Statistics & Machine Learning in Astrophysics

Shantanu Desai
Department of Physics, IIT Hyderabad

Riken-AIP/IITH workshop on machine learning
March 15-16, 2019

Collaborators/Thanks

P.K. Srijith, Suryarao Bethapudi, Aisha Dantuluri, Anirudh Jain, Rahul Maroju, Sristi Ram Dyuthi, Anumandla Sukrutha, Soham Kulkarni, Aishwarya Bhave (NIT Raipur), Anirudh Jain (ISM Dhanbad), Tejas P., Shalini Ganguly, Ashwani Rajan (IIT Guwahati) Dawei Liu (Houston U.), Ben Hoyle (LMU), Markus Rau (CMU) Dark Energy Survey Collaboration, etc.

"One normally associates statistics with large numbers and astronomy is full of large on to believe that increased interaction between statistics and astronomy will be of benefit to both subjects"
J.V. Narlikar to C.R. Rao

Rise of Bayesian analysis in astro literature

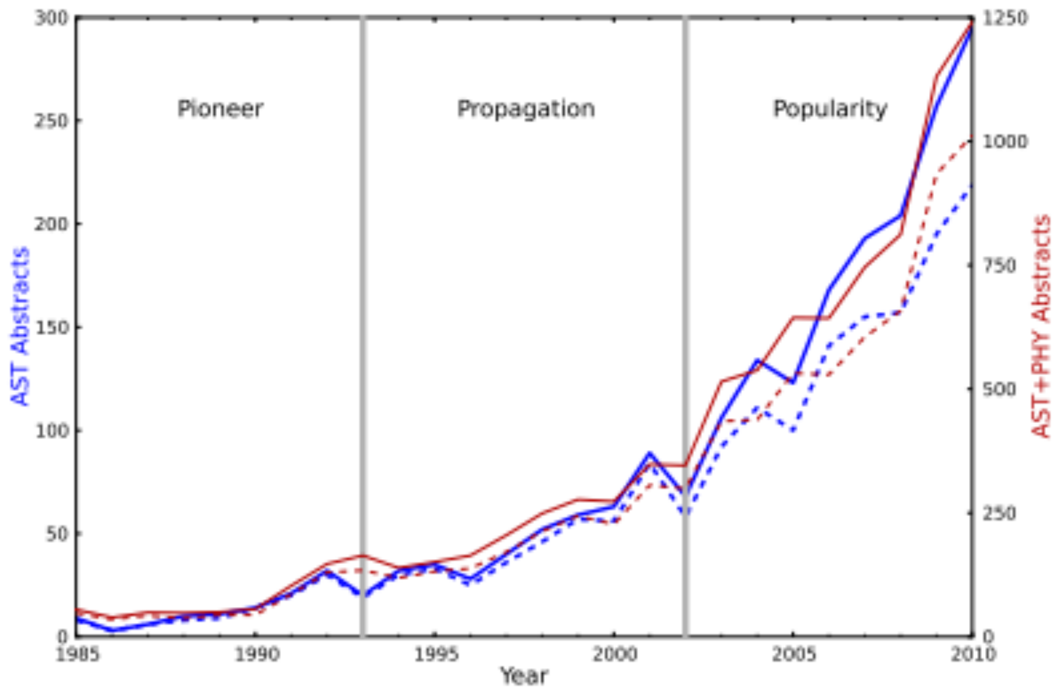


Fig. 1 Simple bibliometrics measuring the growing use of Bayesian methods in astronomy and physics, based on queries of the NASA ADS database in October 2011. Thick (blue) curves (against the left axis) are from queries of the astronomy database; thin (red) curves (against the right axis) are from joint queries of the astronomy and physics databases. For each case the dashed lower curve indicates the number of papers each year that include "Bayes" or "Bayesian" in the title or abstract. The upper curve is based on the same query, but also counting papers that use characteristically Bayesian terminology in the abstract (e.g., the phrase "posterior distribution" or the acronym "MCMC"); it is meant to capture Bayesian usage in areas where the methods are well-established, with the "Bayesian" appellation no longer deemed necessary or notable.

arXiv:1208.3036

Rise of MCMC analysis in astro literature

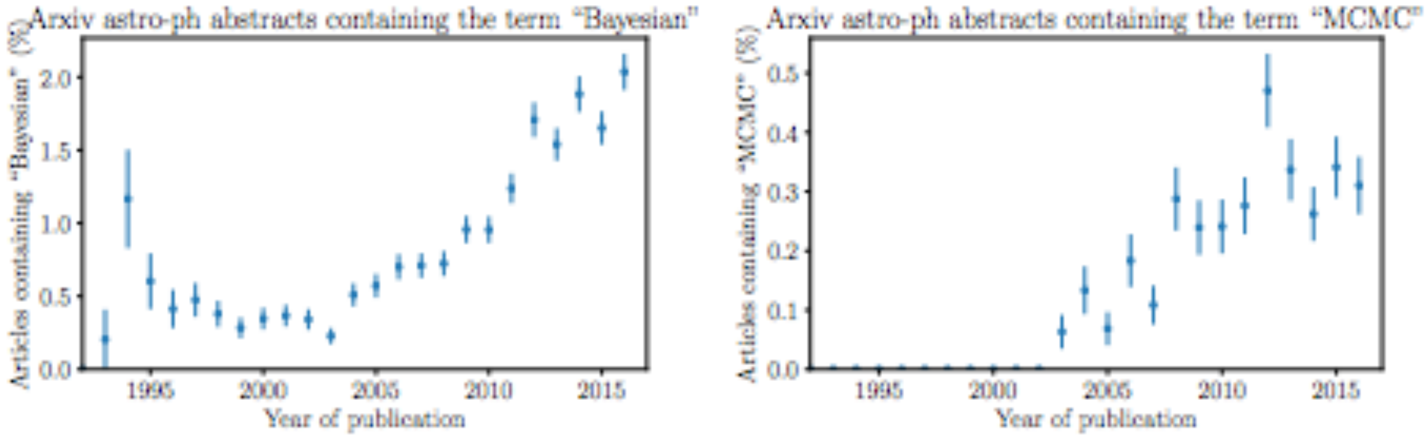


Figure 2

Percentage of articles in Arxiv astro-ph abstracts containing the word Bayesian (left) and MCMC (right). Computed using the code *arxiv.py*, courtesy Dustin Lang.

