Astronomy in the Era of Information Abundance, and the Virtual Observatory Concept

S. George Djorgovski

Lecture 1

Inaugural BRAVO Lecture Series,
São José dos Campos, July 2007

Allow me to introduce myself …

- ***S. George Djorgovski***
- Prof. of Astronomy, and Co-Director, Center for Advanced Computing Research (CACR), at Caltech
- PI of the Digital Palomar Observatory Sky Survey (DPOSS), and Co-PI of the ongoing Palomar-Quest sky survey (PQ)
- Co-founder, Virtual Observatory movement
- Astronomer by training, but really like computational science
  - *How is computation and information technology changing the way we do science - and think about understanding the world?*
- Reinaldo's old buddy

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

The Four Lectures

1. Astronomy in the era of information abundance, and the virtual observatory (VO) concept
2. Examples of VO-enabled science
3. Clustering, classification, and data exploration tools
4. The big picture: information technology revolution, and science in the 21st century

Please interrupt and ask questions!

Your instantaneous feedback will help us all

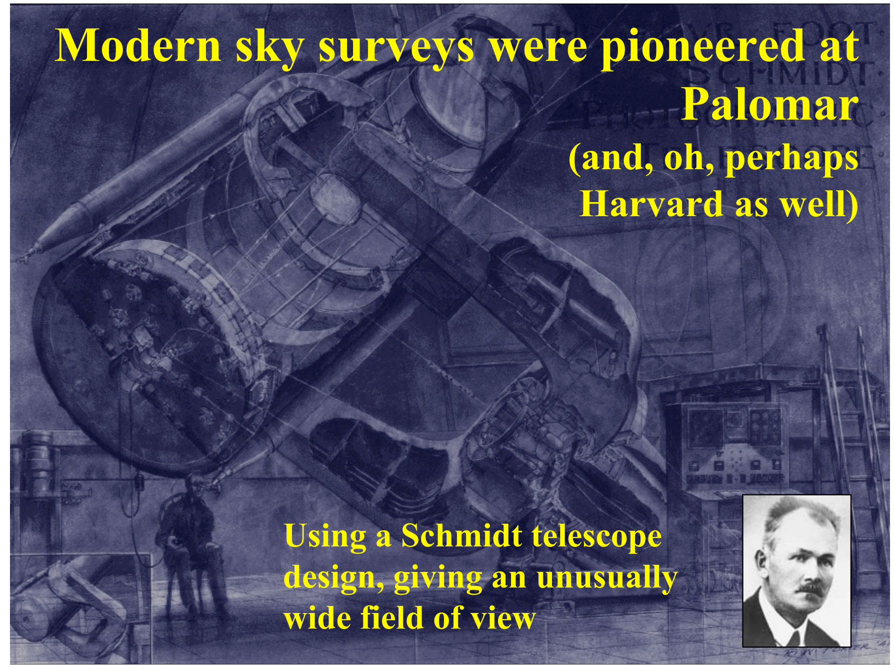
Some material will be familiar or too easy for some of you

Copies of my slides will be made available - you don't need to take furious notes; instead, pay attention and ask questions!

Informal discussions may be more valuable than the lectures …

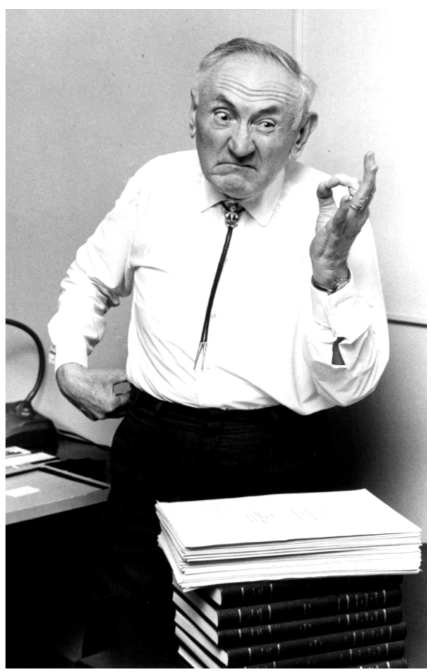
The Key Ideas

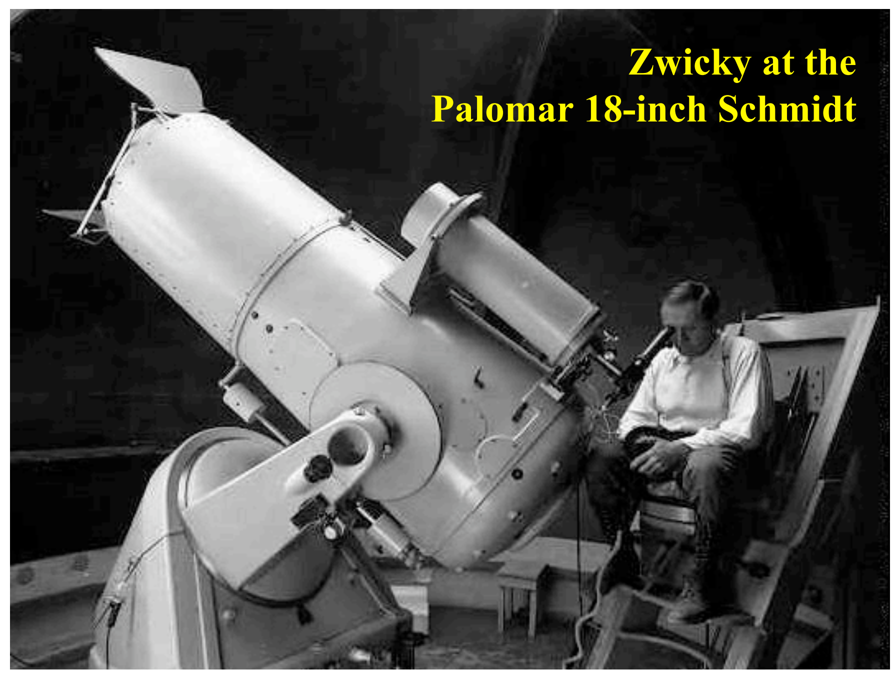
- Astronomy has become an immensely data-rich science, and the volume and complexity of data are growing exponentially
- Major digital sky surveys are the dominant data sources; they enable a broad variety of science
- Our ability to extract knowledge from the data quickly and effectively has been lagging
- The Virtual Observatory (VO) concept was developed as a solution to these challenges: deploy the modern information technology (IT) and applied computer science (CS) in service of a domain science (astronomy) - and drive their development in turn
- The goal is to enable new, qualitatively different science, which would not be practical or even possible otherwise



Fritz Zwicky, the original mad genius of Caltech, **pioneer of sky surveys,** discoverer and predictor of the large-scale structure, the dark matter, gravitational lensing, the nature of supernovae, neutron stars and black holes, and many other weird and wonderful things besides …

… and a champion of systematic exploration of observable parameter spaces





Edwin Hubble at the Palomar 48-inch Schmidt, the surveying telescope for the Hale 200-inch (now named the Samuel Oschin Telescope)

The milestone Palomar Sky Surveys were done with this instrument

Going Going

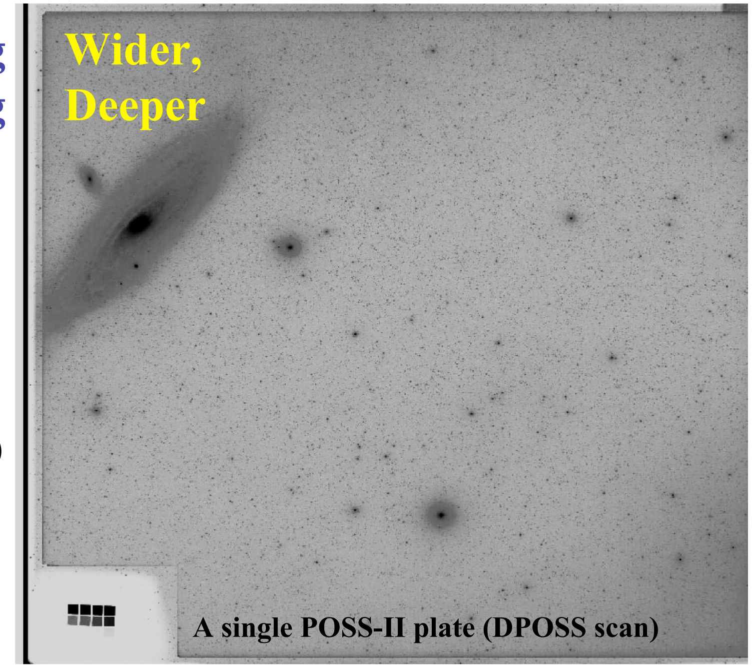
Moon:

Fields of view of:

P200:

Keck:

HST:

