Astrostatistics

Part III Maths & Astrophysics

CMS MR5, Mon, Wed, Fri 12pm

Dr Kaisey S Mandel
University Lecturer on Astrostatistics
Institute of Astronomy
(Kavli Institute for Cosmology, Office 03)
& Statistical Laboratory
Department of Pure Mathematics & Mathematical Statistics
(CMS Pavilion D, Office 1.07)

Office Hours: Tuesdays @ ?pm D.107 or by appointment

kmandel@statslab.cam.ac.uk

Course Website:
https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2020

Lecture 01: 17 January 2020

What is Astrostatistics?

- The application of statistics to analyse data in astronomy, astrophysics & cosmology
- A research field: the interdisciplinary intersection of astronomy & statistics
- How do we properly interpret and analyse increasingly large and complex astronomical datasets?
- Developing and applying advanced statistical and computational methods to meet the unique challenges of astronomical data
- There is no “theory” of astrostatistics, the field is application-driven; we will focus on “real-world” case studies

Scope & Goals

- Learn to think about statistics as more than just a “bag of tricks”: with an ad-hoc recipe to blindly run for each particular data analysis problem in astronomy
- It will be impossible to address every potential statistical task in astronomy in 8 weeks
- Instead, we will focus on general principles to help you think about how to analyse data in your specific cases.
- Where do the data come from? What is the model? What are the (implicit or explicit) assumptions? Are they reasonable?
- Be pragmatic, and very applied. Examine real applications.

Who typically takes this class?

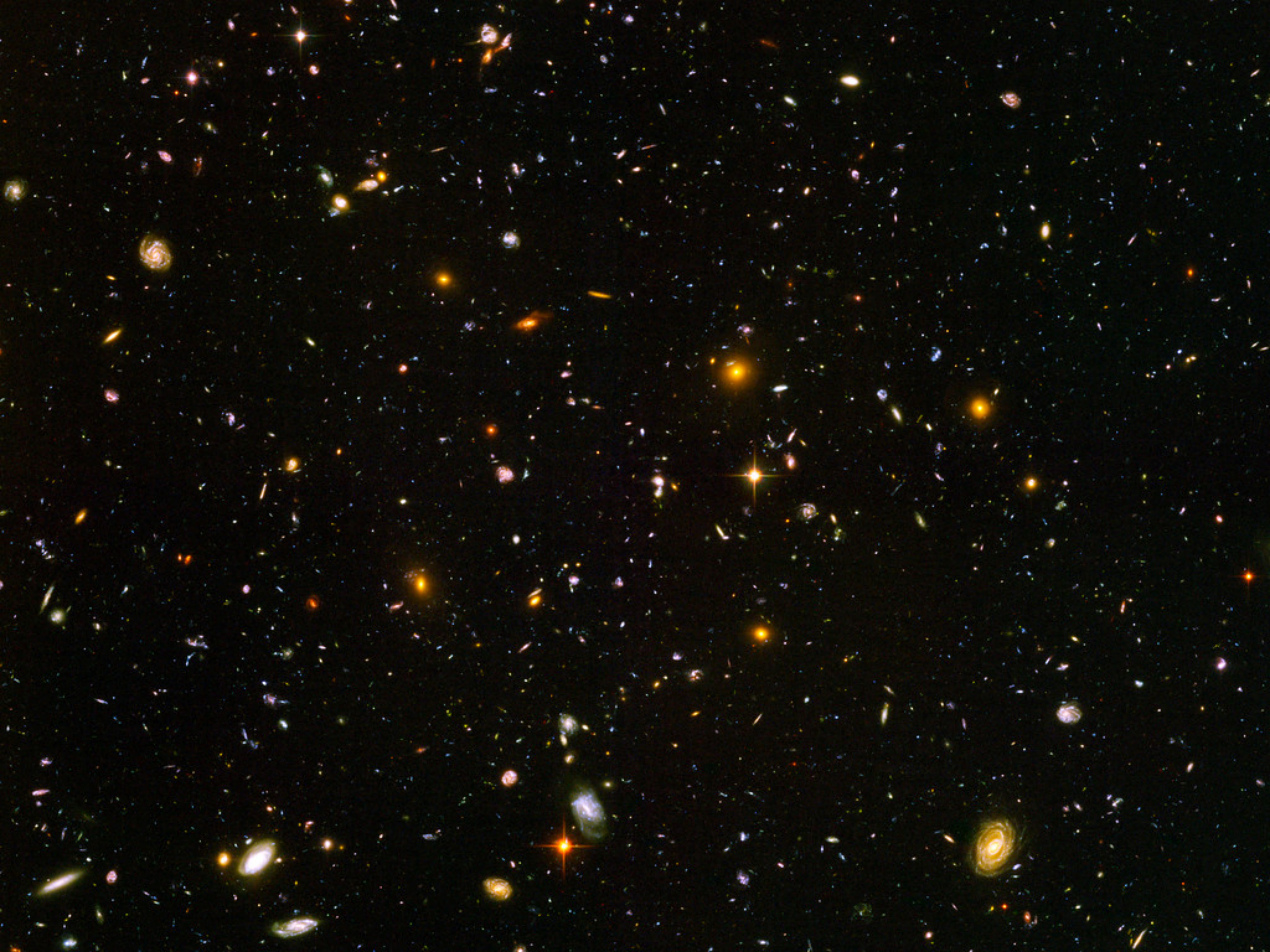
- Inherently an ***interdisciplinary*** course
- Part III Astrophysics students
- Part III Maths students primarily focusing on Statistics
- Part III Maths students primarily focusing on theoretical physics & cosmology
- PhD students in DAMTP, Physics, Astroonmy
    - CDT in Data-Intensive Science
- All are welcome!

For Astrophysicists

- Goal is to help you think critically about your data, rather than blindly applying canned black-box methods
- Understanding your statistical methods is crucial to interpreting their results. When can they go wrong?
- Often your data may be uniquely complex, and may require you to develop a data analysis method optimally suited to your inference problem - this is research!
- Astrostatistics is a creative endeavour!
- Get Jobs!
  - data-intensive astronomy
  - data-science / AI industry

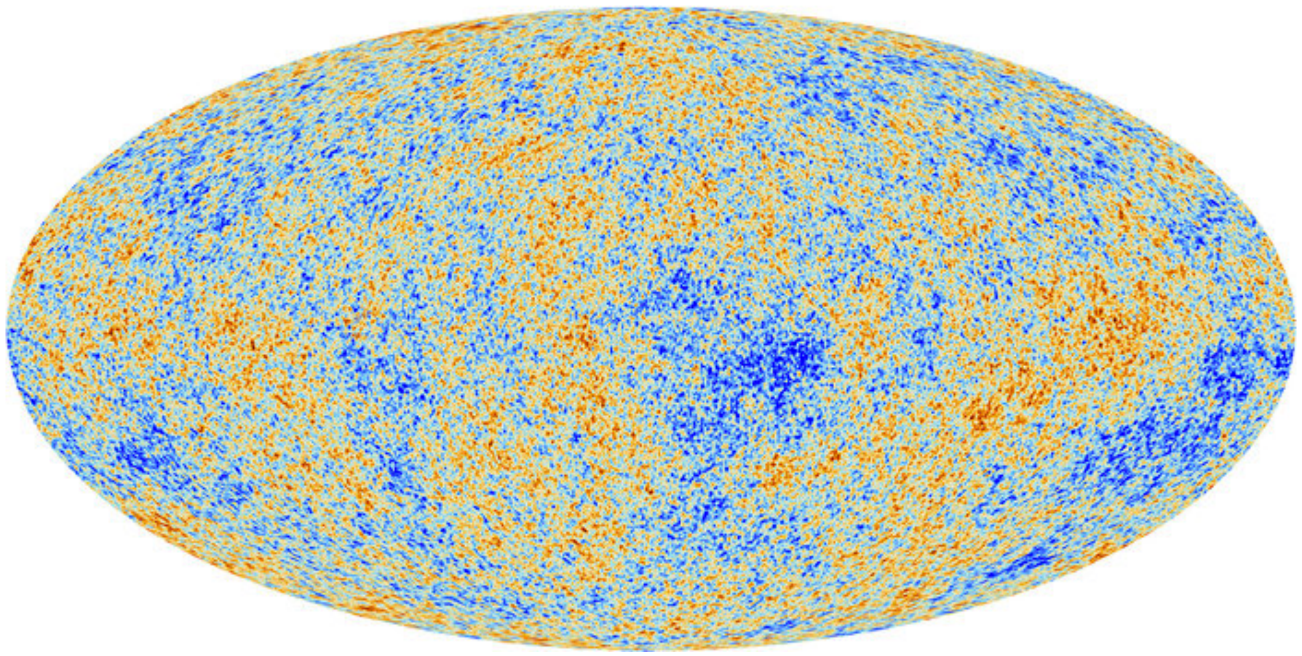
For Statisticians

- Astronomers have complex data sets with unique and challenging inference problems
- Astronomy is an observational science - (usually) no lab experiments
- Cutting-edge is always pushing the limits in the low signal-to-noise regime, where you need statistics to extract the best information
- Measurement errors, selection effects/biased samples, small sample size (One Universe), Big Data (billions of galaxies), combining heterogenous datasets over multiple wavelengths, EM/Gravitational Waves, multi-messenger astronomy
- Billions of $/£/Eur spent on Space Missions, how to get the most scientific value out of them? e.g. LSST, Euclid, WFIRST, TESS, …
- Optimal use of data requires developing, applying, and understanding new statistical approaches
- Play a critical role in answering Deep Questions about the Universe!
- Astronomy Jargon for Statisticians: http://hea-www.harvard.edu/AstroStat/astrojargon.html



