Virtual Observatory, Cyber-Science, and the Rebirth of Libraries

S. G. Djorgovski

(Caltech)

With special thanks to
Roy Williams, Alex Szalay,
Jim Gray, and many other VO founders
and Cyber-Science pioneers …

An Overview:

- Astronomy in the era of information abundance

  The IT revolution, challenges and opportunities
- The Virtual Observatory concept

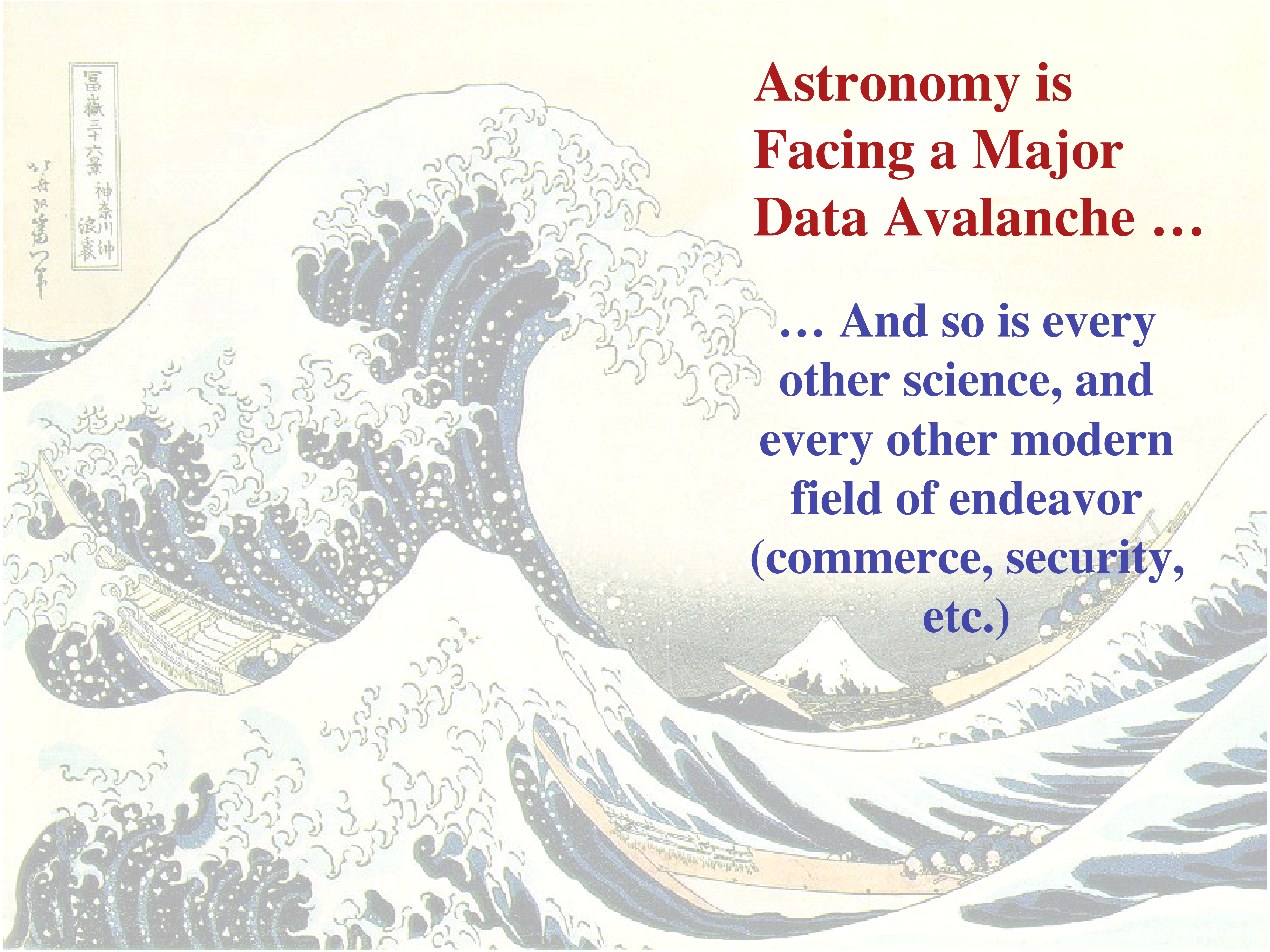
  What is it, how it all got started
- Virtual Observatory status

  Where are we now, where are we going
- From technology to science (and back)

  New tools for the science of 21st century
- Musings on cyber-science in general

  The changing nature of scientific inquiry
- The new roles of resarch libraries

  The changing nature of data and information needs

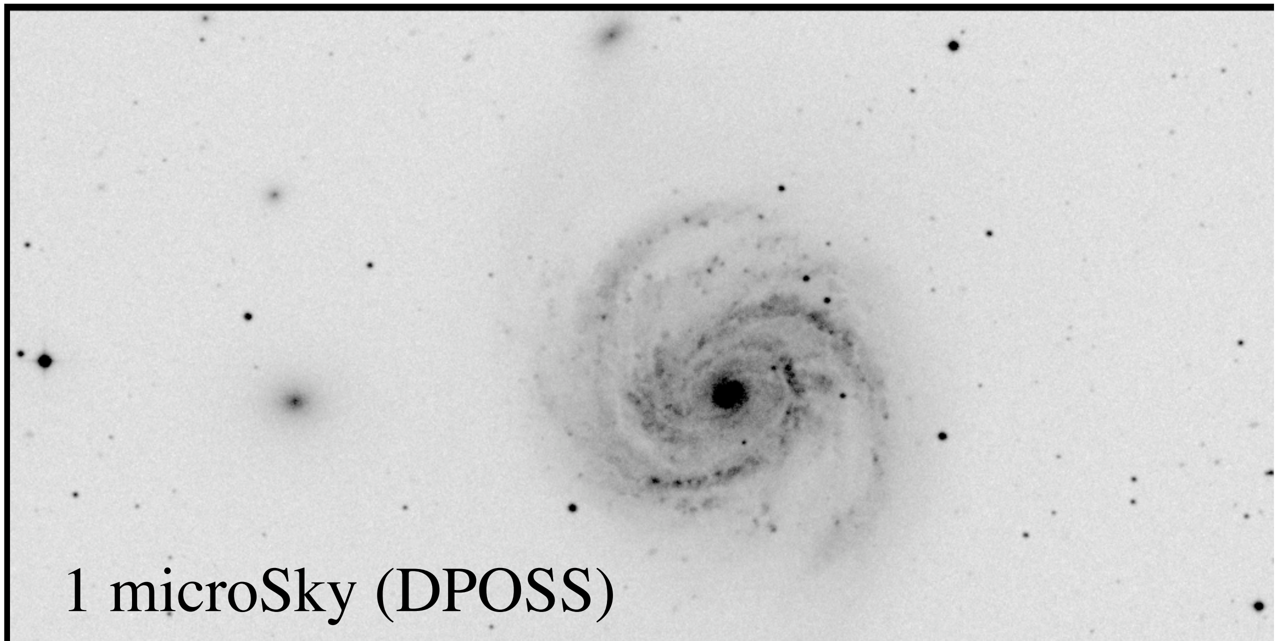


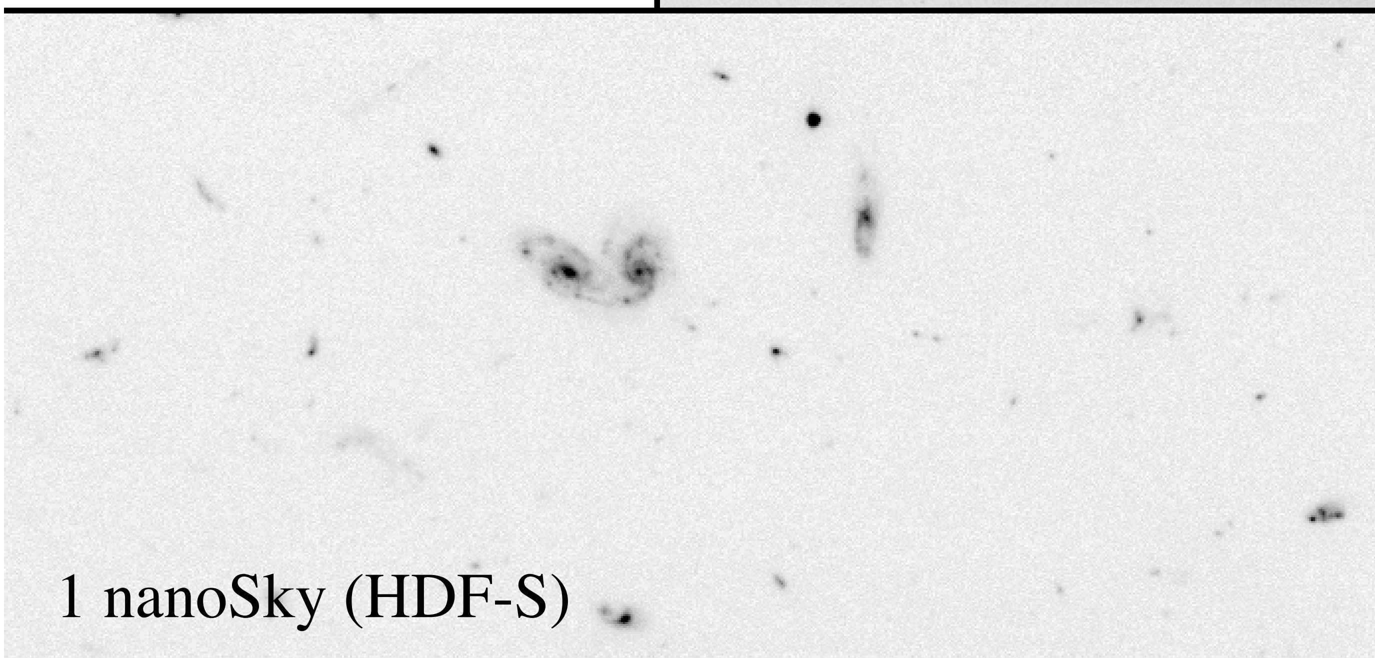
Astronomy is Facing a Major Data Avalanche …

… And so is every other science, and every other modern field of endeavor (commerce, security, etc.)

