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&Eand 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.

Education and Training (Section 2

- Transform the ACMS (Applied and Computational Math Sciences) Program Statistics track
  - two new courses
  - new curriculum
  - mentoring undergraduates
- **Research and mentoring for research ??**
  - ...
  - facilitate entrance of Statistics students in computationally intensive data analysis research
  - mentor Astronomy graduate students
  - mentoring undergraduates by PI's and graduate students
  - undergraduate research seminar
- **Dissemination and Outreach ??**
  - software and datasets infrastructure for teaching computational statistics and machine learning
  - workshop for UW faculty in the use of the infrastructure and our experience with it
  - training workshop with outside participation

This component will be centered around the Applied and Computational Math Sciences (ACMS) Program at the University of Washington. We will significanly 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 continous 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 expend 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 astro-

nomical 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 facilitiate 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 each component 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 incorporating CDS&Econtent in existing courses and develop new courses in CDS&Eareas

- 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 development 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 developing 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 environmnets) for data exploration

- we will address data-intensive scientific problems (arising in astronomy and ...)

3) workshop

- professional development activities centered on CDS&Efor faculty or K-12 teachers

- we will foster interdisciplinary collaborations

- we will create new learning environments and experiences that immerse students in CDS&Ewhile energizing and sustaining the professional growth of faculty in CDS&E

2 Education and Training

2.1 About the ACMS program

The ACMS *(Applied and Computational Mathematical Sciences)* was introduced in 1998 as a joint undergraduate program between the departments of Applied Mathematics, Computer Science and Engineering, Mathematics and Statistics, among others.

The ACMS program is structured into a *core*, totaling 43 credits, and a set of options, or *tracks*. The same set of core courses is required for all options (with some exceptions). Options are either associated with a particular application do-

main (Biological and Life Sciences, Mathematical Economics, Social and Behavioral Sciences, Engineering and Physical Sciences) or with a particular area of specialization in the mathematical sciences (*Discrete Math and Algorithms, Operations Research, Scientific Computing and Numerical Analysis, and Statistics*).

It is the Statistics track of ACMS that we are aiming to tranform. Currently, this track contains, in addition to the core courses (43 credits), the following:

Option Core (37 credits)

- PHYS 121-2-3 (replaceable by other courses in application areas)
- STAT 302 Statistical Software and Its Applications (R course, irregularly offered, 2 credits)
- STAT 340 Intro Probability and Mathematical Statistics
- STAT 341-2 Intro to Probability and Statistical Inference I,II
- STAT 421 Applied Statistics and Experiment Design
- STAT 423 Applied Regression and Analysis of Variance

Option Electives (10 credits)

- MATH/STAT 396 Probability III
- MATH/STAT 491-2 Intro Stochastic Processes
- STAT 403 Intro Resampling Inference
- STAT 427 Intro Analysis of Categorical Data
- STAT/BIOST 529 Sample Survey Techniques
- CSE 373 Data Structures
- MATH 300 Mathematical Reasoning
- MATH 327 Intro Real Analysis I
- MATH 407–8–9 Linear, Nonlinear,& Discrete Optimization
- STAT 428 Multivariate Analysis for the Social Sciences
- GEOG 426 Quantitative Methods in Geography
- QMETH 528 Survey Sampling Applications

The program web page recommends that "[this track] is ideally suited as a second major for students with a primary focus in the biological sciences, earth sciences, social sciences, engineering, or management science." De facto, the track curricullum differs little, most notably by the presence of the computer programming class CSE 142, from the "standard" Statistics major.

Enrollment The ACMS major is competitive, with the number of majors capped at about 200. During the recent academic years, graduation numbers have passed 100 student per year, with the current enrollment at 147[1]. Of these, the Statistics track accounts for 4 students currently enrolled (none as double majors), and of 2 to 6 students graduated in each of the last 5 years. In the same time, enrollment in the Statistics major is at an all times high, with fill of which XX women.

