Southern California
SCGE
Center for Galaxy Evolution

Credit: ESO/Kornmesser

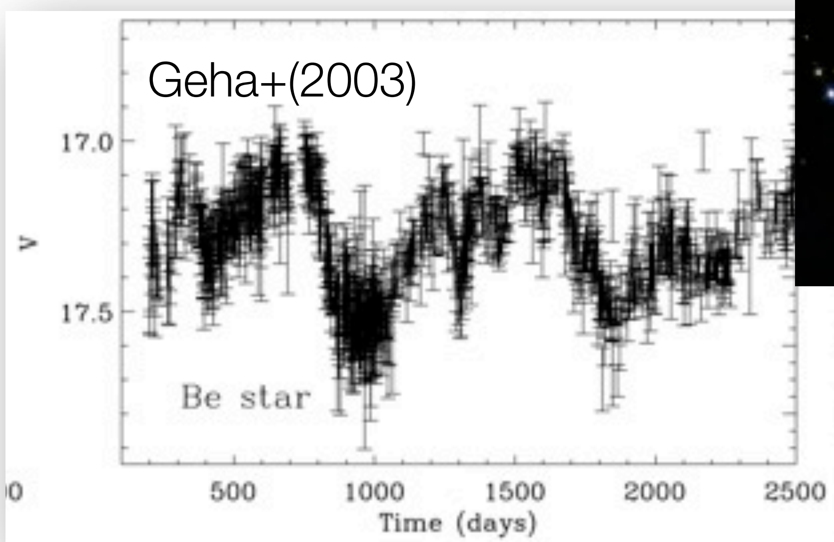
# Stochastic Modeling of Astronomical Time Series

Brandon C. Kelly (UCSB, CGE Fellow, bckelly@physics.ucsb.edu)

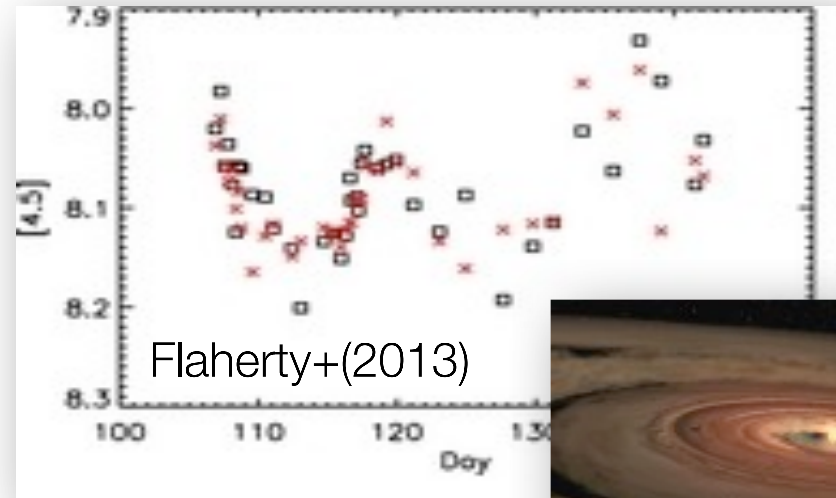
Aneta Siemiginowska (CfA), Malgosia Sobolewska (CfA), Andy Becker (UW), Tommaso Treu (UCSB), Matt Malkan (UCLA), Anna Pancoast (UCSB), Jong-Hak Woo (Seoul), Jill Bechtold (Arizona)

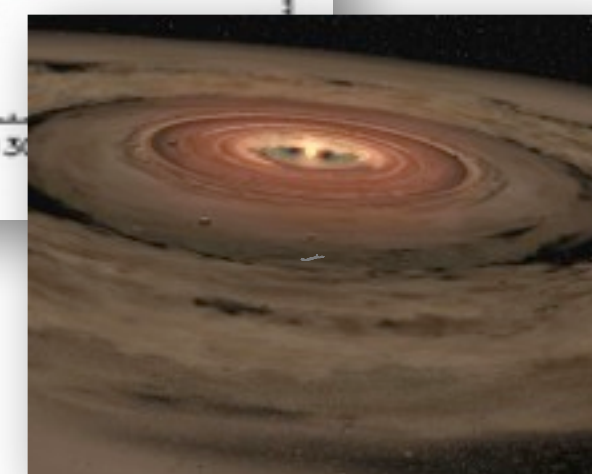
# Aperiodic and Quasi-Periodic Lightcurves (Time Series of Brightness)

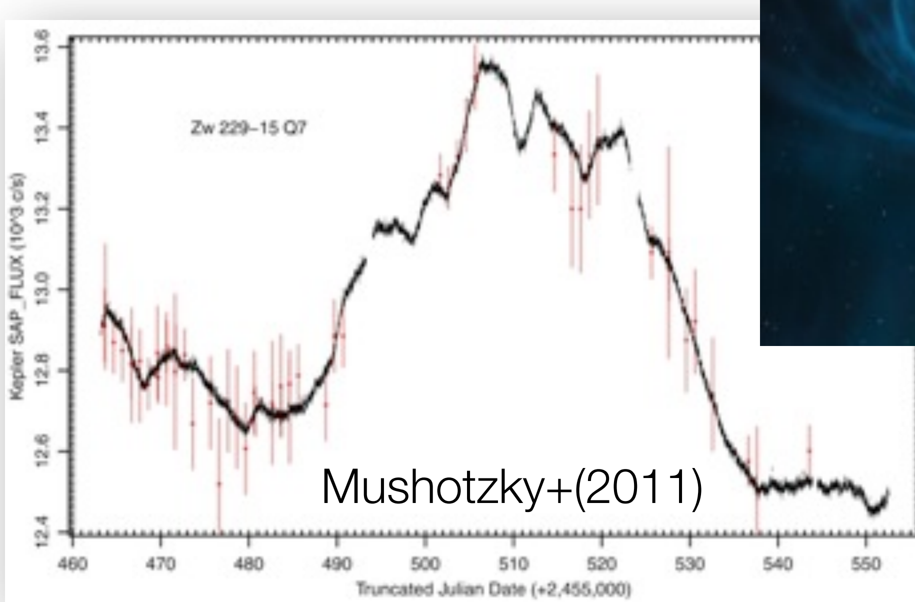
Variable Stars




Protostars






Quasars

# Current and Future Data Sets

# Current and Future Data Sets

- SDSS Stripe 82 (~1998-2008)
    - ugriz (5-d lightcurves), down to r ~ 20, ~60 epochs, ~10,000 quasars

# Current and Future Data Sets

- SDSS Stripe 82 (~1998-2008)
    - ugriz (5-d lightcurves), down to r ~ 20, ~60 epochs, ~10,000 quasars
- Pan-STARRS Medium-Deep Survey (2010-2013)
    - griz, r ~ 23, 300-400 epochs, ~ 7,000 quasars

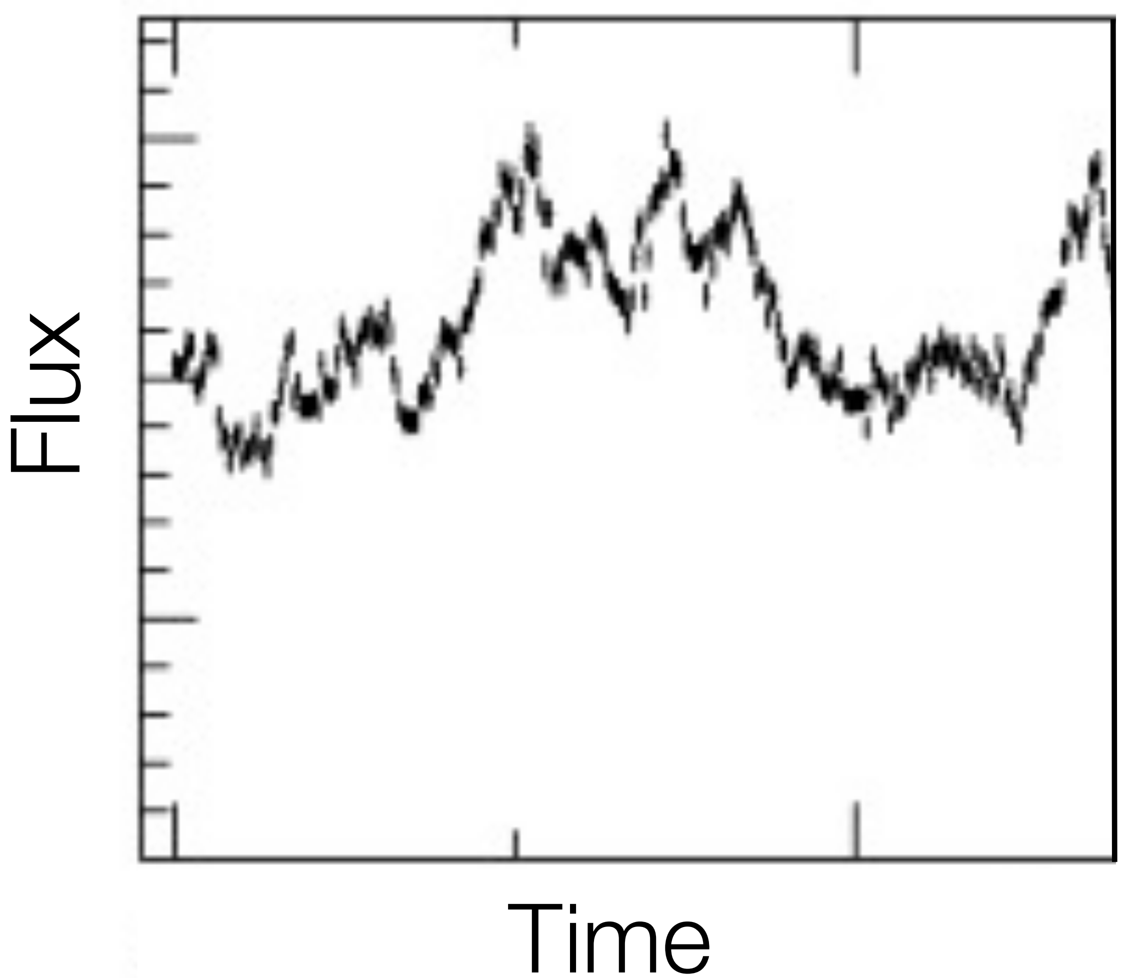
# Current and Future Data Sets

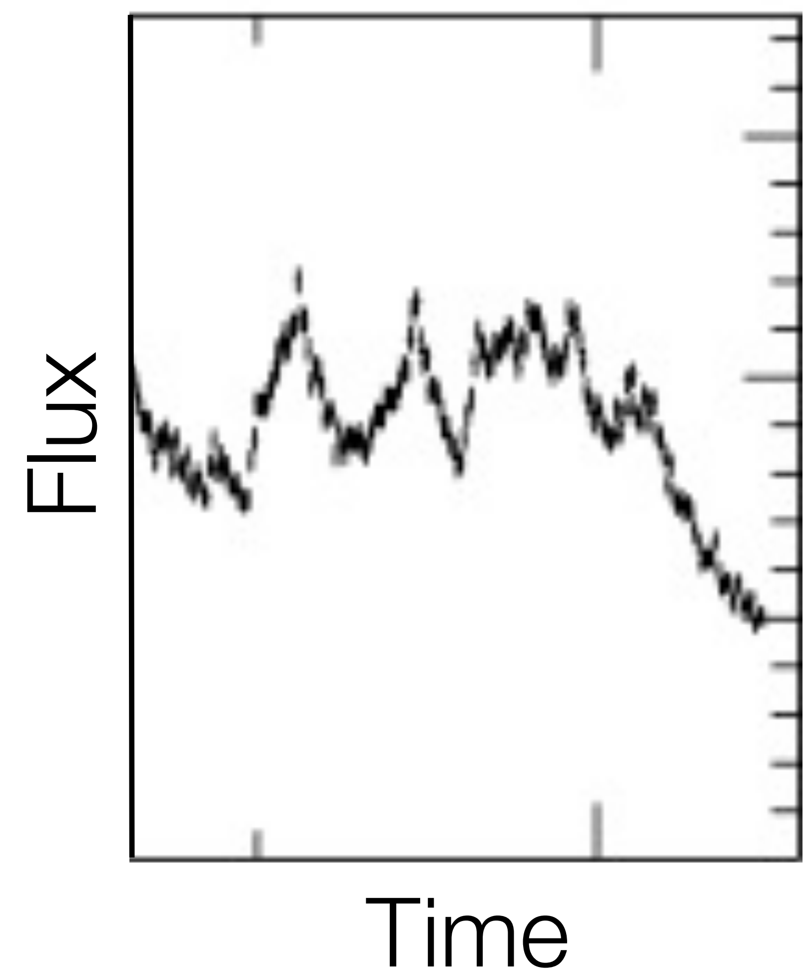
- SDSS Stripe 82 (~1998-2008)
  - ugriz (5-d lightcurves), down to r ~ 20, ~60 epochs, ~10,000 quasars
- Pan-STARRS Medium-Deep Survey (2010-2013)
  - griz, r ~ 23, 300-400 epochs, ~ 7,000 quasars
- DES Supernovae Survey (2012-2017)
  - griz, r ~ 25, ~ 100 epochs, ~ 15,000 quasars
  - Overlaps with Stripe 82 and 2 MDS fields

# Current and Future Data Sets

- SDSS Stripe 82 (~1998-2008)
    - ugriz (5-d lightcurves), down to r ~ 20, ~60 epochs, ~10,000 quasars
- Pan-STARRS Medium-Deep Survey (2010-2013)
    - griz, r ~ 23, 300-400 epochs, ~ 7,000 quasars
- DES Supernovae Survey (2012-2017)
    - griz, r ~ 25, ~ 100 epochs, ~ 15,000 quasars
    - Overlaps with Stripe 82 and 2 MDS fields
- Large Synoptic Survey Telescope (LSST) (2021-2031?)
    - ugrizy, r ~ 24.5, ~ 50-200 epochs (more in ‘deep drilling fields’), millions of quasars

