Learning in an Era of Uncertainty

Applicant/institution:	University of Washington
Street Address:	
Principal Investigator:	Andrew Connolly
Telephone number:	(206) 543 9541
Email:	ajc@astro.washington.edu
Administrative POC name:	Lynnette Arias
Telephone number:	206-543-4043
Administrative POC name:	osp@uw.edu
Funding Opportunity FOA Number:	DE-FOA-0000918
DOE/OSP Office:	Office of Advanced Scientific Computing Research
DOE/Technical Contact:	Dr. Alexandra Landsberg
PAMS Preproposal tracking number:	PRE-0000002147

Collaborating Institutions:
Lead Institution: University of Washington, PI Andrew Connolly
Collaborating Institution: Carnegie Mellon University, PI Jeff Schneider

Lead PI: Andrew Connolly

Learning in an Era of Uncertainty						
	Names	Institution	Year 1 Budget	Year 2 Budget	Year 3 Budget	Total Budget
Lead PI	Andrew Connolly	U Washington	$170K	$170K	$170K	$510K
Co-PI	Jeff Schneider	Carnegie Mellon U	$170K	$170K	$170K	$510K
TOTALS			$340K	$340K	$340K	$1020K

I. INTRODUCTION

A new generation of DOE sponsored data intensive experiments and surveys, designed to address fundamental questions in physics, materials, and biology will come on-line over the next decade. These experiments share many similar challenges in the fields of statistics and machine-learning: how do we choose the next experiment or observation to make in order that we maximize our scientific returns; how do we identify anomalous sources (that may be indicative of new events or potential systematics within our experiments) from a continuous stream of data; how do we characterize and classify correlations and events within data streams that are inherently noisy and incomplete. The goal of this proposal is to address these challenges through the development of machine learning techniques that better quantify the uncertainty in their predictions and corresponding active experiment selection algorithms that utilize those uncertainties to get the most scientific information out of limited data collection budgets.

Active learning algorithms iteratively decide which data points they will collect outputs on and add to a training set. Their goal is to choose the points that will most improve the model being learned. At each step, they consider the current training data, the potential data that might be obtained, and the current learned model, and evaluate what would be the best choice for the next observation, experiment, or feature such that it improves our knowledge of the overall system (according to some objective criterion). The potential impact of active learning algorithms is substantial (optimizing the scientific returns from billion dollar investments in observational facilities). To achieve these breakthroughs requires that we address the challenge of how to scale active learning to the size and complexity of the data expected from this next generation of experiments. For example, our inability to undertake a full lookahead, predicting the impact of all future active learning choices on our learned model, results in the development of myopic heuristics that improve the speed of these learning techniques but at a substantial cost in how well they perform on real data. See, for example, Figure 1 from Garnett *et al.* (2011) and Figures 3 and 4 of Garnett *et al.* (2012a).

By addressing these challenges we propose to develop active learning algorithms that will scale to data sets with hundreds of millions of entries and petabytes of data. This work has the potential to impact many of the data intensive sciences. For this proposal, however, we will focus our work in the context of DOE sponsored cosmology experiments (i.e. the Dark Energy Survey[1], and the Large Synoptic Survey Telescope[2]). These surveys are ideal proxies as their bandwidth (terabytes of data per night and petabytes of data every couple of months) will enable high precision studies of cosmology. The ability to use these data sets to achieve an order of magnitude improvement in constraints on our understanding of cosmology and dark energy will, however, depend on how well we can analyze, optimize, and calibrate data streams that are noisy and incomplete. This will require the development of fundamentally new approaches to the analysis of data at a scale, speed, and complexity beyond the capabilities of current automated machine learning methods.

A. Project Objectives

On-line, data-driven analysis will require model-fitting and classifications that are non-parametric and probabilistic. Gaussian Processes meet these requirements and we will, therefore, build our methods around them. We note, however, objectives 2-4 below will result in active learning algorithms that do not depend on this choice, and should be applicable to any regression or classification method that yields a probability distribution over outcomes. We divide our research program into the following objectives.

(1) Designing Robust Gaussian Processes. Figures 1 and 2 below demonstrate that, "out of the box," Gaussian Processes can already perform much better than other algorithms (both forward-fitting and data-driven) for model inference. Extending them to the future of big-data science will, however, require confronting problems not yet faced by automatic data analysis algorithms. *Model degeneracy* – most model-fitting algorithms, including Gaussian Processes, are written to return an optimal value (whether a latent physical variable or a classification) and some error on that value. This assumes that the true model obeys single-mode Gaussian statistics. As future experiments continue to push the boundaries of our physical understanding, this assumption will cease to be valid and we will be forced to build algorithms that can accommodate multi-modal models obeying a wide range of probability distributions. We propose a series of modifications that will allow Gaussian Processes to function in this regime. *Incomplete training data* – the difficulty of experiments being attempted in high energy physics, astrophysics, and biology often means that, not only is the gathering of data expensive, it may not be possible in some regimes. Future algorithms will need to be able to make inferences about models in regimes where training data is either very sparse or not available at all. Because of their non-parametric design, Gaussian Processes are amenable to this kind of modeling and we propose additional modifications that will enhance their robustness against such incomplete training data.

(2) Active Learning. Supervised regression and classification algorithms require labeled data points for training. These output labels are obtained through labeling by human experts and/or additional data collection. Often, the resources required (both in terms of human-hours and experimental apparatus) are considerable. We propose to develop new algorithms to optimize the use of these resources by determining what event or object labels will maximize the improvement to models from supervised learning algorithms. Our prototype implementation (see Figure 4) using a classical heuristic on the problem of determing astronomical Doppler shifts (or redshifts) from broad-band photometric data already shows the promise of active learning in cosmology. We propose novel algorithms based on new myopic criteria and increasing the computational scalability of lookahead methods that will greatly improve the accuracy and coverage of learned models.

(3) Active Feature Acquisition. Active Learning selects specific unknown events or objects for follow-up observation and classification by human experts. Active Feature Acquisition further refines our inquiry by asking what kinds of follow-up observations will yield the most information about the unknown object. We propose to extend our recent work on using Gaussian Processes (GPs) to detect damped lyman-alpha (DLA) systems [Garnett *et al.* 2012b]. In that work GP regression was used on each observed, noisy spectrum to infer the latent spectrum. The single independent (input) variable was wavelength. In other problems (such as the identification of astronomical transients) there will be several input variables which we can choose to observe or ignore. We will build an information-theoretical framework to make this choice. In the DLA work, a binary choice was made between models with and without a DLA. DLAs were classified by recognizing which model fit best. In the proposed work, we will learn a different model for each class of object and will estimate class probabilities using Bayes rule for combining the prior probability for each class and how well the respective models fit the observations. We will apply these techniques to the challenge of characterizing and classifying the light curves of supernovae (used in measuring the cosmic acceleration) from data with poor temporal sampling, The key advantage of this approach is that the GPs naturally provide a mean and covariance for future unobserved observations (see Section III A for a more detailed discussion). This uncertainty propagates through to class labels and we can use it to estimate the reduction in class uncertainty that will be gained by making a given observation at a specific time. The observation yielding the greatest reduction in entropy for the class and the light curves for this object will be taken.

(4) Active Search. Often it is the anomalous events within large streams of data that lead to fundamental breakthroughs. We, therefore, require methods for rapidly identifying and classifying potentially novel events and objects while staying within the limited budget of additional experiments that might be undertaken to confirm these discoveries. This is an active search problem. The problem and the Bayesian optimal algorithm for it are described in our recent work [Garnett *et al.* 2011, Garnett *et al.* 2012a]. As in active learning, the acquisition of class labels is expensive and we need to learn a model to predict these labels from limited input data. The final performance objective is, however, not the accuracy of the classifier, but rather the number of positives (i.e. objects from interesting classes) identified. Our previous work considered a binary classification problem: "is the object 'interesting' or not?" We will extend this work to allow for classification according to a continuous scalar function or multiple discrete classes. We will design algorithms that consider probability distributions across all possible classes and quantify how new observations will affect these distributions with an eye towards reducing the entropy, or some other information-theoretical measure of uncertainty.

