



The National Virtual Observatory

The heavens at your fingertips

© 2002 Nature Publishing Group NATURE | VOL. 420 | 21 NOVEMBER 2002 | www.nature.com/nature

Dr. Robert Hanisch
Space Telescope Science Institute
Baltimore, Maryland



The National Virtual Observatory



- National Academy of Sciences “Decadal Survey” recommended NVO as highest priority small (<\$100M) project

“Several small initiatives recommended by the committee span both ground and space. The first among them—the National Virtual Observatory (NVO)—is the committee’s top priority among the small initiatives. The NVO will provide a “virtual sky” based on the enormous data sets being created now and the even larger ones proposed for the future. It will enable a new mode of research for professional astronomers and will provide to the public an unparalleled opportunity for education and discovery.”

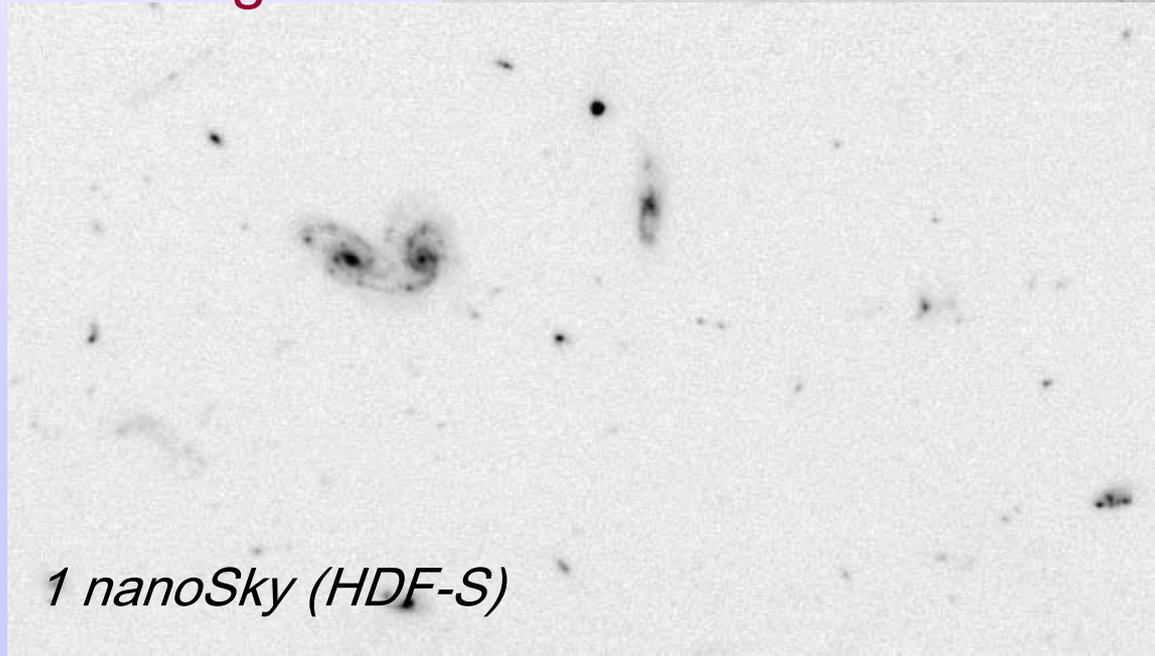
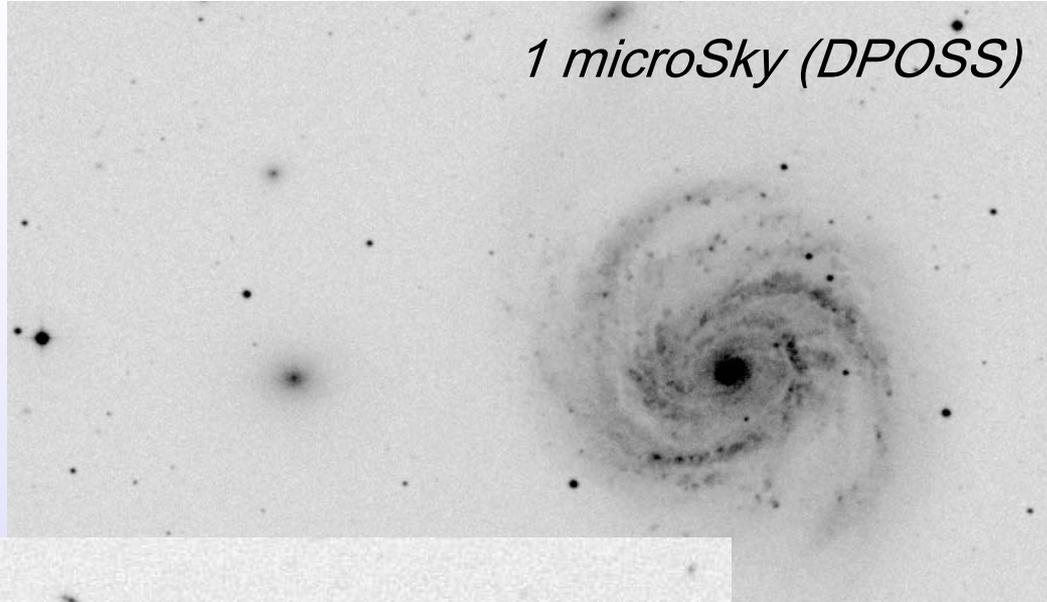
*—Astronomy and Astrophysics in the New Millennium,
p. 14*



Astronomy is Facing a Data Avalanche



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



Billions of
detected
sources,
hundreds of
measured
attributes
per source₃



The Changing Face of Observational Astronomy



- Large digital sky surveys are becoming the dominant source of data in astronomy: > 100 TB, growing rapidly
 - Current examples: SDSS, 2MASS, DPOSS, GSC, FIRST, NVSS, RASS, IRAS; CMBR experiments; Microlensing experiments; NEAT, LONEOS, and other searches for Solar system objects ...
 - Digital libraries: ADS, astro-ph, NED, CDS, NSSDC
 - Observatory archives: HST, CXO, space and ground-based
 - Future: QUEST2, LSST, and other synoptic surveys; GALEX, SIRTf, astrometric missions, GW detectors
- Data sets orders of magnitude larger, more complex, and more homogeneous than in the past
- Roughly 1 TB/Sky/band/epoch
 - Human Genome is < 1 GB, Library of Congress ~ 20 TB



