Lecture 5b: Data Mining

Peter Wheatley

Data archives

- Most astronomical data now available via archives
  - Raw data and high-level products usually available
  - Data reduction software often specific to individual missions/instruments

- MAST – Multimission Archive at STScI
  - HST and ultraviolet missions
  - http://archive.stsci.edu/index.html

- HEASARC - High Energy Astrophysics Science Archive Research Center
  - X-ray and gamma-ray missions
  - http://heasarc.gsfc.nasa.gov/

- ESO Science Archive Facility
  - All ESO instruments
  - http://archive.eso.org/eso/eso_archive_main.html

- And many others…

FAQ

High-Level Science Products ▶

Software ▶

FITS ▶

Archive Manual

Related Sites

NASA Datacenters ▶

MAST Services

MAST and the VO ▶

Newsletters & Reports ▶

Data Use Policy

Dataset Identifiers

Acknowledgments

The Multimission Archive at STScI is a NASA funded project to support and provide to the astronomical community a variety of astronomical data archives, with the primary focus on scientifically related data sets in the optical, ultraviolet, and near-infrared parts of the spectrum.

Search MAST for a Target or Mission

Enter Target name (or Coordinates): `hd189733`

Resolver: ⦿ SIMBAD ◯ NED ◯ Don't Resolve

and/or Band/Data Type(s):   more options

	Extreme UV	Far UV	Near UV	Optical	Near IR	Radio
Images	☐	☐	☐	☐	☐	☐
Spectra	☐	☐	☐	☐	☐	
Other		☐	☐	☐	☐	

[Search]   [Reset]   **Help**

Google™ [   ] [Google Search]
◯ WWW ⦿ MAST

Note: STScI may provide links to Web pages that are not part of the STScI, AURA, NASA, or ESA domain. These sites are managed by organizations, companies or individuals not under our control, and neither STScI, AURA, NASA, nor ESA are responsible for the information or links you may find there. We provide these links as a convenience and the presence of these links is not an endorsement of the site.

NEWS

December 09, 2011:
WFC3 Persistence Project Releases Correction Files

December 08, 2011:
New "Condition Flag" available in Kepler Target Search Forms

December 07, 2011:
Kepler announces Accelerated Data Release schedule

December 03, 2011:
HST/NICMOS Paschen-alpha Survey of the Galactic Center High Level Science Products now available

November 28, 2011:
Enhanced Kepler Target Search Interface now available

RSS 2.0

Missions

Hubble
Hubble Legacy Archive
HSTonline
DSS
GALEX
JWST
KEPLER
XMM-OM
BEFS (ORFEUS)
Copernicus

National Aeronautics and Space Administration

Goddard Space Flight Center

Sciences and Exploration

NASA

GO   Search HEASARC website

[Advanced

HEASARC Quick Links

---Quick Links---

HEASARC Home | Observatories | Archive | Calibration | Software | Tools | Students/Teachers/Pu

Archive

HEASARC Browse

Tip
Archive

Hera

Other Browse interfaces:
Notification Service | Batch | Correlation | Index of all tables | Keyword Search

Query File And Session Uplo

Main Search Form > Search Results > Choose Data Products

[Start Search]   [Reset]   [Detailed Mission/Catalog Search]

1. Do you want to search around a position ... ?
(If you want to search on parameters other than object name or coordinates, select "Detailed Mission/Catalog Search".)

Object Name Or Coordinates: [   ]   **and/or**   **Select Local File:**   [Choose File] no file selected

e.g. **Cyg X-1** or **12 00 00, 4 12 6** or
Cyg X-2; 12.235, 15.345 (Note use of semi-colons
(;) to separate multiple object names or coordinate
pairs)

File should contain objects and/or coordinate pairs one
per line or separated by semi-colons.

Coordinate System: [J2000 ▾]

Search Radius: [Default   ]   [arcmin ▾]

Default uses the optimum radius for each catalog searched.

... and/or search by date?

Observation Dates: [   ]   YYYY-MM-DD hh:mm:ss or MJD: DDDDD.ddd

The time portion of the date is optional. Separate multiple dates/ranges with semicolons (;).
Range operator is '..'. (e.g. 1992-12-31; 48980.5; 1995-01-15 12:00:00; 1997-03-20 .. 2000-10-18)

2. What missions and catalogs do you want to search? (Bold text indicates mission is active)

☐ **Most Requested Missions**

☐ **Chandra** [*CXC,CSC*]   ☐ **Fermi**   ☐ **ROSAT**   ☐ **RXTE**

☐ **Suzaku**   ☐ **Swift**   ☐ **WMAP**   ☐ **XMM-Newton** [*XSA*]

☐ **Other X-Ray and EUV Missions**

☐ Ariel V   ☐ ASCA   ☐ BeppoSAX   ☐ BBXRT/Astro-1

☐ Copernicus   ☐ Einstein   ☐ EUVE   ☐ EXOSAT

ESO Archive Query Form

If you would like to query the Archive for instrument specific parameters, please use the dedicated query forms.
To search for **reduced Data Products**, please have a look at the ESO Data Products page and the Advanced Data Products query form.

The checkboxes on the right of the parameters define whether or not they will be displayed on the query result page.