2.2 Transforming the Statistics ACMS track

We will transform the ACMS Statistics track into a virtually new major, a "computationally minded Statistics major". In other words, we will not aim to produce scientists who also know statistics (which can be well served by Statistics minor and other tracks of ACMS) but full-fledged statisticians who can function autonomously in the cyberworld.

There are several motivating factors:

- Interest in statistics is at an all times high, as witnessed by our enrollment numbers. We expect that the ACMS Statistics track will reach similar enrollment numbers. Note that at the current enrollment in the other tracks, the total ACMS enrollment will be near the 200 level.
- the next items could be shortened if needed The role of the statistician as data ana-

[1]Complete breakdown of ACMS graduation numbers from 1998 on are in the Supplementary material.

lyst has become both more central, and more demanding. As more data analysis and more decisions based on it are being shifted from humans to computers, statistics is called upon to design and validate the procedures used to take these decisions. Thus, the statician's importance in a much wider range of human activities, be they science or buisness. But the statistician's responsability is comensurately growins, as she needs to master the skills required by cyber-enabled science and data analysis.

- These skills include are not limited to computer using and programming skills. The nature of statistical analyses itself is affected get some citations as it has been long recognized. Computationally aware statistical methods need to be designed. Very large samples will support models of a complexity that could not be considered in real-life scenarios a decade or two ago. It is now not possible to perform model selection by explicitly comparing all possible models (i.e. there can be an exponential number of models to compare) and regularization methods are often called into play, as are approximate computational techniques. Leveraging unlabeled data for prediction is now an almost universal necessity. Devising new methods to evaluate complex models in an environment that is not stationary, nor controlled are some of the challenges of modern data analysis.

- The proposal fits in and leverages other efforts and successes at UW in creating a strong collaborative environment for CDS&E (the e-science Institute, the IGERT for graduate education in Big Data and related PhD tracks already existing in CSE and Statistics). The time is ripe to involve undergraduates, and specifically statistics and mathematics majors in this change.

- This proposal is in the spirit of the ACMS original mission, bringing this part up to speed with the demands of the coming decade.

Stages

Stage 1: Design and introduce two new courses STAT 391, ASTR 497, as *electives* in the track. Start the undergraduate research seminar.

Stage 2: Redesign the track: move STAT 391, Stochastic Processes STAT 481, 482 into the core. As this will no longer be regarded as a second major for science students, we will make room by removing the PHYS 121-2-3 courses (15 credits) from it. Reorganizing the electives into two groups: group I (math/stat electives) and group II computing and science electives. Further reorganization of the core and electives in consultation with Statistics faculty and the other ACMS participating departments and Schools.

Stage 3: Incorporate feedback from evaluators and all participants. Continue developing the courses' software, data and (for STAT 391) lesson modules.

here or later? In addition, although we are at the moment not explicitly proposing this, we will investigate if the two new courses could be made accessible to mathematics majors as well. Prof. William Stein, creator and leader of the `Sage` project, teaches a successful Python programming course in Mathematics, which could serve to build a pathway towards one or both of the two new courses, if a way to satisfy the Statistics prerequisite is also found.

move elsewhere To develop a software infrastructure for teaching CDS&E in Python. This will include data sets, data analysis problems, software libraries, and course modules built around the data and problems. This infrastructure will be made available via the web. Due to its modular structure, it will be useable as needed by instructors in other courses. To organize an Undergraduate Research Seminar. In this seminar, unlike the ACMS To organize a 3 day workshop for instructors in statistics and related fields that will teach the basics of using our sofware infrastructure and will impart our experience in the project.

Why this particular approach Currently, computer education is assured primarily through two core courses, CSE 142, 143 [**?**]. These set the foundations in the understanding of computer programming, but they are of a general

nature, and are mainly focuses towards preparing future programmers. (These courses are the same courses that the about 500 CSE majors take). Thus, there is no room for data analysis applications in these courses. The need for a dedicated scientific computing course was recognized, hence the Scientific Computing track and courses developed by AMATH. There is also an R course (STAT 302, 2 credits) offered irregularly. However, for reasons we will develop later, we consider that Python is the more appropriate computer language for our goals. Thus, we will both introduce Python and will use it to teach statitical methodology. As Python will be taught as a second programming language, and as it is similar enough to Java, we expect the students to be assimilate it quickly under our guidance.

to continue after the proposal more fleshed out

put this somewhere? *A fundamental concept at the core of the ACMS program is modeling - casting a real world problem in a way that makes it amenable to mathematical, statistical, or computational analysis. Continuous modeling, while central to many applications, is not part of the CSE undergraduate curriculum, and statistical modeling is only a small component.*

2.3 Description of the courses

- STAT 391"Probability and Statistics for Computer Science" "Computational Statistical Modeling and Machine Learning"
- ASTR 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.

