The Astronomer's IT Toolkit and The Virtual Observatory

Tara Murphy

1st September 2011

Outline

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, 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

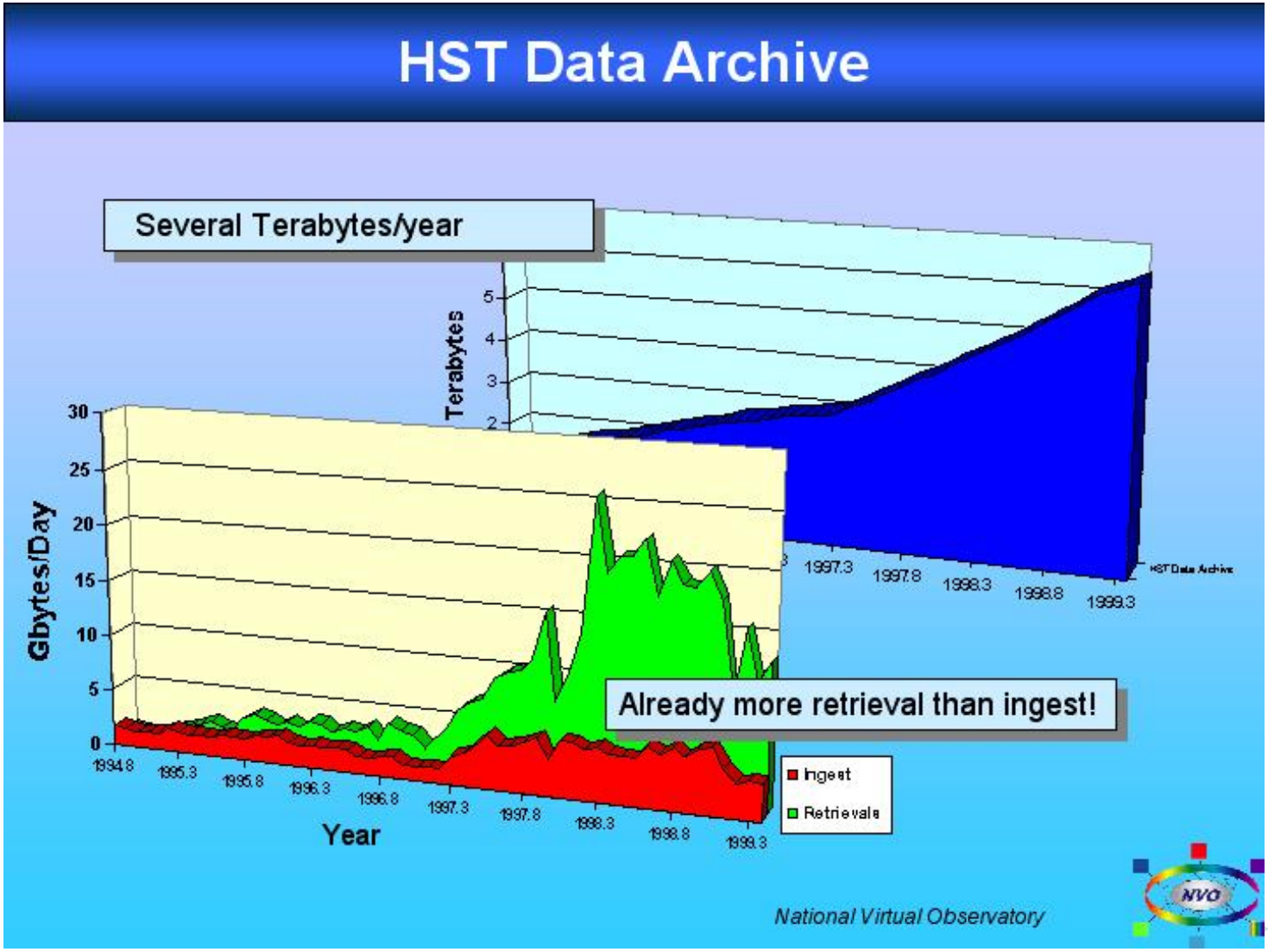
http://www.educationgrid.org/msicii/June06Workshop/Presentations/astro-cyberinfrastructure.ppt

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



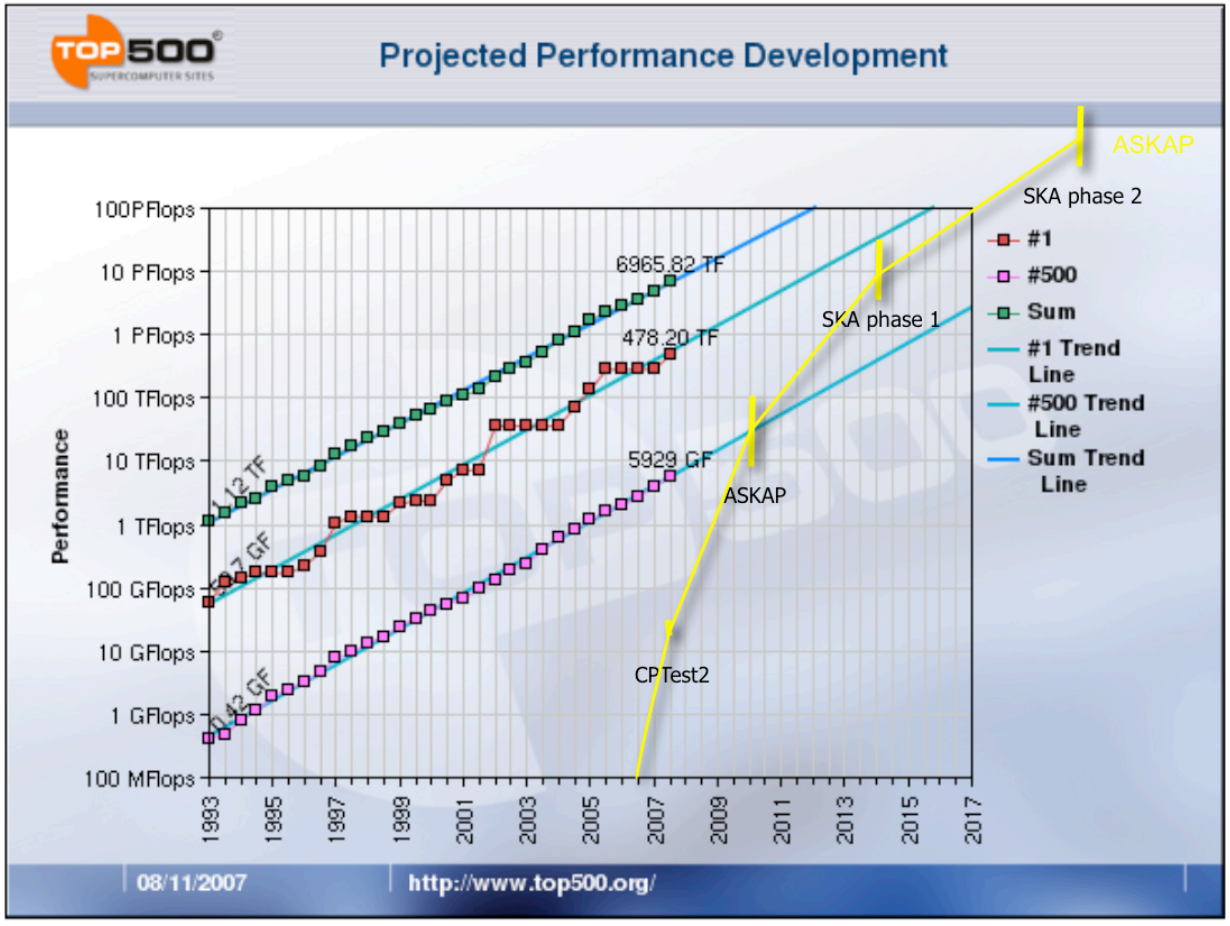
What does this mean for astronomy?

