

Handling the Event Deluge

Roy Williams
California Institute of Technology

We consider how event handling will change as the discovery rates scale up in the future. We summarize here the different types of data that can help to resolve the nature of an astronomical transient, direct its dissemination, and prioritize the follow-up response.

The Coming Flood

More than ten years ago, the gamma-ray burst (GRB) community ignited the excitement over transient astronomical events. GRBs were an enigma until ultra-fast event dissemination allowed optical identification of afterglows, leading to rich data and rich science. The events back then were both valuable and infrequent: every new GRB could make a career for a young astronomer, and they were only detected every few days. However, in the next few years, surveys [1] such as LOFAR, Pan-STARRS, LSST and SKA will produce a flood of hundreds of events every 24 hours, with the scientific jewels surrounded by dross. By analogy, a hundred years ago, the daily newspaper provided a low-rate, high-value summary of the world, but now there is such a flood of “news” information that we must be very selective in what we read. In this article we suggest some ways to build this selectivity for astronomical transients, so that machines can select and deliver the most valuable events to astronomers.

Of course different people ascribe value differently. An astronomer may be interested in a particular type of variable source, or a particular set of sources; or in finding comets with a backyard telescope; or in finding unknown examples of a known type of source. All of us on this planet are interested in finding devastating objects that could collide with the Earth. Thus we expect the infrastructure for astronomical transients to allow end-users to select in advance what types of events they want, and we expect specialist “event brokers” that disseminate only certain kinds of transients to their target audiences. Such dissemination should happen quickly, so that the essential telescope resources can be directed for follow-up, and we expect these decisions to be taken automatically by machines. We also expect multiple repositories of past events, with sophisticated query interfaces, to enable statistical and correlation studies of large numbers of events.

The ideas here apply not just to astronomy, but also in other areas of society where events must be followed by rapid, often expensive actions. For example, earthquake early warning systems could shut down elevators, computers, etc, in the seconds between detection and the time the earthquake waves need to propagate. A Geiger counter at a homeland security installation scans cargo, and issues events, whose importance is compared with warnings from other sources, and may trigger a manual inspection. A power or data network must respond to reports from multiple sources to detect unusual events, and (occasionally) bring in a human to evaluate. Everything said here about astronomical events can be also understood in a general sense, to refer to any event stream.

Event Selection

This selectivity will be predicated on there being enough information to make such selections. However, when an astronomical transient occurs, it may not be clear immediately if it was caused by exciting new astrophysics that should get rapid follow-up, or if it is an ordinary object that gets much less attention. The immediate data that describes the new transient may be no more than a single number, perhaps just a magnitude measurement. Of course, a transient survey is picking up differences between “now” and “the past” which may or may not be due to astrophysics: that measurement may actually be the product of an artifact or processing error, causing an event report that has zero interest.

Follow-up observation will be in short supply in the era of the event deluge. Faint objects can only be observed with the largest telescopes, that are already over-subscribed, and objects with uncertain position require deep wide-field imaging to look for possible counterparts. But selection of “interesting” events is subject to the same quality measures of any selection process, being the false-positive and false-negative rates. False positives are uninteresting things that waste valuable follow-up resources, and false negatives are exciting objects that were not identified as such. It will be important to bring together all possible information, as quickly as possible, and to build new ways of automated information fusion, in order to minimize both false positives and false negatives.

Telescope Selection

It will also be important for telescope resources to be directed effectively. Traditionally, an observing proposal specifies exactly which objects are to be observed; alongside this will emerge a new kind of proposal that specifies a fruitful event stream and a real-time selection algorithm that will fill the observing schedule in near real-time. It is this algorithm that the committee will judge, rather than a set of target objects. Further, the committee should understand that false positives are part of any selection process, and should not punish the observer for past observations that turned out to be sub-optimal. In addition to selecting targets in real time, observers (or their automated surrogates) will optimize which telescope should be used for the follow-up. This will obviously depend on local conditions (horizon, cloud etc), but will also depend on what kind of observation yields the most useful information in classifying the transient. Work is in progress on implementing these ideas[2][3].

