Astronomical data Mining
DAMEWARE and beyond

Giuseppe Longo
Università Federico II
Napoli (Italy)

M. Brescia – INAF OAC
G.S. Djorgovski – Caltech
S. Cavuoti – INAF UFII

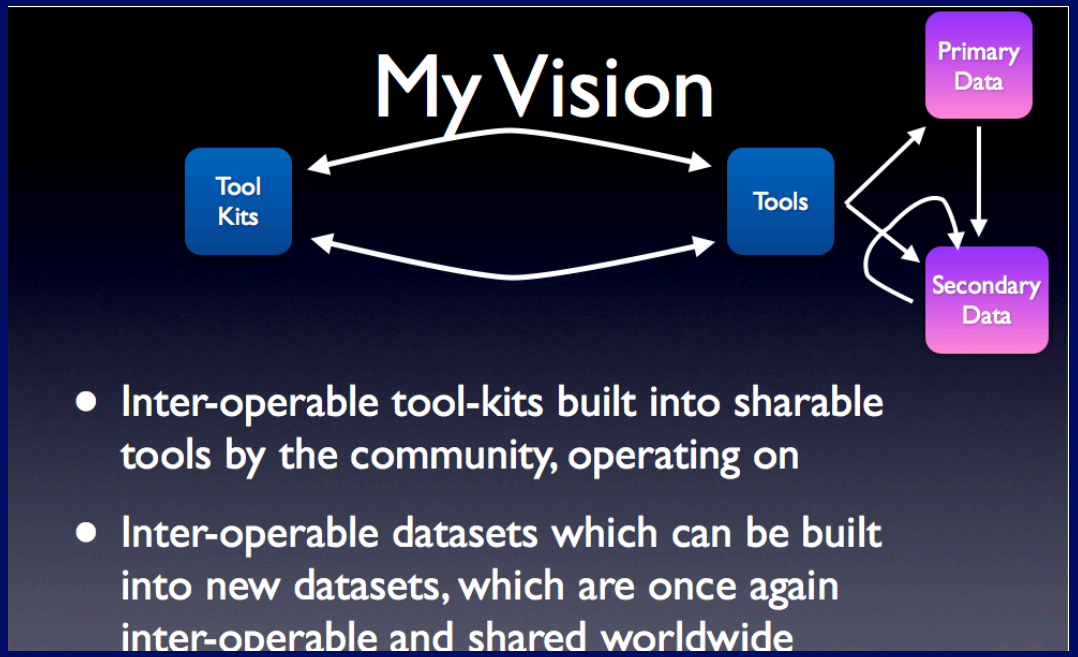
& the

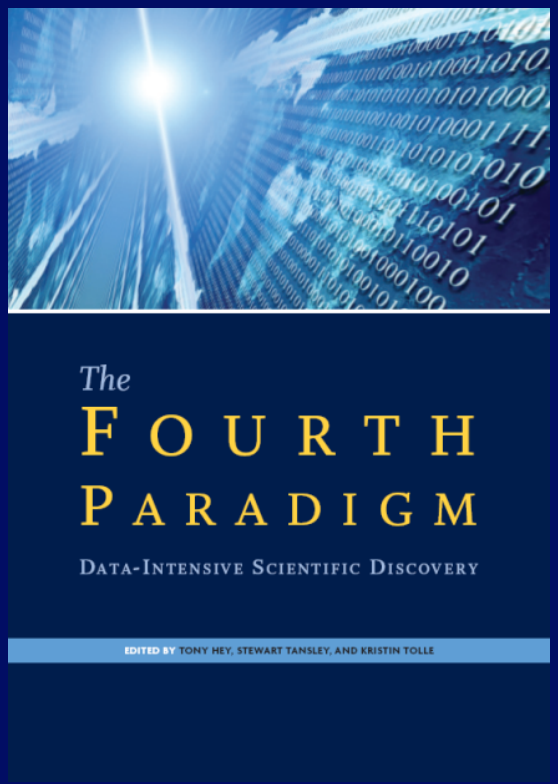
DAMEWARE people

“Prerequisites”

Brian Shmidt’s talk

1. Role of knowledge discovery in big data
2. Interoperability of tools



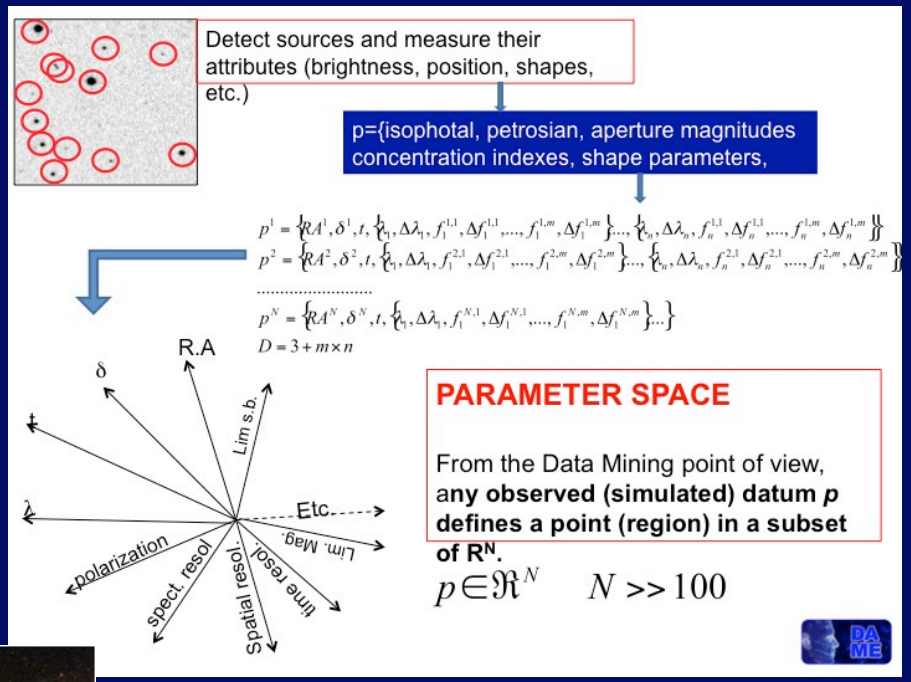


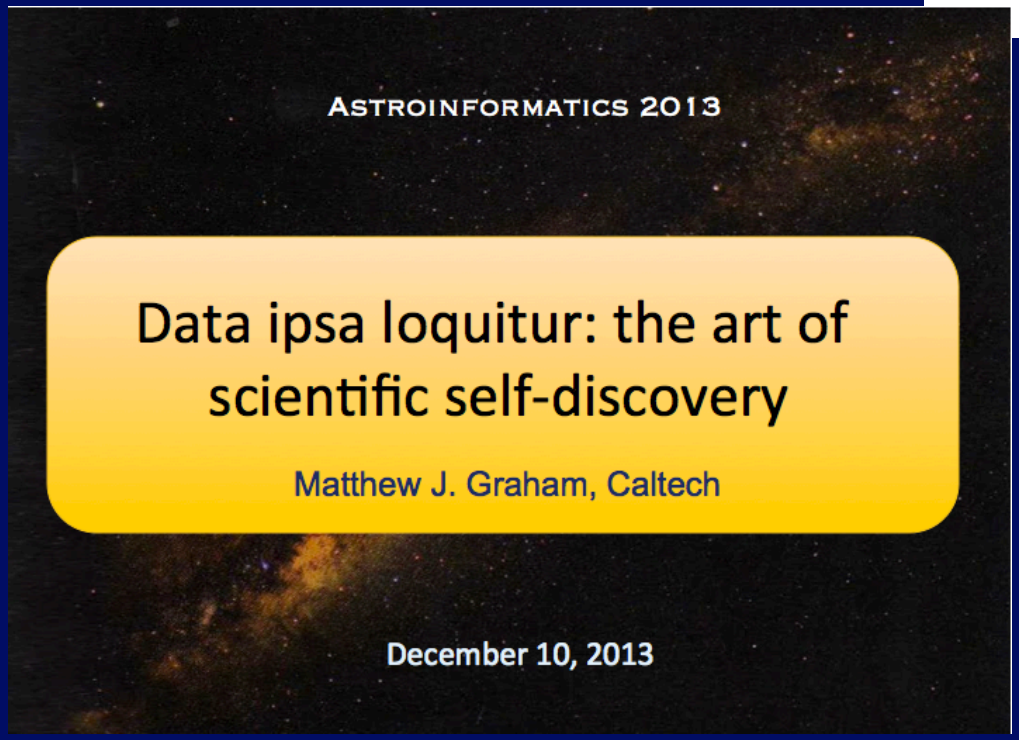
Tony Hey’s talk

Data science as the new path to discovery

George Djorgovsky talk

1. Parameter space concept
2. Role of DM in scientific discoveries of the future





Matthew Graham

Find the data in the haystack
Search for innovative methods for DM

As a result of large surveys, astronomy has entered an era where:

- ***Most data will never be seen by humans!***
  The need for data storage, network, database-related technologies standards, etc.
- ***Most knowledge hidden behind data complexity is potentially lost***
  Most (if not all) empirical relationships known so far depend on 3 parameters …. (e.g. fundamental plane of E galaxies and bulges). Simple universe or rather human bias?
- ***Most data (and data constructs) cannot be comprehended by humans directly!***
  The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery

Pucon, August 2013

Data Mining (machine lerning) is the only possible answer to **AID** humans in the scientific discovery process

What is data mining?

The search of useful information in large data sets or rather …..

