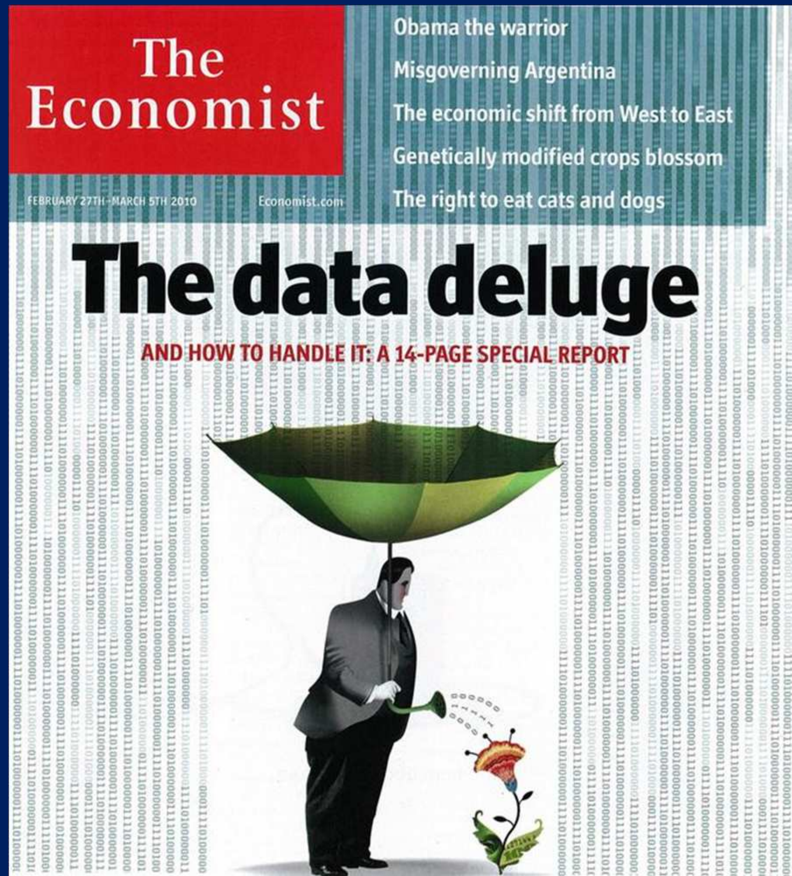
Computational Science and new perspectives for the analysis of massive data sets

Giuseppe Longo

University Federico II – Napoli
Associate
California Institute of Technology

Massimo Brescia

INAF – Capodimonte Observatory in Napoli

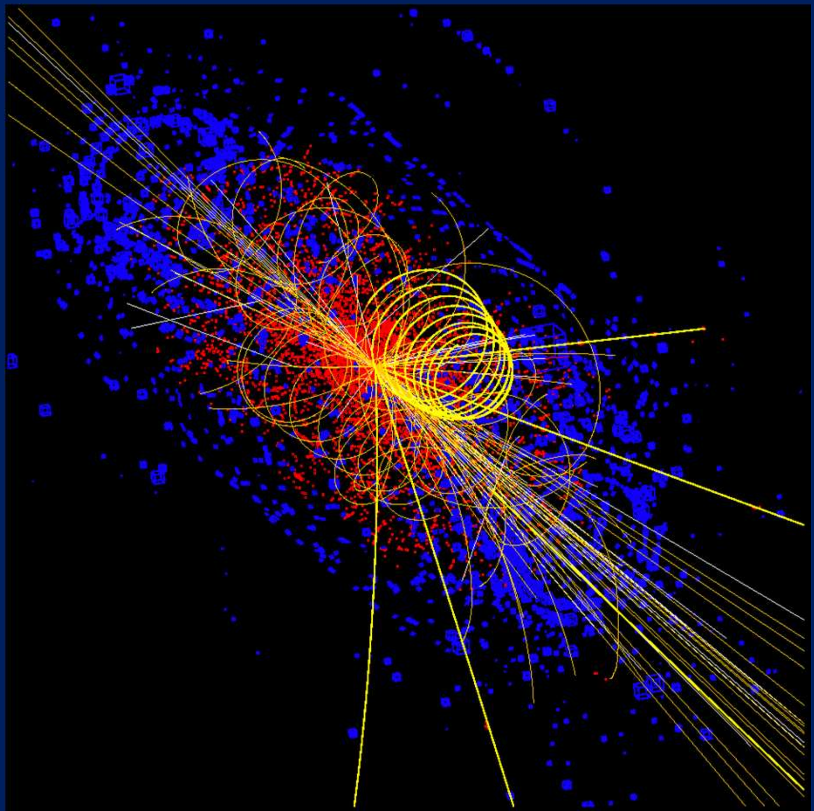


2nd Int. Conf. «Frontiers in Diagnostic Technologies
November 28-30, 2011, INFN National Laboratories
Frascati, Italy

Summary

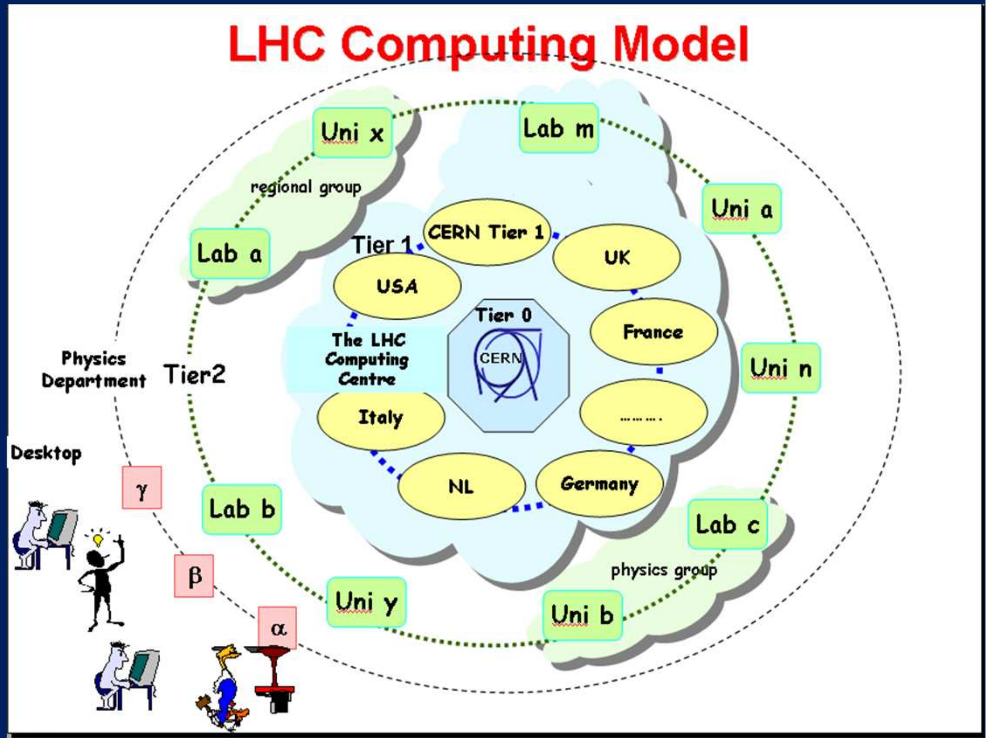
- The data tsunami
- Virtual Organizations
- A new scientific paradigm
- Bottlenecks: moving programs not data
- Knowledge discovery in databases
- Conclusions

