Combining Diverse Classifiers [for astronomical transients] {and in real-time}

Ashish Mahabal, Caltech

IJCAI, 16 July 2009

SG Djorgovski, Ciro Donalek, Baback Moghaddam, Michael Turmon, Andrew Drake, Matthew Graham, Roy Williams, …

- Colors and context info with BNs
- Lightcurves and GPRs
- Combining them

Transformations in astronomy

- Shallow to deep
- Small to large
- Sporadic to repeated

Towards digital movies …

ROTSE, NEAT, DLS, FSVS, …

DPOSS, PQ, CSS, PTF, Pan-STARRS, LSST …

(GALEX, Spitzer, FIRST, …, SKA)

- Area covered
- Depth of coverage
- Number of wavelengths
- Baseline in time
- Number of epochs

Etendue (throughput) ~ dA * dm * dλ * dt *dn

Synoptic skysurveys

Opened up new dimensions

Challenges besides data mining:

- Lots of follow-up observing
- Selecting candidates to follow
- In real-time

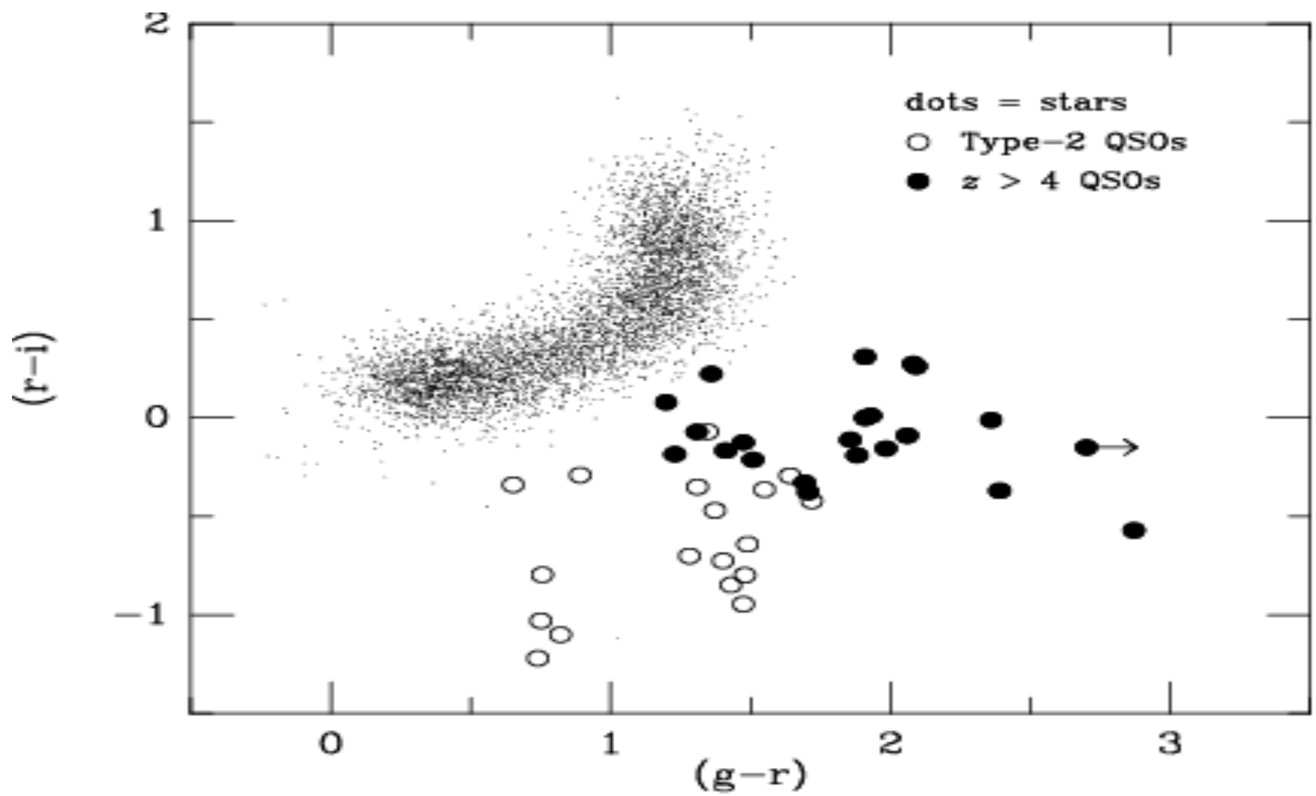
Bayesian Networks to the rescue

- BNs of various flavors can tackle these issues (but there are many barriers to be crossed!)

