Automated Event Classification in Synoptic Sky Surveys

S. G. Djorgovski *(Caltech)*

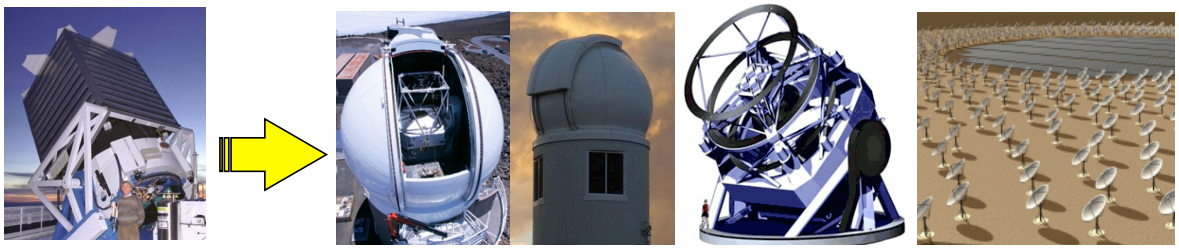
With: A. Mahabal, C. Donalek, M. Graham, A. Drake, B. Moghaddam, M. Turmon, and many students and collaborators

IVOA InterOp, Naples, May 2011

The Evolving Data-Rich Astronomy

- Digital sky surveys have brought us into the Terascale regime, and stimulated:
  - Extensive use of databases in astronomy
  - The Virtual Observatory concept
  - Incipient data-mining-based astronomy

- Synoptic digital sky surveys – i.e., panoramic cosmic cinematography – are moving us into the Petascale regime
  - The same old challenges, only more so
  - New challenges: real time response, event classification, data mining in the time domain…



Astronomy in the Time Domain

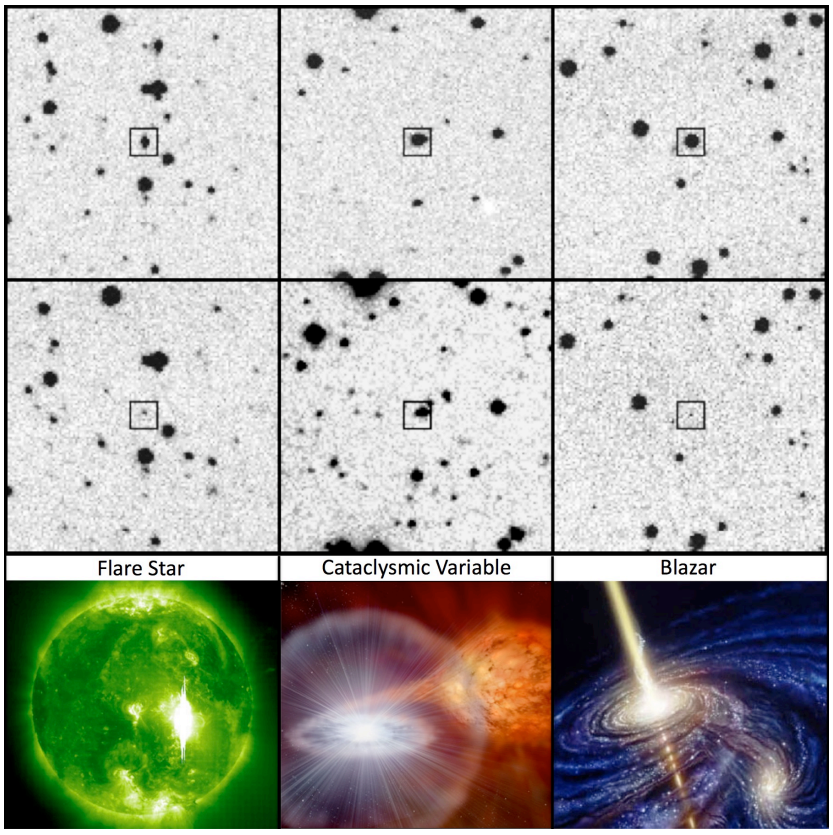
- A major new growth area of astrophysics
- Driven by the new generation of large digital synoptic sky surveys, leading to LSST, SKA, etc.
- Rich phenomenology, from the Solar system to cosmology and extreme relativistic physics
  - For some phenomena, time domain information is a key to the physical understanding
- Transformational in many ways:

  Static → Dynamic sky

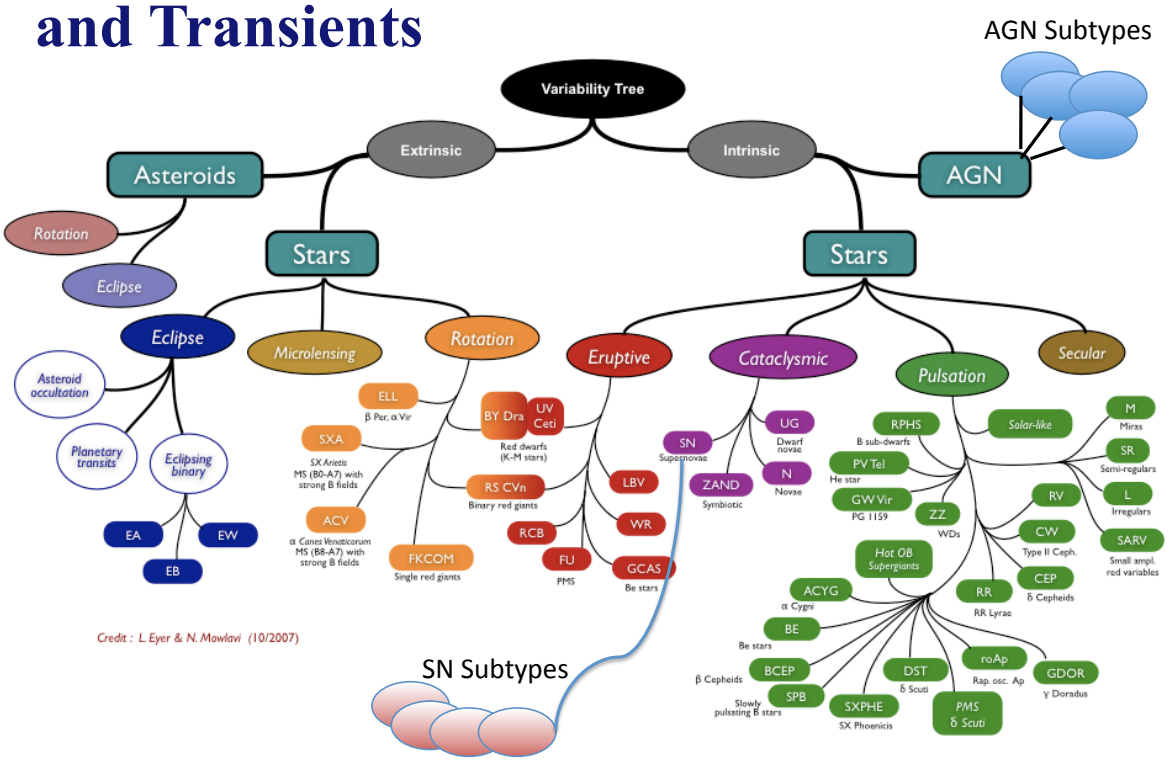
  Sources → Events
- Real-time discovery in massive data streams poses new challenges in automated classification, anomaly detection, decision making, etc.

