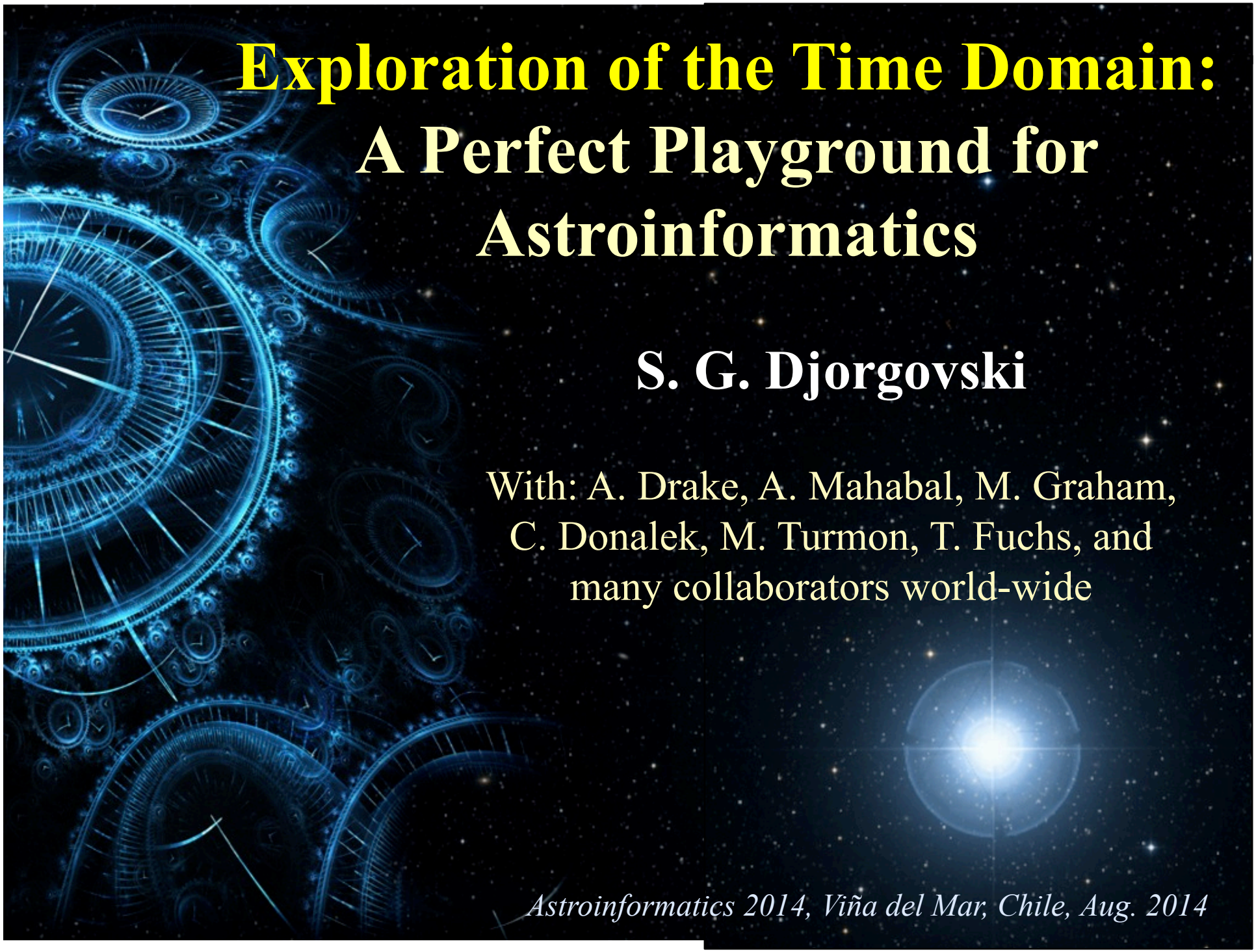
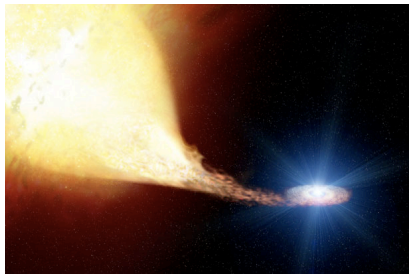

Exploration of the Time Domain:
A Perfect Playground for Astroinformatics
S. G. Djorgovski
With: A. Drake, A. Mahabal, M. Graham, C. Donalek, M. Turmon, T. Fuchs, and many collaborators world-wide
Astroinformatics 2014, Viña del Mar, Chile, Aug. 2014

Astroinformatics in the Time Domain

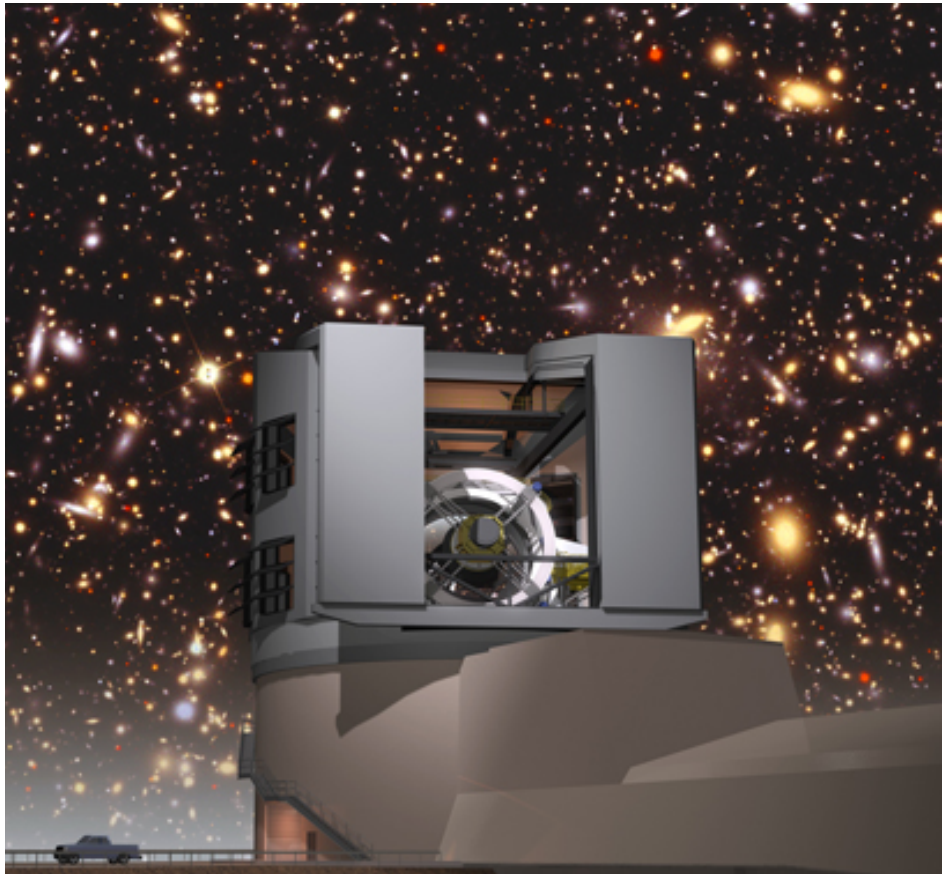
- Time domain is perhaps the most vibrant and growing research arena of astronomy today, from the Solar system to cosmology
- Enabled by the new generation of synoptic sky surveys (PQ, PTF, CRTS, PS1, … LSST, SKA…)
  - Themselves enabled by the continuing progress in computing and information technology
- Great scientific opportunities, but also challenges
  - **Time domain adds complexity** and additional axes of the observable parameter space
  - Some phenomena can be studied *only* in the time domain
- Exercises all aspects of the VO, and adds new ones, e.g., real-time data analysis and follow-up



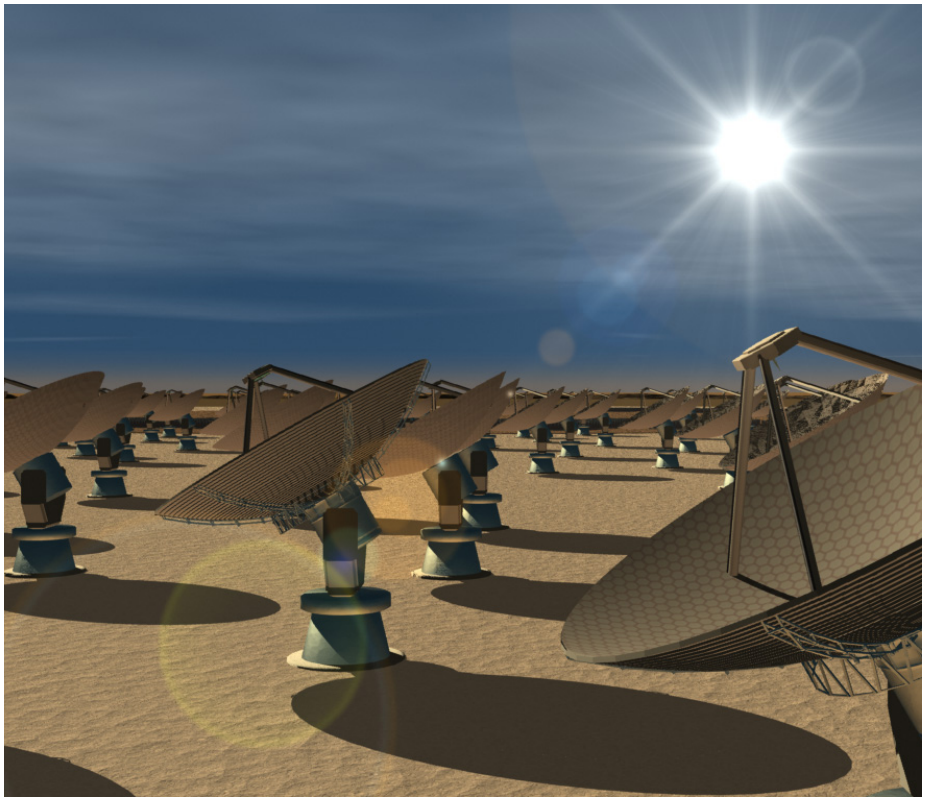



The Panoramic Cosmic Cinematography

Large Synoptic Survey Telescope (LSST) ~ 30 TB / night



Square Kilometer Array (SKA)
~ 1 EB / second (raw data)
(EB = 1,000,000 TB)

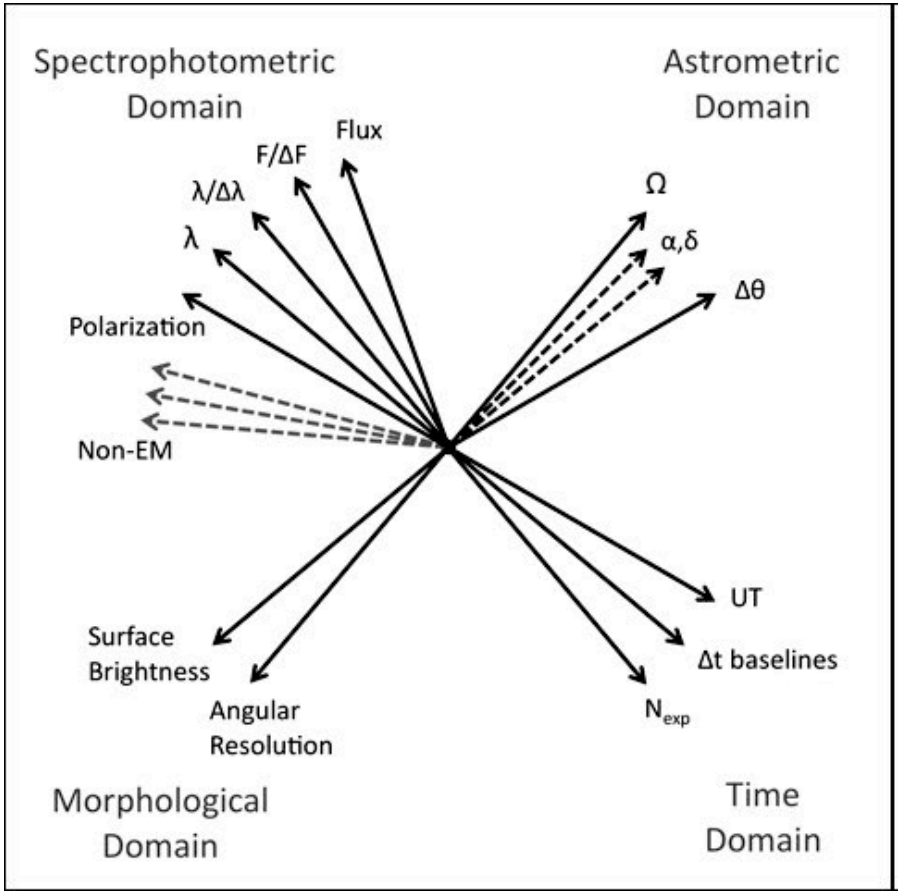


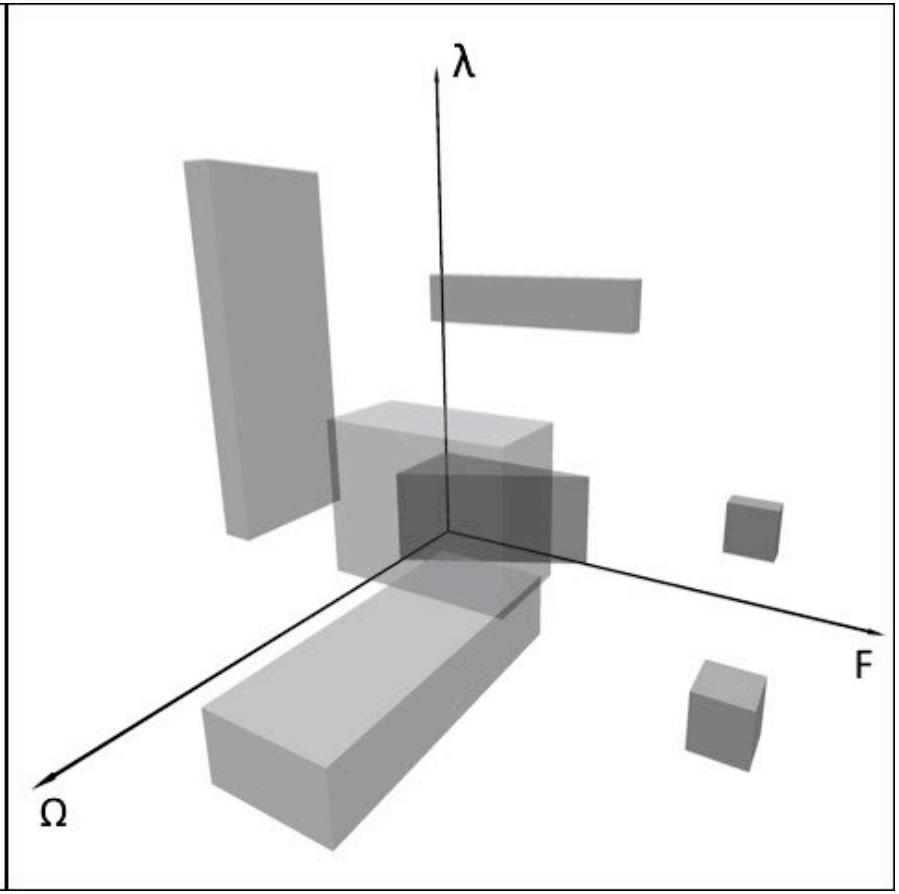
+ many of their precursors – including the current surveys

Systematic Exploration of the Observable Parameter Space (OPS)

Its axes are defined by the observable quantities

Every observation, surveys included, carves out a hypervolume in the OPS

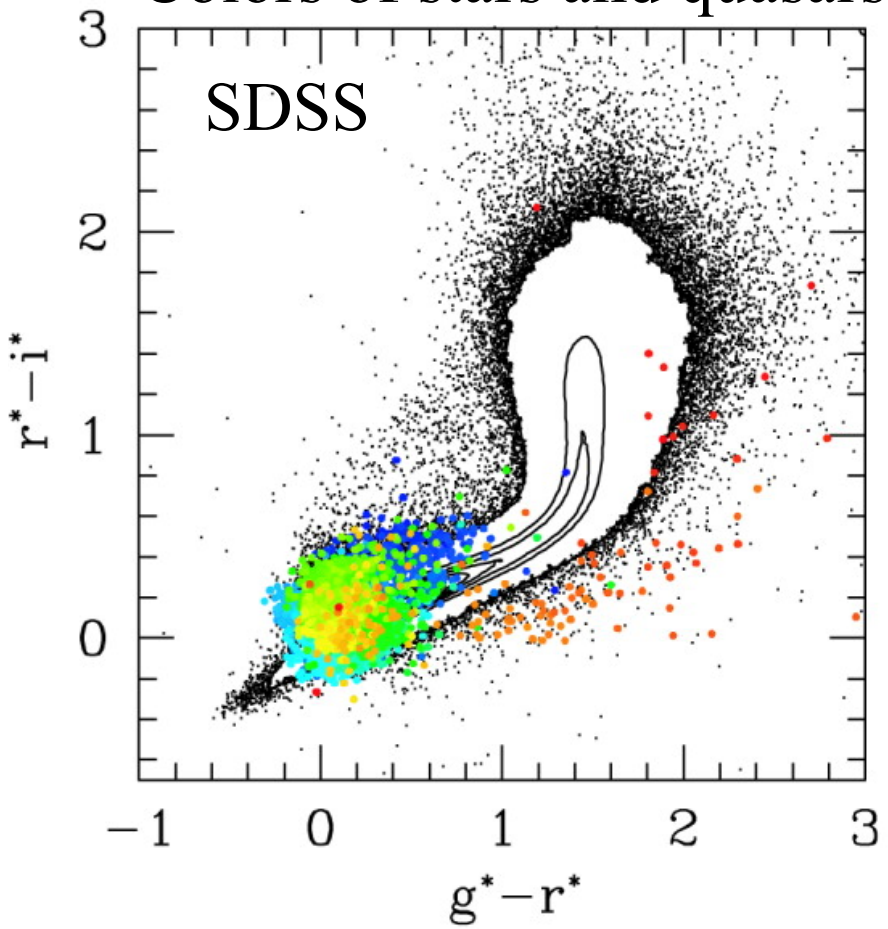




Technology opens new domains of the OPS → **New discoveries**

Measurements Parameter Space

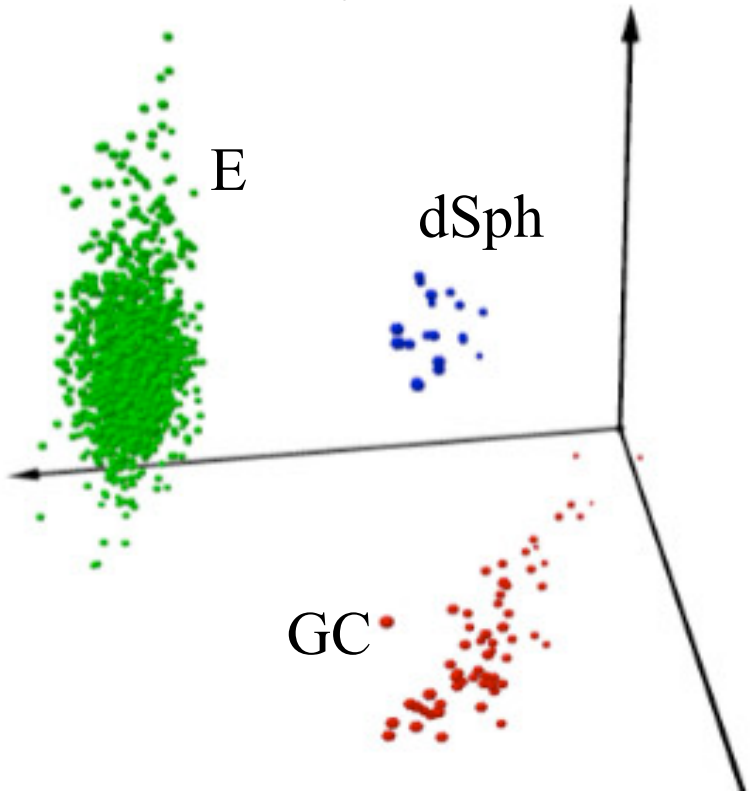
Colors of stars and quasars



Dimensionality ≤ the number of observed quantities

Physical Parameter Space

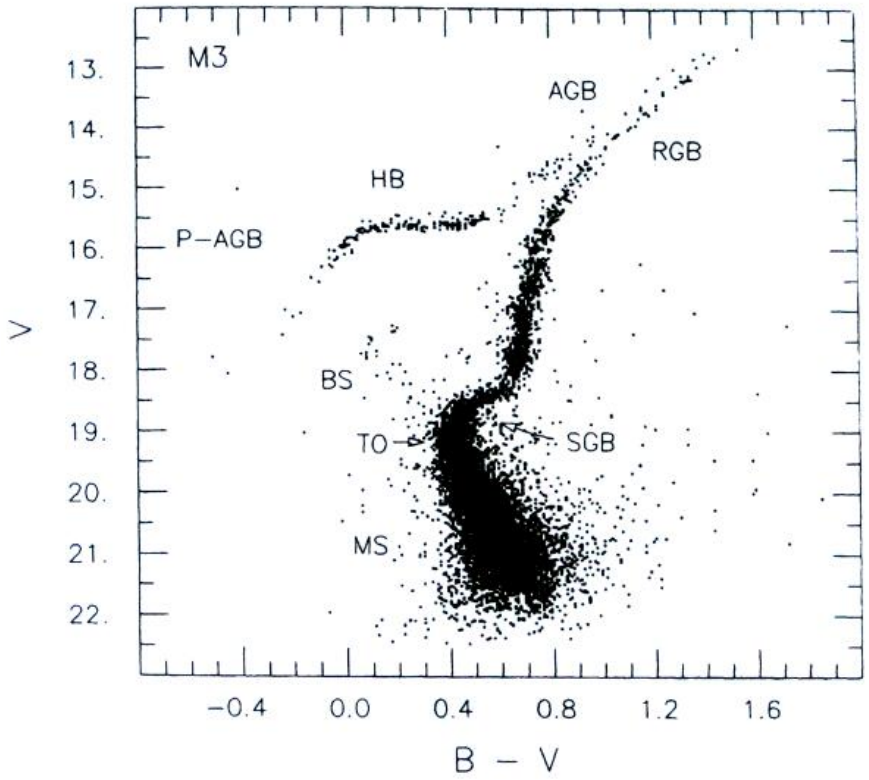
Fundamental Plane(s) of hot stellar systems



Both are populated by **objects or events**

Measurements Parameter Space

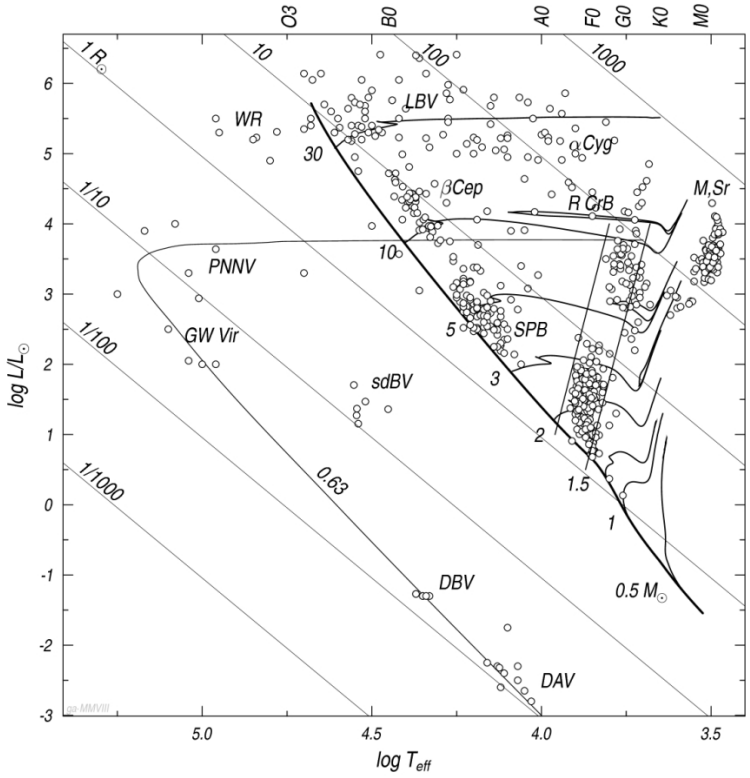
Color-magnitude diagram



Theory
+
Other data

Physical Parameter Space

H-R diagram



- Not filled uniformly: clustering indicates different families
- Clustering + dimensionality reduction ⇨correlations
- High dimensionality poses analysis challenges

Parameter Spaces for the Time Domain

(in addition to everything else: flux, wavelength, etc.)

- For ***surveys***:
  - Total exposure per pointing
  - Number of exposures per pointing
  - How to characterize the cadence?

    ↳ Window function(s)

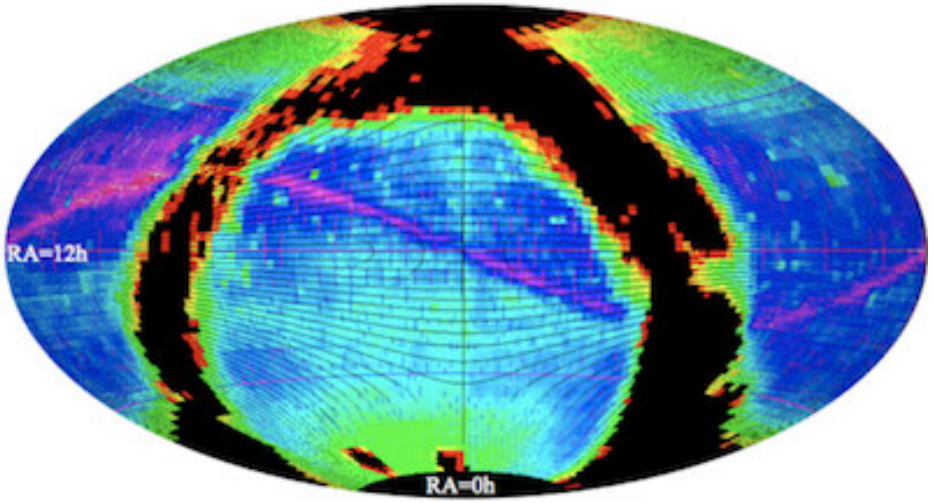
    ↳ Inevitable biases
- For ***objects/events*** ~ light curves:
  - Significance of periodicity, periods
  - Descriptors of the power spectrum (e.g., power law)
  - Amplitudes and their statistical descriptors

  … etc. − over 70 parameters defined so far, but which ones are the minimum / optimal set?

Catalina Real-Time Transient Survey (CRTS)

http://crts.caltech.edu

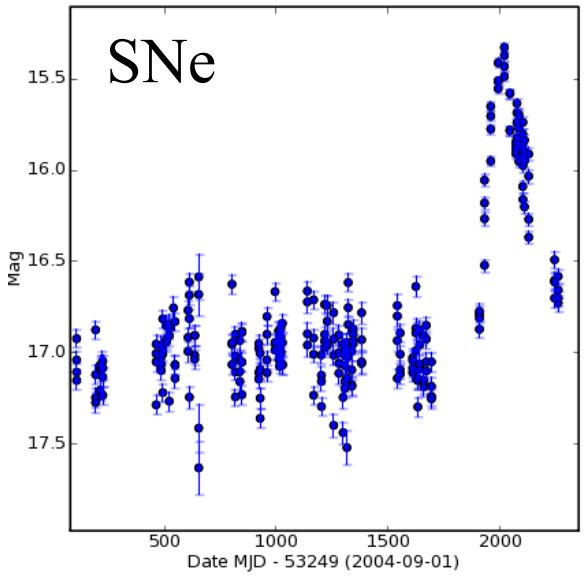
- Data from a search for near-Earth asteroids at UA/LPL; we discover astrophysical transients in their data stream
- 3 telescopes in AZ, Australia – maybe also Chile?