- STAT 391Draft 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 computing
  - examples real applications from engineering and sciences (image analysis, information retrieval, etc.)

  Prerequisites: an introductory programming course (not necessarily in Python) or equivalent programming experience; an introductory statistics course; mathematics multivariate calculus

  Learning goals for STAT 391 Ability to perform computationally intense/authomated 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.

- Motivation for ASTR 497 shorter,more to

the point of grant 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.

- ASTR 497 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 exaamples, why these...

Learning goals for ASTR 497 - familiarity with drivers for and accomplishments of modern astronomical surveys
- 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// - help to develop a diverse STEM workforce//

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, in press) coauthored by the Co-PIs on this prooposal.

2.4 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 ASTR 497: Where the two courses have overlapping topics, ASTR 497 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 ASTR 497 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 Unfortunately, there is no single textbook one can assign for this course. We will rely partly on Meilă's previously developed course notes for *"Probability and Statistics for Computer Science"*, a computatonally minded introductory course, partly on new course notes to be developed by Meilăspecifically for this more advanced course, and partly on the textbook of ASTR 497 which will provide among others the Python exercises. We are also considering [] Daniela's machine learning book.

2.5 Python Packages

We will leverage all the publicly available modern python tools. In particular, seminar work will be built around the *astroML* package (available from http://www.astroml.org) that was developed to support textbook to be used with the proposed ASTR 497course. *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 (there are close to two hundred examples of machine learning and visualization in the code library that supports the textbook alone). In addition to *astroML* package, we will expose students to several other popular and widely used toolkits (e.g. PyMC for Markov chain Monte Carlo methods, and HealPy for spherical coordinates and spherical harmonic transformations).

As an example of methods and exercises available in *astroML*, we single out methods for reducing data dimensionality. Many astronomical analyses must address the question of the complexity as well as size of the data set. Dimensionality reduction methods address the complexity issue by finding the directions within a multivariate data set that contain most of the information. Classical approaches for identifying the principal dimensions include principal com-

