

# **Astronomy - when, why and how to share your data**

## **the experience of large surveys**

---

Jarle Brinchmann  
Leiden Observatory

# Outline

- An overview of data sharing & surveys in astronomy.
- In-depth examples: The Hubble Deep Field and the Sloan Digital Sky Survey.
- Lessons learned
  - $1+1 > 2$
  - Why it is not so bad if people do bad science with your data.
  - It does take effort.
  - ...

But keep in mind: Astronomical data are “worthless”

# Astronomical observations

What do we gather and what can we share?



Observatories



Survey facilities

We mostly record electromagnetic radiation - and we can usually **not** influence what we observe. While classical experiments can be done in **astrochemistry**, generally we have to make do with what we can see.

# Observatories

- A set of observations (imaging, spectroscopy, ...) is proposed by a group (a few nights duration).
  - Time is allocated by a peer review group.
- Data can be gathered by the group or automatically (“service” observing).
  - In most cases, the data will be automatically stored in an archive facility at the telescope/ground facility.
- Typically the data is private for ~1 year.
  - After this proprietary period the **raw** (i.e.. not reduced) data becomes publicly available.
- The group might provide the final reduced/analysed data for others to use.
  - For large requests observatories sometimes require this to be done.

# Survey facilities

- A major project (~years) proposed by a group.
  - Time/money allocated by review panels/research foundations etc.
- Data are typically gathered and stored automatically.
- The proprietary period varies from 0 to several years.
  - Many projects without a proprietary period require massive infrastructure to make use of the data - e.g. GAIA and LSST.
- Some sort of public data release is expected/required, often at regular intervals.
  - A dedicated web site/archive might have to be created by the project.

# Time-variability surveys

Automated surveys, data on **events** are provided freely to everyone who cares to subscribe to the feed.

Also the plan for the upcoming LSST.



This does *not* mean that the **raw data** is freely available.

The reason is that these events must be followed up quickly so spreading the word as widely as possible maximises science output from the survey.

It does mean that chances are that the team behind will be scooped on the one really amazing discovery.

# In-depth 1 - the Hubble Deep Field

Revolutionary idea to dedicate a large amount of time on the Hubble Space Telescope (HST) to **one** project chosen by the community - so community driven.