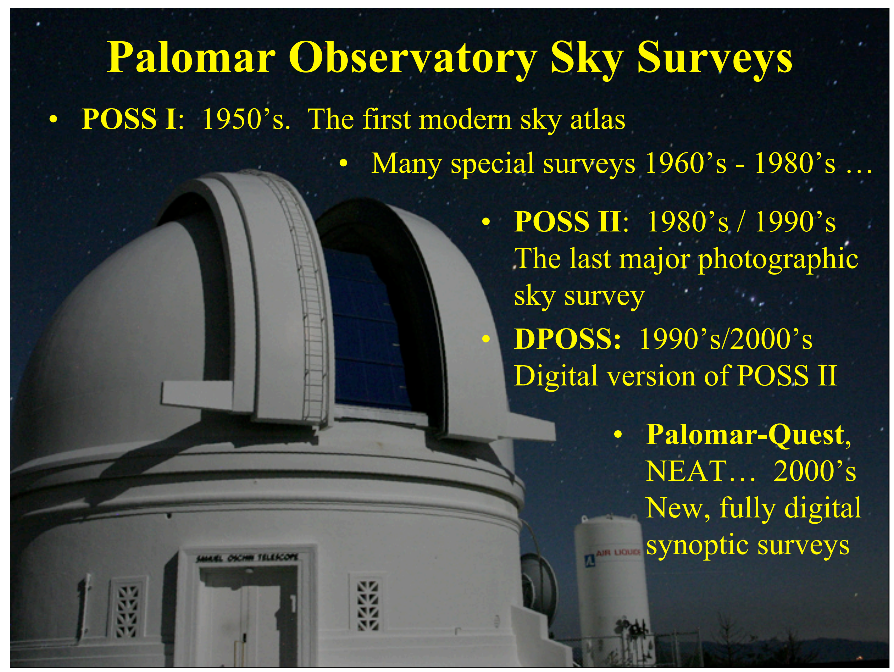


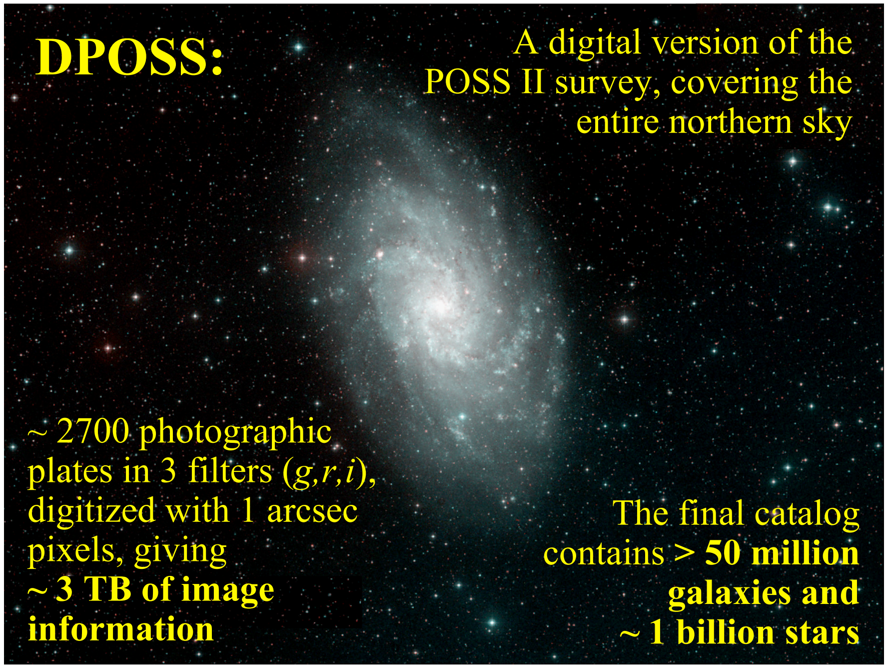
Sky Surveys

Initial Exploration → Systematic Studies → t

Improving Technology

- Division by λ (radio to γ-ray), by type (imaging, spectroscopic, mixed, etc.), or by sampling (targeted, panoramic)
- Goals: (1) exploration of observable parameter space; (2) creation of statistical samples of sources; (3) searches for particular types of objects or phenomena
- Issues: data processing; calibration, testing, validation and calibration; archiving and data distribution



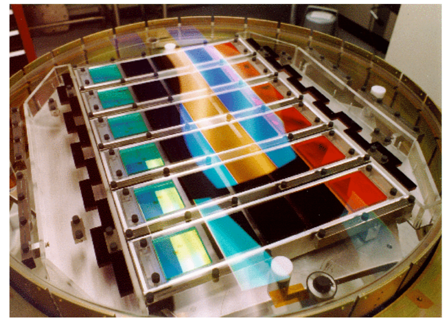


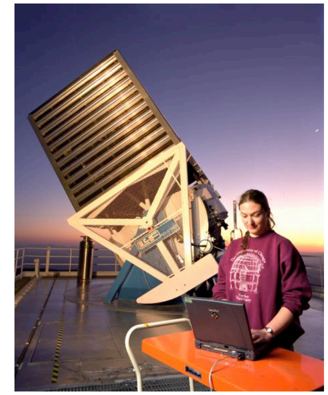
A major technology shift in the 1990's:

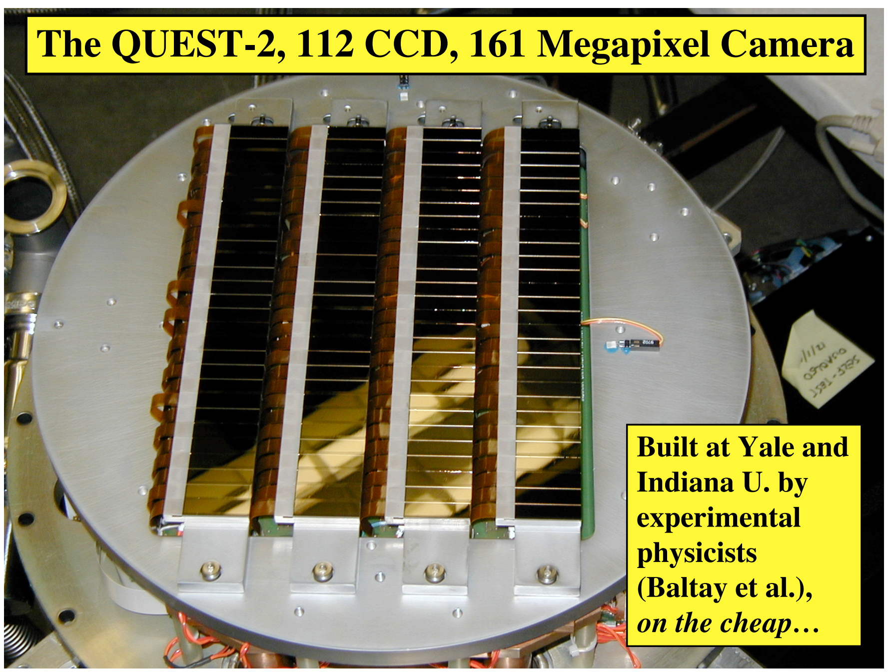
From digitized photographic plates (non-linear, hard to calibrate, poor dynamical range, poor noise characteristics, messy…) to fully digital (e.g., CCD) panoramic detector arrays - a great improvement in data quality!

SDSS was the first major CCD-based survey

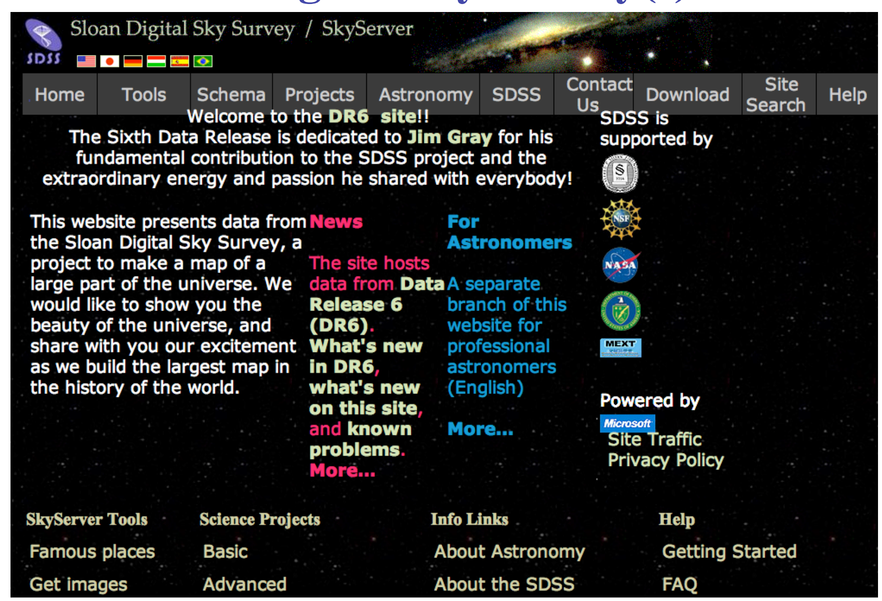
(However, we *still* don't have a fully CCD-based, all-sky survey!)



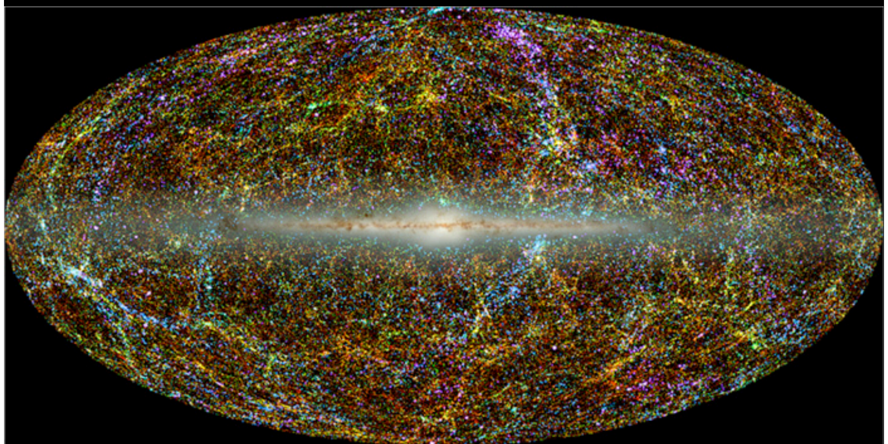




The Sloan Digital Sky Survey(s) = SDSS



Infrared Sky as seen by the 2-Micron All-Sky Survey (2MASS)

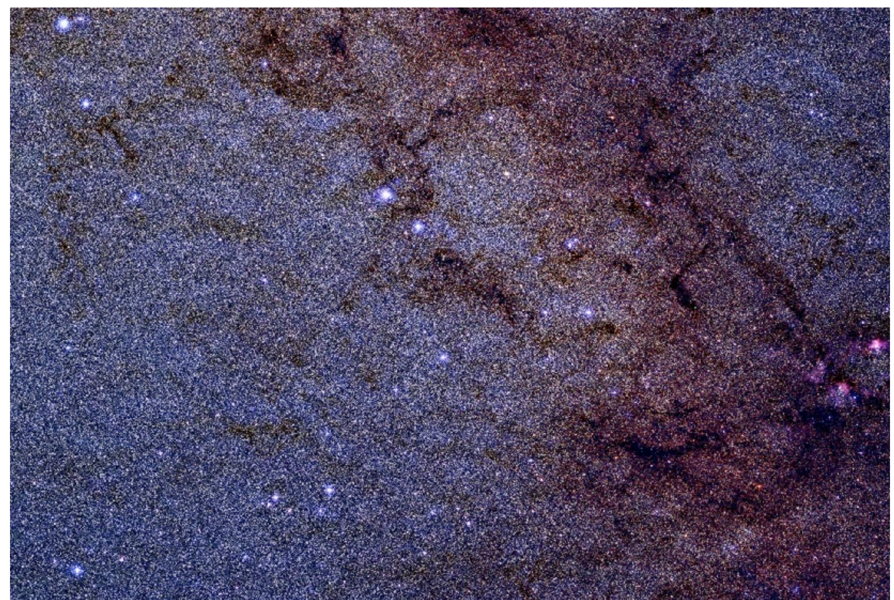


The Science of Sky Surveys

1. Exploring an observable parameter space
   - Especially if some new capability opens up, e.g., a new wavelength regime (radio, x-ray, etc.), a higher angular, temporal, or wavelength resolution, etc.
2. Global and statistical studies
   - Large-scale structure of the universe, structure and contents of our Galaxy, families of astrophysical objects, etc.
3. Selecting interesting objects or samples for detailed follow-up observations
   - Large telescope or space observatory time is too precious, use it well

