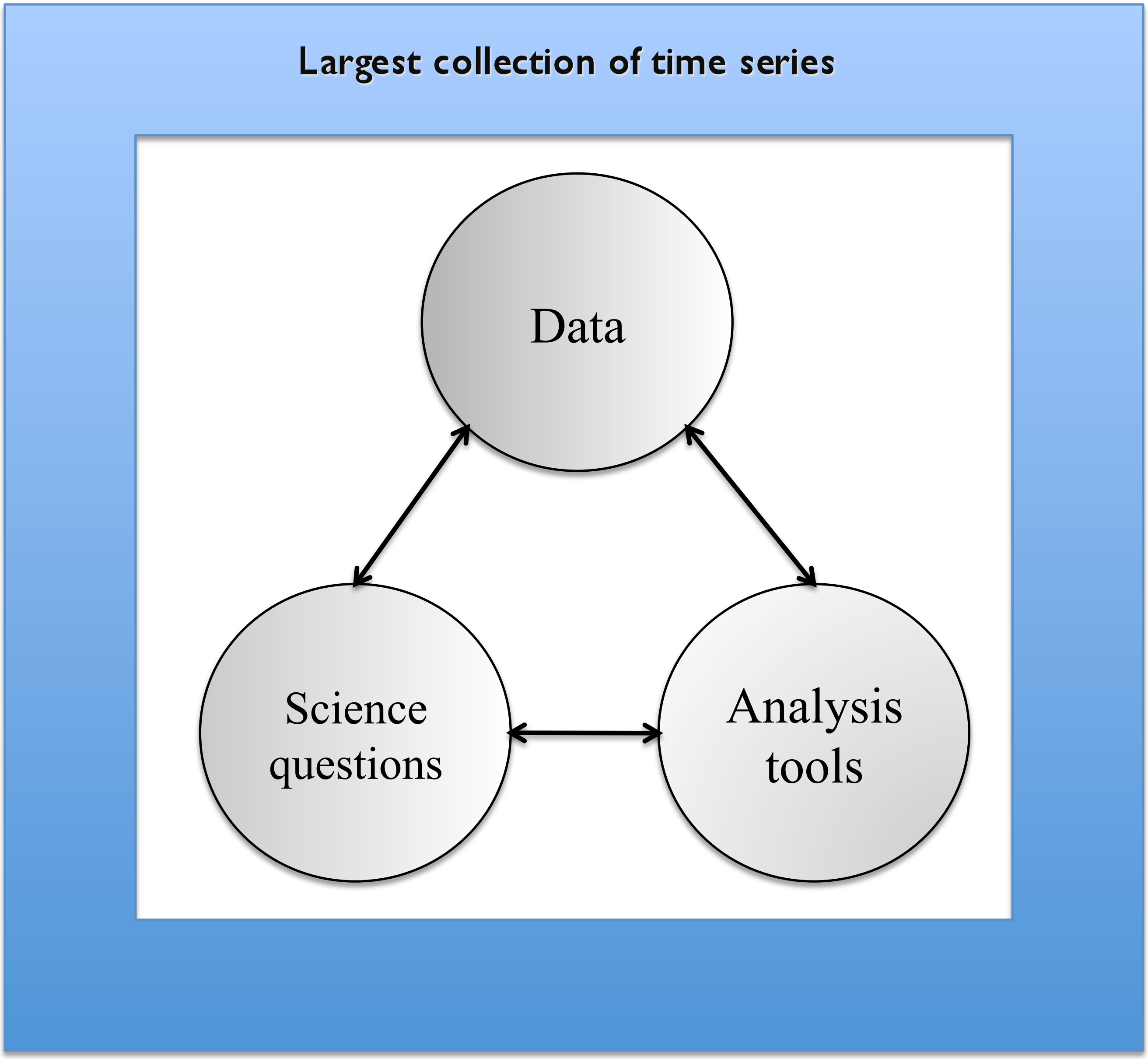
Machine Learning and Statistics Applications in Astronomy

Pavlos Protopapas (CfA and Institute of Applied Computational Science)
The Time Series Group

Time Series Center

- An interdisciplinary group created in 2008 at the Initiative in Innovative Computing (IIC).
- Moved to Institute of Applied Computational Sciences in 2010 (IACS).
- Focus now is in astronomy but since early 2010 we have labor data, real estate data, heart monitor data, archeological data, brain activity etc. and research activities in these areas.
- Vision: Largest time series collection and center of expertise in time series analysis.



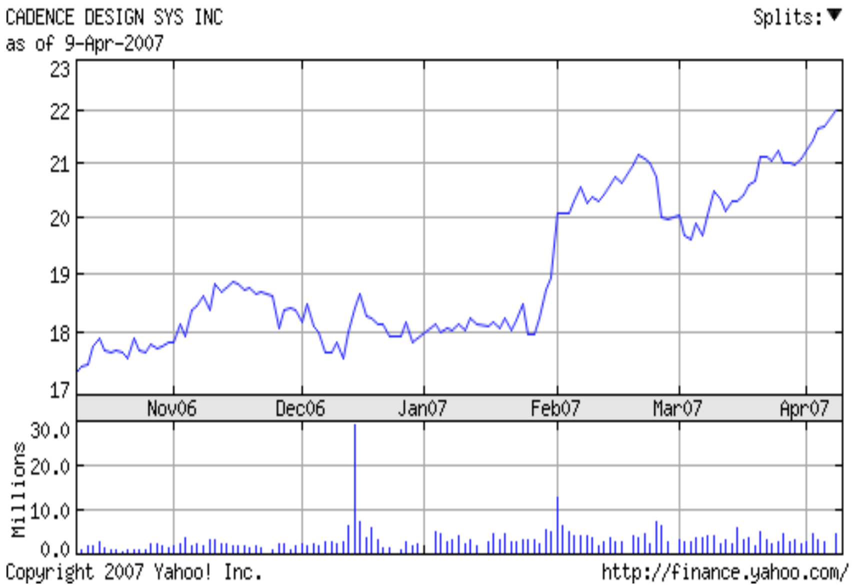
Largest collection of time series

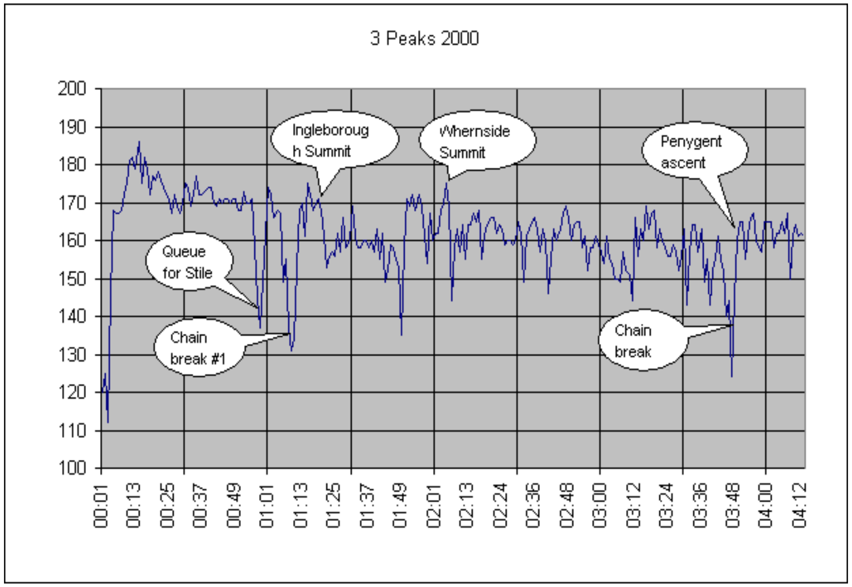
Data

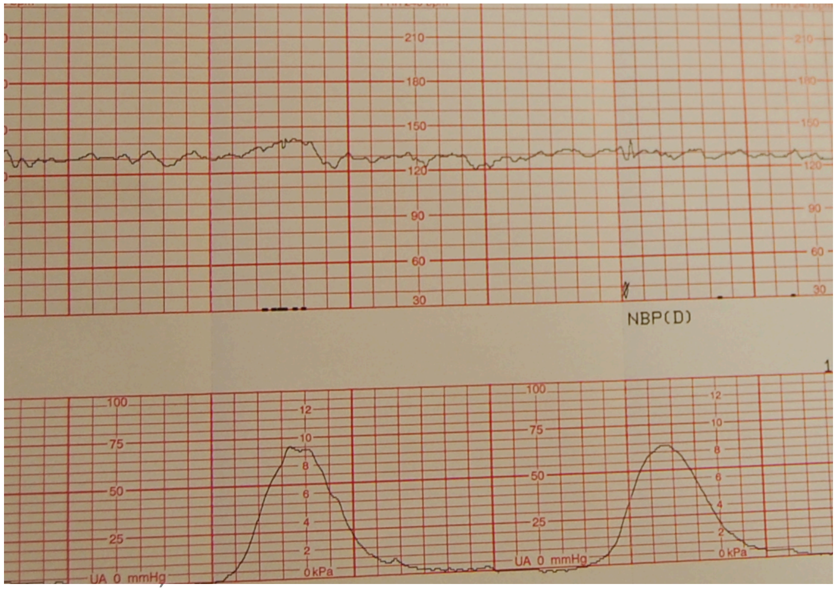
Science questions

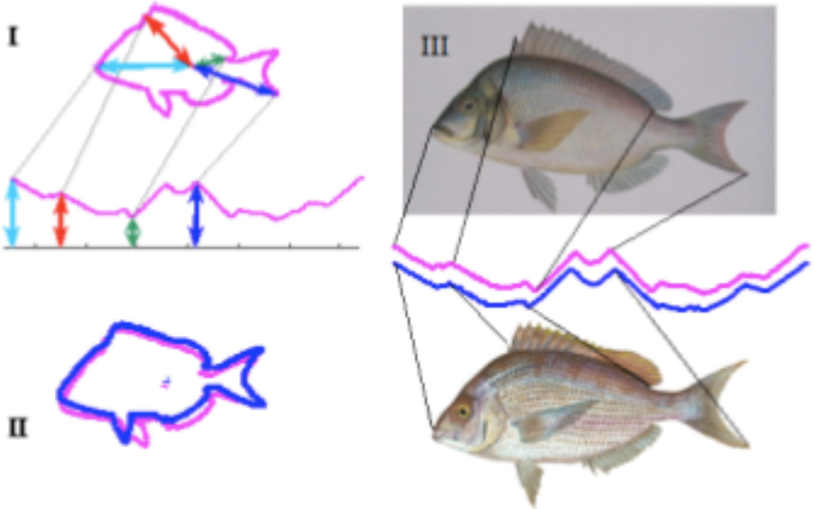
Analysis tools

Why time series









Data (astronomy)

MACHO	66 million objects. 1000 flux observations per object in 2 bands (wavelengths)
EROS	~100 million objects. 1000 flux observations per object in 2 bands
OGLE	few million objects.
TAOS	100000 objects. 100K flux observations per object.
ESSENCE	thousands objects, hundred observations.
Pan-STARRS	billion of objects. (only medium deep fields for now)
MMT	occultation studies, variability studies,
SDSS 82	1 million objects ugriz about 100 epochs
DASCH	Harvard plates, 100 years
LSST	A lot !

Hardware/Computational

Storage: Disk: ~100 TB of disk space

Filesystems: Lustre, NFS

Computing : Odyssey cluster at Harvard (~8000 cores)

GPU cluster with 32 machines with 2xNvidia Tesla T10

db : Postgresql, migrating to BigTable and MapReduce.

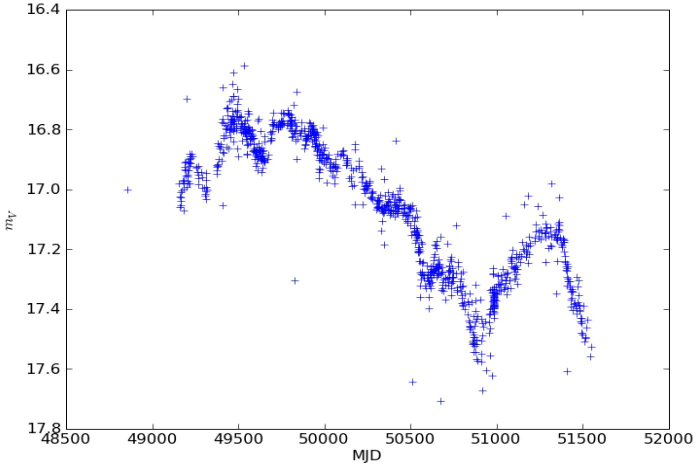
Multiple machines with a lots of memory.

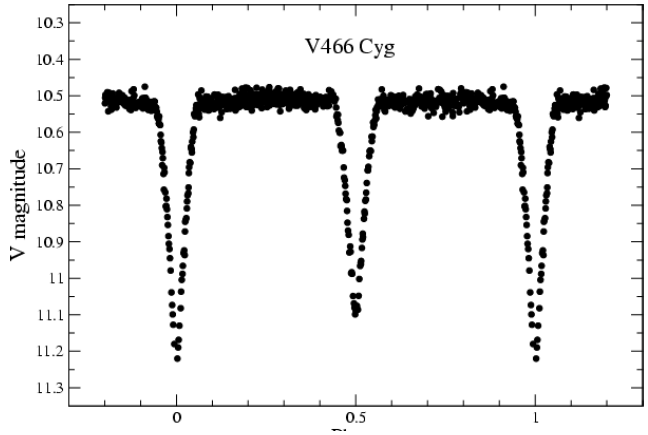
Web server: 2 dual machines

Web site with search engine and online tools and VO standards

Questions/Wish List

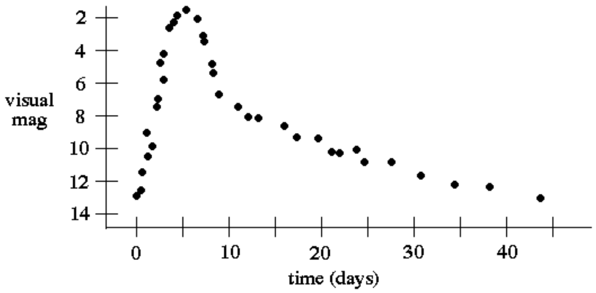
- **Classification**
  - Be able to classify objects based on their variability characteristics: Quasars, Variable Stars, Supernovae, etc





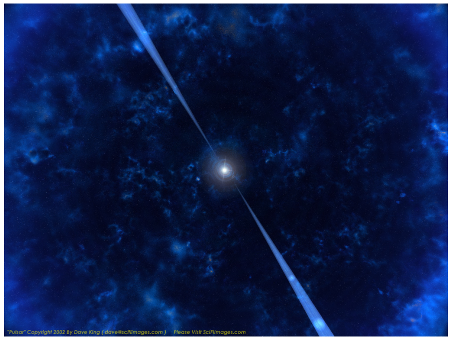
- **Period finding**
  - For *sparse* and noisy data, period determination is not easy
- .