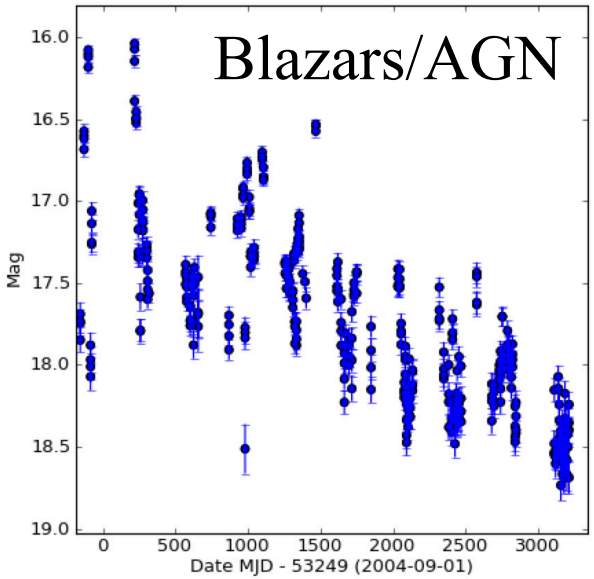


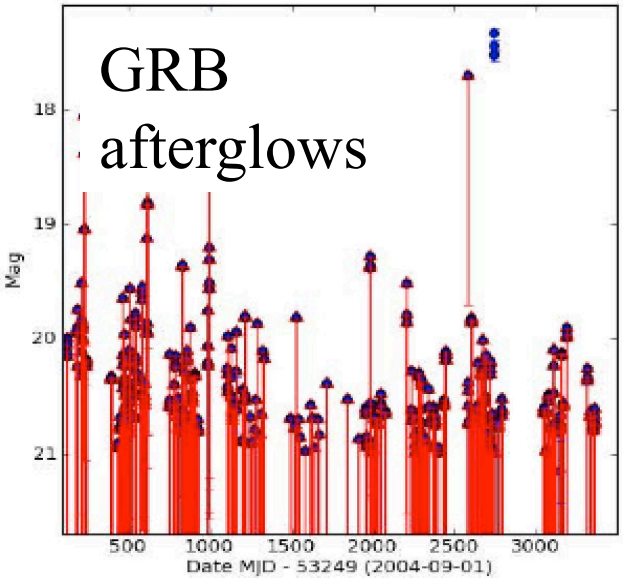
- > 80% of the sky covered ~ 200 – 500 times over ~ 8+ years down to ~ 19 – 21 mag, baselines from 10 min to 8 years, a few hundred exposures per pointing
- **CRTS2:** new cameras, sky coverage rate × 2.5, transients discovery rate × 20; forming an international consortium

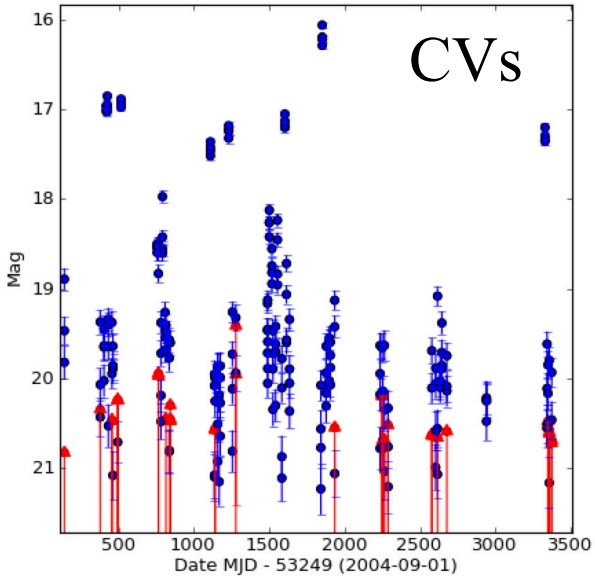
Open data policy: ***all data are made public; transients are published immediately on line,*** **for the entire community**

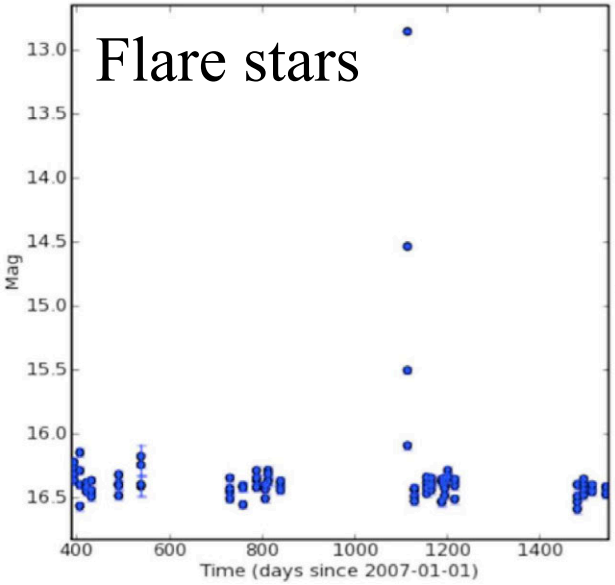
A Variety of CRTS Transients

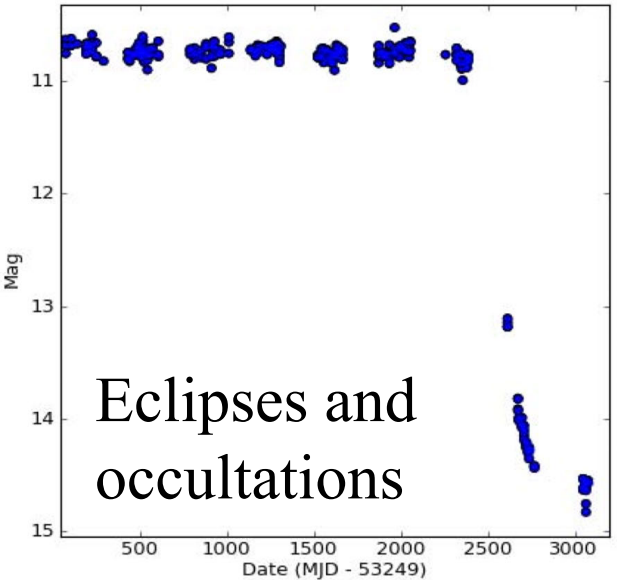












CRTS Transients as of the August 2014

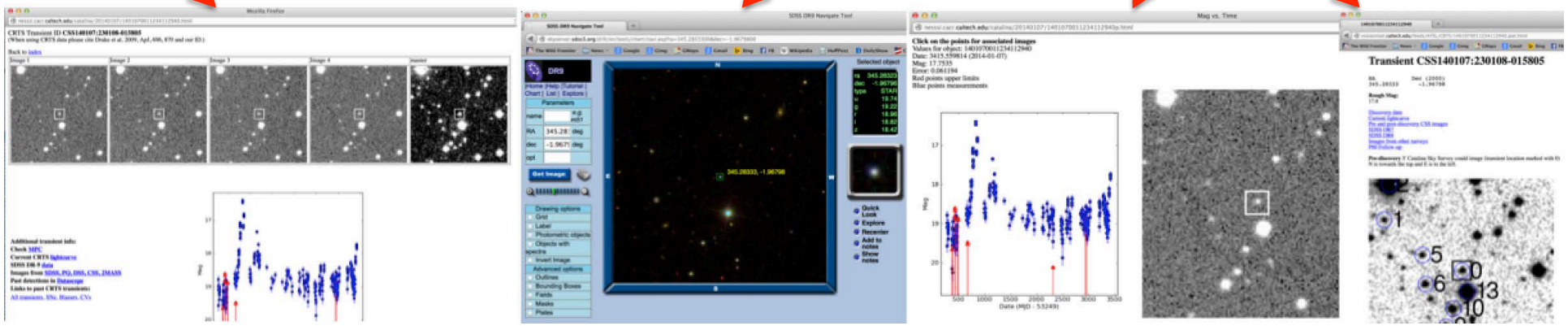
Telescope	All OTs	SNe	SN/ CV?	CVs	Blazar	AGN	Aster./ Flare?	Other
CSS	**4207**	**1281**	**494**	**829**	**236**	**555**	**299**	**603**
MLS	**4374**	**644**	**728**	**85**	**84**	**2080**	**271**	**686**
SSS	**697**	**105**	**109**	**254**	**18**	**33**	**12**	**171**
SNhunt	**197**	**197**	**0**	**0**	**0**	**0**	**0**	**0**
Total	**9475**	**2227**	**1331**	**1168**	**338**	**2668**	**582**	**1460**

- Threshold set deliberately very high (~ 1 – 2 mag, >5 σ), so only the most dramatic transients are pulled out in the real time
- About 1 strong transient per 10^6 source detections
- The rate of all statistically significant transients is at least an order of magnitude higher; available for an archival study
- Many events are re-detected repeatedly (not counted above)
  - Many also detected independently by PTF, PS1, etc.

Event Publishing / Dissemination

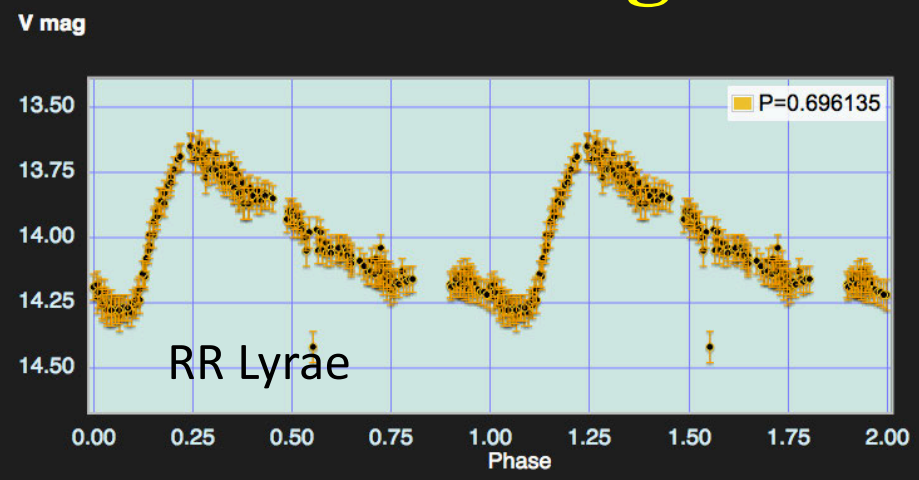
- Real time: VOEvent, RSS, *SkyAlert*, Twitter, iApp, …
- Next day: annotated tables on the CRTS website

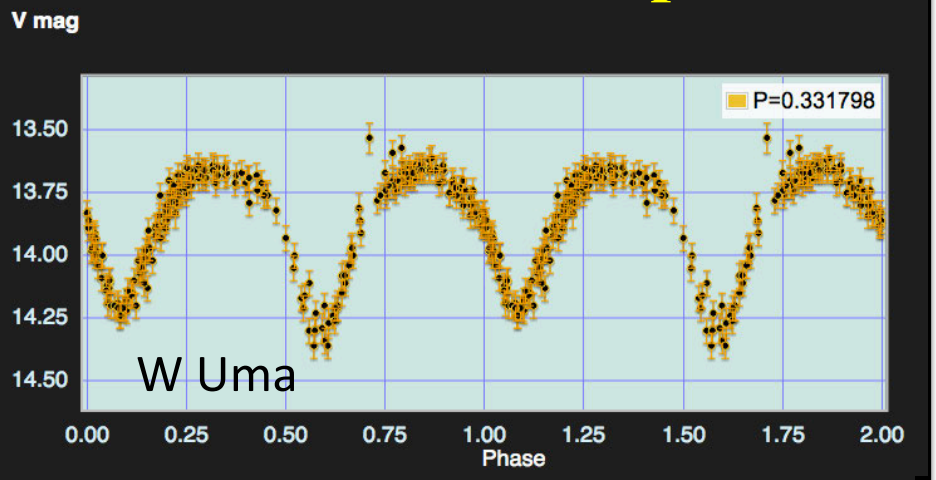
CRTS ID	RA (J2000)	Dec (J2000)	Date	Mag	CSS images	SDSS	Others	Followed	Last	LC	FC	Classification
CSS140107:082255+073033	125.72901	7.50903	20140107	18.88	1401071070454146201	yes	14620	no	2014-01-07	14620	yes	Unknown SDSS mag 22
CSS140107:015854+053524	29.72348	5.59009	20140107	15.33	1401071040114136538	yes	13653	no	2014-01-07	13653	yes	SN 2013hs (Howerton) mag 1
CSS140107:114807+014254	177.02819	1.71506	20140107	19.34	1401071010634128518	yes	12851	no	2014-01-07	12851	yes	QSO SDSS mag 20,5
CSS140107:145029-083859	222.61934	-8.64971	20140107	14.16	1401070090794151988	yes	15198	no	2014-01-07	15198	yes	HPM LHS_381
CSS140107:133002-084233	202.50739	-8.70909	20140107	12.83	1401070090724151417	no	15141	no	2014-01-07	15141	yes	HPM GJ_514
CSS140107:230108-015805	345.28333	-1.96798	20140107	17.75	1401070011234112940	yes	11294	no	2014-01-07	11294	yes	Blazar mag 19,0
CSS140104:013741+220312	24.42141	22.05338	20140104	14.51	1401041210094138406	yes	13840	no	2014-01-04	13840	yes	Unknown SDSS mag 22
CSS140104:020225+144325	30.60214	14.72357	20140104	19.89	1401041150114109714	yes	10971	no	2014-01-04	10971	yes	AGN SDSS mag 21,9
CSS140104:034718+014254	56.82594	1.71497	20140104	19.42	1401041010214127020	no	12702	no	2014-01-04	12702	yes	SN mag 16

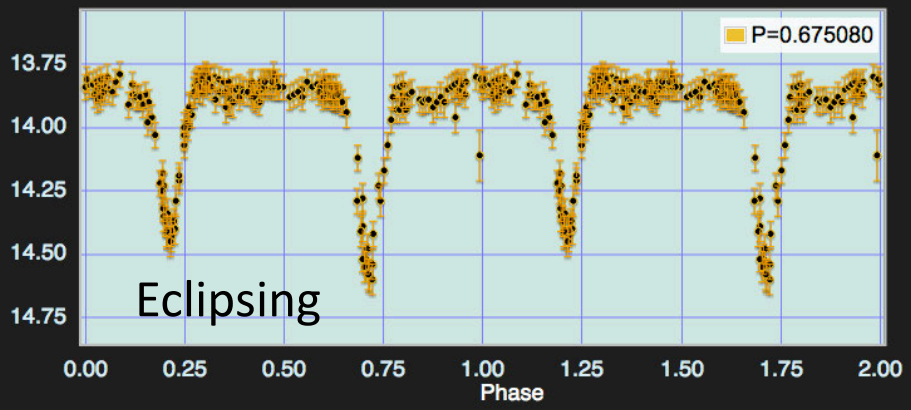


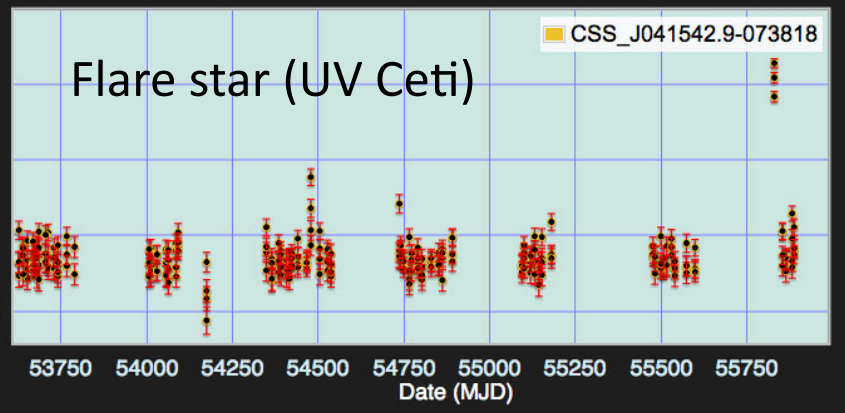
Discovery data Archival data Light curve+images Finding chart

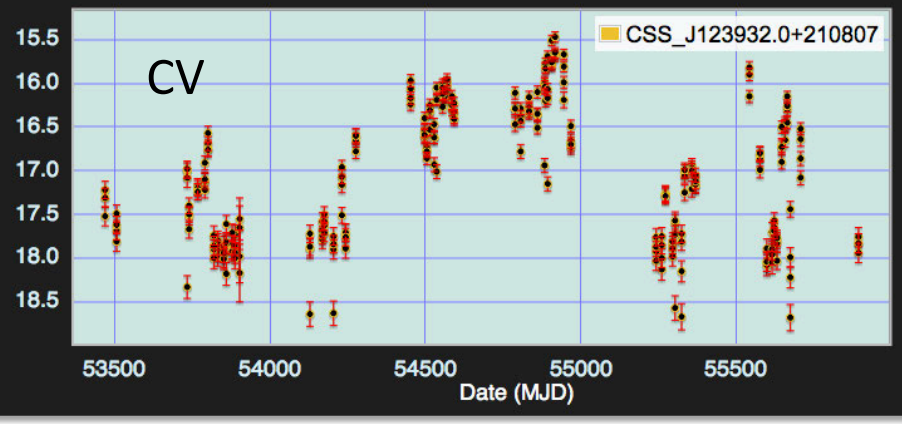
500 Million Light Curves with ~ 10^{11} data points

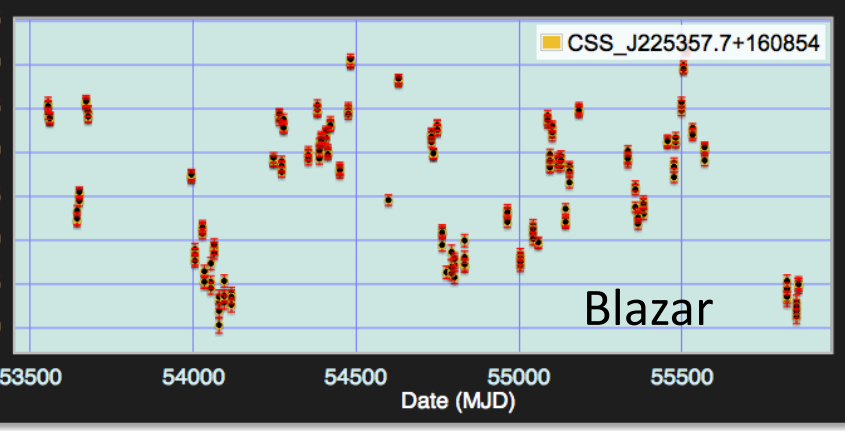






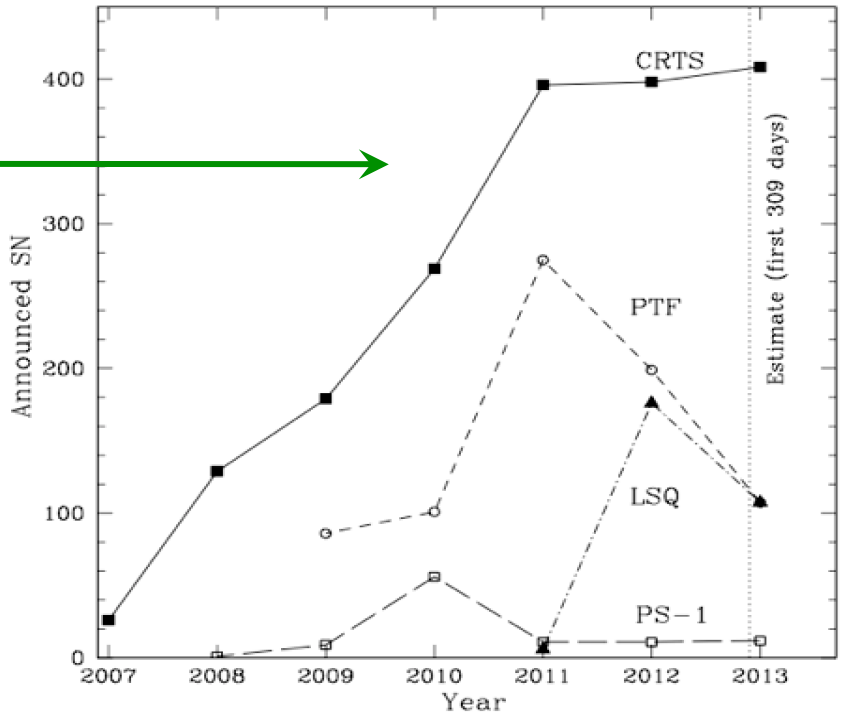


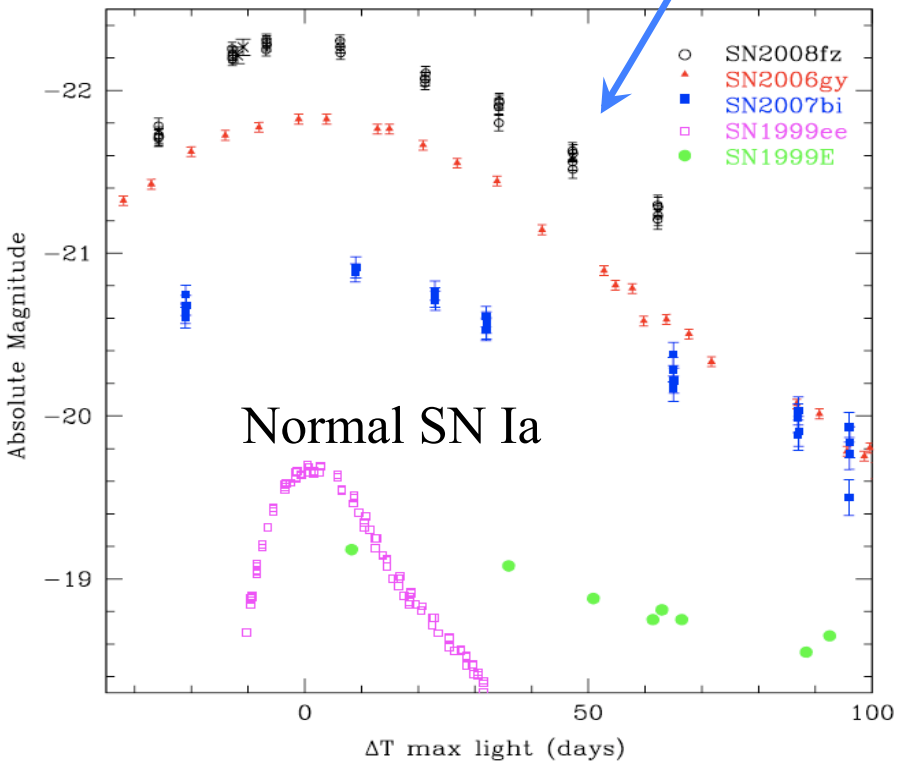


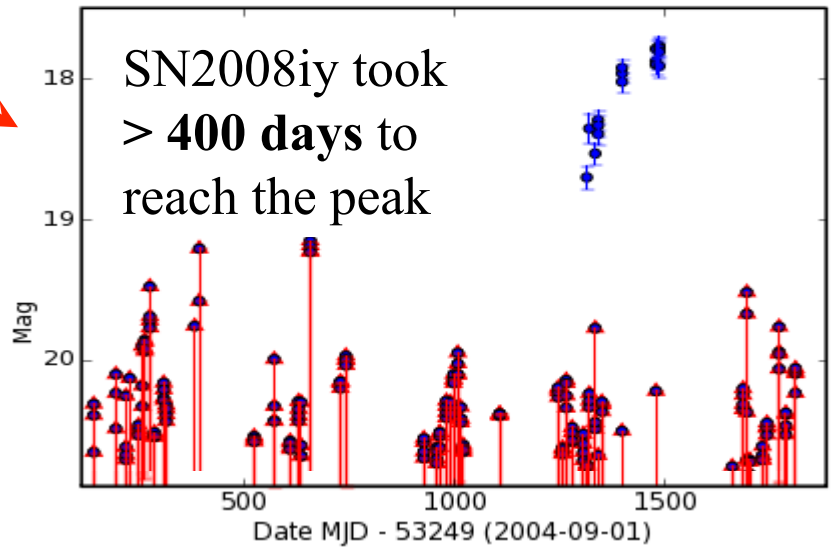


CRTS Supernova Discoveries

- More SNe published in the last 5 years than any other survey
- Extremely luminous and possible pair-production SNe
- Extremely long time-scale SNe

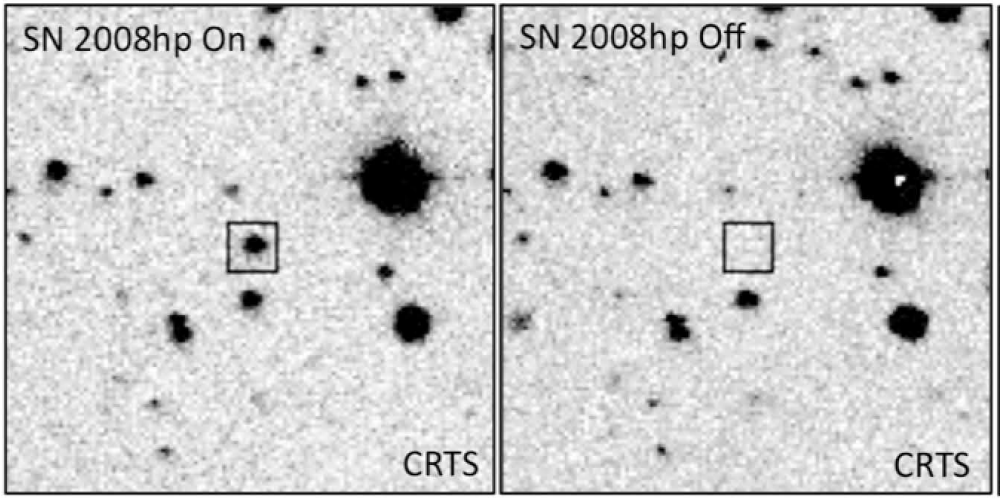




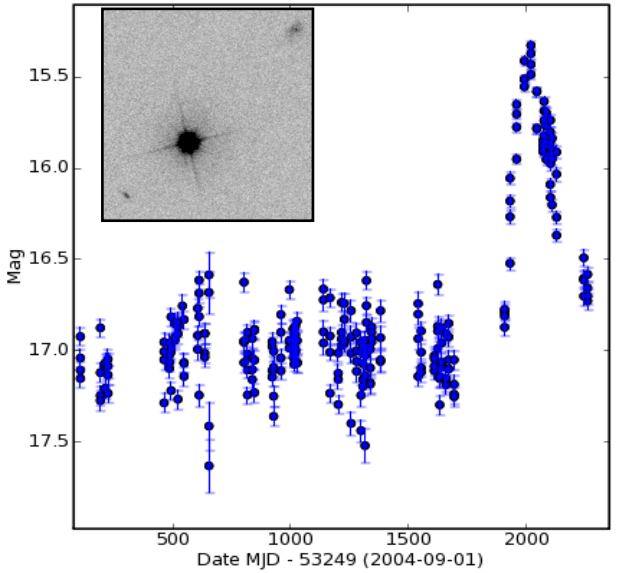


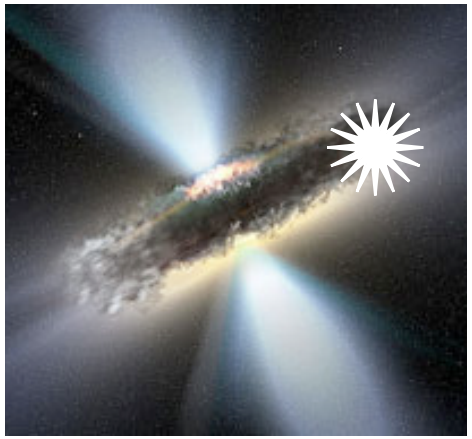
Luminous SNe in Underluminous Hosts

- A number of SNe discovered in extremely faint dwarf galaxy hosts (M ≈ –12 or –13)
  - ⇨ Huge specific SN rates (per unit stellar mass)
- Local Pop. III analogs?



A New Kind of a Supernova?