- Data define network
- No “training” necessary

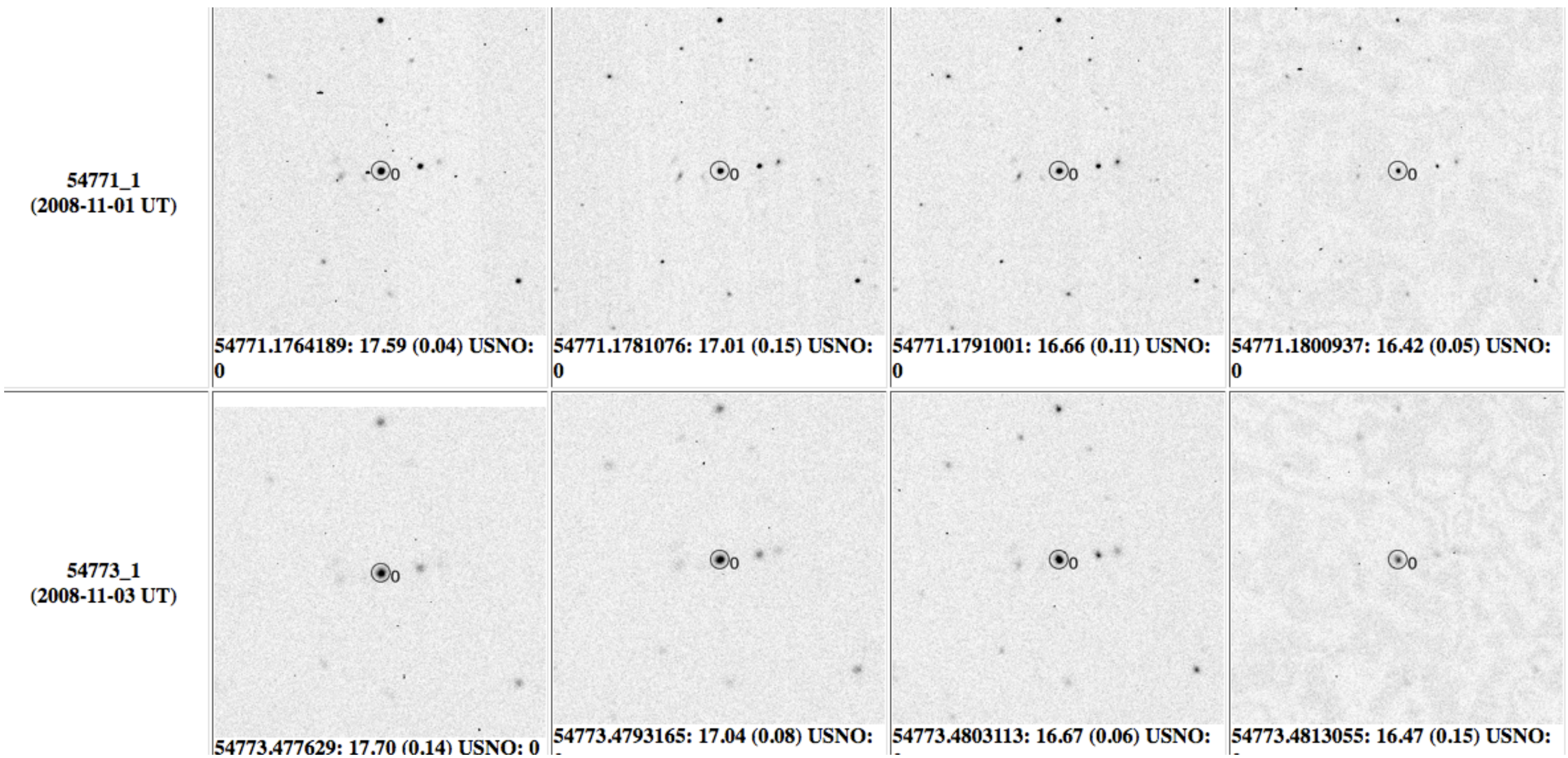
Colors for classification

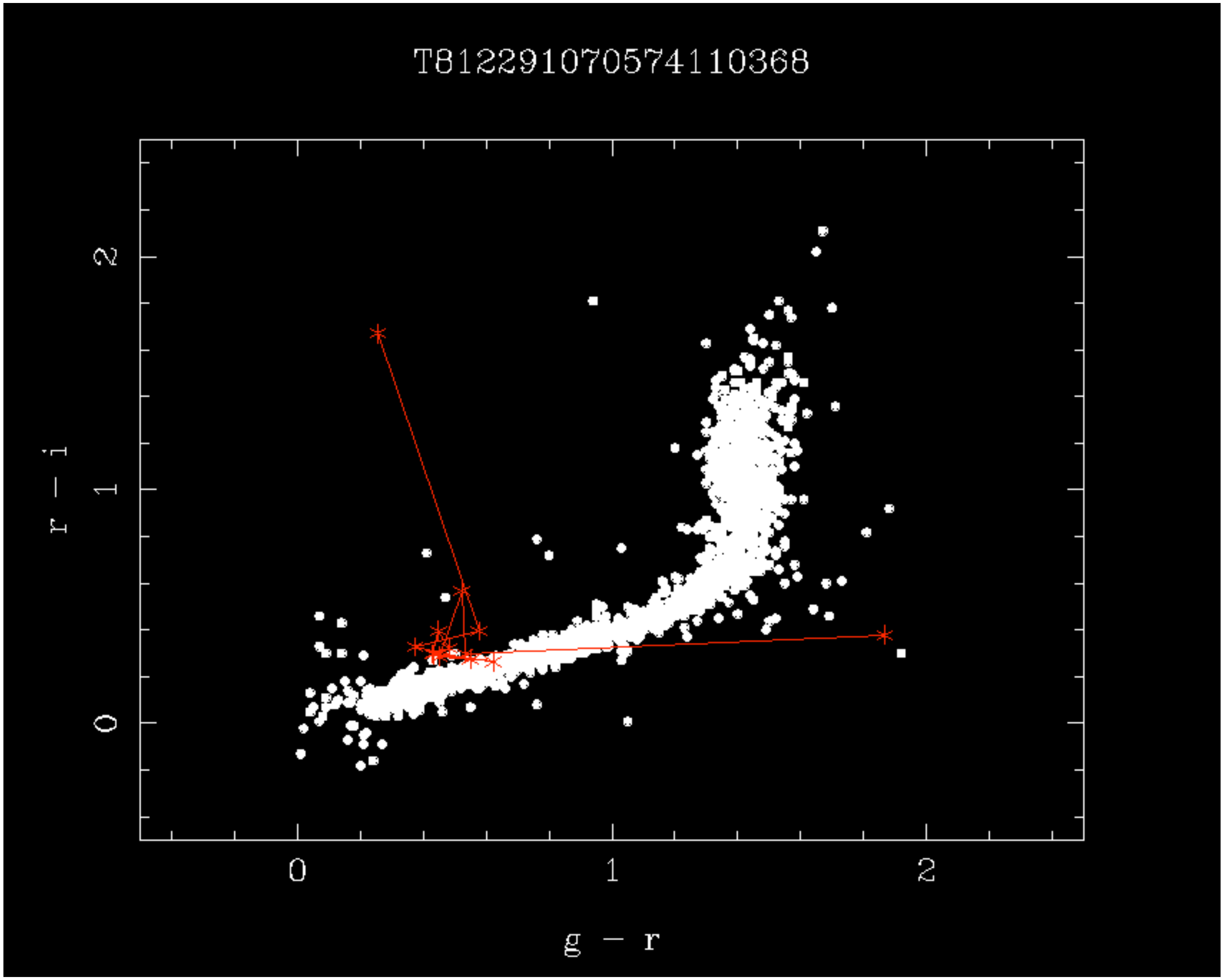
- Magnitude as basic observation (flux)
- Color as flux ratio
- Color-color diagram as a diagnostic



P60 follow-up of a CRTS transient

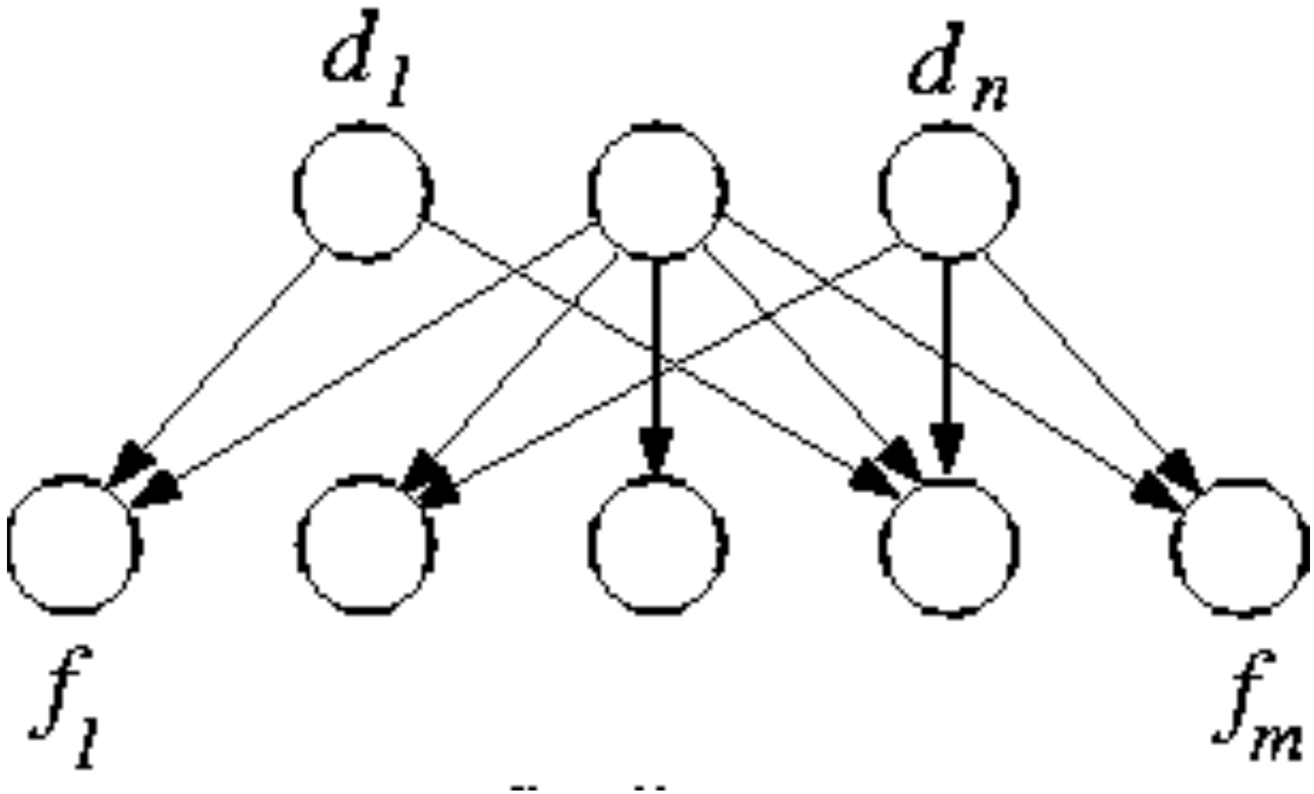
http://www.astro.caltech.edu/P60FollowUp/



