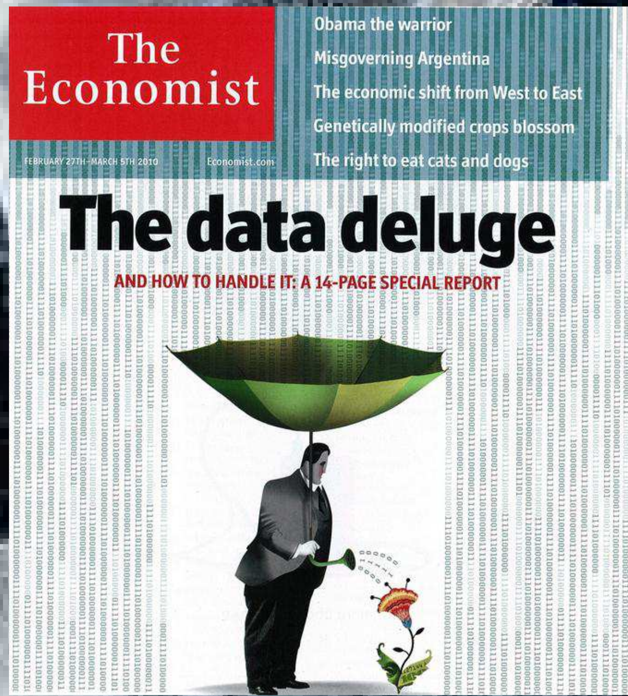
Time Domain Astronomy: a new frontier of astronomy
(and a challenge for computational science)

S.G. Djorgovski, A. Mahabal, C. Donalek, M. Graham, A. Drake

G. LONGO, M. BRESCIA (INAF-OAC),
S. Cavuoti , M. Annunziatella, D. De Cicco and many others from DSF

and

Many collaborators at CfA, JPL, etc.

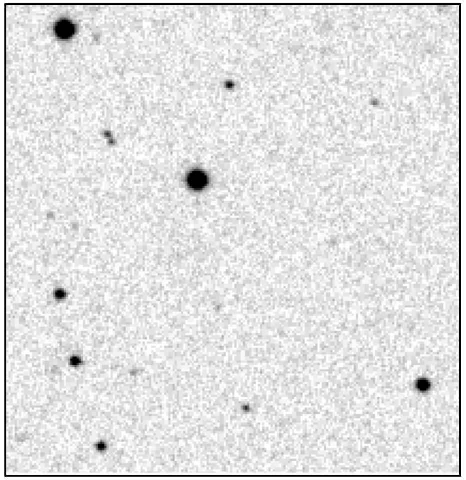


Many slides from G.S. Djorgovski

Outline

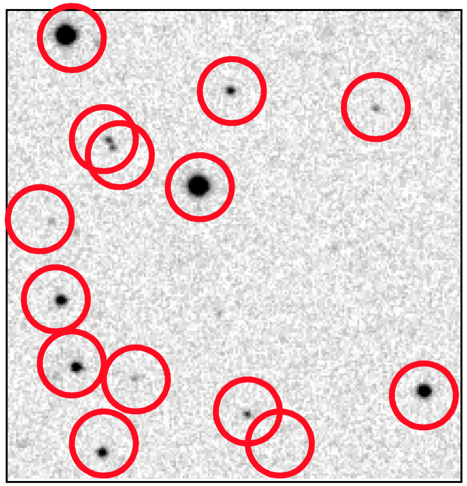
- Introduction: the new astronomical scenario and the key role of ICT
- Why it is not only a “do iy bigger and do it faster” business
- Opening the time domain
- Exploration of highly-dimensional parameter spaces
  - Classification, clustering, and outlier search
  - Multivariate correlation search
- Mining the data
  - Classification of transient events and variable sources
  - Optimal decisions for follow-up observations

OA: from Images to Knowledge

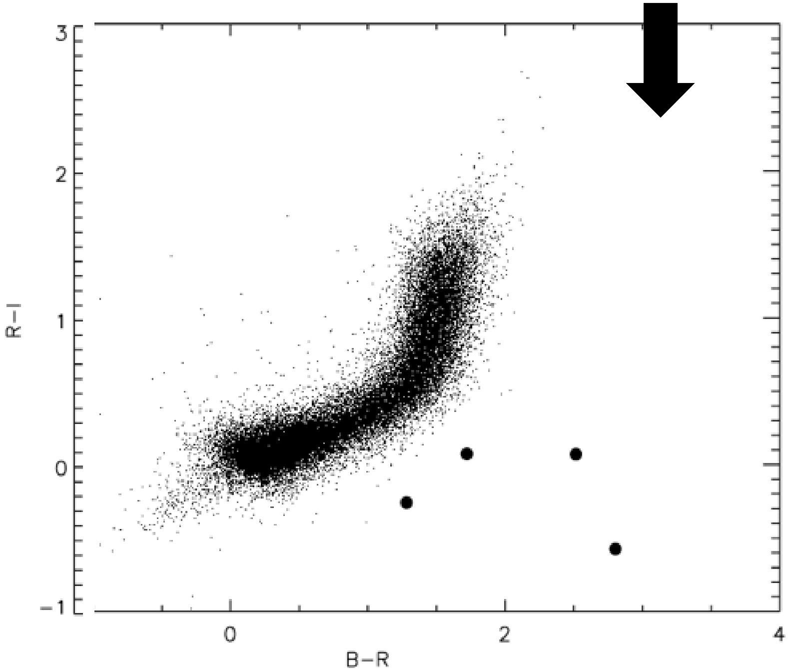


⬅ Raw multi-λ data processed into calibrated images

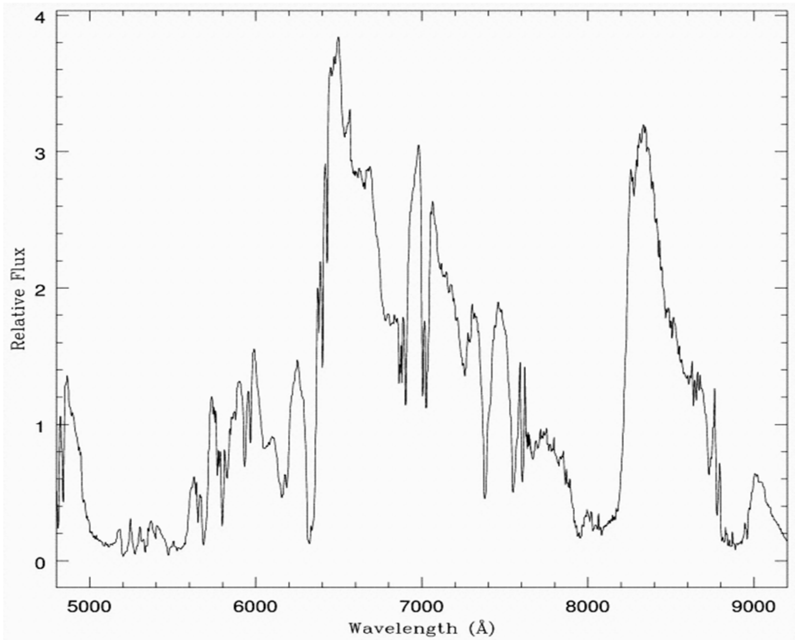
Detect sources and measure their attributes (brightness, position, shapes, etc.) ➡



cross-correlate the data, select interesting targets



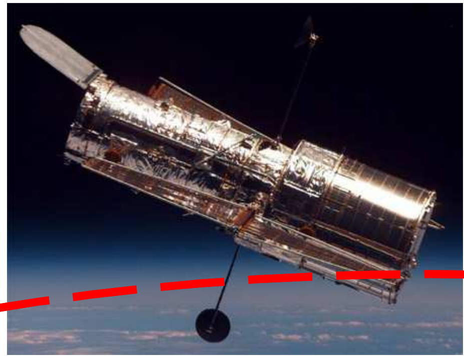
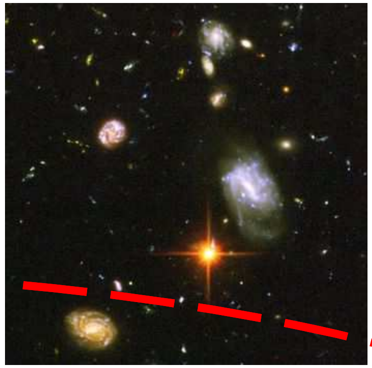
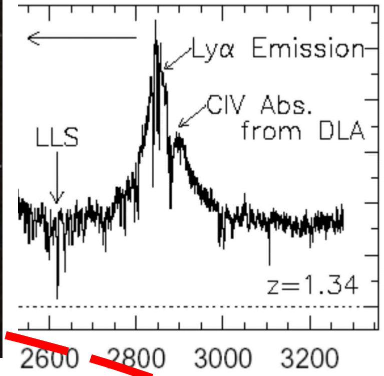
Obtain follow-up spectra ➡



➡ **Physical interpretation and understanding**

Different Modes of observational Astronomy

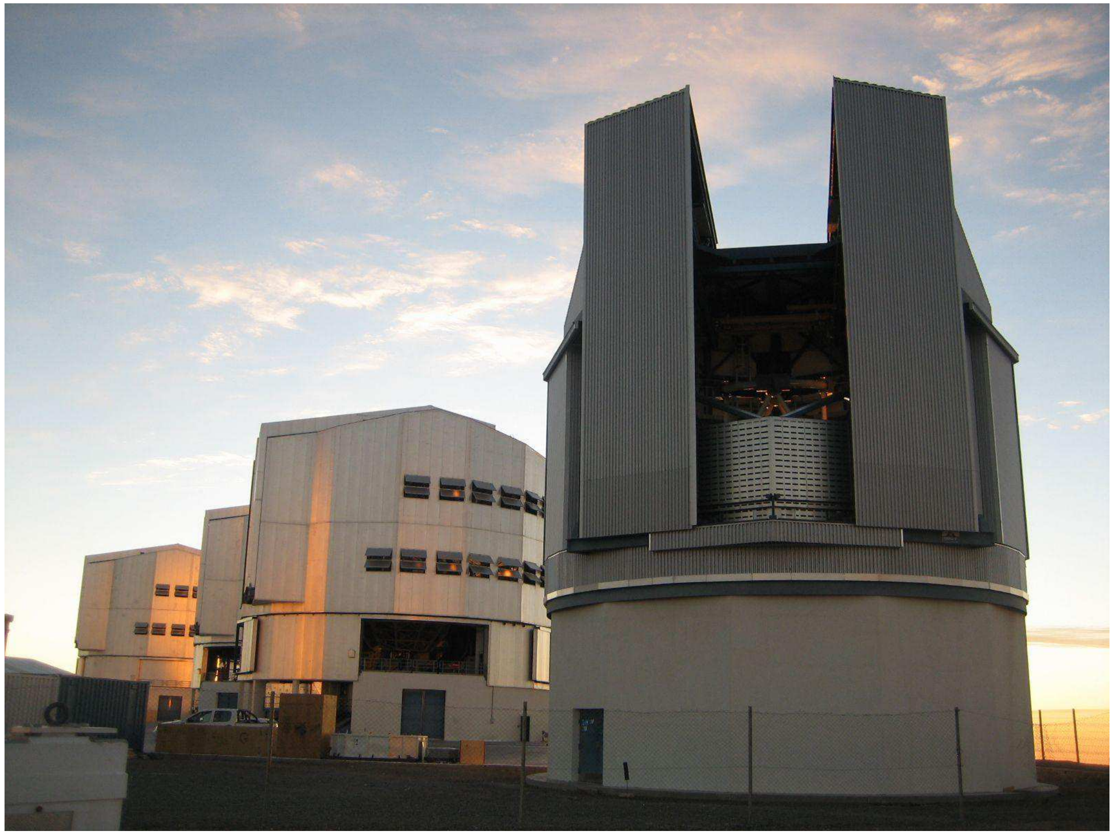
Targeted Observations of Selected Objects

Sky Surveys: A Systematic Exploration of the Sky

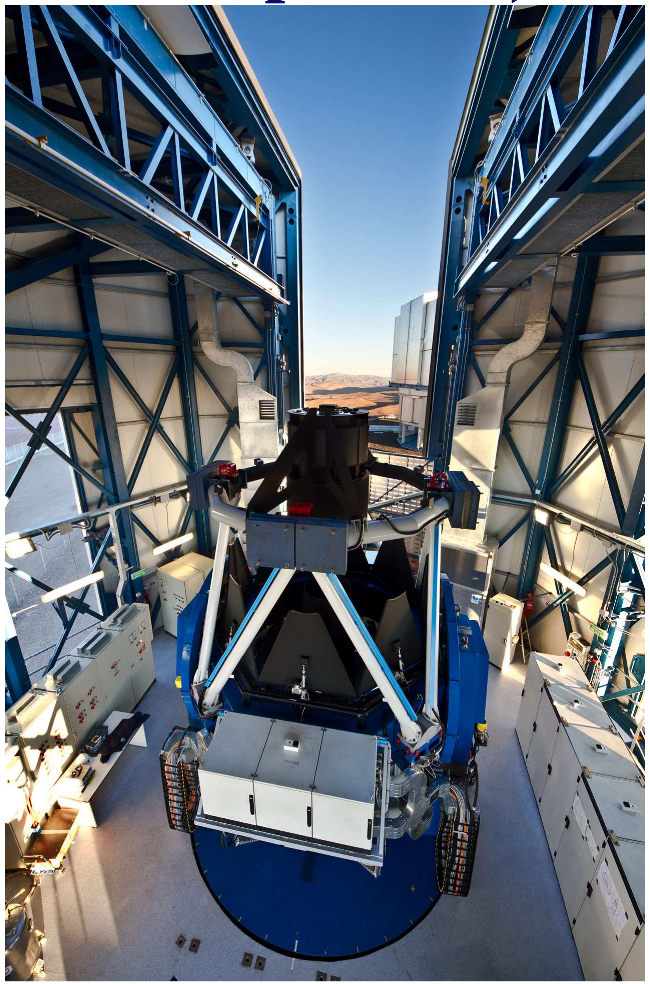


The Two Complement Each Other

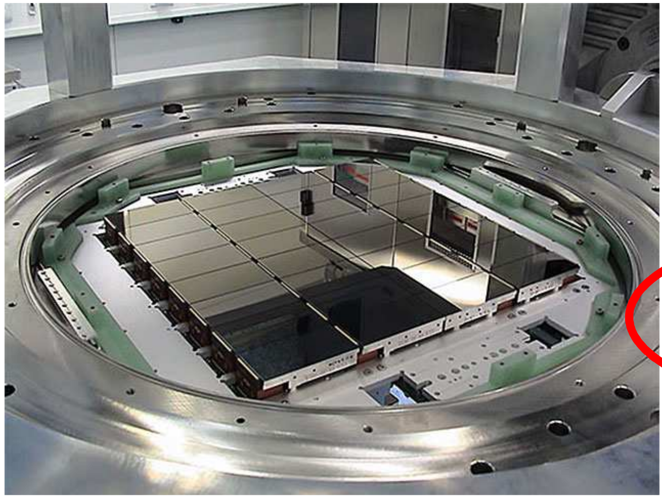


The Main Survey player

VST or VLT survey telescope (P.I: M. Capaccioli)



OMEGACAM



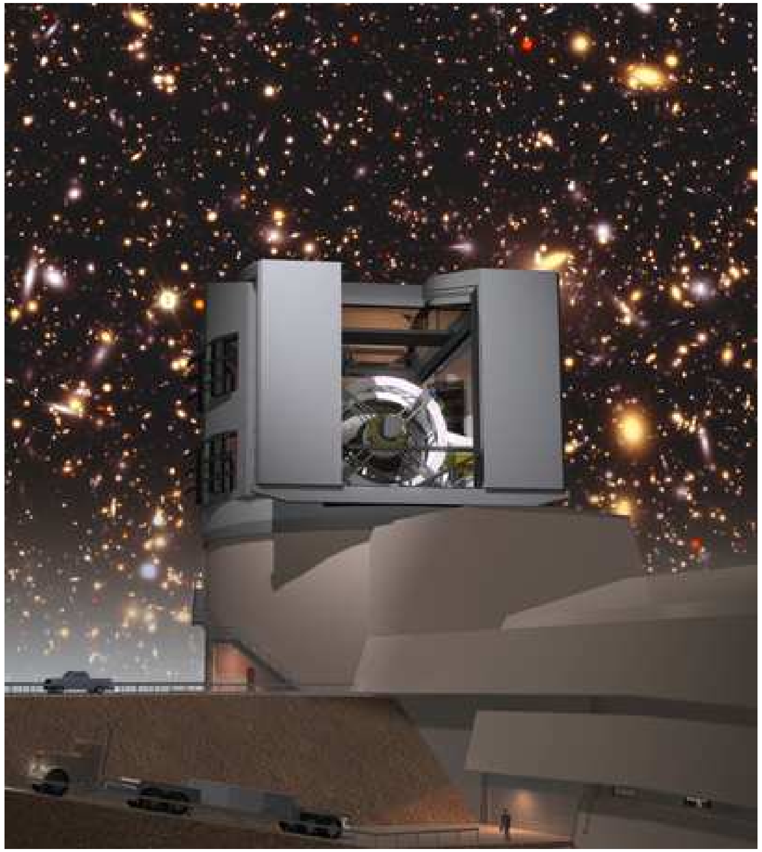
100/Gbyte/night
30 Tbyte/year

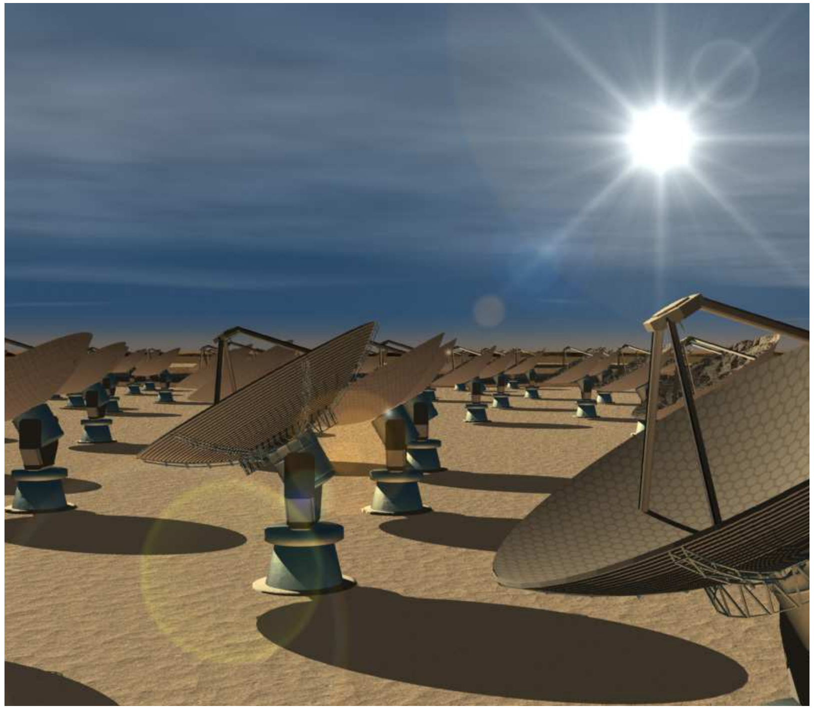
300 Tbyte TOTAL RAW
3 Pbyte TOTAL

… which is still nothing !!!!!

