# Identifying Black Holes and Neutron Stars in Large Astronomical Surveys using Gaussian Processes


International Centre for Radio Astronomy Research

Shih Ching Fu

Co-supervisors:

Dr Arash Bahramian, Dr Aloke Phatak

Associate Supervisors:

Dr James Miller-Jones, Dr Suman Rakshit

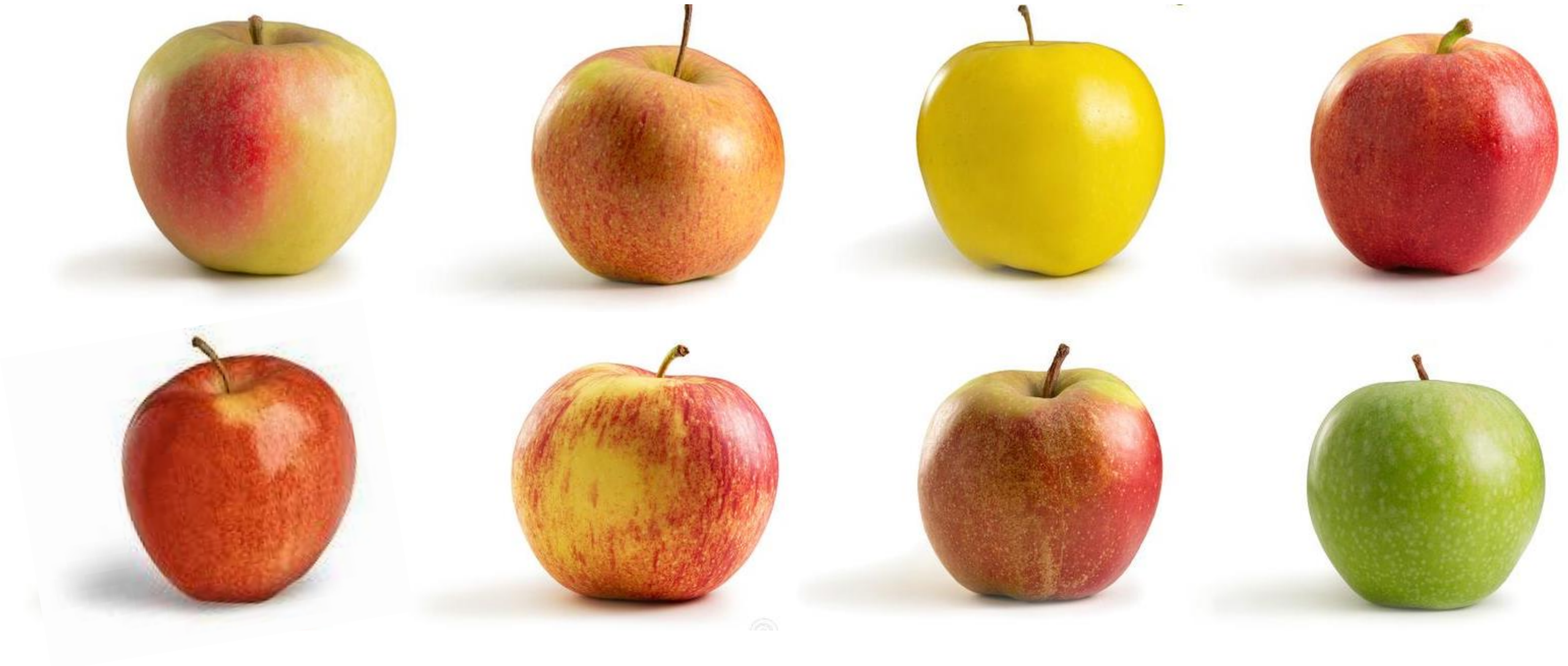
Curtin University

The University of Western Australia

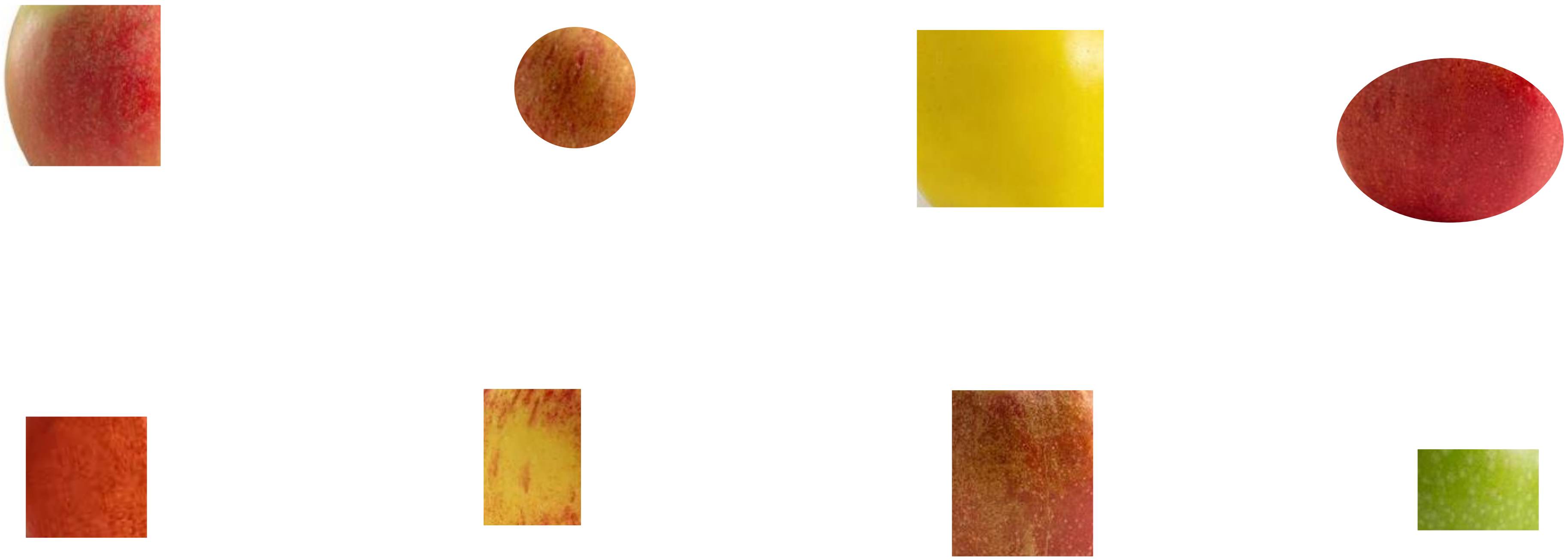
Government of Western Australia
Department of the Premier and Cabinet
Office of Science

# What physical characteristics can be used to distinguish these types of produce?

# … what about now?



# … and now?



# Light Curves in Astronomy

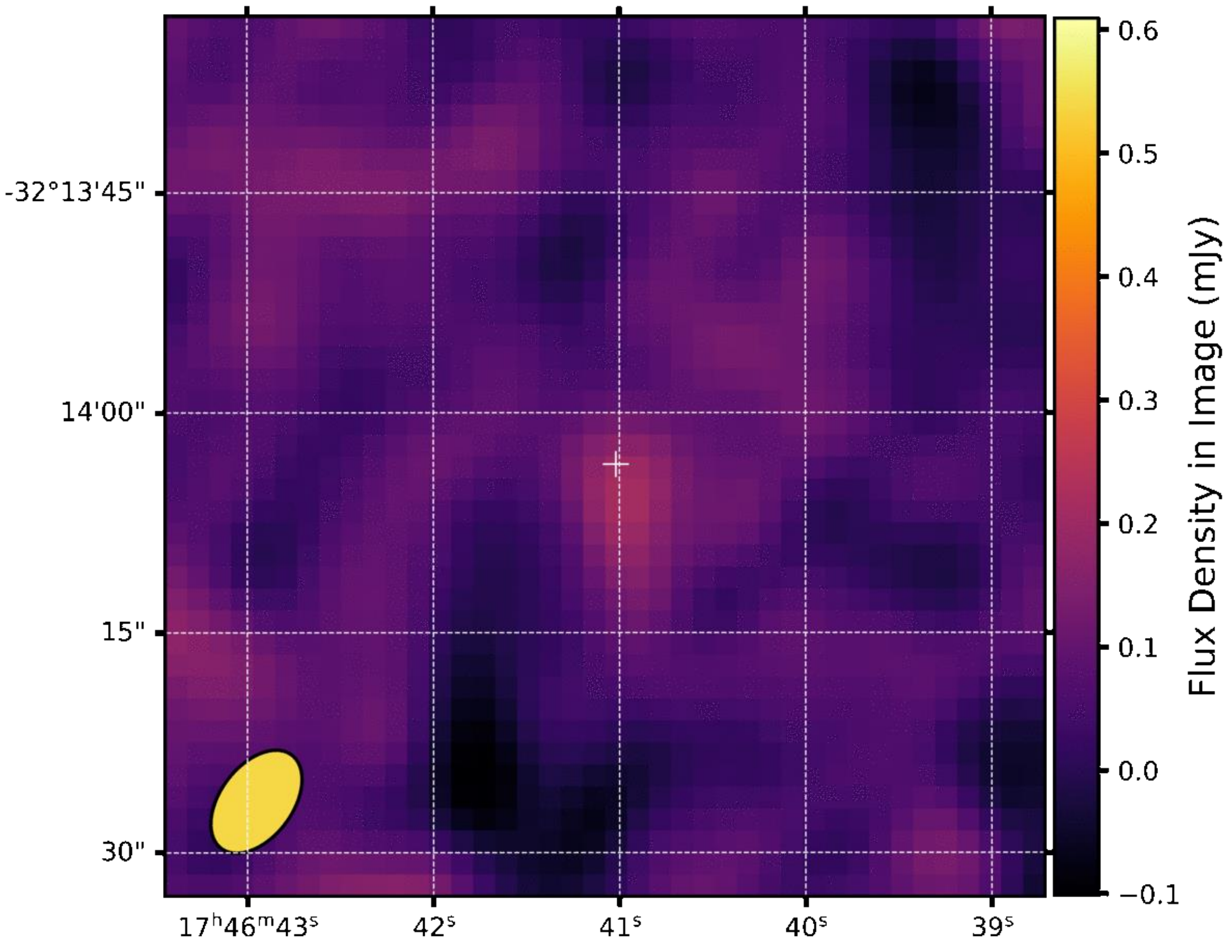
*Light curves* are time-series describing the brightness of a celestial object over time.

