Lecciones en Astroinformática Avanzada (Semester 1 2025)

Automatic Classification of Variable Stars (II)

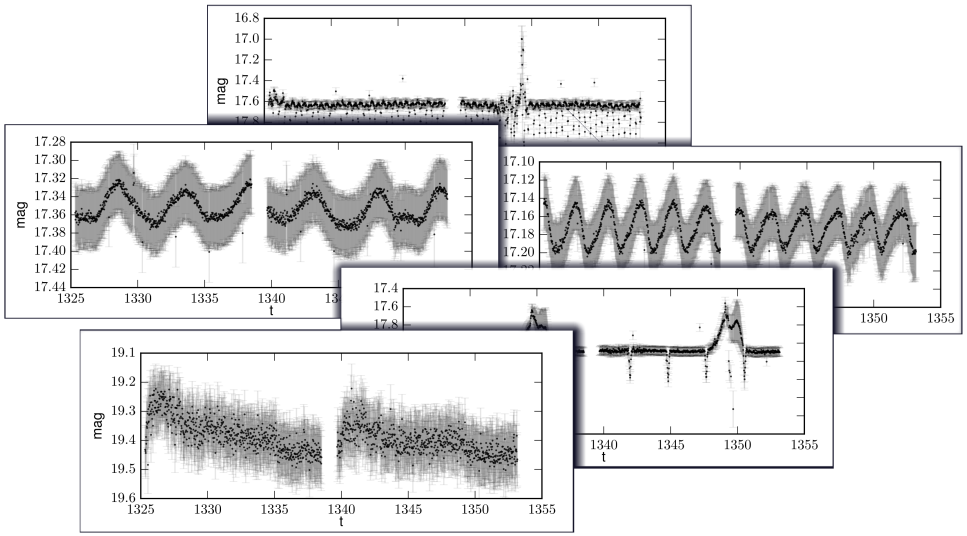
Nina Hernitschek
Centro de Astronomía CITEVA
Universidad de Antofagasta

April 23, 2025

Time Series Analysis

Variable stars can be classified by their **time series**:

Different types of variable stars vary characteristically.



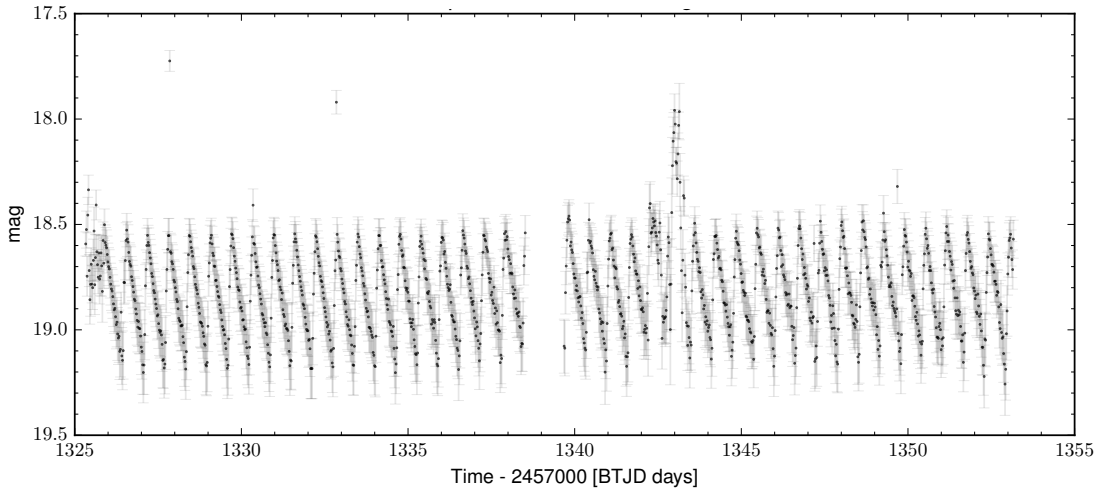
Transforming light curves into **features** will enable automatic classification of variable sources:

calculating features such as colors, amplitudes, periods...

Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

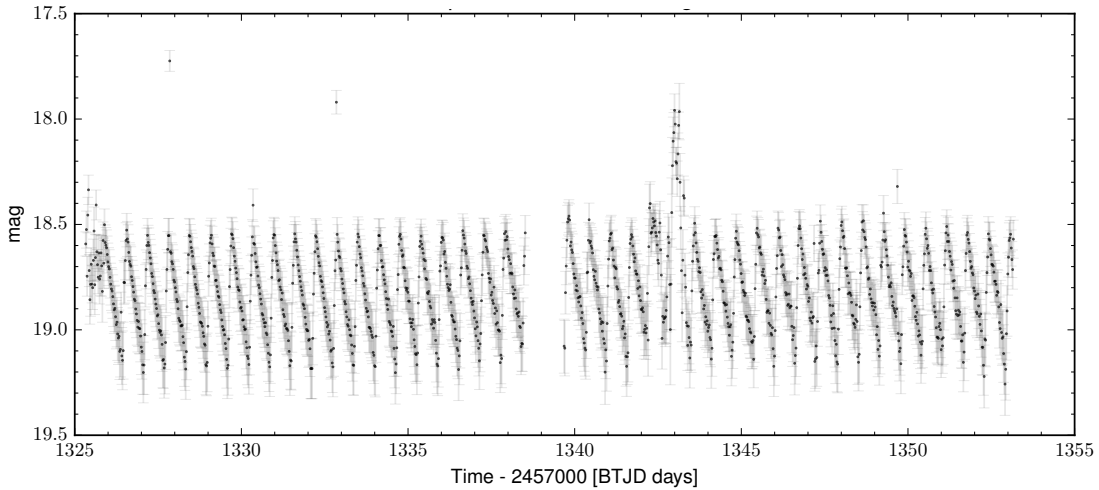
RRab (RR Lyrae type ab):

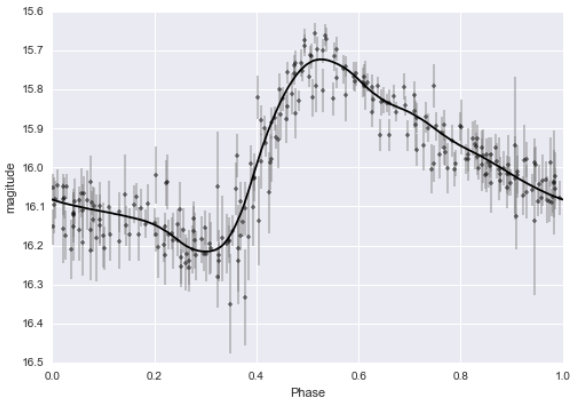


Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

RRab (RR Lyrae type ab):

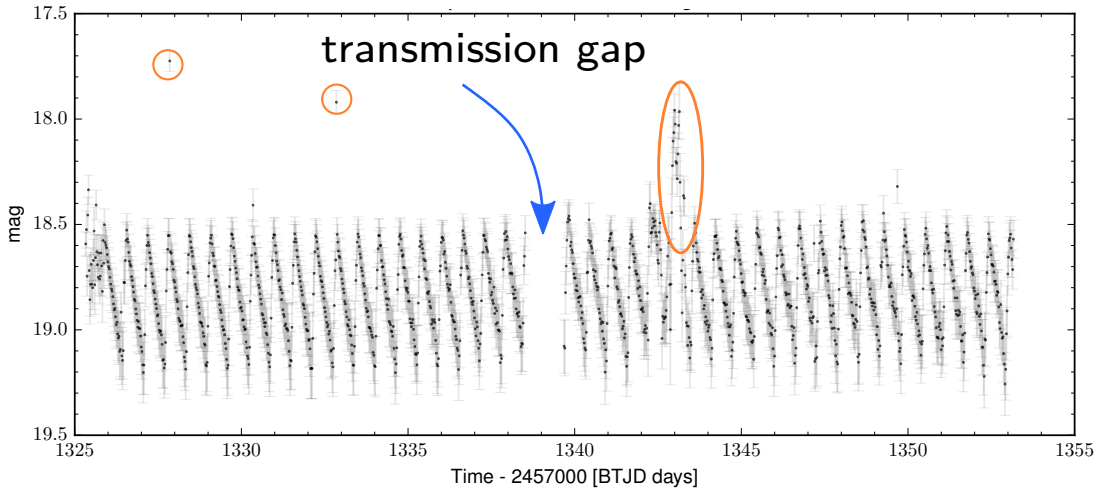


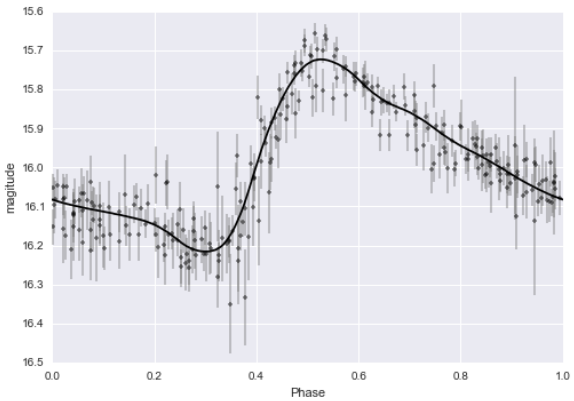


Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

RRab (RR Lyrae type ab):

