

ASTRODATA2020S
PASADENA, 2018 DECEMBER 6

What to do with a billion time series?

Matthew J. Graham
Center for Data-Driven Discovery/ZTF, Caltech
mjg@caltech.edu



The first astronomical time series

1610. Dec. 8. Luna.
Dicitur. 8. Luna.
The altitude of the
Sunne being 7 or 8
degrees. It being
a host of a mile. I saw
the same in this manner.
Instrument. $\frac{10}{1}$. B.
I saw it twice or thrice. once
with the right eye & other times
with the left. In the space of a minute time, after the Sunne was clear.

1610. Luna.
January. 19. In a notable mist. I observed diligently at
sundry times when it was fit. I saw nothing but the cleare
Sunne both with right and left eye.

1611. Dec. 1. Luna. 5. 10. 0.
per eclipsim Solare.
I saw these black spots in face
when as it was expressed as man
as it could be. observed $\frac{10}{1}$
so in time. With the left eye
the same at sundry times all these
I saw it once for halfe an hour
at sundry times and all the morning before
it was with.

The greatest was both reflex, most orientall
& it appeared single about 2'. the other
two, were more or less beyond. & of 1' magnitude.
was seen about

509

Thomas Harriott: Dec 1610

1610
Sex^{mo} Principe.

Galileo Galilei Humilis. Servo della Ser. V. inuigilando
do assiduamente, et co ogni spirito di potere no solo satisfare
alcuno che tiene della lettura di Mathematicis nelle Scu-
ole di Padoua,

Si uolere dauere determinato di presentare al Sex^{mo} Principe
l'occhio et a p essere di giouamento inestimabile p ogni
negozio et ombra marittima o terrestre stimo di tenere que-
sto nuovo artificio nel maggior segreto et solame a disposizione
di V. Ser. L'occhio conato dalle piu re d'ite speculazioni di
prospettua ha il uantaggio di scoprire i legni et vele dell' inimico
p due hore et piu di tempo prima che egli sia sopra noi et distinguendo
il numero et la qualita de i vasselli, giudicare le sue forze
pallestirli alla caccia al combattimento o alla fuga, o pure auere
nella campagna aperta uedere et particularme distinguere ogni suo
moto et preparatione.

Apr. 7. di Gennaio
Gioue si uide con
Adi 8 con
Adi 12. si uide in tale costituzione
Il 13. si uide due uicini: a Gioue 4 stelle
Adi 14 è un glo
Il 15. si uide la pros^a a 74 ora la mig^a la 4^a ora di =
stante dalla 3^a il doppio laira
Lo spazio delle 3 uide si uide no era
maggiore del diametro di 74 et e =
vano in linea retta.

74 * uici: 10. 11.
Adi 8 con
Adi 12. si uide in tale costituzione
Il 13. si uide due uicini: a Gioue 4 stelle
Adi 14 è un glo
Il 15. si uide la pros^a a 74 ora la mig^a la 4^a ora di =
stante dalla 3^a il doppio laira
Lo spazio delle 3 uide si uide no era
maggiore del diametro di 74 et e =
vano in linea retta.

74 long. 71. 38 lat. 1. 19

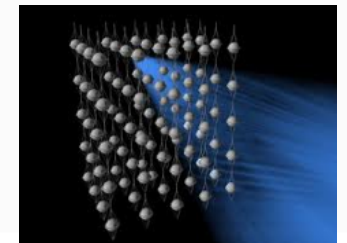
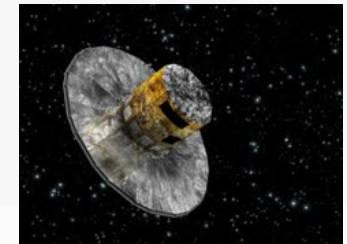
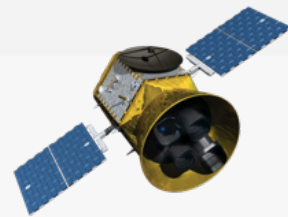
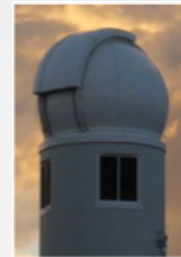
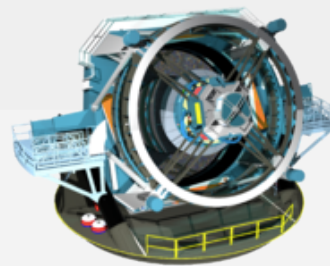
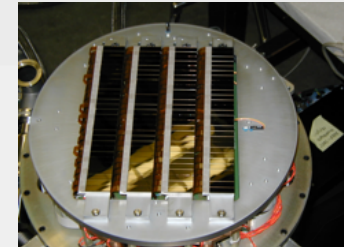
Image credit: University of Michigan Special Collections Library



The burgeoning time domain

- Palomar-Quest Synoptic Sky Survey
- SDSS (Stripe 82)
- Catalina Real-time Transient Survey
- Palomar Transient Factory
- Zwicky Transient Factory
- Pan-STARRs
- SkyMapper
- ASKAP
- ThunderKat (MeerKAT)
- KEPLER
- GAIA
- LIGO
- IceCUBE
- LOFAR
- LSST
- SKA
- TESS
- ASAS-SN
- MASTER
- DES
- ATLAS
- BlackGEM

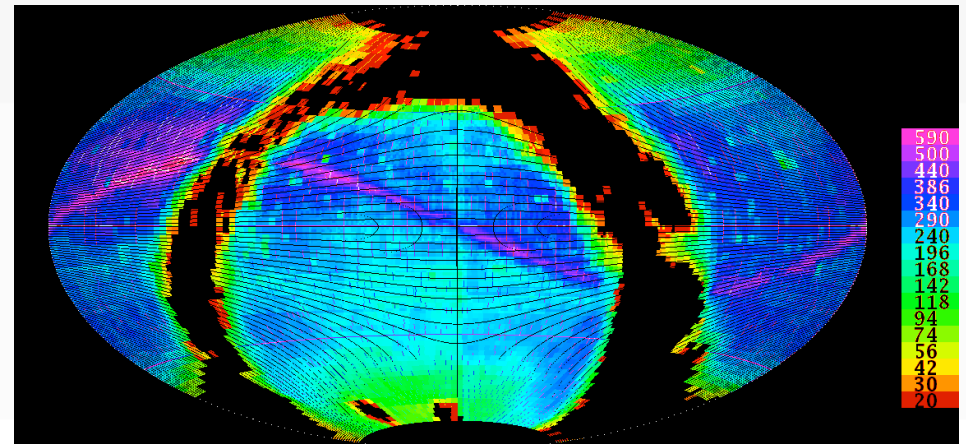
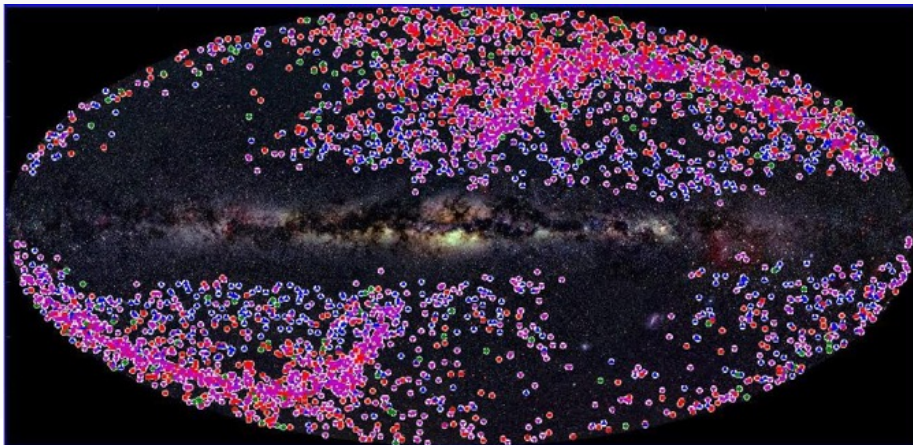
...





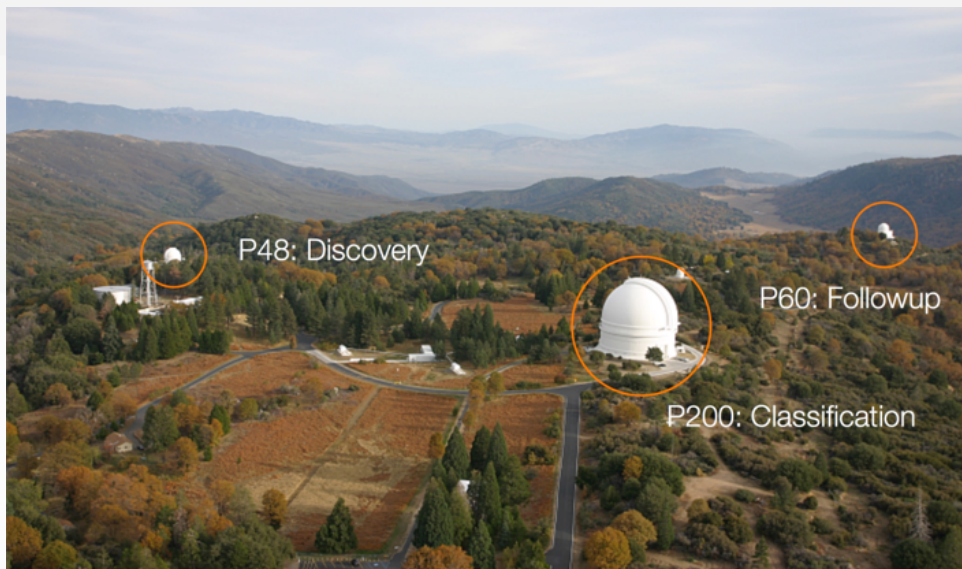
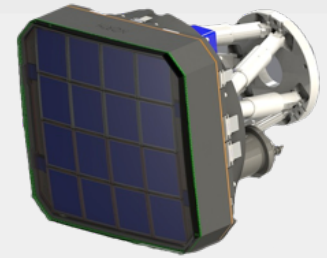
Catalina Real-time Transient Survey (2005-)



- Collaborative survey with Catalina Sky Survey (LPL, UA)
- Unfiltered observations 21 nights/lunation covering up to 2000 deg²/night
- Covers 33000 sq. deg. ($0 < RA < 360$, $-75 < Dec < 70$).
- Calibrated photometry for 500 million objects (> 100 billion data points)
- Depth V = 19 to 21.5
- 100 – 600 observations in most regions (median ~ 320)
- Temporal baselines of 10 min to ~12 years
- More published SNe and CVs than any other survey (public instantly)
- Open data policy (<http://catalinadata.org>)
- ~3% LSST (2^{-5})



Zwicky Transient Facility (2017-)

- New camera on Palomar Oschin 48" with 47 deg² field of view
- 3750 deg² / hr to 20.5-21 mag (1.4 TB / night)
- Full northern sky every three nights in g and r
- Galactic Plane every night in g and r
- Over 3 years: 3 PB, 750 billion detections, ~1000 detections / source
- First megaevent survey: 10⁶ alerts per night (public since June 2018)



		
No. of sources	1 billion	37 billion
No. of detections	1 trillion	37 trillion
Annual visits per source	1000 (2+1 filters)	100 (6 filters)
No. of pixels	600 million (1320 cm ² CCDs)	3.2 billion (3200 cm ² CCDs)
Field of view	47 deg ²	9 deg ²
Hourly survey rate	3750 deg ²	1000 deg ²
Nightly alert rate	1 million	10 million
Nightly data rate	1.4 TB	15 TB

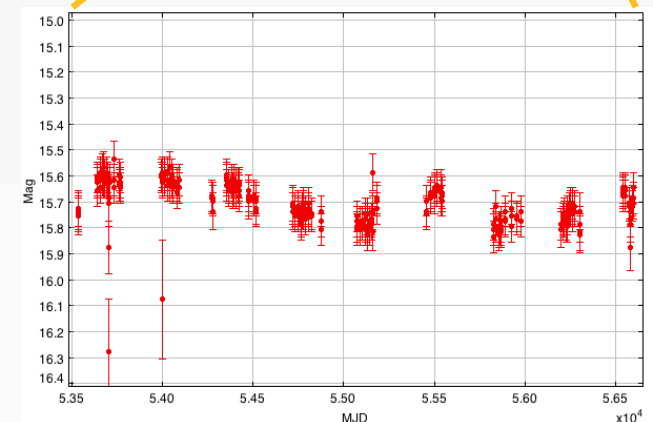
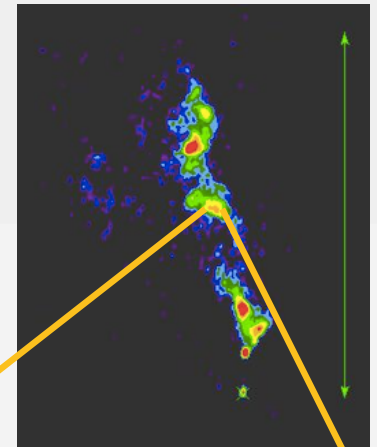


What do we ask of time series data?

