Management of Astronomical Data Archives and their Interoperability through the Virtual Observatory

Fabio Pasian – INAF

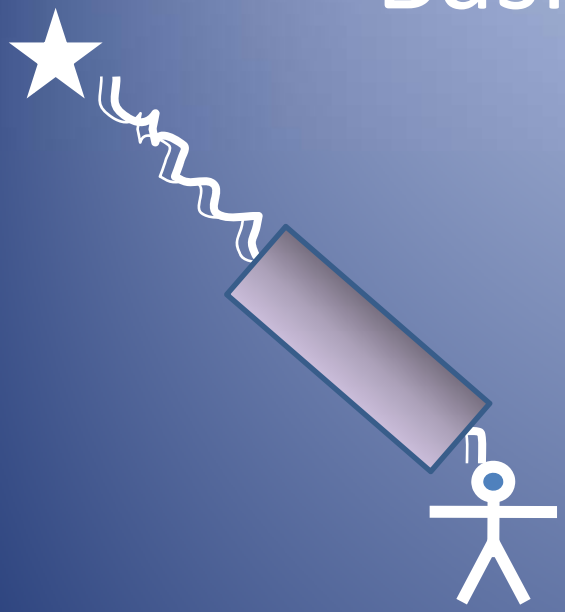
Chair , International Virtual Observatory Alliance

First Workshop on Data Preservation and Long Term Analysis in HEP, DESY, Hamburg, 26-28 Jan 2009

Outline

- Basics on Astronomy
- Astronomy in the XXI century
- Data Archives for Astronomy
- The Virtual Observatory (VObs)
- The VObs from a project's perspective: the EURO-VO experience
- Processing of VObs data
- Generalisation of concepts

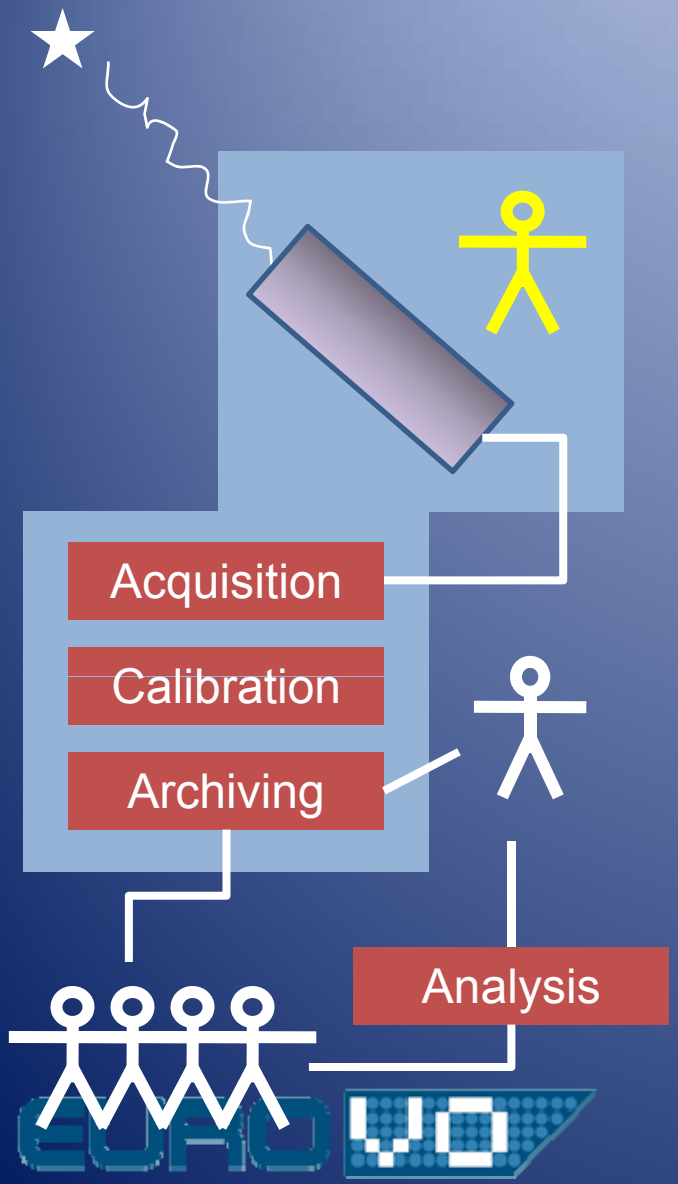
Basics on Astronomy (I)



- Astronomy is an observational science – no repeatability
- Measures I(λ), or I(ν), or I(E)
- We actually observe $I'(\lambda) = I(\lambda)*T(x,\lambda)$

- Most phenomena are in fact variable
- We actually observe $I''(\lambda,t) = I(\lambda,t)*T(x,\lambda)$
- Every single observation must be kept ⇒ preservation!

Basics on Astronomy (II)