Cosmology

Planck CMB



Type Ia Supernovae

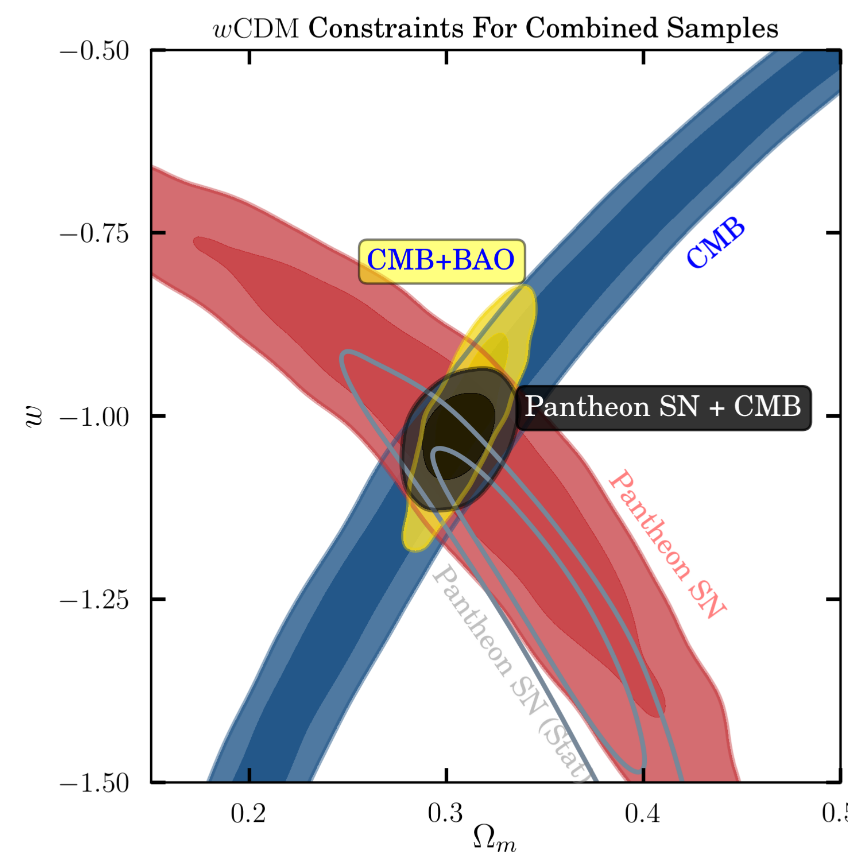
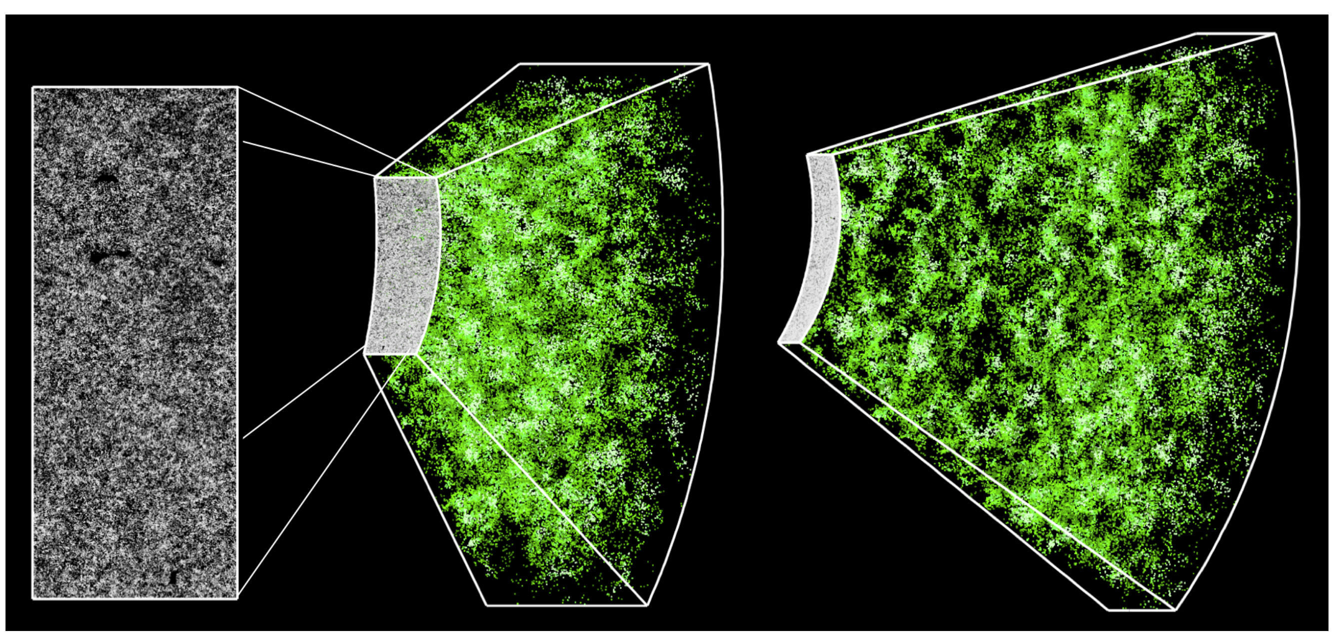
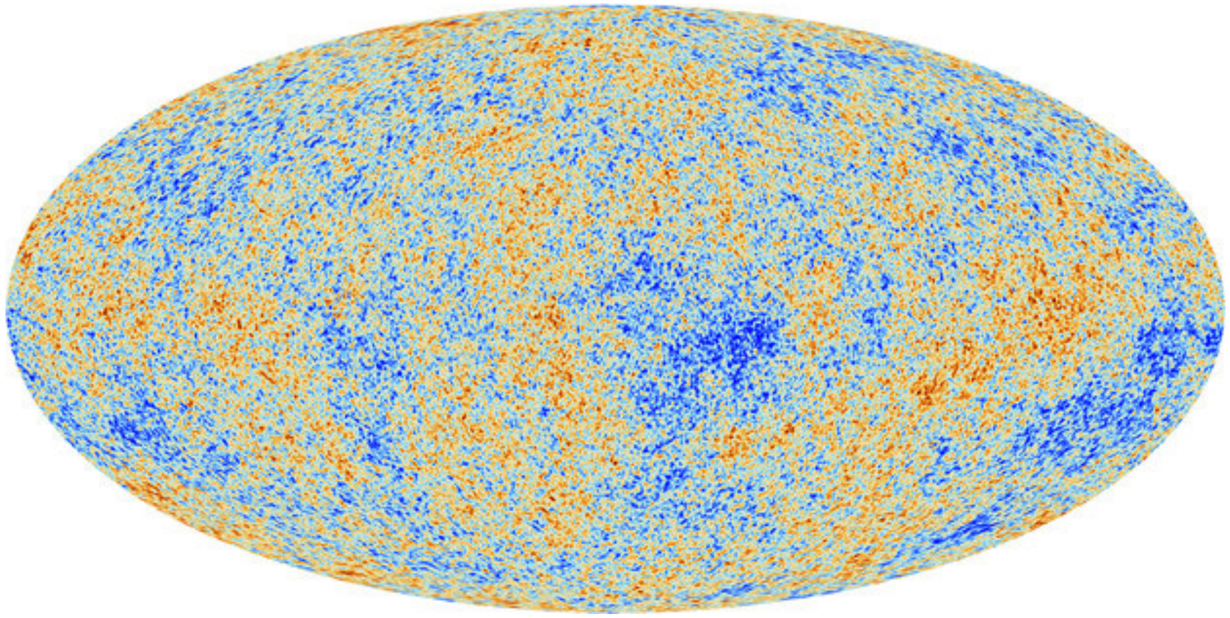


Figure 20. Confidence contours at 68% and 95% for the Ω_m and w cosmological parameters for the wCDM model. Constraints from CMB (blue), SN - with systematic uncertainties (red), SN - with only statistical uncertainties (gray-line), and SN+CMB (purple) are shown.

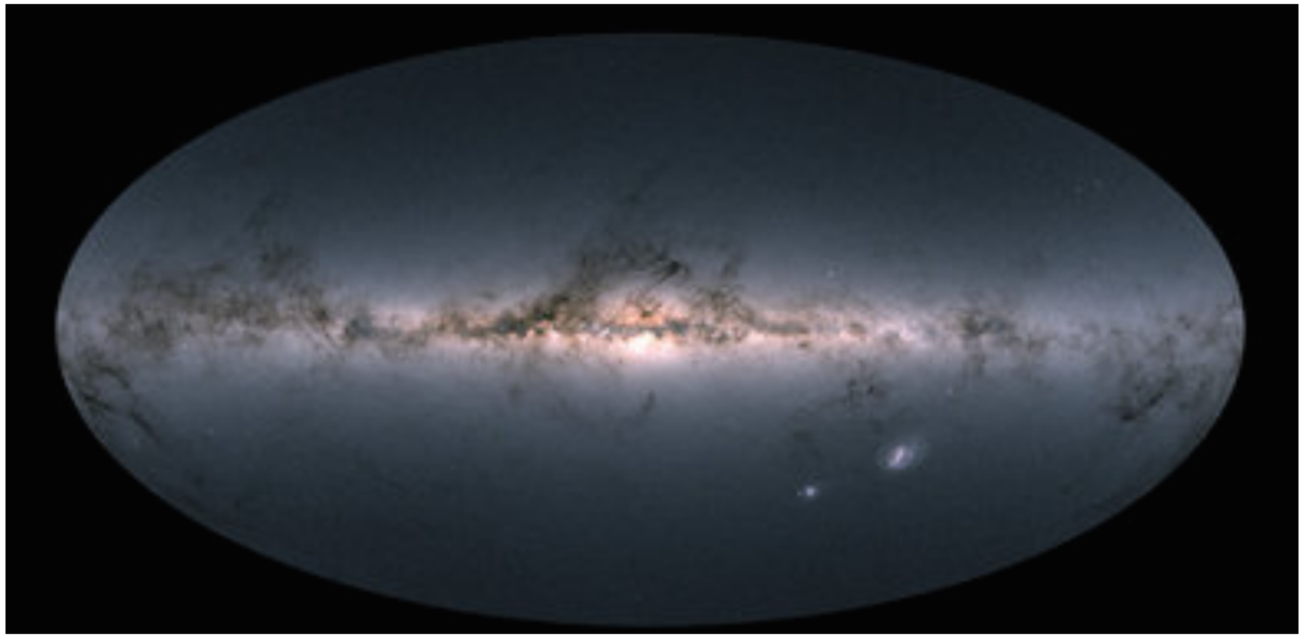


SDSS Baryonic Acoustic Oscillations

Data-Intensive Science in Astronomy: Major Experiments, Satellites & Surveys

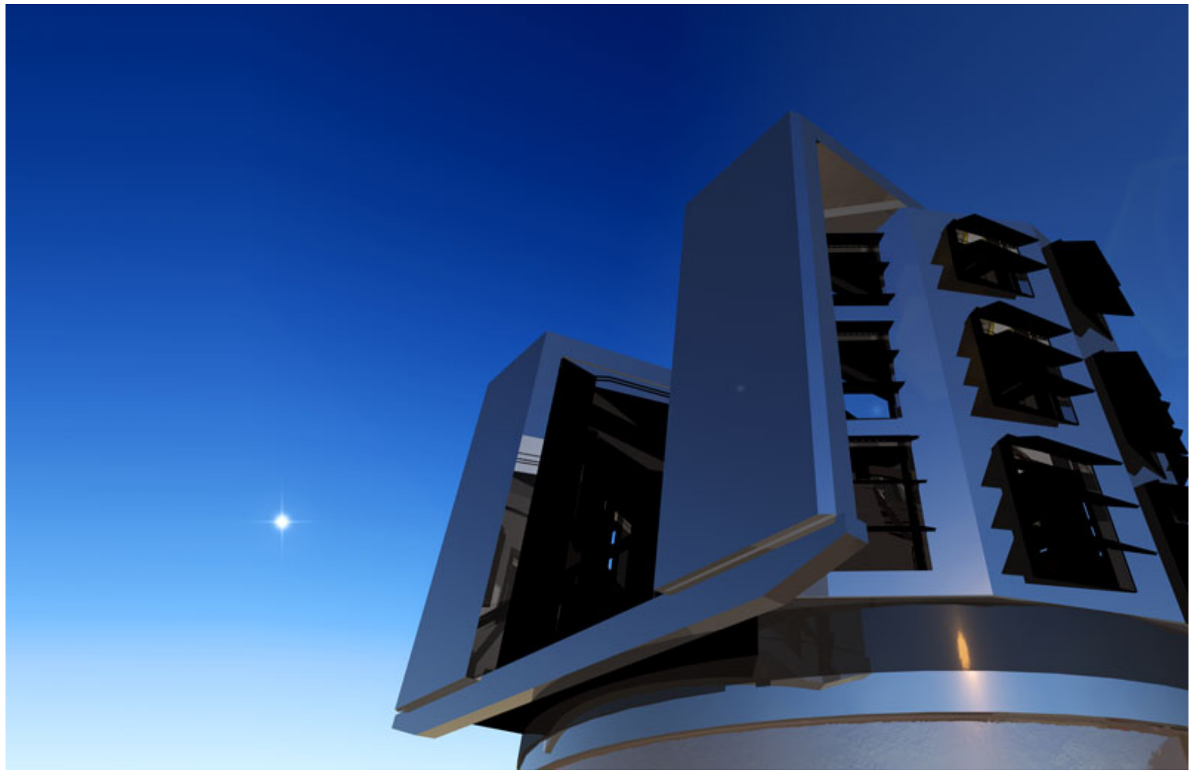


Planck (Cosmic Microwave Background)



Gaia (Milky Way Galaxy)

Square Kilometer Array



Large Synoptic Survey Telescope

Extrasolar Planets

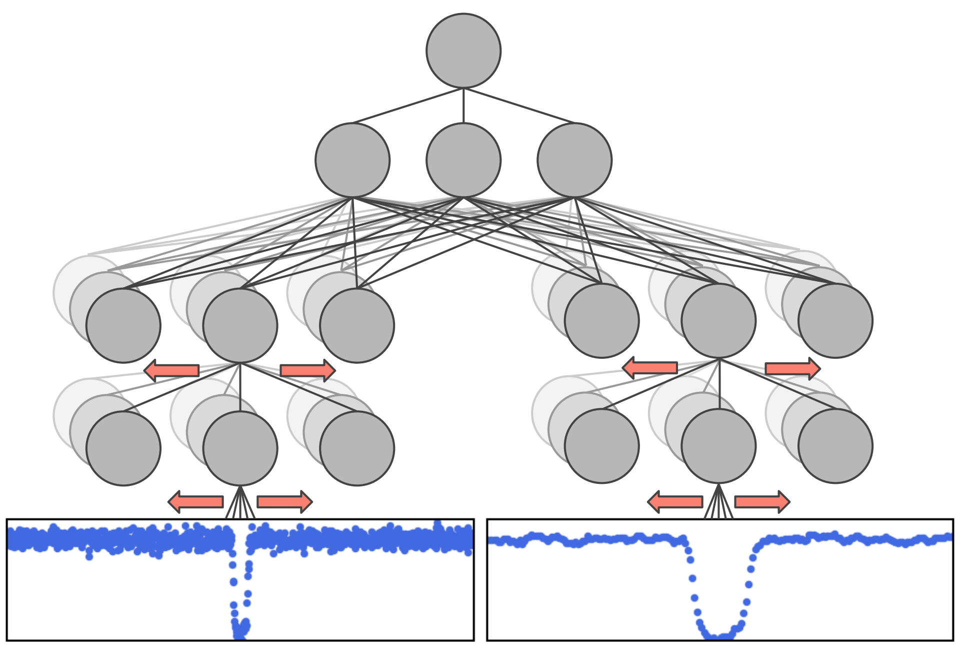


Fig. 5.— Convolutional neural network architecture for classifying light curves, with both global and local input views.

