Data science for (physical) scientists – PHYS 467/667

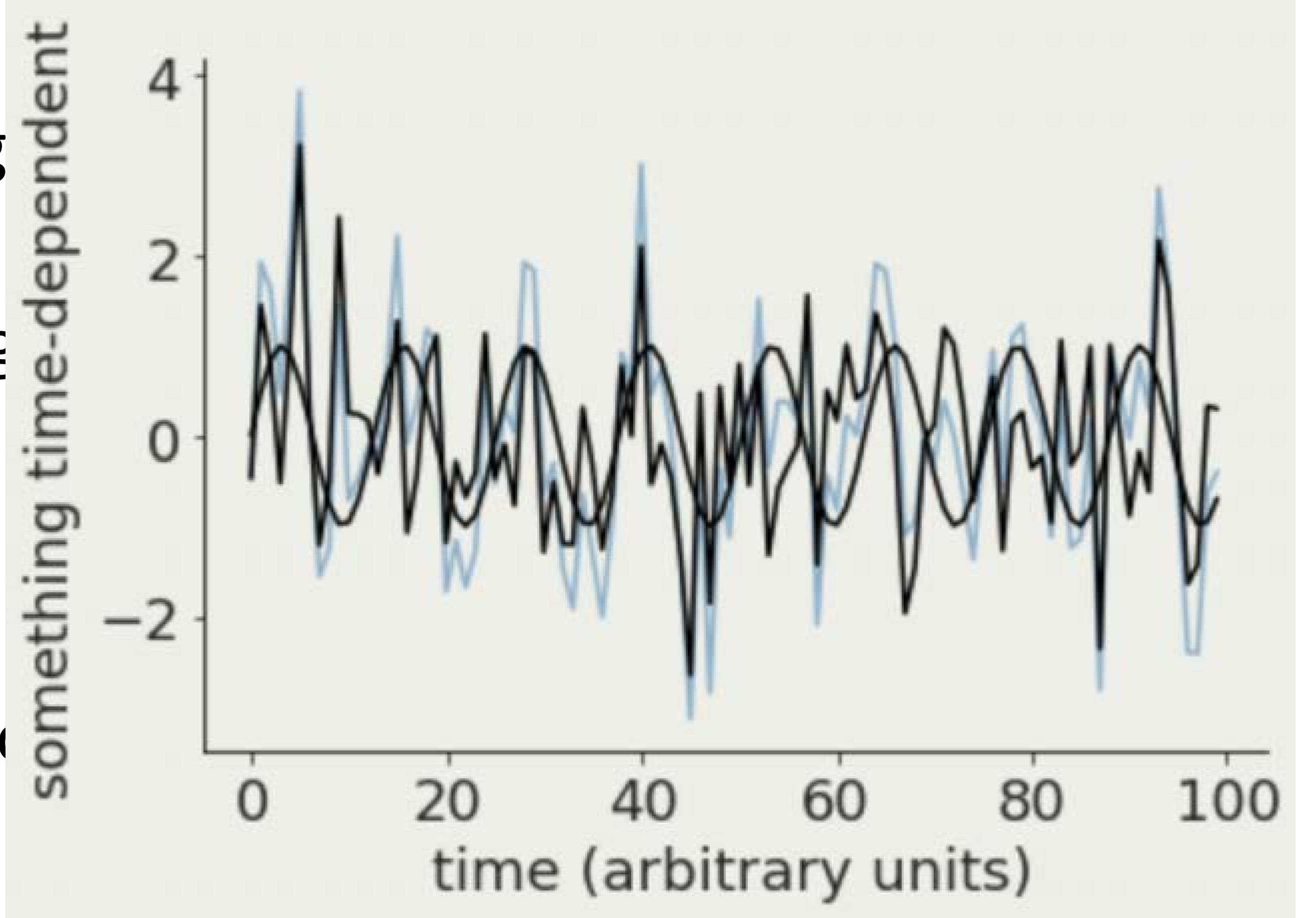
Lecture 10, part 2 – time series analysis

Quick recap

- Stationarity (mean and variance constant)
- Smoothing (rolling average)
- Detrending (by difference, or by fitting the trend)
- Event detection (thresholding)
- Classification (clustering)

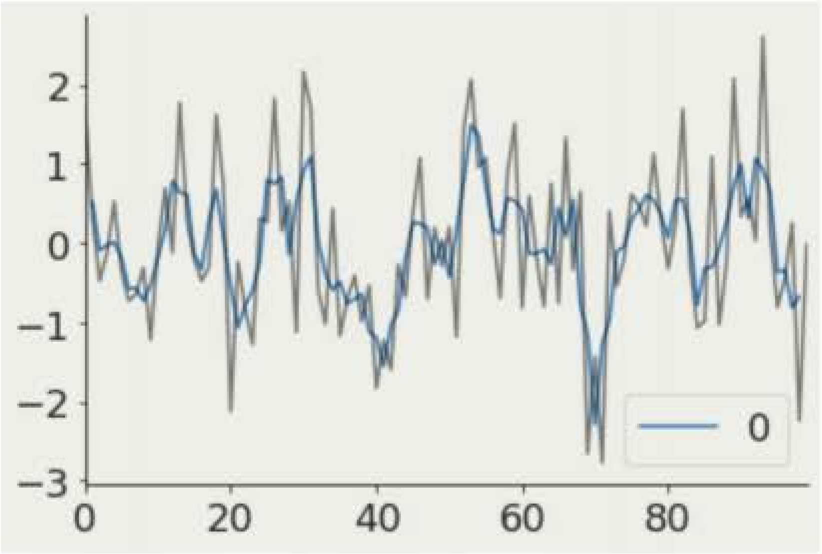
Quick recap

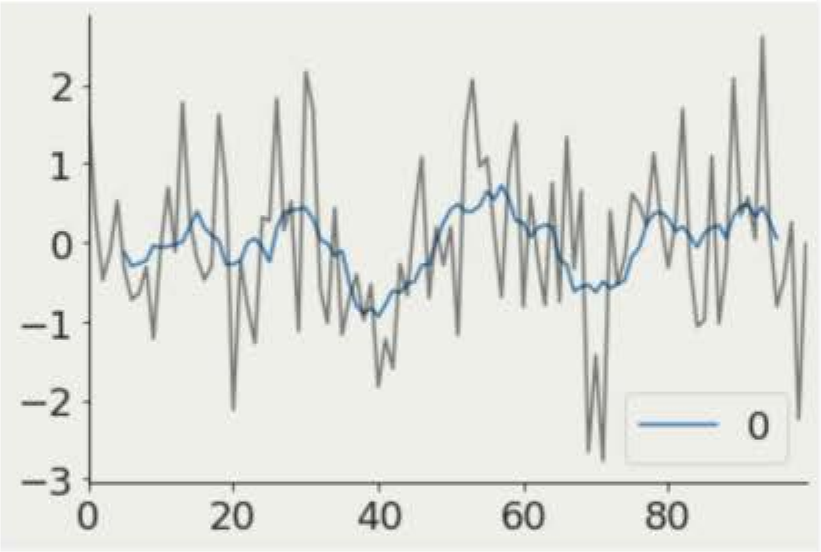
- **Stationarity (mean and variance constant)**
- Smoothing
- Detrending
- Event dete
- Classificati

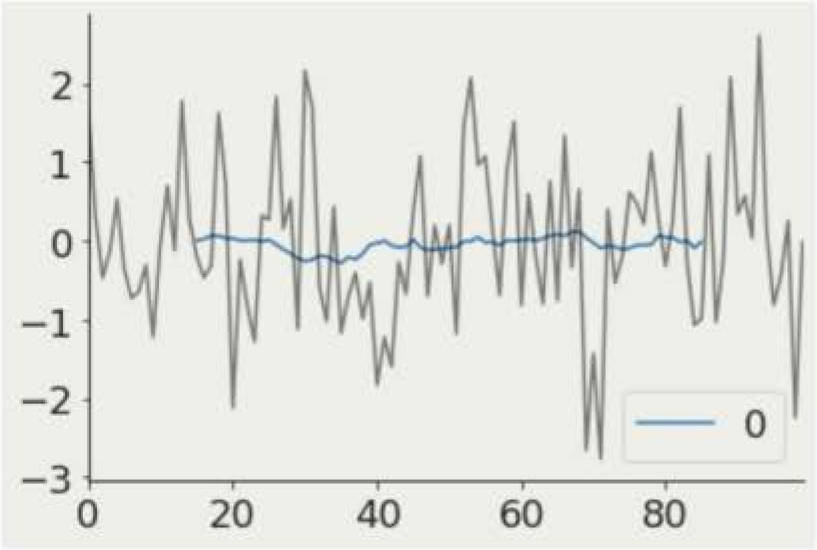


Quick recap

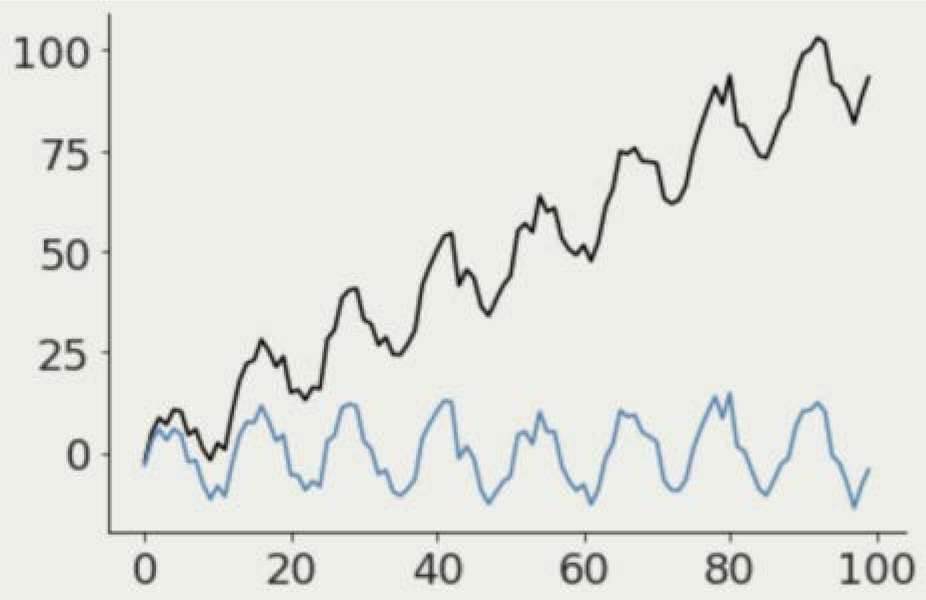
- Stationarity (mean and variance constant)
- **Smoothing (rolling average)**