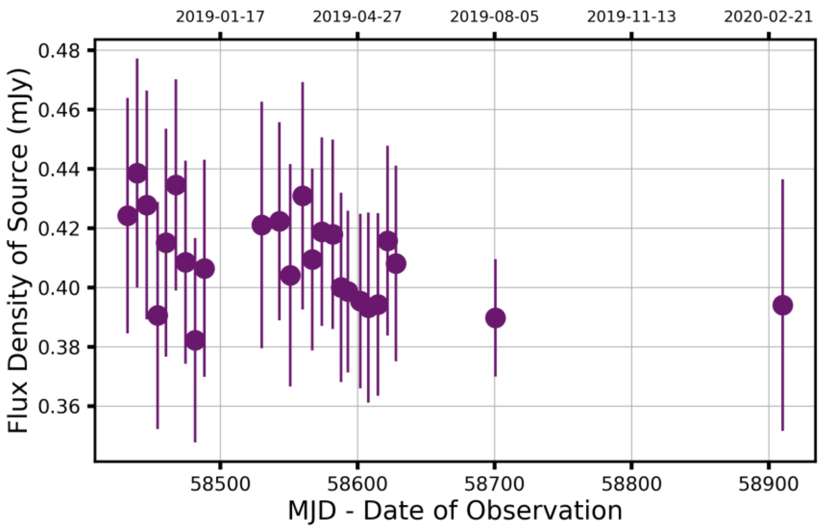
Variability in brightness can reveal information about the processes at work within an object or help identify the category of event being observed.

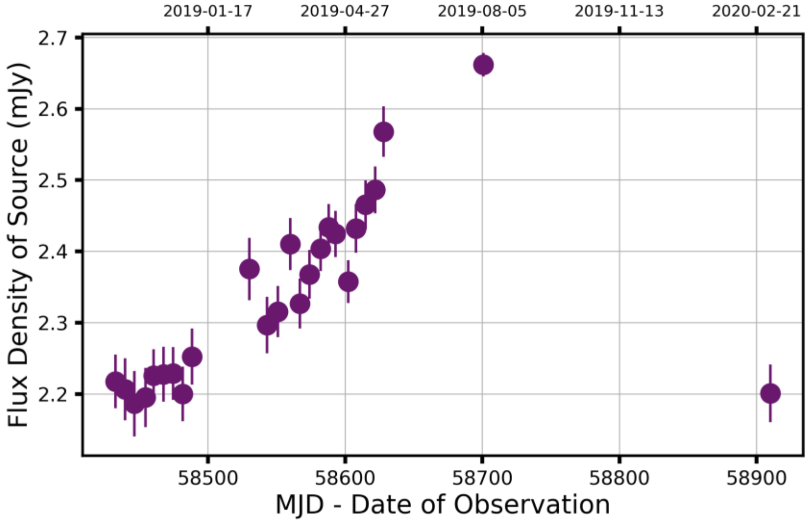
**But beware!**

- Sparsity of observations
- Uneven sampling rates
- Varying noise levels

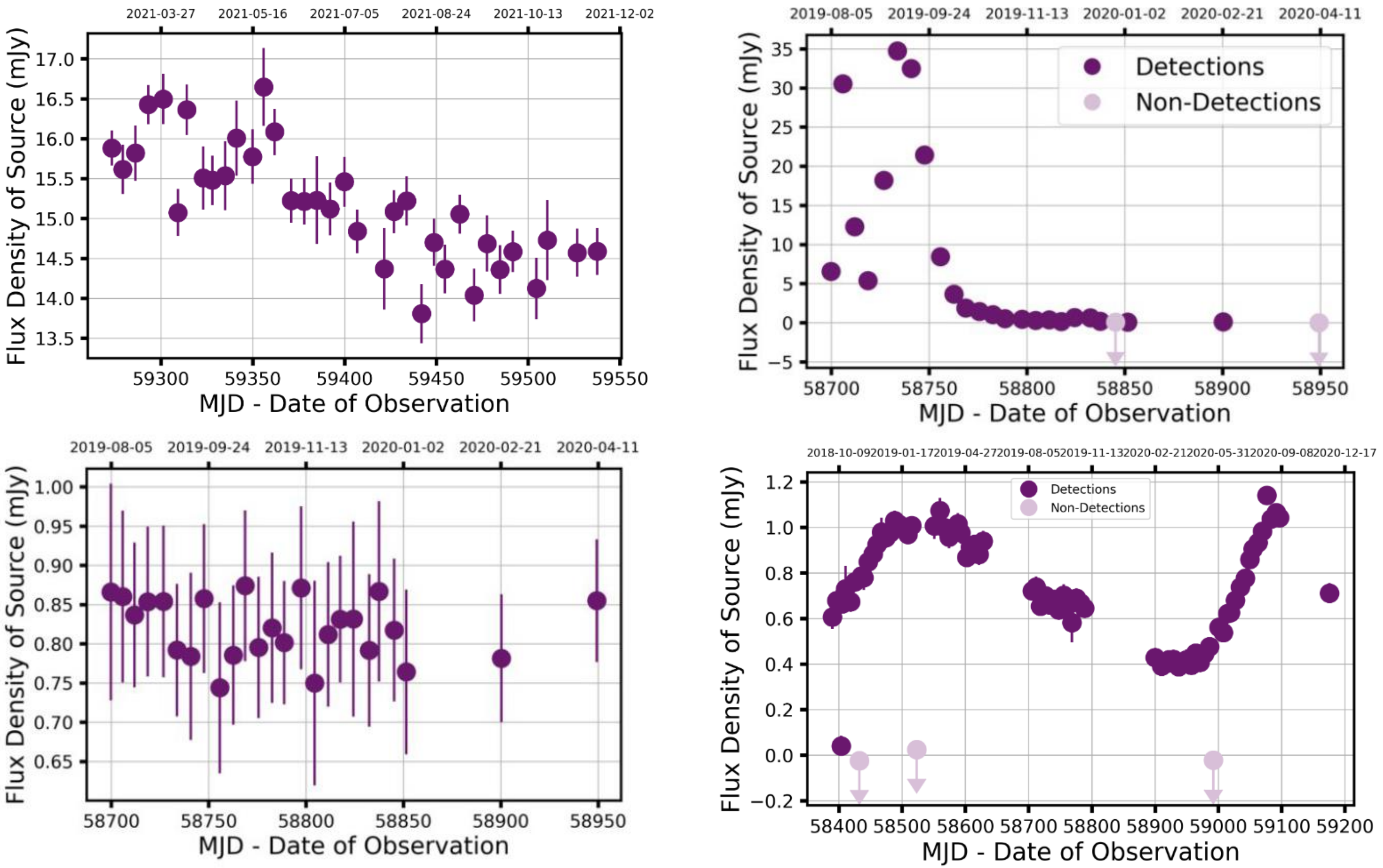
# Hunting for ‘Variables’



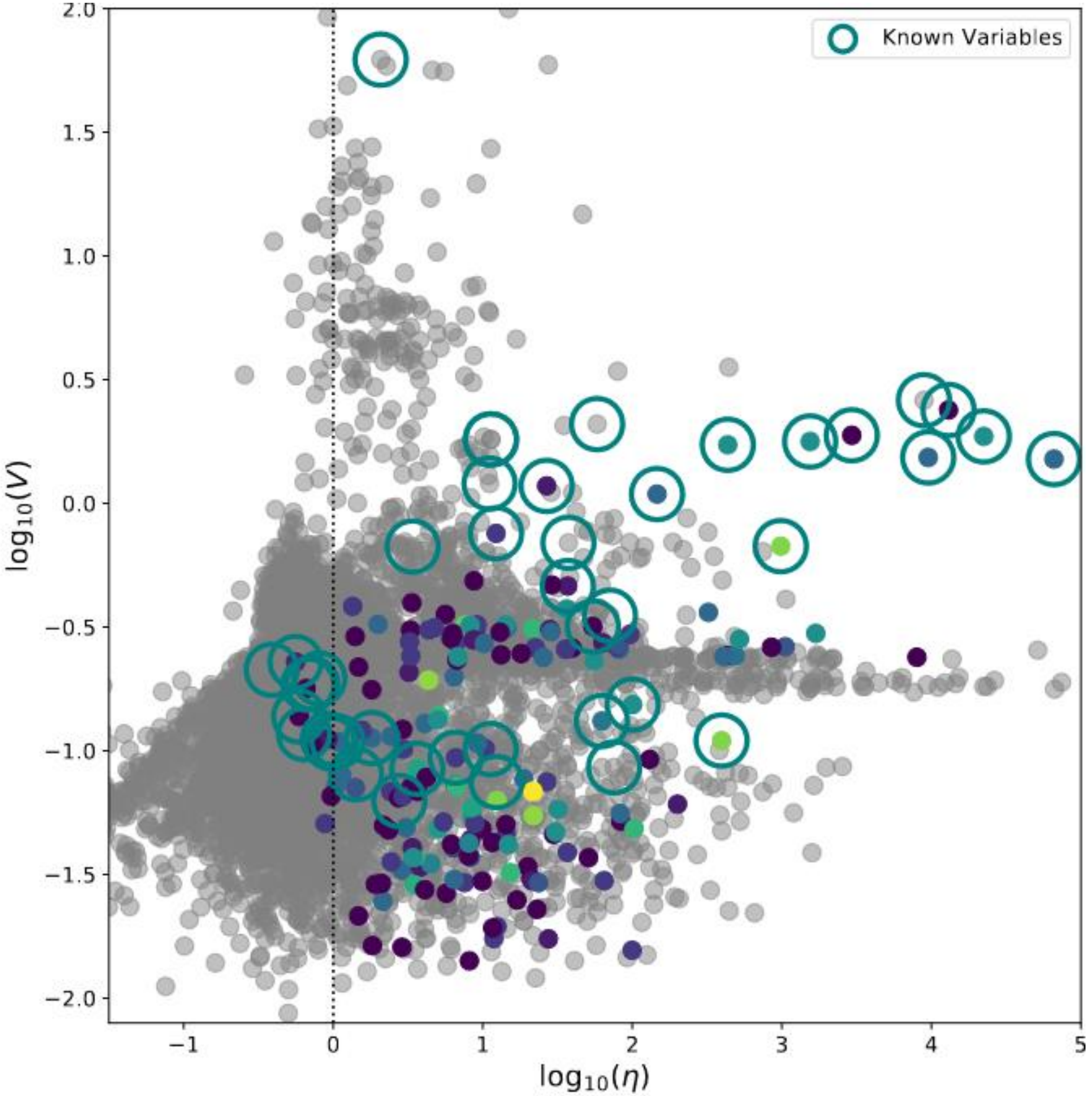




Bursts from Space: MeerKAT (https://www.zooniverse.org/projects/alex-andersson/bursts-from-space-meerkat)




Andersson, et al., 2023. MNRAS, in press.




