Virtual Observatory, Data Intensive Science & 2.0 technologies

Giuseppe Longo

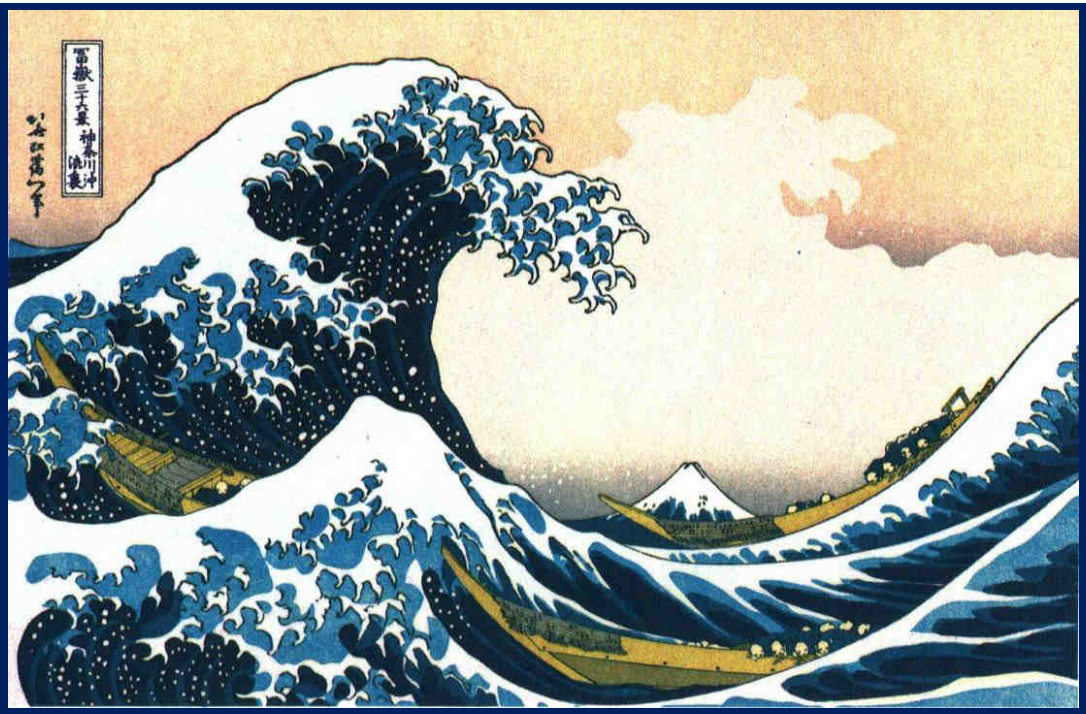
S. Cavuoti & many others
University Federico II – Napoli (ITALY)

M. Brescia
INAF – Capodimonte Observatory in Napoli (ITALY)

G.S. Djorgovski, A. Mahabal, C. Donalek
California Institute of Technology (USA)

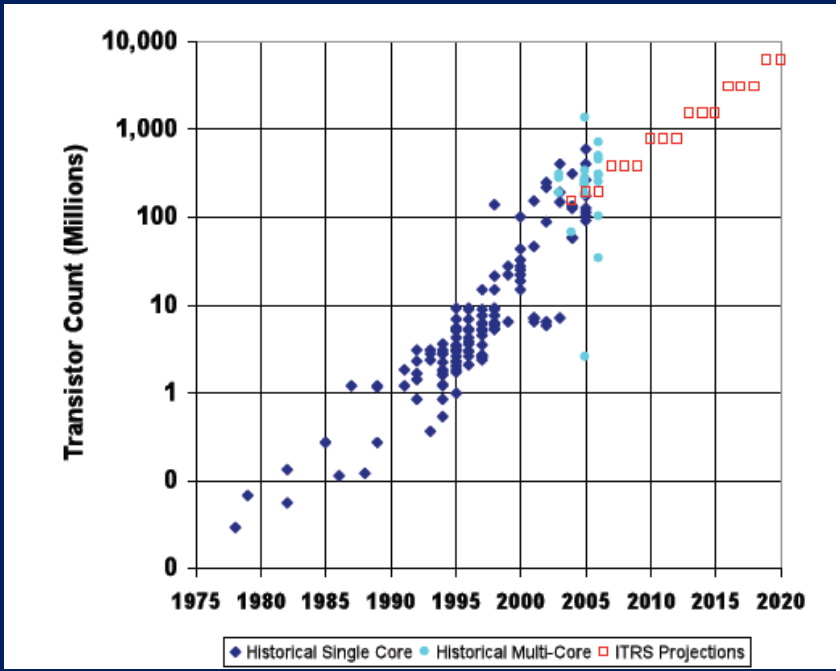
Belissima Conference, Belgrad, September 2012

Summary



- The data tsunami
- Data Intensive Science: a new scientific paradigm
- Virtual Observatory
- Bottleneck - moving programs not data
- Knowledge discovery in databases or data mining
- Astroinformatics
- Conclusions

Small, big, in a network, isolated … telescope produce large amounts of data and EACH DATA which is produced needs to be reduced, analysed, interpreted



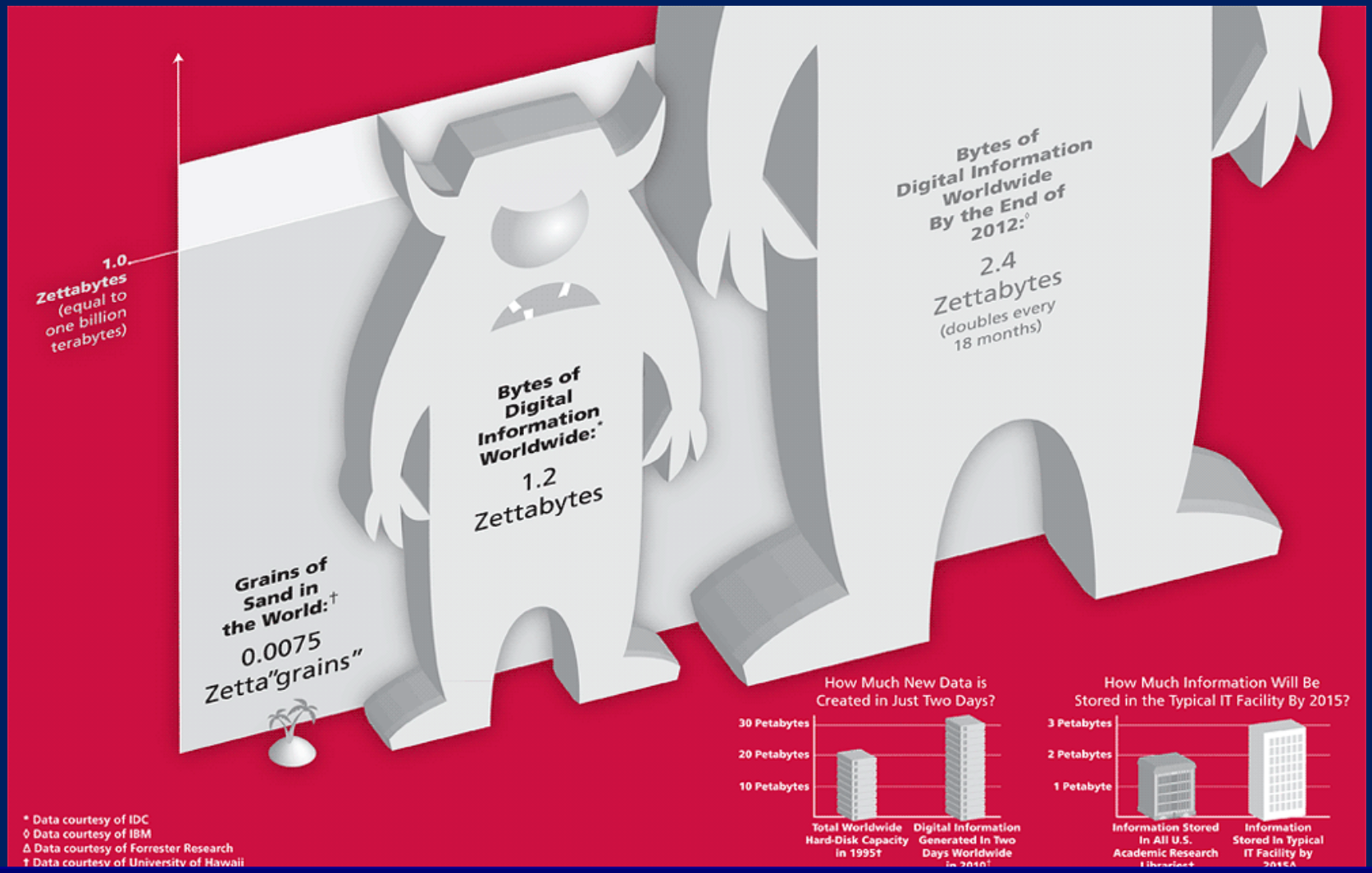
Increase in number of telescopes, improvements or detector size or in efficiency or in number of bands … all cause an increase in pixels (worldwide)

Computing time and costs do not scale linearly with number of pixels

Moore law’s does not apply anymore. Slopes are changed.