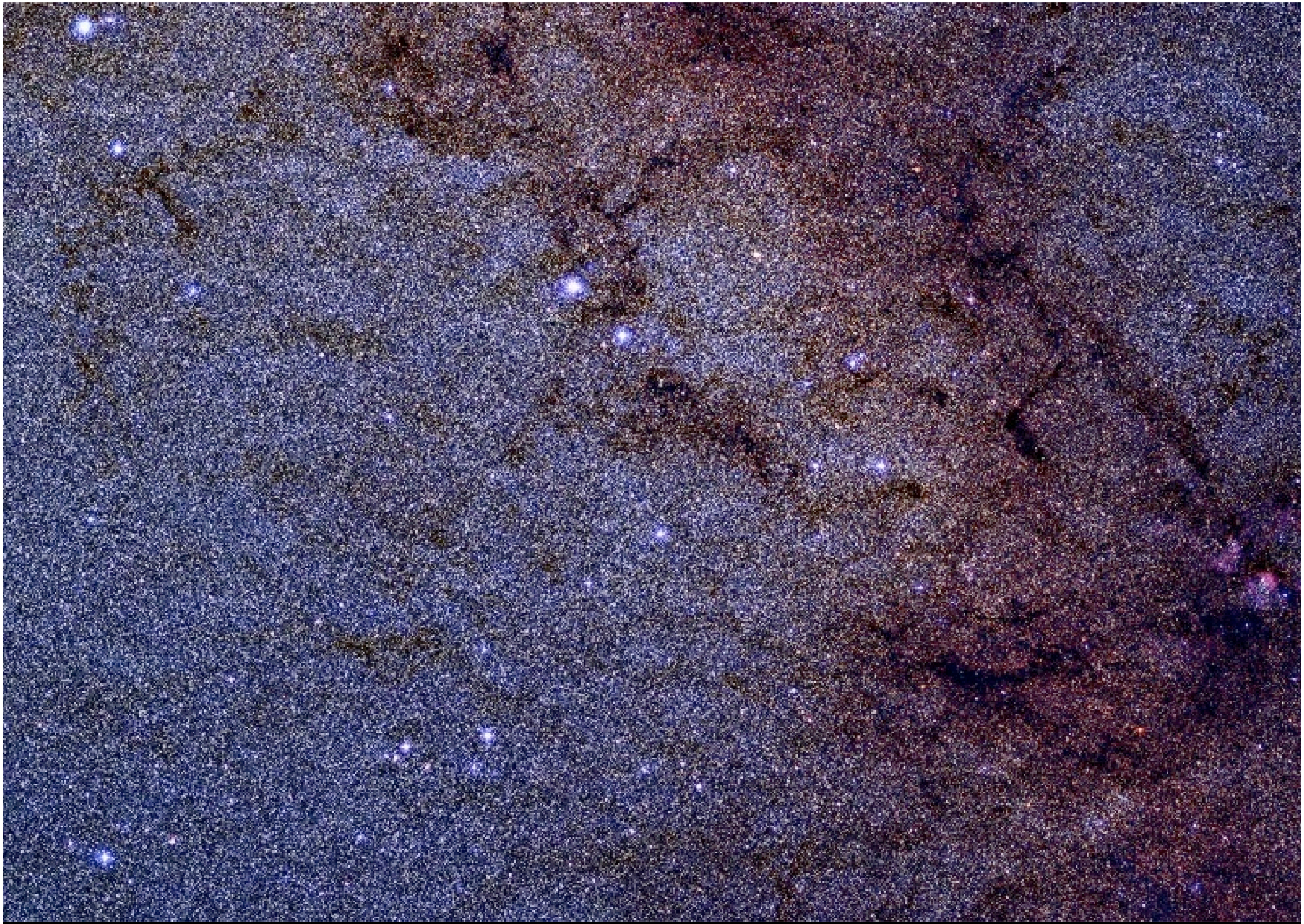
Astronomy is Now a Very Data-Rich Science

Multi-Terabyte (soon: multi-PB) ***sky surveys and archives*** over a broad range of wavelengths …





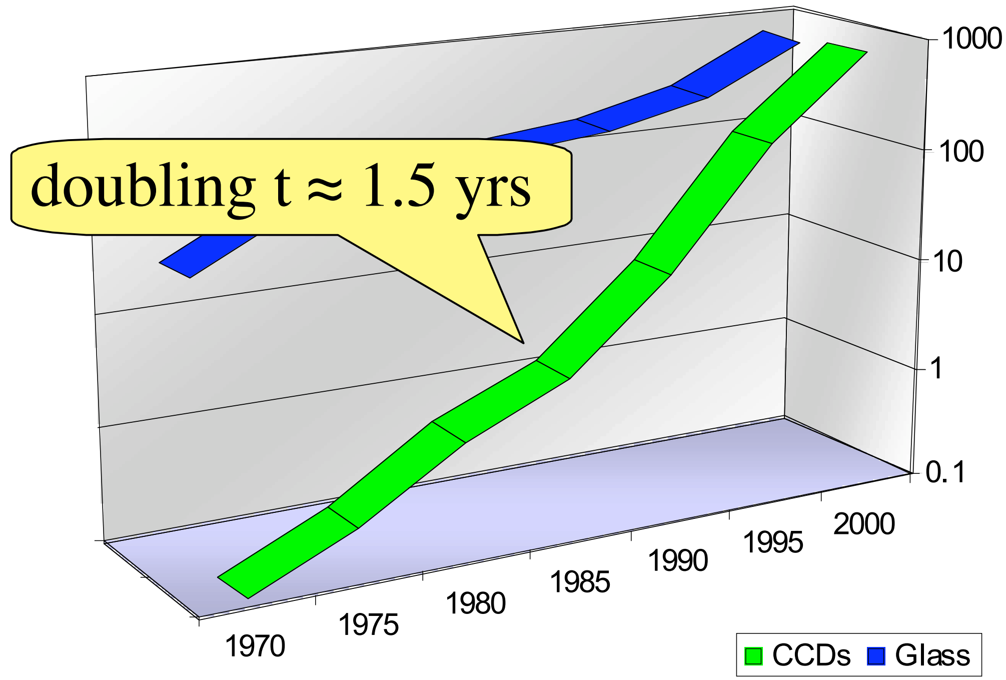
Billions of detected sources, hundreds of measured attributes per source …


Galactic Center Region (a tiny portion)   2MASS NIR Image

- **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**

Data → Knowledge ?

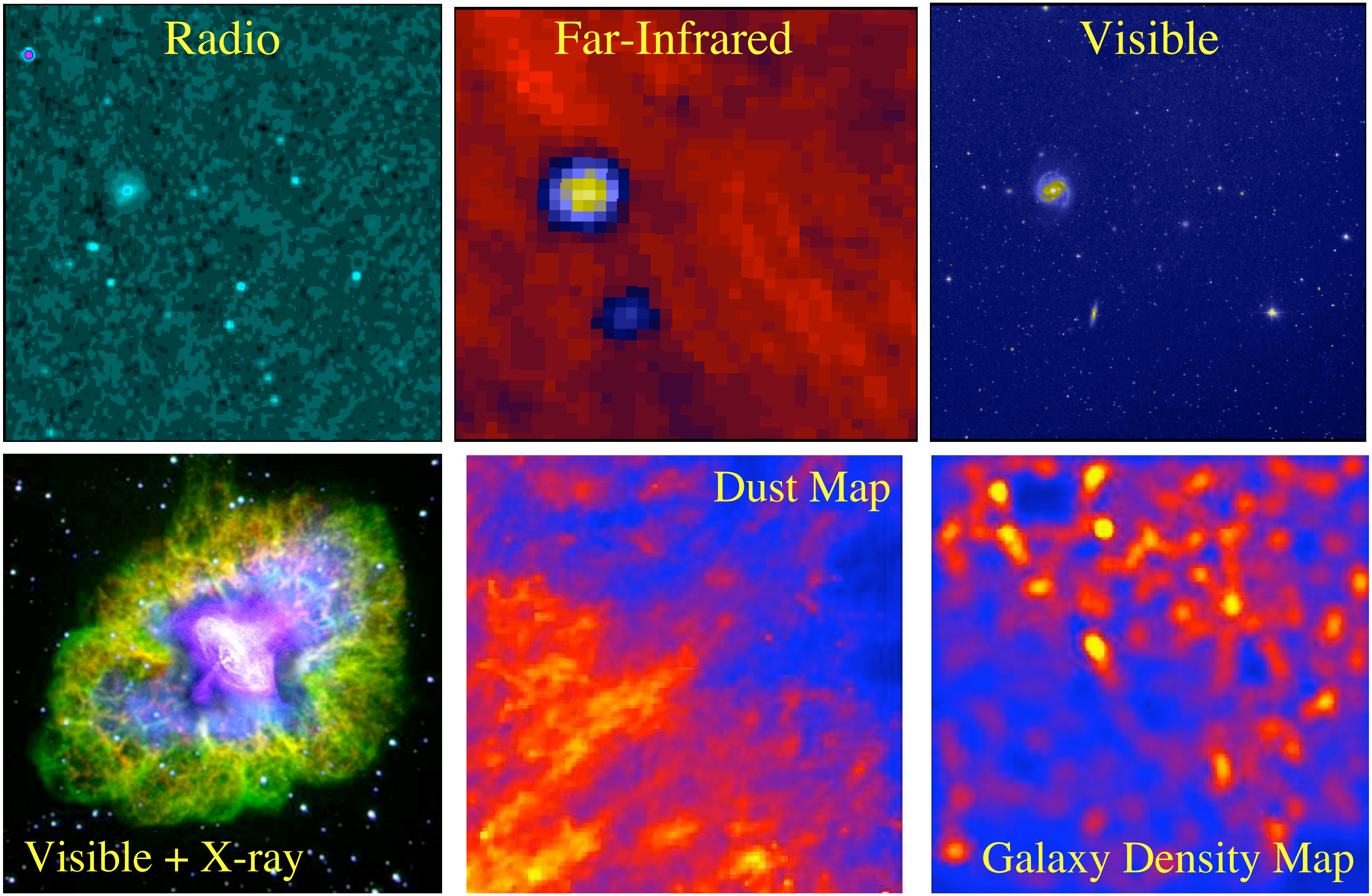
The exponential growth of data volume (and also complexity, quality) driven by the exponential growth in detector and computing technology



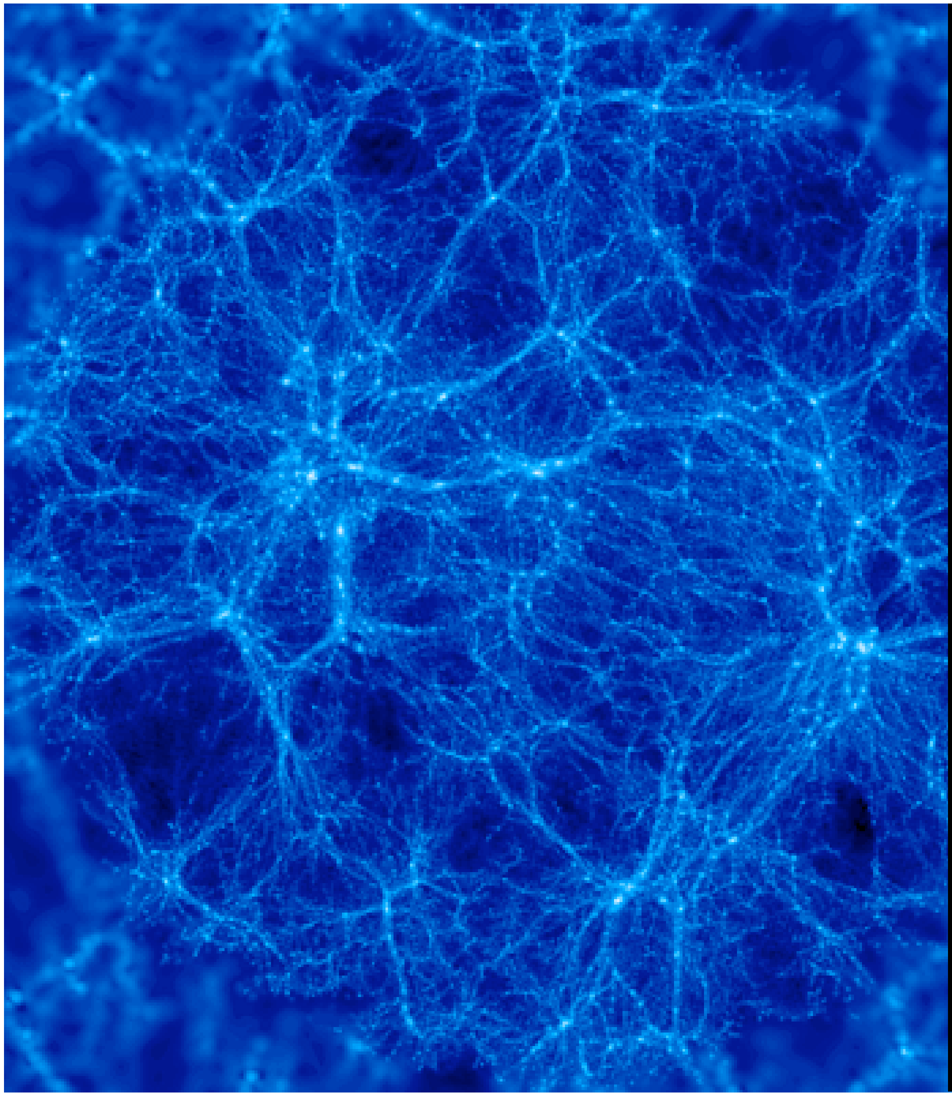
… but our understanding of the universe increases much more slowly!

