DAta Mining & Exploration

DA
ME

we make science discovery happen

searching for KDD in MDS standards…
…the DAME experience

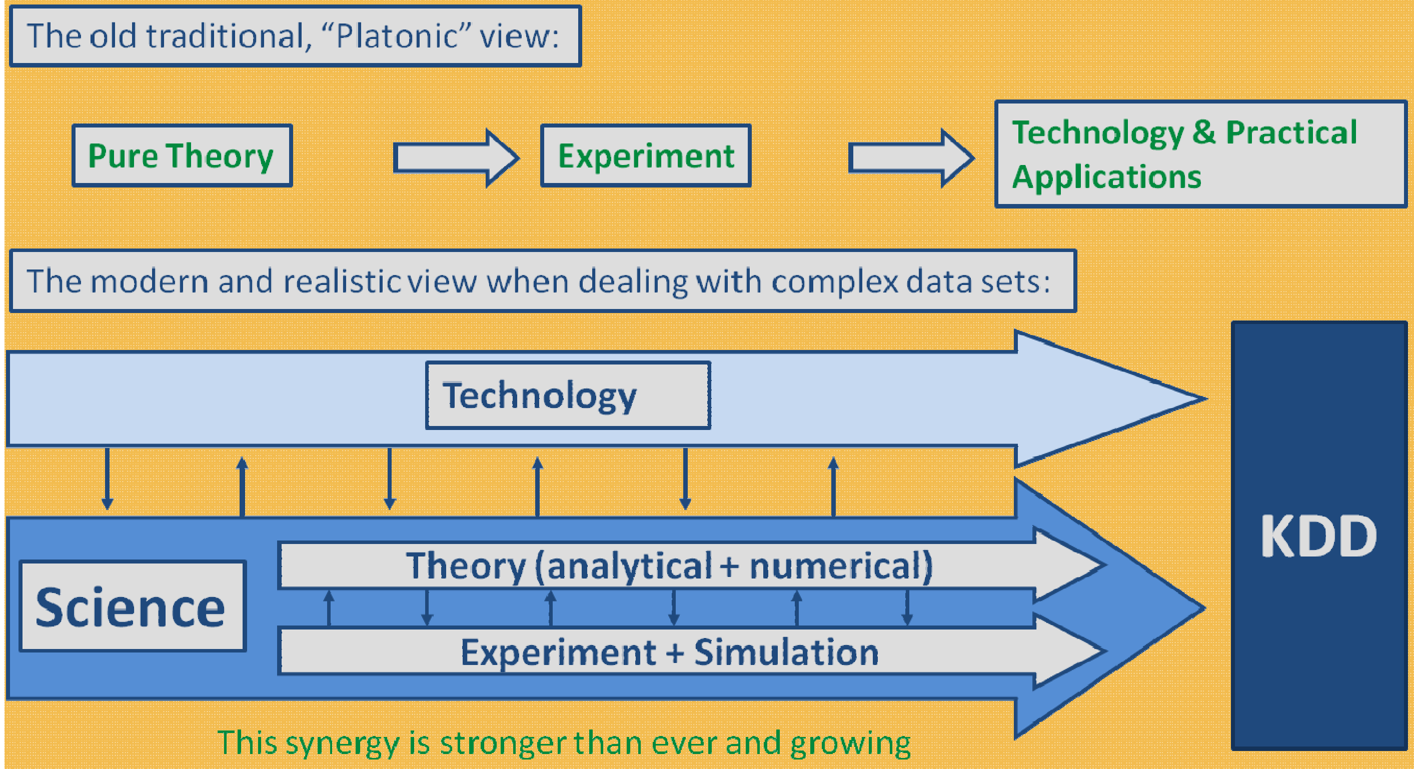
Marianna Annunziatella, **Massimo Brescia**, Stefano Cavuoti, Raffaele D'Abrusco, George S. Djorgovski, Ciro Donalek, Mauro Garofalo , Marisa Guglielmo, Omar Laurino, Giuseppe Longo, Ashish Mahabal, Ettore Mancini, Francesco Manna, Amata Mercurio, Alfonso Nocella, Maurizio Paolillo, Luca Pellecchia, Sandro Riccardi, Giovanni Vebber, Civita Vellucci.

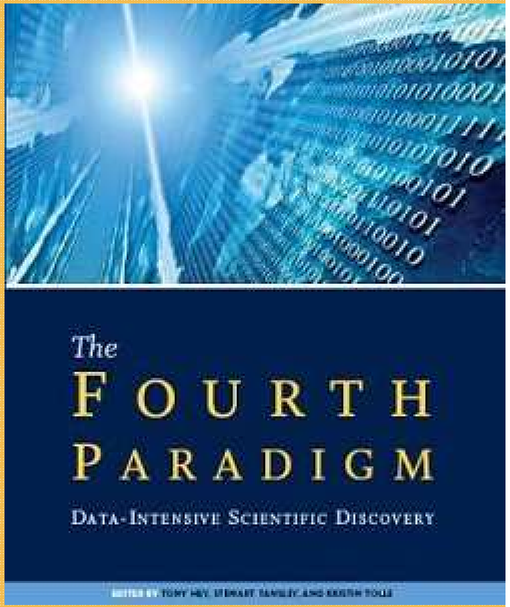
Department of Physics – University Federico II – Napoli

INAF – National Institute of Astrophysics – Capodimonte Astronomical Observatory – Napoli

CALTECH – California Institute of Technology - Pasadena

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:
Technology
Science
Theory (analytical + numerical)
Experiment + Simulation
KDD
This synergy is stronger than ever and growing



The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY

Definition

DM is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules

The BoK's Problem

Limited number of problems due to limited number of reliable BoKs

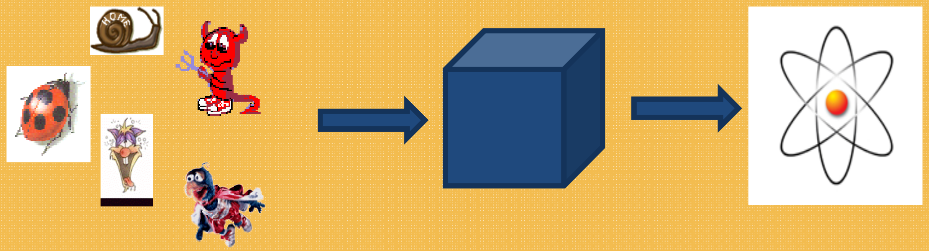
So far

- 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)

