Where statistical methods can help with Transients classification from surveys

Ashish Mahabal, Caltech

Mini-workshop on Computational AstroStatistics

25 Aug 2010

George Djorgovski, Ciro Donalek, Andrew Drake, Matthew Graham,
Roy Williams (Caltech); Baback Moghaddam, Michael Turmon (JPL); …

- Mainly optical based but extends to other wavelengths too (e.g. colors are like flux ratios)
  - Also context from other wavelengths used

- Mainly transients, but extends to archival information as well

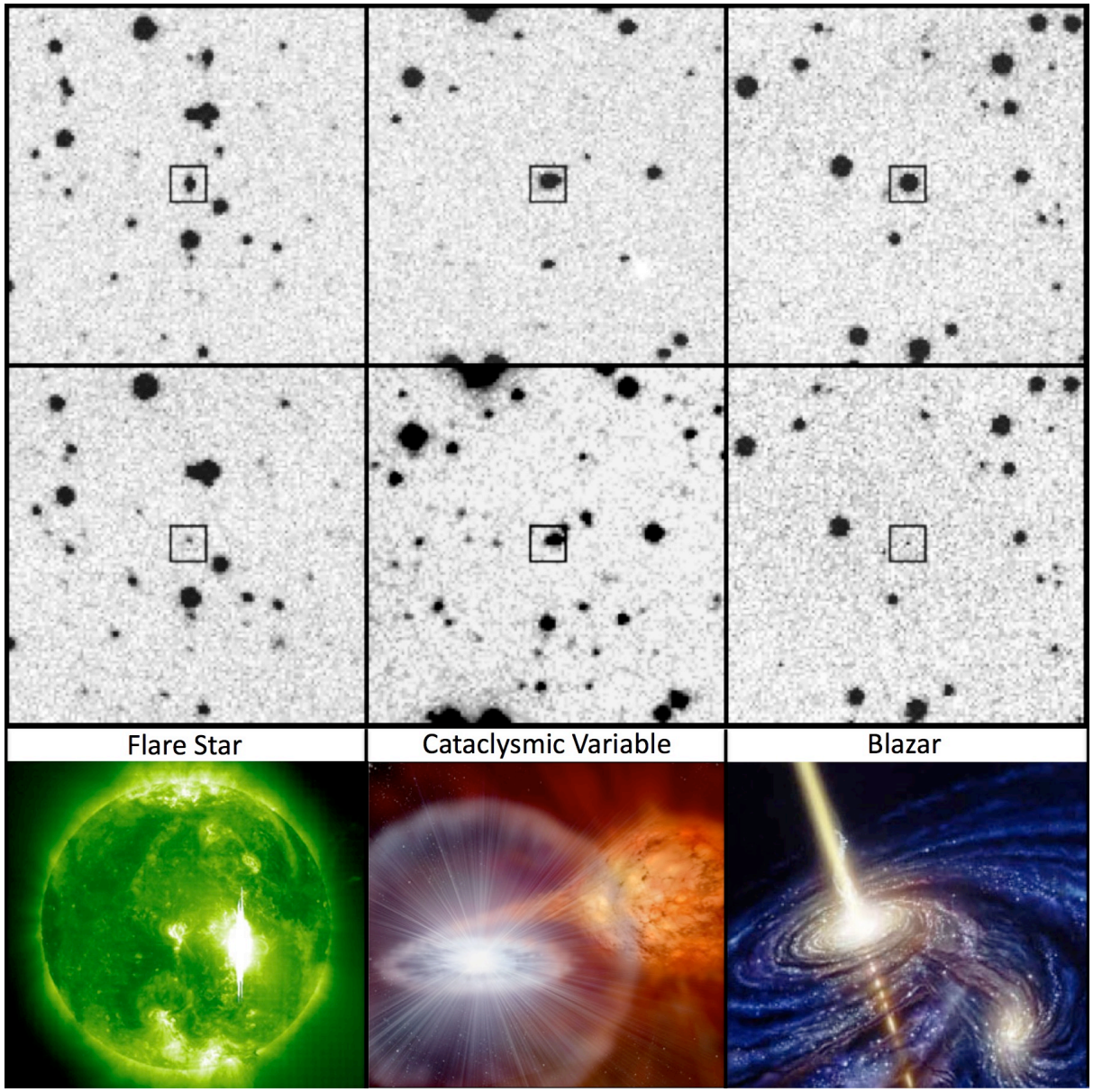
Why explore the time domain

- Moving objects (asteroids, TNOs, KBOs)
- SNe (cosmological standard candles, endpoints of stellar evolution)
- GRB orphan afterglows (constraining beaming models)
- Variable stars (stellar astrophysics, galactic structure)
- AGN (QSOs, fuelling mechanisms, lifetimes)
- Blazars, Cosmic Rays, …