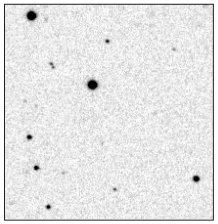
A single survey usually serves multiple scientific needs - most of which were not envisioned by its planners

A Small Portion of the DPOSS Galaxy Distribution Map



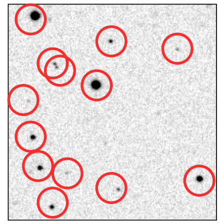
Galactic Center Region (a tiny portion) 2MASS NIR Image

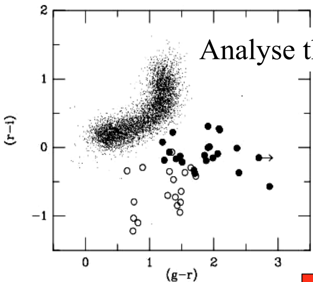
From Pictures to Knowledge



⬅ Raw data into calibrated images

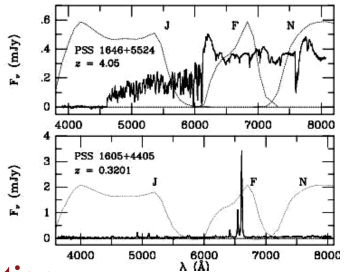
Detect sources and measure their attributes (brightness, position, shapes, etc.) ➡





Analyse the data…

Obtain follow-up spectra ➡

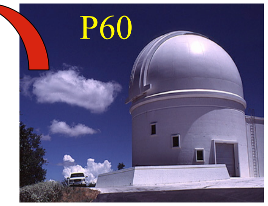


Physical interpretation and understanding

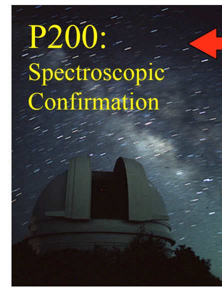
A Synergy of Telescopes

POSS-II Sky Survey

Digital Palomar Obs. Sky Survey (DPOSS)

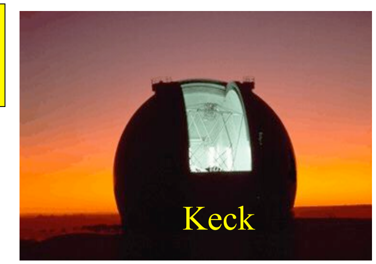


CCD Calibrations



Data Mining: High-z QSO Candidate Selection

Follow-up Studies …

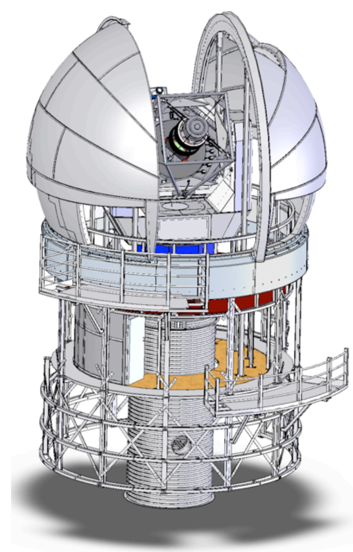


Now: Palomar-Quest Survey

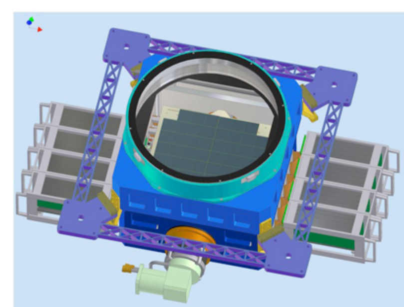
Some Current Projects

- SDSS-II: Legacy survey, SEGUE, SN survey …
- ***Many*** surveys with ~ 0.1 - 2 m class telescopes, generally for SNe, GRB follow-up, microlensing, asteroids, or variable stars: RAPTOR, OGLE, ASAS, CASE, PVS, HAT, DIRECT, Ystar, MACHO, EROS, MEGA, MOA, AGAPE, PLANET , ROTSE, LOTIS, KAIT, Spacewatch, LINEAR, LONEOS, Catalina, MISAO, MPC, VSNET, TAOS, TASS, STARE, STARDIAL, MONET, ConCam, etc.
- And of course, Palomar: NEAT, M. Brown, PQ, SNF
- UKIDSS in near-IR
- Deep surveys: CFHTLS, COSMOS, NDWFS, DLS, Subaru, various deep fields… (typically 1 - few deg^2)
- Many at other wavelengths: radio to X-ray to gamma…

Forthcoming Projects: PanSTARRS



(1-4) 1.8 m diameter telescopes
1.4 Gigapixel cameras
Cover up to 6000 deg^2 / night
Limiting mags ~ 24, *ugrizy*
PanSTARRS-1: 2008?
PanSTARRS-4: 2012?

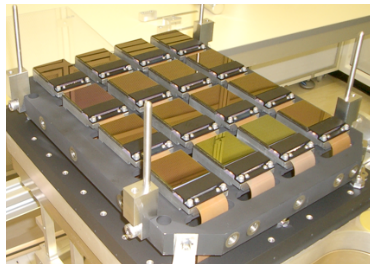


Forthcoming Projects: VST and VISTA

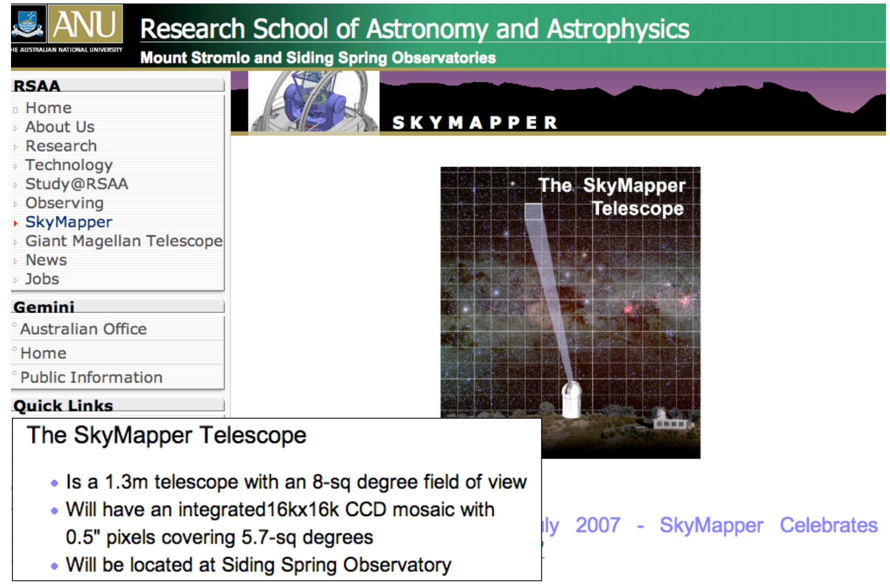
Survey telescopes at ESO

VISTA: ~ 4 m diameter, 67 Mpix, IR (ZYJHK bands), FOV ~ 1.65 deg

VST: ~ 2.6 m diameter, 256 Mpix, optical bands, FOV ~ 1 deg



The SkyMapper Project



The SkyMapper Project

Approximately 75% of the time on SkyMapper will be initially dedicated to the Southern Sky Survey.

Features of the Southern Sky Survey include

- Multi-colour, multi-epoch of all 20000 sq. degrees south of equator (ugriz filter + stromgren-like v)
- Data supplied to the community via Virtual Observatory
- Star and Galaxy photometry (3% absolute calibration)
- Astrometry (better than 50 mas)
- Digital images available for download photometrically calibrated, with accurate World Coordinate Systems, both single images and combined images.
- Cadence: 0, +4 hours, +2-3 days, +1-2 weeks and +1-2 years

Expected Survey Depth (1.5" seeing) for signal-to-noise of 5 in AB mags

	u	v	g	r	i	z
1 epoch	21.5	21.3	21.9	21.6	21.0	20.6
expt. time (sec)	110	110	110	110	110	110
6 epochs	22.9	22.7	22.9	22.6	22.0	21.5

In addition, a 5-Second Survey will be undertaken in photometric conditions for calibration of stars from 9-16th magnitude in all bands. This will provide the calibration of the survey and will allow the survey to be tied to the Hipparchos and Tycho catalogs (and other photometric standard systems that are established in the southern hemisphere) to ensure uniformity across the sky.

Southern Sky Survey

Southern Sky Survey Science Goals include:

- Census of bright end of TNO and Centaur population, especially off the ecliptic plane
- Galactic Census – metallicity, gravity, temperature, variability of 5 billion stars
- Calibration of 2dF, 6dF surveys
- Discovery of up to 50000 SNe over 5 years
- Phot-Z samples of z<0.5 galaxies for studies of Large Scale Structure (e.g. Int-Sachs wolf, Paczynski-Alcock)
- QSO discovery, variability, evolution
- Bright z>6 Quasars
- Digital reference for Radio, X-Ray, GRB instruments
- An astrometric and photometric basis for the Virtual Observatory

Non-survey science identified to be undertaken with the telescope includes

- Planet Transit studies
- Microlensing Studies
- Supernovae
- Widefield surveys in non-survey filters

The Dark Energy Survey

Science Program:

Four Probes of Dark Energy:

- Galaxy Cluster counting: 20,000 clusters to z = 1
- Weak lensing: 300 million galaxies with shape measurements over 5000 deg^2
- Spatial clustering of galaxies: 300 million galaxies to z = 1 and beyond
- Standard Candles: 2000 SN Ia, z = 0.3 - 0.8
- Start in 2009?