• There's a need of standardization and interoperability between data together with DM application

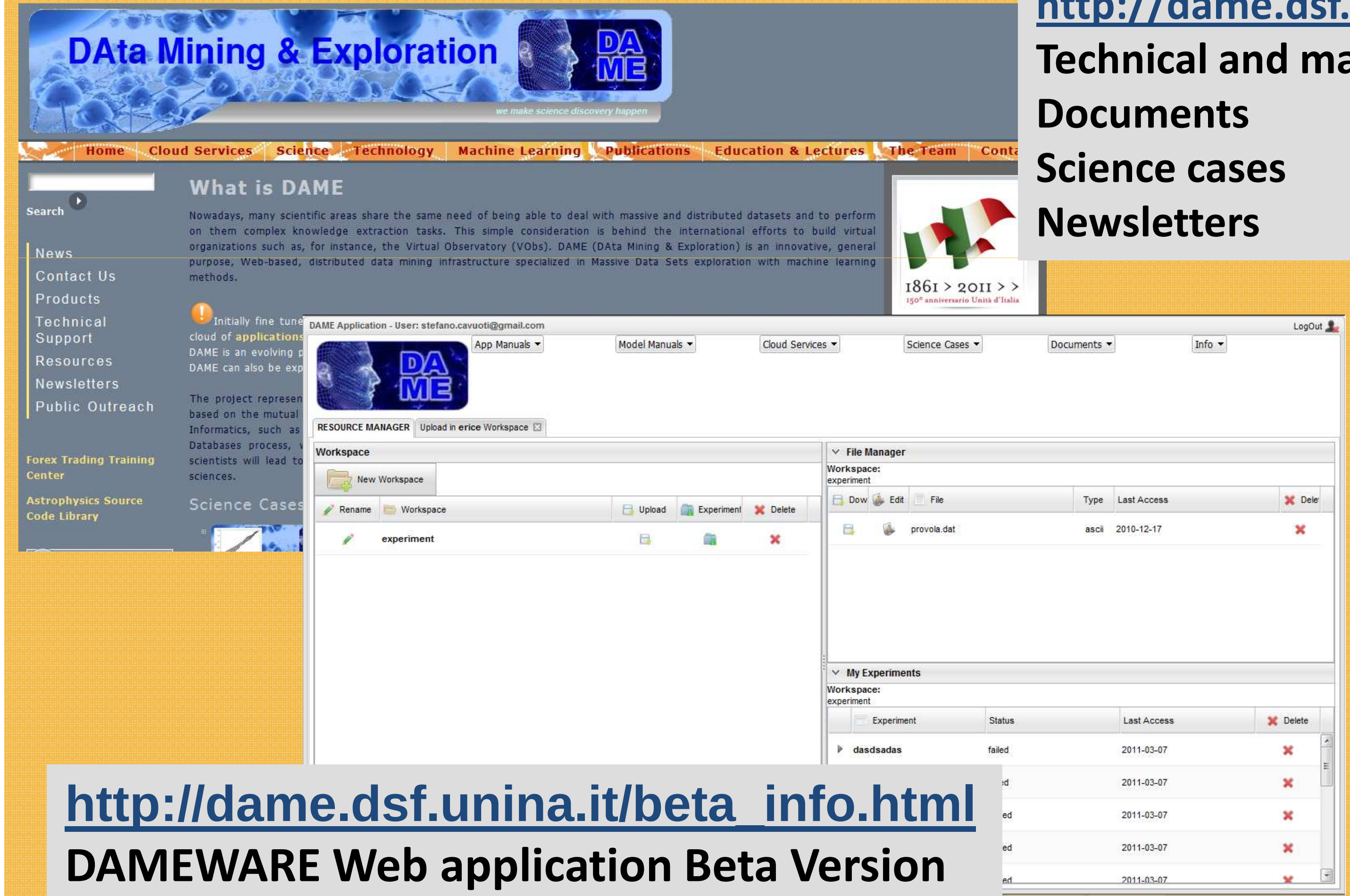
Community believes AI/DM methods are black boxes

You feed in something, and obtain patters, trends, i.e. knowledge….



What DAME is

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



What is DAME

Nowadays, many scientific areas share the same need of being able to deal with massive and distributed datasets and to perform on them complex knowledge extraction tasks. This simple consideration is behind the international efforts to build virtual organizations such as, for instance, the Virtual Observatory (VObs). DAME (DAta Mining & Exploration) is an innovative, general purpose, Web-based, distributed data mining infrastructure specialized in Massive Data Sets exploration with machine learning methods.

Initially fine tune

cloud of applications

DAME is an evolving p

DAME can also be exp

The project represen

based on the mutual

Informatics, such as

Databases process, v

scientists will lead to

sciences.

Science Cases

DAME Application - User: stefano.cavuoti@gmail.com

Workspace

Rename	Workspace	Upload	Experiment	Delete
	experiment			✖

File Manager

Workspace:
experiment

Dow	Edit	File	Type	Last Access	Dele
		provola.dat	ascii	2010-12-17	✖

My Experiments

Workspace:
experiment

Experiment	Status	Last Access	Delete
dasdsadas	failed	2011-03-07	✖
	ed	2011-03-07	✖
	ed	2011-03-07	✖
	ed	2011-03-07	✖
	ed	2011-03-07	✖

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

Technical and management info

Documents

Science cases

Newsletters

01
DAta Mining & Exploration
In this edition
Web Solutions for Data Mining in Astrophysics

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

DAMEWARE Web application Beta Version

DM 4-rule virtuous cycle

- Finding patterns is not enough
- Science business must:

– Respond to patterns by taking action

– Turning:

  - Data into Information
  - Information into Action
  - Action into Value

- **Hence, the Virtuous Cycle of DM:**

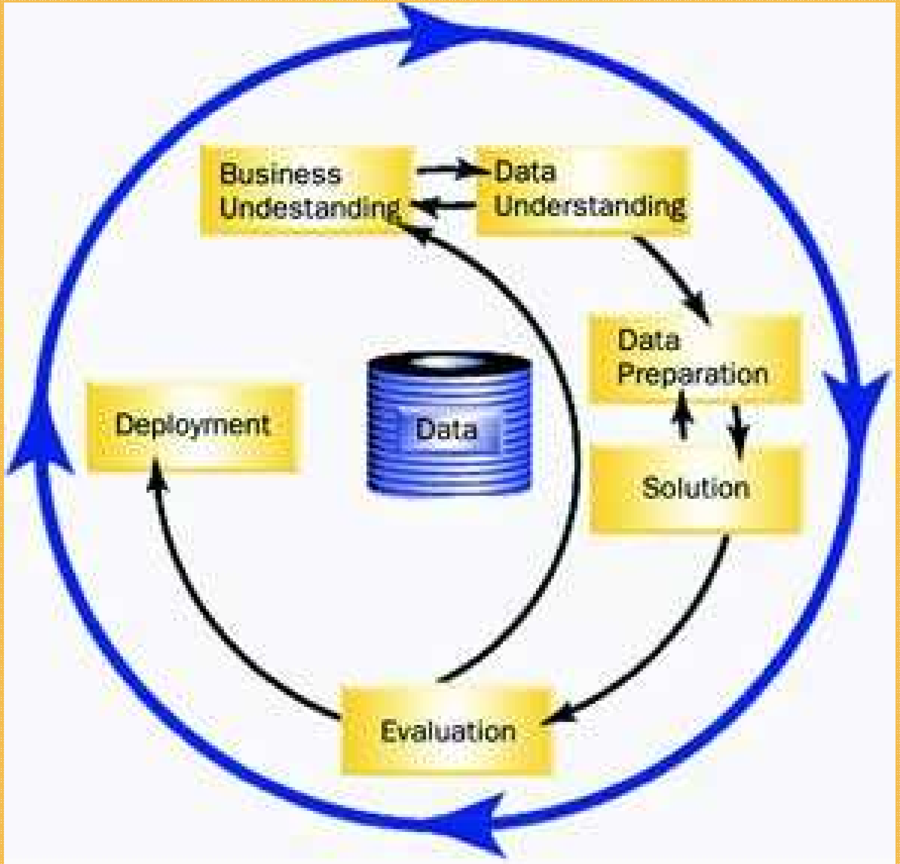
- **Virtuous cycle implementation steps:**
  - **Transforming data into information via:**
    - **Hypothesis testing**
    - **Profiling**
    - **Predictive modeling**
  - **Taking action**
    - **Model deployment**
    - **Scoring**
  - **Measurement**
    - **Assessing a model's stability & effectiveness before it is used**

1. **Identify the problem**
2. **Mining data to transform it into actionable information**
3. **Acting on the information**
4. **Measuring the results**

