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. This 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, DES[1] and the Large Synoptic Sky Survey, LSST[2]). 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 7.5x10^8 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.

RR-Lyrae - figure showing sampling

4. Experimental Design and Optimization through Active Learning

Experimental design description

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 **?**. 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).

The demands of next generation cosmological experiments will require that our photometric redshift determinations be accurate to within $\leq 2 \times 10^{-3}(1+z)$ **?**. 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% **???**. 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 **?**. 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 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.

Focus on these as our primary test to develop these algorithms

4.2. 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 evalute 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)