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						
			Year 1	Year 2	Year 3	Total
	Names	Institution	Budget	Budget	Budget	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

1. 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 look ahead to the end of all possible experiments 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 our 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 requires developing fundamentally new approaches to the analysis of data at a scale, speed, and complexity beyond the capabilities of current automated machine learning methods.

1.1. 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 therefore we will build our methods around them. We note, however, objectives 2-4 below will result in algorithms that do

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

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

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 will 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 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 preliminary implementation 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 3.1 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.

(3) 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 3.2) and improved classification and online follow-up of transient astronomical objects. 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").

1.2. 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 (**?**) and that enable rapid characterization of orbital tracks in sparsely sampled temporal data (**?**). 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 **(author?)** (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 spectroscopic 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. On sabbatical at Google, Connolly led the development of Sky in Google Earth (aka Google Sky; http://earth.google.com/sky), which was similarly successful at engaging the public.

2. 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 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).

3. 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 these 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.

3.1. Gaussian Processes

Gaussian processes (GPs) model the output of an unknown, noisy function in multi-dimensional data 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 data 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 astrophysics 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 stars. Huisje *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 (**? ? ?**).

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 function, 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:

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

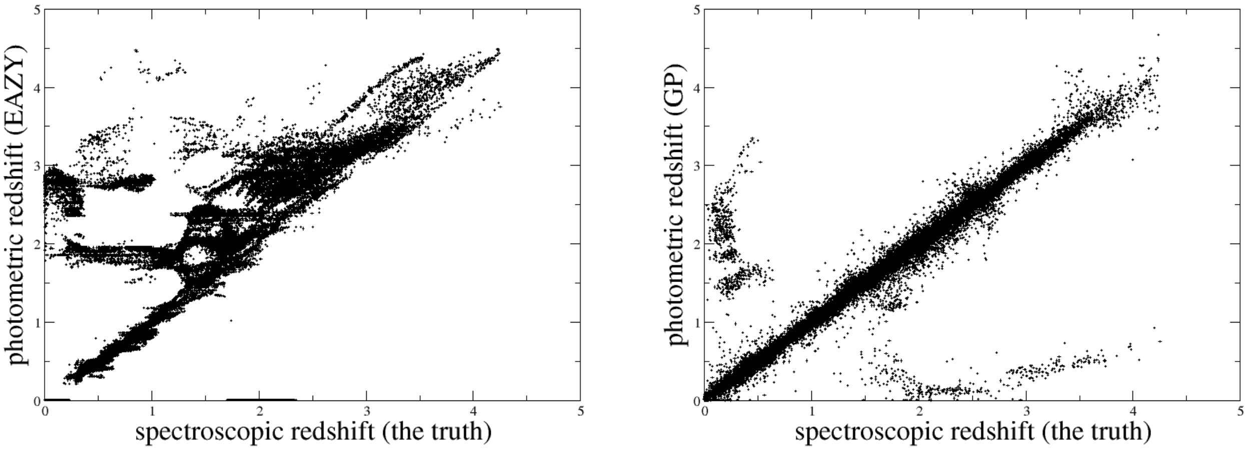
Here, K is the matrix of covariances between all training points, σ^2 is the variance of Gaussian noise added to each observed value of y, $\vec{y}$ is the training data outputs, $\bar{y}$ is the algebraic mean of the elements of $\vec{y}$, and K_q is the vector of covariances between the query point 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.

3.2. 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. Astronomers 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 $z_{\mathrm{photometric}} = z_{\mathrm{spectroscopic}}$ line, there is significant scatter in the results. Figure 1(b) plots the results from running the same simulated galaxies through GP-based algorithm that attempts to learn the full probability distribution $P(z_{\mathrm{photometric}})$. 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 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 publically-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 only 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 $z_{\mathrm{photometric}}$ derived from multiple instantiations of ANNz. There is no guarantee that it represents the true, probabilistic uncertainty in $z_{\mathrm{photometric}}$. Figure 2 plots the mean value of $\ln[P(\mathrm{truth})]$, i.e. the value of the logarithm $P(z_{\mathrm{photometric}})$ at the point $z_{\mathrm{photometric}} = z_{\mathrm{spectroscopic}}$, 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 $z_{\mathrm{photometric}} = z_{\mathrm{spectroscopic}}$ 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 $z_{\mathrm{spectroscopic}}$-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 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.