(5) Science Outcomes. Data from the Dark Energy Survey will become publicly available over the term of this project (together with detailed simulations of the expected data flow from the LSST). We will, therefore, use these data to demonstrate that the algorithms developed above will improve the cosmological reach of the DES and LSST in the areas of improved determination of photometric redshifts (see Section III B) and improved classification and online follow-up of transient astronomical objects (see Section V). We will release our software publicly and seek collaborations where our algorithms may be spread to other sciences – such as climatology and microbiology – wherein complex systems can be modeled or measured at great expense and algorithms are required to determine which models and experiments are actually worth performing.

Each of these goals will require us to develop algorithms to simultaneously learn the latent physical model underlying a given data set, the uncertainties around that model, and the potential information to be gained by futher studying a specific instance of the model ("event" or "object").

B. The Collaboration

To accomplish the objectives laid out in this program we have assembled a team experienced in algorithm design, data structures, and in developing and delivering data mining algorithms that are actively used by the cosmology community. The PIs of this research proposal have a proven track record for propagating research ideas in educational and multidisciplinary environments. Connolly and Schneider have been part of an ongoing collaboration between machine learning, cosmology and statistics for over a decade (noted by the President of the American Statistical Association (ASA) as an exemplary interdisciplinary research team [Straf 2003]).

Highlights from this collaboration include n-tree searching algorithms that make the calculation of n-point correlation functions scale to the size of current surveys [Gray *et al.* 2004] and that enable rapid characterization of orbital tracks in sparsely sampled temporal data [Kubica *et al.* 2007]. The software associated with these algorithms was made publicly available and has been used to compute the 2–point function on over 10^6 galaxies and the 3–point correlation function of 400,000 galaxies from the SDSS survey [Scranton *et al.* 2002, Szapudi *et al.* 2002, Nichol *et al.* 2006, McBride *et al.* 2011a, McBride *et al.* 2011b] as well as in measures of the marked correlation functions [Skibba *et al.* 2006]. As part of this collaboration [Yip *et al.* 2004, Vanderplas and Connolly 2009, Daniel *et al.* 2011] introduced to astrophysics signal compression and analysis techniques that are now regularly applied to the analysis of spec-

troscopic surveys. More recently this collaboration has developed algorithms for automatically classifying astronomical objects [Vanderplas and Connolly 2009, Daniel *et al.* 2011] as well as an initial set of papers that use a simplified active learning method to accelerate the exploration of complex parameter spaces [Daniel *et al.* 2012]. Schneider's group has led the development of several new methods of regularization to learn complex models with only a small amount of training data [Zhang and Schneider 2010a, Zhang *et al.* 2010, Zhang and Schneider 2010b, Zhang and Schneider 2011, Zhang and Schneider 2012], for non-parametric estimators for divergences and dependencies [Poczos and Schneider 2011, Poczos *et al.* 2011, Poczos *et al.* 2012], and for graphical models for finding groups of collectively anomalous records [Xiong *et al.* 2011a, Xiong *et al.* 2011b].

Connolly leads the University of Washington data management group that develops the algorithms and techniques for the real-time analysis of LSST data (including the detection and characterization of transients and anomalies). Schneider heads a machine learning group that has done extensive work on automatic anomaly detection and pattern-recognition in complex data sets [Xiong *et al.* 2011a, Poczos *et al.* 2012]. Schneider, and Connolly are members of the Large Synoptic Survey Telescope collaboration with Connolly the software coordinator for the DOE sponsored Dark Energy Science Collaboration (a collaboration of over 150 scientists working on the characterization of dark energy).

On the educational front, all PIs have developed and taught computational techniques at the graduate and undergraduate level including data-mining for Astrophysics. Connolly recently completed a text book "Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data" that will be published by Princeton University Press and provides a comprehensive introduction to cutting-edge statistical methods together with the software associated with these techniques.

II. THE ROLE OF ACTIVE LEARNING IN COSMOLOGY

Over the last decade a concordance model has emerged for the universe that describes its energy content. The most significant contribution to the energy budget today comes in the form of "dark energy", which explains the observation that we reside in an accelerating universe. Despite its importance to the formation and evolution of the universe there is no compelling theory that explains the energy density nor the properties of the dark energy. Understanding the nature of dark energy remains as one of the most fundamental questions in Physics today, impacting our understanding of cosmology, particle physics, and potentially theories of gravity itself. As noted in the report of the Dark Energy Task Force (DETF; constituted jointly by DOE, NSF and NASA), "the nature of dark energy ranks among the very most compelling of all outstanding problems in physical science".

To address the question of the nature of dark energy a new generation of DOE sponsored experiments are entering service (e.g. the Dark Energy Survey and the Large Synoptic Survey Telescope). These surveys will represent a 40-fold increase in data rates over current experiments (generating over 100 Petabytes of data over a period of 10 years) and decreasing the uncertainties on our measures of the underlying properties of dark energy by more than a factor of ten. At the scale of these experiment, statistical noise will no longer determine the accuracy to which we can measure cosmological parameters. The control and correction of systematics will ultimately determine our final figure-of-merit. For example, systematic errors in the estimation of cosmological distance or in the identification and classification of high-energy transient events

(e.g. supernovae) lead to biases in the derived cosmological parameters. A bias of just 1% in distance (at a redshift $z = 1$) degrades measures of the properties of dark energy by over 50% [Kitching *et al.* 2008, Huterer *et al.* 2006, Nakajima *et al.* 2012].

III. A PROBABILISTIC FRAMEWORK FOR SCIENTIFIC INFERENCE

Current state-of-the art algorithms attempt to learn a model as a one-to-one relationship between the input data and the output function. Uncertainties are also learned, but these are usually heuristics derived by considering multiple attempts at model-fitting such as by a committee of aritficial neural networks [Collister and Lahav 2004]. We propose to replace this framework with one that directly models the probability distributions underlying the data. We believe that this framework is the most robust and informative available. It will allow us to learn, not only the models underlying the data, but the uncertainties surrounding those models, and the potential information to be gained from different follow-up observations. These three inferences will be critical in an age of research with low tolerance for uncertainty and limited budgets for follow-up observation and will allow us to extend our algorithms' application beyond realm of function regression and into that of object classification. We will build this new framework on the foundation of Gaussian Processes.

A. Gaussian Processes

Gaussian processes (GPs) model the output of an unknown, noisy function in multi-dimensional input space such that any set of samples from it has a joint multivariate Gaussian distribution [Rasmussen and Williams 2006]. Given a set of observed samples from the function's input space, a GP can make predictions about a set of other locations and assign these predictions a multivariate Gaussian distribution. An important feature of GPs is that they do not make parametric assumptions about the form of the function they are modeling and thus are well suited to nonlinear regression problems.

GPs have been used successfully to describe a wide range of physical phenomena without having to assume a model of the underlying process, even in the case of sparse measurements. Examples in cosmology include the expansion history of the universe [Shafieloo *et al.* 2012] and interpolating point spread functions across large images [Bergé *et al.* 2012]. Mahabal *et al.* (2008b) and Wang *et al.* (2011, 2012) use a mixture of Gaussian Processes to model light curves and identify periodically varying sources. Huijse *et al.* (2011, 2012) improve upon their methods, using information-theoretical quantities to separate the true period of these variations from systematics introduced by observational apparatus. The PIs have used GPs to accelerate the search of high-dimensional likelihood functions on the cosmological parameters by efficiently selecting sample points [Daniel *et al.* 2012], to detect damped Lyman alpha systems in the spectra of quasars [Garnett *et al.* 2012b], and to optimize the performance of complex robots [Tesch *et al.* 2011a, Tesch *et al.* 2011b, Tesch *et al.* 2013].

