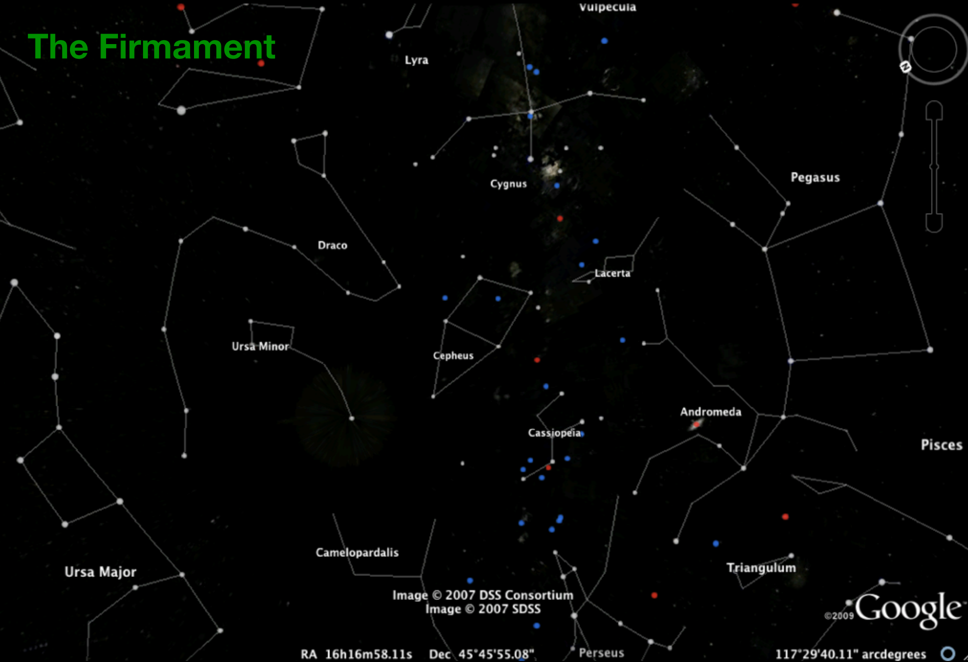
Welcome!

AAS237 Workshop:

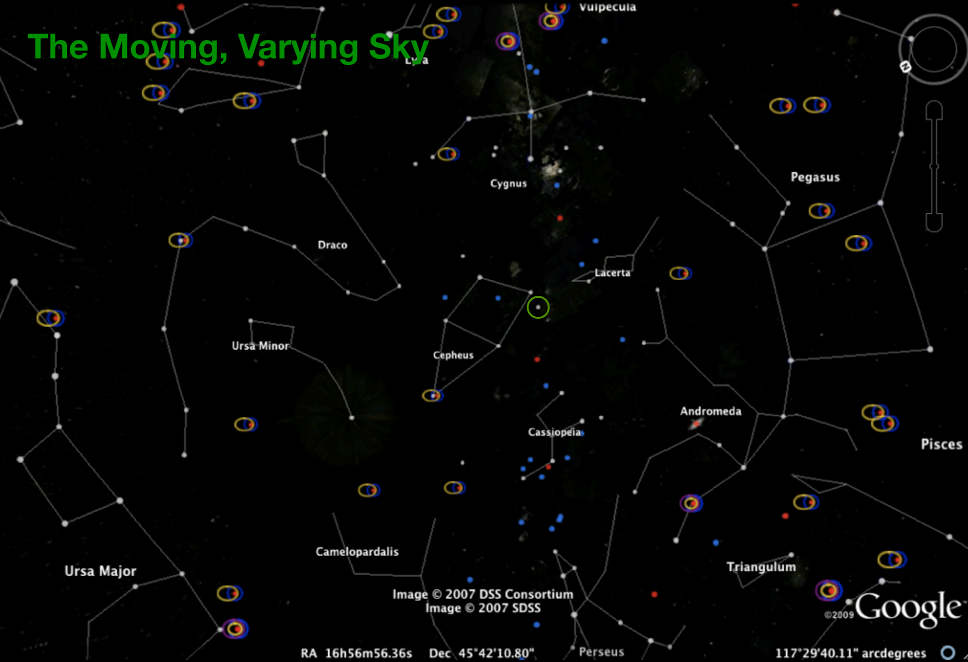
Exploring and modeling astrononomical time series data

Session material public GitHub repo:
`https://github.com/tloredo/AAS237-TimeSeries`

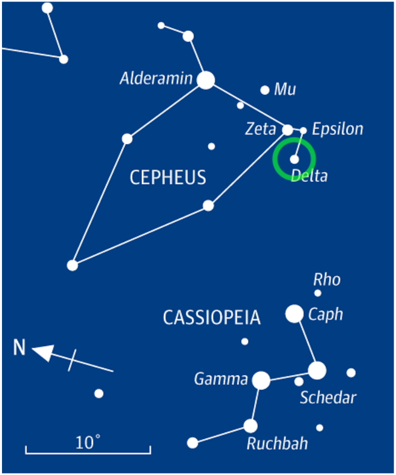
The Firmament



The Moving, Varying Sky



Delta Cephei — Variability!



