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. CD³ – 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
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.

*Toulouse – MAESTRO School – July 6-th 2016*
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…
Sorry to burst your bubble, dudes, but yes, that’s the future.

No offense, Future Man, but is everyone in your time retarded?

And “upload” pictures of their breakfasts to a “facebook”?

And other people will look at the pictures and make comments?

So, Future Man, you are saying people will “tweet” what they are eating for breakfast?
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.…

Data of the Internet of Things

- **BrontoByte**: The digital universe of tomorrow, 2020
- **YottaByte**: The digital universe today: 250 trillion DVD’s
- **ZettaByte**: In 2016, 1.3 ZB will cross our digital networks daily
- **ExaByte**: At the moment, every day 1 EB of data is created on the internet, that is the equivalent of 250 million DVD’s
- **PetaByte**: The CERN LHC generates 1 PB per second

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)*

**The Third Paradigm**
Numerical simulations
*(early 40’s)*

**The Fourth Paradigm**
Data Driven Discovery
*(Now)*

\[ \nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon} \quad \text{(Gauss' Law)} \]
\[ \nabla \cdot \mathbf{H} = 0 \quad \text{(Gauss' Law for Magnetism)} \]
\[ \nabla \times \mathbf{E} = -\frac{\partial \mathbf{H}}{\partial t} \quad \text{(Faraday's Law)} \]
\[ \nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{E}}{\partial t} \quad \text{(Ampere’s Law)} \]
Big data accordingly to Borat

Big Data is anything 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

MOST, IF NOT ALL, ASTRONOMICAL DATABASES

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

IVOA
INTERNATIONAL VIRTUAL OBSERVATORY ALLIANCE

The Virtual Observatory
Data re-use: a market fact

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.
Virtual Observatory Science Examples

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
What should we do to extract patterns (i.e. laws or ordering relationships) in a \( \mathbb{R}^n \) space (\( n >> 100 \))?
Traditional way to look for candidate QSO in 3 band survey

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…
Artificial Intelligence – SF or truth?

Watson - IBM Jeopardy

Deep blue - IBM Automatic driving

Automatic landing A.I. failure but… how about Human Intelligence failure?
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

1/160,000 of the sky, moderately deep (25.0 in r)
55,000 detected sources (0.75 mag above $m_{\text{lim}}$)

Multiwavelength Digital Surveys

Band 1
Band 2
Band 3
Band n

Calibrated data

30 arcmin

Measure attributes (brightness, position, shapes, etc.) of detected sources
Each object becomes a record $p^i$ defined by $n$ parameters (or features)

$$p^i \equiv \left\{ t, \lambda_1, \left\{ f_1^1, f_2^1, \ldots, f_k^1 \right\}, \lambda_2, \left\{ f_1^2, f_2^2, \ldots, f_k^2 \right\}, \ldots, \lambda_n, \left\{ f_1^n, f_2^n, \ldots, f_k^n \right\} \right\}$$

Hence an observation is a point in a high dimensionality parameter space
The history of astronomical discoveries can be reconstructed in terms of better coverage or sampling of the Parameter Space.
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 x N x I x D
Expectation Maximisation: K x N x I x D^2
Monte Carlo Cross-Validation: M x K_{max}^2 x N x I x D^2
Correlations ~ N log N or N^2, ~ D^k (k ≥ 1)
Likelihood, Bayesian ~ N^m (m ≥ 3), ~ D^k (k ≥ 1)
SVM > ~ (NxD)^3

Lots of computing power
Parallelization?
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

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

Extensions
- DAME-KNIME
- ML Model plugin

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
Some Thoughts About e-Science

• **Computational** science ≠ **Computer** 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

\[ 1 + z = \frac{\lambda_{\text{obs}}}{\lambda_0} \approx \frac{v}{c} \]

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

**Crucial cosmological probe**
- Large scale structure
- Weak lensing
- Tests of cosmological models

QSO; $z=3.81$  
QSO; $z=5.31$
Mathematically simple: to find the mapping function $f(x) \rightarrow y$, where: $x \in \mathbb{R}^n$, $y \in \mathbb{R}$

- **Input vector**
  $\{X_j \{x_1, ..., x_n\} j = 1, ... m\} \in OPPS \subset \mathbb{R}^n$

- **Target vector**
  $\{Y_j \{x_1, ..., x_n\} \in OPPS \subset \mathbb{R}^n$

**Physical redshift**

- KB from VO set of templates
- Mapping function
- Phot-z’s

**OPPS** = Observable Photometric Parameter Space

**OSPS** = Observable Spectroscopic Parameter Space

**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

3x10^6 galaxies

The SDSS Genealogy

Terra Server

SDSS SkyServer

- Cinco Space
- Lifes Under Your Feet
- CASJobs MyDB
- GalaxyZoo
- Turbulence DB
- Super COSMOS
- Hubble Legacy Arch
- SkyQuery