We now describe how the underlying function is modeled with a GP based on a sample of training data. Assume that each training set datum is of the form $\{\vec{\theta}, y\}$, where $\vec{\theta}$ is an N_p-dimensional vector representing the measured data (input) and $y = f(\vec{\theta})$ is the latent quantity (output) we are trying to infer. f is assumed to be a probabilistic function on the N_p-dimensional space with some covariance function relating pairs of points on the input space, such as a squared exponential

covariance,

$$K_{ij} \equiv \mathrm{Cov}\left[f(\vec{\theta}_i), f(\vec{\theta}_j)\right] = \exp(-\frac{1}{2}|\vec{\theta}_i - \vec{\theta}_j|^2/\ell^2) \tag{1}$$

where ℓ is a characteristic length scale set by cross-validation. Under those assumptions, one can derive a posterior probability distribution for f at a new query point $\{\vec{\theta}_q\}$ by marginalizing over the measured points $\{\vec{\theta}\}$. This gives a Gaussian distribution with mean:

$$f(\vec{\theta}_q) = \bar{y} + K_q \left(K + \sigma^2 I\right)^{-1} (\vec{y} - \bar{y}) \tag{2}$$

and variance:

$$\Sigma_q = \mathrm{Cov}(f_q) = K_{qq} - K_q^T (K + \sigma^2 I)^{-1} K_q \tag{3}$$

Here, K is the matrix of covariances between all training points, σ^2 is the variance of Gaussian noise added to each observed value of y, $\vec{y}$ is the training data outputs, $\bar{y}$ is the algebraic mean of the elements of $\vec{y}$, and K_q is the vector of covariances between the query point and all training points. Eq. 3 extends directly to the case of multiple query points by taking the variables as matrices where appropriate, and the result is a full covariance matrix for the query points. Readers looking for a more detailed explanation of GPs should consult Rasmussen and Williams (2006).

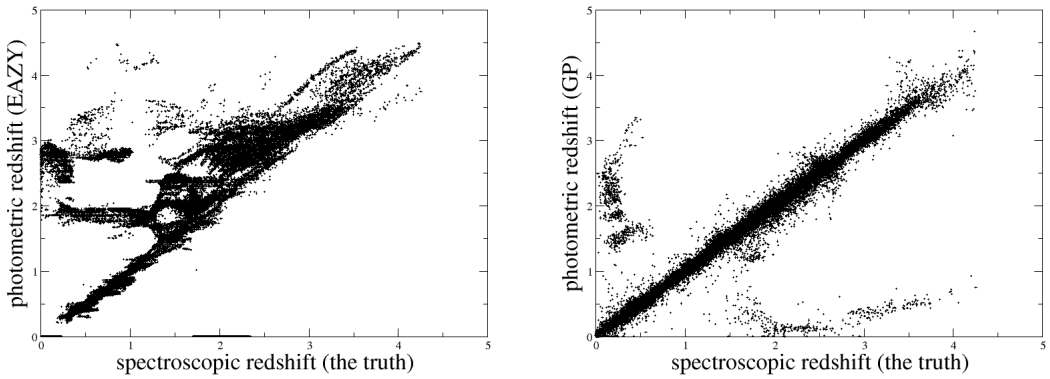
The inferences in equations 2 and 3 are non-parametric: they do not assume a form for the latent model they represent. This makes GPs especially powerful at learning models over data with complex degeneracies. We demonstrate this with the following problem.

B. Photometric redshifts: a Gaussian Process case study

One science use case for the algorithms proposed here is the problem of learning photometric redshifts. Any attempt to solve the astrophysical problem of dark energy requires measuring the cosmological Doppler shift of light from hundreds of millions of galaxies. Directly measuring the spectra of these galaxies is prohibitively expensive. Measuring their brightness in a handful of broadband filters (i.e., taking their photometry), is several thousand times cheaper. For this reason, surveys such as the DES and the LSST are designed to exclusively take photometric data and rely on algorithmic researchers to deliver ways of inferring a relationship between a galaxy's observed photometry (usually in five or six bands) and its true redshift.

Photometric redshifts are principally determined using forward-fitting models. Cosmologists assume that they can model the rest frame spectra of any galaxy. These spectral models are redshifted and integrated over the profile of an experiment's photometric filters until a good fit to the observed photometric data is found. The redshift of the galaxy is taken as that which produces the best fit between model and data. Many publicly available codes such as EAZY [Brammer *et al.* 2008] implement this method. While it is straightforward in principle, it requires accurate foreknowledge to select the appropriate rest frame spectral models. If the chosen models are not representative of the population of observed galaxies, the algorithm will fail to give accurate redshifts and cosmological inferences will be inaccurate [Budavári 2008]. The effects of this shortcoming can be

seen in Figure 1(a), which plots the results of running EAZY on a set of simulated galaxy observations designed to represent data expected from the LSST. While many of the galaxies fall near the $z_{\rm photometric} = z_{\rm spectroscopic}$ line, there is significant scatter in the results due to degeneracies between the color and redshift of a galaxy. Figure 1(b) plots the results from running the same simulated galaxies through GP-based algorithm that attempts to learn the full probability distribution $P(z_{\rm photometric})$ over the output. There is much less scatter in this case than in the case of the forward-fitting modeling of EAZY.



(a) (b)

FIG. 1: Photometric redshift plotted against true spectroscopic redshift for 48,000 simulated LSST galaxy observations. Photometric redshifts are derived using the EAZY template-fitting algorithm (a forward-fitting model) in Figure 1(a) and our GP based algorithm in Figure 1(b).

Other data-driven algorithms for photometric redshift determination do exist. An example of these is the publicly-available code ANNz [Collister and Lahav 2004], which is based on an artificial neural network scheme. In this case, the principal shortcoming the method is that the artificial neural network is designed to return only a photometric redshift value and an uncertainty. This uncertainty is a heuristic combination of the uncertainties in the photometric data as well as the scatter in $z_{\rm photometric}$ derived from multiple instantiations of ANNz. There is no guarantee that it represents the true, probabilistic uncertainty in $z_{\rm photometric}$. Figure 2 plots the mean value of $\ln[P(\text{truth})]$, i.e. the value of the logarithm $P(z_{\rm photometric})$ at the point $z_{\rm photometric} = z_{\rm spectroscopic}$, as a function of photometric redshift for EAZY, ANNz, and our GP algorithm. We see that both EAZY and ANNz consistently assign lower probabilities to $z_{\rm photometric} = z_{\rm spectroscopic}$ than does our GP algorithm.

Other works have attempted to apply GPs to the problem of photometric redshfifts [Kaufman *et al.* 2011, Bonfield *et al.* 2010], however, they have treated the problem as one of learning the form of a one-to-one scalar function. We propose to use the probabilistic nature of GPs to learn the full probability distribution that a given galaxy is at a given redshift. The resulting prediction is simultaneously more robust against sparse, noisy or degenerate training data, more usable in follow-on scientific analyses, and more amenable to model improvement by the introduction of active learning.

The tests considered above represent idealized cases. The algorithm resulting in Figure 1 was trained on high signal-to-noise data with a training set sampled everywhere in $z_{\rm spectroscopic}$-space. The algorithm resulting in Figure 2 is well-behaved in that the low redshift regime considered exhibits no degeneracies in the photometry-to-redshift relationship. This will not be the case for

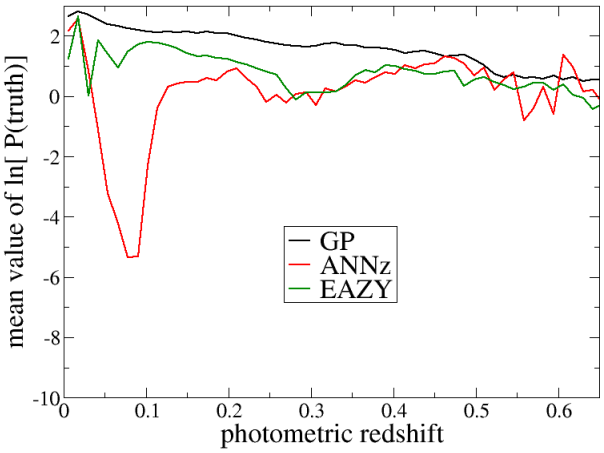


FIG. 2: The mean value of $\ln[P(\text{truth})]$ as a function of photometric redshift (the vertical axes in Figures 1) for all three algorithms under consideration using real data taken from the Sloan Digital Sky Survey [Abazajian *et al.* 2009]. In this latter case, the algorithms are trained on 21,000 galaxies and tested on 191,000 galaxies.

the large, deep surveys of the future. In order to extend the favorable results of GPs above, further development of the framework will be required.

C. Fortifying Gaussian Processes against incomplete training data

GPs are already designed to handle input data that is sparsely sampled. Equation 3 gives the GP the freedom to declare a large uncertainty in regions of input space where there is no training data. However, in order to deliver on the science results required by DES, LSST, and other big-data surveys, it is not enough that Σ_q be large where the model is uncertain. We require that the inferred value $f(\vec{\theta}_q)$ in equation 2 be an accurate representation of the true underlying model. We propose to achieve this in several ways.