Andersson, et al., 2023. MNRAS, in press.

# TRAP Variability Metrics (Swinbank, et al., 2015)

- Flux density coefficient of variation, $V_\nu$
- Statistic of flux density variability, $\eta_\nu \sim \chi^2_{N-1}$

- As $V_\nu \rightarrow 0$ and $\eta_\nu \rightarrow 0$, consistent with a stable source.

**Variable sources are spread across the 2D parameter space**

# Characterising Light Curves

**Oversimplified** ⟷ **Overspecified**

**Oversimplified**

- Fewer parameters
- Scales easily
- High information loss

**Overspecified**

- Many parameters
- High discriminatory power
- Overfitting

# Gaussian Processes
# (GPs)

# Multivariate (Normal) Gaussian

$\boldsymbol{Y}$ is a vector of $n$ Gaussian random variables.

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \boldsymbol{Y} \sim \mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{n\times n})$$

where $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ and $\boldsymbol{\Sigma}$ is a $n \times n$ covariance matrix.

# Covariance Matrix

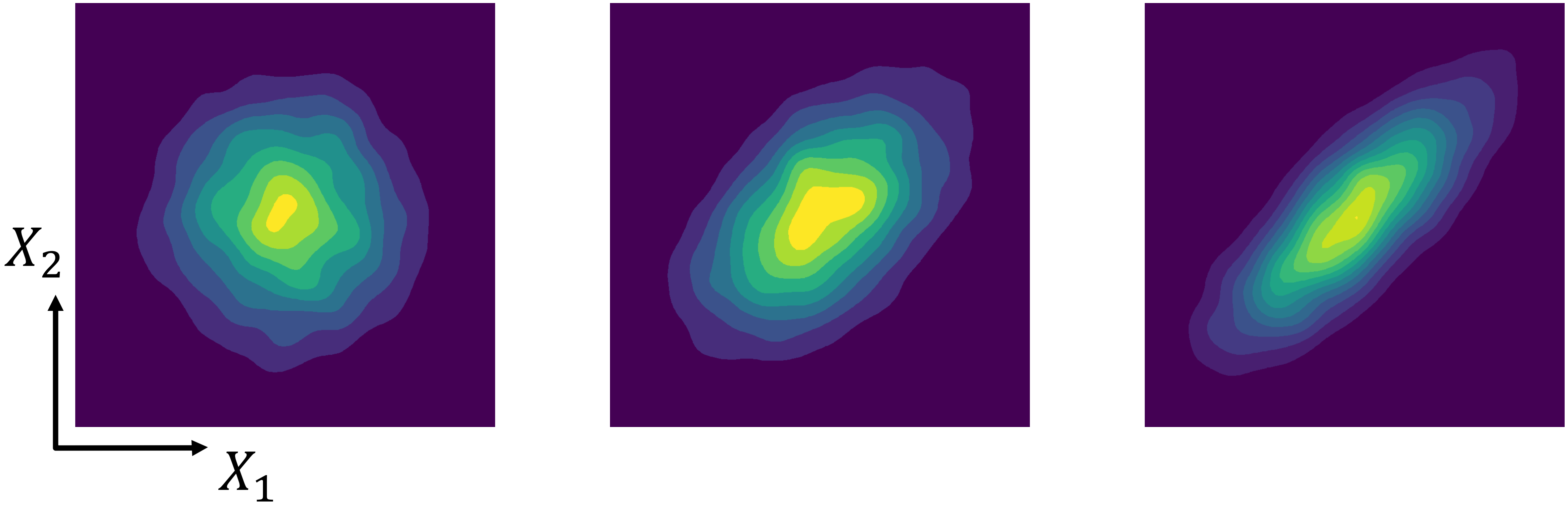
Each entry $\Sigma_{ij} = \mathrm{Cov}(Y_i, Y_j)$ describes how much $Y_i$ and $Y_j$ co-vary or influence each other.

$$\boldsymbol{\Sigma}_{n\times n} = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1n} \\ \vdots & \ddots & \vdots \\ \Sigma_{n1} & \cdots & \Sigma_{nn} \end{bmatrix} \qquad \Sigma_{ii} = \sigma_i^2$$

- Symmetric, positive semi-definite matrix.
- Linear combinations of covariance matrices are also valid covariance matrices.

# Bivariate Gaussian $\boldsymbol{X} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_{2\times 2})$

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



# Gaussian Processes

Extend multivariate Gaussian to 'infinite' dimensions.

- Mean function, $\mu()$
- Covariance or kernel function, $k()$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \end{bmatrix} = \boldsymbol{Y} \sim GP(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu} = \mu(t_i)$ and $\boldsymbol{\Sigma} = k(t_i, t_j)$, for $i, j = 1, 2, \ldots$

Rather than specifying a fixed covariance matrix with fixed dimensions, compute covariances using the kernel function.

# Kernel Functions

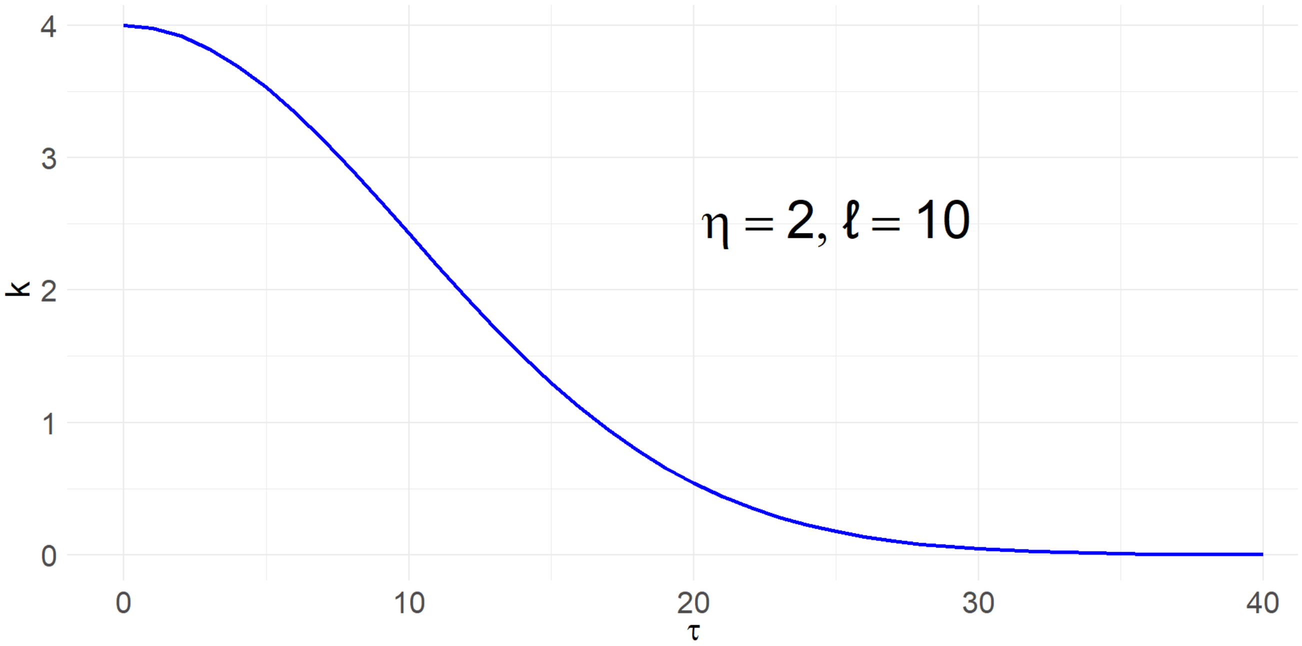
$$\Sigma_{ij} = \mathrm{Cov}(Y_i, Y_j) = k(t_i, t_j) \qquad i, j = 1, \ldots$$

There are many kernel functions to choose from!

- Squared Exponential (Radial Basis), (Absolute) Exponential, Matern-3/2, Matern-5/2, Rational Quadratic, Cosine, Sine Squared Exponential (Periodic), Stochastic Harmonic Oscillator, etc.
- or combinations thereof

Each has its own functional form and parameterisation.