Panchromatic Views of the Universe:

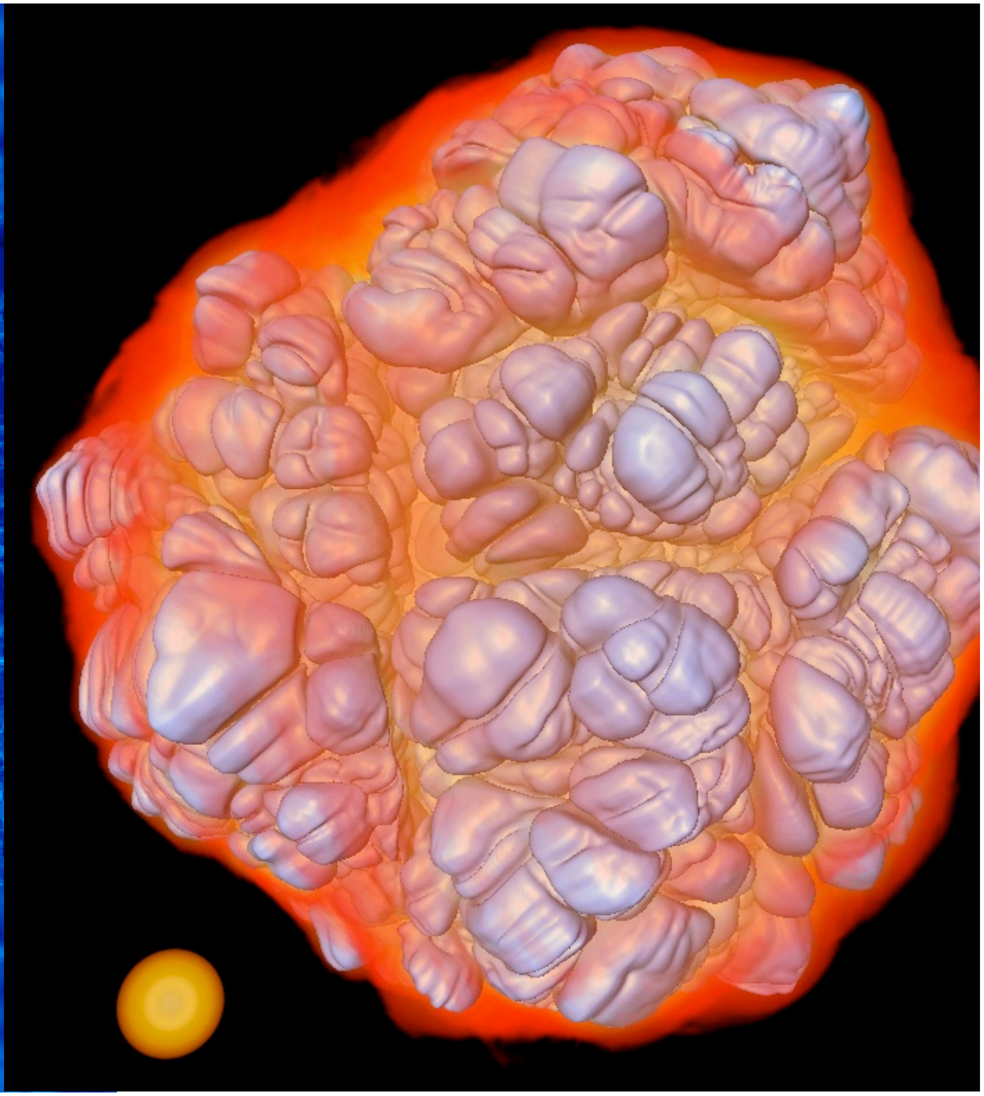
Data Fusion ➔ A More Complete, Less Biased Picture



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



Structure formation in the Universe

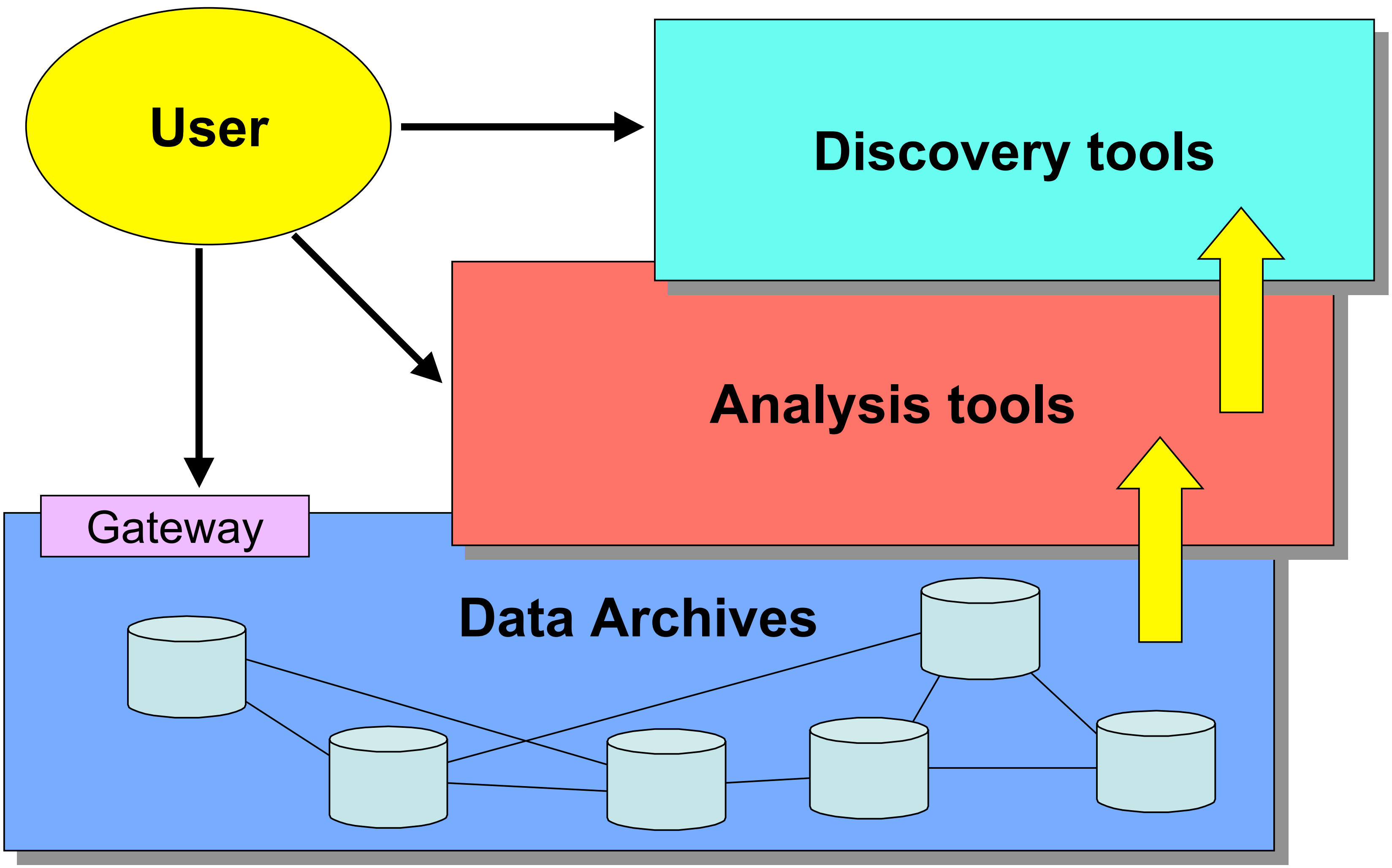


Supernova explosion instabilities

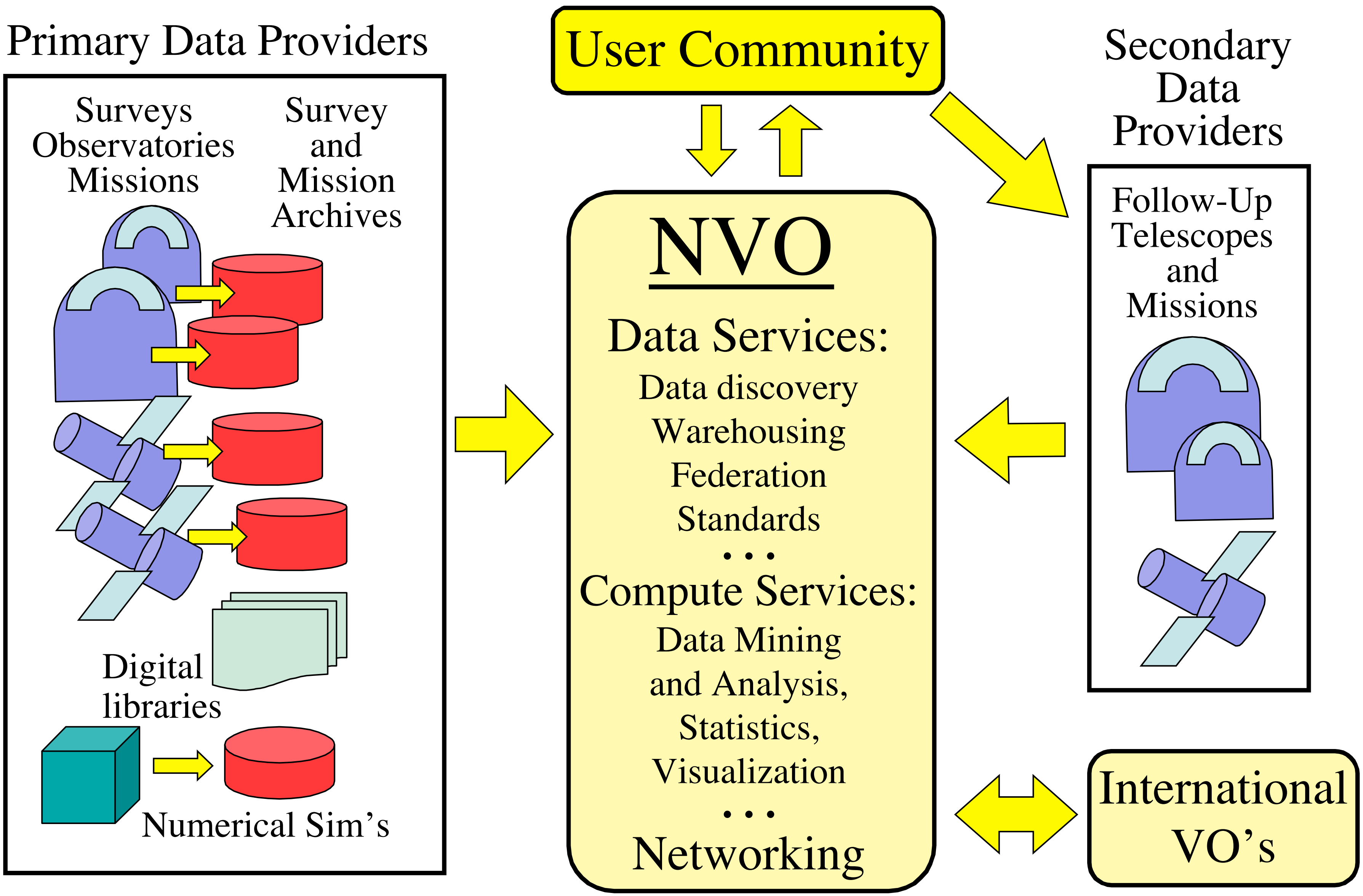
The Virtual Observatory Concept

- Astronomy community response to the scientific and technological challenges posed by massive data sets
  - Harness the modern information technology in service of astronomy, and partner with it
- A complete, dynamical, distributed, open ***research environment for the new astronomy with massive and complex data sets***
  - Provide content (data, metadata) services, standards, and analysis/compute services
  - Federate the existing and forthcoming large digital sky surveys and archives, facilitate data inclusion and distribution
  - Develop and provide data exploration and discovery tools
  - ***Technology-enabled, but science-driven***

VO: Conceptual Architecture



A Systemic View of the NVO



Why is VO a Good Scientific Prospect?

