Classification with Sparse Timeseries

Ashish Mahabal
aam@astro.caltech.edu
7 Sep 2011
HEAD 2011, Newport, RI

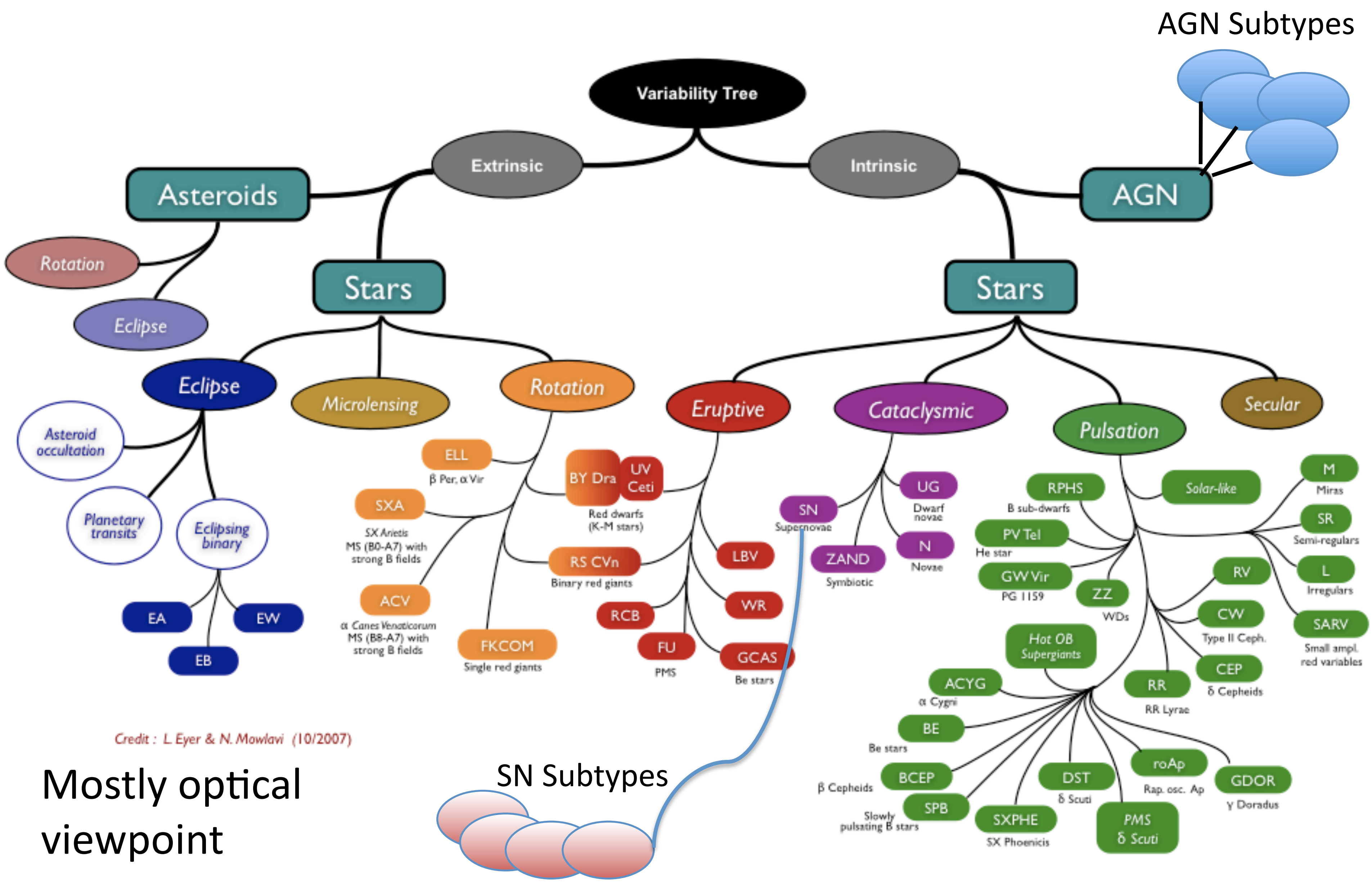
Collaborators

- Caltech
  - George Djorgovski
  - Ciro Donalek
  - Andrew Drake
  - Matthew Graham
  - Roy Williams
- JPL
  - Baback Moghaddam
  - Mike Turmon

Plus at various other institutes all over, but especially in US, India and Italy

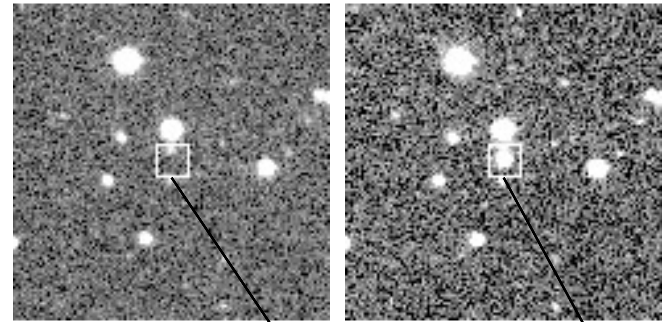
http://pardington10.wikis.birmingham.k12.mi.us/Collaboration+Techniques

