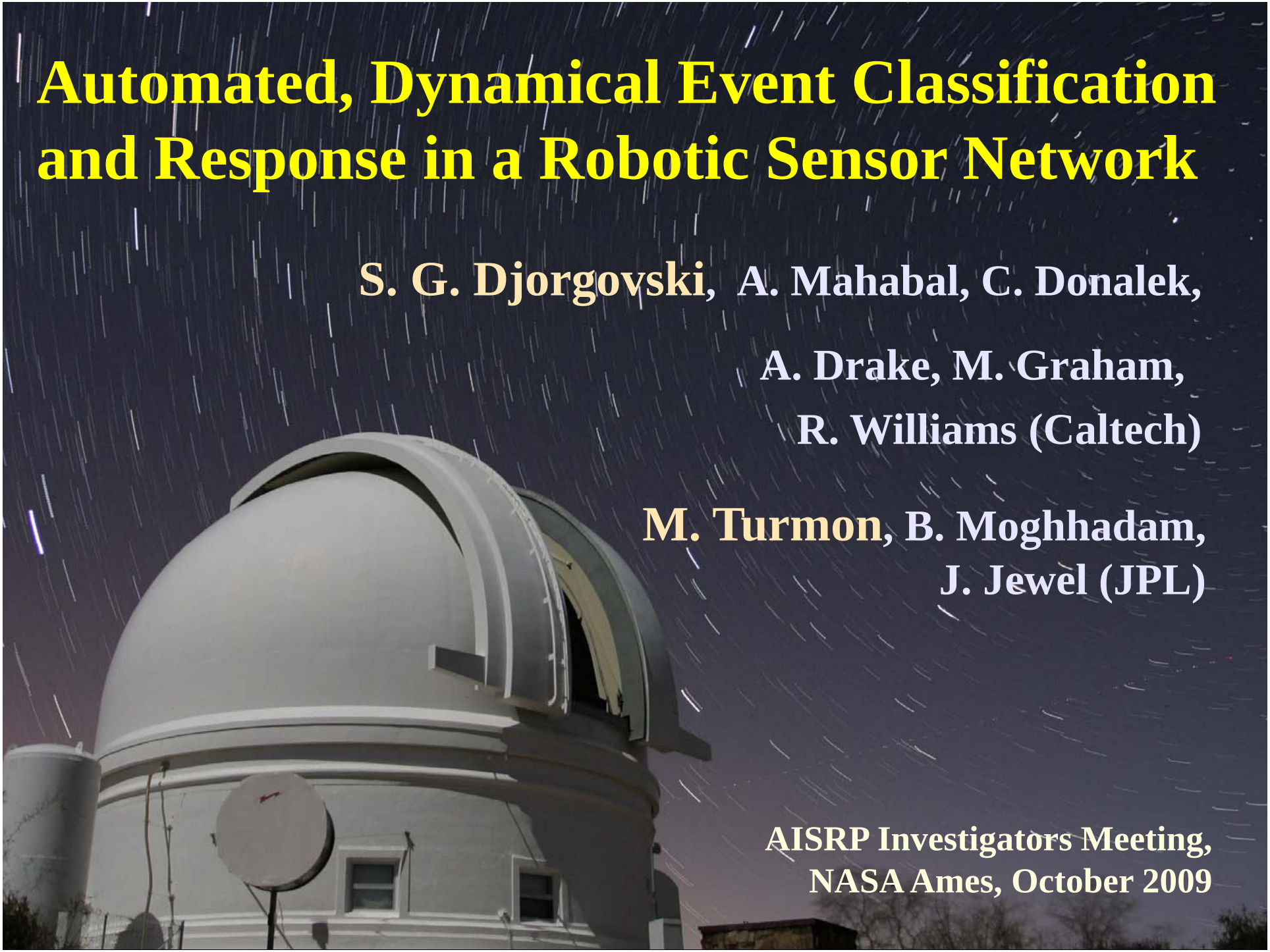


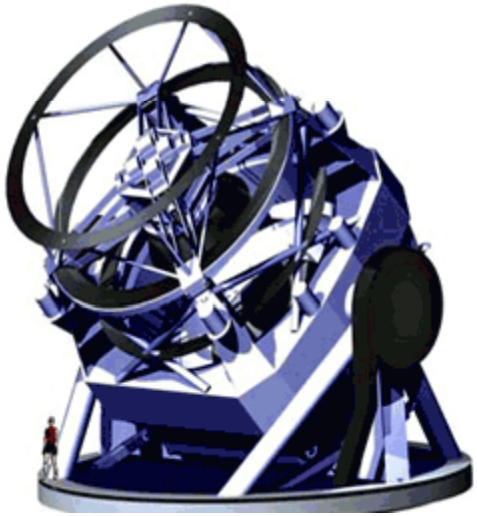
Automated, Dynamical Event Classification and Response in a Robotic Sensor Network

S. G. Djorgovski, A. Mahabal, C. Donalek,
A. Drake, M. Graham,
R. Williams (Caltech)
M. Turmon, B. Moghhadam,
J. Jewel (JPL)

AISRP Investigators Meeting,
NASA Ames, October 2009

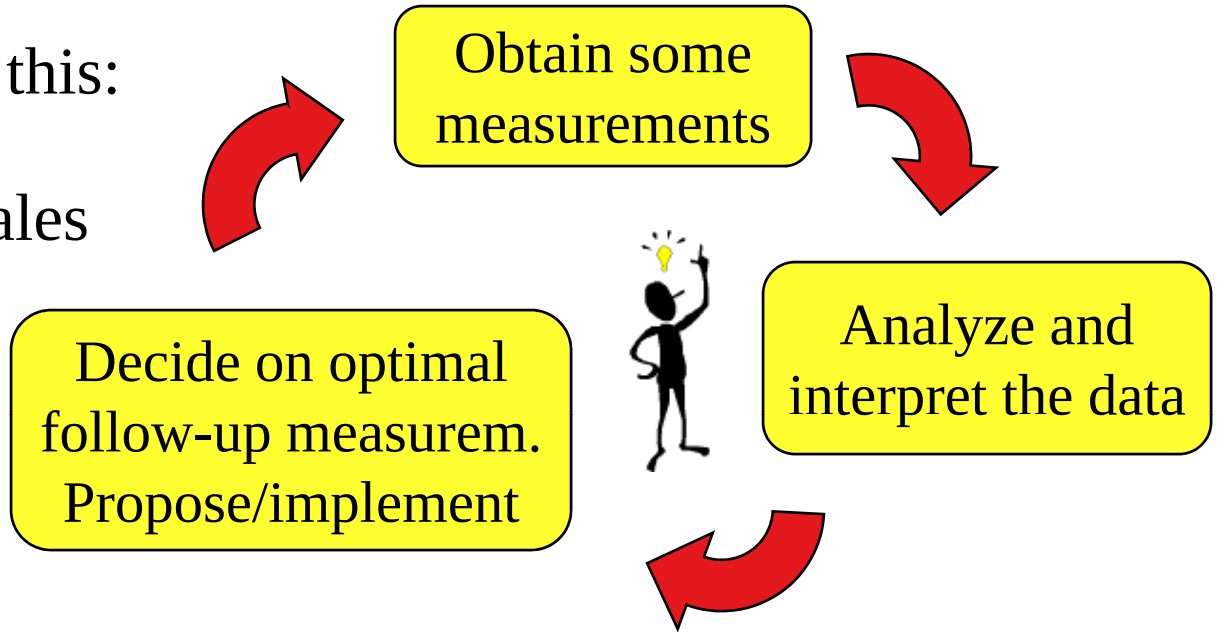
Real-Time Mining of Petascale Data Streams

