DAME (DAta Mining and Exploration)

An extensible, astronomer friendly data mining platform

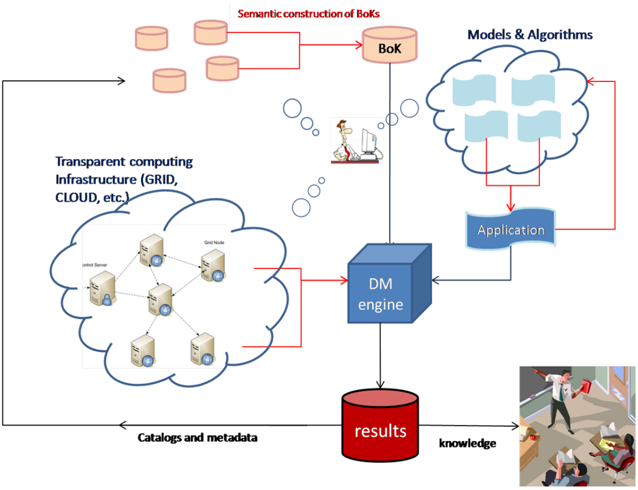
Massimo Brescia Omar Laurino Giuseppe Longo
and the DAME Team

INAF - OATS (Trieste) - OAC (Napoli)
University Federico II (Napoli)
Caifornia Institute of Technology

IVOA InterOp, Victoria - May 18th 2010

Data Mining and the Virtual Observatory: our reference framework

Key components:
BoK, App framework, Deployment environment.



Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts…

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts…

  ⟶ provide functionalities, not (just) models

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts...

  ⟶ provide functionalities, not (just) models
- Deployment environment (execution and storage) abstraction...

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts. . .
  ⟶ provide functionalities, not (just) models
- Deployment environment (execution and storage) abstraction. . .
  ⟶ Abstract Drivers to implement for GRID, SA, . . .

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts. . .

  ⟶ provide functionalities, not (just) models

- Deployment environment (execution and storage) abstraction. . .

  ⟶ Abstract Drivers to implement for GRID, SA, . . .

  ⟶ Serializable plugins

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts…
  ⟶ provide functionalities, not (just) models
- Deployment environment (execution and storage) abstraction…
  ⟶ Abstract Drivers to implement for GRID, SA, …
  ⟶ Serializable plugins
- Allow computer scientists and power users to contribute…

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts. . .
  ⟶ provide functionalities, not (just) models
- Deployment environment (execution and storage) abstraction. . .
  ⟶ Abstract Drivers to implement for GRID, SA, . . .
  ⟶ Serializable plugins
- Allow computer scientists and power users to contribute. . .
  ⟶ Plugin development kit

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts…

  ⟶ provide functionalities, not (just) models

- Deployment environment (execution and storage) abstraction…

  ⟶ Abstract Drivers to implement for GRID, SA, …

  ⟶ Serializable plugins

- Allow computer scientists and power users to contribute…

  ⟶ Plugin development kit

- Client customization…

Requirements

- Let KDD unaware researchers access KDD without becoming KDD experts. . .

  ⟶ provide functionalities, not (just) models
- Deployment environment (execution and storage) abstraction. . .

  ⟶ Abstract Drivers to implement for GRID, SA, . . .

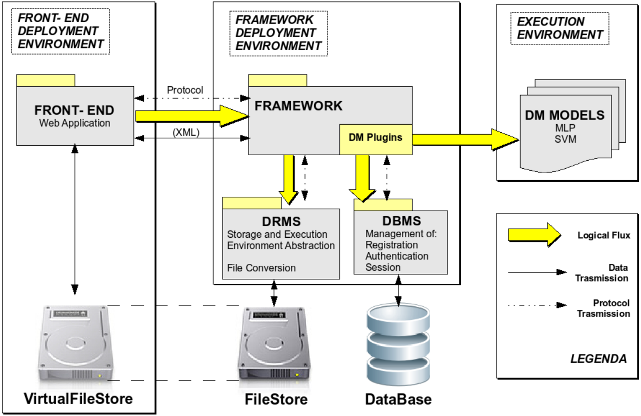
  ⟶ Serializable plugins
- Allow computer scientists and power users to contribute. . .