Science of a Qualitatively Different Nature



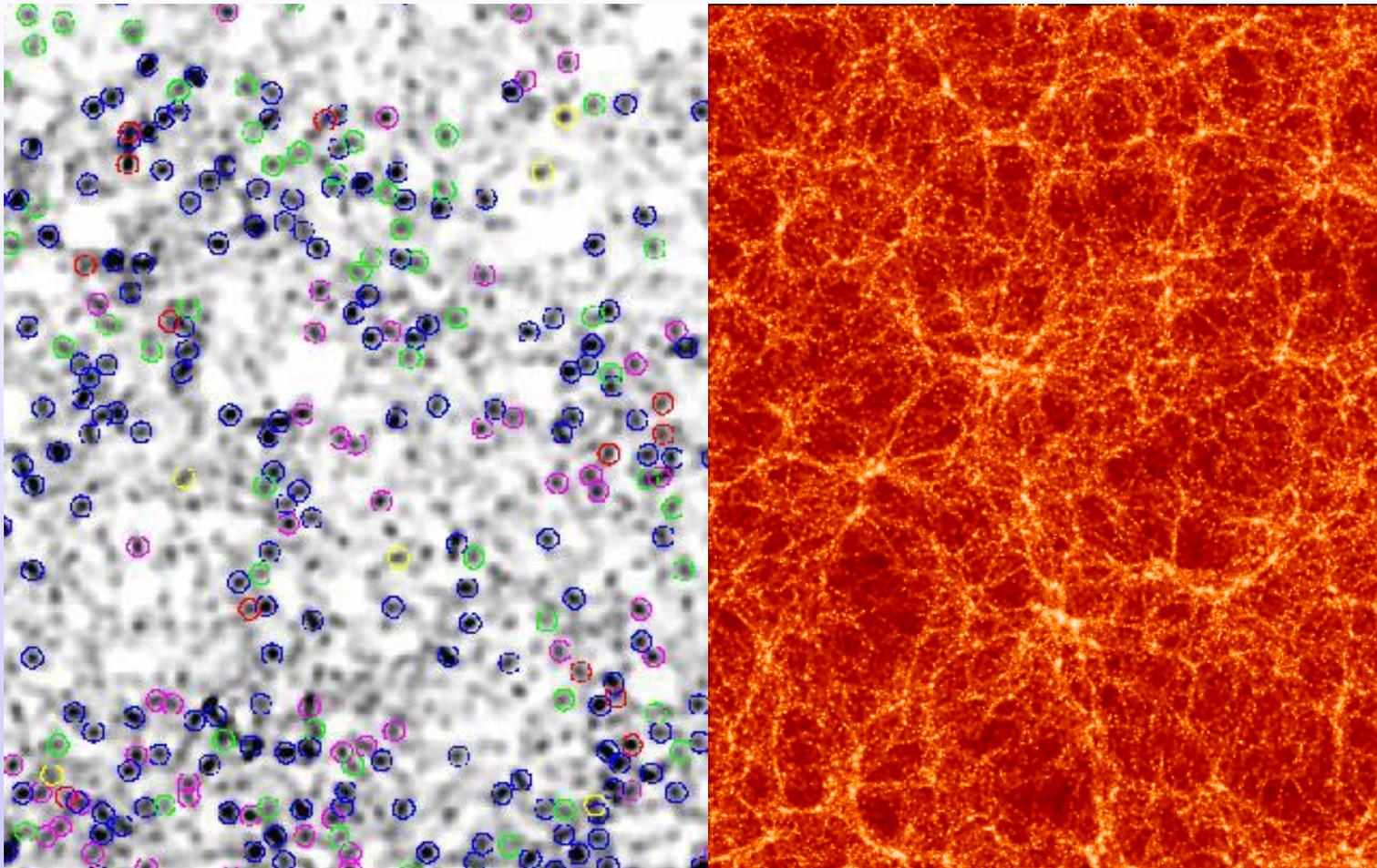
- **Statistical astronomy done right**
 - Precision cosmology, Galactic structure, stellar astrophysics ...
 - Discovery of significant patterns and multivariate correlations
 - Poissonian errors unimportant
- **Systematic exploration of the observable parameter spaces**
 - Searches for rare or unknown types of objects and phenomena
 - Low surface brightness universe, the time domain
 - Confronting massive numerical simulations with massive data sets