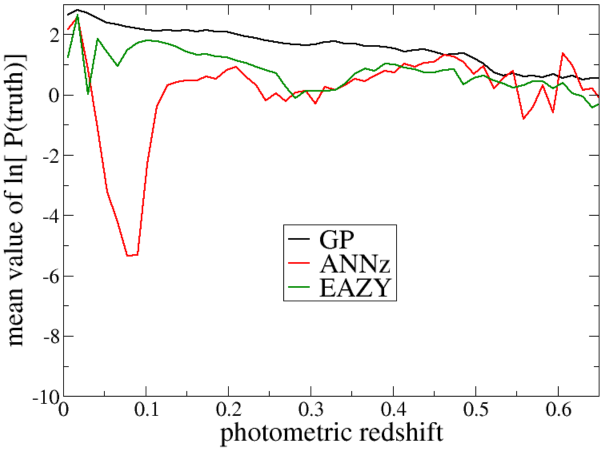


Fig. 2.— The mean value of ln$[P(\text{truth})]$ as a function of photometric redshift (the vertical axes in Figures 1) for all three algorithms under consideration using real data taken from the Sloan Digital Sky Survey (Abazajian *et al.* 2009). In this latter case, the algorithms are trained on 21,000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

3.3. 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 data 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_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 parameters $\vec{\theta}_q$ starts from the same value and equation 2 expands the departures from 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, attempting to find a more robust foundation off of which to build our GPs. One proposal would be to performa principal component analysis on the input $\{\vec{\theta}\}$ and set $\bar{y}$ according to a linear regression on the principal component most correlated with the outputs $\vec{y}$. Preliminary results indicate that this gives a 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 data space. We will develop dynamic algorithms again 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. While this certainly resulted in acceptable behavior in our test cases, it is by no means guaranteed that the choice we made (a Gaussian function in data space) will always be the best choice. We will explore ways to help the GP learn the most appropriate form of the covariogram based on the training data at hand.. The simplest such possibility is to set the form of the covariance function using cross-validation. We will also experiment with information-theoretical approaches in an attempt to find a relationship between the form of the covariance function and the entropy of the distribution produced by the GP.

3.4. Fortifying Gaussian Processes against degenerate training data

Additionally, it will be necessary to build GPs that are robust against the possibility that degeneracies 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 $z_{\rm spectroscopic}$ galaxies that appear at low $z_{\rm photometric}$ 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 satisfied that you have found the best fit to the training data.

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

4. 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 (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

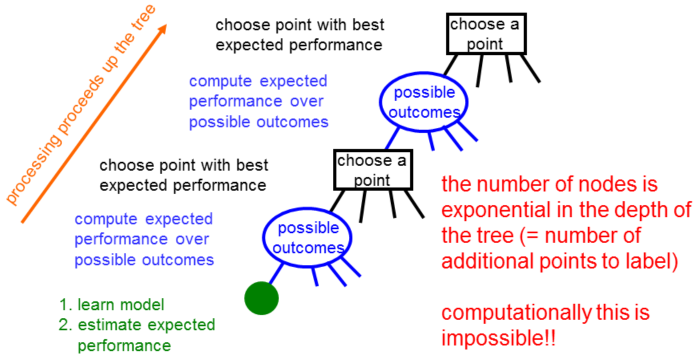

processing proceeds up the tree
choose point with best expected performance
choose a point
compute expected performance over possible outcomes
possible outcomes
choose point with best expected performance
choose a point
the number of nodes is exponential in the depth of the tree (= number of additional points to label)
compute expected performance over possible outcomes
possible outcomes
1. learn model
2. estimate expected performance
computationally this is impossible!!

