Time Series Analysis for Synoptic surveys and Gravitational Wave Astronomy

Issues in time domain astronomy

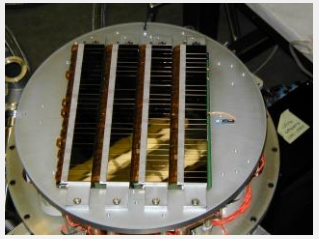
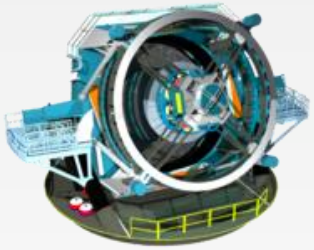
Matthew J. Graham
Center for Data-Driven Discovery, Caltech (and NOAO)
mjg@caltech.edu / graham@noao.edu
Andrew Drake, S.G. Djorgovski, Ashish Mahabal, Ciro Donalek

March 21, 2017

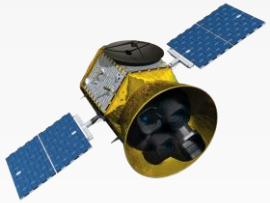
CENTER FOR DATA-DRIVEN DISCOVERY

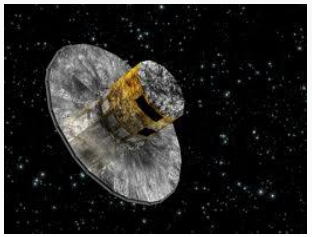
The burgeoning time domain

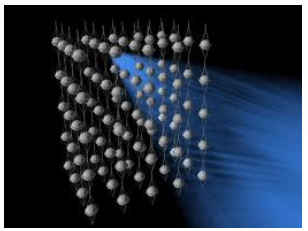
- Palomar-Quest Synoptic Sky Survey
- SDSS (Stripe 82)
- Catalina Real-time Transient Survey
- Palomar/Zwicky Transient Factory
- Pan-STARRs
- SkyMapper
- ASKAP
- ThunderKat (MeerKAT)
- KEPLER
- GAIA
- LIGO
- IceCUBE
- LOFAR
- LSST
- SKA
- TESS
- ASAS-SN
- MASTER
- DES
- …







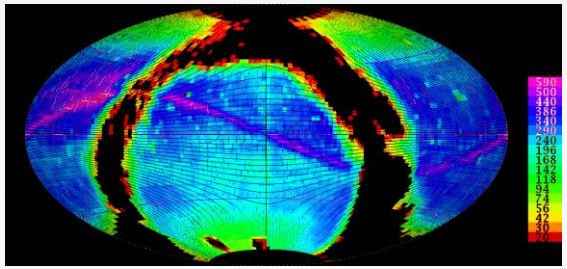
Optical surveys mentioned in ATELs

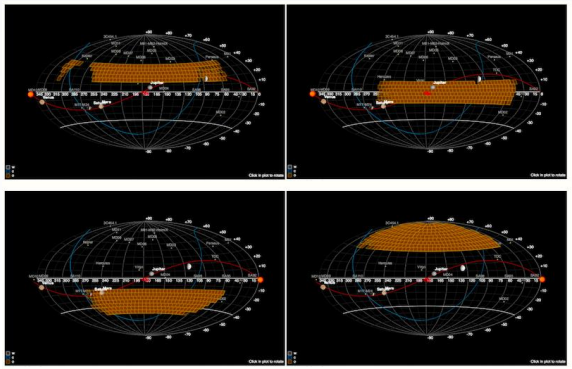
Photometric

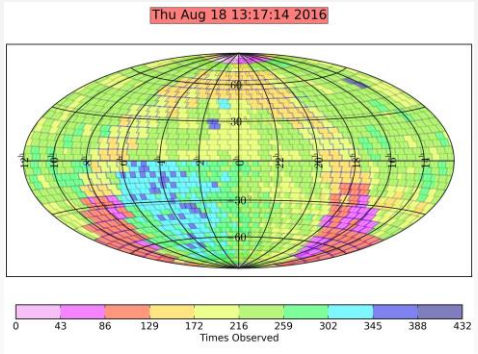
- Palomar-Quest Synoptic Sky Survey
- SDSS (Stripe 82)
- Catalina Real-time Transient Survey
- LINEAR
- MACHO
- EROS ?
- OGLE ?
- Kepler
- (i)PTF (2016-)
- Pan-STARRs (2016-)
- SkyMapper (2016-)
- DES (2018-)
- GAIA (2016-)
- ATLAS ?
- MASTER ?
- ASASSN ?

