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, number, email:	
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:
University of Washington, PI Andrew Connolly
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

Learning in an Era of Uncertainty

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 can 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 noisy and incomplete. The goal of this proposal is to address these challenges through the development of a broad class of novel and scalable machine-learning techniques centered around the theme of active learning.

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 improves 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 improve the 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 the 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 approaches but at a substantial cost in performance. 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 inherently 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

We divide the problem of on-line big-data analysis into the following three problems.

Active Learning. Automatic regression and classification algorithms require input from human experts against which to calibrate themselves. Often, the resources required to generate this input (both in terms of human-hours and experimental apparatus) are considerable. We propose to develop algorithms to optimize the use of these resources by determining what events' or objects'

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

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

human classification will maximize the improvement of the automatic classifiers. We have already implemented a toy version of such a model on the problem of determing astronomical Doppler shifts (or redshifts) using broad-band photometric data. Our results (shown in Figure 4 below) give us hope that greater attention will significantly increase the efficiency of future experiments.

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 different model was learned for spectra 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 the light curves of supernovae (used in measuring the cosmic acceleration) from data with poor temporal sampling (i.e. increasing the number of supernovae that can be used for cosmological experiments by an order of magnitude). The key advantage of this approach is that the GPs naturally provide a mean and covariance for future unobserved observations. This uncertainty propagates through to class labels and we can use it to estimate the reduction in class uncertainty that will be gained by observing a given observation at a certain time. The observation yielding the greatest reduction in entropy for the class and the light curves for this object will be taken.

Active Search. Often, the bulk of data yielded by any experiment will be expected. Groundbreaking research requires that we develop a method for rapdily identifying and classifying novel events and objects while staying within a budget of follow-up observations. 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. However, the final performance objective is not the accuracy of the classifier, but rather the number of positives (i.e. objects from interesting classes) identified. This 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 and multiple discrete classes. We will design an algorithm that considers probability distributions across these classifications and how new observations affects these distributions with an eye towards reducing the entropy, or some other information-theoretical measure of uncertainty.

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 (an event or object as described above). We believe that this problem can be best addressed by treating our models non-parametrically and probabilistically, i.e. by solving for the probability distribution over the data's features. For this reason, we propose to base our work on the formalism of Gaussian Processes discussed in Section 3.1 below.

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 (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 to automatically classify 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 real-time 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 the 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 draft 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, DES[3] and the Large Synoptic Sky Survey, LSST[4]). 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.

Impact

On these scales, 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. Prime amongst these systematics is the estimation of distances to extragalactic sources, the identification of anomalous events within the temporal universe (e.g. detecting optical flashes and supernovae at cosmological distances), and the real time classification of data in the presence of uncertainties and gaps within the data stream.

With surveys such as the LSST we will detect 250,000 SNe per year increasing the accuracy of measures of the energy content of the universe by an order of magnitude.

more...

3. A Probabilistic Framework for Scientific Inference

Current state-of-the art regression algorithms attempt to learn 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 regressionl 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 that model, 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] http://www.darkenergysurvey.org

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

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, it 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 true period of these variations from systematics introduced by observing systems. 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), detect damped Lyman alpha systems in the spectra of quasars (Garnett *et al.* 2012b), and 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 $y = f(\theta)$ 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, $K_{ij} \equiv \mathrm{Cov}\left[f(\vec{\theta}^i), f(\vec{\theta}^j)\right] = \exp(-\frac{1}{2}|\vec{\theta}^i - \vec{\theta}^j|^2/\ell^2)$ 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:

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

and variance:

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

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, and K_q is the vector of covariances between the query point and all training points. Eq. 2 extends directly to the case of multiple query points by making 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 Gaussian processes should consult Rasmussen and Williams (2006).

The inferences in equations 1 and 2 are non-parametric: they do not assume a form for the latent model they represent. This means that Gaussian Processes can be used for model regression even in the cases of noisy, sparse data. The returned inferences will not be constrained by any model assumptions, and the deleterious effects of noise will be propagated through Σ_q. We demonstrate this behavior on the following problem.

