

1 INTRODUCTION

This three-year project is an enthusiastic response to the EXTREEMS-QED (Expeditions in Training, Research, and Education for Mathematics and Statistics Through Quantitative Explorations of Data) solicitation for proposals that aim to train the next generation of statistics and mathematics undergraduate students for confronting new challenges in computational and data-enabled science and engineering (CDS&E). The proposed work addresses and includes all the main required project components: Education and Training, Research, and Faculty Professional Development. The proposed program will provide opportunities for undergraduate research and hands-on experiences centered on CDS&E and will result in significant changes to the undergraduate mathematics and statistics curriculum at the University of Washington. We have secured broad institutional support and buy-in from two major departments (Statistics and Astronomy), and the proposed work includes a workshop centered on professional development activities for faculty from other institutions wishing to emulate the proposed program.

The Education and Training component will be centered around the Applied and Computational Math Sciences (ACMS) Program at the University of Washington. We will significantly enhance this program in the context of computational and data-enabled science and engineering by including two new specialized courses in the ACMS statistics track. The first course, STAT 391 “Computational Statistical Modeling and Machine Learning”, will include essential statistical methodology, such as regression and probability models for discrete and continuous data, as well as topics crucial for data-enabled science, such as classification and clustering. The second course, ASTR 497 “Data Intensive Astronomy and Astrophysics”, will apply methods introduced in STAT 391 to contemporary massive datasets collected by modern astronomical sky surveys, and further expand them with domain-specific methodologies. These courses are designed (1) to give students a hands-on experience with statistical modeling through programming

and performing real data analyses on a computer, and (2) to introduce students to the machine learning methodology in particular, with specific attention to the issues of big data, with astronomy as an attractive core science example.

The Research component will build upon existing close collaboration between Statistics and Astronomy departments at the University of Washington, led by the PIs of this proposal. In addition to three faculty and an NSF postdoctoral Fellow, the proposed program will also include two graduate students and a large number of undergraduate students. We will utilize a suite of statistical and machine learning methods to attack a number of unsolved challenges in data-intensive astronomy posed by recent data avalanches coming from modern astronomical sky surveys.

The proposed program, including course work and supporting research efforts, will represent a paradigm-shifting model that may be easily adapted by other institutions. To facilitate such adoption, the Faculty Professional Development component will utilize several communication and dissemination techniques, culminating with a workshop for faculty from other institutions wishing to emulate the proposed program. The workshop, and a supporting website, will disseminate all the teaching materials (including datasets and code to perform hands-on research exercises) and the results of program effectiveness evaluation.

In the remainder of this proposal, we describe the new courses in detail, how they fit within the ACMS program and the overall “Big Data” efforts within the University of Washington, and the budget and execution schedule for the proposed program.

This is from an old email summarizing this solicitation; some statements that we should disperse through the text at some point:

1) course/class work

- we will contribute to education of the next generation of mathematics and statistics undergraduate students to confront new challenges in computational and data-enabled science and engineering (CDS&E)

- we will also include math and stat minors
 - our efforts will result in significant changes to the undergraduate curriculum
 - student training will incorporate computational tools for analysis of large data sets and for modeling and simulation of complex systems
 - we will incorporate CDS&E content in existing courses and develop new courses in CDS&E areas
 - we will create resources for scientific education, including cyber-enabled pedagogies (eBooks, online resources, etc.).
 - we will foster interdisciplinary collaborations aiming to transform both departmental and institutional culture.
 - we have broad institutional support and department-wide commitment that encourage collaborations within and across disciplines
- 2) research work
- research work will be broadly defined, long-term, team-based, interdisciplinary, and will include with other institutions
 - we will develop tools and theory for analyzing massive data sets
 - we will use cyberinfrastructure to model and visualize complex scientific and engineering concepts;
 - we will create resources for scientific investigation, including state-of-the-art tools and theory for knowledge discovery from massive, complex, and dynamic data sets
 - we will foster interdisciplinary collaborations
 - we will promote undergraduate research and hands-on experiences centered on CDS&E
 - the hands-on research work will develop CI competences (programming, data management, simulation-building)
 - we will leverage and advance the use of cyberinfrastructure resources (e.g. data archives, networks, advanced computing systems, visualization environments) for data exploration
 - we will address data-intensive scientific problems (arising in astronomy and ...)
- 3) workshop
- professional development activities centered on CDS&E for faculty or K-12 teachers
 - we will foster interdisciplinary collaborations
 - we will create new learning environments and experiences that immerse students in CDS&E while energizing and sustaining the professional growth of faculty in CDS&E

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](#)

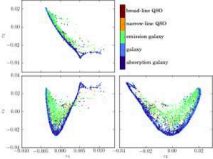
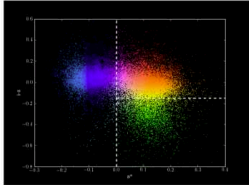
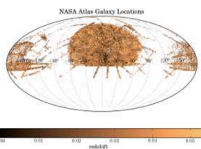
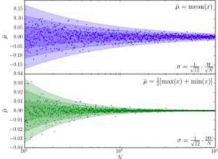
Videos

[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.

