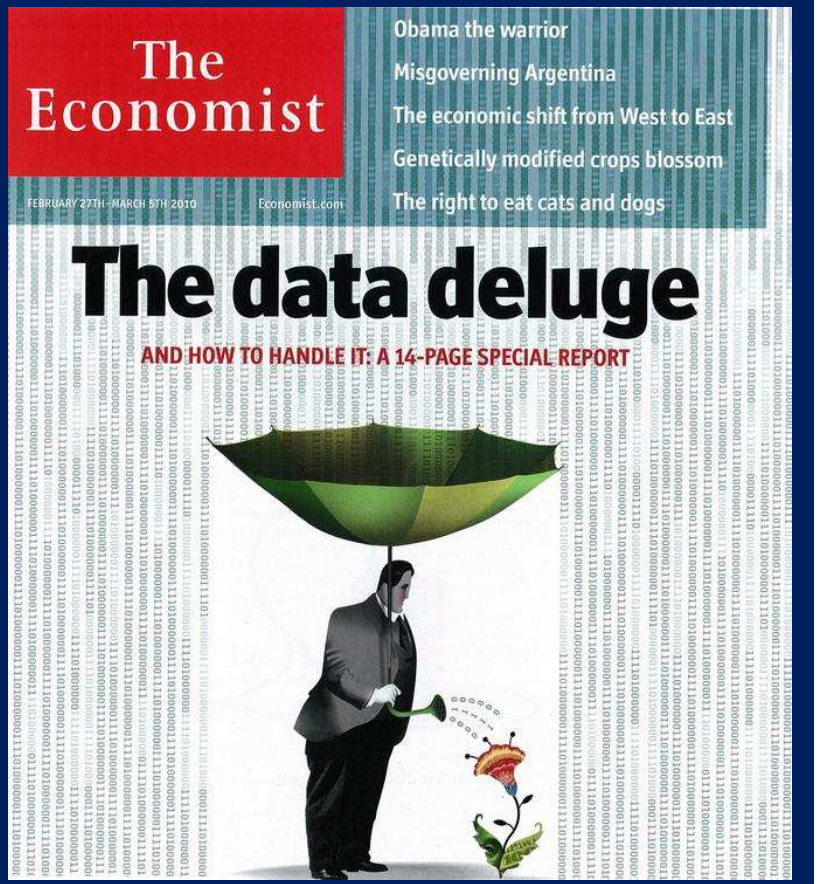
Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age

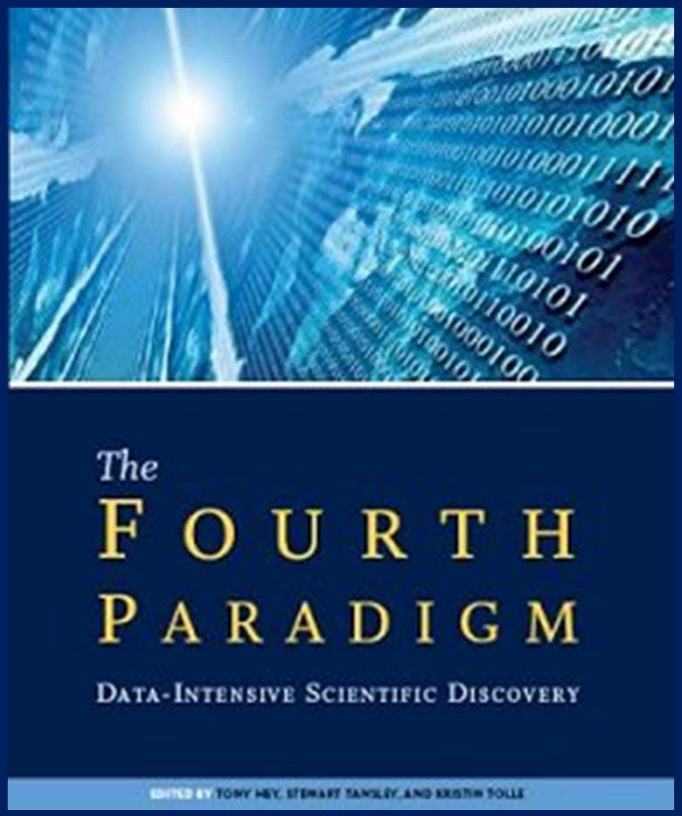
Stefano Cavuoti
Department of Physics – University Federico II – Napoli
INAF – Capodimonte Astronomical Observatory – Napoli

Massimo Brescia
INAF – Capodimonte Astronomical Observatory – Napoli
Department of Physics – University Federico II – Napoli

Giuseppe Longo
Department of Physics – University Federico II – Napoli
INAF – Capodimonte Astronomical Observatory – Napoli
Visiting Associate – Department of Astronomy – CALTECH – Pasadena



SPIE - Software and Cyberinfrastructure for Astronomy II – Amsterdam, July 1 2012



1. **Experiment** (ca. 3000 years)
2. **Theory** (few hundreds years) mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)
3. **Simulations** (few tens of years) Complex phenomena
4. **Data-Intensive science** (**now!!!**)

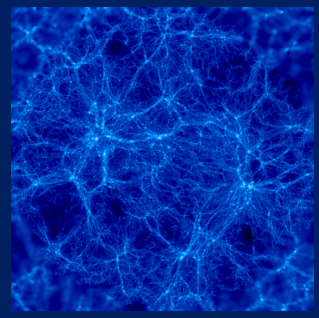
"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.
This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

http://research.microsoft.com/fourthparadigm/

The fourth paradigm relies upon….