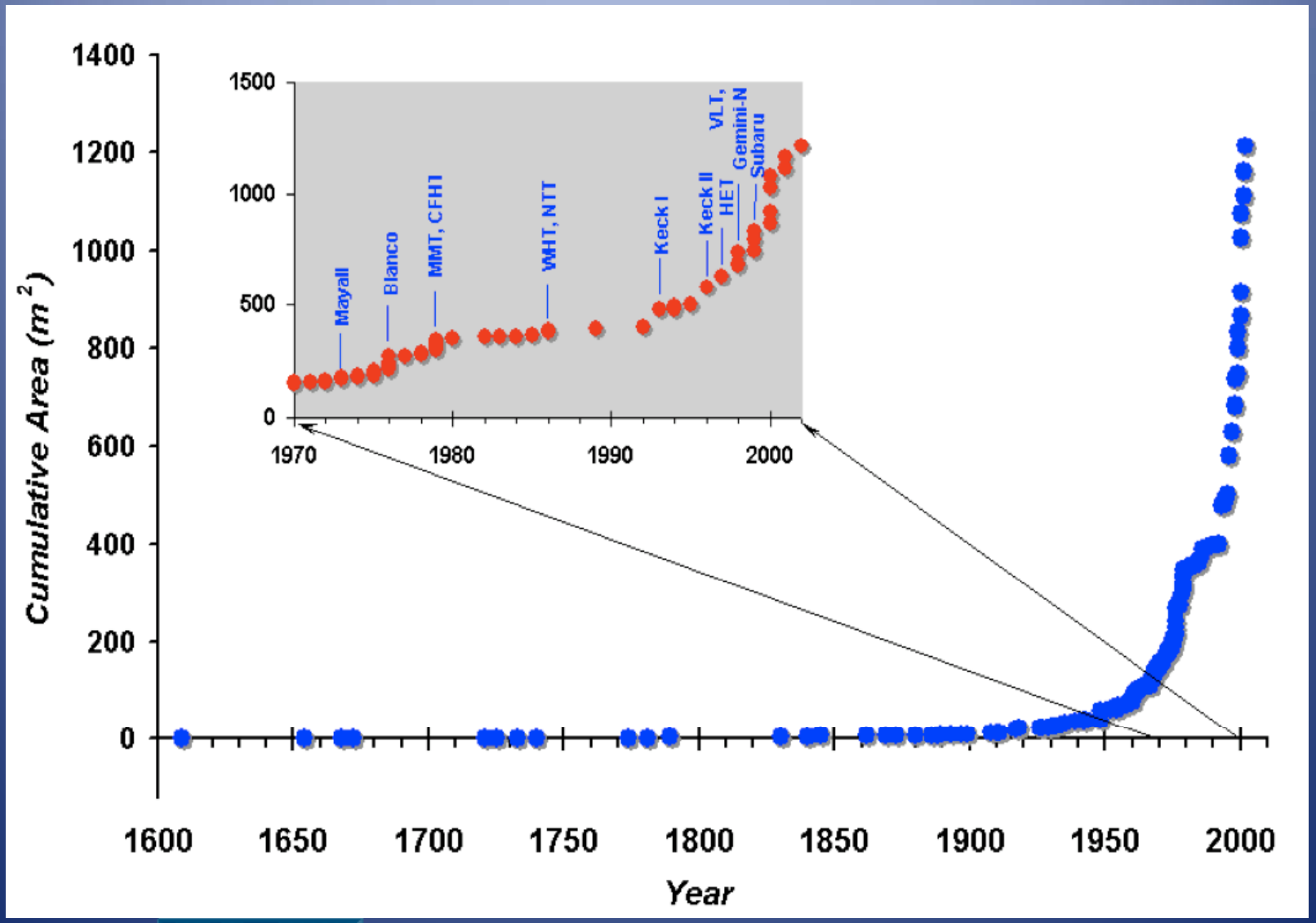


- Service observing due to:
  - complexity of instrumentation
  - optimisation of obs. conditions
- Data handling and basic processing (calibration) c/o observatory
- Owner of observation gets data from archive
- After some time, data become public to be analysed
- Since 1977, there is a standard format for data files (FITS)

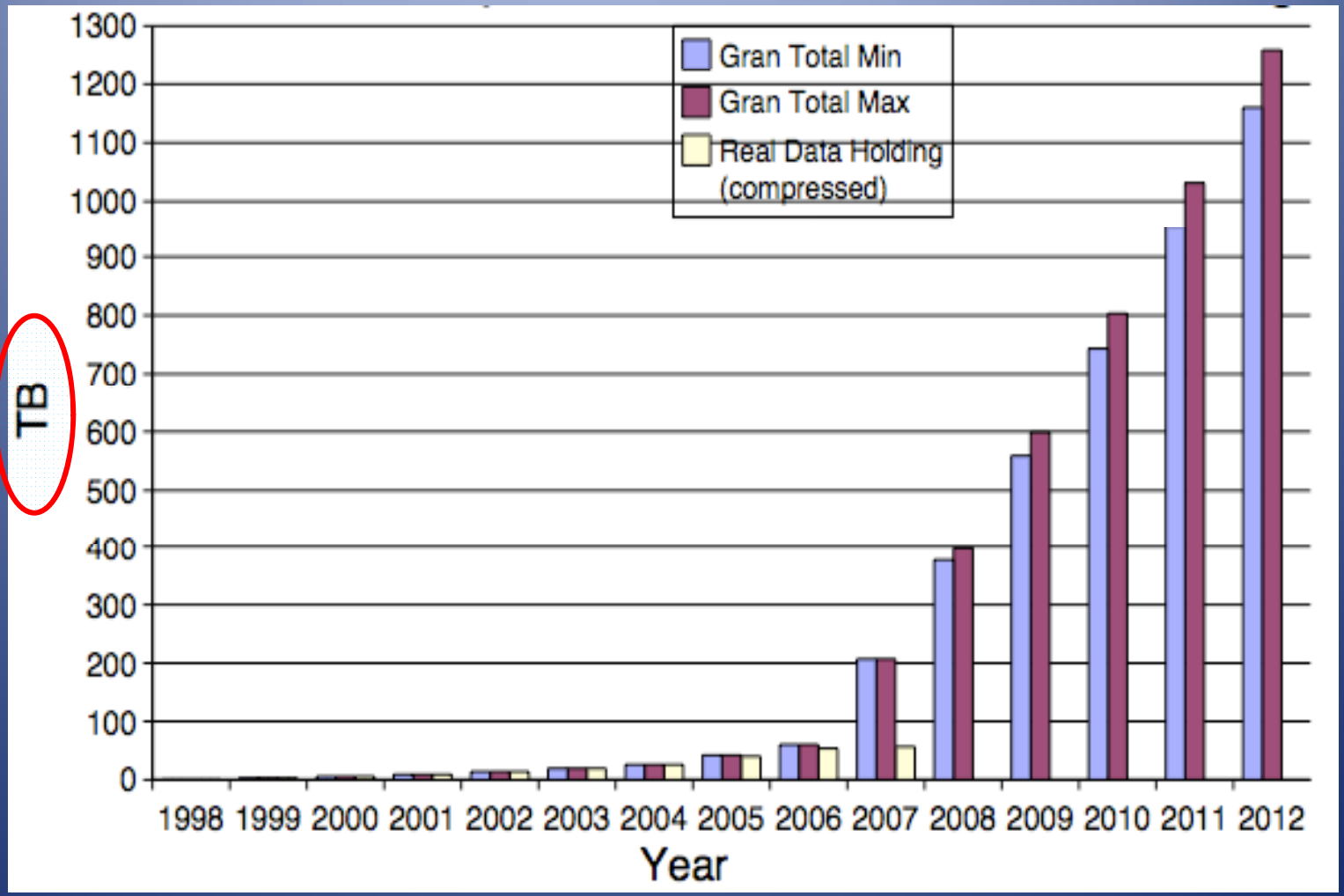
Need for Archives

- Monitor time variability of phenomena (digitization of photographic plates)
- Need to reprocess raw data given better knowledge of instrumental effects
- Compare phenomena in different bands (multi-λ astronomy)
- Increase return for investment (data re-use, educational, outreach, ...)
- Statistical analysis / mining of large quantities of data
- Cope with data avalanche

Telescope Collecting Area Increase



Archive Growth (e.g. ESO)



The way Astronomy works

- Telescopes (ground- and space-based, covering the full electromagnetic spectrum) ⇒ Observatories
- Instruments (telescope/band dependent) ⇒ Observatories/Consortia
- Data analysis software (instrument dependent) ⇒ Observatories/Consortia/Researchers
- *Active* Archives ⇒ Observatories/Agencies
- Publications ⇒ Journals
- Data curation (metadata + tables & catalogues) ⇒ Data curators
- … and Public Outreach ⇒ Observatories/Agencies

The Good News

- Observational data normally stored in *astronomical archives*, freely available *on-line* after ~ 1 year
- Results published in academic journals, all *available on-line* (full content in general freely available after ~ 3 years)
- One *single entry point for journals*: ADS
- Two-way links between most archives and publications
- Data curators (object metadata + catalogues) *on-line*; some links to archives and publications
- Analysis software maintained and made available by Observatories/Archives *on-line*
- Press release and outreach material (pictures, movies) *on-line*

The “not-so-Good” News (I)

- Different astronomical archives have *widely different access/search interfaces and standards/conventions*; serving mainly raw data
- Widely specialized, *complex analysis software* for various sub-branches; steep learning curve, but multi-wavelength is now the norm to produce science
- Publication - Archive links point to *raw, unprocessed data*
- Object *metadata not homogeneously defined*; links with archives and publications not complete
- Press release and *outreach material disconnected*

The “not-so-Good” News (II)

- *Preservation policies* depend on the individual data centres
- *Policies on providing science-ready data* depend on the individual data centres

- *Information avalanche*:
  - ✓ *Huge surveys*: 100M sources at <3k/night ⇒ >100 yr to ID them! (Ever fainter sources, routinely beyond limits of 8 - 10m telescopes [R ≈ 25])
  - ✓ *Huge data collections*: download and data analysis on desktop problematic/impossible (TB dataset: ~ 1 week at 10 Mbps)