Instrument Description:

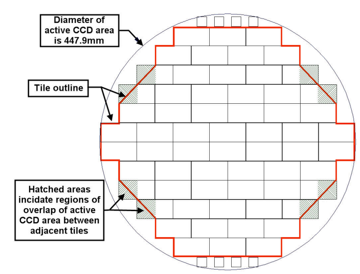
3 deg^2 camera at the CTIO 4-m

≥ 2.2 deg FOV

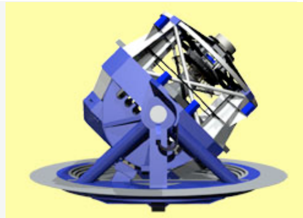
62 CCDs, 2k x 4k = 0.5 Gpix

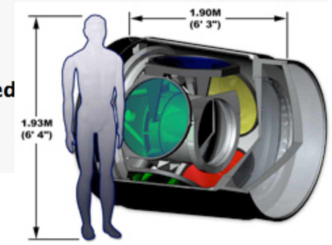
SDSS *griz* filters

Limiting mags ~ 24



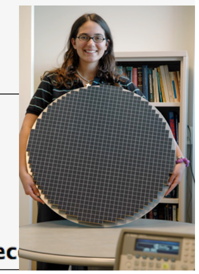
- Representative Key Science Missions:
    1. Dark energy
    2. Solar system survey
    3. Optical transients
    4. Galactic map
- First light schedule: Spring, 2014
- Sky coverage: 20,000 degrees^2 (General Survey)
- Standard cadence (per visit): 15 sec expose + 1 sec shutter overhead, 2 sec read, 15 sec expose + 1 sec shutter overhead, 2 sec read, 5 sec slew = 39 sec total
- Etendue (AΩ): 319 meter^{2}degrees^2
- Field of View: 3.5 degrees (9.6 square degrees)
- Effective clear aperture (On-Axis): 6.68m (adjusted
- Wavelength coverage: 320nm to 1080nm
- Number of active filters: five (ugrizY)
- Site: Cerro Pachon, Chile





- Nightly data generation rate
    - Raw pixel data: 15 Tbytes (16 bit)
    - Image through pipelines: 30 Tbytes raw science (32 bits) + 108 TB (32 bit) intermediate images
    - Archived images + metadata: 15 + 1 Tbytes (32 bits compressed to 16 bits)
    - Catalogs (transient phenomena): 1 Tbyte (32 bits compressed to 16 bits)
- Data release volume (average per release)
    - Source Catalog: 560 TB
    - Deep Object Catalog: 140 TB
- Yearly data archive rate (average)
    - Images: 6.5 Pbytes
    - Catalogs: 6.5 Pbytes
    - Metadata: 0.5 Pbytes

- Pixel count: 3.2 Gpixels
- Pixel pitch: 10 microns
- Readout time: 2 sec
- Dynamic range: 16 bits
- Nominal exposure time: 15 sec



Etendue (AΩ) Comparison

NB: This is just one figure of merit. Consider also the number of visits per unit time, data rate, site quality, etc.

Telescope	Diam. (m)	Area (m^2)	Solid Angle	AΩ
SDSS	2.5	4.9	1.5	7.3
P48/PQ camera	1.2	1.1	9+	10+
CFHT/Megacam	3.6	10	1.0	10
Subaru/Suprimecam	8.0	50	0.25	13
MMT/1-deg camera	6.5	33	1.0	33
Discovery Chan. Tel.	4.2	14	3.1	43
SkyMapper	1.3	1.3	5.7	7.6
PanSTARRS-1	1.8	2.5	1.7	15
PanSTARRS-4	3.6	10	7.0	60
LSST	6.5	33	7.0	230?

The Evolution of Technology

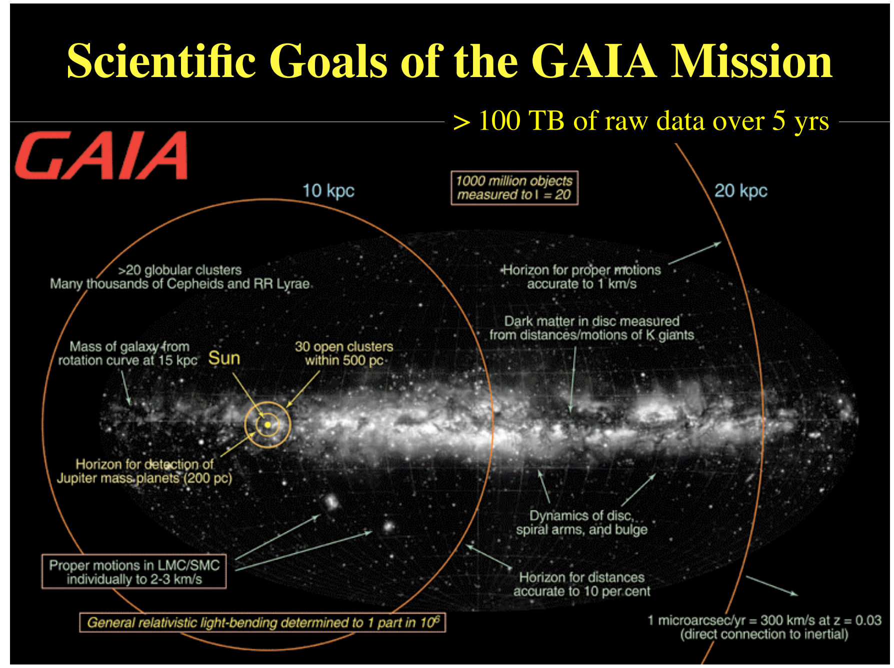
- The art of telescope building and optics improves slowly
- Detectors grow at a nearly Moore’s law rate
- Computing capabilities double on a ~ 12-18 month scale
- Software improves much more slowly

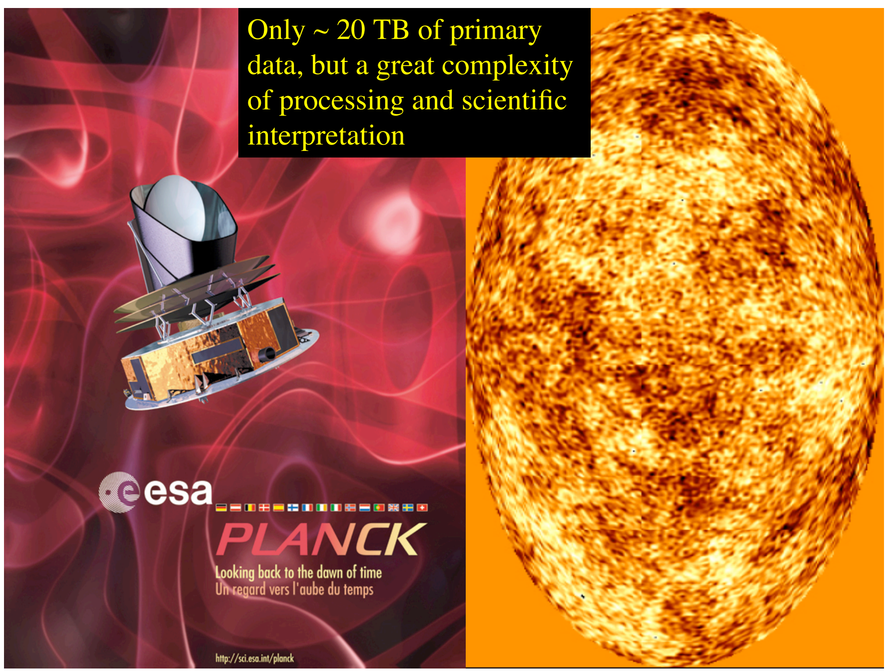
Thus:

- It is more efficient to ride on the exponentially growing technology: detectors and data processing
- Importance of computing and software increases relative to the hardware (telescopes and detectors)
- Software is the key bottleneck, and increasingly the main cost driver (at least 30%, sometimes 80%?)

Some General Comments

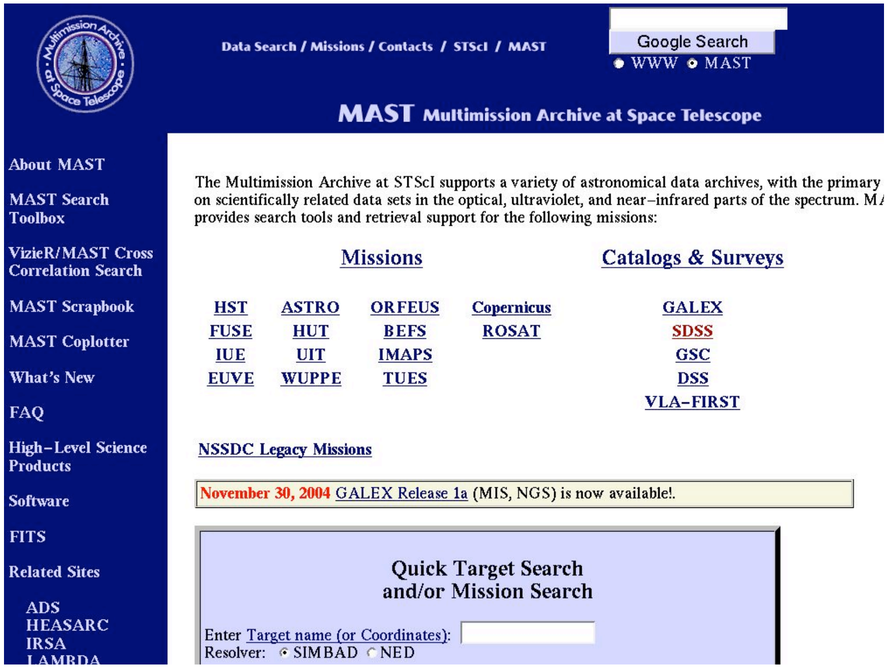
- Distinguish ***surveys*** (general data sets, could be motivated by a particular scientific goal, but have enough information content to serve many others) and ***experiments*** (optimized for a single scientific goal, e.g., find supernovae for cosmology, killer asteroids, etc.)
- Surveys are *expensive*: major projects in tens of M$ now, going to hundreds of M$ with LSST, etc. Implies large collaborations …
- Panoramic surveys are now a dominant data source in astronomy
  - Typical size ~ 10 - 100 TB, PB scale data sets are coming
  - Current archives ~ 1 PB
- There is a growing overabundance of imaging surveys; spectroscopic follow-up (where most of the physics is) is now a major bottleneck, and it will get worse
- Synoptic surveys are the new (major) trend





The Archive Archipelago

- As the data sets kept increasing, a number of archives, data depositories, and digital library services were created
- All of them are mission-, domain-, or observatory-specific, distinct and independent scientifically, technologically, institutionally, heterogeneous in look-feel, usage, etc.
  - There was a considerable replication of effort
  - There was some functional redundancy
  - There was almost no interoperability
- All of them were primarily designed for single-object (or single-pointing) queries - and thus ***inherently unsuitable for the science enabled by the massive and complex data sets***
- The next step was clearly to connect them in a functional manner, and develop interoperability standards, formats, etc.