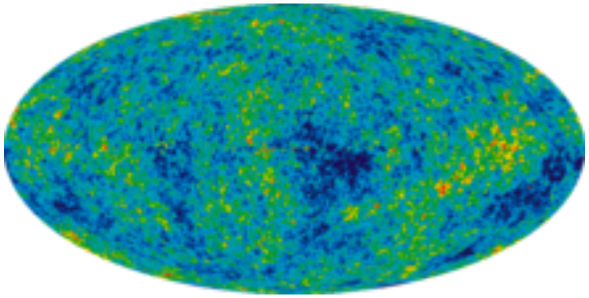
Nova Light Curve

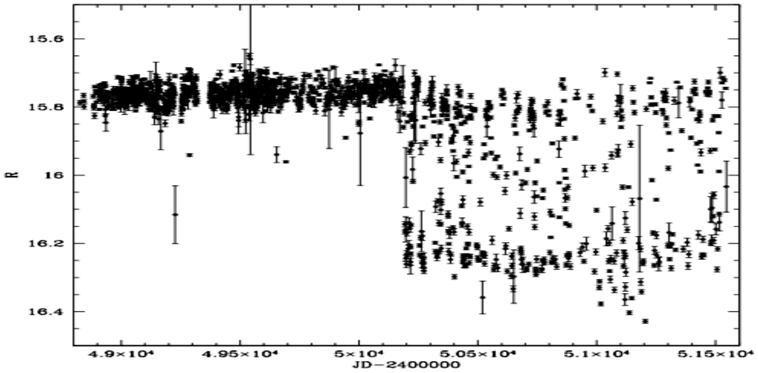


Questions/Wish List

- **Novelty detection**
  - Classify something as novel
  - Pulsars (Jocelyn Bell Burnell and Antony Hewish 1967 while looking for Quasars).
  - CMB (Arno Penzias and Robert Wilson 1967 using a horn antenna designed to relay telephone calls via satellite
  - Four Jovian moons (Galileo 1609)


"Pulsar" Copyright 2002 By Dave King (dave@scifiimages.com)   Please Visit SciFiImages.com





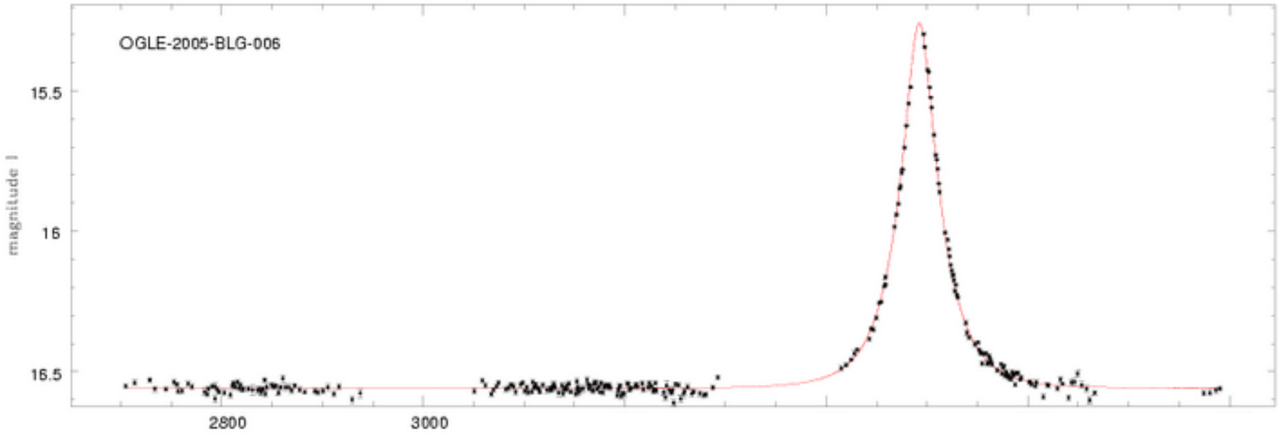
Questions/Wish List

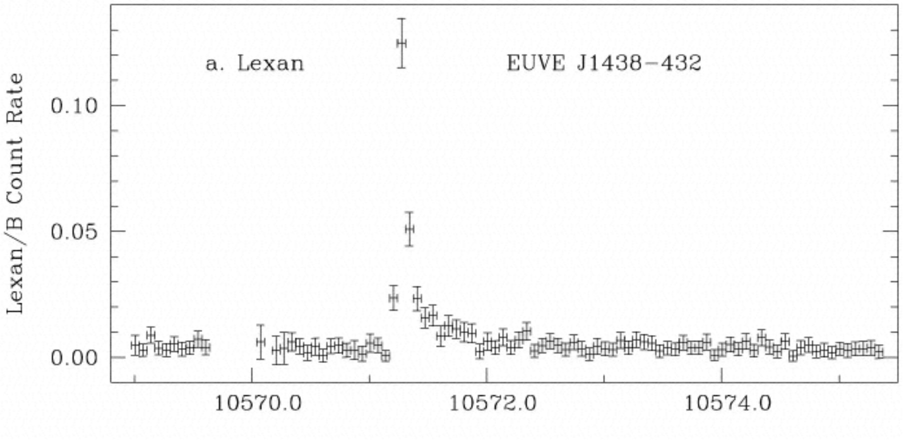
- **Event detection**
  - Rare, low signal-to-noise ratio

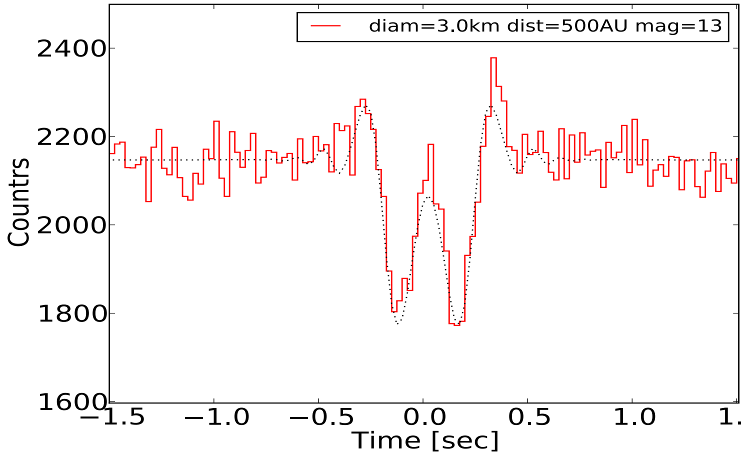
Occultation

Microlensing

Stellar Flares







Questions/Wish List

- **Time Series modeling**

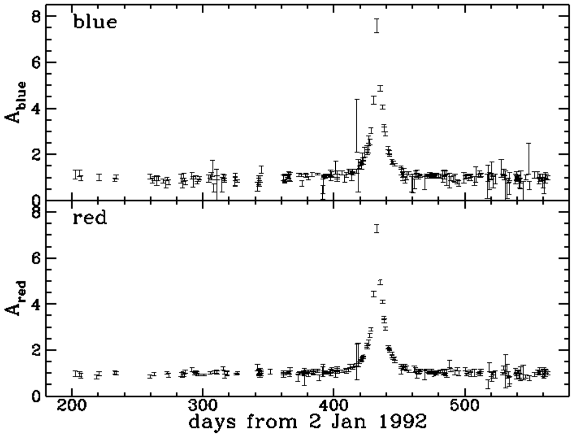
  autoregresssion coefficients

The challenges

- Educate astronomers in machine learning/statistics
- Large dataset
  - MACHO 1990s 1TBytes
  - SDSS 40TBytes
  - Pan-STARRS 2011: 2 Tbytes per night
- Irregular sampling
- White, red, blue, purple noise that

sometimes is not even stable

- Dealing with the cultures.



Classification

- SVN for classifying periodic variables; use:
  - color,
  - magnitude,
  - period estimations and
  - lightcurve morphology

  analyze the whole MACHO and EROS database of 120 million lightcurves (Wachman et al 2009, Protopapas et al 2011)
- Cross-correlation

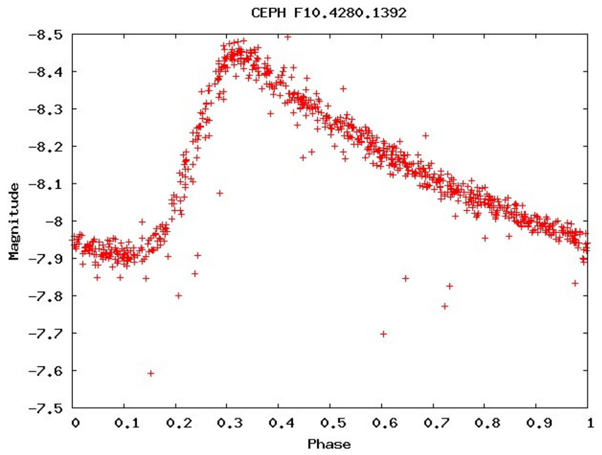
$$K(f,g) = \max_{\tau} \sum_{i} (f_i - g_{i+\tau})^2$$

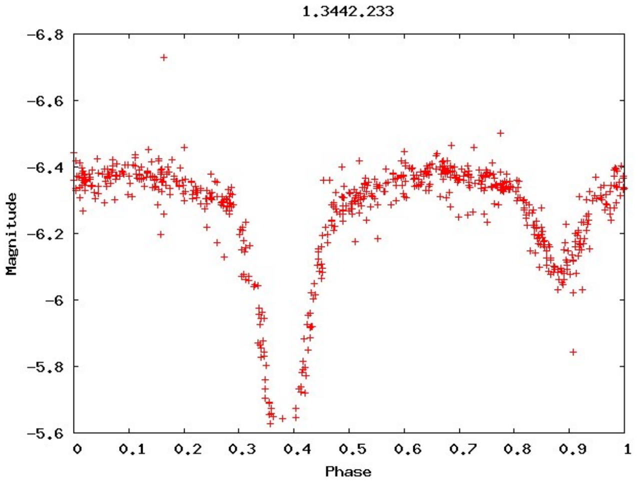
- Improve kernel

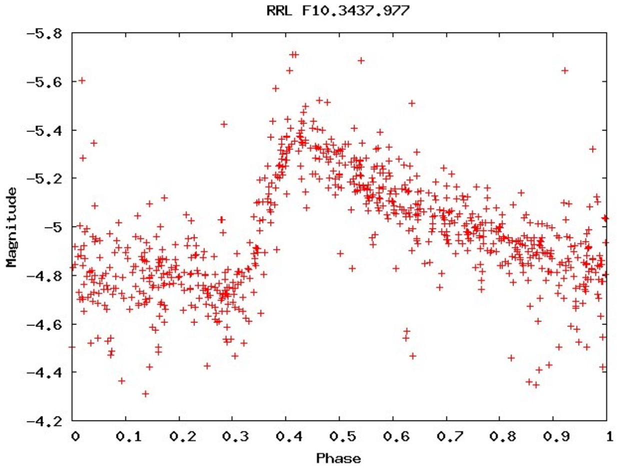
$$K(f,g) = \sum_{\tau} \exp\left(\sum_{i} (f_i - g_{i+\tau})^2 \right)$$

- Results: 14000 new periodic variable stars, doubling the number of known variables in the LMC

