



Malte Marquarding
CASS, ASKAP
Tara Murphy
CAASTRO University of Sydney



Overview

- Introduction
- IT Toolkit
- Managing data
- Scripting
- Virtual Observatory (VO)
- Advanced VO

We are undergoing a data explosion

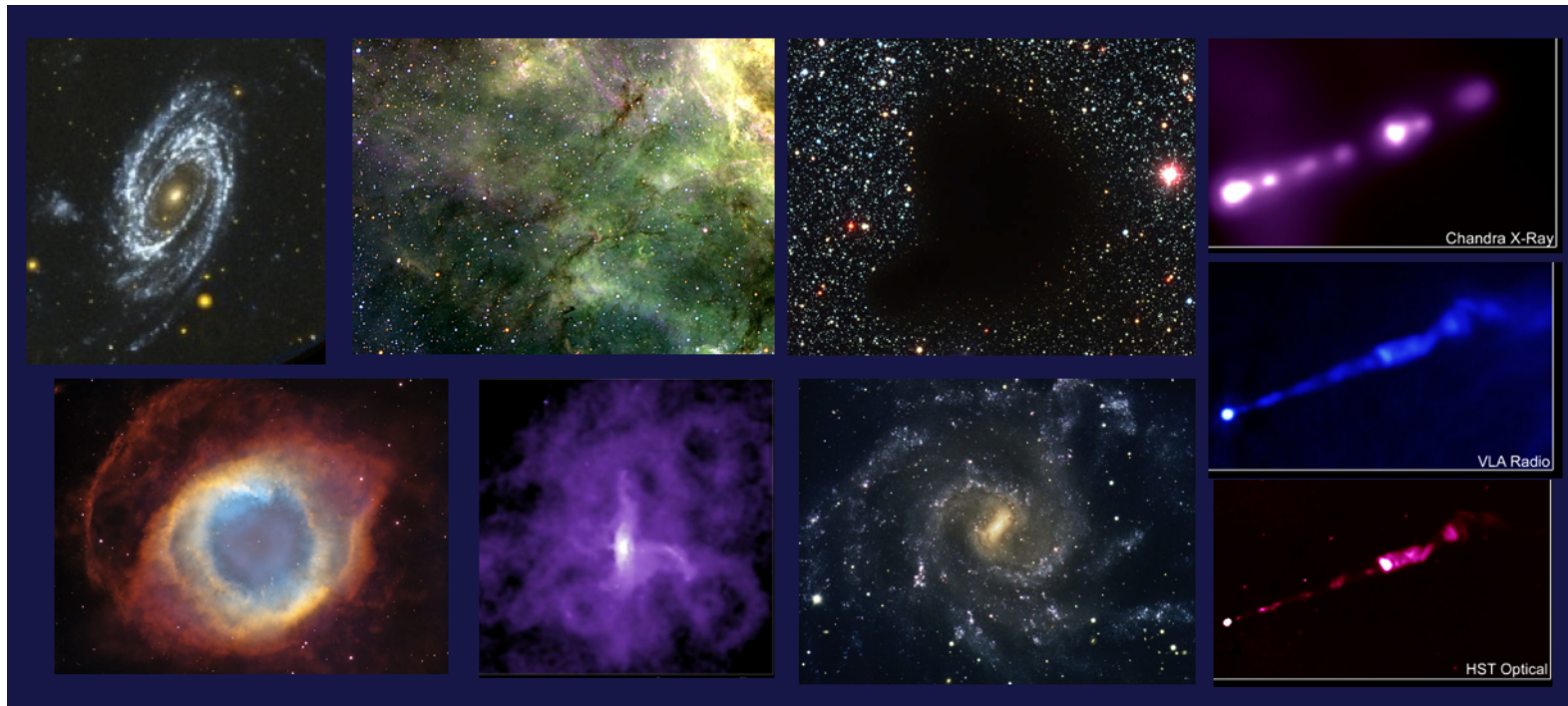
- We live in the age of the mega-survey
- Datasets are orders of magnitude larger and more complex than in the past
 - Surveys: SDSS, 2dF, 2MASS, DPOS, WMAP. . .
 - Digital libraries: ADS, astro-ph, NED, CDS
 - Observatory archives: HST, ATOA, MAST
 - Future examples: LSST, GALEX, ASKAP, SKA. . .

These surveys will produce terabytes per night

- For comparison:
 - The Library of Congress is ~ 20 TB
 - SDSS has publicly released ~ 10 TB
 - The Human Genome is < 10 GB

Datasets are becoming more complex

Multi-wavelength astronomy is critical



<http://www.educationgrid.org/msicii/June06Workshop/Presentations/astrocyberinfrastructure.ppt>

The way we do science is changing

- Massive datasets
- Multi-wavelength datasets
- New techniques
- Giant collaborations
- Use and reuse of archives
- A typical scientist will have to deal with more and more data:

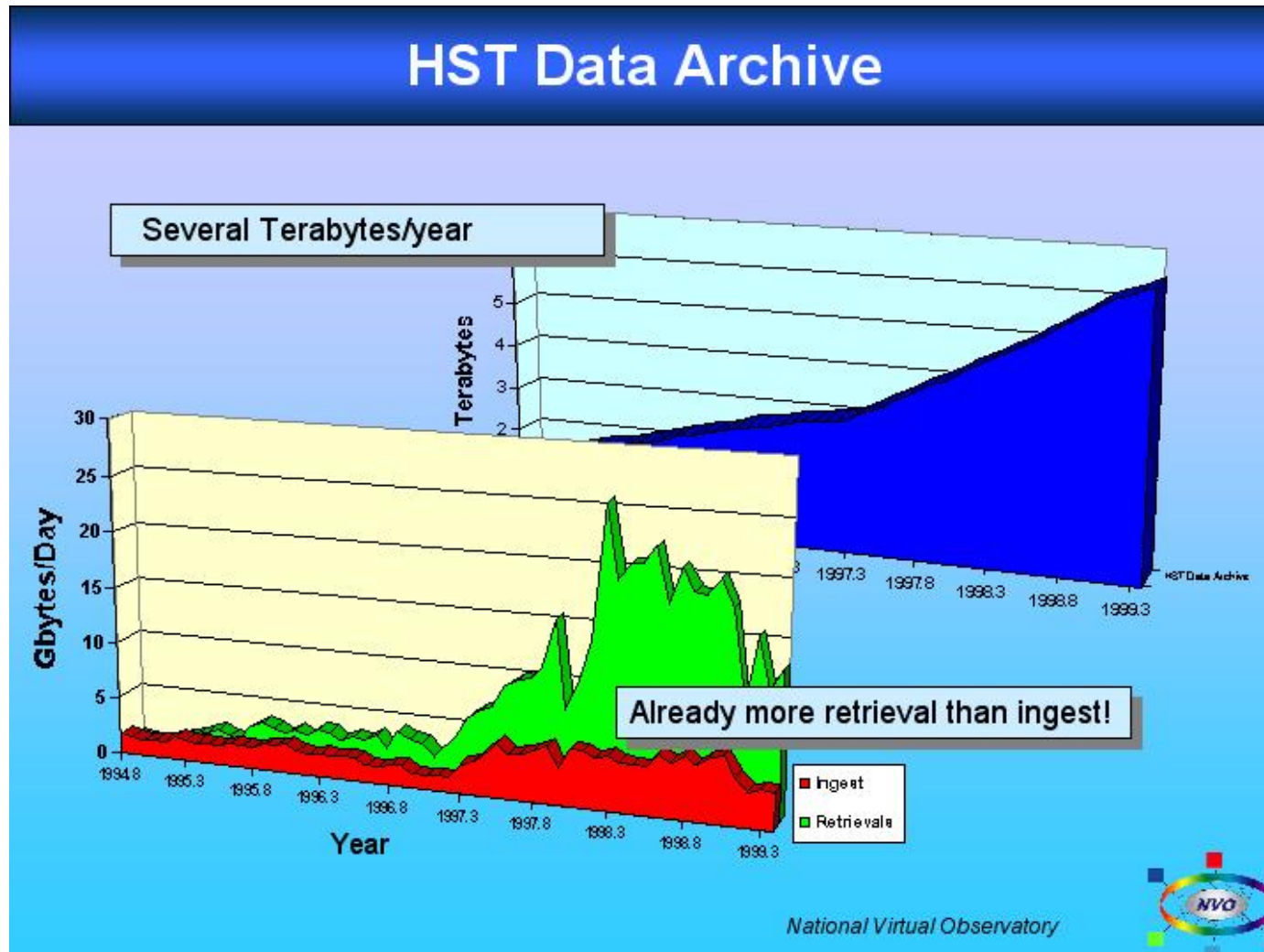
You can `grep` 1 MB in a second

You can `grep` 1 GB in a minute

You can `grep` 1 TB in 2 days

You can `grep` 1 PB in 3 YEARS

The way we do science is changing

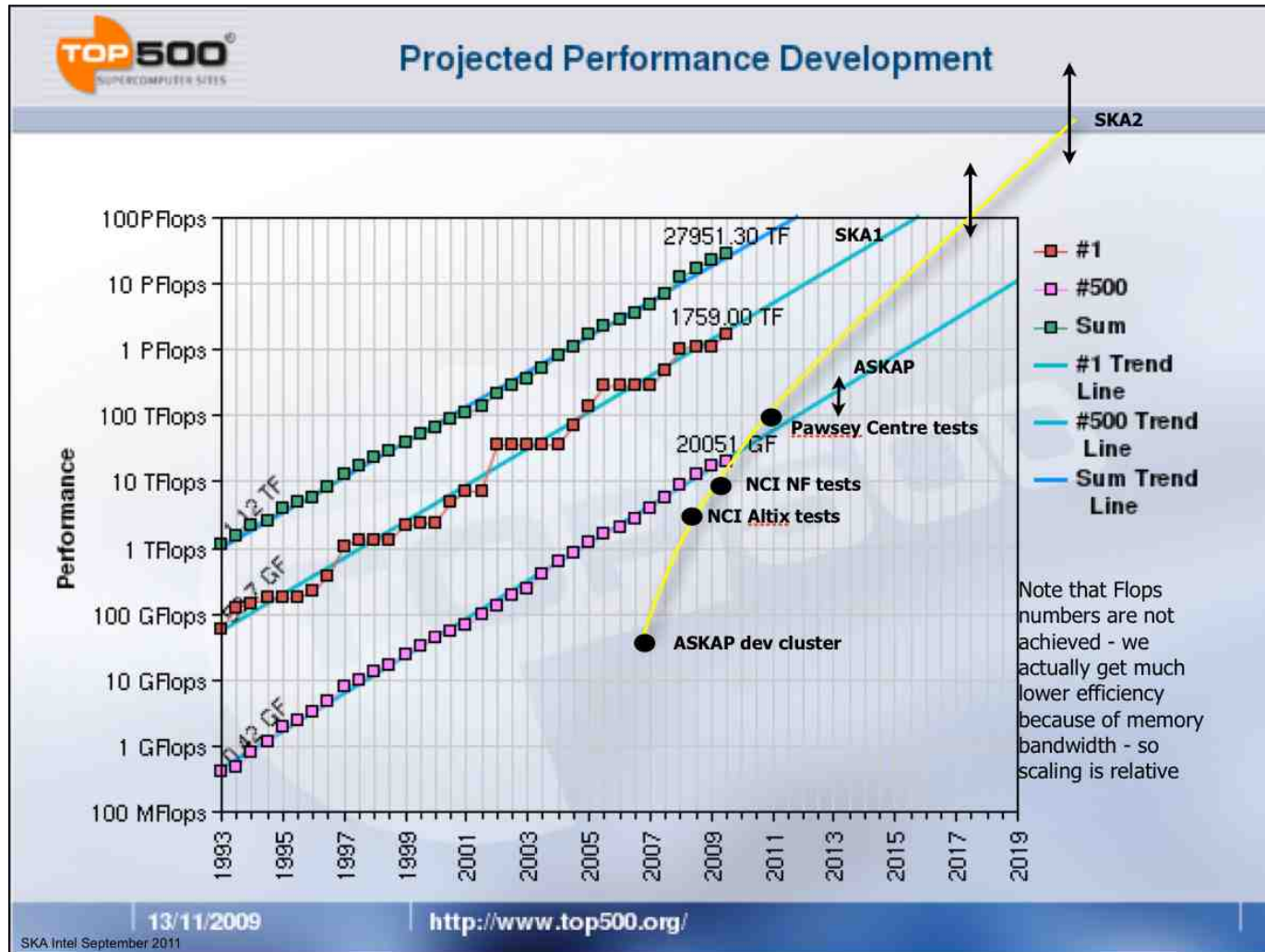


What does this mean for astronomy?

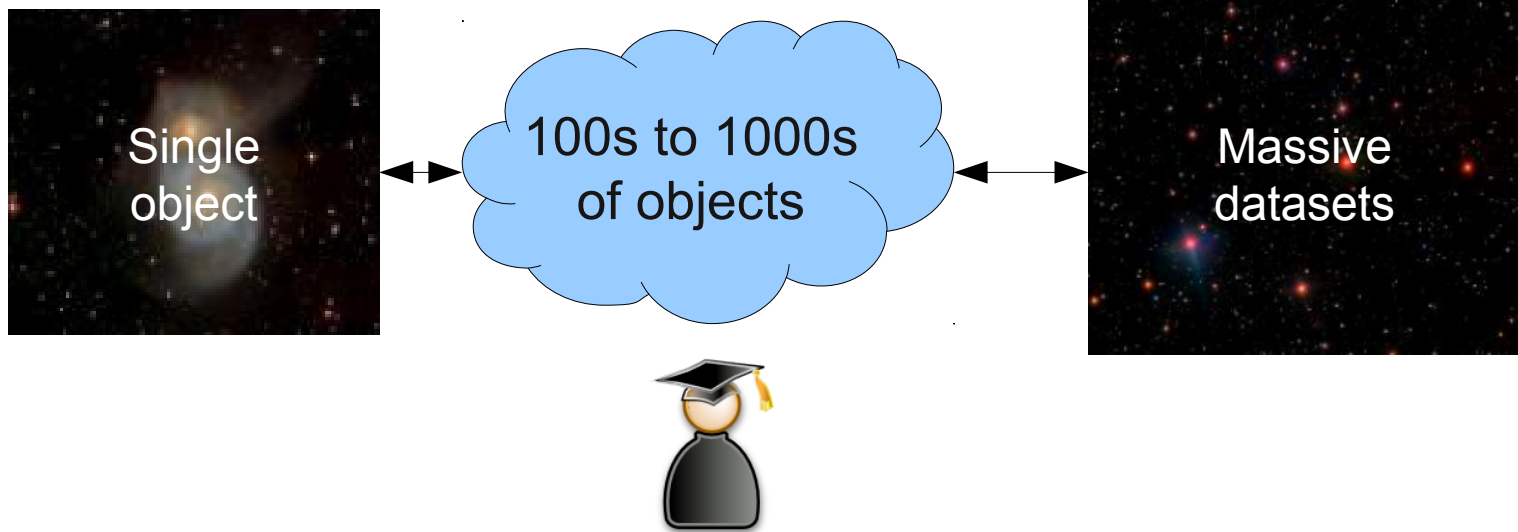
- Software pipelines
- Massive simulations
- Distributed computing
- Supercomputing
- Robotic telescopes
- Online communication
- The Virtual Observatory

- Some paradigm shifts in the way we do astronomy
 - The archive is the telescope!
 - We can't store all data!

