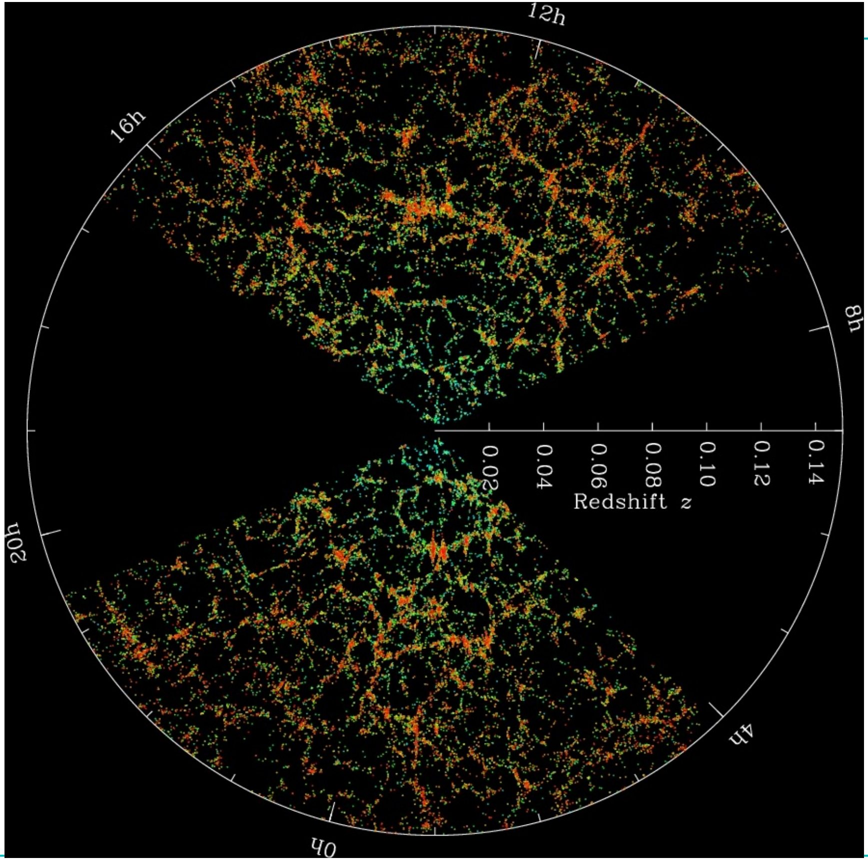
Winter ‘23, Astr 597A:
Astronomy with Rubin Observatory and LSST

Outline for today (these slides: ls.st/fs2):

1) **How to estimate distances to stars using LSST data: photo-D**
   - photo-z methods to estimate distances to galaxies (and quasars)
   - photo-D methods to estimate distances to stars

2) **A pitch for UW course Astr 598**: "Astro-statistics and Machine Learning”
   - very useful skills for analysis of Big Data in astronomy, such as LSST
   - it will be offered next time probably in Spring 23/24 by Connolly & Ivezic
   - a year from now but still before Rubin first light
   - a few practical examples...

Spatial distribution of SDSS galaxies



Left: each dot is one galaxy from SDSS

Note that the galaxy distribution is highly **inhomogeneous:** statistical details of that distribution contain rich cosmological information

For most LSST galaxies, distances (i.e. redshifts) will be estimated using LSST's broad-band photometry (as opposed to from spectra): photo-z

LSST redshift limit for galaxies: about an order of magnitude larger than for SDSS

Distance estimates for stars: photo-D

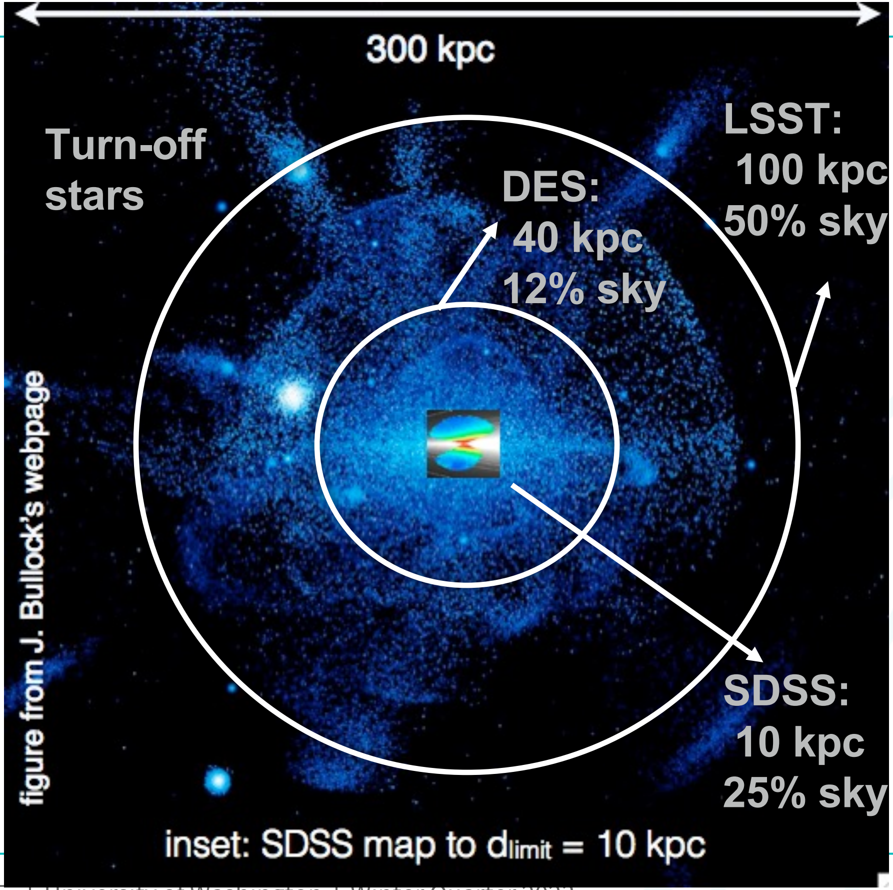
To maximize science output with **20 billions stars** measured by LSST, we need to estimate their **distances**: go from 2D to 3D studies.

Today: a brief introduction to astrophysics and statistics of stellar photo-D

This is an ongoing research project, and well suited for dissertation work (please let me know if you interested).