Precision Cosmology



and a better marriage of theory and observations



DPOSS Clusters

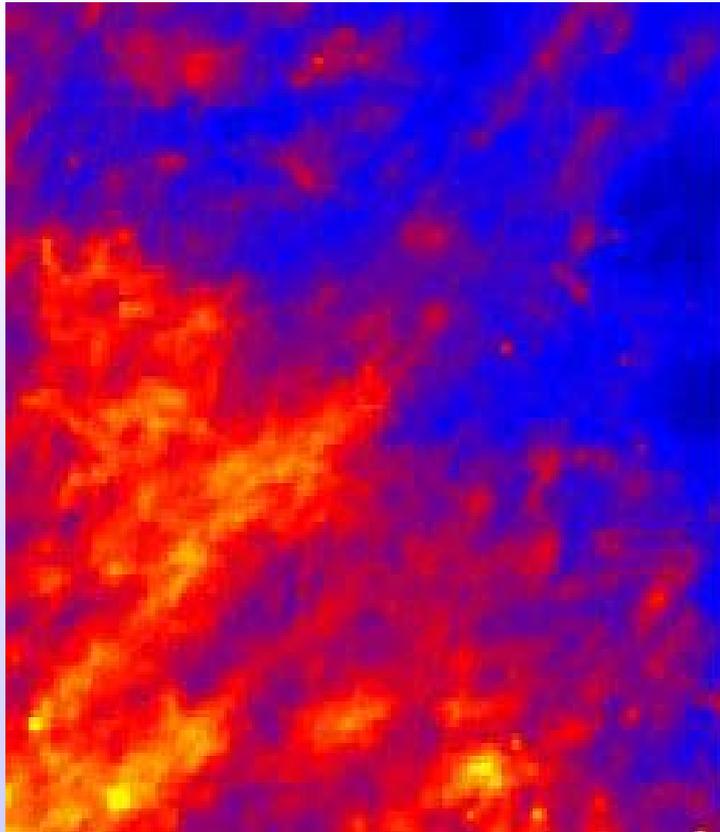
Numerical Simulation



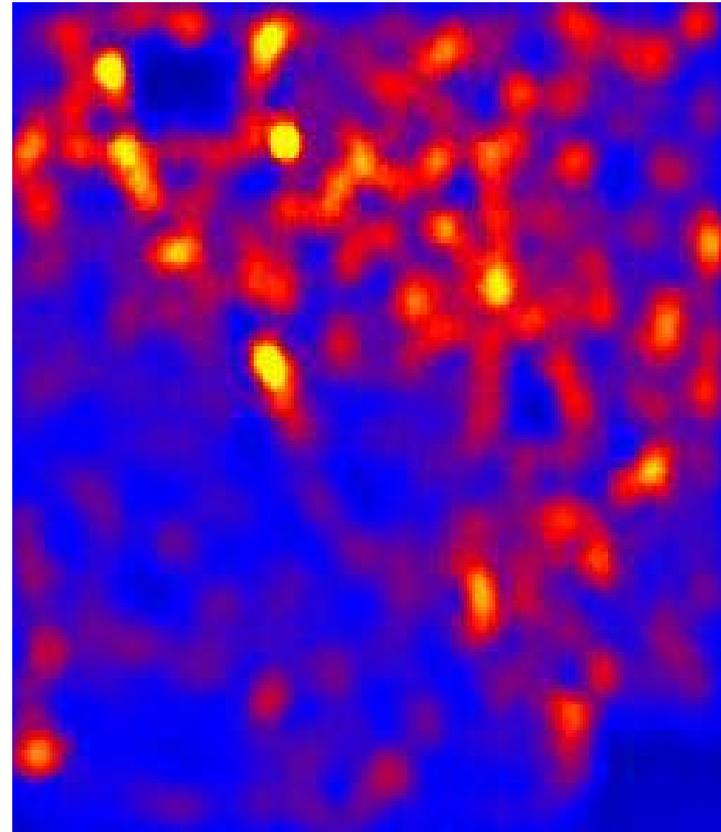
Multi-Wavelength Data



paint a more complete
(and more complex!) picture of the universe



*Infrared emission from
interstellar dust*



*Smoothed galaxy
density map*



A Panchromatic Approach to the Universe...

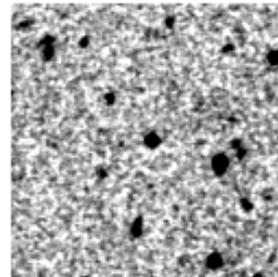


...reveals
a more complete
physical picture

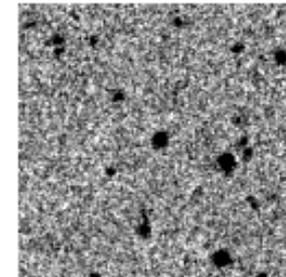
The resulting
complexity of
data translates
into increased
demands for
data analysis,
visualization, and
understanding



*Megaflares
on normal
main
sequence
stars
(DPOSS)*

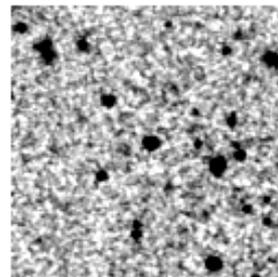


1988.3697

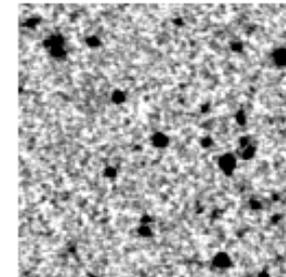


1988.4487

J

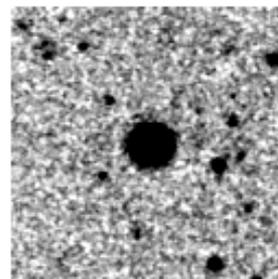


1991.2723

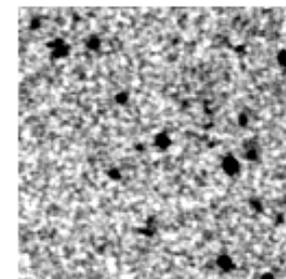


1994.3679

F



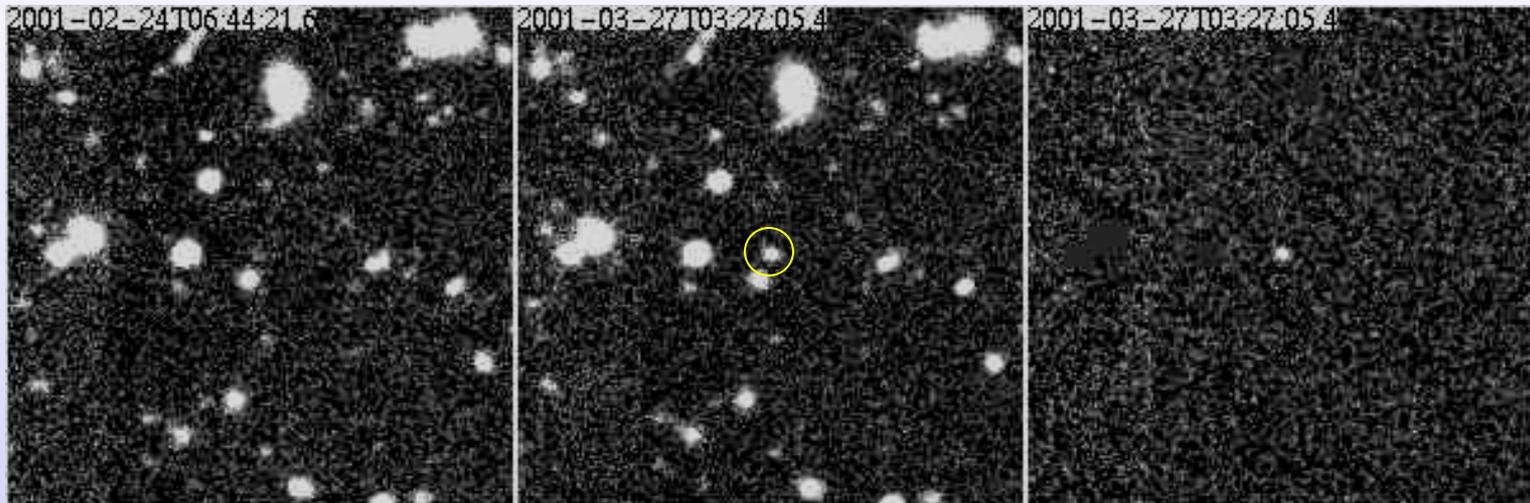
1990.1793



1997.3408

N

Faint, Fast Transients





Trends



- Astrophysical data is growing exponentially
 - Doubling every year (Moore's Law): both data sizes and number of data sets
- Computational resources scale the same way
 - Constant \$\$\$ will keep up with the data
- Main problem is the software component
 - Currently components are not reused
 - Software costs are increasingly larger fraction
 - Aggregate costs are growing exponentially