All transients *look* the same, but can represent *vastly different* physical phenomena

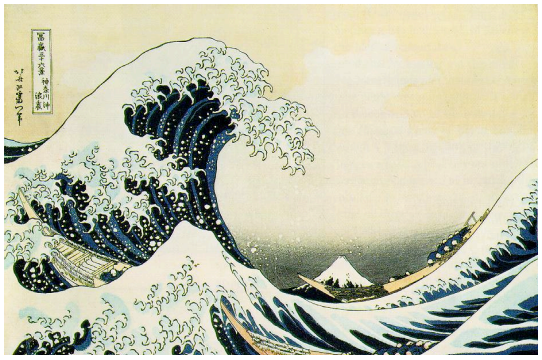
Which ones are the most interesting and worthy of follow-up efforts?



Semantic Tree of Astronomical Variables and Transients



The Tsunami Wave of the Future



- Now: data streams of ~ **0.1 TB / night, ~ 10^2 transients / night** (CRTS, PQ, PTF, various SN surveys, asteroid surveys)
- Forthcoming on a time scale ~ 1 - 5 years: ~ **1 TB / night, ~10^4 transients / night** (PanSTARRS, Skymapper, VISTA, VST…)
- Forthcoming in ~ 8 - 10 years: LSST, ~ **30 TB / night, ~ 10^5 - 10^6 transients / night**

A major, qualitative change!

- Observational follow-up needs:
  - Rapid photometric/positional monitoring
  - Rapid spectroscopy
  - Information/computation infrastructure

Transient classification technologies are essential

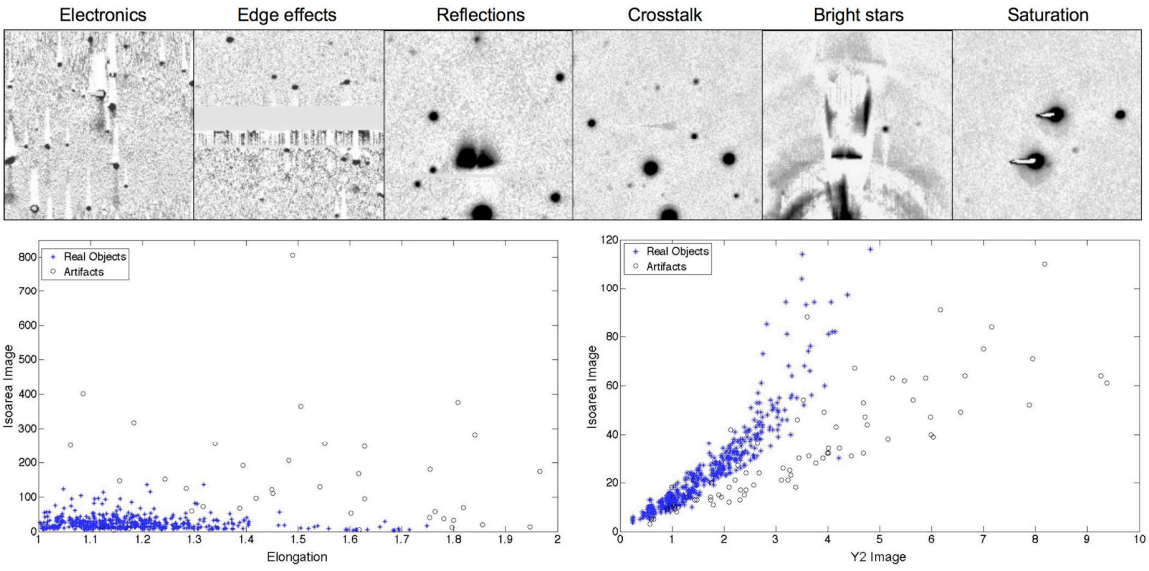
Event Classification is a Hard Problem

- 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 future Petascale data streams

- This is hard:
  - Data are sparse and heterogeneous: feature vector approaches do not work; using Bayesian approach
  - Completeness vs. contamination ☯
  - Follow-up resources are expensive and/or limited: only the most interesting events
  - Iterate classifications dynamically as new data come in

- Traditional DP pipelines do not capture a lot of the relevant contextual information, prior/expert knowledge, etc.

