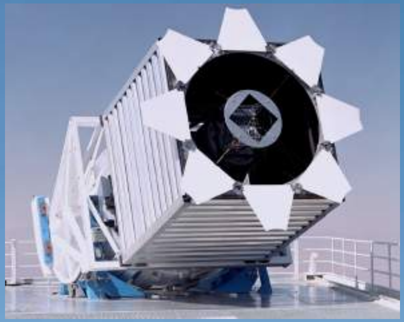
Big Data Sets in Astronomy

Željko Ivezić, University of Washington

LSST



SDSS

Sao Paulo School of Advanced Science on Learning from Data, July 31 - Aug 2, 2019

Main Topics:

Day 1: Introduction
- who I think you are?
- who I am?
- why do astronomers need Big Data?
- Large Synoptic Survey Telescope: Big Data!
- astroML

Day 2: Density Estimation, Clustering and Classification in Astronomy

Day 3a: Dimensionality reduction, Regression and Time Series Analysis in Astronomy

Day 3b: Schedule reserve and free-form discussions

1) Introduction

- who I think you are: "About 200 computer science graduate students who do python"

I am assuming that you like astronomy but didn't take (m)any college-level classes. Therefore, today I am only going to provide astronomical context for Big Data.

I will talk about astronomical Big Data analysis in more detail tomorrow and the third day.

But first I need to ask you a few questions (to help me optimize Days 2 and 3)…

Please raise your hand if:

- you are a computer-science graduate student
- you ever took a college-level astronomy class
- you are a python user
- you used jupyter (ipython) notebooks
- you used SQL language and databases
- you did quantitative model parameter estimation (e.g. fitting a gaussian to a histogram, or fitted a straight line to y(x) data)
- you are familiar with Bayesian statistics
- you used any clustering algorithm
- you used any classification algorithm
- you did time series analysis (e.g. Fourier analysis)

• Some tools and methods…

- o Correlation coefficients (many dimensions, missing data)
- o The bootstrap and the jackknife methods
- o Maximum Likelihood Method
- o The goodness of fit and model selection
- o Bayesian statistics
- o Markov Chain Monte Carlo methods
- o Regression (“fitting”, LSQ, outliers, regularization)
- o Density estimation (“multi-dimensional histograms”)
- o Clustering (kernel, parametric)
- o Classification (supervised and unsupervised, active learning)
- o Dimensionality Reduction (PCA, ICA, LLE and friends)
- o Time-series analysis (periodogram, stochastic processes)

These topics are covered in lectures available at
https://github.com/dirac-institute/uw-astr598-w18

• Some tools and methods…

- o Correlation coefficients (many dimensions, missing data)
- o The bootstrap and the jackknife methods
- o Maximum Likelihood Method
- o The goodness of fit and model selection
- o Bayesian statistics
- o Markov Chain Monte Carlo methods
- o Regression (“fitting”, LSQ, outliers, regularization)
- o Density estimation (“multi-dimensional histograms”)
- o Clustering (kernel, parametric)
- o Classification (supervised and unsupervised, active learning)
- o Dimensionality Reduction (PCA, ICA, LLE and friends)
- o Time-series analysis (periodogram, stochastic processes)

My main goal for these lectures: to give you a taste of the use of the last six methods in astronomy.

1) Introduction

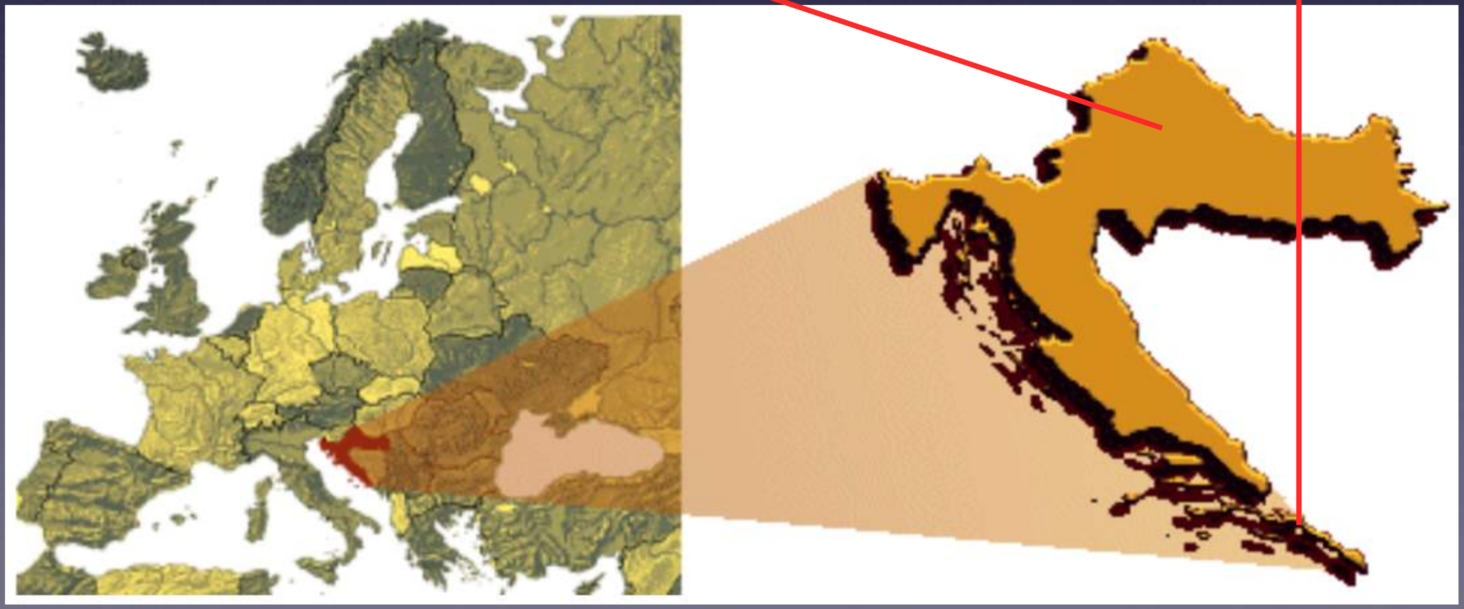
- who I am: a professor of astronomy, a former software (pipeline) developer for the Sloan Digital Sky Survey (SDSS), and the Project Scientist and Deputy Director for the Large Synoptic Survey Telescope (LSST) project (more details about LSST later today).

