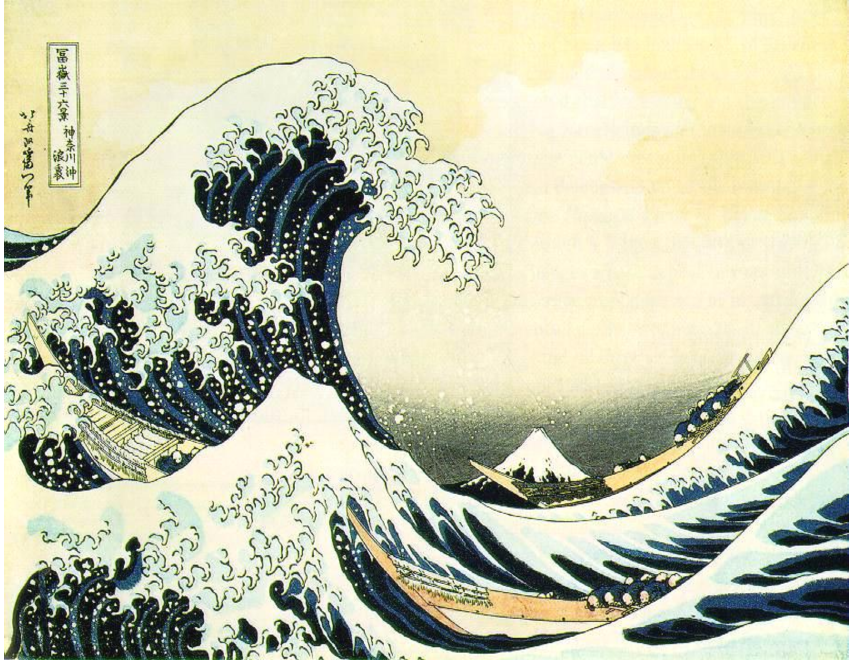
Knowledge discovery in astrophysics: massive data sets, virtual observatory and beyond

astrophysics and the data tsunami

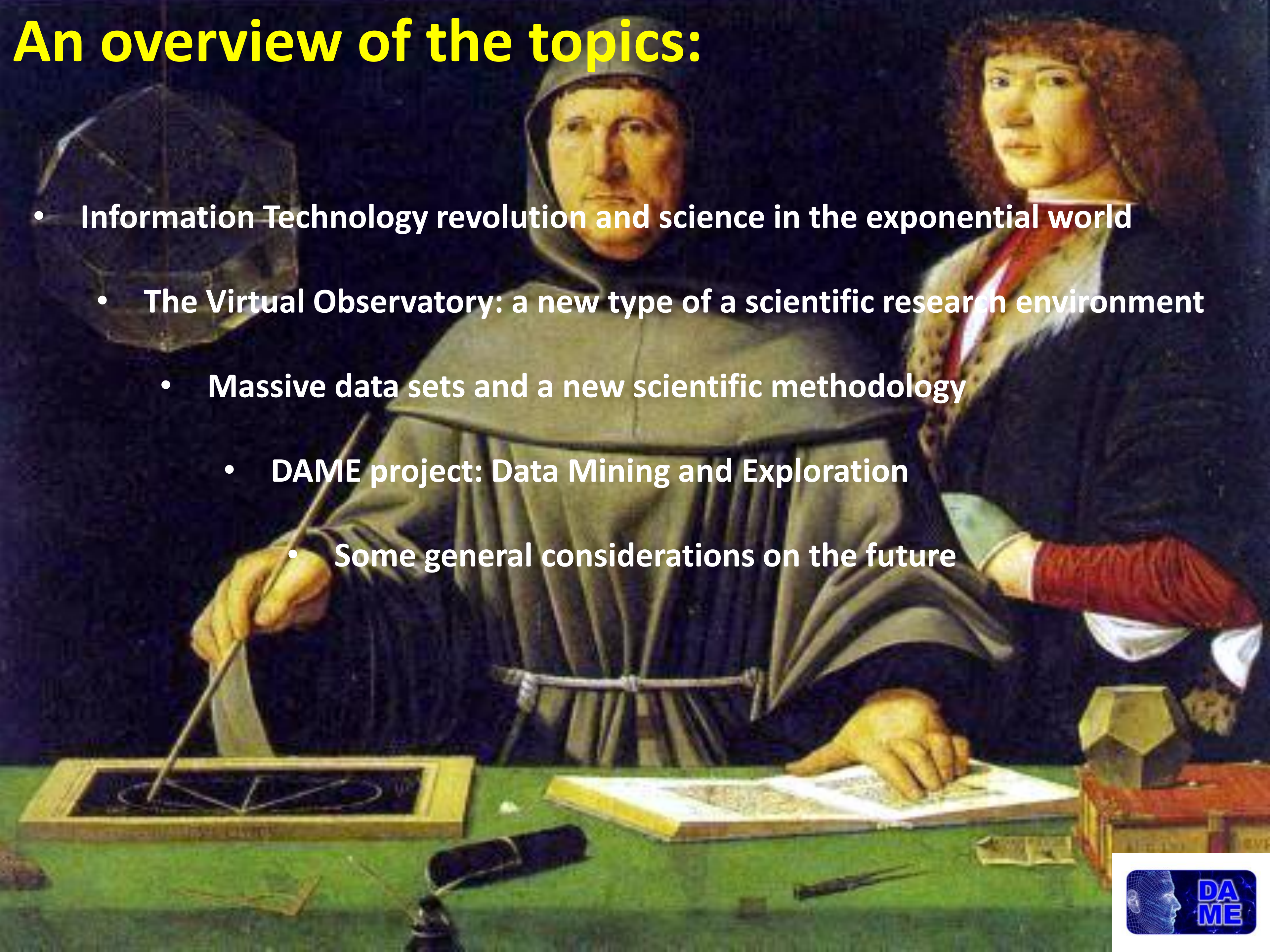


M. Brescia[1], G. Longo[2]

1- INAF – Astronomical Observatory of Capodimonte in Napoli (longo@na.infn.it)
2 - Department of Physical Sciences - University Federico II Napoli

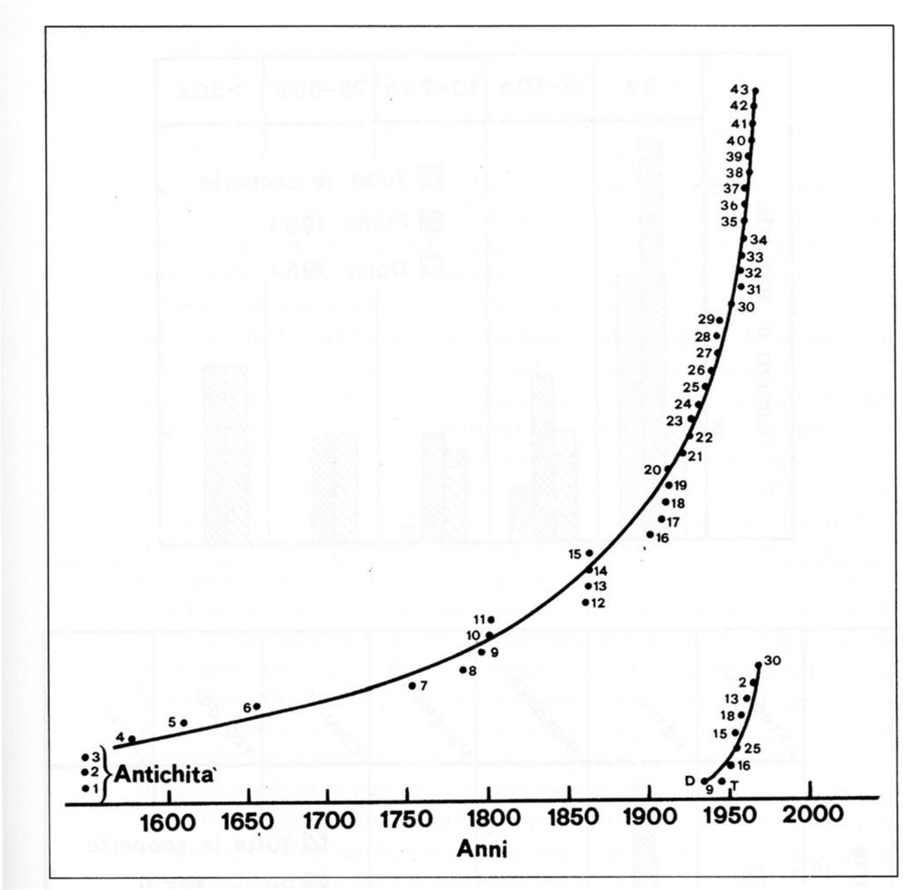
Napoli, february 3-rd, 2010

An overview of the topics:



- Information Technology revolution and science in the exponential world
- The Virtual Observatory: a new type of a scientific research environment
- Massive data sets and a new scientific methodology
- DAME project: Data Mining and Exploration
- Some general considerations on the future

Discoveries in astronomy

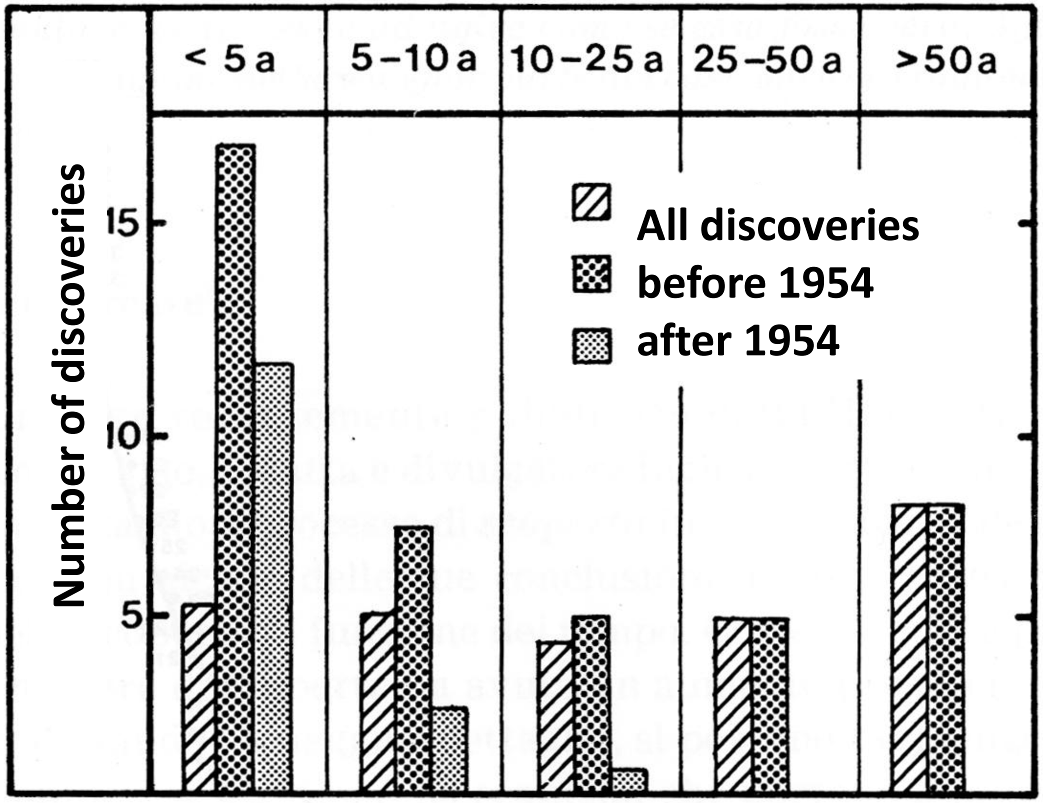


From M.Harwit, Cosmic discoveries

1. Stars
2. Planets
3. Novae
4. Comets
5. Satellites
6. Rings
7. Galactic clusters
8. Galaxy clusters
9. Interplanetary dust
10. Asteroids
11. Binary stars
12. Variable stars
13. Planetary nebulae
14. Globular clusters
15. HII regions
16. Cold ISM
17. Giant stars
18. Cosmic rays
19. Pulsating variables
20. White dwarfs
21. Galaxies
22. Expansion of universe
23. Cosmic dust
24. Supernovae/novae
25. Gas in galaxies
26. SN remnants
27. Radiogalaxies
28. Magnetic variables
29. Flare stars
30. Intergalactic magnetic fields
31. X stars
32. X background
33. Quasar
34. CMB
35. Masers
36. Infrared stars
37. X galaxies
38. Pulsar
39. Gamma background
40. IR galaxies
41. Superluminal sources
42. GRB
43. Unidentified radio sources
44. …
45. ….

The role of technology

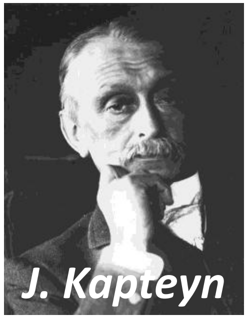
Most discoveries take place immediately after a technological breaktrough



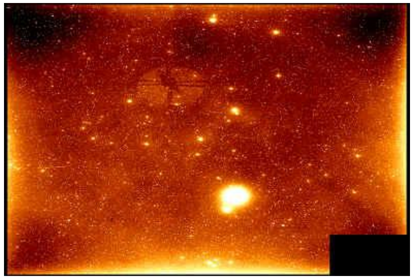
An historical perspective

1910
Final settling of stellar statistics, by the work of Kapteyn, Oort, etc.)

S.I.L.



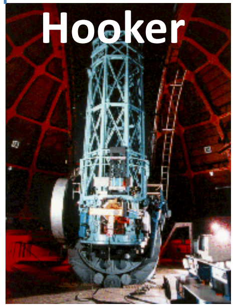
1960's
Photographic wide field plates (PSS)



XXI century
Renaissance of statistical astronomy (synoptic surveys)

Rush for the larger and the bigger

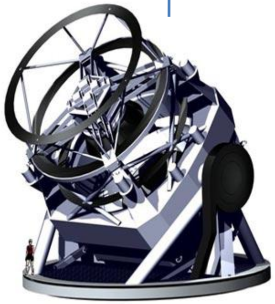
20's



Palomar

80's

Virtual Obs.



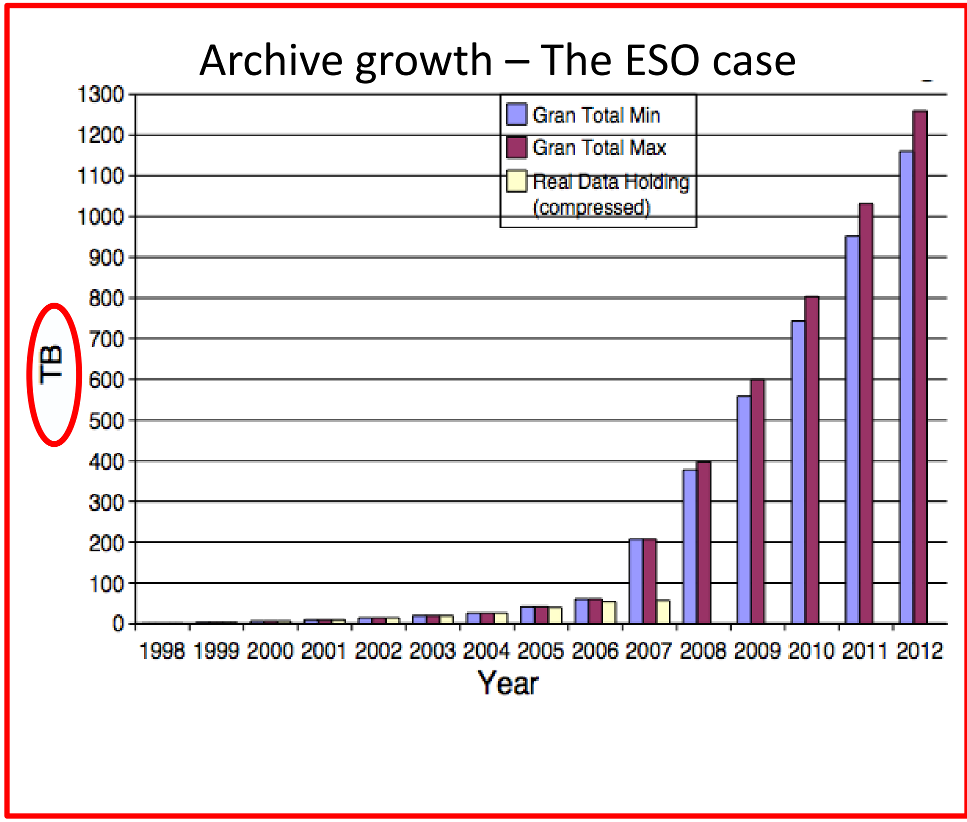
LSST (2013)

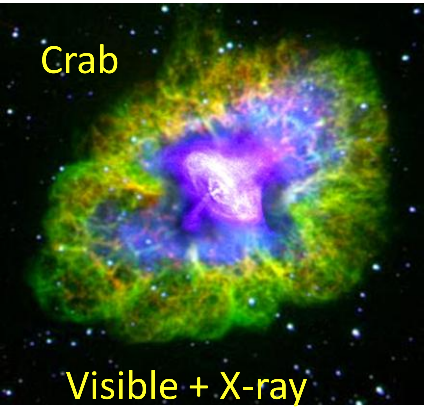
Few objects, few λ's, heterogeneous data