SKA Computing Requirements



What does this mean for you?



Scenario: How would you solve this problem?

Your supervisor gives you a data file that they've dug out of the archives. They say that it contains Nobel prize winning data. . . if only you could analyse it. . . you take a look. . .

```
010.1002587 -78.0749976 3.6 4.7 J0000M32 0.000 D
322.2776209 -64.6831876 3.1 3.6 J0000M52 0.108 C
328.0067189 -68.5874347 2.1 2.5 J0000M48 0.193 D
325.6616556 -67.1618942 1.8 2.0 J0000M48 2.539 B
314.1262894 -55.5638273 9.0 2.8 J0000M60 1.266 A
314.6473335 -56.3390663 3.3 3.5 J0000M60 2.522 C
334.7202054 -71.7342040 1.8 2.3 J0000M40 5.693 B
321.3525638 -63.9139186 2.0 2.1 J0000M52 2.080 B
325.7259722 -67.2143813 2.9 3.0 J0000M48 3.311 D
347.0515173 -75.2447016 2.3 3.4 J0000M36 4.345 C
325.5104290 -67.0742957 2.9 3.1 J0000M48 0.105 C
001.9296920 -77.4361767 5.9 6.9 J0000M32 1.530 B
307.3447442 -41.6675454 1.8 2.1 J0000M76 6.400 B
```



... and its 10,000 lines long.

Your task:

- Find the number of galaxies of each type (A, B, C, D) that are in the declination range $-70 < \delta < -60$.

What do you do next?

- Run, screaming, from the room, cursing the astronomers of yesteryear. Then brew a strong coffee and prepare yourself for several days of manual adjustments.
- Sigh and take out some old FORTRAN code your supervisor gave you. Comment out 5 lines, uncomment 10 lines, make a few random tweaks, recompile and hope that it works.
- Spend a few hours remembering Perl and writing a script to reformat it, then sit basking in your own glory, wishing there were other people around to see how brilliant you are.
- Write a one line Unix script in a couple of minutes, then move on to the Nobel prize winning research. You solve this kind of problem hundreds of times a day.

Solution - python

```
types = {'A':0, 'B':0, 'C':0, 'D':0}
for line in open('catalogue.txt'):
    cols = line.split()
    dec = float(cols[1])
    gtype = cols[-1]
    if dec > -70 and dec < -60:
        types[gtype] += 1
for gtype in types:
    print gtype, types[gtype]
```

Solution - shell

```
%> awk '($2>-70 && $2<-60){print $7}' \  
catalogue.txt | sort | uniq -c | sort
```

The astronomer's IT toolkit

- What IT skills do you need to do your research effectively?
 - A data reduction/processing (or simulation) package
 - (e.g. Miriad, CASApy, IRAF)
 - A FITS visualisation package
 - (e.g. kvis, ds9, aladin)
 - A range of Un*x tools
 - (e.g. cut, paste, grep, sed, awk, for loops)
 - A scripting language
 - (e.g. Python, Perl)
 - A plotting package
 - (e.g. IDL, matplotlib, Matlab)
 - Familiarity with accessing large online resources
 - (e.g. NED, SIMBAD, Vizier, ADS)
 - Version control software
 - (e.g. git, subversion, mercurial)

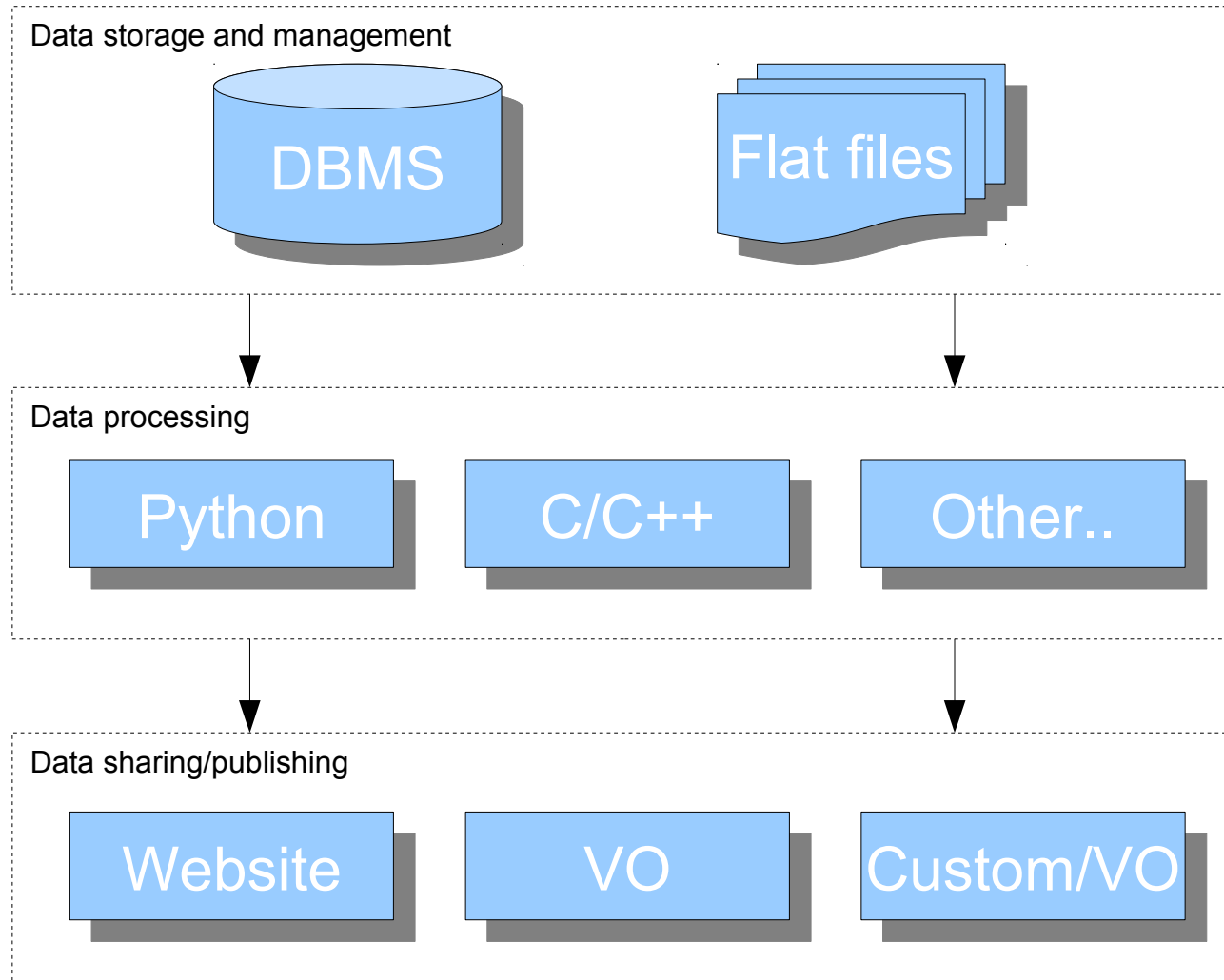
The astronomer's IT toolkit

- Plus sometimes you need to
 - Write software in C, C++, FORTRAN
 - Read other people's code. . .
 - Query databases using SQL
(e.g. SDSS, 6dF)
 - Use a wiki for collaboration
 - Set up a website
 - Set up a website with forms and CGI scripts
 - Set up a database to share/organise your data
 - Use VO Tools for complex queries