- Technological revolutions as the drivers/enablers of the bursts of scientific growth
- Historical examples in astronomy:
  - 1960's: the advent of electronics and access to space

    Quasars, CMBR, x-ray astronomy, pulsars, GRBs, …
  - 1980's - 1990's: computers, digital detectors (CCDs etc.)

    Galaxy formation and evolution, extrasolar planets, CMBR fluctuations, dark matter and energy, GRBs, …
  - **2000's and beyond: information technology**

The next golden age of discovery in astronomy?

VO is the mechanism to effect this process

Information Technology → New Science

- The information volume grows exponentially

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 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 …

- VO is the framework to effect this for astronomy

A Modern Scientific Discovery Process

Data Gathering

↳ **Data Farming:**

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability

} Database Technologies

↳ **Data Mining** (or Knowledge Discovery in Databases):

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

Key Technical Challenges →

↳ **Data Understanding**

↳ **New Knowledge**

← Key Methodological Challenges

How and Where are Discoveries Made?

- **Conceptual Discoveries:** e.g., Relativity, Quantum Mechanics, Strings, Inflation … *Theoretical, may be inspired by observations*
- **Phenomenological Discoveries:** e.g., Dark Matter, Dark Energy, QSOs, GRBs, CMBR, Extrasolar Planets, Obscured Universe … *Empirical, inspire theories, can be motivated by them*

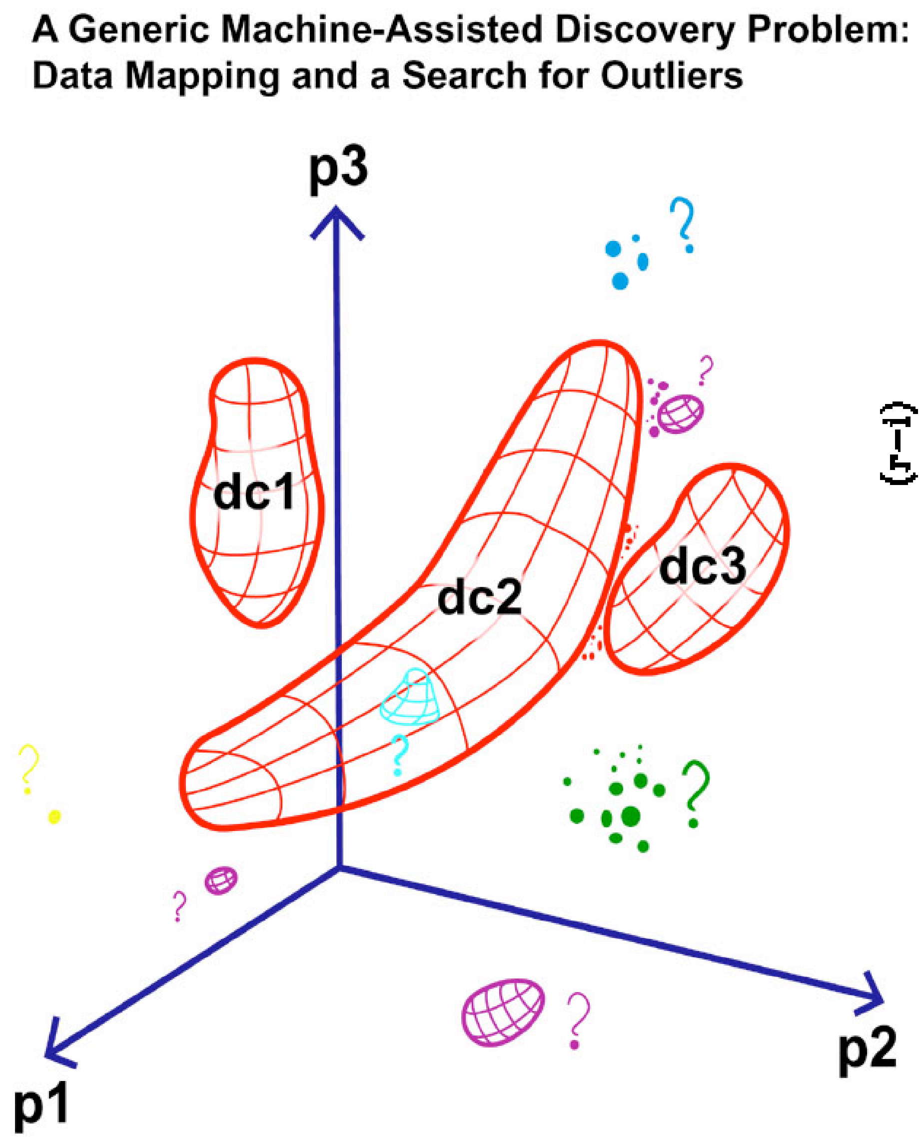
New Technical Capabilities → (IT/VO) → Observational Discoveries ↔ (VO) ↔ Theory

Phenomenological Discoveries:

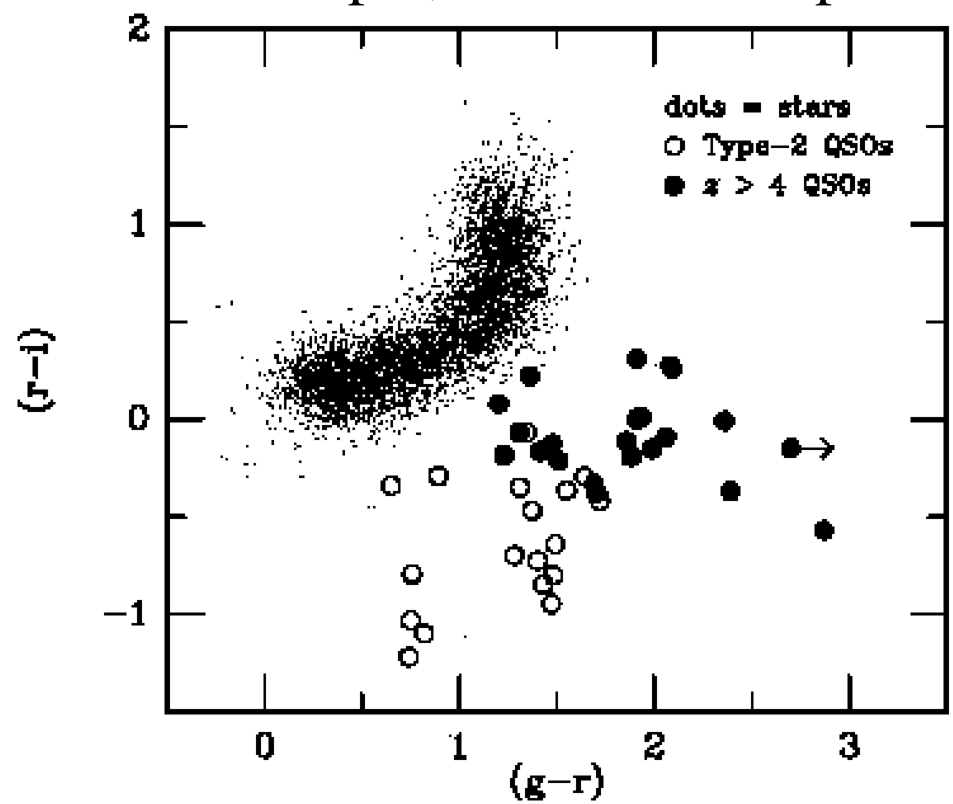
- Pushing along some parameter space axis ⇐ VO useful
- Making new connections (e.g., multi-□) ⇐ **VO critical!**

Understanding of complex astrophysical phenomena requires complex, information-rich data (and simulations?)

Exploration of observable parameter spaces and searches for rare or new types of objects



A simple, real-life example:



Now consider ~ 10^9 data vectors in ~ 10^2 - 10^3 dimensions …

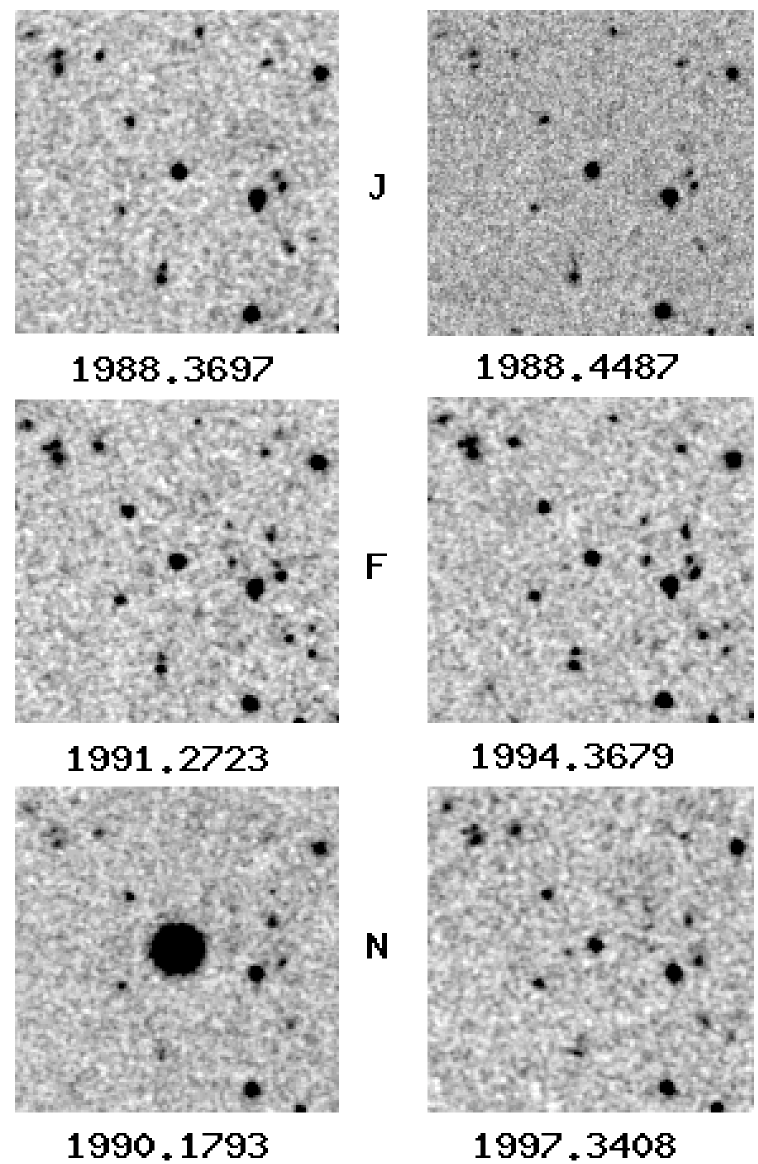
Exploration of the Time Domain …

… and the advent of
Synoptic Sky Surveys

An example (from DPOSS) of a new type of a phenomenon which may be discovered in a systematic exploration of the **Time Domain**:

A normal, main-sequence star which underwent an outburst by a factor of > 300. There is some anecdotal evidence for such **megaflares** in normal stars.

The cause, duration, and frequency of these outbursts is currently **unknown**.