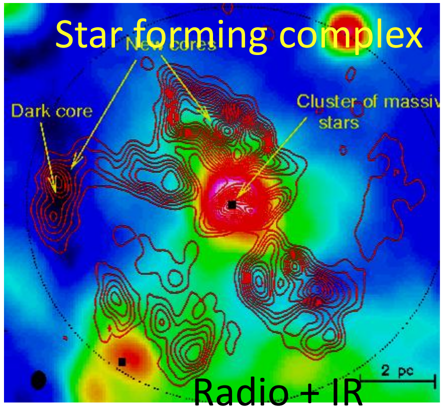
Astrophysics as a data rich science

- Telescopes (ground- and space-based, covering the full electromagnetic spectrum)
- Instruments (telescope/band dependent)

- **Large digital sky surveys** are becoming the dominant source of data in astronomy: ~ 10-100 TB/survey (soon PB), ~ 10^6 - 10^9 sources/survey, many wavelengths…
- **Data sets many orders of magnitude larger, more complex, and more homogeneous than in the past**



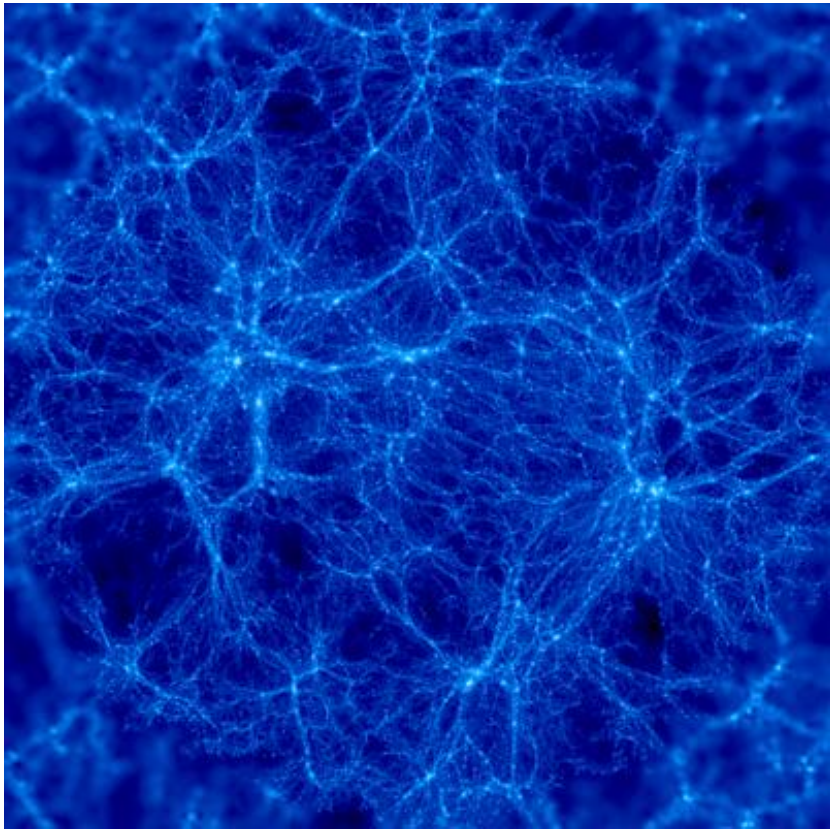




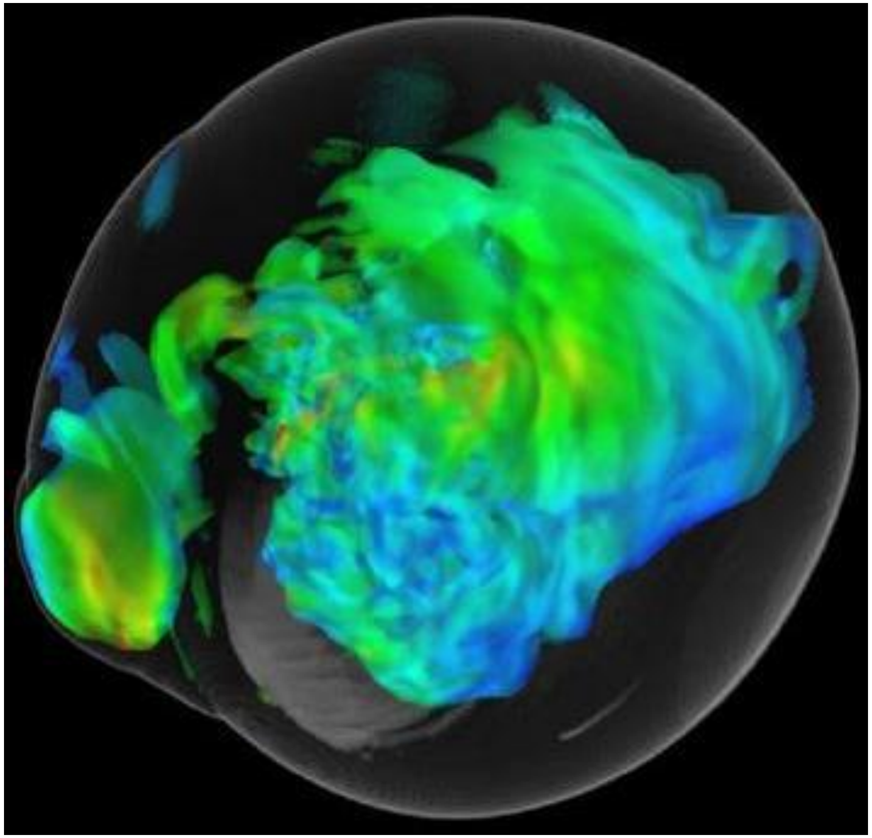
Panchromatic Views of the Universe:
Data Fusion - A More Complete, Less Biased Picture

2. The astronomical data tsunami:

Theoretical Simulations Are Becoming More Complex and Generate TB's of Data …



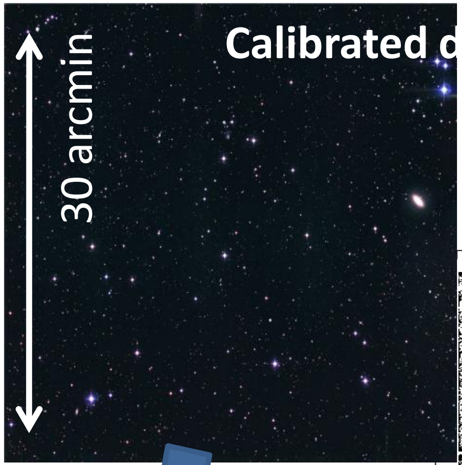
Structure formation in the Universe



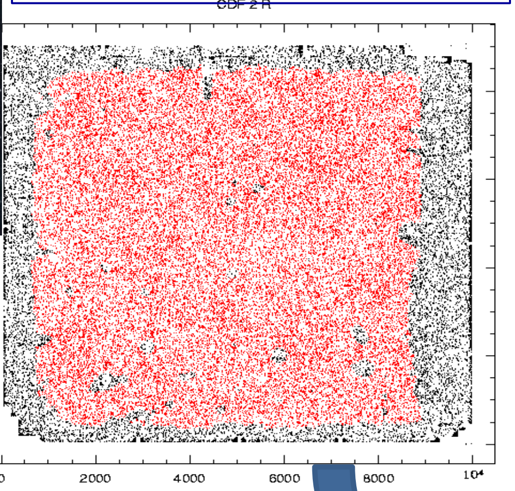
Supernova explosion instabilities

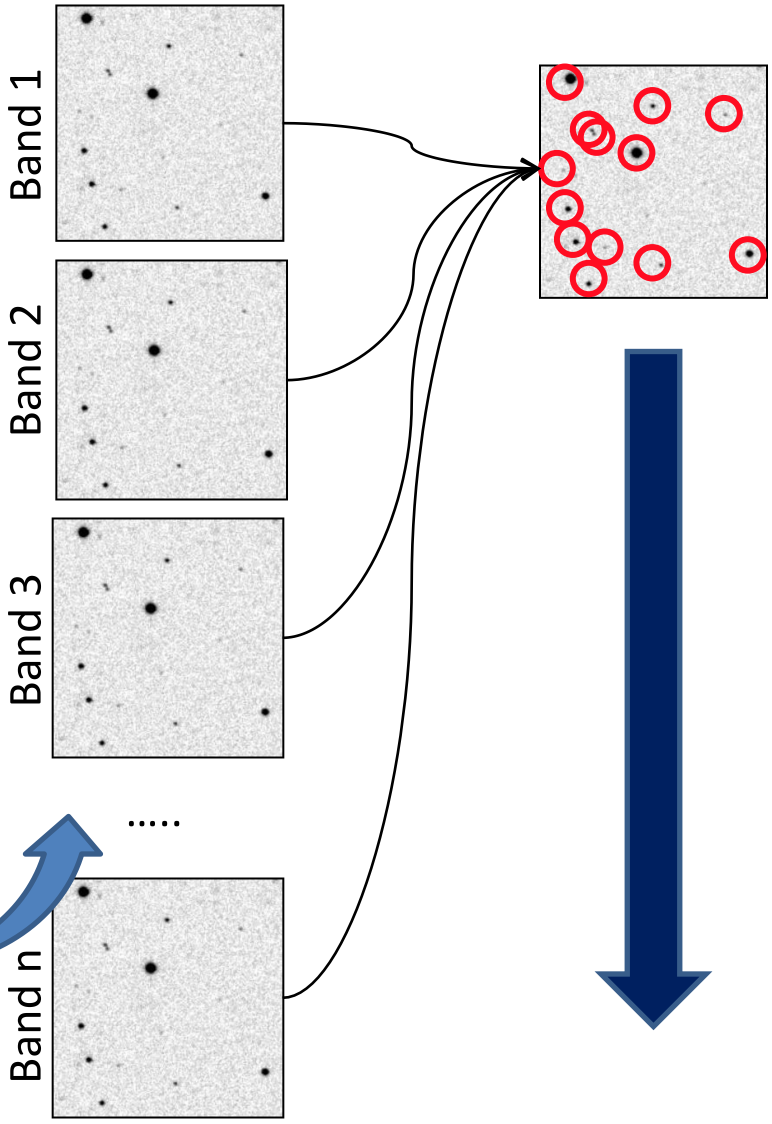
Comparing the massive, complex output of such simulations to equally massive and complex data sets is a non-trivial problem!

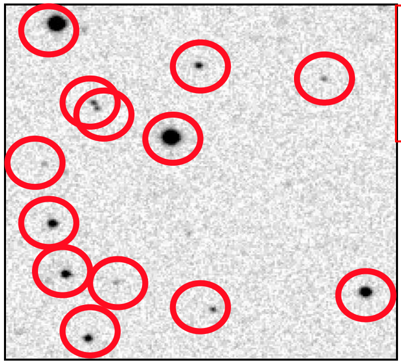
3. The data complexity: the parameter space



1/160.000 of the sky, not sodeep (25.0 in r)
55.000 detected sources
(0.75 mag above m lim)







Detect sources and measure their attributes (brightness, position, shapes, etc.)

p={isophotal, petrosian, aperture magnitudes concentration indexes, shape parameters, etc.}

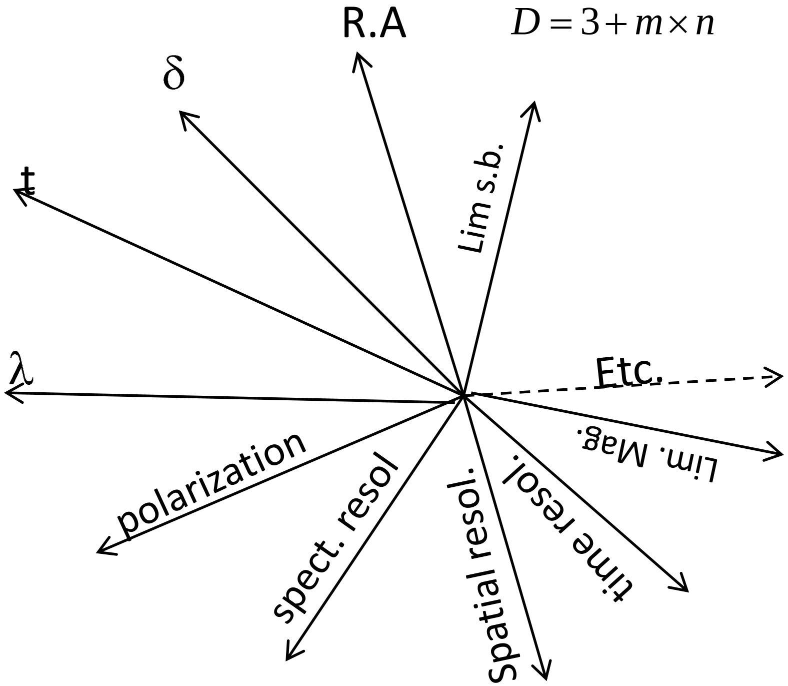
$$p^1 = \{RA^1, \delta^1, t, \{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m}\}\}$$

$$p^2 = \{RA^2, \delta^2, t, \{\lambda_1, \Delta\lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \ldots, f_1^{2,m}, \Delta f_1^{2,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \ldots, f_n^{2,m}, \Delta f_n^{2,m}$$

....

$$p^N = \{RA^N, \delta^N, t, \{\lambda_1, \Delta\lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \ldots, f_1^{N,m}, \Delta f_1^{N,m}\}, \ldots\}$$

$$D = 3 + m \times n$$



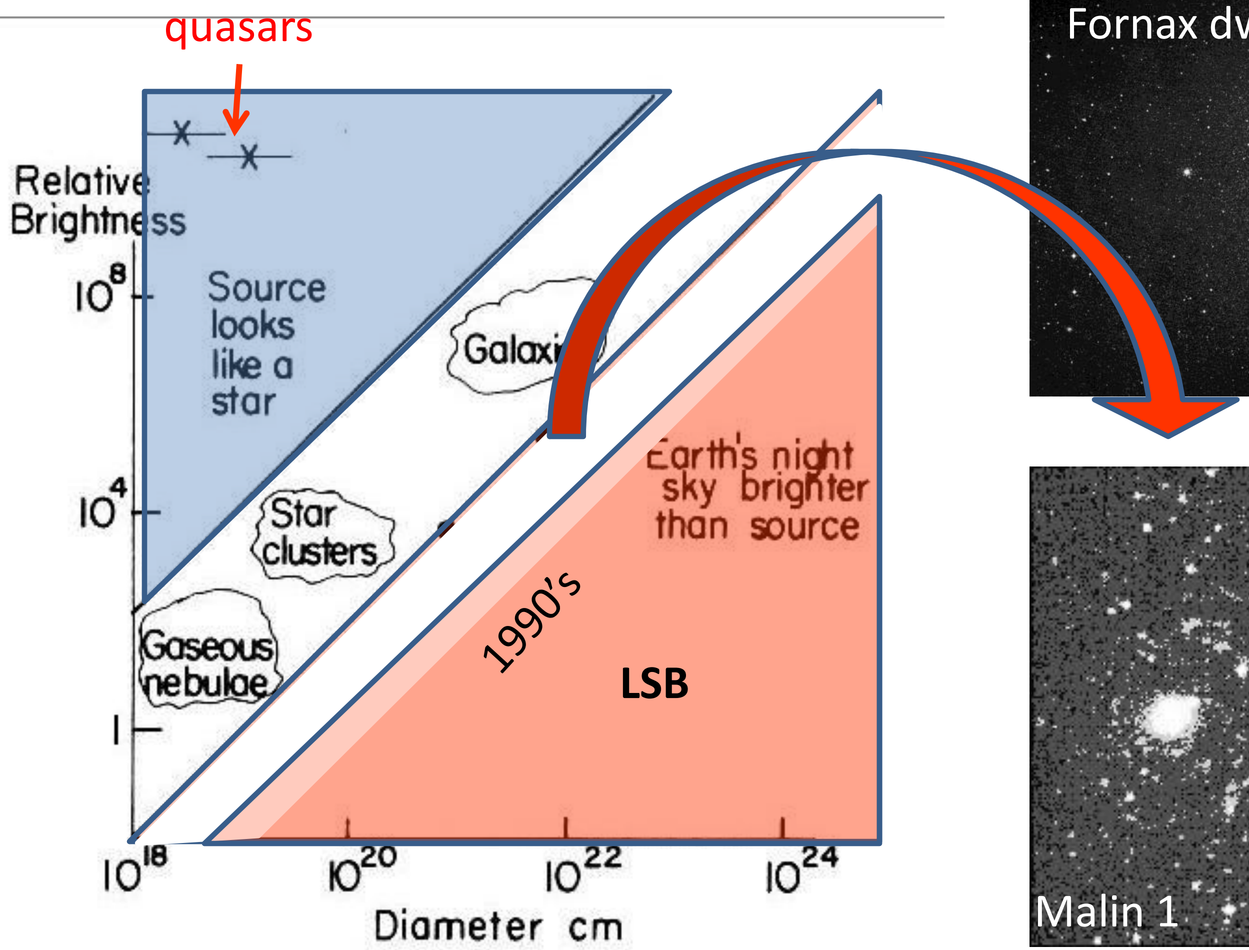
PARAMETER SPACE

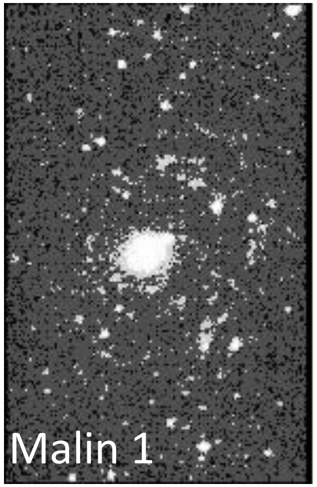
From the Data Mining point of view, a**ny observed (simulated) datum *p* defines a point (region) in a subset of R^N.**

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

Every time you improve the coverage of the PS….

