DATA-rich astronomy: mining synoptic sky surveys

Stefano Cavuoti

Supervisors:

M. Brescia [1,2], G.Longo[1]

(1) Department of Physics – University Federico II – Napoli

(2) INAF – National Institute of Astrophysics – Capodimonte Astronomical Observatory - Napoli

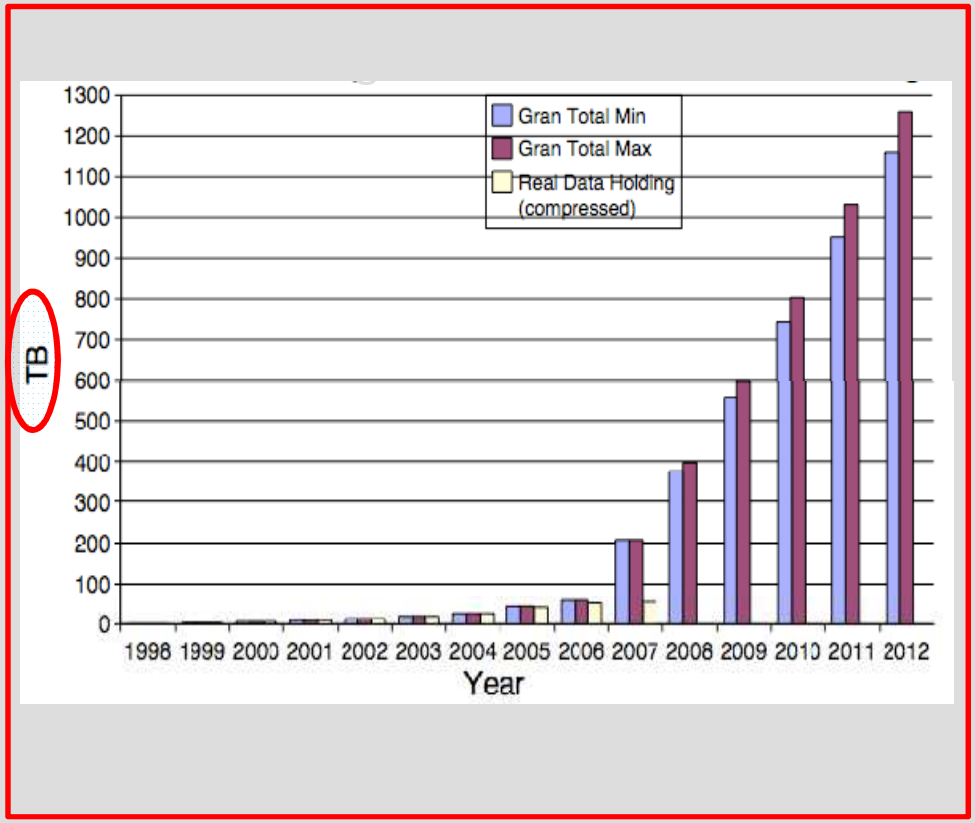
Astrophysics as a Data Rich Science

- Telescopes (ground-based and space-based, covering the full electromagnetic spectrum)

- Instruments (telescope/band dependent)

- Large digital sky surveys are becoming the dominant source of data in astronomy: ~ 10-100 TB/survey (soon PB), ~ 10^6 - 10^9 sources/survey, many wavelengths…

- Data sets many orders of magnitude larger, more complex, and more standardized than in the past



The General Astrophysical Problem

Due to new instruments and new diagnostic tools, the information volume grows exponentially

Most data will never be seen by humans! (BLADE RUNNER)

➡ The need for data storage, network, database-related technologies, standards, etc.

<u>Information complexity is also increasing greatly</u>

Most knowledge hidden behind data complexity is lost

➡ Most (all) empirical relationships known so far depend on 3 parameters ….
Simple universe or rather human bias?

Most data (and data constructs) cannot be comprehended by humans directly!

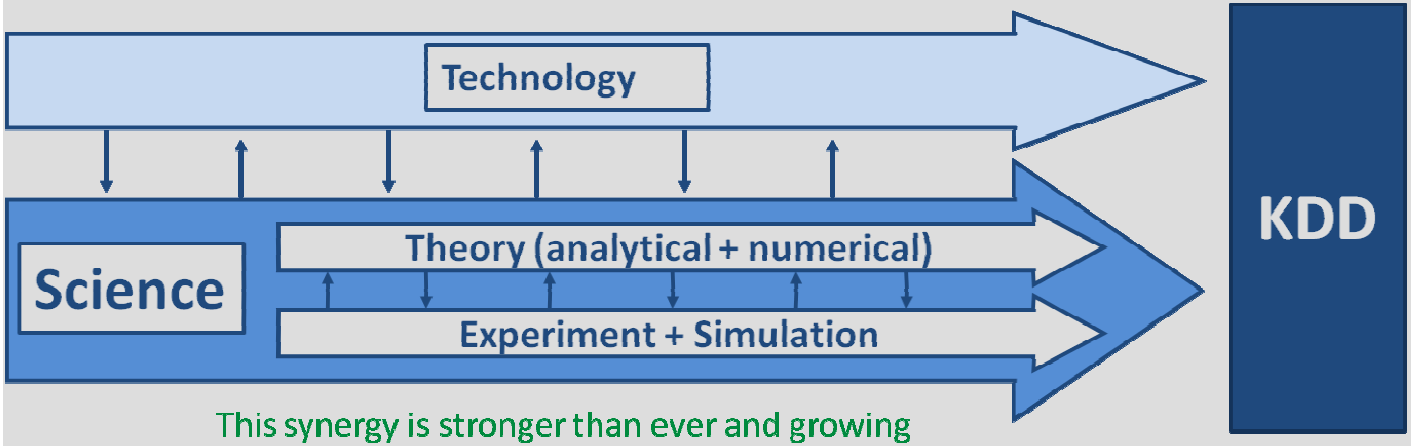
➡ The need for data mining, KDD (Knowledge Discovery in Databases), data understanding technologies, hyper dimensional visualization, AI/Machine-assisted discovery