Imagine a scalar function f distributed over a 6-dimensional parameter space. We would like to learn a model for this function by sampling a sub-set of that space and using a Gaussian Process to infer the value of f at points we did not sample. Suppose, however, that we are unable to sample points from a given region of the parameter space. This situation arises in many sciences due to the limitation of experimental apparatus. It occurs in the case of astronomy, where cosmological Doppler shifting causes the identifying spectral features of galaxies to be shifted out of the observational range of most spectrographs. We simulate this scenario in Figure 1. Training data is taken from the parameter space according to a representative distribution of the truth, except that no training data is taken for values $1.4 \leq f \leq 2.5$. This data is then fed through both a Gaussian Process based algorithm (black curves) and an artificial neural network (red curves). The models learned by these algorithms are evaluated by considering the probability that they return the true value when sampled at points distributed throughout parameter space. The thin curves in Figure 1 show that Gaussian Processes perform worse that neural networks in the case where no training data exists between $1.4 \leq f \leq 2.5$. However, if we include training data in that region that is only 5% as densely populated as the truth, we see a significant improvement in the performance of Gaussian Processes (the thick curve). The Gaussian Processes realize that the training data is under-sampled and calibrate themselves accordingly. We will harness this behavior in designing our active learning algorithms. Figure 1 already shows (via the difference between the thin and thick curves) that the potential gain in performance from using active learning to select just a few new objects or follow-up observation is signitifcantly greater for a Gaussian Process than current state of the art classification and regression algorithms. We will further use the probabilistic nature of equations 1 and 2 to quantify the information contained in the Gaussian Process models and guide the active learning selection process itself. Thus we will optimize future observations constrained by limited follow-up budgets.

3.2. Optimizing Gaussian Processes

Despite the felicitous features already noted, we hope to improve the performance of our regression and classification algorithms by further optimizing the GPs themselves. One obvious difficulty in impelementing GP regression is the need to invert the matrix $(K + \sigma^2 I)$ in equations (1) and (2). This is straightforward in the case of small datasets, but represents a significant bottleneck in the case of big-data. Several solutions have been proposed. Kaufman *et al.* (2011) design a covariance function K with sparse support so that sparse matrix inversion methods may be used. Foster *et al.* (2009) advocate an approximate Gaussian Process method in which the matrix to be inverted is decomposed via a QR decomposition and only the most signficant columns of Q are used to compute the actual GP. In the work presented above and below, we adopt a solution wherein, for any given unknown $\{\vec{\theta}_q\}$, the GP is only given the k nearest neighbor data points as inputs, and k is a parameter fixed for all $\{\vec{\theta}_q\}$ and optimized by cross-validation. We have already shown that this method gives acceptable performance. Significant gains, however, may be made by allowing k to vary depending on the region of data space being probed. One can imagine an algorithm in which many values of k are tested for each $\{\vec{\theta}_q\}$ and the one which minimizing the entropy of the output probability distribution is finally chosen. This is another way in which we can use the non-parametric, probabilistic nature of GPs to optimize performance on our classification and regression problems.

There is also the question of choosing the functional form of K_{ij}. We have adopted a squared-exponential form in our preliminary work, but this is not the only possible form for K_{ij} and it does not have to be the most optimal choice. In the era of big-data, it should be possible to learn the form of K_{ij} directly from the data so that the covariances in our GP more truly reflect the covariances in the input data. Indeed, in the photometric redshift problem considered in Section 4 below, we set the characteristic length scale ℓ in K_{ij} dynamically so that different regions of data space imply different values of ℓ and the width of our posterior probability distributions is more responsive to the data. Further research focusing on the root functional form of K_{ij} as expressed in the data will likely yield additional improvements in performance.

4. Active learning: applications in cosmology

We will develop and test our basic active learning routines with an eye towards application to the astronomical problem of determining photometric redshifts.

The accurate determination of an object's distance or redshift is central to every test of cosmology that happens outside of a particle accelerator. The comparison of redshifts and luminosities of standard candles enabled the discovery of dark energy and cosmic acceleration (**?**). Redshifts serve as a proxy for radial distance from Earth to the observed object. Redshifts thus are necessary for building three dimensional maps of the distribution of galaxies in the Universe. Such maps will help and have helped us to constrain how galaxies formed over the history of the Universe, and thus can tell us much about how gravity operates at the largest scales and what the parameters are that govern the behavior of dark energy, dark matter, and the cosmic acceleration (Daniel and Linder 2010; de Putter *et al.* 2010; Das *et al.* 2011; Linder 2013). Accurately determining the redshifts of distant galaxies is a requirement if we are to answer some of the most vexing problems in fundamental physics today.