Milky Way science with coadded LSST data

To make such maps of the Milky Way with LSST, we need first to estimate distances to stars (~20 billion stars in LSST)



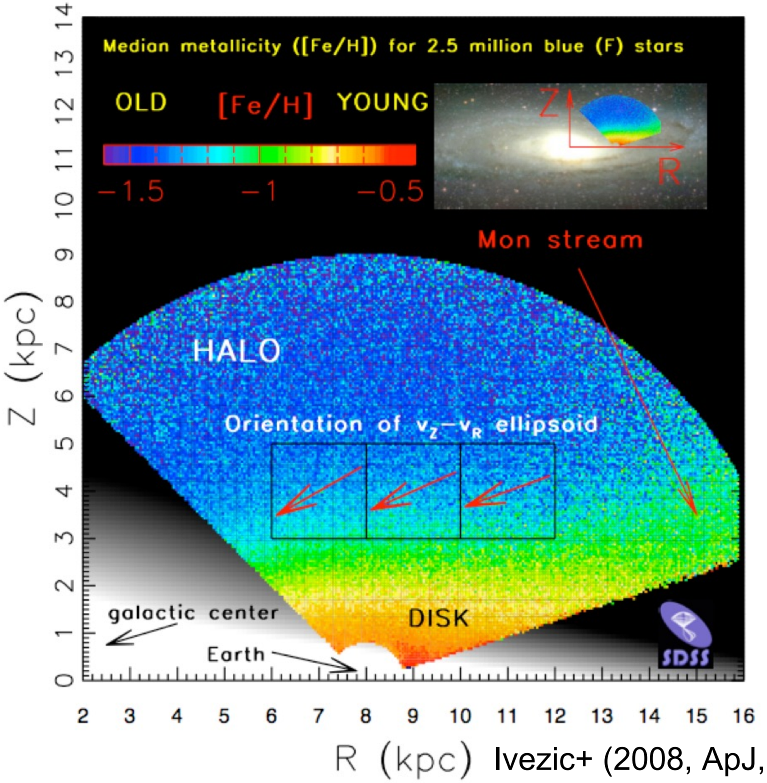
One of the four LSST Science Themes:

The Milky Way structure (stars as tracers of the structure and evolution of our Galaxy, interstellar matter, the physics of stars)

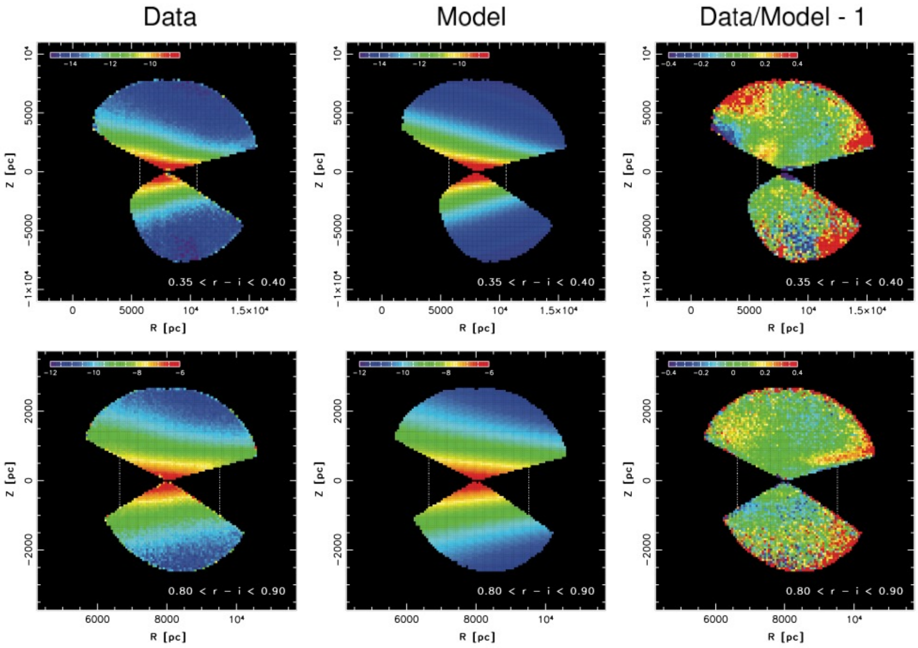
SDSS example: Jurić et al. (2008, *ApJ*, 673, 864) data-based map of stellar counts shown in the center

photo-D: science motivation

- if we know stellar distances, we can study the Milky Way structure

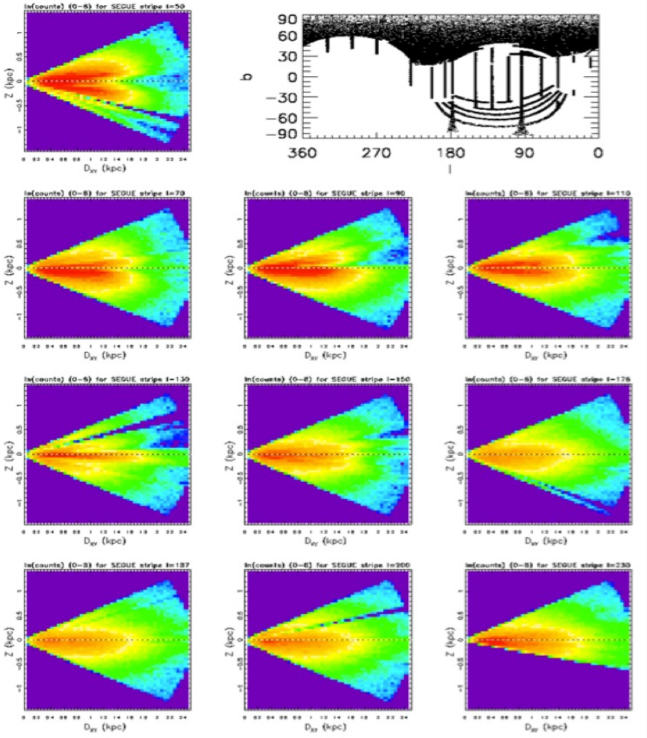


Ivezic+ (2008, ApJ, 684, 287)



Juric+ (2008, ApJ, 673, 864)

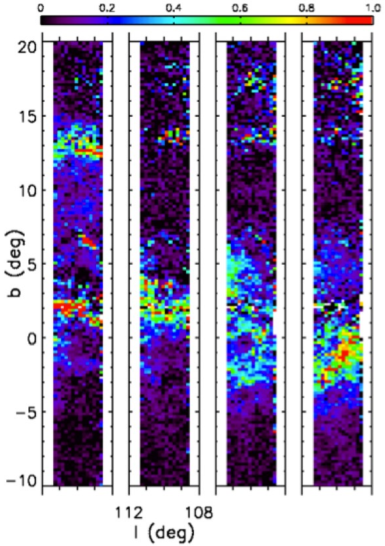
photo-D: science motivation



Stellar counts

Berry+ (2012, ApJ, 757, 166)