My interest in Big Data comes from my work with the SDSS data (more details later today). This work led to me teaching related courses with a number of colleagues, and then we turned our lectures into a textbook, with worked-out open-source examples coded in python, available as astroML.

Disclaimer: I am only an astronomer, not a computer scientist!

Zagreb

Dubrovnik



Europe

Croatia

World Cup 2018: silver medal!

France was better in the final game. Congratulations!

World Cup 2018: silver medal!

1) Introduction
- why do astronomers need Big Data?

- What is astronomy about?
  - search for life elsewhere
  - understanding the Universe

Generally speaking, astronomy (or astrophysics - but not astrology!) studies the formation and evolution of structure in the Universe (we apply laws of physics to observations).

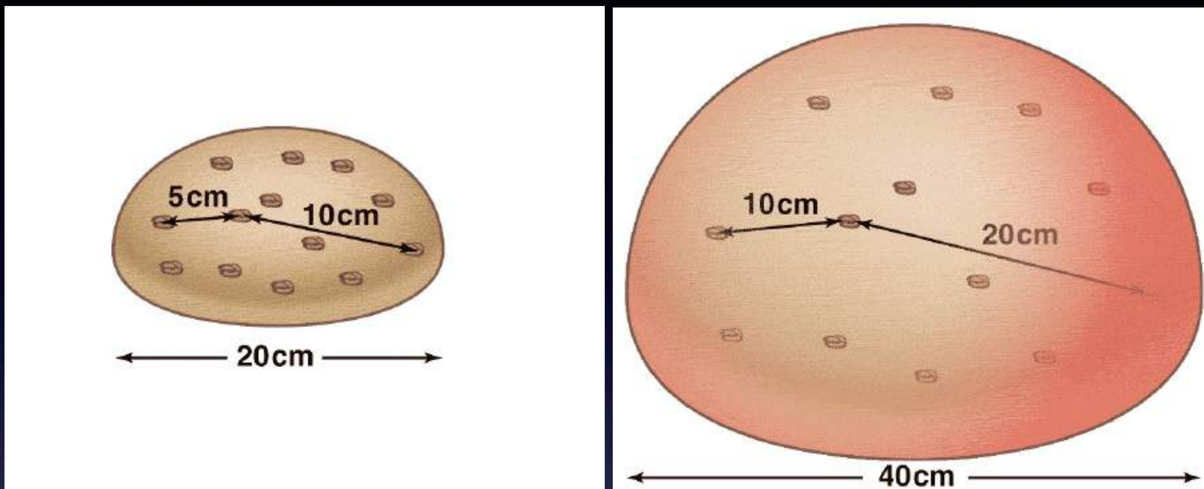
1) Introduction

- why do astronomers need Big Data?

- What is astronomy about?
  - search for life elsewhere
  - understanding the Universe

Over the last three of decades, astronomers have discovered about 4,000 extra-solar planets (or exoplanets). These are planets outside of our Solar System, with its 8 planets. It is possible that some of them could support life. Are we alone?