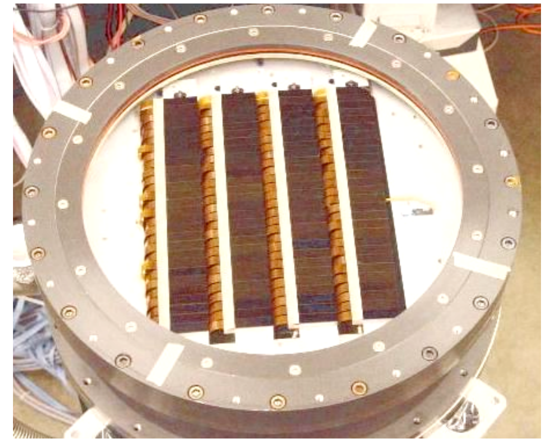
An Example of a Synoptic Sky Survey: Palomar-Quest

A Caltech-Yale-JPL collaboration

Huge data rate:

~ 1 TB/month (but in < 10 yrs, we'll have > 1 TB/day)

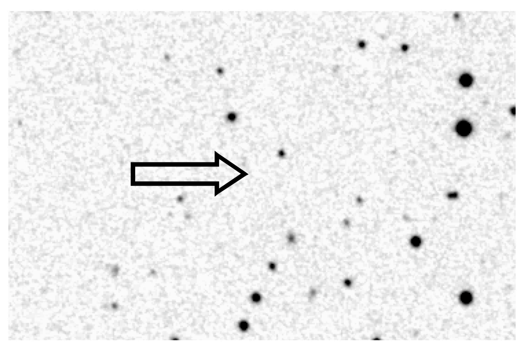
Palomar 40-inch telescope

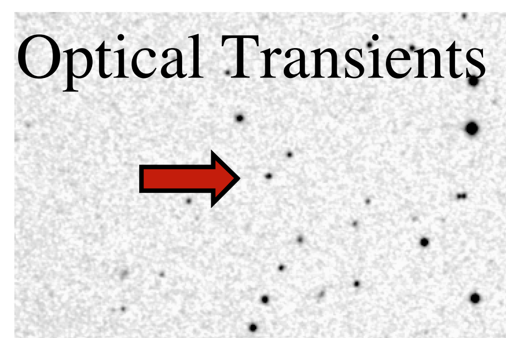


The 112-CCD camera

Look for things that move…

… and things that go **Bang!** in the night





Scientific Roles and Benefits of a VO

- **Facilitate science with massive data sets** (observations and theory/simulations) ⟹ **<u>efficiency amplifier</u>**
- Provide an **<u>added value</u>** from federated data sets (e.g., multi-wavelength, multi-scale, multi-epoch …)
  - Discover the knowledge which is present in the data, but can be uncovered *only* through data fusion
- **Enable and stimulate some <u>*qualitatively new* science</u>** with massive data sets (not just old-but-bigger)
- **Optimize the use of expensive resources** (e.g., space missions, large ground-based telescopes, computing …)
- Provide R&D drivers, application testbeds, and stimulus to the **partnering disciplines** (CS/IT, statistics …)

VO Developments and Status

- The concept originated in 1990's, developed and refined through several conferences and workshops
- Major blessing by the National Academy Report
- **In the US:** National Virtual Observatory (NVO)
  - Concept developed by the NVO Science Definition Team (SDT). See the report at ***http://www.nvosdt.org***
  - NSF/ITR funded project: ***http://us-vo.org***
  - A number of other, smaller projects under way
- **Worldwide** efforts: International V.O. Alliance
- A good synergy of astronomy and CS/IT
- Good progress on data management issues, a little on data mining/analysis, first science demos forthcoming


US National Virtual Observatory

search

Home | Registry | Tools | Data Access | Publish | Education | NVO in Use | Grid Computing | Architecture | Contact Us

News

- NVO Summer School
- Data Inventory Service
- Discovery by VO Demo
- VO Alliance Formed
- NVO News Archive

About

- What is the NVO?
- Who is Involved?
- Science Objectives

Community

- NVO Meetings
- International VO Alliance

Documents

Recent NVO Documents:

- Conesearch definition
- Quarterly Report Q104
- Management Plan
- VO Resource Registry
- All NVO Documents
- IVOA Documents

Supported by the National Science Foundation

Member of the International

NVO - Facilitating Scientific Discovery

NVO's objective is to enable new science by greatly enhancing access to data and computing resources. The NVO is developing tools that make it easy to locate, retrieve, and analyze astronomical data from archives and catalogs worldwide, and to compare theoretical models and simulations with observations.

These tools are based upon international standards developed in collaboration with the International Virtual Observatory Alliance.

We expect to deliver the first production quality services in early 2005. Some examples of existing prototypes:

- Use the VO Spectrum Services to analyze over 500,000 spectra.
- Cross-correlate objects from more than 15 surveys with SkyQuery
- Use YourSky to make custom infrared sky images based on DPOSS or 2MASS.

The NVO also provides software libraries and sample code of VO Services for people who want to write their own VO-enabled applications.

NVO - Data Access

The NVO encourages astronomical research organizations to make their data collections and source catalogs available via the standard VO protocols. These include image access, spectrum access, and catalog search.

A number of astronomical research facilities and survey projects are already making use of NVO interfaces and protocols in support of data processing, analysis, and distribution.

Available collections and services can be located through the NVO Registries -- the Yellow Pages of astronomical resources, with regularly updated entries. Try the different interfaces at NCSA, STScI, or Caltech to the NVO registries already containing more than 6000 entries!

NVO - Education and Public Outreach

Astronomical images are treasured by the public for their beauty, and thus are an excellent vehicle for science education at all levels. We seek partnerships with educational organizations, museums, and planetariums to help them use our tools to incorporate NVO-ready data into their programs and curricula. Sample projects:

- Project LITE is an interactive environment to study astronomical spectra
- ...nyOne, a next-generation web-browser providing encyclopedic access to ...ence information.

...site is a community-maintained collection with content control by the NVO Executive Committee. ...ged by the extent to which it: (a) reflects an aspect of the Virtual Observatory, such as astronomy ... data, (b) uses VO standards or software, or (c) exemplifies grid-based astronomical computing. If you would like a description of your project, data, or software included here, please write to web at us-vo.org with a short description of your work.

Summer School

Aspen CO, Sep 13-17. More Information

Interop Meeting

Sep 27-29, 2004, IUCAA Pune, India More Information

Data Inventory

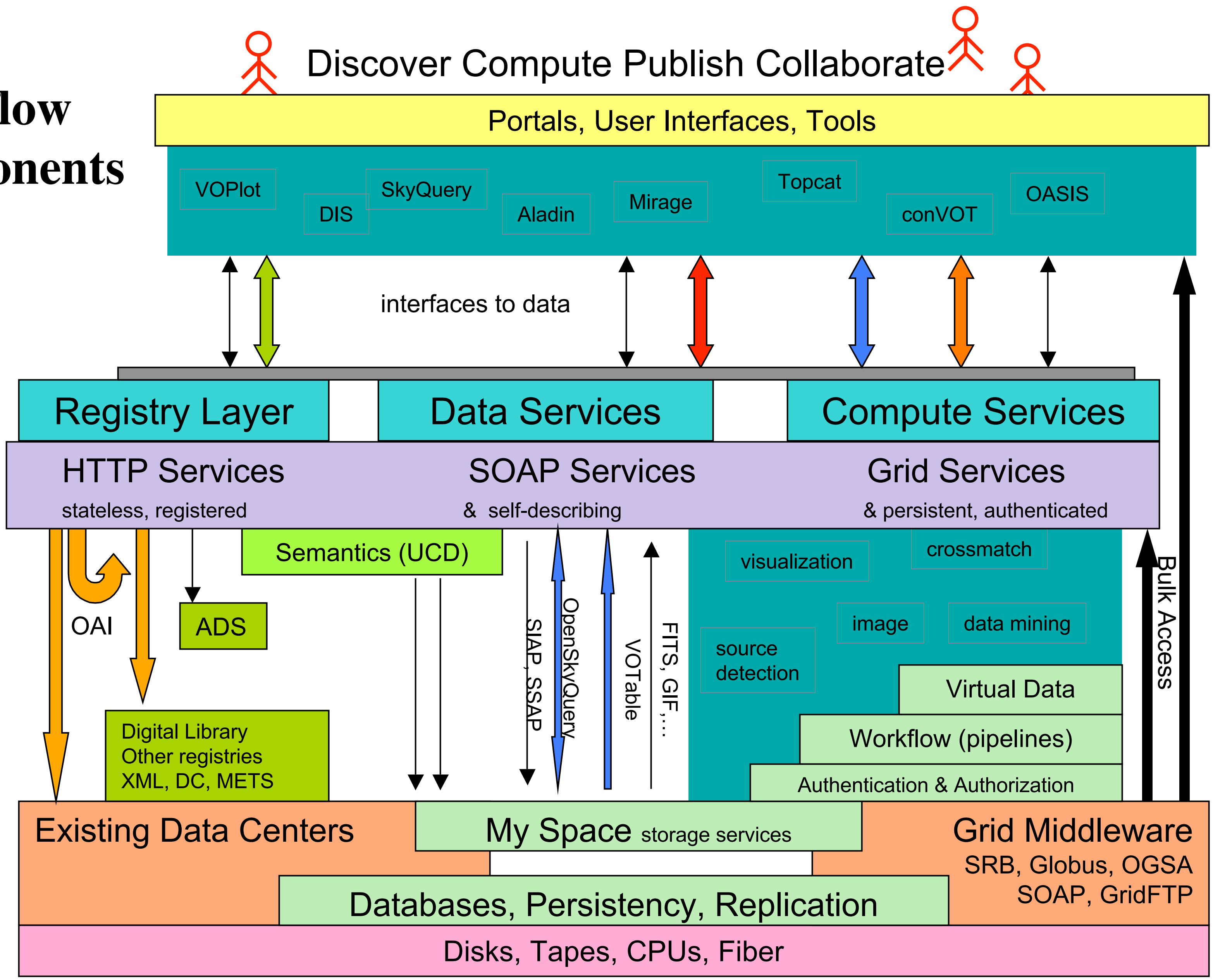
Find images and catalog objects around a given sky position with the Data Inventory Service.

Project LITE

Project LITE is an instructional environment for astronomical spectra

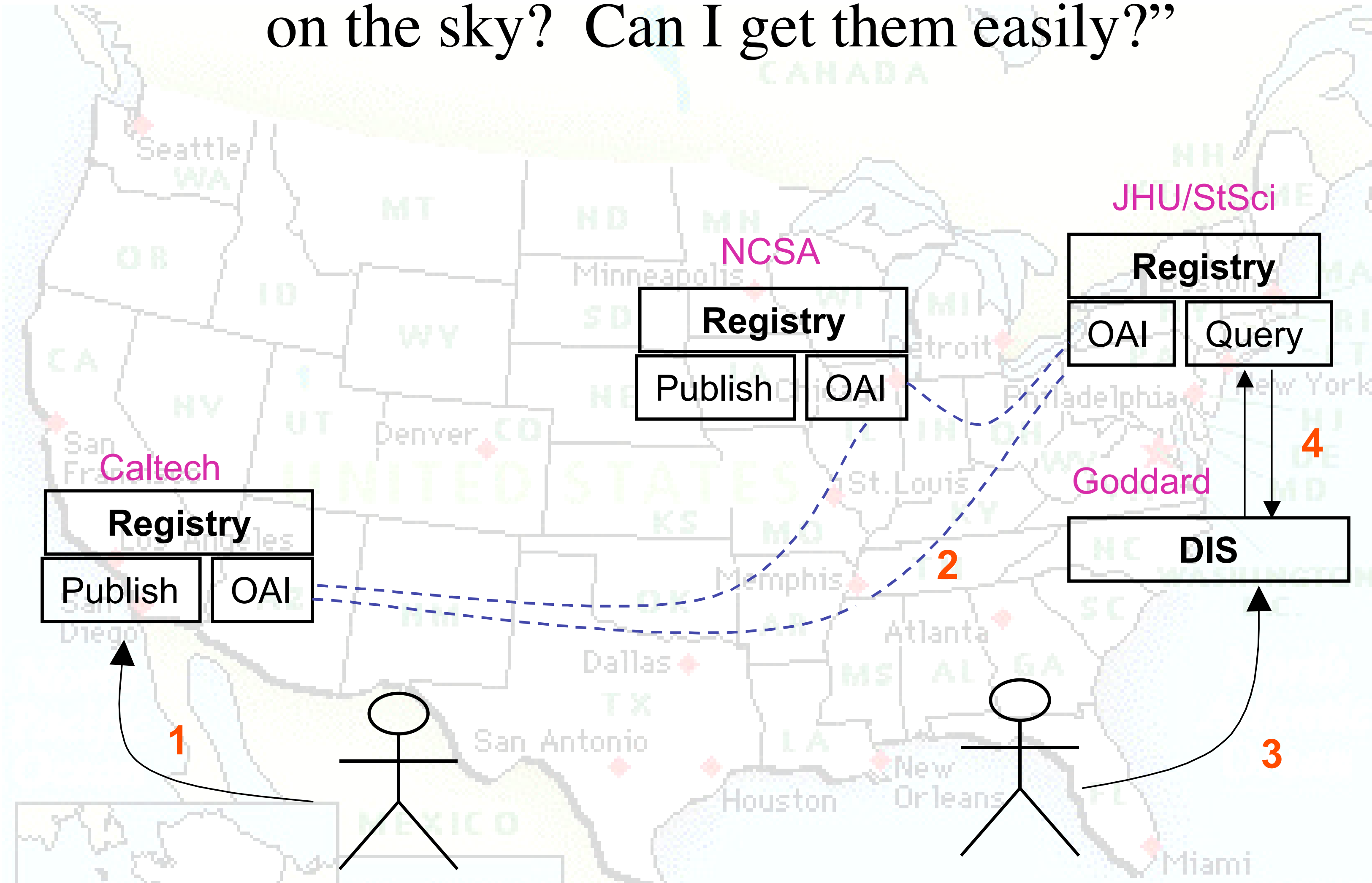
http://us-vo.org

NVO Workflow Components



NVO: A Prototype Data Inventory Service

"What data are available for some object or some region on the sky? Can I get them easily?"