- Distributed sensor networks – both ground-based and space-based – and major new scientific instruments (e.g., the LHC) are starting to generate Petascale data streams
- In addition to the huge data volumes, this brings new challenges along with the opportunities: ***detecting and recognizing interesting events/phenomena in real time, and responding to them in some way***
- Synoptic digital sky surveys are becoming the dominant data provider in astronomy, leading towards the LSST
  - A broad range of exciting astrophysics
  - New challenges: reliable real time processing and event detection, event classification, directed follow-up…
  - Broader relevance, e.g., autonomous spacecraft networks



Scientific Measurement Cycle

Typically it looks like this:



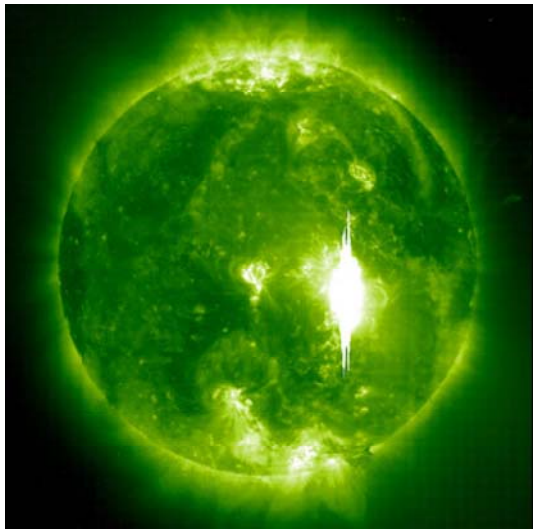
Characteristic time scales
typically ~ a *year*
(or at best, days)

But what if the
phenomena we study last/change on time scales of *minutes/hours*?
… and the data rates are measured in TB's per day or higher?
… and the measurement, data, computation assets are distributed?

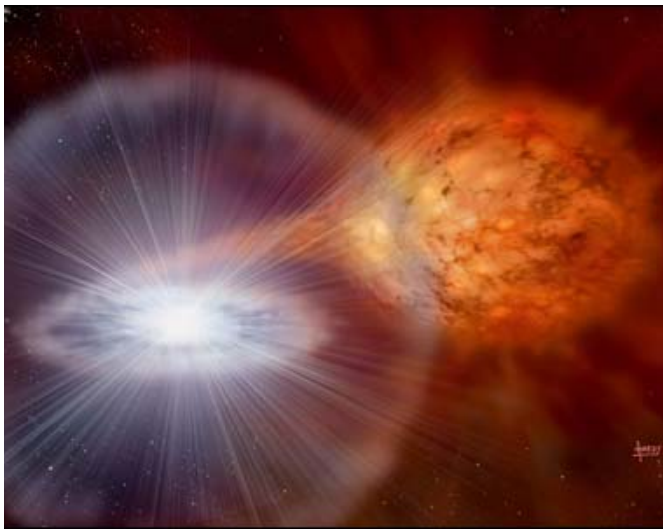
What is required is a system which is:

- Fully automatic/robotic, with no humans in the loop
- Draws on a number of important computational technologies