Discoveries



- **When and where are discoveries made?**
 - Always at the edges and boundaries
 - Going deeper, using more colors....
 - Physicists make many measurements and discard most; Astronomers make many measurements and find wealth in their entirety and combination (J. Ostriker, 6/14/02)
- **Metcalfe's law**
 - Utility of computer networks grows as the number of possible connections: $O(N^2)$
- **VO: Federation of N archives**
 - Possibilities for new discoveries grow as $O(N^2)$
- **Current sky surveys have proven this**
 - Very early discoveries from SDSS, 2MASS, DPOSS



Data Publishing Roles



Roles

Traditional

Emerging

Authors

Scientists

Collaborations

Publishers

Journals

Project www site

Curators

Libraries

Bigger Archives

Consumers

Scientists

Scientists

read->analyze

query-> analyze



Changing Roles



- **Exponential growth**
 - Data will be never centralized
- **More responsibility on projects**
 - Becoming Publishers and Curators
 - Larger fraction of budget spent on software
 - Lot of development duplicated, wasted
- **More standards are needed**
 - Easier data interchange, fewer tools
- **More templates are needed**
 - Develop less software on your own



Evolving Standards



- Astrophysics has good track record
- FITS: universally used to share low level data
 - Individual images, tables, files
- But: new industry standards emerging
 - XML, SOAP
- Required by modern data exchange
 - More dynamic (streams, queries)
 - Merging heterogeneous sources
- Good time to adopt...



Accessing Data: Today



1. Locate data from user supplied source
2. Download and study documentation
3. Identify necessary data components
4. Copy data to local machine
5. Read and filter data locally
6. Perform the analysis locally



Accessing Data: Soon



Phase 1

1. Auto-discovery of data, and documentation
2. Study documentation
3. Filter (query) data from remote source
4. Analyze incoming data stream directly

Phase 2

- Perform even analysis remotely, close to the data source



Emerging Concepts



- **Standardizing access to distributed data**
 - Web Services, supported on all platforms
 - Custom configure remote data dynamically
 - XML: Extensible Markup Language
 - SOAP: Simple Object Access Protocol
 - WSDL: Web Services Description Language
- **Standardizing access to distributed computing**
 - Grid Services
 - Custom configure remote computing dynamically
 - Build your own remote computer, and discard
 - Virtual Data: new data sets on demand



NVO: How Will It Work?



- Define commonly used `atomic' services
- Build higher level toolboxes/portals on top
- We do not build `everything for everybody'
- Use the 90-10 rule:
 - Define the standards and interfaces
 - Build the framework
 - Build the 10% of services that are used by 90%
 - Let the users build the rest from the components



Roles of the NVO



Professional:

- 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

Science Strategy and Planning:

- Identify gaps in our coverage of the observable parameter space; which new missions, instruments, experiments are likely to have the largest impact?

Education and Public Outreach:

- Unprecedented opportunities in terms of the content, broad geographical and societal range, for all educational levels
- Engaging the amateur astronomy community
- Astronomy as a magnet for CS/IT education
- Creating a new generation of science and technology leaders



The VO Gets Underway



- In October 2001 NSF awarded \$10M to a collaboration of 17 organizations to begin building the framework for the NVO
 - Astronomy data centers
 - National observatories
 - Supercomputer centers
 - University departments
 - Computer science/information technology specialists
- PI: Alex Szalay (JHU), CoPI: Roy Williams (Caltech/CACR), PM: Bob Hanisch (STScI), PS: Dave De Young (NOAO)
- Focus is on adapting emerging information technologies to meet the astronomy research challenges
 - Metadata, standards, protocols (XML, http)
 - Interoperability
 - Database federation
 - Web Services (SOAP, WSDL, UDDI)
 - Grid-based computing (OGSA)



Technical Developments (1)



- VOTable format is XML-based mark-up for astronomical tables and catalogs
 - V1.0 format definition released 15 April 2002; international agreement (NVO, AVO, AstroGrid)
 - Four s/w libraries have been implemented (Perl, Java, C++, C#) demonstrating platform independence and robustness of the format



Technical Developments



- **ConeSearch**
 - Search for catalog objects around a point
 - Returns data in VOTable format
 - Requires a registered profile
 - Point of the exercise
 - A learning experience
 - Existing archives test and implement VOTable
 - Understand service description issues
 - **Cross-Identification service built on top**
 - Accepts URLs of two ConeSearch services and returns VOTable of cross-matched objects
 - **Cone search services being utilized in science demos**



Technical Developments (3)



- Metadata management framework
- Data models
- Spatial-temporal metadata definitions
- Bandpass metadata definitions
- UCDs (Uniform Content Descriptors) mapped onto SDSS catalogs, topic maps
- Service and resource metadata definitions (Dublin Core)
 - To support service and resource registration
 - UDDI, OGSA, OAI
- Formed international “registry”, “dm” (data models), “semantics”, and “dal” (data access layer) discussion groups



Technical Developments (4)



- **Simple Image Access Protocol definition**
 - Generalization of ConeSearch
 - Image metadata returned in VOTable, links to images via URLs
 - Input specification: RA, DEC, SZ, +options
 - Output may be cutout, mosaic, atlas image, or pointed observation
 - Several implementations completed and will be utilized in science demonstrations



Simple Image Access Service



STScI Virtual Observatory Prototype Service

All input values should be in decimal degrees.