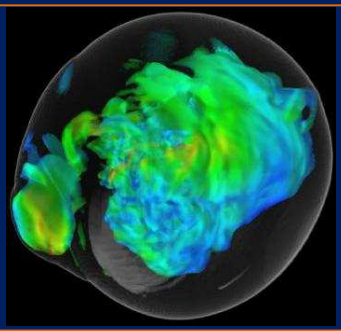
1. Most data will never be seen by human

→ Need for ML, KDD ecc.



2. Complex correlations *(precursors of physical laws)* cannot be visualized and recognized by the human brain

→ Most if not all empirical correlations depend on three parameters only: …
Simple universe or rather human bias?



3. Real world physics is too complex. Validation of models requires *accurate simulations, tools to compare simulations and data,* and better ways to deal with complex & massive data sets

→ Need to increase computational and algorithmic capabilities beyond current and expected technological trends

Data Intensive Science

Data Gathering (e.g., from sensor networks, telescopes…)

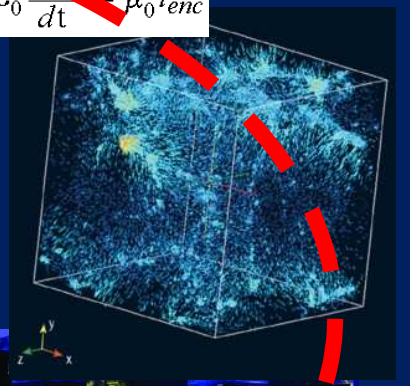
→ Data Farming:

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability, ontologies, etc.

$$\oint \mathbf{E}\cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$
$$\oint \mathbf{B}\cdot d\mathbf{A} = 0$$
$$\oint \mathbf{E}\cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$
$$\oint \mathbf{B}\cdot d\mathbf{s} = \mu_0\varepsilon_0\frac{d\Phi_E}{dt} + \mu_0 i_{enc}$$

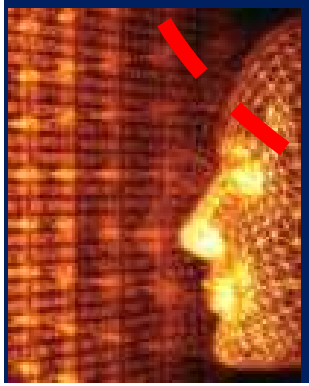
→ Data Mining (or Knowledge Discovery in Databases):

Pattern or correlation search
Clustering analysis, automated classification
Outlier / anomaly searches
Hyperdimensional visualization



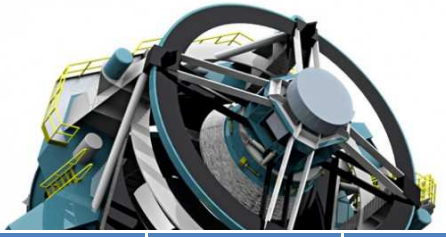
→ Data understanding

Computer aided understanding
KDD
Etc.

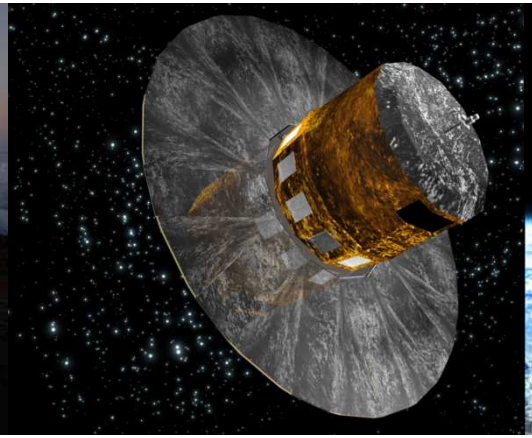


→ New Knowledge

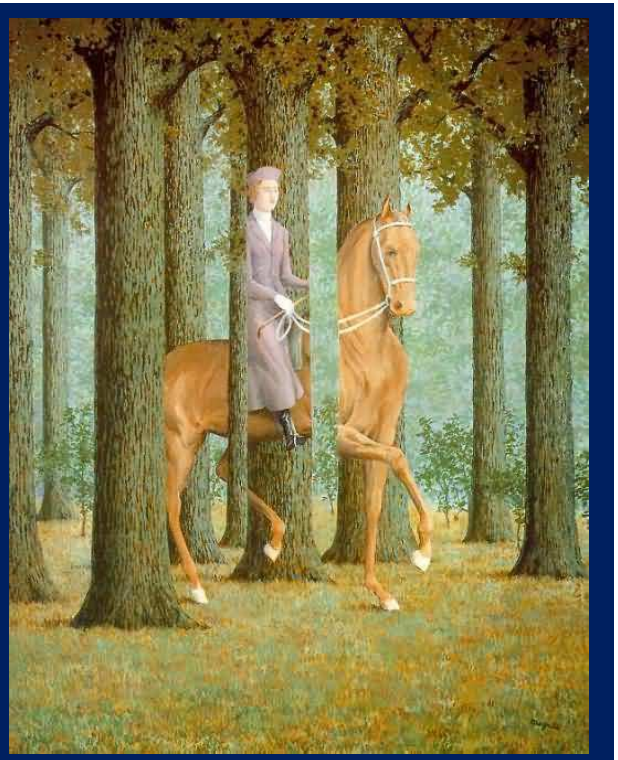
Data Sources



	TB	Total	epochs	parameters
VST	0.15 TB/day	100 TB	tens	>100
HST		120 TB	few	>100
PANSTARRS		600 TB	Few-many	>>100
LSST	30 TB/day	> 10 PB	hundreds	>>100
GAIA		1 PB	many	>>100 heterogeneous
SKA	1.5 PB/day		>> 10^2	hundreds



