Digging Deeper:
Algorithms for Computationally-Limited Searches in Astronomy

Report of a Study Program

Study start date: June 7, 2011

Study end date: December 15, 2011

Team Leads:

George Djorgovski (Caltech), george@astro.caltech.edu
Curt Cutler (JPL, Caltech), curt.j.cutler@jpl.nasa.gov
Bruce Elmegreen (IBM), bge@us.ibm.com

Co-Leads:

Michele Vallisneri (JPL, Caltech), vallis@caltech.edu
Ashish Mahabal (Caltech), aam@astro.caltech.edu

California Institute of Technology

December 2012

Table of Contents

Executive Summary

Astronomy, like most other fields, is being deluged by exponentially growing streams of ever more complex data. While these massive data streams bring a great discovery potential, their full scientific exploitation poses many challenges, due to both data volumes and data complexity. Moreover, the need to discover and characterize interesting, faint signals in such data streams quickly and robustly, in order to deploy costly follow-up resources that are often necessary for the full scientific returns, makes the challenges even sharper. Examples in astronomy include transient events and variable sources found in digital synoptic sky surveys, gravitational wave signals, faint radio transients, pulsars, and other types of variable sources in the next generation of panoramic radio surveys, etc. Similar situations arise in the context of space science and planetary exploration, environmental monitoring, security, etc. In most cases, rapid discovery and characterization of interesting signals is highly computationally limited. The goal of this study was to define a number of interesting, often mission-critical challenges of this nature in the broader context of time-domain astronomy, but with an eye on their applicability elsewhere. Three types of challenges were identified and followed through the duration of this study:

1. Searching for Long, Weak Gravitational Wave Chirps and for Microlensing Events. The first part of this problem is of a critical importance for the nascent field of gravitational wave astronomy, but it is also highly relevant for the searches for heavily dispersed pulsar signals in radio data cubes, or in γ-rays. The second aspect of the problem is to find gravitational microlensing events with characteristic signatures of planets around the lensing star. We invented of a couple new techniques to increase search efficiency, and the effort continues, with another technique added since the study's completion. The current set of methods for this analysis yet has to be optimally combined into a full data analysis pipeline, requiring manpower, and this remains a very worthy and a attainable goal for future work in the near-to-mid-term.

2. Intermittent, Sub-Significant Detections in Data Cubes. In a series of images where the third axis represents time or different wavelengths, there may be sources that appear only intermittently, but that are not statistically significant in any one epoch or channel. If the *right subset* of these were to be averaged, the detection would be significant, but averaging all of them would dilute the signal. An easier version of the problem is if the position of a possible source is already defined; a more challenging application is to blind searches. A solution to this problem could increase the effective depth of multi-epoch sky surveys from both ground or space. A novel, statistically based method was developed for this purposes, and implemented as a software package. It is now being scientifically validated on the data from actual sky surveys.

3. Rapid, Automated Classification of Variable and Transient Sources. Scientific returns from synoptic sky surveys are now increasingly limited by the ability to follow up the most interesting sources and events. Given the time-critical nature of such events, their rapid characterization or classification is essential for an optimal deployment of limited follow-up resources. The problem is complicated by the sparsity and heterogeneity of the data, and the presence of rtifacts that may masquerade as transient signals. The process has to be complete (no good signals are missed) and with a low contamination by false alarms. Automated classification of light curves is also essential for the archival exploration of synoptic sky survey archives. We explored and developed a number of new statistical and Machine learning approaches, that are now being scientifically validated on the actual sky survey data streams. Work continues along all of these avenues that were started or substantially expanded during the KISS study.

1. Motivation and Goals of This Study

In several areas of astronomy the sensitivity of our searches for some types of signals is computationally limited. That is, either faster computers or better algorithms would lead to more discoveries in the same datasets. This is certainly true for many cases in gravitational-wave data analysis. For example, for LIGO, the current searches for unknown gravitational-wave (GW) pulsars are strongly computationally limited. For the proposed space-based GW detector, LISA (the Laser Interferometer Space Antenna), the most extreme such example will be searches for signals from stellar-mass black holes inspiraling into very massive black holes in galactic nuclei.

Improved algorithms are also critical in the rapidly developing field of time-domain astronomy, where transient signals from a variety of interesting astrophysical phenomena, ranging from the Solar System to cosmology and extreme relativistic objects, must be discerned in massive data streams. Some examples include: (1) searches for millisecond pulsars, flaring blazars, or other transient sources in Fermi (gamma-ray) data; (2) searches for short-period binary pulsars in radio data from the current or soon forthcoming surveys, e.g., GBT, Arecibo, EVLA, ASKAP and MeerKAT; (3) discoveries and characterization of transient sources detected in the current (e.g., CRTF, PTF, etc.) or future (e.g., LSST) synoptic sky surveys; (4) all-sky searches for radio sources in (future) SKA data; (5) searches for near-Earth asteroids, and (6) SETI. In addition to the many known types of objects in the time domain (e.g., supernovae and variable stars and AGN of various kinds), there is a real possibility of discovery of heretofore unknown types of objects or phenomena.

Yet, the scientific returns from these missions and experiments are often limited by the ability to detect, recognize, and classify interesting signals in them. *Given the cost of the acquisition of the data, these scientific opportunity costs are unacceptable.* The overall goal of this study was to help increase or even optimize the scientific returns from these massive data streams, especially in S/N-limited situations, through a development of novel and faster algorithms.

To keep the scope manageable, we limited our investigations to time series, which could be light curves of sources detected in multiple images, or a detector output like LIGO. There are two types of related challenges: (1) detection of faint and/or transient signals, and (2) their classification/characterization, which by itself can inform the detection process through a design of optimal detection algorithms, and is essential for the follow-up prioritization of the detected signals and events.

While a lot of work has gone into developing different methods in the various areas listed above, there has been little inter-comparison of methods, and even less development of understanding and intuition regarding which known methods are best for which sorts of searches. Our initial technical goal was to develop a few realistic, benchmark problems on which the methods can be compared, keeping in mind computational resources and available architectures. In practice, we planed to define 2 or 3 specific, timely, astrophysically motivated challenges to guide our thinking and serve as methodological testbeds.

For example, since many astrophysical signals expected in the GW domain are “chirps” (i.e., with a time-dependent frequency, with periodic signals being a special case of chirps), we planed that one benchmark problem will be to search for weak, long-lived, multi-parameter chirps imbedded in noise. We intended to compare hierarchical, grid-based methods; various flavors of

MCMC methods; time-frequency methods, and likely other methods as well. For example, the parameter space may be divided into regions where different methods are preferable.

Detection and characterization (i.e., classification) of transient events in synoptic surveys poses multiple challenges (see, e.g., Mahabal et al. 2008ab, Donalek et al. 2008, Djorgovski et al. 2006). Classification is essential for an optimal follow-up strategy in situations where resources are limited (e.g., telescope time, computing, etc.), especially if a timely response is essential (as it usually is): there are the opposing demands for a high completeness (don't miss any interesting events) and a low contamination (minimize the false alarms). It also informs the detection process, as different filters can be deployed to optimize the detection of particular kinds of transients, or eliminate particular types of artifacts or backgrounds. Since the original data are often sparse and/or incomplete, sometimes with an uneven sampling and S/N, use of all available archival multi-wavelength, multi-epoch, multi-scale, textual, contextual, and therefore even more heterogeneous data is essential. Combining such heterogeneous information in a systematic and statistically justified manner, and a proper quantitative encoding of the relevant contextual information (e.g., is there a possible host galaxy nearby, how unlikely is it to find a particular type of a transient in a given part of the sky, what is the effect of crowding in that particular image, etc.), is a real challenge. There are currently *no* established approaches and practices that effectively optimize this problem.

In order to exploit the full informational content of the data (i.e., find fainter sources), we have to go deeper. This is relevant for the construction of past time histories for any detected event (e.g., a source may have been flickering just under the detection threshold, but multiple weak detections add up to a statistically significant one). A straight co-addition washes out the transients in the data. To retain signal that exists in just a subset of images, better algorithms are needed (e.g., using the Mahalanobis distance; see Babu et al. 2006).

Our aim was to compile a practical guide to the best available methods of attacking these problems. Our work should lead to an improved understanding and useful rules of thumb regarding the advantages and scaling properties of different methods, which can be carried over to other data analysis challenges, and that could also lead to some immediate scientific results.

This study could have a significant impact on the science reach of *both* ground-based and space-based instruments (including Earth-science as well as Astronomy/Astrophysics), and could well have an impact in other areas, such as intelligence/defense, health care and finance. Though interest in this area is clearly ramping up (witness, e.g., the new "Time Series Center" at Harvard), in many ways this field is still in a very immature state. Individual researchers have been developing their own tricks and techniques in a rather ad hoc way, without much input from researchers working on similar problems in other fields. The field cries out for consolidation, including a rigorous comparison of different methods against benchmark problems, and the development of useful rules of thumb to guide algorithm developers.

Past experience shows that work devoted to optimizing algorithms can very often lead to factors ~10 – 100 reductions in computational cost. For example, Cutler et al. (2005) showed that three-stage, hierarchical searches for gravitational-wave pulsars would be ~100 times more efficient than the one-stage methods generally used. Given the current states of both science and data analysis, paying one scientist to develop and implement a better algorithm can often be far more cost effective than paying for bigger and better telescopes/instruments.

At least a couple recent developments are practically demanding this kind of a study. First, very sensitive ground-based gravitational-wave (GW) detectors (LIGO and Virgo) came on-line in the last few years, and are currently being upgraded; no detections have been made so far, and the first detections will likely be just above threshold, so GW data analysts are by necessity "pushing the envelope" to develop the most sensitive algorithms for finding weak signals in noisy (and very expensive) data. There is a substantial joint effort in the emerging field of GW astronomy, involving both LIGO and LISA, with a very strong computational component in numerical relativity. Better algorithms (which would also benefit LISA) would greatly increase the scientific returns from these facilities.

There are also ongoing efforts in electromagnetic time domain astronomy, both with ground-based surveys (PQ, CRTS, PTF) and space missions (GALEX, WISE, NuSTAR, WFIRST, etc.). Ever larger synoptic sky surveys (in both area and time, and in several different wavebands) are at the cutting edge of astronomy today, enabled by the large-format detectors and computation technology. They are dominating the current data volumes, leading to the future facilities like the LSST and SKA, which are of a great interest to the Caltech-JPL community. Faster and more effective ways of extracting, characterizing, and following astrophysically interesting transients in these upcoming Petascale data streams would produce increased scientific returns for all, and the novel software methodology which may be seeded here would place Caltech and JPL in a more competitive position in the era of a massive, panoramic, multi-wavelength cosmic cinematography, and help make JPL an attractive NASA Center partner to PI-led missions.

It is worth noting a broader significance and context of this work. Analysis and mining of massive data streams is not confined to astrophysics, and similar challenges occur in many other arenas of space science, remote sensing, environmental monitoring, as well as electronic commerce, security, etc.

While our focus was on the algorithms, we note that there have been substantial recent advances on the custom and commodity hardware front as well (e.g., GPU-based clusters and pipelines), and we kept these possibilities in mind, in terms of the optimal hardware-software combinations.

We started the study with a 5-day opening workshop, for brainstorming and selecting benchmark problems. The workshop started with a half-day short course for students and postdocs, to help get them up-to-speed, especially since many of them were coming from a broad range of backgrounds. We expect that these young scientists will become the real practicing experts in this field. The remaining part of the opening workshop was used to lay out the challenges and identify some specific problems to be tackled, and the possible paths toward their solutions.

In the following 6 months (June – December 2011) the work continued on these challenges in the form of distributed collaborations, using a variety of mechanisms for discussions and exchanges of ideas and results. These ranged from email and Skype, the KISS study wiki, to immersive VR meetings, enabled by a separately funded experimental program by the Caltech group.

The study was concluded with a 3.5-day closing workshop where some results and the status of the ongoing studies were presented to a broader audience, and the discussions among the study team about the next steps to be taken, and the possible follow-up projects. Indeed, the work is still ongoing, and we anticipate that a number of projects initiated or defined during this KISS study will continue in the future.

2. The Opening Workshop, June 7 – 10, 2011

This Workshop marked a formal start of the study. The participants are listed in the Appendix. They included astrophysicists working on various aspects of time-domain astronomy, statisticians, and computer scientists.

2.1 Workshop Agenda

The workshop started with a short course on "Looking for Nuggets in Massive Datastreams", open to a broad audience that included students, postdocs, and other researchers, with four invited lectures:

Speaker	Title
Badri Krishnan	Gravitational Wave Data Analysis
Jeff Scargle	New Developments in Time Series Analysis
Ashish Mahabal	Automated Classification of Transients
Pavlos Protopapas	Machine Learning and Statistics Applications

It then continued with a number of technical presentations for the audience restricted to the invited study participants. These were intended to set the stage and initiate the discussions:

Speaker	Title
George Djorgovski	Real-time mining of Petascale data streams
Curt J. Cutler, and Bruce Elmegreen	Presenting areas to be reviewed, and proposed benchmark problems
Giuseppe Longo	Computationally limited tasks in astronomy?
Kiri L. Wagstaff	Machine Learning Methods for Astronomy
Joe Lazio, and Mike Turmon	2 lightning talks on research areas related to problems, and discussion on Machine Learning Methods for Astronomy
John A. Rice	Heirarchical Resolution
Baback Moghaddam	What can Biostatistics do for Time-Domain Astronomy
Rosanne Di Stefano	Extracting Periods from Binary-Lens Events: A Slightly Modified Lomb-Scargle Approach
Ciro Donalek	Objects Classification in Synoptic Sky Surveys:

	Contextual and External Information
David Thompson	Hidden Markov Models in 10 Seconds
Raffaele D'Abrusco	The Exploration of Multi-Wavelength Astronomical Datasets: AGNs in the Chandra Source Catalog and Unsupervised Clustering
Walter Max-Moerbeck	What is the Connection Between Radio and Gamma-Ray Emission in Blazars?
Guillermo Cabrera	Automated Detection of Objects Based on Sérsic Profiles
Jeff Scargle	A Few Comments on Time Series Representations

The schedule, with links to most of the presentations (slides, and videos of the short course lectures) can be found at http://kiss.caltech.edu/workshops/digging2011b/schedule.html.

An emphasis was given to unrestricted discussions during and after the talks, intended to create ideas and stimulate the subsequent work. The talks also included several talks by postdocs (e.g., Donalek, Thompson, D'Abrusco) and graduate students (Max-Moerbeck, Cabrera), presenting their research projects. In addition, there were several informal, ad hoc presentations on the emerging topics that were not formally scheduled.

2.2 Summary of the Selected Presentations

It is useful to recap some of these presentations, as they established the scientific and technological context for the study, describing both the challenges and some possible approaches and methods that may be used in tackling them.

In the public (open) session, four reviewers described the state of the art and the outstanding challenges in the analysis of GW signals (B. Krishnan), time series analysis (J. Scargle), automated classification of transient events in synoptic sky surveys (A. Mahabal), and machine learning and statistical methods for the automated classification of light curves (P. Protopapas).

In setting the stage for the remaining part of the workshop, the organizers identified some specific problems that should be addressed: detection of transient or variable sources from multiple, sub-significant detections in synoptic sky surveys, and automated classification and classification-informed detection of transient or variable sources (Djorgovski 2011), and searches for long, weak "chirps" in gravitational wave astronomy (Cutler). These challenges formed the principal foci of the subsequent discussions. We describe the relevant presentations in more detail below.

2.2.1 Gravitational Wave Astronomy: Searching for Chirps

While the interaction of gravitational waves (GWs) with ordinary matter is notoriously weak, astrophysical GW events can be extremely energetic. For instance, the final merger of two comparable-mass black holes is roughly two orders of magnitude more luminous than all the rest of the observable Universe – 100 times brighter than the combined energy flux from all the stars

in all the galaxies within a Hubble volume. The mergers of $\sim 10^6$ solar-mass black holes (the prototypical sources for a LISA-like GW observatory) last approximately one hour.

The effect of these waves on a GW detector on or near the Earth is to modulate the distance between freely floating test masses. The fractional modulation is tiny, usually less than 10^{-21} even for the strongest sources. This is why GW observatories have taken 40 years of instrument development to (almost) reach the sensitivity necessary to detect GW sources. Even so, the first detected signals are likely to be buried in detector noise – essentially invisible to the eye in a plot of the measured time series – and will require optimal *matched filtering* to be "dug out" from the noisy data streams. The GW community has therefore put a premium on the development of the most sensitive possible data-analysis techniques.

For some kinds of sources, GW searches will be computationally limited, such that if we had greater computer power or more efficient algorithms we could dig even deeper into the noise. For ground-based detectors such as LIGO and VIRGO, the prototypical example of a computationally limited search is an all-sky search for continuous, nearly sinusoidal GWs from a rapidly rotating neutron star (NS) in the Galaxy. Note that a perfectly axisymmetric NS rotating around its symmetry axis would not emit any GWs, but there are several physical mechanisms that could distort a NS into slight nonaxisymmetry: a fractional distortion $\sim 10^{-7}$ could already be detected by the "Advanced" ground-based GW detectors that will be coming online starting in 2014. These "GW pulsars" provide an illustrative example of a computationally limited search, so it is worthwhile to examine them here. For simplicity, we will limit our discussion to order-of-magnitude accuracy.

A "typical" detectable GW-pulsar signal would have dimensionless amplitude $\sim 10^{-26}$, four orders of magnitude below the detector noise of Advanced LIGO/VIRGO. Fortunately, these sources are always "on," and (for a typical GW frequency ~300 Hz) emit $\sim 10^{10}$ cycles over a year. The huge number of cycles compensates for the tiny instantaneous GW amplitude. Since in matched filtering the amplitude signal-to-noise ratio (SNR) grows as $(N_{cyc})^{1/2}$, the SNR from a year-long integration is thus of $\sim (10^{-4})(10^5) \sim 10$.

In computationally limited all-sky searches, the GW pulsars have no known electromagnetic counterparts (e.g, no associated radio or X-ray sources), so they could be anywhere on the sky. They could also have any spin frequency and any spindown rate. If the search includes young pulsars (say, younger than 100 years), then the second and third derivative of the of spin period are also significant parameters. In that case, the detector data must be filtering with $\sim 10^{22}$ independent signal templates. (Here we are not including the $\sim 10^{10}$ different frequencies that must be searched over, which can be spanned simultaneously using the "magic" of the FFT; otherwise the total number of templates would be $\sim 10^{32}$. With the FFT the cost of searching over frequencies scales only as the logarithm of the number of frequency bins.) It follows that a straightforward matched-filtering search over such a template bank, even with a Teraflop computer, would take longer than the age of the Universe.

To use computationally practical methods, we must accept nonoptimal sensitivity. Currently, there are two broad approaches to building practical, nonoptimal searches. The first approach is to search parameter space hierarchically: we first cover parameter space with a very coarse template grid, and look for SNR outliers that could represent true GW signals. We then form finer grids around these outliers, and note for further examination any templates in the finer grid that exceed a higher threshold. Possibly, we repeat this iteration through multiple stages. This

strategy is illustrated in Fig 1. Its usefulness in GW searches was demonstrated by Cutler and colleagues (2005); Meinhausen et al. (2009) provided an elegant formulation of this idea in a more general setting.

The second approach is to divide the total observation time into relatively short stretches of data of duration ΔT (typically ~day for GW searches), and to perform matched-filtering searches on the individual segments. The SNR^2 of each template is calculated for each segment, and summed through. This strategy is illustrated in Fig. 2. The GW community borrowed this strategy from pulsar radioastronomers, who call it "power stacking," but GW data analysts call it "semicoherent search". In a fully coherent search, the detection statistic would be the square of the sum of the complex SNRs from all the short segments, while the semicoherent detection statistic is the sum of the squares. The advantage of the latter is that this function is much less narrowly peaked in parameter space: while the maximum of the statistic is lowered, its peaks are broadened. This means that parameter space can be covered with a much coarser grid, greatly reducing the overall computational cost of the search. In fact, the most sensitive current searches for GW pulsars incorporate both approaches.

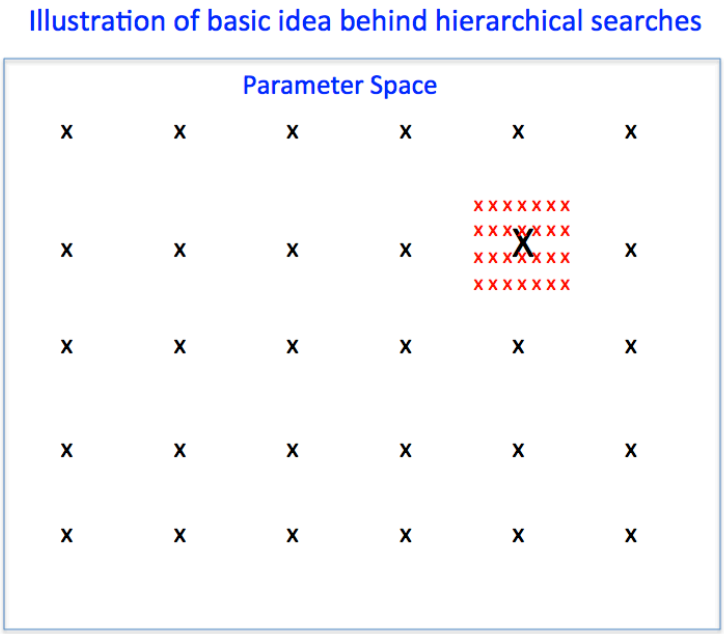

Figure 1. Basic strategy of a typical hierarchical search. The small black x's are the original coarse grid. The large black X is a threshold-exceeding "candidate," which is then followed up with finer grid of red x's.

Last, we mention that the all-sky search for GW pulsars is hardly the most computationally intensive search that the GW data-analysis community needs to confront. Proposed space-based GW missions similar to LISA (NASA and ESA's "Laser Interferometer Space Antenna") would be able to detect the GWs emitted by stellar-mass black holes spiraling into ~million-solar-mass black holes in galactic nuclei, out to redshift $z \sim 1$. These sources are known as "extreme-mass-ratio inspirals," or EMRIs. The parameter space for a single EMRI signal is effectively (at least) 14-dimensional, compared to ~6 dimensions for the GW pulsar case; furthermore, EMRI signals are considerably more complicated. The higher dimensionality is often tackled with stochastic

search methods, such as Markov Chain Monte Carlo integration, that walk quasirandomly across parameter space instead of sampling it regularly.

Indeed, within the LISA community one would generally regard GW-pulsar searches as mere warm-ups for EMRIs. It seems likely that any new ideas developed for improving the efficiency of GW pulsar searches would also be useful for the EMRI problem. For a basic discussion of the EMRI search problem we refer the reader to Gair et al. (2004).

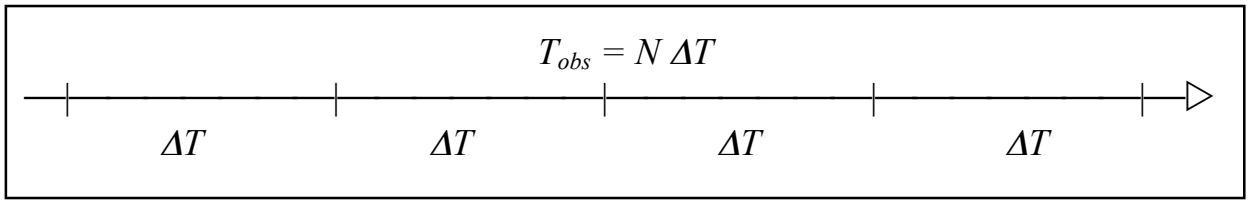

Figure 2. In a semicoherent search, the full data set of duration T_{obs} is divided into N short segments of length ΔT. Each short segment is searched coherently, and the resulting complex SNRs are summed incoherently. This effectively smoothes out the detection statistic, allowing parameter space to be covered by a much coarser grid, which very significantly reduces the computational load.

2.2.2 Synoptic Sky Surveys: Detection and Classification of Transient Events

S. G. Djorgovski placed these challenges in the context of real-time mining of massive data streams. Such streams are being generated by a new generation of scientific measurement systems (e.g., survey telescopes, instruments or sensor networks), that are now moving into the Petascale regime. This exponentially growing wealth of data can enable significant new discoveries, provided that the relevant knowledge is extracted efficiently and rapidly. Often, the interesting phenomena are objects, sources, or events where a rapid change occurs. They have to be identified, characterized, and possibly followed by new measurements in the real time. The requirement to perform the analysis rapidly and objectively, coupled with huge data rates, implies *a need for automated classification and decision making*.

