

Learning in an Era of Uncertainty

Principal Investigator: Andrew Connolly
Applicant/institution: University of Washington
Telephone number: (206) 543 9541
Email: ajc@astro.washington.edu
Funding request: \$170K (year one), \$170K (year two), \$170K (year three)
DOE/OSP Office: Office of Advanced Scientific Computing Research
DOE/Technical Contact: Dr. Alexandra Landsberg

CoPrincipal Investigator: Jeff Schneider
Applicant/institution: Carnegie Mellon University
Telephone number: (412) 268 2339
Email: schneide@cs.cmu.edu
Funding request: \$170K (year one), \$170K (year two), \$170K (year three)

Total Funding request: \$340K (year one), \$340K (year two), \$340K (year three)

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 and biology, will come on-line over the next decade. These experiments share many similar challenges in the field of statistics and machine-learning: how can 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 on 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 training data that could be collected, 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 system (according to some objective criterion).

While the algorithms and methodologies we propose will impact many of the data intensive sciences, we will focus our work in the context of DOE sponsored cosmology experiments (i.e. the Dark Energy Survey, and the Large Synoptic Survey Telescope). 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 impacting the understanding of cosmology, particle physics, and potentially theories of gravity. Our ability to achieve these scientific goals relies on analyses at a scale, speed, and complexity beyond the capabilities of current automated machine learning methods.

2. Active learning for calibrating 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, DES¹ and the Large Synoptic Sky Survey, LSST²). 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. On these scales, statistical noise will no longer determine the accuracy to which we can measure cosmological

¹<http://www.darkenergysurvey.org>

²<http://www.lsst.org>

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.

3. Inference in the Presence of Noise and Gaps in a Data Stream

3.1. Anomaly Detection and Classification in Massive Data Streams

The next generation of astrophysical surveys we will visit the same region of sky many thousands of time. 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 followup 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 7.5×10^8 sources **every night**. Even for the most numerous transient events (SNe) this corresponds to less than 10^{-5} of the total number of sources identified being transient. For the most energetic bursters the magnitude of the challenge is 500-fold larger. Algorithms for identifying anomalies and variability must, therefore, be robust to false positives and missing data and must account for the cadence in how we sample the time domain, variations in the quality of the data due to atmospheric conditions, changes in the performance of the telescope and camera and the possibility that we observe sources at different wavelengths at different times.

4. Experimental Design and Optimization through Active Learning

Experimental design description

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.* 2012; 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. 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 possible approaches are available for the automated classification of transient objects, and not all of them rely upon highly detailed observations to function.

The input features for transient object classification will be photometric and morphologic measures taken at different increments in time. The output is a categorical variable indicating the class. This is in contrast to photometric redshifts where the output is continuous. Fortunately, the underlying covariance for Gaussian process classification is the same as for regression and we will adopt a similar prediction model.

The active learning problem for transient objects contains three subproblems, active learning, active search, and active feature acquisition. Finally, one additional difference between the transient object method and the photometric redshift method is that transient object decisions must be made in an online, streaming fashion. Rather than considering an entire pool of test objects, they appear

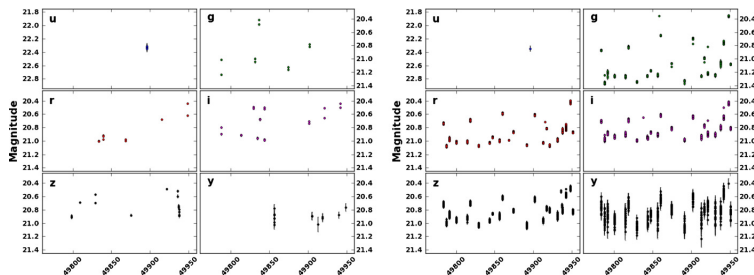


Fig. 1.— Two hundred days of simulated LSST of an RR Lyrae star template, taken from (Oluseyi *et al.* 2012). The left panel shows sampling at the universal cadence. The right panel shows sampling at the deep drilling cadence.

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. We first describe algorithms for each of the three pieces and then how to combine them into a single algorithm for streaming transient detections.