Variability is known on time scales from ms to 10^{10} yr

Synoptic, panoramic surveys → event discovery

Rapid follow-up and multi- λ → keys to understanding


Flare Star
Cataclysmic Variable
Blazar

Large datasets

- CRTS

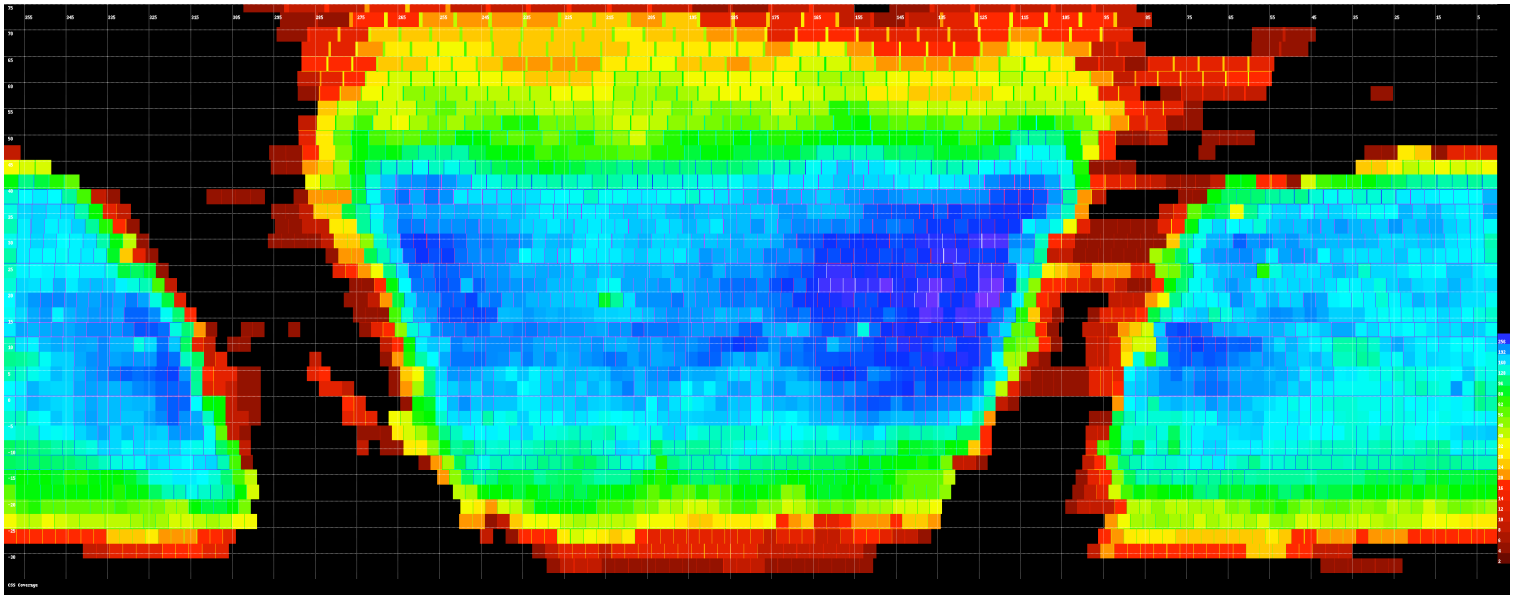
- LSST
- GAIA
- VAST/ASKAP

- BNs, GPRs, NNs, …

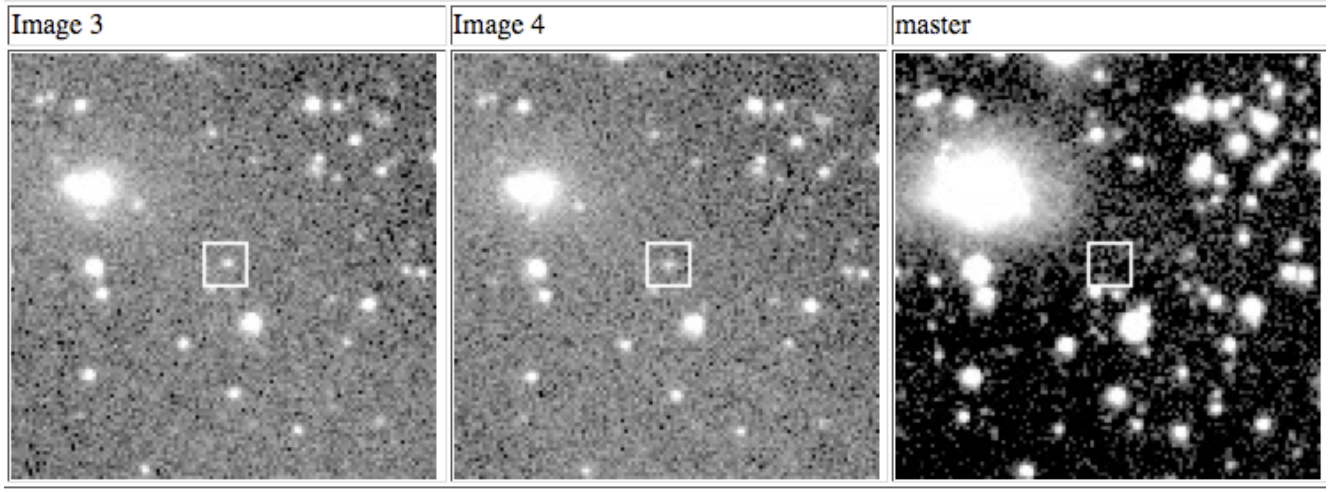
- DAME; VAO compatible

CRTS

Mt. Bigellow
4kx4k CCD
1200 deg^2/nt
4x10 min exp

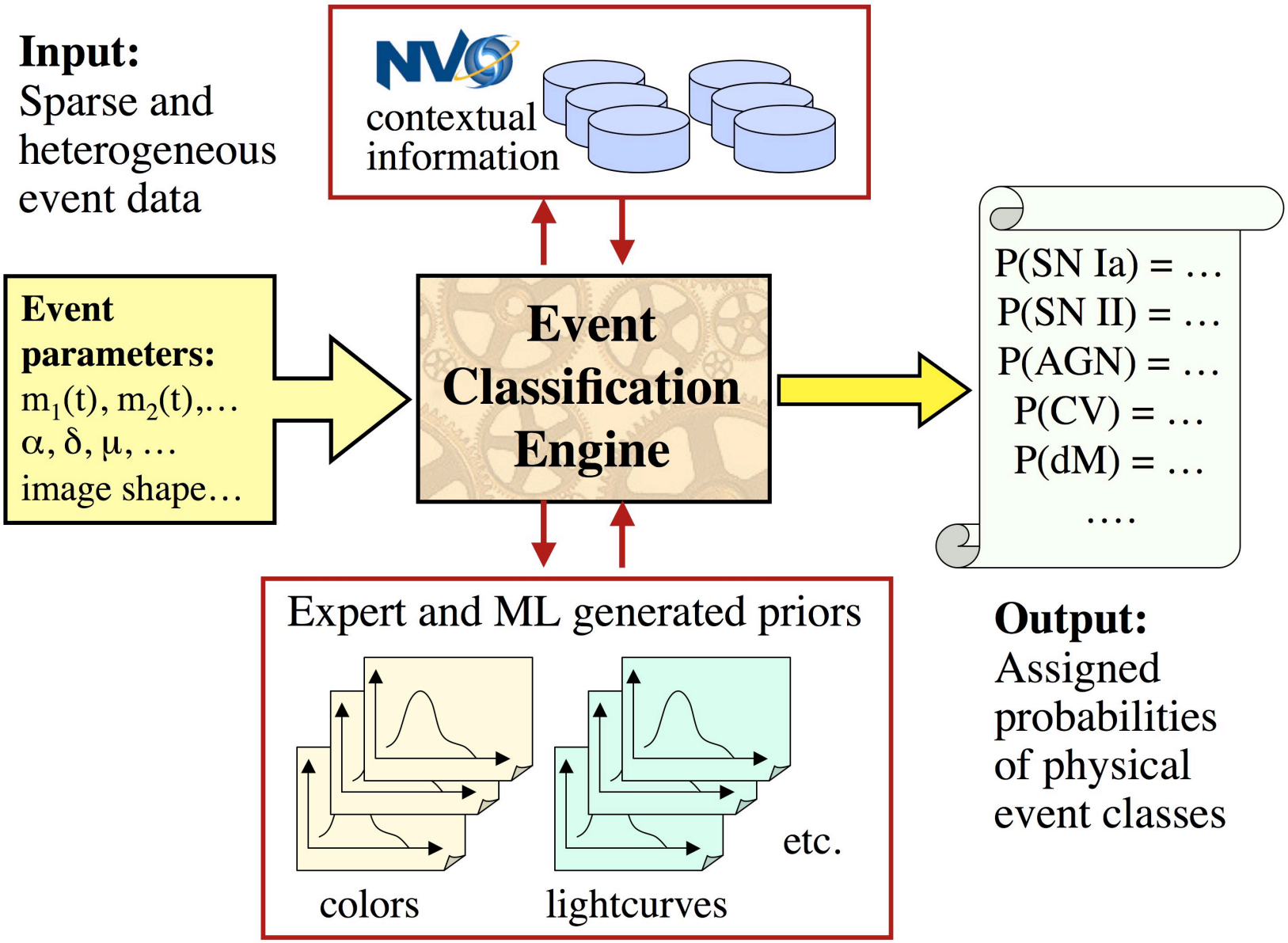


- **~ 2000 optical Transients so far.**
- **~70% SNe and CVs**
- Others include Blazars, AGN, variable stars and flare stars, and HPM stars.