- Population behaviors
 - Characterize, categorize, classify
- Outliers
 - Extreme sources
- Physical models
 - Predictions

Case study: quasar variability

- First quasar identified 3C 48 – most striking feature was that the optical radiation varied
- Physical origin of photometric variability in optical/UV is unclear:
 - Instabilities in the accretion disk
 - Supernovae
 - Microlensing
 - Stellar collisions
 - Thermal fluctuations from magnetic turbulence
- Many studies based on small sample size or (very) sparse time sampling
- Complementary to SED studies

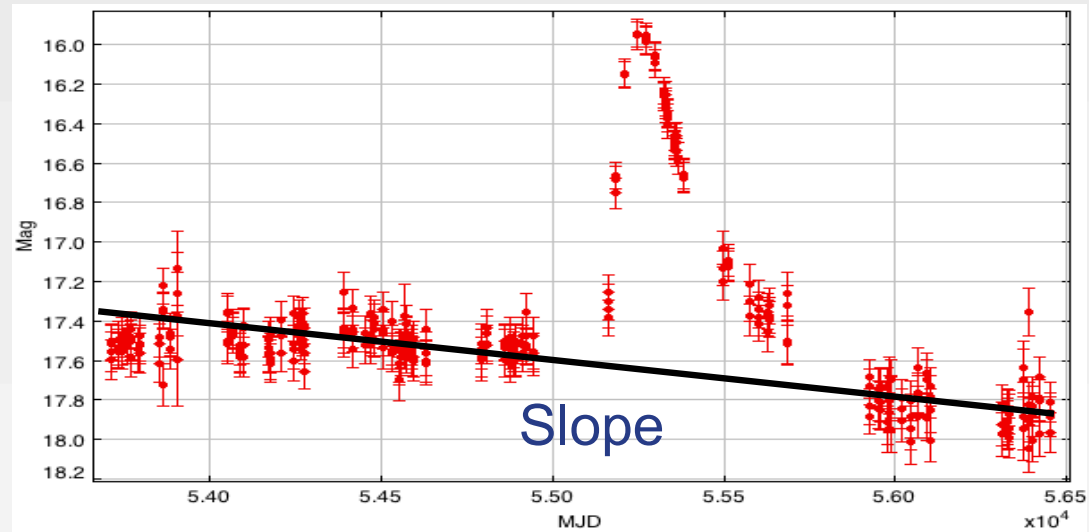




Characterization – extracting data features

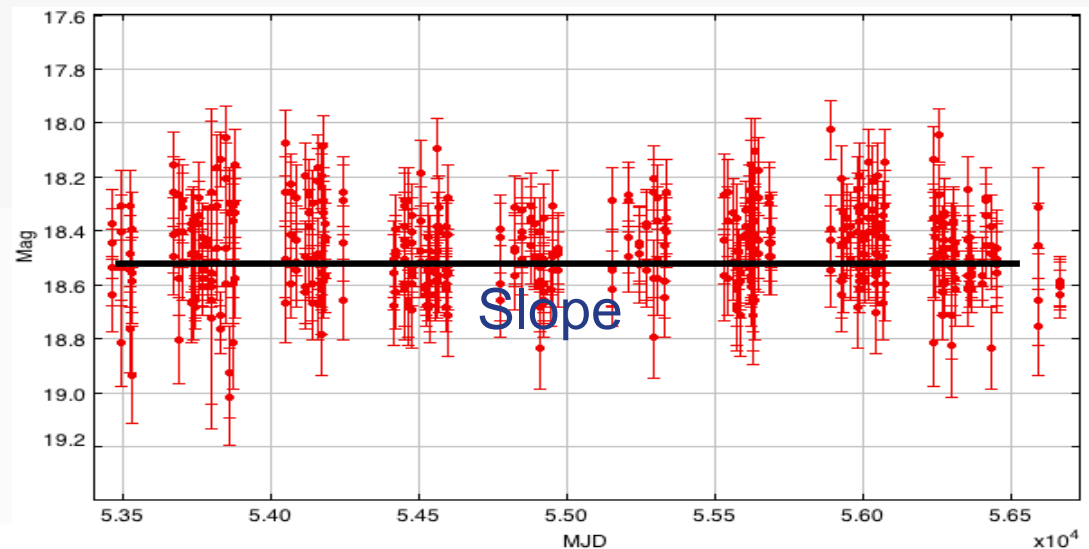
$$\sum_{i=1}^n A_i \sin(\omega t + \phi_i)$$

