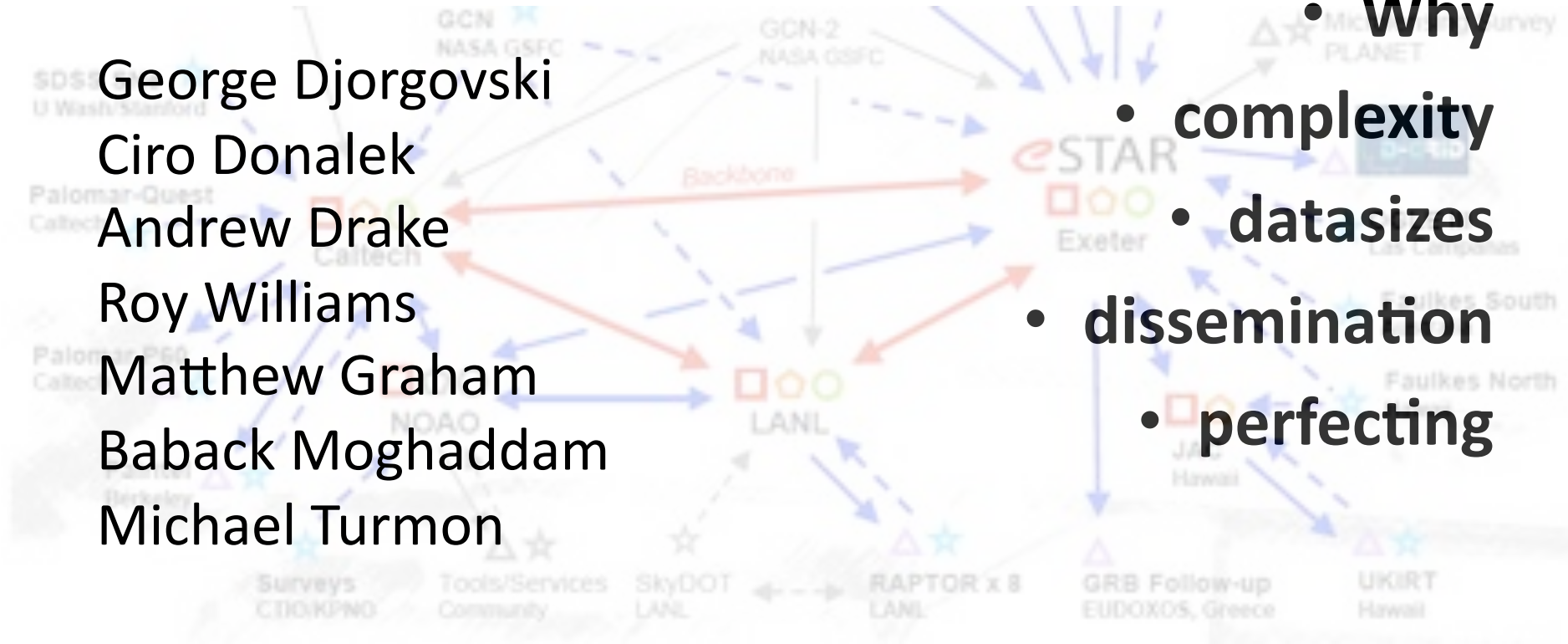
# Real-time classification of transients

http://www.mylespinkney.com/

Ashish Mahabal (aam@astro.caltech.edu),
Caltech; LSST transients working group
Colloque GAIA,
8 Jun 2010

# Real-time classification of transients: Overview

George Djorgovski
Ciro Donalek
Andrew Drake
Roy Williams
Matthew Graham
Baback Moghaddam
Michael Turmon




- **Why**
- **complexity**
- **datasizes**
- **dissemination**
- **perfecting**

# Why?

- Why real-time classification is needed
  - From 0.1 TB to tens of TBs
  - Hundreds of parameters for tens of millions of objects
  - From tens to hundreds of thousands of transients
  - Various wavelengths
  - Many kinds of studies possible (detailed structure of solar system, Galaxy and the universe at different scales and depths)
  - One of the exciting, emerging  sub-field is of temporal astronomy

# The Palomar-Quest Event Factory

Detect ~ 1 - 2 *$10^6$ sources per half-night scan

Compare with the baseline sky

Find ~ $10^3$ apparent transients (in the data)

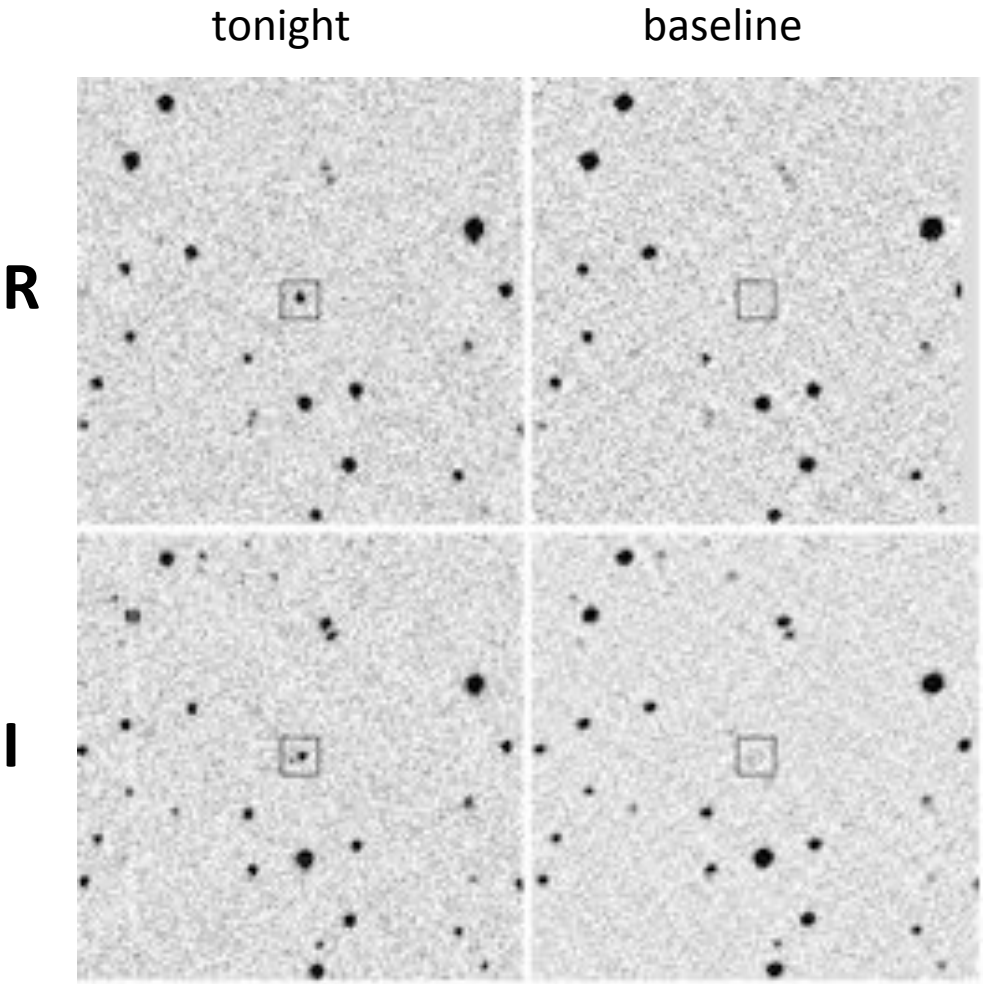
Remove instrum. artifacts

Identify ~ 2 - 4 *$10^2$ real transients (on the sky)

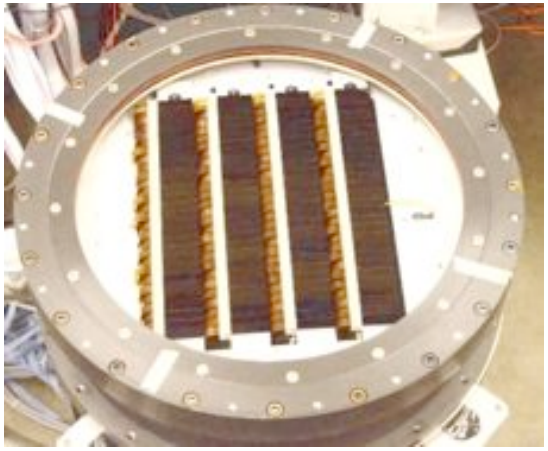
Remove asteroids

Identify ~ 1 - 10 possible Astrophysical transients

Classification and follow-up

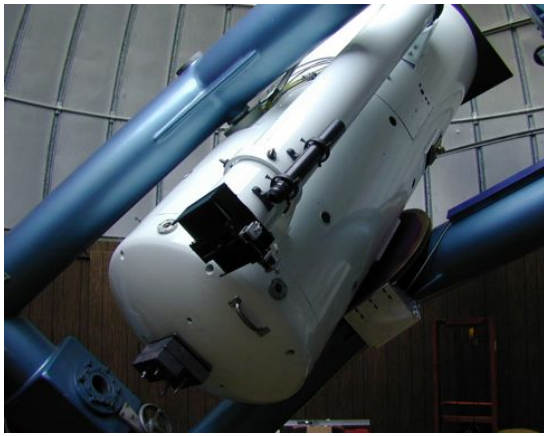


## Shallow to deep; Small to large; Sporadic to repeated; more wavelengths



PQ

CSS



GALEX, Spitzer, FIRST, …

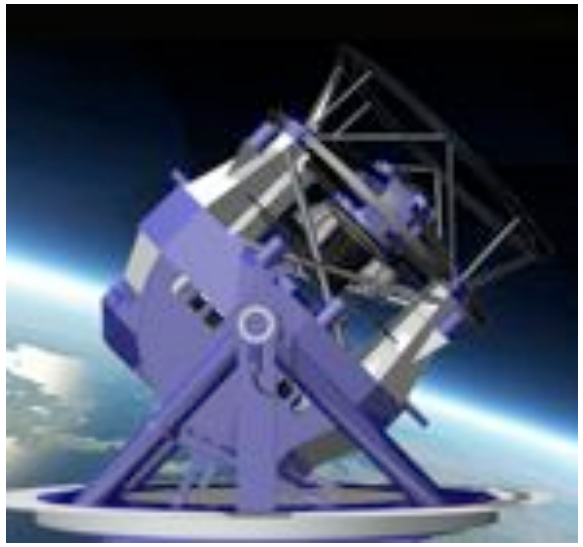
**Recent, current and future multiepoch surveys**