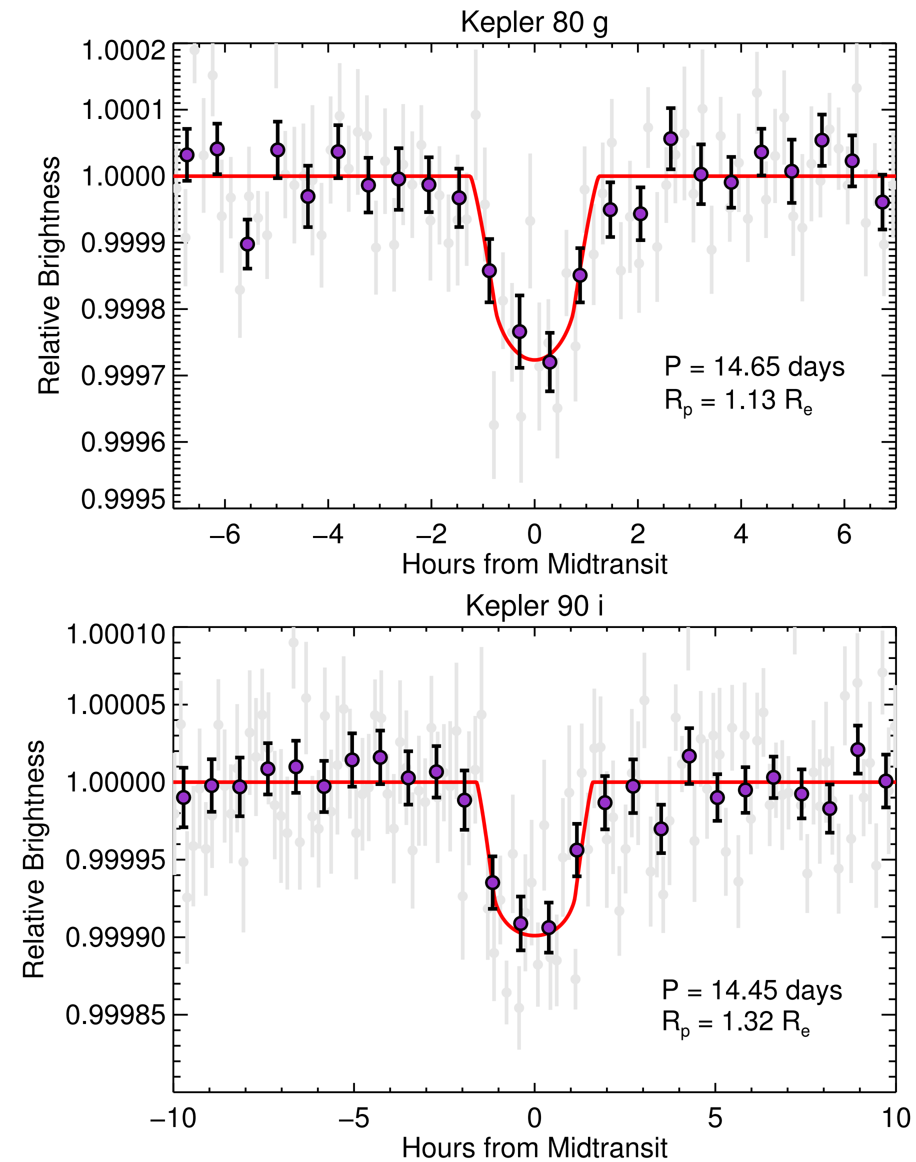
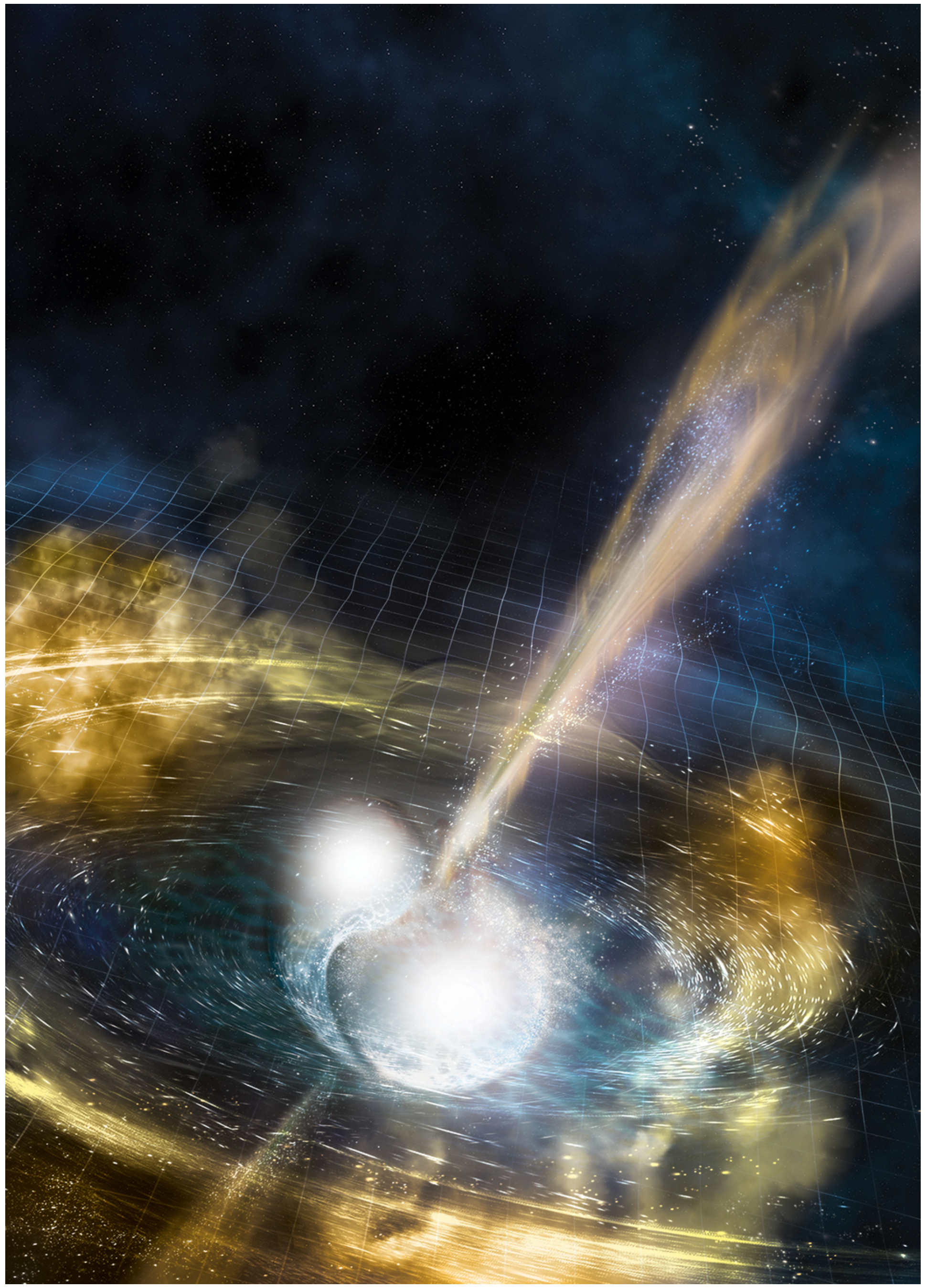


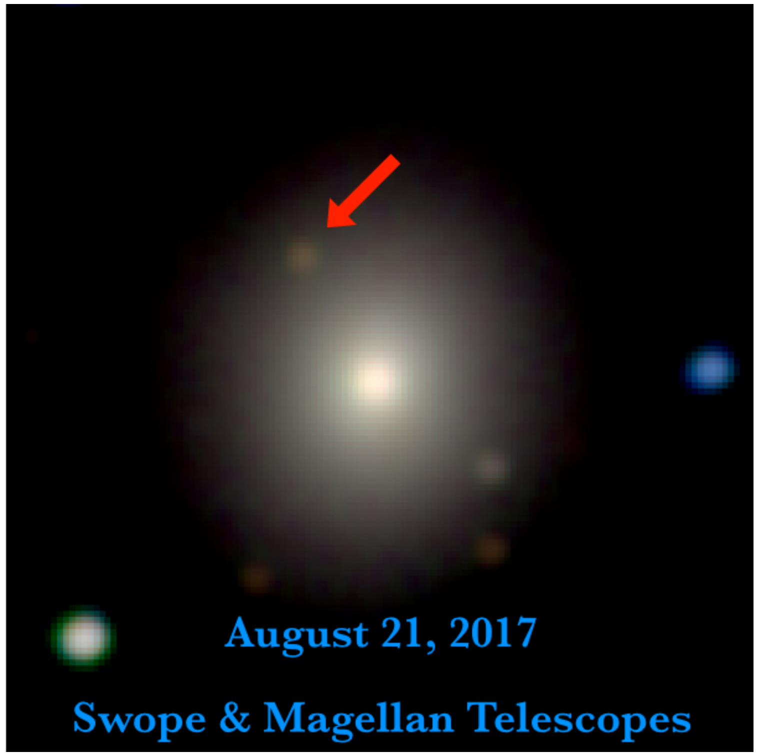
Fig. 12.— Transit light curves and best-fit models for the newly discovered planets around *Kepler*-80 and *Kepler*-90. In these plots, the grey points are robust averages of bins with width of approximately 10 minutes. The purple points are robust averages of bins with size about 1/4 the transit duration of the planet (bins of about 30 minutes for *Kepler*-80 g and about 45 minutes for *Kepler*-90 i.

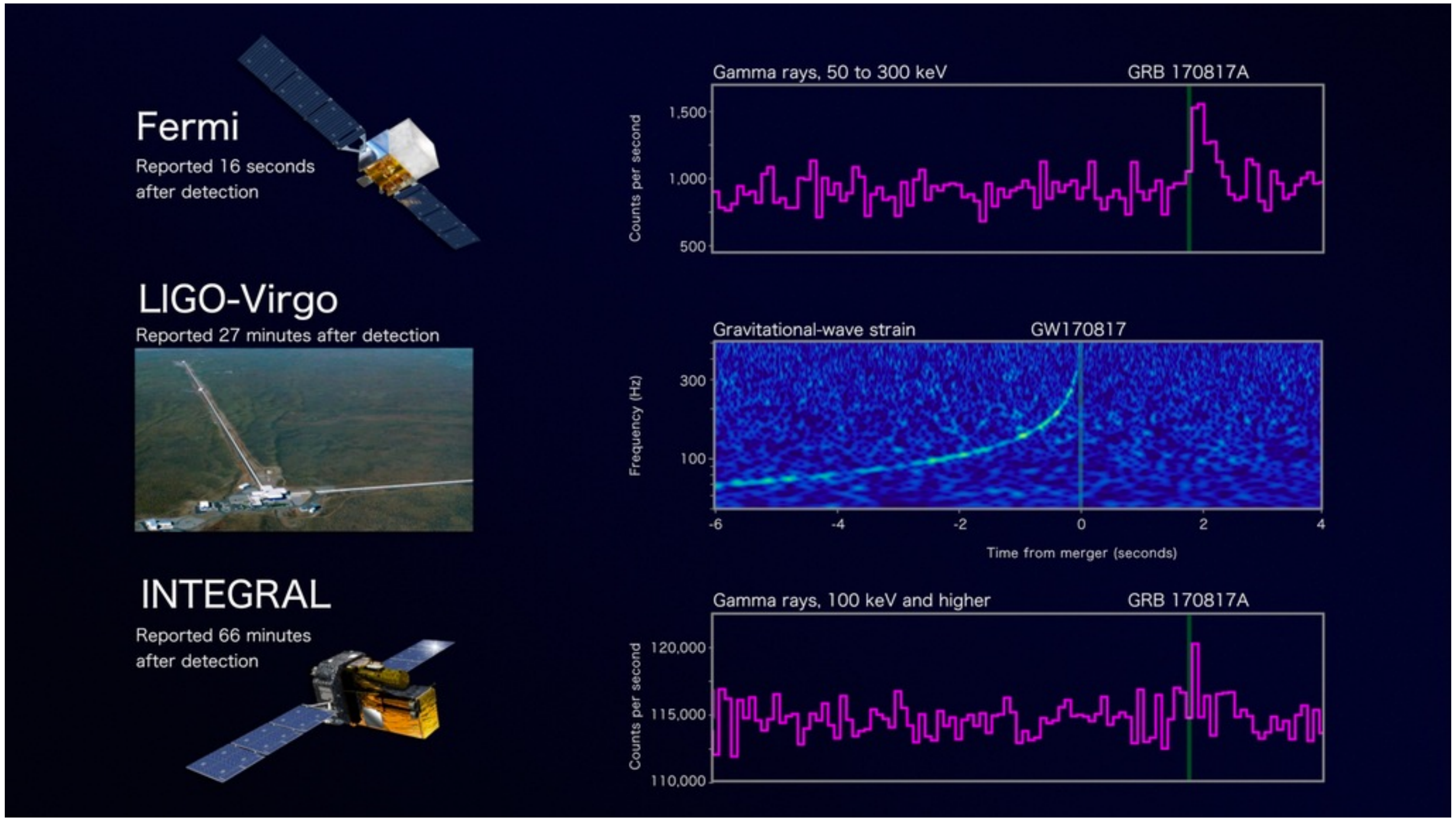
Deep Learning to Find New Exoplanets
(Shallue & Vanderberg 2017)

Gravitational Waves and Transients (Stellar Explosions)



Colliding Neutron Stars!





Astrostatistics (L24)

Kaisey Mandel

This course will cover applied statistical methods necessary to properly interpret today's increasingly complex datasets in astronomy. Particular emphasis will be placed on principled statistical modeling of astrophysical data and statistical computation of inferences of scientific interest. Statistical techniques, such as Bayesian inference, sampling methods, hierarchical models, Gaussian processes, and model selection, will be examined in the context of applications to modern astronomical data analysis. Topics and examples will be motivated by case studies across astrophysics and cosmology.