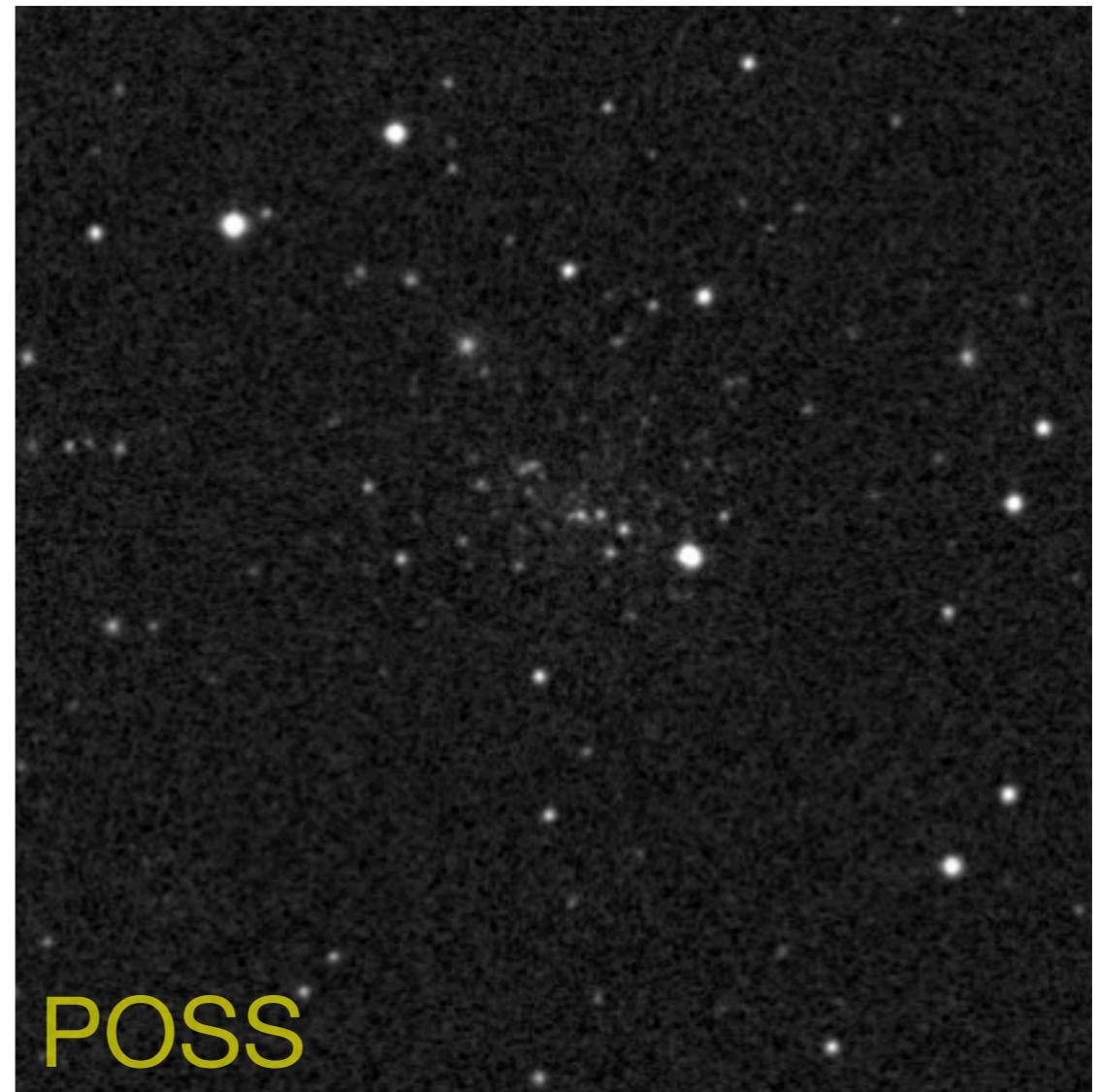
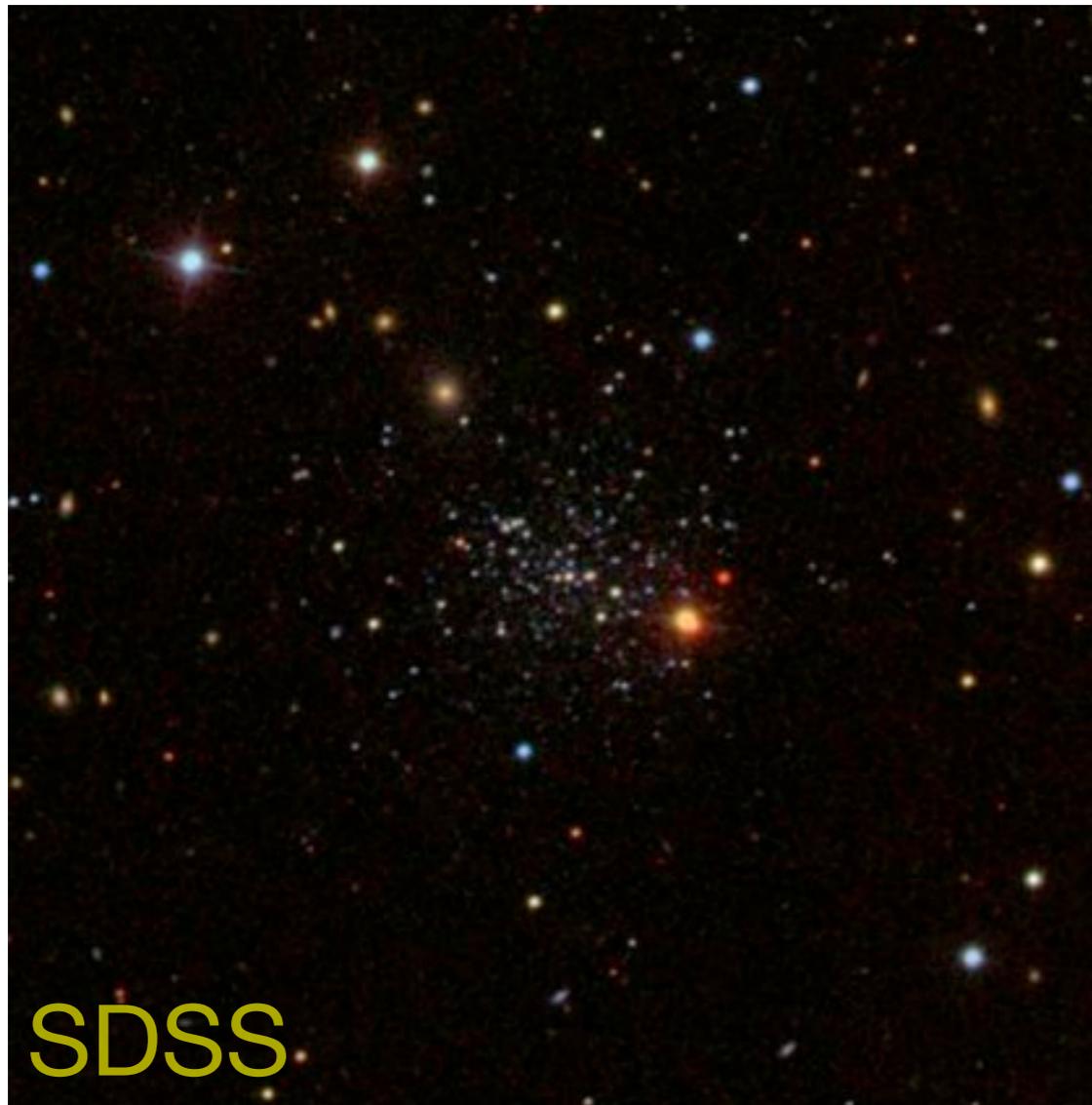
One small patch of sky [ $3 \times 10^{-8}$  of the sky] - 10 days exposure on the HST. ~1000 papers

People jumped on the data - lots of work was done in a short time (days) and it gave a major boost to investigations in this area.

**Lesson 1:** Excellent data provided freely can give a major impetus to a scientific area.

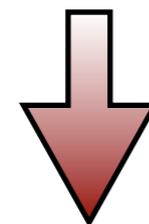
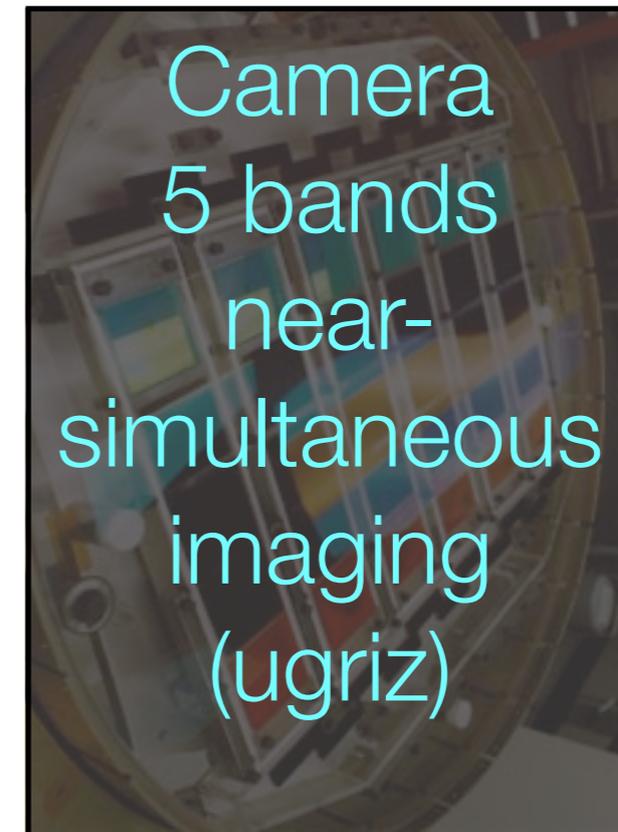
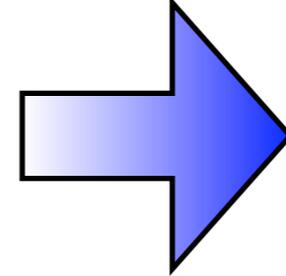
# The Sloan Digital Sky Survey

