AstroML: Machine Learning for Astronomy

Jake Vanderplas
Andrew Connolly
Zeljko Ivezic
Alex Gray

[Preaching to the choir]

Python is becoming a new standard tool in Astronomy, and will remain important for the foreseeable future

Machine Learning / Statistical Data Analysis tasks in Astronomy:

- Photometric Redshifts
- Source Classification
- Dimensionality Reduction / Visualization
- Clustering
- N-point Statistics
- Period Finding
- Transient Detection
- Outlier Detection
- Matched Filtering
- Source Extraction
- Cross-matching
- ...

Machine Learning / Statistical Data Analysis tasks in Astronomy:

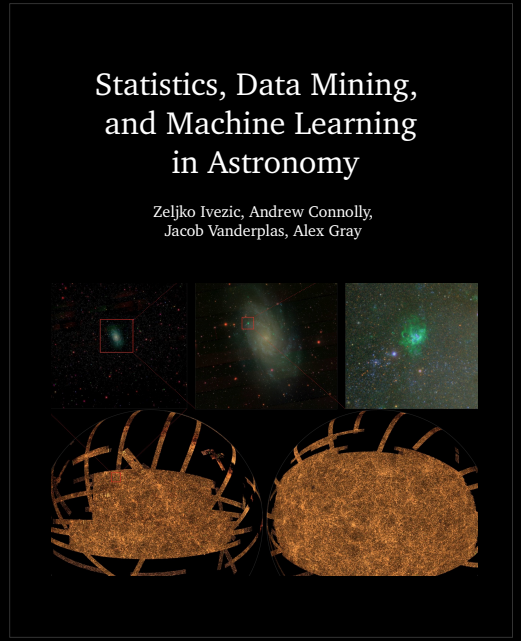
- Photometric Redshifts
- Source Classification
- Dimensionality Reduction / Visualization
- Clustering
- N-point Statistics
- Period Finding
- Transient Detection
- Outlier Detection
- Matched Filtering
- Source Extraction
- Cross-matching
- ...

Every astronomer needs these sorts of tools, and existing python packages provide an easy interface to powerful algorithms.

Statistics, Data Mining and Machine Learning in Astronomy

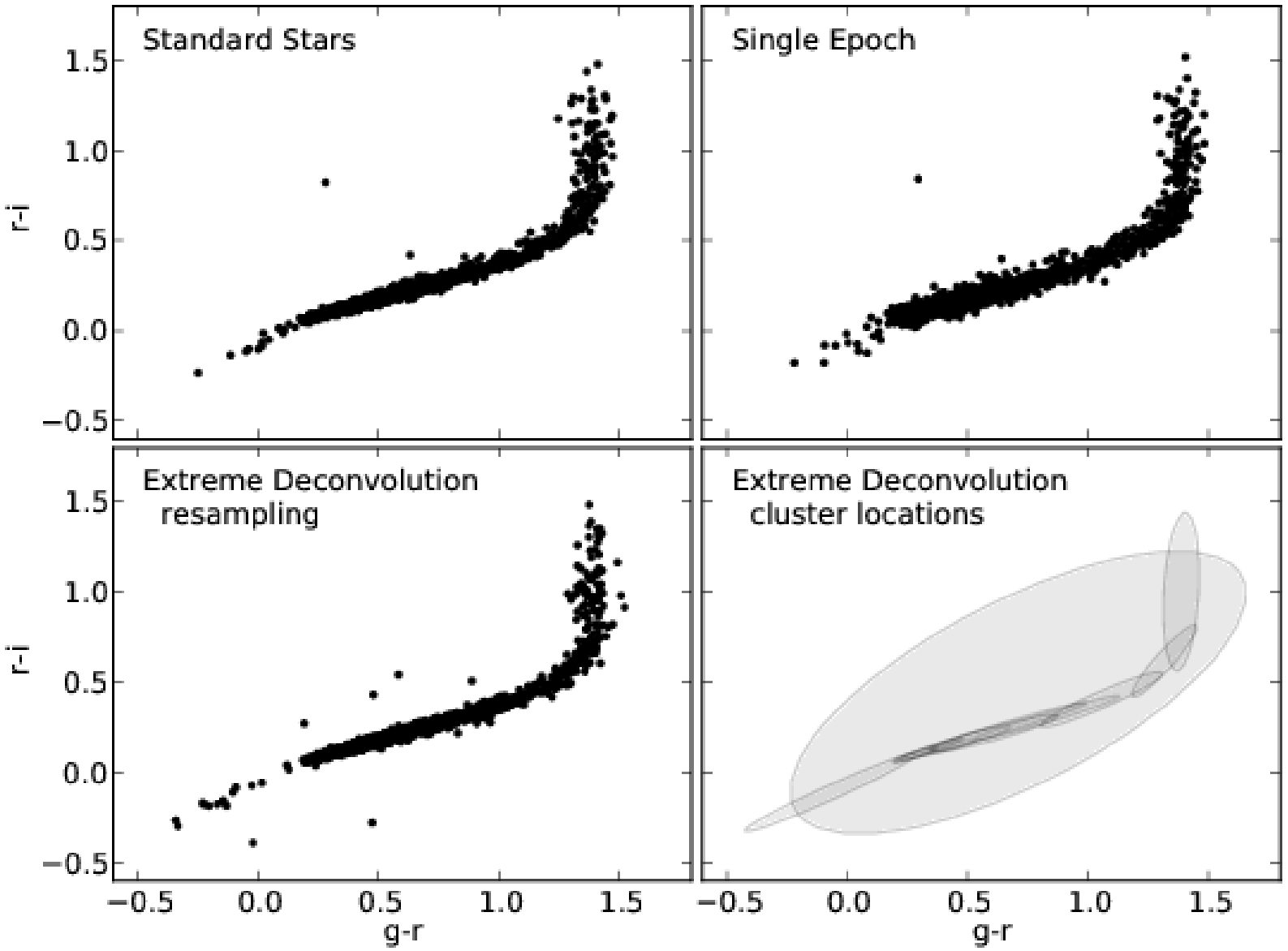
Zeljko Ivezic, Andrew Connolly,
Jacob Vanderplas, Alex Gray