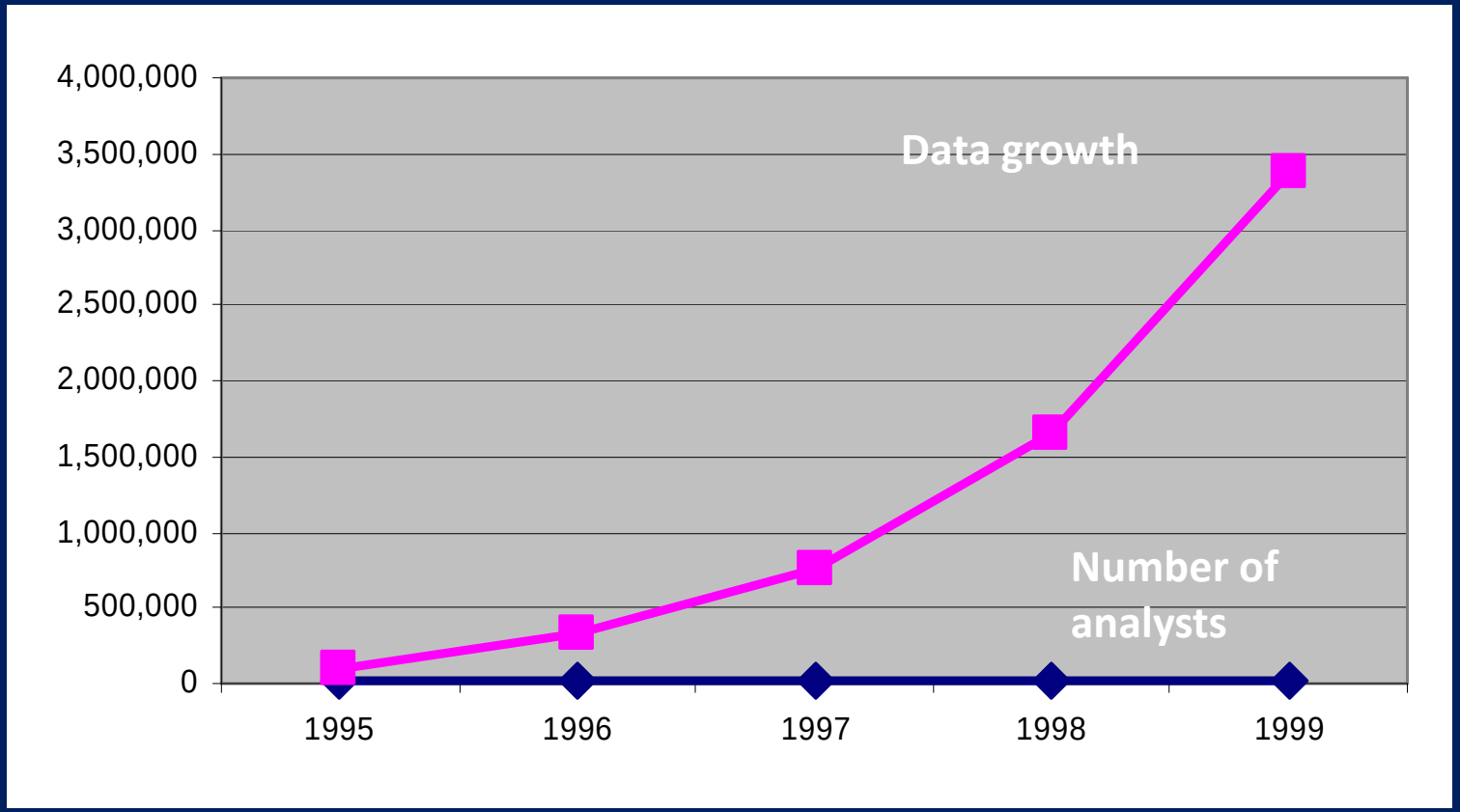
Growth of digital data worldwide

1 ZB or 1.000.000.000.000 GB = 10^9 Terabyte



SIZE, DISTRIBUTED, COMPLEX, HETEROGENEOUS

R. Grossman, C. Kamath, V. Kumar, «Data Mining for Scientific and engineering applications»

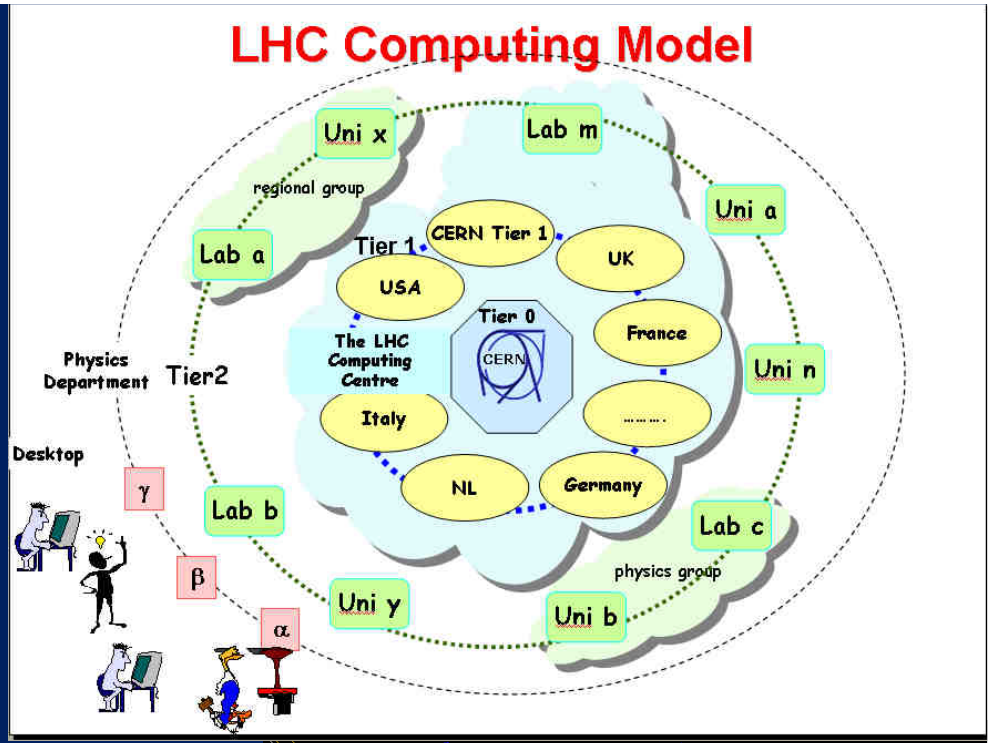


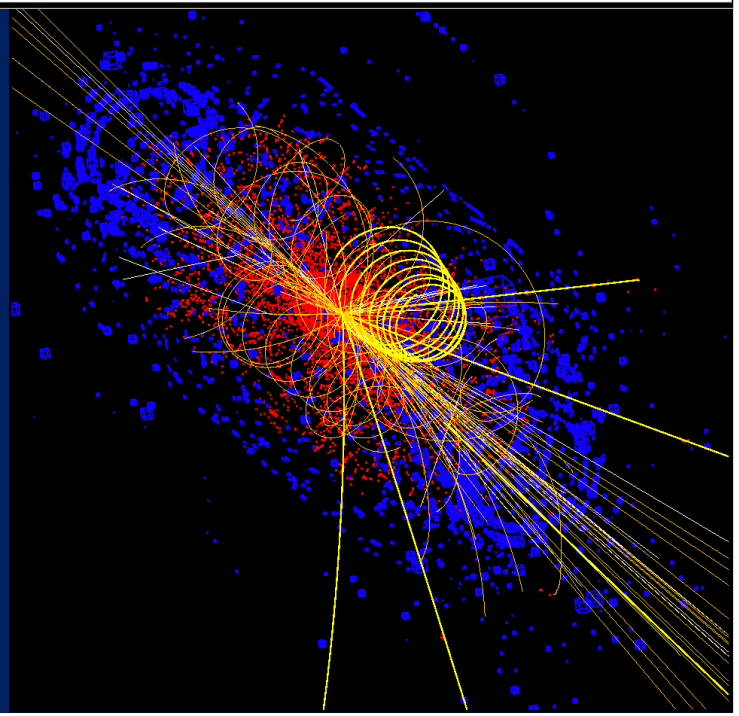
Most data will never be seen by humans…

The forerunner: LHC

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

each CPU gets one event at a time and needs to perform simple tasks





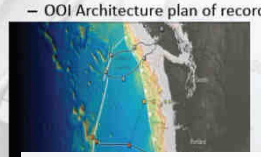
Data Stream: 330 TB/week

ATLAS detector event

Data source	Data acq rate	Data Proc	epochs	parameters
USA meteo	13 TB/day		Cont.	Ca. 50
You Tube	2.2 TB/day			
Google		100 PB/day (multicore)		Text
VST	0.15 TB/day		tens	>100
CRTS				
HST	120 TB (total)		few	>100
PANSTARRS	600 TB		Few-many	>>100
LSST	30 TB/day		hundreds	>>100
GAIA	1 PB (total)		many	>>100 heterogeneous
SKA	1.5 PB/day	exaflops	>> 10^2	hundreds

Supporting Smart Sensors and Data Fusion

- The NSF Ocean Observatory Initiative
  - Hundreds of cabled sensors and robots exploring the sea floor
  - Data to be collected, curated, mined
  - OOI Architecture plan of record, store this data in the cloud



Data collected from:
- Ocean floor sensors, AUV tracks, ship-side cruises, computational models

Data moves from **ocean** to shore side **data center** to the **Azure cloud** to your **computer**.

The Swiss Experiment (EPFL, Marc Parlange)

- Climate change affects on the regional hydrologic cycle will have profound implications for the Alps and therefore Europe
- Need for field measurements remains crucial to test sim
  models
  - The
    pote
  - Larg
    with
- Partner
  deploye
  - 100
  - 'tou
    perc
  - b

"Our abili
regional s
societal n

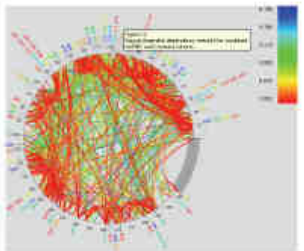
ChronoZoom – History in its broadest possible context …

The challenge: exploration of all known time series, and smoothly transition from billions of years down to individual nanoseconds…

This is what Walter Alvarez, Professor of Earth and Planetary Science at University of Berkeley set out to do. And he did it, with the help of Microsoft Research and the Live Labs team.

Fighting HIV with ML and HPC

- PhyloD.Net is a Bayes-net-based tool that deciphers evolution of HIV within a patient
- Developed by eScience research group and published in *Science*, March 2007
- Now used by dozens of HIV research groups
- Led to discovery of two key insights to fight HIV:
  - Our immune system attacks frameshift epitopes, which may be useful to include in a vaccine (*JEM*, 2010)
  - Natural killer cells directly attack HIV (*Nature Medicine*, in review)
- Typical job
  - 10 – 20 CPU hours with extreme jobs requiring 1K – 2K CPU hours
  - Requires a large number of test runs for a given job (1 – 10M tests)



PhyloD.Net on cover of *PLoS Comp Bio*, Nov 2008
Carlson, Kadie, & Heckerman et al.

DATA INTENSIVE SCIENCE HAS BECOME A REALITY IN ALMOST ALL FIELDS and poses worse problems

- Huge data sets (ca. Pbyte)

In astronomy as in many other sciences

- Thousands of different problems
- Many, many thousands of users

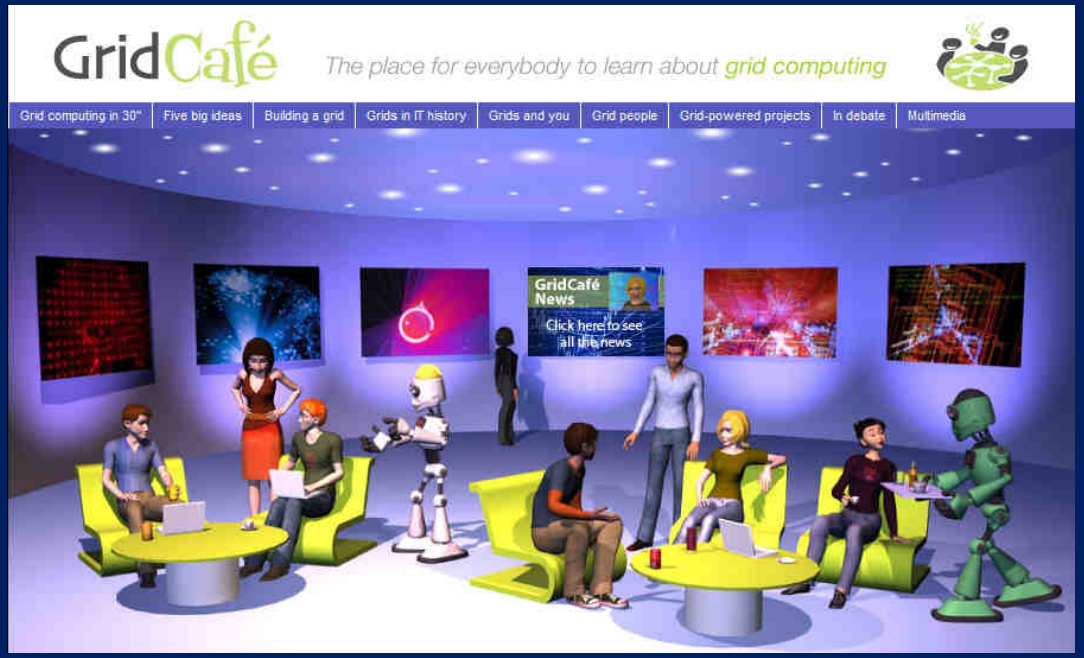
i.e. LHC is a “piece of cake”

(simple computational model)

GRID – LHC evolved into EGEE

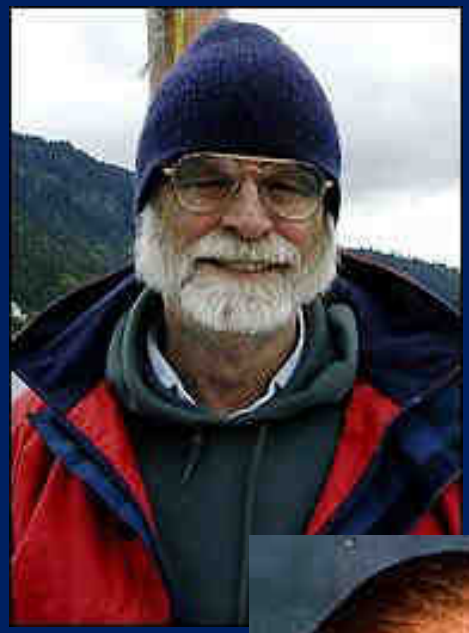
LHC was a forerunner also in many other technologies …

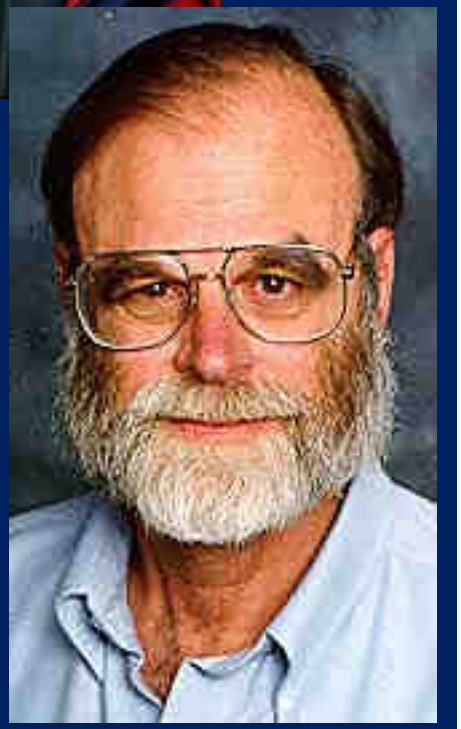
Remember this ….



GRID Café - CERN

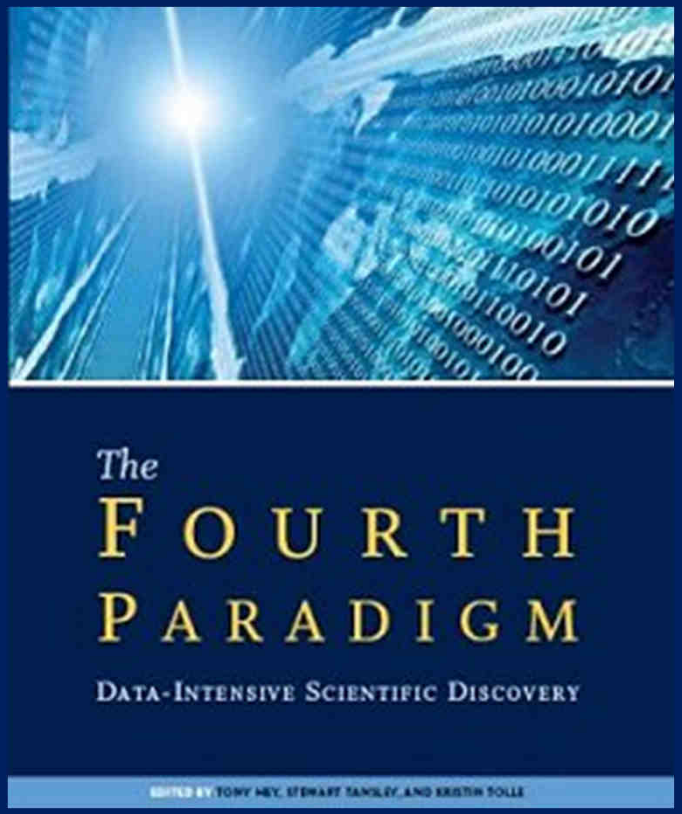
Jim Gray





"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**

(and it is happening **now!!**)