The natural successor to the Palomar sky survey

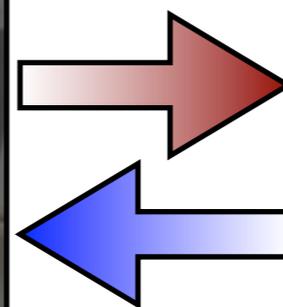
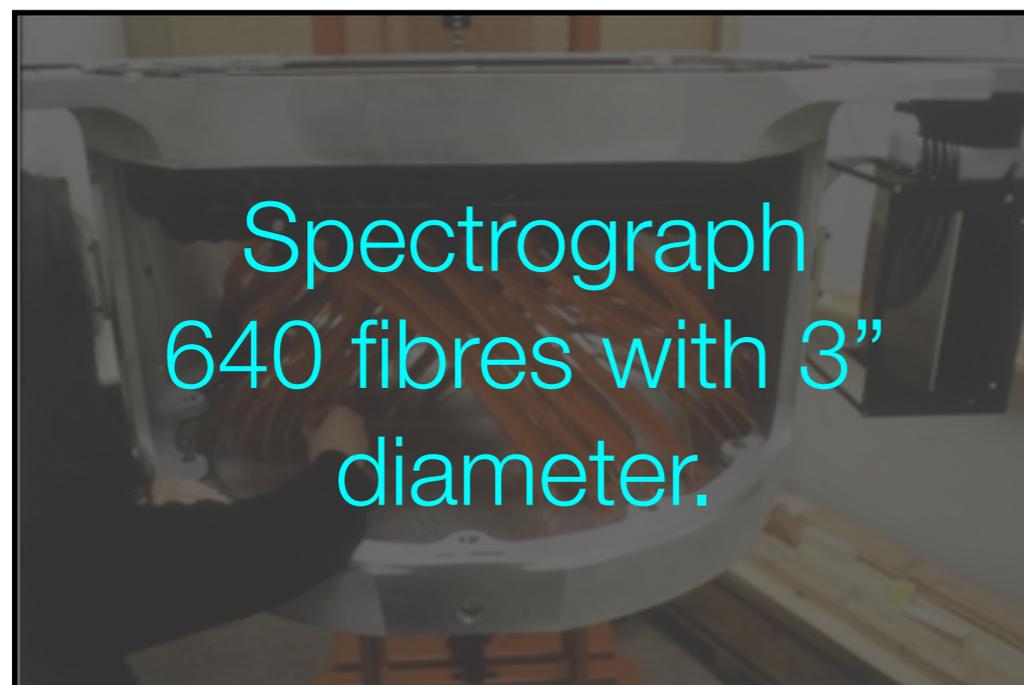
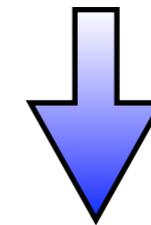




Images



Spectra



Fermilab - where the  
imaging data was  
analysed.  
And the  
spectroscopic data

# The Sloan Digital Sky Survey

The natural successor to the Palomar sky survey

- 30 Tb of reduced data - delivered  $>100$  times that to the world.
- At the start, the largest **hard disks** (8Gb!) were **filled** every **25 mins** when doing imaging.
- $>300$  million web hits with over  $10^6$  unique users (there are about  $\sim 10,000$  astronomers in the world).
- About 200 million SQL queries made since 2001.
- Integral to the development of Google Sky (and WorldWide Telescope).