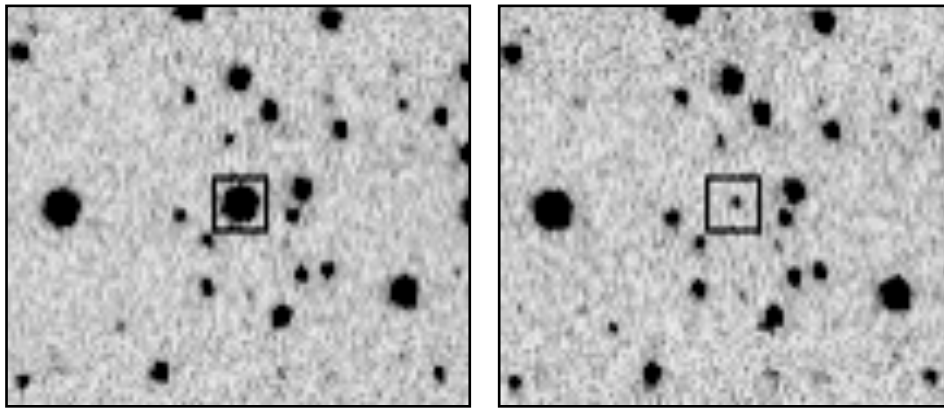


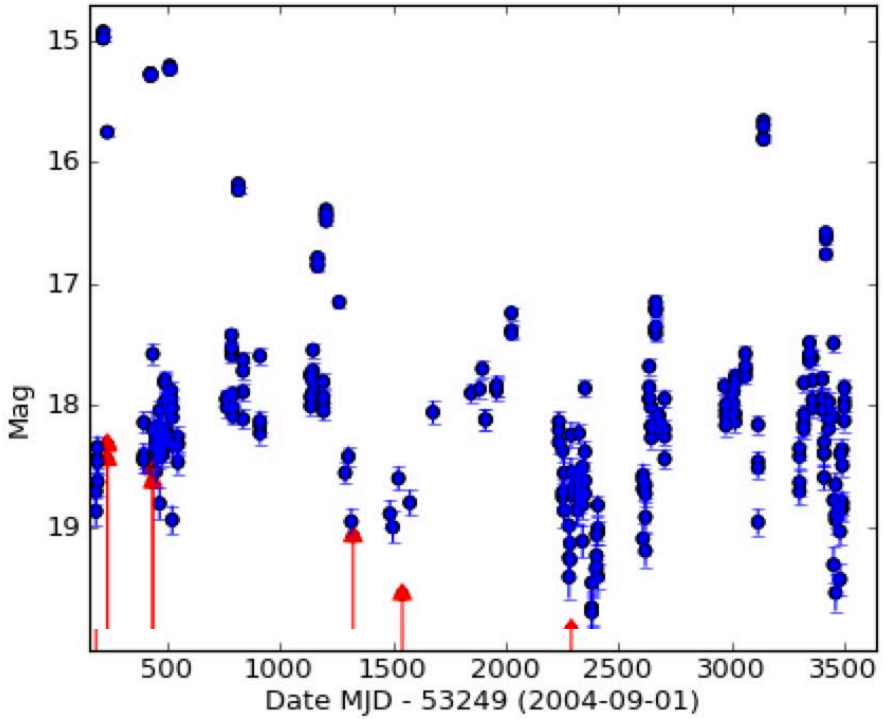


- The first case of a SN from an AGN accretion disk? (Predicted by theory, but never seen before)
- The most luminous SN ever seen

Cataclysmic Variables and Dwarf Novae

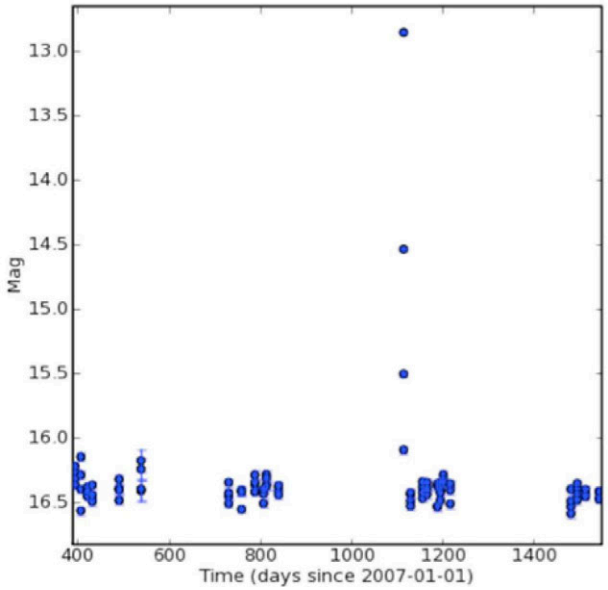
> 1,000 detected so far, > 75% are new discoveries



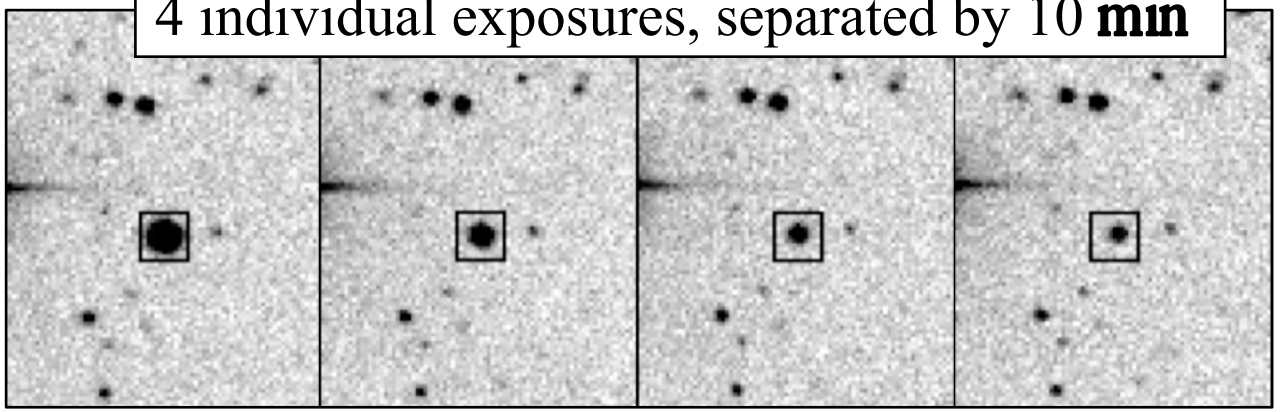


Fast Transients

Mostly dM



4 individual exposures, separated by 10 **min**



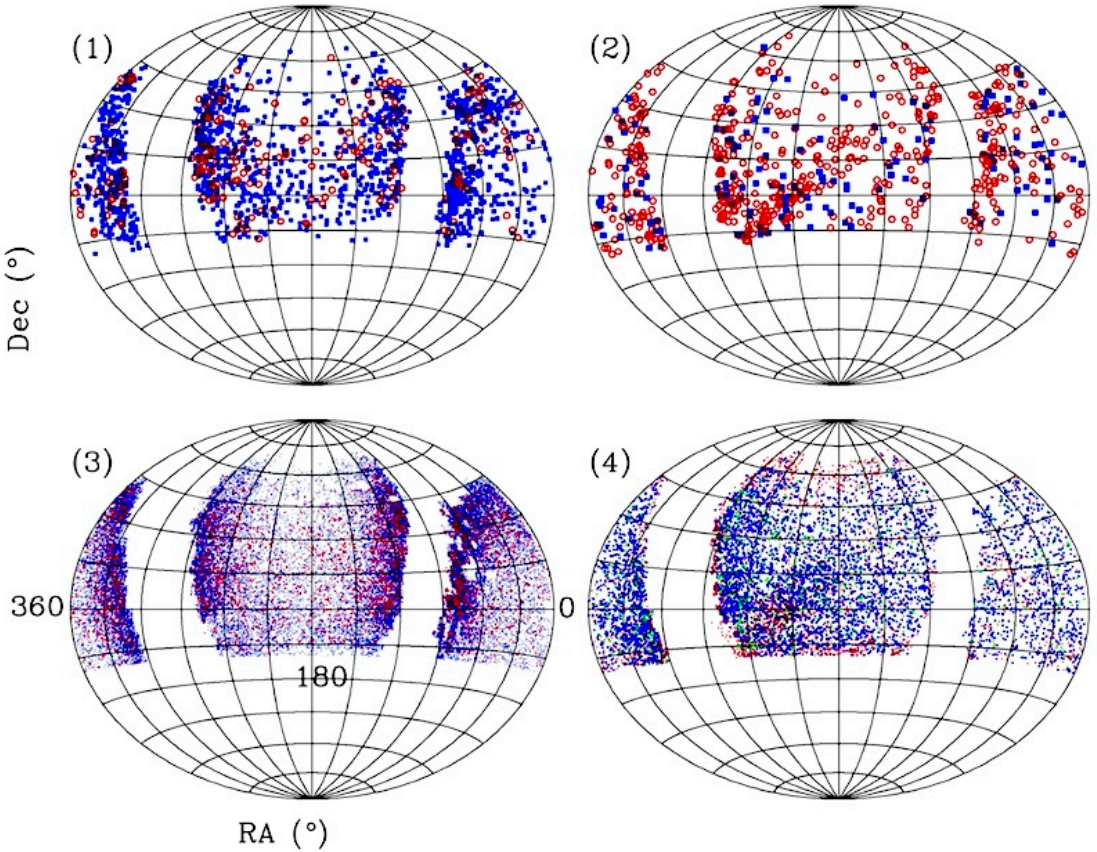
>60,000 New Periodic Variables

Simple data mining of the CRTS DR1 light curves archive:

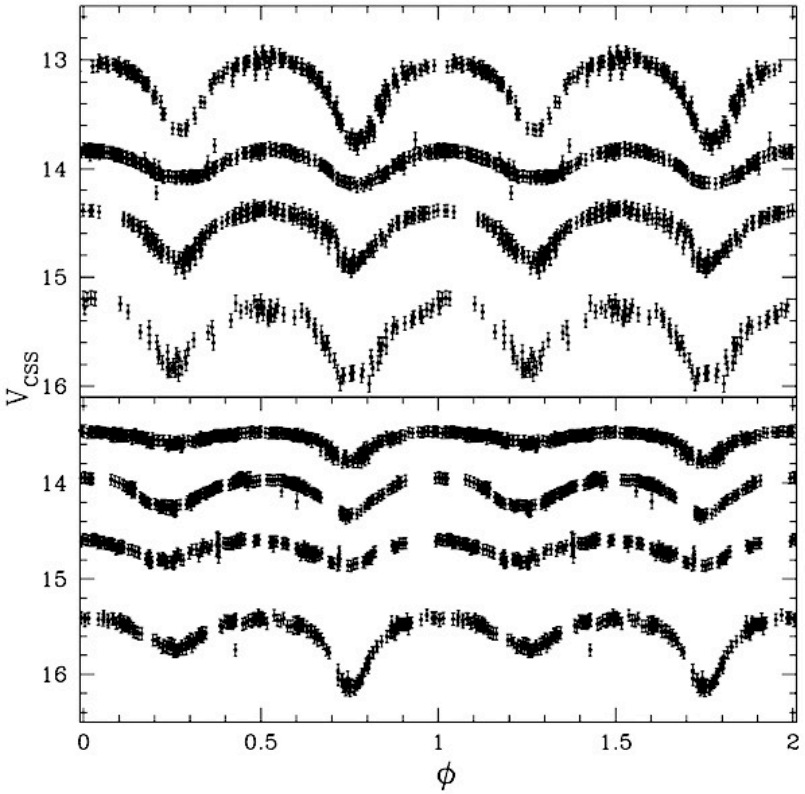
200 million light curves

➪ 5.4 million variable star cand.

➪ 61,000 new *periodic* var's



Examples of eclipsing contact binaries



← Sky distribution of various types of periodic variables

(Drake et al. 2014)

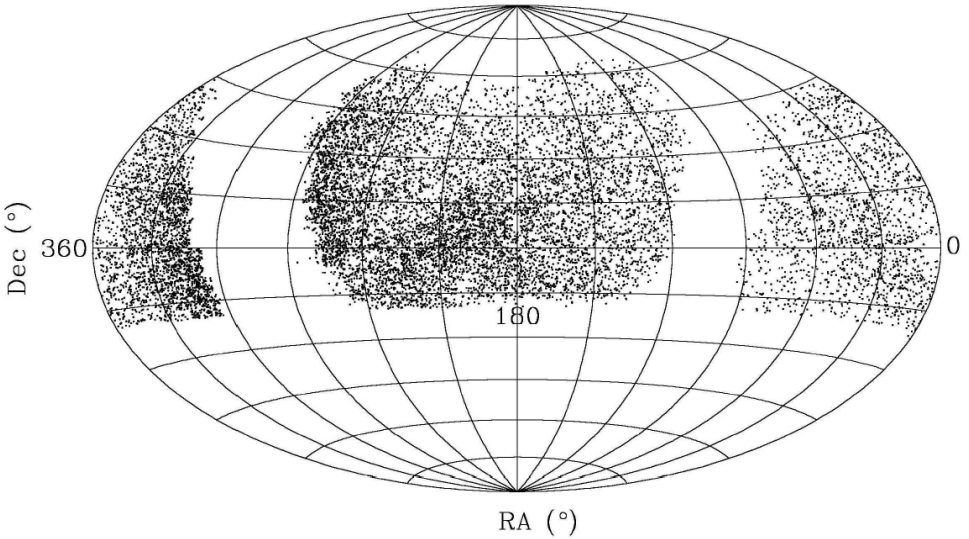
~ 23,000 New RR Lyrae Stars

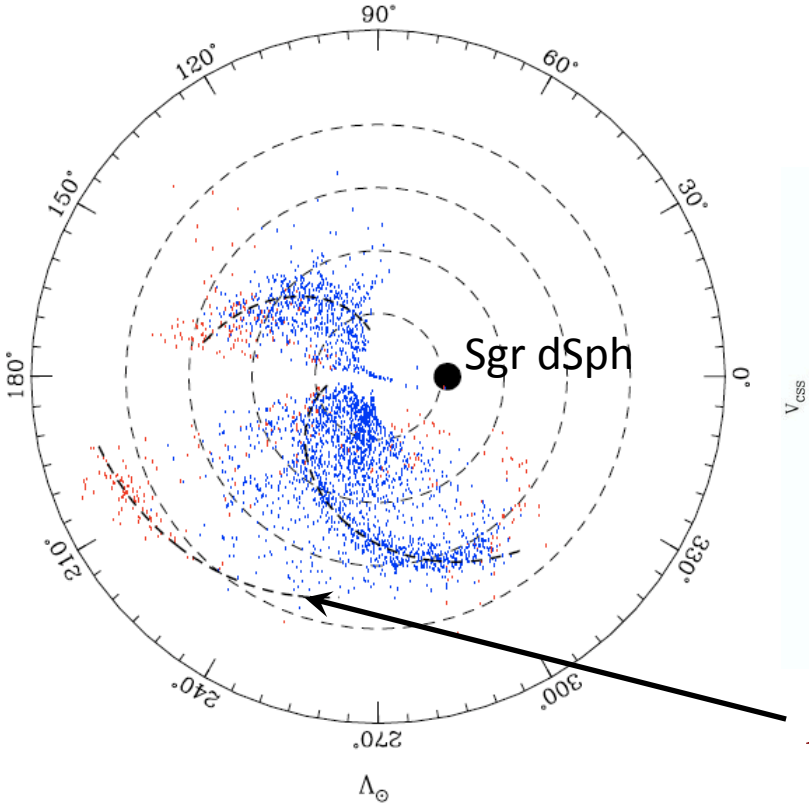
Simple data mining of the CRTS light curves archive:

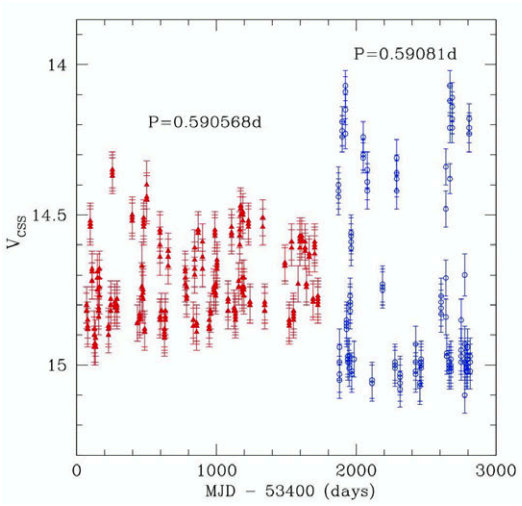
(Drake et al. 2012, 2013)

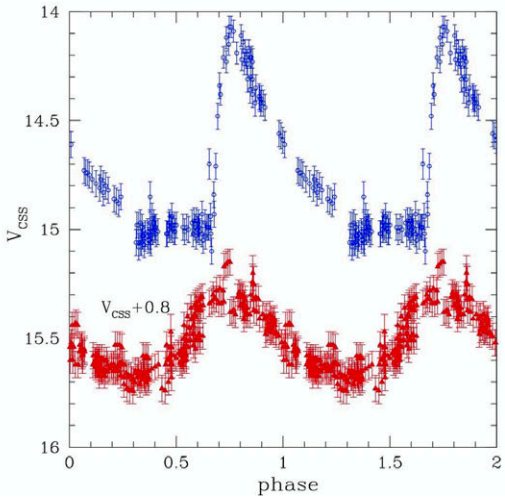
500 million light curves

- ➪ 13 million variable star cand.
  - ➪ 23,000 new RR Lyrae



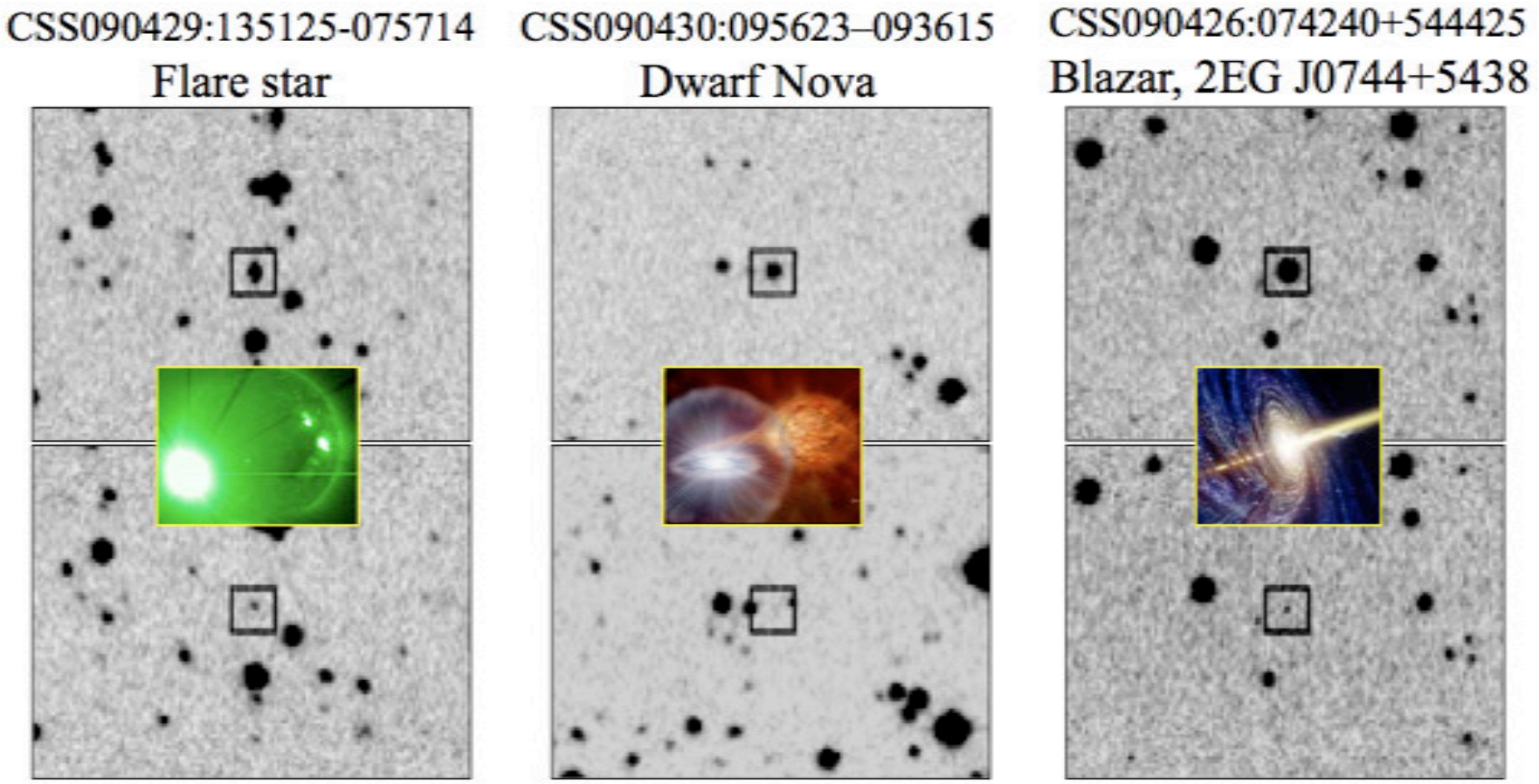






A new tidal stream reaching out to 100 kpc

Automated Classification of Transients



Vastly different physical phenomena, and yet they look the same!
Which ones are the most interesting and worthy of follow-up?

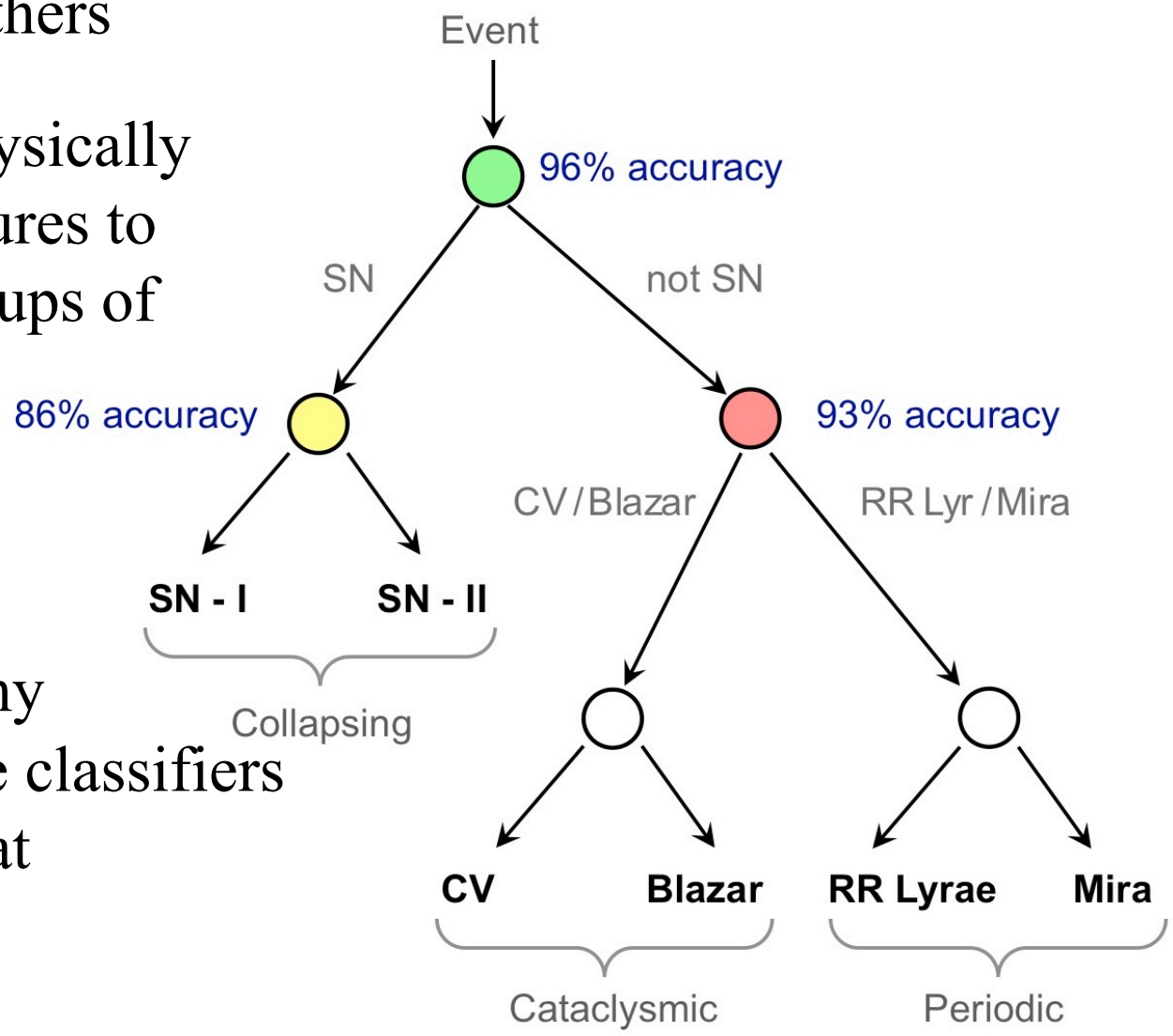
Rapid, automated transient classification is a critical need!
(especially as the discovery rates increase by orders of magnitude)

A Hierarchical Approach to Classification

Different types of classifiers perform better for some event classes than for the others

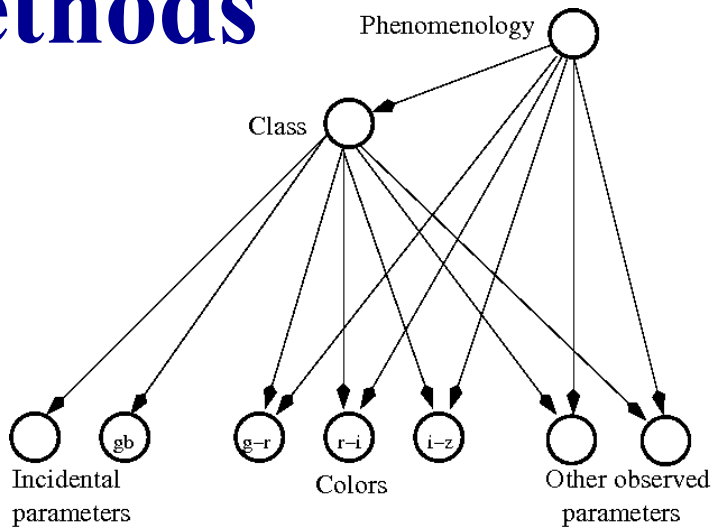
We use some astrophysically motivated major features to separate different groups of classes

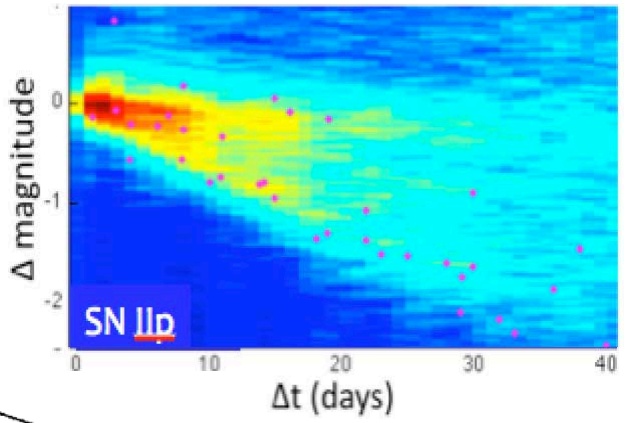
Proceeding down the classification hierarchy every node uses those classifiers that work best for that particular task

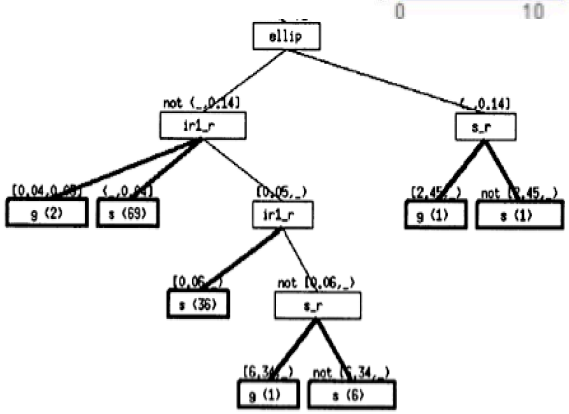


A Variety of Methods

- **Bayesian Networks**
  - Can incorporate heterogeneous and/or missing data
  - Can incorporate contextual data, e.g., distance to the nearest star or galaxy
- **Probabilistic Structure Functions**
  - A new method, based on 2D [Δt_1, Δm] distributions
  - Now expanding to data point triplets: Δt_{12}, Δm_{12}, Δt_{23}, Δm_{23}, giving a 4D histogram
- **Random Forests**
  - Ensembles of Decision Trees
- **Feature Selection Strategies**
  - Optimizing classifiers
- **Machine-Assisted Discovery**



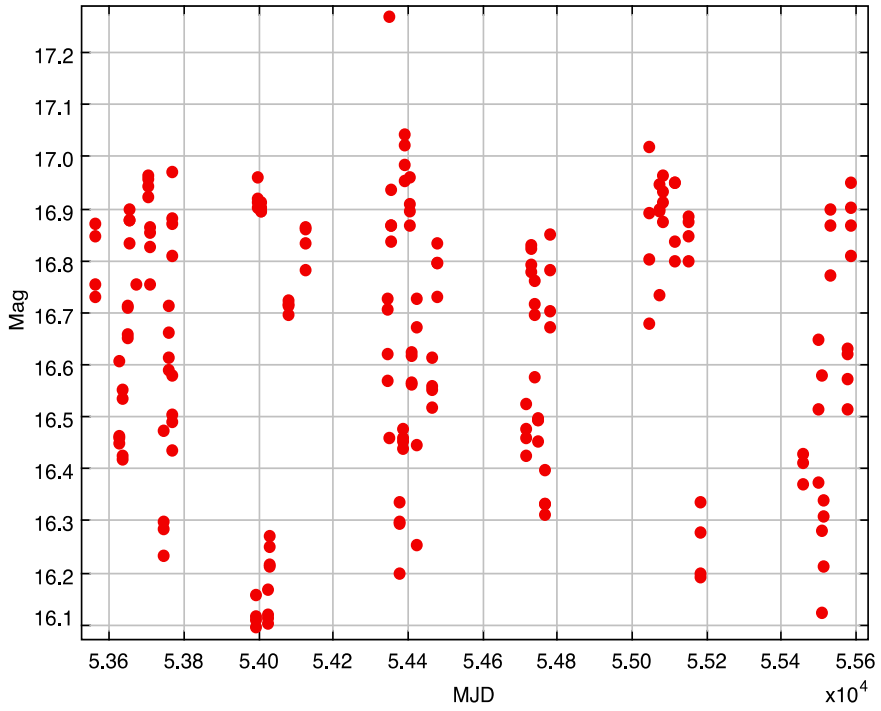


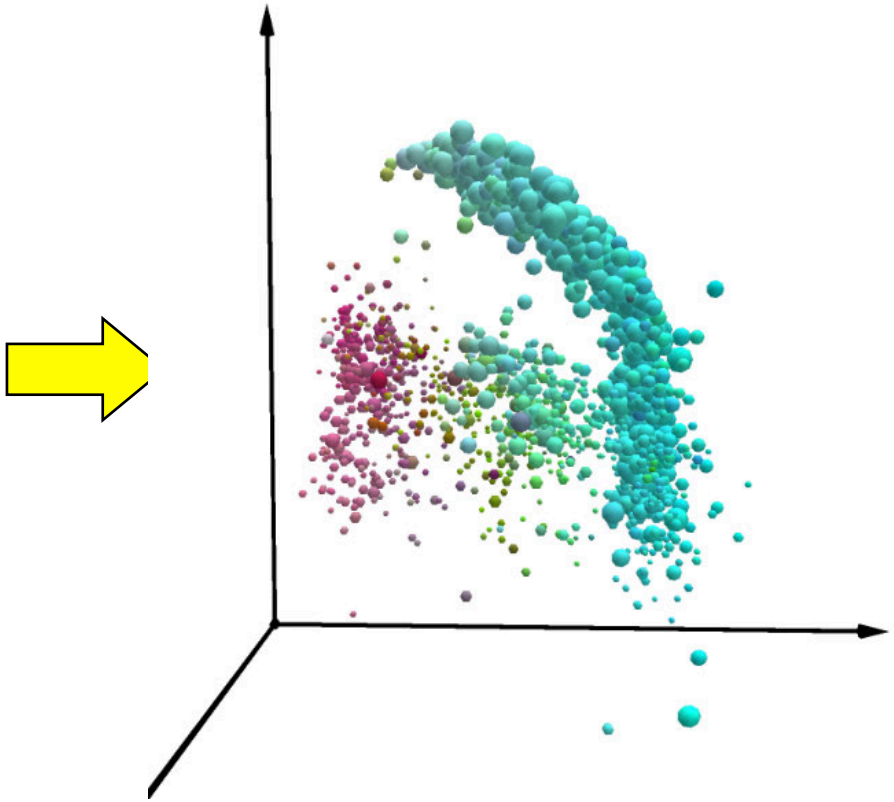


etc., etc.