arXiv:1706.01629

Rise of ML analysis in astro literature

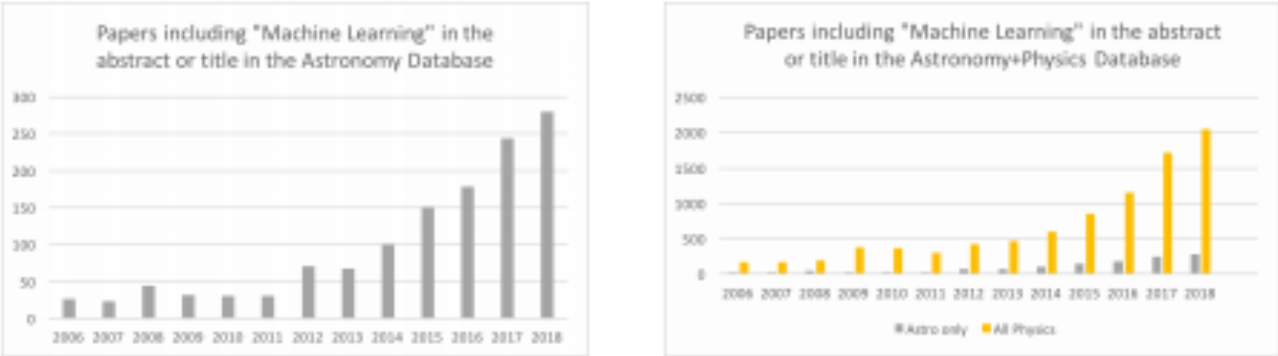


Figure 1. Machine learning related papers on the NASA/ADS archive from 2006 to today.

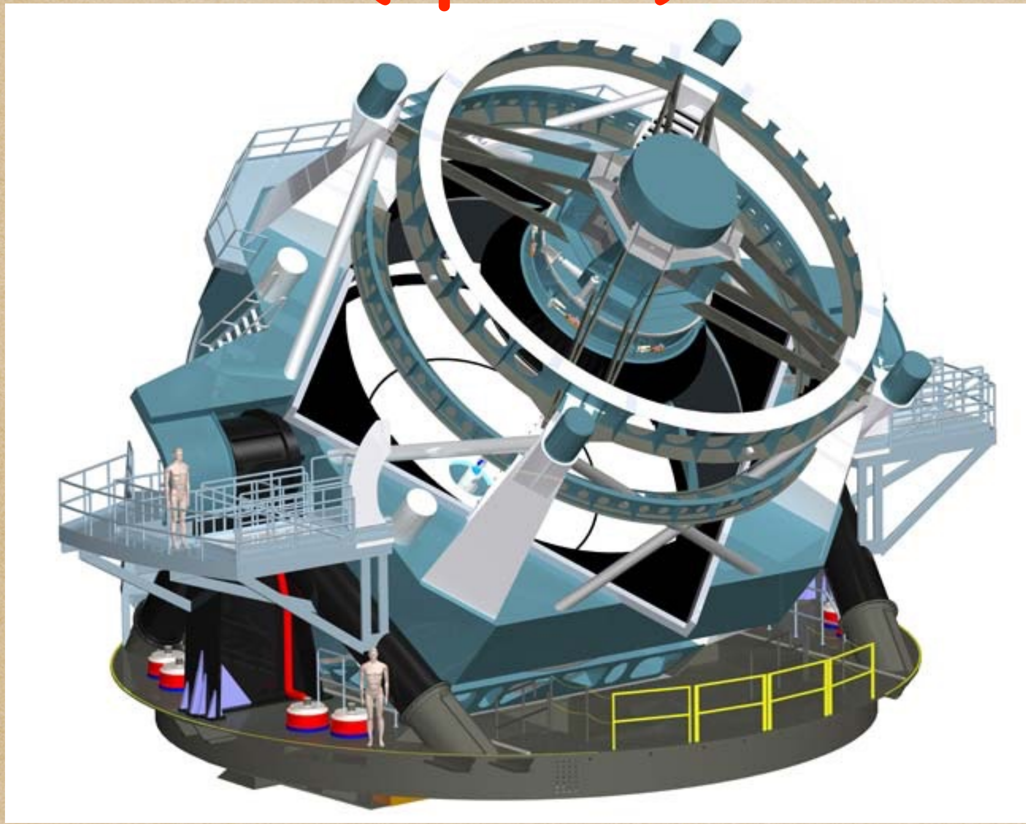
1

About 90 astro-ph papers with deep learning in abstract after 2014

arXiv:1901.05978

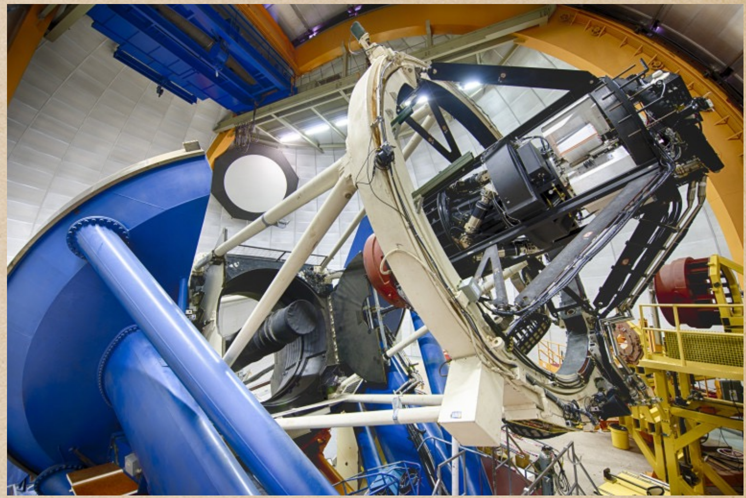
Astronomy in the by-gone era
~ 1930s

Where (optical) astronomy is today



Large Synoptic Survey Telescope (2022+)

Mirror Size ~ 8.4 m
Camera 3.2 Gigapixels
Field of view 10 sq. degrees
30 TB of raw data per night
Total raw data + catalogs ~ 100 PB!