Data Mining (KDD) as the Fourth Paradigm Of Science

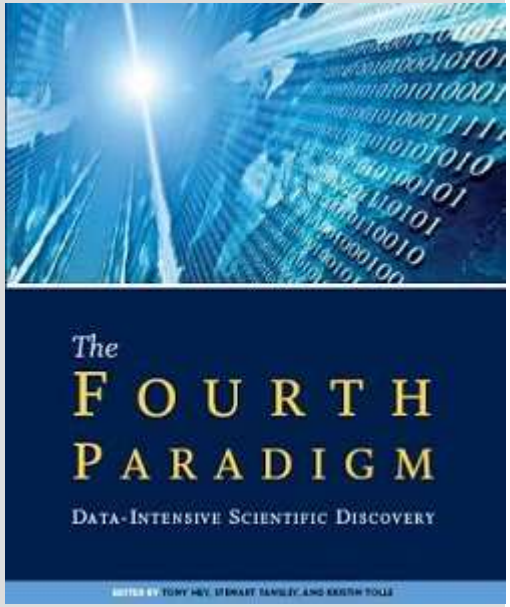
The old traditional, "Platonic" view:

Pure Theory ⇒ Experiment ⇒ Technology & Practical Applications

The modern and realistic view when dealing with complex data sets:



This synergy is stronger than ever and growing



The BoK's Problem

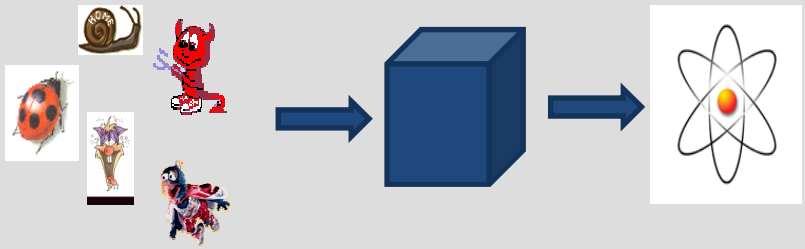
Limited number of problems due to limited number of reliable BoKs

(BoK) Bases of knowledge
(set of well known templates for supervised (training) or unsupervised (labeling) methods

So far
- Limited number of BoK (and of limited scope) available
- Painstaking work for each application (es. spectroscopic redshifts for photometric redshifts training)
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

Bases of knowledge need to be built automatically from Vobs Data repositories
•There's a need of standardization and interoperability between data together with DM application

Community believes AI/DM methods are black boxes
You feed in something, and obtain patters, trends, i.e. knowledge….