- if we know stellar distances, we can study the Milky Way structure
- furthermore, at low galactic latitudes, we can map dust (and its properties), too



- left: differences in median A_r for D~ 1, 1.5, 2, 2.5 kpc
- dust at b ~ 2° and b~13° is confined to D~1-1.5 kpc
- dust at -3° < b < 0° is at D~2 kpc

Dust tomography

SDSS
gri
3.5’x3.5’
r~22.5

3 arcmin is 1/10 of the full Moon’s diameter

HSC
gri
3.5’x3.5’
r~27

like LSST depth
(but tiny area)

LSST will deliver
5 million such
images

LSST will deliver colors for about 20 billion stars

LSST filter complement: ugrizy

THE ASTROPHYSICAL JOURNAL, 873:111 (44pp), 2019 March 10

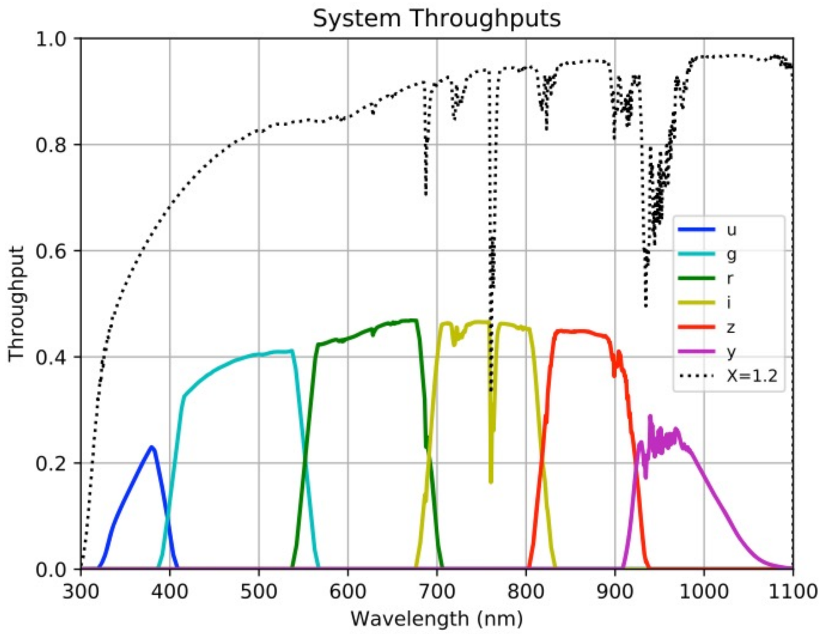


Figure 4. LSST bandpasses. The vertical axis shows the total throughput. The computation includes the atmospheric transmission (assuming an air mass of 1.2; dotted line), optics, and the detector sensitivity.

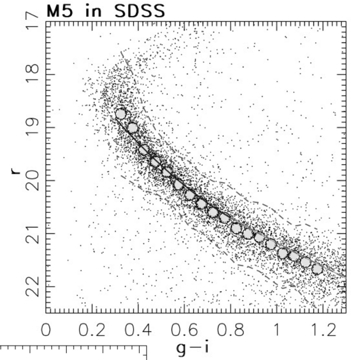
Per-band survey time allocations:
u: 8%, g:10%, r: 22%
i: 22%, z: 19%, y:19%

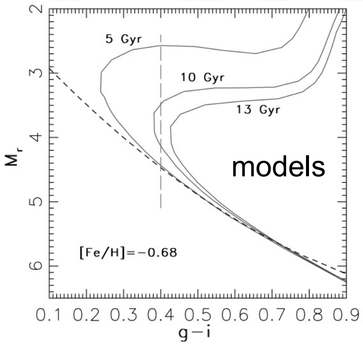
Optimized using photo-z for galaxies but consistent with star-quasar separation and stellar [Fe/H] estimates.

Similar, but not identical, to SDSS.

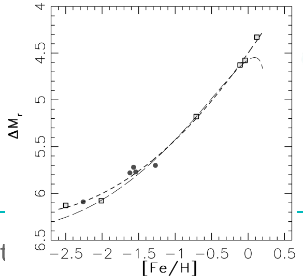
photo-D methodology: stellar astrophysics 0

- Data include apparent magnitudes: one magnitude and many colors
- Given apparent and absolute magnitudes (and perhaps extinction): Distance follows
- Stellar colors determined by: T_{eff}, [Fe/H], log(g) – or alternatively: Mr, [Fe/H], age
  different "populations" along an "isochrone": MS, giants, WDs (binaries)





- Given Mr, [Fe/H] and age, can "predict" colors from theoretical or empirical isochrones, so given observed colors, can place constraints for Mr and [Fe/H] (and sometimes on age)
- Colors can also constrain dust extinction
- Distance from: r = Mr + Ar + 5*log(D) where Mr and Ar constrained by colors



Used globular clusters to derive Mr as a function of metallicity [Fe/H] and (g-i)

photo-D methodology: stellar astrophysics 1

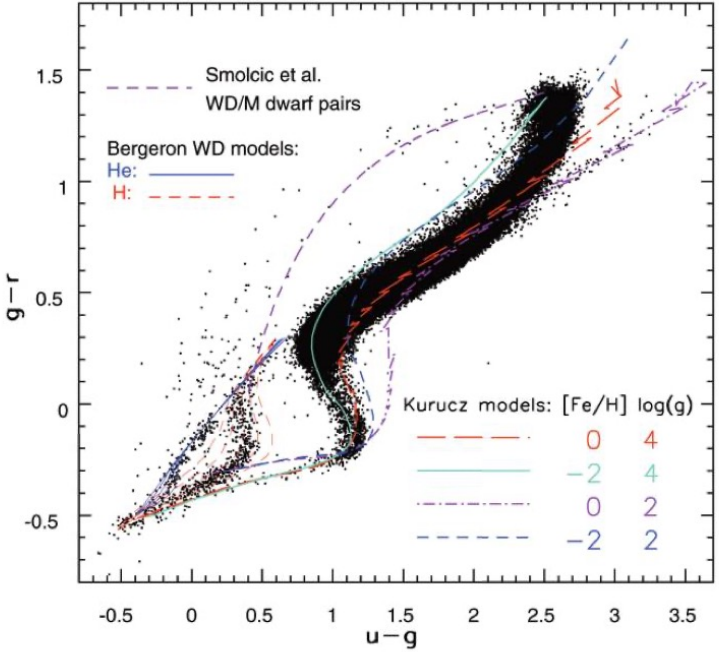


FIG. 23.—The $g - r$ vs. $u - g$ color-color diagrams for all nonvariable point sources constructed with the improved averaged photometry (*dots*). Various stellar models (Kurucz 1979; Bergeron et al. 1995; Smolčić et al. 2004) are shown by lines, as indicated in the figure. Berry+ (2012, ApJ, 757, 166)