  ⟶ Plugin development kit
- Client customization. . .

  ⟶ (RESTful) webservices.

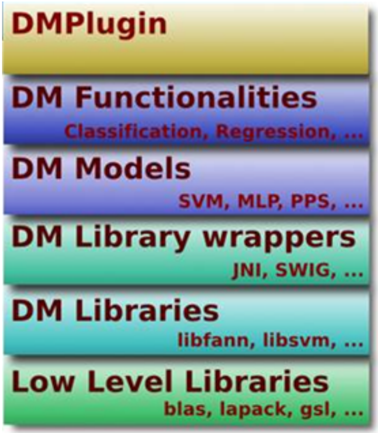
Our solution: an extensible architecture

Each component is replaceable except the framework, which can be extended by means of plugins.



Data Mining Plugins stack

- A Data Mining plugin implements a scientific use case (e.g. classification, regression, multi-layer clustering, . . .).
- Each DMPlugin may use one or more different low level Data Mining Models.
- DAME exposes DMPlugins to the user, and DMModels to devs through a Software Development Kit.
- Models are reusable, plugins are not (unless we decide to implement workflows).



DMPlugin

DM Functionalities

Classification, Regression, ...

DM Models

SVM, MLP, PPS, ...

DM Library wrappers

JNI, SWIG, ...

DM Libraries

libfann, libsvm, ...

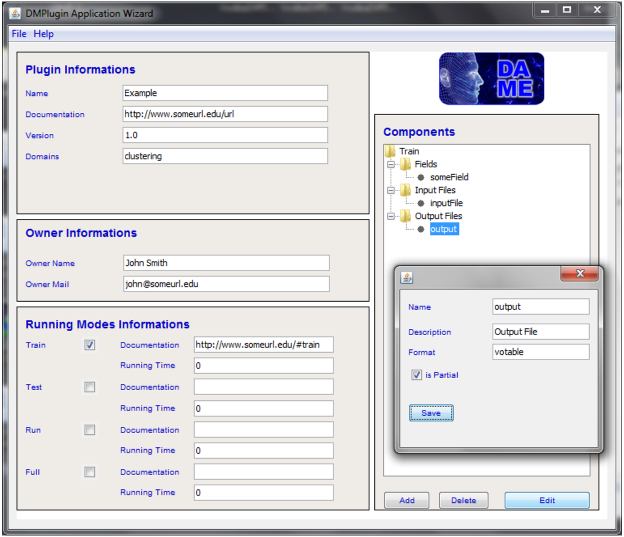
Low Level Libraries

blas, lapack, gsl, ...

DMPlugin basic features

- Dynamically loaded at runtime, but have to be registered by an admin;
- they implement abstract methods of an abstract Java class;
- they are unaware of the environment in which they are going to be launched;
- they are also unaware of where the files actually are;
- they know whether they are long or short running tasks;
- interaction with the environment through abstract drivers.

Contributing DMPlugins



Plugin Informations

Name: Example
Documentation: http://www.someurl.edu/url
Version: 1.0
Domains: clustering

Owner Informations

Owner Name: John Smith
Owner Mail: john@someurl.edu

Running Modes Informations

Train	☑	Documentation	http://www.someurl.edu/#train
		Running Time	0
Test	□	Documentation	
		Running Time	0
Run	□	Documentation	
		Running Time	0
Full	□	Documentation	
		Running Time	0

Components

- Train
  - Fields
    - someField
  - Input Files
    - inputFile
  - Output Files
    - output

Name: output
Description: Output File
Format: votable
☑ is Partial

The DEAL: deployment environment abstraction layer

- Both execution and storage are abstract Java classes which must be implemented for the specific storage and execution platforms.
- We abstracted the idea of job execution running time so that DMPlugin developers can tell the driver, at runtime, whether the code is a long running task or not. That way different drivers can be instantiated in order to launch the task.
- We have implemented the Stand Alone driver and we are currently developing an EGEE GLite compliant driver for GRID.
- To contribute with a driver you just have to implement the abstract methods of an abstract class.