Some event data may be released immediately and publically, some may be held private. The advantage of private data is that scientific credit for discoveries can be closely controlled. But public release and openness generally produces a greater total scientific return, because of the wider variety of observations and data that accrue. Furthermore, the funding sources for astronomical surveys often require public data release.

Annotation

All this close selection implies that enough information is available to make a selection. It is analogous to a patient presented to a doctor with a bad headache: that information alone is not enough for diagnosis or treatment until further information is available, such as medical history or further tests. That first data item is not enough, so the doctor builds a portfolio of different types of information, which is used for judgment. The same happens with astronomical transients, each portfolio including information from archives, follow-up observations, comments, results from software, and other sources.

Thus annotation is the process of adding information to a first detection, in order to give a more complete picture of what is happening in the sky. Archives may be used to select observations of the source in the past, either from the same survey, or from published surveys, building up a panchromatic light curve of the object. Follow-up observation can occur, perhaps automatically, adding photometry and spectroscopy to the portfolio. People can examine the portfolio and add interpretations and evaluations. Software can run on the portfolio, generating new data that is put into the portfolio.

Heterogeneous and Multisource Data

One of the difficulties encountered in portfolio building is that the data is heterogeneous: for example the source may or may not be in the footprint of an archival dataset, or the number of points in the light curve may be variable, there may be zero or more existing classifications (eg. Galaxy, recurrent nova, etc), or there may be multiple human evaluations. Standard data mining algorithms expect each portfolio to have exactly the same format -- a “database record” or a “point in a feature space” -- so these do not work well in the context of heterogeneous portfolios. Prioritizing these portfolios is often like the choice offered in so many stories and movies: between the ‘well known but bland’ candidate, versus the ‘mysterious but exciting’ candidate (See Mahabal et. al. in Section 1.3).

Another difficulty arises because the data comes from multiple sources. Different authors of information may use different names for the same thing, different units, different ways to group or transform data, etc. This can make it very difficult to compare multi-sourced data and reach actionable conclusions. We can solve this by asking authors to define their metadata in advance of the arrival of the transient event, and in a structured fashion. Only if that metadata and structure can be kept simple enough, and yet also sufficiently expressive, then trusted, quantitative comparisons can be made.

When the data is both heterogeneous in content and in source, it becomes quite difficult to work with many of the software solutions that exist to handle event streams. The StreamSQL paradigm[4] uses extensions to the SQL language, designed for tables, but repurposed for real-time streams of records. However, the IVOA[5] has settled on a flavor of XML called VOEvent for the transmission of astronomical events, and processing those would need XML streaming and filtering technologies. There is though a further complication here: we really want to be processing portfolios (sets of events with citation structure).

Let us consider some processes that can happen once a survey detects and announces a new transient. First a new portfolio is created to hold whatever data is available, and then some of these may happen:

- **Adding rich data:** for example from previous data of the same survey or fetching from remote archives, adding results of software evaluation, or adding human writing.
- **Follow-up observation:** when observations are taken of the event, the results should be fed back into the portfolio, so that a complete picture is available. This is an important class of rich data.
- **Summarizing:** New data is added to the portfolio that converts earlier annotation into higher level semantic concepts, such as presence of host galaxy or radio source, association with known nova, etc. A summary is not “rich” data, but rather a condensation of rich data into perhaps a single number that expresses its meaning.
- **Triggering:** The portfolio is used in a Boolean function to determine what actions to take. Different purposes and different people will use triggers to get precisely the events they want.
- **Taking Action:** Each trigger is associated with an action to take if that trigger evaluates True. Examples of actions are messaging, execution of software, remote robotic observation, etc. Actions can also generate new data that is added to the portfolio.

These processes can drive the entire enterprise of event dissemination through a cyclic workflow: a trigger fires, causing new data to be added to the portfolio, so that another trigger fires, causing something else, etc. At the survey itself, triggers can determine whether the event is even made public, and receivers of the event can use triggers to determine the accumulation of data to the portfolio – for example “if the event is in the footprint of survey X, then fetch and annotate with what survey X has.” Requesting follow-up, deciding that an event is interesting, and even waking up a human for an opinion, can all be done through this trigger-action paradigm.