From Light Curves to Feature Vectors

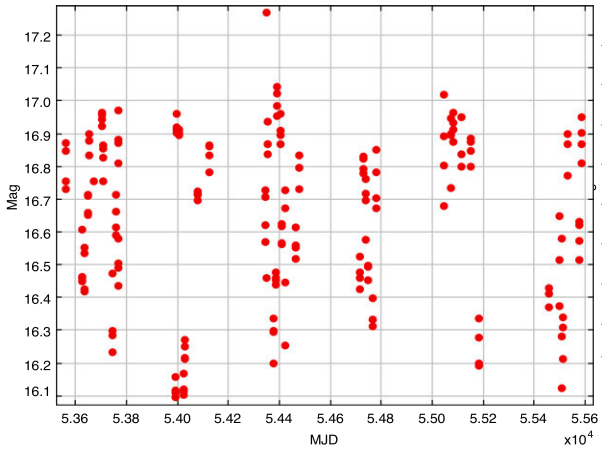
- We compute ~ 70 parameters and statistical measures for each light curve: amplitudes, moments, periodicity, etc.
- This turns heterogeneous light curves into homogeneous ***feature vectors*** in the parameter space
- Apply a variety of automated classification methods



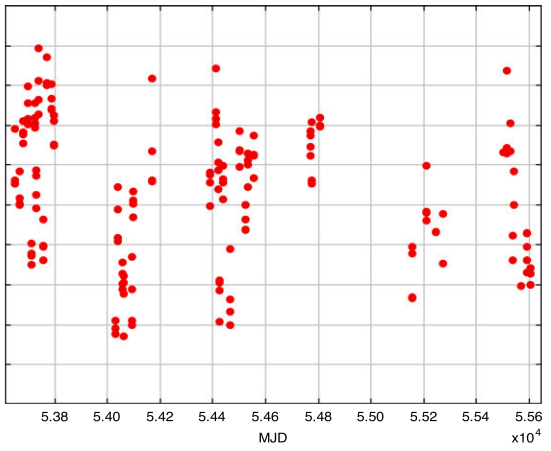


Optimizing Feature Selection

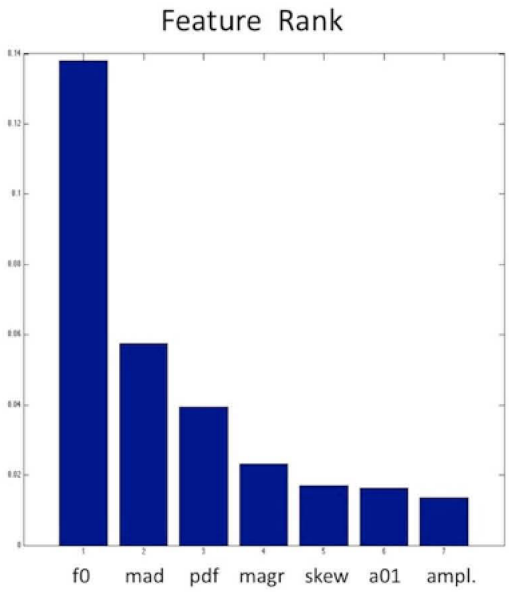
Rank features in the order of classification quality for a given classification problem, e.g., RR Lyrae vs. WUMa

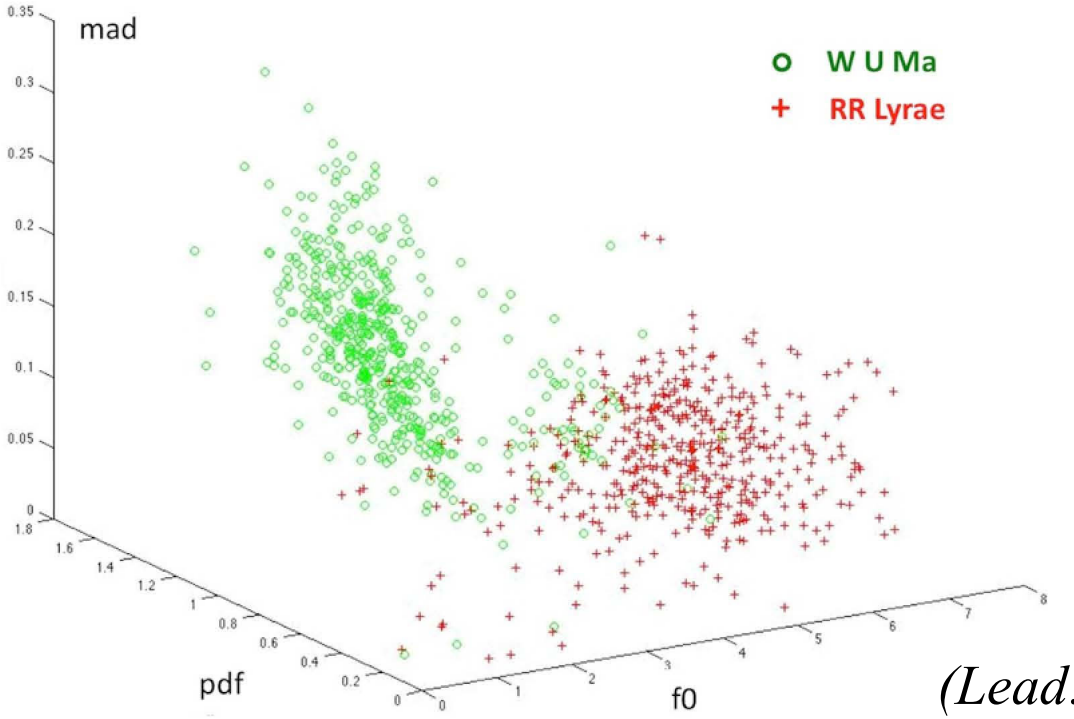


RR Lyrae



Eclipsing binary (W U Ma)



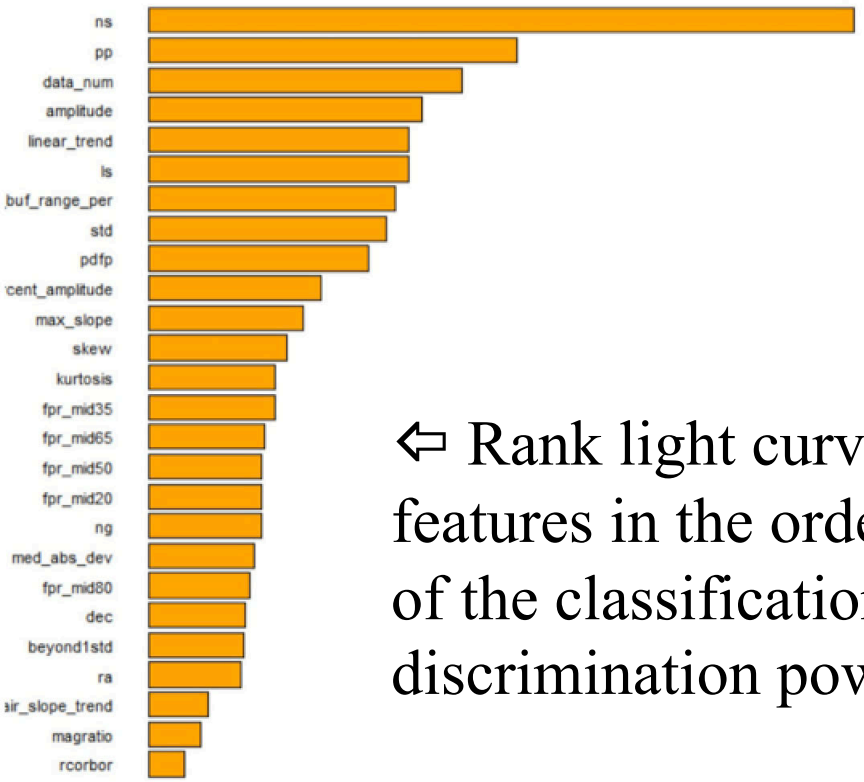


(Lead: C. Donalek)

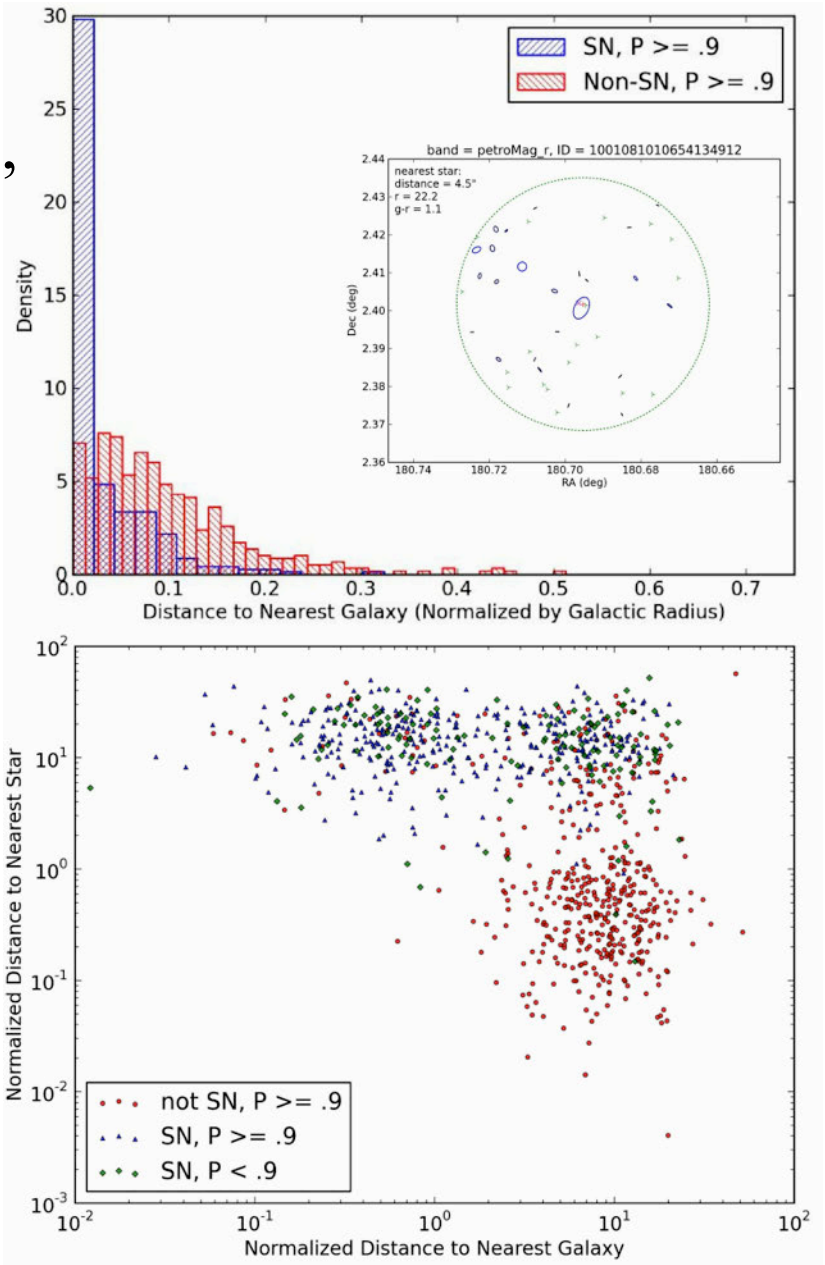
Bayesian Networks: An Example

(Lead: A. Mahabal)

Can incorporate contextual parameters, e.g., the normalized distances to the nearest star and the nearest galaxy as one of the BN variables ⇨



⇦ Rank light curve features in the order of the classification discrimination power



From the Information Technology to the *Cognition Technology*: Towards a Human-Computer Collaborative Discovery

As the complexity of data increases, we may be increasingly reliant on the uses of machine intelligence to help us discover interesting patterns in the data

AS WE MAY THINK

A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLI
IN WHICH MAN-MADE MACHINES WILL START TO THINI

by VANNEVAR BUSH

Vannevar Bush (1945)

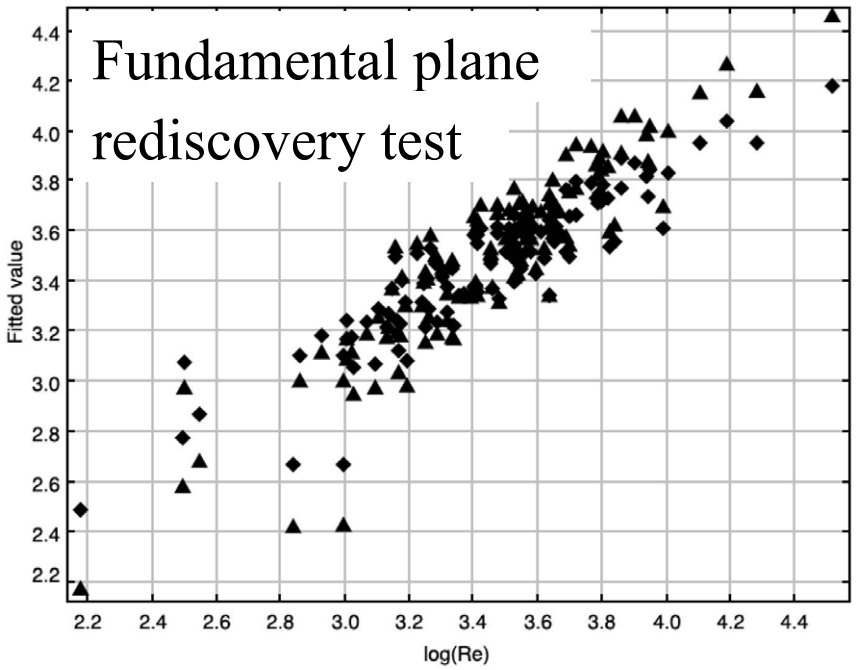
Man-Computer Symbiosis

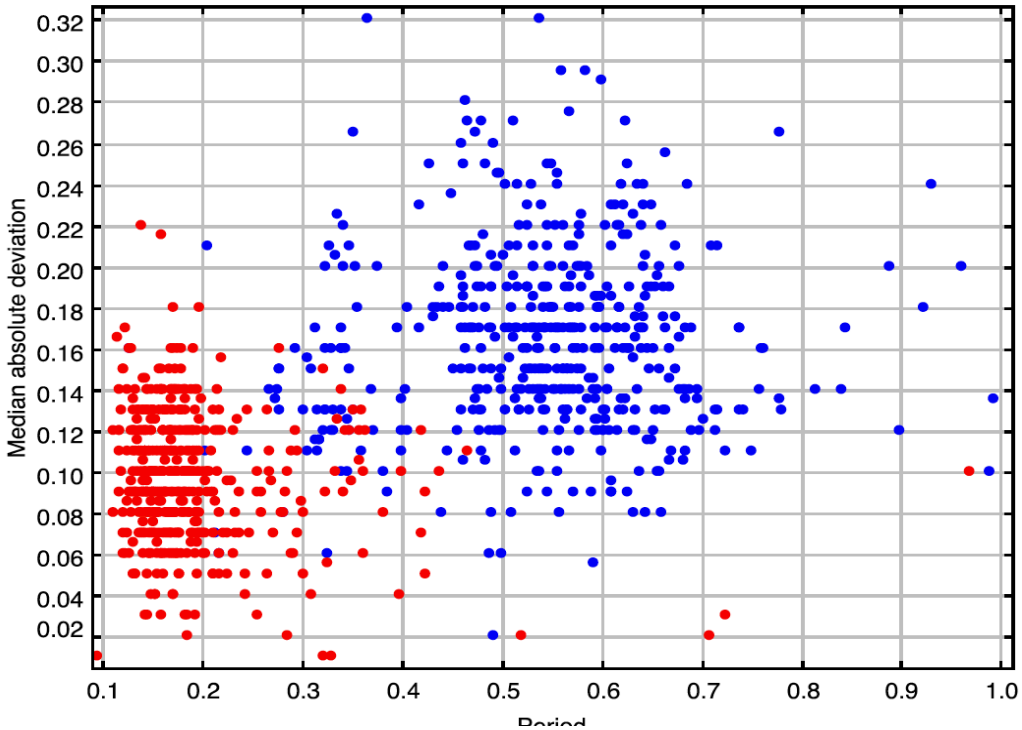
J.C.R. Licklider (1960)

Machine Discovery Using *Eureqa*

Lipman et al., http://creativemachines.cornell.edu/eureqa

- Employs symbolic regression to determine best-fitting functional form to data and its parameters simultaneously
- Specify building blocks to be used: algebraic operators, analytical functions, constants
- See Graham et al. (2013)

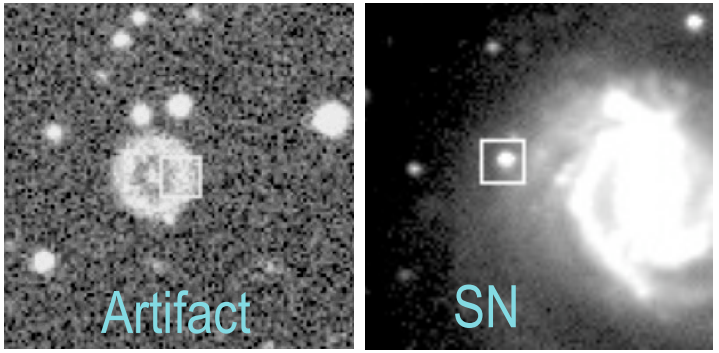


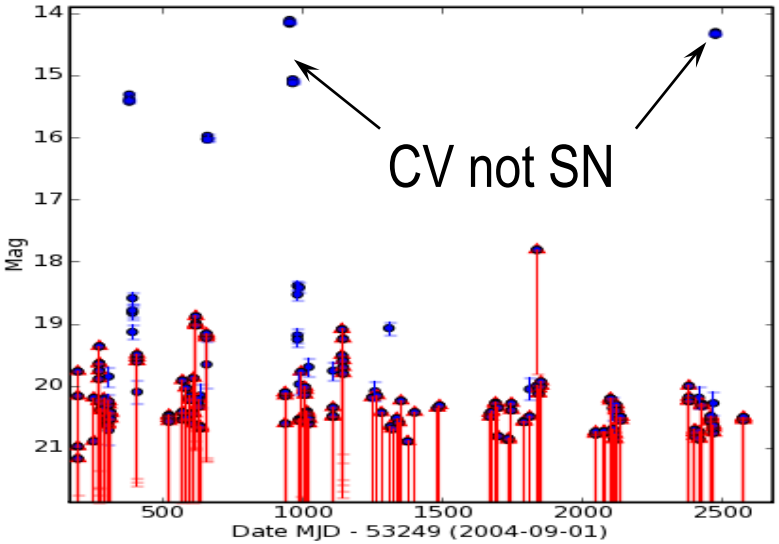


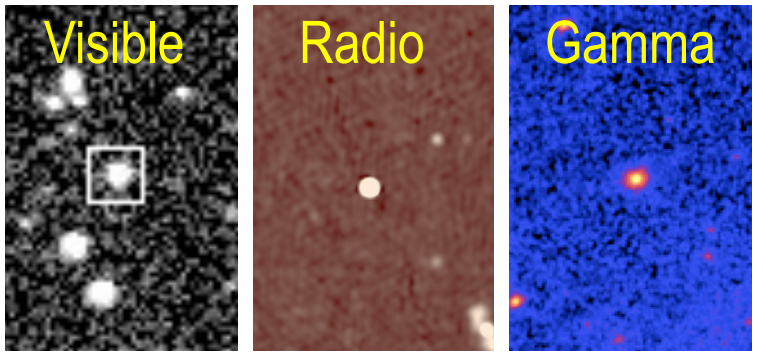
Classification of variable stars

Contextual Information is Essential

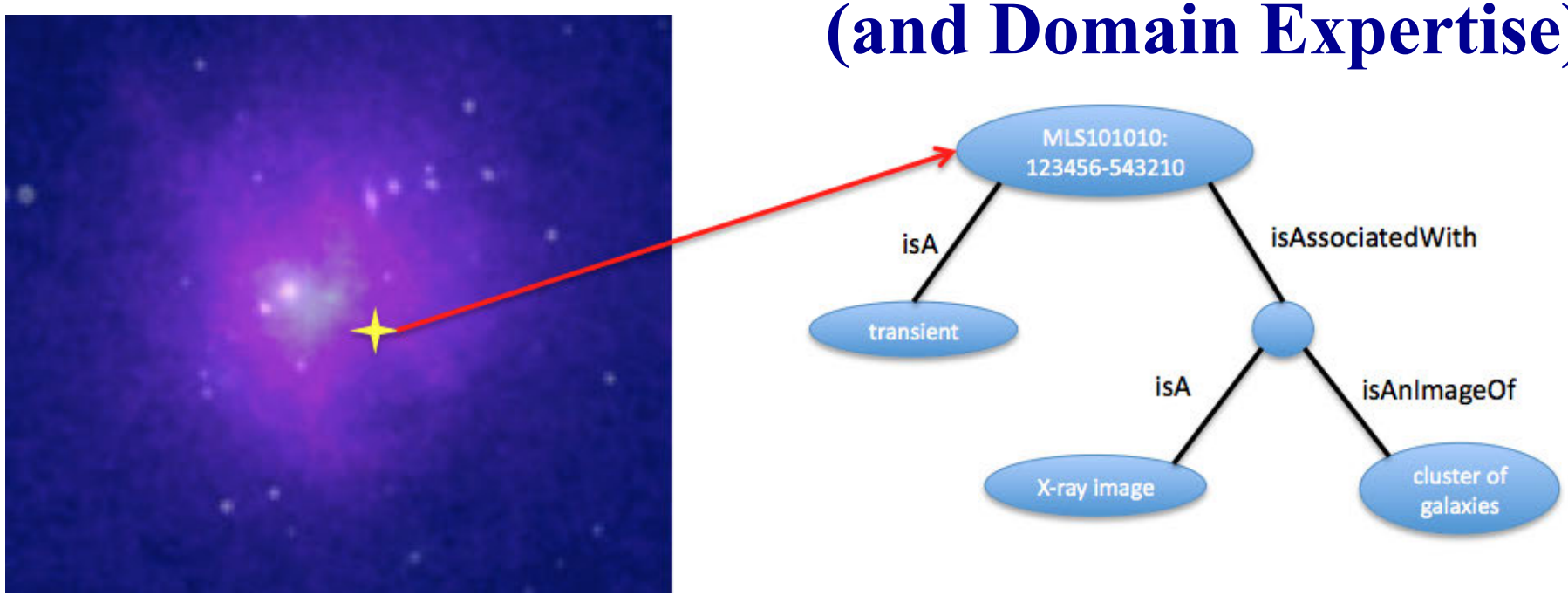
- **Visual context** contains valuable information about the reality and classification of transients
- So does the **temporal context**, from the archival light curves
- And the **multi-λ context**
- Initial detection data contain little information about the transient: α, δ, m, Δm, (t_c). ***Almost all of the initial information is archival or contextual;*** follow-up data trickle in slowly, if at all
- VO data grid is an essential resource







Harvesting the Human Pattern Recognition (and Domain Expertise)



Human-annotated images (via *SkyDiscovery.org*)

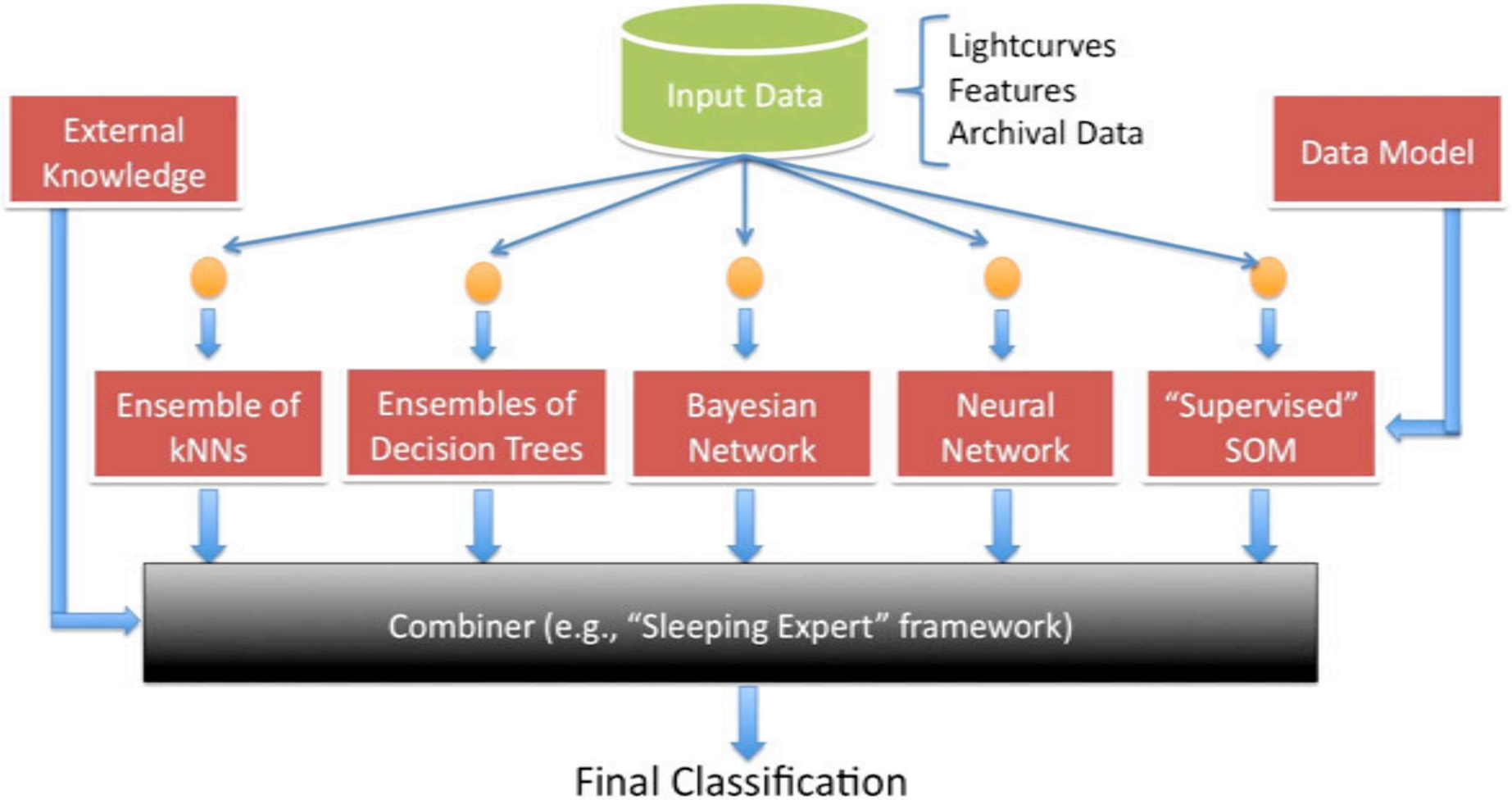
- ➥ Semantic descriptors
  - ➥ Machine processing
    - ➥ Evolving novel algorithms

… and iterate

Challenges: Optimizing for different levels of user expertise; optimal input averaging; encoding contextual information; etc.