Princeton University Press, 2013

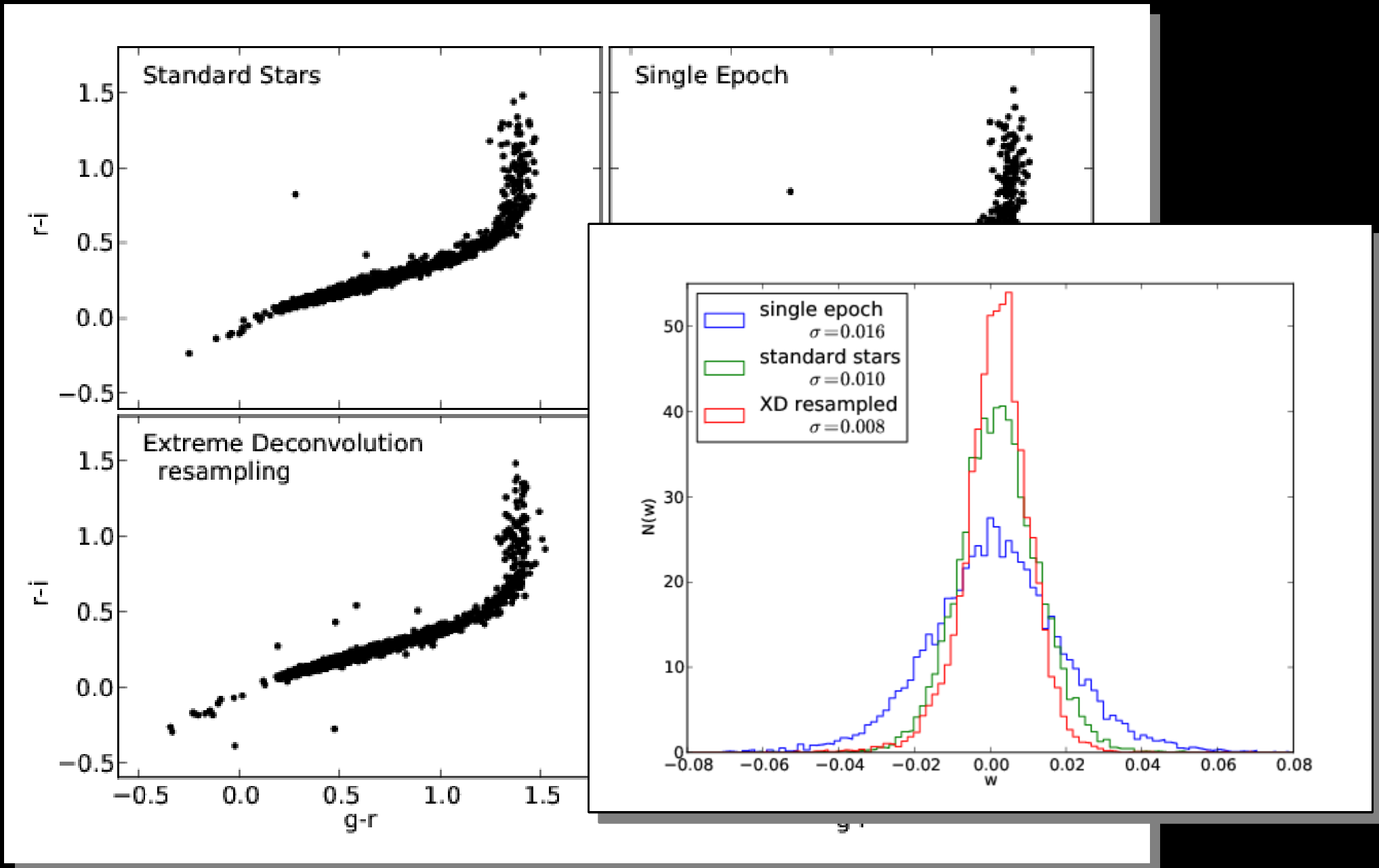


- Complete *Practical* guide to statistical analysis, data exploration, and machine learning
- Example-driven approach, using real data (SDSS, LIGO, LINEAR, WMAP, and others)