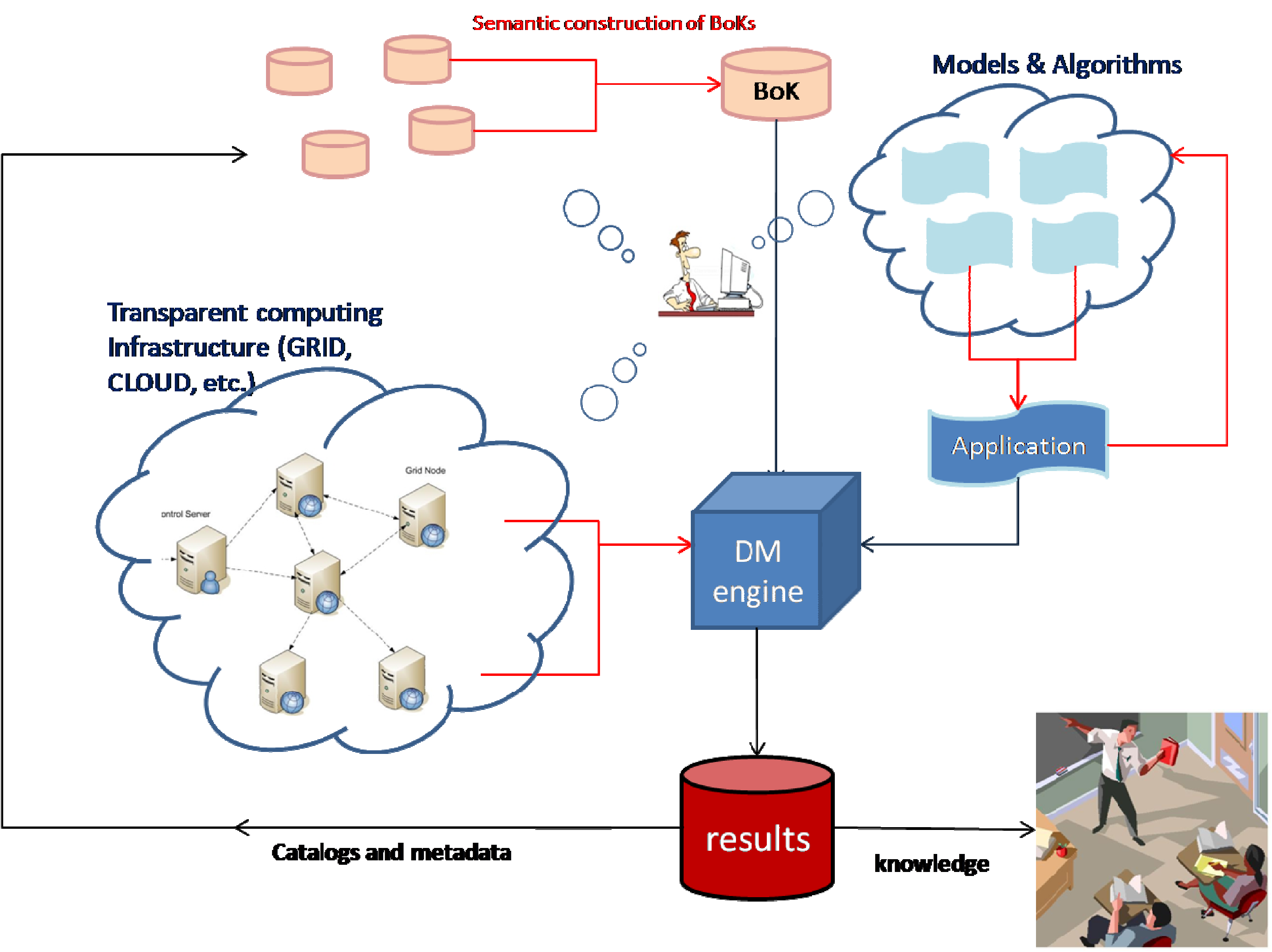
DM: 11-step Methodology

The four rules reflect into an 11-step exploded strategy, at the base of DAME concept

1. **Translate any opportunity (science case) into DM opportunity (problem)**
2. **Select appropriate data**
3. **Get to know the data**
4. **Create a model set**
5. **Fix problems with the data**
6. **Transform data to bring information**
7. **Build models**
8. **Assess models**
9. **Deploy models**
10. **Assess results**
11. **Begin again (GOTO 1)**



Effective DM process break-down



The Black box Infrastructure

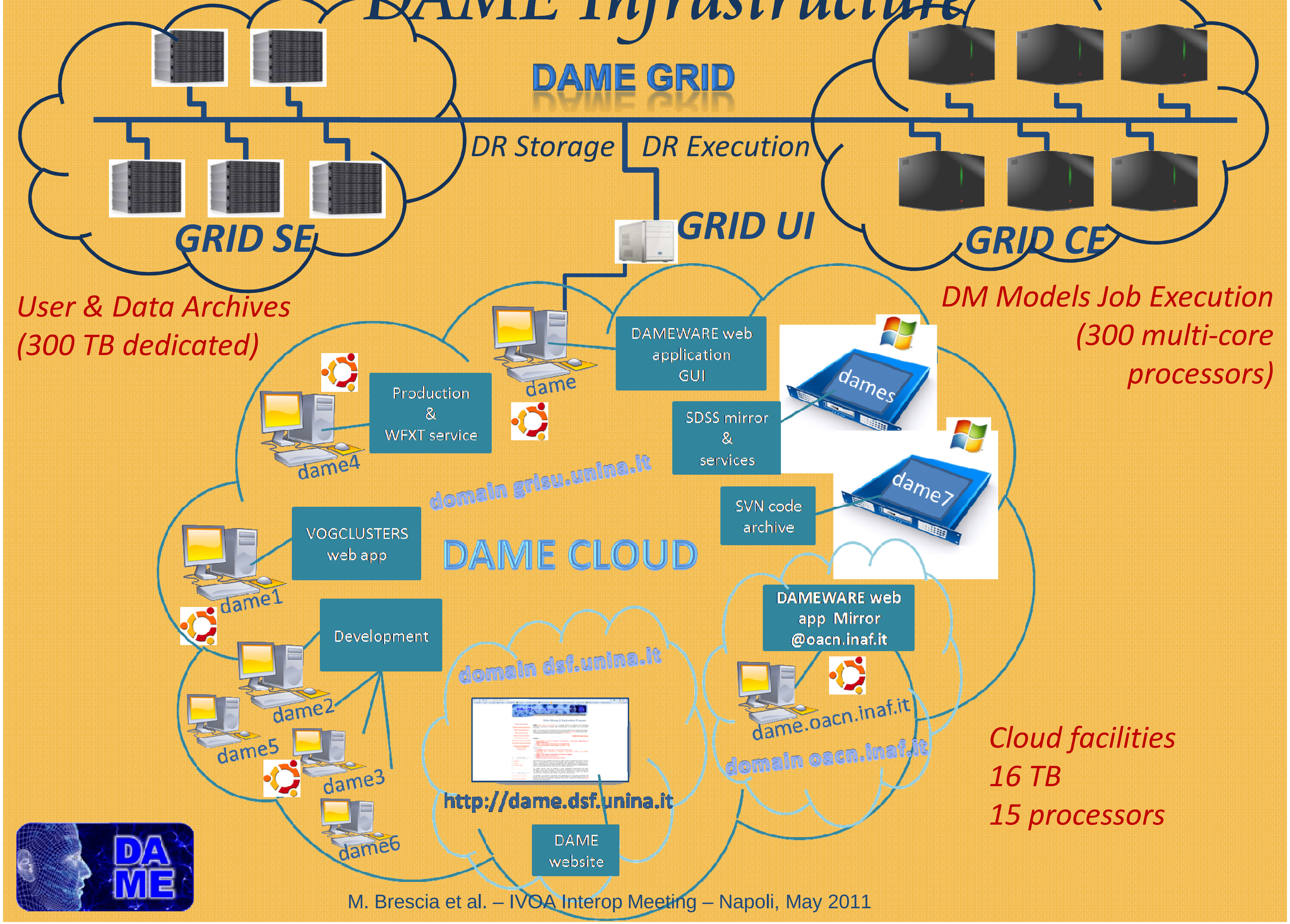
In this scenario DAME (Data Mining & Exploration) project, starting from astrophysics requirements domain, has investigated the Massive Data Sets (MDS) exploration by producing a taxonomy of data mining applications (hereinafter called **functionalities**) and collected a set of machine learning algorithms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

This association functionality-model is made of what we defined "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 the computing infrastructure and almost nothing about the internal mechanisms of the chosen machine learning model..

input
?
black-box
output

DAME Infrastructure



DAME GRID

DR Storage

DR Execution

GRID SE

GRID UI

GRID CE

User & Data Archives
(300 TB dedicated)

DM Models Job Execution
(300 multi-core
processors)