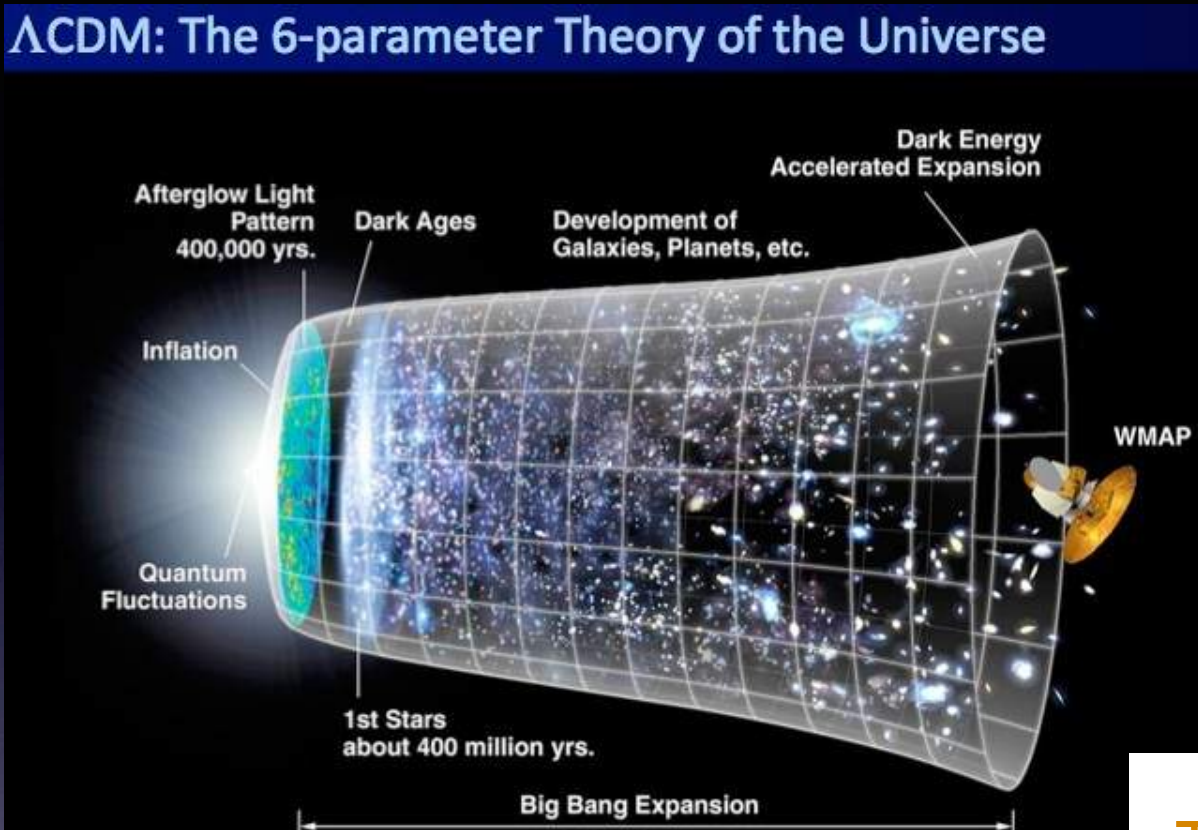
We have known for about 100 years that the Universe is expanding.