The forerunner: LHC



ATLAS detector event

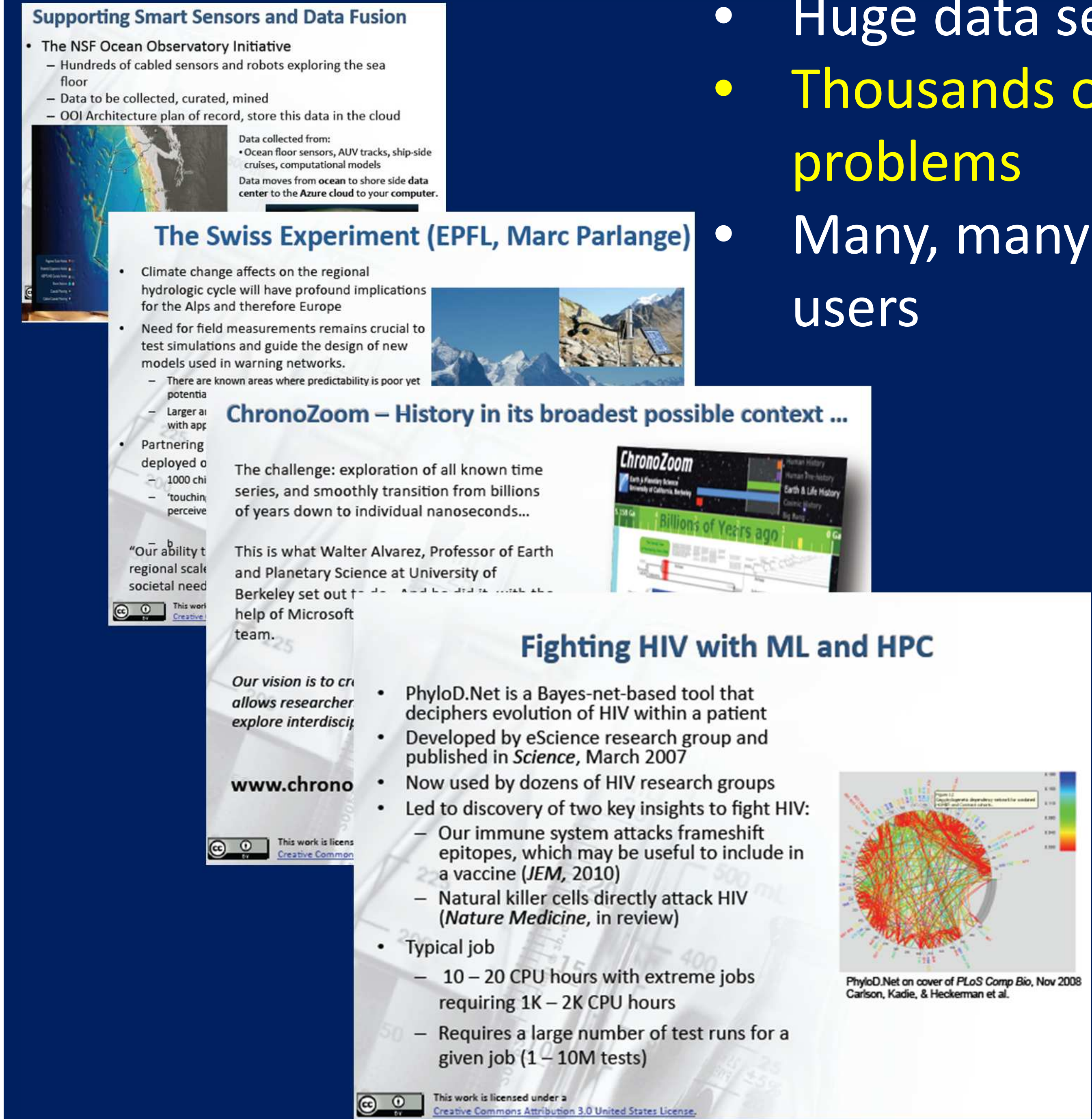
Data Stream: 330 TB/week



Computationally demanding but still a relatively simple (embarassingly parallel) KDD task

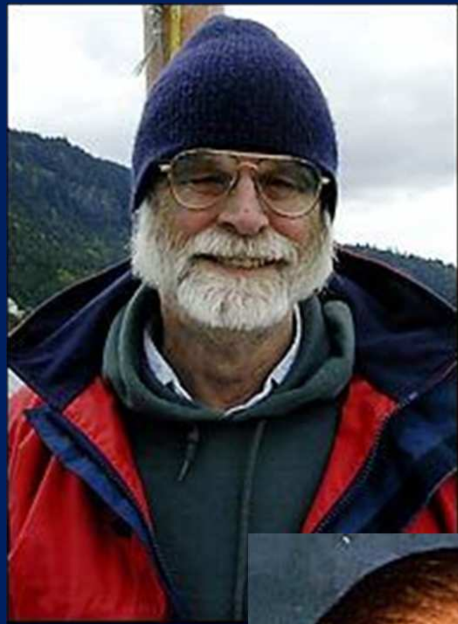
Pruning of uninteresting events and detection of specific ones either known from simulations or outliers

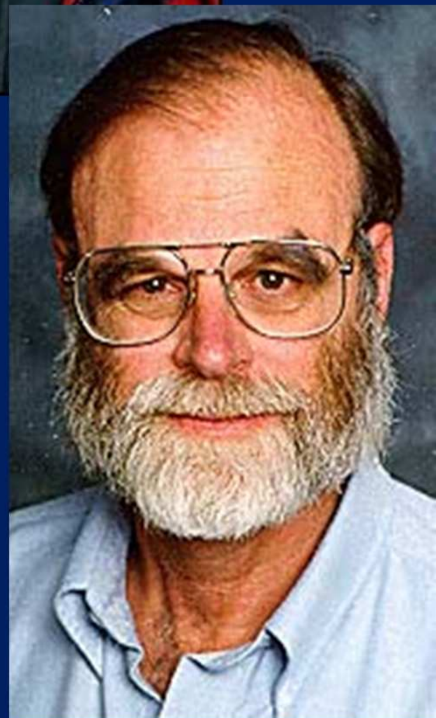
- Huge data sets (ca. Pbyte)
- Thousands of different problems
- Many, many thousands of users

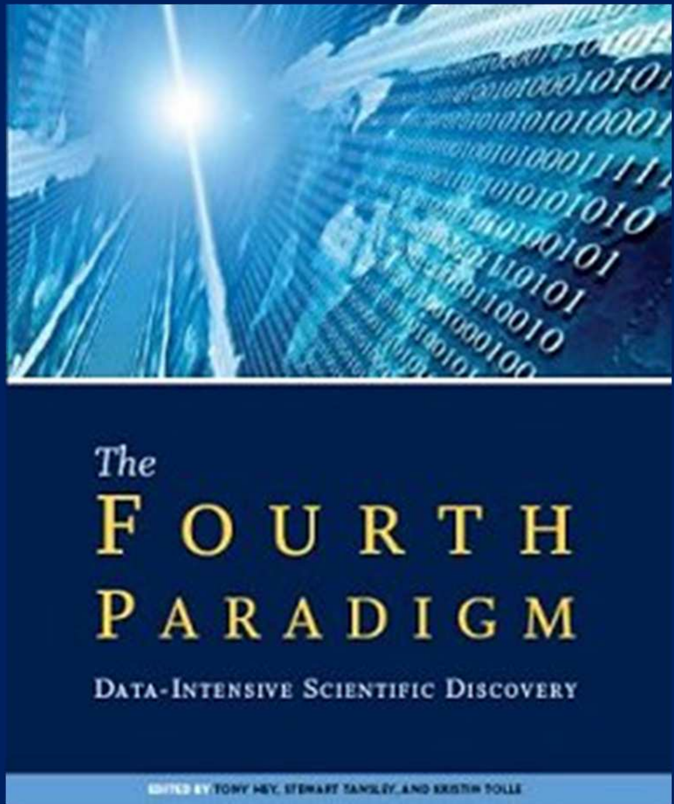


Jim Gray

"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."







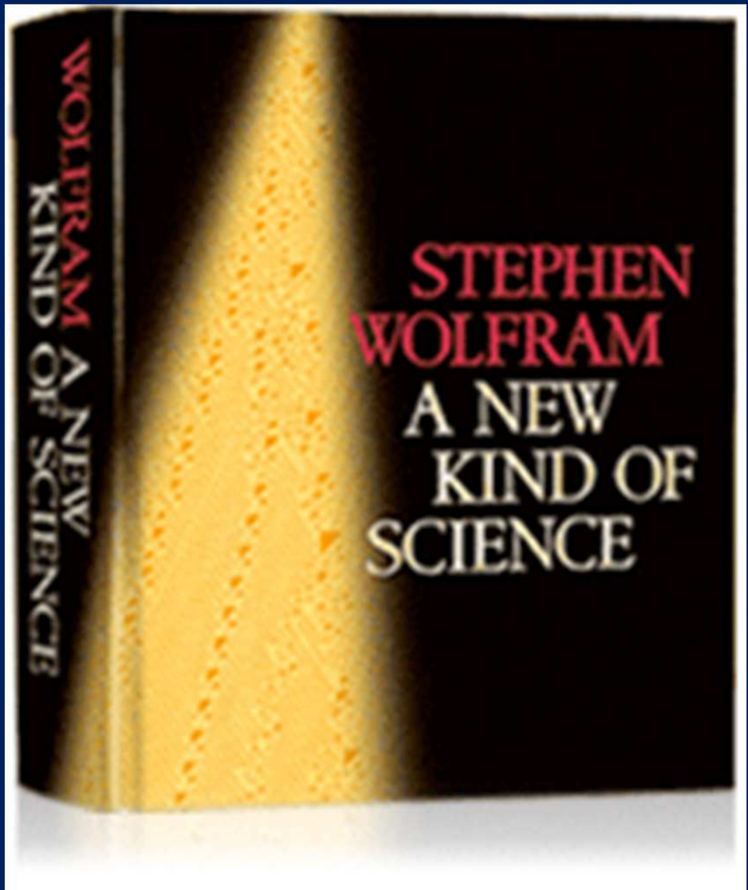
1. **Experiment** (ca. 3000 years)

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!!**)

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



An Outline of Basic Ideas

Three centuries ago science was transformed by the dramatic new idea that rules based on mathematical equations could be used to describe the natural world. My purpose in this book is to initiate another such transformation, and to introduce a new kind of science that is based on the much more general types of rules that can be embodied in simple computer programs.