DAME CLOUD

domain grisu.unina.it

dame

DAMEWARE web application GUI

Production & WFXT service

dame4

SDSS mirror & services

dames

SVN code archive

dame7

VOGCLUSTERS web app

dame1

Development

dame2

dame5

dame3

dame6

domain dsf.unina.it

http://dame.dsf.unina.it

DAME website

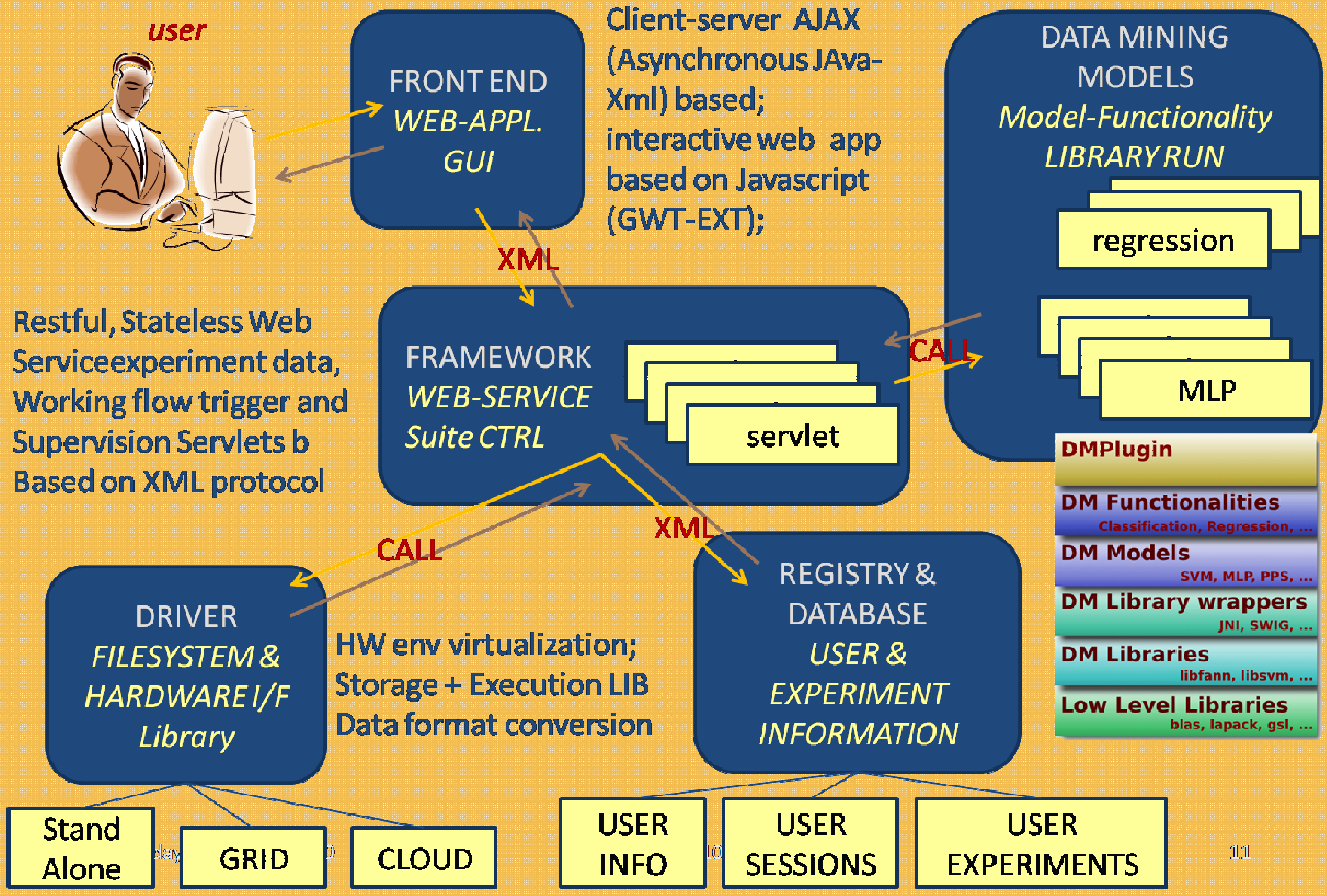
DAMEWARE web app Mirror @oacn.inaf.it

dame.oacn.inaf.it

domain oacn.inaf.it

Cloud facilities
16 TB
15 processors

DAME SW Architecture



user

FRONT END
WEB-APPL.
GUI

Client-server AJAX (Asynchronous JAva-Xml) based; interactive web app based on Javascript (GWT-EXT);

DATA MINING MODELS
Model-Functionality LIBRARY RUN

regression

MLP

XML

CALL

Restful, Stateless Web Serviceexperiment data, Working flow trigger and Supervision Servlets b Based on XML protocol

FRAMEWORK
WEB-SERVICE
Suite CTRL

servlet

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, ...

CALL

XML

DRIVER
FILESYSTEM & HARDWARE I/F Library

HW env virtualization; Storage + Execution LIB Data format conversion

REGISTRY & DATABASE
USER & EXPERIMENT INFORMATION

Stand Alone

GRID

CLOUD

USER INFO

USER SESSIONS

USER EXPERIMENTS

The Available Services

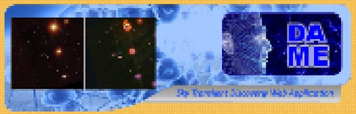
DAMEWARE Web Application Resource

Main service providing via browser a list of algorithms and tools to configure and launch experiments as complete workflows (dataset creation, model setup and run, graphical/text output):

- *Functionalities: Regression, Classification, Image Segmentation, Multi-layer Clustering;*
- *Models: MLP+BP, MLP+GA, SVM, MLP+QNA, **K-Means (through KNIME)**, PPS, SOM, NEXT-II;*

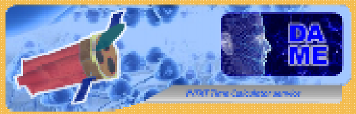
VOGCLUSTERS

Web Application for data and text mining on globular clusters;



STraDiWA (Sky Transient Discovery Web Application)

detect variable objects from real or simulated images (under R&D);



