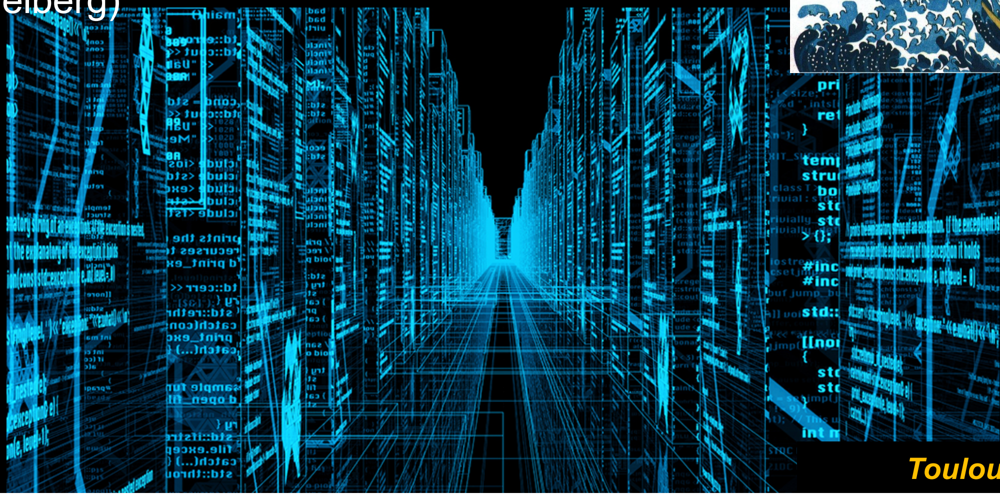
BIG DATA IN ASTROPHYSICS… an overview

Giuseppe Longo

1. Department of Physics, University Federico II in Naples (I)
2. INAF – Astronomical Observatory of Capodimonte (I)
3. CD[3] – Center for Data Driven Discovery – Caltech (USA)

Largely based on work done in collaboration with:

Massimo Brescia
Stefano Cavuoti
George S. Djorgovski (Caltech)
Kai Polsterer (ITHS – Heidelberg)
Valeria Amaro
Civita Vellucci



Università degli Studi Federico II

Istituto Nazionale di Astrofisica - INAF

CENTER FOR DATA-DRIVEN DISCOVERY

California Institute of Technology

TD-1403

COST Action "Big-Sky Earth"

Astrostatistics vs astroinformatics

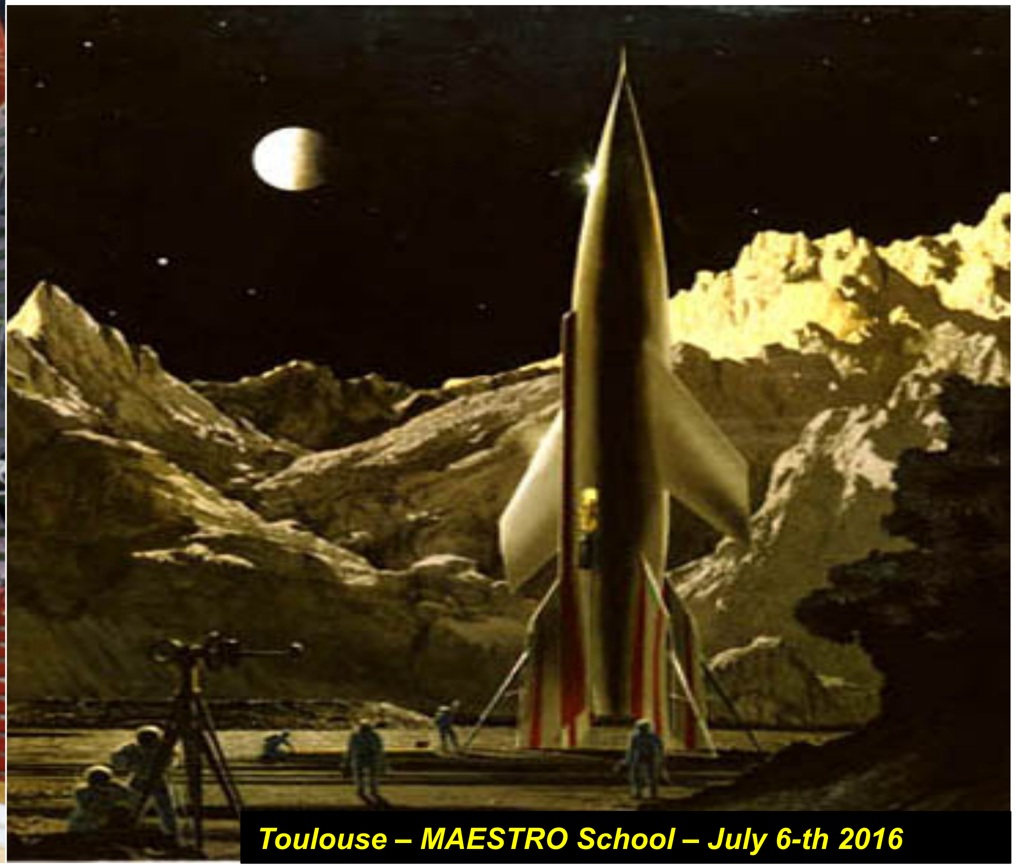
ASTROSTATISTICS:

is a discipline which spans statistical analysis and data mining. It is used to characterize complex datasets, and to link astronomical data to astrophysical theory using the vast amount of data produced by automated scanning of the cosmos.
Many branches of statistics are involved in astronomical analysis including nonparametrics, multivariate regression and multivariate classification, time series analysis, and especially Bayesian inference

ASTROINFORMATICS:

All the rest: data storage and distribution, data processing and data analysis, data mining, data standardization, data re-use, data interoperability, distributed computing, HP computing, visualization, citizen science, etc.

The Future Isn't What It Used To Be



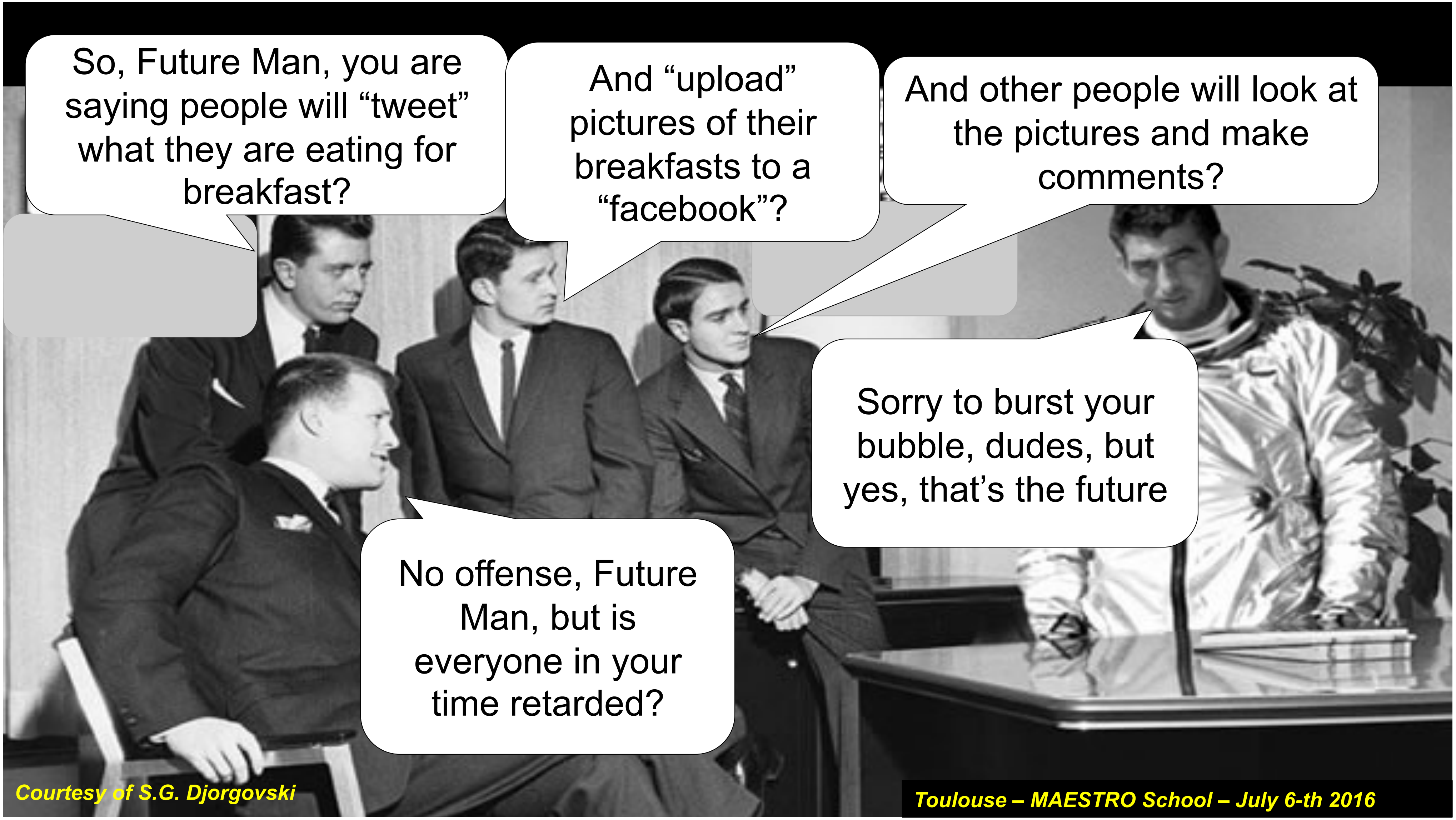
Big Data is like teenage sex:

Everyone talks about it,
Nobody really knows about it,
Everyone thinks everyone else is doing it,
So everyone claims they are doing it ….

Dan Ariely

Big data is not only about **size** but also (may be even more) about **complexity** of data, **heterogeneity** of the data, **data rates**, **variety of tasks** and of the community of users, etc.

Astronomers are makers and users of big data (with some very interesting peculiarities) but they are not the main drive behind data science…

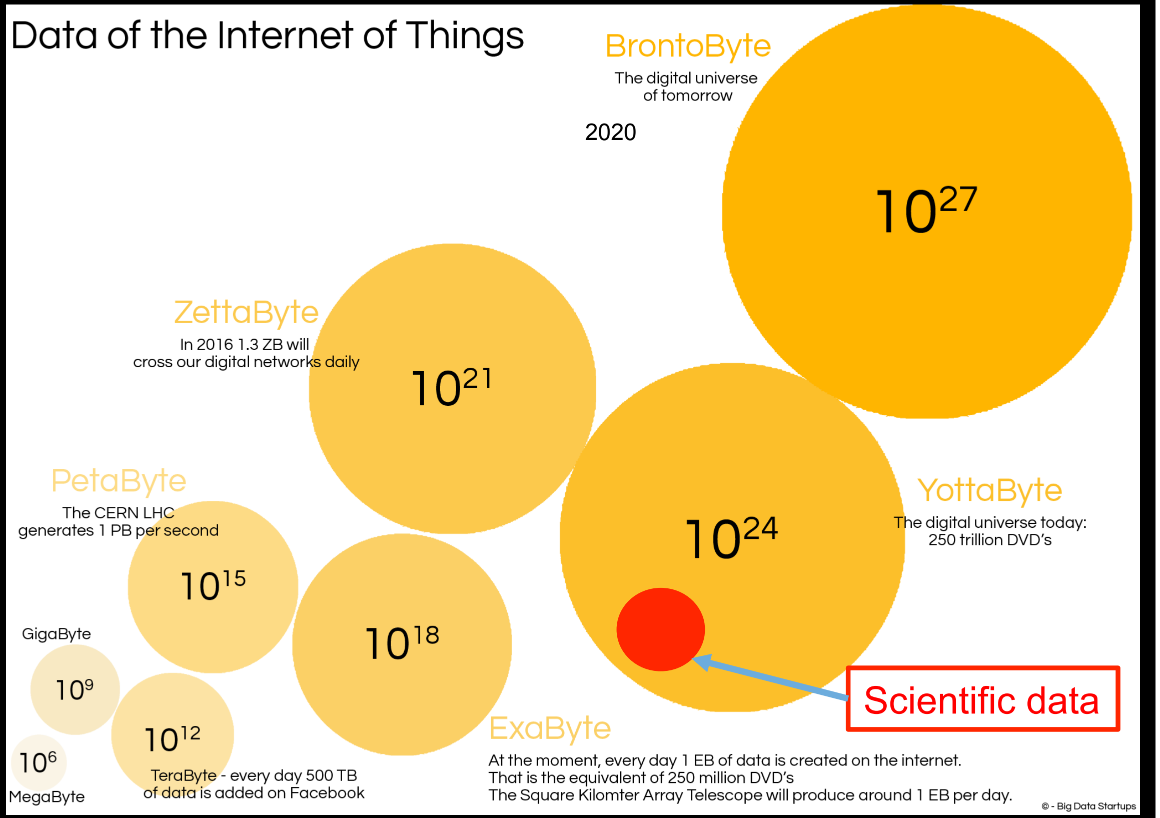


Courtesy of S.G. Djorgovski

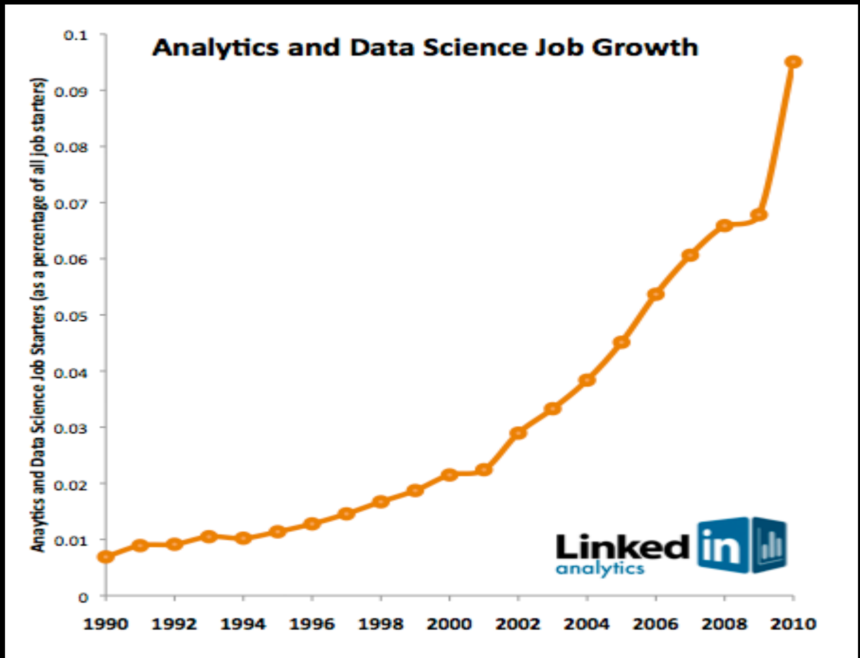
Overwhelmingly large data sets are produced for:

everything….

Finance
Marketing
Domotics
Environmental sensors
Meteorology
Tele-health
Genomics
Bioinformatics
Astrophysics
Physics
Biology
Engineering
Smart cities
Public Administration
Social Sciences
Human Sciences and Digital libraries
Etc….



The request of data scientists is exponentially increasing



BUT:
What are big data?