This entails some special challenges beyond traditional automated classification approaches, which are usually done in some feature vector space, with an abundance of self-contained data derived from homogeneous measurements. Here, the input information is generally sparse and heterogeneous: there are only a few initial measurements, and the types differ from case to case**,** and the values have differing variances; the contextual information is often essential, and yet difficult to capture and incorporate in the classification process; many sources of noise, instrumental glitches, etc., can masquerade as transient events in the data stream; new, heterogeneous data arrive, and the classification must be iterated dynamically. Requiring a high completeness (don't miss any interesting events) and low contamination (a few false alarms), and the need to complete the classification process and make an optimal decision about expending valuable follow-up resources (e.g., obtain additional measurements using a more powerful instrument at a certain cost) in real time are challenges that require some novel approaches.

While this situation arises in many domains, it is especially true for the developing field of time domain astronomy. Telescope systems are dedicated to discovery of moving objects (e.g., potentially hazardous, Earth-crossing asteroids, transient or explosive astrophysical phenomena,

e.g., supernovae (SNe), γ-ray bursts (GRBs), etc. – each requiring rapid alerts and follow-up observations. The time domain is rapidly becoming one of the most exciting new research frontiers in astronomy (Paczynski 2000, Griffin et al. 2012, Djorgovski et al. 2012a), broadening substantially our understanding of the physical universe, and perhaps lead to a discovery of previously unknown phenomena (Djorgovski et al. 2001ab, 2006).

The key to progress in time-domain astrophysics is the availability of substantial event data streams generated by panoramic digital synoptic sky surveys, coupled with a rapid follow-up of potentially interesting events (photometric, spectroscopic, and multi-wavelength). Physical classification of the transient sources is the key to their interpretation and scientific uses, and in many cases scientific returns come from the follow-up observations that depend on scarce or costly resources (e.g., observing time at larger telescopes). Since the transients change rapidly, a rapid (as close to the real time as possible) classification, prioritization, and follow-up are essential, the time scale depending on the nature of the source, which is initially unknown. In some cases the initial classification may remove the rapid-response requirement, but even an archival (i.e., not time-critical) classification of transients poses some interesting challenges.

A number of synoptic astronomical surveys are already operating (see, e.g., Djorgovski et al. 2012b for a review and references); examples include Palomar-Quest (PQ; Mahabal et al. 2005, Djorgovski et al. 2008), Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009, Mahabal et al. 2011, Djorgovski et al. 2012a; http://crts.caltech.edu), Palomar Transient Factory (PTF; Rau et al. 2009, Law et al. 2009; http://www.astro.caltech.edu/ptf/), or PanSTARRS (Kaiser 2004; http://pan-starrs.ifa.hawaii.edu/). Much more ambitious enterprises such as the LSST (Tyson et al. 2002, Ivezic et al. 2009; http://lsst.org) and SKA (http://skatelescope.org) will move us into the Petascale regime, with hundreds of thousands of transient events per night, implying a need for an automated, robust processing and follow-up, sometimes using robotic telescopes.

Thus, *a new generation of scientific measurement systems is emerging* in astronomy, telescope and computational networks. A similar situation exists in many other fields: connected sensor networks which gather and analyze data automatically, and respond to outcome of these measurements in the real-time, often redirecting the measurement process itself, and without human intervention.

A full scientific exploitation and understanding of astrophysical events requires a rapid, multi-wavelength follow-up. The first challenge is to associate classification probabilities that any given event belongs to a variety of known classes of variable astrophysical objects and to update such classifications as more data come in, until a scientifically justified convergence is reached. Perhaps an even more interesting possibility is that a given transient represents a previously unknown class of objects or phenomena that may register as having a low probability of belonging to any of the known data models. The process has to be *as automated as possible, robust, and reliable*; it has to operate from *sparse and heterogeneous data*; it has to maintain a *high completeness* (not miss any interesting events) yet a *low false alarm rate*; and it has to *learn* from the past experience for an ever improving, evolving performance. The next step is development and implementation of an automated follow-up event prioritization and decision making mechanism, which would actively determine and request follow-up observations on demand, driven by the event data analysis. This would include an automated identification of the most discriminating potential measurements from the available follow-up assets, taking into

account their relative cost functions, in order to optimize both classification discrimination, and the potential scientific returns.

While some machine learning approaches have been used for elimination of artifacts in synoptic sky survey data streams (e.g., Romano et al. 2006, Bailey et al. 2007, Donalek et al. 2008), typically using the image morphology alone, the problem of physical classification of transient events is much harder. All transient events look the same in the images (star-like), so that information other than image morphology must be used. One problem is that in general, not all parameters would be measured for all events, e.g., some may be missing a measurement in a particular filter, due to a detector problem; some may be in the area on the sky where there are no useful radio observations; etc.

A more insidious problem is that many observables would be given as upper or lower limits, rather than as well defined measurements; for example, "the increase in brightness is > 3.6 magnitudes", or "the radio to optical flux ratio of this source is < 0.01". One approach is to treat them as missing data, implying a loss of the potentially useful information. A better approach is to reason about "censored" observations that can be naturally incorporated through a Bayesian model by choosing a likelihood function that rules out values violating the bounds.

Additional approaches to an automated classification of transient events include, e.g., Mahabal et al. (2008ab, 2009, 2010c), Donalek et al. (2008), Bloom & Richards (2011), Djorgovski et al. (2011), etc.

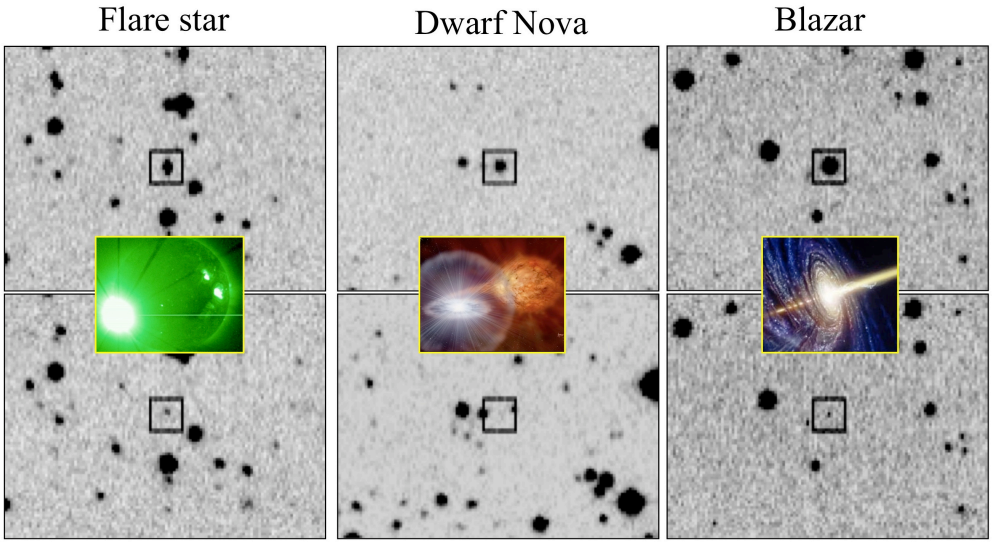

Figure 3. Examples of transient events from the CRTS sky survey. Images in the top row show objects that appear much brighter that night, relative to the baseline images obtained earlier (bottom row). On this basis alone, the three transients are observationally indistinguishable, yet the subsequent follow-up shows them to be three vastly different types of phenomena: a flare star (left), a cataclysmic variable (dwarf nova) powered by an accretion to a compact stellar remnant (middle), and a blazar, flaring due to instabilities in a relativistic jet (right). Accurate transient event classification is the key to their follow-up and physical understanding.

2.2.3 Automated Classification of Light Curves

J. Scargle described the analysis of light curves (LCs) with a degree of abstraction: a time series can be regarded as a sequence of *N* arbitrary time-ordered data cells, which contain all information relevant to any analysis task, with individual data cells representing individual measurements or observations. Data models that are represented by LCs depend on the physical nature of the phenomenon (e.g., a supernova, or a variable star of some type). The Wold Decomposition Theorem states that any stationary process can be represented as the sum of two parts, one of which is completely deterministic and the other which is completely random (called the moving average, essentially a white noise process run through a filter). Astronomical LC observations depend on a number of separate random processes both at origin, propagation and detection – luminosity fluctuations, photon emission and detection, scintillation, dispersion, etc. LC classification process can thus be viewed as a search for most likely data models.

P. Protopapas described the work done at the Time Series Center at the Harvard-Smithsonian Center for Astrophysics, that started in 2008. Light curves for over 100 million objects from surveys like MACHO, EROS, and now Pan-STARRS are being analyzed using a variety of approaches. The principal scientific goals of the Center include: (1) classification of variable stars, supernovae etc., (2) period finding from light curves, (3) outlier detection, and (4) time-series modeling in general.

Several techniques have been applied to these large datasets. These include: SVN was used with MACHO and EROS datasets to find 14,000 new periodic variable stars (Protopapas 2012); Euclidian distances (Protopapas 2006) and active learning (Majidi 2012) have been used to locate anomalous lightcurves; Wavelets for event detection are being; Gaussian Processes (Wang et al. 2012) and Correntropy method (Huise et al. 2011) are being used for period finding.

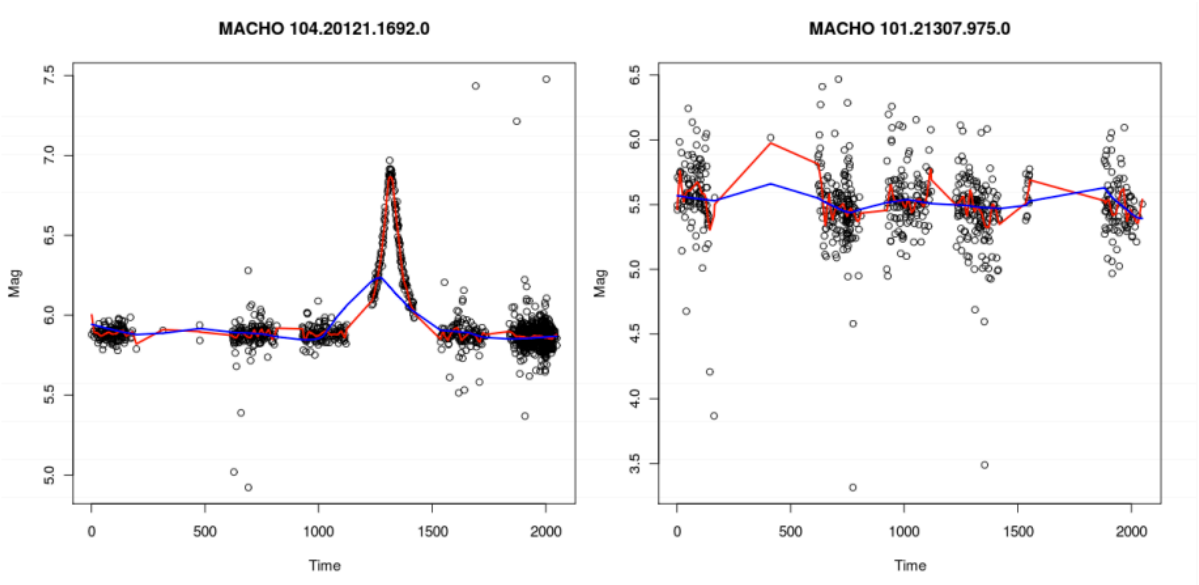

Figure 4. Event detection using wavelets. Wavelets are fit to the time-series and low frequencies compared to high frequencies. The difference in maximized log-likelihoods of low and high components then leads to the detection of events: $2 \times (L_{high} - L_{low})$.

2.2.4 Gravitational Microlensing: Searching for Planets

In some ways, searches for gravitational lensing events that also include signatures of planets around the lensing stars, described here by R. Di Stefano, pose challenges similar to those in the searches for GW chirps (Sec. 2.2.1).

Gravitational lensing offers a potentially transformative way of studying the (mass) distribution of nearby (compact) objects. Irrespective of whether a gravitational lens is dark or luminous, we can measure its mass and tell if it has dark or dim companions (see Fig. 5).

Nearby lenses – *mesolenses* – have larger Einstein angles (astrometric effects may be seen) and higher proper motions so we are more likely to be able to detect them (the event rate is higher for nearer lenses than for more distant ones), or even predict them in some cases. By focusing on systematically identifying these type of lenses, we can discover more of the 10^6 neutron stars and black holes within about 1 kpc and determine the mass distributions of low-mass objects, stellar remnants and stars and the frequency of binaries and planetary systems. The lensed sources can also be better studied, especially if the magnification is extreme or if there are finite source-sized effects.

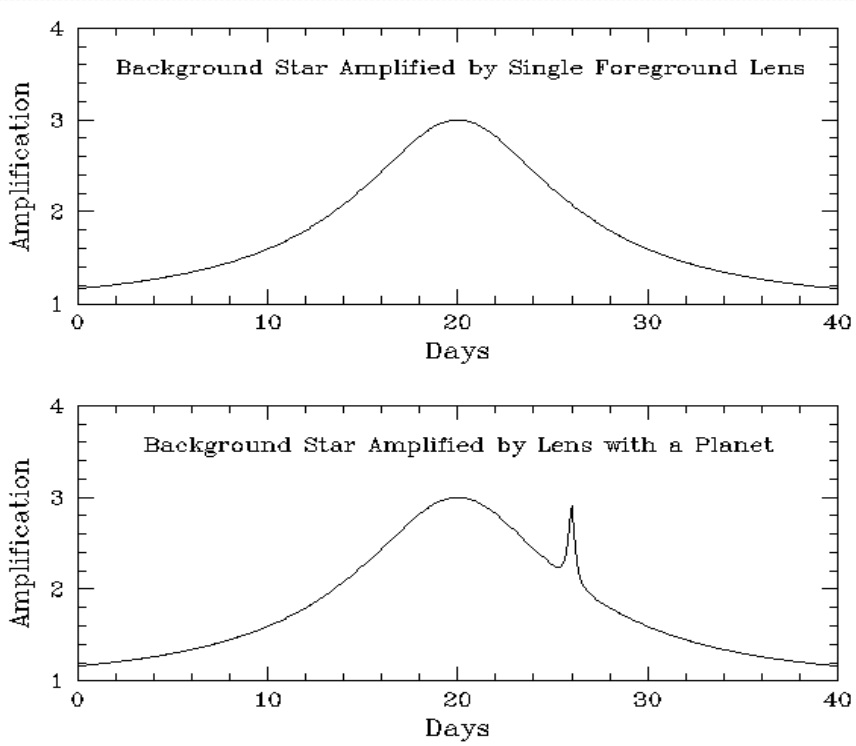

Figure 5. This shows the difference in brightness variation between a lensing event by a single object and one with a companion.

Mesolensing can be found in data from monitoring programs – to date, there are about 8500 known lensing event candidates with a current discovery rate of >1500/year and potentially ~10000/year over the whole sky. Between 10% and 20% of the lenses lie within ~1 kpc and a large fraction of these are dwarf stars. Many of these are likely to have planets and a few percent are compact objects. Additional events can also be predicted for HST, Magellan and other

programs. Kepler data is of sufficient resolution to detect subtle effects and so far has produced about ten serendipitous events among the target stars. Of particular note are those events caused by white dwarfs (or neutron stars or even black holes) which have the appearance of "anti-transits" – short-lived enhancements in the amount of light received from the monitored star. Lastly, there is the potential for finding mesolensing events in data from wide-field surveys such as Pan-STARRS.

The MEarth project (Nutzman & Charbonneau 2008) aims to discover a 2 R_+ transiting habitable planet around a nearby M dwarf. It has been photometrically monitoring 2000 nearby, low-mass M dwarf stars since 2008, with each star being observed once every 20 minutes. Extrapolation of the astrometric motion of the nearby low-mass high-proper-motion star VB 10 indicates that sometime in late 2011/early 2012, it will make a close approach to a background point source (a distant field star much bluer than VB 10 and 1.5 magnitudes dimmer in B band). MEarth will observe VB10 until late November 2011 and again after mid-February 2012 with 5-6 millimag precision. If VB10 has planets, they could produce lensing signatures that enhance the detectability of the stellar-lens event and/or produce distinct planet-lens signatures.

2.2.5 Challenges of the Dynamic Radio Sky

J. Lazio addressed the challenges posed by the new generation of radio surveys. He focused on two particular areas: fundamental physics from pulsar observations and pulsar surveys, including their use as gravitational wave "detectors" through precision timing, and an booming new field of radio transients.

Radio pulsars are a class of neutron stars that emit radio pulses at periodic intervals, typically with pulse periods of order 1 s, in some cases approaching 1 ms. Because of their large moments of inertia and high rotation rates, they can serve as "clocks", with the arrival times of pulses timed to extremely high precision, in some cases better than 100 ns. The precision with which pulses can be timed has enabled a suite of fundamental physics investigations in two broad classes, studies of the theory of gravity and studies of the nuclear equation of state. Observations of pulsars have resulted in two Nobel Prize in Physics (1974, A. Hewish, and 1993, J. Taylor and R. Hulse). Of the current census of ~ 2,000 pulsars, only ~ 10% are stable enough for the precision timing, and only ~ 1% are really good. Thus, there is a need to find more of them, in order to improve the fundamental measurements. We do miss ~ 90% of the total estimated population of pulsars in the Galaxy that could be as large as 20,000.

The first indirect detection of gravitational waves was obtained through the timing analysis of a binary pulsar, resulting in a Nobel prize. More such systems can complement direct detection studies such as those with LIGO and related instruments. In addition, networks of precisely timed pulsars on the sky can be used to search for very long wavelength gravitational waves that are outside the range that can be probed by terrestrial observations.

Finding new pulsars involves a search of at least a four-dimensional parameter space:

- Position on the sky, (α, δ). Consider a search of some area of the sky Ω visible to a radio telescope, which has some resolution element or pixel θ. The area to be searched can be as large as the entire sky. The number of positions to be searched is then Ω/θ. Current radio telescopes used for pulsar searching might have resolution elements of order 10 arcmin. Thus, a search of a significant fraction of the sky accessible to a radio telescope, e.g., 2π sr or about 20,000 deg^2 involves a search of about 10^6 positions on the sky; next generation

telescopes may have higher resolutions, e.g., 1 arcmin angular resolution, implying a factor of 10^2 increase in the number of pixels to be searched.

- Pulsars are faint objects, motivating relatively long observations Δt of each pixel on the sky in order to increase the signal-to-noise ratio. However, in order to detect millisecond pulsars with the shortest periods, fast sampling in time δt must be used. As illustrative values, a typical observation might be $\Delta t \approx 20$ min. with a time sampling $\delta t \approx 70\mu s$, implying time series in excess of 10^7 samples.
- Pulsars are broadband objects and large instrumental bandwidths $\Delta\nu$ are used in order to increase the signal-to-noise ratio. However, the radio pulses travel through the interstellar plasma, which has a dispersive effect. (Higher frequencies arrive sooner than lower frequencies.) There are "de-dispersion" techniques that can mitigate much of these propagation effects, but they require relatively high spectral resolution $\delta\nu$. Currently, typical values are $\Delta\nu \approx 500$ MHz and $\delta\nu \approx 20$ kHz, implying spectra in excess of 10^4 points.

In some cases, it is possible to reduce the number of parameters to be searched. For example, objects detected at other wavelengths (e.g., gamma rays) can be searched for periodic pulsations. In these cases, the position of the object is known. However, a non-negligible fraction of pulsars are found in binaries. A pulsar in a binary is accelerated, which has the effect of changing the apparent pulse period and requiring additional parameters to be used in the search.

The typical processing approach begins with a two dimensional matrix for a particular position in the sky. The matrix or *dynamic spectrum* is of order 10^4 x 10^7 points representing radio power as a function of frequency and time, respectively. The interstellar dispersion causes a pulsed signal to be "chirped". The magnitude of the chirp can be described by the *dispersion measure* DM, which is effectively the total electron column density between the Earth and the pulsar. A set of trial DMs are used to compensate for possible frequency chirps; the number of trial DMs used depends upon the details of the actual observation, but can be several thousand. For each trial DM, the dynamic spectrum is de-dispersed and summed over frequency to produce a time series. The time series is then Fourier transformed to search for statistically significant peaks in the power spectrum. Because pulsar pulses are not perfectly sinusoidal, their power within the power spectrum can be distributed over many harmonics. Typically, a small set of harmonics are also summed, for a range of possible periods. Finally, if one also attempts to find pulsars in binary orbits, accelerations or frequency drifts must also be searched.

This is a huge computational task, not unlike the search for GW signals. Better algorithms are needed in order to make it practical.

A new generation of radio surveys has revealed a number of highly variable or transient radio sources. Some classes of radio transients have been known for decades, while others are only recently discovered (e.g., Berger et al. 2001; Hyman et al. 2005; McLaughlin et al. 2006; Hallinan et al. 2007). Further, there are a host of hypothesized classes of sources, based on such considerations such as extrapolations from behaviors of sources at other wavelengths, analogs of known sources, or simply extrapolations of known physics. Examples include prompt radio emission from gamma-ray bursts and radio emission from extrasolar planets. At radio wavelengths, transients can be divided into two broad classes:

Coherent or "fast": These are produced by particles radiating in phase, e.g., in non-thermal sources, such as the synchrotron emission. Fast transients can produce intense emission, with a luminosity from N particles expected to scale as $L \propto N^2$. They are typically observed at lower

frequencies or longer wavelengths as the radiating volume should grow as λ^3, for an observing wavelength λ. Typical pulse or flare durations are less, potentially much less, than 1 s, and de-dispersion due to interstellar propagation effects is generally required. Pulsars are the examplar of this class. Searches for fast transients typically use processing approaches nearly identical to that for pulsars, except that periodicity and acceleration searches are not important.

Incoherent or "slow": These are produced by independently radiating particles, e.g., in thermal sources. Due to optical depth effects, these typically are brighter at higher frequencies, and propagation effects are not generally important. The typical pulse or duration of the transient is longer than, potentially much longer than, 1 s. Finding slow transients typically involves the formation of light curves, tracking the flux density of sources as a function of time, recognizing that a new source in an image may have appeared, and classifying light curves. While the time scale for slow transients may be substantial, days to weeks or even months, the number of sources in a region of the sky being monitored can be substantial, requiring efficient algorithms to identify sources, recognize new sources, and classify light curves.

The identification of radio transients – or the astrophysical systems from which they originate – on other wavelengths may be essential for their classification. That may be difficult for the fast transients, unless they originate from a source that can be reliably identified in some other way, e.g., a pulsar. Slow transients allow a practical follow-up on other wavelengths; an example of those may be the GRB afterglows. In any case, radio data alone are probably not going to be sufficient for a proper exploration of the transient radio sky.

2.2.6 Statistics and Machine Learning Approaches

G. Longo provided a general overview of the data mining (DM) needs in astronomy. K. Wagstaff then covered three machine learning (ML) techniques that hold promise for analyzing astronomical data.

First, cost-sensitive learning methods can adapt system behavior to accommodate known costs appropriately. Here, "cost" could mean a variety of things: computational cost, misclassification cost, the opportunity cost of delay, or the cost of acquiring additional features or information. Computational cost is an obvious one, since everyone prefers to get results sooner rather than later, all other things being equal. A cost-sensitive learning method might therefore prefer simpler models over the more complex ones, if they can output results sooner. However, it is not always (if ever) the case that "all other things are equal," and in many cases one must trade solution quality against speed of response. Sometimes the desire for more efficient methods is a hard requirement rather than just a preference, as when dealing with massive data sets or streaming, online settings in which slower methods are simply infeasible.

For astronomy applications, especially in online systems that are analyzing data in real time, accommodating other kinds of costs is also important. Misclassification costs may not be the same for all kinds of errors. In most cases, the cost of missing an interesting detection is much higher than falsely classifying noise as signal. Real-time alerting systems must accommodate delay cost, since waiting to observe more of a particular event as it is happening before issuing an alert might preclude any useful follow-up by other assets, yet too many false alerts could saturate available resources and decrease trust in the alerting system. Finally, while better classification results can often be obtained by acquiring or computing more features, there is usually a cost (computational, financial, delay, etc.) involved in obtaining them, and again there

is a decision-theoretic question of when to get more information and when to make a conclusion.

