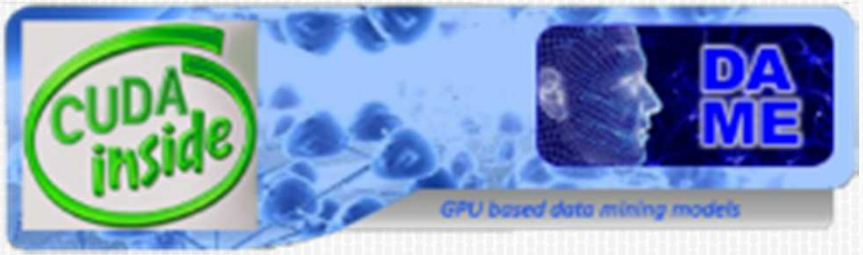
GENETIC ALGORITHM MODELING WITH GPU PARALLEL COMPUTING TECHNOLOGY

S. Cavuoti[(1,3)], M. Garofalo[(2)], M. Brescia[(3,1)], M. Paolillo[(1)], G. Longo[(1,4)], A. Pescapè[(2)], G. Ventre[(2)]
and DAME Working Group[(1,3,4)]

(1) *Dept. of Physics, Faculty of Sciences, University Federico II*
(2) *Dept. of Informatics & Systems, Faculty of Engineering, University Federico II*
(3) *INAF Capodimonte Astronomical Observatory*
(4) *California Institute of Technology*



THE SCIENTIFIC USE CASE

NGC1399 Galaxy Dataset

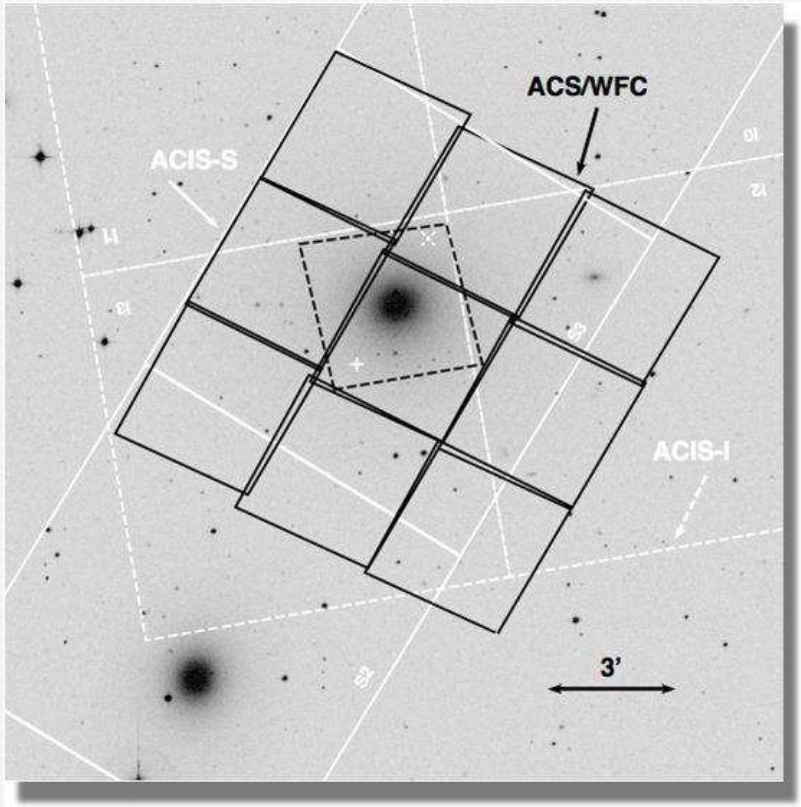
NGC1399 (~20 Mpc) is an ideal target because allows to probe a large fraction of the galaxy and still resolve GC sizes.

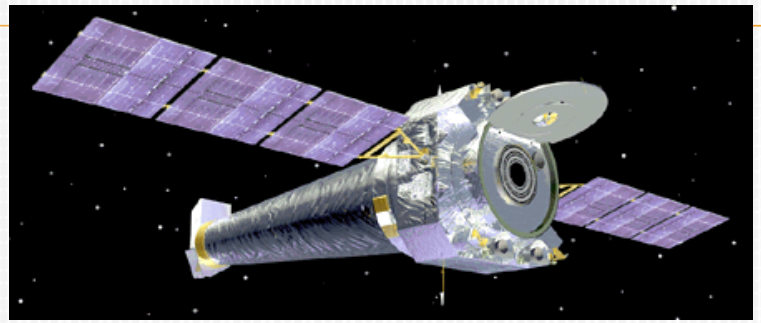
9 HST V-band (f606w) observations, drizzled to super-Nyquist sampling the ACS PSF (2.9 pc/pix).

