I. Introduction

The field of time–domain astrophysics is currently experiencing a renaissance, driven by the fusion of large datasets, computational infrastructure, and astro–statistical tools. Research efforts may be broadly cast into two types of analyses – data–mining efforts (discovery) and follow–up of interesting events (characterization). The most ambitious of these efforts utilize real–time discovery tools on survey data streams to drive autonomous follow–up resources. As these data streams become increasingly more vast and complex, the statistical tools required to efficiently drive these resources must themselves increase in complexity. They must be sensitive to all quadrants of the Rumsfeldian "known" vs. "unknown" characterization scheme, with the majority of phenomena falling in the "known known" category, and the most interesting of events falling within the "unknown unknown". To optimally do this sifting, the complexity of the models much match the complexity of the data stream to enable a single classification assessment.

In this proposal, we will expand upon the current state–of–the art in event classification to generate astrophysical variability models with the same dimensionality as the data being collected. Next generation time–domain data streams will provide constraints on a given event at irregular times, and in one of several passbands. Wide–field time–domain spectroscopic surveys are also on the horizon. The event models themselves must therefore be inherently temporal *and* spectral, to federate the ensemble of survey and follow–up data into a single statistical assessment of event type. The building of these models requires adopting a unifying description of astrophysical variability. **We propose sets of spectral–temporal surfaces, generated for all event types, as the optimal description of astrophysical variability.** The process will incorporate the diversity of existing data into a single statistical model representing the mean behavior of each event class, as well as higher–order moments about this mean that reflect the intrinsic dimensionality of the phenomena. These models will be optimal in the sense that they incorporate extant multi–passband knowledge (photometric *and* spectroscopic) into an empirically–derived event model. They will be optimal in these sense that they represent each class of phenomena at all times and integrable over any photometric passband, a level of complexity needed to evaluate the event streams of next–generation surveys such as the Large Synoptic Survey Telescope (LSST)[1]. And they will be optimal in the sense that they will unify the statistical description of astronomical variability from today's heterogeneous standards into a common language. To provide the broadest possible impact from this effort, we will build a classification infrastructure incorporating these models into a real–time event broker that will be available to the astronomical community.

A Brief History of Time–Domain Astronomy

The field of time–domain astronomy began in earnest in the 1990s with wide–field (10 square degree) microlensing surveys such as MACHO (Alcock et al., 2000), OGLE (Udalski et al., 1994), and EROS (Afonso et al., 2003). Early on, team members recognized that there was value in a quick concerted response to on–going events: followup after the event completed would yield none of the exotic effects such as parallax, resolution of the lensed star's disk, or lens or source binarity. The microlensing surveys implemented rapid alert

[1] http://www.lsst.org

streams that fed (through emails or phone calls) into follow–up networks such as GMAN (Becker, 2000) and PLANET (Albrow et al., 1998), which resolved for the first time many of the exotica expected of complex gravitational lens systems.

Subsequent time–domain efforts turned their focus to precision cosmology measurements through surveys for Type Ia supernovae (Hamuy et al., 1996), which most prominently yielded the discovery of the acceleration of the expansion of the Universe (Riess et al., 1998; Perlmutter et al., 1999). These surveys also required quick (∼few day) analysis of their data to acquire spectroscopic redshifts of the events. In these cases, the most efficient use of resources came both by discovery of the event *and* with an estimate of the event type (Ia or otherwise). Contextual clues were used to assist in event classification, including proximity to host galaxy and host galaxy type.

Gamma-ray bursts, intense but evanescent flashes from exploding and colliding stars, are observable across the electromagnetic spectrum but only if discovered, disseminated, and followed up very rapidly (timescale of seconds). Intense multi-color imaging coupled with synoptic spectroscopy in the minutes to hours after events are the state-of-the-art in the field. The most precious (scientifically) tend to be those events from the lowest and highest redshift, providing vistas on star formation, dust obscuration, and potentially serving as the bridge between the electromagnetic and gravity wave landscapes. However, at a discovery rate of ∼100/yr (primarily from NASA/Swift), the community cannot adequately followup all events with the available resources for target-of-opportunity observations. Knowing *a priori* which GRBs are likely to have the largest scientific return given substantial telescope investment is crucial. We (Morgan et al., submitted) have begun to using machine learning tools to try to predict high-redshift events from immediately available burst data, couching the prediction as a resource maximization problem.

In the early 2000s surveys began to categorize and release *all* types of events found in their data, in near real–time. This includes the Deep Lens Survey (Becker et al., 2004) and the Faint Sky Variability Survey (Groot et al., 2003). In these cases, coarse attempts were made at event classification for all objects displaying astrometric or photometric variability. Primarily contextual information were used (e.g. ecliptic latitude, distance from host galaxy) and classifications were done by humans at the telescope or remotely using web–based visual classification tools. More recent time–domain surveys such as the Palomar–Quest survey (Djorgovski et al., 2008) and Catalina Real-Time Transient Survey (Djorgovski et al., 2011b) have begun to adopt modern statistical techniques to classify events. This includes morphological classification of the pixel–level flux that triggered the event (Donalek et al., 2008), as well as contextual information regarding the location of the variability (Mahabal et al., 2010). We note that colors of the quiescent objects are used in these classification schemes, but *not* the time–varying color of the variable flux.