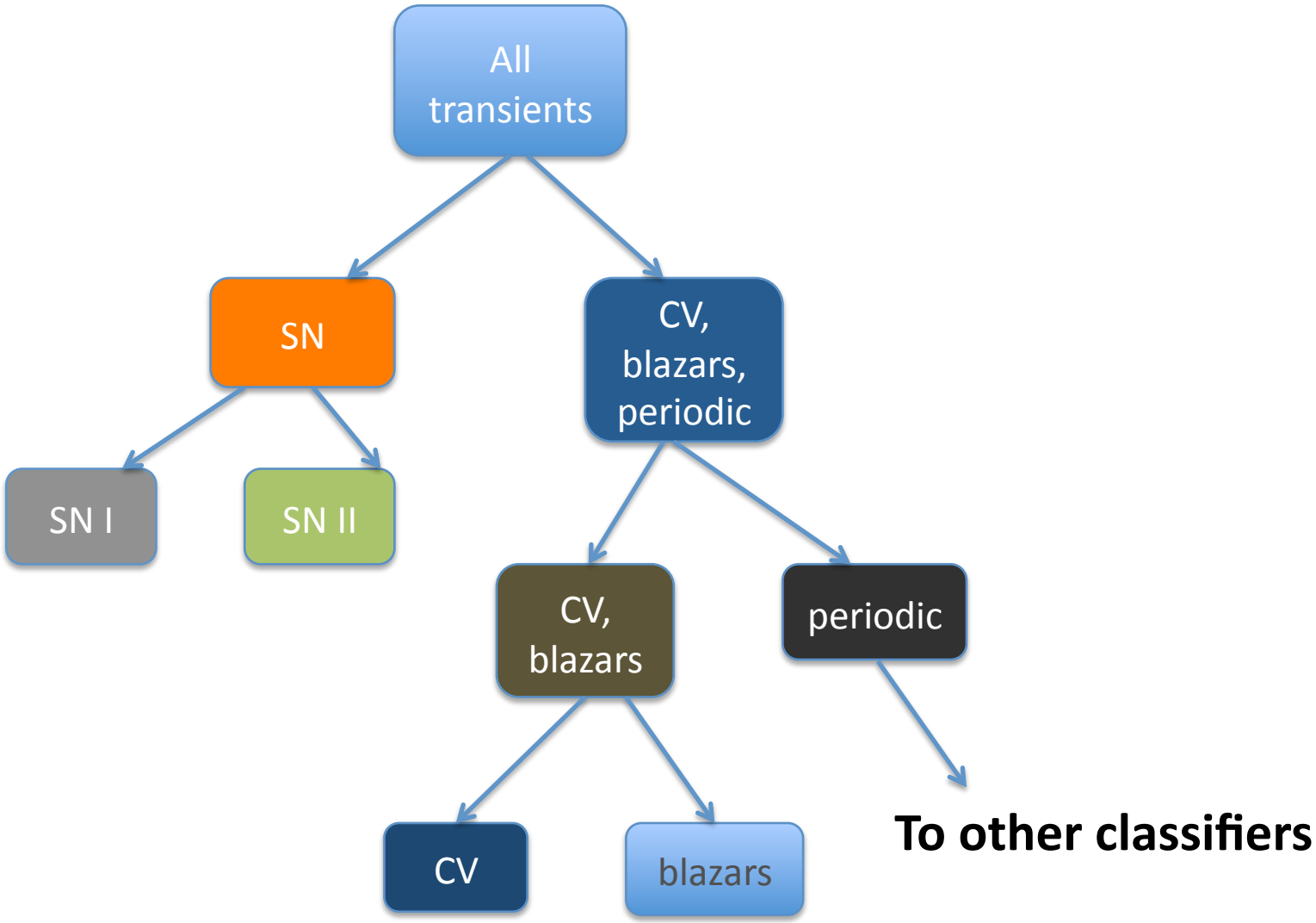


SN z=0.05
CSS 20090711

http://www.crts.org


Input:
Sparse and heterogeneous event data
contextual information
Event parameters:
$m_1(t)$, $m_2(t)$,…
α, δ, μ, …
image shape…
Event Classification Engine
P(SN Ia) = …
P(SN II) = …
P(AGN) = …
P(CV) = …
P(dM) = …
….
Output:
Assigned probabilities of physical event classes
Expert and ML generated priors
colors
lightcurves
etc.

Broad, incomplete hierarchy

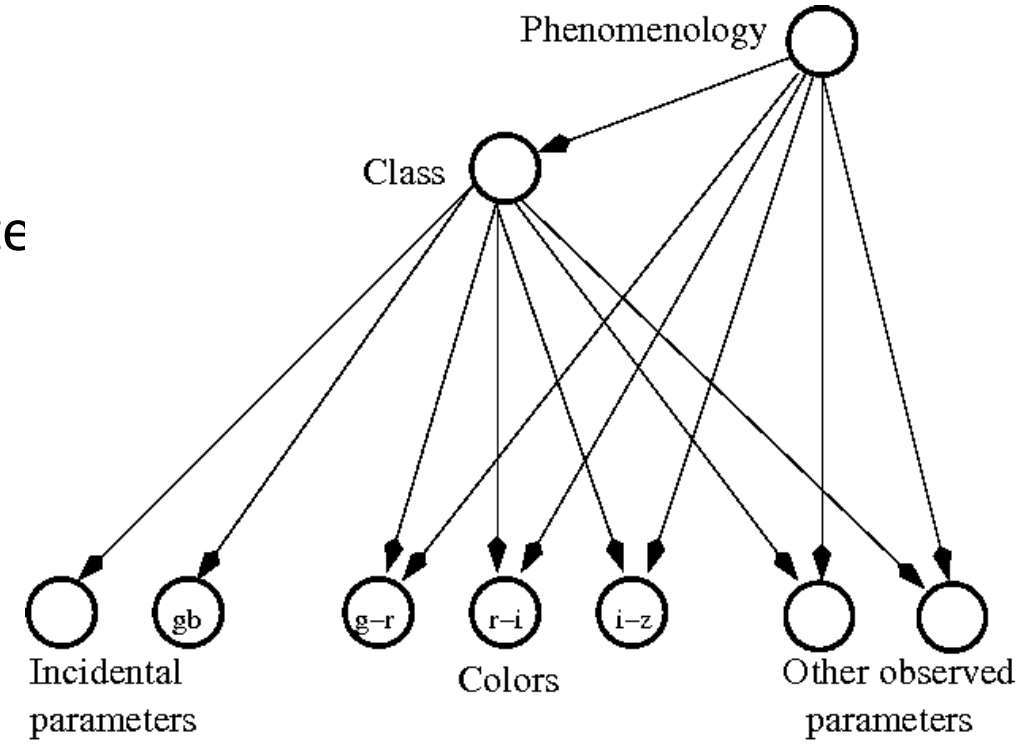


Dealing with …

- Deciding a winner
  - Winner take all; 50+; 40-10 rule
- Upper limits
  - Non-detections; missing data (galaxy proximity)
- Combining classifiers
  - Sleeping expert; multi armed bandit; but how?
- Error-bars
  - SN challenge; 4 slopes; early data challenge
- Choosing follow-up

Building Bayesian Networks

- Local dependencies, irrelevancies are evaluate using modeling
- Priors, likelihoods are obtained
- Data define network



Radio, for instance

Questions raised by the Data paucity regime

- How many classes?
- Too few: probabilities incorrect (where do objects belonging to unrepresented classes go?)
- Too many: overlaps increase (e.g. SN of different types; variables of different types) and probability splits into smaller fractions
- What kind of winner?

Priors based on CRTS data (dm > 2)

3 colors + gb (WTA)	CV (0.65)	SN (0.71)	BL (0.33)	REST (0.23)
CV	0.72	0.08	0.08	0.13
SN	0.23	0.46	0.12	0.19
BL	0.24	0.03	0.49	0.24
REST	0.34	0.18	0.21	0.26

8% CV classified as SN, 65% of objects classified as CV are actually CV

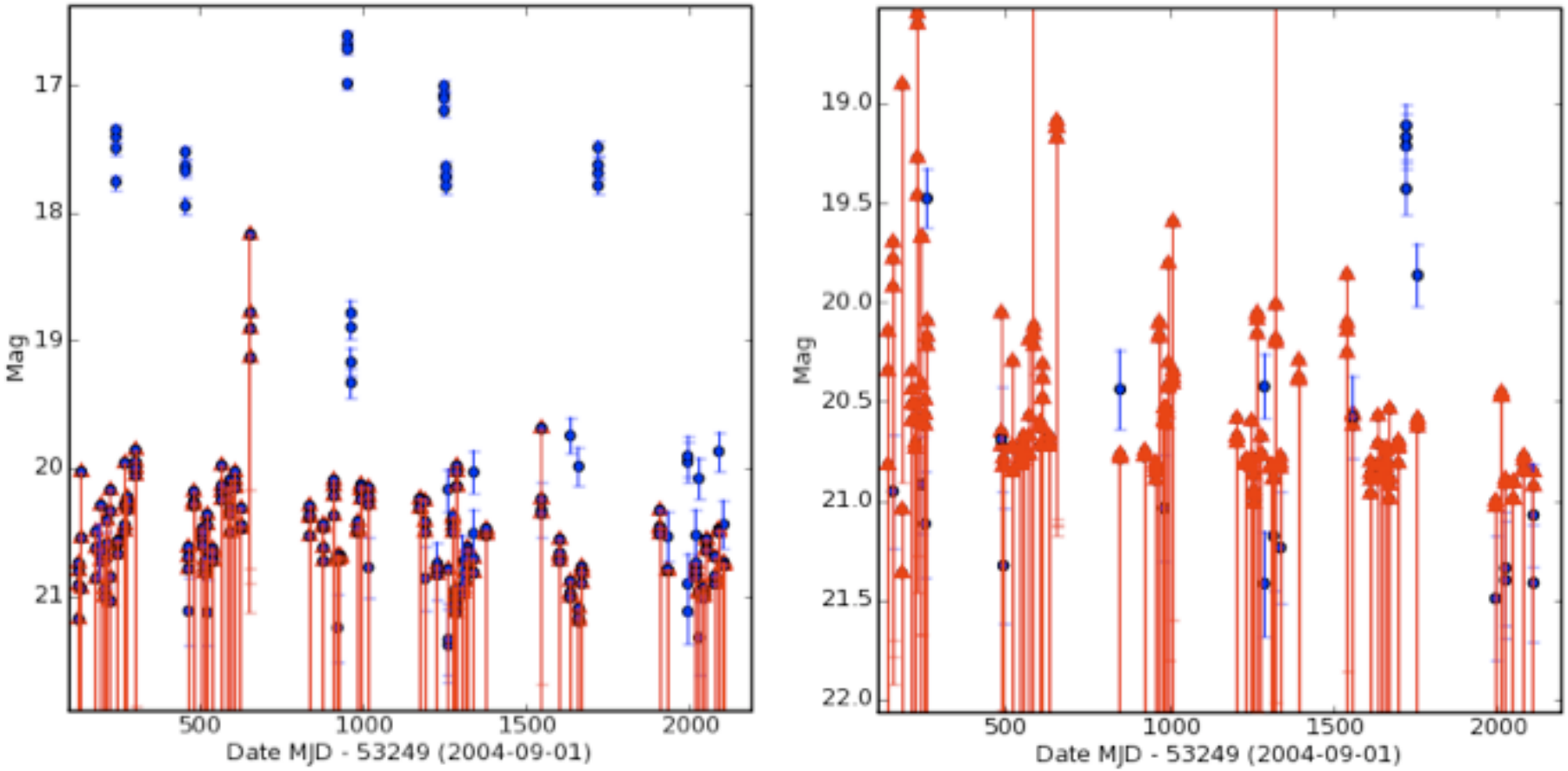
- Winner-take-all
- At least 50%
- 40%+ and 10% diff
- allowing missing info

