Databases in Astronomy

Program for today:

Astronomical Data Catalogs – properties and how to use them

An in-depth look at the Sloan Digital Sky Survey SDSS

A briefer look at the Millenium Database

Some more on SQL

Doing Science with databases – some examples

Databases with Astronomical Data

- Collections of data, ranging from small tables to huge databases

- Have always been there
  (example: Book of Fixed Stars by Abd al-Rahman al-Sufi (AD 964))

- Backbone of astronomical research

- Most of them to be found in the VIZIER database
  (http://vizier.u-strasbg.fr/)

List of most important catalogs by David Montes at UCM:

http://www.ucm.es/info/Astrof/servicios_a/catalogos.html

Categories:

- Data Centers (tomorrow)
- Photometric Surveys
- Catalogs of Stars (except spectra)
- Observatories (tomorrow)
- Fundamental atomic data
- Standard stars
- Stellar libraries

Properties of catalogs:

1 Deliverables
2 Selection criteria (wavelength, depth,completeness)
3 Spectral Coverage and Resolution
4 Spatial Coverage and Resolution
5 Quality

- Availability of material (raw spectra, etc.)

Know your catalog parameters,before using it!!!

Example: Catalog of Markarian Galaxies (1967-1981)

1. Markarian Identification number; S(eyfert) or Q(uasar) flag; cross identifications to other catalogs; B1950 positions; major and minor axis (arcsec); magnitude and spectral types, including a code for the strength of the UV continuum.
2. Selection: moderate to strong ultraviolet continuum as detected by an objective-prism survey
   Depth: no galaxies brighter than mag 13.0 or fainter than mag 17.5 (which Band?)

3 Broadband magnitude; unclear in which band

4. Unclear
5. Unclear

Some widely used catalogs

Photometric Surveys

- Digitized Sky Survey
- Automatic Plate Measuring (APM) catalogue
- SDSS, Sloan Digital Sky Survey - (SDSS-III)
- 2MASS (2 Micron All Sky Survey)
- 2dF Galaxy Redshift Survey
- UKIRT, Infrared Deep Sky Survey

Stellar Catalogs

- Bright Star Catalogue, 5th Revised Ed.
- FK6, Sixth Catalogue of Fundamental Stars
- GCVS, General Catalogue of Variable Stars 4th edition
- USNO-B1.0 Catalogue
- PPM-Extended (PPMX) - a catalogue of positions and proper motions
- The General Catalogue of Photometric Data (GCPD)
- The HIPPARCOS and TYCHO Catalogues

Galaxy Catalogs

- RC-3 (3rd Ref. Cat. of Bright Gal.)
- Uppsala General Catalog of Galaxies
- ESO/Uppsala Survey B
- Frei et al Galaxy Catalog
- Huchra et al CfA Redshift "Zcat"
- Huchra et al CfA "Zbig"

Fundamental Atomic Data

- Allen, astrophysical quantities (1973, 1983)
- Atomic data for astrophysics (Dima Verner)
- Physical Constants at http://physics.nist.gov/cuu/Constants/
- Digital Library of Mathematical Functions

Standard Stars

- SOFA, Standard Objects For Astronomy
- HST Spectrophotometric Standards
- Geneva Radial-Velocity Standard Stars
- ESO Photometric Standards

Libraries of Stellar Spectra

- Spectrophotometric atlas of synthesis standard spectra, PICKLES, A. J.
- The ELODIE archive
- A library of 0.5 to 2.5 μm spectra of luminous cool stars (Lancon & Wood)
- UVES-POP (Paranal Observatory Project)
- MILES: A Medium resolution INT Library of Empirical Spectra
- Spectrum Services for the Virtual Observatory (see later in the course)

More on Stellar Spectra

- Links to many catalogs about Stellar Classification
- High-Resolution Solar Spectrum
- SpectroWeb
- New Grids of ATLAS9 Model Atmospheres (Kurucz)
- BaSeL (Basel Stellar Library)
- A grid of MARCS model atmospheres for late-type stars
- PASTEL, PAramètres STELlaires
- Atomic Spectroscopic Data (NIST)

Program for today:

Astronomical Data Catalogs – properties and how to use them

An in-depth look at the Sloan Digital Sky Survey SDSS

More on Data Centers

A briefer look at the Millenium Database

Some more on SQL

Doing Science with databases – some examples

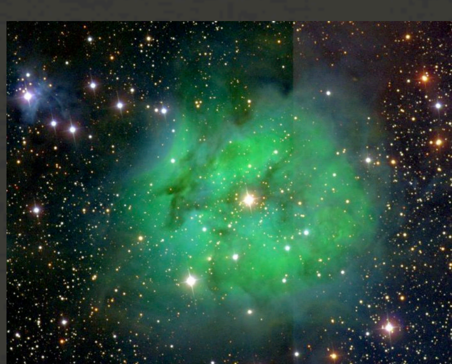
The SDSS

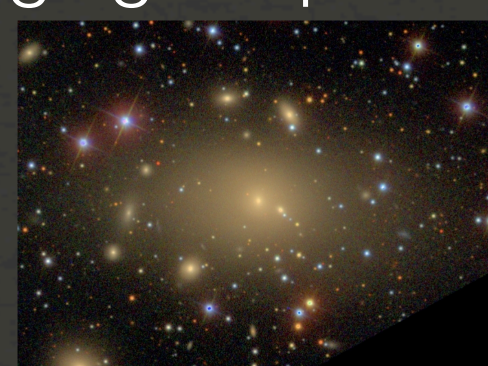
The most ambitious survey of the sky ever undertaken.

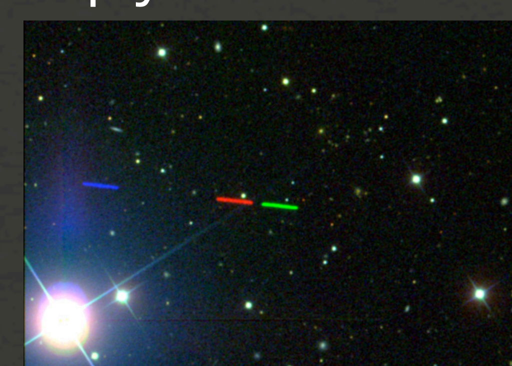
Imaging survey of 8600 square degrees.

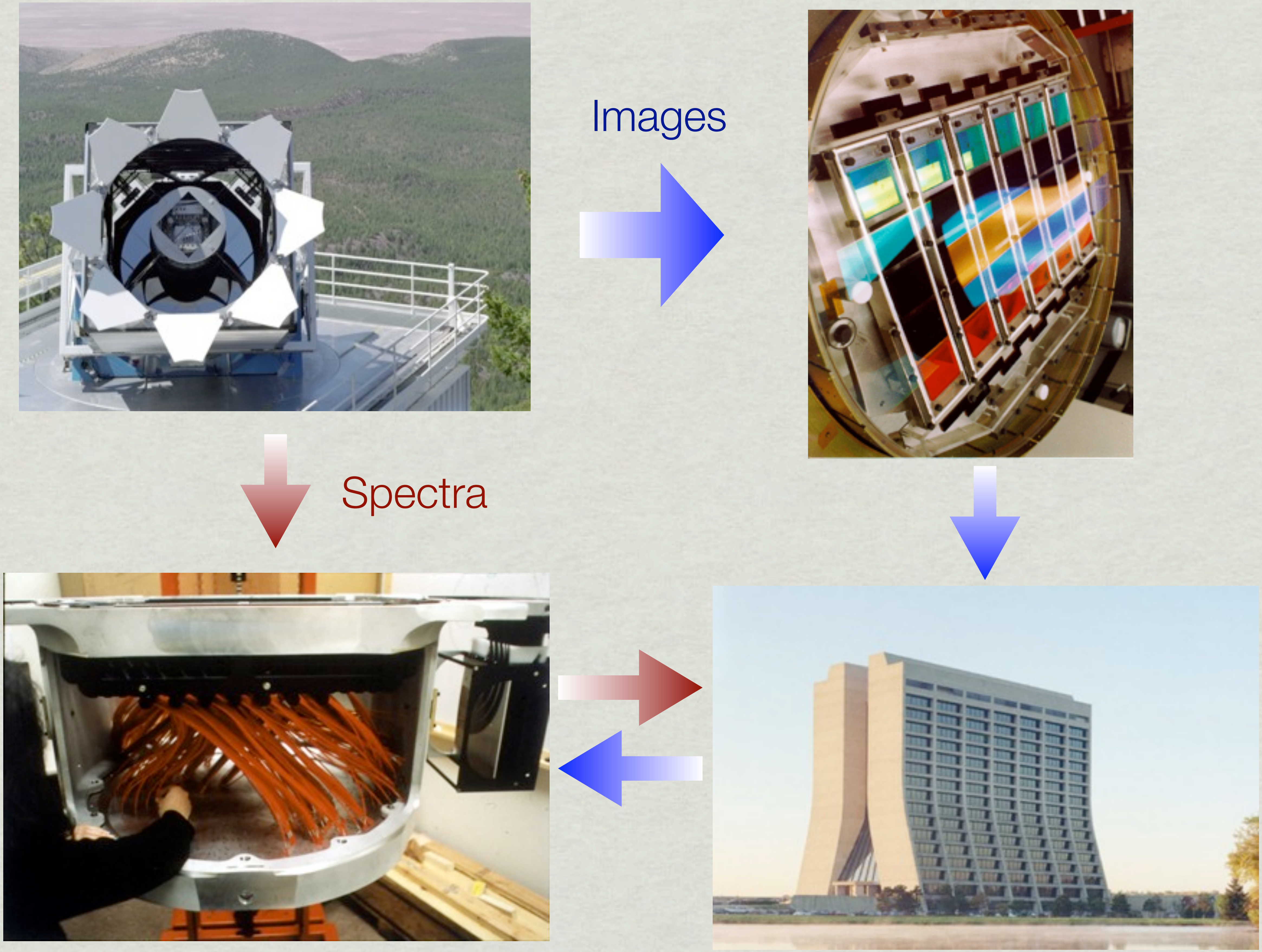
Redshifts of more than 1,000,000 galaxies & QSOs.

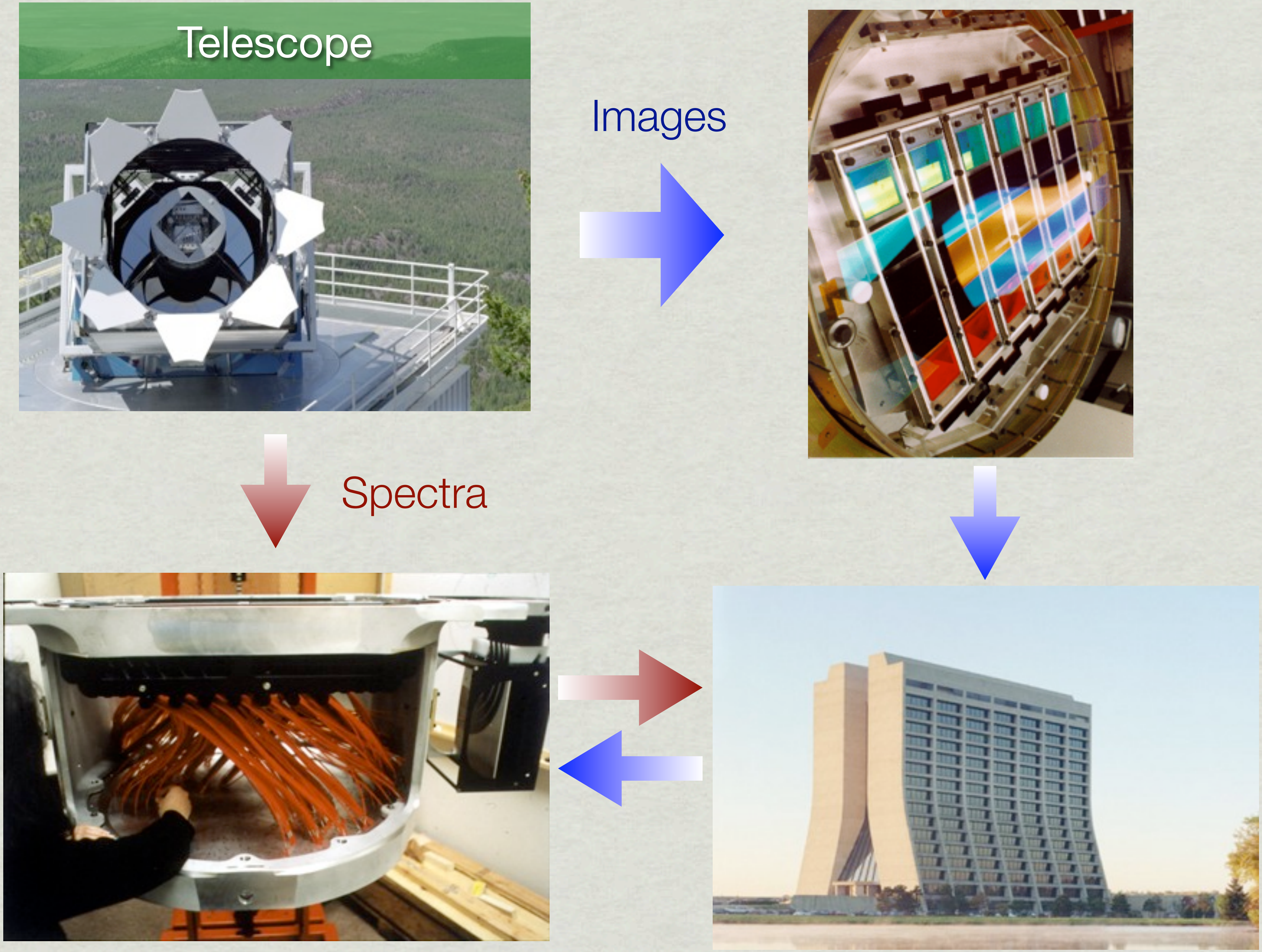
Robotic 2.5m telescope - imaging & Spectrscopy








Images
Spectra

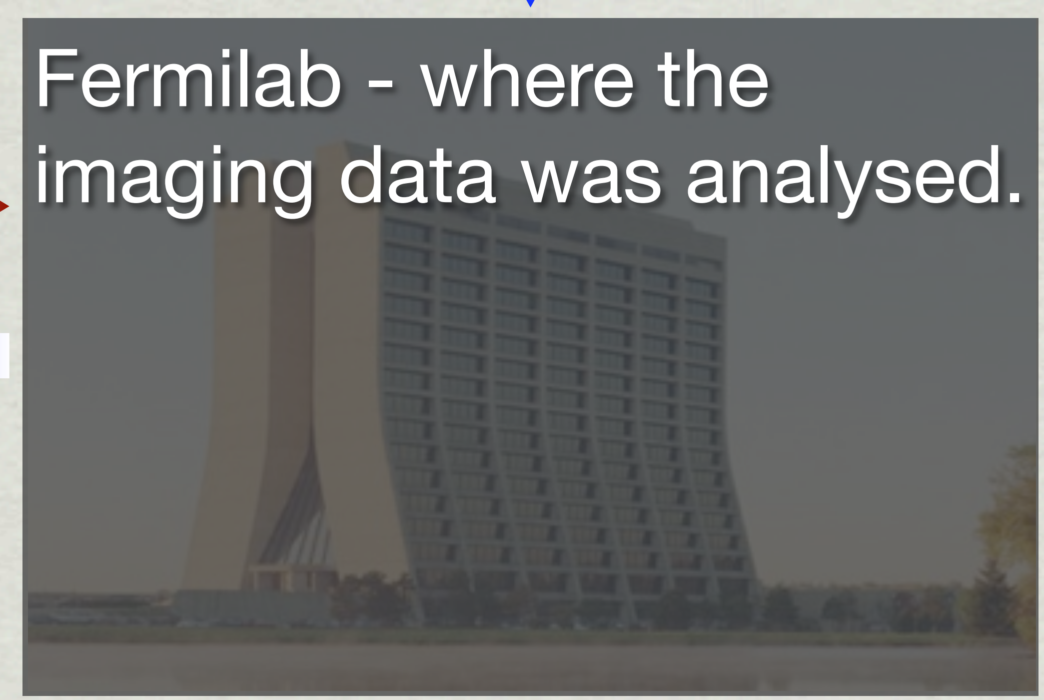

Telescope
Images
Spectra

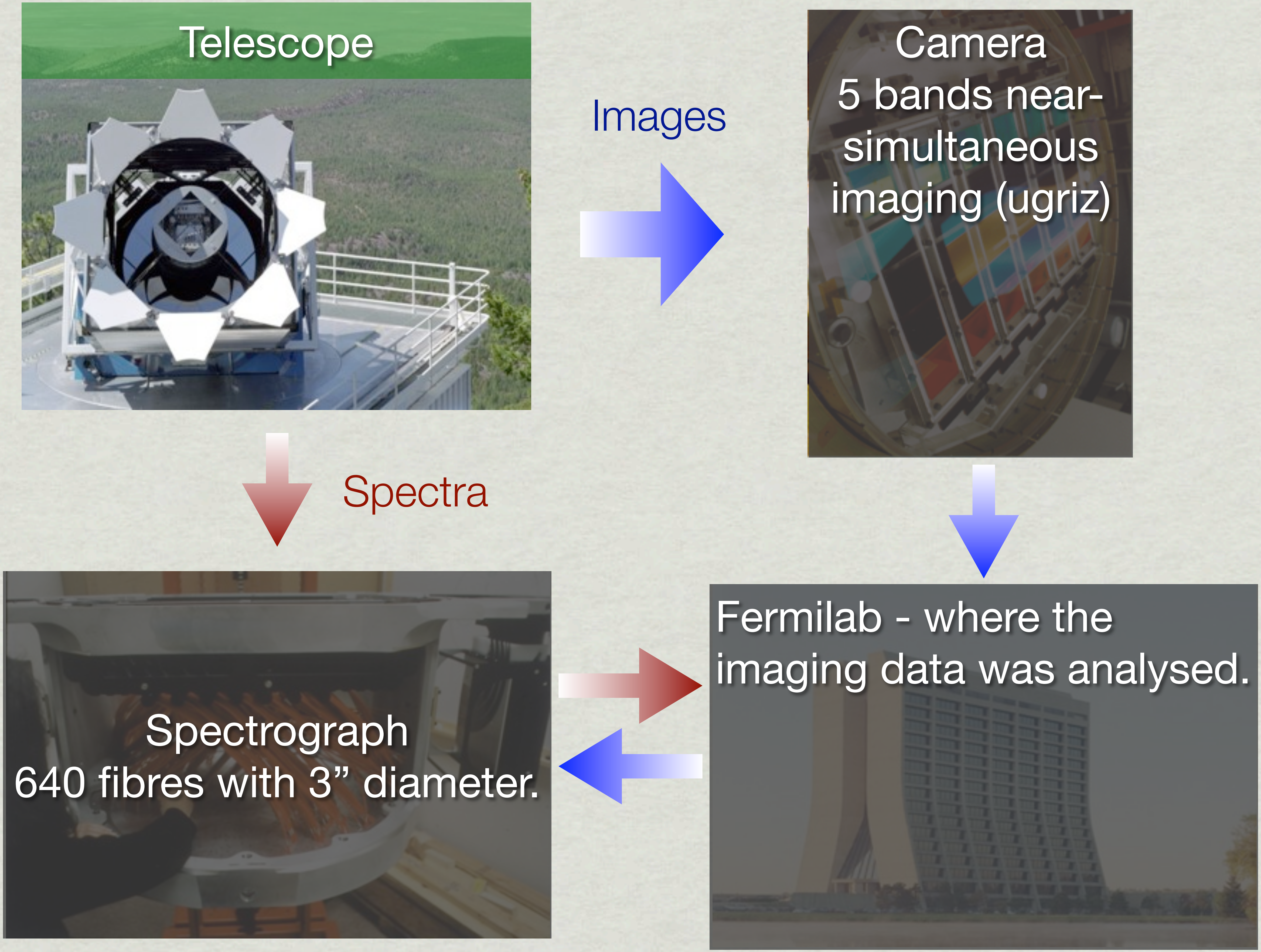
Images

Spectra

Images

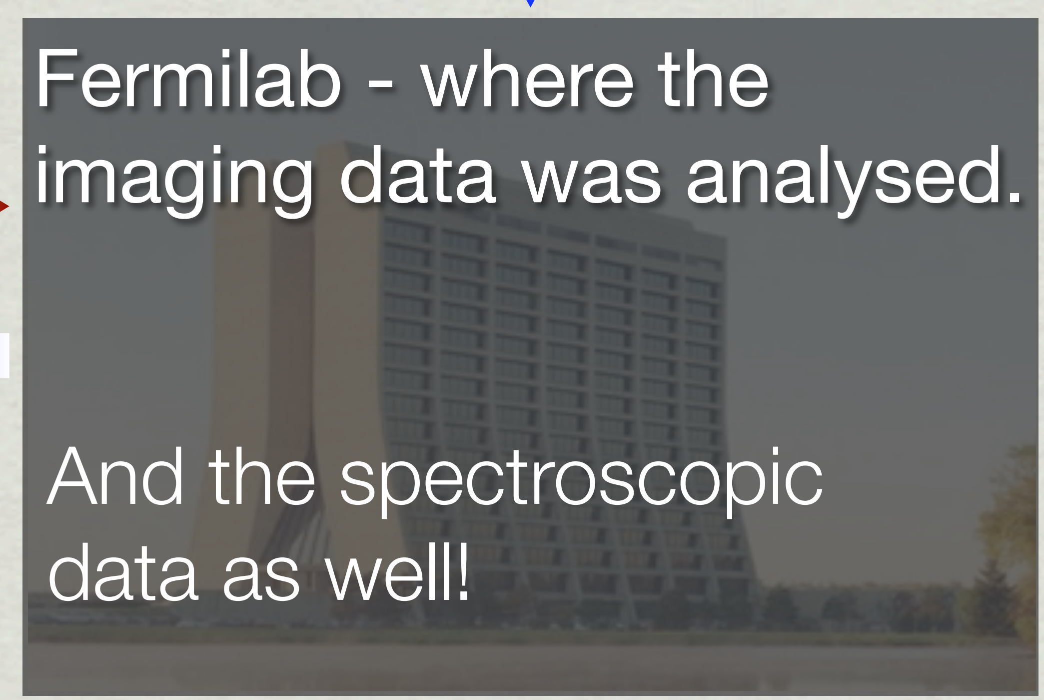
Spectra




Telescope
Images
Camera
5 bands near-simultaneous imaging (ugriz)
Spectra
Spectrograph
640 fibres with 3” diameter.
Fermilab - where the imaging data was analysed.

