

# Statistics, Data Mining, and Machine Learning in Astronomy

Andrew Connolly, Brigitta Sipocz, Zeljko Ivezic  
University of Washington

# Welcome

**DATE:** Monday, 6 January 2020

**TIME:** 13:00 - 17:00pm

**LOCATION:** ROOM 323 A at the Hawai'i Convention Center

This workshop will introduce the astronomical community to the 2nd edition of the book *Statistics, Data Mining, and Machine Learning in Astronomy* and the associated software package *astroML*. The goal is to introduce participants to a variety of statistical and machine learning tools available within the open source *astroML* library. The format will be interactive, including short presentations on different machine learning methodologies followed by instructor-guided, Jupyter notebook-based tutorials. In these tutorial sessions participants will be able to try out the tools and to ask questions from expert users and developers. Our primary focus will be on the new material and applications in the 2nd edition of the book.

# While we discuss astroML

- <http://astroml.dirac.institute>
  - Enter your GitHub credentials (accept the authorization request)
  - Start the server
  - If successful you should see (select the astroML-workshop folder)



The screenshot shows the JupyterLab interface. At the top left is the 'DIRAC Jupyter' logo. To the right are 'Logout' and 'Control Panel' buttons. Below the logo are tabs for 'Files', 'Running', and 'Clusters'. A message says 'Select items to perform actions on them.' To the right of this message are 'Upload', 'New', and a refresh icon. The file browser shows a root directory with a dropdown showing '0' items. Below this, there are two folders: 'astroML-workshop' (last modified 16 minutes ago) and 'lost+found' (last modified 7 days ago). The columns are 'Name', 'Last Modified', and 'File size'.

If you have an error and the log says “cant resolve github.com” relaunch the server

# Outline of this workshop

- Brief introduction to astroML
  - Whats new in astroML (Andy Connolly)
- Example astronomy cases using <http://github.com/astronomy-commons>
  - Hierarchical Bayes (Zeljko Ivezić)
  - Approximate Bayesian Computation (Zeljko Ivezić)
  - Regression with Uncertainty (Brigitta Sipocz)
  - Deep Learning examples (Andy Connolly)
- The next year(s) of development

# STATISTICS, DATA MINING & MACHINE LEARNING IN ASTRONOMY

A PRACTICAL PYTHON GUIDE FOR  
THE ANALYSIS OF SURVEY DATA  
UPDATED EDITION

ŽELJKO IVEZIĆ,  
ANDREW J. CONNOLLY,  
JACOB T. VANDERPLAS  
& ALEXANDER GRAY

AstroML: is a Python module for machine learning and data mining built on numpy, scipy, scikit-learn, matplotlib, and astropy, 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.

Books available at Princeton University Press booth

# What is new in astroML

- <https://github.com/astroML/astroML>
  - Brigitta Sipocz is the new maintainer of astroML
  - Release v0.4 and new edition of the book
    - Updated to latest versions of scikit-learn, astropy, numpy, scipy
    - Ended Python 2 support
    - Included jupyter notebook examples for:
      - **Approximate Bayesian Computation**
      - **Hierarchical Bayes**
      - **Neural Networks**
      - **Convolutional Neural Networks**
      - Autoencoders (variational)

# The year(s) ahead

- Development of astroML
  - We see astroML becoming a community driven package for machine learning tools that can be shared and (eventually) moved into more stable organizations such as scikit-learn and astropy (see the development of the time series period finding code in astropy)
  - In the next year we expect to release a package for regression that works on data with errors in all variables (see preview here); creation and release of deep learning examples and architectures; example packages for scaling machine learning using dask and spark; and transition of our example code into jupyter notebooks
  - We would welcome ideas for areas you would like to see developed (or would like to help develop)