Chandra ACIS-I + ACIS-S

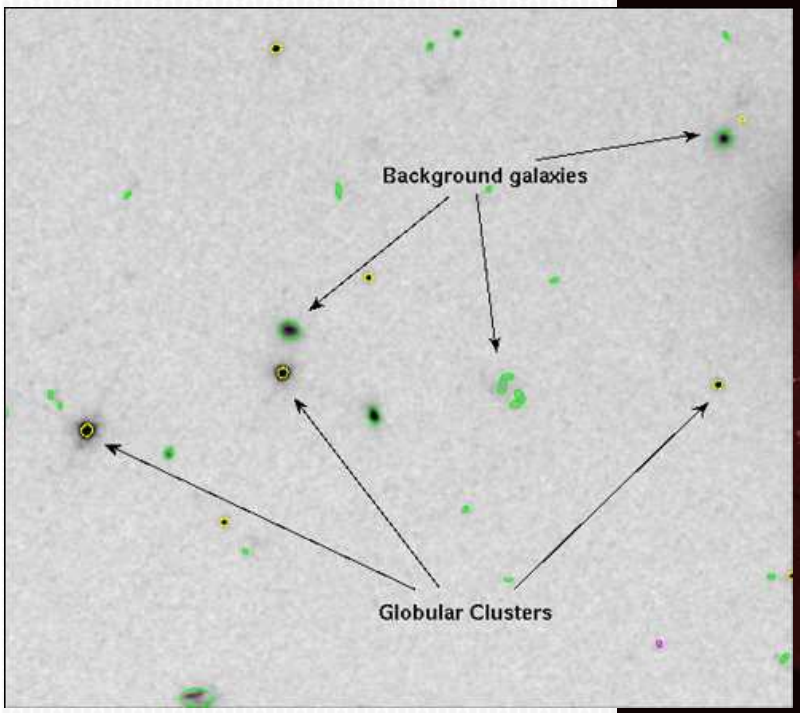
ACS *g-z* colors for central region

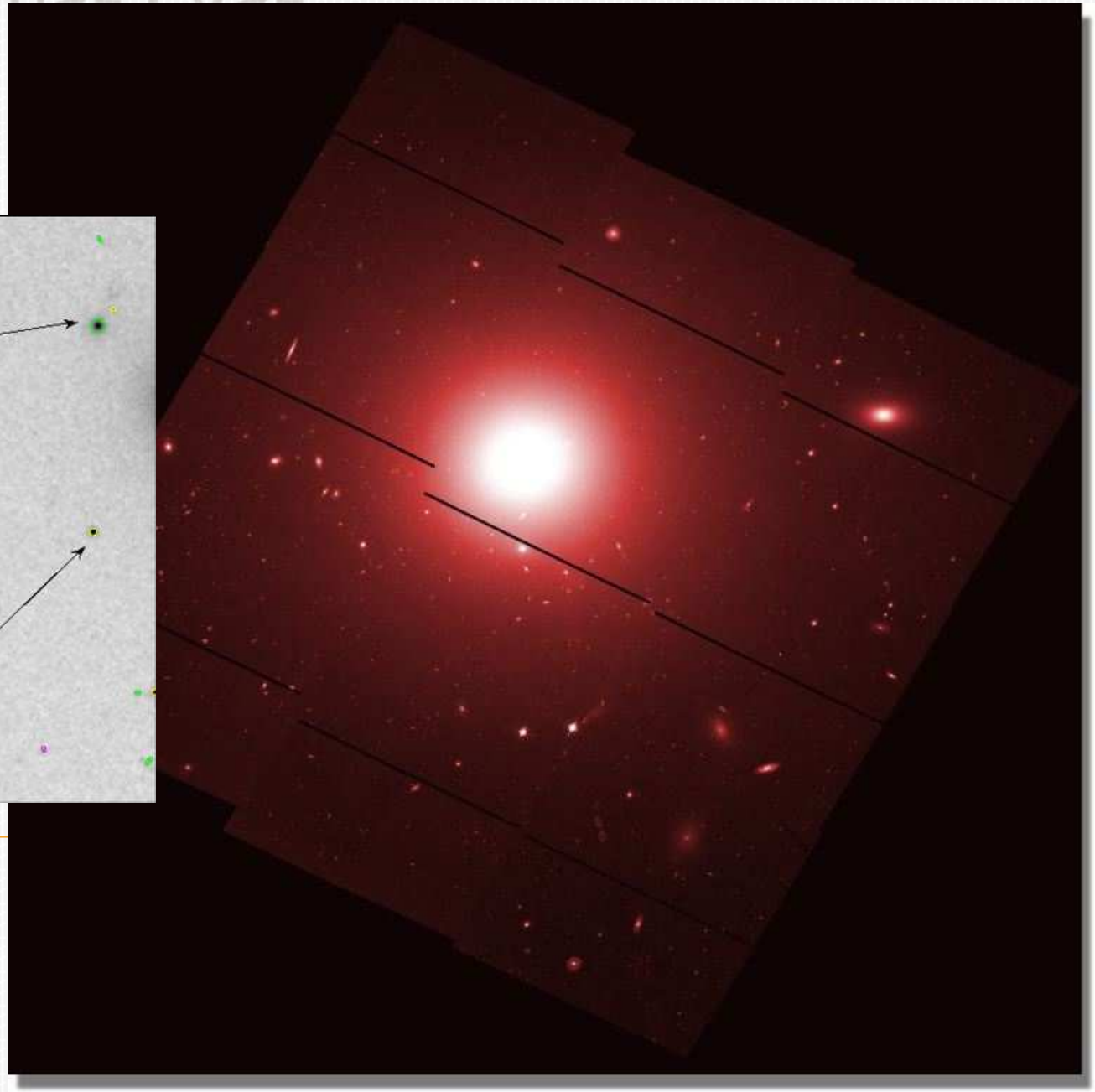
Ground-based *C-R* photometry for part of the sources over the whole field