# Squared Exponential Kernel




$$k(\tau;\eta,\ell) = \eta^2 \exp\left\{-\frac{1}{2}\left(\frac{\tau}{\ell}\right)^2\right\}$$

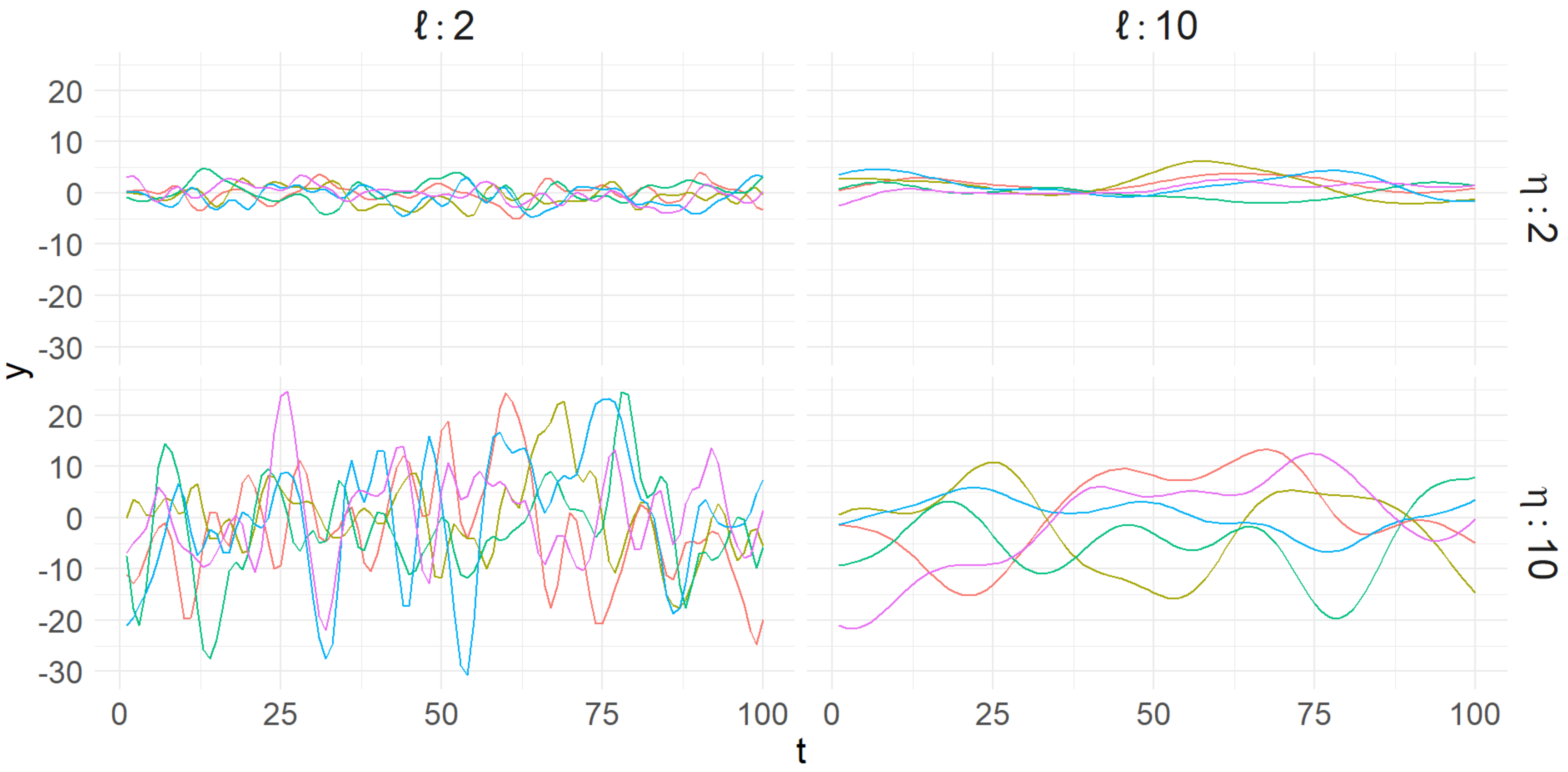
$$\eta, \ell > 0$$

$$\tau = |t_i - t_j|$$

- As distance (in time) increases ↗, the covariance decreases ↘.
- Stationary time-series

# Amplitude $\eta$, Lengthscale $\ell$

$$k(\tau;\eta,\ell) = \eta^2 \exp\left\{-\frac{1}{2}\left(\frac{\tau}{\ell}\right)^2\right\}$$




NB: Squared Exponential kernel is not periodic!

# Research Proposal

# Motivations

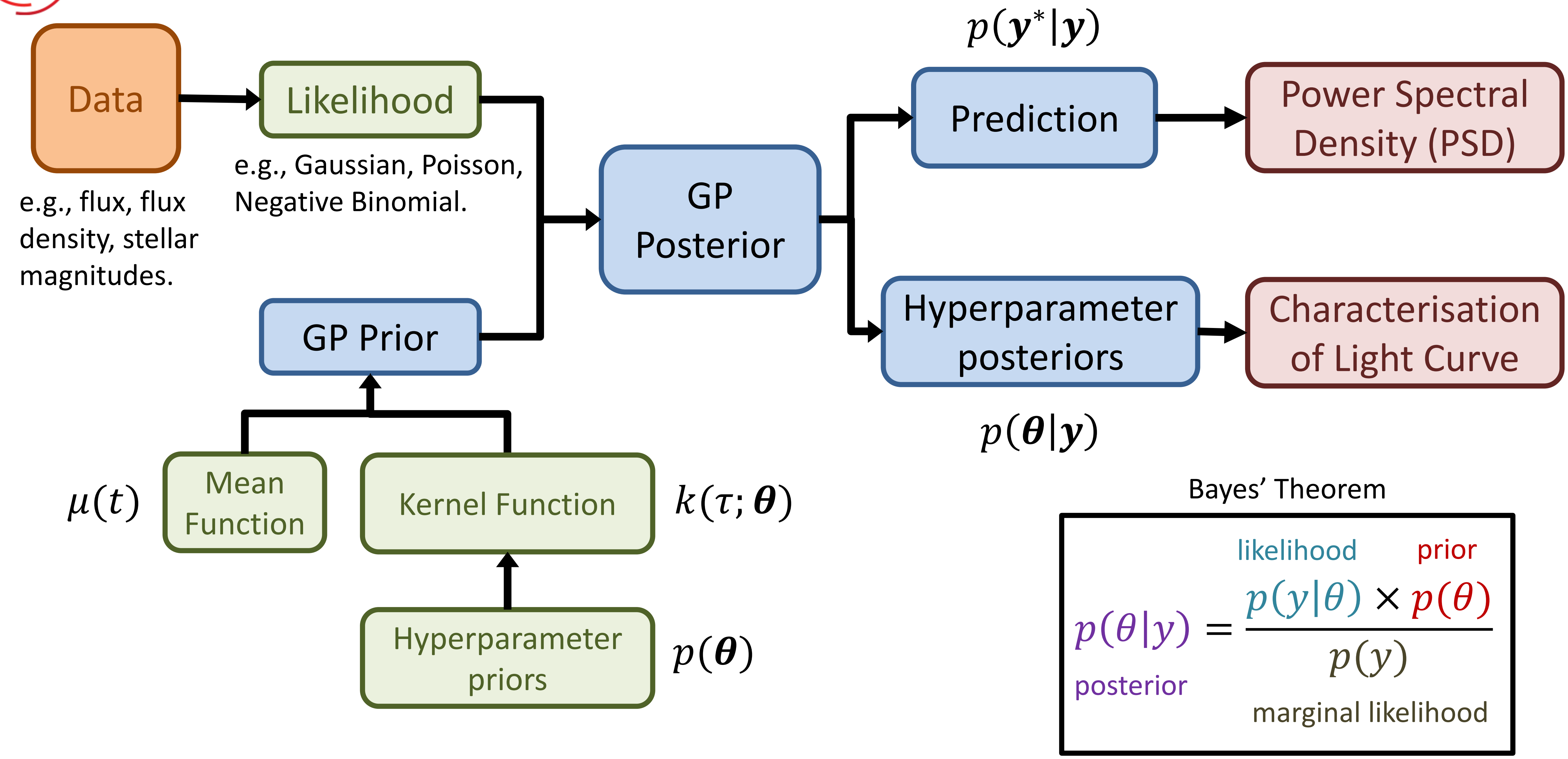
1. Black Holes are eluding us!
    - Estimated population is $> 10^5$ but only found $\approx 100$.
    - Need to find more to get better understanding.
2. Emerging large astronomical surveys
    - Need techniques to analyse these new large datasets.
3. Gaussian Processes are showing promise in astronomy
    - Handles sparse, unevenly sampled data.
    - Flexible
    - Lends itself to Bayesian inference.

# Objectives

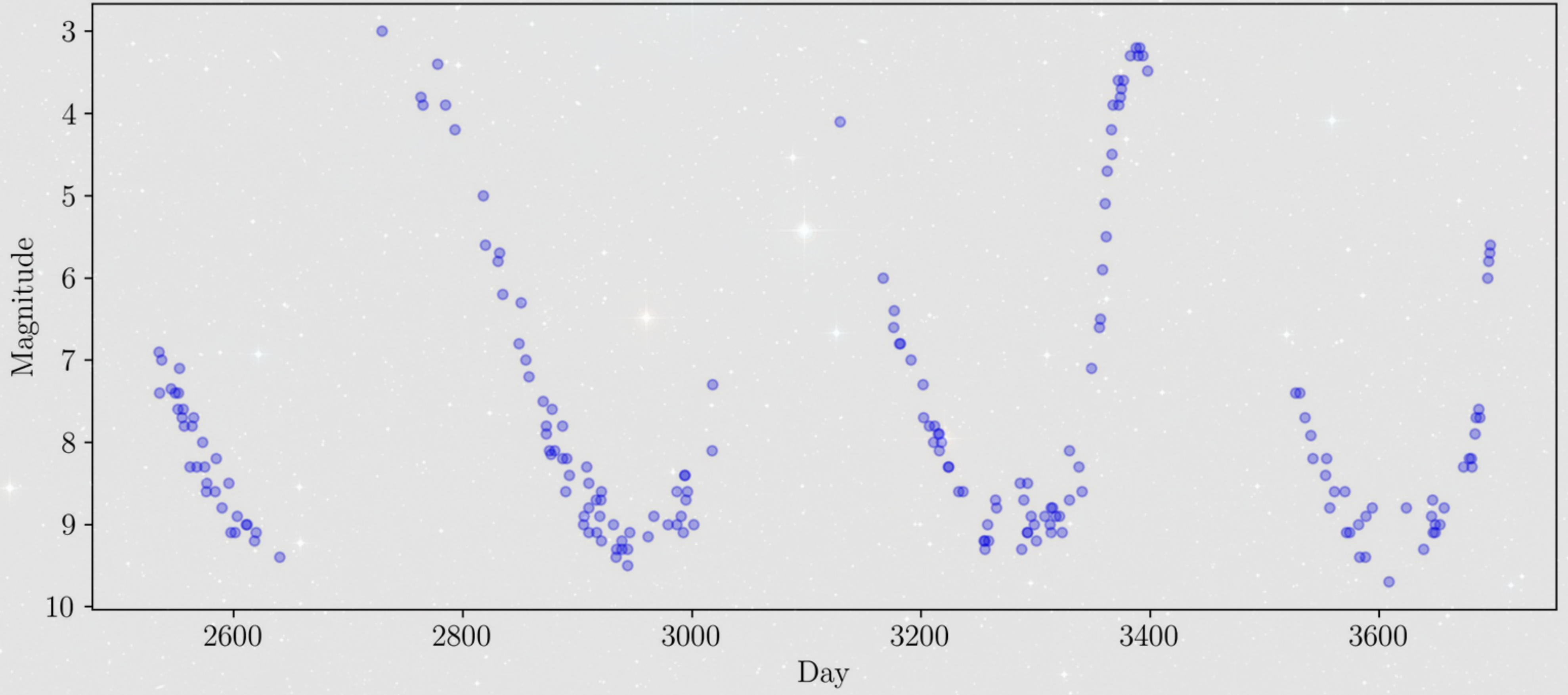
1. Develop a framework using Gaussian Processes for characterising light curves.
2. Apply this framework to identify black hole and neutron star candidates in large astronomical surveys, e.g., LSST, SKA.
3. Build software tools that will be adopted by astronomers.