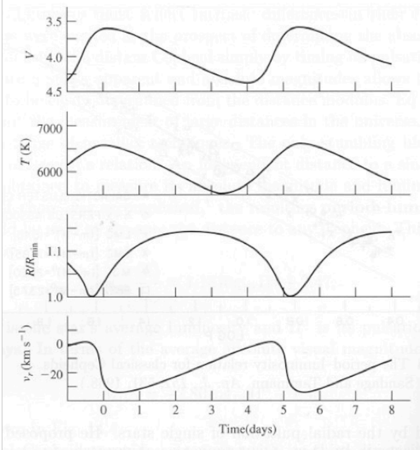
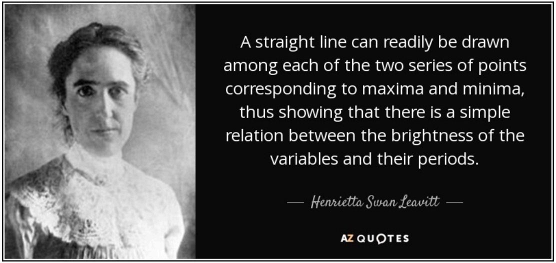


Figure 14.5 Observed pulsation properties of δ Cephei.

Discovered in 1700s; 5.4 d period, 0.9 mag ampl
(Mira & Algol periodic variables discovered in 1600s; "Mira" = "wonderful," "astonishing")

Leavitt law for Cepheids

An early time-domain astronomy triumph



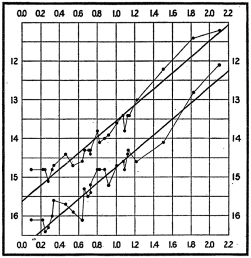
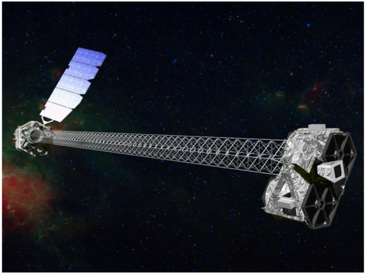


Fig. 2.

Context: Growing role of time-domain astronomy

- Automated, large-scale time-domain surveys: CRTS, PTF, ZTF, Pan-STARRS, Rubin Observatory LSST...
- Space-based observatories with large time-domain datasets: CGRO/BATSE, RXTE, Fermi, COROT, Kepler, TESS, NuStar, NICER, Gaia...



Context: Time series software for astronomy

Packages with generality/breadth/depth (recent/maintained):

- VARTOOLS: Command-line light curve analysis (C)
- SITAR: S-lang/ISIS Timing Analysis Routines
- CULSP: Lomb-Scargle periodograms on GPUs
- LightcurveMC: LC simulation, testing tools in C++, R
- BGLS: Bayesian generalized Lomb-Scargle periodogram
- FATS: Feature analysis for time series
- gatspy: General tools for astro time series (AstroML)
- Spectra: Power spectra for unequally-spaced data
- agatha: Period finding in correlated noise (R)
- Gaussian process packages: George, celerite
- Mission/project-specific tools: *Fermi* tools, *Kepler*/*TESS* lightkurve, Starlink...
- Julia: JuliaAstro/LombScargle, cerite, CARMA.jl...
- carma_pack: Bayesian CARMA modeling via MCMC (C++, Python)
- Stingray: Next-generation spectral-timing software (Python)
- TSE Project: Python and MATLAB packages by Scargle & Loredo

Documentation, VCS, appealing API are essential for buy-in

R packages (mainly by statisticians)

"Best of" list c/o Eric Feigelson

Base-R functions

- acf-pacf-ccf: correlation functions with significance levels
- arima-prewhiten: autoregressive modeling
- Box.test: test for autocorrelation
- density-spline-loess: kernel & local polynomial interpolations
- fft & convolve : Fast Fourier Transform, convolutions
- fitdistr: maximum likelihood fitting of statistical distributions
- plot: display time series
- runmed-smooth-supsmu: running median-like smoothers
- spec.pgram: Fourier periodogram with tapering & smoothing