Data Inventory Service

National Virtual Observatory: Hosted at the HEASARC

What do we know about regions of sky?
Using new Virtual Observatory protocols we can gather and organize information efficiently on a given region of sky.

Enter a position(or name) and the maximum size of the region of sky you're interested in.

Object Position or Name:	cen a	(degrees or sexagesimal)
Size:	0.25	(in decimal degrees)

Send Request Reset Form

☐ Ignore cache! The DIS will reproccess an identical request rather than linking to the existing cache results.

DIS user interface

Example Inputs for the Object Position or Name

- 13.29, -18.47 [Object Position: Decimal degrees]
- 6 45 10.8, -16 41 58 [Object Position:Sexagesimal format; RA in hours]
- 3c273 [Object name]
- Use a comma to delimit J2000 RA and Dec pair.

About Data Inventory Service

1. A user request is broadcast to sites scattered all over the world using two simple common protocols.
2. Catalog data and lists of available images are returned using the new VOTable XML standard.
3. Image, observation and catalog data from these sites are collected and organized for immediate viewing.
4. Data may be analyzed or visualized in Aladin or OASIS

Participating sites currently include: NRAO, NOAO, JHU, ST ScI, HEASARC, NCS, IRSA, CDS, NED, ESO, SDSS, CXC.

A service of the Laboratory for High Energy Astrophysics (LHEA)
and the High Energy Astrophysics Science Archive Research Center (HEASARC) at NASA/ GSFC

NVO

Data Inventory Results: cen a

Data missing - Instructions   Home

National Virtual Observatory: Hosted at the HEASARC

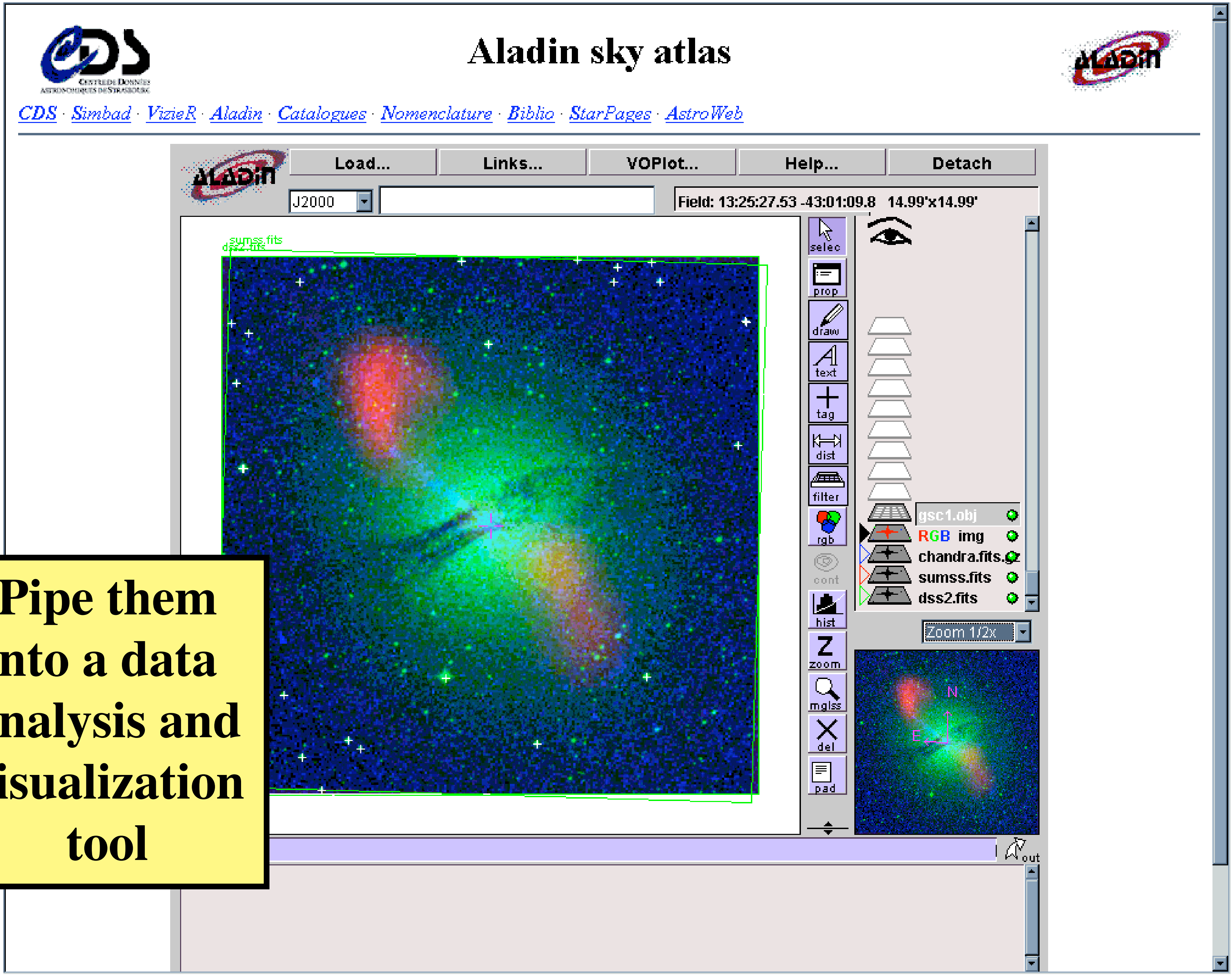
Note: Inventory request completed

DIS search results

RA	Dec	Size
13 25 27.62	-43 01 08.8	0.25

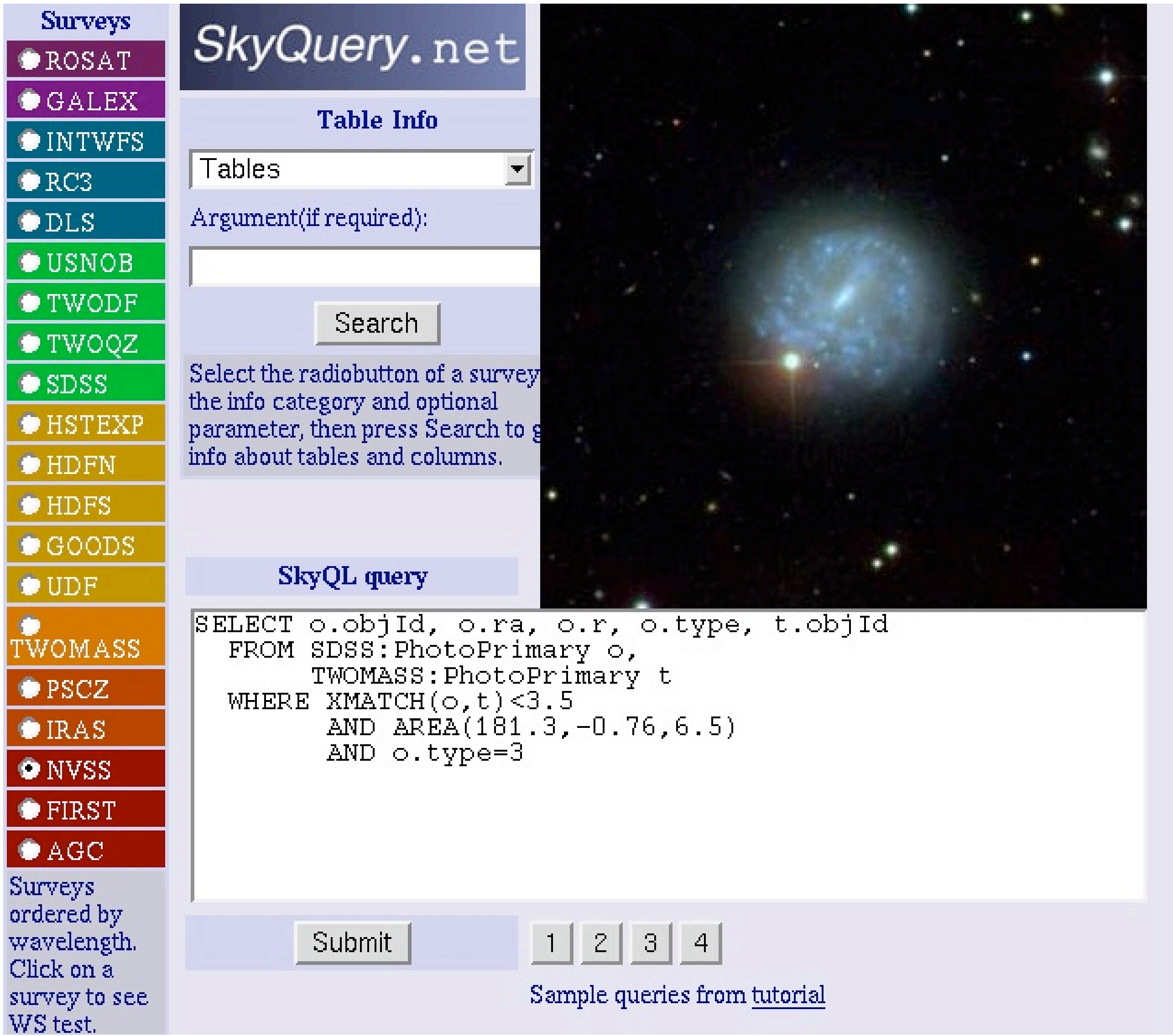
☐ Check All

Images (FITS/GIF)					
Optical	☐ DSS1 SV	☑ DSS2	☐ DSS2B	☐ DSS2IR	☐ DSS2R
Infrared	☐ 2MASS-H	☐ 2MASS-J	☐ 2MASS-K		
Radio	☑ SUMSS				
X-ray	☐ RASS B	☑ Chandra(6)			
Observations (VOTable)					
Optical	☐ HST(100)	☐ STIS(100)	☐ WFPC2(100)	☐ WFPC1(22)	☐ HSTG(394)
Infrared	☐ NICMOS(100)				
X-ray	☐ ASCA(3)	☐ ROSAT(9)	☐ ROSPUBLC(10)	☐ RXTE(23)	☐ EXOSAT(12)
	☐ CHANMAST(10)	☐ Einstein(5)	☐ XMMMAST(3)	☐ ASCAMAST(3)	☐ XTEINDEX(5)
Gamma-ray	☐ OSSE(29)				
UV	☐ FUSE(1)	☐ FOC(20)	☐ HUT(2)	☐ IUE(41)	☐ UIT(7)
	☐ WUPPE(1)				
Objects (VOTable)					
Surveys	☐ USNO-A2.0(1197)	☐ USNO-SA2.0(1197)	☑ GSC1(289)	☐ GSC2.2(2259)	☐ UCAC1(305)
	☐ USNO-A2.0 CDS(999)				
Galaxies	☐ SGC(1)	☐ PGC(1)	☐ NBG(1)	☐ RC3(1)	☐ RNGC(1)
	☐ PSCz(3)				
Stars	☐ HIP(1)	☐ SAO(2)	☐ WDS(1)	☐ AC2000.2(30)	☐ ASCC-2.5(21)
	☐ HD(4)				
Misc.	☐ EGRET3(45)	☐ WGACAT(35)	☐ Radio Catalogs(69)	☐ 2MASS-PSC(CDS)(999)	☐ Veron-Veron(1)
	☐ TYCHO-2(22)				



Pipe them into a data analysis and visualization tool

SkyQuery: NVO Prototype Catalog Cross-Matching Service (Data Federation)



… and much more is coming!