In the NSF-sponsored Palomar Transient Factory, we have begun to use machine learning techniques to discover new events through image differencing, identifying roughly 1000 variable stars and transients out of 1.5 million candidates per night. These events are, in real-time, classified with an ML classifier that makes use of contextual information (such as color of the nearest galaxy) and temporal metrics (such as difference in magnitude between the event and the quiescent counterpart brightness). We have shown (Bloom et al., 2011) the ability to distinguish between transients and variable stars with a 3.8% overall

error rate (with 1.7% errors for imaging within the Sloan Digital Sky Survey footprint). At >96% classification efficiency, the samples achieve 90% purity. Determining just what sort of transient is found (e.g., supernova, nova, QSO event) has proven more challenging. Only with retrospective analysis, once sufficient data on a light curve has been obtained, that the classification errors begin to drop significantly (Richards et al., 2011) (see Bloom & Richards 2011 for a review).

In all cases mentioned above, the volume of data being collected made the surveys sensitive to new types of astrophysical phenomena, with a rapid response component that enabled immediate and early study. It is also the case that many of these efforts designed their event filters based upon the particular survey design they were being applied to. To continue this trend into the future, where surveys such as Gaia[2] and LSST will report on *all* classes of variability, a class of models needs to be generated that match the ambitious designs of the surveys (massive data rates, sparse sampling in time, inhomogeneous sampling in wavelength). This is particularly important for driving the next generation of autonomous follow–up resources, such as LCOGT (Hidas et al., 2008), that must algorithmically sift through the data streams to achieve their particular science goals.

Strides Towards Spectral–Temporal Event Classification

While teams such as the Harvard Time Series Center[3] and the Berkeley Transients Classification Pipeline[4] have made substantial advances in automated event classification, their methods remain inherently one–dimensional. That is, their models of event behavior are relevant for data in a single passband, but do not take into account color evolution during the events.

One field where there has been substantial progress in spectral–temporal modeling is in the field of supernova studies. The pioneering work of Nugent et al. (2002) integrated a large, inhomogeneous sample of spectroscopy of Ia supernovae into a single averaged representation of the Ia phenomena. This integrated model represented the behavior of a typical "Branch–normal" Type Ia supernova, and defined the spectrum of a typical event at all integer wavelengths between 1000 Å and 25000 Å, 1 per day for 90 days after explosion. Nugent subsequently expanded his template set to include separate models for intrinsically bright and faint Ia, supernovae Type Ib/c, and Types II P/L/n. Such templates played a crucial role in real–time event classification for the SDSS-II Supernova Survey (Frieman et al., 2008), and enabled a 90% targeting efficiency for Type Ia supernova (Sako et al., 2008). However, since they only existed for supernova–class events, all events that did not conform to the spectral and temporal behavior of these models were subsequently ignored.

This notion of spectral–temporal event representations was expanded upon by Guy et al. (2007) in the generation of their SALT–II supernova model (see also Hsiao et al. (2007)). This model incorporated spectroscopic *and* photometric data from low– and high–redshift supernovae to yield a set of lightcurve templates describing the behavior of *all* Type Ia supernovae. The spectroscopic data constrain the templates sparsely in time (being acquired infrequently compared to photometry) but finely in wavelength, while the photometric data

[2]http://gaia.esa.int
[3]http://timemachine.iic.harvard.edu/
[4]http://dotastro.org/about/tcp.php

constrain the templates more densely in time, but only coarsely in wavelength. The photometric data were also able to be calibrated, in an absolute sense, to a much higher accuracy than the spectroscopic data, and were used to bootstrap the overall calibration of the model. These data in combination yielded a highly constrained two–dimensional surface in time and wavelength that describes the temporal evolution of the rest-frame spectral energy distribution (SED) for SNe Ia. Importantly, the SALT–II model yields not just the average behavior of the Type Ia sample, equivalent to the Nugent et al. (2002) data, but also includes a first "principal component" about the mean, which captures the spectral–temporal behavior of the diversity (intrinsically bright to intrinsically faint) of Ia. This single model – mean surface plus a first moment – contains the behaviors and correlations seen in the Ia population such as the Δm_{15} brightness–decline (Phillips, 1993) or brightness–stretch (Riess et al., 1998) relations. The need for only a single principal component reflects the fact that Ia supernova are intrinsically a one–parameter family; the SALT–II model captures this in a spectral–temporal representation.

Figure 1 displays the mean surface (*left*) and first principal component surface (*right*) from the SALT–II model. The behavior for a given Ia lightcurve may be generated by an overall scaling of the mean surface (representing distance modulus) with a contribution from the secondary surface representing its place within the 1–parameter family of Ia supernovae. Importantly, an incoming spectral–temporal data stream can be compared to these surfaces, and a statistical assessment made whether or not it is behaving in accordance with this model. This likelihood may be generated using the surfaces, the photometric uncertainty on the incoming data, and (optionally) priors on the amplitude of the second surface indicated by the best–fit. This is the process we envision resulting from this proposal : **an incoming event stream will be evaluated against an ensemble of these surfaces to yield a statistical likelihood that the event is of each class.**

II. Proposed Work

We propose to make spectral–temporal (ST) surfaces for *all* major classes of astronomical variability. The individual components of this effort include:

- the aggregation of photometric and photometrically–calibrated spectroscopic data from a diverse range of input sources;
- the generation of a spectral–temporal variability model from all data on a given class;
- an investigation into the dimensionality of each class, to determine how many principal components to include in the model;
- a statistical framework to generate likelihoods that an incoming event stream is consistent with each variability model;
- and an on–line classification service that includes our spectral–temporal models as part of an overall (single–epoch and multi–epoch) event assessment.