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

Data Federation – Virtual Observatory

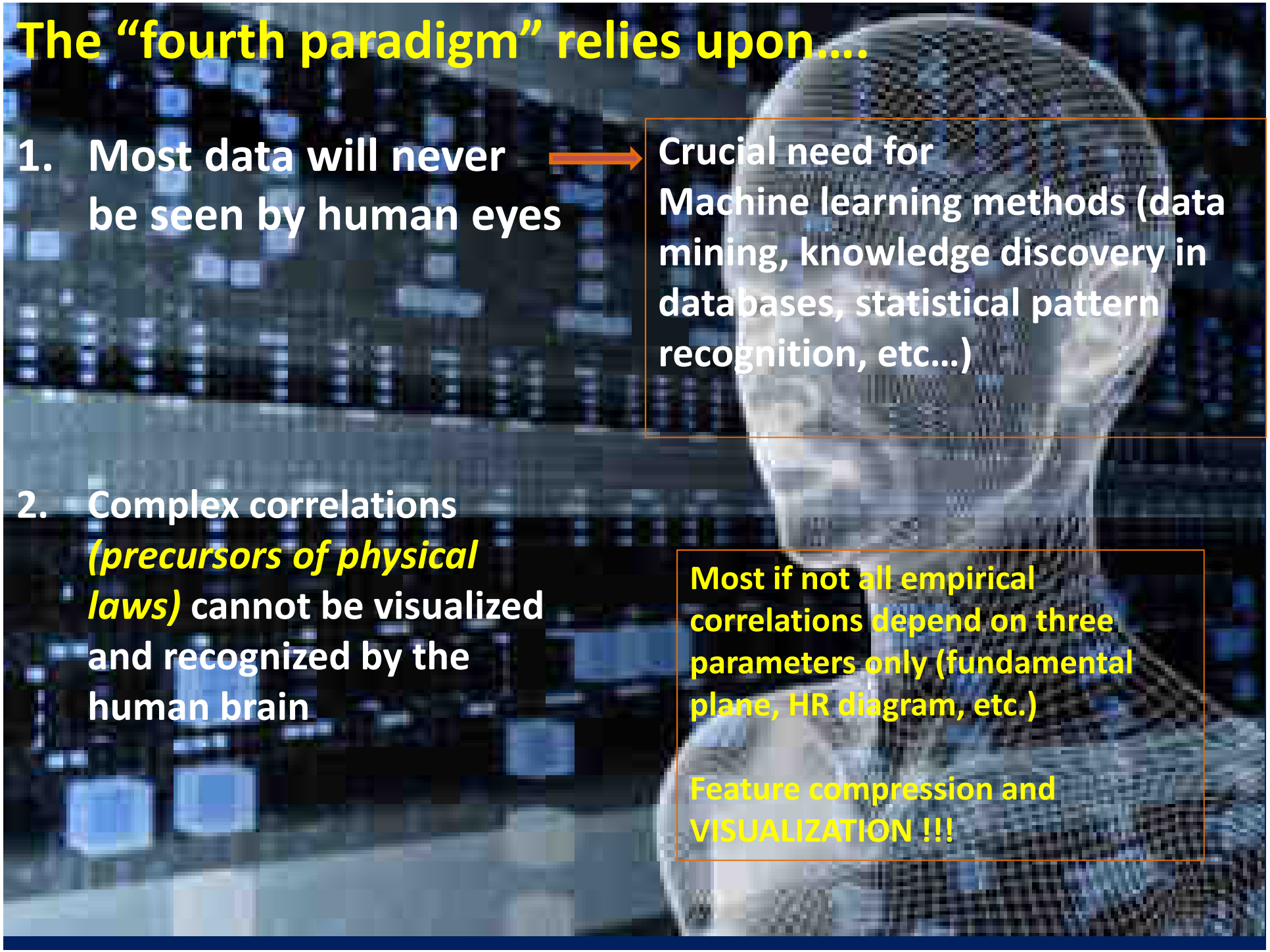
The International Virtual Observatory Alliance (>20 countries)

Data federation (standards) and interoperability has been completed

Every one can publish his data in the VO with simple tools

Data analysis and understanding (KDD) is in the making but requires a change in perspective

IVOA – IG on KDD (Longo)
Astroinformatics exploratory Initiative (Djorgovski)
Astroinformatics WG at the IAU 2012 (Longo) and at the AAS 2012 (Djorgovski)



The "fourth paradigm" relies upon….

1. Most data will never be seen by human eyes

→ Crucial need for Machine learning methods (data mining, knowledge discovery in databases, statistical pattern recognition, etc…)

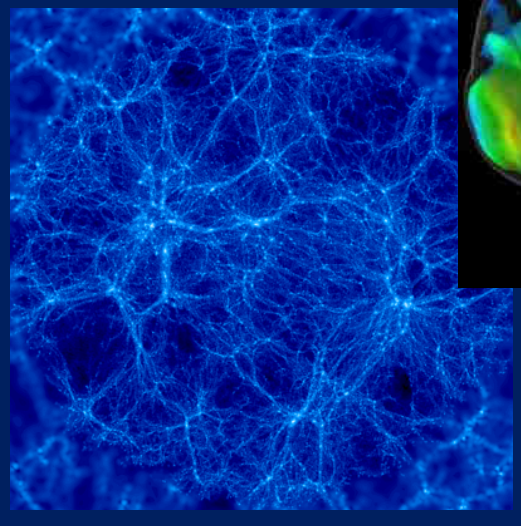
2. Complex correlations *(precursors of physical laws)* cannot be visualized and recognized by the human brain

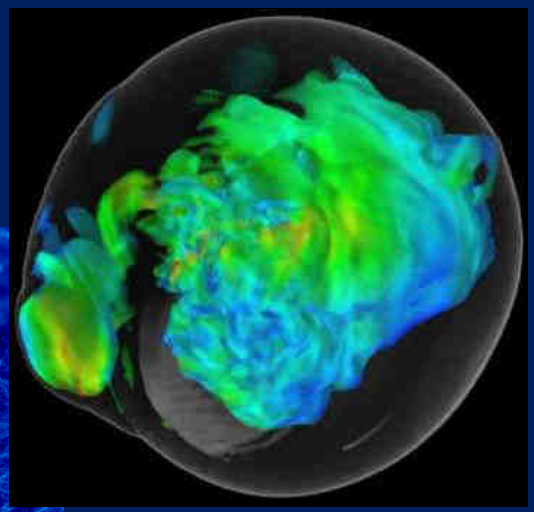
Most if not all empirical correlations depend on three parameters only (fundamental plane, HR diagram, etc.)

Feature compression and VISUALIZATION !!!

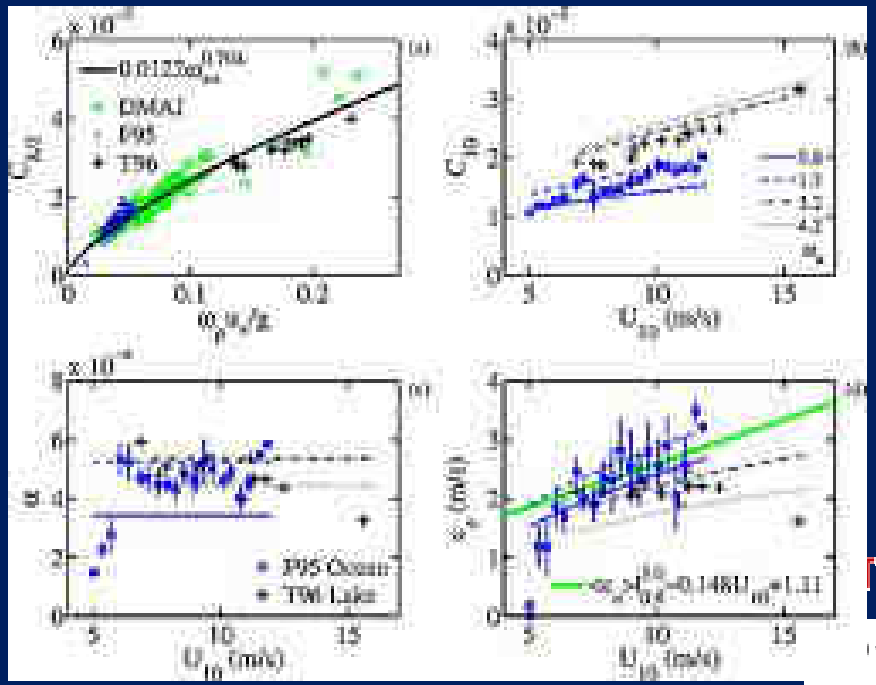
3. **Real world physics is too complex. Validation of models requires *accurate simulations, tools to compare simulations and data,* and better ways to deal with complex & massive data sets**

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

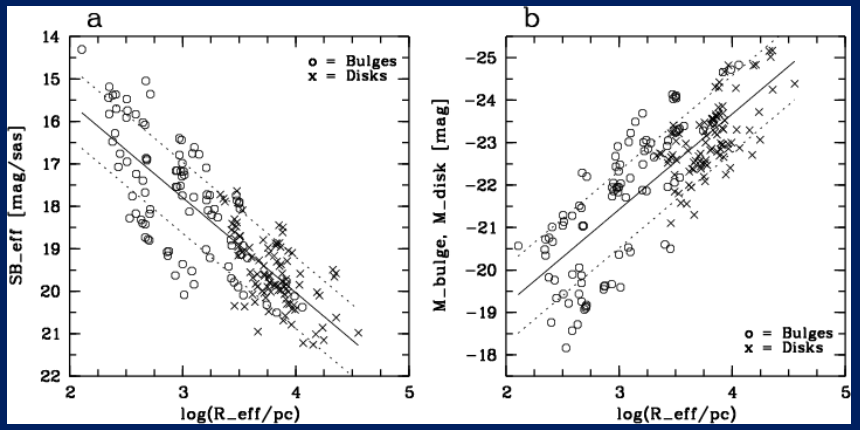




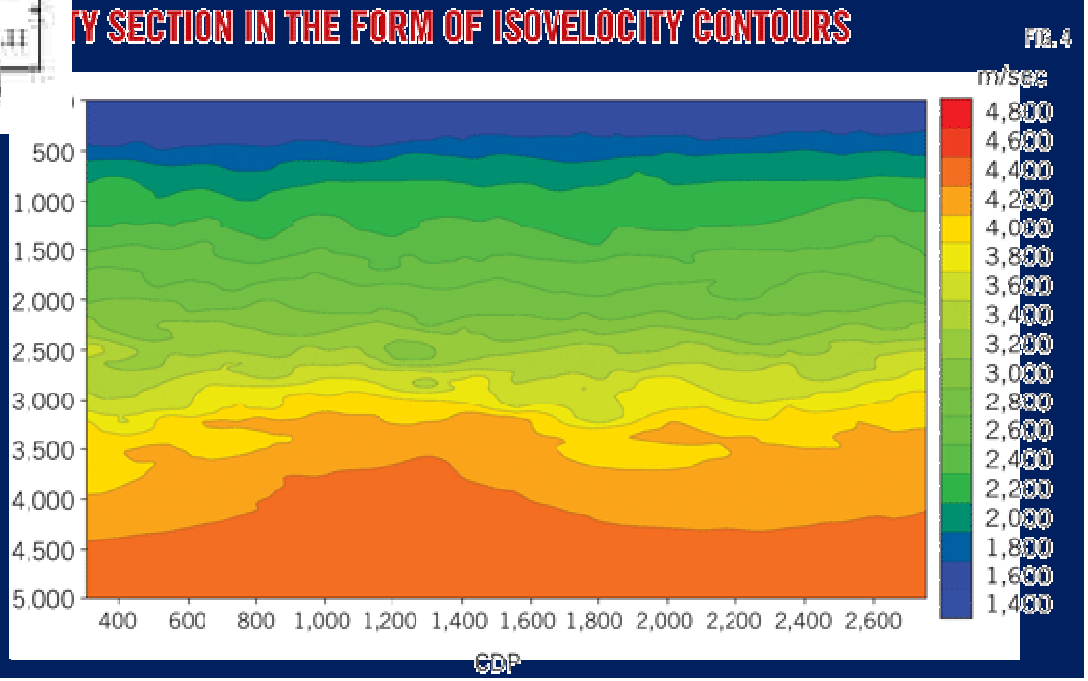
First hint about the need for complex visualization