Fourier



$$\sum_{i=1}^n A_i \sin(\omega t + \phi_i)$$

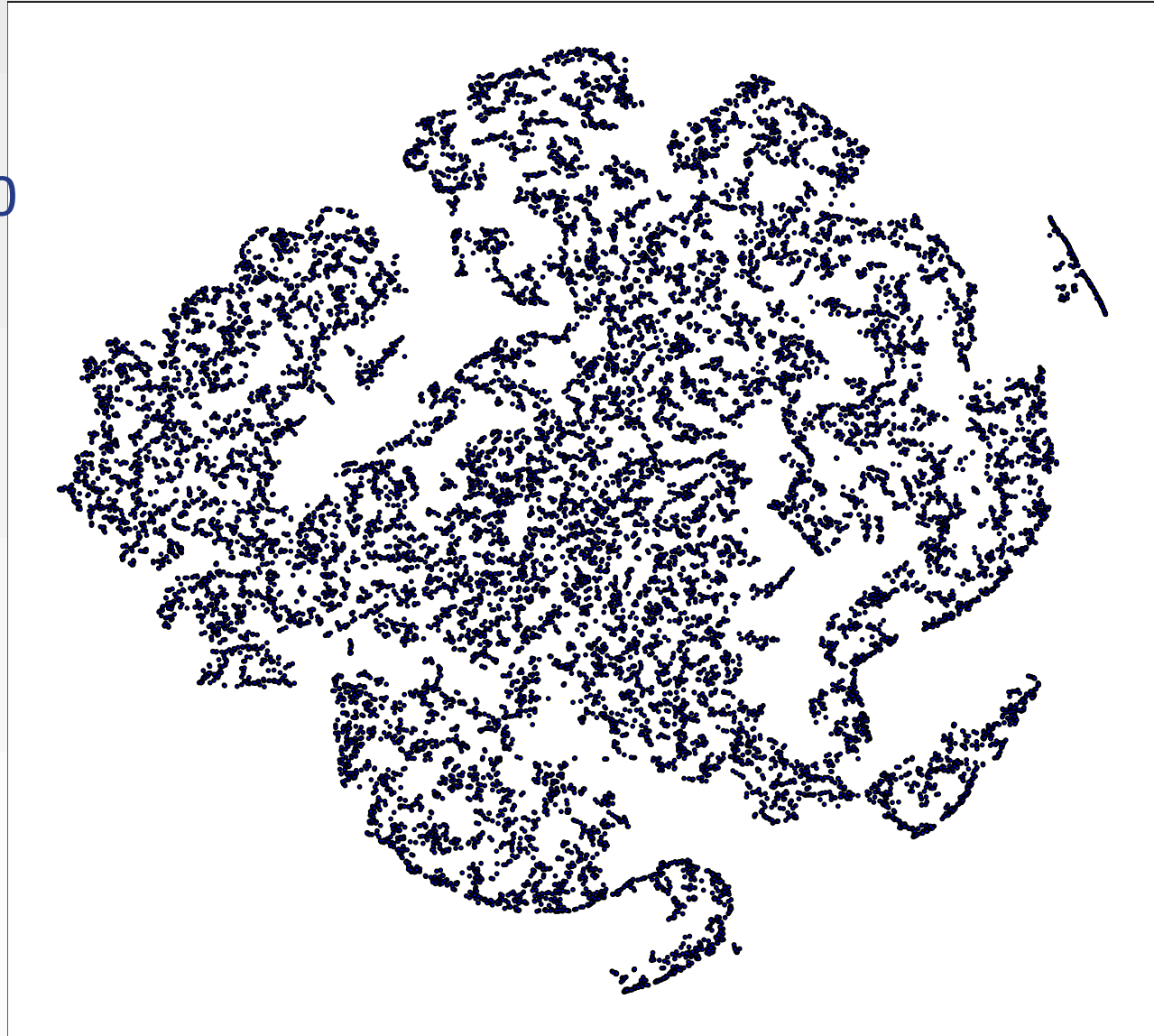
Fourier





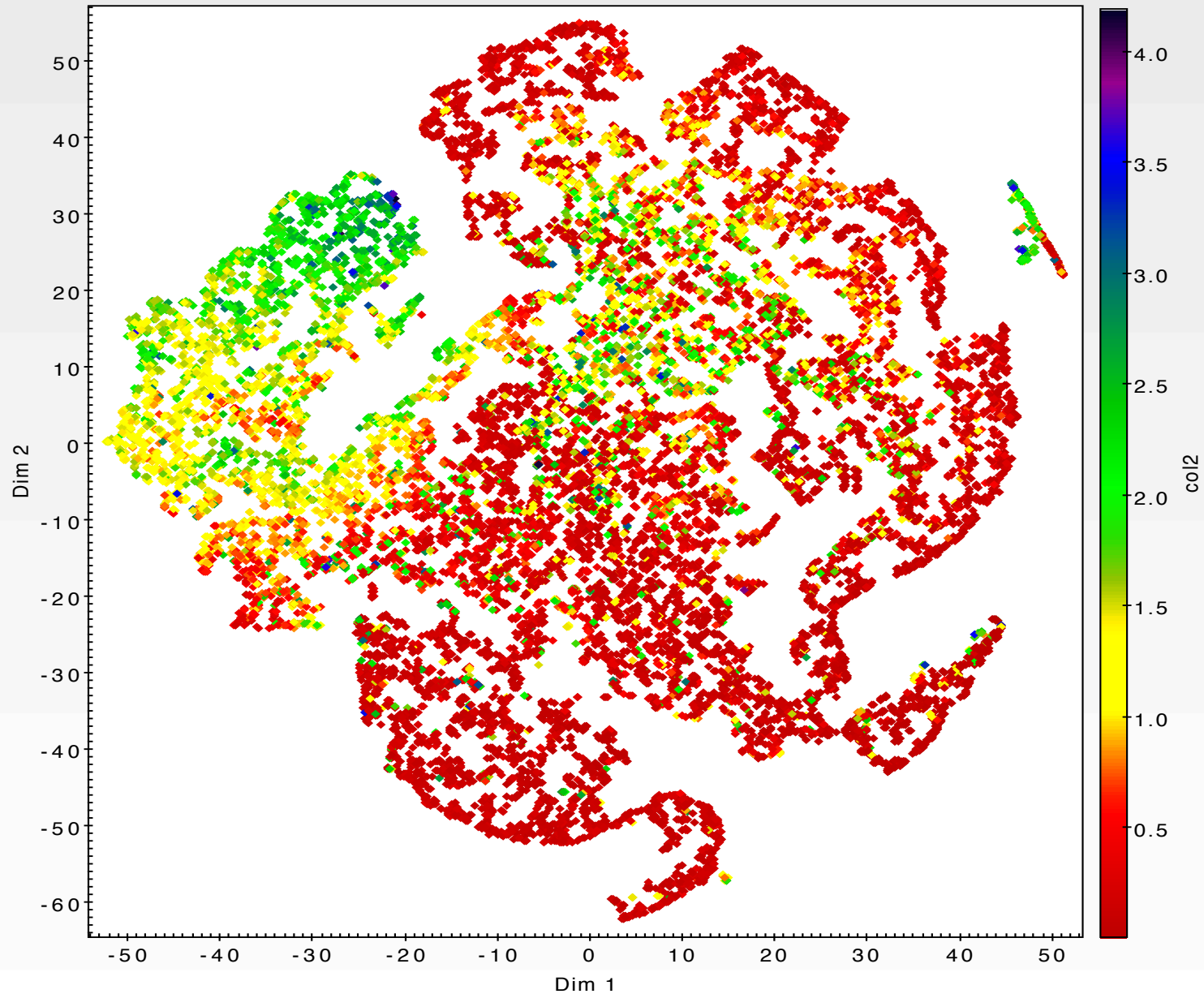
Categorization – clustering the data

A 2-D topology preserving representation of a 6-D parameter space for 20000 characterized time series



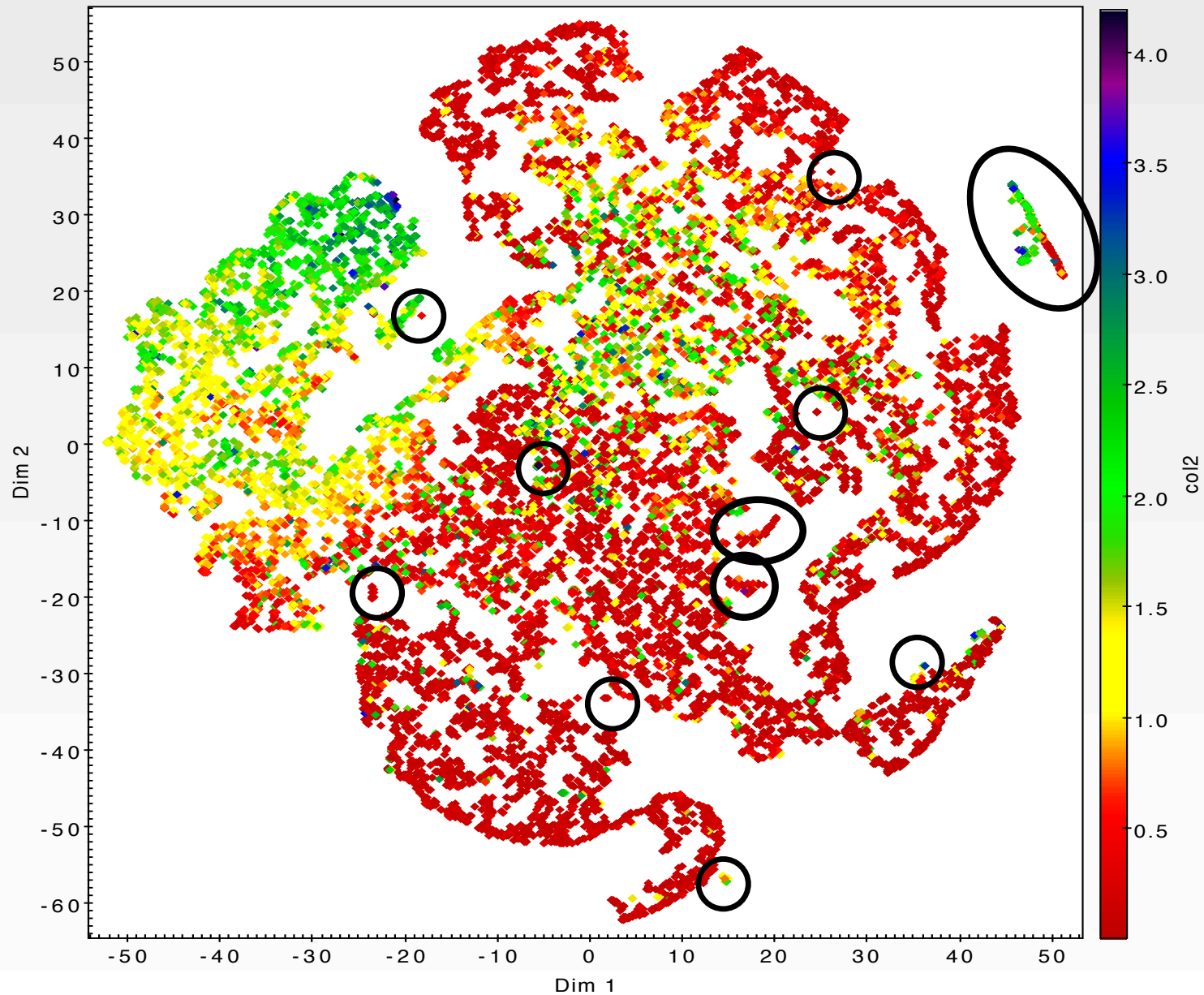


Classification – identifying the clusters





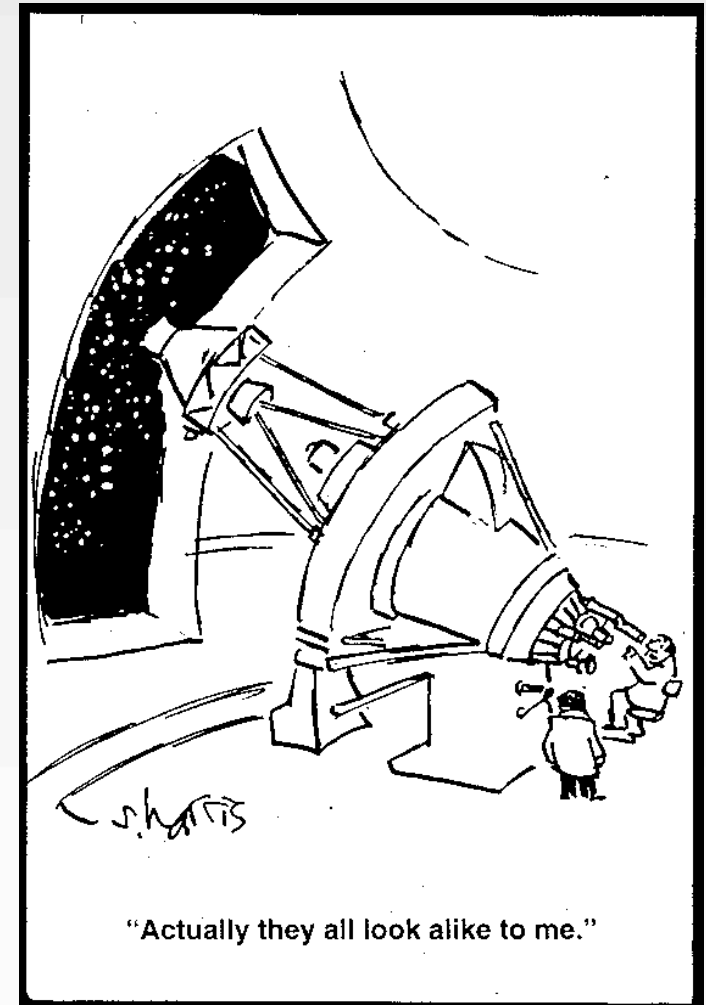
Outliers – the things that do not fit





Common statistical features

- Timescales:
 - Lomb-Scargle
- Variability:
 - von Neumann variability (phase-folded)
 - Stetson K index
- Morphology:
 - Skewness
 - Kurtosis
 - IQR
 - Cumulative sum index (phase-folded)
 - Ratio of magnitudes brighter/fainter than mean
- Trends:
 - Slope percentiles (phase-folded)
- Model:
 - Fourier amplitude ratios
 - Fourier phase differences
 - Fourier amplitude
 - Shapiro-Wilk normality test





Unstated assumptions

- **Homoskedasticity**

- All errors are drawn from the same process (same variance model for all data points)

- **Non-IID**

- Data is sequential

- **Stationarity**

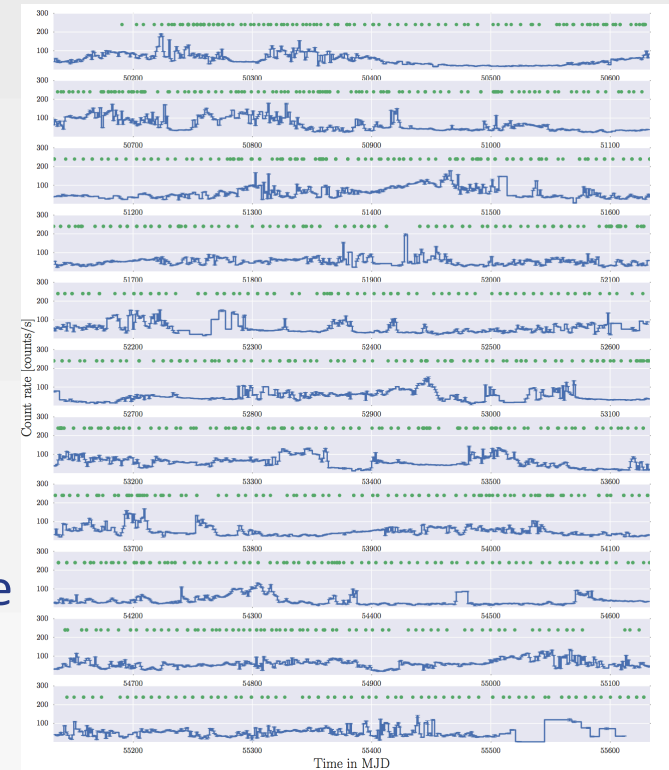
- The generating distribution is time independent
- GRS 1915+215 has ~20 variability states
- GARCH models: variance is a stochastic function of time
- Nonstationary time series do not have to stationary in any limit