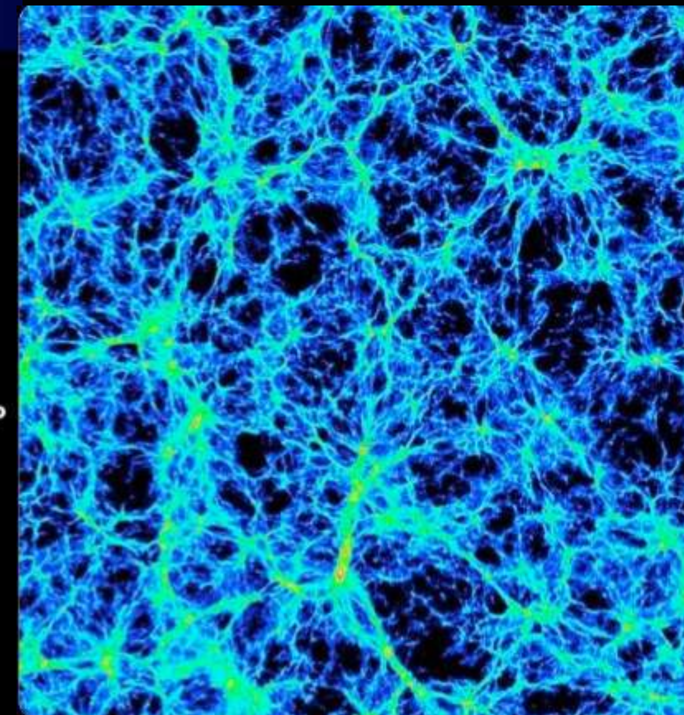


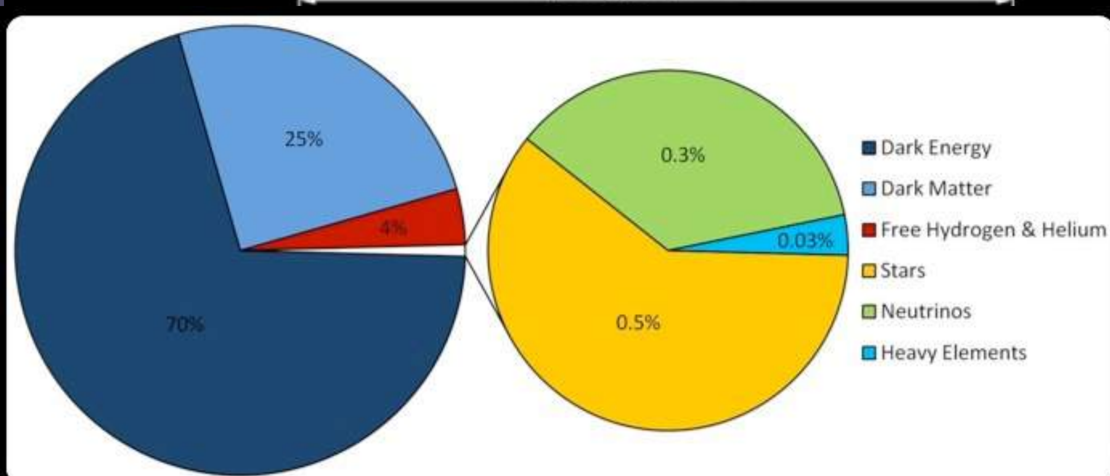
Edwin Hubble (1929)

About a decade ago, it was discovered that this expansion is accelerating. We are uncertain about what this acceleration means; the two most plausible explanations are some mysterious and weird fluid called dark energy, or perhaps Einstein's general theory of relativity fails!

A New Cosmological Puzzle: an Accelerating Universe

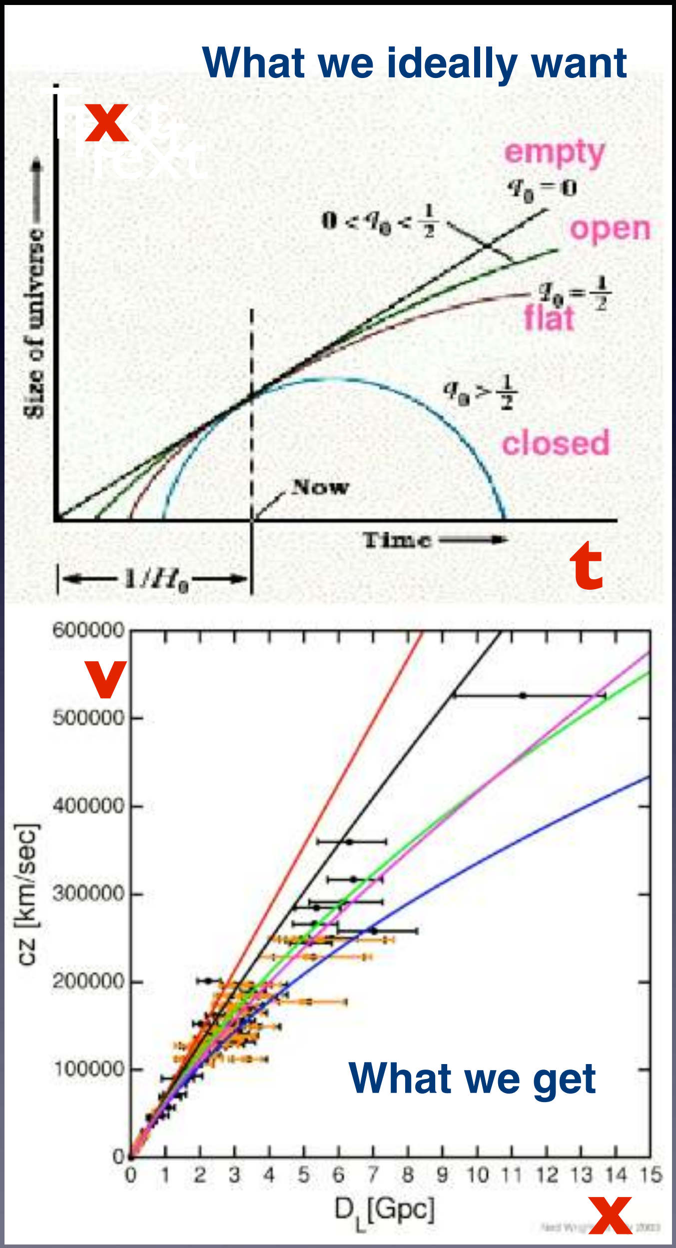






The modern cosmological models can explain all observations, but need to **postulate** dark matter and dark energy (though gravity model could be wrong, too)

How do we measure expansion of the Universe?



Ideally, we'd like to measure the size of the Universe as a function of time, x(t), but we can't.

Instead, we measure the distance to objects, x, and their velocity, v. That is, we have v(x).

And then we use our knowledge of physics (v = dx/dt) and models of the Universe (given what we assume the Universe is made of, how should it expand?) to get x(t) and v(t): **dt = dx / v(x)**

In other words, our knowledge of physics enables us to interpret astronomical measurements using models of the Universe and in turn, understand the makeup and history of the Universe!