# Handling the data

1<sup>st</sup> goal: Get Microsoft interested in the project

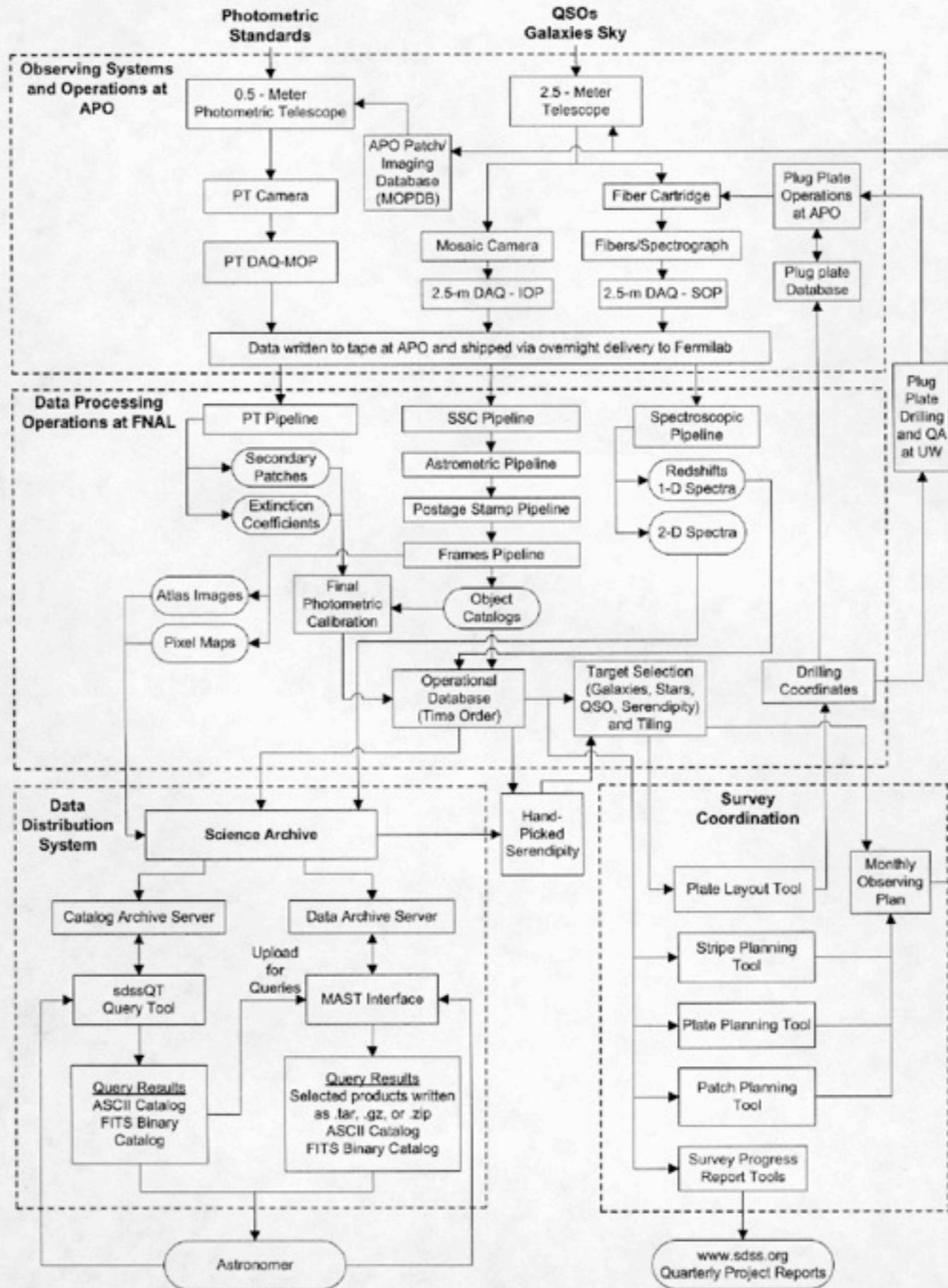
Raw data is stored on data-servers - originally a robotic tape archive, now on a few hard disks.

Metadata = catalogues, characteristics of the observations, locations of files etc etc, is stored in a relational database.

Web interfaces to database and data server, including a SQL workbench and visual tools for browsing the sky. Easily linked to educational projects.