PTF; Skymapper; Pan-STARRS

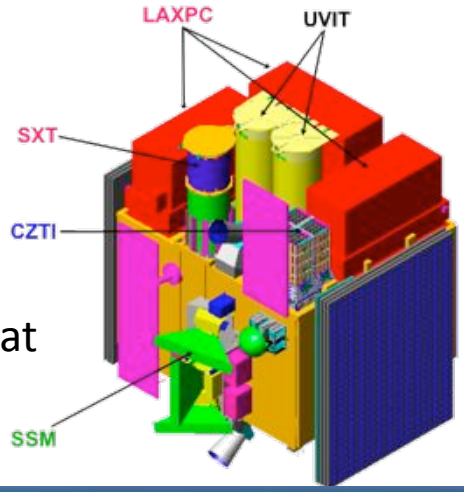
**Orders of magnitudes different.**

**Move towards digital movies!**

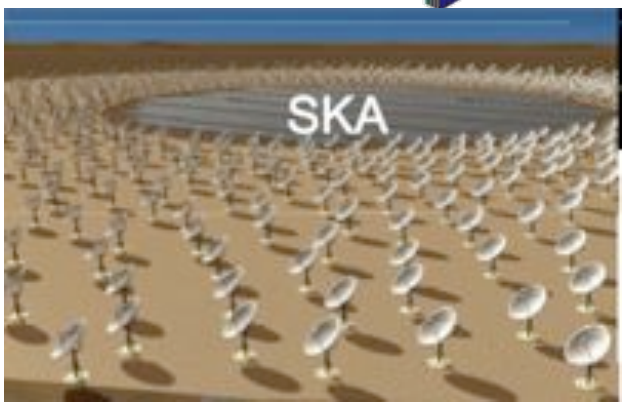
**… and of course GAIA**



LSST




Astrosat

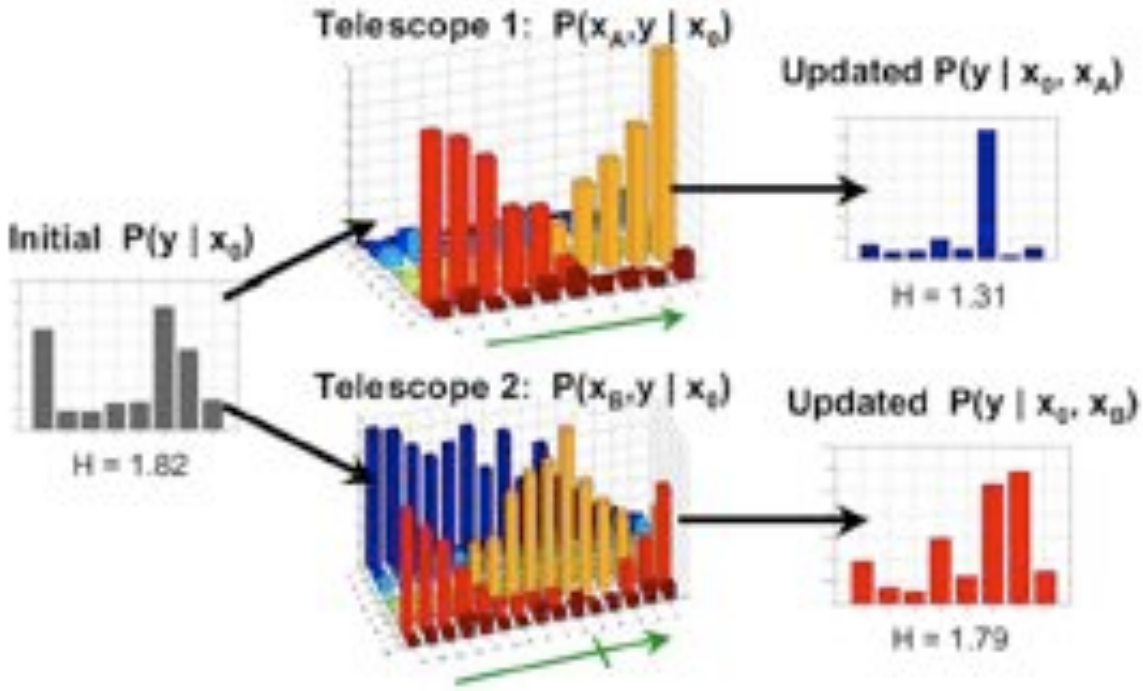


# Why?

- Why realtime classification is needed
  - From tens to hundreds of thousands of transients
  - Various wavelengths
  - **Limited follow-up opportunities (for confirmation and getting interesting physics)**
- **How fast should the response be**
  - **Selection has to be quick**
- **What observations to take needs to be decided**

# 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)



# 3 kinds of complexities

- Size of data (rows, N)
- Number of dimensions (columns, D)
- Types of applicable models (M)

b07

The dream is to reduce M to 1 for a given object (single row in a perhaps crossmatched catalog where D may have risen much higher)