MASSIVE, COMPLEX DATA SETS with: $N > 10^9$, $D >> 100$, $K > 10$



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

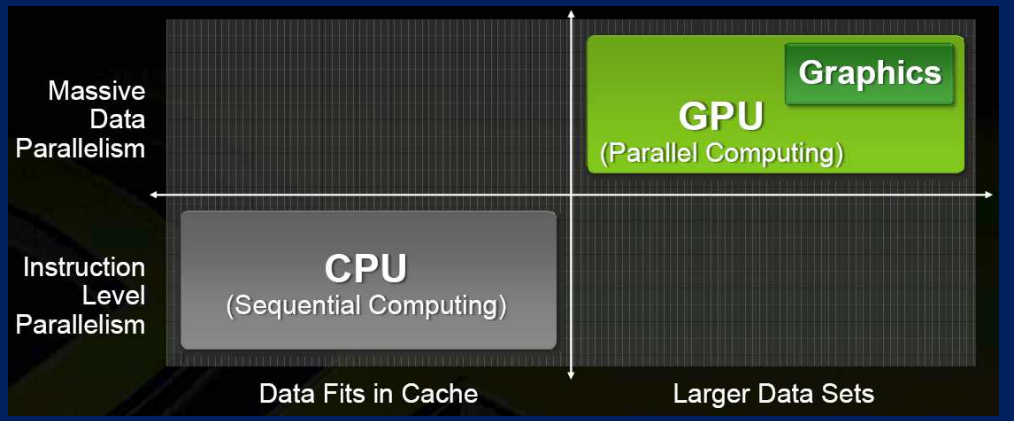
K-means: $K \times N \times I \times \mathbf{D}$
Expectation Maximization: $K \times N \times I \times \mathbf{D^2}$
Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$
Correlations ~ N log N or N^2, ~ D^k ($k \geq 1$)
Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k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

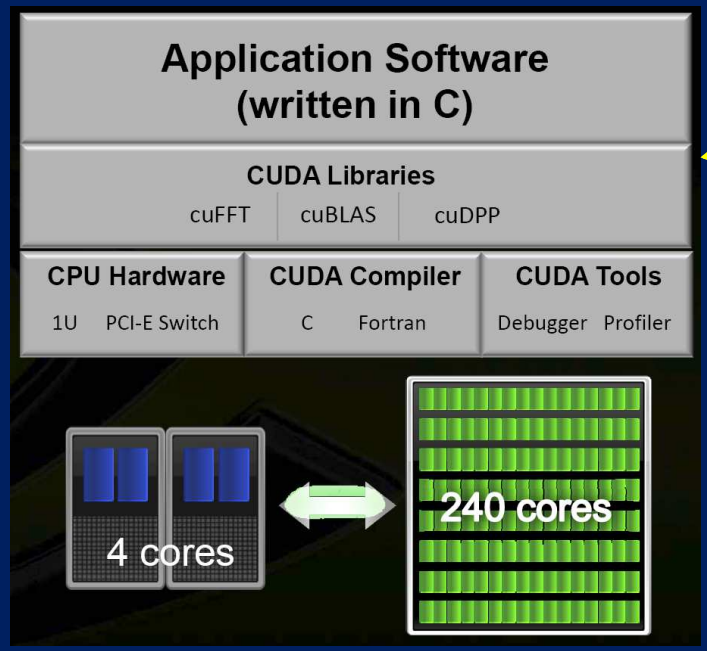
Lots of computing power

… 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 »



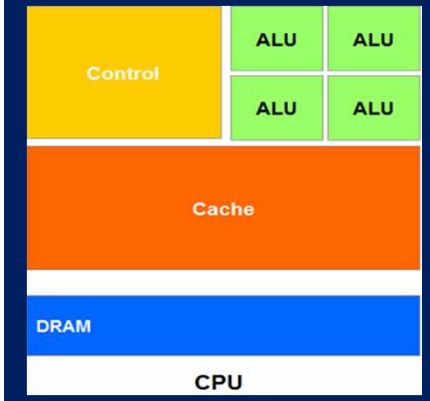


DAME - GAME
Genetic Algorithm Mining Experiment

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).

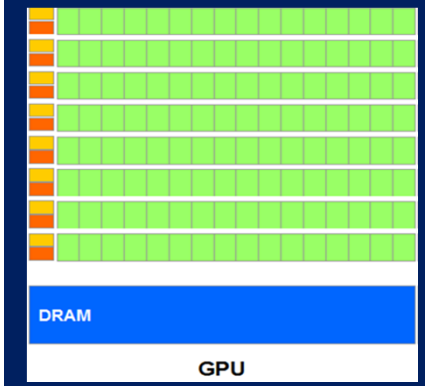
It is intrinsically parallel.