- All book figures and examples generated in python (matplotlib), with code available online – for free!
- Makes use of *numpy, scipy, matplotlib, scikit-learn, pymc, healpy,* and others
- Supporting python package: *astroML*

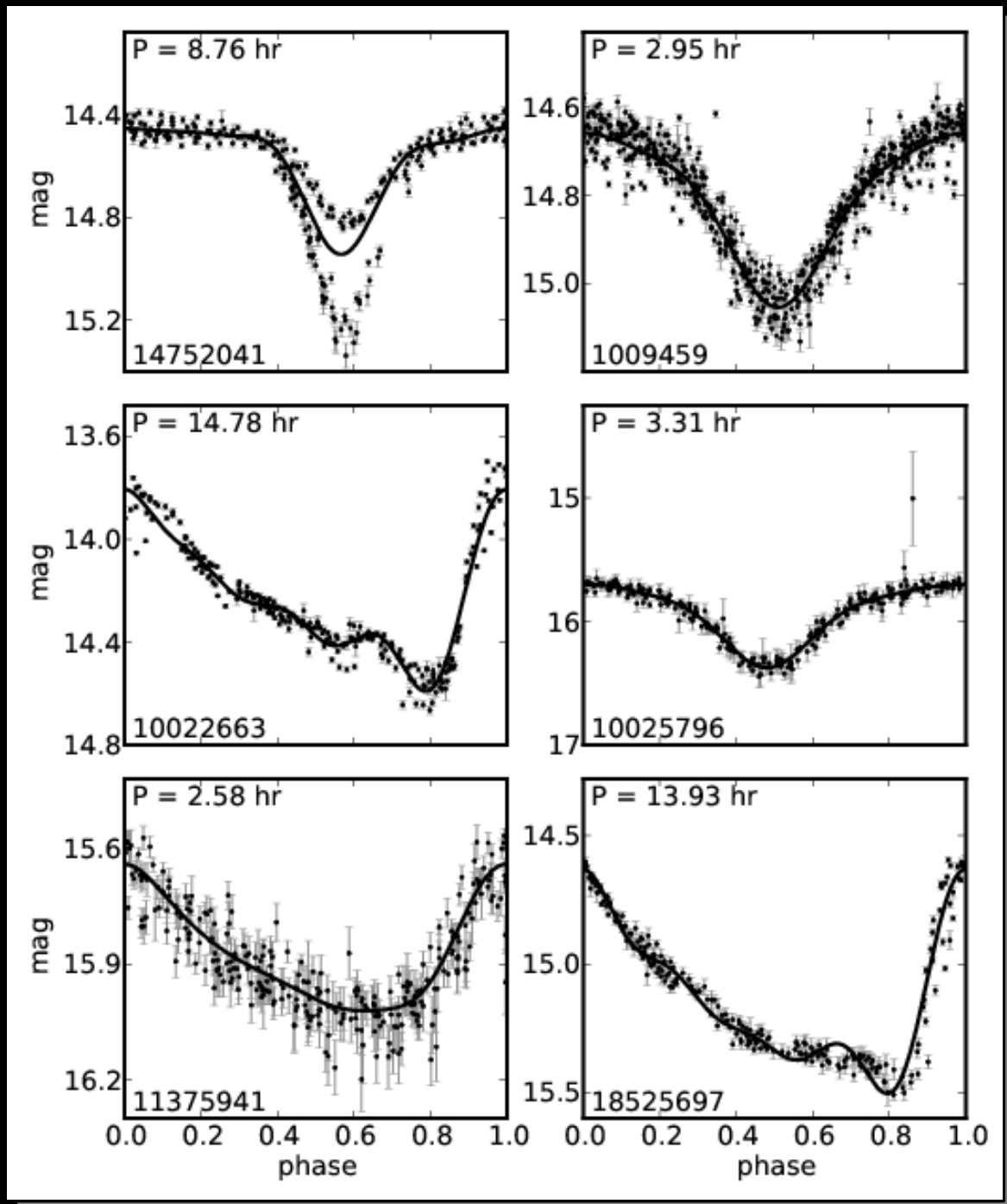
“Extreme Deconvolution” (GMM + errors): SDSS main sequence