Large Synoptic Survey Telescope (2015)
(LSST) ~ 30 TB / night

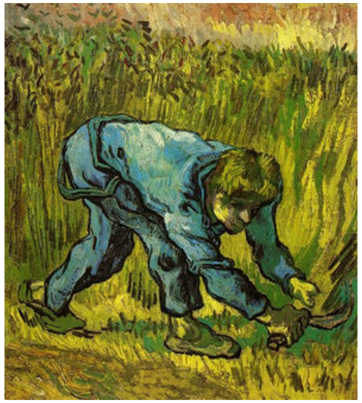
Square Kilometer Array (SKA)
~ 1 EB / second (raw data)
(EB = 1,000,000 TB) … will
require Exaflops … **????**





Why is it a revolution?

From data poverty and subsistence to an **exponential overabundance**



- For the first time in history not all data will be stored
- Most (99.99%) data will never be seen by humans
- Huge increase in data complexity (beyond current visualization capability and possibly even understanding

- Telescope+instrument likely to become “just” a front end to data systems, where the real action will be

Modern sky surveys obtain ~ 10^{12} – 10^{15} bytes of images,
catalog ~ 10^{8} – 10^{9} objects (stars, galaxies, etc.),
and measure ~ 10^{2} – 10^{3} numbers for each

- Astronomy today has ~ a few PB of archived data, and generates ~ 10 TB/day
  - Both data volumes and data rates grow exponentially, with a ***doubling time ~ 1.5 years***
  - Even more important is the growth of ***data complexity***

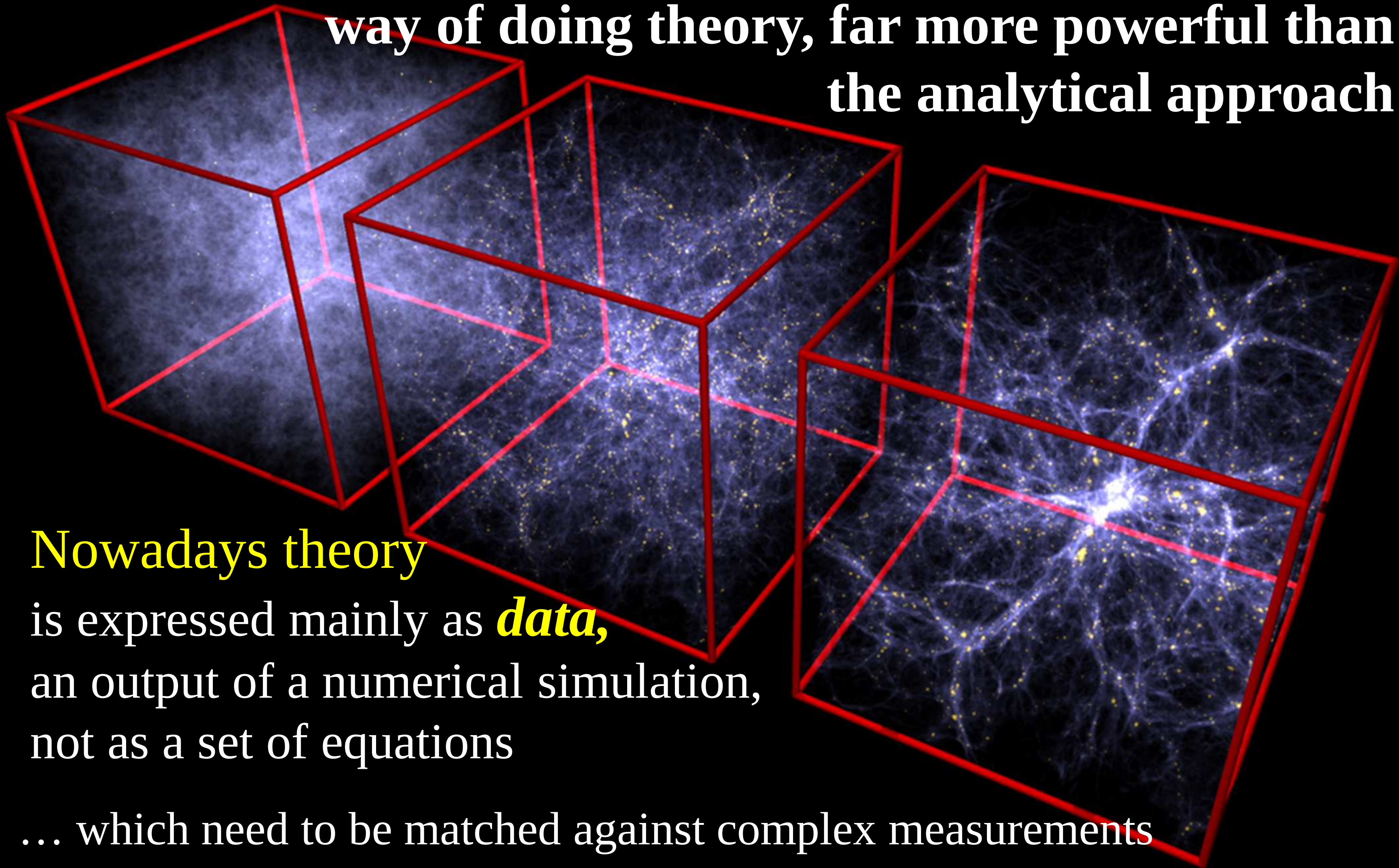
- For comparison:

  Human Genome < 1 GB; Human Memory < 1 GB (?)

  1 TB ~ 2 million books, Human Bandwidth ~ 1 TB / year (±)

Numerical Simulations:

A necessary and qualitatively different way of doing theory, far more powerful than the analytical approach



Nowadays theory
is expressed mainly as ***data,***
an output of a numerical simulation,
not as a set of equations

… which need to be matched against complex measurements

Statistical Approaches are Inevitable

- Large numbers of sources enable populations studies, e.g.,
  - Stars ⇨ Galactic structure,   Galaxies ⇨ Large-scale structure
  - Evolution of galaxies or quasars; etc.
- Imaging is relatively cheap, but spectroscopy is expensive. New indicators are substituting old ones.
  - EXAMPLE: We cannot obtain spectra of all detected sources … implies CRUCIAL need for photometry based estimators of “spectroscopic” quantities, e.g., redshifts, metallicities, star formation rates…

Large number of objects imply lower accuracy

- Comparison with theory’s predictions becomes intrinsically statistical
- Time domain astronomy (see later)

This poses huge HW and SW problems
… still largely unsolved

Hardware:

- dedicated computing infrastructures (GRID, Cloud, etc) and need for new approaches to HPC (GPU's)
- How to store, manage and transfer Pbyte from data producers to data centers and how to distribute data products to final users

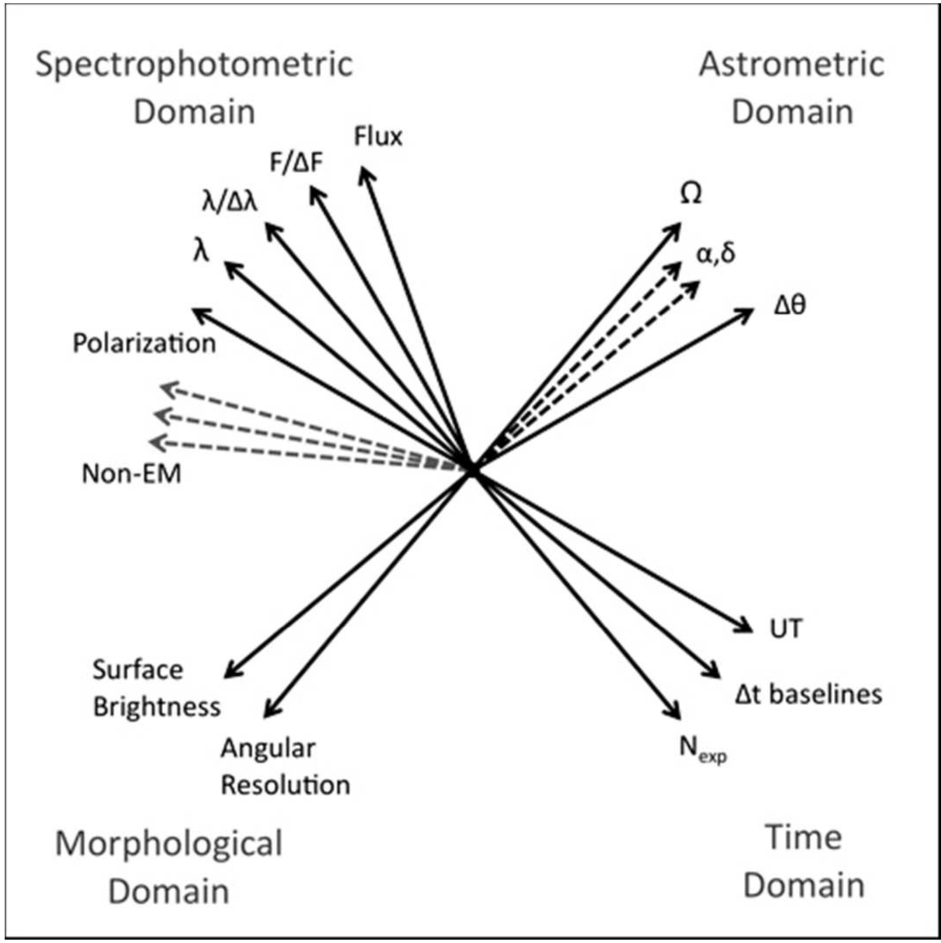
Software:

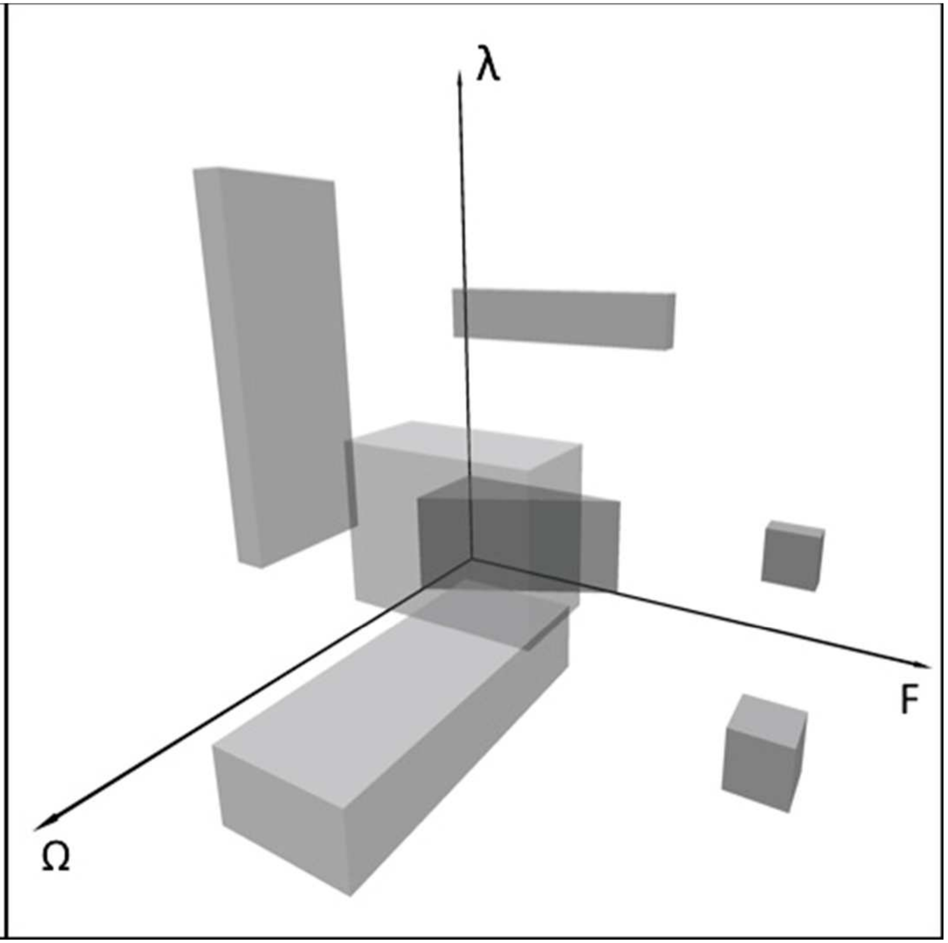
- How to visualize, compress and analyze Pbyte of heterogeneous data in a distributed data repository and computing environment

Systematic Exploration of the Observable Parameter Space (OPS)

Its axes are defined by the observable quantities

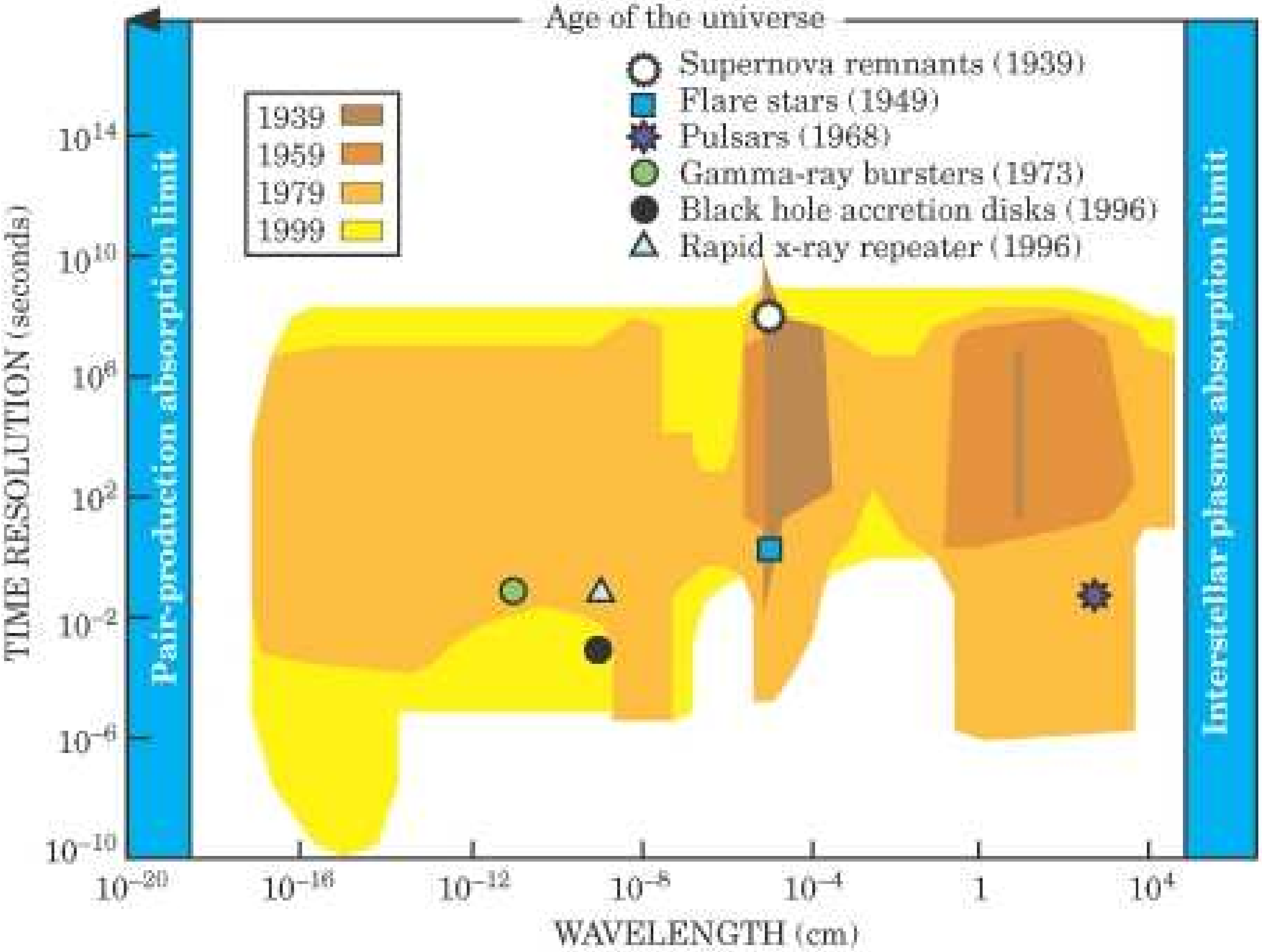
Every observation, surveys included, carves out a hypervolume in the OPS