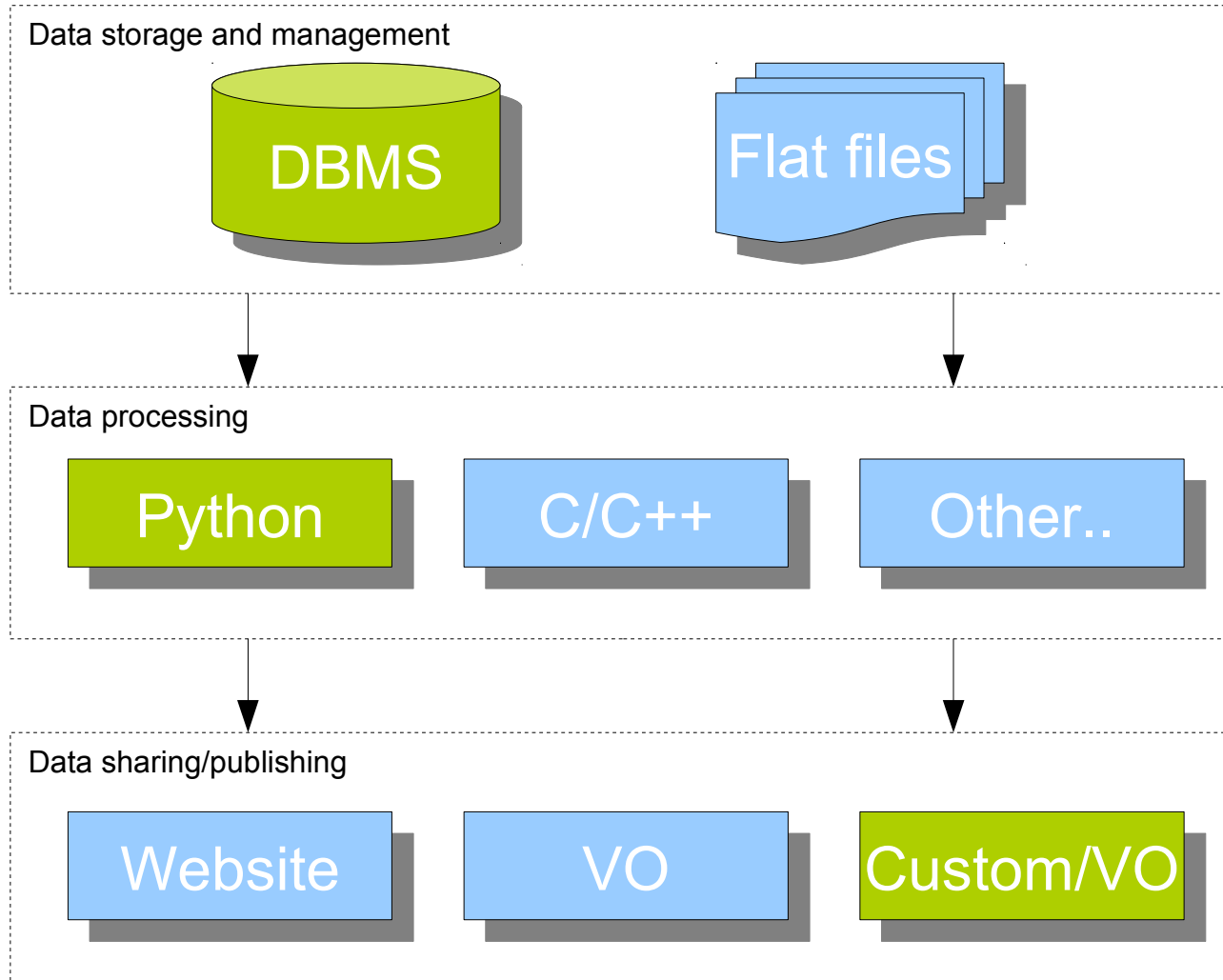
What should I do with my data?

- ... You've had an idea!
- ... You've written an observing proposal
- ... You've slaved away at the telescope for many nights
- ... You've processed your data
- ... You've analysed your images/spectra
- And finally...
- **You have a catalogue**
 - You want to store it
 - You want to put it on the web
 - You want to make it easily accessible to others

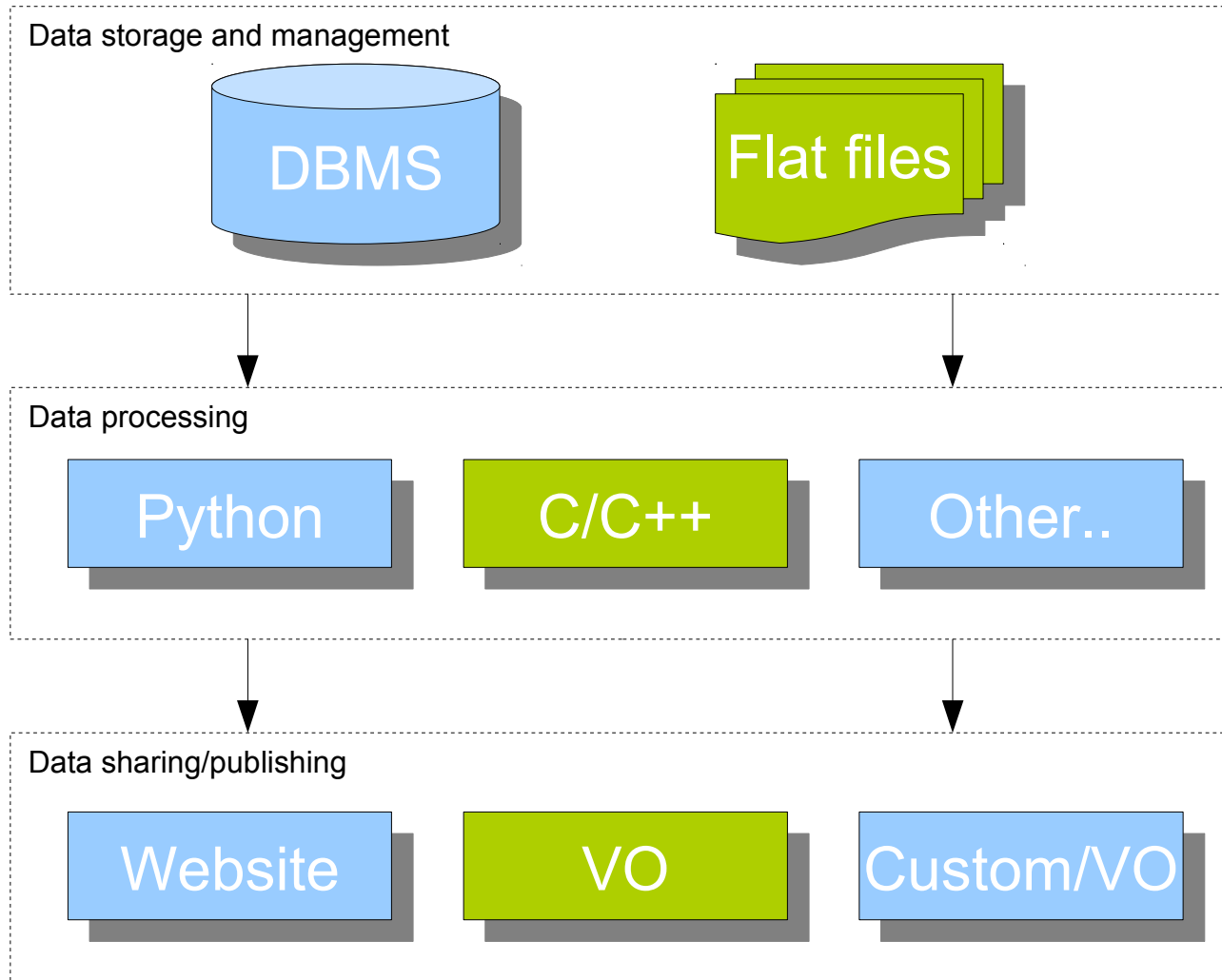
This involves (at least) 3 distinct stages



Your technology choice at each stage is independent



Your technology choice at each stage is independent



Why use databases?