The Virtual Observatory

- The Virtual Observatory (VObs) is an innovative, still evolving, system to:
  - take advantage of astronomical data explosion (e.g., use statistical identification to diminish need for a spectrum ⇒ multi-wavelength, multi-parameter analysis)
  - allow astronomers to interrogate multiple data centres in a seamless and transparent way and to utilize at best astronomical data
  - permit remote computing and data analysis
  - foster *new science*

- Web: all documents inside PC; *VObs: all astronomical databases inside PC*
- VObs ⇒ democratization of astronomy!
- 16 national projects world-wide
- All of the above requires the various players to speak the same language ⇒ *VObs standards and protocols defined and adopted within the IVOA (International Virtual Observatory Alliance), which includes national projects*

A&A 424, 545–559 (2004)
DOI: 10.1051/0004-6361:20041153

Astronomy & Astrophysics

Discovery of optically faint obscured quasars with Virtual Observatory tools

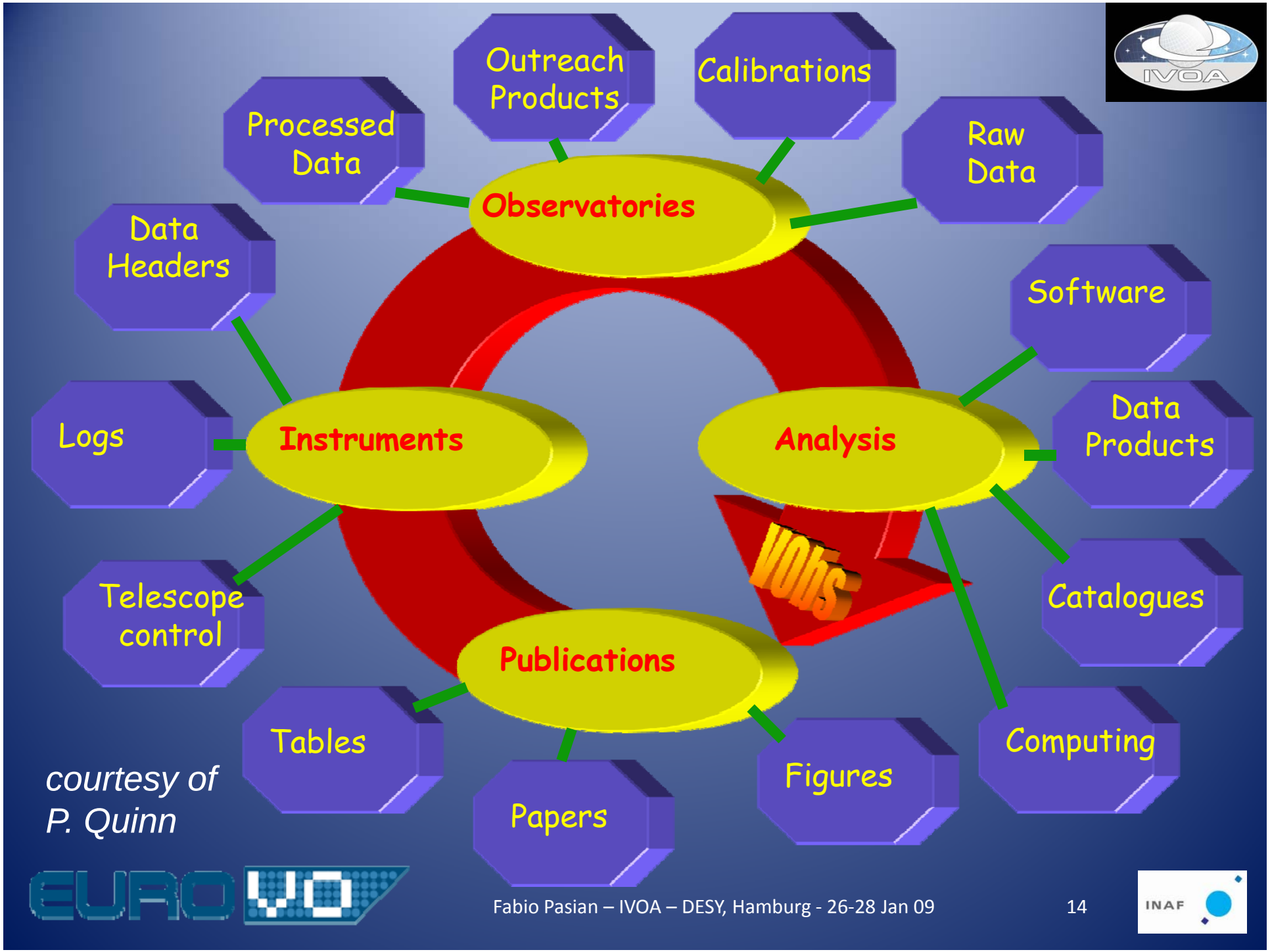
P. Padovani[1], M. G. Allen[2], P. Rosati[3], and N. A. Walton[4]

[1] ST-ECF, European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching bei München, Germany
e-mail: Paolo.Padovani@eso.org
[2] Centre de Données astronomiques de Strasbourg (UMR 7550), 11 rue de l'Université, 67000 Strasbourg, France
e-mail: allen@astro.u-strasbg.fr
[3] European Southern Observatory, Karl-Schwarzschild-Str. 2, 85748 Garching bei München, Germany
e-mail: Piero.Rosati@eso.org
[4] Institute of Astronomy, Madingley Road, Cambridge CB3 0HA, UK
e-mail: naw@ast.cam.ac.uk

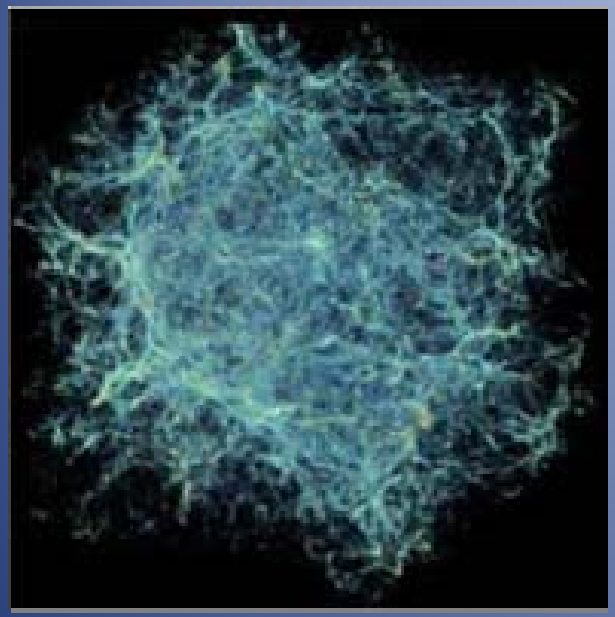
Received 23 April 2004 / Accepted 2 June 2004