Semantic Tree of Astronomical Variables and Transients



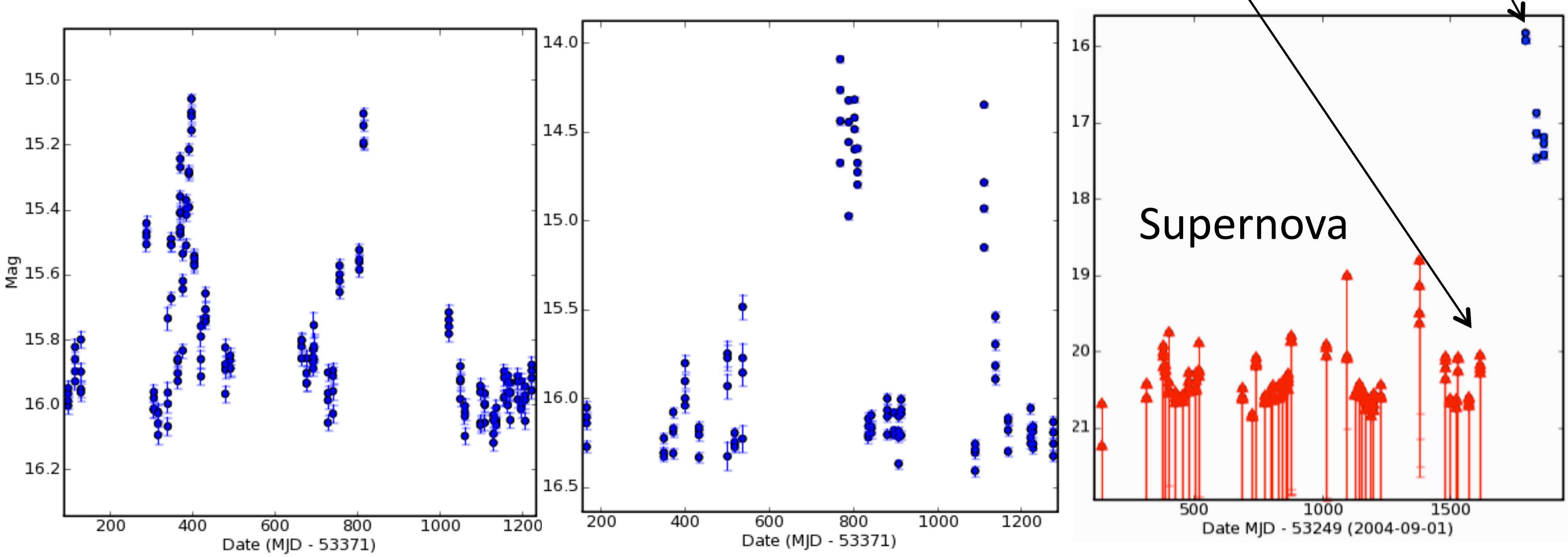
Mostly optical viewpoint

Sample Light Curves



Blazar PKS0823+033

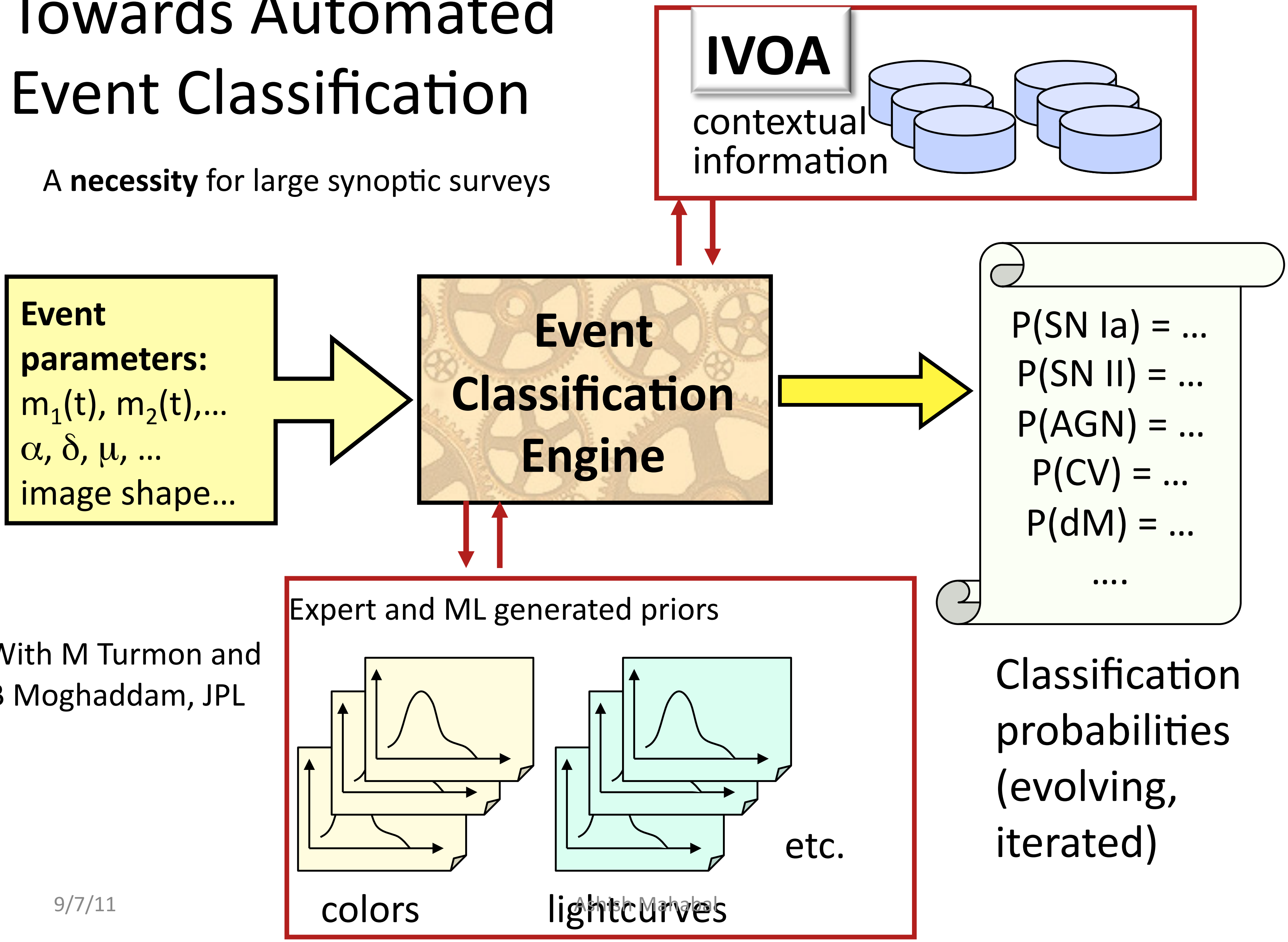
CV 111545+425822



Variables and transients – the distinction is one of perception, and your aims

Towards Automated Event Classification

A **necessity** for large synoptic surveys



With M Turmon and B Moghaddam, JPL

Making optimal use of sparse data sets

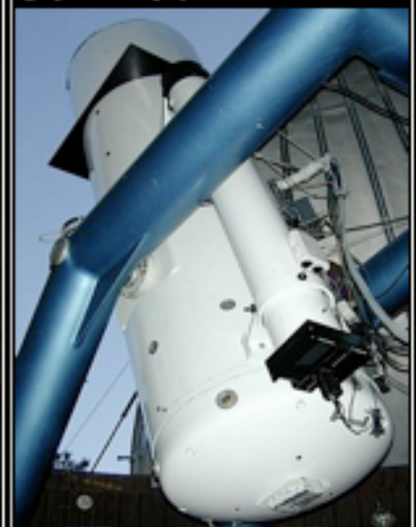
- sparse light curves
  - analysis of different types
- few colors/sparse SEDs
- any contextual information
- priors for different kinds of objects

Holistic approach

Catalina Sky Survey(s):

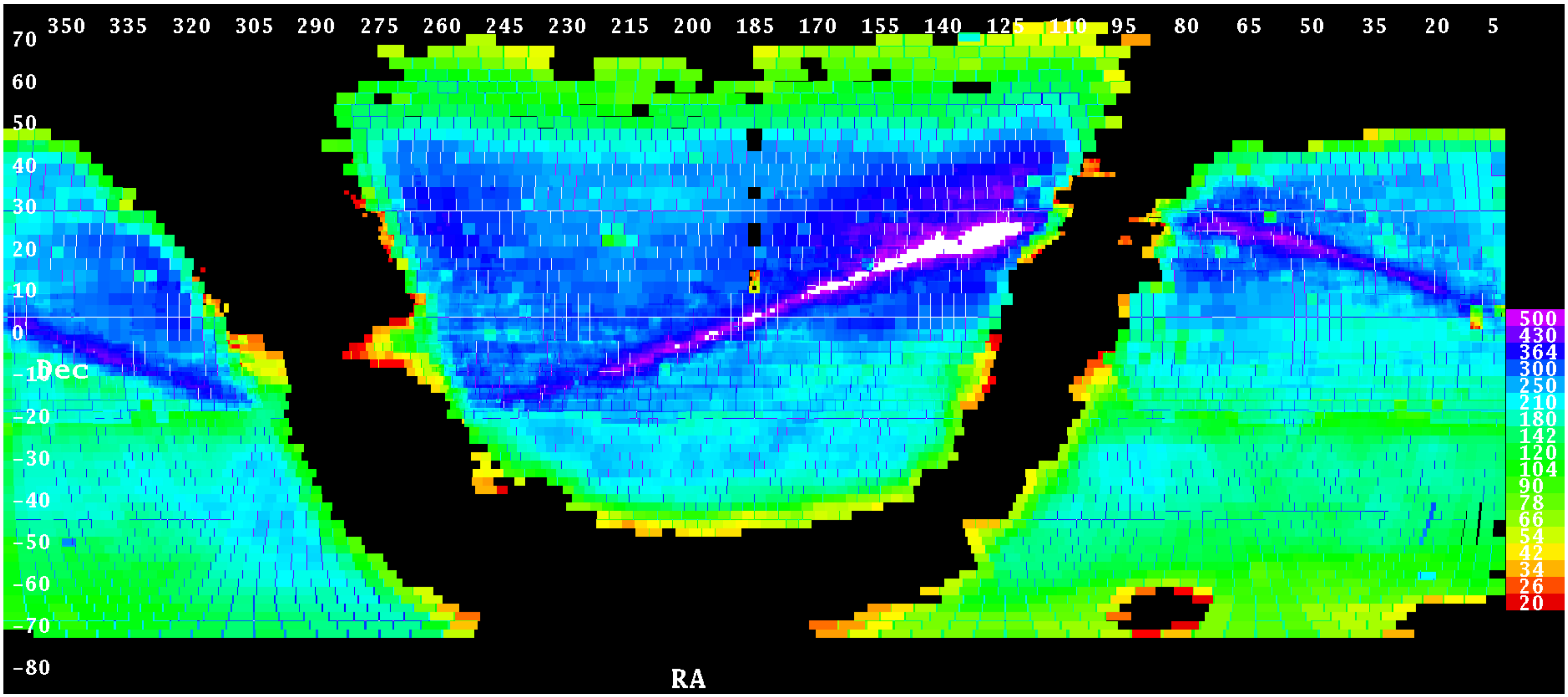
NEO survey Co-PI's:
E. Beshore & S. Larson (LPL)

CRTS uses the data from all three Catalina NEO surveys, with a coverage of up to 2,500 deg^2 / night, and the total area coverage of ~ 32,000 deg^2

	MLS The Mt. Lemmon Survey 1.5m Cass	**CSS** Catalina Sky Survey 0.7m Schmidt	**SSS** Siding Springs Survey 0.5m Schmidt
			
Survey region (deg)	+/- 5 deg ecliptic	-25 < Dec < +70	-80 < Dec < -25
Field of View (square deg)	1.2	8.1	4.2
Mag limit (V)	21.5	19.5	19.0

We are processing the Catalina data streams in real time to look for astrophysical transients