It has taken me the better part of twenty years to build the intellectual structure that is needed, but I have been amazed by its results. For what I have found is that with the new kind of science I have developed it suddenly becomes possible to make progress on a remarkable range of fundamental issues that have never successfully been addressed by any of the existing sciences before.

If theoretical science is to be possible at all, then at some level the systems it studies must follow definite rules. Yet in the past throughout the exact sciences it has usually been assumed that these rules must be ones based on traditional mathematics. But the crucial realization that led me to develop the new kind of science in this book is that there is in fact no reason to think that systems like those we see in nature should follow only such traditional mathematical rules.

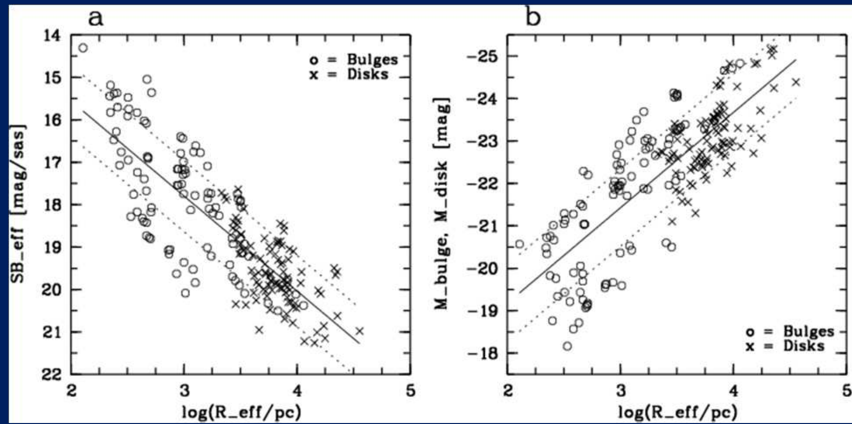
http://www.wolframscience.com/nksonline/toc.html

The fourth paradigm relies upon….

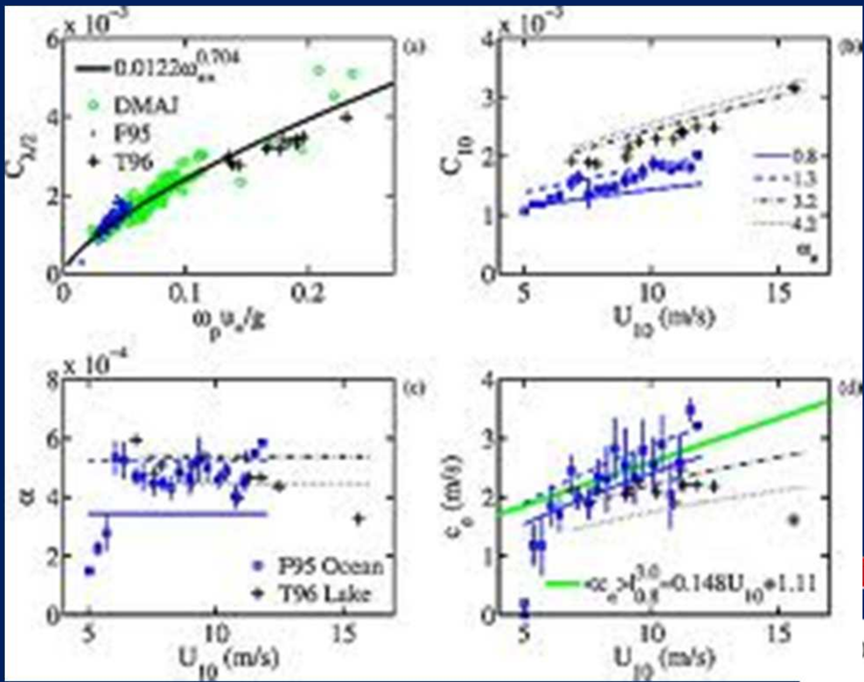
1. Most data will never be seen by human → **Need for ML, KDD ecc.**

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?

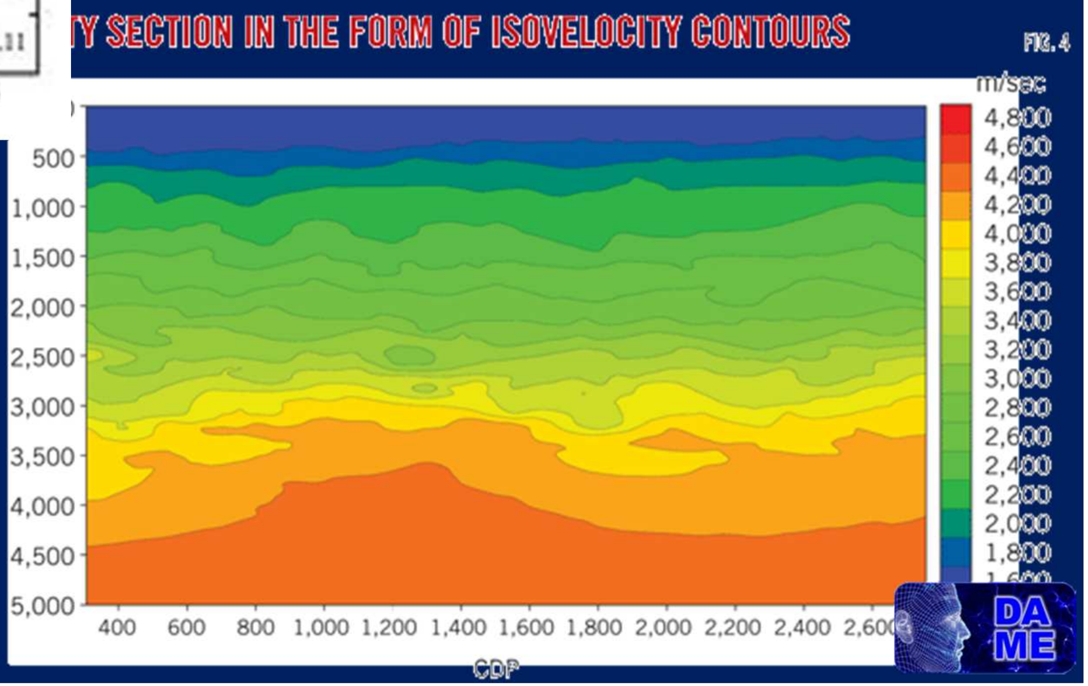
First hint about the need for complex visualization



Astronomy



Oceanography

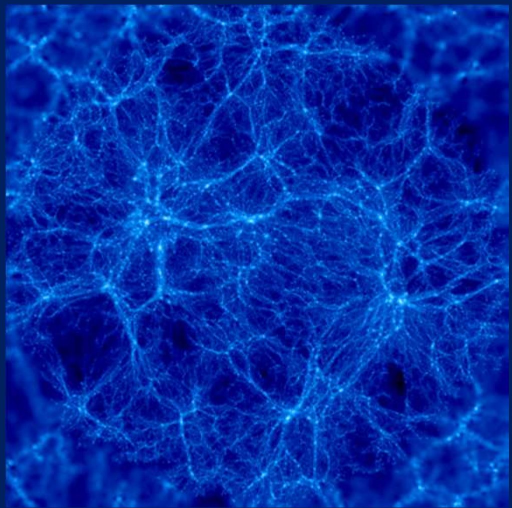


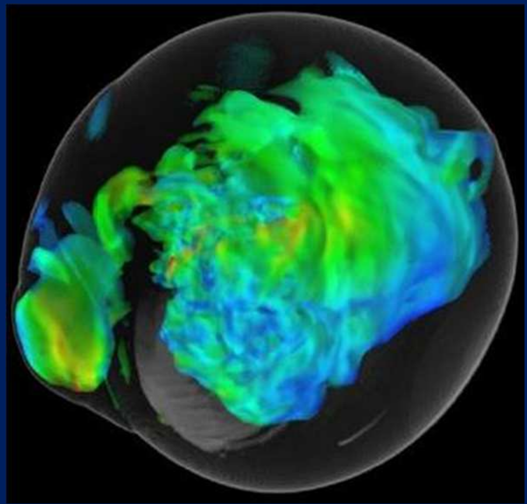
Petrology

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

$$\oint \mathbf{E}\cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$

$$\oint \mathbf{B}\cdot d\mathbf{A} = 0$$

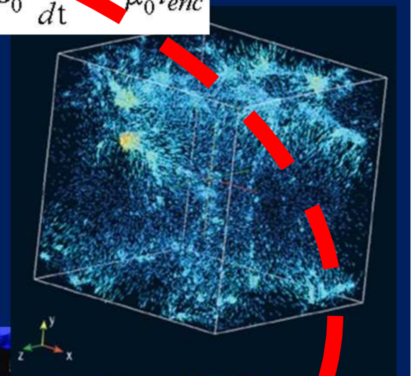
$$\oint \mathbf{E}\cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$