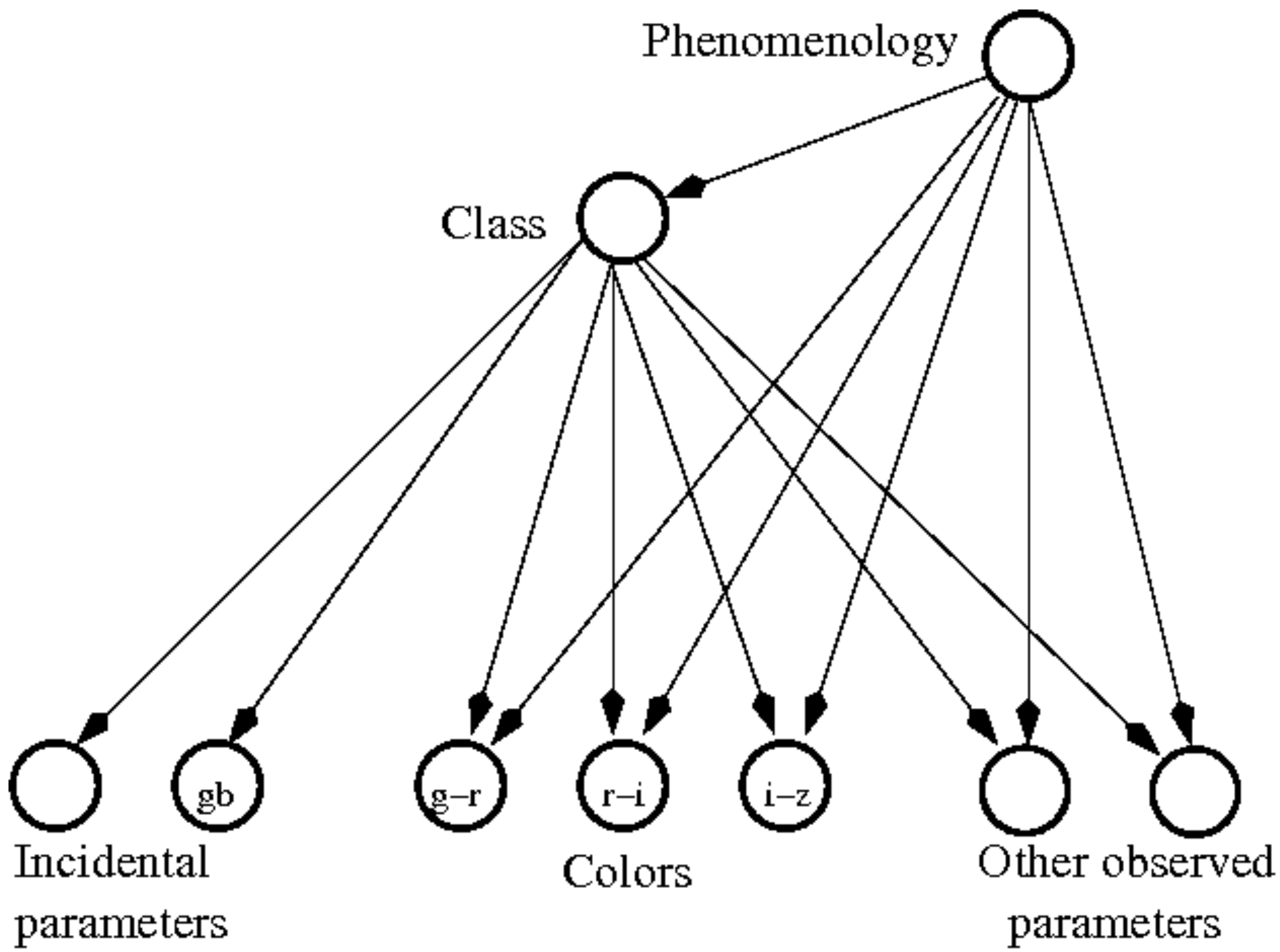


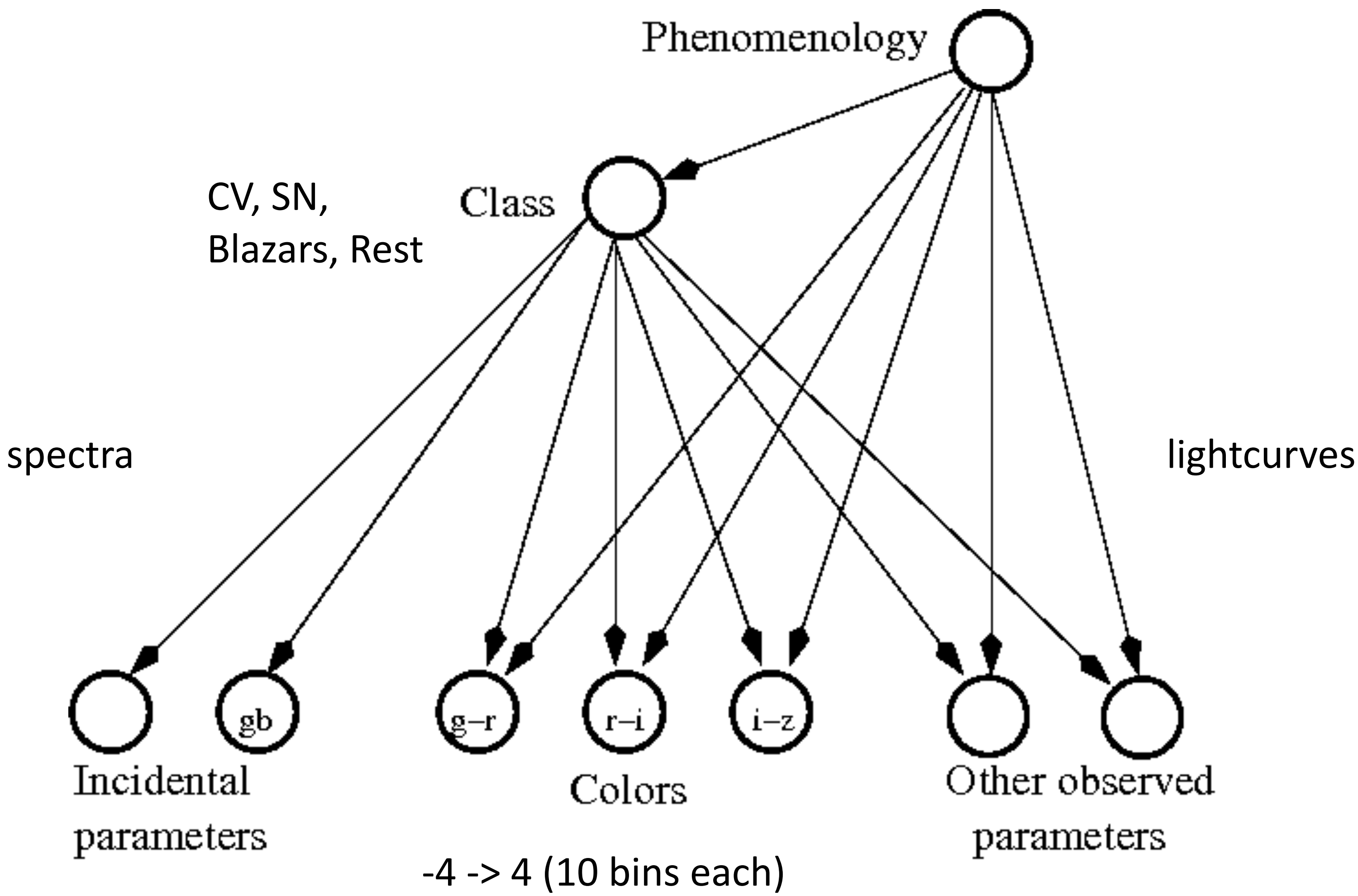
Class can be a function of time

DAG for variables like QMR-DT



Variables and observed properties


Phenomenology
Class
gb
g–r
r–i
i–z
Incidental
parameters
Colors
Other observed
parameters


Phenomenology
Class
CV, SN,
Blazars, Rest
spectra
lightcurves
gb
g–r
r–i
i–z
Incidental
parameters
Colors
Other observed
parameters
-4 -> 4 (10 bins each)

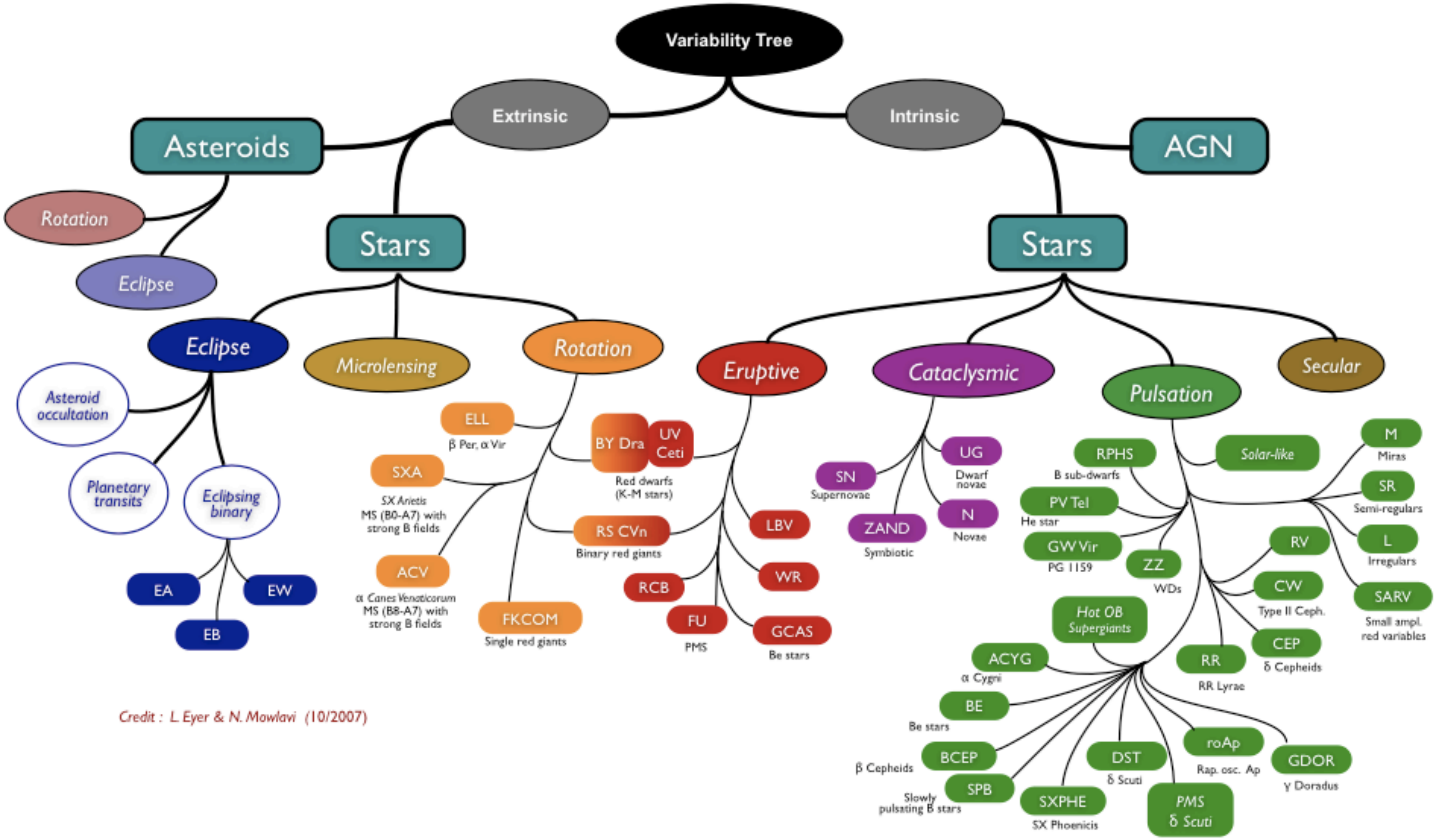
8% CV classified as SN, 60% of objects classified as CV are actually CV