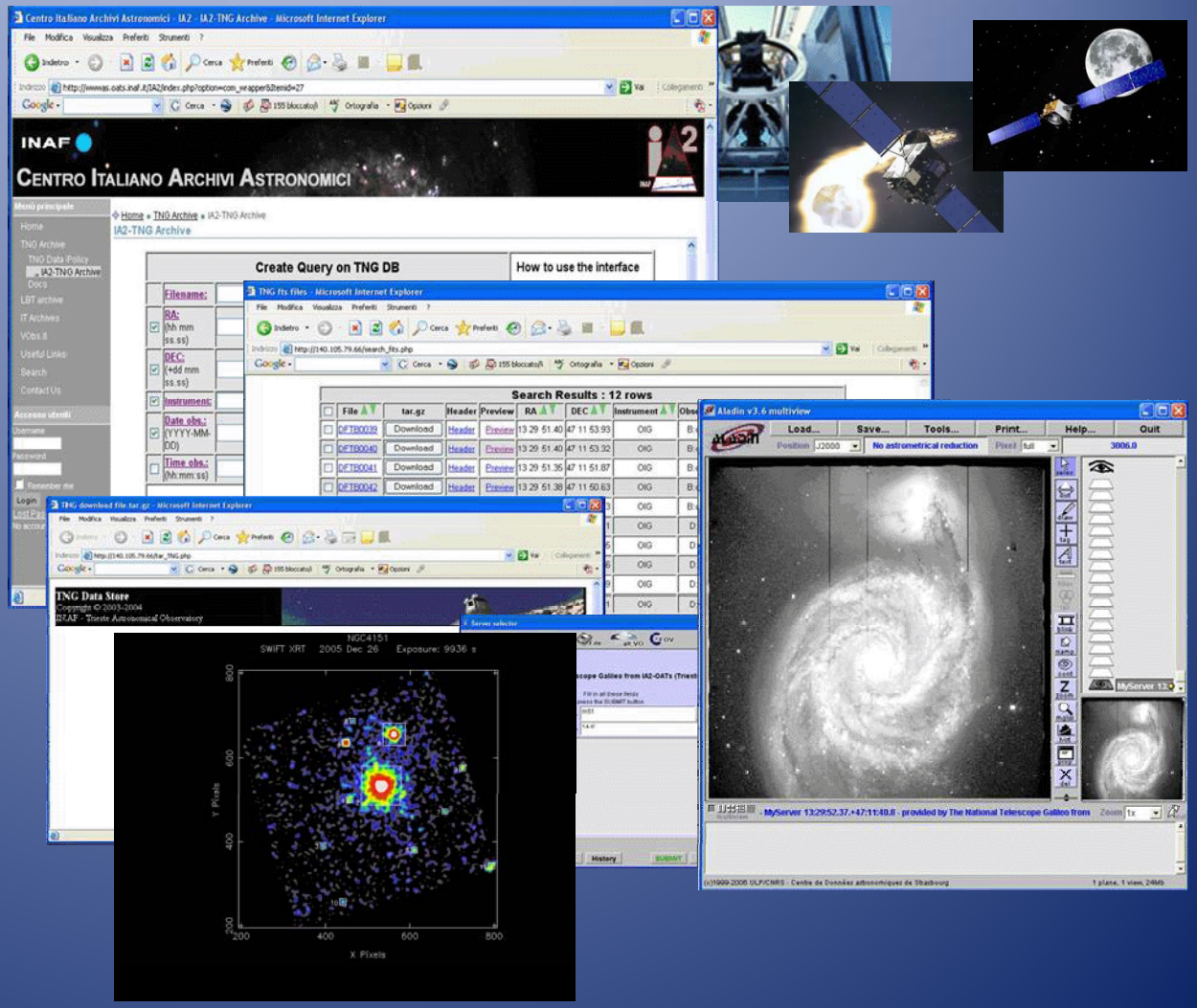
Abstract. We use Virtual Observatory (VO) tools to identify optically faint, obscured (i.e., type 2) active galactic nuclei (AGN) in the two Great Observatories Origins Deep Survey (GOODS) fields. By employing publicly available X-ray and optical data and catalogues we discover 68 type 2 AGN candidates. The X-ray powers of these sources are estimated by using a previously known correlation between X-ray luminosity and X-ray-to-optical flux ratio. Thirty-one of our candidates have high estimated powers ($L_x > 10^{44}$ erg/s) and therefore qualify as optically obscured quasars, the so-called "QSO 2". Based on the derived X-ray powers, our candidates are likely to be at relatively high redshifts, $z \sim 3$, with the QSO 2 at $z \sim 4$. By going ~3 mag fainter than previously known type 2 AGN in the two GOODS fields we are sampling a region of redshift – power space which was previously unreachable with classical methods. Our method brings to 40 the number of QSO 2 in the GOODS fields, an improvement of a factor ~ 4 when compared to the only 9 such sources previously known. We derive a QSO 2 surface density down to 10^{-15} erg cm^{-2} s^{-1} in the 0.5–8 keV band of $\gtrsim$330 deg^{-2}, ~30% of which is made up of previously known sources. This is larger than current estimates and some predictions and suggests that the surface density of QSO 2 at faint flux limits has been underestimated. This work demonstrates that VO tools are mature enough to produce cutting-edge science results by exploiting astronomical data beyond "classical" identification limits ($R \lesssim 25$) with interoperable tools for statistical identification of sources using multiwavelength information.

Key words. astronomical data bases: miscellaneous – methods: statistical – galaxies: quasars: general – X-rays: galaxies


Outreach Products
Calibrations
Processed Data
Raw Data
Observatories
Data Headers
Software
Logs
Instruments
Analysis
Data Products
VObs
Telescope control
Catalogues
Publications
Tables
Computing
Papers
Figures
courtesy of P. Quinn

Theory in the Virtual Observatory





Comparing numerical simulations with observations from ground-based instruments or space-borne experiments