# Work to Date

# Modelling Workflow



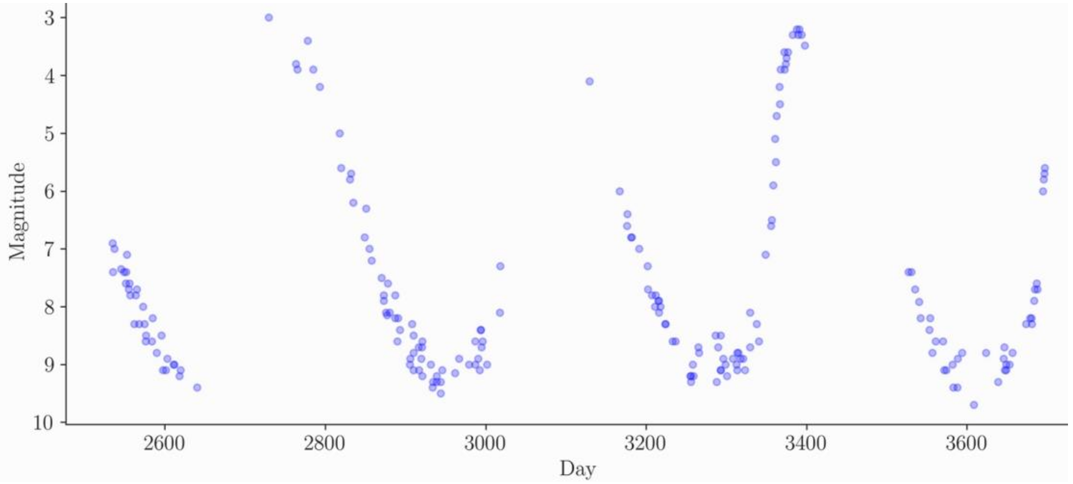
# Benchmarking: Mira




Data source: AAVSO (American Association of Variable Star Observers)

Image credit: ESO/Davide De Martin

# Mira: Model



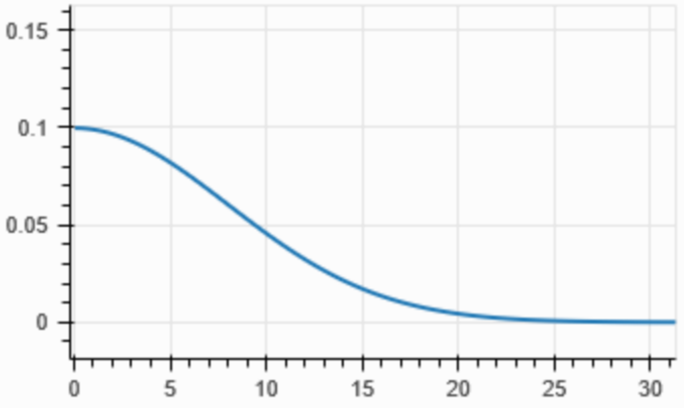

$$\mu(t) = \bar{y}$$

Mean function

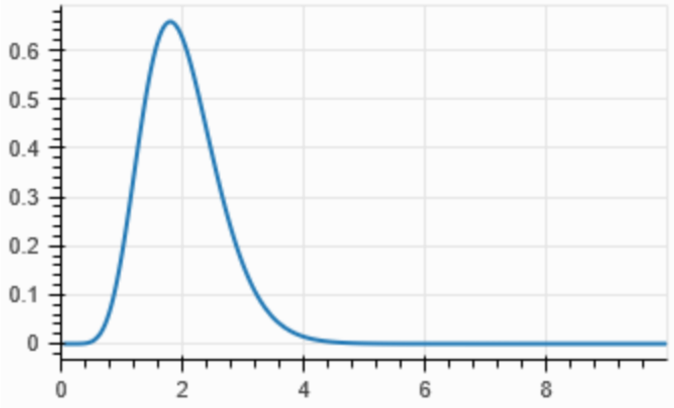
$$k(\tau) = \eta^2 \exp\left\{-\frac{1}{2}\left[\frac{\sin\left(\pi \frac{\tau}{T}\right)}{\ell}\right]^2\right\}$$