3 colors, no gb (WTA)	CV (0.60)	SN (0.73)	BL (0.30)	REST (0.22)
CV	0.72	0.08	0.11	0.08
SN	0.42	0.41	0.08	0.09
BL	0.30	0.00	0.49	0.21
REST	0.32	0.13	0.34	0.21

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

Type	CV	SN	BL	REST
3,n,w	0.60	0.73	0.30	0.22
3,y,w	0.65	0.71	0.33	0.23
2+,n,w	0.58	0.77	0.31	0.17
2+y,w	0.65	0.78	0.41	0.19
3,y,50	0.75	0.83	0.37	0.19
3,y,40-10	0.73	0.82	0.37	0.18
2+,y,50	0.74	0.88	0.43	0.19

- More context information will help
- Heterogeneity
- How to maintain uniformity [from the same source]?
- What when filters change?
- Uniformity of priors?
    - Number of objects
    - Their magnitude range
    - Spread over time
- Ground truth?



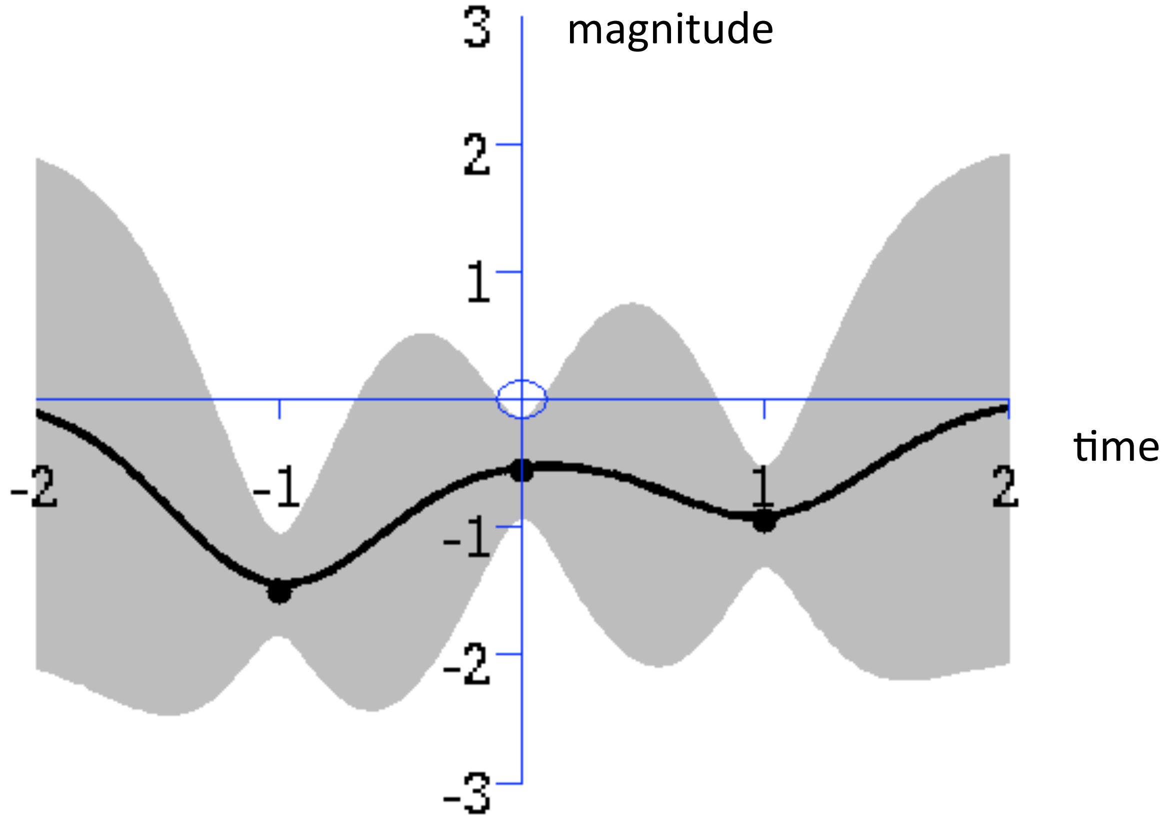
Credit : L. Eyer & N. Mowlavi (10/2007)

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?

- May be evaluate probability for each class independently?
- With GPR something like that may be possible.

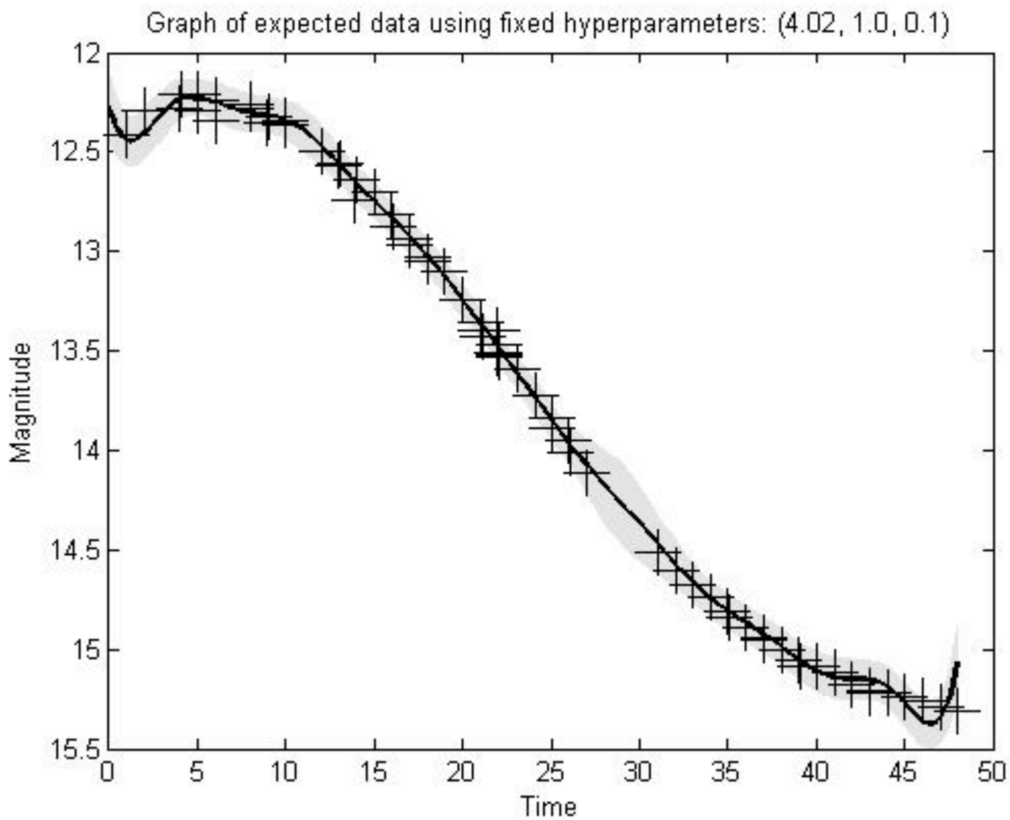
GPR schematic



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

Graph of a SN lightcurve ready to be fitted using GPR (using Squared exponential covarience function from Matlab's GPML).