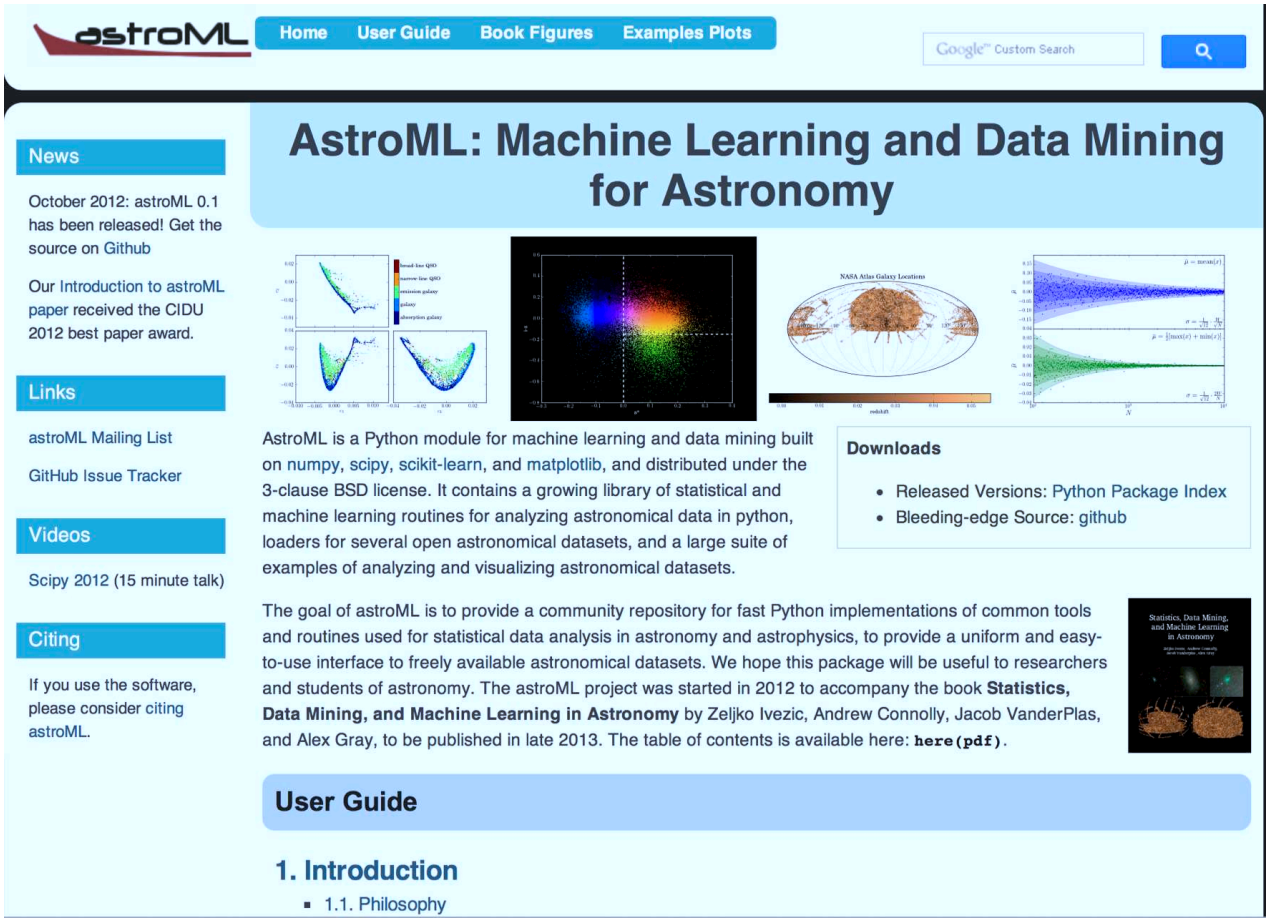

Figure 1: We will leverage all the modern python tools available in *astroML* and other packages, including a large number of practical data-intensive exercises developed to support textbook that will be used with the proposed ASTR 497 course.

ponent analysis (PCA), independent component analysis (ICA), and non-negative matrix factorization (NMF). These methods are implemented in *astroML*, with a user-friendly interface and adequate documentation (see Figure 1). Furthermore, *astroML* also includes easy-to-use code to automatically access and download spectra collected by the Sloan Digital Sky Survey (currently a "gold standard" for modern astronomical surveys and big data sets; see sdss.org). Therefore, an undergraduate student will not only be exposed to modern statistical methods and a cutting-edge astronomical data set, but will be empowered to actually apply these methods to a real complex and massive data set. The result of this exercise is shown in Figure **??**. With such a positive experience, it is very likely that such a student would not have difficulties applying the same methods and tools later to potentially unrelated problems.

2.6 What we are building on

Meilă has extensive previous experience teaching computational statistics courses at all levels. She developed the course "Probability and Statistics for Computer Science" an introductory course aimed at computer science majors, complete with exercises and demos in Matlab, revised the Statistical Computing graduate course sequence, later developed the Statistical Learning