5

3 Description of the courses

- STAT 391 “Probability and Statistics for Computer Science” “Computational Statistical Modeling and Machine Learning”
- ASTRO 497 “Data Intensive Astronomy”

These courses aims are explicitly to (1) to give students a hands on experience, through programming, and performing real data analyses on a computer, with the computational aspects of statistical modeling in general, and (2) to introduce them to the machine learning methodology in particular, with specific attention to the issues of big data.

The material covered will be partly overlapping with other courses (e.g. regression, probability models for discrete data) and partly new (e.g. classification, clustering). However, the treatment of the material will stress on the interaction of computational and statistical aspects in modeling and prediction with scientific and engineering data. In this sense, the overlap has not been avoided; so that the student can gain a new, computational perspective, on areas already studied from a more theoretical point of view.

3.1 STAT 535 and ASTRO 397 Syllabi

- STAT 391 Draft Syllabus
 - a review of the concept of likelihood and Max Likelihood estimation (cases in which MLE has no closed form, gradient ascent/Newton estimation of MLE)
 - a review of basic probability models with focus on ML estimation of these models, supported heavily by simulation (e.g demonstrating gaussianity of MLE for certain models, and non-gaussianity for other models, including Zipf’s law type distributions)
 - models for statistical prediction, with focus on classification
 - in less detail: intro and computational aspects of other statistical topics like density estimation, clustering, testing, model selection and validation
 - intro to programming in Python, and to Python libraries supporting scientific com-

puting

- examples real applications from engineering and sciences (image analysis, information retrieval, etc.)

Learning goals for STAT 391 Ability to perform computationally intense/automated and efficient data analysis. Ability to combine existing tools and libraries with programming in a general purpose language (python). Working knowledge of the most important/main machine learning tools and methods, as well as their probabilistic interpretation. Understanding of the practical implications of theoretical results like independence, overfitting, consistency of an estimator.

- ASTRO 487 Draft Syllabus
 - Computational Challenges in data-intensive astronomy and astrophysics: -data types and data management systems -types of computational problems and strategies for speeding them up - data visualization challenges - selection effects and truncated/censored data in astronomical context

Exploratory techniques and searching for structure (e.g non-parametric density estimation, finding clusters - focus on non-parametrics and large data)

Dimensionality reduction

- review of principal component analysis in a large data context - non-negative matrix factorization - independent component analysis and projection pursuit

Regression and model fitting for large data

Basics of time series analysis

- applications using real data from large sky surveys

Adoption and development of cross-disciplinary tools (e.g. numerical algorithms, visualization methods, data-human interaction) in the context of big data, astronomical or otherwise **fill in examples, why these...**

Learning goals for ASTRO 597

3.2 Format and student experience

The courses will consist of lectures, homework assignments, 1–2 miniprojects, and a final exam. The TA will hold recitations; about half of these will be in a (virtual) computer lab environment.

The Computer Lab recitations will offer support for learning Python, as well as specific data analysis tools **examples: libraries, tools, from the book** The student will practice working in groups, the technique of extreme programming, using a debugger.

Another experience in the computer lab will be actual data analysis and visualization using the tools.

the postdoc will train/supervise the TA's for both courses

Homework assignments There will be 4-5 weekly homework assignments. They will contain concept problems, algorithms problems, programming assignments, and data analysis assignments.

TODO: credits for each course
put in some pictures (from 391..)
sample student evaluations
why python
what support we have for python
why astronomy good testbed

A note on the overlap between STAT 391 and ASTRO 497: Where the two courses have overlapping topics, ASTRO will consider the big data case explicitly, while STAT 391 will be considering the connections between statistical theory and computation. STAT 391 will support more basic Python, while ASTRO 597 will support Python libraries for big data.

Lectures We will blend in computer demos, class question and answer, and group discussion with the standard lecture format.

Textbooks **TBW**

3.3 Precursors (and about the PI's— here or later?)

A previous version of **STAT 391** was developed and taught by Meilăfor 11 years as “Probability and Statistics for Computer Science”.

Recently, (i.e. 2010) CSE introduced their own introduction to probability and statistics,

CSE 312. While STAT 391 continues to be taught and recommended as an elective, it was clear that the course could not remain in its original form. Therefore, in Spring 2013, the PI Meilăwith Hoyt Koepke, revised the course, having in mind that

- the audience was now literate in statistics and probability
- the course could now be opened to a larger audience

I opted to replace the introductory material with more advanced topics, and for these I chose a set of basic machine learning topics. I also introduced more substantial data analysis assignments. These changes were implemented by Hoyt Kopke, who taught the class. The course web page is at . The student feedback to this pilot experiment was very encouraging. **specifics: how many students, what depts, they liked being made to learn python, loved the projects too, level was demanding**

Meilăwith Connolly co-taught an extremely well received course at CMU in 1999-200. **fill in 1-2 more sentences**

ASTRO 497 A,Z say something

3.4 Short term: how/where will these classes fit

some of this needs revision

- These courses fit very well with the original goal and mission of the ACMS program. The revision we propose will take it to a level corresponding to the current state
- Crosscultural diversity: ACMS students will share the class with CS majors (STAT 391) or Astronomy and Physics majors
- General increased interest from employers in computational statistics, big data, machine learning, together or separately. We expect that ACMS students, too, will be well served by these courses.