Based on a single set of observations

- Adding peripheral parameters like gb and distance to nearest galaxy helps
- Having additional colors is good
- More context info helps (flux in radio, x-ray etc.)
- Need to inculcate temporal information

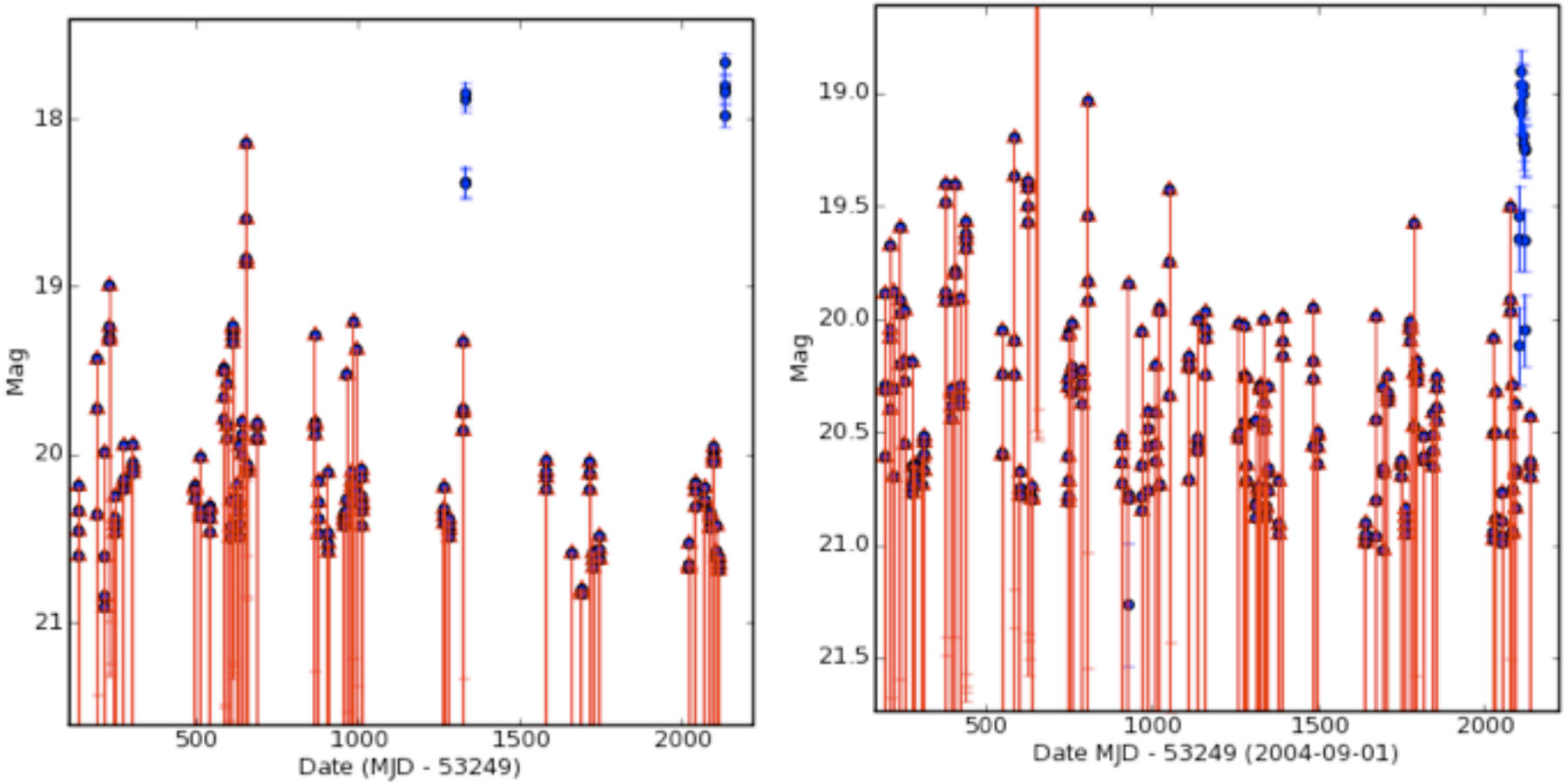
3 colors + gb + galaxy prox. (WTA)	CV (0.74)	SN (0.84)	BL (0.31)	(1-contam.)
CV	0.74	0.08	0.16	
SN	0.21	0.50	0.27	
BL	0.19	0.00	0.80	
				completeness