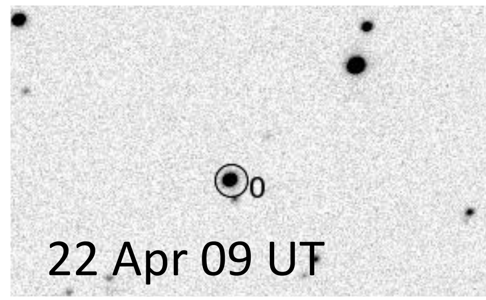
CSS coverage

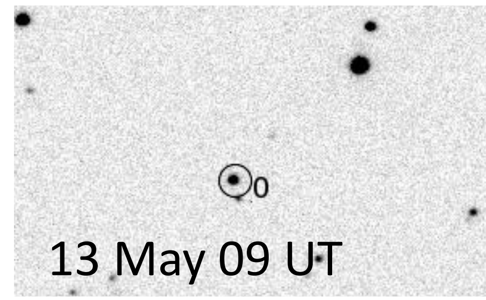


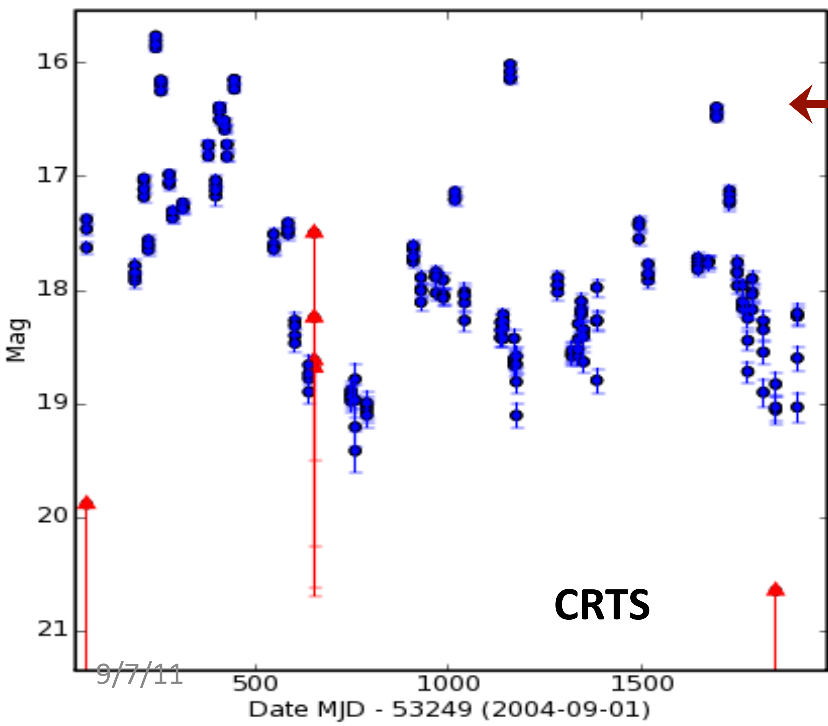
Follow-Up Observations:

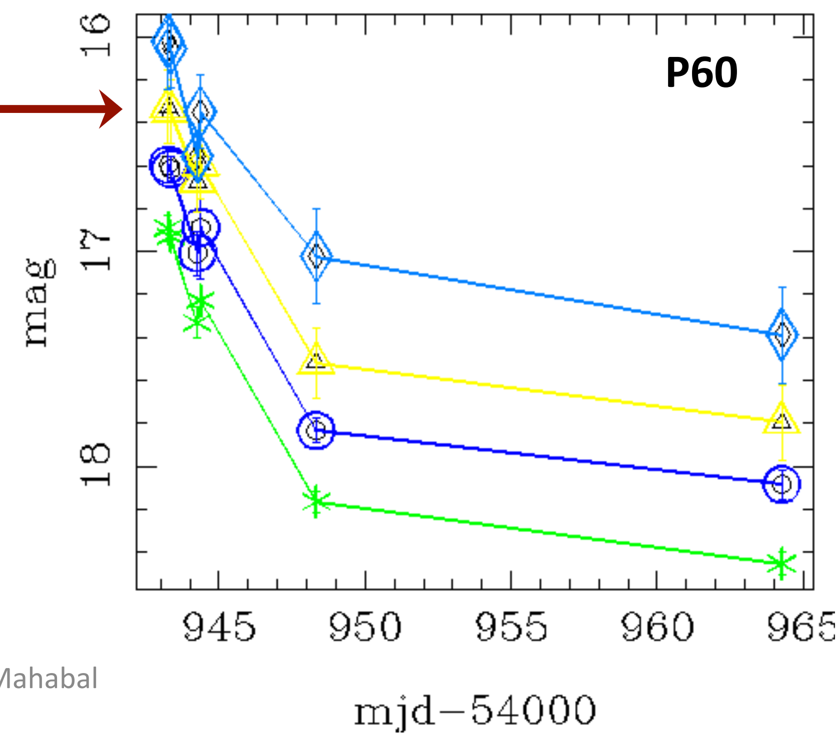
- **Photometry** (P60, NMSU, DAO, HTN, India, Mexico, etc.)
- **Spectroscopy** (Gemini N+S, Keck, P200, SMARTS, IGO, MDM)

CSS090421:174806+340401 A blazar, also monitored at OVRO in radio









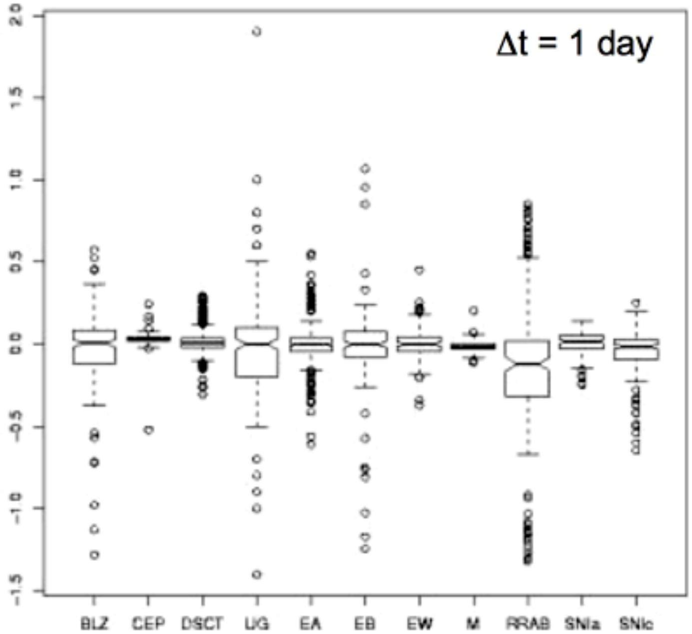
CRTS Event Detections

A Drake

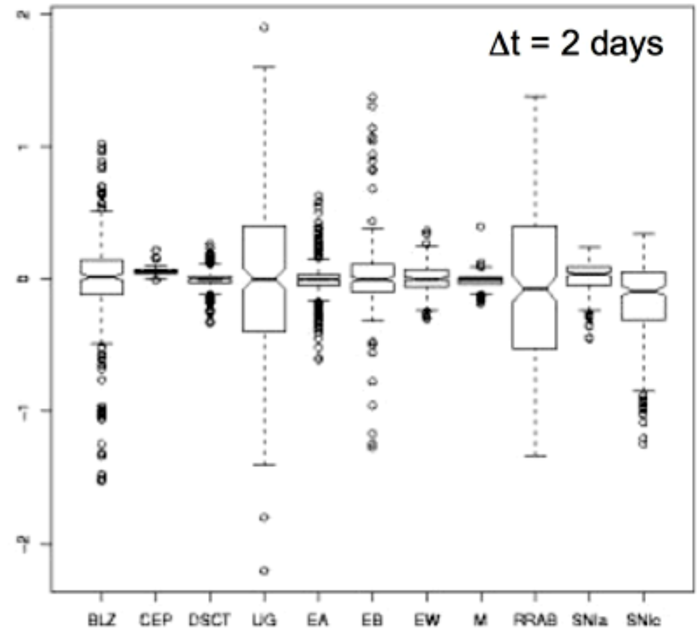
Distinct Events Detection Statistics as of 5 Jun 2011 UT:

Tel	All OTs	SNe	CVs	Blazars	Ast/ flares	CV/ SN	AGN	Other
CSS	2033	596	501	113	184	275	229	195
MLS	1560	183	38	12	122	374	744	214
SSS	227	24	93	7	5	43	16	42
Total	3820	803	632	132	311	692	989	451

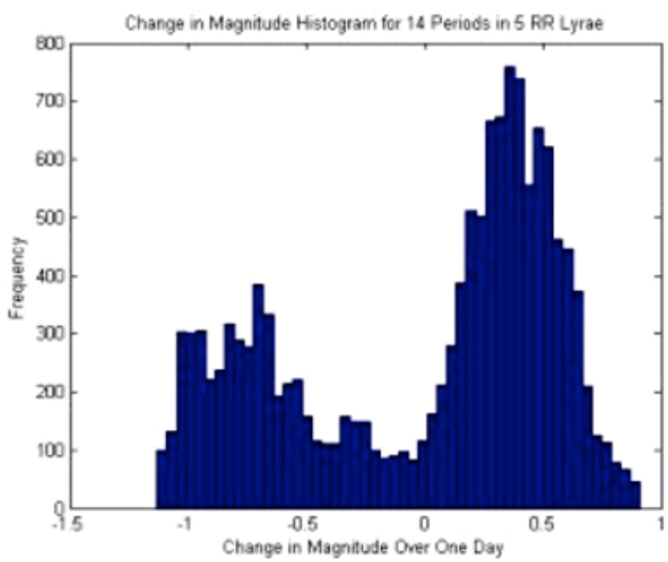
- Threshold set deliberately very high – only the most dramatic transients are pulled out in the real time
- About 1 strong transient per 10^6 source detections
- The rate of significant transients/variables is at least an order of magnitude higher
- Many events are re-detected repeatedly (not counted above)