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

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.

4.1. 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. 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 (**?**) 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 ($tr(\Sigma_q)$, also known as V-optimality). A further approximation is to do no look ahead and simply choose the data point corresponding to the largest diagonal element in $tr(\Sigma_q)$. 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.

- 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 look ahead 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 look ahead allowed more than compensates for the errors induced by pruning.

4.2. Active learning for photometric redshifts

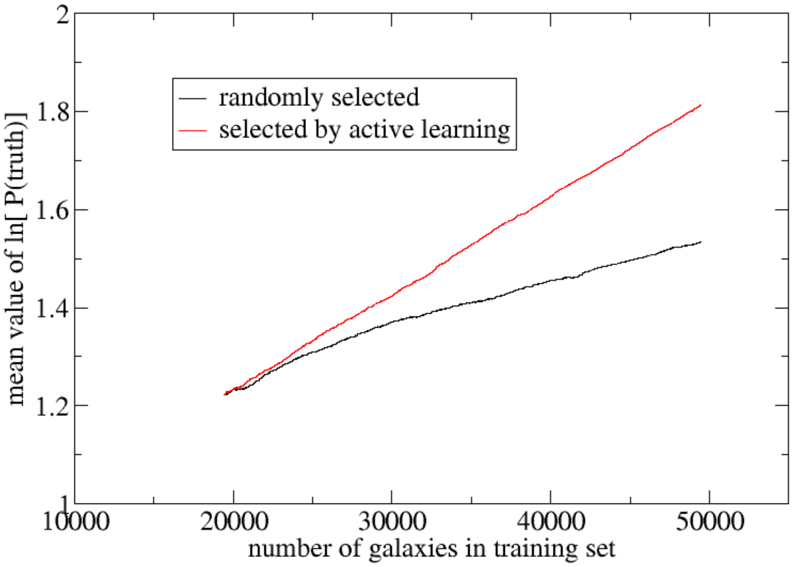


Fig. 4.— Active learning applied to classification according to a scalar function on a 6-dimensional data space. The horizontal axis is the size of the training data set. The vertical axis is the mean value of the output probability distribution at the true value of the scalar function. The black curve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

In both empirical and forward-fitting photometric redshift codes, biases arise from the fact that the distribution of training samples (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 $\ln[P(\text{truth})]$ 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. We will demonstrate our novel methods on real data sets that contain all these additional challenges.

5. 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. Fortunately, survey astronomy 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. It must be able to predict what additional information might be needed to improved (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 7.5x10^8 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.

5.1. Active Learning for Transient Classification

I have just copied-and-pasted the "active learning for transients" paragraphs from the previous draft

Real-time automatic classification of objects is already widely acknowledged as a necessary support technology for the forthcoming age of survey astronomy (Djorgovski *et al.* 2011; Richards *et al.* 2011; Richards *et al.* 2012a; Graham *et al.* 2012; Mahabal *et al.* 2008a; Mahabal *et al.* 2011a). Objects