Transients with extra points after discovery



CV and SN from CRTS

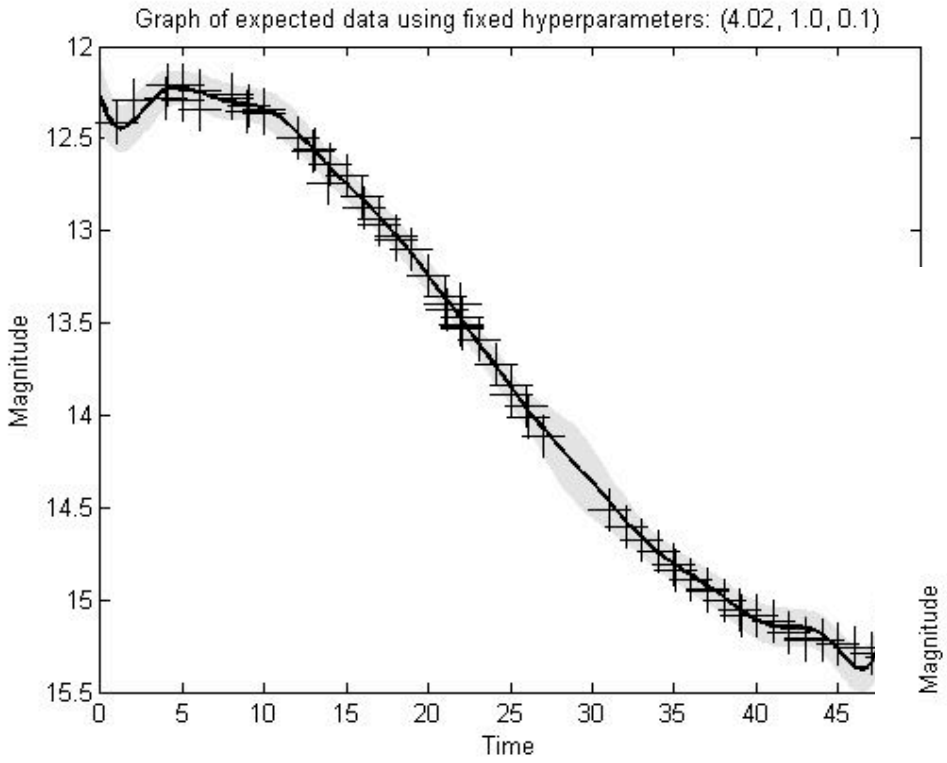
Fresh transients



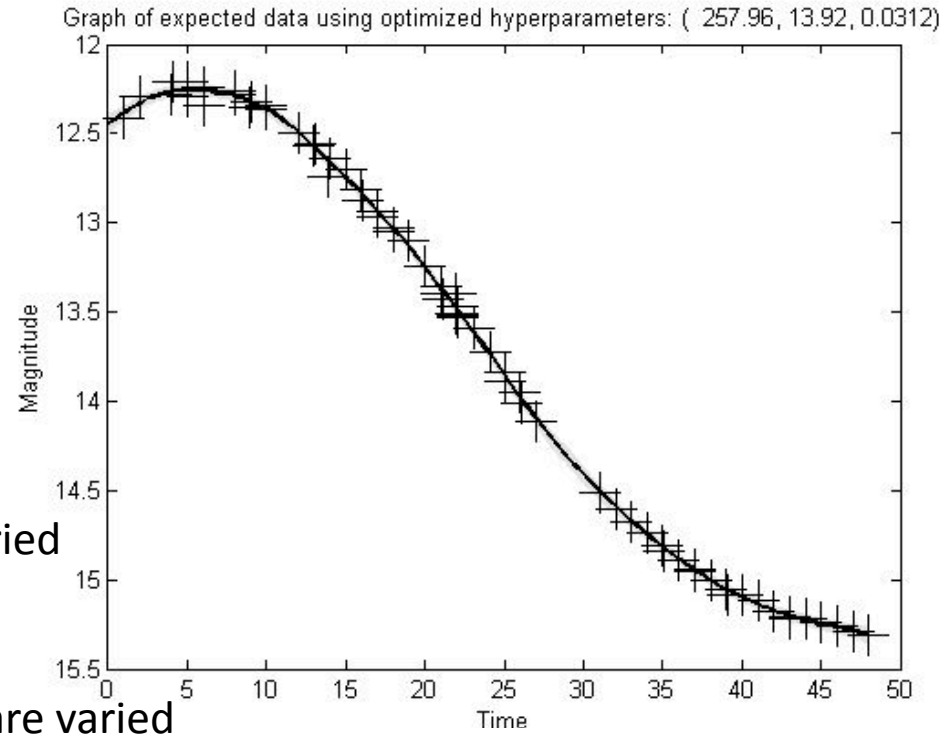
CV and SN from CRTS

GARCH and Kalman filtering

Using GPR with lightcurves



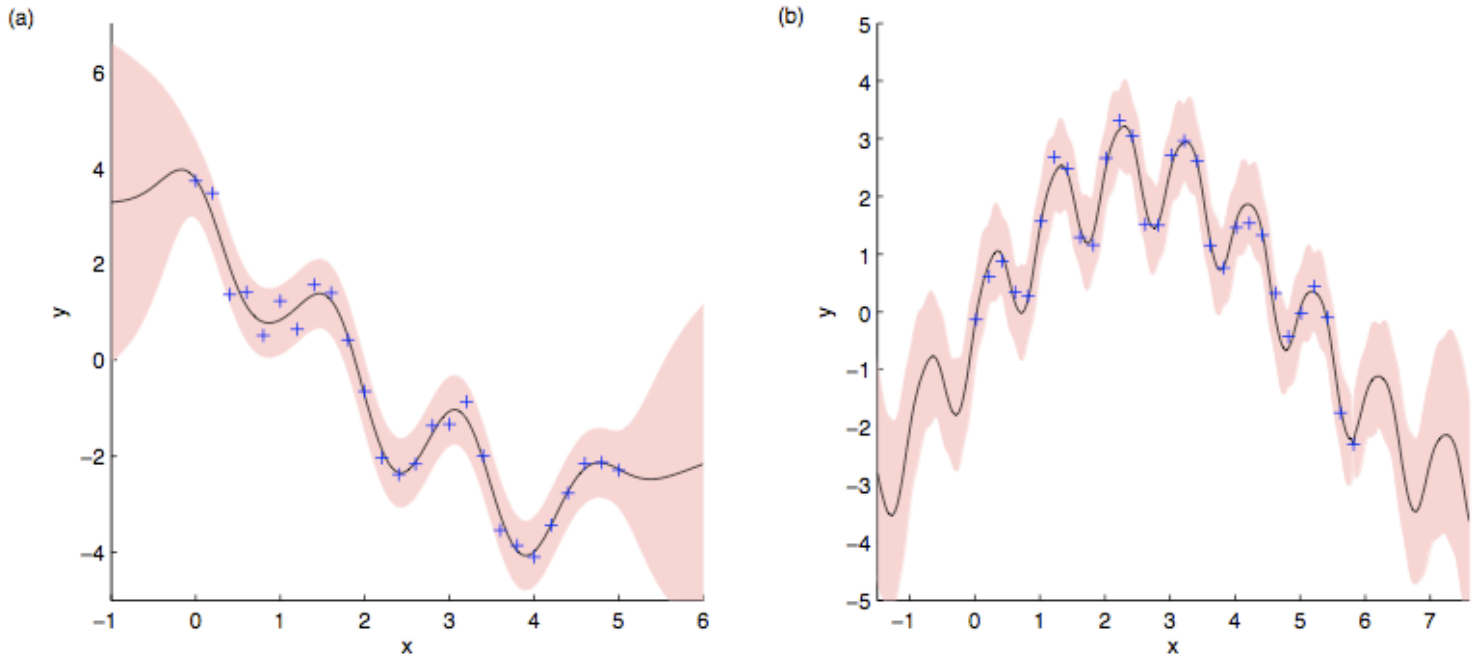
Given several epochs and corresponding magnitudes, estimate the likelihood of a particular magnitude for a new epoch (using some covariance function)



The 3 hyperparameters are “free” and are varied

The 3 hyperparameters are “free” and are varied

$$Cov(f(x_p), f(x_q)) = k_y(x_p, x_q) = \sigma_f^2 e^{-\frac{1}{2}l^2(x_p - x_q)^2} + \sigma_n^2 \delta_{pq}$$

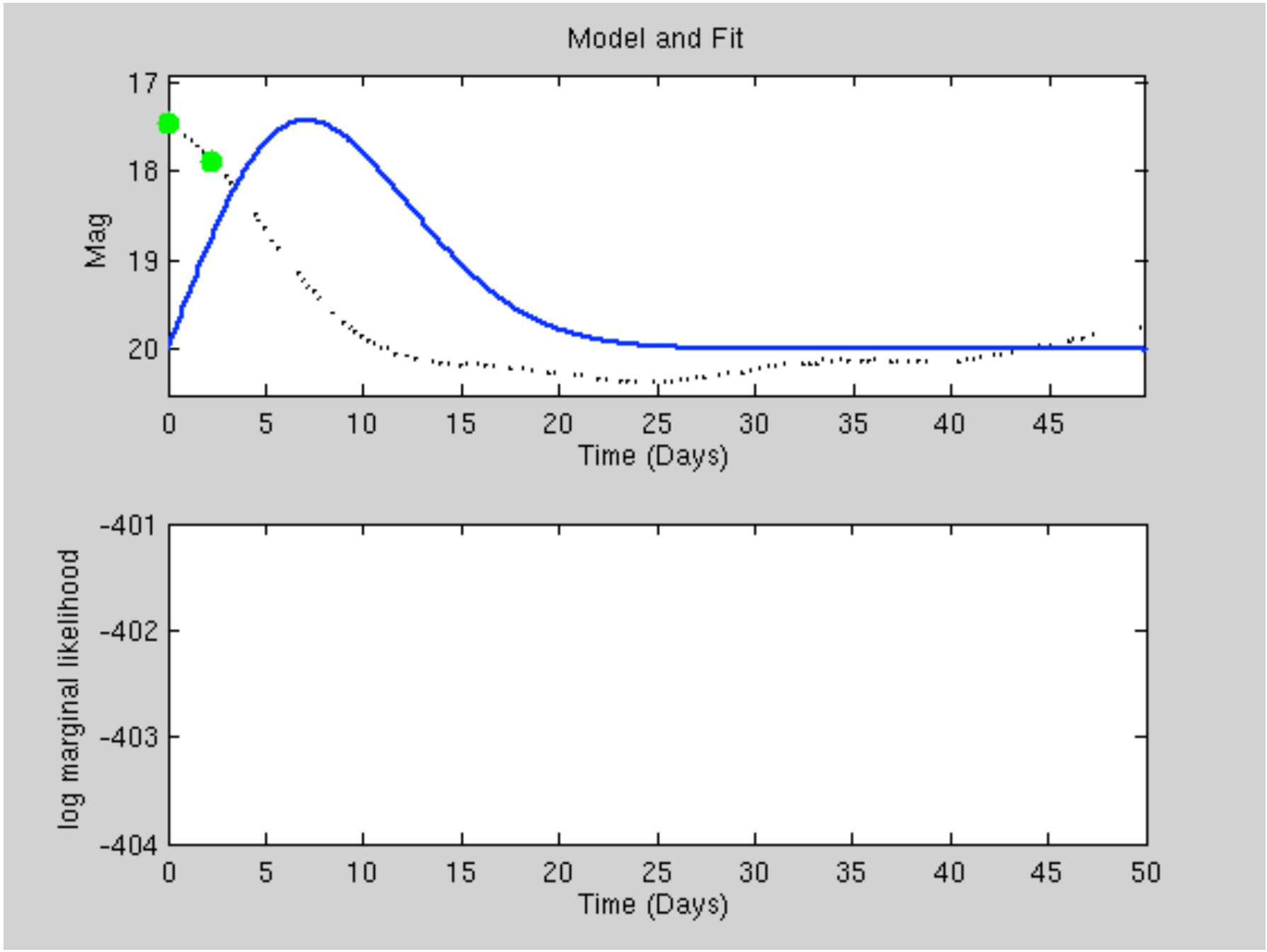


M. Ebden

Figure 3: Estimation of y_* (solid line) for a function with **(a)** short-term and long-term dynamics, and **(b)** long-term dynamics and a periodic element. Observations are shown as crosses.

