New Horizons in Time-Domain Astronomy
Proceedings IAU Symposium No. 285, 2011
R.E.M. Griffin, R.J. Hanisch & R. Seaman, eds.

doi:10.1017/S1743921312000531

Technical and Observational Challenges for Future Time-Domain Surveys

Joshua S. Bloom
University of California, Berkeley, CA 94720, USA
email: jbloom@astro.berkeley.edu
Invited Talk

Abstract. By the end of the last decade, robotic telescopes were established as effective alternatives to the traditional role of astronomer in planning, conducting and reducing time-domain observations. By the end of this decade, machines will play a much more central role in the discovery and classification of time-domain events observed by such robots. While this abstraction of humans away from the real-time loop (and the nightly slog of the nominal scientific process) is inevitable, just how we will get there as a community is uncertain. I discuss the importance of machine learning in astronomy today, and project where we might consider heading in the future. I will also touch on the role of people and organisations in shaping and maximising the scientific returns of the coming data deluge.

Keywords. methods: data analysis—methods: statistical—techniques: photometric—surveys—stars: variables—sociology of astronomy—history and philosophy of astronomy

1. Introduction

Though the scientific interests in this conference were bewilderingly diverse, we were brought to Oxford with a common interest in understanding objects and events that *change* with time. To be sure, *all* astrophysical entities change with time—in brightness, in position on the sky, in physical size, in colour—but our perceptions of such changes have only just begun to broaden, as modern instrumentation techniques have matured. Indeed, we learned of the ambitions in the stellar variability community in studies at the micro-magnitude level (enabled so elegantly by Kepler and CoRoT), and at time-scales ranging from seconds to centuries. Explorations of (intrinsically) faint and fast (<week) transients at optical wavebands (enabled by projects like PTF) have revealed new classes of events (e.g. Kasliwal 2011). Likewise, the radio community has been in hot pursuit of new classes of variability at sub-second time-scales.

One of most exciting endeavours for time-domain astronomers (and, frankly, for most of the maverick-minded) is the discovery of the unknown. However, given the rising complexity and expense of new surveys, blind exploration as a singular goal is a dangerous impetus. As Tom Prince claimed over a particularly engrossing dinner, "If it's not worth doing, it's not worth doing well." In other words, the significant technical and observational challenges of pulling off a successful synoptic survey are only worth tackling when the science is compelling. That said, it is indeed tempting to look at the Kasliwal or Cordes phase-space plot in time-scale and peak luminosity and wonder what else might be lurking in the white-space. But there is, I claim, no serendipity without bread and butter science. And it is precisely the bread and butter science that provides the important technical challenges which, when met, will enable new (unexpected) discoveries.

Aside from presenting *Time* as the the unifying thematic approach, we were also struck by the similarity of challenges imposed by the sheer volume of the data now collected

(and to be collected by the behemoth synoptic surveys envisaged over the next decade). The proverbial "data deluge" has indeed inundated the astronomical community. It is this enormity of high-quality digital data that forces us to address a sea-change in the way we conduct ourselves as scientists going forward. Yes, there are interesting technical hurdles in the acquisition, movement, management and access of those data, but the real paradigm change comes in the need for fundamentally different approaches to discovery and inference regarding those data. The eight-hundred million light curves updated almost daily by LSST cannot all be scrutinised by astronomers, nor by their spectrographs. Yet some small fraction, perhaps just a few rapidly-evolving events per night, will need the full involvement of the world's largest and most precious telescopes to extract the most science. In this context, the crucial question for time-domain science is this: "How do we do discovery, follow-up and inference when the data rates (and requisite time-scales) necessarily preclude human involvement?"

2. The Autonomous Data-Driven Workflow

In this modern data-driven workflow, where people are abstracted from the real-time loop, I distinguish the acts of *finding*, *discovery* and *classification* of astrophysical events. "Finding" might be considered the process of the extraction of candidate events from raw data into a more abstract (and compact) form. In the optical domain, an example would involve the reduction and subtraction of two frames, the identification of significant changes in the subtraction image, and the recording of metadata about each candidate into a database. "Discovery" would be the recognition that a candidate is indeed of some astrophysically varying source (and not an artefact) that might be of interest for further scrutiny. "Classification" would be the act of understanding and quantifying what that event is likely to resemble among the classes of known (and hypothesised) events. In that sense, Galileo was the first to *find* Neptune (Kowal & Drake 1980). But since he recorded it in his notebook as an uninteresting source (at least as compared to Jupiter)†, he is credited with neither its discovery nor classification. Just think how famous Galileo could have been!