Direct spectroscopic redshift measurements of enough galaxies to constrain dark energy parameters to the precision required by next generation experiments would be thousands of times more expensive than taking the corresponding photometric (or imaging) data. Large digital cameras (e.g. the 3.2 Gigapixel camera for the LSST) can observe $\sim 10^6$ sources every 15 seconds (several orders of magnitude more efficient that spectroscopic observations). Our task, then, is to construct algorithms whereby we can convert these much cheaper photometric data into accurate redshifts (i.e. photometric redshifts).

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 template 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 spectra 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 2(a), which plots the results of running the publically available EAZY algorithm (Brammer *et al.* 2008) on a set of simulated galaxy observations designed to represent results expected from the Large Synoptic Survey Telescope (LSST[5]). While many of the galaxies fall near the $z_{\text{photometric}} = z_{\text{spectroscopic}}$ line, there is significant scatter in the results. We propose to overcome this difficulty with an exclusively data-driven algorithm based on Gaussian Processes.

---

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

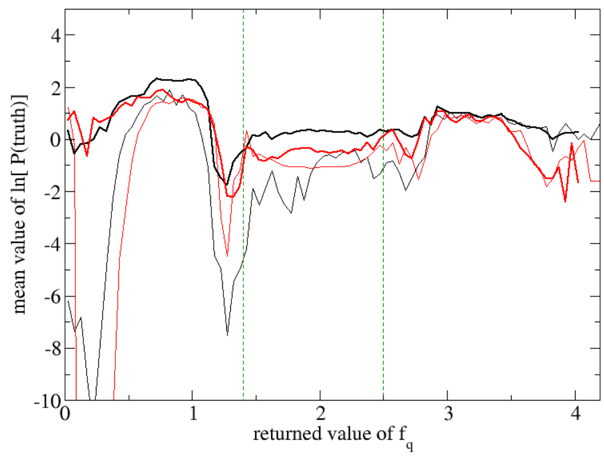
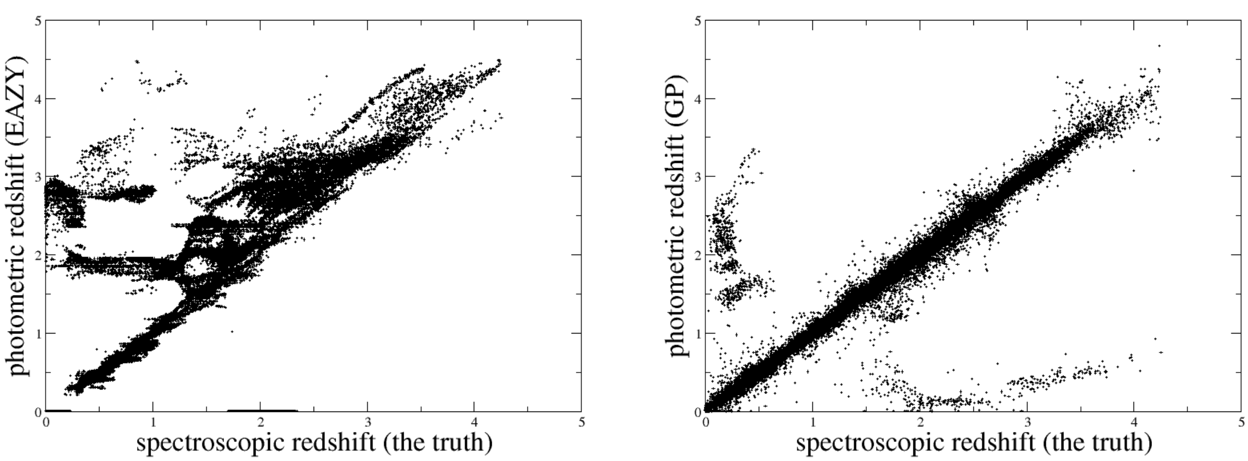


Fig. 1.— We compare the Gaussian Process algorithm outlined in Section 4 to an artificial neural network in the case of sparse training data. The red curves represent an artificial neural network scheme. The black curves represent the Gaussian Process algorithm presented in Section 4. The thin curves represent the case where all of the training data between the dashed lines has been excised. The thick curves represent the case where 5% of that training data has been restored.



(a)

(b)

Fig. 2.— Photometric redshift plotted against true spectroscopic redshift for 48,000 simulated LSST galaxy observations. Photometric redshifts are derived using the EAZY template-fitting algorithm in Figure 2(a) and our Gaussian Process based algorithm in Figure 2(b).