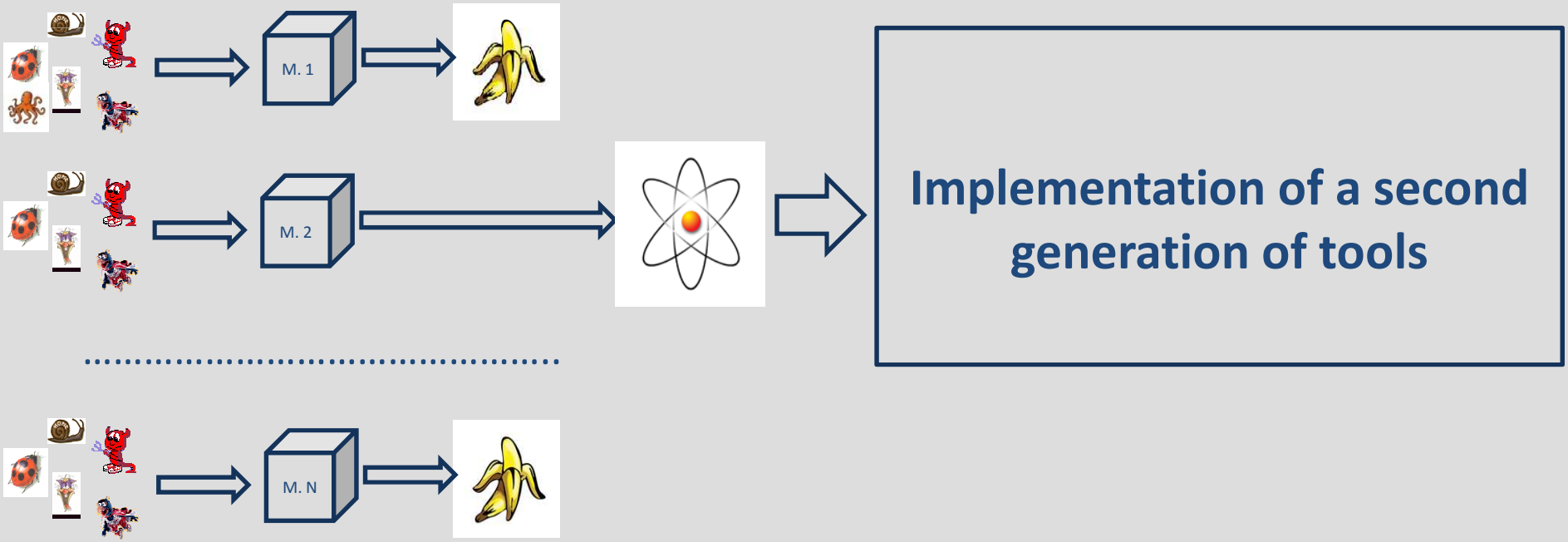


The Choice Problem

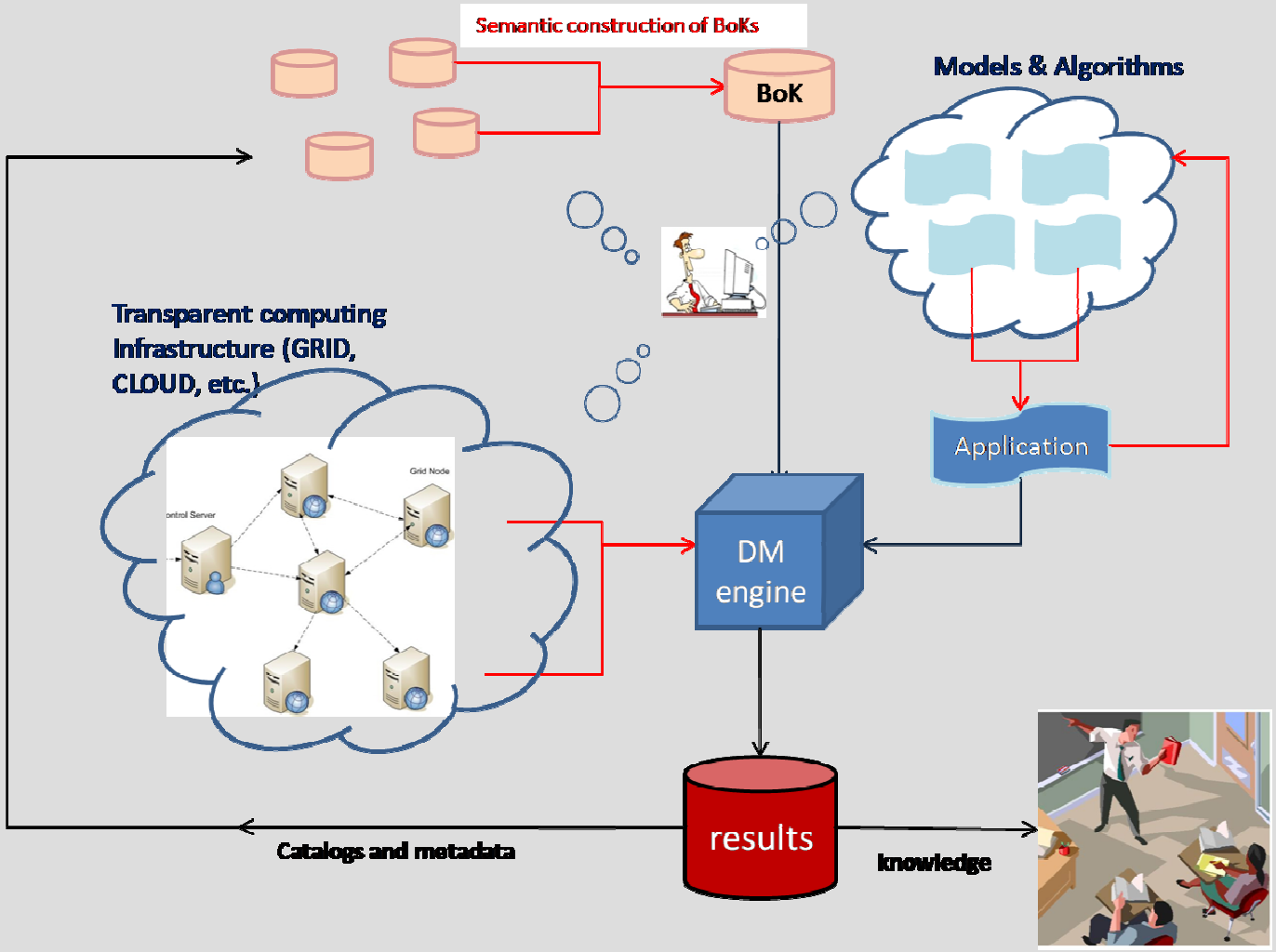
Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is not willing to make an effort to learn them ….
The r.m.s astronomer doesn't want to become a computer scientist or a mathematician
(large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud no personal certificates, no deep understanding of the DM tool etc.

Allow each astronomer to build his black box using own algorithm to run on the infrastructure without knowledge about the infrastructure