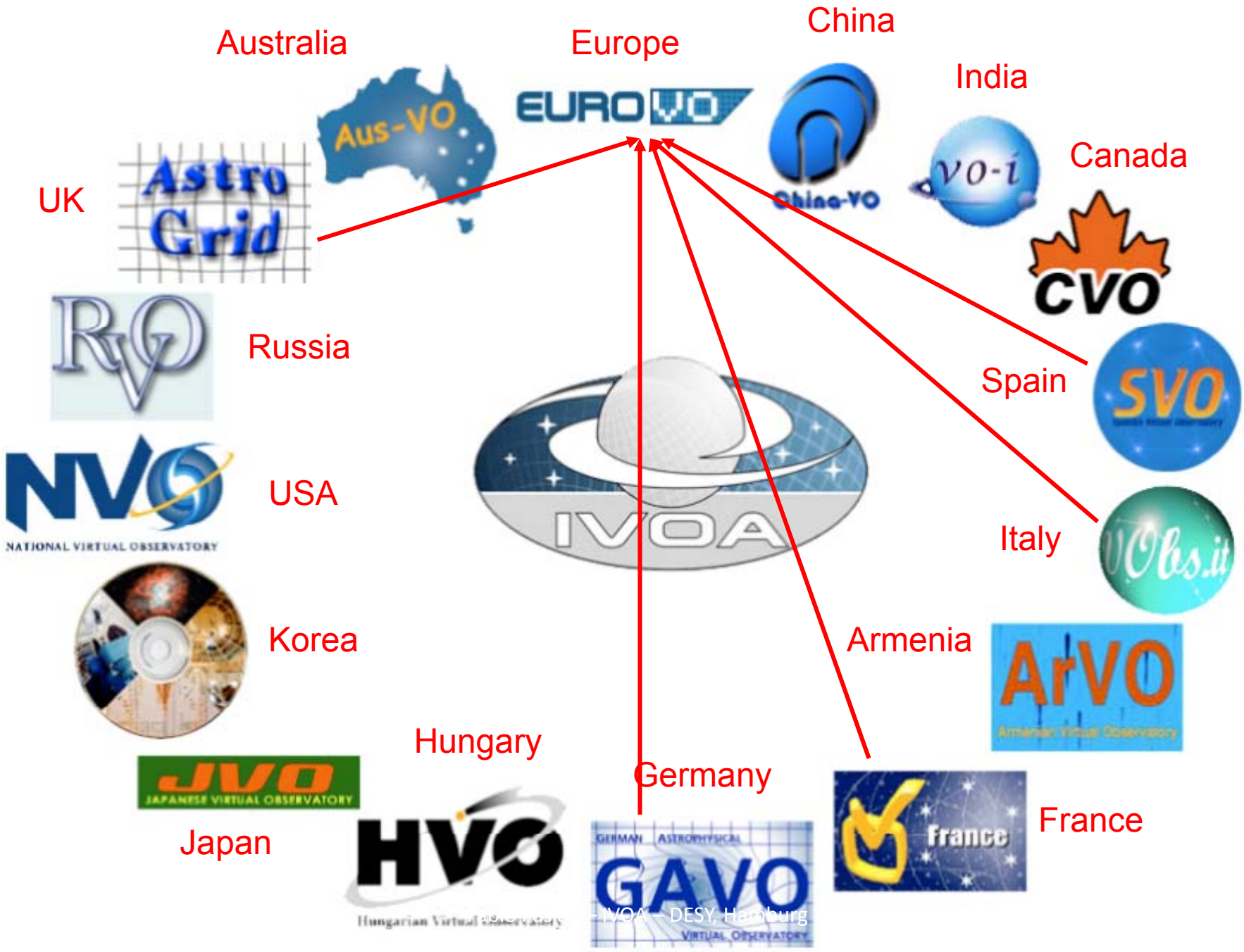
The IVOA: http://ivoa.net

- **Mission:** *"To facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory"*
- Works by telecons, "TWiki" pages, and bi-annual meetings (last one in Baltimore [October 2008], next in Strasbourg [May 2009])
- **Needs:** standardization of data/metadata/sw, data interoperability methods, and list of available services (provided by projects)
- Structure:
  - ✓ IVOA Executive Board includes representatives from all VObs projects
  - ✓ Working and Interest Groups
- Slow convergence on standards: personal / project competition

The IVOA: http://ivoa.net

- Organization: working groups to tackle various aspects
  - ✓ Applications (VObs software)
  - ✓ Data Access Layer (VObs standards for remote data access)
  - ✓ Data Modelling (data characterization)
  - ✓ Data Curation and Preservation (long-term preservation of data)
  - ✓ Grid and Web Services
  - ✓ Resource Registry (VObs resources: “yellow pages”)
  - ✓ Semantics (meaning/interpretation of words, sentences, etc. in astronomy)
  - ✓ VOEvent (definition of immediate event [e.g., GRB])
  - ✓ VObs Query Language (to be used by applications)
  - ✓ VOTable (XML format for VObs data exchange)
- plus Theory and Astronomical Grid (OGF) Interest Groups

16 Member Organizations



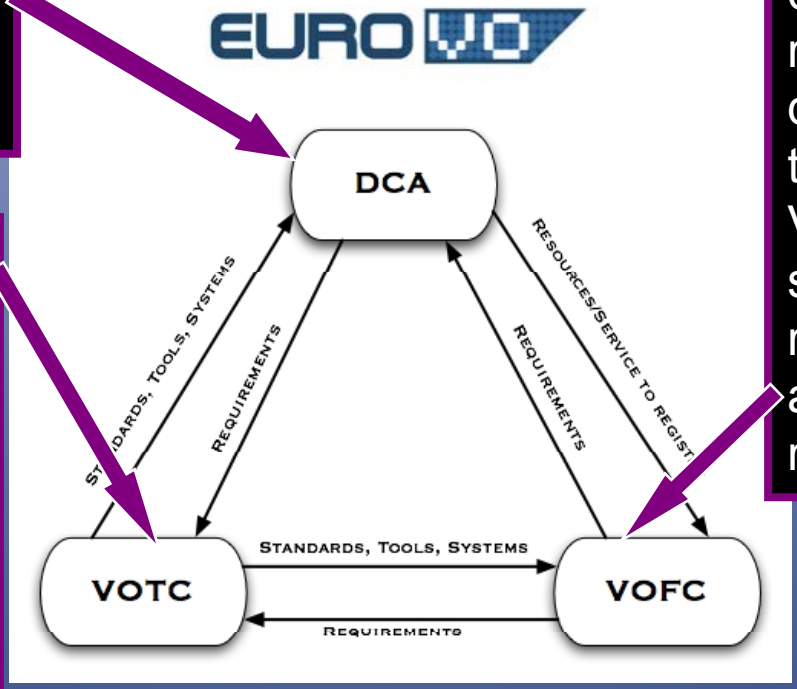
A project's perspective: the EURO-VO

http://www.euro-vo.org

- Successor to the Astrophysical Virtual Observatory (AVO), which was a 5 M€, Phase A study (2001 - 2004/5) on the scientific requirements and technology for building the VObs in Europe, *50%* funded by European Community (Fifth Framework Programme [FP5])
- Includes 8 partners: ESO, European Space Agency (ESA), plus six national nodes: INAF (Italy), INSU (France), INTA (Spain), NOVA (Netherlands), PPARC (UK), and MPG (Germany)
- Partly funded by the EC, but substantial (~ 50%) partner support
- Has three components: Data Centre Alliance, Technology Centre, Facility Centre

An alliance of European data centres who will populate the EURO-VO with data, provide the physical storage and computational fabric and who will publish data, metadata and services to the EURO-VO using VObs technologies

An operational organization, that provides the EURO-VO with a persistent, centralized registry for resources, standards and certification mechanisms as well as community support for VObs technology take-up and scientific programs. EURO-VO's "public face"



A distributed organization that coordinates a set of research and development projects on the advancement of VObs technology, systems and tools in response to scientific and community requirements

The EURO-VO Project (I)

- Data Centre Alliance co-funded by the EC (EuroVO-DCA) at 1.5 M€ level (FP6) for 2.5 yrs since Sept. 2006; *8.5 FTE/yr*. Lead by CDS, Strasbourg, France.
  - Workshops for astronomers and for developers ; coordination
- Technical Centre co-funded by the EC (VO-TECH) at the 3.3 M€ level (FP6) for 4.5 years since Jan. 2005; *21 FTE/yr*. Lead by AstroGrid, UK.
  - "Design Studies", meetings every 6 months
- Facility Centre (FC), located at ESO, co-managed by ESO & ESA; support at "best-effort" level [~ *2 FTE/yr*]

  (but successful FP7 proposal – EuroVO-AIDA)
  - Workshops, Web pages, Research Initiative
  - Selection of EURO-VO Science Advisory Committee

The EURO-VO Project (II)

- The EURO-VO proposal “Astronomical Infrastructure for Data Access (EuroVO-AIDA)” approved within the EC first Framework Programme 7 (FP7) Infrastructure call INFRA-2007-1.2.1 “Scientific Digital Repositories” funded with 2.7 M€; same partners as the EURO-VO. Started Feb 2008.
- Ensures continuation of European-wide VObs activities until 2010
- AIDA is a combination of DCA, VOTECH, and FC activities
- AIDA aims at
  - unifying the digital data collection of European astronomy
  - integrating their access mechanisms with evolving e-technologies
  - enhancing the science extracted from these data-sets
- *VObs is moving worldwide from development to operations*

Data Centres in the VObs Era

- The VObs needs data ⇒ astronomical data centres lie at its foundation
- The VObs is more than a system: also a “frame of mind” ⇒ modern access to better data
- The VObs is “convenient” for data centres as well. Various reasons:
  1. old technology has hard time keeping up with current data volume and complexity
  2. broadens user base
  3. exposes highly processed data in a direct way through VObs protocols

What is a VObs-compliant archive?