These surfaces will serve as common currency in future event classification efforts, alongside extant efforts such as: Artificial Neural Network (Djorgovski et al., 2008), Support

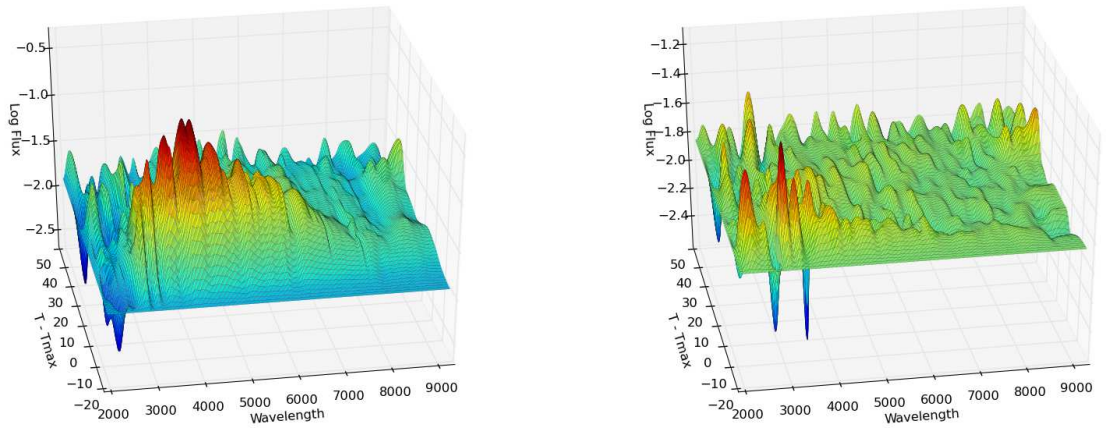

Figure 1: Spectral–temporal surfaces of Type Ia supernovae from the SALT–II model of Guy et al. (2007). Surfaces are generated from the aggregation of photometric and spectroscopic data on several hundred vetted Ia supernova. The leftmost figure represents the mean behavior of the sample, defined every 10 Å between 2000 Å and 9200 Å, and in daily intervals from -20 days from B–band peak brightness to +50 days. The right figure displays the first moment of these data about the mean distribution, derived using principal component analysis. The diversity of Ia lightcurves (across all passbands and at all times) can be represented by a scaled addition of this secondary surface to the primary surface, reflecting the observation that Ia are intrinsically a 1–parameter family. **We propose here to make similar models for *all* classes of astronomical variability, which will serve as the common currency in general event classification efforts.**

Vector Machine (Bailey et al., 2007), and human–vetted (Bloom et al., 2011) classifiers of pixel–level artifacts; contextual understanding of new single–epoch events using Markov Logic Networks (Djorgovski et al., 2011a) and feature–based classifiers (Bloom et al., 2011); and shape–based Random Forest (Richards et al., 2011) classifiers of lightcurves in a single passband. Our effort will add an extra dimensionality to these state–of–the–art classification efforts, allowing the incorporation of information at multiple epochs and wavelengths into the overall classification scheme. Below we describe in detail each step in this process.

Aggregation of Data

Explore classes.

The first step in this process is to aggregate the diversity of time–domain data above phenomena. Since we will be incorporating data taken on different instruments, in different filters, and at different epochs, we must enforce a strict set of standards to ensure we can merge these data into a single ST model. In practice, this will limit our sources of data to archival resources where calibration metadata is included in the data curation.

All incoming photometric data must be accompanied by a filter profile corresponding to the transmission profile of the filter of observation, as well as a time stamp corresponding to the epoch of observation. We will convert each filter profile into an overall atmosphere plus system throughput representing the dimensionless probability that a photon of wavelength λ reaches the detector. This filter profile will be used to weight each data point's contribution to the ST surface. We will also use each filter profile to convert (if necessary) the flux into the AB magnitude system. We will convert each input time stamp to correspond to the

mid–point of observation, in Barycentric Julian Date in the Barycentric Dynamical Time standard (Eastman et al., 2010).

All incoming spectroscopic data will have the same requirement on epoch of observation, with the additional requirement that it has been spectrophotometrically calibrated. This is difficult.

An additional requirement for objects at cosmological distances is that any photometric data must be accompanied by a redshift to translate the observer–frame filter profile into a rest–frame window. This has the disadvantage that K–correction of the photometry (Hogg et al., 2002) requires an instantaneous spectrum of the event, which leads to a chicken–and–egg problem since that exact spectrum is expected to result from the ST modeling. This will require a boot–strapping of the cosmological K–corrections with an initial ST surface. However, these K–corrections will allow us to build ST surfaces that are well–calibrated in the rest–frame u–band, where ground–based calibration is difficult, by using photometry of objects at moderate redshift.

Identified input data sources include:

- ■ **andy:** These could also be put in a normal paragraph to save space. ■
- ■ **andy:** ASAS? ■
- SDSS Stripe–82 where we will use the six *ugriz* per–camera column filter profiles for SDSS (Ivezić et al., 2007).
- Variable stars from the OGLE and Hipparcos surveys cataloged by Debosscher et al. (2007).
- SDSS Stripe–82 supernova
- Carnegie SN project
- The on–line supernova spectrum archive
- Saurabh's data ■ **andy:** ?? ■
- Time–domain spectroscopy projects TDSS if it happens.
- PTF
- ■ **andy:** Probably also want to include some language about the inclusion of theoretical curves of known unknowns. There are pair instability models, kilonovae, etc. These will be important to find. ■

Generation of Spectral–Temporal Surfaces

We proceed using the following underlying model for astrophysical variability: the flux from an astronomical object may be represented as $F_\nu(\lambda, t)$ with F_ν the specific flux of an object above the atmosphere, in Janskys (1 Jy $= 10^{-23}$ erg cm^{-2} s^{-1} Hz^{-1}), as a function of wavelength and time. Our spectral–temporal models will be defined over the grid of λ

and t, with the bin sizes dependent on the amount of input data available. These models will be defined in the event rest–frame, and (to the best of our ability) above the Earths's atmosphere, meaning the input data will need to be calibrated in an absolute sense. Data on a single event will only coarsely sample this surface. However, the phase space will be more densely sampled by an ensemble of data of different events of the same class.