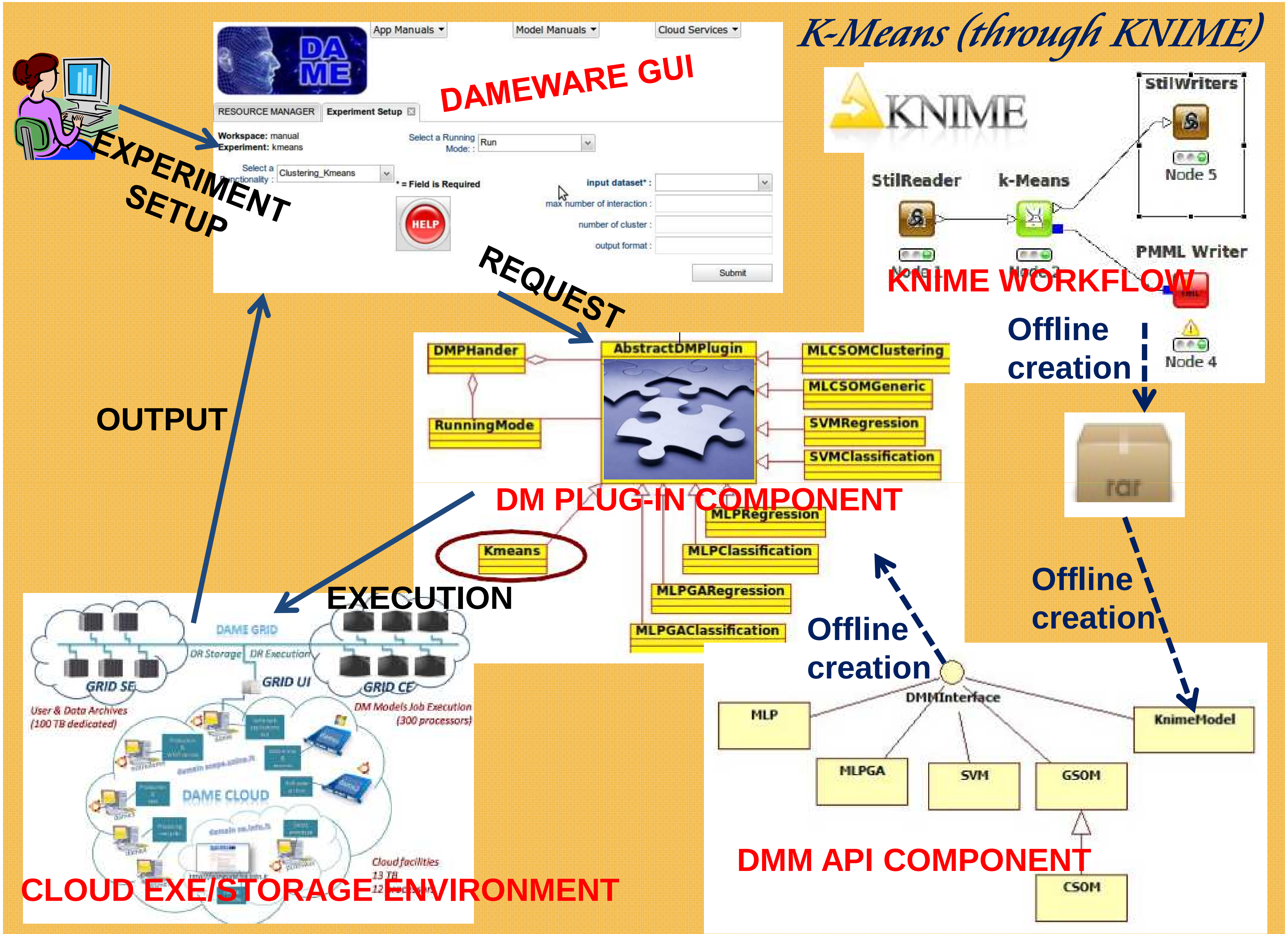
WFXT (Wide Field X-Ray Telescope) Transient Calculator

Web service to estimate the number of transient and variable sources that can be detected by WFXT within the 3 main planned extragalactic surveys, with a given significant threshold;

SDSS (Sloan Digital Sky Survey)

Local mirror website hosting a complete SDSS Data Archive and Exploration System;

K-Means (through KNIME)



DAMEWARE GUI

App Manuals ▾ | Model Manuals ▾ | Cloud Services ▾

RESOURCE MANAGER | Experiment Setup ☒

Workspace: manual
Experiment: kmeans

Select a Running Mode: : Run

Select a Functionality : Clustering_Kmeans

* = Field is Required

HELP

input dataset* :
max number of interaction :
number of cluster :
output format :

Submit

Web 2.0 Features in DAME

***Web 2.0? It is a system that breaks with the old model of centralized Web sites and moves the power of the Web/Internet to the desktop*. [J. Robb]**

***the Web becomes a universal, standards-based integration platform.* [S. Dietzen]**

software and storage facilities, all through a simple browser

Service Oriented App (SOA) Growing functionalities integration via app service interoperability

client-side browser with asynchronous Javascript/Ajax, JDOM and XML standard technologies

unification in a single framework of:

❖ Client-server structure

❖ WYSWYG Dynamical content

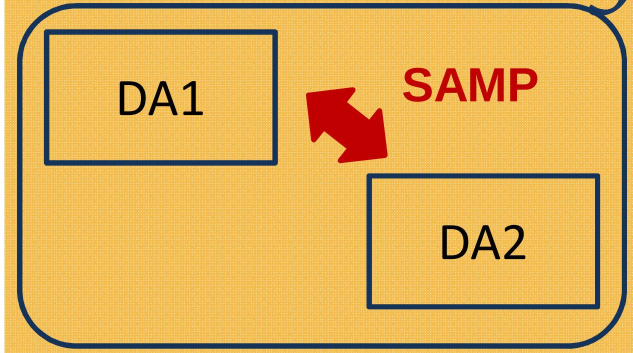
❖ Network protocols

Rich Internet App (RIA) Network as a process platform Desktop app → Web app

Web as a participating and sharing information platform

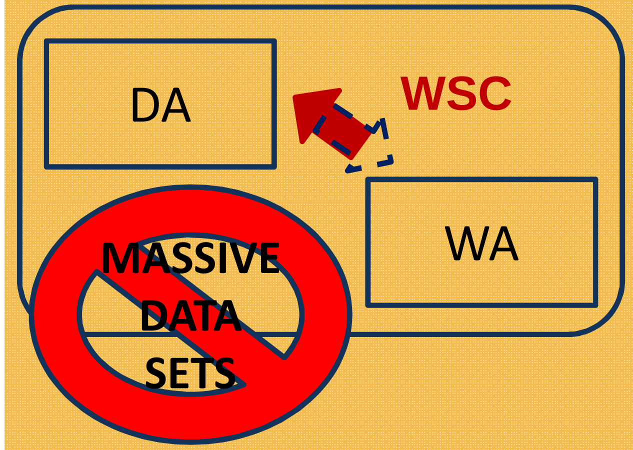
Machine-based interactions (REST, SOAP) based on standards (PMML, WSDL, XML)

VO Interoperability scenarios



Full interoperability between DA (Desktop Applications)

Local user desktop fully involved (requires computing power)

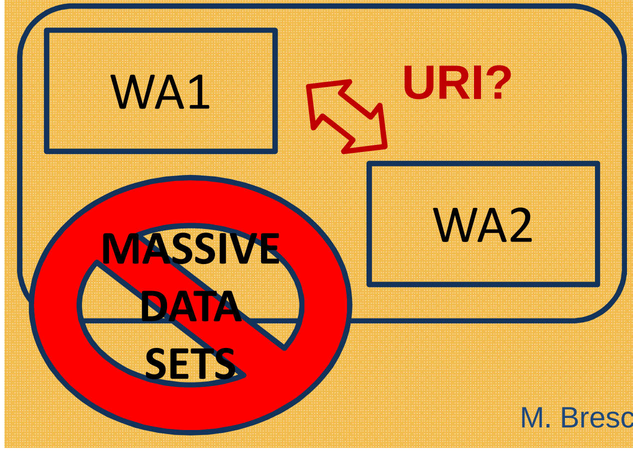


Full WA → DA interoperability

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

MDS must be moved between local and remote apps

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



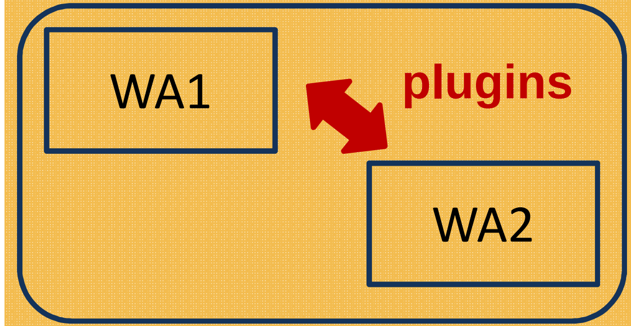
Except from URI exchange, no standard interoperabilty

Different accounting policy

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

No local computing power required

Our vision: improving aspects



DAs has to become WAs

Unique accounting policy (google/Microsoft like)

To overcome MDS flow apps must be plug&play (e.g. any WAx feature should be pluggable in WAy on demand)

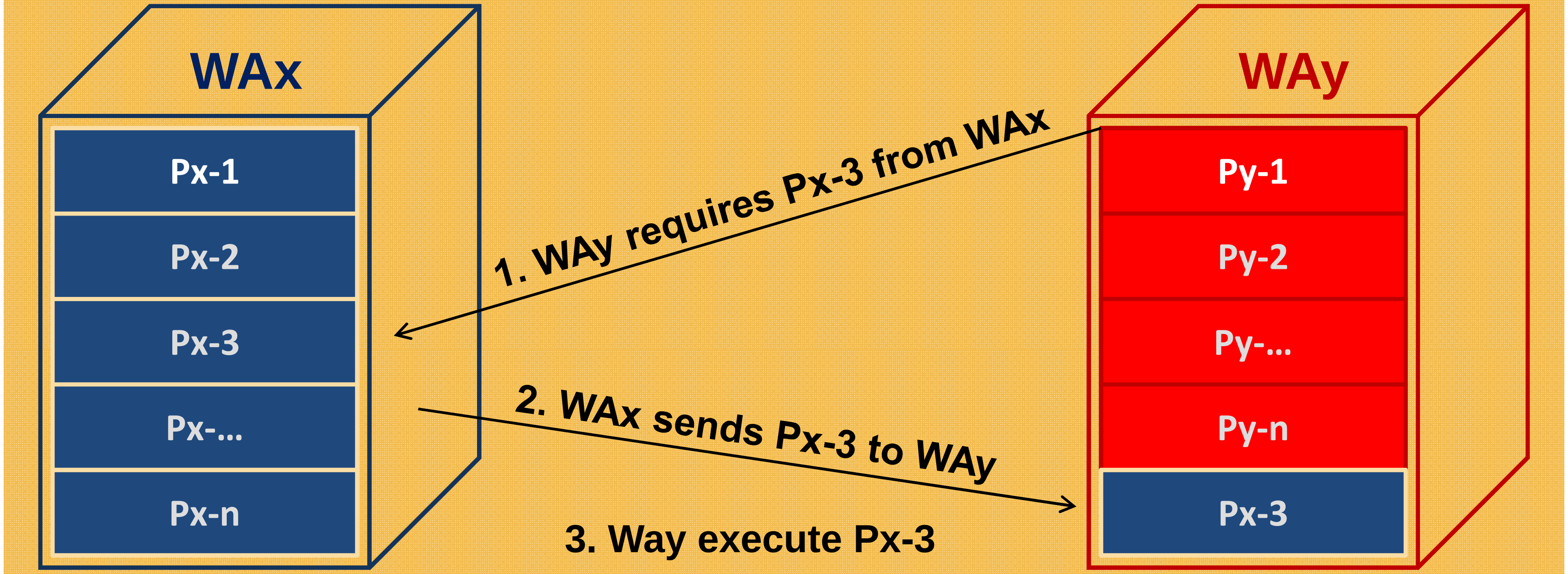
No local computing power required. Also smartphones can run VO apps