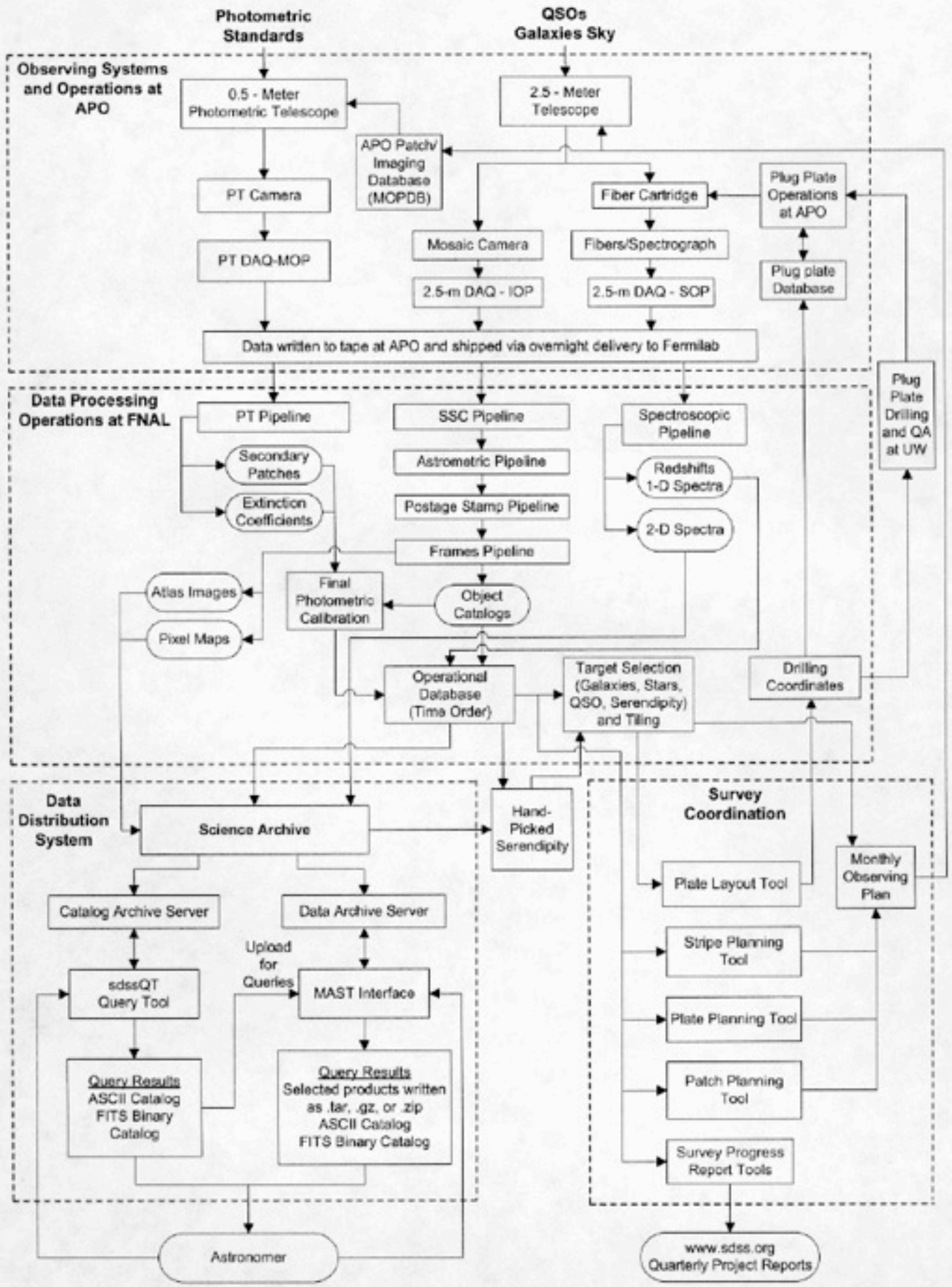
Images

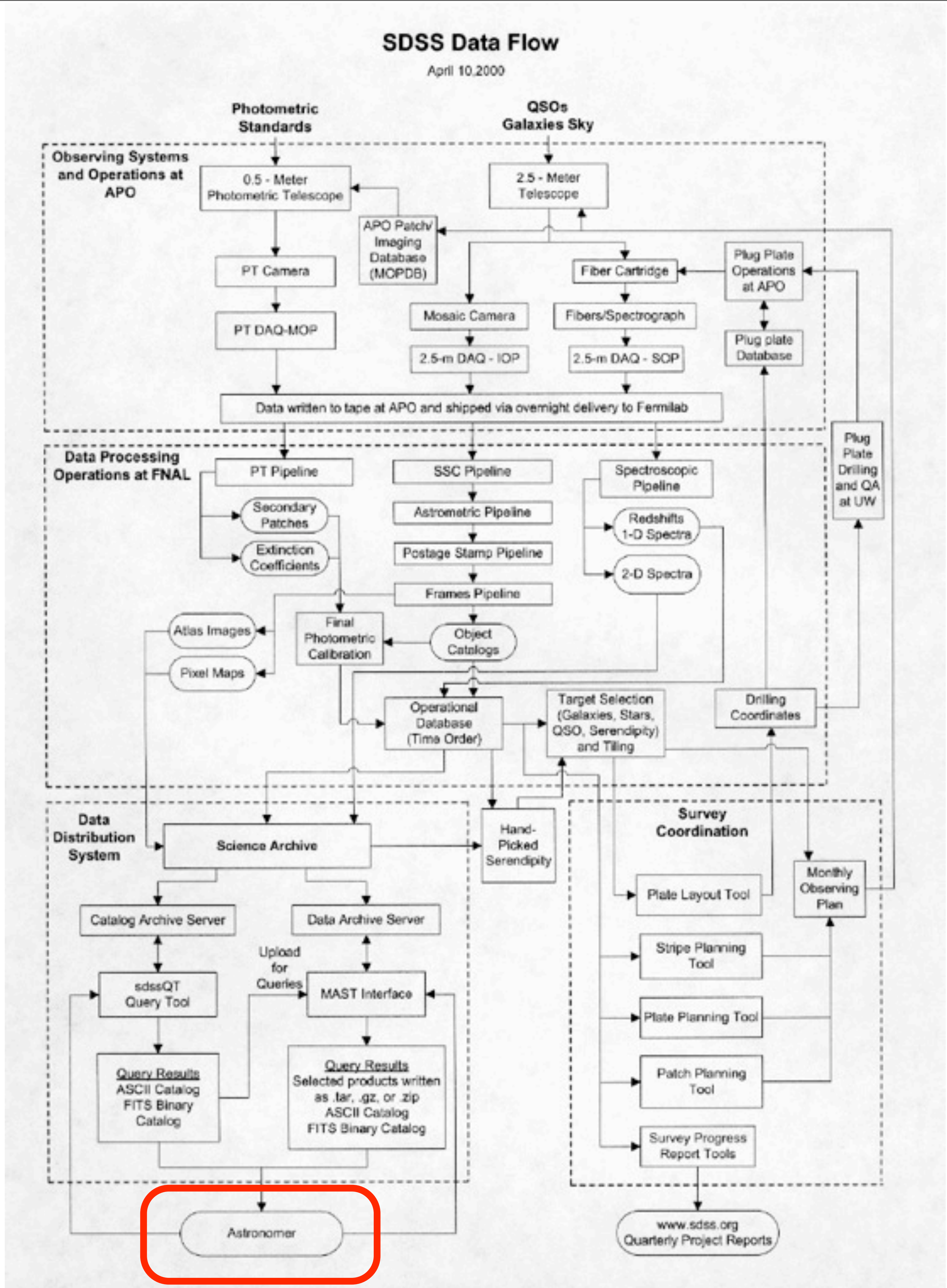
Spectra



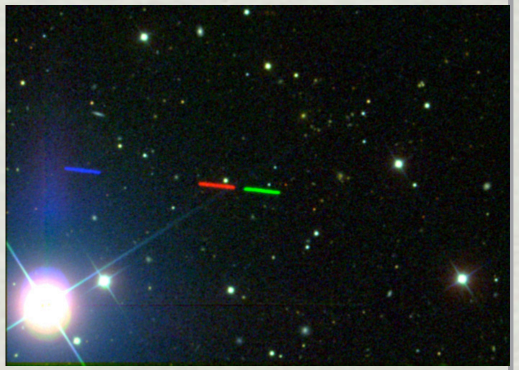
SDSS Data Flow

April 10,2000




SDSS Data Flow
April 10,2000
Photometric Standards
QSOs Galaxies Sky
Observing Systems and Operations at APO
0.5 - Meter Photometric Telescope
2.5 - Meter Telescope
APO Patch/ Imaging Database (MOPDB)
PT Camera
PT DAQ-MOP
Mosaic Camera
2.5-m DAQ - IOP
Fiber Cartridge
Fibers/Spectrograph
2.5-m DAQ - SOP
Plug Plate Operations at APO
Plug plate Database
Data written to tape at APO and shipped via overnight delivery to Fermilab
Data Processing Operations at FNAL
PT Pipeline
Secondary Patches
Extinction Coefficients
SSC Pipeline
Astrometric Pipeline
Postage Stamp Pipeline
Frames Pipeline
Spectroscopic Pipeline
Redshifts 1-D Spectra
2-D Spectra
Plug Plate Drilling and QA at UW
Atlas Images
Pixel Maps
Final Photometric Calibration
Object Catalogs
Operational Database (Time Order)
Target Selection (Galaxies, Stars, QSO, Serendipity) and Tiling
Drilling Coordinates
Data Distribution System
Science Archive
Hand-Picked Serendipity
Catalog Archive Server
Data Archive Server
sdssQT Query Tool
Upload for Queries
MAST Interface
Query Results ASCII Catalog FITS Binary Catalog
Query Results Selected products written as .tar, .gz, or .zip ASCII Catalog FITS Binary Catalog
Astronomer
Survey Coordination
Plate Layout Tool
Monthly Observing Plan
Stripe Planning Tool
Plate Planning Tool
Patch Planning Tool
Survey Progress Report Tools
www.sdss.org Quarterly Project Reports

The Workings

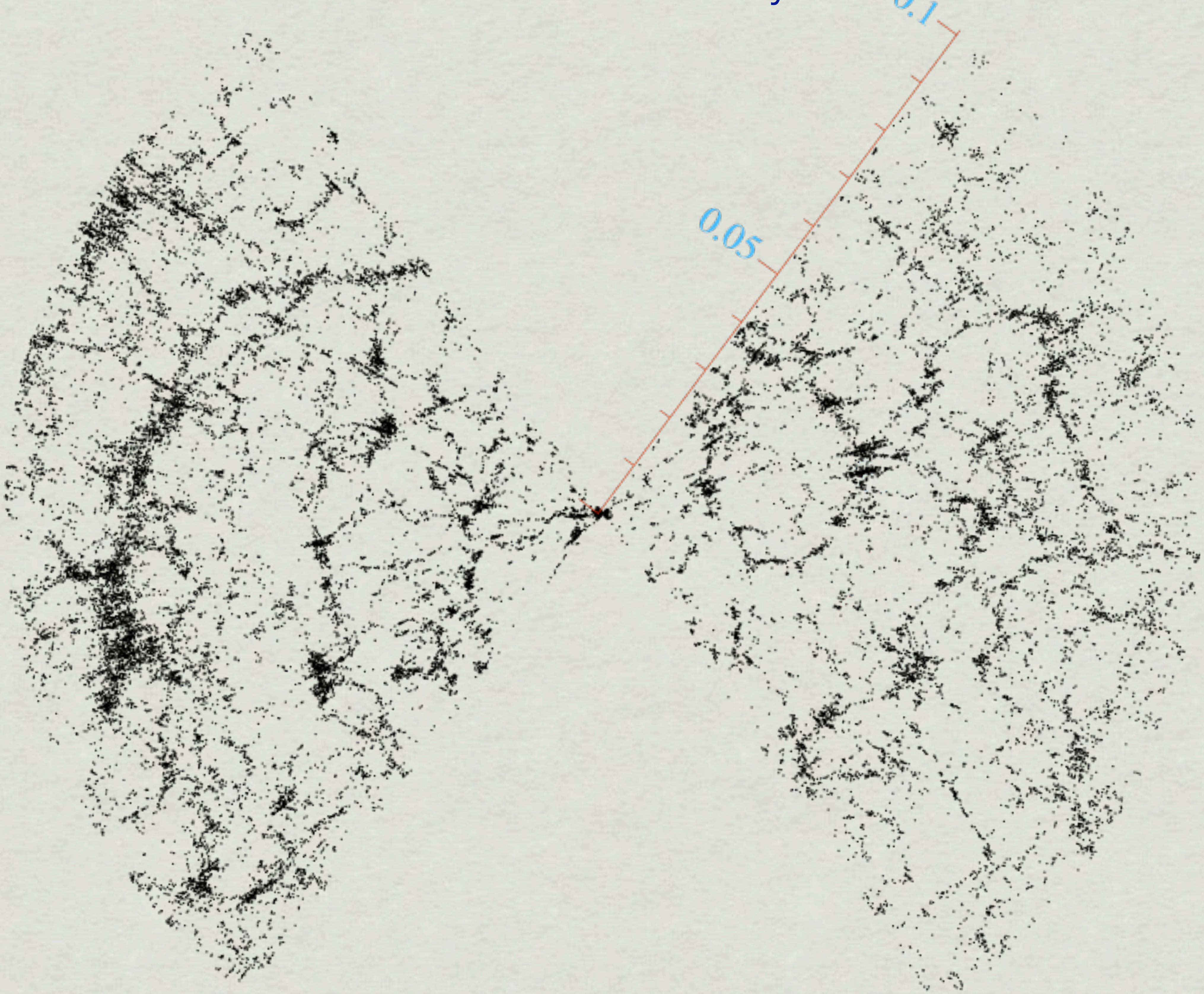


- Images are taken in a drift-scan using a mosaic camera with five filters. Each scan is called a strip, but to cover gaps a second scan is made and the combination is called a stripe. The exposure time is ~54s with ~73s between each filter.
- The images are then analysed using a code called Photo and based on the measurements objects are selected for spectroscopy.
- Spectroscopy is carried out the following season using a fibre spectrograph with **3” fibres** covering 3800Å-9000Å.
- Spectroscopic data are analysed using two pipelines and imaging & spectroscopy is released to the public.

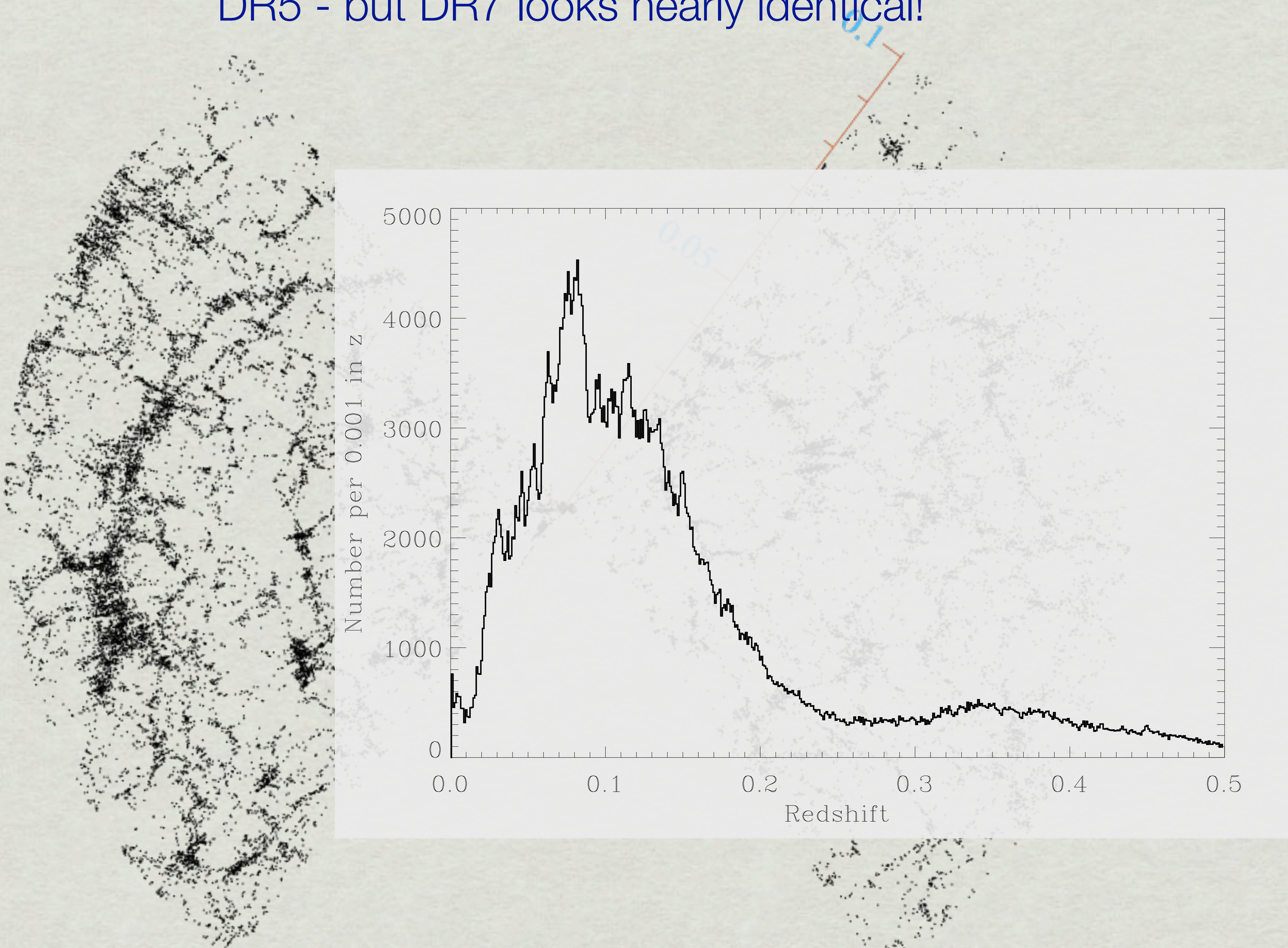
The SDSS DR7 - (Autumn 2008)