The most basic cost-sensitive machine learning method is the Cost-Sensitive Decision Tree (CSDT; Ling et al. 2004). The CSDT modifies the standard decision objective function (information gain) to both minimize errors and incorporate the cost of feature acquisition (addressing the final cost mentioned above). For example, when classifying Parkes radio telescope observations, specifying a high de-dispersion (feature acquisition) cost leads to a shallow tree with fewer node tests, while specifying a high misclassification cost creates deeper trees. However, the CSDT only factors in feature acquisition cost when classifying new items; it does not try to minimize the cost while training or building the tree. For very large data sets or archives, this consideration is also important. A general Confidence-Based Feature Acquisition (CFA) method developed by the JPL group uses a cascade ensemble to minimize the cost of both training and testing (classifying new items) (desJardins et al. 2010). The ensemble can consist of any type of classifier, not just decision trees. The only requirement is that the base classifiers produce a posterior probability or confidence in their classifications. The system trains a series of classifiers, each with an increasing number of features (and therefore increased feature acquisition cost), but only uses the cheapest subset needed to classify a new observation with sufficient confidence. Another relevant development is that of "reliable" or abstaining classifiers that produce an output classification only if they are sufficiently confident in that decision (Vanderlooy et al. 2009).

The second area of machine learning featured in this talk was that of collaborative analysis. This is a general class of methods that leverages observations from multiple vantage points, or views, to make a more robust collective decision. The JPL group have previously investigated this approach using observations from the Mount Erebus Volcano Observatory, a network of seismic sensors. They developed a related method for radio astronomy and deployed it at the Very Long Baseline Array (VLBA) where it has been operational as part of the V-FASTR system since July 2011 (Thompson et al. 2011ab, 2012). Using observations from multiple radio telescopes spread out across the U.S., we estimate an ensemble CDF to describe "normal" data, which allows for robust detection of any deviations from normal behavior. Further, since this CDF is data-driven and updated every second, it adapts to the current noise environment automatically.

Finally, anomaly detection is a machine learning method of great use for a variety of astronomy investigations. Some anomalies (such as RFI in radio astronomy) are to be ignored, while others (such as optical or radio transients) may provide new scientific insights. We have developed two methods that do anomaly, or novelty, detection in an intelligent fashion, by learning from user input. The first method is Semi-Supervised Eigenbasis Novelty Detection (SSEND; Thompson et al. 2011ab, 2012). It is designed for novelty detection when analyzing streaming data, such as that coming from a radio telescope. Here, the user is able to specify up front some examples of anomalous but scientifically uninteresting observations (like RFI). The system incorporates those examples into its model of the uninteresting background, then builds an adaptive model of the streaming data as it is observed. Combining these two models allows for the detection of novel signals while ignoring locally anomalous but known uninteresting signals like RFI.

The second method is Discovery through Eigenbasis Modeling of Uninteresting Data (DEMUD) (Wagstaff et al., submitted). This interactive discovery method is aimed instead at the analysis of large archives of data, those that are too voluminous to permit manual review of every element. DEMUD first ranks all observations with a generic PCA-based analysis, then presents the most anomalous item to the user for feedback. If the item is deemed interesting, DEMUD

moves on to the next item. However, if the item is uninteresting, DEMUD incorporates it into its model of "what to ignore" and re-ranks the remaining items. In this way, DEMUD quickly adapts to user priorities. Its efficacy was shown in finding items of interest such as exoplanets in Kepler time series and magnesite (a carbonate) in CRISM hyperspectral data. DEMUD is most powerful for settings in which the items of interest are rare.

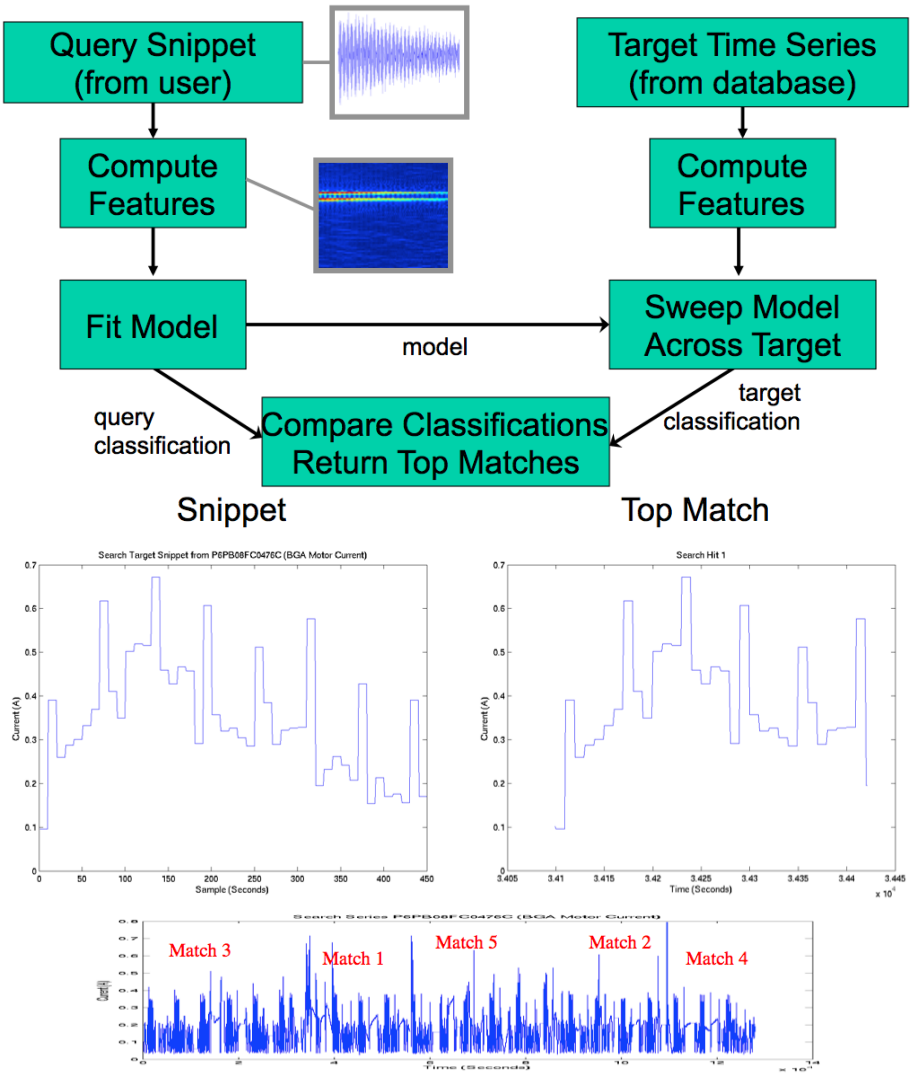

Figure 6. An example of a HMM-based classifier, applied on the GPS and seismograph data (Turmon et al.). Snippets of the time series signals are processed as to represented as feature vectors, which are then compared to a set of models for different underlying phenomena. The system then returns the top matches across the entire input time series. This approach could, in principle, be applied to astronomical data, such as the light curves, GW signals, etc.

M. Turmon and D. Thompson described applications of Hidden Markov Models (HMM) in the analysis of astronomical and space data. They are a natural generalization of clustering algorithms from the standard ML tools. A HMM is a statistical Markov model used to represent

systems that behave like processes with hidden (i.e., unobserved) data, and that can be seen as a generalization of a mixture model where the hidden variables are not independent, but related through a stochastic Markov process. This simulates a situation where there is some underlying dynamic system running along according to simple and uncertain dynamics, but we can't see it; all that we can see are some noisy signals arising from the underlying system. For example, we may be observing some transient source, e.g., a supernova, by sampling its underlying light curve at different times. From those noisy observations we want to predict the most likely underlying system state, or the time history of states, or the likelihood of the next observation. Since HMM can deal with noisy, irregularly spaced data, and yet giving a computationally efficient representation of some temporal relationships, they may be very useful algorithms to tackle the problems addressed in this study. A schematic outline of an application of a HMM based classifier is shown in Fig. 6.

Finally, two postdoc talks described some specific applications of ML and DM in astronomical classification problems of interest: C. Donalek discussed several approaches to inclusion of external (a priori, or contextual) knowledge in the classification process, both in the time domain, and image domain, and described an experimental crowdsourcing project, http://skydiscovery.org , designed to harvest human pattern recognition skills for elimination of artifacts and classification of transients in synoptic sky surveys. R. D'Abrusco described several unsupervised classification/clustering methods as applied to discovery of quasars in multi-wavelength data sets from sky surveys.

2.2.7 Other Scientific Use Cases

A number of other scientific cases and methods have been discussed through shorter presentations. J. Scargle elaborated on the representation of time series, and J. Rice explained the concept of Hierarchical Resolution (e.g., Meinshausen et al. 2009). B. Moghaddam reviewed some methods of biostatistics, with an emphasis on Bayesian techniques, and their potential applications in astronomy, in the context of this workshop.

J. Babu addressed the issues of faint signal detection in astronomical data cubes, such as those expected from ALMA, and other radio astronomical facilities, both present and future, leading to SKA. In an earlier work, Babu, Mahabal, and collaborators explored the use of a new metric for transient detection, referred to as the *Mahalonobis* distance. This is just the square-root of the χ^2 metric that includes covariance estimates between the measurements:

$$D_M = \sqrt{[(\boldsymbol{P_j} - \boldsymbol{M_j})^T \Omega^{-1} (\boldsymbol{P_j} - \boldsymbol{M_j})]},$$

where $\boldsymbol{P_j}$, $\boldsymbol{M_j}$ are the data and mean column vectors respectively and Ω is the error-covariance matrix. The details of the method are given in Babu et al. (2006). Fig. 7 illustrates the application of the method for the detection of faint or moving (transient) sources and subtle image artifacts, using the data from the PQ survey.

Two student presentations were especially noteworthy:

W. Max-Moerbeck described a project aimed to constrain the location of the gamma-ray emission site in blazars by the monitoring and cross-correlation of large number of sources in the radio and γ-ray bands. The observations show large variability in both bands, with features that are sometimes identified as flares, in which the flux of the source increases by a large factor with respect to what seems to be a base level. The variability can be modeled as a noise process with a

power-law power spectral density. These processes show flare like features, so it is not unusual to find that by correlating two unrelated time series a large value of the correlation coefficient is found. In order to distinguish these chance coincidences from real ones, Monte Carlo simulations are used to estimate the probability of occurrence of chance coincidences. Simulating a large number of unrelated time series with a given power spectral density model significance of the possible cross-correlations can be evaluated.

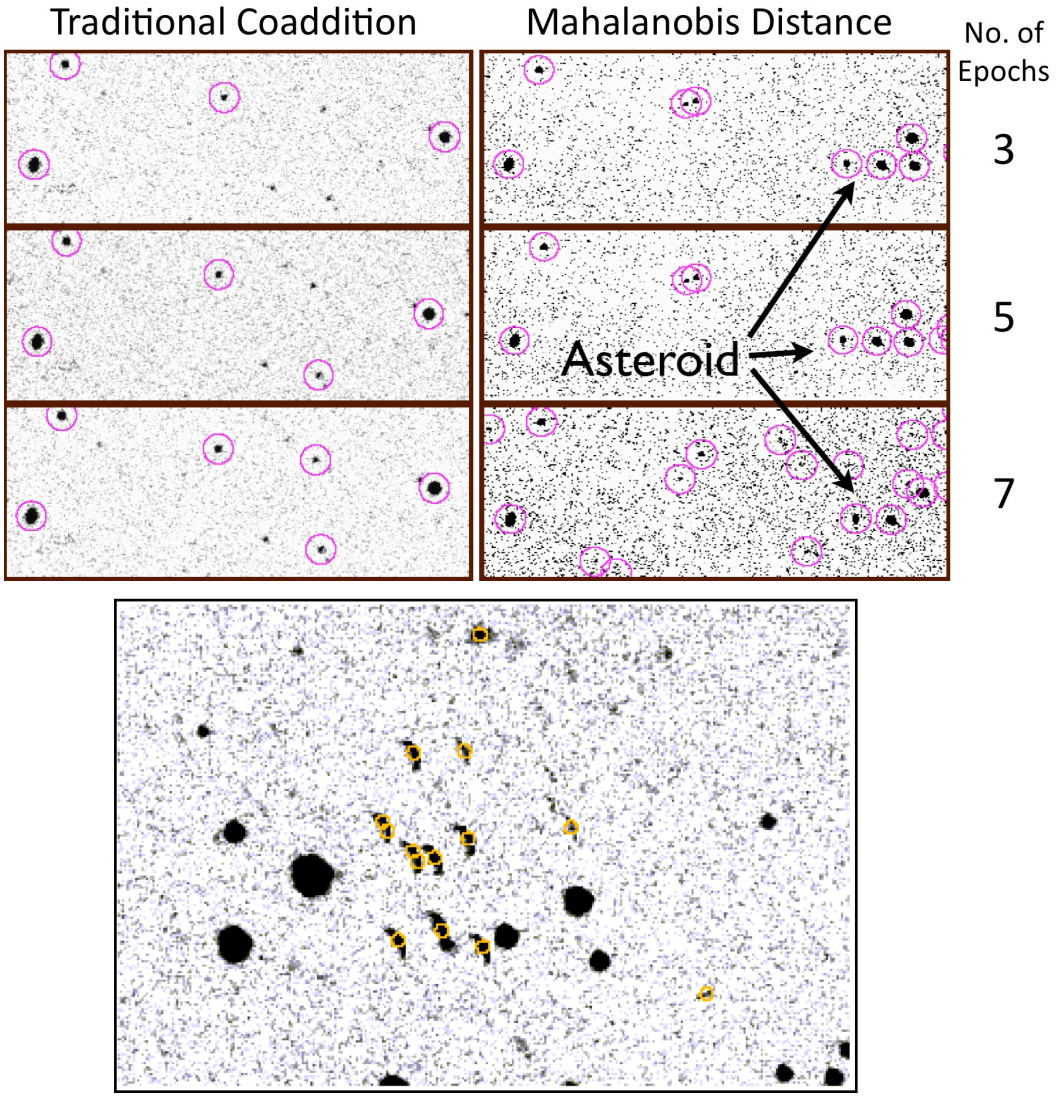

Figure 7. *Top:* Faint source detection using Mahalanobis distance, applied to multi-epoch images of a small portion of the sky from the PQ sky survey. The three panels show co-adds from 3 (top), 5 (middle) and 7 (bottom) epochs. The images on the left are formed by a traditional pixel-by-pixel averaging method with 4-σ sources circled. The images on the right are formed by the statistic based on Mahalanobis distance with significant sources circled. This technique can detect more sources near the detection threshold. The "new" objects in the right-hand panels show a passage of an asteroid through the field. *Right:* This method can also be used to detect and flag faint artifacts (crescent-like structures) in this image coadd from the CRTS survey. Once identified, the artifacts can be used to form training samples for an automated removal using a supervised classification method, such as ANN (Donalek et al. 2008).

G. Cabrera presented the results of a new astronomical object detection and deblending algorithm when applied to Sloan Digital Sky Survey data. Our algorithm fits PSF-convolved Sérsic profiles to elliptical isophotes of source candidates. The main advantage of our method is that it minimizes the amount and complexity of real-time user input relative to many commonly used source detection algorithms. Our results are compared with 1D radial profile Sérsic fits. Our long-term goal is to use these techniques in a mixture-model environment to leverage the speed and advantages of machine learning. This approach will have a great impact when re-processing large data-sets and data-streams from next generation telescopes, such as the LSST.

Overall, these presentations and discussions from the opening workshop laid out a rich scientific territory, touching on a broad range of astrophysical issues, and identified some key, common challenges that must be addressed through a development of better and faster algorithms.

2.3 Working Groups and Identified Research Topics

The outcome of these presentations and accompanying discussions was that the goals of the study evolved in scope and focus. Three distinct challenges were identified:

1. Searching for Long, Weak Gravitational Wave Chirps and for Microlensing Events:

At the first workshop, a small group ("Group 1") was formed to work on problems of very weak, deterministic signals buried in noise, with C. Cutler as the official group leader. The group's main interests fell into roughly two classes, and so we decided to focus on two distinct problems. The first was to develop improved methods for detecting weak, long-lasting, but extremely simple chirps in Gaussian noise. Our extremely simple chirps were simple sinusoids with monotonically varying frequency, parametrizable by just: an amplitude and overall phase, the signal frequency at some instant, the first few time-derivatives of the frequency at that instant. That is, our signal model was simply:

$$s(t) = A\cos\phi(t)\,, \text{ with } \phi(t) = \phi_0 + \sum_{k=1}^{k_{\max}} \frac{2\pi}{k!} f^{(k)}(t - t_0)^k$$

with $k_{\max} = 3$. While this signal model is somewhat simpler than the physical ones of most interest, this problem still exhibits the main difficulties, and so we considered it to be a good prototype. I.e., it seemed hard to imagine a search idea/algorithm for this signal-type that would not also be useful for the more complicated problems of greatest astrophysical interest.

The second problem was to develop more efficient methods of identifying microlensing events in which the lensing profile reveals a planet in orbit around the lensing star. The planet (when observable) typically contributes a narrow spike in amplitude modulation, corresponding to the moment when the planet passes near to the light rays that have been “bent” towards the observer’s telescope by the lensing star. The class of exoplanet-lensed light curves is in practice too large for the data analyst to search through all of them directly, especially since the accurate computation of these light curves is numerically costly, so there is a substantial motivation to develop better search methods. Also, while this problem is important in itself, we also regarded it as a first step towards developing efficient microlensing searches for lensing systems with multiple planets.

2. ***Intermittent, Sub-Significant Detections:*** Consider a scenario where a variable or transient source appears intermittently in a synoptic sky survey, just below the significance cutoff in *some* of the exposures, but is undetected in a majority of others. Averaging of all such individually sub-significant exposures would yield a statistically significant "global" detection, but averaging of all exposures would dilute the signal below the threshold, since noise, but not the signal, is added from the bulk of the exposures. Generally, the variability should be considered to be stochastic, with periodic and transient (e.g., a faint supernova) being the special cases (it is of course not known a priori in which exposures the source may be found). There are two sub-scenario: (a) where the position of the putative source is known for some reason (e.g., it was detected on some other wavelength), which makes the problem described as a time sequence of flux measurements in a particular beam; and (b) if it is not known a priori that a source may be present, which can be described as a set of the cases (a), over a densely sampled grid of possible beam centers (source positions). Another way of thinking about it is that a source appears along a particular temporal line in a data cube consisting of all available, co-aligned images of the field. An algorithm for detection of such intermittent sources could greatly increase the effective depth of synoptic sky surveys.
3. ***Classification of Variable and Transient Sources:*** Variable sources and transient events detected in synoptic sky surveys can be caused by a large number of different phenomena, some of which may be more interesting than others. Their potential scientific value lies in determining their (likely) physical nature, and may require follow-up observations, e.g., spectroscopy, additional photometry, etc. Moreover, given their ephemeral nature, and the fact that follow-up resources are generally costly or limited, there is a premium in deciding as quickly and reliably as possible whether any given event justifies the use of such resources. *Thus, the effective scientific depth of synoptic sky surveys is not determined by their flux detection limit, but rather by their classification depth.* Physical classification or characterization of transient events can also inform the detection process itself: a detection algorithm may be optimized for a particular type of variability, e.g., for supernovae, or for periodic variables, etc. Transient event detection and classification are intertwined, and both determine the scientific returns from a given synoptic sky survey. Here we also distinguish two regimes: (a) non-time-critical, where a reasonably sampled light curve with tens or hundreds of measurements exists; and (b) time-critical, where at first only a few flux measurements are available, and the follow-up decision needs to be made on the basis of a preliminary classification (which will evolve dynamically, as more data come in).

The participants divided themselves into three working groups (with some overlaps) that focused on these particular challenges. The results of their deliberations represent the major outcome of this study, and are described below.

3. Studies of the Three Key Problems

The initial discussions took place during the initial study week (June 7 – 10), continued through group interactions (in person, by email, and various forms of telepresence) in the months between the opening and the closing workshops, and even continued in the months following the formal end of the KISS study.

Some of the incidental files, papers, and test data sets have been posted on https://kisscaltech-digging.pbworks.com/w/page/37168614/KISS%20Digging%20Deeper%20Wiki%20-%20Home.

3.1 Searching for Long, Weak Gravitational Wave Chirps and for Microlensing Events

As mentioned above, Rice proposed to start by tackling the simple question of whether it can ever be advantageous to leave some of the data unanalyzed, when computational constraints are a limiting factor. He posted a note on the KISS wiki with a very simple argument why analyzing all the data is always better, at least for a one-stage search. However he assumed some approximate scaling relations that Cutler believes are overly restrictive; i.e., they do not always hold for realistic searches. Coincidentally, around the same time, a paper appeared by Prix & Shaltev (2012) which examines this very question in depth and claims that in some cases it is better to start with all the data, and in some cases it is not. Cutler suspects that the Prix–Shaltev analysis is right, and that when Rice's scaling assumptions are generalized, Rice's formulation of the argument will give basically the same result, but in a much more clear, direct, and intuitive manner.

As also mentioned above, Cutler pursued several ideas for improving the sensitivity of searches for weak chirps. We now describe those ideas, and the progress made on them:

(i) The first idea was motivated by the observation that current searches are optimized using a local metric on the parameter space to determine the spacing of template grid points. The local metric encodes the falloff in the overlap between a template signal and the true signal, as the former is moved away from the latter. The metric is a purely local construct, which encodes how the overlap is affected for very small parameter errors. However, once the parameter errors become large the falloff in overlap is actually far milder than one would guess from the metric alone, and far milder than most workers in the field seem to have realized. For instance, consider just the two parameters $(f_0,\dot{f}_0)$, and consider how the overlap function (between true signal and template) scales with the errors $(\Delta f_0\,,\Delta\dot{f}_0)$. For very small errors, the overlap must take the form $1-a(\Delta f_0)^2-b(\Delta\dot{f}_0)^2-c(\Delta f_0)(\Delta\dot{f}_0)$. For moderate-sized errors, it is easy to show that the overlap falls off only like $(\Delta\dot{f}_0)^{-1/2}$ [for fixed Δf_0] and is actually approximately independent of Δf_0 [for fixed $\Delta\dot{f}_0$]. This suggests that our current "optimized" searches, which incorporate only the metric information, probably use grids that are too fine. That is because if one makes the grid spacing significantly larger (to save on computational cost), the decrease in sensitivity is much less than estimated using the standard approximations.

Cutler tried for a few months to turn this observation into an improved algorithm, but it eventually became clear that the path he was on was just leading back to the algorithms and statistics that we currently have. (An early estimate by Cutler showing that the result had to be different turned out to be an error, resulting from employing an approximate relation outside its domain of validity.)

(ii) Cutler's second new "trick" was to use much more of the information generated in semi-coherent searches, as follows. Given N coherent segments of length ΔT, the data analysis pipeline returns N complex amplitudes A_i, and then forms the semi-coherent detection

statistic $\rho^2_{semi} = \sum_i |A_i|^2$. But this statistic clearly "throws away" all the phase information of the complex amplitudes. These phases follow a pattern that provides good estimates of the parameter errors $(\Delta f_0 , \Delta \dot{f}_0)$. With these estimates, one can construct an improved detection statistic that turns out to be rather close to the fully coherent detection statistic. Of course, this scheme works only when there actually *is* a signal of detectable amplitude embedded in noise, and when one is sitting at a point on the semi-coherent grid that is close to the actual parameter values. Thus this trick is probably best suited for use as a "follow up" step in a hierarchical search, to be applied at grid points that have already yielded an anomalously high value of ρ^2_{semi} .

Cutler has been working on this with Vallisneri. They have derived a method for quickly estimating $(\Delta f_0 , \Delta \dot{f}_0)$ from the A_i, which turns out to work very well when the (matched filtering) signal-to-noise ratio (snr) per segment is ~3 or higher, but becomes quite unreliable when the snr-per-segment is less than 2. The regime of greatest interest is the one where snr-per-segment is ~1, so this appears less promising than Cutler and Vallisneri had hoped for. However there are many variants on this idea which still appear promising, and the numerical tools that they have already developed should allow them to evaluate those ideas much more quickly than for their "first try". They have a list of ideas to explore via numerical simulation, and have started those investigations.