SEARCH   ShowAll   ShowNone   Reset

query Help | Status of Requests

Target, Program and Scheduling Information

Target Name ☑ [   ] [Resolved by SIMBAD ⇅]

RA [   ] **DEC** [   ] *J2000*

Search Box [00 10 00] **Input** [RA(h) DEC(deg) ⇅]

Output ☑ [Sexagesimal (h, deg) ⇅]

List of Targets (Choose File) no file selected

Night ☐ [   ] (DD MM YYYY)

OR give a query range using the following start/end dates:

Start [   ] [12 hrs [UT] ⇅] **End** [   ] [12 hrs [UT] ⇅]

Program ID ☑ [   ] **Program Type** ☐ [Any ⇅]

PI CoI ☐ [   ] **SV** ☐ [Any ⇅]

Title ☐ [   ]

Observing Information

Imaging	Spectroscopy	Interferometry	Other
ALL NONE	ALL NONE	ALL	ALL NONE
☐VLT/FORS1	☐VLT/CRIRES	NONE	☐APEX/HET
☐VLT/FORS2	☐VLT/FORS1	☐VLTI/VINCI	☐APEX/BOL
☐VLT/HAWKI	☐VLT/FORS2	☐VLTI/MIDI	☐
☐VLT/ISAAC	☐VLT/GIRAFFE	☐	UKIRT/WFCAM
☐VLT/NACO	☐VLT/ISAAC	VLTI/AMBER	☐LGSF
☐VLT/VIMOS	☐VLT/NACO		☐MASCOT
☐VLT/VISIR	☐VLT/SINFONI	**Polarimetry**	
☐VISTA/VIRCAM	☐VLT/UVES	ALL	
☐	☐VLT/VIMOS	NONE	
VST/OMEGACAM	☐VLT/VISIR	☐VLT/FORS1	

Data Product Info

Type ☑ [Any ⇅]
User defined input: [   ]

Mode ☑ [Any ⇅]
User defined input: [   ]

Dataset ID ☑ [   ]

Orig Name ☐ [   ]

Release Date ☑ [   ]

OB Name ☐ [   ]

Category ☑

☐ SCIENCE
☐ CALIB

Data file formats

- ASCII – text files
  - Commonly used for small / simple datasets

- FITS - Flexible Image Transport System
  - The most common format for astronomical data
  - Ascii headers and binary data storage
  - Designed for images, but extended for 2d (and even 3d) tables

- HDS - Hierarchical Data System
  - UK Starlink standard, still used by many starlink tools
  - Highly versatile, with arbitrarily complex data structures
  - No longer widely used

- VOTable – virtual observatory table
  - XML standard for astronomical data tables
  - Designed for easy exchange of data for Virtual Observatory tools
  - Software can *understand* rather than just *read* (e.g. defn of exposure time)