Periodic kernel function

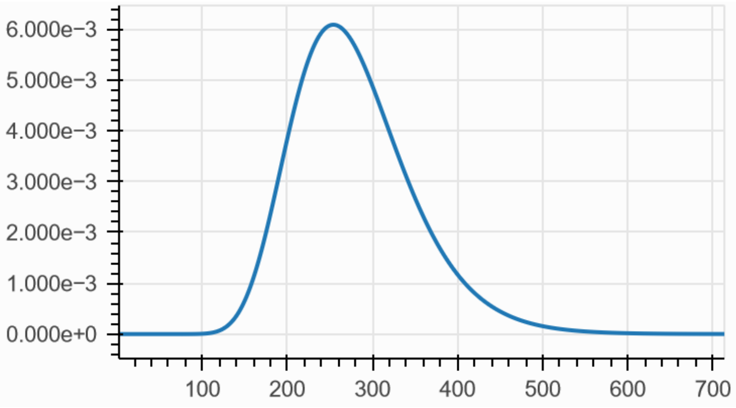
$\eta \sim \mathrm{HalfNormal}(8)$




$\ell \sim \mathrm{Gamma}(10, 5)$

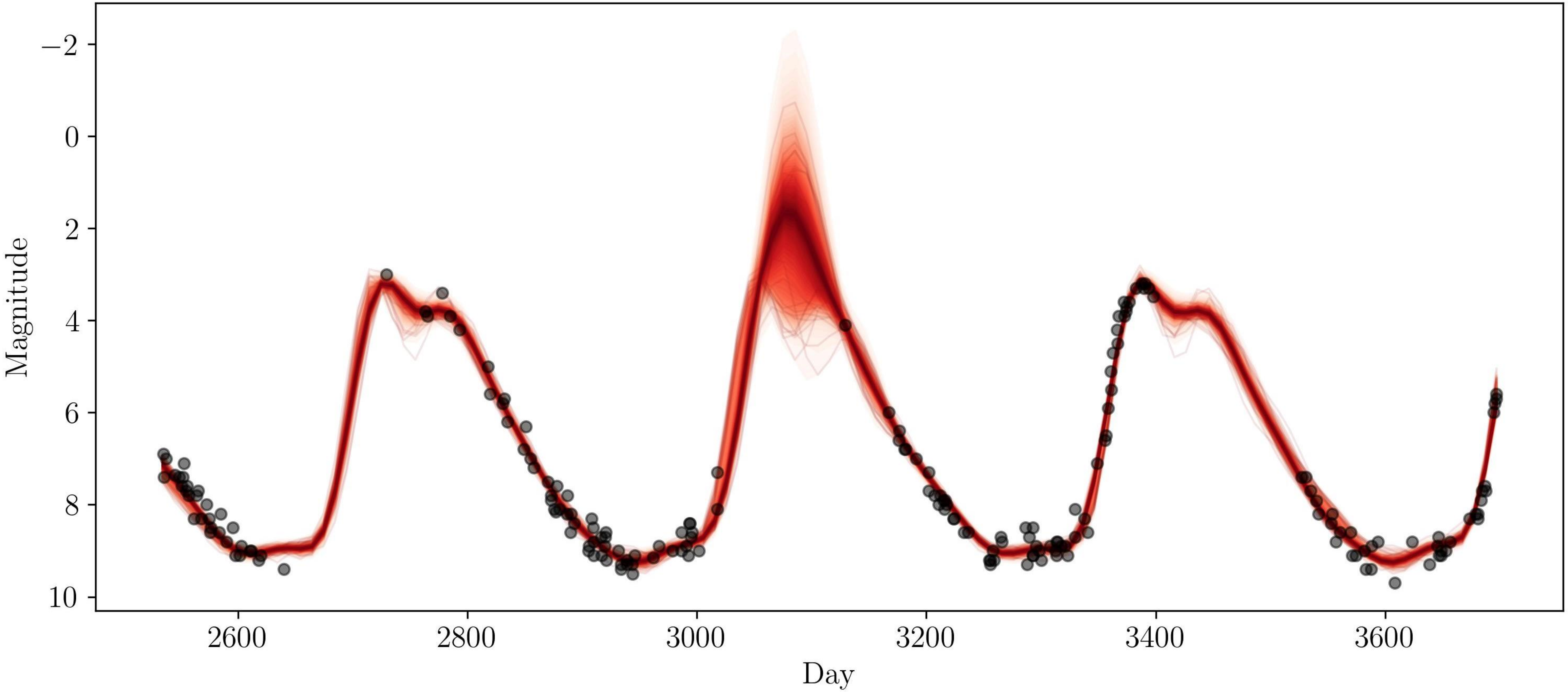



$T \sim \mathrm{LogNormal}(6, 0.25)$

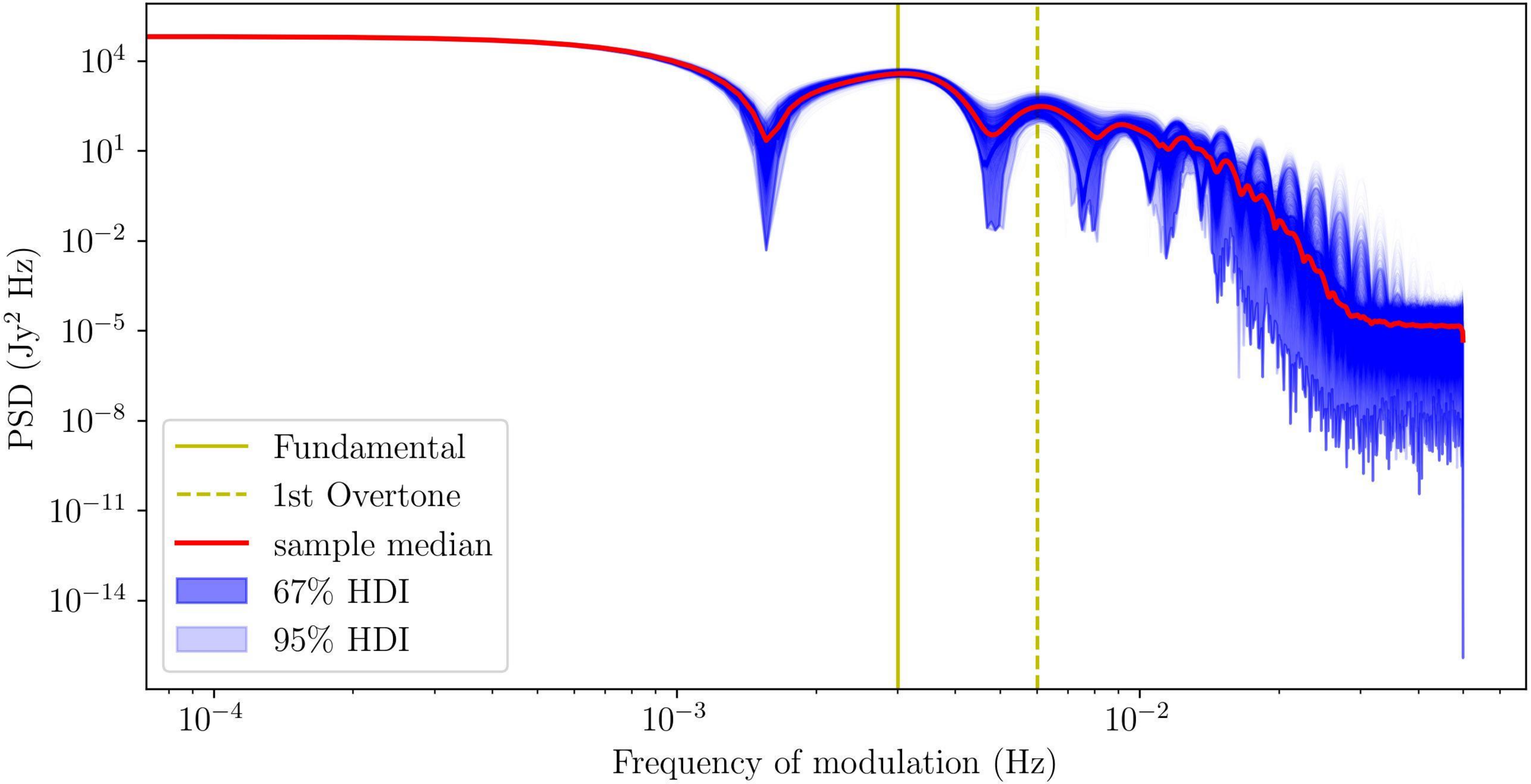



Hyperpriors

# Mira: Posterior Predictive

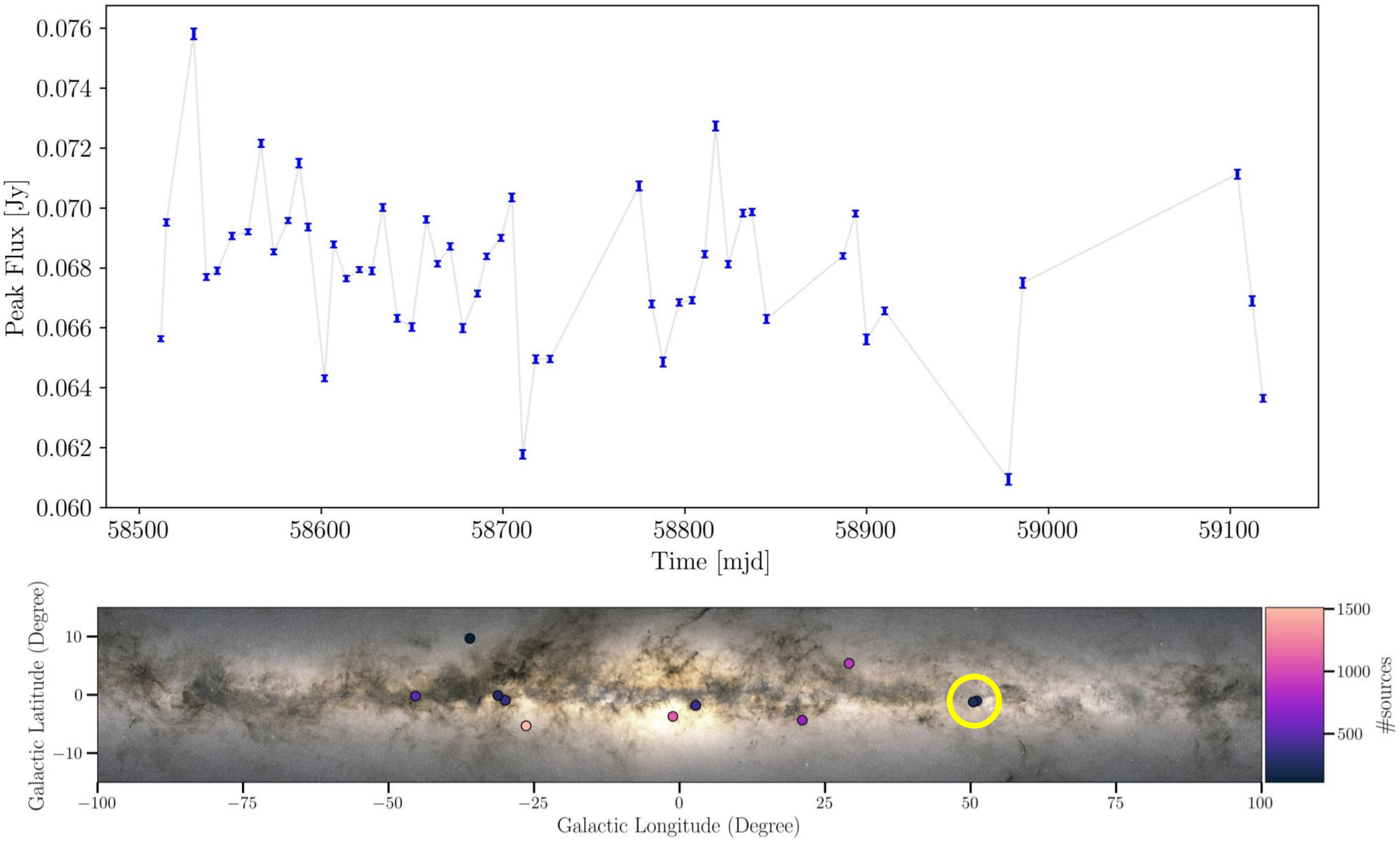


# Mira: PSD



PSD (Jy$^2$ Hz)
$10^4$
$10^1$
$10^{-2}$
$10^{-5}$
$10^{-8}$
$10^{-11}$
$10^{-14}$
Fundamental
1st Overtone
sample median
67% HDI
95% HDI
$10^{-4}$
$10^{-3}$
$10^{-2}$
Frequency of modulation (Hz)

# ThunderKAT Survey



# Model

## Likelihood + Noise

$$y(t) \sim N(f(t), \sigma^2)$$

## GP Prior

$$f(t) \sim \underbrace{GP_1(\bar{y}, k_1(\tau))}_{\text{Quasiperiodic}} + \underbrace{GP_2(\bar{y}, k_2(\tau))}_{\text{Noise}}$$

## Kernel

$$k_1(\tau) = \underbrace{\eta_1^2 \left[1 + \sqrt{5}\left(\frac{\tau}{\ell_1}\right) + \frac{5}{3}\left(\frac{\tau}{\ell_1}\right)^2\right] \exp\left\{-\sqrt{5}\left(\frac{\tau}{\ell_1}\right)\right\}}_{\text{Matern 5/2}} \times \underbrace{\exp\left\{-\frac{1}{2}\left[\frac{\sin\left(\pi \frac{\tau}{T}\right)}{\ell_p}\right]^2\right\}}_{\text{Periodic}}$$

$$k_2(\tau) = \underbrace{\eta_2^2 \left[1 + \sqrt{3}\left(\frac{\tau}{\ell_2}\right)\right] \exp\left\{-\sqrt{3}\left(\frac{\tau}{\ell_2}\right)\right\}}_{\text{Matern 3/2}}$$