- **Ergodicity**

- The time average for one sequence is the same as the ensemble average

$$\hat{f}(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x).$$

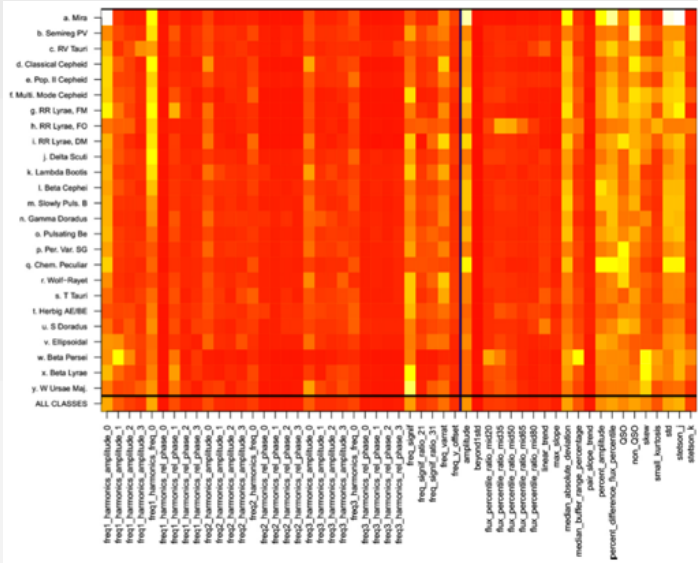
- Observations sufficiently far apart in time are uncorrelated and new observations give extra information



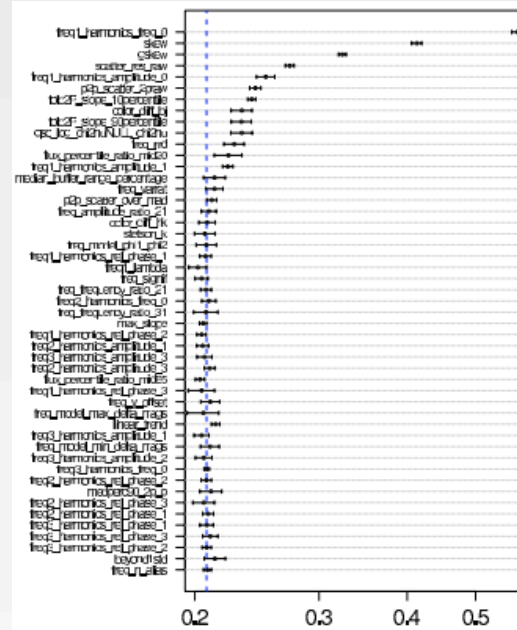
(Huppenkothen et al. 2016)



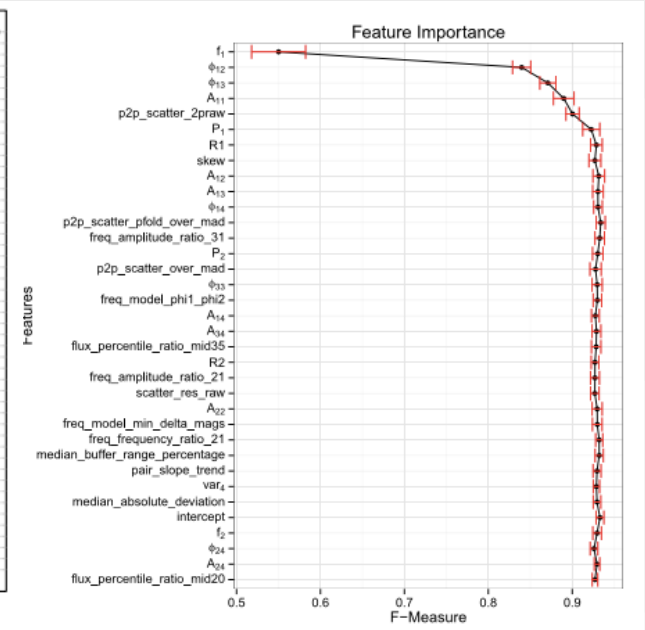
Not all features are equal



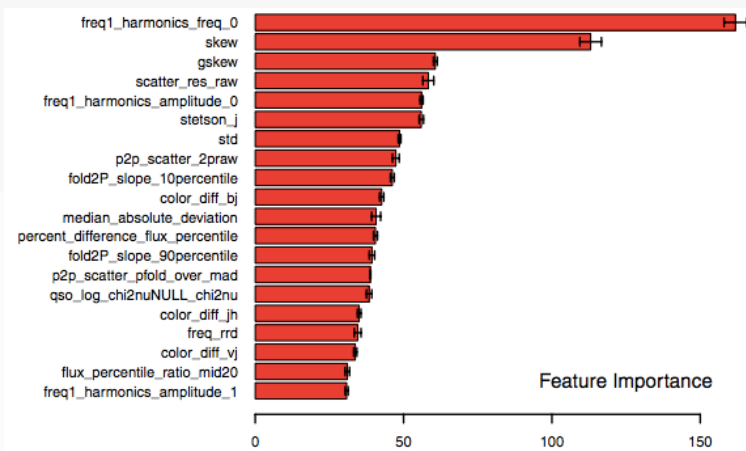
Richards et al. 2011



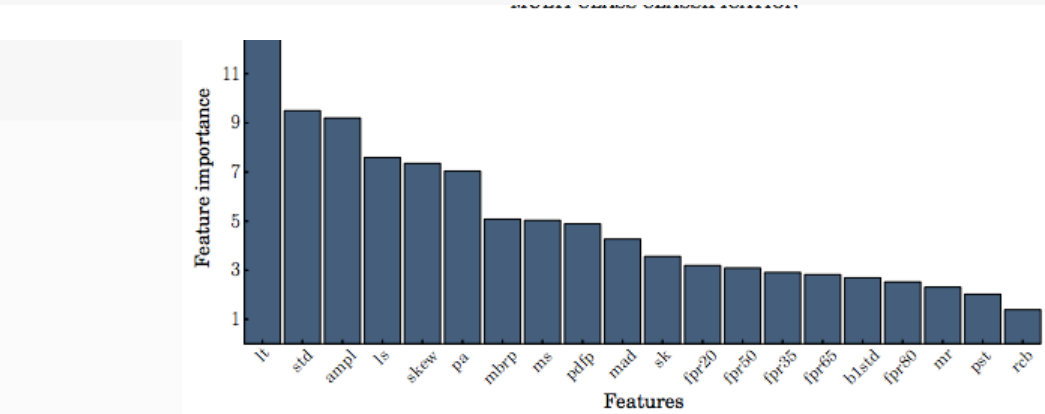
Dubath et al. 2012



Elorietta et al. 2016



Richards et al. 2012

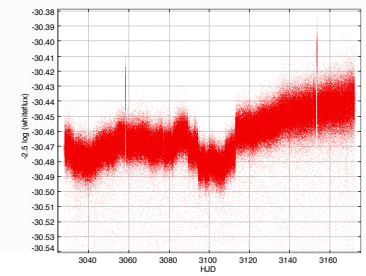
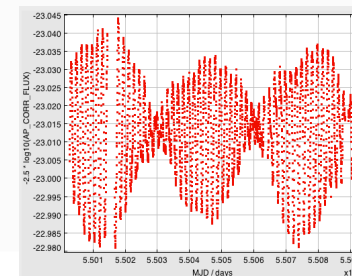
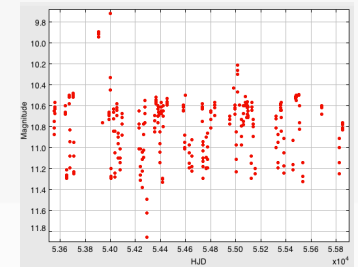
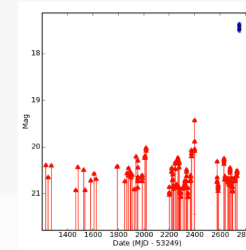


D'Isanto et al. 2016



The most important feature: period

- Many features used to characterize light curves rely on a derived period:
 - Dubath et al. (2011) show a 22% misclassification error rate for non-eclipsing variable stars with an incorrect period
 - Richards et al. (2011) estimate that periodic feature routines account for 75% of computing time used in feature extraction
 - Deep learning still applied to folded light curves
- Domain knowledge constraints
 - RR Lyrae: Blazho behavior (30%), small amplitude cycle-to-cycle modulations (RRabs)
 - Close binaries, LPVs: cyclic period changes over multidecade baselines
 - Semi-regular variables: double periods, multiperiodicity
 - ARMA models: quasi-periodicity
- Trustworthiness of quoted periods





Period finding is not a single algorithm

- Minimized (least-squares) fit to a set of basis functions:
 - Lomb-Scargle and its variants
 - Wavelets
- Minimize dispersion measure in phase space:
 - Means (PDM)
 - Variance (AOV)
 - String length
 - Entropy
- Rank ordering (in phase space)
- Bayesian
- Neural networks
- Gaussian process regression
- Convolved algorithms





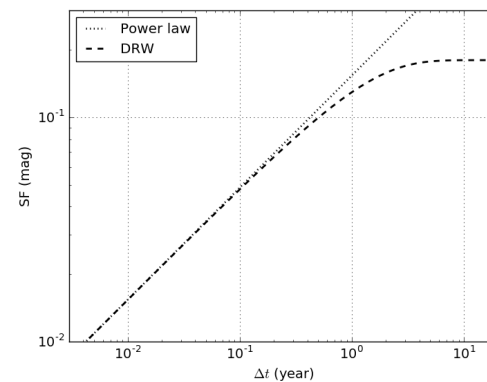
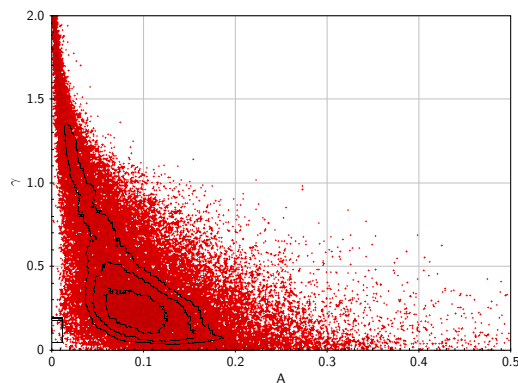
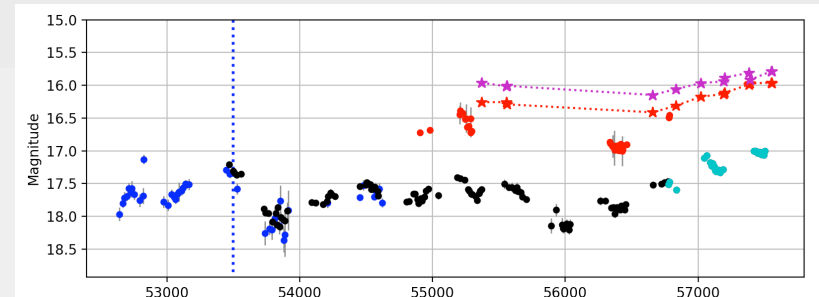
What can we say about period finding

- No algorithm is generally better than ~60% accurate
- All methods are dependent on the quality of the light curve and show a decline in period recovery with lower quality light curves as a consequence of:
 - fewer observations
 - fainter magnitudes
 - noisier data and an increase in period recovery with higher object variability;
- All algorithms are stable with a minimum bin occupancy of ~10 ($\Delta\phi = 0.1$)
- A bimodal observing strategy consisting of pairs (or more) of short Δt observations per night and normal repeat visits is better
- The algorithms work best with pulsating and eclipsing variable classes
- LS/GLS are strongly effected by half-period issue (eclipsing binaries)
- Specific algorithms work better with irregular sampling, bright magnitudes (containing saturated values), or with performance constraints



Describing quasar photometric variability

- $|\Delta m| > x$
 - DPOSS vs. SDSS (Stripe 82) vs. PS1
- Excess variability
- Structure function
 - Variability amplitude as a function of the time lag between compared observations
 - Historic descriptor of variability and a variety of estimators
 - Not much information