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

Sampling of the parameter space



The Era of Cosmic Cinematography
Time Domain Astronomy

[redacted]

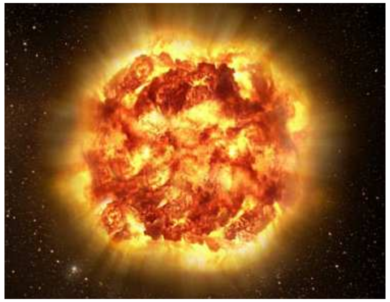
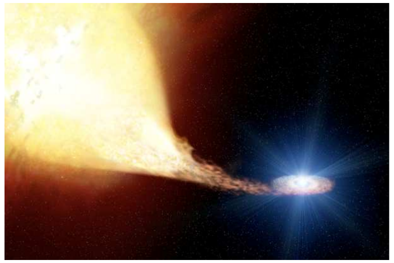
Nowadays, averything which is successfull needs to become a movie

Quote from M. Paolillo

TDA: rich phenomenology

TDA Touches essentially every field of astronomy

- Asteroids (NEO), KB Objects
- Extrasolar planets (microlensing, transits, etc…)
- Variable stars (stellar evolution)
- Better understanding of physical phenomena
- High energy phenomena (SN, GRB)
- Cosmology (AGN, Blazars, etc)


TDA: rich physical phenomenology

Physical causes of intrinsic variability:

- Evolution: structural changes etc., long time scales
- Internal processes, e.g., turbulence inside stars
- Accretion/collapse: protostars, CVs, GRBs, QSOs
- Thermonuclear explosions (SNe)
- Magnetic field reconnections, e.g., stellar flares
- Line of sight changes (rotation, jet instabilities…)

Variability is known on time scales from ms to 10^{10} yr

A broad, diverse range of interesting physics

Catching the Fireworks

- So far: data streams of **~ 0.1 TB / night, ~ 10^2 transients / night** (CRTS, PTF, various SN surveys, microlensing, etc.)
  - ✧ We are already in the regime where we *cannot follow them all*
  - ✧ Spectroscopy is the key bottleneck now, and it will get worse

- Forthcoming on a time scale ~ 1 - 5 years: **~ 1 TB / night, ~ 10^3 - 10^4 transients / night** (PanSTARRS, Skymapper, VISTA, VST, SKA precursors…)
- Forthcoming in ~ 8 – 10 (?) years: LSST, **~ 30 TB / night, ~ 10^5 - 10^7 transients / night**, SKA

A major, qualitative change!

- So… which ones will you follow up?

- Follow-up resources will likely remain limited

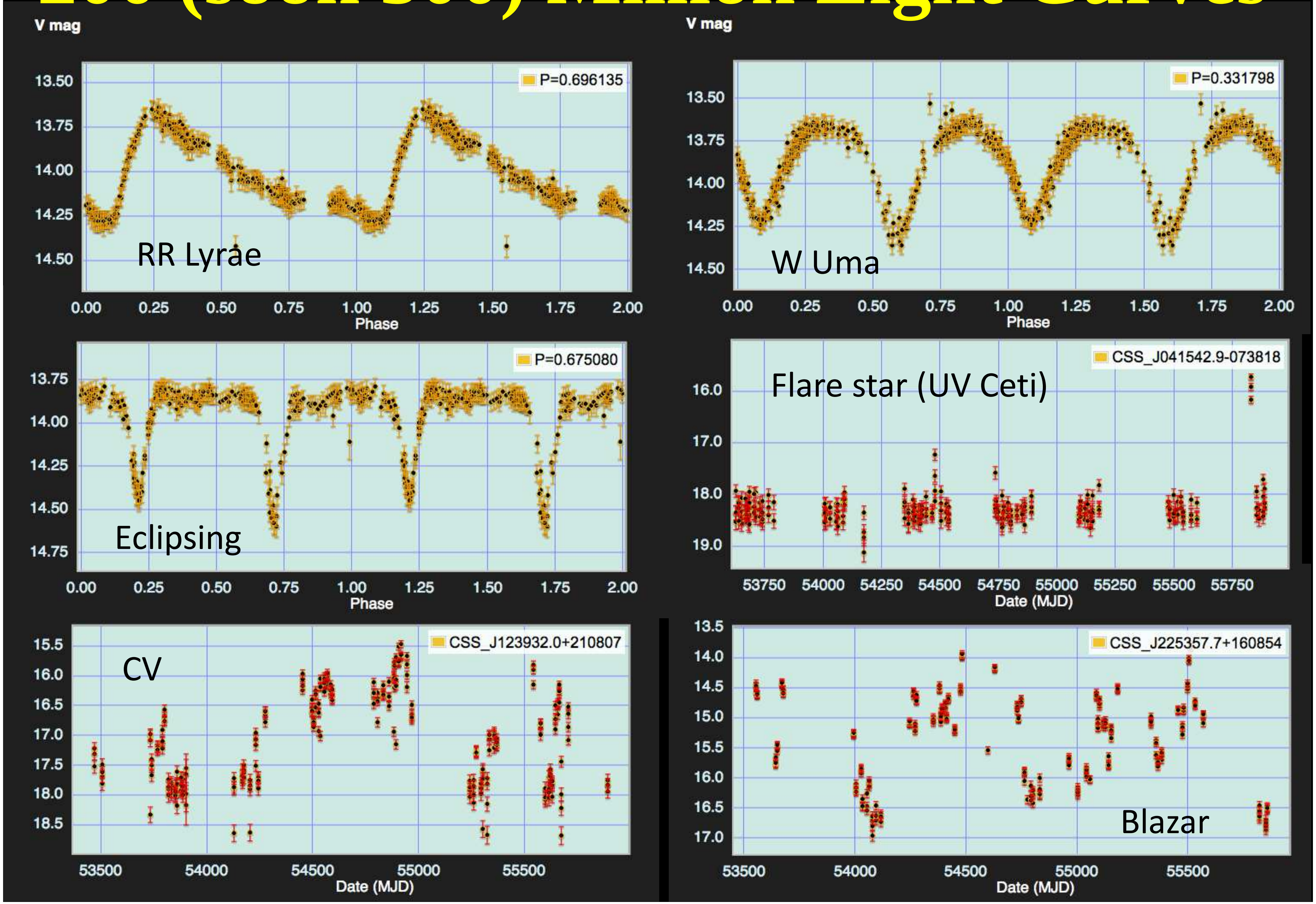
Transient classification is essential

The Catalina Real-Time Transient Survey (CRTS) *(co-PI. G. Djorgovski)*

http://crts.caltech.edu

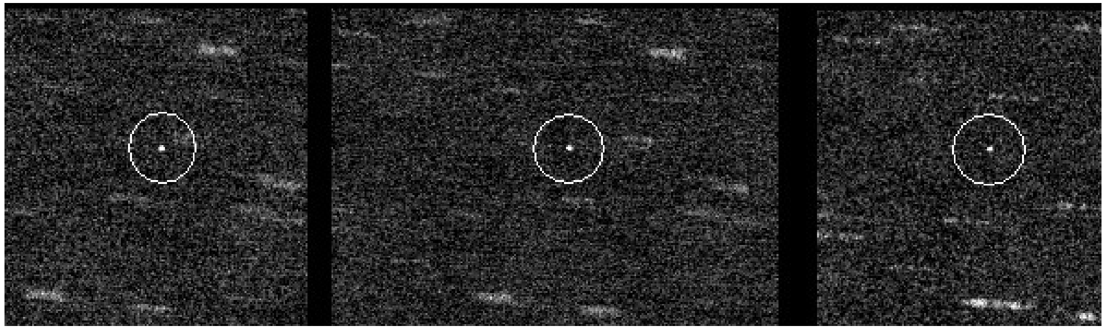
- Collaboration with a search for near-Earth asteroids at UA/LPL; we discover astrophysical transients in their data stream
- 3 telescopes in AZ, Australia
- About 6,000 unique, strong transients to date, including > 1,500 supernovae, > 1,000 CVs, ~ 2,000 flaring blazars/AGN, etc.
- > 80% of the sky covered ~ 300 – 500 times over ~ 7+ years
- **Open data policy: *all data are made public immediately***

200 (soon 500) Million Light Curves

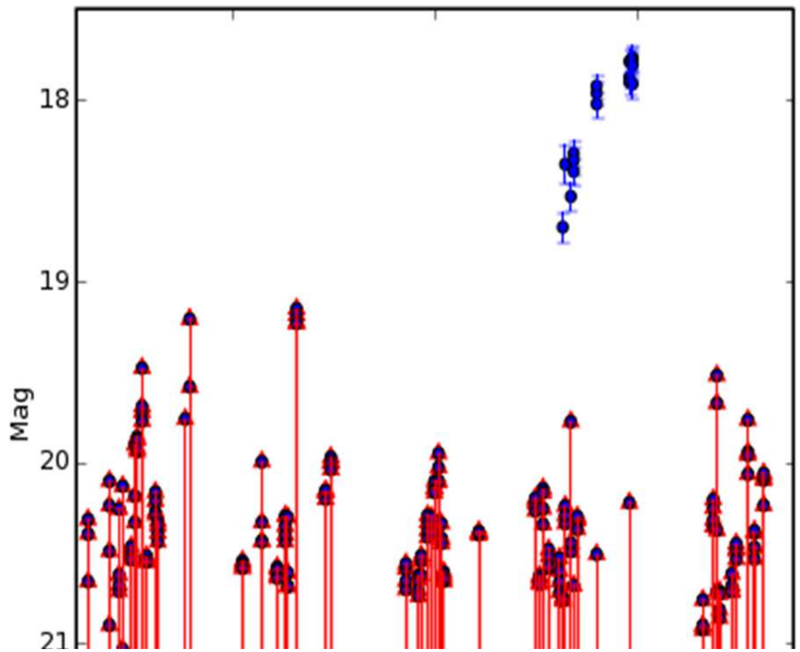


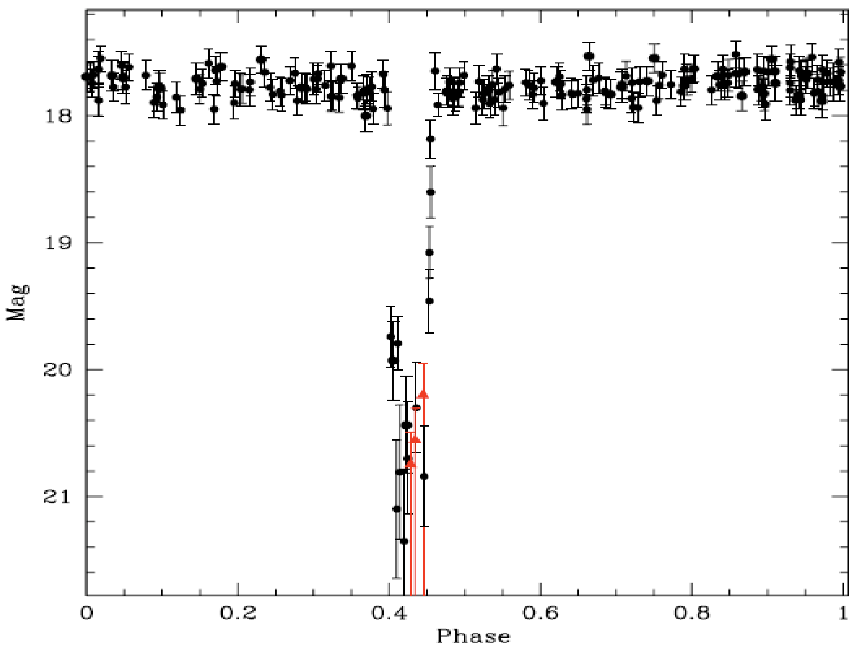
Scores of dwarf novae, blazars, etc.

CSS Discoveries of Earth-Grazing Asteroids

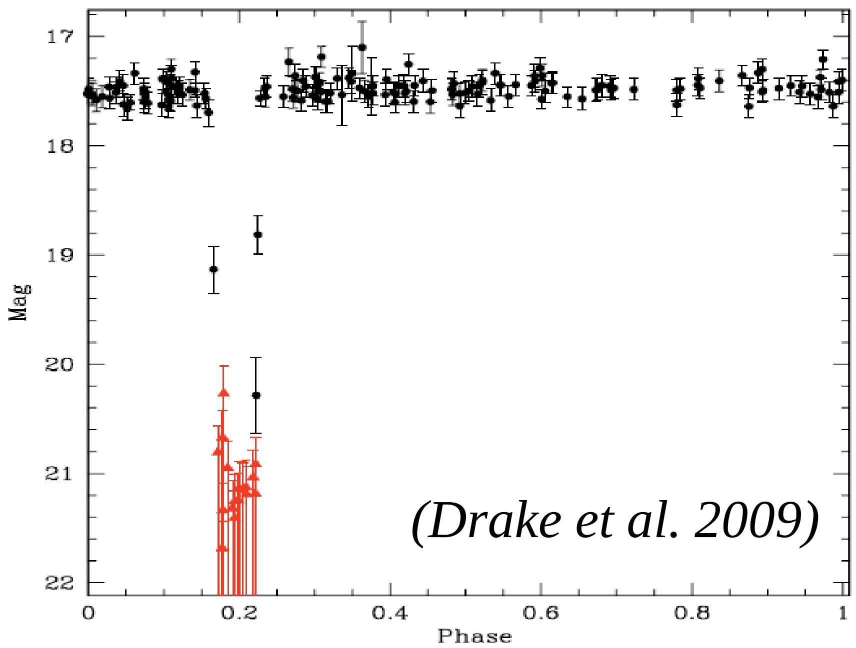


The slowest SN ever





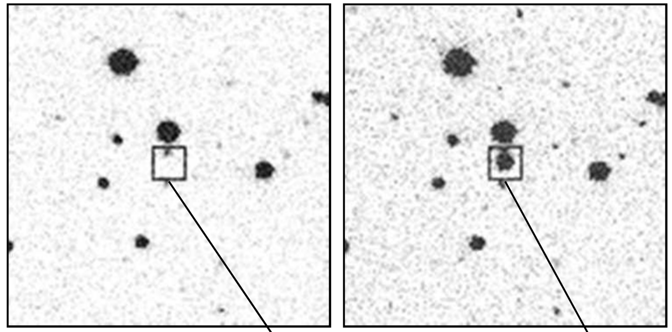
18
19
20
21
Mag
0
0.2
0.4
0.6
0.8
1
Phase



17
18
19
20
21
22
Mag
0
0.2
0.4
0.6
0.8
1
Phase
(Drake et al. 2009)