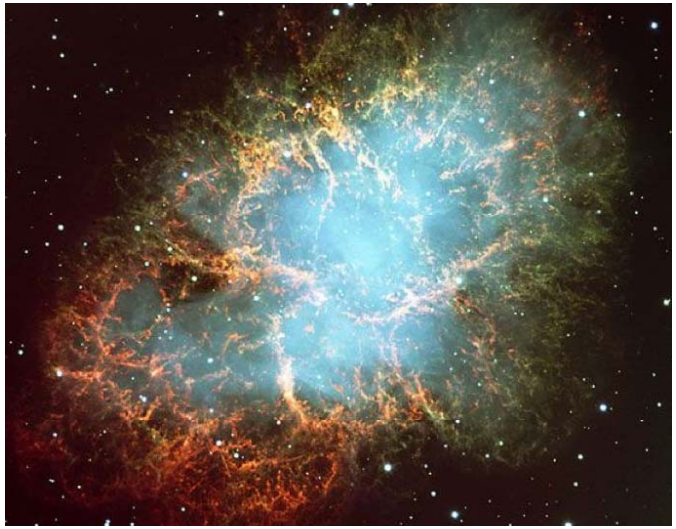
A Broad Variety of Phenomena



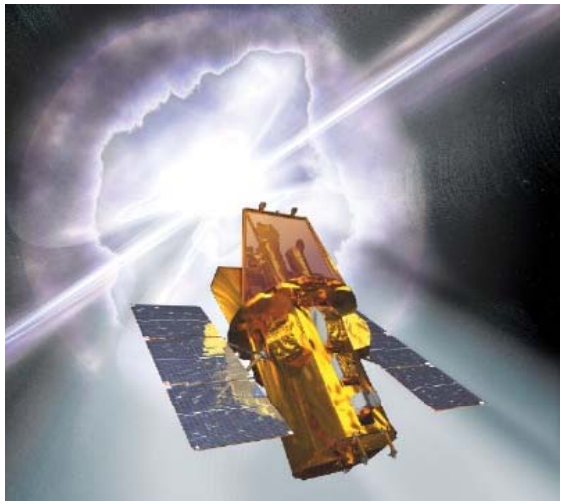
Flaring stars



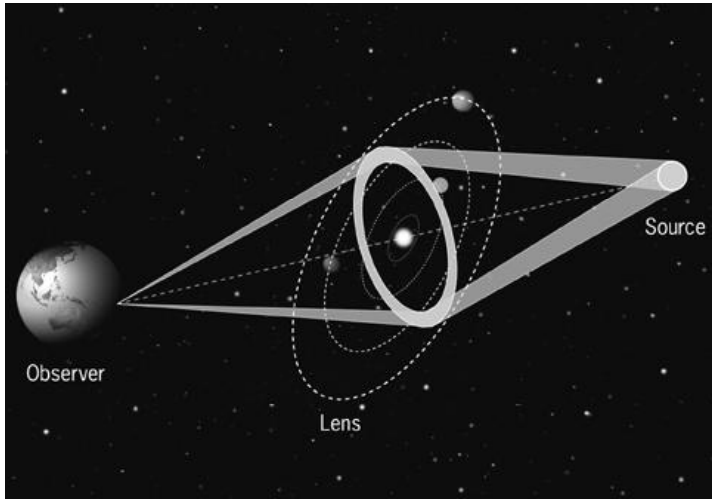
Novae, Cataclysmic Variables



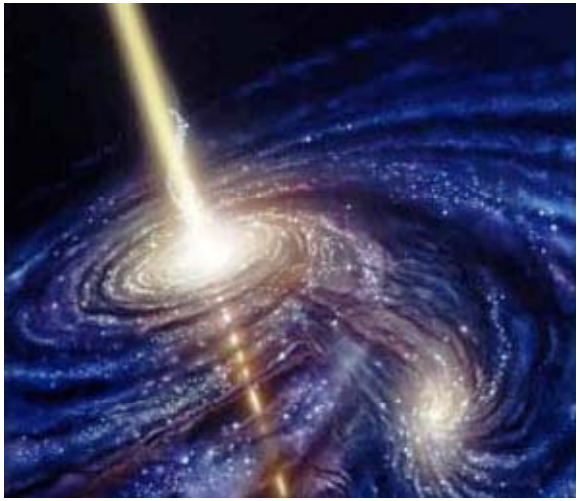
Supernovae



Gamma-Ray Bursts



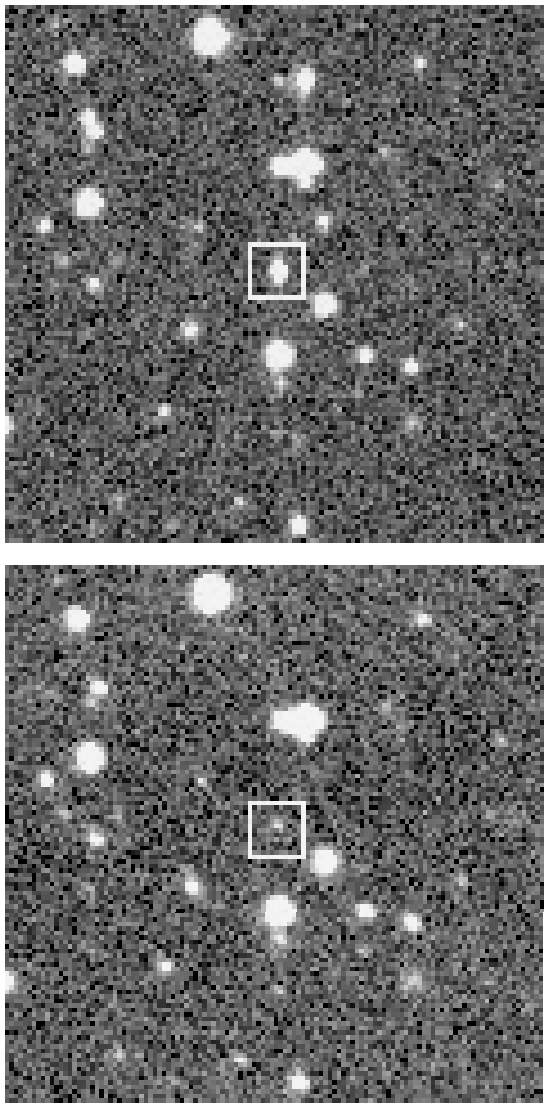
Gravitational Microlensing



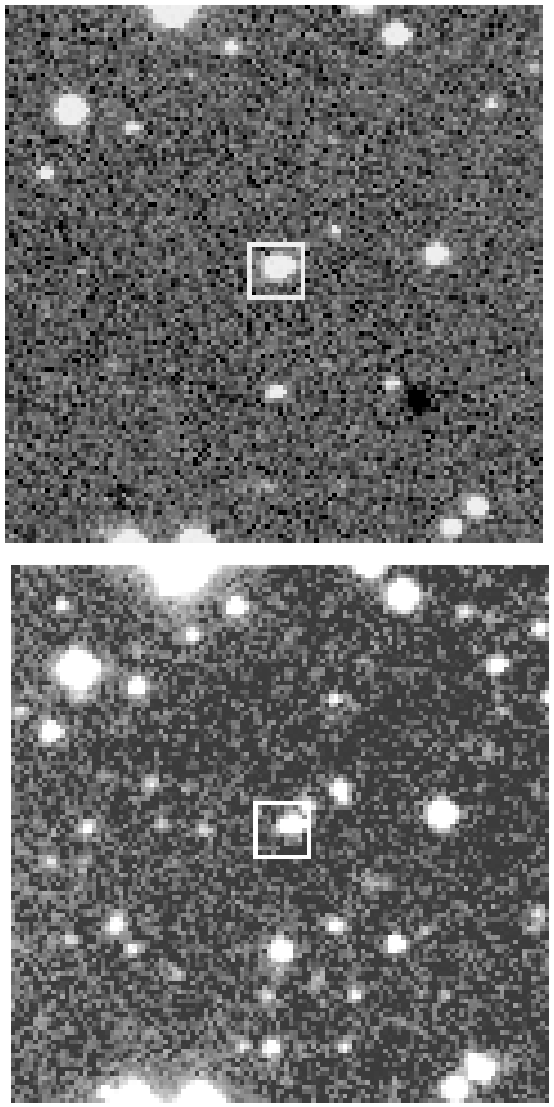
Accretion to SMBHs

Examples of CRTS Transients

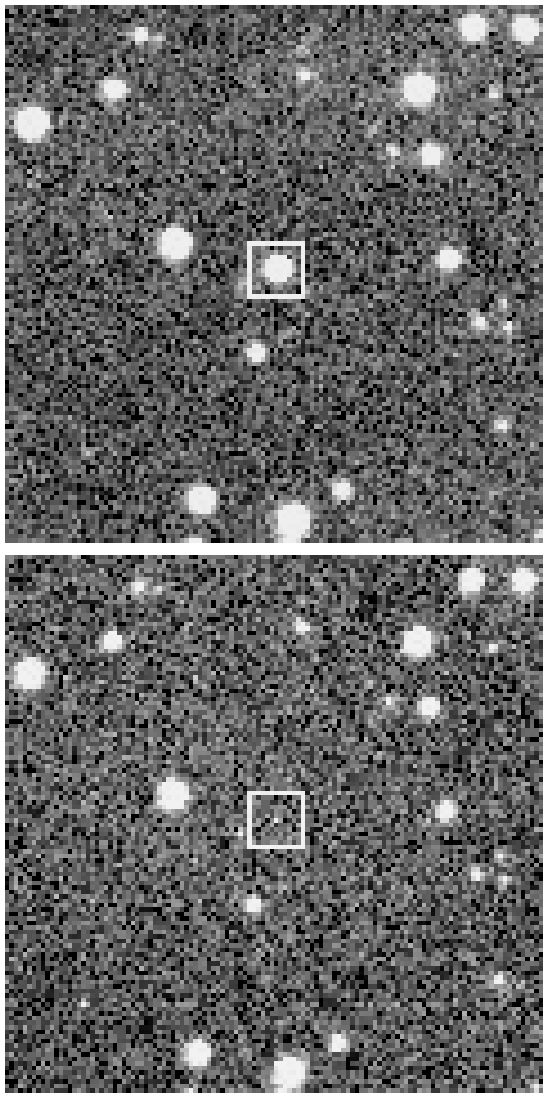
CSS090429:135125-075714
Probable flare star