- Persistence
- Access control and security
- Atomicity and concurrent access
- Standard queries
- Avoids inconsistency
- Avoids redundancy
- Avoids data isolation

However. . .

- Some overhead in setting up
- Need to learn new skills
- **Need to evaluate whether it is worth it for your data**

Why use Python scripting?

- Easy to code
- Easy to read
- Supports sophisticated programming (e.g. OO)
- Many built-in functions for modern tools (e.g. databases, web)
- Increasing uptake in astronomy community
- Wide community support

However. . .

- Takes time to learn a new language
- Language X solves all of my problems
- **Need to evaluate whether it is worth it for you**

Why use web/VO?

- You want your data to be used as widely as possible
- Hence you want to share it in ways that are easy to access
- Writing custom solutions is intensive

However. . .

- It is easier just to put your catalogue on a website
- If you want VO-lite upload your catalogue to Vizier
- **Need to evaluate whether it is worth it for your data**

Why should I use scripting?

- Advantages vs. manual processing
 - Speed
 - Reproducibility
 - Documentation
 - Collaboration
- Advantages vs. 'real programming'
 - Speed of development
 - Flexibility
 - Easier for a beginner to understand

Everybody stand back, I know regular expressions!



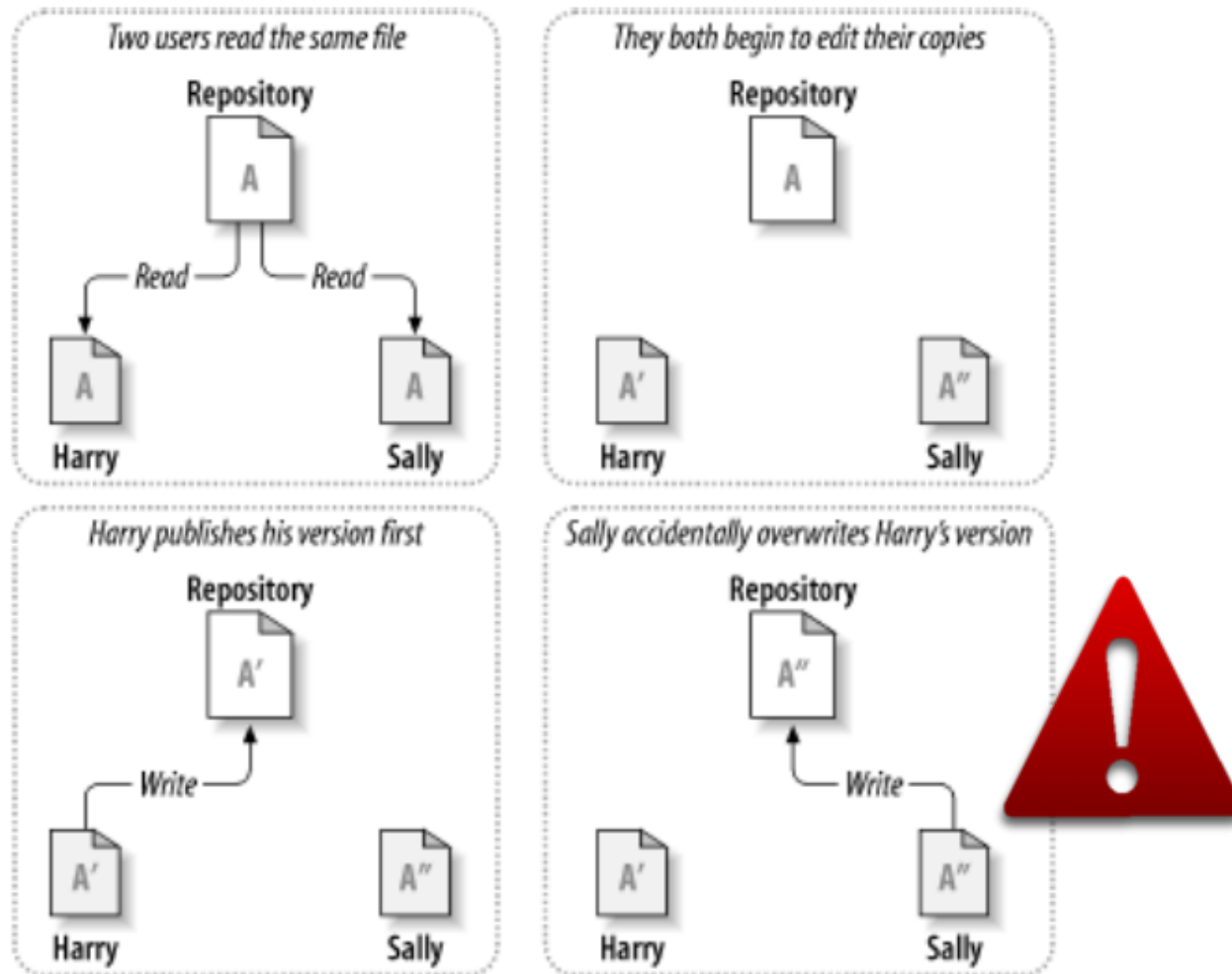
Why should I use version control?

- Works as a constant backup (accessible anywhere at anytime)
- Allows syncing between laptop, work and home desktops
- Allows collaboration on source code, papers, schedule files
- Allows students and supervisors to share code/resources
- Keeps a record of who made changes and why
- Version control works like an e-version of a log book