CRAN packages

- bspec: Bayesian autocorrelation & spectral analysis
- cobs: cobs quantile spline interpolation
- changepoint-Rseg-segmented-strucchange: changepoint detection & segmented regression
- dlm: Bayesian dynamic modeling
- dtw-dtwclust: dynamic time warping & clustering
- dyn-dyn.lm: regression for irregular time series
- forecast: ARIMA modeling with model selection & 1/f-noise
- imputeTS: na.Kalman ARIMA interpolation
- its, xts & zoo: infrastructure for irregular time series
- locfit: locfit local interpolation with bootstrap, weighting & censoring
- *lomb: Lomb-Scargle periodogram*

CRAN packages

- meboot: bootstrap for nonstationary time series
- MSBVar: dynamic multivariate autoregressive modeling
- msl.trend: linear, spline, SSA interpolation of gaps for irregular time series
- mvtsplot: visualization of multivariate time series
- nortest: ad.test test for normality
- robfilter: robust treatments of outliers
- RobPer: RobPer robust periodograms: PDM, LSP, etc
- sde: stochastic differential equations
- tseries-TSA: extensive time series analysis & testing
- TSDist-TSClust: distance and clustering ensembles of times series
- wavelets-wavethresh-wmtsa-adlift: wavelet transform, denoising & analysis
- WeightPortTest: tests for autocorrelation with heteroscedastic weights
- *spritzer: Bayesian period detection in red noise (Simon Vaughan; non-CRAN)*

This workshop

- Time series intro; understanding periodograms (Tom Loredo)
- Methods for analyzing irregularly sampled time series and point data (Jeff Scargle)
- Tools for spectro-temporal analysis of X ray time series data (Daniela Huppenkothen)

Astrophysical time series: Basic phenomenology and terminology

Tom Loredo
Cornell Center for Astrophysics and Planetary Science
http://www.astro.cornell.edu/staff/loredo/bayes/

AAS237 — 8 Jan 2021

Agenda

1. **Types of astrophysical variability**

2. **Time series data & signal types**

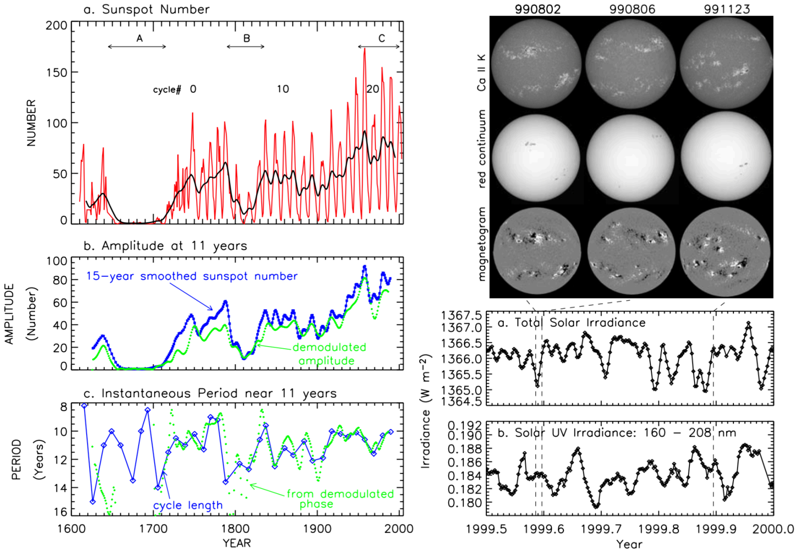
3. **Statistical models: Deterministic vs. stochastic signals**

Agenda

1. **Types of astrophysical variability**
2. **Time series data & signal types**
3. **Statistical models: Deterministic vs. stochastic signals**