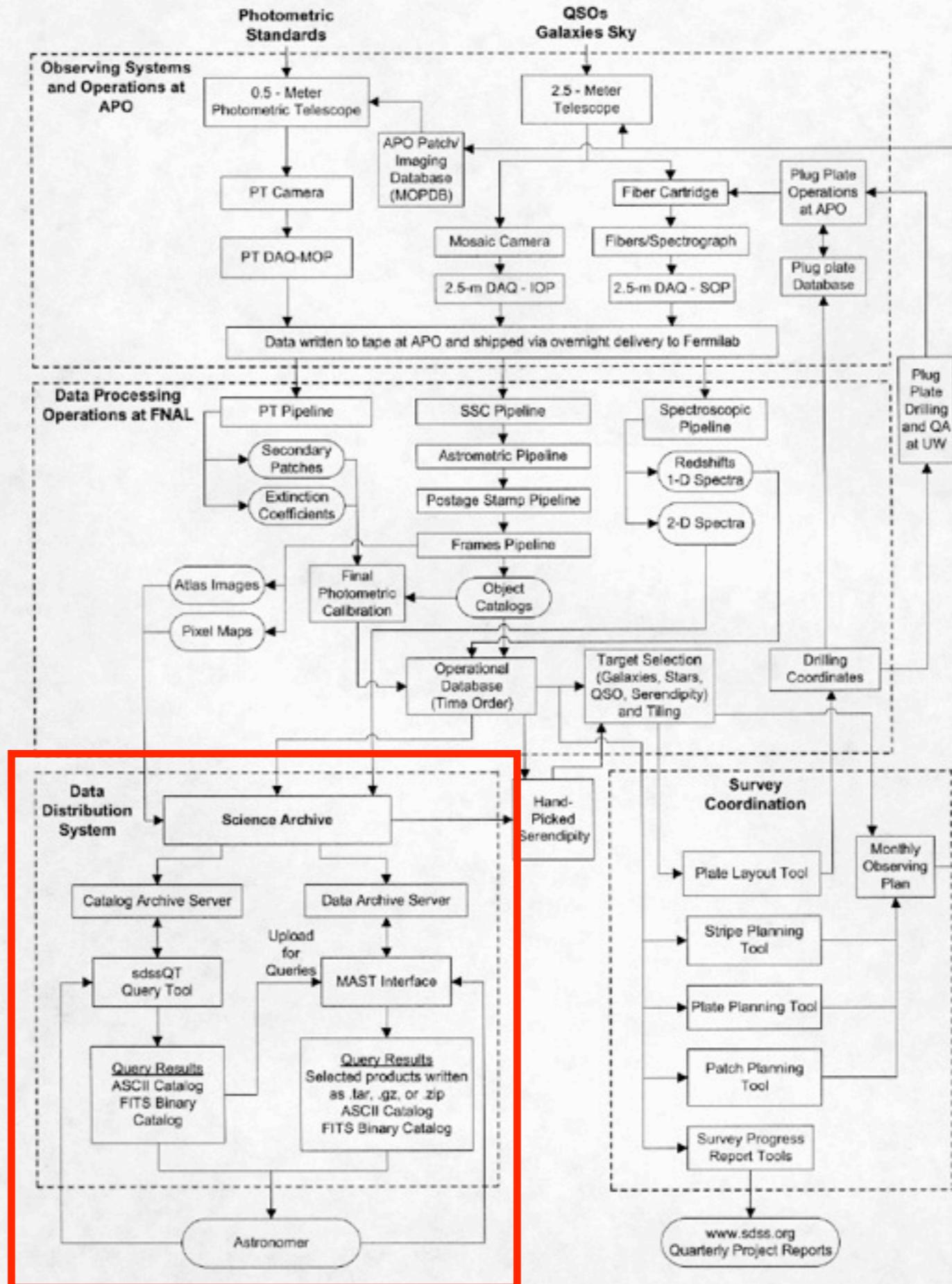
# SDSS Data Flow

April 10, 2000



# SDSS Data Flow

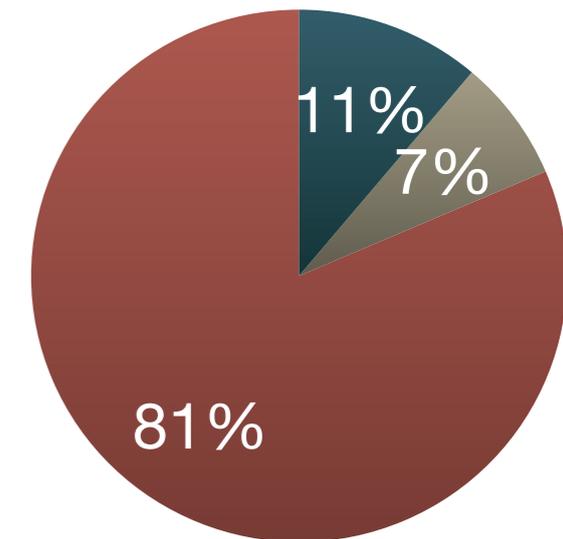
April 10, 2000



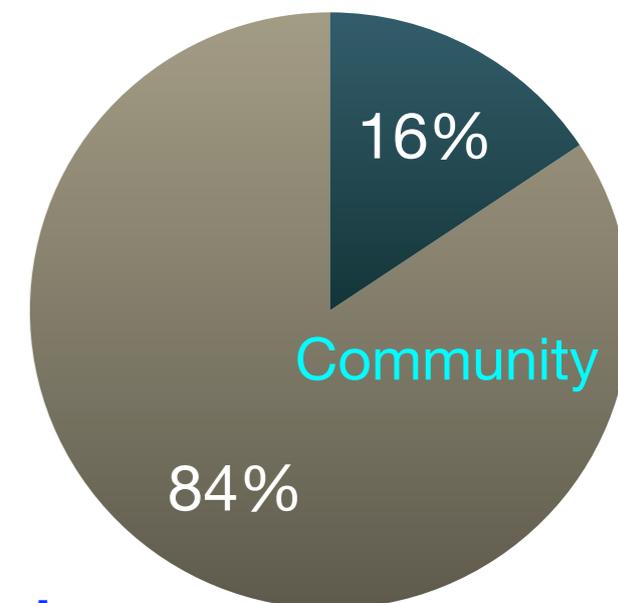
# SDSS lessons

Out of 834 “official” SDSS journal papers:

Area	# papers	Percentage
Cosmology	93	11.2%
Supernovae	62	7.4%
Legacy	679	81.4%

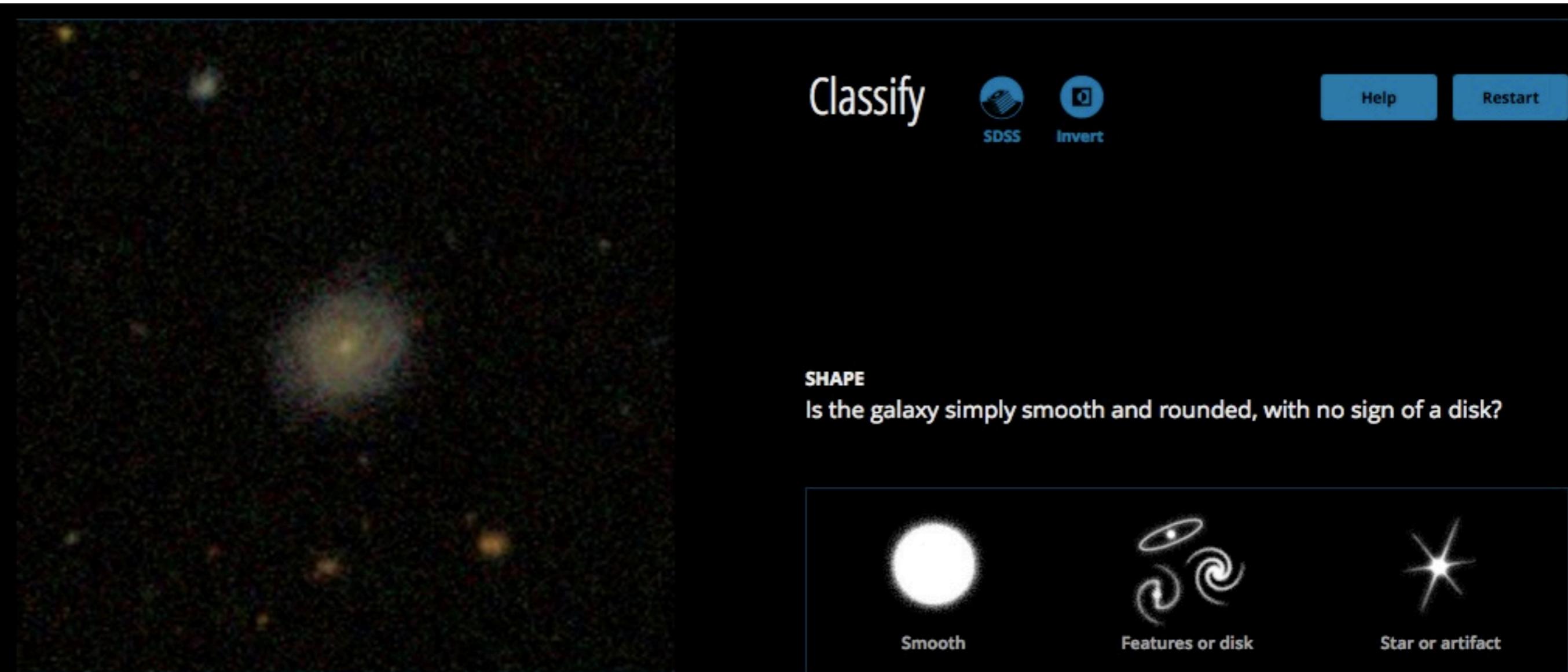


But many more (>4500) papers have been written using SDSS data by others.