What is data mining according to an expert….

There are known knowns,
There are known unknowns, and
There are unknown unknowns

Classification
Morphological classification of galaxies
Star/galaxy separation, etc.

Regression
Photometric redshifts

Clustering
Search for peculiar and rare objects,
Etc.

Two machine learning paradigms:

Supervised & Unsupervised

Functionality

Classification

Regression

Clustering

Feature selection

DM models

GAME S, C,R
MLPBP S, C,R
MLPGA S, C,R
MLPQNA S, C,R
SVM S, C,R
K-Means U, Cl
ESOM U, Cl
SOFM U, Cl
SOM U, Cl
PPS U, Cl, FS
…..
….

Use cases….

Use case

Functionality

Classification

Regression

Clustering

Feature selection

DM models

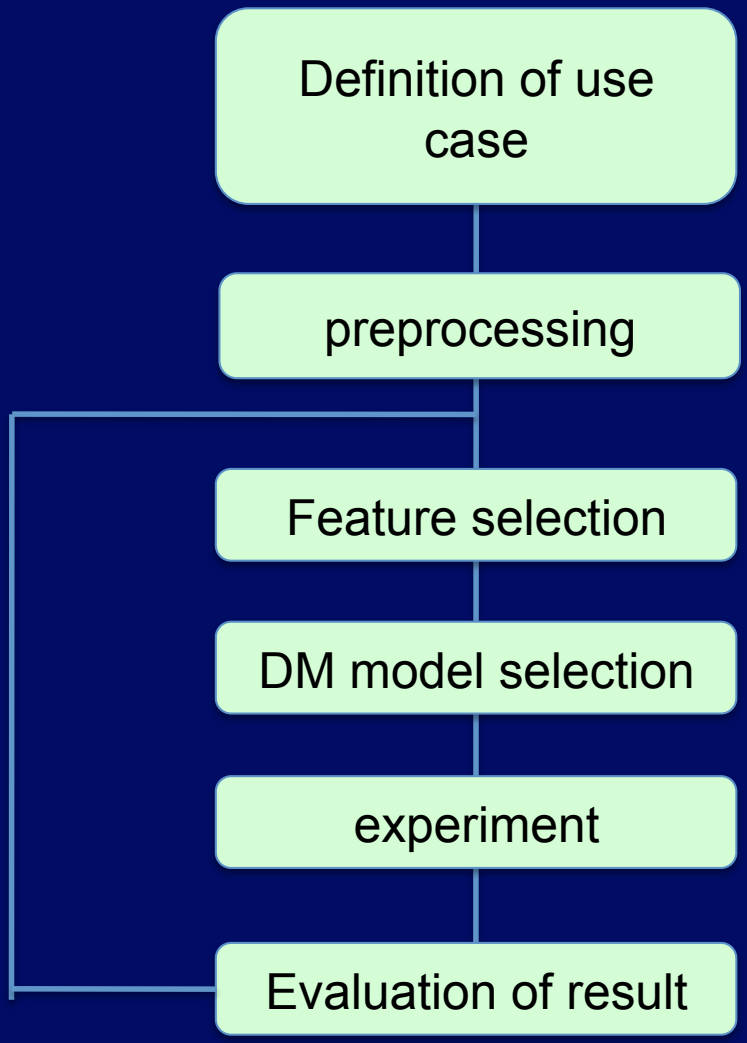
Decision trees, S, C, R
GAME S, C,R
MLPBP S, C,R
MLPGA S, C,R
MLPQNA S, C,R
SVM S, C,R
K-Means U, Cl
ESOM U, Cl
SOFM U, Cl
SOM U, Cl
PPS U, Cl, FS
…….
…….
……
Hundreds of models to choose from

Experiments

1-st
2-nd
3-rd
4-th
….

N-th

Effective DM (ML) requires complex workflows



….. iterated many many times in order to find the optimal, model, the optimal combination of parameters, etc….

- **Speed**
- **Computational power**
- **Expertise**

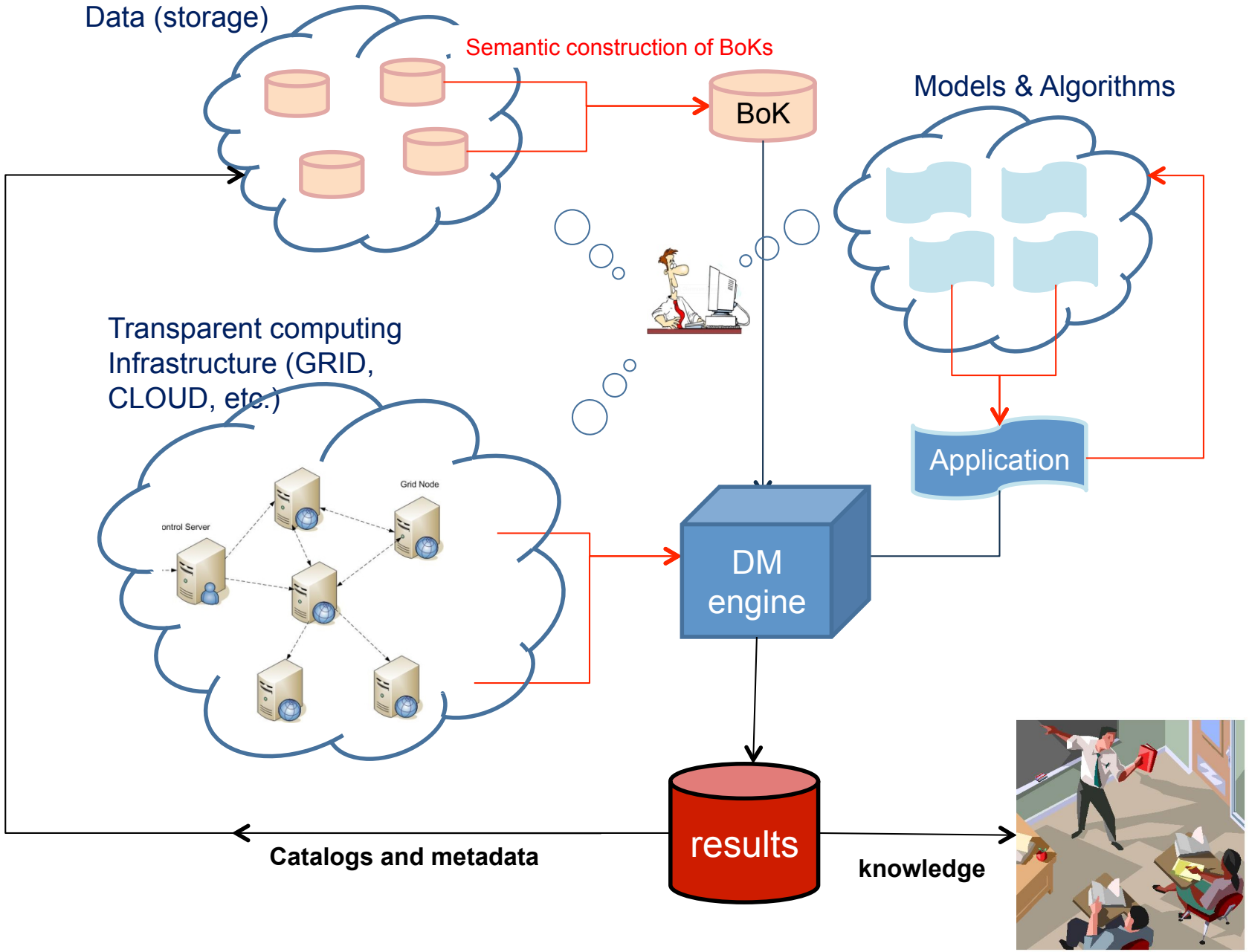
Exploration of PS with N >10^9, D>>100, K>10 Is anything but simple

N = no. of data vectors,
D = no. of data dimensions
K = no. of clusters chosen,
K_{max} = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