GPU vs CPU



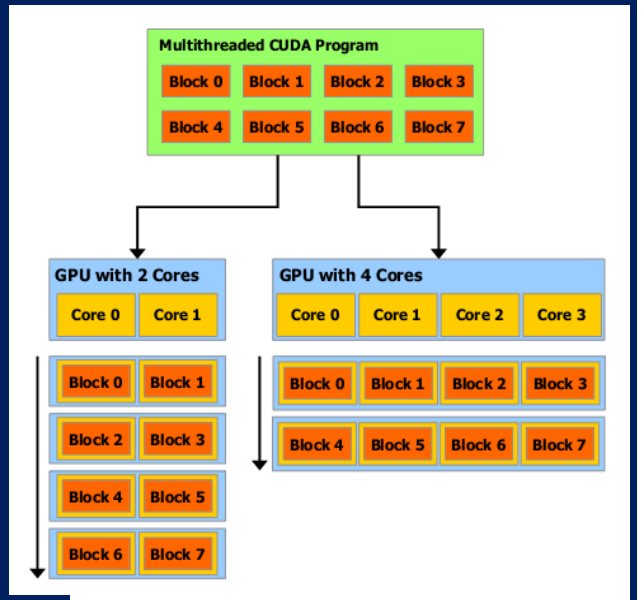
Multi-core CPU

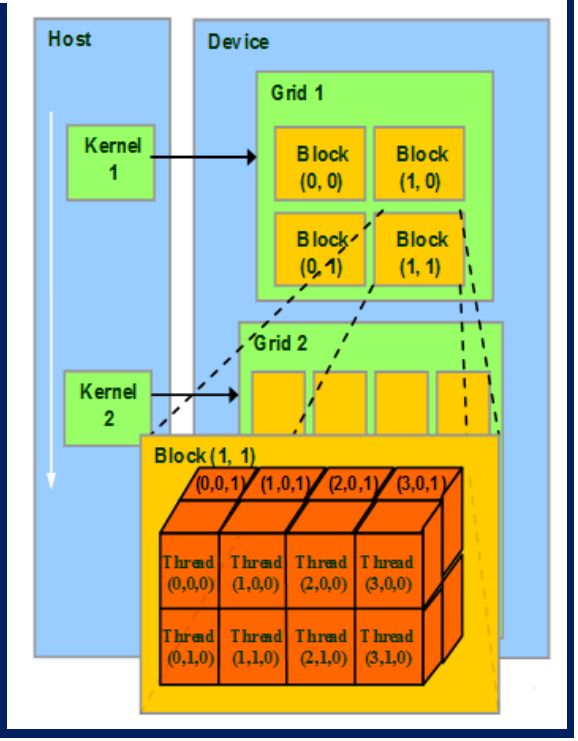
- Composed by few cores, designed to maximize the sequential code efficiency;
- Large cache memory to reduce latency time to access data and/or complex instruction execution;
- Sophisticate control logic to handle instruction flow (pipelining and multi-threading).



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);



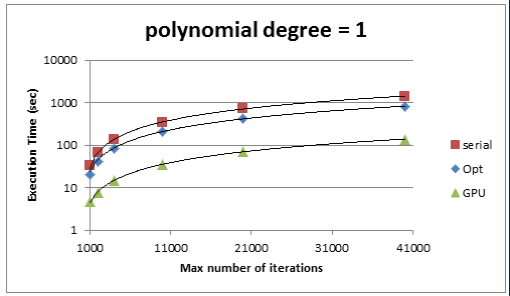


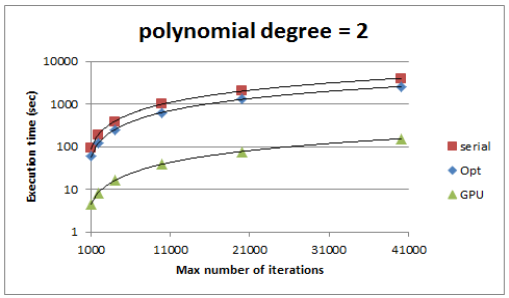
GAME GPU performances

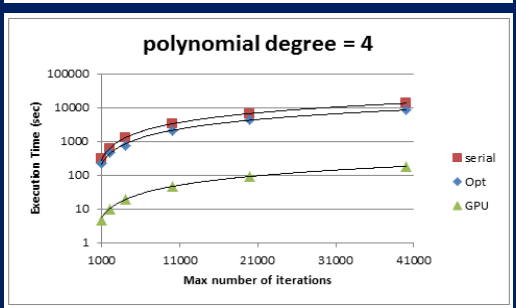
ID	CPU	GPU	Pol. Degree	DATASET	iterations	Exe time
1	2.0 GHz Intel i7 2630QM quad core		8	2100 patterns 11 features	40000	31092 sec (~9 h)
1		GeForce Tesla TM C1060 (240 cores)				231 sec (~0.064 h) 0,7% of CPU time
2	3.4 GHz Intel i7 2600 dual core					76000 sec (~21 h)
2		GeForce GTX 460 (336 cores)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

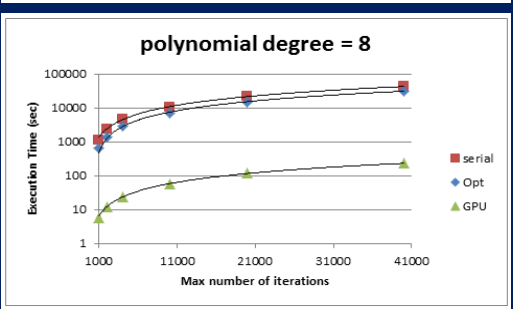
GPU Speedup (Tesla vs i7)		
degree	vs. Serial	vs. Opt
1	8x	6x
2	23x	16x
4	66x	45x
8	200x	125x