**Lesson 2:** Providing data to the community in an easy to use way allows **much** more science to be done - and often science you didn't think of.

# SDSS lessons



The screenshot shows the SDSS Classify interface. On the left is a large image of a galaxy. On the right, there is a navigation bar with the word "Classify" and two icons: "SDSS" and "Invert". Further right are two buttons: "Help" and "Restart". Below the navigation bar, the text "SHAPE" is followed by the question "Is the galaxy simply smooth and rounded, with no sign of a disk?". At the bottom, there are three options with corresponding icons: "Smooth" (a white circle), "Features or disk" (a white spiral galaxy), and "Star or artifact" (a white star with diffraction spikes).

**Lesson X:** Publicly available data is a major boon for Citizen science projects - but this is an entirely separate endeavour from survey science,

# SDSS lessons

## Space

Sort by



### How do galaxies form?

NASA's Hubble Space Telescope archive provides hundreds of thousands of galaxy images.

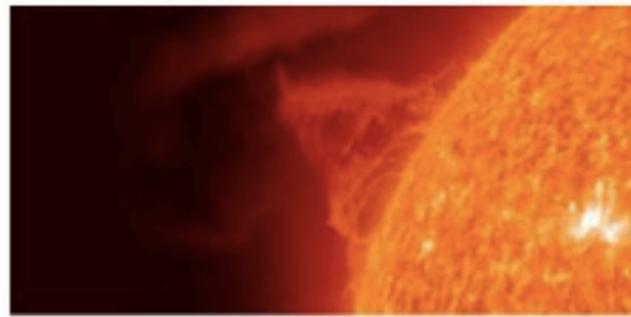
GALAXY ZOO



### Explore the surface of the Moon

We hope to study the lunar surface in unprecedented detail.

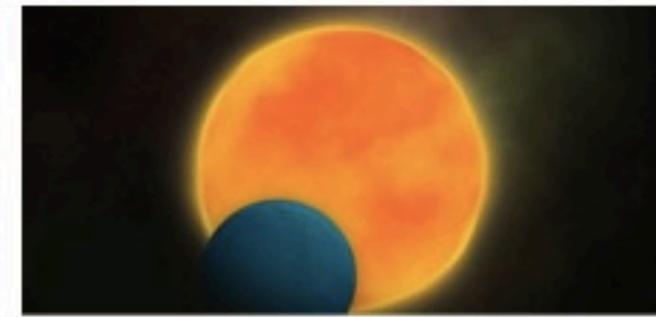
MOON ZOO



### Study explosions on the Sun

Explore interactive diagrams to learn out about the Sun and the spacecraft monitoring it.

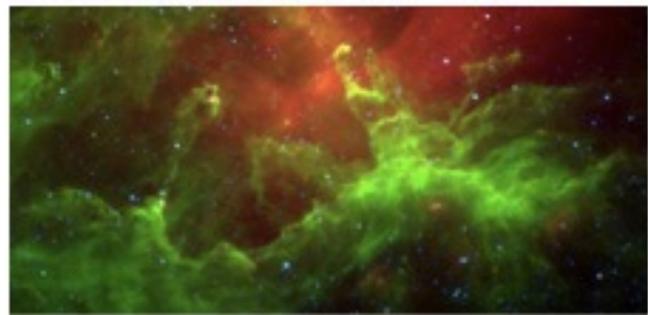
SOLAR STORMWATCH



### Find planets around stars

Lightcurve changes from the Kepler spacecraft can indicate transiting planets.

planethunters.org



### How do stars form?

We're asking you to help us find and draw circles on infrared image data from the Spitzer Space Telescope.

THE MILKY WAY PROJECT



### We're on a collision course with Andromeda

Help researchers understand the awesomeness of the Andromeda galaxy, because one day we'll be in it...

THE ANDROMEDA PROJECT

# Types of data & approaches to take

## Massive surveys:

Data rates are often so large (e.g. LOFAR - Tb/s) that a dedicated facility (database, archive) must be constructed by the survey. This can be a major undertaking, but can be tailored to the data.

## Smaller projects:

The overhead in providing the data for the community, and hosting it for a long time can be prohibitive for individuals/small groups.

What can be done? One solution: Dedicated archives/data centres that deal with this.

# General archiving facilities in astronomy

Infrastructure operated by some groups that aim to archive catalogues, tables etc (primarily). A main European hub is Centre de Données astronomiques de Strasbourg (CDS)

Portal Simbad VizieR Aladin X-Match Other Help

### VizieR Service

Find catalogs among 10731 available

Clear  Find...

Expand search

*?* Catalog, author's name, word(s) from title, description, etc. e.g.: AGN, Veron, I/239, or bibcodes...

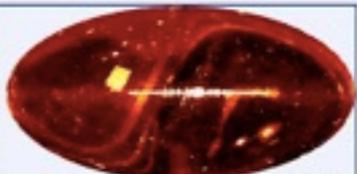
Search for catalogs by column descriptions (UCD) *?*

Search by Position across 11197 tables

Target Name (resolved by [Sesame](#)) or Position:  J2000  2  arcmin  Go!