THE SCIENTIFIC USE CASE





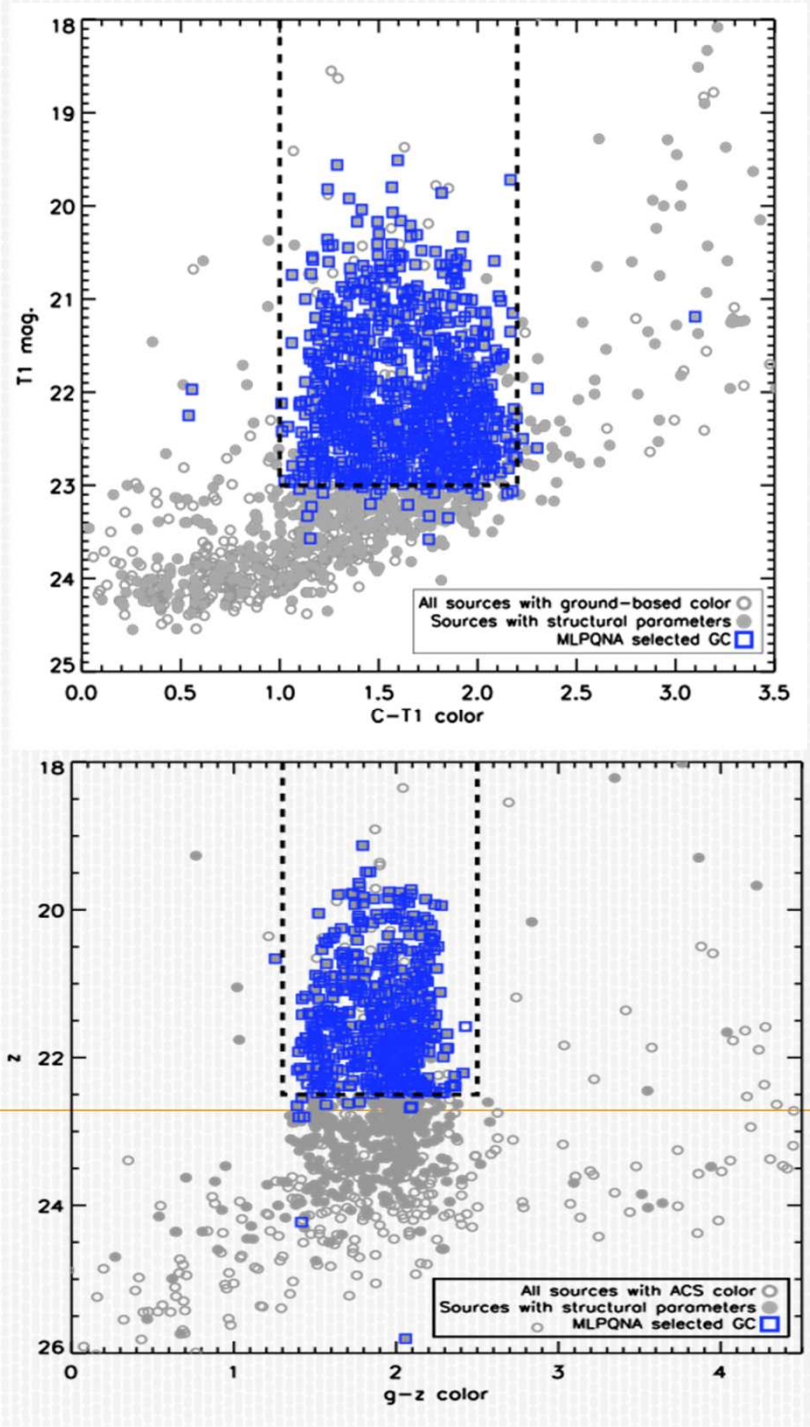
THE SCIENTIFIC USE CASE

THE DATA-MINING SELECTION APPROACH

Wide-field, multi-band observations are expensive. Easier to get single-band mosaics and possibly ground-based colors, or colors on subset of sources.

Using the subsample of sources with colors, we trained a Neural Network and Genetic Algorithm to recognize GCs, not relying on arbitrary criteria.

The experiments yields a sample 97% complete and <5% contaminated, with respect to color selected samples.



(Brescia et al. 2012 - MNRAS,
Volume 421, Issue 2, pp. 1155-1165)

http://dame.dsf.unina.it/dame_gcs.html

THE KNOWLEDGE BASE FOR GCs

All experiments were performed on the Knowledge Base (KB) sample presented in the introduction, assuming that bona fide GCs are represented by sources selected according to the discussed color cuts. We used as features (columns of patterns) the following quantities:

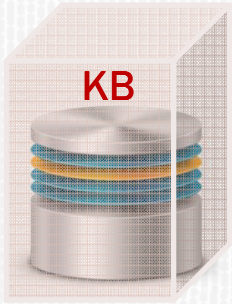
- **isophotal magnitude** (feature 1);
- **3 aperture magnitudes** (features 2–4) obtained through **circular apertures of radii 2, 6** and **20 arcsec**, respectively;
- **Kron radius**, **ellipticity** and the **FWHM** of the image (features 5–7);
- **4 structural parameters** (features 8–11) which are, respectively, the **central surface brightness**, the **core radius**, the **effective radius** and the **tidal radius**;
- One target value **ONLY** for training set: **class labels 0 (no GC), 1 (yes GC)**;

2100 training patterns

```
24.4753,26.7468,24.3789,0.0205,3.72,0.067,4.12,16.25,-0.1139,1.822,51.29,0
24.2342,26.5263,24.1632,0.0196,3.5,0.027,4.01,16.61,0.1321,1.856,35.38,0
23.1554,25.5964,23.1654,0.016,3.5,0.032,4.09,14.47,-0.3295,2.638,129.2,1
22.6316,25.3519,22.6808,0.0151,3.5,0.039,4.69,16.33,0.8065,5.002,80.45,1
22.4708,24.4951,22.4699,0.0216,3.5,0.066,3.45,12.81,-0.3912,-7.425,5.66,0
23.9033,27.5896,23.9168,0.0255,4.49,0.272,9.63,19.99,8.397,14.79,88.5,1
24.1972,26.4219,24.0978,0.0192,3.7,0.079,4.04,15.72,-0.1447,1.514,44.77,0
20.2423,22.1866,20.2963,0.017,3.5,0.03,3.23,6.68,-0.6999,-0.1492,1.899,0
23.5134,26.0983,23.511,0.0167,3.76,0.05,4.55,16.6,0.3777,4.75,105.8,1
...
```

THE GC EXPERIMENT

http://dame.dsf.unina.it/dame_gcs.html



2100 training patterns 11 features each

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

Support Vector Machines

GAME

5 supervised classifiers

- **Classification accuracy**: fraction of patterns (objects) correctly classified (either GCs or non-GCs), with respect to the total number of objects in the sample;
- **completeness**: fraction of objects correctly classified as GCs;
- **contamination**: fraction of non-GC objects erroneously classified as GCs

3 quality evaluation criteria

425 pruning experiments (85 for each classifier), by alternately removing subsets of features, in order to evaluate the minimal set of required (highly correlated) parameters.

- ✓ K-fold (k=10) cross validation to avoid overfitting;
- ✓ Cross entropy formula for statistical evaluation of training error (not simple MSE):

$$H(T,q) = -\sum_{i=1}^{N} \frac{1}{N} log_2 q(x_i)$$

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

Mon. Not. R. Astron. Soc. (2012) doi:10.1111/j.1365-2966.2011.20375.x

The detection of globular clusters in galaxies as a data mining problem

Massimo Brescia,[1] Stefano Cavuoti,[2] Maurizio Paolillo,[2⋆] Giuseppe Longo[2†] and Thomas Puzia[3]

[1]*INAF - Astronomical Observatory of Naples, via Moiariello 16, I-80131 Napoli, Italy*
[2]*Dipartimento di Scienze Fisiche, University Federico II, via Cinthia 6, I-80126 Napoli, Italy*
[3]*Department of Astronomy and Astrophysics, Pontificia Universidad Católica de Chile, 7820436 Macul, Santiago, Chile*

Accepted 2011 December 12. Received 2011 November 28; in original form 2011 October 10