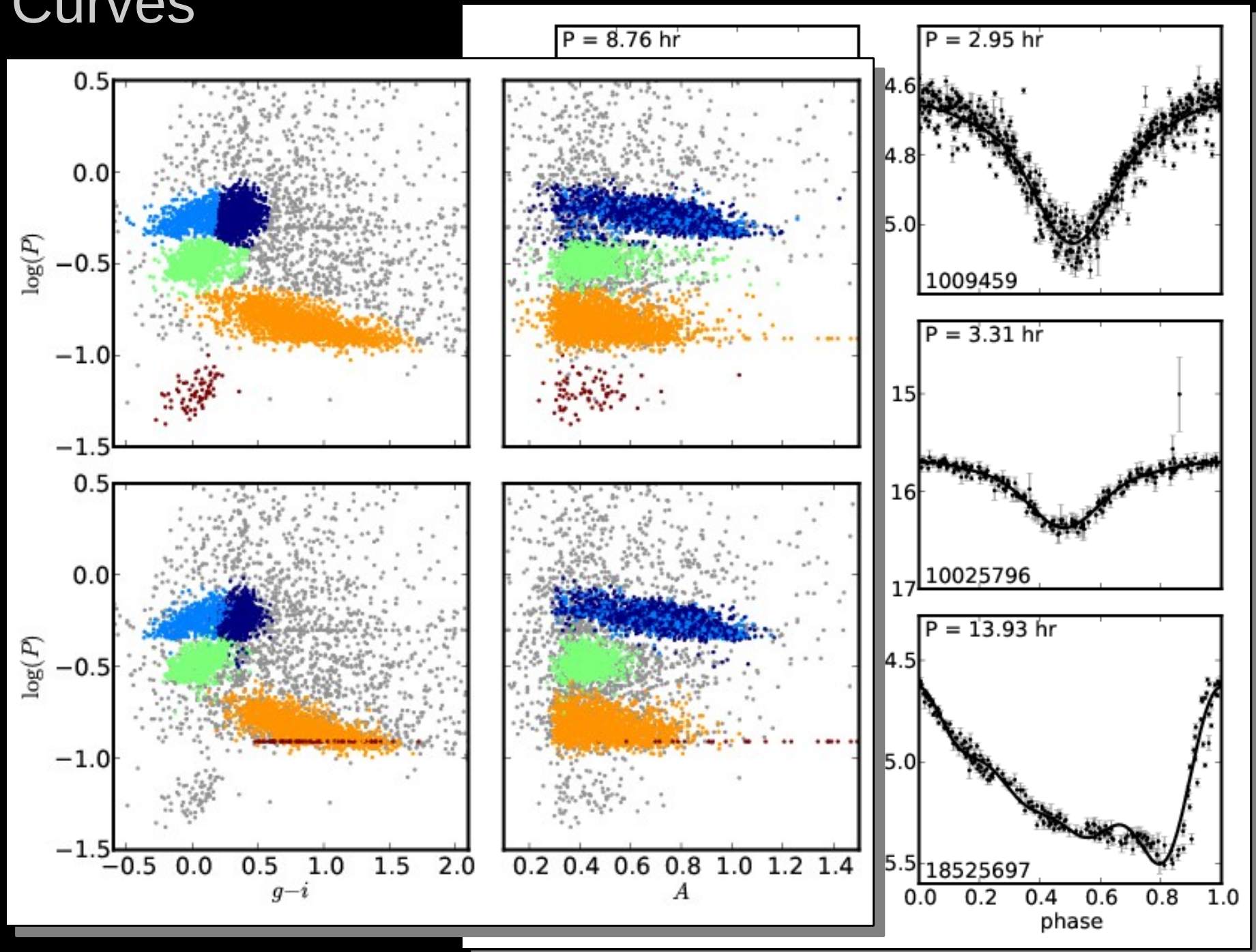
“Extreme Deconvolution” (GMM + errors): SDSS main sequence