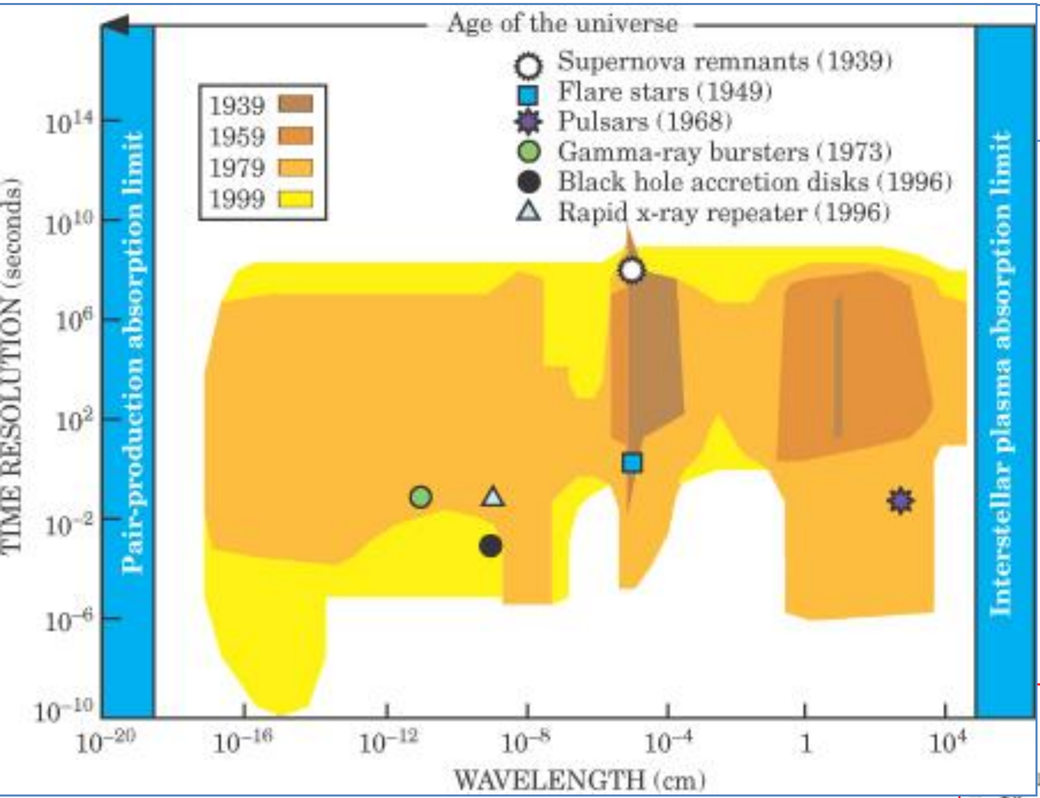
Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place



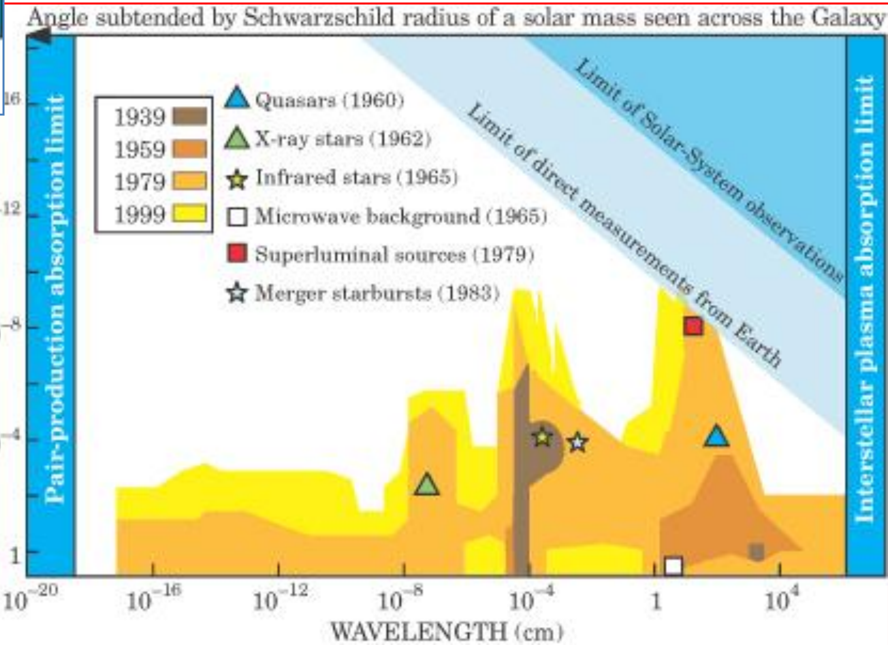


Discovery of Low surface brightness Universe

Improving the coverage of the Parameter Space - II



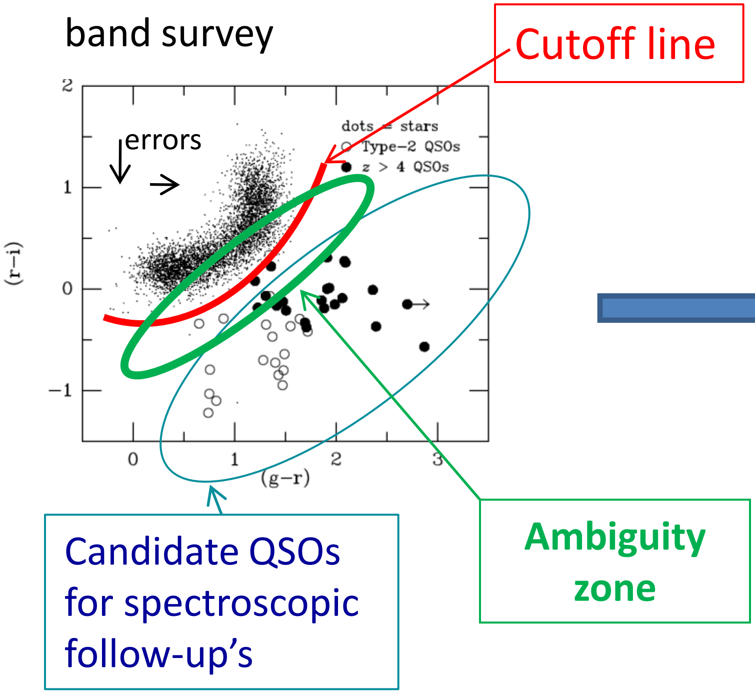
Projection of parameter space along (time resolution & wavelength)



Projection of parameter space along (angular resolution & wavelength)

More dimensions allow better disentanglement

Traditional way to look for candidate QSO in 3 band survey

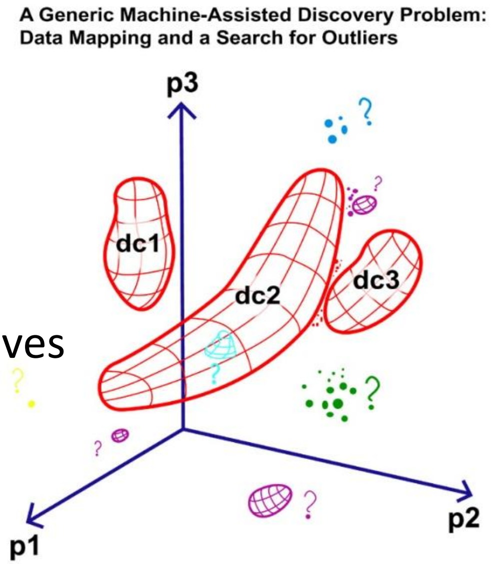


Cutoff line

Ambiguity zone

Candidate QSOs for spectroscopic follow-up's

Adding one feature improves separation...


A Generic Machine-Assisted Discovery Problem:
Data Mapping and a Search for Outliers
p3
p2
p1
dc1
dc2
dc3
?

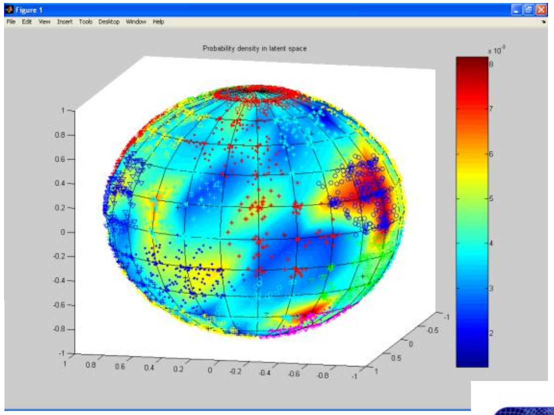


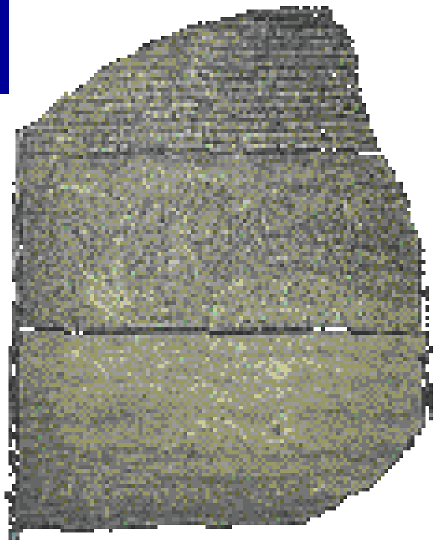
Figure 1
File Edit View Insert Tools Desktop Window Help
Probability density in latent space

PPS projection of a 21-D parameter space showing as blue dots the candidate quasars.Notice better disentanglement

And now, the question is….

Where to search … for the next discoveries?
And what has mathematics got to do with it?

Considerations on the next breakthroughs



- We have reached the physical limit of observations (single photon counting) at almost all wavelenght…
- Detectors are linear & all electromagnetic bands have been opened

Hence

Our capability to gain new insights on the universe will depend mainly on:

- Capability to recognize patterns or trends in the parameter space (i.e. physical laws) being not limited by human 3-D visualization

- Capability to extract patterns from very large multiwavelenght, multiepoch, multi-technique parameter spaces

➡ ***Most data will never be seen by humans!***

The need for data storage, network, database-related technologies, standards, etc.

Information complexity is also increasing greatly

➡ ***Most knowledge hidden behind data complexity is lost***

Most (all) empirical relationships known so far depend on 3 parameters ….
Simple universe or rather human bias?

➡ ***Most data (and data constructs) cannot be comprehended by humans directly!***

The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery

The answer is Data mining …. matching Donald Rumsfeld's epistemology

There are known knowns,
There are known unknowns, and
There are unknown unknowns

Donald Rumsfeld's about Iraqi war

(AP/FILE)

Classification

Morphological classification of galaxies

Star/galaxy separation, etc.

Regression

Photometric redshifts

Clustering

Search for peculiar and rare objects,

Etc.

Extracting knowledge

The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for **hidden patterns**, trends, etc. **among N points in a DxK dimensional parameter space**:

MASSIVE, COMPLEX DATA SETS with:

$N > 10^9$, $D >> 100$, $K > 10$

The computational cost of Data Mining:

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 computing power**

Need for a new science: Astroinformatics

Knowledge Discovery in Databases

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

→ Data Farming:

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

Database technologies

→ Data Mining (or Knowledge Discovery in Databases):

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

Key mathematical issues

→ Data understanding

Computer aided understanding
KDD
Etc.

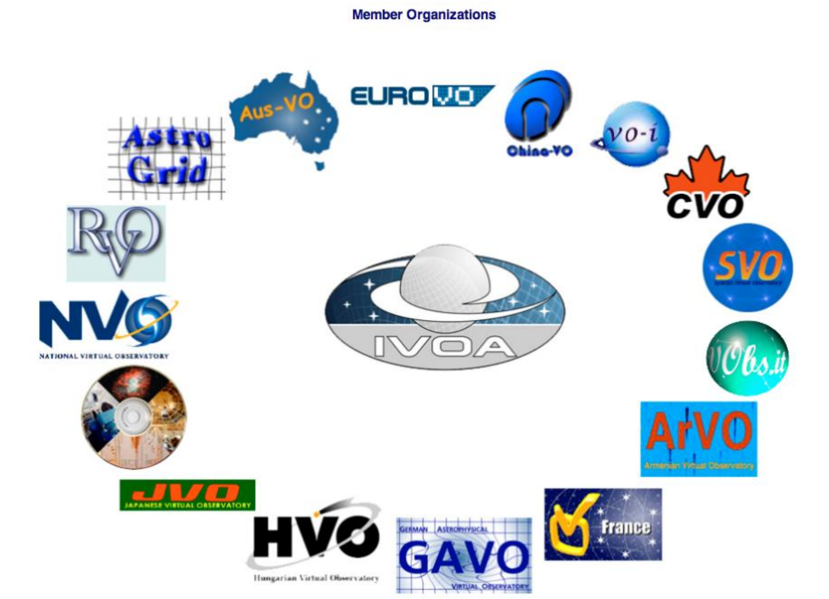
Ongoing research

→ New Knowledge

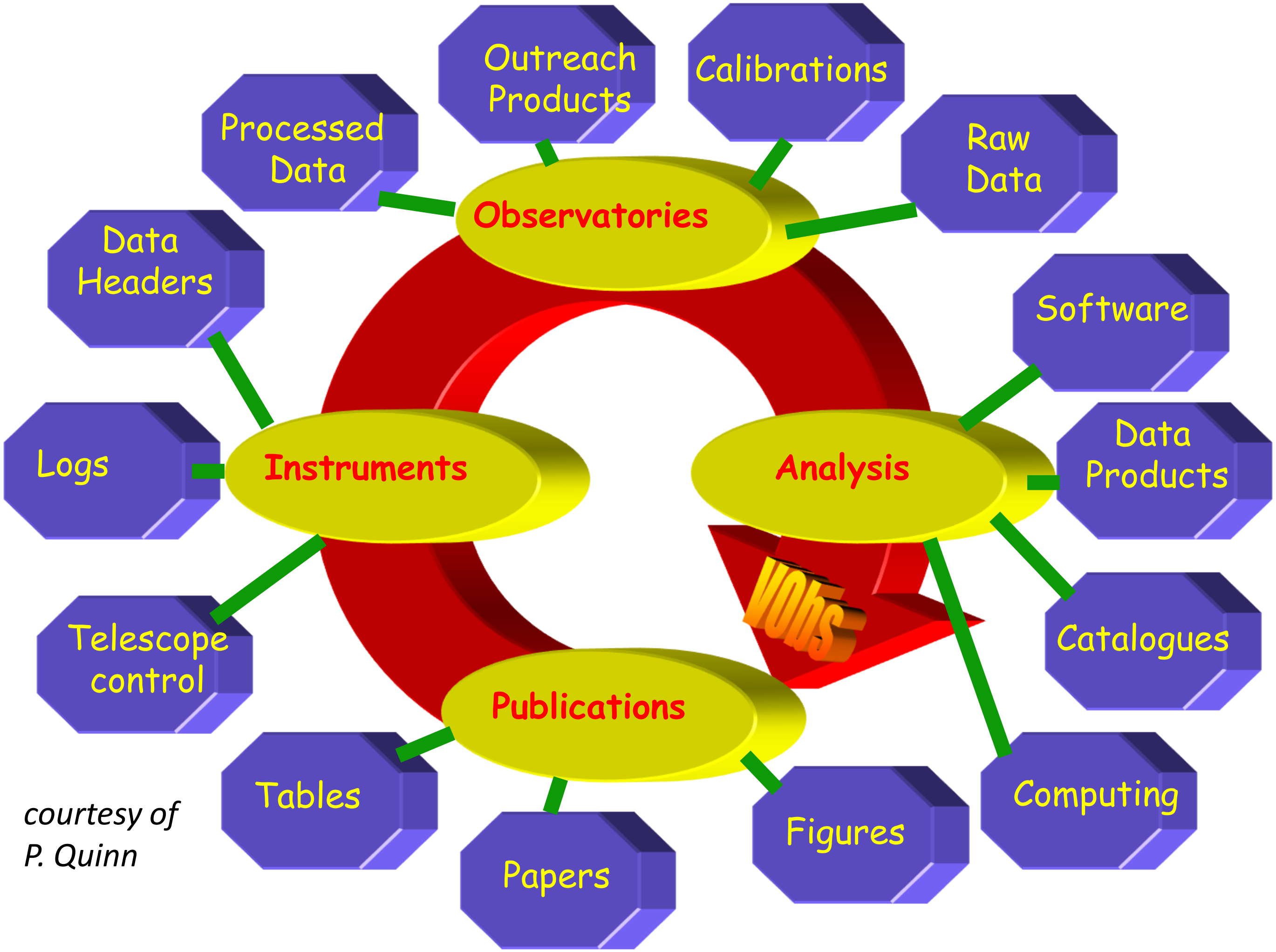
OK, So …

Which is the answer of the astronomical community?

