1 INTRODUCTION

This three-year project aims to deliver

2 PLAN

What will we do...

2.1 Teaching Program

What will we do...

2.2 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[1]. 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/`

2.3 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.