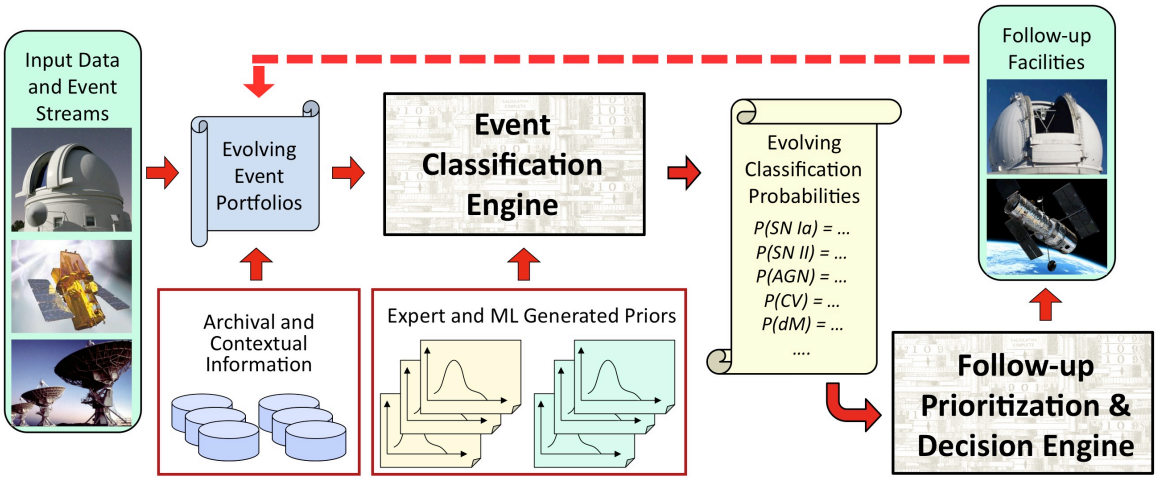
Automated Detection of Artifacts



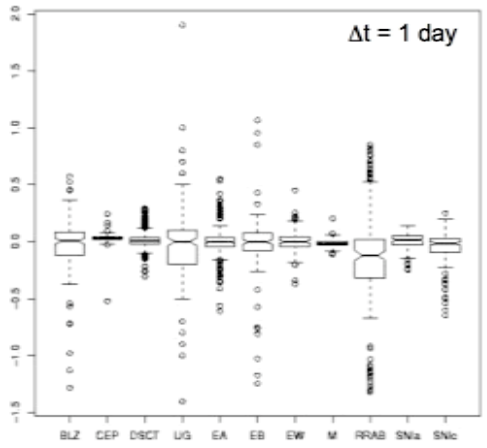
Automated classification and rejection of artifacts masquerading as transient events in the PQ survey pipeline, using a Multi-Layer Perceptron ANN

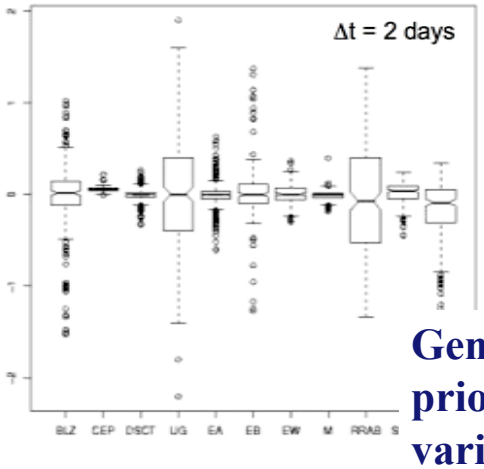
Lead:
C. Donalek

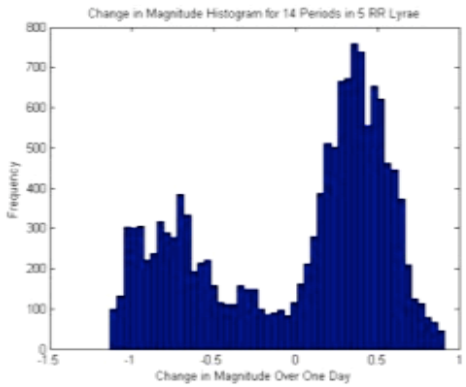
Towards the Automated Event Classification

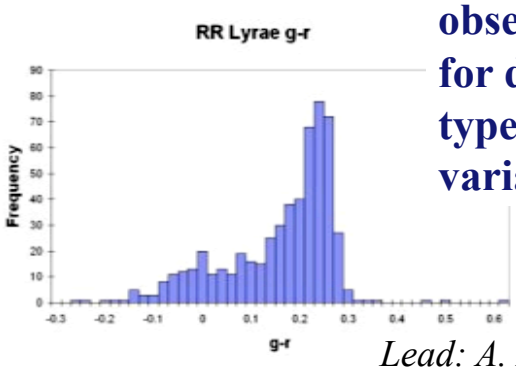


- Incorporation of the contextual information (archival, and from the data themselves) is essential
- Automated prioritization of follow-up observations, given the available resources and their cost
- A dynamical, iterative system









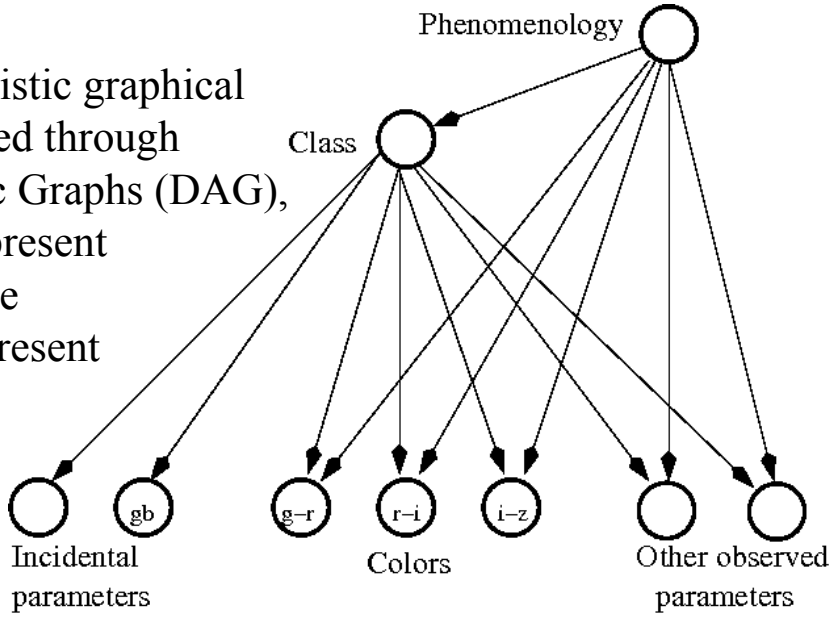
Generating priors for various observables for different types of variables