CSS090429:101546+033311
Probable dwarf nova



CSS090426:074240+544425
Blazar, 2EG J0744+5438



The Palomar-Quest Event Factory

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

Compare with the baseline sky

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

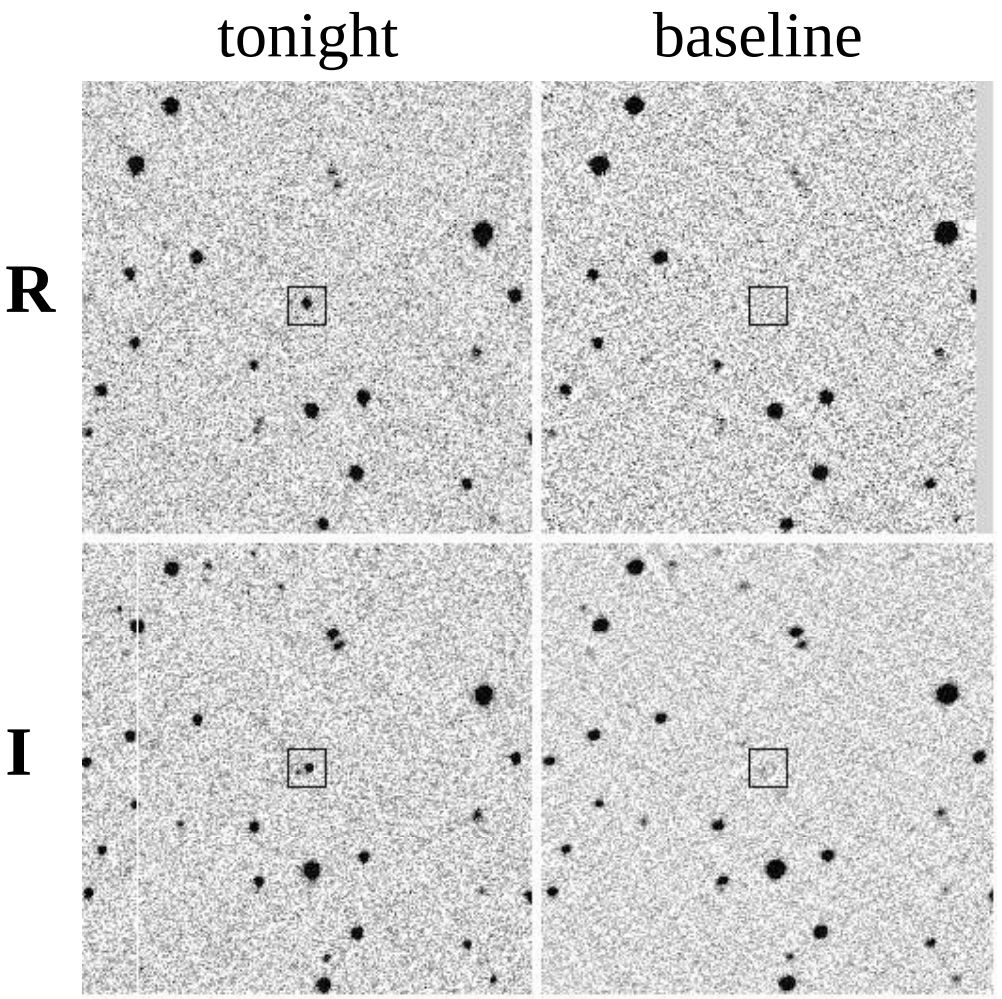
Remove instrum. artifacts

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

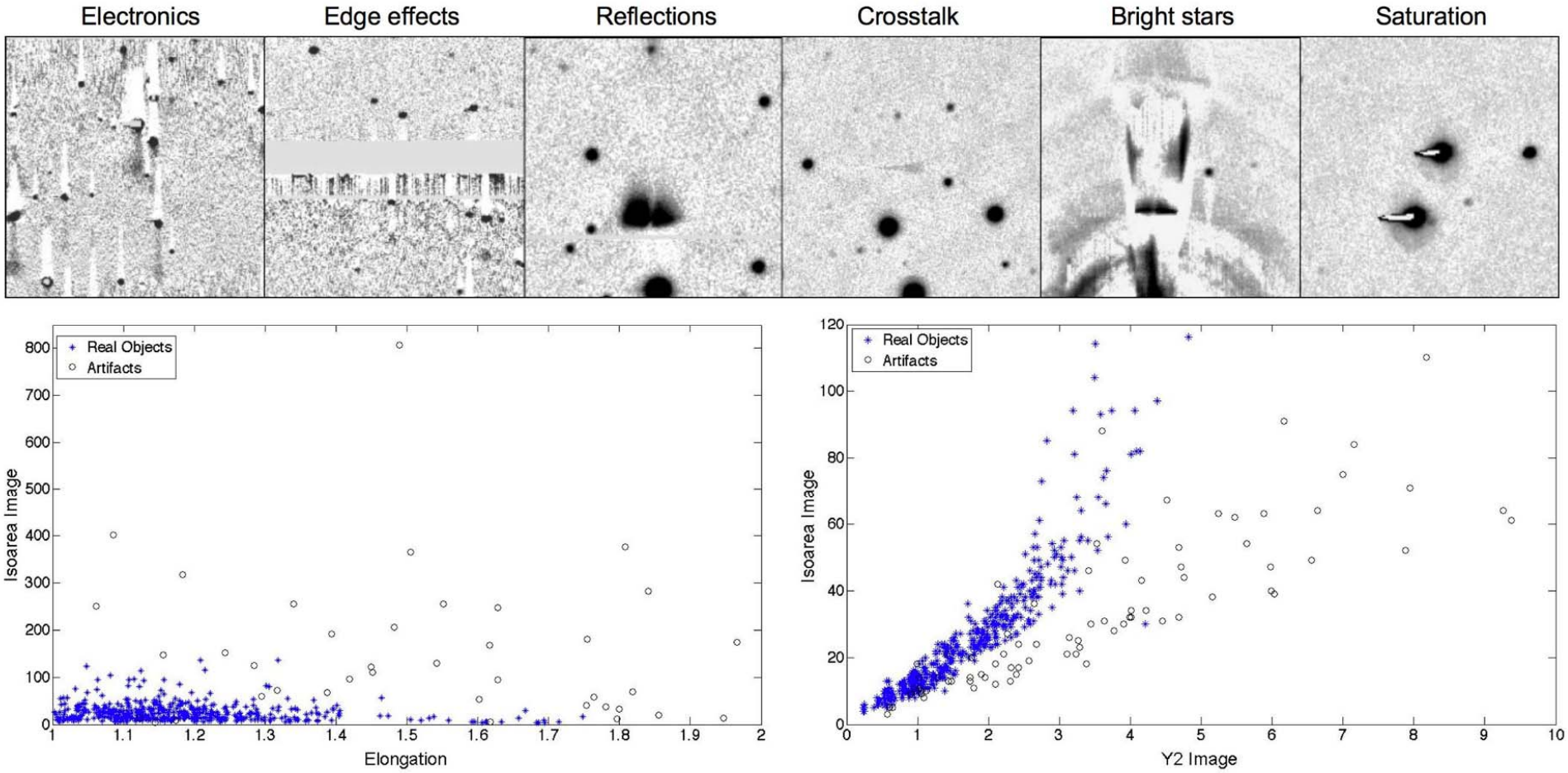
Remove asteroids

Identify ~ 1 - 10 possible Astrophysical transients

Classification and follow-up

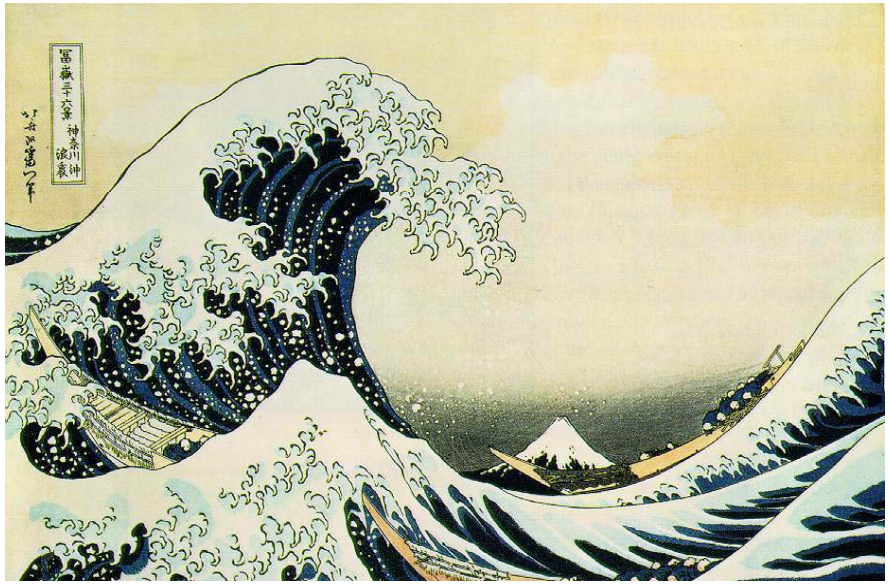


Automated Detection of Artifacts



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

The (Tsunami) Wave of the Future



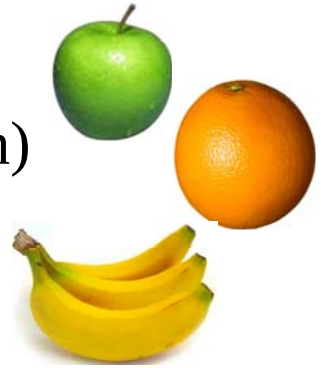
- Now: data streams of **~ 0.1 TB / night, ~ 10 - 10^2 transients / night** (SDSS, PQ, various SN surveys, asteroid surveys)
- Forthcoming on a time scale ~ 1 - 5 years: **~ 1 TB / night, ~10^4 transients / night** (PanSTARRS, Skymapper, VISTA, VST…)
- Forthcoming in ~ 5 - 10 years: LSST, **~ 20 TB / night, ~ 10^5 - 10^6 transients / night**
- Observational follow-up needs:
  - Rapid photometric/positional monitoring
  - Rapid spectroscopy
  - Information/computation infrastructure

A major, qualitative change!