The VObs cannot (and does not) dictate how to manage archives

- The VObs requires data centres to have a "VObs layer" to:
  - ✓ "translate" any locally defined parameter to the standard (IVOA compliant) ones (e.g., RA can be called in many different ways)
  - ✓ hide any observatory/telescope/instrument specific detail and work in astronomical units: e.g., *wavelength range/band* (not grism or filter name), *spectral resolution, signal-to-noise ratio, field of view, limiting magnitude* ⇒ provide the correct meta-data (i.e. data about data, data description)
- The VObs will work at best with high level "science-ready" data ⇒ data centres should make an effort to provide such data

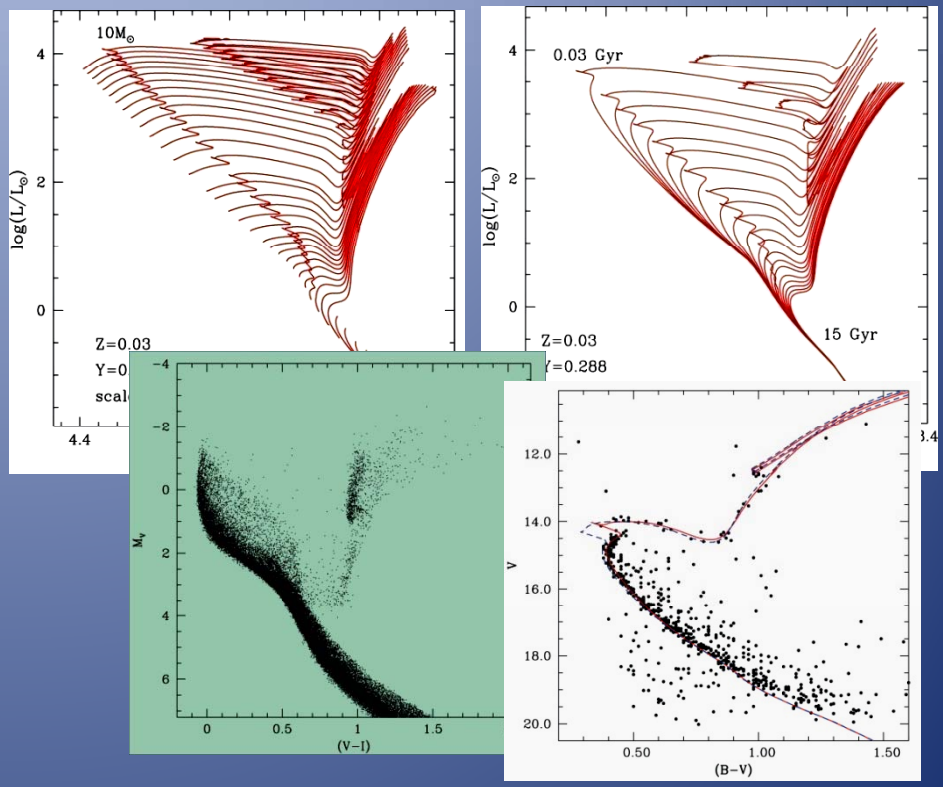
Processing of VObs data

- Examples:
  - Data processing (reduction, calibration, ...)
  - Astro data on demand (e.g. numerical simulations, MC, theory data, calibration on-the-fly, ...)
  - Data mining
- Need for computing power (!)
  - c/o Data Centres
  - on the Grid

Astro data on demand

BaSTI – VObs-compliant tool providing numerical models for evolutionary tracks, isochrones, luminosity functions, synthetic color-magnitude diagrams, tables with relevant data.

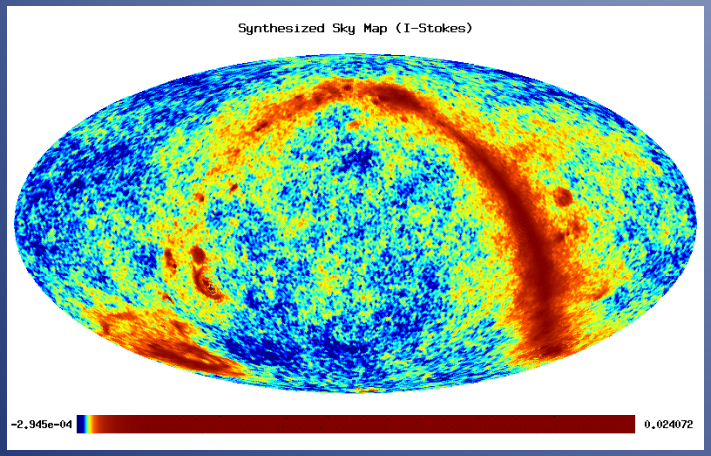
BaSTI is also a database/ centre, which provides numerical models on request to astro users. Model production can be computationally heavy.

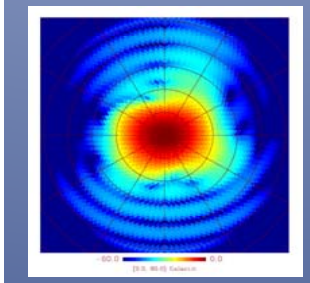


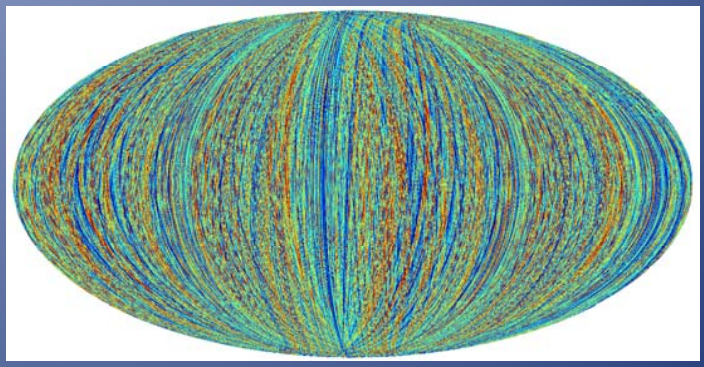
The Virtual Observatory meets the EGEE Grid

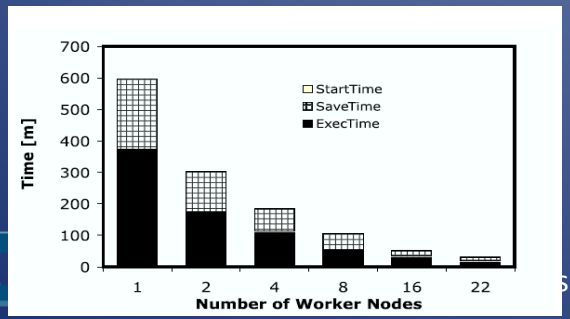
Simulations of the Planck mission

Evaluation of the impact of instrumental systematic effects on the scientific results of the mission (code ported on 400-CPU Planck-LFI DPC local cluster + GILDA, Grid.it, TriGrid, EGEE, CINECA, CSC, DEISA, NERSC).