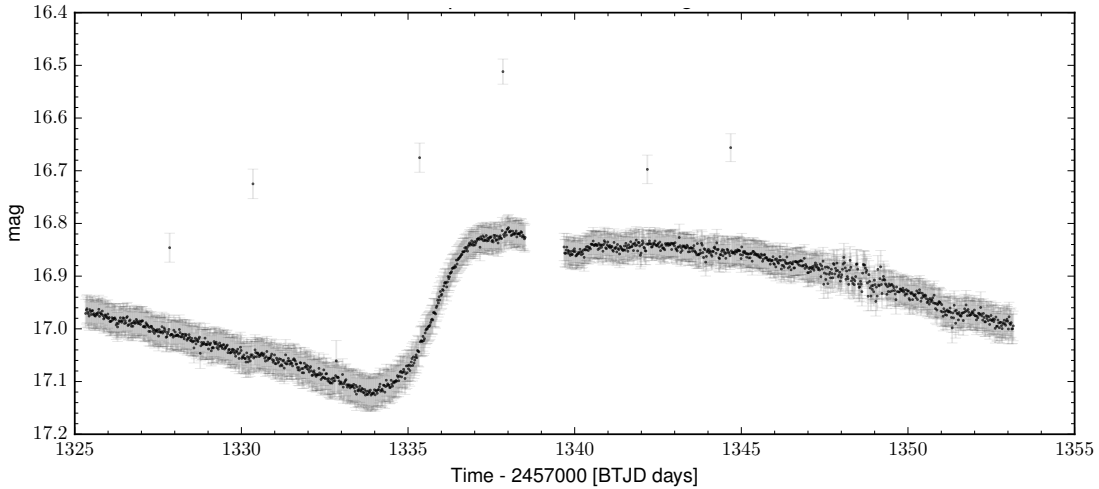




Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

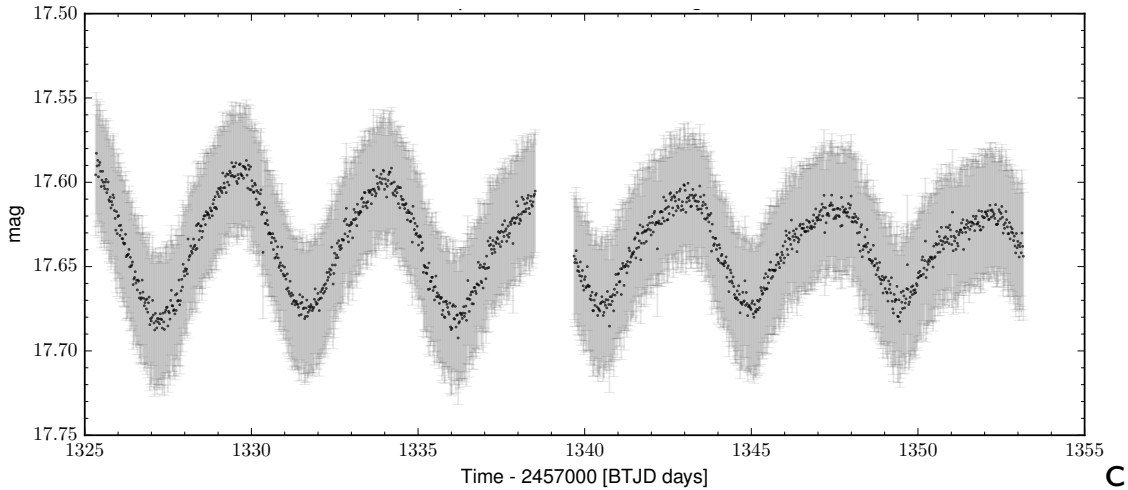
Cepheid:

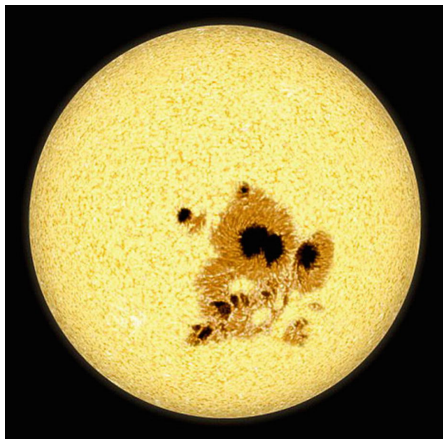


Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

rotational variable star:



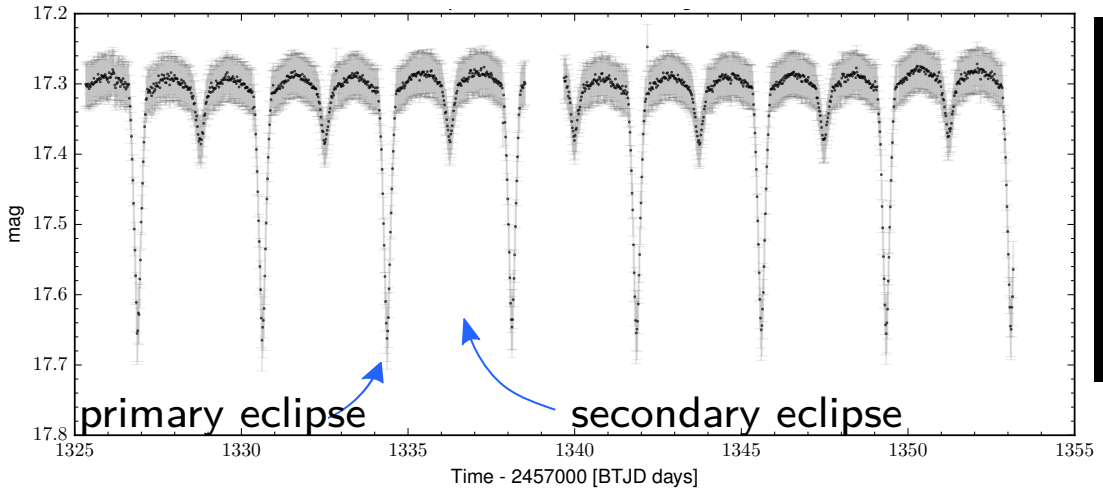


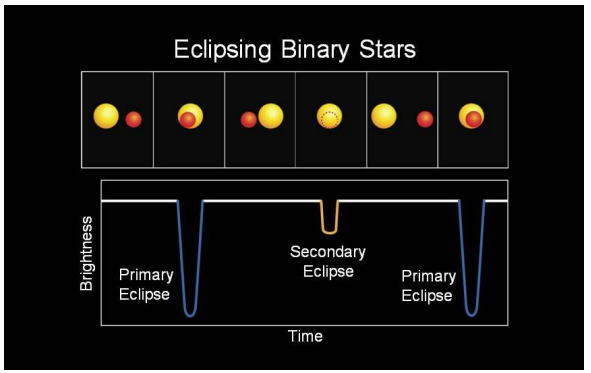
credit: Observer's Guide to Variable Stars, M. Griffiths

Astronomical Light Curves

example light curves with a high **cadence** ($\Delta t = 30$ min) from TESS:

eclipsing binary star:





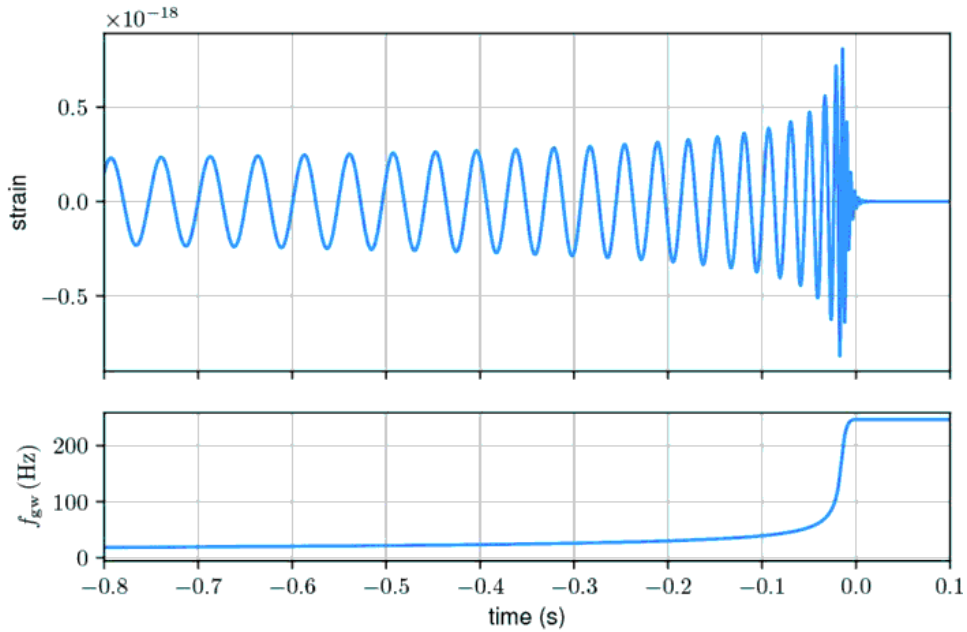
credit: Wikimedia, NASA

When the smaller star partially blocks the larger star, a primary eclipse occurs, and a secondary eclipse occurs when the smaller star is occulted.

Other Astronomical Time Series Data

Light curves show variability from electromagnetic sources.
In addition: gravitational-wave variability as time series data

Gravitational Wave Signal:



Typical GW signal of a compact binary coalescence. The GW strain (above) and the GW frequency (below) are plotted as function of the time before merging. credit: Vallisneri et al. (2015).