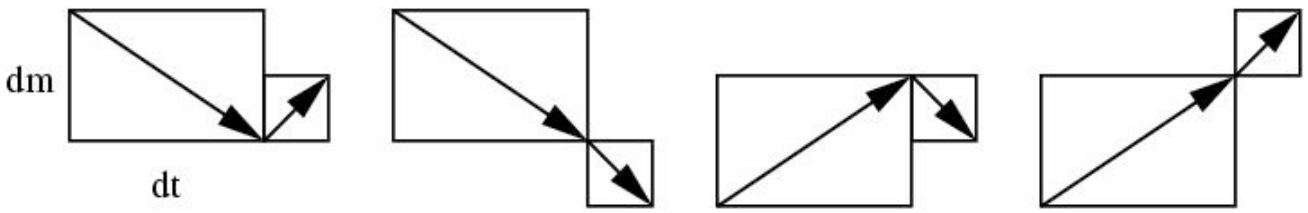
$$k(x, x') = {\sigma_f}_1^2 \exp\left[\frac{-(x-x')^2}{2l_1^2}\right] + {\sigma_f}_2^2 \exp\left[\frac{-(x-x')^2}{2l_2^2}\right] + \sigma_n^2 \delta(x, x')$$

$$k(x, x') = \sigma_f^2 \exp\left[\frac{-(x-x')^2}{2l^2}\right] + \exp\{-2\sin^2[\nu\pi(x-x')]\} + \sigma_n^2 \delta(x, x')$$

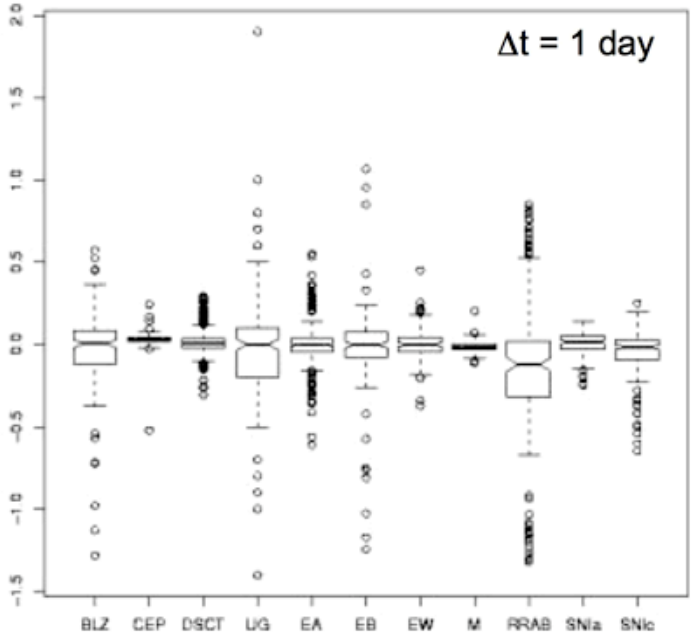

Model and Fit
Mag
17
18
19
20
0
5
10
15
20
25
30
35
40
45
Time (Days)
log marginal likelihood
-401
-402
-403
-404
0
5
10
15
20
25
30
35
40
45
50
Time (Days)

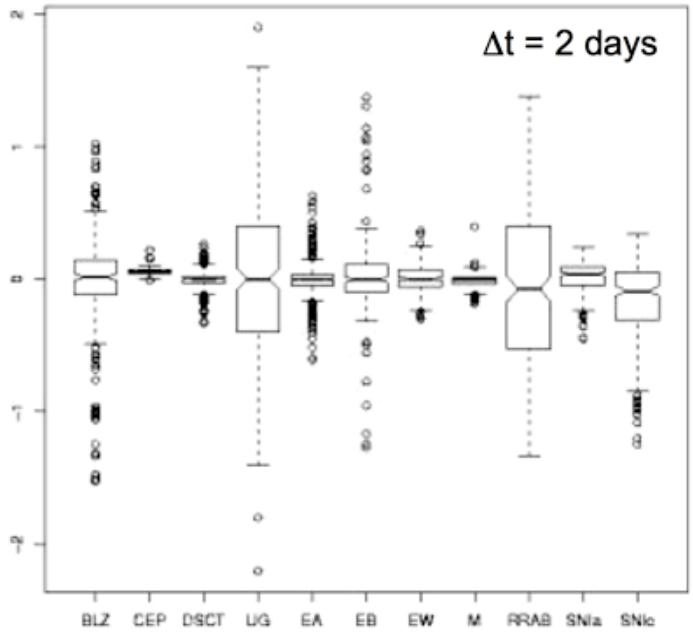
Characterization Vs. Classification

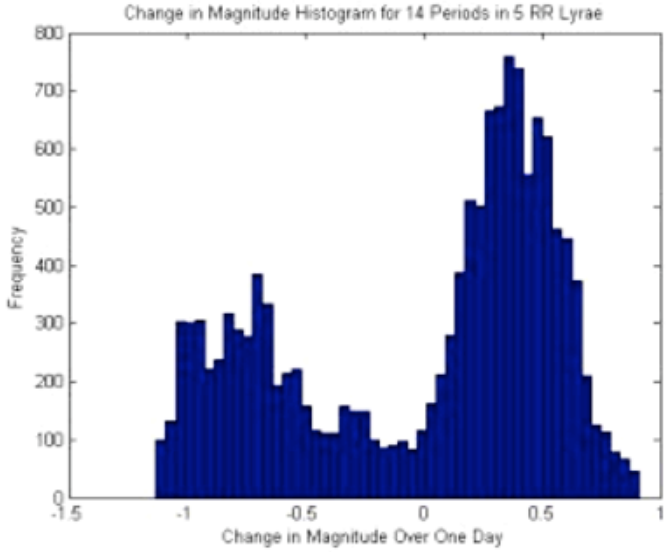
- Early focus on the extraction and dissemination of time series
- Characterizations is important
  - dm/dt
  - change of direction per unit time
  - change in periodicities (e.g,. wavelet or fourier decomposition);
  - variation in dm/dt
  - acceleration in dm/dt

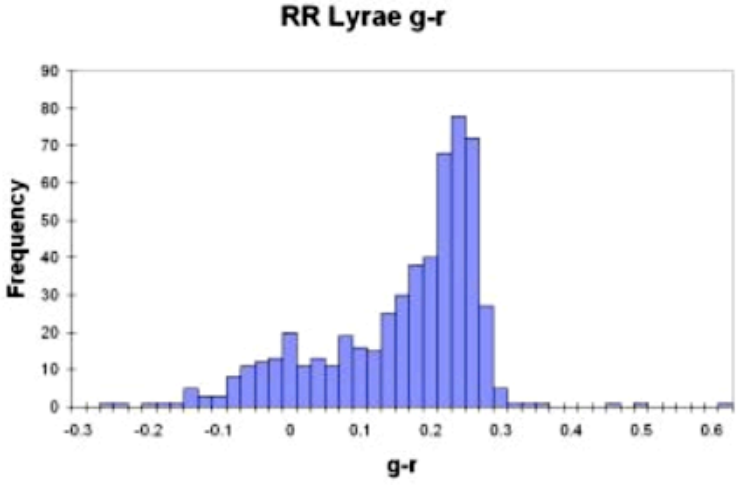


Most SNe will not become fainter and then brighten up


Δt = 1 day
BLZ
CEP
DSCT
UG
EA
EB
EW
M
RRAB
SNIa
SNIc


Δt = 2 days
BLZ
CEP
DSCT
UG
EA
EB
EW
M
RRAB
SNIa
SNIc


Change in Magnitude Histogram for 14 Periods in 5 RR Lyrae
Frequency
Change in Magnitude Over One Day