Between emission and measurement, F_ν is attenuated by a (typically unknown) atmospheric transmission function $T^{atm}(\lambda, t)$, the probability that a photon of wavelength λ successfully propagates through the atmosphere, and a (typically measured, and occasionally well–measured) system transmission probability per unit photon $T^{sys}_{filt}(\lambda, t)$. This latter term should include the wavelength dependence of the mirror reflectivity, lens and filter transmission, and detection sensitivity. The product of these two defines the overall transmission profile of a given observation in a given filter $T_{filt}(\lambda, t)$.

Many surveys provide only measurements of $T^{sys}_{filt}(\lambda, t)$ (e.g. Stubbs & Tonry, 2006) meaning we must synthesize the atmospheric transmission function for a given observation. This is a difficult task, since the atmospheric components that contribute to $T^{atm}(\lambda, t)$ (including water vapor, aerosol scattering, Rayleigh scattering and molecular absorption) may vary on the order of 10% per hour (Stubbs et al., 2007). So as to not fall down the rabbit hole in terms of atmospheric calibration, we will adopt fiducial MODTRAN atmospheres (Berk et al., 1999) at the airmass of observation (when reported) or adopt a fiducial airmass of 1.3 when not. We note that some surveys such as SDSS define their filter profiles with the atmospheric component rolled in (Ivezić et al., 2007).

The total counts transmitted through the optical system are:

$$F^{obs}_{filt}(t) = \int F_\nu(\lambda, t) \;\; T_{filt}(\lambda, t) \;\; \lambda^{-1} d\lambda$$

where $T_{filt}(\lambda, t)$ is the overall transmission probability per unit photon [5]. To place the final flux measurement on the AB magnitude system, this integral is normalized by $F_{AB} = 3631$ Jy.

From the available data on each input training event, we will create a sparse surface labelled $f(\lambda, t)$ that constrains the overall model. Spectral data will be rebinned to the wavelength resolution of our model, and photometric flux will added to the model across wavelengths with a weighting of $T_{filt}(\lambda, t)\ \lambda^{-1}$. To generate the ST model from these data, we proceed with the standard assumption that each input dataset $f_0(\lambda, t), f_1(\lambda, t)\ \ldots\ f_i(\lambda, t)$ may be recovered from a linear combination of (as–yet unknown) basis surfaces

$$f_i(\lambda, t) = \sum_{j=0} x_{ij} \times \mathbf{S_j}(\lambda, t)$$

where $\mathbf{S_j}(\lambda, t)$ are said basis surfaces and x_{ij} are their weightings, which will be different for each given event. By choosing the right basis, we may approximate each event with a finite number of surfaces $\mathbf{S_j}(\lambda, t)$. Principal Component Analysis (PCA) is an optimal way to select this basis from a ensemble of input data (e.g. Connolly et al., 1995, for astrophysical

[5] The term λ^{-1} comes from the conversion of energy per unit frequency into the number of photons per unit wavelength

application to galaxy spectra), and "sparse" or "gappy" PCA is a particular implementation to be used when the inputs sample only a small fraction of the model (Zou et al., 2006), such as here. As is standard when using PCA decomposition, we will first create a "mean" surface $\mathbf{S_M}(\lambda, t)$ and find the principal variations about this mean, yielding the model:

$$f_i(\lambda, t) = x_0 \times \left(\mathbf{S_M}(\lambda, t) + \sum_{j=1} x_j \times \mathbf{S_j}(\lambda, t) \right)$$

where x_0 represents an overall brightness scaling (i.e. distance modulus), and x_j represents the location of the phenomena within the family of events.

We emphasize that these basis surfaces ($\mathbf{S_M}(\lambda, t)$,$\mathbf{S_1}(\lambda, t)$,$\mathbf{S_2}(\lambda, t)$ ... $\mathbf{S_j}(\lambda, t)$) are the fundamental products of this proposal.

Algorithmically, this process to generate them is as follows: accumulate data on each instance of the event type to be modelled; for each instance, create a sparse data surface $f(\lambda, t)$; from the ensemble of $f_i(\lambda, t)$, find the mean $\mathbf{S_M}(\lambda, t)$; subtract the mean $\mathbf{S_M}(\lambda, t)$ from all $f_i(\lambda, t)$; run a Karhunen–Loeve decomposition (Karhunen, 1947; Loève, 1948) on the mean–subtracted $f_i(\lambda, t)$, yielding principal surfaces $\mathbf{S_j}(\lambda, t)$.

Intrinsic Dimensionality of each Basis: The Principal Component Analysis is fundamentally an eigenvector analysis, with eigenvalues that represent the variance in the population that is represented by each surface. Surfaces that represent effects that dominate the behavior of the population (e.g. the brightness–decline relationship for Supernova Ia) will have large eigenvalues, with secondary effects less important and thus having smaller corresponding eigenvalues in the PCA. One can thus derive the intrinsic dimensionality of the population from the spectrum of eigenvalues. Correspondingly, one can well approximate any event within this population by using the first few bases (sorted by eigenvalue). We will use such a truncated expansion when creating our basis models; only surfaces representing the majority of the variance (90% to 99%, depending on the amount of input data) will be retained in the event modeling process described above.