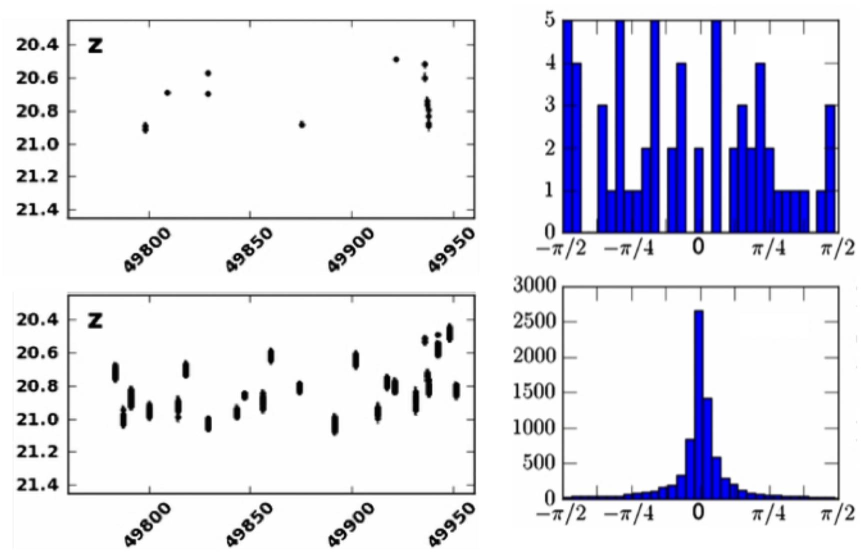


Fig. 5.— Taken from **(author?)** (Oluseyi *et al.* 2012) the left panels shows the expected sampling of a light curve by the LSST at its main survey cadence (top left) and for a higher sampling rate that will be employed for 10% of the LSST survey (bottom left). The right panels shows a measure of the 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.

will need to be categorized into known science classes so that novel or rare objects can be flagged for detailed follow-up observations. For transient events, algorithms must be able to make rapid decisions so that sources can be targeted for follow-up and classifications learned before objects return to their quiescent phases.

A great deal of work has already been done developing algorithms that can learn the classification of an object given a fixed set of observations and training data. Mahabal *et al.* (2008a,2011a,2011b) propose to break down the observations of a given object into $\{\Delta m, \Delta t\}$ pairs (where m is magnitude and t is time) and use the density of observations in this two-dimensional space as the basis for a Bayesian classification algorithm. Mahabal *et al.* (2008b) alternatively propose to use those same $\{\Delta m, \Delta t\}$ pairs as the input to a GP regression by which they will reconstruct the object's entire light curve as a function of time, and then classify the object based on that reconstruction. Anomaly detection has been attempted by decomposing light curves into basis functions of different timescales and looking for events that occur more rapidly than some fit baseline (Preston *et al.* 2009; Blocker and Protopapas 2013). Richards *et al.* (2011) use observations of transient objects to extract periodic (e.g. the amplitude and frequency of the first two Fourier modes of the object's light curve) and non-periodic (e.g. the variance and skewness of all of the magnitude observations taken, regardless of their separation in time) and feed those features into several tree-based classifiers. They find misclassification rates lower than 30% with their best method yielding a misclassification rate of 22.8%. Using only non-periodic features, which will be especially easy for survey telescopes to gather, rather than full light curves, they find a misclassification rate of between 26% and 28%. Bloom *et al.* (2011) also use a tree-based automatic classifier on Palomar Transient Factory data and find a 3.8% error rate when discriminating between four major classifications. Richards *et al.* consider a more complete set of 25 possible classifications. Clearly, many possibile approaches are available for the automated classification of transient objects, and not all of them rely upon highly detailed observations to function. None of the above algorithms, however, make any promises regarding their ability to deliver rapid recommendations for optimum follow-up observations in real time. This will be a significant contribution to all time-sensitive sciences.

Similarly, a large number of classifiers have been developed which 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 at redshift greater than 4 or is it not?" (Morgan *et al.* 2011), "Is an object a real astrophysical transient, or is it an instrumental artifact?" (Brink *et al.* 2012). The methods developed (Random Forests; Support Vector Machiens; etc.) are all useful and provide guidance for what can and should be attempted, however, they do not deliver the robust, probabilistic, multi-class identifiers that will be required by the rapid, big-data experiments of the next decades.