- New mean function models – as presented, equation 2 sets the value of $\bar{y}$ to the algebraic mean of the training outputs $\vec{y}$. This assumes that the GP for all query inputs $\vec{\theta}_q$ starts from the same value and equation 2 expands the departures about that value. We will evaluate models more complex than the algebraic mean but simpler than the GP, which can be used to give a more informative value for $\bar{y}$ such that it interpolates over gaps in the training data. We propose to experiment with such functions including the development of a basis function representation of the $\bar{y}$ models. A preliminary analysis using principal component analysis on the input $\{\vec{\theta}\}$ to set $\bar{y}$ according to a linear regression on the principal components most correlated with the outputs $\vec{y}$ gives an slight improvement on the peformance of GP as a function regressor.
- Dynamic hyper parameters – the value of the characteristic length scale ℓ in the covariance function K_{ij} is presented as a parameter to be optimized by cross-validation. This, again, selects a single value for ℓ for all possible $\vec{\theta}_q$, regardless of $\vec{\theta}_q$'s position in the input space. We will develop dynamic algorithms based on the covariance structure present within the input $\{\vec{\theta}, y\}$ for setting ℓ so that highly-sampled training data will return small ℓ (tightly coupling

the GP to the training data) and sparsely-sampled training data will give larger ℓ, allowing for broader interpolation across gaps in the training data.

- Learned K_{ij} – in the above discussion, the functional form of the covariance function K_{ij} in equation 1 was assumed. It is, however, by no means guaranteed that the choice we made (a Gaussian function in input space) will always be the best. We will explore ways to enable the GP to learn the most appropriate form of the covariance function based on the training data at hand. Examples of this include setting the form of the covariance function using cross-validation and information-theoretical approaches that attempt to find a relationship between the form of the covariance function and the entropy of the distribution produced by the GP.

D. Fortifying Gaussian Processes against degenerate training data

It will be necessary to build GPs that are robust against degeneracies that exist in the relationship between input and output data. This pitfall can already be seen in the simulated photometric redshifts plotted in Figure 1(b). Note the high $z_{\text{spectroscopic}}$ galaxies that appear at low $z_{\text{photometric}}$ and vice-versa. To correct for this error, we will harness the full probabilistic nature of GPs, building a system which tests the simplest, single-mode Gaussian model against multi-modal models. The proposed algorithm will work as follows:

1. Fit the training data to a single-mode GP.
2. Use Bayesian model selection [MacKay 1992] to compute the model likelihood of this single-mode GP.
3. Divide the training data into two sub-populations based on their output values.
4. Fit each sub-population to its own GP.
5. Compute the model likelihood of this two-mode model and compare to the model likelihood of the single-mode model.
6. Repeat steps 3-5 for three-mode, four-mode, etc. models until the system converges to the best fit to the training data.

Through this iterative approach, we expect to build GPs such that, even when they succumb to the degeneracies seen in Figure 1(b), they will still place significant probability density at the true $z_{\text{photometric}} = z_{\text{spectroscopic}}$ value. This will represent a signtificant improvement over current model-learning algorithms (i.e. neural networks), which are principally concerned with returning a learned value $f(\vec{\theta}_q)$ for the model function and some heuristic estimate for the uncertainty Σ_q.

IV. ACTIVE LEARNING

A generic active learning algorithm takes the following form:

1. Learn a model (e.g. a Gaussian process) using the current set of labeled training data.
2. Evaluate the set or space of unlabeled data according to some active learning criterion, usually beginning by requesting the predictive distribution from the learned model.
3. Choose the data point/experiment with the highest evaluation and obtain the label for it

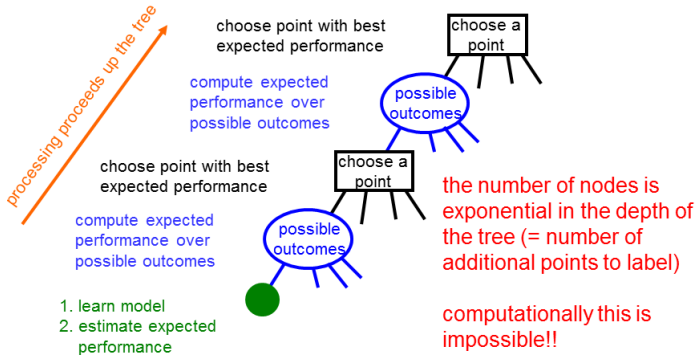


FIG. 3: The search tree for an optimal active learning algorithm that will be allowed to choose two more data points.

(e.g. by requesting a data collection experiment and/or asking a human expert to provide the label).

4. Add the resulting data to the training data and repeat.

The performance of an active learning algorithm is evaluated by some measure of the model's quality (e.g. accuracy or log-likelihood on a test set of data) as a function of the number of data points collected for the training set. The key algorithmic components are the model used in step 1 and the selection criterion used in step 2. The previous section proposed work on improving GP models. In this section we propose new selection criteria.

A. Novel active learning criteria

It is useful to see the challenges involved in developing good active learning criteria by considering what an optimal algorithm might need to do. Figure 3 shows the computation it would require. The root of the tree contains the current decision on which point to choose and an optimal algorithm would consider them all. For each choice, it would then consider every possible outcome (label) that might be obtained and weight them by the estimated probability of each outcome. For each outcome, it would then consider what next point it would choose if it received that outcome. The tree expands down to a depth equal to the number of data points that will be chosen and at the leaf every possible model would be learned and evaluated. The algorithm does not actually choose a new point to label (i.e., make the decision signified by the root of the tree) until it has extrapolated all possible choices out to every possible leaf and determined which of the original choices is optimal. Since the number of leaves is exponential in the depth of the tree, this quickly become intractable. A typical solution (referred to as myopic) is to truncate the tree to zero or one level. A second speedup is often obtained by replacing a full model evaluation with a fast heuristic (see [Settles 2009] for a summary of common heuristics).

The evaluation of the model, whether in a full or truncated tree, is also a challenge. If we knew the true labels in our test set, we could use them to evaluate the accuracy or log-likelihood of the true labels. In a simulated problem this is possible, but in a real application we do not have the labels. The usual alternative is to consider the uncertainty in the predictions on the test points. In the case of basic GPs, the distribution is a multivariate Gaussian with covariance Σ_q from equation 3 and the query is the entire test set.

A myopic information gain (change in entropy) strategy looks ahead one step and chooses the data point that minimizes the expected entropy (monotonic in $\det(\Sigma_q)$). Even this can be expensive since it requires the computation of the determinant. An alternative is minimizing the average variance $\mathrm{tr}(\Sigma_q)$ (also known as V-optimality). A further approximation is to do no lookahead and simply choose the data point corresponding to the largest diagonal element in $\mathrm{tr}(\Sigma_q)$. This is known as uncertainty sampling.

We propose the following new methods:

- The usual implementations of the criteria listed above are based on Gaussian predictive distributions which means only the covariance need be considered. We will develop efficient ways to estimate these quantities from our improved models of non-Guassian distributions. Two obvious criteria to try are the entropy of the output distribution or the logarithm of the probability distribution evaluated at its peak (which we use in the toy model generating Figure 4).
- Our preliminary experiments on classification in graphs indicate that the sum of all entries in the covariance matrix is a better criterion than the trace [Ma *et al.* 2012]. We propose to develop an analogous criterion for euclidean spaces and for regression problems. This criterion has not been previously proposed in the experiment design literature and we will also analyze theoretically when and why it outperforms the alternatives.
- It is clear that we can improve performance by looking ahead further (e.g. see Garnett *et al.* (2012a)). We propose to develop pruning rules that will allow us to lookahead further using our improved models. The rules will have the ability to trade off the aggressiveness of pruning, and thus the amount of look ahead allowed, against the approximation accuracy. We will investigate when the additional lookahead allowed more than compensates for the errors induced by pruning.

B. Active learning for photometric redshifts

In both empirical and forward-fitting photometric redshift codes, biases arise from the fact that the distribution of training samples (or templates) is not the same in color and redshift space as the data the model will be applied to. Additional issues include the fact that some regions of color space and some redshifts are more difficult to learn (require more training samples). We will show how our novel active learning algorithms can achieve further improvements by choosing the most useful galaxies for follow-up spectroscopy.