Other works have already attempted to apply Gaussian Processes 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 Gaussian Processes to learn the full probability distribution that a given galaxy is at a given redshift. This will produce an algorithm that is simultaneously more robust against sparse, noisy or degenerate training data and more amenable to improvement by the introduction of active learning. Figure 2(b) shows preliminary results from our algorithm when trained on spectroscopic data from 50,000 galaxies and tested on the same 48,000 galaxies as Figure 2(a).

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 leaves its results sensitive to degeneracies in the photometric data whereby low redshift galaxies look similar to high redshift galaxies, confusing the algorithm. Figure 3 plots the mean value of $\ln[P(\text{truth})]$, i.e. the value of the logarithm $P(z_{\text{photometric}})$ at the point $z_{\text{photometric}} = z_{\text{spectroscopic}}$ as a function of photometric redshift for EAZY, ANNz, and our Gaussian Process algorithm. We see that both EAZY and ANNz consistently assign lower probabilities to $z_{\text{photometric}} = z_{\text{spectroscopic}}$ than does our Gaussian Process algorithm.

4.1. Experimental Design and Optimization through Active Learning

The demands of next generation cosmological experiments will require that our photometric redshift determinations be accurate to within $\leq 2\times10^{-3}(1+z)$ (LSST Dark Energy Science Collaboration 2012). This is a hard limit, as a bias in redshift determination of just 0.01 can degrade dark energy constraints by as much as 50% (Kitching *et al.* 2008; Huterer *et al.* 2006; Nakajima *et al.* 2012). Testing present template and empirical methods on a sample of 5,482 galaxies from the 2df-SDSS LRG and Quasar survey, Abdalla *et al.* (2011) find biases of order 0.05 (see their Figure 4). This level of bias can degrade dark energy constraints by as much as a factor of 3 (Ma *et al.* 2006). Considering 3,000 galaxies from the DEEP2 EGS and zCOSMOS surveys and using Bayesian methods, Mandelbaum *et al.* (2008) find a bias in redshift determination of order 0.01 (see their Table 2). While this is an improvement, it still an order of magnitude larger than what is required.

How can we resolve these problems? In both the empirical and template photometric redshift codes, biases arise from the fact that the training samples (templates) do not occupy the same color and redshift space as the data. The solution, therefore, lies in correctly the choosing the most useful galaxies for follow-up spectroscopy. Simple sampling strategies (e.g. random or stratefied) are not efficient. We need a technique to identify the next best observation to take to best reduce the redshift estimation bias of our algorithm. Active learning will provide this technique.

A classic active learning method, called uncertainty sampling, uses the uncertainty of each test point as the criterion for choosing the next experiment (e.g. the next spectroscopic measurement of a source in the training sample). Teh algorithm considers the modeled galaxies and chooses the most uncertain for follow-up Figure 4 demonstrates a variant of this scheme assigned to the problem presentedin Figure 1. From a total of 97,000 data points, we start with 20,000 training points and assess the efficacy of our Gaussian Process classifier by considering the mean value of $\ln[P(\text{truth})]$ as in Figure 3. The training set is then grown by selecting new points either randomly (black curve) or according to the maximum value of $(-\ln[P(\text{mode})])$ (red curve). As you can see, assembling the training set with active learning leads to a significant improvement in the classifier's performance.

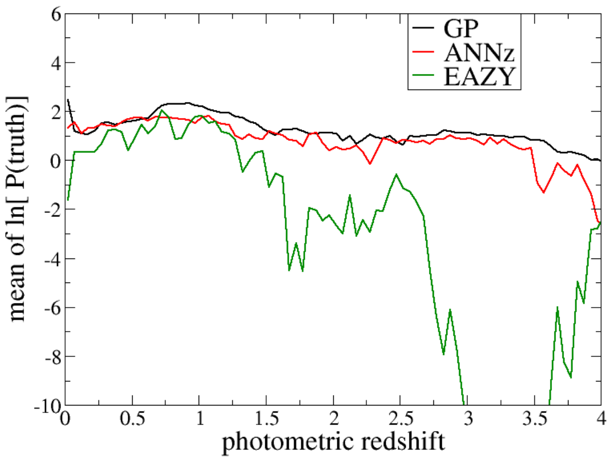
We will expand upon these approaches using our recent work on the problem of optimal surveying or polling (Garnett et al 2012a). Rather than having a goal of correctly predicting the output for each point in a test set, the goal is to predict the average output (or the class proportions in classification problems) over the test set. This dramatically increases the efficiency of the active learning. In preliminary experiments on graphs and other domains, minimizing this survey variance not only performs well on the surveying problem, but also outperforms the trace criterion and other popular active learning methods such as uncertainty and density sampling on active learning problems. This result is consistent with the findings of Richards *et al.* (2012b), who consider a similar problem for a Random Forest classifier on the static data set produced by the All Sky Automated Survey. Intuitively, it seems reasonable that considering the entire covariance matrix of the modeled data might lead to better performance than choosing only based on its diagonal. We have, however, little theoretical understanding of why this is better than the trace criterion which directly optimizes the quantity on which we will ultimately measure performance. We will seek a better theoretical understanding of this phenomenon as part of this work.