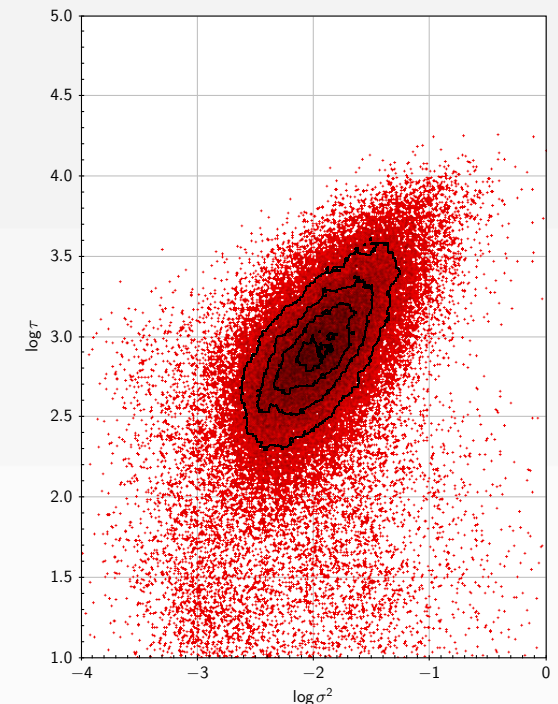
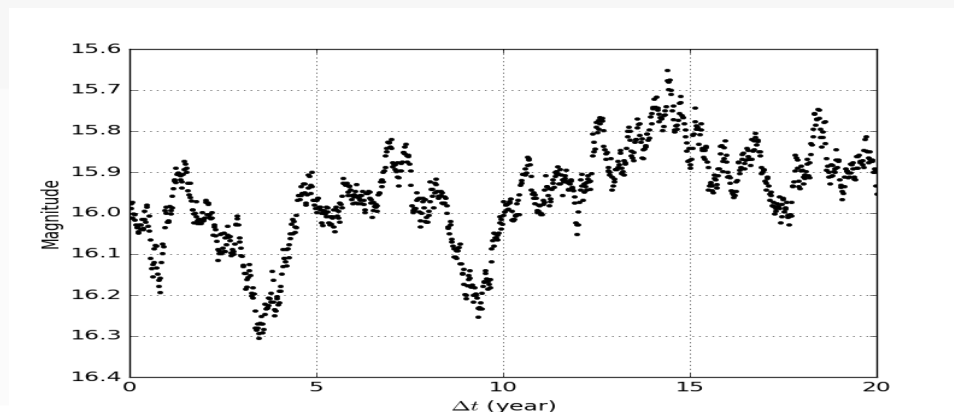




Damped random walk

$$dX(t) = -\frac{1}{\tau} X(t)dt + \sigma \sqrt{dt} \varepsilon(t) + bdt \quad \tau, \sigma, t > 0$$

- Characterized by variability amplitude and timescale
- Basis for stochastic models of variability
- Deviations noted (e.g., Mushotzky 2011, Zu et al. 2013, Graham et al. 2014)
- Degenerate model – can be best fit for a non-DRW process (Kozłowski 2016)

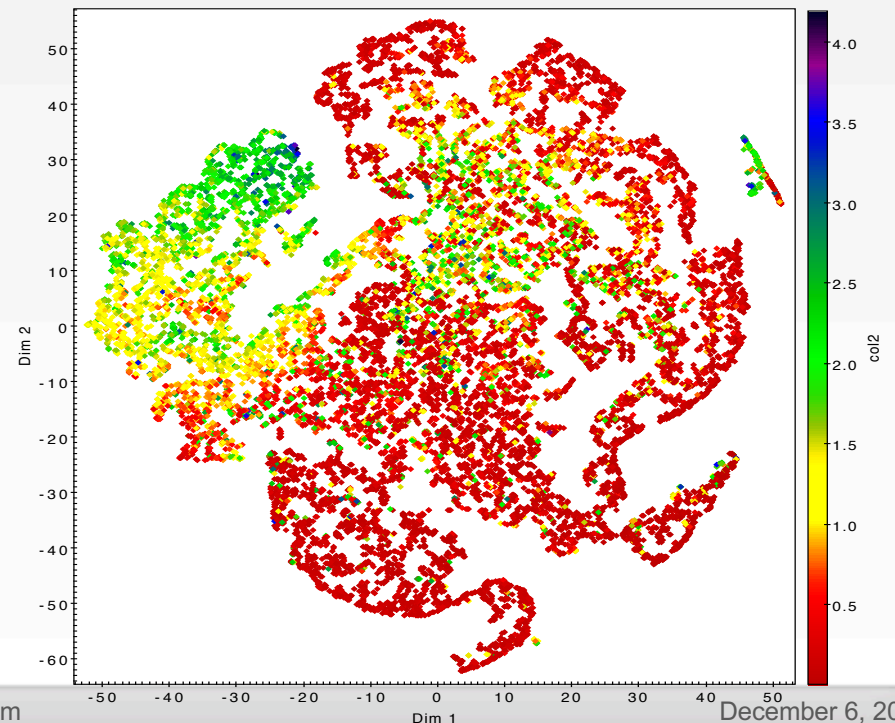
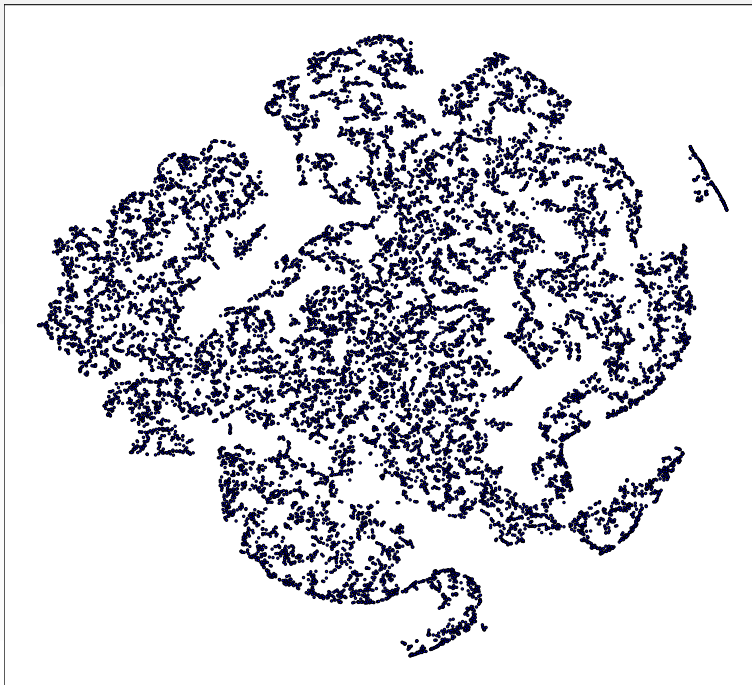


DRW+ = CARMA

- Better model is continuous time autoregressive moving average (CARMA; Kelly et al. (2015), Kasliwal et al. (2016))

$$\frac{d^p y(t)}{dt^p} + \alpha_{p-1} \frac{d^{p-1} y(t)}{dt^{p-1}} + \dots + \alpha_0 y(t) = \beta_q \frac{d^q \varepsilon(t)}{dt^q} + \beta_{q-1} \frac{d^{q-1} \varepsilon(t)}{dt^{q-1}} + \dots + \varepsilon(t)$$

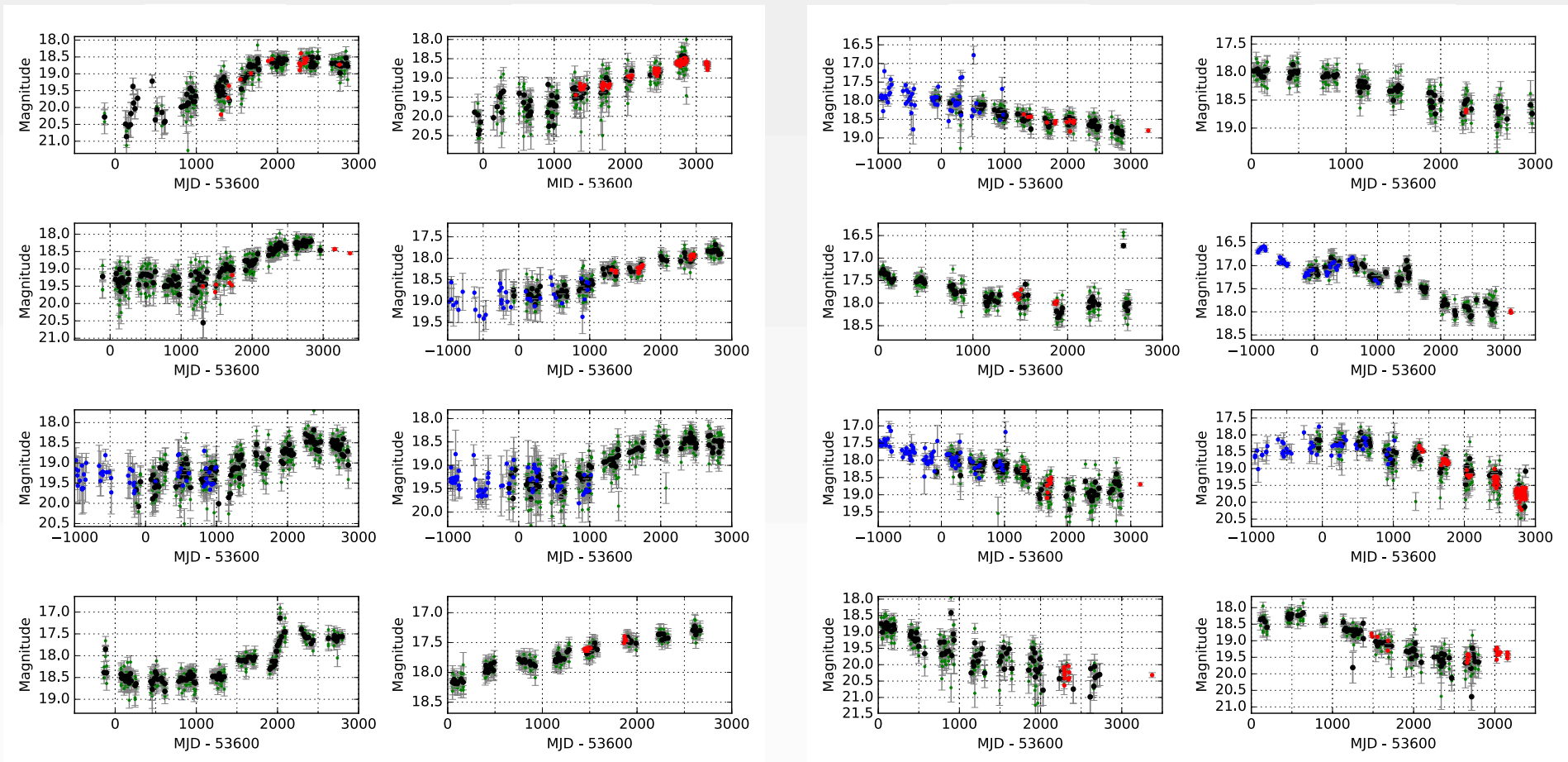
TSNE (t-distributed stochastic neighbor embedding) plot of restframe CARMA(3,2) parameters for 16498 AGN