- IJHU 1K Genomes
- GALEX
- Pan-STARRS
- Millennium
- Palomar QUEST

- UKIDDS
- VOServices
- Open SkyQuery

- INDRA Simulation
- Potsdam
- Milky Way Laboratory
- MHD DB
- VO Footprint
- VO Spectrum
SDSS – Data Release 10

OPPS
3x10^8 objects
> 100 features
> 100 flags

OSPS
3x10^6

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

Spectroscopic Knowledge Base

WGE: Weak Gated Expert

Data from the unresolved objects SDSS catalogue
Figure 15. In the upper panel, it is shown the scatter plot of the spectroscopic versus photometric redshifts evaluated with the WQ method for the members of the KiDS 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 versus photometric redshifts evaluated with the WQ method for the quasars extracted from the SDSS catalogue with optical and UV photometry. The vertical dashed lines represent the redshift at which the most luminous emission lines characteristic of quasar spectra shift off the SDSS photometric limits due to metallicity. Most of the features of the plots are associated to one or more of those lines.
Photo-z’s for SDSS QSO’s with MLPQNA

<table>
<thead>
<tr>
<th>Survey</th>
<th>Bands</th>
<th>Name of features</th>
<th>Synthetic description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALEX</td>
<td>near, far</td>
<td>mag, magAper, magAper_2 magAper_3 magAper_5</td>
<td>Near and far UV total and imaged maps plus through 2, 4.5 and 7.75 arcsec aperture magnitudes and FWHM radius in units of A or B</td>
</tr>
<tr>
<td>SDSS</td>
<td>u, g, r, i, z</td>
<td>peMag</td>
<td>FTF fitting magnitude in the u, g, r, i, z bands</td>
</tr>
<tr>
<td>UKIDSS</td>
<td>Y, J, H, K</td>
<td>F TF fitting magnitude in Y, J, H, K bands; aperture photometry through 2, 3.8 and 5.7'' circular aperture in each band; Callivari magnitude within circular aperture radius and Petrosian magnitude in Y, J, H, K bands</td>
<td></td>
</tr>
<tr>
<td>WISE</td>
<td>W1, W2, W3, W4</td>
<td>W1, W2, W3, W4</td>
<td>Magnitudes measured with profile fitting photometry at the 50% level, brightness upper limit if the flux measurement has SNR &lt; 2</td>
</tr>
</tbody>
</table>


Lenghty feature selection procedure

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{\text{clean}}(\Delta z_{\text{norm}})$ and $\text{NMAD}(\Delta z_{\text{norm}})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{\text{norm}}| > 2\sigma(\Delta z_{\text{norm}})$. The standard deviation $\sigma_{\text{clean}}(\Delta z_{\text{norm}})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{\text{norm}}| \leq 2\sigma(\Delta z_{\text{norm}})$.
Different Machine Learning methods of different complexity (MLPQNA is conceptually simpler than WGE) lead to similar results with a slight edge for MLPQNA.

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 NMAID calculated on $\Delta z_{\text{norm}} = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$ related to the test sets. For the definition of the parameters and for discussion see text.

<table>
<thead>
<tr>
<th>Exp</th>
<th>$BIAS(\Delta z_{\text{norm}})$</th>
<th>$\sigma(\Delta z_{\text{norm}})$</th>
<th>MAD$(\Delta z_{\text{norm}})$</th>
<th>RMS$(\Delta z_{\text{norm}})$</th>
<th>NMAID$(\Delta z_{\text{norm}})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDSS</td>
<td>0.002</td>
<td>0.15</td>
<td>0.028</td>
<td>0.017</td>
<td>0.008</td>
</tr>
<tr>
<td>Lauer et al.</td>
<td>0.005</td>
<td>0.16</td>
<td>0.041</td>
<td>0.19</td>
<td>-</td>
</tr>
<tr>
<td>Ball et al.</td>
<td>0.006</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Richards et al.</td>
<td>0.115</td>
<td>0.28</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SDSS + GALEX</td>
<td>0.012</td>
<td>0.11</td>
<td>0.029</td>
<td>0.11</td>
<td>0.003</td>
</tr>
<tr>
<td>Lauer et al.</td>
<td>0.008</td>
<td>0.29</td>
<td>0.059</td>
<td>0.11</td>
<td>-</td>
</tr>
<tr>
<td>Ball et al.</td>
<td>0.09</td>
<td>0.12</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Richards et al.</td>
<td>0.071</td>
<td>0.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SDSS + UKIDSS</td>
<td>0.005</td>
<td>0.11</td>
<td>0.027</td>
<td>0.11</td>
<td>0.040</td>
</tr>
<tr>
<td>MLPQNA</td>
<td>0.005</td>
<td>0.067</td>
<td>0.022</td>
<td>0.038</td>
<td>0.032</td>
</tr>
<tr>
<td>SDSS + GALEX + UKIDSS</td>
<td>0.005</td>
<td>0.067</td>
<td>0.022</td>
<td>0.038</td>
<td>0.032</td>
</tr>
<tr>
<td>MLPQNA</td>
<td>0.004</td>
<td>0.069</td>
<td>0.020</td>
<td>0.009</td>
<td>0.029</td>
</tr>
</tbody>
</table>