(Lead: M. Graham)

Metaclassification: An optimal combining of classifiers

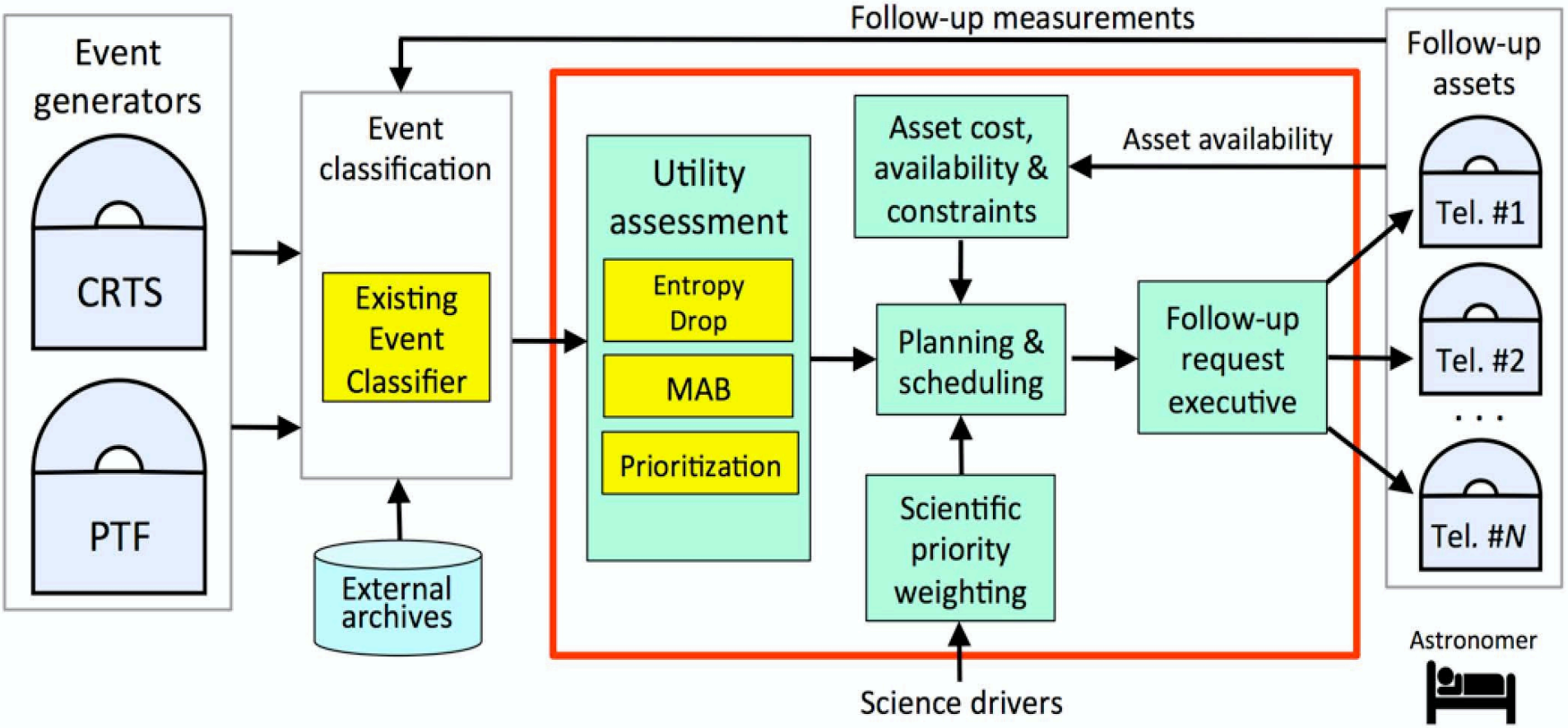


Exploring a variety of techniques for an optimal classification fusion: Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit, Sleeping Expert…

(Lead: C. Donalek)

Automating the Optimal Follow-Up

For the *potentially most interesting events*, what type of follow-up observations has the greatest potential to discriminate among the competing event classes, given the available assets, and the potential scientific value?



Conclusions

- Time domain astronomy is ***here now*** (CRTS, PTF, PS1, *SkyMapper*, ASCAP, *Kepler*, *Fermi*, …), and it is a vibrant new frontier
  - Lots of exciting and diverse science already under way, from the Solar system to cosmology – something for everyone!
  - Surveys today are science and methodology precursors and testbeds for the LSST, and they are delivering science now
  - CRTS data stream is *open* – use it! (and free ≠ bad)
- It is ***astronomy of telescope and computational systems***, requiring a ***strong cyber-infrastructure*** (VO, astroinformatics)
  - ***Automated classification is a core problem***; it is critical for a proper scientific exploitation of synoptic sky surveys
  - Data mining of Petascale data streams both in real time and archival modes is ***important well beyond astronomy***
- CRTS II consortium now forming – join us!

Some Useful Links:

- CRTS survey (includes links to the data, etc.):

  http://crts.caltech.edu/

- Keck Institute for Space Studies (KISS) study, "Digging Deeper (and Faster): Algorithms for Computationally-Limited Searches in Astronomy":

  http://kiss.caltech.edu/study/digging/

  (links to presentations from 2 workshops, and the final report)

- Papers in MNRAS, astro-ph …

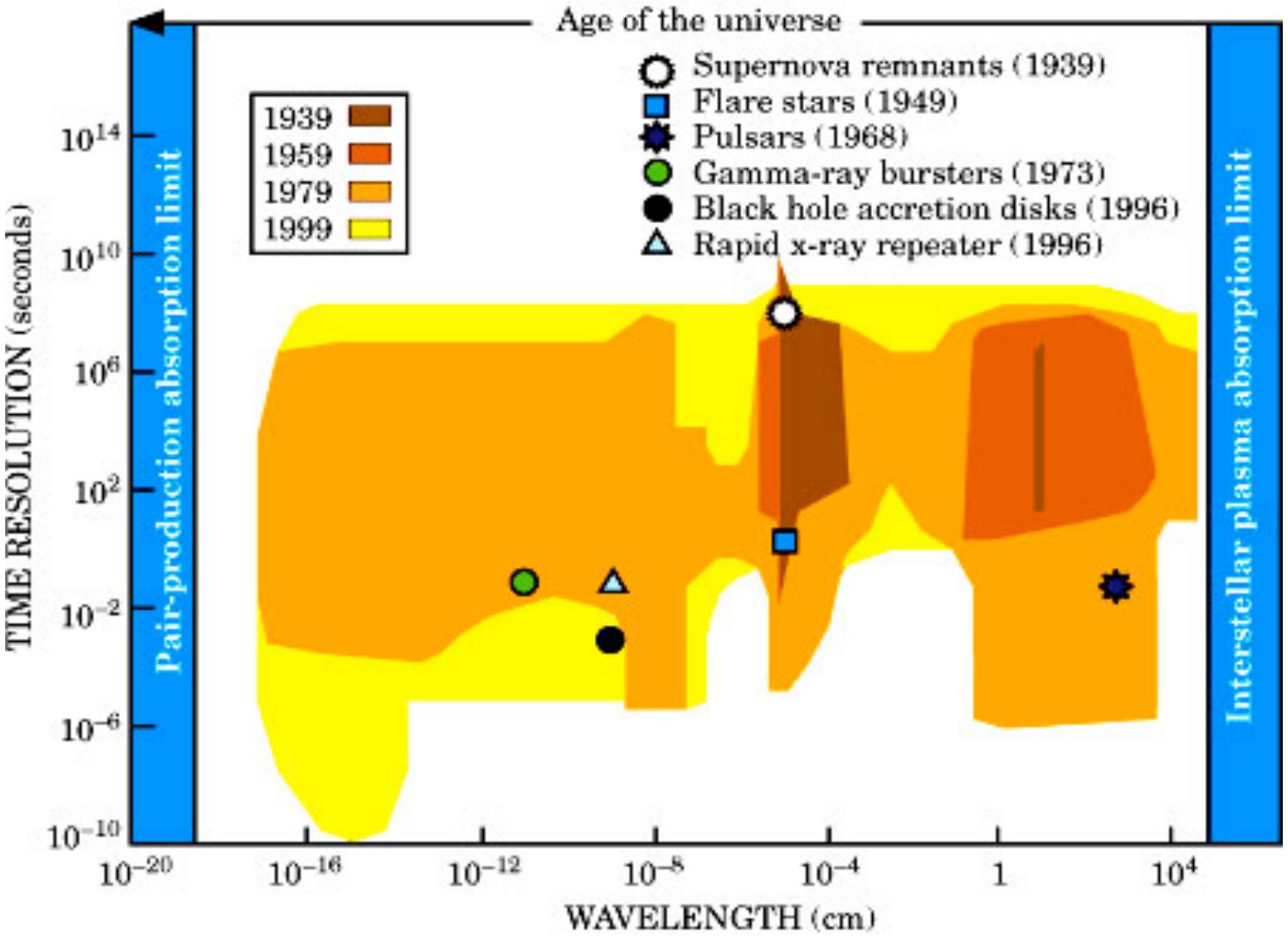
Expanding the Observable Parameter Space

Technology advances ➞ Expanded domain of measurements

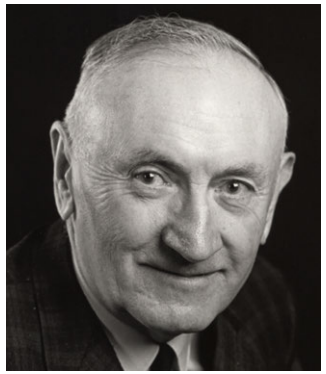
➞ Discovery of new types of phenomena

M. Harwit

Recent example: FRBs

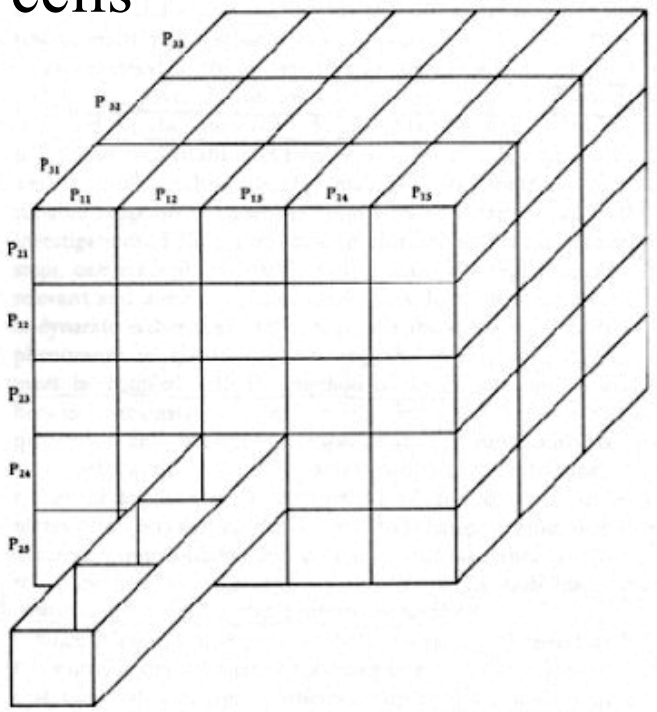


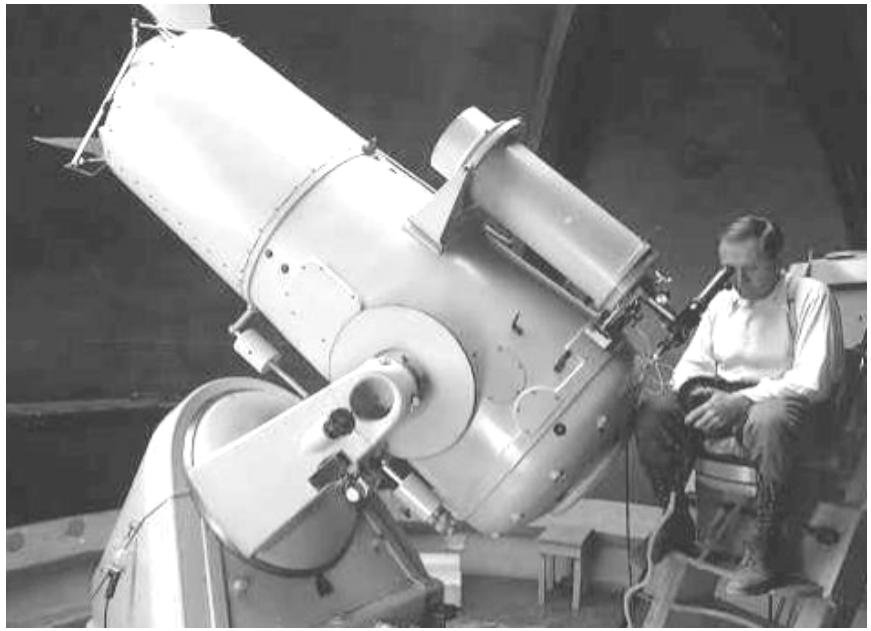
From "Morphological Box" to the Observable Parameter Spaces



Fritz Zwicky

Zwicky's concept: explore all possible combinations of the relevant parameters in a given problem; these correspond to the individual cells in a "Morphological Box"





Example: Zwicky's discovery of the compact dwarfs

Characterizing Synoptic Sky Surveys

Define a measure of **depth** (roughly ~ S/N of indiv. exposures):

$$D = [\, A \times t_{exp} \times \varepsilon \,]^{1/2} \;\; / \, FWHM$$

where A = the effective collecting area of the telescope in m^2

t_{exp} = typical exposure length

ε = the overall throughput efficiency of the telescope+instrument

$FWHM$ = seeing

Define the **Scientific Discovery Potential** for a survey:

$$SDP = D \times \Omega_{\rm tot} \times N_b \times N_{avg}$$

where $\Omega_{\rm tot}$ = total survey area covered

N_b = number of bandpasses or spec. resolution elements

N_{avg} = average number of exposures per pointing

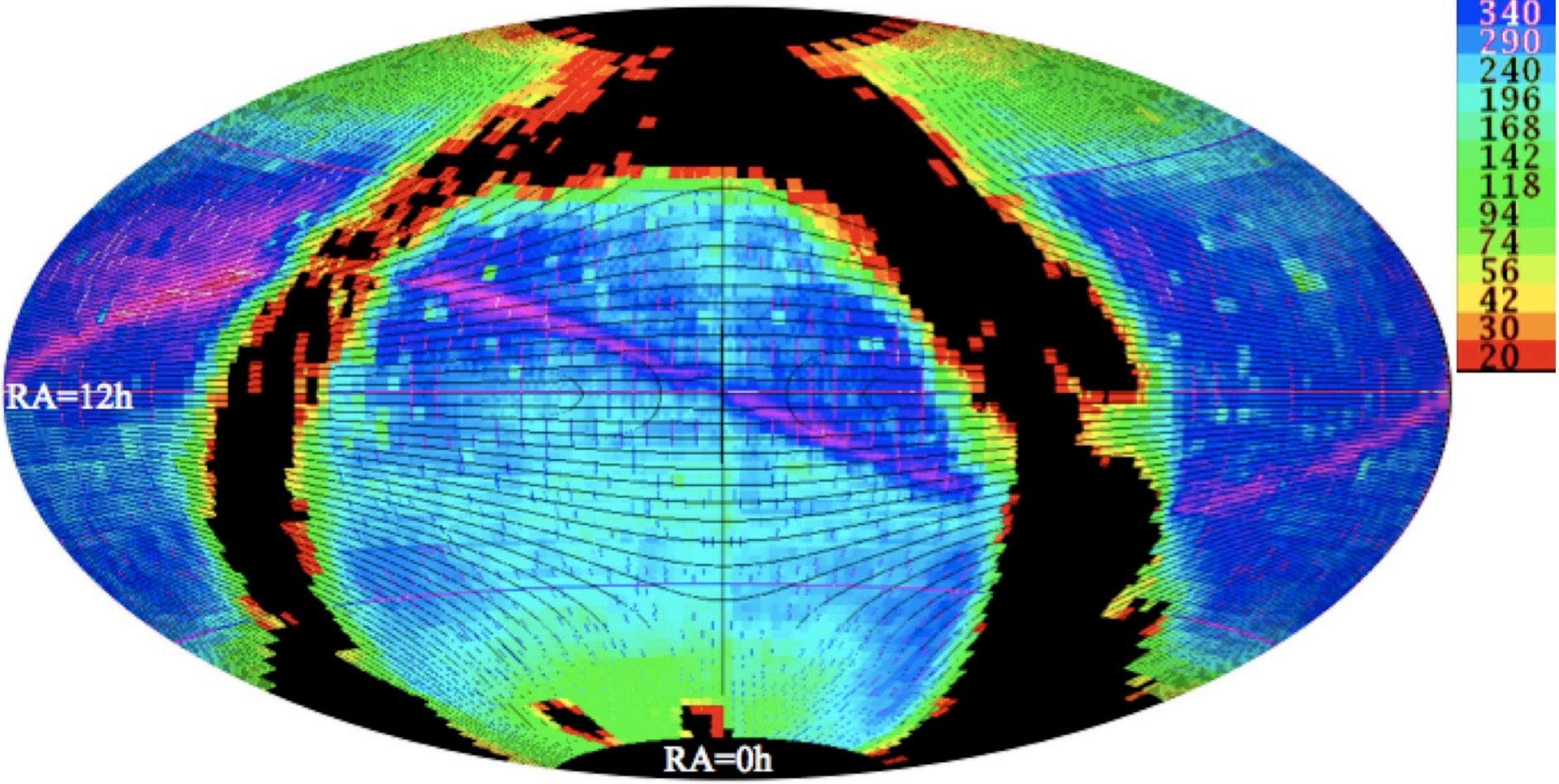
Transient Discovery Rate:

$$TDR = D \times R \times N_e$$

where $R = d\Omega/dt$ = area coverage rate

N_e = number of passes per night

CRTS Sky Coverage (as of Aug. 2011)



Total ≈ 33,000 deg^2 ≈ 82% of the entire sky (more than any other survey)

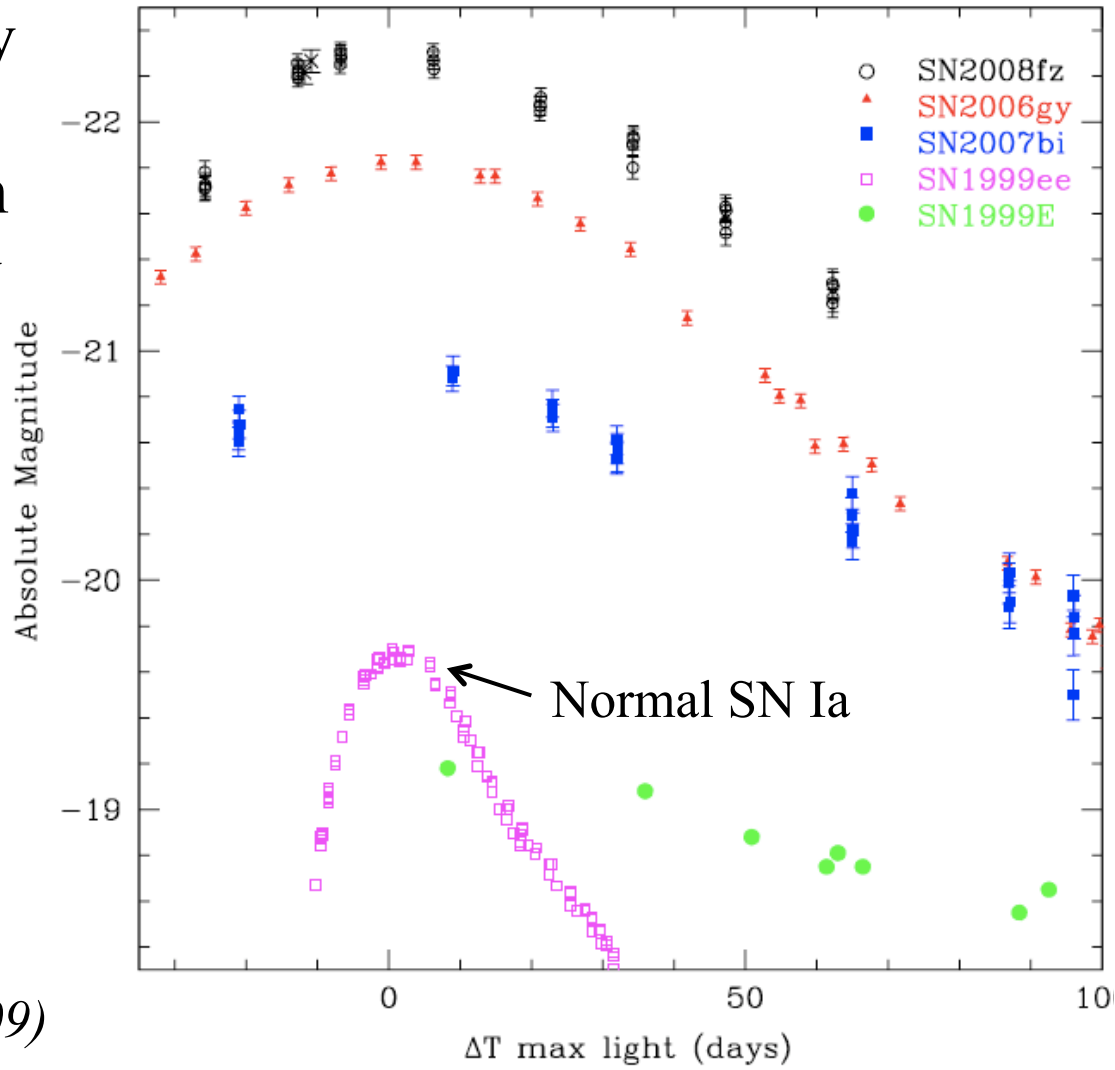
Limiting mags ~ 19 – 21 per pass. Image coadds reach r ~ 24 mag

Time baselines from 10 min to ~ 8 years (and growing)

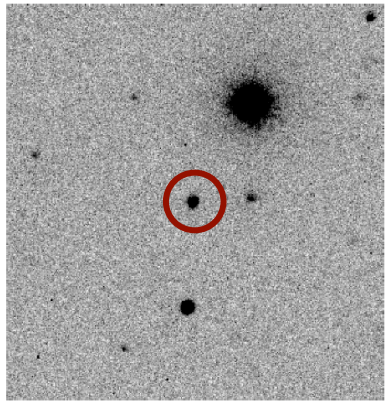
2008fz: The Most Luminous Supernova

(Then)

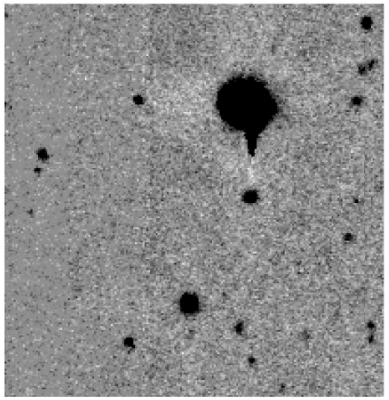
- Brightest type IIn known (5 times brighter than the Milky Way); surpassed by subsequent CRTS discoveries
- Host galaxy > 50 times fainter than Milky Way
- A possible example of a pair-production SN?



Discovery

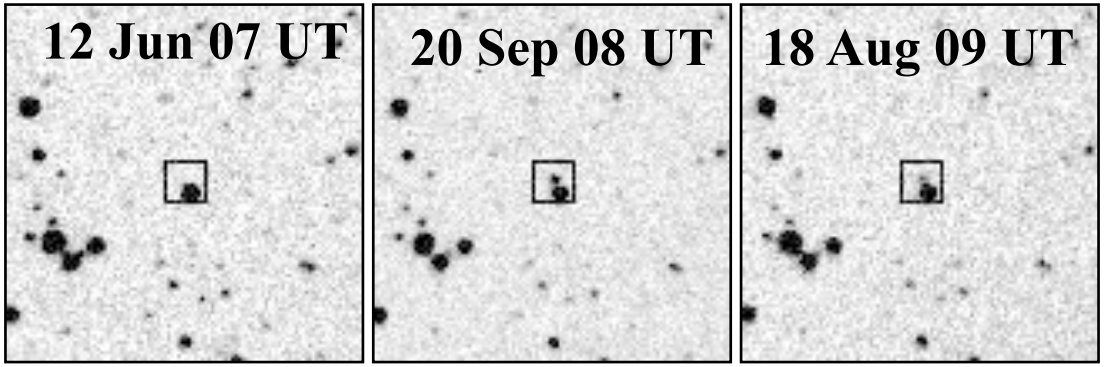


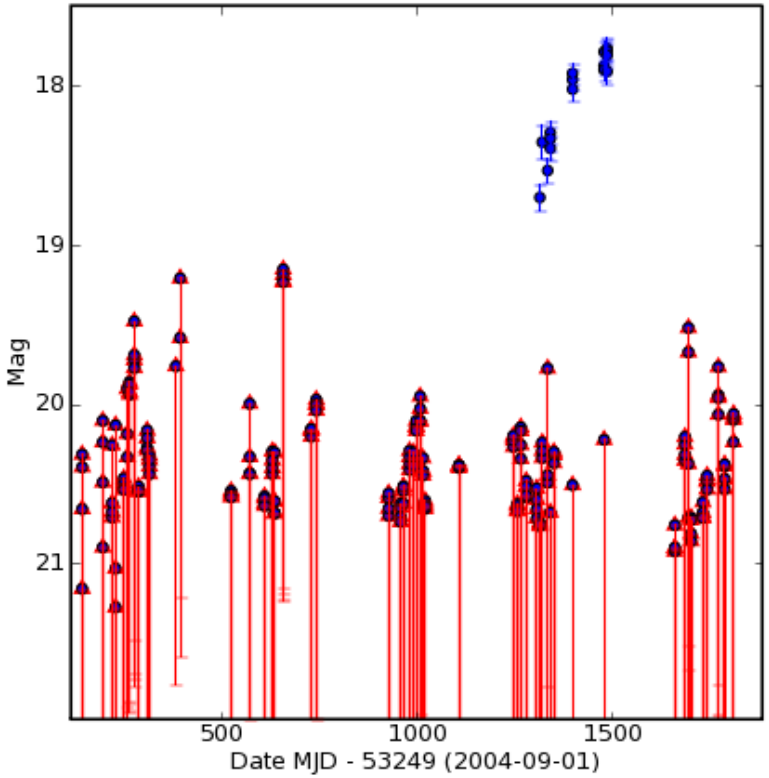
Comparison



(Drake et al. 2009)

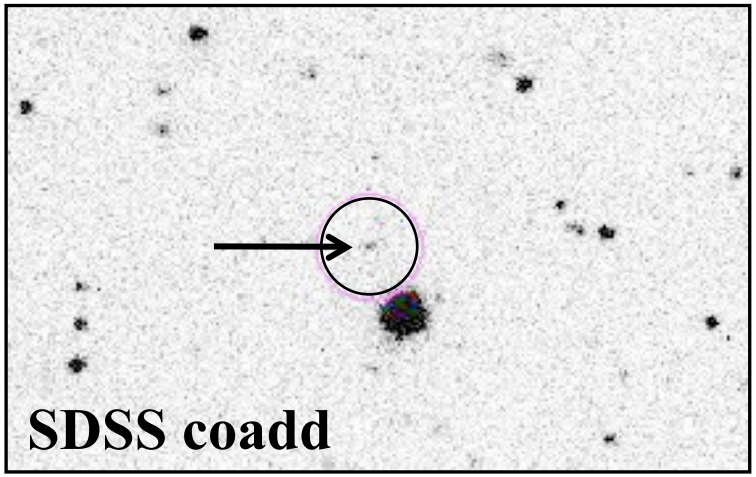
The Ultra-Slow SN 2008iy = CSS080928:160837+041627



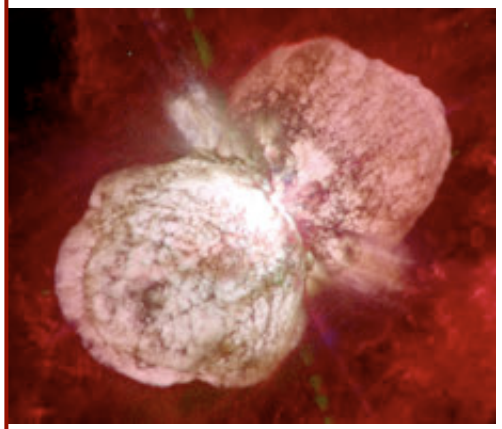


Longest-lasting type IIn at z = 0.041
it took **> 400 days** to reach the peak!

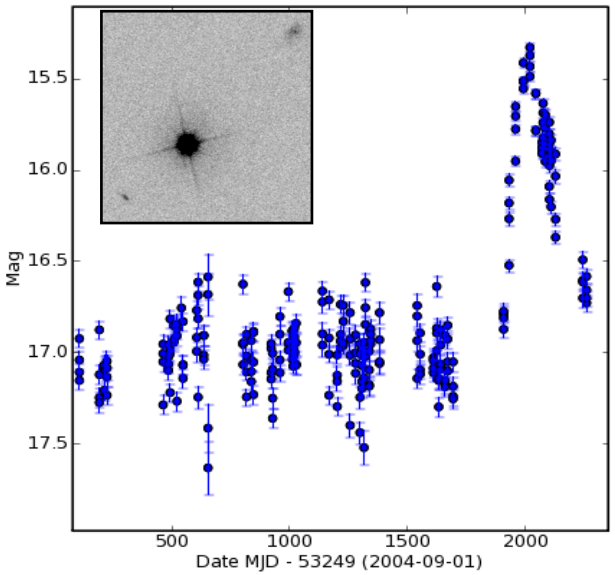
Host galaxy **> *500 times fainter*** than the Milky Way (M ≈ −13)

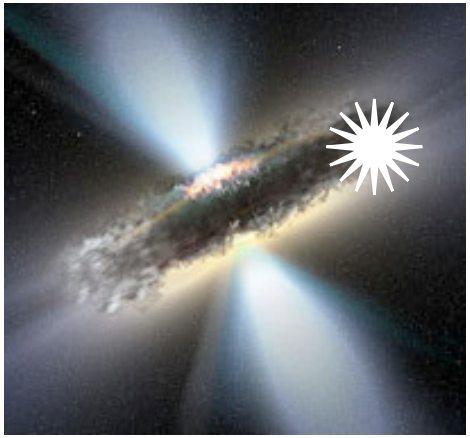


Possibly from an ~ η Carinae type progenitor: expanding SN interacts with the material from past outbursts



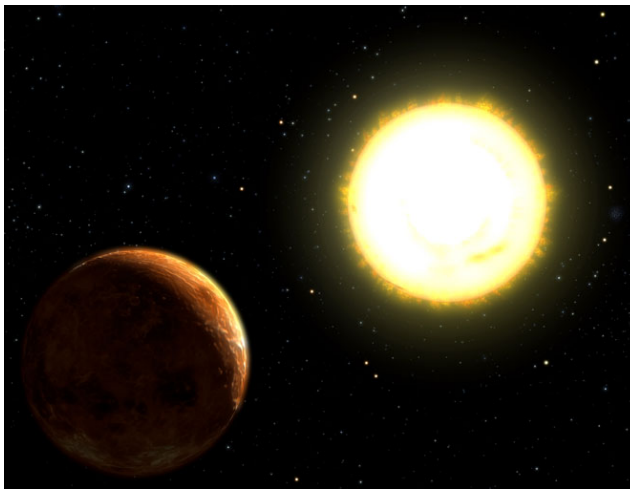
A New Kind of a Supernova?