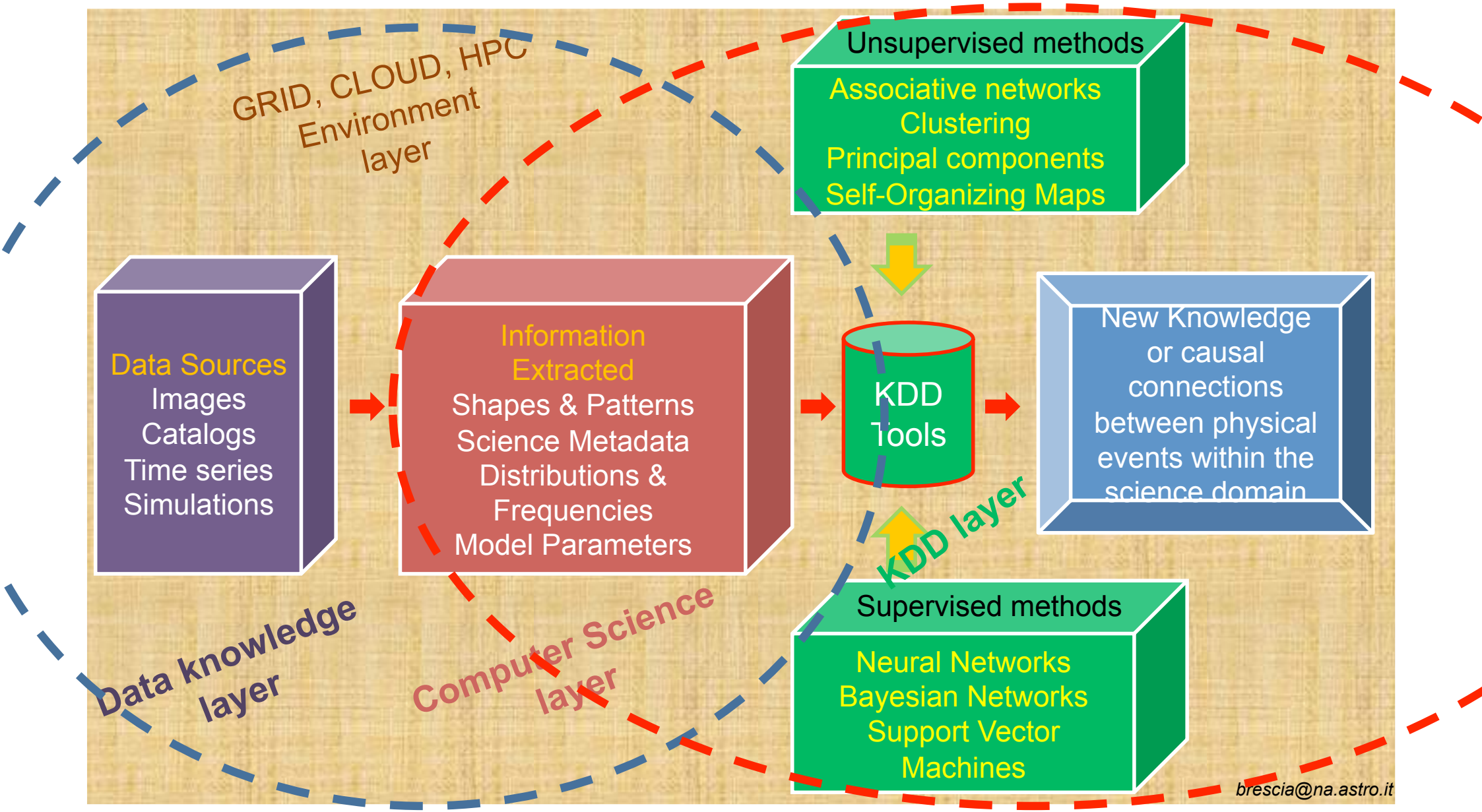
K-means: $K \times N \times I \times \mathbf{D}$
Expectation Maximisation: $K \times N \times I \times \mathbf{D^2}$
Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$
Correlations ~ N log N or N^2, ~ D^k ($k \geq 1$)
Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)
SVM > ~ $(NxD)^3$

Lots (…truly lots and lots…) of computing power

A break-down of an what you need for effective DM process



Vobs standards and infrastructure



brescia@na.astro.it

Many packages are available

Weka (astroweka)
Orange
Rapid Miner
Roots
R
Matlab
Mathematica
….

Many more NON packages are available

ANN2
Individual implementations,
etc…

Comparison in terms of user friendliness, scalability to big data, accuracy, etc…

Donalek et al. 2010, IVOA document

DAMEWARE TURNED OUT TO BE THE MOST SCALABLE PACKAGE

DAMEWARE

Is a web−based application (FREE AND OPEN TO THE PUBLIC) for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure

A joint effort between University Federico II, INAF−OACN & Caltech, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration

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

Science and management

PROGRAMMING STANDARDS

Technical documents (MORE THAN 200)

Template science cases

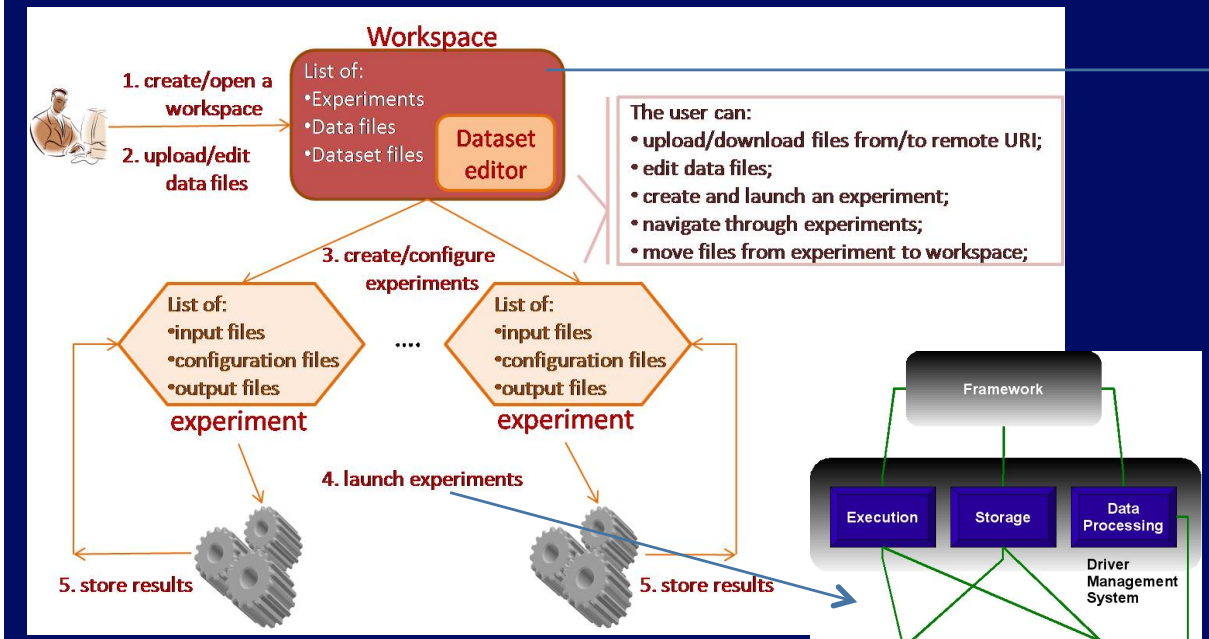
Tutorials

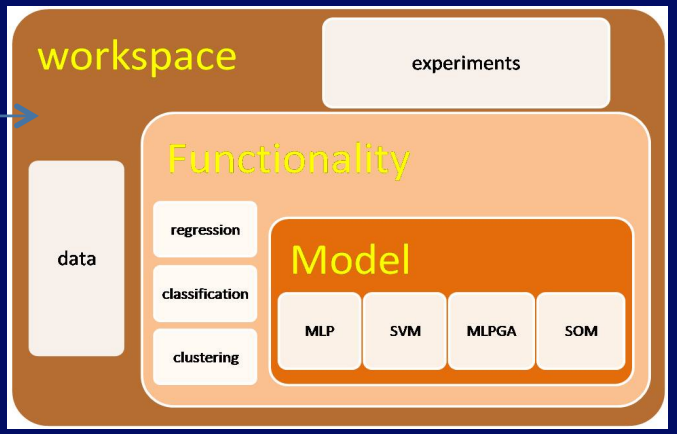
DAMEWARE

It is multi-disciplinary platform (astronomy, bioinformatics and medical diagnostics)

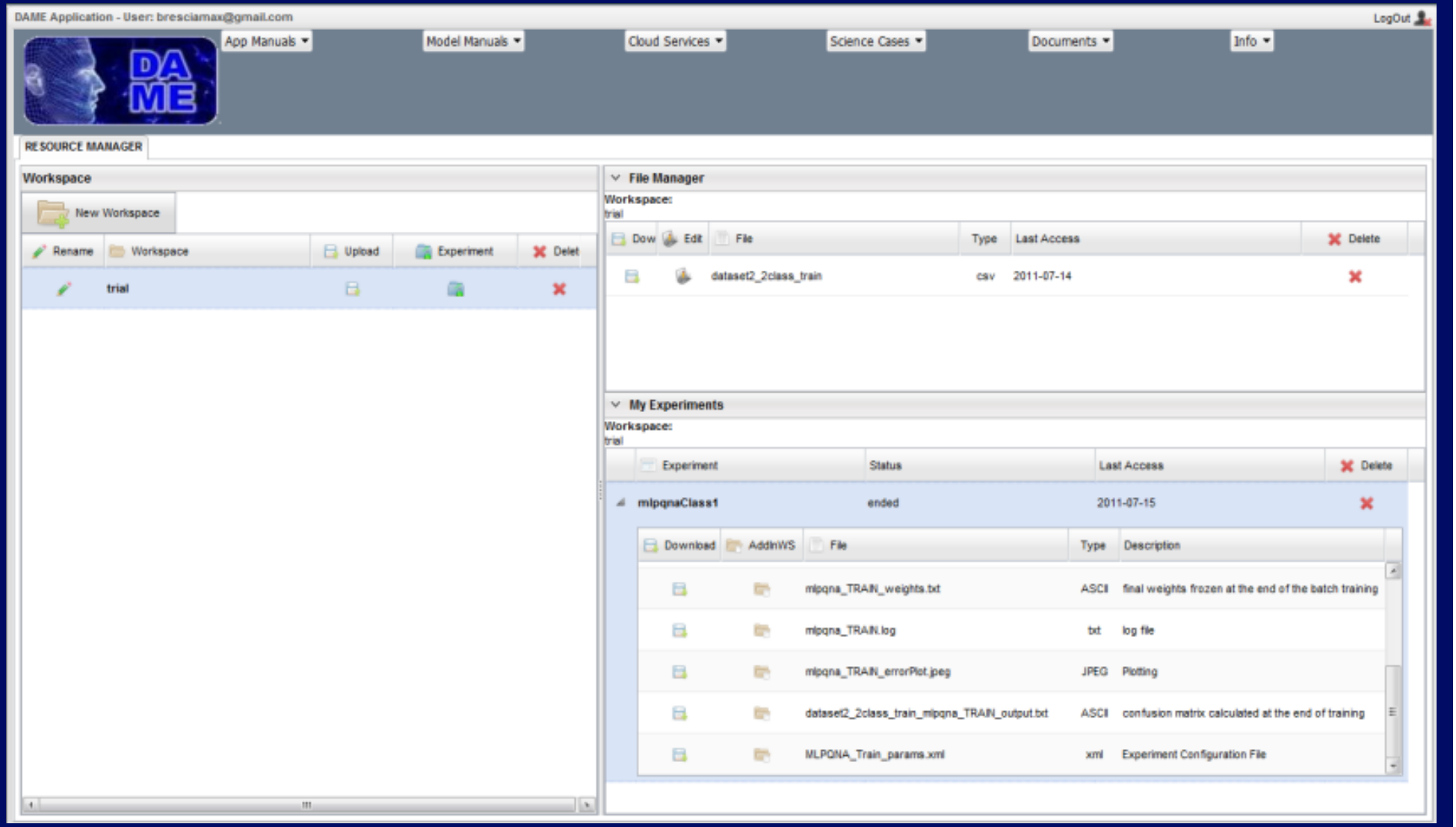
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