- 5 band (ugriz) imaging of 11,663 deg^2 [~16Tb].
- 3.5×10^8 unique objects.
- R~1800 spectroscopy covering 3800-9200Å for 1,640,960 objects [929,555 galaxies, 121,373 QSOs, 380,214 stars including 84,047 M and later].
- Data kept in a database which can efficiently be queried using SQL.
- **Main galaxy sample: r < 17.77** - also a QSO sample, Luminous Red Galaxies sample + stars & special objects.
- DR7 was the final release for SDSS-II - surveys are still on-going but different in goal & structure.
- SEGUE & stellar parameters: [Fe/H], T_{eff}, log g for all stars. Also Legacy & SDSS SNe.
- Median seeing for images: 1.43” (50%: [1.3”, 1.58”])

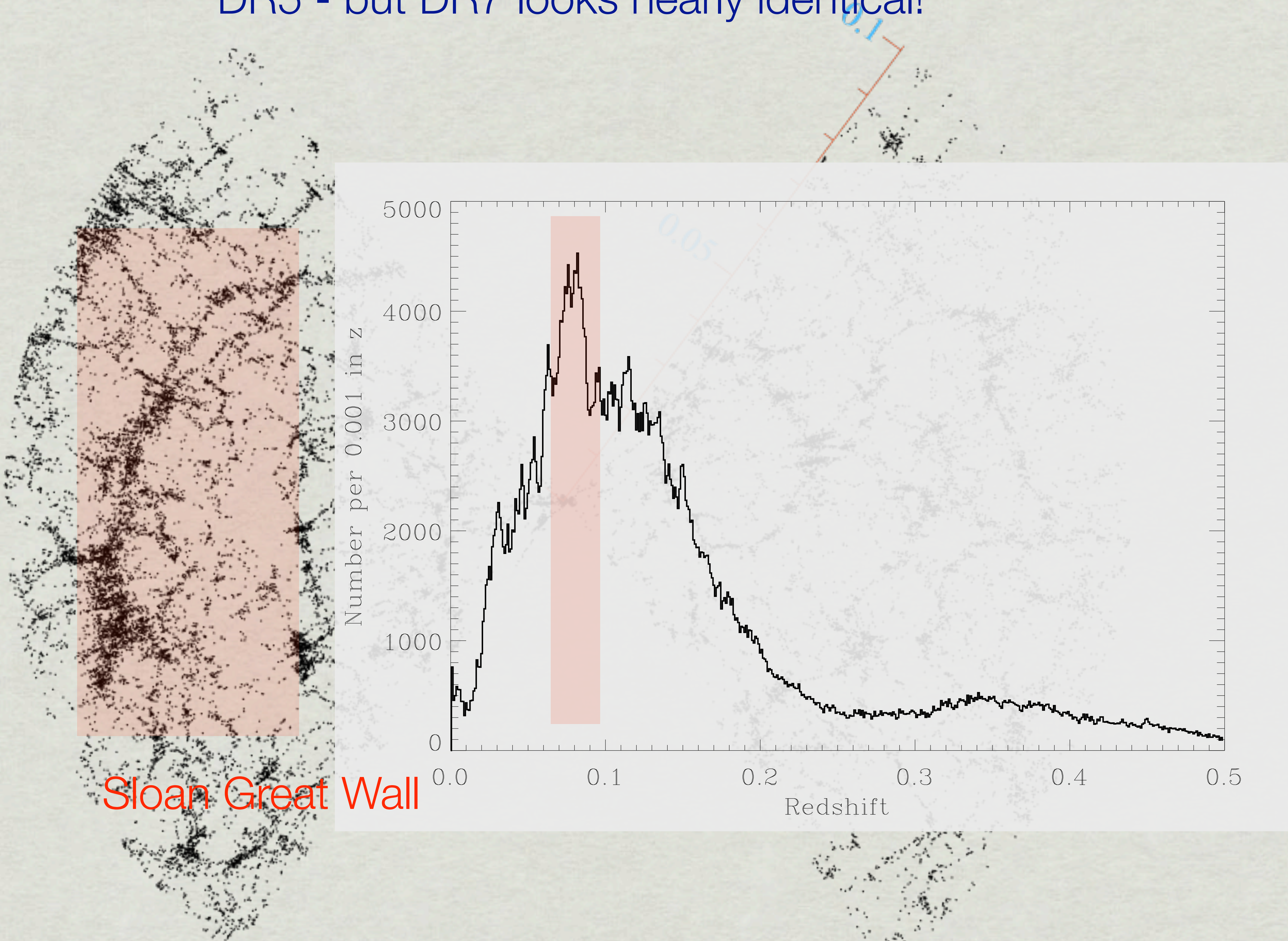
DR5 - but DR7 looks nearly identical!



DR5 - but DR7 looks nearly identical!



DR5 - but DR7 looks nearly identical!



Magnitude System(s)

Normal magnitudes:

$$m = -2.5 \log_{10} f + \mathrm{zeropoint}$$

asinh magnitudes (luptitudes):

$$m = -\frac{2.5}{\ln 10}\left[\mathrm{asinh}\left(\frac{f/f_0}{2b}\right) + \ln b\right]$$

The difference is small (irrelevant) for bright objects but is very important at faint flux levels!

nano-maggies:

A linear flux measurement such that 1 nMgy corresponds to a conventional magnitude of 22.5. Widely used in the NYU Value Added Galaxy Catalogue (VAGC)

Magnitude System(s)

Relationship to other photometric calibrations:

The SDSS is a near AB magnitude system. Thus conversion to Janskys is fairly easy:

$$m_{\rm AB} = 2.5 \times (23 - \log_{10} F\,[\rm Jy]) - 48.6$$

But there are some small offsets, such that:

$u_{AB} = u_{SDSS} - 0.04$ & $z_{AB} = z_{SDSS} + 0.02$

To convert to Johnson etc systems:

http://www.sdss.org/dr7/algorithms/sdssUBVRITransform.html

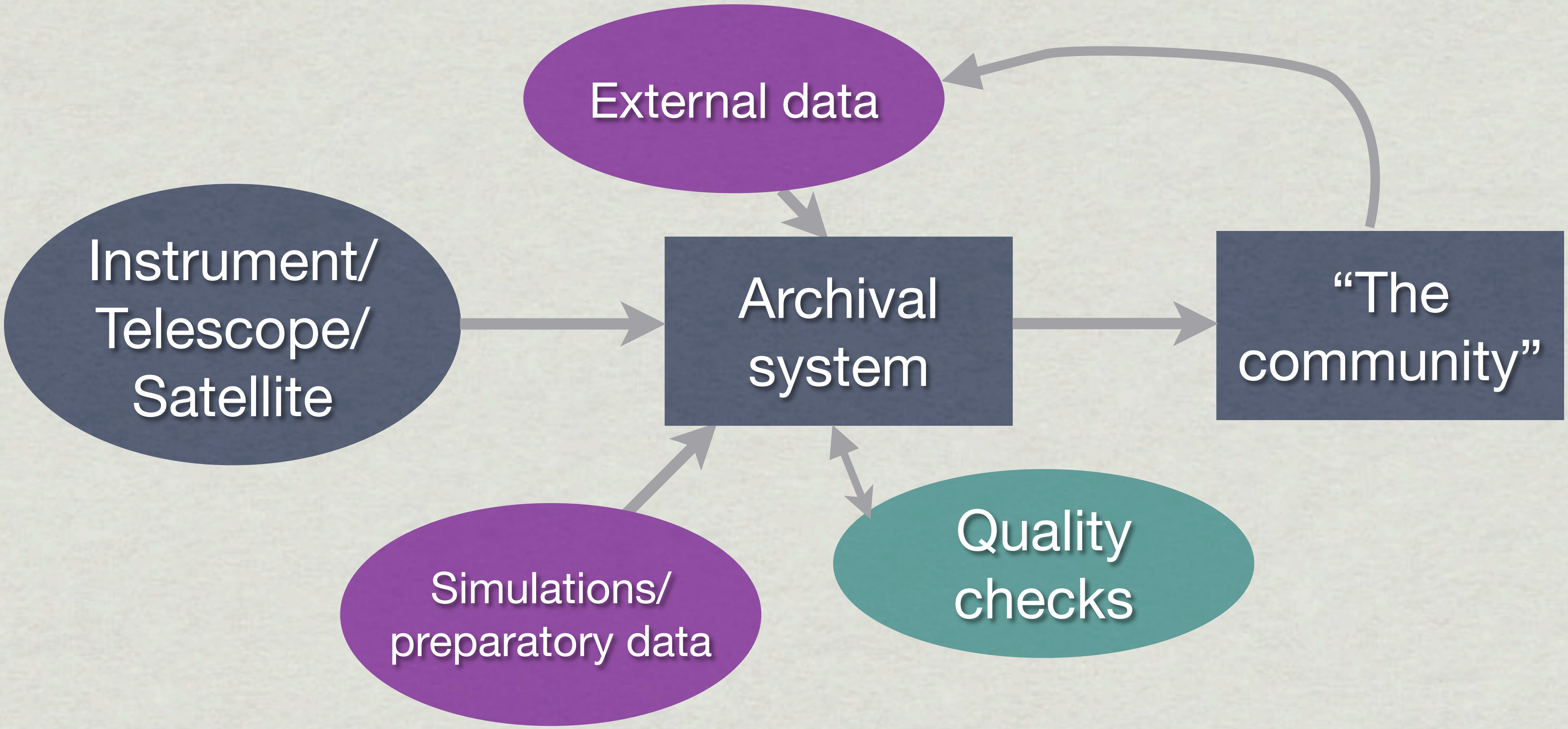
An example (for stars, from Robert Lupton):

```
B = g + 0.3130*(g - r) + 0.2271
V = g - 0.5784*(g - r) - 0.0038
R = r - 0.2936*(r - i) - 0.1439
I = i - 0.3780*(i - z)  -0.3974
```

Be aware that there now is Ubercalibration: Reduced systematic effects across the entire survey (~1%).

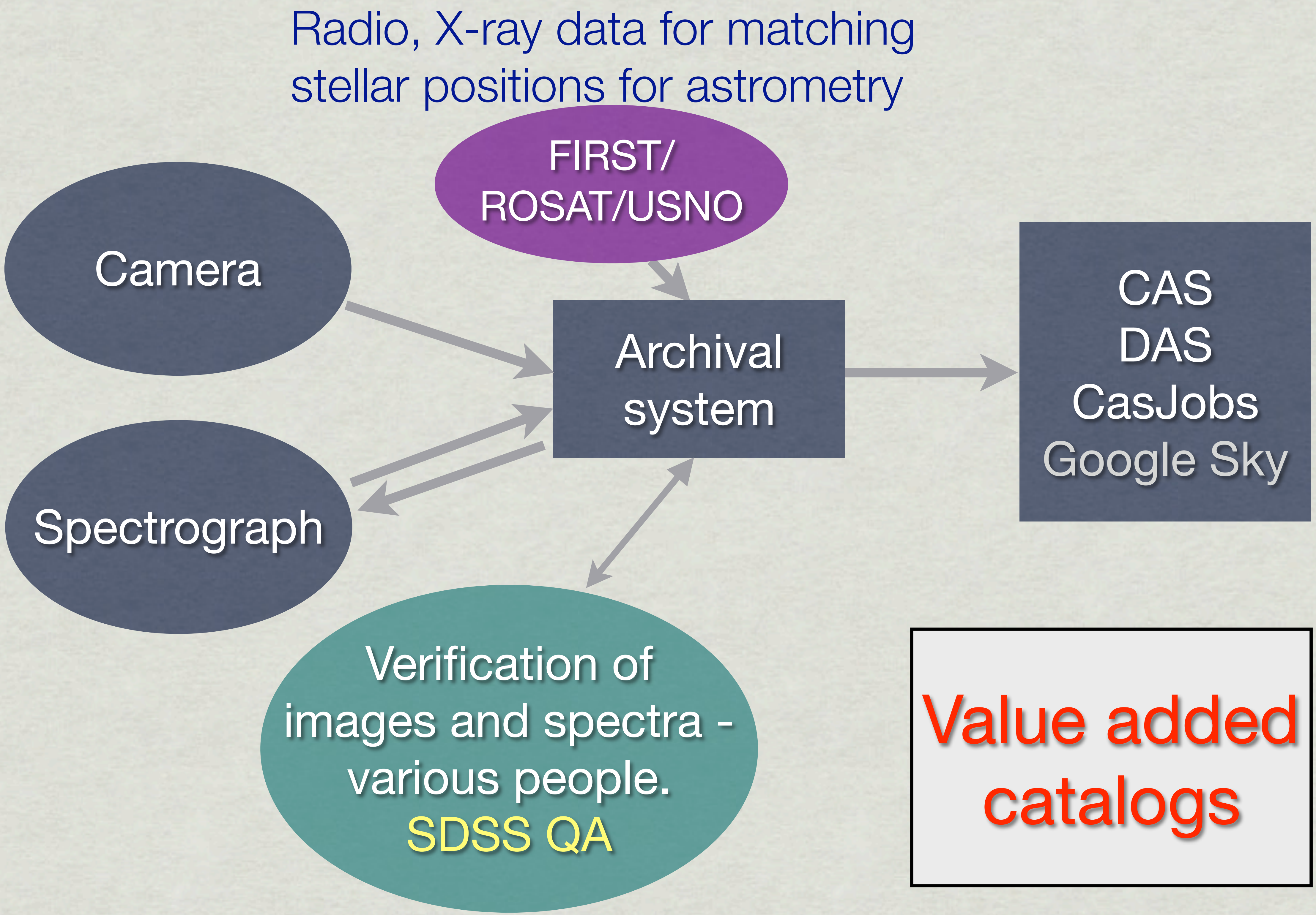
(Padmanabhan et al 2007)

Storing & organising the data

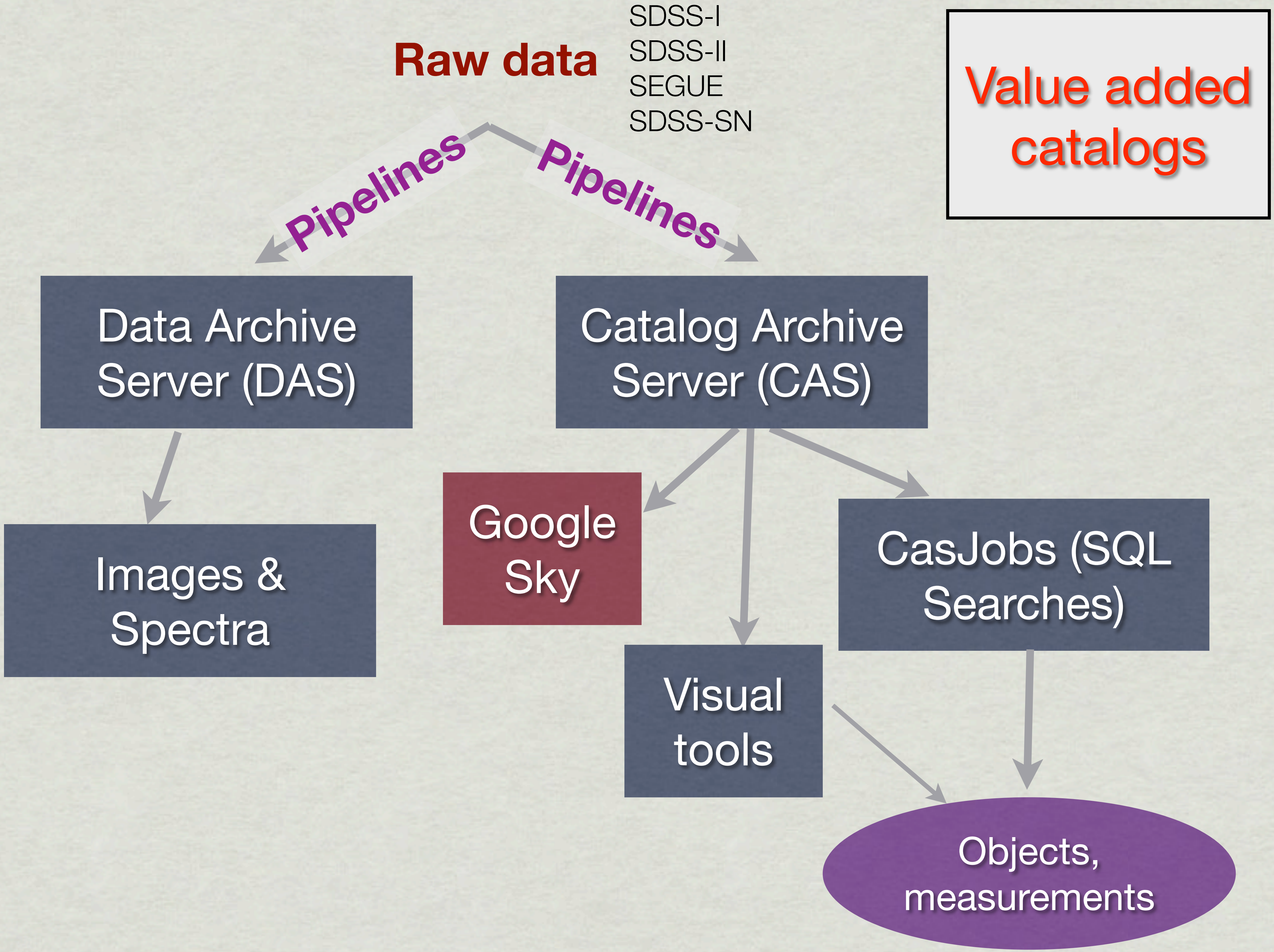


How does this look for the SDSS?

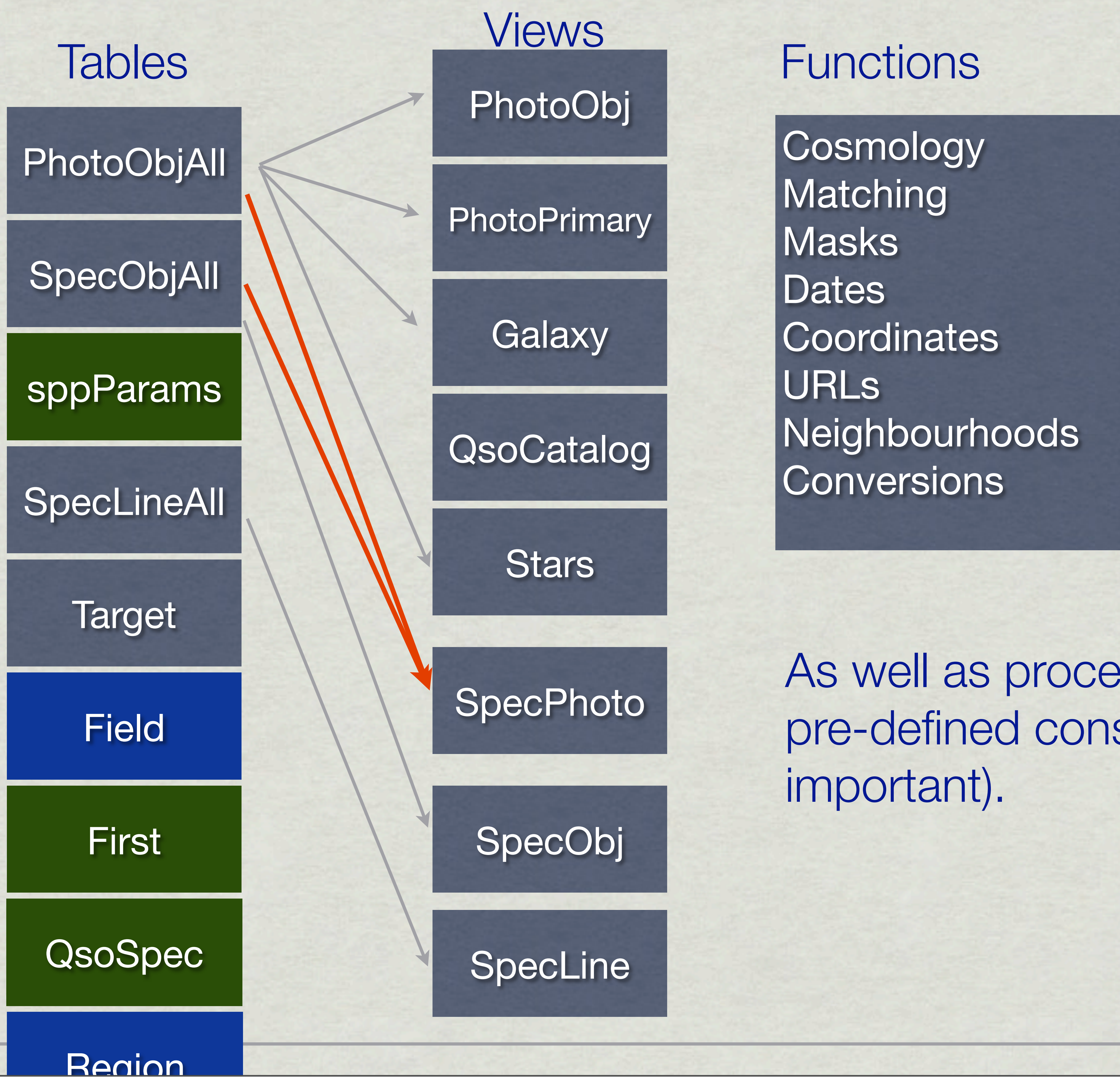
Storing & organising the data in the SDSS



Data access/archive



The structure of the database



As well as procedures and pre-defined constants (can be important).