# Hyperparameter Inference

## Priors

$\eta_1 \sim \mathrm{HalfNormal}(8)$

$\ell_1 \sim \mathrm{Gamma}(10, 0.1)$

$T \sim \mathrm{LogNormal}(3, 0.5)$

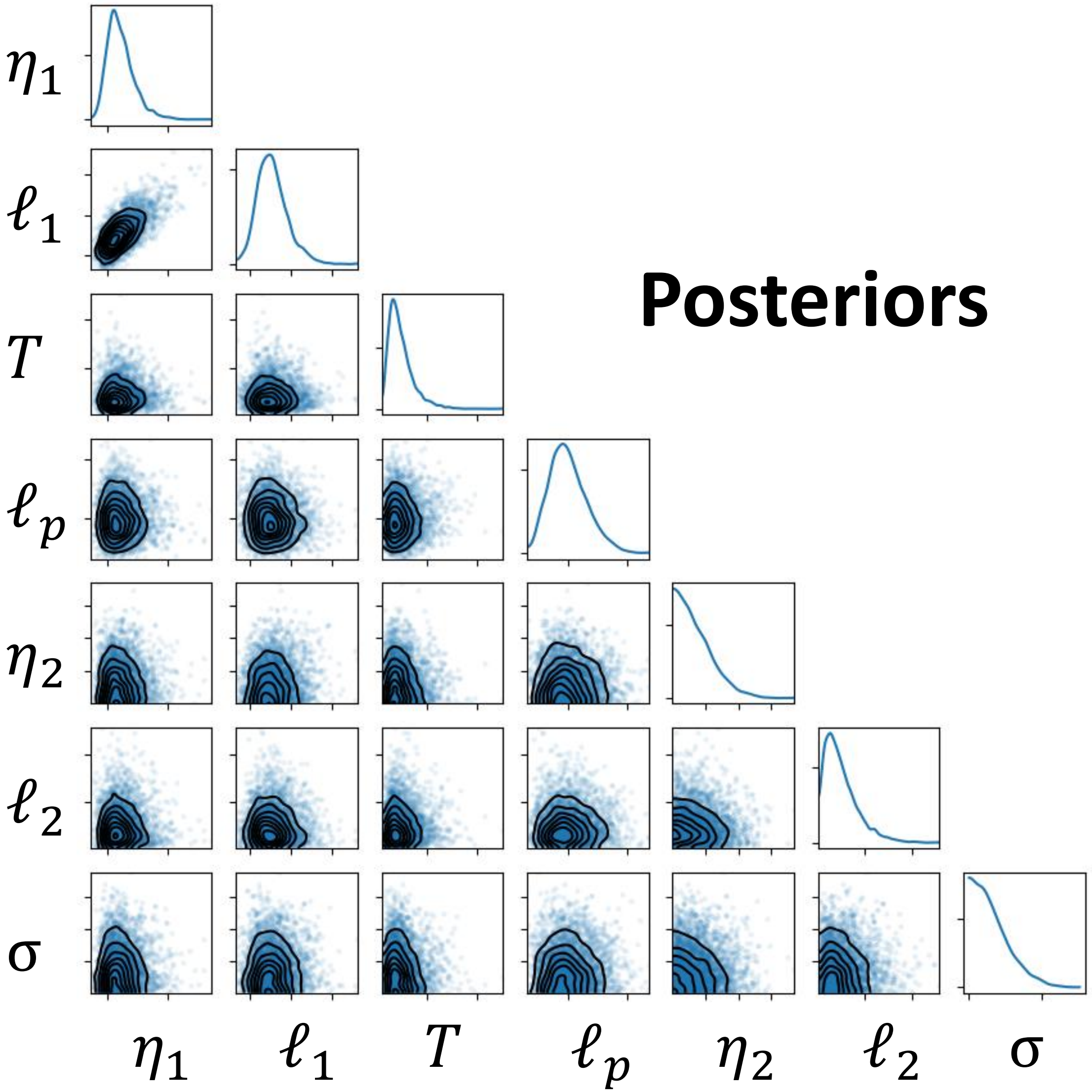
$\ell_p \sim \mathrm{Gamma}(10, 0.1)$

$\eta_2 \sim \mathrm{HalfNormal}(0.0002)$

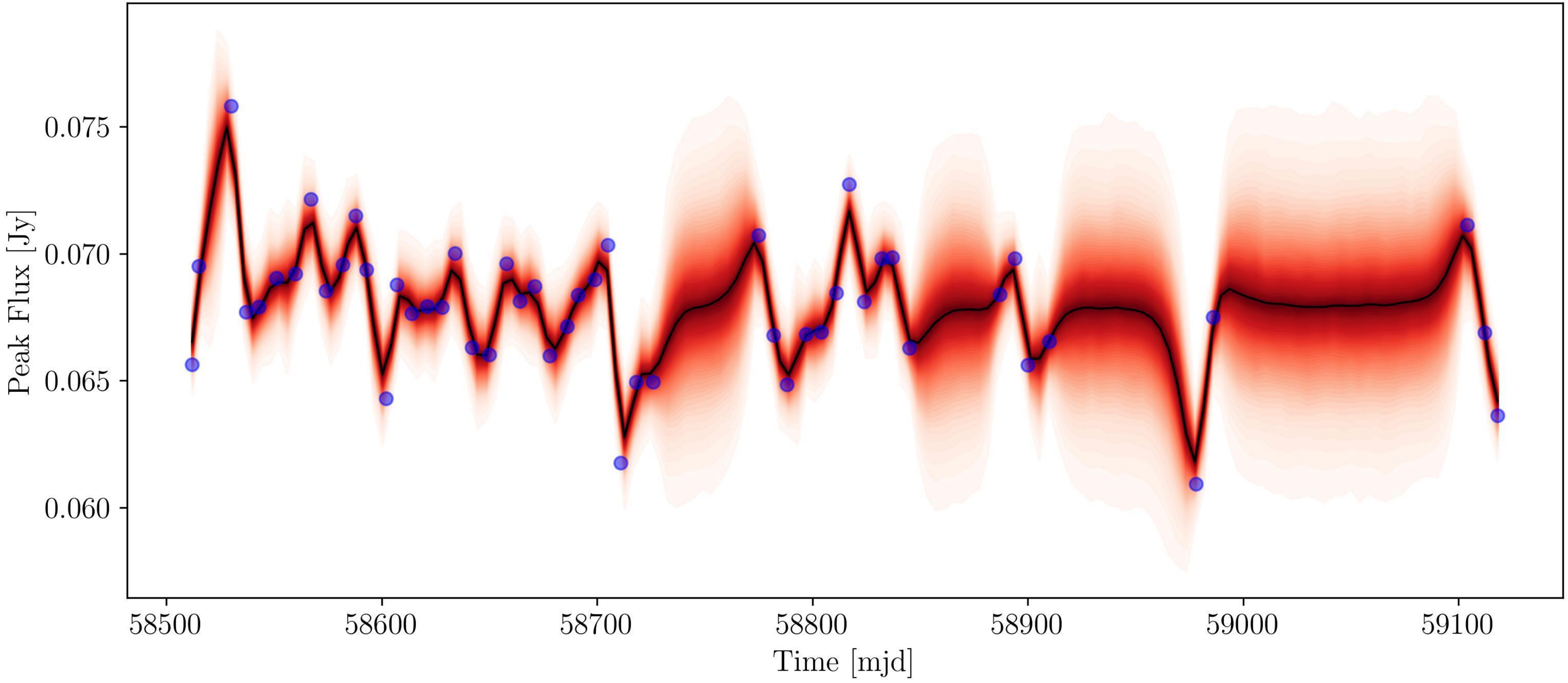
$\ell_2 \sim \mathrm{Gamma}(2, 4)$

$\sigma \sim \mathrm{HalfNormal}(0,1)$

## Posteriors



# Posterior Predictive Samples



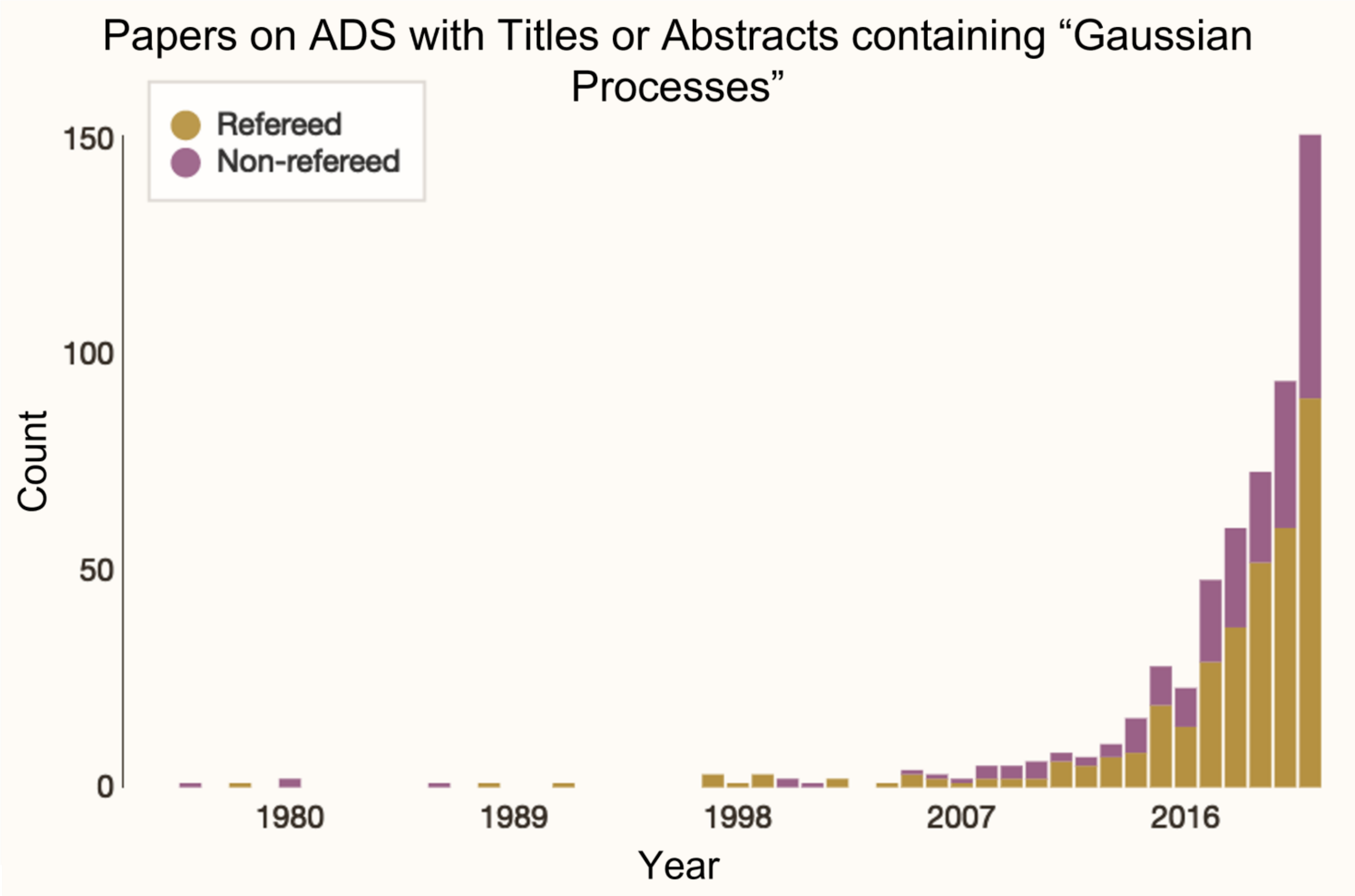
# Planned Work

| Project | Astronomy | Statistics |
|---|---|---|
| **Modelling radio light curves from ThunderKAT** | Identifying black hole candidates in commensal radio surveys in the SKA era | • Univariate Gaussian Processes<br>• Gaussian likelihood<br>• Sparse, unevenly sampled |
| **Modelling LSST light curves** | Identifying black hole candidates in multi-wavelength light curves across the optical band | • Multivariate Gaussian Processes<br>• High noise and nuisance artefacts |
| **Modelling light curves from large X-ray surveys (eROSITA, Swift)** | Characterisation of black hole accretion through light curve modelling | • Non-Gaussian likelihood<br>• Non-Gaussian noise |
| **Tools for GPs in Astronomy** | • Software (Python)<br>• Guidance for using GPs, e.g., kernels, hyperparameters, etc. | |