Oceanography

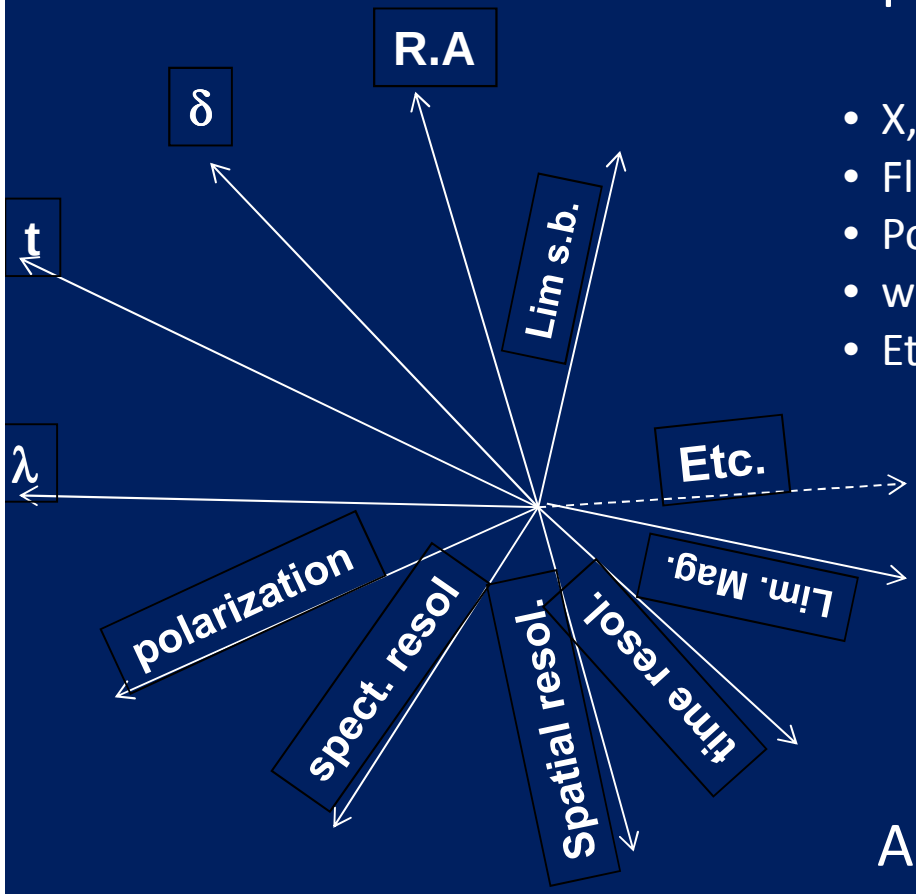


Astronomy



Petrology

The measurable parameter space of KDD



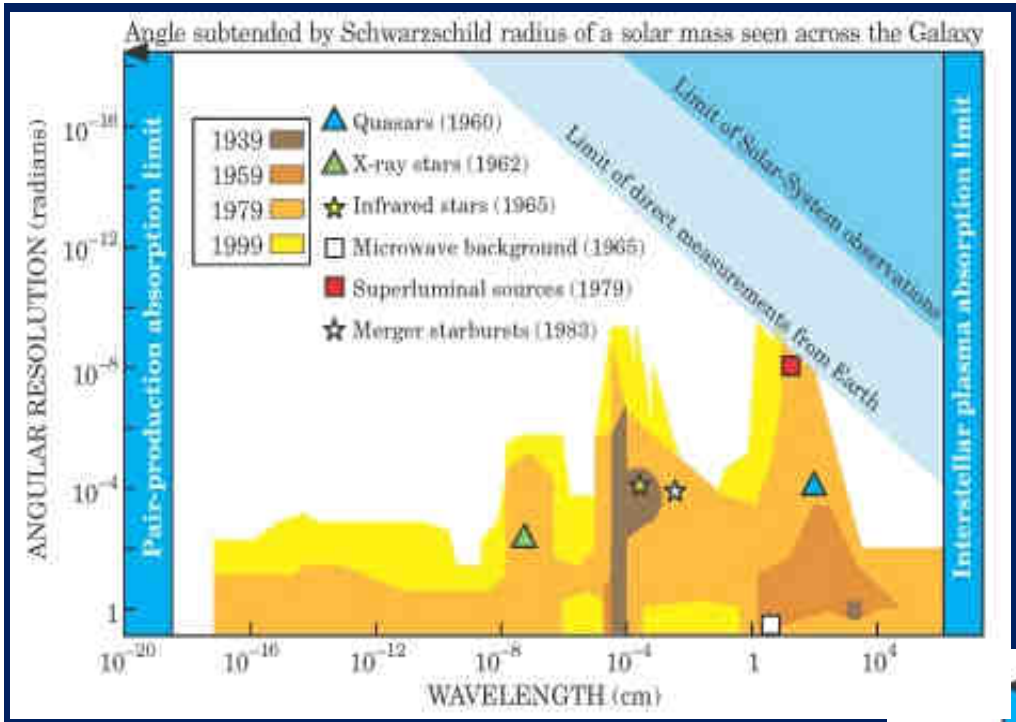
Each datum is defined by n measured parameters.

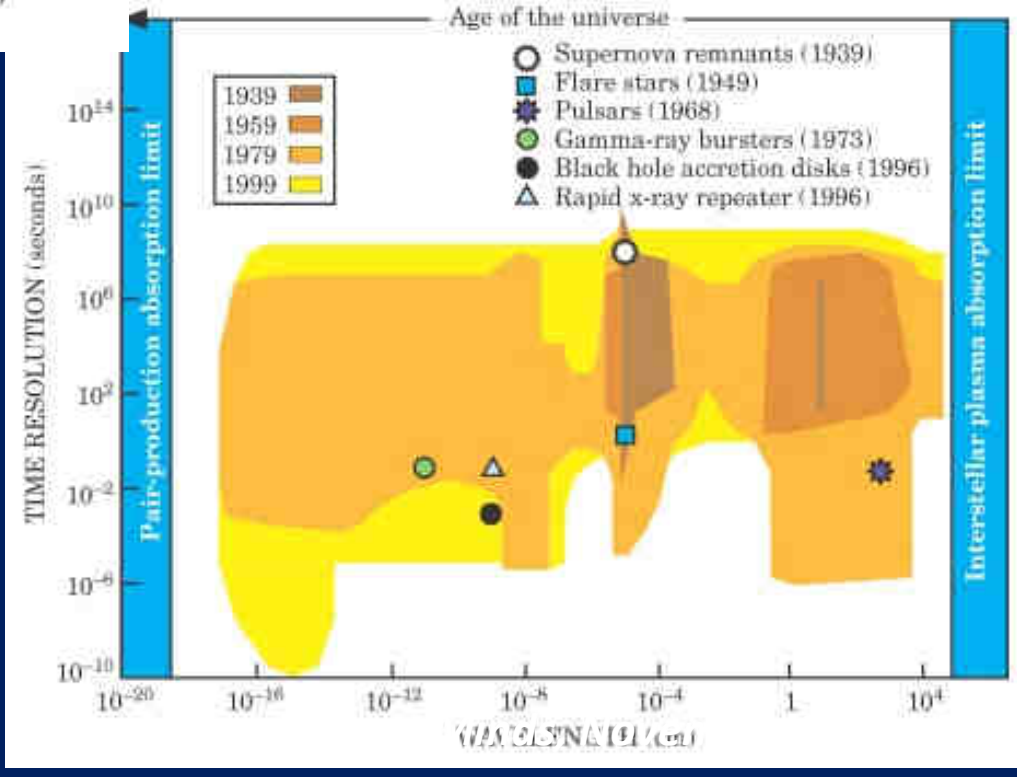
- X,y,t
- Flux
- Polarization
- wavelength
- Etc..

New sensor technologies:

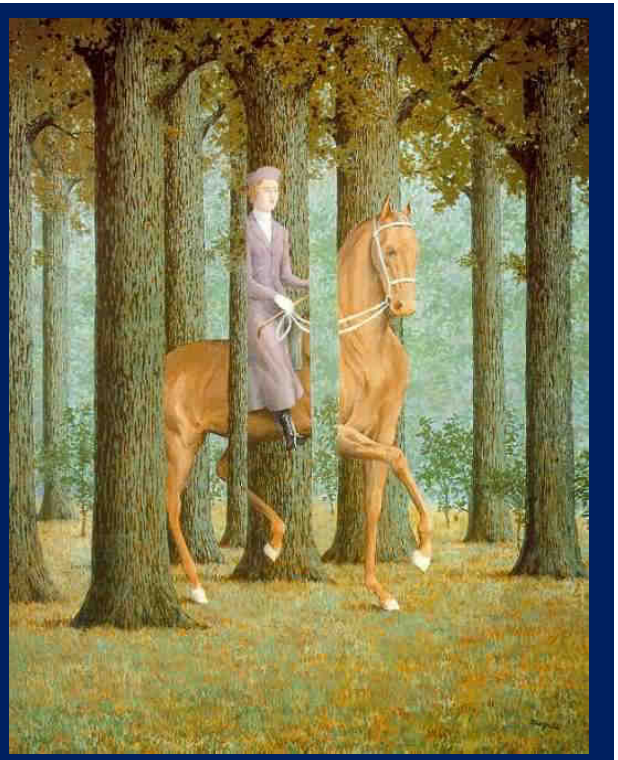
$$p \in \Re^N \qquad N >> 100$$

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


Angle subtended by Schwarzschild radius of a solar mass seen across the Galaxy
ANGULAR RESOLUTION (radians)
WAVELENGTH (cm)
Pair-production absorption limit
Interstellar plasma absorption limit
Limit of Solar-System observations
Limit of direct measurements from Earth
1939
1959
1979
1999
Quasars (1960)
X-ray stars (1962)
Infrared stars (1965)
Microwave background (1965)
Superluminal sources (1979)
Merger starbursts (1983)


Age of the universe
TIME RESOLUTION (seconds)
WAVELENGTH (cm)
Pair-production absorption limit
Interstellar plasma absorption limit
1939
1959
1979
1999
Supernova remnants (1939)
Flare stars (1949)
Pulsars (1968)
Gamma-ray bursters (1973)
Black hole accretion disks (1996)
Rapid x-ray repeater (1996)

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 of distributed computing power

AstroNeural – 1992-2001

VO- Neural 2002 – 2006

DAME 2007 – td

ALL STARTED WITH

ELSEVIER

EPSL

Earth and Planetary Science Letters 139 (1996) 33–45

Neural net aided detection of astronomical periodicities in geologic records

M. Brescia [a,b], B. D'Argenio [c,d], V. Ferreri [d], G. Longo [b], N. Pelosi [c], S. Rampone [e,f], R. Tagliaferri [a,f,*]

[a] *Dipartimento di Informatica ed Applicazioni, Università di Salerno, Salerno, Italy*
[b] *Osservatorio Astronomico di Capodimonte, Napoli, Italy*
[c] *Istituto di Ricerca GEOMARE sud, CNR, Napoli, Italy*
[d] *Dipartimento di Scienze della Terra, Università di Napoli "Federico II", Napoli, Italy*
[e] *Dipartimento di Fisica Teorica, Università di Salerno, Salerno, Italy*
[f] *INFM, Unit of Salerno, Salerno, Italy*

Received 1 May 1995; accepted 20 December 1995

Abstract

Astronomically controlled variations in the Earth's climate induce cyclic trends in the sedimentary process and re (Milankovitch periodicity). One of the main difficulties to be solved in order to choose among the registered periodiciti the conversion from the spatial (i.e. recurrent variations along the stratal sequences) to the temporal domains of astronomically induced frequencies present in the rock record. We discuss here how this problem can be circumvente teaching a neural net how to recognize periodicities in the signal. The application to two sequences of shallow w carbonate deposits from the Cretaceous of Southern Italy has shown this approach to be particularly effective, confirming existence of Milankovitch-type periodicities in the records examined, where climate, sediments and biota concomit react to the variation in the solar constant induced by secular perturbations of the Earth's orbital elements.

Keywords: Milankovitch theory; paleoclimatology; Cretaceous; Southern Apennines

The detection of Milankovic cycles
In stratigraphic records

Spec. Publs Int. Ass. Sediment. (1994) **19**, 77–85

Fourier evidence for high-frequency astronomical cycles recorded in Early Cretaceous carbonate platform strata, Monte Maggiore, southern Apennines, Italy

G. LONGO*, B. D'ARGENIO†, V. FERRERI† *and* M. IORIO‡

* *Osservatorio Astronomico, Napoli, Italy;*
† *Dipartimento di Scienze della Terra, Università Federico II, Napoli, Italy; and*
‡ *Geomare, Istituto di Geologia Marina del CNR, Napoli, Italy*

ABSTRACT

Carbonate peritidal deposits of Early Cretaceous age, widely outcropping in the carbonate platform sequence of southern Italy, carry distinct signals of cyclicity in the Milankovitch band. We have studied the depositional and diagenetic facies organization of a *c.*100-m-thick sequence of Barremian age, at Monte Raggeto (Monte Maggiore Mountains, near Naples), where, from a total of 60 m analysed at centimetre scale, two sedimentary modules have been recognized.

1 *Depositional cyclothems*: rare, made of one or more subtidal–supratidal couplets and topped by supratidal intervals.

2 *Diagenetic cyclothems*: very common, made of dominantly subtidal intervals which show emersion-generated features (karst, reddened surfaces) at the top.

Cyclothems along the sequence tend to group into fairly regular intervals, each about 10 m thick, formed by sets of seven to nine cyclothems. This trend is confirmed by Fourier analysis of the data, showing periodicities at 105 and 950 cm. Moreover, the mathematical processing of the total data set shows also two shorter periodicities at 40 and 72 cm. The algorithm used for the analysis is a modified version of the Deeming code, first written for astrophysical applications.

The set of periodicities obtained (40, 72, 105 and 950 cm) can be related to the variations of the insolation constant computed for the Cretaceous; moreover the ratios between the two sets of periodicities, expressed in centimetres and in years respectively, show a very high degree of correlation with those predicted for the main orbital periods in the Cretaceous. We propose that the observed cyclicities indicate Barremian sea-level oscillations induced by high-frequency eustatic control under climatic forcing.

Moreover the link between the ratios of time and depositional periodicity sets appears to be a useful method of assigning a duration to a given periodicity in a sequence, quite independently from a precise determination of biostratigraphic age and/or total thickness of a stage.