$$\oint \mathbf{B}\cdot d\mathbf{s} = \mu_0\varepsilon_0\frac{d\Phi_E}{dt} + \mu_0 i_{enc}$$

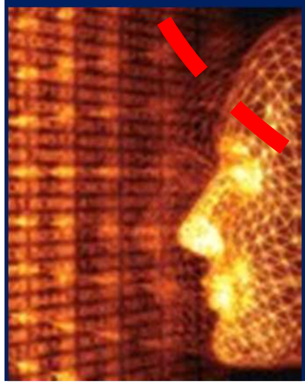
Data Gathering (e.g., from sensor networks, telescopes…)

→ Data Farming:
- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability, ontologies, etc.

→ 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

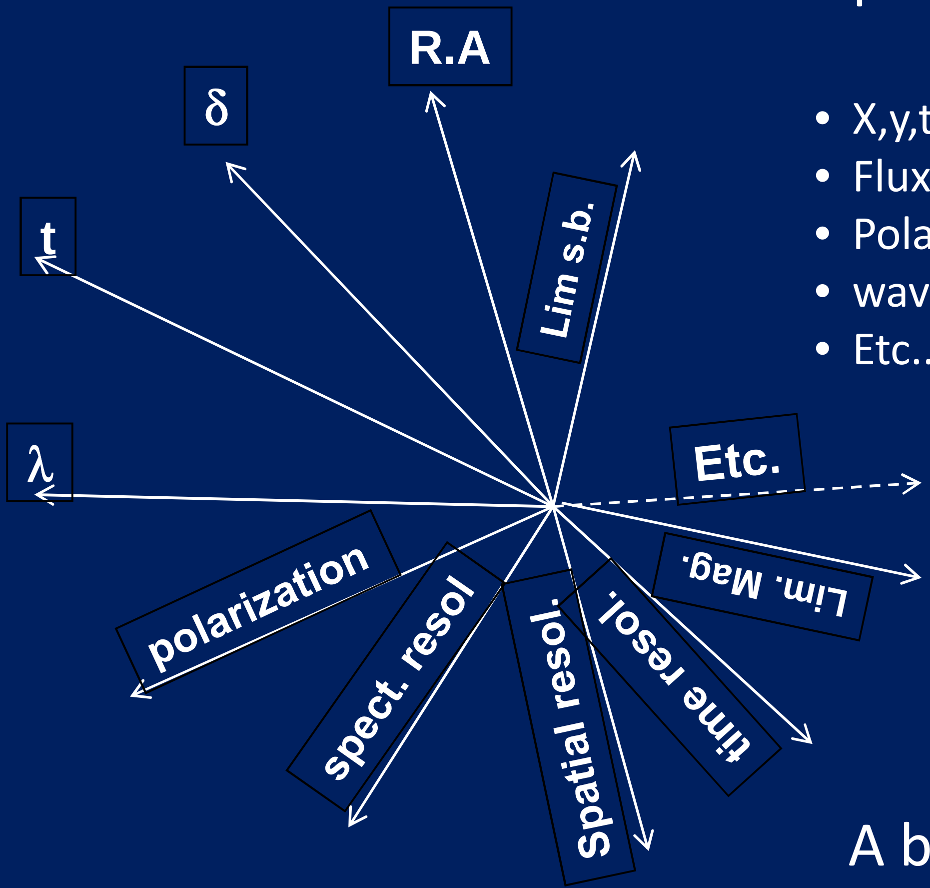
Distributed data sets and virtual organizations

The International Virtual Observatory Alliance

	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
US-Meteo		460 TB/yr		Hundreds heterogeneous
You Tube		530 TB		
Google	1 Pbyte/min			heterogeneous

The measurable parameter space of KDD

Each datum is defined by n measured parameters.



- X,y,t
- Flux
- Polarization
- wavelength
- Etc..

New sensor technologies:

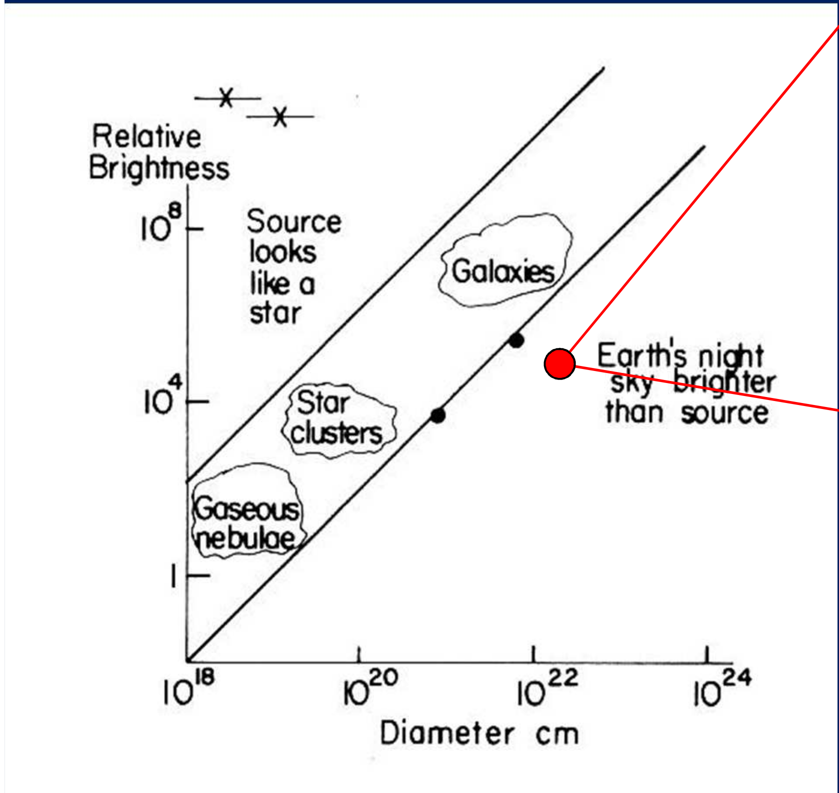
$$p \in \Re^N \qquad N >> 100$$

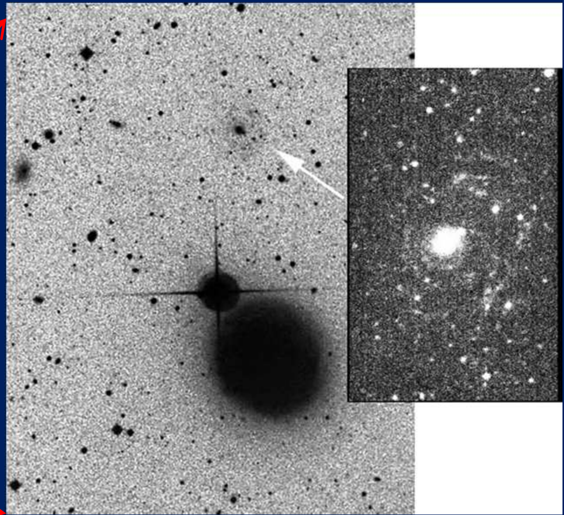
A better exploration and sampling of an ever increasing parameter space of **data intensive science**

An astronomical example

The astronomical parameter space is of high dimensionality, still sparsely covered and poorly sampled:

every time you improve either coverage or sampling you make new discoveries





Malin 1

a new type of low surface brightness galaxies (Malin, 1991)

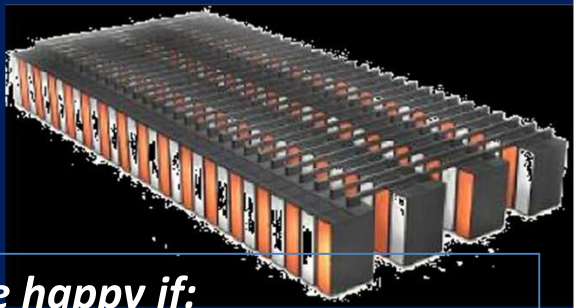
MASSIVE, COMPLEX DATA SETS with: $N > 10^9$, $D >> 100$, $K > 10$

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 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 > ~ $(N \times D)^3$

→ **Lots of computing power**

Scalability: 1-st bottle neck



Exaflop (are needed for simulations, metereology, data fusion, data mining, etc.)