(iii) Current semi-coherent searches divide the data in segments that all have the same duration ΔT. But this choice seems to be motivated more by simplicity of implementation than maximization of sensitivity. This year, Cutler devised an argument for why ΔT should *not* be the same for every segment. A brief version goes as follows. Imagine that initially we allow all the lengths ΔT_i to be independent, and then vary them all to maximize search sensitivity at fixed computing cost. Using the method of Lagrange multipliers, it is easy to show that the optimum values satisfy

$$\frac{\partial(Sensitivity)/\partial(\Delta T_i)}{\partial(Cost)/\partial(\Delta T)_i} = const \ (independent \ of \ i) .$$

It is also easy to see that if ΔT = *const*, then the above ratio is *not* constant, but instead decreases moving outwards from the center of the total observation time. So a better choice would be to have ΔT larger towards the center and decreasing towards both ends of the observing period. Cutler plans to quantify the gains achievable by varying ΔT in this way.

Once Cutler and Vallisneri know which of the new methods have value (based on simulations of their sensitivity and of the cost of implementing them), our goal will be to incorporate the ideas optimally into a search pipeline, using the sort of optimization scheme illustrated in Cutler, Gholami & Krishnan (2005). Unfortunately, since the various optimizations are all coupled, until the final optimization is done, it will be difficult to know how much these improvements will "buy us" in terms of increased science payoff. But the fact that none of these has even been considered before illustrates how little thought has gone into the current schemes. This in itself suggests that there could be a lot of low-hanging fruit in this field, and that our new efforts could yield a substantial payoff.

Finally, regarding work on microlensing searches for exoplanets, despite several interesting

discussions and ambitious planning at the beginning, Group 1 ended up barely making a dent in its task plan, due mainly to the usual press of other obligations on people's time.

In summary, the results from Group 1 were frankly rather modest—partly due to the small size of the group and the diversity of interests within it, and partly due to some promising-looking ideas not panning out. However other promising Group 1 ideas, developed under the auspices of this Study, are still being explored, and we still expect that one or more of them will "pay off" in the not-too-distant future.

3.2 Intermittent, Sub-Significant Detections

Much of this effort was led by F. Masci and A. Mahabal. In this study we focused on optimizing transient detection at low to moderately low S/N levels across multiple single-exposure observations. We have explored some optimal image-combination metrics in the maximum-likelihood sense according to the noise-distribution followed by the input measurements. Our method is optimized for optical/IR data where the underlying photon-noise is well into the Gaussian limit (setting aside systematic error sources). An important question is: how low a S/N can we reach in a series of single-epoch observations to ensure a moderately high significance in our combined metric-image space? We address this using Monte-Carlo simulations. Our focus is *reliable* identification of faint, low S/N transient candidates. We have implemented our method in a prototype software tool "*imtrandetect*" which we briefly describe below.

The problem of detecting faint (usually low-significant) events in single epoch observations entails devising a statistic that combines information from multiple consecutive epochs in time that we can test for significant excursion above the null hypothesis (H0) of pure noise fluctuations. The difficulty is finding a statistic which is most sensitive to repeated (systematic) behavior in a series of measurements (different from that expected by the underlying noise) that may suggest a faint transient. This assumes the underlying noise (including any correlated behavior over time) is well characterized beforehand.

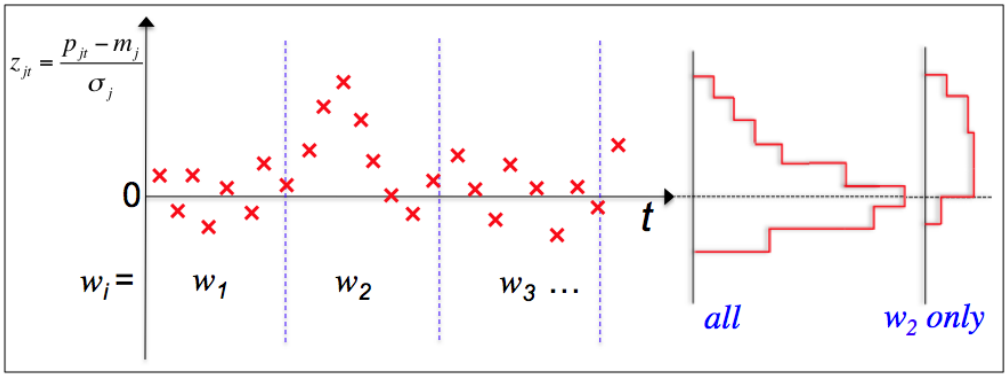

Figure 8. Schematic of windowing scheme for a series of noisy pixel measurements (relative to some baseline m_j and normalized by the long run noise-sigma, σ_j, over time on a sky location j). A hypothetical transient appears above the noise in window w_2. On the far right are two marginal distributions (collapsed along the time axis) formed by *all* measurements and only those in window w_2.

Figure 8 shows a schematic of a transient whose peak signal is shown to be relatively strong for clarity and the purposes of this discussion. The measurements may be that of a single pixel j through a stack of registered, time-ordered images, which we represent as z-scores, i.e., the

number of sigma above the pixel's 'long-run' baseline level, m_j. The sigma value here (σ_j) is characteristic of the pixel over time (see below for estimation methods). Figure 8 also shows marginal distributions by collapsing the time axis and binning the *z*-scores. One can see that if the time series is windowed in time, the distribution of measurements for the window containing a suspect transient will be more skewed or rather, have a relatively greater fraction of values with large excursions from the underlying baseline level above the *long-run* noise (σ_j) than if the distribution from all measurements (i.e., from all windows) is used.

Windowing a series (with some optimum window-length; see below) therefore reduces dilution from the underlying noise. For relatively faint short-lived transients (compared to the available history of measurements on a sky location) the baseline noise will dominate if many 'null' measurements over a long time-span were combined, hence rendering reliable detection difficult. Therefore, windowing increases our chances of detecting faint transients on *local* time-scales if one has a good handle on the historical noise at a given sky location. This is an improvement over traditional single-epoch image differencing (which is a point-wise process in time) since by combining multiple consecutive epochs (assuming they are relatively closely separated in time; i.e., to provide good sampling of the target transients), will increase the detection S/N. This is the crux of our method. We expand on the details and limitations below.

We have constructed several image-combination metrics for "collapsing" a series of time-ordered pixel measurements (*z*-scores) from a set of sky-registered images. These "metric-images" are generated for each window along the time-sequence. At first, we experimented with four metrics (per pixel stack in a window): (i) the maximum pixel *z*-score; (ii) the fractional excess of *z*-score values above some threshold relative to that expected from noise alone (e.g., a Gaussian distribution); (iii) the classic reduced chi-square; and (iv) the third central moment (which we refer to as the *skew* from now on, with symbol *S*). These are respectively defined as follows for a pixel stack *j* in window *i* containing N_{wi} images:

$$z_{ij,\max} = \max\left\{ z_{ijt} \quad \forall\ t = 1,2,3\ldots N_{wi} \right\} \tag{1}$$

$$R_{ij} = \frac{Frac\left(z_{ijt} \geq z_{thres}\right)}{0.5\left[1 - \mathrm{erf}\left(z_{thres}/\sqrt{2}\right)\right]} \tag{2}$$

$$\chi^2_{ij} = \frac{1}{N_{wi} - 1} \sum_{t=1}^{N_{wi}} z^2_{ijt} \tag{3}$$

$$S_{ij} = \frac{\sqrt{N_{wi}\left(N_{wi} - 1\right)}}{N_{wi}\left(N_{wi} - 2\right)} \sum_{t=1}^{N_{wi}} z^3_{ijt} \tag{4}$$

where

$$z_{ijt} = \frac{p_{ijt} - m_j}{\sigma_j} \tag{5}$$

for a pixel signal p_{ijt} falling in window *i*, at sky location *j*, and measured at time *t* that exhibits a *long-run* "static" baseline level m_j and noise-sigma σ_j. The functions of N_{wi} pre-multiplying Eqs. (3) and (4) make these metrics *unbiased* estimators of the respective *normal* population values

since each is based on sample estimates of the location (baseline level) and noise-variance (e.g., Pearson 1931). A χ^2 statistic (similar to Eq. 3) was used by Szalay et al. (1999) for generic source detectio by combining images across multiple passbands. It is used here in a somewhat different context.

Furthermore, a related metric was explored by Babu et al. (2006) for transient detection, referred to as the *Mahalonobis* distance. Here we ignore correlations since our method only combines pixel measurements in the temporal domain where they are expected to be largely independent, i.e., Ω is diagonal, while Babu et al. also combine measurements in the spatial domain which are not necessarily independent.

The metric-images formed by metrics (1) – (4) can then be thresholded to identify transient candidates through use of a matched filter (e.g., that appear PSF-like), or searching for spatially contiguous hi-values above some local spatial-noise threshold. As a detail, one may want to place these metrics on an equal footing for thresholding purposes, e.g., by converting them to probabilities per pixel (i.e., of getting a value larger than that observed by "chance" under a H0 of pure noise). The best approach is to use empirical null probability distributions derived from the data at hand. This will capture the noise structure and properties inherent in the data itself, including systematics, correlated-noise etc. The calibration of null empirical probability distributions is cumbersome, although it need only be done once for the detector/instrument being used. In this initial study, we opted to threshold the metric-image values directly relative to the mean and sigma of a *sample* metric expected under a H0 that measurement errors are distributed as Gaussian. Taking metrics (3) and (4) for instance, we convert these to equivalent *z*-scores that can be thresholded:

$$z(\chi^2_{ij}) = \frac{\chi^2_{ij} - 1}{\sqrt{2/(N_{wi} - 1)}} \tag{6}$$

$$z(S_{ij}) = \frac{S_{ij}}{\sqrt{\sigma^2_{Sij}}} \tag{7}$$

where

$$\sigma^2_{Sij} = \frac{6\left(N_{wi} - 2\right)}{\left(N_{wi} + 1\right)\left(N_{wi} + 3\right)}. \tag{8}$$

Equation (6) follows from the fact that the mean and variance of the reduced χ^2 are 1 and $2/dof$ respectively, where the number of degrees of freedom is $dof = N_{wi} - 1$. Equation (7) uses the fact that the *skew* (third moment) for a Gaussian population of errors is zero, and the expected variance for the *sample skew* as computed using the unbiased estimator in Eq. (4) for Gaussian-distributed errors is given by Eq. (8). This took some effort to verify via simulations, but discussions can be found in Cramer (1946) and Pearson (1931). Furthermore, we find that predictions for the sample mean and variance as used in Eqs (6) and (7) conform very well to empirical (histogram-derived) estimates in real optical/IR image data in noisy background regions, testament that the Gaussian-noise assumption is acceptable (when systematics and instrumental glitches are at bay!).

Experiments on real and simulated data revealed that metrics (1) and (2) did not perform as well as the reduced χ^2 and *skew* in Eqs (3) and (4) respectively. The results showed that the metric

images from Eqs (1) and (2) were very noisy and generated a plethora of false positives when thresholded. The two that looked most promising (in terms of maximizing detection S/N in metric-image space; see §2.2) were the χ^2 and *skew* metrics. For the remainder of this paper, we focus on these last two metrics. Even though related, these metrics reinforce each other in that the *skew* preserves the *sign* of an event (or events), i.e., whether it is a positive or negative excursion relative to the baseline level. Negative excursions are obviously unphysical (e.g., instrumental glitches) and can be immediately flagged as unreliable. The χ^2 however, depends on the square of fluctuations and cannot be used on its own to reject negative excursions.

Earlier we fleetingly mentioned long-run estimates of the underlying baseline-level and noise-sigma per pixel-stack at sky position j (i.e., m_j and σ_j respectively in Eq. 5) for computing z-scores. By "long-run", we mean over the available history of pixel measurements, or a large number of them to properly capture the average temporal behavior of a detector's pixel when collecting real flux from the sky and possibly a *static* astrophysical source. Before using the pixels in a set of images acquired at different times to compute m_j and σ_j, we first stabilize the pixels in each image against possible temporal variations in the sky background by subtracting a local estimate of the background (at low spatial frequencies) from each respective image (details are given in §3). We do not consider possible *uncalibrated* multiplicative effects over time (e.g., changing instrumental throughput, atmospheric transparency, etc.), nor possible changes in the noise properties of a detector (including photon noise). Once the *single-epoch* images have been stabilized against local offset variations (i.e., effectively a de-trending operation), the challenge then is to estimate the m_j and σ_j images as robustly as possible from the global image stack (over all windows). The goal is to be robust against the possible presence of transients that may bias m_j and/or σ_j for a pixel relative to that expected in the steady state, i.e., containing a static or null signal that fluctuates according to the properties of the detector and photon collection process. One may resort to using a well characterized noise model for σ_j, but from experience, we have found such models difficult to tune over the full dynamic flux range of a detector. We have decided to estimate m_j and σ_j directly from the data (with some caveats in mind, see below). We adopted a simple median for m_j and half the difference in 15.85 and 84.13 percentile values in each temporal pixel-stack:

$$\sigma_j = 0.5\left[p_j(84.13\%) - p_j(15.86\%)\right]. \tag{9}$$

This is equivalent to the standard deviation of a Gaussian population. We ignore the convergence properties of sample estimates based on Eq. (9) with respect to unbiased population estimates for now. The important thing is that this is robust and its accuracy improves appreciably as more data is used. In our software implementation (§3), we have an option to globally regularize estimates from Eq. (9) for pixels j that fall on sources whose flux varies significantly on regular (or perhaps irregular) timescales. These sources will inevitably inflate estimates of σ_j. We regularize the σ_j image by computing its *mode* and robust spatial *RMS* over all pixels j and then winsorize (reset) σ_j values exceeding some threshold: *mode* + *n*RMS* to equal this threshold itself. This is still an approximation, but it reduces the incidence of high stack sigmas due to the presence of real astrophysical variables and intermittent transients, bringing down their σ_j, increasing their z-scores (Eq. 5), and increasing their chance of detection in the metric images.

At this stage, some limitations of the above method are worth noting. Many of these have been fleshed out during the course of testing on CRTS data (details are expanded in the software

description in §3). We stress that this method is not a generic tool for detecting all flavors of astrophysical transients and variables. Aside from the limitations imposed by the data (e.g., separation of observation epochs, Earth's atmosphere, instrumental glitches), there are assumptions in our design that severely limit the physical transient phase space. Our methodology is intended to complement other more generic search methods (e.g., single-epoch image differencing), with the goal of extending discovery space.

First, the method is ideally suited to detecting *faint* (possibly intermittent) transients close to the background level, and not continuous variables (that may vary regularly or sporadically). By faint, we mean below some S/N threshold in the median-combined global stack image (m_j). Pixel signals above this threshold (e.g., typically 5 to 7) are masked and excluded from all the windowed metric images (i.e., via Eqns 3 and 4), and do not participate in the transient search. The reason for this is to minimize contamination from detector artifacts associated with bright sources (e.g., diffraction spikes, noisy PSF wings, charge bleeds, etc.). This masking is defined using the global median-combined image since the majority of sources in this image will be static, or more precisely, will have been active for ≥ 50% of the time spanned by the single-epoch images used to compute the median. Therefore, potentially transient or variable sources with a *long-run* median signal exceeding some user-specified S/N threshold (either static or periodically varying with ≥ 50% of its "peak" phases above the threshold) will be missed. This includes bona fide transients superimposed on extended sources with high apparent surface brightness, e.g., supernovae that explode in nearby galaxies. Our design therefore severely sacrifices completeness for reliability, since the latter is of utmost importance at faint, low S/N levels when searching for rare events.

Second, a related issue is the use of a median to estimate the long-run baseline-level per pixel (m_j), which enters in computation of the z-scores (Eq. 5) and eventually the windowed metrics for transient detection. An astrophysical transient must persist for <50% of the entire historical length of the series of single-epoch images under investigation (from which m_j is computed) to stand a chance of detection. A signal persisting for ≥ 50% over the span of all epochs will be pegged to m_j (the median) and treated as static, resulting in null z-score values and metrics. Therefore it is advised that a sufficiently long historical set of observations be used in order to be sensitive to the longest transient timescales of interest, i.e., up to half the historical span.

Tied to the previous point is selection of an optimal window size for computing the metric images. One may get the impression that a specific size will bias against certain types of transients, but this is not the case. Note that the window size is defined as the *number* of single-epoch observations in a partition, regardless of their separation in time, regular or irregular. The frequency of observations obviously determines what types of transients can be detected. The window must be small enough so the metrics are sensitive to the shortest-lived transients of interest, given limitations imposed by the observing frequency. That is, as discussed above, such that dilution from noisy measurements within the window is minimized and the metric S/N is maximized (see Figure 1). However, the window size must be big enough to ensure good statistics are accumulated for the faintest longer-lived transients so the metric S/N is maximized as well. Once a window size is selected, the metrics will then be sensitive to *all* transient timescales exceeding the window size, but < 50% the full history of observations from which the baseline median m_j and noise-sigma σ_j in Eq. (5) were computed. As discussed above, transients persisting longer than this will not be detected since they'll be pegged at the value m_j resulting in a z-score of 0. Tuning of the window size may be done via simulations.

In practice, the windowed, time-collapsed metric-images from either Eq. (3) or (4) (or *z*-score equivalents in Eq. 6 and 7) may be generated from a historical set of images in an archive, or in real time as new observations become available and some minimum number of images is reached within the window to trigger generation of a new metric-image. Another possible caveat is that depending on the observing cadence, window size, and transient time-scale of a source, this process may incur a longer lag-time for alerting that an event has occurred (or *is* occurring) compared to the traditional single-epoch image differencing method.

We stress that the important elements for this method to work optimally are the derivation of unbiased, robust estimates of the baseline level m_j and noise-sigma σ_j to capture the *long-run* steady state behavior of the instrument, including any fluctuations in throughput (and detector gain) that controls the level of photon-noise observed. The more data, the better, but a long enough history of observations must exist to enable these parameters to be determined in the first place. They can be refined as more observations are accumulated.

We have performed Monte Carlo simulations to explore the sensitivity of the χ^2 and *skew* epoch-combination metrics (Eqs 3 and 4 respectively) to the number of times a transient is measured at or above some single-epoch S/N level within a window of noisy observations. We assumed a window containing 15 hypothesized observation epochs, which could be part of a much longer history of observations from which *long-run* estimates of the baseline level m_j and noise-sigma σ_j were derived, as discussed above. Any window length would suffice, with the number of hypothesized transient events scaled accordingly to illustrate our point. We assume uncorrelated Gaussian noise throughout. Figure 9 shows a schematic of two transient signals: one reaches ~ 3σ at three epochs (left) and another reaches ~ 2σ at five epochs (left). The measurements could be of a single pixel or a source integrated over a region. Either scenario in Figure 9 would pose a challenge to the single-epoch image-differencing method, i.e., by differencing against a deeper, higher S/N template image and examining the detections above some threshold. How high a S/N can we achieve by transforming the measurements to a new space formed by combining the epochs according to the χ^2 or *skew* metric? Furthermore, what is the minimum number of times a transient must persist (or be intermittently elevated) above some single-epoch S/N within a window in order to achieve an appreciable S/N for detection in the metric-space?

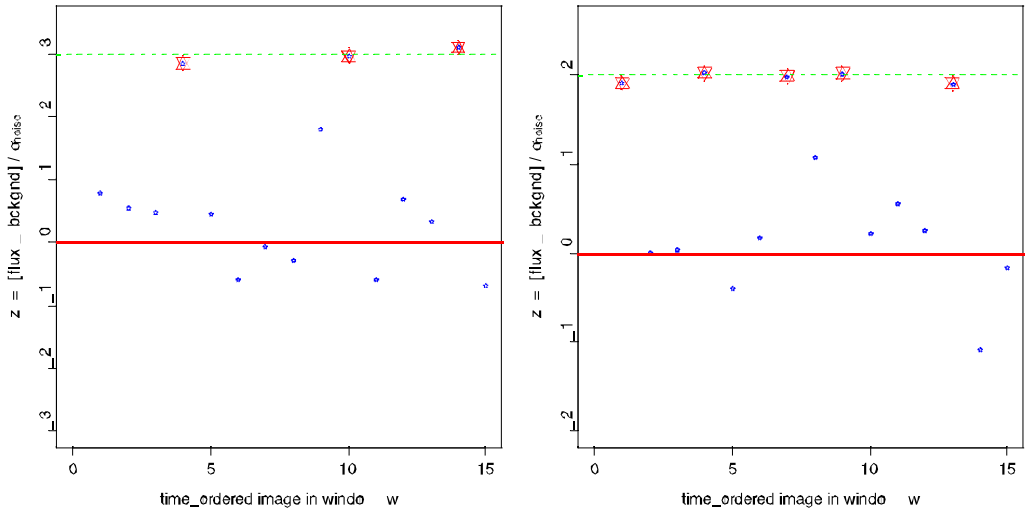

Figure 9. A window containing 15 simulated measurements of a transient where three are at ~ 3σ (*left*) and five are at ~ 2σ (*right*).

Figure 10 shows the results of our simulations for the χ^2 metric (Eq. 3). We considered a transient exhibiting 2, 3, 5, 7, 10, and 15 events (shown labeled) with single-epoch S/N running from 1 to 10 within a window of 15 measurements each affected by Gaussian noise $\varepsilon \sim N(0,1)$. When the measurement is not elevated as an "event" with some S/N, it is assigned a pure noise fluctuation at the zero baseline. We converted the single-epoch S/N and χ^2 values to equivalent probability measures assuming Gaussian statistics, i.e., as the probability of obtaining at least the observed values by chance. For a given number of events at some single-epoch significance, the χ^2 value obtained is actually a random variable, attaining a slightly different value for a new realization of the noise across the 15 measurements. Therefore, we show the mean χ^2 (high-tail) probability values (solid blue lines) and the 10 – 90 percentile ranges (error bars) obtained over 500,000 simulation trials. The single-epoch significance levels are equal to the χ^2 significance levels along the red-dashed line. Overall, the χ^2-combined measurements outperform the single-epoch measurements – effectively what one would obtain from the image-differencing method. One can see that for $N = 3$ single-epoch events hovering at S/N ~ 3 out of 15 measurements, the χ^2-combined measurements can attain a significance (probability of occurring by chance) of $\alpha <\sim 10^{-4}$. Can we do better?

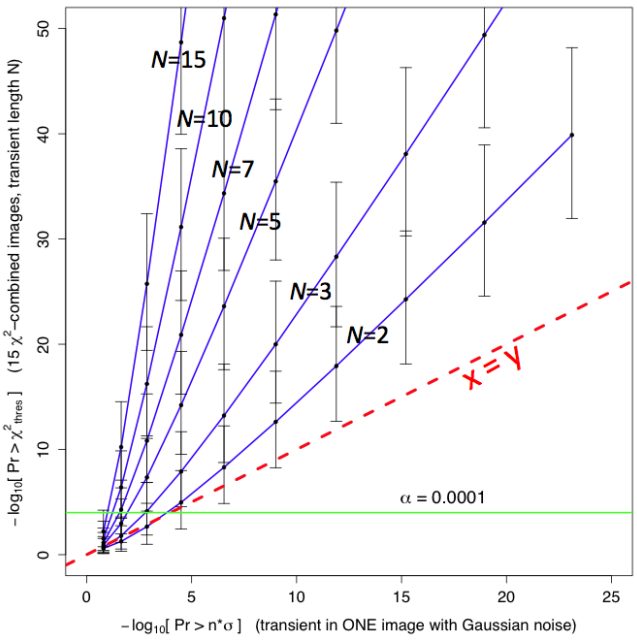

Figure 10. Simulation illustrating the significance (or effective sensitivity) of the χ^2 metric (Eq. 3) represented as the probability of obtaining a χ^2 value larger than that measured by chance if a sequence of 15 images contains a transient exhibiting N events with a single-epoch significance of n (= S/N) $\geq$ 1, 2, 3, … 10 running along the horizontal axis. Blue lines are average χ^2 probability values and the error bars span the 10 – 90 percentile range in probabilities obtained over 500,000 simulation trials for each N and n. The red dashed line is the line of equality.