User can automatically plug-in his/her own algorithm and launch experiments through the Suite via a simple web browser





DAMEWARE the GUI

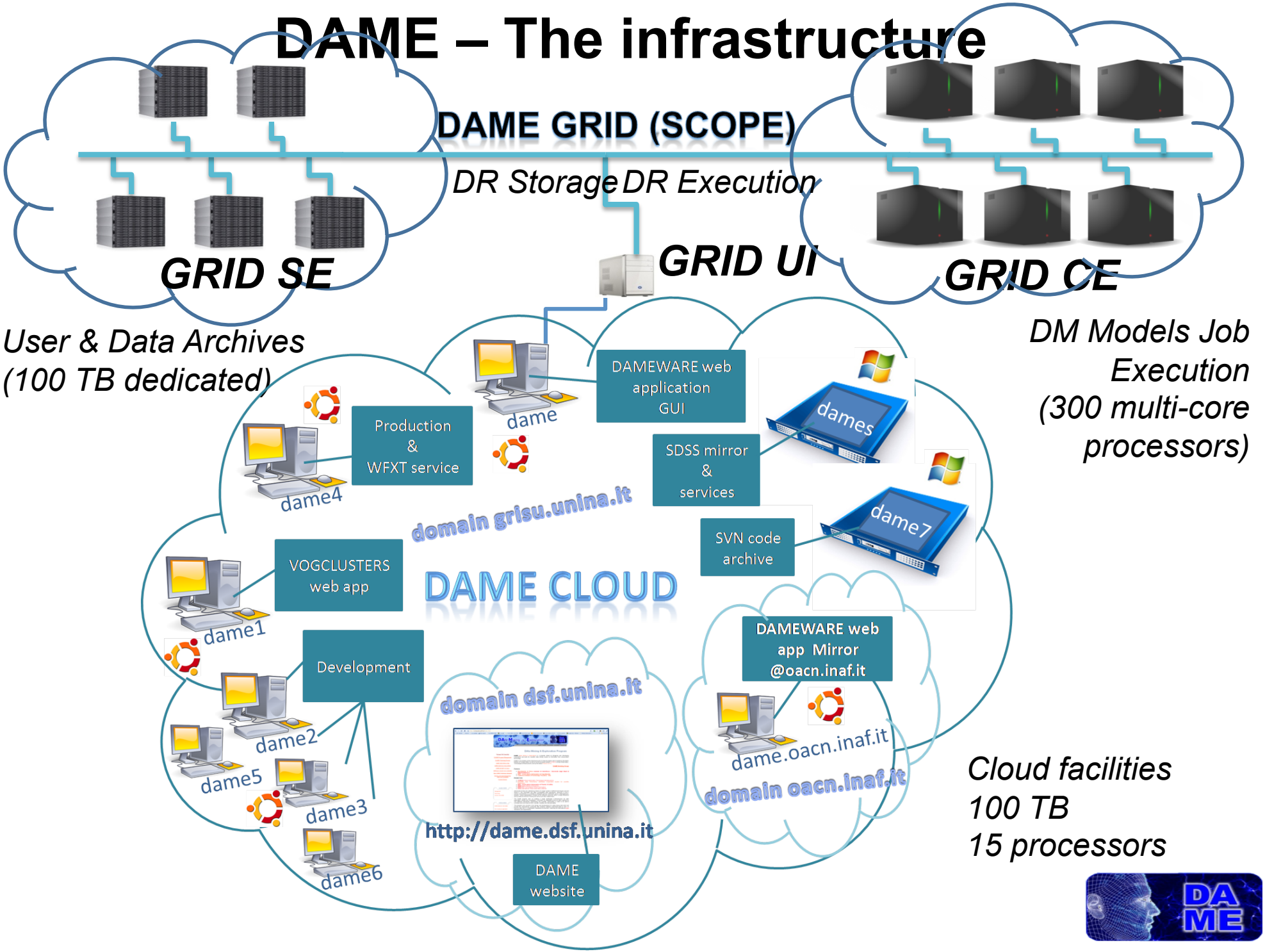


DM models in DAME (released)

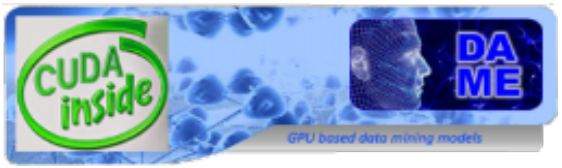
Model	Category	Functionality
MLPBP	Supervised	Classification, regression
FMLPGA	Supervised	Classification, regression
MLPQNA	Supervised	Classification, regression
SVM	Supervised	Classification, regression
ESOM	Unsupervised	Clustering
K-Means	Unsupervised	Clustering
SOFM	Unsupervised	Clustering
SOM	Unsupervised	Clustering
PPS	Unsupervised	Feature Extraction

Table 1: Data mining models and functionalities available in the DAMEWARE framework

DAME – The infrastructure

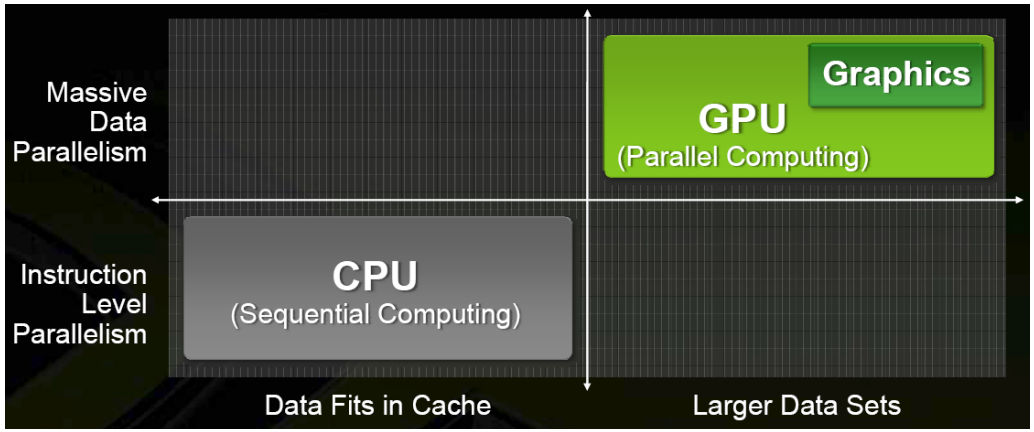


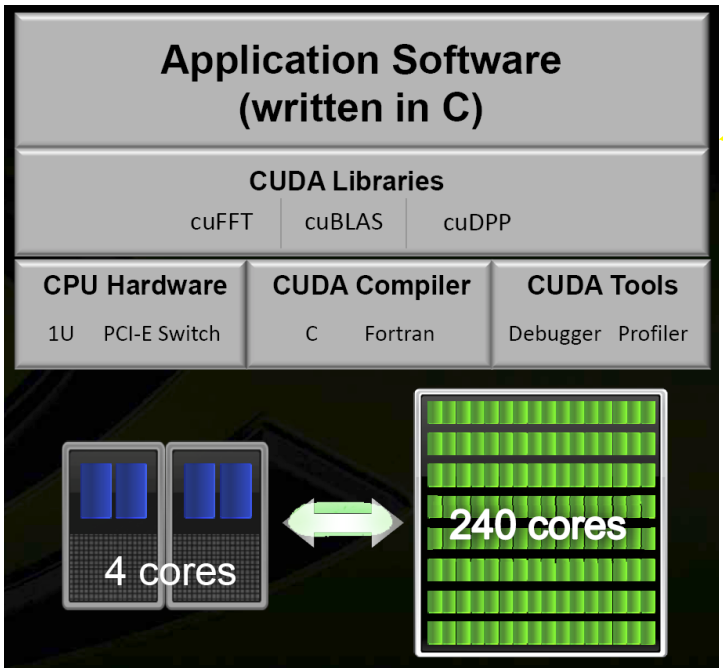
GPU technology ... sometime useful



The Graphical Processing Unit is specialized for compute-intensive, highly parallel computation (exactly what graphics rendering is about).

« 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.»