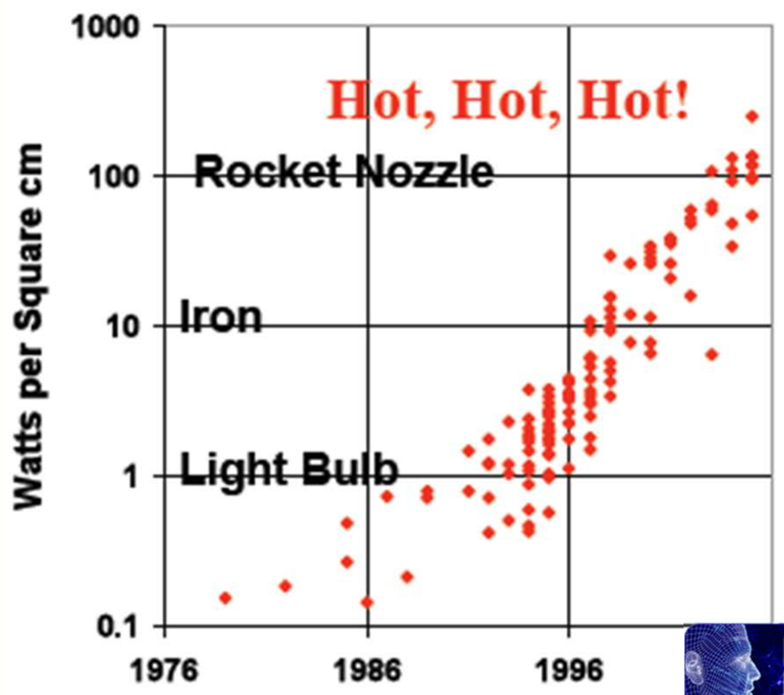
Exaflop = *100 x present capability*

Exascale != Exaflops but
Exascale at the data center size
Exascale at the "rack" size
embedded => ***Teraflops in a cube***

To reach exaflops required 14 yrs
But…

We should be happy if:

- *1.000.000 CPU's*
- *power supply of a nuclear plant*
- *Minimum changes in software*

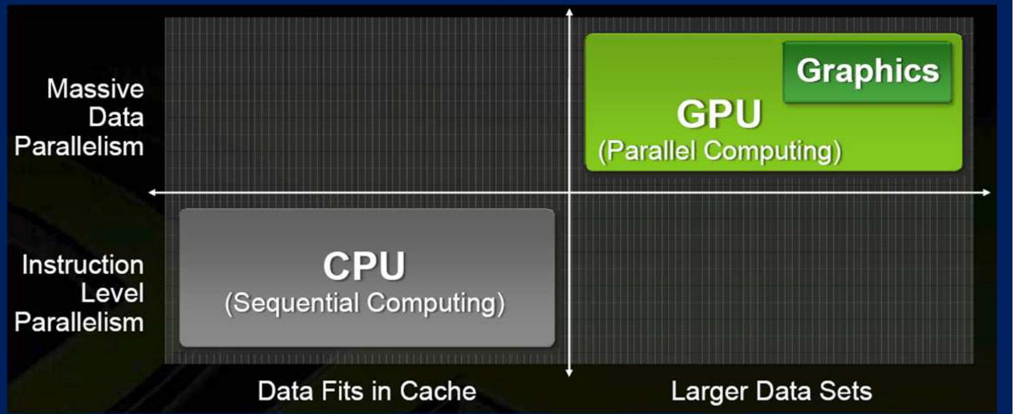


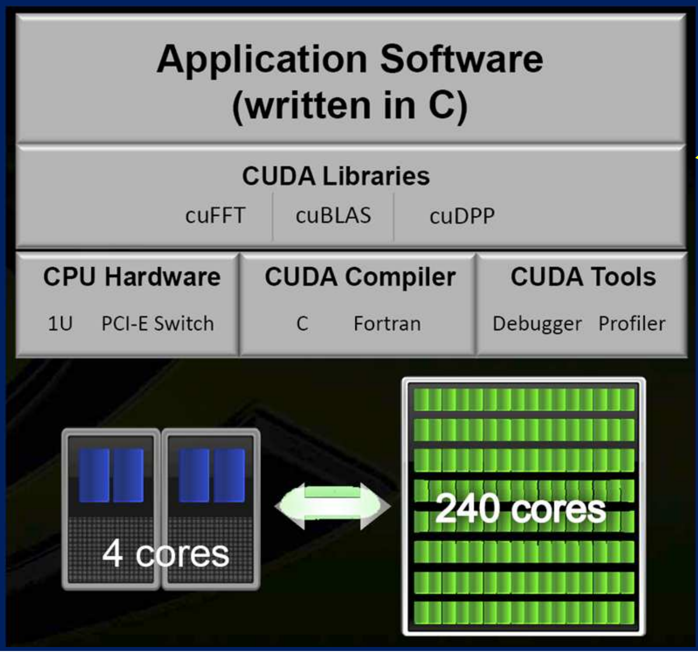
Adapted from Kogge– Astroinformatics 2010

… 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 and it is now under GPU+CUDA implementation.

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

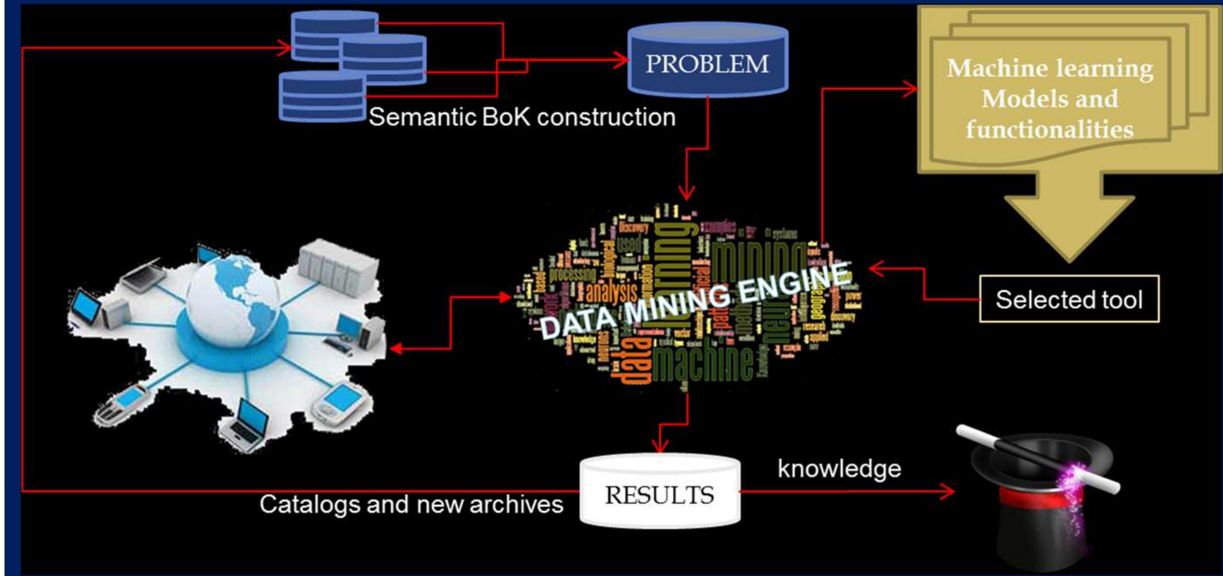
http://www.youtube.com/user/DAMEmedia
DAMEWARE Web Application media channel