RR Lyrae g-r
Frequency
g-r

Principal component analysis on different “simultaneous” bands

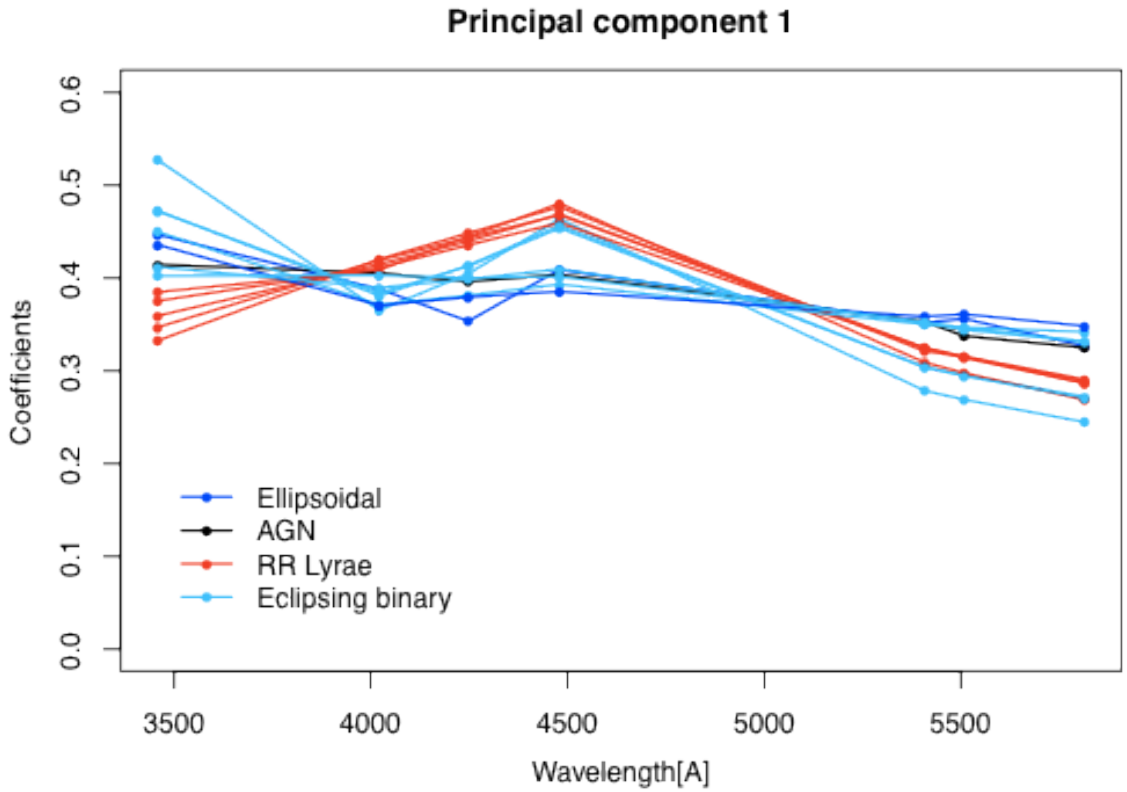
Proposed by Paul Bartholdi 2005: Applied to the Geneva photometry constant stars
(results: some are variable!)

- Perform the period search on the “new magnitude” (first component)
- Characterize the physical properties of stars
- Current tests on Geneva Photometry and on SDSS stripe 82

$$\overline{s}_j = \frac{1}{N}\sum_i s_{ij}$$

$$d_{ij} = s_{ij} - \overline{s}_j$$

$$C_{lm} = \frac{1}{N}\sum_i d_{il}d_{im}$$

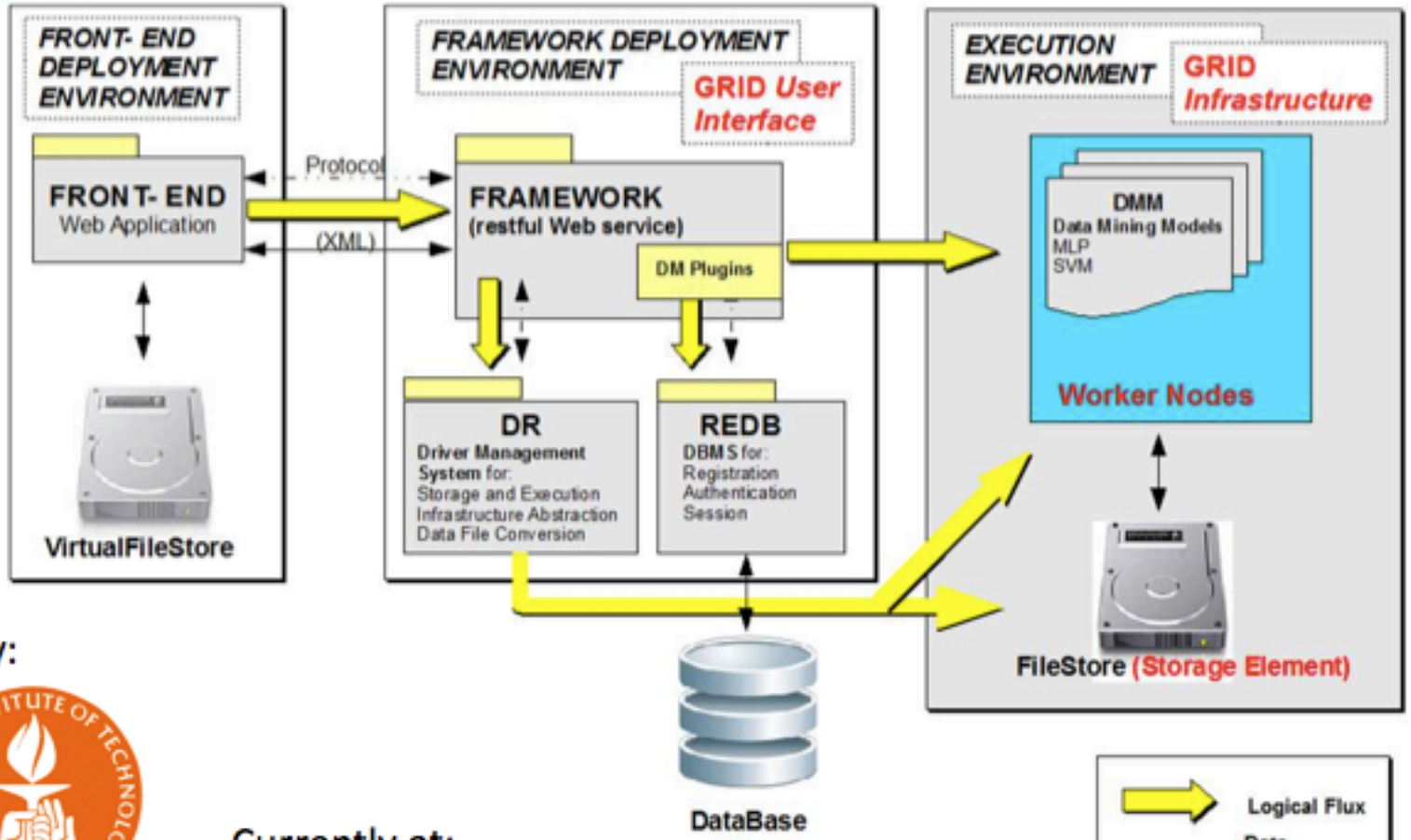


Eyer et al. 2010

DAta Mining & Exploration

DA ME

A powerful, expandable, web-based user-friendly data mining service for astronomy