Accessing FITS files

- cfitsio - set of sub-routines for reading/writing fits files
  - Often built into scripting languages, e.g. python
- FV- Useful tool for viewing contents of FITS files
  - Part of the HEASOFT software, or available stand-alone
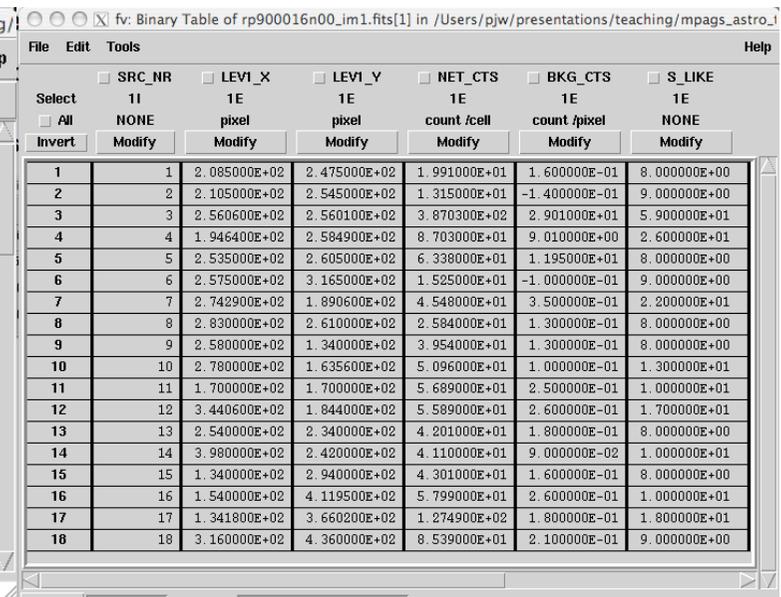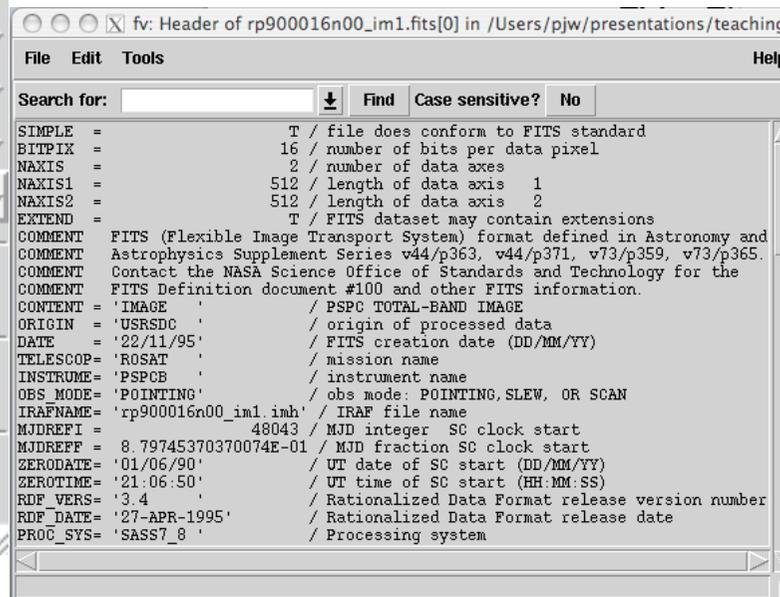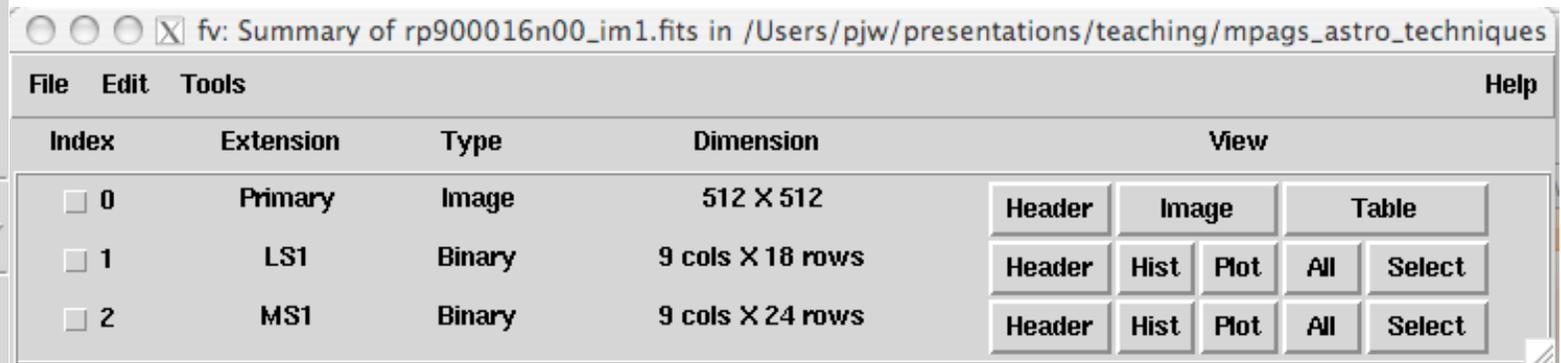
Astronomical catalogues

- CDS Simbad – data and bibliography for objects mentioned in papers

CDS — CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

| Simbad | VizieR | Aladin | Catalogs | Dictionary | Biblio | Tutorials | Resources |

SIMBAD query result
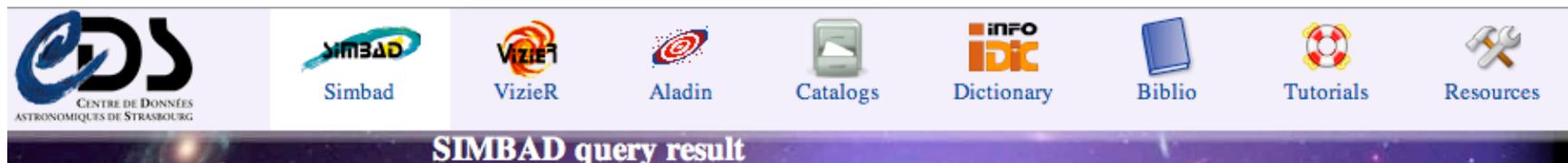
| *other query modes :* | Identifier query | Coordinate query | Criteria query | Reference query | Basic query | Script submission | Output options | Help |

Object query : vega

C.D.S. - SIMBAD4 rel 1.186 - 2011.12.16CET08:14:51

Available data : Basic data • Identifiers • Plot & images • Bibliography • Measurements • External archives • Notes • Annotations

Basic data :

NAME VEGA -- Variable Star

query around with radius 2 arcmin

Other object types:	*
	(*,AG,ASCC,BD,CSI,FK5,GAT,GC,GCRV,GEN#,GJ,HD,HGAM,HIC,HIP,HR,JP11,N30,8pc,PLX,PMC,F
	,** (ADS,CCDM,IDS,WDS) ,IR (IRAS,IRC,2MASS,RAFGL) ,PM* (LSPM,LTT,NLTT) ,V* (V*,CSV,
ICRS coord. (ep=J2000) :	18 36 56.33635 +38 47 01.2802 (Optical) [3.51 2.81 90] A 2007A&A...474..653V
FK5 coord. (ep=J2000 eq=2000) :	18 36 56.336 +38 47 01.28 (Optical) [3.51 2.81 90] A 2007A&A...474..653V
FK4 coord. (ep=B1950 eq=1950) :	18 35 14.67 +38 44 09.8 (Optical) [20.31 16.25 90] A 2007A&A...474..653V
Gal coord. (ep=J2000) :	067.4482 +19.2373 (Optical) [3.51 2.81 90] A 2007A&A...474..653V
Proper motions *mas/yr* [error ellipse]:	200.94 286.23 [0.32 0.40 0] A 2007A&A...474..653V
Radial velocity / Redshift / cz :	V(km/s) -13.9 [0.9] / z(~) -0.000046 [0.000003] / cz -13.90 [0.90] (~) A 1979IAUS
Parallaxes *mas*:	130.23 [0.36] A 2007A&A...474..653V
Spectral type:	A0V C ~
Fluxes (7) :	B 0.03 [~] C ~
	V 0.03 [~] C ~