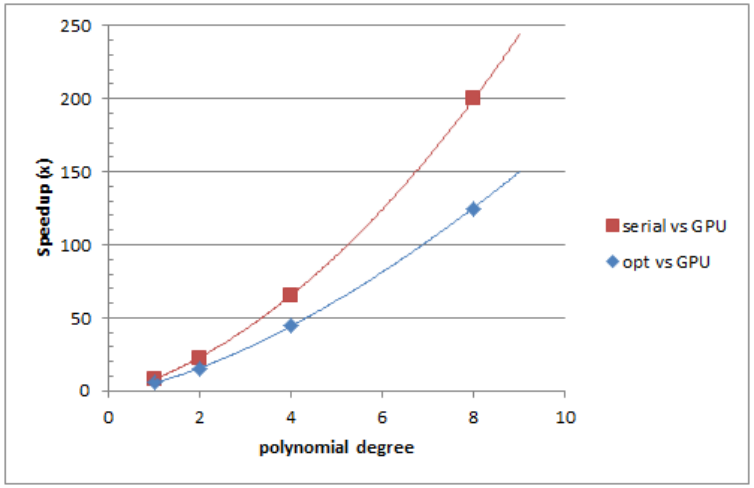
- The increase of the polynomial degree enhances the speedup difference.
- It results evident that the speedup is as much as highest is the complexity of the fitness function











DAME Program

DAME Program is a joint effort between University Federico II, Caltech and INAF-OACN, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration on top of a virtualized distributed computing environment.

Multi-purpose data mining with machine learning Web App REsource

Extensions
- DAME-KNIME
- ML Model plugin

Specialized web apps for:
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality

Web Services:
- SDSS mirror
- WFXT Time Calculator
- GAME (GPU+CUDA ML model)

http://dame.dsf.unina.it/
Science and management
Documents
Science cases
Newsletters

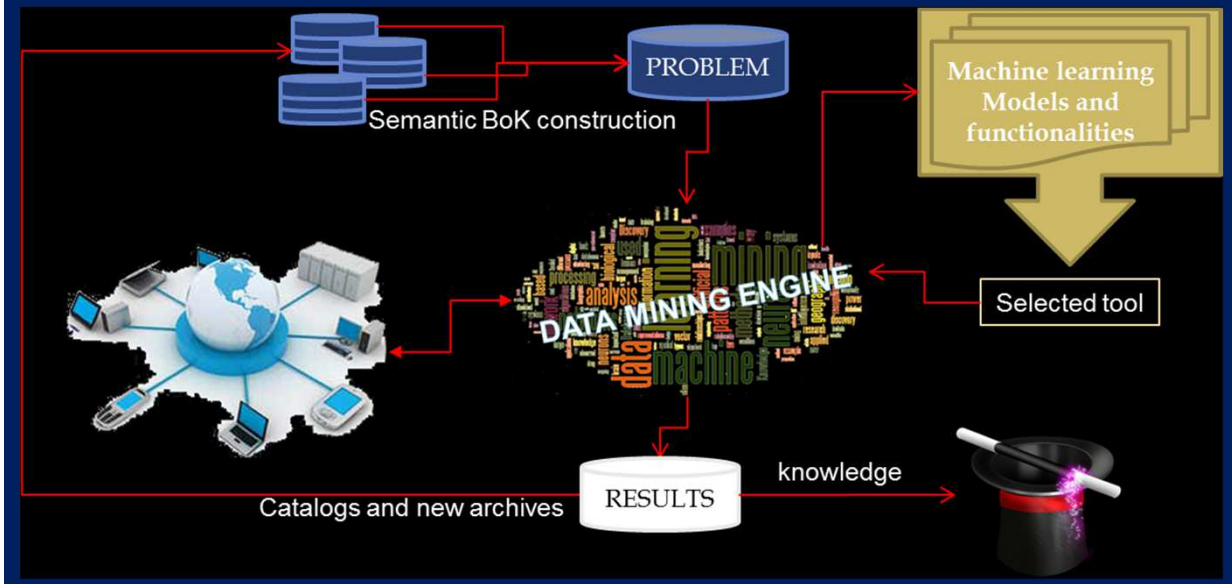
http://www.youtube.com/user/DAMEmedia
DAMEWARE Web Application media channel