Radius  Box size

Wavelength	Mission	Astronomy
Radio	AKARI	AGN
IR	ANS	Abundances
optical	ASCA	Ages
UV	BeppoSAX	Associations
EUV	CGRO	Atomic_Data
X-ray	COBE	BL_Lac_objects
Gamma-ray	Chandra	Binaries:cataclysmic

 Find Catalogs 

[More about VizieR](#)

Browsing modes: [Designation](#), [Acronyms](#), [Favorites](#), [Dates](#), [Image spectra](#), [Kohonen](#)  
Or list [the large surveys](#)

→ Thanks for acknowledging the VizieR Service

© UDS/CNRS  
[Contact](#)

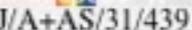
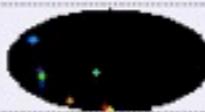
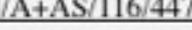
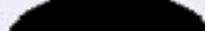
# General archiving facilities in astronomy

Infrastructure operated by some groups that aim to archive catalogues, tables etc (primarily). A main European hub is Centre de Données astronomiques de Strasbourg (CDS)

**Catalog Selection Page**

Looking for catalogs .. ■■■■■  
358 catalogs found having potential matches

ALL   or

<input type="checkbox"/>	 	<input type="checkbox"/>	(c) 36	M31 CO(2-1) spectra (Melchior+, 2012)	<a href="#">spectrum</a>	<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	502	Long slit spectroscopy in M31 (Saglia+, 2010)	<a href="#">profile</a>	<a href="#">Objects</a>	<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>
<input type="checkbox"/>		<input type="checkbox"/>	(c) 1k	A Survey of HII Regions in M31 (Pellet+ 1978)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 4	Colors and extinction across the disk of M31 (Montalto+, 2009)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	1k	M31 globular clusters photometry (Barmby+, 2000)	<a href="#">Objects</a>	<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 155	AGNs and QSOs behind nearby galaxies (Crampton+, 1997)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	695	M31 UBVRI photometry (Hodge+, 1988)	<a href="#">Objects</a>	<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	210	M31 globular clusters ellipticities (Staneva+, 1996)	<a href="#">Objects</a>	<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 518k (density 170)	M31 and M33 UBVRI photometry (Massey+, 2006)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 606k (density 170)	UBVRI phot. in seven Local Group dwarfs galaxies (Massey+, 2007)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 23k (density 168)	WeCAPP Survey. M31 variables (Fliri+, 2006)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 24M (density 75)	2MASS 6X Point Source Working Database / Catalog (Cutri+ 2006)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	
<input type="checkbox"/>		<input type="checkbox"/>	(c) 47k (density 74)	BVRI CCD photometry in the field of M 31 (Magnier+ 1992)		<a href="#">Similar Catalogs</a>	<a href="#">ReadMe+ftp</a>	

# General archiving facilities in astronomy

Infrastructure operated by some groups that aim to archive catalogues, tables etc (primarily). A main European hub is Centre de Données astronomiques de Strasbourg (CDS)

This provides storage of tables and results from published papers. In some cases (too few!), the data are automatically harvested from the journal - otherwise the authors need to provide the data.

This is now integrated into other software so I can look at an image and then ask to get an overview of all data in the literature for that region of the sky.

In principle a similar facility could be constructed for very many research areas - but you do need to establish the centre and then get funding.

# The Virtual Observatory

International effort to provide access to all astronomical data. You could use this to “observe” any patch of the sky from your computer.

Most up-coming surveys will be ‘VO-compatible’ and so should be able to make use of this.

In principle very appealing - but exchange of data requires standards to be adhered to. For images and spectra this was mostly solved long time ago, but tables required some effort. Software essential but often written by temporary staff and not maintained/not always what astronomers need.

**Lesson 3:** Generic facilities are **very useful**, but challenging to set up - standards must be defined and software development and maintenance must be funded.

# Lessons learned

- *Useful* data sharing is not free - it has a cost and this must be judged against the benefits.
  - Big projects must have helpdesks, long-term storage facilities etc.
- Data for public use usually require **higher** quality standards than for internal use (but this is not a bad thing!).
- You can be scooped - but if your data have wide applicability this is unlikely to be detrimental to science overall. But some protection mechanism (proprietary period, fencing off for PhD projects) might be necessary.
- **More users means more science - true for all astronomical surveys. And more checks on the results!**
- Dumping data on a web site is (almost) a waste of everyone's time.

# **1+1 > 2 - Will inclusion of other institutes' data boost your research?**

- To get a full understanding of an astronomical object we need data over much of the electromagnetic spectrum.
- Combination of surveys (e.g. SDSS + radio surveys) can yield a vast amount of additional information.
- The combined data has more information than each on their own.
- When surprising claims are made, the ability to check these using data obtained on different telescopes, analysed using different pipelines, is essential - access to other people's (raw) data is necessary.

# What about loosing your competitive edge?

## Personal experience:

If you produce a useful dataset, then you will be the expert in the world on it. You might not be able to do all the science you want to do, but you will be well cited.

We created something called the MPA-JHU value added catalogue - a catalogue of physical parameters of galaxies.

This led to ~4 main papers (more later, as well as follow-up projects), that each are cited >700 times and a couple >1000.

By not hoarding the data and hogging the science, all kinds of groups were encouraged to pick it up and work with it. Yet the effort put into preparing the release ensured we were ahead of the competition.