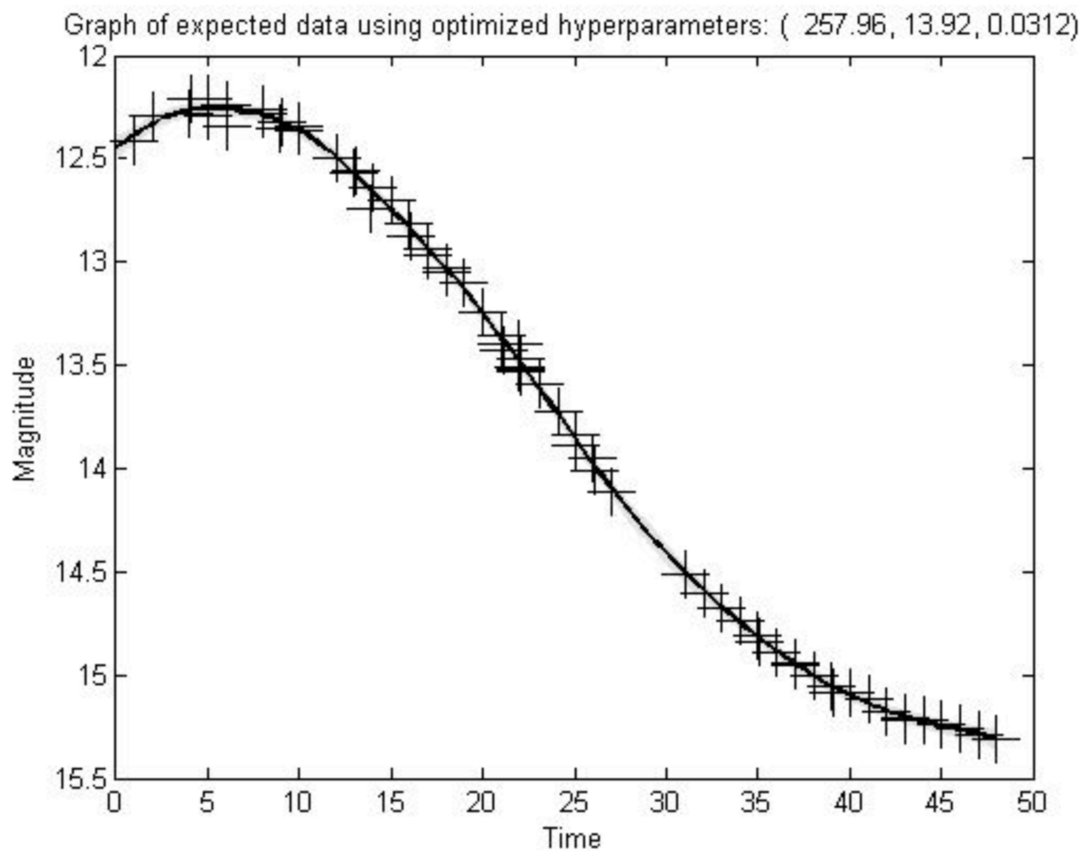
$$Cov(f(x_p), f(x_q)) = k_y(x_p, x_q) = \sigma_f^2 e^{-\frac{1}{2} l^2 (x_p - x_q)^2} + \sigma_n^2 \delta_{pq}$$



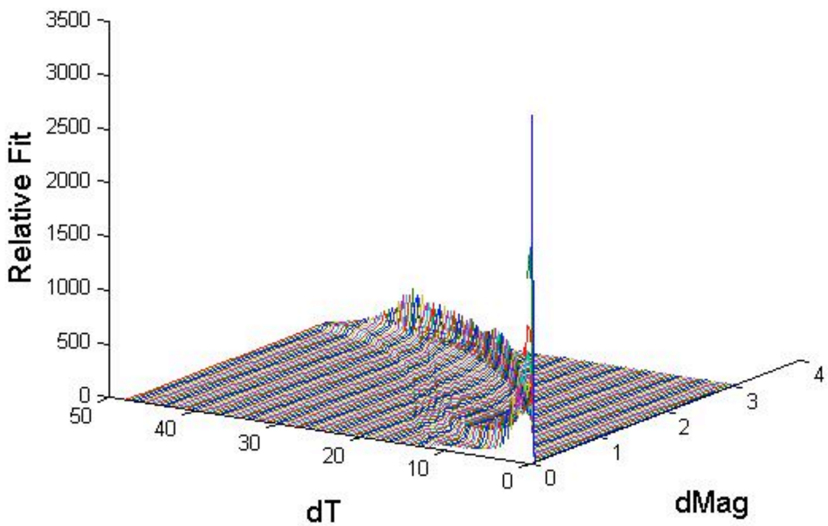
The 3 hyperparameters are "free" and are varied

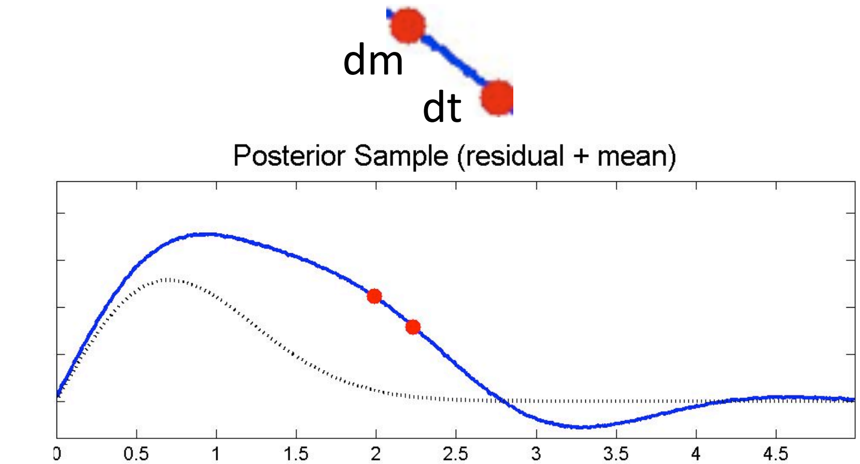
Graph of a SN lightcurve fitted using GPR (using Squared exponential covarience function from Matlab's GPML).

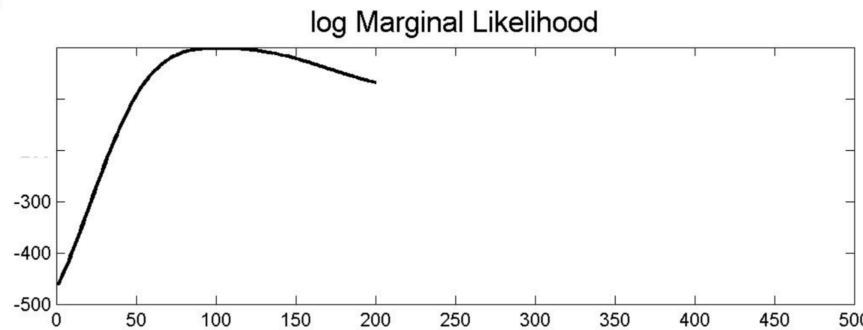
$$Cov(f(x_p), f(x_q)) = k_y(x_p, x_q) = \sigma_f^2 e^{-\frac{1}{2} l^2 (x_p - x_q)^2} + \sigma_n^2 \delta_{pq}$$