- SDSS color-color diagram (corrected for dust):
  - stellar SEDs determined by: T_{eff}, [Fe/H], log(g)
  - different populations: MS, giants, WDs, binaries
  - for given Mr and [Fe/H] can "predict" colors:

THE ASTROPHYSICAL JOURNAL, 783:114 (16pp), 2014 March 10

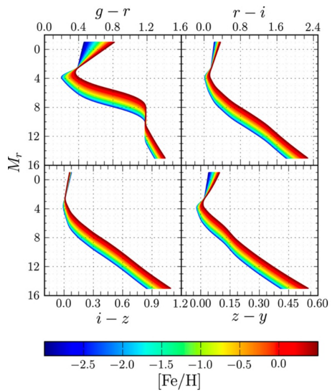


Figure 1. Model stellar colors as a function of absolute r magnitude and metallicity in Pan-STARRS 1 passbands. The stellar templates are based on PS1 color–color relations, and color is related to absolute magnitude and metallicity by SDSS observations of globular clusters (Ivezić et al. 2008a). Our empirical templates therefore assume an old stellar population. While the main sequence below the turnoff is nearly invariant with age, the giant branch and the location of the turnoff do, in reality, vary considerably with age. For this reason, we expect our inferences for main-sequence stars to be more accurate than those for giants. The narrowness of the kink at $M_r \simeq 2.4$ is an artifact of our models (see Section 4.1).

Given observed colors, can estimate Mr and [Fe/H] (and Ar) by chi2 minimization:

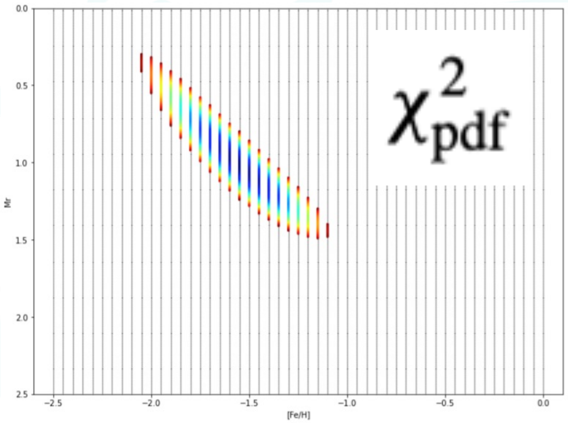


photo-D methodology: stellar astrophysics 2

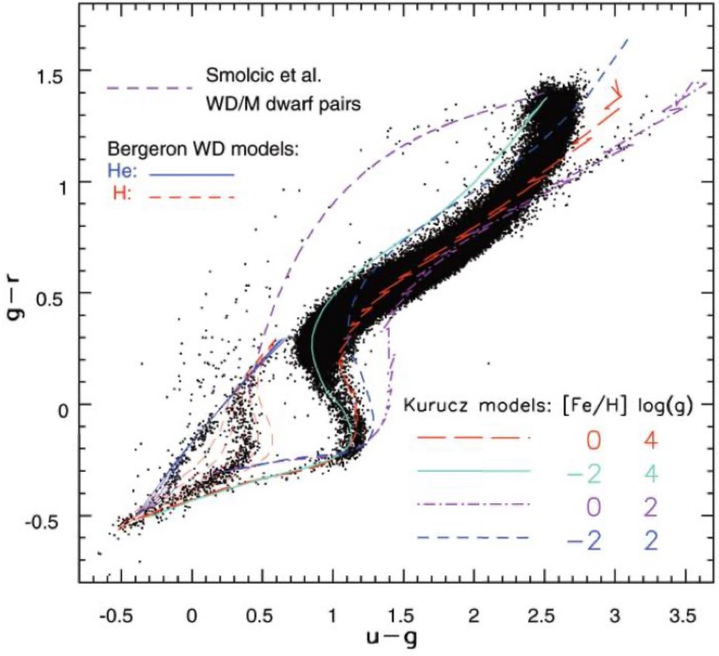
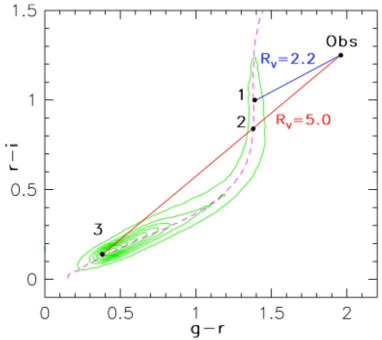


FIG. 23.—The $g - r$ vs. $u - g$ color-color diagrams for all nonvariable point sources constructed with the improved averaged photometry (*dots*). Various stellar models (Kurucz 1979; Bergeron et al. 1995; Smolčić et al. 2004) are shown by lines, as indicated in the figure. Berry+ (2012, ApJ, 757, 166)

- SDSS color-color diagram (corrected for dust):
  - stellar SEDs determined by: T_{eff}, [Fe/H], log(g)
  - different populations: MS, giants, WDs, binaries
  - **assuming** pop: from colors get best-fit SED
  - best-fit SED gives **L**uminosity constraint
  - pop **probability** can be gauged from goodness of fit, variability, priors, and other information
  - but there is dust:



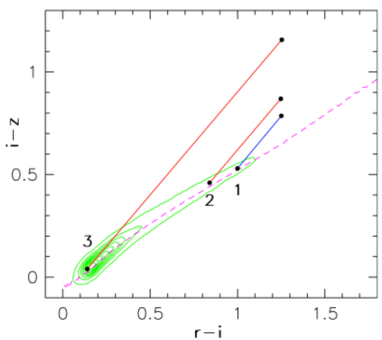


photo-D methodology: impact of dust

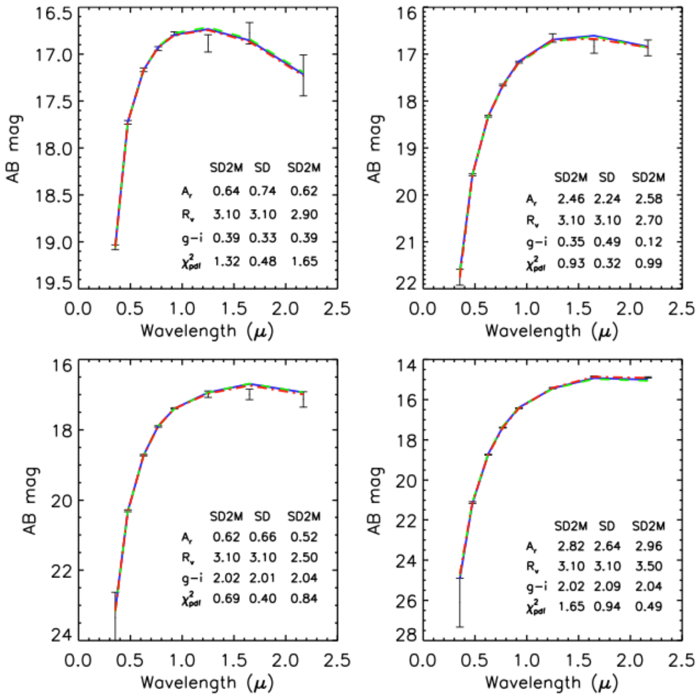


Figure 31. Comparison of three different types of best-fit SEDs: using only-SDSS data with fixed $R_V = 3.1$ (blue line) and using joint SDSS–2MASS data set with fixed R_V (green line) and with free R_V (red line). As demonstrated by the similarity of best-fit lines, the differences in best-fit parameters, listed in each panel, are due to degeneracies between intrinsic stellar color, amount of dust, and R_V. The shown cases correspond to blue and red stars (top row vs. bottom row), and small and large A_r (left column vs. right column).

- There are degeneracies with dust:
  - need to adopt an extinction curve (usually 1-parameter family, R_V)
- Two fitting philosophies:
  1) use stellar models to fit SEDs, or
  2) use high-latitude observations to fit dust-extincted low-latitude data

- The role of priors…
- Hierarchical Bayes…
- Robust and fast implementation

photo-D methodology: statistical treatment

2D projections (Ar and “Mr”) of the 3D parameter space (fixed [Fe/H])

$\chi^2_{\rm pdf}$

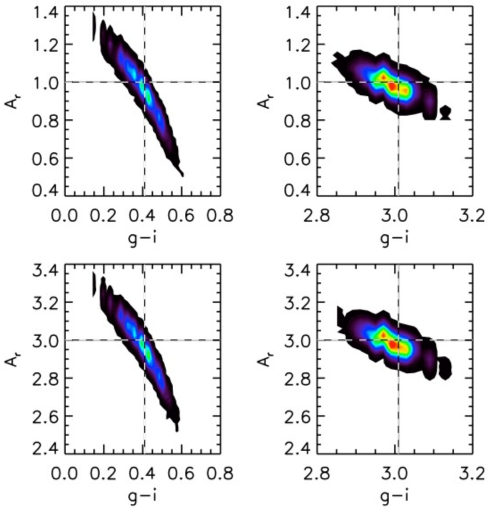


Figure 12. Analysis of the covariance in the best-fit values for A_r and $g-i$ using a simulated data set. The panels show the distributions of the best-fit values for A_r and $g-i$ for two different fiducial stars (left column: a blue star with true $g-i=0.4$; right column: a red star with true $g-i=3.0$), and two different extinction values (top panels: $A_r=1$; bottom panels: $A_r=3$). Photometric errors in the *ugriz* bands are generated using Gaussian distributions with $\sigma=0.02$ mag (uncorrelated between different bands). Note that the A_r vs. $g-i$ covariance is larger for the blue star, and does not strongly depend on assumed A_r.

- Berry+ (2012): Fit SEDs constructed from high-b observations and a dust model parametrized by R_V (shape vs. wavelength) and Av (how much dust) to SDSS data (very similar to LSST)
- Compute best-fit via a brute force χ^2-minimization process:

$$\chi^2_{\rm pdf} = \frac{1}{N-k}\sum_{i=1}^{N}\left(\frac{c_i^{\rm obs}-c_i^{\rm mod}}{\sigma_i}\right)^2$$

- c_i and σ_i are N adjacent colors and errors (e.g., u - g, g - r, etc
- the number of fitting parameters is k = 2 (the position along the locus and A_r) for fixed-R_V (R_V=3.1), and k = 3 for free-R_V
- model colors are constructed by:

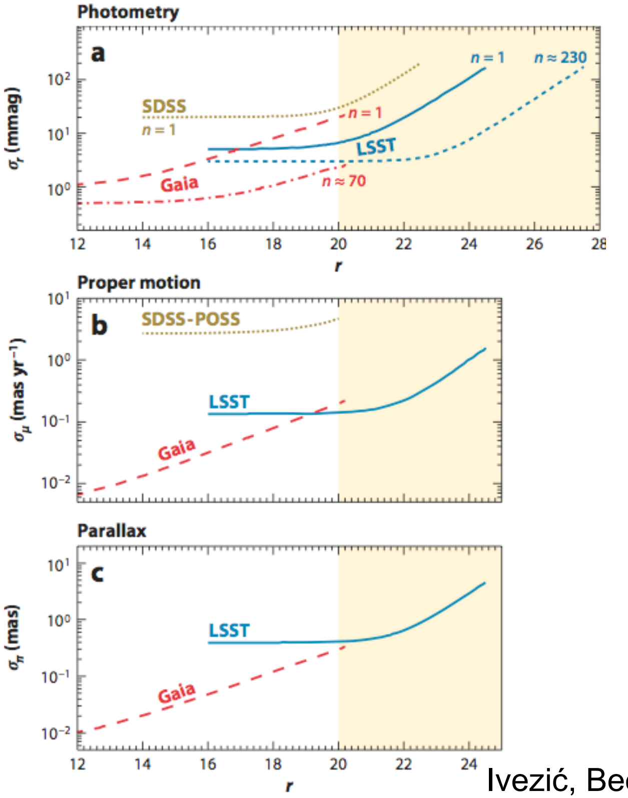
$$c^{\rm mod} = c^{\rm lib}(t) + [C_{\lambda 2}(R_V) - C_{\lambda 1}(R_V)]\ A_r$$

photo-D methodology: ongoing work...