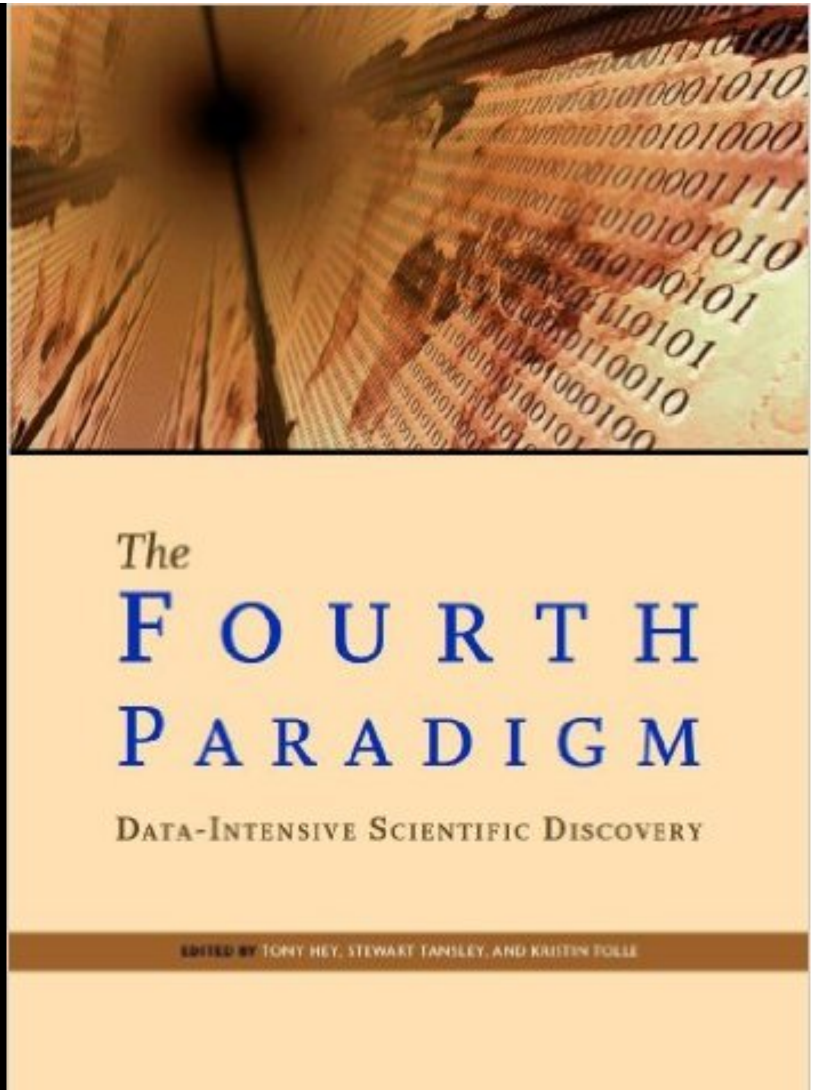
Turning point

The Fourth Paradigm – T. Hey et al., Microsoft Research, 2009

Kindle Download from Amazon

e-Science
X-informatics
Data Science,

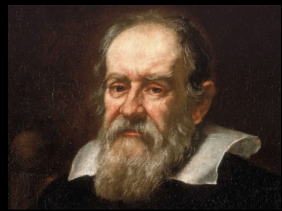
etc

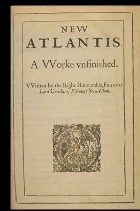


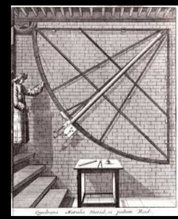
The evolving paths to knowledge

(Jim Gray)

The First Paradigm
Experiments/measurements
(XVII century)



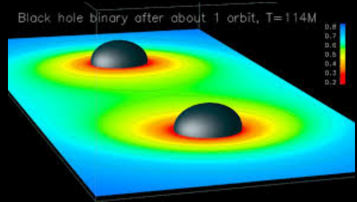


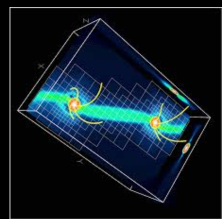


The Second Paradigm
Analytical theory
(XVIII century)

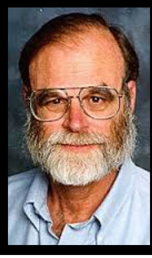
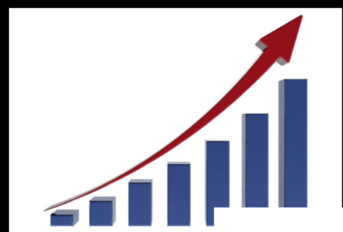
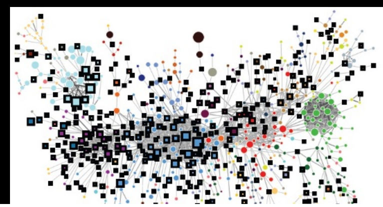
$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\varepsilon} \qquad \text{(Gauss' Law)}$$
$$\nabla \cdot \mathbf{H} = 0 \qquad \text{(Gauss' Law for Magnetism)}$$
$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \qquad \text{(Faraday's Law)}$$
$$\nabla \times \mathbf{H} = \mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampere's Law)}$$

The Third Paradigm
Numerical simulations
(early 40's)



The Fourth Paradigm
Data Driven Discovery
(Now)

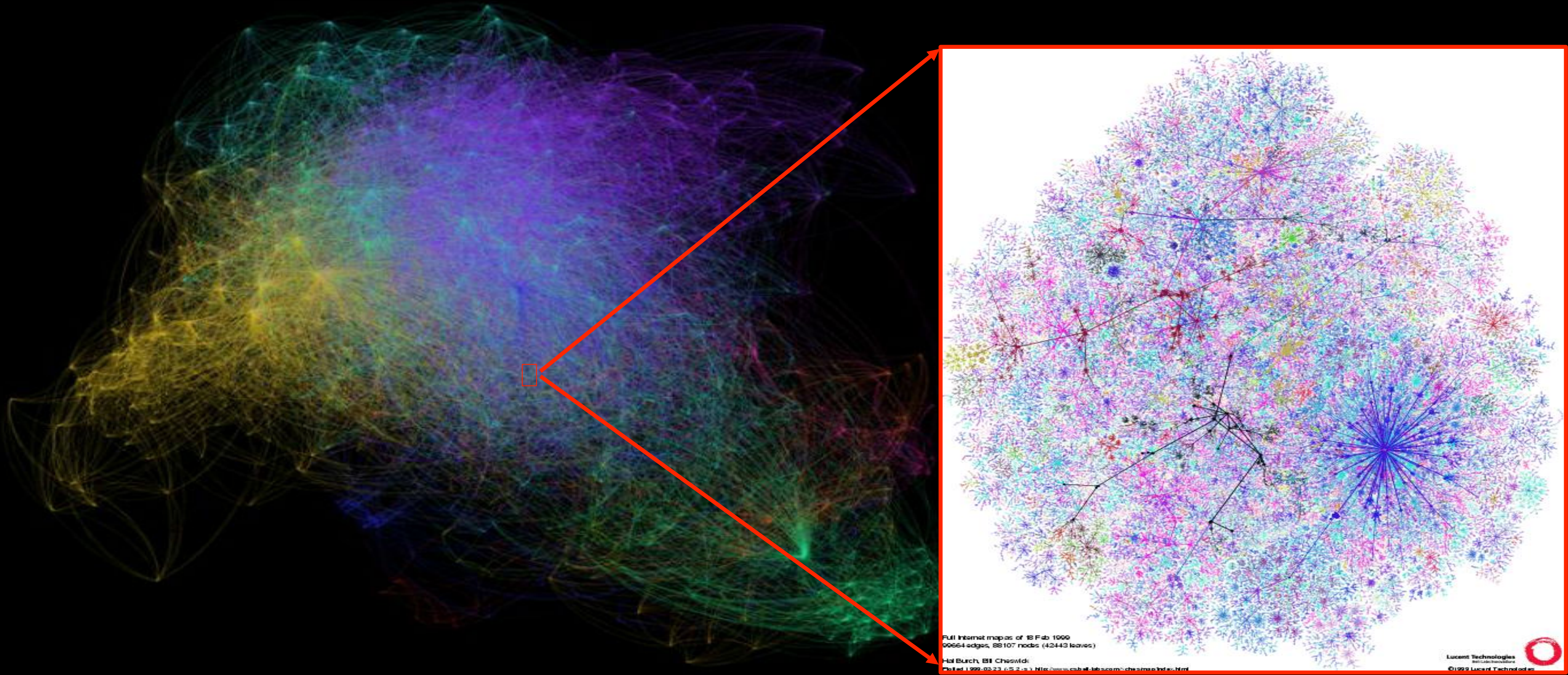
Big data accordingly to Borat

Big Data is any thing which is crash Excel.

Small Data is when is fit in RAM. Big Data is when is crash because is not fit in RAM.

Or, in other words, Big Data is data in volumes too great to process by traditional methods.

WHAT ARE BIG DATA? - connectivity



… 2 billions of nodes in 2014 connecting data but also … computing power --- in the CLOUD !

Cloud computing is "…computing based on the internet…"

Where in the past, people would run applications or programs from software downloaded on a physical computer or server in their building, cloud computing allows people access to the same kinds of applications through the internet.

When you update your Facebook status, you're using cloud computing. Checking your bank balance on your phone? You're in the cloud again.

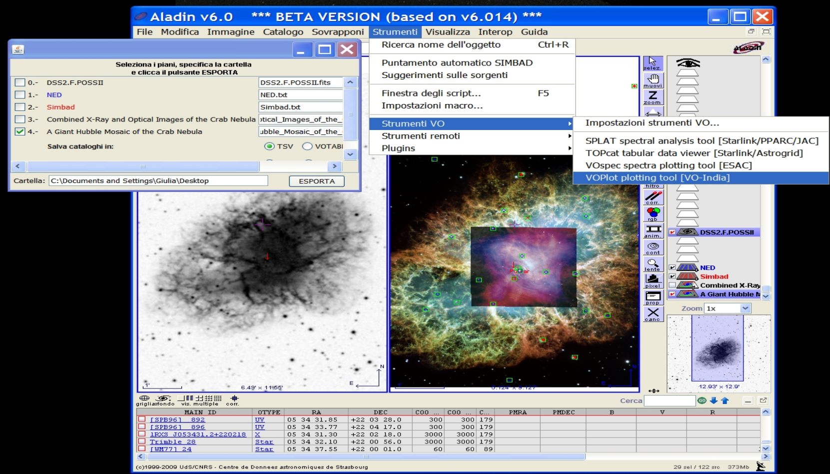
In short, cloud is fast becoming the new normal. By the end of 2016 it was estimated that 90% of UK businesses will be using at least one cloud service.

STANDARDIZATION AND INTEROPERABILITY

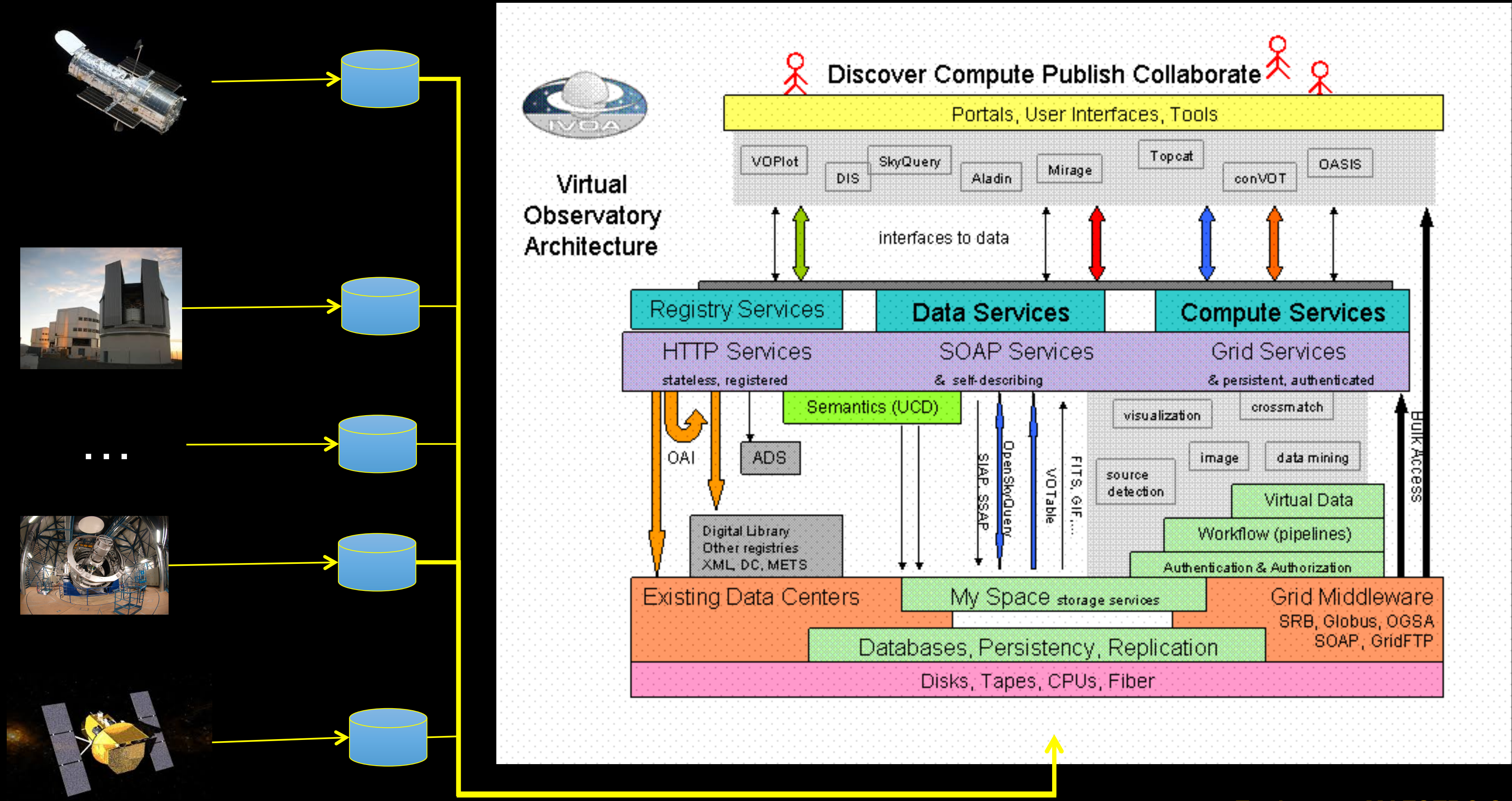
IVOA
INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE

MOST, IF NOT ALL, ASTRONOMICAL DATABASES

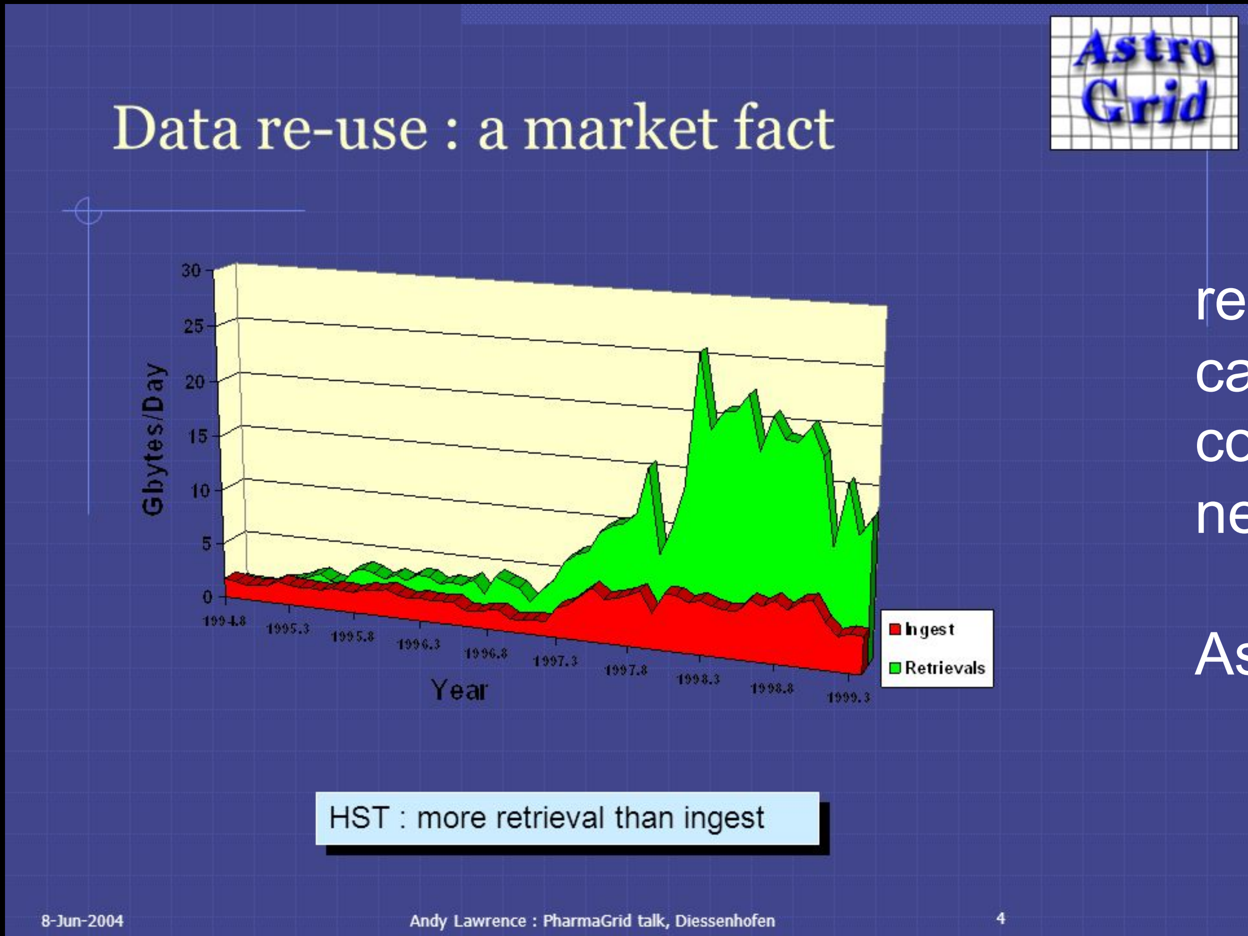
- Common standards (data structures)
- Common Formats (FITS, VOTable)
- Uniform descriptors
- Common resource registry
- Compliant tools
- Samp (Interoperability)



The Virtual Observatory



1. DATA RE-USE: VO as a new type of telescope…

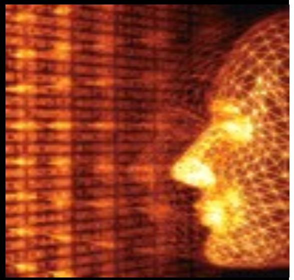


retrieving data in most cases will be much more convenient than obtaining new observations…

AstroInformatics 2012

Exploiting big data complexity in science calls for:

Statistics, Statistical Pattern Recognition, Data Mining (Machine learning and artificial Intelligence),….

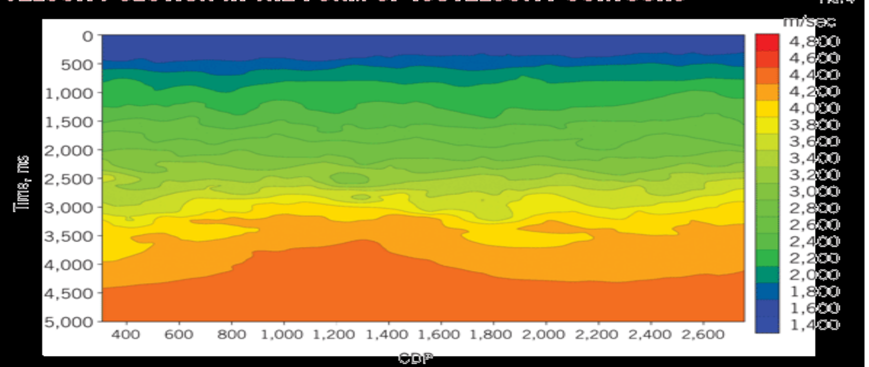


- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches

Advanced visualization:

- Data compression (dimensionality reduction)
- Immersive and virtual reality
- Etc.

VELOCITY SECTION IN THE FORM OF ISOVELOCITY CONTOURS

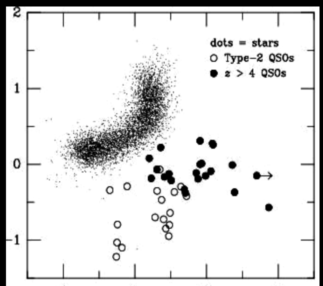


Virtual Observatory Science Examples

Courtesy of S.G. Djorgovski

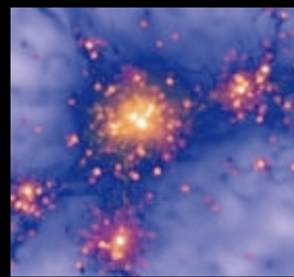
Combine the data from multi-TB, billion-object surveys in the optical, IR, radio, X-ray, etc.

- Precision large scale structure in the universe
- Precision structure of our Galaxy



- Discover rare and unusual (one-in-a-million or one-in-a-billion) types of sources
  - E.g., extremely distant or unusual quasars, new types, etc.

Match Peta-scale numerical simulations of star or galaxy formation with equally large and complex observations



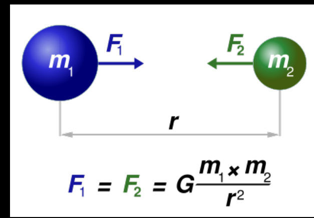
… etc., etc.

Implications of complexity of big data for science

Physical, social, economic, biological laws are derived from data patterns

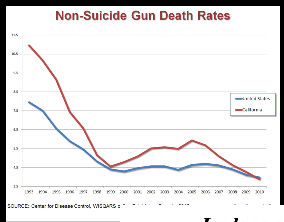
$$f(x,y,z) = 0$$

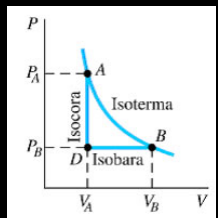
No empirical law depends on more than 3 independent parameters !!!



A simple universe… or rather **an intrinsic human bias …**

… affecting our knowledge and our understanding of the physical laws





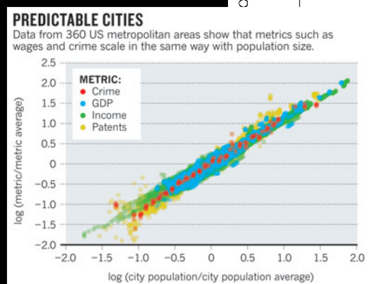
La legge dei gas ideali

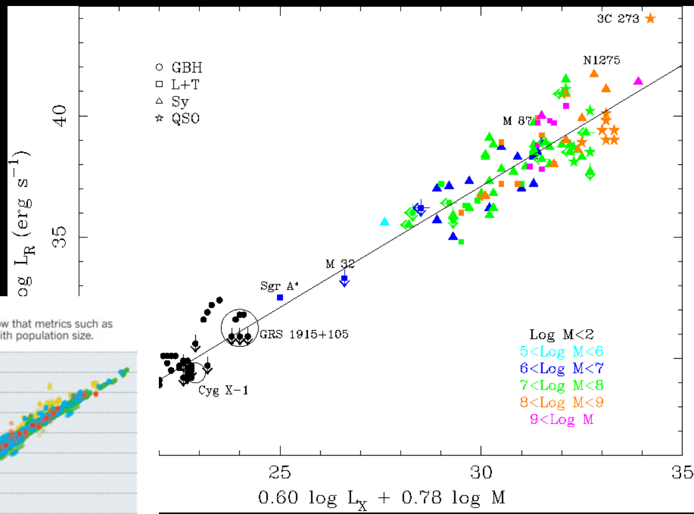
Unendo le leggi di Boyle, Charles e Avogadro si ha un'unica legge, approssimativamente valida per tutti i gas

PV= nRT

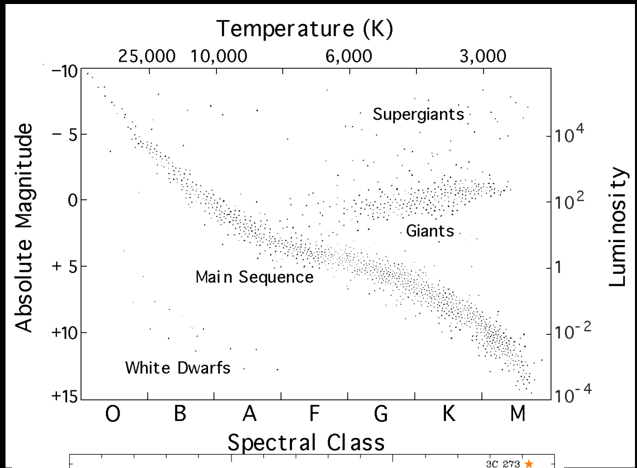
la costante R, costante dei gas, ha il medesimo valore per ogni gas (cioè è «universale»).
Alla temperatura di 0°C (273,15 K) e alla pressione di 1,00 atm, 1,00 mol di un gas occupano il volume di 22,414 L,

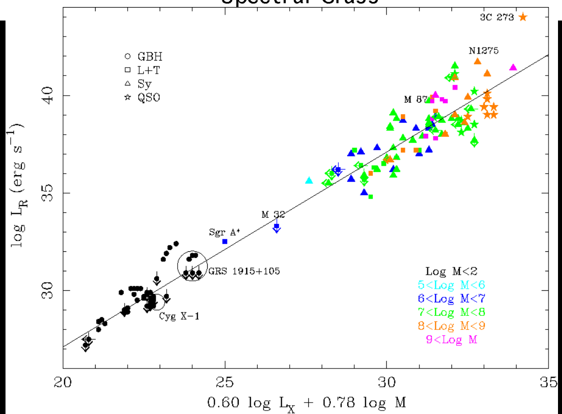
R = 0.08206 L . atm/(K . mol)
R = 62.37 L . Torr/(K . mol)
R = 8.314 L . Pascal/(K . mol)



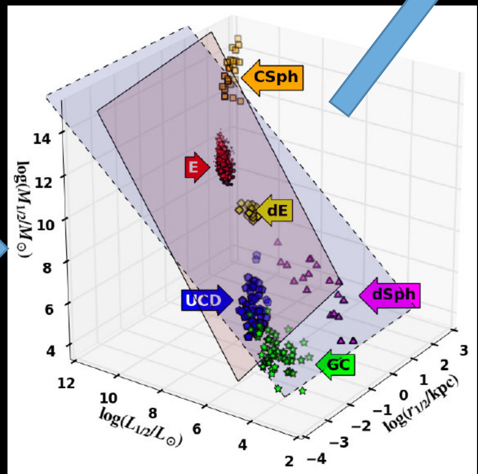


$$\mathrm{SFR} \propto \begin{cases} (1+z)^{\beta} & \text{for } z < 1\,, \\ (1+z)^{\alpha} & \text{for } 1 \le z < 5\,, \\ 0 & \text{for } z \ge 5\,. \end{cases}$$

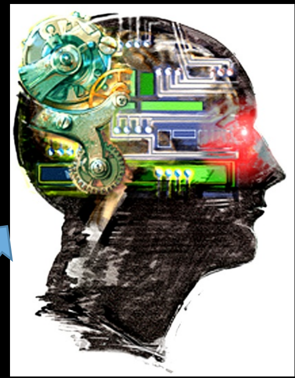




2-d diagnostics



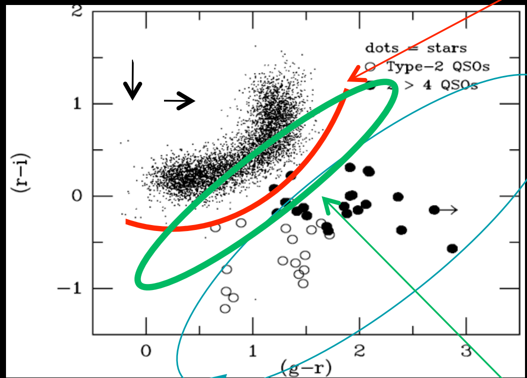
3-d diagnostics



What should we do to extract patterns (i.e. laws r ordering relationships) in a R^n space (n>>100) ?