Figure 11 shows our simulation results for the *skew* metric (Eq. 4) using the same method and inputs as for the χ^2. The only difference is a computational detail in how the probabilities are computed. While the distribution of a χ^2 random variable is well known, the distribution for *skew* when sampling from a *normal* population is not. We resorted to estimating probabilities from analytical fits to distributions for the sample *skew* derived from bootstrap resampling of a *normal*

population. Figure 11 shows that the *skew* metric is more sensitive than the χ^2 metric (Figure 10) at detecting transients for the same range of single-epoch S/N levels and number of events that may occur at these levels. For example, an intermittent transient exhibiting S/N ~ 2 single-epoch events needs to occur on average >~ 5 times out of 15 to give an average *skew*-metric significance of $\alpha <\sim 10^{-8}$. The χ^2 metric will require it to occur >~ 10 times out of 15 to achieve the same level of significance. Pushing the *skew*-metric further, a S/N ~ 1 event will need occur >~ 7 times out of 15 to give an average *skew*-metric significance of $\alpha <\sim 10^{-4}$. This is very encouraging. In general, the lower the single-epoch S/N, the *longer* a transient must persist at >~ S/N (or exhibit more events at or above this threshold) for it to be detected with a high significance in the epoch-combined metric space. Note that there may be other more sensitive metrics. From experimenting on several metrics, we found that the *skew* is the most sensitive at detecting low S/N transient events, presumably due to it's ability to detect slight asymmetries in an appropriately windowed, time-collapsed distribution of measurements relative to some long-run baseline.

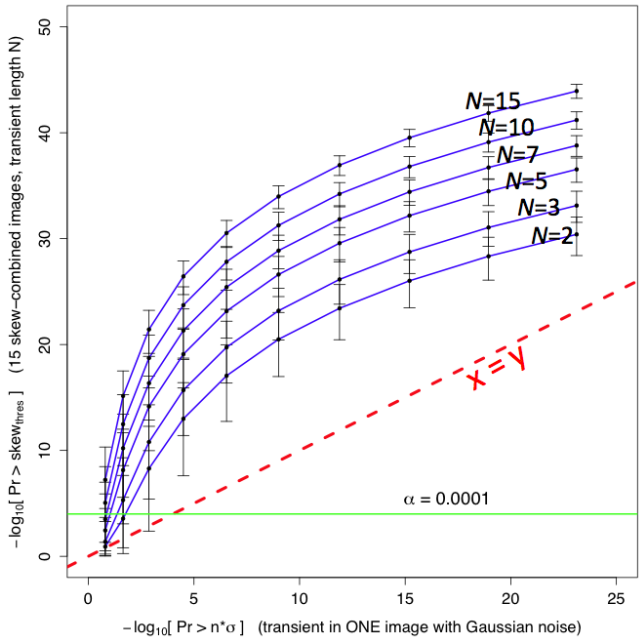

Figure 11. Simulation illustrating the significance (or effective sensitivity) of the *skew* metric (Eq. 4) represented as the probability of obtaining a skew value larger than that measured by chance if a sequence of 15 images contains a transient exhibiting N events with a single-epoch significance of n (= S/N) ≥ 1, 2, 3, … 10 running along the horizontal axis. Blue lines are average *skew* probability values and the error bars span the 10 – 90 percentile range in probabilities obtained over 500,000 simulation trials for each N and n. The red dashed line is the line of equality.

We have implemented the methodology outlined above in a prototype software tool called *imtrandetect*. This tool is still in a developmental, alpha-testing phase, although it implements all crucial elements of the transient search algorithm with a few extra features to assist with reliability. As discussed, we only focus on the χ^2 and *skew* metrics, and the tool is being made flexible enough to run stand-alone on image data acquired in real-time. Details and software usage will be outlined in a future publication. Below we summarize some of the features. The rational for most of the steps was described above, with caveats and limitations outlined above.

The main processing steps in *imtrandetect* are shown in Figure 12. The most CPU-intensive steps are the first two: reprojection and interpolation of the input images onto a common sky grid, and the estimation/subtraction of a local background at low-spatial frequencies to ensure stationary pixel baselines versus time. The first of these may not be needed if the images are from fixed predefined survey fields and the telescope pointing is reasonably accurate. If the software is to be run on an image archive, all steps in Figure 12 are massively parallelizable, with certain steps being triggered as intermediate products become available (e.g., when a window's worth of data has been preprocessed). If processing on a incoming data stream in real-time, one will still have to pre-process a historical subset of archival data in order to obtain initial long-run estimates of the baseline-level and noise-sigma per pixel (last box on the top row of Figure 12). This "calibration" need only be done once, and perhaps refined later. The incoming image-data can then be processed serially as a new window's worth of data becomes available.

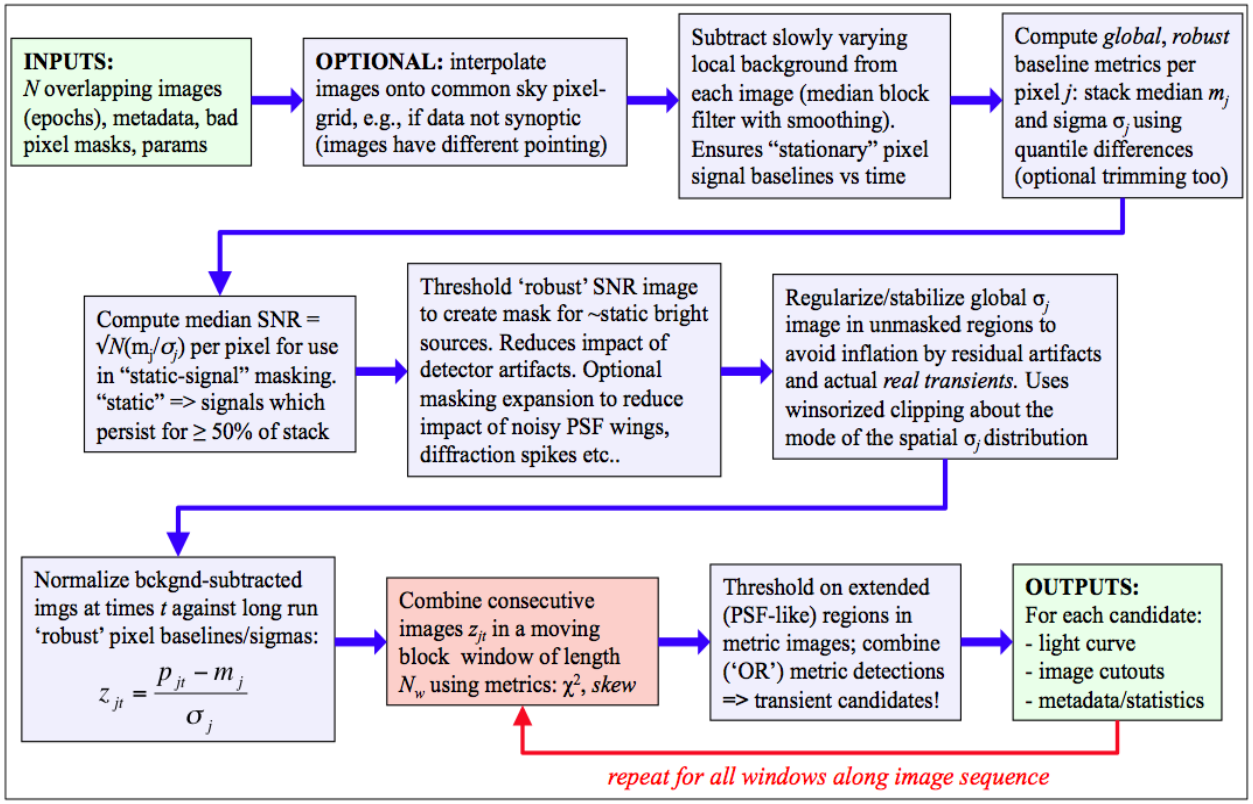

Figure 12. Processing flow in *imtrandetect* version 1.0. Details are expanded in the text.

Some features of the *imtrandetect* tool are as follows:

- Overall, the tool emphasizes masking of instrumental artifacts through use of dynamic image masking of bright "static" sources and their artifacts.
- There is minimal impact from temporal and spatial PSF variations. Hence there are no spurious PSF-related residuals since no image-differencing is involved.
- There is the ability to combine images acquired simultaneously across different filters within a window, in order to further improve S/N.

- It can handle image data with irregularly-spaced observation times and large gaps provided one is aware of the limitations.
- It can handle images with non-uniform overlap (hence spatially-varying depth) across epochs, where it is assumed that images will be to be reprojected and registered prior to use.
- Generates light-curves that are photometrically calibrated if calibration information is available, otherwise internal relative photometry is performed.
- Generates image-cutouts of transient candidates through an image stack over a specifiable time-range, as well as the window-combined metric detection images.
- *Under development*: optional use of priors (e.g., light-curve templates) to assist with reliability, isolating specific transient candidates, or omitting undesired types.
- *Under development*: optional moving object (asteroid) filtering.
- Other constraints to maximize reliability: e.g., require n consecutive (or intermittent) events above some single epoch S/N spanning some Δt.

We are currently testing our methodologies on optical image data acquired from the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009). The primary objective of CRTS is to search for Near-Earth Objects (asteroids), although there are parallel searches for SNe, CVs, Novae, and a wealth of other astrophysical transients and variables, both new and previously identified in other surveys. The search for SNe in particular has uncovered some rare types (Drake et al. 2011), a large fraction being extremely luminous with a tendency to favor very faint host galaxies. Our methodology is well suited to discovering these types since it relies on minimal contamination from host galaxy light, or other bright "static" underlying/nearby emission to avoid being masked for reliability purposes.

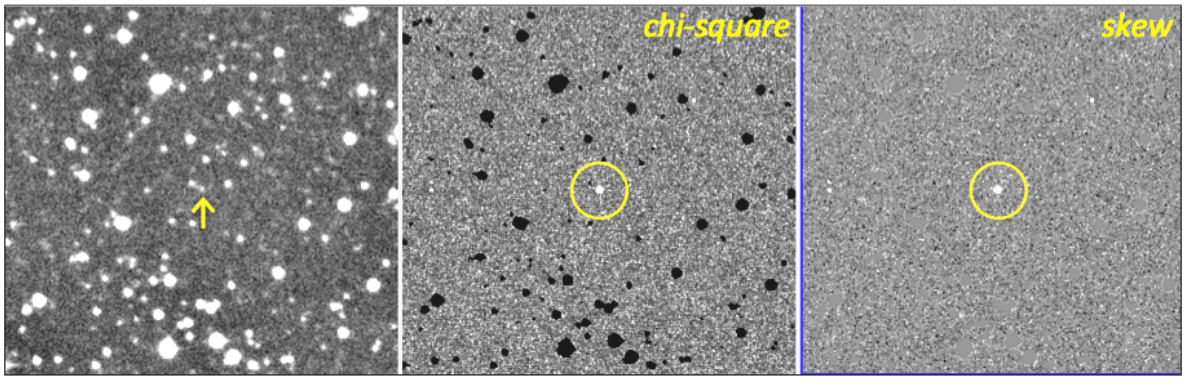

Figure 13. *Left*: long-run median-combined image containing a barely visible galaxy, and *right*: Metric images from *imtrandetect* of a ~ 3′ × 3′ field centered on the type IIn supernova SN 2011cw discovered by CRTS on May 5, 2011. A running window of 15 images was used.

Our testing is very preliminary, although we have managed to recover several previously discovered SNe, e.g., Figures 13 and 14. Guided by the simulations described above, a moving block window of 15 images was used throughout. We pushed down to an effective single-epoch S/N of ~ 3 and uncovered a false-positive rate of ~ 6%, comprising mostly instrumental glitches. This isn't too bad compared to other traditional approaches (e.g., image differencing) down to the same S/N level. We also uncovered a plethora of faint asteroids which at the time of writing,

may or may not have been previously discovered. These are detected by virtue of their motion, however slight. The metrics are sensitive to differences in flux at fixed sky positions across an image stack. Objects moving at speeds of typically > an effective PSF width between epochs will inevitably appear "compact" and trigger a detection in the metric-image. They show up as "events" on a light-curve since the photometry is *forced* at the fixed sky location. Light-curves and thumbnails for two asteroids uncovered with *imtrandetect* are shown in Figure 15.

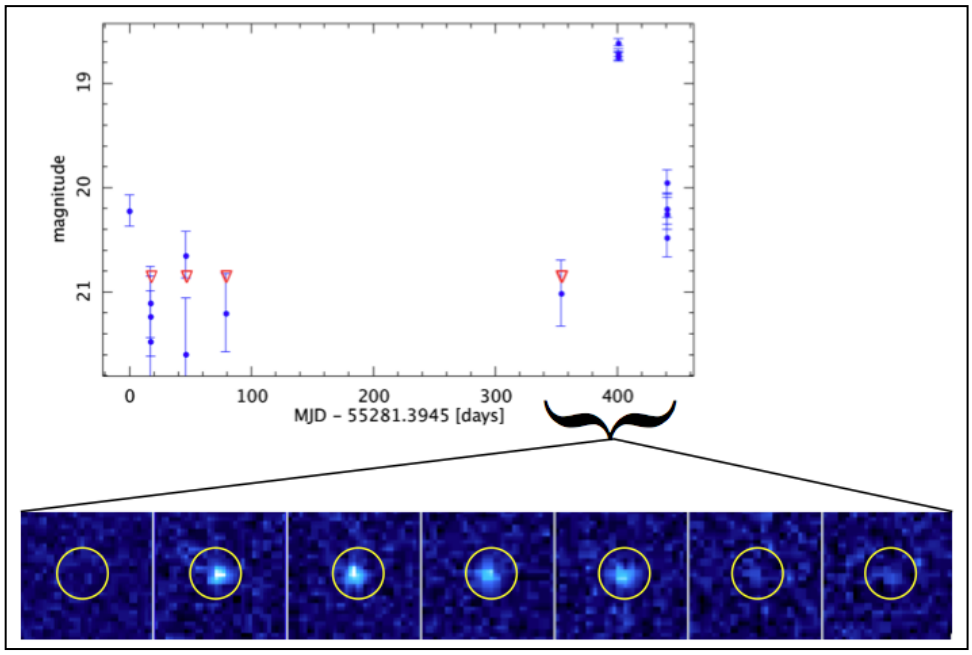

Figure 14. Light curve and single-epoch thumbnail images from *imtrandetect* of SN 2011cw detected off the metric images shown in Figure 13.

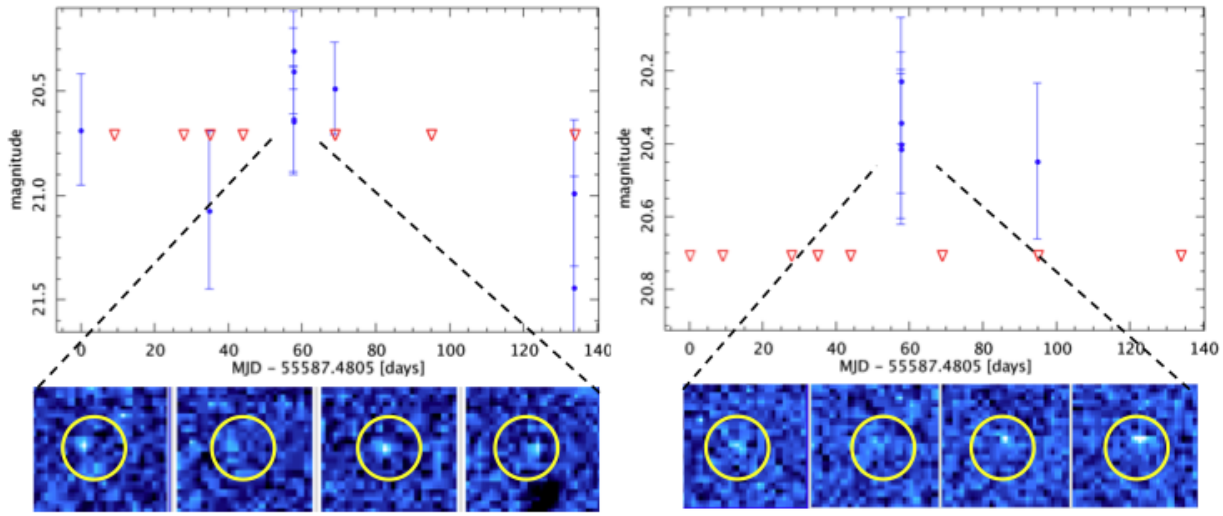

Figure 15. Asteroid candidates. Estimated speeds are ~15 and 9 arcsec/hr for the left and right objects respectively.

We have described a methodology and tool for optimally detecting low S/N (possibly intermittent) transient events from an incoming data stream or an image archive that may have escaped detection using traditional methods. The goal is not to replace existing methods but extend them (perhaps in parallel) to maximize the scientific returns of a dataset given the observational and technological limitations. We emphasize reliability over completeness since we are interested in detecting rare events on the surface of a sea of noise. Only by judiciously combining observations where a transient may be active do we stand a chance of going beyond what sequential single-epoch searches can offer.

Setting aside technological improvements, there may be other more optimal methods and metrics (in the maximal S/N sense) than what we presented here. We will continue the search. Furthermore, we plan to optimize and extend the *imtrandetect* tool with more functionality, in particular to enable the use of prior information to assist with reliability, weeding out "uninteresting" transients, and/or targeting a specific class of transient for further study.

A complementary approach to this problem was initiated by D. Thompson et al. They treated weak event detection based on an entire time series (without the assumption of any single event above the "catalog threshold" cutoff). The most general case enatils optical transient phenomena in *time/space image cubes*, e.g. without any catalog at all. The ability to consider the complete time series at once can potentially improve the sensitivity of transient searches while possibly revealing phenomena that are not apparent to a traditional threshold test. One example from their initial analysis appears in Figure 16. This work continues.

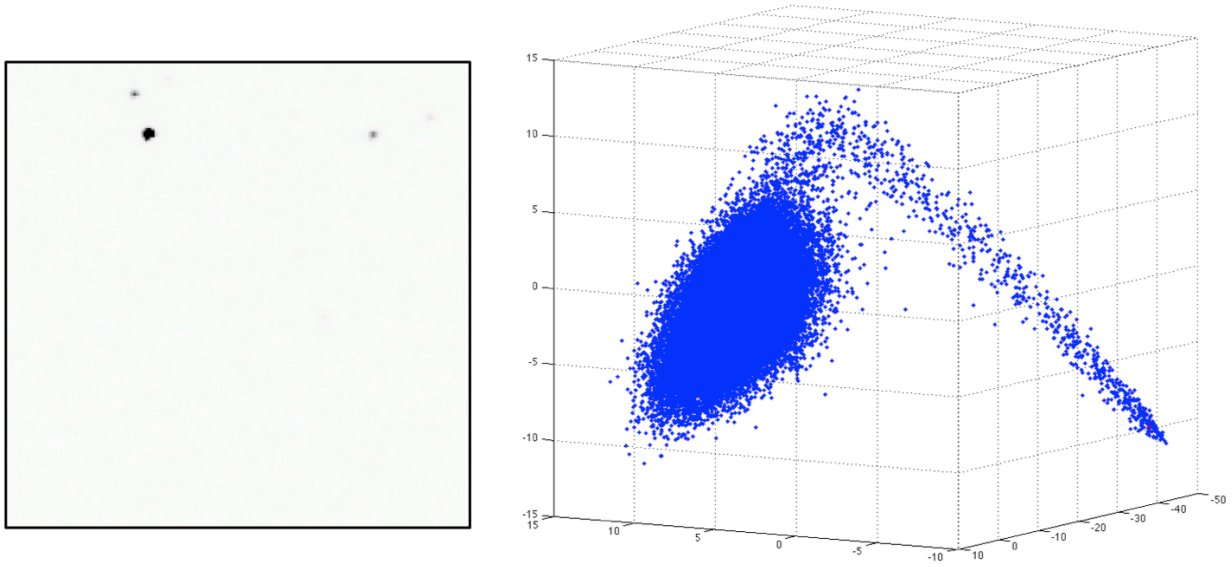

Figure 16. *Left:* One image segment a time series of images. A few bright sources are apparent, but the fainter ones are not easily discernible in the image itself. *Right*: A Principal Component projection of all image pixel intensities over the time series, capturing both the magnitude as well as its temporal evolution. The large "blob" at center corresponds to the majority of pixels which are the background, while the arm of the distribution identifies physical sources of varying magnitudes. Outlier points in these distributions represent pixel locations having unique or distinctive time/intensity signatures. A cut in this eigenspace can then isolate the pixels of interest that can be processed by the object-finding algorithms.

3.3 Classification of Variable and Transient Sources

We approached the problem in two distinct regimes:

(1) *Rapid, iterative classification of transient events.* In this regime, very little is known about a transient when it is detected: its position on the sky, flux, change from the baseline data, and whatever heterogeneous archival information may be available for that location on the sky. The sparsity and the heterogeneity of the available information make this a very challenging problem, sharpened by the time-critical nature of the process (e.g., in order to make the optimal decisions about the use of follow-up resources).

(2) *Classification of archival light curves of transient and variable sources.* In this regime, a sufficient number of flux measurements have been made, e.g., tens to hundreds, and this becomes a time series classification problem. However, the sampling may be very non-uniform, different for different sources in the survey, and different sources may have different numbers of measurements (i.e., not all light curves are created equal). Generally, while challenging on its own, this is a much easier problem than (1), due to the larger amount of available data, and the non-time-critical nature of the problem. However, insights gained in this regime can inform the approaches to (1), and may be used in the design of detection algorithms optimized for particular types of sources or transients.

This working group conducted a semi-regular series of meetings and discussions, many of them using a novel telepresence platform, immersive virtual reality (Djorgovski et al. 2009; see also http://www.mica-vw.org). Email correspondence was conducted through a dedicated list server, classtronomy@astro.caltech.edu, and some of the test data sets were posted on the KISS wiki.

3.3.1 Rapid Classification of Transient Events

The problem of astrophysical classification of genuine transient events is highly complex and challenging. When first discovered, all transients look the same in the images (star-like), and the initial information is very limited: the flux, the difference from a baseline, and the position on the sky. Even as data acummulate, not all parameters would be measured for all events, e.g., some may be missing a measurement in a particular filter, due to a detector problem; some may be in the area on the sky where there are no useful radio observations; many observables may be given as upper or lower limits. The sparsity, incompleteness, and heterogeneity of the data for individual events precludes use of feature vector-based techniques, and suggests Bayesian methods, as they can deal with missing data effectively.

We continued to explore Bayesian Networks (BN) for classification of transients found in the CRTS survey (Mahabal et al. 2010a, and in prep.). A Bayesian Network is a probabilistic graphical model represented through directed acyclic graphs (DAG), whose nodes represent parameters, and the missing arcs represent conditional independence assumptions. These networks can be used to compute the probability distribution of a subset of parameters when other parameters are observed (so-called probabilistic inference). To describe a BN we need to specify the graph topology and the parameters of each conditional probability distribution. An attractive feature of BNs is their ability to learn from the data. In the Bayesian approach we generate (and subsequently, update) a library of prior distributions capturing, for example, brightness changes in a certain filter over a certain time interval, conditional on object type such as type Ia supernova. Such distributions need to be estimated for each type of variable astrophysical phenomenon that we want to classify. Then an estimated probability of a new

event belonging to any given class can be evaluated from such pieces of information as are available. We make use of both Naive BNs as they are straightforward to implement, and more complex BNs partially structured by learning from the data, and partly advised by domain knowledge. The parameters, or nodes, can be directly observed quantities, such as the brightness changes or colors, derived quantities, e.g. light curve characteristics, context-dependent parameters, e.g., distance to the nearest radio source, or even the score from other classifiers dealing with some subset of the data. This work now continues.

Another approach uses *Probabilistic Structure Functions* (Djorgovski et al. 2011, 2012; Moghaddam et al., in prep.), see Figure 18. Since typical survey (flux-only) observations come in the form of magnitude changes over time increments – $(\Delta t, \Delta m)$ – we focus on modeling the joint distribution of all such pairs of values for a given LC (note: we consider all *causal* increments, corresponding to $\Delta t > 0$, therefore n LC observations lead to $n(n\text{-}1)/2$ pairs of Δ "change events"). By virtue of being increments these $(\Delta t, \Delta m)$ change values and their PDF will be *invariant* to absolute magnitude and time as well as corresponding shifts in each (since distance to an object and "true" onset time of its LC are unknown). These densities allow flux upper limits to be encoded rather easily – *e.g.*, under poor seeing conditions we may only have bounded observations, such as $m > 18$, which leads to a bounded Δm (which maps to a vertical segment in the histogram, as opposed to a single bin).