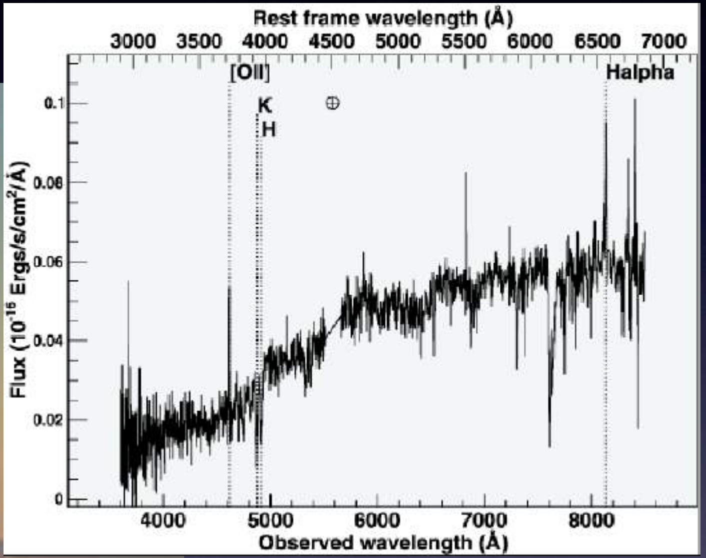
Modern observational methods in astronomy and astrophysics

- **Telescopes above the atmosphere:** high angular resolution (e.g., the Hubble Space Telescope) and other wavelength regions (X-ray, radio, infrared)

The HST in orbit and an example of a galaxy image

Modern observational methods in astronomy and astrophysics

- Large telescopes (~10m): faint objects, especially spectroscopy



The Keck telescopes on Mauna Kea (Hawaii)

Modern observational methods in astronomy and astrophysics

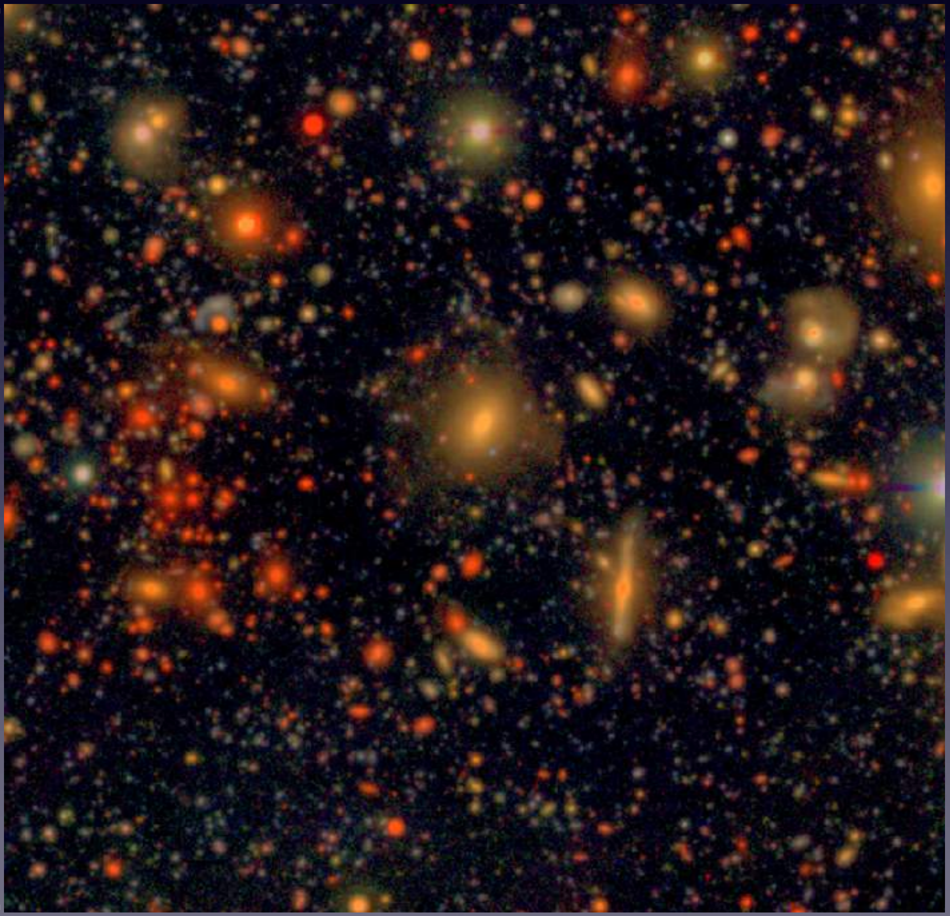
- Large telescopes (~10m): faint objects, especially spectroscopy
- Telescopes above the atmosphere: high angular resolution (e.g., the Hubble Space Telescope) and other wavelength regions (X-ray, radio, infrared)
- Large sky surveys and sky maps: digital sensor technology(CCD: charge-coupled device), information technology (data processing and data distribution)

Key point: modern sky surveys make all their data (images and catalogs) publicly available

- What is astronomy about?
  - understanding the Universe

I work on a project called LSST, that aims to obtain the greatest ever ”movie of the Universe”: the image of the sky will be recorded about 1000 times over 10 years (about 100,000,000 GB of data).