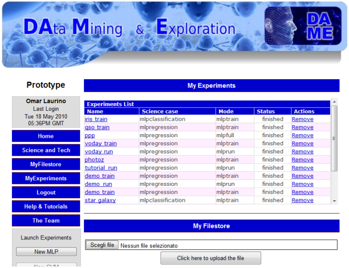
The DEAL: an example

We developed a proof of concept GLite driver to demonstrate how plugins can be launched in different execution environments:

- according to user inputs the plugin is instantiated on the framework machine (where all the business logic is running);
- according to the running time the correct Driver is instantiated;
- if it is a long run task, the GRID driver is instantiated and its `run()` method invoked;
- the GRID implementation just serializes the plugin and calls a GRID executable through GLite middleware;
- the executable takes the serialized plugin as input, deserializes it and lauch it on the Computing Element;
- during the execution the DMPlugin is also able to send messages back to the Framework.

The first prototype

- Currently running at `http://dame.na.infn.it`. It was meant to be just a prototype for validating user interface requirements.
- It was successfully used during Italian VODays (>200 users).
- It is soon retiring, as soon as the new rich web application front end is ready (matter of days).
- It just exposes an MLP model. But it can be used to show KDD potential on actual science cases (refereed papers).



Omar Laurino
Last Login
Tue 18 May 2010
05:36PM GMT

My Experiments

Experiments List

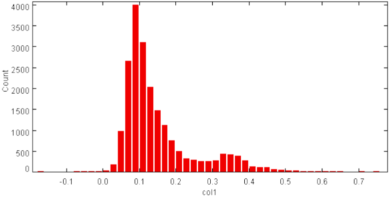
Name	Science case	Mode	Status	Actions
iris_train	mlpclassification	mlptrain	finished	Remove
qso_train	mlpregression	mlptrain	finished	Remove
ppp	mlpregression	mlpfull	finished	Remove
voday_train	mlpregression	mlptrain	finished	Remove
voday_run	mlpregression	mlprun	finished	Remove
photoz	mlpregression	mlptrain	finished	Remove
tutorial_run	mlpregression	mlprun	finished	Remove
demo_train	mlpregression	mlptrain	finished	Remove
demo_run	mlpregression	mlprun	finished	Remove
demo_train	mlpregression	mlptrain	finished	Remove
star_galaxy	mlpclassification	mlptrain	finished	Remove

My Filestore

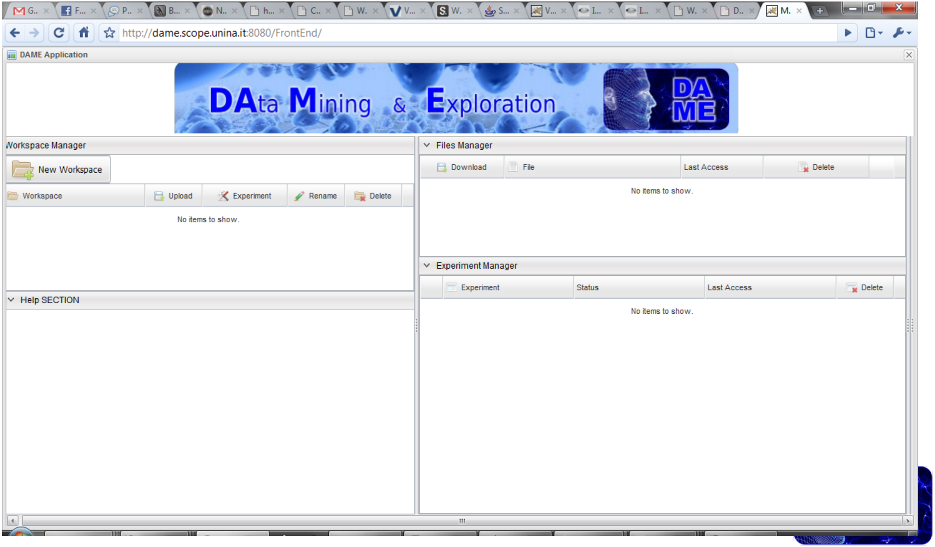
Nessun file selezionato

DAME @ VODay

- Scientific Use Case: photometric redshifts of SDSS galaxies:
  1. Train a Neural Network and test it.
  2. Download a catalogue data about Abell2255 from the VO.
  3. Use your own neural network to verify that Abell2255 is actually a cluster of galaxies and find its redshift.
- Data Mining was unknown to almost all the participants.
- Who found DAME interesting found it easy to use, too!
- Several researchers wondered it they could use KDD for their science.



