

1 INTRODUCTION

This three-year project aims to deliver

2 PLAN

What will we do...

2.1 ASTR 490

Was it 490? Andy remembers...

Copied text from that email, which was copied from our book, to see how many pages it would take...

(a) Motivation for Astr 490

Astronomy and astrophysics are witnessing dramatic increases in data volume as detectors, telescopes, and computers become ever more powerful. During the last decade, sky surveys across the electromagnetic spectrum have collected hundreds of terabytes of astronomical data for hundreds of millions of sources. Over the next decade, the data volume will enter the petabyte domain, and provide accurate measurements for billions of sources. Astronomy and physics students are not traditionally trained to handle such voluminous and complex data sets. Furthermore, standard analysis methods employed in astronomy often lag far behind rapid progress in statistics and computer science. The main goal of this course is to contribute to efficient training of next generations of students to handle the fast growing data sets, not only in astronomy, but in other quantitative sciences as well.

This course will be aimed at physical and data-centric math, statistics, science and engineering students who have an understanding of the science drivers for analyzing large data sets but may not be aware of appropriate statistical techniques for doing so. The course work will provide to students a connection between scientific data analysis problems and modern statistical methods. We will limit theoretical discussions to the minimum required to understand the algorithms and will build the courses upon an example-driven compendium of modern statistical and data mining methods, together with carefully chosen examples based on real modern data sets, and of current astronomical applications that will illustrate each method introduced in the book. Discussion of the advanced material will be supported by appropriate (publicly

available) Python code and data which will enable students to perform exercises, evaluate the techniques, and adapt them to their own fields of interest. We chose to use Python, a powerful and flexible programming language that is quickly becoming a standard in data-intensive sciences (and elsewhere).

The target audience for our course includes undergraduate students with scientific or engineering background, but it is likely that graduate students would benefit from it too. Familiarity with calculus and other basic mathematical techniques will be assumed, but no extensive prior knowledge in statistics will be required.

The course outline:

1. Computational Challenges in data-intensive astronomy and astrophysics
 - data types and data management systems
 - analysis of algorithmic efficiency
 - types of computational problems and strategies for speeding them up
 - data visualization challenges
 - selection effects and truncated/censored data in astronomical context
2. Searching for structure in astronomical point data
 - non-parametric density estimation
 - nearest-neighbor density estimation
 - parametric density estimation
 - finding clusters in data
 - correlation functions
3. Dimensionality reduction
 - principal component analysis in astronomical context
 - non-negative matrix factorization
 - independent component analysis and projection pursuit
 - manifold learning
4. Regression and model fitting

- regression for linear models
- non-linear regression
- kernel and principal component regression
- methods for handling heteroscedastic and non-Gaussian errors
- Gaussian processes
- overfitting, underfitting and cross-validation

5. Classification

- generative classification methods
- discriminative classification method
- evaluation and comparison of classifiers: ROCcurves

6. Time series analysis in astronomy

- main concepts and tools for time series analysis
- analysis of periodic time series
- temporally localized signals
- analysis of stochastic processes

We will textbook *“Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data”* (Princeton Series in Modern Observational Astronomy) coauthored by Co-PIs on this prooposal.

2.2 Python Packages

We will leverage all the publicly available modern python tools. In particular, seminar work will be built around the *astroML* package that was developed to support textbook to be used with the proposed course.

2.3 Workshop

At the last meeting, we assumed a 3-day workshop for about 30 faculty from other institutions of higher learning who would want to emulate our program. Further assuming \$20/day/person for lunch and coffee, and \$75/person for conference dinner, \$1,500 for the workshop venue, and four grants of \$500 to young faculty and postdocs, we need about \$7,500 for the workshop.


2.4 Budget

We assumed a 3-year long project.

We assumed 2.5 months of summer salary for Marina, and a postdoc. About \$150,000/year.

We assumed 1 month of summer salary for both ŽI and Andy to demonstrate seriousness, and a graduate student. About \$100,000/year.

All together, about \$750,000 for 3 years. If this is above the limit, perhaps we could descope faculty to only the first two years?



[Home](#)
[User Guide](#)
[Book Figures](#)
[Examples Plots](#)

News

October 2012: astroML 0.1 has been released! Get the source on [Github](#)

Our Introduction to astroML paper received the CIDU 2012 best paper award.