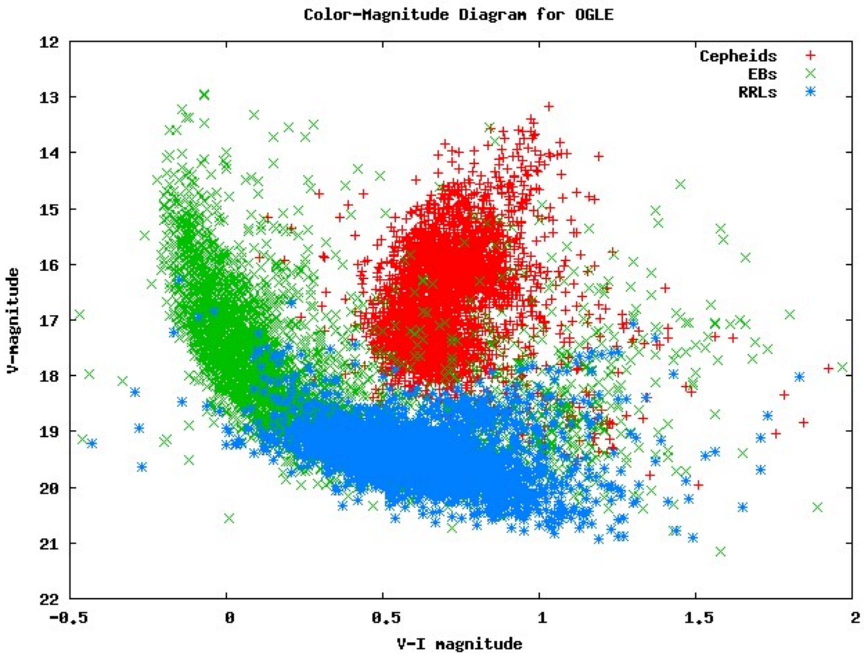
typical light curves

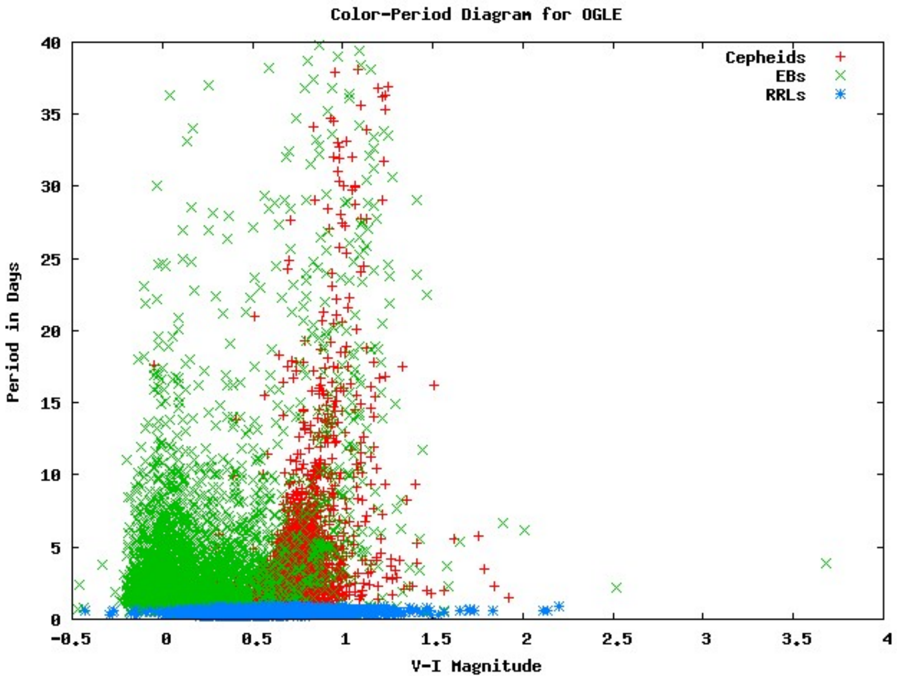

CEPH F10.4280.1392


1.3442.233


RRL F10.3437.977

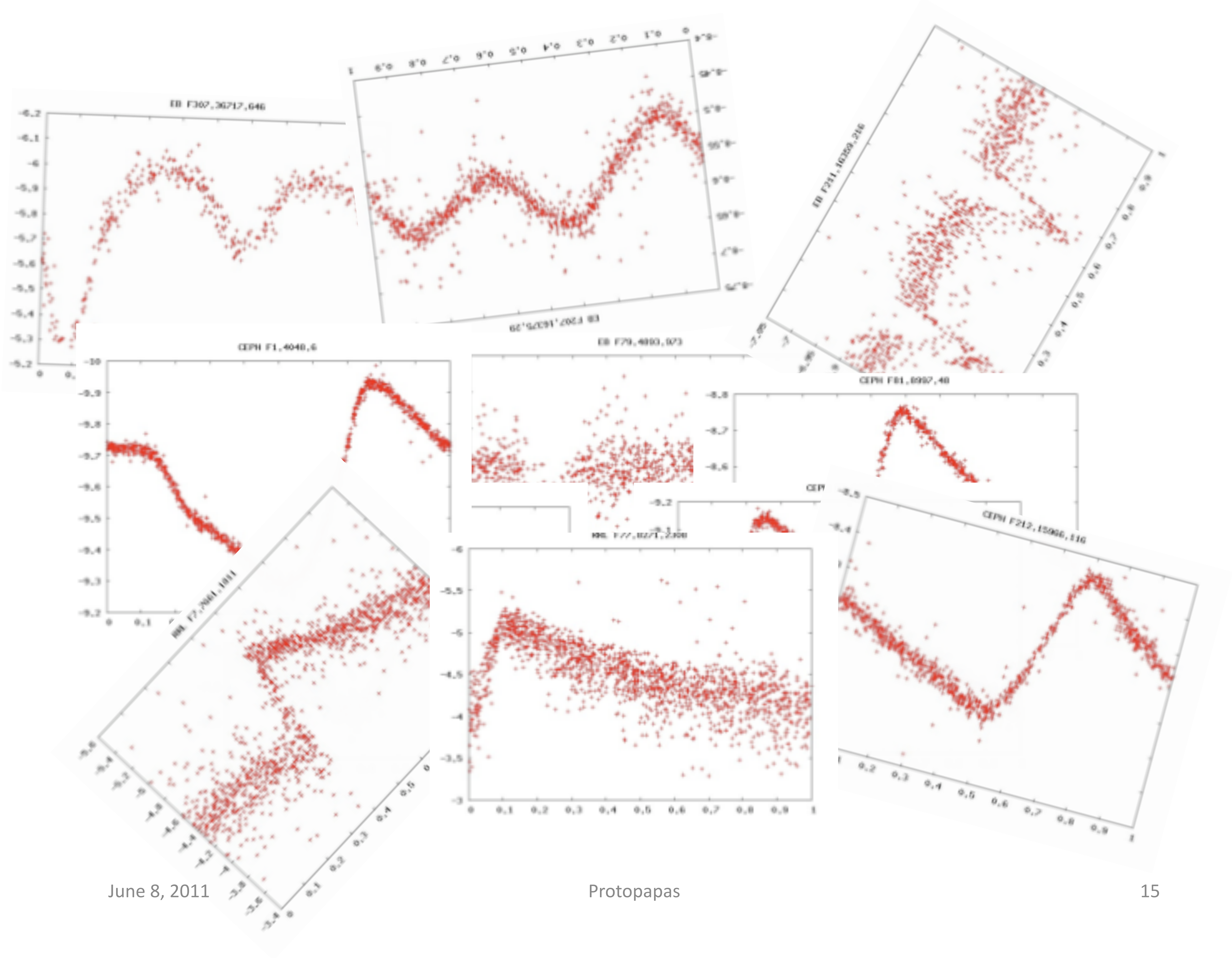
Other features: brightness, color, period





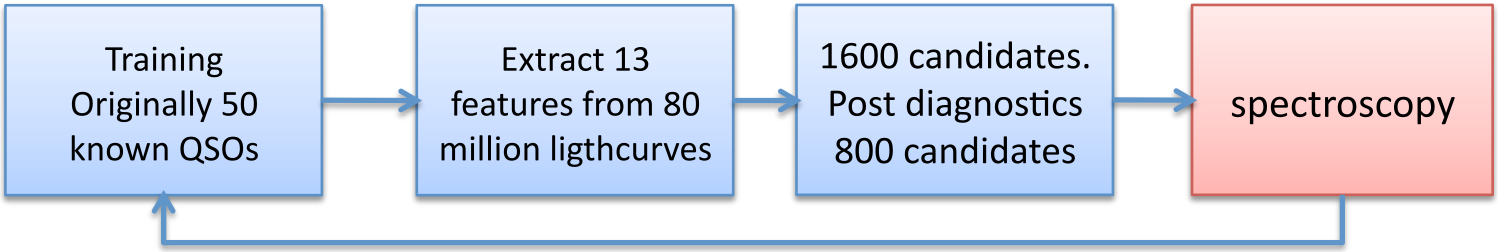
- Combine the kernels in linearly and optimized for coefficients

$$K = \alpha_{CC} K_{cc} + \alpha_{COL} K_{COL} + \alpha_P K_P + \alpha_M K_M$$


EB F307,36717,646
EB F207,16375,29
EB F211,16259,216
CEPH F1,4048,6
EB F79,4083,973
CEPH F81,8997,48
RRL F7,7661,1011
RRL F77,8271,2388
CEPH F212,15986,116

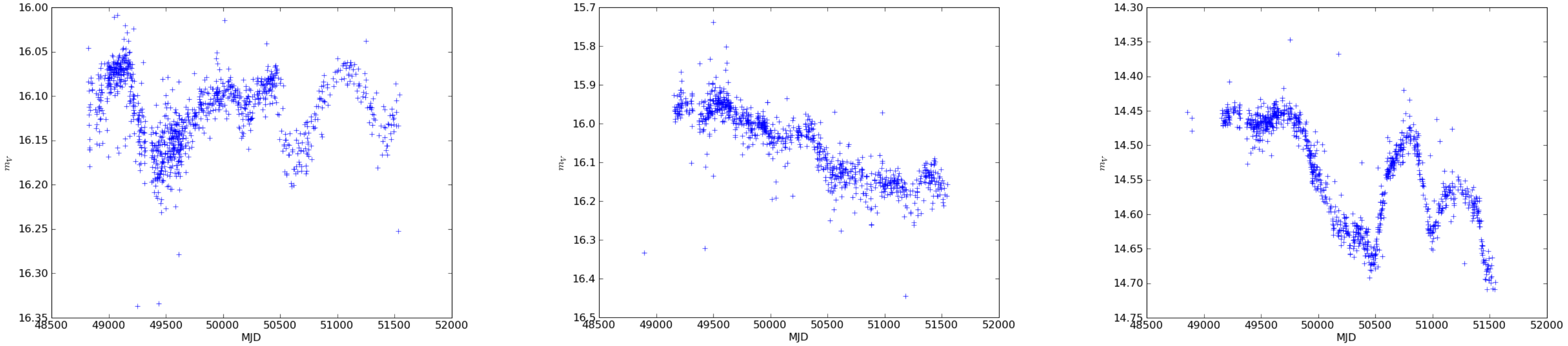
Classification (non-periodic)

- Use support vector machine (SVM) using many extracted features from the time series (Kim et al 2011) for QSO classification.

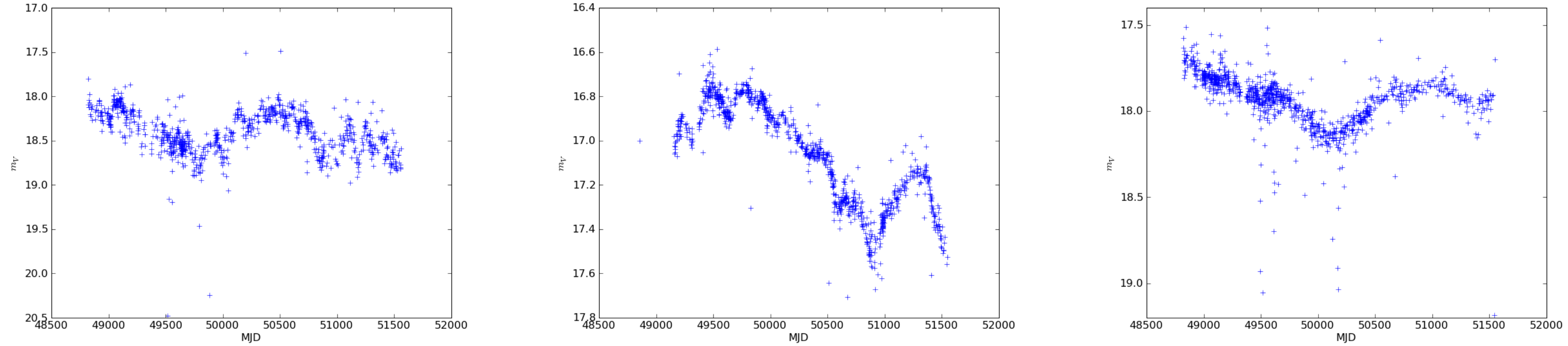


- Previous time series work shows very low efficiency
  - Geha et al. 2003 selected total ~2,500 QSOs candidates
  - Manually removed 2,140 false positives and observed only 260 targets spectroscopically47 of them were confirmed as QSOs; ~2%
- False positives are Be stars
- Training set:
  - Use 50 known MACHO QSOs to train (again using SVN)
  - 58 QSOs
  - 128 Be stars
  - 582 Microlensings
  - 193 Eclipsing Binaries
  - 288 RR Lyraes
  - 73 Cepheids
  - 365 Long period variables
  - 4,288 Non-variable stars

Example Light-Curves of MACHO QSOs and Be Stars

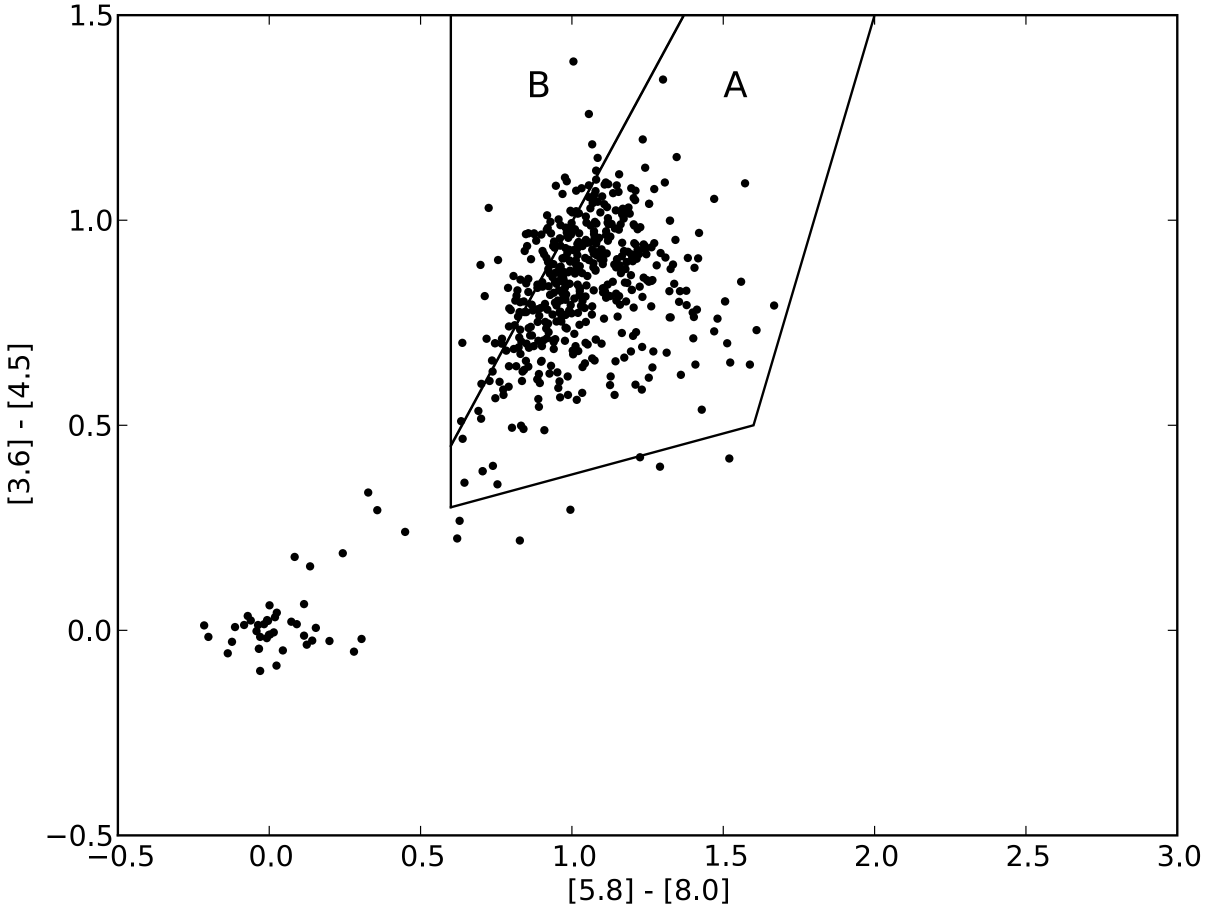


Be stars



QSOs

Stern et al. 2005



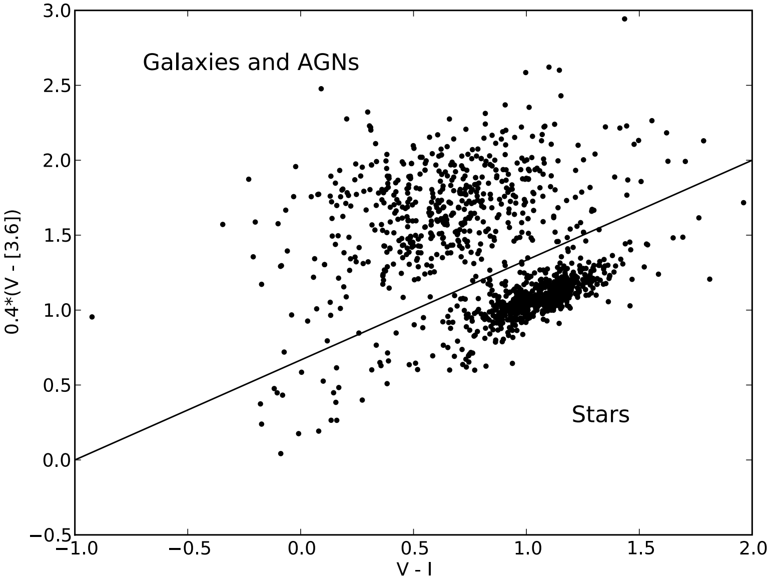
Photometric redshift

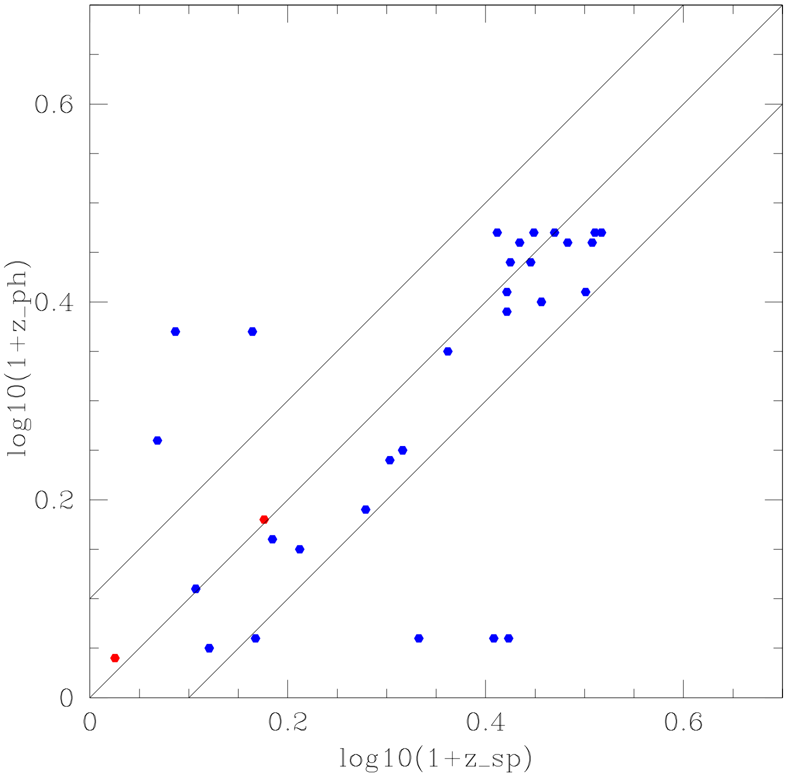
MACHO UBVI catalog (Zaritsky et al. 2004)

2MASS catalog

Separated stars and ‘Galaxies and AGNs using criteria from Eisenhardt et al. (2004) and Rowan-Robinson et al. (2005).