Transient classification technologies are essential

Why Is This Hard?

1. Data are sparse and heterogeneous
   - Different measurements for different events, random(ish) sampling, variable data quality, archival coverage…
   - Feature vector methodology generally does not work

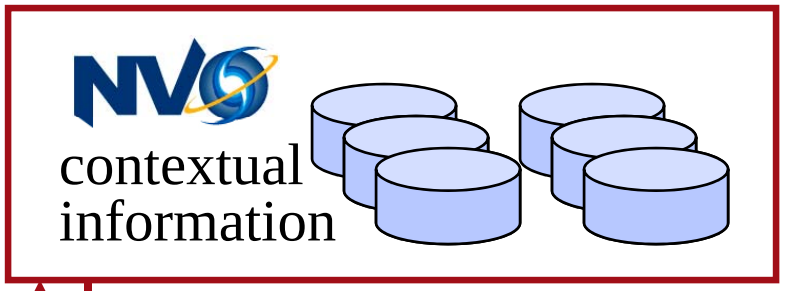


2. High completeness / low contamination requirement ☯
3. Must be done in real time and iterated dynamically
4. Follow-up resources are expensive and/or limited
   - Only the most interesting/valuable events
   - Decide on the optimal follow-up to resolve ambiguities

5. Could be resource-limited (compute power, bandwidth, etc.)
   - E.g., space-based sensor systems, spacecraft networks
6. Huge and growing data volumes
7. Must be scalable to more and different data inputs

Towards Automated Event Classification

A **necessity** for large synoptic surveys

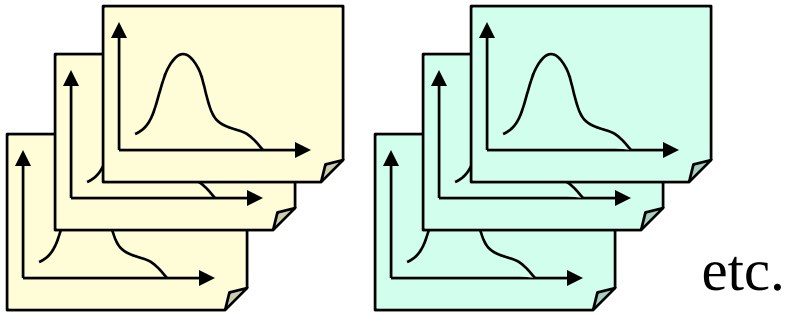


Event parameters:
$m_1(t)$, $m_2(t)$,…
α, δ, μ, …
image shape…

Event Classification Engine

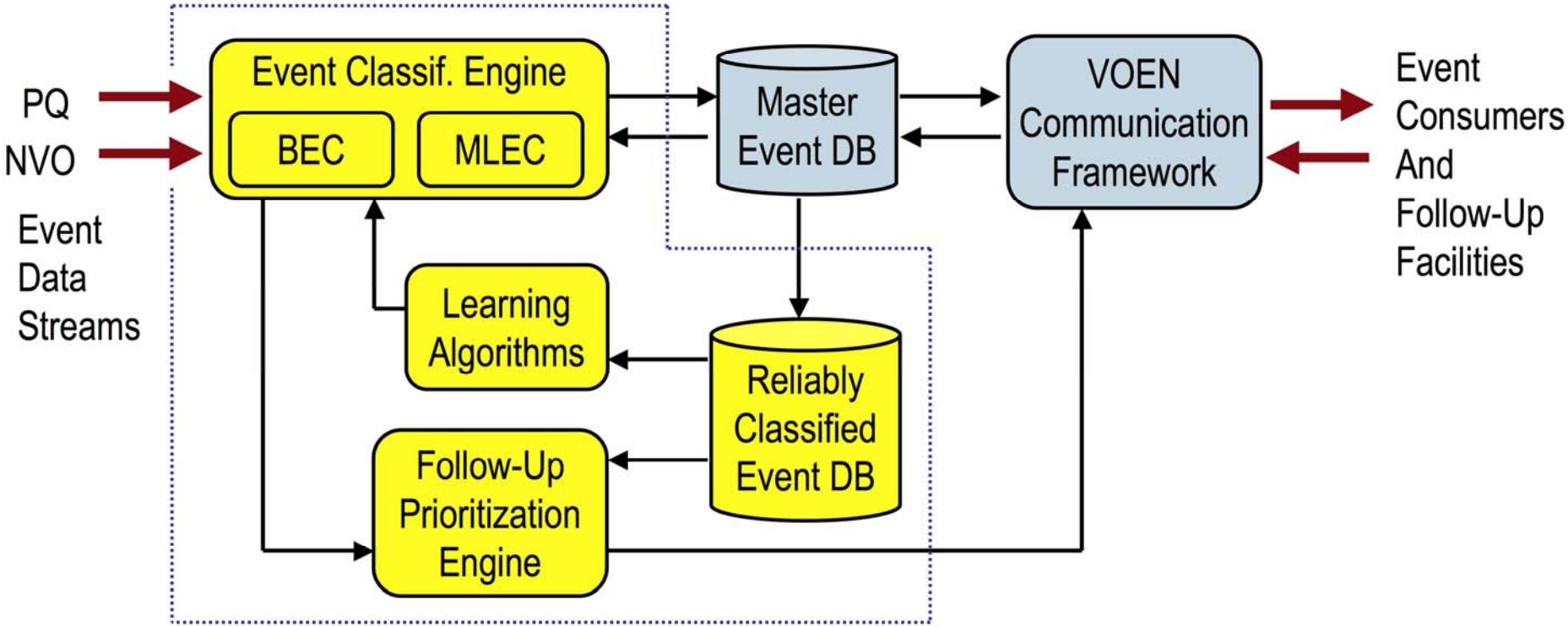
P(SN Ia) = …
P(SN II) = …
P(AGN) = …
P(CV) = …
P(dM) = …
….