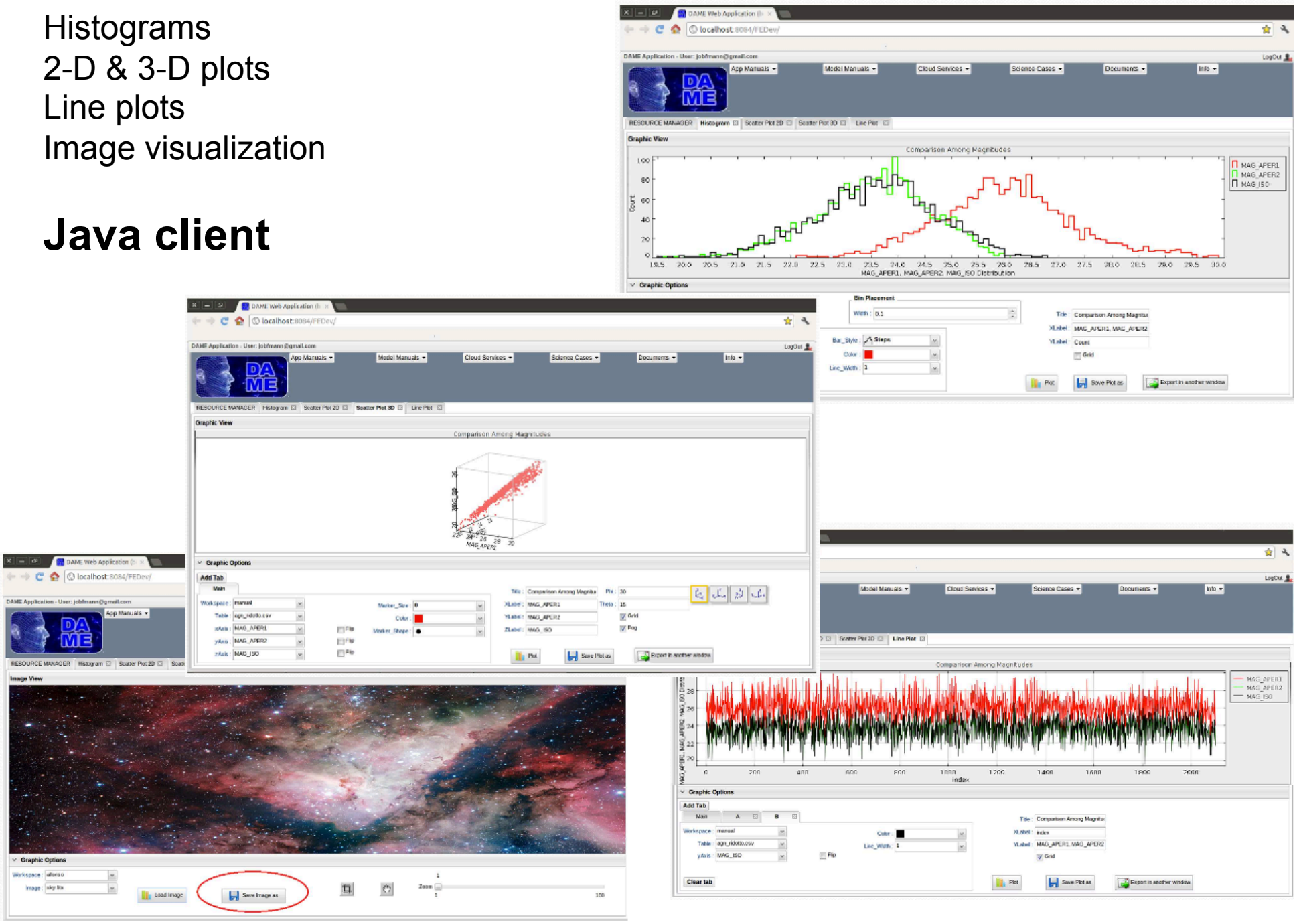
DAME - GAME
Genetic Algorithm Mining Experiment

GAME is a pure genetic algorithm developed in order to solve supervised problems of regression or classification, able to work on Massive Data Sets (MDS).

Graphical capabilities in DAMEWARE

Histograms
2-D & 3-D plots
Line plots
Image visualization

Java client



AGN identification and classification

Photometric classification of emission line galaxies with Machine Learning methods, Cavuoti et al., 2013, MNRAS, submitted

Star/Galaxy separation

The detection of globular clusters as a data mining problem, Brescia et al., 2012, MNRAS, 421, 1155-1165 (arXiv:1110.2144)

GPUs for astrophysical data mining. A test on the search for candidate globular clusters in external galaxies. S. Cavuoti, et al., New Astronomy, april 20, 2013, http://dx.doi.org/10.1016/j.newast.2013.04.004 (astro-ph: 1304.0597)

Photometric redshifts

Mining the SDSS archive. I. Photometric redshifts in the nearby universe, D'Abrusco, Logno G., Walton N., 2007, ApJ, 663, 752

Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation , O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

Photometric redshifts with Quasi Newton Algorithm (MLPQNA) Results in the PHAT1 context, Cavuoti et al. 2012, , Astronomy and Astrophysics 546, 13, (ArXiv:1206.0876)

Photometric redshifts for quasars in multiband surveys, M. Brescia et al., 2013, ApJ, 772, 140 (astro-ph: 1305.5641)

Inside catalogs: a comparison of source extraction software, M. Annunziatella, et al., 2012, PASP, 125, 68 (astro-ph:1212.0564).

Other

Astroinformatics, data mining and the future of astronomical research, M. Brescia & G. Longo, 2012, invited to appear in proceed. of IFDT2 - 2nd International conference frontiers on diagnostic technologies (arXiv:1201.1867)

CLASPS: a new methodology for knowledge extraction from complex astronomical data sets, R. D'Abrusco, G. Fabbiano, S.G. Djorgovski, C. Donalek, O. Laurino & G. Longo, 2012, ApJ, 755, 92 (ArXiv: 1206.2919)

MLPQNA in DAME
Use Cases

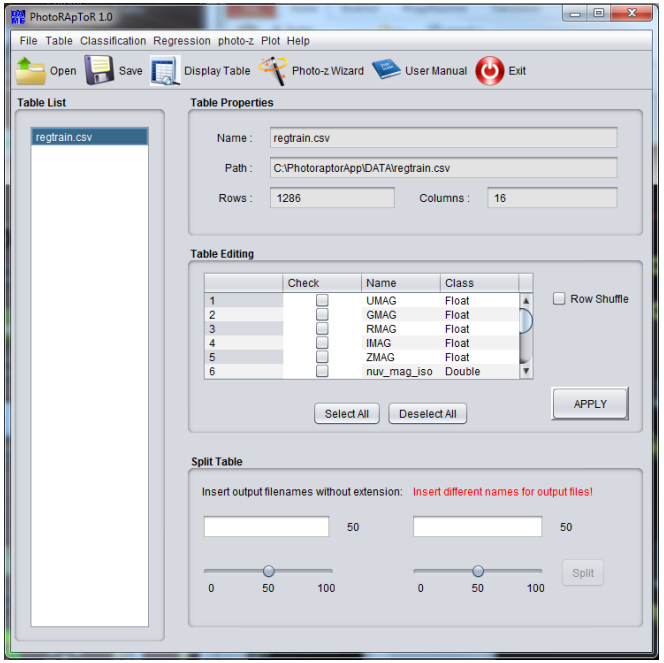


Photo-z regression

CLASH-VLT: The mass, velocity-anisotropy, and pseudo-phase-space density profiles of the z = 0.44 galaxy cluster MACS J1206.2-0847,
Biviano et al. 2013, A&A 558, A1

Photometric Redshifts for Quasars in Multi-band Surveys,
Brescia et al. 2013, ApJ 772, 2, 140

Photometric redshifts with the quasi Newton algorithm (MLPQNA)
Results in the PHAT1 contest,
Cavuoti et al. 2012, A&A 546, A13

Classification

Photometric classification of emission line galaxies with machine-learning methods,
Cavuoti et al. 2013, MNRAS (in press)

The detection of globular clusters in galaxies as a data mining problem, Brescia et al. 2012, MNRAS 421, 2

Current Applications in other fields

Medical diagnosis of alzhaimer
(S. Cocozza et al.)
Brain tomography analysis
(Bellotti M. et al.)
Real time classification of ethernet data flows
(G. Ventre et al.)
DATA Mining Quality Tools for the Euclid Mission

Etc…

During its commissioning period, ended in August 2013,

100 scientists from 27 countries registered as users and performed many different experiments.

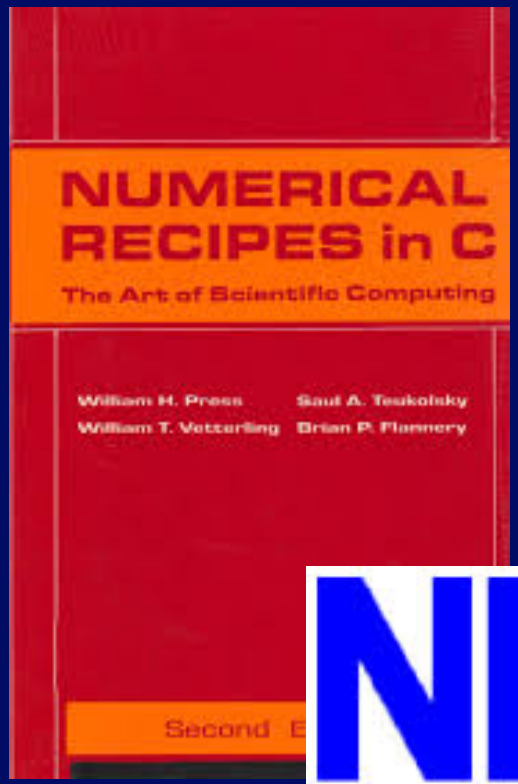
In the same period, the project web site hosted ca.12.000 independent accesses.