Traditional way to look for candidate QSO in 3 band survey

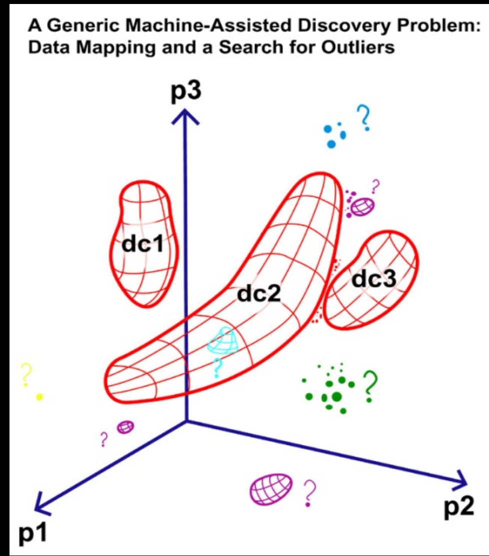
Cutoff line



Candidate QSOs for spectroscopic follow-up's

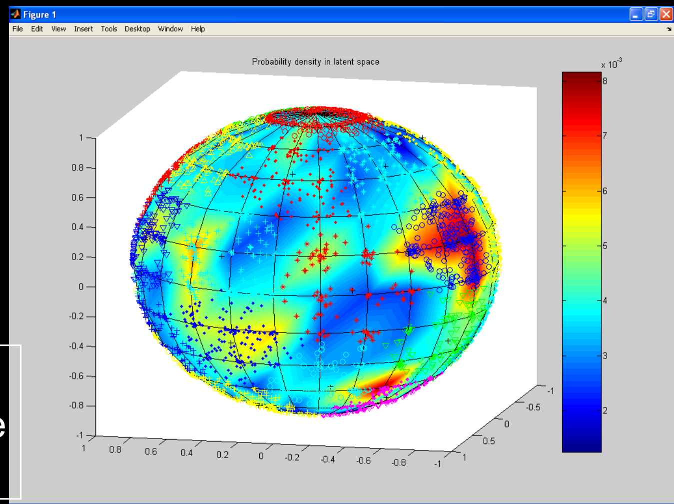
Ambiguity zone

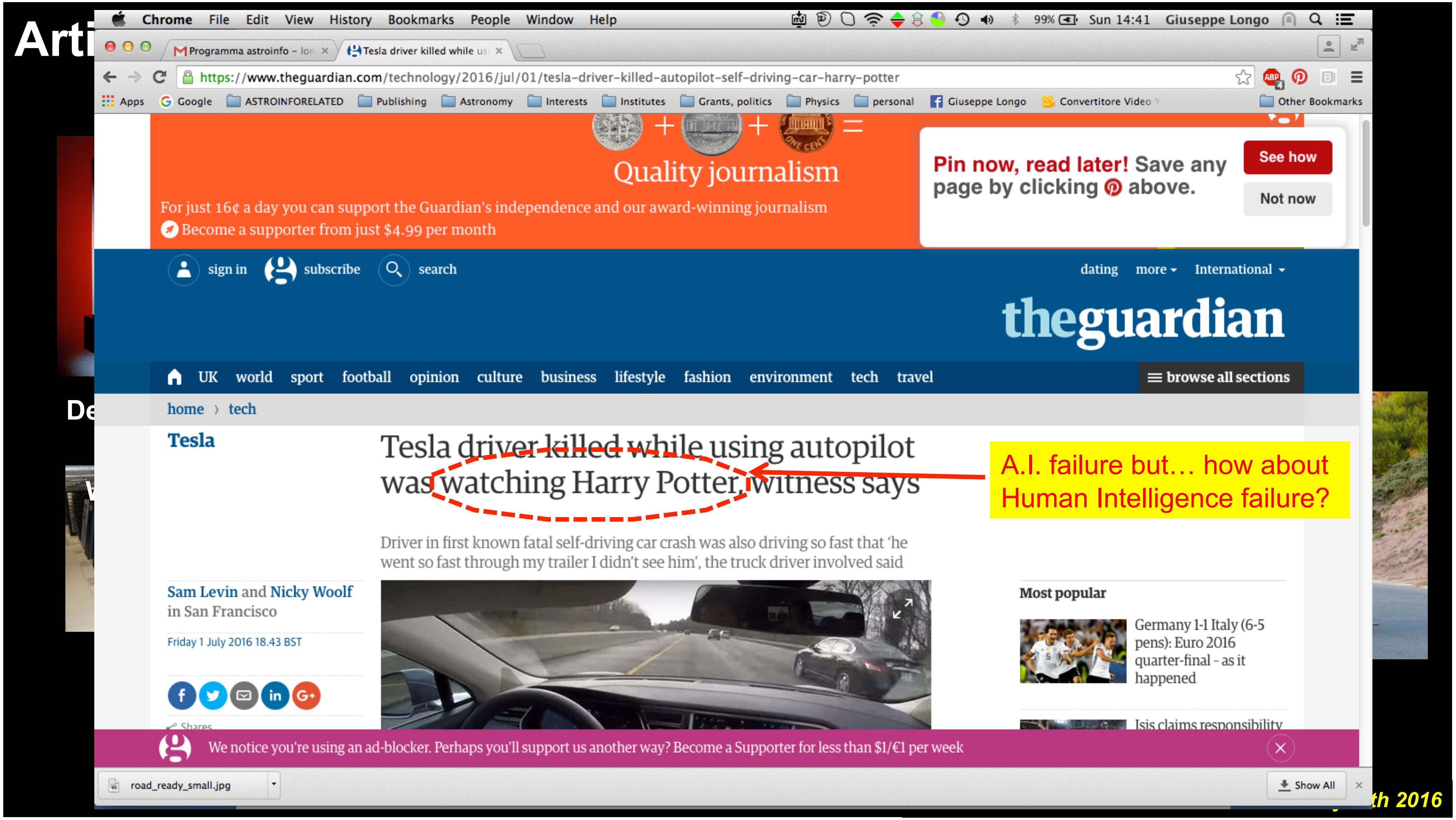
Adding one feature improves separation…



Need for Machine learning and AI

Probabilistic Principal Surfaces + clustering projection on a sphere of a 21-D parameter space showing as blue dots the candidate quasars…




Arti
De
Chrome File Edit View History Bookmarks People Window Help
Sun 14:41 Giuseppe Longo
Programma astroinfo
Tesla driver killed while us
https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter
Apps Google ASTROINFORELATED Publishing Astronomy Interests Institutes Grants, politics Physics personal Giuseppe Longo Convertitore Video Other Bookmarks
Quality journalism
For just 16¢ a day you can support the Guardian's independence and our award-winning journalism
Become a supporter from just $4.99 per month
Pin now, read later! Save any page by clicking above.
See how
Not now
sign in subscribe search
dating more International
theguardian
UK world sport football opinion culture business lifestyle fashion environment tech travel
browse all sections
home › tech
Tesla
Tesla driver killed while using autopilot was watching Harry Potter, witness says
A.I. failure but… how about Human Intelligence failure?
Driver in first known fatal self-driving car crash was also driving so fast that 'he went so fast through my trailer I didn't see him', the truck driver involved said
Sam Levin and Nicky Woolf in San Francisco
Friday 1 July 2016 18.43 BST
Most popular
Germany 1-1 Italy (6-5 pens): Euro 2016 quarter-final - as it happened
Isis claims responsibility
We notice you're using an ad-blocker. Perhaps you'll support us another way? Become a Supporter for less than $1/€1 per week
road_ready_small.jpg
Show All
th 2016

Find the sentence

…. Lucky is the people with blonde air and blu eyes which never goes to that far end where everything becomes meaningless and terrible …

Or…:

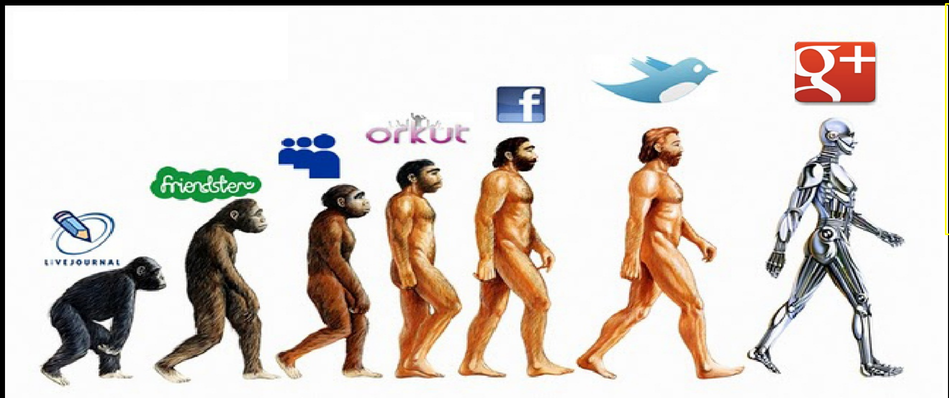
Find all sentences which are semantically equivalent to the previous one …. (e.g. not everyone can grasp the difficulties of life)

Or.…

Find a specific pattern (i.e. a chicken) in all images, videos and frames existing in the WWW

Or…

Translate all books in any of the 294 existing languages and dialects…



Google does all this and much more hundreds of millions times everyday

Complexity of astronomical data. Some implications

Multiwavelenght Digital Surveys

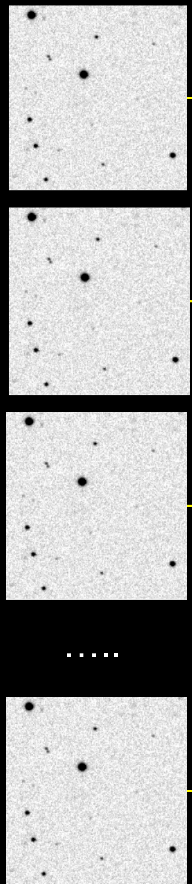
Band 1

Band 2

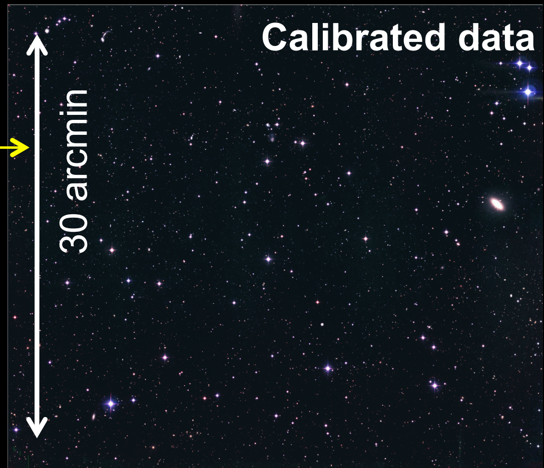
Band 3

.....

Band n

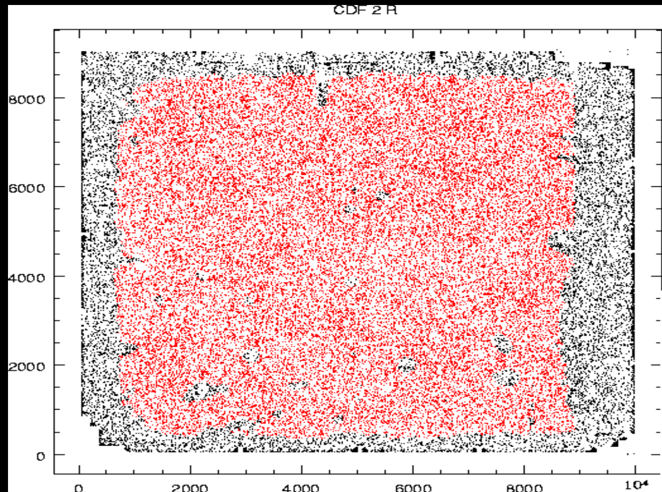


Calibrated data



1/160.000 of the sky, moderately deep (25.0 in r)

55.000 detected sources (0.75 mag above m lim)

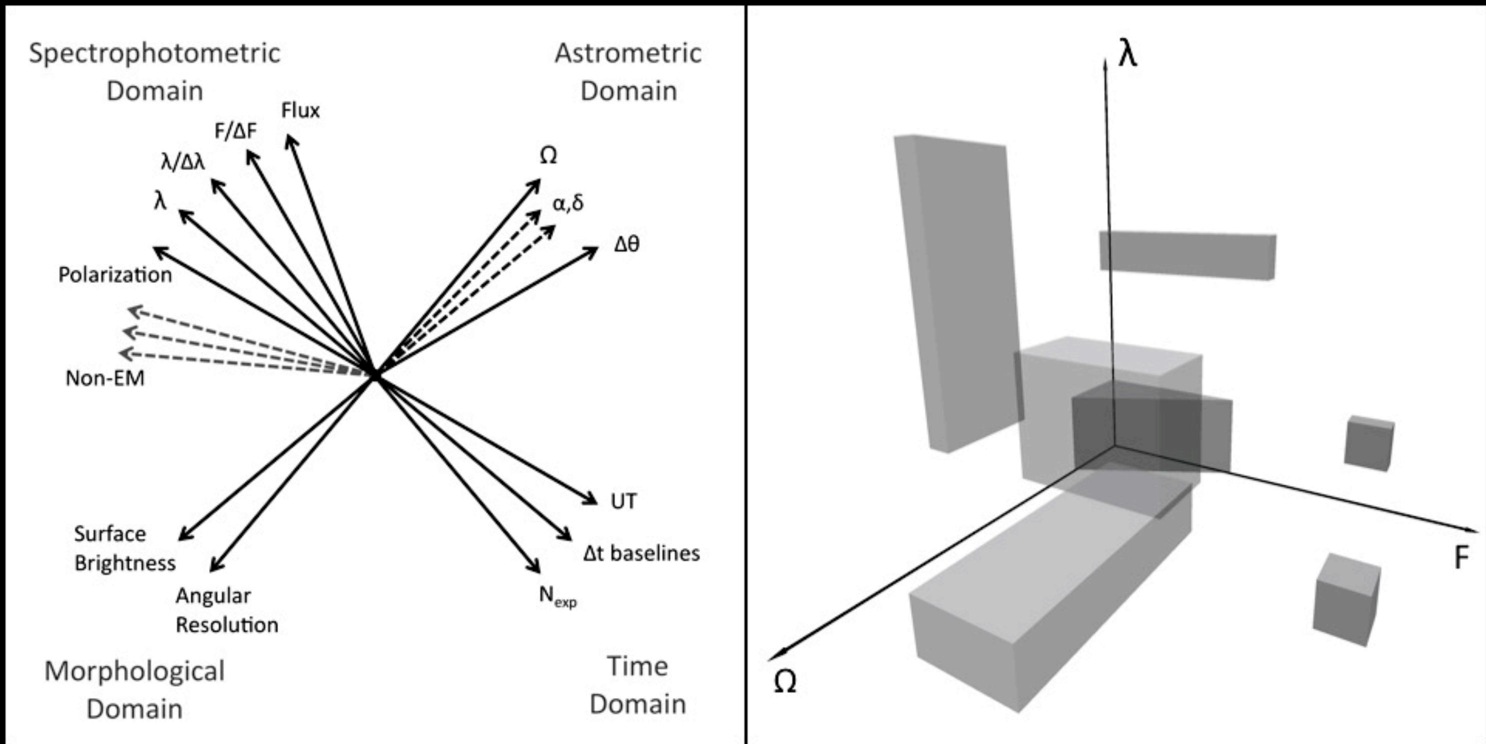


Measure attributes (brightness, position, shapes, etc.) of detected sources ➡

Each object becomes a record p^i defined by n parameters (or features)

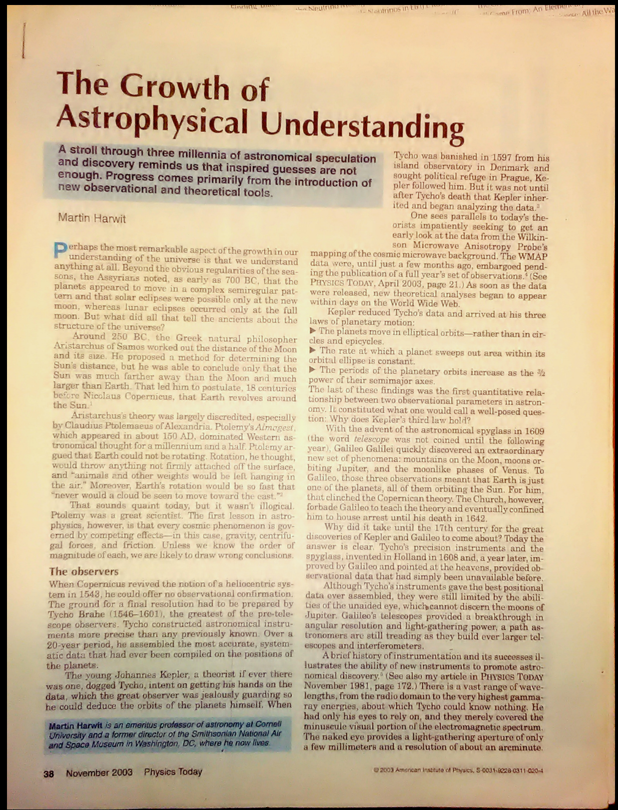
$$p^i \equiv \left\{ t, \lambda_1, \left\{ f_1^1, f_2^1, \ldots, f_{k'}^2 \right\}, \lambda_2, \left\{ f_1^2, f_2^2, \ldots, f_{k''}^2 \right\}, \ldots, \lambda_n \left\{ f_1^n, f_2^n, \ldots, f_{k^n}^n \right\} \right\}$$

Hence an observation is a point in a high dimensionality parameter space



Why understanding the PS is particularly important in astronomy?