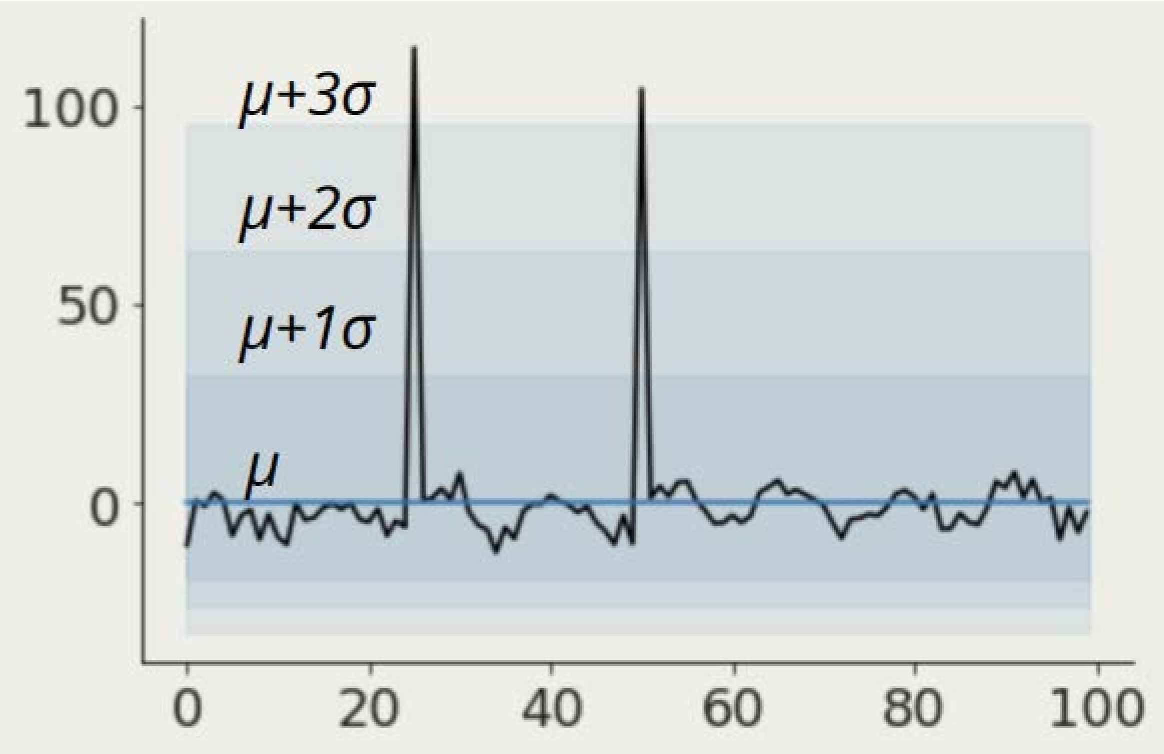
Quick recap



- Stationarity (mean and variance co
- Smoothing (rolling average)
- **Detrending (by difference, or by fitting the trend)**
- Event detection (thresholding)
- Classification (clustering)

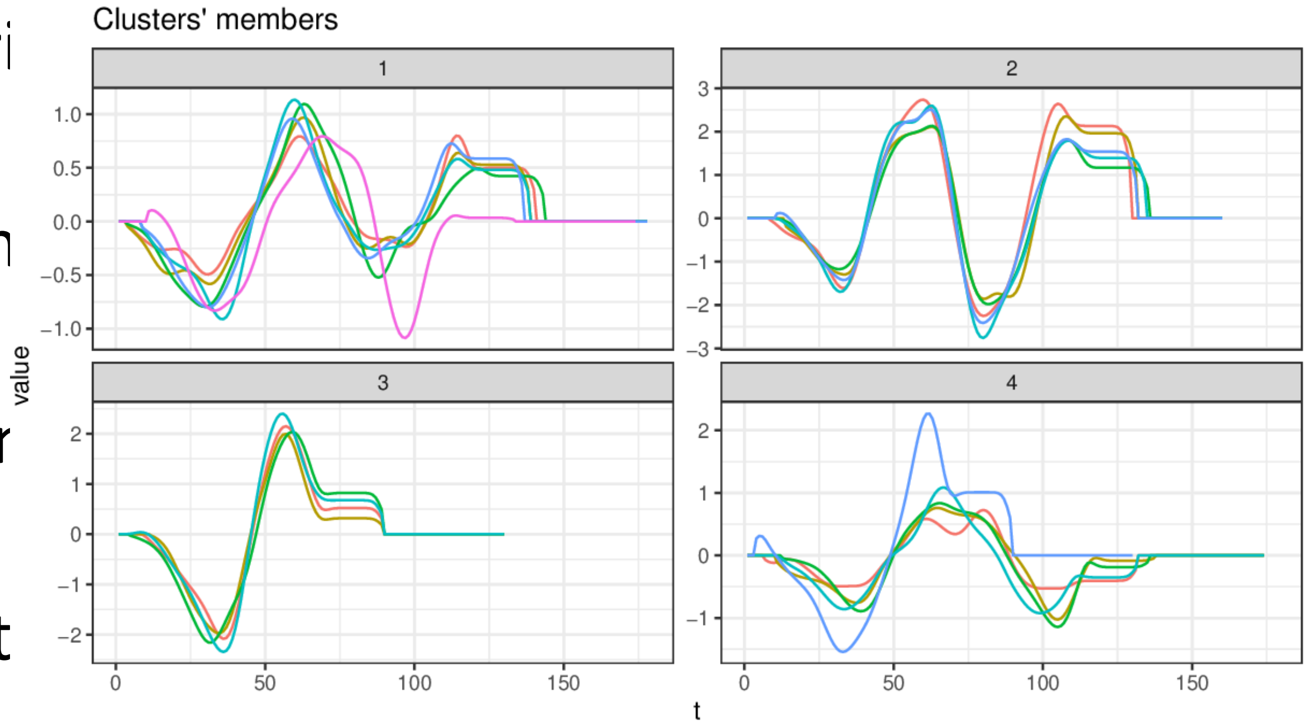
Quic

- Static
- Smoo
- Detre   trend)
- **Event detection (thresholding)**
- Classification (clustering)



Quick recap

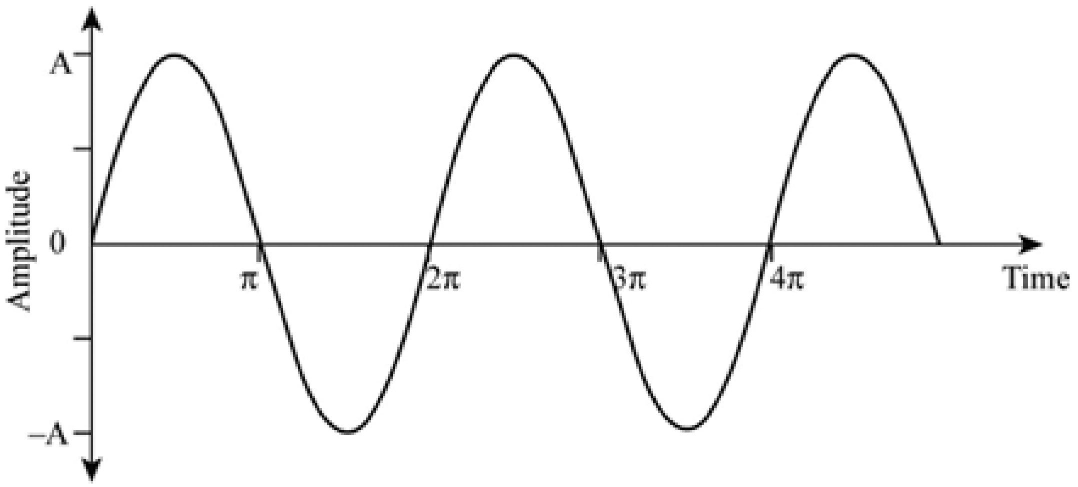
- Stationari
- Smoothin
- Detrendir
- Event det



- **Classification (clustering)**

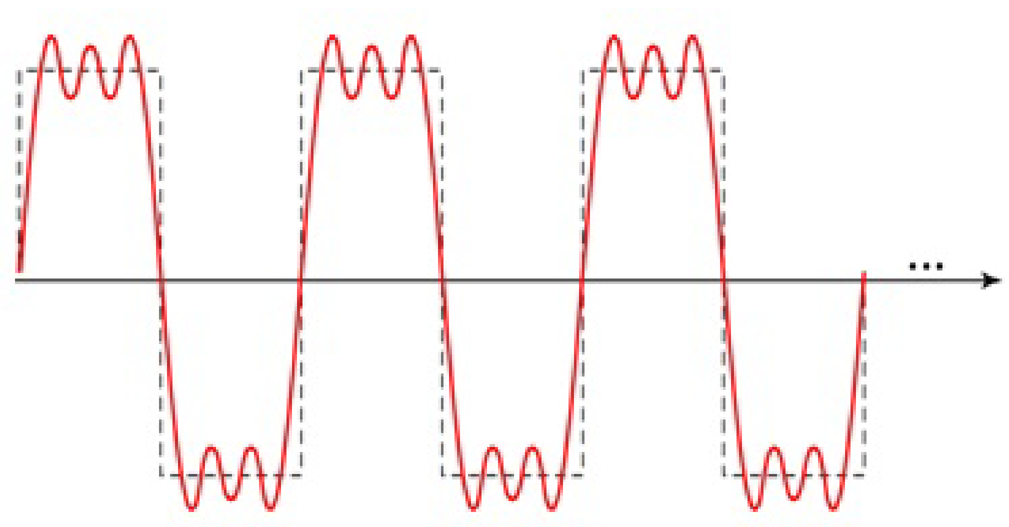
Non-stationary time series - Types of variability