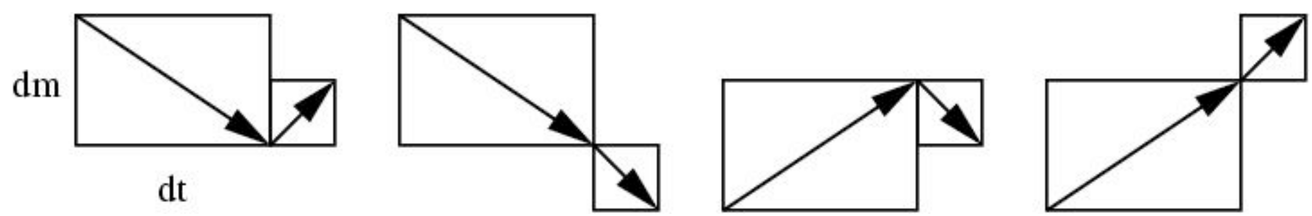


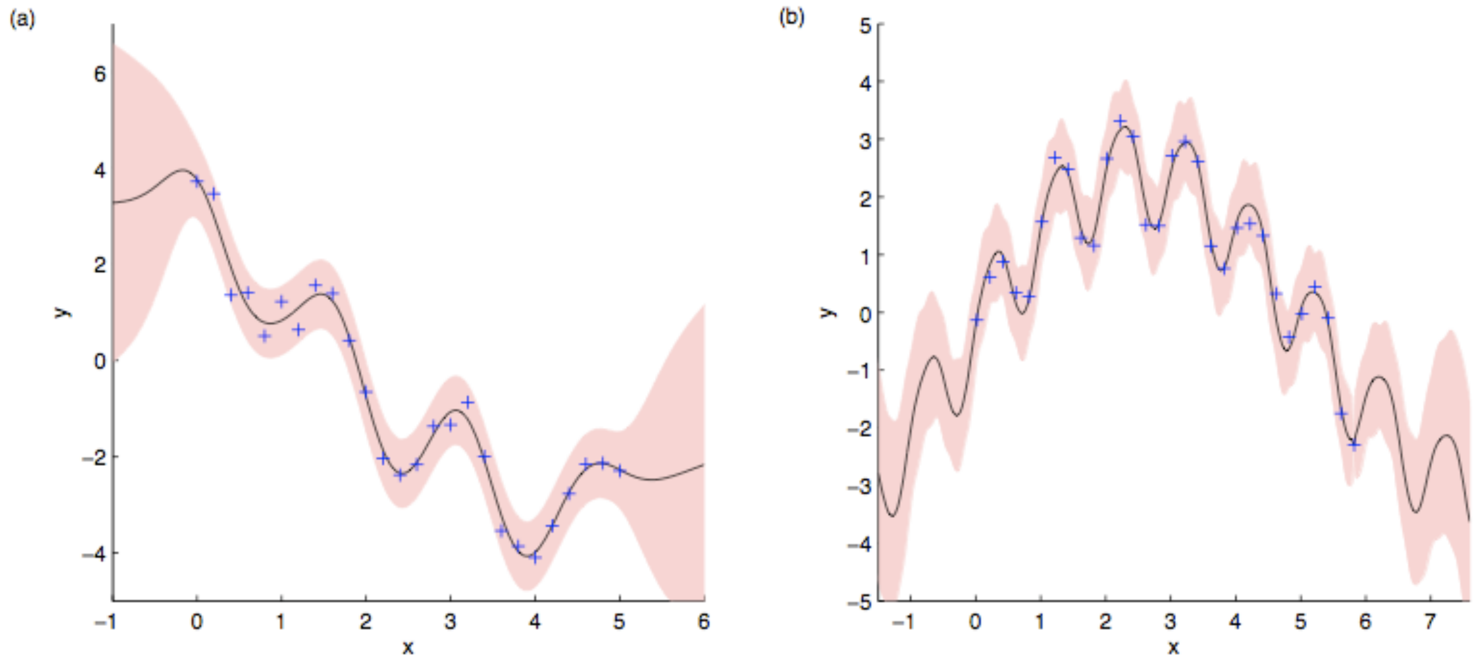
The 3 hyperparameters are "free" and are varied


Relative Fit
3500
3000
2500
2000
1500
1000
500
0
50
40
30
20
10
0
dT
0
1
2
3
4
dMag


dm
dt
Posterior Sample (residual + mean)
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5


log Marginal Likelihood
-300
-400
-500
0
50
100
150
200
250
300
350
400
450
500


dm
dt



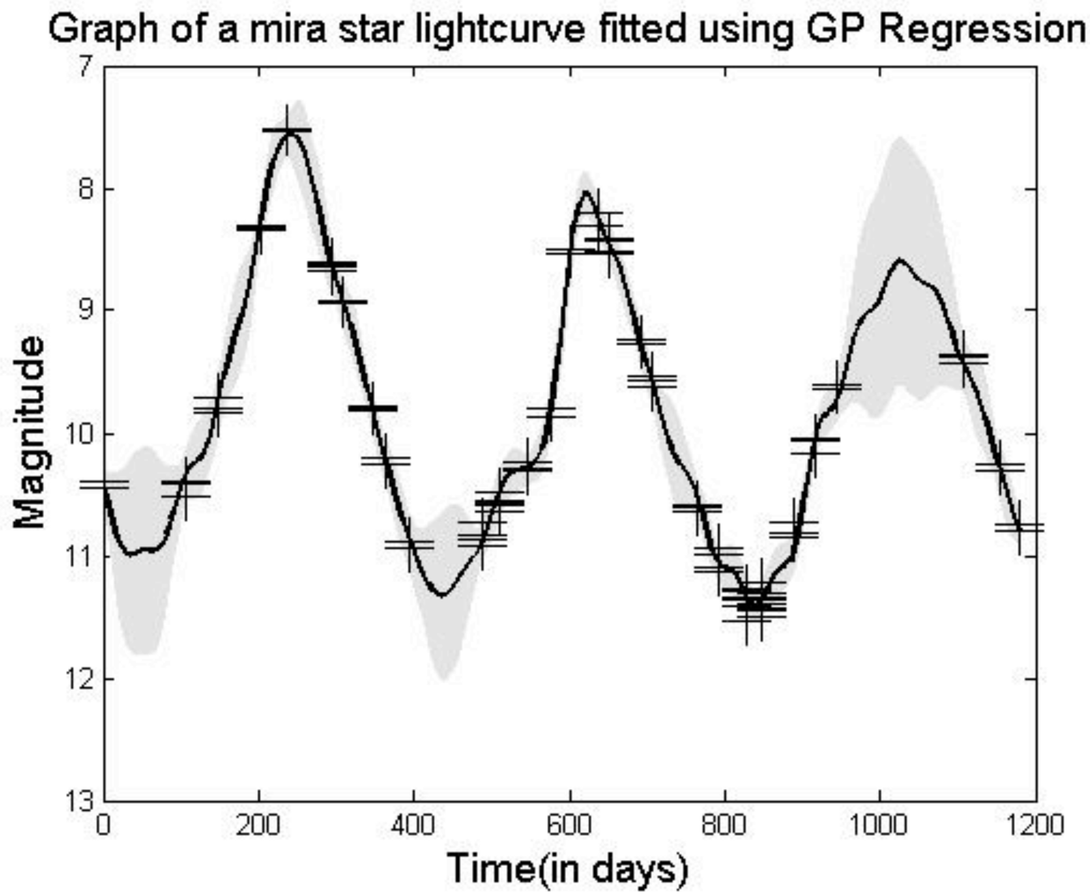
M. Ebden

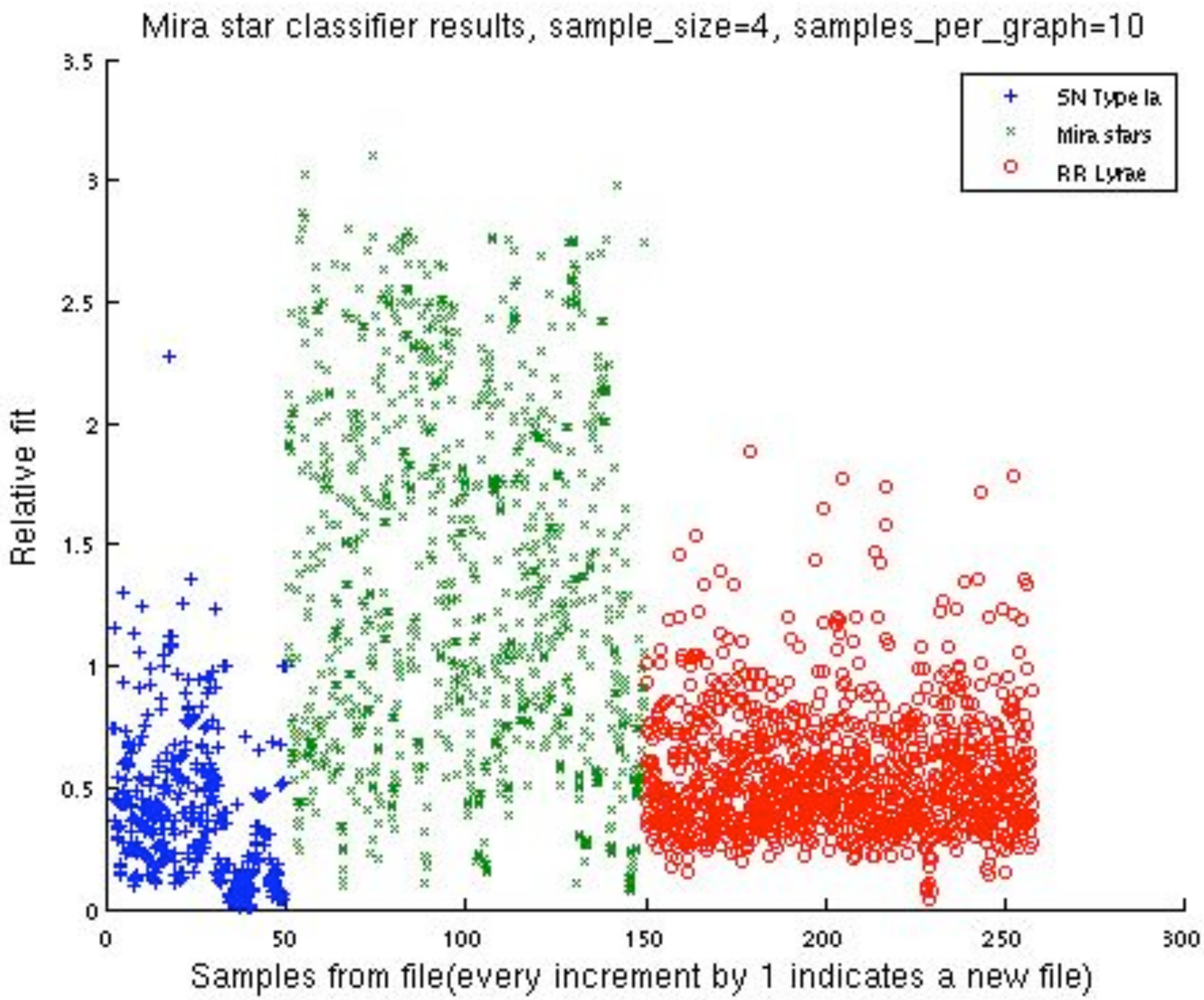
Figure 3: Estimation of y_* (solid line) for a function with **(a)** short-term and long-term dynamics, and **(b)** long-term dynamics and a periodic element. Observations are shown as crosses.