Because astronomy is based on observations and not experiments….



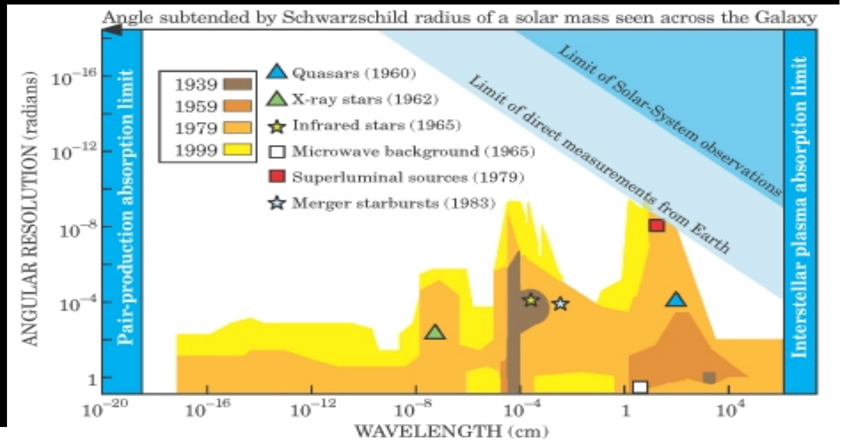
The Growth of Astrophysical Understanding

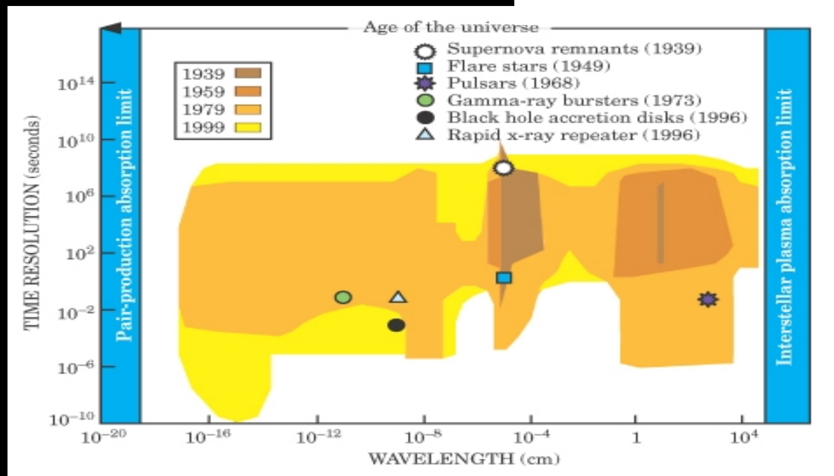
A stroll through three millennia of astronomical speculation and discovery reminds us that inspired guesses are not enough. Progress comes primarily from the introduction of new observational and theoretical tools.

Martin Harwit

38 November 2003 Physics Today

The history of astronomical discoveries can be reconstructed in terms of better coverage or sampling of the Parameter Space





DATA MINING instead of Serendipity

M. Harwit, Physics Today, 2003

BUT: Exploration of high dimensionality PS ($N > 10^9$, $D >> 100$, $K > 10$) 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

MOST DATA MINING methods scale poorly

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**

Parallelization ?

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

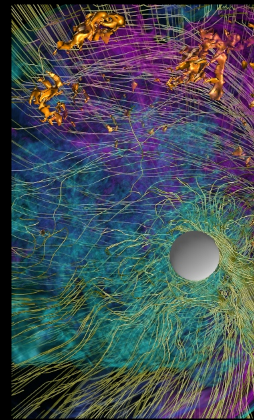
Web Services:
- SDSS mirror
- WFXT Time Calculator
- GAME (GPU+CUDA ML model)

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

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

Courtesy of S.G. Djorgovski

Some Thoughts About e-Science



- Comput***ational*** science $\neq$ Comput***er*** science
- Data-driven science is *not* about data, it is about ***knowledge extraction***
- Information and data are (relatively) cheap, but the expertise is expensive
  - Just like the hardware/software situation

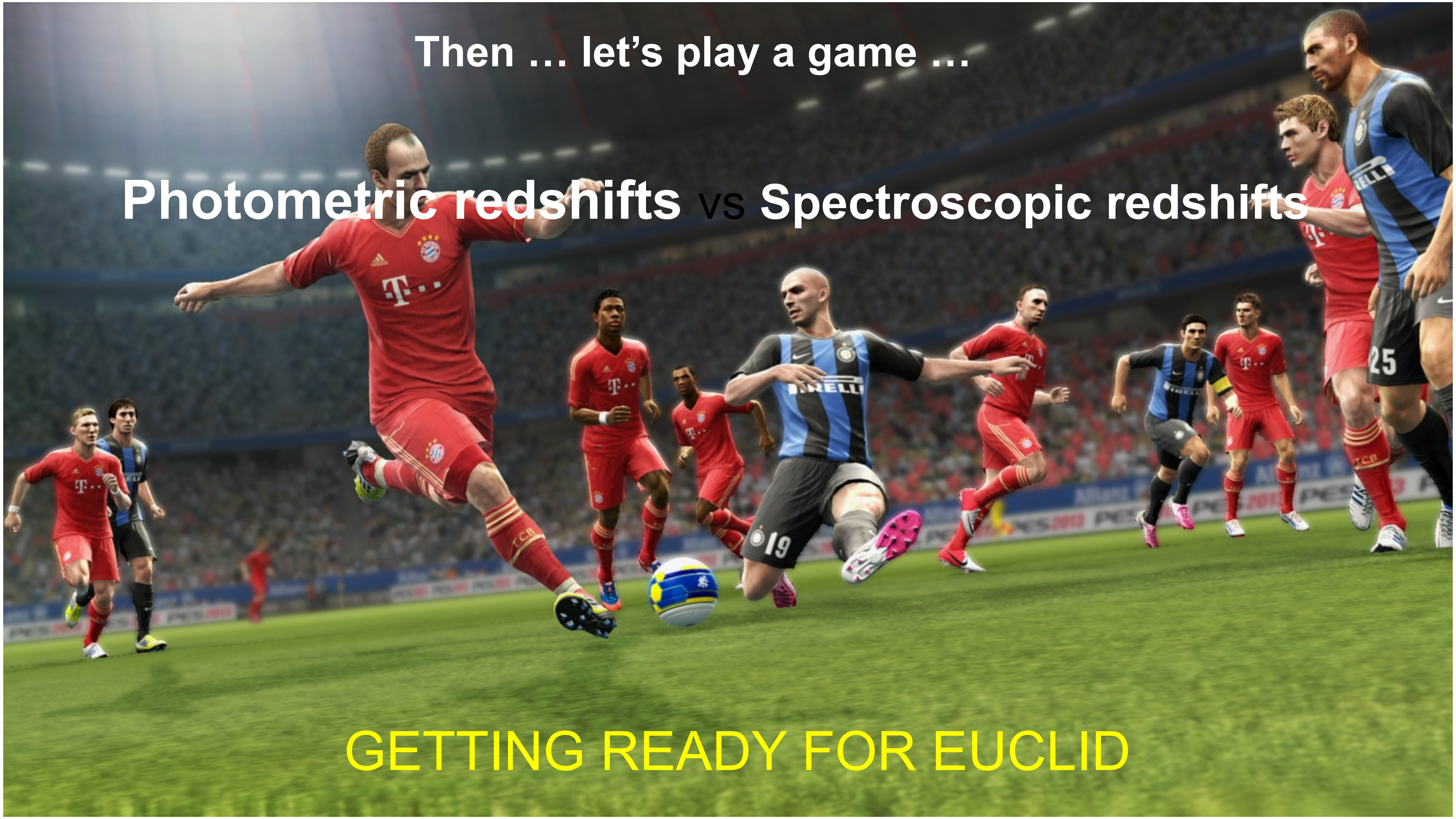
- Computer science as the "new mathematics"
  - It plays the role in relation to other sciences which mathematics did in ~ 17th - 20th century

- Computation: an interdisciplinary glue/lubricant
  - Many important problems (e.g., climate change) are inherently inter/multi-disciplinary

DATA MINING is about rediscovering/discovering known (unknown) useful patterns in the data

DATA DRIVEN DISCOVERY is not «simply» about machine learning…

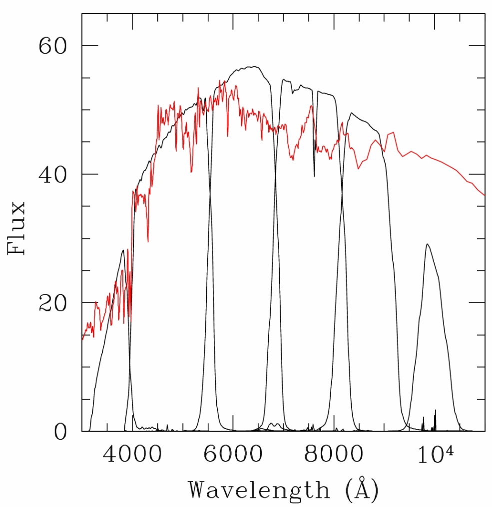
D*^{3}** is about ***letting the data to speak for themselves with minimum use of a-priori assumed models and hypothesis

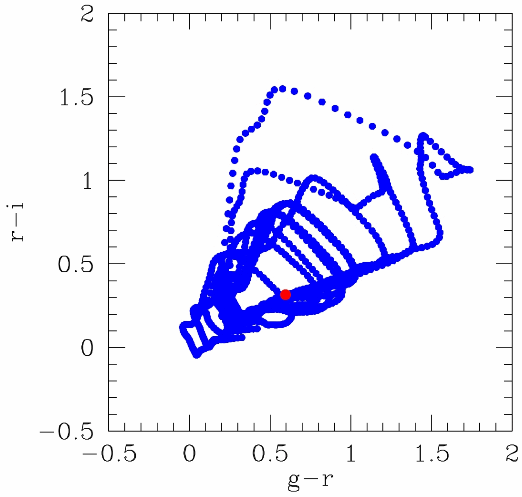

Then … let’s play a game …
Photometric redshifts vs Spectroscopic redshifts
GETTING READY FOR EUCLID

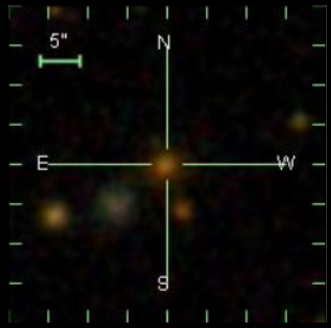
A template case of …. machine learning vs «pure» *D*^{3}

Photometric redshifts for quasars and galaxies

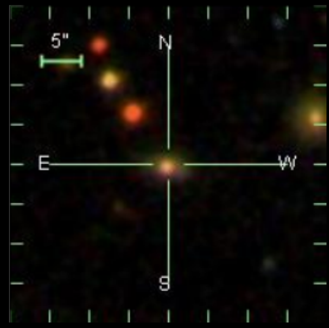
$$1 + z = \frac{\lambda_{obs}}{\lambda_0} \approx \frac{v}{c}$$







QSO; z=3.81



QSO; z=5.31

Crucial cosmological probe

- Large scale structure
- Weak lensing
- Tests of cosmological models

Only viable way to obtain distance info's for large samples of galaxies

Mathematically simple: to find the mapping function $f(\bar{x}) \rightarrow y$, where: $\bar{x} \in \mathbb{R}^n, y \in \mathbb{R}$

Input vector $\left(X_j \{x_1,\ldots,x_n\} j = 1,..m\right) \in OPPS \subset \mathbb{R}^n$

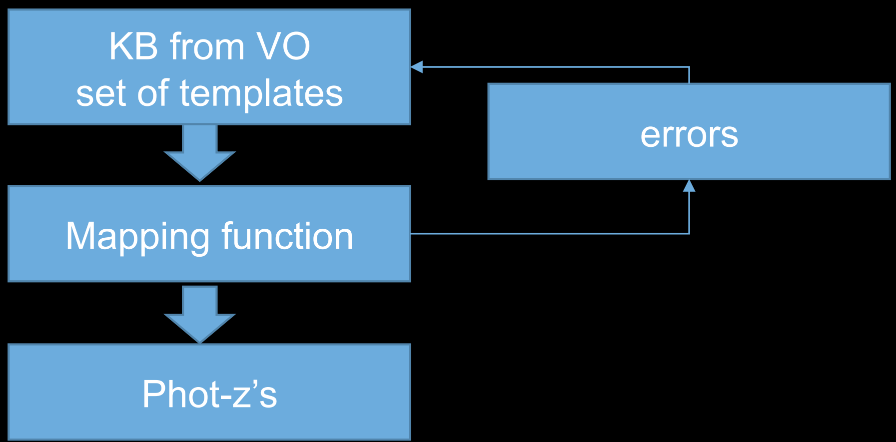
OPPS = Observable Photometric Parameter Space

Target vector $\bar{Y}_j \{x_1,\ldots,x_n\} \in OPPS \subset \mathbb{R}$

OSPS = Observable Spectroscopic Parameter Space

Physical redshift

PPS = Physical Parameter Space



The Sloan Digital Sky Survey *(in its various incarnations)*

Sloan Digital Sky Survey – Sky Server
–2.5 Terapixels of images => 5 Tpx of sky; 10 TB of raw data => 400TB processed; 0.5 TB catalogs => 35TB final

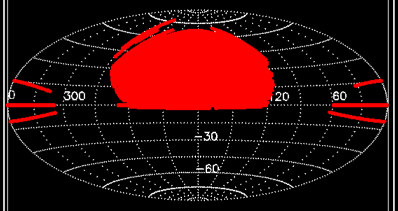
… a Prototype in 21st Century data access
–1.2B web hits in 12 years; 200M external SQL queries; 4,000,000 distinct users vs. 15,000 astronomers

Data products (e.g. **SPECTROSCOPIC and PHOTOMETRIC** catalogues) and raw data were «immediately» made available to the community

The right data set at the right moment

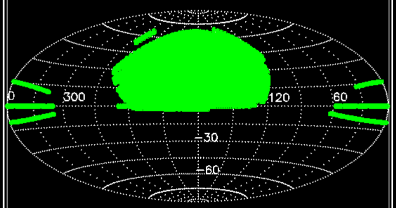
Pioneeristic yet manageable with available technology (10 TB of data products); general in purpose, flexible enough to be useful for a large variety of existing problems, yet capable to rise new ones