Structure - keys & indices

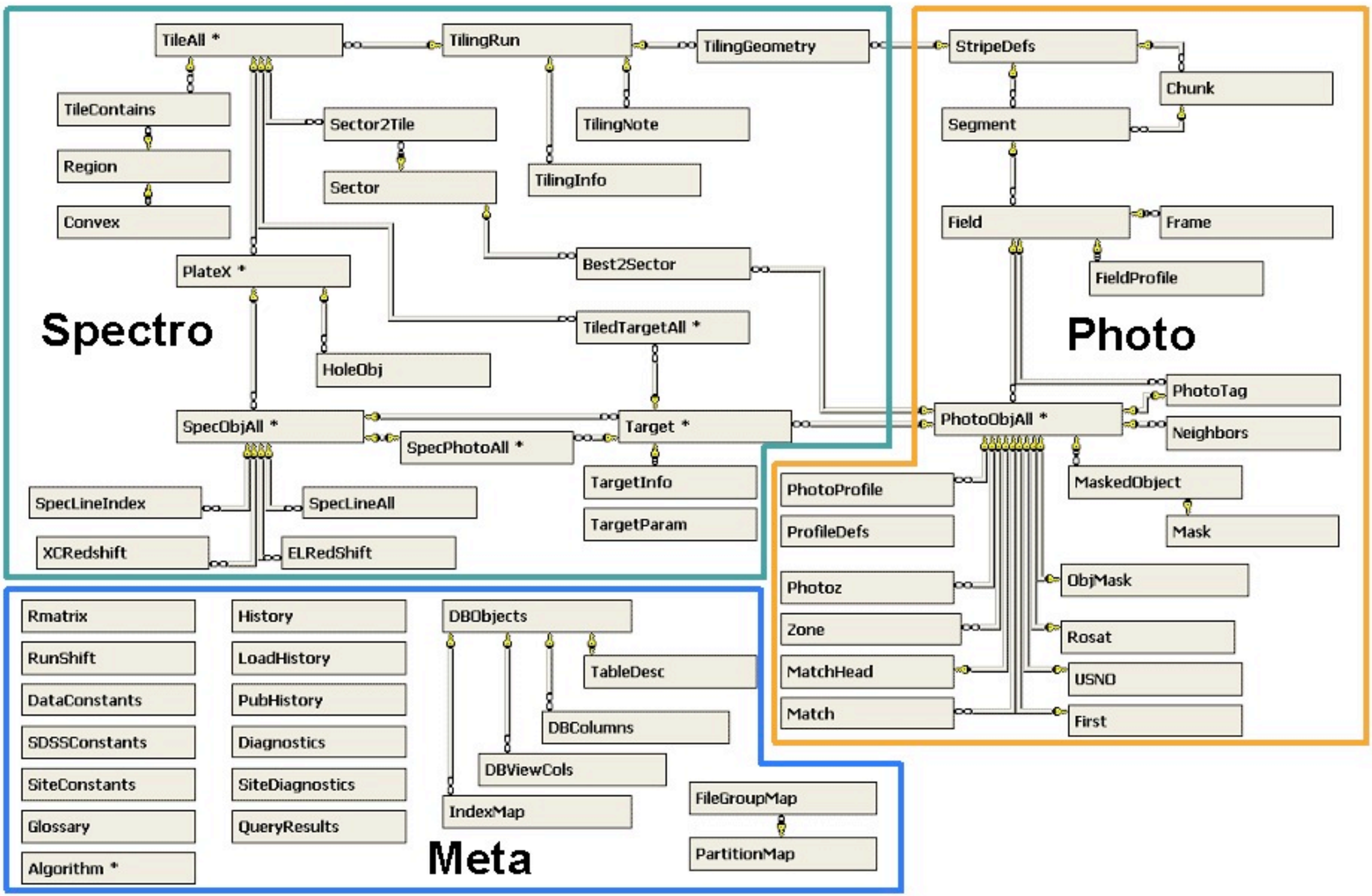
How are the tables/objects linked?

Gray et al 2002: http://fr.arxiv.org/abs/cs/0202014

Szalay et al 2002: http://fr.arxiv.org/abs/cs/0202013

Sloan Digital Sky Survey Data Release 1 (SDSS DR1) Schema

(best)



Structure - keys & indices

How are the tables/objects linked?

That was more detail than you would normally need so let us focus:

Each photometrically detection is assigned an **objID** this depends on the reduction run & the same physical source can have many objIDs!

The objID at time of target selection is called **targetObjID**
The currently best reductions give rise to **bestObjID**.

Each spectroscopic observation is assigned a **specObjID**. This identifies the spectrum.

To link spectra to photometry we need to match on bestObjID or targetObjID depending on what we need.

Structure - keys & indices

How are the tables/objects linked?

To link spectra to photometry we need to **match on bestObjID or targetObjID** depending on what our needs.

But there are also other indices:

You use objID to look up an entry but you usually wants something else (magnitude, position etc.)

These are therefore encoded as **covering indices** and are also fast to search on. See the **Indices** list in the Schema browser.

PhotoObjAll and views thereof

Photometric data comes from Photo:

Magnitudes [ugriz, Petrosian, Model, Apertures, PSF]

e.g.: petroMag_f, modelMag_u, fiberMagErr_z, psfMag_g

Sizes [Petrosian radii, Image moments]

e.g.: PetroR50_g, petroR90_r, isoA_r, deVRad_g

Positions [better than 0.1”, often multiple observations]

e.g.: ra, dec, colc, rowc, rowv, colv, etc.

Shapes, orientations [image moments]

e.g.: mE1_r, mE2_r, isoPhi_r, expAB_r

Images [full field, JPGs, objects only - “atlas images”]

Functions can be used to create a link

Image quality [position dependent PSF]

Most be got from the **FIELD** table.

Photometry - Advice

Reliability:

The r-band is in general the most reliable - u & z might be poor quality.

r<21.5 should in general be fine, but be aware of single-band detections (typical detection limit r_{AB} ~ 22.5).

r<14.5 requires care and sky estimates for very large galaxies are often poor. See e.g. Blanton et al (2005) for details.

Large, actively star forming galaxies are sometimes shredded with multiple spectroscopic targets.

Take care to check imaging **flags**.

Photometry - Advice

Reliability:

The r-band is in general the most reliable - u & z might be poor quality.

r<21.5 should in general be fine, but be aware of single-band detections (typical detection limit r_{AB} ~ 22.5).

r<14.5 requires care and sky estimates for very large galaxies are often poor. See e.g. Blanton et al (2005) for details.

Large, actively star forming galaxies are sometimes shredded with multiple spectroscopic targets.

Take care to check imaging **flags**.

What should I use?

For colours: Model magnitudes or aperture magnitudes.

For total flux: Petrosian or optimal model magnitudes.

Fiber magnitudes are useful to check spectrophotometry.

Imaging Flags...

These are set as individual bits in a long integer to indicate whether all was well with the photometric analysis. It is crucial to check these when doing accurate work! Some examples:

- `SATURATED`: Tells you whether the image of an object was saturated.
- `EDGE`: Is the object on the edge (commonly the case for large galaxies).
- `CHILD`: Is this object part of a larger object that was split?
- `MOVED`: Did the object move?
- etc. etc. See http://www.sdss.org/dr7/products/catalogs/flags.html for details.

Spectroscopy - SpecObj et al

z, plate, mjd, fiberID, zConf

Absorption line indices [Lick + a few more]

In **SpecLineIndex**

Emission lines [Fluxes, EWs, widths]

In **SpecLine**

Velocity dispersions [Two pipelines, Elodie PCA]

velDisp

Spectroscopy - SpecObj et al

z, plate, mjd, fiberID, zConf

Absorption line indices [Lick + a few more]

In **SpecLineIndex**

Emission lines [Fluxes, EWs, widths]

In **SpecLine**

Velocity dispersions [Two pipelines, Elodie PCA]

velDisp

Probably the area where using Value-Added Catalogues is a useful alternative!

In particular the MPA-JHU database with more careful continuum subtraction and the NYU VAGC which has more matching and is an excellent base sample for further work.

Spectroscopy - Advice

Reliability:

[O II]3727Å is not measured for low-z galaxies.

Very strong emission lines are sometimes clipped.

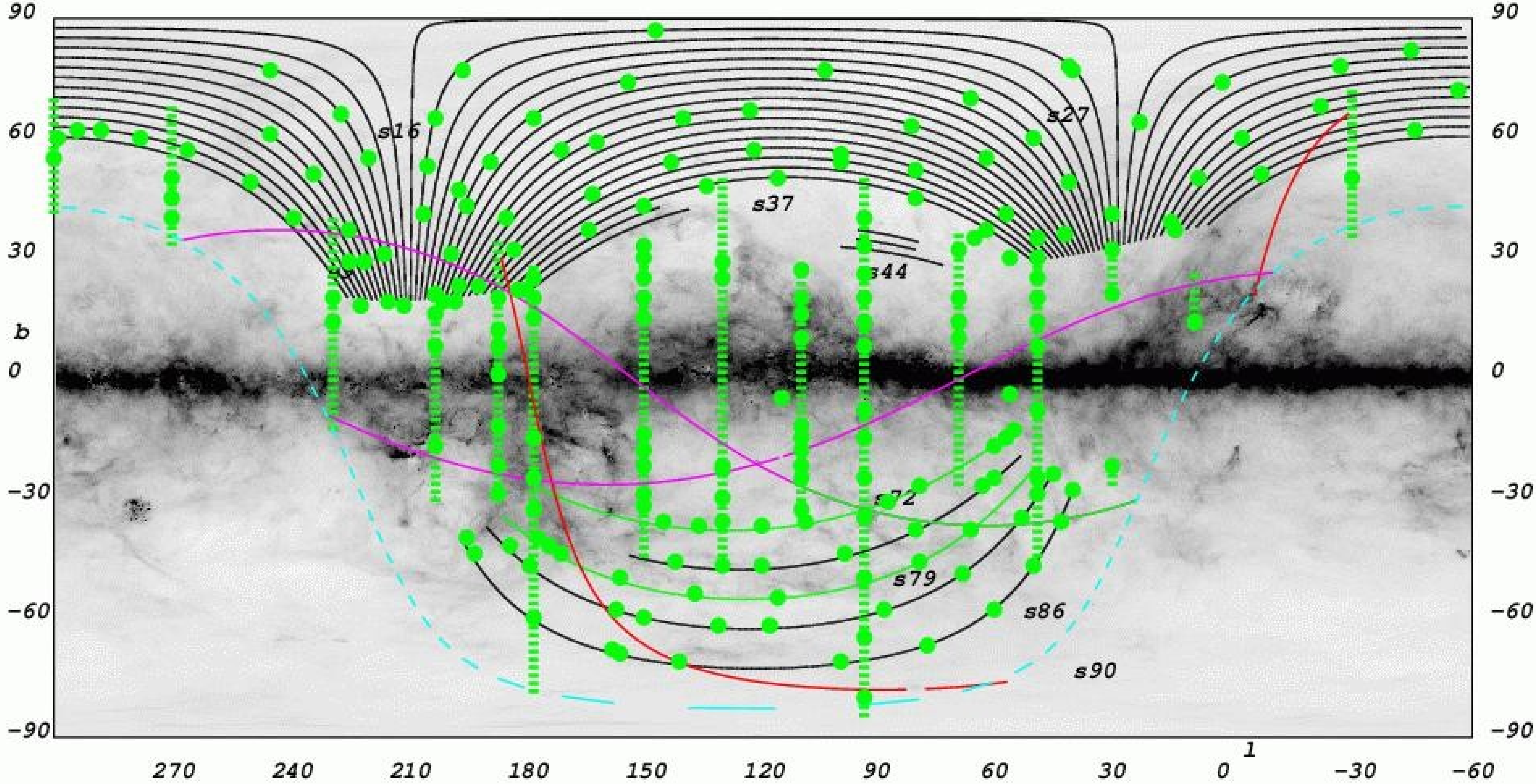
Bright sky and/or CCD defects might cause some lines not to be measured.

Sky subtraction is not very good in the red [c.f. Wild et al 2005]

The spectrophotometric calibration for DR6&7 is tied to PSF magnitudes - watch out for galaxies!

The slope of the spectrum in the blue is only good to the ~few % level.

SEGUE imaging sky coverage and plate layout
July 15, 2008 (Complete)



SEGUE - sppParams & sppLines

Stellar analysis from SEGUE

- **Absorption line indices** [Lick + a few more]
  - All found in sppLines, e.g.: H8, Ca H+K, Hδ, Ca I etc.
- **Stellar parameters** [Teff, log g, [Fe/H], RV]
  - Multiple pipelines - useful for intercomparison but for first look the "adopted" versions are good: teffa, logga, feha, sptypea, elodierv

SEGUE - sppParams & sppLines

Stellar analysis from SEGUE

Absorption line indices [Lick + a few more]

Stellar parameters [Teff, log g, [Fe/H], R_V]

The stellar parameters are derived from a number of different pipelines. Since the spectra in the SDSS are not optimal to derive stellar parameters it is necessary to compare these estimates to assess the quality of the results! And also to get external data with better resolution. ([Fe/H] ~ 0 show noticeable systematic offset).

SEGUE in general targets lower Galactic latitudes and crowded fields and go outside the normal SDSS footprint. Thus much of the data is not in the normal CAS context.

Further data

Repeat observations of Stripe 82. These are sometimes taken in poor conditions and are therefore provided in uncalibrated form. These data were in part used for a supernova search.

Moving objects [from repeat observations - note that fast-moving objects can be found from individual runs by comparing different filters]

Variability [from repeat observations]

Observing conditions [extinction, seeing etc.]

Raw data [raw counts etc.]

Value-Added Data

Very important!

When people have done some careful job they might offer this as a value-added catalogue. Using these, rather than the official SDSS data might **save you a lot of time and effort**!

Finding out about these:

- Read papers!!
- Check the SDSS VAC page:
- Listen carefully - as I will say this only once

Value-Added Data

- The NYU VAGC
  - Large-Scale Structure samples with well characterised selection functions.
  - Cross-matches of SDSS to other surveys.
- MPA-JHU VAGC
  - Improved spectroscopic data reduction, line fluxes and indices.
  - Derived quantities such as SFR, O/H and stellar masses etc.
- QSO catalogues (Schneider et al) & BAL-QSOs.
- White Dwarf catalogues (Eisenstein et al)
- Variable star catalogues from Stripe 82 & CV catalogues
- Moving objects catalogues
- Galaxy cluster catalogues (MaxBCG, cut & enhance)
- & more!

The NYU VAGC

http://sdss.physics.nyu.edu/vagc/

See: Blanton et al (2005)

- Large-scale structure sample.
- Careful characterisation of the survey geometry.
- Low-z galaxy sample (this is non-trivial to construct).
- SDSS data sweeps - files with a useful subset of all the data with manageable size (~100 Gb).
- Sersic fits & a number of other quantities.
- Various pieces of software - in particular `kcorrect` which is the most widely used software to calculate k-corrections for SDSS data.

The MPA-JHU VAGC

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

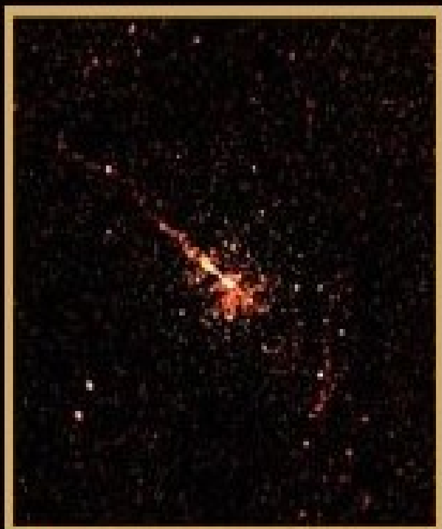
http://www.strw.leidenuniv.nl/~jarle/SDSS- soon

See e.g: Tremonti et al (2004); Brinchmann et al (2004);

- Improved continuum subtraction.
- Pipeline optimised to measure emission lines and absorption line indices on galaxy spectra.
- Emission lines, absorption line indices.
- Star formation rates, oxygen abundances, stellar masses, emission line classification of galaxies and stellar metallicities.
- All the photometric information available in tsObj files for spectroscopic targets.
- Data organised in a set of FITS files and available for DR7.

Centaurus A – Peculiar Galaxy

Distance: 11,000,000 ly light-years (3.4 Mpc) | Image Size = 15 x 14 arcmin | Visual Magnitude = 7.0



X-Ray: Chandra

Ultraviolet: GALEX

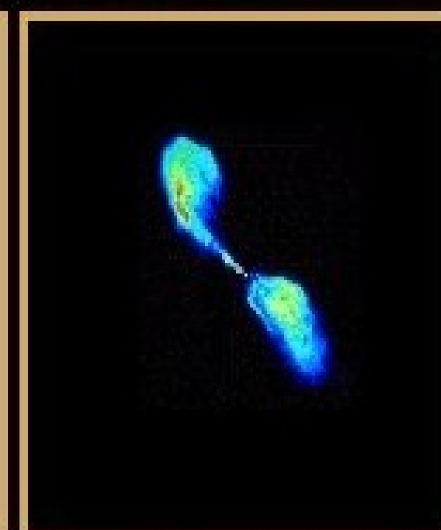
Visible: DSS

Visible: Color ©AAO

Near-Infrared: 2MASS

Mid-Infrared: Spitzer

Far-Infrared: IRAS



Radio: VLA

Data Centres

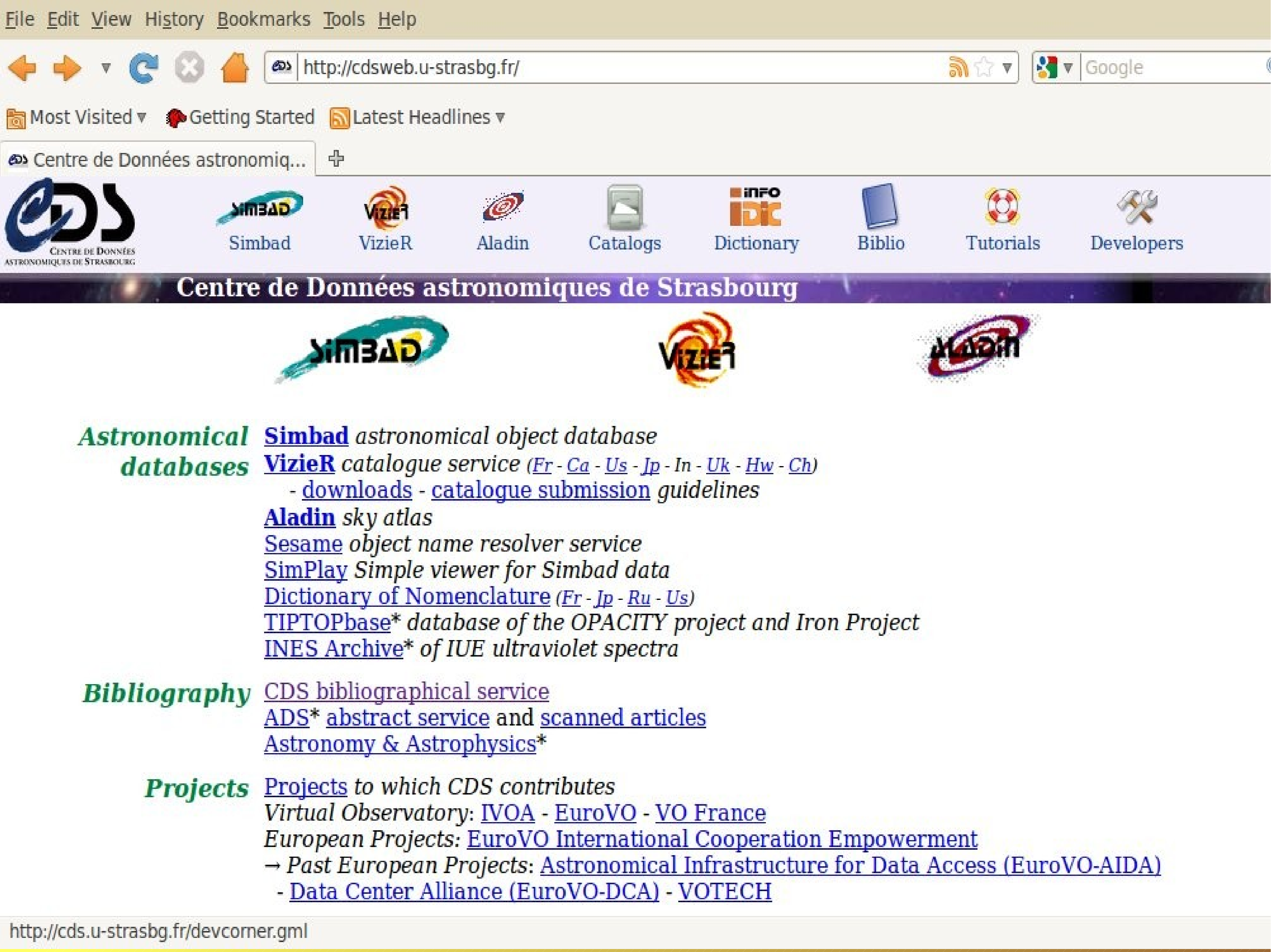
There are 2 general purpose Data Centers in the World:
They try to be Object Centers, that contain all data for certain objects.