# Timeline

| Activity | | Months since 22 November 2022 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Activity | | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 24 | 27 | 30 | 33 | 36 |
| Literature review | | | | | | | | | | | | | |
| Learning software packages, e.g., PyMC | | | | | | | | | | | | | |
| Analysis of ThunderKAT survey data | | | | | | | | | | | | | |
| Papers | I | | | | | | | | | | | | |
| | II | | | | | | | | | | | | |
| | III | | | | | | | | | | | | |
| | IV | | | | | | | | | | | | |
| Thesis preparation | Introduction & Background | | | | | | | | | | | | |
| | Methodology | | | | | | | | | | | | |
| | Paper I | | | | | | | | | | | | |
| | Paper II | | | | | | | | | | | | |
| | Paper III | | | | | | | | | | | | |
| | Paper IV | | | | | | | | | | | | |
| | Discussion & Conclusions | | | | | | | | | | | | |

# Questions & Responses



# Tools

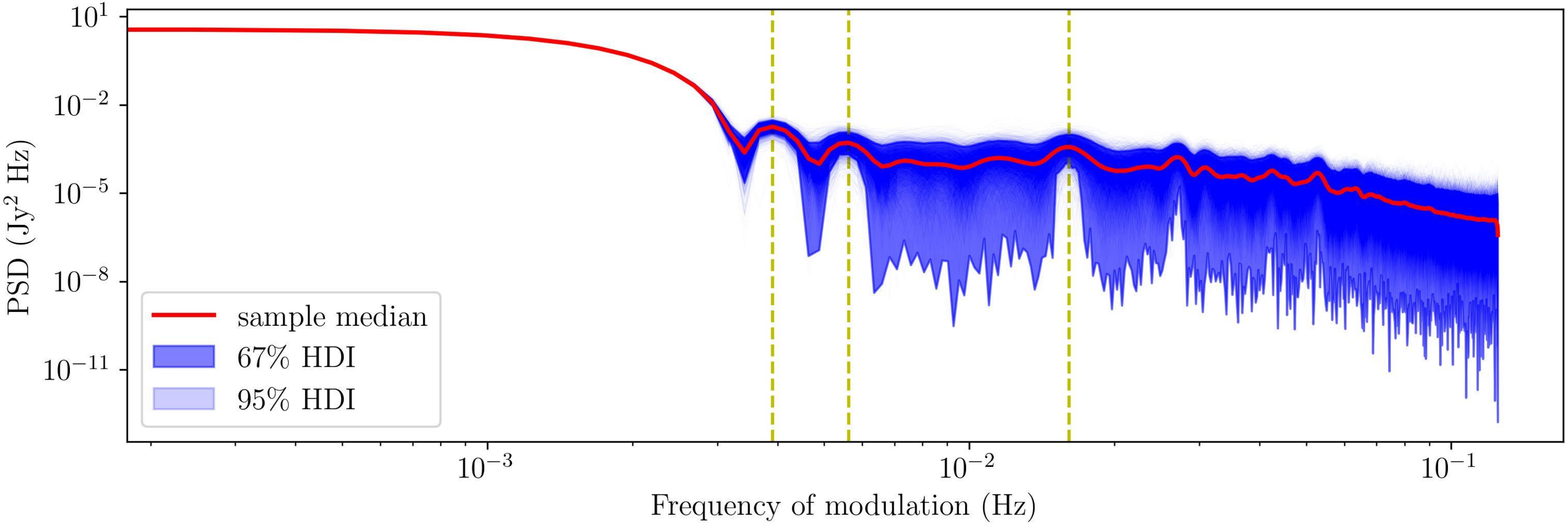
Chosen to use `Python`[1] and `PyMC`[2] for this work.

- Accessible to astronomers
- Probabilistic programming framework
- Well-maintained open-source software

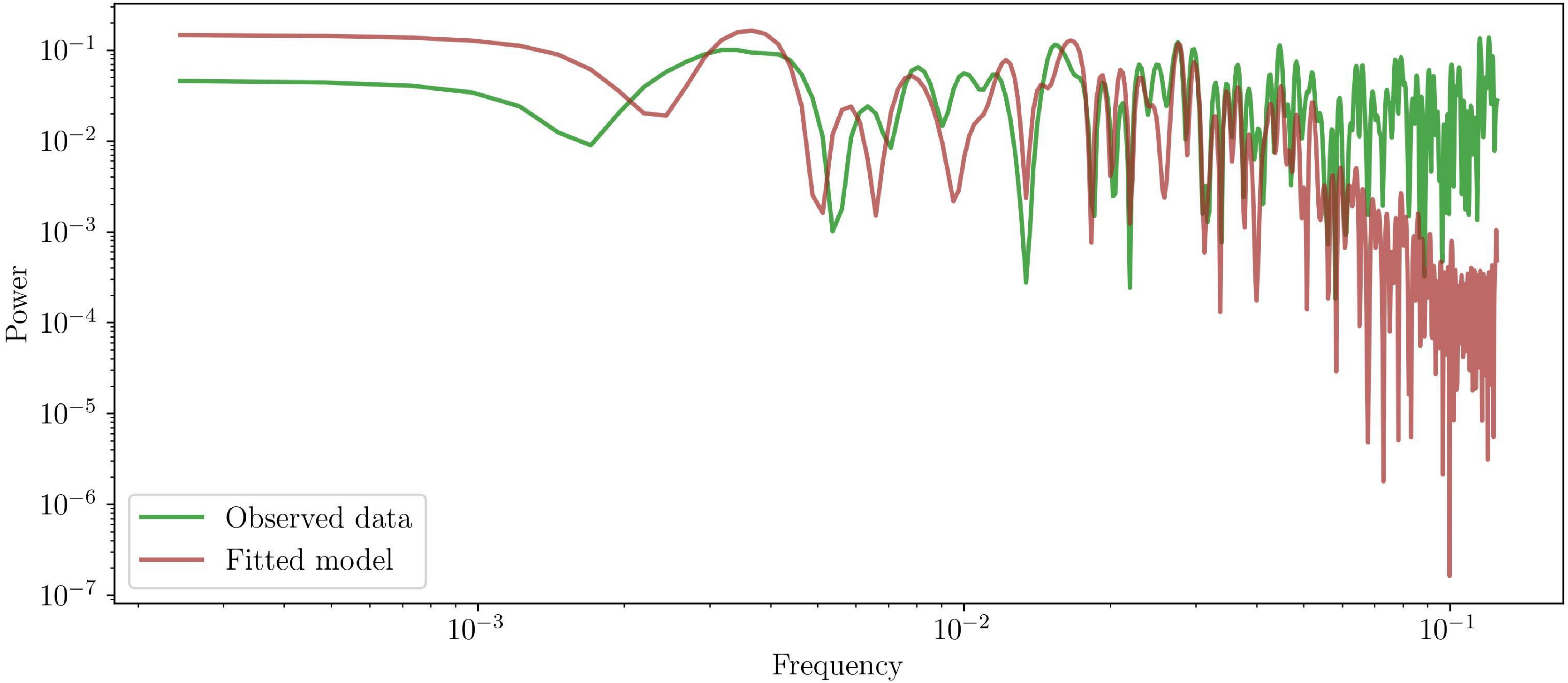
Considered: `R`[3], `Stan`[4], `celerite2`[5], `george`[6].

1. https://www.python.org
2. https://www.pymc.io
3. https://cran.r-project.org/
4. https://mc-stan.org/
5. https://celerite2.readthedocs.io/en/latest/
6. https://george.readthedocs.io/en/latest/

# Posterior Predictive PSD



# Lomb-Scargle Periodogram



# Modelling Workflow