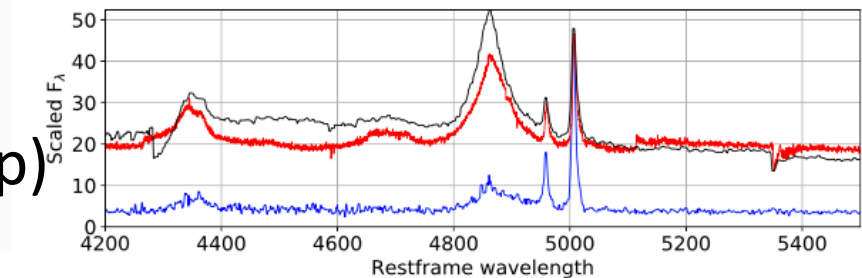
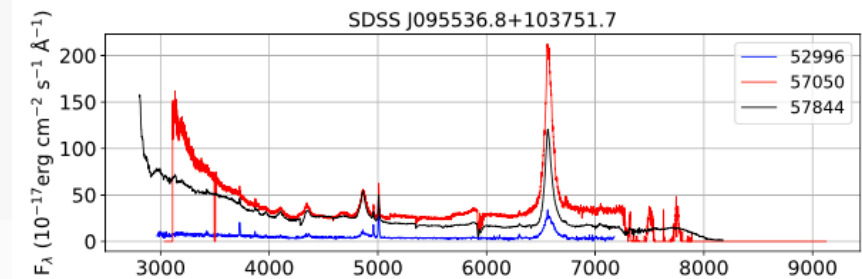
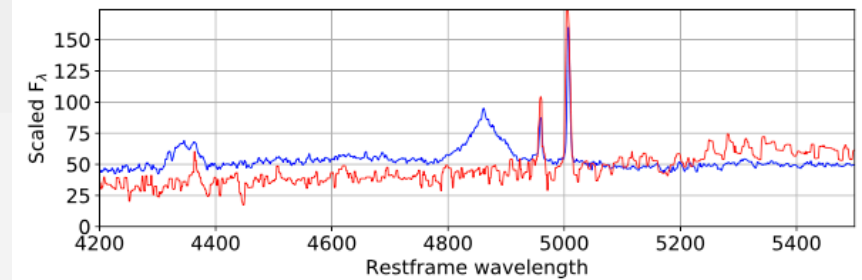
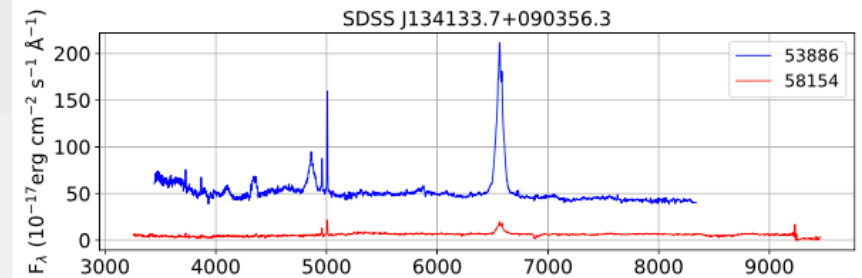
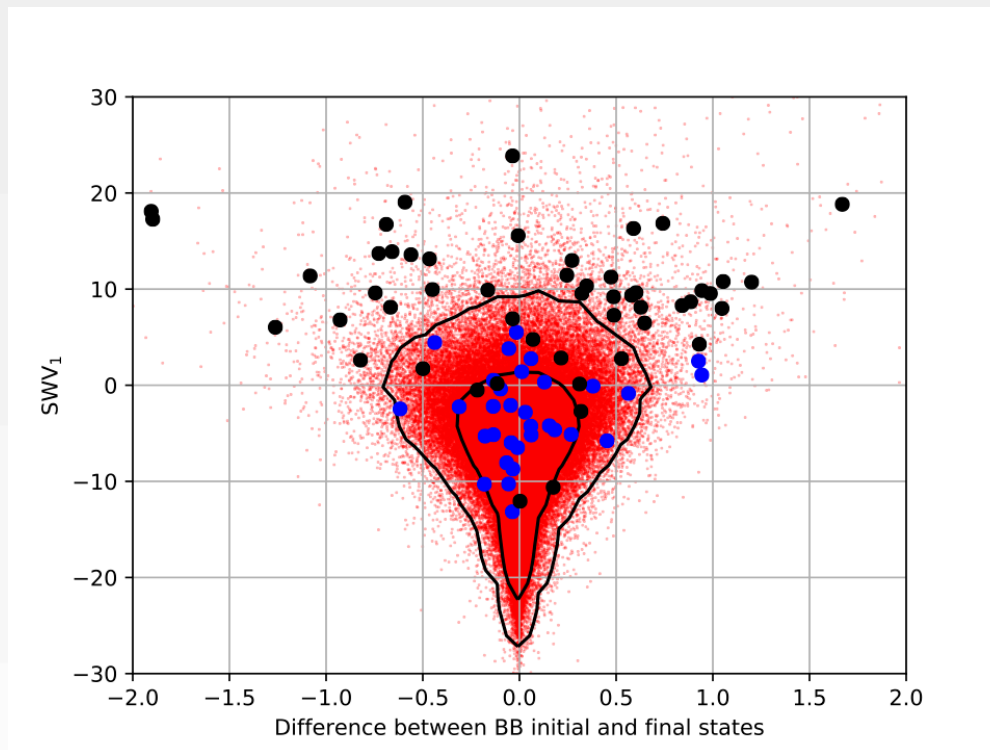
Changing state quasars

- Characterized by a smooth slow photometric rise/decline of ~ 1 mag over several years and some degree of spectral variability





Changing state quasars



10s of new CSQs (Graham et al., in prep)



Extremes: heavy tail or big outlier

- There is no reason why the characterized variability of every type of astronomical source in the observable universe over a decadal baseline should be Gaussian
- For a generic heavy-tailed distribution:

$$\lim_{x \rightarrow \infty} P[|X| > x] x^{-\alpha} = \lambda$$

λ and α cannot be estimated and so the general significance known

- There is no formal statistical definition for an outlier but it can be shown that the presence of outliers has no connection with the existence of heavy tails of an underlying distribution or with experimental errors (Klebanov 2016)



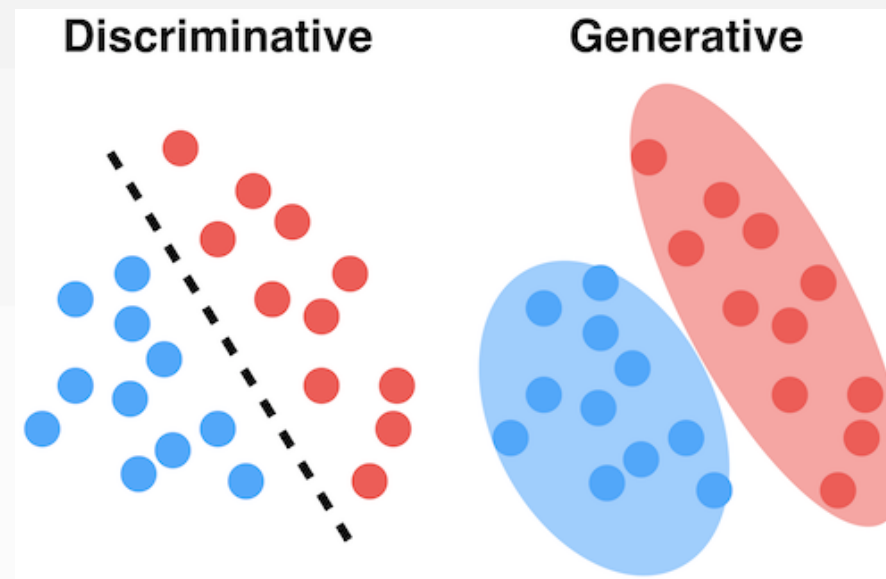


Elements of a 21st century approach



Generative vs. discriminative

- Current statistical models of variability are designed to discriminate between classes, e.g. stars/galaxies – $p(y|x)$
- Better to learn time series (shape) rather than determining some parameterizable form – $p(y, x)$
- Generative approach that supports predictions





Gaussian processes

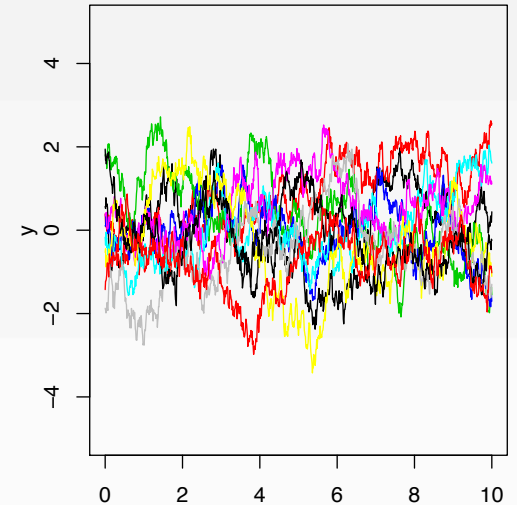
$$f(x) \sim \mathcal{GP}(\boldsymbol{\mu}, k(x, x'))$$

$$\ln p(\mathbf{Y}|\boldsymbol{\theta}) = -\frac{1}{2}(\mathbf{Y} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T K^{-1}(\mathbf{Y} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) - \frac{1}{2} \ln \det K_{\boldsymbol{\theta}} - \frac{N}{2} \ln(2\pi)$$

$$K_{SE}(x, x') = \exp\left(-\frac{r^2}{2l^2}\right), \quad r = \|x - x'\|$$

$$K_{OU}(x, x') = \exp\left(-\frac{r}{l}\right)$$

$$K_P(x, x') = \exp\left(-\frac{2 \sin^2\left(\frac{r}{2}\right)}{l^2}\right)$$



$$K_{celerite} = \sum_{j=1} J [a_j \exp(-c_j r) \cos d_j r + b_j \exp(-c_j r) \sin d_j r]$$

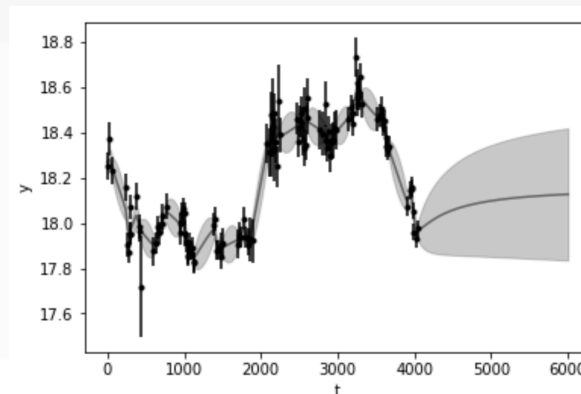


Better Gaussian process kernels

DRW = CAR(1) = CARMA(1,0) = CARIMA(1,0,0) = CARFIMA(1,0,0)