Data Search / Missions / Contacts / STScI / MAST

Google Search
WWW MAST

MAST Multimission Archive at Space Telescope

About MAST
MAST Search Toolbox
VizieR/MAST Cross Correlation Search
MAST Scrapbook
MAST Coplotter
What's New
FAQ
High–Level Science Products
Software
FITS
Related Sites
ADS
HEASARC
IRSA
LAMBDA

The Multimission Archive at STScI supports a variety of astronomical data archives, with the primary on scientifically related data sets in the optical, ultraviolet, and near–infrared parts of the spectrum. M/ provides search tools and retrieval support for the following missions:

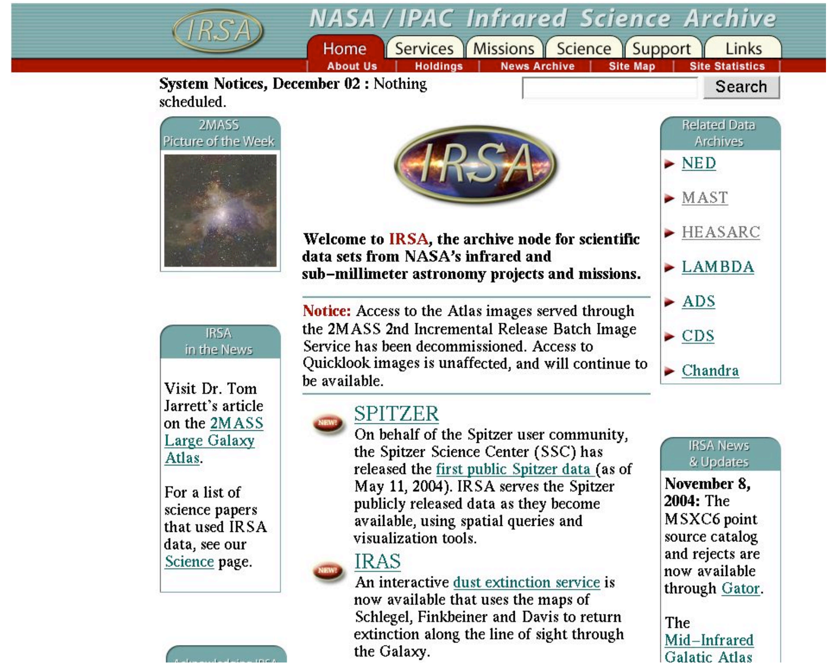
Missions				Catalogs & Surveys
HST	ASTRO	ORFEUS	Copernicus	GALEX
FUSE	HUT	BEFS	ROSAT	SDSS
IUE	UIT	IMAPS		GSC
EUVE	WUPPE	TUES		DSS
				VLA–FIRST

NSSDC Legacy Missions

November 30, 2004 GALEX Release 1a (MIS, NGS) is now available!.

Quick Target Search
and/or Mission Search

Enter Target name (or Coordinates):
Resolver: SIMBAD NED



NASA / IPAC Infrared Science Archive

Home | Services | Missions | Science | Support | Links

About Us | Holdings | News Archive | Site Map | Site Statistics

System Notices, December 02 : Nothing scheduled.

Search

2MASS Picture of the Week

Welcome to IRSA, the archive node for scientific data sets from NASA's infrared and sub–millimeter astronomy projects and missions.

Notice: Access to the Atlas images served through the 2MASS 2nd Incremental Release Batch Image Service has been decommissioned. Access to Quicklook images is unaffected, and will continue to be available.

IRSA in the News

Visit Dr. Tom Jarrett's article on the 2MASS Large Galaxy Atlas.

For a list of science papers that used IRSA data, see our Science page.

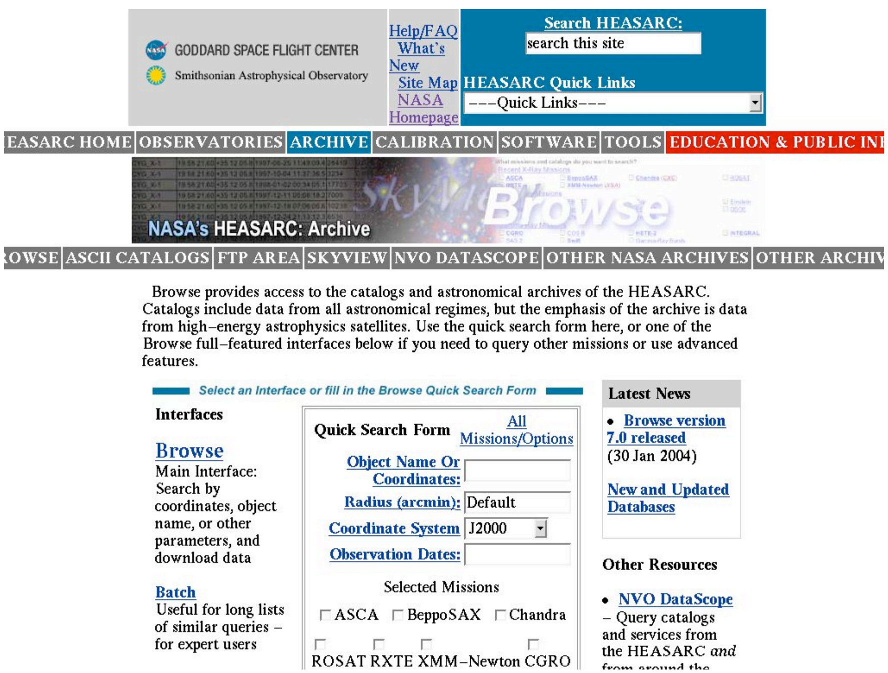
SPITZER
On behalf of the Spitzer user community, the Spitzer Science Center (SSC) has released the first public Spitzer data (as of May 11, 2004). IRSA serves the Spitzer publicly released data as they become available, using spatial queries and visualization tools.

IRAS
An interactive dust extinction service is now available that uses the maps of Schlegel, Finkbeiner and Davis to return extinction along the line of sight through the Galaxy.

Related Data Archives
NED
MAST
HEASARC
LAMBDA
ADS
CDS
Chandra

IRSA News & Updates
November 8, 2004: The MSXC6 point source catalog and rejects are now available through Gator.

The Mid–Infrared Galatic Atlas



GODDARD SPACE FLIGHT CENTER
Smithsonian Astrophysical Observatory

Help/FAQ
What's New
Site Map
NASA Homepage

Search HEASARC:
search this site
HEASARC Quick Links
---Quick Links---

EASARC HOME | OBSERVATORIES | ARCHIVE | CALIBRATION | SOFTWARE | TOOLS | EDUCATION & PUBLIC INI

NASA's HEASARC: Archive
Browse

OWSE | ASCII CATALOGS | FTP AREA | SKYVIEW | NVO DATASCOPE | OTHER NASA ARCHIVES | OTHER ARCHIV

Browse provides access to the catalogs and astronomical archives of the HEASARC. Catalogs include data from all astronomical regimes, but the emphasis of the archive is data from high–energy astrophysics satellites. Use the quick search form here, or one of the Browse full–featured interfaces below if you need to query other missions or use advanced features.

Select an Interface or fill in the Browse Quick Search Form

Interfaces

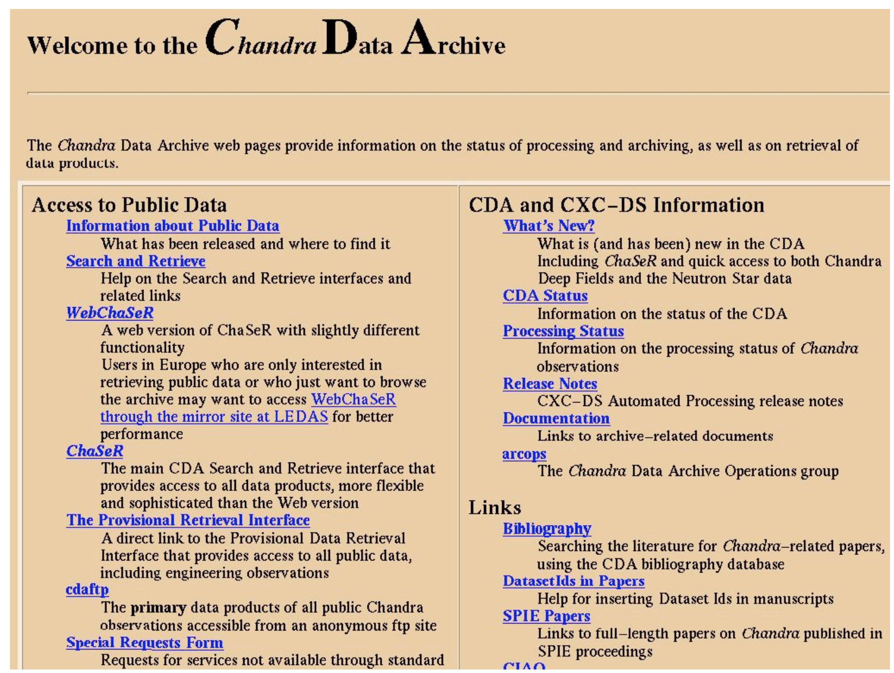
Browse
Main Interface: Search by coordinates, object name, or other parameters, and download data

Batch
Useful for long lists of similar queries – for expert users

Quick Search Form
All Missions/Options
Object Name Or Coordinates:
Radius (arcmin): Default
Coordinate System J2000
Observation Dates:
Selected Missions
ASCA BeppoSAX Chandra
ROSAT RXTE XMM–Newton CGRO

Latest News
- Browse version 7.0 released (30 Jan 2004)
- New and Updated Databases

Other Resources
- NVO DataScope – Query catalogs and services from the HEASARC and from around the



Welcome to the *Chandra* Data Archive

The *Chandra* Data Archive web pages provide information on the status of processing and archiving, as well as on retrieval of data products.

Access to Public Data

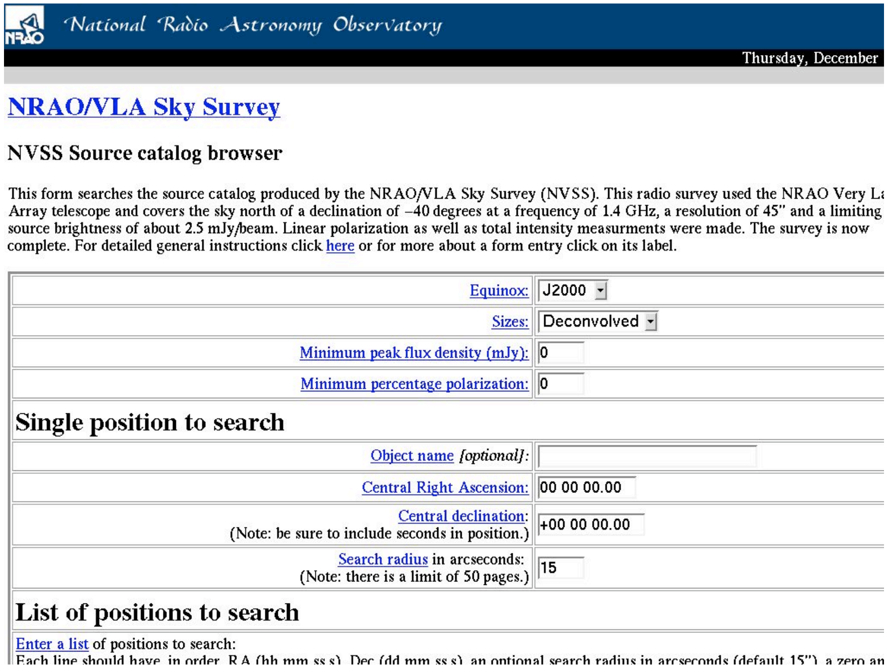
Information about Public Data
What has been released and where to find it
Search and Retrieve
Help on the Search and Retrieve interfaces and related links
WebChaSeR
A web version of ChaSeR with slightly different functionality
Users in Europe who are only interested in retrieving public data or who just want to browse the archive may want to access WebChaSeR through the mirror site at LEDAS for better performance
ChaSeR
The main CDA Search and Retrieve interface that provides access to all data products, more flexible and sophisticated than the Web version
The Provisional Retrieval Interface
A direct link to the Provisional Data Retrieval Interface that provides access to all public data, including engineering observations
cdaftp
The **primary** data products of all public Chandra observations accessible from an anonymous ftp site
Special Requests Form
Requests for services not available through standard