DAME Main Project: DAMEWARE

DAta Mining Web Application REsource

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

web-based app for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure



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

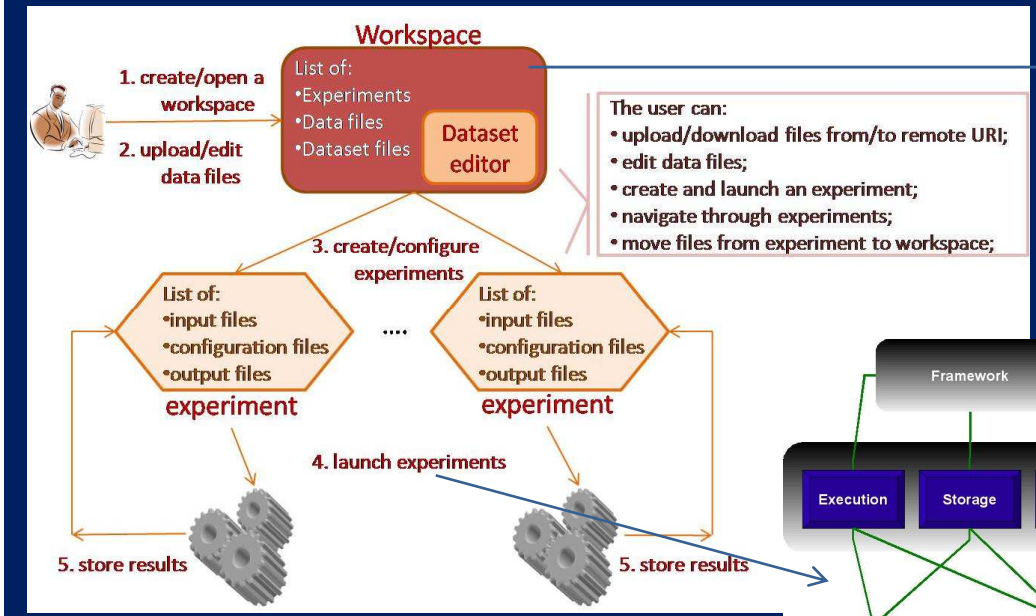
MLP with Levenberg-Marquardt

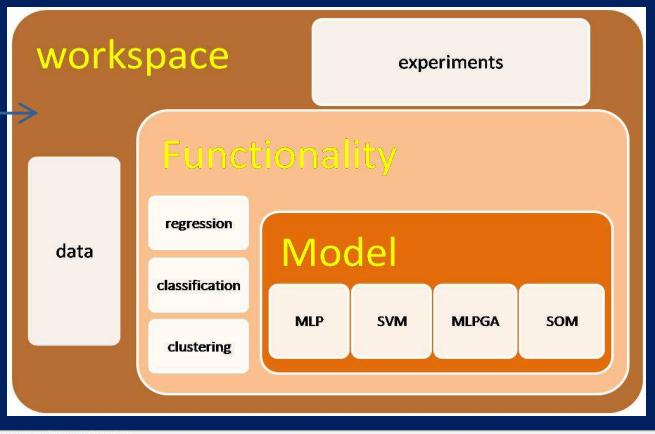
DAMEWARE fundamentals

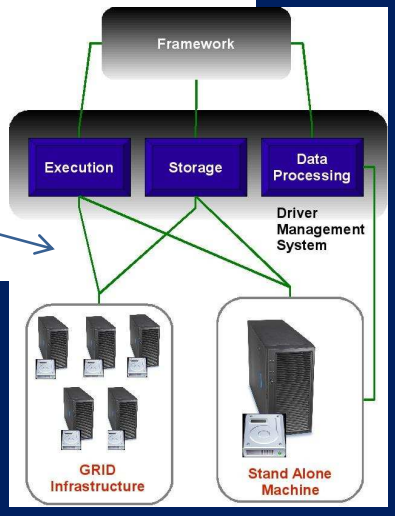
Based on the X-Informatics paradigm, it is multi-disciplinary platform (until now X = Astro)

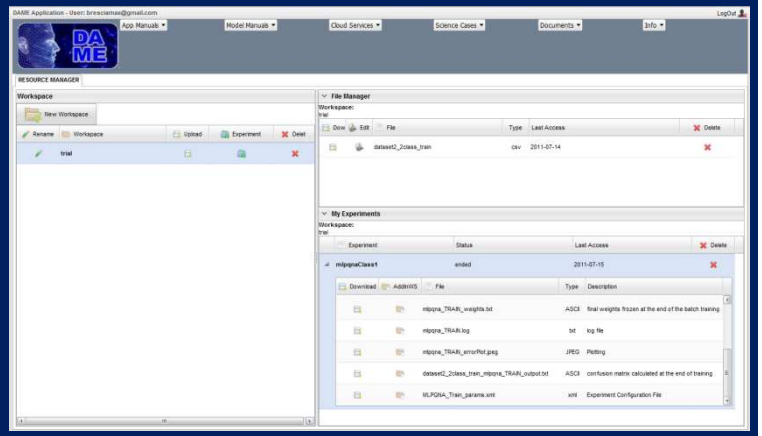
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

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









DAME Science case examples

DAME has been successfully applied to a variety of scientific cases:

AGN identification and classification

Cavuoti, S.; Brescia, M.; D'Abrusco R.; Longo G., *Photometric AGN Classification in the SDSS with Machine Learning Methods*, (in preparation)

Globular Cluster classification

Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T., 2012, *The detection of Globular Clusters in galaxies as a data mining problem*, MNRAS, 421, 2, 1155-1165

Evaluation of photometric redshifts

D'Abrusco et al. 2007, *Mining the SDSS Archive I. Photometric redshifts in the nearby universe*, ApJ., 663, 752

Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A., 2012, *Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest*, Submitted to Astronomy & Astrophysics, arxiv:1206.0876v2

Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A., *High Accuracy Photometric Redshifts for Quasars*, (in preparation);