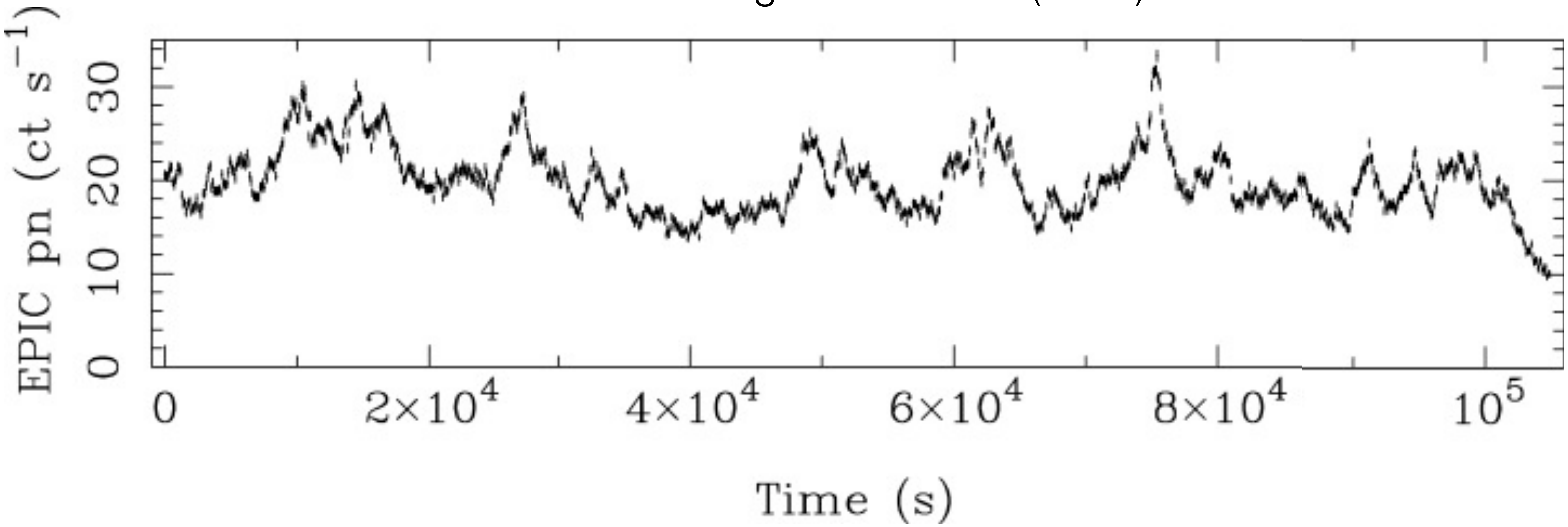
# The Data Analysis Challenge: Aperiodic Lightcurves



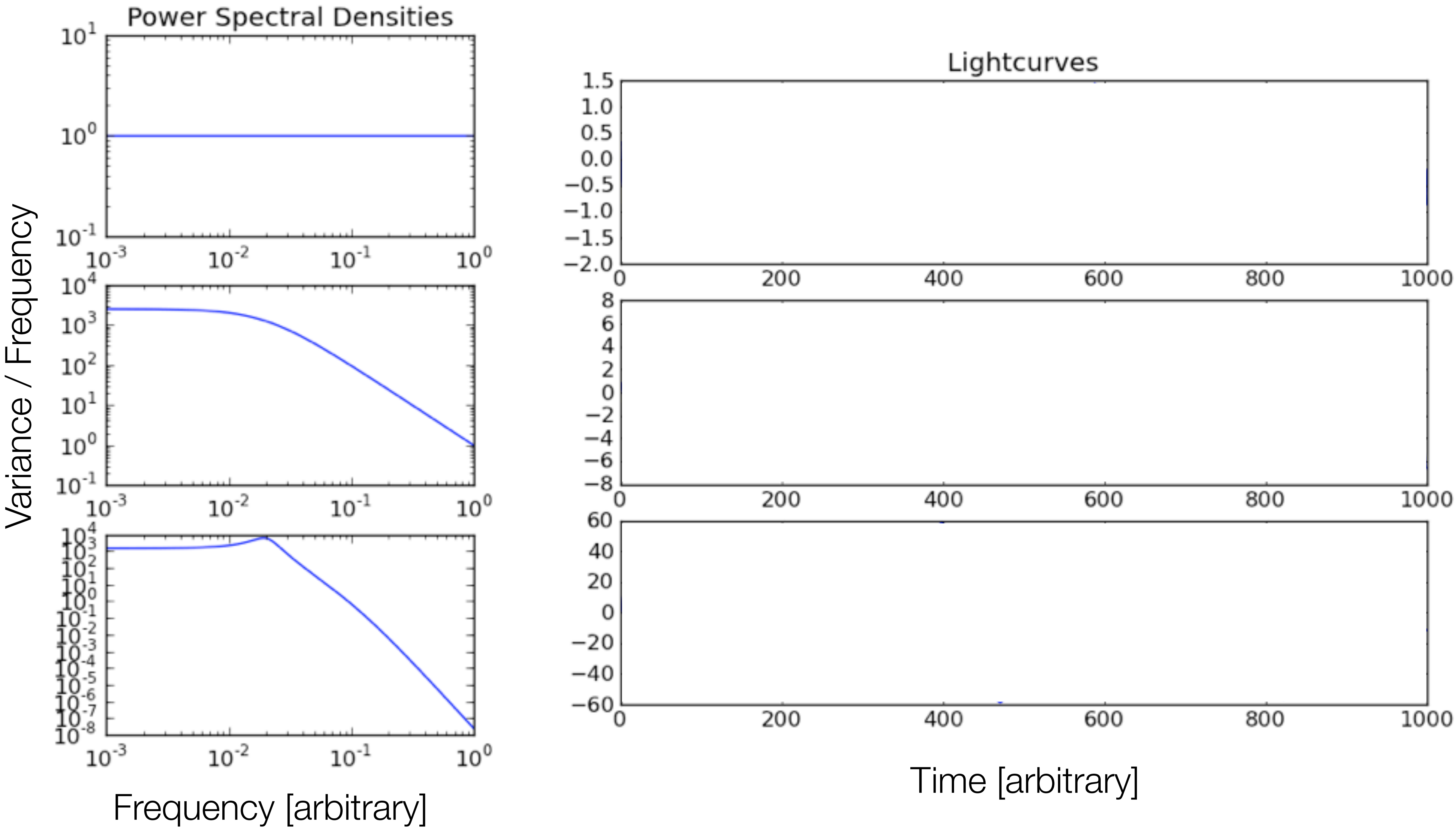
# The Data Analysis Challenge: Aperiodic Lightcurves

Mrk 766 Vaughan & Fabian (2003)

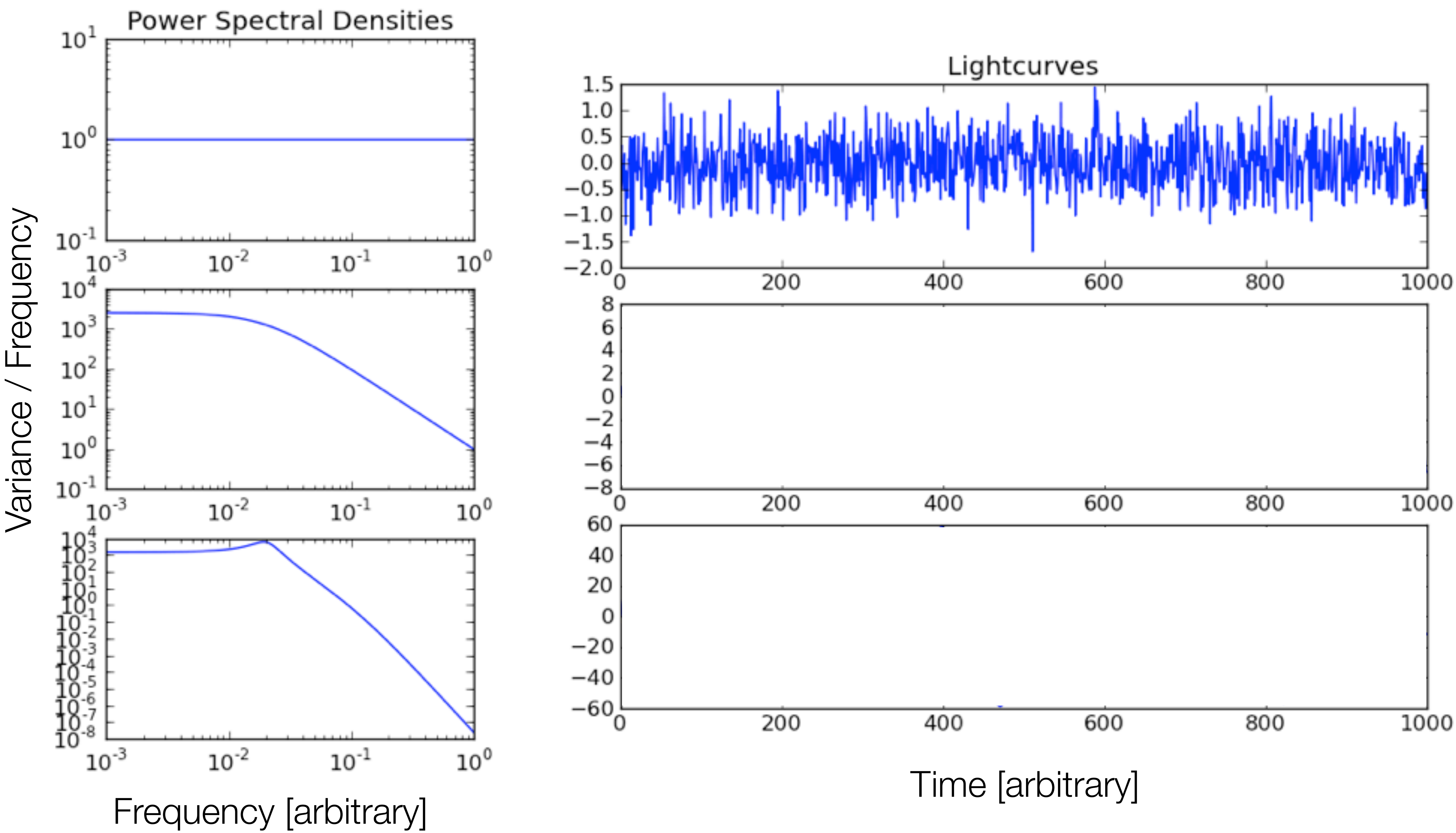



What variability 'features' can we measure for quasar lightcurves?

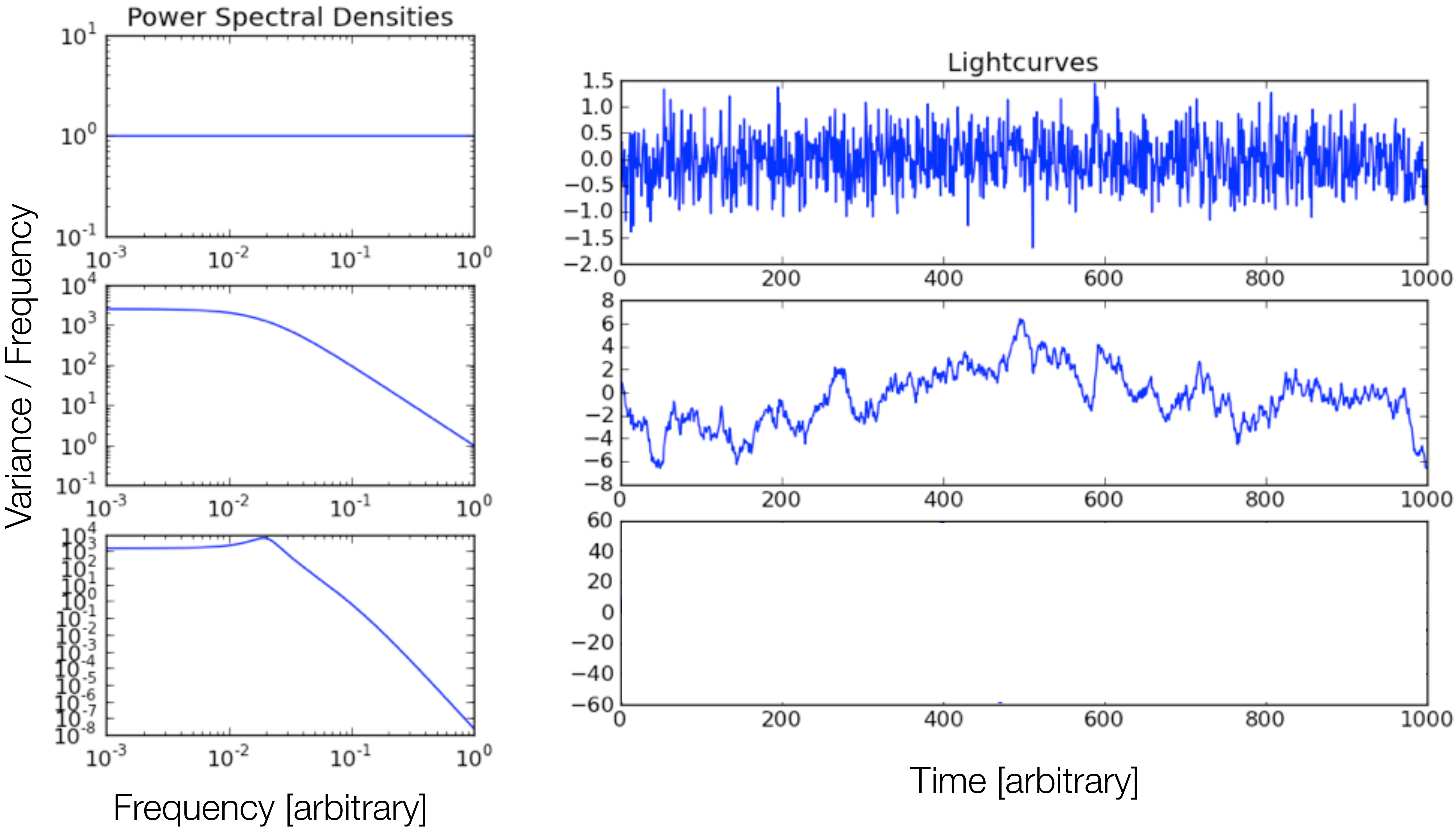
# Quantifying Variability with the Power Spectral Density (PSD)