Time Series Data

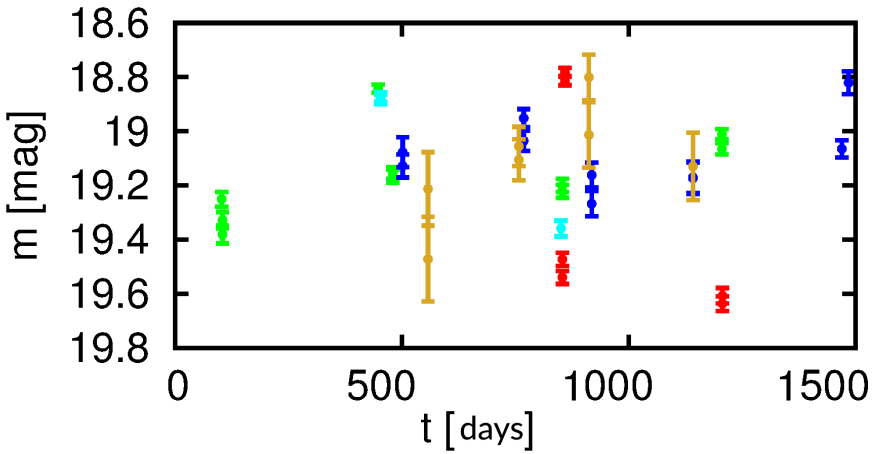
A time series is a sequence of random variables $\{\mathbf{X}_t\}_{t=1,2,\ldots}$.

Thus, a time series is a **series of data points ordered in time**. The time of observations provides a source of additional information to be analyzed.

Astronomical time series are typically assumed to be generated at irregularly spaced interval of time (**irregular time series**).

Time series can have one or more variables that change over time. If there is only one variable varying over time, we call it **univariate time series**. If there is more than one variable it is called **multivariate time series**.

example: light curve from multi-band survey

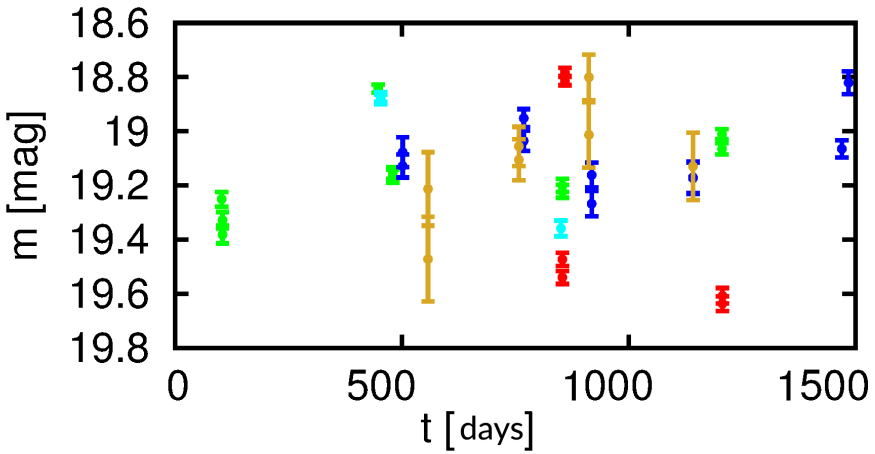


Characteristics of Astronomical Time Series Data

Astronomical time series data in general is:

- irregularly sampled
- multivariate
- not sampled to fully characterize the variability process
- not an independent random variable in their y values: often $y_{i+1} = f(y_i)$

example: light curve from multi-band survey

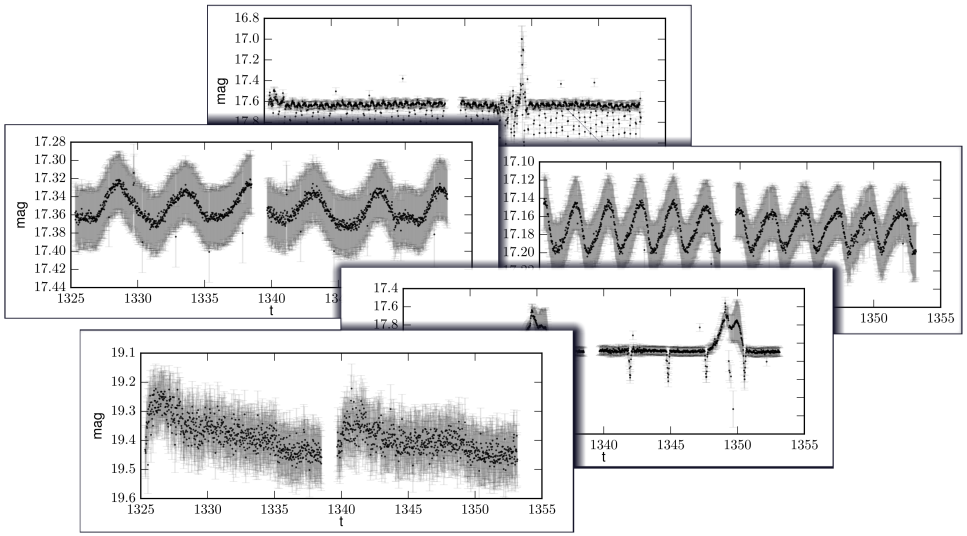


Goals of Time Series Analysis

Time series analysis extracts meaningful statistics and other characteristics of the dataset in order to understand it.

The main tasks of time series analysis are:

- **characterize** the temporal correlation between different values of y, including its significance
  example: classification of variable sources



- **forecast** (predict) future values of y
  example: transient detection, e.g. early supernovae detection

Goals of Time Series Analysis

When dealing with time series data, the first question we ask is ***Does the time series vary over some timescale?*** (if not, there is no point doing time series analysis)

Variability does not mean necessarily periodicity.
Stochastic processes are variable over some timescale, but are distinctly aperiodic through the inherent randomness.

Goals of Time Series Analysis

When dealing with time series data, the first question we ask is ***Does the time series vary over some timescale?*** (if not, there is no point doing time series analysis)

Variability does not mean necessarily periodicity. Stochastic processes are variable over some timescale, but are distinctly aperiodic through the inherent randomness.

If we find that a source is variable (almost all astronomical sources are), then time-series analysis has two main goals:

1. Characterize the temporal correlation between different values of y (i.e., characterize the light curve), e.g. by learning the parameters for a model.
2. Predict future values of y.

Detecting Variability

For known and Gaussian uncertainties, we can compute χ^2 and the corresponding p values for variation in a signal.
For a sinusoidal variable signal $A\sin(\omega t)$, with homoescedastic measurement uncertainties, the data model would be

$$y(t) = A\sin(\omega t) + \epsilon$$

where $\epsilon \sim G(0, \sigma)$ and N observations. The overall data variance is then $V = \sigma^2 + A^2/2$.

Detecting Variability

For known and Gaussian uncertainties, we can compute χ^2 and the corresponding p values for variation in a signal.
For a sinusoidal variable signal $A\sin(\omega t)$, with homoescedastic measurement uncertainties, the data model would be

$$y(t) = A\sin(\omega t) + \epsilon$$

where $\epsilon \sim G(0, \sigma)$ and N observations. The overall data variance is then $V = \sigma^2 + A^2/2$.

If $A = 0$ (no variability, with $\overline{y} = 0$):

- $\chi^2_{\rm dof} = N^{-1}\sum_j (y_j/\sigma)^2 \sim V/\sigma^2$
- $\chi^2_{\rm dof}$ has expectation value of 1 and std dev of $\sqrt{2/N}$

Detecting Variability

If $|A| > 0$ (variability):

- $\chi^2_{\rm dof}$ will be larger than 1.
- probability that $\chi^2_{\rm dof} > 1 + 3\sqrt{2/N}$ is about 1 in 1000 (i.e., $> 3\sigma$ above 1, where 3σ is 0.997).

Detecting Variability

If $|A| > 0$ (variability):

- $\chi^2_{\rm dof}$ will be larger than 1.
- probability that $\chi^2_{\rm dof} > 1 + 3\sqrt{2/N}$ is about 1 in 1000 (i.e., $> 3\sigma$ above 1, where 3σ is 0.997).

If this false-positive rate (1 in a 1000) is acceptable (because even without variability 1 in 1000 will be above this threshold) then the minimum detectable amplitude is $A > 2.9\sigma/N^{1/4}$ (from $V/\sigma^2 = 1 + 3\sqrt{2/N}$, so that $A^2/2\sigma^2 = 3\sqrt{2/N}$).

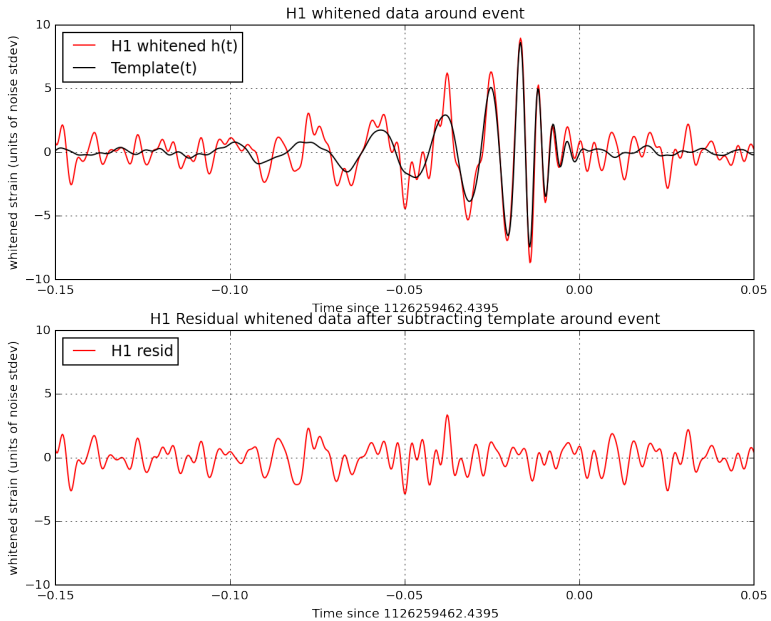
Depending on how big your sample is, you may want to choose a higher threshold. E.g., for 1 million non-variable stars, this criterion would identify 1000 as variable.