Centre des Donnéés Stellaires (CDS)

For stars in our Milky Way and resolved stars in nearby galaxies

NASA's Extragalactic Database (NED)

Extragalactic Database


File Edit View History Bookmarks Tools Help
http://cdsweb.u-strasbg.fr/
Google
Most Visited
Getting Started
Latest Headlines
Centre de Données astronomiq...
Simbad
VizieR
Aladin
Catalogs
Dictionary
Biblio
Tutorials
Developers
Centre de Données astronomiques de Strasbourg
Astronomical databases
Simbad astronomical object database
VizieR catalogue service (Fr - Ca - Us - Jp - In - Uk - Hw - Ch)
- downloads - catalogue submission guidelines
Aladin sky atlas
Sesame object name resolver service
SimPlay Simple viewer for Simbad data
Dictionary of Nomenclature (Fr - Jp - Ru - Us)
TIPTOPbase* database of the OPACITY project and Iron Project
INES Archive* of IUE ultraviolet spectra
Bibliography
CDS bibliographical service
ADS* abstract service and scanned articles
Astronomy & Astrophysics*
Projects
Projects to which CDS contributes
Virtual Observatory: IVOA - EuroVO - VO France
European Projects: EuroVO International Cooperation Empowerment
→ Past European Projects: Astronomical Infrastructure for Data Access (EuroVO-AIDA)
- Data Center Alliance (EuroVO-DCA) - VOTECH
http://cds.u-strasbg.fr/devcorner.gml

File Edit View History Bookmarks Tools Help

http://nedwww.ipac.caltech.edu/

Most Visited | Getting Started | Latest Headlines

NASA/IPAC Extragalactic Datab...

NASA/IPAC EXTRAGALACTIC DATABASE

Latest updates to NED content and interface (September 2010)

- Over 760,000 new data entries
- Additions to Level5, including M. Cignoni and Monica Tosi (2010)
- Spectra from R. Giovanelli et al (2004)
- Compilation of Distances: NED-D complete download available
- Galaxy Morphology: 16,274 entries from The ESO/Uppsala survey of the ESO(B) Atlas
- New features in the *Classifications by Object Name* service

OBJECTS	DATA	LITERATURE	TOOLS	INFO
By Name	Images By Object Name or By Region	References by Object Name	Coordinate Transformation & Extinction Calculator	Introduction Latest News/Updates
Near Name	Photometry & SEDs	References by Author Name	Velocity Calculator	Features FAQ
Near Position	Spectra	Text Search	Cosmology Calculators	Overview (pdf)
IAU Format	Redshifts	Knowledgebase LEVEL 5	Extinction-Law Calculators	Source List
By Parameters	Redshift-Independent	Galaxy Distance	Skyplot	Web Links

http://nedwww.ipac.caltech.edu/forms/OBJatt.html

Observatory Data Centers

Many observatories make their data public after one year, and have an archiving facility
These include:

ESO (all telescopes)
NRAO VLA (Very Large Array)
ING La Palma
Gemini
NOAO
Subaru
AAT

BIBLIOGRAPHICAL SERVICES

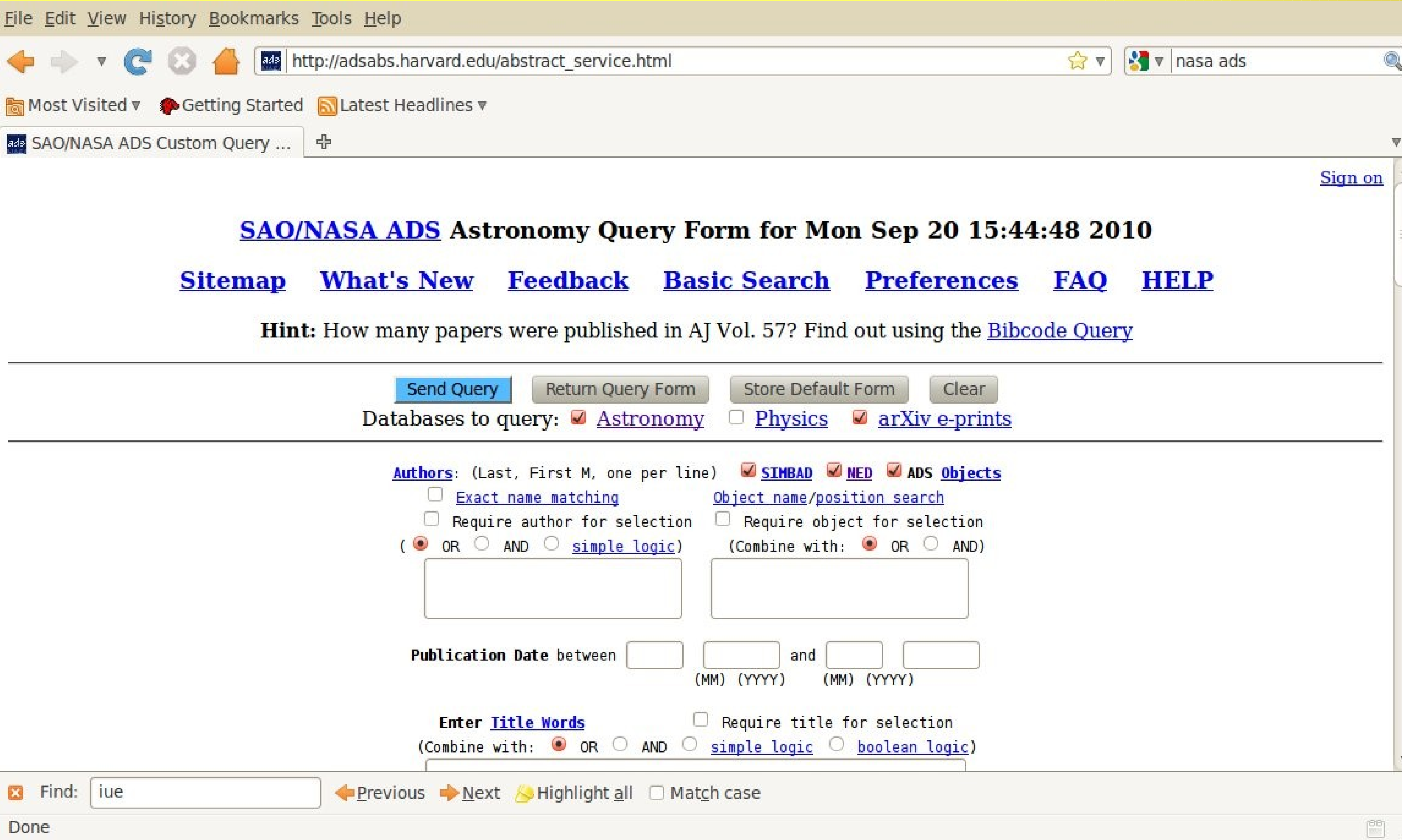
NASA's Astrophysics Data System (http://adswww.harvard.edu/) (1993-) is a unique system in the scientific world, which

- Gives access to electronic journal articles
- Contained scanned copies of almost all journals since their first issue.
- Gives links to references, object databases and other resources
- Gives a unique Refcode (19 char) to every paper
- myADS (configurable service by user)

BIBLIOGRAPHICAL SERVICES

Since 1992 there is a preprint service (ArXiv.org) which contains papers in advance of publication.

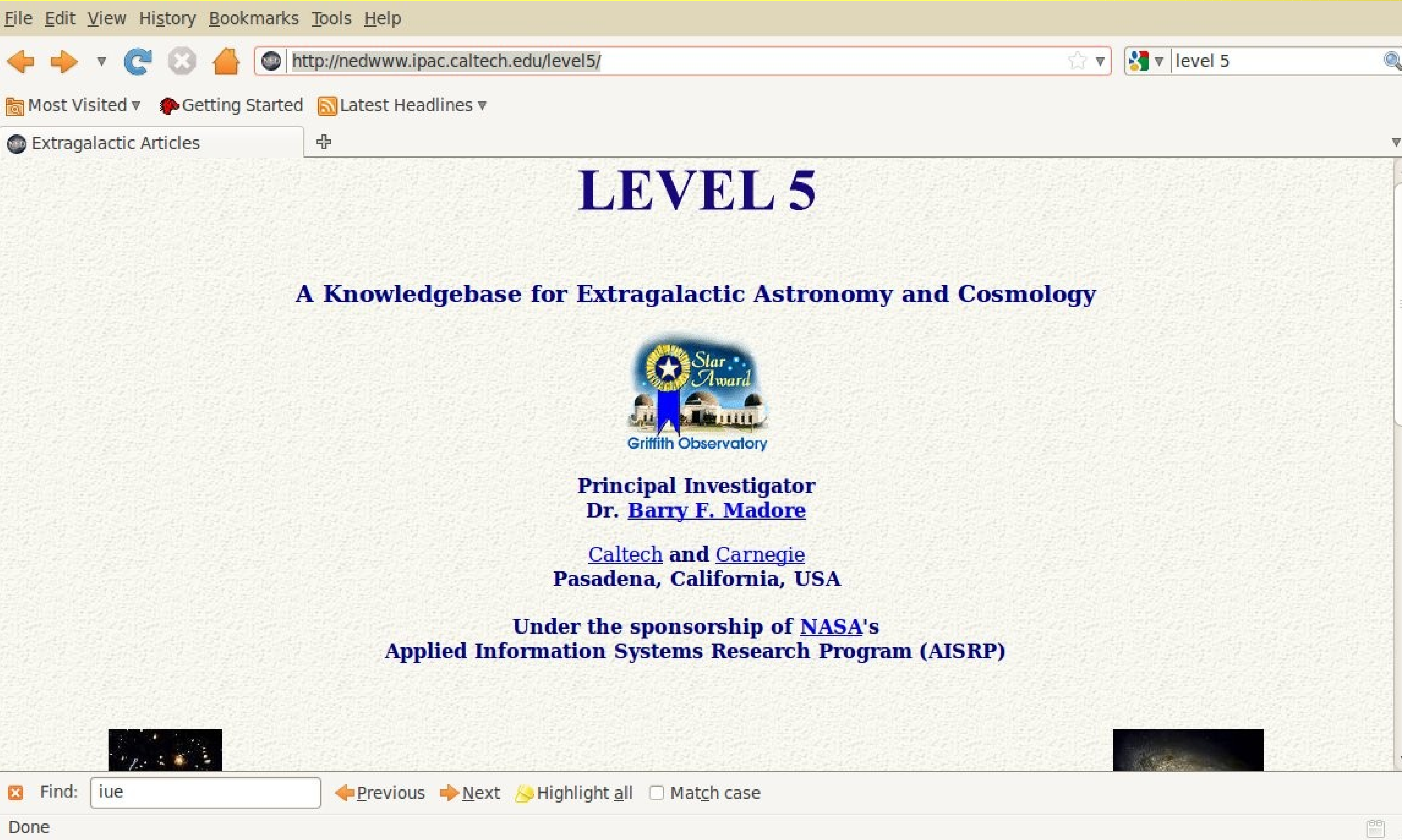
- Linked to ADS
- Free
- Contains most papers, so that access is possible without paying (e.g. from home)


File Edit View History Bookmarks Tools Help
http://adsabs.harvard.edu/abstract_service.html
nasa ads
Most Visited
Getting Started
Latest Headlines
SAO/NASA ADS Custom Query ...
Sign on
SAO/NASA ADS Astronomy Query Form for Mon Sep 20 15:44:48 2010
Sitemap What's New Feedback Basic Search Preferences FAQ HELP
Hint: How many papers were published in AJ Vol. 57? Find out using the Bibcode Query
Send Query
Return Query Form
Store Default Form
Clear
Databases to query: Astronomy Physics arXiv e-prints
Authors: (Last, First M, one per line)
SIMBAD NED ADS Objects
Exact name matching
Object name/position search
Require author for selection
Require object for selection
(OR AND simple logic)
(Combine with: OR AND)
Publication Date between and
(MM) (YYYY) (MM) (YYYY)
Enter Title Words
Require title for selection
(Combine with: OR AND simple logic boolean logic)
Find: iue
Previous
Next
Highlight all
Match case
Done

A KNOWLEDGE DATABASE

NED has tried to create a database with the most
important articles, arranged by subject, of current
and lasting interest to cosmologists, particle physicists
and extragalactic astronomers.

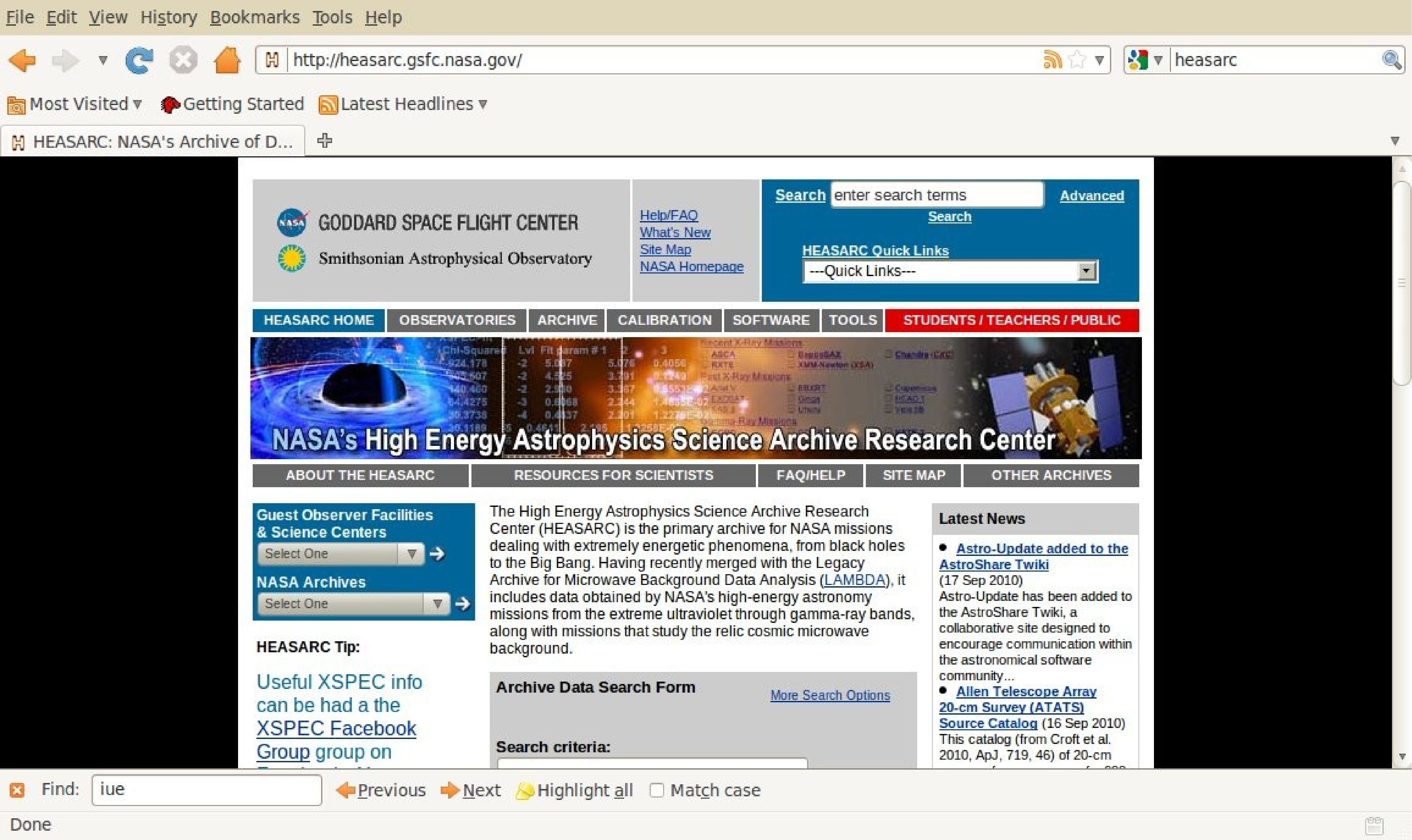
See http://nedwww.ipac.caltech.edu/level5


File Edit View History Bookmarks Tools Help
http://nedwww.ipac.caltech.edu/level5/
level 5
Most Visited
Getting Started
Latest Headlines
Extragalactic Articles
LEVEL 5
A Knowledgebase for Extragalactic Astronomy and Cosmology
Star Award
Griffith Observatory
Principal Investigator
Dr. Barry F. Madore
Caltech and Carnegie
Pasadena, California, USA
Under the sponsorship of NASA's
Applied Information Systems Research Program (AISRP)
Find: iue
Previous
Next
Highlight all
Match case
Done