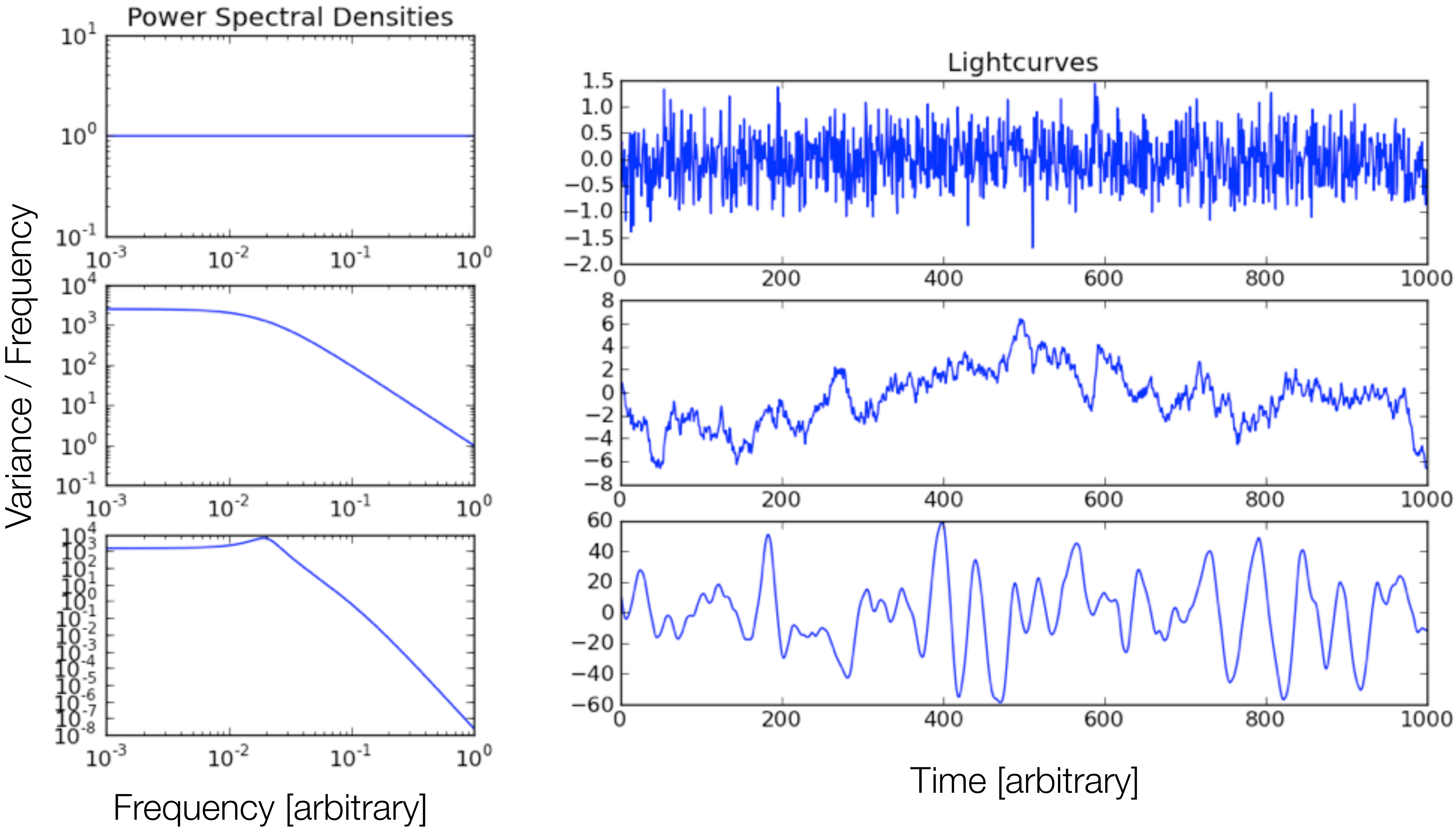
# Quantifying Variability with the Power Spectral Density (PSD)



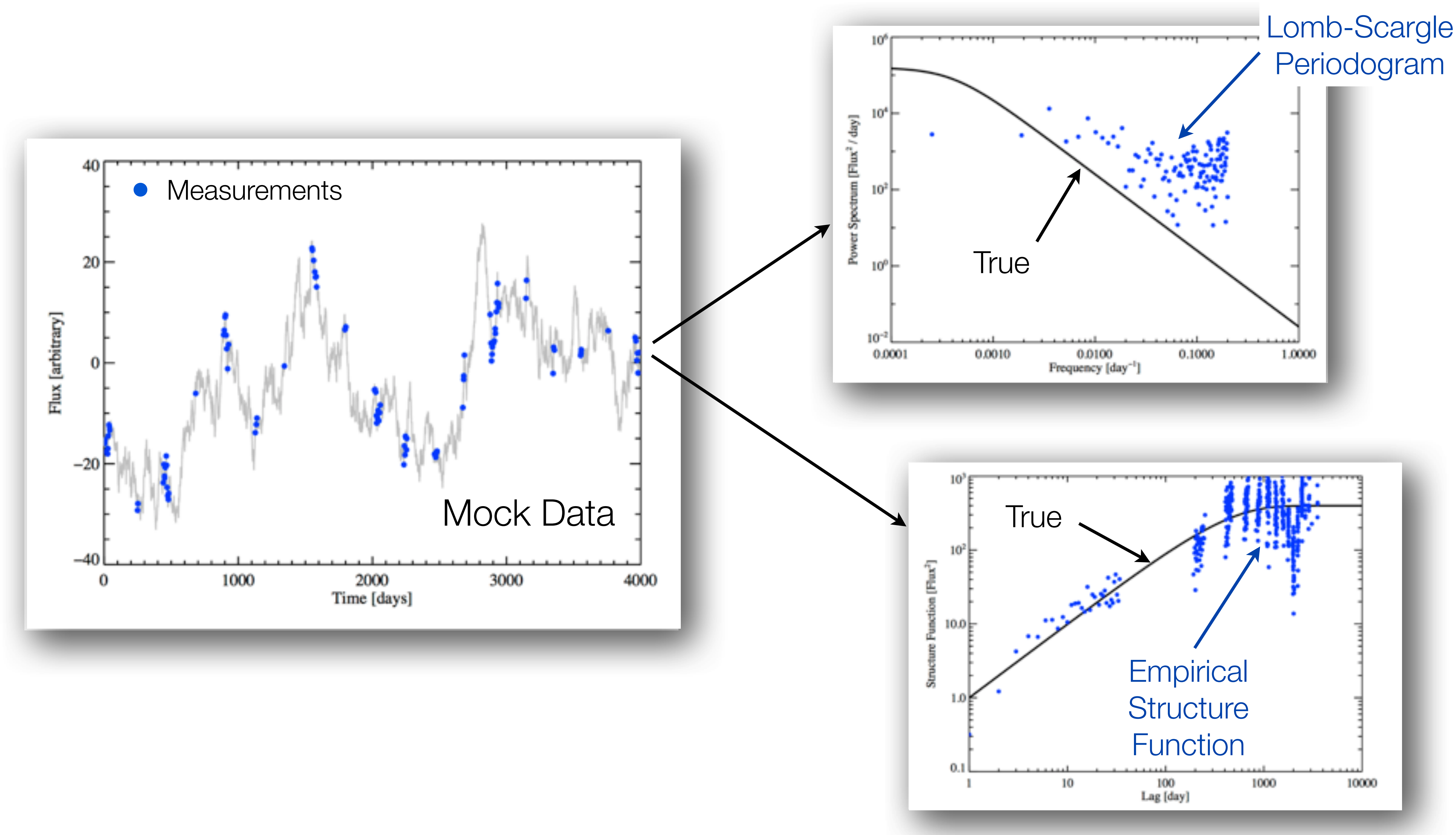
# Quantifying Variability with the Power Spectral Density (PSD)



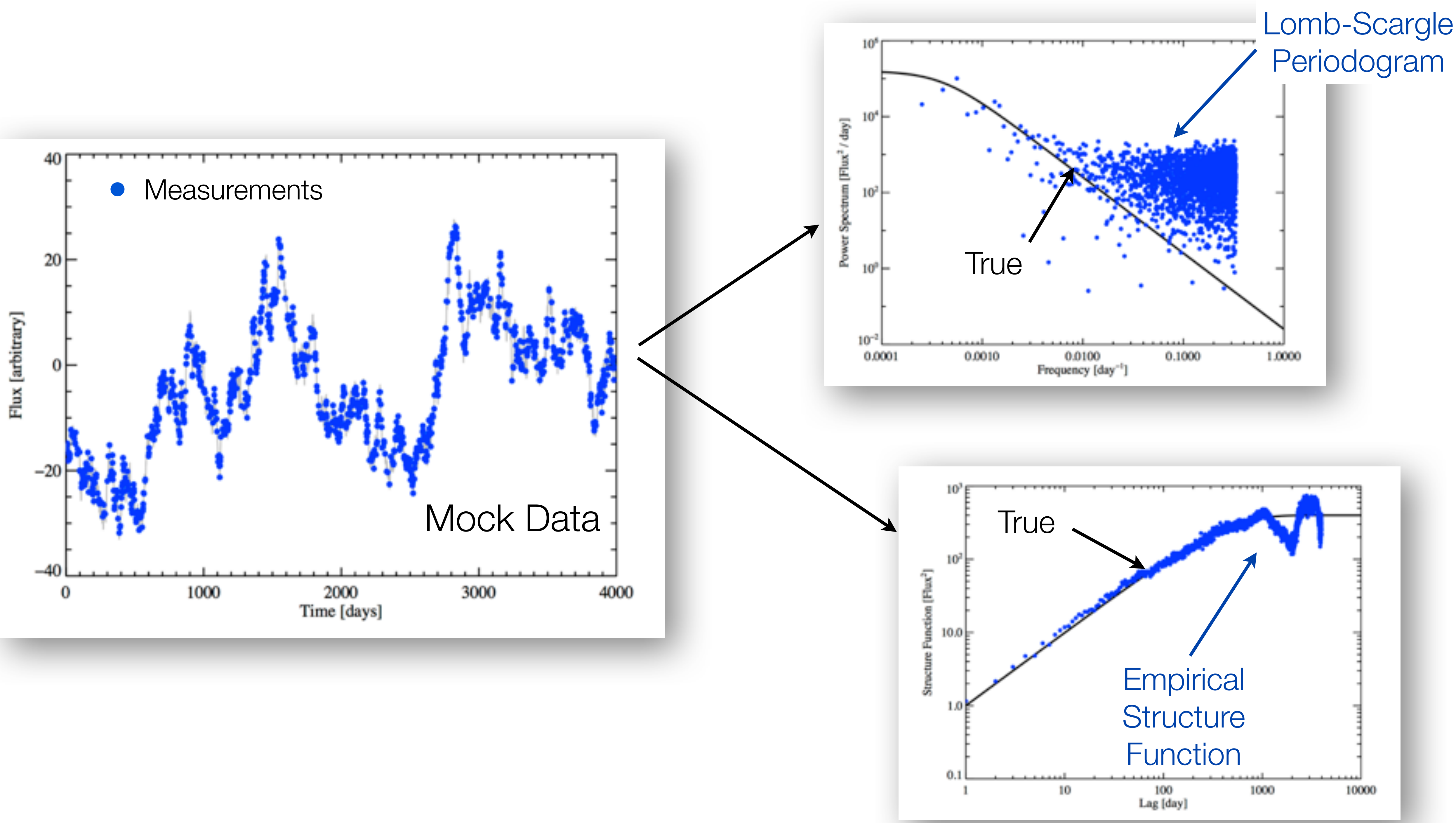
# Quantifying Variability with the Power Spectral Density (PSD)



# Disadvantages of Traditional Non-parameteric Tools for Quantifying Aperiodic Variability



# Disadvantages of Traditional Non-parameteric Tools for Quantifying Aperiodic Variability



# Tools for Characterizing Aperiodic (Quasar) Variability: What should they do?

# Tools for Characterizing Aperiodic (Quasar) Variability: What should they do?

- Handle irregular/arbitrary sampling patterns and measurement errors

# Tools for Characterizing Aperiodic (Quasar) Variability: What should they do?

- Handle irregular/arbitrary sampling patterns and measurement errors
- Produce interpretable results:
    - Connection to physical models
    - Connection to features in the power spectrum

# Tools for Characterizing Aperiodic (Quasar) Variability: What should they do?

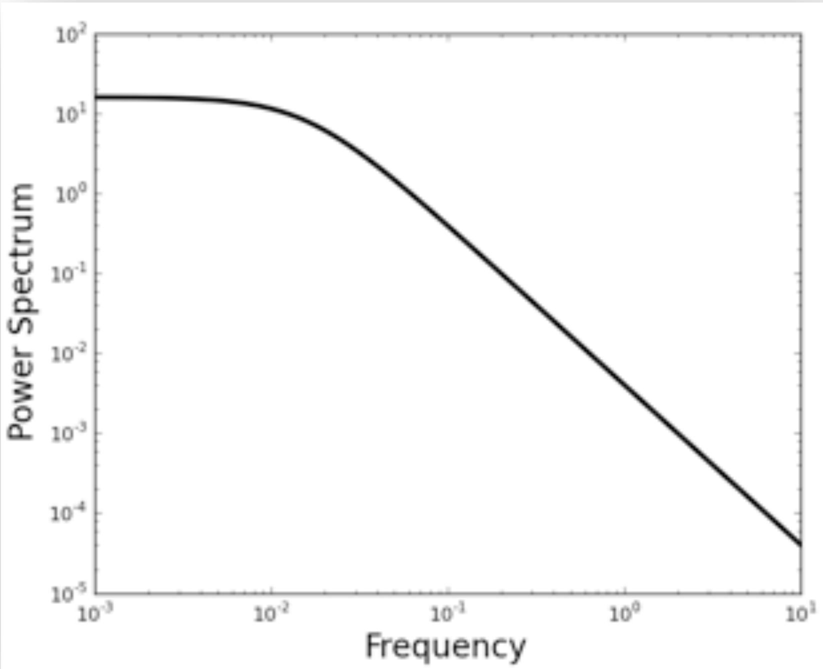
- Handle irregular/arbitrary sampling patterns and measurement errors
- Produce interpretable results:
  - Connection to physical models
  - Connection to features in the power spectrum
- Fast & scalable to massive time domain surveys
  - By the end of LSST we will have billions of multiwavelength lightcurves with ~ 50-1000+ epochs

# Tools for Characterizing Aperiodic (Quasar) Variability: What should they do?

- Handle irregular/arbitrary sampling patterns and measurement errors
- Produce interpretable results:
  - Connection to physical models
  - Connection to features in the power spectrum
- Fast & scalable to massive time domain surveys
  - By the end of LSST we will have billions of multiwavelength lightcurves with ~ 50-1000+ epochs
- Handle multiwavelength/multivariate time series
  - Account for correlations/time lags among lightcurves in different bands