3x10^8 galaxies

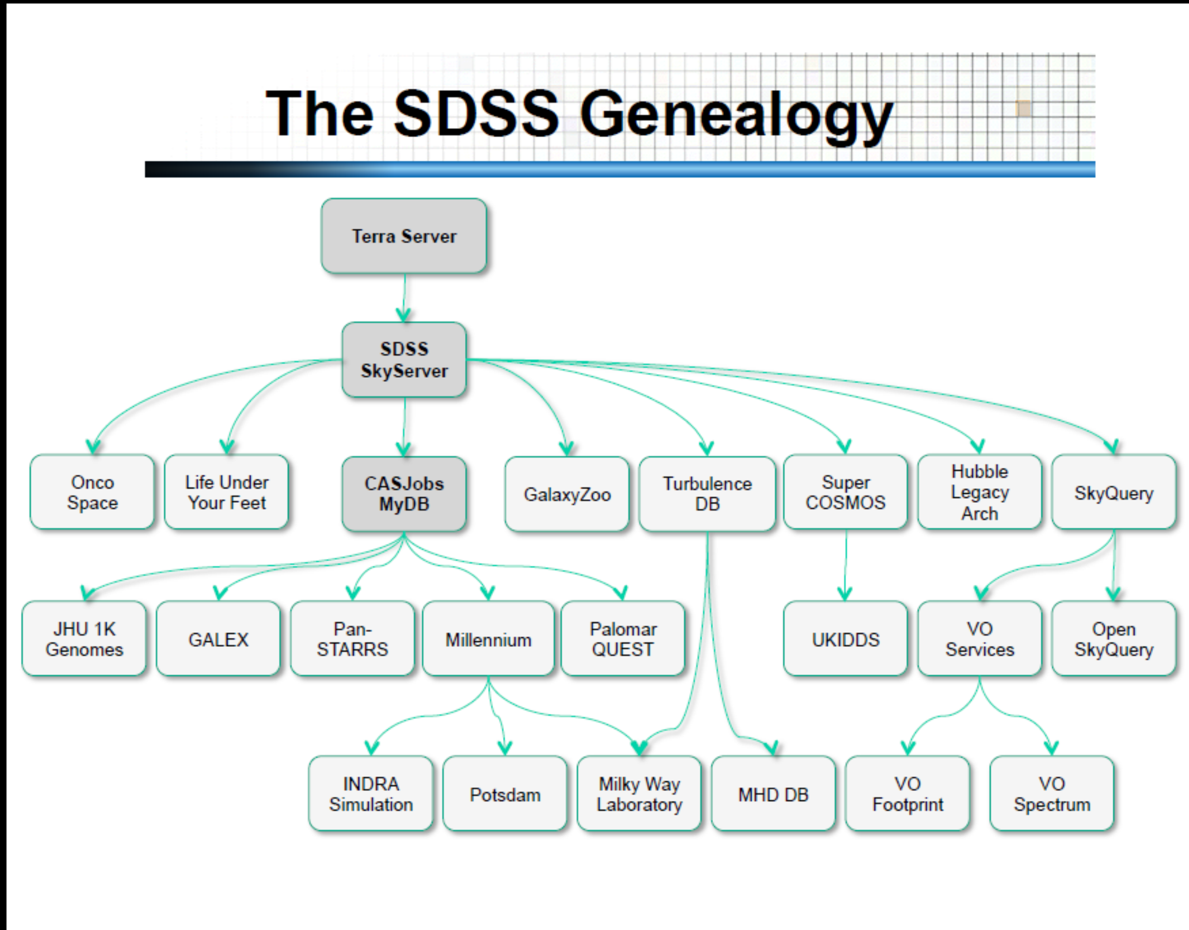


DR10 – photo coverage

3x10^6 galaxies



DR10 – spectral coverage



SDSS – Data Release 10

OPPS		OSPS
3×10^8	objects	3×10^6
> 100	features	>50
> 100	flags	>30

Problem:

To evaluate Photo-z for all SDSS objects using the spectroscopic z's in the KB

The KB is the result of selection criterias and is biased

Not all selections and biases can be mapped in the OPPS

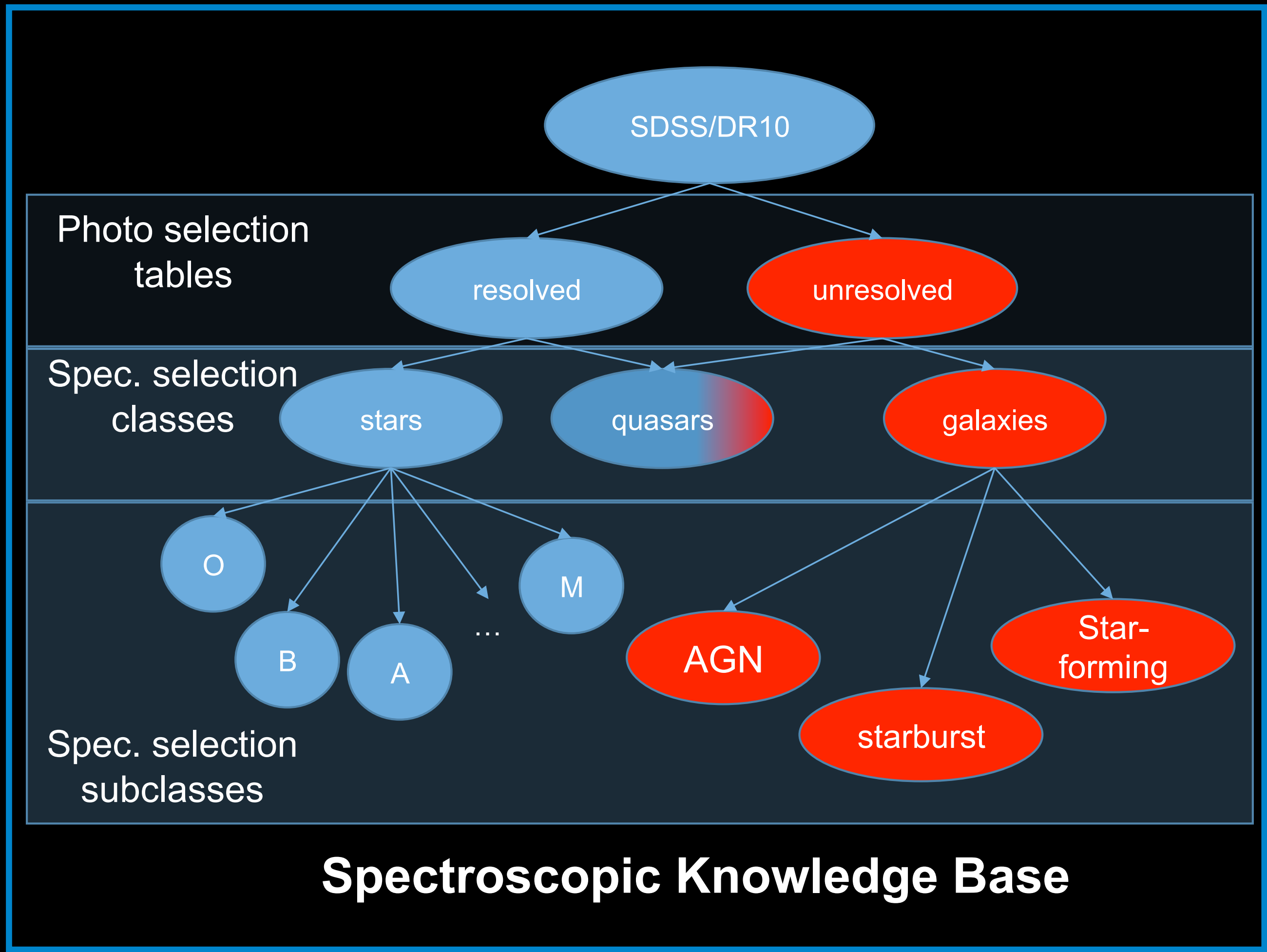
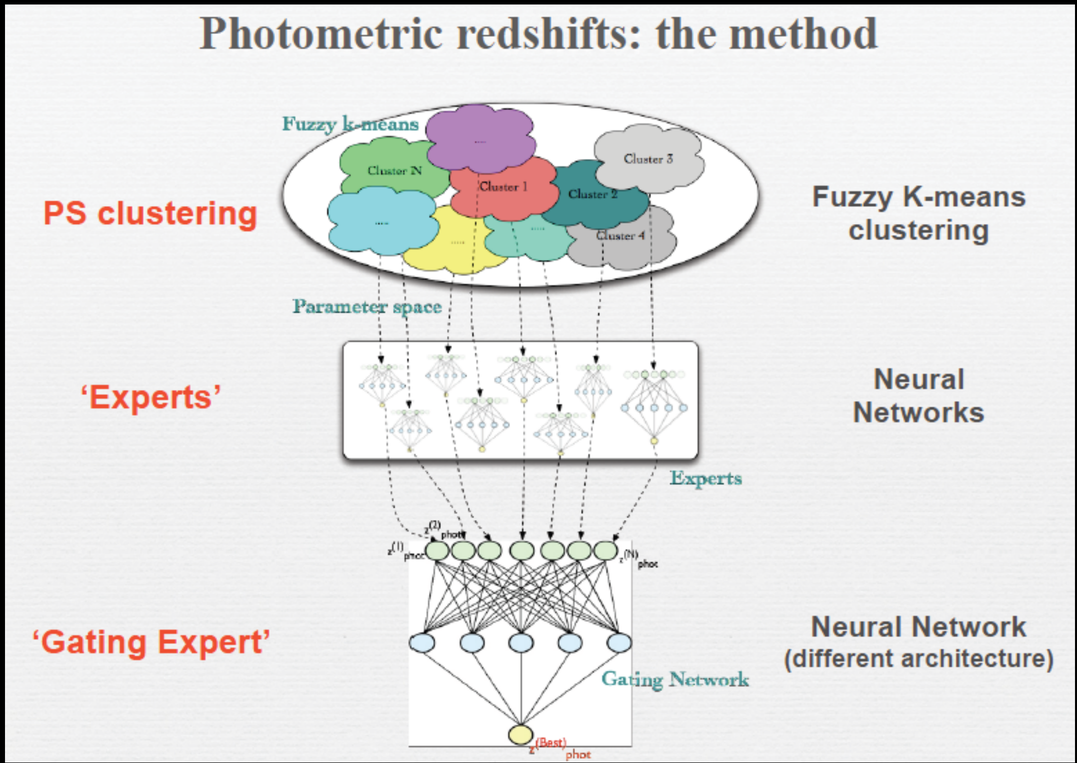


Photo-z for Quasars:

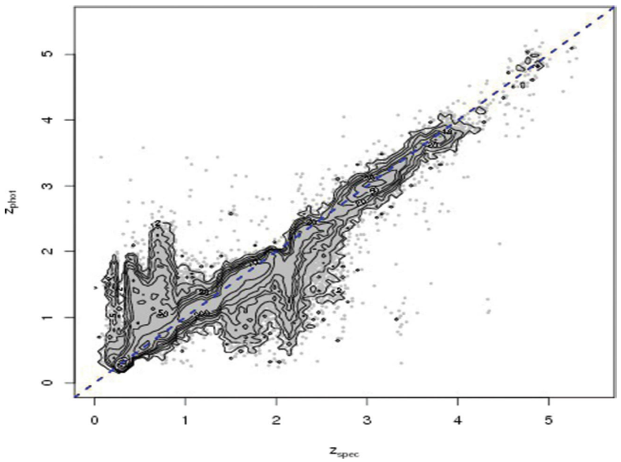
Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation, O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

WGE: Weak Gated Expert

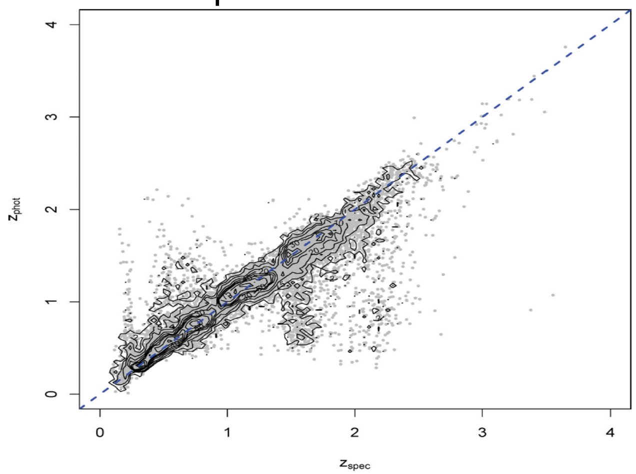
Data from the unresolved objects SDSS catalogue



Optical bands only



Optical + UV bands



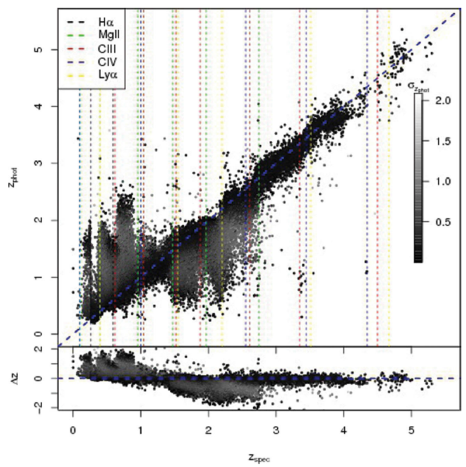


Figure 15. In the upper panel, it is shown the scatter plot of the spectroscopic versus photometric redshifts evaluated with the WGE method for the members of the KB of the experiment for the quasars extracted from the SDSS catalogue with optical photometry, while in the lower panel the scatter plot of the spectroscopic redshift $z_{\rm spec}$ versus Δz variable is shown for the same sources. All points are colour coded according to the value of the errors $\sigma_{z_{\rm phot}}$ as evaluated but the WGE. The vertical dashed lines represent the redshift at which the most luminous emission lines characterizing quasars spectra shift off the SDSS photometric filters due to redshift. Most of the features of the plot are associated to one or more of these lines.

Photo-z's for SDSS QSO's with MLPQNA

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso mag_Aper_1 mag_Aper_2 mag_Aper_3 mag_auto and kron_radius	Near and Far UV total and isophotal mags phot. through 3, 4.5 and 7.5 arcsec apertures magnitudes and Kron radius in units of A or B
SDSS	u, g, r, i, z	psfMag	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y, J, H, K	PsfMag AperMag3, AperMag4, AperMag6 HallMag, PetroMag	PSF fitting magnitude in Y, J, H, K bands aperture photometry through 2, 2.8 & 5.7$''$ circular aperture in each band Calibrated magnitude within circular aperture r_hall and Petrosian magnitude in Y, J, H, K bands
WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro	W1: 3.4 μm and 6.1$''$ angular resolution; W2: 4.6 μm and 6.4$''$ angular resolution; W3: 12 μm and 6.5$''$ angular resolution; 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	-	z_{spec}	Spectroscopic redshift

Photometric redshifts for quasars in multiband surveys, M. Brescia, S. Cavuoti, R. D'Abrusco, A. Mercurio, G. Longo, 2013, ApJ, 772, 140 (astro-ph: 1305.5641)

Lenghty feature selection procedure

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{clean}(\Delta z_{norm})$ and $NMAD(\Delta z_{norm})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{norm}| > 2\sigma\,(\Delta z_{norm})$. The standard deviation $\sigma_{clean}(\Delta z_{norm})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{norm}| \leq 2\sigma\,(\Delta z_{norm})$

Exp	n. obj.	$\sigma\,(\Delta z_{norm})$	% catas. outliers	$\sigma_{clean}\,(\Delta z_{norm})$	$NMAD\,(\Delta z_{norm})$
SDSS	41431	0.15	6.53	0.062	0.058
SDSS + GALEX	17876	0.11	4.57	0.045	0.043
SDSS+UKIDSS	12438	0.11	3.82	0.041	0.040
SDSS+GALEX+UKIDSS	5836	0.087	3.05	0.040	0.032
SDSS+GALEX+UKIDSS+WISE	5716	0.069	2.88	0.035	0.029

Table 4. Comparison among the performances of the different references. MLPQNA is related to our experiments, based on a four-layers network, trained on the mixed (colors + reference magnitudes) datasets. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; columns 2-6, respectively: bias, standard deviation, MAD, RMS and NMAD calculated on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$ related to the test sets. For the definition of the parameters and for discussion see text.