Links

- STAT 391 Spring 2013 web site <http://www.stat.washi>

- AstroML Textbook web site <http://www.amazon.com/Data-Mining-Machine-Learning-Astronomy/dp/0691156224>

3.5 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/>

3.6 The UW E-science program

The educational program described in this proposal fits naturally with a number of initiatives to integrate the mathematical and physical sciences around the concepts of “Big Data”, that are ongoing at the University of Washington. Prime amongst these endeavors are the creation of an eScience Institute, the development of a Data Science Environment program sponsored by the Moore and Sloan foundations, and a new NSF-sponsored IGERT graduate student program in “Big Data”. We will leverage these programs throughout this initiative by enhancing the curriculum and research experiences of the mathematical and statistical students in data-intensive science, by providing the resources to enable hands-on research experiences, and by engaging the IGERT graduate students in working with the undergraduate students.

3.7 The University of Washington eScience Institute

The eScience Institute was created in 2008 with a goal of advancing data-driven techniques and technologies. eScience has a core faculty and permanent research staff with long standing collaborations with Statistics, Applied Math, Computer Science, and the domain sciences. Activities sponsored through eScience include bootcamps, a long-running seminar series on eScience, a new Phd program, graduate and informal education curricula in the emerging area of data science, an established suite of UW-wide research cyber-infrastructure services for computing, data management and scalable analytics tools (for example, SQLShare), the creation of physical infrastructure for high-performance computing (Hyak) and scalable storage (lolo), and significant enhancements to campus and regional networks that facilitated access to cloud services.

This year, the eScience Institute received an award from the Moore and Sloan foundations to create a Data Science Environment at the university. The physical space associated with this award will include classrooms and meeting areas for seminar series on data intensive research, and free cloud computing resources for data stor-

age and analysis. The program itself will focus on career paths for researchers at the interface of science and data but will include the educational and career development of these researchers (though courses and curricula for data science). As part of our program we will utilize the eScience resources. **WE NEED A LETTER FROM ED** Bootcamps and classes will provide students with introductory material for the computational components of our new courses. Seminar series will illustrate how the skills develop through statistics and machine learning might be applied to the broader science community (and the workforce in general). The cloud resources for analyzing data in the student projects will be made available to the ACML students through the Data Science Environment.

The Data Science Environment expects to hire promising undergraduate students from the mathematical and statistical field to work within the physical space on research software projects under the mentorship of the core staff and eScience leadership. Research opportunities for undergraduates have a profound impact on their education and careers; the limiting factor is typically the management overhead. By providing a physical location, a critical mass of mentors, a queue of shovel-ready projects, and a structured management environment, we will significantly increase the number of students we can mentor at one time.

3.8 Graduate education: the “Big-Data” IGERT

Most disciplines, from physical to life sciences, have entered an era, where discovery is no longer limited by the collection and processing of data, but by the management, analysis, and visualization of this information. Novel developments in instrumentation have lead to a tremendous increase in the magnitude of this data, forcing scientists to perform analyses on data that is too big for standard desktop computing tools, i.e., leading to a focus on *Big Data*. While significant steps in the development of statistical methodologies for processing Big Data have been made, these “hammers” are rarely accessible to

domain scientists, either because these scientists lack training in statistics or because the tools, designed for industry, fail to meet their needs.

The recognition of this gap between the needs and capabilities of the current generation of graduate students has led to the development of an IGERT funded program in “Big Data”. The transformative path to address these challenges comprises: developing a new PhD program, with a novel curriculum and practical training, leveraging committed partnerships with 11 of the very best companies and national labs in the field; enabling the development of computational tools and statistical and machine learning models for managing, analyzing and visualizing Big Data. The goal of the IGERT is to create a new breed of scientists: domain scientists proficient in and able to develop tools for Big Data Science, and statisticians and computer scientists versed in the needs and challenges of Big Data Science, and able to develop tools and models to tackle some of the biggest scientific questions of the coming decades. Most importantly, this IGERT will have an immutable focus on multidisciplinary training, thus blurring the distinctions between domain scientists, computer scientists, and statisticians.

The IGERT program naturally maps to our proposed undergraduate big data tract. We expect that the graduate students will undertake some of the supervision of the projects described in Section XX and the mentoring of the mathematical and statistical students. Each of these elements will be integrated in our proposed curricula. One of the key goals is developing a diverse STEM workforce, including strategies for recruiting and retaining traditionally under-represented groups, women and students with disabilities. We will, therefore, leverage the graduate student program to...

4 PROJECT ORGANIZATION

4.1 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.

4.2 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

4.3 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.