The Virtual Observatory (VObs)



VObs Represents a New Type of a Scientific Organization for the era of information abundance

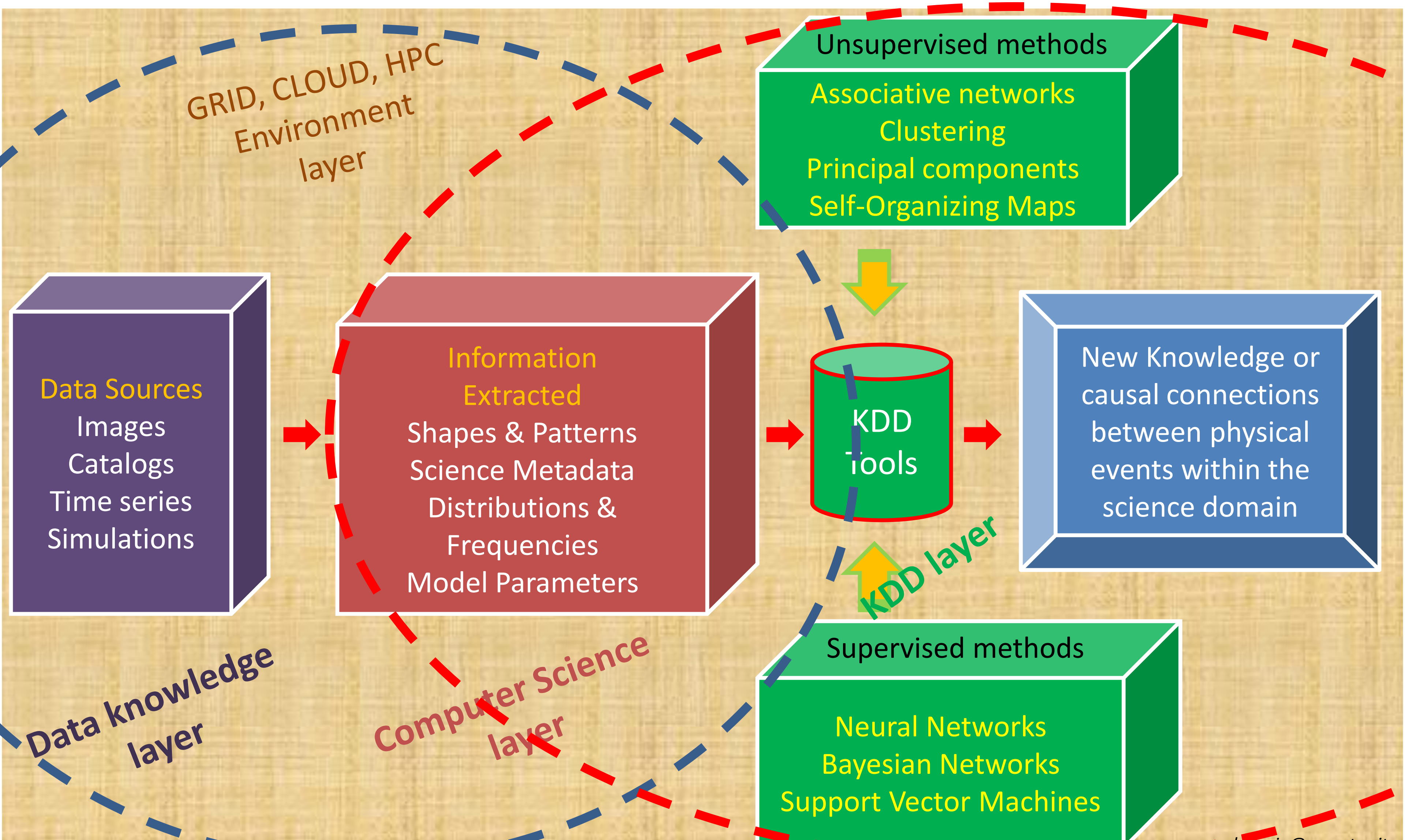


courtesy of P. Quinn

- It is inherently ***distributed,*** and web-centric
- It is fundamentally based on a ***rapidly developing technology*** (IT/CS)
- ***It transcends the traditional boundaries*** between different wavelength regimes, agency domains, etc.
- It has an ***unusually broad range of constituents*** and interfaces
- It is inherently ***multidisciplinary***

Vobs standards and infrastructure

Data mining level



Data Sources
Images
Catalogs
Time series
Simulations

Information Extracted
Shapes & Patterns
Science Metadata
Distributions & Frequencies
Model Parameters

Unsupervised methods
Associative networks
Clustering
Principal components
Self-Organizing Maps

KDD Tools

Supervised methods
Neural Networks
Bayesian Networks
Support Vector Machines

New Knowledge or causal connections between physical events within the science domain

GRID, CLOUD, HPC Environment layer

Data knowledge layer

Computer Science layer

KDD layer

What is DAME

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

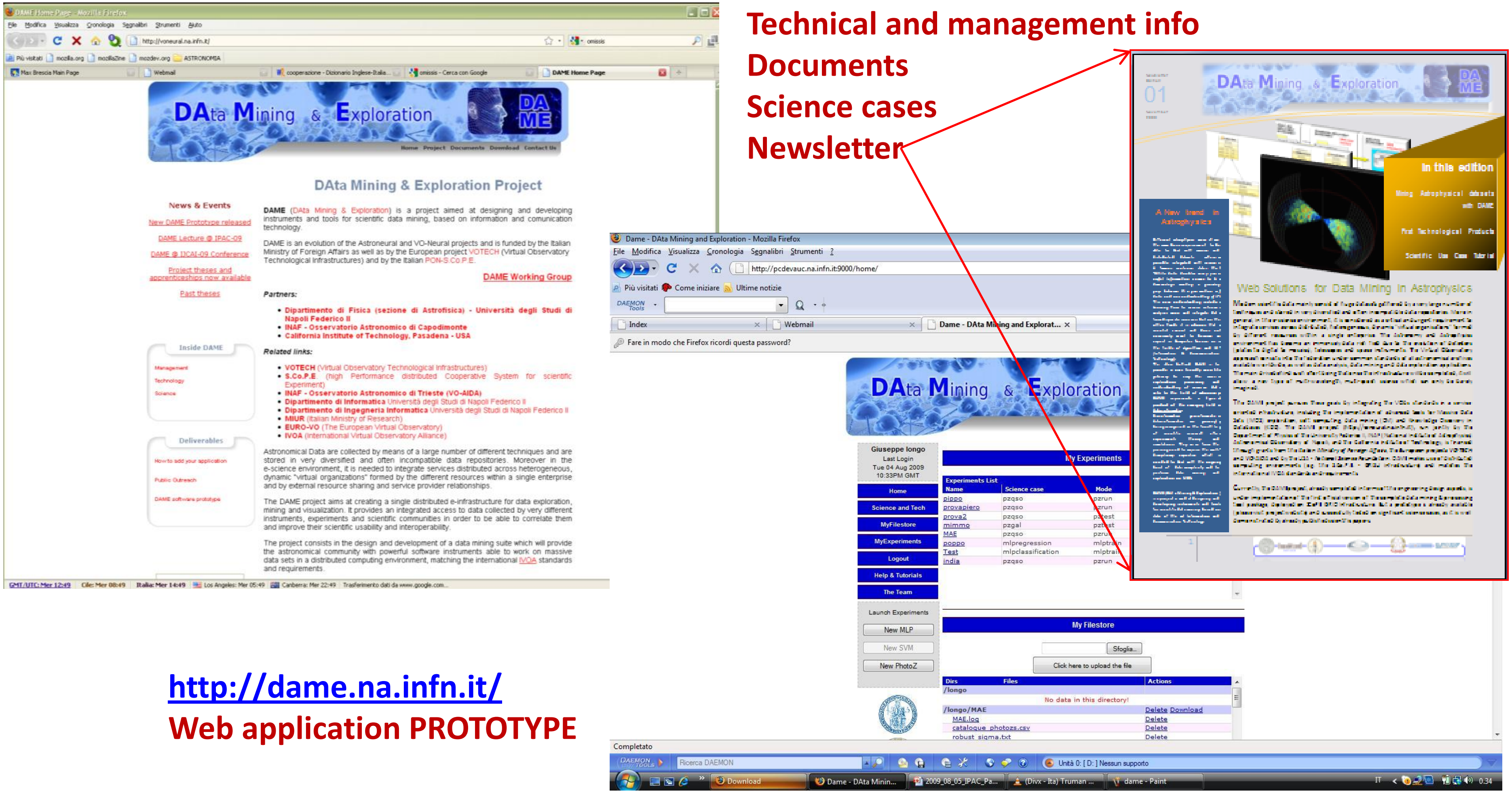
http://voneural.na.infn.it/

Technical and management info

Documents

Science cases

Newsletter



DAME Home Page - Mozilla Firefox
http://voneural.na.infn.it/
DAta Mining & Exploration
Home Project Documents Download Contact Us

DAta Mining & Exploration Project

News & Events

New DAME Prototype released

DAME Lecture @ IPAC-09

DAME @ IJCAI-09 Conference

Project theses and apprenticeships now available

Past theses

Inside DAME

Management

Technology

Science

Deliverables

How to add your application

Public Outreach

DAME software prototype

DAME (DAta Mining & Exploration) is a project aimed at designing and developing instruments and tools for scientific data mining, based on information and comunication technology.

DAME is an evolution of the Astroneural and VO-Neural projects and is funded by the Italian Ministry of Foreign Affairs as well as by the European project VOTECH (Virtual Observatory Technological Infrastructures) and by the Italian PON-S.Co.P.E.

DAME Working Group

Partners:

- Dipartimento di Fisica (sezione di Astrofisica) - Università degli Studi di Napoli Federico II
- INAF - Osservatorio Astronomico di Capodimonte
- California Institute of Technology, Pasadena - USA

Related links:

- VOTECH (Virtual Observatory Technological Infrastructures)
- S.Co.P.E. (high Performance distributed Cooperative System for scientific Experiment)
- INAF - Osservatorio Astronomico di Trieste (VO-AIDA)
- Dipartimento di Informatica Università degli Studi di Napoli Federico II
- Dipartimento di Ingegneria Informatica Università degli Studi di Napoli Federico II
- MIUR (Italian Ministry of Research)
- EURO-VO (The European Virtual Observatory)
- IVOA (International Virtual Observatory Alliance)

Astronomical Data are collected by means of a large number of different techniques and are stored in very diversified and often incompatible data repositories. Moreover in the e-science environment, it is needed to integrate services distributed across heterogeneous, dynamic "virtual organizations" formed by the different resources within a single enterprise and by external resource sharing and service provider relationships.

The DAME project aims at creating a single distributed e-infrastructure for data exploration, mining and visualization. It provides an integrated access to data collected by very different instruments, experiments and scientific communities in order to be able to correlate them and improve their scientific usability and interoperability.

The project consists in the design and development of a data mining suite which will provide the astronomical community with powerful software instruments able to work on massive data sets in a distributed computing environment, matching the international IVOA standards and requirements.

GMT/UTC: Mer 12:49 Cile: Mer 08:49 Italia: Mer 14:49 Los Angeles: Mer 05:49 Canberra: Mer 22:49 Trasferimento dati da www.google.com...
Dame - DAta Mining and Exploration - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti ?
http://pcdevauc.na.infn.it:9000/home/
Più visitati Come iniziare Ultime notizie
Index Webmail Dame - DAta Mining and Explorat...
Fare in modo che Firefox ricordi questa password?
DAta Mining & Exploration

Giuseppe longo
Last Login
Tue 04 Aug 2009
10:33PM GMT

My Experiments

Experiments List

Name	Science case	Mode
pippo	pzqso	pzrun
provapiero	pzqso	pzrun
prova2	pzqso	pztest
mimmo	pzgal	pztest
MAE	pzqso	pzrun
poppo	mlpregression	mlptrain
Test	mlpclassification	mlptrain
india	pzqso	pzrun

My Filestore

Dirs Files	Actions
/longo	
No data in this directory!	
/longo/MAE	Delete Download
MAE.log	Delete
catalogue_photozs.csv	Delete
robust_sigma.txt	Delete

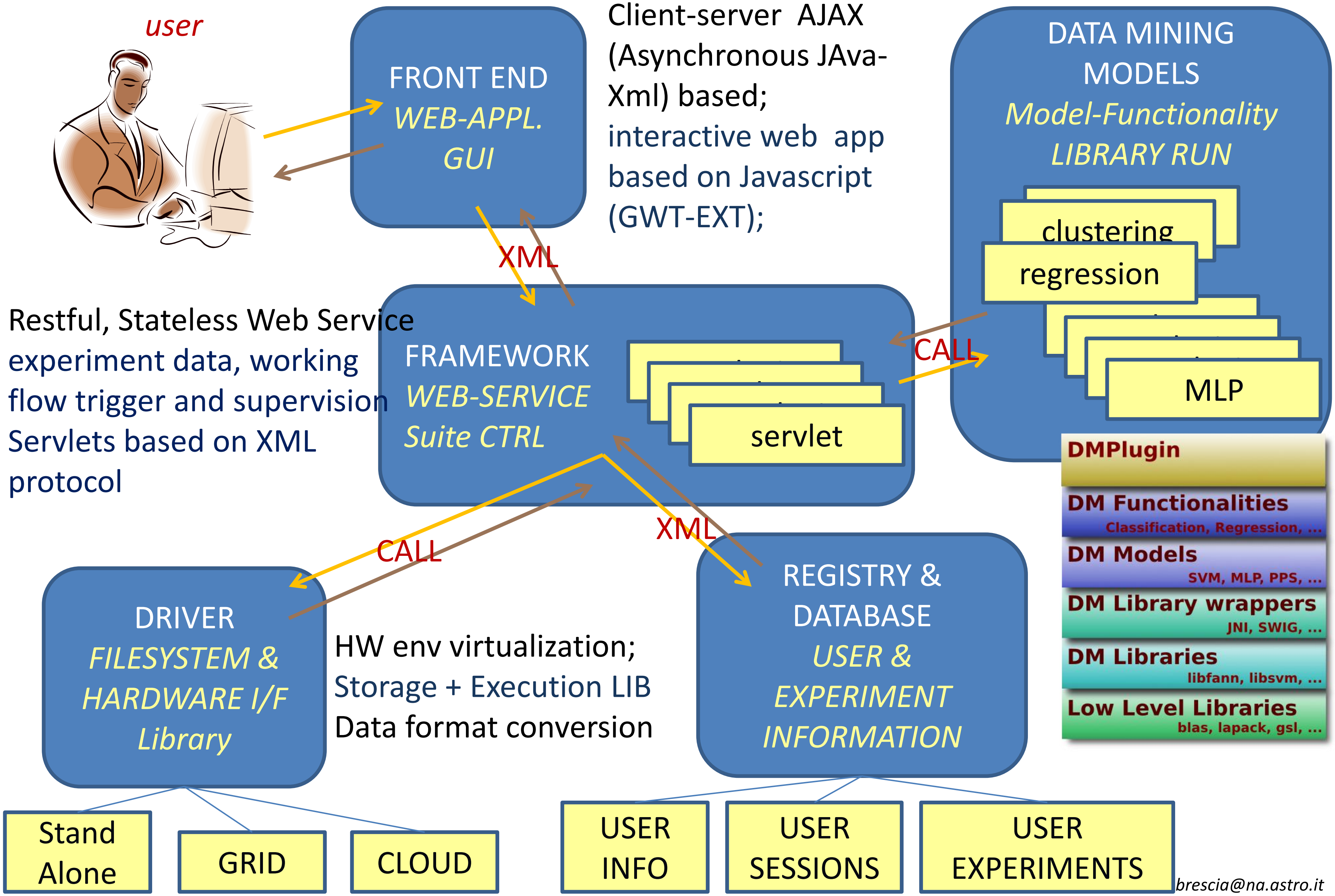
Completato
Ricerca DAEMON
Unità 0: [D:] Nessun supporto
Download
Dame - DAta Minin...
2009_08_05_IPAC_Pa...
(Divx - Ita) Truman ...
dame - Paint
0:34
DAta Mining & Exploration
01
In this edition
Mining Astrophysical datasets with DAME
First Technological Products
Scientific Use Case Tutorial
A New trend in Astrophysics
Web Solutions for Data Mining in Astrophysics