Pre-requisites

Students of astrophysics, physics, statistics or mathematics are welcome. Astronomical context will be provided when necessary. Students without a previous statistics background should familiarise themselves with the material in Feigelson & Babu, Chapters 1-4, and Ivezić, Chapters 1, 3-5, by the beginning of the course. (Note that the two textbooks cover many of the same topics). These texts are freely available online to Cambridge students via the library website.

Literature

1. E. Feigelson and G. Babu. *Modern statistical methods for astronomy: with R applications.* Cambridge University Press, 2012.

2. Z. Ivezić, A. Connolly, J. VanderPlas & A. Gray. *Statistics, Data Mining, and Machine Learning in Astronomy.* Princeton University Press, 2014.

3. C. Schafer. *A Framwork for Statistical Inference in Astrophysics.* 2015. Annual Review of Statistics and Its Application, 2: 141-162.

4. C. Bishop. *Pattern Recognition & Machine Learning.* Springer-Verlag, 2006. Also available at

   https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book

5. D. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. Also available at

   http://www.inference.org.uk/mackay/itila/book.html

Additional support

Four examples sheets will be provided and four associated examples classes will be given. There will be a one-hour revision class in the Easter Term.

Example sheets and Exam: What to expect

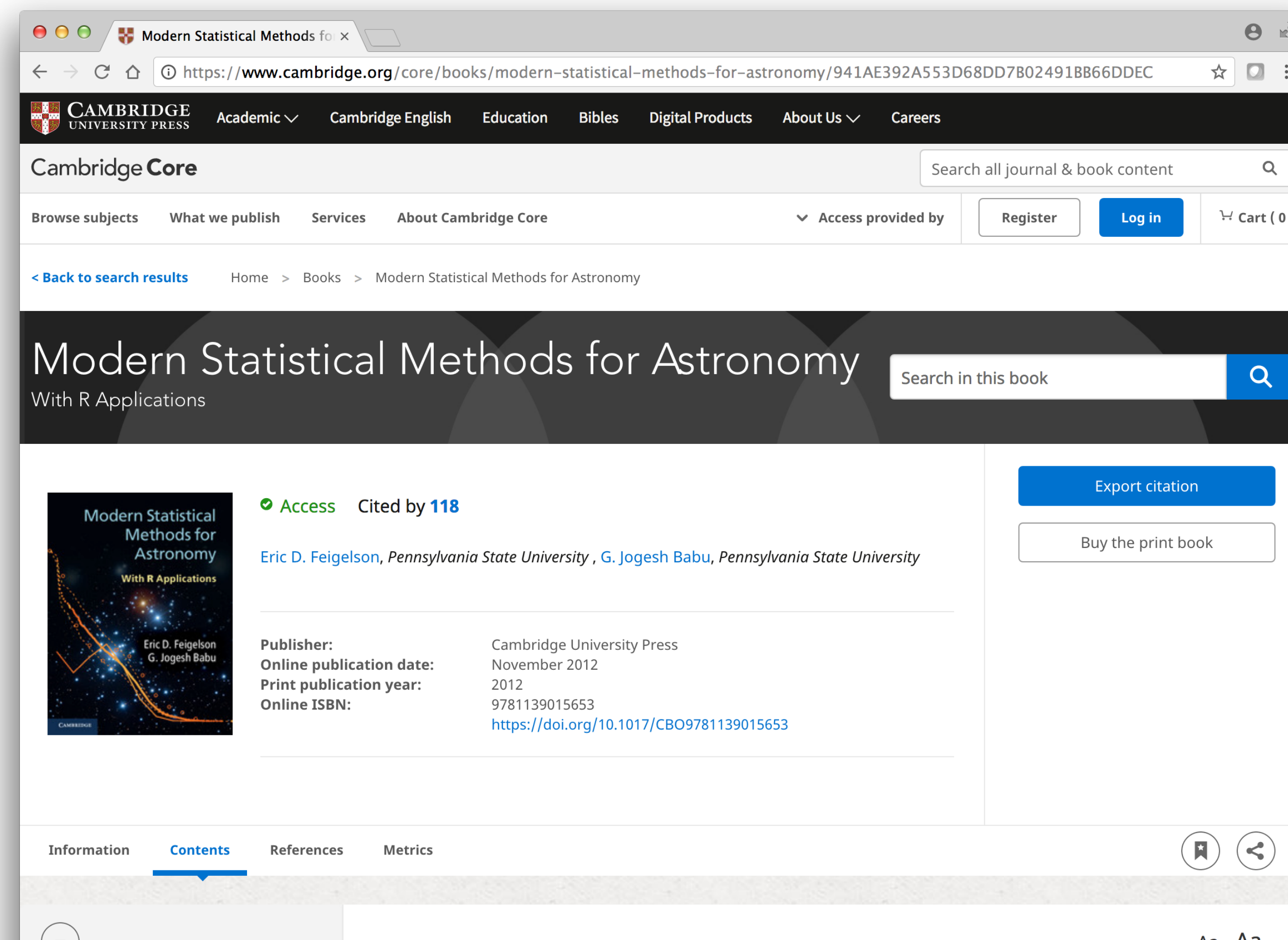
- Example sheets will comprise
  - analytic statistical modelling & derivations / proofs
  - practical implementation of algorithms and data analysis in code (e.g. Python / Matlab)
- Final exam : entirely written (no computer)
  - Analytic modelling, derivations, proofs on paper
  - How would you implement this algorithm or data analysis?

Astrostatistics Texts

E. Feigelson and G. Babu. **Modern statistical methods for astronomy: with R applications**. CUP, 2012.
[Free Cambridge access: search at
https://www.cambridge.org/core/ or
http://idiscover.lib.cam.ac.uk/ or go to the library]

An overview of statistical
Methods for astronomers.
R code available

Recommended Reading:
Chapters 1-4
Intro to Statistics in Astronomy
Review of Probability



Astrostatistics Texts

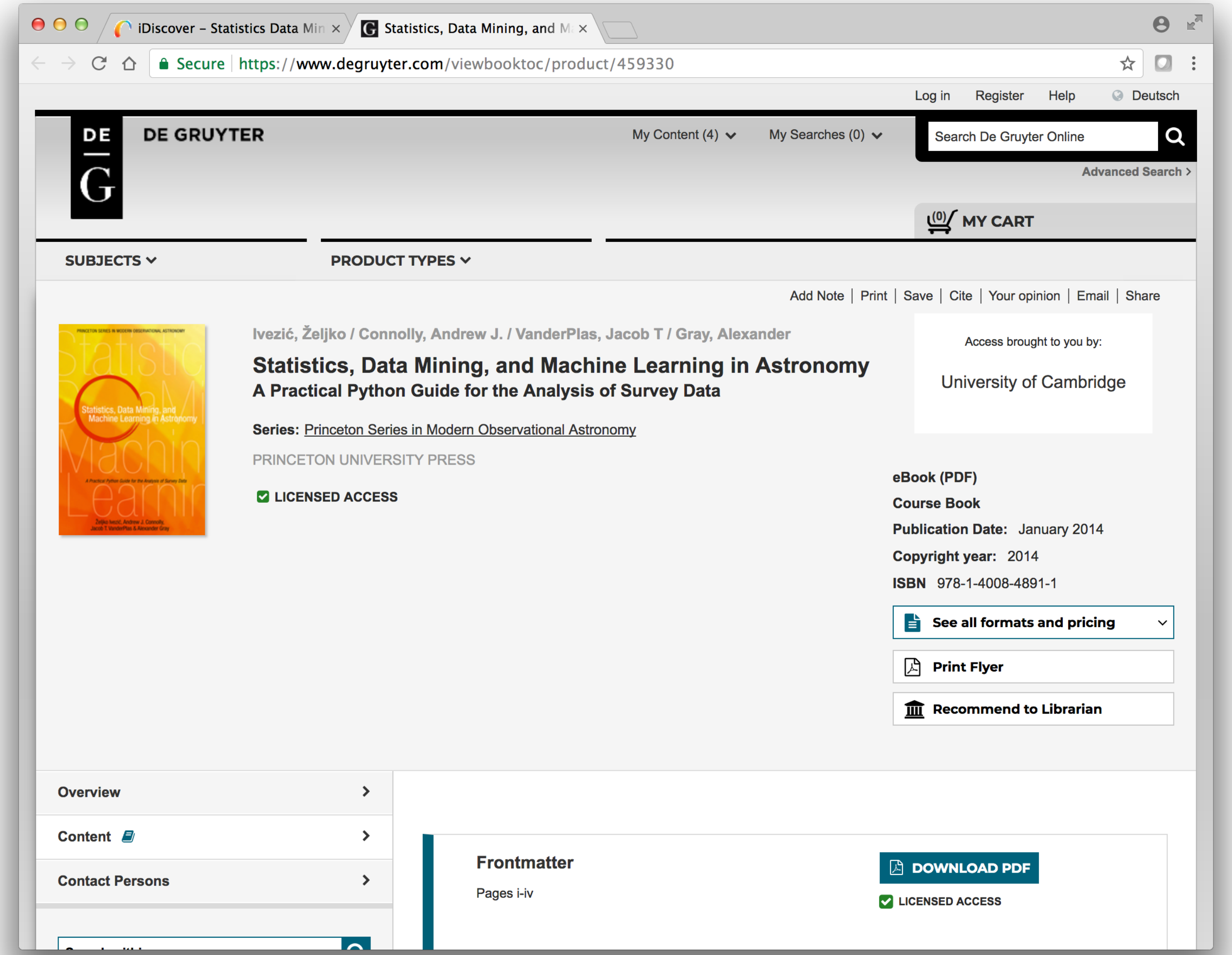
Z. Ivezic et al. **Statistics, Data Mining, and Machine Learning in Astronomy**. PUP, 2014. [Cambridge Library Online Access: Search at http://idiscover.lib.cam.ac.uk/, hard copies also available in library]

A Machine Learning bent. Python package AstroML and datasets to play with. Try it!

Recommended Reading:

Chapters 1, 3-5

Introduction and Basic review of Probability & Statistics



Introductory Reading or Review

- Feigelson & Babu Ch 1-4 and/or Ivezic, Ch 1, 3-5
- Probability foundations:
  - Probability axioms & properties
  - Conditional probability, Bayes' Theorem
  - Limit Theorems (LLN, CLT)
  - Random variables and univariate/multivariate probability distribution functions
  - Random number generation (computational)

Introductory Reading or Review

- Feigelson & Babu Ch 1-4 and/or Ivezic, Ch 1, 3-5
- Statistics foundations
    - Point Estimation: Moments, Least Squares, Maximum Likelihood, Confidence Intervals
    - Hypothesis Tests, Goodness-of-Fit, Model Selection
    - Sampling methods (e.g. bootstrap)
    - Likelihood Principle, Bayesian Inference and Parameter Estimation, Large-Sample Limits

Additional Reading: Intro to Astrostatistics

Short Articles

Roberto Trotta. **Astrostatistics is a field full of opportunities right now**
http://www.statisticsviews.com/details/feature/10741983/Astrostatistics-is-a-field-full-of-opportunities-right-now-An-interview-with-Rob.html

Long & de Souza. **Statistical methods in astronomy.**
https://arxiv.org/abs/1707.05834

C. Schafer. **A Framework for Statistical Inference in Astrophysics.** 2015, Annual Review of Statistics and Its Application, 2: 141-162

Statistics & Machine Learning

Going Deeper…

Gelman et al. **Bayesian Data Analysis**, 3rd Edition, 2013
[Hard copy in library]

Bishop et al. **Pattern Recognition and Machine Learning**, 2006
https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/ [FREE online]

MacKay, D. **Information Theory, Inference, and Learning Algorithms**, 2003
http://www.inference.org.uk/itila/ [FREE online]

Rasmussen & Williams. **Gaussian Processes for Machine Learning**, 2006
http://www.gaussianprocess.org/gpml/ [FREE online]