Lead: A. Mahabal

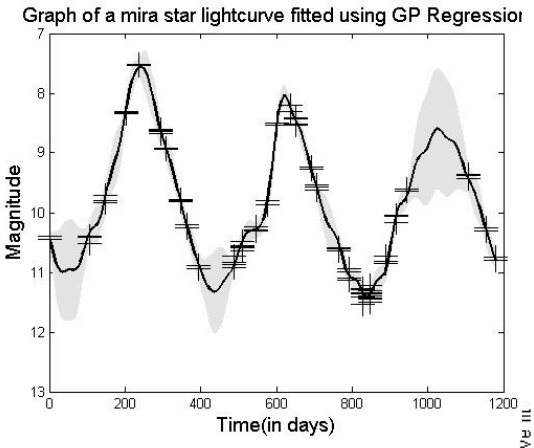
Bayesian Networks (BN)

Bayesian methodology is desirable and attractive for this task, since it can deal with missing or heterogeneous data

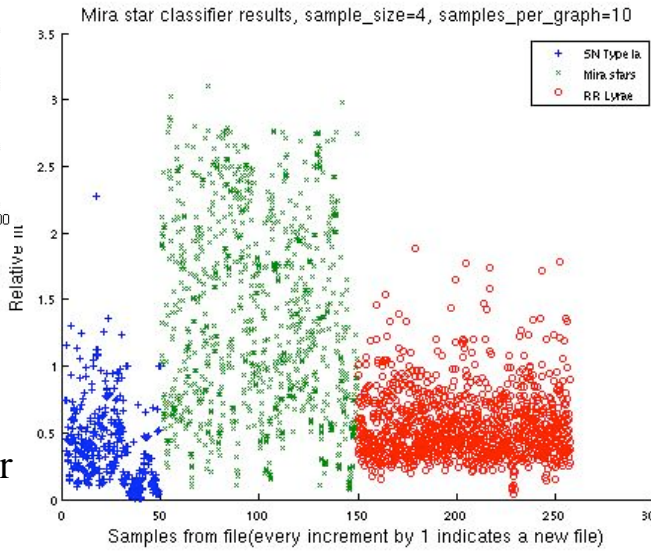
BN is a probabilistic graphical model represented through Directed Acyclic Graphs (DAG), whose nodes represent variables, and the missing arcs represent conditional independence assumptions



Gaussian Process Regression (GPR)



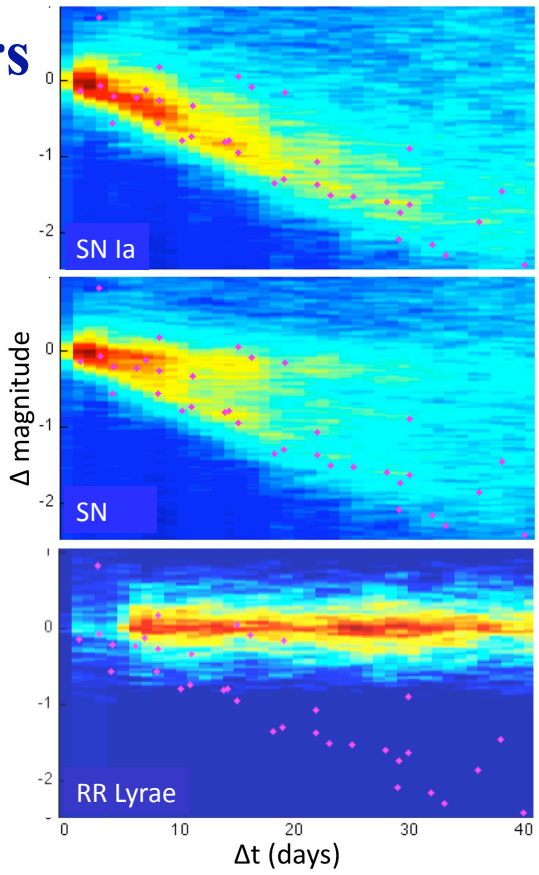
A Mira variable star light curve fitted using GPR



Given 4 random points from the light curve of a Mira variable, the probability of it being a Mira variable is higher than, say, a SN

2D Light Curve Priors

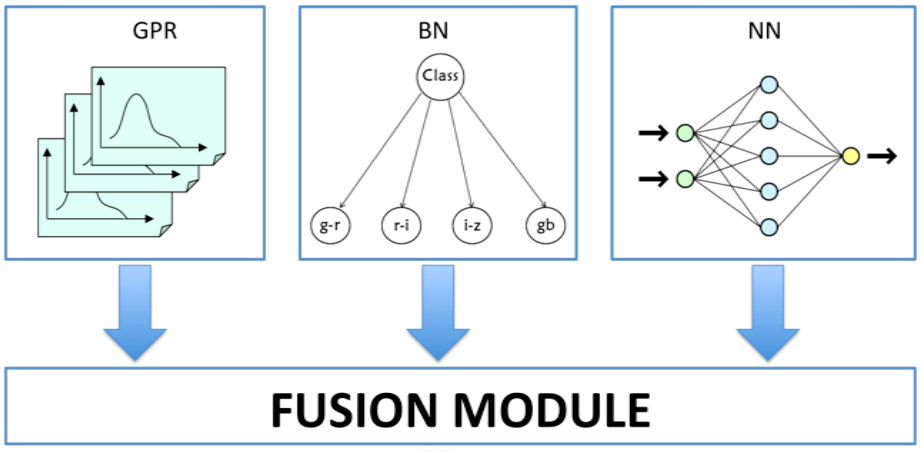
- For any pair of light curve measurements, compute the Δt and Δm, make a 2D histogram
  - Note: N independent measurem. generate N^2 correlated data points
- Compare with the priors for different types of transients
- Repeat as more measurements are obtained, for an evolving, constantly improving classif.