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

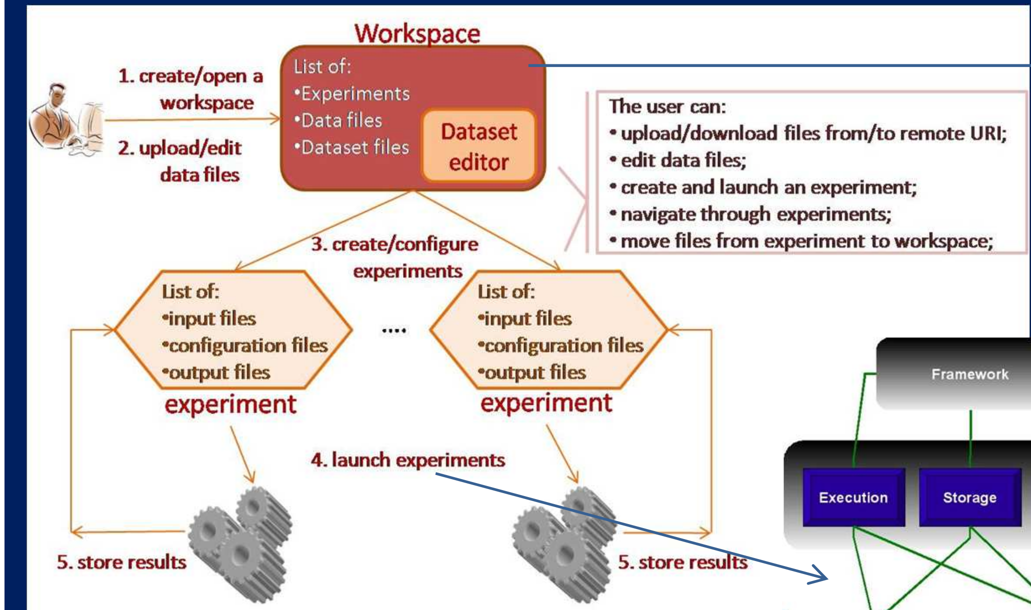
MLP with Levenberg-Marquardt

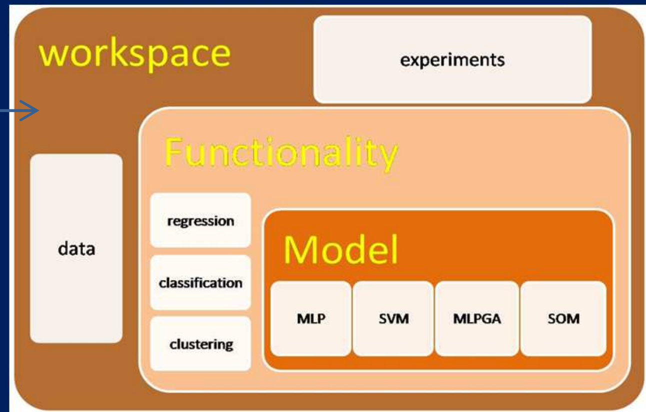
DAMEWARE fundamentals

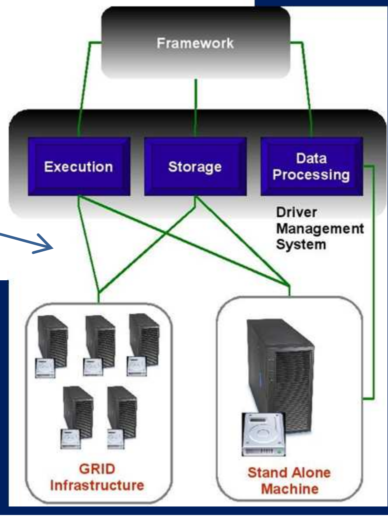
Based on the X-Informatics paradigm, it is multi-disciplinary platform (until now X = Astro)

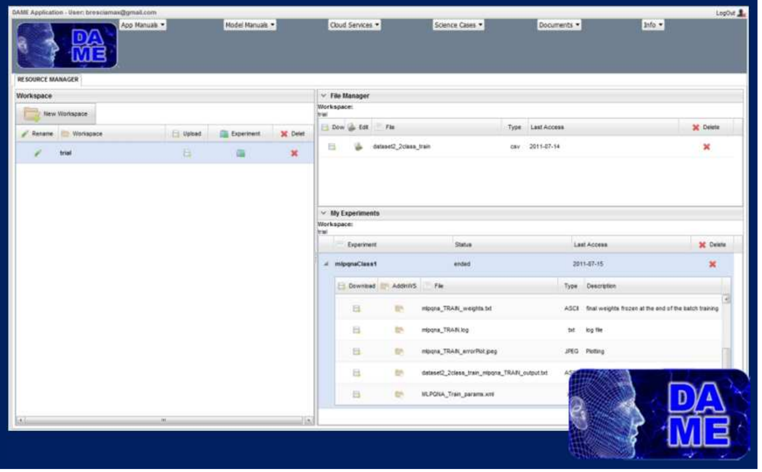
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

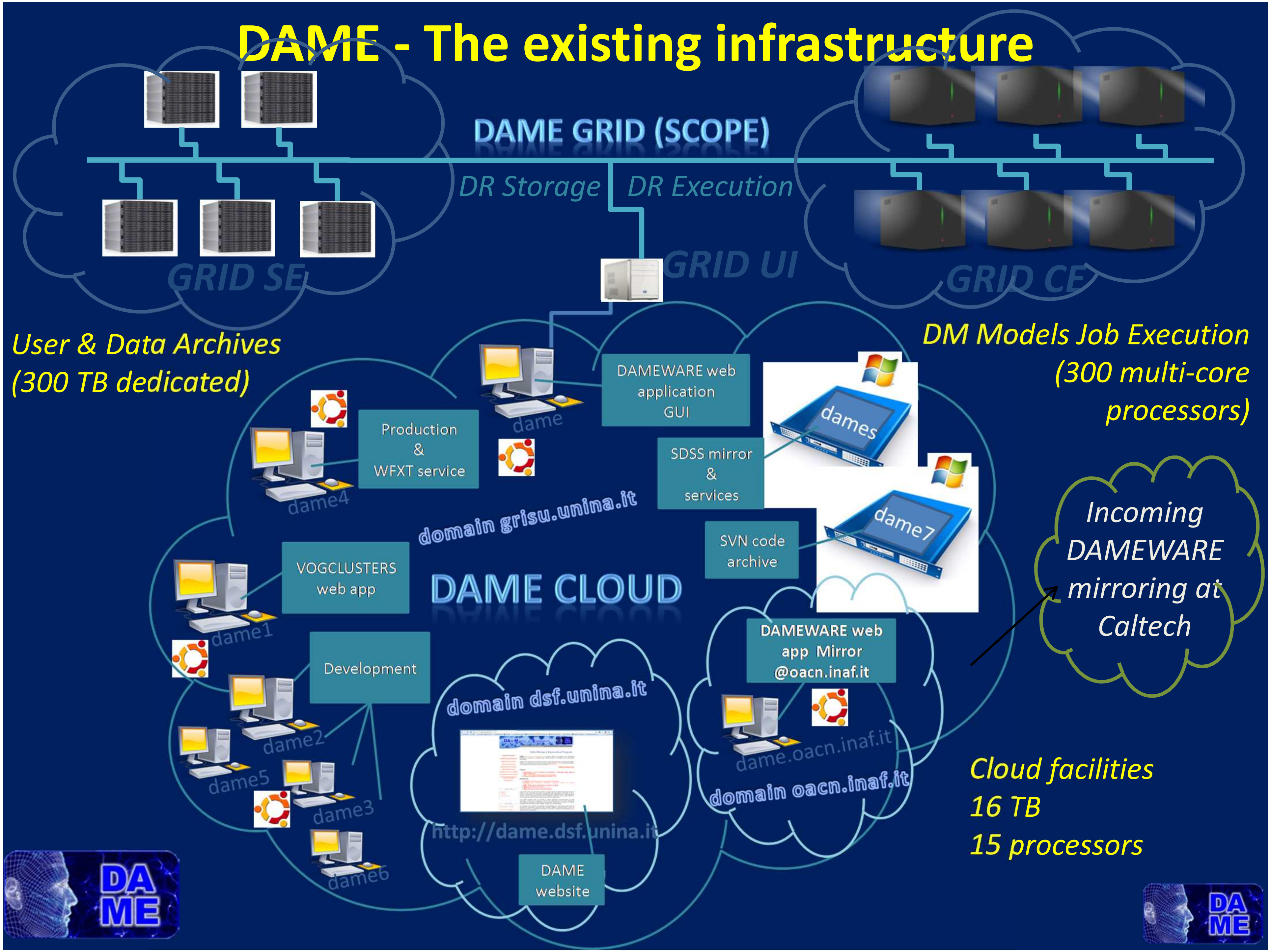
User can automatically plug-in his own algorithm and launch experiments through the Suite via a simple web browser