1. For $N = 100$ data points (not 100 objects), the minimum detectable amplitude is $A_{\rm min} = 0.92\sigma$
2. For $N = 1000$, $A_{\rm min} = 0.52\sigma$

Detecting Variability

We do this under the assumption of the null hypothesis of no variability. If instead we have a model, we can perform a **matched filter analysis** by correlating a known template with an unknown signal to detect the presence of the template in the unknown signal

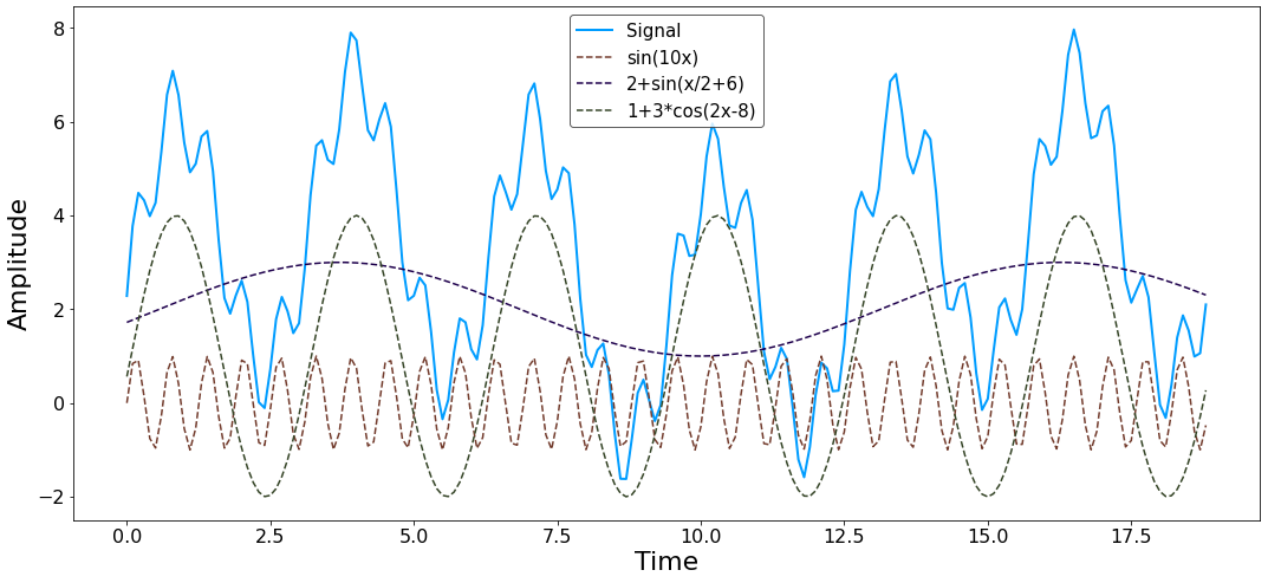
example: gravitational wave event GW150914



credit: https://www.gw-openscience.org/tutorials/

Fourier Analysis

Fourier analysis plays a **major role** in the analysis of time series data. In Fourier analysis, general **functions are approximated by integrals or sums of trigonometric functions**.



For periodic functions, such as periodic light curves in astronomy, often a relatively small number of terms (less than 10) suffices to reach an approximation precision level similar to the measurement precision.

Fourier Analysis

The **Fourier transform (FT)** $H(f)$ of function $h(t)$ is defined as

$$H(f) = \int_{-\infty}^{\infty} h(t) \exp(-i2\pi ft) dt$$

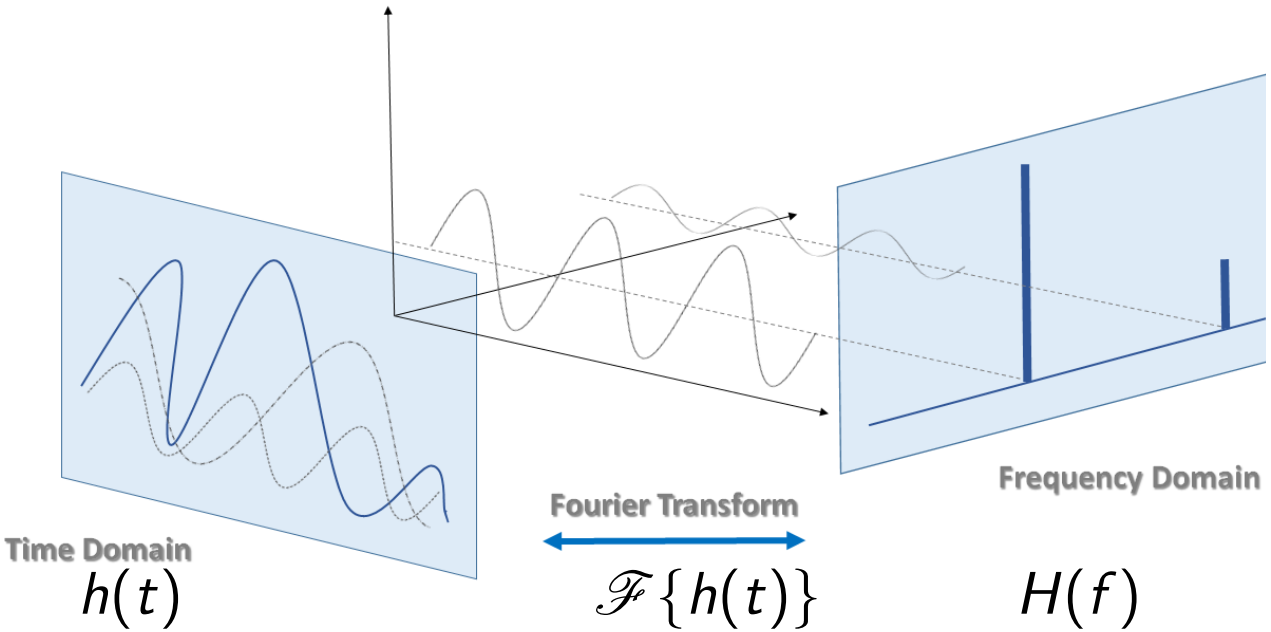
with inverse transformation

$$h(t) = \int_{-\infty}^{\infty} H(f) \exp(-i2\pi ft) df$$

where t is time and f is frequency (for time in seconds, the unit for frequency is hertz, or Hz).

Fourier Analysis

In other words, FT transforms a periodic function in **Time Domain** to a function in **Frequency Domain**:

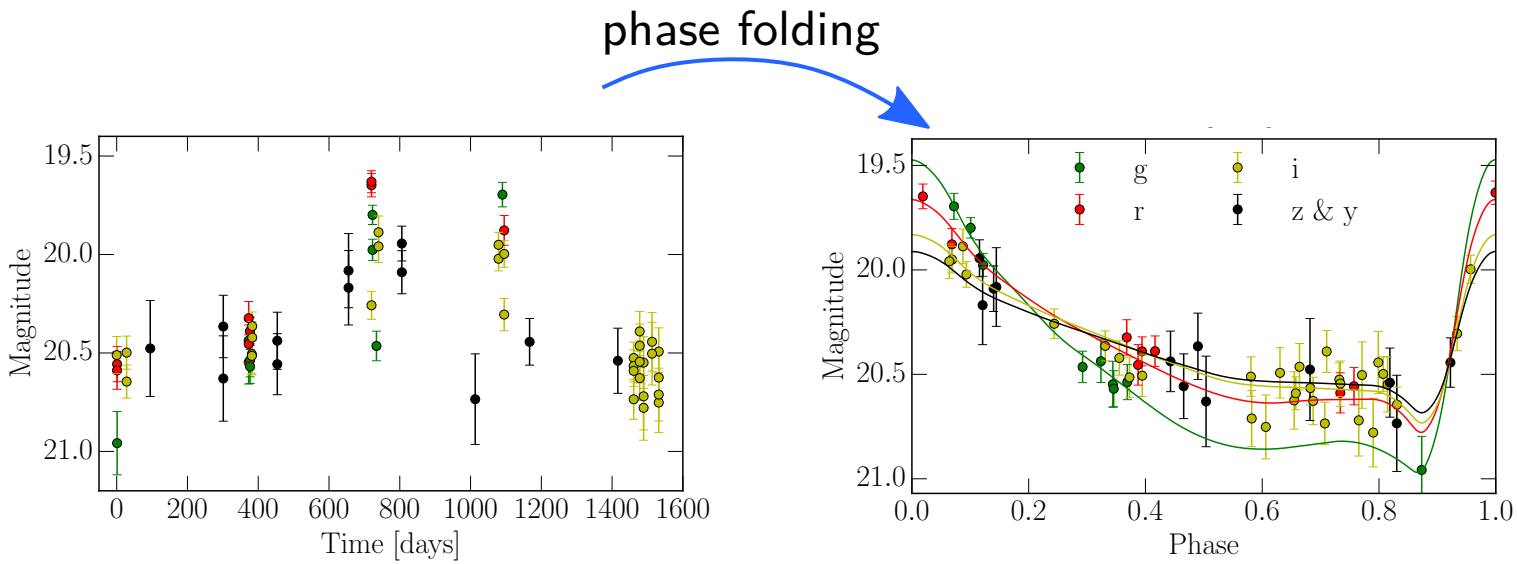


Detecting Periodic Signals

many objects/ systems have periodic signals: e.g., pulsars, RR Lyrae, Cepheids, eclipsing binaries

For a periodic signal, if the period is known

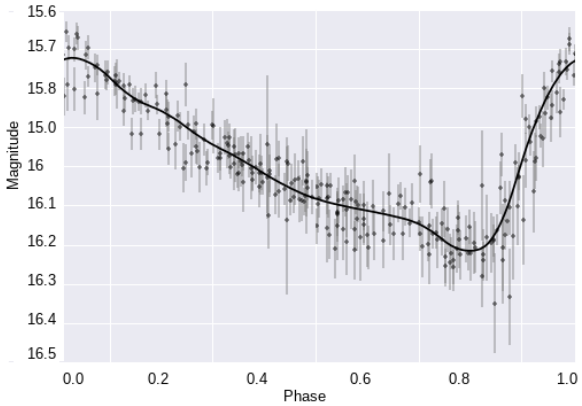
- we can write $y(t + P) = y(t)$, where P is the period.
- we can create a **phased light curve** that plots the data as function of phase: $\phi = \frac{t}{P} - \mathrm{int}\left(\frac{t}{P}\right)$ with $\mathrm{int}(x)$ being the integer part of x.