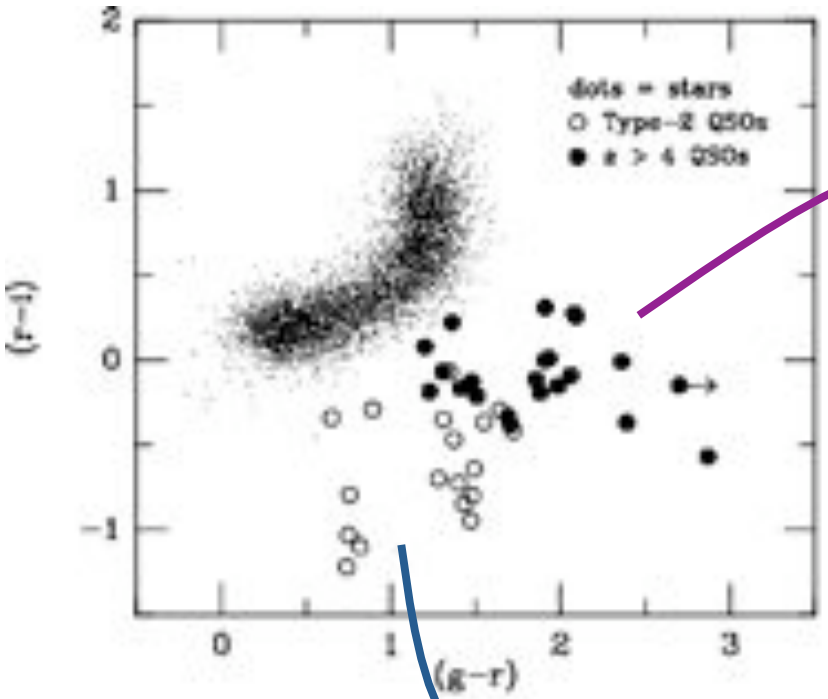
# A simple classification problem

- star – galaxy

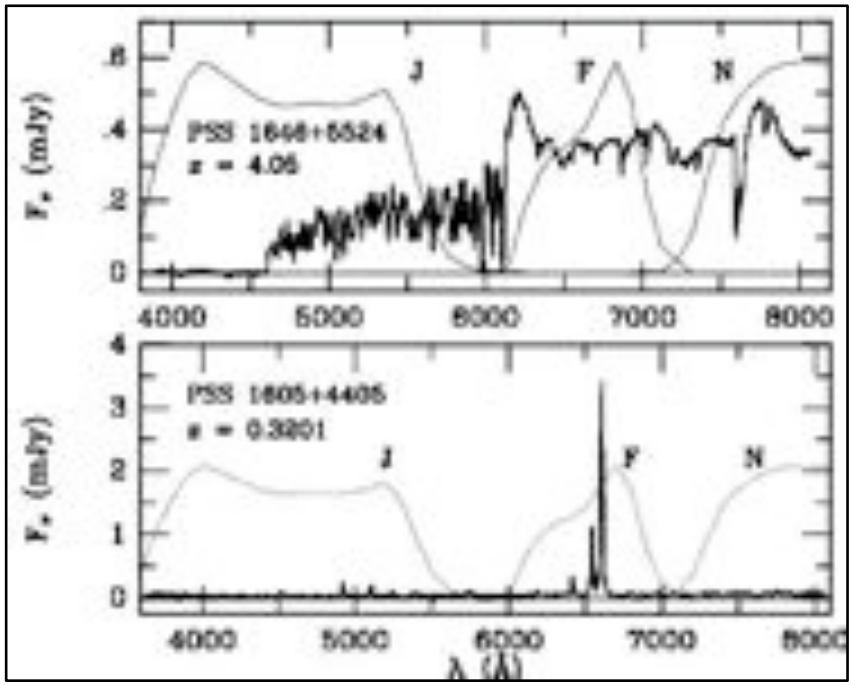
  elongation, concentration index

  M = 2; D = 2

# An Example: Discoveries of High-Redshift Quasars and Type-2 Quasars



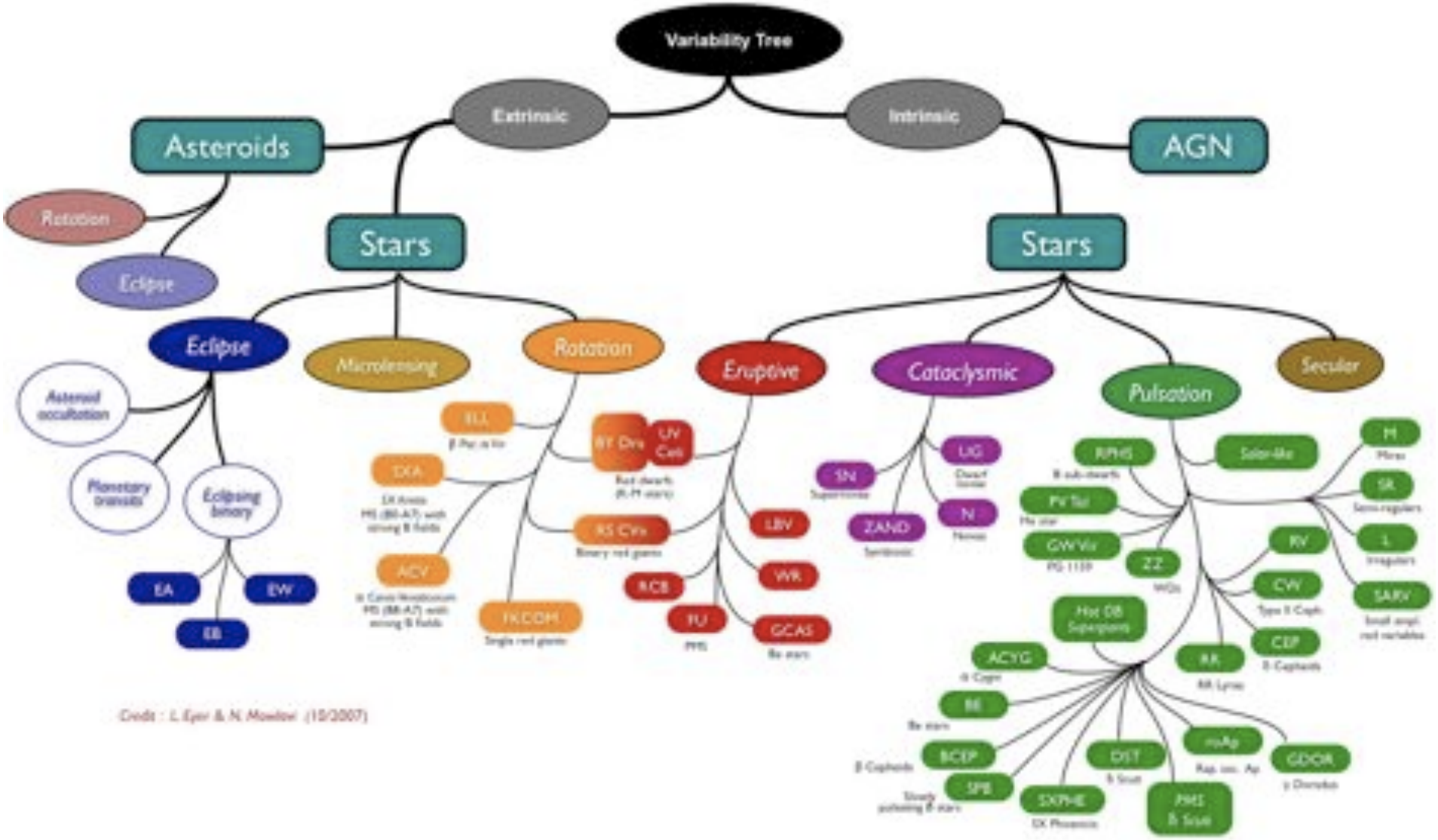

High-*z* QSO




Type-2 QSO

# Simple classification mantra

- Determine the number of classes
- Understand their properties
- Measure parameters that are handles for these properties


Variability Tree
Extrinsic
Intrinsic
Asteroids
AGN
Rotation
Eclipse
Stars
Stars
Eclipse
Microlensing
Rotation
Eruptive
Cataclysmic
Pulsation
Secular
Asteroid occultation
Planetary transits
Eclipsing binary
EA
EW
EB
DCEP
RR
GDOR
SPB
DSCT
SXPHE
PMS
Credit : L.Eyer & N. Mowlavi (10/2007)

# Complications

- How many classes are there?
- Are they cleanly separated?
  - Brighter stars
  - Distant galaxies
  - Grazing cosmic rays
- Do all objects belong to these classes?
- Could we add observables to determine the class better?

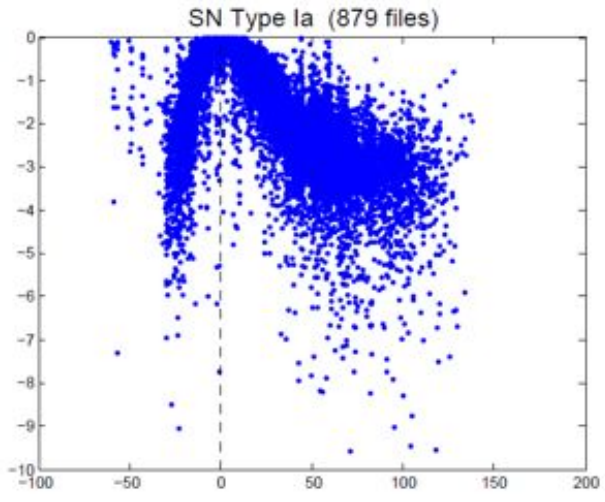
# General Catalog of Variable Stars (GCVS)

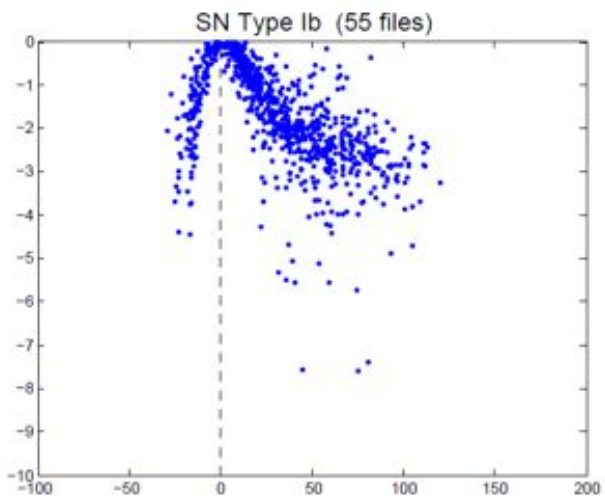
http://www.sai.msu.su/groups/cluster/gcvs/intro.htm

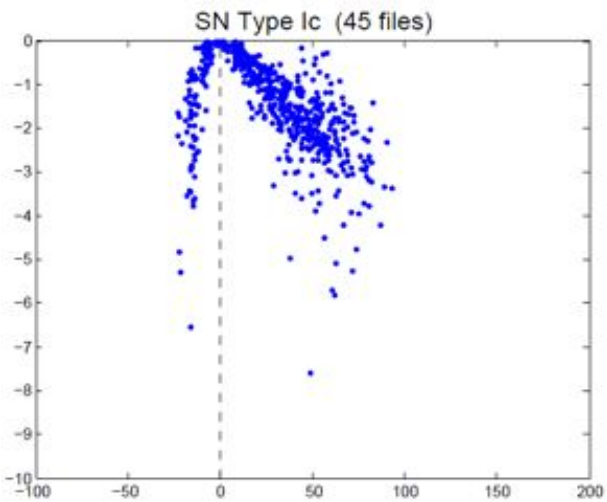
- 40000 variables
- Magnitude range known (for most)
- Periods known for over half
- Duty cycle known for a good fraction
- 500 (!) subclasses

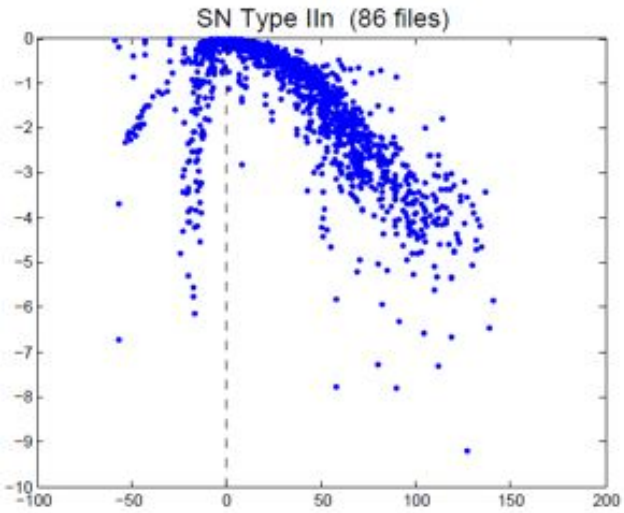
132 Mira from ASAS:
Period: 55 – 523 days
Peak mag: 4.89 – 13.29
Amplitude: 1.63 – 6.65