Effective DM process break-down



The Black box Infrastructure

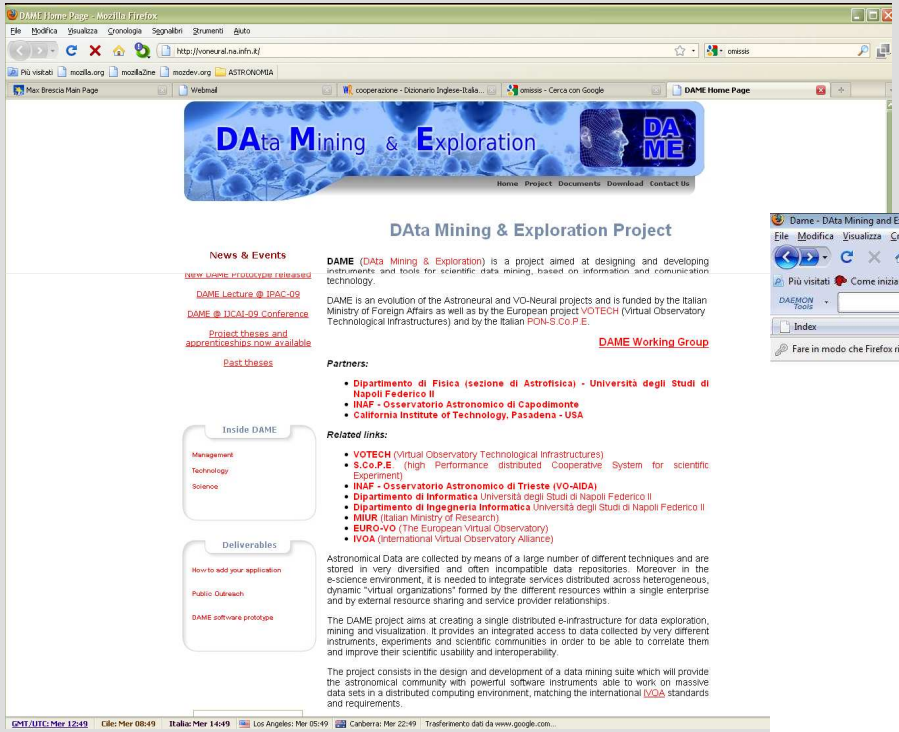
In this scenario DAME (Data Mining & Exploration) project, starting from astrophysics requirements domain, has investigated the Massive Data Sets (MDS) exploration by producing a taxonomy of data mining applications (hereinafter called functionalities) and collected a set of machine learning algorithms (hereinafter called models).

This association functionality-model represents what we defined as simply "use case", easily configurable by the user through specific tutorials. At low level, any experiment launched on the DAME framework, externally configurable through dynamical interactive web pages, is treated in a standard way, making completely transparent to the user the specific computing infrastructure used and specific data format given as input.

So the user doesn’t need to know anything about GRID, Cloud or what else.

What DAME is

DAME is a joint effort between University Federico II, INAF-OACN, and Caltech aimed at implementing (as web application) a scientific gateway for data analysis, exploration, mining and visualization tools, on top of virtualized distributed computing environment.