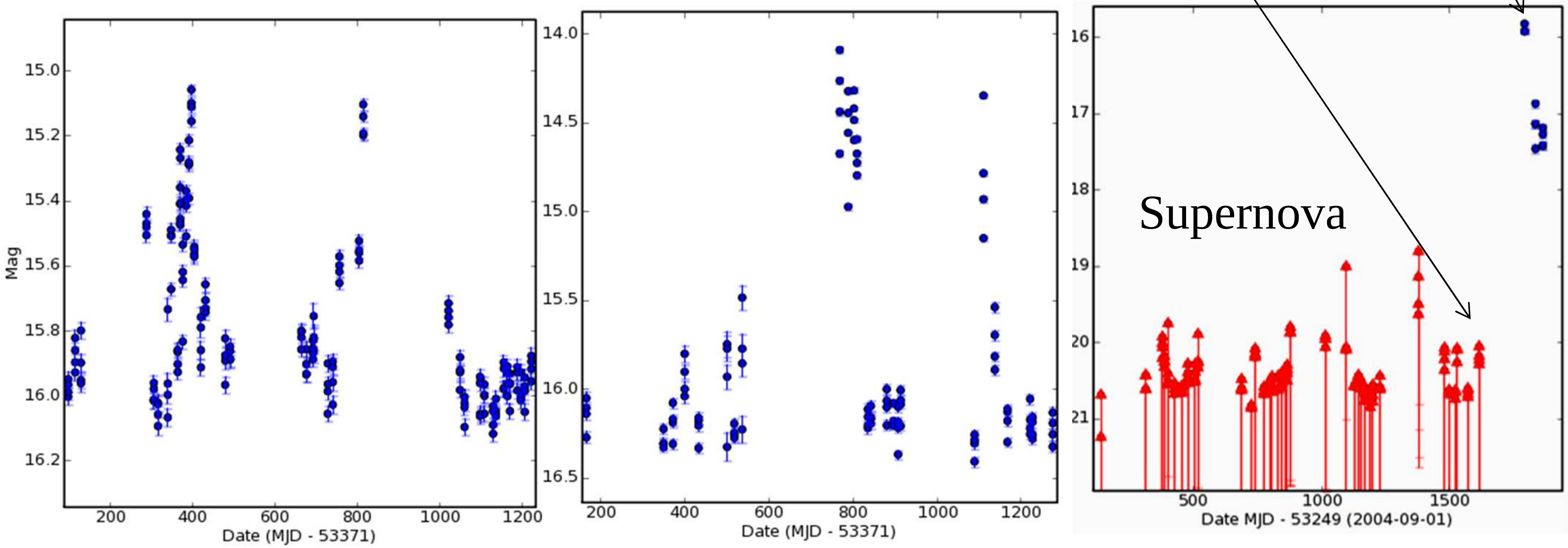
Exoplanets

Sample Light Curves



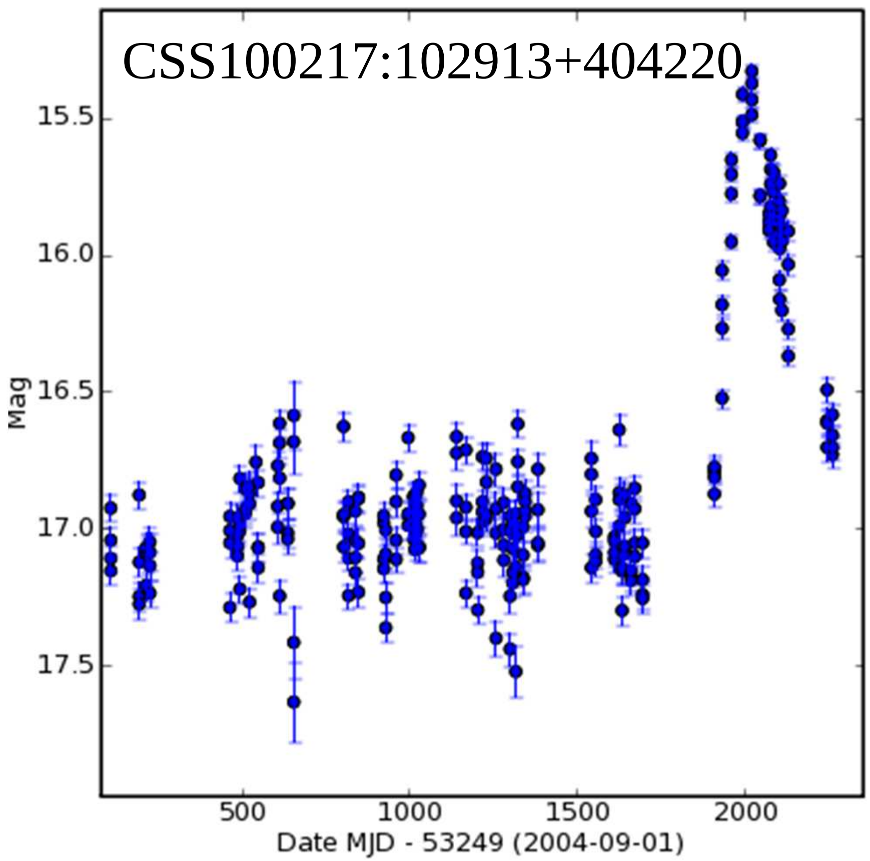
Blazar PKS0823+033

CV 111545+425822

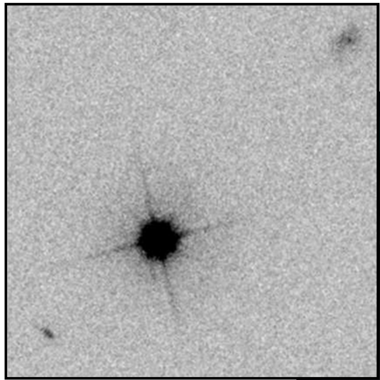


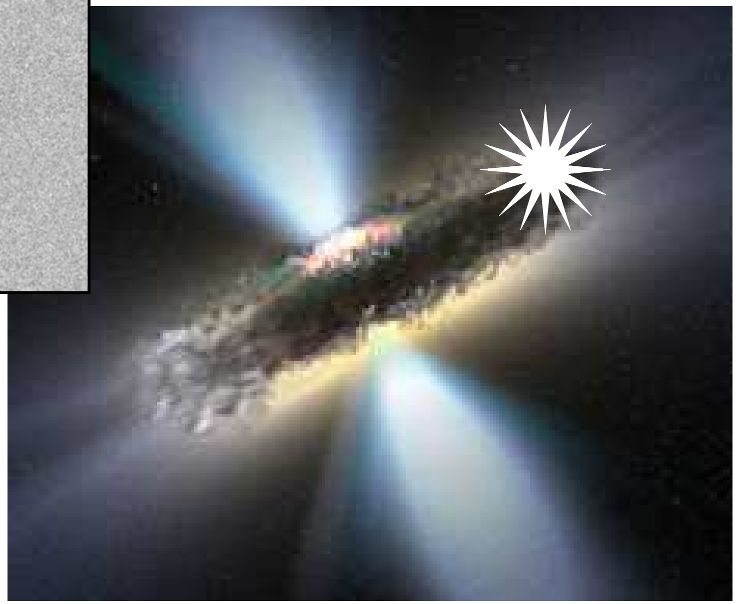
We produce light curves for every detected source in the survey ($> 5 \times 10^8$ sources), and make them publicly available. They are generated automatically for transient sources, blazars, etc. This is an unprecedented data set for time domain astronomy.

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

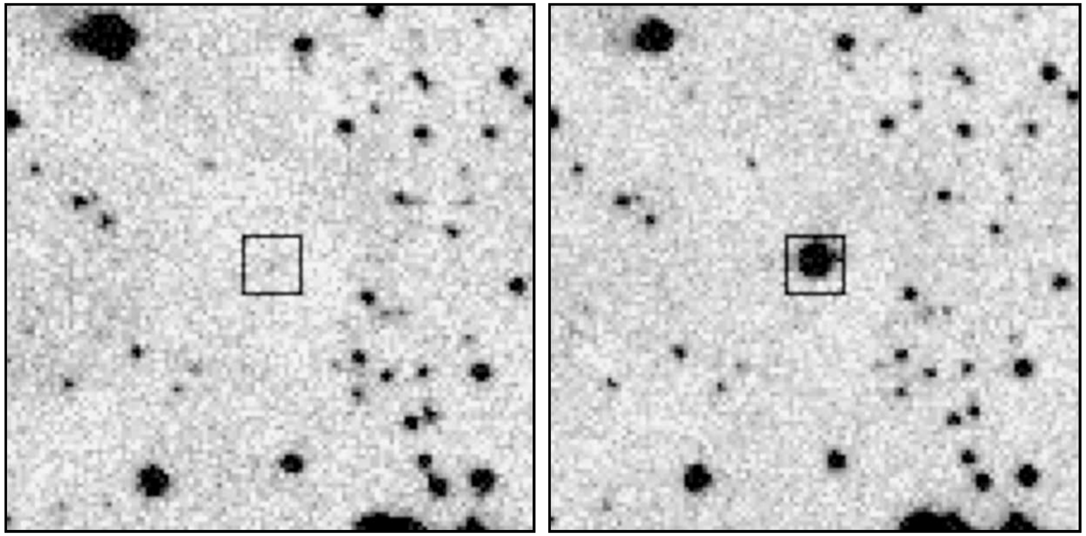




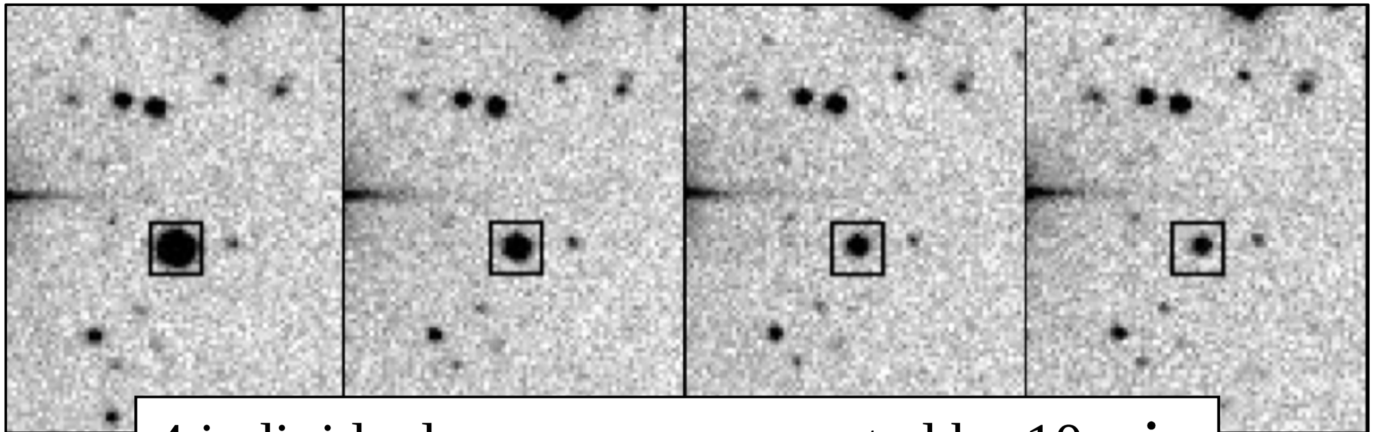
The first case of a Supernova from an Active Galactic Nucleus accretion disk? (Predicted by theory, but never seen before)

Unsettled Stars

Newborn stars, FU Ori objects

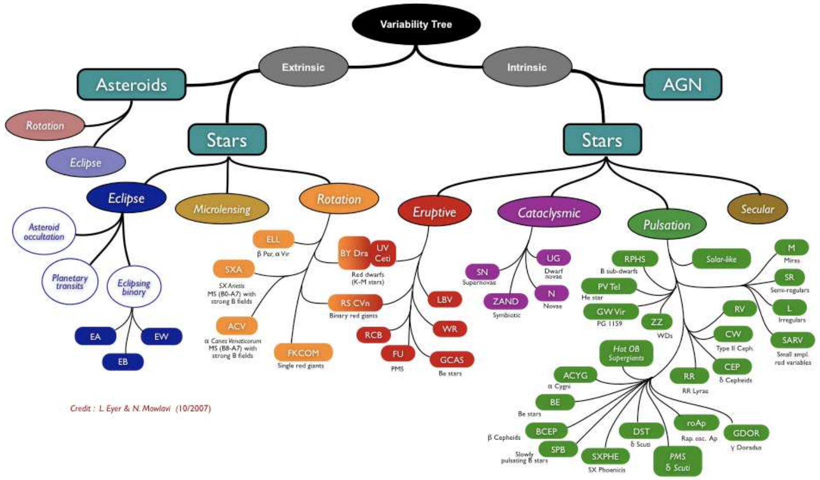


Fast transients (flaring dwarf stars)



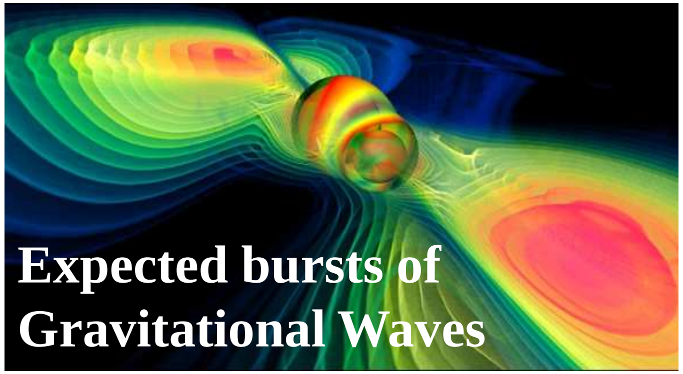
4 individual exposures, separated by 10 **min**

Two different types of problems



Offline TDA.

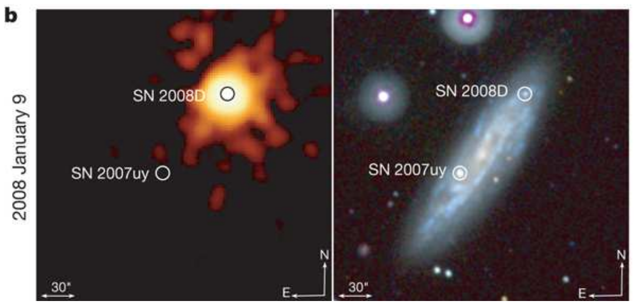
Understanding the variable universe on offline huge amounts of light curves produced by modern surveys



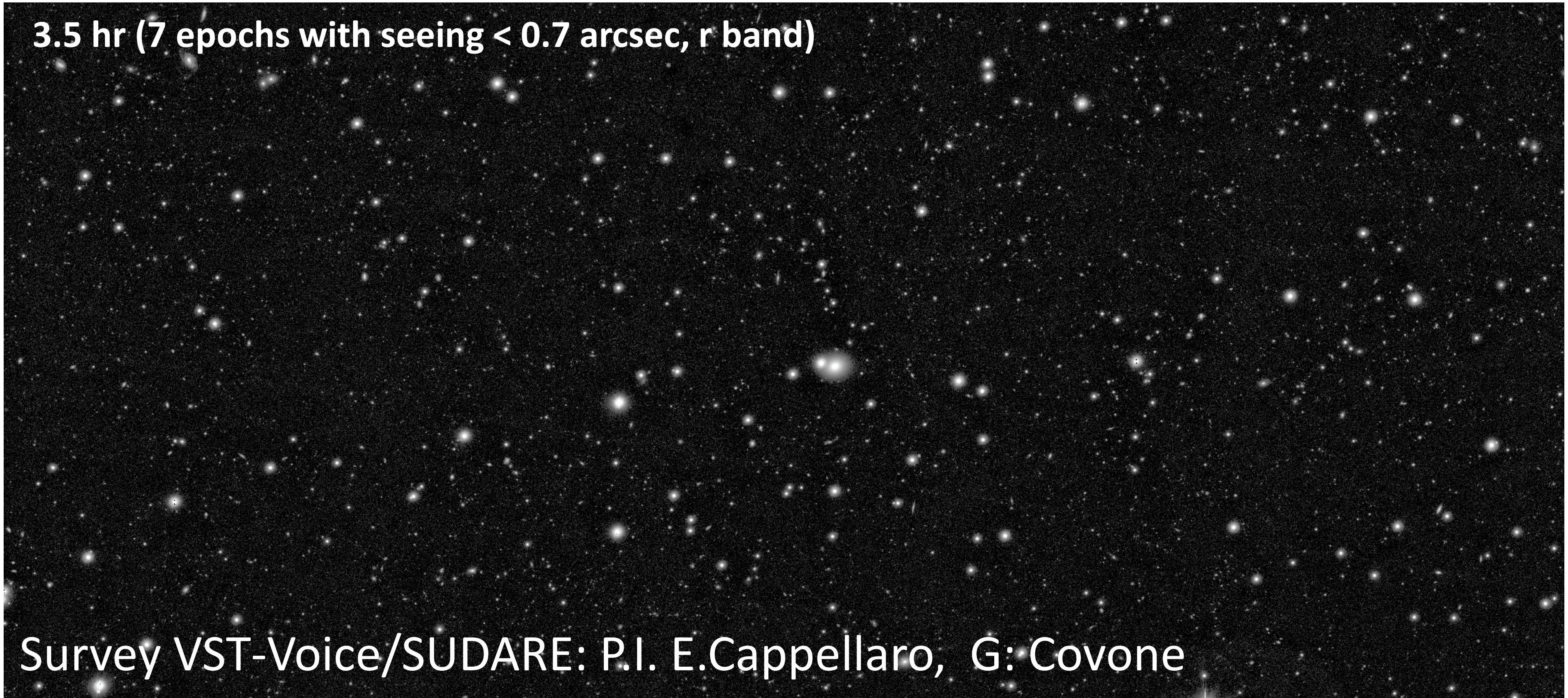
Online TDA.

Detecting and characterizing in real time photometric transients transients («things which go «BOOOM!»)