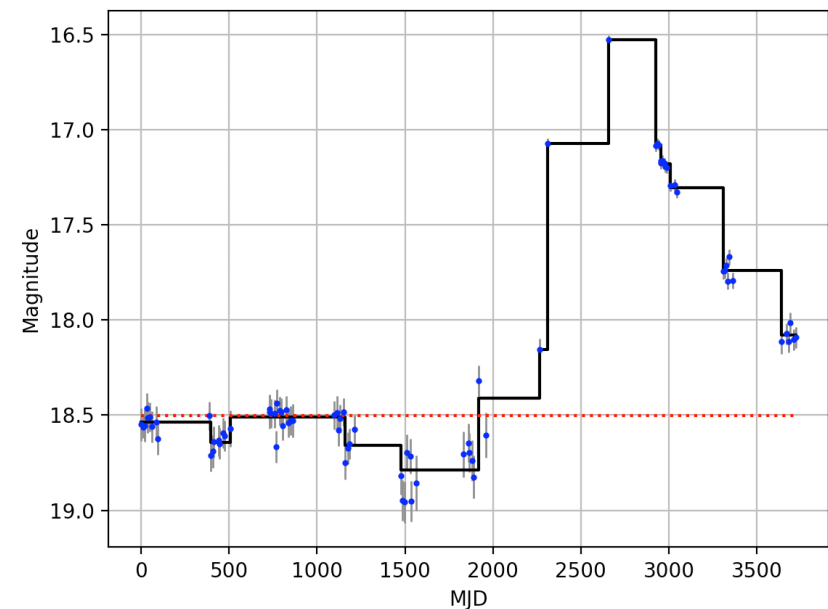
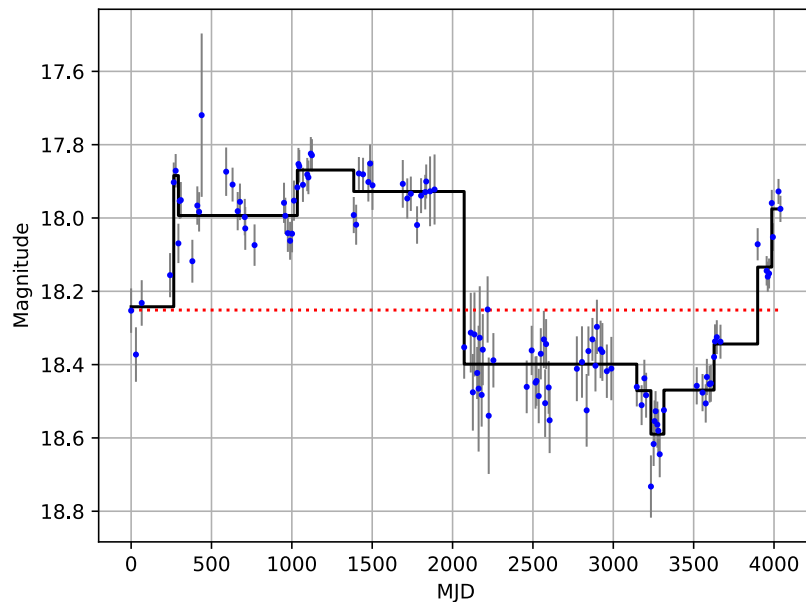
- (Zero mean) Gaussian processes are completely defined by their covariance function:
- No closed form for (super)parent models
- Fractional Brownian motion is equivalent to CARFIMA and a Cauchy class separates characterization of the fractal dimension (roughness) and long range dependence

$$K(x, x') = \sigma^2 (1 + (\theta |x - x'|)^\nu)^{-\nu}$$





Optimal representations

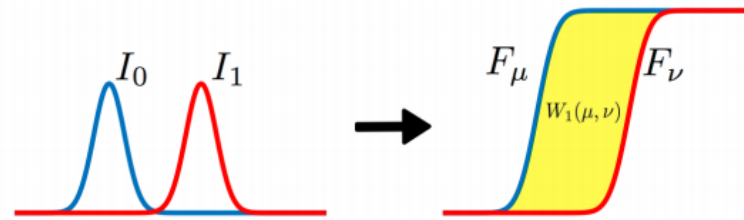
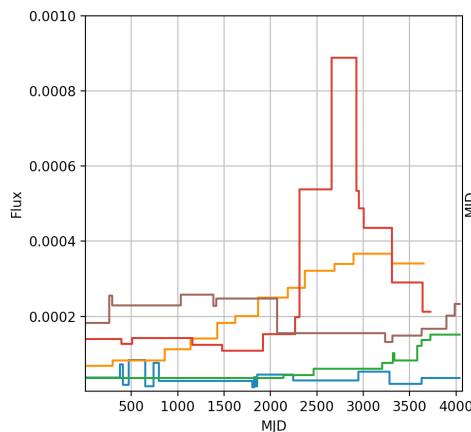


Bayesian Blocks (Scargle 2012) representation gives an optimal segmentation of the data in terms of a set of discontinuous piecewise continuous components



Distances between time series

- Consider the BB representation of the time series as the flux pdf over a time interval



- The p -Wasserstein distance for 1D probability measures:

$$W_p(\mu, \nu) = \left(\int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p dt \right)^{\frac{1}{p}}$$

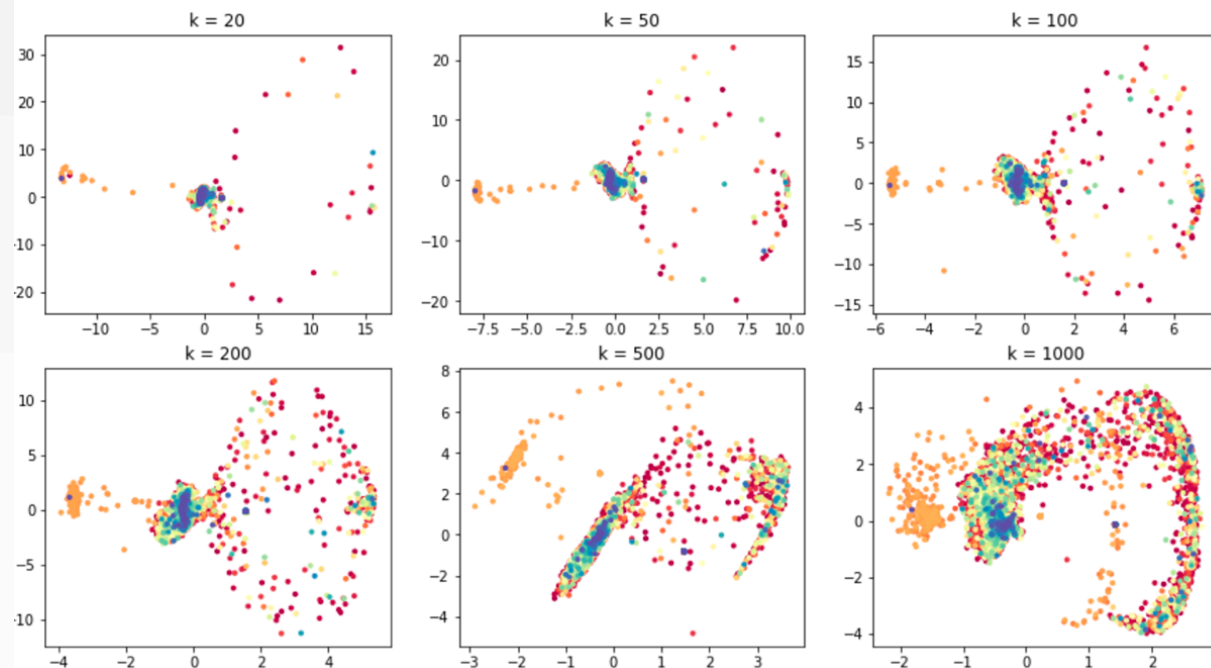
- For BBs, there is an analytical expression for W_2



Spectral embedding of time series

with Tamas Budavari and Tom Loredó

- Use eigenvectors/values of similarity graph to define features, parameterize data, perform clustering/regression
- Use a pairwise similarity measure constrained to local neighborhoods between time series
- Limit adjacency matrix to k nearest neighbors and solve





Combining time series

led by Rachel Buttry and Gordon Richards

What is the best method for combining (temporally overlapping) data from different passbands and/or surveys?

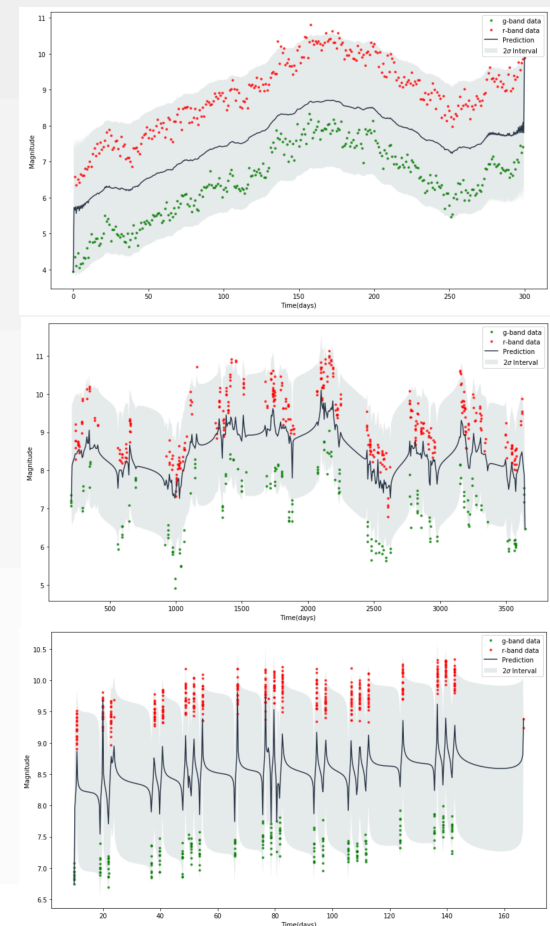
Expect a shift in magnitude/flux and potentially in time

Methods:

- Subtracting the means
- Minimizing standard deviation
- Minimizing distance to nearest neighbor
- Minimizing chi-squared statistic
- Coregionalized regression

Simulate:

- Idealized daily observations: minimizing std
- Wide Fast Deep (default LSST cadence): NN
- Deep Drilling Fields: NN



Dealing with uncertainties

- There are many sources of uncertainty:
 - Time series have observation errors in flux (and time)
 - Regularization and imputation add interpolation uncertainties
 - Model parameters and hyperparameters have uncertainties
- Feature representations do not traditionally deal with these (and will also introduce their own uncertainties)
- Probabilistic classifications tend to only simulate effect of observation errors through choice of priors or parameter space coverage
- Ideally full PDF should be given for each classification
- Uncertainty quantification (UQ) formally considers this:
 - forward uncertainty propagation (simulations and expansion methods)
 - inverse uncertainty quantification (Bayesian)



Detecting outliers

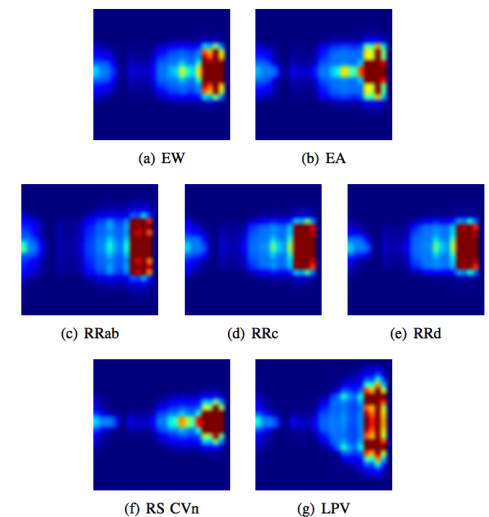
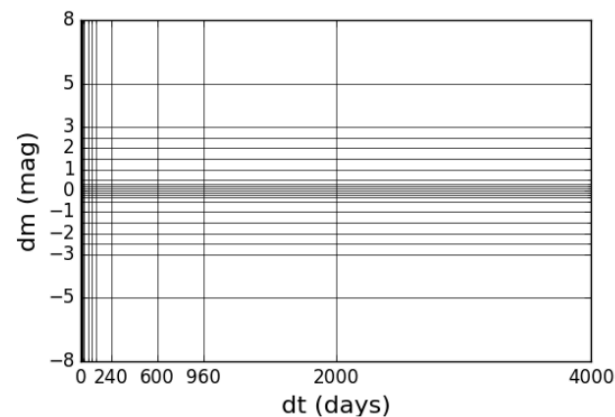
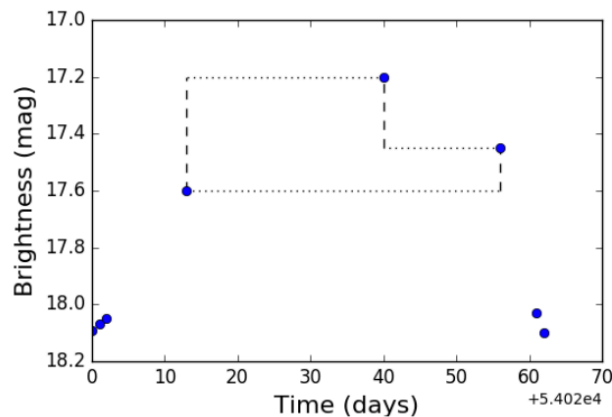