Photometric redshifts (Rowan-Robinson 2003)

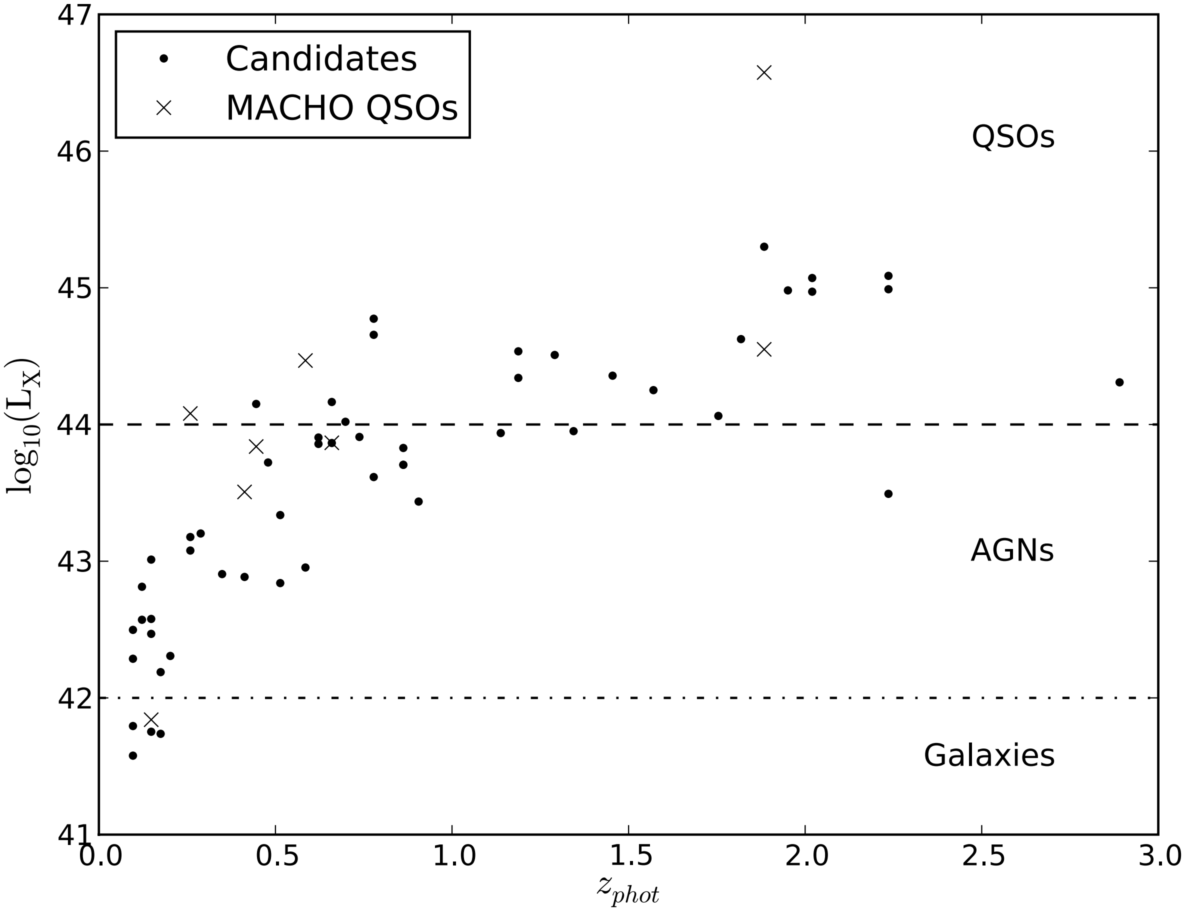




X-ray luminosity.

Cross-match with Chandra and XMM

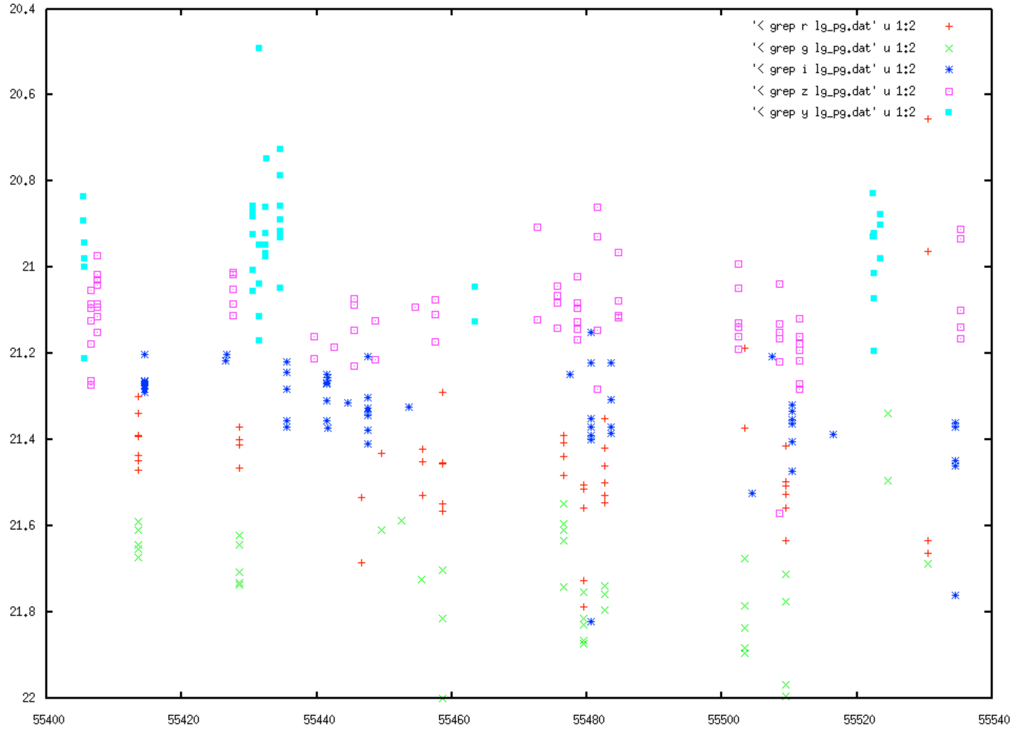
Use photo-z from above



- 650 final candidates that did not fail any of the above tests.

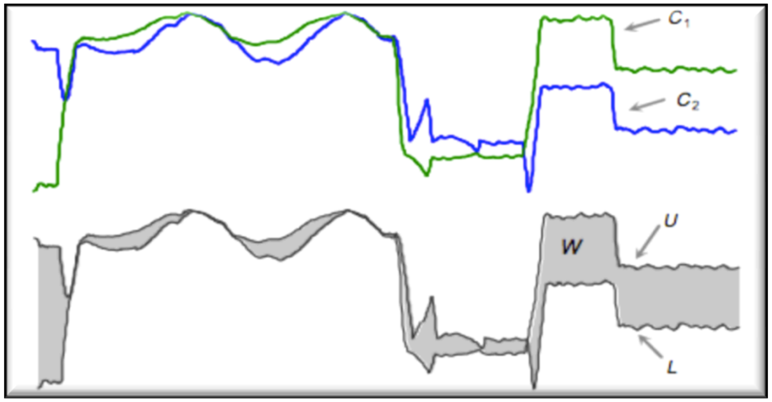
Spectroscopic validation

- Proposed time on IMACS to follow candidates
- Apply model to PanSTARRs MD field

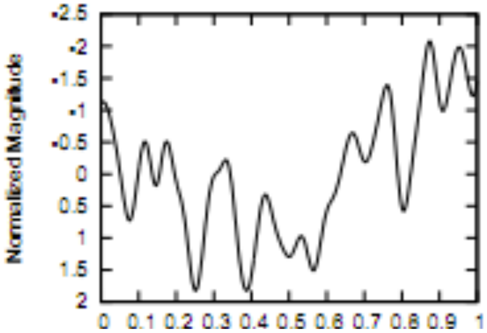


Anomaly detection

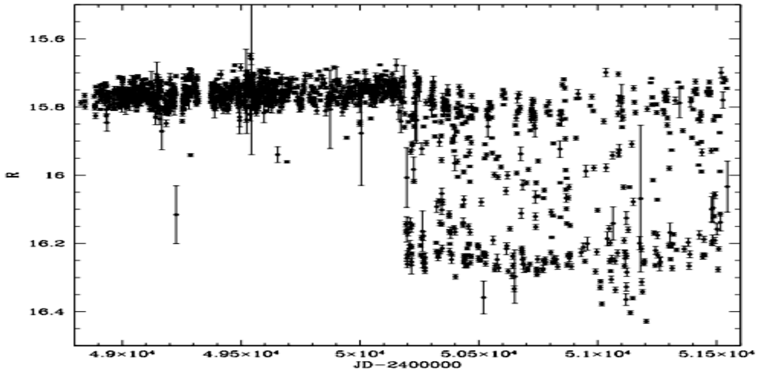
- Protopapas et al 2006 used Euclidean distances of the shape to find anomalous lightcurves in an ensemble of ligthcurves. Use universal phasing to deal with phase invariance of periodic variables



- U. Rebbapragada 2009 extended k-means to pk-means for periodic lightcurves where we need phase invariance



- Majidi 2011 (submitted) using active learning methods

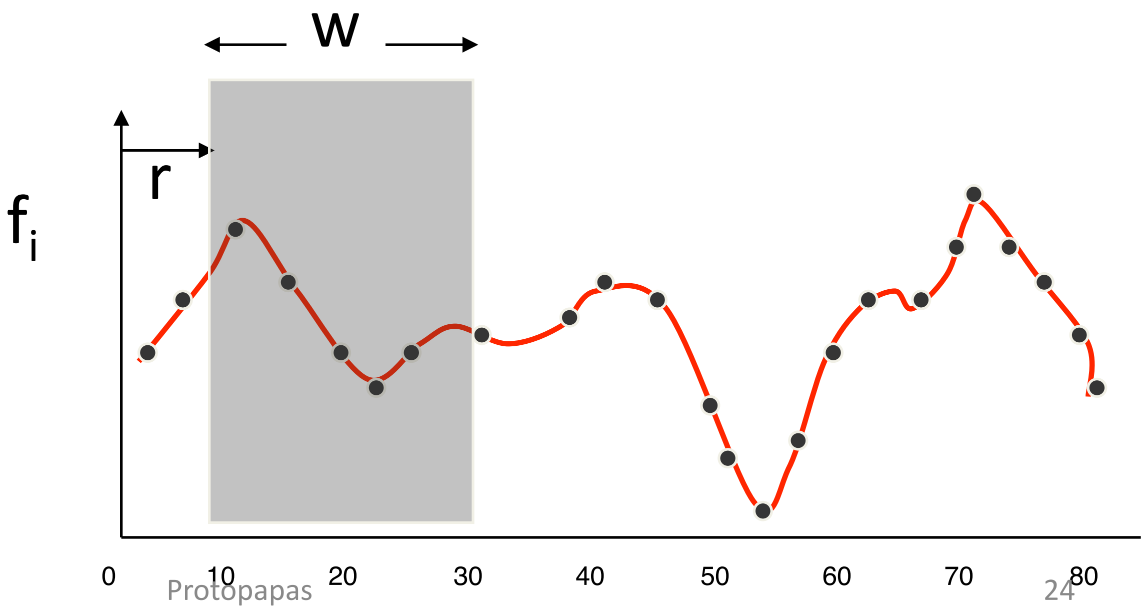


Event detection – rank statistics

Idea: Given a time series find the region [sub sequence of measurements] that are not consistent with noise.

Let f_i be the value at time t_i calculate the statistics

$$S(r,w) = \sum_{i=r}^{r+w} f_i$$



Need p-value. Build the distribution and estimate $P(S(r,w) > S_0)$

We could simulate null time series with the same noise characteristics and build distribution of P(S).

Problem:

1) Too expensive for large datasets

2) Modeling the noise is not as simple

Solution 1:

Reshuffle the sequence (the noise model is taken care).
Do this many times and each time calculate all S(r,w). Choose S_0 and thus $P(S(r,w) > S_0)$. [Or use bootstrapping method]

No good. We need to this for each time series since the noise could be very different (data are taken different times, different filters etc)

We need something that eliminates the need for modeling the noise.

Solution 2: Rank the y-values (flux)

Consider a time series T with n points. We then create a time series T_R by converting each point in T into its ranked value. Thus, the highest point in T_R will be n, the second highest n − 1, third highest n − 2, etc.

$$Q(r,w) = \sum_{i=r}^{r+w} R_i$$

where R's are the rank values and Q is the new statistics (equivalent to S)

Advantage: All time series of n points in rank space have exactly the same $P(Q_0)$. We need to calculate this only once. Then for real data we calculate Q(r,w) and we know the p-value for each point.

To calculate this simply select w numbers out of 1..n and calculate Q. From that build the distribution $P(Q_0)$. This probability depends on n and w but not on the location: P(Q;n,w)

Better. **This distribution can be found analytically**

It is the same problem as finding the number of partitions of S with w distinct parts, each part between 1 and n, inclusive. Consider the values

$e1 , e2 , . . . , e_w$.

If we can find all possible solutions to $0 < e1 < e2 < . . . < e_w \leq n$, we can simply multiply by w! (all possible permutations) and obtain our result.

This will be the same as the following: Subtract 1 from the smallest part, 2 from the second, etc. to get

$0 \leq e1 - 1 \leq e2 - 2 \leq \dots \leq e_w - w \leq n - w$.