Dark Energy Survey (2013-19)
mapped ~ 5000 sq deg of sky

Mirror Size ~ 4m
Camera 570 pixels
Field of view 3 sq deg.
Total data volume 2 PB!

Statistical tasks in Astrophysics

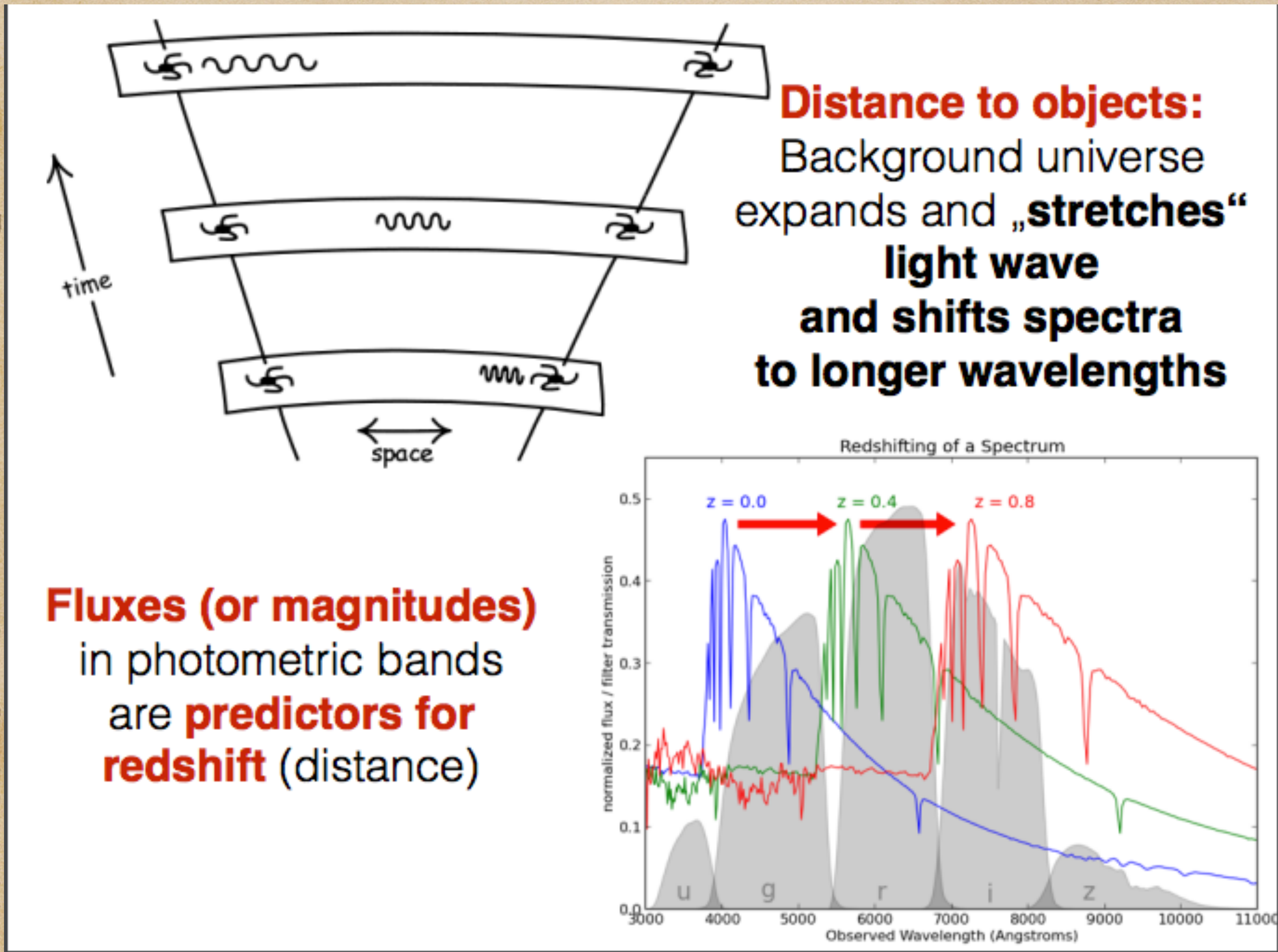
Photometric Redshifts (Regression)
Star/Galaxy classification
Source Classification
Dimensionality Reduction/Visualization
Clustering
N-point statistics
Dealing with censored and truncated data
Transient and Outlier Detection
Density Estimation
Matched Filtering
Source Extraction
(Fast) Cross-Matching
Data/Image Compression
Model Comparison
Forecasts using Fisher matrices
MCMC Methods and alternatives (for parameter estimation)
Nested Sampling Techniques
Cosmic Ray and other artefact removal from images
Time-Series and Time-frequency analysis
Looking for periodicities