Exp	$BIAS(\Delta z_{norm})$	$\sigma(\Delta z_{norm})$	$MAD(\Delta z_{norm})$	$RMS(\Delta z_{norm})$	$NMAD(\Delta z_{norm})$
			SDSS		
MLPQNA	0.032	0.15	0.039	0.17	0.058
Laurino et al.	0.095	0.16	0.041	0.19	-
Ball et al.	0.095	0.18	-	-	-
Richards et al.	0.115	0.28	-	-	-
			SDSS + GALEX		
MLPQNA	0.012	0.11	0.029	0.11	0.043
Laurino et al.	0.058	0.29	0.029	0.11	-
Ball et al.	0.06	0.12	-	-	-
Richards et al.	0.071	0.18	-	-	-
			SDSS + UKIDSS		
MLPQNA	0.008	0.11	0.027	0.11	0.040
			SDSS + GALEX + UKIDSS		
MLPQNA	0.005	0.087	0.022	0.088	0.032
			SDSS + GALEX + UKIDSS + WISE		
MLPQNA	0.004	0.069	0.020	0.069	0.029

Table 5. Comparison in terms of outliers percentages among the different references. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; Column 2-3 are fractions of outliers at different σ based on $\Delta z = (z_{spec} - z_{phot})$; Column 4-5 are the fractions of outliers at different σ based on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$. The column 4 reports our catastrophic outliers, defined as $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$.

Exp	Outliers ($\lvert\Delta z\rvert$)		Outliers ($\lvert\Delta z_{norm}\rvert$)	
	$> 2\sigma(\Delta z)$	$> 4\sigma(\Delta z)$	$> 2\sigma(\Delta z_{norm})$	$> 4\sigma(\Delta z_{norm})$
	SDSS			
MLPQNA	7.68	0.38	6.53	1.24
Bovy et al.		0.51		
	SDSS + GALEX			
MLPQNA	4.88	1.61	4.57	1.37
Bovy et al.		1.86		
	SDSS + UKIDSS			
MLPQNA	4.00	1.73	3.82	1.38
Bovy et al.		1.92		
	SDSS + GALEX + UKIDSS			
MLPQNA	2.86	1.47	3.05	0.23
Bovy et al.		1.13		
	SDSS + GALEX + UKIDSS + WISE			
MLPQNA	2.57	0.87	2.88	0.91

Different Machine Learning methods of different complexity (MLPQNA is conceptually simpler than WGE) lead to similar results with a slight edge for MLPQNA

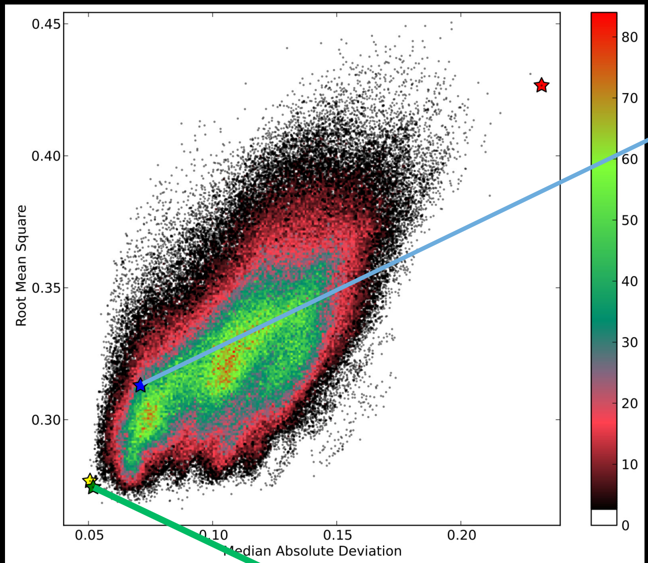
Photometric redshifts for QSO's … a data driven approach

(from K. Polsterer, Heidelberg, 2015)

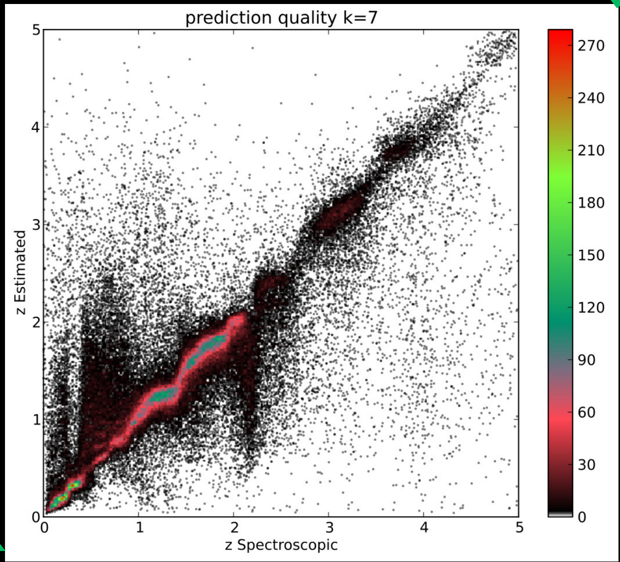
$$\frac{n!}{(n-r)!r!} = 341{,}055 \quad \text{combinations}$$

One does not know a-priori which features are the most relevant

Use all 55 significant photometric features to select the most significant 4



Laurino et al.
Traditional feature selection



Best combination

$u_{model} - g_{model}$
g_{psf}-r_{model}
z_{psf}-r_{model}
i_{psf}-z_{model}

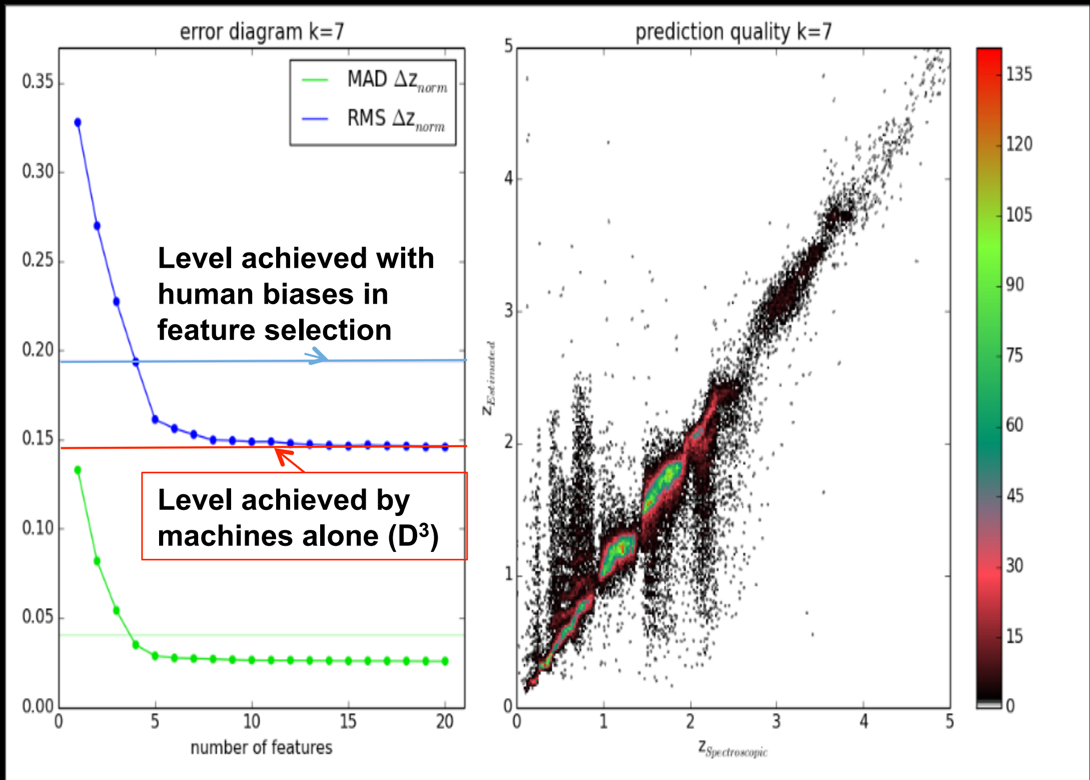
Results comparable to Brescia et al. 2014

Is it possible to do better ?

Photometric redshifts for SDSS QSO

PSF, Petrosian, Total magnitudes + extinction + errors ….. 585 features…. Let us find the best combination of 10, 11, 12

For just 10 features ….. 1,197,308,441,345,108,200,000 combinations



You hit a plateau at 10 features.

Accuracy twice better

These 10 features do not make sense to an astronomer

$$u_{psf} - g_{petr}$$
$$dered(z_{pdf}) - dered(i_{petr})$$
$$dered(g_{psf}) - dered(r_{mod})$$
$$dered(r_{psf}) - dered(z_{mod})$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{model}}}$$
$$dered(r_{mod}) - dered(i_{mod})$$
$$i_{psf} - i_{petr}$$
$$dered(z_{psf}) - dered(r_{petr})$$
$$g_{mod} - g_{petr}$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{petr}}}$$

$u_{psf} - g_{petr}$

$dered(z_{pdf}) - dered(i_{petr})$

$dered(g_{psf}) - dered(r_{\mathrm{mod}})$

$dered(r_{psf}) - dered(z_{\mathrm{mod}})$

$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{\mathrm{mod}el}}}$

$dered(r_{\mathrm{mod}}) - dered(i_{\mathrm{mod}})$

$i_{psf} - i_{petr}$

$dered(z_{psf}) - dered(r_{petr})$

$g_{\mathrm{mod}} - g_{petr}$

$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{petr}}}$

Afterwards … astronomers may find explanations ….
(Capak, private comm.)

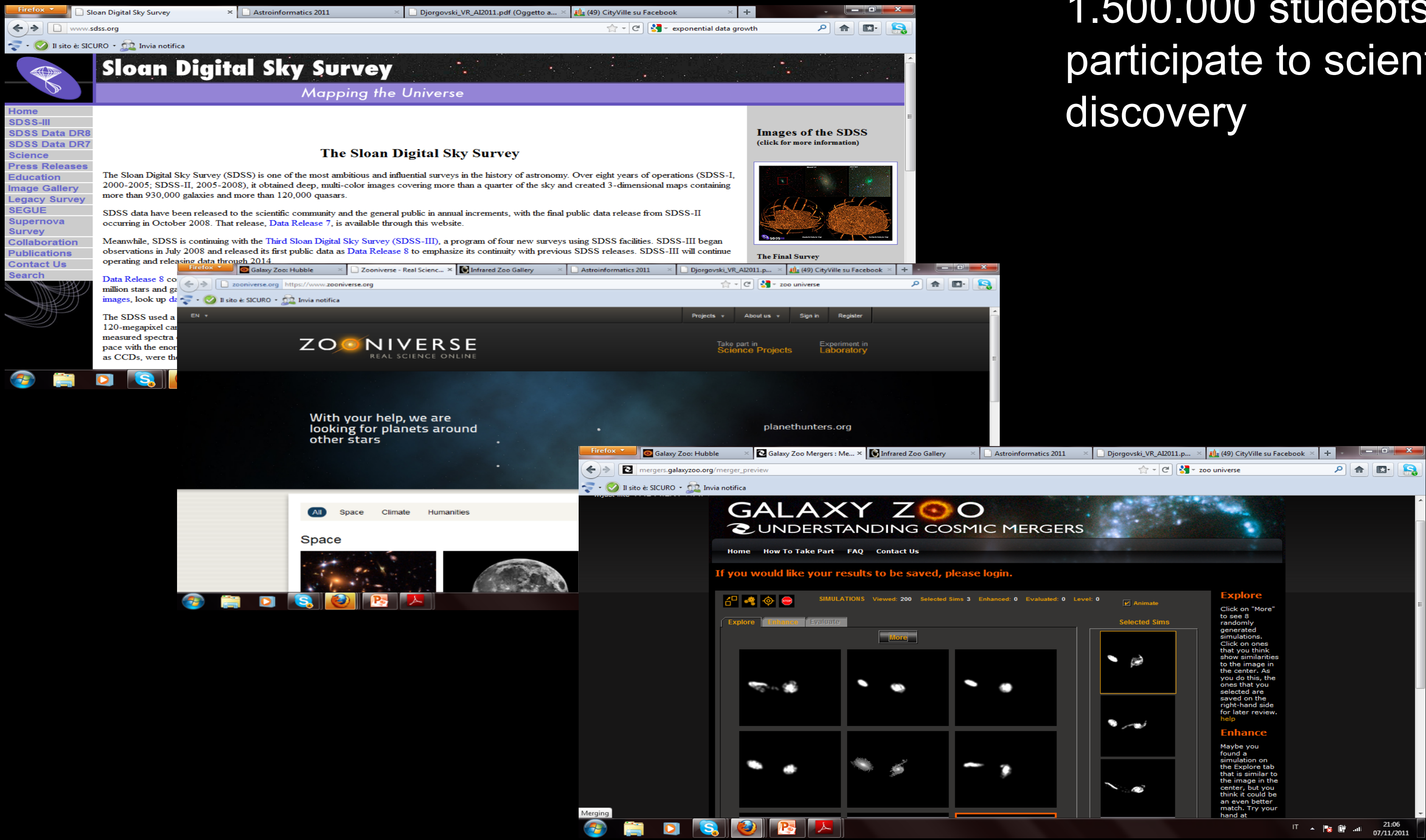
Filter leaks, etc…

Lesson to be learned

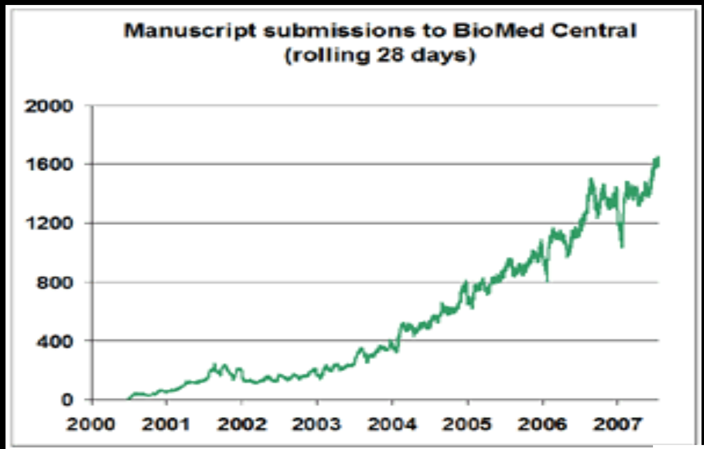
Features which carry most of the information are not those usually selected by the astronomer on the basis of his/her personal experience….