A glimpse of the brandnew shining Front End.



http://dame.scope.unina.it:8080/FrontEnd/

DAME Application

DAta Mining & Exploration

DA ME

Workspace Manager

New Workspace

Workspace	Upload	Experiment	Rename	Delete

No items to show.

Help SECTION

Files Manager

Download	File	Last Access	Delete

No items to show.

Experiment Manager

Experiment	Status	Last Access	Delete

No items to show.

Prioritized functionality

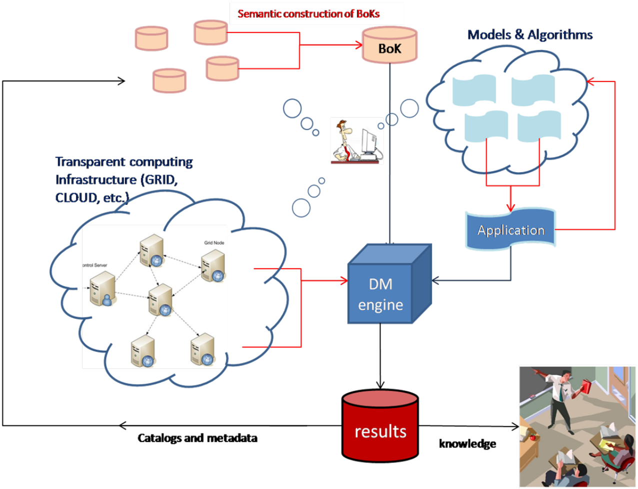
- Supervised regression and classification.
- Unsupervised clustering.
- Multi-Layer clustering.
- Image segmentation.

Guest stars

- VOGClusters: a web application for globular clusters;
- Wide Field X-Ray Telescopes Transient Calculator;

Data Mining and the Virtual Observatory: our reference framework

Key components:
BoK, App framework, Deployment environment.



Semantic construction of BoKs

BoK

Models & Algorithms

Transparent computing Infrastructure (GRID, CLOUD, etc.)

Application

DM engine

results

Catalogs and metadata

knowledge

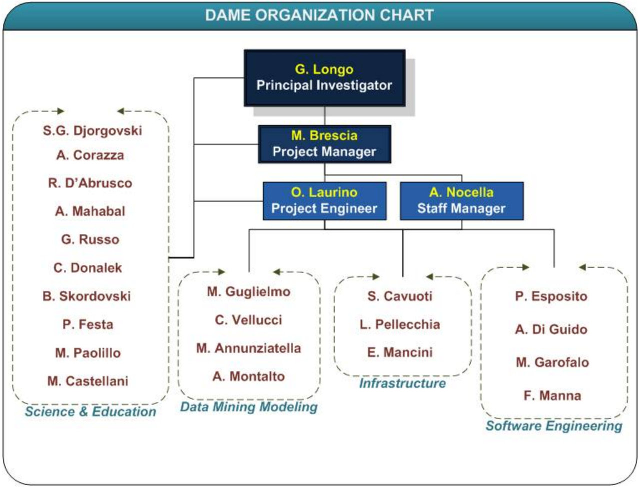
Talking of BoK...

DAME tries to address algorithms and deployment infrastructure: what about the BoK extraction?

The Virtual Observatory is the best environment for BoK selection, and we are now starting to have VO specific tools that will allow to do that. For example:

- Semantics WG (I'm looking forward to hear Matthew's talk!)
- CDS annotation service;
- VOdka (?).

Thank you



DAME ORGANIZATION CHART

G. Longo
Principal Investigator

M. Brescia
Project Manager

O. Laurino
Project Engineer

A. Nocella
Staff Manager

S.G. Djorgovski
A. Corazza
R. D'Abrusco
A. Mahabal
G. Russo
C. Donalek
B. Skordovski
P. Festa
M. Paolillo
M. Castellani
Science & Education

M. Guglielmo
C. Vellucci
M. Annunziatella
A. Montalto
Data Mining Modeling

S. Cavuoti
L. Pellecchia
E. Mancini
Infrastructure

P. Esposito
A. Di Guido
M. Garofalo
F. Manna
Software Engineering