Applications of Machine Learning to astronomy

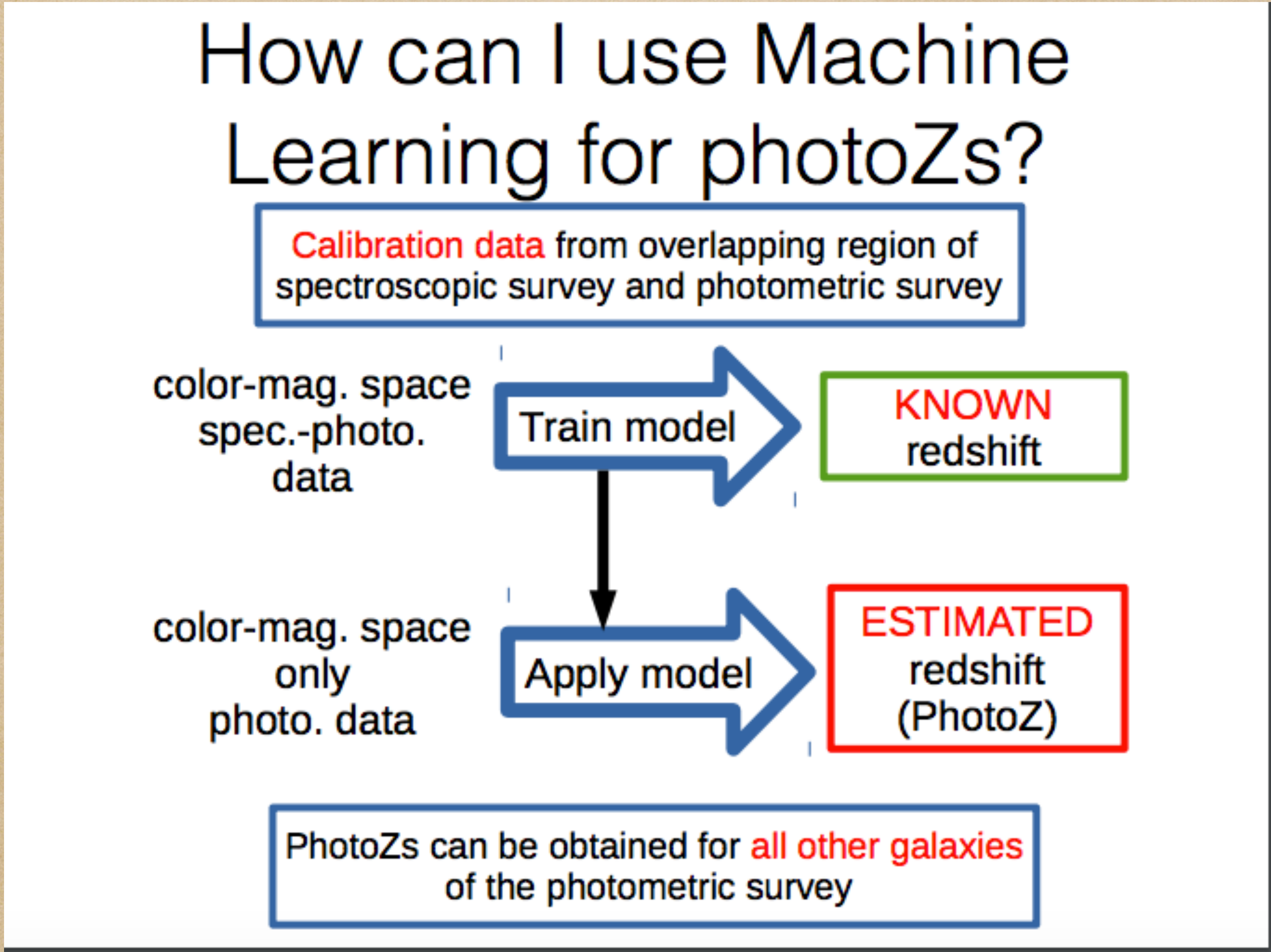
In the context of astronomy, ML is used to:

- Describe complicated relationships
- identify data clusters and data outliers
- reduce scatter by using complex or subtle signals
- generate simulated data
- classify objects
- address sparse data
- explore datasets to understand the physical underpinnings

Photometric Redshifts



Credit : Markus Rau



Credit : Markus Rau

Snapshot of problems in photoz involving ML

Mon. Not. R. Astron. Soc. **000**, 1–?? (2010) Printed 14 June 2018 (MN LaTeX style file v2.2)

Feature importance for machine learning redshifts applied to SDSS galaxies

Ben Hoyle[1,2], Markus Michael Rau[1], Roman Zitlau[1], Stella Seitz[1,3], Jochen Weller[1,2,3]

Stacking for machine learning redshifts applied to SDSS galaxies

Roman Zitlau[1], Ben Hoyle[1], Kerstin Paech[1], Jochen Weller[1,2,3]
Markus Michael Rau[1,3], Stella Seitz[1,3]

Measuring photometric redshifts using galaxy images and Deep Neural Networks

Ben Hoyle

Anomaly detection for machine learning redshifts applied to SDSS galaxies

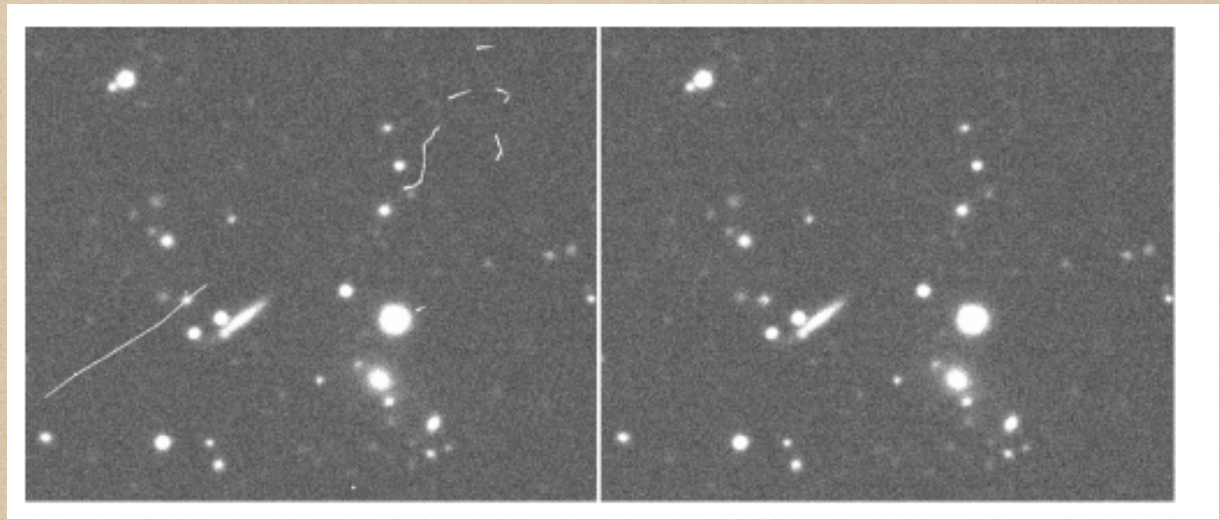
Ben Hoyle[1,2], Markus Michael Rau[1,4], Kerstin Paech[1,2], Christopher Bonnett[3]
Stella Seitz[1,4], Jochen Weller[1,2,4]