Service:	DSS.jsp
RA:	<input type="text" value="195"/>
Declination:	<input type="text" value="29"/>
Radius	<input type="text" value=".25"/>
FMT	<input type="text" value="image/fits"/>
SZ	<input type="text" value="0.5"/>
RASZ	<input type="text" value="0.0"/>
DECSZ	<input type="text" value="0.0"/>
GETIMAGE	<input type="text" value="false"/>
<input type="button" value="Submit Query"/> <input type="button" value="Reset"/>	



<!DOCTYPE VOTABLE (View Source for full doctype...)>

```

- <VOTABLE>
  <DESCRIPTION>DSS 2 Catalogue from STScI</DESCRIPTION>
  - <RESOURCE type="results">
    - <TABLE>
      <FIELD ID="title" ucd="NVOX:IMAGE_TITLE" datatype="double" />
      <FIELD ID="width" ucd="NVOX:IMAGE_PIX_WIDTH" datatype="int" />
      <FIELD ID="height" ucd="NVOX:IMAGE_PIX_HEIGHT" datatype="int" />
      <FIELD ID="size" ucd="NVOX:IMAGE_SIZE" datatype="int" />
      <FIELD ID="RA" ucd="POS_EQ_RA_MAIN" datatype="double" />
      <FIELD ID="Dec" ucd="POS_EQ_DEC_MAIN" datatype="double" />
      <FIELD ID="scale" ucd="NVOX:IMAGE_SCALE" datatype="double" />
      <FIELD ID="format" ucd="NVOX:IMAGE_FILE_FORMAT" datatype="char" arraysize="*" />
      <FIELD ID="url" ucd="NVOX:Image_AccessReference" datatype="char" arraysize="*" />
      <FIELD ID="epoch" ucd="NVOX:IMAGE_EPOCH" datatype="double" />
      <FIELD ID="naxes" ucd="NVOX:IMAGE_NAXES" datatype="int" />
      <FIELD ID="naxis" ucd="NVOX:IMAGE_NAXIS" datatype="int" arraysize="*" />
      <FIELD ID="crtype" ucd="NVOX:WCS_CoordSystem_CoordType" datatype="char" arraysize="*" />
      <FIELD ID="crpix" ucd="NVOX:WCS_CoordSystem_CoordRefPixel" datatype="double" arraysize="*" />
      <FIELD ID="crval" ucd="NVOX:WCS_CoordSystem_CoordRefValue" datatype="double" arraysize="*" />
      <FIELD ID="cdval" ucd="NVOX:WCS_CoordSystem_CDMatrix" datatype="double" arraysize="*" />
      <FIELD ID="pixSizeX" ucd="" datatype="double" />
      <FIELD ID="pixSizeY" ucd="" datatype="double" />
      <FIELD ID="plateRA" ucd="" datatype="double" />
      <FIELD ID="plateDEC" ucd="" datatype="double" />
      <FIELD ID="plateId" ucd="" datatype="char" arraysize="*" />
      <FIELD ID="positionOnPlate" ucd="" datatype="char" arraysize="*" />
    - <DATA>
      - <TABLEDATA>
        - <TR>
          <TD>DSS at STScI RED F</TD>
          <TD>3564</TD>
          <TD>3571</TD>
          <TD>12727044</TD>
          <TD>-10000.0</TD>
          <TD>-10000.0</TD>
          <TD>67.19999694824</TD>
          <TD>image/fits</TD>
        - <TD>
          <![CDATA[ http://archive.stsci.edu/cgi-bin/dss_plate_finder?

```





Education & Outreach



- **Began in earnest with a workshop last summer**
 - Understand requirements on NVO services from perspective of formal education, informal education, commercial/corporate, and public outreach content developers
 - Draft requirements document reviewed by workshop participants and being finalized
- **High priority activities include:**
 - Incorporation of EPO-related metadata into emerging VO standards
 - Coordinate system information in press release images
 - VO-enabled archive for amateur astronomy imaging (Sky & Telescope collaboration)



Science Prototypes

- **Science demonstration projects**
 - Assure that technology responds to science requirements
 - Provide evidence to the user community that progress is being made
- **Gamma Ray Burst Follow-up Service: quickly show me the multi-wavelength sky at position (α, δ)**
- **Brown Dwarf Search: cross-correlate large, distributed multi-wavelength catalogs and look for unusual objects**
- **Galaxy Cluster Morphology and Evolution: measure morphological parameters for large numbers of galaxies using the Grid**

NVO Gamma-Ray Burst Follow-Up Service Science Prototype

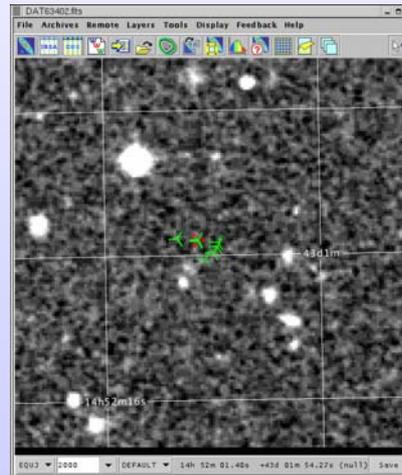


Scientific Motivation: Rapid collection of multi-wavelength imaging, catalog and observation data following an interesting transient event is essential. This service can also be used as a general tool to quickly access all data available on any patch of sky for any science use.

Data Resources: Multi-wavelength data from any number of sites (currently 13 different sites) sampling energies from X-ray to radio, and including images, object lists, and catalogues of observations.

What the VO Brings: Integration and organization of a variety of data sources into an easily comprehensible information set. Scalability to an arbitrary number of data providers. Integrates data with multiple data visualization services.

Event	Time	RA	Dec	Size	Source
141	2001-02-22 12:00:00	14 52 12.00	+43 01 01.6	0.25	Other
Images					
Optical	<input checked="" type="checkbox"/> DSS(Blue)	<input checked="" type="checkbox"/> DSS(Red)			
X-ray	<input checked="" type="checkbox"/> RASS				
Radio	<input checked="" type="checkbox"/> NVSS	<input checked="" type="checkbox"/> WENSS	<input checked="" type="checkbox"/> FIRST		
Observations					
HST	<input checked="" type="checkbox"/> HST(141)				
X-ray	<input checked="" type="checkbox"/> Chandra(2)				
Objects					
Major Catalogs	<input type="checkbox"/> NED(25)	<input type="checkbox"/> GSC2.2(294)			
	<input type="checkbox"/> USNO A2(213)	<input type="checkbox"/> 2MASS(P)(236)	<input type="checkbox"/> 2MASS(X)(2)		
Clusters	<input checked="" type="checkbox"/> CEDAG(3)				
Stars	<input type="checkbox"/> AC2000.2(1)				
Analyze data in Aladin					
Analyze data in OASIS					
Download selected data					



Enabling Technologies: Standard protocols to remote services such as Cone Search and Simple Image Access, standardized VOTables for data retrieval transformation, and standardized semantics encoded as Uniform Content Descriptors (UCDs).

Future Prospects: Automated discovery of data sources; customization and quality control of resources searched; more sophisticated use of metadata.