Solar variability

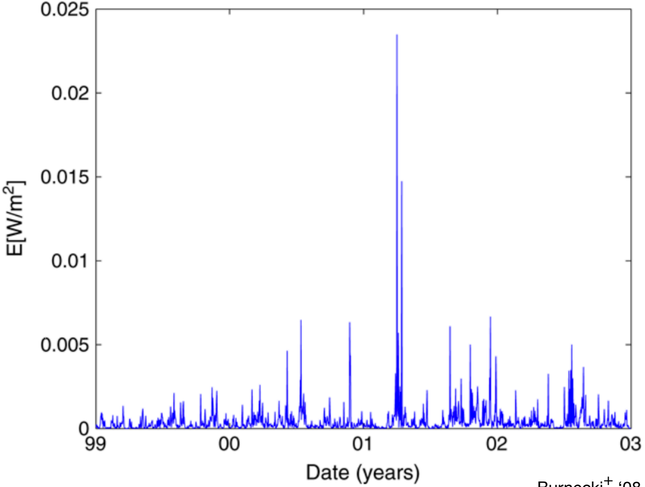
Oscillatory and stochastic variability



Frohlich & Lean '04

Transients: Solar flares

Solar flare X-ray energies from GOES (1-day aggregate)



Burnecki+ '08

Secular: Tree rings & ice cores over ∼1000 y

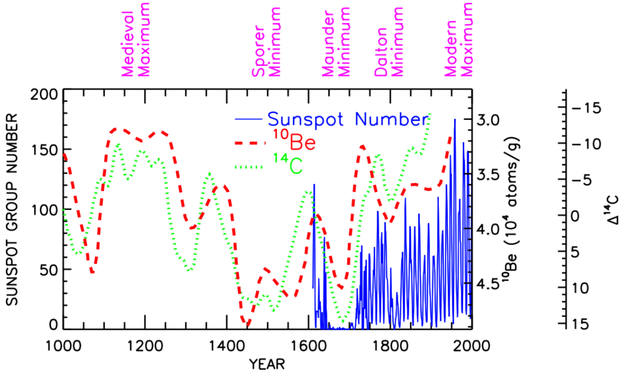
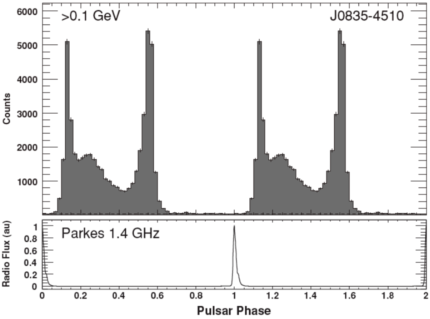


Fig. 29. Shown are the records of cosmogenic isotope fluctuations in tree-rings and ice-cores associated with solar activity during the past millennium. The long-term trends in the cosmogenic isotopes track the envelope of sunspot number amplitudes

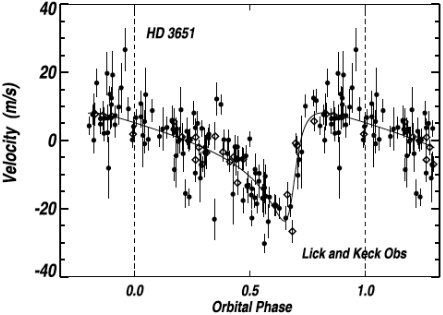
Frohlich & Lean '04

Periodic variability: Pulsars, exoplanets

Vela pulsar folded gamma, radio data



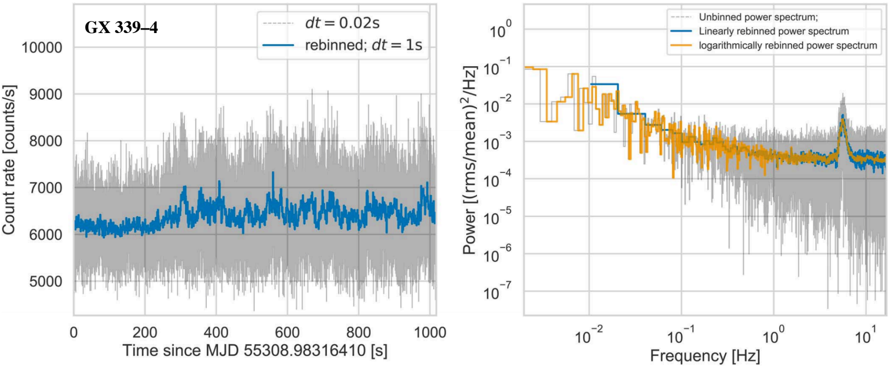
Exoplanet folded RV data



Also periodic variable stars: Cepheid, RR Lyrae, Mira. . .