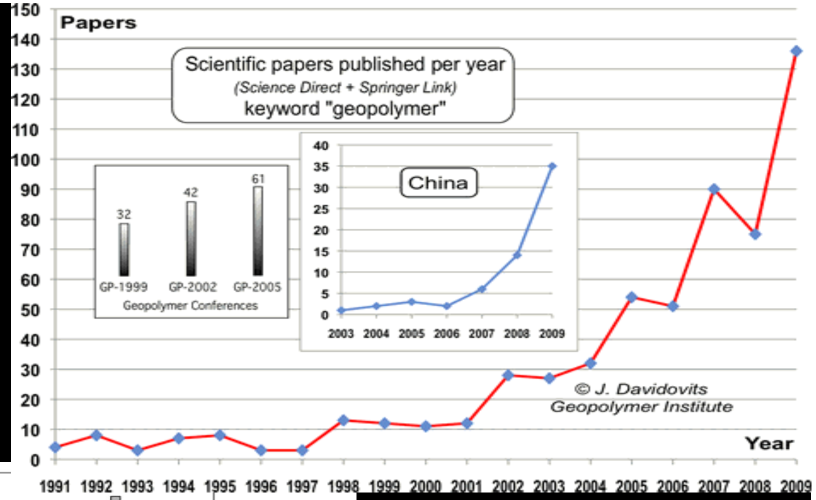
Let the data speak for themselves ?

Crowd sourcing, citizen science, etc

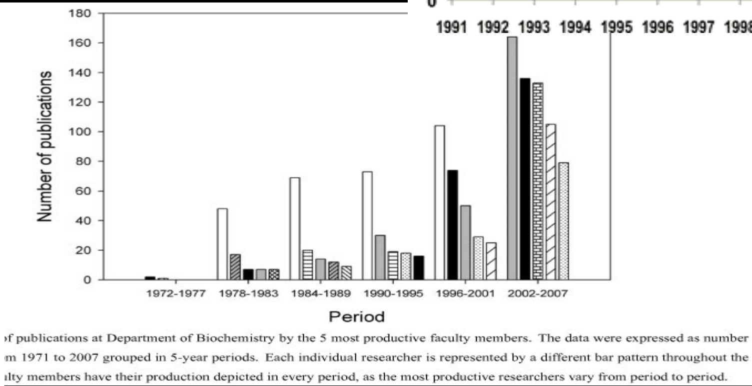


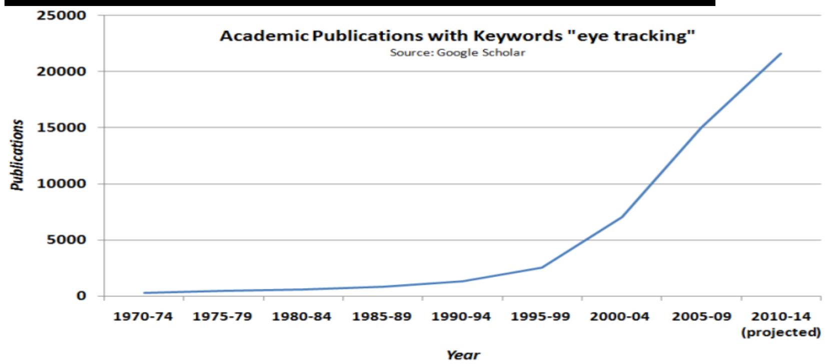
1.500.000 studebts participate to scientific discovery





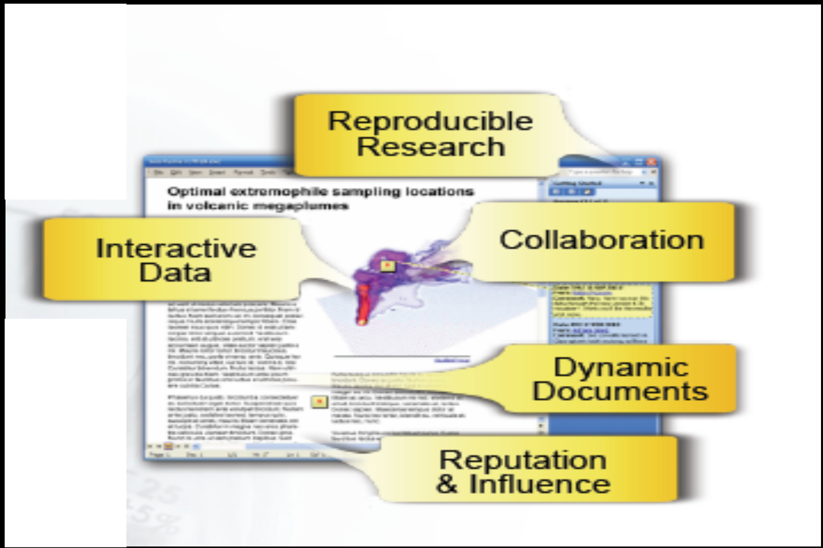
Papers published in specific fields





Most of them will never be read… unless

New avenues for sharing (publishing) results



Sdynamical publications (tailored on the user's needs) …

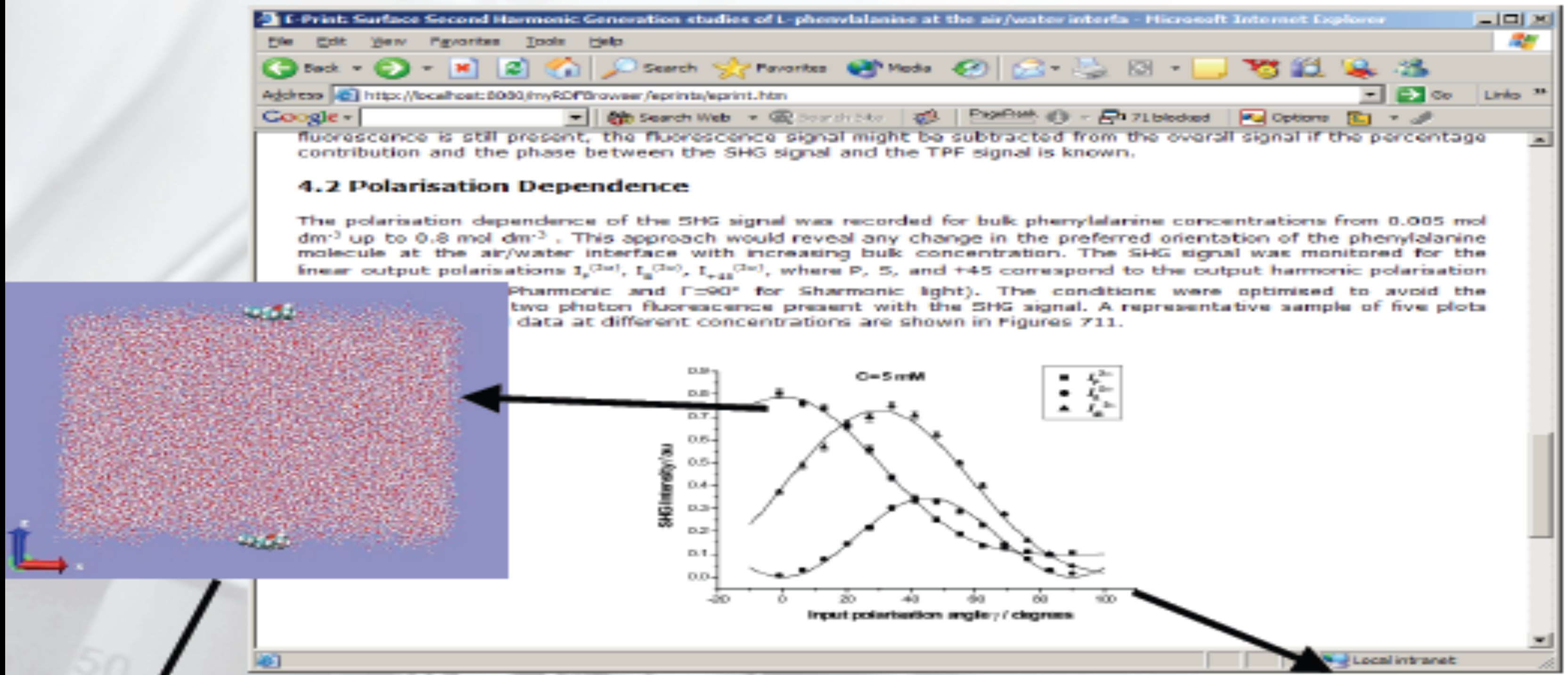
… including research workflows and laboratory results

… can be included in work-benchs to allow repeatability

… Direct real time comparison of similar works

… possibility to apply new tools on same data to validate results

Publications as Live Documents

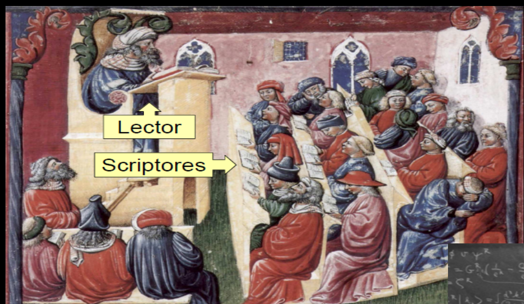


Link to simulation software and data in archive

Link to data, follow links back to the raw data archive

Sociology

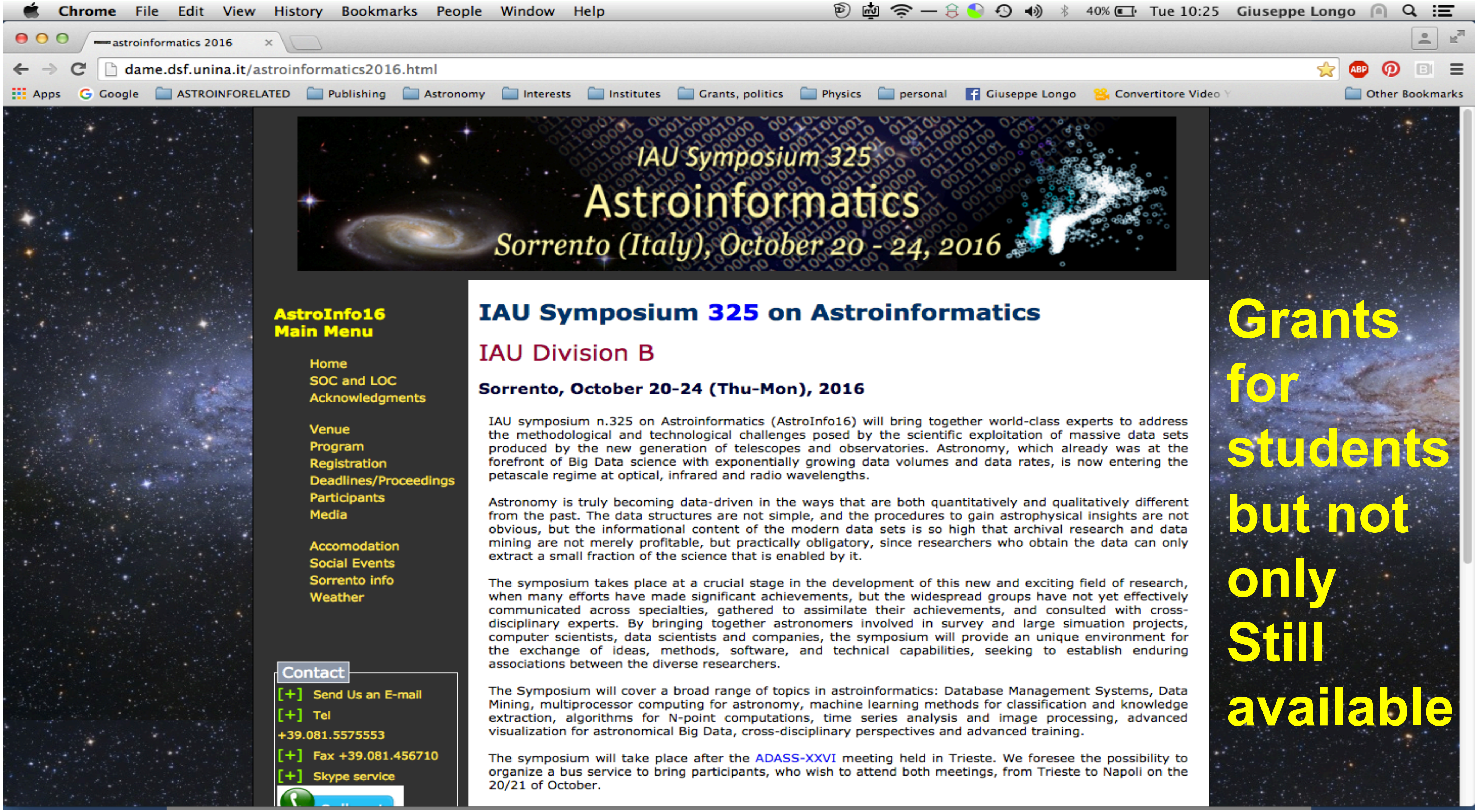
Time scale of scientific change	1-3 years
Time scale of formation process	20 years
Time scale of a career \	50 years
Time scale of academic change	100 – 400 years



XI secolo

XXI secolo

Human knowledge is now available in cyberspace and can be finely tuned to your needs



IAU Symposium 325 on Astroinformatics

IAU Division B

Sorrento, October 20-24 (Thu-Mon), 2016

IAU symposium n.325 on Astroinformatics (AstroInfo16) will bring together world-class experts to address the methodological and technological challenges posed by the scientific exploitation of massive data sets produced by the new generation of telescopes and observatories. Astronomy, which already was at the forefront of Big Data science with exponentially growing data volumes and data rates, is now entering the petascale regime at optical, infrared and radio wavelengths.

Astronomy is truly becoming data-driven in the ways that are both quantitatively and qualitatively different from the past. The data structures are not simple, and the procedures to gain astrophysical insights are not obvious, but the informational content of the modern data sets is so high that archival research and data mining are not merely profitable, but practically obligatory, since researchers who obtain the data can only extract a small fraction of the science that is enabled by it.

The symposium takes place at a crucial stage in the development of this new and exciting field of research, when many efforts have made significant achievements, but the widespread groups have not yet effectively communicated across specialties, gathered to assimilate their achievements, and consulted with cross-disciplinary experts. By bringing together astronomers involved in survey and large simuation projects, computer scientists, data scientists and companies, the symposium will provide an unique environment for the exchange of ideas, methods, software, and technical capabilities, seeking to establish enduring associations between the diverse researchers.

The Symposium will cover a broad range of topics in astroinformatics: Database Management Systems, Data Mining, multiprocessor computing for astronomy, machine learning methods for classification and knowledge extraction, algorithms for N-point computations, time series analysis and image processing, advanced visualization for astronomical Big Data, cross-disciplinary perspectives and advanced training.

The symposium will take place after the ADASS-XXVI meeting held in Trieste. We foresee the possibility to organize a bus service to bring participants, who wish to attend both meetings, from Trieste to Napoli on the 20/21 of October.

Some final thoughts

This an era of profound changes in technology, methodology, objectives and strategies. 1.5 billion USD invested in the next 5 years.

E-Science will become more and more important in the coming years. Scientists of the future will be obliged to have an in-depth understanding of these technologies.

Better Interfacing between humans (scientists) and computing infrastructures will become crucial.

Data Driven Science is still in its infancy but it is clear that it opens a whole new range of possibilities and discoveries, but it is also clera that it calls for a re-thinking of the way we collect and analyse data

Academy is beginning to adjust but it does so very slowly and in a non effective way


Would you rather have taken the blue pill? ….
Thanks for listening