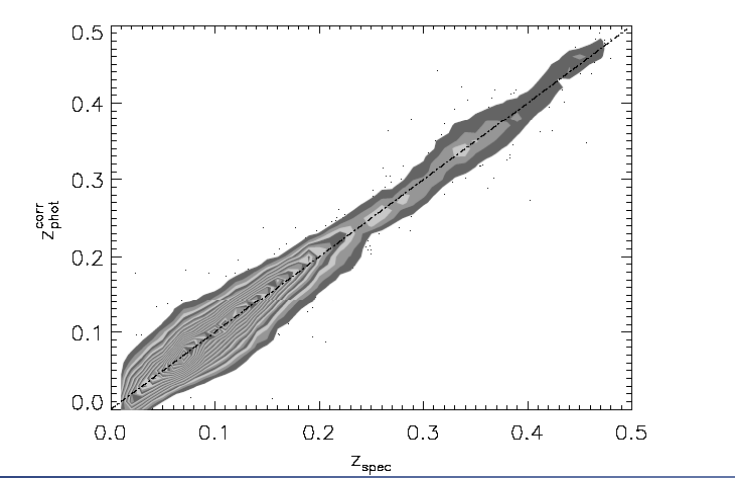




The scalability of running the Planck simulations on EGEE – retrieval of all simulated data depends on network bandwidth

Data mining

VO-Neural / DAME – Evaluation of Photometric redshifts using Neural networks



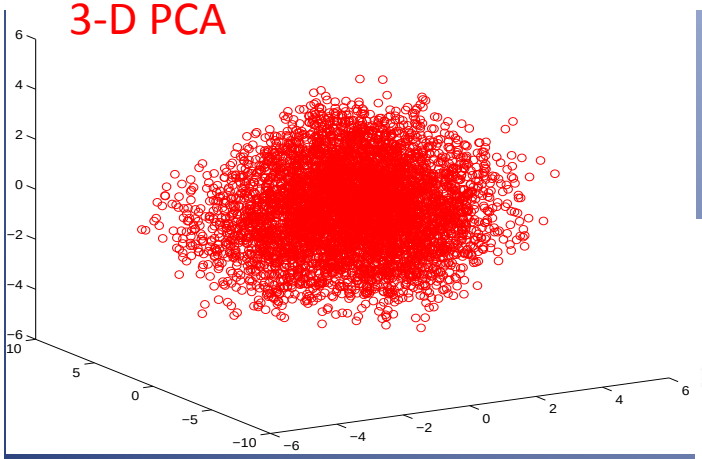
Trend of spectroscopic versus photometric redshifts for the spectroscopic datasets in the Main Galaxy sample (i.e. all galaxies regardless they are LRG's or not). Due to the large number of points to be displayed we show them as isocontours.

Exploit the data wealth of the Sloane Digital Sky Survey to train a super-vised neural network to recognize photometric redshifts.

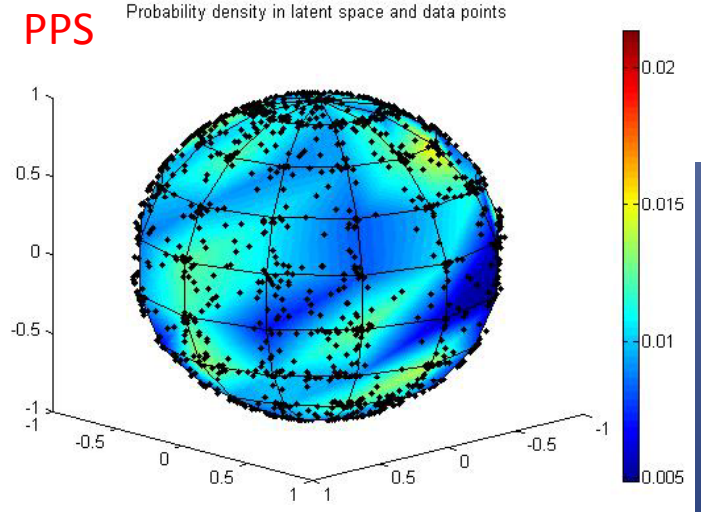
Given the size of the dataset (30 M galaxies) and the complexity of computations, the campus Grid developed at UniNA within the SCoPE project is used to perform the computations, triggered by the user through a GUI

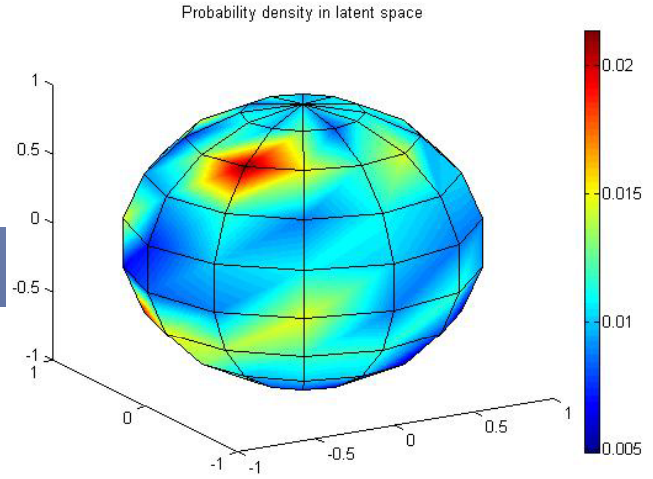
Looking for AGN candidates in SDSS+UKIDS

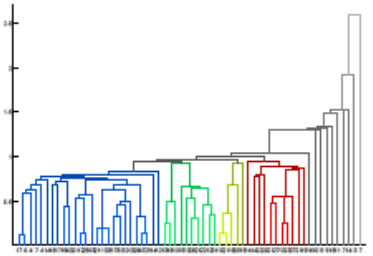
3-D PCA

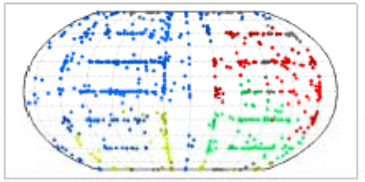


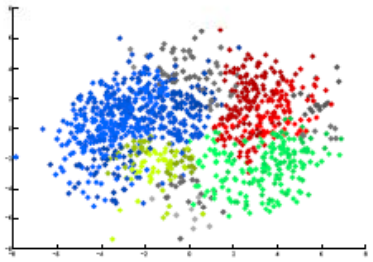
PPS











c)

Figure 9. NEC colored dendrogram.PPS 2-dimensional map and labeling.MDS 2-dimensional projection and labeling.