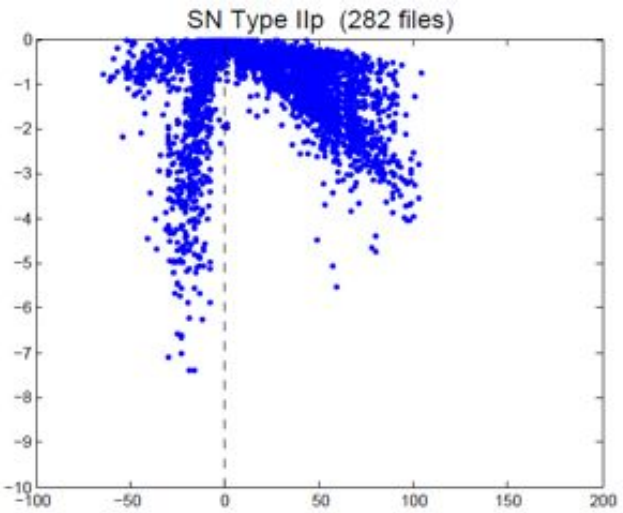
SN challenge and
Improving priors


SN Type Ia (879 files)


SN Type Ib (55 files)


SN Type Ic (45 files)


SN Type IIn (86 files)

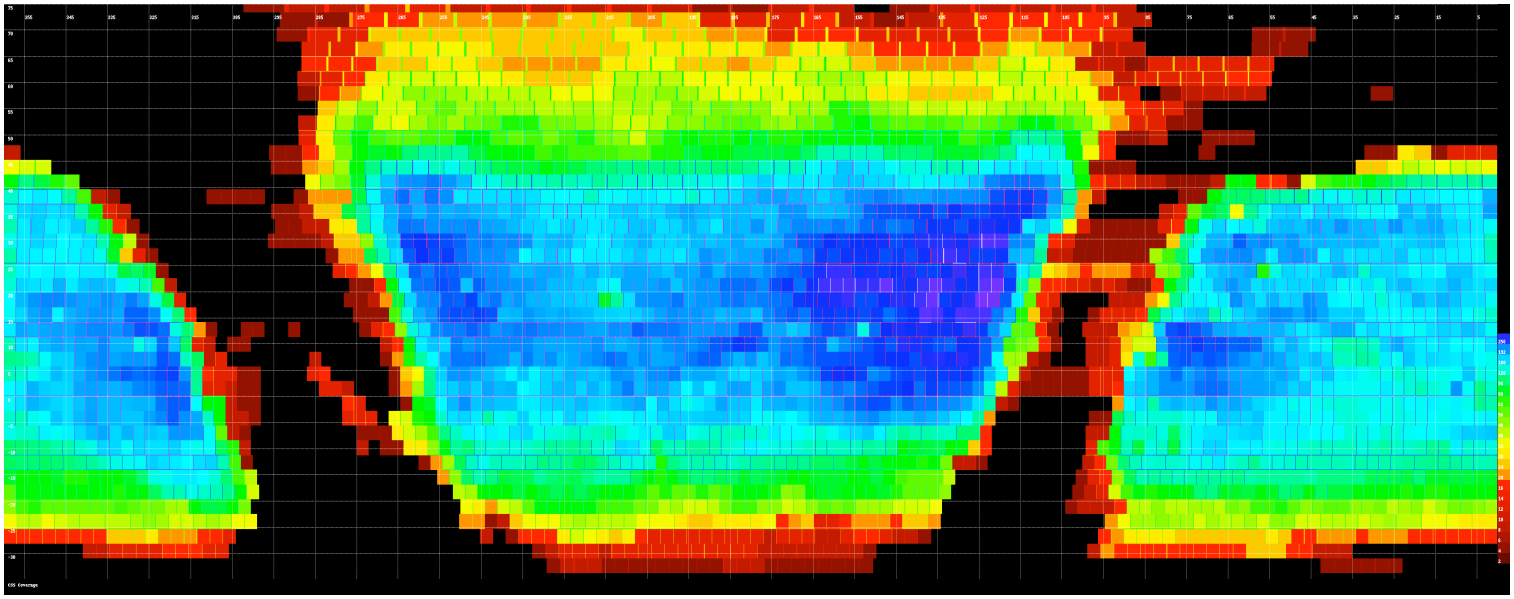

SN Type IIp (282 files)

# Existing surveys and cadence

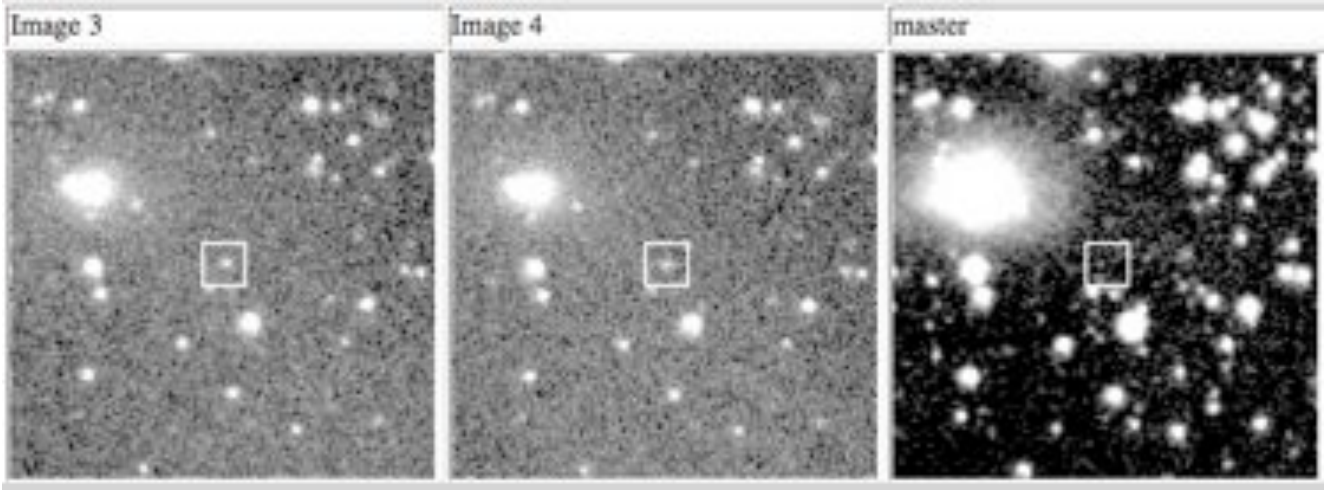
- PQ – 4 colors, nearly simultaneous
  - That's where our real-time work began
- CRTS: 4 images, 10 minutes apart, no filter
  - 3 telescopes now, including one from Australia
  - Follow-up colors with Palomar 60-inch
  - **Some aspects similar to GAIA**
- Applying some of the aspects to PTF
- (skymapper, Pan-STARRS), LSST to follow

# CRTS

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



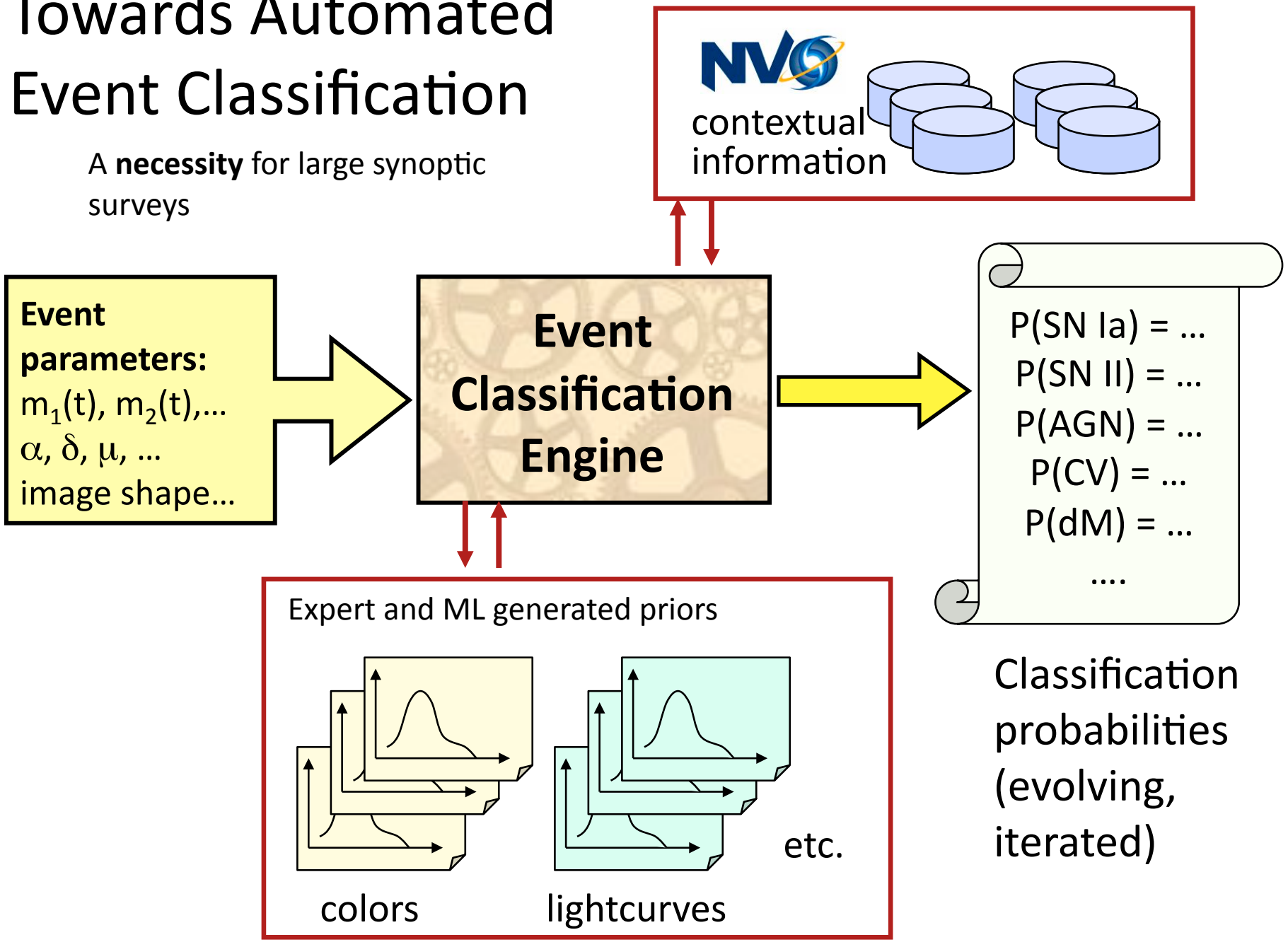
- **Over 1500 optical Transients so far.**
- **~70% SNe and CVs**
- Others include Blazars, AGN, variable stars and flare stars, and HPM stars.