Uncertainties on the Surfaces: We will additionally capture excess variance in the models that is not represented by the surfaces, e.g. portions of large model uncertainty that we will want to de-weight when fitting. This excess variance will be modelled using a jackknife resampling procedure: for each input event, we will remove it from the training sample and re–create the model. The variance of these residuals across all jackknife resamplings $V(\lambda, t)$ will be used as an empirical estimate of the excess variance about our model. ■ **andy:** Uncertainties of theoretical surfaces could be generated from ranging over plausible (but unknown) input parameters that are used to generate the models. ■

Variants to this Procedure: We recognize that for certain phenomena, this basic derivation of the surfaces is not sufficient to completely describe the events. For example, objects with periodic variability require folding at the correct period before fitting to these surfaces. We must therefore construct periodograms for these lightcurves, optimally using the ensemble of event data to yield a single period assessment. This will require development of multi–band period assessment tools. Two options we will pursue are: the generation of

periodograms for each passband of information, the conversion of these periodograms into likelihood functions using a false–alarm probability analysis (e.g. Zechmeister & Kürster, 2009, and references therein), and using the product of these likelihood functions in a final period assessment; or to use priors on color–evolution derived from our ST surfaces to wrap all data into a periodogram analysis. Other classes of events such as supernovae undergo substantial extinction from their host galaxy, which will create a color–dependent non–intrinsic reddening of the underlying ST surface. This will require an assessment of the reddening of each input vector, and an additional fit parameter on each incoming lightcurve representing the degree of extinction (e.g. Equation 1, Guy et al., 2007). Finally, the evolution of objects at cosmological distances will behave like a stretched spectral–temporal surface. When fitting such classes of events, we will be comparing the observed data to the ST models with an extra degree of freedom in the redshift z:

$$f_i(\lambda, t) = \mathbf{S_M}(\lambda/(1+z), t/(1+z)) + \sum_{j=1} x_j \times \mathbf{S_j}(\lambda/(1+z), t/(1+z)).$$

Previous Work: We note that this procedure follows very closely the SALT–II model of Guy et al. (2007). In the SALT–II model, the degree to which the process captured the behavior of the training data is represented in the RMS dispersion of the Ia distance moduli in the Hubble diagram of 0.16 magnitudes **SO WHAT**. In addition, for events where the redshift of the event was not known, this extra redshift parameter could be included in the χ^2 calculation to provide a photometric redshift estimate. Guy et al. (2007) finds an RMS dispersion of $\Delta z/(1+z) = 0.01 - 0.02$, which is comparable to the redshift that may be inferred from the spectra alone (see also Kessler et al., 2010). Spectral reconstruction in practice (Asensio Ramos & Allende Prieto, 2010).

Statistical Infrastructure

We expect to construct the set of $\mathcal{S}_k = \{(\mathbf{S_M}, \mathbf{S_j})\}_k$ for k classes of variables and transients. While the actual number of classes will ultimately be subject to availability and quality of the input data streams (§XXX), we expect $k \approx 100$ and may be much more if we include a variety of theoretical curves. The fundamental question we wish to answer is: *given a new light curve, $f_{i+1}(\lambda, t)$, what is the probability that it arises from class k?*

A straight forward approach is to represent a notion of distance of the lightcurve from each of the k templates as $\mathrm{D}_k(f_{i+1})$, where the normalization and start time t_0 (or phase offset ϕ_0, or redshift z) are varied so as to minimize the χ^2 for each $(\mathbf{S_M}, \mathbf{S_j})$. The most likely class is then that which minimizes D [we will explore both parametric and non-parametric interpolations of the surfaces to facilitate this distance measure]. In the limit with well–sampled and multicolor light curves, this approach should be sufficient for class *selection*. However, for probabilistic lightcurve classification, especially in the few–data limit, a more complex approach is required: in particular, there must be some notion of a minimal distance subject to likelihood of having observed the particular dataset. In particular, each spectral–temporal training set yields the expected range of contributions from the secondary (tertiary, etc.) surfaces, which may be used as priors when doing absolute classification. We will explore such a Bayesian approach to distance minimization. All software tools (written in Python and C++) used to determine $\mathrm{D}_k(f_{i+1})$, as well as the surfaces themselves, will be made open source.