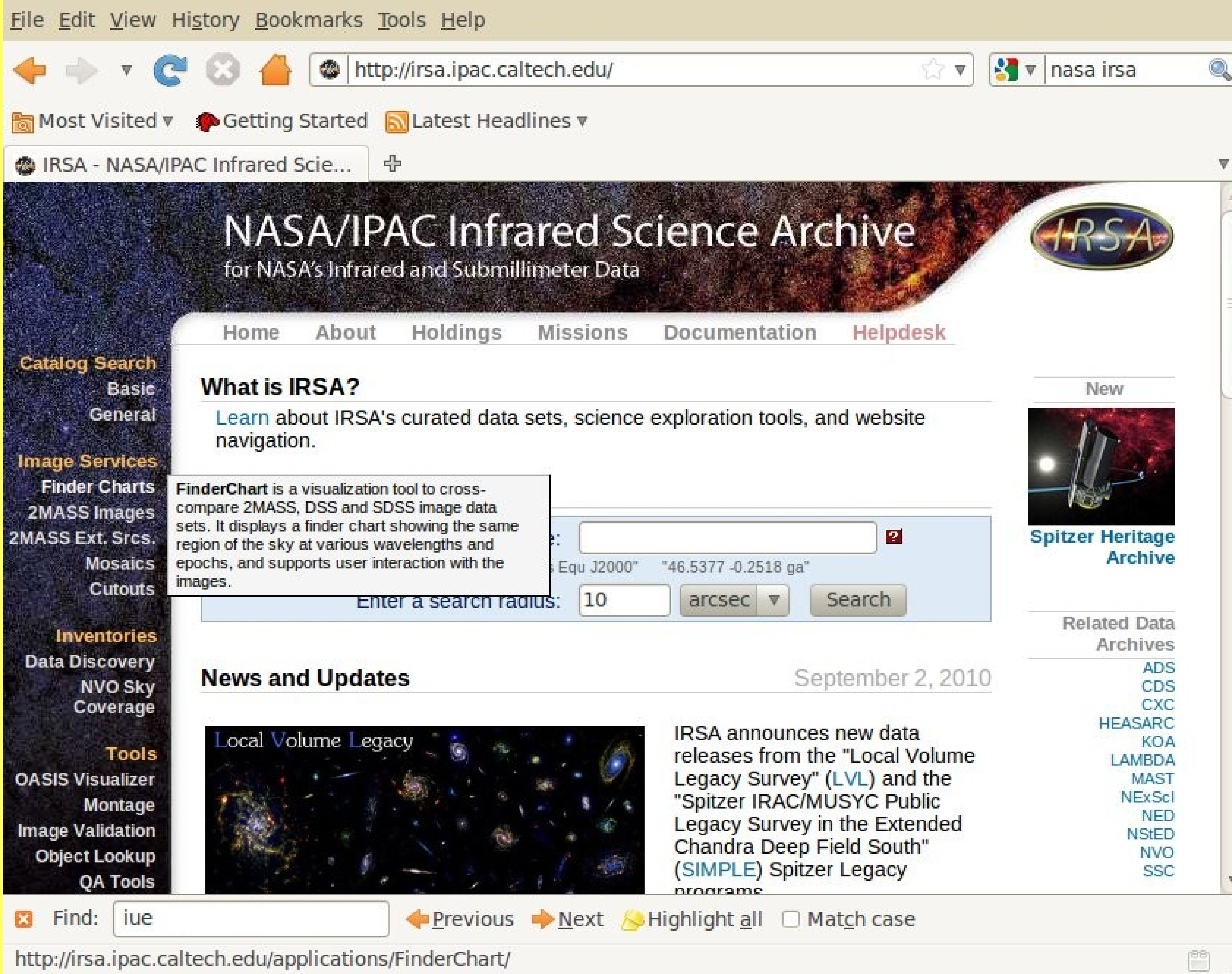
MULTI-MISSION SERVERS

The increasing number of space missions have led to the creation of a number of multi-mission data centers, with data from space missions

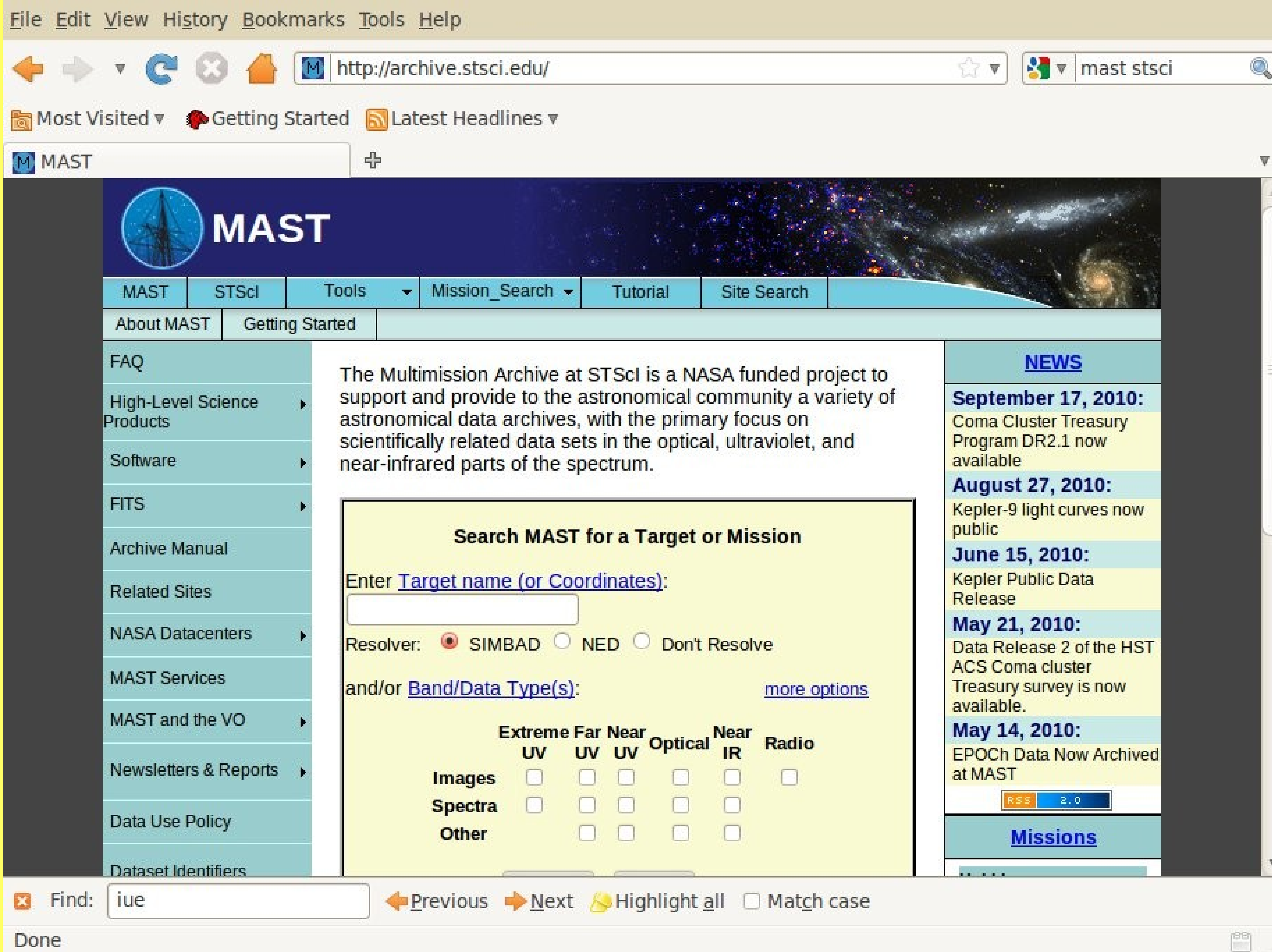
- IPAC (http://www.ipac.caltech.edu/) - Infrared
- HEASARC (http://heasarc.gsfc.nasa.gov/) High Energy
- MAST (http://archive.stsci.edu/) Optical/Multi Purpose
- CADC (http://cadcwww.dao.nrc.ca/cadc/) Mult. Purpose



HEARSRC is the primary archive for high-energy astronomy missions, In the extreme UV, X-ray and Gamma rays. Instruments: ASCA, BeppoSAX,CGRO, Chandra, EUVE, INTEGRAL, etc.



At IRSA data can be found from NASA's Infrared and Submm missions.For example IRAS, ISO, Spitzer, 2MASS.

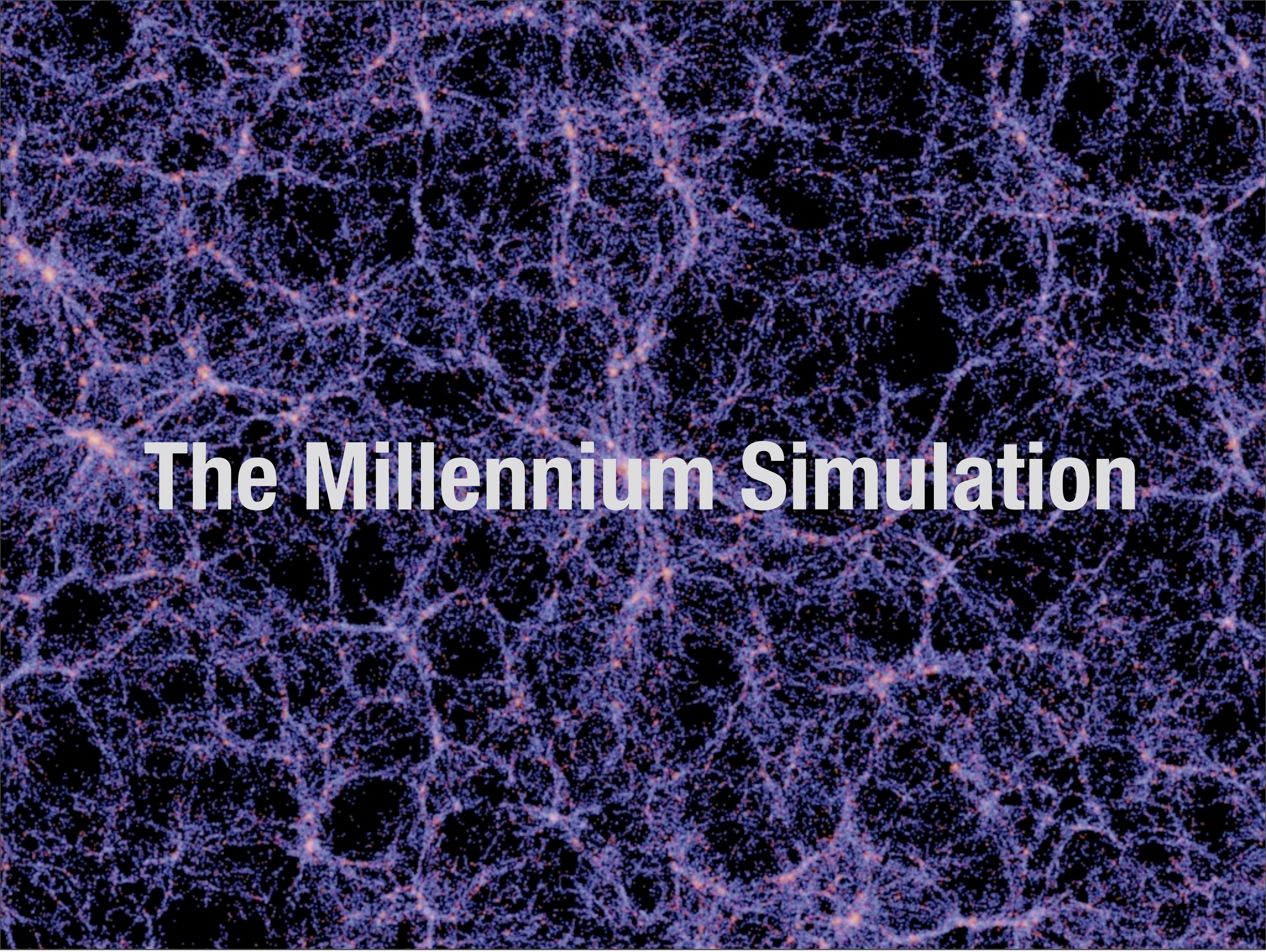


MAST supports many astronomical archives, in Opt, UV and IR. Examples are HST, GALEX , IUE, FUSE.

Site with many archives (Space and Ground-based). Has the meeting server.
Examples of archives: CFHT, JCMT, GEMINI

Caveats when using databases

- Documentation might be wrong or incomplete
- Data quality varying (seeing, photometric conditions)
- Varying depth in data (difficult to calculate statistically complete samples)
- Instrumental settings might be varying without the user knowing this.
- All this is worse for ground-based than for space-based archives.


The Millennium Simulation

The Millennium Simulation

A very large N-body simulation of the Universe (Springel et al 2005).

To make predictions for galaxy evolution semi-analytic models can be run “on-top”. The Millennium database has two such models - from Bower et al (2006) and de Lucia et al (2006)/de Lucia & Blaizot (2007)

Contents:

Stellar masses, star formation rates, metal content, star formation histories, assembly histories, dark matter halos, predictions on observables such as absolute magnitudes.

http://www.g-vo.org/Millennium

Example query:

```
SELECT *
  FROM millimil..DeLucia2006a
WHERE snapnum=63
     AND mag_b BETWEEN -26 AND -18
     AND stellarMass > 10
```

Example query:

```
SELECT *
  FROM millimil..DeLucia2006a
WHERE snapnum=63
     AND mag_b BETWEEN -26 AND -18
     AND stellarMass > 10
```

The catalogue to use

Example query:

```
SELECT *
  FROM millimil..DeLucia2006a
WHERE snapnum=63
     AND mag_b BETWEEN -26 AND -18
     AND stellarMass > 10
```

The snapshot number corresponds to a redshift

Example query:

```
SELECT *
  FROM millimil..DeLucia2006a
WHERE snapnum=63
     AND mag_b BETWEEN -26 AND -18
     AND stellarMass > 10
```

See the schema for the physical parameters

Example query:

Also:
Bower2006a

```
SELECT *
  FROM millimil..DeLucia2006a
WHERE snapnum=63
     AND mag_b BETWEEN -26 AND -18
     AND stellarMass > 10
```

Some issues:

This table contains only 1/512 of the full Millennium simulation - you can get access to the full one by **registering**.

There is a limit on the max number of rows and execution time

Download is a bit awkward - but there are examples on how to download straight to R/IDL and you can access it directly from Topcat which might be the best option!

Other astronomical databases

- ★The The Cornell Digital HI Archive has a SQL interface to search their database of HI detections in galaxies.
- ★The Debris disk database contains much information about debris disks but no SQL interface.
- ★The 6 degree field galaxy redshift survey does contain a nice search form for running SQL searches on their data.
- ★The RAVE survey is a radial velocity survey of up to 1,000,000 stars. It has a query interface for the public (follow the link for Data access) but no SQL.
- ★The GALEX mission has a searchable archive using the same CasJobs technology as the SDSS.
- ★The HyperLEDA catalogue for galaxy properties.

But there are of course many, many more! How do you deal with it all? Well, that is a topic for the Virtual Observatory!

Simple SQL searches

```
SELECT
    objID, u, g, r,

FROM PhotoPrimary

WHERE u - g < 0.4
    and g - r < 0.7
```

Select something

From a table/view

According to a criterion

Ordering & Groupings

If your search would return, say 100 million objects and you only want to count them in some bins, you might not want to download it all. The solution? GROUP BY.

Do you want the closest, 10 closest, brightest, 10 brightest etc.? Then ORDER BY is necessary

Ordering

I would like to find the 10 highest redshift (most distant) objects that are brighter than r = 14 [perhaps because I want to study them at high spectral resolution?]

Ordering

I would like to find the 10 highest redshift (most distant) objects that are brighter than r = 14 [perhaps because I want to study them at high spectral resolution?]

```
Select TOP 10 plate, mjd, fiberid, z
From SpecPhoto
Where
   petroMag_R < 14
AND
   petroMagErr_R < 0.1
ORDER BY -z
```

Ordering

I would like to find the 10 highest redshift (most distant) objects that are brighter than r = 14 [perhaps because I want to study them at high spectral resolution?]

```
Select TOP 10 plate, mjd, fiberid, z
From SpecPhoto
Where
  petroMag_R < 14
AND
  petroMagErr_R < 0.1
ORDER BY -z
```

(but too simple for real use! See the practical class!)

Ordering & Groupings

Counting objects in bins:

```
SELECT .2*(.5+floor(g.r/.2)) as mag,
       count(*) as num
FROM GALAXY
GROUP BY .2*(.5+floor(g.r/.2))
ORDER BY mag
```

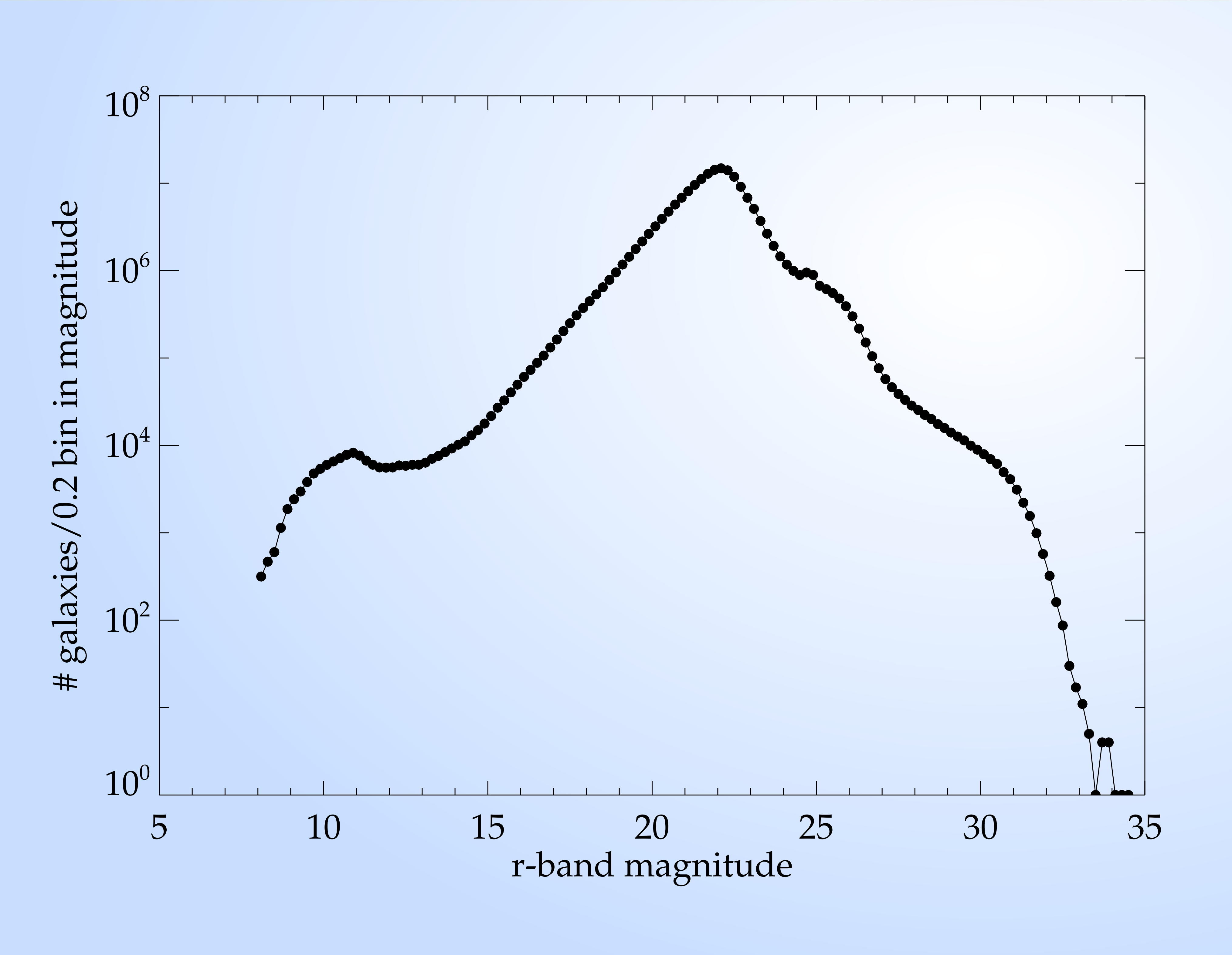
In general if you want to make C bins per unit for variable x, you use:

$$\frac{1}{C}\left(0.5 + \lfloor Cx \rfloor\right)$$

Ordering & Groupings

Cour

In g   u use:



Temporary tables - nesting

Sometimes when you want to make a query it is logical to divide it in two or more steps. This is particularly true in many cases when you need to combine one or more tables.