1/10 of Moon’s diameter



LSST will obtain 8 million such images!

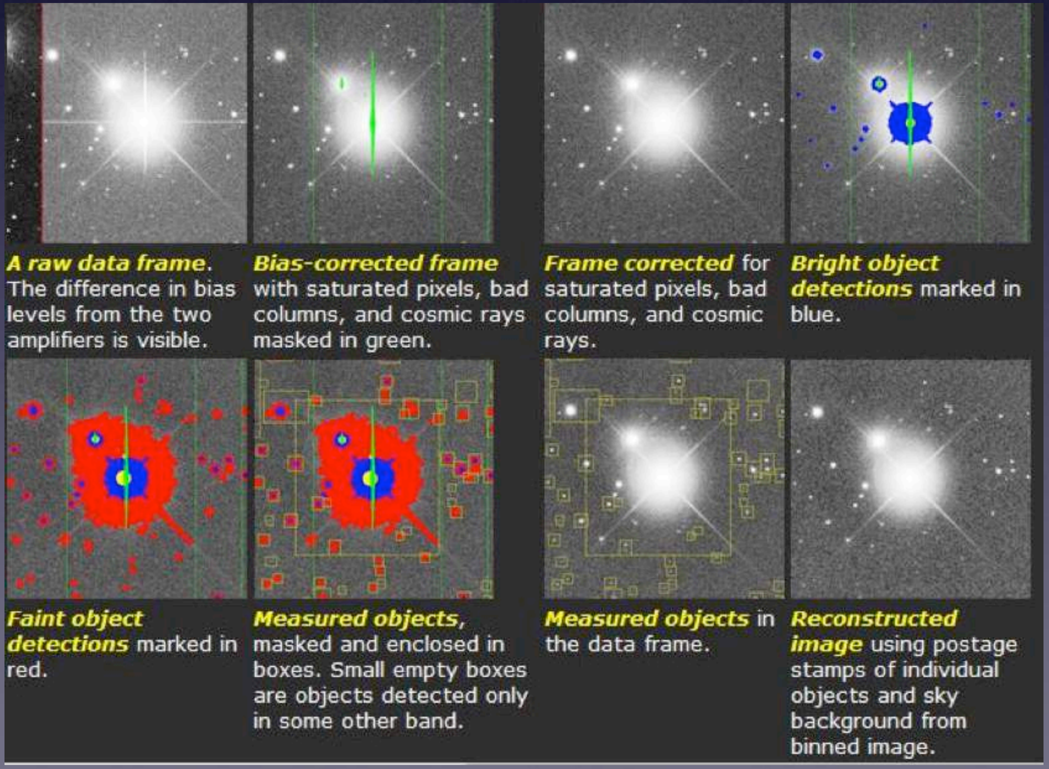
There are about 5,000 objects in this small image; LSST will detect 40 billion objects over half the sky!

What is a sky map?
Why are sky maps useful?

- Sky map:
  - a list of all detected objects (stars, galaxies, ...)
  - measured parameters (size, color, brightness,...)

Basic steps in astronomical image processing (example: Sloan Digital Sky Survey):

All these (complicated) steps are already done: "science-ready database"