3. Discovery

Each wavelength domain (and spectrum, for that matter) presents its own set of difficulties when trying to automate discovery. High-energy missions like Fermi are working at the Poisson noise limit, gravity-wave astronomers are working in the low signal-to-noise regime where template matching algorithms increase the number of effective statistical trials, and radio surveys contend with complex radio frequency interference (RFI) that can mimic the signal of interest. My group has focused its effort in automating discovery at optical wavebands, on candidates found in image differences. While image differencing is an improvement over catalogue-based discovery (such as minimising bias against discovery in crowded fields or around galaxies), the number of spurious ("bogus") candidates vastly outnumbers the *bona fide* ("real") astrophysical sources. In the Palomar Transient Factory (PTF) there about 1000 bogus for each single real candidate. We have developed a framework that uses a training set—based either on the aggregation of expert opinion (Bloom *et al.* 2011) or on retrospective samples with ground truth (e.g. from spectroscopic confirmation)—that allows us to identify rapidly and automatically the most promising real candidates during real-time runs of PTF. We use the labels to train a random forest (RF) classifier that allows us to identify astrophysical candidates reasonably (Bloom

† He did, however, note its apparent motion.

https://doi.org/10.1017/S1743921312000531 Published online by Cambridge University Press

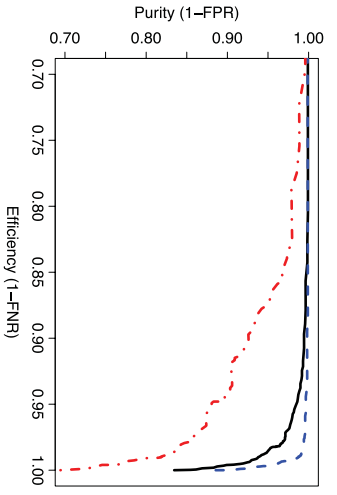

Figure 1. Machine-learned on-the-fly purity vs. efficiency in determining whether a new source from the PTF is a transient (i.e. an explosive event) as opposed to a variable star. The dashed (dashed-dotted) line shows this trade-off for sources within (outside) the SDSS footprint. The solid line shows the aggregate results for all PTF sources in the training sample.

et al. 2011). Just one month before this Symposium our real/bogus framework promoted a candidate event in M101 to the top of the scanning list of "local-universe" events, leading to the discovery of the very nearby Type Ia supernova SN 2011fe (Nugent *et al.* 2011). Future surveys, dealing with many more candidates per night than the 1.5 million found in PTF, will need to employ a similar approach to discovery.

4. Classification

Once a source makes it through the discovery hoop of being "astrophysical†", follow-up decisions must be made. Generating a probabilistic classification based upon all (but still limited) available data is used to inform follow-up decisions. One generic approach to classification is to map the available data, however heterogenous, into an m-dimensional real-number set; this is called "feature space", and once the data are coerced into this space we can apply existing machine-learned frameworks for classifying new sources. There are modern techniques (e.g. Stekhoven & Bühlmann 2011) for imputing missing values into feature space by predicting what those values would have been if they were available and not censored.

4.1. *On the Fly*

In the limit of only little data, context (the location of an event on the sky and what other sources it is physically associated with) becomes an important discriminator of class. Almost all of the initial classification in PTF, built upon another machine-learned trained set, relies on contextual information coerced into a set of well-defined features. Some of those data are available locally, within the same databases as those that house the candidates, but many of the data are distributed throughout the Web. The quality and information content in those ancillary datasets is critical for making accurate classification statements. This is illustrated in Fig. 1, where the lack of a rich set of SDSS information clearly diminishes the classification accuracy of new events.

† In PTF, so as to avoid asteroid discovery, we actually require two astrometrically coincident candidates to be of high real/bogus value as the criteria for discovery.

https://doi.org/10.1017/S1743921312000531 Published online by Cambridge University Press

Since we rely heavily on Simbad, USNO-B1.0, NED and SDSS for contextual data, the main bottleneck in making rapid classification statements about a new event is in the speed and reliability of foreign Web-services. One clear lesson for future surveys is that having physically co-located and unrestricted access to information from other surveys and at other wavebands is crucial. That statement is somewhat contrary to the current push to standardised access protocols for remotely-managed data.