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

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

Graph of a Mira lightcurve fitted using GPR (using a function that has a periodic component).

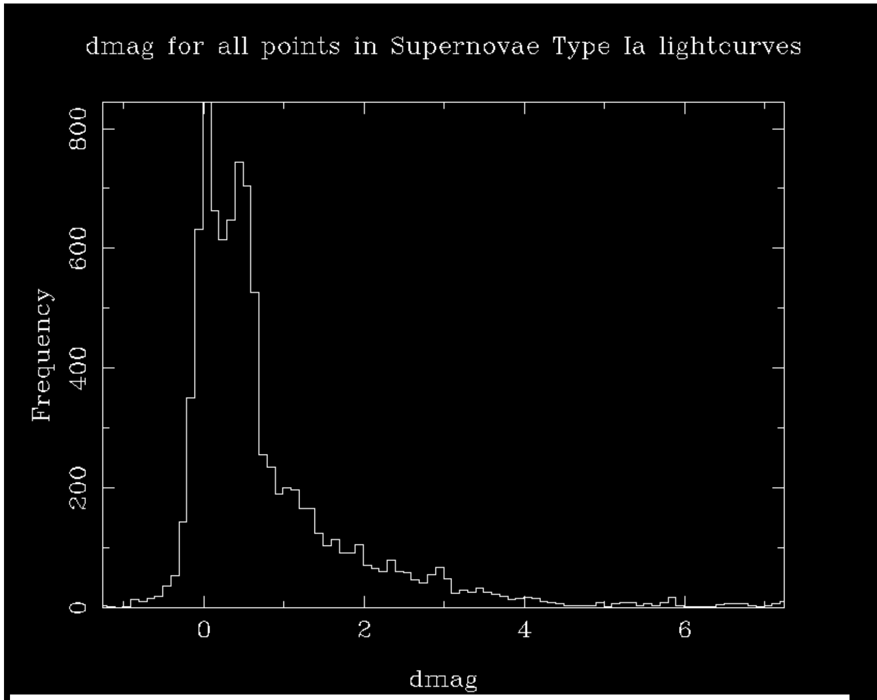


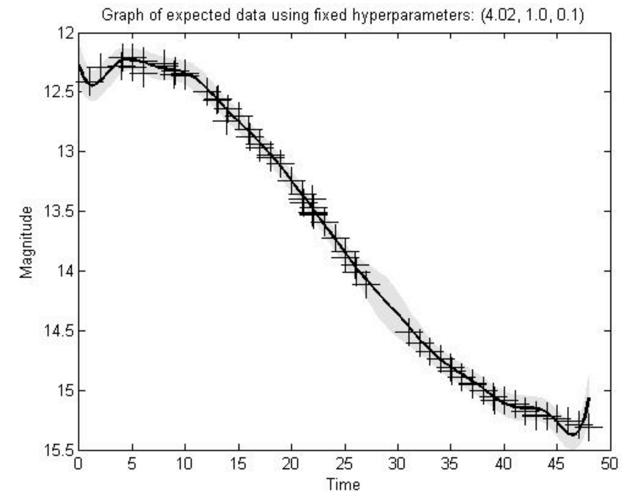

Mira star classifier results, sample_size=4, samples_per_graph=10
SN Type Ia
Mira stars
RR Lyrae
Relative fit
Samples from file(every increment by 1 indicates a new file)
0
0.5
1
1.5
2
2.5
3
3.5
0
50
100
150
200
250
300

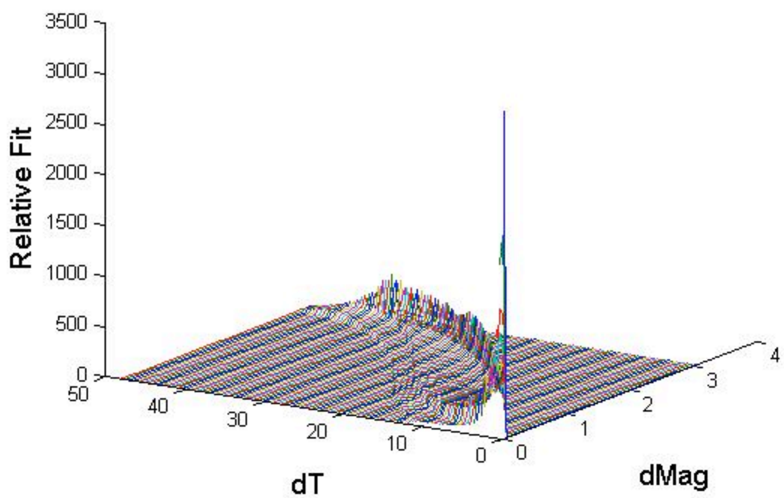
Question of normalization

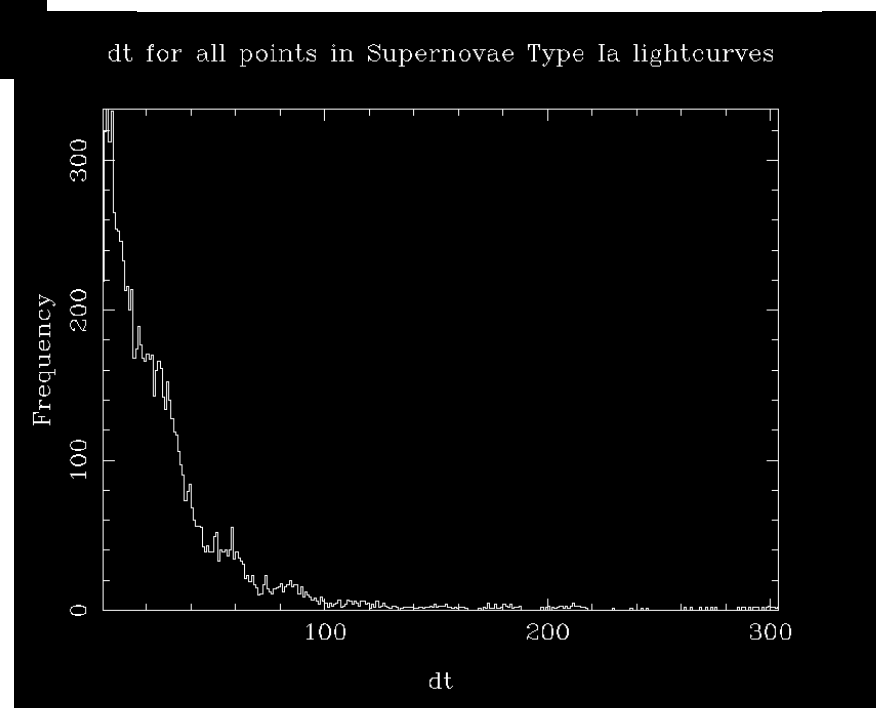
- First epoch after detection (for SN, at what distance from peak?)
- Number and frequency of epochs
- Periodic but non-constant periods