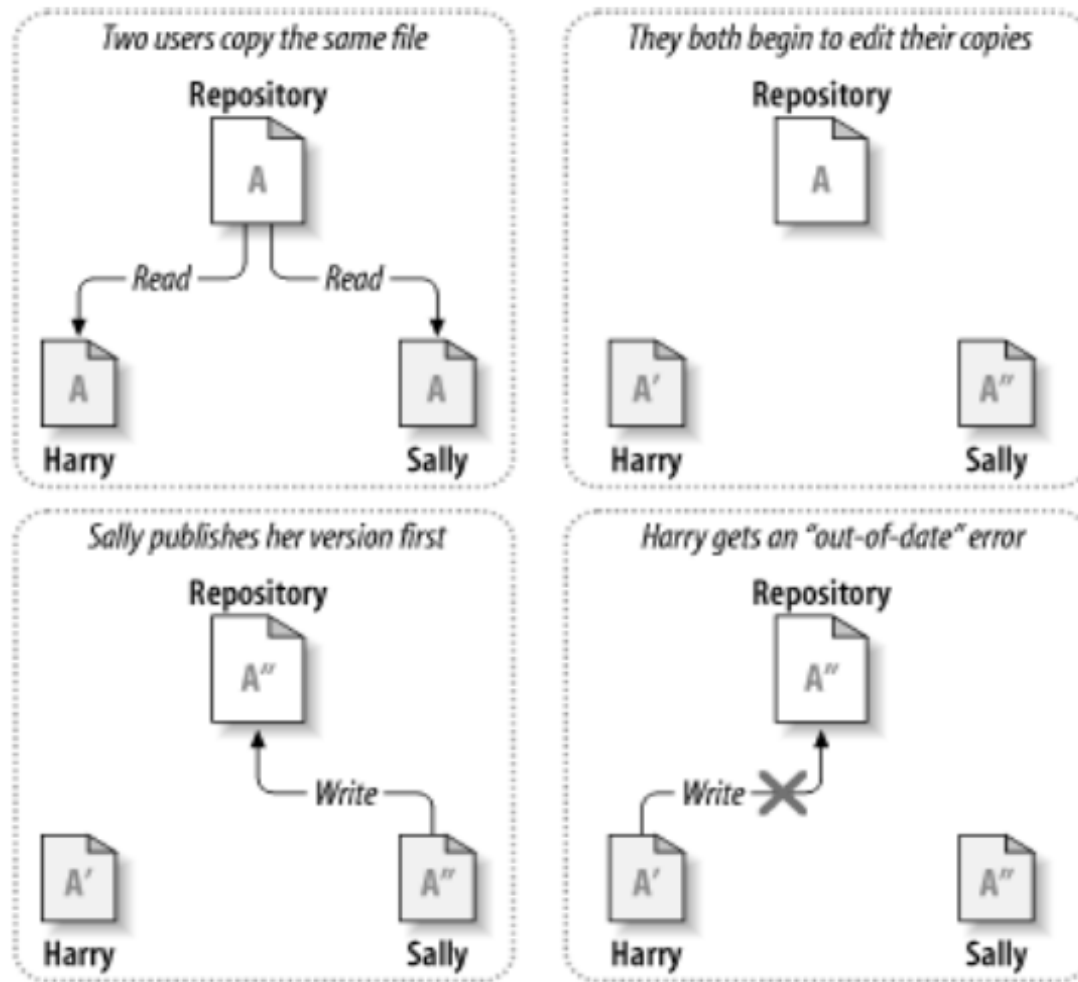
- **It makes you a better coder!**

- For a free (private) repository see <http://www.bitbucket.org>
- Contact your institutes system administrator

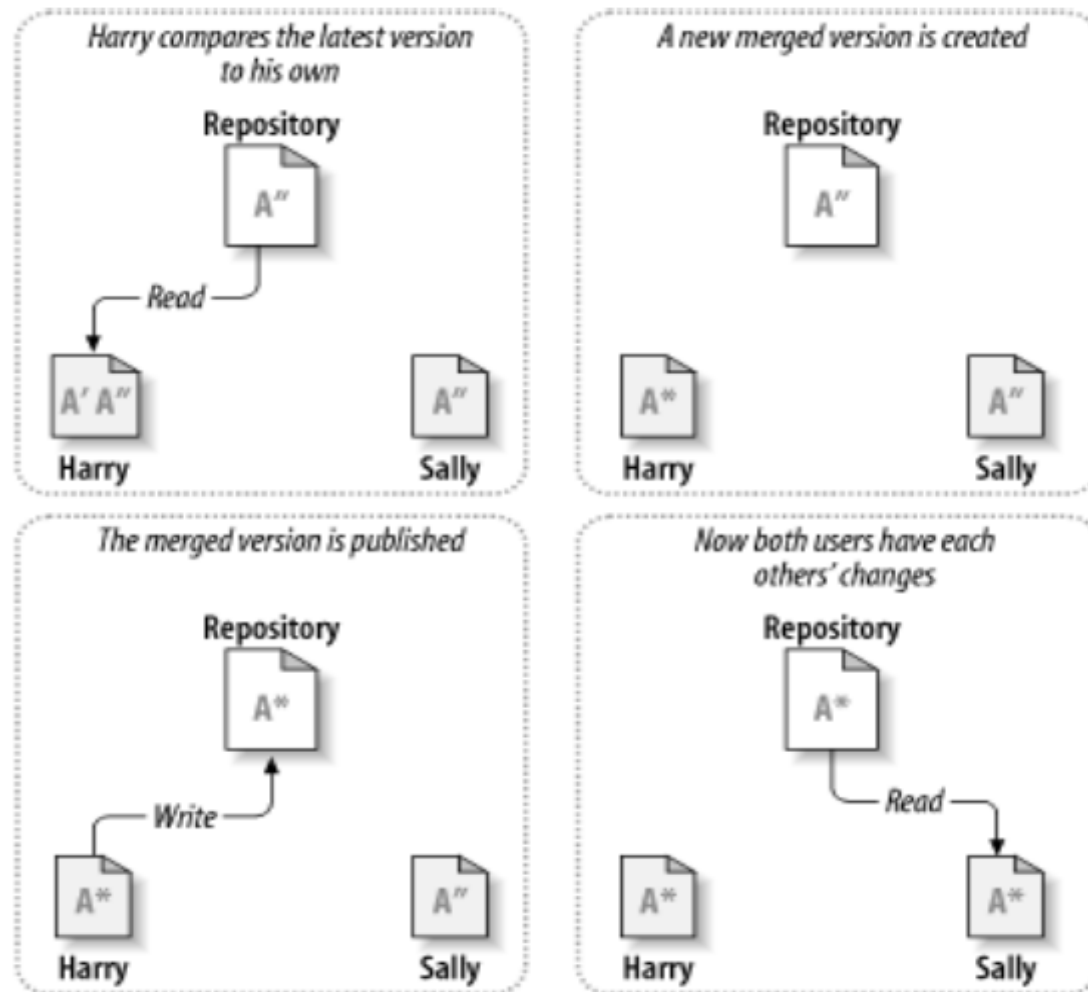
Revision Control – Uh oh



Revision Control – Copy, modify and merge



Revision Control - Copy, modify and merge



What is the Virtual Observatory?

- The VO addresses the data management, analysis, distribution and interoperability challenges of modern astronomy
- The main drivers are
 - Data growth: volume and richness
 - Desire to work online
 - Multi-archive science
 - Large database science
- The Virtual Observatory is a distributed collection of
 - Data resources
 - Software resources
 - Computing (grid) resources
 - Telescopes

The International Virtual Observatory Alliance



What are VO tools?

Open SkyQuery

Nodes

- Rosat
- GALEX
- DL5
- RC3
- SDSS
- SDSSDR2
- TwoDf
- Twoqz
- USNOB
- GOODS
- HDFN
- HDFS
- UDF
- TWOMASS
- IRAS
- PSCz
- ADIL
- FIRST
- NVSS
- NVORegistry

```
SELECT o.objid, o.ra,
o.dec, o.type, t.objid,
t.j_m, o.z
FROM
SDSSDR2:PhotoPrimary o, TWOMASS:PhotoPrimary t
WHERE XMATCH(o, t) < 2.5 AND
Region('CIRCLE 32000 16.031 -0.891 30') AND
(o.z - t.j_m) > 2
```

Spectrum Services

National Virtual Observatory

Spectrum Advanced Search Results

Found 447 objects. Displaying from 1 to 12

<input type="checkbox"/> 1. SDSS J101549.00+002020.00 0271-51878-01 class: Qso, Z= 4.4013 ra = 153.954180, dec = 0.338888	<input type="checkbox"/> 2. SDSS J101549.00+002020.00 0271-51878-01 class: Qso, Z= 4.4027 ra = 153.954180, dec = 0.338888	<input type="checkbox"/> 3. SDSS J102043.82+000105.77 0271-51878-01 class: Qso, Z= 4.2073 ra = 155.182580, dec = 0.018269
<input type="checkbox"/> 4. SDSS J102043.82+000105.77 0271-51878-01 class: Qso, Z= 4.2073 ra = 155.182580, dec = 0.018269	<input type="checkbox"/> 5. SDSS J103432.72-002702.57 0273-51957-01 class: HII_Qso, Z= 4.3771 ra = 158.836330, dec = -0.450713	<input type="checkbox"/> 6. SDSS J103432.72-002702.57 0273-51957-01 class: HII_Qso, Z= 4.3771 ra = 158.836330, dec = -0.450713