Astronomical catalogues

- CDS Vizier –inc. sky survey catalogues and tables from papers

Image and catalogue display

- ds9
  - Image display and comparison
  - Image and catalogue download
  - Catalogue overlay
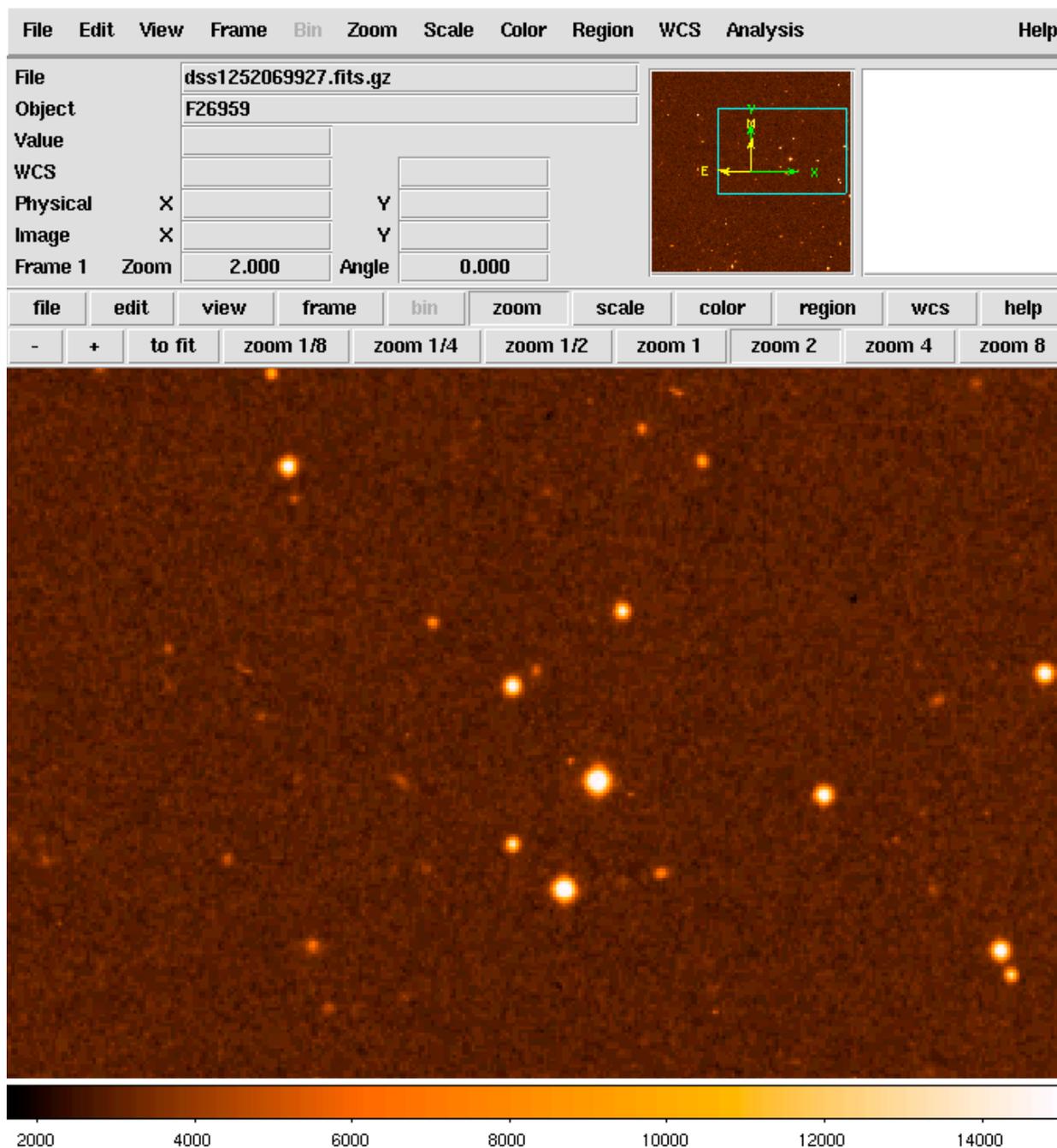
Image and catalogue display