Spectral

- SDSS
- PESSTO
- LAMOST

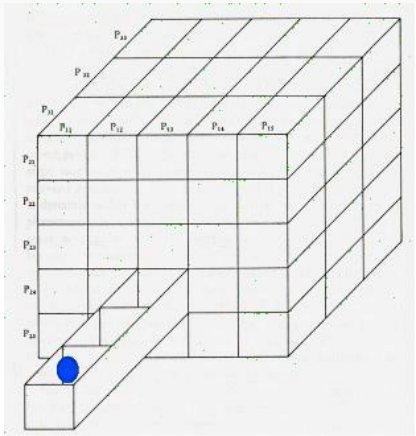






Automated classification work so far

- ASAS: Eyer & Blake (2005); Richards et al. (2012)
- CoRoT: Debosscher et al. (2009)
- Kepler: Blomme et al. (2010)
- Hipparcos: Dubath et al. (2011), Blomme et al. (2011)
- OGLE: Sarro et al. (2009)
- Stripe 82: Suveges et al. (2012)
- LINEAR: Palversa et al. (2013)
- CRTS: Drake et al. (2014, 2015, in prep.)
- VVV: Catelan et al. (2013); Elorrieta et al. (2016)
- EROS-2: Kim et al. (2014)
- WISE: Masci et al. (2014)



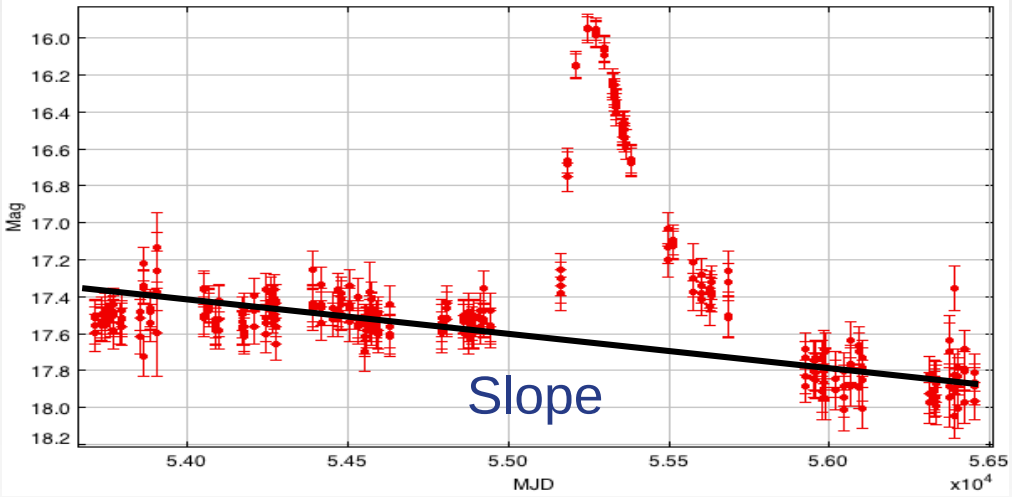
How to automatedly classify a data set

- Convert inhomogeneous data (irregularly sampled, noisy, gappy time series) to a homogeneous representation
  - Regularization
  - Feature representation
- Define ground truth on training set:
  - Specify labels
  - Identify clusters
- Train a classifier(s) on the training set
  - Random forest, support vector machine, convolutional neural network
  - Ensemble methods
- Verify on test set
  - False negative/positive rates
  - ROC curve
- Apply to full data set
  - Population statistics
  - Outlier detection

