poznanski, Joshua S Bloom, John Rice, Sahand Negahban, and Martin Wainwright, "Using Machine Learning for Discovery in Synoptic Survey Imaging," *Monthly Notes of the Royal Astronomical Society*, pp. 1–16, Sept. 2012.

[5] Martin Elvis, Jonathan McDowell, Jeffrey A Hoffman, and Richard P Binzel, "Ultra-low delta-v objects and the human exploration of asteroids," *Planetary and Space Science*, vol. 59, no. 13, pp. 1408–1412, Oct. 2011.

[6] Adam T Deller, S J Tingay, M Bailes, and C West, "DiFX: a software correlator for very long baseline interferometry using multiprocessor computing environments," *Publications of the Astronomical Society of the Pacific*, vol. 119, no. 853, pp. 318–336, 2007.

[7] David R Thompson, Kiri L Wagstaff, Walter F Brisken, Adam T Deller, Walid A Majid, Steven J Tingay, and Randall B Wayth, "Detection of fast radio transients with multiple stations: a case study using the very long baseline array," *The Astrophysical Journal*, vol. 735, no. 2, pp. 98, 2011.

[8] Randall B Wayth, Walter F Brisken, Adam T Deller, Walid A Majid, David R Thompson, Steven J Tingay, and Kiri L Wagstaff, "V-fastr: The vlba fast radio transients experiment," *The Astrophysical Journal*, vol. 735, no. 2, pp. 97, 2011.

[9] Randall B Wayth, Steven J Tingay, Adam T Deller, Walter F Brisken, David R Thompson, Kiri L Wagstaff, and Walid A Majid, "Limits on the Event Rates of Fast Radio Transients from the V-FASTR Experiment," *The Astrophysical Journal*, vol. 753, no. 2, pp. L36, June 2012.