http://voneural.na.infn.it/alpha_info.html

Web application PROTOTYPE

http://voneural.na.infn.it/

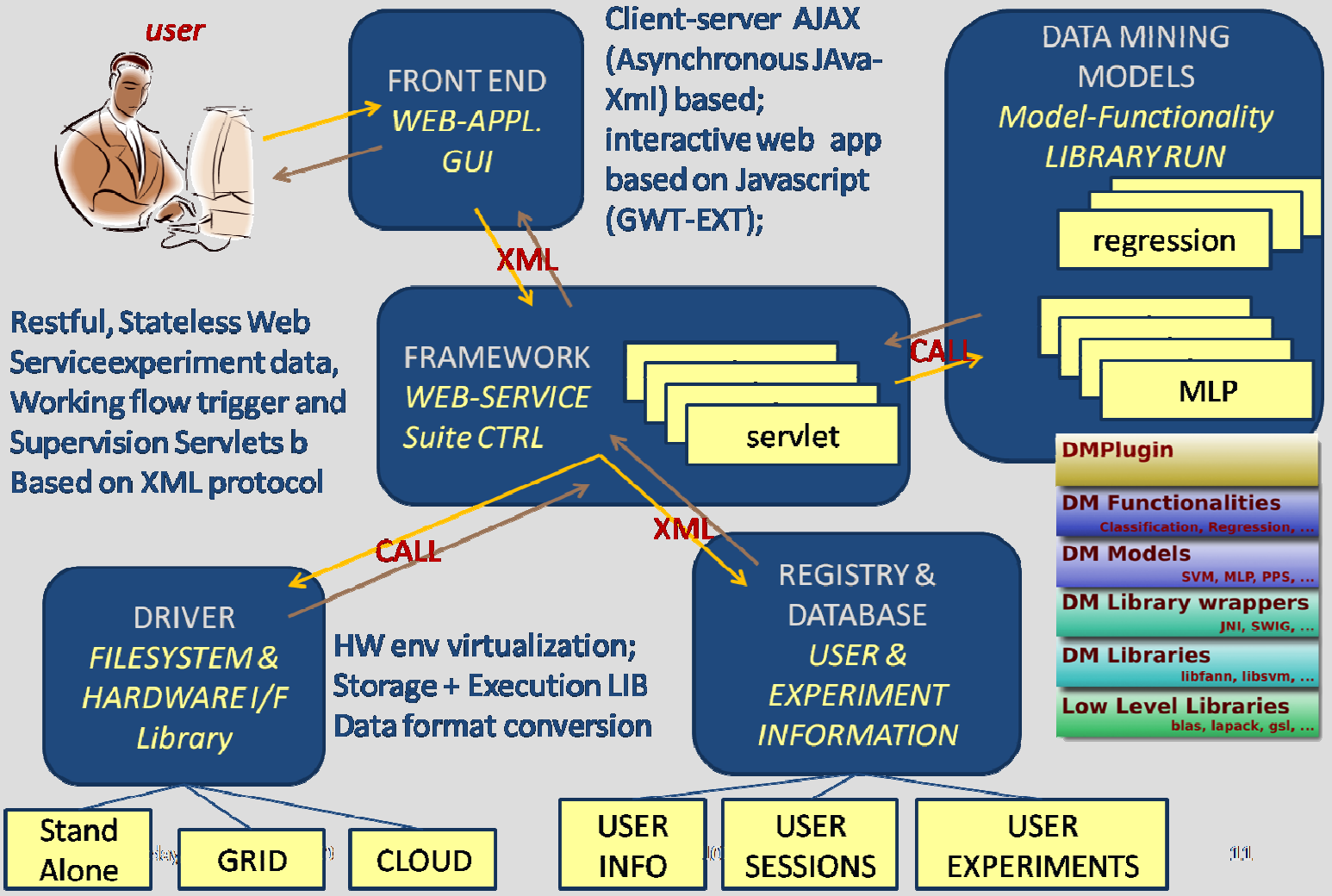
Technical and management info

Documents

Science cases

Newsletter

The DAME Architecture



Globular Clusters Search

(in coll. with M. Paolillo & DAME coll.)

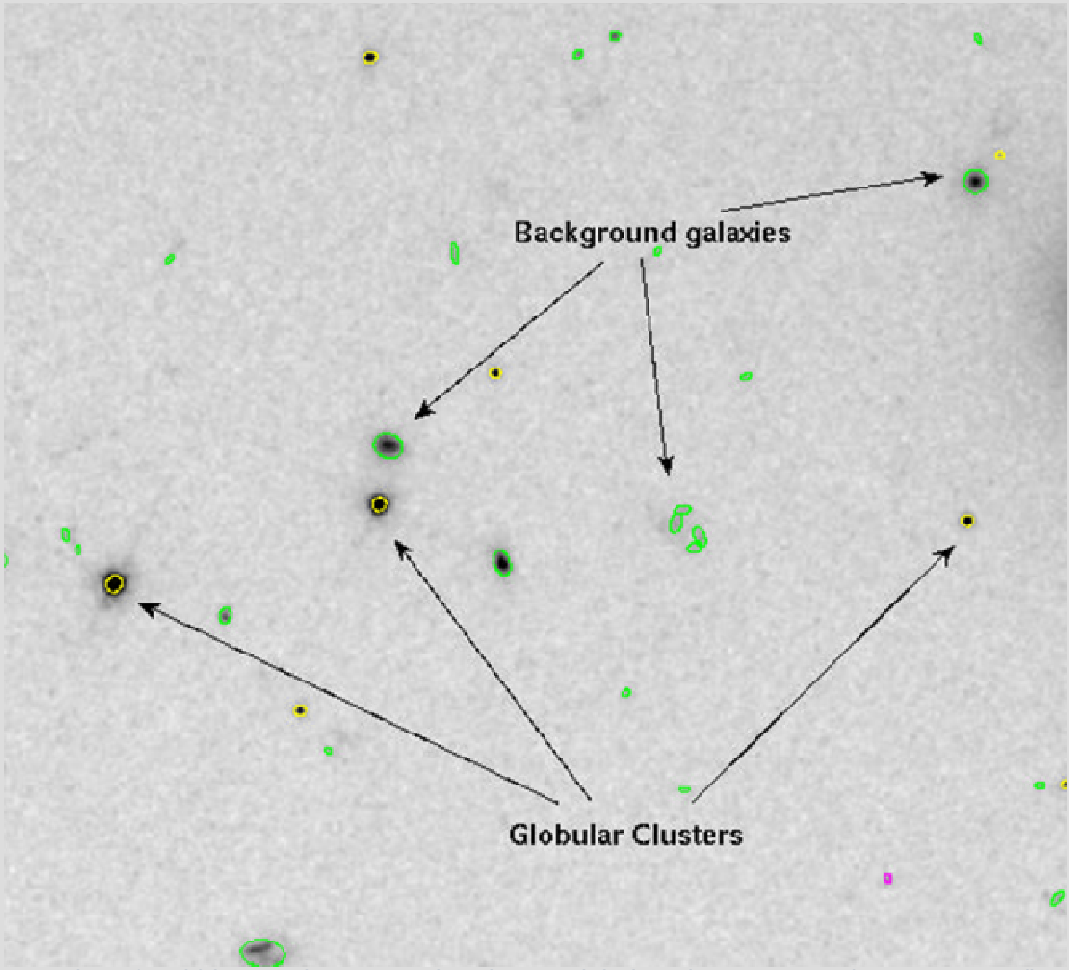
The study of Globular Clusters populations in external galaxies requires the use of wide-field, multi-band photometry.

In order to minimize contamination high-resolution data are required

BoK (TRUE CLUSTERS) selection in color

Single images only one flux and morphology (reduces the cost in terms of observing time)

http://voneural.na.infn.it/dame_gcs.html



Section of Hubble ACS image used to detect Globular Clusters around N1399. GCs (in yellow)
GC are difficult to distinguish from background galaxies (in green) based only on single band images.