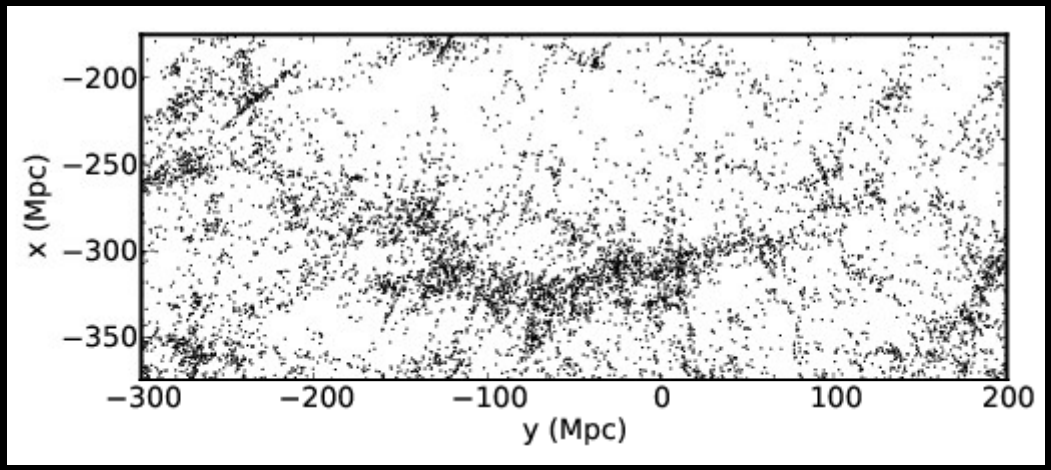
Lomb-Scargle Periodograms:
Light Curves



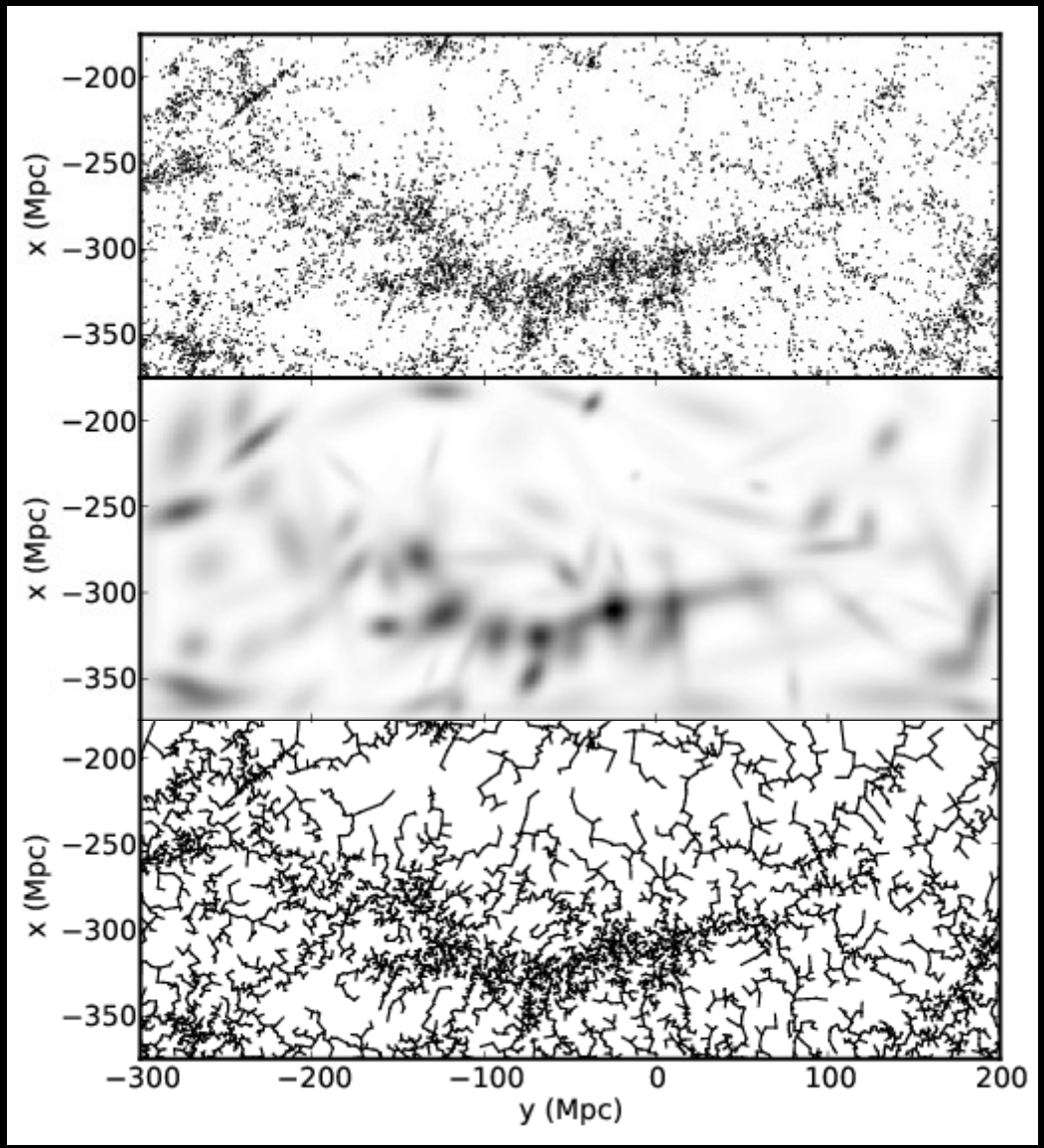
Lomb-Scargle Periodograms:
Light Curves



Clustering and Density Estimation: SDSS Great Wall



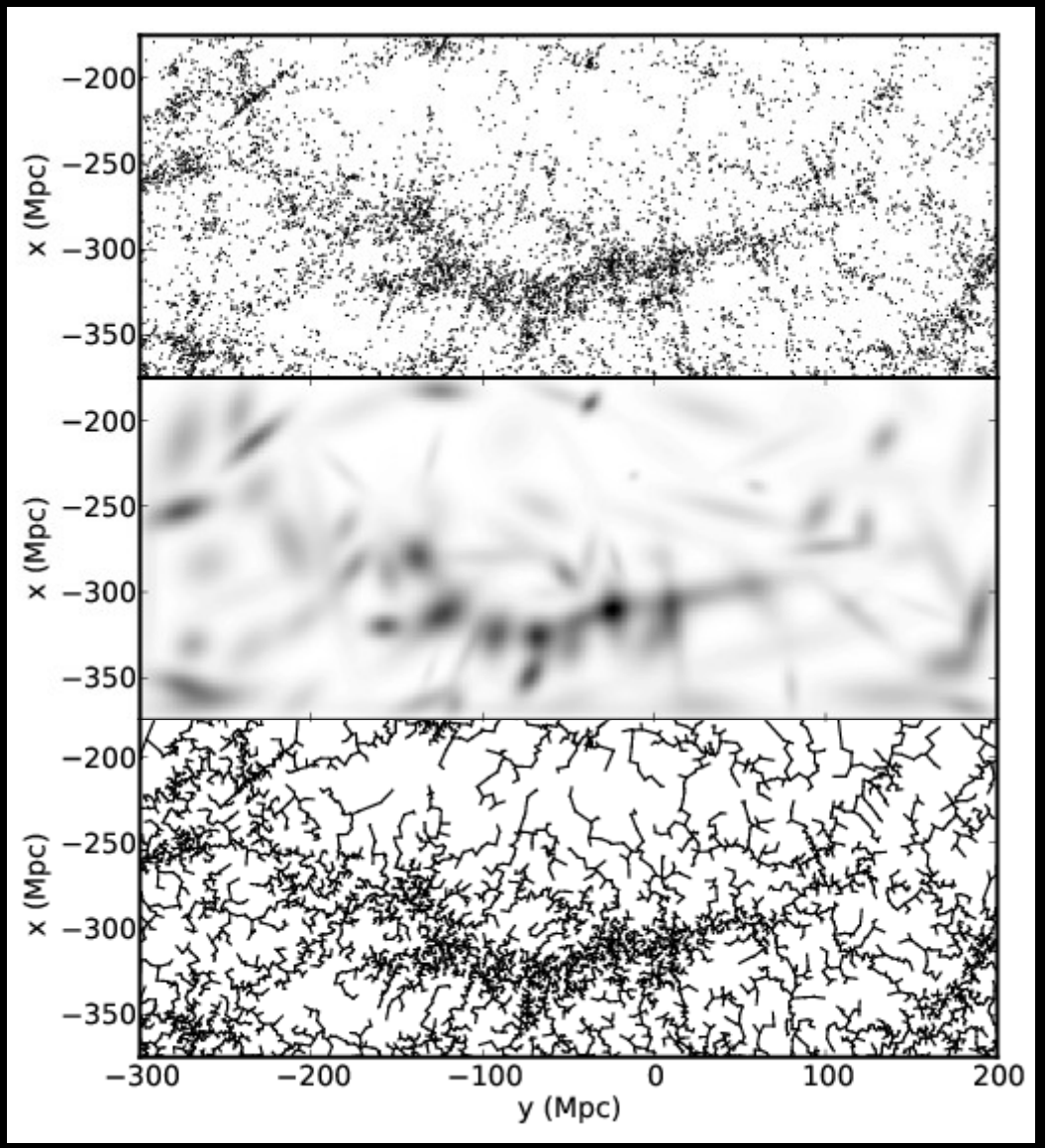
Clustering and Density Estimation: SDSS Great Wall