DAME - The existing infrastructure
DAME GRID (SCOPE)
DR Storage
DR Execution
GRID SE
GRID UI
GRID CE
User & Data Archives
(300 TB dedicated)
DM Models Job Execution
(300 multi-core
processors)
DAMEWARE web
application
GUI
dame
dames
Production
&
WFXT service
dame4
SDSS mirror
&
services
dame7
SVN code
archive
domain grisu.unina.it
VOGCLUSTERS
web app
DAME CLOUD
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
Incoming
DAMEWARE
mirroring at
Caltech
Cloud facilities
16 TB
15 processors
DA
ME

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

DA1 ↔ DA2 — Data+apps Exchange

Full interoperability between DA (Desktop Applications)

Local user desktop fully involved (requires computing power)

NO MDS

DA ↔ WA — Data+apps Exchange

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)

NO MDS

WA ↔ WA — Data+apps Exchange

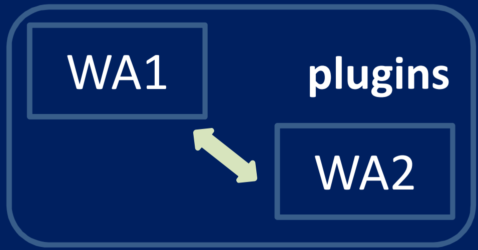
Except from URI exchange, no interoperability and different accounting policy

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

No local computing power required

NO MDS

The new vision for KDD



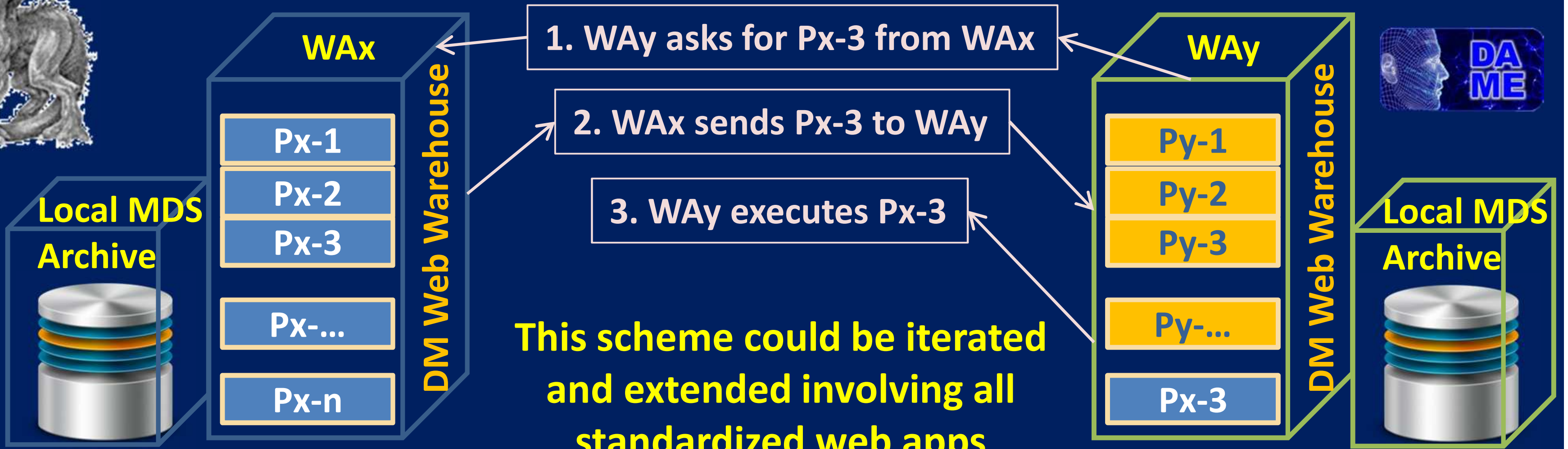
All DAs must become WAs
Unique accounting policy (google/Microsoft like)
To overcome MDS flow, apps must be plug & play
(*e.g. any WAx feature should be pluggable in WAy on demand*)

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);

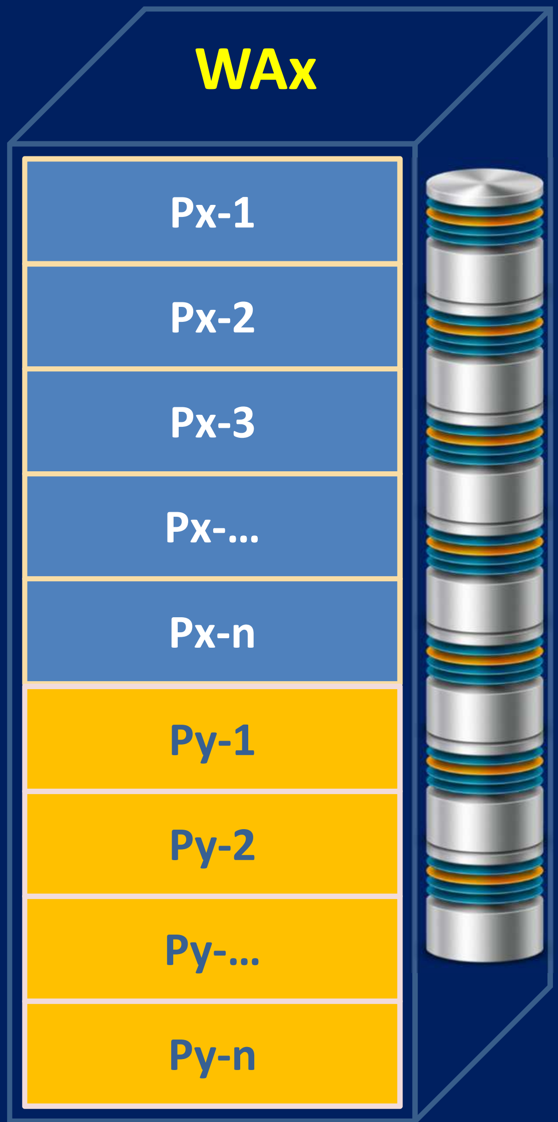
The Lernaean Hydra DAME KDD (*plugin granularity*)



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:

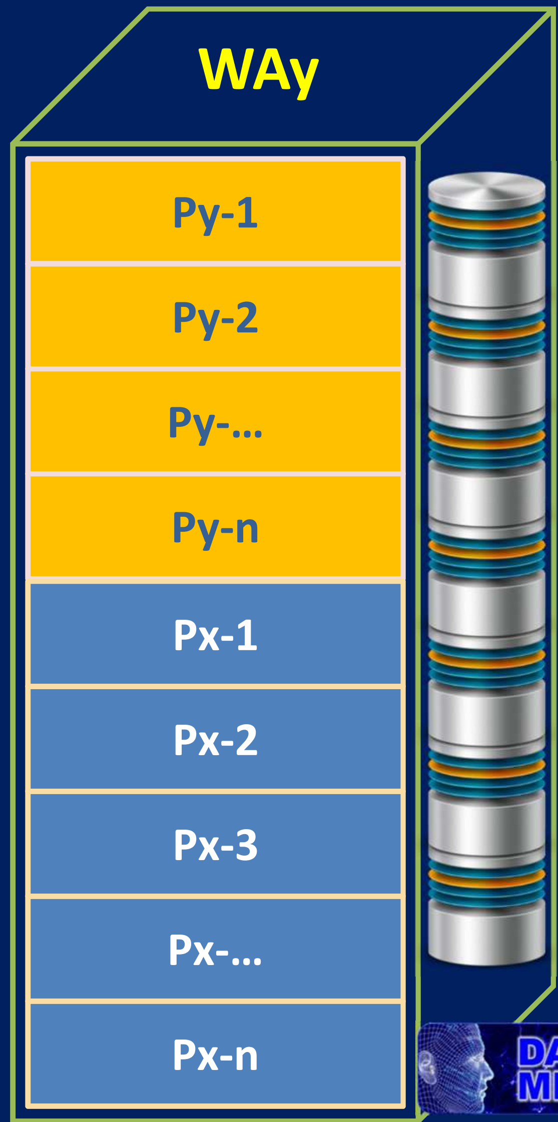
No different WAs, but simply one WA with several sites (eventually with different GUIs and computing environments)

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)