Deriving Photometric Redshifts using Fuzzy Templates and Self-Organizing Maps. II. Comparing Sampling Techniques Using Mock Data

Joshua S. Speagle[1,2*] and Daniel J. Eisenstein[1]
[1] *Harvard University Department of Astronomy, 60 Garden St., MS 46, Cambridge, MA 02138, USA*
[2] *Kavli IPMU (WPI), UTIAS, The University of Tokyo, Kashiwanoha 5-1-5, Kashiwa, Chiba, Japan*

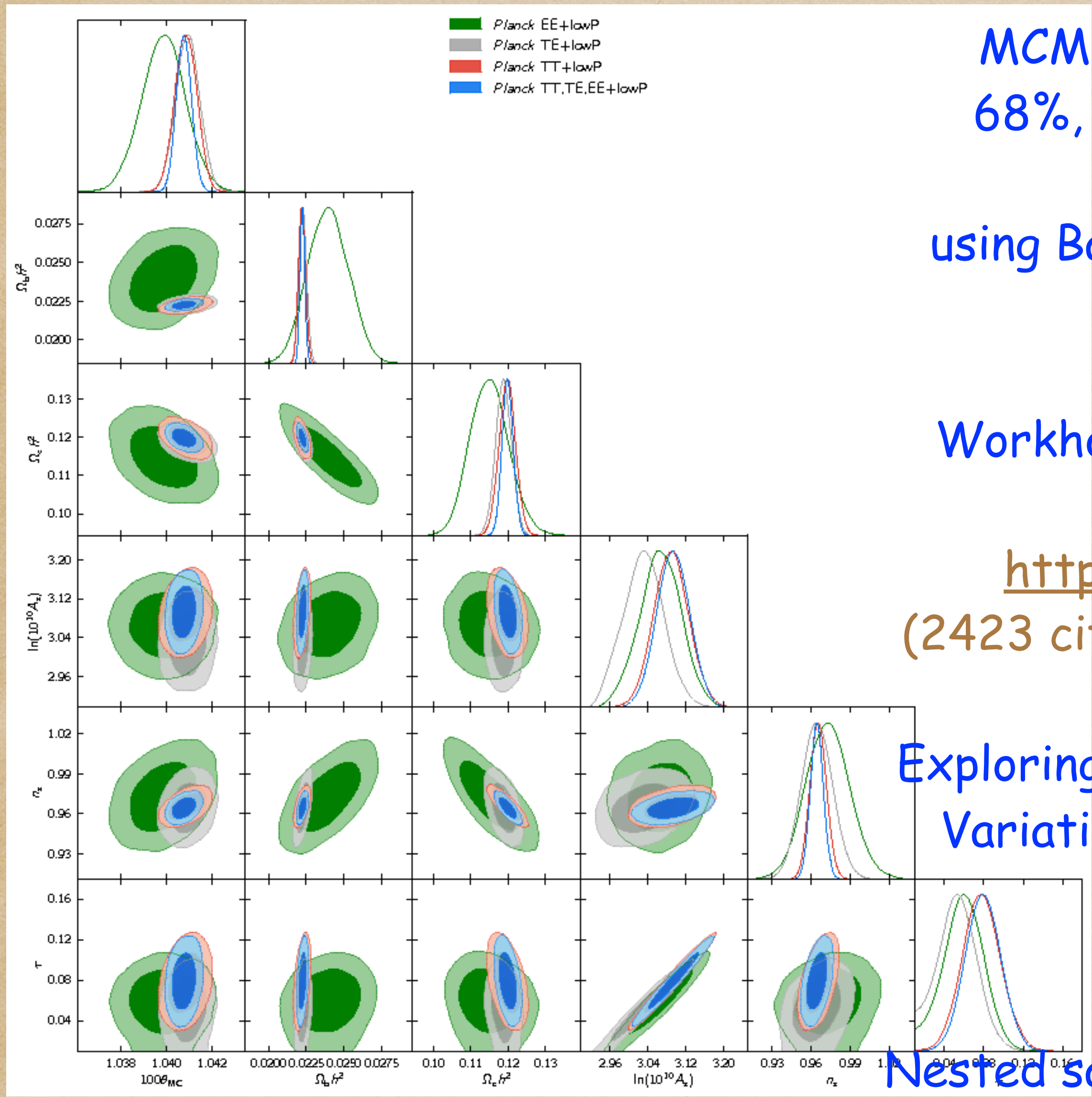
~400 papers on photo-z in title on arXiv!

Removal of cosmic rays/satellite trails from images



SD et al (arXiv:1601.07182)

Parameters estimation/Regression



MCMC methods used to calculate
68%, 90%, etc credible intervals
on various parameters
using Bayesian regression techniques.