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

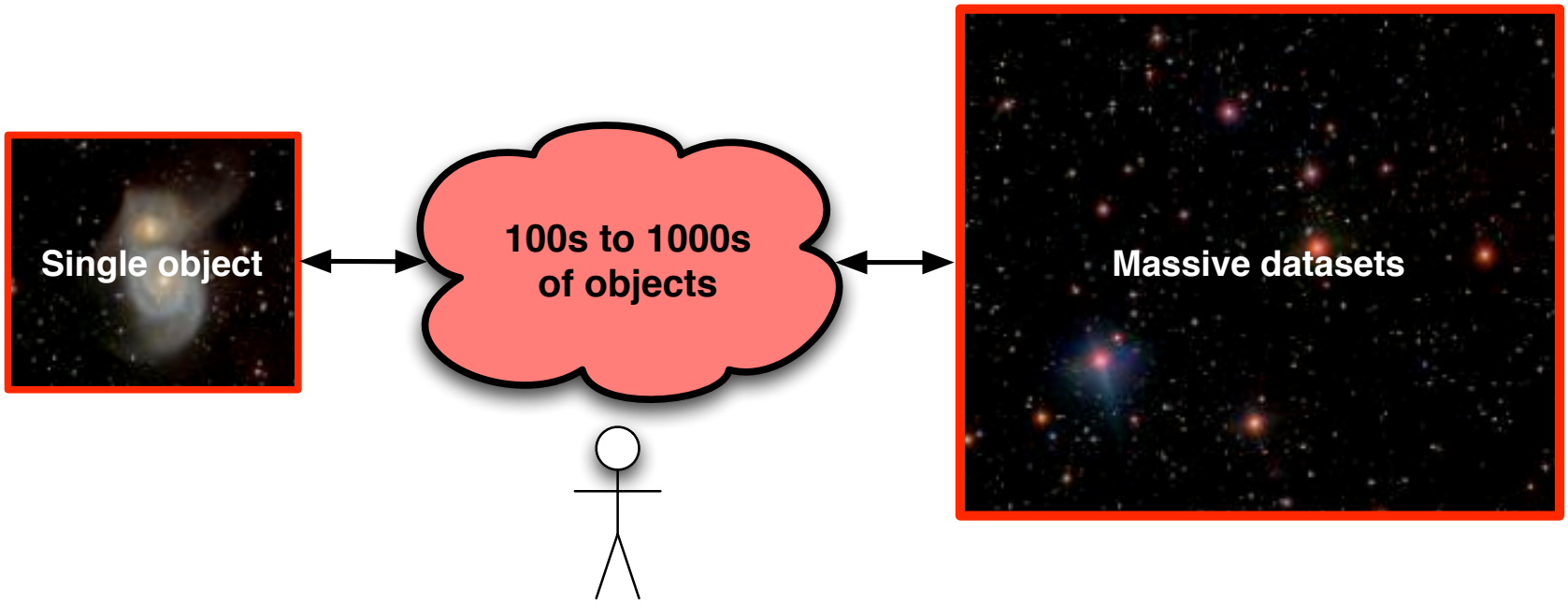
- Some paradigm shifts in the way we do astronomy
  - The archive is the telescope!
  - We can't store all data!

SKA computing requirements (Tim Cornwell)

Climbing Mount Exaflop



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. . .

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?

1. Run, screaming, from the room, cursing the astronomers of yesteryear. Then brew a strong coffee and prepare yourself for several days of manual adjustments.

What do you do next?

1. Run, screaming, from the room, cursing the astronomers of yesteryear. Then brew a strong coffee and prepare yourself for several days of manual adjustments.
2. 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 works.

What do you do next?

1. Run, screaming, from the room, cursing the astronomers of yesteryear. Then brew a strong coffee and prepare yourself for several days of manual adjustments.
2. 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 works.
3. 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.

What do you do next?

1. Run, screaming, from the room, cursing the astronomers of yesteryear. Then brew a strong coffee and prepare yourself for several days of manual adjustments.
2. 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 works.
3. 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.
4. 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.