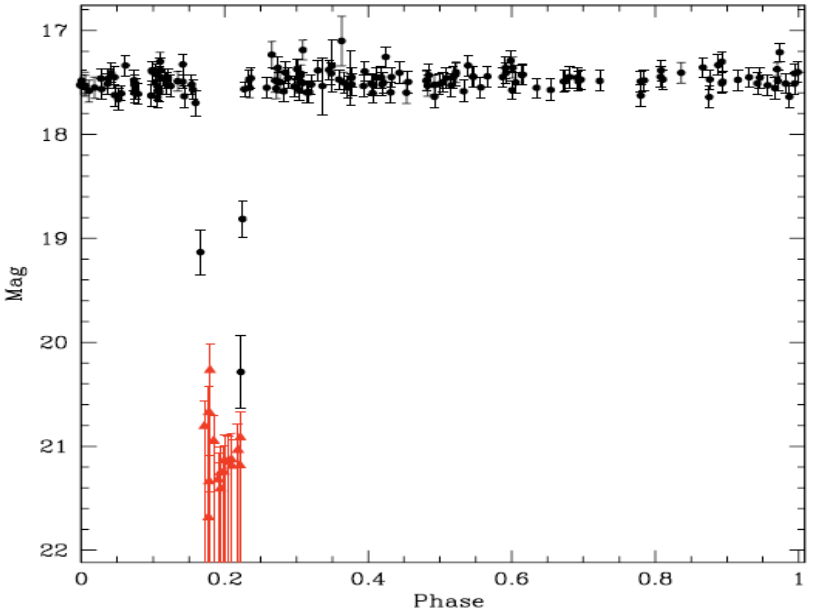




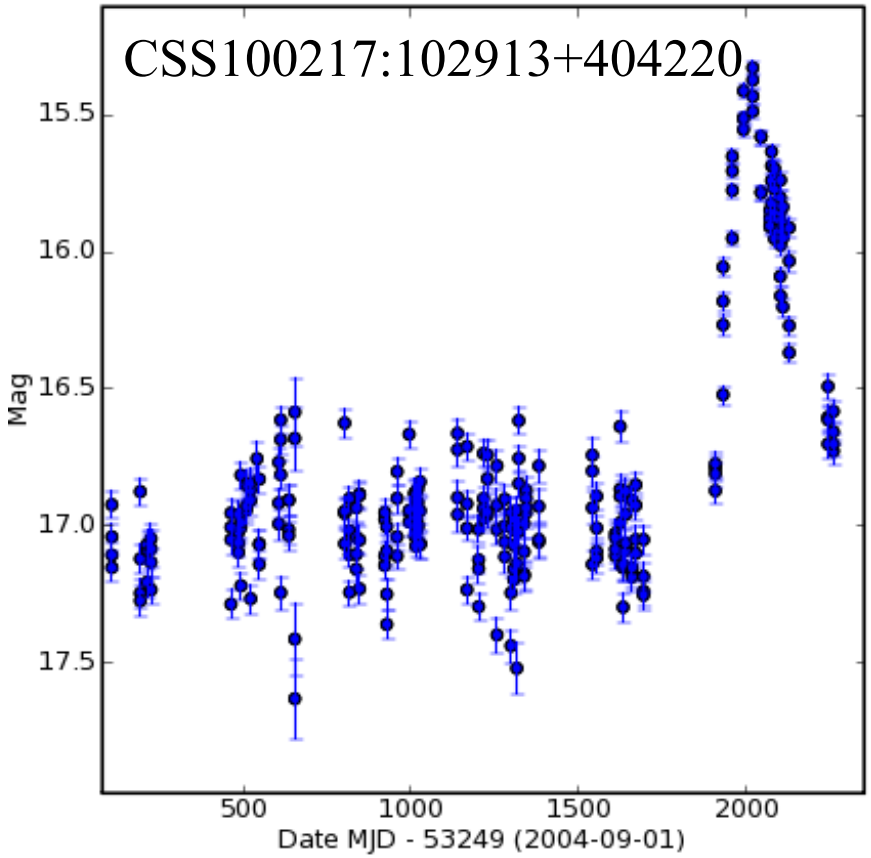
- The first case of a Supernova from an Active Galactic Nucleus accretion disk? (Predicted by theory, but never seen before)
- The most luminous SN ever seen

Planets Around White Dwarfs



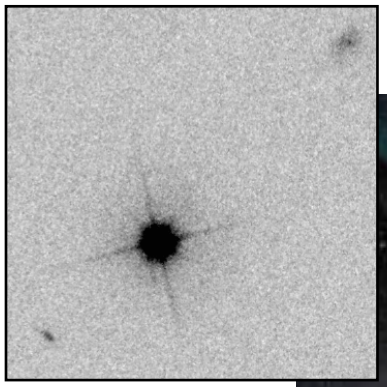


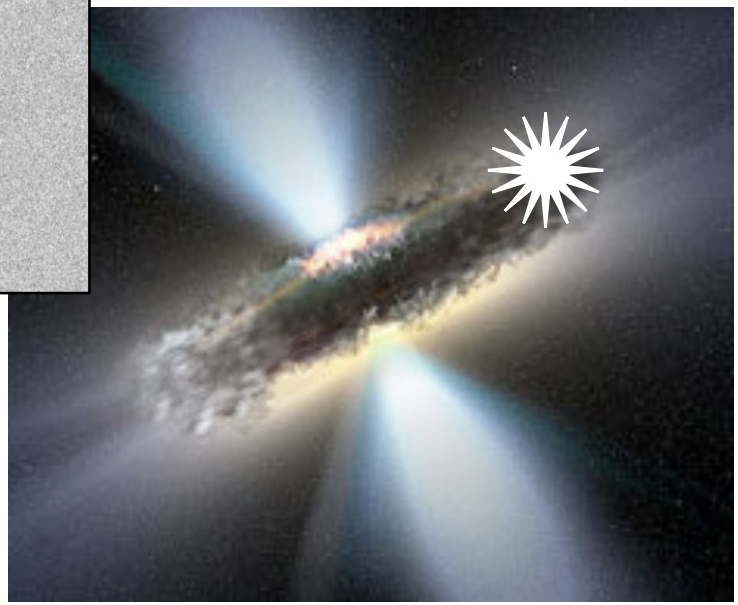
A New Kind of a Supernova?



- Transient in an active galaxy
- All data consistent with it being a Type II SN – but it would be the ***most luminous SN ever seen!***
- HST and Keck AO imaging shows that the event occurred within 150 pc of the active nucleus

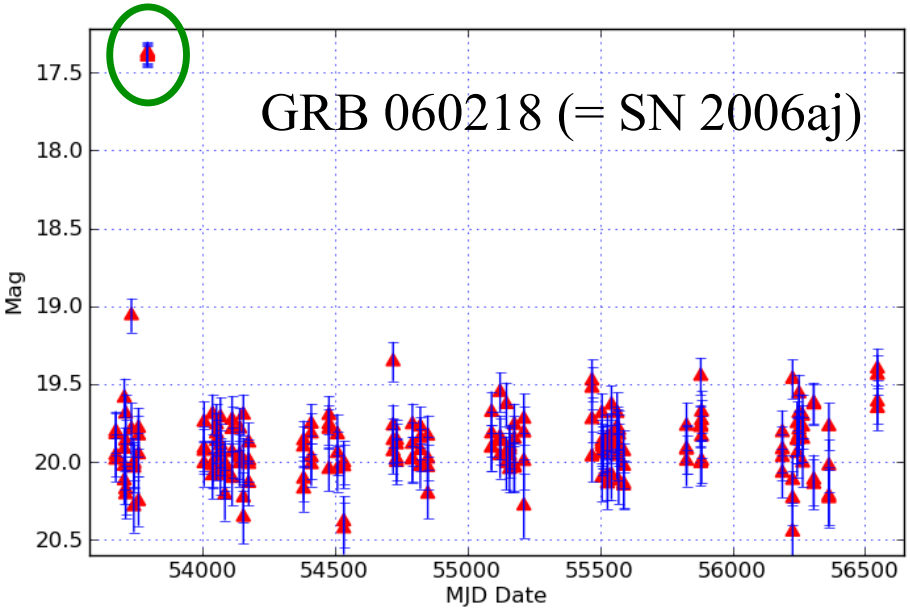
(Drake et al. 2011)





The first case of a Supernova from an Active Galactic Nucleus accretion disk? (Predicted by theory, but never seen before)

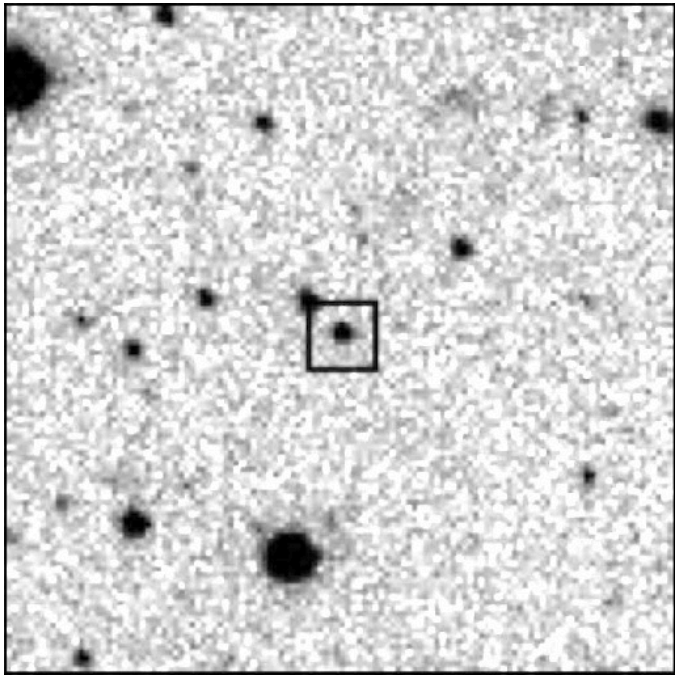
GRB Afterglows

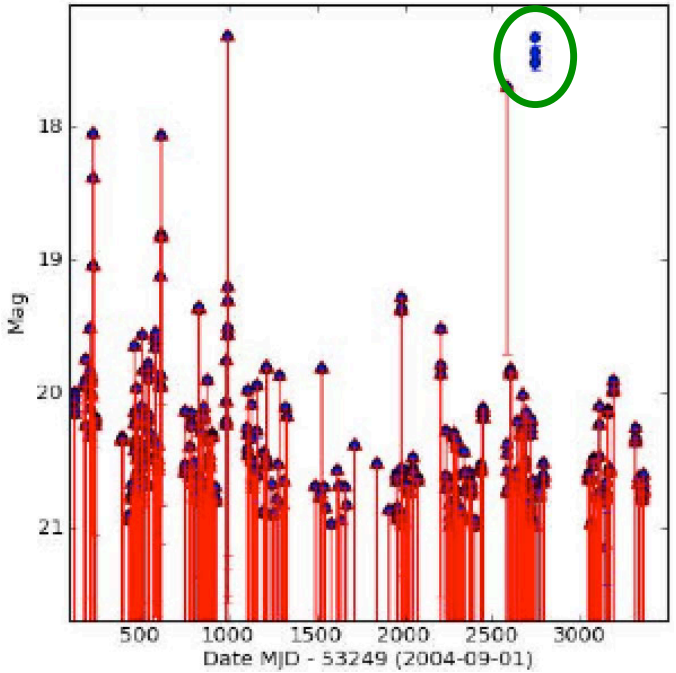


- A simple spatiotemporal search
- Every time we could have detected an afterglow, we did
- Fermi GBM bursts: a capability demo for the searches for EM counterparts of LIGO events

Fermi GBM GRB120302A:

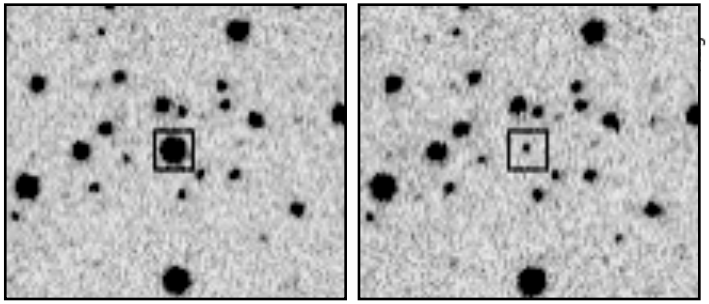
Detected 90 min after the trigger; Swift detection 1 day later

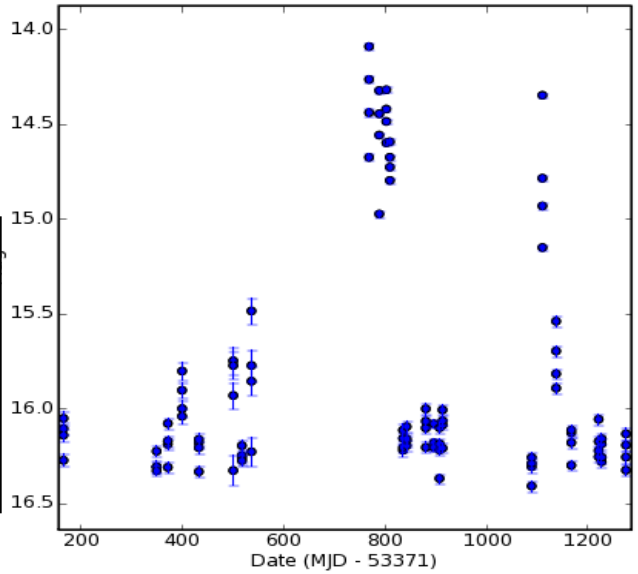


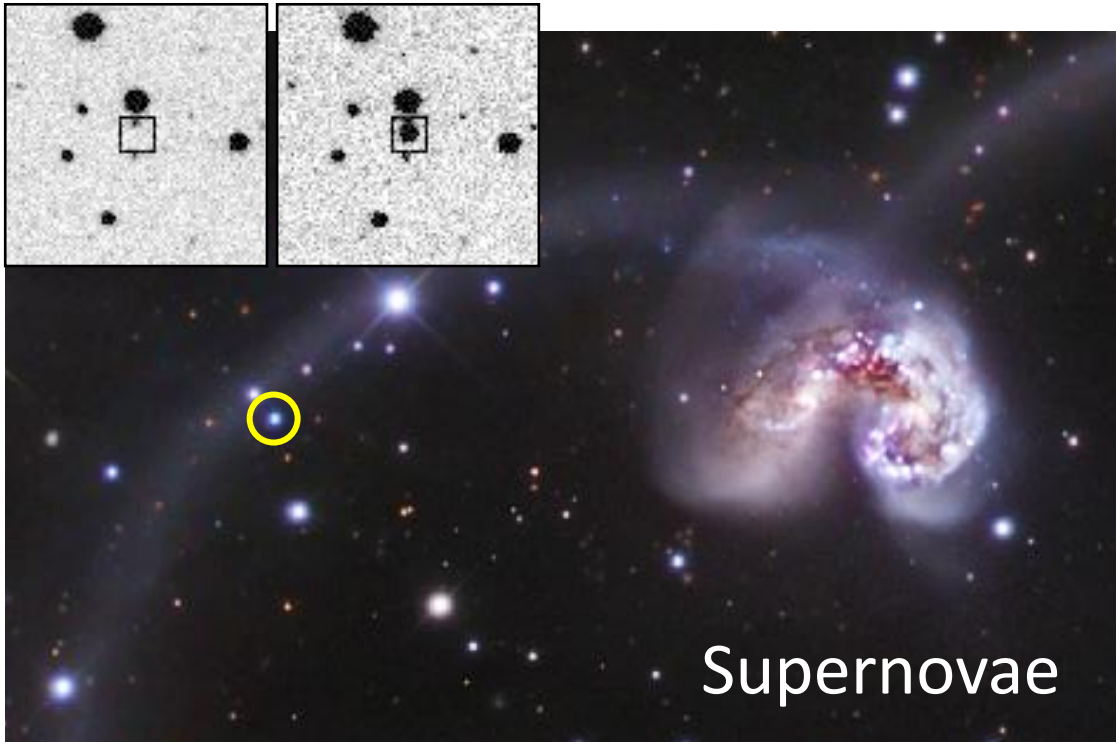


Examples of CRTS Discoveries

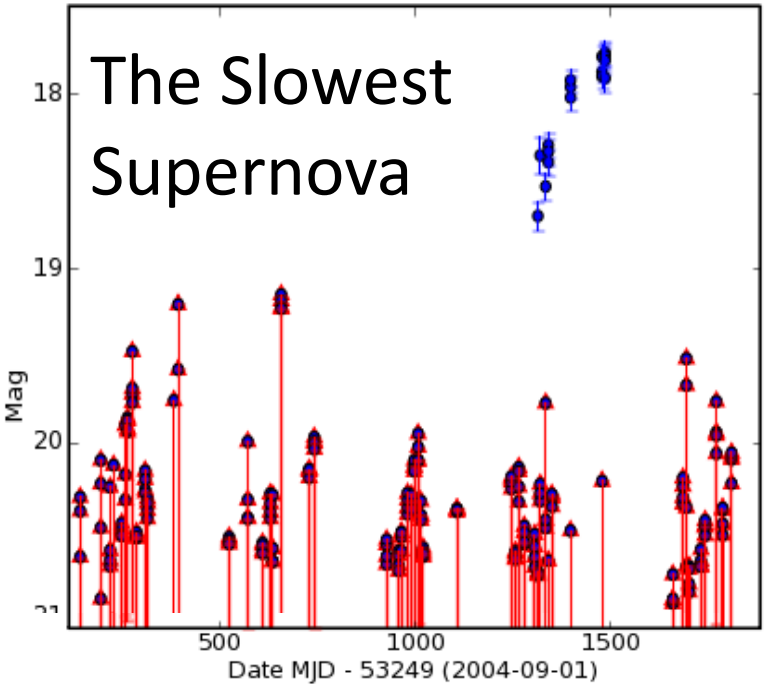
Cataclysmic Variables
and Dwarf Novae



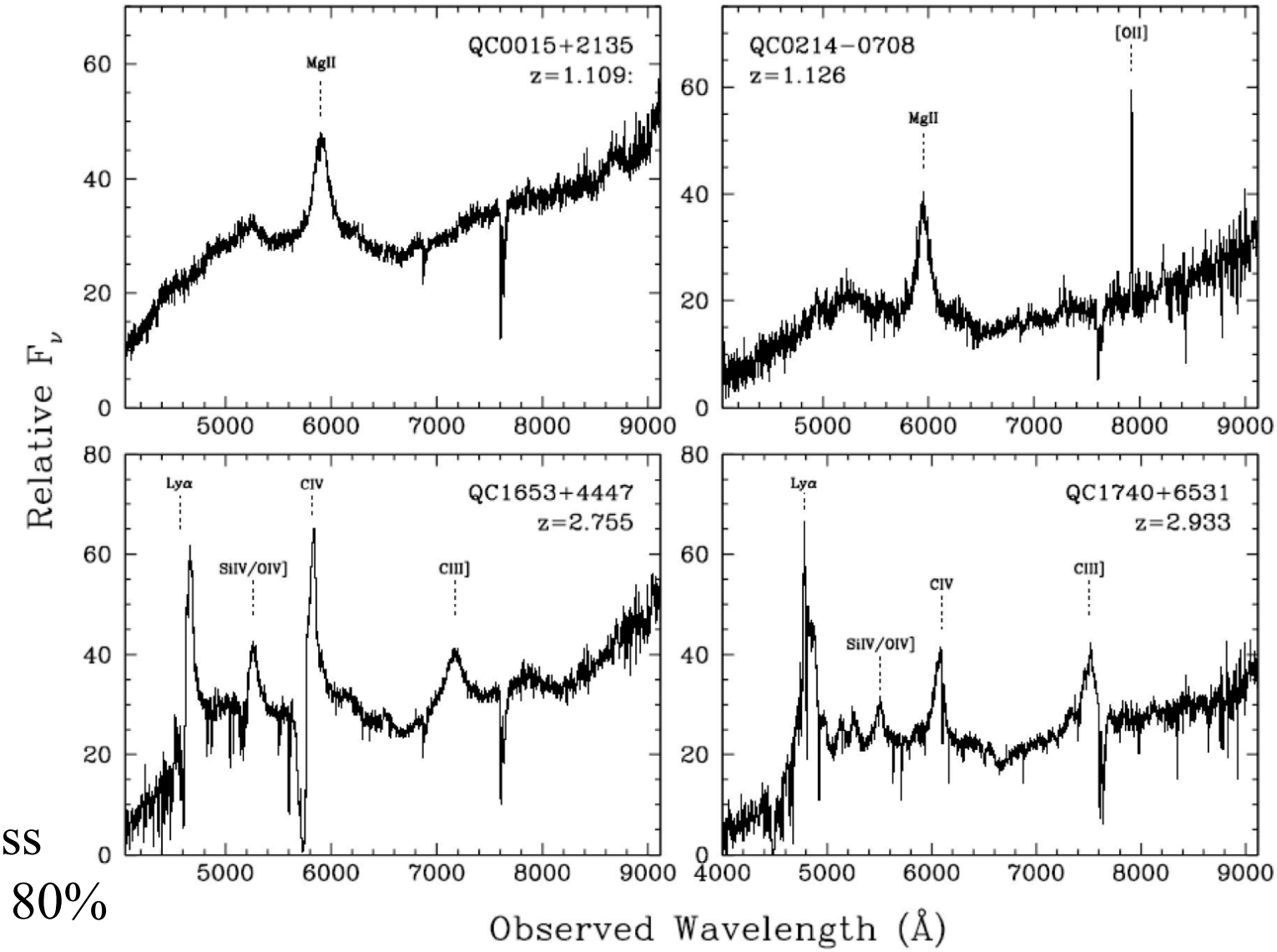




The Slowest
Supernova

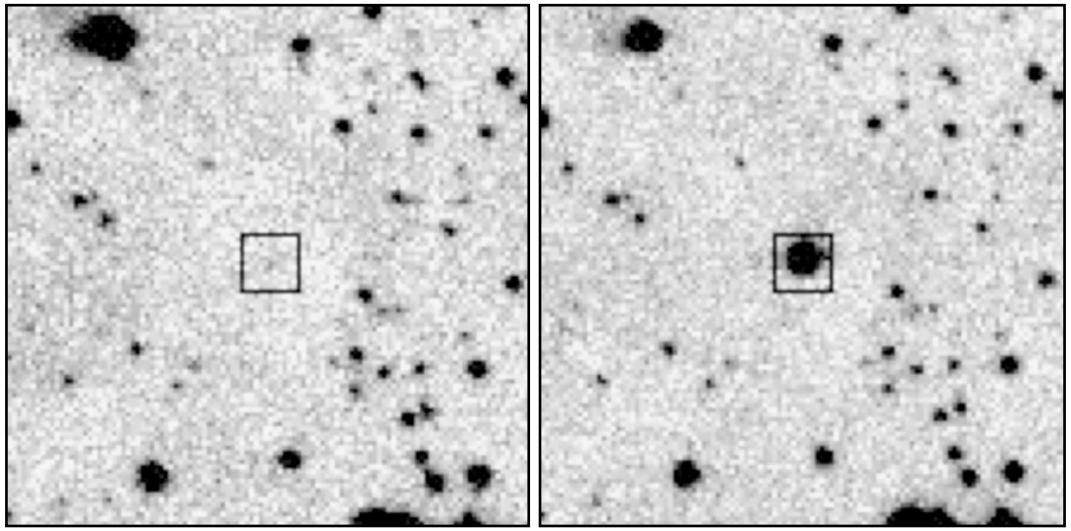


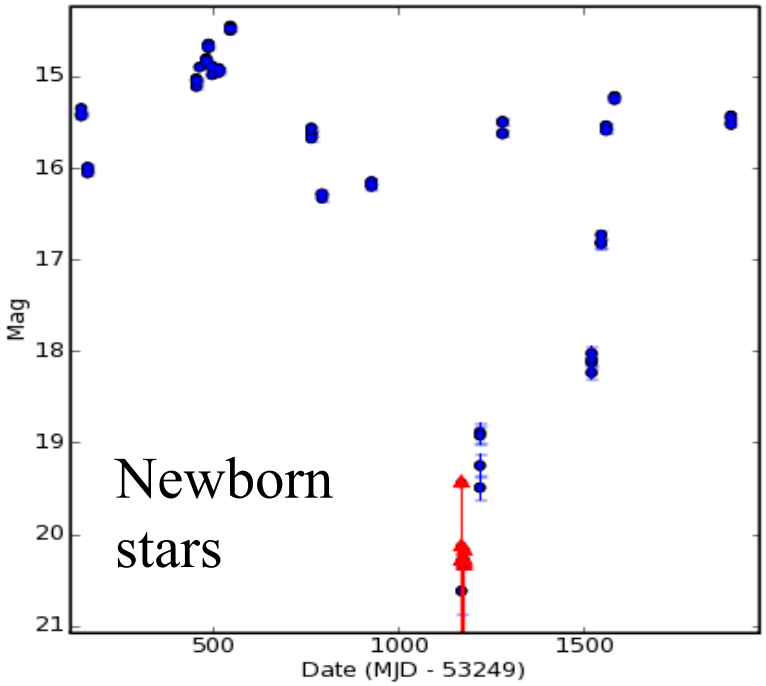
Spectroscopic Confirmation of Variability-Selected QSO Candidates