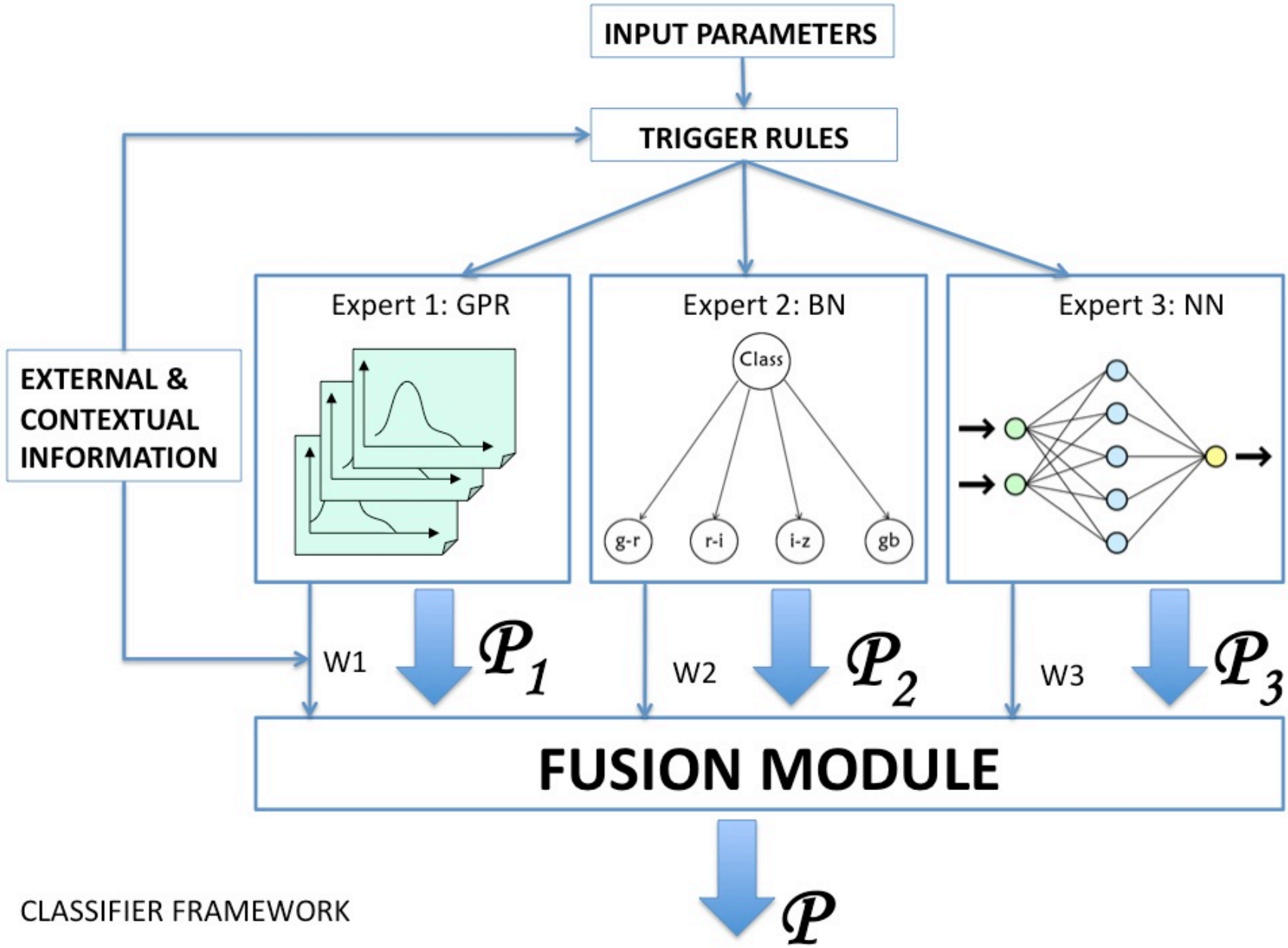


Brought to you by:

... and donated to VO

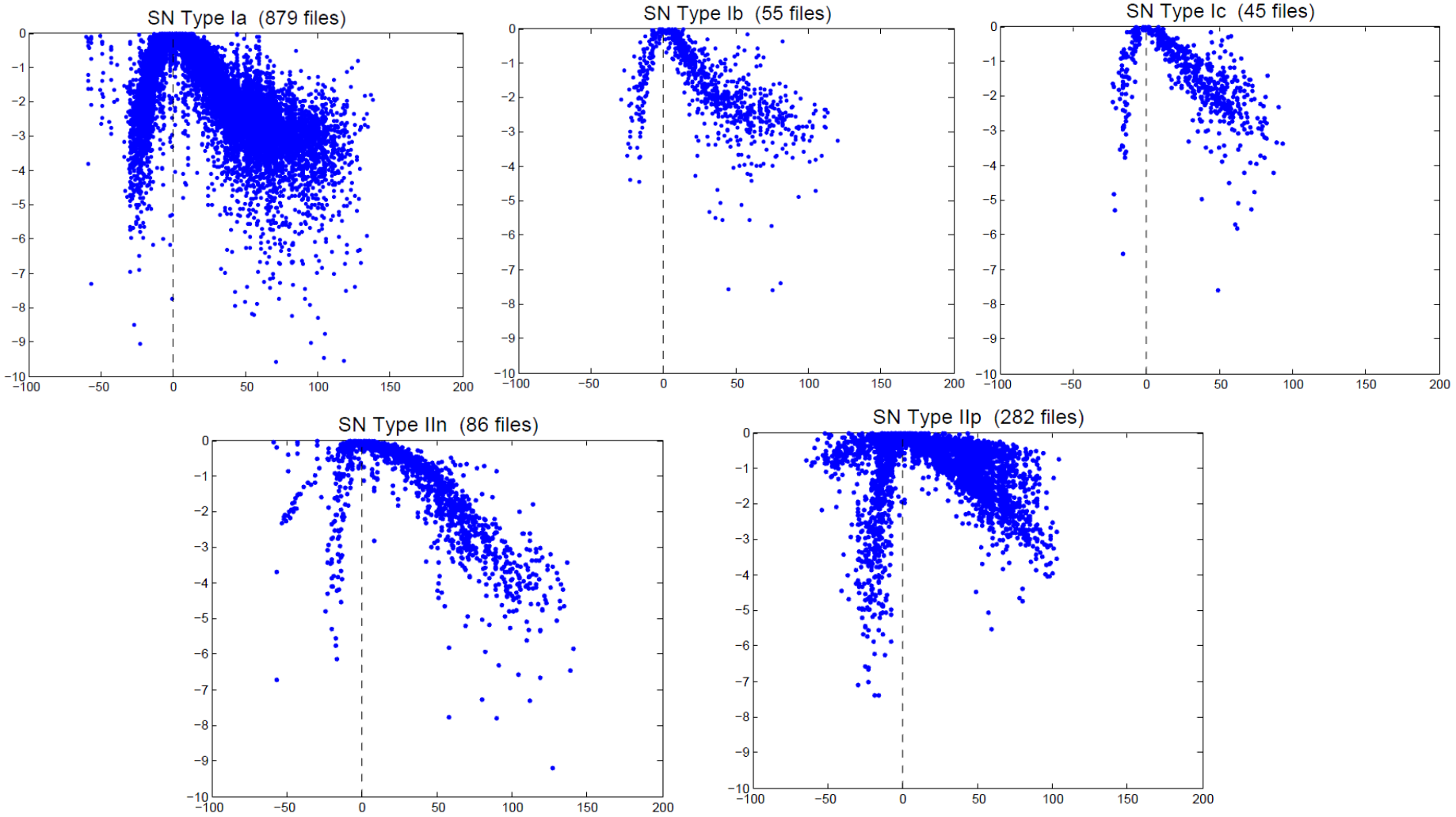
Currently at:

http://voneural.na.infn.it



CLASSIFIER FRAMEWORK

- Non-sparse time series (Many methods; relatively easy)
- Sparse time series (Non-trivial)
  - Non-gridded
  - Error-bars
  - Upper limits

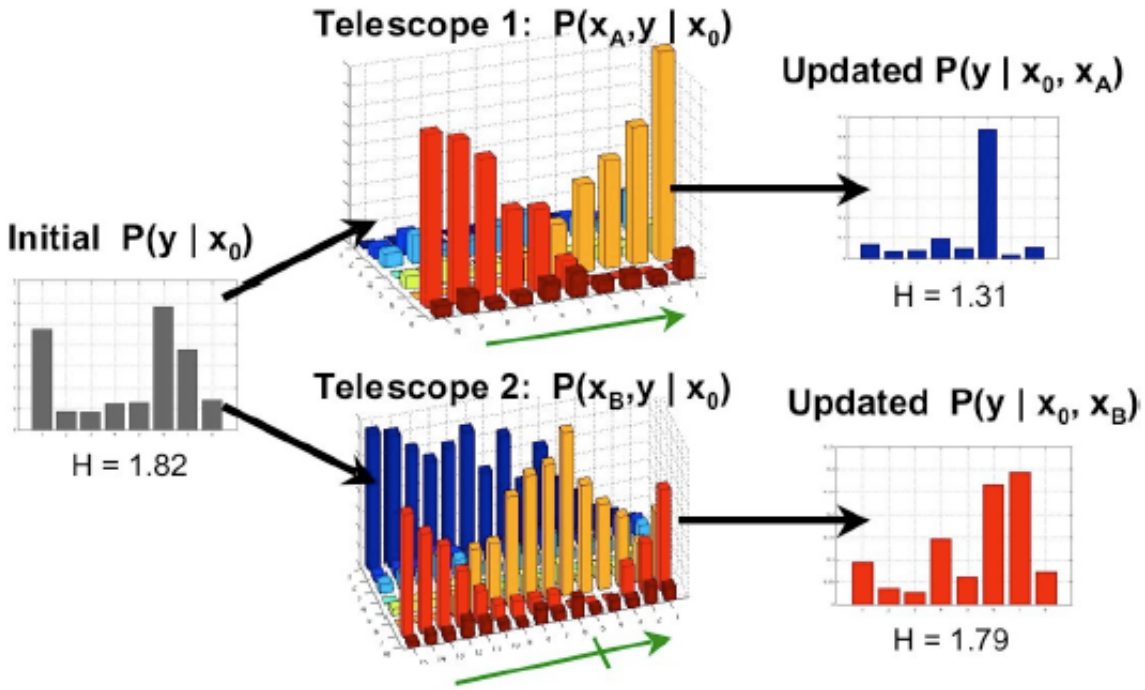


Supernova challenge

- with photo-z for host galaxy
- without redshift information
- early-epoch challenge

Follow-up (for missing values)

- Such that it will help discriminate better
- Serve probabilities so that consumers can choose their types of transients
- Widest possible models
- (resource uniformity)
- (well connectedness)



Further work needed for …

- Deciding a winner
  - Winner take all; 50+; 40-10 rule
- Upper limits
  - Non-detections; missing data (galaxy proximity)
- Combining classifiers
  - Sleeping expert; multi armed bandit; but how?
- Error-bars
  - SN challenge; 4 slopes; early data challenge
- Choosing follow-up