QUALITY AND PRUNING RESULTS

Type of experiment	Missing features	Figure of merit	MLPQNA	GAME	SVM	MLPBP	MLPGA
Complete patterns	–						
		class.accuracy	**98.3**	82.1	90.5	59.9	66.2
		completeness	**97.8**	73.3	89.1	54.1	61.4
		contamination	**1.8**	18.7	7.7	42.2	35.1
No par. 11	11						
		class.accuracy	**98.0**	81.9	90.5	59.0	62.4
		completeness	**97.6**	79.3	88.9	56.1	62.2
		contamination	**1.6**	19.6	7.9	43.1	38.8
Only optical	8, 9, 10, 11						
		class.accuracy	93.9	86.4	90.9	70.3	76.2
		completeness	91.4	78.9	88.7	54.0	65.1
		contamination	5.9	13.9	8.0	33.2	24.6
Mixed	5, 8, 9, 10, 11						
		class.accuracy	94.7	86.7	89.1	68.6	71.5
		completeness	92.3	81.5	88.6	52.8	63.8
		contamination	5.0	16.6	8.1	37.6	30.1

- **isophotal magnitude** (feature 1);
- **3 aperture magnitudes** (features 2–4) obtained through **circular apertures of radii 2, 6** and **20 arcsec**, respectively;
- **Kron radius**, **ellipticity** and the **FWHM** of the image (features 5–7);
- **4 structural parameters** (features 8–11) which are, respectively, the **central surface brightness**, the **core radius**, the **effective radius** and the **tidal radius**;

GAME MODEL MATHEMATICS

Given a generic dataset with N features and a target *t*, *pat* a generic input pattern of the dataset, $pat = (f_1, \cdots, f_N, t)$ and *g(x)* a generic real function, the representation of a generic feature f_i of a generic pattern, with a polynomial sequence of degree *d* is:

$$G(f_i) \cong a_0 + a_1\, g(f_i) + \cdots + a_d\, g^d(f_i)$$

Hence, the k-th pattern (pat_k) with N features may be represented by:

$$Out(pat_k) \cong \sum_{i=1}^{N} G(f_i) \cong a_0 + \sum_{i=1}^{N} \sum_{j=1}^{d} a_j\, g^j(f_i) \qquad (1)$$

The target t_k, concerning to pattern pat_k, can be used to evaluate the approximation error of the input pattern to the expected value:

$$E_k = (t_k - Out(pat_k))^2$$

With NP patterns number (k = 1, ..., NP), at the end of the "forward" phase (batch) of the GA, we have NP expressions (1) which represent the polynomial approximation of the dataset.

In order to evaluate the fitness of the patterns as extension of (9) Mean Square Error (MSE) or Root Mean Square Error (RMSE) may be used:

$$MSE = \frac{\sum_{k=1}^{NP} (t_k - Out(pat_k))^2}{NP} \qquad RMSE = \sqrt{\frac{\sum_{k=1}^{NP} (t_k - Out(pat_k))^2}{NP}}$$

GAME MODEL FINAL EQUATIONS

We use the trigonometric polynomial sequence, given by the following expression,

$$g(x) = a_0 + \sum_{m=1}^{n} a_m \cos(m\,x) + \sum_{m=1}^{n} b_m \sin(m\,x)$$

$$NUM_{CHROMOSOMES} = (B \cdot N) + 1$$

B = 2

where *N* is the number of features of the patterns and *B* is a multiplicative factor that depends from the *g(x)* function, in the simplest case is just 1, but can arise to 3 or 4

$$NUM_{GENES} = (d \cdot B) + 1$$

where *d* is the degree of the polynomial.

With 2100 patterns, 11 features each, the expression for the single (*k-th*) pattern, using (1) with degree 6, will be:

$$Out(pat_k) \cong \sum_{i=1}^{11} G(f_i) \cong a_0 + \sum_{i=1}^{11}\sum_{j=1}^{6} a_j\, cos(j\, f_i) + \sum_{i=1}^{11}\sum_{j=1}^{6} b_j\, sin(j\, f_i)$$

for *k* = 1,…,2100.

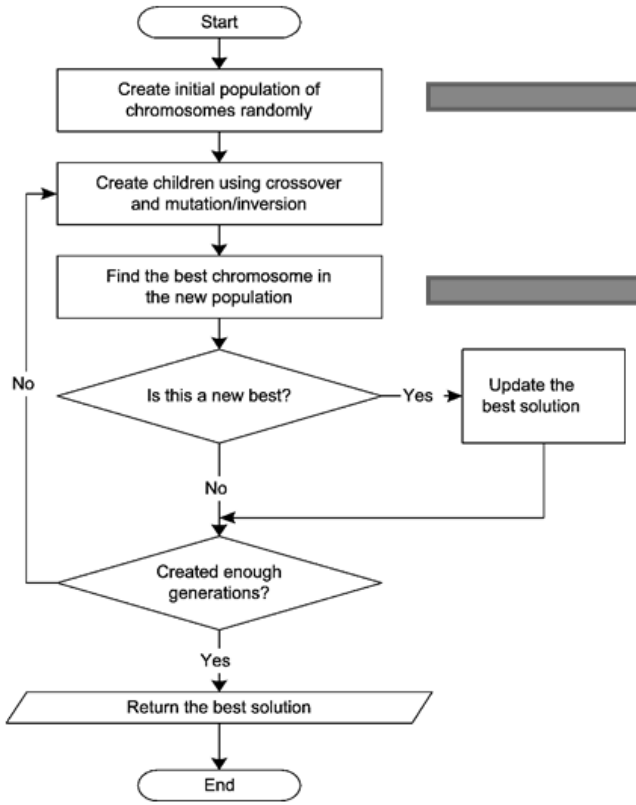
$$NUM_{CHROMOSOMES} = (2 \cdot 11) + 1 = 23$$

$$NUM_{GENES} = (6 \cdot 2) + 1 = 13$$

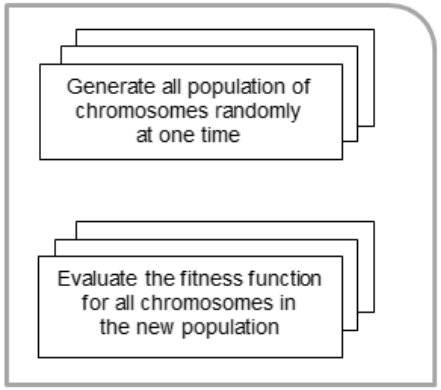
THE GAME ON GPU EXPERIMENT

The general-purpose GA has been internally designed for classification and regression problems

Genetic Algorithms are embarrassingly parallel (granularity + repetitive operations)