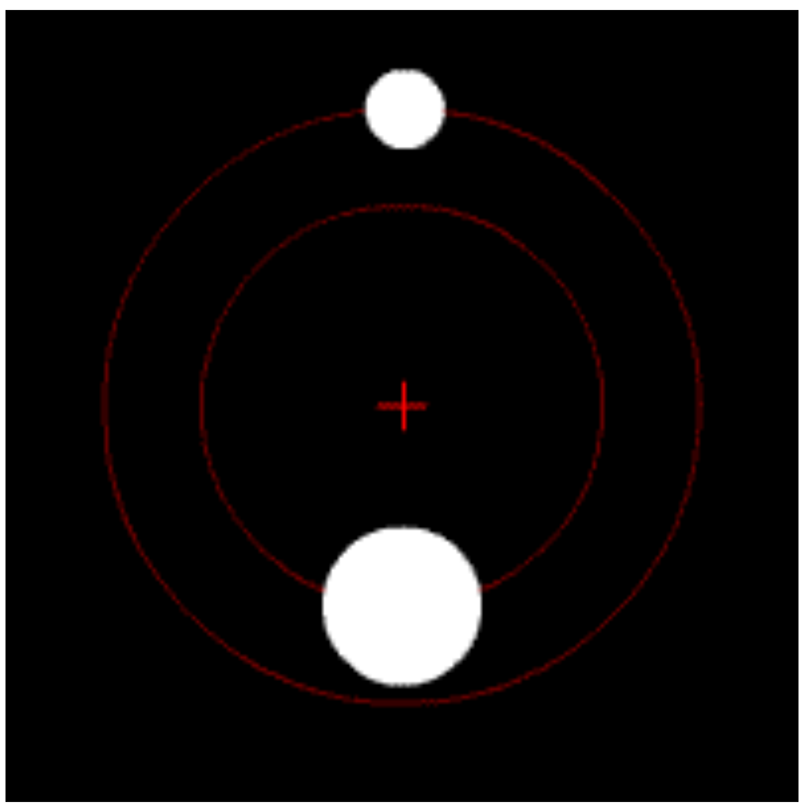
Hastie, Tibshirani and Friedman. **The Elements of Statistical Learning** (2nd Ed), 2009.
https://web.stanford.edu/~hastie/ElemStatLearn/ [FREE online]

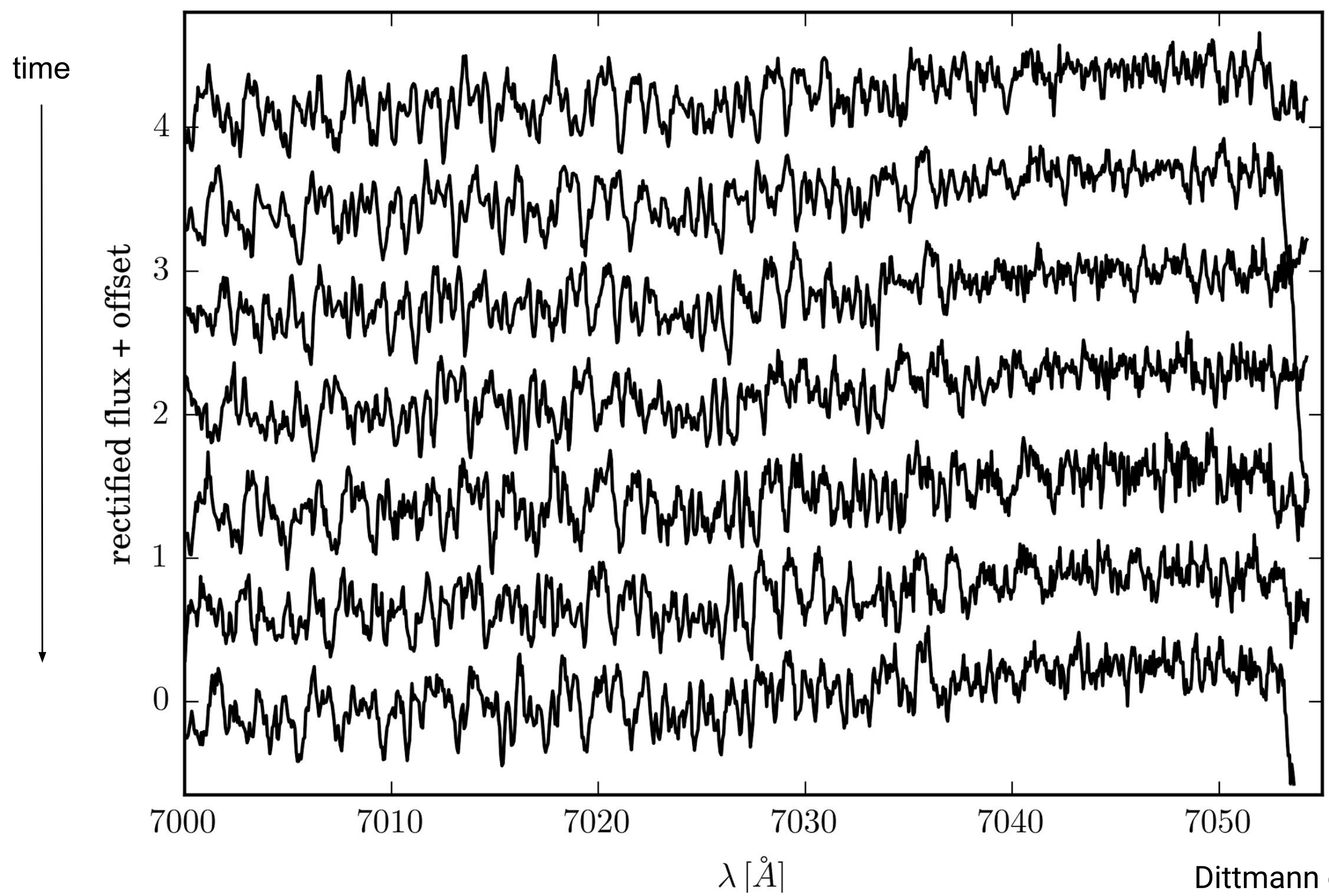
Topics

- Preliminaries / Statistics & Astronomy Background
- Regression / Fitting Models to Astronomical Data
- Generative / Forward Modelling
- Bayesian Inference
- Gaussian Processes / Nonparametric Bayes
- Time Series Analysis
- Hierarchical Bayesian Modelling
- Probabilistic Graphical Models
- Statistical Computation:
  - Markov Chain Monte Carlo
    - (Metropolis-Hastings, Gibbs, Hamiltonian)
  - Nested Sampling
  - Approximate Bayesian Computation (ABC)
- Model Selection
- Classification

Astrostatistics Case Studies:
Disentangling Time Series Spectra with Gaussian Processes: Applications to Radial Velocity Analysis
(Czekala et al. 2017)

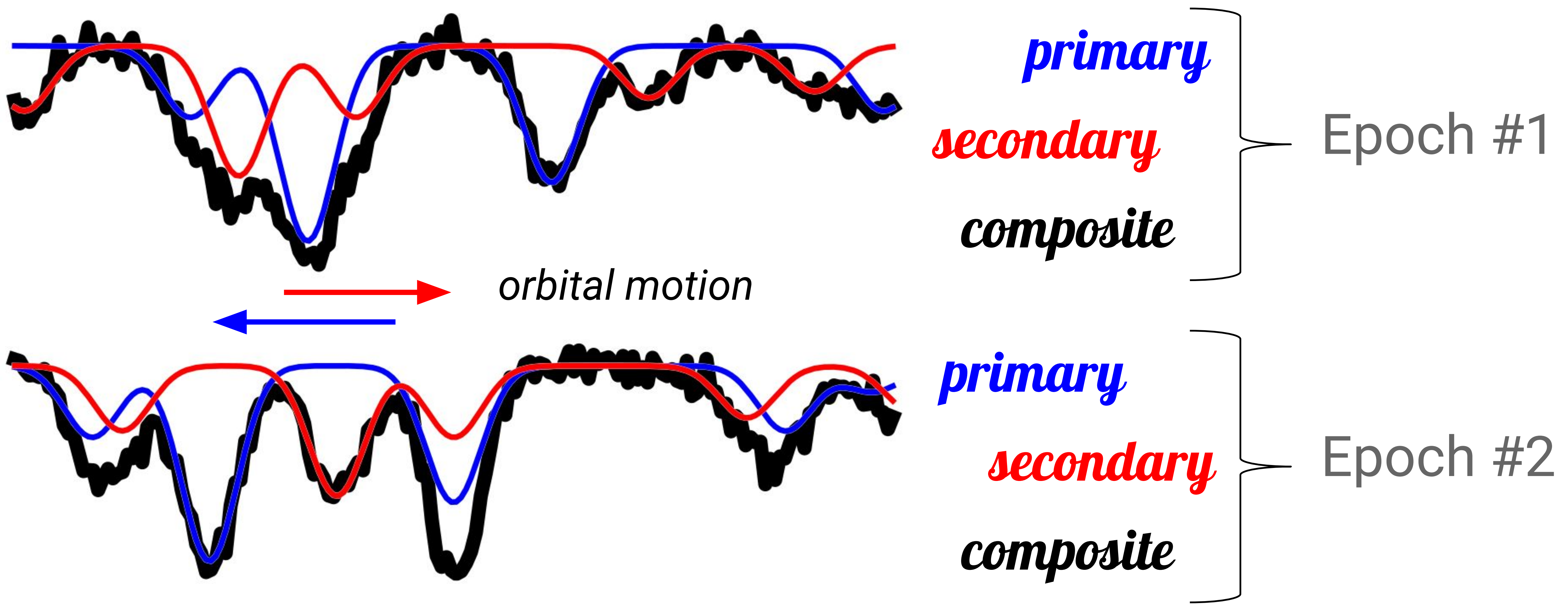
Raw Observations of the LP661-13 M4 Binary



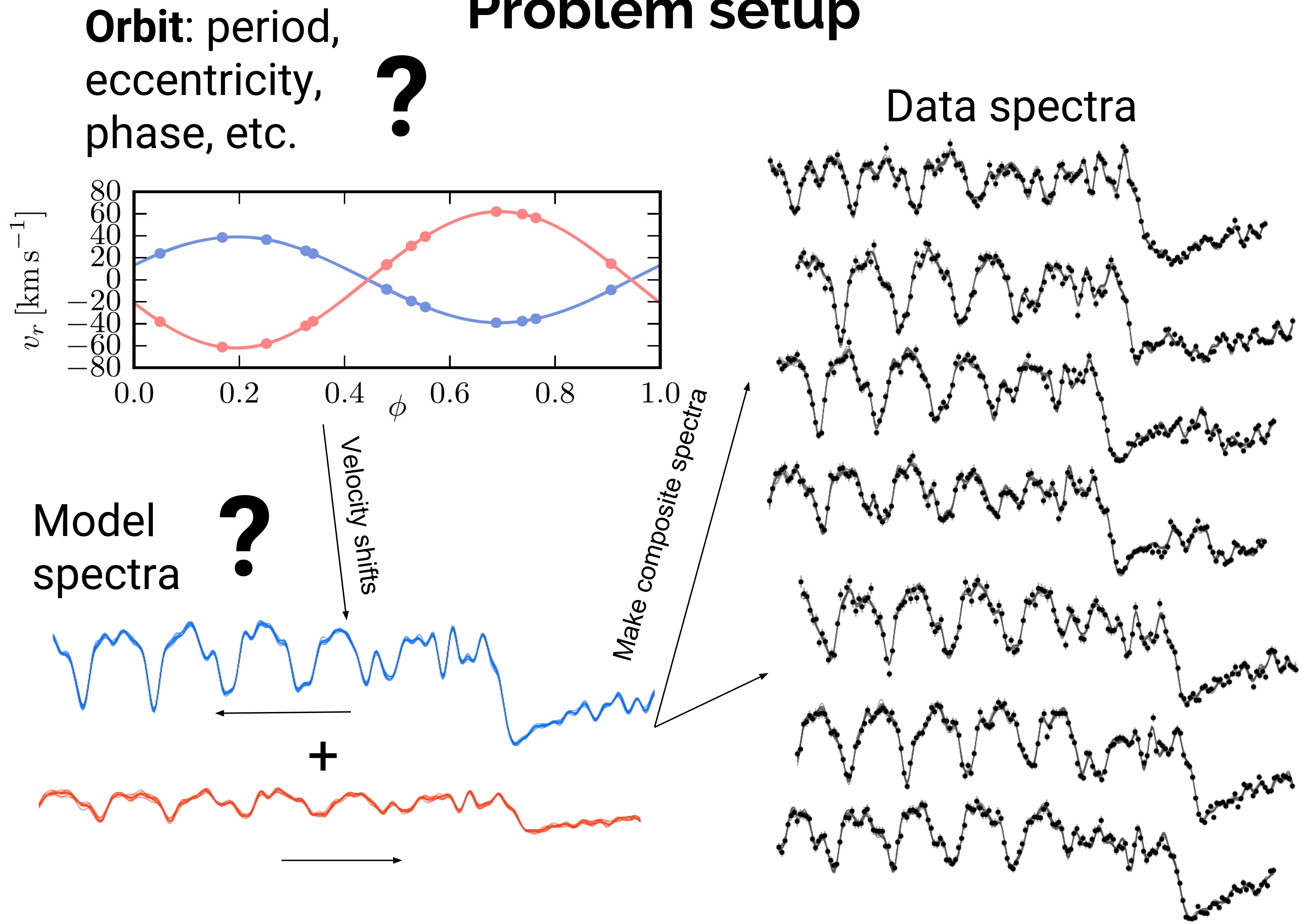


Dittmann et al. 17
Czekala et al. 17a

Spectroscopic Binary Stars



Problem setup



https://www.youtube.com/watch?v=kHjN42ft6aU

Goal: Go Backwards and Infer the Component Spectra & Orbital Parameters from noisy, observed (composite) spectra time series