We have subtracted a total of

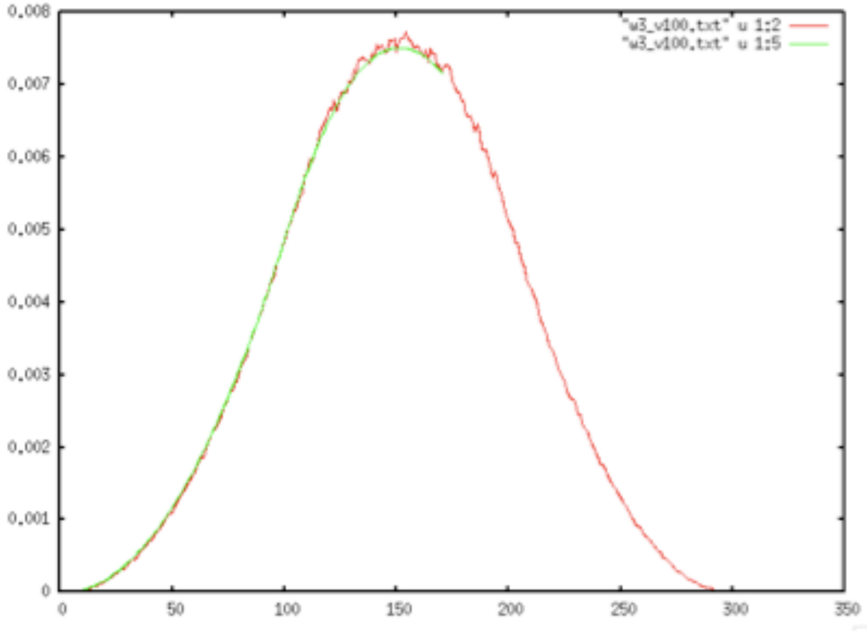
$1 + 2 + . . . + w = w(w + 1)/2$, so we are now looking for the number of partitions of $- w(w + 1)/2$ with at most w parts, and with largest part at most $n - w$.

To find the number of partitions, we consider a specific application of q-binomial coefficients.

A similar problem is finding the number of distinct partitions of k elements which fit inside an m by n rectangle. This can be found by finding the coefficient of q_k in the q-binomial coefficient

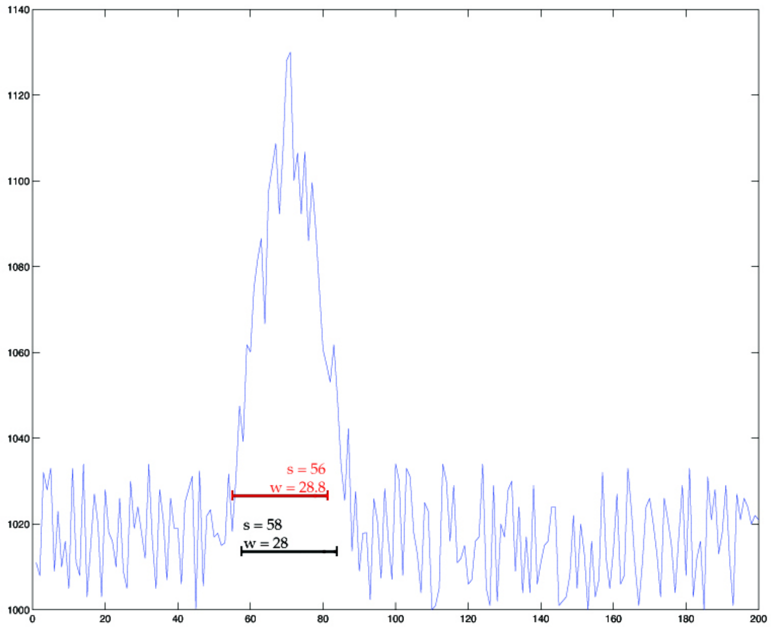
An iterative solution does exist. To do so, we create the following recurrence

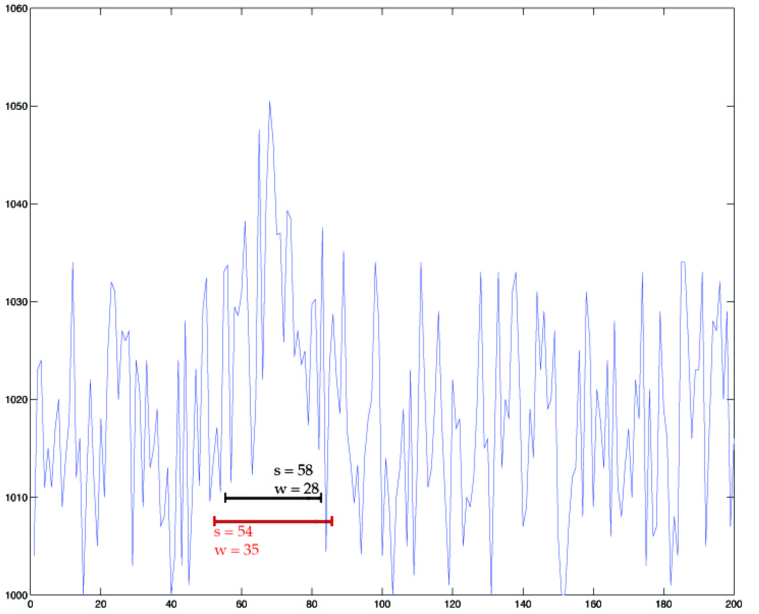
$$P(Q;n,w) = \sum_{j=1}^{n} P(Q-j;w-1,j-1)$$



probability distribution build “red” by simulation and “green” with the analytical formula

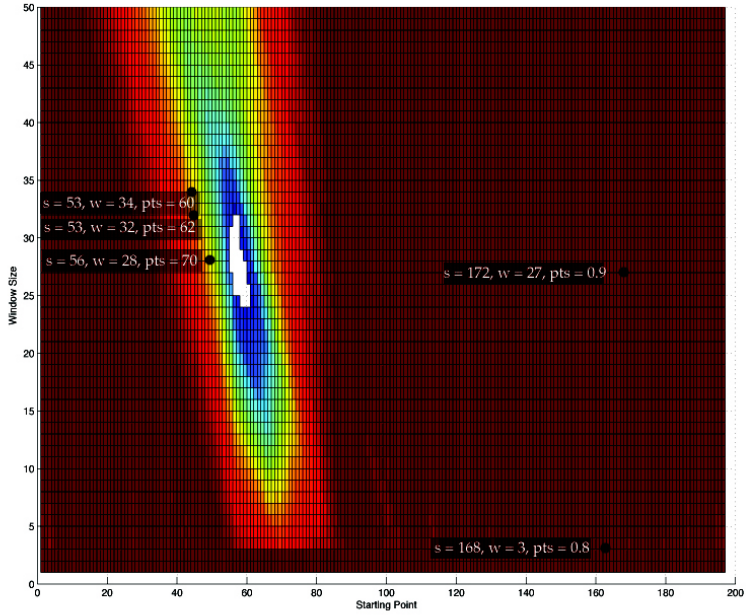
P(Q(r,w);w,n) for synthetic lighcurves



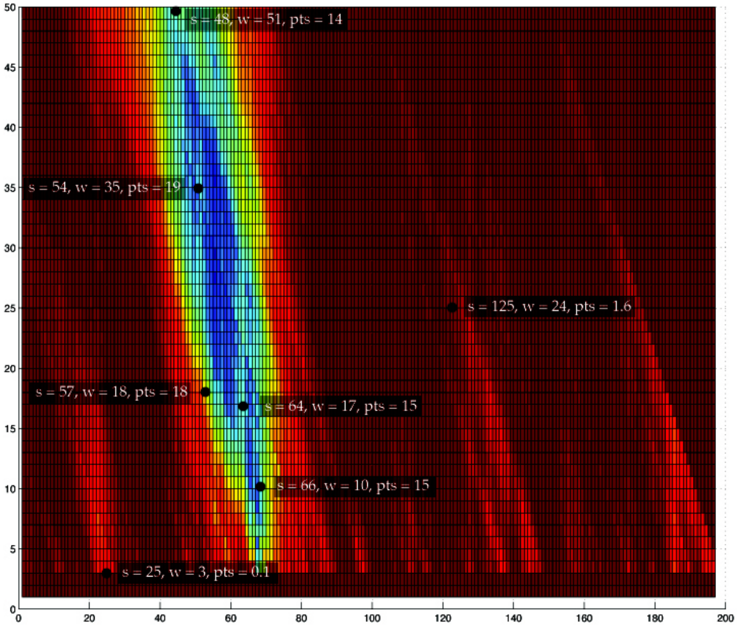


Color Legend: Black (Goal, theoretical) Red (Actual, found)

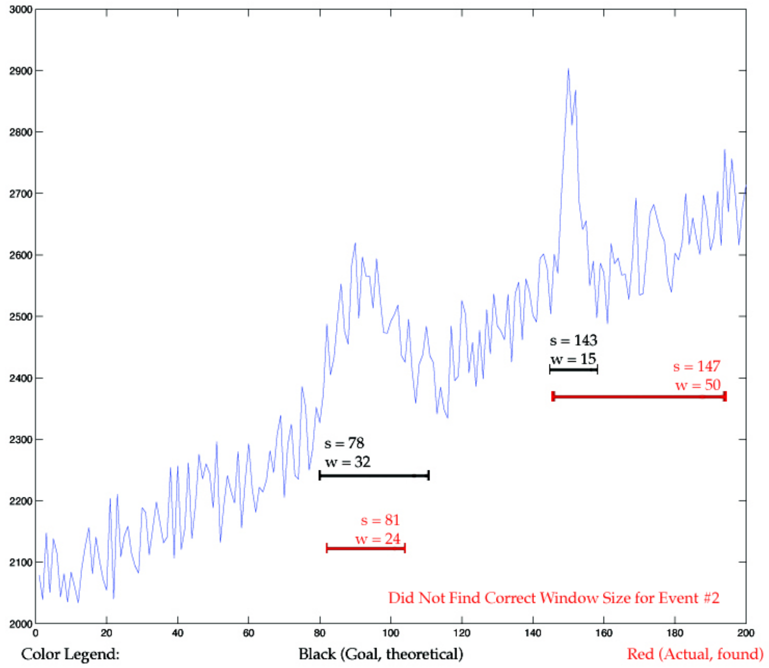
Color Legend: Black (Goal, theoretical) Red (Actual, found)

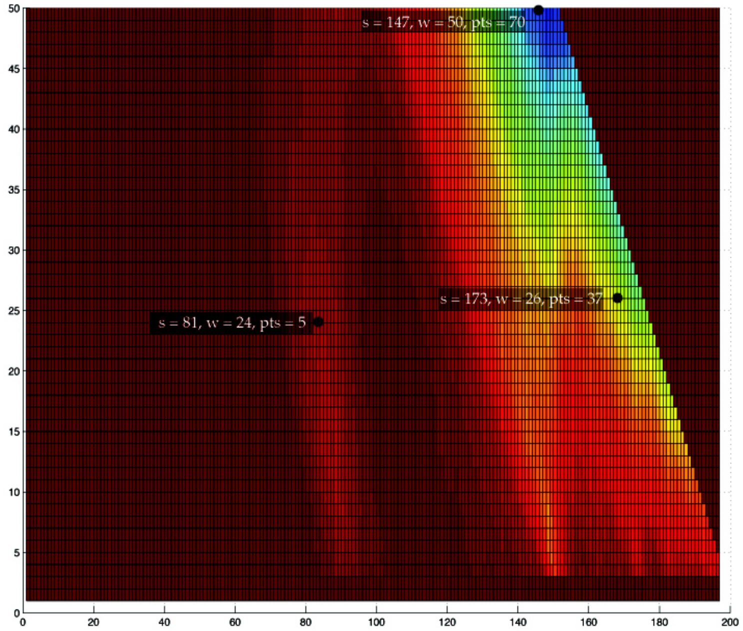


Protopapas

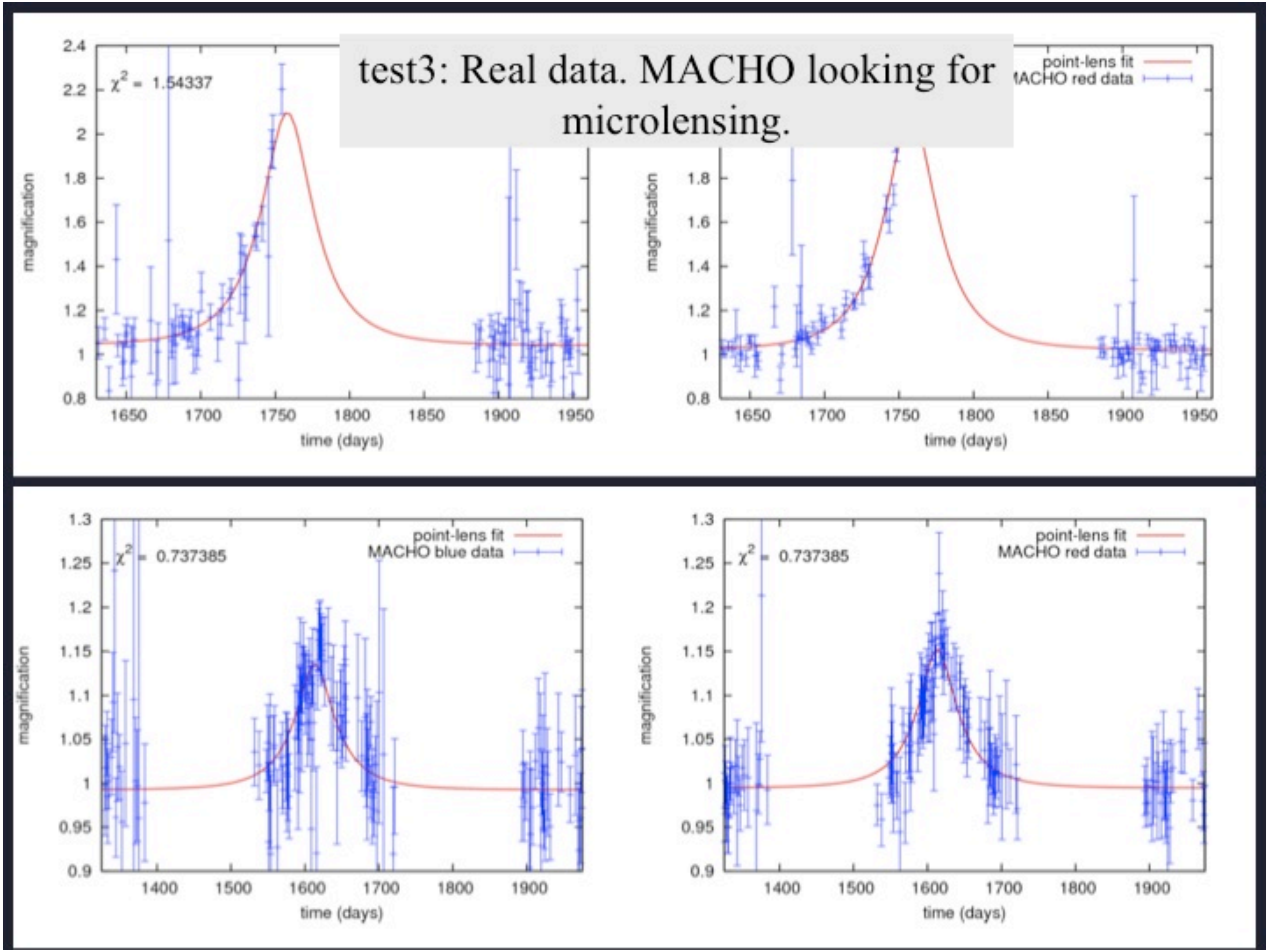


Note that the surface of the p-value is smooth with respect to r and w.
We therefore used optimization method to determine the solutions.