Serial on multi-core CPU



Parallel on many-core GPU

Needed faster execution to become scalable for MDS

Genetic Algorithm Modeling with GPU Parallel Computing Technology

S. Cavuoti[1], M. Garofalo[2], M. Brescia[3,1], A. Pescapè[2], G. Longo[1,4], G. Ventre[2]

1 Department of Physics, University Federico II, Via Cinthia 6, I-80126 Napoli, Italy

2 Department of Computer Engineering and Systems, University Federico II, Via Claudio 21, I-80125 Napoli, Italy

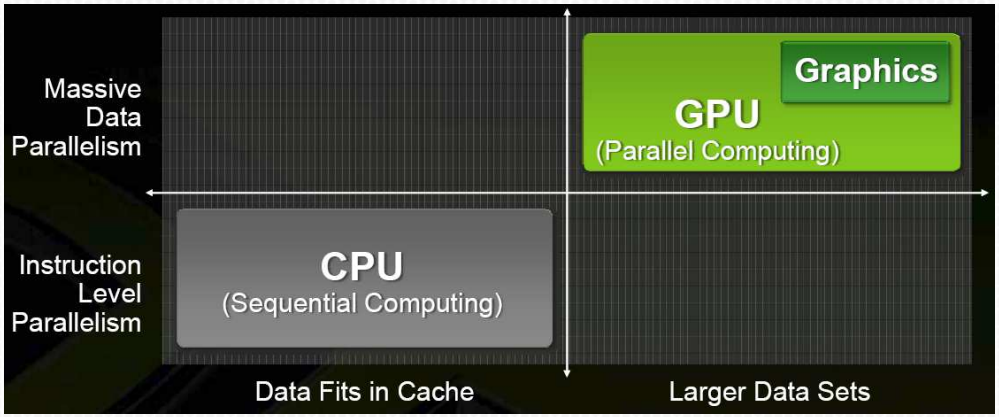
3 INAF, Astronomical Observatory of Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy

4 Visiting Associate, California Institute of Technology, Pasadena, CA 91125, USA

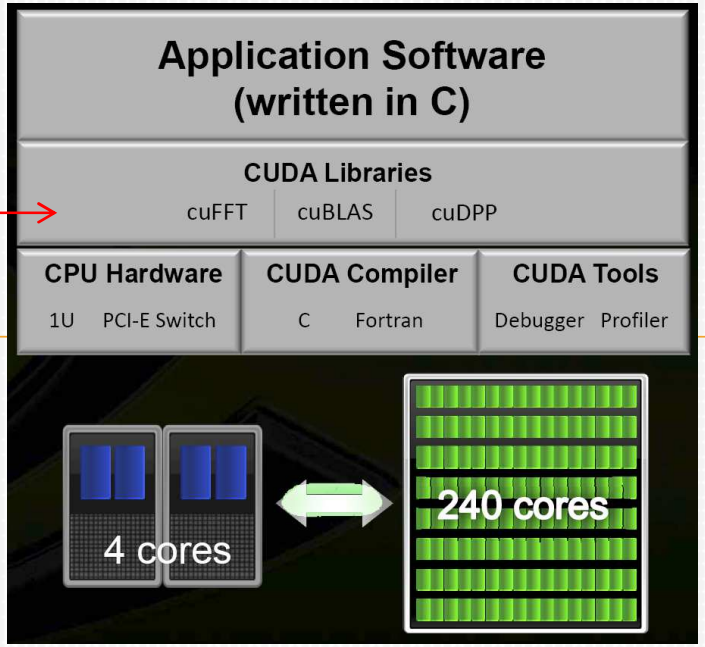
THE GPU TECHNOLOGY

The Graphical Processing Unit is specialized for compute-intensive, highly parallel computation (exactly what graphics rendering is about). So, more transistors can be devoted to data processing rather than data caching and flow control.

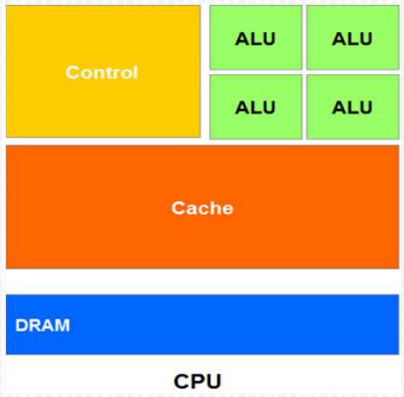
« GPU have evolved to the point where many real world apps are easily implemented on them and run significantly faster than on multi-core systems.
Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs »



DAMEWARE – GAME
FIRST PARALLEL PROGRAMMING EXPERIMENT

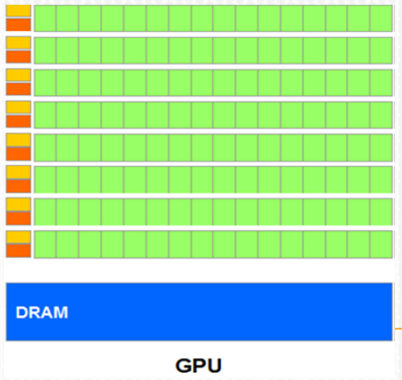


GPU VS CPU



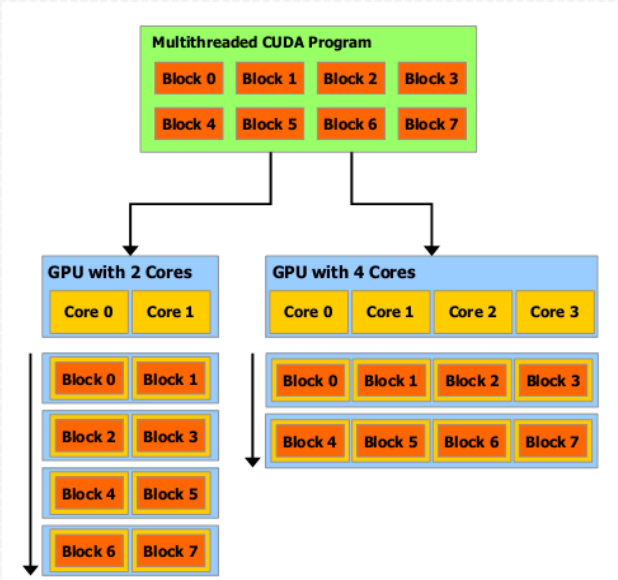
Multi-core CPU

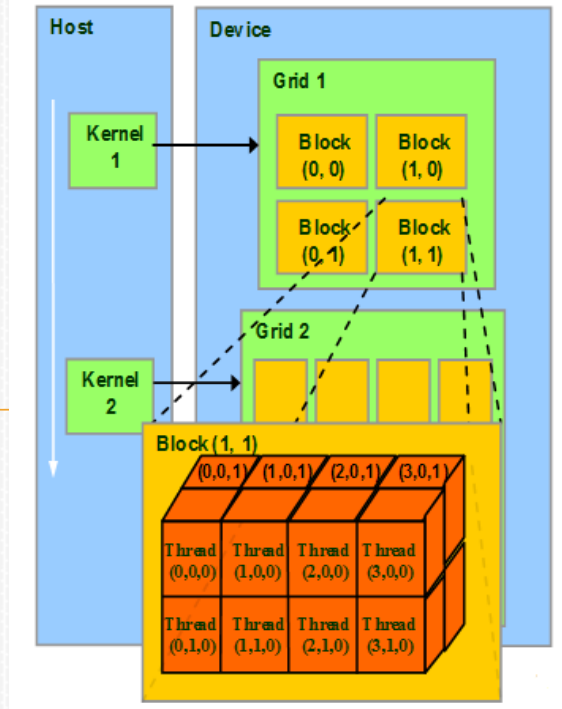
- Composed by few cores, designed to maximize the sequential code efficiency;
- Large cache memory to reduce latency time to access data and/or complex instruction execution;
- Sophisticate control logic to handle instruction flow (pipelining and multi-threading).



Many-core GPU

- Composed by many cores (hundreds), designed to execute parallel code;
- Memory structures with negligible access time to perform contemporary simple instructions;
- Simple control logic (the only bottleneck could be the communication with the CPU host);