- Use of model SEDs or **empirical isochrones?**
  - many pros and cons here...
- For a fixed (assumed population), use of **priors** for fitted parameters
  - can rely on TRILEGAL models that are available from NOIRLab's DataLab, see https://arxiv.org/abs/2208.00829v1
  - what we want to do is nicely described in **Green et al. 2014 (ApJ**, 783:114)
    (but with LSST twice as small distance errors due to u band constraining [Fe/H]!)
- Quality assurance using Gaia data products (so-called Bailer-Jones+ distances)
  - distances must be consistent with Gaia results
- **Better code:** fast and robust, configuration and metadata management
- Documentation!

N.B. There should be many similarities with photo-z frameworks.

Stellar astrometry & photometry from LSST



Photometric accuracy: random errors 0.005 mag, calibration to 0.01 mag; for light curves, LSST "takes over from Gaia" around r ~ 17

Time-resolved measurements: photometric variability, and parallax and proper motions from astrometric measurements

Gaia vs. LSST: complementarity of the two surveys: photometric, proper motion and trigonometric parallax errors are similar around r=20

Ivezić, Beers, Jurić 2012, ARA&A, 50, 251

photo-D in Green et al. (2014)

- Left: **empirical isochrones** (colors on Mr-FeH grid)
- Middle: [Fe/H] prior as a function of distance from the Galactic plane (Z)
- Right: distance prior

The Astrophysical Journal, 783:114 (16pp), 2014 March 10

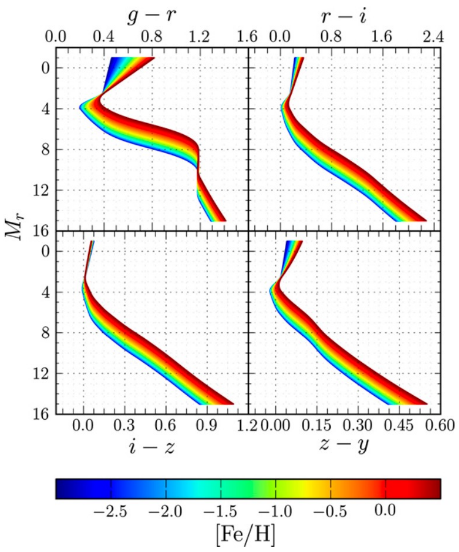


Figure 1. Model stellar colors as a function of absolute r magnitude and metallicity in Pan-STARRS 1 passbands. The stellar templates are based on PS1 color–color relations, and color is related to absolute magnitude and metallicity by SDSS observations of globular clusters (Ivezić et al. 2008a). Our empirical templates therefore assume an old stellar population. While the main sequence below the turnoff is nearly invariant with age, the giant branch and the location of the turnoff do, in reality, vary considerably with age. For this reason, we expect our inferences for main-sequence stars to be more accurate than those for giants. The narrowness of the kink at $M_r \simeq 2.4$ is an artifact of our models (see Section 4.1).

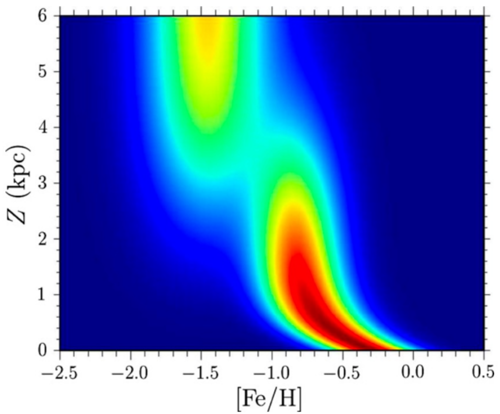


Figure 3. Metallicity prior, $p([\mathrm{Fe/H}]|\,Z)$, in the solar neighborhood ($R = 8$ kpc). High above the plane of the Galaxy, where the halo dominates, the metallicity distribution has a constant mean and variance. In the plane, where the disk dominates, the mean decreases with scale height. Adapted from Figure 9 of Ivezić et al. (2008a).

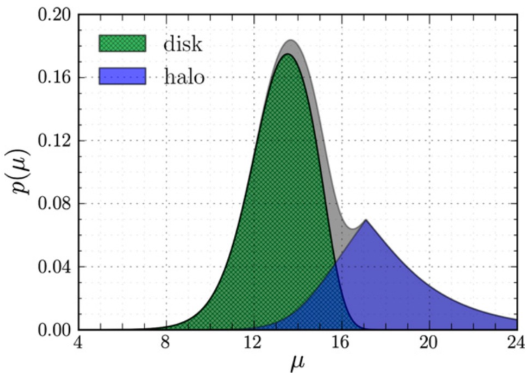


Figure 2. Distance prior for $(\ell, b) = (90°, 10°)$. The contributions of the disk and halo are shown individually in green and purple, respectively, while the total prior is given by the gray contour. The break in the contribution from the halo is due to the use of a broken power law for the number density of stars in this component.

photo-D in Green et al. (2014)

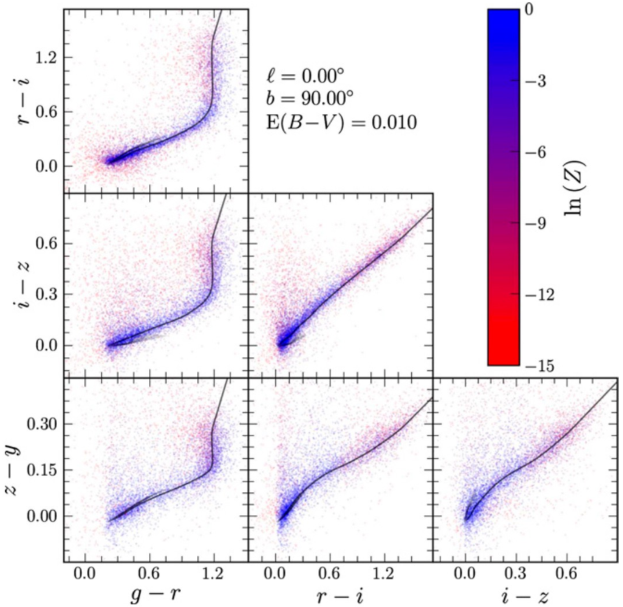


Figure 10. Comparison of PS1 stellar colors in the vicinity of the North Galactic Pole with our model colors. Each object is colored according to the evidence Z we compute. Objects represented by red dots have a low probability of being drawn from our stellar model and are rejected for the line-of-sight reddening determination. The solid black line traces our model stellar colors. Our main-sequence model colors do not depend on metallicity, while the model colors for the giant branch have a slight metallicity dependence.