Active Learning. In order to get a ground-truth class for a transient object, repeated observations are taken to estimate a full light curve. A human expert then assigns a class label. Both the repeated telescope observations and human time are expensive and thus we want to learn a good classification model using limited training data. We propose a corresponding trace and survey criteria for active learning on transients as in Section ??.

Active Feature Acquisition. When considering photometric redshifts, we assumed that all photometric inputs were observed for every object of interest. For transients this will not be the case. A large component of the problem is deciding for each object, whether observations of additional colors and/or additional times would be valuable in classifying or characterizing the object.

We propose to extend our recent work on using 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. For transient objects we will have 5 color magnitudes that are a function of time and are coupled to each other. 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. The key advantage of this approach is that the GPs naturally provide a mean and covariance for future unobserved colors. 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 certain color 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. Many transient objects will be from common and/or well-understood classes that do not have much observational value. The ultimate objective in following up detected transients is to maximize the number of interesting transients classified and characterized 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. We propose to use the simple myopic algorithm described in that work. It computes the probability of each point belonging to the positive class and chooses the largest.

A combined streaming method. The three goals above will be combined in a staged set of decisions. When a new transient is detected, the light curve model will be used to provide estimated light curves for the object. Those light curves 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.

The active search and active learning algorithms each compute an objective criterion score for each object. For photometric redshifts the scores are used to create a ranked list and observations are scheduled proceeding down the list. When a budget is given for follow-ups on each batch of newly detected transients, going down a ranked list in each batch until that budget is exhausted is appropriate.

However, when the budget is an aggregate over a longer time period a streaming decision on how much of the budget to spend on each batch must be made on that batch in isolation. This will be done by choosing a score threshold. The threshold will be set by evaluating the historical stream and setting it at a value that would yield a number of follow ups equal to an available budget for following up. The threshold will be adjusted continuously as more observations are taken and the models and scientific goals change (e.g. the object types designated as “interesting” are changed).

Since all of the algorithm components are based on GP models, it is possible to merge them together into a single model. We hypothesize that additional performance improvements can be made through an integrated model and decision algorithm. For example, an object with a modest active learning score that can be easily characterized with only one more follow up might be promoted over one with a higher score that can not be easily characterized even with many more observations. After we have implemented the staged method described above, we will investigate and compare an integrated model.

4.1. Active learning for calibrating cosmology

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 and the cosmic acceleration (Daniel and Linder 2010; de Putter *et al.* 2010; Das *et al.* 2011). 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 than 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).

Presently, 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 basis spectra. 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 (LSST³). 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.

We treat the photometric redshift problem probabilistically. Rather than taking the spectroscopic data and learning a one-to-one photometric redshift relationship, we will use Gaussian Processes to model a probability density function in $P(z_{\text{photometric}})$ which can characterize the probability that a given galaxy is at a given redshift. This method will not only be resilient against the pitfalls demonstrated in Figure 2(a), but will allow us to quantify the information content of our trained data in such a way as to optimize the gathering of additional spectroscopic data to further improve on our results. Figure 2(b) shows preliminary results from our algorithm when trained on spectroscopic data from an additional 50,000 galaxies. We discuss specifically how we model $P(z_{\text{photometric}})$ below.

4.2. Gaussian Processes

Gaussian processes (GPs) model the output of an unknown, noisy, sparsely-measured function as though it was generated by a random process with an underlying Gaussian probability distribution (Rasmussen and Williams 2006). An important feature of GPs is that they do not make parametric assumptions about the form of the function they are modeling. They have received attention as a means of describing physical phenomena (e.g. the expansion history of the universe) without having to assume a model of the underlying process (Shafieloo *et al.* 2012), of interpolating point spread functions across large images (Bergé *et al.* 2012), and of accelerating the search of high-dimensional likelihood functions on the cosmological parameters by efficiently selecting sample