5.2. Active Feature Acquisition and Active Search for Transient Classification

The input features for transient object classification will be a variable combination photometric and morphologic measures taken at different increments in time. The output is a categorical variable indicating the class. Brink *et al.* (2012) and Richards *et al.* (2012b) consider this diversity of features in designing automatic classifiers on static data sets. They use cross-validation to find that some features are more useful than others and that the inclusion of all features degrades the performance of their classifier (see Figure 4 of both references). This determination of optimal inputs is at the heart of active feature acquisition. Our intention is to introduce an information-theoretical approach based on the output covariance matrix of equation 3 which will allow us to perform a similar determination in real time as observations are made, directing experiments as they are performed. We will not choose our follow-up observations 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. An information-theoretical approach has already been tried in the case of a static (i.e. off-line) dataset by Huisje *et al.* (2011, 2012) who use a the correntropy as a measure of the correlation between observed intensities at different time lags in already observed light curves. We will consider the effect on entropy of as-of-yet untaken observations. This will be a significant development towards the goal of real-time object 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 culmination of our proposed program of research will be to combine the methods devised above into a single combined streaming algorithm as described below.

The three goals of active learning, active search, and active feature acquisition will be combined in a staged set of decisions. When a new event or object is detected, the data will be used to provide and estimated physical model for the object. These models are the input variables for this object in the active learning and active search algorithms. In parallel, the active learning and active search methods will decide whether to follow-up on this object. If either of them selects the object, it is advanced to active feature acquisition. There additional observations on the object are selected and the process for this object repeats. An object that initially seemed interesting to one algorithm may cease to be so after additional observations or may be adopted by the other one. The process for one object terminates when neither active learning nor active search remains interested in it or the object class and light curves are characterized well enough that no more observations are required.

6. Timeline and Development Plan

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 (**?**).

Year 1:

- Develop GP algorithms that are robust against sparse training data and capable of modeling multi-modal probability distributions. Build a photometric redshift algorithm based on these next generation GPs. Test the algorithm against data from the SDSS, DES, and simulated LSST.
- Develop active learning algorithms to discover the optimum follow-up strategy for our photometric redshift algorithms. We will initially base these algorithms on myopic, single-step criteria (i.e., choosing the object whose follow-up, independent of all other galaxies, yields the greates scientific pay-off). We will move on to modeling one- and two-step-look-ahead methods to quantify the effect of follow-up observations on the entire learned photometric redshift model.
- Develop a software package that provides access to our next generation GPs for any researcher using any data set.

Year 2:

- Develop active feature acquisition algorithms to determine what additional information (e.g., morphological or angular correlation function) provides the most helpful supplement to traditional photometric measurements.
- 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.
- 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.
- 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:

- 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.
- 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.

REFERENCES

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

Abdalla, F. B., Banerji, M., Lahav, O., and Rashkov, V. 2011, Monthly Notices of the Royal Astronomical Society **417**, 1891

Abramo, L. R., Strauss, M. A., Lima, M., Hernández-Monteagudo, C., Lazkoz, R., Moles, M., de Oliveira, C. M., Sendra, I., Sodré Jr., L., and Storchi-Bergmann, T. 2012, Monthly Notices of the Royal Astronomical Society **423**, 3251

Albrecht, A., Bernstein, B., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., Huth, J., Kamionkowski, M., Kolb, E., Knox, L., Mather, J. C., Staggs, S., Suntzeff, N. B. (Dark Energy Task Force) 2006, "Report of the Dark Energy Task Force," `http://jdem.gsfc.nasa.gov/science/DETF_Report.pdf`

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

Blocker, Alexander W. and Protopapas, Pavlos 2013, arXiv:1301.3027

Bloom, J. S., Richards, J. W., Nugent, P. E., Quimby, R. M., Kasliwal, M. M., Starr, D. L., Posnanski, D., Ofek, E. O., Cenko, S. B., Butler, N. R., Kulkarni, S. R., Gal-Yam, A., and Law, N. 2011 [arXiv:1106.5491]

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, G. B., van Dokkum, P. G., and Coppi, P. 2008, The Astrophysical Journal **686**, 1503

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

Bryan, B., 2007, Ph.D. thesis `http://reports-archive.adm.cs.cmu.edu/anon/ml2007/abstracts/07-122.html`