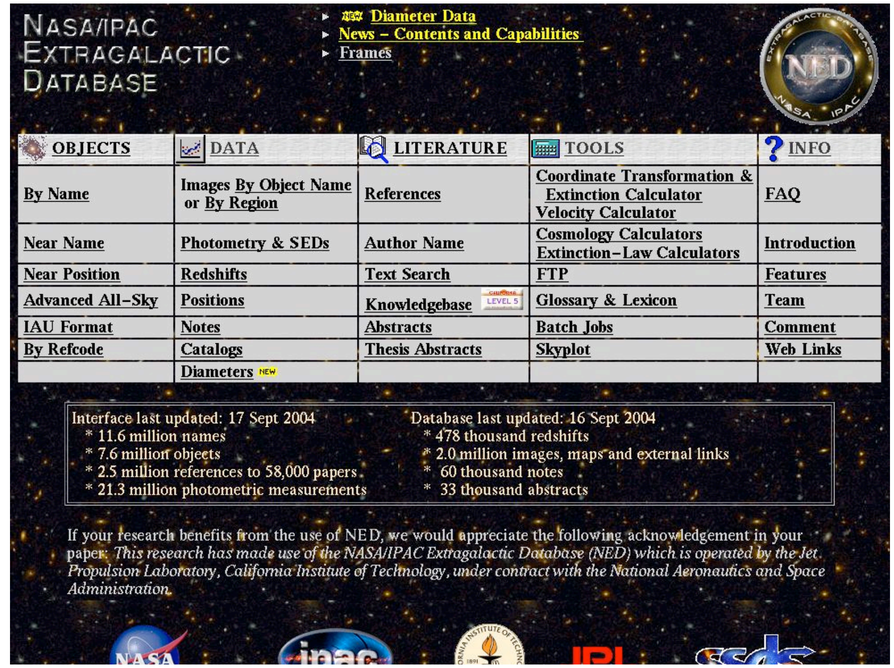
CDA and CXC–DS Information

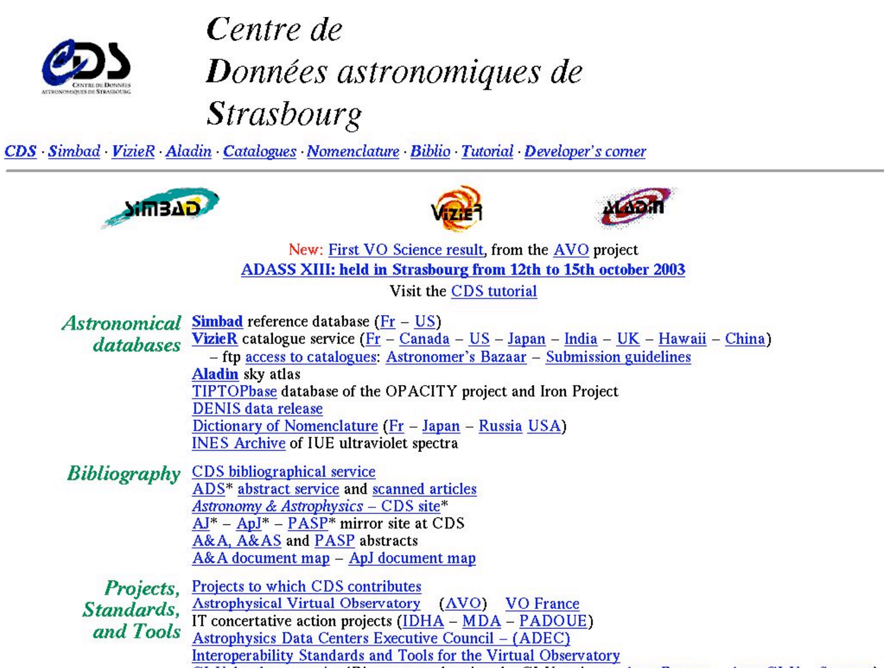
What's New?
What is (and has been) new in the CDA Including *ChaSeR* and quick access to both Chandra Deep Fields and the Neutron Star data
CDA Status
Information on the status of the CDA
Processing Status
Information on the processing status of *Chandra* observations
Release Notes
CXC–DS Automated Processing release notes
Documentation
Links to archive–related documents
arcops
The *Chandra* Data Archive Operations group

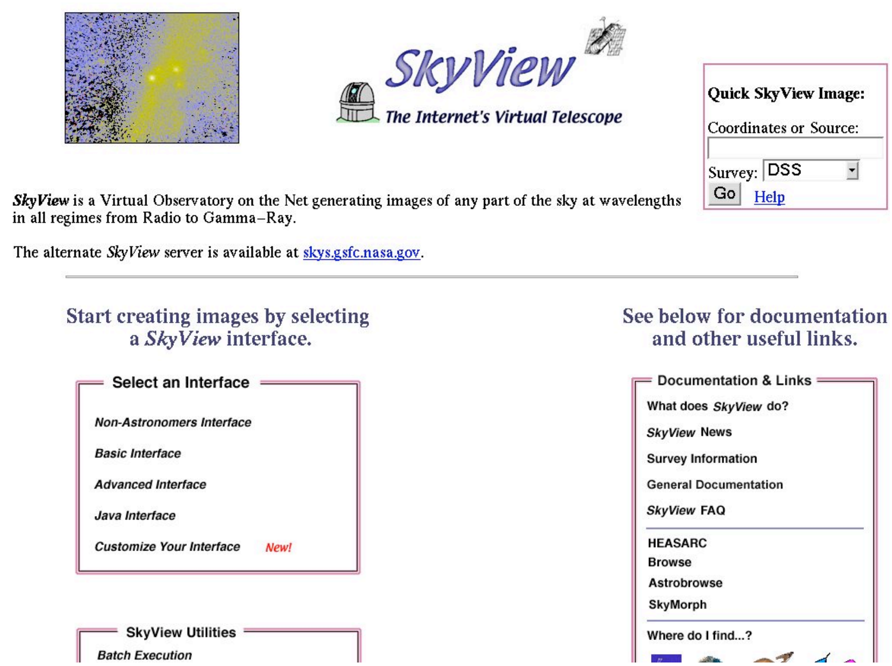
Links

Bibliography
Searching the literature for *Chandra*–related papers, using the CDA bibliography database
DatasetIds in Papers
Help for inserting Dataset Ids in manuscripts
SPIE Papers
Links to full–length papers on *Chandra* published in SPIE proceedings
CIAO


National Radio Astronomy Observatory
Thursday, December
NRAO/VLA Sky Survey
NVSS Source catalog browser
This form searches the source catalog produced by the NRAO/VLA Sky Survey (NVSS). This radio survey used the NRAO Very La
Array telescope and covers the sky north of a declination of −40 degrees at a frequency of 1.4 GHz, a resolution of 45" and a limiting
source brightness of about 2.5 mJy/beam. Linear polarization as well as total intensity measurments were made. The survey is now
complete. For detailed general instructions click here or for more about a form entry click on its label.
Equinox: J2000
Sizes: Deconvolved
Minimum peak flux density (mJy): 0
Minimum percentage polarization: 0
Single position to search
Object name [optional]:
Central Right Ascension: 00 00 00.00
Central declination: +00 00 00.00
(Note: be sure to include seconds in position.)
Search radius in arcseconds: 15
(Note: there is a limit of 50 pages.)
List of positions to search
Enter a list of positions to search:

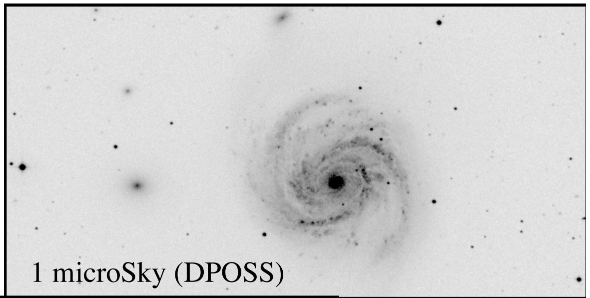

NASA/IPAC EXTRAGALACTIC DATABASE
Diameter Data
News – Contents and Capabilities
Frames
OBJECTS
DATA
LITERATURE
TOOLS
INFO
By Name
Images By Object Name or By Region
References
Coordinate Transformation & Extinction Calculator
Velocity Calculator
FAQ
Near Name
Photometry & SEDs
Author Name
Cosmology Calculators
Extinction–Law Calculators
Introduction
Near Position
Redshifts
Text Search
FTP
Features
Advanced All–Sky
Positions
Knowledgebase
Glossary & Lexicon
Team
IAU Format
Notes
Abstracts
Batch Jobs
Comment
By Refcode
Catalogs
Thesis Abstracts
Skyplot
Web Links
Diameters NEW
Interface last updated: 17 Sept 2004
* 11.6 million names
* 7.6 million objects
* 2.5 million references to 58,000 papers
* 21.3 million photometric measurements
Database last updated: 16 Sept 2004
* 478 thousand redshifts
* 2.0 million images, maps and external links
* 60 thousand notes
* 33 thousand abstracts
If your research benefits from the use of NED, we would appreciate the following acknowledgement in your paper: This research has made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration.