Tidal disruption flares

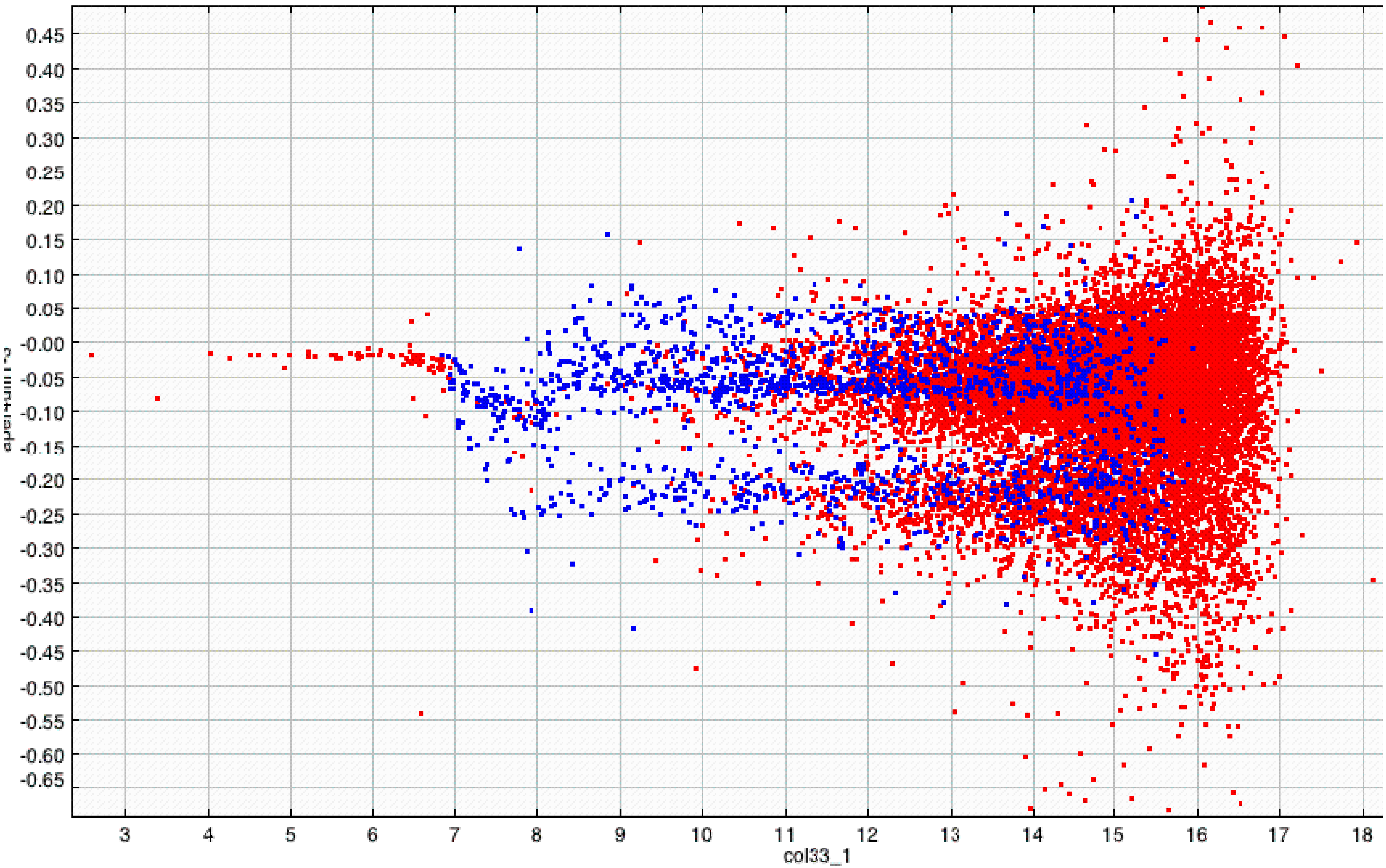


Supernova breakout shocks

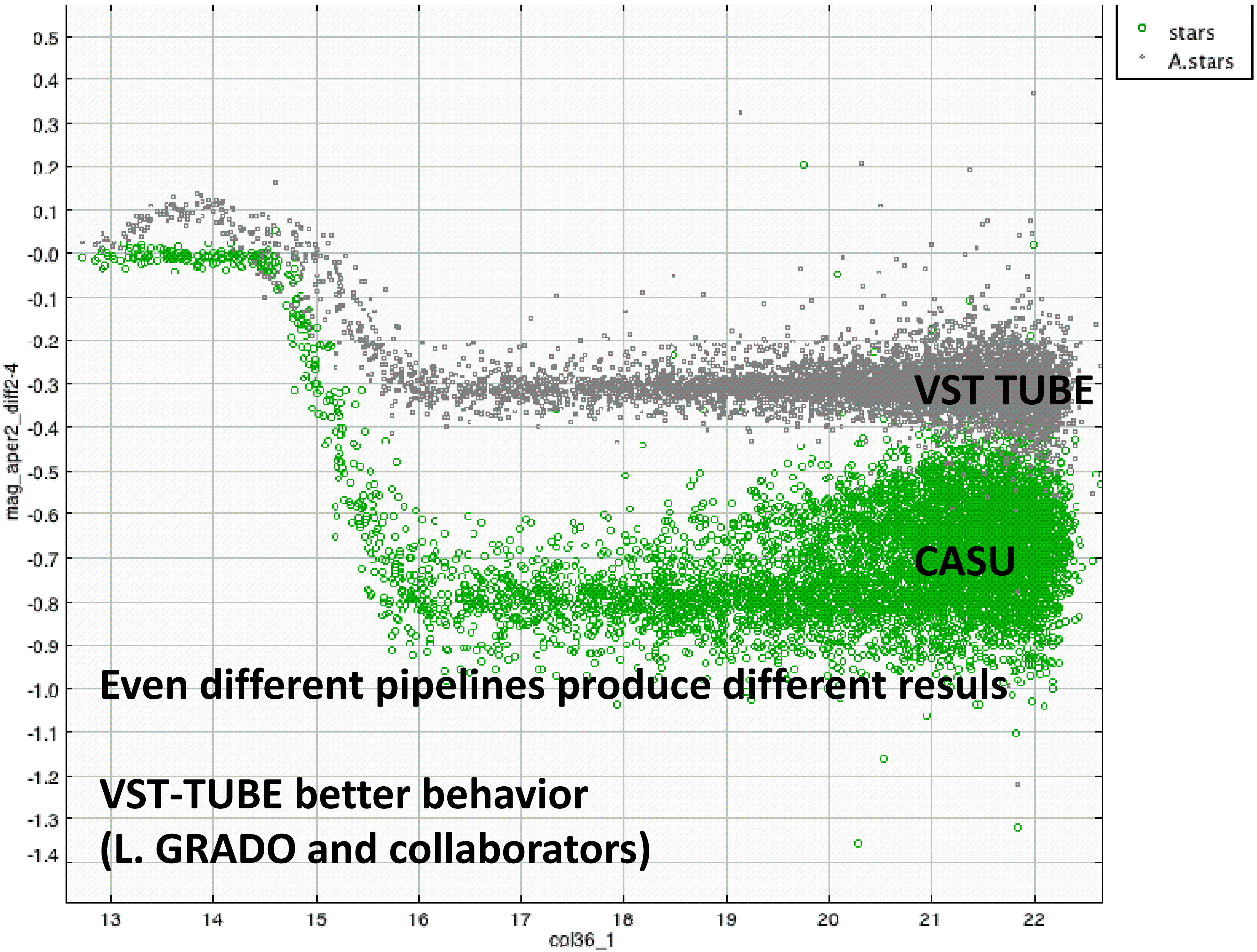


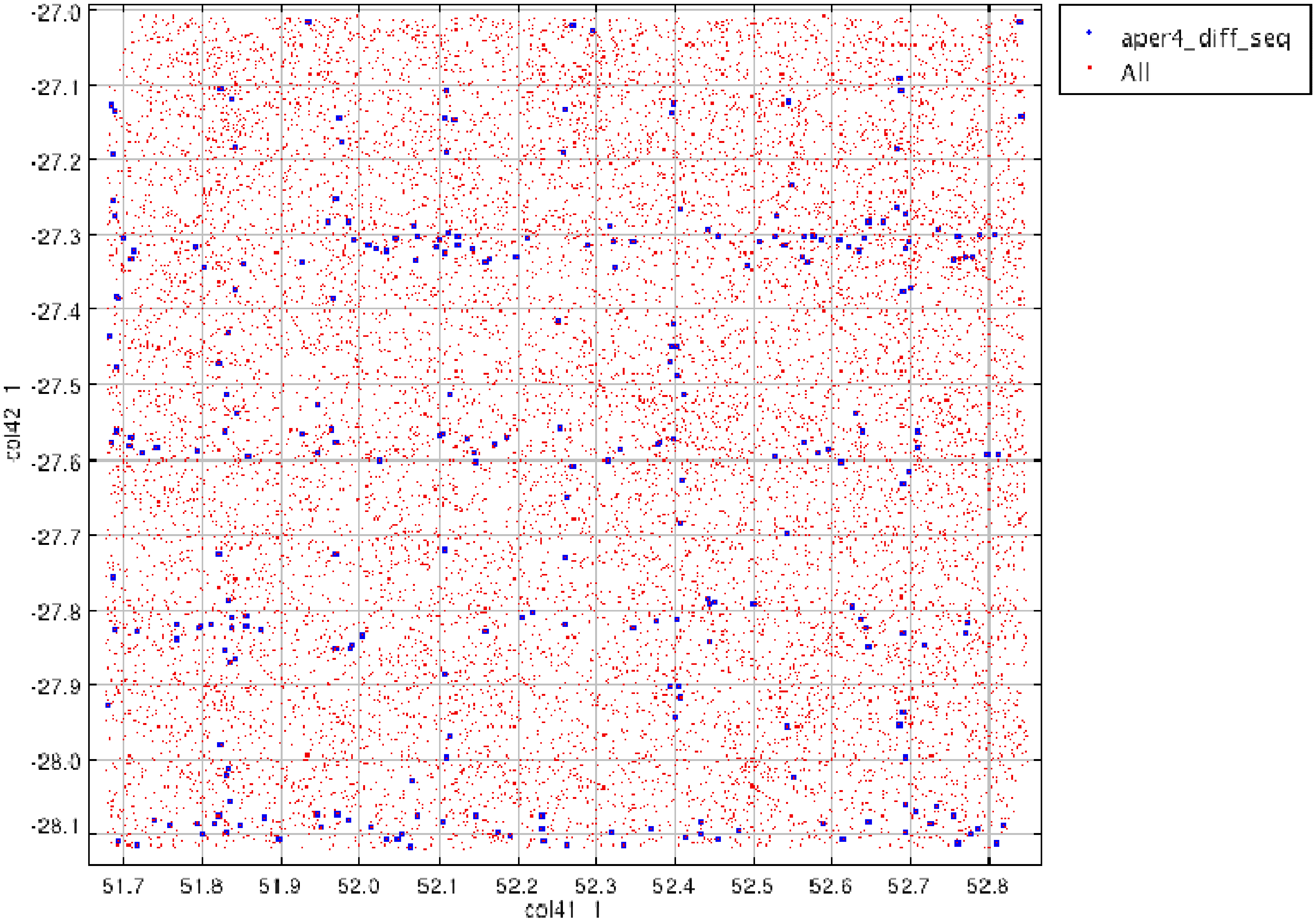
VOICE/SUDARE cover 3 fields: CDFS (4 sq deg), COSMOS (1 sq deg), ELAIS South 1 (4 sq deg)
Started in October 2011

r band r: ca 60 epochs e, exp time 0.5 hr, sampling every 2-3 nights
g,i bands : sampling 5-6 nights, 28-30 epochs



EVEN THE DETECTION OF VARIABLE OBJECTS IS NOT AN EASY TASK


stars
A.stars
VST TUBE
CASU
Even different pipelines produce different resuls
VST-TUBE better behavior
(L. GRADO and collaborators)
mag_aper2_diff2-4
col36_1
0.5
0.4
0.3
0.2
0.1
-0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.3
-1.4
13
14
15
16
17
18
19
20
21
22

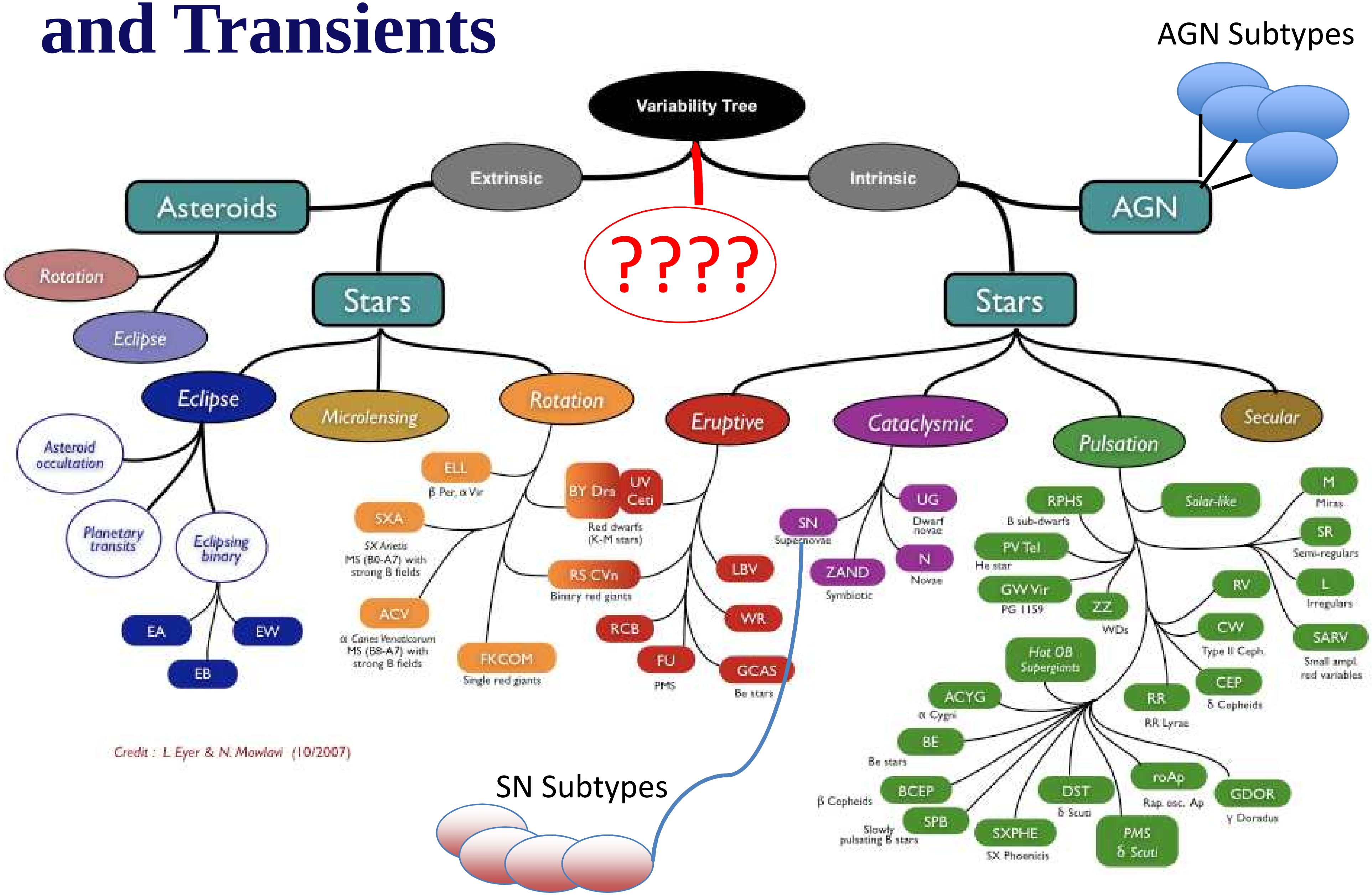

aper4_diff_seq
All
col42 1
col41 1
-27.0
-27.1
-27.2
-27.3
-27.4
-27.5
-27.6
-27.7
-27.8
-27.9
-28.0
-28.1
51.7
51.8
51.9
52.0
52.1
52.2
52.3
52.4
52.5
52.6
52.7
52.8

Stay Tuned…..

Off Line -TDA

- At the moment ca. 250 million light curves are available.
- Less than 5% have been classified
- Less than 40% of the variable objects in the MACHO survey have been identified.
- Most of the micro-lensing phenomena n the MACHO survey were not identified
- In the Harvard light curve collection (3x106 objects) more than 60% are of unknown type and a large **fraction is impossible to fit into any known cataegory…**

Semantic Tree of Astronomical Variables and Transients



Why Is This Hard?

("Not your grandma's classification problem")

1. ***Data are sparse and heterogeneous***
   - Different measurements for different events, random sampling, variable data quality, archival coverage…
   - Feature vector methodology generally does not work
   - Most of the initial information is archival and/or contextual
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 follow the most interesting/valuable events*
   - *Decide on the optimal follow-up resource use*

5. *Could be also computationally resource limited (processing power, bandwidth, etc.)*
6. *Huge and growing data volumes/rates*

   ➞ ***Must take the humans out of the loop!***

7. *Must be scalable to more and different data inputs*

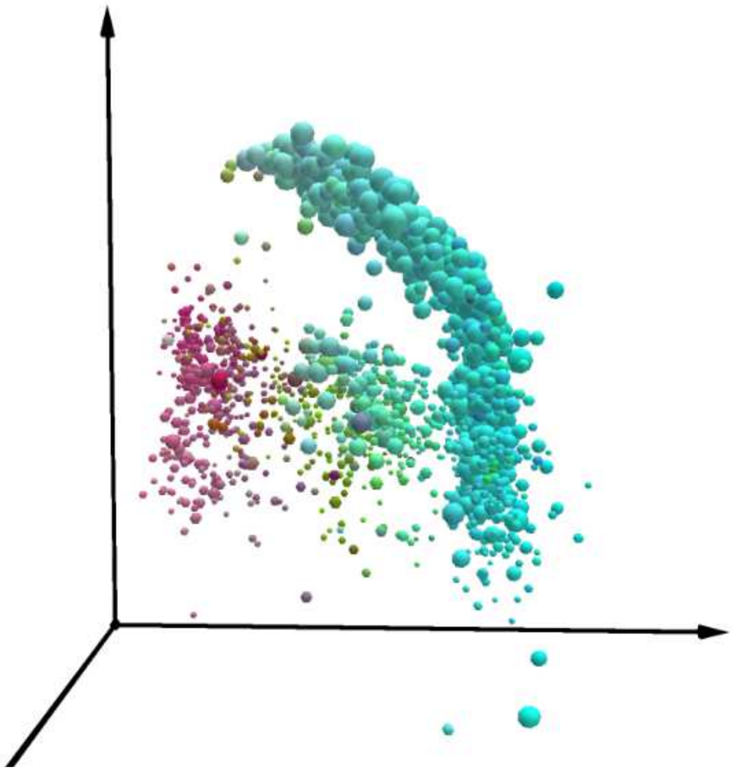
Problems

- How to characterize light curves (KISS WG)?
- Knowledge base built on the data themselves (CRTS) or rather on simulated ones (STRADIWA) ?
- How to solve the computational challenge (DAMEWARE)?
- How to find the unknown? Throwing away all the known (classification) or searching for intrinsic partitions of the OP (clustering)?
- How to do this real time (on-line TDA… not addressed here)

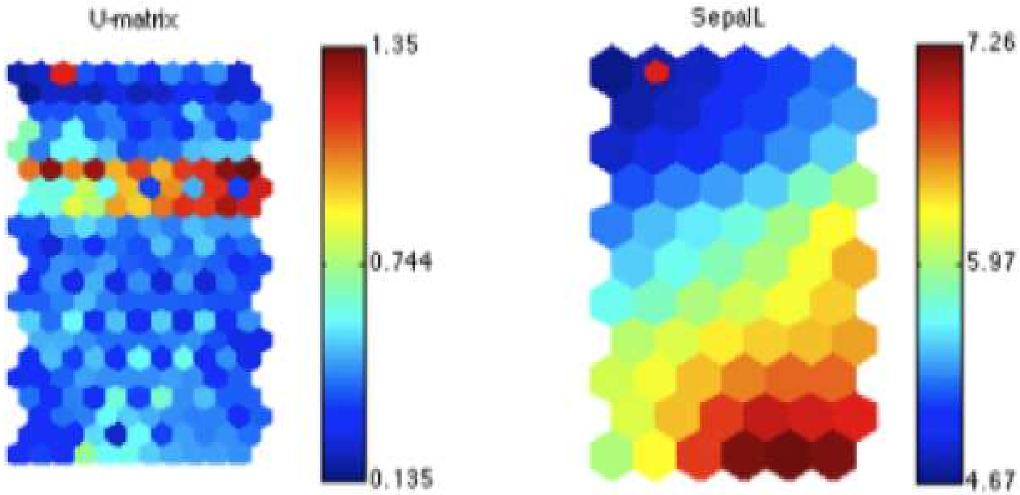
Unsupervised Clustering in Feature Space

(Lead: Ciro Donalek - Caltech)

- Unsupervised Machine Learning
- Can be used to determine the number of classes and cluster the input data in classes on the basis of their statistical properties only
- Search for Outliers, Trajectories, etc.
- Methods: SOM, K-means, Hierarchical Clustering, etc.