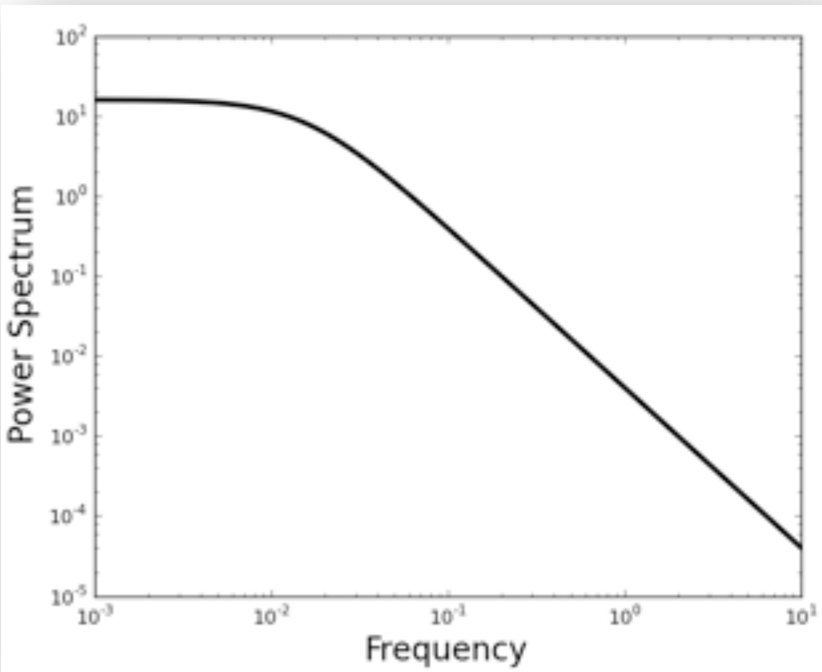
# Two approaches to (stochastic) modeling of real lightcurves: Frequency Domain and Time Domain

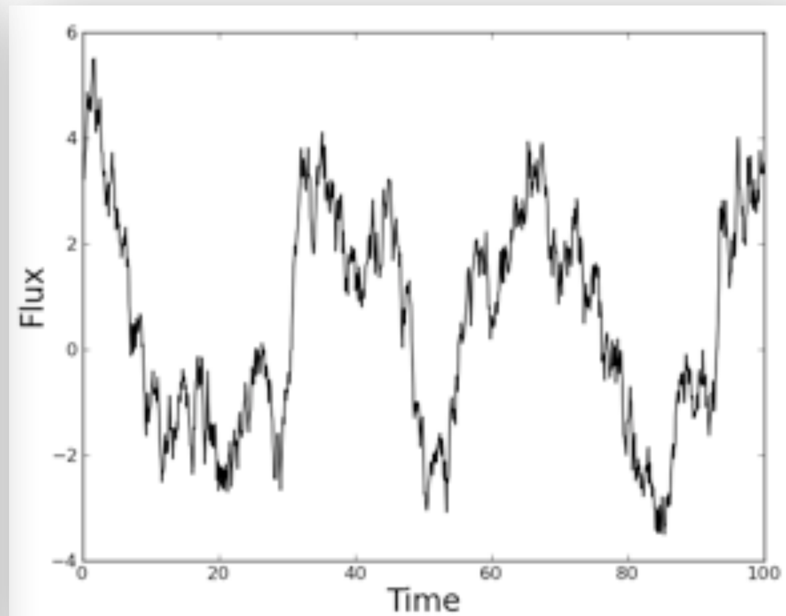
# Monte-Carlo Methods (Done+1992,Uttley +2002,Emmanaloupolous 2010,2013)




- Extremely flexible, limited only by ability to do simulation
- Can be computationally expensive
- Reliance on $\chi^2$ may not provide optimal use of information in lightcurve

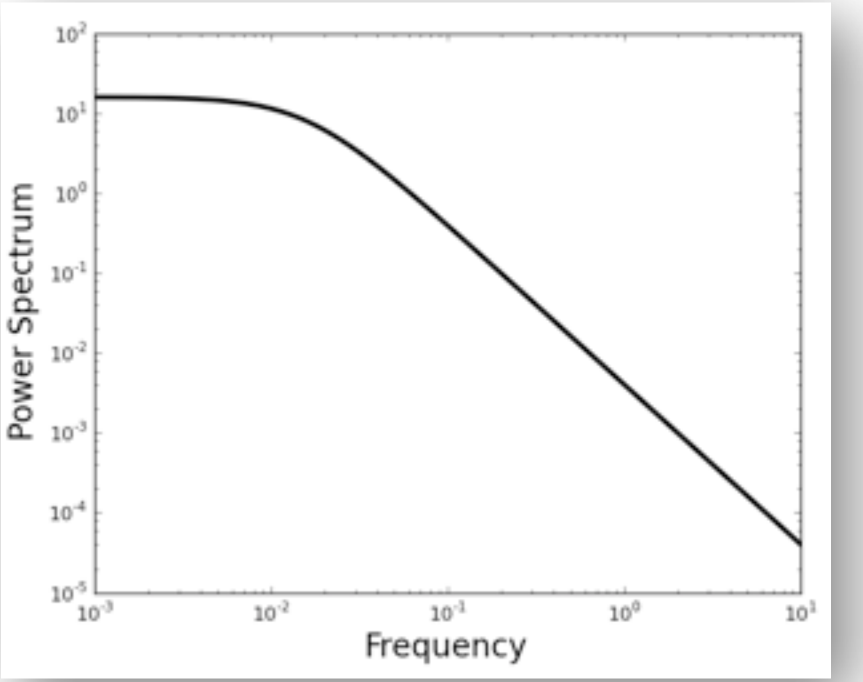
# Monte-Carlo Methods (Done+1992,Uttley +2002,Emmanaloupolous 2010,2013)

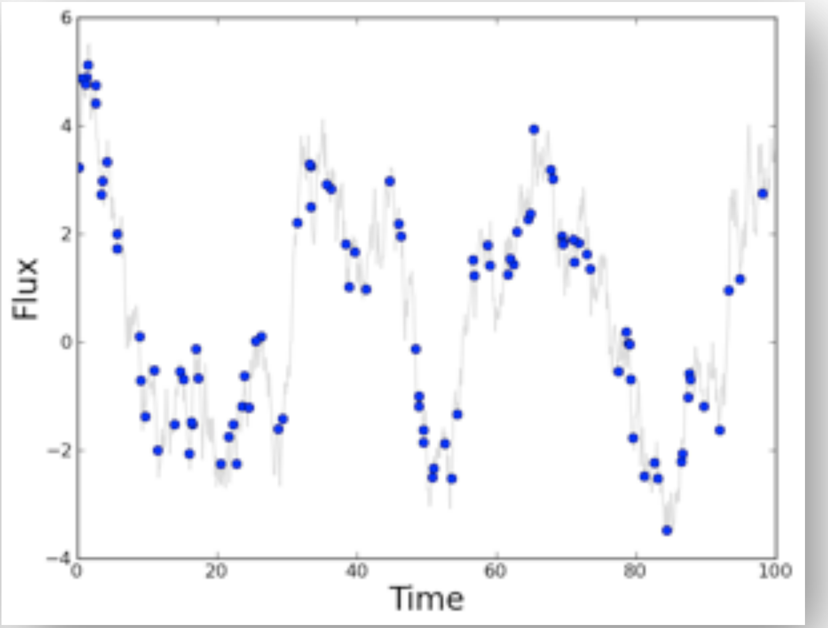



- Extremely flexible, limited only by ability to do simulation
- Can be computationally expensive
- Reliance on $\chi^2$ may not provide optimal use of information in lightcurve

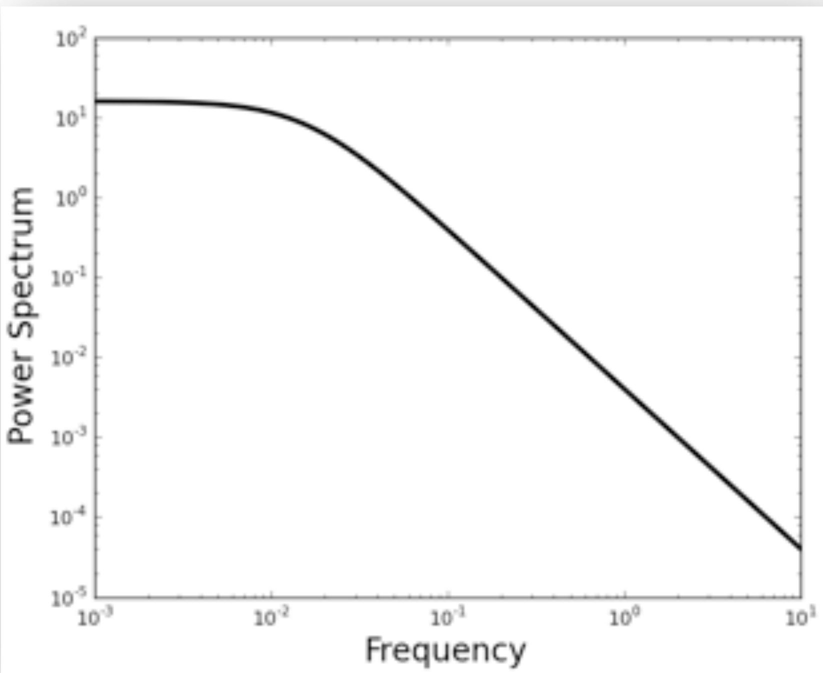
# Monte-Carlo Methods (Done+1992,Uttley +2002,Emmanaloupolous 2010,2013)

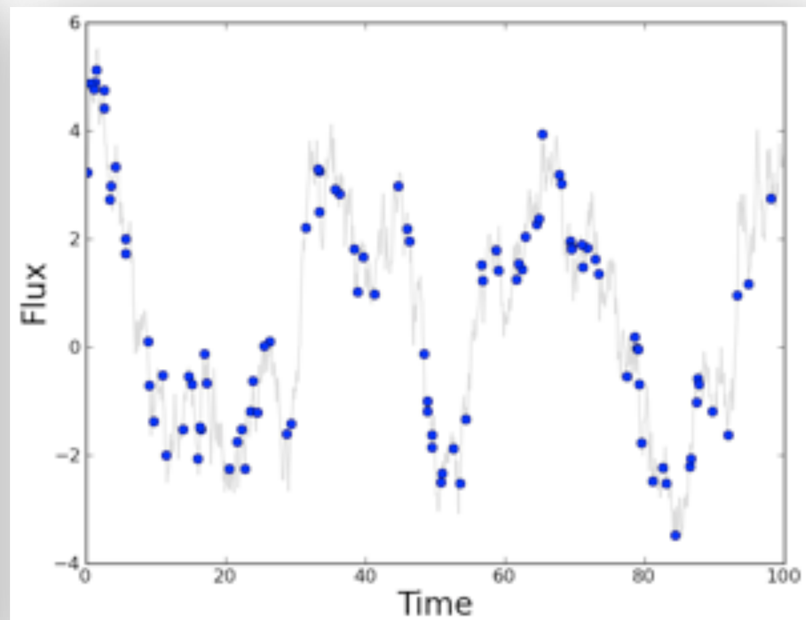


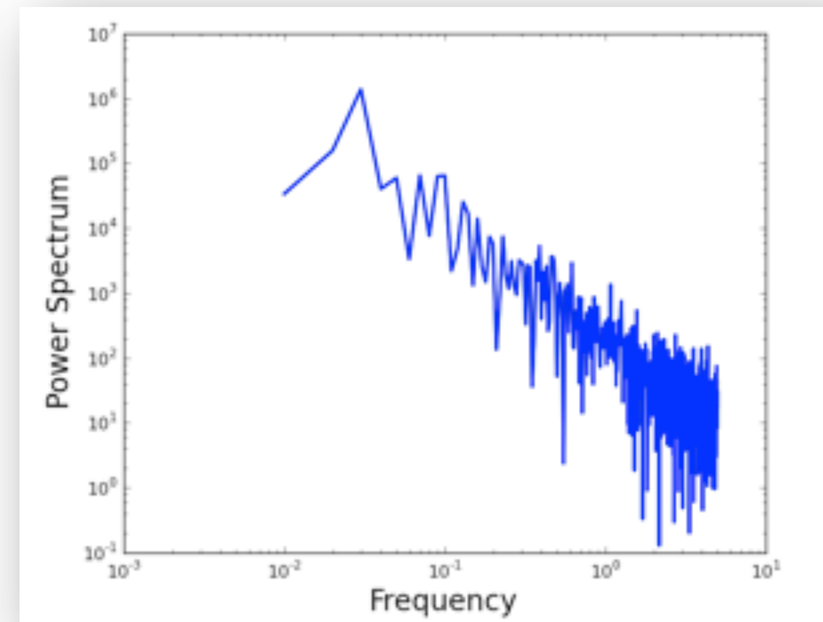




- Extremely flexible, limited only by ability to do simulation
- Can be computationally expensive
- Reliance on $\chi^2$ may not provide optimal use of information in lightcurve

# Monte-Carlo Methods (Done+1992,Uttley +2002,Emmanaloupolous 2010,2013)




- Extremely flexible, limited only by ability to do simulation
- Can be computationally expensive
- Reliance on $\chi^2$ may not provide optimal use of information in lightcurve

# Gaussian Processes (Rybicki & Press 1992, Kelly+ 2009,2011, Miller+2010)

$$\mathrm{loglik} = -\log|\Sigma| - \frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)$$

- Likelihood-based approach, enables Bayesian inference
- Statistically powerful, but limited by Gaussian assumption
- In general, computationally expensive ($O(n^3)$)

# Gaussian Processes (Rybicki & Press 1992, Kelly+ 2009,2011, Miller+2010)

$$\mathrm{loglik} = -\log|\Sigma| - \frac{1}{2}(\mathbf{y}-\mu)^T\Sigma^{-1}(\mathbf{y}-\mu)$$

$$\Sigma = \left( \quad n \times n \quad \right)$$

$$\Sigma_{ij} = \mathrm{Cov}(y(t_i), y(t_j))$$
$$= \int_{-\infty}^{\infty} PSD(f) e^{2\pi i f|t_i - t_j|} df$$

- Likelihood-based approach, enables Bayesian inference
- Statistically powerful, but limited by Gaussian assumption
- In general, computationally expensive ($O(n^3)$)