Algorithms

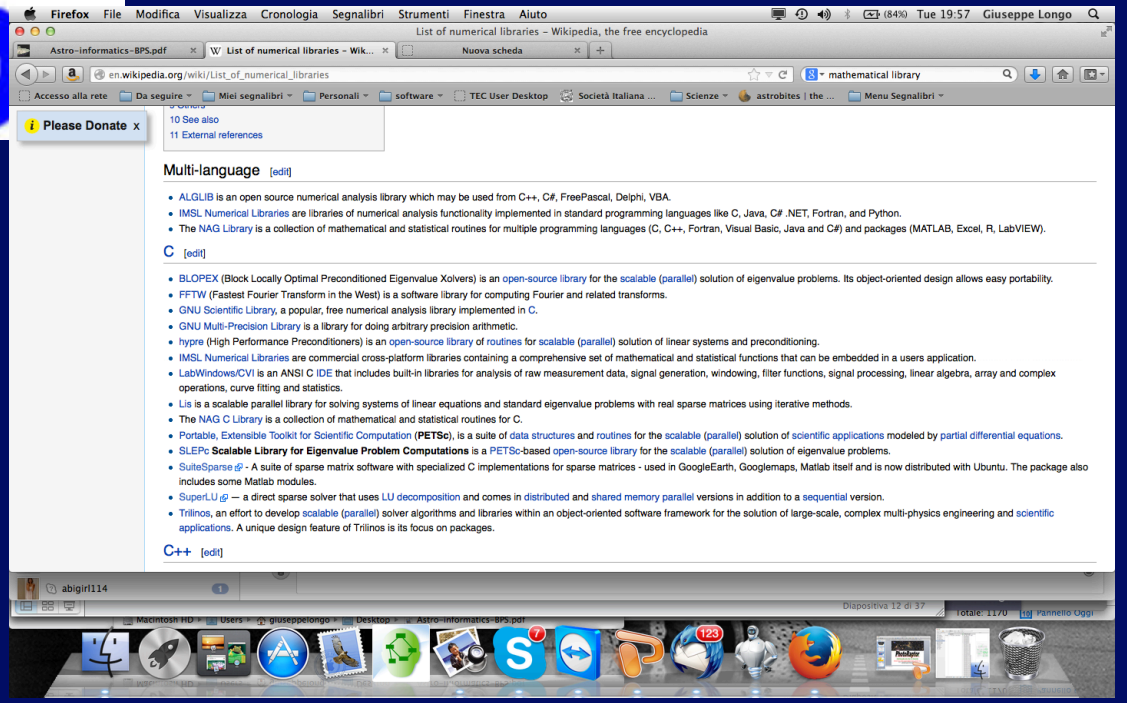
Restricted choice of algorithms (MLPs, SVM, Kernel methods, Genetic algoritms (few models), K Means, PPS, SOM, random forest…)

Astronomers know little statistics, forget about SPR, DM, etc…
Just a few astronomers go beyond the introductory chapters of the Bishop.

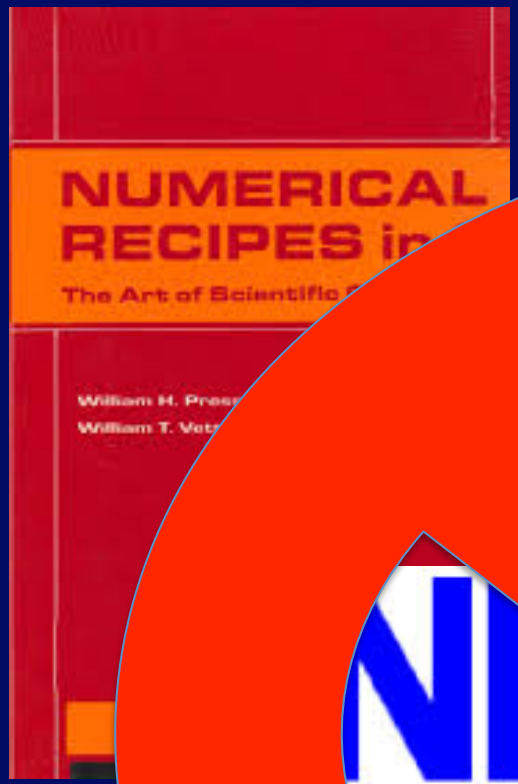
Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	Y
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	-----	
Noise removal *(grav. waves, pixel lensing, images)*	-----	
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSO's)*	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	Time domain	
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects	Y
Planetary studies (asteroids)	Planetary studies (asteroids)	Y
Solar activity	Solar activity	Y
Interstellar magnetic fields	----	
Stellar evolution models	----	

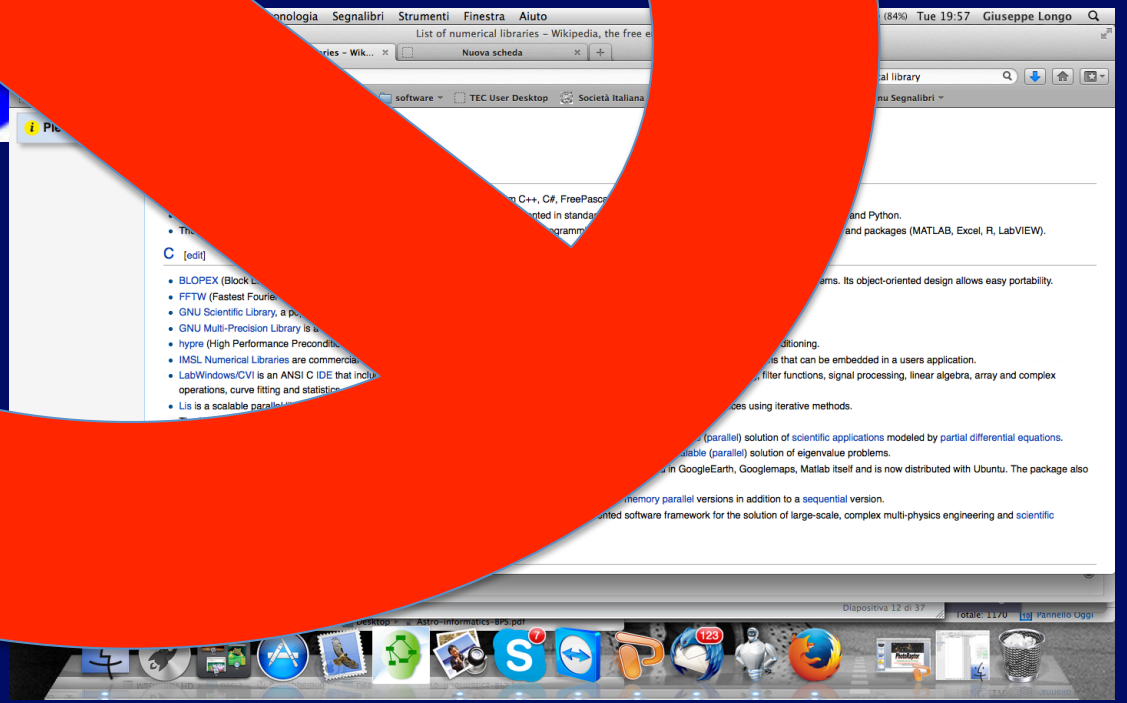


But astronomers have a lot of fine tuned algorithms which are unbeatable on specific tasks

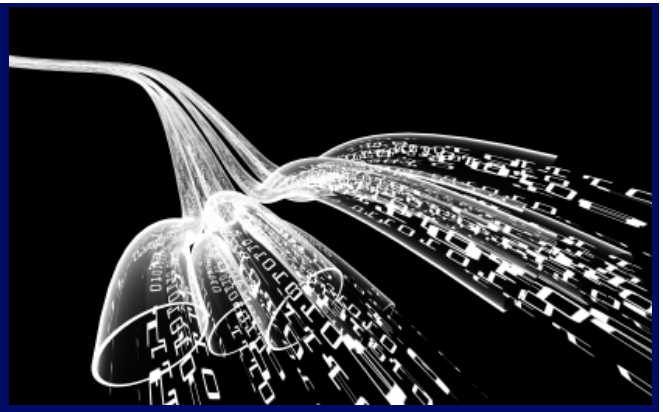


But astronomers have a lot of fi… algorithms which are …cific tasks





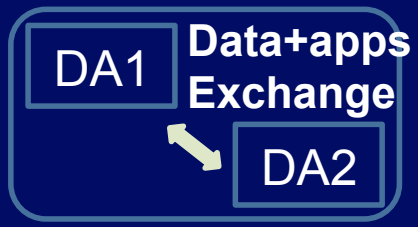
Moving programs not data: The 1-st true bottle neck



Data Mining + Data Warehouse =
Mining of Warehouse Data

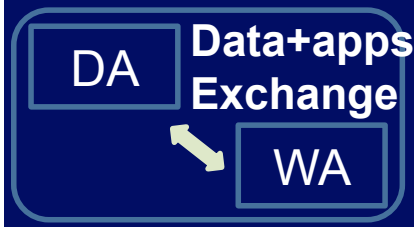
- For organizational learning to take place, data from must be gathered together and organized in a consistent and useful way – hence, Data Warehousing (DW);
- DW allows an organization to remember what it has noticed about its data;
- Data Mining apps should be interoperable with data organized and shared between DW.