- Given a set of features, which ones are the most discriminating between different classes?

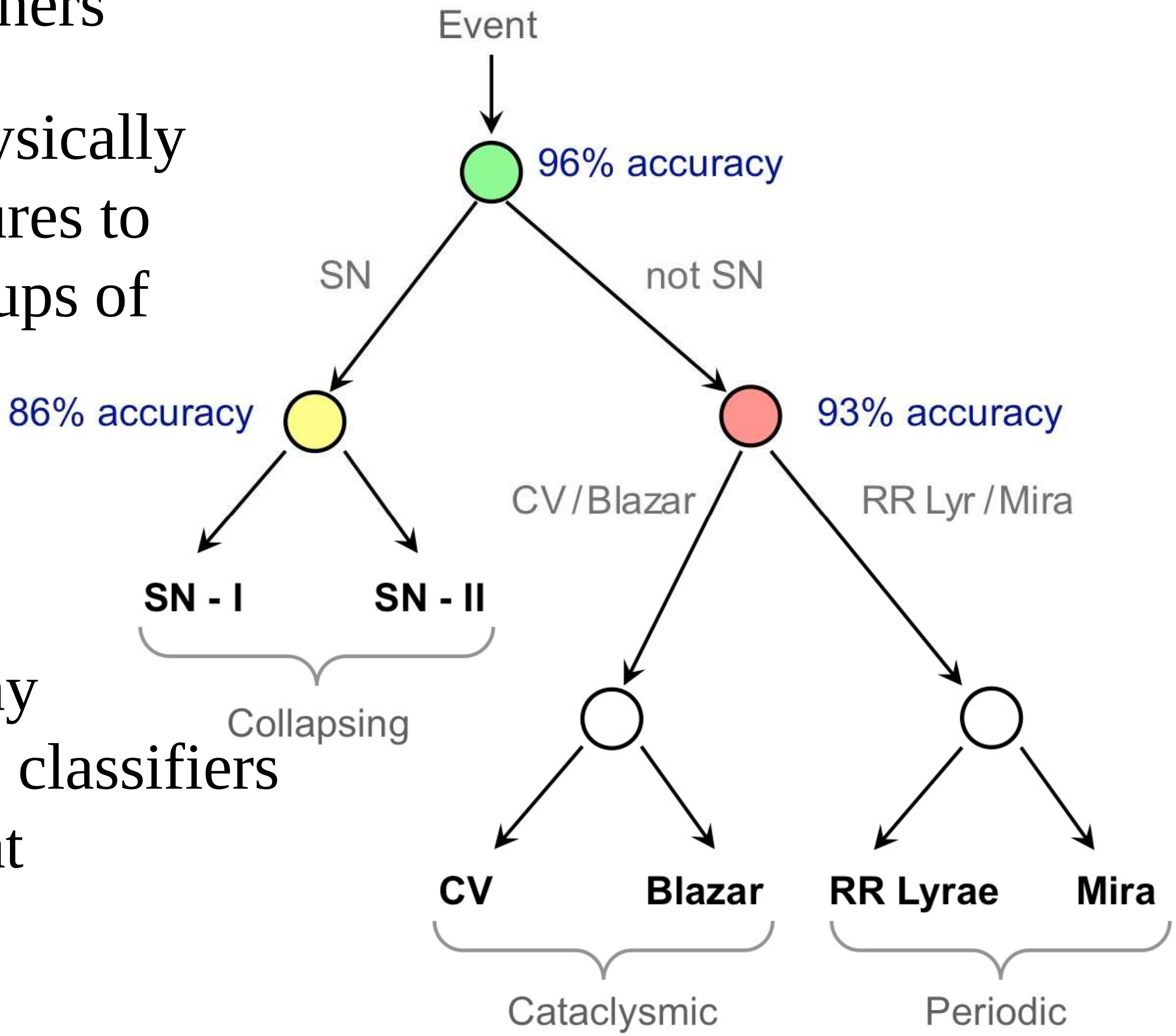


A Hierarchical Approach to Classification

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

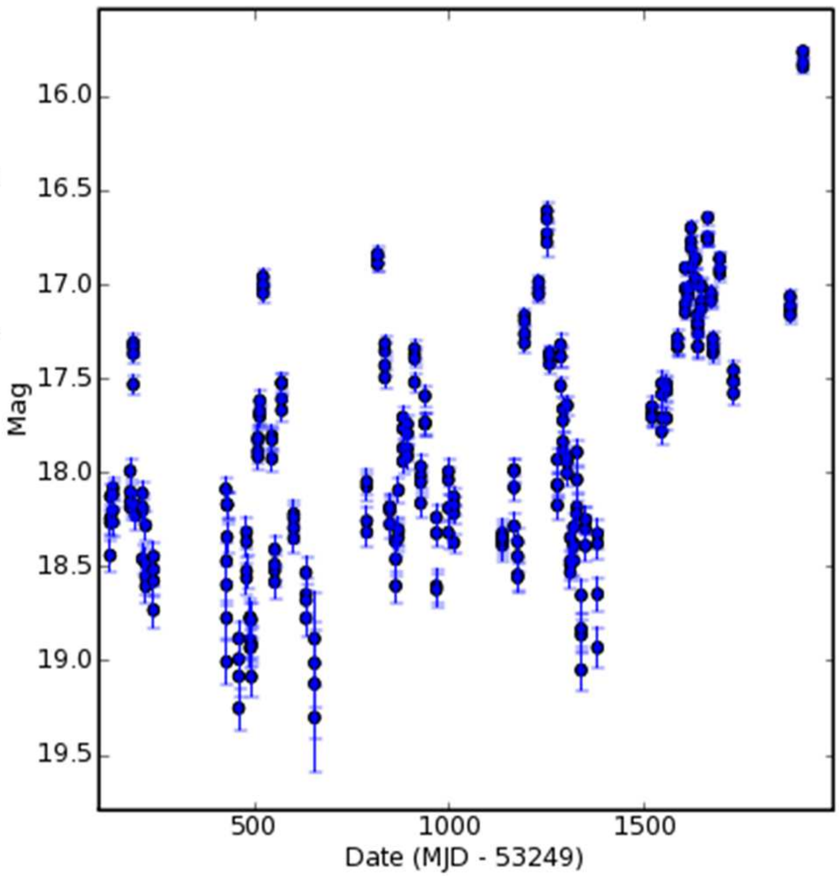
We use some astrophysically motivated major features to separate different groups of classes

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



LC Feature Vectors

- Light curves have different numbers of points and different sampling
- Compute a set of parameters for any given light curve; they form feature vectors (Richards et al. 2011)
- Apply various classifiers: random forests, SVM, ANN, SOM…
  - Also experimenting with *Eureqa (M. Graham, CARC)*



Feature	Description
amplitude	Half the difference between the maximum and the
beyond1std	Percentage of points beyond one st. dev. from the
flux_percentile_ratio_mid20	Ratio of flux percentiles (60th - 40th) over (95th -
flux_percentile_ratio_mid35	Ratio of flux percentiles (67.5th - 32.5th) over (95t
flux_percentile_ratio_mid50	Ratio of flux percentiles (75th - 25th) over (95th -
flux_percentile_ratio_mid65	Ratio of flux percentiles (82.5th - 17.5th) over (95t
flux_percentile_ratio_mid80	Ratio of flux percentiles (90th - 10th) over (95th -
linear_trend	linear fit to the light curve fluxes
max_slope	absolute flux slope between two consecu

… etc. ~ 60 features total

Automated Classification of Variable Stars

Dubath et al. (2011):

Used **random forests** on a set of 14 light curve features to recover 26 classes of variable stars from the *Hipparcos* catalog

Confusion matrix ➪

Predicted Class

True Class	EA	EB	EW	ELL	LPV	RV	CWA	CWB	DCEP	DCEPS	CEP(B)	RRAB	RRC	GDOR	DSCT	DSCTC	BCEP	SPB	BE+GCAS	ACYG	ACV	SXARI	BY+RS
EA	214	13									1												
EB	19	191	28	2	1				2					1		4		3		2	2		
EW		30	76							1													
ELL		14			1									1		1		3			5		2
LPV					285																		
RV		1			1				2	1													
CWA		2				1			5														1
CWB		1						2	2	1													
DCEP									183	5	1												
DCEPS		1							11	17													2
CEP(B)		1							4		6												
RRAB		1										69	1					1					
RRC		2	4									1	12		1								
GDOR														27									
DSCT		1	1									1			32	12							
DSCTC		1													1	77					2		
BCEP		1	1													1	26	1					
SPB				1													1	74		1	4		
BE+GCAS	1									1								5		2	4		
ACYG		1																	1	13	2		1
ACV		3								1				1				6			66		
SXARI		2																2			3		
BY+RS		1							1														33

Similar results by the Berkeley group (Richards et al. 2011)

Bayesian Networks Implementation

(C. Donalek lead, Caltech)

X = input measurements of individual kinds (e.g., mags, colors, etc.)
Y = classes of events, $Y = 1, \ldots k$

Then:

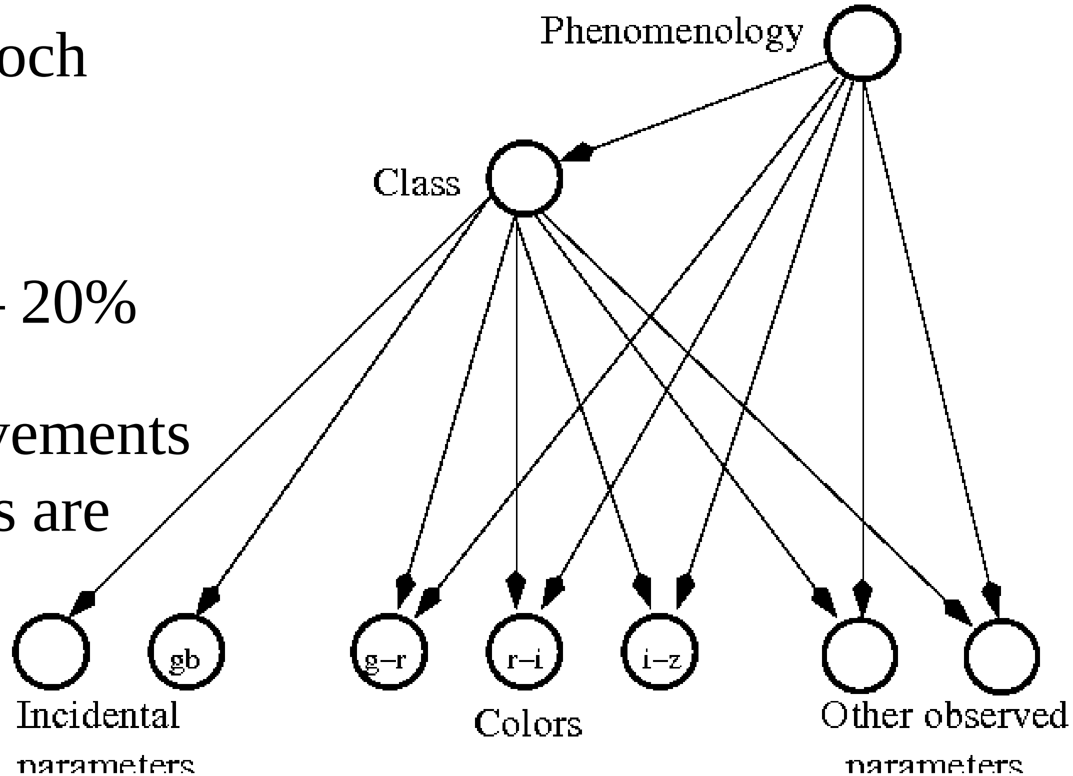
$$P(y=k \mid x) = P(x \mid y=k)P(k)/P(x) \propto$$

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

Initial results using single-epoch color measurements:

Typical accuracy ~ 80%
Typical contamination ~ 15 – 20%

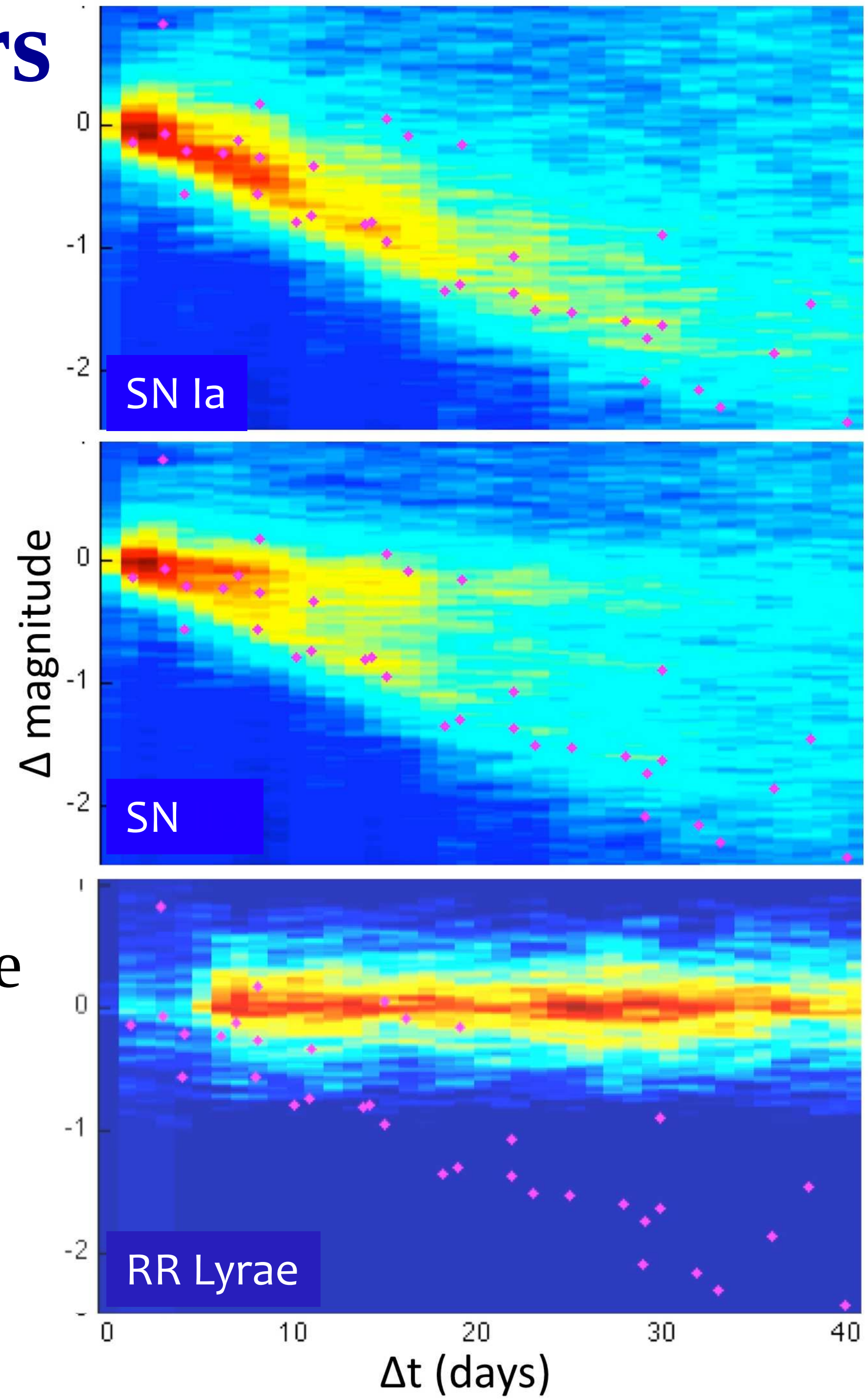
Expecting significant improvements when more observed features are used