- Periodic variability: repeatable pattern with a well defined *period*



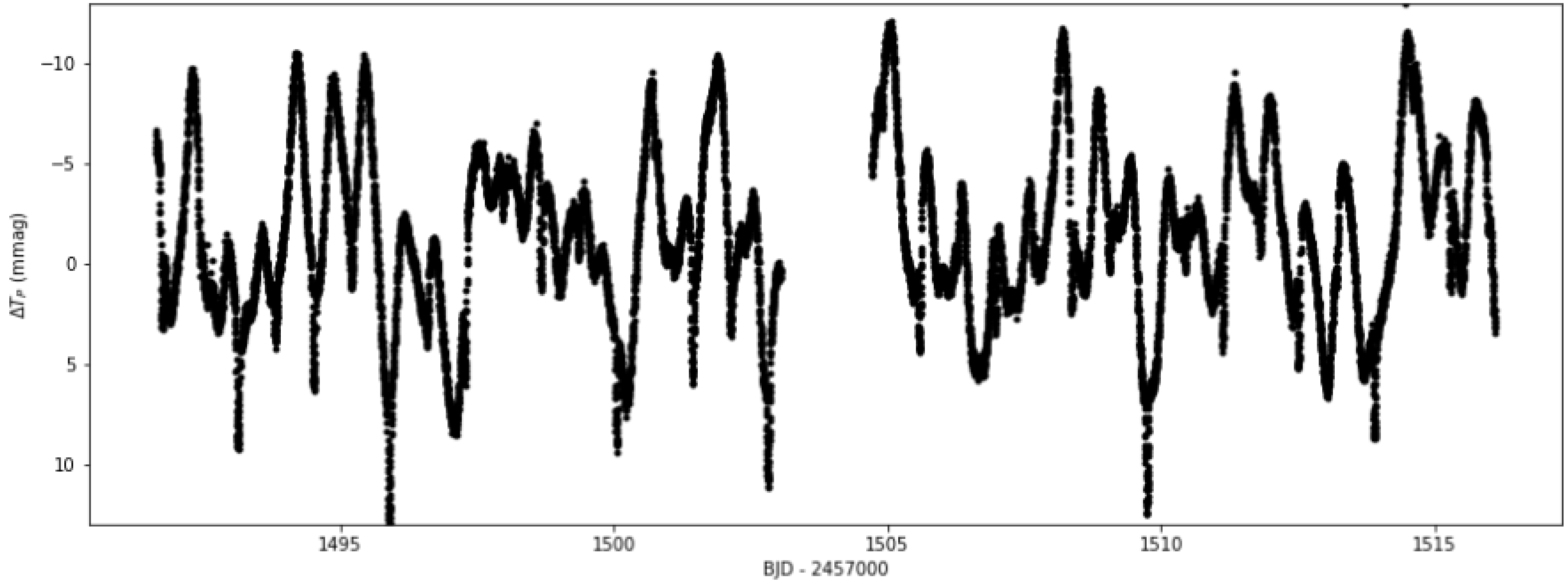
Non-stationary time series - Types of variability

- Periodic variability: repeatable pattern with a well defined *period*



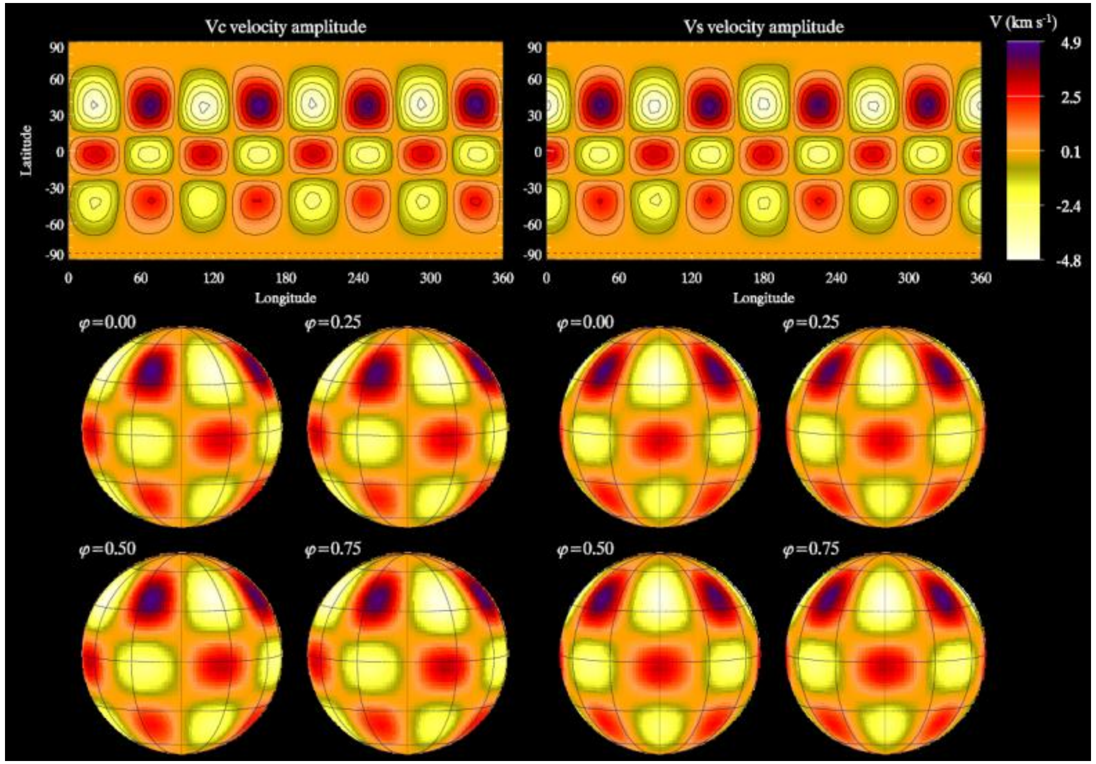
Non-stationary time series - Types of variability

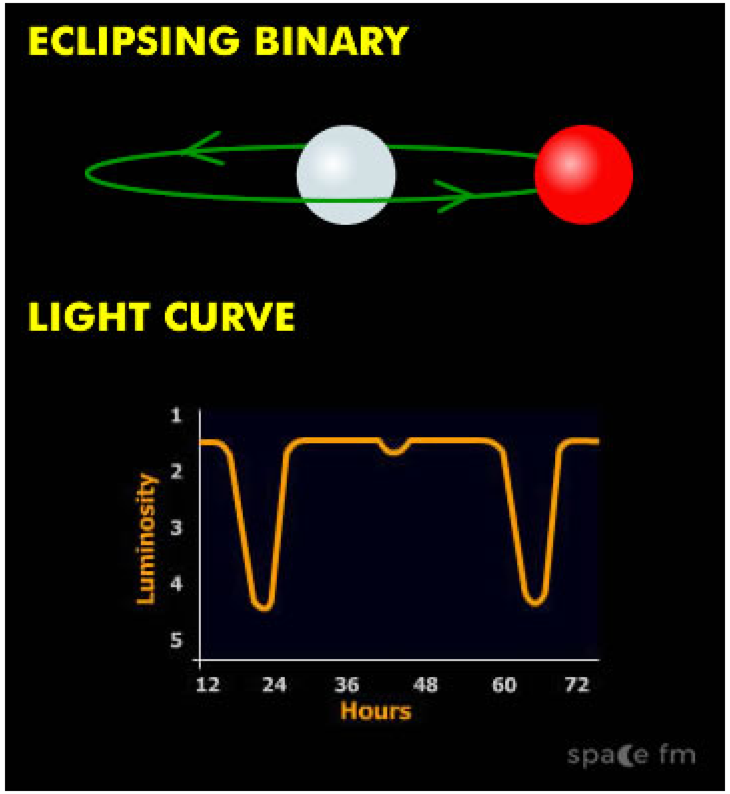
- Time series can be multiperiodic, i.e. defined by multiple periods (we will come back to this example later…)



Non-stationary time series - Types of variability

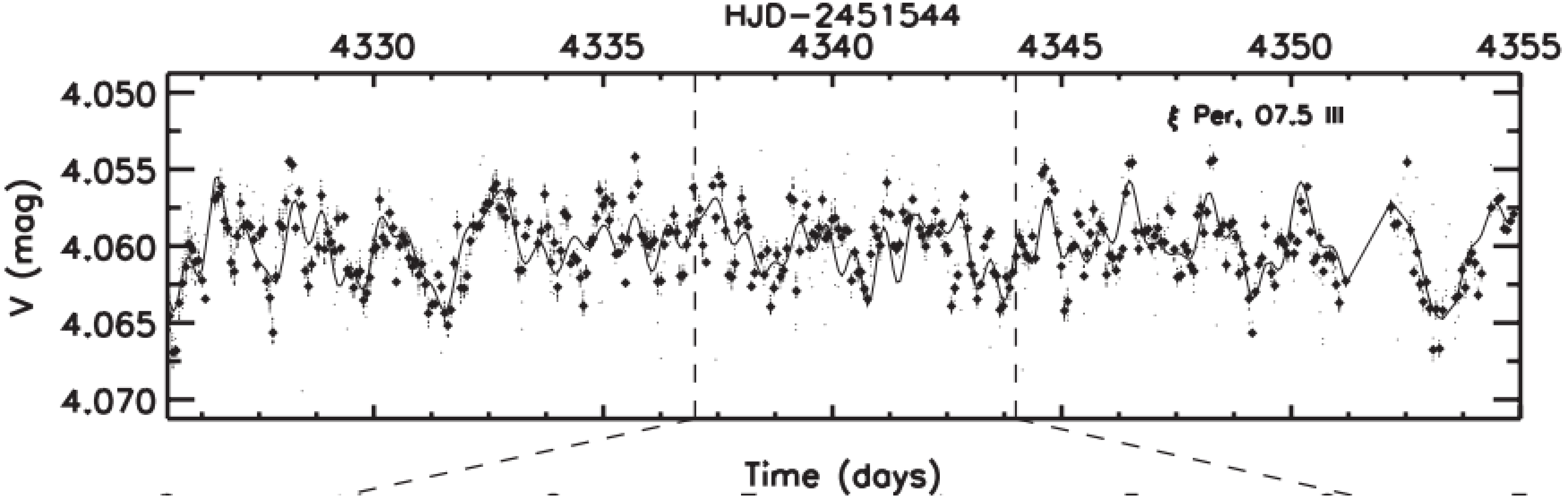
- Astrophysical example of a periodic signal: pulsations, eclipses/transits





Non-stationary time series – Types of variability

- Cyclical variability: well-defined *timescale*, but cannot be well *phased* (similar to multi-periodic, but we'll see later how they can be distinguished)