In a preliminary study we used a fully labeled data set containing galaxies with both color measurements and true (spectroscopically determined) redshift labels. We simulated active learning by hiding the labels from the GP learner and then providing them as requested by an active learning algorithm. We compared uniform random selection against uncertainty sampling. Figure 4 demonstrates the results using the problem presented in Figure 1. We start with 20,000 training points (consistent with the size of current training sets used for photometric redshifts) and assess the efficacy of our GP classifier by considering the mean value of $\ln[P(\mathrm{truth})]$ as in Figure 2. Using this simplified active learning to add to our training sample (as opposed to a random sampling strategy) leads to a significant improvement in the classifier's performance.

This study was done in a high signal-to-noise regime. It takes no account of the relative cost of doing one measurement or another. No gaps exist in the available training data. These are all challenges that real-world machine learning algorithms will face (see Section III C). We will

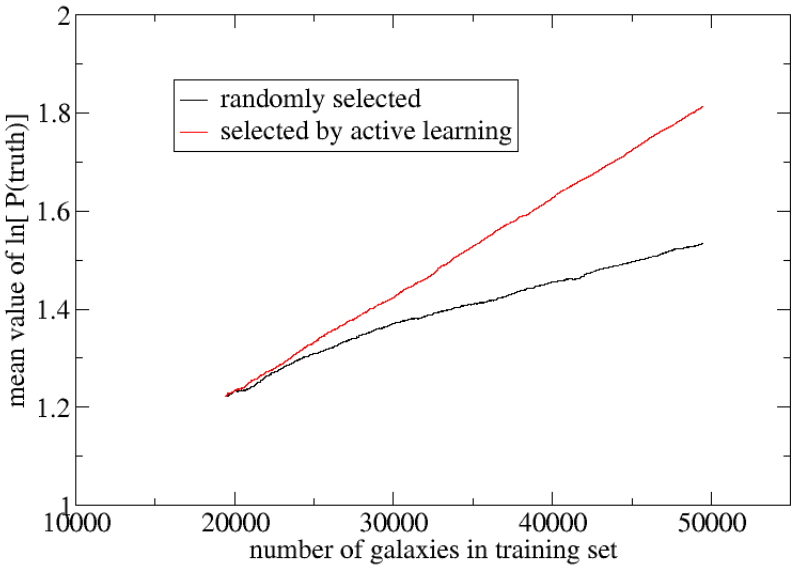


FIG. 4: Active learning applied to classification according to photometric redshifts, i.e. a scalar function on a 6-dimensional input space. The horizontal axis is the size of the training data set. The vertical axis is the mean value of the output probability distribution at the true value of the redshift. The black curve assembles the training set randomly. The red curve selects new training points that maximize the figure of merit $(-\ln[P(\text{mode})])$.

demonstrate our novel methods on real data sets that contain all these additional challenges.

V. ACTIVE FEATURE ACQUISITION: ANOMALY DETECTION AND CLASSIFICATION IN MASSIVE DATA STREAMS

While the problem of photometric redshift determination is fertile ground for the development of active learning algorithms, it poses little challenge for active feature acquisition: for a given experiment, the photometric filters are often immutable. Survey cosmology, however, presents us with another problem, the identification and classification of transient sources, which raises both the questions "which objects should we follow-up?" and "what observations should we make of them?" With timescales as short as seconds to hours we require the ability to identify, classify and report any detection in time to allow for follow-up observations before the initial outburst fades. Identification and classification must, therefore, be undertaken in almost real-time with probabilistic classifications that incorporate our uncertainties about our classification together with the ability for algorithms to learn based on a posteriori information from earlier classifications. Algorithms must be able to predict what additional information might be needed to improve (or exclude) the likelihood of a given classification and to specify which parameters led to the source being classified as anomalous. Small errors in the identification and classification of these anomalous sources will swamp any underlying signal. The LSST will detect $7.5\text{x}10^8$ sources **every night**. Even for the most numerous transient events (SNe) this corresponds to less than 10^{-5} of the total number of sources identified being transient. For the most energetic bursters the magnitude of the challenge is 500-fold larger. Algorithms for identifying anomalies and variability must, therefore,

be robust to false positives and missing data and must account for the cadence in how we sample the time domain, variations in the quality of the data due to atmospheric conditions, changes in the performance of the telescope and camera and the possibility that we observe sources at different wavelengths at different times.

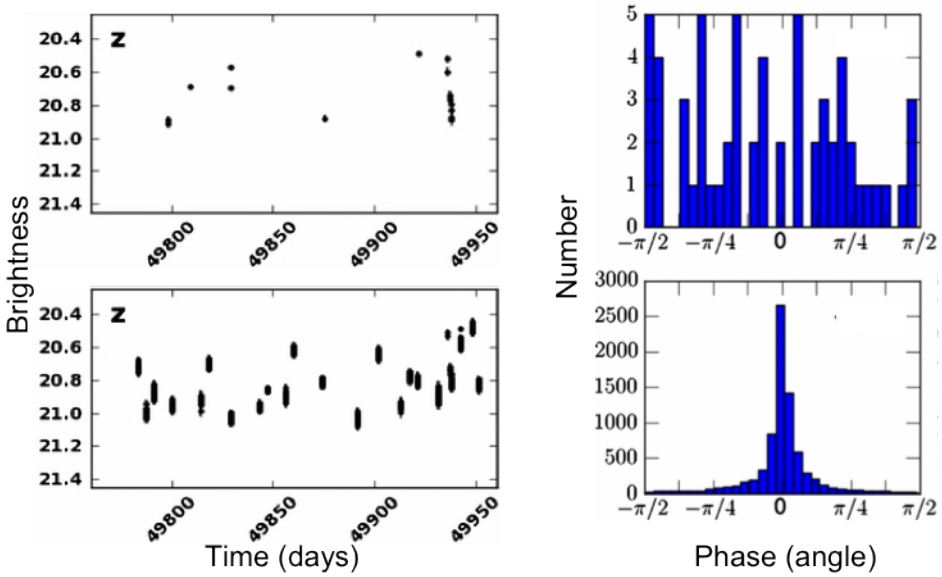


FIG. 5: Taken from **(author?)** [Oluseyi *et al.* 2012] the left panels shows the expected sampling of a light curve by the LSST at its main survey cadence (top left) and for a higher sampling rate that will be employed for 10% of the LSST survey (bottom left). The right panels shows a measure of the how well the light curves can be phased (i.e. combining the observations while solving for the periodicity of the data). This illustrates the significant effect on scientific output of more and more targeted data.

A. Active Learning for Transient Classification

Real-time automatic classification of objects is already widely acknowledged as a necessary support technology for the forthcoming age of survey cosmology [Djorgovski *et al.* 2011, Richards *et al.* 2011, Richards *et al.* 2012a, Graham *et al.* 2012, Mahabal *et al.* 2008a, Mahabal *et al.* 2011a]. A great deal of work has already been accomplished developing algorithms that can learn the classification of an object given data that are well sampled and that are analysed in a batch mode. Even with these idealized data sets misclassification rates are typically 30% or greater. For the DES and the LSST data streams these challenges are exacerbated as the data are poorly sampled in the temporal domain and the classifications must be undertaken in almost real-time (within 60s of an observation for the case of the LSST). For the case of the LSST 90% of the data will be sampled at a rate 20-times less than the highest cadence (with the highest sampling rates undertaken in the "Deep Drilling Fields"). Figure 5 shows how this reduction in sampling degrades our ability to constrain a light curve (and thereby classify a source).

We propose to break down the observations of a given object into $\{\Delta m, \Delta t\}$ pairs (where m is brightness and t is time) as an input to a GP regression and then classify the object based on that reconstruction. Extension of the GP to return a probabilistic reconstruction will enable **XXX**

something. For lower signal-to-noise observations we propose to build the classification model using basis functions derived from a decomposition of light curves learned from the "Deep Drilling Fields" data (i.e. data with high temporal sampling).

B. Group-based Active Feature Acquisition and Active Search for Transient Classification

The active learning, active search, and active feature acquisition choices for transient classification all must be made in an online, streaming fashion. Rather than considering an entire pool of test objects, they appear one at a time as they are detected and the algorithm must decide whether and how to follow up on each immediately as they are detected. The input features for transient object classification will be a variable combination photometric and morphologic measures taken at different increments in time. With this proposal we will introduce an information-theoretical approach based on the output probability distributions generated by our next generation GPs which will allow us determine the utility of different types of follow-up observation in real time as observations are made, directing experiments as they are performed. These determinations will not be based on cross-validation testing, but rather based on which observations we expect to yield the most significant decreases in the entropy of our probabilistic model. Huijse *et al.* (2011,2012) use a similar scheme to find the correlation between observations within a single light curve. We will attempt to find the type of observation with the greatest information content.