A Python solution

```
1  types = {'A':0, 'B':0, 'C':0, 'D':0}
2  for line in open('catalogue.txt'):
3    cols = line.split()
4    dec = float(cols[1])
5    gtype = cols[-1]
6    if dec > -70 and dec < -60:
7      types[gtype] += 1
8
9  for gtype in types:
10   print gtype, types[gtype]
```

A Shell solution

```
1 %> 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, AIPS++, IRAF)
  - A FITS visualisation package
    (e.g. kvis, ds9)
  - 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. Subversion)

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

{girls, guys} like {girls, guys} who have skills

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. . .

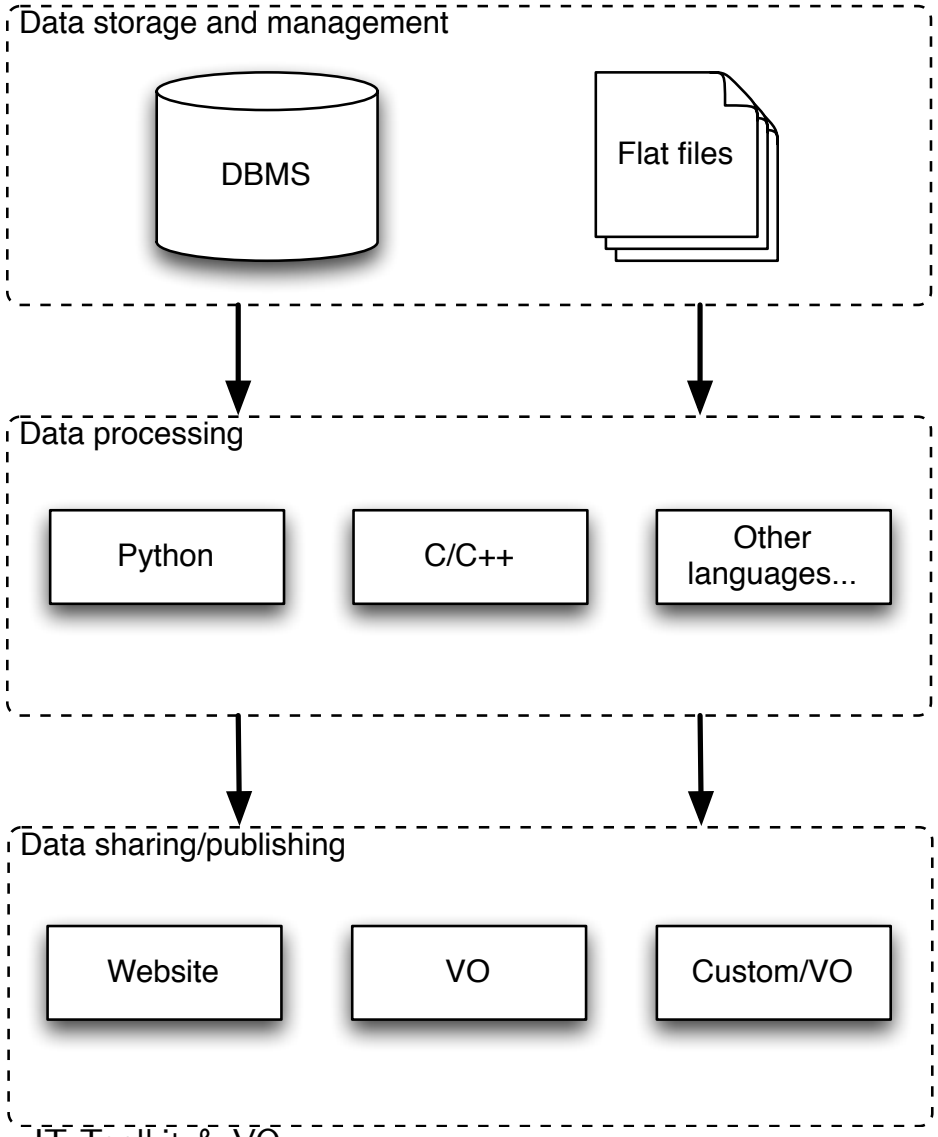
What should I do with my data?

- . . . 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**

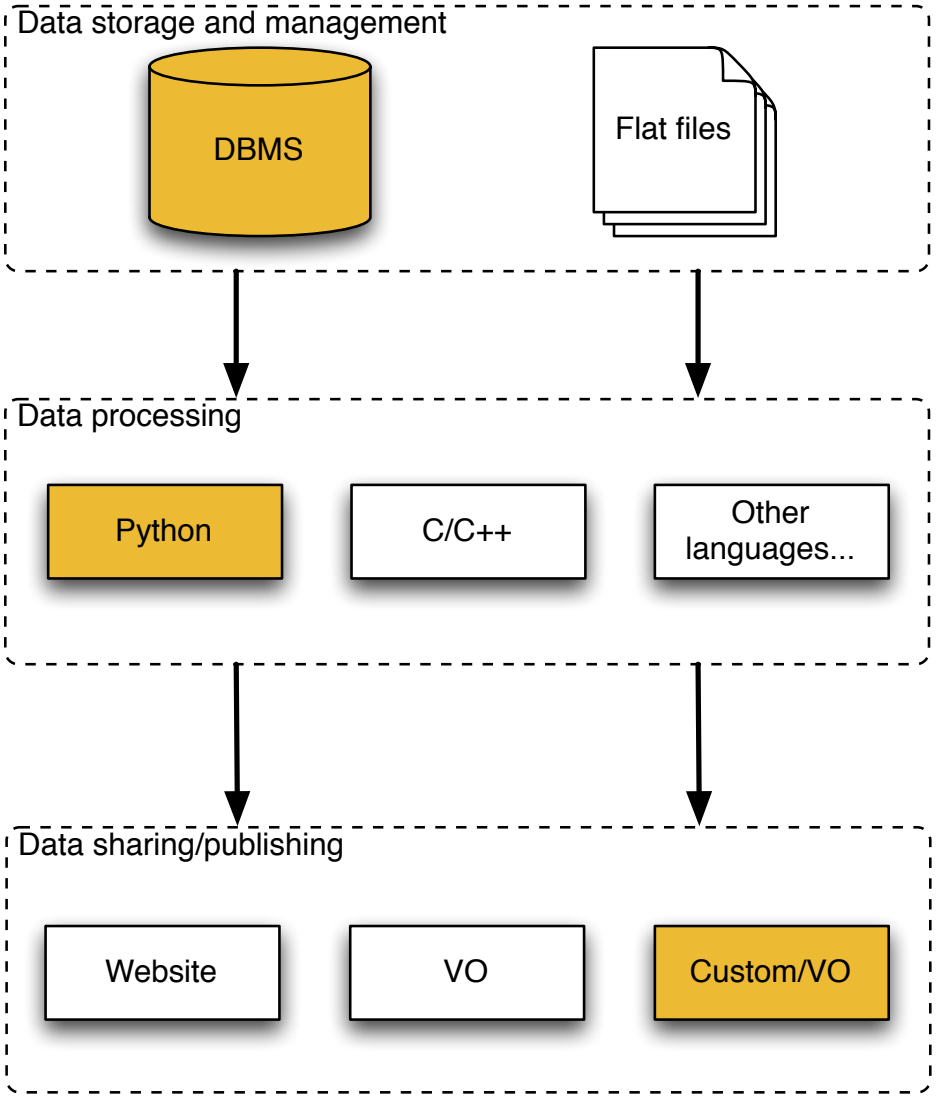
What should I do with my data?

- . . . 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**
  1. You want to store it
  2. You want to put it on the web
  3. You want to make it easily accessible to others

This involves (at least) 3 distinct stages



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!

http://xkcd.com

However. . .