Globular Clusters Search

First Results can be summarized as it follows:

Classification Statistics results on 2100 patterns
(in the classification case, confusion matrix column values are conditioned probabilities)

	P(class 1)	P(class 2)
target class 1 -->	1197	22
target class 2 -->	21	860

total classification percentage: 95.5%
class 1 classification percentage: 96.4% (class 1: not globular cluster)
class 2 classification percentage: 94.3% (class 2: it is a globular cluster)

Classical methods bring to results about 10% worse than this

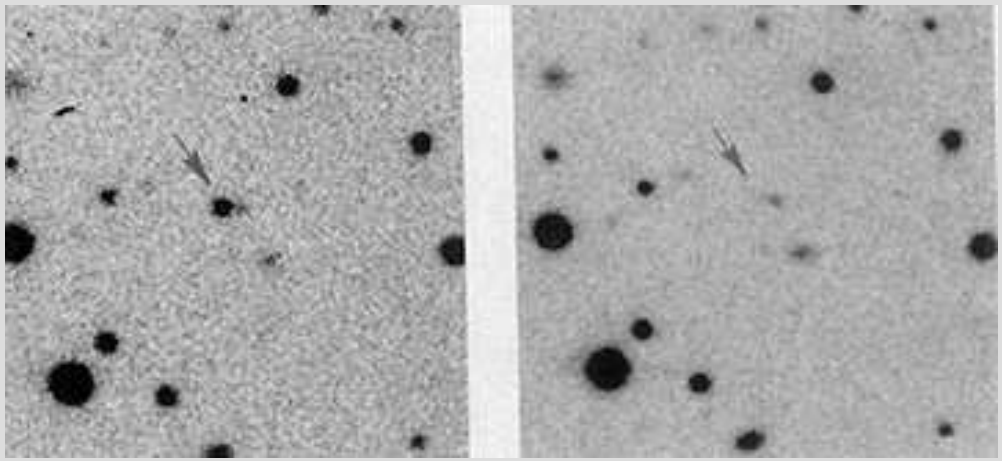
FUTURE DEVELOPMENTS: finding a method to recognize from photometric observations in the optical bands only, GC containing X-ray emitters, a problem which has been widely discussed in the literature for the following reasons:
Constraining which observables play a role in the existence or not of Low Mass X-ray Binary(LMXB): irradiation-induced winds (Maccarone et al. 2004), magnetic breaking (Ivanova 2006) or IMF variations (Grindlay 1987, also see Jord´n et al. 2004) can explain the a LMXB formation likelihood as a function of the host GC color in terms of a metallicity effect, while other dynamical models (e.g Kim et al. 2006; Jord´n et al. 2007a; Sivako et al. 2007) suggest that this color dependence may reflect the higher LMXB formation efficiency in more centrally-concentrated red GCs.

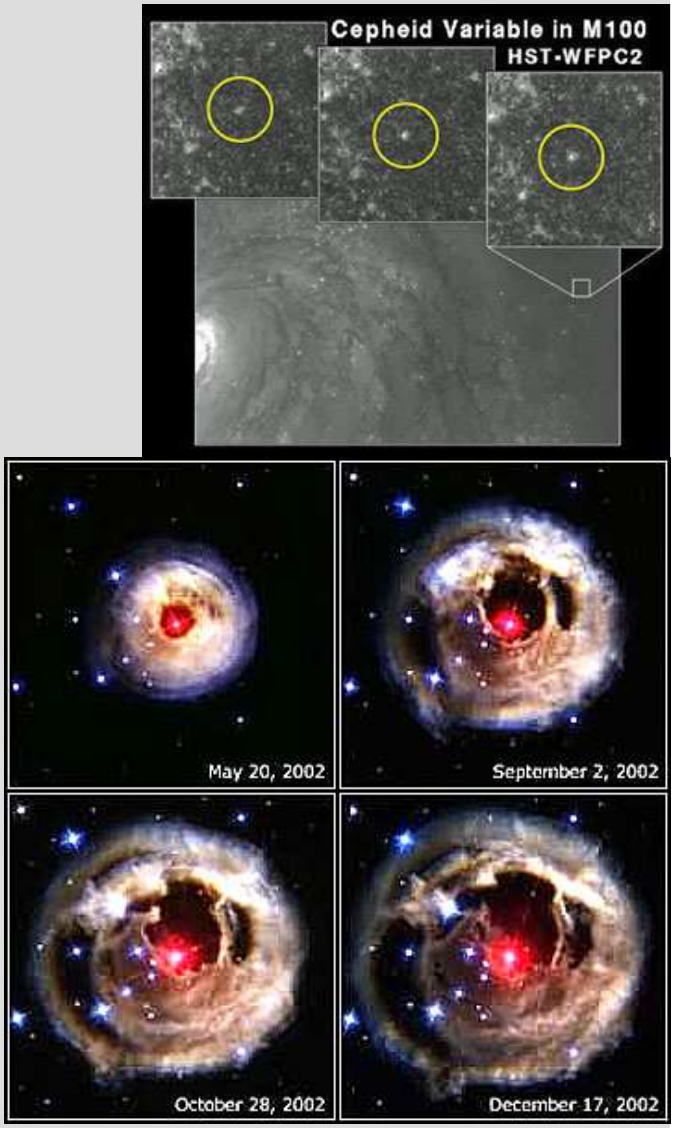
Transient Discovery

(in coll. with M. Annunziatella & DAME coll.)

A object is classified as transient if its apparent brightness as seen from Earth changes over time, whether the changes are due to variations in the actual luminosity, or to variations in the amount of the light that is blocked from reaching Earth.