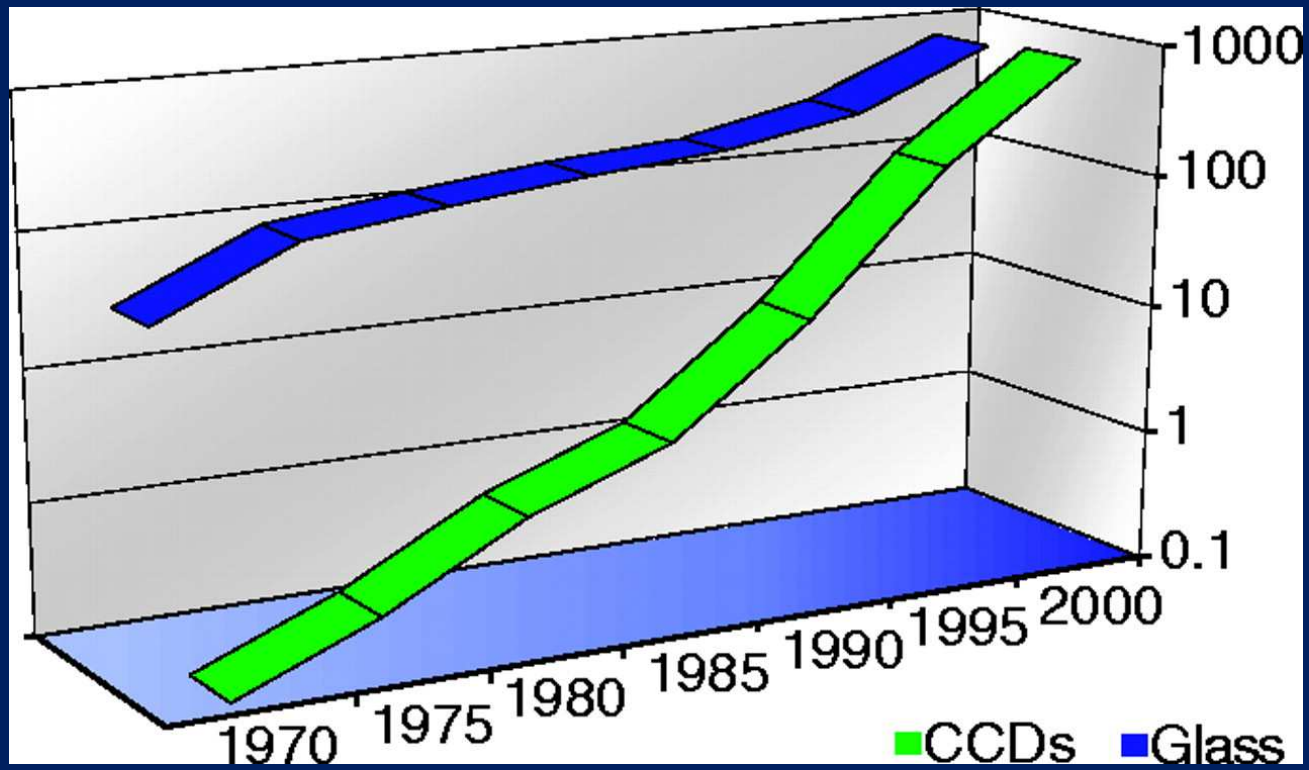
Candidate quasar identification

D'Abrusco R., Longo G., Walton N.A., *Quasar candidate selection in the Virtual Observatory era*, 2009, MNRAS, 396, 223

We refer the interested readers to these papers.

MDS Moving

According to the Nielsen's Law Network bandwidth double every 21 month; instead data in astrophysics are growing exponentially, with a doubling time of 1.5 years.
(courtesy G. S. Djorgovski)

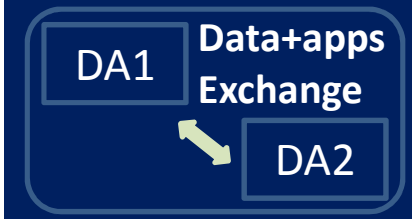


Moving programs not data: the true bottle neck

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)

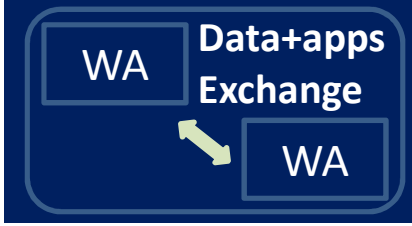
Local user desktop fully involved (requires computing power)

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)

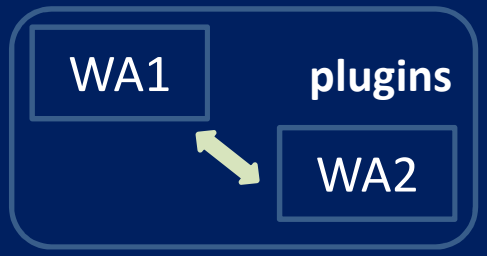


Except from URI exchange, no interoperability and different accounting policy

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

No local computing power required

The new vision for KDD



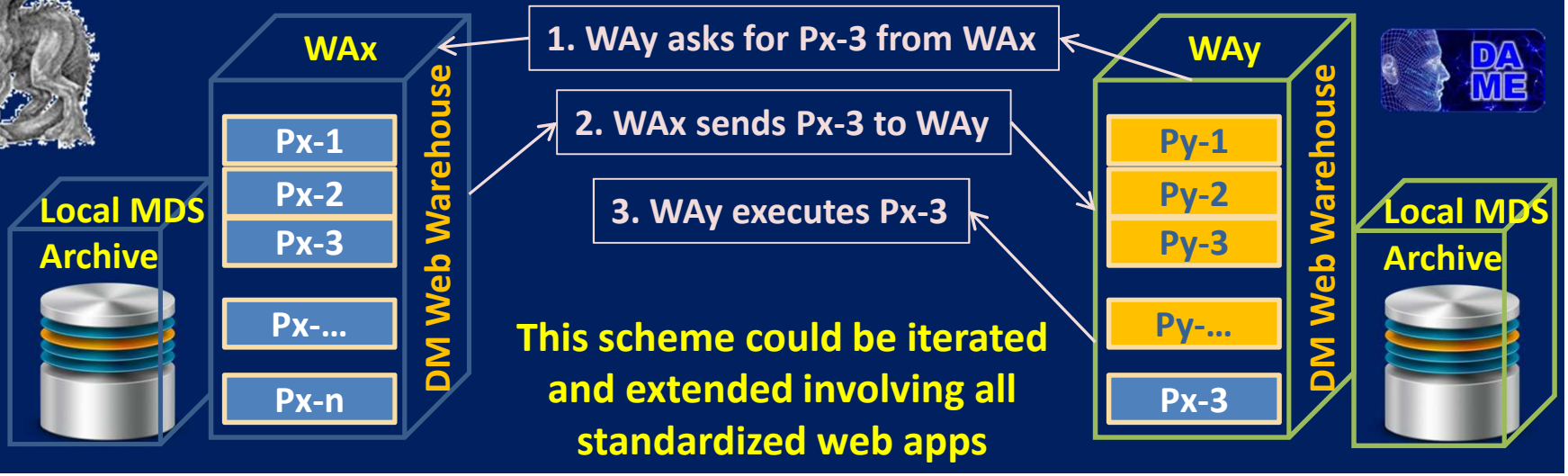
All DAs must become WAs
Unique accounting policy (google/Microsoft like)
To overcome MDS flow, apps must be plug & play
(*e.g. any WAx feature should be pluggable in WAy on demand*)