Using dMags and dT



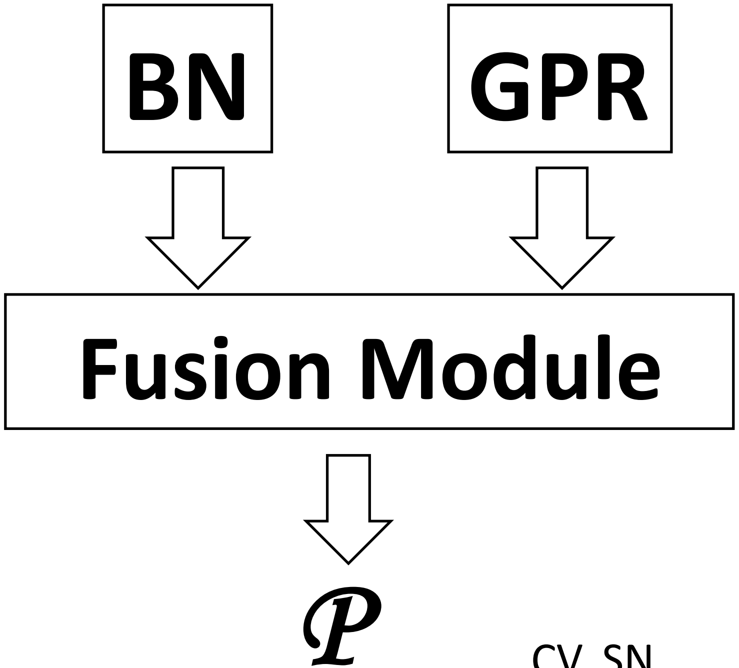


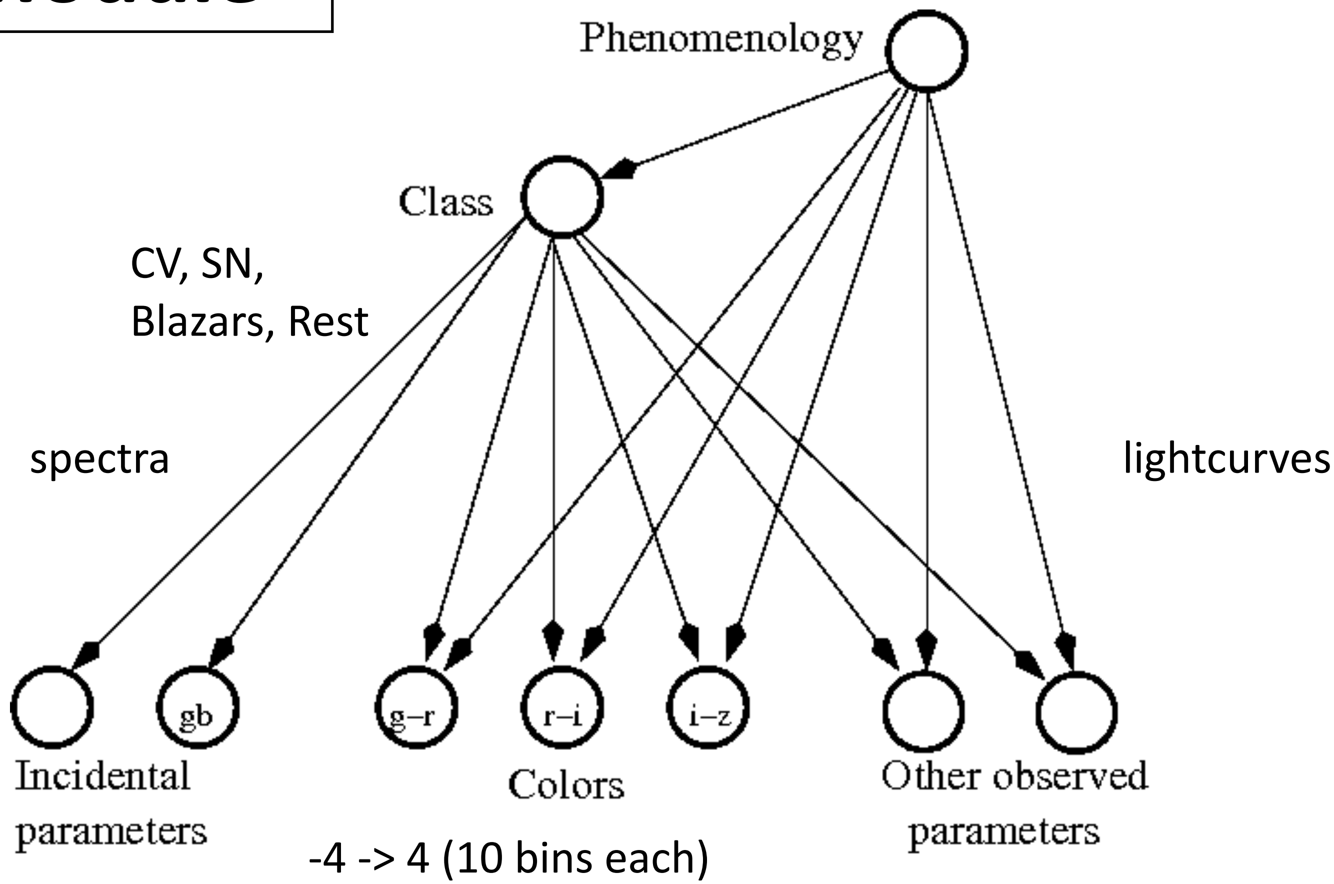




- Combining lightcurves + colors + contextual info in the right proportion can be tricky (just like the WTA, 50%, 40-10% classifications)

Next steps include combining CRTS, P60, PTF etc. data

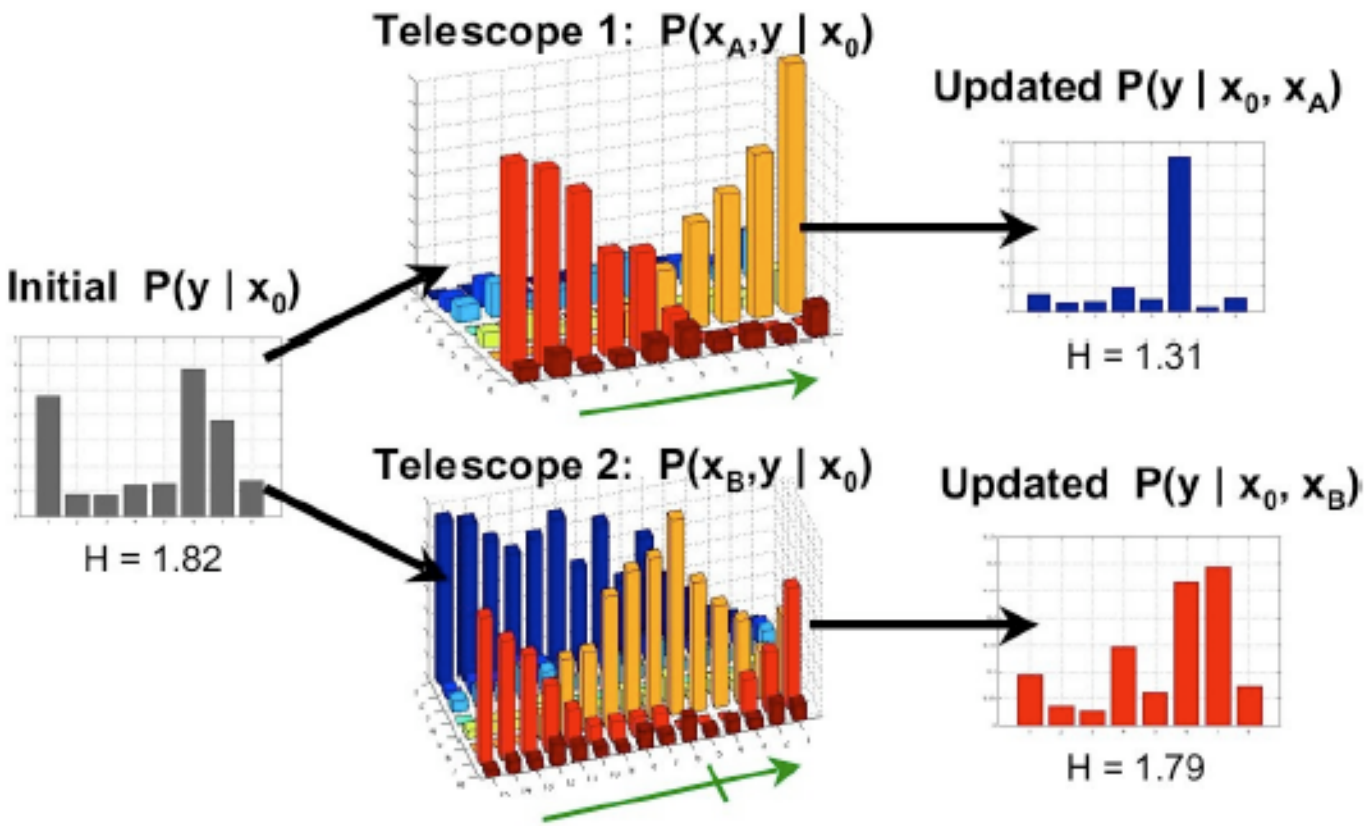




Follow-up (for missing values)

- Such that it will help discriminate better
- Serve probabilities so that consumers can choose their types of transients
- Widest possible models
- (but we need a proper listing of all follow-up resources)
- (and they need to be uniformly connected and talking to each other)

Choosing follow-up configs



r-i color, hi-z quasar, blue star

Summary questions

- Number of classes (may have to be hierarchical)
- Normalization
- Combining diverse probabilities

Preprocessing + processing = incorporating domain knowledge and knowledge about the nature of data right into the DM/AI methodologies

Classification based on minimal data -> killer app.

Synoptic, panoramic surveys → Event discovery

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

Early classification → Selective follow-up

A *very rich variety of astrophysical phenomena*: from asteroids to cosmology, from extrasolar planets to extreme relativistic physics

All interesting things are outliers in some parameter space.

Event discovery is just a start: 99% of the astrophysics is in the follow-up, and classification