VObs-Grid Integration

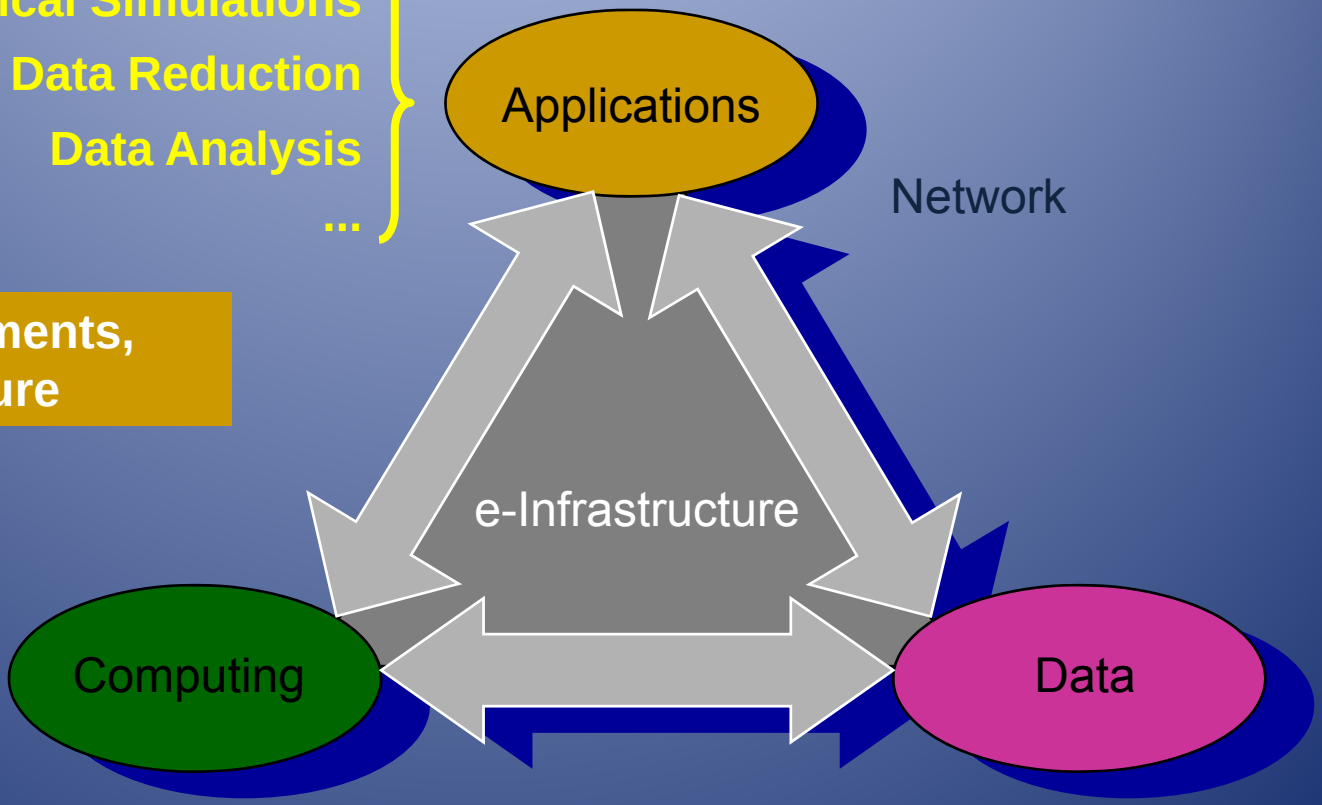
VObs	⇒	Grids
• Single-sign-on		• Authentication & Authorization
• VOSpace		• Data Management
• Workflows		• Job Management
• Information System (Registries)		• Information system

... plus development of a “native” way of accessing databases from the Grid through a Query Element (similar in structure to the CE).

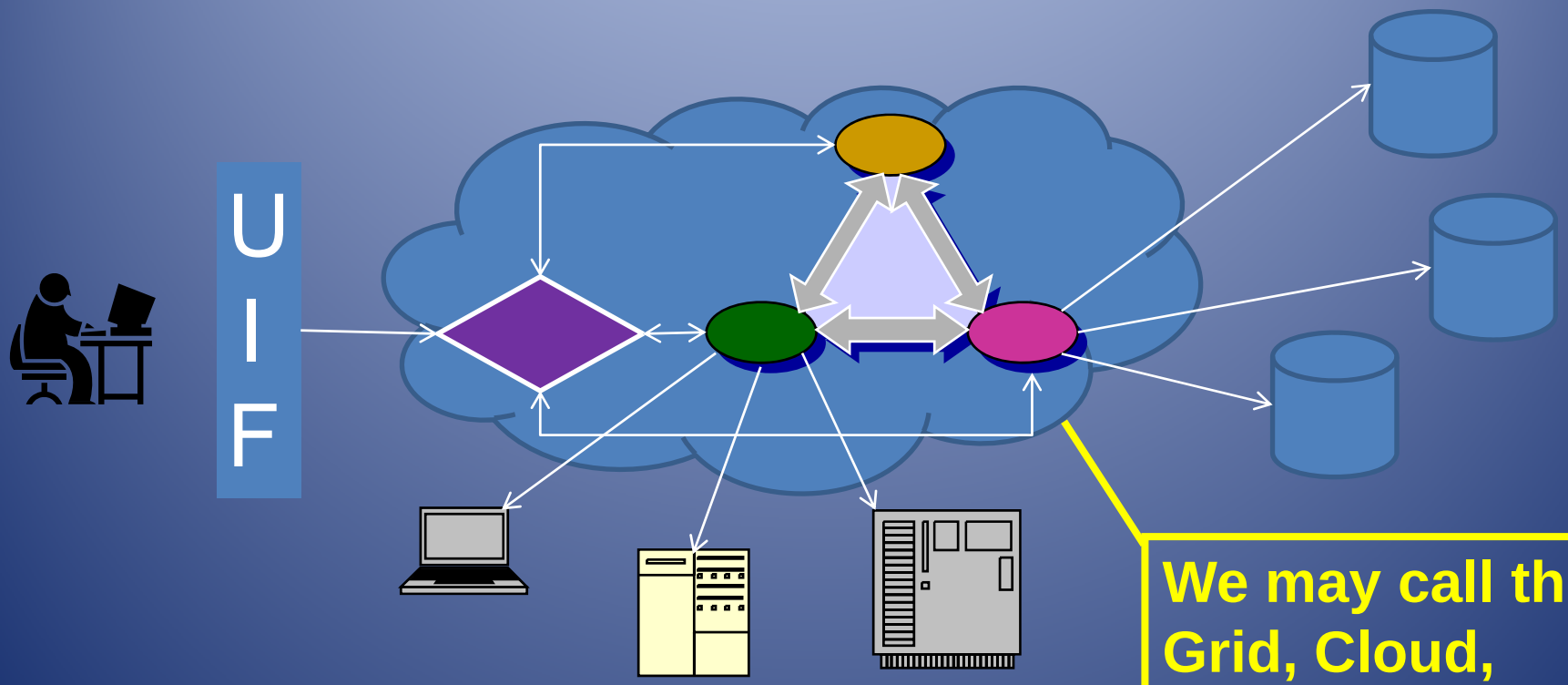
e-Infrastructure: conceptual schema

Theory
Numerical Simulations
Data Reduction
Data Analysis
...

different requirements, same infrastructure



User's viewpoint



We may call this Grid, Cloud, VObs, …

User is not concerned about (should not know?) what is in the Cloud

Summary

- Astronomy has changed and grown considerably
  $\Rightarrow$ archives needed
- Some work is required to integrate and make the various data archives interoperable $\Rightarrow$ the Virtual Observatory
- Goal: all astronomical databases "one click away"
  $\Rightarrow$ democratization of Astronomy!
- To make sense, the Virtual Observatory needs to be an international effort, which requires involvement at the *project* but also at the *data centre* level
- The Virtual Observatory concept can be re-used in different domains
- The final goal is **Science**

Thanks to:

- The organisers for the invitation
- P.Padovani (ESO), C.Arviset (ESA/ESAC), F.Genova (CDS), G.Rixon (AstroGrid) ⇒ Euro-VO
- C.Vuerli, U.Becciani, S.Cassisi, G.Taffoni (INAF), G.Longo (Univ. Napoli "Federico II"), P.Benvenuti (Univ. Padova) ⇒ VObs.it
- C.Loomis (CERN), M.Mazzuccato (INFN), R.Barbera (Univ. Catania) ⇒ EGEE

Thank you for your attention!

pasian@oats.inaf.it