- Left: test of isochrones in PS1 color-color diagrams
- Below: test using PS1 data for globular clusters, **indicates need for isochrone improvements**

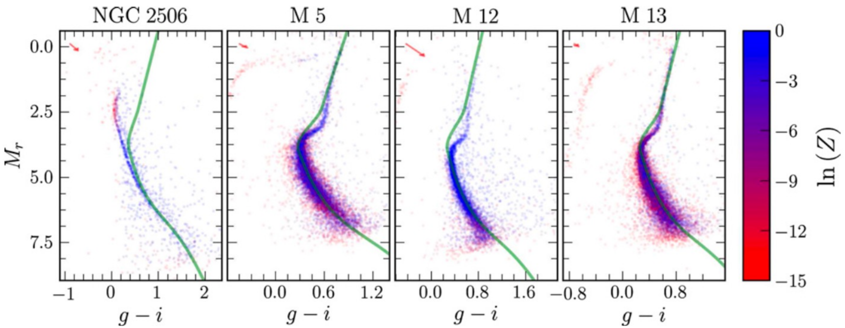
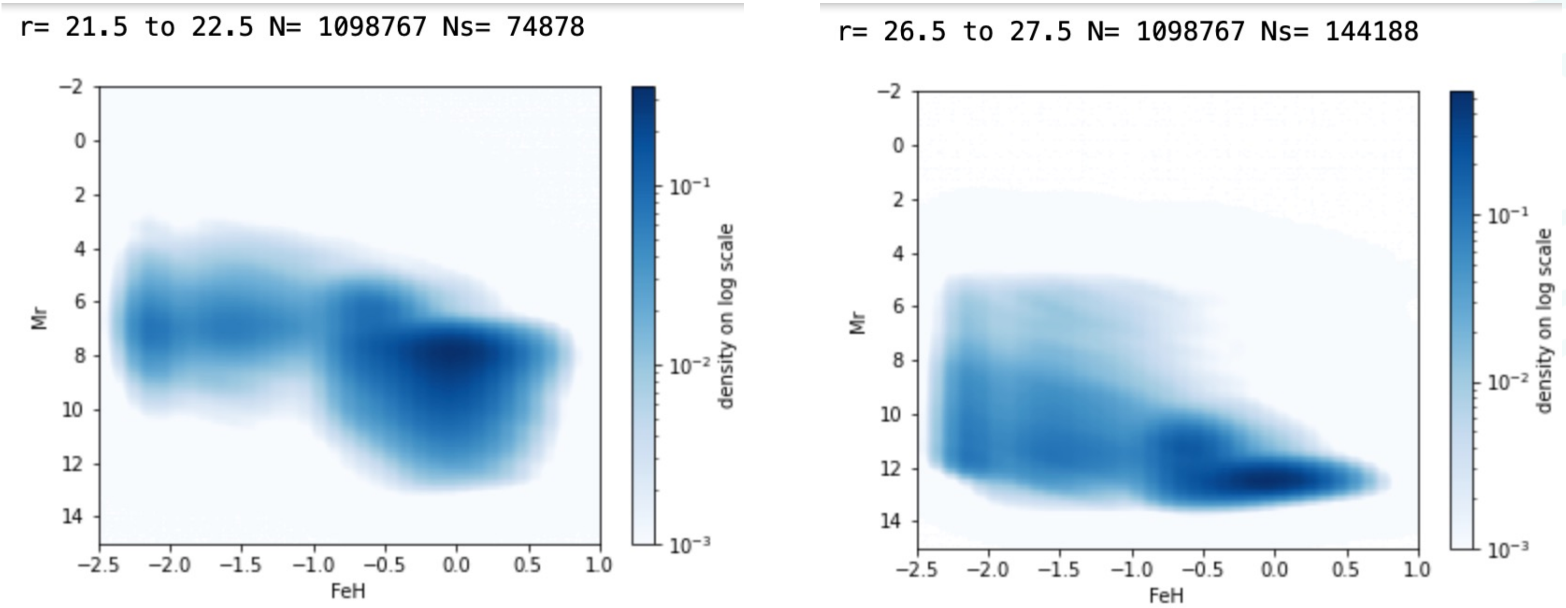


Figure 11. PS1 color–magnitude diagrams of three globular and one open cluster. For each cluster, the model isochrone with the catalog metallicity of the cluster is overplotted. The stellar photometry has been de-reddened and shifted by the catalog distance modulus to produce absolute magnitudes. The reddening vector is plotted in the top left corner of each panel in red for reference. Each star is colored by its evidence, with red stars unlikely to be drawn from our stellar model. In particular, stars which are blueward of the main-sequence turnoff, which are bluer than any star in our template library, have low evidence.

- An advantage of LSST: the u band photometry will provide much stronger constraints for [Fe/H]!

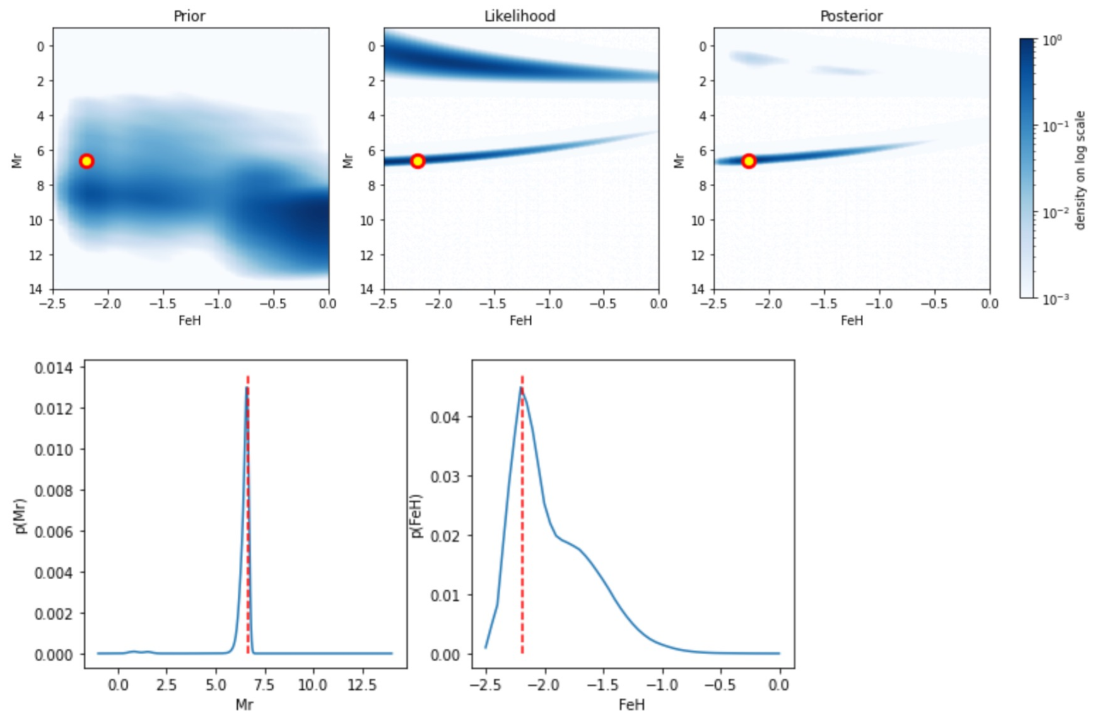
Priors from TRILEGAL (Galaxy simulation code)