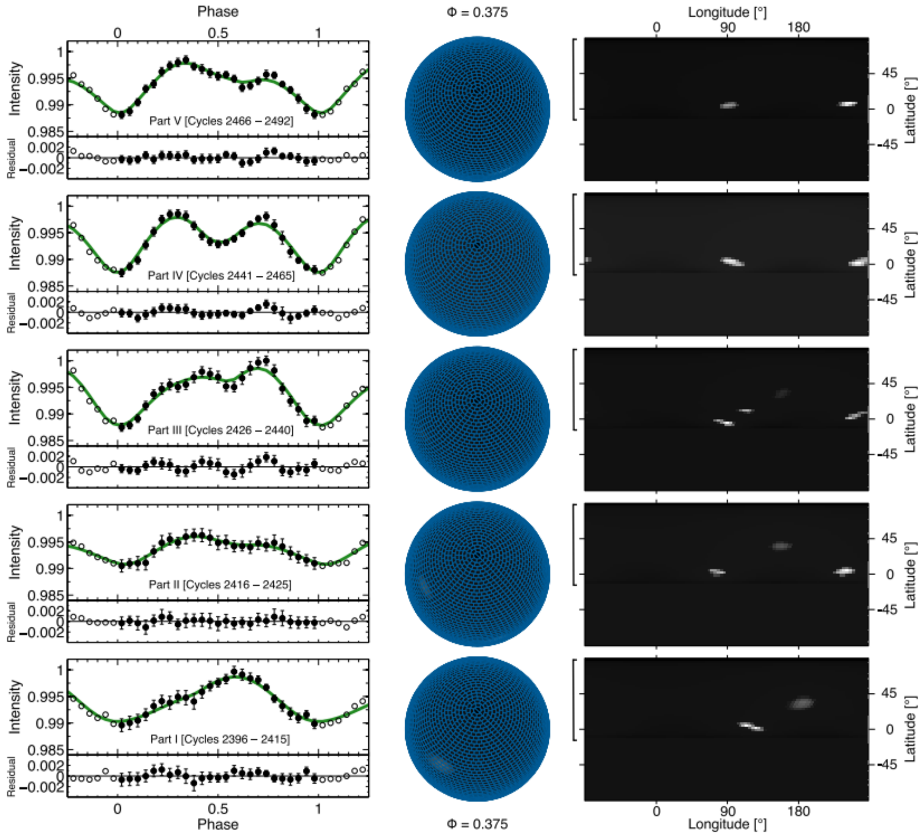
Parenthesis: phase-folding

- This is accomplished by taking the fractional part of a rescaled exogenous variable, e.g.:

$$\varphi = \mathrm{frac}\left(\frac{t - t_0}{P}\right)$$

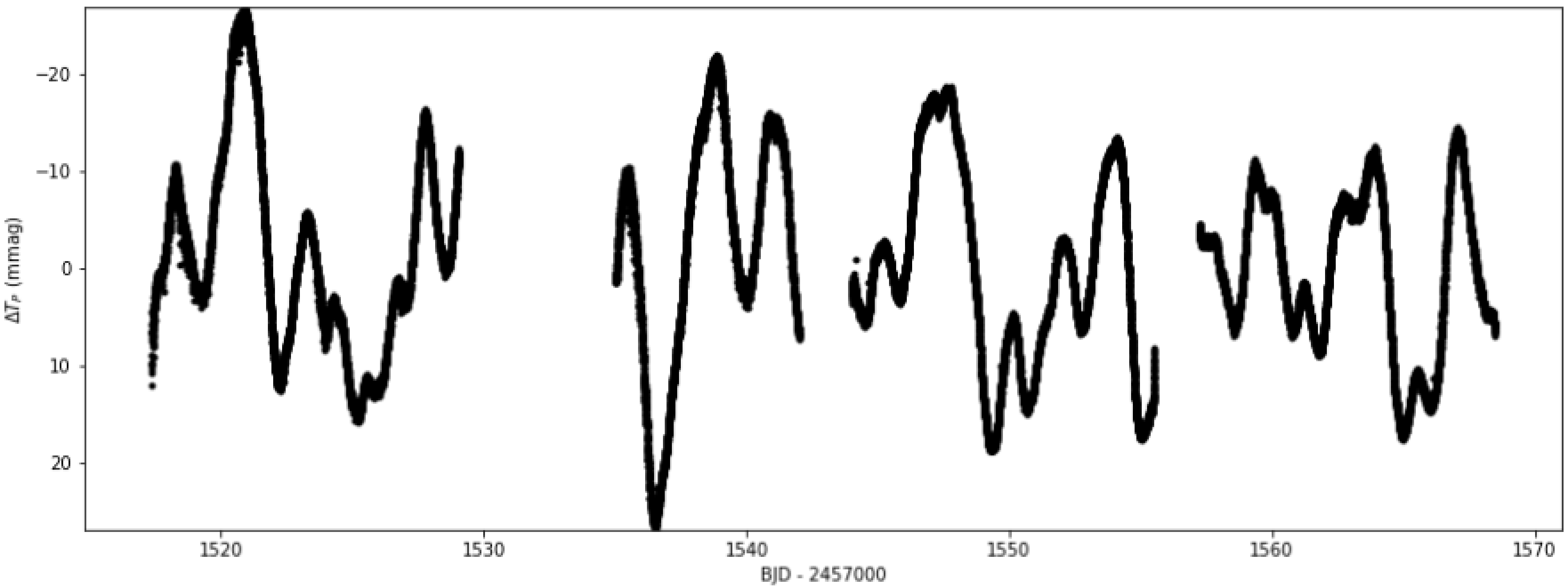
Non-stationary time series – Types of variability

- Astrophysical example of a cyclical signal: spots



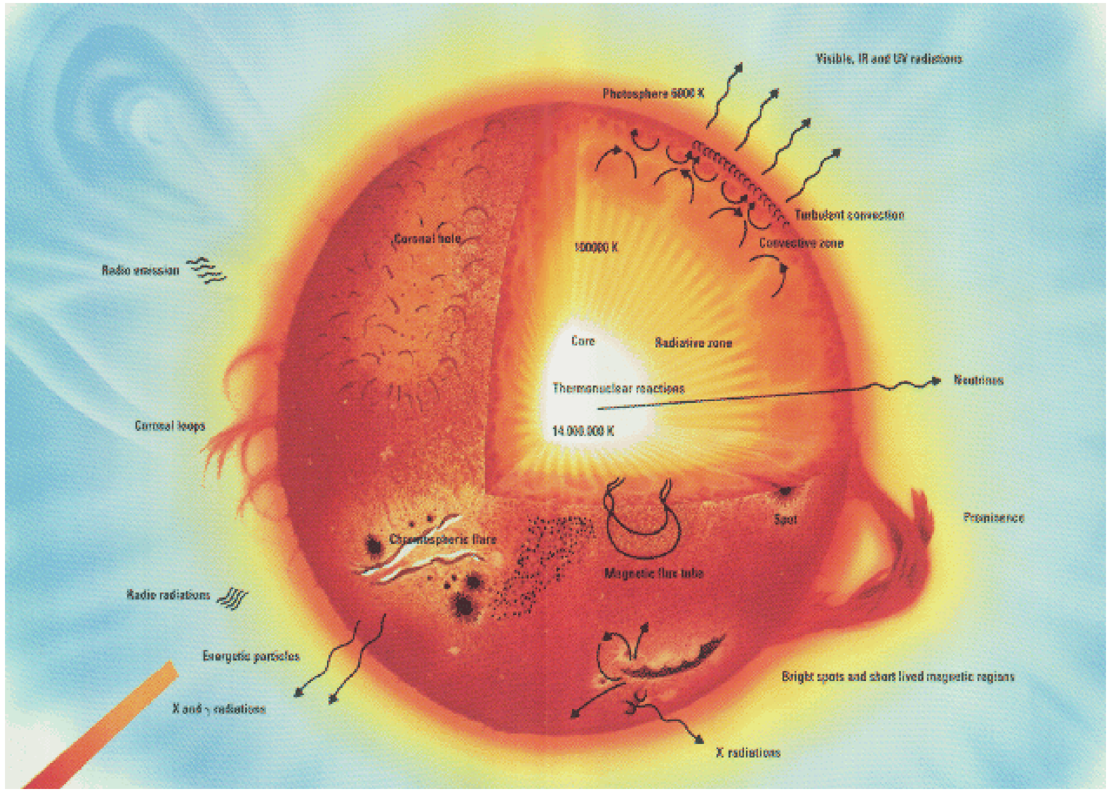
Non-stationary time series – Types of variability

- Stochastic variability: this type of variability is more or less random and cannot be associated with a well-defined period



Non-stationary time series – Types of variability

- Astrophysical example of a periodic signal: convection/granulation



How to identify variability type?