Characterizing astronomical time series

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

Fourier

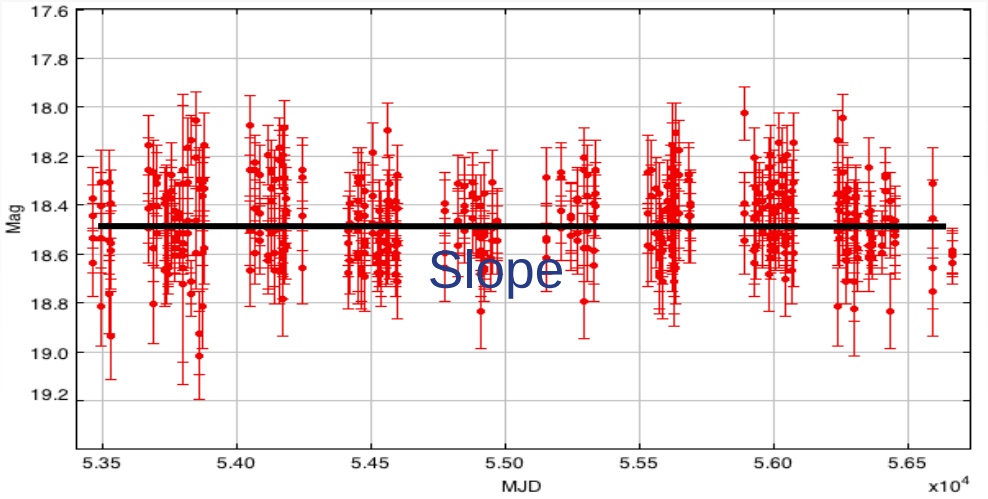


Amplitude

- Not all objects are this interesting…

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

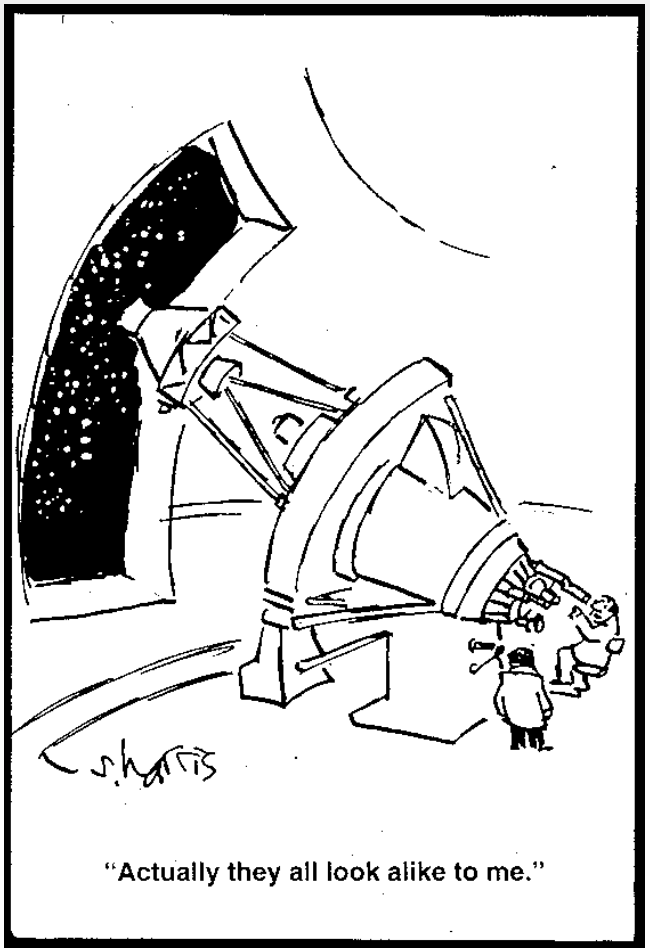
Fourier



Amplitude

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

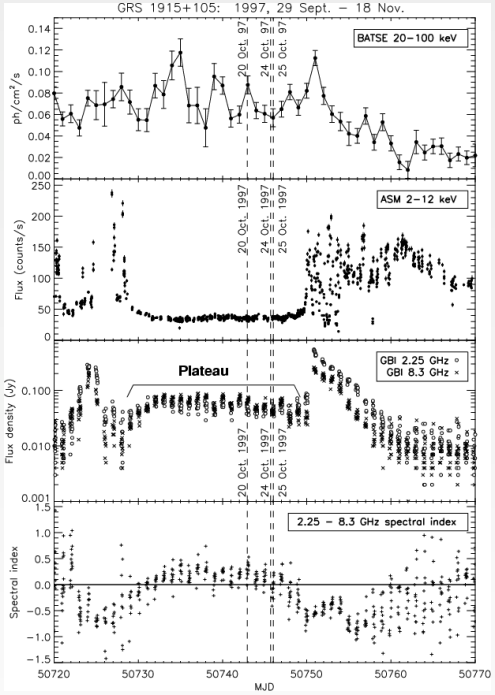


Unstated assumptions

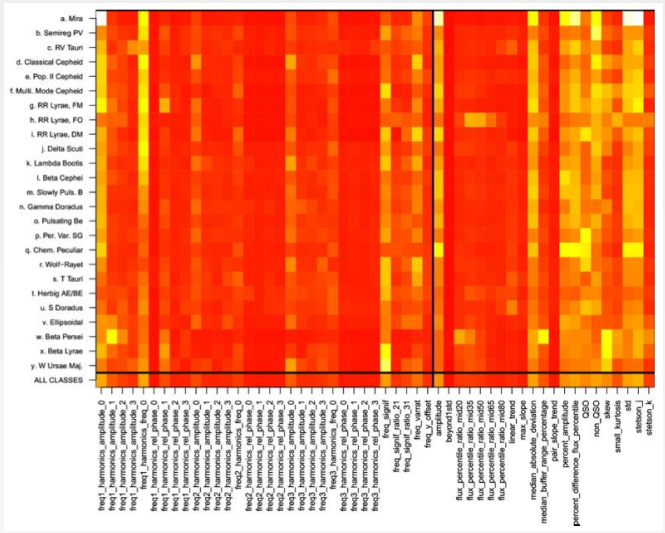
- Homoskedasticity