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 $\sigma$ based on $\Delta(z) = (z_{\text{spec}} - z_{\text{phot}})$; Column 4-5 are the fractions of outliers at different $\sigma$ based on $\Delta z_{\text{norm}} = (z_{\text{spec}} - z_{\text{phot}})/(1 + z_{\text{spec}})$. The column 4 reports our catastrophic outliers, defined as $|\Delta z_{\text{norm}}| > 2\sigma(\Delta z_{\text{norm}})$.

<table>
<thead>
<tr>
<th>Exp</th>
<th>Outliers ($\Delta z$)</th>
<th>Outliers ($\Delta z_{\text{norm}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$&gt; 2\sigma(\Delta z)$</td>
<td>$&gt; 4\sigma(\Delta z)$</td>
</tr>
<tr>
<td>SDSS</td>
<td>MLPQNA</td>
<td>7.68</td>
</tr>
<tr>
<td></td>
<td>Lauer et al.</td>
<td>0.51</td>
</tr>
<tr>
<td>SDSS + GALEX</td>
<td>MLPQNA</td>
<td>4.88</td>
</tr>
<tr>
<td></td>
<td>Lauer et al.</td>
<td>1.86</td>
</tr>
<tr>
<td>SDSS + UKIDSS</td>
<td>MLPQNA</td>
<td>4.90</td>
</tr>
<tr>
<td></td>
<td>Lauer et al.</td>
<td>1.92</td>
</tr>
<tr>
<td>SDSS + GALEX + UKIDSS</td>
<td>MLPQNA</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td>Lauer et al.</td>
<td>1.13</td>
</tr>
<tr>
<td>SDSS + GALEX + UKIDSS + WISE</td>
<td>MLPQNA</td>
<td>2.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Photometric redshifts for QSO’s … a data driven approach
(from K. Polsterer, Heidelberg, 2015)

One does not know a-priori which features are the most relevant
Use all 55 significant photometric features to select the most significant

Best combination
- \( u_{\text{model}} - g_{\text{model}} \)
- \( g_{\text{psf-r}}_{\text{model}} \)
- \( z_{\text{psf-r}}_{\text{model}} \)
- \( i_{\text{psf-z}}_{\text{model}} \)

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

\[ \mu_{\text{psf}} - g_{\text{petr}} \]
\[ \text{dered}(z_{\text{psf}}) - \text{dered}(i_{\text{petr}}) \]
\[ \text{dered}(g_{\text{psf}}) - \text{dered}(r_{\text{mod}}) \]
\[ \text{dered}(r_{\text{psf}}) - \text{dered}(z_{\text{mod}}) \]
\[ \sqrt{\sigma_{\text{petr}}^2 - \sigma_{\text{model}}^2} \]
\[ \text{dered}(r_{\text{mod}}) - \text{dered}(i_{\text{mod}}) \]
\[ i_{\text{psf}} - i_{\text{petr}} \]
\[ \text{dered}(z_{\text{psf}}) - \text{dered}(r_{\text{petr}}) \]
\[ g_{\text{mod}} - g_{\text{petr}} \]
\[ \sqrt{\sigma_{\text{petr}}^2 - \sigma_{\text{petr}}^2} \]
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 students 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-benches to allow repeatability

… Direct real time comparison of similar works

… possibility to apply new tools on same data to validate results
4.2 Polarisation Dependence

The polarisation dependence of the SHG signal was recorded for bulk phenylalanine concentrations from 0.005 mol dm$^{-3}$ up to 0.8 mol dm$^{-3}$. This approach would reveal any change in the preferred orientation of the phenylalanine molecule at the air/water interface with increasing bulk concentration. The SHG signal was monitored for the linear output polarisations $r_{11}^{(3)}$, $r_{12}^{(3)}$, $r_{13}^{(3)}$, where P, S, and T45 correspond to the output harmonic polarisation $r_{11}^{(2)}$, $r_{12}^{(2)}$, and $r_{13}^{(2)}$ for Harmonic light). The conditions were optimized to avoid the two photon fluorescence present with the SHG signal. A representative sample of five plots data at different concentrations are shown in Figures 711.

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.
Grants for students but not only Still available
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