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

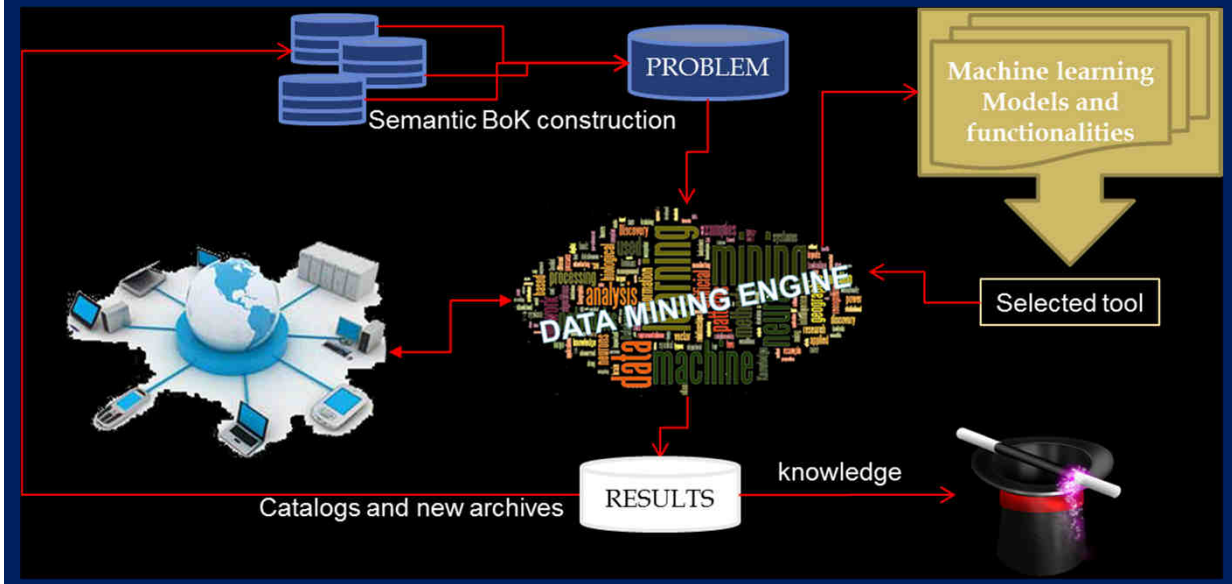
http://www.youtube.com/user/DAMEmedia
DAMEWARE Web Application media channel

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

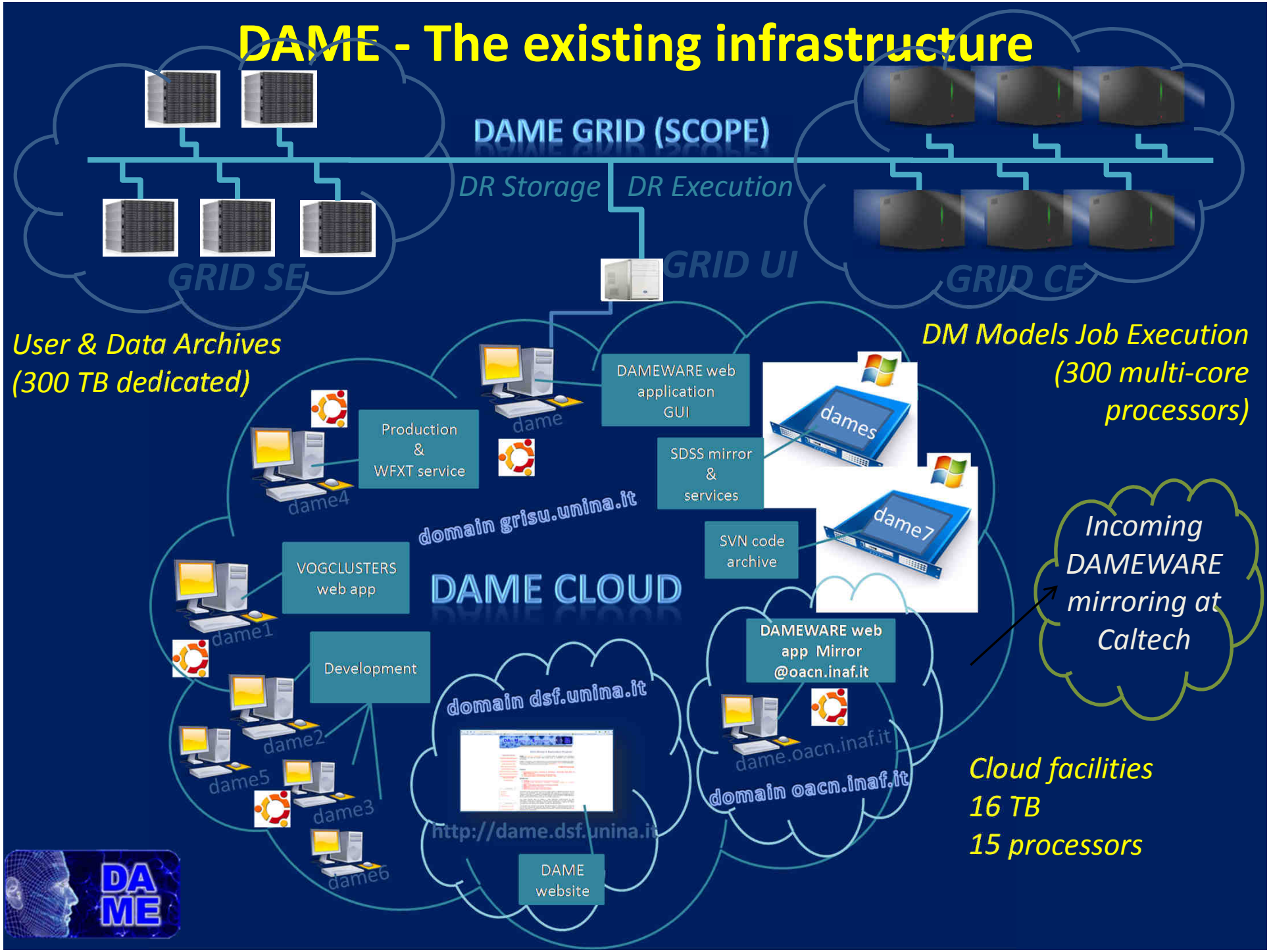
next …

Bayesian Networks

Random Decision Forest

MLP with Levenberg-Marquardt

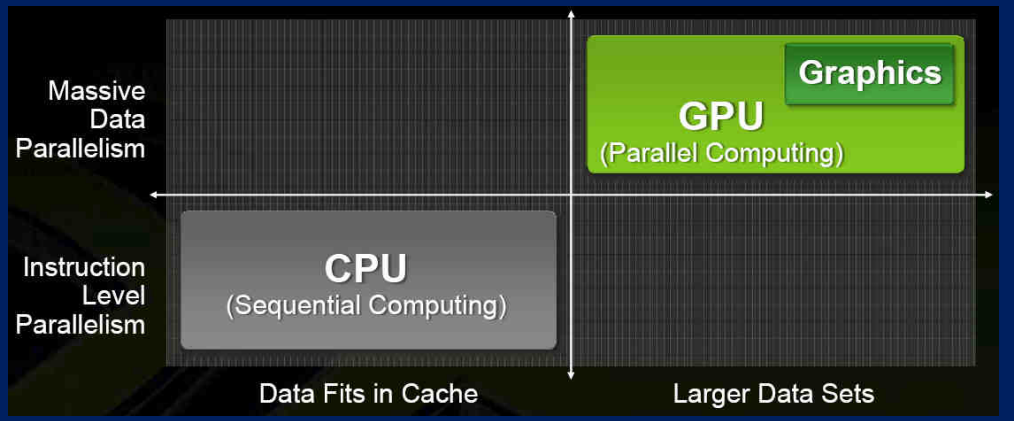
DAME - The existing infrastructure

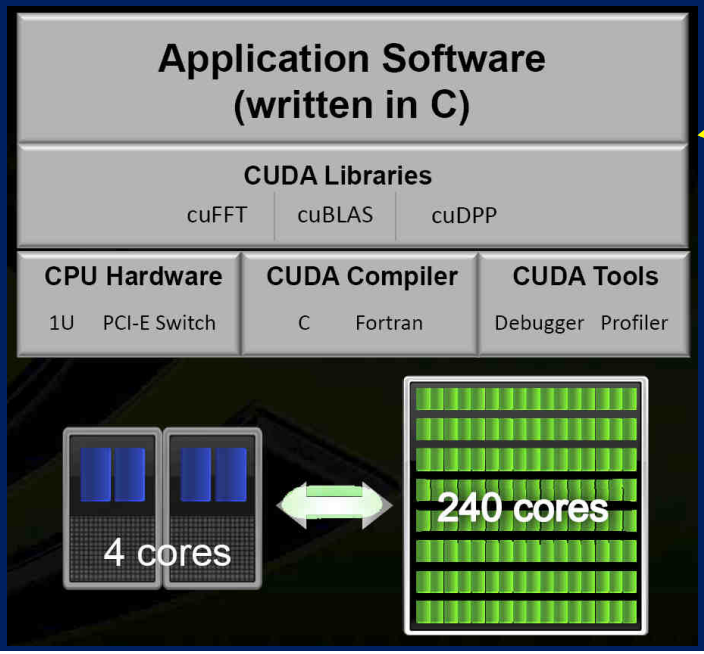


… GPU technology?

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

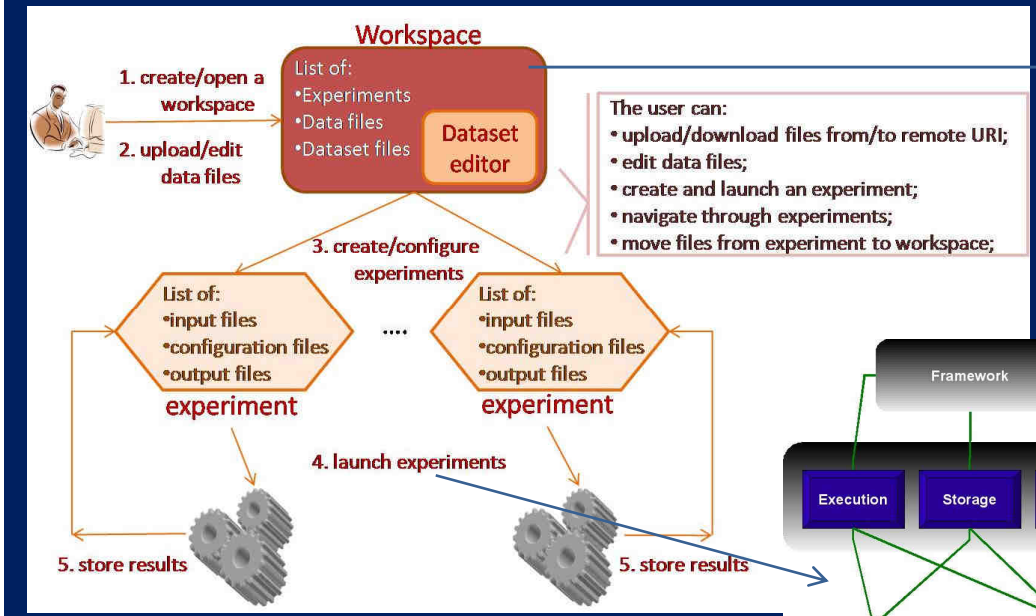
GAME is a pure genetic algorithm developed in order to solve supervised problems of regression or classification, able to work on Massive Data Sets (MDS).

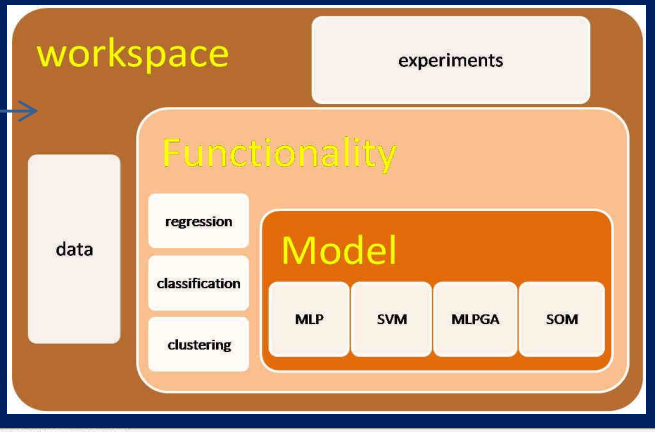
DAMEWARE fundamentals

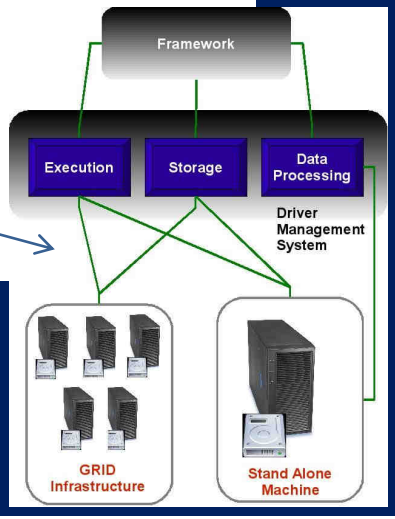
It is multi-disciplinary platform (astronomy, bioinformatics and medical diagnostics)

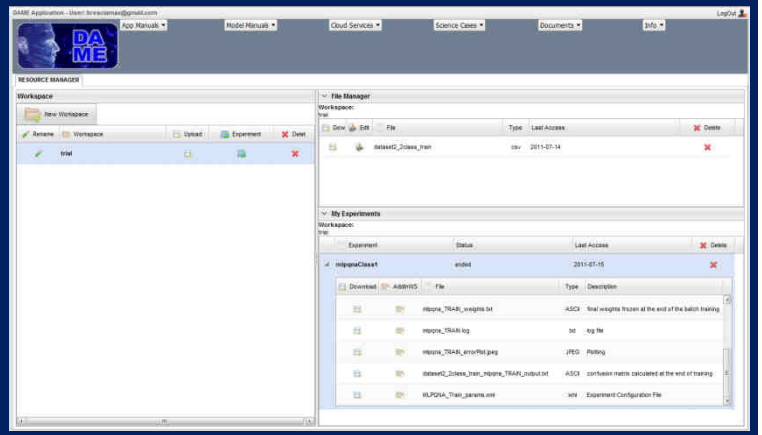
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

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









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.