No local computing power required.
Also smartphones can run DM 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);

The Lernaean Hydra DAME KDD (*plugin granularity*)



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

The Lernaean Hydra DAME KDD

After a certain number of such iterations…

The scenario will become:

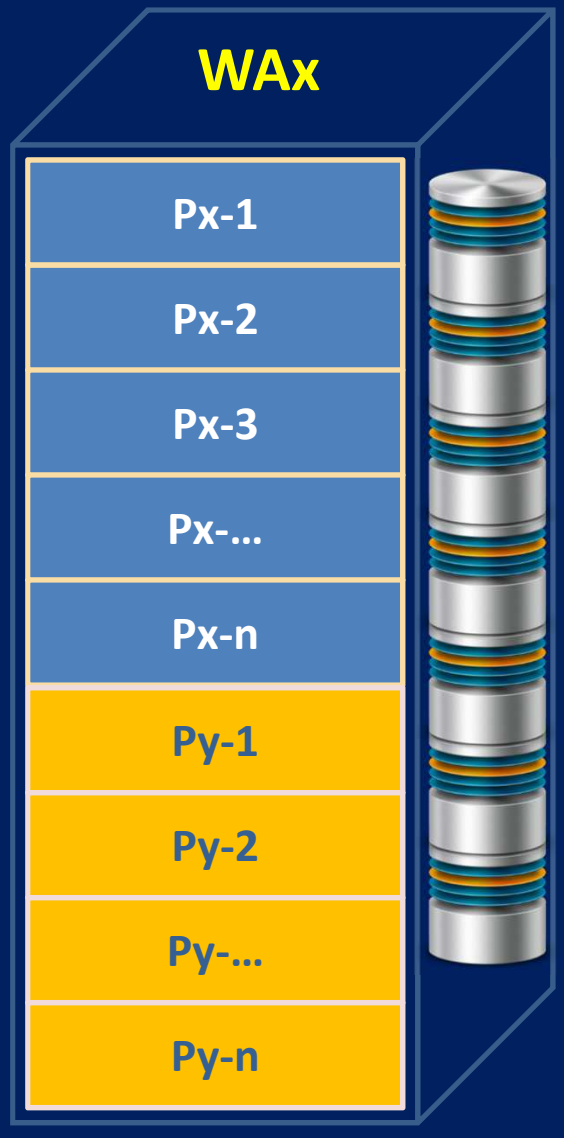
No different WAs, but simply one WA with several sites (eventually with different GUIs and computing environments)

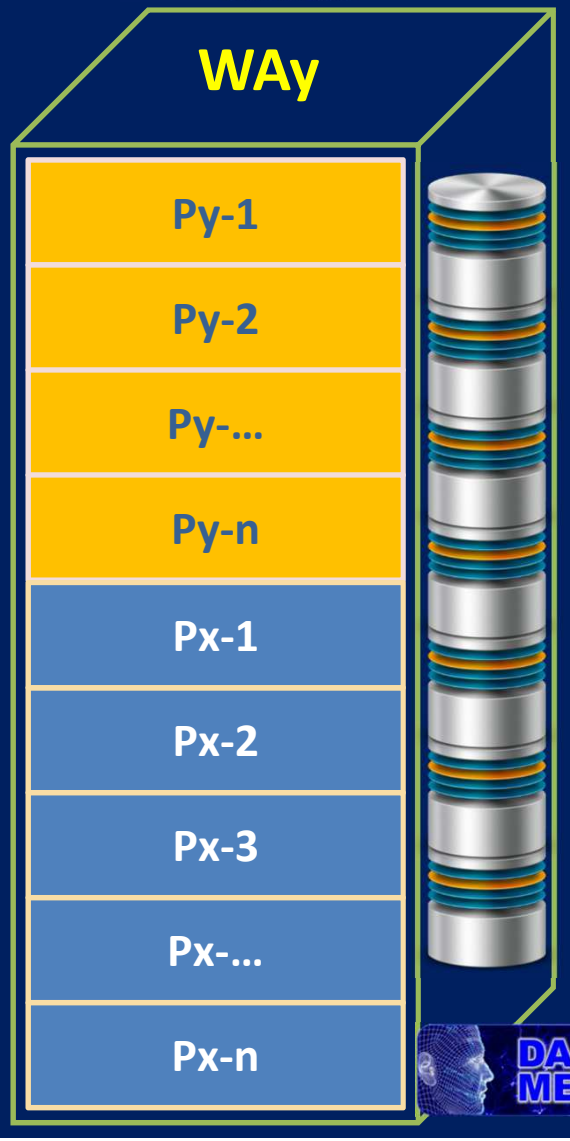
All WA sites can become a mirror site of all the others

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)

YES MDS!





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 also in a VO environment.

1. DAME is just an example of what new technologies (GPU and 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).