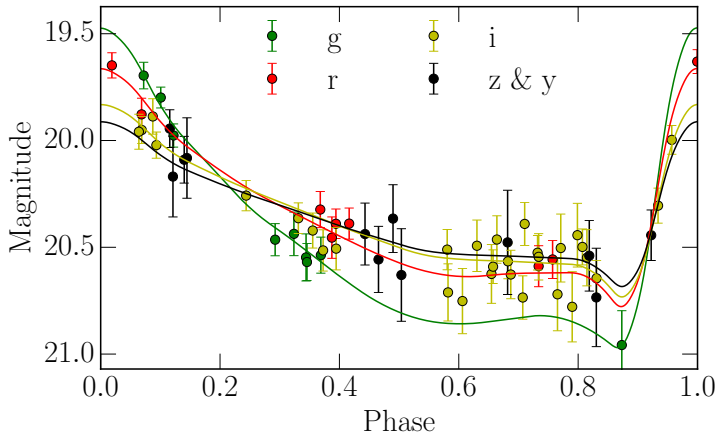
Detecting Periodic Signals

for well-sampled, high-cadence data: easy, standard methods can be applied

for sparse, low-cadence data: harder, specialized methods like template fitting necessary



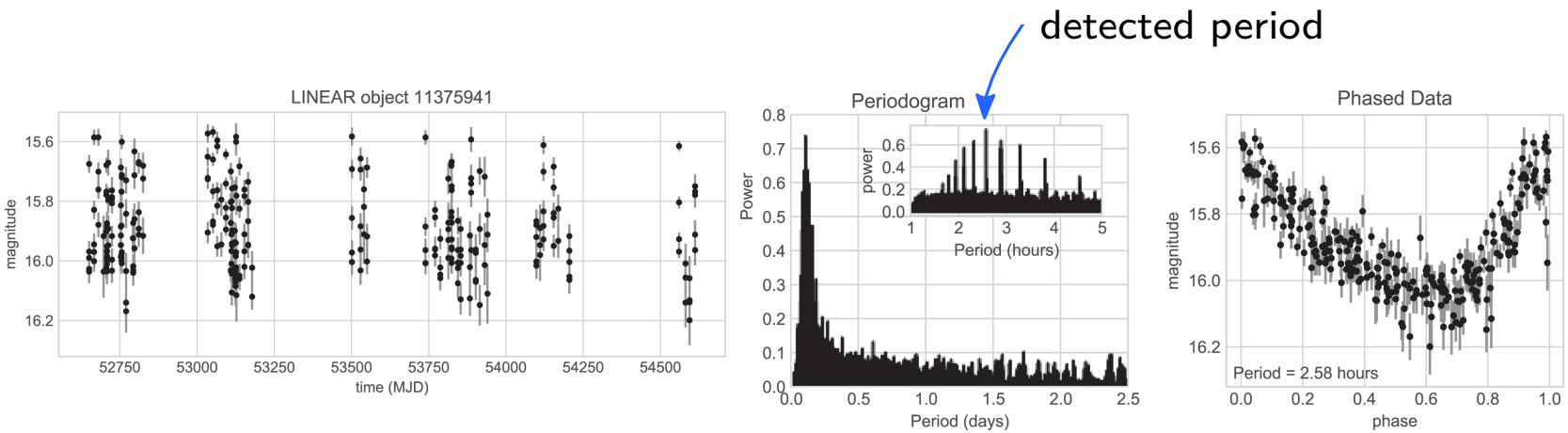
vs.



measure the period and amplitude in the face of both noisy and incomplete data

Detecting Periodic Signals

A **periodogram** is a plot of the *power* in the time series at each possible period (as illustrated below):



left panel: observed light curve from LINEAR object ID 11375941
middle panel: periodogram computed from the light curve
right panel: light curve folded over the detected 2.58 hr period
credit: VanderPlas (2018)

Detecting Periodic Signals

The periodogram is defined as

$$P(\omega) = \frac{1}{N}\left[\left(\sum_{j=1}^{N} y_j \sin(\omega t_j)\right)^2 + \left(\sum_{j=1}^{N} y_j \cos(\omega t_j)\right)^2\right]$$

The **best value** ω is given by

$$\chi^2(\omega) = \chi_0^2\left[1 - \frac{2}{N\,V}P(\omega)\right],$$

where $P(\omega)$ is the periodogram, V the variance of the data y, and χ_0^2 is the χ^2 for the null-hypothesis model $y(t) = \text{const}$:

$$\chi_0^2 = \frac{1}{\sigma^2}\sum_{j=1}^{N} y_j^2 = \frac{N\,V}{\sigma^2}$$

Detecting Periodic Signals

We can renormalize the periodogram, defining the **Lomb-Scargle periodogram** as

$$P_{\mathrm{LS}}(\omega) = \frac{2}{NV} P(\omega),$$

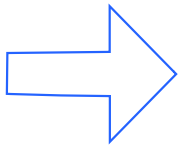
where $0 \leq P_{\mathrm{LS}}(\omega) \leq 1$.

Detecting Periodic Signals

How to determine if our source is variable or not:

- compute Lomb-Scargle periodogram $P_{\rm LS}(\omega)$
- model the odds ratio for our variability model vs. a non-variability model.

If our variability model is correct, then the **peak** of $P(\omega)$ (found by grid search) gives the best period ω.



The Lomb-Scargle periodogram (Lomb 1976; Scargle 1982) is the **standard method** to search for periodicity in unevenly-sampled time-series data.

Classification

Detected **features** are what classification relies on.

How does classification happen?

Data products?

Science cases?

The Data

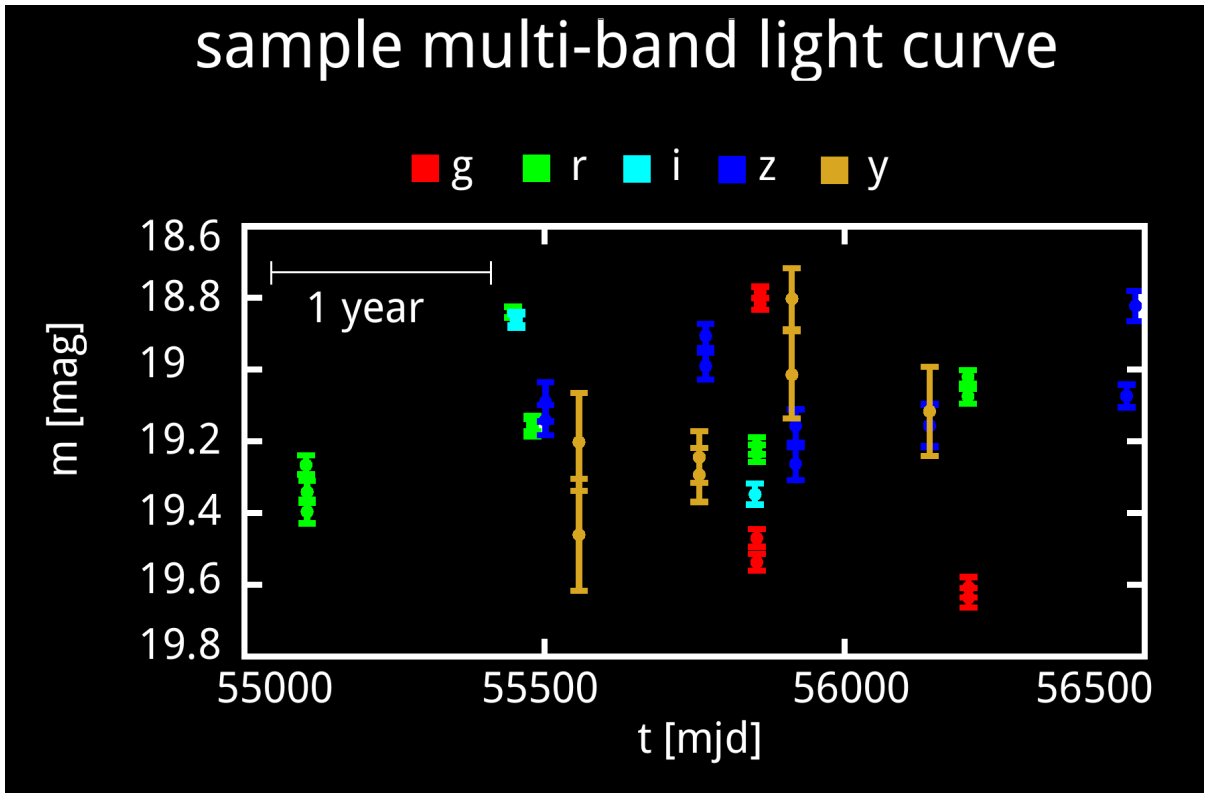
The Pan-STARRS1 3π Survey in one sentence:

An optical/near-IR survey covering 3/4 of the sky in non-simultaneous *grizy* to $r\sim$21.8 with $\sim$70 visits over 5.5 years.