We can also *smooth* our 2D histograms in order to model uncertainties in $(\Delta t, \Delta m)$. Hence, this yields a computationally simple and effective way to implement a nonparametric density model that is flexible enough for the variety of object classes under consideration. Note that our histograms can be viewed as *probabilistic structure functions*: a standard structure function simply gives flux variance (a scalar quantity) as a function of Δt, whereas here we have a full PDF on Δm, indexed by Δt (from which a standard structure function can be easily derived). Figure 5 shows examples of these 2D histograms for three classes of transient objects.

When a new transient is detected, its $(\Delta t, \Delta m)$ histogram starts to be accumulated. After each new measurement, it is compared to a set of template histograms for different classes of transients. We apply a set of metrics that produces relative likelihoods of the new transient belonging to any given class. As the data accumulate, the classification accuracy improves.

The next step in the development of this classifier is to use 4D histograms of data point triplets. For example, if we measure magnitudes m_1, m_2, and m_3 at times t_1, t_2, and t_3, the histogram axes are now $(\Delta t_{12}, \Delta m_{12}, \Delta t_{23}, \Delta m_{23})$. These 4D histograms are sparsely populated, but separate the different classes more clearly. While the preliminary demonstration has been made, we are still working towards the speeding up of the algorithm that generalizes this approach to data point triplets and higher multiplets.

One outstanding issue is the possible sensitivity to window functions, i.e., LC sampling patterns. The annual/seasonal cycle is seen in the histograms. To the extent that all of the data in a given survey may be sampled in a similar way, this may not be a large problem. However, using template histograms generated from the data in a different survey may introduce some biases.

Obviously, this technique is equally applicable to the archival classification of LCs.

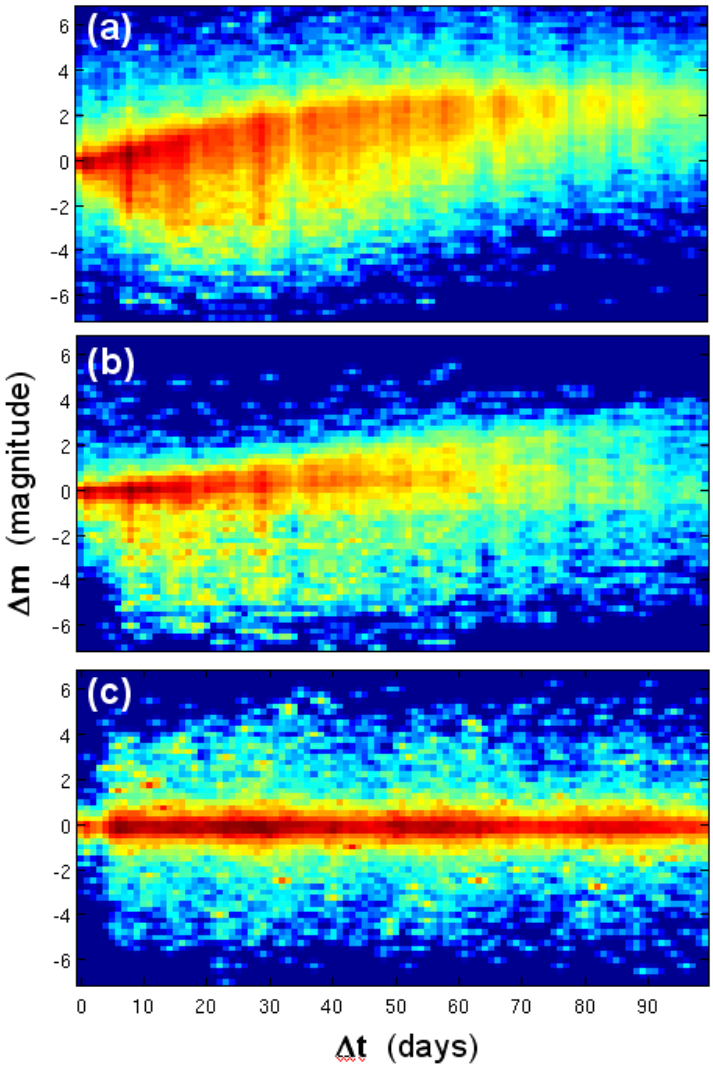

Figure 18. Probabilistic structure functions representing the joint distribution of (Δt, Δm) values from all (Δt > 0) paired observations in LCs, shown here as discretized 2D histograms of 3 classes of transients: (a) supernovae of type SN-Ia, (b) supernovae of type SN-IIp and (c) Cataclysmic Variables, using bin widths of Δt = 1 day, and Δm = 0.25, smoothing with an anisotropic convolution kernel, and with pixel intensity corresponding to log-probability. These class prototype histograms were obtained by pooling several hundred LCs from the corresponding 3 object classes. Note that the upward "arch" of the supernovae is due to their sustained flux decay (increasing Δm) and that the temporal/flux shape structure of all 3 classes forms a distinct signature. The probability of observed (Δt, Δm) values from a new (unknown) object's LC can therefore be easily "read off" (scored) by each histogram. Probabilistic structure functions can thus be viewed as "generative models" of (Δt, Δm) for their respective LC classes (i.e., as nonparametric likelihood functions).

In the analysis of radio transients, the JPL group (Wagstaff, Thompson, et al.) enhanced the strategy used in their SSEND system (Thompson et al. 2011ab) with a better interference model than the original approach presented at the opening workshop. The interference itself is treated as the component of the event most orthogonal to its local background. This yields a more faithful representation and notable algorithm performance improvements. The original result

won a best paper award at the Conference on Intelligent Data Understanding (October 2011). The later, improved version and final results are included in a paper, now in press. This is illustrated in the application of the SSEND algorithm to simulated data from the ASKAP radio telescope (Figure 19).

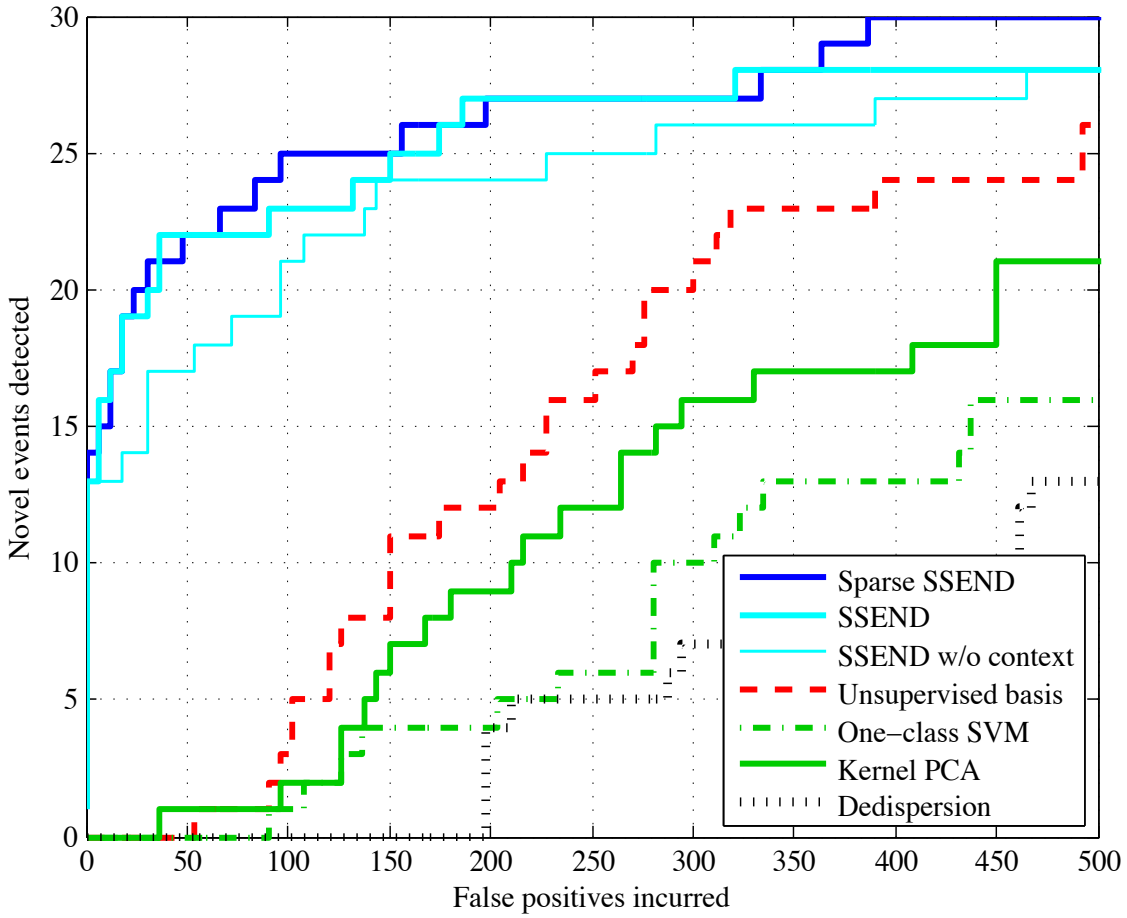

Figure 19. Performance for detection of anomalous "peryton" transients in radio time series data. The adaptive model performance is represented by blue lines, while other colors show state of the art alternatives. The blue lines approach the upper left, representing a high tolerance to false positive anomalies. The improvement in the algorithm performance produced between the opening and the closing workshops is apparent in the difference between thin and thick blue lines.

3.3.2 Automated Classification of Light Curves

If a variable or transient source has been monitored sufficiently long so that its light curve has at least a few tens of epochs, classification of light curves becomes a considerably better defined task. In addition, purely archival studies of variable sources of different kinds can be conducted with large sets of LCs.

In general, light curves can show tremendous variation in their temporal coverage, sampling rates, errors and missing values, etc., which makes comparisons between them difficult and training classifiers even harder (Figure 20). A LC for a newly detected supernova may just have a few points whilst those from monitoring projects, such as CoRoT or SuperWASP, can have

tens or hundreds of thousands of data points. Any classification algorithms must deal with such heterogeneity and sparsity of data, and we recognized up front that different algorithms may perform optimally in different regimes.

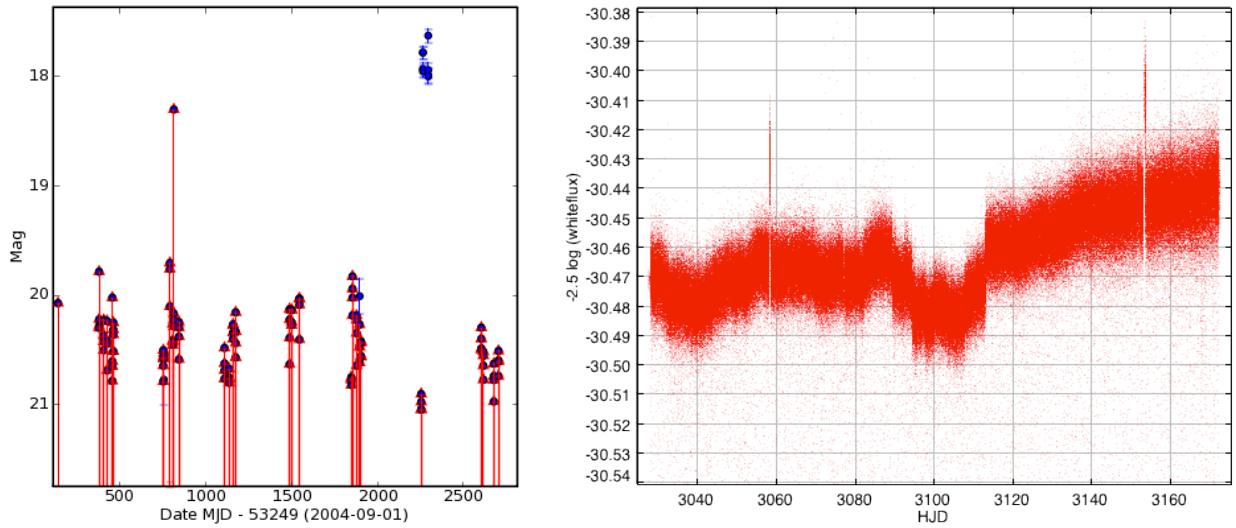

Figure 20. This illustrates the extremes of light curves. Only the black points in the figure on the left (a supernova) are real observations – all other points are just upper detection limits. In contrast, the densely sampled light curve on the right shows ~150,000 points from the CoRoT project.

In order to confront this heterogeneity, we can replace the LCs with a set of common statistical or morphological descriptors, that effectively form feature vectors (see, e.g., Richards et al. 2011). We can then use this alternate, homogeneous representation as the basis for further analysis or training of ML algorithms. Many different types of feature are used in the literature to capture information contained in the light curve: moments, flux and shape ratios, variability indices, periodicity measures, model representations, e.g., HMM, as well as more sophisticated techniques such as segmentation methods and discretization. The Caltech Time Series Characterization Service (http://nirgun.caltech.edu:8000) aims to extract a comprehensive set of features from any supplied light curve - currently over 60 features can be supplied. Vectors of such features derived from the light curves of known classes of objects can then be used as the training sets for particular supervised classifiers.

We experimented with Decision Trees (DT; see, e.g., Breiman et al. 1984) for an optimal classification of these feature vectors; C. Donalek was the lead in this effort. In a DT each internal node denotes a test on an attribute, each branch represents the outcome of the test and each leaf holds a class label. In our tests, DTs have been trained using the feature vectors for various combination of classes. To reduce the dimensionality of the input space, we have applied a forward feature selection strategy that consists in selecting a subset of features from the training set that best predict the test data by sequentially selecting features until there is no improvement in prediction.

Each tree is built using the Gini diversity index (gdi) as criterion for choosing the split; the splitting stops when there is no further gain that can be made. To avoid overfitting we use a 10-fold cross validation approach: the original sample is randomly partitioned into 10 subsamples. Each time a single subsample is retained as test, and the remaining records are used as training

data. This process is then repeated 10 times with each of the subsamples used exactly once as test. Moreover, the DTs are pruned in order to choose the simplest one within one standard error of the minimum. Tables 1 and 2 show the results obtained applying this procedure to a data set composed of LCs of Blazars, CV and RR Lyrae from the CRTS survey, and SNe from the SN Challenge data set (NEED REF).

	Completeness	Contamination
Blazar	83%	13%
CV	94%	6%
RR Lyrae	97%	4%

Table 1. Results obtained using an optimized DT classifier on a set of LCs of Blazars, CVs and RR Lyrae from CRTS. Best discriminating feature set consists of: Amplitude, beyond1std, flux_percentile_ratio_mid65, max_slope, qso, std, lomb-scargle.

	Completeness	Contamination
SNIa (879)	96%	9%
SNIb (55)	33%	25%
SNIc (45)	33%	32%
SNIIn (86)	75%	24%
SNIIp (282)	83%	15%

Table 2. Results obtained using SN LCs from the SN Challenge data set. The numbers in parenthesis show the sample sizes for each given class. Best discriminating feature set consists of: flux_percentile_ratio_mid50, median_absolute_deviation, pair_slope_trend, percent_amplitude, percent_diff_flux_percentile. The classifier séparâtes SNe Ia well from the rest, which is a meaningful physical différence. It does not do as well in separating the remaining types of core collapse SNe, due to a général similarity of their LCs.

Comparable results have been obtained by the Berkeley group (Richards et al. REFS), who used Random Forests.

In a complementary approach, J. Scargle elaborated on the use of Bayesian Blocks (BB; Scargle 2005), a non-parametric method that can achieve this with the only assumptions being generic priors on amplitudes and number of blocks used. Given data consisting of N observations, $\{X_n, n = 1, \ldots, N$, how can we estimate the probability distribution, $p(X)$, i.e., the optimal data model for this time series? The simplest possible data model of a variable source is a piecewise-constant model of the time series. Standard histogram techniques to estimate the density assume that X_n are independent draws from the same distribution, equal size bins and have bin size, number and location as parameters to set. A BB-based histogram only assumes independent draws; in addition, it can handle multivariate data, treats gaps gracefully and takes account of exposure variations. A dynamic programming-based algorithm exists to calculate the BB representation of a time series in an optimal fashion. Essentially any analysis method which involves data binning will work with this data representation: for example, the generally useful Discrete Correlation Function algorithm (Edelson & Krolik 1988) is easy to code up with data cells (which could be anything). This can also be easily extended to higher dimensions using Voronoi cells for the data points. LCs in their BB representation can then also be treated as feature vectors, and classified using any of the standard ML techniques.

Another novel approach that we explored in the course of this study is the use of Machine Discovery, i.e., software that can formulate and test data models. The particular package that we used, with M. Graham as the lead, is Eureqa (Schmidt & Lipson 2009, software available at http://nutonian.com).

This is a software tool which aims to describe a data set by identifying the simplest mathematical formulae which could describe the underlying mechanism that produced the data. It employs symbolic regression to search the space of mathematical expressions to determine the best-fitting functional form – this involves fitting both the form of the equation and its parameters simultaneously. Binary classification can be cast as a problem amenable to this tool – the "trick" is to formulate the search relationship as: $class = g(f(x_1, x_2, x_3, \ldots, x_n))$ where g is either the Heaviside step function or the logistic function, which gives a better search gradient. Eureqa finds a best-fit function, f, to the data that will get mapped to a 0 or a 1, depending on whether it is positively or negatively valued (or lies on either side of a specified threshold, say 0.5, in the case of the logistic function.)

We considered three specific binary light curve classification problems using Eureqa: RR Lyrae vs. W UMa, CV vs. blazar, and Type Ia vs. core-collapse supernovae. For each case, we compiled data sets of light curves from the CRTS survey for the appropriate classes of objects, and derived ~30 – 60 dimensional feature vectors for each object. A set of 10 Eureqa runs was performed for each case with each run omitting 10% of the data and the best-fit solution for that run then applied with the omitted data as the validation set so giving us 10x-cross-validation on the resulting solutions. Some of the preliminary results are given in the confusion matrix shown in Table 3; the main diagonal gives the completeness fractions, and the orthogonal diagonal gives the contamination fraction for each type.

	RRLyrae / CV / SNeIa	WUMa / Blazar / CCSNe
RRLyrae / CV / SNeIa	98.3% / 91.1% / 92.5%	1.7% / 8.9% / 7/5%
WUMa / Blazar / CCSNe	3.6% / 37.5% / 58.6%	96.4% / 62.5% / 41.4%

Table 3. Results obtained Eureqa for the discrimination of three pairwise classifications, as noted in the text. The best discrimination is between the two types of pulsating variables, and the worst is between the two types of SNe, due to an overall similarity of their LCs; with Blazar/CV in between the other two cases.

As these preliminary results show, at least in some cases Eureqa can identify and characterize physically meaningful structures in feature vector data to a sufficient degree that it can be employed for binary classification. An advantage of this is that Eureqa provides an analytical expression to separate the classes rather than relying on application of a trained black box algorithm. The work on this project continues.

4. The Closing Workshop, December 12 – 15, 2011

This workshop represented a formal end of the study, although many of the collaborative research activities are still continuing. It began with an open technical workshop, attended by about 60 participants, with the following presentations. Most of them were summarizing the challenges and reporting on the results of the study to date, but some were addressing other relevant issues in the computational science.

4.1 Workshop Agenda

Speaker	Title
C. Cutler	Detecting weak, long-lived chirps
V. Dergachev	Loosely coherent algorithms - robust and computationally efficient search of large parameter spaces
F. Masci	Imtrandetect: a new tool and methodology for digging out transients down to low SNR levels
A. Mahabal	Extracting Faint Intermittent Transients
D. Thompson	Interference-resistant real time adaptive detection
P. Protopapas	Variable classification and beyond
J. Richards	Automated Discovery and Classification for the PTF
B. Moghaddam	A Stochastic Structure Function for Light Curves

C. Donalek	Classification of Transient Events in Synoptic Sky Surveys
M. Graham	Deconstructing classifiers - a postmodern approach to data mining?
U. Rebbapragada	Classification of VAST Radio Transients and Variables
J. Scargle	Tao of Better Histograms: 1D, 2D and Higher
V. Kashyap	Project Tanagra: Timing Analysis of Grating Data from X-ray observatories
J. Babu	Analysis of astronomical datacubes
M. Stalzer	Trends in Scientific Discovery Engines
Y. Xu	Discovery of Hidden Patterns in Data Through Interactive Search
G. Rocha	PowellSnakes: a fast Bayesian approach to discrete object detection in multi-frequency astronomical data sets
R. DiStefano	From planets to black holes: searching for lensing events in data from wide-field surveys
C. Law	All Transients, All the Time: A New Algorithm for Interferometric Radio Transient Detection
J. Hartman	All-sky transient searches with the Long Wavelength Array
N. Fotopoulos	Toward Early-Warning Detections of Compact Binary Coalescence
L. Singer	Optimization and Coordination of Electromagnetic Followup
L. Eyer	Classification of the variable stars of Hipparcos and Gaia
P. Huijse Heise	Finding periodicities in astronomical light curves using information theoretic learning
F. Bianco	LCOGT/LIHSP: A Robotic system for Lucky Imaging
Y. Xu	Building a better scientist

The agenda with the links to slides can be also found at
http://www.astro.caltech.edu/digging/index.php?mode=agenda .

4.2 Summary of the Selected Presentations and Working Group Summaries

Working group leads described the activities covered in Sec. 3 above. In addition, many speakers addressed other, related issues and challenges that may naturally fit in the follow-up studies. Some of them are described briefly here.

4.2.1 Event and Light Curve Classification for GAIA

Dr. Laurent Eyer, described some of the challenges from the upcoming GAIA mission, that are closely related to the subjects covered in the KISS workshop. This mission is expected to revolutionize our understanding of Galactic astronomy, and also have a significant impact on the stellar physics, cosmological distance scale, etc.

Variability on the sky is one of the key scientific goals of GAIA, and the first step is to detect variable phenomena. To this effect, in addition to the traditional approaches like the chi-square, the mission team seeks to implement specific algorithms that take advantage of what we know about a particular type of variability, which would be missed by classical global statistical tests. Examples of such phenomena are planetary transits or small-amplitude periodic variability.

One of the most important tasks is the classification of the variable sources observed by GAIA. The classification is structured into a three-step process: (1) a number of attributes are first computed to characterize the source and the variability properties, (2) these attributes are then fed into the classification algorithms, and (3) finally specific processing is applied to the sources of each of the different class to validate and possibly refine the classification result.

The different types of classification approaches considered by the GAIA team are categorized as supervised, unsupervised, and extractors:

(1) Supervised classification. One of the most important factors of supervised classification is to find best attributes in order to build the training set for classification algorithms. The strategy is to iteratively refine the training set by using different attribute/method combinations, then compare the results. The most meaningful attribute set, both in terms of separation power and independence, can thus gradually be derived. Two steps are identified in order to build a representative training set for the GAIA variability classification task: first, a list of sources representing different variability classes has to be established; second, light curves, similar to the ones thaht are expected to be obtained by the GAIA mission, must be collected for selected source classes from the existing data. Eventually, GAIA measurements during the mission can also be used in that step. The final training set is likely to emerge from the iterative process as described above. Several classification methods were compared and Random Forest appears, at the moment, as the most convenient and reliable classification algorithm (Dubath et al. 2011).

(2) Unsupervised classification. The challenge of the unsupervised classification with the GAIA data may well come from the potentially large number of variable sources detected, probably of the order of hundreds of millions. Several algorithms have been explored by the GAIA team, mostly based on the K-means technique. This topic remains a subject of active work by their team.

(3) Extractors. These are tools that take advantage of the knowledge gained about the light curve behaviors for certain types of variable objects. Scanning the complete set of variable objects, they try to identify specific light curve behaviors, and thus they can "extract" candidate sources of a given class. Examples of these include various types of transients, and gravitational microlensing events.

The different approaches for the classification and the work done for specific objects will help to determine better the aspects of contamination and completeness, which are fundamental in any classification scheme.

4.2.2 Light Curve Classification and AGN Selection

P. Protopapas provided an update on the classification work being done at the CfA Time Series Center. Of a particular interest is their novel approach to detection of AGN using optical variability. Variability-based searches complement the standard color-based techniques, ostensibly leading to more complete samples. Correlations of AGN variability with other physical properties can also lead to some new insights into the physics of AGN.