Workhorse software used for MCMC
in Astronomy is emcee
http://dfm.io/emcee/current/
(2423 citations, including outside astro)

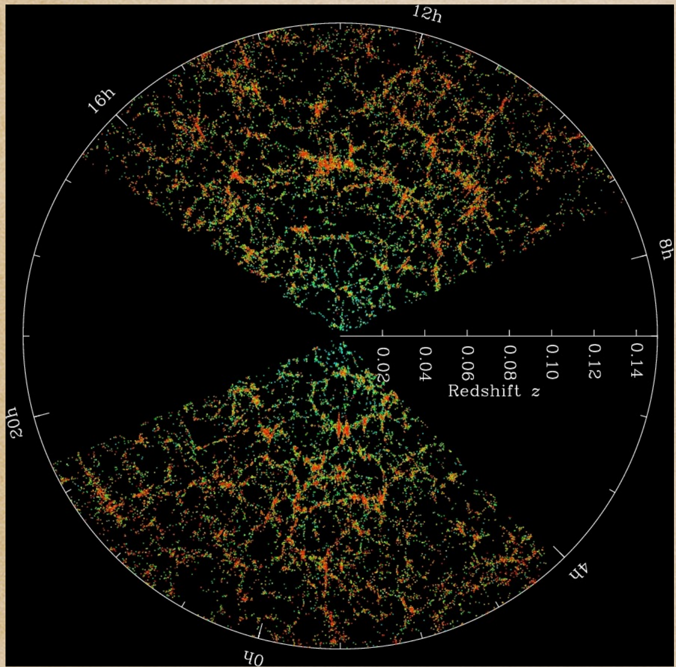
Exploring alternatives to MCMC, such as
Variational Inference for parameter
estimation

Jain, Srijith, SD, arXiv:1803.6473

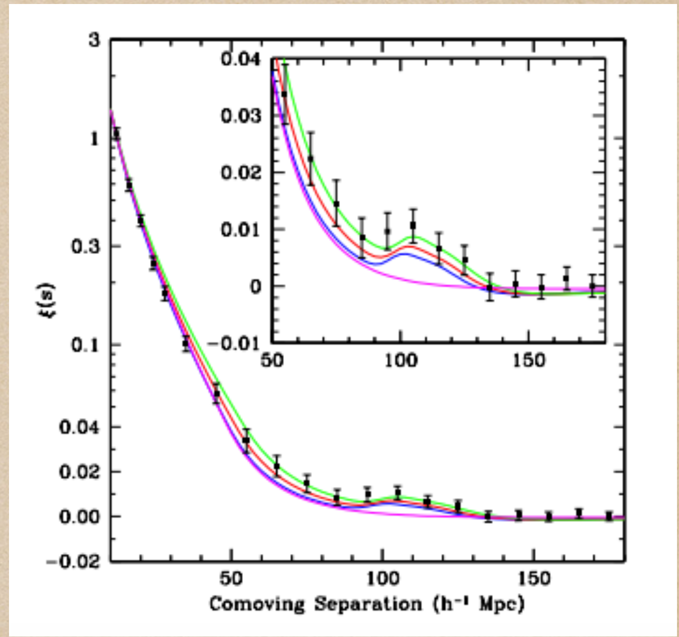
Nested sampling techniques for calculating
Bayesian evidence

Two-point (n-point) correlation functions

$$dP_{12} = \rho^2 dV_1 dV_2 (1 + \xi(r))$$ ξ (r) is called two-point correlation function



SDSS



Eisenstein et al (2005)

CUTE solutions for two-point correlation functions from large cosmological datasets

David Alonso[1]
[1] *Instituto de Física Teórica UAM-CSIC, Universidad Autónoma de Madrid, 28049 Cantoblanco, Spain*

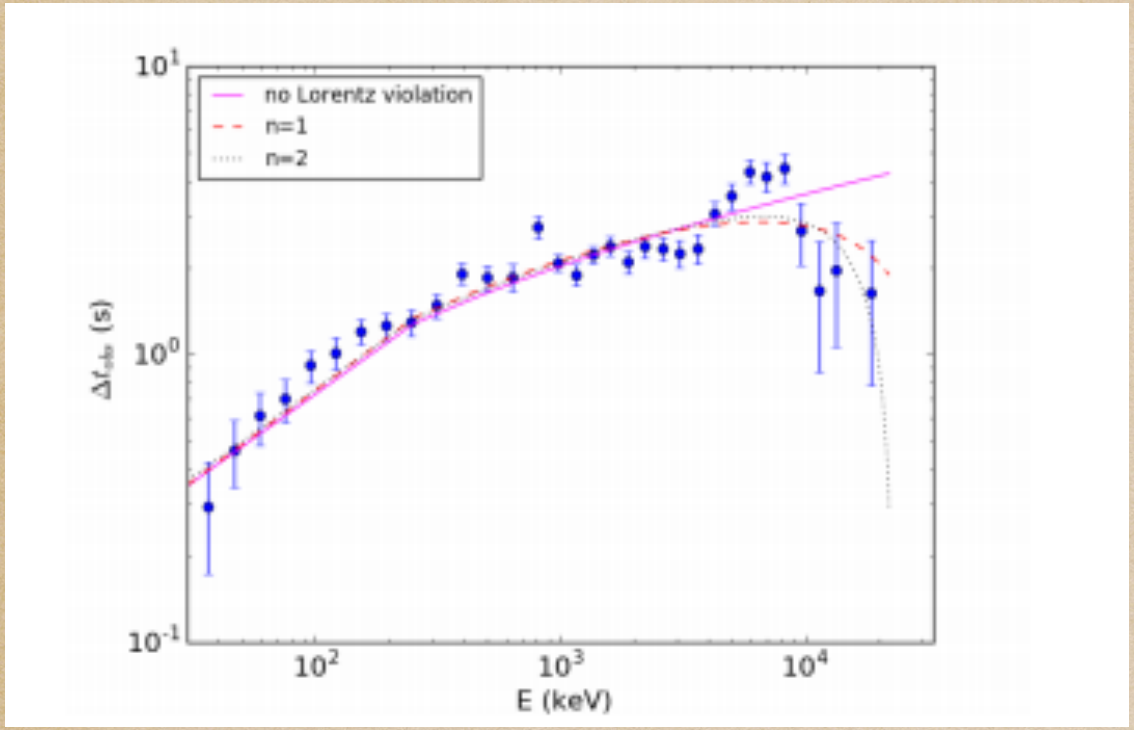
(Dated: June 21, 2013)

In the advent of new large galaxy surveys, which will produce enormous datasets with hundreds of millions of objects, new computational techniques are necessary in order to extract from them any two-point statistic, the computational time of which grows with the square of the number of objects to be correlated. Fortunately technology now provides multiple means to massively parallelize this problem. Here we present a free-source code specifically designed for this kind of calculations. Two implementations are provided: one for execution on shared-memory machines using OpenMP and one that runs on graphical processing units (GPUs) using CUDA. The code is available at `http://members.ift.uam-csic.es/dmonge/CUTE.html`.