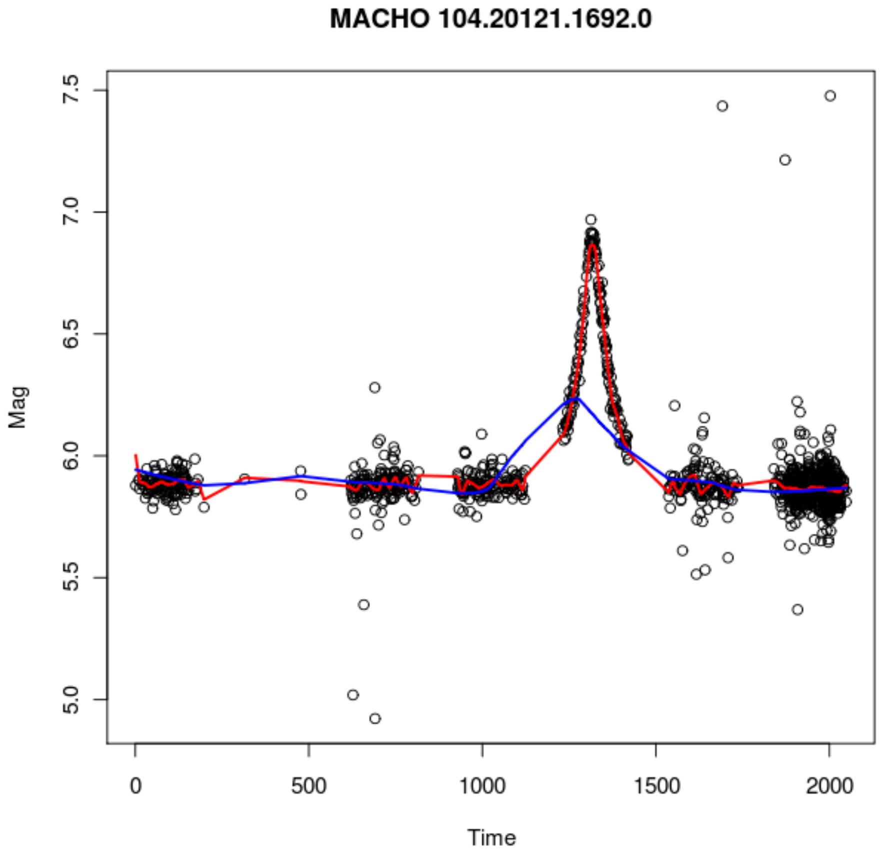


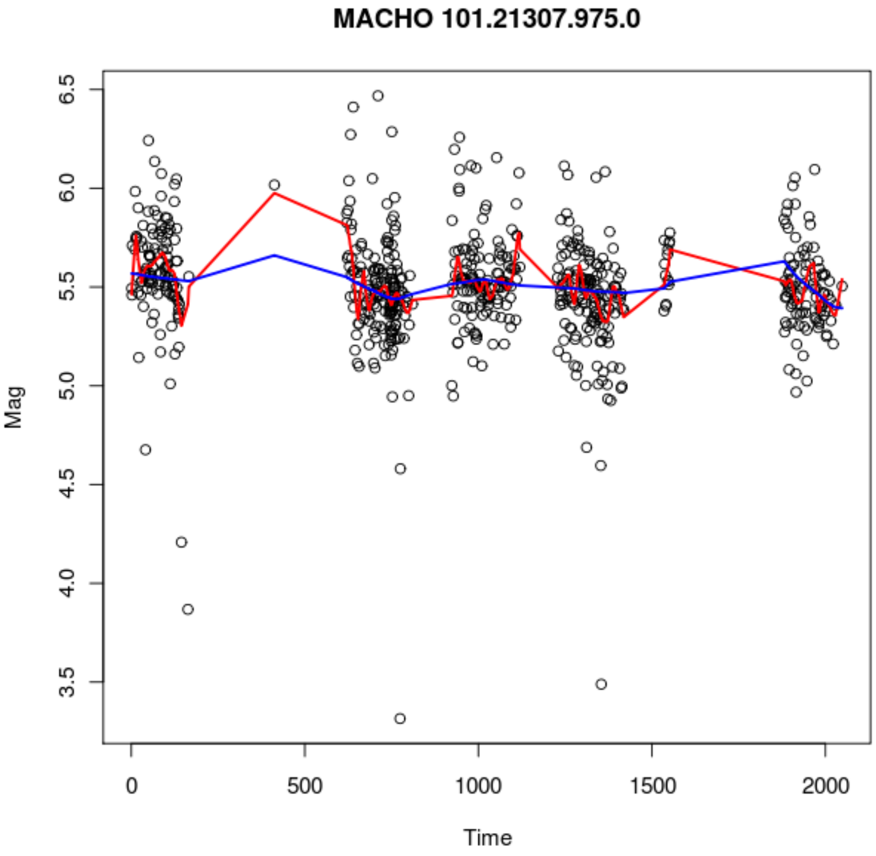
test3: Real data. MACHO looking for microlensing.



Event detection: wavelets

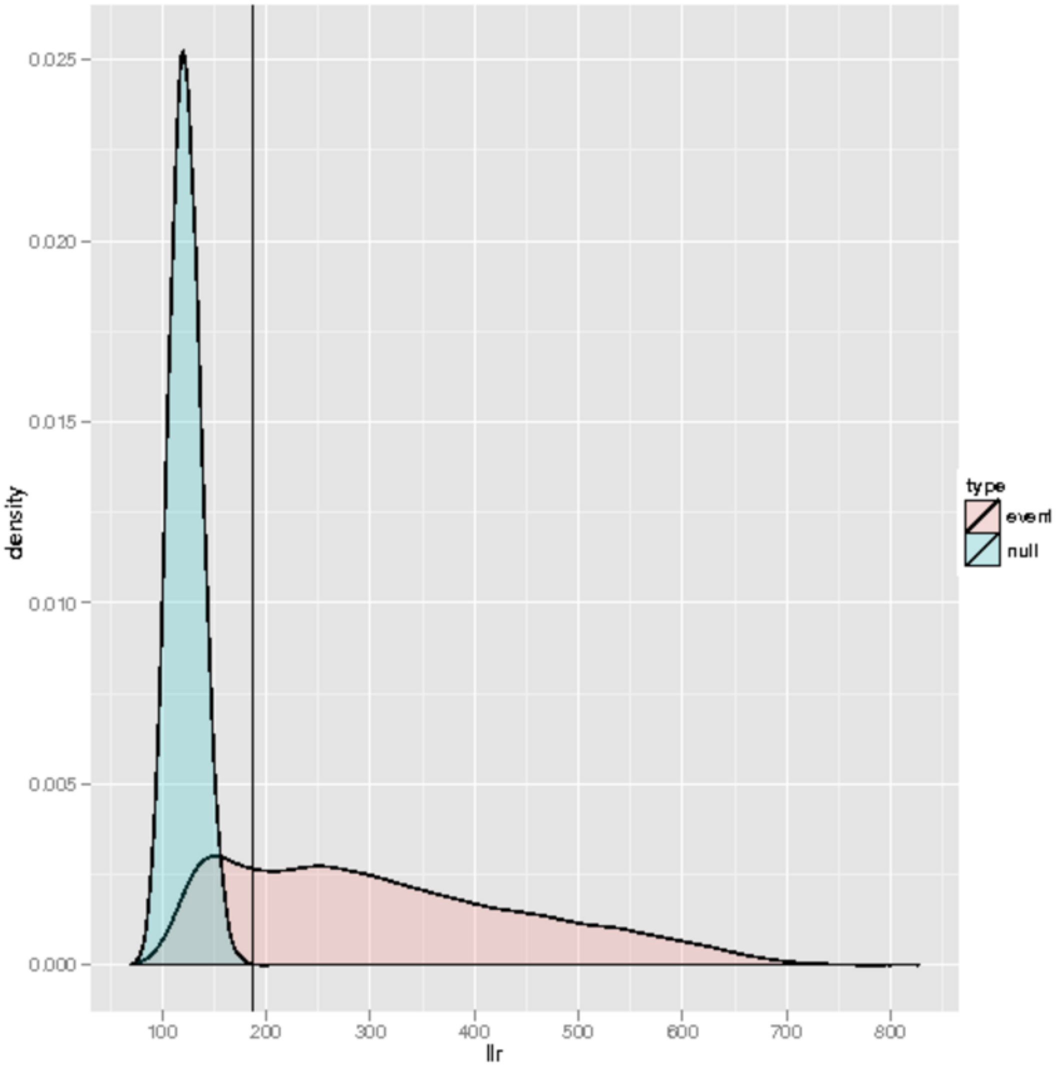
- Fit wavelets to the timeseries and compare the low frequency to the high frequency coefficients.
- We measure the difference in maximized log-likelihoods (low and high components) $=2*(l_{high}-l_{low})$



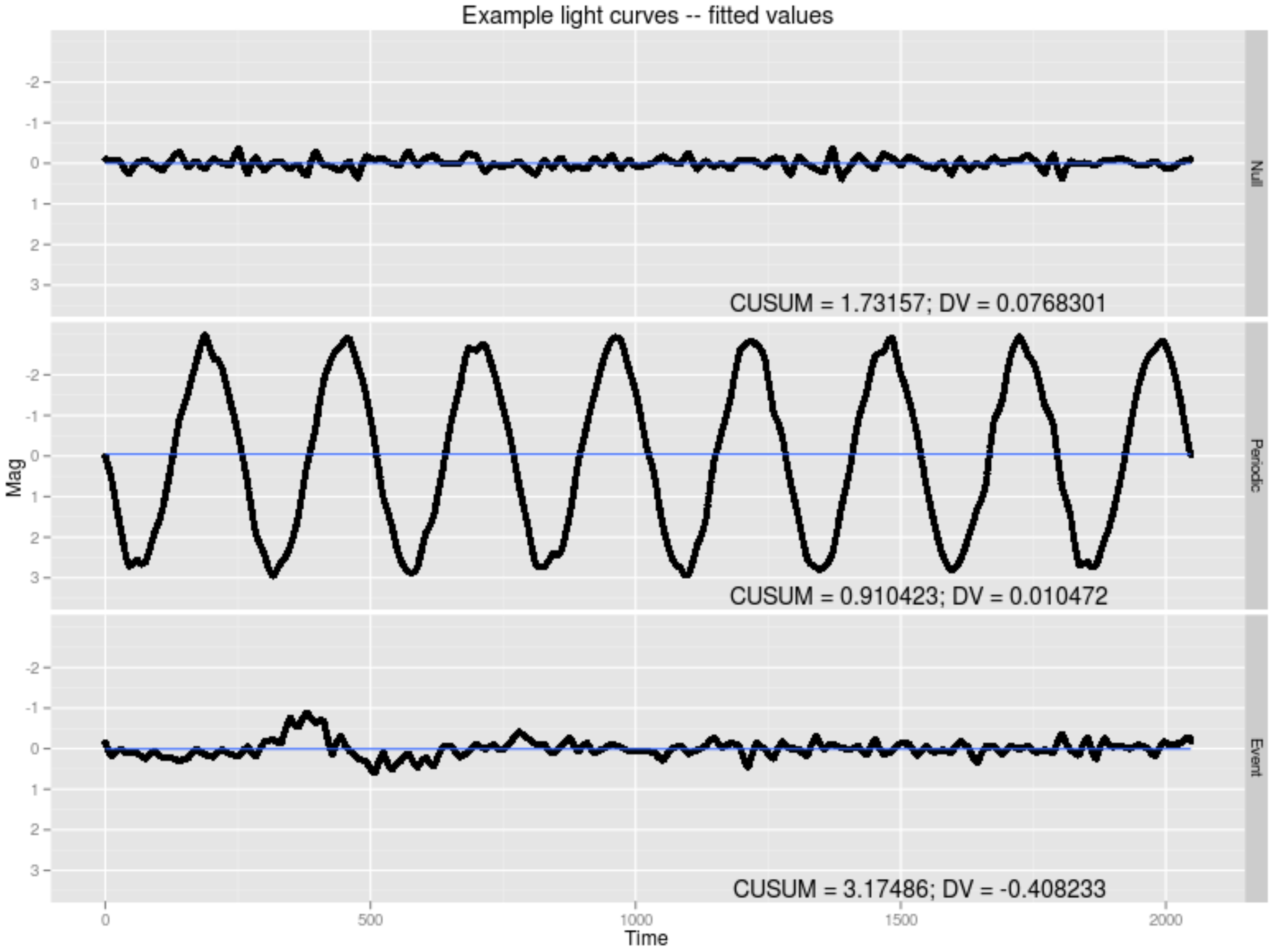


False positives

- To assess how well this statistic perform, we simulated 50,000 events from physics-based model and 50,000 null time series
- We use False Recovery Rate (FDR) of 10^{-4} (Benjamini-Hochberg approach)



Isolated vs diffused events



Use two features to discriminate between diffuse and isolated variability

1. CUSUM statistics: Let $\{\mathrm{z_t}\}$ be the normalized fitted values (without the trend component). We define:

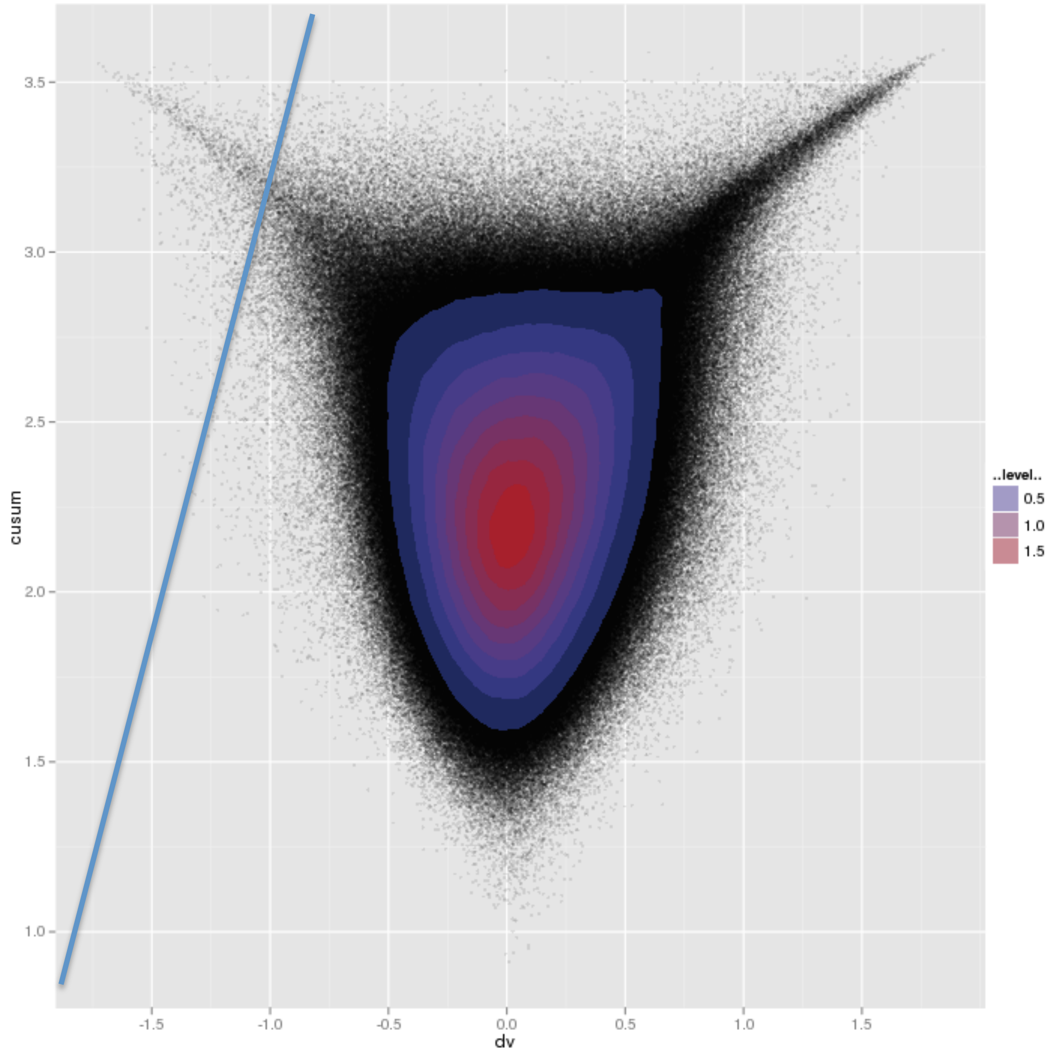
$$S_t = \sum_{k=1}^{t} \left(z_k^2 - 1 \right)$$

$$CUSUM = \max_t S_t - \min_t S_t$$

2. The second feature tries to cacpture the deviation from symmetric variation. Let z_{med} is the median of $\mathrm{z_t}$