Links

[astroML Mailing List](#)

[GitHub Issue Tracker](#)

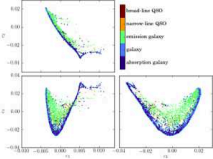
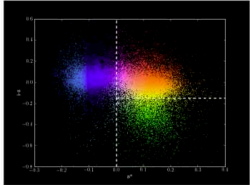
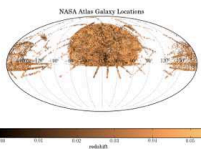
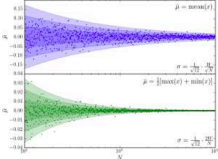
Videos

[Scipy 2012 \(15 minute talk\)](#)

Citing

If you use the software, please consider citing [astroML](#).

AstroML: Machine Learning and Data Mining for Astronomy







AstroML is a Python module for machine learning and data mining built on `numpy`, `scipy`, `scikit-learn`, and `matplotlib`, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. The astroML project was started in 2012 to accompany the book **Statistics, Data Mining, and Machine Learning in Astronomy** by Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alex Gray, to be published in late 2013. The table of contents is available here: [here \(pdf\)](#).

Downloads

- Released Versions: [Python Package Index](#)
- Bleeding-edge Source: [github](#)



User Guide

1. Introduction

1.1. Philosophy

Figure 1: We will leverage all the modern python tools available in *astroML* and other packages.

2.5 Teaching Program

What will we do...

2.6 Leveraging Current NSF Funding

Co-PI Vanderplas is currently supported by a 3-year NSF postdoctoral fellowship through the interdisciplinary CI-TraCS program. Though Vanderplas' background is in Astronomy, the sponsoring professor is in the Computer Science and Engineering department. The focus of the fellowship is research on the computational side of Astronomy, especially on efficient statistical analysis of very large datasets.

A full 20% of the fellowship time is devoted to teaching and course preparation, and as part of this requirement Vanderplas has developed and taught a Fall 2013 graduate seminar course through the Astronomy department: Astr 599, *Scientific Computing with Python*. The purpose of the course is to offer a comprehensive introduction to scientific computing in the Python programming language, geared toward graduate and advanced undergraduate students in Astronomy. After stepping through the fundamental tools of scientific computing, the course scratches the surface of statistical, machine learning, and datamining methods made available through various packages in the scientific Python ecosystem. The entirety of the curriculum material is made available on the course website¹. In the remaining two years of his fellowship, Vanderplas will expand this curriculum and offer the course to a wider audience of students through the University's inter-disciplinary eScience Institute.

This curriculum is in many ways a fundamental component of the goals of the current proposal. Practical statistical analysis and data mining requires a certain level of proficiency in a scientific computing platform: this course equips students with that foundational knowledge from which they can explore the use of data mining and machine learning algorithms within their own field.

As the current proposal moves forward, we will...

¹ <http://www.astro.washington.edu/vanderplas/Astr599/>

2.7 E-science

What will we do...

3 PROJECT ORGANIZATION

3.1 Key Project Aims

This project will **i)** develop something, **ii)** apply these methods, and **iii)** synthesize and compare...

These deliverables will have an impact on the community that is much broader than the focus of this proposal.

3.2 Responsibilities and Schedule

The PI, Meila, will be responsible for the overall success of the project.

The co-PIs, Connolly, Ivezić, and Vanderplas will be grossly irresponsible.

In summary, the project schedule is:

Year 1: Think

Year 2: Do

Year 3: Analyze

3.3 Results from Prior NSF Support

The Co-PI Ivezić was recently PI on four projects supported by NSF that are indirectly related to the work proposed here (mostly through data mining aspects, public release practice for all data products, and through engaging large numbers of students in research and publication process).

The projects “Towards a Panoramic 7-D Map of the Milky Way” (AST-070790) and “Mapping the Milky Way: Data-miners, Modelers, Observers, Unite!” (AST-1008784) quantified statistical behavior of a few tens of millions of Milky Way stars observed by the Sloan Digital Sky Survey in multi-dimensional position–velocity–chemical composition space. The results were published in over a dozen refereed papers, and the work engaged four graduate students (including two Ph.D. theses) and 11 undergraduate students. A team of three undergraduate students has developed an education and public outreach site.

The key project aims for the NSF award AST-0507529 “Interpretation of Modern Radio Surveys: Test of the Unification Paradigm” were unification of several modern radio catalogs into a single public database containing several million sources and morphological classification of the matched sources. This three-year long project

has produced six journal publications, two Ph.D. theses, and has engaged six undergraduate students in data analysis and publications.

The project “Statistical Description and Modeling of the Variability of Optical Continuum Emission from Quasars” (AST-0807500) used time-domain data for the exploration of quasar physics. This three-year project has produced four journal publications, a Ph.D. thesis, and has engaged four undergraduates in data analysis and publications.