  - All errors are drawn from the same process (same variance model for all data points)
- Non-IID
  - Data is sequential
- Stationarity
  - The generating distribution is time independent
  - GRS 1915+215 has ~20 variability states
  - GARCH models: variance is a stochastic function 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

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

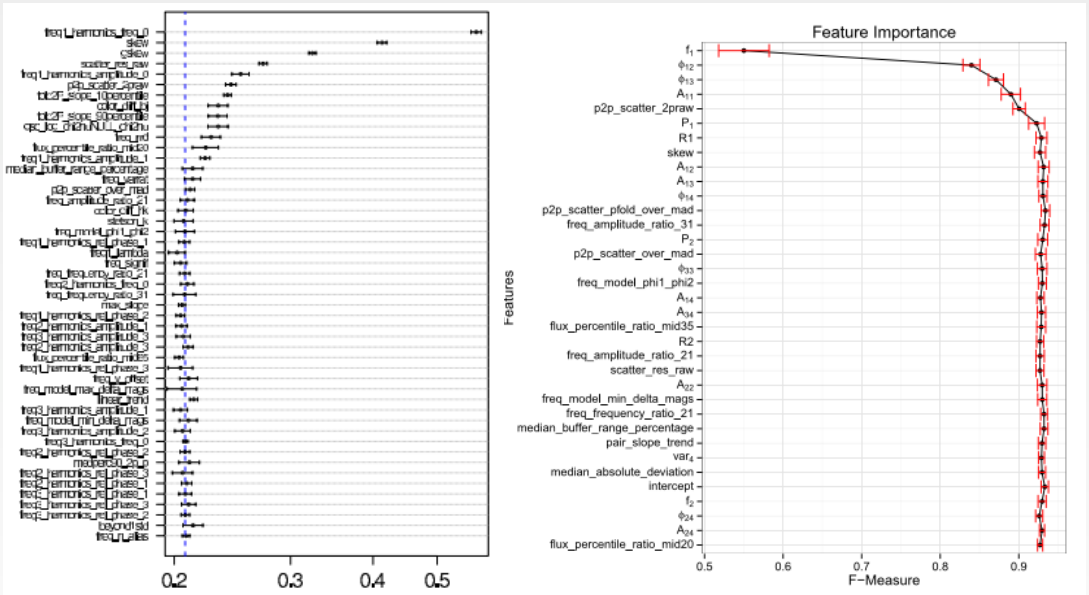
  - Observations sufficiently far apart in time are uncorrelated and new observations give extra information