http://dame.na.infn.it/

Web application PROTOTYPE

brescia@na.astro.it

The DAME architecture

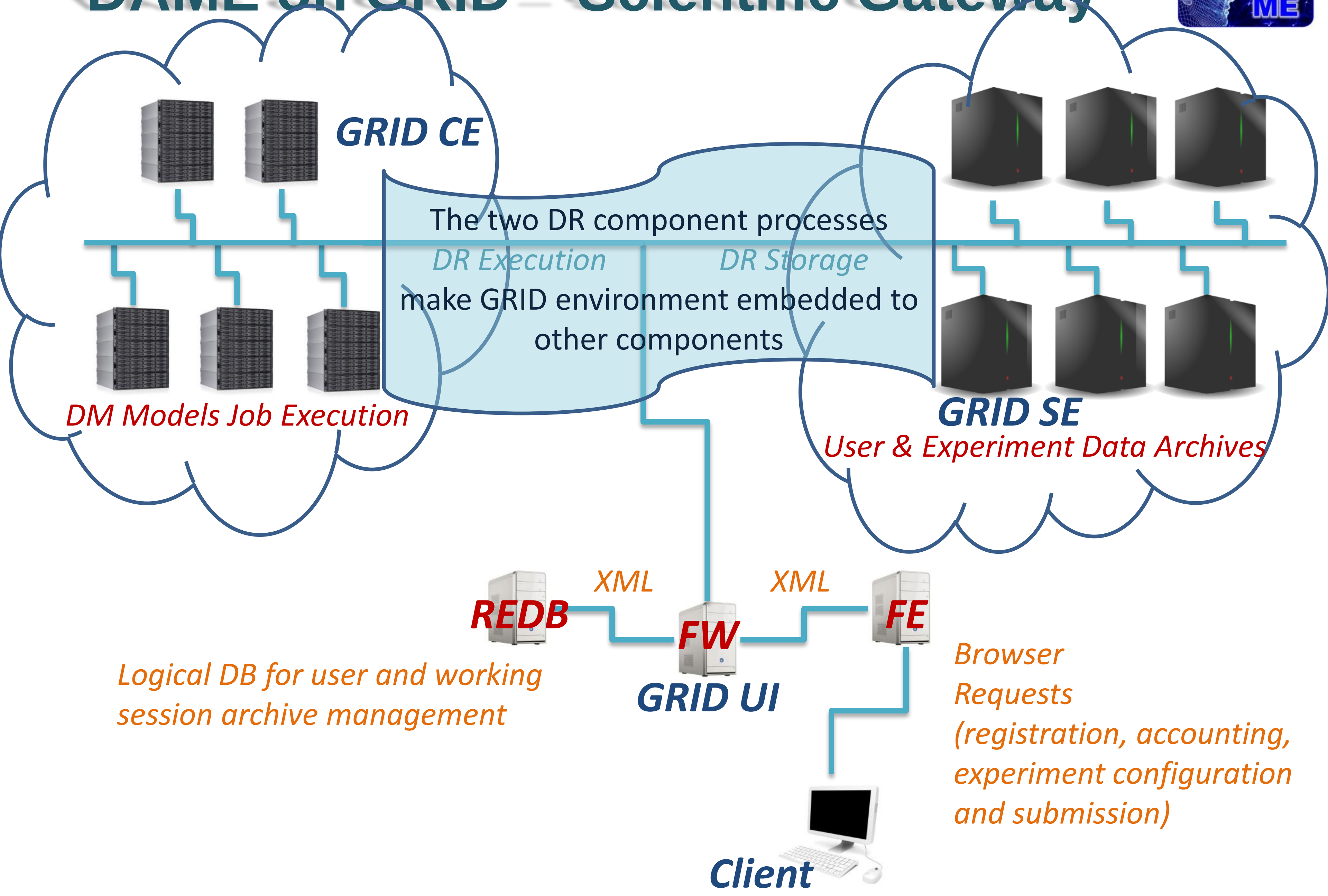


Client-server AJAX (Asynchronous JAva-Xml) based; interactive web app based on Javascript (GWT-EXT);

Restful, Stateless Web Service experiment data, working flow trigger and supervision Servlets based on XML protocol

HW env virtualization; Storage + Execution LIB Data format conversion

DAME on GRID – Scientific Gateway



The two DR component processes

DR Execution *DR Storage*

make GRID environment embedded to other components

DM Models Job Execution

User & Experiment Data Archives

Logical DB for user and working session archive management

Browser Requests (registration, accounting, experiment configuration and submission)

How to spread the word within the community

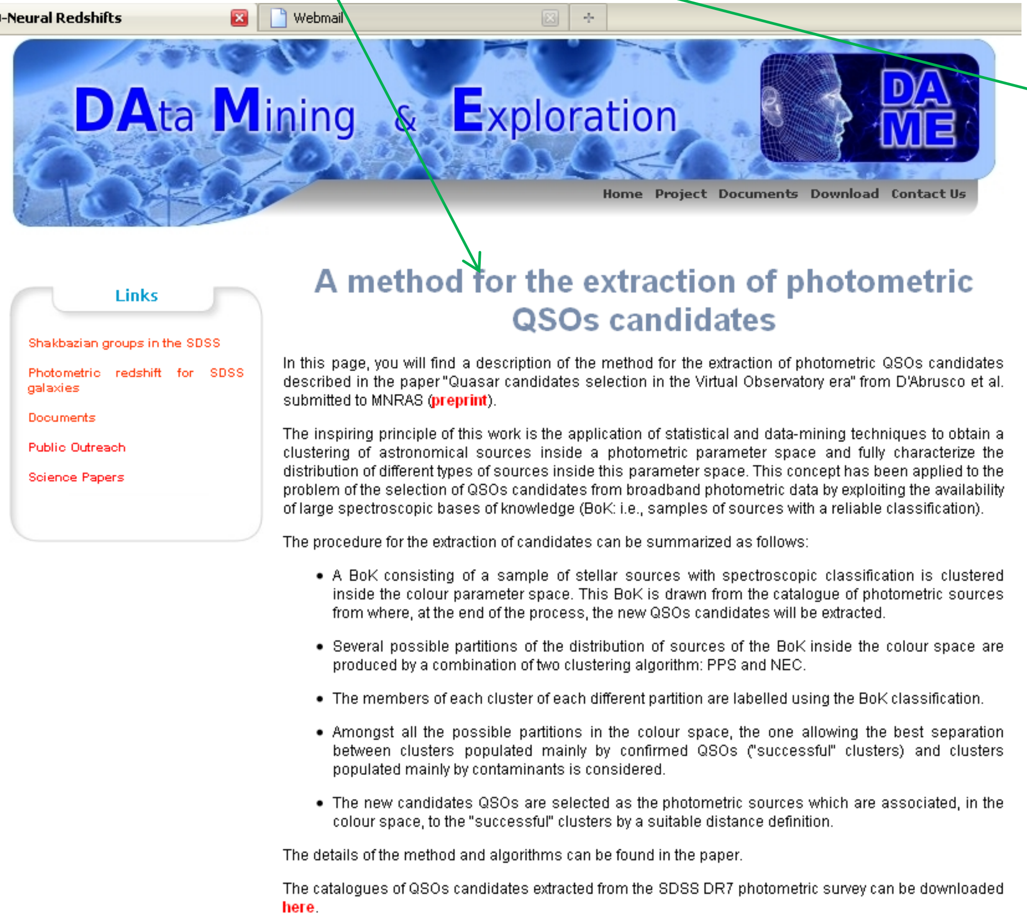
In parallel with the Suite R&D process, all data processing algorithms (foreseen to be plugged in) have been massively tested on real astrophysical cases.

http://voneural.na.infn.it/

Technical and management info

Documents

Science cases



0-Neural Redshifts
Webmail
DAta Mining & Exploration
DA ME
Home Project Documents Download Contact Us
Links
Shakbazian groups in the SDSS
Photometric redshift for SDSS galaxies
Documents
Public Outreach
Science Papers

A method for the extraction of photometric QSOs candidates

In this page, you will find a description of the method for the extraction of photometric QSOs candidates described in the paper "Quasar candidates selection in the Virtual Observatory era" from D'Abrusco et al. submitted to MNRAS (**preprint**).

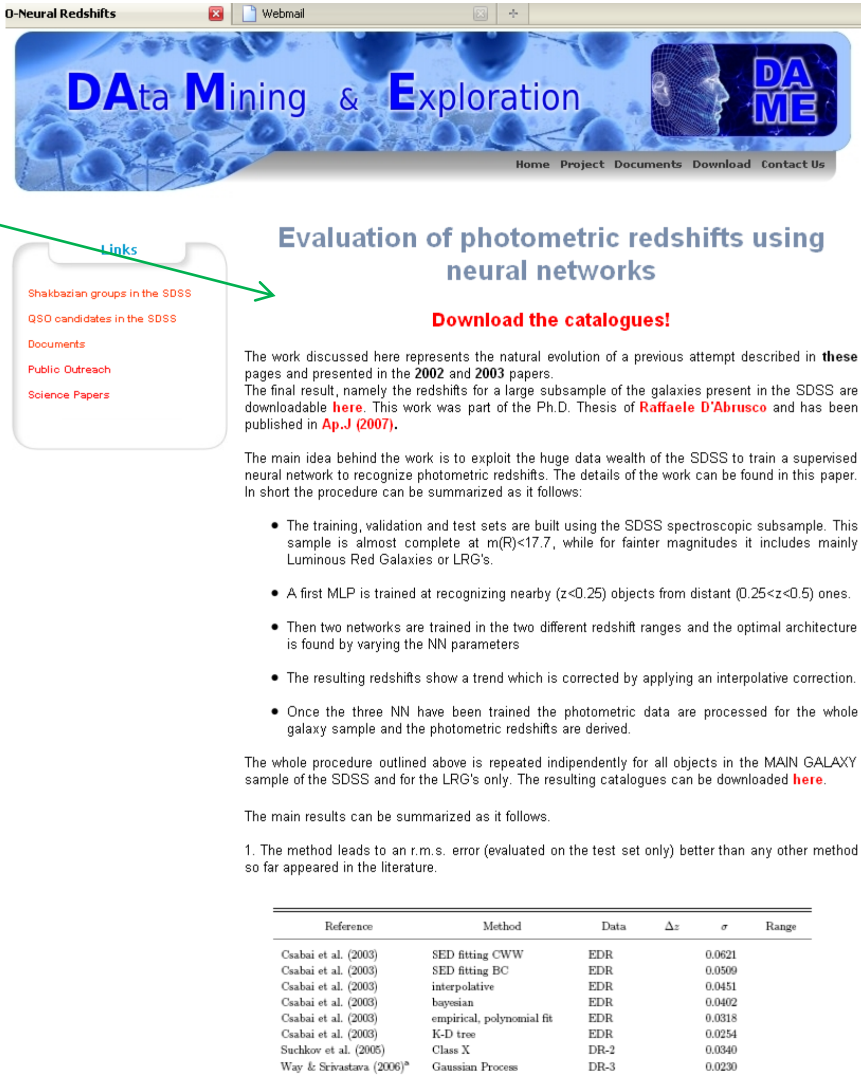
The inspiring principle of this work is the application of statistical and data-mining techniques to obtain a clustering of astronomical sources inside a photometric parameter space and fully characterize the distribution of different types of sources inside this parameter space. This concept has been applied to the problem of the selection of QSOs candidates from broadband photometric data by exploiting the availability of large spectroscopic bases of knowledge (BoK: i.e., samples of sources with a reliable classification).

The procedure for the extraction of candidates can be summarized as follows:

- A BoK consisting of a sample of stellar sources with spectroscopic classification is clustered inside the colour parameter space. This BoK is drawn from the catalogue of photometric sources from where, at the end of the process, the new QSOs candidates will be extracted.
- Several possible partitions of the distribution of sources of the BoK inside the colour space are produced by a combination of two clustering algorithm: PPS and NEC.
- The members of each cluster of each different partition are labelled using the BoK classification.
- Amongst all the possible partitions in the colour space, the one allowing the best separation between clusters populated mainly by confirmed QSOs ("successful" clusters) and clusters populated mainly by contaminants is considered.
- The new candidates QSOs are selected as the photometric sources which are associated, in the colour space, to the "successful" clusters by a suitable distance definition.

The details of the method and algorithms can be found in the paper.

The catalogues of QSOs candidates extracted from the SDSS DR7 photometric survey can be downloaded **here**.



0-Neural Redshifts
Webmail
DAta Mining & Exploration
DA ME
Home Project Documents Download Contact Us
Links
Shakbazian groups in the SDSS
QSO candidates in the SDSS
Documents
Public Outreach
Science Papers

Evaluation of photometric redshifts using neural networks

Download the catalogues!

The work discussed here represents the natural evolution of a previous attempt described in **these** pages and presented in the **2002** and **2003** papers.
The final result, namely the redshifts for a large subsample of the galaxies present in the SDSS are downloadable **here**. This work was part of the Ph.D. Thesis of **Raffaele D'Abrusco** and has been published in **Ap.J (2007).**

The main idea behind the work is to exploit the huge data wealth of the SDSS to train a supervised neural network to recognize photometric redshifts. The details of the work can be found in this paper. In short the procedure can be summarized as it follows:

- The training, validation and test sets are built using the SDSS spectroscopic subsample. This sample is almost complete at m(R)<17.7, while for fainter magnitudes it includes mainly Luminous Red Galaxies or LRG's.
- A first MLP is trained at recognizing nearby (z<0.25) objects from distant (0.25<z<0.5) ones.
- Then two networks are trained in the two different redshift ranges and the optimal architecture is found by varying the NN parameters
- The resulting redshifts show a trend which is corrected by applying an interpolative correction.
- Once the three NN have been trained the photometric data are processed for the whole galaxy sample and the photometric redshifts are derived.

The whole procedure outlined above is repeated indipendently for all objects in the MAIN GALAXY sample of the SDSS and for the LRG's only. The resulting catalogues can be downloaded **here**.

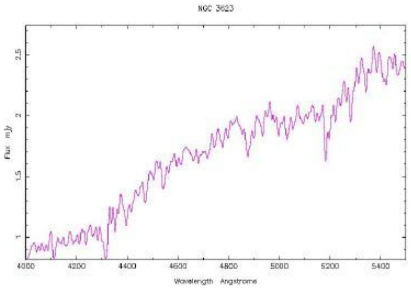
The main results can be summarized as it follows.

1. The method leads to an r.m.s. error (evaluated on the test set only) better than any other method so far appeared in the literature.

Reference	Method	Data	Δz	σ	Range
Csabai et al. (2003)	SED fitting CWW	EDR		0.0621	
Csabai et al. (2003)	SED fitting BC	EDR		0.0509	
Csabai et al. (2003)	interpolative	EDR		0.0451	
Csabai et al. (2003)	bayesian	EDR		0.0402	
Csabai et al. (2003)	empirical, polynomial fit	EDR		0.0318	
Csabai et al. (2003)	K-D tree	EDR		0.0254	
Suchkov et al. (2005)	Class X	DR-2		0.0340	
Way & Srivastava (2006)[a]	Gaussian Process	DR-3		0.0230	

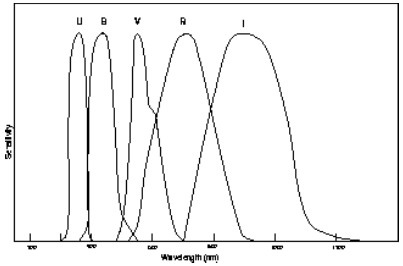
An EXAMPLE: photometric redshifts of SDSS galaxies

$$z \times c \equiv \frac{\Delta\lambda}{\lambda_0}$$



Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

$$m_U = -2.5\log_{10}\frac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u$$

$$m_B = -2.5\log_{10}\frac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B$$

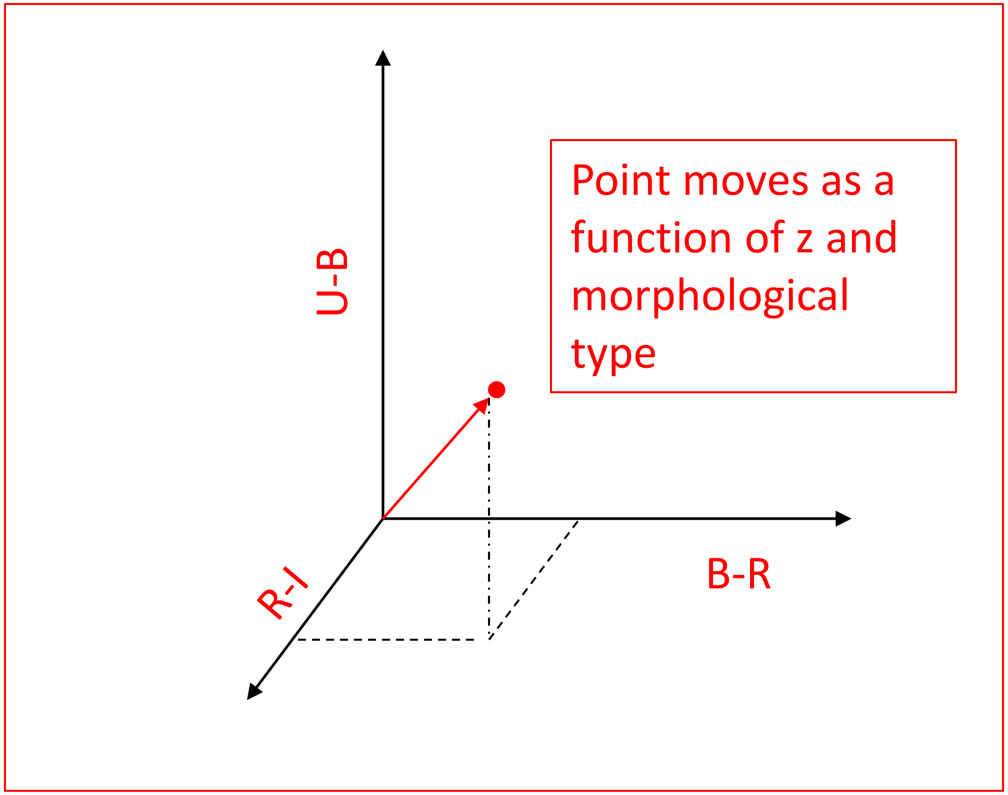
Etc…

Color indexes

$$U - B \equiv m_U - m_B$$

$$B - R \equiv m_B - m_R$$

etc.