Mon. Not. R. Astron. Soc. **000**, 1–11 (2011) Printed 6 September 2011 (MN LATEX style file v2.2)

The detection of globular clusters in galaxies as a data mining problem

Massimo Brescia[1], Stefano Cavuoti[2], Maurizio Paolillo[2], Giuseppe Longo[2,3], Thomas Puzia[4]

1 - INAF-Astronomical Observatory of Naples, via Moiariello 16, I-80131 Napoli, Italy
2 - Dipartimento di Scienze Fisiche, University Federico II, via Cinthia 6, I-80126 Napoli, Italy
3 - Visiting Associate, California Institute of Technology, Pasadena USA
4 - Catholica Universidad del Chile

MLP-QNA as classifier

NGC3199 GC system

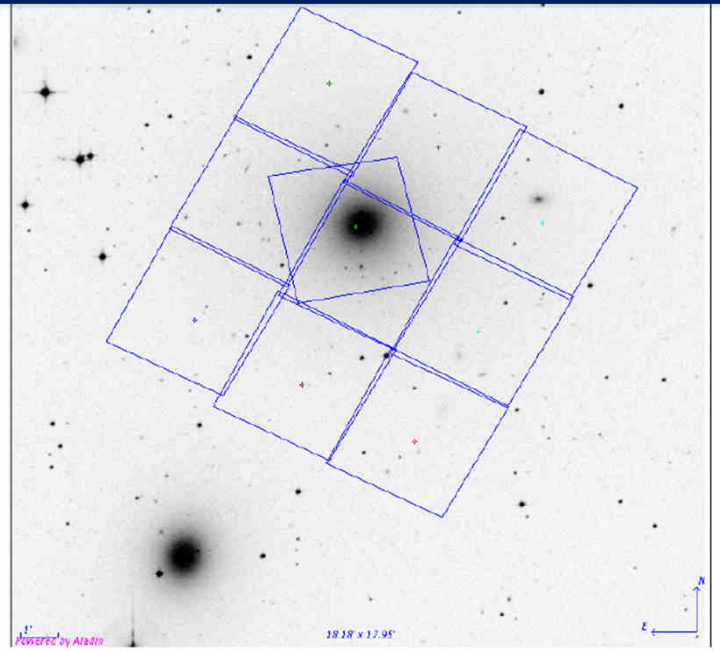


Figure 1. The field of view covered by the 3x3 HST/ACS mosaic in the F606W band. The central field, with a different orientation, shows the region covered by previous archival ASC observations in g and z bands.

11 measured parameters

(magnitude, colors, parametric measures)

Use KB based on color selection for central field to identify GC's in external field using only 1 band data

2146 GC in central field

type of experiment	missing features	MLPQNA	GAME	SVM	MLPBP	MLPGA
complete patterns		98,3	82,2	90,5	59,9	66,2
no par. 11	11	98,9	81,9	90,5	59,0	62,4
only optical	8 9 10 11	93,9	86,4	90,9	70,3	76,2
mixed	5 8 9 10 11	94,7	86,7	89,1	68,6	71,5

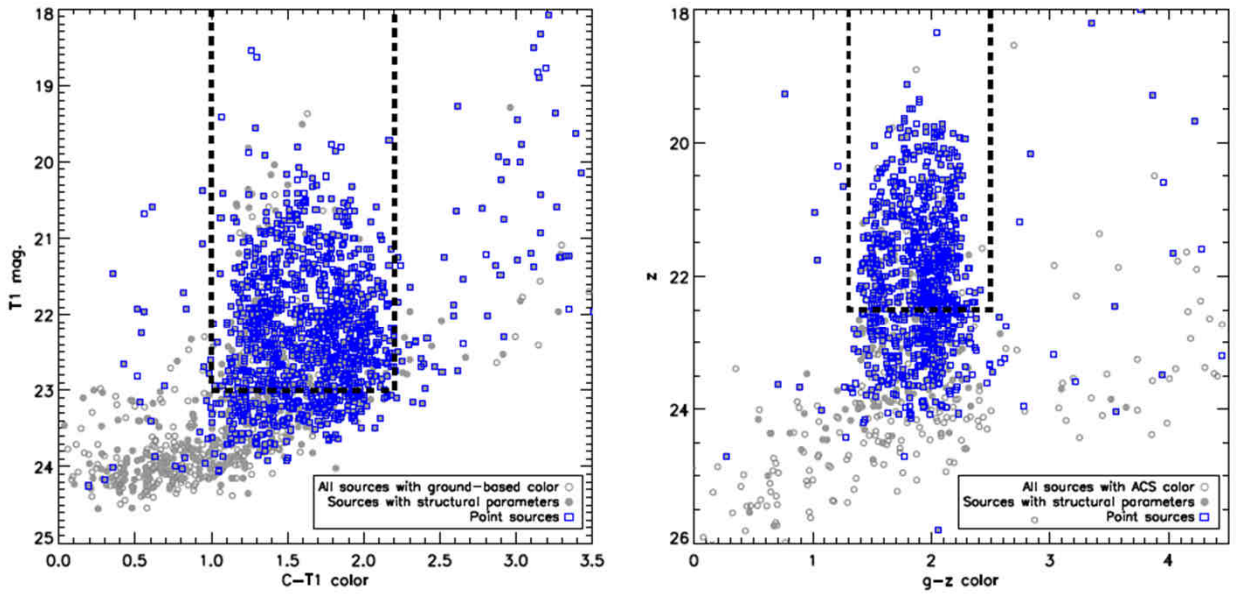


Figure 3. Color-magnitude diagrams using $C-T1$ ground-based (left panel) and $g-z$ HST photometry (right panel). Ground-based photometry covers the whole FOV of our ACS mosaic, while HST colors are limited to the central ACS field ($\sim 200'' \times 200''$, Figure 1). Open grey dots represent all sources in color catalogs while solid ones refer to the subsample with both color and structural parameters that represents our Knowledge Base. Blue squares mark pointlike sources, i.e. sources with stellarity index > 0.9, while the dashed line highlights the parameter space (Table 1) used to select bona-fide GC.

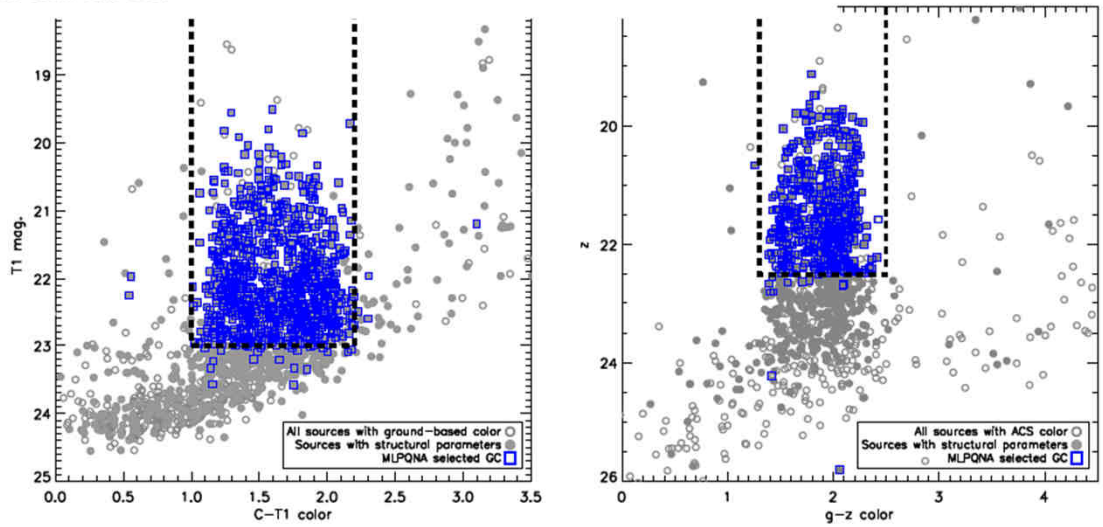


Figure 4. Same as Figure 3 showing the color distribution of the MLPQNA selected sample. The MLPQNA sample (blue squares) reproduces the properties of the color-selected GC population (i.e. the KB) with much less contaminants than, e.g., the pointlike population shown in Figure 3.

MLP-QNA as a regressor

LETTER TO THE EDITOR

Photometric redshifts with MLP-QNA. I. Results on the PHAT1 dataset

S. Cavuoti[1,2], M. Brescia[2], G. Longo[2,1,3], and A. Mercurio[2,]

[1] Dipartimento di Scienze Fisiche, University Federico II, via Cinthia 6, I-80126 Napoli, Italy e-mail: cavuoti@na.infn.it
[2] INAF-Astronomical Observatory of Naples, via Moiariello 16, I-80131 Napoli, Italy
[3] Visiting associate - Department of Astronomy, California Institute of Technology, CA 90125, USA

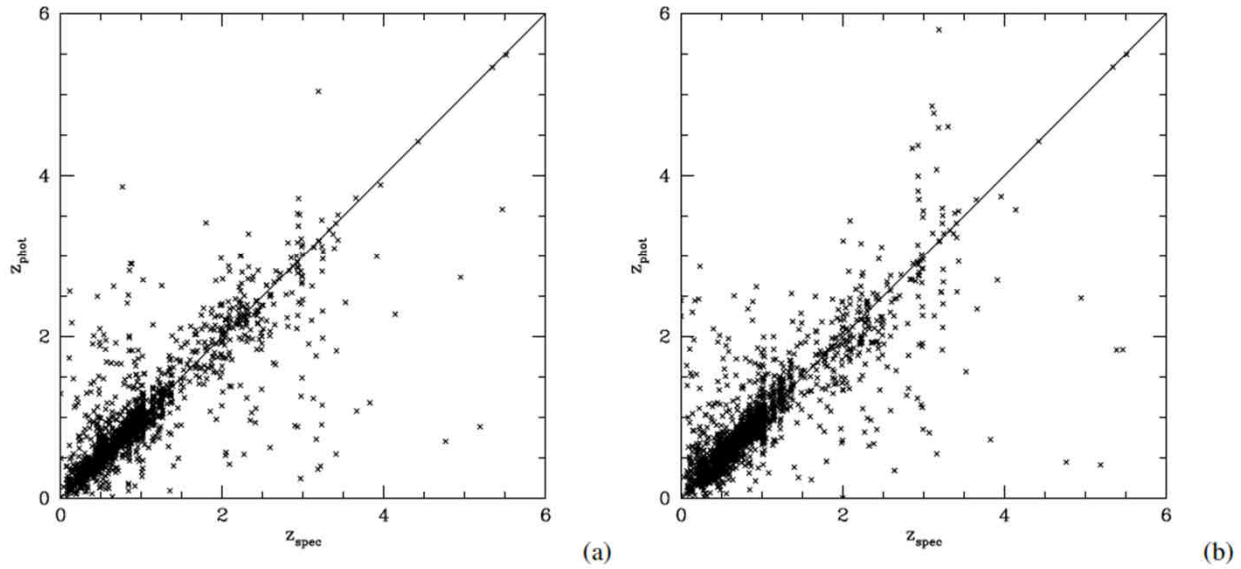


Fig. 1. Results obtained using the analysis described in this paper by the PHAT contest group. In the (a) panel are plotted the photometric vs. spectroscopic redshifts for the whole dataset using 10 photometric bands (Experiment 1). In panel (b) the same but using only 14 photometric bands (Experiment 2).

Good				quite good				acceptable				medium				poor			
B	ID	F	T	B	ID	F	T	B	ID	F	T	B	ID	F	T	B	ID	F	T
I	C5	0	0	B	C2	2	1	U	C1	20	6	m8	C18	69	18	m5.8	C17	384	89
Z	C6	0	0	V	C3	1	1	J	C11	23	8	F435W	C7	53	9	K	C14	340	96
F775W	C9	0	0	R	C4	3	1					H	C12	47	13	HK	C13	112	32
F850LP	C10	0	0	F606W	C8	1	0												
m4.5	C16	0	0	m3.6	C15	1	0												

Table 1. Number of NaN's in the various bands. Columns B and ID denote the photometric band and the sequential numbering of features used in the experiments. For each band/feature the column F gives the number of missing values (NaN's) in the full dataset, while the column T gives the same number for the training set. As it is explained in the text, the features are grouped in good, quite good, acceptable ,medium and poor.

exp. n	missing feat.	feat.	hid.	wstep	res.	de	MxIt	CV	rms	out%	bias
37	13,14,17,18	14	29	0,0001	30	0,1	3000	10	0,057	22,61%	-0,0077
26	13,14,15,16,17,18	12	25	0,0001	30	0,1	3000	10	0,062	17,39%	0,0078

Table 2. Description of the best experiments for the 18 bands (Exp. n. 37) and the 14 bands datasets (Exp. n. 26), respectively Column 1: sequential experiment identification code; column 2: features which wer not used in the experiment; columns 3-4: number of input and hidden neurons; column 5–9: parameters of the MLP-QNA used during the experiment; column 10: rms error evaluated as described in the text; column 11: fraction of outliers; column 12: bias.