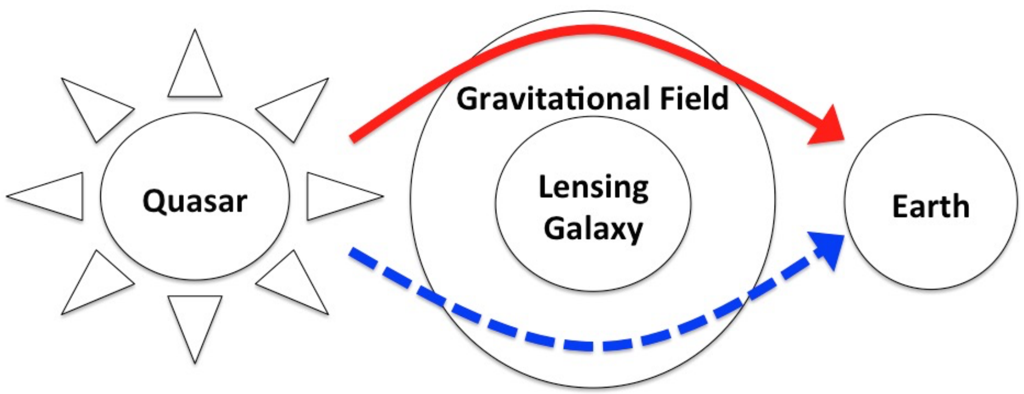
Astrostatistics Case Study 1: Disentangling Time Series Spectra with Gaussian Processes: Applications to Radial Velocity Analysis (Czekala et al. 2017, arXiv:1702.05652)

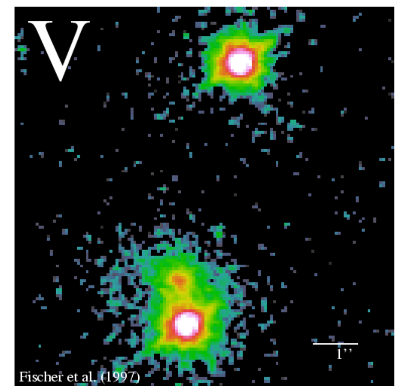
http://psoap.readthedocs.io/en/latest/

- Statistics:
  - Parametric Modelling (Orbit)
  - Nonparametric Modelling (Gaussian Process Spectrum)
  - Bayesian Inference
  - Markov Chain Monte Carlo
- Astronomy:
  - Applications to Radial Velocity Analysis of Stars/Exoplanets

Astrostatistics Case Study 2:

Bayesian Estimates of Astronomical Time
Delays Between Gravitationally Lensed Stochastic Light Curves
(Tak et al. 2017, Annals of Applied Statistics, arXiv:1602.01462)





Estimating time delays between noisy, irregularly sampled, gappy astronomical time series —> determine expansion rate of Universe (H_0)

- Bayesian Inference
- Stochastic Processes
- MCMC
- Gibbs Sampling

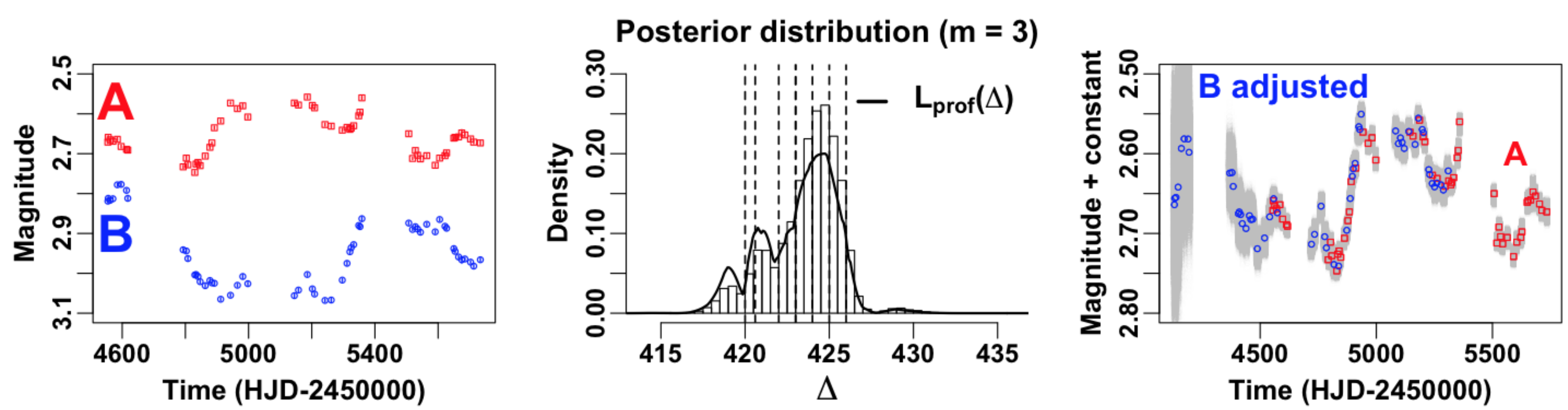
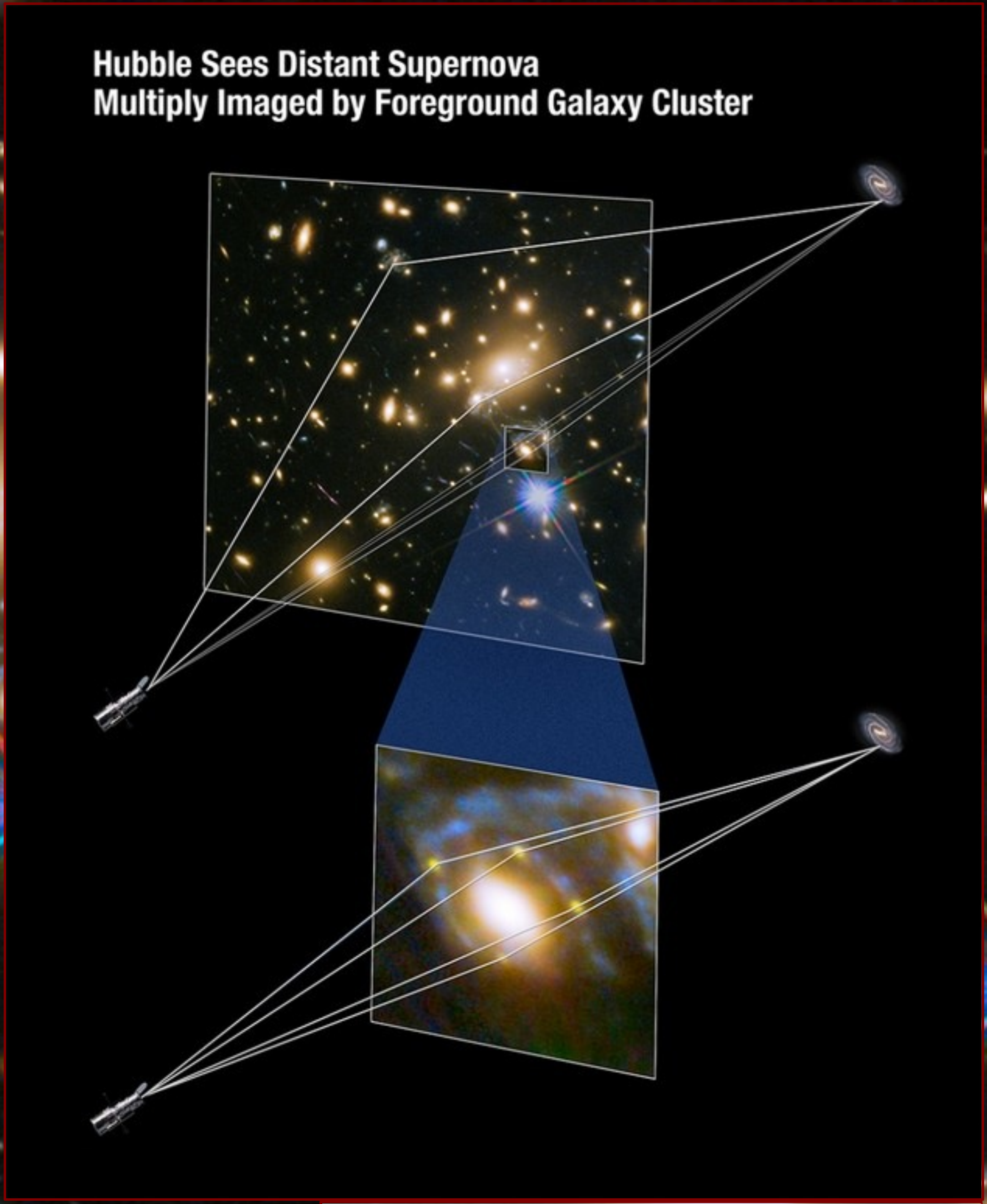


FIG 13. *Observations of Quasar Q0957+561 from Hainline et al. (2012) are plotted in the first panel. The second panel exhibits the marginal posterior distribution of Δ with*

Bayesian Gaussian Process Modelling of Supernova Refsdal

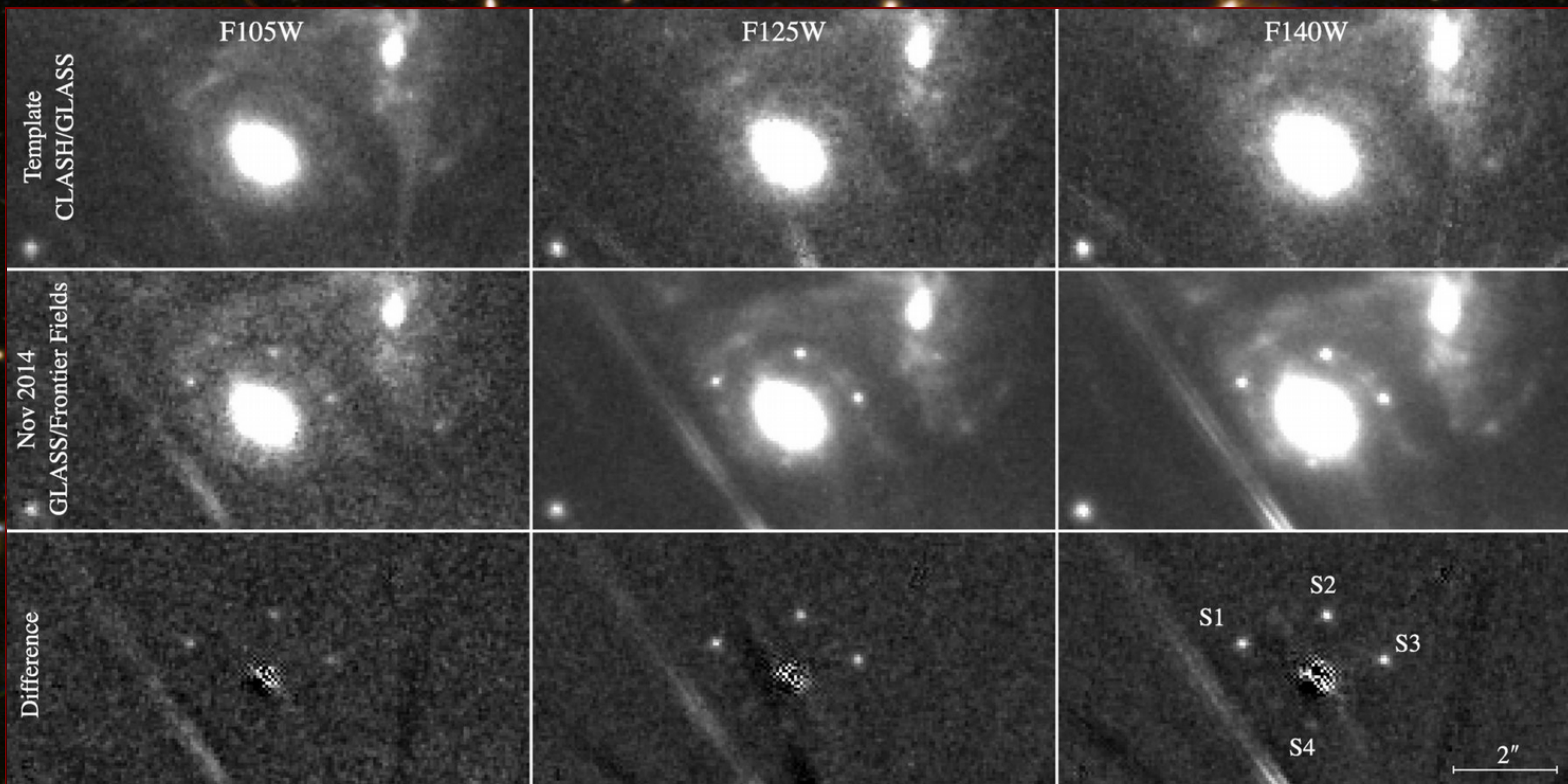
Stephen Thorp (IoA)

- Supernova: star explodes in a distant galaxy
- Light deflected by gravitational lens (intervening galaxy cluster)
- Multiple images formed by light on different paths
- Different travel times ⇒ each image shows different stage of explosion
- Allows us to measure time delay between images
- Can be used to infer H_0 ⇒ expansion rate of Universe



NASA/ESA/A. Feild (STScI)

- Refsdal: 1st supernova seen to be multiply imaged by a strong grav lens
- Initially 4 images, 5th followed as predicted
- Hubble Space Telescope (HST) observations made, giving partial light curves for each image