Point moves as a function of z and morphological type

Phot-z are an inverse problem

Photometric redshifts: the DM approach

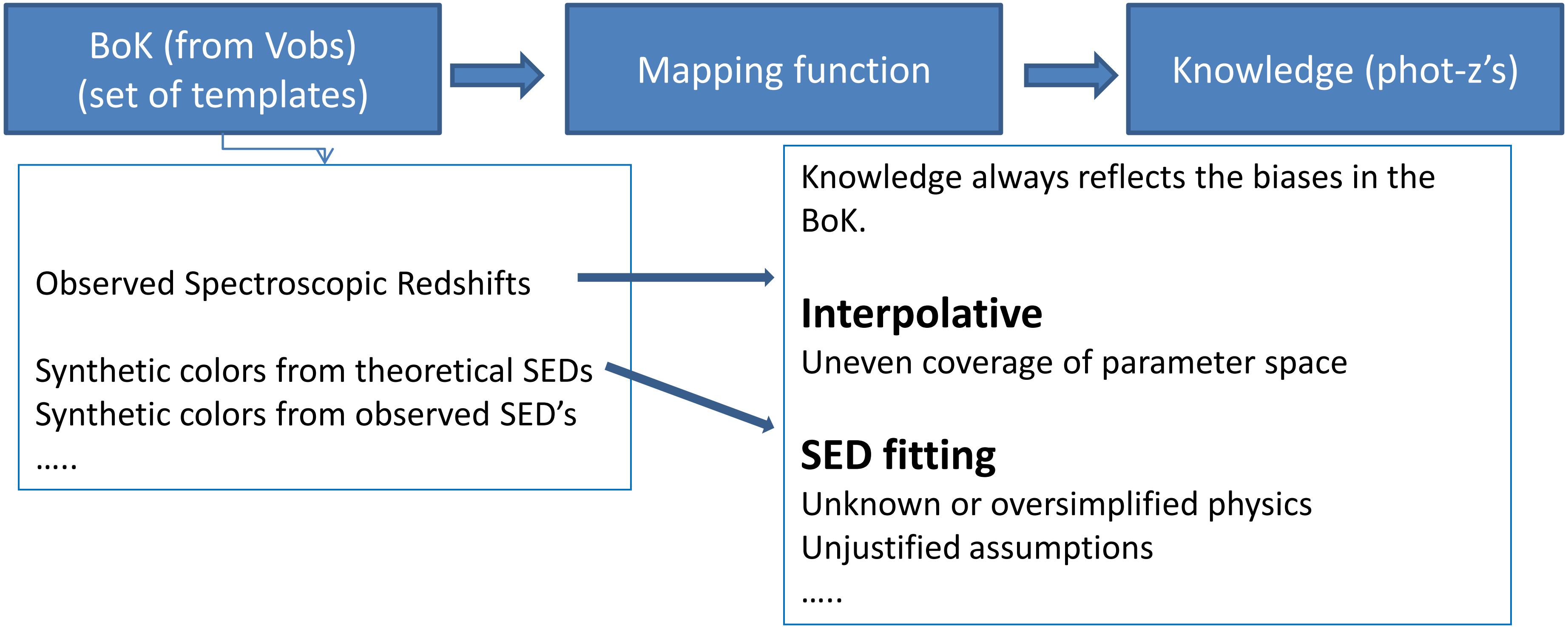
Photometric redshifts are always a function approximation hence a DM problem:

$$\mathbf{X} \equiv \{x_1, x_2, x_3, \ldots x_N\} \ \textbf{input vectors}$$

$$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\} \ \textbf{target vectors} \ M \ll N$$

$$\textbf{find} \quad \hat{f}: \ \hat{\mathbf{Y}} = \hat{f}(\mathbf{X}) \ \textbf{is a good approximation of } \mathbf{Y}$$

BoK = Base of Knowledge



BoK (from Vobs)
(set of templates)

Mapping function

Knowledge (phot-z's)

Observed Spectroscopic Redshifts

Synthetic colors from theoretical SEDs
Synthetic colors from observed SED's
…..

Knowledge always reflects the biases in the BoK.

Interpolative

Uneven coverage of parameter space

SED fitting

Unknown or oversimplified physics
Unjustified assumptions
…..

Data used in the science case:

SDSS: 10^8 galaxies in 5 optical bands;

BoK: spectroscopic redshifts for 10^6 galaxies → **Spectroscopic BoK**

BoK: incomplete and **biased**.

UKIDDS: overlap with SDSS

3 infrared bands.

GALEX: overlap with SDSS

Ultraviolet bands;

SDSS

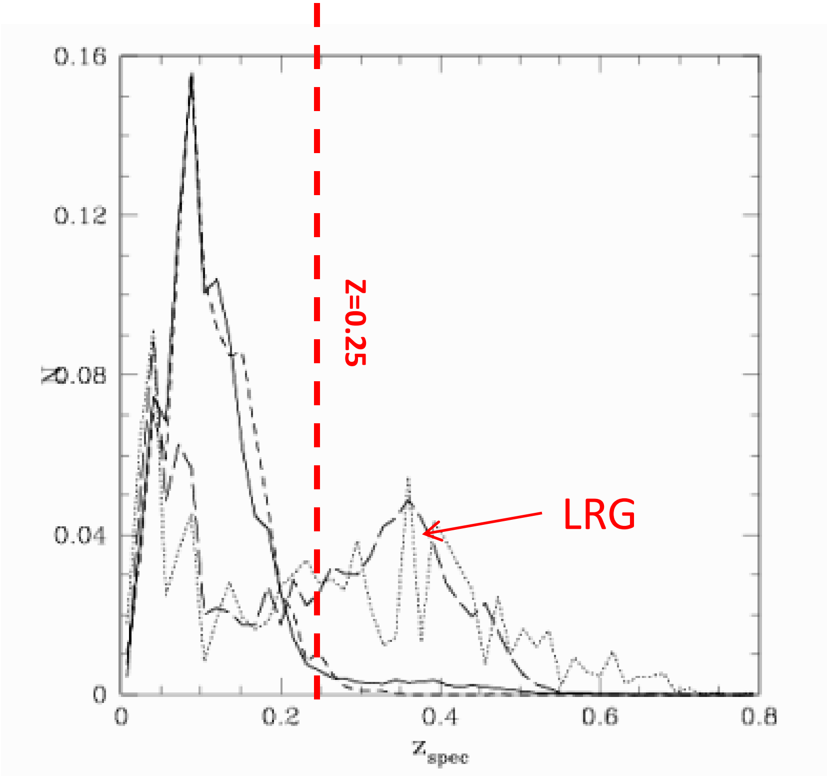
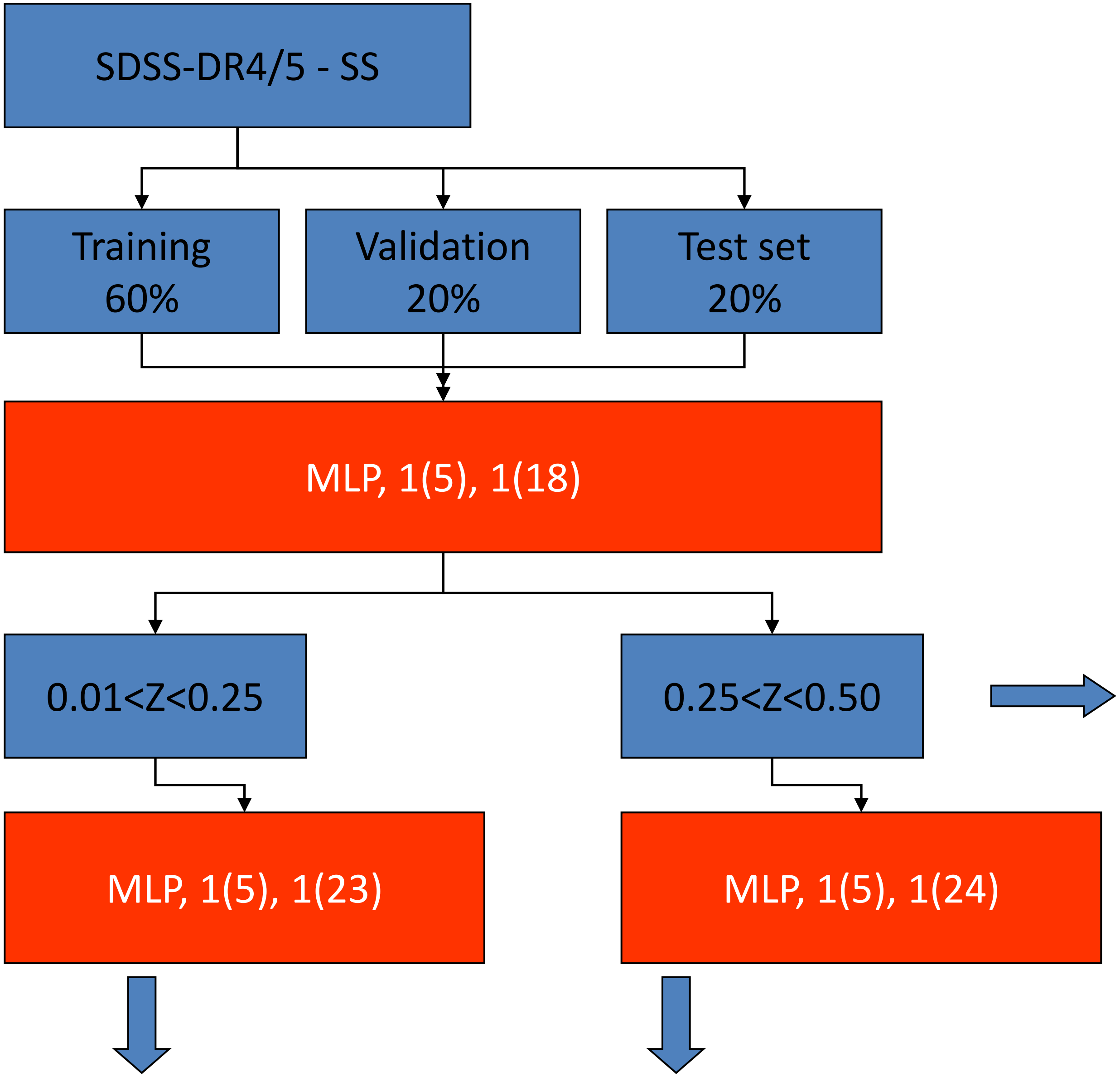


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.



SDSS-DR4/5 - SS

Training 60%

Validation 20%

Test set 20%

MLP, 1(5), 1(18)

0.01<Z<0.25

0.25<Z<0.50

MLP, 1(5), 1(23)

MLP, 1(5), 1(24)

99.6 % accuracy

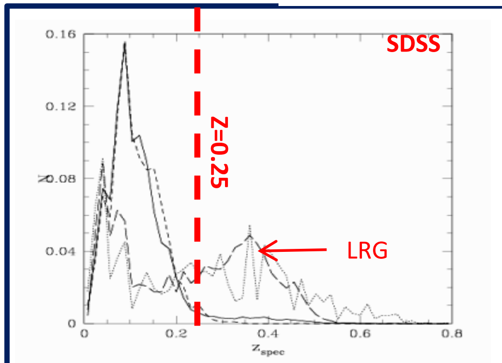
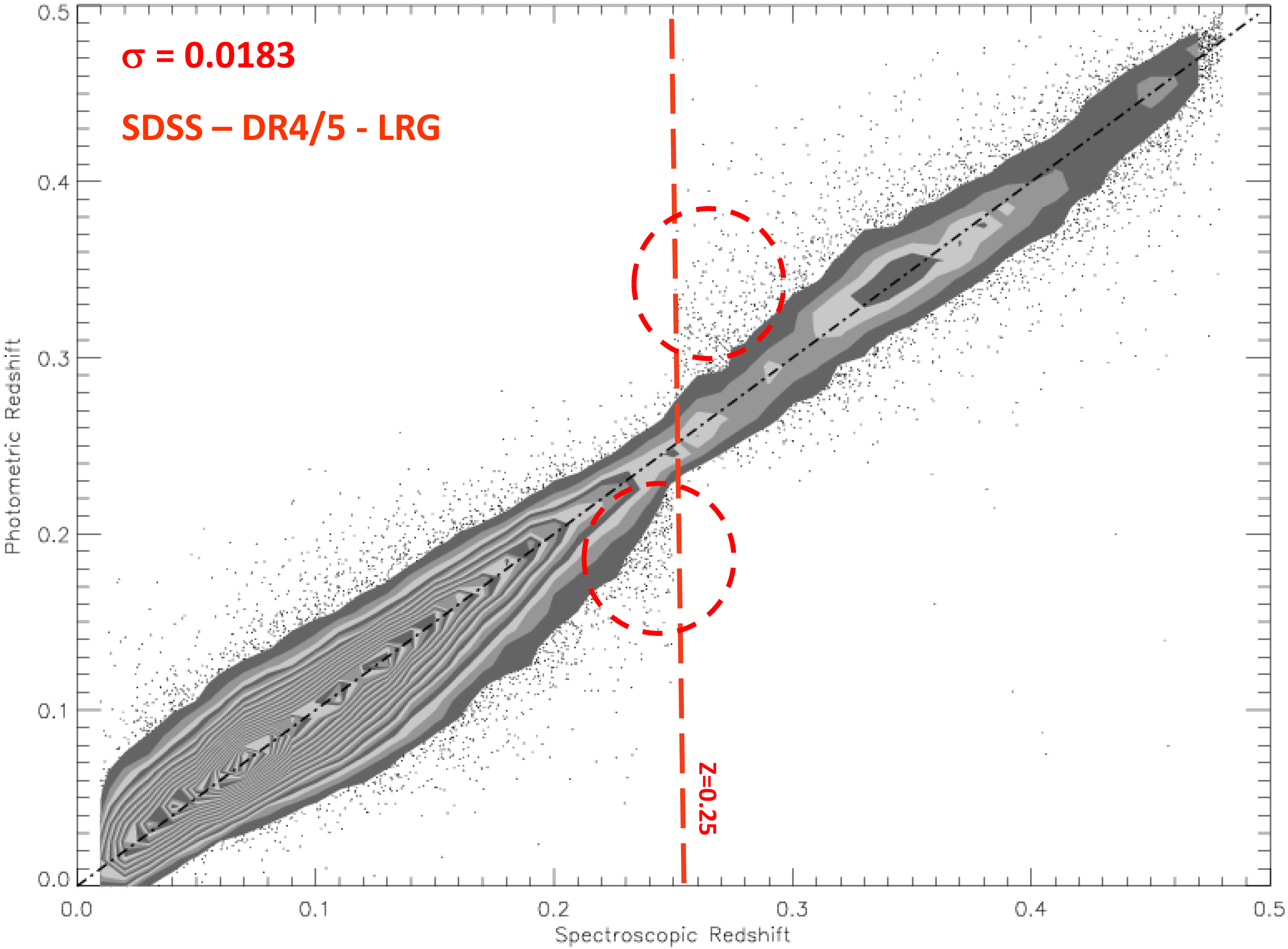


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.