A	18-band; $\|\Delta z\| \leq 0.15$			14-band; $\|\Delta z\| \leq 0.15$			18-band; R< 24; $\|\Delta z\| \leq 0.15$			14-band R< 24; $\|\Delta z\| \leq 0.15$		
Code	bias	scatter	out	bias	scatter	out	bias	scatter	out	bias	scatter	out
AN-e	-0,0100	0,074	31,00	-0,0060	0,078	38,50	-0,0130	0,071	24,40	-0,0070	0,076	32,80
EC-e	-0,0010	0,067	18,40	0,0020	0,066	16,70	-0,0060	0,064	14,50	-0,0030	0,064	13,50
PO-e	-0,0090	0,052	18,00	-0,0070	0,051	13,70	-0,0090	0,047	10,70	-0,0080	0,046	7,10
RT-e	-0,0090	0,066	21,40	-0,0080	0,067	24,20	-0,0120	0,063	16,40	-0,0120	0,064	18,40
QNA	0,0006	0,056	16,33	0,0028	0,063	19,35	0,0002	0,053	11,70	0,0016	0,060	13,75
B	18-band; $\|\Delta z\| \leq 0.5$			14-band; $\|\Delta z\| \leq 0.5$			18-band; R< 24; $\|\Delta z\| \leq 0.5$			14-band; R< 24; $\|\Delta z\| \leq 0.5$		
AN-e	-0,0360	0,15	3,10	-0,035	0,173	4,20	-0,047	0,13	1,40	-0,047	0,13	1,40
EC-e	-0,0070	0,12	3,60	-0,003	0,114	3,60	-0,015	0,11	1,90	-0,015	0,11	1,90
PO-e	-0,0130	0,12	3,10	0,001	0,107	2,30	-0,020	0,10	1,20	-0,020	0,10	1,20
RT-e	-0,0310	0,12	3,20	-0,028	0,137	3,60	-0,034	0,11	1,40	-0,034	0,11	1,40
QNA	-0,0028	0,11	3,78	-0,005	0,125	3,83	-0,004	0,10	1,66	-0,004	0,10	1,66

MLP-QNA: regression

Photometric redshifts of Quasars: Longo et al. 2012 in prep.

Survey	Bands	Name of feature	Synthetic description
SDSS	u, g, r, i, z	psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_ z	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y,J,H,K	yPsfMag, j_1PsfMag, hPsfMag, kPsfMag	PSF fitting magnitude in Y, J, H, K bands
	Y,J, H,K	Es. for y band: yAperMag3, yAperMag4, yAperMag6	aperture photometry through 2, 2.8 & 5.7″ circular aperture
	J,H,K	Es. for J band: jHallMag, JPetroMag	Calibrated magnitude within circular aperture r_hall and Petrosian magnitude
GALEX	NUV	Nuv_mag, Nuv_mag_iso;	Respectively: Near UV total and isop. mags
	NUV	Nuv_mag_Aper_1 Nuv_mag_Aper_2 Nuv_mag_Aper_3	aperture photometry through 2,3 & 5 pxl apertures
	NUV	Nuv_mag_auto and Nuv_kron_radius	magnitudes and Kron radius in units of A or B
	FUV	Fuv_mag, Fuv_mag_iso;	Respectively: Far UV total and isop. mags
	FUV	Fuv_mag_Aper_1 Fuv_mag_Aper_2 Fuv_mag_Aper_3	aperture photometry through 2,3 & 5 pxl apertures
	FUV	Fuv_mag_auto and Fuv_kron_radius	magnitudes and Kron radius in units of A or B
WISE	W1, W2, W3, W4	W1mpro, W2mpro, Wmpro, Wmpro4	W1: 3.4 μm and 6.1″ angular resolution, W2: 4.6 μm and 6.4″ angular resolution. W3 12 μm and 6.5″ W4 22 μm and 12″ angular resolution Magnitudes measured with profile-fitting photometry at the 95% level. Brightness upper limit if the flux measurement has SNR<2
SDSS	-	zspec	Spectroscopic redshift

1	2	3	4	5	6	7	8	9	10	11	12
E1	X	X	X	X	All	0,0033	0,174	15,96%	4,75%	2,24%	0,92%
E2	X	X	X	X	UKIDSS: hall CALEX: mag + mag_iso	-0,0001	0,152	19,66%	4,49%	1,85%	0,92%
E3	X	X	X	X	UKIDSS: hall CALEX: Aper 1, 2 ,3	-0,0016	0,165	15,83%	3,96%	1,98%	1,19%
E4	X	X	X	X	UKIDSS: hall CALEX: mag	0,0054	0,151	16,23%	4,75%	1,98%	1,06%
E5	X	X	X	X	UKIDSS: hall CALEX: mag_iso	-0,0026	0,151	18,47%	4,62%	2,37%	0,79%
E6	X	X	X	X	UKIDSS: hall CALEX: mag_auto + kron radius	-0,0008	0,152	17,81%	5,15%	2,64%	0,79%
E7	X	X	X	X	UKIDSS: hall CALEX: mag + mag_iso + Aper 1, 2, 3	0,0041	0,163	19,39%	4,22%	2,51%	0,66%
E8	X	X	X	X	UKIDSS: hall CALEX: mag_iso + Aper 1, 2, 3	-0,0033	0,155	19,26%	5,01%	1,98%	0,92%
E9				X	All	0,0165	0,297	22,16%	5,80%	2,11%	0,53%
E10	X				All	-0,0162	0,338	19,66%	7,26%	2,37%	0,40%
E11		X			UKIDSS: hall + petro	-0,0091	0,299	23,75%	4,88%	1,58%	0,66%
E12			X		CALEX: mag + mag_iso	0,055	0,419	29,68%	4,75%	0,79%	0,26%
E13		X			UKIDSS: petro	0,0111	0,465	34,43%	3,43%	0,40%	0,00%
E14		X			UKIDSS: hall	-0,0081	0,294	22,82%	5,94%	1,85%	0,66%
E15		X		X	UKIDSS: hall	0,0045	0,236	17,94%	4,75%	2,11%	1,06%
E16	X	X	X		UKIDSS: hall CALEX: mag_iso	-0,0046	0,152	21,11%	4,88%	1,98%	0,79%
E17	X		X	X	CALEX: mag_iso	0,0025	0,162	16,23%	3,69%	2,37%	1,06%
E18	X	X		X	UKIDSS: hall	-0,0032	0,179	14,38%	4,49%	2,11%	1,32%
E19		X	X	X	UKIDSS: hall CALEX: mag_iso	0,011	0,203	19,26%	4,88%	1,72%	0,79%
E20			X	X	CALEX: mag_iso	0,0175	0,288	22,96%	4,88%	1,45%	0,53%
E21	X	X			UKIDSS: hall	-0,0027	0,21	15,96%	5,15%	2,24%	1,06%
E22	X			X	All	-0,0039	0,197	13,85%	3,43%	2,37%	1,58%
E23	X		X		CALEX: mag_iso	-0,0055	0,24	17,56%	6,73%	2,51%	0,79%
E24		X	X		UKIDSS: hall CALEX: mag_iso	0,0133	0,238	23,22%	6,20%	1,72%	0,40%

Feature selection

WISE substantially useless

Mag_iso substantially useless

TEST	MEAN	σ	out. 1σ	out. 2σ	out. 3σ	out. 4σ	TOTAL OBJECTS
E5	0,0005	0,118	18,67%	4,01%	1,51%	0,87%	3787
E16	-0,0004	0,154	18,11%	4,75%	1,98%	0,98%	3787

Table 3. Summary of the statistical indicators alread used in Table xx (bias, σ and the percentage of outliers at, respectively, 1,2,3 and 4 σ computed as in citebovy2012 on all objects (test and training set).

ZSPEC BIN	EXP	BIAS	SIGMA	$\|\Delta Z\| > 0.1$	$\|\Delta Z\| > 0.2$	$\|\Delta Z\| > 0.3$	$\|\Delta Z\| > 0.4$	OBJECTS
TRAIN Only								
[0.2, 1.0]	E23	-0.0897	0.206	44.94%	22.78%	13.29%	8.86%	316
[0.2, 1.0]	E5	-0.0183	0.118	27.53%	7.28%	2.22%	1.27%	316
[0.2, 1.0]	E16	-0.029	0.127	29.43%	10.76%	3.48%	1.90%	316
[0.2, 1.0]	E10	-0,1807	0,281	64,87%	39,87%	26,58%	18,67%	316
[1.4, 3.0]	E23	0.1209	0.273	59.05%	32.33%	21.98%	14.22%	232
[1.4, 3.0]	E5	0.0364	0.18	38.36%	14.66%	8.19%	4.74%	232
[1.4, 3.0]	E16	0.0408	0.183	40.09%	15.52%	8.62%	4.74%	232
[1.4, 3.0]	E10	0,2188	0,367	62,50%	41,38%	28,88%	22,84%	232
TRAIN+TEST								
[0.2, 1.0]	E23	-0.0911	0.23	46.24%	23.18%	13.77%	9.03%	1583
[0.2, 1.0]	E5	-0.0174	0.101	21.04%	4.17%	1.58%	0.82%	1583
[0.2, 1.0]	E16	-0.0326	0.142	30.01%	9.85%	4.67%	2.65%	1583
[0.2, 1.0]	E10	-0,1877	0,287	63,93%	39,55%	27,48%	19,08%	1583
[1.4, 3.0]	E23	0.1238	0.269	56.24%	30.74%	18.69%	12.49%	1145
[1.4, 3.0]	E5	0.0271	0.139	31.44%	9.61%	3.93%	2.10%	1145
[1.4, 3.0]	E16	0.0492	0.183	39.83%	14.93%	7.95%	4.37%	1145
[1.4, 3.0]	E10	0.2488	0,37	64,28%	44,02%	32,23%	24,72%	1145

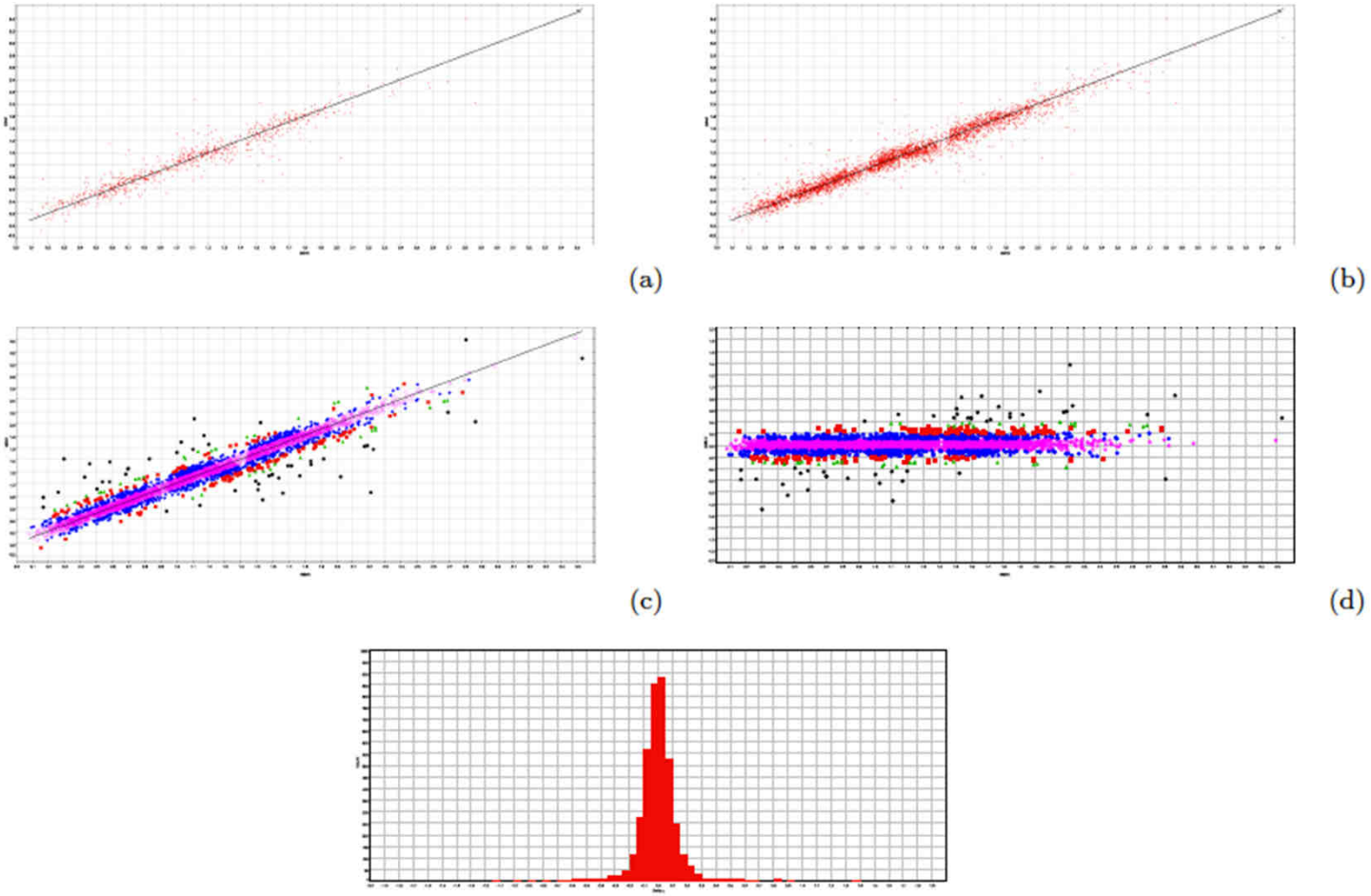


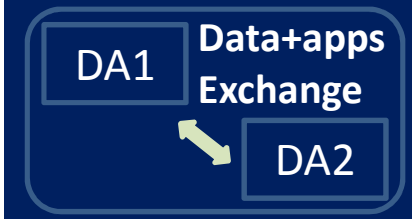
Figure B1. Experiment E5: (a) Spectroscopic vs photometric redshift distribution on the test set(758 objects); (b) spectroscopic vs photometric redshift distribution on the whole set, train + test set (3787 objects); (c) binned redshift distribution zspec vs zphot on the whole set; (d) binned redshift distribution zspec vs Δz on the whole set; (e) histogram of Δz