MACHO data over 7.4 years covering the LMC and SMC provide an excellent sample of light curves that can be used for this purpose. Light curves are parametrized with features like variance, auto-correlation, structure function, Stetson coefficients, cumulative sums, colors, magnitudes, forming feature vectors. Various ML methods were used to select AGNs from this sample. For example, using SVN method on 40 million lightcurves yielded 1620 QSO candidates, more than thrice based on SDSS surface density estimates. These were cross-matched with other samples like Spitzer, 2MASS, Chandra etc., using their published color parameter space bounds for AGN (e.g., Kozlowski & Kochanek 2009 for mid-IR Spitzer comparison). Additional methods like AGN-galaxy separation, fitting AGN SED templates, photo-z methods and looking for X-ray counterparts were used to further subselect the best candidates. Finally training using the 58 known MACHO QSOs was used to get the final high confidence set of 663 AGN candidates. Techniques like random forests were used separately to get a similar sample. The cross-match selects candidates with ~98% overlap. A variety of supervised and unsupervised methods can thus be used to select a high-confidence sample of AGNs based on known criteria and a small training sample.

In a related presentation, P. Huijse described a fully automated and robust method to discriminate periodic versus non-periodic light curves. The method uses concepts from the Information theoretic learning framework (ITL) to solve the period detection problem. In a nutshell ITL statistical descriptors, such as Renyi quadratic entropy and correntropy, are generalizations of second order moment statistics such as variance and correlation. ITL metrics have been used in the field of machine learning to develop training algorithms that are superior to conventional second order algorithms. The first approach uses the ITL generalized correlation function or correntropy. This function was extended using a slotting scheme in order to evaluate the unevenly sampled light curves. It was compared with conventional methods such as the Lomb-Scargle and AoV periodograms in a set of periodic light curves from the MACHO survey, and outperformed the traditional methods on period estimation of eclipsing binary stars, while performing equally well in pulsating variable stars (Cepheids, RRLyrae). To solve the problem of periodic light curve discrimination we propose a metric called correntropy kernelized periodogram (CKP), which does not require folding, re-sampling or slotting schemes, which is currently being tested on the EROS database. The goal is to achieve the false positive rates below 0.1%, while being computationally efficient.

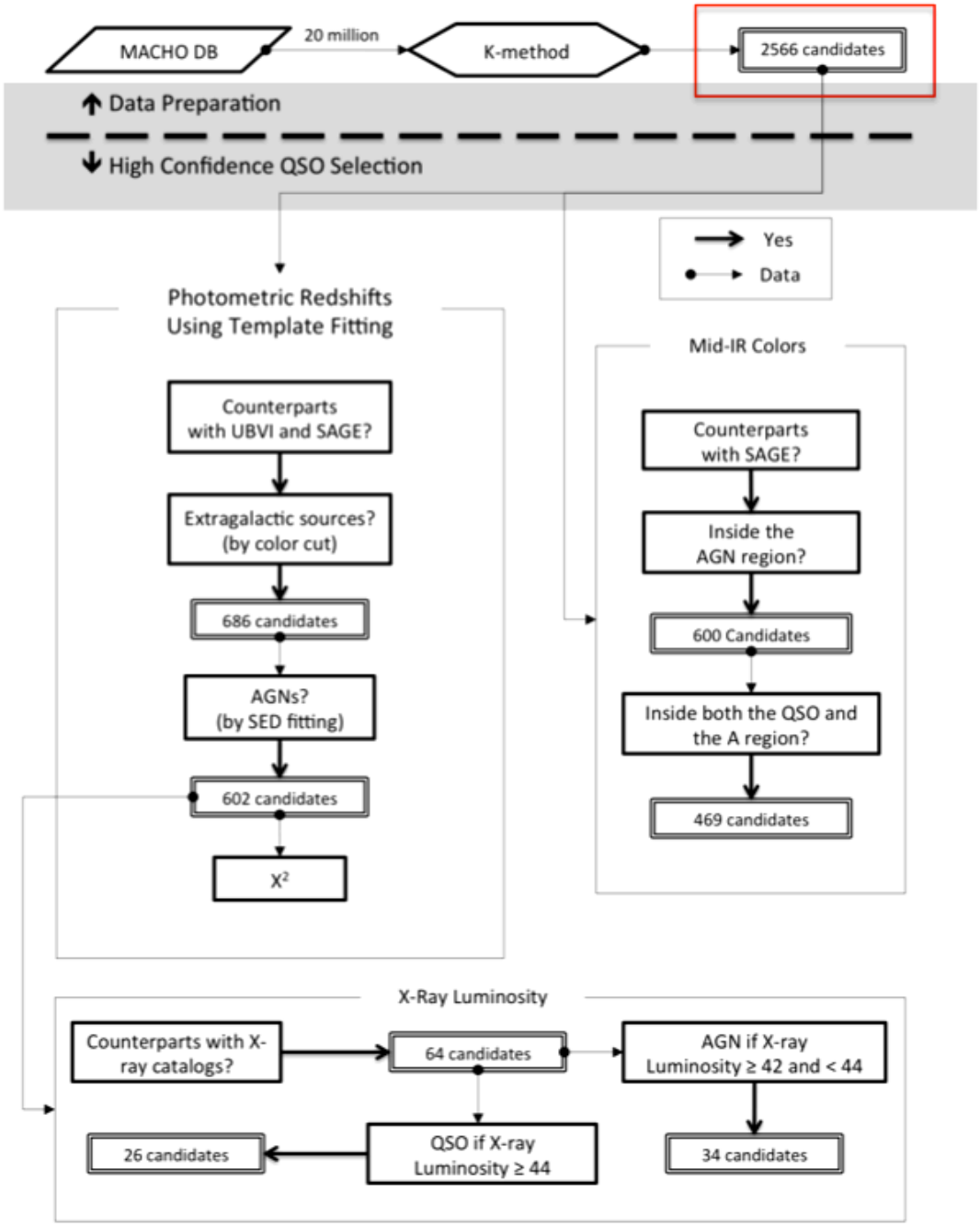

Figure 21. An overall flowchart for high-confidence AGN sample selection using different crossmatches and various model fits. A combination of these methods can be used to subselect a few hundred best AGN candidates from tens of millions of light curves.

4.2.3 Other Selected Presentations

Dr. Yan Xu presented a talk "Discovery of Hidden Patterns in Data through Interactive Search," introducing the *Environmental Informatics Framework* (EIF), a strategy and technology platform that the Microsoft Research Connections *Earth, Energy, and Environment* group developed to help advance data exploration in environmental research. Dr. Xu also demonstrated *Microsoft*

PivotViewer, a faceted search technology included in EIF that enables users to visually and interactively search and discover hidden patterns in massive data or image sets.

Dr. Xu also presented a talk, "Building a Better Scientist," where she discussed how the "Fourth Paradigm" for data-intensive scientific discovery is changing the way scientists conduct research, and is, therefore, creating a need for a new generation of scientists with advanced computational mindsets. The presentation stimulated passionate discussions, given the eminent need for training of scientist with computational skills needed to extract maximum knowledge from the data quickly and effectively, the very subject of our workshop.

5. Education and Public Outreach

A KISS public lecture titled "Science in Cyberspace" was given by Prof. Djorgovski in the evening of Dec. 13, 2011, in the Hameetman Auditorium of the Cahill Center for Astronomy and Astrophysics at Caltech. The abstract of the lecture is as follows:

> *"Science, scholarship, and education are being transformed by the advances in computation and information technology. Much of the scholarly work, including data, tools for their exploration and theoretical modeling, literature, and collaboration tools, are now moving to virtual environments. The exponential growth of data volumes, and the simultaneous increase in the data complexity offer both new scientific opportunities and new challenges for knowledge discovery in massive and complex data sets and data streams. We are now developing new methodologies for the scientific research in the 21st century."*

The video of the lecture was posted on the KISS website, http://kiss.caltech.edu/workshops/digging2011b/video/djorgovski/djorgovski_13dec11.html, as well as at Caltech's iTunes website, linked at http://itunes.apple.com/us/itunes-u/keck-institute-for-space-studies/id422626460.

These archived videos can reach a much wider audience than that was present in the auditorium.

In addition, the workshop inspired a more formal and more advanced educational effort, preparation of the first textbook on the emerging field of Astroinformatics. Training the next generation of students and postdocs to understand and use effective new computation and information technologies for research in the era of exponential data overabundance is a critical and growing need. These challenges are of course not confined to astronomy, but are common to all sciences today.

To this effect, Profs. Longo and Djorgovski proposed to develop the first textbook on this subject, tentatively titled *"Practical Astroinformatics: Methods and Tools",* aimed at the upper level undergraduate and graduate students (and also postdocs). The textbook would be primarily Web-based and freely available, with an accompanying traditional hardcopy version as well, possibly done through a print-on-demand service (we are exploring the possibilities with several major ppublishers). The electronic nature of the book would allow for its steady evolution and improvements, links to the relevant and useful resources (texts, codes, data, etc.), social media for continued discussions, feedback, exchange of ideas, interest groups, etc. This e-textbook can spur development of the new curricula in this arena, form a basis for courses and summer schools, and it may serve as a leading example for a new kind of textbooks for the 21st century.

The intended coverage includes: best programming practices and code maintenance; computing environments; databases, archives, and data structures; Web/grid/cloud services and applications; data mining and knowledge discovery tools and methods; scientific data visualization; commonalities with other fields; available software resources; uses of on-line/virtual environments for scientific collaboration and communication; introduction to the semantic web; etc. The content is envisioned to evolve, as the scientific and educational needs evolve.

In order to gather the necessary expertise of such broad variety of planned topics, our plan is to engage a number of invited chapter authors, each of whom would be a world-class expert in their subject. Profs. Longo and Djorgovski, in addition to providing some of the content, would serve as chief editors of the volume, and assure the coherence and uniformity of the coverage.

A class at Caltech, Ay 119 "Methods of Computational Science", taught by Prof. Djorgovski and the scientists in his group in Spring 2012, was used as a testbed for the curriculum development. We expect to have the first draft of the textbook by the end of this calendar year.

6. Participant Feedback

The organizers and most of the participants expressed a great satisfaction about the workshops and the overall stimulating effects of the study. Here are some of the responses we got:

From Dr. Yan Xu of Microsoft Research:

> I enjoyed both of the "Digging Deeper" workshops and learned a great deal. In particular, listening to astronomers presenting their data problems was a very valuable learning experience for me as a computer science researcher.
>
> I was pleased to receive positive feedback from attendees about the work that Microsoft Research is doing for data-intensive sciences. As one participant noted to me in email, "I have to admit that I wasn't aware of the work that Microsoft Research was doing, but I was very impressed with what I saw yesterday. The work you've been doing on data visualization can only be described as stunning!"
>
> My observation was that "Digging Deeper" really stimulated discussions around data and technologies on data mining. It was very informational for me to have the Q&As following my presentation on "Discovery of Hidden Patterns in Data through Interactive Search". I appreciated the opportunity to present the researchers what Microsoft technology can do for their data sciences. I also truly enjoyed the discussions with the audience following my presentation on "Building a Better Scientist". Actually, I published a blog post about this on MSDN, http://blogs.msdn.com/b/msr_er/archive/2011/12/19/coping-with-data-deluge.aspx.
>
> It would be good to see follow-up actions from "Digging Deeper", the report you are gathering, and perhaps workshops with specific focus on topics such as "new technologies for data mining", "creating a generation of data scientists", etc. Please let me if I can be any help on behalf of Microsoft Research.

From Dr. L. Eyer, representing the GAIA mission:

> Brainstorming is probably the best word to characterize the KISS meeting. It is the first time I attend such a meeting and I found it particularly interesting. It was a bit chaotic sometimes, which is probably necessary for creative thinking. I would describe the atmosphere as friendly and also frank. Such a meeting is important also on social contacts. Several

generated ideas are probably very interesting, but now the ball is in our hands to make them real improvement of knowledge in astrophysics.

From Dr. K. Wagstaff:

I thought the opening workshop was excellent, from the short course to the lightning talks to the breakout sessions to the individual discussions. The provision of offices and work spaces was very conducive to collaboration and the development of new ideas. I learned a lot.

Between the opening and closing workshops, I participated in weekly meetings of "group 3" (focused on source classification). There were interesting and varied. Mostly we used Second Life as a virtual meeting venue, mostly with success.

While the workshops and related discussions have been great and led to interesting collaborations, we certainly haven't "solved" the source type classification problem. I think it'll take more than six months of coordinated effort. Our group is perhaps too large and diverse for an ongoing weekly meeting to be practical. But I would welcome some mechanism that would motivate us to stay in touch, and even better if we can spark interest in these problems for the wider research world. Perhaps brainstorming challenge problems to be posted publicly? Certainly there's enough data to support such a thing!

From Dr. D. Thompson:

Overall, the workshop was a great experience. It was a rare chance to collaborate with scientists at the cutting edge of many distinct new application domains such as gravitational wave detection and optical sky transient surveys. I am led toward the conclusion that each these communities have detection methods that are very highly refined to their specific instruments and problems. This prevents a generalist or an outsider from another field from making a dramatic breakthrough on their own in the short time available and the pressing demands of the participants' own research efforts. This is not to say that the meeting was not useful – quite the opposite… I think that we did a great job of carving out some soluble niche problems. Moreover, I think the true value of the workshop is to provide cross-disciplinary exposure so that I can apply these other fields' common practices to my own research.

I think that overall the KISS process was effective, and the Institute rules and protocol seem effective for these intense think tank sessions. I wouldn't change them! Given the particularly diverse, multidisciplinary backgrounds of the participants in this workshop, it might have been further improved with more bottom-up, grassroots problem definitions. Defining three "problem areas" in advance as a team forced us to create big-tent challenges that could incorporate everyone. We did a good job with this approach, though it did lead us to tackle some very big and general problem areas, and resulted occasionally in full-group discussions which were only relevant to a subset. Alternatively, letting participants freely propose their own problems during the workshop, and attract small asynchronous teams around them, would have created more space to be opportunistic, and ensured that all the required knowledge, skills, and contacts were present for each problem.

Overall, I commend the KISS organizers as well as the workshop leads for pulling this off!

From Dr. M. Turmon:

I thought the collection of people who attended the workshop was well-chosen and worked effectively together. The workshop allowed me to learn more about the problem of detecting exponential chirps in noise at challenging SNR. I had not been aware of the work of Bickel,

P., Rice, J., and Meinshausen, M. 2009, "Efficient blind search: optimal power of detection under computational cost constraints", *Annals of Applied Statistics* 3, 38-60) on this problem. I believe that the search problem in this domain (looking for maxima in a detection statistic that cannot be evaluated everywhere on a very fine grid, but rather, is evaluated on a succession of coarse-to-fine grids) could be helped by being able to characterize the shape and size of the excursion regions. I pointed out some related work on this problem (e.g., the monograph by David Aldous, *Probability Approximations via the Poisson Clumping Heuristic*) to Curt, and perhaps it was of interest.

Although this was not related to the topic of the workshop, Vinay Kashyap of CfA, who attended the workshop, and I discussed our work on classification of the morphology and behavior of solar active regions. This work, which has been separate but related, involved (on the one hand) code I wrote for extracting the entire paths of specific solar active regions, and (on the other hand) joint work Vinay has done with David van Dyk and David Stenning on classification of active region type. This resulted in me participating in a workshop at CfA in February of 2012 (http://hea-www.harvard.edu/AstroStat/SolStat2012/), and giving a talk there ("Algorithms for Solar Active Region Identification and Tracking").

From W. Max-Moerbeck:

The workshop was a great opportunity to learn about other problems where time series analysis is an important component, like the study of gravitational wave signals, classification and follow-up of transients, gravitational lensing and many other topics. The quality of the speakers was very good and the interactions very intense. I really like the informal style of the workshop, because I think I got to really talk about research and listen to others in a relaxed environment. A couple of ideas for research came out of the workshop, which I hope I can materialize after I am done with my thesis work.

From G. Cabrera:

The Workshop was a great opportunity for discussing about algorithms for astronomical data processing. Although my primary line of investigation is not time-series driven, I met very interesting people whose main objective is very similar to mine: addressing the astronomical data deluge problem. The expertise of people dedicated to time-series was of great help for understanding their problems and find similarities with other problems of astronomical data processing.

Some ideas were conceived and some others were further developed during the coffee breaks and long discussion times. In particular, we had the chance to exchange not only ideas, but also data and expertise on how to work on it. All this was achieved by promoting interdisciplinary work between astronomers, computer scientists, mathematicians and statisticians, a key element for these initiatives to work. This kind of work is relatively new for astronomical data processing. Sometimes you feel alone swimming against the tide, but workshops like Digging Deeper make you realize there are others there too. We just need to know and help each other to swim through the astronomical data deluge.

From R. DiStefano:

I can say that new collaborations have begun, that the KISS experience helped me to design an ongoing seminar on wide-field surveys, and that I have a richer skill set on which to draw when I work on projects involving data.

7. Conclusions and Recommendations for the Future Work

The study helped crystallize some of the outstanding challenges in time-domain astrophysics, and identified some promising algorithmic and methodological approaches to their solutions. Not surprisingly, many of these challenges turned out to be far harder and more complex than we originally hoped for. Several tangible results have been produced, and the work continues along several directions, inspired at least in part by this study.

Regarding the search for long, weak chirps, we did not come much closer to a "final theory" for how to maximize their computational efficiency (or, equivalently, maximize their sensitivity at fixed cost). What we did accomplish in this Study was the invention of a couple new techniques (or "clever tricks") to increase search efficiency. And this effort has continued; recently we invented yet another trick, the usefulness of which we are now investigating.

Of course, the granddaddy of all such "tricks" was of the FFT, which for current searches decreases the computational cost by factors of millions. In time series analysis, no algorithm that has been invented since then has given us that sort of revolutionary increase in power. Is there another idea out there that could buy us another factor of a million? We suspect that the answer is 'no', but we also think it is still worth looking. Also, even with our current basket of methods, there has not been enough work on how to optimally combine them into a full data analysis pipeline. That latter question could be answered without a stroke of genius. It would just take manpower, and we believe that it remains a very worthy and a quite attainable goal for future work in the near-to-mid-term.

Regarding the challenge of detecting faint, intermittent signals in imaging surveys, a novel method was developed, led mainly by F. Masci and A. Mahabal, and described in detail earlier in this report. The method has resulted in a practical software package, and it will be used for the analysis of data from a variety of imaging surveys, archival, current, or forthcoming. Additional ideas, based on the statistical properties of tails of distributions, are still being investigated.

Perhaps most of the work done in this study was made on the challenge of a rapid, automated classification of transient events, and the related, but somewhat easier challenge of an objective classification of light curves. A number of new possible approaches have been explored. One of them is the use of Probabilistic Structure Functions (see Sec. 3.3.1), which are now being generalized to higher data point multiplets in higher dimensionality spaces. We did an extensive experimentation with the use of Bayesian Blocks, devised by J. Scargle, and find them to be very promising for the analysis of time series in general. We implemented a comprehensive set of statistical descriptors of light curves, to generate feature vectors that can be clustered using a number of different supervised and unsupervised classification methods. One radically new approach is the use of machine discovery tools (e.g., *Eureqa*), that we evaluated for the first time in the astronomical context. Potential uses of Hidden Markov Models have been explored. Work continues along all of these avenues that were started or substantially expanded during the KISS study.

In all, we now have a much better understanding of the classification problems in time domain astronomy. However, it is clear that much more needs to be done; different methods are optimal in some use cases, but not in the others, and we have started mapping that space. The challenge of a rapid, automated, robust classification of astronomical transient events is still with us. Along with the challenge of an automated optimal decision making for the follow-up observations, this will remain to be a very active and timely research area in the coming years.

This study clarified many issues, refocused efforts by several groups, produced some tangible results, but perhaps more importantly, it opened many interesting new questions.

8. Publications and Presentations

The work performed or initiated during this study was reflected in a number of publications and conference presentations, as listed below. Several additional technical papers will be submitted to refereed journals based at least in part on this work. Obviously, the role of the KISS study is acknowledged where appropriate.

8.1 Publications

Brescia, M., Cavuoti, S., Djorgovski, S.G., Donalek, C., Longo, G., & Paolillo, M. 2012, "Extracting Knowledge from Massive Astronomical Data Sets", in: *Astrostatistics and Data Mining in Large Astronomical Databases*, eds. L.M. Barrosaro et al., *Springer Series on Astrostatistics*, New York: Springer, **31**.

D'Abrusco, R., Fabbiano, G., Djorgovski, S.G., Donalek, C., Laurino, O., & Longo, G. 2012, "CLaSPS: A New Methodology for Knowledge Extraction from Complex Astronomical Datasets", *Astrophys. J.*, **755**, 92.

Djorgovski, S.G., Mahabal, A., Donalek, C., Graham, M., Drake, A., Moghaddam, B., & Turmon, M. 2012, "Flashes in a Star Stream: Automated Classification of Astronomical Transient Events", in ref. proc. *e-Science 2012*, IEEE press.

Djorgovski, S.G., Donalek, C., Mahabal, A., Moghaddam, B., Turmon, M., Graham, M., Drake, A., Sharma, N., & Chen, Y. 2011, "Towards an Automated Classification of Transient Events in Synoptic Sky Surveys", in: *Statistical Analysis and Data Mining*, ref. proc. *CIDU 2011* conf., eds. A. Srivasatva, N. Chawla, & A. Perera, NASA Ames Res. Ctr. publ., **174**.

Djorgovski, S.G., Mahabal, A., Drake, A., Graham, M., Donalek, C., & Williams, R., 2012, "Exploring the Time Domain With Synoptic Sky Surveys", in *Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy*, eds. E. Griffin et al., Cambridge: Cambridge Univ. Press, **141**.

Djorgovski, S.G., Mahabal, A., Drake, A., Graham, M., & Donalek, C. 2012, "Sky Surveys", in: *Astronomical Techniques, Software, and Data* (ed. H. Bond), Vol.2 of *Planets, Stars, and Stellar Systems* (ser. ed. T. Oswalt), Berlin: Springer Verlag, in press.

Graham, M., Djorgovski, S.G., Mahabal, A., Donalek, C., Drake, A., & Longo, G. 2012, "Data Challenges of Time Domain Astronomy", in: special edition of *Distributed and Parallel Databases*, eds. Qiu, X., & Gannan, D., in press.

Graham, M., Djorgovski, S.G., Drake, A., Mahabal, A., Williams, R., & Seaman, R. 2012, "The VAO Transient Facility", in: Proc. *IAU Symp. 285, New Horizons in Time Domain Astronomy,* eds. E. Griffin et al., Cambridge: Cambridge Univ. Press, **318**.

Graham, M., Djorgovski, S.G, Donalek, C., Drake, A., Mahabal, A., Plante, R., Kantor, J., & Good, J. 2012, "Connecting the Time Domain Community with the Virtual Astronomical Observatory", in: *Observatory Operations: Strategies, Processes and System IV*, eds., Peck, A., Seaman, R., & Comeron, F., *Proc. SPIE*, **5884**, in press.

Graham, M.J., Djorgovski, S.G., Mahabal, A., Donalek, C., Drake, A.J. 2012, "Miachine-Assisted Discovery of Relationships in Astronomy", *MNRAS*, in press.

Graham, M.J., Drake, A.J., Djorgovski, S.G., Donalek, C., Mahabal, A. 2012, "A Comparison of Period Finding Algorithms", MNRAS, in prep.

Huijse P., Estevez P.A., Protopapas P., Zegers P., Principe J.C. 2012, "An Information Theoretic Algorithm for Finding Periodicities in Stellar Light Curves", *IEEE Trans. Signal Proc*., **60**, No. 10, 5135-5145.

Mahabal, A., Djorgovski, S.G., Drake, A., Donalek, C., Graham, M., Moghaddam, B., Turmon, M., Williams, R., Beshore, E., & Larson, S. 2011, "Discovery, Classification, and Scientific Exploration of Transient Events From the Catalina Real-Time Transient Survey", *Bull. Astr. Soc. India*, **39**, 387.

Mahabal, A., Donalek, C., Djorgovski, S.G., Drake, A., Graham, M., Williams, T.,Chen, Y., Moghaddam, B., & Turmon, M. 2012, "Real Time Classification of Transient Events in Synoptic Sky Surveys", in: *Proc. IAU Symp. 285, New Horizons in Time Domain Astronomy*, eds. E. Griffin et al., Cambridge: Cambridge Univ. Press, **355**.