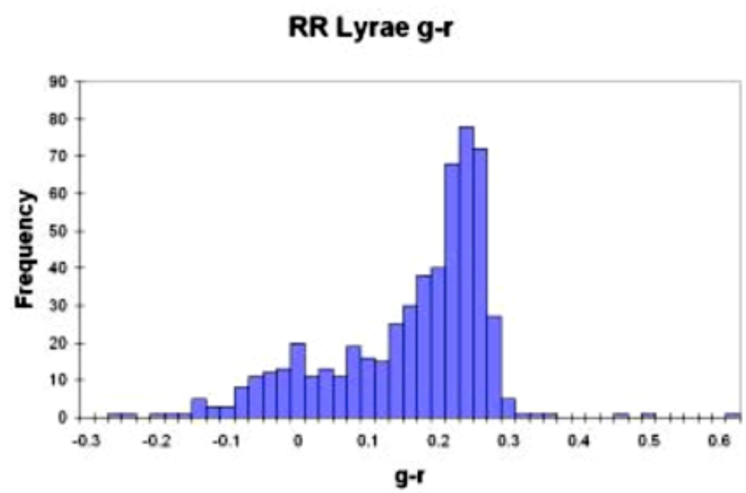
Δt = 1 day
BLZ CEP DSCT UG EA EB EW M RRAB SNIa SNIc



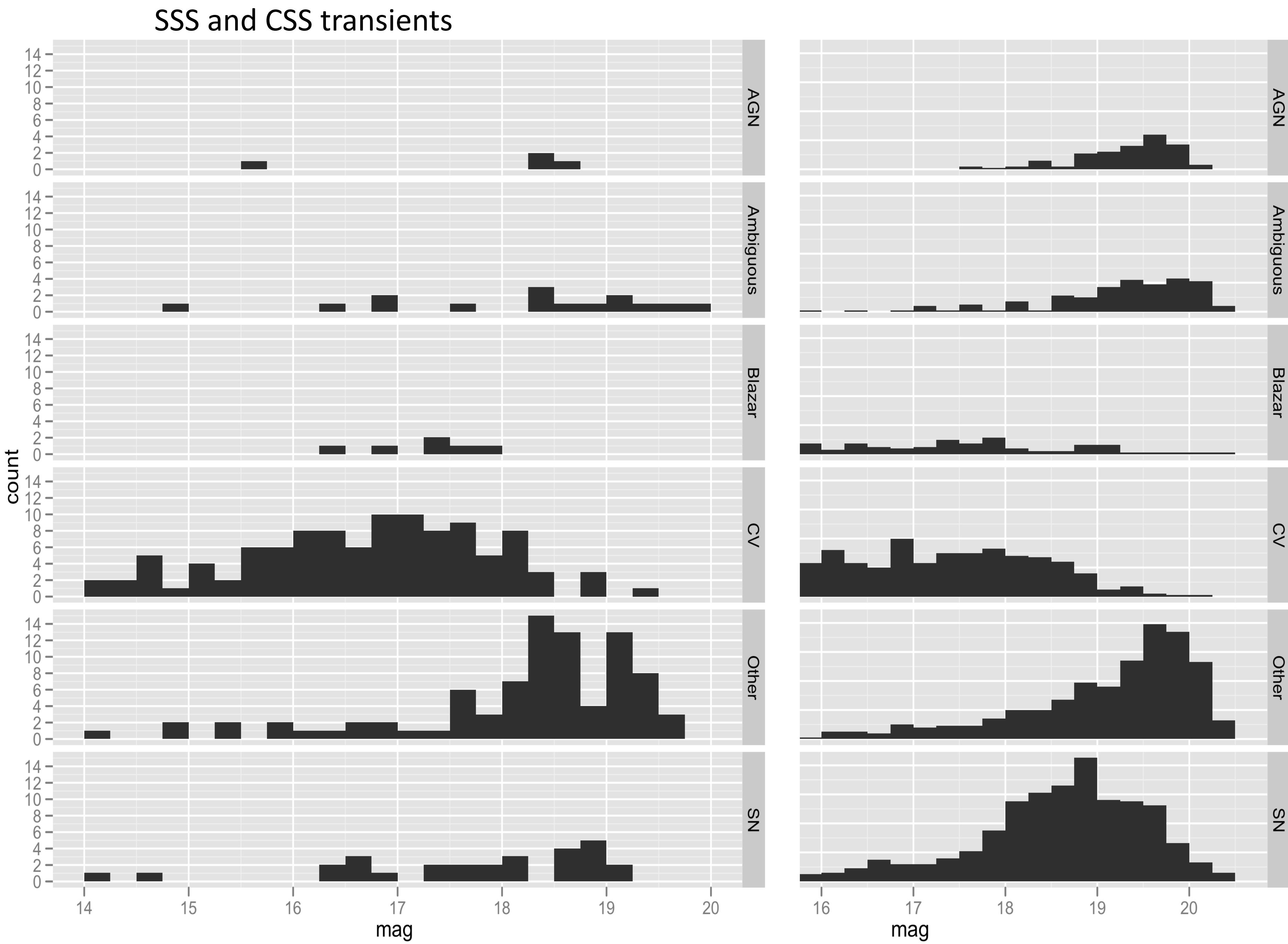
Δt = 2 days
BLZ CEP DSCT UG EA EB EW M RRAB SNIa SNIc



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

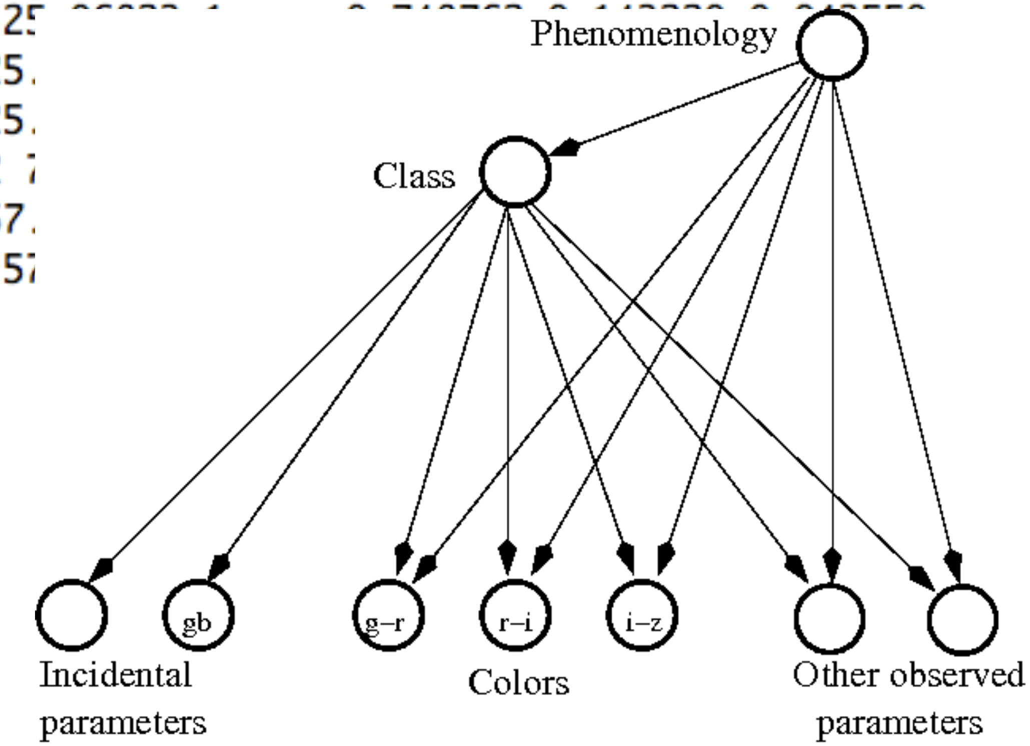

SSS and CSS transients
count
mag
mag
AGN
Ambiguous
Blazar
CV
Other
SN
14
15
16
17
18
19
20

Sample data input to BN

C Donalek,
N Sharma

```
id gminr rmini iminz gb class                    pCV       pSN       pblazar

1 801301180124103586 0.20 0.49 -1.06 41.570266 1   0.433000 0.221294 0.343222
2 801301180124103586 0.72 0.43 0.30 41.570266  1   0.114421 0.130915 0.754664
3 801301230184144420 0.16 0.50 -0.30 25.068228 1   0.945996 0.015071 0.038933
4 801301230184144420 0.18 0.54 -0.38 25.068228 1   0.959667 0.024743 0.015591
5 801301230184144420 0.19 -99.0 -99.0 25
6 801301230184144420 1.01 0.69  0.55 25.
7 801301230184144420 1.72 0.69 -0.07 25.
8 802011320554107996 -0.70 -0.16 -0.82
9 802191230754114380 0.76 0.14 -0.02 57.
10 802191230754114380 0.79 0.12 -0.16 57
```



Radio, for instance

The output is BN class which is fed to skyalert as an annotation to the original event