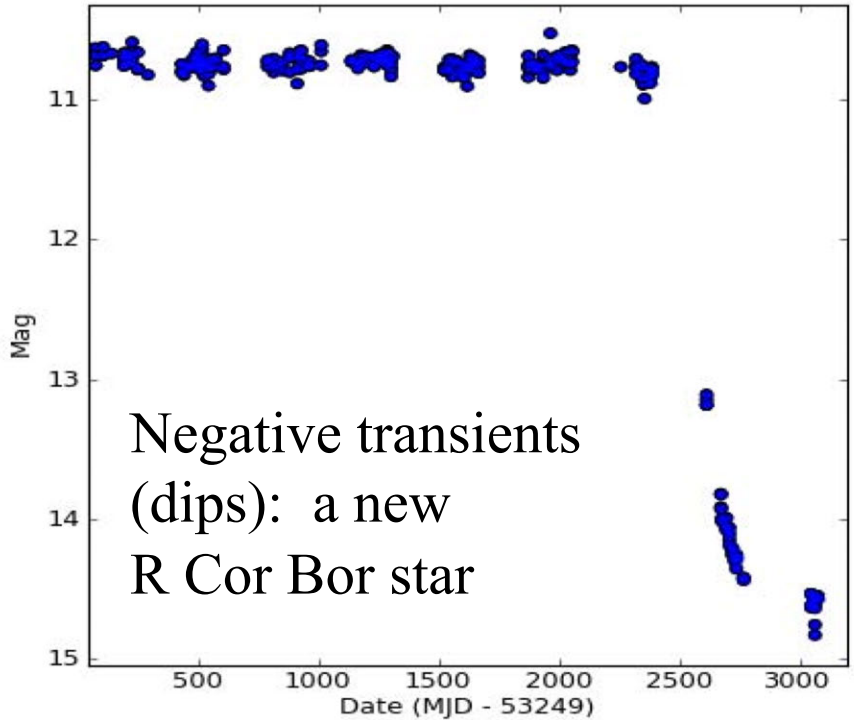


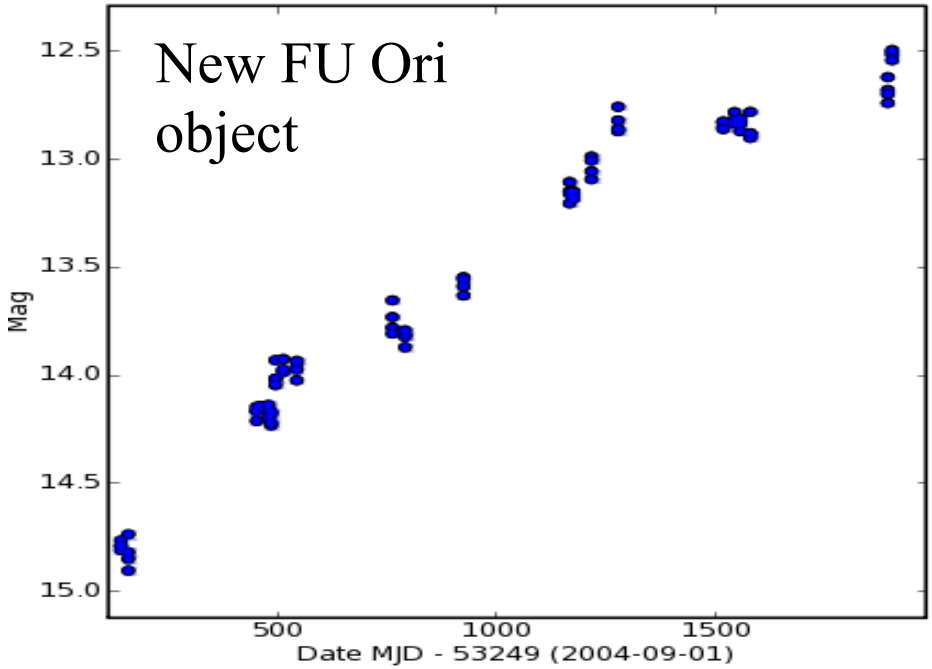
Success rate > 80%

Unsettled Stars



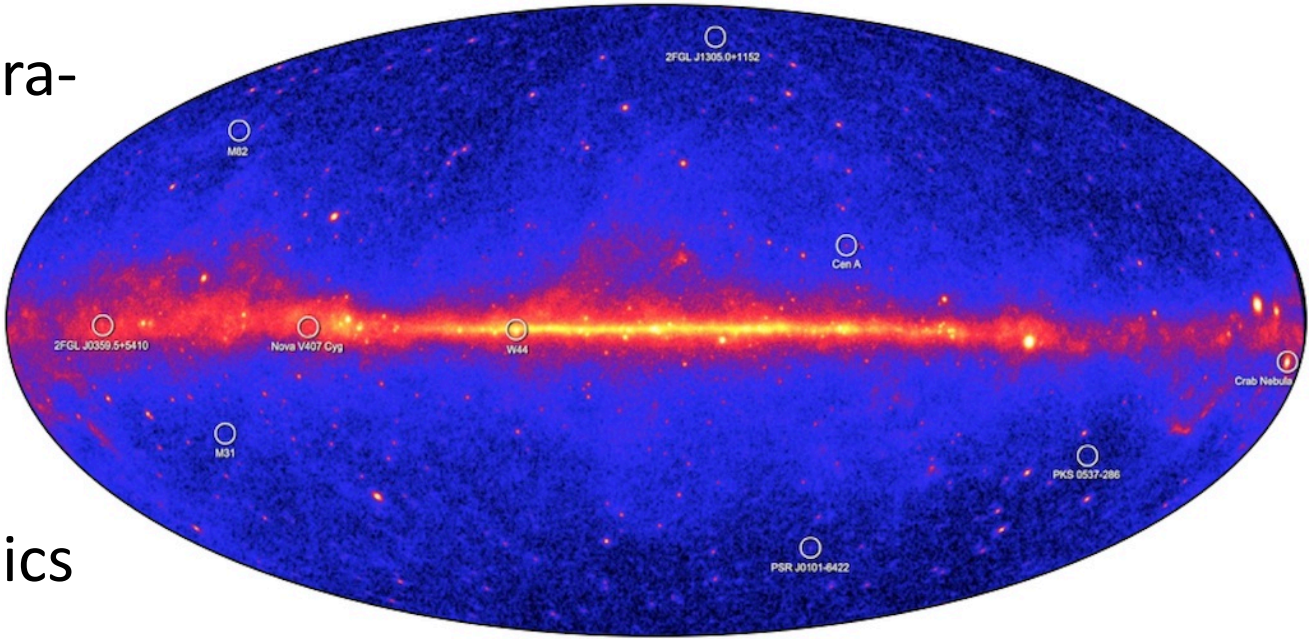




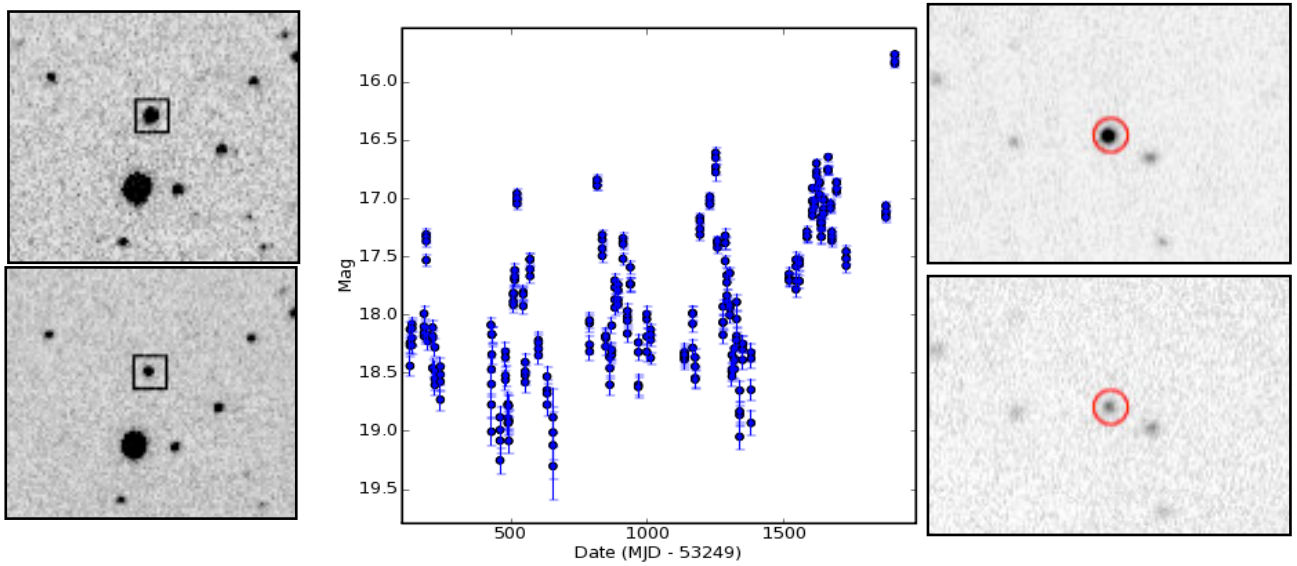


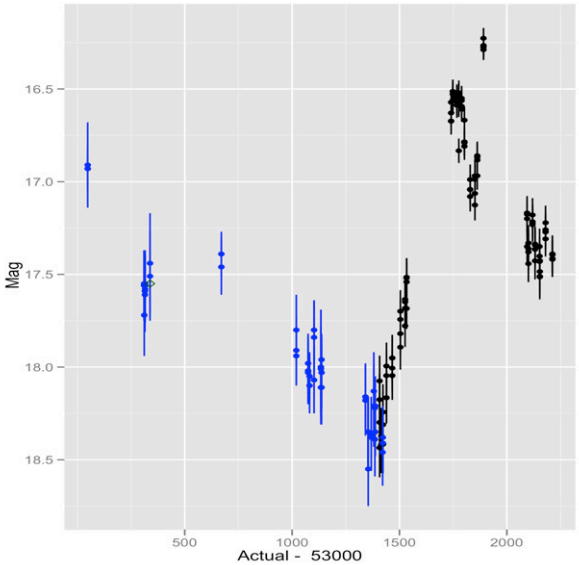
Blazars: the Cosmic Accelerators

- The principal extra-galactic γ-ray sources
- Probes of relativistic jet physics
- AGN demographics and evolution

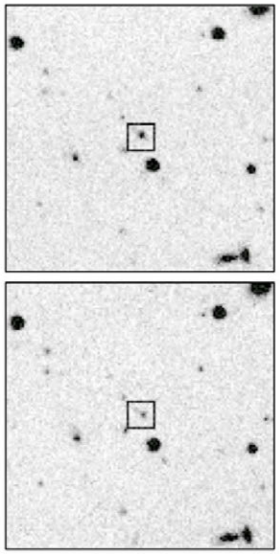


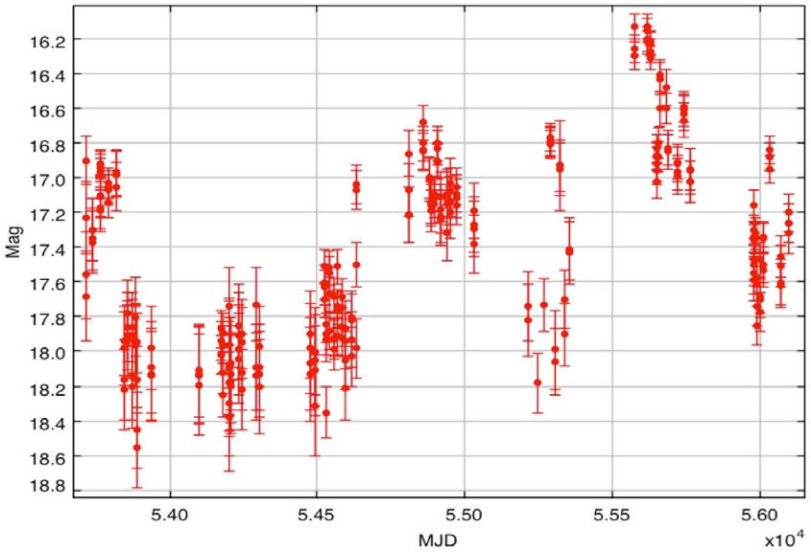
Variability-based IDs of Fermi sources





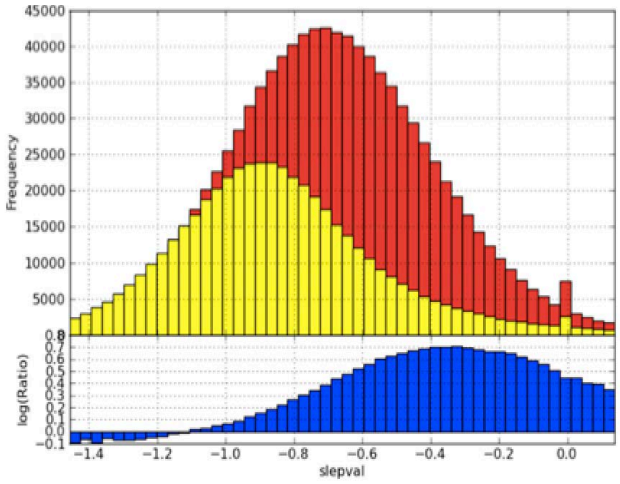
Variability-Based Selection of Quasars

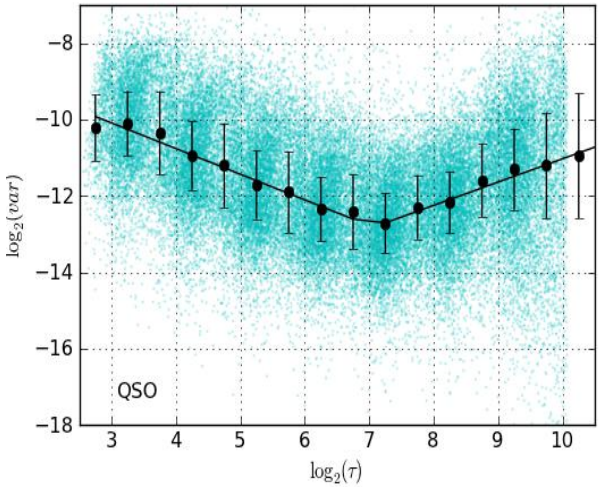


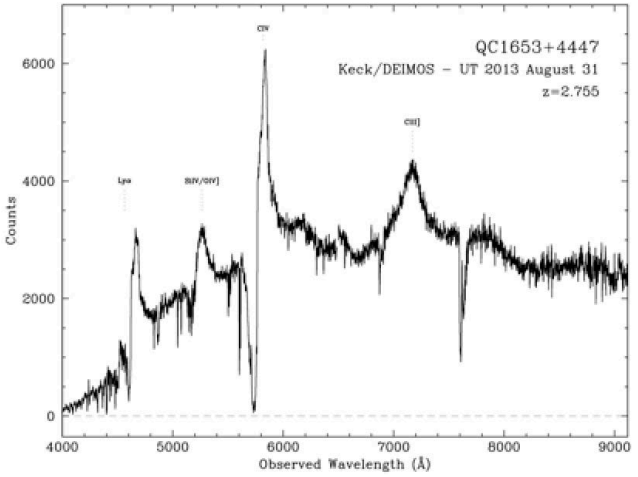


Data mining the light curve variability parameter space to select quasar candidates

(Lead: M. Graham)

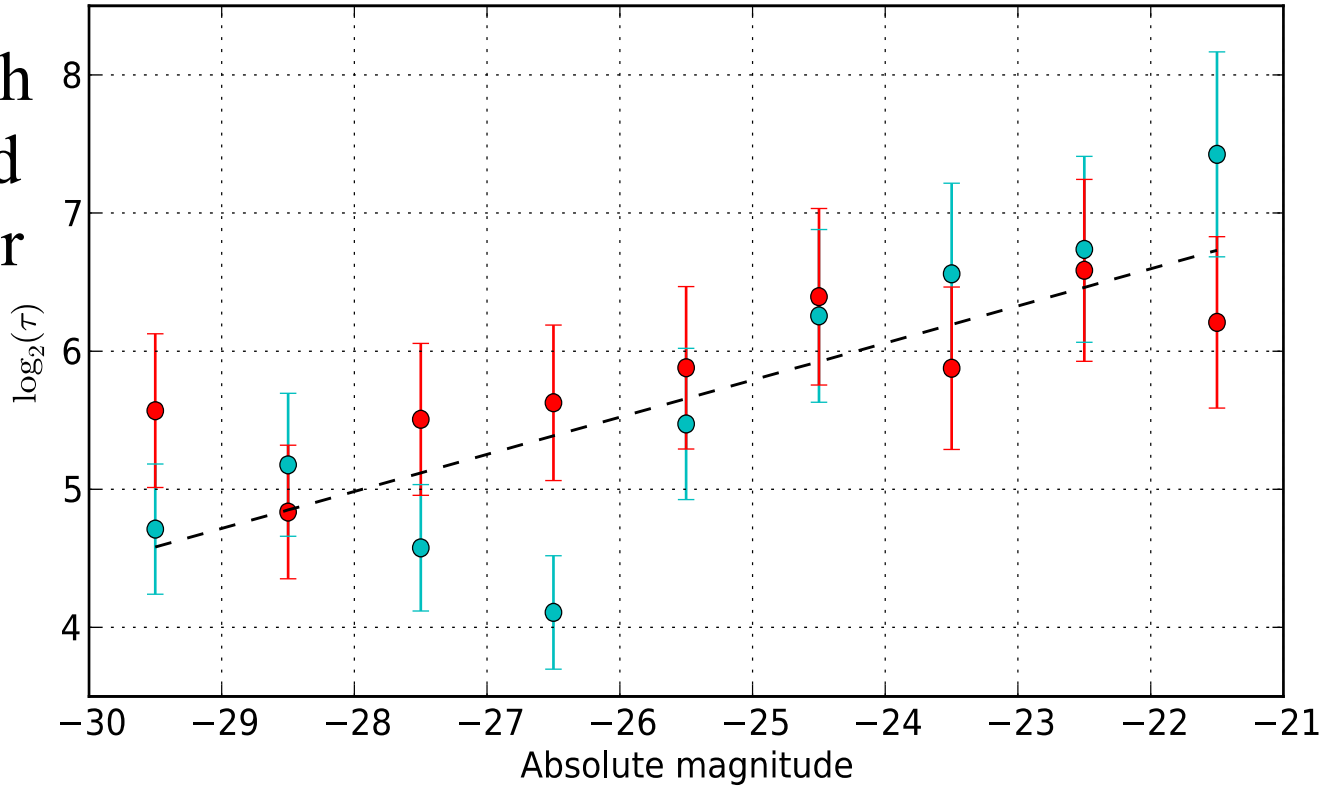






Evidence for a Characteristic Time Scale

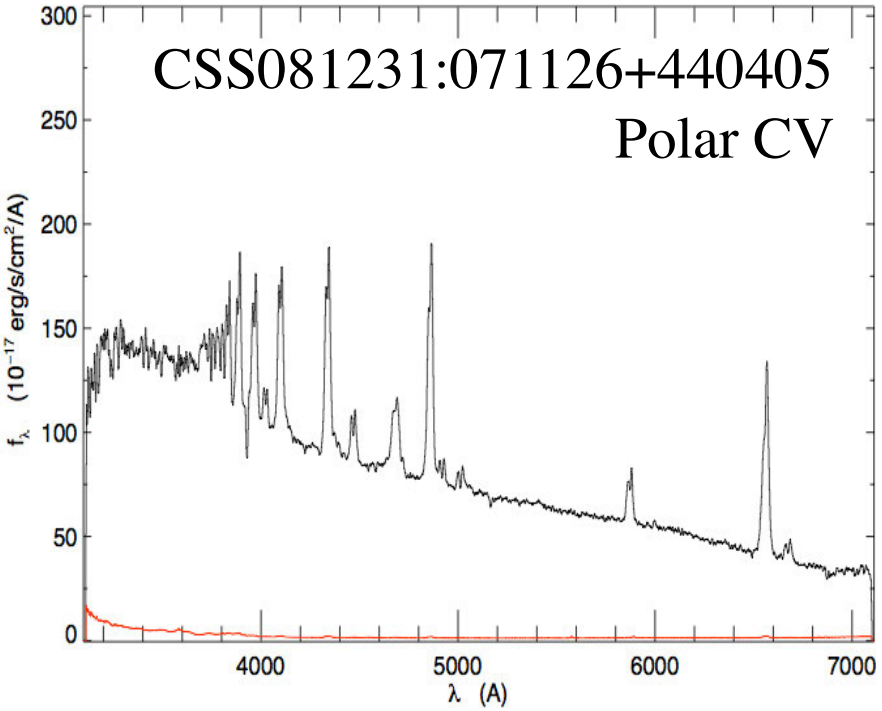
- First solid evidence for a characteristic time scale (~ 50 days) associated with the quasar variability in the visible
  - Previous indications in the X-ray
- Possible probe of the accretion disk physics
  - Diffusion time scale in the outer regions of the accretion flow?
- Anticorrelated with the luminosity, and possibly with other parameters (work in progress)

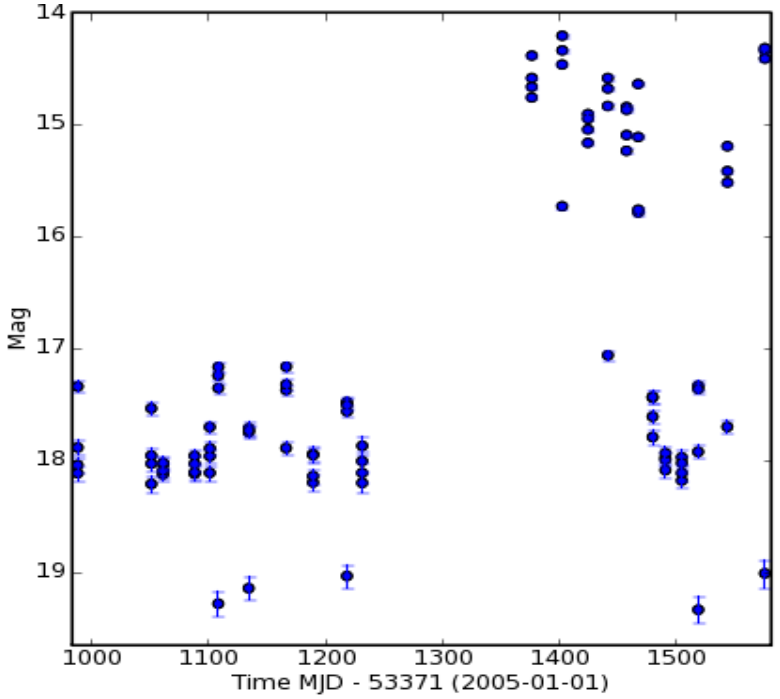


(Graham et al. 2013)

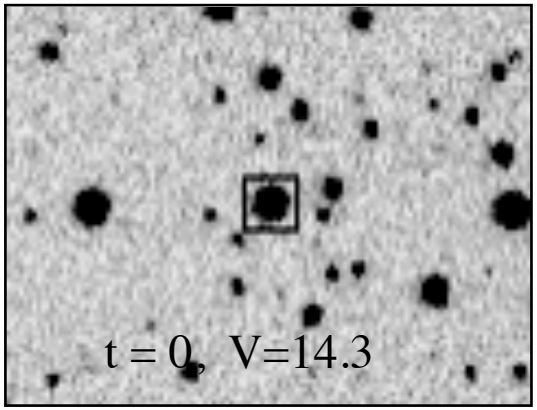
Cataclysmic Variables and Dwarf Novae

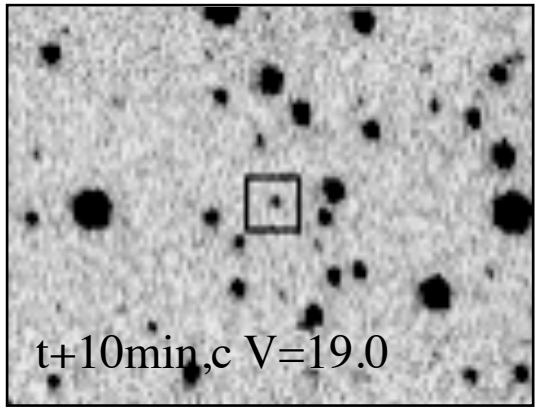
> 1,000 detected so far, > 75% are new discoveries

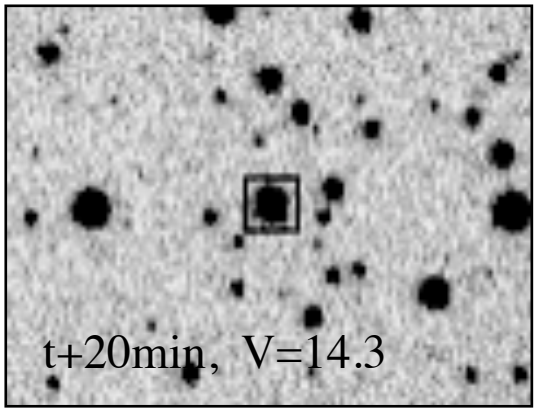




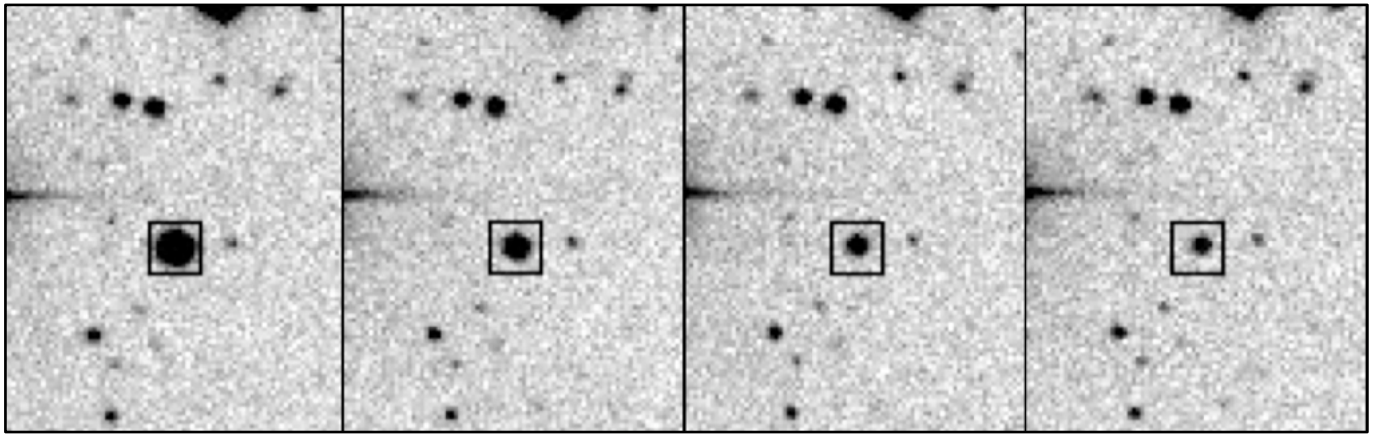
Eclipsing Polar CSS081231:071126+440405



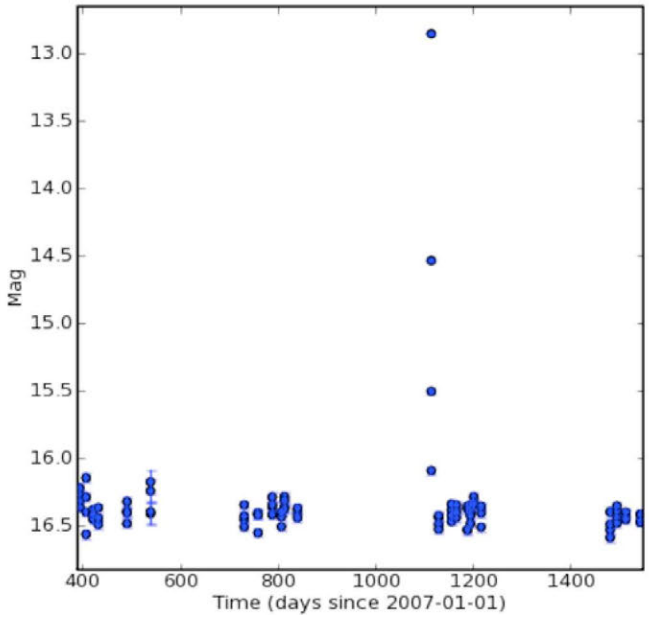


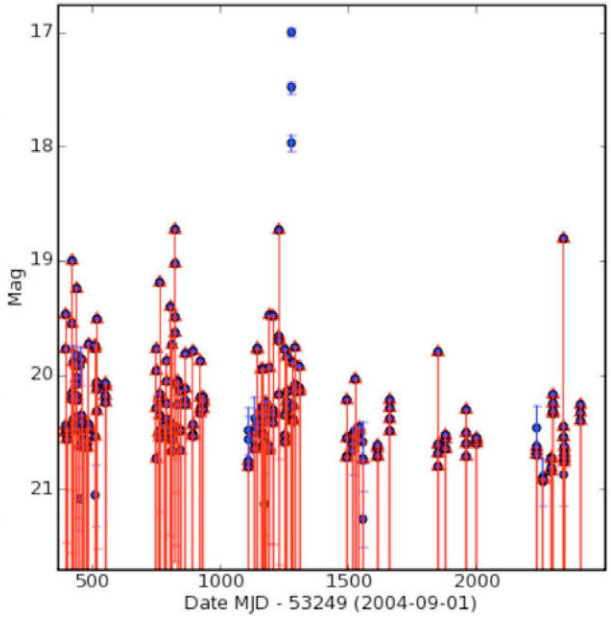


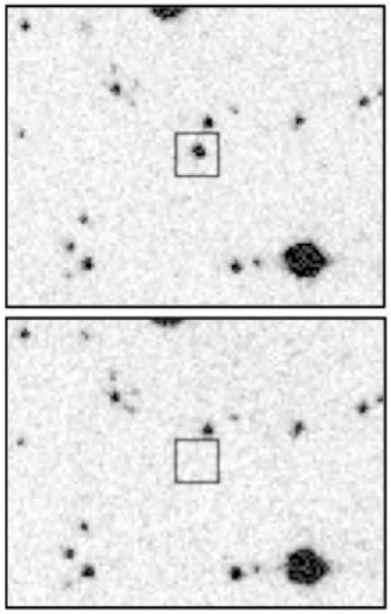
Fast Transients



4 individual exposures, separated by 10 **min**

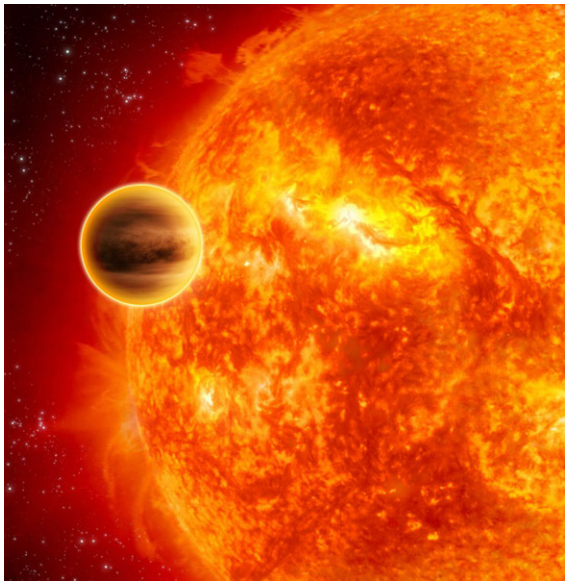






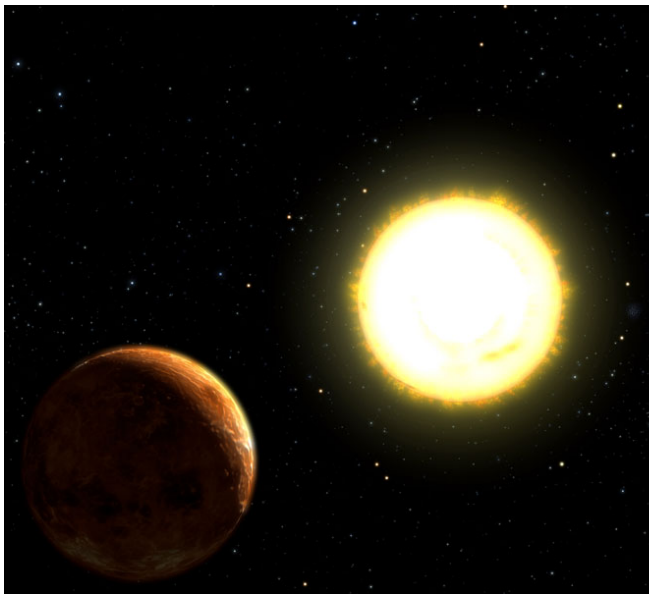
Most (but not all!) are flaring dwarf stars (UV Ceti)

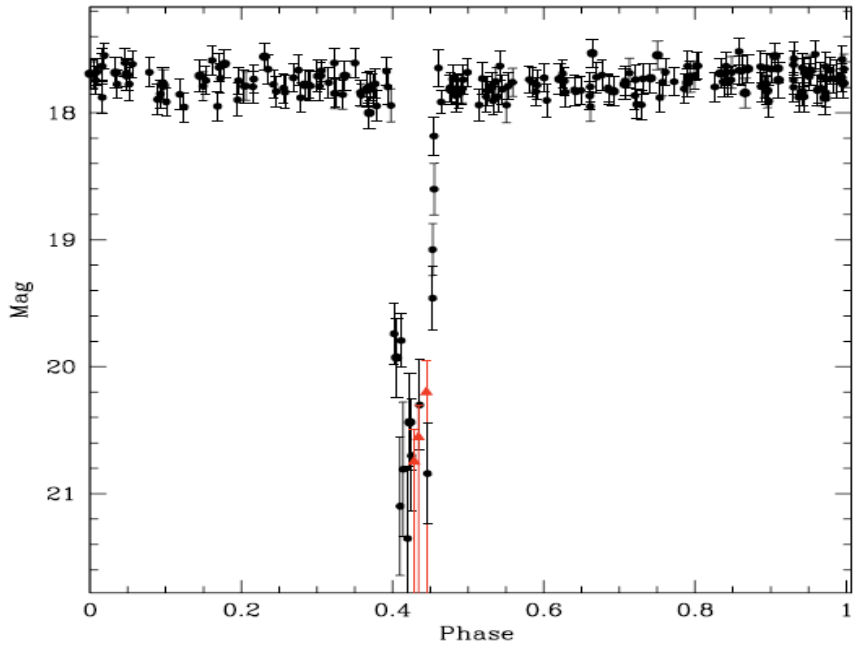
Eclipsing White Dwarfs: Planets?