5. Active Feature Acquisition: Anomaly Detection and Classification in Massive Data Streams

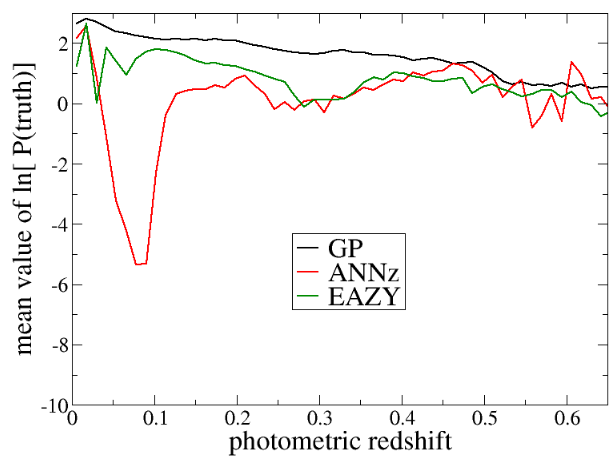
While the problem of photometric redshift determination is fertile ground for the development of active learning algorithms, it poses little challenge for active feature acquisition: for a given experiment, the photometric filters are often immutable. Fortunately, survey astronomy presents us with another 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?"

The next generation of astrophysical surveys will visit the same region of sky many thousands of times. This opening of the temporal domain in astrophysics offers the potential to discover new classes of physical phenomena while coming with many associated computational challenges. Variability within the universe is believed to be present on time scales of seconds through to tens of years. The shortest time scales correspond to the explosion of the most massive stars within the universe which produce short but intense optical and gamma-ray flashes. These outbursts provide direct tests of General Relativity and of high energy physical processes (at energies far beyond those accessible on the Earth). For example the rate at which these events occur constrains the age at which the first stars within the universe came into being. Intermediate timescale variability comes in the form of supernovae (SNe) which detonate, brighten and then dim. These exploding stars are known to have a narrow range of intrinsic brightnesses; they act as standard candles that can be used to determine the rate at which the universe expands and thereby measure its mass and energy content (**?**). With surveys such as the LSST we will detect 250,000 SNe per year increasing the accuracy of measures of the energy content of the universe by an order of magnitude.

With timescales as short as seconds to hours we need to be able 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.



(a)



(b)

Fig. 3.— Figure 3(a) shows the mean value of $\ln[P(\mathrm{truth})]$ as a function of photometric redshift (the vertical axes in Figures 2) for all three algorithms under consideration using simulated LSST data. Figure 3(b) considers 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.

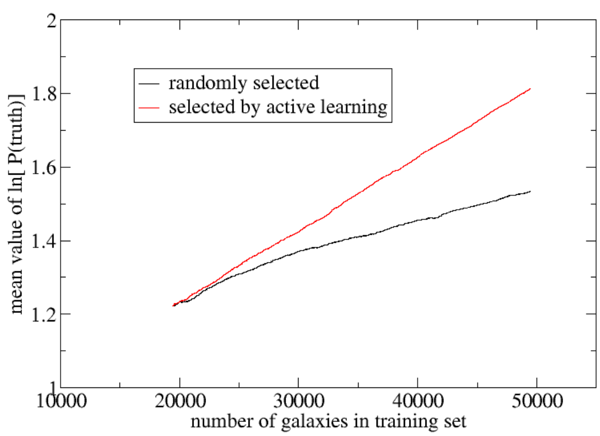


Fig. 4.— Active learning applied to classification according to a scalar function on a 6-dimensional data space. The horizontal axis is the size of the training data set. The vertical axis is the mean value of the output probability distribution at the true value of the scalar function. The black curve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

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.