Murphy, T., et al. (the VAST team) 2012, "VAST: An ASKAP Survey for Variables and Slow Transients", *Publ. Astr. Soc. Australia*, in press.

Philip, N., Mahabal, A., Abraham, S., Williams, R., Djorgovski, S.G., Drake, A., Donalek, C., & Graham, M. 2012, "Classification by Boosting Differences in Input Vectors", in: *Proc. International Workshop on Stellar Spectral Libraries*, eds. P. Prugniel & H. P. Singh, *Ast. Soc. India Conf. Ser*., in press.

Rebbapragada, U., Lo, K., Wagstaff, K.L., Reed, C., & Murphy, T., 2012, "Offline and Online Classification of Simulated VAST Transients", VAST Memo No. 5.

Thompson, D.R., Brisken, W., Burke-Spolaor, S., Deller, A., Majid, W.A., Tingay, S., & Wagstaff, K.L., 2012, Real-time learning for adaptive weak event detection in data streams – lessons from the V-FASTR project *IEEE Intelligent Systems, manuscript in preparation.*

Thompson, D.R., Majid, W.A., Reed, C., & Wagstaff, K.L., 2012, "Semi-supervised novelty detection with adaptive eigenbases, and application to radio transients", *Statistical Analysis and Data Mining,* in press.

Thompson, D.R., Majid, W.A., Reed, C., & Wagstaff, K.L., 2011, "Semi-supervised novelty detection with adaptive eigenbases, and application to radio transients", in: *Statistical Analysis and Data Mining*, ref. proc. *CIDU 2011* conf., eds. A. Srivasatva, N. Chawla, & A. Perera, NASA Ames Res. Ctr. (Best Paper Award).

Vallisneri, M. 2011, "Beyond Fisher: exact sampling distributions of the maximum-likelihood estimator in gravitational-wave parameter estimation", *Phys. Rev. Lett.*, **107**, 191104

Wagstaff, K.R., Thompson, D.R., & Dietterich T.G., 2012, "Eyes Wide Open: Iterative Discovery in Large Data Sets without Premature Specialization", in prep.

Wayth, R., Tingay, S., Deller, A., Brisken, W., Thompson, D.R., & Wagstaff, K.L., 2012, "Limits on the event rates of fast radio transients from the V-FASTR experiment", *Astrophys. J.* in press.

8.2 Conference Presentations

Listed here if there was no conference proceedings paper, such as those listed in Sec. 8.1.

Cutler, C., "Improved version of the F-statistic for more efficient GW pulsar searches", Texas Symposium on Relativistic Astrophysics, Sao Paulo, Brazil, Dec. 2012.

Djorgovski, S.G. et al., "Automated Event Classification in Synoptic Sky Surveys", Computational science Workshop, Pucon, Chile, Aug. 2011.

Djorgovski, S.G. et al., "Some Discovery and Classification Challenges in the Exploration of Massive Data Sets and Data Streams", Progress on Statistical Issues in Searches, SLAC/Stanford, June 2012.

Djorgovski, S.G. et al., "Exploring the Time Domain: Some Things We Have Learned and Some Strategic Considerations", IAU Symp. 285, Oxford, UK, Sept. 2011.

Djorgovski, S.G. et al., "Virtual Observatory in a Broder Context of e-Science, KDD/DM Needs and Resources", IVOA KDD IG, Sao Paulo, Brazil, Oct. 2012.

Djorgovski, S.G. et al., "Exploring the Time Domain", Future Science With Metre-Class Telescopes, Belgrade, Serbia, Sept. 2012.

Djorgovski, S.G. et al., "e-Science, Virtual Observatory, and Astroinformatics: Enabling the Data-Intensive Astronomy for the 21st Century", SKA/MeerKAT HQ, Cape Town, South Africa, Apr. 2012

Djorgovski, S.G. et al., "Exploration of the Time Domain", Caltech Time Domain Forum, Nov. 2011.

Djorgovski, S.G. et al., "Exploring the Variable Sky with the Catalina Real-Time Transient Survey", NAOC Colloquium, Beijing, China, Nov. 2011.

Djorgovski, S.G. et al., "Exploring the Time Domain with the Catalina Real-Time Transient Survey", Univ. Federico II, Napoli, Italy, May 2012.

Djorgovski, S.G., Longo, G., Brescia, M., Donalek, C., Cavuoti, S., Paololllo, M., D'Abrusco, R., Laurino, O., Mahabal, A., & Graham, M. 2012, "DAta Mining and Exploration (DAME): New Tools for Knowledge Discovery in Astronomy", American Astronomical Society Meeting Abstracts

Donalek, C., Fang, K., Drake, A., Djorgovski, S.G., Graham, M., Mahabal, A., & Williams, R. 2011: "SkyDiscovery: Humans and Machines Working Together", American Astronomical Society Meeting Abstracts, 217, #334.02.

Donalek, C., Graham, M.J. , Mahabal, A., Djorgovski, S.G., Drake, A.J., Moghaddam, B., Turmon, M.J., Sharma, N., & Chen, Y., 2012, "Lightcurve Based Classification of Transient Events", American Astronomical Society Meeting Abstracts, #219, #446.16.

Graham, M., Conwill, L., Djorgovski, S.G., Mahabal, A., Donalek, C., & Drake, A. 2011, "Extracting Meaning From Astronomical Telegrams", American Astronomical Society Meeting Abstracts, 217, #344.11

Graham, M., Zhang, M., Djorgovski, S.G., Drake, A., Donalek, C., & Mahabal, A. 2012, "Constructing Concept Schemes From Astronomical Telegrams Via Natural Language Clustering", American Astronomical Society Meeting Abstracts.

Graham, M.J., “The transient sky and the VO”, Brazilian Astronomical Society meeting, Brazil, Oct. 2012.

Masci, F.J. & Hoffman, D., “Imtrandetect: a new tool/methodology for detecting astronomical transients in large image-data streams down to low S/N”, 219th AAS Meeting, Austin TX, Jan. 2012.

Thompson, D.R., “Machine Learning for Exploring Data Streams”, SETI Institute Talk, Mountain View, CA, 2012.

Thompson, D.R., “On-line data mining and event detection in petascale data streams”, Special session on Cyber-Discovery and Science for the Decade, 219th AAS Meeting, Austin TX, Jan. 2012.

Vallisneri, M., “Bayesian methods for stochastic gravitational-wave detection in pulsar-timing data (and more)”, North American Nanohertz Observatory for Gravitational Waves online seminar, Oct 2012.

Vallisneri, M., “Between Fisher and Monte Carlo: mapping the distribution of the maximum-likelihood estimator in gravitational-wave observations”, *Gravitational Wave Physics and Astronomy Workshop*, Hannover, Germany, Jun 2012.

Vallisneri, M., “All the rest is noise: gravitational waves and statistical inference”, astrophysics seminar, NYU, New York, Oct 2011.

Vallisneri, M., “Space-based GW astronomy in the next decade, with a coda of statistics”, astrophysics seminar, Cambridge, UK, Jul 2011.

Vallisneri, M., “Mapping the distribution of the maximum-likelihood estimator for GW source parameters”, *CaJAGWR* seminar, Caltech, May 2011.

References:

Babu, G.J., Mahabal, A., Djorgovski, S.G., & Williams, R. 2006, "Object Detection in Multi-Epoch Data", in *Proc. Astronomical Data Analysis IV*, eds. J.-L. Starck & T. Loredo, *Stat. Methodology*, **5**, 299.

Bailey, S., et al. 2007, "How to Find More Supernovae with Less Work: Object Classification Techniques for Difference Imaging", *Astroph. J.,* **665**, 1246.

Beim, A., Elmegreen, B. et al. 2010, "A streaming approach to Radio Astronomy Imaging", *Proceedings of ICASSP 2010.*

Berger, E., Ball, S., Becker, K., et al. 2001, "Discovery of Radio Emission from the Brown Dwarf LP944-20", *Nature*, **410**, 338.

Bloom, J., & Richards, J. 2011, "Data Mining and Machine-Learning in Time-Domain Discovery & Classification", in: *Advances in Machine Learning and Data Mining for Astronomy*, in press; arXiv/1104.3142.

Brady, P.R., Creighton, T., Cutler, C., & Schutz, B.F. 1998, PRD, **57**, 2101.

Brady, P.R., & Creighton, T. 2000, PRD, **61**, 082001.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1984, "*Classification and regression trees".* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Cutler, C., et al. 2005, "Improved Stack-Slides Searches for Gravitational-Wave Pulsars", *Phys Rev D*, **72**, 042004.

desJardins, M.. MacGlashan, J., & Wagstaff, K.L. 2010 "Confidence-Based Feature Acquisition to Minimize Training and Test Costs", in *Proceedings of the SIAM Conference on Data Mining*, p. 514-524.

Djorgovski, S. G., et al. 2001a, "Exploration of Parameter Spaces in a Virtual Observatory", in: *Astronomical Data Analysis*, eds. J.-L. Starck & F. Murtagh, *Proc. SPIE*, **4477**, 43.

Djorgovski, S. G., et al. 2001b, "Exploration of Large Digital Sky Surveys", in: *Mining the Sky*, eds. A.J. Banday et al., ESO Astrophysics Symposia, p. 305, Berlin: Springer Verlag.

Djorgovski, S. G., et al. 2006, "Some Pattern Recognition Challenges in Data-Intensive Astronomy", *Proc. ICPR 2006*, Vol. **1**, IEEE Press, p. 856.

Djorgovski, S.G., et al. (PQ survey team) 2008, "The Palomar-Quest digital synoptic sky survey", *Aston. Nachrichten*, **329**, 263.

Djorgovski, S.G., Donalek, C., Mahabal, A., Moghaddam, B., Turmon, M., Graham, M., Drake, A., Sharma, N., & Chen, Y. 2011, "Towards an Automated Classification of Transient Events in Synoptic Sky Surveys", in: *Proc. Conf. on Intelligent Data Understanding 2011*, eds. A. Srivasatva, N. Chawla, & A. Perera, p. 174, Mountain View, CA: NASA Ames Res. Ctr.

Djorgovski, S.G., Drake, A., Mahabal, A., Graham, M., Donalek, C., Williams, R., Beshore, E., Larson, S., Prieto, J., Catelan, M., Christensen, E., & McNaught, R. 2012a, "The Catalina Real-Time Transient Survey (CRTS)", in: *The First Year of MAXI: Monitoring Variable X-ray Sources*, eds. T. Mihara & M. Serino, Special Publ. IPCR-127, p. 263, Tokyo: RIKEN.

Djorgovski, S.G., Mahabal, A., Drake, A., Graham, M., Donalek, C., & Williams, R., 2012b, "Exploring the Time Domain With Synoptic Sky Surveys", in: *Proc. IAU Symp. 285: New Horizons in Time Domain Astronomy*, eds. E. Griffin et al., p. 141. Cambridge: Cambridge Univ. Press.

Djorgovski, S.G., Mahabal, A., Drake, A., Graham, M., & Donalek, C. 2012c, "Sky Surveys", in: *Astronomical Techniques, Software, and Data*, ed. H. Bond, Vol.2 of *Planets, Stars, and Stellar Systems*, ser. ed. T. Oswalt, Berlin: Springer Verlag, in press.

Donalek, C. et al. 2008, "New Approaches to Object Classification in Synoptic Sky Surveys", in *AIP Conf. Ser.,* **1082**, 252.

Drake, A., Djorgovski, S.G., Mahabal, A., et al. (the CRTS team) 2009, "First Results from the Catalina Real-time Transient Survey", *Astroph. J.,* **696**, 870.

Dubath et al., 2011, *MNRAS,* **414**, 2602.

Edelson, Krolik, 1988, "The Discrete Correlation Function, a New Method for Analyzing Unevenly Sampled Variability Data", *Astroph. J.*, **333**, 646.

Gair, J.R., et al. 2004, *Class. Quant. Grav.,* **21**, S1595.

Griffin, E., Hanisch, R., & Seeman, R. (eds.) 2012, in *Proc. IAU Symp. 285: New Horizons in Time Domain Astronomy*, Cambridge: Cambridge Univ. Press.

Hallinan, G., Bourke, S., Lane, C., et al. 2007, "Periodic Bursts of Coherent Radio Emission from an Ultracool Dwarf", *Astroph. J.*, **663**, L25.

Huise, P., et al., 2011, IEEE Signal Processing Letters, **18**, #6, 371.

Hyman, S., Lazio, T.J.W., Kassim, N., Ray, P., Markwardt, C., & Yusef-Zadeh, F. 2005, "A Powerful Bursting Radio Source Towards the Galactic Centre", *Nature*, **434**, 50.

Ivezic, Z., et al. (the LSST team) 2009, *The LSST Science Book*, v2.0, arXiv:0912.0201, http://www.lsst.org/lsst/scibook.

Kaiser, N. 2004, "Pan-STARRS: a wide-field optical survey telescope array", in: G*round-based Telescopes*, ed. J. Oschmann, *Proc. SPIE*, **5489**, 11.

Kozlowski and Kochanek, 2009, *Astroph. J.*, **701**, 508

Ling, C.X., Yang, W., Wang, J. & Zhang, S. 2004, "Decision trees with minimal costs," in *Proceedings of the Twenty-first International Conference on Machine Learning*, p. 544-551.

Mahabal, A. et al. 2008a, "Automated Probabilistic Classification of Transients and Variables", *Astron. Nachr*., **329**, 288.

Mahabal, A., Djorgovski, S.G., Williams, R., Drake, A., Donalek, C., Graham, M., Moghaddam, B., Turmon, M., Jewell, J., Khosla, A., & Hensley, B. 2008b, "Towards Real-Time Classification of Astronomical Transients", in *AIP Conf. Ser*., **1082**, 287.

Mahabal, A., Wozniak, P., Donalek, C., & Djorgovski, S.G. 2009, "Transients and Variable Stars in the Era of Synoptic Imaging", in: *LSST Science Book*, eds. Z. Ivezic, et al., Ch. 8.4, p. 261; available at http://www.lsst.org/lsst/scibook.

Mahabal, A., Djorgovski, S.G., Donalek, C., Drake, A., Graham, M., Moghaddam, B., Turmon, M., & Williams, R. 2010a, "Mixing Bayesian Techniques for Effective Real-time

Classification of Astronomical Transients", in *Proc. ADASS XIX*, ed. Y. Mizumoto, *A.S.P. Conf. Ser*., **434**, 115.

Mahabal, A., et al. 2010b, "Classification of Optical Transients: Experiences from PQ and CRTS Surveys", in *Gaia: At the Frontiers of Astrometry*, eds. C. Turon, et al., *EAS Publ. Ser.* **45**, 173, Paris: EDP Sciences.

Mahabal, A., et al. 2010d, "The Meaning of Events", in: *Hotwiring the Transient Universe*, eds. S. Emery Bunn, et al., Lulu Enterprises Publ. http://www.lulu.com/, p. 31.

Mahabal, A., Djorgovski, S.G., Drake, A., Donalek, C., Graham, M., Moghaddam, B., Turmon, M., Williams, R., Beshore, E., & Larson, S. 2011, "Discovery, Classification, and Scientific Exploration of Transient Events From the Catalina Real-Time Transient Survey", *Bull. Astr. Soc. India*, **39**, 387.

Mahabal, A., Donalek, C., Djorgovski, S.G., Drake, A., Graham, M., Williams, R., Chen, Y., Moghaddam, B., & Turmon, M. 2012, "Real Time Classification of Transient Events in Synoptic Sky Surveys", in *Proc. IAU Symp. 285: New Horizons in Time Domain Astronomy*, eds. E. Griffin et al., p. 355, Cambridge: Cambridge Univ. Press.

Majid, W. 2012, Submitted.

McLaughlin, M., Lyne, A., Lorimer, D., et al. 2006, "Transient radio bursts from rotating neutron stars", *Nature*, **439**, 817.

Meinshausen, N., Bickel, P., & Rice, J. 2009, Ann. Appl. Stat., **3**, 38.

Nutzman, P., & Charbonneau, D. 2008, in *PASP*, **120**, 317

Paczynski, B. 2000, "Monitoring All Sky for Variability", in *PASP*, **112**, 1281.

Prix, R. & Shaltev, M. 2012, PRD, **85**, 084010.

Protopapas, 2006, *MNRAS*, **369**, 677.

Protopapas, 2012, submitted.

Romano, R., Aragon, C. & Ding, C. 2006, "Supernova Recognition using Support Vector Machines", LBNL-61192, in *Proc. 5th Int'l. Conf. Machine Learning Applications*.

Scargle, J., 2005, "An Algorithm for the Optimal Partitioning of Data on an Interval", *IEEE Signal Processing Letters*, **12**, 105.

Schmidt, M. & Lipson H., 2009, *Science*, **324**, 81.

Thompson, D.R., Wagstaff, K.L. et al 2011a, "Detection of fast radio transients with multiple stations: A case study using the Very Long Baseline Array", *Astroph. J.*, **735**(2).

Thompson D.R., Majid, W.A., Reed, C.J. & Wagstaff K.L. 2011b, "Semi-Supervised Novelty Detection with Adaptive Eigenbases, and Application to Radio Transients", in *Proceedings of the Conference on Intelligent Data Understanding.*

Thompson D.R., Majid, W.A., Reed, C.J. & Wagstaff K.L. 2012, "Semi-Supervised Eigenbasis Novelty Detection", *Statistical Analysis and Data Mining*, in press.

Tyson, J.A. 2002, "Large Synoptic Survey Telescope: Overview", in: *Survey and Other Telescope Technologies and Discoveries*, eds. J.A. Tyson & S. Wolf, *Proc. SPIE*, **4836**, 10.

Vanderlooy, S., Sprinkhuizen-Kuyper, I.G., Smirnov, E.N., & van den Herik, J. H. 2009 "The ROC isometrics approach to construct reliable classifiers", *Intelligent Data Analysis*, **13**, 3–37.

Wang, 2012, Submited (KDD), arXiv/1203:0970

Appendix A: Workshop Participants

A joint list of participants in the opening and closing workshops, with their affiliations, listed alphabetically. Some of them attended only a part of the technical workshops and study discussions. In addition, several tens of other scientists and students have attended the open public parts of the workshops.

- Jogesh Babu Gutti - The Pennsylvania State University
- Guillermo Cabrera - University of Chile
- Curt J. Cutler - Jet Propulsion Laboratory, Caltech
- Raffaele D'Abrusco - Harvard-Smithsonian Center for Astrophysics
- Rosanne Di Stefano - Harvard-Smithsonian Center for Astrophysics
- George Djorgovski - California Institute of Technology
- Ciro Donalek - California Institute of Technology
- Andrew Drake - California Institute of Technology
- Bruce G. Elmegreen - IBM Research Division
- Matthew J. Graham - California Institute of Technology
- Pablo Huijse - University of Chile
- Laurent Eyer – University of Geneva, Switzerland
- Vinay Kashyap - Harvard-Smithsonian Center for Astrophysics
- Badri Krishnan - Albert Einstein Institute, Germany
- Joseph Lazio - Jet Propulsion Laboratory, Caltech
- Giuseppe Longo - University Federico II, Napoli, Italy
- Ashish Mahabal - California Institute of Technology
- Frank J. Masci - IPAC, California Institute of Technology
- Walter Max-Moerbeck - California Institute of Technology
- Baback Moghaddam - Jet Propulsion Laboratory, Caltech
- Pavlos Protopapas - Harvard-Smithsonian Center for Astrophysics
- Umaa D. Rebbapragada - Jet Propulsion Laboratory, Caltech
- John A. Rice - University of California, Berkeley
- Graca Rocha - Jet Propulsion Laboratory, Caltech
- Jeff D. Scargle - NASA Ames Research Center
- Mark Stalzer - California Institute of Technology
- David R. Thompson - Jet Propulsion Laboratory, Caltech
- Mike Turmon - Jet Propulsion Laboratory, Caltech
- Michele Vallisneri - Jet Propulsion Laboratory, Caltech
- Eduardo S. Vera - University of Chile
- Kiri L. Wagstaff - Jet Propulsion Laboratory, Caltech
- Yan Xu – Microsoft Research

Dr. Jeff Scargle was also a KISS Distinguished Visiting Scholar for the duration of this study.

In memoriam: During the process of preparation of this report, one of our participants, Dr. Baback Moghaddam, tragically passed away after a brief illness. His contributions live on.

Appendix B: Selected Acronyms and the Associated Websites

2MASS = Two Micron All-Sky Survey, http://www.ipac.caltech.edu/2mass
AGN = Active Galactic Nucleus
ALMA = Atacama Large Milllimeter/Submillimeter Array, http://www.almaobservatory.org
ANN = Adaptive Neural Network
ASKAP = Australian Square Kilometer Array Pathfinder,
http://www.atnf.csiro.au/projects/askap
ATA = Allen Telescope Array, http://www.seti.org/ata
BB = Bayesian Blocks
BN = Bayesian Network
CFA = Confidence-based Feature Acquisition
CKP = Correntropy Kernelized Periodogram
CoRoT = Convection Rotation and Planetary Transits, http://smsc.cnes.fr/COROT/
CPU = Central Processing Unit
CRTS = Catalina Real-Time Transients Survey, http://crts.caltech.edu
CSDT = Cost-Sensitive Decision Tree
CV = Cataclysmic Variable
DAG = Directed Acyclic Graph
DEMUD = Discovery through Eigenbasis Modeling of Uninteresting Data
DM = Dispersion Measure; Data Mining
DT = Decision Tree
EIF = Environmental Informatics Framework
EMRI = Extreme Mass Ratio Inspiral
EROS = Expérience pour la Recherche d'Objets Sombres, http://eros.in2p3.fr/
EVLA = Expanded Very Large Array, http://www.aoc.nrao.edu/evla/
GAIA = http://gaia.esa.int
GALEX = Galaxy Evolution Explorer, http://galex.caltech.edu
GBT = Green Bank Telescope, https://science.nrao.edu/facilities/gbt/
GDI = Gini Diversity Index
GPU = Graphics Processing Unit
GRB = Gamma-Ray Burst
GW = Gravitational Wave
HMM = Hidden Markov Model
HST = Hubble Space Telescope, http://www.stsci.edu/hst
Kepler = http://kepler.nasa.gov/
LCOGT = Las Cumbres Observatory Global Telescope Network, http://lcogt.net
LIGO = Laser Interferometry Gravitational Observatory, http://ligo.caltech.edu
LISA = Laser Interferometer Space Antenna, http://lisa.nasa.gov
LOFAR = Low Frequency Array, http://www.lofar.org/
LSST = Large Synoptic Survey Telescope, http://www.lsst.org
ITL = Information-theoretic Learning Framework
LC = Light curve
MACHO = Massive Astrophysical Compact Halo Object, http://wwwmacho.anu.edu.au/
MCMC = Markov Chain Monte Carlo
MeerKAT, http://www.ska.ac.za/meerkat/index.php

ML = Machine Learning
NuSTAR = Nuclear Spectroscopic Telescope Array, http://www.nustar.caltech.edu
PanSTARRS = Panoramic Survey Telescope and Rapid Response System, http://pan-starrs.ifa.hawaii.edu
PCA = Principal Component Analysis
PQ = The Palomar-Quest Sky Survey, http://palquest.org
PSF = Point Spread Function
PTF = Palomar Transient Factory, http://www.astro.caltech.edu/ptf
RFI = Radio Frequency Interference
SED = Spectral Energy Distribution
SETI = Search for Extraterrestrial Intelligence, http://www.seti.org
SKA = Square Kilometer Array, http://www.skatelescope.org
S/N, SNR = Signal-to-noise ratio
SNE = Supernova
SSEND = Semi-Supervised Eigenbasis Novelty Detection
SVM = Support Vector Machine
SuperWASP = Super Wide Angle Search for Planets, http://www.superwasp.org/
VLA = NRAO Very Large Array, http://www.vla.nrao.edu
VLBA = Very Long Baseline Array, http://www.vlba.nrao.edu
VR = Virtual Reality
WFIRST = Wide-Field Infrared Survey Telescope, http://wfirst.gsfc.nasa.gov
WISE = Wide-Field Infrared Survey Explorer, http://wise.ssl.berkeley.edu