In *active search* the ultimate objective in following up detected anomalies is to maximize the number of interesting anomalies classified and characterized while staying within a budget of follow-up observations or experiments. The problem and the Bayesian optimal algorithm for it are described in Garnett et al 2011 and Garnett et al 2012a. We propose to expand on these approaches using scan statistics to consider not just individual anomalies but also group anomaly detection algorithms that consider arbitrary groupings of self-similar anomalous records (Neil and Moore 2005).

Techniques such as GPs offer two distinct advantages which render them particularly amenable to integration into active learning frameworks. Variance can be used to assign an uncertainty to the predicted value and the covariance provides the structure between all pairs of predicted and observed outputs. Bryan et al (2007) and Daniel et al (2012) use these aspects to great affect, treating the determined variances as a measure of the information that can be learned by promoting a query point to a new training point and thus learning the likelihood surface of a theory space with greater efficiency than traditional MCMC methods.

A limitation of current classification approaches is that they are optimized for the case of binary classifications: "Is something a quasar, or is it not?" [Kim *et al.* 2011, Pichara *et al.* 2012], "Is an object a real astrophysical transient, or is it an instrumental artifact?" [Brink *et al.* 2012] rather than learning a model over the full range of classifications.

C. Caching structures for computational efficiency for real-time classification

Given the size and dimensionality of data from the next generation of surveys efficient caching and search techniques are required for both anomaly detection and classification and to optimize sampling strategies whose cost function is, effectively, our ability to constrain the underlying cosmology. We will address these computational issues with a set of caching, sampling, and search techniques.

a. AD-Trees. The AD-Tree [Moore and Lee 1998] caching structure has the feature that it can answer any count query (i.e. how many records match query X?) in time independent of the number of records. It is similar to an OLAP datacube, except that it caches the answer to all possible queries in a single structure rather than requiring the building of many cubes. Specialized pruning tricks keep its memory footprint modest. Our algorithms will rely heavily on this structure. An extension called TCube stores time series in the AD-Tree [Sabhnani *et al.* 2006]. This allows queries of the form: *return the aggregated time series of counts for all records matching query X.* Additional pruning tricks based on sparse vector representations allow this to be memory efficient. The TCube structure will be used primarily for identifying time-varying anomalies, and will be extended to images.

We propose additional modifications that will cache sufficient statistics. In addition to the counts of records matching a query, it will store means and covariances of continuous-valued attributes of those records. This is especially useful for search-based comparison techniques because those sufficient statistics can be used to build classifiers such as LDA or QDA, which can quickly assess whether there is a difference between two populations of objects. This allows an efficient rule-based search for sub-populations with differences between them in the same way it was done in the Radsearch algorithm [Moore and Schneider 2002].

b. KD-Trees. KD-Trees are classic data structures used extensively in machine learning and N-body simulations. A traditional use is to retrieve nearest neighbors in $log\ N$ time. By storing sufficient statistics at the internal nodes of the trees, one can retrieve the statistics needed for operations such as kernel regression very efficiently [Moore *et al.* 1997]. A more interesting extension for this work is the use of KD-Trees to speed up the EM-based mixture model algorithm, which achieved as much as three orders of magnitude speedup over previous methods [Moore 1999]. We will modify this approach to speed up the EM algorithm proposed for learning dynamical models.

c. Hierarchical search structures. The field of scan statistics [Kulldorff 1997, Neill and Moore 2003] has developed a set of hierarchical structures specially designed for spatial-temporal data. We will extend these for use in identifying interesting space-varying and time-varying phenomenon. The key feature is that when searching for spatio-temporal anomalies or matches to patterns, the right choice of scoring metric makes it possible to significantly prune the search space (i.e. based on the sufficient statistics for this region, I can determine that no sub-region will fit my criteria).

The algorithms we propose may be further sped up by parallelizing across CPUs on a grid. We will use this technique when possible using standard cloud computing paradigms [Gardner *et al.* 2007] and expect our efforts on efficient algorithms and parallelization to yield speedups that compound.

VI. TIMELINE, DEVELOPMENT PLAN, AND RESPONSIBILITIES

We will focus on multiple projects in each of the three years of this program. Algorithms will be released as they are developed in a twice-yearly release cycle. Algorithm development will be undertaken using data from the SDSS, the DES survey, and a series of detailed simulations developed for the LSST [Connolly *et al.* 2010]. Each item below will indicate which of the PIs will take the lead in that stage of the program.

Year 1:

- (Schneider) Develop GP algorithms that are robust against sparse training data and capable of modeling multi-modal probability distributions.

- (Connolly) Build a photometric redshift algorithm based on these next generation GPs. Test the algorithm against data from the SDSS, DES, and simulated LSST.
- (Schneider) Develop active learning methods optimized for our next generation GPs. Initially focus on myopic criteria, selecting data for follow up whose classification considered in a vacuum is the most advantageous. Subsequently develop methods to implement one- and two-step-lookahead criteria.
- (Connolly) Test the developed active learning methods on photometric redshift data.
- (Schneider) Develop a software package that provides access to our next generation GPs for any researcher using any data set.

Year 2:

- (Schneider) Develop active feature acquisition algorithms based on the probability distributions output by our next generation GPs. Use them to determine what additional information (e.g., morphological or angular correlation function) provides the most helpful supplemental information in the case of photometric redshifts.
- (Connolly) Extend our next generation GPs to the problem of transient-classification. Analyze stellar sources where there are known populations with defined variability signatures (e.g. Cepheid variables) to provide validation for the analysis.
- (Connolly) Extend the active feature acquisition algorithms developed for photometric redshifts to the problem of determining the next best measurement for transient classification. Validate these algorithms on SDSS, DES, and simulated LSST observations, witholding data from the algorithm to simulate the case of incomplete, ongoing observations.
- (Schneider) Generalize our active learning and active feature acquisition algorithms so that they can accept output from any model-fitting algorithm, not just our next-generation GPs. Develop a software package providing access to these algorithms.

Year 3:

- (Connolly) Integrate our active learning and active feature acquisition algorithms to produce an active search algorithm that can identify interesting objects and optimize their follow-up in real time data streams. Begin running this algorithm on DES data as it is made public.
- (Schneider) Develop a software package that will provide access to all of our developed algorithms in a general way so that researchers can provide input and output data from any experiment and receive appropriate follow-up recommendations.

[Abazajian *et al.* 2009] Abazajian, K. N.*et al.* [SDSS Collaboration] 2009, The Astrophysical Journal Supplement Series **182**, 543 [arXiv:0812.0649 [astro-ph]].

[Bergé *et al.* 2012] Bergé, J., Price, S., Amara, A., and Rhodes, J. 2012, Monthly Notices of the Royal Astronomical Society **419**, 2356

[Bonfield *et al.* 2010] Bonfield, D. G., Sun, Y., Davey, N., Jarvis, M. J., Abdalla, F. B., Banerji, M., Adams, R. G. 2010, Monthly Notices of the Royal Astronomical Society **405** 987

[Brammer *et al.* 2008] Brammer, G. B., van Dokkum, P. G., and Coppi, P. 2008, The Astrophysical Journal **686**, 1503

[Brink *et al.* 2012] Brink, Henrik, Richards, Joseph W., Poznanski, Dovi, Bloom, Joshua S., Rice, John, Negahban, Sahand, and Wainwright, Martin 2012, arXiv:1209.3775

[Bryan *et al.* 2007] Bryan, B., Schneider, J., Miller, C. J., Nichol, R. C., Genovese, C., and Wasserman, L., 2007, The Astrophysical Journal **665**, 25

[Budavári 2008] Budavári, T. 2008 The Astrophysical Journal **695**, 747

[Collister and Lahav 2004] Collister, A. A. and Lahav, O. 2004, Publications of the Astronomical Society of the Pacific **116**, 345

[Connolly *et al.* 2010] Connolly, A. J., Peterson, J., Jernigan, J. Garrett, Abel, Robert, Bankert, Justin, Chang, Chihway, Claver, Charles F., Gibson, Robert, Gilmore, David K., Grace, Emily, Johnes, R. Lynne, Ivezic, Zeljko, Jee, James, Juric, Mario, Kahn, Steven M., Krabbendam, Victor L., Krughoff, Simon, Lorenz, Suzanne, Pizagno, Jame,s Rasmussen, Andrew, Todd, Nathan, Tyson, J. Anthon, Young, Mallory 2010, Proceedings of the SPIE **7738** 773810

[Daniel *et al.* 2011] Daniel, S. F., Connolly, A. J., Schneider, J., Vanderplas, J. and Xiong, L. The Astronomical Journal **142**, 203 (2011) [arXiv:1110.4646 [astro-ph.SR]].