Broader and Societal Benefits of a VO

- **Professional Empowerment:** Scientists and students anywhere with an internet connection would be able to do a first-rate science ⇒ A broadening of the talent pool in astronomy, democratization of the field
- **Interdisciplinary Exchanges:**
  - The challenges facing the VO are common to most sciences and other fields of the modern human endeavor
  - Intellectual cross-fertilization, feedback to IT/CS
- **Education and Public Outreach:**
  - Unprecedented opportunities in terms of the content, broad geographical and societal range, at all levels
  - Astronomy as a magnet for the CS/IT education

“Weapons of Mass Instruction”


http://virtualsky.org (R. Williams et al.)
Decrease map size
Show Marker
Clicking on sky will:
Zoom In
Zoom Out
Recenter
Galaxy lookup
Star lookup
You are at:
Right Ascension
14.679072
Declination
60.933364
(clean)
Go Here
30 arcmin
Image courtesy of Digital Palomar Observatory Sky Survey
Image center is RA,Dec=14.679072 60.933364
Image width is 53.715627 minutes
Get VS Jpeg Get VS FITS
virtualSky.org

International Virtual Observatory Alliance

Member Organizations

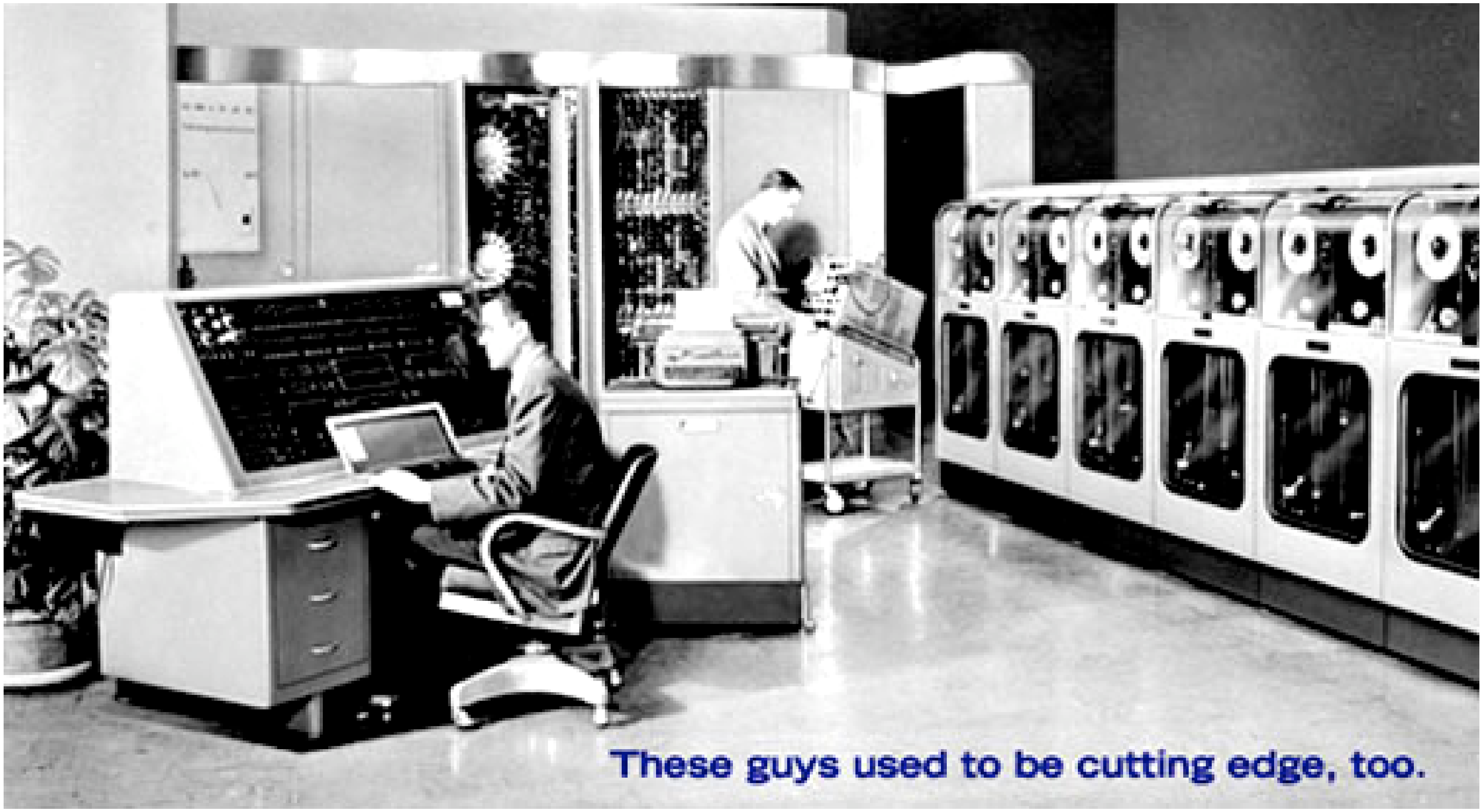
http:// ivoa.net

A Coalition of the Willing?

Do We Know How to Run a VO?

- The VO is ***not*** yet another data center, archive, mission, or a traditional project ⇨ ***It does not fit into any of the usual structures today***
  - 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
  - It has an ***unusually broad range of constituents*** and interfaces
  - It is inherently ***multidisciplinary***
- The VO represents ***a novel type of a scientific organization*** for the era of information abundance

Now Let's Take A Look At Some Relevant Technology Trends …



The rate of the overall computing power has been amazingly growing for more than one hundred years

Computing efficiency in ops/s/$ had 3 growth curves:

1890-1945

- Mechanical
- Relay
- 7-year doubling

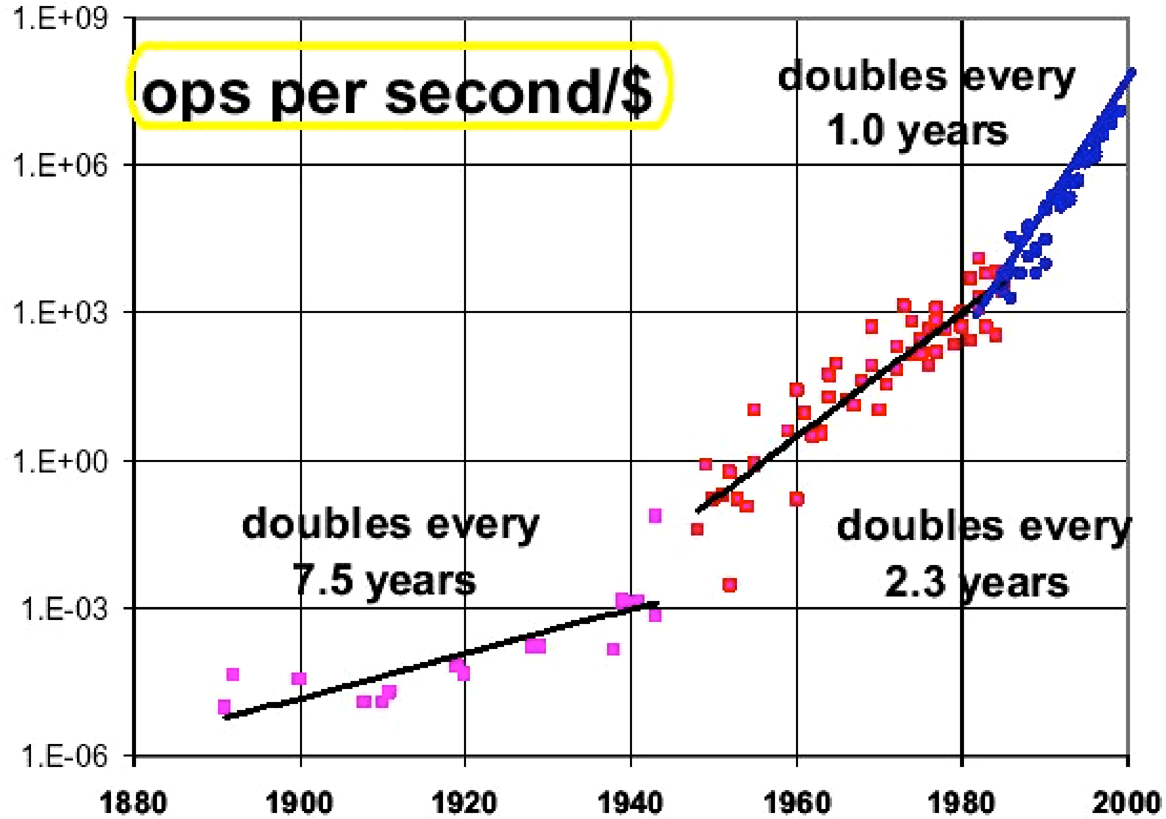
1945-1985

- Tube, transistor,..
- 2.3 year doubling

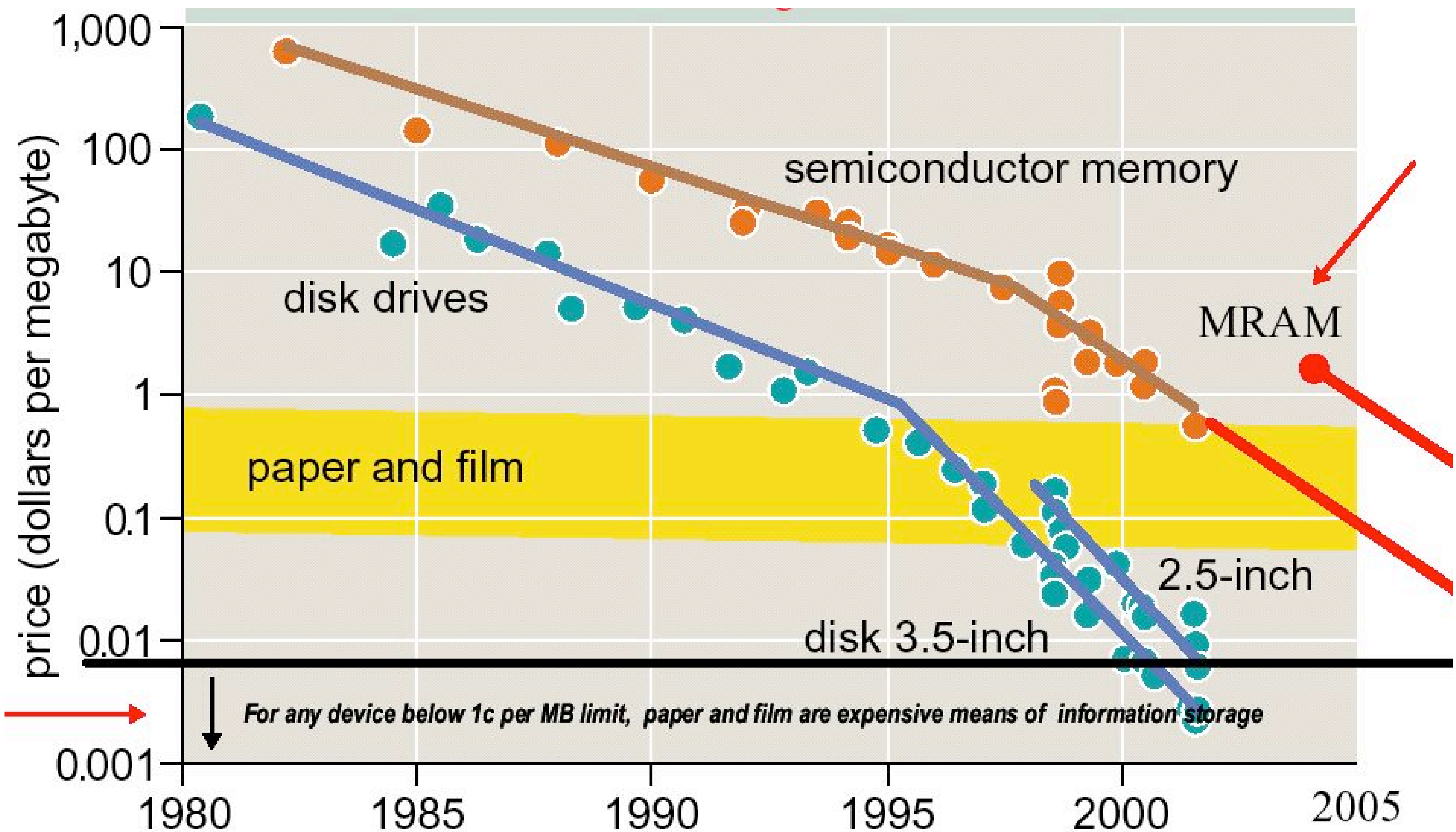
1985-2000

- Microprocessor
- 1.0 year doubling

Combination of Hans Moravac + Larry Roberts + Gordon Bell
WordSize*ops/s/sysprice