Expert and ML generated priors



Classification probabilities (evolving, iterated)

Bayesian and Machine Learning Event Classification

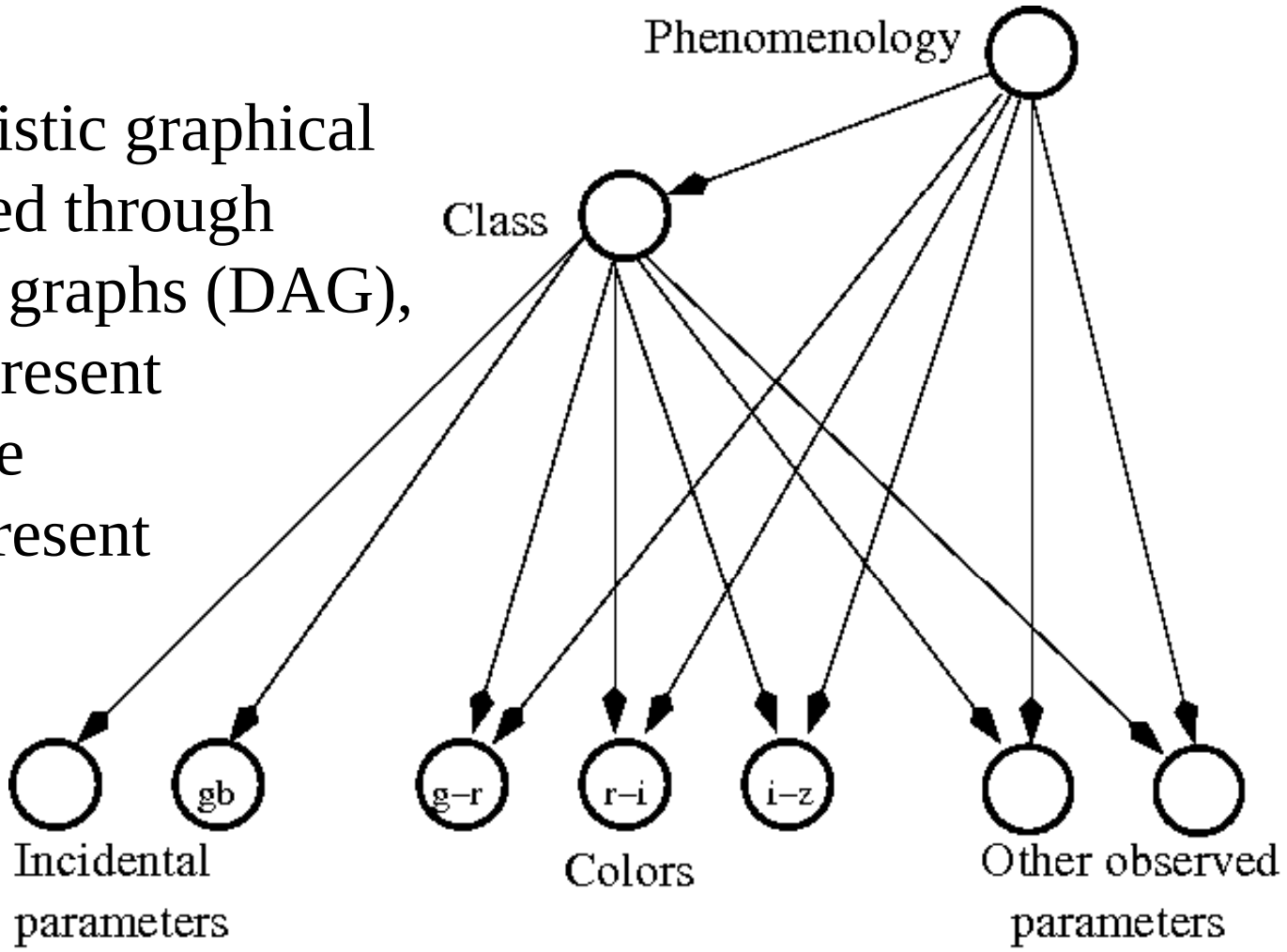


- Bayesian methods are more tolerant of heterogeneous or missing data; easy to add new event classes
- Machine learning approach (ANN and SVM, unsupervised classif.) will get better as the database of known events grows

Bayesian Networks (BN)

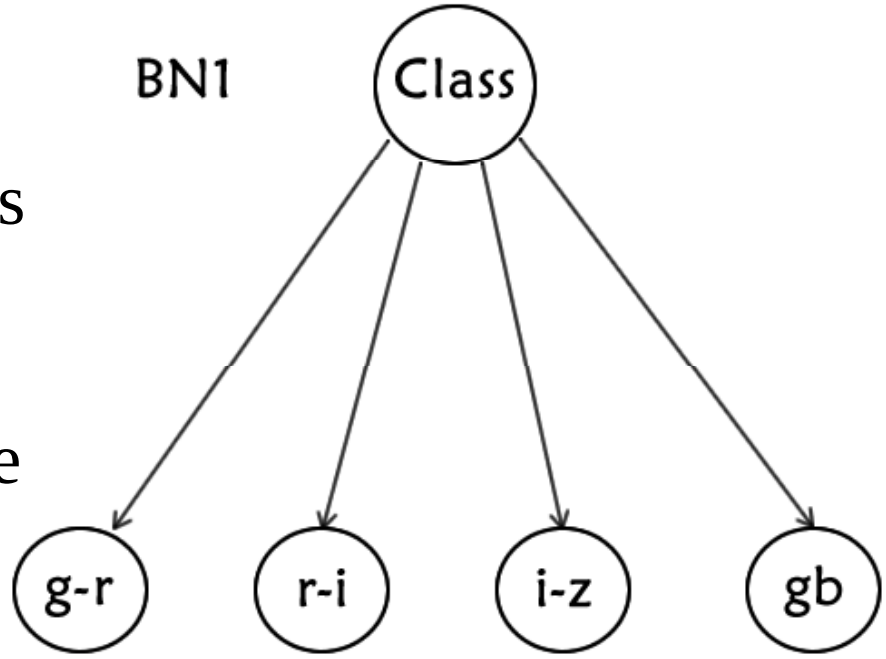
Bayesian methodology is desirable and attractive for this task, since it can deal with missing or heterogeneously sampled data.

BN is a probabilistic graphical model represented through Directed acyclic graphs (DAG), whose nodes represent variables, and the missing arcs represent Conditional independence assumptions.



BN: First Results

In the network shown here, colors and Galactic latitude have been used to generate priors. For testing purposes four classes have been used: CVs, Supernovae, Blazars, and "Rest"



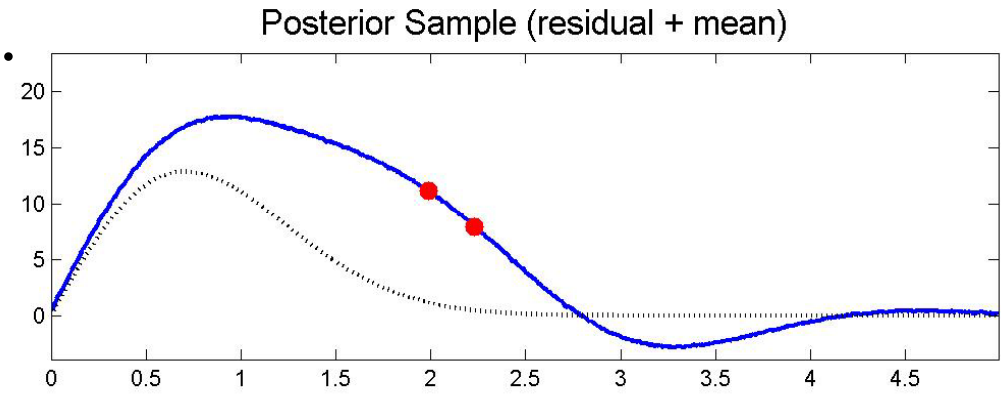
Confusion matrix: rows are the true classes, columns are the predicted classes

Classes	CV	SN	Blazars	Rest
CV	110 (0.80)	5 (0.04)	7 (0.05)	15 (0.11)
SN	22 (0.19)	64 (0.56)	12 (0.10)	17 (0.15)
Blazars	4 (0.13)	0 (0)	19 (0.64)	7 (0.23)
Rest	12 (0.39)	4 (0.13)	6 (0.19)	9 (0.29)