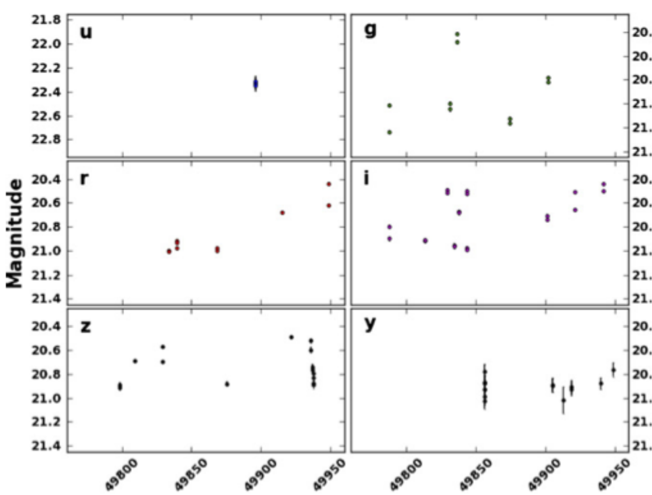
5.1. Active Learning for Transient Classification

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

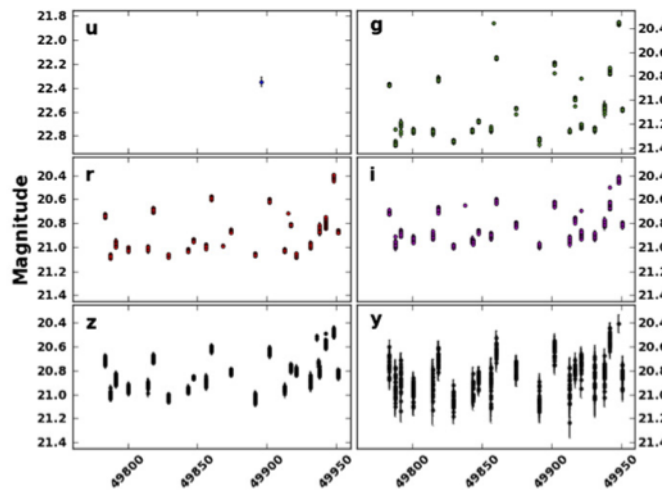
Real-time automatic classification of objects is already widely acknowledged as a necessary support technology for the forthcoming age of survey astronomy (Djorgovski *et al.* 2011; Richards *et al.* 2011; Richards *et al.* 2012a; Graham *et al.* 2012; Mahabal *et al.* 2008a; Mahabal *et al.* 2011a).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

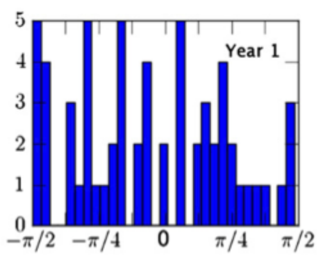
Similarly, a large number of classifiers have been developed which are optimized for the case of binary classifications: "Is something a quasar, or is it not?" (Kim *et al.* 2011; Pichara *et al.* 2012),



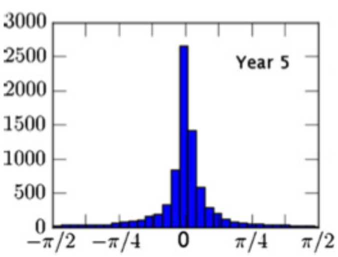
(a)



(b)



(c)



(d)

Fig. 5.— Taken from (Oluseyi *et al.* 2012). Figure 5(a) shows the sampling of a template RR Lyrae light curve at the LSST Universal Cadence. Figure 5(b) shows the same for the Deep Drilling Cadence. Figures 5(c) and 5(d) compare the scatter in determining the phase of an RR Lyrae light curve after 1 year of Universal Cadence data and 5 years of Universal Cadence data. This illustrates the significant effect on scientific output of more and more targeted data.

"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. 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 2 which will allow us to perform a similar determination in real time as observations are made, directing experiments as they are performed. 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 to determine a light curve's true period. The method should be successful when extended to the case of multiple observed and latent quantities.

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. Timetable of Activities

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

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.