Human Computing

While machines are capable of many types of information processing, they are not so good where something new is present that has not been programmed. Certainly we expect the future flood of events to be mostly handled by machines, with the uninteresting ones never seen by a human expert, but some may be escalated in importance and come to the attention of such experts through a message, or even being urgently awoken in the night.

There could be a large number of other people also involved in the enterprise, volunteers recruited from the internet, with some, but by no means expert, ability. All people have excellent image analysis capabilities: they could, for example, look at an image of a star field and determine quickly and accurately if there is an artifact, such as a satellite trail, interference from a nearby bright star, or one of many Earthbound artifacts: from the telescope, camera, or electronics. While many of these types of common artifacts can be detected by machine, there are always new types, or artifacts that are a combination of

known types. Since the transient detection software is looking for differences between new and past observations, such artifacts, though rare, will be inevitably found and thus pollute the event stream.

This type of ‘citizen science’ has been both popular and extremely useful in GalaxyZoo[6] and CitizenSky[7], and we expect it to be the same with events. A new aspect with events, different from the traditional web-based citizen science, could be that events are ‘pushed’ to the volunteers, so they can respond with their mobile device immediately. Another novel aspect to citizen science could be the recruiting of a cadre of dedicated volunteers to work at a more expert level, looking at light curves or other non-image data; they would need to be sufficiently motivated to study and take a test, to be inducted to this higher level.

Data Mining

Software components can be executed, perhaps in response to a trigger, and the results added back into the event portfolio. For example, taking four images of the sky, separated by several minutes in time, is a good way to find asteroids; this is because software can compare the four source lists and detect the moving object. In this case, the result is a probability that an asteroid is present, together with position, velocity etc. Thus complex data is reduced to *meaning*: a single numerical “probability of asteroid”. Other annotation components can be used to determine the association between the new transient and a host galaxy, which increases the chance that the transient is from a supernova, or between the transient and a known radio source, which increases the chance that the transient is from an active galaxy.

Information Fusion

When all the annotation is assembled into a portfolio, a difficult problem arises, which is summarizing all that information into a classification of the transient (supernova, asteroid, cataclysmic variable etc), and further, converting that classification into a decision about follow-up priority. Even when the problems of multi-sourced data are solved, it is still very difficult for a machine to fuse all the information into a trustworthy decision. Much work has been done on such problems in the machine learning community. However the best successes have come when there is a comprehensive training set that includes good coverage of all possible outcomes; or when each transient is represented with the same high-quality data, such as a long and regularly-sampled light curve, high-accuracy photometry, or a spectrum. Unfortunately, transient events are not like this: the most interesting objects are the rare ones, and the high-quality data is that which appears as a result of the expensive follow-up. Another difficulty is that astronomers are less interested in the objects that are already known and classified: rather they want the ones of a rare class that are not yet identified by others.

References

- [1] LOFAR: <http://www.lofar.org/>, Pann-STARRS: <http://pan-starrs.ifa.hawaii.edu/>, LSST: <http://www.lsst.org/> and SKA: <http://www.skatelescope.org/>
- [2] R.C. Smith, R. Seaman, and P. Warner, *Integrating VOEEvent into the ground-based O/IR system at NOAO*, *Astron. Nachr.* 329 (2008) 241.
- [3] Alasdair Allan et. al., *eSTAR: Building an Observational GRID*, <http://adass.org/adass/proceedings/adass02/reprints/FO1-3.pdf>
- [4] StreamSQL
<http://en.wikipedia.org/wiki/StreamSQL>
- [5] Rob Seaman, Roy Williams, Alasdair Allan et. al, *Sky Event Reporting Metadata (VOEvent)*, a Recommendation of the IVOA,
<http://www.ivoa.net/Documents/latest/VOEvent.html>

[6] GalaxyZoo
<http://www.galaxyzoo.org/>

[7] CitizenSky
<http://www.citizensky.org/>