Interoperability scenarios



Full interoperability between DA (Desktop Applications)

Local user desktop fully involved (requires computing power)

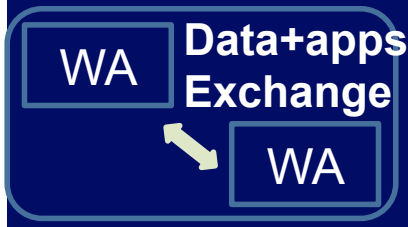


Full WA → DA interoperability

Partial DA → WA interoperability (such as remote file storing)

MDS must be moved between local and remote apps

user desktop partially involved (requires minor computing and storage power)



Except from URI exchange, no interoperability and different accounting policy

MDS must be moved between remote apps (but larger bandwidth)

No local computing power required

The Lernaean Hydra KDD

After a certain number of such iterations…

The scenario will become:

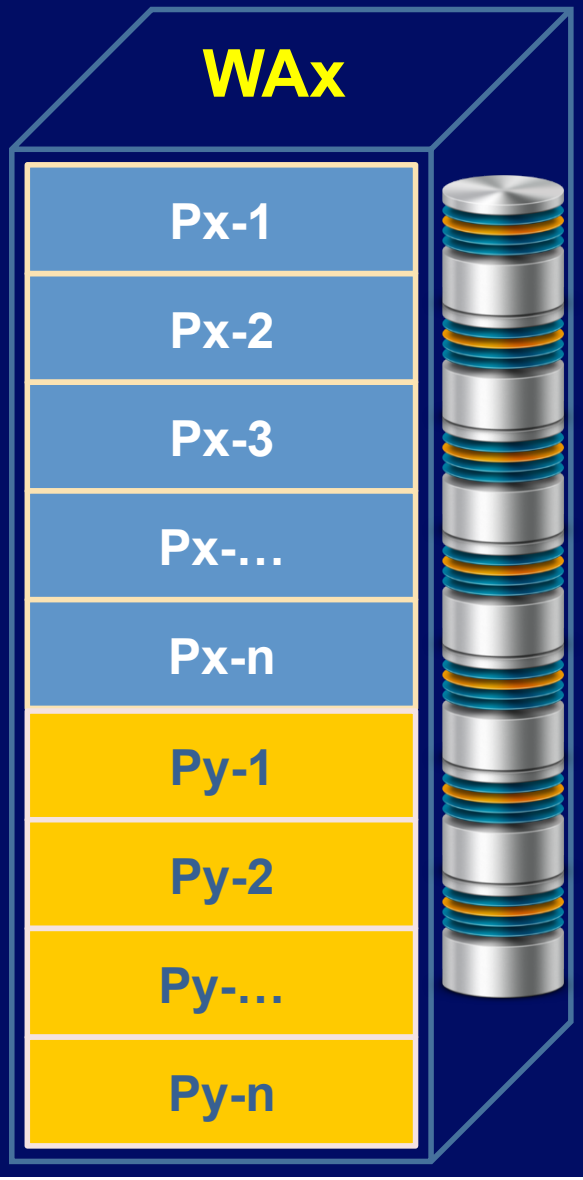
No different WSs, but simply one WS with several sites (eventually with different GUIs and computing environments)

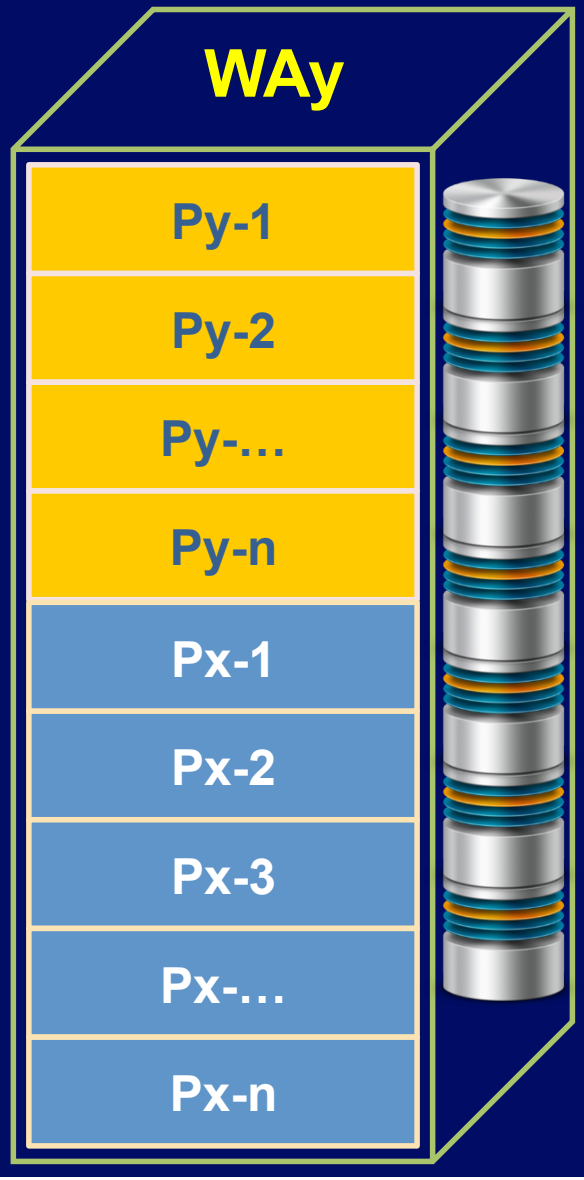
All WS sites can become a mirror site of all the others

The synchronization of plugin releases between WSs is performed at request time

Minimization of data exchange flow (just few plugins in case of synchronization between mirrors)

YES MDS!





Advantages:

- No need to move data
- Large computing power available server side
- Re-usability of codes (acts as a SW repository)
- Large variety of models

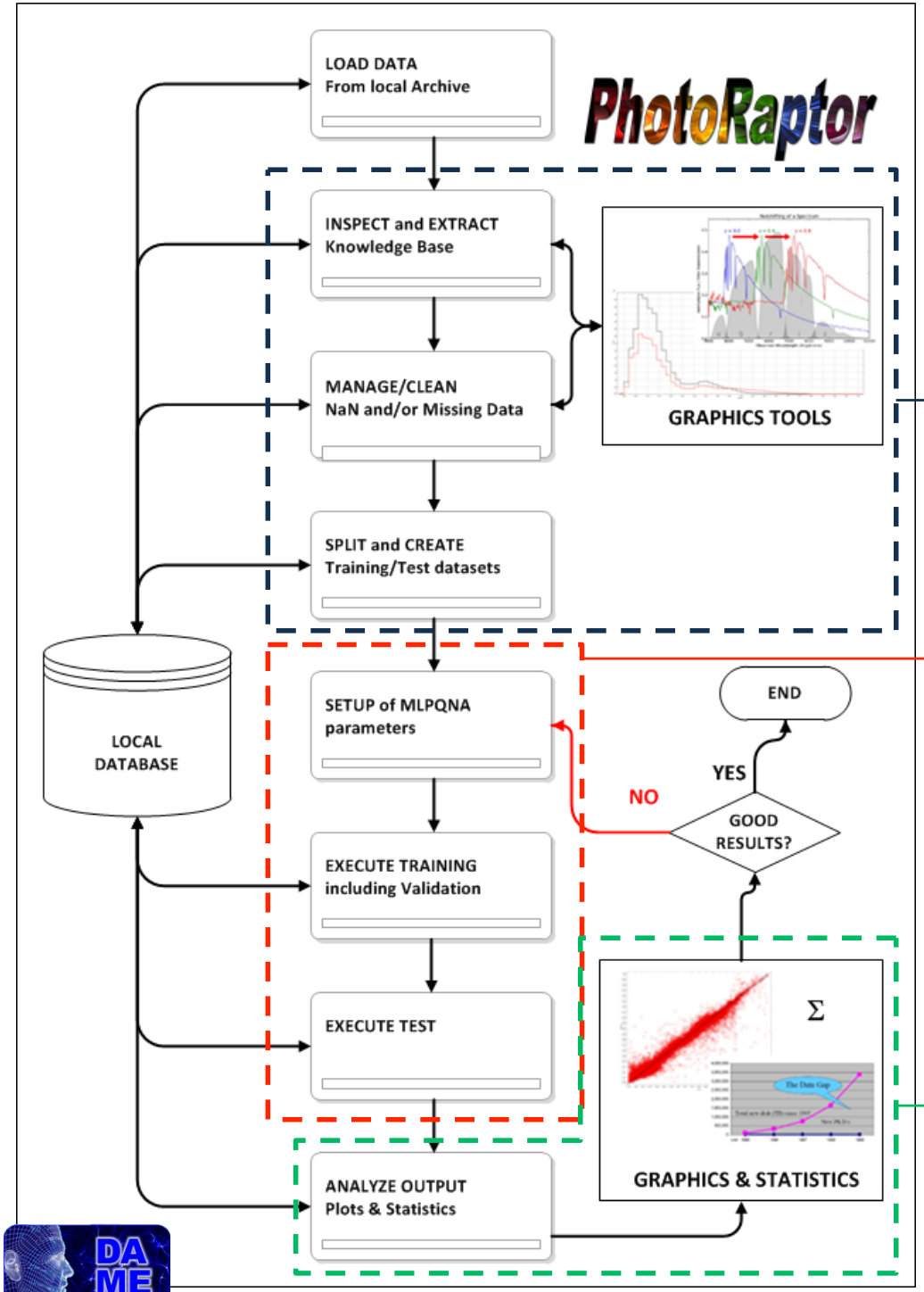
Disadvantages

- Need for better programming practice
- Need for accurate exhaustive documentation