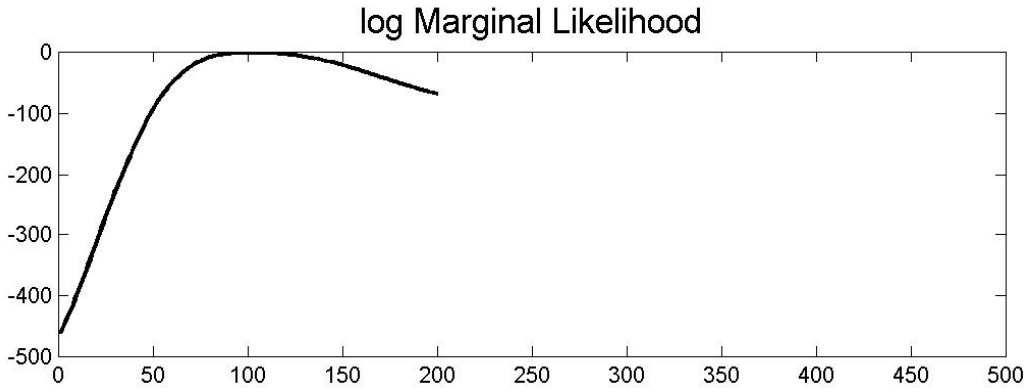
Gaussian Process Regression (GPR)

A generalization of a Gaussian probability, specified by a mean function and a positive definite covariance function.

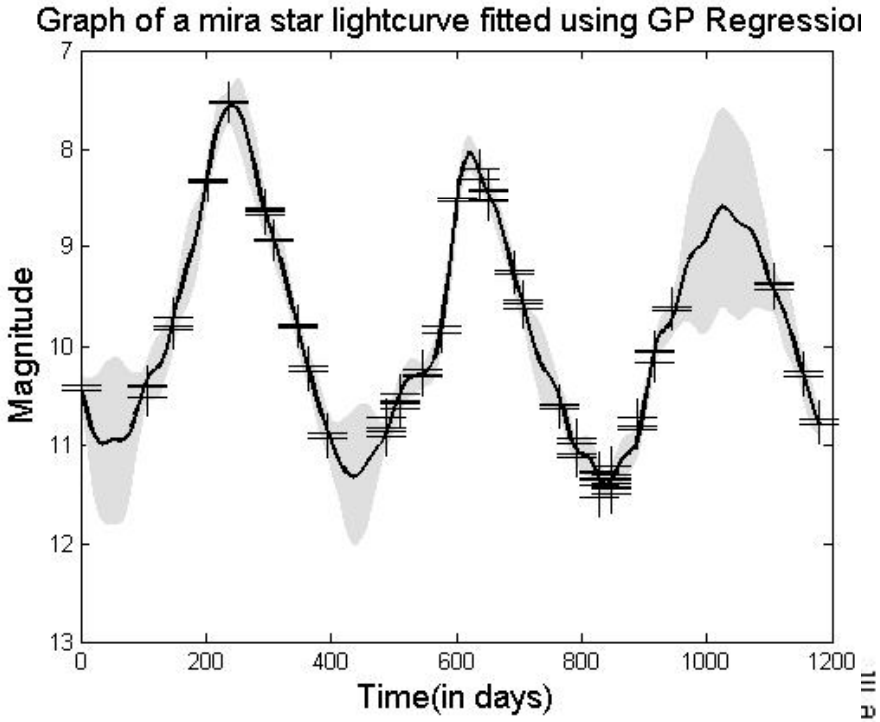
Given two flux measurement points for a new transient we can then ask which of the different models it fits, and what stage of their period or phase. The more points you have, the better the estimate.



Log marginal likelihood of a pair of points corresponding to different parts of a lightcurve.