GRAPH DATABASE SOLUTION FOR HIGHER ORDER SPATIAL STATISTICS IN THE ERA OF BIG DATA

Cristiano G. Sabiu,[1] Ben Hoyle,[2,3] Juhan Kim,[4] and Xiao-Dong Li[5]

Model Comparison

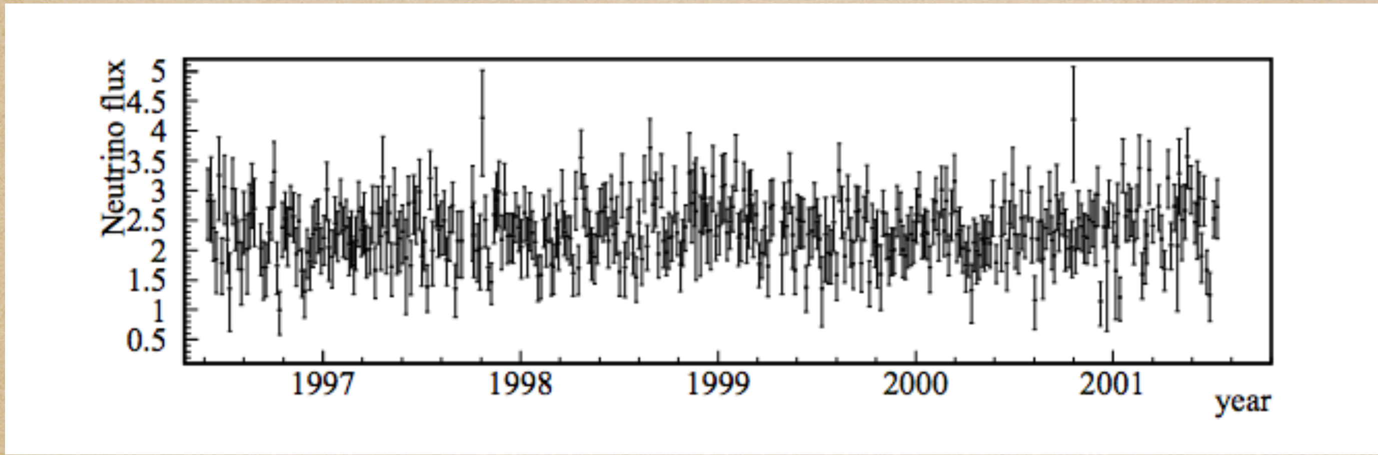


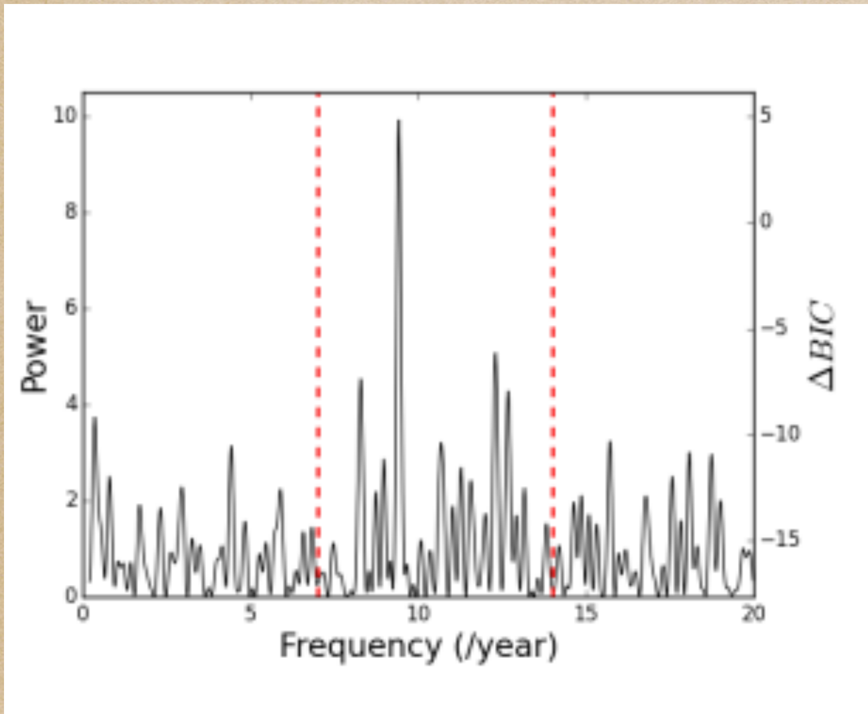
Shalini Ganguly, SD
arXiv:1706.01202

	No LIV[a]	(n=1)[b]	(n=2)[c]
Frequentist			
DOF	35	34	34
χ^2/DOF	2.6	2.37	2.23
χ^2GOF	2.2×10^{-7}	3.7×10^{-6}	1.5×10^{-5}
p-value		0.0014	9.2×10^{-5}
significance		3.05σ	3.74σ
Δ AIC		8.2	12.9
Δ BIC		6.9	11.7

[a] *No Lorentz Invariance*
[b] *Lorentz Invariance up to linear (n=1) order*
[c] *Lorentz Invariance up to quadratic (n=2) order*

Searching for periodicities in noisy unevenly sampled data.