SN z=0.05
CSS 20090711

# Towards Automated Event Classification

A **necessity** for large synoptic surveys



# Transient classification mantra

- Obtain a couple of epochs in one or more filters
- Assign probabilities for different classes
- Choose [to do more] observations (filters, wavelengths) for best discrimination
- Feed the new observations back in
- Revise probabilities, choose observations, …
- Based on confirmed class revise priors

# Different techniques

- Various machine learning techniques
- Highlighting Bayesian Network since it is robust even with missing data
- Other: GPR, RF, NN
- Combining with a fusion module
- Emphasis: early classification based on minimal data which can then be gradually improved (feeds into optimal follow-up as well)

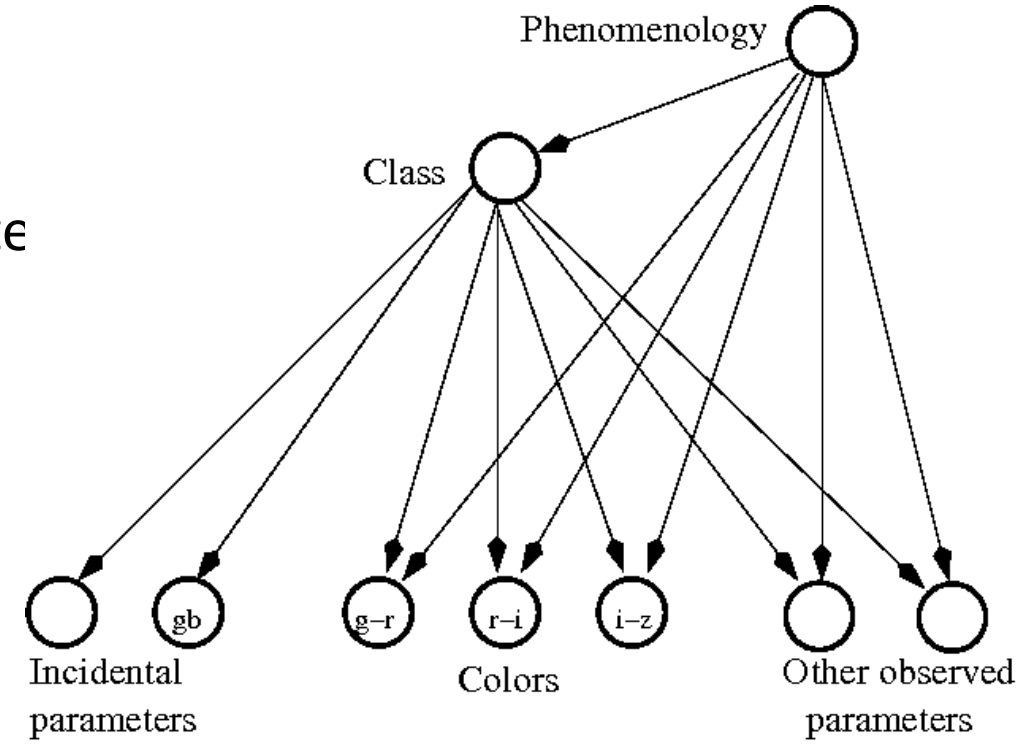
# Naïve Bayes

$$P(y = k \mid x) = P(x \mid y = k)P(k)/P(x) \propto P(k)P(x \mid y = k) \approx P(k)\prod\prod_{b=1}^{B} P(x_b \mid y = k)$$

- x: feature vector of event parameters
- y: object class that gives rise to x (1<y<k)
- Certain features of x known: (position, flux)
- Others will be unknown: (color, delta-mag)
- Assumption: based on y, x is decomposable into B distinct independent classes (labeled $x_b$)
- This helps with the curse of dimensionality
- Also allows us to deal with missing values

# Building Bayesian Networks

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




Radio, for instance

# 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 an 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 |

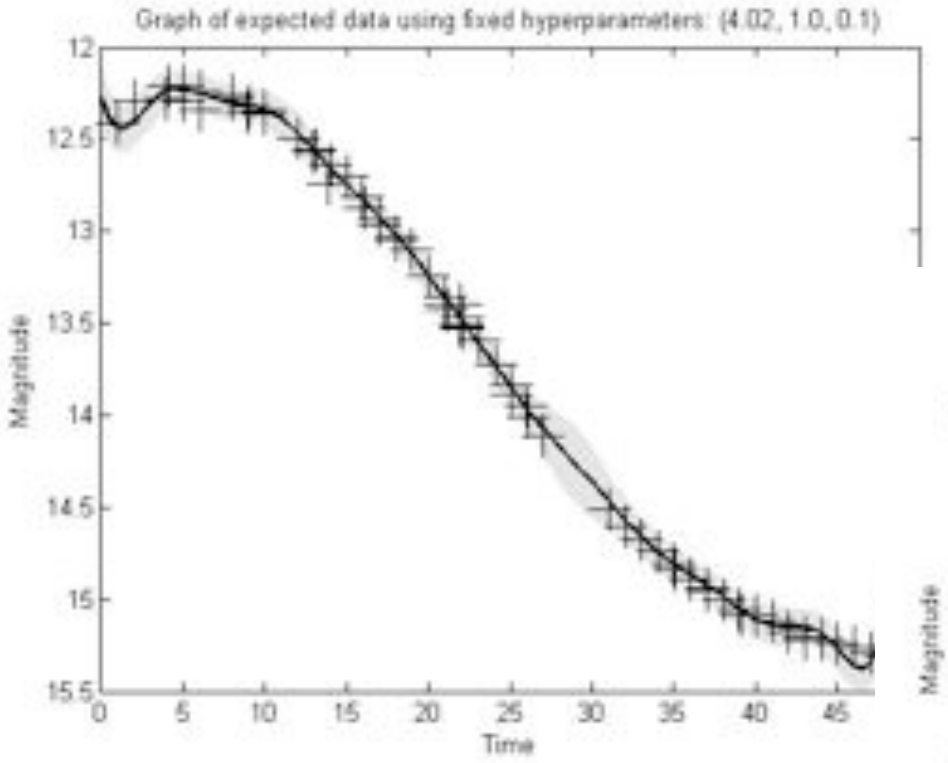
- Number of classes
  - The REST class
    - If too few classes used, too many objects get classified wrong
    - If a large number of classes used, probabilities split into small fractions unless classification is unambiguous (SN subtypes, for instance)
  - Uniformity of priors
    - Number of objects
    - Their magnitude range
    - Spread over time

    Changes:

    delta-t after transient

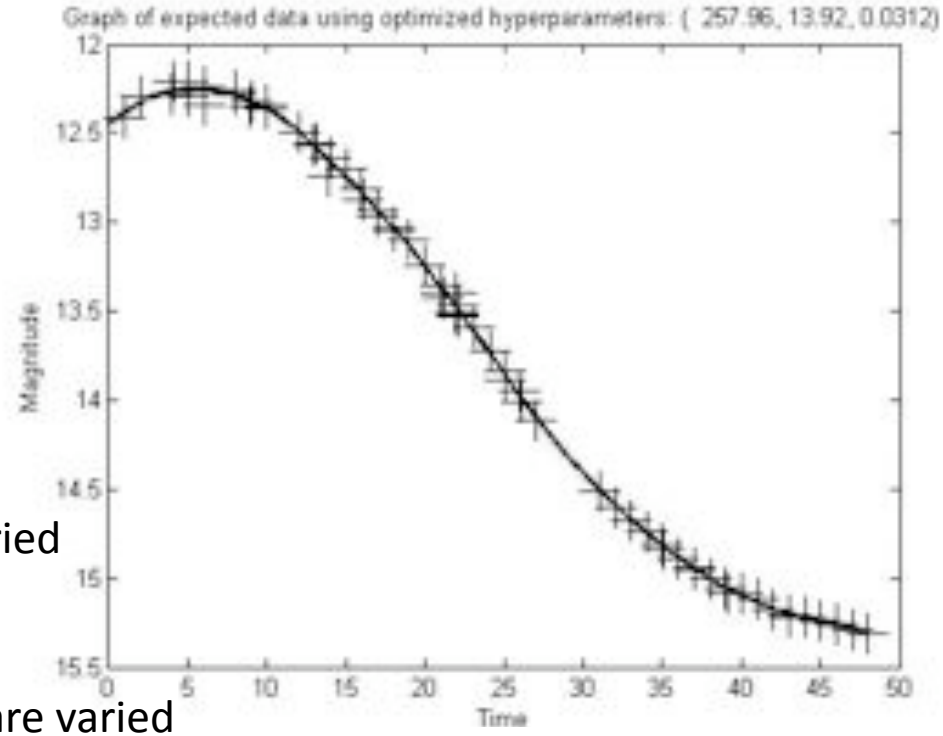
    delta-t between filters
- Sky truth?
- Application of CRTS priors to other surveys