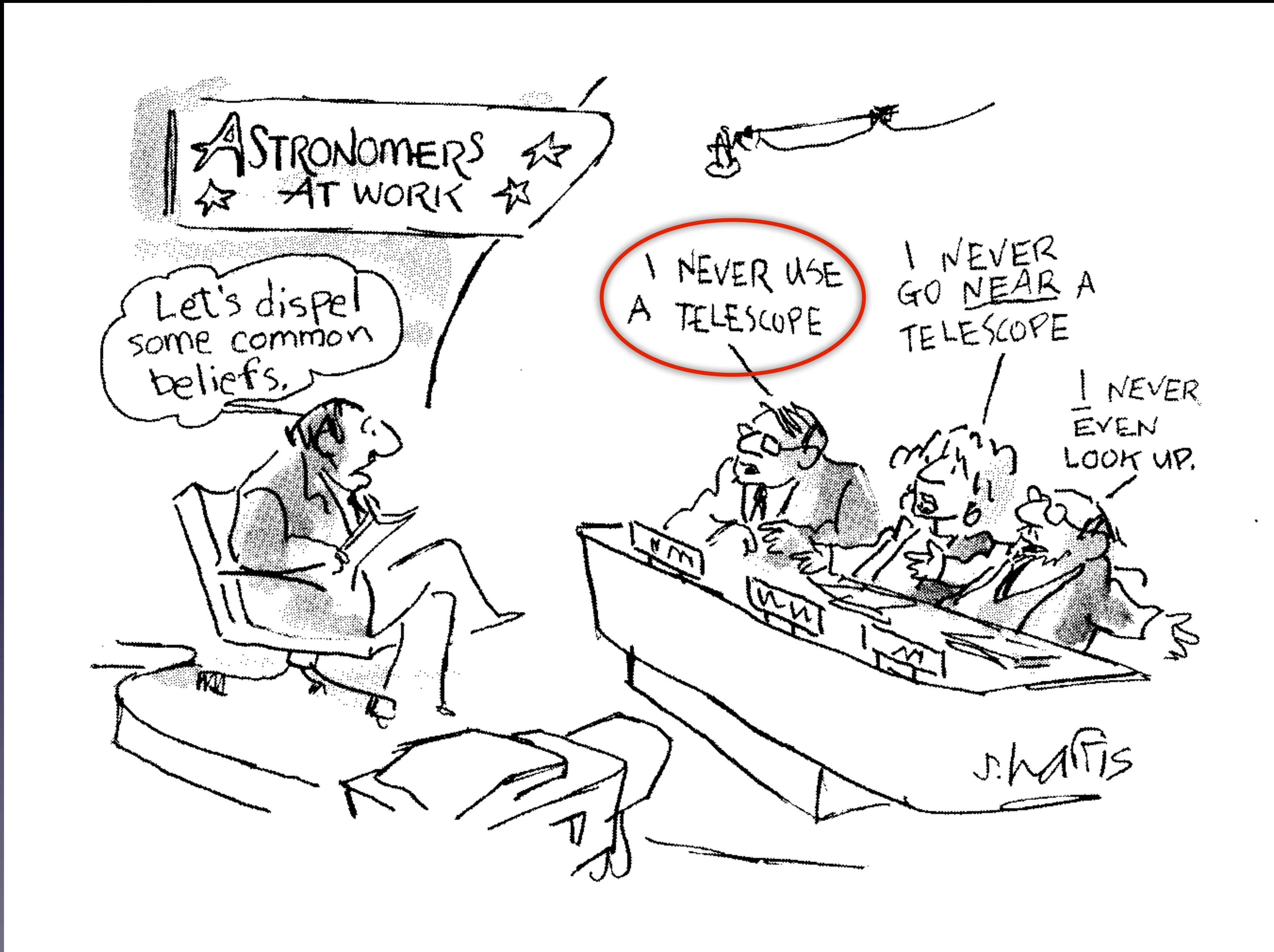
What is a sky map? Why are sky maps useful?

- Sky map:
  - a list of all detected objects (stars, galaxies, ...)
  - measured parameters (size, color, brightness,...)

- The utility of sky maps:

  Discoveries of new objects: "Is this a new asteroid, or is it already cataloged?"
  Object classification: "What types of galaxies exist?"
  Statistical population studies: "Do quasars change their properties with time?"
  Search for unusual objects: "Is this star very weird?"
  Cosmological measurements: "How fast does the Universe expand?"

"Science-ready database": measurements can be (simply) analyzed without the need for (complex) image processing

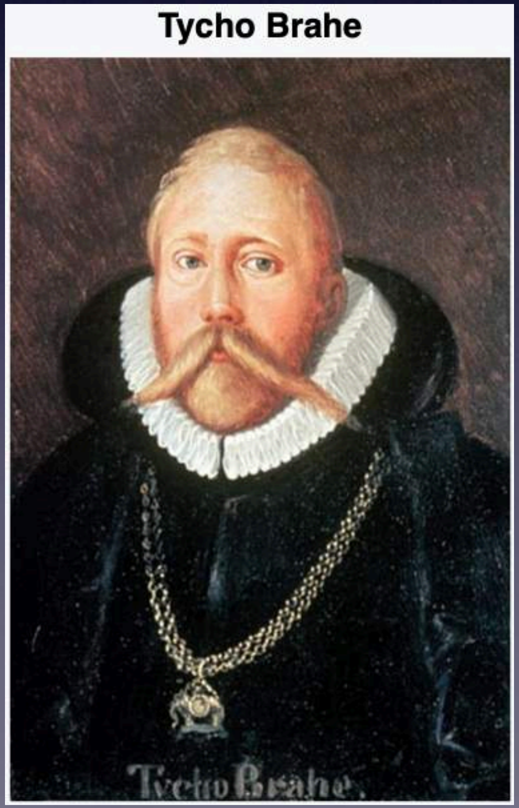


ASTRONOMERS AT WORK
Let's dispel some common beliefs.
I NEVER USE A TELESCOPE
I NEVER GO NEAR A TELESCOPE
I NEVER EVEN LOOK UP.
S. Harris

Short history of sky mapping

- Hipparchos
  - about 3,000 years ago
  - all stars visible from Greece: about 3,000
  - the main source of astronomical measurements for the next 2,500 years!