- TRILEGAL & LSST paper: Dal Tio+ (2022, arXiv:2208.00829)
- for now let's assume distant halo stars with **known Ar** (dust extinction)
- priors in Mr vs. [Fe/H] plane as a function of **(R.A., Dec) and r magnitude** from TRILEGAL:



Bayes: posterior ∝ likelihood * prior

```
rmagStar = 23.68 true Mr= 6.63 true FeH= -2.19
Mr= 6.39754402240025 +- 0.7662628917031663
FeH= -1.9333472394579039 +- 0.3349617551726989
```



ALGORITHM:

- for given healpixel (from RA, Dec) get TRILEGAL simulated sample
- select stars with similar r band magnitudes (~0.5 mag) and construct prior map
- with given isochrones, construct likelihood map
- multiply likelihood and prior maps to get posterior pdf, and then marginalize to get 1-D posteriors for Mr and [Fe/H]
- demonstrably working!

(a lot of) Remaining work...

WORK:

- automate the production of maps with priors from TRILEGAL
- complete the isochrone library
- implement “unknown Ar” use case
- develop **better code:** fast and robust, with configuration and metadata management
- test, test, test!
- documentation
- papers

ALGORITHM:

- for given healpixel (from RA, Dec) get TRILEGAL simulated sample
- select stars with similar r band magnitudes (~0.5 mag) and construct prior map (Mr – [Fe/H])
- with given isochrones, construct likelihood map (Mr – [Fe/H])
- multiply likelihood and prior maps to get posterior pdf, and then marginalize to get 1-D posteriors for Mr and [Fe/H]
- demonstrably working!

Astr 598: “Astro-statistics and Machine Learning in Astronomy”

Topics:

Introduction to statistics (probability, distributions, robust statistics, Central Limit Theorem, hypothesis testing).

Maximum likelihood and applications in astronomy (point-spread-function photometry, astrometry)

Bayesian statistics and introduction to Markov Chain Monte Carlo

Model parameter estimation and model selection

Regression and Time series analysis

Dimensionality reduction

Density estimation and clustering

Supervised Classification

Class repository: https://github.com/dirac-institute/uw-astr598-w18

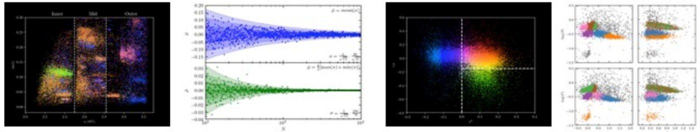
Astr 598: Astro-statistics and Machine Learning in Astronomy”

These Astr 598 topics follow this book:

All numerical examples from the book are fully reproducible. They rely on **astroML.org:**

astroML User Guide Examples Notebooks Book Figures Development

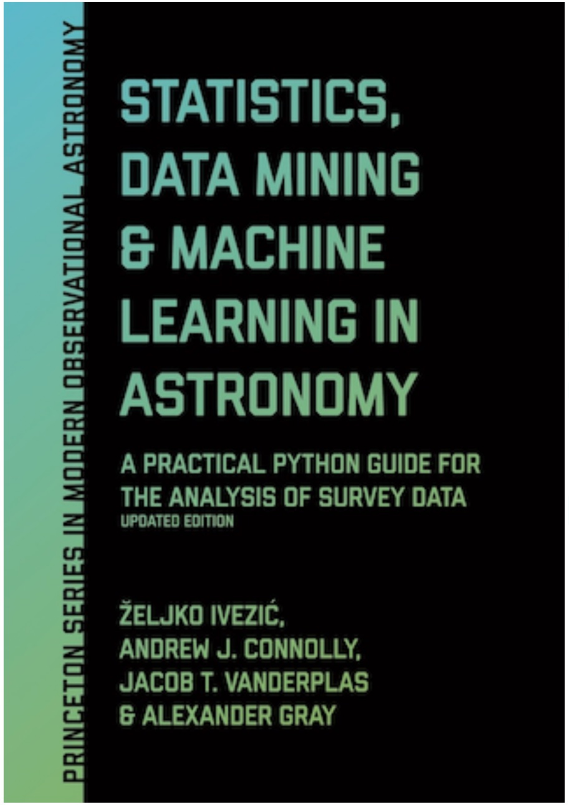
Search the docs ...



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.

Downloads

- Released Versions: Python Package Index
- Bleeding-edge Source: github



Example Notebooks

astroML

AstroML Interactive Book

Search the docs ...

AstroML Interactive Book

astroML is a Python module for machine learning and data mining that accompanies the book "Statistics, Data Mining, and Machine Learning in Astronomy", by Željko Ivezić, Andrew Connolly, Jacob Vanderplas, and Alex Gray. astroML is built on numpy, scipy, scikit-learn, matplotlib, and astropy, and contains a growing library of statistical and machine learning routines for analyzing astronomical data.

In this interactive book we provide notebooks that describe the statistical and machine learning methods used in astroML together with code that runs these methods on existing astronomical data sets. The structure of this interactive book follows the chapters of "Statistics, Data Mining, and Machine Learning in Astronomy". Each notebook can viewed through the browser (with navigation links at the side of the page), be downloaded to your own computer, or be executed directly using Binder or Google Colab

And now an astroML-based notebook with a few examples...

1) Model selection using Bayesian Information

2) Bayesian Blocks Algorithm

Notebook available as: ls.st/f23

full link:

https://github.com/ivezic/Notebooks/blob/master/Astr597A_astroMLexamples.ipynb