- Sometimes, we can get more information in frequency space

- To access this information, we perform a *Fourier transform*

Fourier transform

$$g(f) = \int_{-\infty}^{\infty} f(t) e^{-i2\pi f t} dt$$

Fourier transform

$$g(f) = \int_{-\infty}^{\infty} f(t)\boxed{e^{-i2\pi ft}}dt$$

Fourier transform

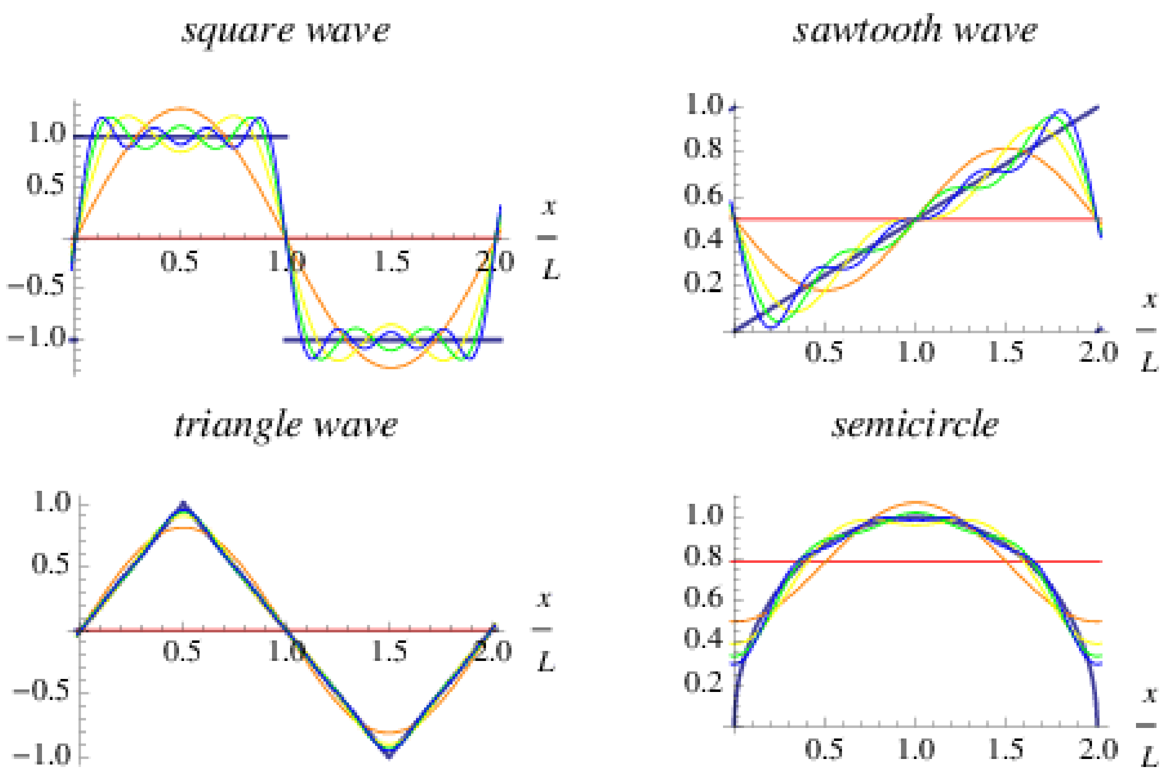
Time series

Sinusoidal functions

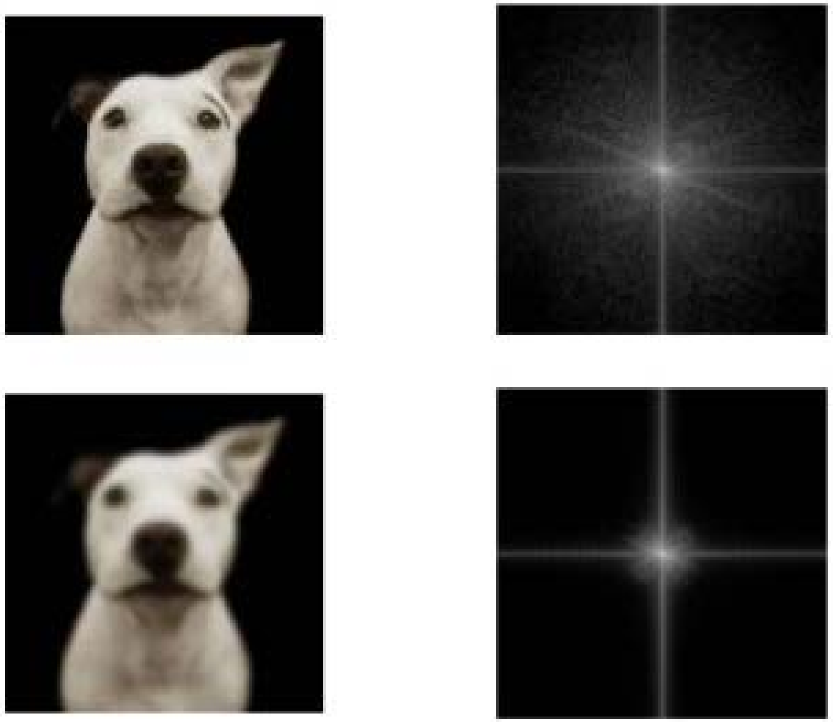
Fourier transform

- For a basic but quite intuitive tutorial: http://www.thefouriertransform.com/

Fourier transform



Fourier transform – image analysis applications



Original image

Gaussian blur with sigma value of 3

Fourier transform – additional considerations

- Such as presented above, the Fourier transform applies to continuous functions
- Time series data are discrete, therefore a slightly different technique is applied – the *Discrete Fourier Transform*
- Some additional considerations must be taken into account when dealing with unevenly spaced data, both in executing the DFT and in interpreting the results (e.g. *aliasing*)
- Improvements can be achieved with, e.g., least-square fitting techniques such as the *Lomb-Scargle* method

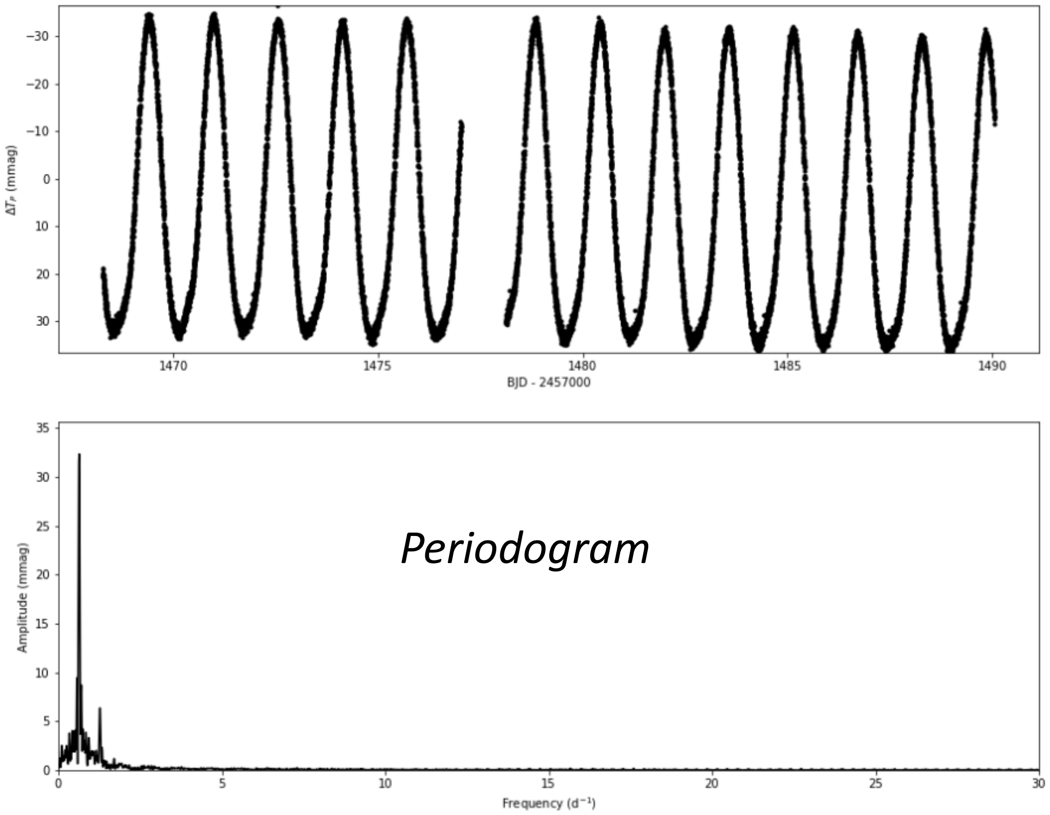
Fourier transform – additional considerations

- Such as presented above, the Fourier tran … ntinuous functions
- Time series data are discrete, … chnique is applied – the *Discrete F* …
- Some additional c … nto account when dealing with … executing the DFT and in interpre …
- Improv … with, e.g., least-square fitting techniq … *mb-Scargle* method

Thankfully, we will not worry about the mathematical details that each of these issues involves, since there exist advanced Python packages that do the work for us!

Fourier transform – Types of variability

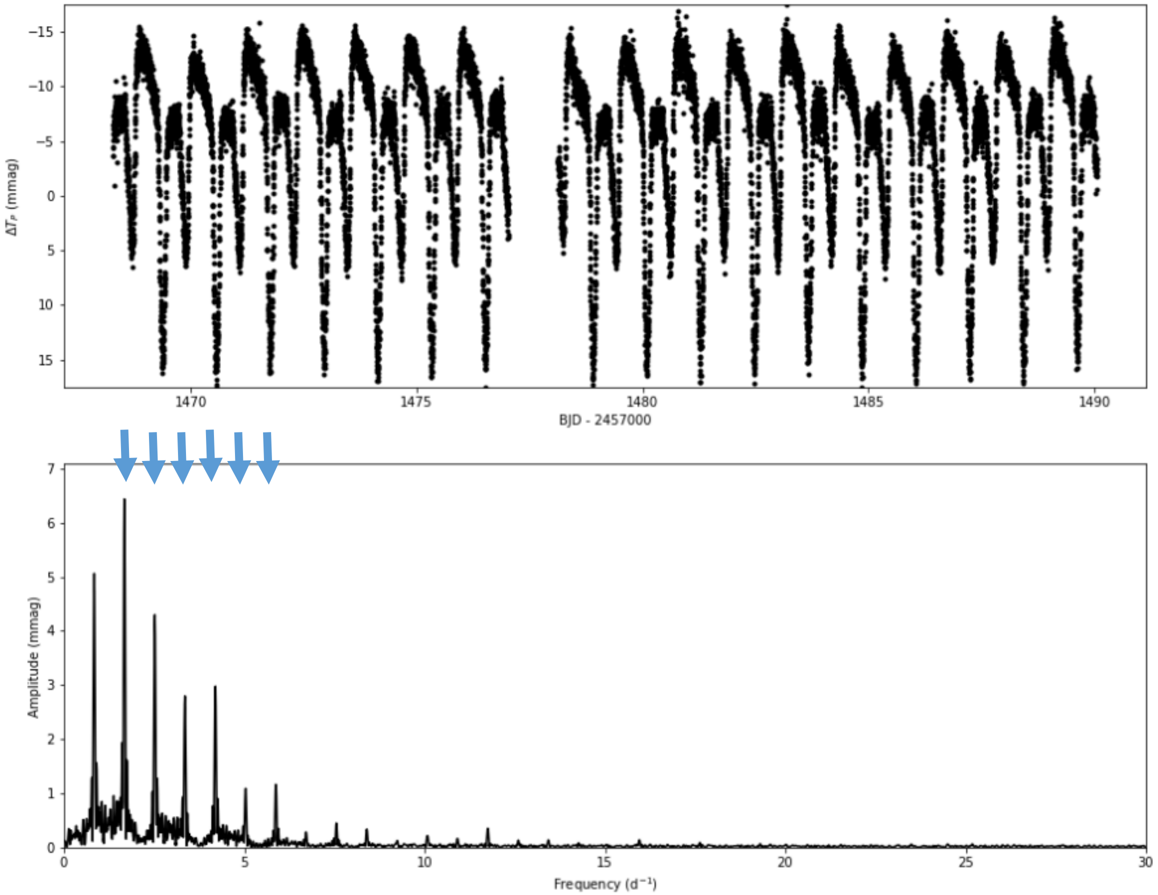
- Periodic signal: fundamental frequency (+ *harmonics*)