# Using GPR with lightcurves



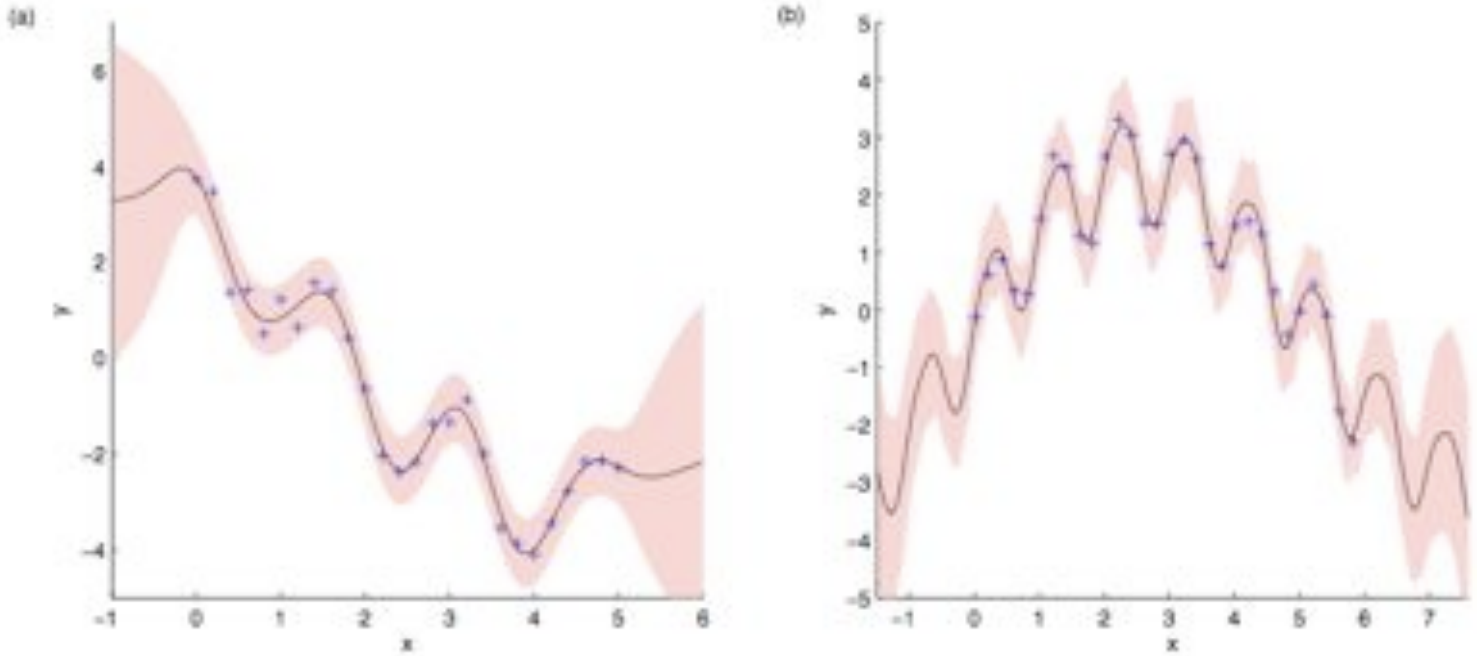

Given several epochs and corresponding magnitudes, estimate the likelihood of a particular magnitude for a new epoch (using some covariance function)

The 3 hyperparameters are “free” and are varied




The 3 hyperparameters are “free” and are varied

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

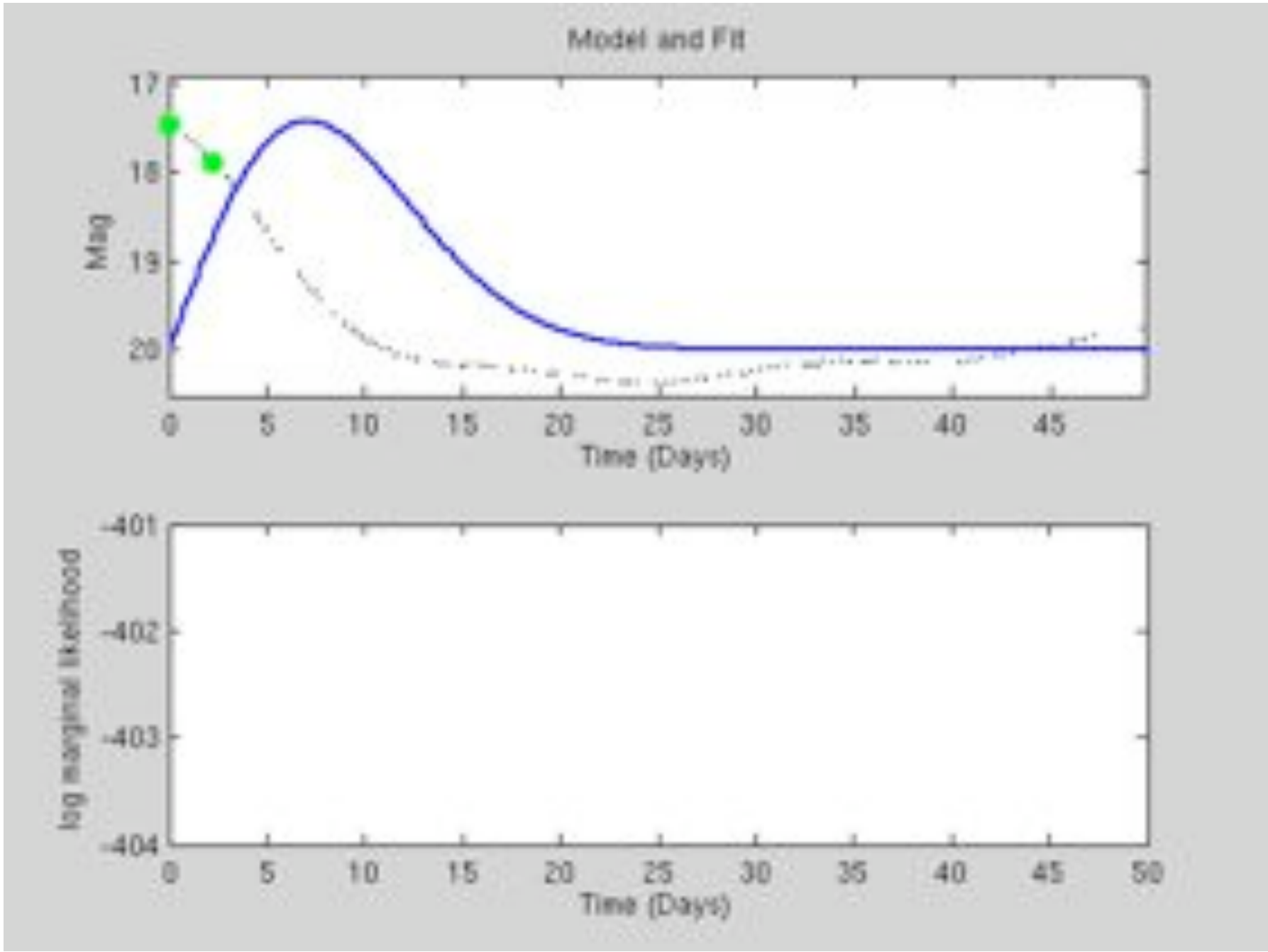


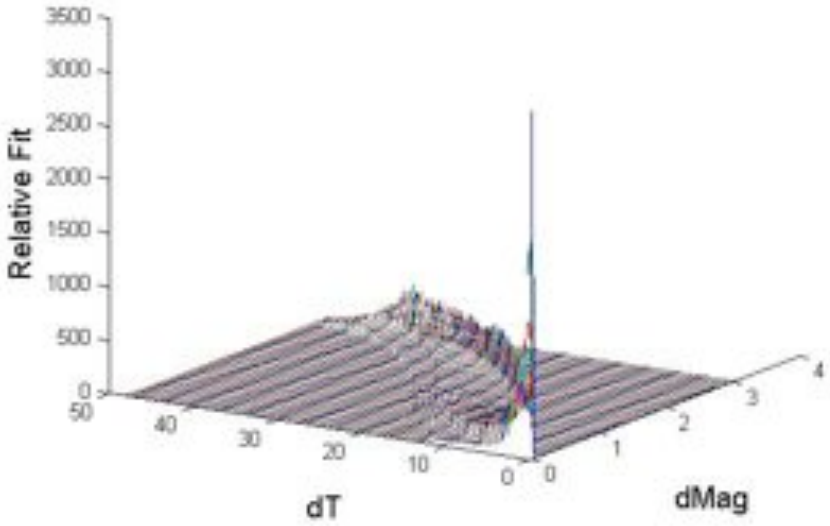

M. Ebden

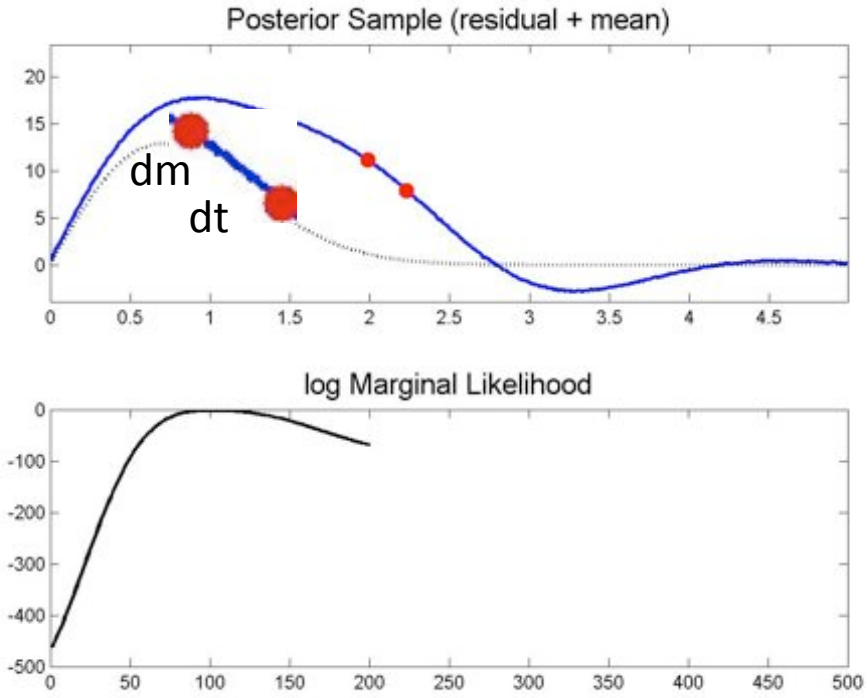
Figure 3: Estimation of $y_*$ (solid line) for a function with **(a)** short-term and long-term dynamics, and **(b)** long-term dynamics and a periodic element. Observations are shown as crosses.