Positions of HST and Chandra observations for GRB010222



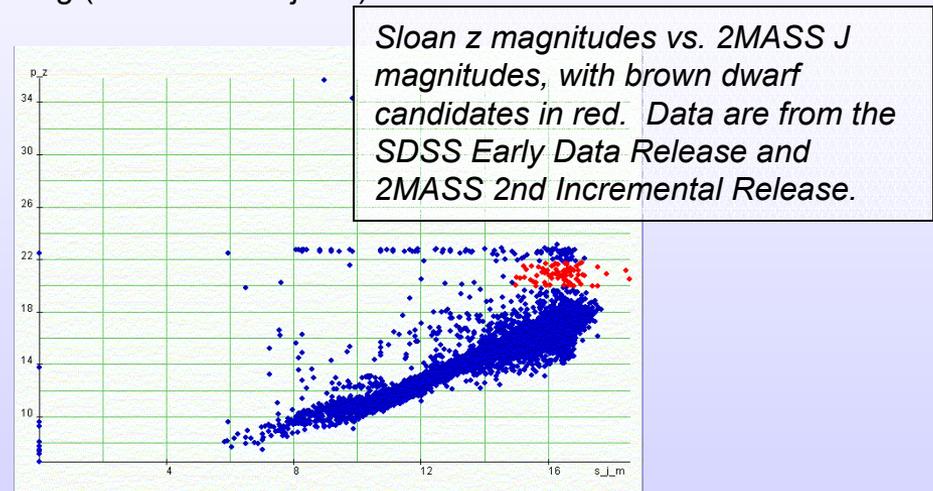
Brown Dwarf Search Science Prototype



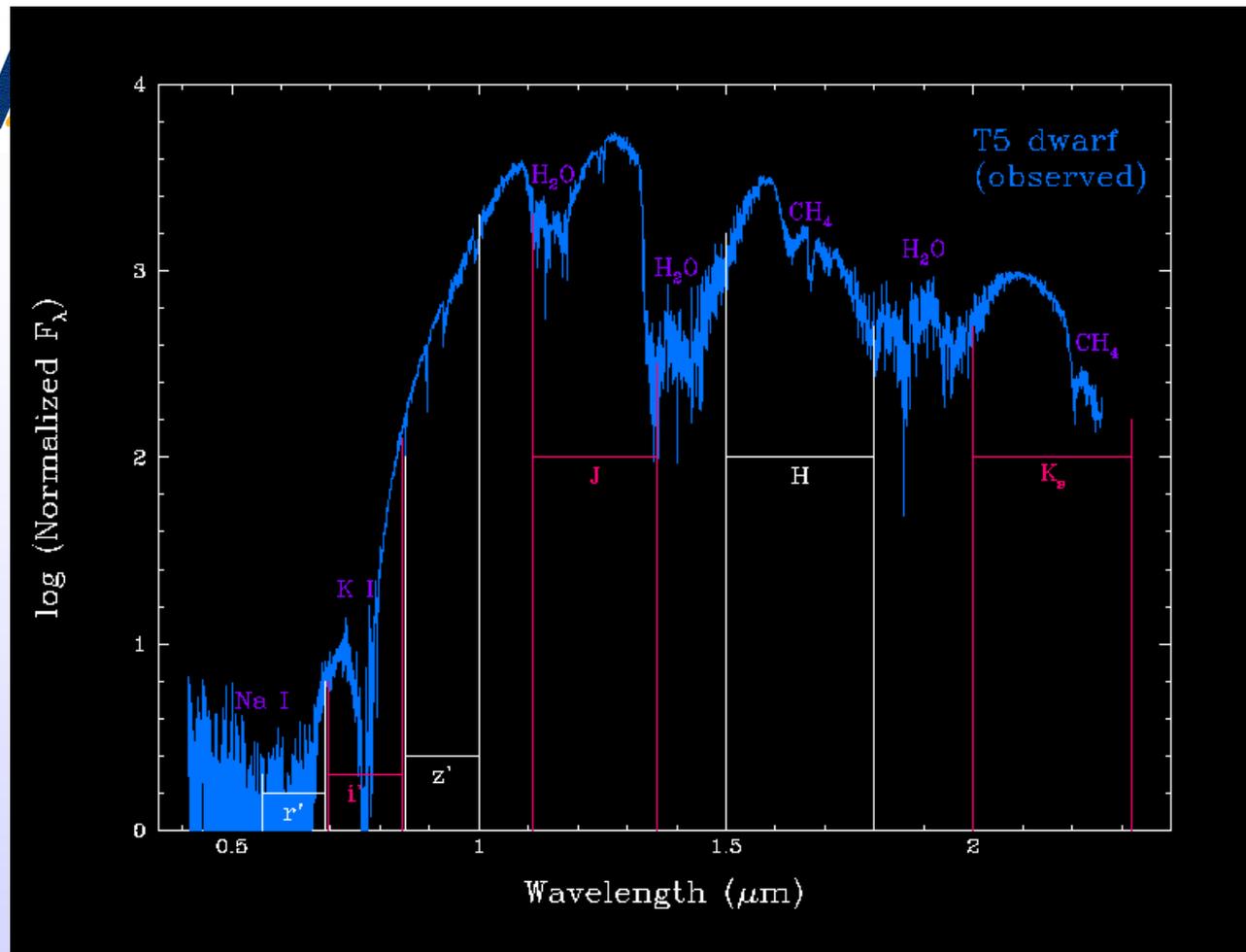
Scientific Motivation: The search for brown dwarfs has been revolutionized by the latest deep sky surveys. A key attribute to discovering brown dwarfs is the federation of many surveys over different wavelengths. Such matching of catalogs is currently laborious and time consuming. This matching problem is generic to many areas of astrophysics.

Data Resources: Sloan Digital Sky Survey (SDSS) Early Data Release (15 million objects) 2-Micron All Sky Survey (2MASS) 2nd Incremental Point Source Catalog (162 million objects)

What the VO Brings: Today, doing datasets is user-intensive and is replicated by many different users. Also, the correlation of these two datasets can take years of CPU time if not done correctly. The NVO brings two key aspects to this problem. First, it removes the need for the user to download large data to their machine, making direct use of distributed data. Second, the matching algorithm used here is computationally efficient and designed to give answers in minutes rather than hours; results can be returned to the user in real-time.



Future Prospects: Catalog matching of large datasets is a generic problem in astrophysics. Therefore, making the matching facility available to any user for use on any dataset will greatly enhance the productivity of scientists. Standard I/O formats allow developers to create tools to use the matched data and easily integrate with existing visualization and analysis tools (anomaly detector). Bringing these data together on remote machines with enough CPU to perform analysis (Grid technology) will allow cross-comparisons of unprecedented scale.