4.2. *In Retrospect*

As a survey continues to accumulate data on the same varying source, eventually the time-series data outweigh the importance placed on context data. For example, a source near the outskirts of a galaxy might rightfully be called a supernova or nova based on context, but if the source is found to be varying periodically with a period around 0.5 days with a certain peak-to-peak amplitude then we will eventually gain confidence in classifying the source as an RR Lyrae star. Much of the body of work on time-domain classification has focused on a *retrospective classification* of completed or nearly completed time-domain surveys (see Bloom & Richards 2011 for a review). On a set of thousands of HIPPARCOS and OGLE variable stars across 26 different classes, for instance, we found a classification error rate of about 22% using 53 periodic and aperiodic features (Richards *et al.* 2011b). Adding more features and dealing better with missing data reduces the error rate to ∼15%. It is interesting that, when the taxonomy of variable stars is accounted for, the overall mis-classification rate across three broad classes (pulsating, eruptive and multi-star) plummets to about 5%. In other words, the machine-learned frameworks are better at distinguishing between grossly different physical processes even if the feature set was not specifically encoded to capture those physical processes.

The generation of features can be a non-negligible expense both in time (for Web-based features) and computationally (e.g. for periodogram analysis). However, what are even more costly are traditional fitting routines, such as those for eclipsing binaries, microlensing events and supernovæ. In that context, my suspicion is that the best retrospective classifiers will use a hybrid of generic classification tools on computed features and science-specific fitting routines on sources that are likely to belong to certain families of sources. Just how—from a perspective of classification, accuracy and expense—to architect optimally such a system is an open question.

4.3. *On People vs. Machine*

Despite (or perhaps because of) the enormity of the data, there is considerable interest in using coordinated, collaborative public input to aid classification. Such crowdsourcing† has been carried out in earnest with static-sky images, and is now being used in some time-domain applications (e.g. Smith *et al.* 2011). Though the outreach aspect of that effort may be incredibly important, it remains to be seen whether truly novel science will flow from astronomy crowdsourcing at a sustained level. One worry on the time-domain front is that, unlike software and hardware, expert opinions do not scale easily. To get increasingly refined classification statements on more and more data, for example, the number of those capable and willing to give opinions cannot keep pace with the data growth. Another concern I have is that, unlike algorithms, people do not behave in deterministic ways; I can re-run my code on all previous data and get back the same

† A note for future generations: "crowdsourcing," a blend of the words crowd and outsourcing, is a vernacular term used to describe the act of public collaboration in a project to create and annotate content. As of late 2011, the term had not yet been added to the Oxford-English Dictionary.

https://doi.org/10.1017/S1743921312000531 Published online by Cambridge University Press

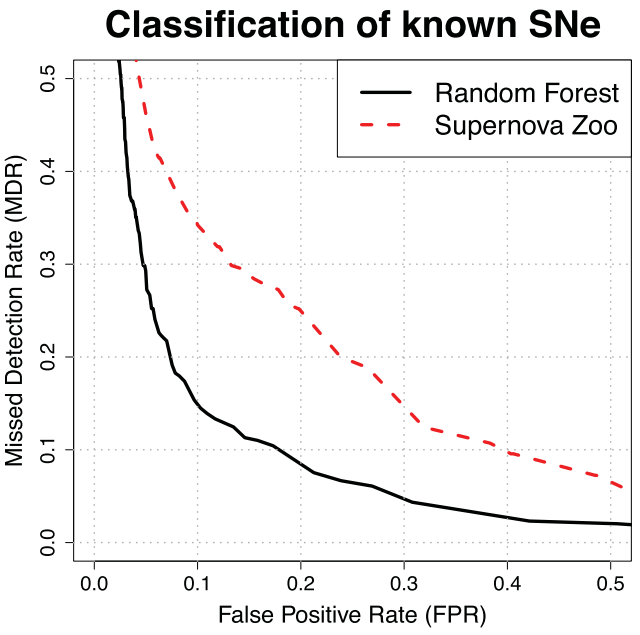

Figure 2. Supernova Zoo versus random-forest supernova discovery. We trained a classifier based on human labelling of PTF sources to predict which sources would turn out to be spectroscopically-typed SNe. By changing the SN discovery threshold from the random forest model, we trade a lower missed detection rate (MDF) for a higher false-positive rate (FPR). For a sample of 345 known PTF SNe, employing the RF score to select objects is *uniformly better*, in terms of MDF and FPR, than using the SN Zoo score. At 10% FPR, the RF criterion (threshold = 0.035) attains a 14.5% MDR compared to 34.2% MDR for SN Zoo (threshold = 0.4). From Richards *et al.* (2012).