The Data

The Pan-STARRS1 3π Survey in one sentence:

An optical/near-IR survey covering 3/4 of the sky in non-simultaneous *grizy* to $r\sim$21.8 with $\sim$70 visits over 5.5 years.



The Data

goal: a catalog of variable sources in PS1 3π

The Data

goal: a catalog of variable sources in PS1 3π

to **model a survey**, tools are needed for

- describing data quality $\rightarrow$ outlier might fake or hide true variability
- describing light curve characteristics $\rightarrow$ "features" with scientific relevance
- classifying sources $\rightarrow$ catalogs others can use
- finding substructure $\rightarrow$ clumps, overdensities, ... the science we want to do

The Data

goal: a catalog of variable sources in PS1 3π

to **model a survey**, tools are needed for

- describing data quality $\rightarrow$ outlier might fake or hide true variability
- describing light curve characteristics $\rightarrow$ "features" with scientific relevance
- classifying sources $\rightarrow$ catalogs others can use
- finding substructure $\rightarrow$ clumps, overdensities, ... the science we want to do

challenge:
processing $\sim 10^9$ rather sparse, noisy light curves

The Methods

Classification of variable sources relies fundamentally on algorithms quantifying different aspects of variability.

feature extraction:

light curve $\xrightarrow{\text{signal processing}}$ numbers

$\Rightarrow$ features should be as discriminative and informative as possible

The Methods

Classification of variable sources relies fundamentally on algorithms quantifying different aspects of variability.

feature extraction:

light curve $\xrightarrow{\text{signal processing}}$ numbers

$\Rightarrow$ features should be as discriminative and informative as possible

challenges:

- non-simultaneous multi-band data
- noise & uncertainties
- foreground effects
- not all variables are periodic: QSOs, supernovae...
- time-sampling can act as window function (hiding variability)
- many period-finders are computationally very expensive: pre-selection

The Methods

multi-band structure-function variability model:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within Δt?

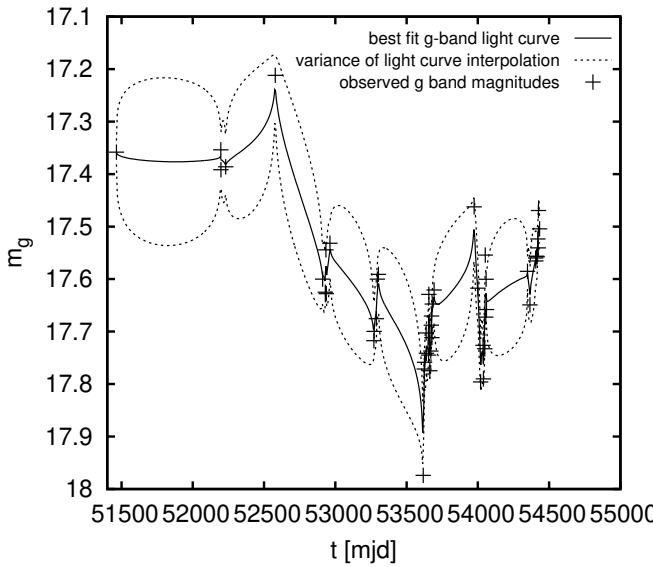
$$\mathrm{V}(|\Delta t|) \equiv \mathrm{E}[(m(t) - m(t + \Delta t))^2]$$

assume functional form

$$V(\Delta t) \overset{\text{model}}{\equiv} \omega_i(\lambda_i)\omega_j(\lambda_j)\left(1 - \exp\left[-\frac{|\Delta t|}{\tau}\right]\right)$$

with $\tilde{m}_\lambda(t) = m_\lambda(t) - \bar{m}_\lambda$, $\omega_k(\lambda_k) = \omega_r \left(\frac{\lambda_k}{\lambda_r}\right)^\alpha$

(Hernitschek et al. 2016)



The Methods

multi-band structure-function variability model:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within Δt?

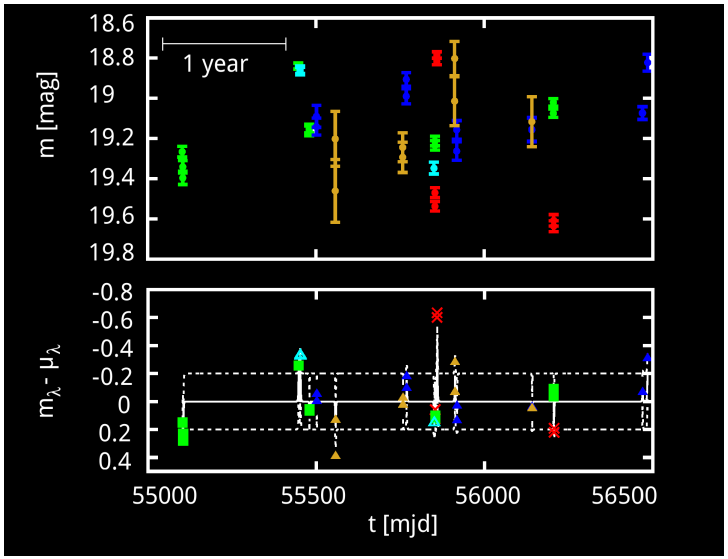
$\Rightarrow$ fit $\Rightarrow$ characteristic variability timescale & amplitude

The Methods

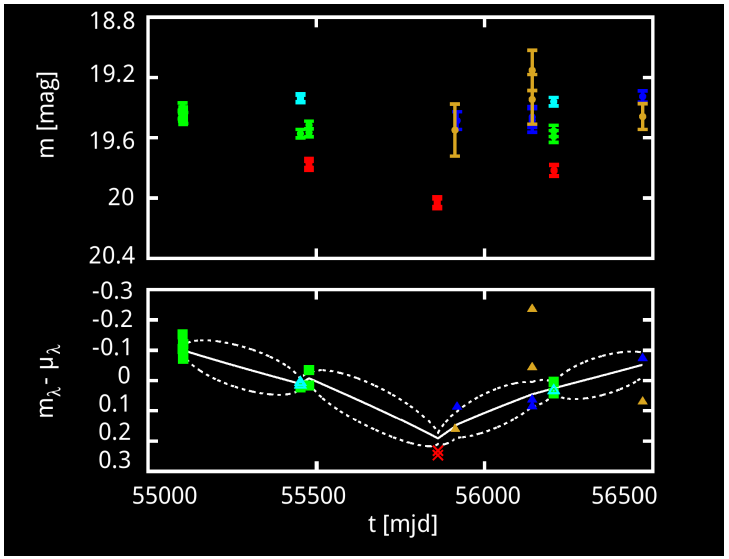
multi-band structure-function variability model:

describe light curves as stochastic processes: how much should you expect a multi-band source to vary within Δt?

$\Rightarrow$ fit $\Rightarrow$ characteristic variability timescale & amplitude



RR Lyrae, ω_r=0.3, τ=1.5 days



QSO, ω_r=0.13 , τ=560 days

The Methods

period fitting:

RR Lyrae period crucial for distance determination:
Period-Luminosity-Metallicity (PLZ) relation

L=f(P,Z), D=f(L,m) $\Rightarrow$ D = f(m,P,Z)

$\Rightarrow$ goal: 3D map of Milky Way's RR Lyrae
from measured magnitude m and derived period P (and Z)

The Methods

period fitting:

RR Lyrae period crucial for distance determination:
Period-Luminosity-Metallicity (PLZ) relation

L=f(P,Z), D=f(L,m) $\Rightarrow$ D = f(m,P,Z)

$\Rightarrow$ goal: 3D map of Milky Way's RR Lyrae
from measured magnitude m and derived period P (and Z)

However:

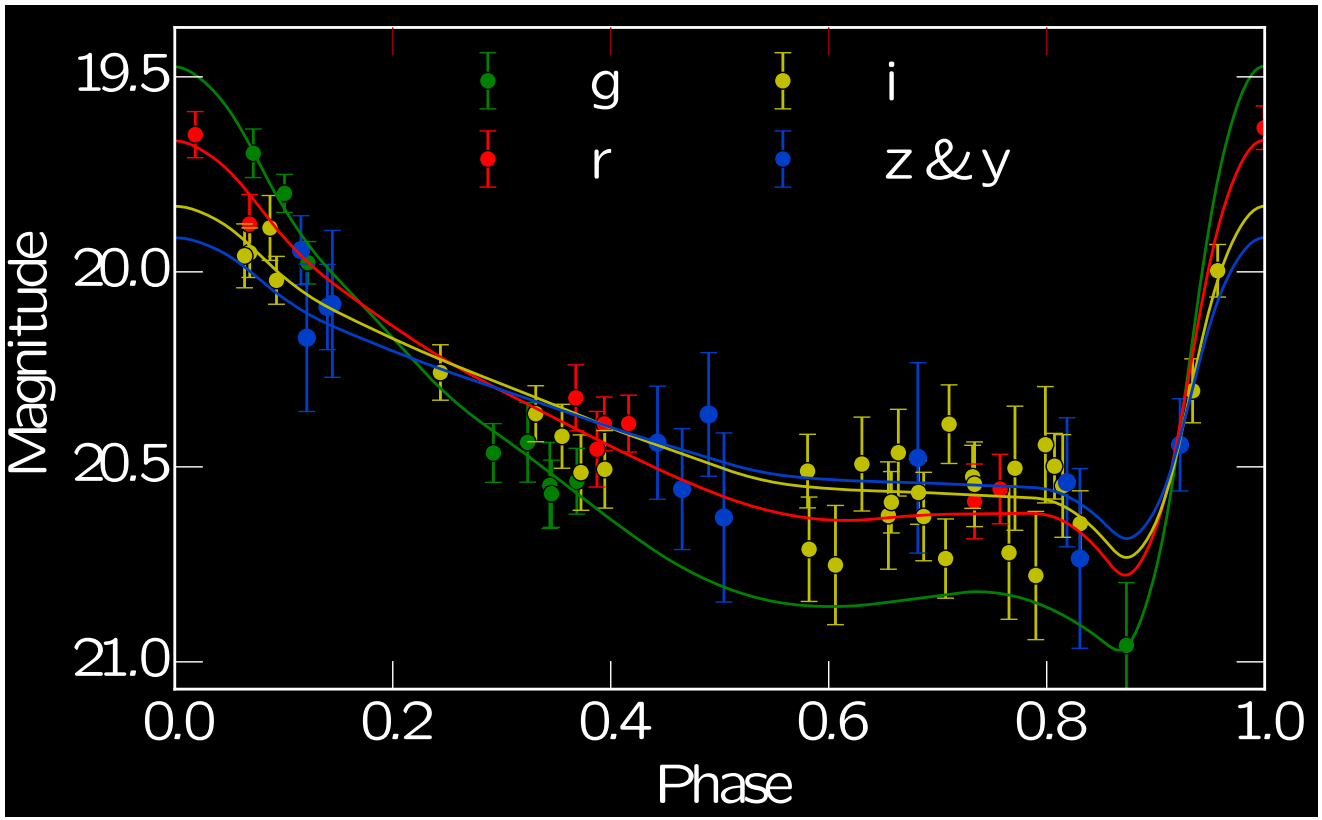
- sparse light curves
- computationally expensive

$\Rightarrow$ apply methods suitable for sparse and unevenly sampled multi-band data to **pre-selected sources**

The Methods

Period and Phase from Template Fitting

example: RR Lyrae period/phase fitting, using light curve templates from SDSS Stripe 82 (Sesar et al. 2010)



Results

10^9 light curves from PS1 3π

outlier cleaning using a machine-learning method

feature extraction: structure functions, mean magnitudes

first machine-learning classification

RR Lyrae

1.5×10^5 RR Lyrae candidates (80% purity, 80% completeness)

feature extraction: template fitting

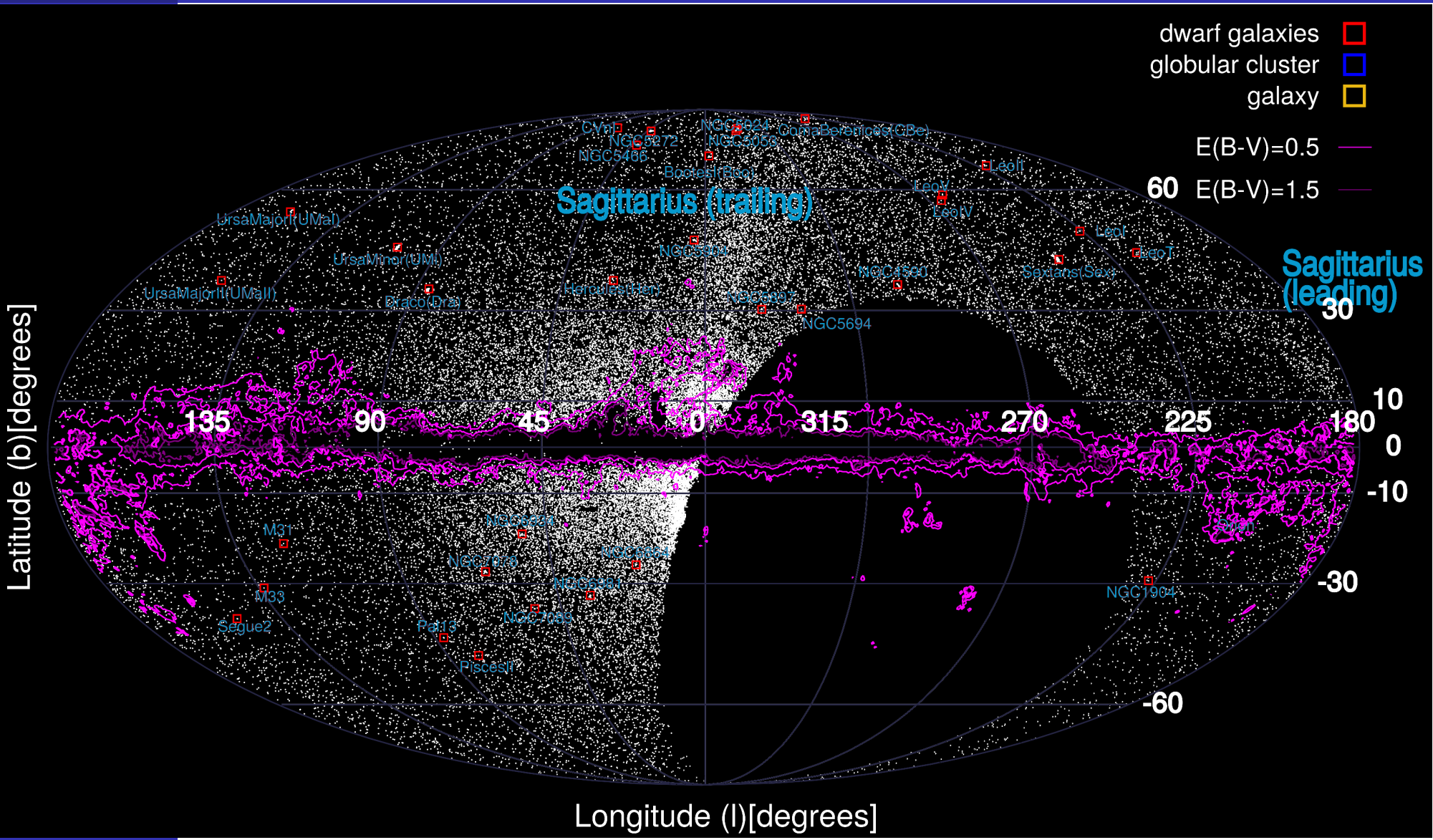
second machine-learning classification

4.4×10^4 RR Lyrae stars (90% purity, 80% completeness), **distances** up to $\sim$140 kpc, $\sigma = 3\%$

QSO

3.8×10^6 QSO candidates (85% purity, 85% completeness)

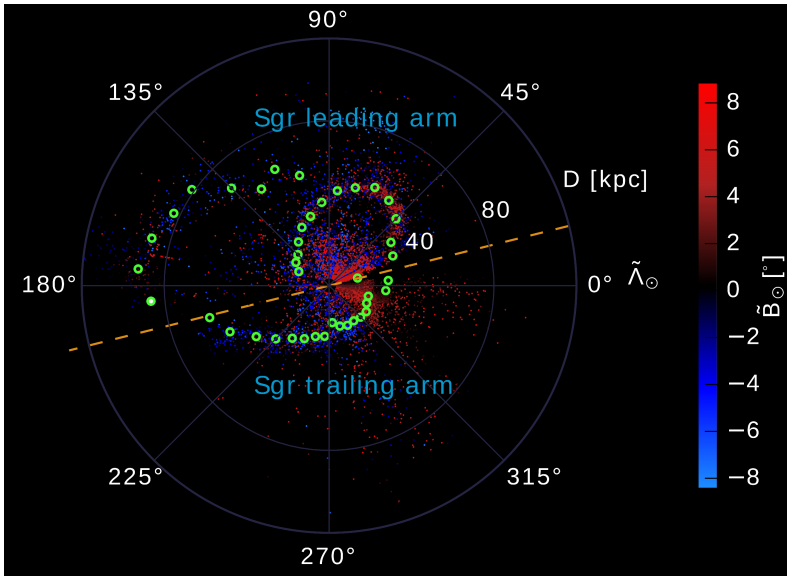
The Results

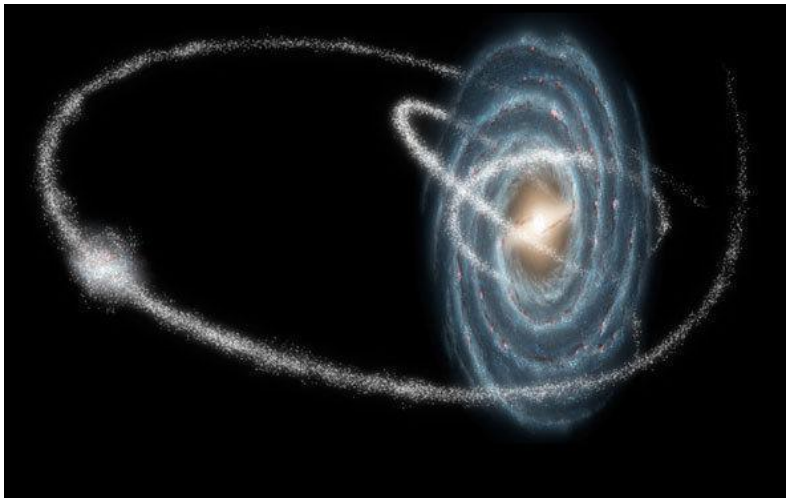

dwarf galaxies
globular cluster
galaxy
E(B-V)=0.5
E(B-V)=1.5
Sagittarius (trailing)
Sagittarius (leading)
Latitude (b)[degrees]
Longitude (l)[degrees]
60
30
10
0
-10
-30
-60
135
90
45
0
315
270
225
180
UrsaMajorI(UMaI)
UrsaMinor(UMi)
UrsaMajorII(UMaII)
Draco(Dra)
Hercules(Her)
NGC5694
NGC4590
Sextans(Sex)
Leo I
Leo II
Leo V
CVn I
NGC5824
NGC5272
NGC5466
NGC5053
Coma Berenices(CBe)
Boötes I(Boo)
M31
M33
Segue2
Pal13
Pisces I
NGC1904