[Daniel *et al.* 2012] Daniel, S. F., Connolly, A. J., and Schneider, J. 2012 [arXiv:1205.2708]

[Djorgovski *et al.* 2011] Djorgovski, S. J., Donalek, C., Mahabal, A. A., Moghaddam, B., Turmon, M., Graham, M. J., Drake, A. J., Sharma, N. and Chen, Y. 2011 [arXiv:1110.4655] to appear in Statistical Analysis and Data Mining, ref. proc. CIDU 2011 conf., eds. A. Srivastava and N. Chawla

[Gardner *et al.* 2007] Gardner, J. P., Connolly, A. J., and McBride, C. 2007, "Enabling rapid development of parallel tree search applications," in "Proceedings of the 2007 Workshop on Challenges of Large Applications in Distributed Environments (CLADE)"

[Garnett *et al.* 2011] Garnett, R., Krishnamurhty, Y., Wang, D., Schneider, J., and Mann, R. 2011, "Bayesian Optimal Active Search on Graphs," KDD Workshop on Mining and Learning with Graphs

[Garnett *et al.* 2012a] Garnett, R., Krishnamurthy, Y., Xiong, X., Schneider, J., and Mann, R. 2012a, "Bayesian Optimal Active Search and Surveying," International Conference on Machine Learning

[Garnett *et al.* 2012b] Garnett, R., Ho, S., and Schneider, J. 2012b, "Gaussian Processes for Identifying Damped Lyman-alpha Systems in Spectroscopic Surveys," Neural Information Processing Systems workshop on Modern Nonparametric Methods in Machine Learning

[Graham *et al.* 2012] Graham, M. J., Djorgovski, S. G., Mahabal, A., Donalek, C., Drake, A., Longo, G. 2012 [arXiv:1208.2480] to appear in special issue of Distributed and Parallel Databases on Data Intensive eScience

[Gray *et al.* 2004] Gray, A. G., Moore, A. W., Nichol, R. C., Connolly, A. J., Genovese, C., and Wasserman, L. 2004, Astronomical Society of the Pacific Conference Proceedings **314** 249

[Huijse *et al.* 2011] Huijse, Pablo, Estévez, Pablo A., Zegers, Pablo, Príncipe, Jose C., and Protopapas, Pavlos 2011, IEEE Signal Processing Letters **18**, 371

[Huijse *et al.* 2012] Huijse, Pablo, Estévez, Pablo A., Protopapas, Pavlos, Zebers, Pablo, and Príncipe, José C. 2012, arXiv:1212.2398

[Huterer *et al.* 2006] Huterer, D., Takada, M., Bernstein, G., and Jain, B. 2006, Monthly Notices of the Royal Astronomical Society **366**, 101

[Kaufman *et al.* 2011] Kaufman, Cari G., Bingham, Derek, Habib, Salman, Heitmann, Katrin, and Frieman, Joshua A. 2011, The Annals of Applied Statistitcs **5** 2470

[Kim *et al.* 2011] Kim, Dae-Won, Protopapas, Pavlos, Byun, Young-Ik, Alcock, Charles, Khardon, Roni, and Trichas, Markos 2011, arXiv:1101.3316

[Kitching *et al.* 2008] Kitching, T. D., Taylor, A. N., and Heavens, A. F. 2008, Monthly Notices of the Royal Astronomical Society **389** 173

[Kubica *et al.* 2007] Kubica, J., Denneau, L., Jr., Moore, A., Jedicke, R., Connolly, A. 2007, Astronomical Society of the Pacific Conference Series **376** 395

[Kulldorff 1997] Kulldorrff, M. 1997, Communications in Statistics – Theory and Methods **26** 1481

[Ma *et al.* 2012] Ma, Y., Garnett, R., and Schneider, J. 2012, "Submodularity in Batch Active Learning and Survey Problems on Gaussian Random Fields," Neural Information Processing Systems workshop on Discrete Optimization in Machine Learning

[MacKay 1992] MacKay, David 1992, Neural Computation **4** 415

[Mahabal *et al.* 2008a] Mahabal, A., Djorgovski, S. G., Turmon, M., Jewell, J., Williams, R. R., Drake, A. J., Graham, M. G., Donalek, C., Glikman, E., and the Palomar-QUEST Team 2008a, Astronomische Nachrichten **329**, 288

[Mahabal *et al.* 2008b] Mahabal, A., Djorgovski, S. G., Williams, R., Drake, A., Donalek, C., Graham, M., Moghaddam, B., Turmon, M., Jewell, J., Khosla, A., and Hensley, B. 2008b [arXiv:0810.4527] to appear in proceedings fo the Class2008 conference (Classification and Discovery in Large Astronomical Surveys, Ringberg Castle, 14-17 October 2008)

[Mahabal *et al.* 2011a] Mahabal, A. A., Donalek, C., Djorgovski, S. J., Drake, A. J., Graham, M. J., Williams, R., Chen, Y., Moghaddam, B., and Turmon, M. 2011a, [arxiv:1111.3699] to appear in Proc. IAU 285, "New Horizons in Transient Astronomy," Oxford, September 2011

[McBride *et al.* 2011a] McBride, C. K., Connolly, A. J., Gardner, J. P., Scranton, R., Newman, J. A., Scoccimarro, R., Zehavi, I., and Schneider, D. P. 2011a, The Astrophysical Journal, **726**, 13

[McBride *et al.* 2011b] McBride, C. K., Connolly, A. J., Gardner, J. P., Scranton, R., Scoccimarro, R., Berlind, A. A., Marín, F., and Schneider, D. P. 2011b, The Astrophysical Journal **739**, 85

[Moore *et al.* 1997] Moore, A., Schneider, J., and Deng, K. 1997, "Efficient Locally Weighted Polynomial Regression Predictions," in "International Conference on Machine Learning"

[Moore and Lee 1998] Moore, A., and Lee, M. 1998, Journal of Artificial Intelligence Research **8** 67

[Moore 1999] Moore, A. 1999, "Very Fast EM-based Mixture Model Clustering using Multiresolution Kd-trees," in "Neural Information Processing Systems 11" pp. 543-549

[Moore *et al.* 2000] Moore, A., Connolly, A., Genovese, C., Grone, L., Kanidoris, N., Nichol, R., Schneider, J., Szalay, A., Szapudi, I., and Wasserman, L. 2000, "Fast Algorithms and Efficient Statistics: N-point Correlation Functions," in MPA/MPE/ESO Conference on Mining the Sky [arXiv:astro-ph/0012333]

[Moore and Schneider 2002] Moore, A. W., and Schneider, J. G. 2002, "RADSEARCH: A new approach for finding optimal rules efficiently from dense datasets," in "AAAI-2002"

[Morgan *et al.* 2011] Morgan, A.N., Long, James, Richards, Joseph W., Broderick, Tamara, Butler, Nathaniel R., and Bloom, Joshua S. 2011, arXiv:1112.3654

[Nakajima *et al.* 2012] Nakajima, R., Mandelbaum, R., Seljak, U., Cohn, J. D., Reyes, R., and Cool, R. 2012, Monthly Notices of the Royal Astronomical Society **420**, 3240 [arXiv:1107.1395]

[Neill and Moore 2003] Neill, D. and Moore, A. 2003, "A fast multi-resolution method for detection of significant spatial overdensities," Carnegie Mellon University Computer Science Department paper 2175

[Nichol *et al.* 2006] Nichol, R. C., Sheth, R. K., Suto, Y., Gray, A. J., Kayo, I., Wechsler, R. H., Marin, F., Kulkarni, G., Blanton, M., Connolly, A. J., Gardner, J. P., Jain, B., Miller, C. J., Moore, A. W., Pope, A., Pun, J., Schneider, D., Schneider, J., Szalay, A., Szapudi, I., Zehavi, I., Bahcall, N. A., Csabai, I., Brinkmann, J. 2006, Monthly Notices of the Royal Astronomical Society **368**, 1507