- Gaia
  - Starlink software
  - Includes photometry and astrometry software
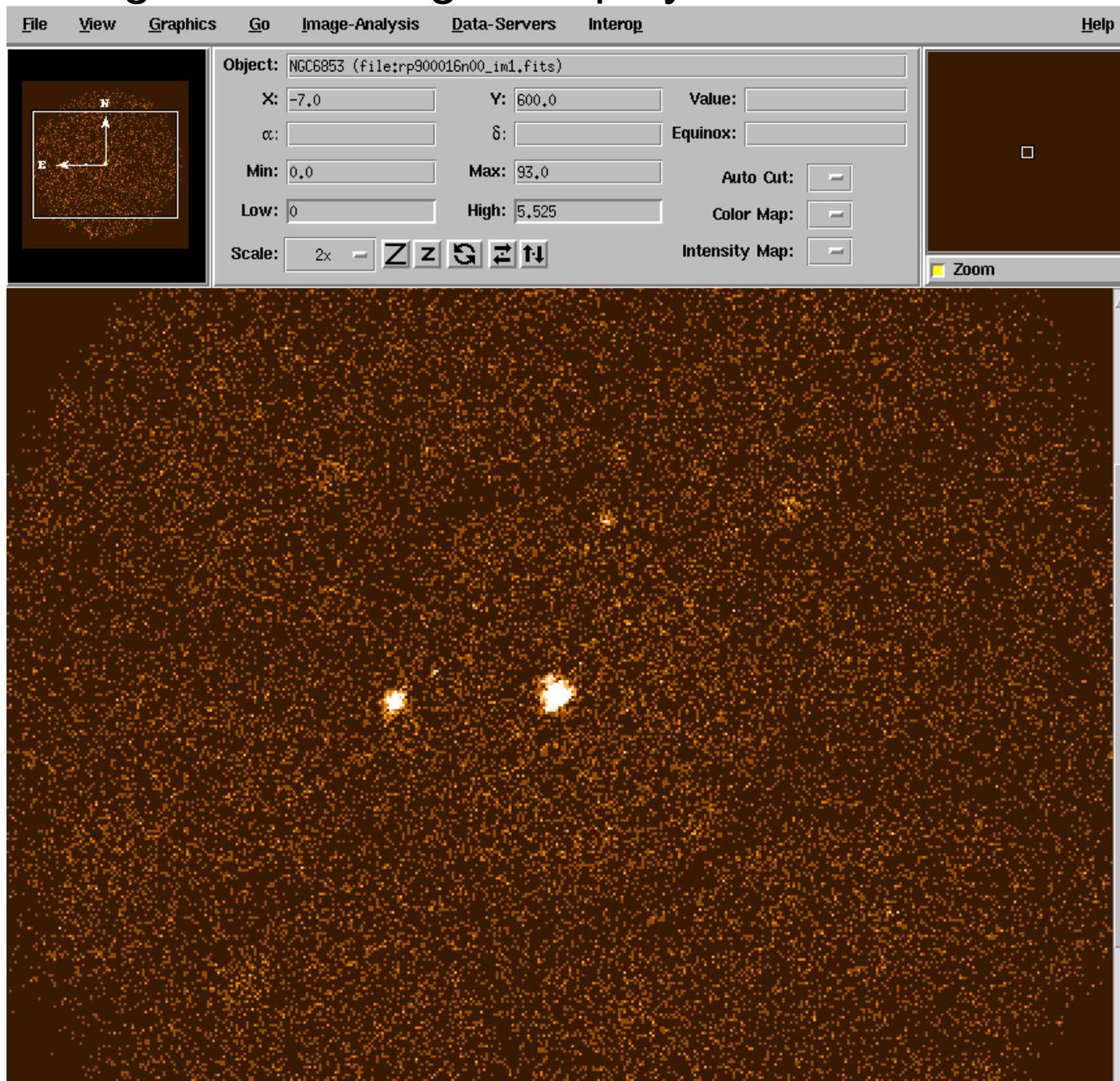
Lecture 5 Assignment, Part 1 of 2

- Use SIMBAD to determine the nature of the star HD189733 and any companion objects
- Download from HEASARC a ROSAT PSPC image of the star HD189733
- Display the ROSAT image aligned with an optical image of the same field (downloaded from within ds9 or Gaia)
- Overlay SIMBAD sources (also downloaded within ds9/Gaia)
- Send us a screenshot
- Determine the name and nature of the famous nebula nearby

TOPCAT & STIL(TS)

TOPCAT Java GUI
http://www.star.bris.ac.uk/~mbt/topcat/

STIL library for dealing with tabular data
http://www.star.bristol.ac.uk/~mbt/stil

STILTS command-line tool
http://www.star.bristol.ac.uk/~mbt/stilts/

Key features:

- Import/export catalogs in various formats

- Cross-matching and merging

- Plotting and manipulating data

- Search and retrieve data from archives

THE UNIVERSITY OF
WARWICK

TOPCAT : VO retrieval

Table Data

add custom columns

THE UNIVERSITY OF
WARWICK

TOPCAT Example

3 tables loaded

scatter plot

sky matched pairs

table matching

THE UNIVERSITY OF WARWICK

TOPCAT
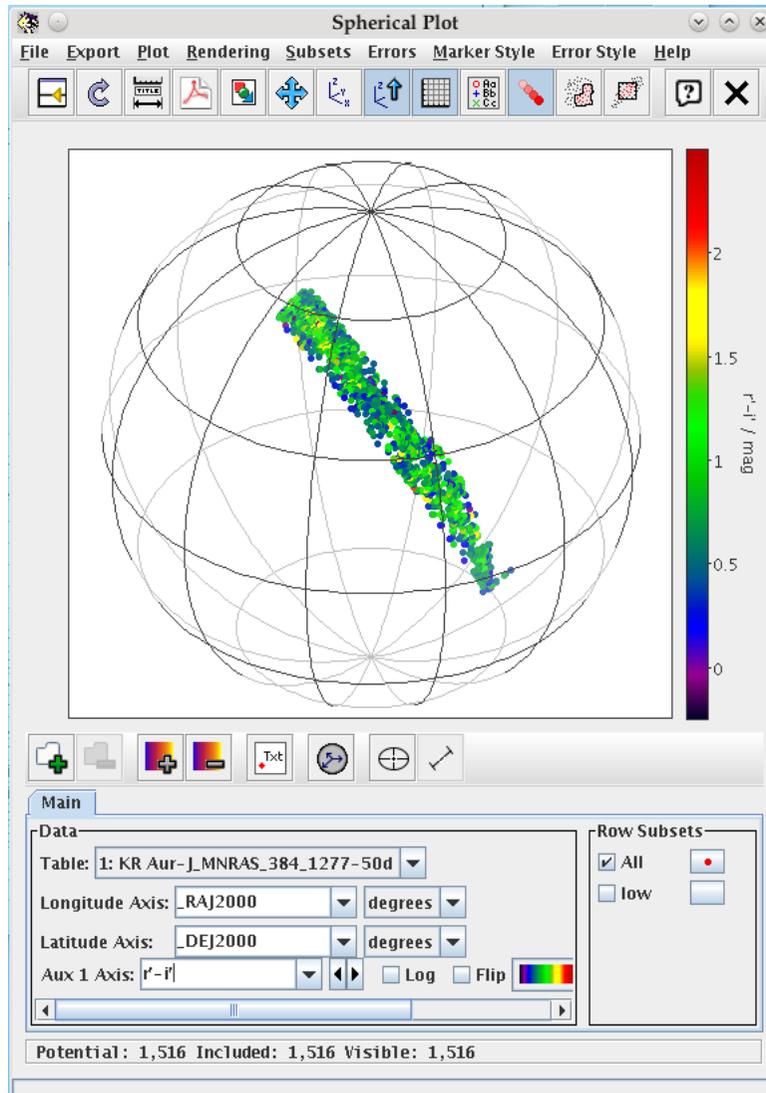
sky plot, colour coded   subsets

THE UNIVERSITY OF
WARWICK
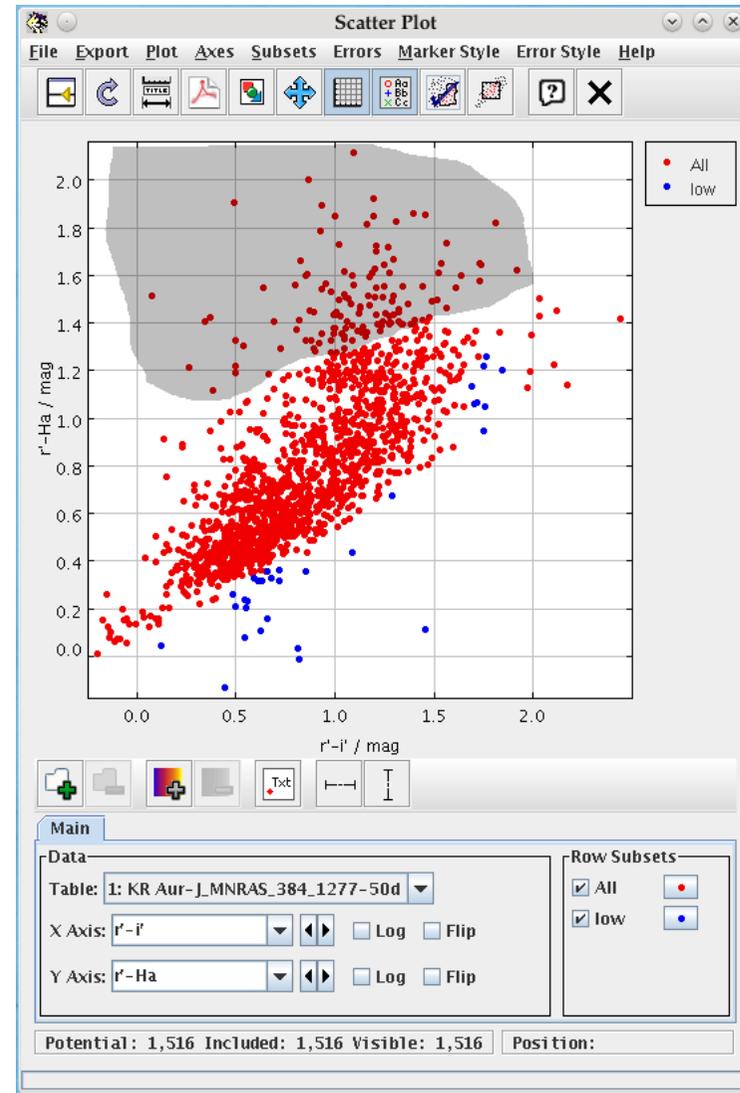
SQL Queries

- **SQL** refers to query syntax allowing you to search and interact with large databases stored by a SQL-compatible server