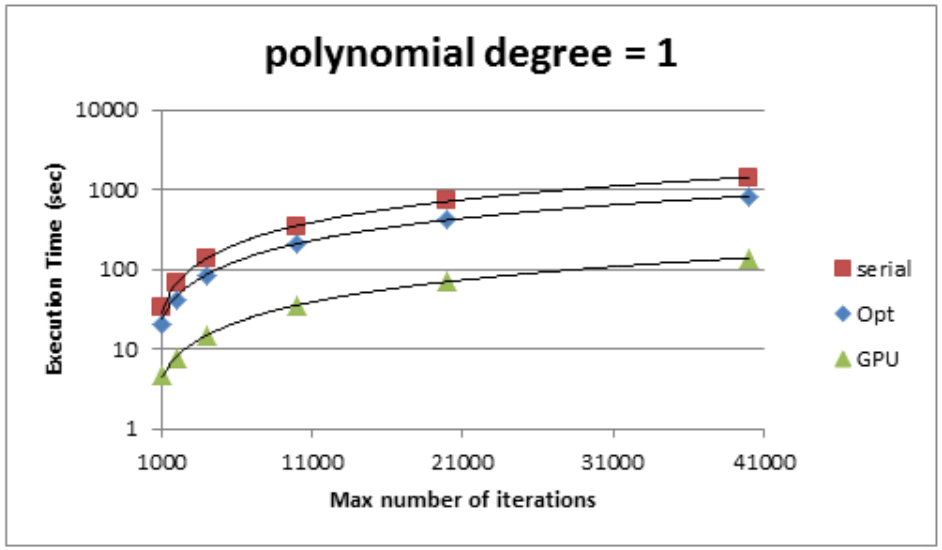


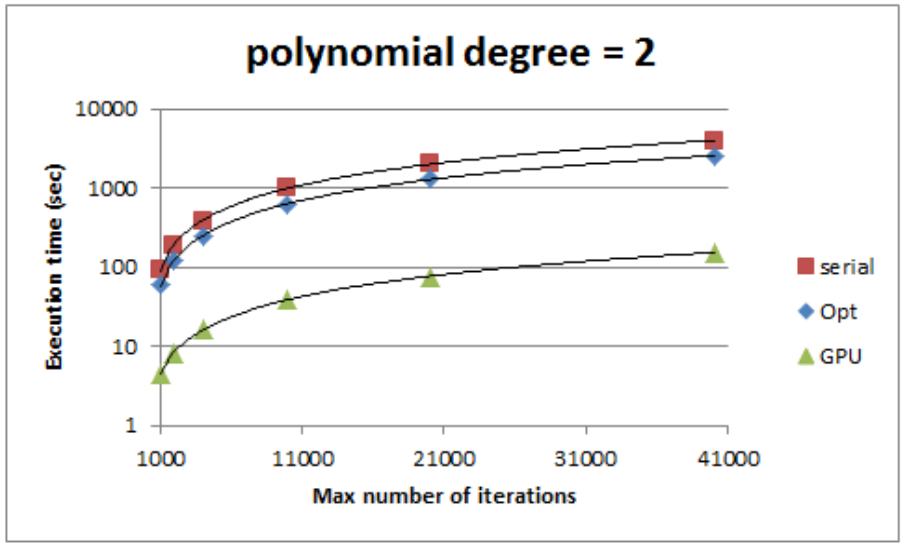


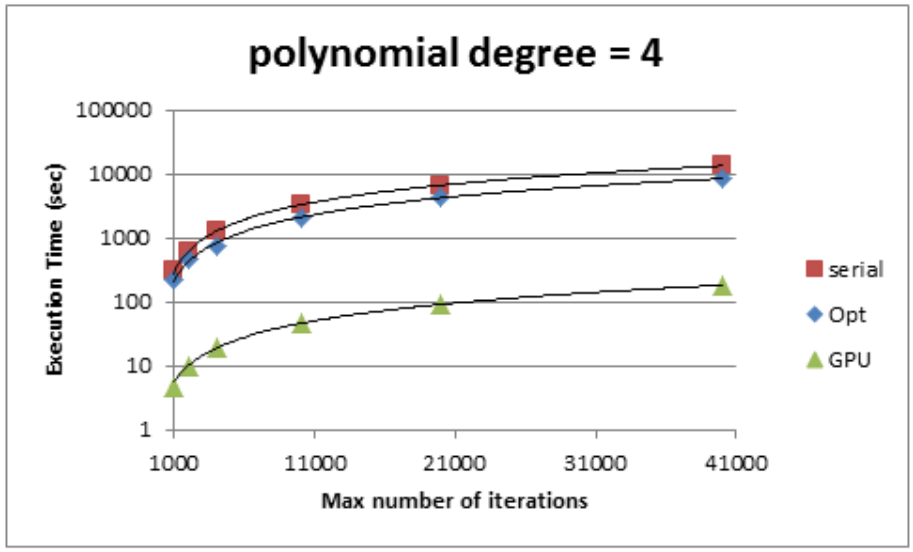
GAME HW PERFORMANCES

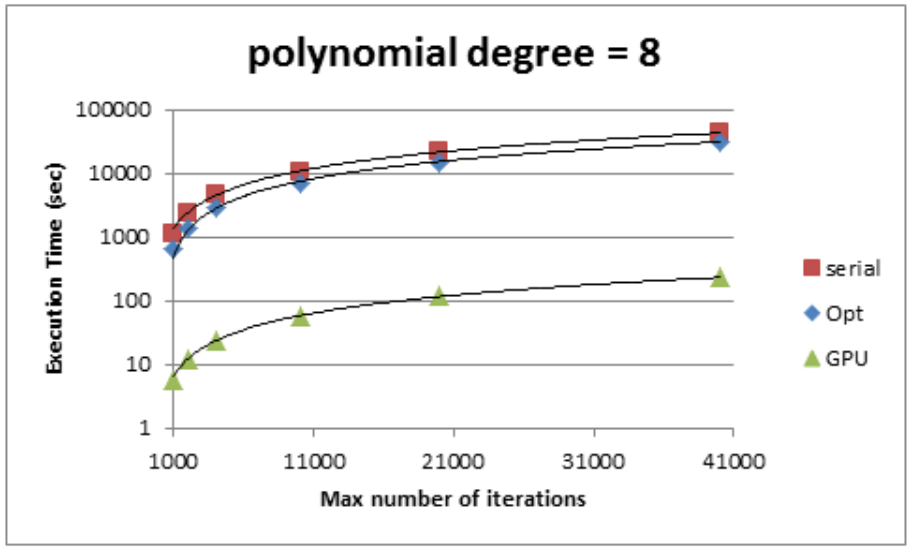
ID	CPU	GPU	Pol. Degree	DATASET	iterations	Exe time
1	2.0 GHz Intel i7 2630QM quad core		8	2100 patterns 11 features	40000	31092 sec (~9 h)
1		GeForce Tesla TM C1060 (240 cores)				231 sec (~0.064 h) 0,7% of CPU time
2	3.4 GHz Intel i7 2600 dual core					76000 sec (~21 h)
2		GeForce GTX 460 (336 cores)				165 sec (~0.046 h) 0,2% of CPU time
3	2.27 GHz Intel i5 M430 dual core					258400 sec (~72 h)
3		GeForce GT 320M (72 cores)				2489 sec (~0.691 h) 1% of CPU time

GAME GPU TESLA VS CPU I7





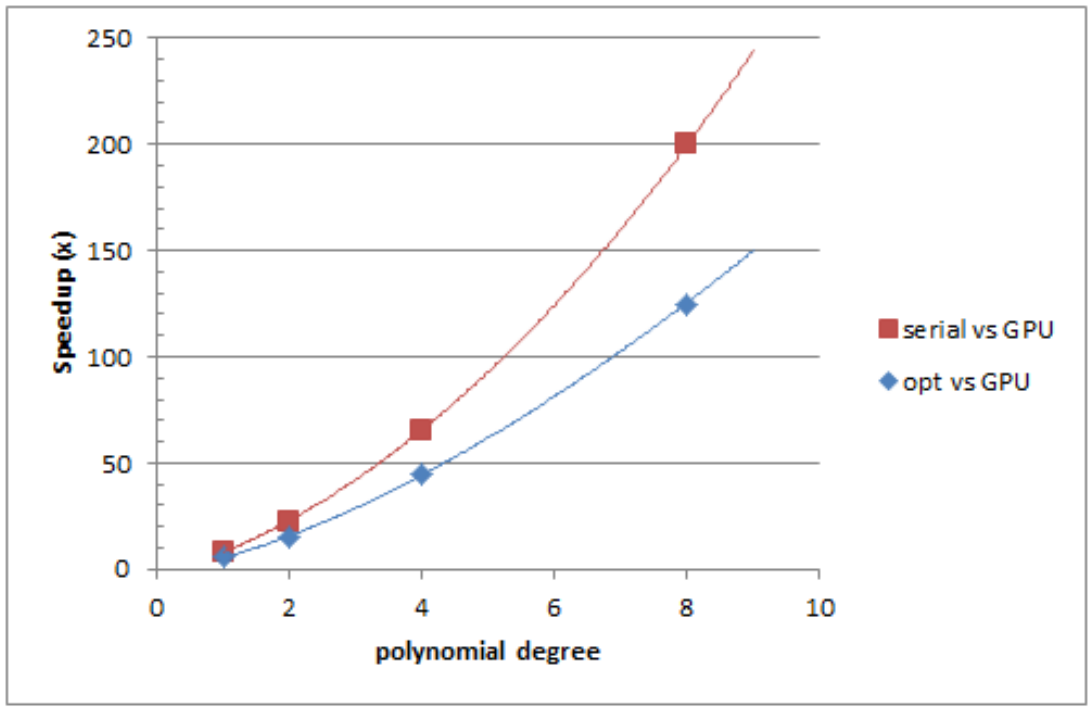




GAME GPU TESLA VS CPU I7

GPU Speedup		
degree	vs. Serial	vs. Opt
1	8x	6x
2	23x	16x
4	66x	45x
8	200x	125x

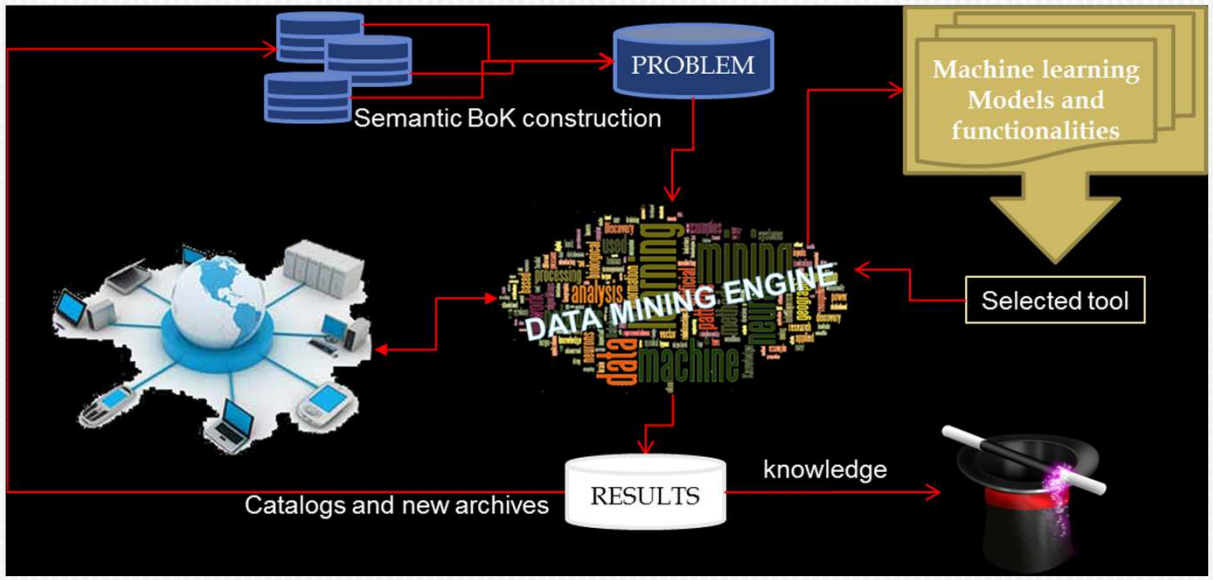
- The increase of the polynomial degree enhances the speedup difference.
- It results evident that the speedup is as much as highest is the complexity of the fitness function