Lomb-Scargle periodogram

D. Liu, SD (1604.06758)

Contribution of astrophysicists to Statistics

Extension of Gaussian mixture model to incorporate errors -> "Extreme Deconvolution"

The Annals of Applied Statistics
2011, Vol. 5, No. 2B, 1657–1677
DOI: 10.1214/10-AOAS439

EXTREME DECONVOLUTION: INFERRING COMPLETE DISTRIBUTION FUNCTIONS FROM NOISY, HETEROGENEOUS AND INCOMPLETE OBSERVATIONS

By Jo Bovy[1], David W. Hogg[1,2] and Sam T. Roweis[3]

New York University

EMPIRICISN: RE-SAMPLING OBSERVED SUPERNOVA/HOST GALAXY POPULATIONS USING AN XD GAUSSIAN MIXTURE MODEL

Thomas W.-S. Holoien[1,2,3,4], Philip J. Marshall[1,2], Risa H. Wechsler[1,2]

Draft version November 18, 2018

Incorrect use of statistics in Astronomy

How proper are Bayesian models in the astronomical literature?

Hyungsuk Tak,[1*] Sujit K. Ghosh,[2] and Justin A. Ellis[3]
[1] *Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, USA*
[2] *Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA*
[3] *Infinia ML, Durham, NC 27701, USA*

ABSTRACT
The well-known Bayes theorem assumes that a posterior distribution is a probability distribution. However, the posterior distribution may no longer be a probability distribution if an improper prior distribution (non-probability measure) such as an unbounded uniform prior is used. Improper priors are often used in the astronomical literature to reflect a lack of prior knowledge, but checking whether the resulting posterior is a probability distribution is sometimes neglected. It turns out that 23 articles out of 75 articles (30.7%) published online in two renowned astronomy journals (*ApJ* and *MNRAS*) between Jan 1, 2017 and Oct 15, 2017 make use of Bayesian analyses without rigorously establishing posterior propriety. A disturbing aspect is that a Gibbs-type Markov chain Monte Carlo (MCMC) method can produce a seemingly reasonable posterior sample even when the posterior is not a probability distribution (Hobert and Casella 1996). In such cases, researchers may erroneously make probabilistic inferences without noticing that the MCMC sample is from a non-existing probability distribution. We review why checking posterior propriety is fundamental in Bayesian analyses, and discuss how to set up scientifically motivated proper priors.

Key words: Markov chain Monte Carlo (MCMC) – improper flat prior – vague prior – uniform prior – inverse gamma prior – non-informative prior – scientifically motivated prior

arXiv:1712.03549

Table 1. Classification of 75 articles published online in *ApJ* and *MNRAS* between Jan 1, 2017 and Oct 15, 2017 according to their prior distributions.

	ApJ	*MNRAS*
(a) Jointly proper priors	18	34
(b) Jointly improper priors	1	2
(c) Unclear priors	11	9
Total	30	45

Training/educational/collaborative resources in statistics for astrophysicists

- Improved access to statistical software. R/CRAN public-domain statistical software environment with thousands of functions. Increasing capability in Python.

- Papers in astronomical literature doubled to ~500/yr in past decade ("Methods: statistical" papers in *NASA-Smithsonian Astrophysics Data System*)

- Short training courses (Penn State, India, Brazil, Greece, China, Italy, France, Germany, Spain, Sweden, LSST, IAU/AAS/CASCA/… meetings)

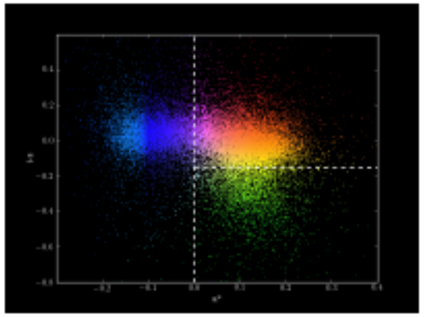
- Cross-disciplinary research collaborations (Harvard/ICHASC, Carnegie-Mellon, Penn State, NASA-Ames/Stanford, CEA-Saclay/Stanford, Cornell, UC-Berkeley, Michigan, Imperial College London, Swinburne, Texas A&M, JPL, LANL, …)

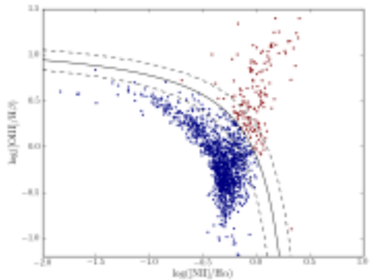
- Cross-disciplinary conferences (*Statistical Challenges in Modern Astronomy, Astronomical Data Analysis 1991-2016, PhysStat*, SAMSI 2006/2012, *Astroinformatics 2012-16, IAU Symposia 2014--, IEEE Symposia 2018--*)

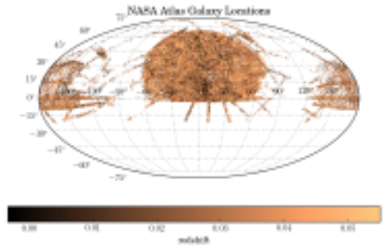
- Scholarly society working groups and a new integrated Web portal http://asaip.psu.edu serving: Int'l Stat Institute's Int'l Astrostatistical Assn, Int'l Astro Union Working Group (Commission), Amer Astro Soc Working Group, Amer Stat Assn Interest Group, LSST Science Collaboration, IEEE Astro Data Miner Task Force)

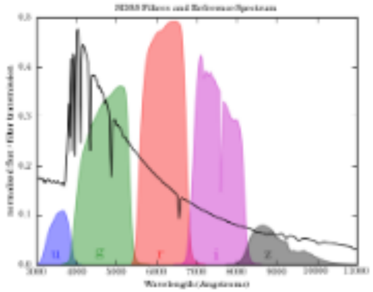
Credit : Eric Feigelson opening lecture at Penn State astrostatistics school

AstroML: Machine Learning and Data Mining for Astronomy









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

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. If you have an example you'd like to share, we are happy to accept a contribution via a GitHub Pull Request: the code repository can be found at http://github.com/astroML/astroML.

astroml.org

Conclusions

- Lot of synergy between Machine Learning, data mining, advanced statistical tools and Astrophysics
- Contact me or Srijith for more details.
- Compilation of interesting astrostatistics/ astroinformatics papers in goo.gl/4FY9qg

Thank you for your attention!!!