result. That lack of determinism also means that the science-sacred concept of repeating and evolving experiments, notwithstanding the large "human cost", is lost in crowdsourcing‡.

The John Henrys of this world have learned that we should never do with people what can be done as well (or nearly as well for a fraction of the cost) with machines. That is not to say that crowdsourcing might not have its useful role in time-domain science, but that great care must be placed on experimental design both from the perspective of "human subjects" and in the uniqueness of the expected result. One crucial role that humans will continue to play in enabling the automated real-time loop is that of *label experts*. Just as we did in the real/bogus exercise, trained humans can label a small subset of the data so that supervised learning algorithms can be performed on the data and applied in real time. At this Symposium we presented the results of machine-learned discovery/classification of PTF sources that are likely to be supernovæ. We used the Supernova Zoo (Smith *et al.* 2011) mark-up of tens of thousands of candidate events (collected through a specialised DB query at the end of each PTF night) to construct a classifier capable of discovering supernovæ efficiently. The results of that exercise are depicted in Fig. 2 showing that the classifier outperforms the human labelling.

‡ Peng (2011) provides a recent and useful discourse on repeatability in modern science.

https://doi.org/10.1017/S1743921312000531 Published online by Cambridge University Press

5. Parting Thoughts

As the automated discovery and classification business matures, a number of interesting lines of inquiry will need to be studied. For example:

- How do we bootstrap the machine-learning process from one time-domain survey to the next, given the inherent differences in the ways that those surveys are conducted? Active learning (using expert opinions at optimised moments in the learning process) looks promising (e.g. Richards *et al.* 2011a).
- How do we detect and quantify real time-domain outliers—novel events that are not part of the established taxonomy? Clustering and semi-supervised learning seems an appropriate start (Protopapas *et al.* 2006; Rebbapragada *et al.* 2009). My view is that the way we get good at finding needles in a haystack is by getting really good at identifying hay.
- How can we imbue domain knowledge (and physics) into the learning process without having to use traditional domain-specific fitting routines?

No discovery or classifier engine will ever be perfect, in a sense of making statements precisely about the underlying origin of the observed variability. The best we can hope for are well-calibrated probabilities that can allow us to make informed decisions about moving to the next stage of the scientific process (see, e.g., Morgan *et al.* 2012). Viewed that way, classification is a maximisation tool: science in the time-domain will increasingly be conducted with major resource limitations—in the computational power available for discovery and classification, in follow-up telescope availability, in people's time and, ultimately, in capital cost.

Acknowledgements

I am grateful to the organisers and participants for the many enlightening discussions during the conference. I acknowledge support from a CDI grant (#0941742) from the National Science Foundation.

References

Bloom, J. S. & Richards, J. W. 2011, arxiv/1104.3142
Bloom, J. S., *et al.* 2011, ArXiv e-prints
Kasliwal, M. M. 2011, PhD thesis, California Institute of Technology
Kowal, C. T. & Drake, S. 1980, *Nature*, 287, 311
Morgan, A., *et al.* 2012, *PASP*, in press
Nugent, P.E., *et al.* 2011, *Nature*, in press
Peng, R. D. 2011, *Science*, 334, 1226
Protopapas, P., *et al.* 2006, *MNRAS*, 369, 677
Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. 2009, in: D.A. Bohlender, D. Durand and P. Dowler (eds.), Astronomical Data Analysis Software and Systems XVIII, ASP Conf. Ser. Vol. 411 (San Francisco: ASP), p. 264
Richards, J., *et al.* 2012, in prep
Richards, J. W., *et al.* 2011a, arxiv/1106.2832
Richards, J. W. 2011b, *ApJ* , 733, 10
Smith, A.M., *et al.* 2011, *MNRAS*, 412, 1309
Stekhoven, D. J.. & Bühlmann, P. 2011, arxiv/1105.0828