As a T dwarf becomes cooler (i.e., methane and water absorptions increase) or more distant...

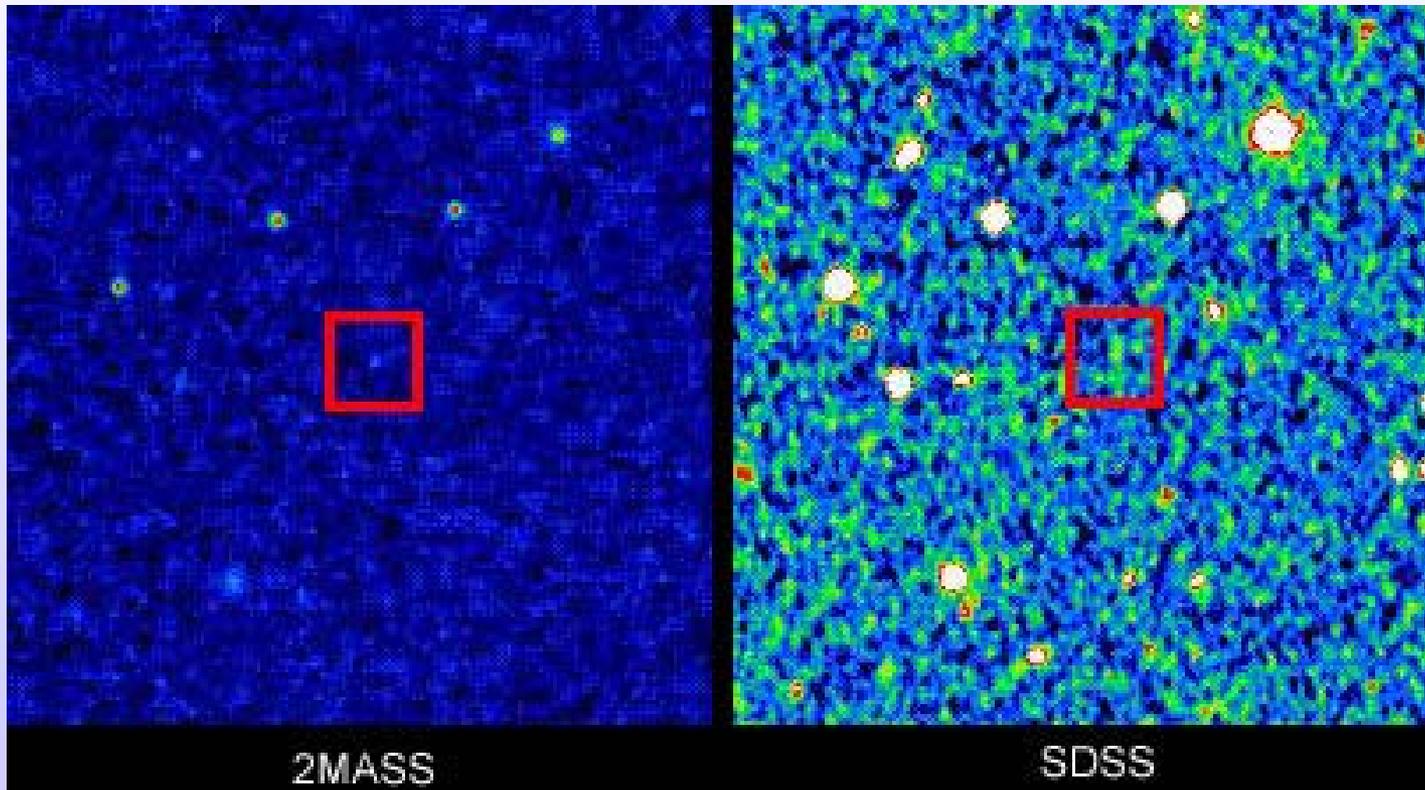
- SDSS detects it only at z' band
- 2MASS detects it only at J band



Demo Leads to Discovery!



- New brown dwarf candidate confirmed spectroscopically with Keck Observatory





Galaxy Morphology Science Prototype



Scientific Motivation: Investigate the dynamical state of galaxy clusters and galaxy evolution within the context of large-scale structure. Use galaxy morphology as a probe of dynamical history by calculating, for each galaxy in a cluster:

- Surface brightness
- Concentration index
- Asymmetry index

These parameters are analyzed with other indicators such as magnitude, color, peculiar velocity, position in cluster, and cluster large-scale structure.

Data Resources:

Chandra X-ray image (SAO/CXC)
ROSAT image (GSFC/HEASARC)
DSS image (STScI/MAST)
Galaxy cluster catalogs (NED)
CNOC1 cluster images and catalogs (CADC)

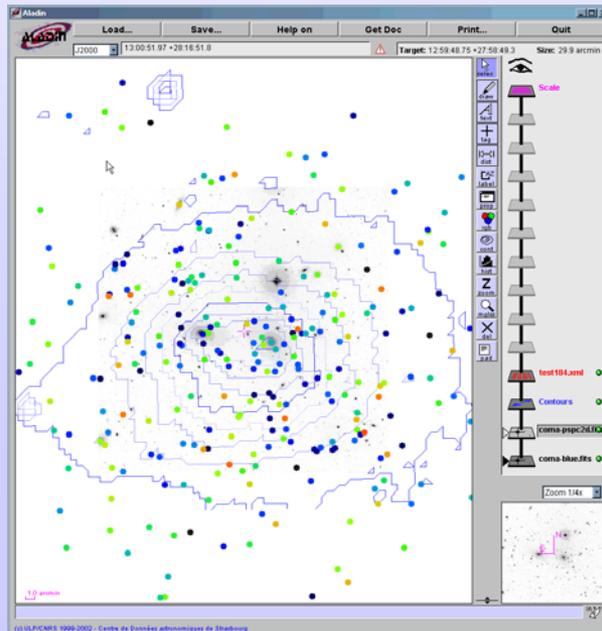
Computing Resources:

USC/ISI
UW-Madison/NCSA
Fermilab

What the VO Brings: Distributed data access and Grid-based computing make possible for the first time effective integration of multiple datasets and real-time computing. Integration of data from diverse sources is enabled by standardized data objects and standardized remote computing services. Flexibility of access means that further NVO-compliant images and catalogs can be added easily. Users can select their visualization portal (Aladin, OASIS, DS9).