PhotoRaptor

PHOTOmetRic APp TOol for Redshifts

Massimo Brescia, Stefano Cavuoti, Virgilio De Stefano, Giuseppe Longo
INAF – Capodimonte Astronomical Observatory
University Federico II of Naples



PRE-PROCESSING

- Metadata inspection
- Feature selection
- NaN handling
- Row shuffling
- Row splitting
- Table visualization
- Table plotting (histo, 2D/3D scatter plot)

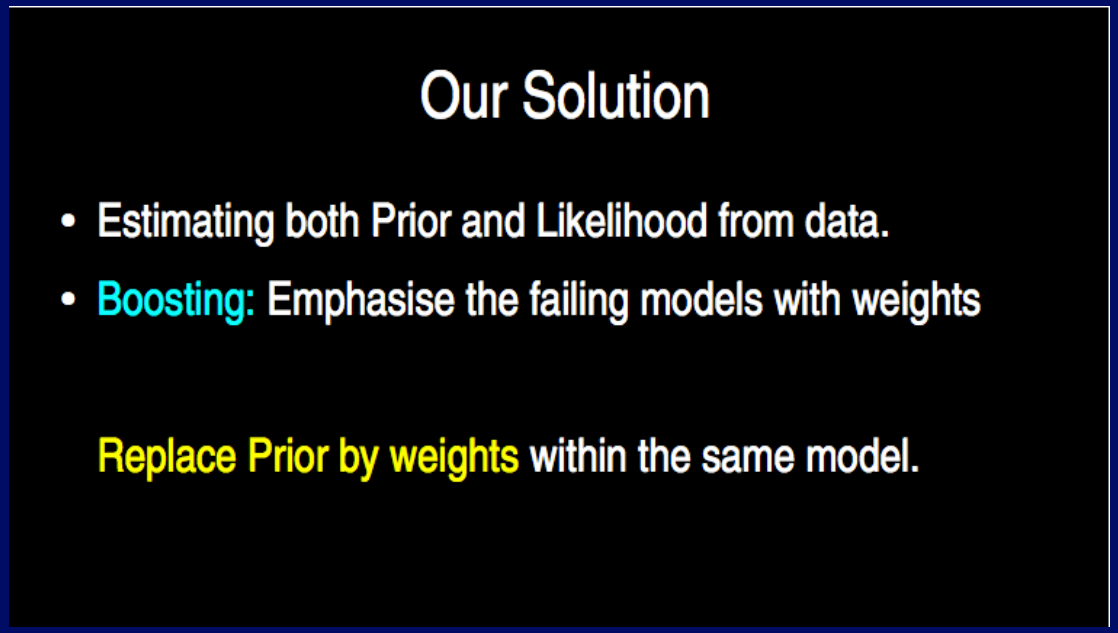
PROCESSING

- Regression/classification by MLPQNA
- Experiment setup
- Leave-one-out cross validation
- Training
- testing

POST-PROCESSING

- Regression statistics
  - Bias
  - Standard deviation
  - NMAD
  - Outlier percentage (at different σ)
- Classification statistics
- Outlier Handling
- Table plotting (histo, 2D/3D scatter plot)

Other Open Issues and conclusions. 1



From Ninan Sajeet Philip talk

1. How ML method can deal with Massive Data Sets with Missing data

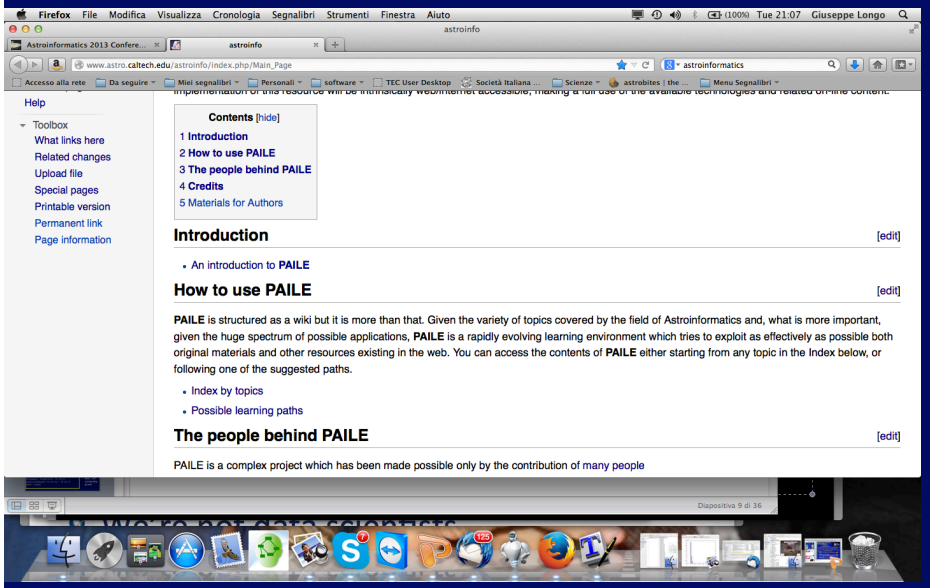
Other Open Issues and conclusions. 2

1. How to exctract reliable data sets to be used as knowledge base for learners…

2. Lack of template data sets for algorithm comparison (crucial for Euclid and other ongoing projects)

3. How to render this techniques user friendly for a non CS community and how to build expertise

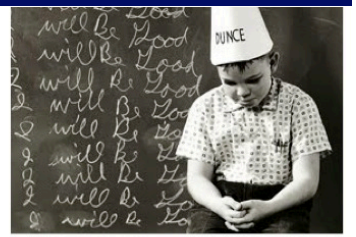
PAILE: Practical Astroinformatics "i" Learning Environment

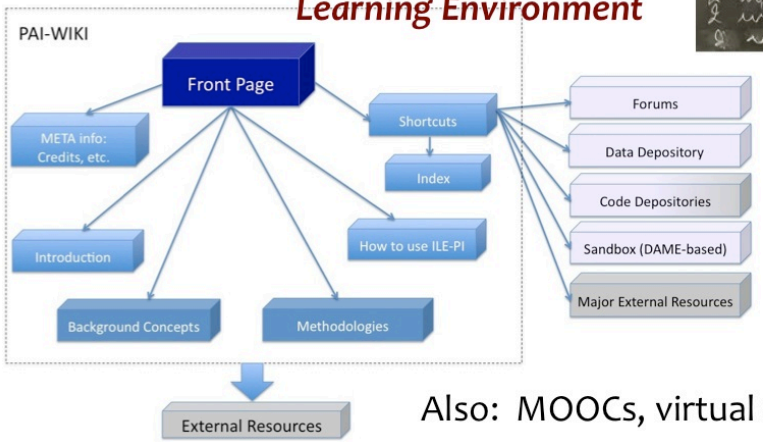


computational skills needed for the 21st century

Our answer:

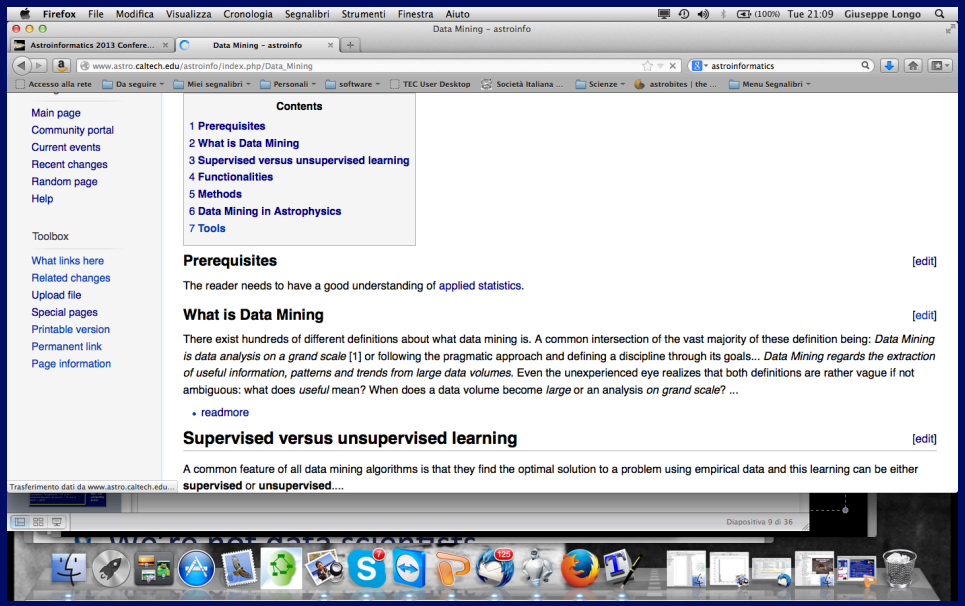
Practical Astroinformatics Interactive Learning Environment





Longo, Djorgovski, et al., in prep.

Also: MOOCs, virtual summer schools




NSA listening station in Bluffdale, South Dakota
1 YB (Yottabyte) of storage = 10^{12} TB
Indexed, searched, queried, mined…
With mainly unknown technology
which will slowly leak out to the scientific community


NSA listening station in Bluffdale, South Dakota
1 YB (Yottabyte) of storage = 10^{12} TB
Indexed, searched, queried, mined…
With mainly unknown technology
which will slowly leak out to the scientific community
And this is
…. The end
… in all meanings of the word