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

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


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)




Difficult tool to train for periodic variables.

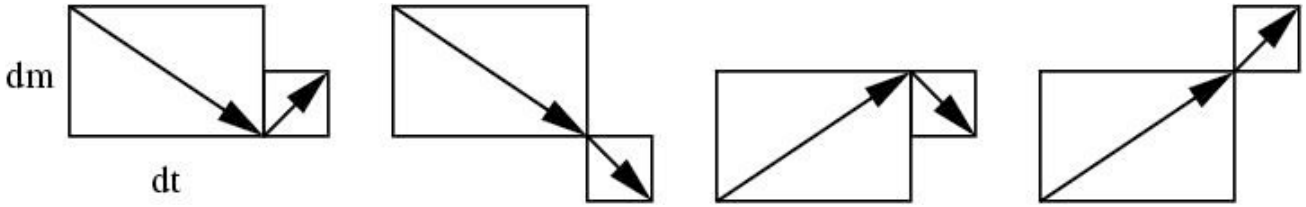
132 Mira from ASAS:

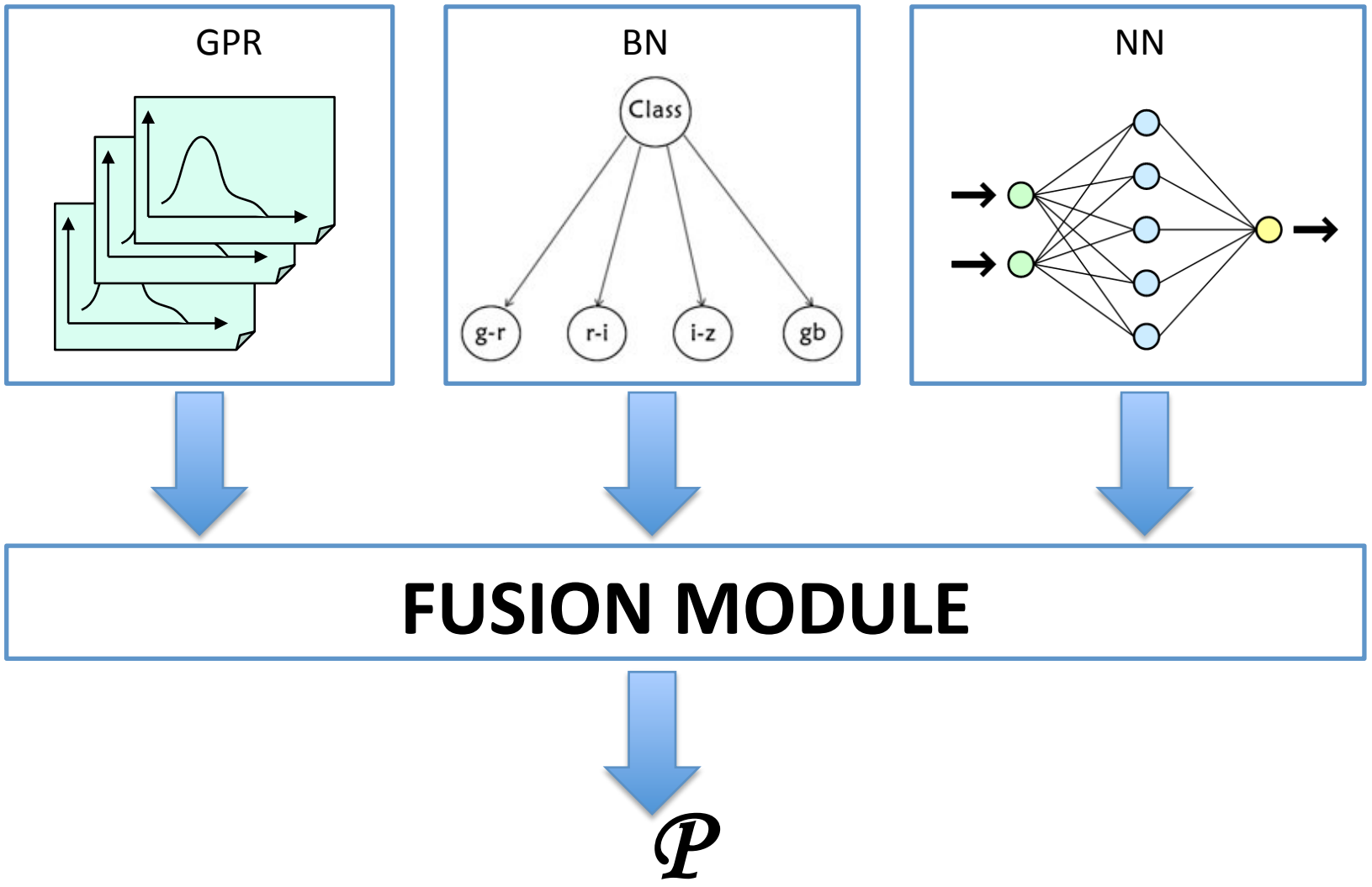
Period: 55 – 523 days

Peak mag: 4.89 – 13.29

Amplitude: 1.63 – 6.65

More points => better performance



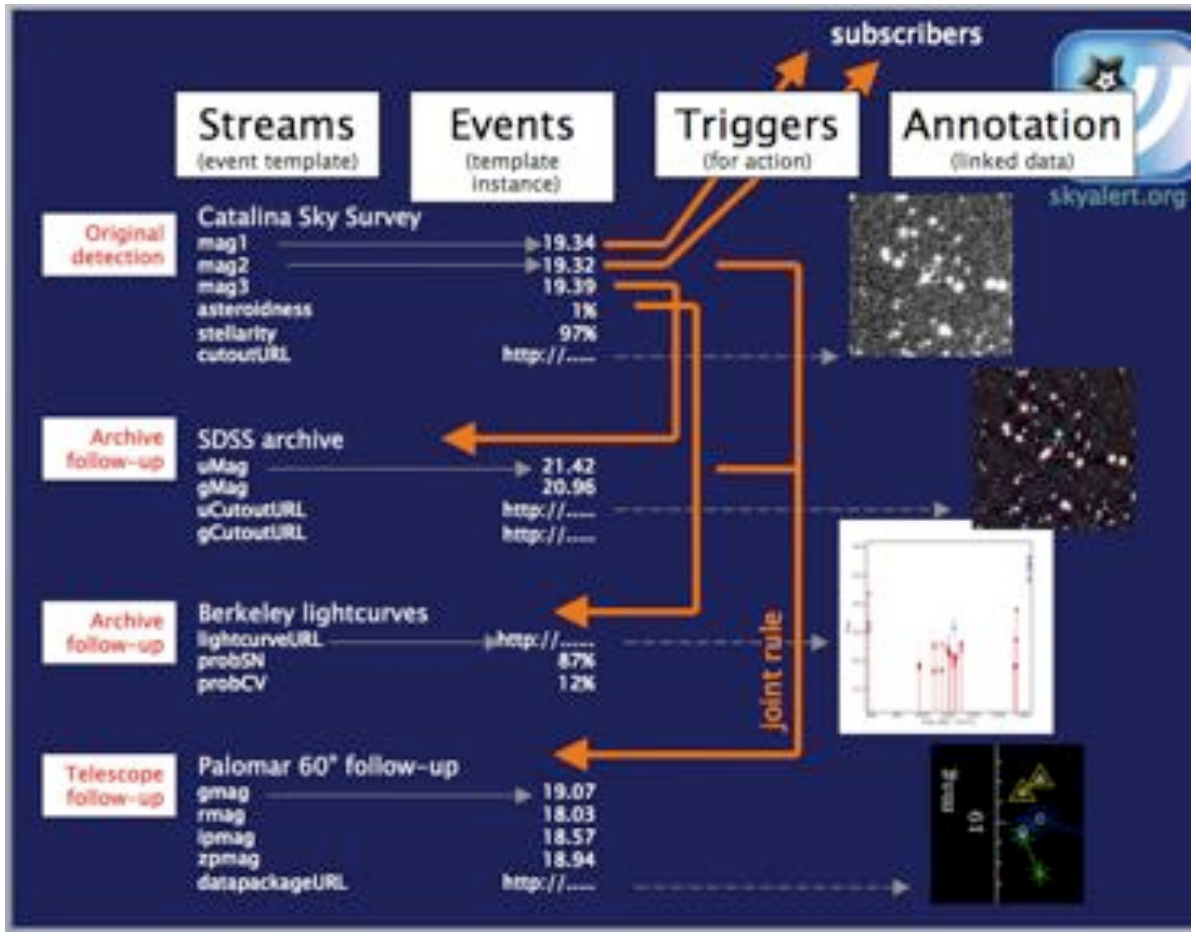

GPR
BN
Class
g-r
r-i
i-z
gb
NN
FUSION MODULE
$\mathcal{P}$

# Disribution

- VOEventNet
  - Jabber clients (perl, c, java)
- Skyalert
  - Email alerts with xml links
  - Possible to put filters at users end
  - Streams, annotations, portfolios
  - Active and passive follow-up (e.g. radio data)