2D Light Curve Priors

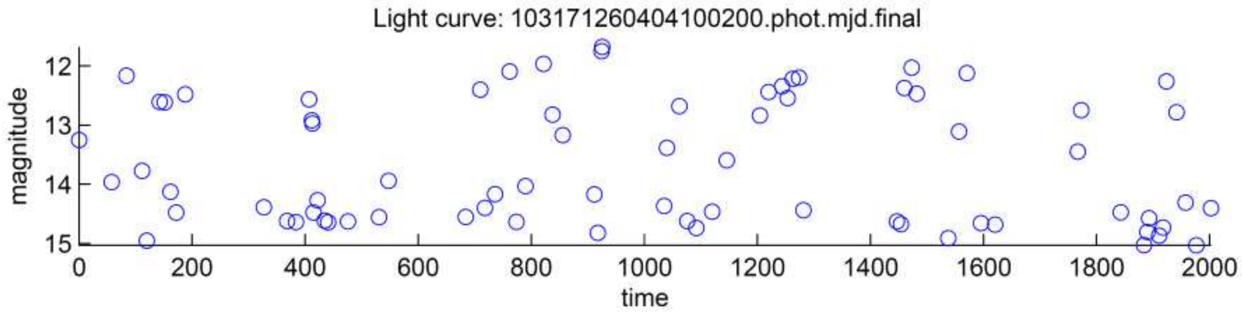
Lead : A. Mahabal

- For any pair of light curve measurements, compute the Δt and Δm, make a 2D histogram
  - Note: 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

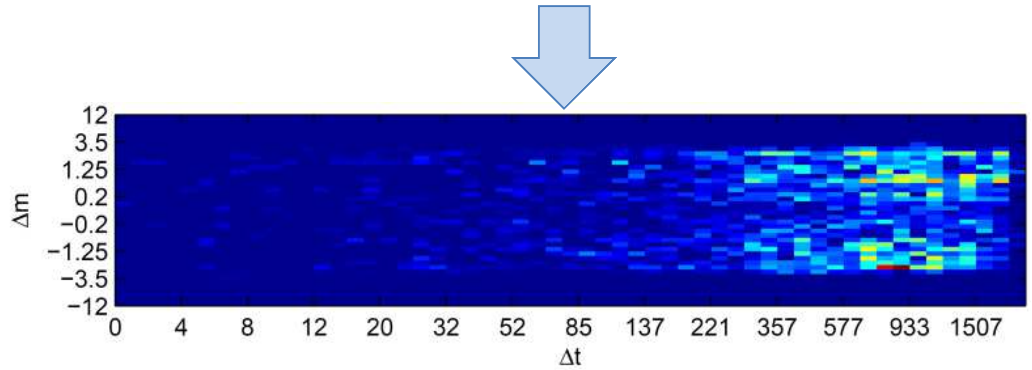


Lead: B. Moghaddam, JPL

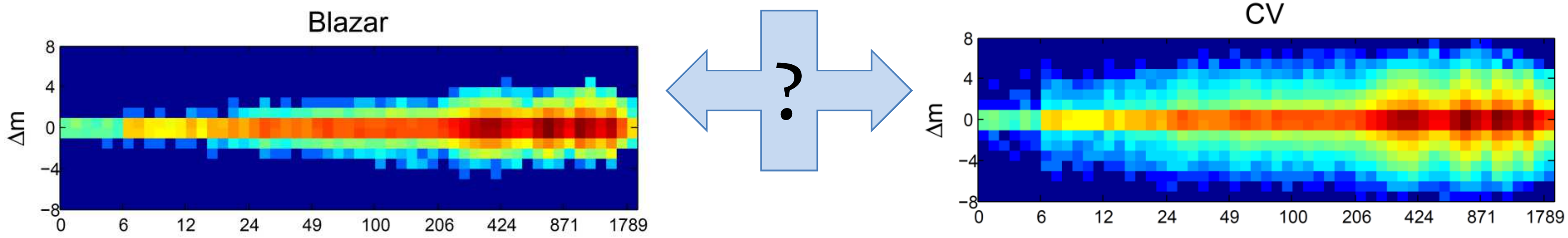
Applying *Δm vs. Δt* Histograms



Unclassified light curve



Its Δm vs. Δt histogram

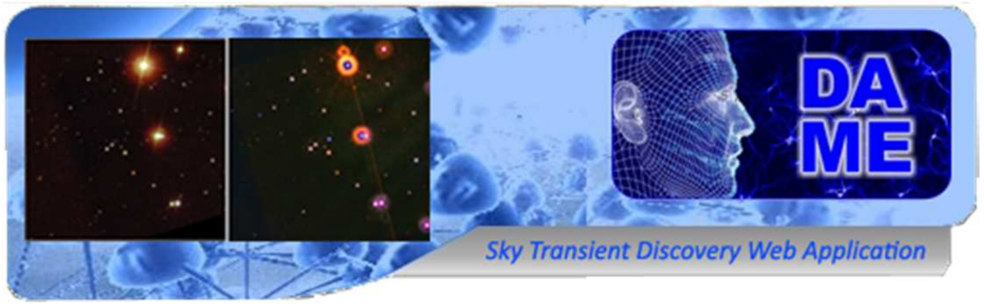


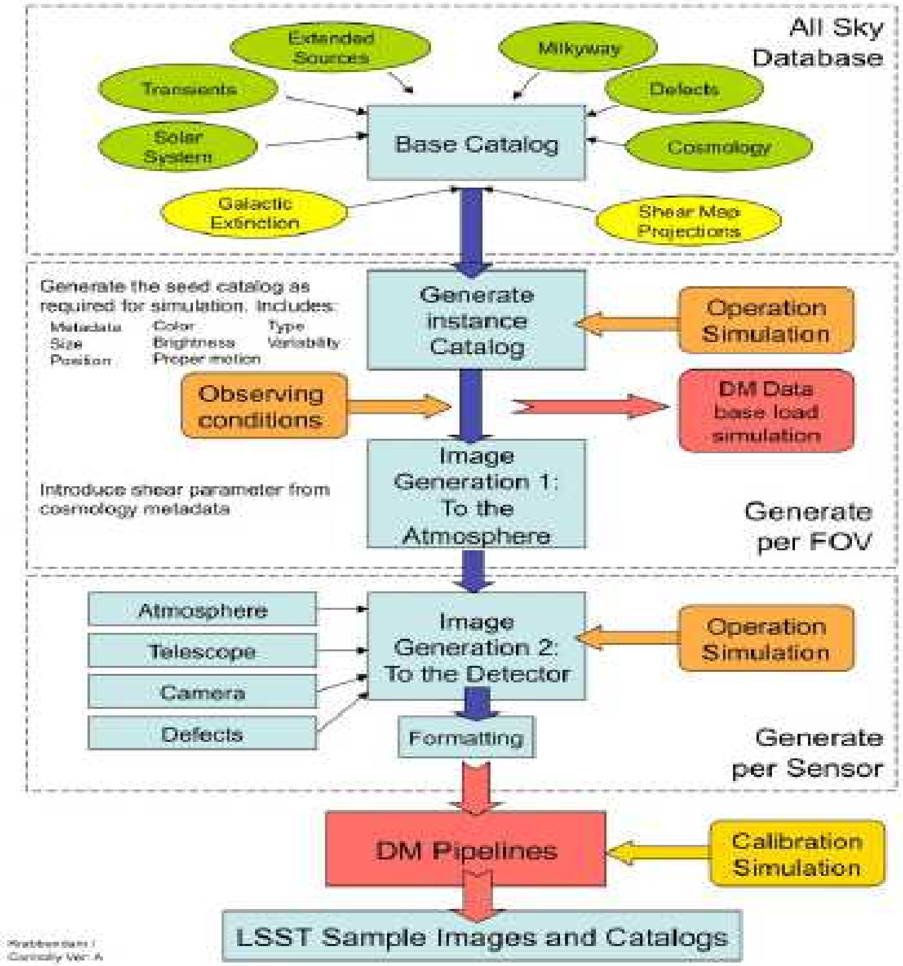
- Measure of a divergence between the unknown transient histogram and two prototype class histograms