AND A LEAN, SILENT FIGURE SLOWLY FADES INTO THE GATHERING DARKNESS, AWARE AT LAST THAT IN THIS WORLD, WITH GREAT POWER THERE MUST ALSO COME -- GREAT RESPONSIBILITY!

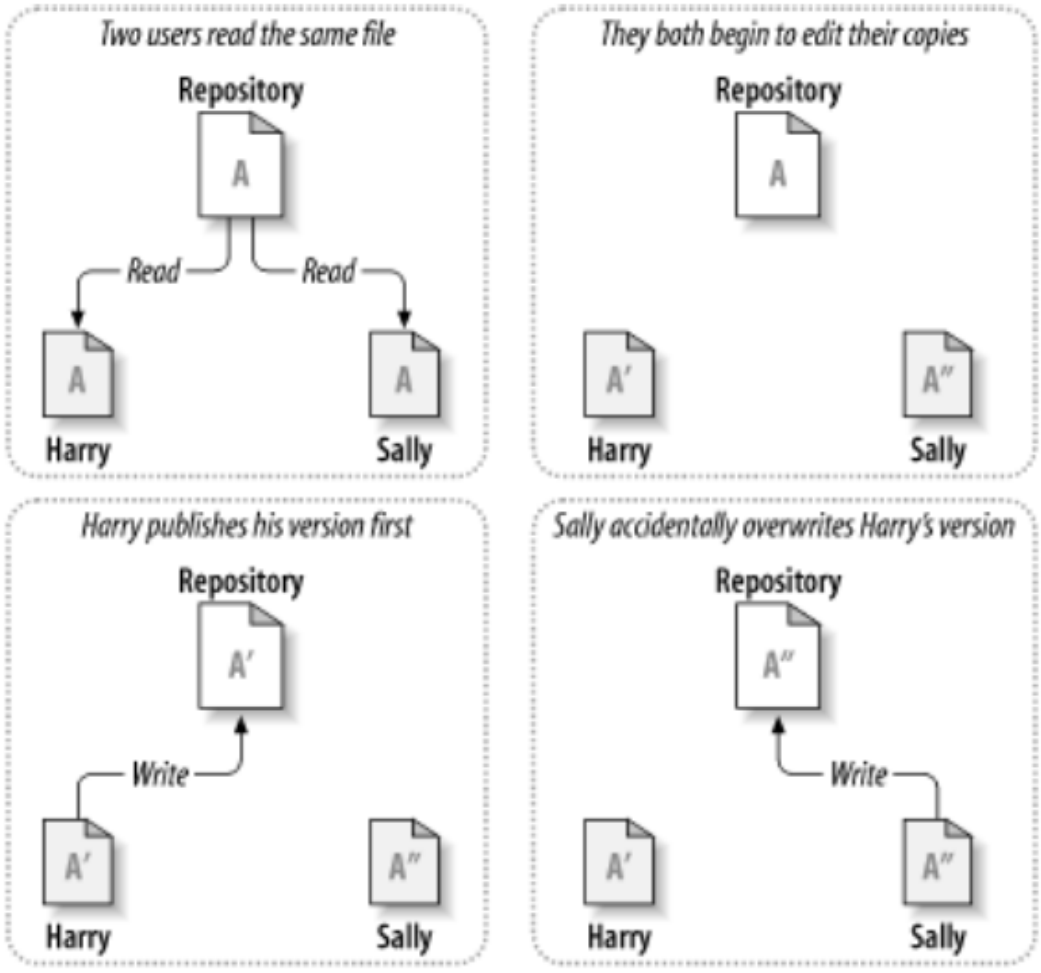
Why should I use version control?

- 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

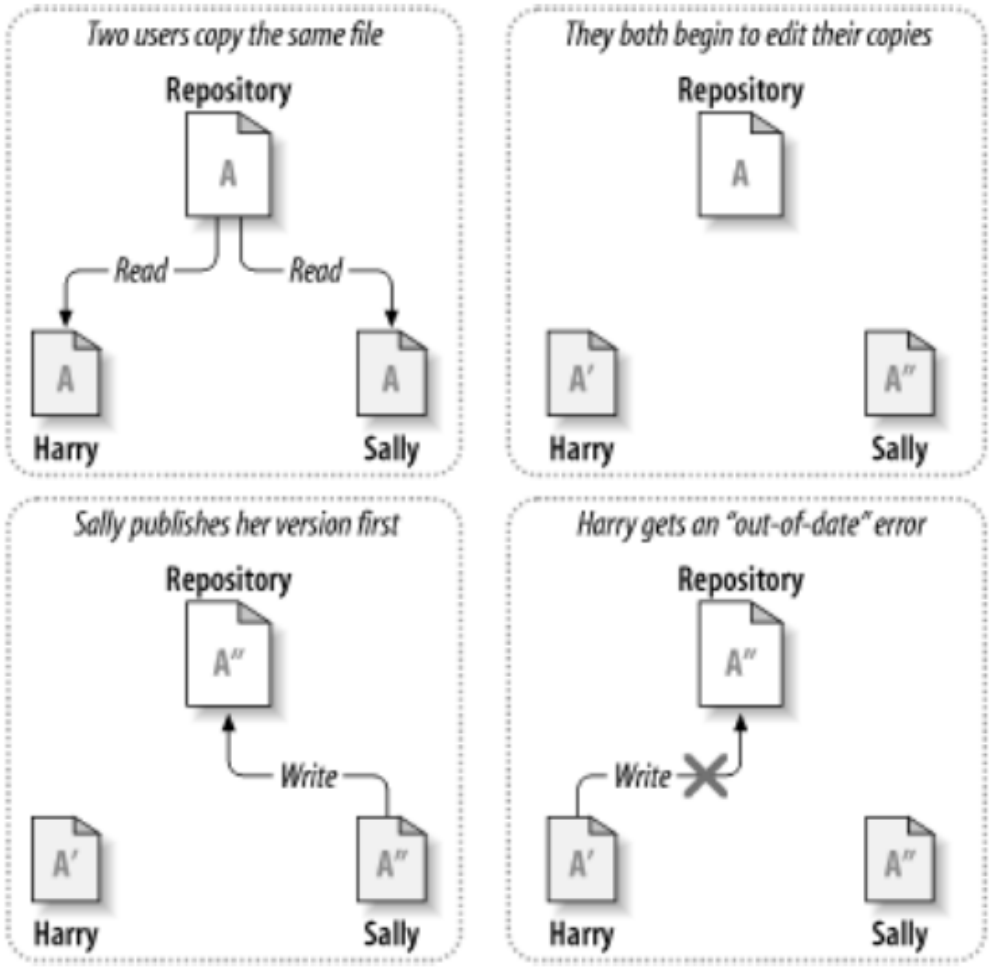
- **It makes you a better coder!**

The problem to avoid



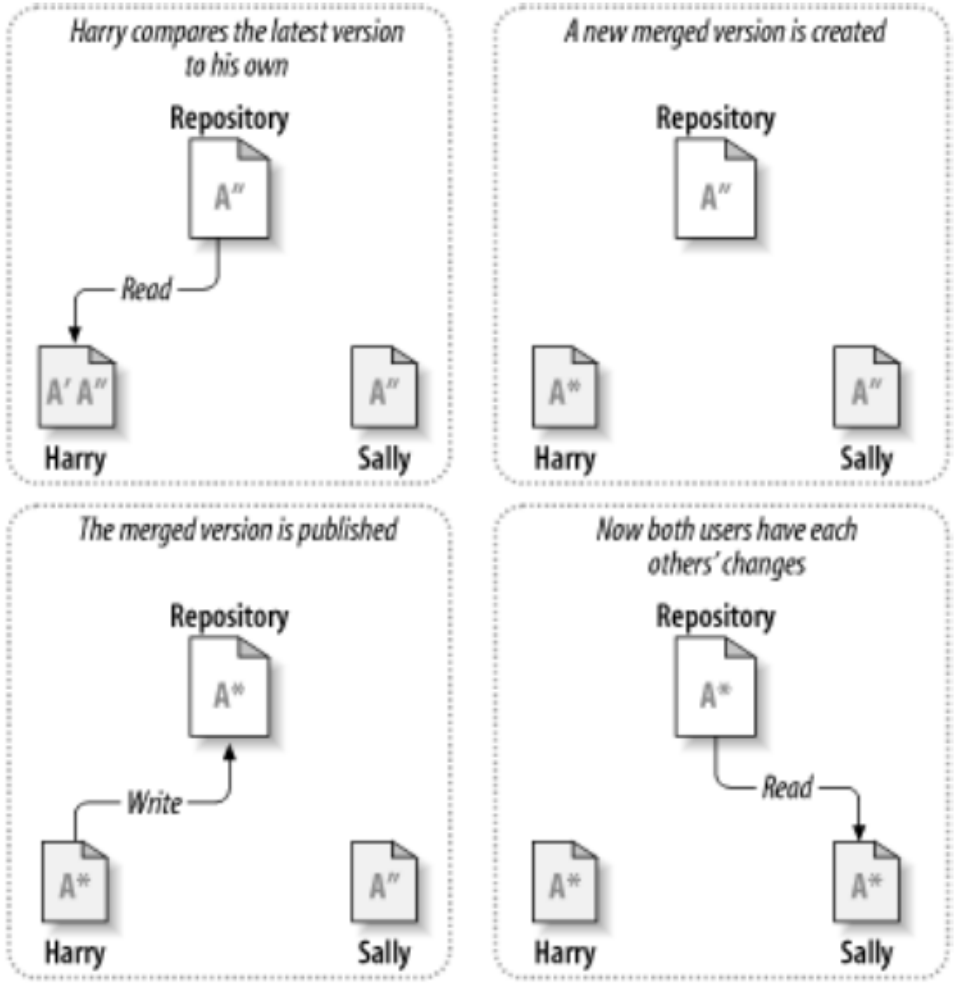
http://svnbook.red-bean.com

The Copy-Modify-Merge solution



http://svnbook.red-bean.com

The Copy-Modify-Merge solution



http://svnbook.red-bean.com

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

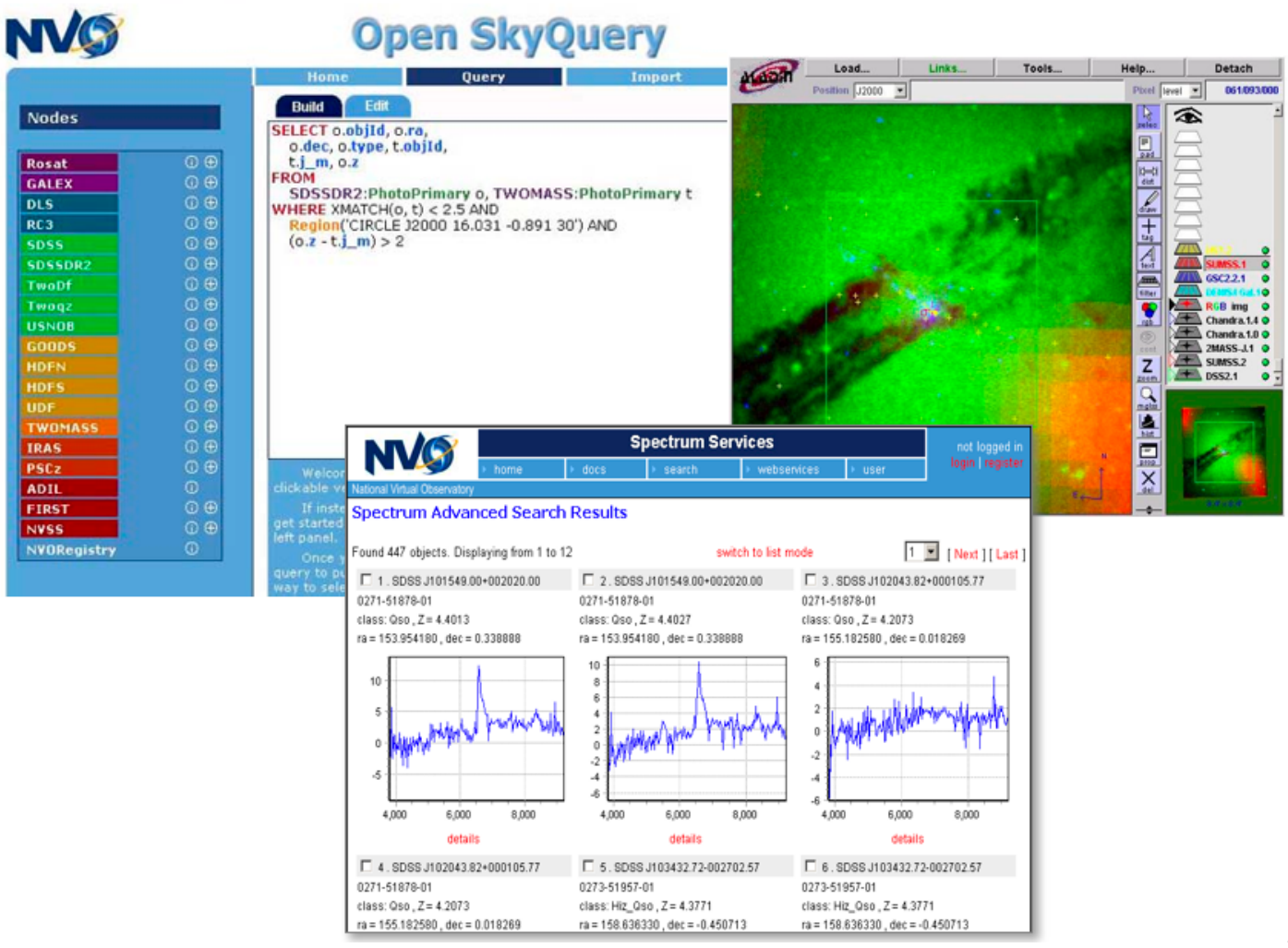
The Virtual Observatory is a distributed collection of