³<http://www.lsst.org>

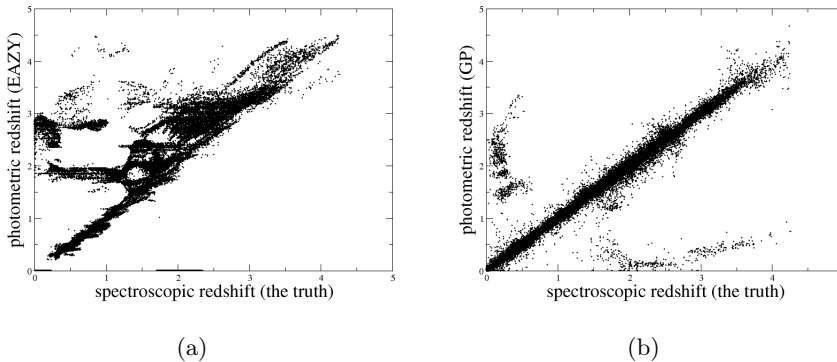


Fig. 2.— Photometric redshift plotted against true spectroscopic redshift for 48,000 simulated LSST galaxy observations. Photometric redshifts are derived using the EAZY template-fitting algorithm in Figure 2(a) and our Gaussian Process based algorithm in Figure 2(b).

points (Daniel *et al.* 2012). Other works have already attempted to apply Gaussian Processes to the problem of photometric redshifts (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. We describe our method below.

Assume that each galaxy training set datum is of the form $\{\vec{\theta}, y\}$, where $\vec{\theta}$ is an N_p -dimensional vector representing the magnitude of the galaxy in each of the survey’s filters (the galaxy’s position in photometric color space) and $y = f(\theta)$ is the physical quantity (redshift) we are trying to infer. Gaussian processes assume that f is a probabilistic function on the N_p -dimensional space with some covariance function such as a squared exponential covariance, $K_{ij} \equiv \text{Cov}[f(\vec{\theta}^i), f(\vec{\theta}^j)] = \exp(-\frac{1}{2}|\vec{\theta}^i - \vec{\theta}^j|^2)$. Under those assumptions, one can derive a posterior probability distribution for the redshift at a new query point $\{\vec{\theta}^q\}$ by marginalizing over the measured galaxies. This gives a Gaussian distribution with mean

$$f(\vec{\theta}^q) = \sum_{i=1}^n \alpha_i k(\vec{\theta}^i, \vec{\theta}^q) \quad (1)$$

and variance

$$\Sigma^q = \text{Cov}(f^q) = K_{qq} - K_q^T (K + \sigma^2 I)^{-1} K_q \quad (2)$$

Here, $\vec{\alpha} = (K + \sigma^2 I)^{-1} \vec{y}$, 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} represents the training data redshifts, and where 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 matrices as 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).

We apply this method to the case of photometric redshifts as follows. For each test galaxy $\{\vec{\theta}_q\}$, we find the k nearest neighbor training galaxies in flux space (k is treated as a parameter to be optimized by our algorithm). We divide these neighbor galaxies into two sub-populations based on their spectroscopic redshifts and fit a Gaussian Process to each sub-population. This gives us a bi-modal probability distribution for $P(z_{\text{photometric}})$. We compare the likelihood of this hypothesis to a single-mode $P(z_{\text{photometric}})$ in which all of the neighbor galaxies are fit together in one Gaussian Process. The final $P(z_{\text{photometric}})$ is a linear combination of these two hypotheses, weighted according to their respective model likelihoods.

Figure 3 shows preliminary results of our proposed probabilistic Gaussian Process algorithm. Figure 2(b) plots photometric redshift versus spectroscopic redshift for the same simulated galaxy observations used in Figure 2. According to this metric, Gaussian Processes perform better than the EAZY algorithm and comparably to the neural network of ANNz. However, if we instead measure performancy by considering $\ln[P(\text{truth})]$, i.e. the logarithm of the value of the posterior probability density function at the known spectroscopic redshift of our galaxies, we see that Gaussian Processes consistently outperform neural networks (see Figure 3(a)). Figure 3(b) recreates Figure 3(a) using actual data taken from the Sloan Digital Sky Survey (Abazajian *et al.* 2009).

Other data-driven algorithms for photometric redshift determination do exist. The most popular 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 present algorithms are designed only to return a photometric redshift value and an uncertainty. This leaves their results sensitive to degeneracies in the photometric data whereby low redshift galaxies look similar to high redshift galaxies, confusing the algorithm. We see this in the lower mean values of $\ln[P(\text{truth})]$ returned by ANNz in Figure 3.