This is particularly common when you deal with disjoint/dissimilar selections. Say you wanted to find the fraction of galaxies that have u-g > 2 as a function of redshift:

```
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   modelMag_u-modelMag_g > 2
AND
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
ORDER BY redshift
```

Temporary tables - nesting

Say you wanted to find the fraction of galaxies that have u-g > 2 as a function of redshift:

```
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   modelMag_u-modelMag_g > 2
AND
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
ORDER BY redshift
```

Gets the distribution of the u-g> 2 galaxies

```
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
ORDER BY redshift
```

Gets the distribution of the all galaxies

Temporary tables - nesting

Say you wanted to find the fraction of galaxies that have u-g > 2 as a function of redshift:

```
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   modelMag_u-modelMag_g > 2
AND
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
ORDER BY redshift
```

Gets the distribution of the u-g> 2 galaxies

How do you combine them?

```
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
ORDER BY redshift
```

Gets the distribution of the all galaxies

Temporary tables - nesting

Say you wanted to find the fraction of galaxies that have u-g > 2 as a function of redshift:

```
SELECT red.redshift, cast(red.num AS float)/cast(b.num AS float)
as fraction
FROM  (
  Select floor(0.5+z*50.)/50 as redshift, count(*) as num
   From SpecPhoto
  Where
   modelMag_u-modelMag_g > 2
  AND
   z between 0.001 and 0.5
  GROUP BY floor(0.5+z*50.)/50
) as red JOIN
(
Select floor(0.5+z*50.)/50 as redshift, count(*) as num
From SpecPhoto
Where
   z between 0.001 and 0.5
GROUP BY floor(0.5+z*50.)/50
) AS b ON b.redshift = red.redshift

ORDER BY red.redshift
```

Example of the NULL issue

Calculating the Tully-Fisher relation (Luminosity versus Rotation velocity) from the Millennium simulation:

```
SELECT vVir, mag_b, mag_k
    FROM millimil..DeLucia2006a
WHERE (bulgeMass < 0.1*stellarMass OR
       bulgeMass IS NULL)
AND
    snapnum = 41
```

What would happen if we left this out?

We would miss out on all the galaxies without a bulge!

Matching stuff

Matching data

Two standard scenarios:

Joining tables:

The database contains several tables or views and we want to select objects in one table but also get some information about these objects/observations from another table.

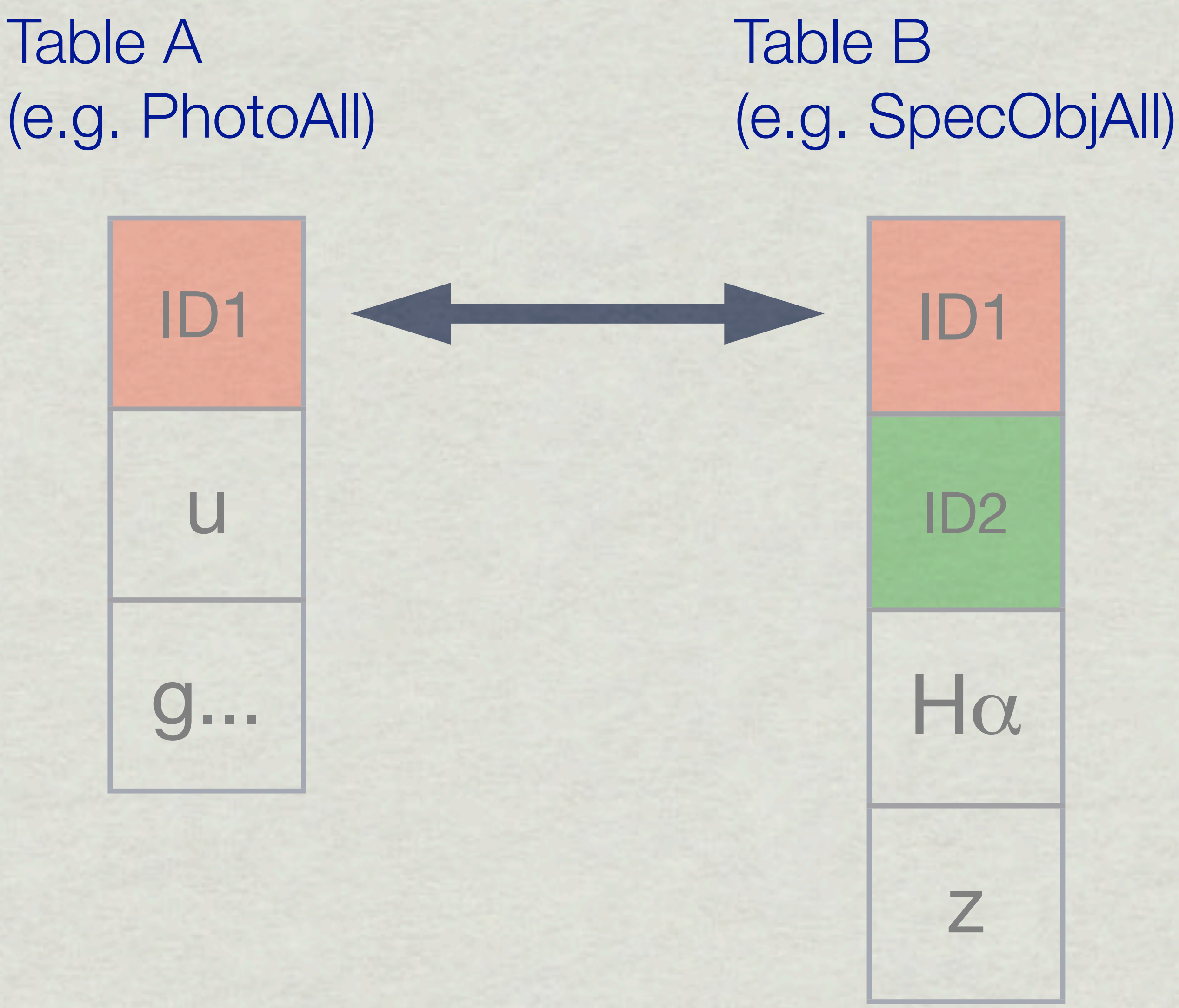
Database JOINs

Matching separate databases:

We might have observations from another database (e.g. the 2MASS survey) and want to match this to our main database (e.g. the SDSS)

More complex - what do we match on, what are the criteria to accept a match etc.

Joins - Matching tables up



The idea:

Get z from Table B and g from Table A for the same object.

SQL:

Need a quantity that is *common* between two tables. Here ID1.

Combining tables - JOINs

Photometric table

ID	1
u	19.3

ID	2
u	17.5

ID	3
u	20.5

Spectroscopic table

ID	1
EW(Ha)	75Å

ID	3
EW(Ha)	0.5Å

How do you combine them?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

How do you combine these?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

How do you combine these?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

Union: $\{x_i, y_j | i=1, n; j=1, m\}$ elements must be the same

How do you combine these?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

Union: $\{x_i, y_j | i=1, n; j=1, m\}$ elements must be the same

Cross-join: $\{(x_i, y_j) | i=1, n; j=1, m\}$ ie. all possible pairs

How do you combine these?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

Union: $\{x_i, y_j | i=1, n; j=1, m\}$ elements must be the same

Cross-join: $\{(x_i, y_j) | i=1, n; j=1, m\}$ ie. all possible pairs

Outer join: $\{(x_i, y_i)$ if y_i exists, (x_i, NULL) otherwise$\}$

How do you combine these?

First a theoretical view:

Two sets of values: $\{x_i\}$ $\{y_j\}$ (the elements can be vectors/matrices etc)

Possible ways to combine:

Union: $\{x_i, y_j | i=1, n; j=1, m\}$ elements must be the same

Cross-join: $\{(x_i, y_j) | i=1, n; j=1, m\}$ ie. all possible pairs

Outer join: $\{(x_i, y_i)$ if y_i exists, (x_i, NULL) otherwise$\}$

Inner join: $\{(x_i, y_i)$ if y_i exists$\}$

UNION

It must make sense to glue the tables together!

```
Select TOP 10
ra, dec
FROM SpecPhoto
WHERE ra > 120
AND DEC < 0
```

Table 1

```
UNION
```

```
Select TOP 10
ra, dec
From SpecPhoto
WHERE ra < 10
AND DEC > 0
```

Table 2

UNION

It must make sense to glue the tables together!

```
Select TOP 10
ra, dec
FROM SpecPhoto
WHERE ra > 120
AND DEC < 0
```

Table 1

Ra	Dec

```
UNION
```

```
Select TOP 10
ra, dec
From SpecPhoto
WHERE ra < 10
AND DEC > 0
```

Table 2

UNION

It must make sense to glue the tables together!

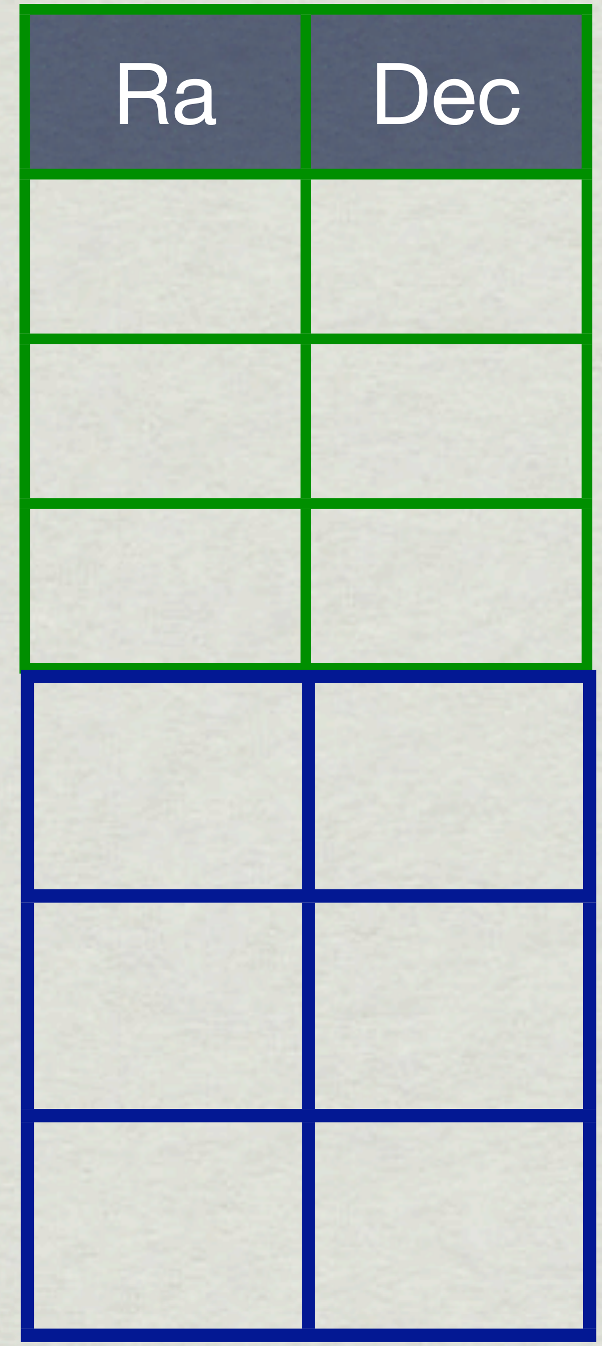
```
Select TOP 10
ra, dec
FROM SpecPhoto
WHERE ra > 120
AND DEC < 0

UNION

Select TOP 10
ra, dec
From SpecPhoto
WHERE ra < 10
AND DEC > 0
```

Table 1

Table 2



How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

```
SELECT P.u, S.z
FROM Photo as P
  JOIN  Spectro as S
ON P.ID=S.ID
```

How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

```
SELECT P.u, S.z
FROM Photo as P
  JOIN  Spectro as S
ON P.ID=S.ID
```

OR

```
SELECT P.u, S.z
FROM Photo as P,
Spectro as S
WHERE P.ID=S.ID
```

How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

```
SELECT P.u, S.z
FROM Photo as P
  JOIN  Spectro as S
ON P.ID=S.ID
```

Explicit INNER JOIN

OR

```
SELECT P.u, S.z
FROM Photo as P,
Spectro as S
WHERE P.ID=S.ID
```

Implicit INNER JOIN

(or *old-style* INNER JOIN)

Explicit vs implicit JOINs

JOIN ... ON a=b
or
WHERE a = b

Mostly up to you - there should be no significant difference between the two.

The main disadvantage of an implicit JOIN is that you have less control of the order things are done if you have more than two tables.

I personally prefer explicit JOINs because they show more clearly what your intention is and if you have a problem because your query runs too slowly, you can more easily figure out the execution order.

Inner join gets us those that are in common in both tables

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

Inner join gets us those that are in common in both tables

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

Inner join gets us those that are in common in both tables

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

But if we want to keep **all** possible pairs
we need an **OUTER JOIN**

Inner join gets us those that are in common in both tables

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

JOIN →

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

But if we want to keep **all** possible pairs
we need an **OUTER JOIN**

```
SELECT P.u, S.z
FROM Photo as P
LEFT OUTER JOIN  Spectro as S
ON P.ID=S.ID
```

How do you combine these?

ID	1
u	19.3
ID	2
u	17.5
ID	3
u	20.5

ID	1
EW(Ha)	75Å
ID	3
EW(Ha)	0.5Å

Join

ID	u	EW(Ha)
1	19.3	75Å
3	20.5	0.5Å

But if we want to keep all possible pairs
we need an **OUTER JOIN**

```
SELECT P.u, S.z
FROM Photo as P
LEFT OUTER JOIN  Spectro as S
ON P.ID=S.ID
```

ID	u	EW(Ha)
1	19.3	75Å
2	17.5	NULL
3	20.5	0.5Å

Looking at an SDSS Join

Find galaxies with EW(Ha) > 40Å

```
SELECT G.ObjID                      -- we want the photometric ObjID

FROM Galaxy as G,
     SpecObj as S,
     SpecLine as L

WHERE G.ObjID = S.bestObjID  -- the spectroscopic object should
                             -- be (photometrically) a galaxy
and S.SpecObjID = L.SpecObjID -- and spectral line L is
                                 detected in spectrum
and L.LineId = 6565          -- and line L is the H alpha line
and L.ew > 40                -- and H alpha is at least
                             -- 40 angstroms wide
```

Looking at an SDSS Outer Join

Find stars with & without spectra in a range:

```
SELECT TOP 20 S.ra, S.dec, S.objID, S.specObjID
FROM Star as S
   LEFT OUTER JOIN
     SpecObj as Sp  ON S.ObjID = Sp.bestObjID

WHERE S.ra > 180 AND S.ra < 181 AND
 abs(S.dec) < 1
```

Matching data - speeding it up

How do you find stuff quickly?

Look-up tables/indices, clever organisation

- Unique IDs (all members of database have this)
- Sorting the data
- Maybe duplicate data to allow for several sortings

Don't create huge intermediate tables

Matching data - speeding it up

Don't create huge intermediate tables

E.g.: Table P contains the phone numbers of every person in the world
Table H contains the height of every person in Leiden

And we want the phone number of every person in Leiden that is above 2.00m tall (a common problem....)

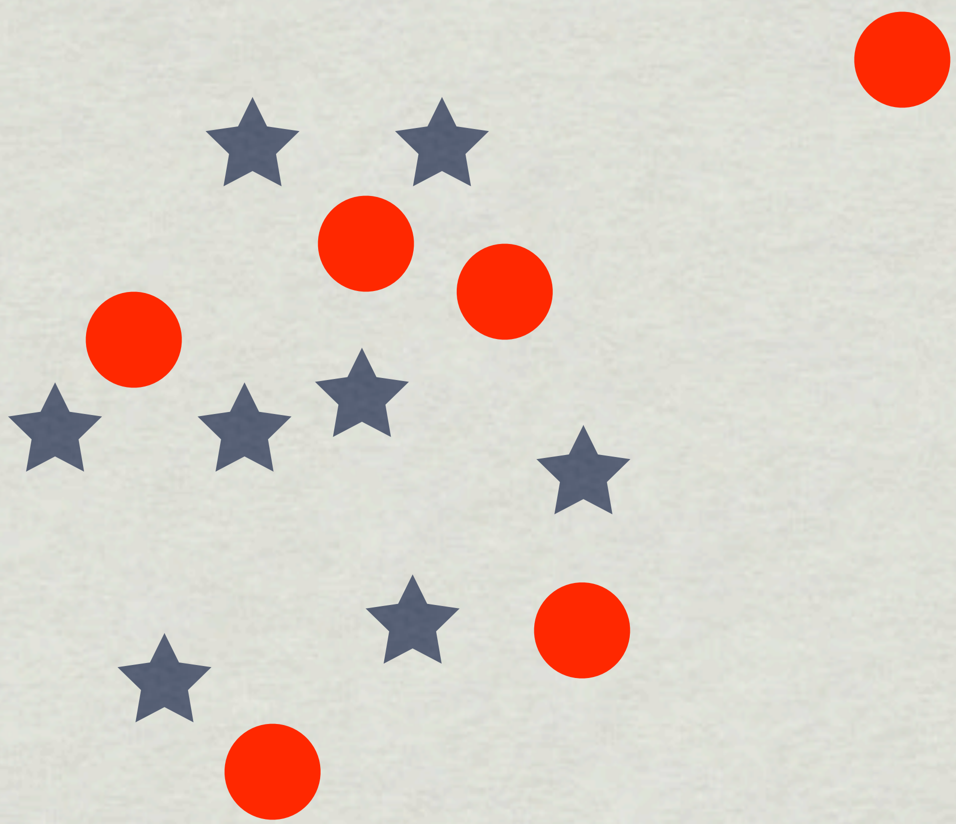
```
  P
LEFT OUTER JOIN
  H
ON P.NAME = H.NAME
WHERE
  H.Height > 2.0
```

```
  P
JOIN
  H
ON P.NAME = H.NAME
WHERE
  H.Height > 2.0
```

Usually a lot less obvious - but important to watch out for!

But what about in space?

- Table A
- Table B



For N & M objects we would naively need to calculate NxM distances - time-consuming for large databases.

So what can be done?

But what about in space?

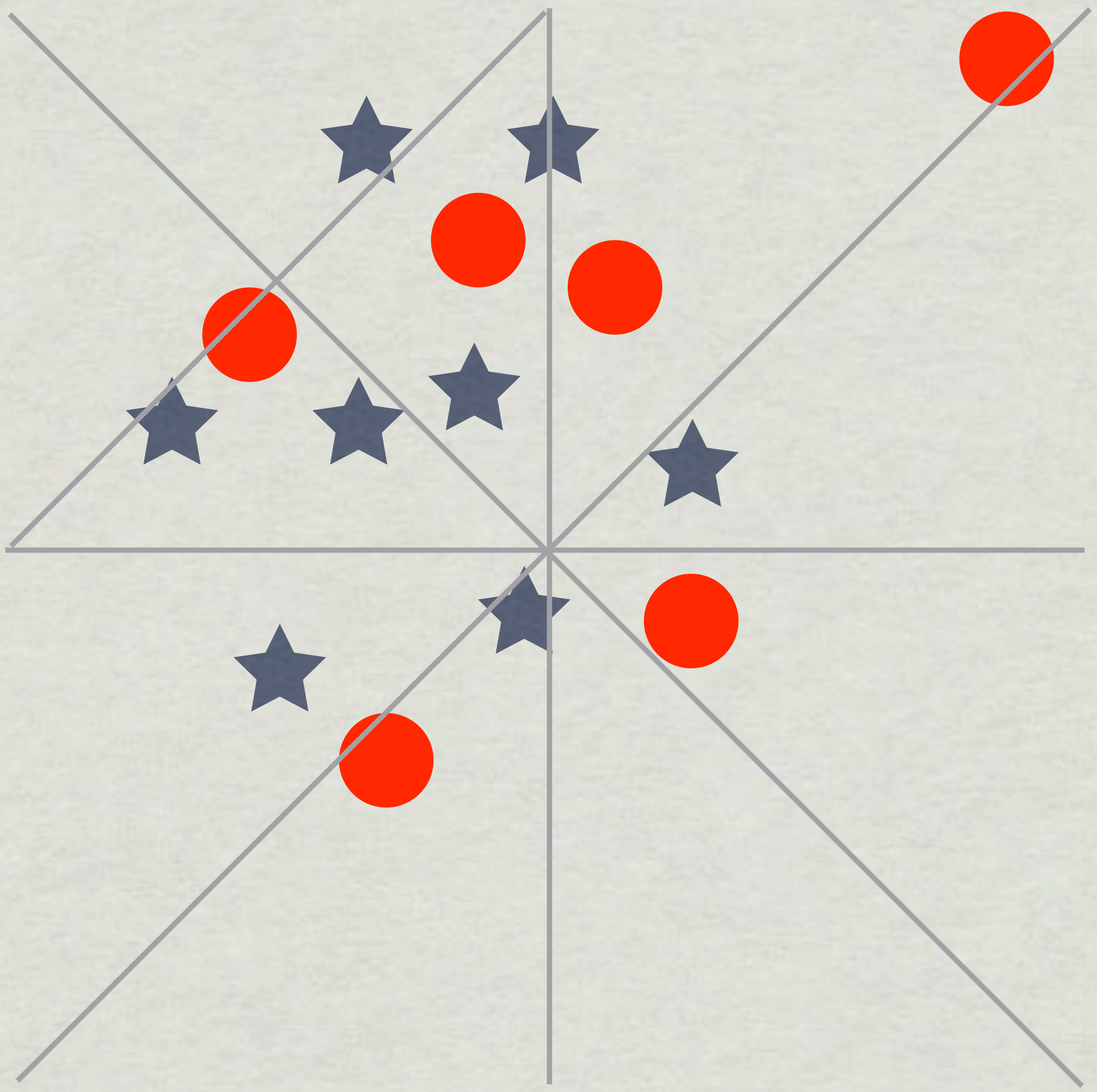
For N & M objects we would naively need to calculate NxM distances - time-consuming for large databases.