# Simple and fast tool: First order continuous autoregressive process (CAR(1), Kelly+2009)

Lightcurve

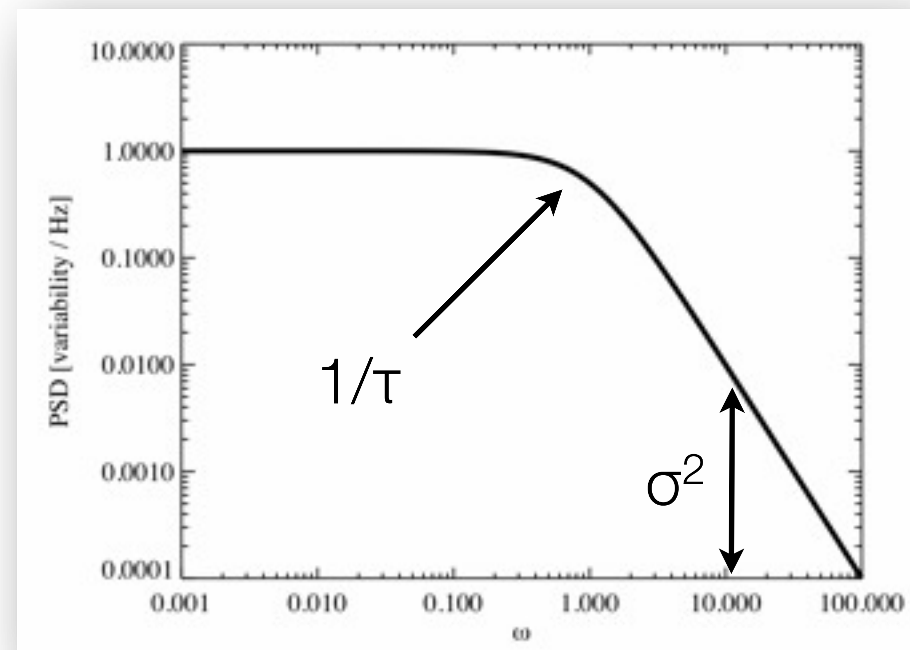
White Noise

$$dL(t) = -\frac{dt}{\tau}(L(t) - \mu) + \sigma dW(t)$$

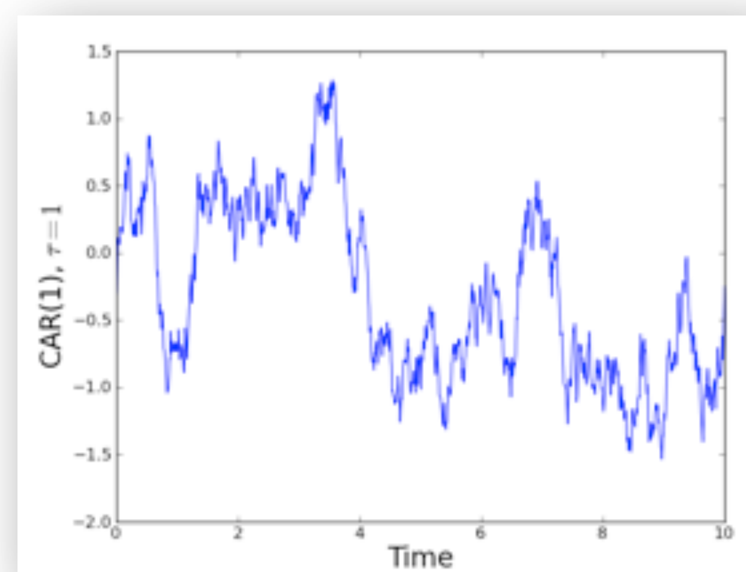
Characteristic Time Scale

LC Mean

Power Spectrum

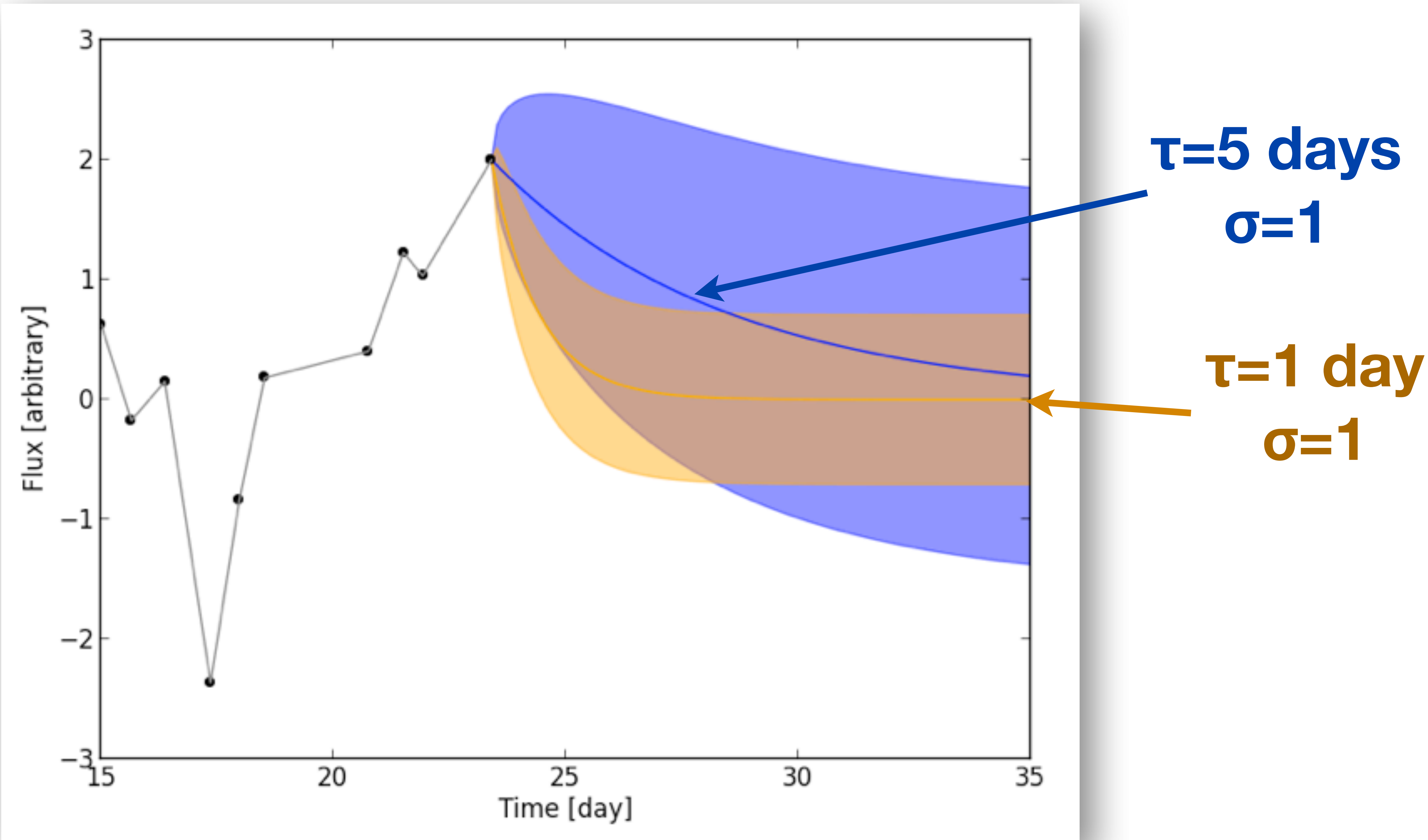



Frequency

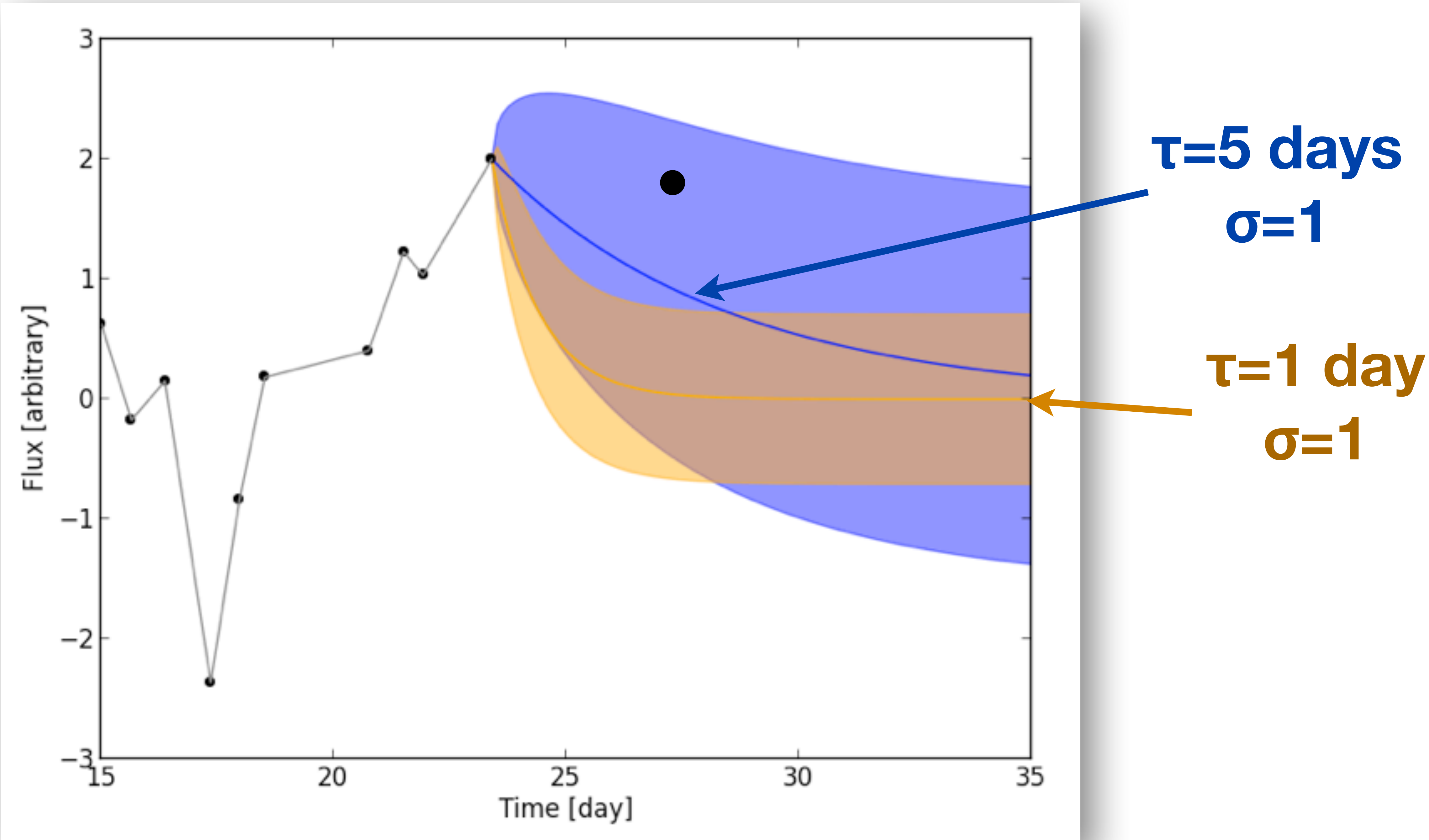



- Solution provides likelihood function, enables maximum-likelihood or Bayesian inference
- Fitting is fast! Only O(n) operations to evaluate likelihood function (e.g., Kelly+2009, Kozlowski +2010) or do interpolation

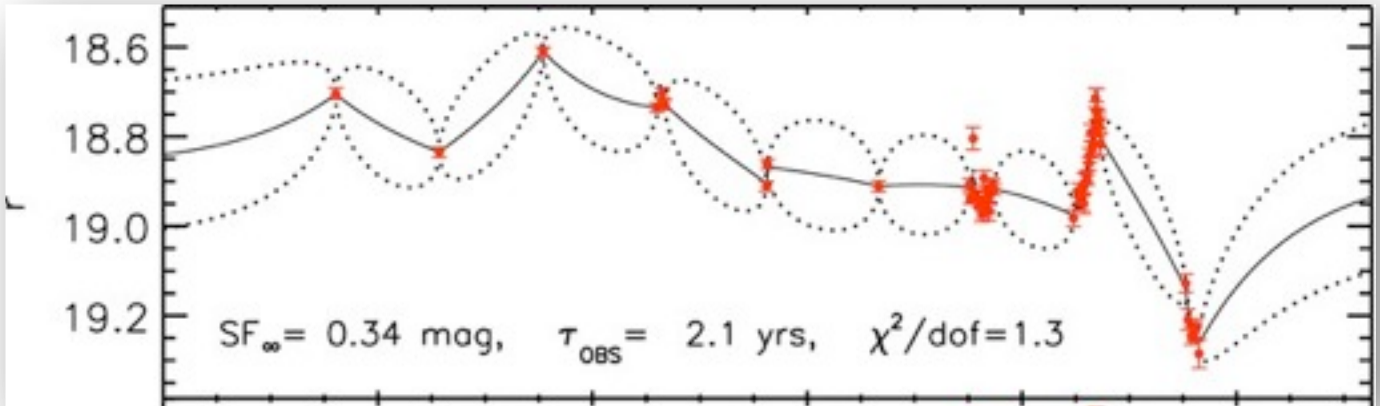
# Fitting the CAR(1) model: Illustration