Black hole X-ray binary variability

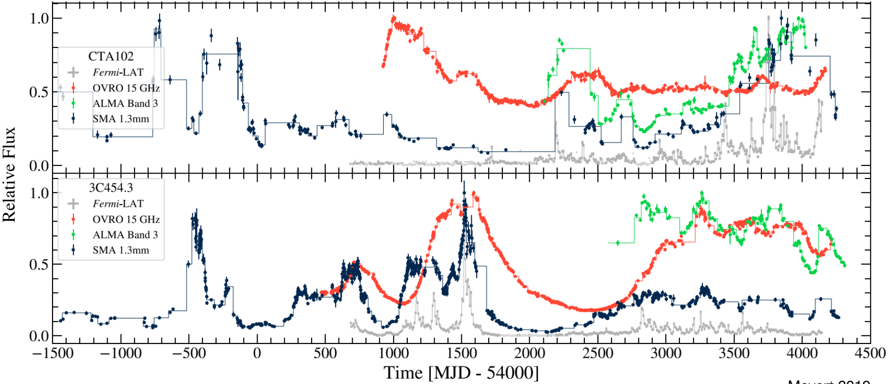
Quasi-periodic oscillation (QPO) and colored noise



Huppenkothen+ 2019

Quasar multi-wavelength variability

Dependence and lags

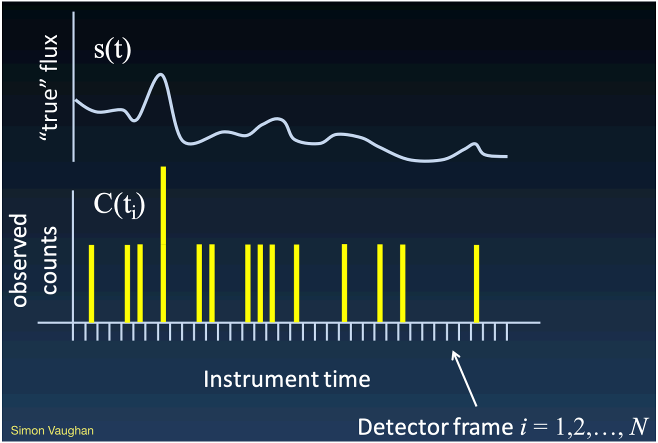


Meyer[+] 2019

Agenda

1. Types of astrophysical variability
2. Time series data & signal types
3. Statistical models: Deterministic vs. stochastic signals

Signal vs. data



It's tempting to try to massage the data to make it look like the signal, e.g., via binned rate estimation or smoothing.

Avoid this temptation! This type of data reduction typically reduces the information content in the data.

Time series data modes

- **Magnitude/amplitude data:** Real-valued amount/strength/size of something
  - *Pointwise:* Amount *at a point in time* (or perhaps locally averaged)
    Ex: radial velocity (RV), v_i at time t_i; radio polarization
  - *Cumulative:* Amount *accumulated in an interval*
    Ex: bolometer data, energy measured in Δt_i
- **Event/point data/count data:** Integer-valued count of discrete events or objects *in an interval*
  Ex: Time-tagged even (TTE) photon data, binned photon data, time-to-spill data
- **Marked event data:** Event data, in a setting where each event has a strength/size associated with it; can be viewed as point data in > 1D
  Ex: flare or burst energy time series, photon time + energy

Data mode for a derived data product may differ from that of raw data

Signal representations

What type of mathematical object should represent the signal?

- **Intensive signals:** Represent a quantity meaningful only at *a point in time*, not accumulating in time: use a **continuous function** (a mapping from a point in time to an amplitude)
  Ex: velocity, $v(t)$; temperature $T(t)$
- **Extensive signals:** Represent a quantity that accumulates over an *interval*: use a **measure** (a mapping from an interval to an amount in the interval)
  Ex: flux (photon counts or energy per unit time), flux density (photon counts per unit time and energy or wavelength)

A measure $\mu(\Delta t)$ is typically specified in terms of a (nonegative) *rate or intensity* **function**, $\mu(\Delta t) = \int_{\Delta t} dt\, r(t)$

Key distinction is *transformation under a time scale change*: Would the signal level change if you change the units of time labeling the measurements?