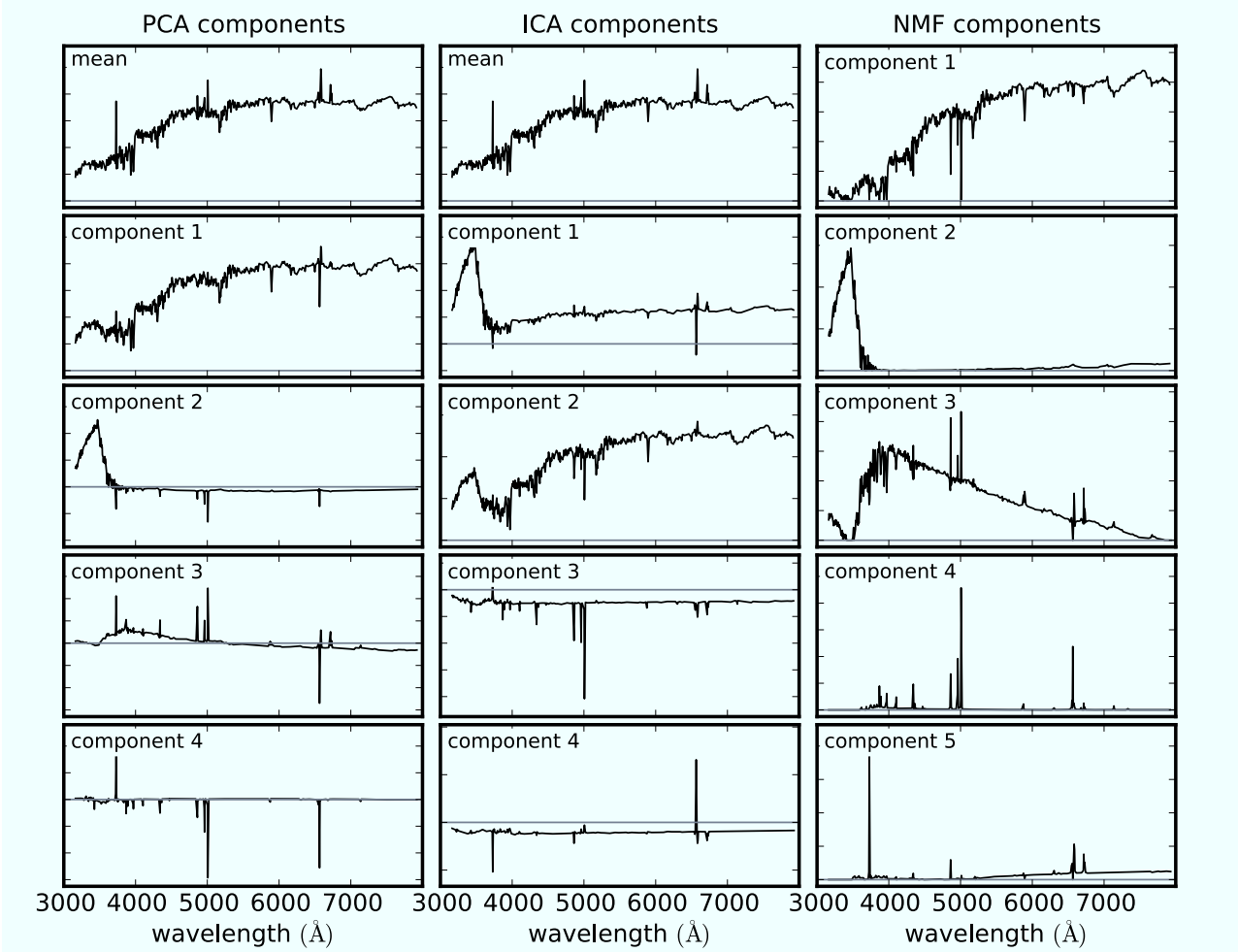

Figure 2: An example of sophisticated tools available in *astroML* and exercises that will be used in practical seminar work. The figure shows a comparison of the decomposition of SDSS spectra using PCA (left panel), ICA (middle panel) and NMF (right panel). The rank of the component increases from top to bottom. For the ICA and PCA the first component is the mean spectrum (NMF does not require mean subtraction). All of these techniques isolate a common set of spectral features (identifying features associated with the continuum and line emission). The ordering of the spectral components is technique dependent.

graduate sequence and lead, with E. Fox, the effort to introduce the Machine Learning/Big Data PhD track in the Statistics department.