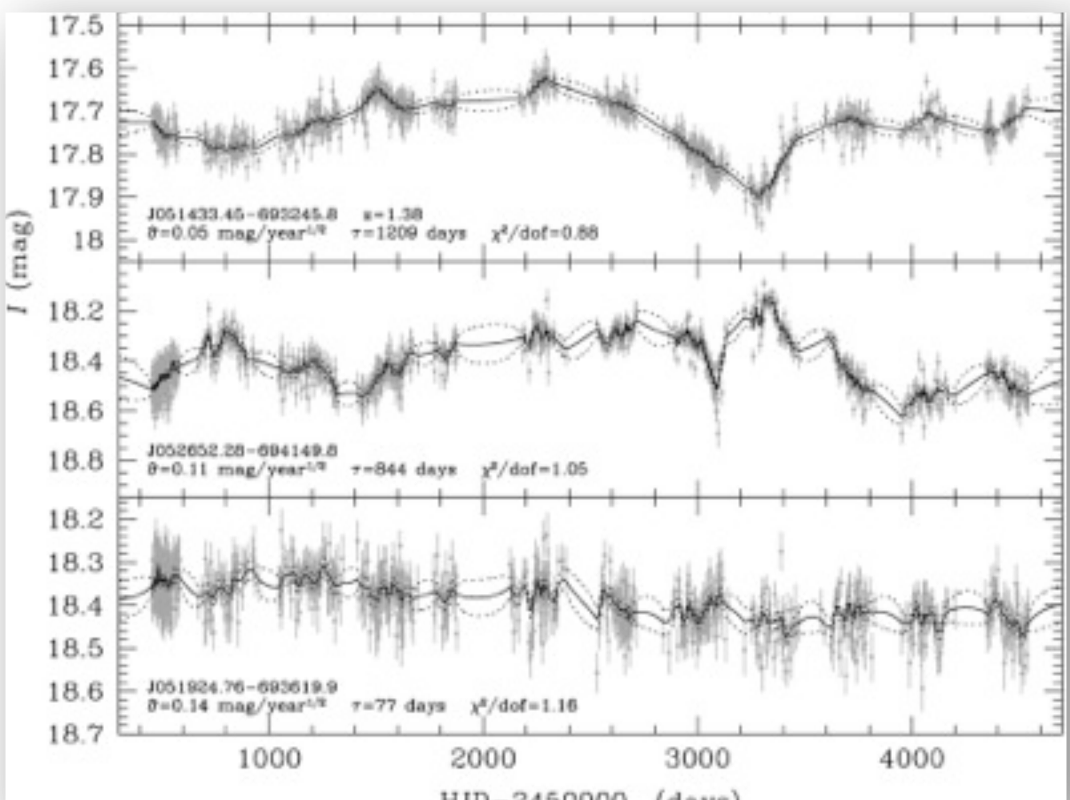
# Fitting the CAR(1) model: Illustration



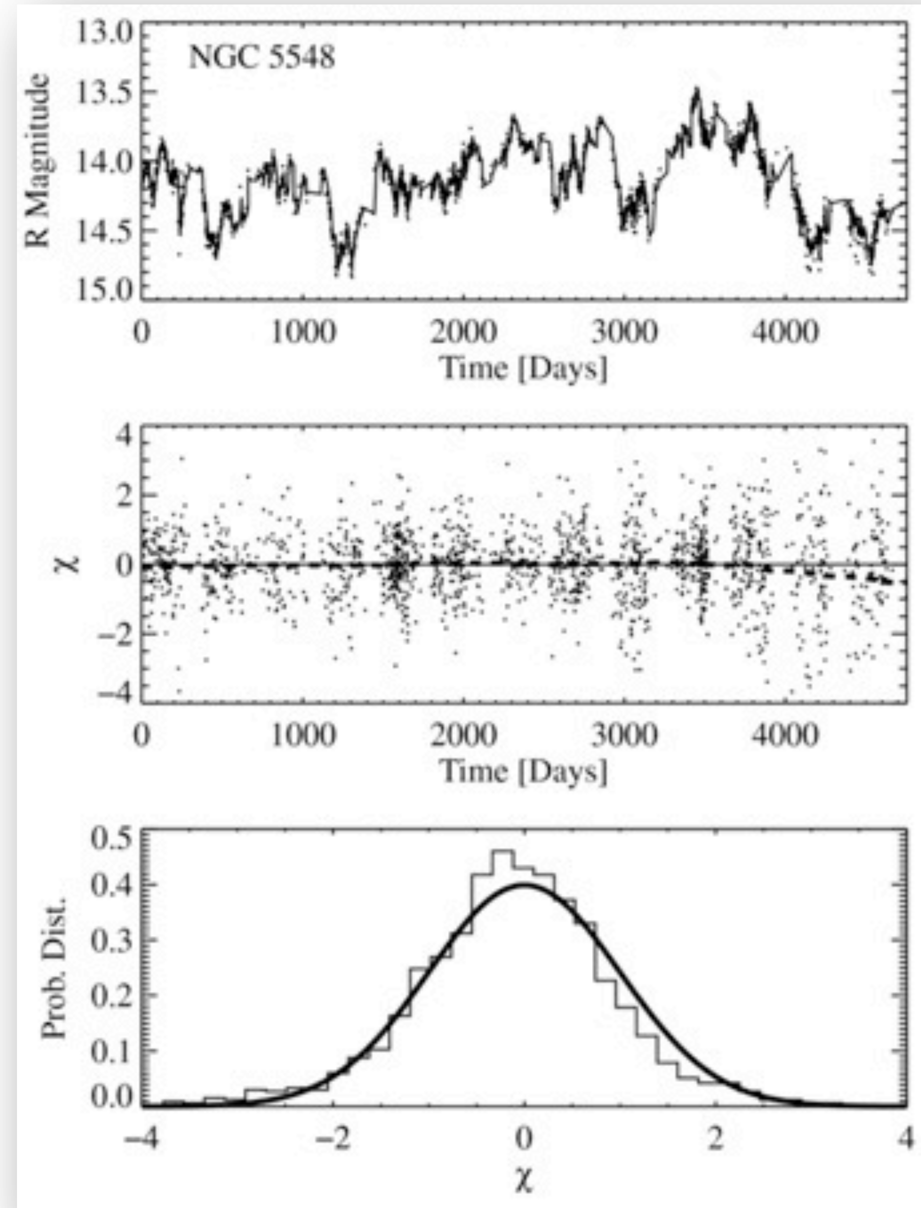
# CAR(1) does well on optical lightcurves with typical sampling of current surveys




MacLeod+(2010), ~10,000 quasars from stripe 82

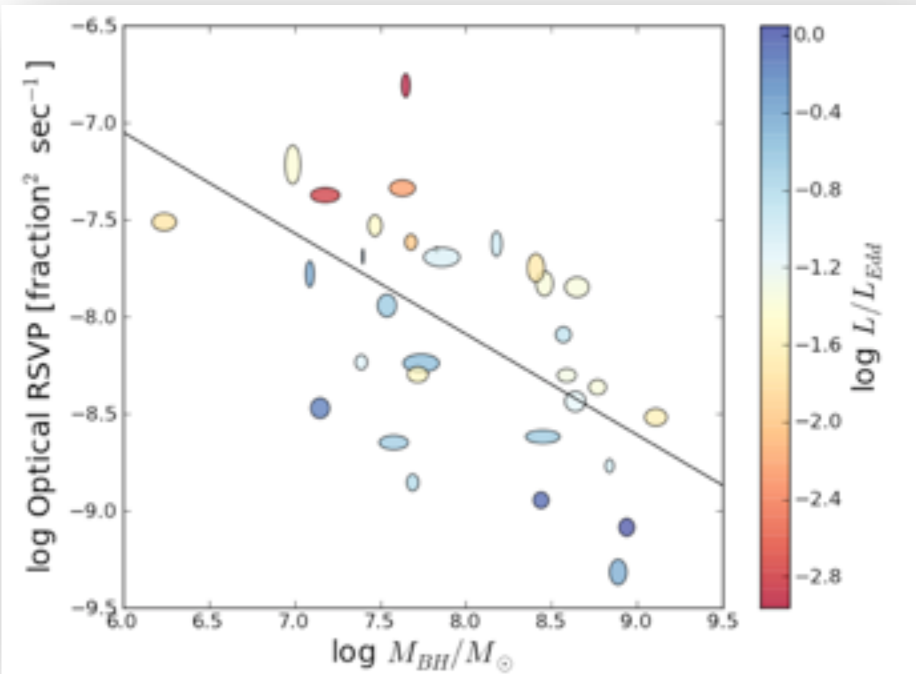


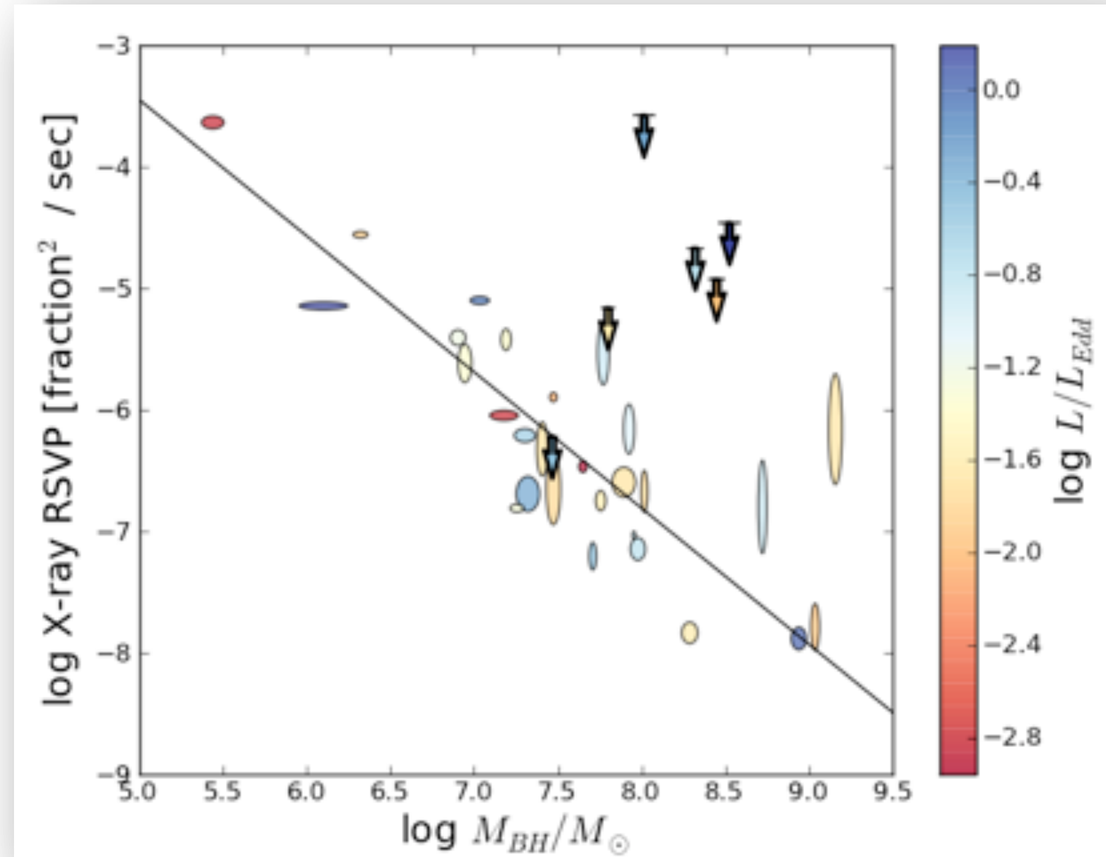

Kozlowski+(2010), Ogle-III



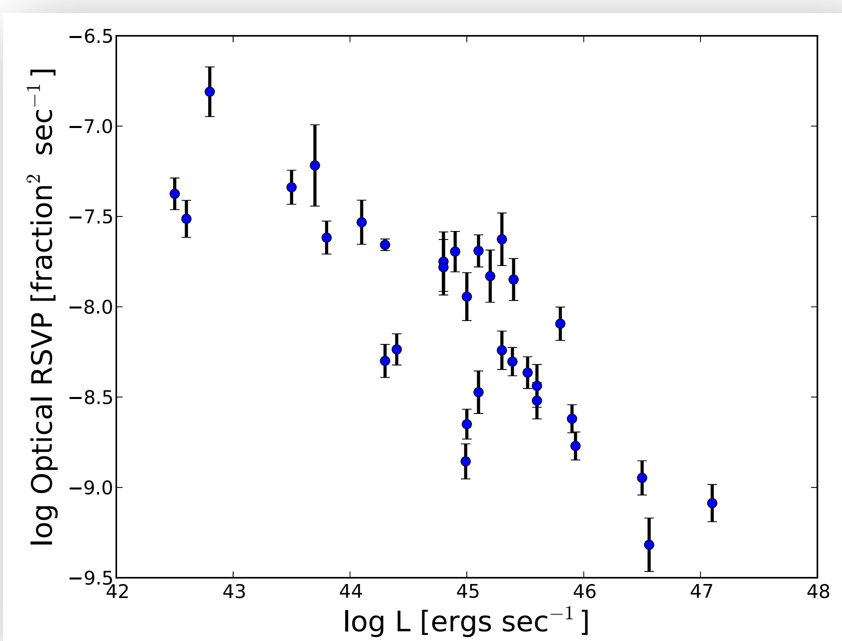

Kelly+(2009), AGN Watch

# Trends involving the CAR(1) process parameters



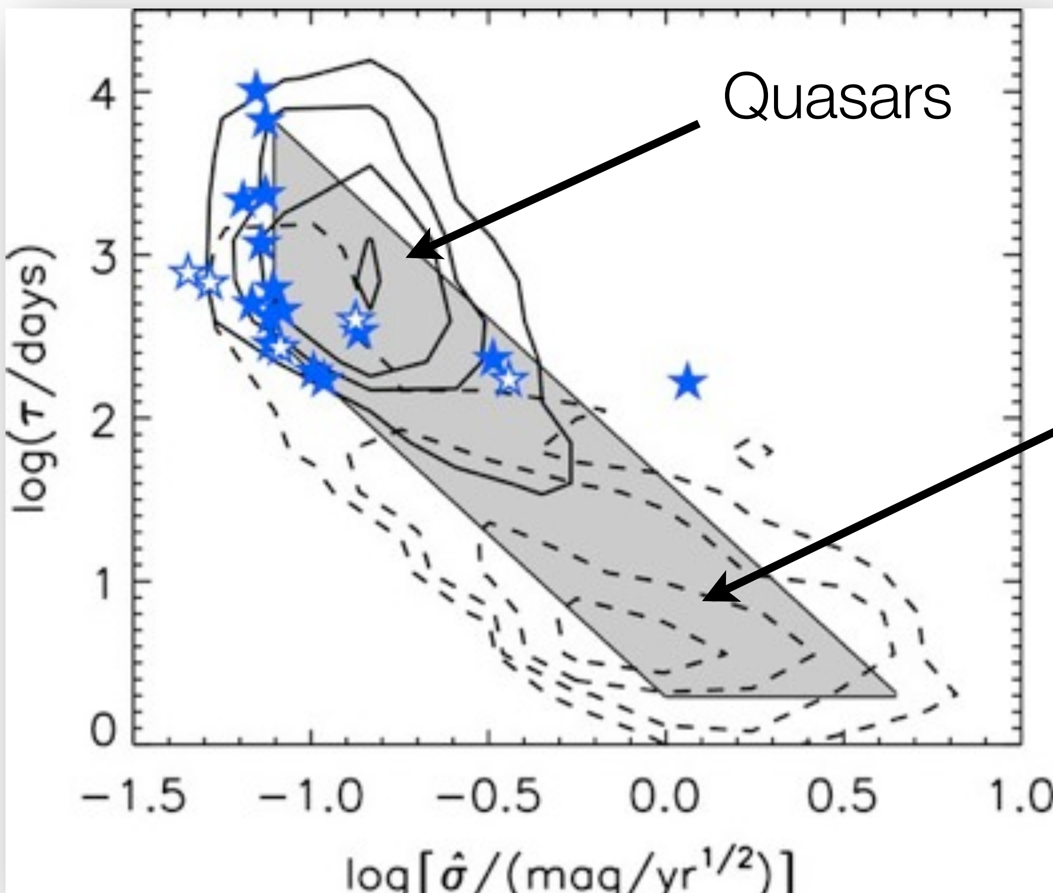




Kelly+(in prep)