So what can be done?

A grid is one possibility (used in the Millennium database for instance). Naming becomes important so that closeness can be defined. Well-suited to Euclidean spaces

But what about in space?

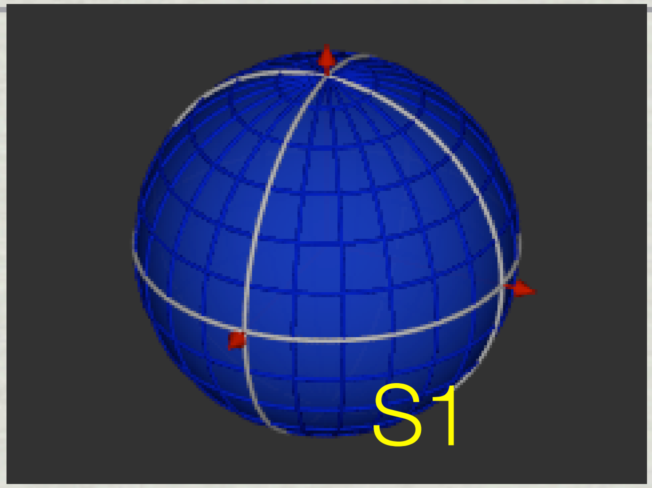
- Table A
- Table B



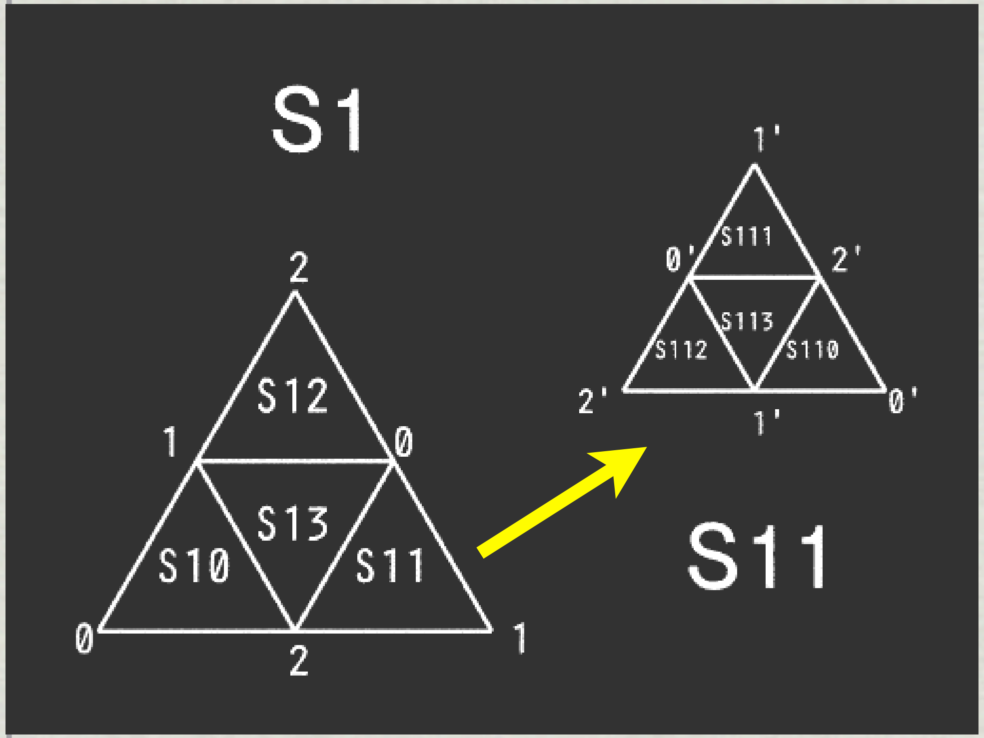
For N & M objects we would naively need to calculate NxM distances - time-consuming for large databases.

So what can be done?

Triangles is another possibility - this is used by the SDSS & quite a few other databases in what is called Hierarchical Triangular Mesh (HTM)



The sphere can be divided into 8 triangular regions. By starting on the midpoint of each side of a triangle you can subdivide it and create new triangles & then continue.



Triangles allow for a compact & (relatively) easy to understand naming scheme. The "name" is known as the HTMId.

Since subdivisions keep the "name" of their parents it is easy to find regions that are close to a given one.

Do I need to know?

- It is useful as a technique in many programming situations.
- Finding objects that are close together in one way or another (don't just think about x, y, z or ra & dec!) is very often important.
- Usually you can use database functions to do the dirty work but you need to know what **not** to do.
- But, you will rarely need to know the details unless you need to analyse neighbourhoods and correlation functions.

Matching objects - not so simple?

The standard techniques described above are ok for matching in simple situations but the world/sky is not always simple:

- What about moving objects? Asteroids, high proper motion stars?
- Many satellites have very different resolution. E.g. a single dish radio telescope might have a beam that includes many SDSS sources. The IRAS survey also had a very big beam. How do you find the right match?
- Gamma-ray bursts are often located with very poor positional accuracy - how do you know what to look at?

We can improve on the standard techniques with more sophisticated techniques incorporating statistical modelling and physical knowledge.

Doing science with databases

Some examples

Focus on big databases - the SDSS in particular

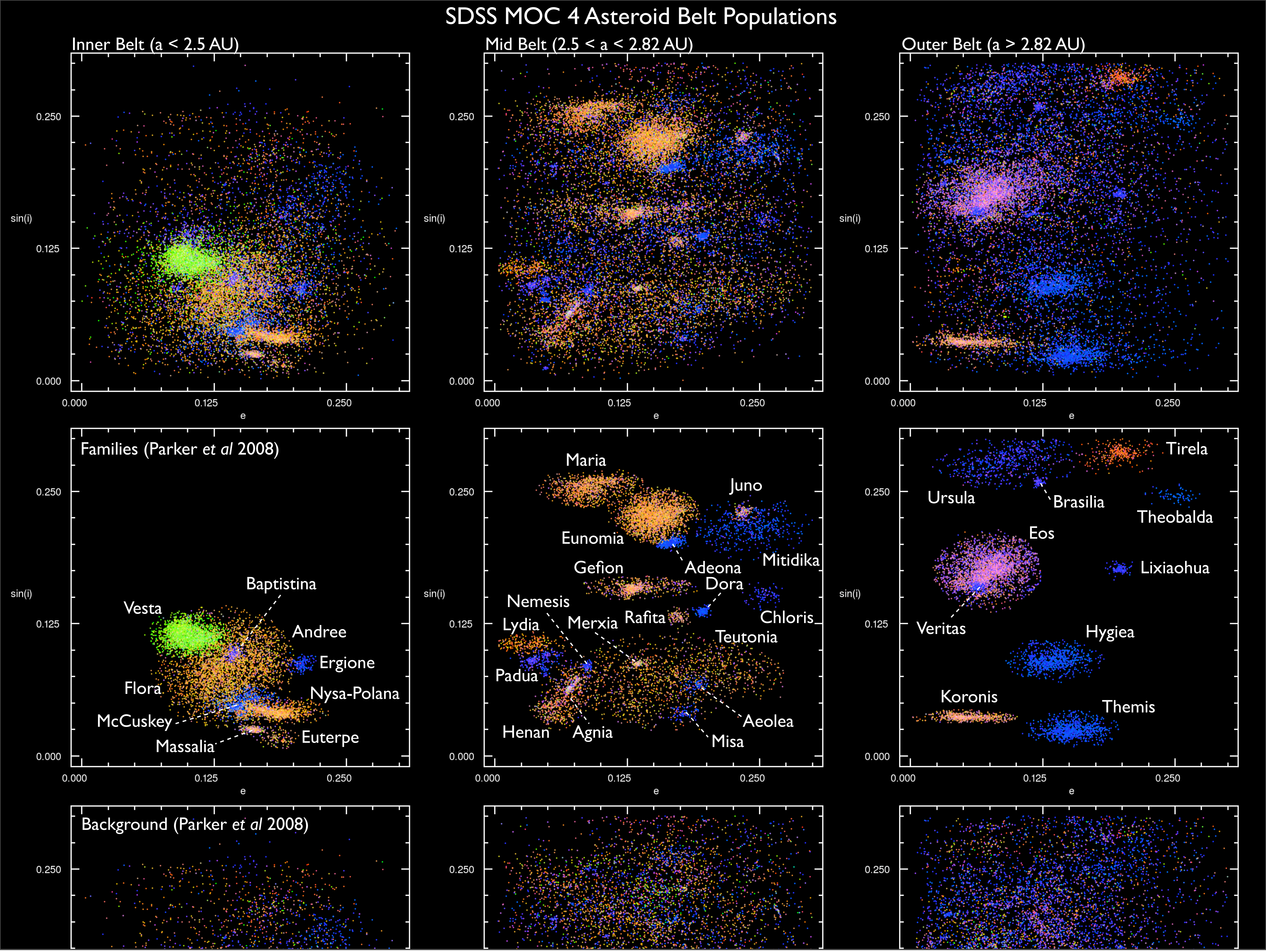
Finding asteroids

e.g. Ivezic et al (2001); Juric et al (2002); Parker et al (2008)

and your Werkcollege...

Any survey that take exposures with some separation in time will have a chance to find moving objects. The SDSS used a drift-scan strategy and the typical time between each filter is ~72s so a total time-arm of about 5 minutes.

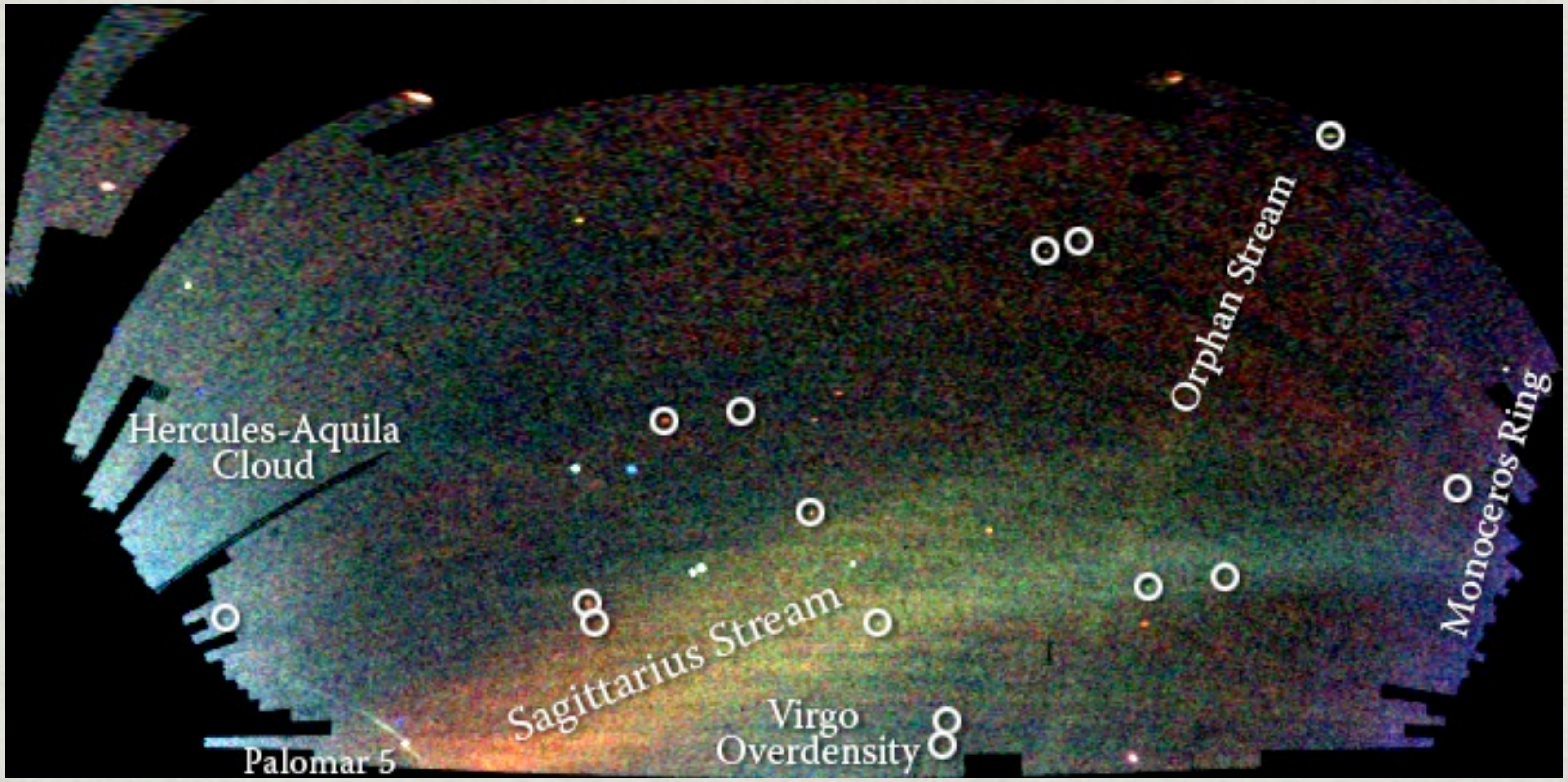
Can then use CasJobs to write an SQL survey to find these moving objects (see homework), this can be matched to known asteroids and orbital families can be found - Parker et al (2008) found 50.



Wednesday, 3 March 2010

Stars & satellites of the MW

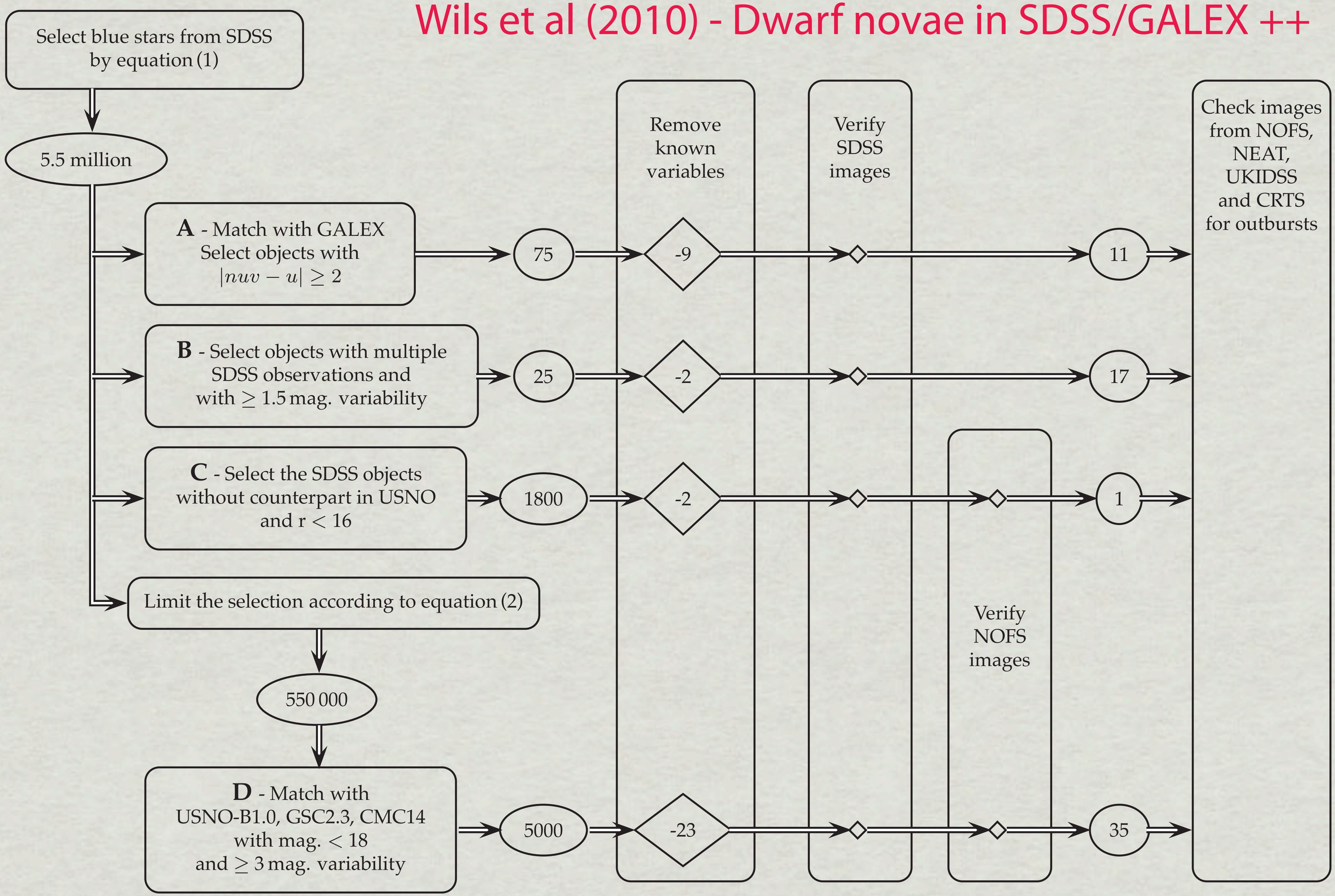
By selecting stars as a function of magnitude, position and colour you can find structure in the Milky-Way:



This is from Belokurov et al (2006) - the colour indicate distance from us and the brightness the density of sources. This can be used to find streams in the halo of the Galaxy as well as discover new nearby satellites.

Dwarf novae, cataclysmic variables and other variable objects

Wils et al (2010) - Dwarf novae in SDSS/GALEX ++



Future work:

A time-domain survey will provide:

Time-variability of positions: $\alpha(t)$, $\delta(t)$

Time-variability of fluxes: $f(t; \lambda)$

What can be done & what do we need to do it?

One lesson is what could be done with the SDSS - we have already seen the asteroids. But variable stars & moving stars are also possible.

Future work:

A time-domain survey will provide:

Time-variability of positions: $\alpha(t)$, $\delta(t)$

Time-variability of fluxes: $f(t; \lambda)$

What can be done & what do we need to do it?

Variable stars & bursts etc.

With something like LSST: Will have $\alpha(t)$, $\delta(t)$, $f(t; \lambda)$ sampled 1000 times, over 10 years for ~2×10^{10} objects.
Homework: Classify these data! Not easy!

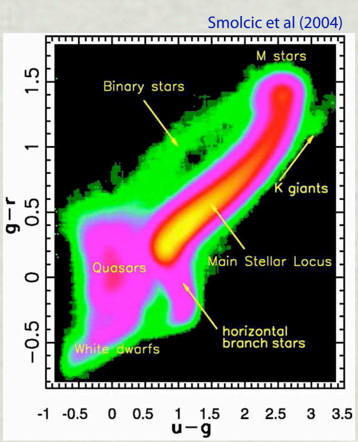
Future work:

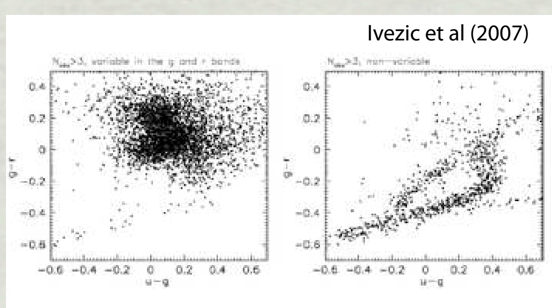
A time-domain survey will provide:

Time-variability of positions: $\alpha(t)$, $\delta(t)$

Time-variability of fluxes: $f(t; \lambda)$

What can be done & what do we need to do it?





Variable stars & bursts etc.

With something like LSST: Will have $\alpha(t)$, $\delta(t)$, $f(t; \lambda)$ sampled 1000 times, over 10 years for ~2×10^{10} objects.
Homework: Classify these data! Not easy!