Not all features are equal

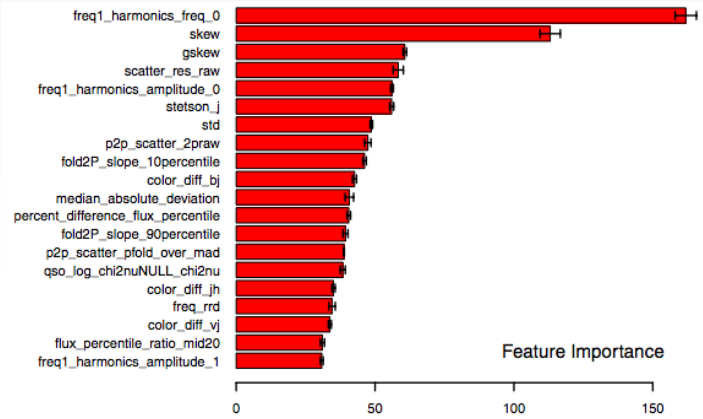


Richards et al. 2011

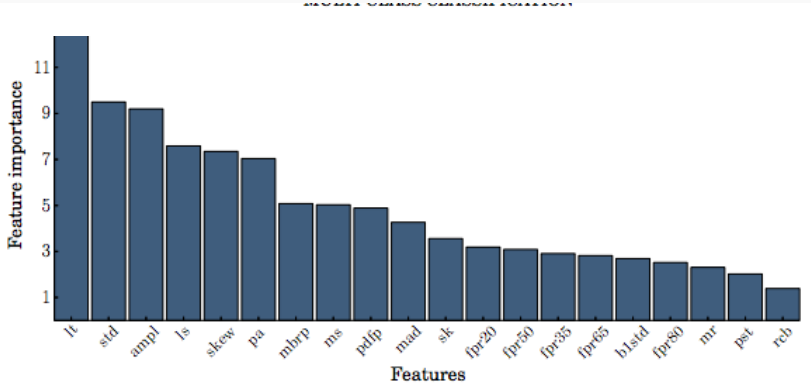


Dubath et al. 2012

Elorietta et al. 2016



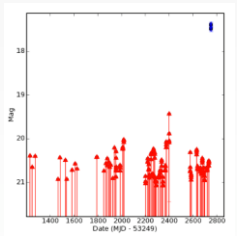
Richards et al. 2012

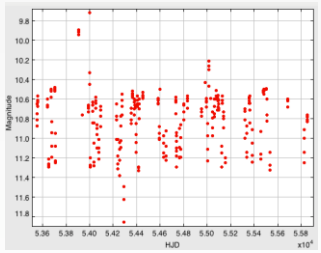


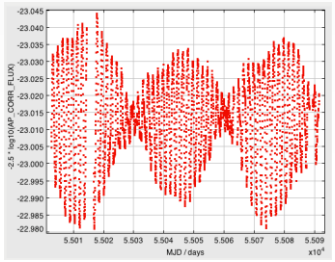
D'Isanto et al. 2016

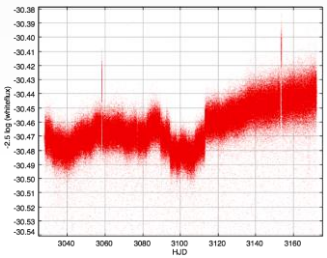
The most important feature: period

- 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
- 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

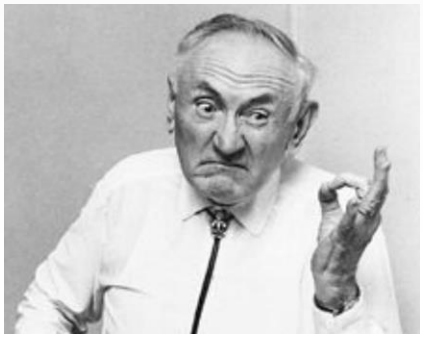
- Minimized (least-squares) fit to a set of basis functions:
  - Lomb-Scargle and its variants
  - Wavelets
- Minimize dispersion measure in phase space:
  - Means (PDM)
  - Variance (AOV)
  - String length
  - Entropy
- Rank ordering (in phase space)
- Bayesian
- Neural networks
- Gaussian process regression
- 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 ($\Delta\varphi = 0.1$)
- 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