Centre de
Données astronomiques de
Strasbourg
CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · Tutorial · Developer's corner
New: First VO Science result, from the AVO project
ADASS XIII: held in Strasbourg from 12th to 15th october 2003
Visit the CDS tutorial
Astronomical databases
Simbad reference database (Fr – US)
VizieR catalogue service (Fr – Canada – US – Japan – India – UK – Hawaii – China)
– ftp access to catalogues: Astronomer's Bazaar – Submission guidelines
Aladin sky atlas
TIPTOPbase database of the OPACITY project and Iron Project
DENIS data release
Dictionary of Nomenclature (Fr – Japan – Russia USA)
INES Archive of IUE ultraviolet spectra
Bibliography
CDS bibliographical service
ADS* abstract service and scanned articles
Astronomy & Astrophysics – CDS site*
AJ* – ApJ* – PASP* mirror site at CDS
A&A, A&AS and PASP abstracts
A&A document map – ApJ document map
Projects, Standards, and Tools
Projects to which CDS contributes
Astrophysical Virtual Observatory (AVO) VO France
IT concertative action projects (IDHA – MDA – PADOUE)
Astrophysics Data Centers Executive Council – (ADEC)
Interoperability Standards and Tools for the Virtual Observatory

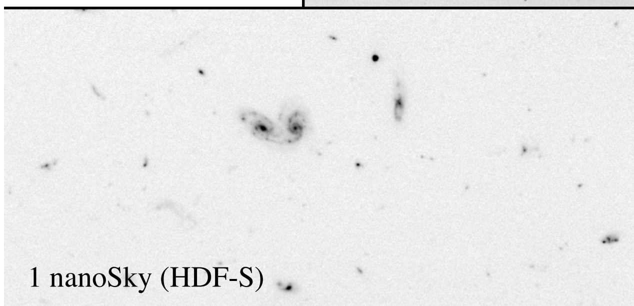

SkyView
The Internet's Virtual Telescope
Quick SkyView Image:
Coordinates or Source:
Survey: DSS
Go Help
SkyView is a Virtual Observatory on the Net generating images of any part of the sky at wavelengths in all regimes from Radio to Gamma–Ray.
The alternate SkyView server is available at skys.gsfc.nasa.gov.
Start creating images by selecting a SkyView interface.
See below for documentation and other useful links.
Select an Interface
Non-Astronomers Interface
Basic Interface
Advanced Interface
Java Interface
Customize Your Interface New!
Documentation & Links
What does SkyView do?
SkyView News
Survey Information
General Documentation
SkyView FAQ
HEASARC
Browse
Astrobrowse
SkyMorph
Where do I find...?
SkyView Utilities
Batch Execution

From Data-Poor to Data-Rich Astronomy

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

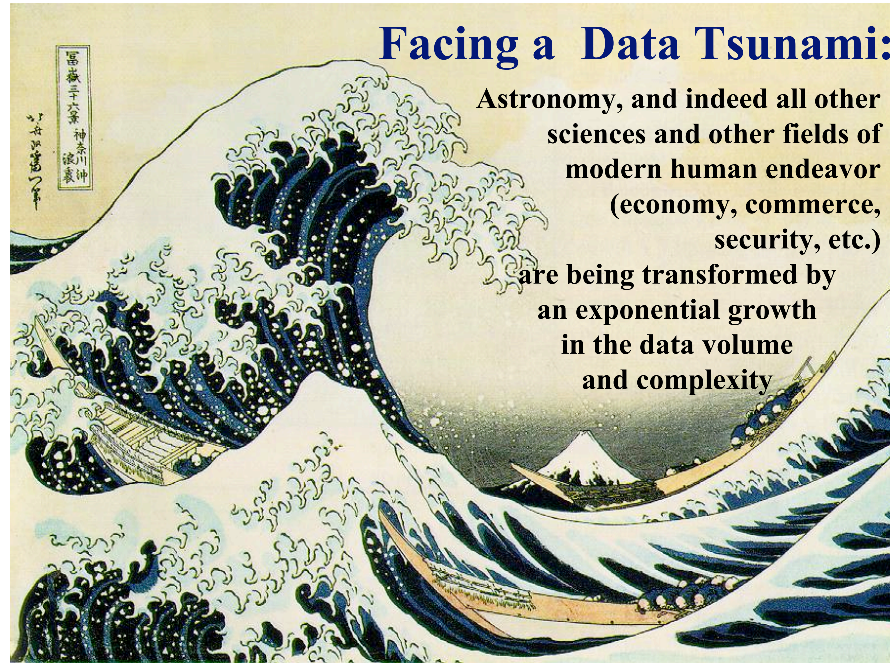


1 microSky (DPOSS)



1 nanoSky (HDF-S)

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



Facing a Data Tsunami:

Astronomy, and indeed all other sciences and other fields of modern human endeavor (economy, commerce, security, etc.) are being transformed by an exponential growth in the data volume and complexity

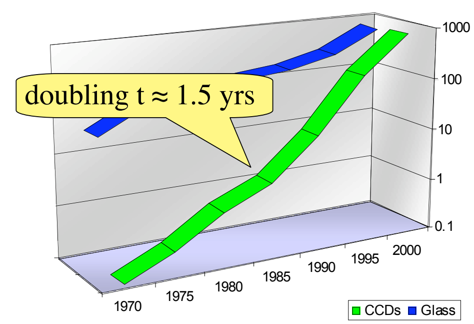
So, How Much Data Is This?

- Typical sky survey now generates ~ 10 - 50 TB
- PB-scale data sets are coming (e.g., LSST, PanSTARRS…)
- There is now ~ 1 - 2 PB of archived data in all of astronomy
- New data generation rate is a few TB / day
- For comparison:
  - Human memory ~ a few hundred MB
  - Human Genome < 1 GB
  - 1 TB ~ 2 million books
  - Library of Congress (print only) ~ 30 TB
- But the real issue is not so much data volume (or rate), as data *complexity*

- **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**

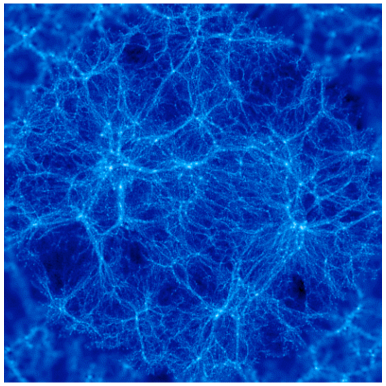
Data ➔ Knowledge ?

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

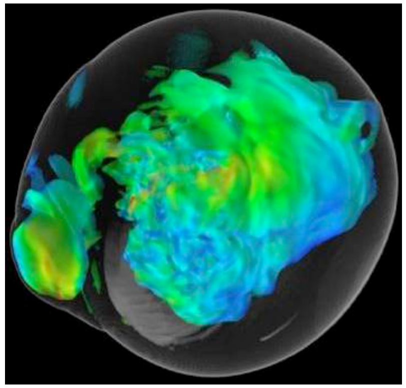


… but our understanding of the universe increases much more slowly!

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



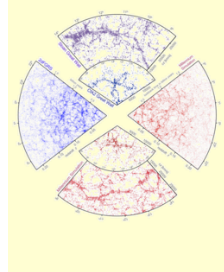
Structure formation in the Universe



Supernova explosions

Numerical simulations are not just a weak substitute for the analytical theory - they are an inevitable methodology to study theoretically many complex phenomena, e.g., star or galaxy formation, etc.

Simulation Data via VO Access



Halo and Galaxy Formation Histories from the Millennium Simulation

Public release of a VO-oriented and SQL-queryable database for studying the evolution of galaxies in the ΛCDM cosmogony

Gerard Lemson & the Virgo Consortium

astro-ph/0608019
full description of release (PDF)
database mirror site at ICC, Durham University

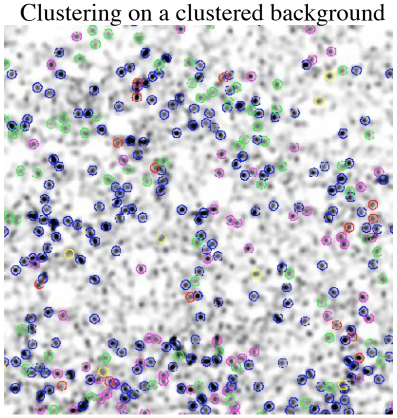
- Database Access
- Visual Material
- Related Links
- Publications

http://www.mpa-garching.mpg.de/millennium

The Millennium Run is the largest simulation of the formation of structure within the ΛCDM cosmogony so far carried out. It uses 10^{10} particles to follow the dark matter distribution in a cubic region $500h^{-1}$Mpc on a side, and has a spatial resolution of $5h^{-1}$kpc. Application of simplified modelling techniques to the stored output of this calculation allows the formation and evolution of the ~10^7 galaxies more luminous than the Small Magellanic Cloud to be simulated for a variety of assumptions about the detailed physics involved. As part of the activities of the German Astrophysical Virtual Observatory we have created relational databases to store the detailed assembly histories both of all the haloes and subhaloes resolved by the simulation, and of all the galaxies that form within these structures for two independent models of the galaxy formation physics. We have implemented a Structured Query Language (SQL) server on these databases.

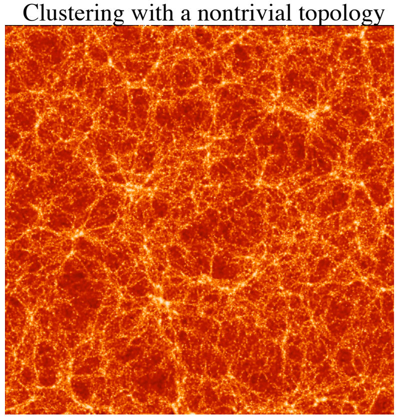
Comparing the output of numerical simulations to equally massive and complex observational data sets is a very non-trivial problem …

Clustering on a clustered background



DPOSS Clusters (Gal et al.)