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

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

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

Connolly, A. J., Peterson, J., Jernigan, J. G., Abel, R., Bankert, J., Chang, C., Claver, C. F., Gibson, R., Gilmore, D. K., Grace, E., Jones, R. L., Ivezic, Z., Jee, J., Juric, M., Kahn, S. M., Krabbendam, V. L., Krughoff, S., Lorenz, S., Pizagno, J., Rasmussen, A., Todd, N. Tyson, J. A., and Young, M. 2005, Society of the Photo-Optical Instrumentation Engineers (SPIE) Converence Series **7738**, 53

Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., and Wechsler, R. H. 2012, [arXiv:1207.3347]

Daniel, S. F. and Linder, E. V. 2010, Physical Review D **82**, 103523

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, S. F., Connolly, A. J., and Schneider, J. 2012 [arXiv:1205.2708]

Das, S., de Putter, R., Linder, E. V., and Nakajima, R. 2011, [arXiv:1102.5090]

Davis, T. M., Mörtsell, E., Sollerman, J., Becker, A. C., Blondin, S., Challis, P., Clocchiatti, A., Filippenko, A. V., Foley, R. J., Garnavich, P. M., Jha, S., Krisciunas, K., Kirshner, R. P., Leibundgut, B., Li, W., Matheson, T., Miknaitis, G., Pignata, G., Rest, A., Riess, A. G., Schmidt, B. P., Smith, R. C., Spyromilio, J., Stubbs, C. W., Suntzeff, N. B., Tonry, J. L., Wood-Vasey, W. M., and Zenteno, A. 2007, The Astrophysical Journal, **666**, 716

de Putter, R., Huterer, D. and Linder, E. V. 2010, Physical Review D **81**, 103513

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

Foster, Leslie, Waagen, Alex, Aijaz, Nabeela, Hurley, Michael, Luis, Apolonio, Rinsky, Joel, Satyavolu, Chandrika, Way, Michael J., Gazis, Paul, Srivastava, Ashok 2009, Journal of Machine Learning Research **10** 857

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

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, 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

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

Huijse, Pablo, Estévez, Pablo A., Protopapas, Pavlos, Zebers, Pablo, and Príncipe, José C. 2012, arXiv:1212.2398

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

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

Kim, Dae-Won, Protopapas, Pavlos, Byun, Young-Ik, Alcock, Charles, Khardon, Roni, and Trichas, Markos 2011, arXiv:1101.3316

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

Linder, Eric V. 2013, Journal of Cosmology and Astropartical Physics **1304** 031

Long, J. P., El Karoui, N., Rice, J. A., Richards, J. W., and Bloom, J. S. 2012, Publications of the Astronomical Society of the Pacific **124** 280

LSST Collaboration 2011, [arXiv:0805.2366] `http://www.lsst.org/lsst/overview/`

LSST Dark Energy Science Collaboration 2012, [arXiv:1211.0310]

LSST Science Collaborations 2009, "LSST Science Book", http://www.lsst.org/lsst/science/scibook

Ma, Z., Hu, H., and Huterer, D. 2006, The Astrophysical Journal **636**, 21

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, David 1992, Neural Computation **4** 415

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, 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, 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

Mahabal, A. A., Djorgovski, S. G., Drake, A. J., Donalek C., Graham, M J., Williams, R. D., Chen, Y., Moghaddam, B., Turmon, M., Beshore, E., and Larson, S. 2011b, Bulletin of the Astronomical Society of India **39**, 387

Mandelbaum, R., Seljak, U., Hirata, C. M., Bardelli, S., Bolzonella,!M., Bongiorno, A., Carollo, M., Contini, T., Cunha, C. E., Garilli, B., Iovino, A., Kambczyk, P, Kneib, J.-P., Knobel, C., Koo, D. C., Lamareille, F., Le Fèvre, O., Leborgne, J.-F., Lilly, S. J., Maier, C., Mainieri, V., Mignoli, M., Newman, J. A., Oesch, P. A., Perez-Montero, E., Ricciardelli, E., Scodeggio, M., Silverman, J., and Tasca, L. 2008, Monthly Notices of the Royal Astronomical Society **386**, 781