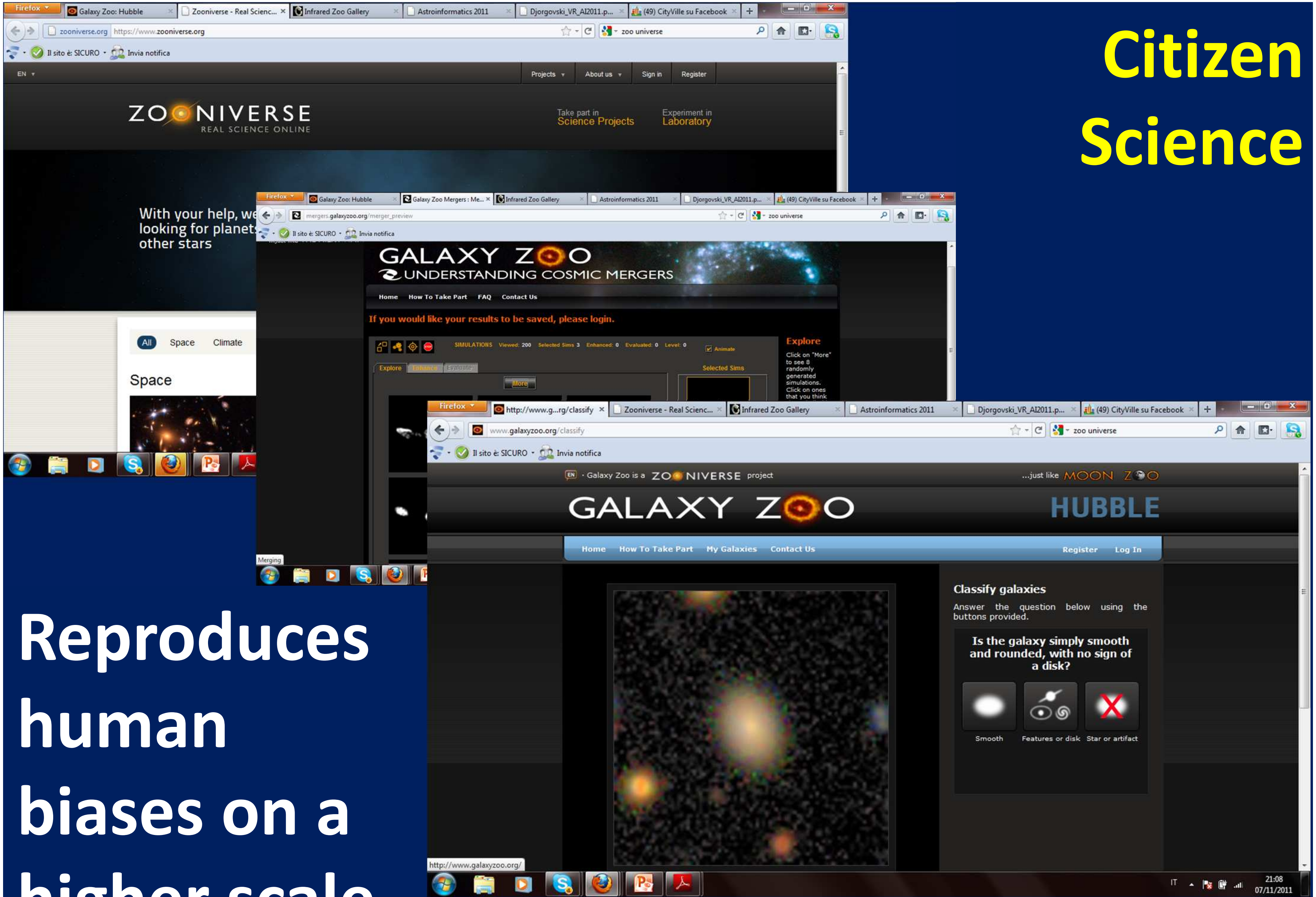
YES MDS!



Third bottle neck: lack of reliable a priori information

Longo et al. 2003	Ball & Brunner 2009	BoK
S/G separation In its various implementations	S/G separation In its various implementations	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	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	----	

Citizen Science



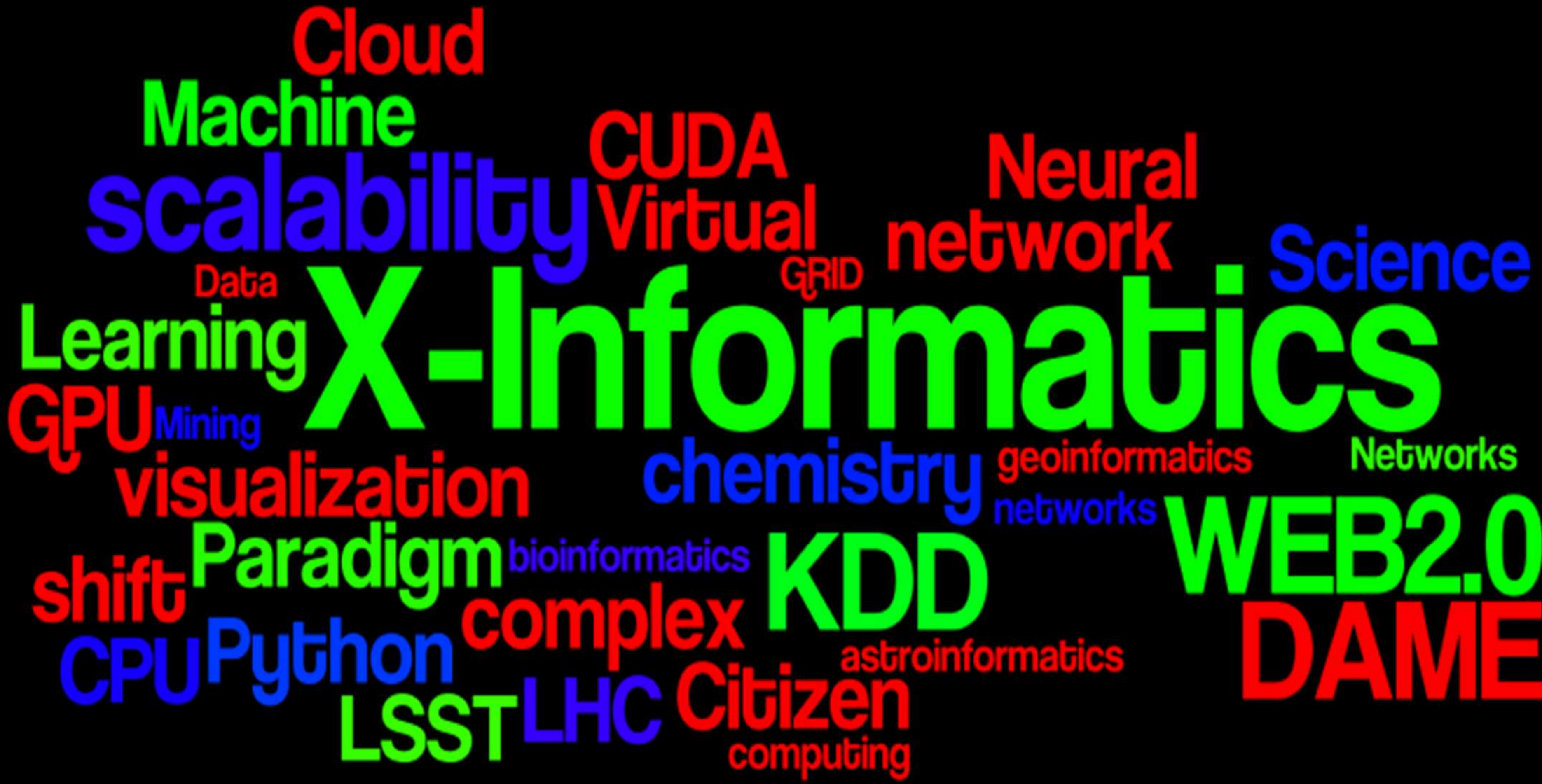
Reproduces human biases on a higher scale

Last and most serious problem: Need for a new generation of scientists
(NSF panel for interdisciplinary computing)

- Domain experts (scientists) do not want and must not become computer scientists

- The exploitation of MDS requires a much deeper understanding of computing infrastructures and of ITC technologies than what is currently done

    - **Large , crossdisciplinary teams?**
    - **New university curricula?**
    - **More user friendly SW and HW infrastructures?**


Cloud
Machine
CUDA
Neural
scalability
Virtual
network
Science
Data
GRID
Learning
X-Informatics
GPU
Mining
chemistry
geoinformatics
Networks
visualization
networks
WEB2.0
Paradigm
bioinformatics
KDD
shift
complex
DAME
CPU
Python
astroinformatics
LSST
LHC
Citizen
computing

…THE END