An interactive sky atlas/viewer: Aladin

Aladin v3.0 multiview

Load... Save... Tools... Print... Help... Quit

Position J2000 18:02:13.65 -23:01:02.4 Pixel full 0.7594

Trifid nebula

selec

dist

draw

tag

text

filter

rgb

blink

isamp

cont

zoom

mgles

hist

prop

del

Drawing 1

Circ.Mag

HST

USNO-B1

RGB img

SERC.S.MAMA.521

SERC.I.MAMA.521

SERC.V.DSS.521

11.19' x 11.2'

11.19' x 11.2'

11.19' x 11.2'

11.46' x 11.46'

[View B1] - SERC.S.MAMA.521 - Provided by CDS Aladin image server

0669-0683137	270.586123	-23.039303	1964.9	0	0	2	9.84				10.99	0	0
0669-0683138	270.596381	-23.033381	2000.0	0	-12	0	8.64	8.66	8.65	8.66	8.68	0	0
0669-0683139	270.598134	-23.030853	2000.0	-2	-10	0	7.48	7.61	7.55	7.62	7.68	0	0
0669-0683140	270.598778	-23.029200	2000.0	-6	-2	0	10.53	11.48	11.12	11.50	11.71	0	0

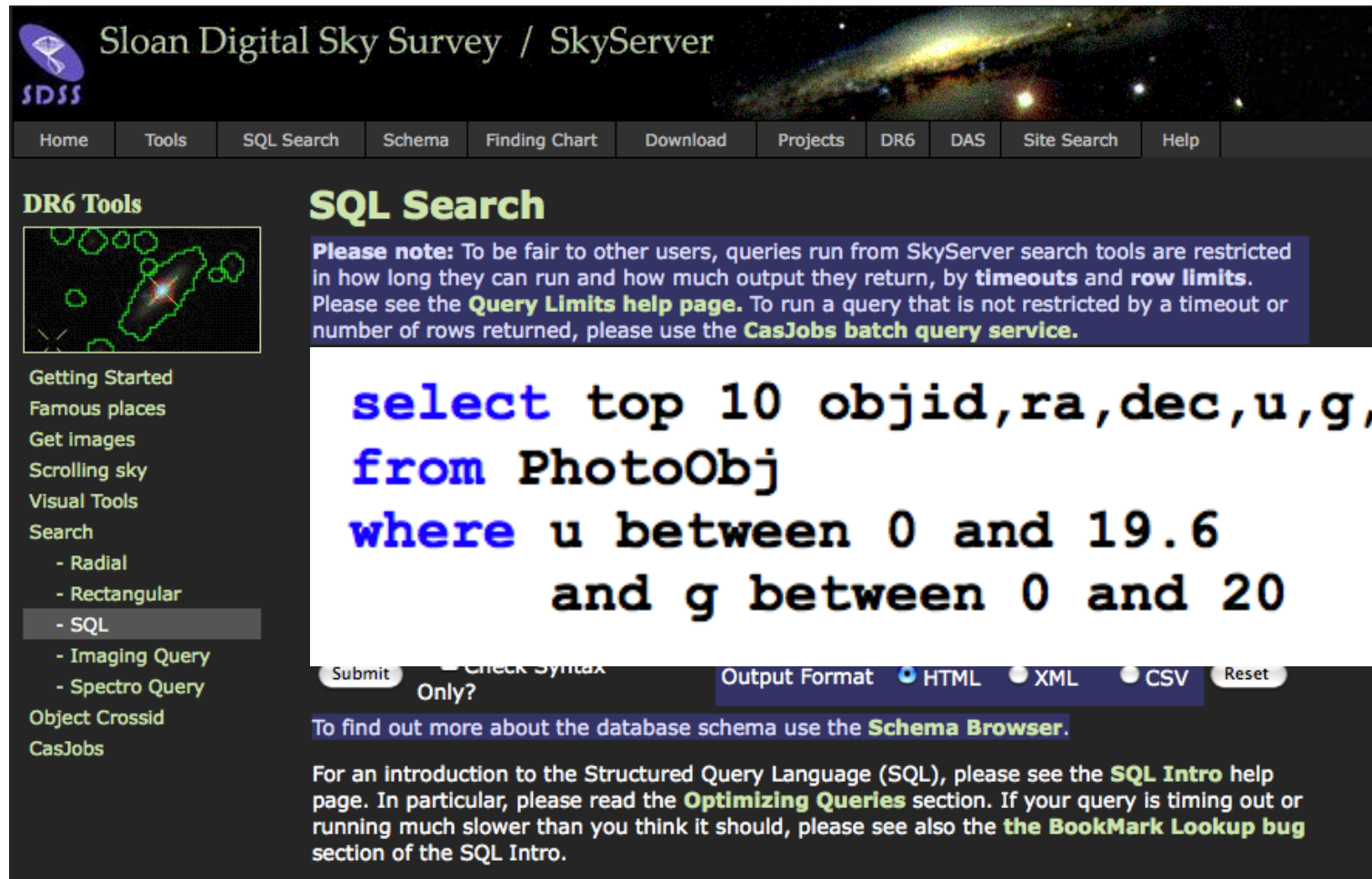
(c)1999-2005 ULP/CNRS - Centre de Données astronomiques de Strasbourg

8 planes, 4 views, 30Mb

An interactive sky atlas/viewer: Aladin

- Visualize digitized astronomical images
- Superimpose entries from catalogues or databases
- Interactively access online data from SIMBAD, NED, VizieR
- Fully VO aware — access other VO resources
- <http://aladin.u-strasbg.fr>
- You can also write your own plug-ins
- The developers are very keen to get feedback from users - they are happy to make suggested changes!

Querying online databases: SDSS



The screenshot shows the Sloan Digital Sky Survey (SDSS) SkyServer website. The header includes the SDSS logo and the text "Sloan Digital Sky Survey / SkyServer". A navigation menu contains links for Home, Tools, SQL Search, Schema, Finding Chart, Download, Projects, DR6, DAS, Site Search, and Help. The main content area is titled "SQL Search" and features a "Please note" box with restrictions on query execution. Below this is a large text input field containing an SQL query. At the bottom of the input area are buttons for "Submit", "Check Syntax", "Output Format" (with radio buttons for HTML, XML, and CSV), and "Reset". A "Only?" label is positioned between "Submit" and "Check Syntax". Below the input area is a "To find out more about the database schema use the Schema Browser." link, followed by an introductory paragraph about SQL.

DR6 Tools

- Getting Started
- Famous places
- Get images
- Scrolling sky
- Visual Tools
- Search
 - Radial
 - Rectangular
 - SQL
 - Imaging Query
 - Spectro Query
- Object Crossid
- CasJobs

SQL Search

Please note: To be fair to other users, queries run from SkyServer search tools are restricted in how long they can run and how much output they return, by **timeouts** and **row limits**. Please see the **Query Limits help page**. To run a query that is not restricted by a timeout or number of rows returned, please use the **CasJobs batch query service**.

```
select top 10 objid,ra,dec,u,g,r,i,z
from PhotoObj
where u between 0 and 19.6
and g between 0 and 20
```

Submit Check Syntax Only? Output Format HTML XML CSV Reset

To find out more about the database schema use the **Schema Browser**.

For an introduction to the Structured Query Language (SQL), please see the **SQL Intro** help page. In particular, please read the **Optimizing Queries** section. If your query is timing out or running much slower than you think it should, please see also the **the BookMark Lookup bug** section of the SQL Intro.