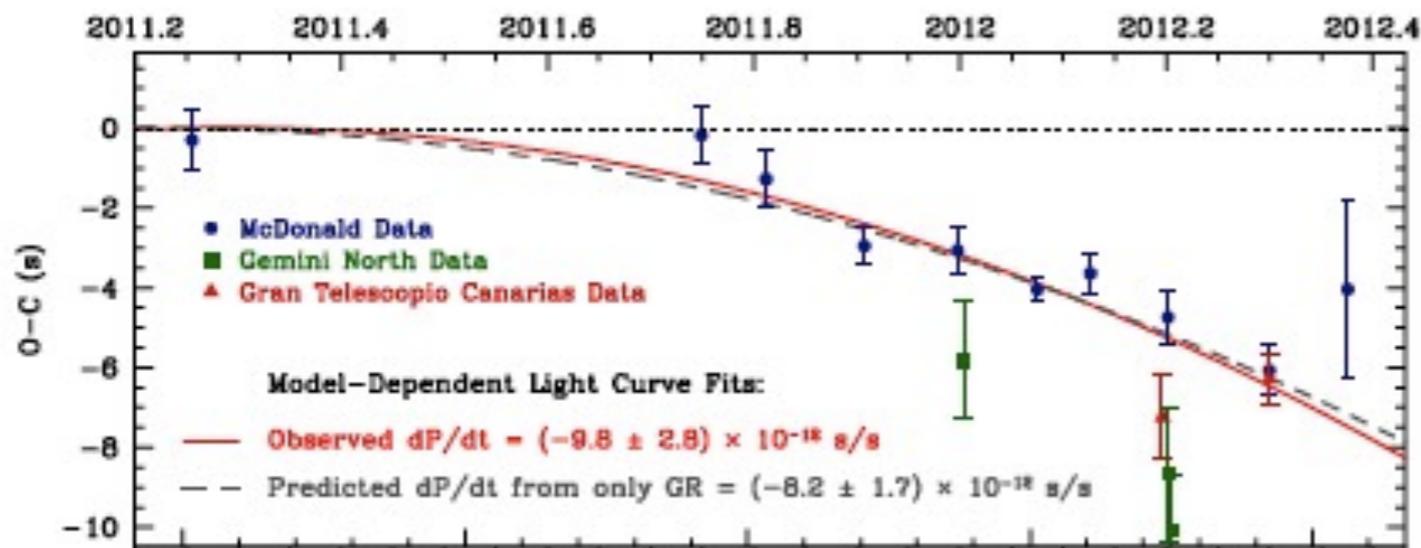
# What about loosing your competitive edge?

But sometimes astronomers do not share immediately:

The binary White Dwarf J0651+2844



By monitoring the orbital period of such a system, one can constrain general relativity - but to do that you need to observe it over time.



Hermes et al (2012)

This is easy to do, so the discoverers did not publish the coordinates immediately when they found it - they now have a head-start, but only by making their data public will people believe their findings.

# Doing bad science with your data

It happens.

But **bad science is the fault of the scientist** - not the data. You have to provide documentation for your data but as long as that is good (=most people can use your data to do good science), then in my opinion you just have to live with people doing bad science.

And if your data is generally available other researchers can **check** the results - and they *will* if it is sufficiently interesting (if it is bad and uninteresting, do you really care?)

# **Benefits from sharing your data**

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.
3. Shared and used data means better checks on your data.

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.
3. Shared and used data means better checks on your data.
4. Large, public datasets might spur cross-disciplinary collaboration (e.g. Computer Science/Statistics).

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.
3. Shared and used data means better checks on your data.
4. Large, public datasets might spur cross-disciplinary collaboration (e.g. Computer Science/Statistics).
5. Citation rates usually go up, students/post-docs that worked on the project become attractive to others.

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.
3. Shared and used data means better checks on your data.
4. Large, public datasets might spur cross-disciplinary collaboration (e.g. Computer Science/Statistics).
5. Citation rates usually go up, students/post-docs that worked on the project become attractive to others.
6. Open access makes citizen science projects/education much easier [e.g. Galaxy Zoo, Google Sky etc]

# Benefits from sharing your data

1. Reproducibility of science - the default must be that any scientific result is reproducible by others (eventually).
2. More science from the same data.
3. Shared and used data means better checks on your data.
4. Large, public datasets might spur cross-disciplinary collaboration (e.g. Computer Science/Statistics).
5. Citation rates usually go up, students/post-docs that worked on the project become attractive to others.
6. Open access makes citizen science projects/education much easier [e.g. Galaxy Zoo, Google Sky etc]
7. It is the right thing to do: Beneficial for poorer countries, the ones that paid for it (indirectly) can access the results, ...

# Questions to ask

- Who will use the data - are there enough of them?
- What effort is needed to make the data **easily** accessed/used?
  - Data that are hard to access will only be used in rare cases, so no citation gain, no boost to the area of study.
- Who will pay for the initial setup & for the maintenance?
  - It costs money/effort to set up a good data repository - and even keeping it spinning costs money (electricity, hosting costs, management when web servers/computer infrastructure is upgraded).
- Is there an international facility that can host your data?

# What should you share?

## Small projects:

Provide tables & final data in a convenient electronic format.  
Raw/unanalysed data should be possible to provide.

## Large projects:

Plan for release of data from the start of the project.  
Identify what the potential users and what they would need.  
Provide data through a data management infrastructure  
(typically a web page + database & data server backend)  
Do not forget to have a system for quality check - it will  
benefit yourself!

## Both: Software

Don't forget that often the software you use is a crucial  
ingredient in your research - make it public when you can!

# Thank you

- ADS - integrated interface to most astronomical & related physics literature (back to ~1850s).
  - <http://www.adsabs.harvard.edu/>
- NASA/IPAC Extragalactic Database - links data, publications, images ++
  - <http://ned.ipac.caltech.edu/>
- CDS data archive++
  - <http://cdsweb.u-strasbg.fr/>
- The SDSS web pages
  - <http://cas.sdss.org/astrodr7/en/> <http://www.sdss3.org/>
- The NASA multi-mission archive (space mission primarily)
  - <http://archive.stsci.edu/>
- ESO archive
  - <http://archive.eso.org/cms.html>
- Google Sky - Worldwide Telescope
  - <http://www.google.com/sky/>
  - <http://www.worldwidetelescope.org/home.aspx>