The demands of next generation cosmological experiments will require that our photometric redshift determinations be accurate to within $\leq 2 \times 10^{-3}(1+z)$ (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. Improving on this through targeted observations is, however, expensive. Simple sampling strategies (e.g. random or stratified) 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.

Focus on these as our primary test to develop these algorithms

4.3. Active learning

In the field of machine learning, *active learning* algorithms iteratively decide which data points they will collect outputs on and add to a training set. The 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 training data that could be collected, and the current learned model, and evaluate each potential new point according to some objective criterion. They are particularly valuable when the data in question are expensive to acquire (in time or resources). Many examples of such problems exist in physics and cosmology: from defining training sets for estimating the distances to galaxies based on their photometric data (Connolly et al 1995), to choosing which observation will most improve a cosmological signal, to picking the next anomaly to follow-up, or the next piece of data to obtain about an anomaly that would enable its accurate classification.

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). 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. Intuitively, it seems reasonable that considering the entire covariance matrix 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.

Information gain and astronomy (Jake's stuff)

REFERENCES

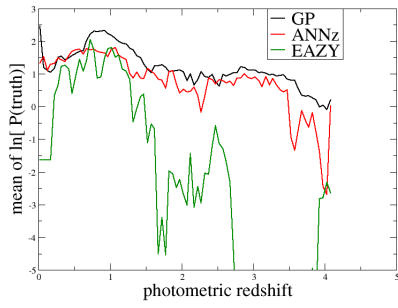
- Abazajian, K. N. *et al.* [SDSS Collaboration] 2009, The Astrophysical Journal Supplement Series **182**, 543 [arXiv:0812.0649 [astro-ph]].
- Abdalla, F. B., Banerji, M., Lahav, O., and Rashkov, V. 2011, Monthly Notices of the Royal Astronomical Society **417**, 1891
- Abramo, L. R., Strauss, M. A., Lima, M., Hernández-Monteagudo, C., Lazkoz, R., Moles, M., de Oliveira, C. M., Sendra, I., Sodr   Jr., L., and Storch-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
- Berg  , J., Price, S., Amara, A., and Rhodes, J. 2012, Monthly Notices of the Royal Astronomical Society **419**, 2356
- 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
- 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
- Budav  ri, 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., Ivezić, 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) Conference 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., Mörtzell, 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
- Garnett, R., Krishnamurthy, 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*
- 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 Statistics* **5** 2470
- Kitching, T. D., Taylor, A. N., and Heavens, A. F. 2008, *Monthly Notices of the Royal Astronomical Society* **389** 173
- 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

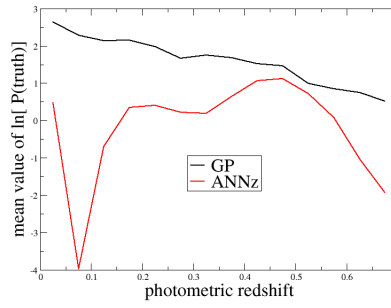
- Mahabal, A., Djorgovski, S. G., Williams, R., Drake, A., Donalek, C., Graham, M., Moghaddam, B., Turmon, M., Jewell, J., Khosla, A., and Hensley, B. 2008b [arXiv:0810.4527] to appear in proceedings for the Class2008 conference (Classification and Discovery in Large Astronomical Surveys, Ringberg Castle, 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., Kambezzyk, 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., Scodreggio, 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]
- 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., Ivezic, Zeljko, Jones, R. Lynne, Krughoff, K. Simon, and Sesar, Branimir (2012), The Astronomical Journal **144** 9
- 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
- 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. 2012 *The Astrophysical Journal* **744**, 192
- 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., Ivezić, 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., Soccimarro, 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., Stoughton, 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
- 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



(a)



(b)

Fig. 3.— Figure 3(a) shows the mean value of $\ln[P(\text{truth})]$ as a function of photometric redshift (the vertical axes in Figures 2) for all three algorithms under consideration. Figure 3(b) compares our Gaussian Process method to the artificial neural network code ANNz (Collister and Lahav 2004) on real data taken from the Sloan Digital Sky Survey. In this latter case, the algorithms are trained on 70,000 galaxies and tested on 715,000 galaxies.