- Data resources
- Software resources
- Computing (grid) resources
- Telescopes

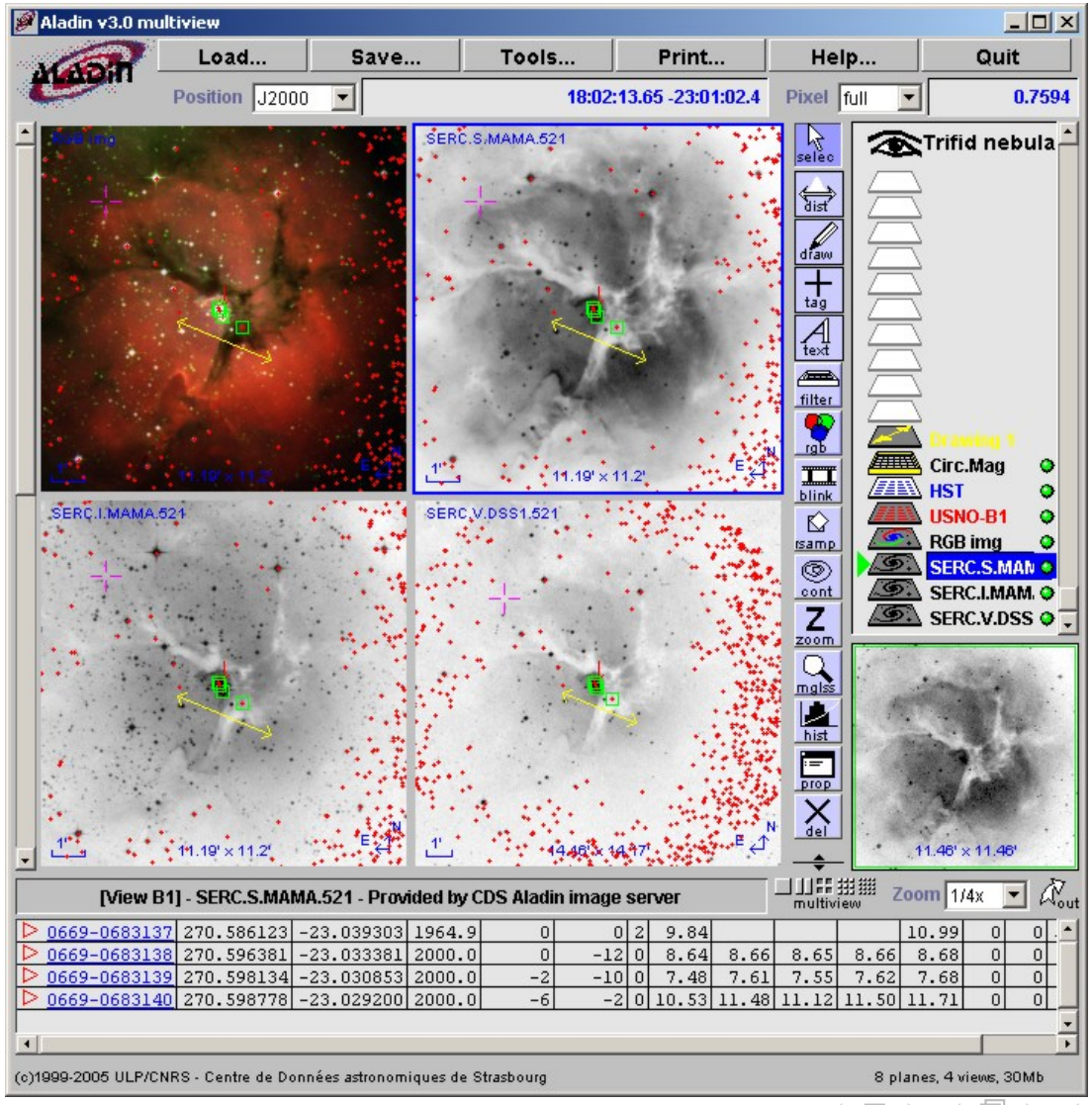
The International Virtual Observatory Alliance

http://www.ivoa.net

What are VO tools?



An interactive sky atlas: Aladin

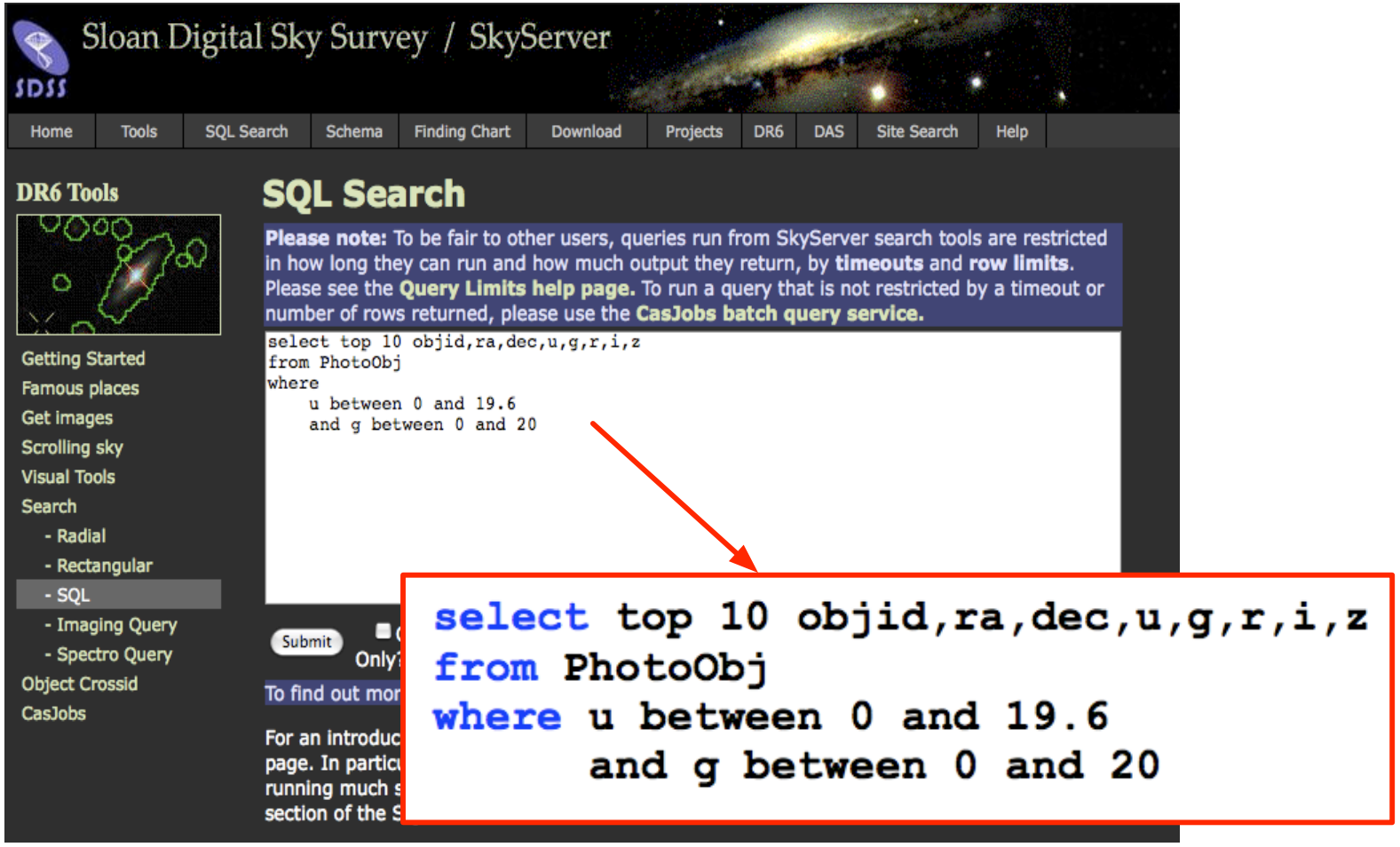


An interactive sky atlas: Aladin

- 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
- **See demo**
- 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



http://cas.sdss.org/astrodr6/en/tools/search/sql.asp

Querying online databases: Open Sky Query

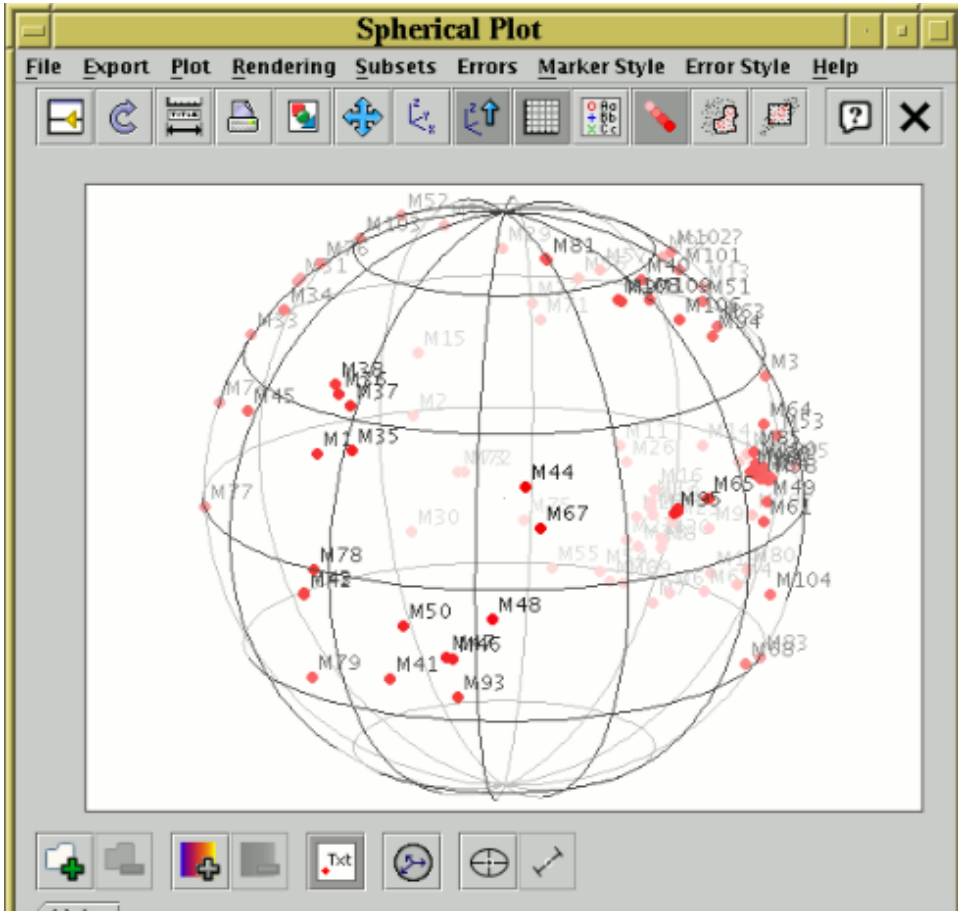
http://openskyquery.net/Sky/SkySite