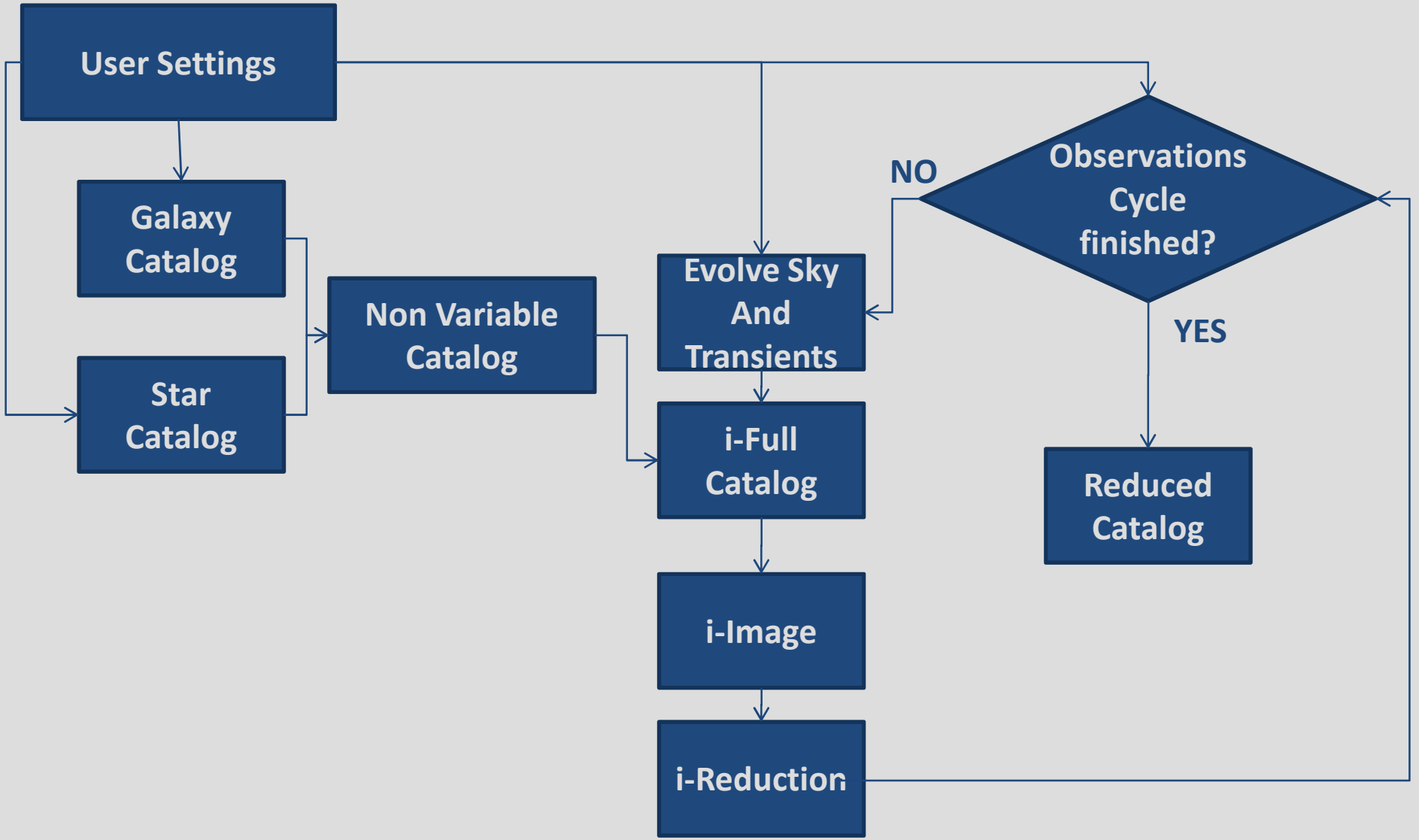
The study of Transient populations in external galaxies requires the use of real-time classificators that may identify a Transient object when it's changing