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.

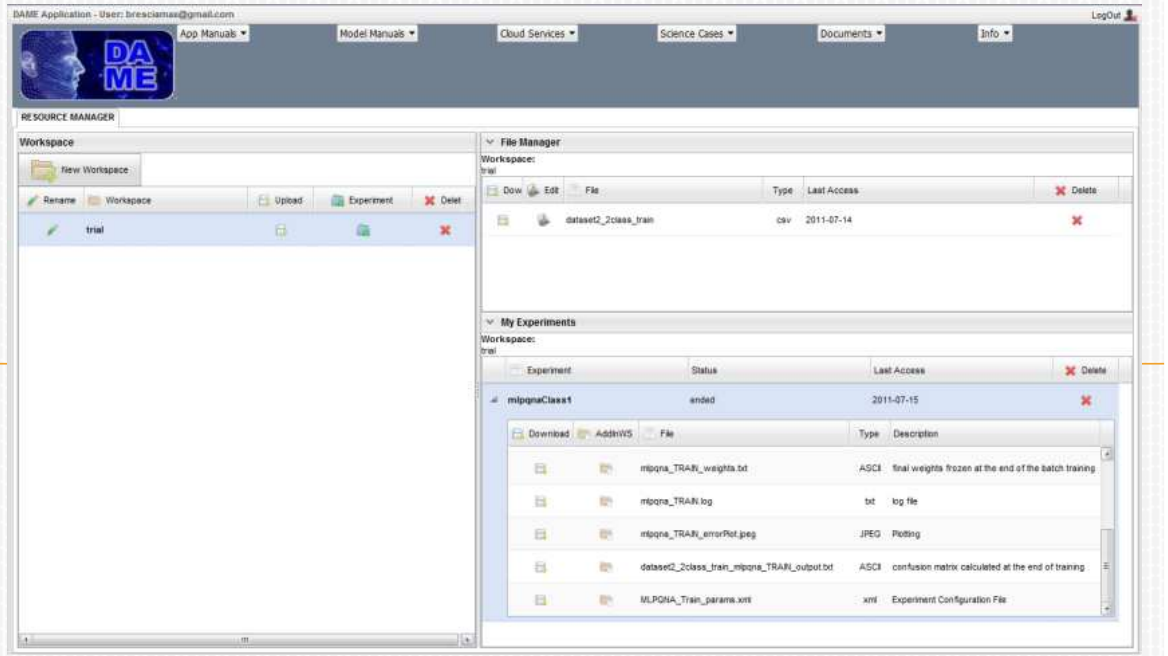
Intro page + manuals: http://dame.dsf.unina.it/beta_info.html

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

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

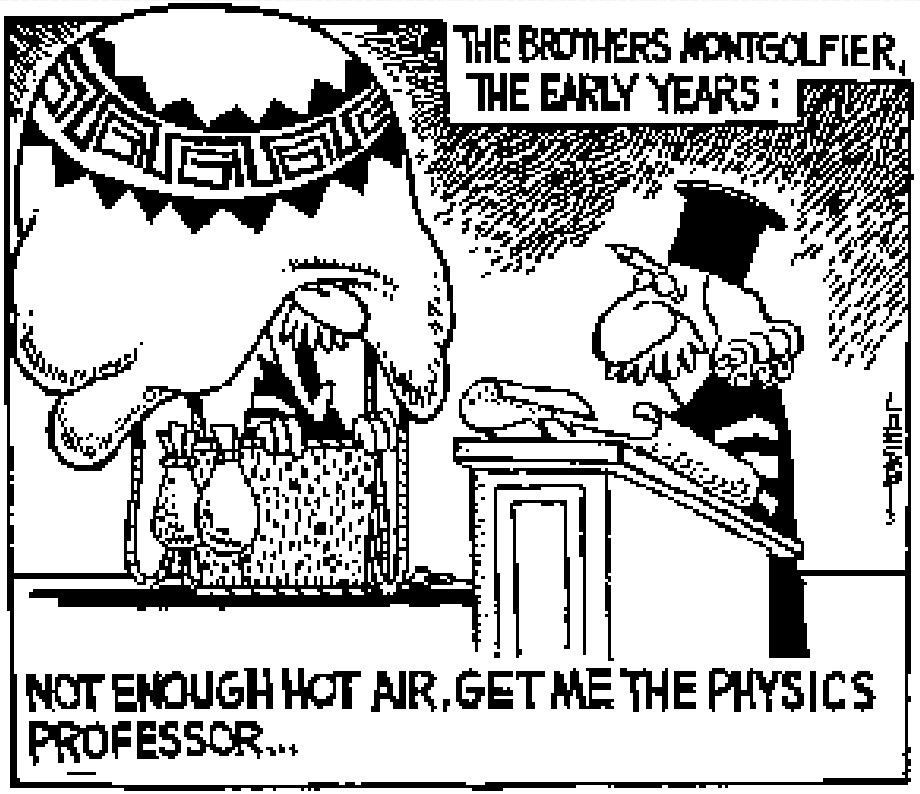
The release 1.0 will be deployed on a CLOUD, including GRID farm of S.Co.P.E.

- 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;



KIND INVITATION

You don’t have to believe our words, but follow St. Thomas rule: try us!



http://dame.dsf.unina.it