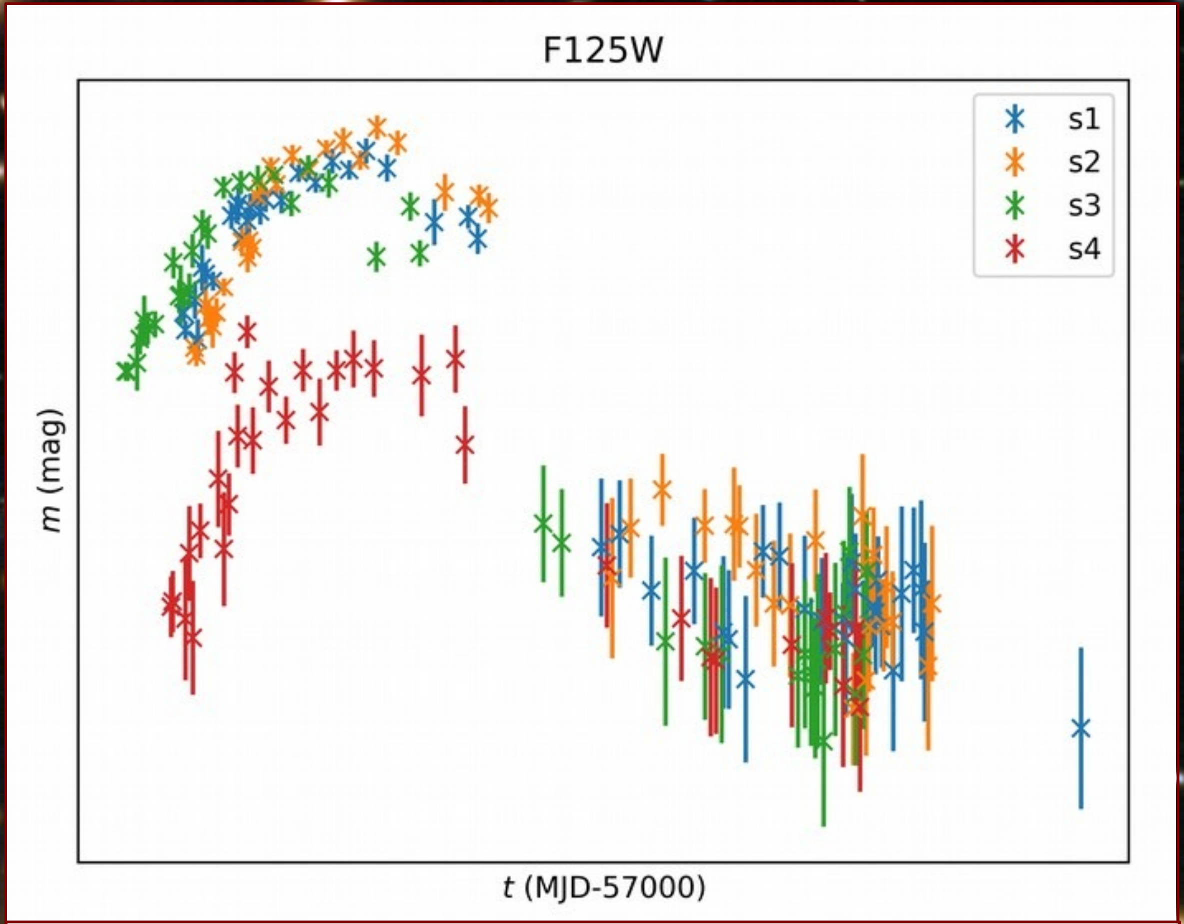


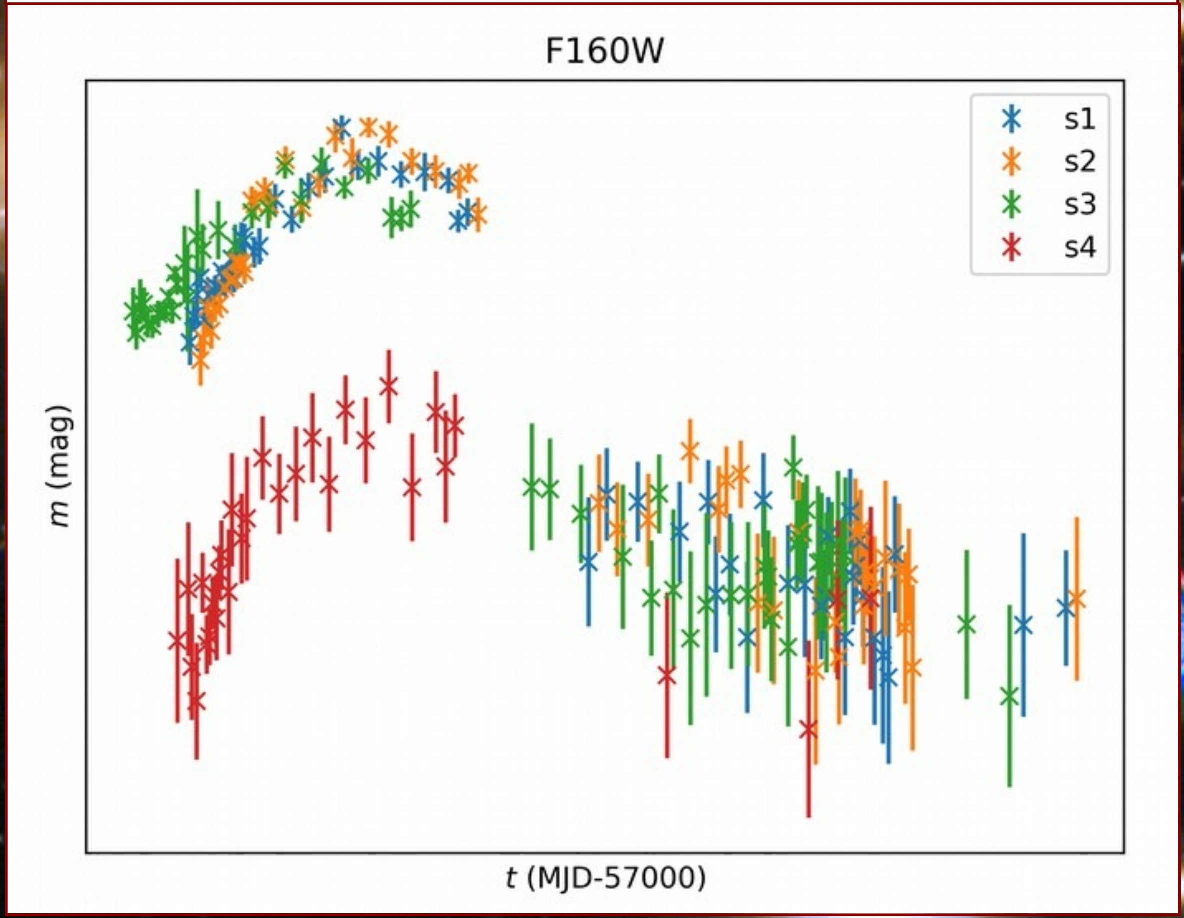
Kelly et al. (2015)

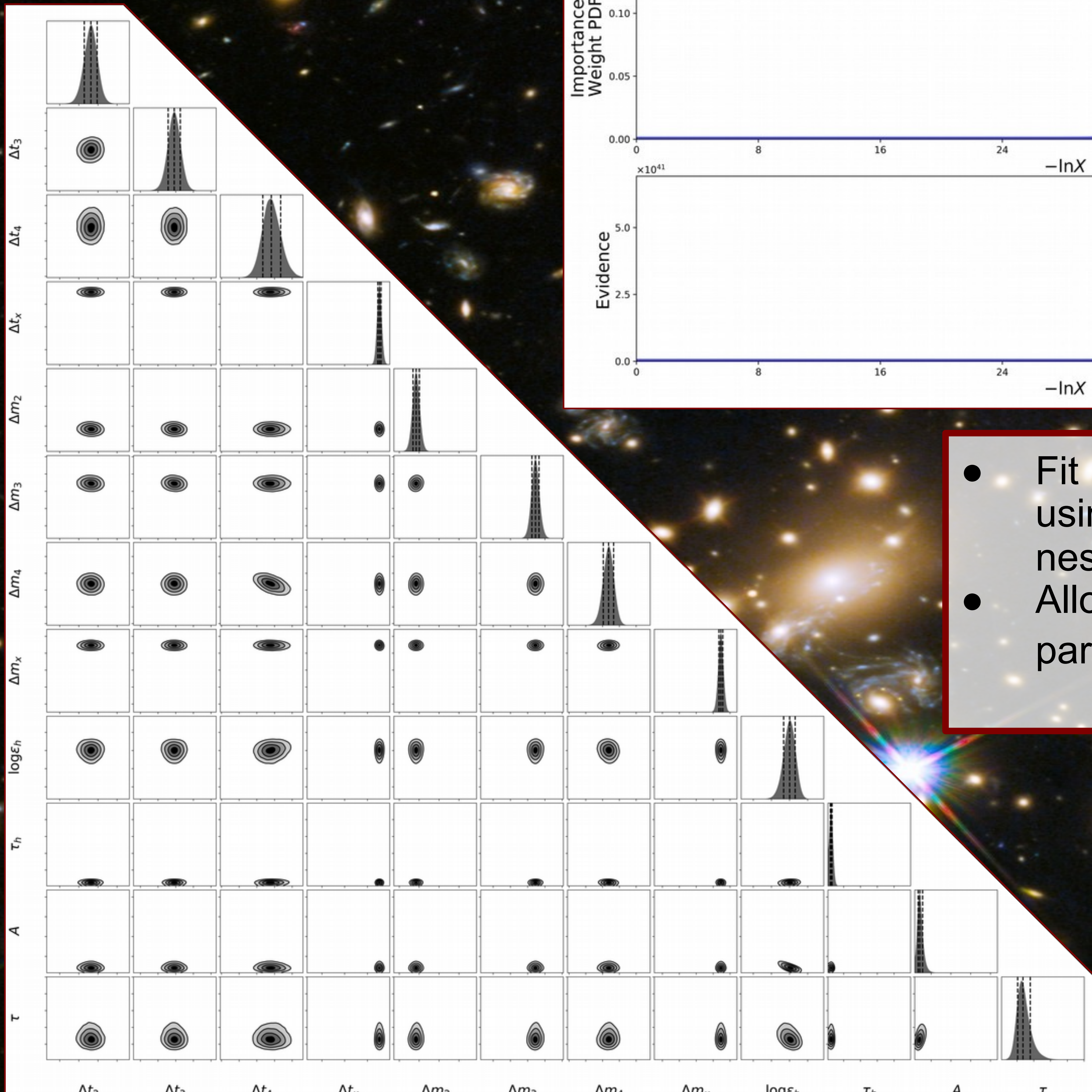
- Measuring time delay between the images' time series allows H_0 estimation

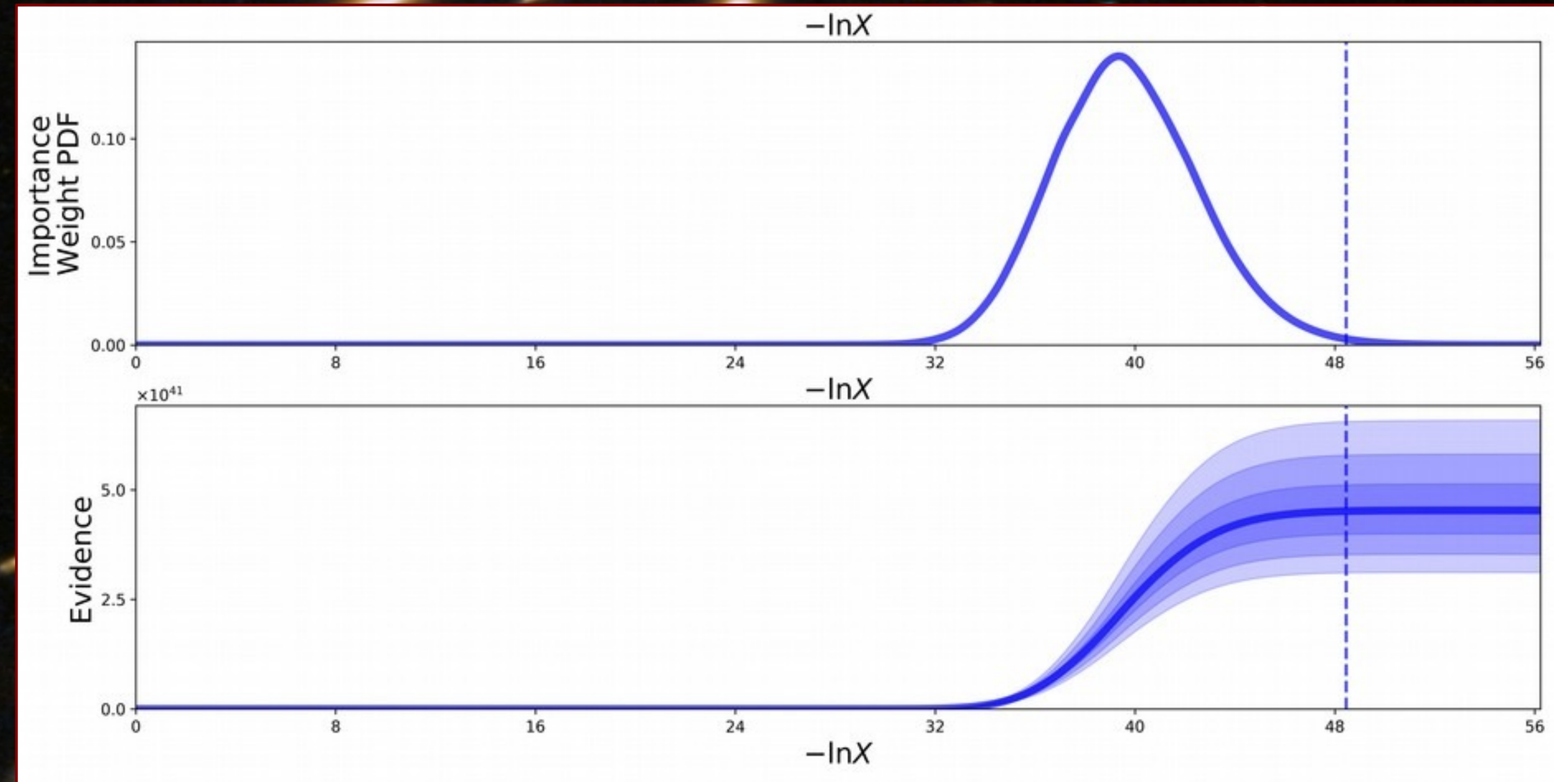
- Bayesian non-parametric approach, modelling latent light curve (time series) as a Gaussian process