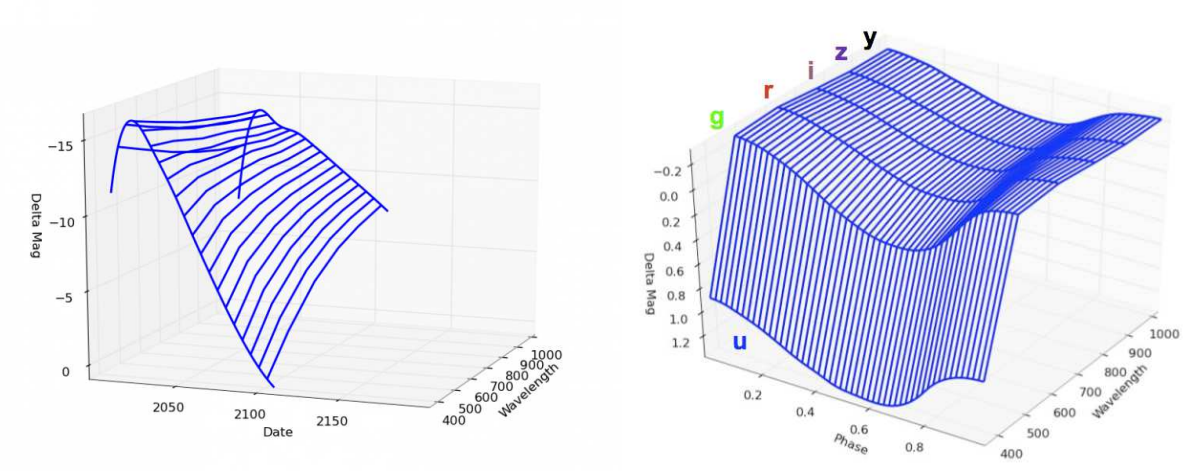

Figure 2: Example theoretical spectral–temporal surfaces of astronomical phenomena. The *left* panel shows the lightcurve of a failed "fallback" supernovae from Fryer et al. (2009), with total energy of 1.7×10^{51} erg, ^{56}NI yield of $1.0 \times 10^{-13}\,M_\odot$ and total ejecta mass of $3\,M_\odot$. The *right* panel shows a 5–band RR Lyrae lightcurve template from Sesar et al. (2010). While these phenomena are intrinsically quite different, they may be represented in the same spectral–temporal fashion. While we anticipate building our surfaces primarily from experimental data, models like the ones shown here will be used to fill–in–the blanks when there are insufficient constraints.

The lightcurve classification problem raises a very interesting statistical question: how can we account for the taxonometric (hierarchical, tree-like) structure of the landscape of astronomical events? A increasingly more detailed set of questions may be asked about such an event: is it variable or non-variable? Is it a pulsating variable or an eclipsing binary? Is it a Cepheid or an RR Lyrae? What kind of RR Lyrae is it? With limited data it may be possible to make highly probable classifications at the high levels of the taxonomy but not at the lowest levels.

The General Catalog of Variable Stars (GCVS; Kholopov et al. 1996) has developed a classification tree based upon an admixture of observed properties (e.g., "slow" and "red") and inferred physical characteristics (e.g., "eclipsing"). While it is not crucial for our purposes that a taxonomy be based on intrinsic physical properties alone, the lack of a coherent taxonometric structure for variable sources presents a challenge when one of the goals is to make a probabilistic statements about nature of an unknown event. We propose to *explore possible taxonomies with particular attention to their computational consequences.* The existing taxonometric structure is certainly relevant to training a classifier based upon $\mathcal{S}_k$, since not all lightcurves in the training database will be confidently identified at the leaf level, but may contain valuable information about the relationships of features to higher level vertices in the taxonomy.

Making robust probabilistic statements based on a graphical structure has received some attention in the statistics literature (Koller & Sahami, 1997; Sun & Lim, 2001; Dekel et al., 2004; Cesa-Bianchi et al., 2006; Bartlett et al., 2005), with much of the interest spurred by the problems of identifying text (e.g., Weigend et al. 1999). We propose to explore two general ways of attacking this problem. One approach is to incorporate a fixed (e.g.

GCVS) taxonomy by changing our measures of loss to reflect taxonomic information (through either simple tree-based distances or more involved Laplacian-based ideas), going beyond the standard 0–1 loss described above. (Under 0–1 loss, the cost of misclassifying an RR Lyrae as a Cepheid is the same as failing to distinguish between two types of RR Lyrae.) Another approach addresses the taxonomy by doing a "multi–scale" or "multi–resolution" classification. Much of the exploration will be in finding various metrics on the tree (e.g. Chung 1997).

Another approach to incorporating taxonomy into classification, which we propose to explore, could make use of recent work on statistical phylogenetics, where the taxonomy is provided by a phylogenetic tree. The idea (**?**) is to use the taxonomy to induce a taxonomically meaningful inner–product between observations. This is done by, for instance, using the Laplacian of the graph and its spectral decomposition to find a taxonomically meaningful way of representing functions on the graph, which for us are really vectors of observations on the graph. For instance, when using Principal Component Analysis, one can then incorporate graph information into classical multivariate statistical techniques Purdom (2008). A natural question is to investigate whether a similar idea can be used in the context of max–margin classifiers, though there are classification–specific problems that need to be addressed. We note that ideas from structured classification work (Bartlett et al., 2005) might become helpful in our context. In this line of inquiry, one might also ask if one can derive linear combinations of features that are taxonomically–meaningful and use those in growing decision trees.

Another possible approach that takes into account the tree structure of the problem would consist of pruning the tree $\mathcal{T}_0$ at different levels and constructing recursively classifiers for each level. For example, we could first construct a random forest that would classify the observations into the first level of categorizations of the tree $\mathcal{T}_0$. Call $\{\mathcal{T}_1^{(i)}\}_{i=1}^k$ the corresponding subtrees. We could then fit a random forest in the same fashion to classify the observations assigned to each of these new trees into their first level of categorization. A procedure like this one clearly takes into account the taxonomy information (cf. Cesa-Bianchi et al. 2006). We note that this sequence of classifiers could be constructed with a sequence of distance-based loss functions described above which assign edge lengths equal to zero for lower levels of the tree.

The notion of a classifier being "calibrated" is a relevant line of inquiry. This has long been a concern in probabilistic weather forecasting—does it rain on 30% of the days for which the forecast probability of rain was 30%? The "probabilities" a classifier produces may not correspond to the actual relative frequencies in the training set on which it was based (Niculescu-Mizil & Caruana, 2005). In our context it would be desirable that the proportion of objects labeled as pulsating variables is a good estimate of the actual proportion of such variables in the data set. For example, we seek to meaningfully combine the reported probabilities of the various types of pulsating variables to obtain the marginal probability that an object is of this type. We plan to investigate this issue of calibration within a taxonomic structure.