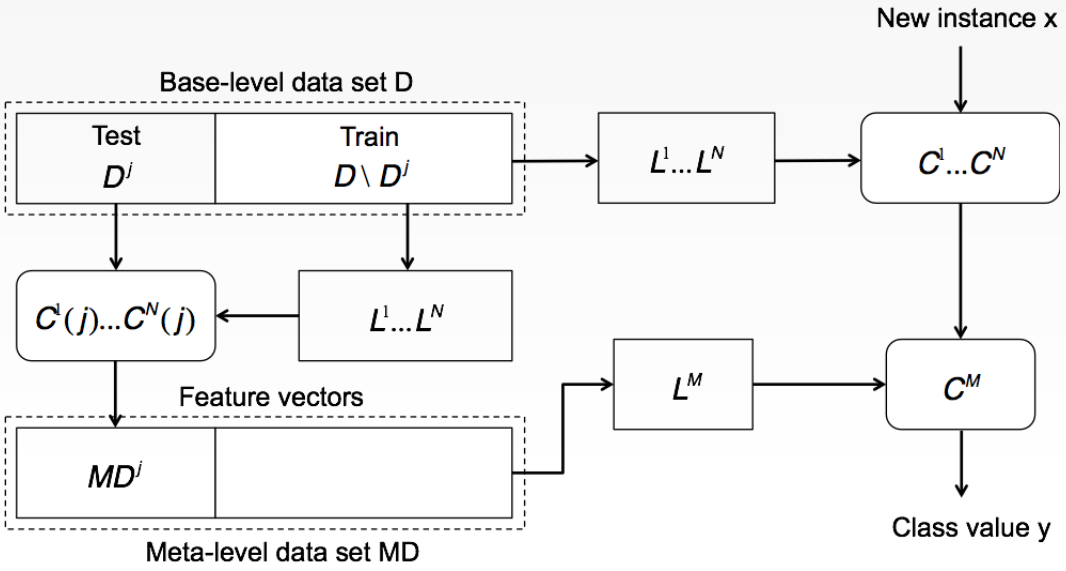
Are we using the best features?

- A lot of the “standard” features are correlated
- Use domain knowledge to define phenomenologically motivated features:
  - Autocorrelation: Durbin-Watson statistic
  - Long-term memory processes: Hurst exponent
  - Chaos: Lyapunov exponent
  - Nonlinearity: Teraesvirta measure
- Model time series via generative models that can capture deterministic and stochastic components:
  - ARFIMA (autoregressive fractionally integrated moving average) models
  - Echo state networks



Which classifier?

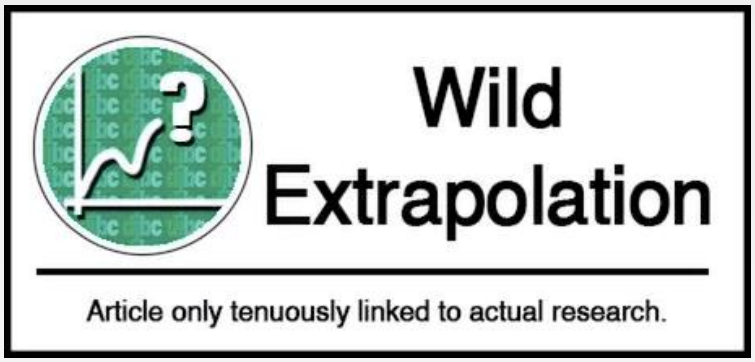
- Most individual classifiers will give broadly the same results (precision and recall) for the same feature set, possibly with slight preferences for certain classes
- The state of the art is random forest (although see also support vector machine, Bayesian networks and self-organizing maps)
- Better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



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)