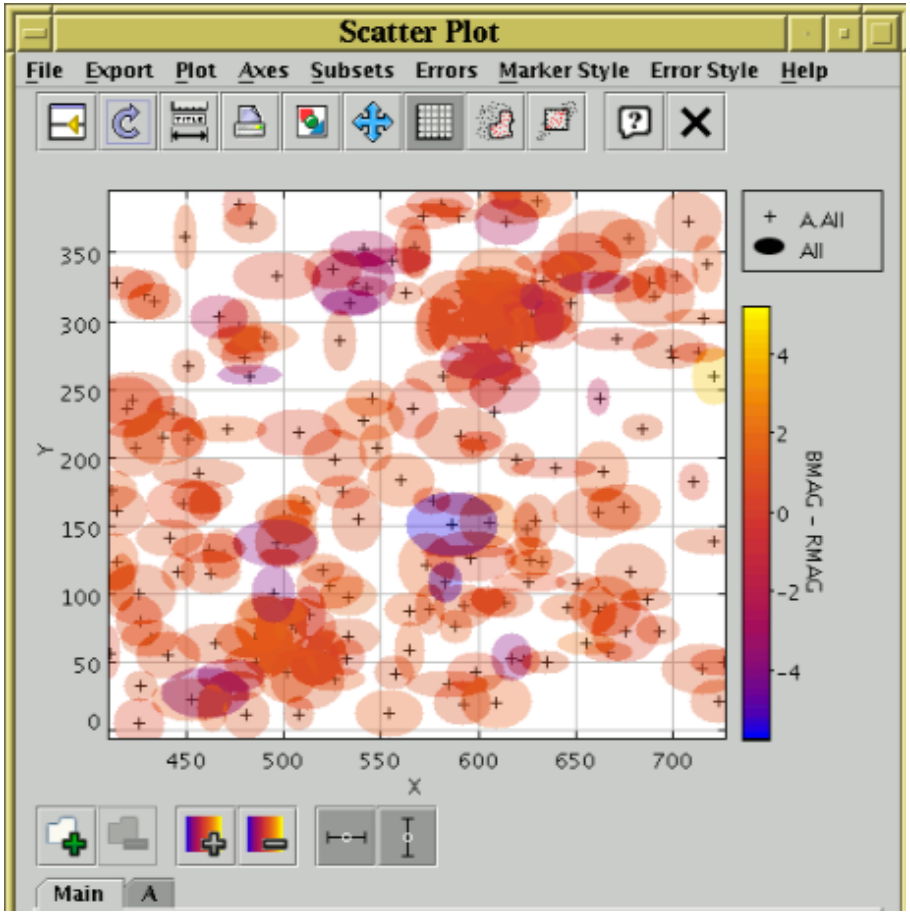
VO enabled plotting

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

- 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



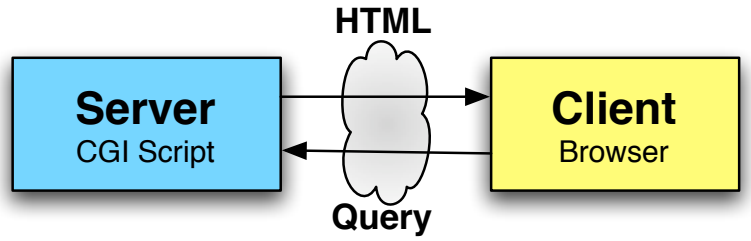


Other tools worth looking at

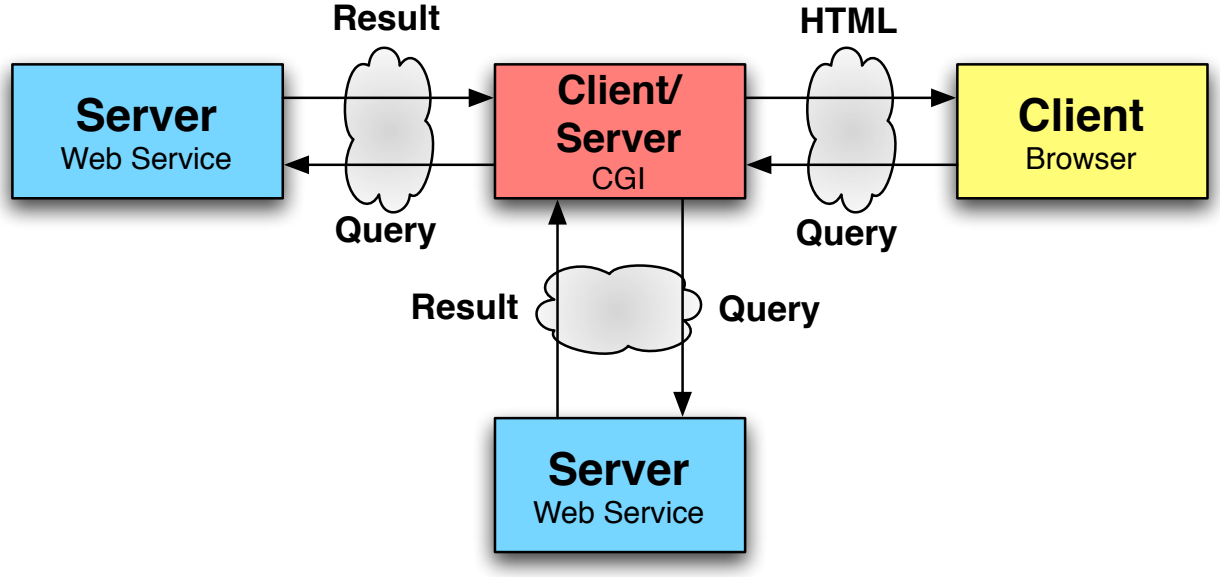
- DataScope
  heasarc.gsfc.nasa.gov/vo
- SkyView
  http://skyview.gsfc.nasa.gov
- MyADS
  http://myads.harvard.edu
- AstroGrid
  http://www2.astrogrid.org/science
- Google Sky
  http://www.google.com/sky

Under the bonnet of the VO

- Many dynamic websites use the CGI client-server interaction



- The Web Service-client interaction



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

Getting your data into the VO

1. Upload to existing service (e.g. Vizier)
   http://vizier.u-strasbg.fr/cgi-bin/VizieR
2. Set up your own VO compliant server
   Implement VO protocols such as Cone Search
   http://www.ivoa.net/Documents/latest/ConeSearch.html
3. Off the shelf setup (e.g. Astrogrid DSA)
   http://www.astrogrid.org/maven/docs/HEAD/pal/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 in 'everyday' science
- It is important to learn/improve these skills now!

- 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

- Look out for the Astroinformatics School in 2012