The nexus of statistical methods, modern computational resources, and massive data sets is leading to revised notions of statistical efficiency that take into account computational constraints (**?**). Convex surrogates for 0–1 loss as mentioned above are but one of a growing

number of examples. Mindful of the importance of computational efficiency in the LSST era, *we expect that the methods we develop to meet the unique challenges posed by lightcurve classification will contribute to this evolving literature.*

Computational constraints are less important during the learning/training phase than at other stages. A decision tree would clearly be preferable to a Bayesian method that required expensive Markov Chain Monte Carlo (MCMC) runs for every object. Constraining a classification procedure by a computational cost would amount to adding a Lagrange multiplier times the computational cost to the statistical loss function. The key is that the resulting optimization problem be computationally tractable. **?** show that by using dynamic programming, tree searches from the top node downward could produce enormous gains in computation with very little loss of statistical efficiency.

An especially intriguing and challenging aspect of real-time lightcurve classification is that the $\mathcal{D}_k$ of each object *changes* with every new data time–series data point. Here the taxonomic approach can be used to advantage. For example, some strategies arise naturally for an approach, such as that described above, which composes classification procedures at every vertex of the tree. For a given object it may not be too important to classify more finely in the taxonomy at every new observation. On the other hand, if there is some non–negligible probability of an event belonging to a rare class or a never–observed phenomena, it would be crucial to update frequently. Thus, updating policies can vary among the vertices and the updating choices will depend dynamically on the availability of computational resources. We propose to explore this technique both at the theoretical level and by example.

Event Classification Service

Members of our collaboration are already running a discovery and classification engine for the Palomar Transient Factory and will soon be adding the Dark Energy Survey and La Silla Supernova Search as part of the input streams. This service, now private, is highly trained to PTF cadences and peculiarities, making use of light curve and context information when available. The $\mathcal{D}_k$ metrics on each source will serve as additional features for the classification engine.

Both DES and La Silla are expected to run during much of the lifetime of the proposed work, allowing us to expand the classification effort with this additional input. In addition, as new surveys start producing public streams (e.g., PTF2 and Gaia) we intend to add a classification engine for those sources as well. For all non–proprietary data streams we propose to build a public classification service, allowing subscribers to receive (via push and pull mechanisms) real–time classification statements about each source that match certain criteria. This web service will be operated out at UC Berkeley and make use of best practices for web front ends, database architecture (e.g., postgresql, Hadoop), push mechanisms (e.g., Jabber), and astronomical event packaging (e.g., VOEvent and other VO tools).

IIIa. Team Qualifications

Dr. Becker

PI Becker has an extensive history working on classifying event streams from within several time–domain projects, including the MACHO project (Becker, 2000), the Deep Lens Survey (Becker et al., 2004), and most recently the SDSS–II Supernova Survey (Frieman

et al., 2008; Sako et al., 2008). His most relevant work to this proposal was in leading the SALT–II cosmology analysis in Kessler et al. (2009). His familiarity with the SALT–II software provided the inspiration to extend these models to all classes of variability, as proposed here. He has been aggregating spectral–temporal surfaces for the LSST image simulation effort (Connolly et al., 2010) for the purposes of added realistic stellar and cosmological variability to the simulations. He has been working since 2004 on the real–time nightly processing pipeline for LSST at the University of Washington.

Dr. Bloom

Co–PI Bloom is director of the Berkeley Center for Time–Domain Informatics, which focuses statisticians, computer scientists, and astronomers on matters of classification and regression on astronomical time–series. The primary focus has been on the real-time and retrospective classification of the Palomar Transient Factory survey and other public datasets (e.g., Stripe 82). He has also worked on efficient discovery techniques of quasars through time variability and fast implementation of Lomb-Scargle periodograms with Bayesian cross validation. He has worked extensively on gamma–ray burst followup and characterization. He is co-chair of the LSST science working group on transients and variable stars. VOEvent, the IVOA standard for astronomical event publication, was his brainchild.

Dr. Connolly

Co-PI Connolly is simulation scientist for the Large Synoptic Survey Telescope, and lead of the UW Data Management group. His previous work includes investigating the dimensionality of large astronomical data sets using PCA and Locally Linear Embedding, developing and releasing applications for data intensive cosmology, and for integrating research and education (e.g. Connolly was the technical lead for the development of Sky in Google Earth).

IIIb. Previous Support

Dr. Becker

"The LSST FaST Program : Expanding Participation of Underrepresented Minorities in LSST", funded through Specific Program Order 9 (AST-0551161) to the NSF-AURA (Association of Universities for Research in Astronomy) Cooperative Agreement AST-0132798. The project involved simulating tens of millions of RR Lyrae lightcurves to investigate LSST's ability to recognize the periods and types of these events, as a function of distance (faintness) and survey duration. This team delivered a technical report to the LSST Transients and Variable Stars working group at the LSST Fall 2009 "All Hands" meeting and has submitted a paper to the Astrophysical Journal on the final results (Oluseyi et al., 2011). Importantly, three of the six students funded by this proposal successfully applied to graduate school, with a fourth expected to apply this next year.

Dr. Bloom

"Real-time Classification of Massive Time-series Data Streams" (PI, J. Bloom; NSF grant No 0941742; amount $1,573,550; expires 07/12). Three postdocs, six graduate students, and several undergraduates have been supported as part of that project, resulting in more than 10 publications to date. The classification framework built as part of this proposal is used

in the real-time pipeline of the Palomar Transient Factory, and has been responsible for the discovery of more than 15,000 new variable stars and transients.