CV/SN classification ~80% with single epoch

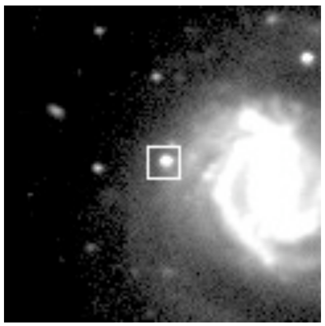
Naïve Bayes

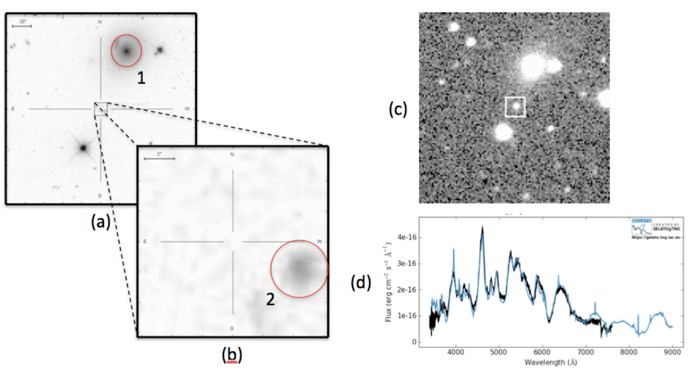
$$P(y=k \mid x) = P(x \mid y=k)P(k)/P(x) \propto P(k)P(x \mid y=k) \approx P(k)\prod_{b=1}^{B} P(x_b \mid y=k)$$

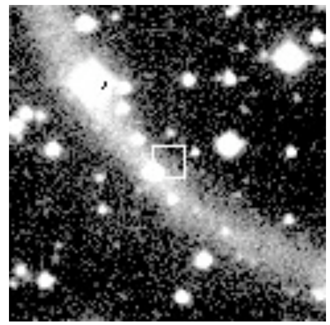
- x: feature vector of event parameters
- y: object class that gives rise to x (1<y<k)
- Certain features of x known: (position, flux)
- Others will be unknown: (color, delta-mag)
- Assumption: based on y, x is decomposable into B distinct independent classes (labeled x_b)
- This helps with the curse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

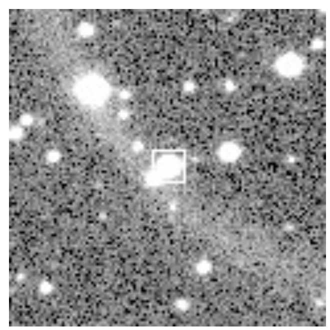
The importance of context

Which galaxy does a supernova belong to?

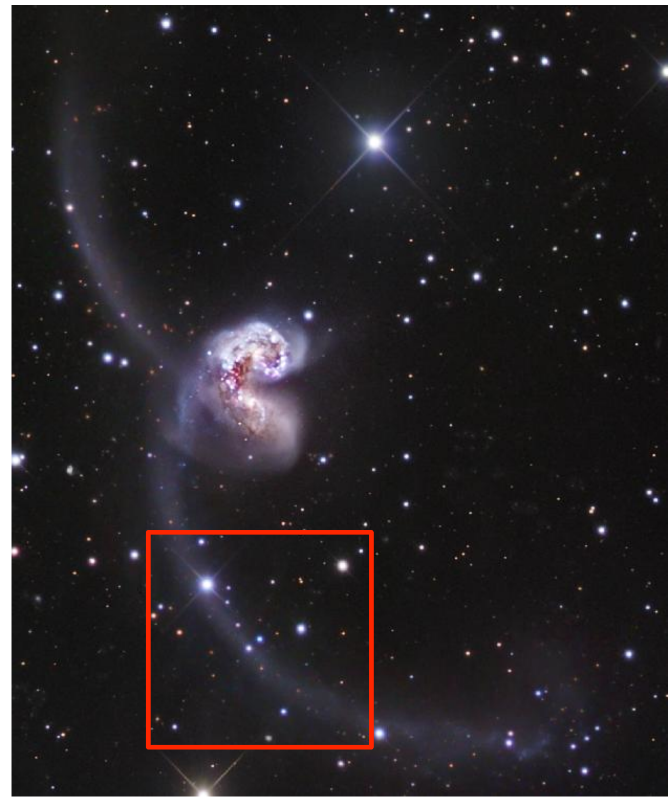






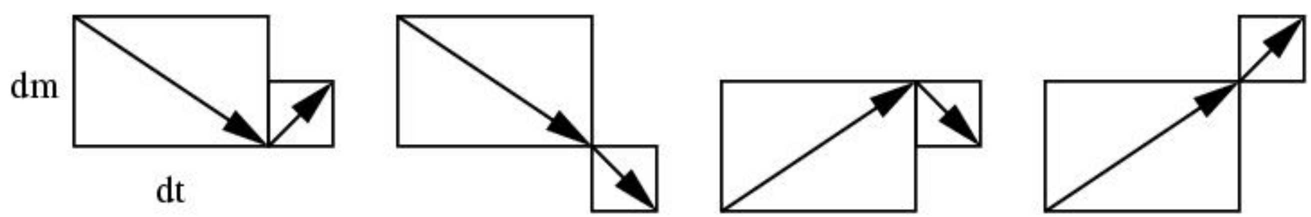


The need to see the big picture



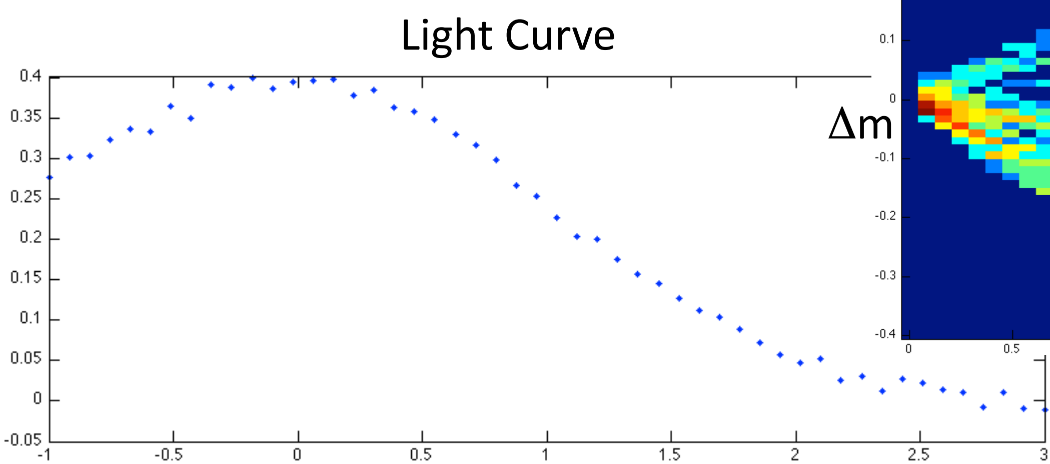
Characterization Vs. Classification

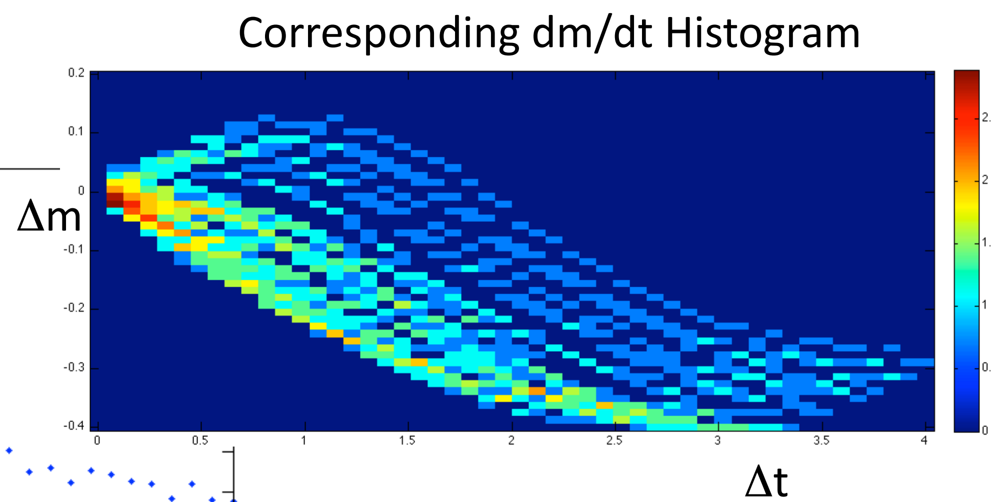
- Early focus on the extraction and dissemination of time series
- Characterizations is important
  - dm/dt
  - change of direction per unit time
  - change in periodicities (e.g,. wavelet or fourier decomposition);
  - variation in dm/dt
  - acceleration in dm/dt