Lead: B. Moghaddam

Fusion Module

Colors and light curve information can be combined in one network. This "fusion module" combines the probabilistic results from each constituent classifier



Exploring a variety of techniques for optimal classification fusion:

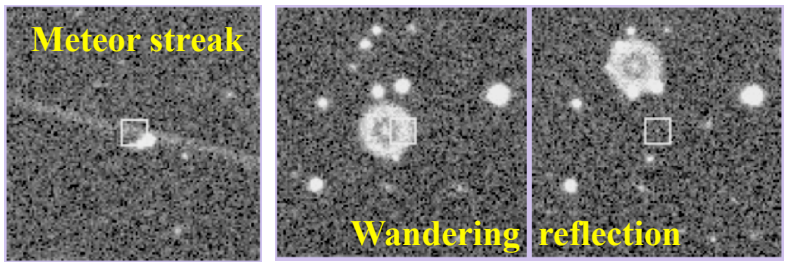
$\mathcal{P}_{\text{class}}$

Markov Logic Networks, Diffusion Maps, Multi-Arm Bandit, Sleeping Expert…

Harvesting the Human Pattern Recognition

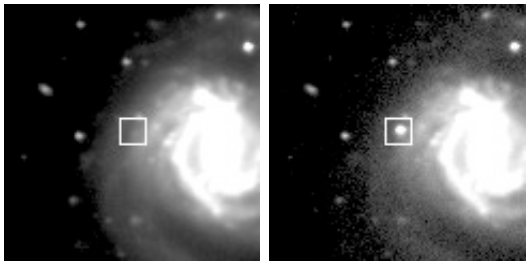
Recognizing the artifacts (false transients)

Contextual information is essential



A more sophisticated case uses a **prior (expert) knowledge:**

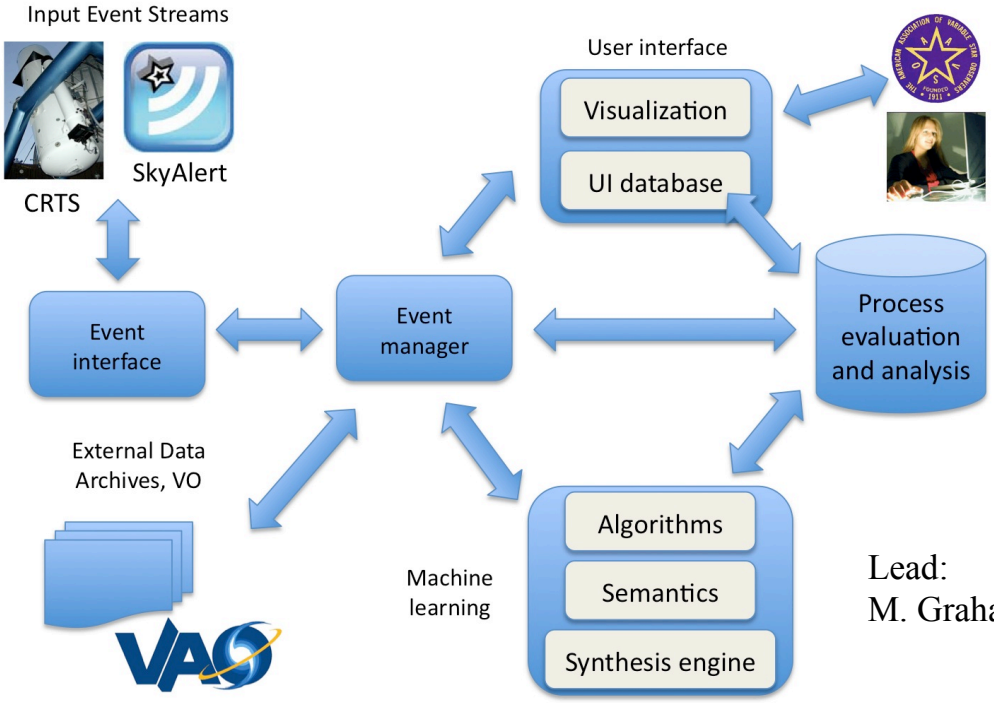
Star-like transient apparently associated with a non-coincident galaxy a likely Supernova



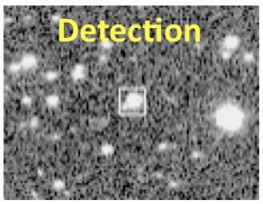
Spiral host galaxy
a possible Type II

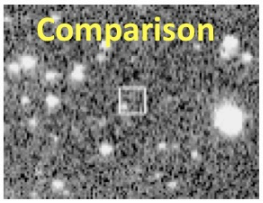
How to capture this and teach a machine to do the same thing?

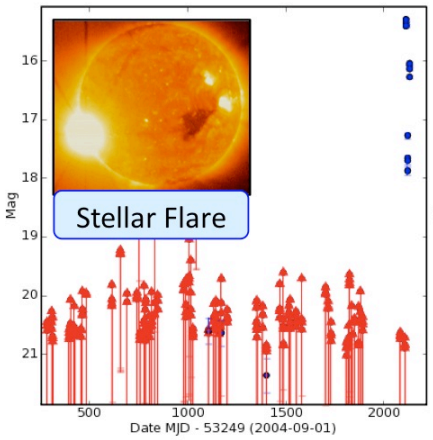
AstroCollation: Towards Harvesting Human Pattern Recognition and Domain Expertise