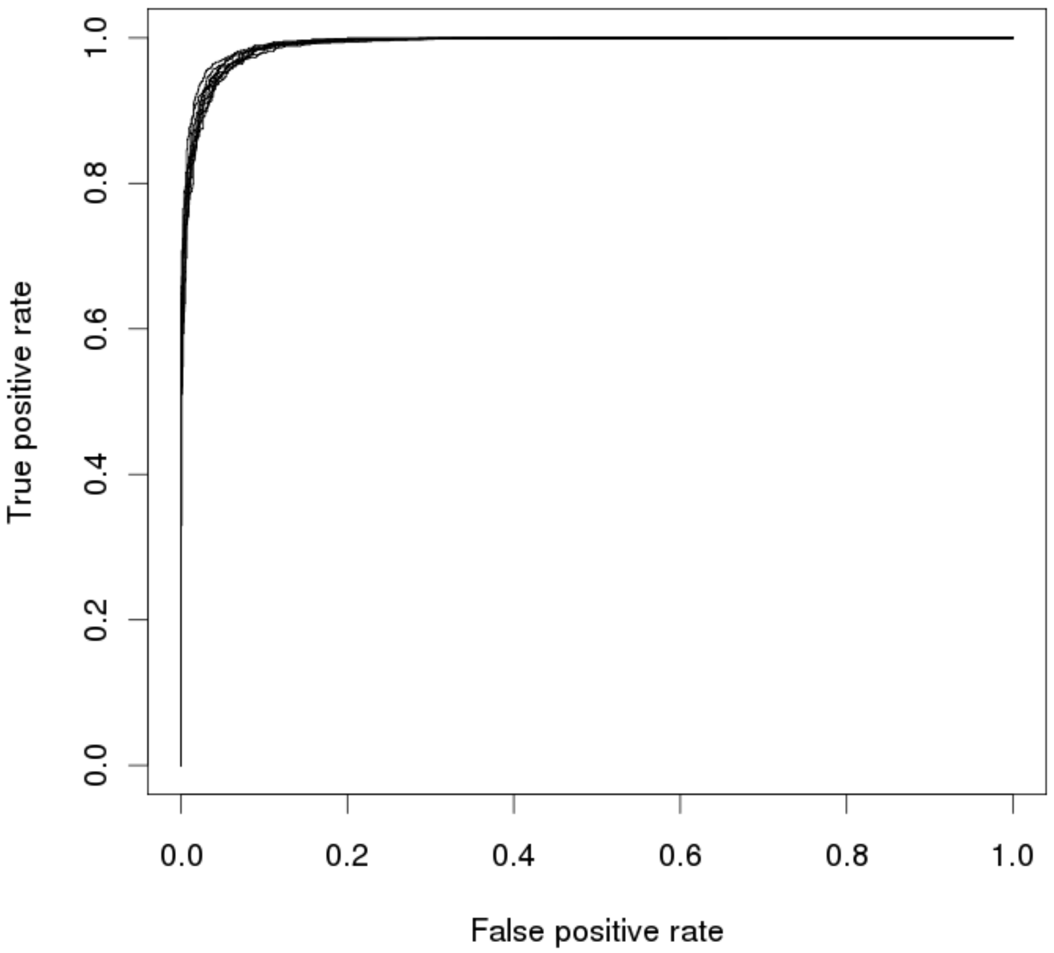
$$DV = \frac{1}{\#\{t : z_t > z_{med}\}} \sum_{t:z_t > z_{med}} z_t^2 - \frac{1}{\#\{t : z_t < z_{med}\}} \sum_{t:z_t < z_{med}} z_t^2$$

Distribution of features on MACHO data: TRAINING



Training

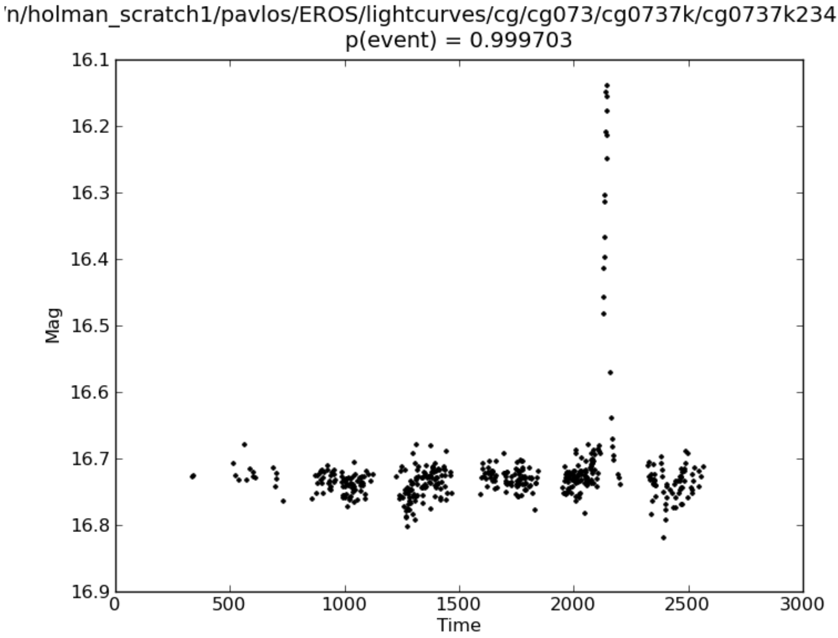
- Cross validation on MACHO training data

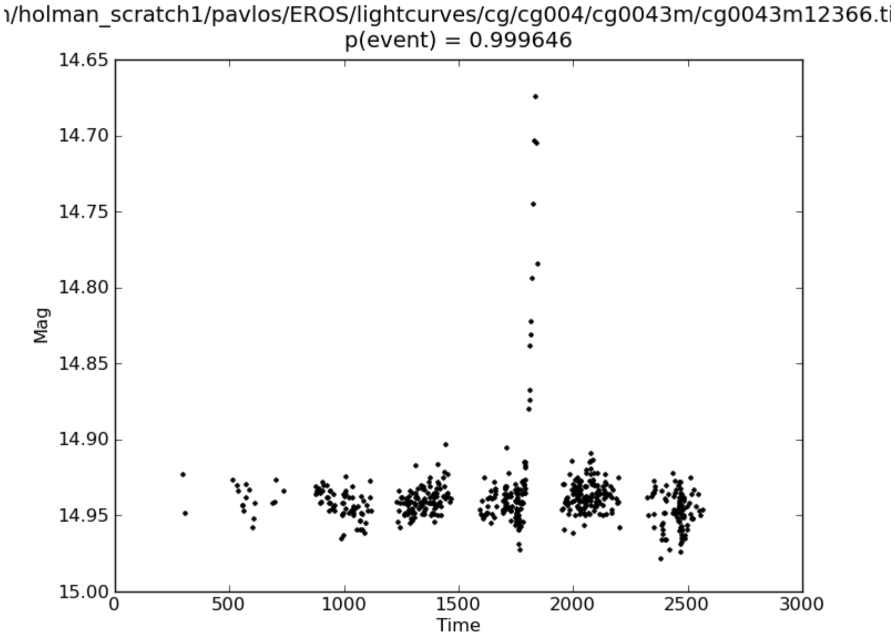


Results

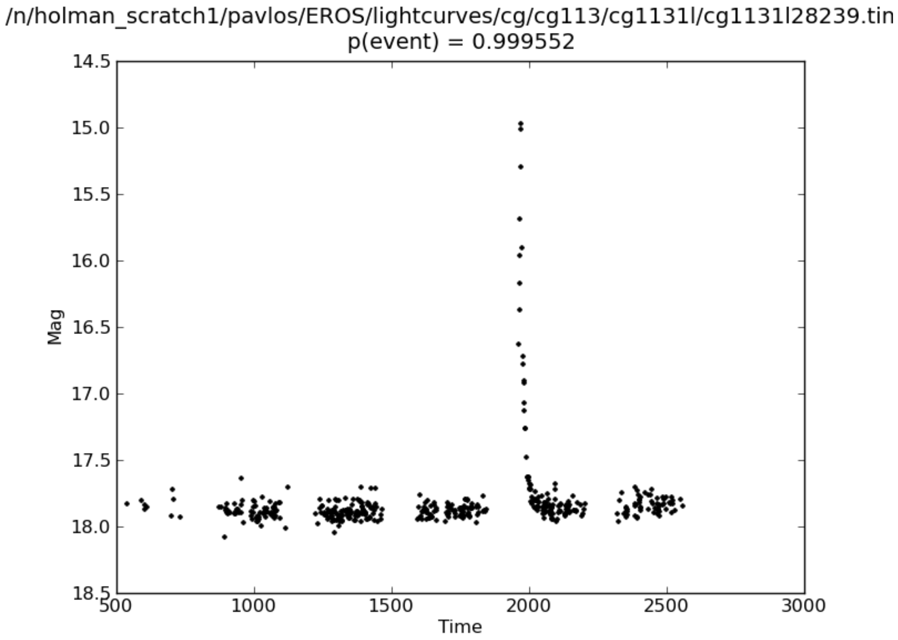
- Started with 87.2 million candidate lightcurves in EROS db.
- Reduce it to 500,000 after likelihood-ratio test
- Approximately 5000 are likely isolated events based on the isolated event stage
- Currently pursuing scientific follow-up on top candidates

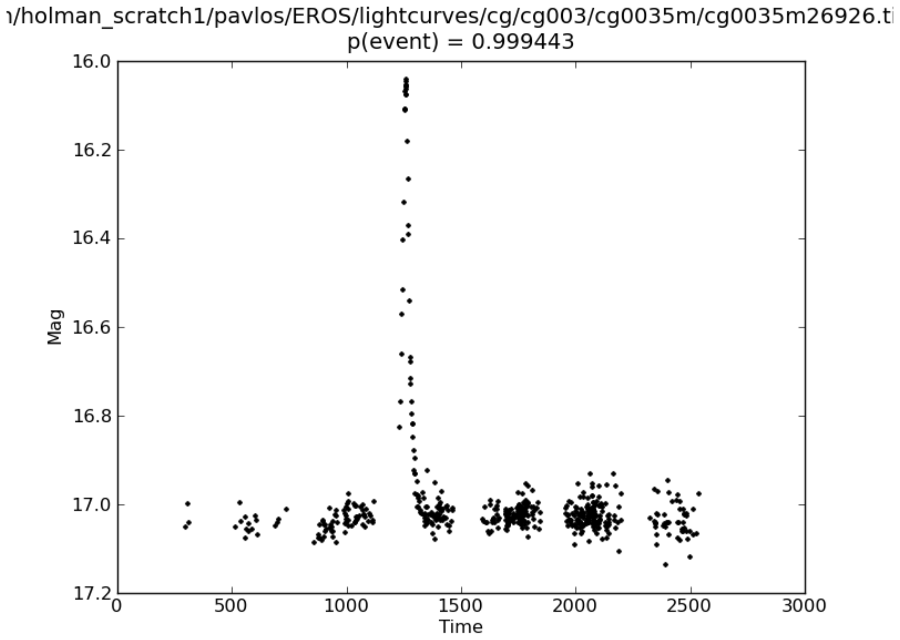
Examples top 6



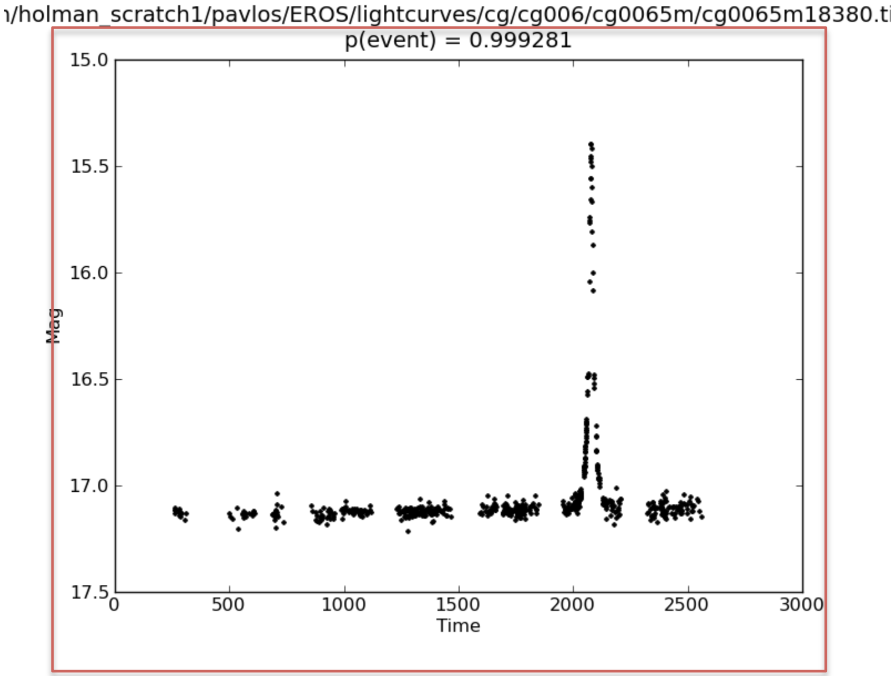


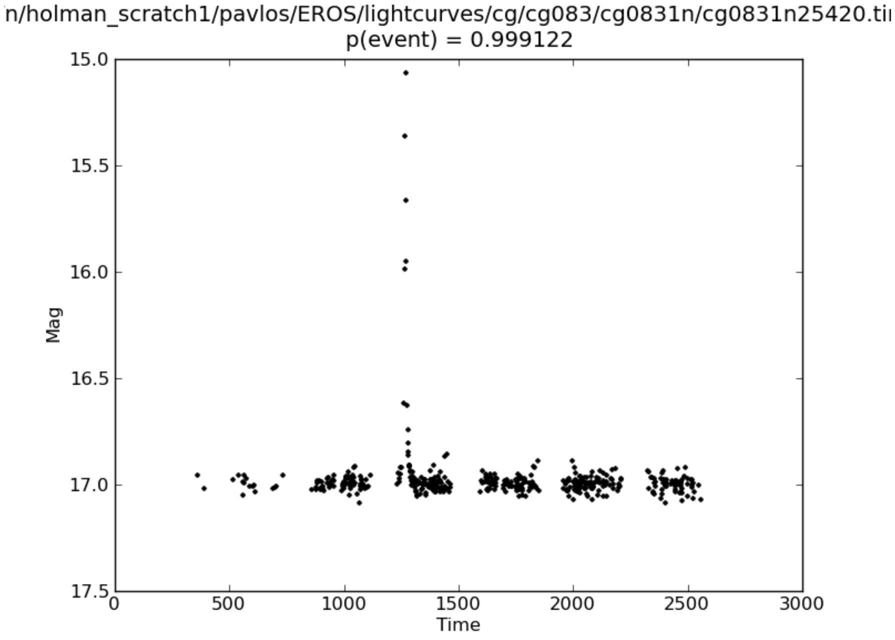
Examples top 6





Examples top 6





Period finding

Lesson learned during periodic variables is that period estimation is not so easy if we require high accuracy