Agenda

1. Types of astrophysical variability

2. Time series data & signal types

3. **Statistical models: Deterministic vs. stochastic signals**

Deterministic vs. stochastic signal models

Statistical models have random/uncertain elements, described by probability distributions, but the randomness may be decoupled from the signal

Consider an intensive model for pointwise amplitude data with additive Gaussian noise:

$$y_i \equiv f(t_i) + \epsilon_i; \qquad \epsilon_i \sim \mathrm{Norm}(0, \sigma)$$

Also, an extensive model for Poisson-distributed binned count data:

$$n_i \equiv \mu(\delta t_i); \qquad n_i \sim \mathrm{Pois}\left(\mu(\delta t_i)\right)$$

with expectation values

$$\mu(\delta t_i) = \int_{\delta t_i} dt\, r(t)$$

Deterministic (parametric) models

Specify $f(t;\theta)$ or $r(t;\theta)$, fully specified functions of time, once we specify values for parameters θ

Ex: Sinusoid, $f(t;A,\omega,\phi) = A\cos(\omega t + \phi)$; $\theta = (A,\omega,\phi)$

Stochastic models (non- and semi-parametric)

Specify a *stochastic process*, $\mathrm{SP}(\psi)$, with (hyper)parameters ψ, and

$$f(t) \sim \mathrm{SP}_f(\psi), \quad \text{a random function} \tag{1}$$

or one of

$$r(t) \sim \mathrm{SP}_r(\psi), \quad \text{a random function} \tag{2}$$

$$\mu(\delta t) \sim \mathrm{SP}_\mu(\psi), \quad \text{a random measure} \tag{3}$$

Nonparametric: There are no hyperparameters, or their values are uninteresting; we're interested in estimating $f(t)$ or $r(t)$

Semiparametric: The detailed form of the signal isn't of direct interest; one or more of the hyperparameters are (e.g., frequency for a periodic signal of complex shape)

Stochastic processes

Probabilistic models for a quantity evolving unpredictably in time or space

Stochastic process:

- A collection of random variables (RVs)/uncertain quantities (UQs)
- With the variables labeled by values of an index variable (e.g., time)
- With each variable taking values *in the same space* (e.g., with the same units)

Discrete SP: Index is an integer

Continuous SP: Index is real-valued

Typically the collection is of arbitrary size, e.g., arbitrarily extensible in time

A SP is a special kind of multivariate distribution—a joint distribution for an arbitrary number of similar quantities. Many have *dependence/correlation* across the index space.

Types based on index space

- *Temporal process:* Time (univariate and directional)
- *Spatial process:* Any non-temporal continuous index/indices, e.g., Euclidean space, energy
- *Spatio-temporal:* Time and one or more spatial dimensions, e.g., physical space + time, or time + energy or wavelength (dynamic spectrum/spectral timing)
- *Random field:* A spatial SP on R^2 or R^3

Point processes (PPs)

- *Basic:* A realization spreads points over a continuous index space, e.g., photon arrivals in time
- *Marked PP:* A realization spreads points, and assigns each one a random mark; e.g., photon time & energy; burst time and fluence
- *Compound:* A marked PP where one sums up the total mark amount in intervals; e.g., amount of rain over an area (depending on drop rate & size dist'n)

Kolmogorov extension theorem

A SP may be fully specified by giving a rule telling you how to write the joint distribution for the values at any fixed *finite* number of index points or intervals (as long as the rule satisfies a trivial consistency condition)

This is enough *even for continuous SPs*

SPs are often named for the univariate distribution that applies to a single index point or interval

Important SPs

- *Bernoulli process*: Binary outcomes in discrete time, e.g., sequence of (biased) coin flips
- *Random walk*: Continuous outcomes in discrete time (drunkard's walk)
- *Binomial PP*: N points spread in continuous time with $t \sim p(t)$, independently, for a probability density function (PDF) $p(t)$
- *Poisson PP*: Like binomial, but with $N \sim \mathrm{Pois}$, and $t \sim$ proportional to rate
- *Markov process*: "No memory"/minimal dependence processes, continuous and discrete
- *Gaussian process (GP)*: Real-valued quantity as a function of time/space; *random functions*
- *Levy process*: Positive or signed functions with with independent, stationary increments in continuous time/space, including instantaneous jumps ("continuous random walk"); good for *random functions or measures*
- *Dirichlet process*: Positive functions with unit normalization; good for *random probability measures*

Multiple meanings

Two bits of potentially confusing terminology…

- **Sample:**
  - Take a measurement at a (deterministic) time, as in "sample rate" or "regularly sampled" ("measure", "probe"?)
  - Take a random sample of the entire time series
  - Take a random sample from a stochastic process signal model—a *sample path*
- **Asymptotics:** How does an algorithm behave as you accumulate more and more data?
  - *Infill asymptotics:* Get more data in an observing interval of fixed duration (raise the sampling rate, increase the collecting area)
  - *Extended domain asymptotics:* Get more data by observing for more time; there can be unusual behavior (e.g., period uncertainty shrinks *faster* than $1/\sqrt{N}$)