- **Model**: Time series for each image assumed to be time-shifted and magnified version of latent (true) light curve, plus measurement errors



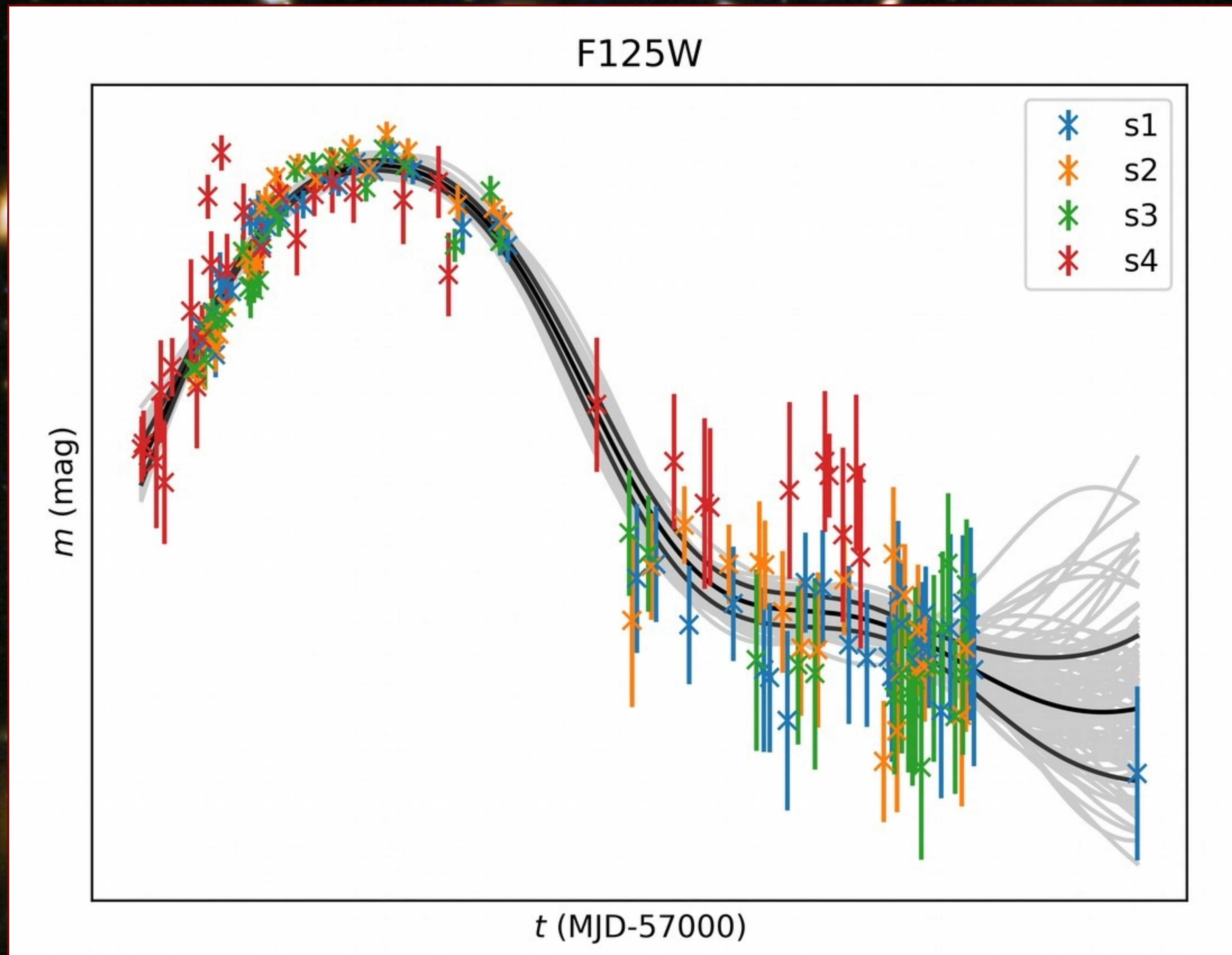






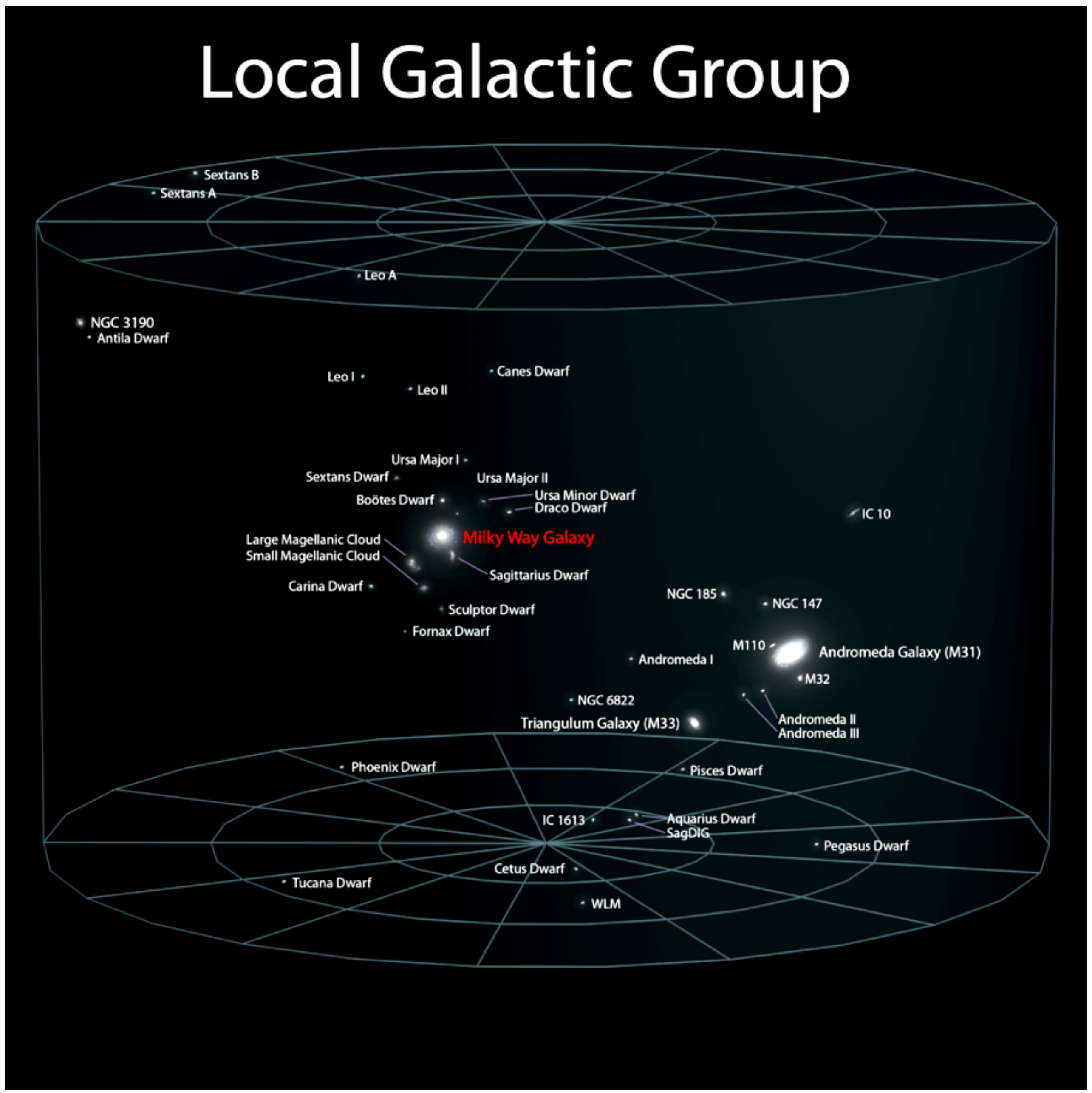
- Fit model to data using MCMC or nested sampling
- Allows inference of best parameter values

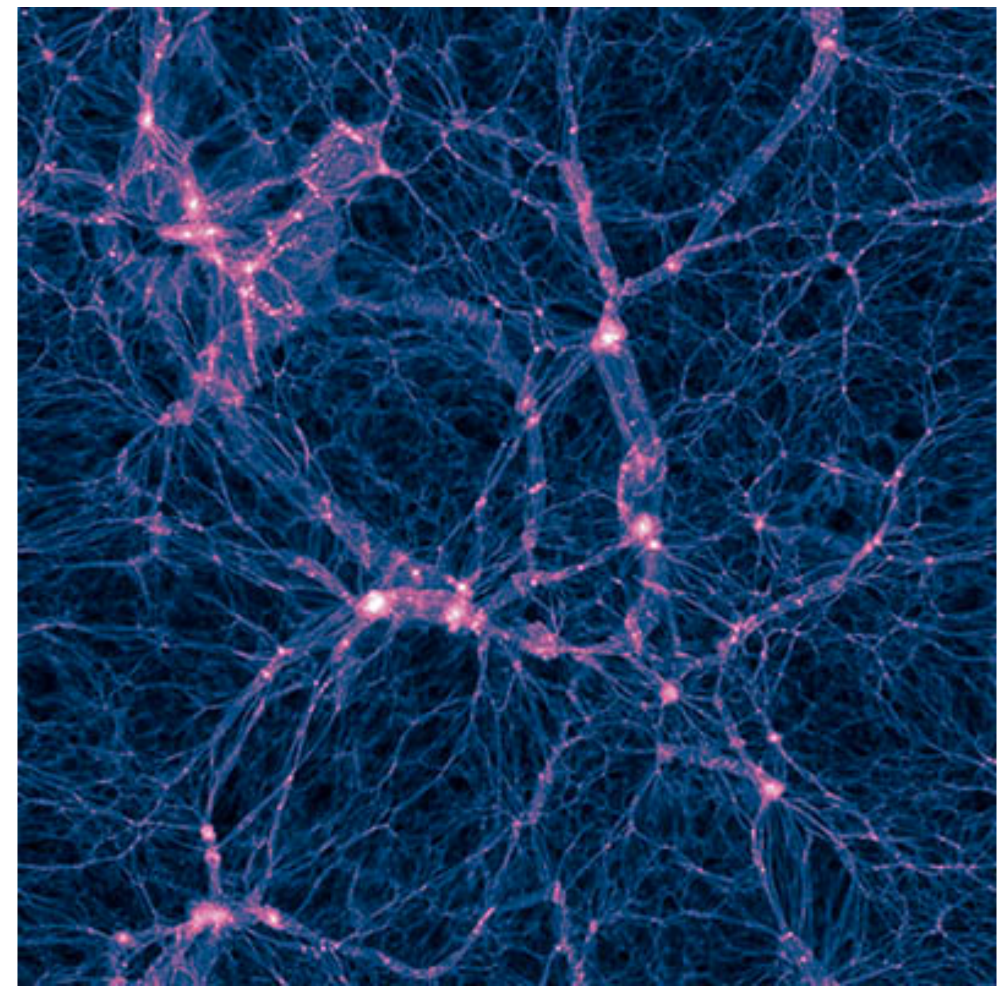
- Successful parameter inference matches data for images back to mean latent light curve
- Can sample realisations from posterior GP



Astrostatistics Case Study 3:

Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations (Patel et al. 2017, arXiv:1703.05767)





Illustris
Cosmological Simulation of Galaxy Formation

Illustris Cosmological Simulation
Movie

http://www.illustris-project.org/movies/illustris_movie_cube_sub_frame.mp4