Some of the more advance

ZI, AC, JVdP

Meilăwith Connoly co-taught an extremely well received course at CMU in 1999-2000, "Computational Statistics of Multi-Dimensional Scientific Databases", which reunited students from Statistics, Computer Science and Astronomy, as well as faculty from these departments.

The Statistics department runs a *Virtual computer lab* that will be used by STAT 391 students in their quiz sessions and assignments. The clusters `newton1,2`, consisting of 8, respectively 2 high memory nodes, accquired from a UW Student Technology Grant, will serve as basis for the graduate student research. shall i buy more nodes??

The *AstroML* book and web site will provide the starting point for the new courses' software and data infrastructure.

move to university context or remove **Short term: how/where will these classes fit**

- These courses fit very well with the original goal and mission of the ACMS program.
- Crosscultural diversity: these classes can also serve science and CSE majors as well as other quantitatively able students across the UW.

Links

STAT 391Spring 2013 web site `http://www.stat.washington.edu/courses/stat391/spring13/`

AstroML textbook web site `http://www.amazon.com/Data-Mining-Machine-Learning-Astronomy/dp/0691151691`

3 Dissemination and Outreach

3.1 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 wor kshop venue, and four grants of $500 to young faculty and postdocs, we need about $7,500 fo r the workshop.

4 Research

to be filled with awesome stuff

5 Activities

tentative title

5.1 Grad students involvement

For the statistics graduate students funded by this grant, I envision

- to fund several students for a relatively short time (2 quarters to 1 year) We adopt this "rotation" plan recognizing that developing tools . (This "rotation" model will also assure that the "API" of our data infrastructure is truly functional, as each departing grad student will have to ensure the smoothe transition to her/his successor.)
- Student helps develop the software and data infrastructure for the program. Searches for available data sets, curates them, writes preprocessing software if necessary, designes tasks and exercises.
- In the same time, student gets practice and training with analyzing large data.
- The student can be advised by the PI, by another Statistics faculty, or by another UW scientist with interests in statistical analysis of big data. Gradually, a research problem is formulated, and the student formulates on the data and methodology related to this problem. Although this "apprenticeship", some students continue their research supervised by other advisors.
- The students will also be strongly recommended to TA the STAT 391 class, thus rounding their preparation and self-confidence.
- This will enable stat PhD students to play useful roles in the other NSF funded initiatives at UW. For instance, in the CSNE, with with the PI is involved, where the data collected, far from reaching Tera byte sizes, represents a daunting challenge for the average Statistics graduate student who hasn't accquired a CS degree before. This plan also harmonizes with the IGERT plan of offering graduate students the experience of working with domain scientists, and with outside big data companies via internships.
- Finally, the graduate student will participate in the Undergraduate Research Seminar, and will mentor 1-2 undergraduates.

5.2 PI Involvement

- First year: program coordination and curriculum planning. Within the department and between the other departments participating in ACMS. Preparation for phase 2 of the project happens now.

  Develops (with the RA and the co-PI's) a plan for the core Python numerical and data structures libraries to be taught/presented.

  Starts developing course notes for STAT 391.

- Second year: Evaluates the success of the first phase and incorporates lessons learned. Writes the bulk of the STAT 391 course notes. Supervises the undergraduate research seminar. Major work (with RA) selecting/curating the data sets to be used in STAT 391.

- Third year: Evaluation of the second phase and fine-tuning of the curriculum and program requirements. Explores possibilities to open this pathway to math majors.

5.3 Undergraduate involvement

(sketch) The PI and co-PI's have extensive track records of involving undergrads in research.

Ugrads who take either of the courses will be involved in research projects either (1) along with the funded graduate students, or (2) in the UW units that provide data sets. All undergraduates involved in research under this project, along with the graduate students will participate in an *Undergraduate Research Seminar* where they will present and discuss their work.

as well as any undergraduate students involved in research with Statistics faculty who would like to participate, will

5.4 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.

5.5 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 cyberinfrastructure 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 storage 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.

5.6 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...

5.7 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 website[2]. 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...

6 PROJECT ORGANIZATION

6.1 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.

6.2 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

6.3 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.

[2] `http://www.astro.washington.edu/vanderplas/Astr599/`