- The **S**tructured **Q**uery **L**anguage is very popular, originating outside of astronomy, and can handle very large databases plus many clients (commercial and open source implementations available)
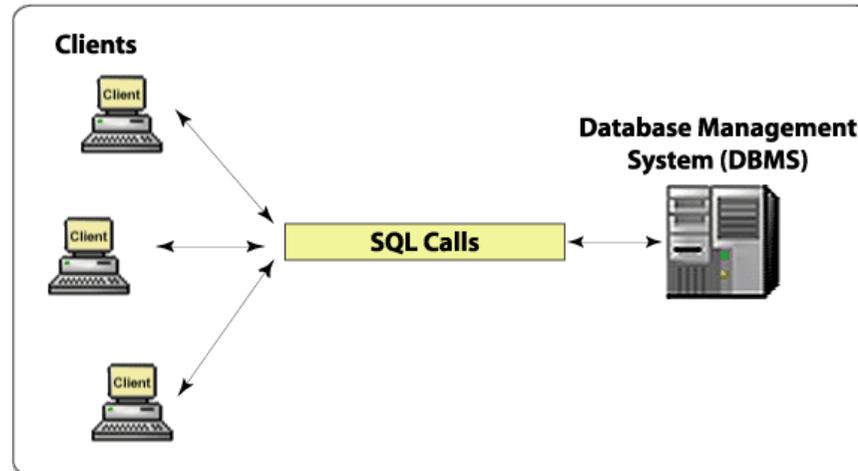
- Many astronomical databases powered by SQL even if you may query it by other means (a webform or VO query)

- When simpler frontends fail, or once tables involved become very large, you may need to resort to performing your own SQL queries

Database lingo

- Tables with rows and columns = Database with records and fields

ID	Source Class	Magnitude	Filter
1	1	12.34	1
2	3	16.89	2

- Records in separate tables may be connected to each other:

Class ID	Value
1	Star
2	Galaxy
3	Unknown

Relational database

Filter ID	Value
1	U
2	B
3	V
4	R

THE UNIVERSITY OF
WARWICK

SQL Queries : SDSS

- SDSS prime and early example of a large astronomical data resource that offered multiple ways of retrieving data, including custom SQL queries

- Tutorial covering performing SQL queries on SDSS data:

  `http://skyserver.sdss3.org/dr8/en/help/howto/search/`

Table browser

Schema Browser

Need to know what Tables are available and which fields are stored in those tables

Basic Query

```
SELECT
  ra,dec
FROM
  specObj
WHERE
  ra BETWEEN 140 and 141
AND
  dec BETWEEN 20 and 21
AND
  class='STAR'
```

which fields do you want returned?

from which table?

boolean constraints
on fields that records
must satisfy

THE UNIVERSITY OF
WARWICK

Relational Query

```
SELECT
  specObj.fiberID,
  PhotoObj.modelMag_u,
  PhotoObj.modelMag_g,
  PhotoObj.modelMag_r,
  PhotoObj.modelMag_i,
  PhotoObj.modelMag_z,
  PhotoObj.ra, PhotoObj.dec,
  specObj.z, PhotoObj.ObjID
FROM
  PhotoObj, specObj
WHERE
  specObj.bestObjid = PhotoObj.ObjID
AND
  specObj.class = 'qso'
AND
  specObj.zWarning = 0
AND
  specobj.z between 0.3 and 0.4
```

which fields do you want returned?
[Table.Field]

from which tables?

need to ensure we compare same object

boolean constraints on fields that records must satisfy
[Table.field = Value]

THE UNIVERSITY OF
WARWICK

Result

Your SQL command was:

```
SELECT
 specObj.fiberID, PhotoObj.modelMag_u, PhotoObj.modelMag_g,
 PhotoObj.modelMag_r, PhotoObj.modelMag_i, PhotoObj.modelMag_z,
 PhotoObj.ra, PhotoObj.dec, specObj.z, PhotoObj.ObjID
FROM
 PhotoObj, specObj
WHERE
 specObj.bestObjid = PhotoObj.ObjID
AND
 specObj.class = 'qso'
AND
 specObj.zWarning = 0
AND
 specobj.z between 0.3 and 0.4
```

fiberID	modelMag_u	modelMag_g	modelMag_r	modelMag_i	modelMag_z	ra	dec	z	ObjID
225	21.608122	20.919027	20.106173	19.625843	19.161301	215.97274128	-0.96643655	0.391129	1237648720172745195
640	20.437267	20.066856	19.869516	19.683071	19.237711	195.66483383	0.40463547	0.374446	1237648674510995921
221	14.929588	12.896034	12.344924	12.255587	13.574942	6.40105152	-0.46343099	0.383457	1237657190371164165
218	20.620367	19.742416	18.909544	18.444504	17.982805	6.55951741	-0.98912516	0.372924	1237657189834358887
512	17.007906	16.777861	16.785234	16.746496	16.370136	23.57580286	0.26019863	0.399601	1237657071157444676
189	18.119993	17.833719	17.874315	17.67757	17.250725	33.24829998	-0.50817531	0.394522	1237657070087962632
151	20.841267	20.093006	18.863449	18.384741	17.953033	337.42612345	-0.71003756	0.374328	1237656567586095748
192	19.786795	19.552595	19.321449	19.210716	18.773647	233.74375738	2.70391221	0.389854	1237654879671550196
332	19.542379	19.182779	18.850351	18.602491	18.349787	143.53824274	2.5436351	0.381111	1237654599949942970
513	18.413069	18.271866	18.284575	18.26753	17.632139	145.17885592	2.26594343	0.385768	1237653665256767497
85	20.295122	19.745787	19.058146	18.573439	18.002174	166.94072826	60.1539464	0.397834	1237653616398696678
42	20.057171	19.952911	19.487123	19.267277	18.884089	348.67848826	-10.51547783	0.377506	1237653436008759550
382	19.270554	19.143511	18.722548	18.43214	18.025404	49.6711526	-6.67914789	0.390237	1237652901306499242
601	20.580416	20.056355	19.482079	19.14917	18.849903	45.02323702	-7.20430752	0.374826	1237652901304467725
585	20.303259	19.847921	19.321499	19.023911	18.627033	28.93484124	-8.68499508	0.381934	1237652901297520731