Fourier transform – Types of variability

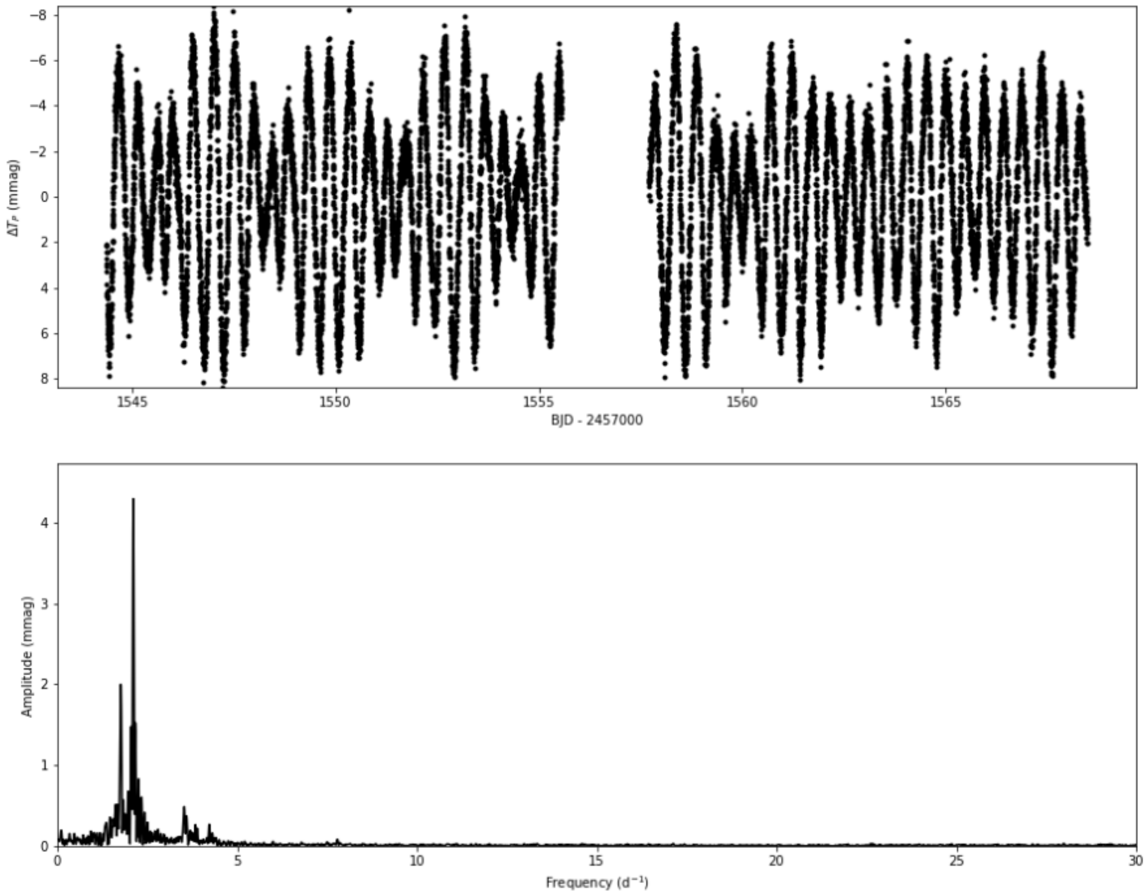
- Periodic signal: fundamental frequency (+ *harmonics*)

Multiple integers of the fundamental frequency



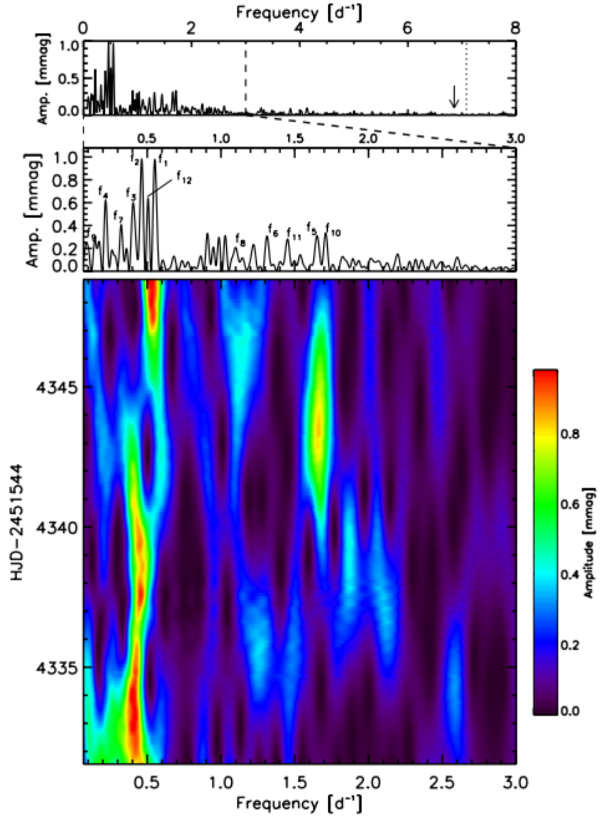
Fourier transform – Types of variability

- Multiperiodic signal: two or more *independent* periods



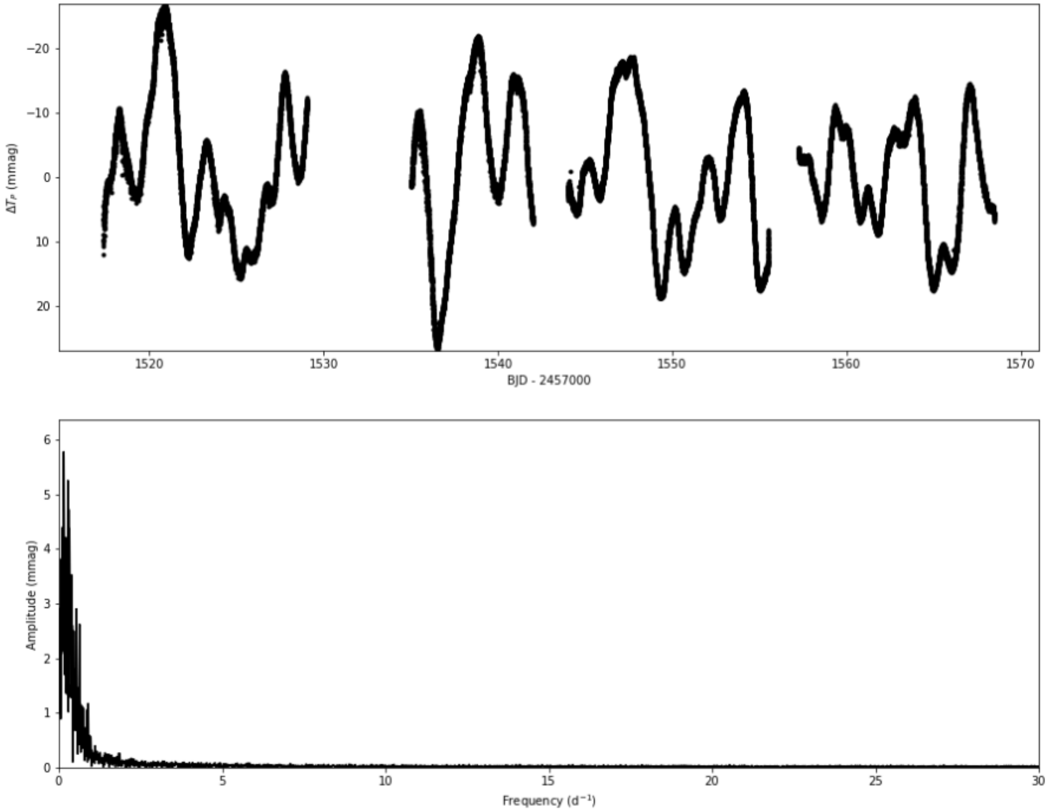
Fourier transform – Types of variability

- Cyclical signal: dominant period(s), but signal does not phase over time (can also be shown with a Short Term Fourier Transform)

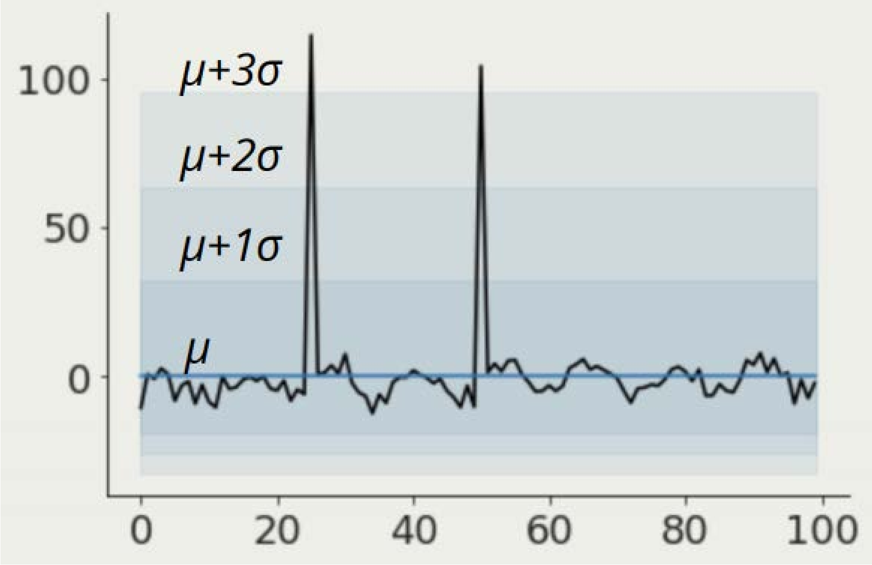


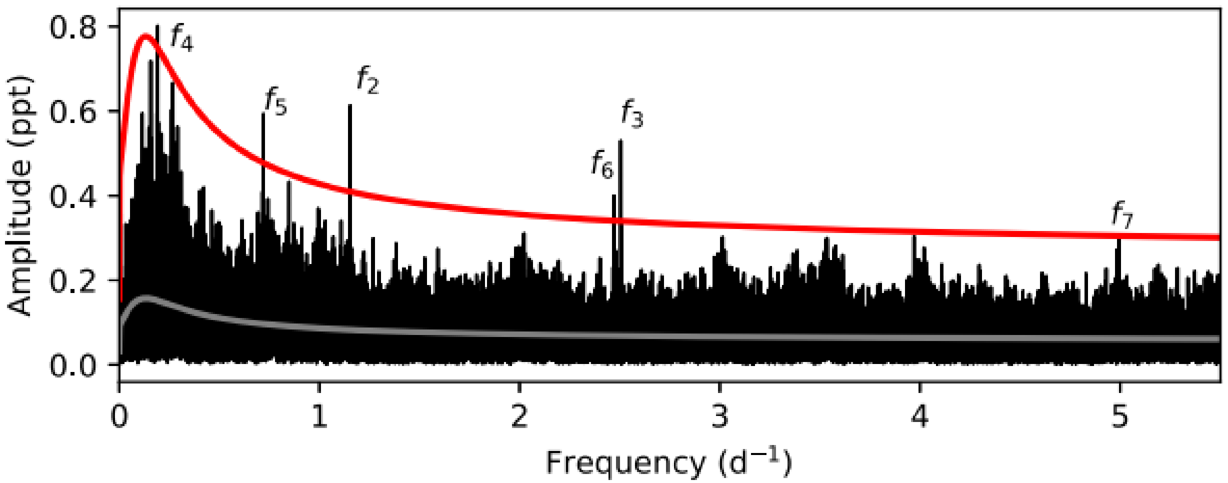
Fourier transform – Types of variability