# Using the CAR(1) model to find quasars

MacLeod+(2011)

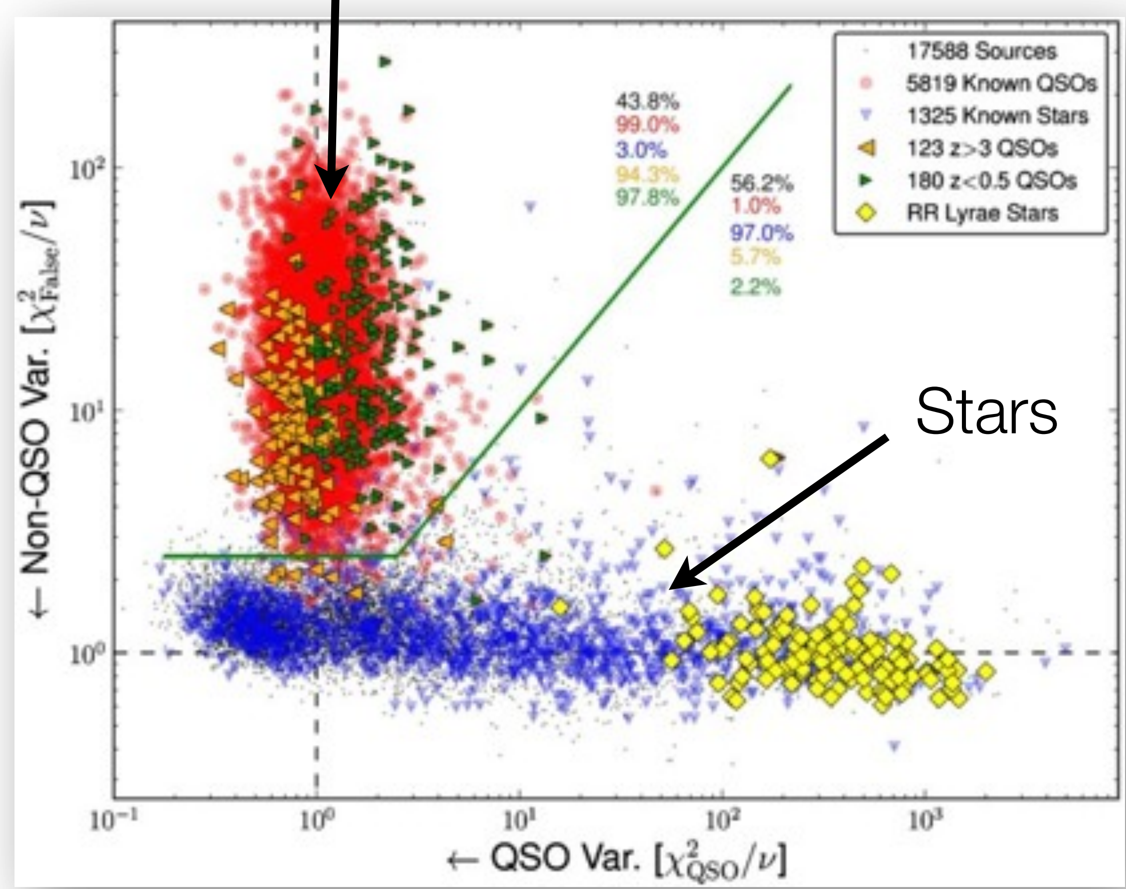



Stars

Based on Stripe82 variable point sources

Quasars

Butler & Bloom (2011)




Works because quasars have more correlated variability on longer time scales compared to stars

# Current Work: More Flexible Stochastic Models

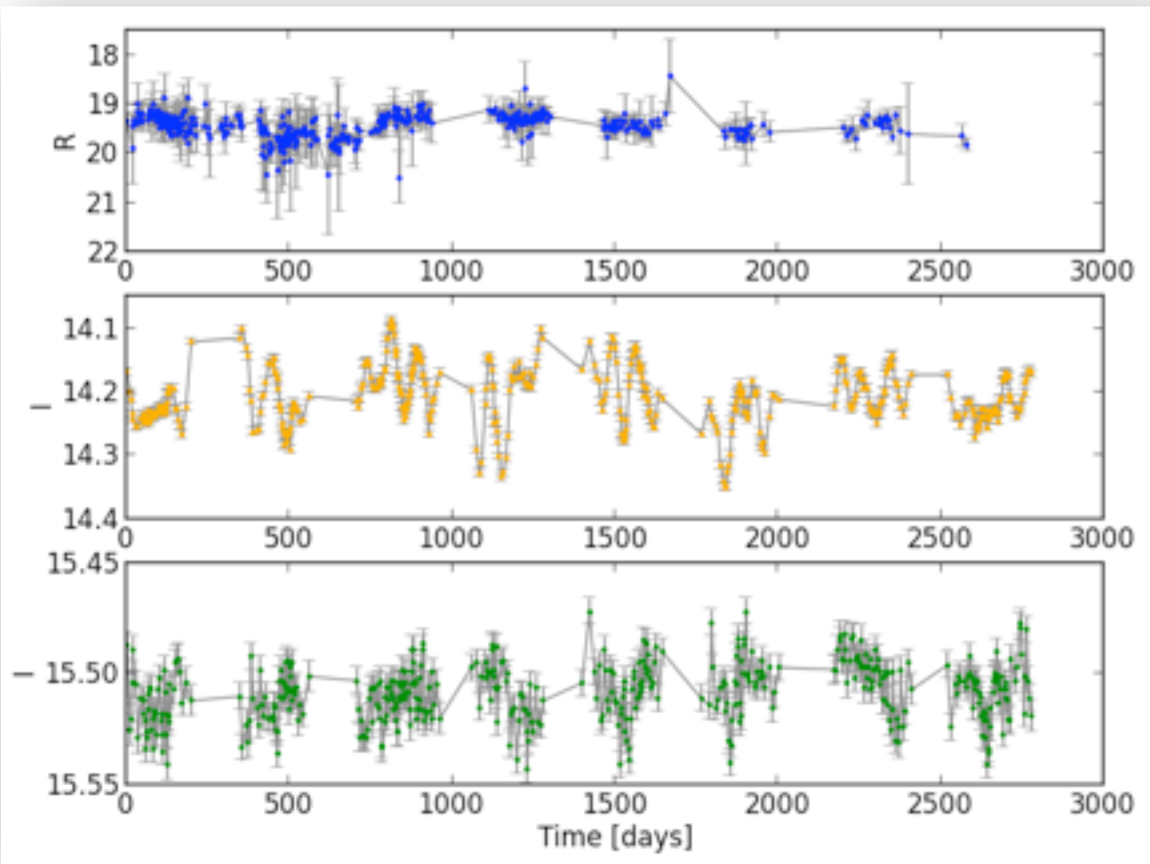
$$\frac{dL^p(t)}{dt} + \alpha_p \frac{dL^{p-1}(t)}{dt^{p-1}} + \ldots + \alpha_1 L(t) = \delta_q \frac{d^q \epsilon(t)}{dt^q} + \delta_{q-1} \frac{d^{q-1}\epsilon(t)}{dt^{q-1}} + \ldots + \epsilon(t)$$

- Continuous-time autoregressive moving average models (CARMA(p,q)) provide flexible modeling of variability

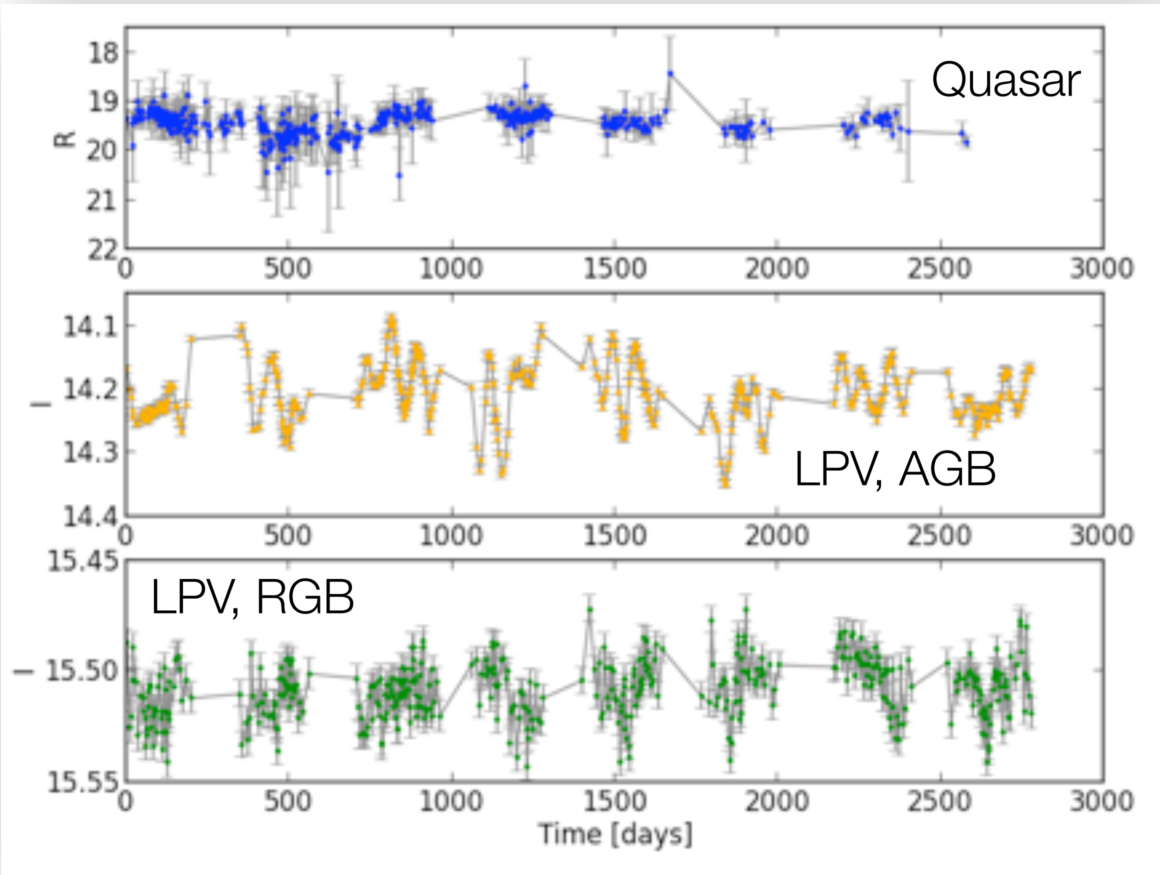
- Power spectrum is a rational function

$$P(\omega) = \sigma^2 \frac{\left|\sum_{k=1}^{q} \delta_k (i\omega)^k\right|^2}{\left|\sum_{j=1}^{p} \alpha_j (i\omega)^j\right|^2}$$

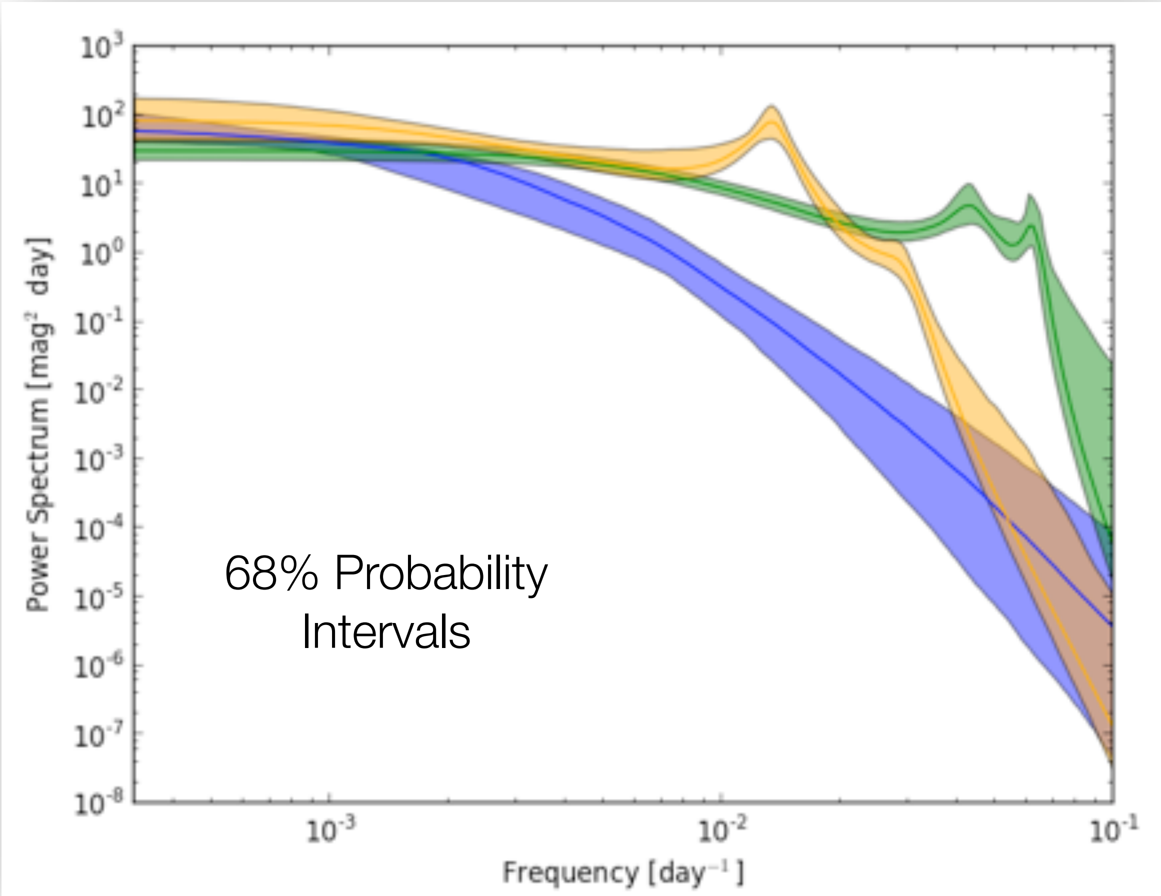
# Example: Quasar vs Variable Stars



# Example: Quasar vs Variable Stars




Kelly+(in prep)