Most SNe will not become fainter and then brighten up

Aspects of dm/dt processing

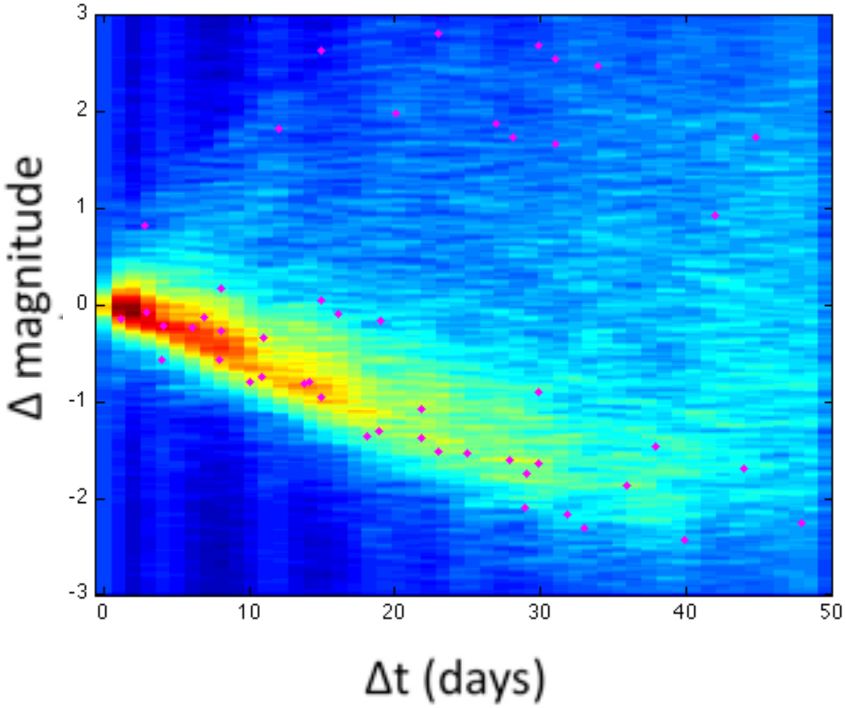




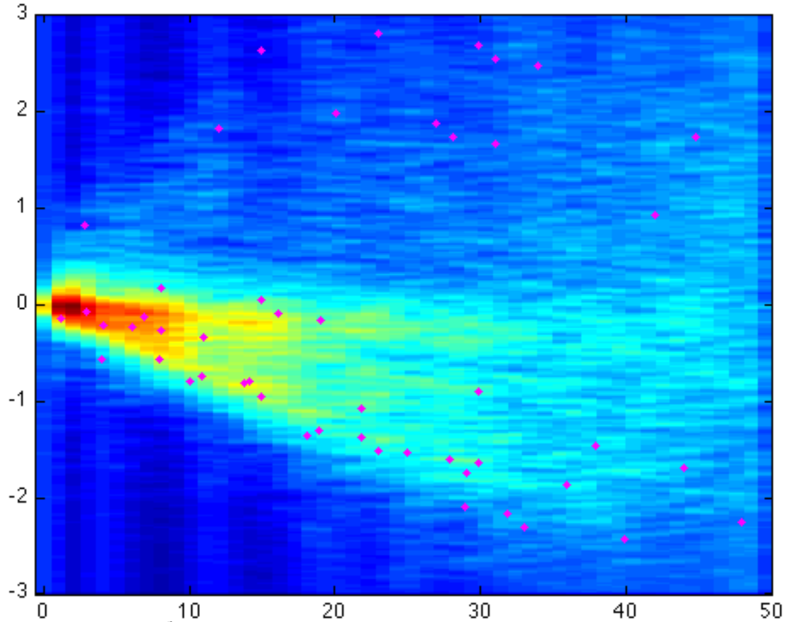
- dm/dt features capture sparse or irregular LCs
- The features, and thus the underlying density models, are invariant to absolute magnitude and time shifts
- Features & densities allow bound-only flux observations
  - Under poor seeing, we obtain only bounds like m > 18

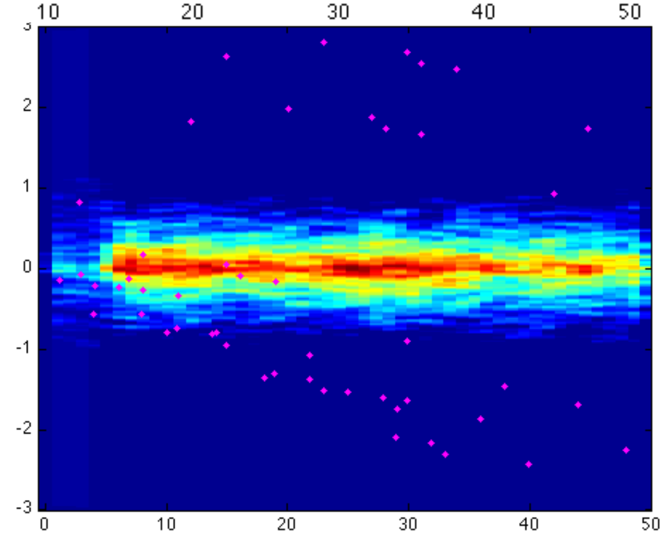
B Moghaddam, Y Chen

SN Ia



SN IIP





RR Lyrae

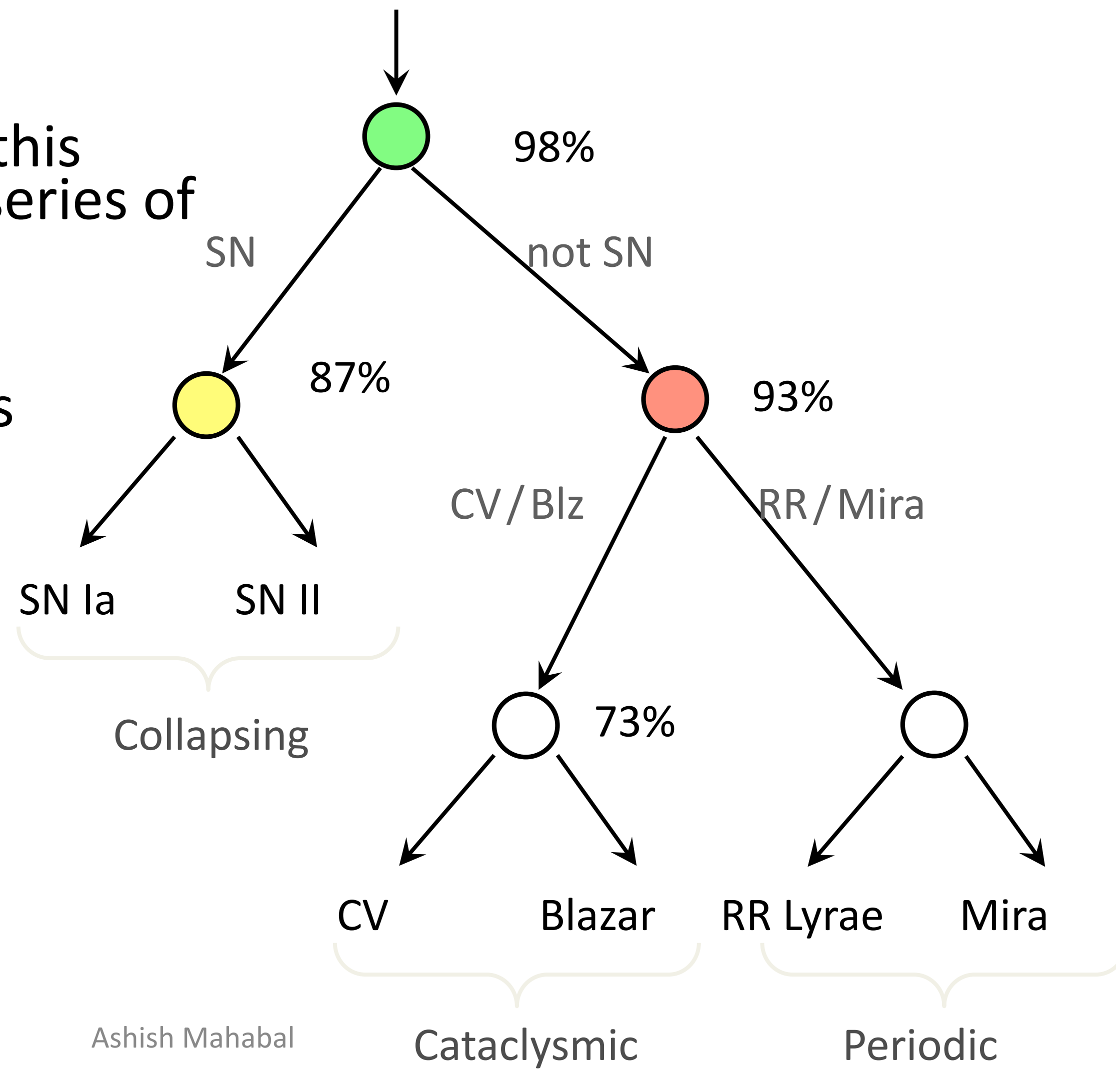
By taking subsections of dt/dm space determine which area is characteristic for which kind of variable