STraDiWA

Sky Transient Discovery Web Application

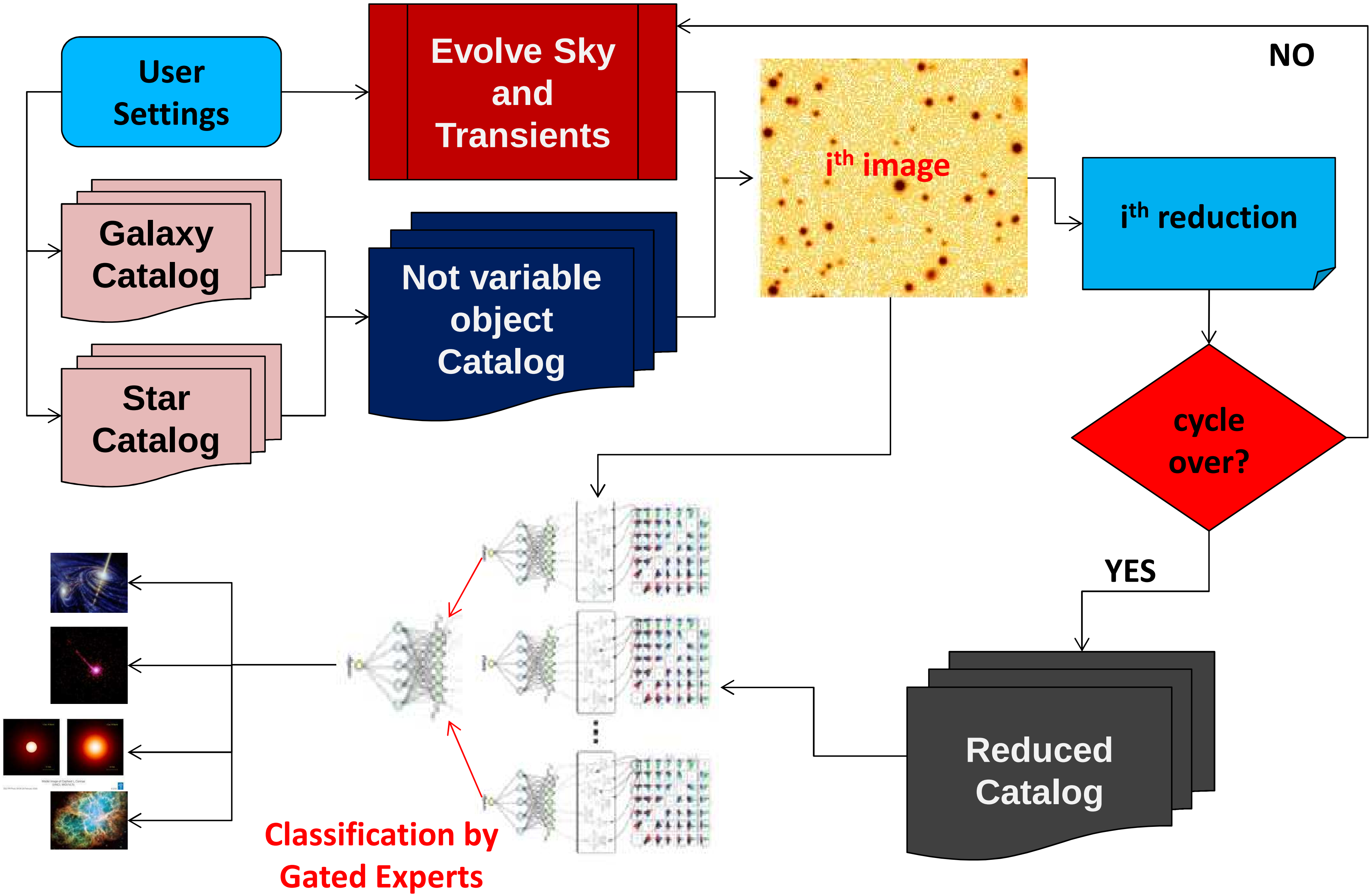
Lead: M. Brescia,



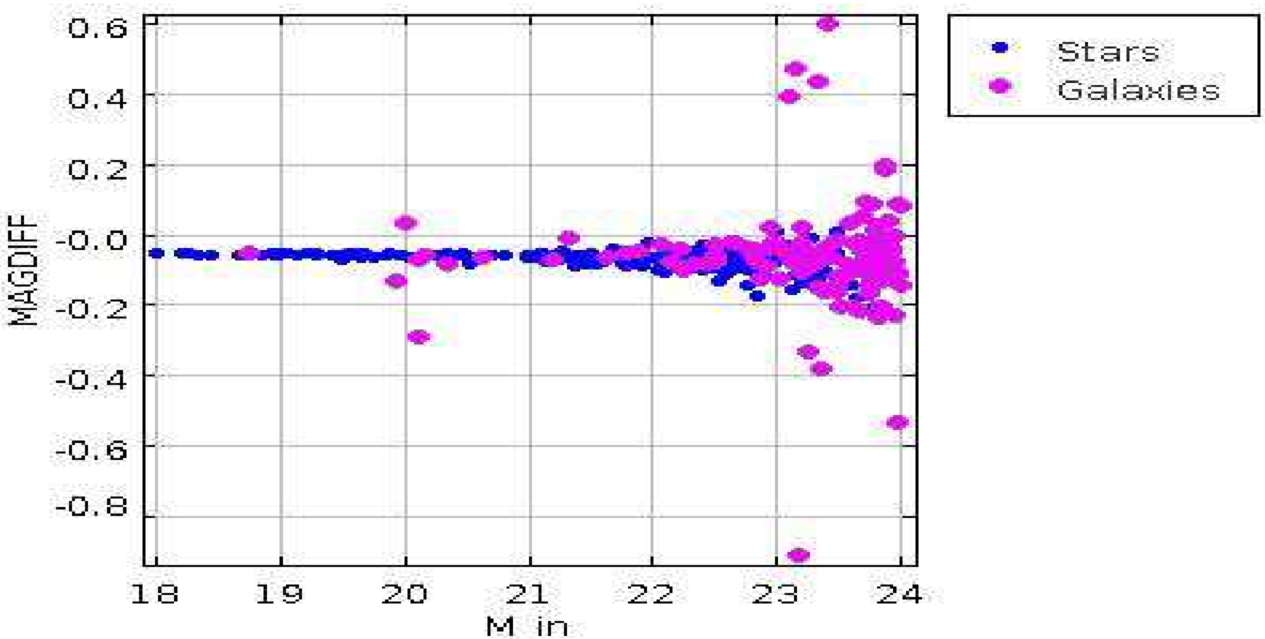


http://www.noao.edu/lsst/opsim/

STraDiWA Workflow



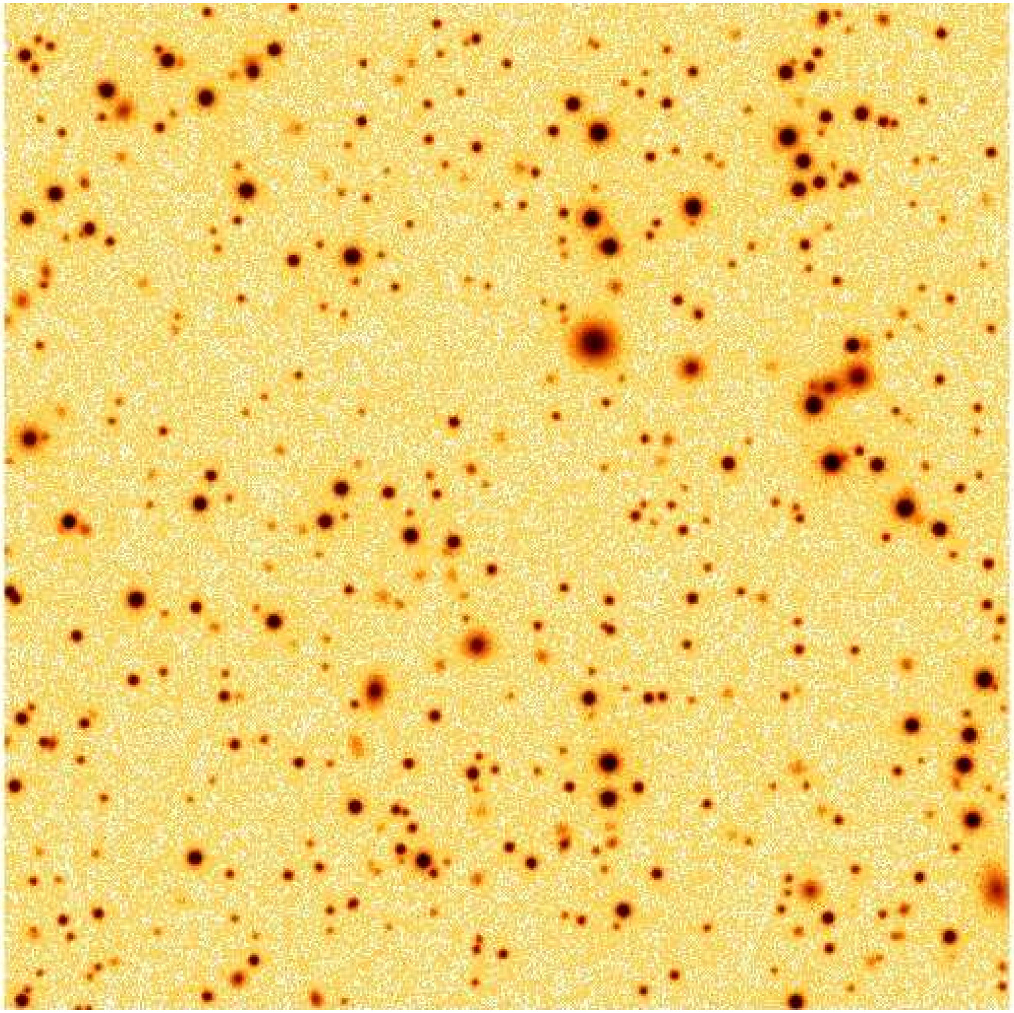
Catalog Extraction - III



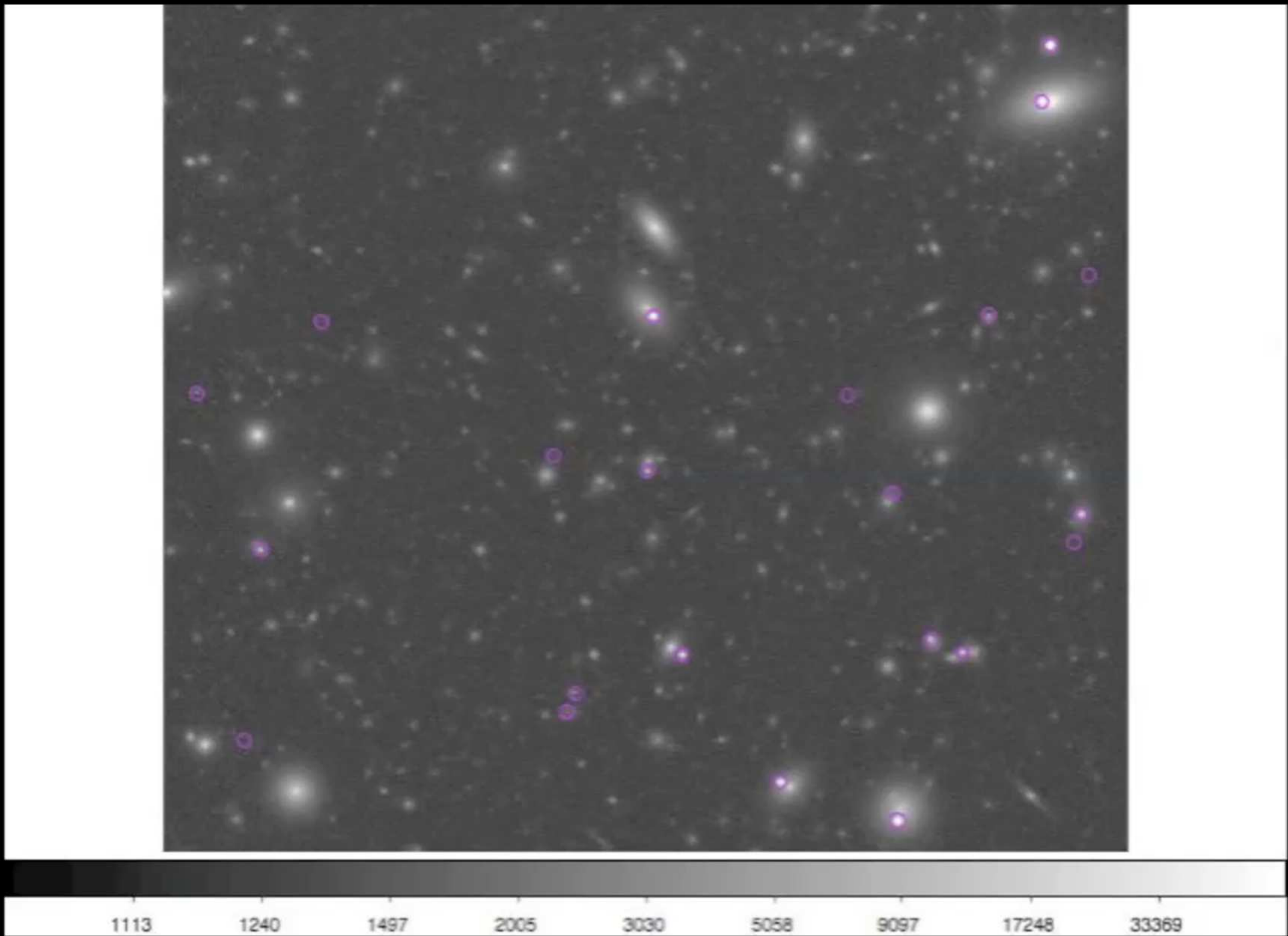
We report the magnitude difference ($M_{output} – M_{input}$) vs the input magnitude.

We can observe that there is a systematic shift below the zero that, for brightest objects, is of the order of 0.05 mag.

Catalog Extraction - IV



Example of simulated image wih magnitude of objects < 24, and seeing 1.07.


1113
1240
1497
2005
3030
5058
9097
17248
33369

Some thoughts

- Data mining ≈ Statistics expressed as algorithms
- *Scalability* with the number of data vectors and the number of dimensions is a looming issue:

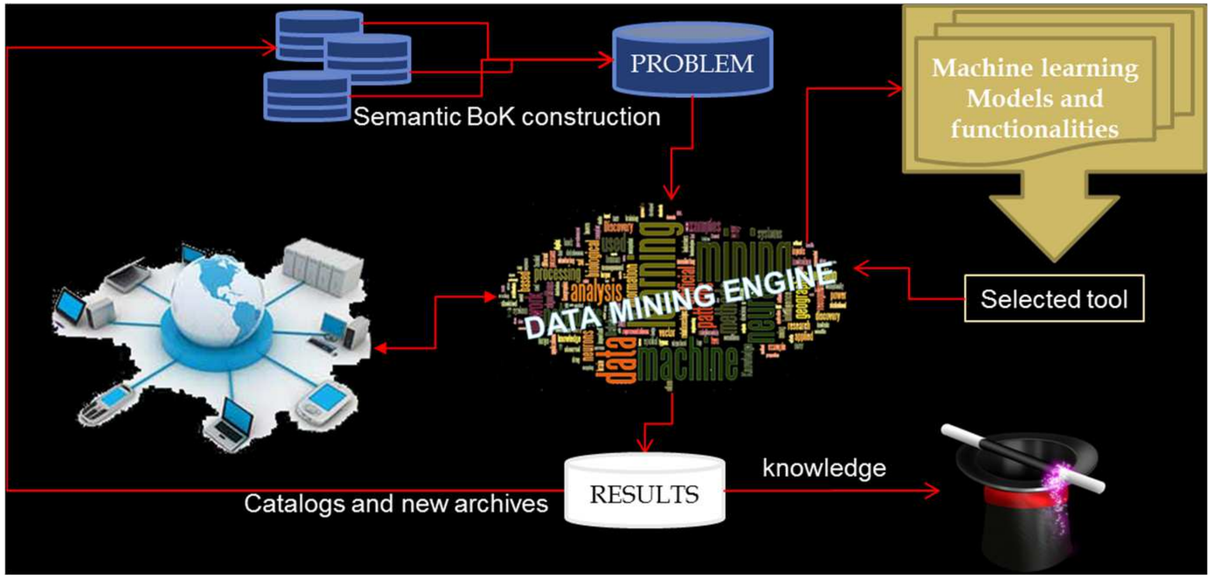
  N = data vectors, ~ 10^8 - 10^9, D = dimension, ~ 10^2 - 10^3

  – Clustering ~ N log N ⌫ N^2, ~ D^2
  – Correlations ~ N log N ⌫ N^2, ~ D^k ($k \geq 1$)
  – Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)
- Maybe we need to learn to live with *an incomplete analysis*: stop when you reach a desired accuracy
- Clusters are seldom Gaussian – beware of the assumptions
- Can we develop some entropy-like criterion to isolate portions or projections of hyper-dimensional parameter spaces where "something non-random may be going on"?
- Need to account for heteroskedadic errors
- Visualization in >>3 dimensions is a *huge* problem

data mining & Exploration Tool

Inspired by human brain features: high-parallel data flow, generalization, robustness, self-organization, pruning, associative memory, incremental learning, genetic evolution.

It is a web application for data mining experiments, based on WEB 2.0 technology



Multi Layer Perceptron trained by:
- Back Propagation
- Quasi Newton
- Genetic Algorithm

Support Vector Machines

Genetic Algorithms

Self Organizing Feature Maps

K-Means

Multi-layer Clustering

Principal Probabilistic Surfaces

Classification

Regression

Clustering

Feature Extraction

next ...

Bayesian Networks

Random Decision Forest

MLP with Levenberg-Marquardt

The data mining WEB Application

DAMEWARE - DAta Mining Web Application REsource

web-based app for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure.

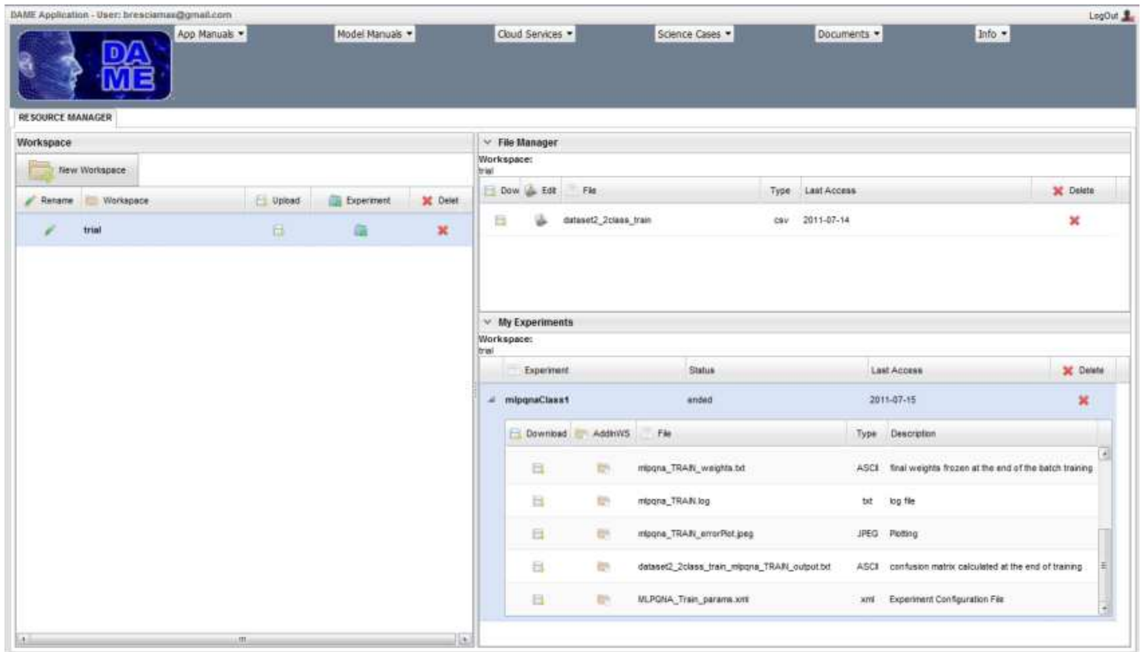
Intro page + manuals http://dame.dsf.unina.it/beta_info.html

Beta Release available http://dame.na.astro.it:8080/MyDameFE/

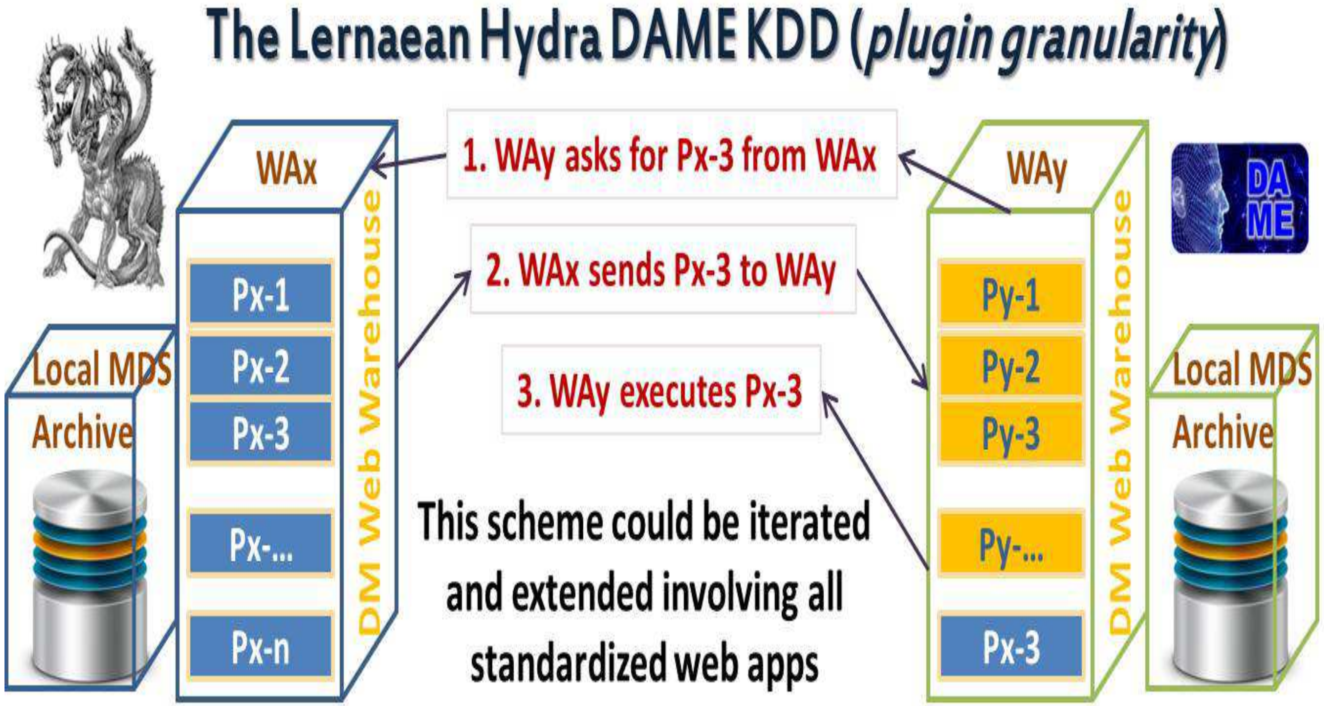
demo videos http://www.youtube.com/user/DAMEmedia

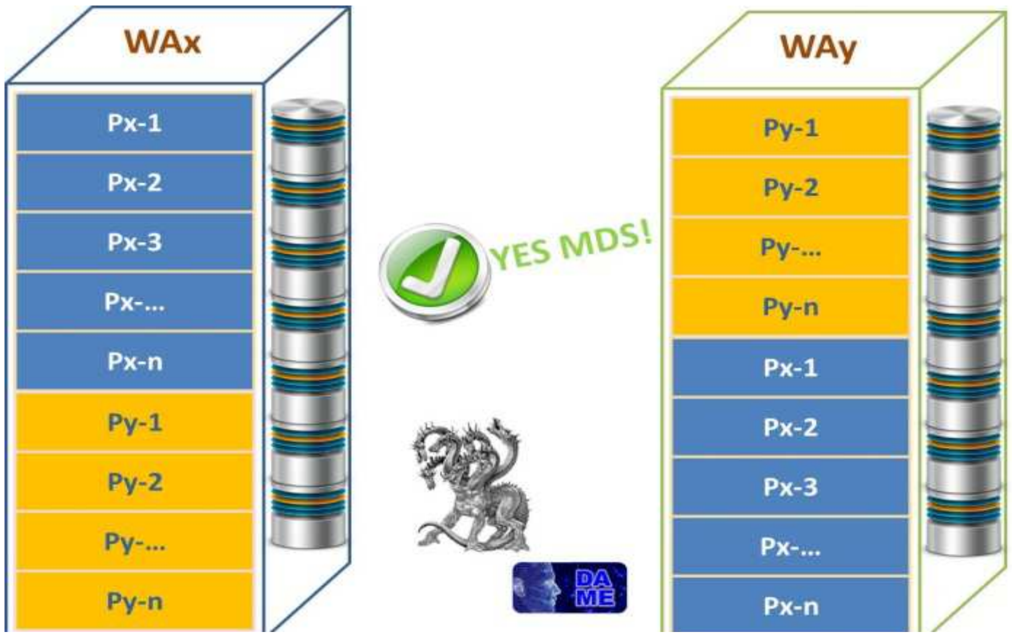
The release 1.0 will be deployed on a CLOUD, including GRID farm of S.Co.P.E.

- Private user account after registration;
- Data files (CSV, ASCII, FITS-image, FITS-table, VOTable);
- Classification models;
- Regression models;
- Clustering models;
- Feature Extraction models;
- Editing files for experiment setup (join, split, sort, shuffle, scale etc.);
- Output scatter plots and text data;



EVEN DAMEWARE IS NOT ENOUGH





M. Brescia et al.

Time Domain Astronomy

… is a new research frontier, touching all subfields of astronomy, from the Solar system to cosmology

… is ***here now*** (CRTS, PTF, PanSTARRS, ASCAP, *Kepler*, *Fermi*, …)

… **can be done here** and now **(archives and VST surveys!)**

- The **low-hanging fruit picking season is in a full swing**
- Lots of exciting and diverse science already under way

- ever growing importance of VO data grid, archives, and astroinformatics tools … TDA is ***astronomy of telescope and computational systems***, requiring a ***strong cyber-infrastructure*** *SCOPE !*
- ***Automated classification is a core problem***; it is critical and indispensable 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***

Some open challenges:

- Detection of faint, intermittent, sub-significant sources
- Sources/clusters on a correlated/clustered background
- Scalable clustering with a non-Gaussian geometry, errors
- Efficient discovery of “interesting” regions in parameter spaces
- Need for new metrics in the OPS
- Fast efficient and robust algoritms for the classification of transient events (sparse, heterogeneous data)
- Optimal decision making for limited follow-up resources

Thanks to:

Massimo Brescia (INAF-OAC)

G.S. Djorgovski, A. Mahabal and M. Graham (Caltech)

Former students who are collaborating:
C. Donalek (Caltech), R. D'Abrusco (CfA-Harvard)

Students:
S. Cavuoti, M. Annunziatella, D. De Cicco, M. Garofalo, A. Nocella, and …

the past, present and future DAME team

All the VST survey and pipelines people