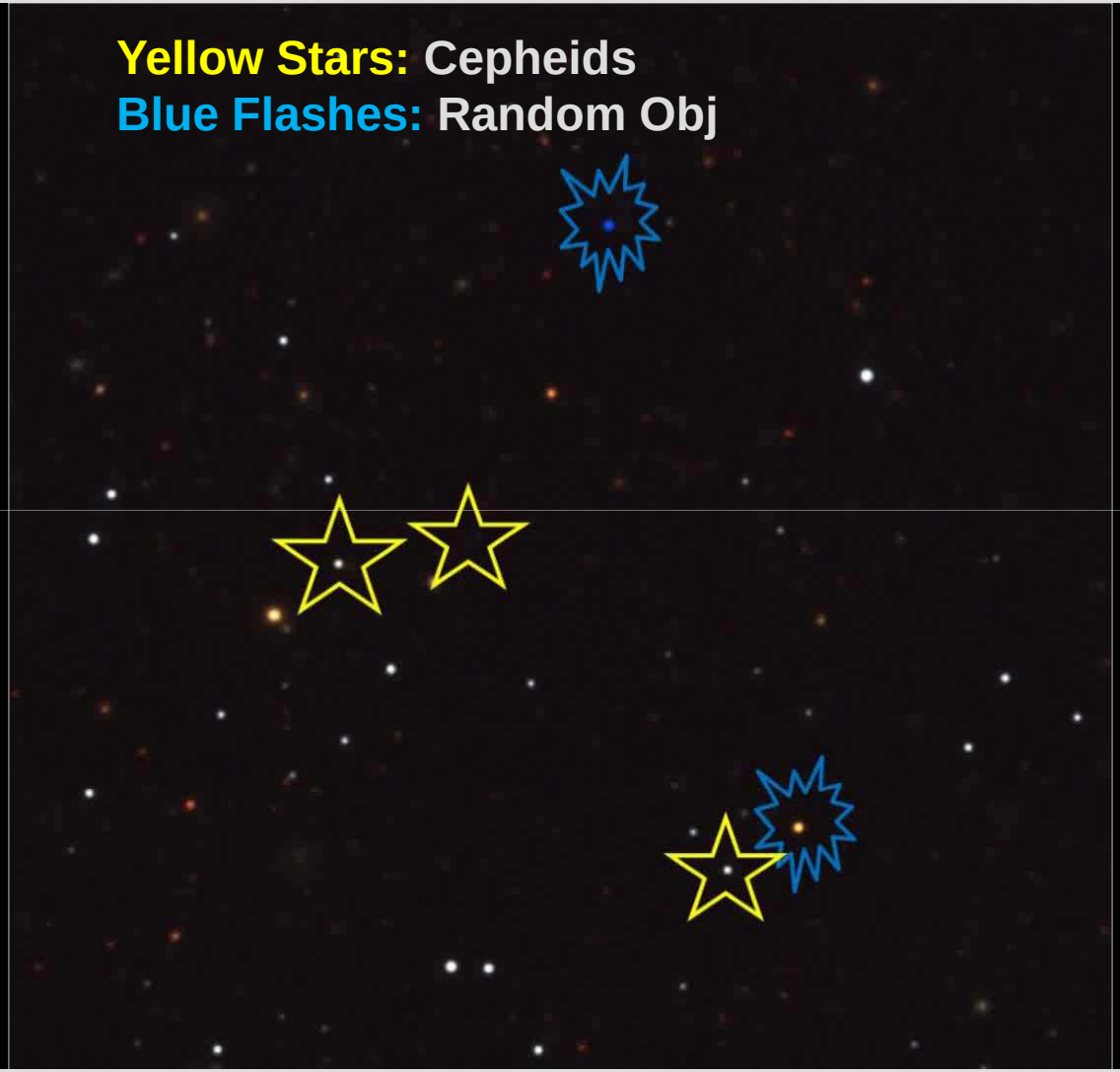
V838 - HST

Transient Discovery



Prototype simulation: the movie

We have simulated 12 images using VST features
Mag: 18 - 26.0
Pixel Size: 0.213 arcsec
Exp Time: 1500.0(s)
Seeing: 0.6 - 1.1
Image Size: 1024x1024



the video is available at: http://voneural.na.infn.it/dame_td.html

Summary of PhD activities

- **Courses**:
  - theoretical and observational cosmology (Dr. Gianpiero Mangano) 6 credits
  - tecnologie astronomiche (Dr. Massimo Brescia) 8 credits
  - ingegneria del software (Prof. Sergio Di Martino) 6 credits
- **Workshop attended so far:**
  - Lecture at INGRID 2010 (http://www.ingrid.cnit.it): DAME: A Distributed Data Mining & Exploration Framework within the Virtual Observatory
- **External Collaborations**:
  - Massimo Brescia (INAF OAC) – DAME, Transient Discovery, Globular Cluster Search
  - George S. Djorgovski (CALTECH) – DAME, Transient Discovery
  - Ciro Donalek (CALTECH) – DAME, Bayesian network and citizen science
  - Ashish Mahabal (CALTECH) – DAME, Transient Discovery
  - Raffaele D'Abrusco (Harvard-Smithsonian Center for Astrophysics) – DAME
  - Omar Laurino (INAF OAT) – DAME

Publications (1st year)

- **Refereed Publications :**

  - **DAME: A Web Oriented Infrastructure for Scientific Data Mining & Exploration**, Massimo Brescia, Giuseppe Longo, George S. Djorgovski, Stefano Cavuoti, Raffaele D'Abrusco, Ciro Donalek, Alessandro Di Guido, Michelangelo Fiore, Mauro Garofalo, Omar Laurino, Ashish Mahabal, Francesco Manna, Alfonso Nocella, Giovanni d'Angelo, Maurizio Paolillo, submitted at Journal of Computational Science, Elsevier, ISSN: 1877-7503 (**arXiv:1010.4843v1)**
  - **DAME: a VO compliant platform for astrophysical data mining**, M. Brescia1*;2, G. Longo2;3;1, G. Djorgovski3, S. Cavuoti2, R. D'abrusco4, O. Laurino5, and* A. Mahabal, PASP, in preparation
  - **Selection of candidate globular clusters: a data mining approach**, Brescia, Cavuoti, Longo, Paolillo, Putzia, MNRAS, in preparation

- **Conference proceedings:**

  - **DAME: A Distributed Web Based Framework for Knowledge Discovery in Databases**, M. Brescia, G. Longo, M. Castellani, S. Cavuoti, R. D'Abrusco, O. Laurino, Mem. S.A.It., Vol. 75, 282, INAF OAC Napoli, Italy, 2010
  - **DAME: A Distributed Data Mining & Exploration Framework within the Virtual Observatory**, M. Brescia, S. Cavuoti, R. D'Abrusco, O. Laurino, G. Longo, INGRID 2010 Workshop on Instrumenting the GRID, Springer Editor (accepted in press), Poznan, Poland, May 12-14, 2010

Conclusion

• Dame β release is coming and as soon as possible new models and new functionality will be released. Dame α release is available at http://voneural.na.infn.it/alpha_info.html **During my PhD I will implement new plugins and new functionalities.**

• First Results on the selection of candidate Globular Clusters in external galaxies. We are tackling the second part of the work, about the X-Ray binaries selection.

• The pipeline for the transients simulation is in progress and soon we shall tackle the most challenging problem of detection and classification of the transient from images and then switch from the simulated images to real ones.

Final goal of my PhD is to obtain:
a complete pipeline for the real time classification of transients in synoptic (multiepoch multiwavelength surveys) and to apply it to real data from the Catalina Sky Survey and the VST surveys