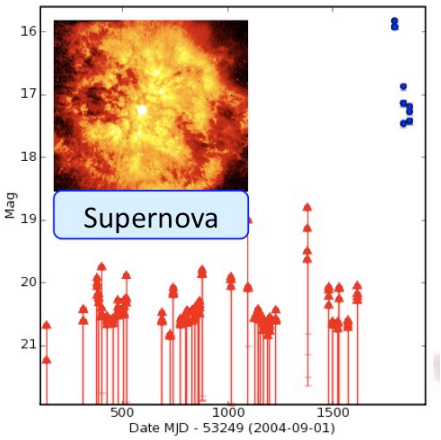


Lead:
M. Graham



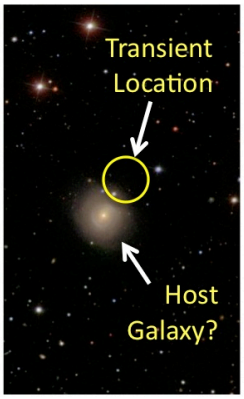






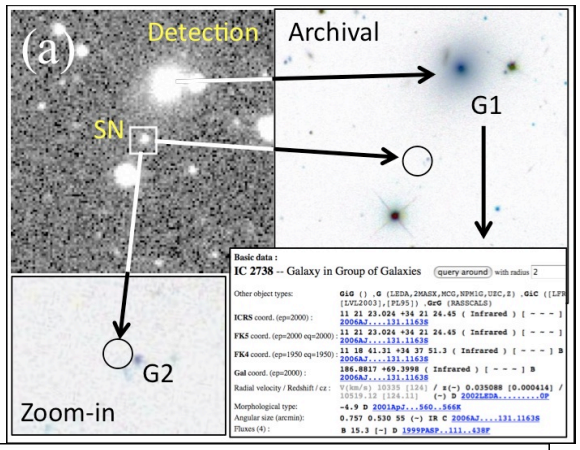
Use Case Scenario:

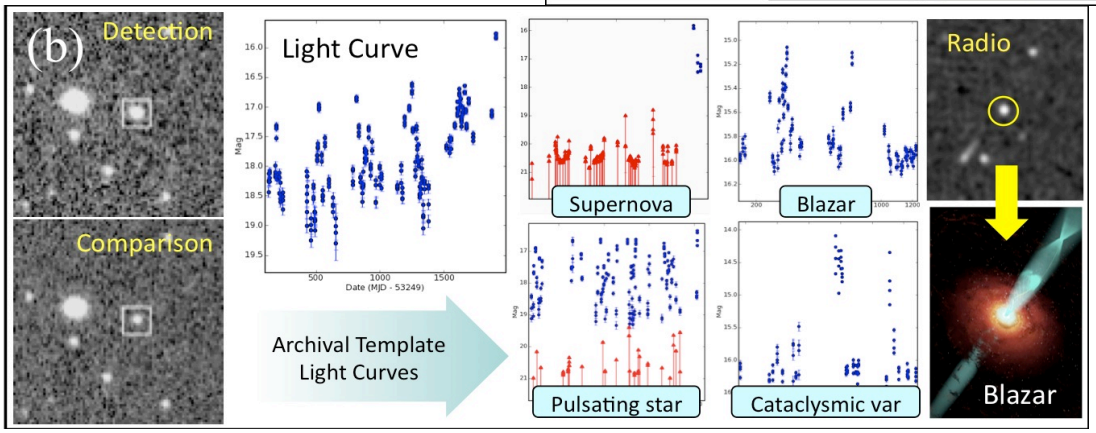
Light curves are ambiguous, but the presence of a possible host galaxy suggest that it is a Supernova