# Calculation of the Likelihood function

- CARMA models have a state space representation:

Measured Lightcurve

Moving Average Coefficients

Measurement Error

$$y_i = b^T X_i + \epsilon_i, \quad \epsilon_i \sim N(0, V_i)$$

$$X_i = A_i X_{i-1} + u_i, \quad u_i \sim N(0, \Sigma_i)$$

Vector of lightcurve derivatives

Depends autoregressive coefficients

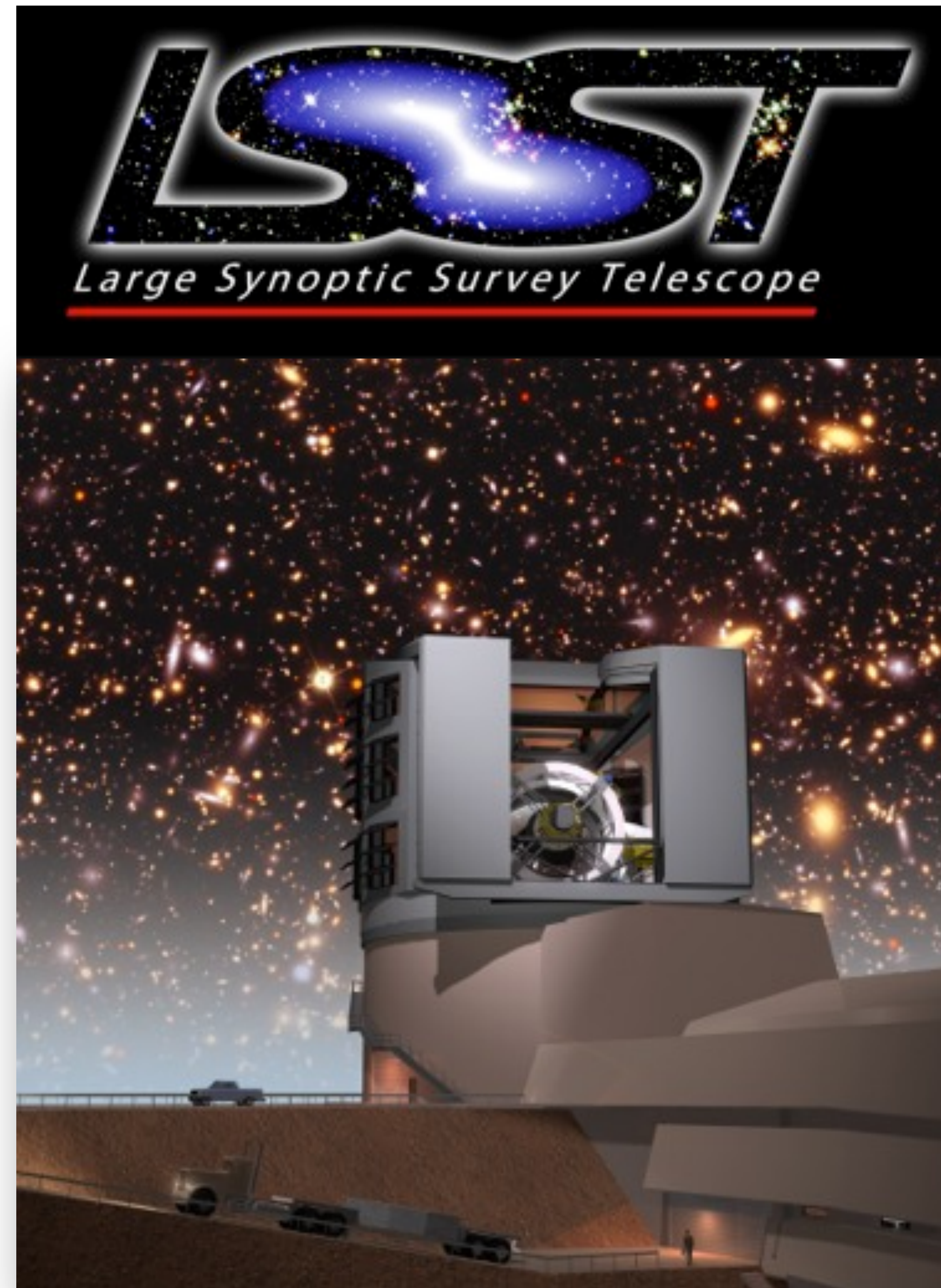
'Innovation' (source of stochasticity)

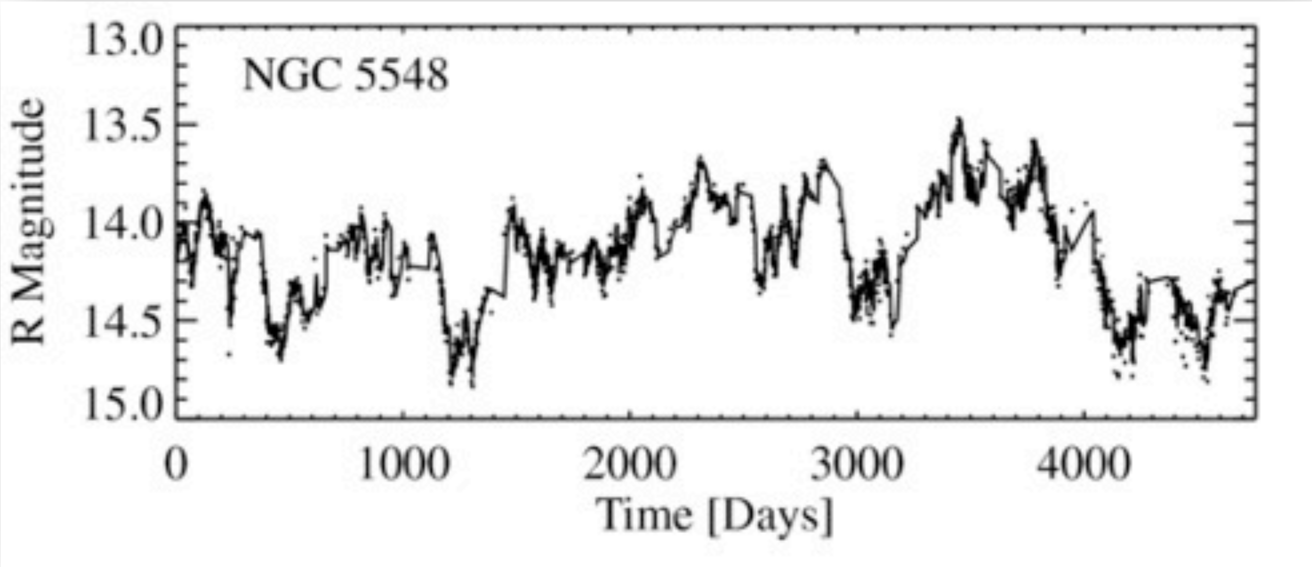
- Likelihood calculated from Kalman Recursions in O(n) operations:

$$p(y_1, \ldots, y_n | \delta, \alpha, \sigma^2) = p(y_1 | \delta, \alpha, \sigma^2) \prod_{i=2}^{n} p(y_i | y_{i-1}, \ldots, y_1, \delta, \alpha, \sigma^2)$$

# Computational Techniques

- Use Robust Adaptive Metropolis Algorithm (Vihola 2012)
- Likelihood space often multimodal, so also do parallel tempering
- Can be slow (~minutes for ~ 100,000 iterations for ~ 100 epochs) due to complicated posterior space
- Exploring alternative parameterizations for improving efficiency
- Sampling methods/global optimization algorithms can be efficiently parallelized, exploit high-performance computing, GPUs?






Credit: ESO/Kornmesser

# Main Takeaway Point:

Stochastic modeling provides a useful and powerful framework to quantify quasar variability that can be applied to lightcurves of arbitrary sampling and with measurement error.

# Time Domain Stochastic Modeling: Outstanding issues and directions for future work

# Time Domain Stochastic Modeling: Outstanding issues and directions for future work

- Multivariate lightcurves:
    - Explicit modeling of time lags and correlation structure between different wavelengths
    - Vector-valued CARMA(p,q) processes may provide general framework

# Time Domain Stochastic Modeling: Outstanding issues and directions for future work

- Multivariate lightcurves:
    - Explicit modeling of time lags and correlation structure between different wavelengths
    - Vector-valued CARMA(p,q) processes may provide general framework
- Moving beyond a single stationary Gaussian process:
    - Direct modeling of stochastic process + flares, other ‘state’ changes (Sobolewska+, in prep)
    - Using alternatives to Gaussian noise (Emmanaloupolous+2013)
    - Non-stationary and non-linear models

# Tir deling: Outstanding iss ıre work

• Mu

• ıre between different wavelengths

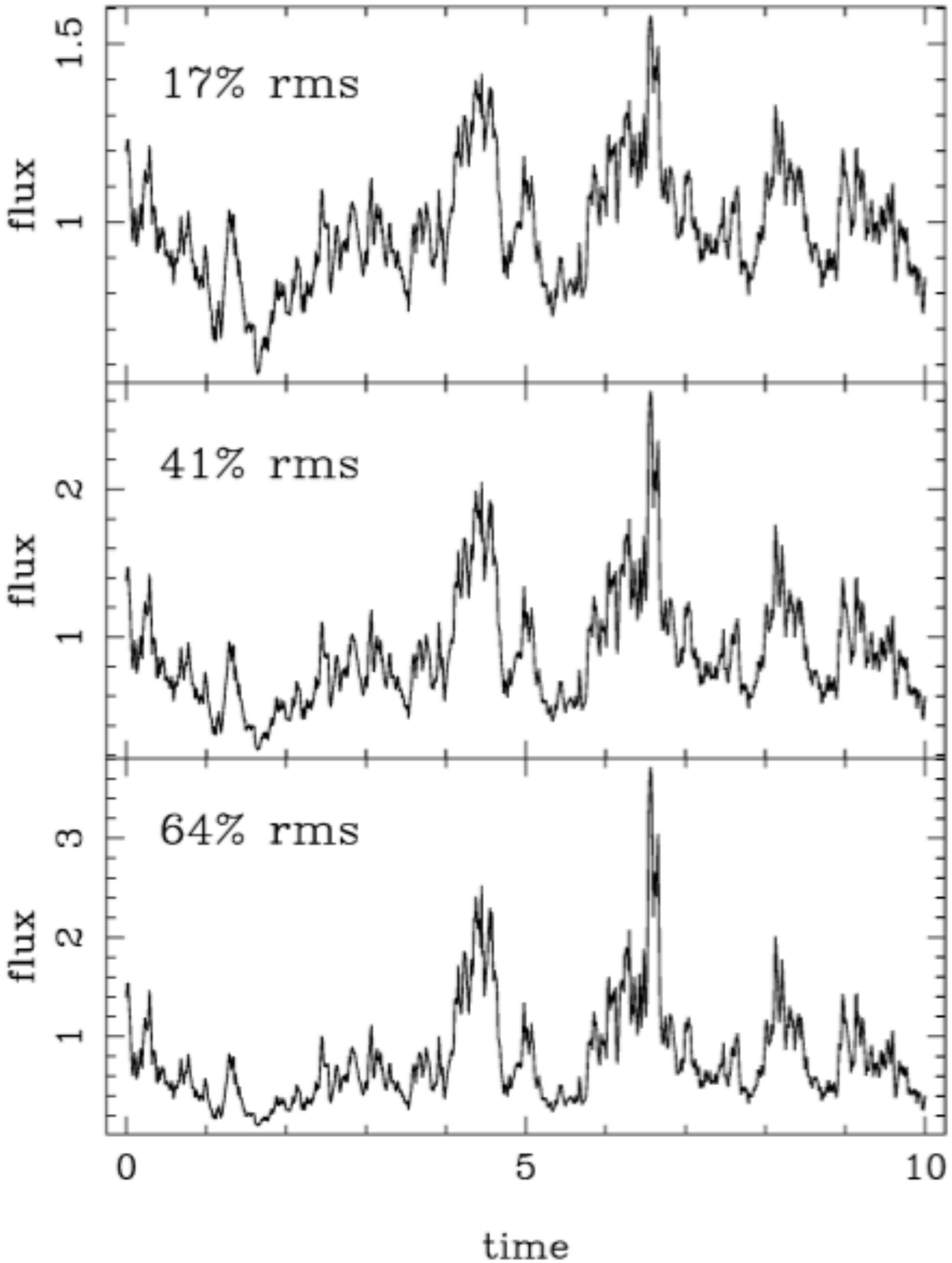
• general framework

• Mo

• r 'state' changes (Sobolewska+, in prep)

• ılous+2013)

•




Uttley+(2005)

# Time Domain Stochastic Modeling: Outstanding issues and directions for future work

- Multivariate lightcurves:
  - Explicit modeling of time lags and correlation structure between different wavelengths
  - Vector-valued CARMA(p,q) processes may provide general framework
- Moving beyond a single stationary Gaussian process:
  - Direct modeling of stochastic process + flares, other ‘state’ changes (Sobolewska+, in prep)
  - Using alternatives to Gaussian noise (Emmanaloupolous+2013)
  - Non-stationary and non-linear models

# Time Domain Stochastic Modeling: Outstanding issues and directions for future work

- Multivariate lightcurves:
  - Explicit modeling of time lags and correlation structure between different wavelengths
  - Vector-valued CARMA(p,q) processes may provide general framework
- Moving beyond a single stationary Gaussian process:
  - Direct modeling of stochastic process + flares, other ‘state’ changes (Sobolewska+, in prep)
  - Using alternatives to Gaussian noise (Emmanaloupolous+2013)
  - Non-stationary and non-linear models
- Building astrophysically-motivated stochastic models
  - Stochastic partial differential calculations + accretion flow models?