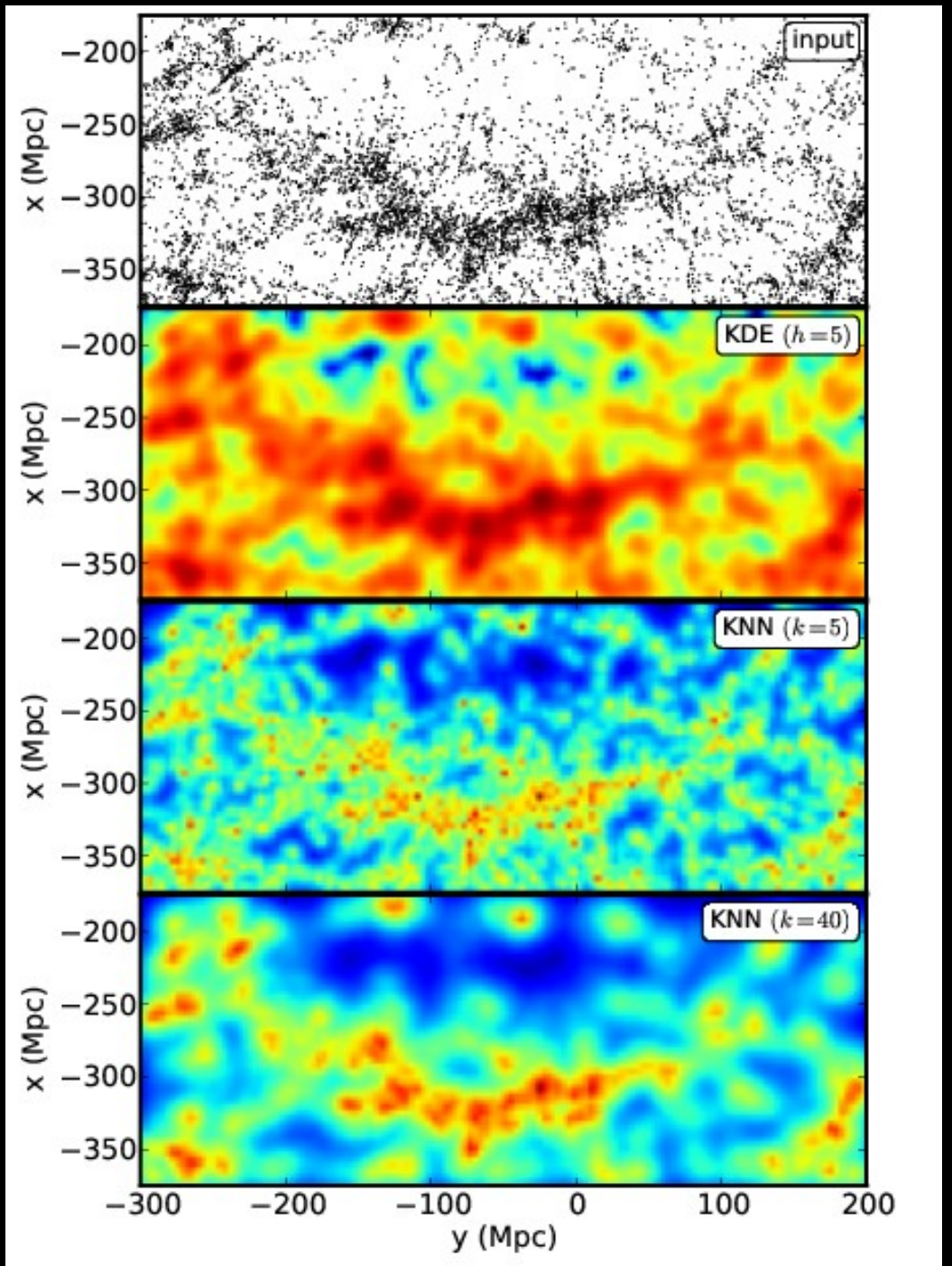


Gaussian Mixtures
(scikit-learn)

Minimum Spanning Tree
(scipy.sparse.csgraph)