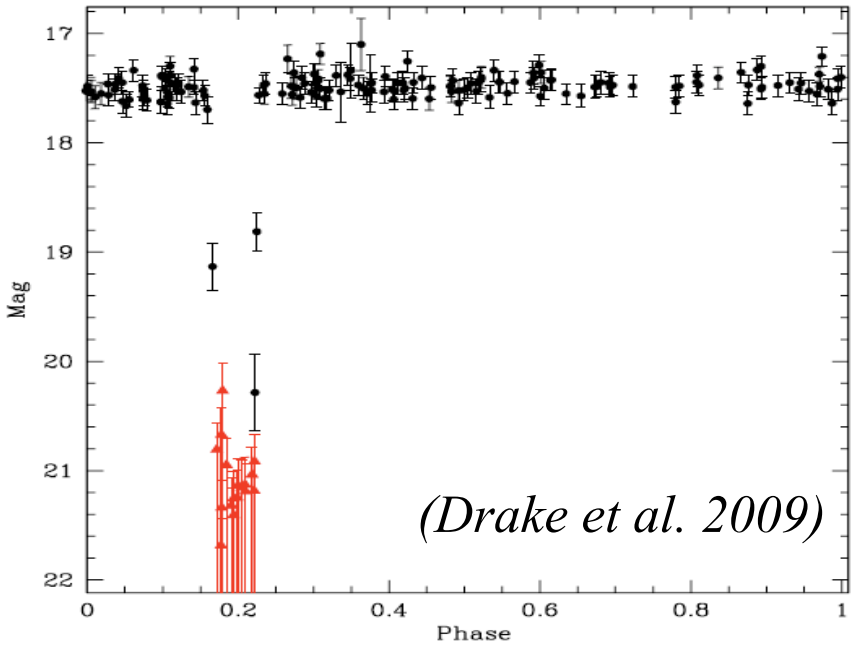


Earth-like planets cause ~ 10^{-4} eclipses for the main-sequence stars…

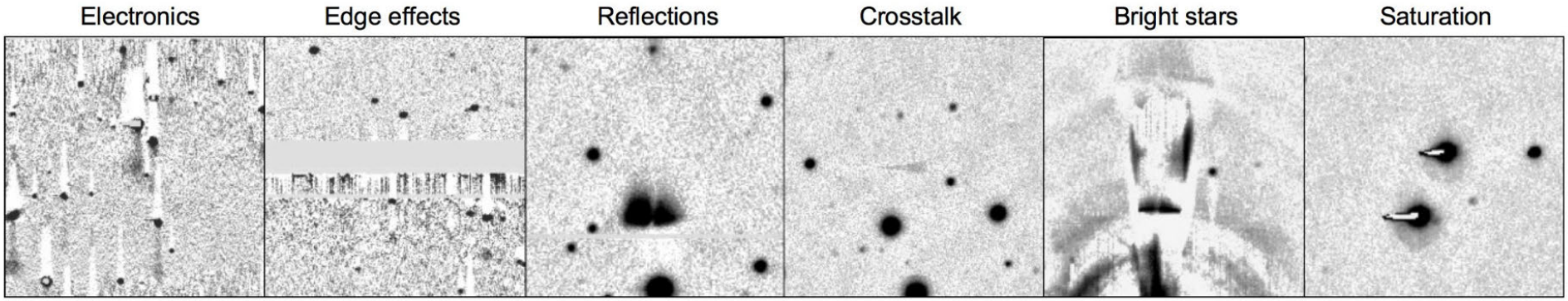
But it could be ~100% eclipses for the white dwarfs!

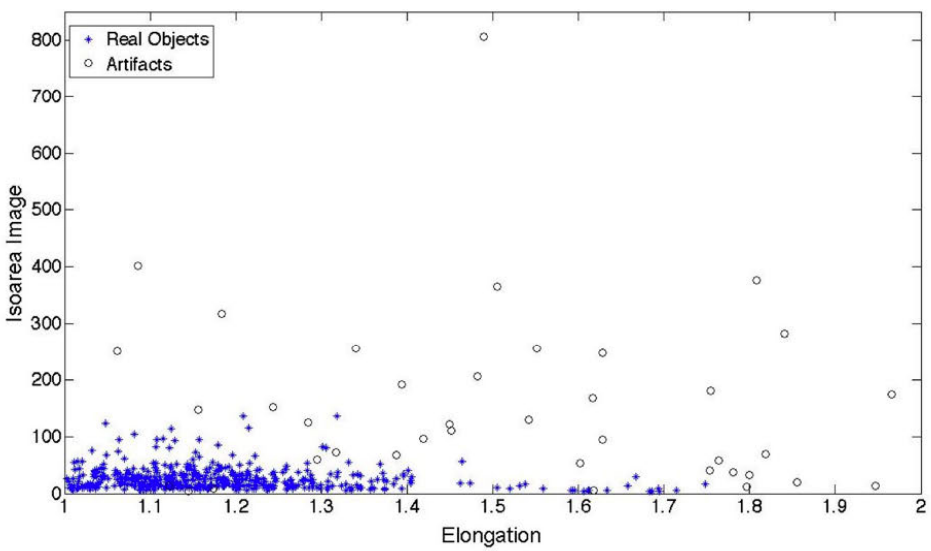


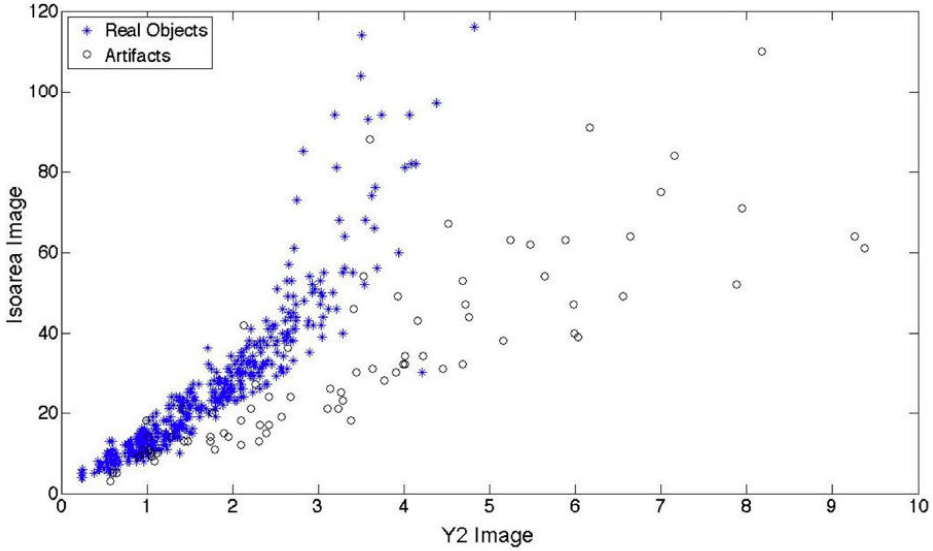




Automated Detection of Artifacts



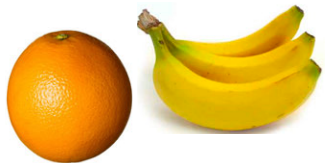




Automated classification and rejection of ~ 95% of artifacts masquerading as transient events in the PQ survey pipeline, using a Multi-Layer Perceptron ANN

(Lead: C. Donalek)

Event Classification is a Hard Problem

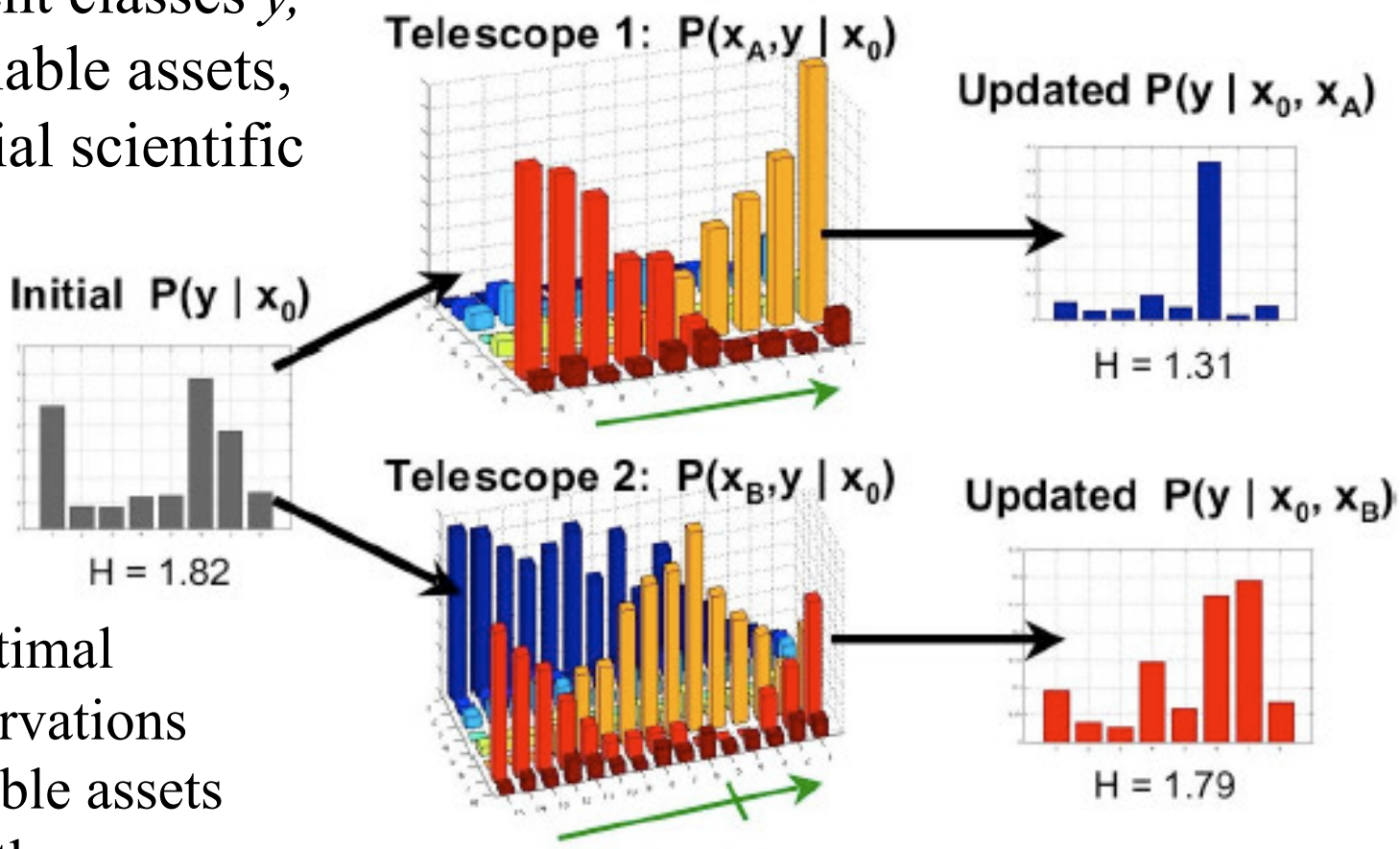


- Physical classification of transient events is essential for their astrophysical interpretation and uses
  - Must be done in real time and iterated dynamically
- Human classification is already unsustainable, and will not scale to the Petascale data streams
- This is ***hard:***
  - Data are sparse and heterogeneous: feature vector approaches are limited; Bayesian approaches work
  - Completeness vs. contamination ☯
  - Follow-up resources are expensive and/or limited: follow only the most interesting objects

    ⇒ Need automated decisions for follow-up
  - Traditional DP pipelines do not capture a lot of the relevant contextual information, prior/expert knowledge, etc.

Automating the Optimal Follow-Up

For the *potentially most interesting events*, what type of follow-up observations a *x* has the greatest potential to discriminate among the competing event classes *y*, given the available assets, and the potential scientific value?

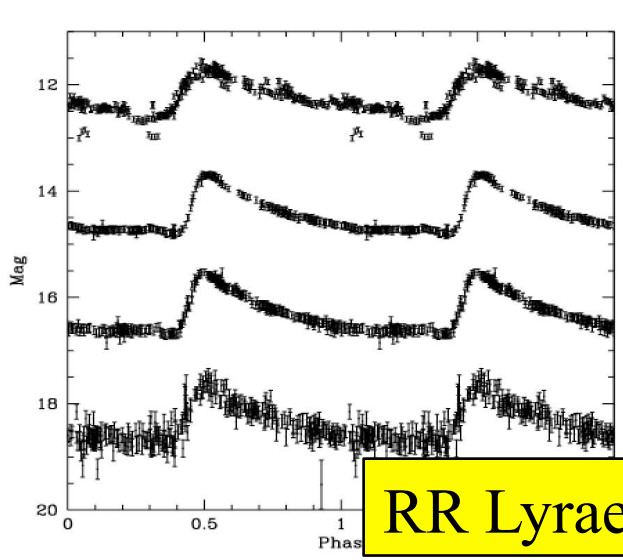


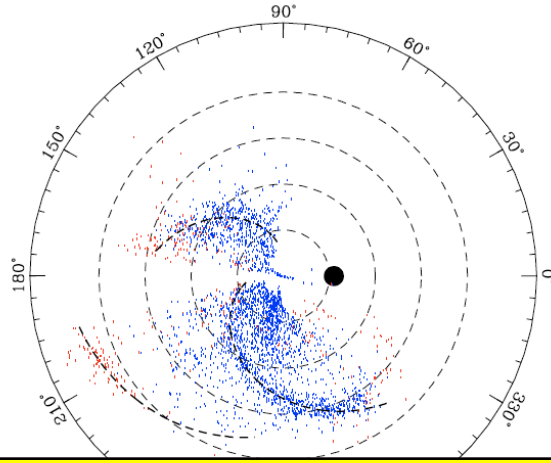
Request the optimal follow-up observations from the available assets that maximize the entropy drop:

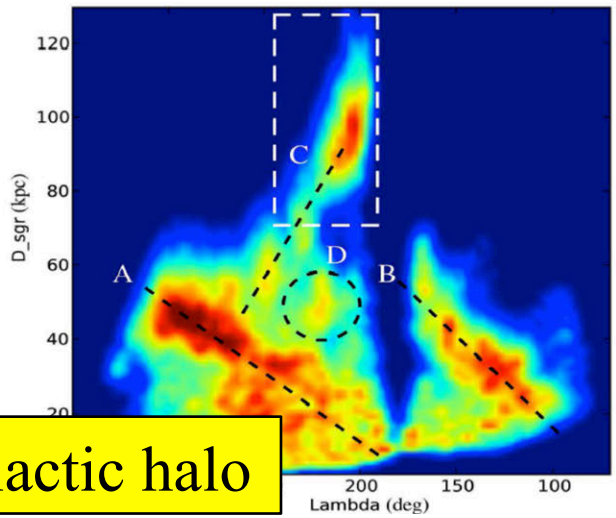
$$H[p(y \mid x_+, x_0)] = -\sum_{y,x_+} p(y, x_+ \mid x_0) \log p(y \mid x_+, x_0)$$

Real-Time *vs.* Non-Time-Critical

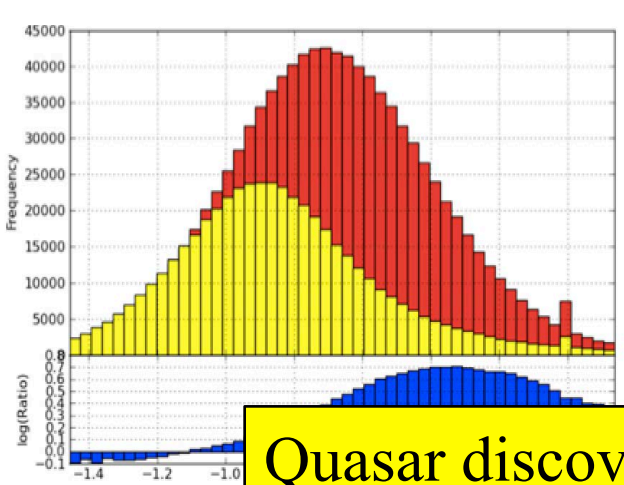
- Transients may be overemphasized; there is a lot of good science in the archival studies, and that can only get better in time

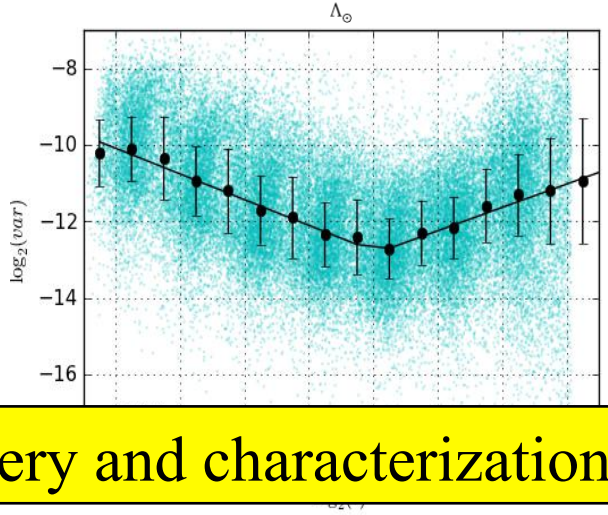


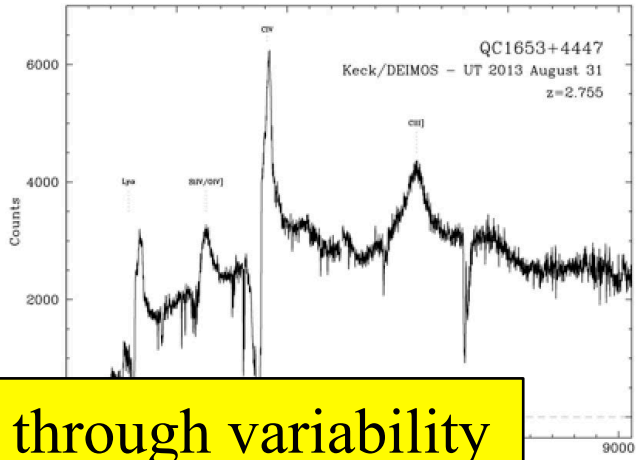




RR Lyrae and mapping of the Galactic halo



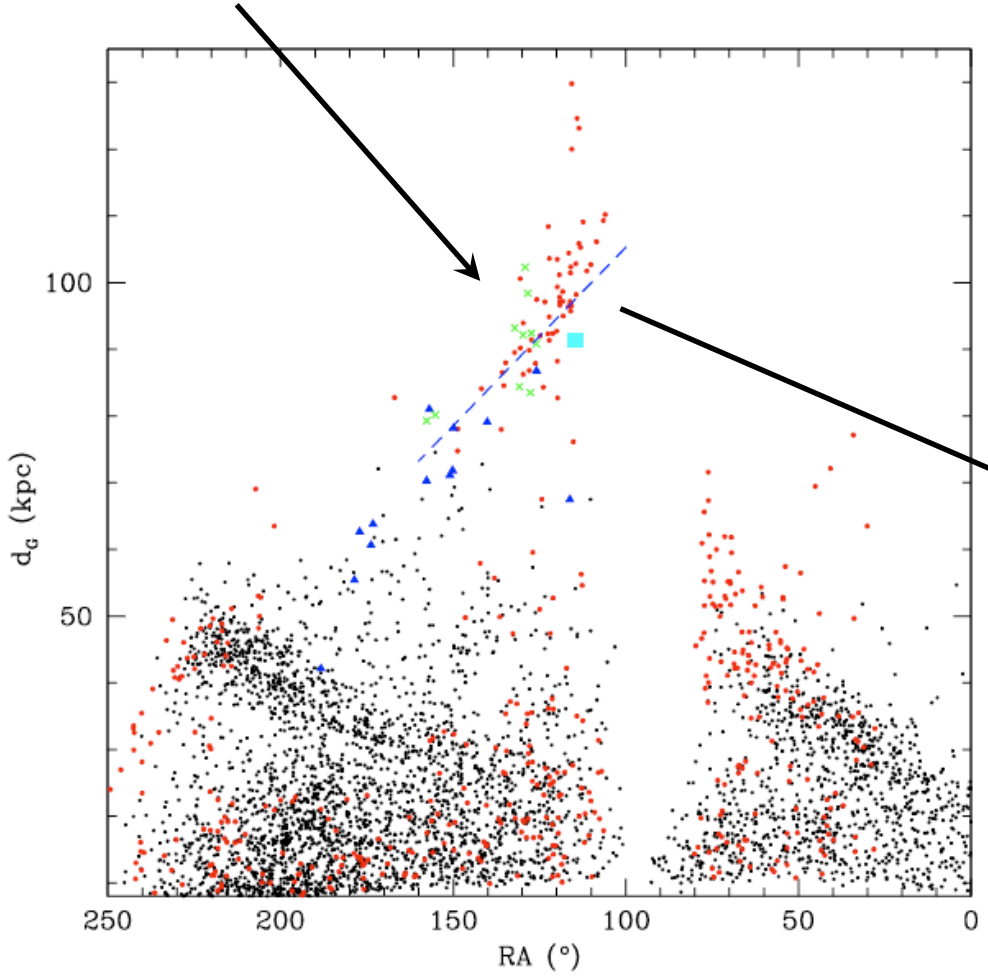


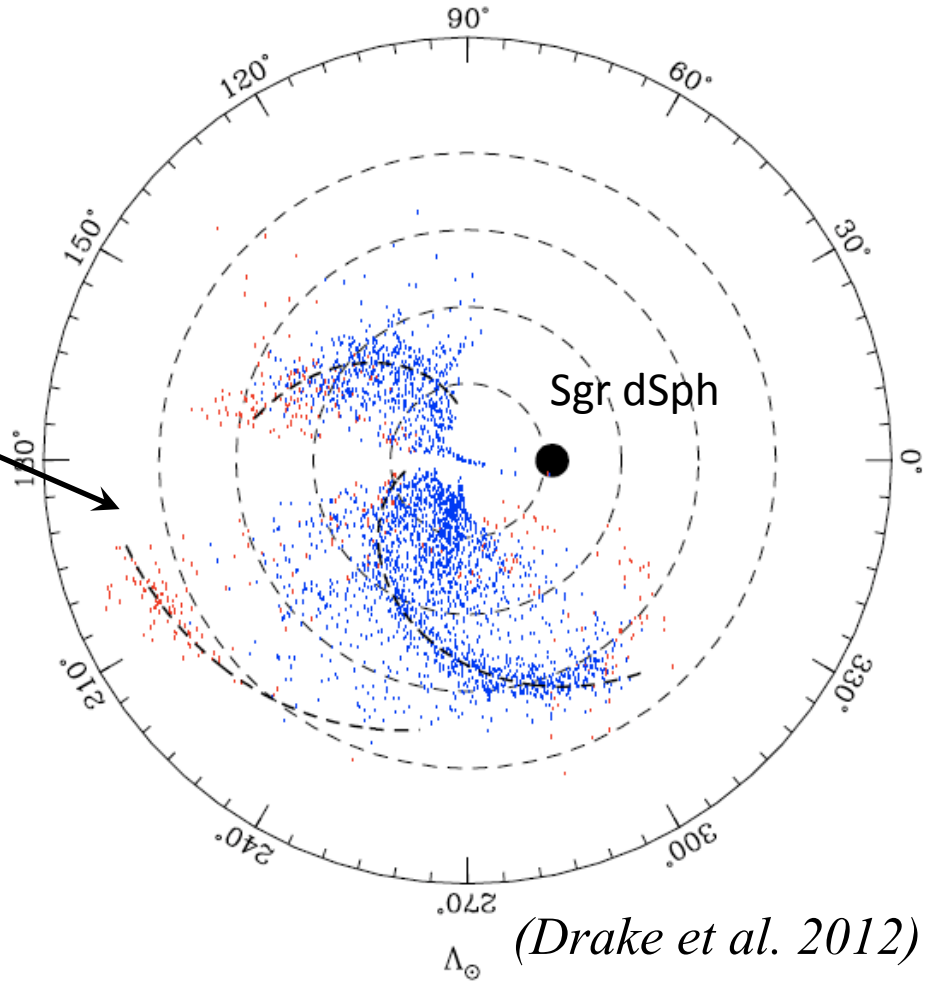


Quasar discovery and characterization through variability

Tidal Streams in the Galactic Halo

Using > 12,000 RR Lyrae selected from the CRTS light curves archive to map the Galactic halo: **discovery of a new tidal stream reaching out to 100 kpc** (Gemini stream)

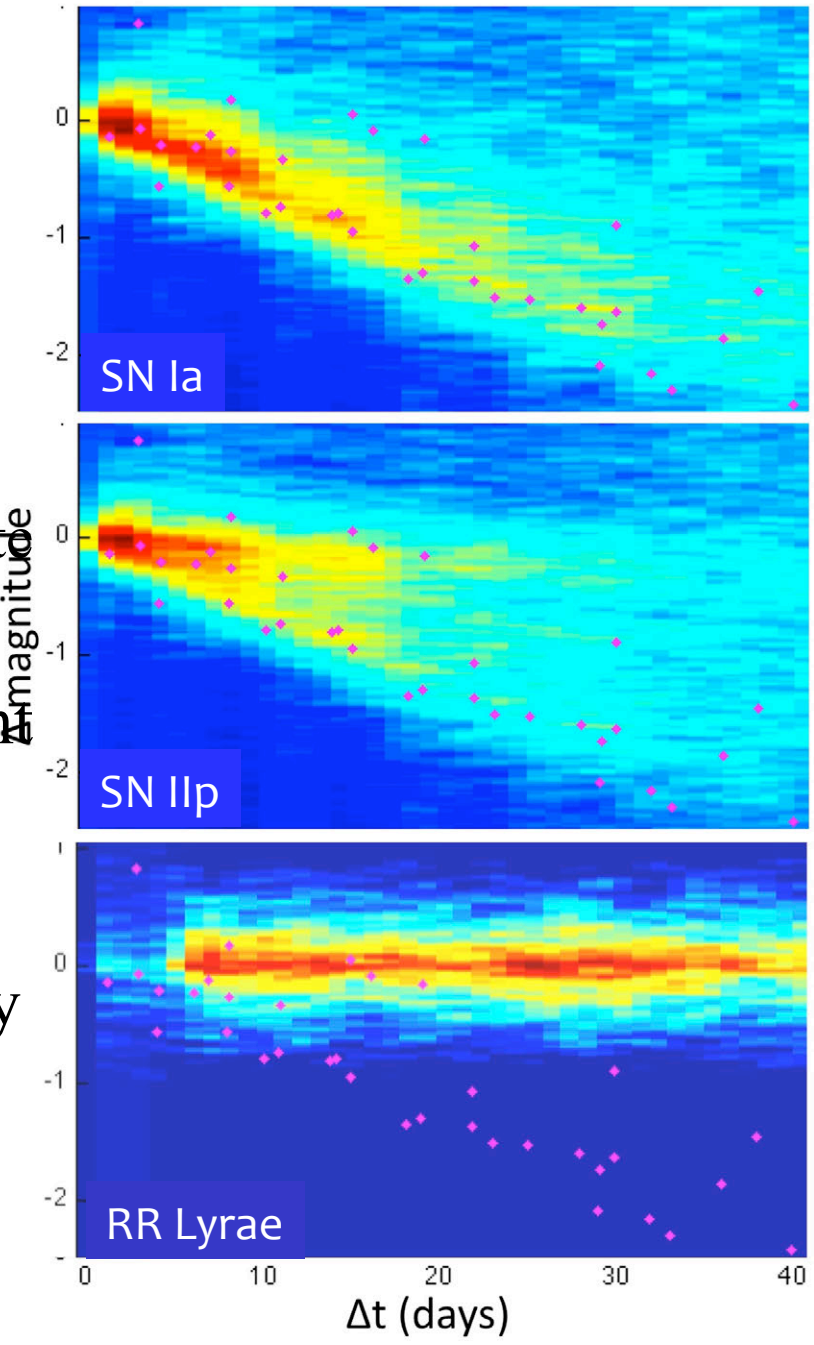




(Drake et al. 2012)

2D Light Curve Priors

- For any pair of light curve measurements, compute the Δt and Δm, make a 2D histogram
  - N independent measurements generate N^2 correlated data points
- Compare with the priors for different types of transients
- Repeat as more measurements are obtained, for an evolving, constantly improving classification
- Now expanding to consecutive data point triplets: Δt_{12}, Δm_{12}, Δt_{23}, Δm_{23}, giving a 4D histogram



(Lead: B. Moghaddam)