⬇

Requirements

- Standard accounting system;
- No more MDS moving on the web, but just moving Apps, structured as plugin repositories and execution environments;
- standard modeling of WA and components to obtain the maximum level of granularity;
- Evolution of SAMP architecture to extend web interoperability (in particular for the migration of the plugins);

Our vision: plugin granularity flow

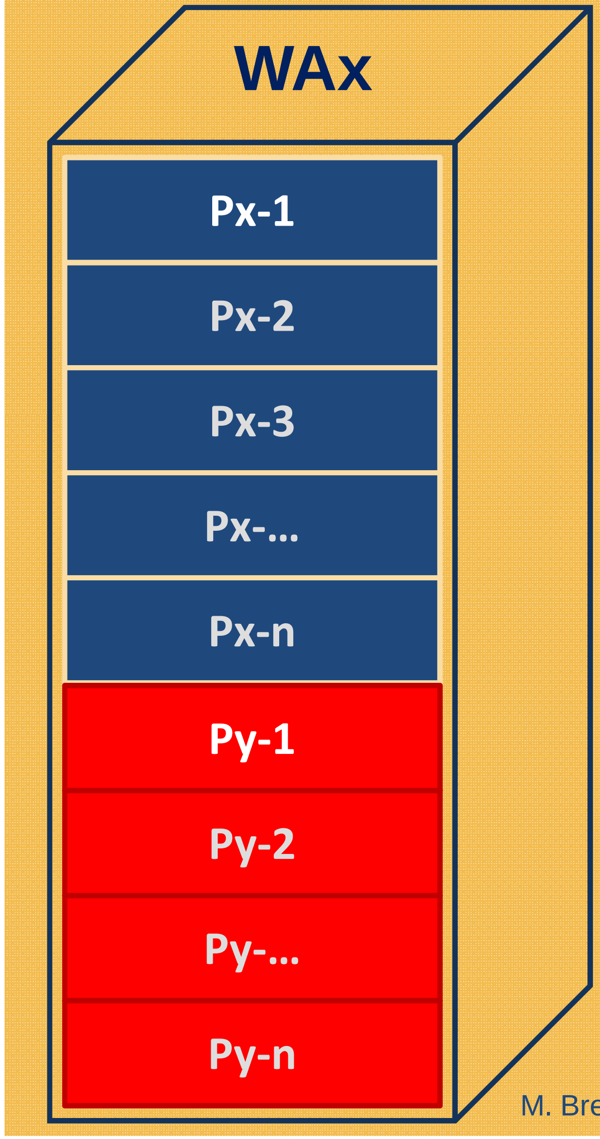


This scheme could be iterated and extended involving all standardized web apps

The Lernaean Hydra VO KDD App

The Lernaean Hydra VO KDD App

After a certain number of such iterations…



WAx

Px-1

Px-2

Px-3

Px-…

Px-n

Py-1

Py-2

Py-…

Py-n

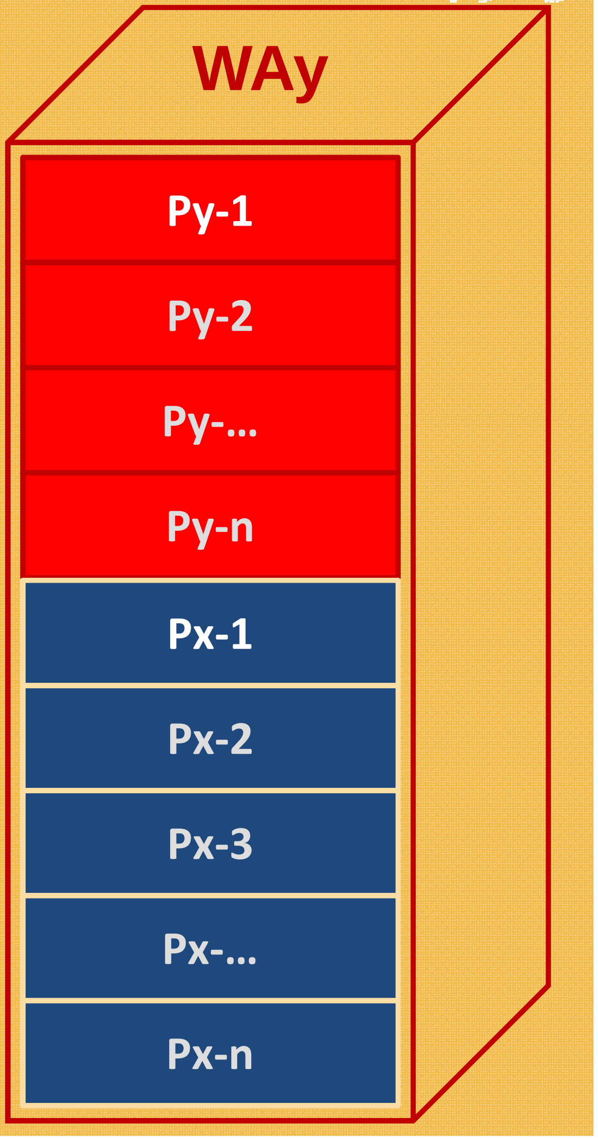
The VO KDD App scenario will become:

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

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

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

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



WAy

Py-1

Py-2

Py-…

Py-n

Px-1

Px-2

Px-3

Px-…

Px-n

Conclusions

DAME was not originally conceived (for the lack of suitable standards) to be interoperable with the VO, but offers a good benchmark to plan for the future developments of KDD on MDS in a VO environment.

1. DAME is just an example of what new ICT (Web 2.0) can do for A&A KDD problems.
2. A new vision of the KDD App approach, suitable for VO must be based on the minimization of data transfer and maximization of interoperability within the VO community.
3. If implemented, the new scheme can reach a wider science community by giving the opportunity to share data and apps worldwide, without any particular infrastructure requirements (i.e. by using a simple smartphone with a low-band connection).

DAME group is currently involved in the definition of standards and rules and is working to modify and adapt the present infrastructure to become compliant with the VO.