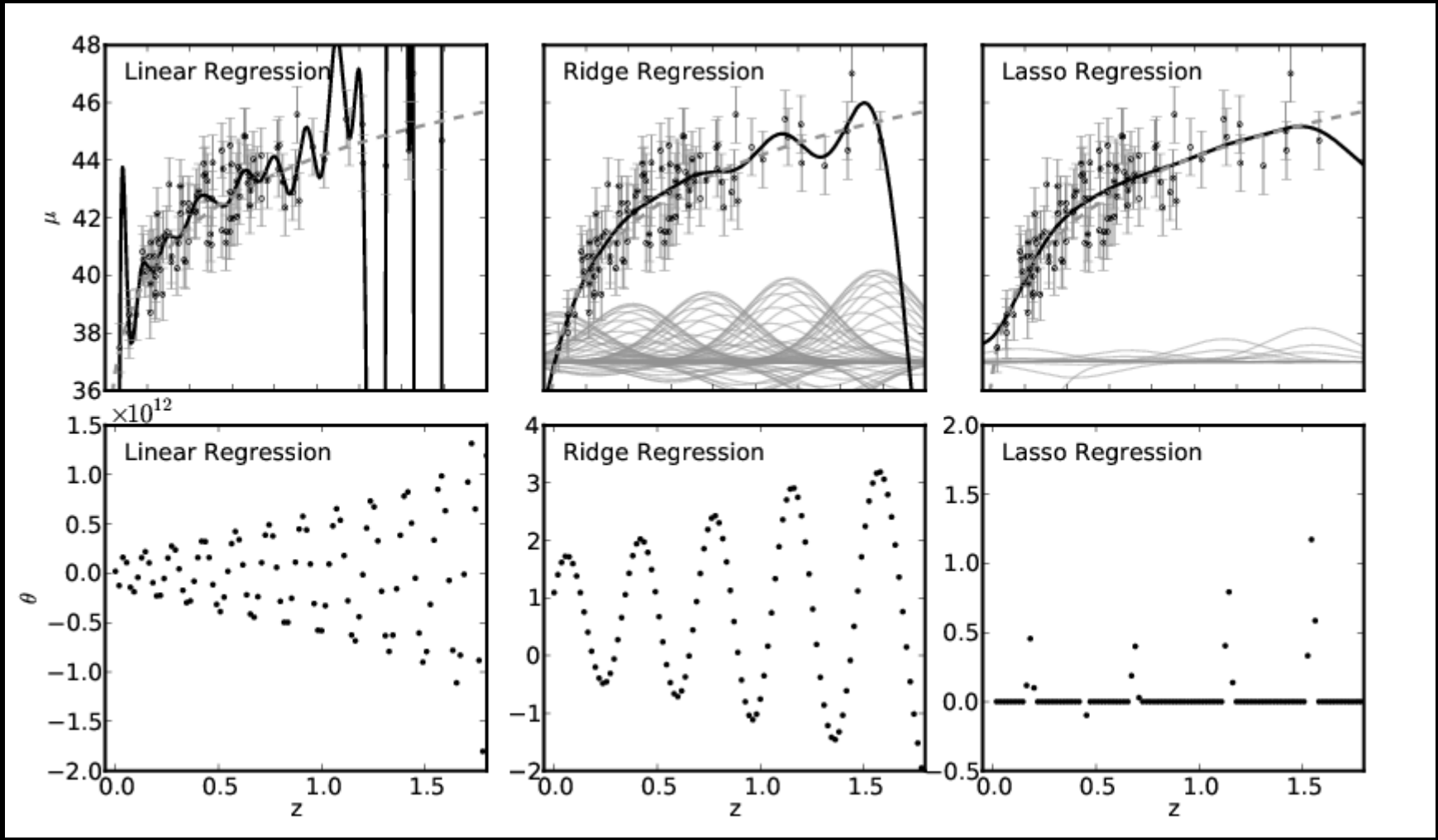
Clustering and Density Estimation: SDSS Great Wall





Exploring Regularization in regression

Radial Basis Function Regression + gaussian errors (numpy.linalg)



The Vision:

- Maintain a coherent & well-written set of examples of data processing in python *using real data.*

- Offer a standard repository of high-performance python code for astronomy.

- Complement the effort of the AstroPy team.

- Let the cream rise to the top: useful code – once battle-tested – can be moved upstream and become useful in other fields:
  - Ball Tree & two-point statistics (scikit-learn 0.10)
  - Minimum Spanning Tree (scipy 0.12)
  - binned_statistics (scipy 0.12)

Big Goal: community involvement! astroML should be much more than simply a textbook software package. It should become a community repository.

Thank You

http://www.github.com/astroML
http://astroML.github.com/sklearn_tutorial

*astroML 0.1 will be released in October 2012