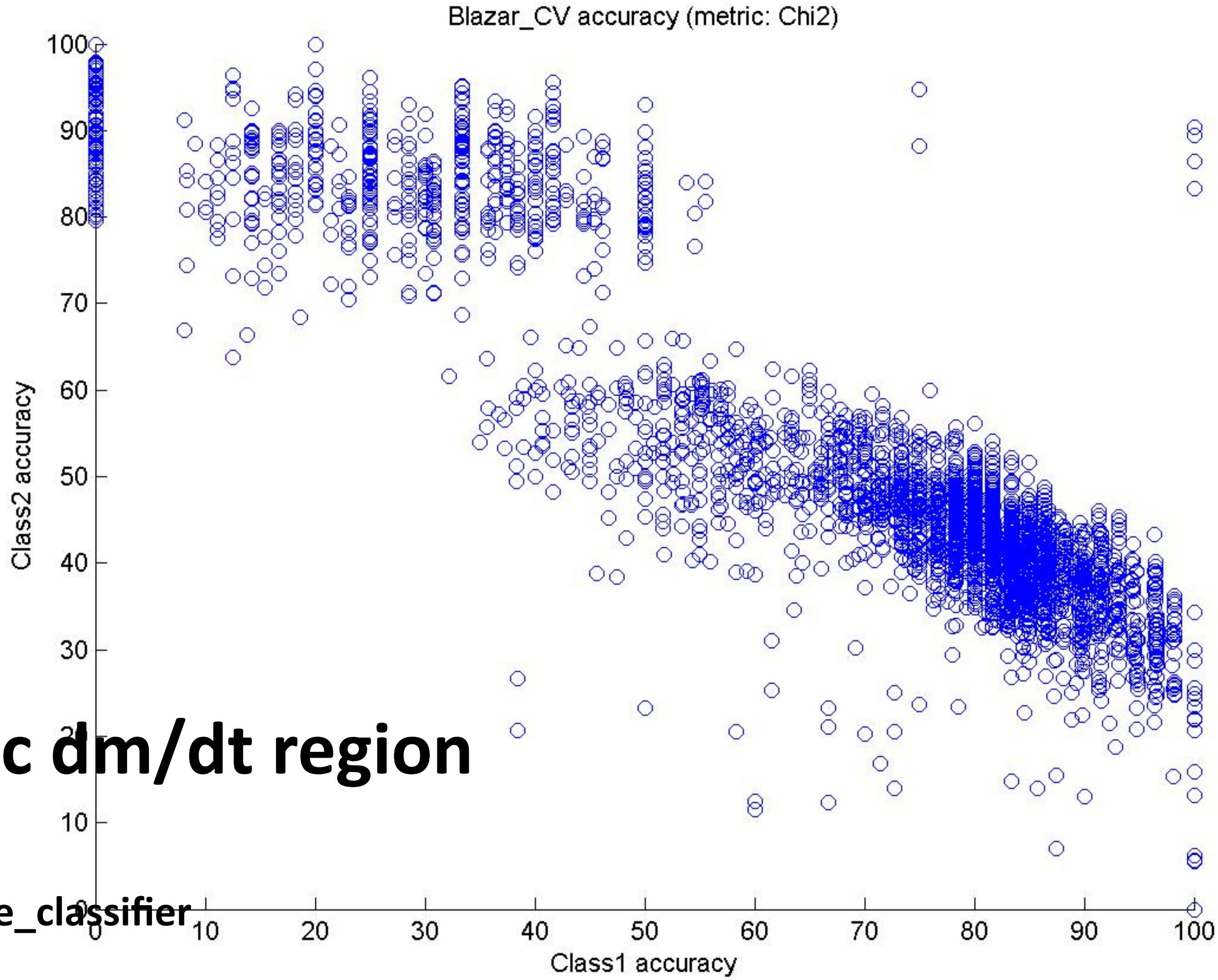
Classifier Architecture

Decision Tree decomposes this multi-class classifier into a series of binary discrimination tasks.

This specific DT follows the stratification that seems natural to astronomers.

All nodes shown were Implemented via dm/dt histogram binary classifiers.





For a specific dm/dt region

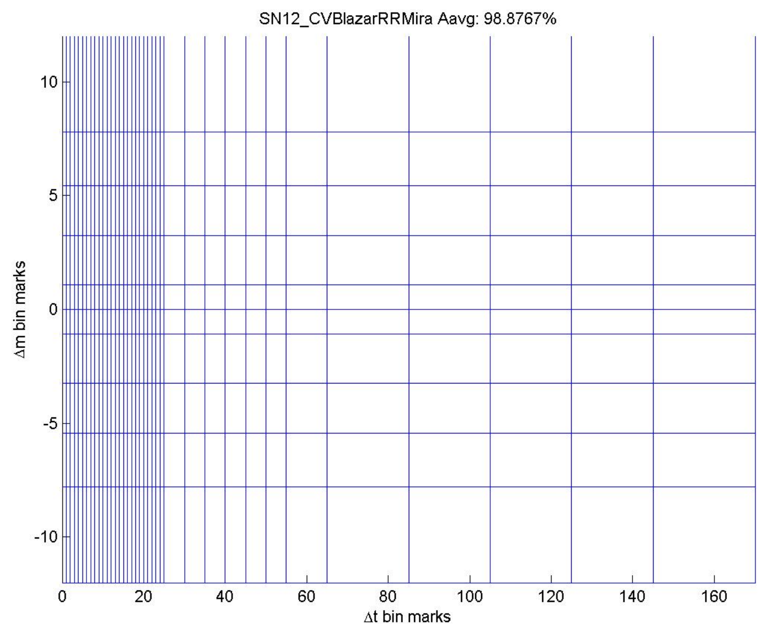
>> prep_Blazar_CV
>> RUNME_prototype_classifier

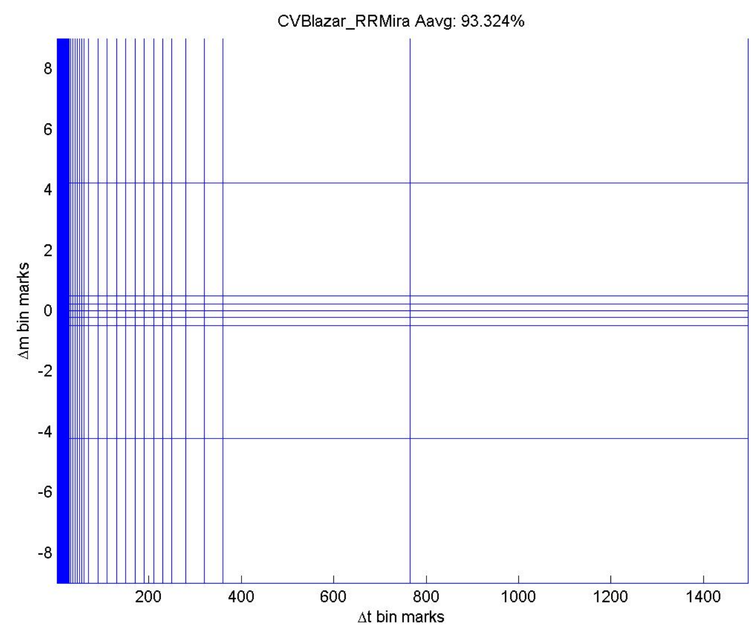
Blazar accuracy = 83.33 %
CV accuracy = 58.85 %
Total accuracy = 64.31 %
Average accuracy = 71.09 %

Using GAs to determine intervals

- *dmagbins*
  - generate array elements (intervals) based on a normal distribution around 0
  - remove negative elements in array
  - sort in ascending order
  - bin marks determined by cumulative sum of array elements
  - reflect over 0 to build symmetric *dmagbins*
- *dtbins*
  - 1-day interval for first 25-35 days (chosen uniformly at random)
  - 5-day interval for next 30 days
  - 10-day interval for next 60 days
  - 20-day interval for next 240 days
  - 365-day interval until end
- σ_{dm}: chosen uniformly at random from [0.2, 1.5]
- σ_{dt}: chosen uniformly at random from [0.2, 1.5]
- *SymDir*: 0 or 1
- *Alpha*: chosen log-uniformly at random from $[10^{-3}, 10^{2}]$