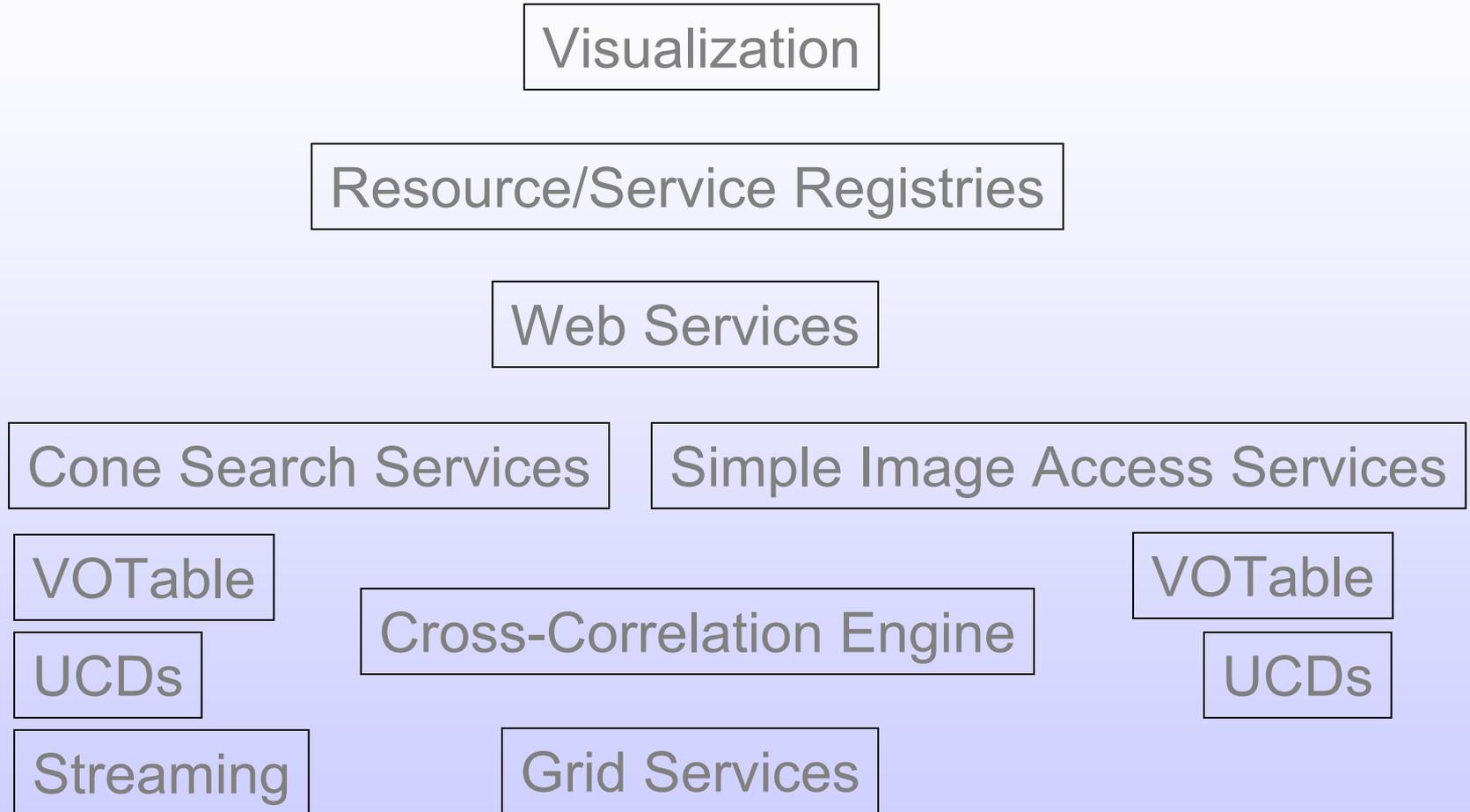
Enabling Technologies: VOTable, NVO-compliant catalog and image access, standard semantics, Grid computing infrastructure.

Future Prospects: Dynamic discovery and selection of image, catalog, and computing resources. User-selection of analysis tools and ability to publish data to the NVO framework.





NVO Components





NVO Components



*GRB Event
Follow-Up
Service*

Visualization

Resource/Service Registries

Web Services

Cone Search Services

Simple Image Access Services

VOTable

VOTable

UCDs

UCDs

Cross-Correlation Engine

Streaming

Grid Services



NVO Components



*Brown Dwarf
Search*

Visualization

Resource/Service Registries

Web Services

Cone Search Services

Simple Image Access Services

VOTable

VOTable

UCDs

Cross-Correlation Engine

UCDs

Streaming

Grid Services



NVO Components



*Galaxy
Morphology*

Visualization

Resource/Service Registries

Web Services

Cone Search Services

Simple Image Access Services

VOTable

VOTable

UCDs

Cross-Correlation Engine

UCDs

Streaming

Grid Services

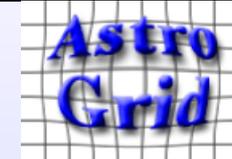


International Collaboration



- **European initiatives underway**

- Astrophysical Virtual Observatory funded by European Commission (€3.3 million, three years)
- AstroGrid, funded by UK e-science program (£5 million, three years)



- **Other international efforts:**



– Canada (C\$4M)

– German AVO (~ €2M)



– VO India
(4 programmers)

– Russian VO 



– VO Japan

– e-Astronomy Australia
(~ A\$500k x 2-3 years)



– VO China

– IVOA





VO: Here, There, & Everywhere



The idea of the Virtual Observatory is building rapidly and developing concurrently in many fields

- Astronomy: US NVO, CVO, UK AstroGrid, EC AVO, GAVO, VOIndia, RVO, e-Astronomy Australia, ChinaVO, JapanVO; heritage from AstroBrowse, ISAIA, SkyView, DigitalSky, etc., and NASA archives
- Solar: US Virtual Solar Observatory (NSO, Stanford, Montana State Univ., NASA), European Grid for Solar Observations
- Atmospheric Science: National Virtual Aeronomical Observatory
- Space Physics: Space Physics VO, PhysiBrowse (CDPP, France, and NASA), SSDS initiatives
- Physics: GriPhyN, iVDGL
- Bioinformatics
- Digital Library



The VO Vision



- The VO is the “semantic web” for astronomy (Tim Berners-Lee)
- The VO democratizes astronomical research
- The VO brings the universe to your desktop
 - The professional astronomer
 - Graduate students
 - Undergraduates
 - K-12
 - Amateurs
 - The public