- Stochastic signal: no clearly defined period, “pink” noise



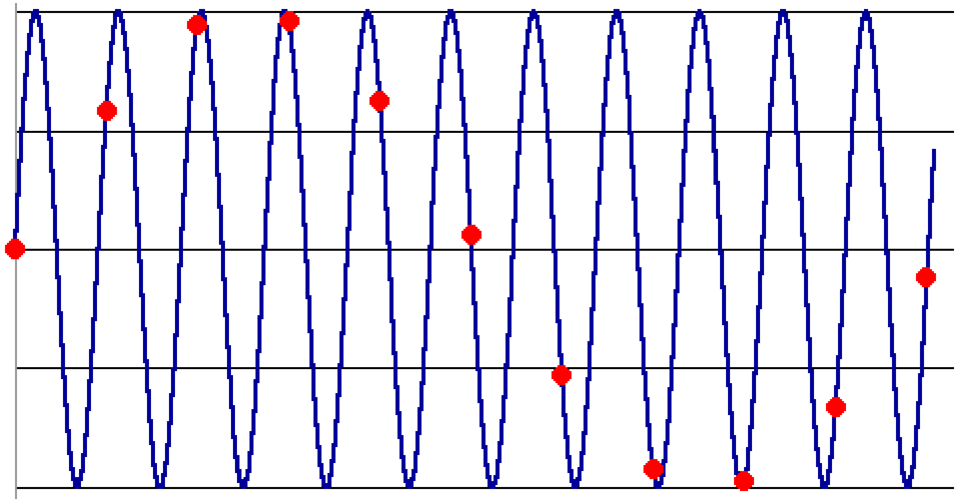
Period detection – thresholding (in frequency space)





Limitations – sampling (aliasing)

- The Fourier transform can appear to include periods which are in fact due to the sampling



Limitations – sampling (cadence)

- Signals with a frequency comparable or higher than the sampling frequency will not be properly detectable

- The Nyquist frequency determines the maximum frequency that is sufficiently sample and is defined as half the sampling frequency, or in other words, given a sampling rate of Δt:

$$f_N = \frac{1}{2\Delta t}$$

Limitations – dataset length

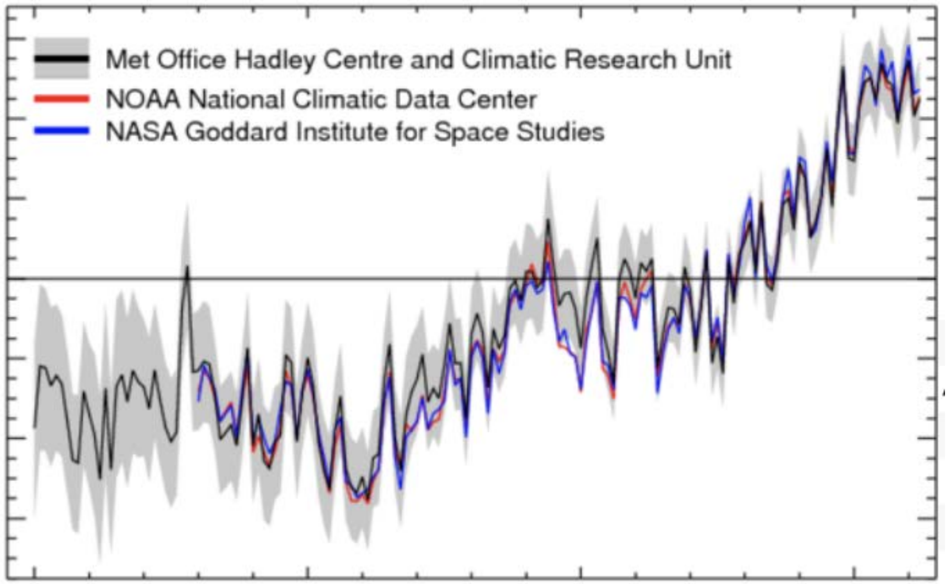
- Similarly, signals with a period that is longer than the total length of the dataset cannot be properly recovered

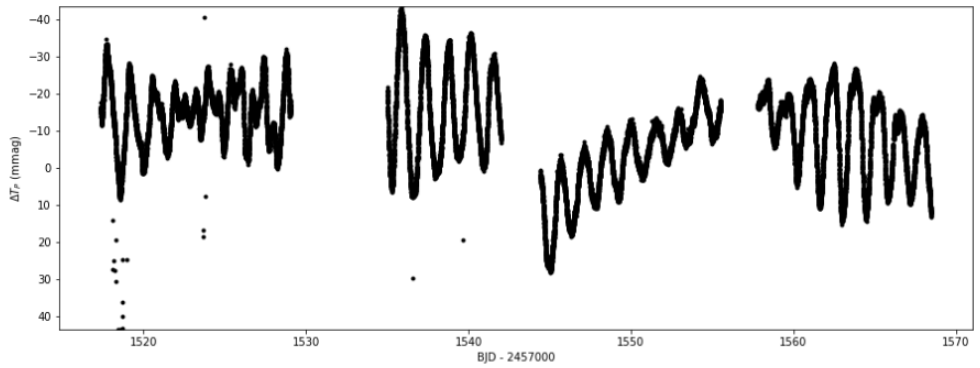
- Given a total length of the dataset, ΔT, the minimum period that can be recovered using a Fourier transform is:

$$f_{min} = \frac{1}{\Delta T}$$

Limitations - detrending

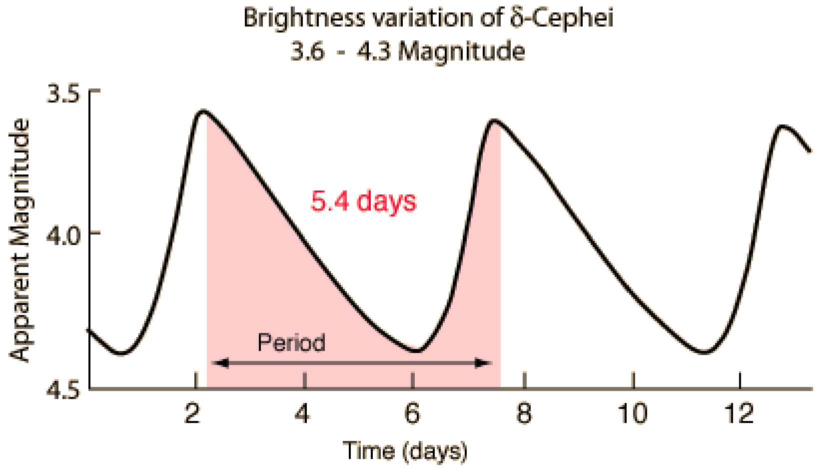
- Therefore, detrending can limit our ability to interpret long-term trends (unless we have some prior information about their origin)



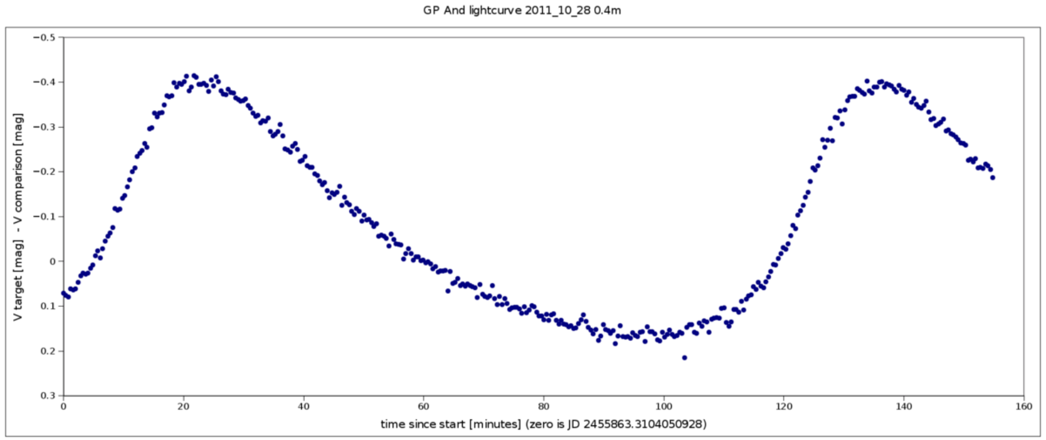


ML applications

- If specific types of variability are expected, they can sometimes be described in terms of more abstract characteristics (period, “shape”, etc.)