Querying online databases: Open Sky Query

NVO Open SkyQuery

Home Simple Query Advanced Query Import Tutorial Help

National Virtual Observatory

Build Edit Submit

Nodes

- Rosat
- XMM
- GALEX
- GALEXGR1
- DLS
- RC3
- GSC2
- NBCKDEDR1
- SDSS
- SDSSDR2
- SDSSDR3
- SDSSDR4
- TwoDf
- Twoqz
- TWOSLAQLRGEDR
- TWOSLAQQSOEDR
- USNOB
- GOODS
- HDFN
- HDFS
- UDF
- TWOMASS
- IRAS
- PSCz
- FIRST
- NVSS

```
SELECT o.objId, o.ra,
       o.dec, o.r, o.type,
       t.objId, t.ra, t.dec
FROM
  SDSS:PhotoPrimary o, TWOMASS:PhotoPrimary t
WHERE XMATCH(o, t) < 3.5 AND
      Region('CIRCLE J2000 181.3 -0.76 6.5') AND
      o.type = 3
```

Welcome to the Open SkyQuery interactive query builder. You should see a parsed, clickable version of your entered query in the pane directly above this one.

If instead you see 'Query is empty', this means that builder needs a node or two to get started. You can add nodes to the builder by clicking the desired node's '+' icon in the left panel.

Once you have some sql in the above panel, you can then click on a token in that query to pull up a menu with options appropriate for that specific token. For example, one way to select an additional column from a mythical 'mytable' is to click on 'mytable' and then chose 'Add Selection', then pick the desired column from the given choices.

You can switch between 'edit' and 'build' modes at any time by using the tabs at the top of the query panel. Your changes from one will carry over to the other. Most menu options have additional mouse-over info.

Sample Queries

- XMatch/Region
- XMatch/Region 2
- Three Node Match
- Brown Dwarf Search
- MyData XMatch (upload)
- Xmatch t* (upload)
- ABELL Xmatch (upload)
- Single Node Query
- Single Node Join

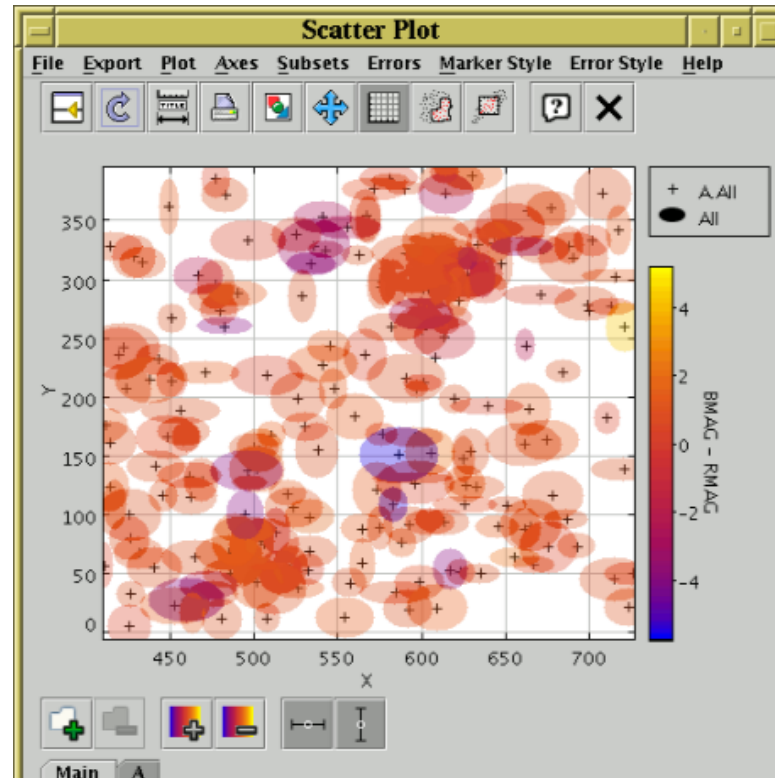
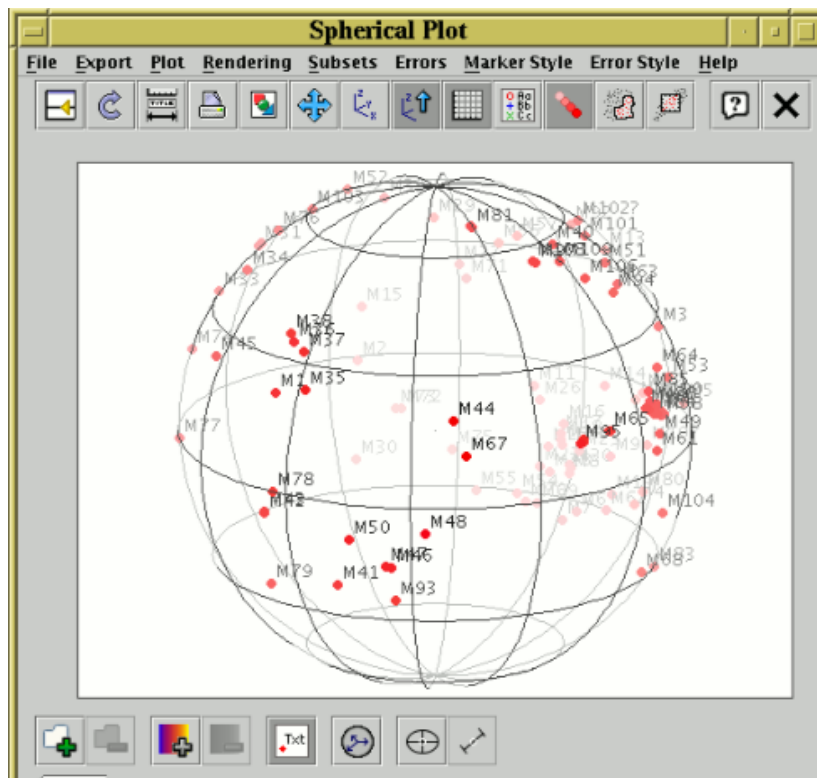
<http://openskyquery.net/Sky/SkySite>

VO enabled plotting

- Many VO tools let you select sources and plot them
- All VO tools allow you to retrieve data as an XML VO table
- Many VO tools can communicate (data)

- TOPCAT is an interactive graphical tool for analysis and
- manipulation of tabular data
- TOPCAT manifesto: Does what you want with tables
<http://www.star.bris.ac.uk/~mbt/topcat>

VO enabled plotting: TOPCAT



Other tools worth looking at

- **DataScope**
 - heasarc.gsfc.nasa.gov/vo
- **SkyView**
 - <http://skyview.gsfc.nasa.gov>
- **MyADS**
 - <http://myads.harvard.edu>
- **AstroGrid**
 - <http://www2.astrogrid.org/science>
- **Google Sky**
 - <http://www.google.com/sky>

Accessing the VO using scripts

- VO protocols make it possible to access the VO automatically
- You can integrate online database queries into your programs
- Libraries available for most major languages
- Java and Python probably best supported at the moment
- <http://www.us-vo.org/summer-school/2008/index.html>

Interesting things I didn't cover

- Machine learning

`http://www.cs.waikato.ac.nz/ml/weka`

- High performance computing
- Object oriented programming
- Web services

IT is critical in future astronomy

- IT is ~~becoming increasingly important~~ essential in 'everyday' science
- It is important to learn/improve these skills now!
 - Attend Astroinformatics 2013 (ASA announcement)
- Your PhD is the best opportunity you'll ever get
- (Your first postdoc is the second best opportunity ;)
- Resources available from the Astroinformatics School website
<http://www.icrar.org/news/pastevents/astroinformatics>
- Resources available from the NVO Summer School website
<http://www.us-vo.org/summer-school>