The Results

Sagittarius stream: an example for structure finding

globular cluster or dwarf galaxy $\rightarrow$ torn apart and stretched out along its orbit by tidal forces $\rightarrow$ stellar stream





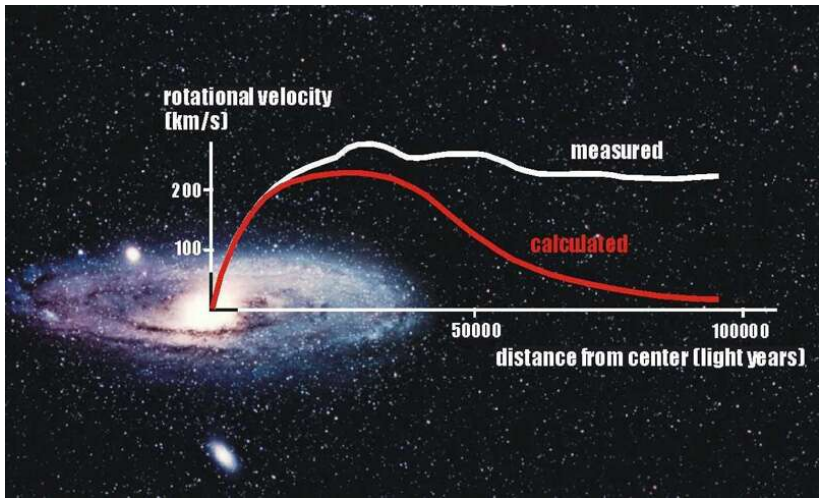
artistic image,

www.spitzer.caltech.edu

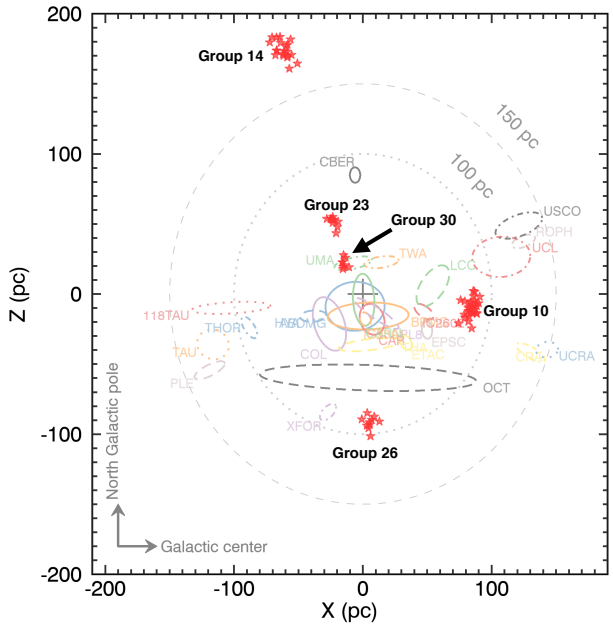
The Follow-Up Survey

Milky Way dynamics: get 3D velocities

Dark Matter



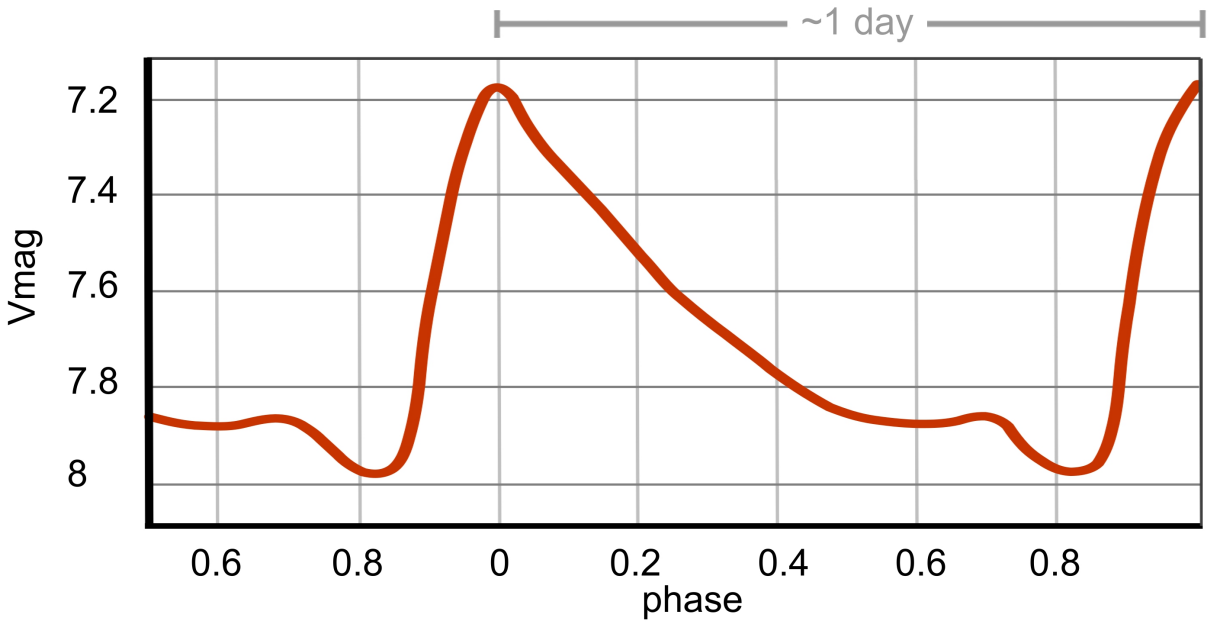
comoving groups and clusters in the Milky Way

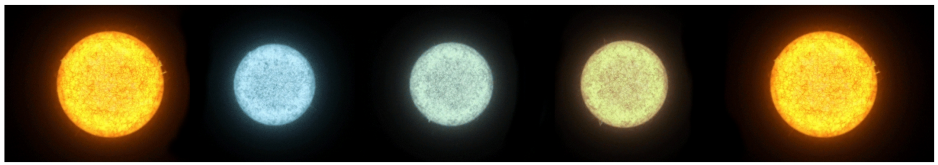


Faherty+2018

The Follow-Up Survey

crucial for RR Lyrae (and pulsators in general): **timing**

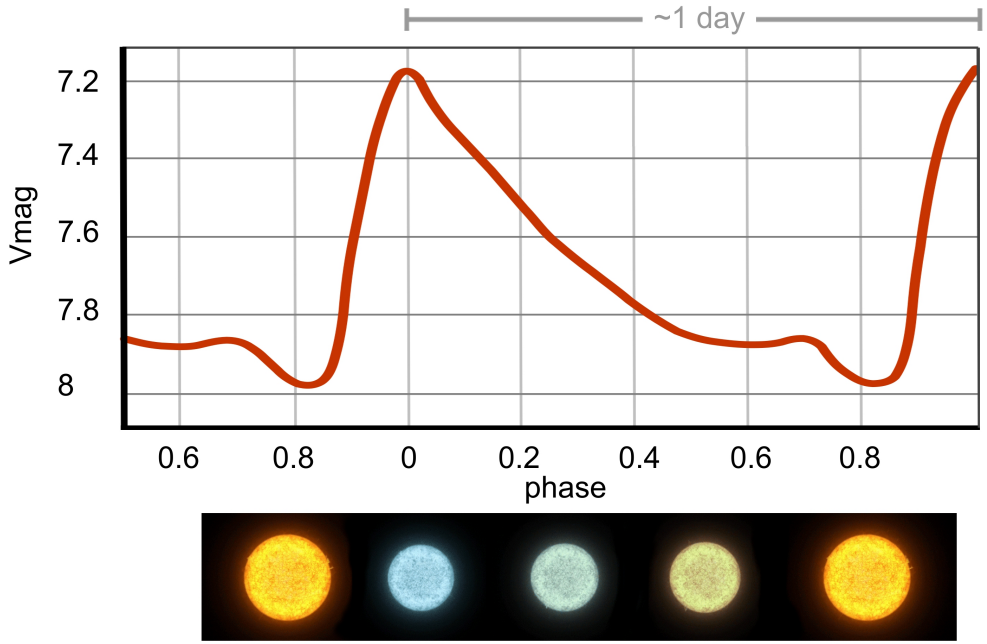




lack of hydrostatic equilibrium drives pulsation and thus periodic change in brightness

The Follow-Up Survey

crucial for RR Lyrae (and pulsators in general): **timing**



$$v_{\mathrm{obs}} = v_{\mathrm{systemic}} + v_{\mathrm{photospheric}}$$

from pulsation models:

observe at $\phi = 0.37$ where $v_{\mathrm{photospheric}} \sim 0$

Data Challenges

Data challenges: PLAsTiCC /ELAsTiCC

Photometric LSST Astronomical Time-series Classification Challenge
and its extension
The Extended LSST Astronomical Time-Series Classification Challenge

ELAsTiCC uses simulated alerts, delivered to the alert brokers, to mimic the future rate, volume, and complexity of the LSST prompt data products. Realistic contextual information is incorporated into synthetic alerts.

The Challenge:

- Types are unbalanced
- Small number in the training set
- The training set is not representative of the test data
- Seasonal gaps
- Non-uniform cadence