Moving programs not data: the true bottle neck

Data Mining + Data Warehouse =
Mining of Warehouse Data

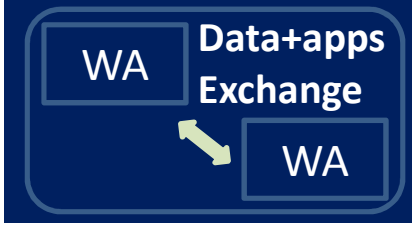
- For organizational learning to take place, data from must be gathered together and organized in a consistent and useful way – hence, Data Warehousing (DW);
- DW allows an organization to remember what it has noticed about its data;
- Data Mining apps should be interoperable with data organized and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

After a certain number of such iterations…

The scenario will become:

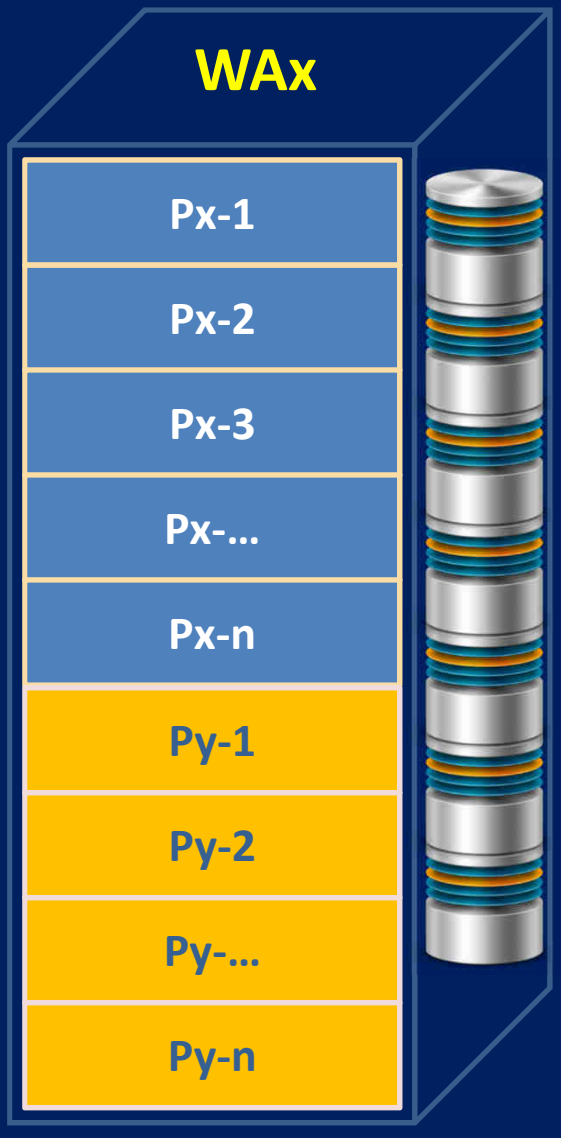
No different WSs, but simply one WS with several sites (eventually with different GUIs and computing environments)

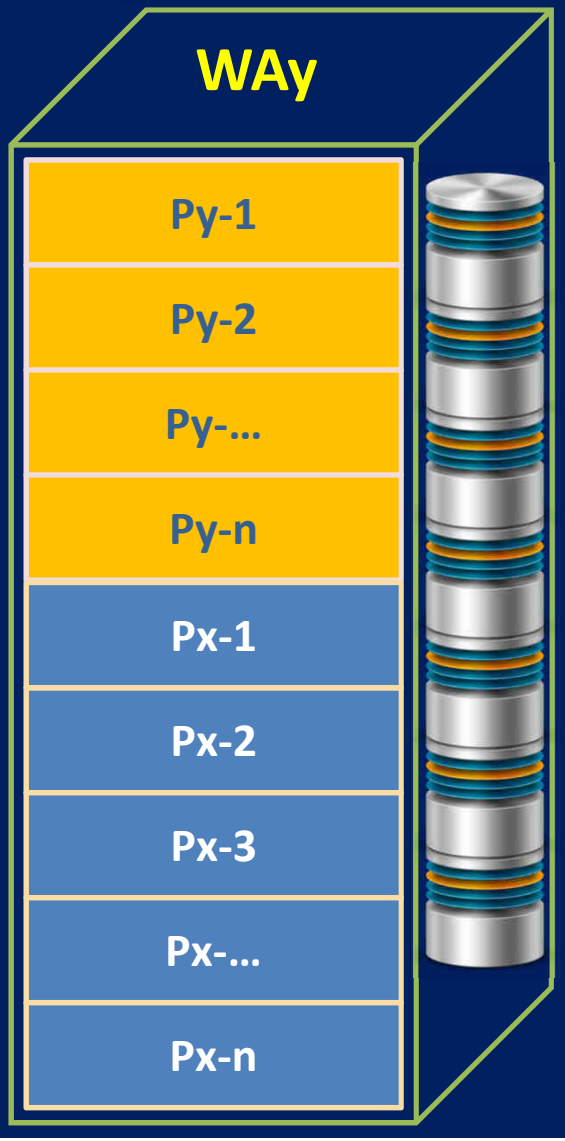
All WS sites can become a mirror site of all the others

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

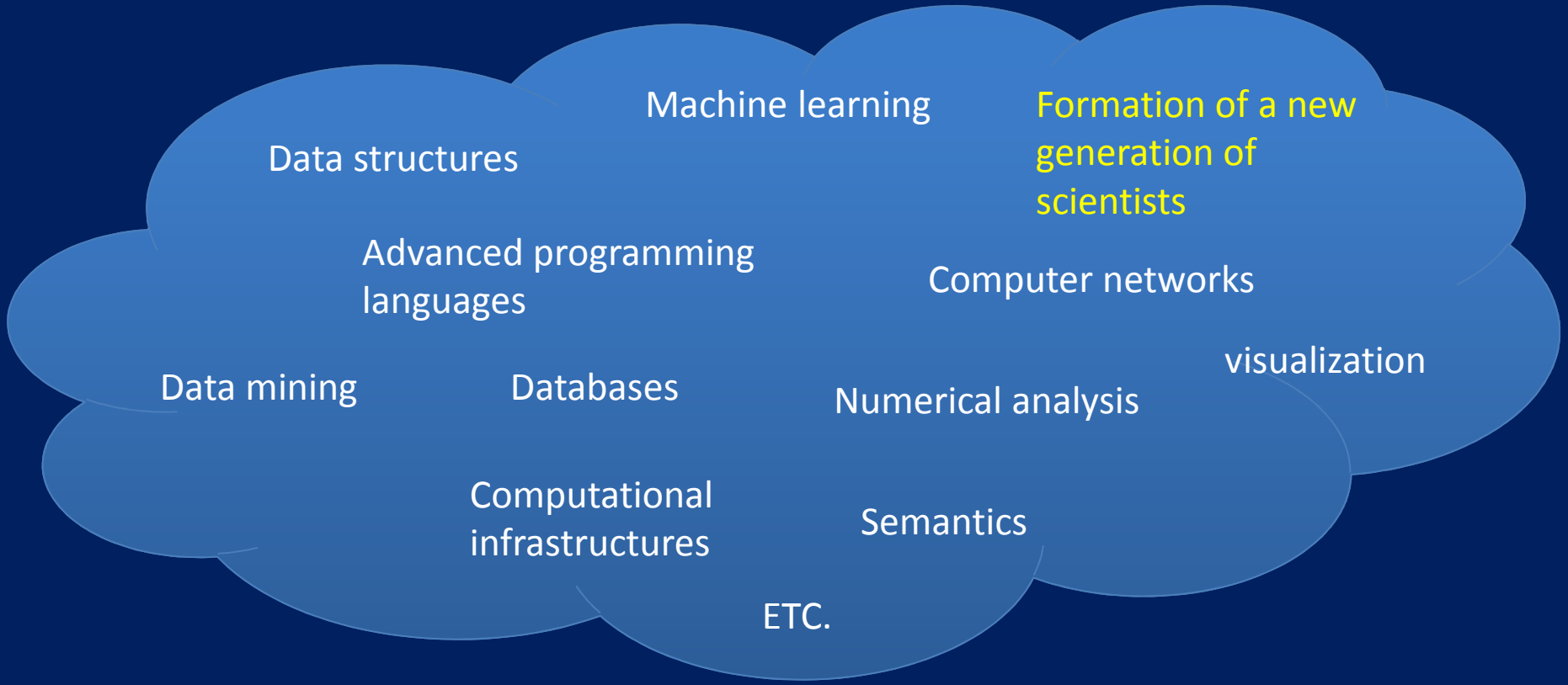
Minimization of data exchange flow (just few plugins in case of synchronization between mirrors)

YES MDS!





A new discipline in the making: AstroInformatics



Very lively Community - AstroInformatics International Conferences

2010 – Pasadena
2011 – Napoli
2012 – Redmond (Microsoft)
2013 – South Africa

Join us on Astroinformatics page on Facebook

IVOA – IG on KDD WIKI

CONCLUSIONS

1. Astronomy has become data rich.

Bad things: most data will be lost if new technologies (ITC) are not exploited to their best, need for computational power, need to transform the profession

Good things: enormous potential for new discoveries (also from archives),
New attraction power for many (larger and richer) communities (astronomical data are complex, large and FREE….), possibility to exploit advances in similar fields

2. Telescopes (also robotic networks) in order to be competitive need to be cutting edge technology and are expensive.
Their optimal use makes them very very expensive (software engineers, programmers, data managers, computing infrastructures, etc.)

CONCLUSIONS

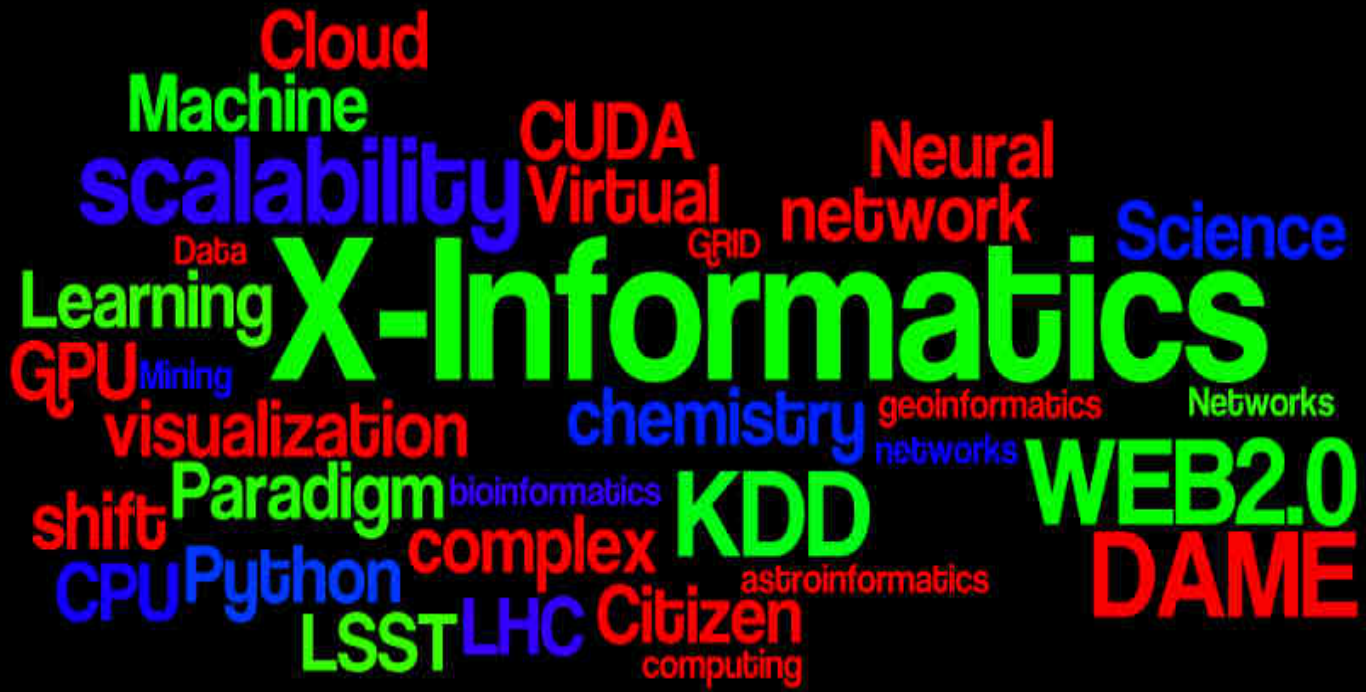
3. Resources are concentrated in few countries, intelligence and know how are not … and both trends will continue in the future

4. data overabundance calls for shorter (-> 0) proprietary periods and therefore the capability of making discoveries will more and more depend on the capability to extract information from complex data …

Data producers will require advanced know-how in data processing and powerful computational facilities (could be also exchanged for time?)
Astronomers of the future will see less and less of telescopes and more and more of large computational infrastructures (need for investements in changing the way we teach it ...)

5. ASTROINFORMATICS
Is not just using computers for astronomy

THE END


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