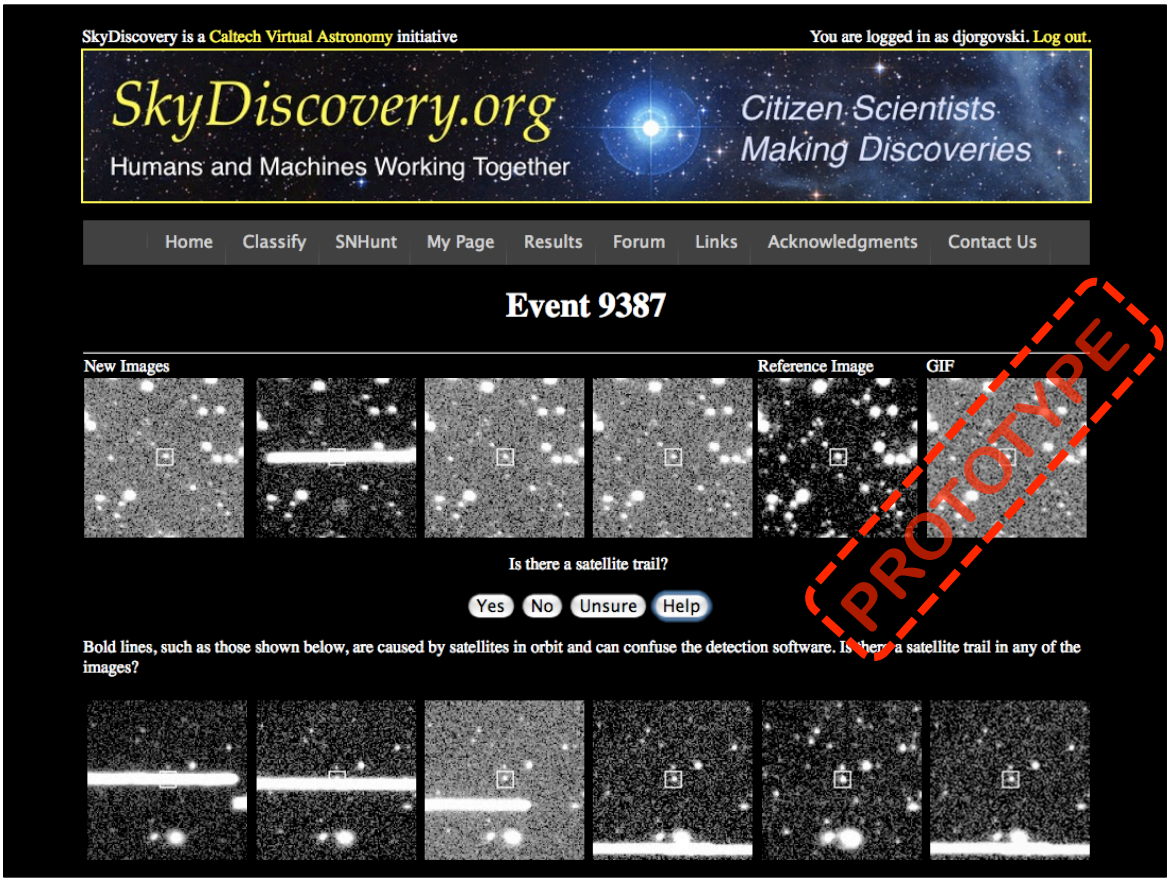


Use Case Scenarios:

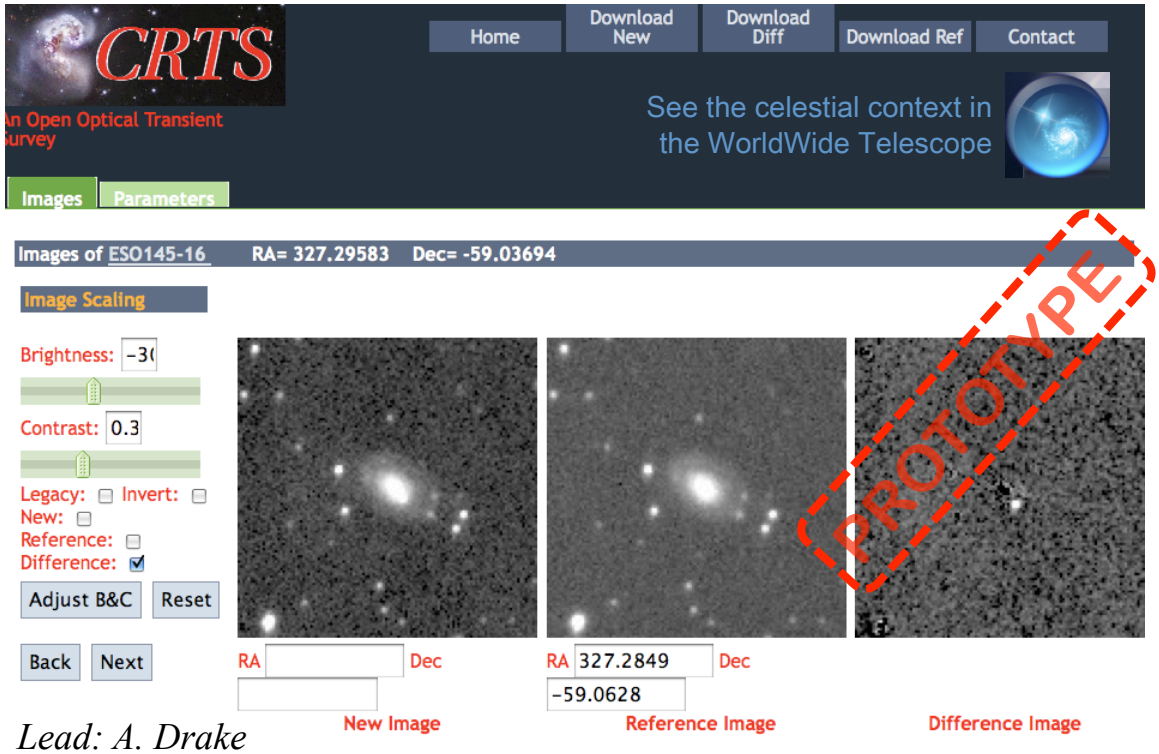
(a) Archival data on potential host galaxies provides the more likely choice,
(b) Presence of a radio source discriminates between a CV and a blazar