[Oluseyi *et al.* 2012] Oluseyi, Hakeem M., Becker, Andrew C., Culliton, Christopher, Furqan, Muhammad, Hoadley, Keri L., Regencia, Paul, Wells, Akeem J., Ivezìc, Zeljko, Jones, R. Lynne, Krughoff, K. Simon, and Sesar, Branimir (2012), The Astronomical Joural **144** 9

[Pichara *et al.* 2012] Pichara, K., Protopapas, P., Kim, D.-W., Marquette, J.-B., and Tisserand, P. 2012, Monthly Notices of the Royal Astronomical Society, **427** 1284

[Poczos and Schneider 2011] Poczos, B. and Schneider, J. 2011, "On the Estimation of alpha-Divergences," Artificial Intelligence and Statistics (AISTATS)

[Poczos *et al.* 2011] Poczos, B., Xiong, L., and Schneider, J. 2011, "Nonparametric Divergence Estimation with Applications to Machine Learning on Distributions," Uncertainty in Artificial Intelligence

[Poczos *et al.* 2012] Poczos, B., Xiong, L., Sutherland, D., and Schneider, J. 2012, "Nonparametric Kernel

Estimators for Image Classification," IEEE Conference on Computer Vision and Pattern Recognition
[Rasmussen and Williams 2006] Rasmussen, C. E. and Williams, C. K. I., 2006, "Gaussian Processes for Machine Learning" `http://www.GaussianProcess.org/gpml/`
[Richards *et al.* 2011] Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., and Rischard, M. 2011, The Astrophysical Journal **733**, 10
[Richards *et al.* 2012a] Richards, J. W., Starr, D. L., Brink, H., Miller, A. A., Bloom, J. S., Butler, N. R., James, J. B., Long, J. P., and Rice, J. 2012a The Astrophysical Journal **744**, 192
[Sabhnani *et al.* 2006] Sabhnani, M., Moore, A., and Dubrawski, A. 2006, "T-Cube: A Data Structure for Fast Extraction of Time Series from Large Datasets," technical report number CMU-ML-06-104
[Scranton *et al.* 2002] Scranton, R., Johnston, D., Dodelson, S., Frieman, J. A., Connolly, A., Eisenstein, D. J., Gunn, J. E., Hui, L., Jain, B., Kent, S., Loveday, J., Narayanan, V., Nichol, R. C., O'Connell, L., Soccimarro, R., Sheth, R. K., Stebbins, A., Strauss, M. A., Szalay, A. S., Sapudi, I., Tegmark, M., Vogeley, M., Zehavi, I., Annis, J., Bahcall, N. A., Brinkman, J., Csabai, I., Hindsley, R., Ivezic, Z., Kim, R. S. J., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T., Munn, J., Peoples, J., Pier, J., Richards, G. T., Rockosi, C., Schlegel, D., Schneider, D. P., Stoughton, C., Tucker, D. L., Yanny, B., York, D. G. 2002, The Astrophysical Journal **579**, 48
[Settles 2009] Settles, B. 2009, "Active Learning Literature Survey," Computer Sciences Technical Report 1648, University of Wisconsin-Madison, `http://pages.cs.wisc.edu/~bsettles/active-learning/`
[Shafieloo *et al.* 2012] Shafieloo, A., Kim, A. G., and Linder, E. V. 2012, Physical Review D **85**, 123530 [arXiv:1204.2272]
[Skibba *et al.* 2006] Skibba, R., Sheth, R. K., Connolly, A. J., and Scranton, R. 2006, Monthly Notices of the Royal Astronomical Society, **369**, 68
[Straf 2003] Straf, M. L. 2003, Journal of the American Statistical Association **98**, 1
[Szapudi *et al.* 2002] Szapud, I., Frieman, J. A., Scoccimarro, R., Szalay, A. S., Connolly, A. J., Dodelson, S., Eisenstein, D. J., Gunn, J. E., Johnston, D., Kent, S., Loveday, J., Meiksin, A., Nichol, R. C., Scranton, R., Stebbins, A., Vogeley, M. S., Annis, J., Bahcall, N. A., Brinkman, J., Csabai, I., Doi, M., Fukigita, M., Ivezić, Ž., Kim, R. S. J., Knapp, G. R., Lamb, D. Q., Lee, B. C., Lupton, R. H., McKay, T. A., Munn, J., Peoples, J., Pier, J., Rockosi, C., Schlegel, D., Stoughtfon, C., Tucker, D. L., Yanny, B., York, D. G. 2002, The Astrophysical Journal **570**, 75
[Tesch *et al.* 2011a] Tesch, M., Schneider, J., and Choset, H. 2011a, "Adapting Control Policies for Expensive Systems to Changing Environments," in "IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)"
[Tesch *et al.* 2011b] Tesch, M., Schneider, J., and Choset, H. 2011b, "Using Response Surfaces and Expected Improvements to Optimize Snake Robot Gait Parameters," in "IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)"
[Tesch *et al.* 2013] Tesch, M., Schneider, J., and Choset, H. 2013, "Expensive Funciton Optimization with Stochastic Binary Outcomes," in "International Conference on Machine Learning (ICML)"
[Vanderplas and Connolly 2009] Vanderplas, J. and Connolly, A. J. 2009, Astronomical Journal **138**, 1365
[Wang *et al.* 2011] Wang, Yuyang, Khardon, Roni, and Protopapas, Pavlos 2011 arXiv:1111.1315
[Wang *et al.* 2012] Wang, Yuyang, Khardon, Roni, and Protopapas, Pavlos 2012 arXiv:1203.0970
[Xiong *et al.* 2011a] Xiong, L., Poczos, B., Schneider, J., Connolly, A., Vanderplas, J. 2011a, "Hierarchical Probabilistic Models for Group Anomaly Detection," Artificial Intelligence and Statistics (AISTATS)
[Xiong *et al.* 2011b] Xiong, L., Poczos, B., and Schneider, J. 2011, "Group Anomaly Detection using Flexible Genre Models," Neural Information Processing Systems
[Yip *et al.* 2004] Yip, C. W., Connolly, A. J., Szalay, A. S., Budavári, T., SubbaRao, M., Frieman, J. A., Nichol, R. C., Hopkins, A. M., York, D. G., Okamura, S., Brinkmann, J., Csabai, I., Thakar, A. R., Fukugita, M., and Ivezić, Ž. 2004, The Astronomical Journal **128**, 585
[Zhang and Schneider 2010a] Zhang, Y. and Schneider, J. 2010a, "Projection Penalties: Dimension Reduction without Loss," International Conference on Machine Learning
[Zhang and Schneider 2010b] Zhang, Y. and Schneider, J. 2010b, "Learning Multiple Tasks with a Sparse Matrix-Normal Penalty," Neural Information Processing Systems
[Zhang *et al.* 2010] Zhang, Y., Schneider, J., and Dubrawski, A. 2010, "Learning Compressible Models," Proceedings of SIAM Data Mining Conference
[Zhang and Schneider 2011] Zhang, Y. and Schneider, J 2011, "Multi-label Output Codes using Canonical

Correlation Analysis," Artificial Intelligence and Statistics
[Zhang and Schneider 2012] Zhang, Y. and Schneider, J. 2012, "Maximum Margin Output Coding," International Conference on Machine Learning
[1] http://www.darkenergysurvey.org
[2] http://www.lsst.org

Facilities and Other Resources

Physical Facilities:
The Astronomy Department has ample office space for faculty, staff, and students.

General Compute Resources:
The astronomy department maintains a wide variety of state-of-the-art computing facilities for research and instructional use. General-purpose research computing is provided by over 70 Unix-based workstations and servers, located in laboratories, machine rooms and offices. The back-end infrastructure is comprised of general-purpose compute, file, web, mail and print servers, operating as a well-integrated Linux and environment. Departmental networking utilizes 1 and 10 gigabit Ethernet connections to servers and desktop machines, and a wireless network provides 802.11b/g connectivity throughout the entire building.

Research-specific Resources:
The PI has access to multiple clusters including a dedicated 500-code Linux cluster with Infiniband for low latency communication, and a multi-core database system attached to 150TB of storage.