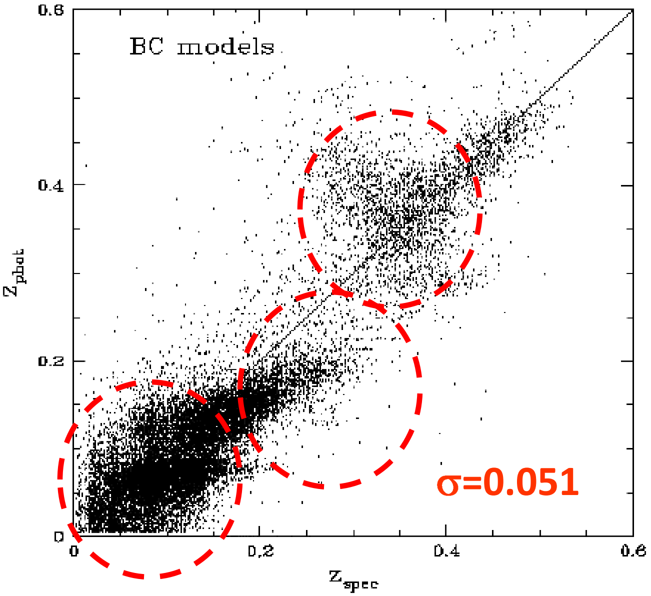


σ = 0.0183

SDSS – DR4/5 - LRG

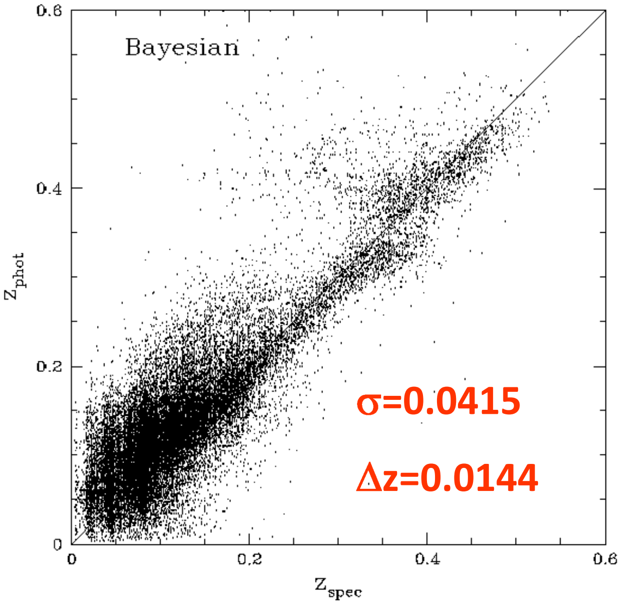
D'Abrusco et al. 2007

Traditional approaches: interpolation based on BoK



BoK from Spectral Energy Distribution (SED) fitting

Templates from synthetic colors obtained from theoretical SED's
Mapping function from simple interpolation

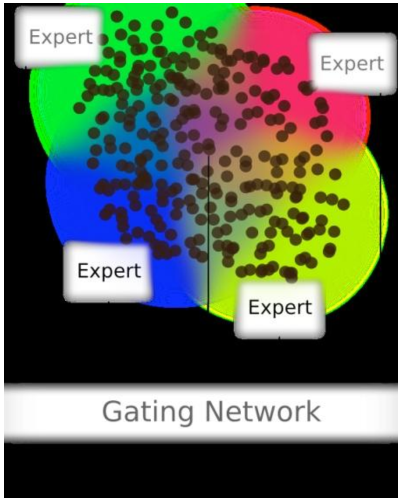


BoK from Spectral Energy Distribution (SED) fitting Interpolative

Templates from synthetic colors obtained from theoretical SED's
Mapping function from Bayesian inference

What do we learn if the BoK is biased:

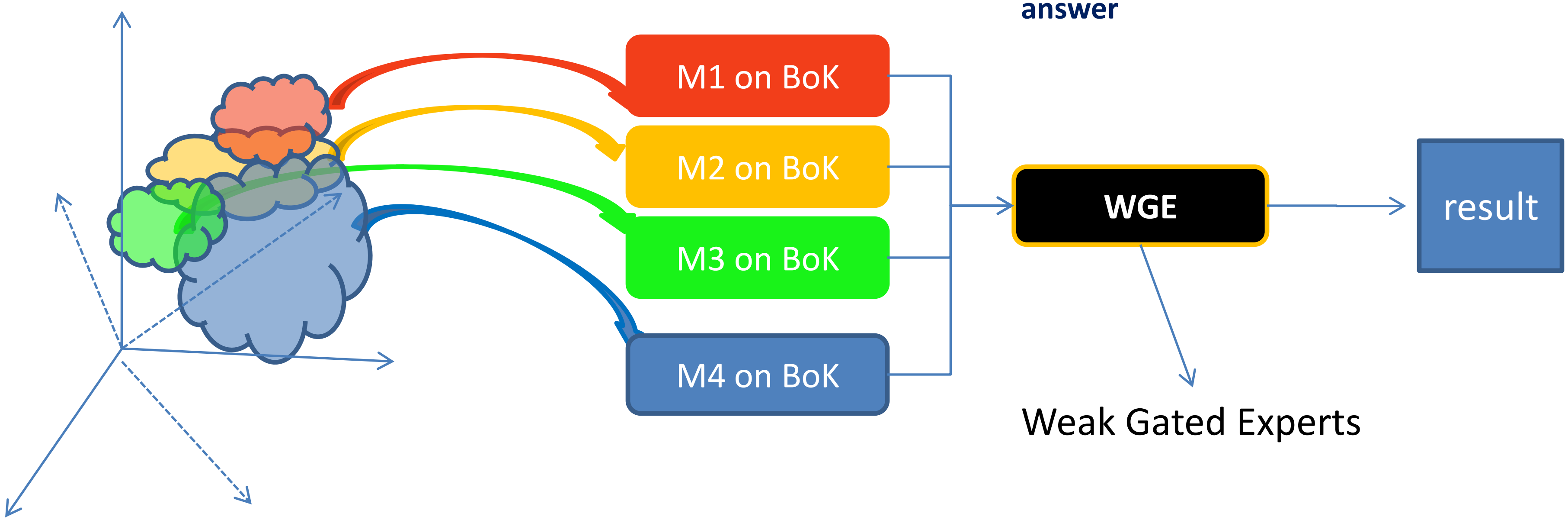
- At high z LRG dominate and interpolative methods are not capable to "generalize" rules
- An unique method optimizes its performances on the parts of the parameter space which are best covered in the BoK



Step 1:
unsupervised clustering in parameter space

Step 2:
supervised training of different NN for each cluster

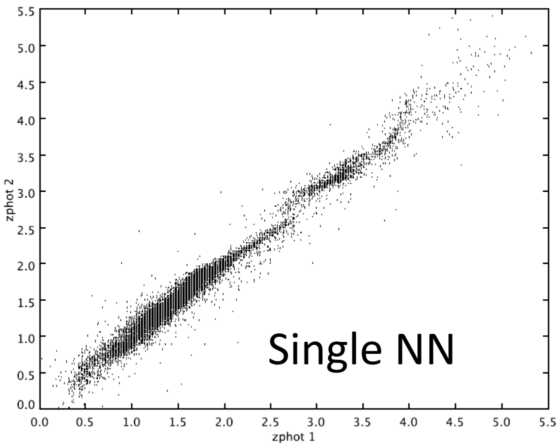
Step 3:
output of all NN go to WGE which learns the correct answer

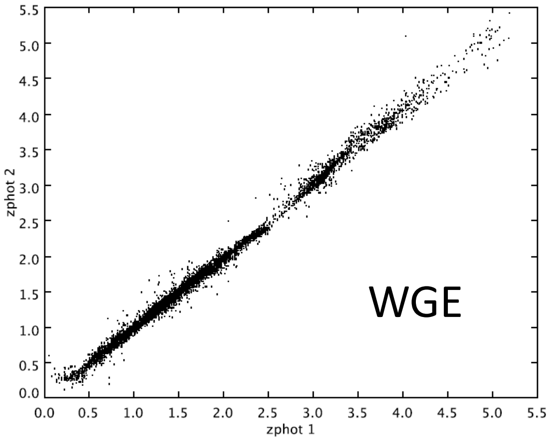


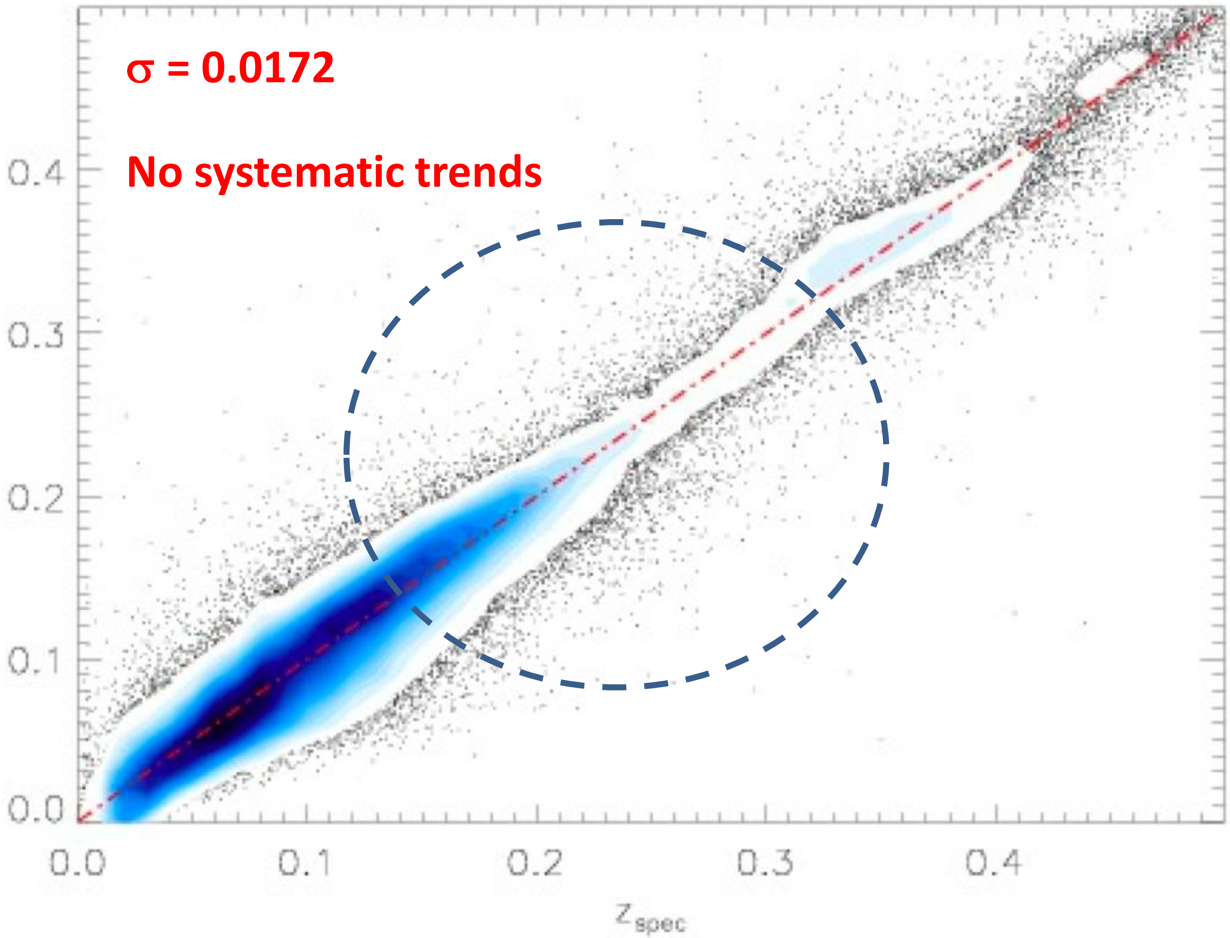
Weak Gated Experts

Laurino et al. 2009a,2009b

Laurino et al. 2009a,2009b


Single NN

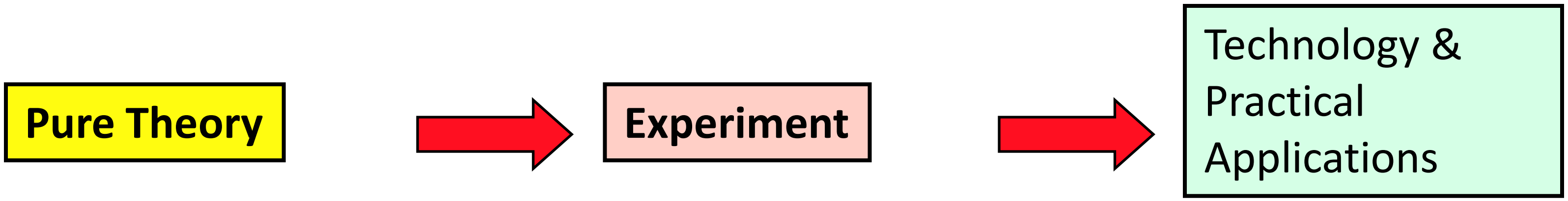

WGE


σ = 0.0172

No systematic trends

Conclusion I. I.T. is changing the methodology of science

The old traditional, “Platonistic” view:

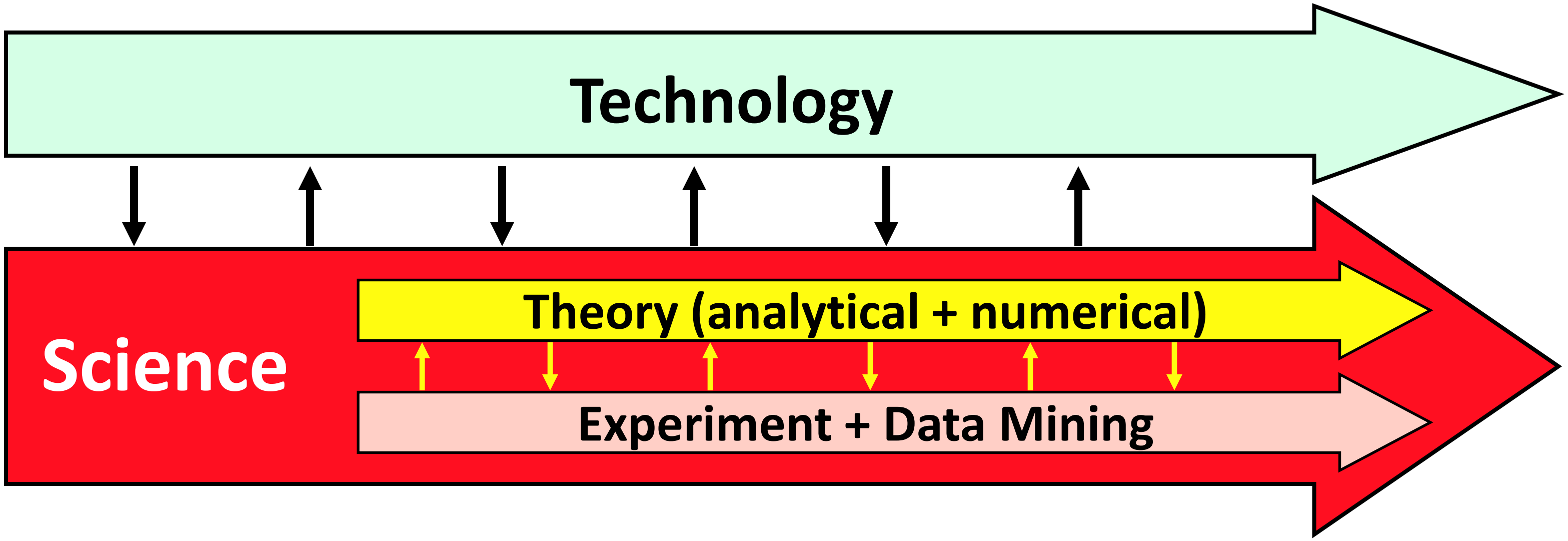


Pure Theory

Experiment

Technology & Practical Applications

The modern and realistic view when dealing with complex data sets:



Technology

Science

Theory (analytical + numerical)

Experiment + Data Mining

This synergy is stronger than ever and growing

Open problems to be addressed soon:

- Scalability
- Robustness
- Reliability
- Choice of optimal models
- Connection: semantics -> Ontologies -> Bases of knowledge
- Visualization

Algorithms

Restricted choice of algorithms (MLPs, SVM, Kernel methods, Genetic algoritms (few models), K Means, PPS, SOM …)

Astronomers know little statistics, forget about SPR, DM, etc… Just a few astronomers go beyond the introductory chapters of the Bishop.

Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	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	-----	
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	----	

Limited number of problems due to limited number of reliable BoKs

Bases of knowledge

(set of well known templates for supervised (training) or unsupervised (labeling) methods

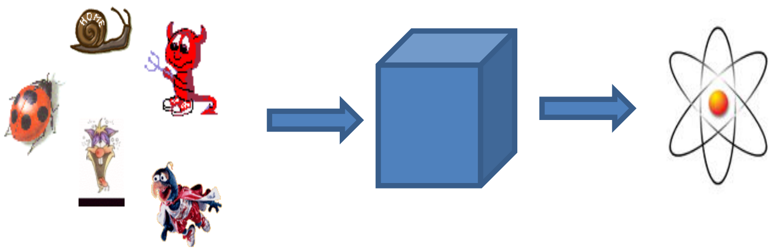
So far

- 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)

Bases of knowledge need to be built automatically from Vobs Data repositories

Community believes AI/DM methods are black boxes

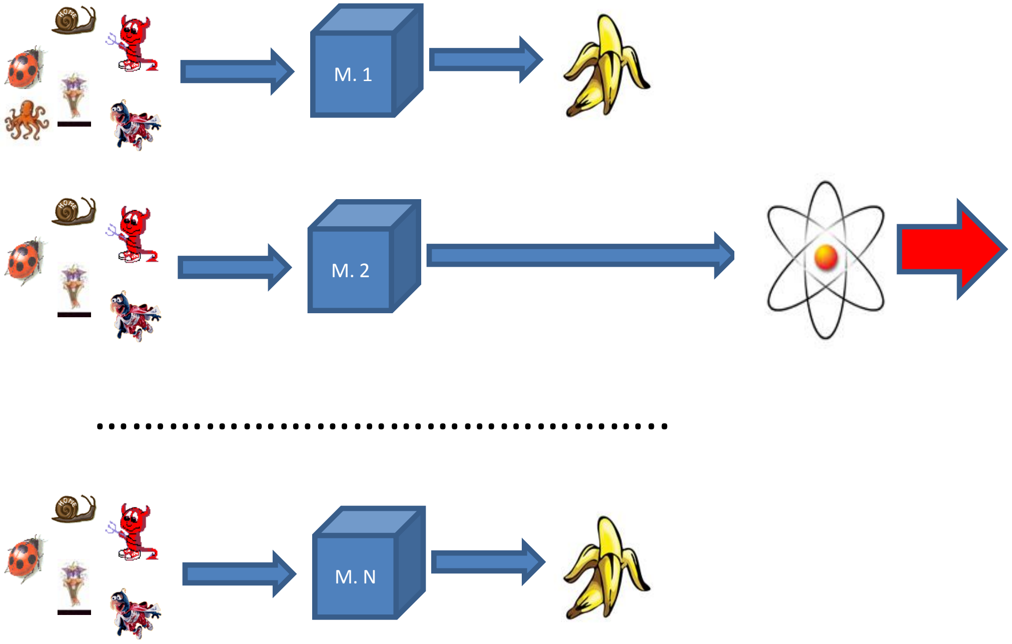
You feed in something, and obtain patters, trends, i.e. knowledge….



Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is not willing to make an effort to learn them ….

The r.m.s astronomer doesn’t want to become a computer scientist or a mathematician (large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud no personal certificates, no deep understanding of the DM tool etc.)



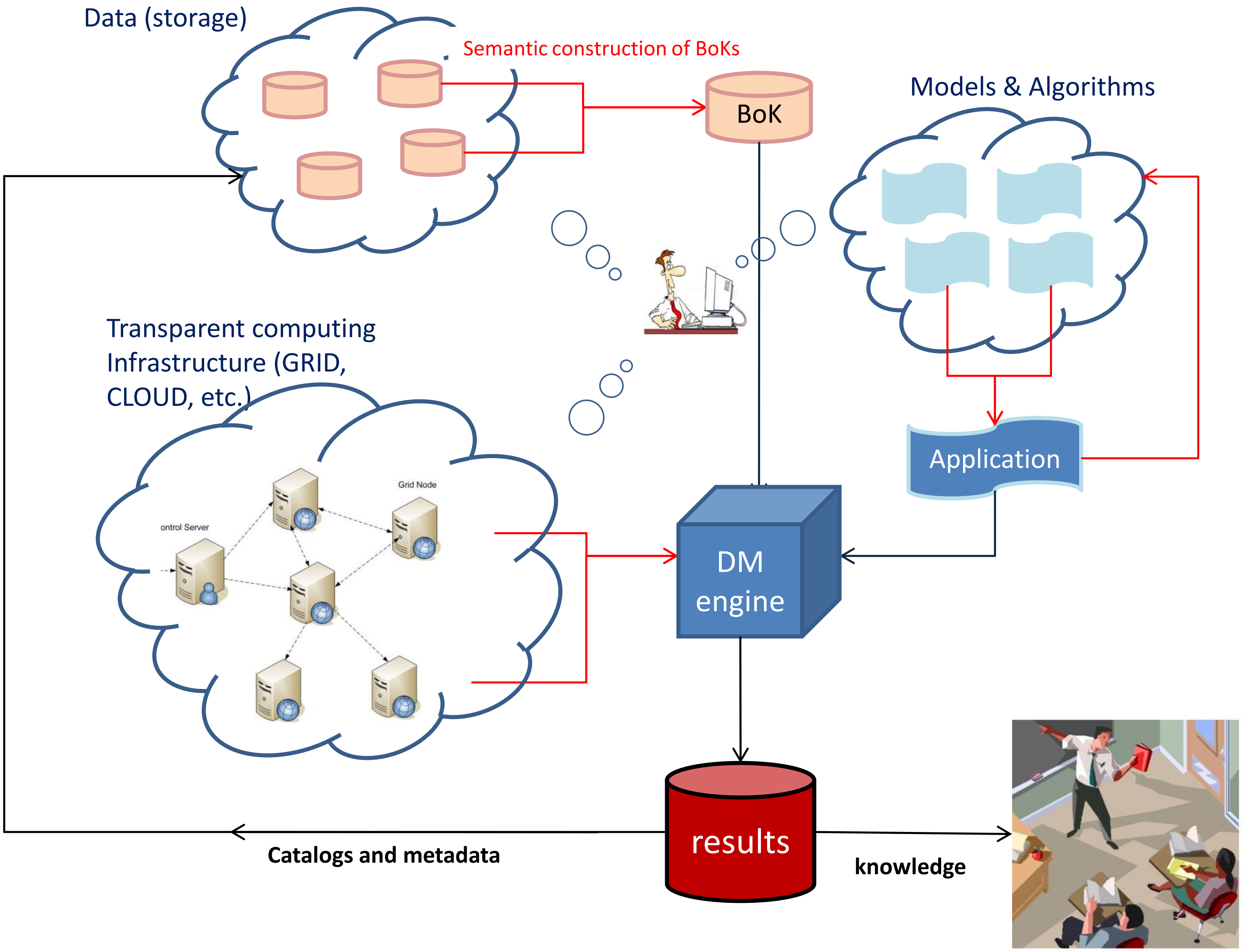
Formation of a new generation of experts

(…… suggesting the solutions)

AND/OR

Implementation of a second generation of tools

A break-down of an effective DM process



Data (storage)

Semantic construction of BoKs

BoK

Models & Algorithms

Transparent computing Infrastructure (GRID, CLOUD, etc.)

Application

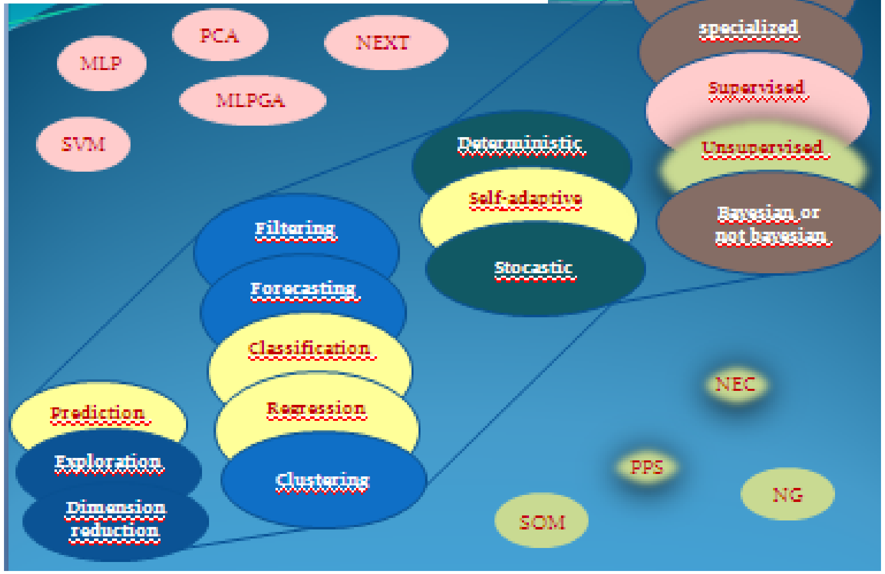
DM engine

results

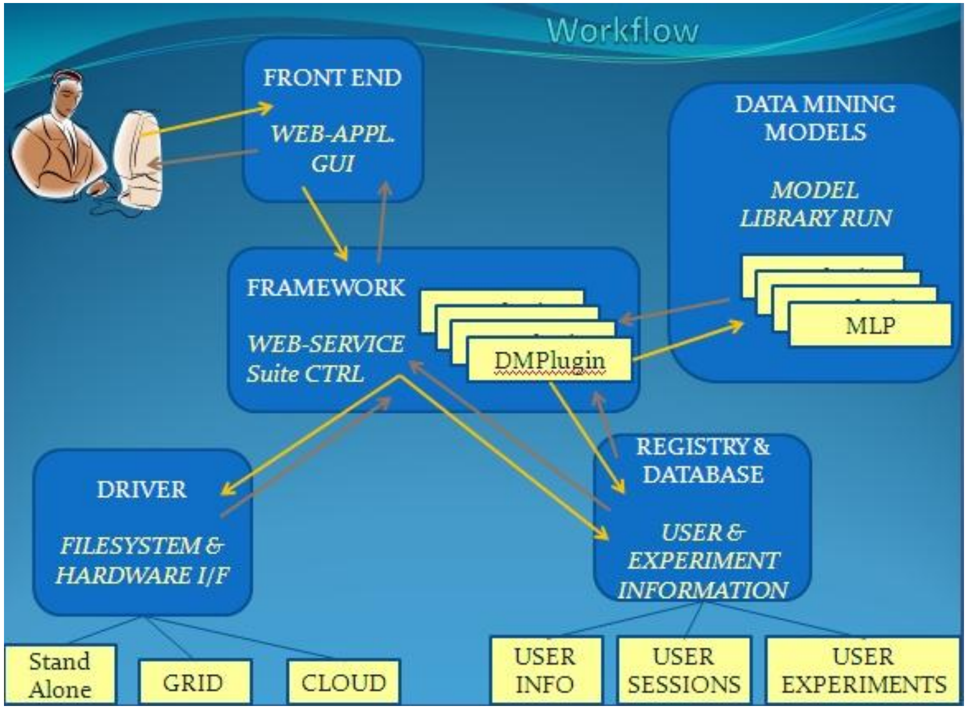
Catalogs and metadata

knowledge

Functionality Taxonomy



MLP
PCA
NEXT
MLPGA
SVM
Expert systems
specialized
Supervised
Unsupervised
Bayesian or not bayesian
Deterministic
Self-adaptive
Stocastic
Filtering
Forecasting
Classification
Regression
Clustering
Prediction
Exploration
Dimension reduction
NEC
PPS
NG
SOM



Workflow

FRONT END
WEB-APPL. GUI

DATA MINING MODELS
MODEL LIBRARY RUN
MLP

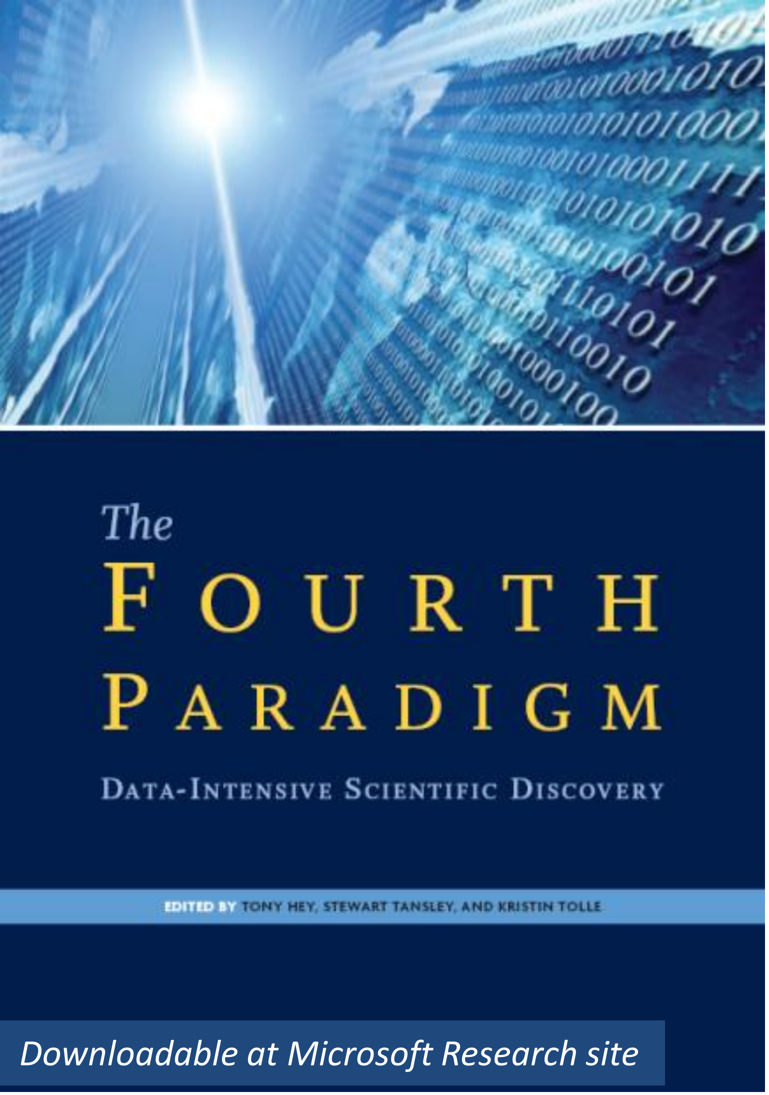
FRAMEWORK
WEB-SERVICE Suite CTRL
DMPlugin

DRIVER
FILESYSTEM & HARDWARE I/F

REGISTRY & DATABASE
USER & EXPERIMENT INFORMATION

Stand Alone
GRID
CLOUD
USER INFO
USER SESSIONS
USER EXPERIMENTS

X-informatics

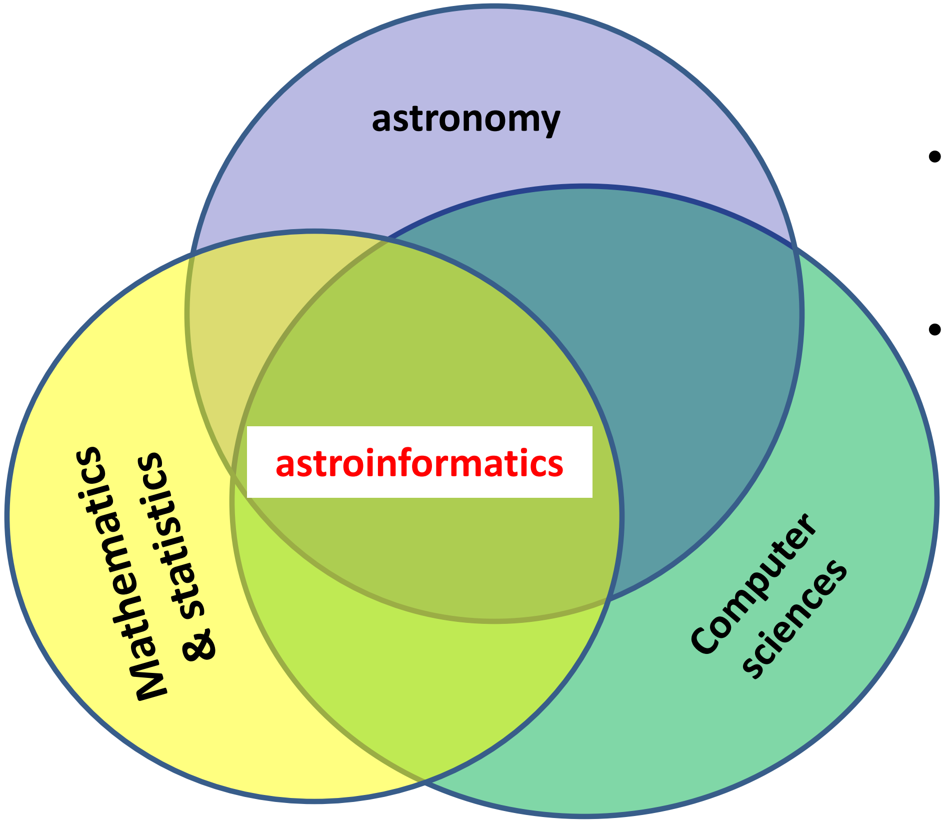

The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Downloadable at Microsoft Research site

The changing methodology of science

- Data Mining, computer science, etc. have become the “fourth leg of science” (besides theory, experimentation and simulations)
    - Sinergy between different worlds is required
    - Sociological issues to be solved (formation, infrastructures, and so on)

Experimental astronomy has become a three players game



- **astronomy**: problems, data, understanding of the data structure and biases
- **mathematics**: evaluation of the data, falsification/validation of theories/models, etc
- **computer science**: implementation of infrastructures, databases, middleware, scalable tools, etc

- **Astroinformatics: AAS n. 215, Washington, December 2009,** chairperson: K. Borne
- **Astroinformatics 2010: Caltech (USA) June 16-19 2010;**co-chairpersons: S.G. Djorgovski, G. Longo
- **Astroinformatics 2011: UNINA – Sorrento,** co-chairpersons: S.G. Djorgovski, G. Longo