# Portfolios, semantic linking and skyalert (http://www.skyalert.org)

- Active follow-up
  - New images
  - New colors
  - Better astrometry
  - Spectra
- Passive follow-up (annotators)
  - Galaxy distance
  - Classification
    - Program
    - Expert



# Annotator examples

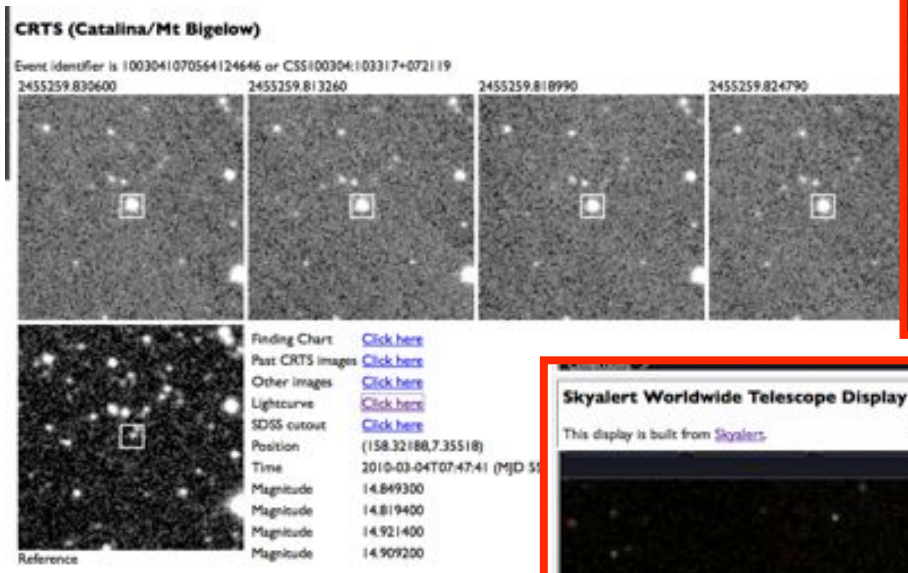
- add intelligence
  - input is lightcurve, or one or more colors
  - output is likelihood for known object classes
- get data
  - input is position
  - output is survey catalogs and cutouts and parameters like distance to nearest galaxy
- telescopic followup
  - Connect to scheduler
  - output can be linked back

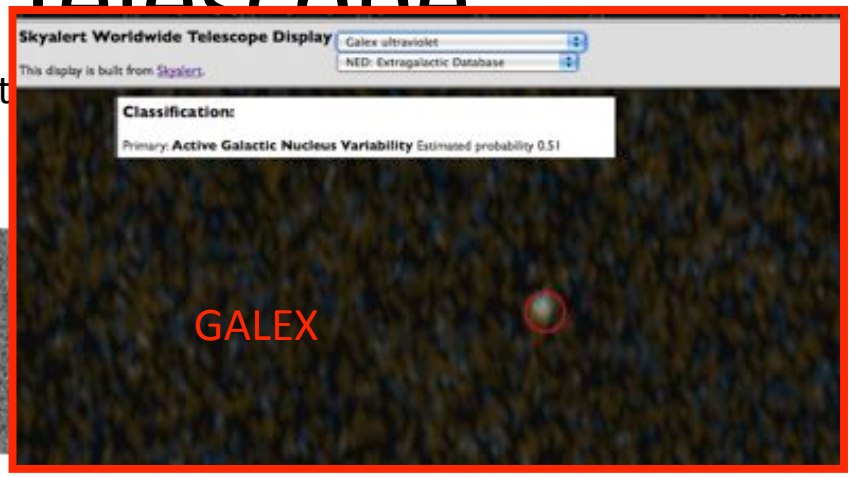
# iPhone app

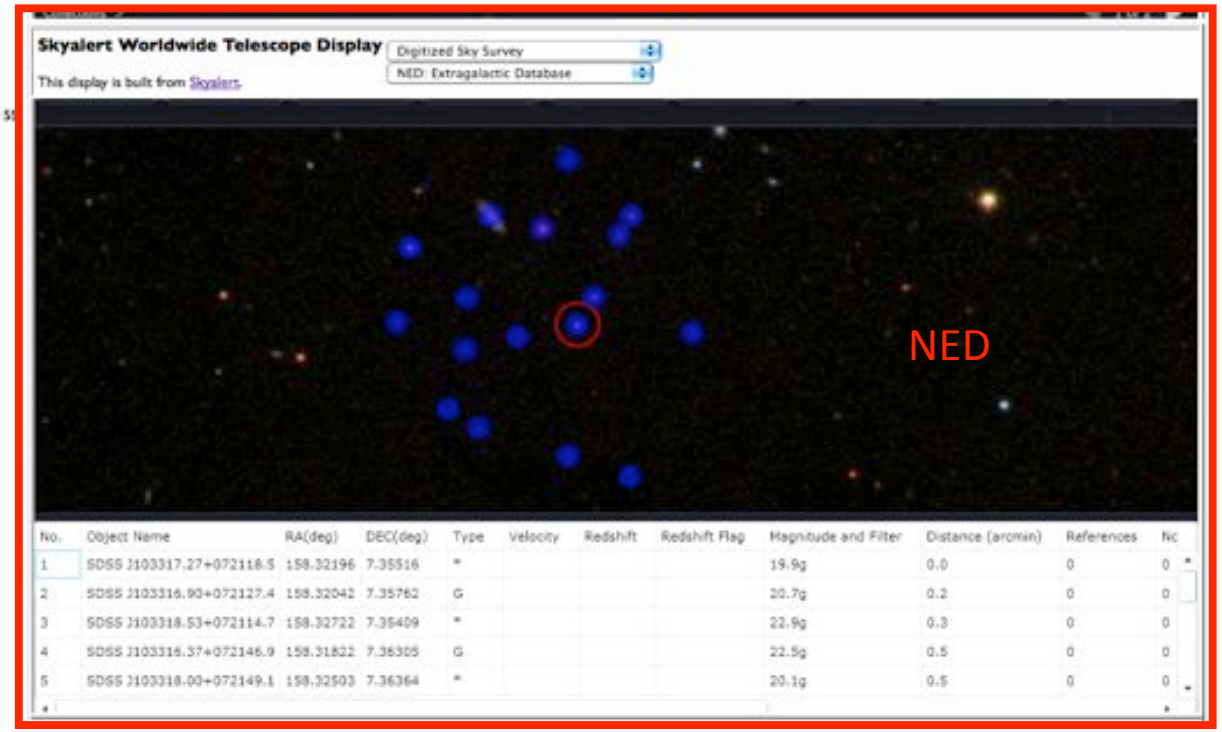
## “Transient Events”

# Worldwide Telescope

view of event

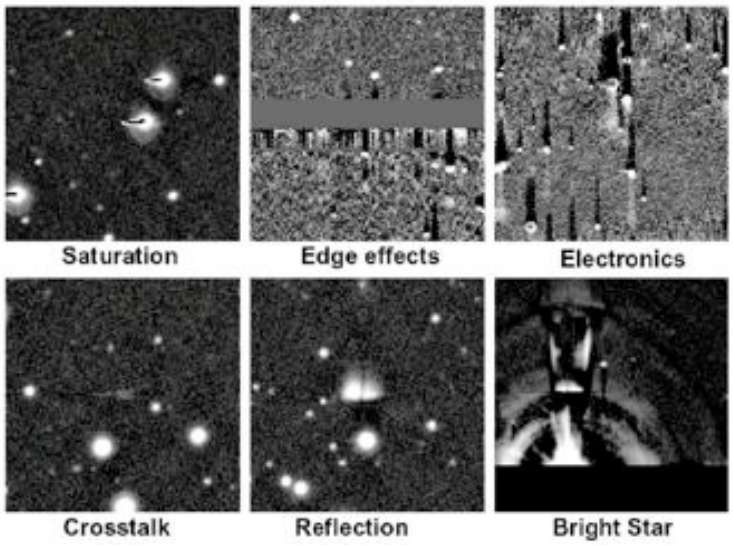


## Perfecting with human power

- A good combo of machines and humans
  - Present humans with choices difficult for machines (class boundaries, for instance)
  - Learn from their response to make better classifiers




- Uniform large priors at different flux levels need to be put together
- Some may come from ongoing programs, but specific campaigns may have to be run for more
- Especially to ensure that faint flaring variables do not get mistaken for rarer classes of transients.

**Natural and artificial classifiers; varied inputs; unified output**
**Difficult but exciting problem**
**To boldly classify what no one has classified before …**

# GAIA connection

- 1 billion objects
- G, BP, RP – will help narrow classes
- Variability models (deep coadded images)
- Existing telescopes can start follow-up
- Priors of objects of interest
- Richer portfolio semantics

# Bottomline

- Real-time classification of transients is important
- A lot of work on classification is going on
- Some aspects are similar to those of GAIA
- We would love to collaborate, get involved and contribute
- aam@astro.caltech.edu