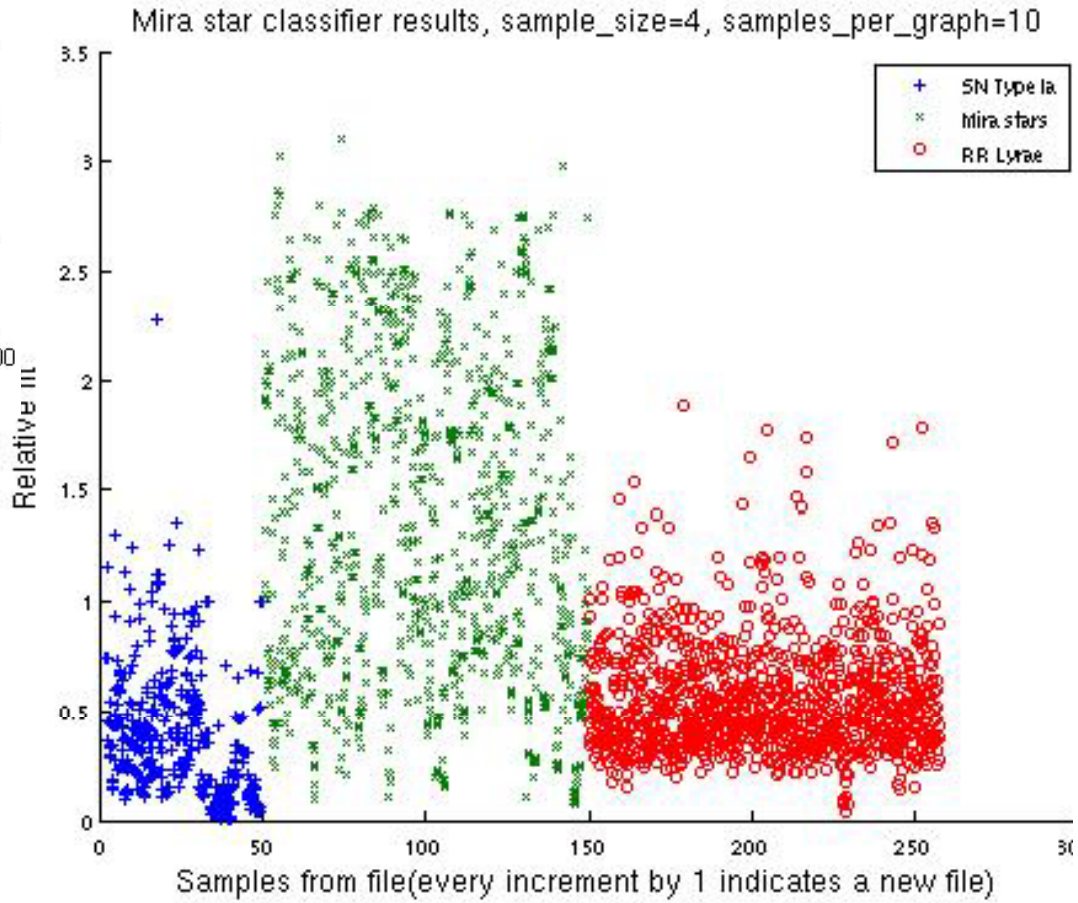
GPR: Preliminary Results



A Mira variable star light curve fitted using GPR

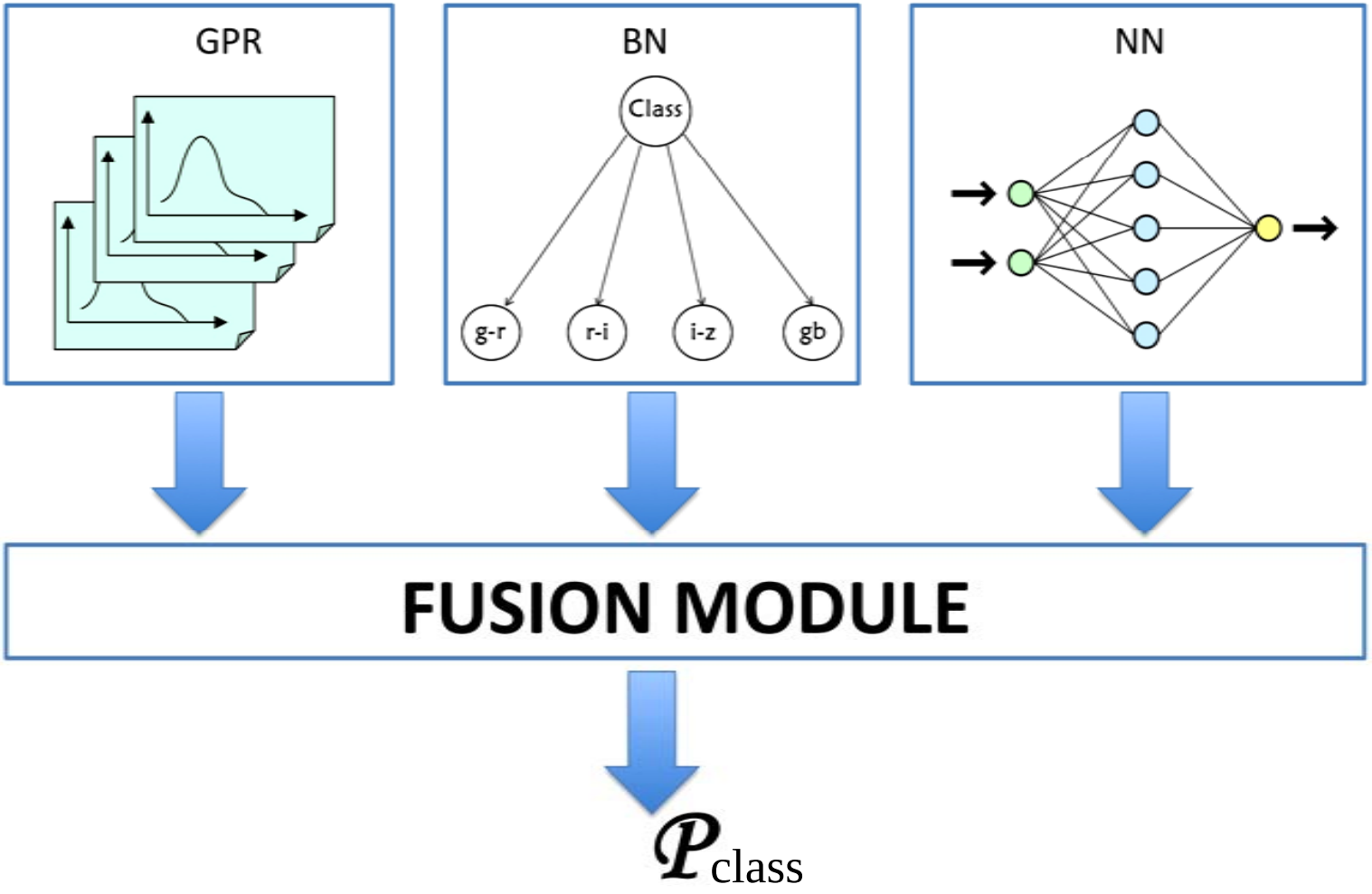
↵

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



Fusion Module

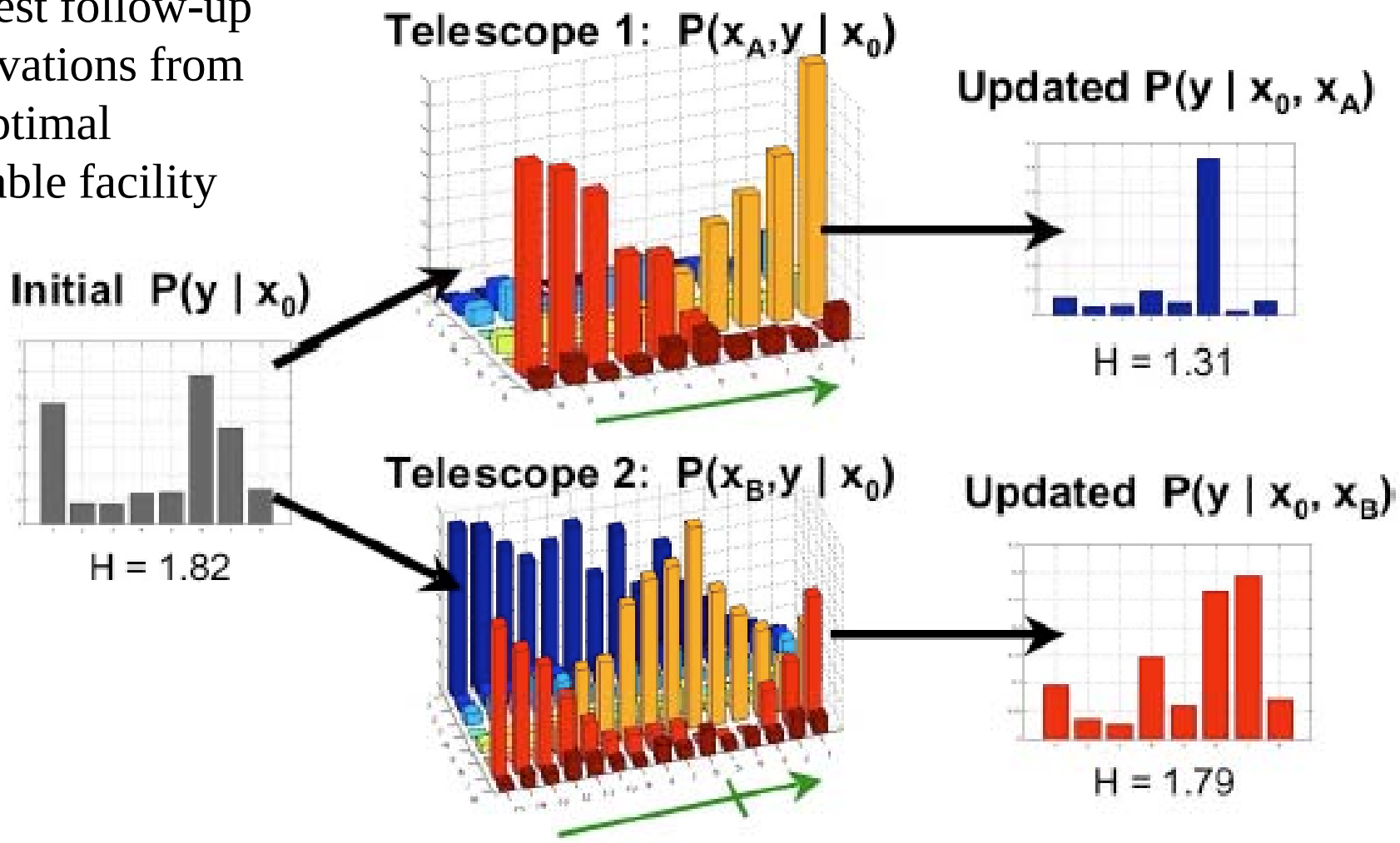
Colors and light curve information can be combined in one network. This "fusion module" combines the probabilistic results from each constituent classifier.



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



Summary

- Real-time mining of massive data streams offers great opportunities and challenges
  - Synoptic sky surveys and real-time astronomy are an excellent 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
- A broader relevance, e.g.:
  - Autonomous spacecraft networks
  - Environmental sensor networks; etc.