Powerful Queries

```
SELECT TOP 10
run,camCol,rerun,field,objID,ra,dec
FROM Galaxy
WHERE ( ( flags & (dbo.fPhotoFlags('BINNED1')
| dbo.fPhotoFlags('BINNED2')
| dbo.fPhotoFlags('BINNED4')) ) > 0
and ( flags & (dbo.fPhotoFlags('BLENDED')
| dbo.fPhotoFlags('NODEBLEND')
| dbo.fPhotoFlags('CHILD')) ) != dbo.fPhotoFlags('BLENDED')
and ( flags & (dbo.fPhotoFlags('EDGE')
| dbo.fPhotoFlags('SATURATED')) ) = 0
and petroMag_i > 17.5
and (petroMag_r > 15.5 or petroR50_r > 2)
and (petroMag_r > 0 and g > 0 and r > 0 and i > 0)
and ( (petroMag_r-extinction_r) < 19.2
and (petroMag_r - extinction_r <
(13.1 + (7/3) * (dered_g - dered_r) + 4 * (dered_r - dered_i)
- 4 * 0.18) )
and ( (dered_r - dered_i - (dered_g - dered_r)/4 - 0.18) < 0.2)
and ( (dered_r - dered_i - (dered_g - dered_r)/4 - 0.18) > -0.2)
-- dered_ quantities already include reddening
and ( (petroMag_r - extinction_r +
2.5 * LOG10(2 * 3.1415 * petroR50_r * petroR50_r)) < 24.2) )
or ( (petroMag_r - extinction_r < 19.5)
and ( (dered_r - dered_i - (dered_g - dered_r)/4 - 0.18) > (0.45 - 4 *
(dered_g - dered_r)) )
and ( (dered_g - dered_r) > (1.35 + 0.25 * (dered_r - dered_i)) ) )
and ( (petroMag_r - extinction_r +
2.5 * LOG10(2 * 3.1415 * petroR50_r * petroR50_r) ) < 23.3 ) )
```

THE UNIVERSITY OF
WARWICK

CasJobs

- No limit on number of records returned or the time it takes for your Query to run

- You can also upload and create your own Tables, then join those in SQL queries
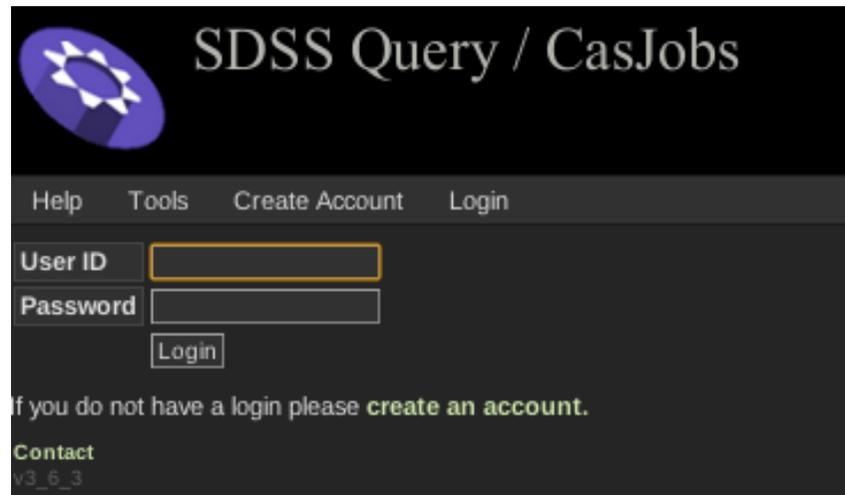
- Note: it is very easy to write a very complex Query so appreciate what you are asking from the server

- The SDSS/SQL model is used by others, such as GALEX, MAST, VISTA etc

THE UNIVERSITY OF
WARWICK

Other example ; UKIDSS @ WFAU

Lecture 5 Assignment (part 2 of 2)

- Using TOPCAT Retrieve a 5' cone of data from the SDSS DR7 photometry (`sdssdr7-sda` service, `PhotoObjAll` catalog) centered on 01:30:56, +43:55:49

- Merge the resulting table with 2MASS (`2mass-psc`) using a sky matching radius of 2"

- Make a scatter plot of SDSS (u-g) colour versus 2MASS (J-K) colour, limited to sources brighter than SDSS g=20

- Consider the source that is most red in the near-infrared and comment on its optical colour relative to other sources in the field