Exponentially Declining Cost of Data Storage



Hayes, Grochowski: American Scientist, 2002

Computing is Cheap …

Today (~2004), 1 $ buys:

- 1 day of CPU time
- 4 GB (fast) RAM for a day
- 1 GB of network bandwidth
- 1 GB of disk storage for 3 years
- 10 M database accesses
- 10 TB of disk access (sequential)
- 10 TB of LAN bandwidth (bulk)
- 10 KWh = 4 days of computer time

… Yet somehow computer companies make billions: you do want some toys, about $ 10^5 worth ≈ 1 postdoc year

… But People are Expensive!

People ~ Software, maintenance, expertise, creativity …

Moving Data is Slow!

How long does it take to move a Terabyte?

(how about a Petabyte?)

Context	Speed Mbps	Rent $/month	$/Mbps	$/TB Sent	Time/TB
Home phone	**0.04**	**40**	**1,000**	**3,086**	**6 years**
Home DSL	**0.6**	**50**	**117**	**360**	**5 months**
T1	**1.5**	**1,200**	**800**	**2,469**	**2 months**
T3	**43**	**28,000**	**651**	**2,010**	**2 days**
OC3	**155**	**49,000**	**316**	**976**	**14 hours**
OC 192	**9600**	**1,920,000**	**200**	**617**	**14 minutes**
100 Mpbs	**100**				**1 day**
Gbps	**1000**				**2.2 hours**

Source: TeraScale Sneakernet, Microsoft Research, Jim Gray et al.

Disks are Cheap and Efficient

- Price/performance of disks is improving faster than the computing (Moore's law): a factor of ~ 100 over 10 years!
  - Disks are now already cheaper than paper
- Network bandwith used to grow even faster, but no longer does
  - And most telcos are bancrupt …
  - Sneakernet is faster than any network
- Disks make data preservation easier as the storage technology evolves
  - Can you still read your 10 (5?) year old tapes?

An Early Disk for Information Storage

- Phaistos Disk:
  Minoan, 1700 BC

- No one can read it ☺

(From Jim Gray)

The Gospel According to Jim Gray:

- Store everything on disks, with a high redundancy (cheaper than the maintenance/recovery)
  - Curate data where the expertise is
- Do not move data over the network: **bring the computation to data!**
  - The Beowulf paradigm: Datawulf clusters, smart disks …
  - The Grid paradigm (done right): move only the questions and answers, and the flow control
- You ***will*** learn to use databases!
- And we need a better fusion of databases and data mining and exploration

These Challenges Are Common!

- Astronomical data volume *ca*. 2004: **a few × 10^2 TB** (but PB's are coming soon!)
- All recorded information in the world: **a few × 10^7 TB** (but most of it is video, *i.e*., junk)
- The data volume everywhere is growing exponentially, with *e*-folding times ~ 1.5 yrs (Moore's law)
  - NB: the data rate is also growing exponentially!
- So, ***everybody*** needs efficient db techniques, DM (searches, trends & correlations, anomaly/outlier detection, clustering/classification, summarization, visualization, etc.)
- What others discover will help us, and maybe we can also help change the world *(remember the WWW!)*

The Evolution of Science

1600 1700 1800 1900 1950 2000 → t

Empirical/Descriptive

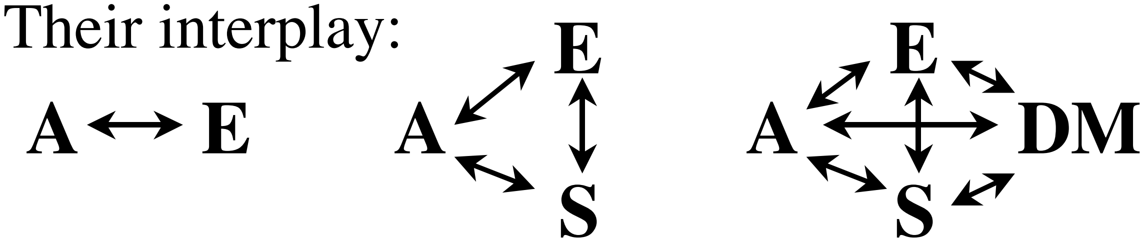
Analytical+Experimental

A+E+Simulations

A+E+S+**DM/DE/KDD**

↓ *Technology*

Their interplay:



Computational science rises with the advent of computers

Data-intensive science is a more recent phenomenon

The Evolving Role of Computing:

Number crunching ➔ Data intensive (data farming, data mining)

Some Musings on CyberScience

- Enables a broad spectrum of users/contributors
  - From large teams to small teams to individuals
  - Data volume ~ Team size
  - Scientific returns $\neq f$(team size)
  - Human talent is distributed very broadly geographically
- Transition from data-poor to data-rich science
  - Chaotic ➔ Organized … However, *some* chaos (or the lack of excessive regulation) is good, as it correlates with the creative freedom (recall the WWW)
- Computer science as the "new mathematics"
  - It plays the role in relation to other sciences which mathematics did in ~ 17^{th} - 20^{th} century

    (The frontiers of mathematics are now elsewhere…)

The Fundamental Roles of Research/University Libraries

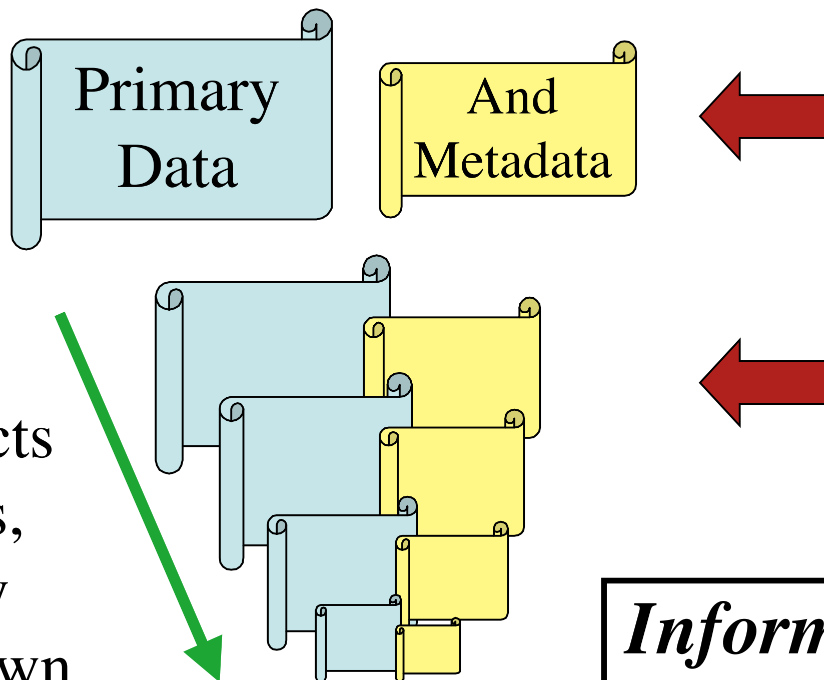
To preserve, organize,
and provide/facilitate access
to scientific and scholarly
data and results

This purpose is constant, but the implementation and functionality evolve.

What should the libraries become in the 21st century?

The Concept of Data (*and* Scientific Results) is Becoming More Complex

Data
- Actual data (preserved)
- Virtual data (recomputed as needed)



Produced and often archived by the primary data providers

Produced and published by the domain experts

Derived Data Products And Results, Increasingly Distilled down

Information is cheap, but expertise and knowledge are expensive!

Scientific Publishing is Changing

- Journals (and books?) are obsolete formats; must evolve to accommodate data-intensive science
- Massive data sets can be only published as electronic archives - and should be curated by domain experts
- Peer review / quality control for data and algorithms?
- The rise of un-refereed archives (e.g., archiv.org): very effective and useful, but highly heterogeneous and unselective
- A low-cost entry to publish on the web
  - Who needs journals?
  - Will there be science blogs?
- Persistency and integrity of data (and pointers)
- Interoperability and metadata standards

Research Libraries for the 21st Century

- How should research libraries evolve in the era of information abundance and complexity?
- What should be their roles / functionality?
  - Data discovery services
  - Data provider federators
  - Primary and/or derived data archivers
  - How much domain expertise should be provided?
  - Quality control?
  - Relationship with web portals and search engines?

  } ***Libraries As Virtual Organizations?*** (applies to the first three items)
- Is this too much for a single type of an institution?
  - Are libraries obsolete (inadequate)?
  - Should they split into several types of institutions?

VO Summary

- National/International Virtual Observatory is an ***emerging framework*** to harness the power of IT for astronomy with massive and complex data sets
  - Enable data archiving, fusion, exploration, discovery
  - Cross the traditional boundaries (wavelength regimes, ground/space, theory/observation …
  - Facilitate inclusion of major new data providers, surveys
  - Broad professional empowerment via the WWW
  - Great for E/PO at all educational levels
- It is ***inherently multidisciplinary***: an excellent synergy with the applied CS/IT, statistics…and it can lead to new IT advances of a broad importance
- It is ***inherently distributed*** and web-based

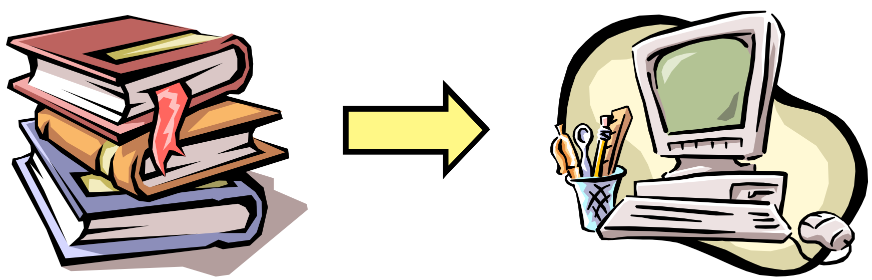
But It Is More General Than That:

- Coping with the data flood and extracting knowledge from massive/complex data sets is ***a universal problem facing all sciences today:***

  Quantitative changes in data volumes + IT advances:
  → ***Qualitative changes in the way we do science***
- (N)VO is an example of ***a new type of a scientific research environment / institution(?)*** in the era of information abundance
- This requires ***new types of scientific management and organization structures***, a challenge in itself
- The real intellectual challenges are methodological: how do we formulate ***genuinely new types of scientific inquiries, enabled by this technological revolution?***

… and the Evolution of Libraries

- Scientific / research ***libraries must evolve***, in order to stay useful in the era of data-intensive, computation-based science
  - Database technologies are essential
  - Fusion with data exploration technologies will be next
  - A growing importance of domain expertise
  - Blending in the web, then semantic web?



For more details and links, please see

http://www.astro.caltech.edu/~george/vo/