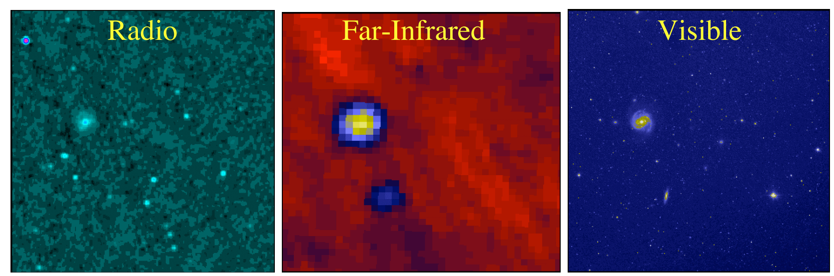
Clustering with a nontrivial topology

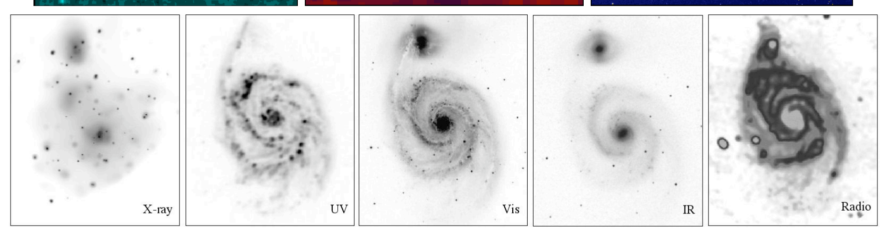


LSS Numerical Simulation (VIRGO)

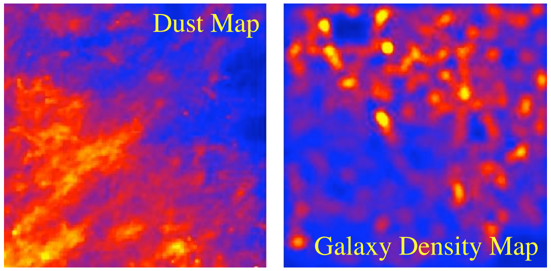
The Universe Looks Different at Different Wavelengths

(since the emission is dominated by different physical processes)

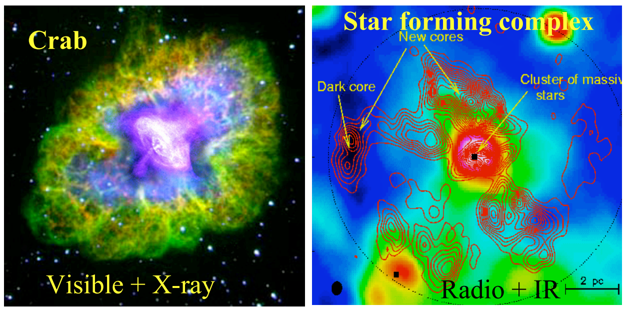




Panchromatic Views of the Universe



Multi-Wavelength Data Fusion leads to a more complete, less biased picture

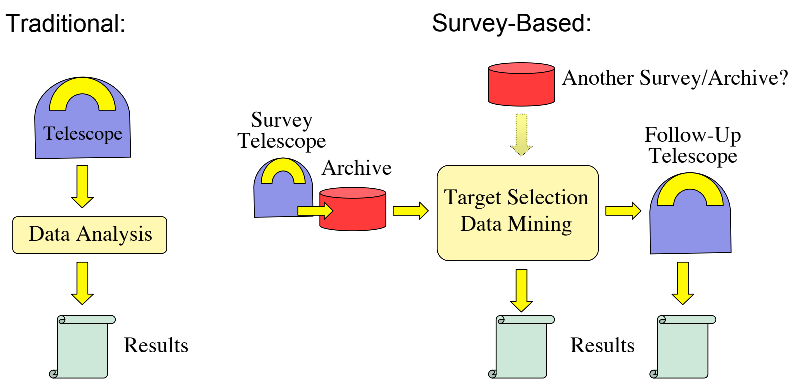


Understanding of complex phenomena requires complex data sets!

The Changing Style of Observational Astronomy

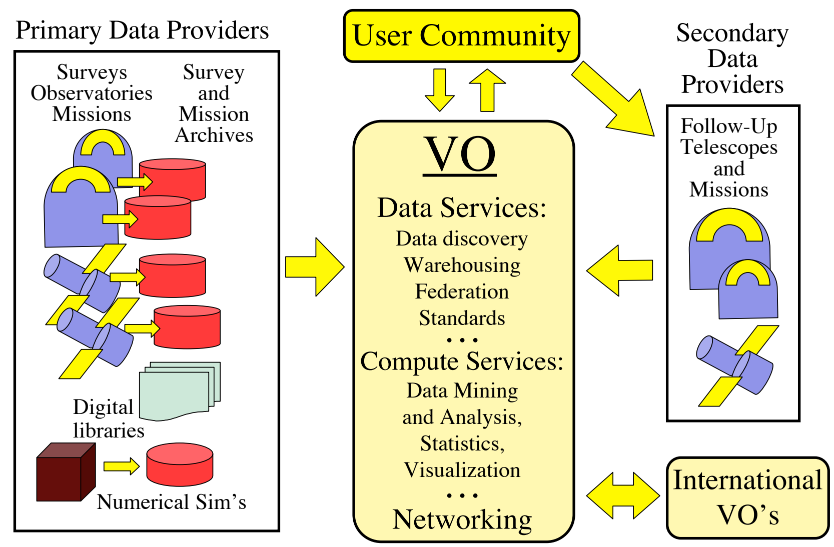
The Old Way:	**Now:**	**Future:**
Pointed, heterogeneous observations (~ MB - GB)	Large, homogeneous sky surveys (multi-TB, ~ 10^6 - 10^9 sources)	Multiple, federated sky surveys and archives (muti-PB)
Small samples of objects (~ 10^1 - 10^3)	Archives of pointed observations (~ few TB)	**Virtual Observatory**

From Traditional to Survey to VO-Based Science



Highly successful and increasingly prominent, but inherently limited by the information content of individual surveys …
<u>*What comes next, beyond survey science is the **VO science***</u>

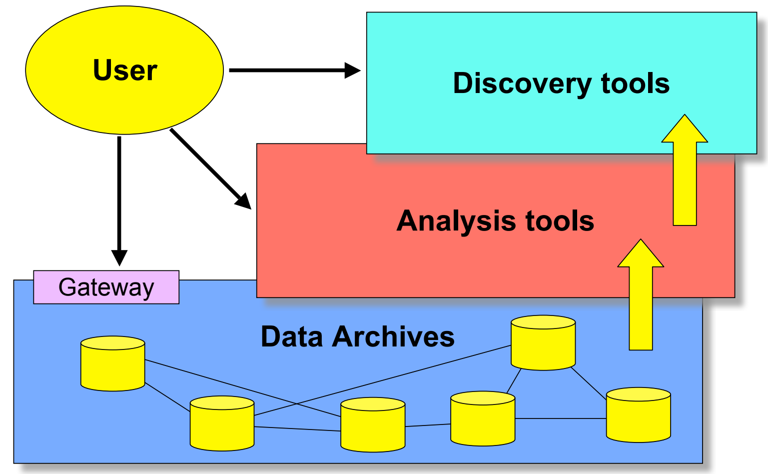
A Systemic View of the VO



The Virtual Observatory Concept

- Astronomy community's response to the scientific and technological challenges posed by the exponential growth of data sets
  - *Technology-enabled, but science-driven:* harness the IT advances in service of astronomy
- A complete, dynamical, distributed, web-based, open ***research environment for astronomy with massive and complex data sets***
  - Provide content (data, metadata) services, standards, and analysis/compute services
  - Federate the existing and forthcoming digital sky surveys and archives, facilitate data inclusion and distribution
  - Develop and provide data exploration and discovery tools
- A new type of a scientific organization

VO: Conceptual Architecture



Information-Rich Astronomy in the 21st Century

- Technological revolutions as the drivers/enablers of the bursts of scientific growth
  - Detectors, computers + WWW, now data technologies
- 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**

<u>The next golden age of discovery in astronomy?</u>

VO is the mechanism to effect this process

This quantitative change in the information volume and complexity will enable the

<u>Science of a Qualitatively Different Nature:</u>

- **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**
  (NB: Energy content ≠ Information content)
  - 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**

+ things we have not thought of yet …

Scientific Roles and Benefits of a VO

- **Facilitate science with massive data sets** (observations and theory/simulations) ⇨ **efficiency amplifier**
- Provide an **added value** 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 *qualitatively new* science** 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 …)

Broader 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

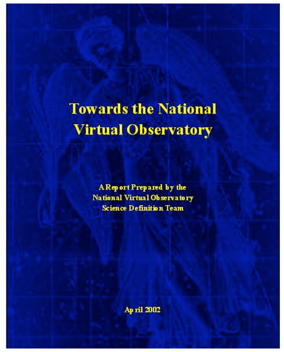
"Weapons of Mass Instruction"

Why is VO a Different Thing

- 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

A Brief History of the VO Concept

- Early (pre-web!) ideas already in the "Astrophyisics Data System" (only the digital library part survives)
- Concept developed through 1990's, mainly from large digital sky surveys (DPOSS, SDSS…), discussions at conferences and workshops in the late 1990's
- Top recommendation in the "small projects" category in the NAS Decadal Astronomy & Astrophysics survey (the McKee-Taylor report), 2001
- The first major VO conference at Caltech in 2000; the NVO White paper
- National Virtual Observatory Science Definition Team, 2001 - 2002
- NSF-sponsored project 2002 - present
- Vigorous international efforts, coordinated via Int'l VO Alliance (IVOA)



arXiv.org > astro-ph > arXiv:astro-ph/0108115

Search or Article-

Astrophysics

Toward a National Virtual Observatory: Science Goals, Technical Challanges, and Implementation Plan

NVO Interim Steering Committee

(Submitted on 7 Aug 2001)

The National Academy of Science Astronomy and Astrophysics Survey Committee, in its new Decadal survey entitled Astronomy and Astrophysics in the New Millennium, recommends, as a first priority, the establishment of a National Virtual Observatory. The NVO would link the archival data sets of space- and ground-based observatories, the catalogs of multi-wavelength surveys, and the computational resources necessary to support comparison and cross-correlation among these resources. This White Paper describes the scientific opportunities and technical challenges of an NVO, and lays out an implementation strategy aimed at realizing the goals of the NVO in cost-effective manner. The NVO will depend on inter-agency cooperation, distributed development, and distributed operations. It will challenge the astronomical community, yet provide new opportunities for scientific discovery that were unimaginable just a few years ago.

Comments: 20 pages, published in the proceedings of the conference, Virtual Observatories of the Future, eds. Robert J. Brunner, S. George Djorgovski, and Alex S. Szalay, ASP Volume 225, p 353

http://us-vo.org

International VO Alliance (IVOA)

http:// ivoa.net

And now BRAVO!

VO Status and Prospects

- Good progress on data grid infrastructure:
  - DB design and implementation
  - Formats, standards, protocols, …
  - Interoperability, etc.
- Good communications and exchanges world-wide (IVOA)
- Community buy-in is slow
- Where we need to go next?
  - Data mining, exploration, analysis tools
  - Scalability, usability
  - That is where the discoveries will come from!
- Fit within a broader e-Science / Cyber-Infrastructure

Some Readings:

- A quick summary:
  - "Virtual Observatory: From Concept to Implementation", Djorgovski, S.G., & Williams, R. 2005, A.S.P. Conf. Ser. **345**, 517, available as ***http://arXiv.org/abs/astro-ph/0504006***
- The original VO White Paper:
  - "Toward a National Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan", in Virtual Observatories of the Future, A.S.P. Conf. Ser. **225**, 353, available as ***http://arXiv.org/abs/astro-ph/0108115***
- The NVO SDT report, available at ***http://www.us-vo.org/sdt***
- Many other good documents available at ***http://us-vo.org*** (especially the *summer school* presentations)
- Technical documents at ***http://www.ivoa.net***

Practical Examples

- Explore the services linked at ***http://us-vo.org***

(A good exercise to do in the afternoons)

- Look through the lectures and student project presentations linked at ***http://us-vo.org/summer-school/index.cfm***

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***
- 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***, which is 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?***