A warning about automated classification



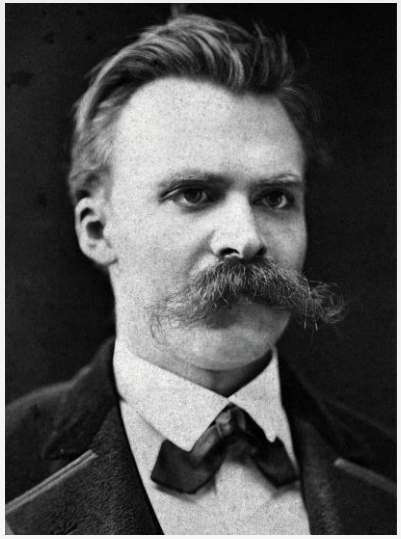
- Kim et al. (2014) used a random forest classifier to automatically detect periodic variables in EROS-II
- Soszynski et al. (2016) report that 4408 sources were missed in their new collection of RR Lyrae in the LMC
- 3234 have a counterpart in the OGLE-IV database and were carefully analyzed:
  - 149 (4.6%) are probable RR Lyrae (mainly noisy RRc)
  - 3085 (95.4%) are rather eclipsing binaries, δ Scuti, Cepheids or constant

Establishing ground truths

- The ground truth will not be 100% correct:
  - Classifier is mistrained
  - Raises false positive/false negative rates in data
  - 1% misclassification is highly significant in a 100 million object data set
- There is a lack of a (comprehensive) community-agreed training set, particularly at fainter magnitudes (V ~ 16)
  - GCVS?
  - VSX?
  - SIMBAD?
- Known class sampling is biased towards certain classes and not representative of real population statistics
  - Periodic variables: RR Lyrae, eclipsing binaries
  - MACC: semiregular variables (~20%), small amplitude red giants (~20%)

The nature of categorization

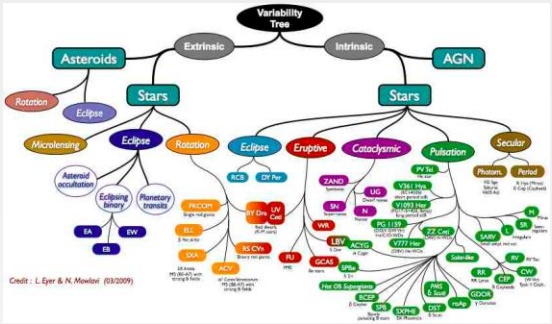
- The definitions of mathematical objects are immutable
- Our definitions are based on snapshot samples with the hope that we have captured enough objects at all phases (training set considerations)
- Analytic (ideal, theoretical) vs synthetic (pragmatic, nuanced) classes
- With LSST, ZTF, etc., we are entering the era of *precision classification*



Nothing that has a history can ever be defined

Class distinctions

- How many classes are in a data set?
- Domain knowledge gives us:
  - Intrinsic vs. Extrinsic
  - Eruptive, Pulsating, Rotating, Cataclysmic, Eclipsing



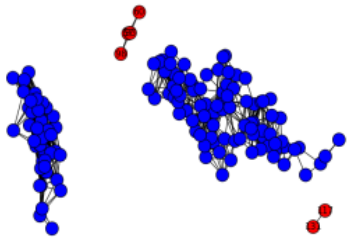
- Standard automated classification assumes mutually exclusive classes but *fuzzy* clustering allows multiple memberships
- Unsupervised learning can determine clusters:
  - Different clustering paradigms: partitional, hierarchical, density-based, grid-based, correlation, spectral, gravitational, herd, …
  - Are these clusters physically realistic?
- The “other” class:
  - MACC has 60% of objects in MISC class
  - Most accurately classified in the CRTS periodic variable catalog is the aperiodic noise class

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:

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

  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)
- "An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism" (Hawkins 1980)
- From a topological perspective, a significant outlier has high persistence and marginal connectivity



Further challenges for automated classification

- Data fusion:
  - Multiband/wavelength data
  - Different surveys in similar passbands
- Domain adaptation:
  - Transferring what has worked for one survey to another
- Performance/hardware:
  - Mean characterization time for a time series
  - Computational resources available
  - Reclassification with new data (streaming)

Summary

- Automated classification is a necessity with large data sets
- There are a number of caveats to be aware of:
  - The choice of features, their inherent assumptions and dependencies
  - How uncertainties propagate through to the final classification
  - The choice of learning algorithm (relatively unimportant)
  - The ground truth (training set) and what it assumes/reflects
  - The class assignments available and their reality
  - Outlier detection
- There are further considerations regarding combining data sets, transferring techniques between data sets, and performance
- The promise is statistically significant samples of 1 in 10^4 – 10^6 phenomena