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, 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, 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]

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

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]

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

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

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

Poczos, B., Xiong, L., Sutherland, D., and Schneider, J. 2012, "Nonparametric Kernel Estimators for Image Classification," IEEE Conference on Computer Vision and Pattern Recognition

Preston, Dan, Protopapas, Pavlos, and Brodely, Carla 2009, arXiv:0901.3329

Rasmussen, C. E. and Williams, C. K. I., 2006, "Gaussian Processes for Machine Learning" `http://www.GaussianProcess.org/gpml/`

Richards, G. T., Nichols, R. C., Gray, A. G., Brunner, R. J., Lupton, R. H., Vanden Berk, D. E., Chong, S. S., Weinstein, M. A., Schneider, D. P., Anderson, S. F., Munn, J. A., Harris, H. C., Strauss, M. A., Fan, X., Gunn, J. E., Ivezić, Z., York, D. G., Brinkmann, J., and Moore, A. W. 2004, The Astrophysical Journal Supplement Series, **155**, 257

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, 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

Richards, Joseph W., Starr, Dan L., Miller, Adam A., Bloom, Joshua S., Butler, Nathaniel R., Brink, Henrik, and Crellin-Quick, Arien 2012b, The Astrophysical Journal Supplement Series **203** 32

Rosenfield, P., Connolly, A., Fay, J., Sayres, C., and Tofflemire, B. 2011, Astronomical Society of the Pacific Conference Series **443**, 109

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

Sesar, B., Stuart, J. S., Ivezić, Ž., Morgan, D. P., Becker, A. C., and Woźniak, P. 2011, The Astronomical Journal **142**, 190

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, A., Kim, A. G., and Linder, E. V. 2012, Physical Review D **85**, 123530 [arXiv:1204.2272]

Skibba, R., Sheth, R. K., Connolly, A. J., and Scranton, R. 2006, Monthly Notices of the Royal

Astronomical Society, **369**, 68

Straf, M. L. 2003, Journal of the American Statistical Association **98**, 1

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

Vanderplas, J. and Connolly, A. J. 2009, Astronomical Journal **138**, 1365

Wang, Yuyang, Khardon, Roni, and Protopapas, Pavlos 2011 arXiv:1111.1315

Wang, Yuyang, Khardon, Roni, and Protopapas, Pavlos 2012 arXiv:1203.0970

Wiley, K, Connolly, A. J., Gardner, J., Krughoff, S., Balazinska, M., Howe, B., Kwon, Y., and Bu, Y. 2011, Publication of the Astronomical Society of the Pacific **123**, 366

Xiong, L., Poczos, B., Schneider, J., Connolly, A., Vanderplas, J. 2011a, "Hierarchical Probabilistic Models for Group Anomaly Detection," Artificial Intelligence and Statistics (AISTATS)

Xiong, L., Poczos, B., and Schneider, J. 2011, "Group Anomaly Detection using Flexible Genre Models," Neural Information Processing Systems

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, Y. and Schneider, J. 2010a, "Projection Penalties: Dimension Reduction without Loss," International Conference on Machine Learning

Zhang, Y. and Schneider, J. 2010b, "Learning Multiple Tasks with a Sparse Matrix-Normal Penalty," Neural Information Processing Systems

Zhang, Y., Schneider, J., and Dubrawski, A. 2010, "Learning Compressible Models," Proceedings of SIAM Data Mining Conference

Zhang, Y. and Schneider, J 2011, "Multi-label Output Codes using Canonical Correlation Analysis," Artificial Intelligence and Statistics

Zhang, Y. and Schneider, J. 2012, "Maximum Margin Output Coding," International Conference on Machine Learning

This preprint was prepared with the AAS LaTeX macros v5.2.

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.