Some images from <http://proforhobo.com>



Going deep

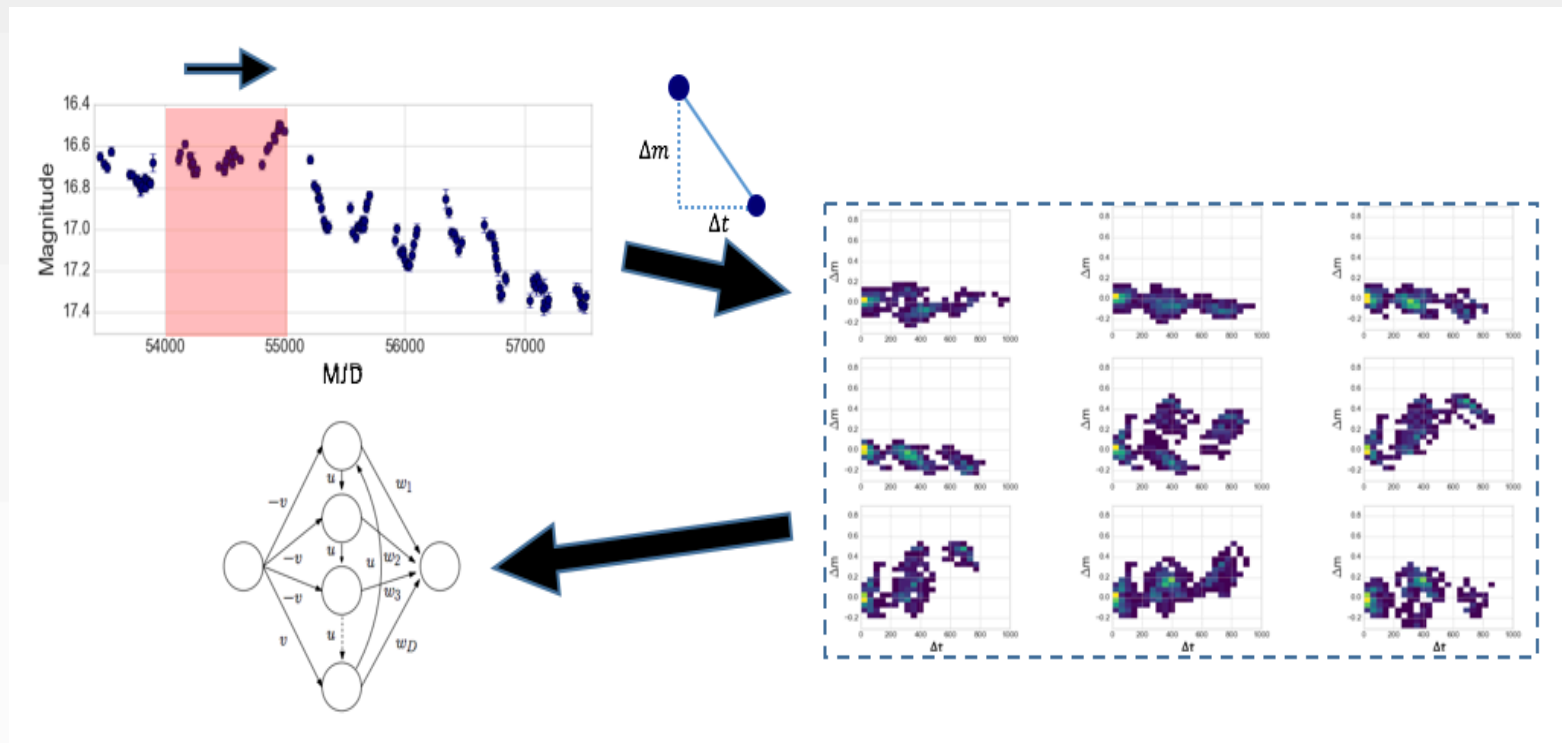
- Trendy, good for funding proposals
- Convolutional neural networks (CNNs) are current game changer for images
- Need to convert time series to image:
 - Wang & Oates (2015), Hatami (2017)
 - Mahabal et al. (2017) use dm-dt mechanism with variable stars





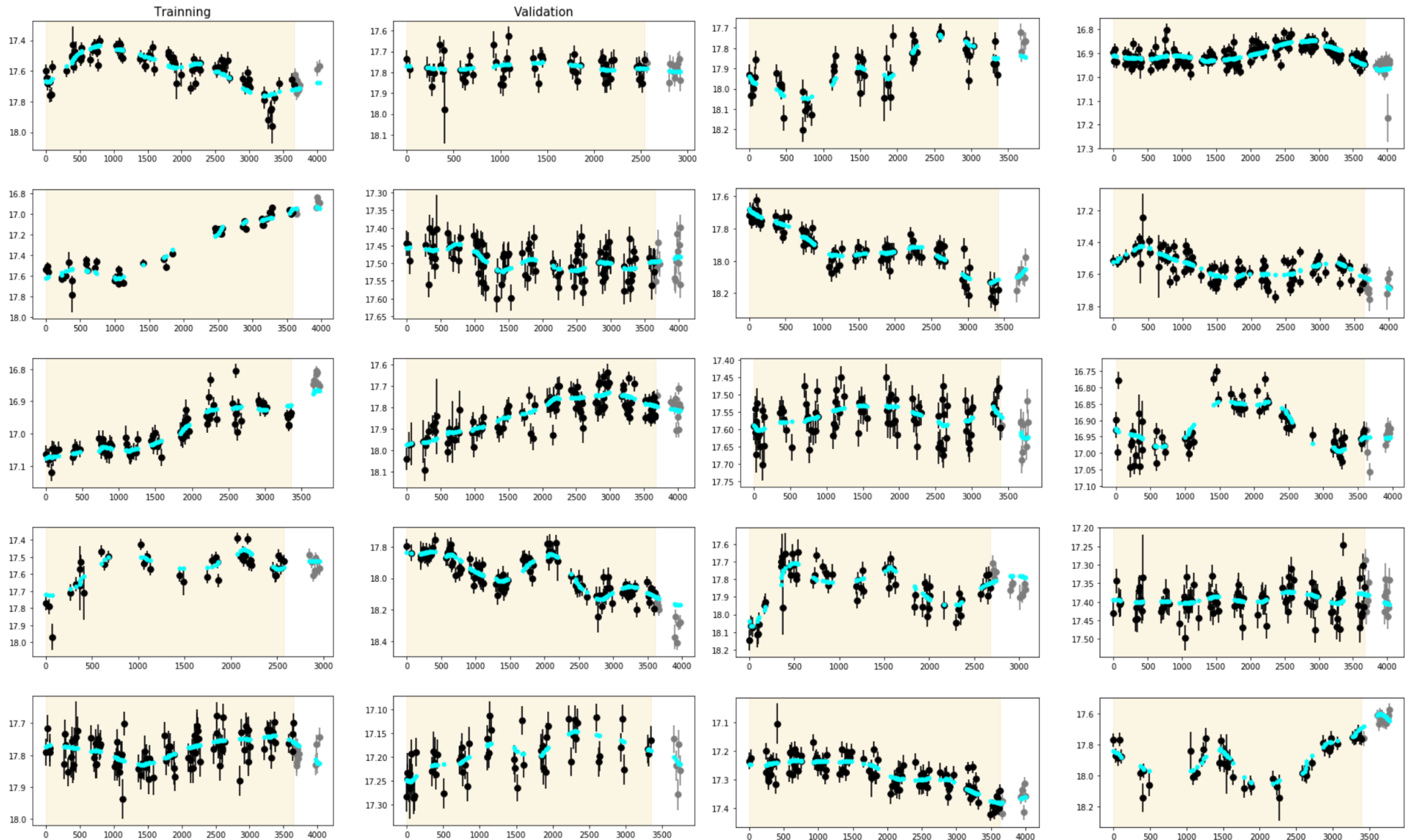
Going deeper

- Some neural networks architectures have “memory” - connections between links forming directed cycles, e.g., RNNs, LSTMs, good for time series



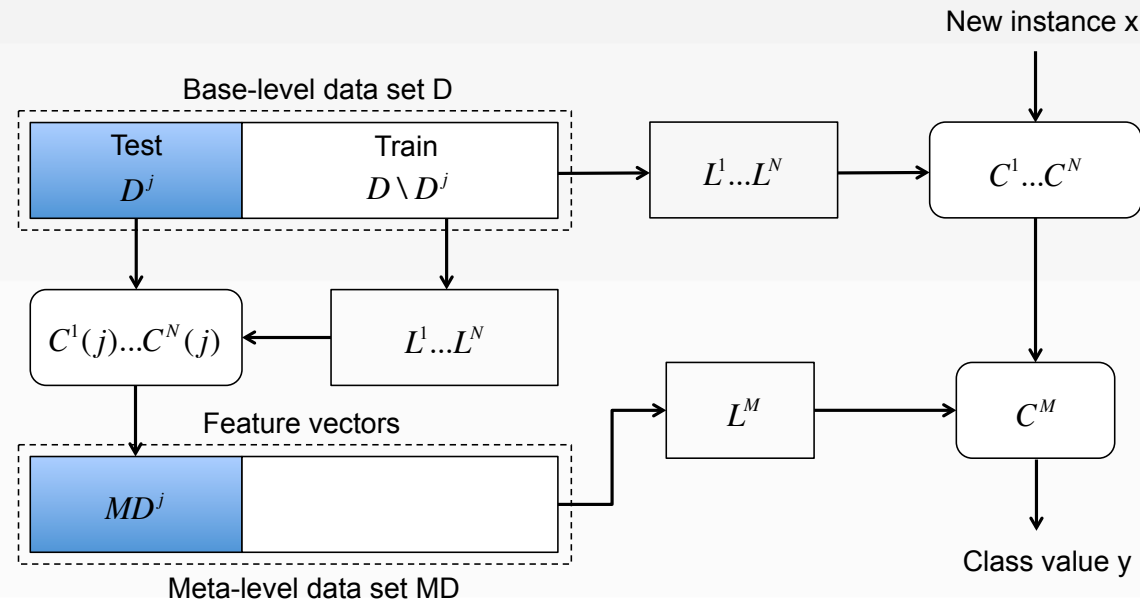
RNNs with QSOs

with Yutaro Tachibana



Which classifier?

- Most individual classifiers will give broadly the same results (precision and recall) for the same feature set, possibly with slight preferences for certain classes
- The state of the art is random forest (although see also support vector machine, Bayesian networks and self-organizing maps)
- Better results can be obtained with an ensemble classifier:





Summary

- Traditional time series analyses in astronomy involve:
 - (simple) discriminative features as (possible) inputs to machine learning algorithms
 - outlier detections based on Gaussian tails
 - little predictive power
- Data volumes now mean that we can *model individual* sources:
 - capturing full time series behavior
 - better identifying extrema
 - with generative approaches
- Next generation surveys enable real-time validation of predicted behaviors and swift identification of deviance