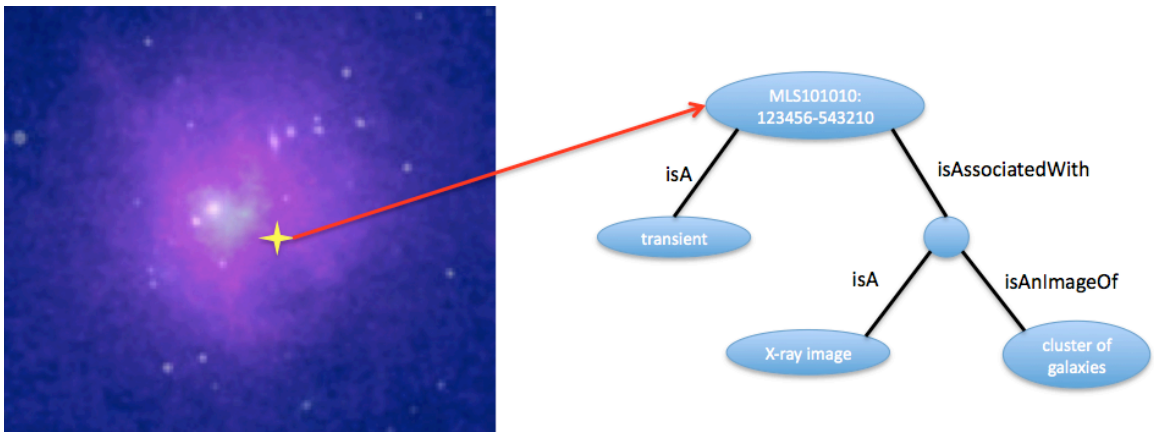


Citizen Science Supernova Hunt



Lead: A. Drake

Developing an Interface Between Carbon-Based and Silicon-Based Minds



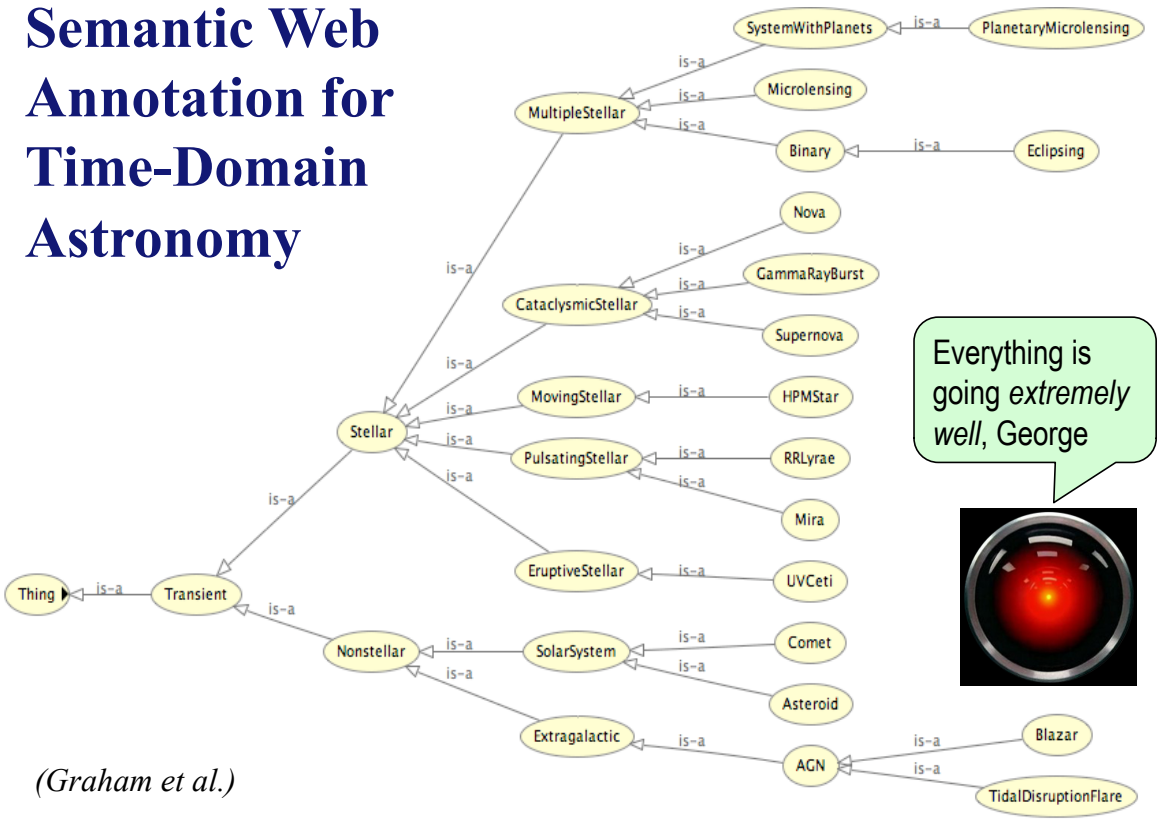
Human-annotated images (via *SkyDiscovery.org*)

⇨ Semantic descriptors

⇨ Machine processing

⇨ Novel algorithms

Semantic Web Annotation for Time-Domain Astronomy

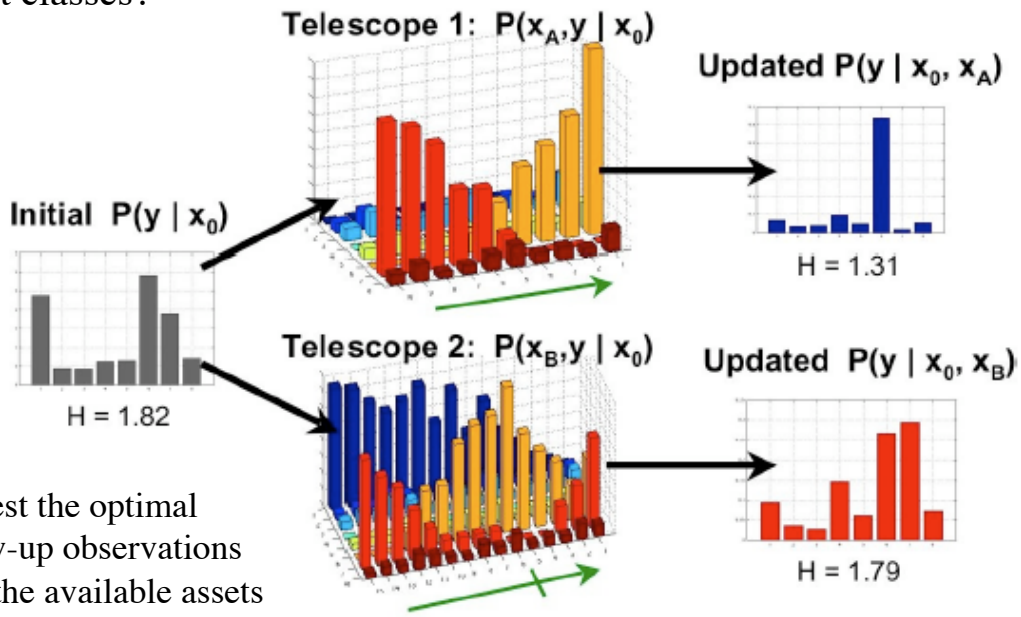


Everything is going *extremely well*, George

(Graham et al.)

Automating the Optimal Follow-Up

For the potentially most interesting events, what type of follow-up data has the greatest potential to discriminate among the competing event classes?



Request the optimal follow-up observations from the available assets

Summary

- Real-time mining of massive data streams offers great opportunities and challenges
  - Synoptic sky surveys and real-time astronomy are an excellent science & technology testbed
- We are making progress on real-time, automated, iterated event classification
  - *Not your grandma's classification problem!*
  - Sparse and heterogeneous data, real time, dynamically iterated, resource-limited
  - Next: an automated decision making for optimal follow-up observations
- Harvesting human pattern recognition skills and expertise using citizen science
- A broader relevance for a real-time mining of massive data streams