“Cepheid” variable star



“delta Scuti” variable star

ML applications

- In this particular case, the shape of the variation is similar, but its amplitude and especially the timescales involved are very different

- The output of a Fourier transform can be used as a feature for certain machine learning methods

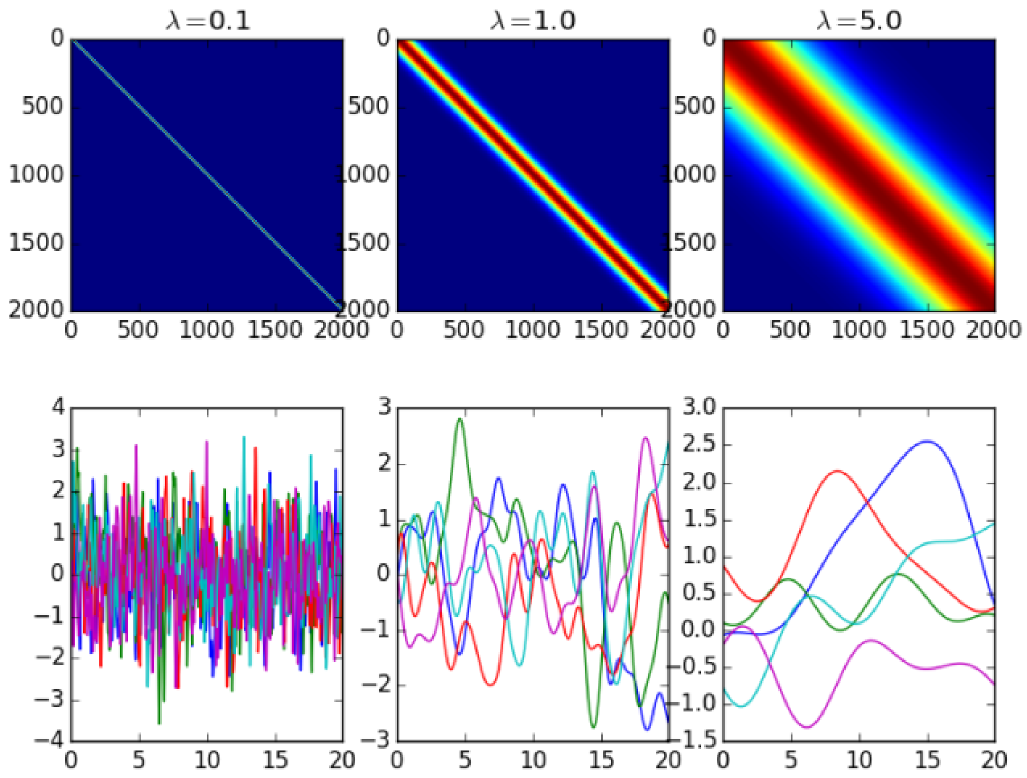
- For instance, recovered periods could then be used in a decision tree to classify variability types

Advanced applications – Gaussian processes

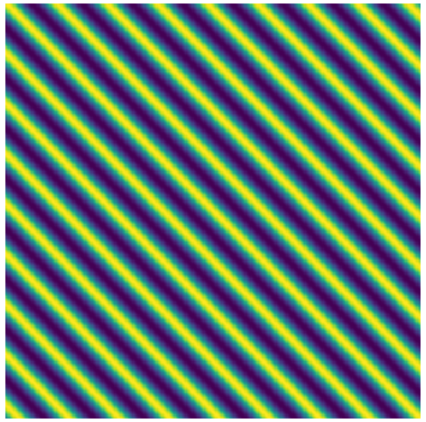
- Time series can also be represented using an autocorrelation matrix

- Certain features of the time series can then be explored using various "kernels", which can use different hyperparameters

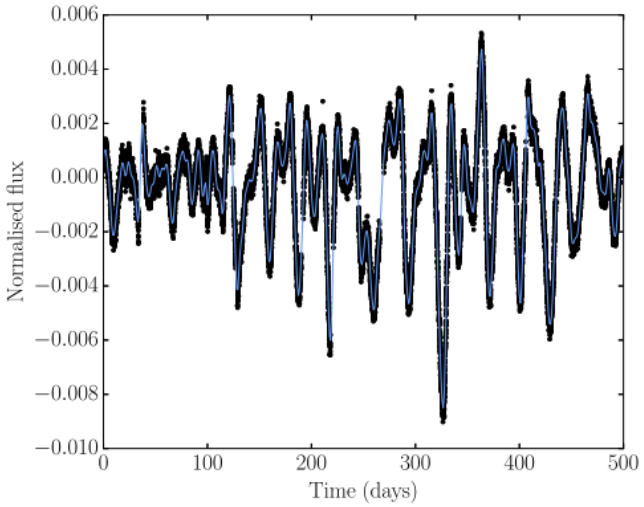
Advanced applications – Gaussian processes



Gaussian kernel informing about "correlation length"



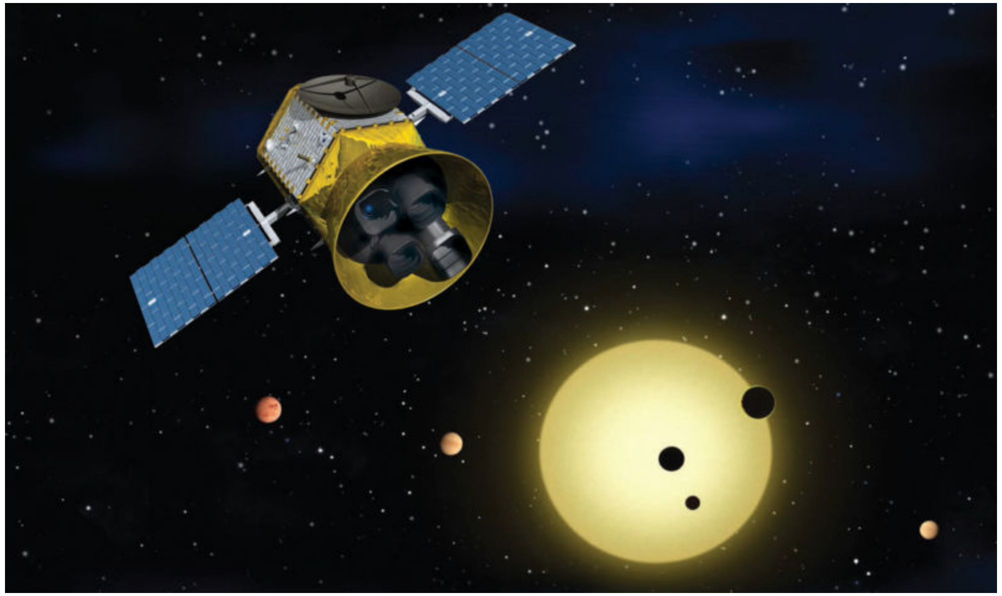
Periodic kernel informing about repeatability

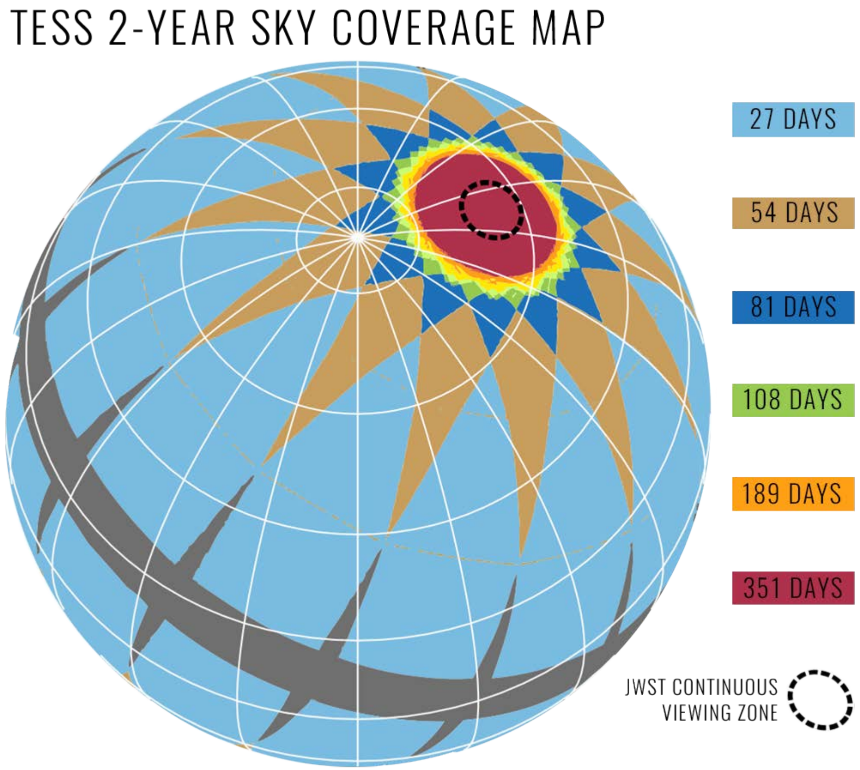


Example of a time series that can be analyzed using both types of kernels combined

Applied example – the TESS mission

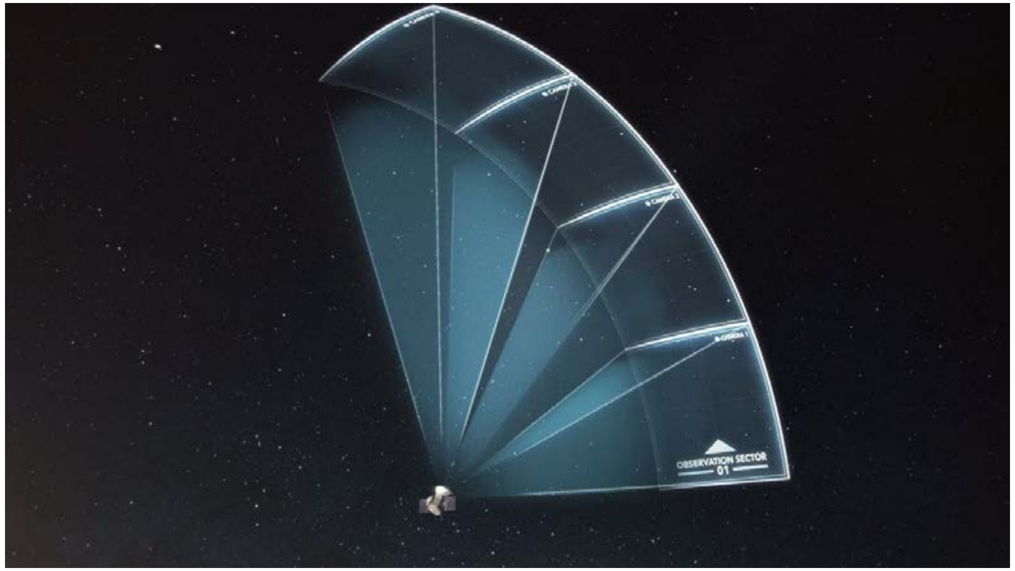
- TESS – Transiting Exoplanet Survey Satellite





TESS – instrumental trends and mitigation

- Certain instrumental drifts can occur within time series obtained with TESS (especially around “gaps”)
- These can be partially removed by looking at time series of other objects nearby on the camera



Homework assignment part 2

- You will use the `LombScargle` package in astropy.stats to compute the periodograms for 3 stars observed with TESS
- For each star, you will have to decide what kind of variability it exhibits
- If a star exhibits periodic variability, you will extract that period and create a phase-folded plot of the time series
- Finally, you will *bin* the phase-folded time series in order to better illustrate the periodic trend

Homework assignment part 2 - hints

```
ls = LombScargle(bjd,1000*dmag)
N = len(bjd)
freq, power = ls.autopower(normalization='psd', nyquist_factor=0.2)
amp = 2*(power/N)**0.5
```

- Binning consists in averaging the time series in *bins* relatively to the exogenous variable (you can use the `df.resample()` method in `pandas`)