dm/dt bins as selected by GA

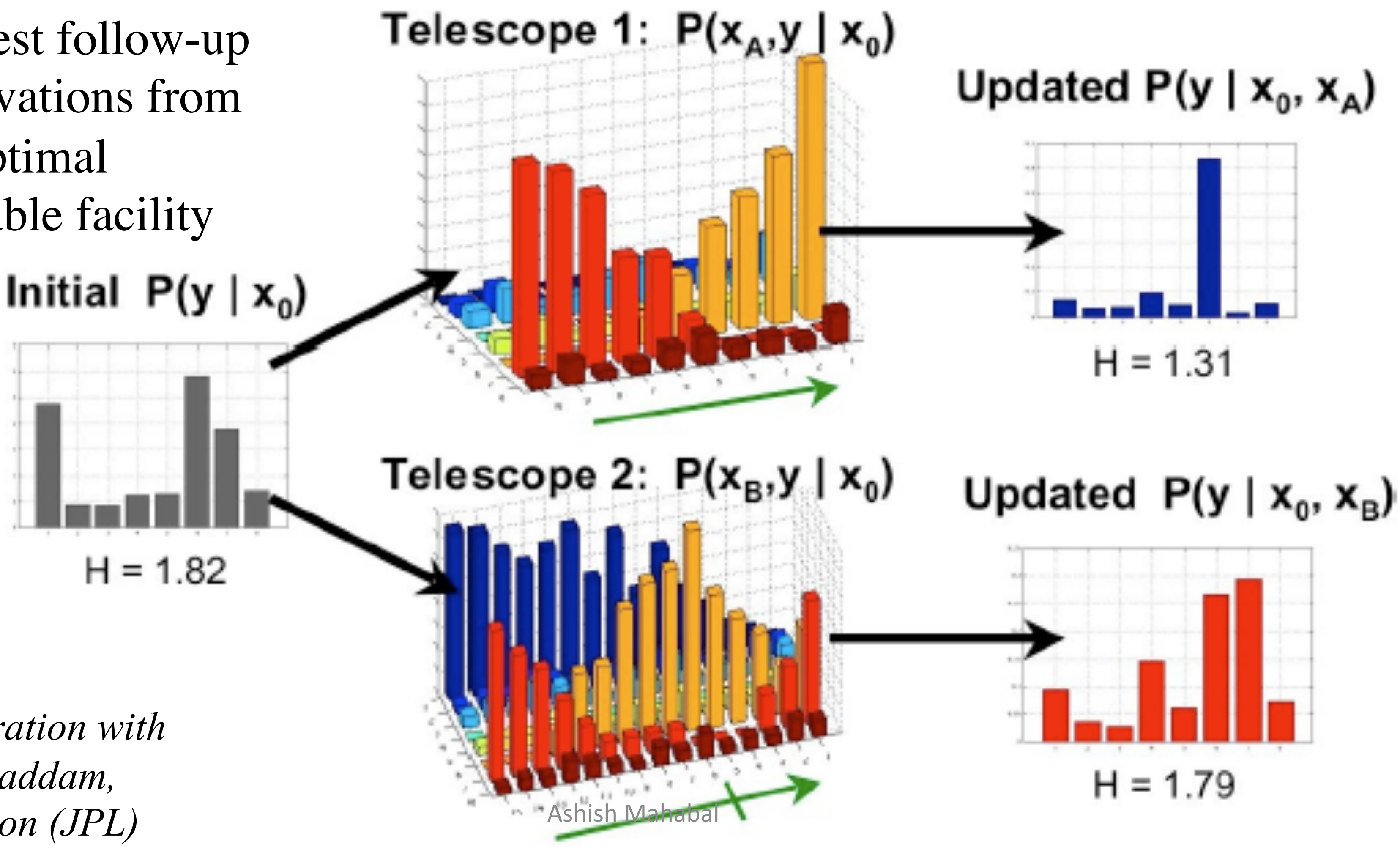




Automating the Optimal Follow-Up

What type of follow-up data has the greatest potential to discriminate among the competing models (event classes)?

Request follow-up observations from the optimal available facility



Collaboration with B. Moghaddam, M. Turmon (JPL)

Event Publishing / Dissemination
skyalert.org

PI: R. Williams

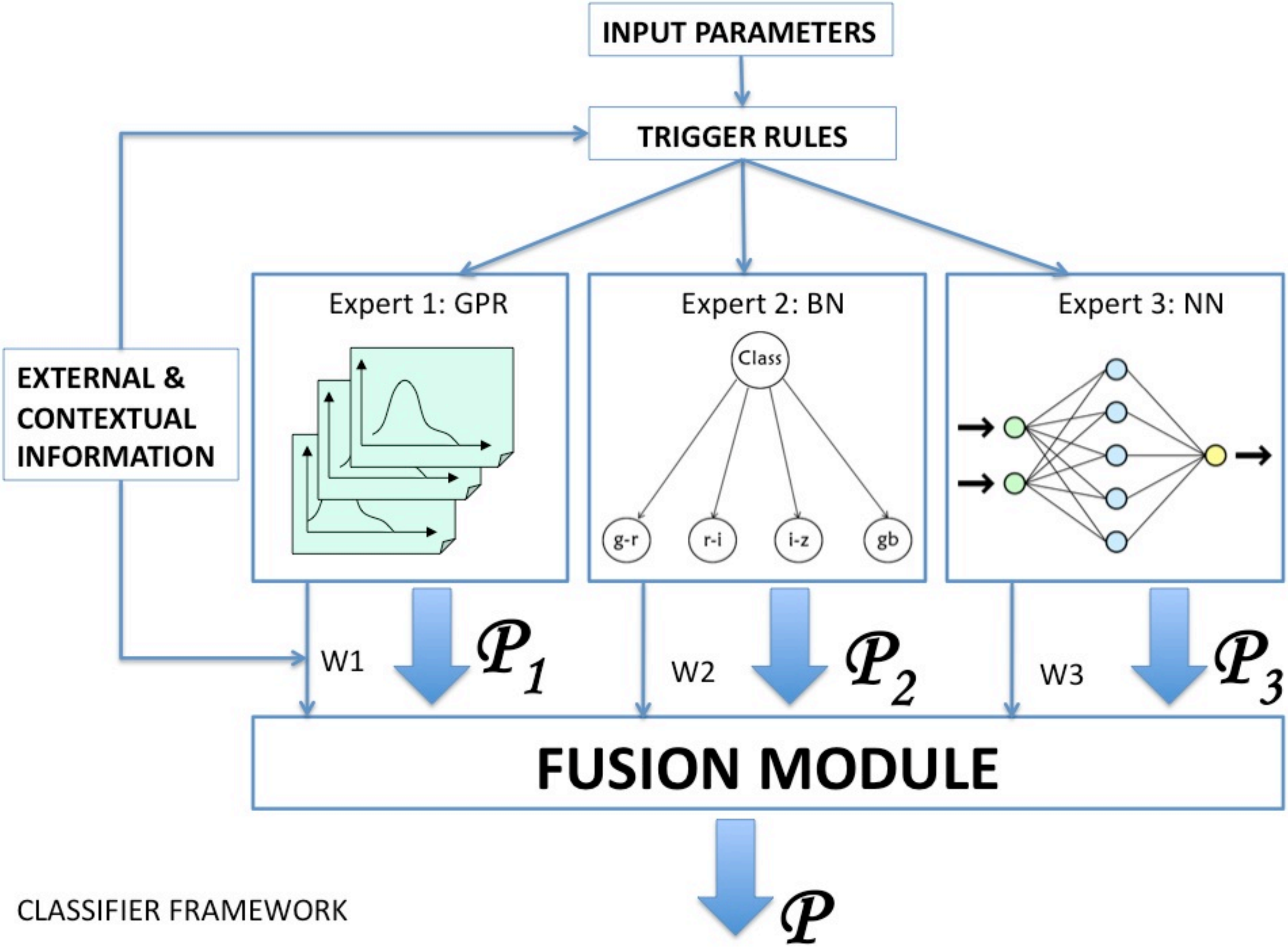
- Real time:
  - VOEvents, Twitter, iApp (thousands of events)
  - Also on SkyAlert.org, feeds to the WWT, GoogleSky
- Next day: annotated tables on the CRTS website

CSS ID	RA (J2000)	Dec (J2000)	Date	Mag	CSS images	SDSS	Others	Followed	Last	LC	Classification
CSS091121:221159+263906	332.99697	26.65153	20091121	18.33	911211261084134848	no	34848	no	2009-11-21	34848	SN/Blazar mag 21
CSS091121:013728+253450	24.36768	25.58061	20091121	17.78	911211260084103595	no	03595	no	2009-11-21	03595	SN/CV
CSS091121:032627+070744	51.61364	7.12902	20091121	16.68	911211070194124436	no	24436	no	2009-11-21	24436	CV mag 21
CSS091121:033232+020439	53.13295	2.07747	20091121	16.93	911211010194134434	no	34434	no	2009-11-21	34434	CV mag 20
CSS091121:085600-051945	133.99922	-5.32906	20091121	18.17	911210040484107252	no	07252	no	2009-11-21	07252	SN CFHT mag 22 gal
CSS091120:100525+511639	151.35223	51.27742	20091120	18.80	911201520354108835	yes	08835	no	2009-11-20	08835	SN SDSS mag 21,9 gal
CSS091120:082908+482639	127.28503	48.44423	20091120	15.69	911201490314109371	yes	09371	no	2009-11-20	09371	CV/SN SDSS mag 21,6 gal?
CSS091120:004417+411854	11.07004	41.31494	20091120	17.00	911201400044145995	yes	45995	no	2009-11-20	45995	Nova M31 2009-11d
CSS091120:001019+410455	2.58044	41.08191	20091120	16.69	911201400014137919	no	37919	no	2009-11-20	37919	CV mag 20,0

Transient classification mantra

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

Bayesian network, dm/dt processing, (DAME, VOStat, VO), Skyalert



CLASSIFIER FRAMEWORK

Bayesian Network/fusion modules are no Cartesian theatre

- Different parameters, methods are separate (though perhaps not independent) probes

(non-)Cartesian theatre

One observation can drive the direction given the large number of possible candidates

Not much scope for error