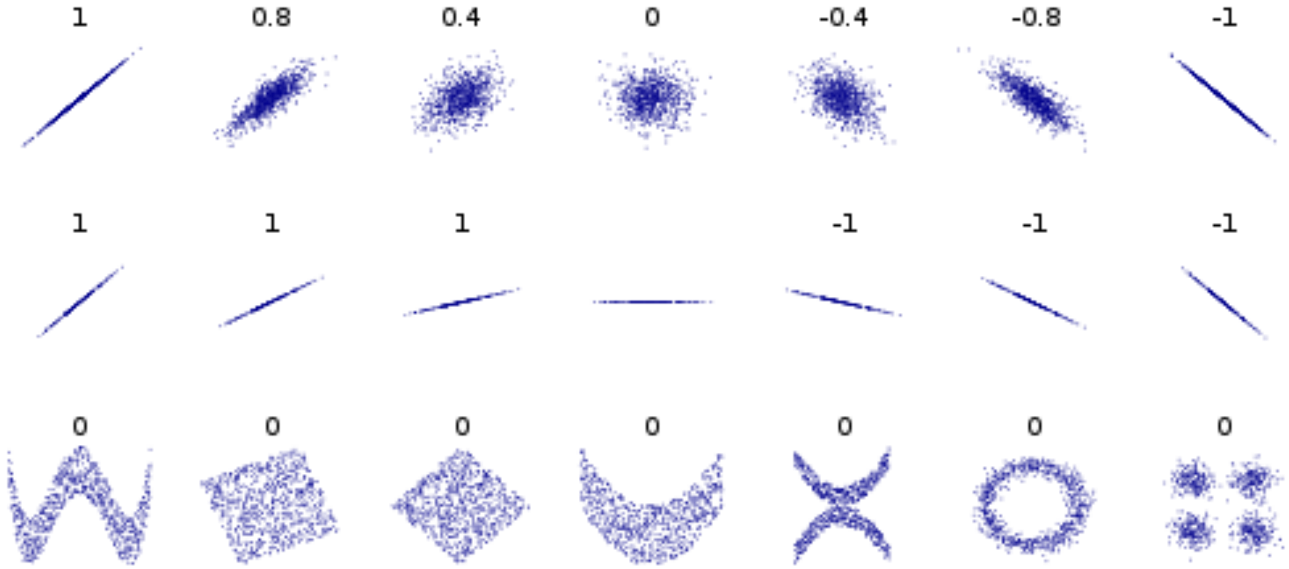
- Gaussian processes (fit multivariate Gaussians):
  - We use the following kernel

$$K(t_1,t_2) \propto \exp\left\{-\frac{2\sin^2(w\pi(t_1-t_2))}{l^2}\right\}$$

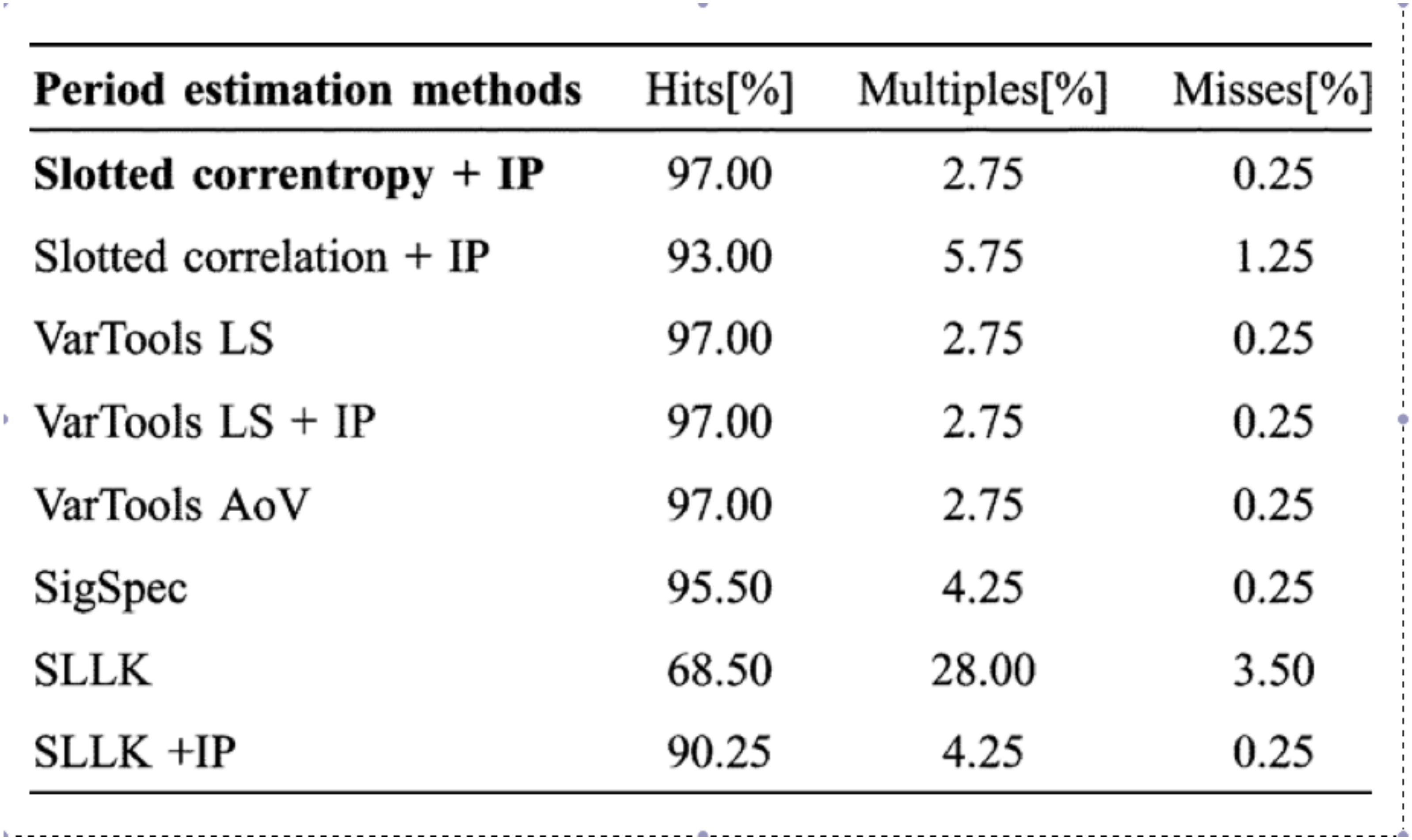
  - And use Bayesian approach to identify the hyper-parameters that maximize the marginal likelihood, where the “marginal” means that we integrate out the latent function that associated with the data. Wang et al 2011 (submitted to KDD)
- Correntropy. Instead of finding the peaks in the power spectrum of autocorrelation we use Correntropy
  - P. Huise et al 2011 (IEEE Signal Processing Letters)

$$V[m] \propto \sum_{n=m}^{N-1} \exp\left\{-\frac{\left|x_n - x_{n-m}\right|^2}{2\sigma^2}\right\}$$

Period finding



Period finding



Period estimation methods	Hits[%]	Multiples[%]	Misses[%]
Slotted correntropy + IP	97.00	2.75	0.25
Slotted correlation + IP	93.00	5.75	1.25
VarTools LS	97.00	2.75	0.25
VarTools LS + IP	97.00	2.75	0.25
VarTools AoV	97.00	2.75	0.25
SigSpec	95.50	4.25	0.25
SLLK	68.50	28.00	3.50
SLLK +IP	90.25	4.25	0.25

Period finding

- New extensions
  - Try periodic kernels of the type

$$V[m] \propto \sum_{n=m}^{N-1} \exp\left\{ -\frac{\left(\left| t_n - t_{n-m} \right| - kP_c \right)^2}{2\sigma^2} \right\}$$

  - Use spatio-temporal kernel (here temporal kernel deals with variable sampling)

$$V[m] \propto \sum_{n=m}^{N-1} \exp\left\{ -\frac{\left| x_n - x_{n-m} \right|^2}{2\sigma_x{}^2} \right\} \sum_{n=m}^{N-1} \exp\left\{ -\frac{\left| t_n - t_{n-m} \right|^2}{2\sigma_t{}^2} \right\}$$

Autoregressive models

- Assume linear response of the variability nd model it using autoregressive process.
- Check partial autocorrelation of lightcurve and determine the scale n of AR(n).
- Then determine the

$$\text{AR}(1)$$

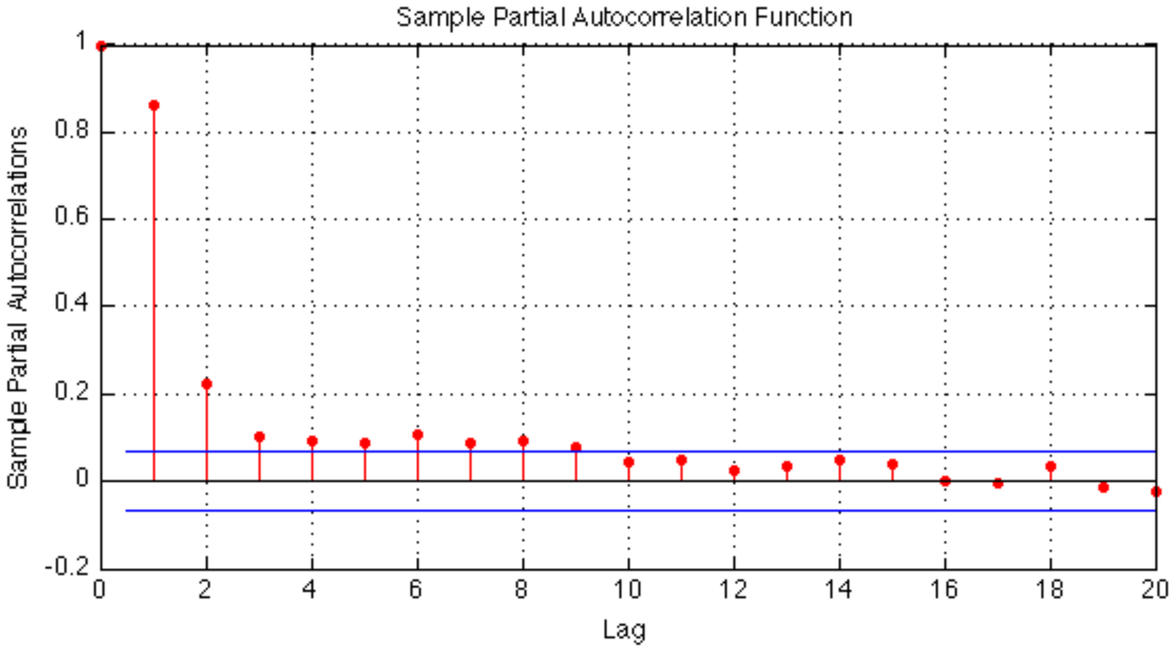
$$X_t = \alpha X_{t-1} + \varepsilon_t \quad \text{where } \alpha < 1$$

$$AR(2)$$

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$$

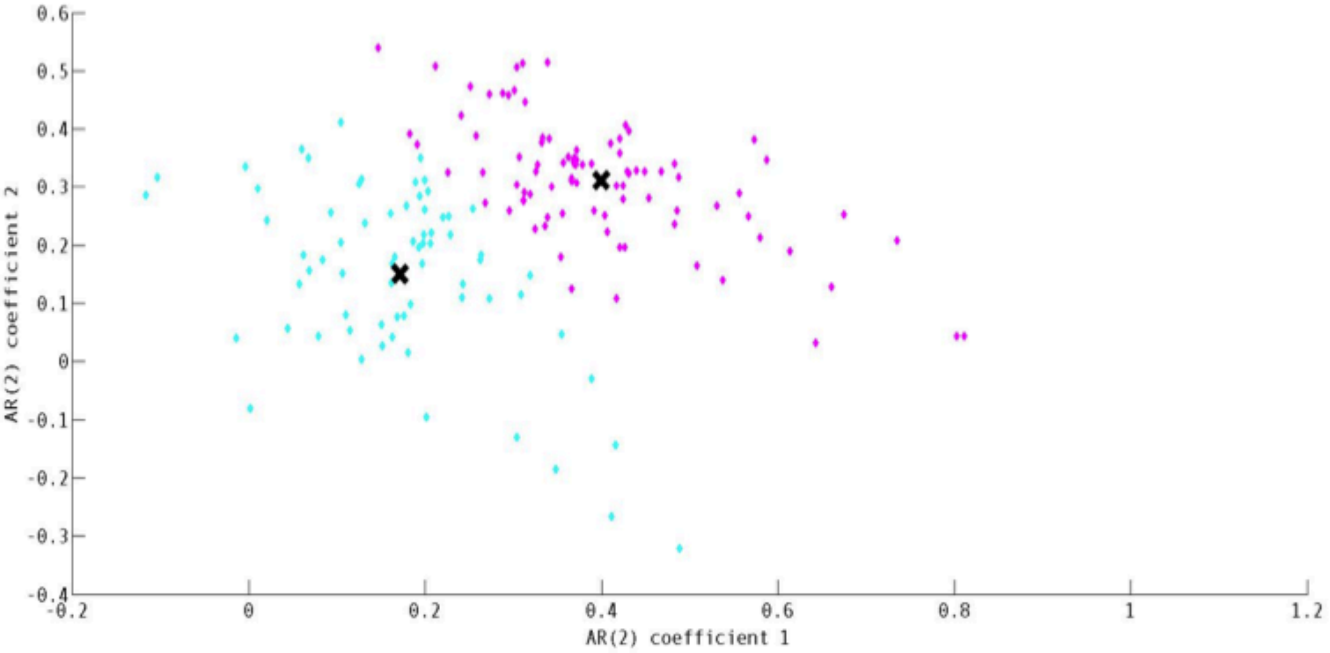
Autoregressive models

- Try it to a sample of AGNs, periodic variables and others
- We found that most AGNs have AR(2) (implication of accretion disk instabilities).



Autoregressive models

- Unfortunately Be stars have the similar length AR(2) but there is hope. HOT OUT OF THE OVEN



Summary

- Time domain astronomy is happening now.
- We should be able to handle the new challenges.
- Serendipitous discoveries are not possible so we need statistics/machine learning
- Need cross disciplinary work
- A lot of exciting new science will come out

THE SUPER CLASSIFICATOR SHOULD BE DOABLE NOW