Dr. Connolly

"Image Coaddition, Subtraction and Source Detection in the Era of Terabyte Data Streams" (PI, A. Connolly; NSF grant AST-0709394; amount $427,933; expires 8/31/2011) is the most closely related grant. Outcomes from this work include: the development of non–parametric techniques for the detection of sources within sequences of astronomical images through the use of image coaddition and subtraction, and algorithms for measuring the clustering of galaxies using n-point correlations functions that scale to high-performance parallel architectures.

IV. Broader Impacts

The proposal will build and strengthen collaboration between current producers and consumers of real–time event data (UCB with PTF) and future producers of same (UW with LSST). It will bring needed statistical rigor to the multidimensional description and interpretation of astrophysical phenomena; this type of interdisciplinary effort typically pays dividends by fostering advances in both fields. In statistics this will occur in the application of sparse PCA on a two–dimensional surface of information, and in the innovation that it will require to operate robustly. In astronomy, it will result in the understanding of the intrinsic dimensionality of all classes of astronomical variability, how the classes may be interrelated, and what sort of observations (in time and wavelength) are optimal for nailing down the class of a given event. These models will also enable photometric redshift estimates for the event types (e.g. Kessler et al., 2010), which may be used in conjunction with photometric redshifts of any host galaxy (e.g. Baum, 1962) to provide a more certain estimate of event redshift without spectroscopic observations, which are expensive to obtain.

Our participation in current and future time–domain surveys will also ensure that these models are not generated in a vacuum, and that they will be developed with practicality in mind. We plan to advertise this project at multiple conferences per year, which will allow for collaboration with other producers and consumers of time–domain data we may have not included here. Importantly, our planned inclusion of these models into extant classification services will ensure that they achieve broad application and relevance in a burgeoning field.

V. Project Management and Development Plan

PI Becker will serve as the lead for supervision of a research associate (at the post–doctoral level) at UW. This postdoc ideally will have expertise in the calibration and management of survey–level volumes of astrophysical data. Co–PI Bloom will serve as the lead for supervision of a second research associate (at the post–doctoral level) at UCB. This postdoc ideally will have expertise in the application of statistical tools in the field of astronomy. We require RAs at the postdoctoral level because the technical challenges posed by this problem require individuals with immediate expertise, and who have fought through such issues before at (at least) the graduate student level. This includes minutiae such as the role of the underling spectral energy distribution on the calibration of photometry (there is no "one" zeropoint per astronomical image, it is source dependent), and the efficient implementation (memory management, disk access, scaling relations) of statistical algorithms on compute

hardware. PI Becker will serve as the technical lead for the project, and will contribute in all arenas astronomical and statistical at the level of 25% of his time.

The difficulties in having a geographically distributed team require a commensurate level of interaction to maintain the group focus. We plan on having weekly phone or video–conferences, moderated by Becker and Bloom, outlining the current state of the project, addressing problems that arise, and setting goals and expectations for future work. We will also arrange a quarterly trip for each postdoc to visit the other's institution, with a rotating host institution, for direct interaction and camaraderie. We budget for 3 domestic trips per year per institution in this proposal, 2 of which will be used for these collaborative meetings. We budget for one domestic and one international trip per institution for the postdocs to advertise our work at appropriate conferences in the fields of Astronomy and Computer Science. This type of model – mixing regular weekly phone conferences with biannual collaboration meetings – has served Becker and Connolly well in working within the distributed LSST Data Management group.

We outline the yearly responsibilities for the PIs, the UW postdoc (PD1), and the USB postdoc (PD2) below.

Year 1

The first year of the project will require the enumeration of the astrophysical phenomena we will model in this project (Bloom), the aggregation and calibration of data on these events from extant data archives (PD1 + Becker), the development of the statistical infrastructure to build the ST models from these input data (PD2 + Bloom + Becker), and importantly the staging of the input data for the Principal Component Analyses (PD1 + PD2).

Year 2

In the second year we come to the meat of the project, where we must build viable models from the aggregate input data. Details such as the model resolution in wavelength and time, and the number of principal components that can be meaningfully extracted from the data, will be highly input data–dependent. The input data themselves will be inhomogeneous, meaning we will have to converge on the optimal configuration of each ST surface dynamically. PD1 will be responsible for understanding how the input data may effectively be translated into output models (surface resolution in λ, t; number of principal components the data will support w/Connolly), and PD2 will be responsible for understanding how to effectively implement a sparse PCA with these data, and how sparsity impacts the uncertainties of the surface at each λ, t bin. Becker will actively participate in both of these efforts using as reference his understanding of the SALT–II algorithm and its implementation details.

Year 3

The final year of the project will include validation of the models, and incorporation of the models into a classification service. PD1 will be responsible for developing a fitting infrastructure that takes as inputs a new data stream and returns the best–fit parameters from each surface (x_{ij} and period or redshift). PD2 will be responsible for taking the results of this fitting process and returning statistical likelihoods that the event is of each class. PD2 will also be responsible for incorporating the overall ST infrastructure into an on–

line classification resource that is available to the astronomical community. PD1 will be responsible for enforcing validation of the input data to this service, while PD2 will be responsible for validating that the probabilities that are returned from this fitting process are accurate representations of our knowledge of the system (through monte–carlo simulations and using sources of vetted data that were not input to the models). ■ **andy:** Becker should do what? ■ ★ **josh:** Does Bloom want summer salary for this part of the project? ★