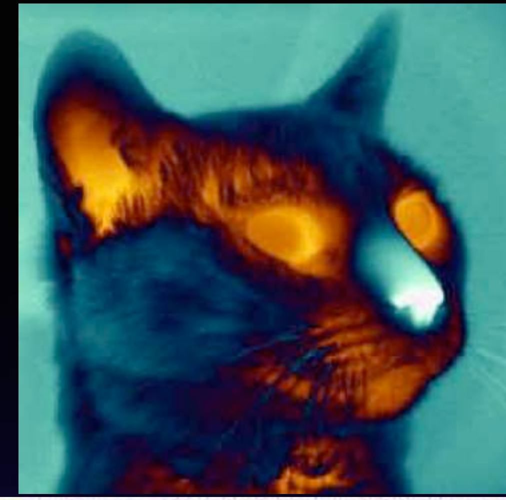
- Tycho Brahe
  - XVI century, much more accurate measurements than Hipparchos
  - still without a telescope: only about 3,000 stars
  - the main results: Kepler's Laws of planetary motions, Newton's theory of gravity

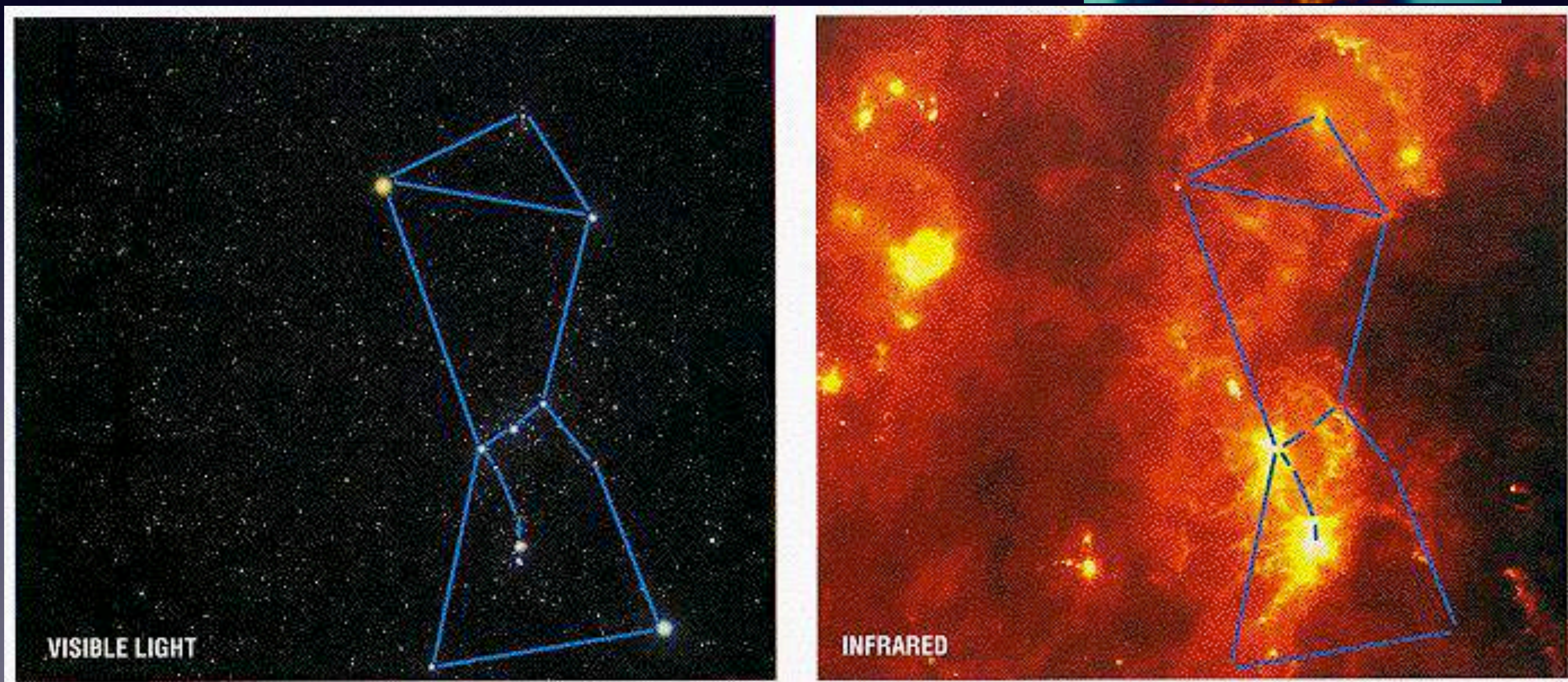


Modern sky mapping

- Palomar Observatory Sky Survey
  (National Geographic Sky Survey):
  – optical wavelengths, two bandpasses
  – 1950-1955 (second phase in 80’s)
  – about 1,000 photographs (whole sky)
- Other wavelengths:
  – X rays (Chandra, XMM-Newton)
  – ultraviolet (GALEX)
  – infrared (2MASS, Spitzer)
  – radio (FIRST, NVSS)

Optical wavelengths reveal only a bit of reality...

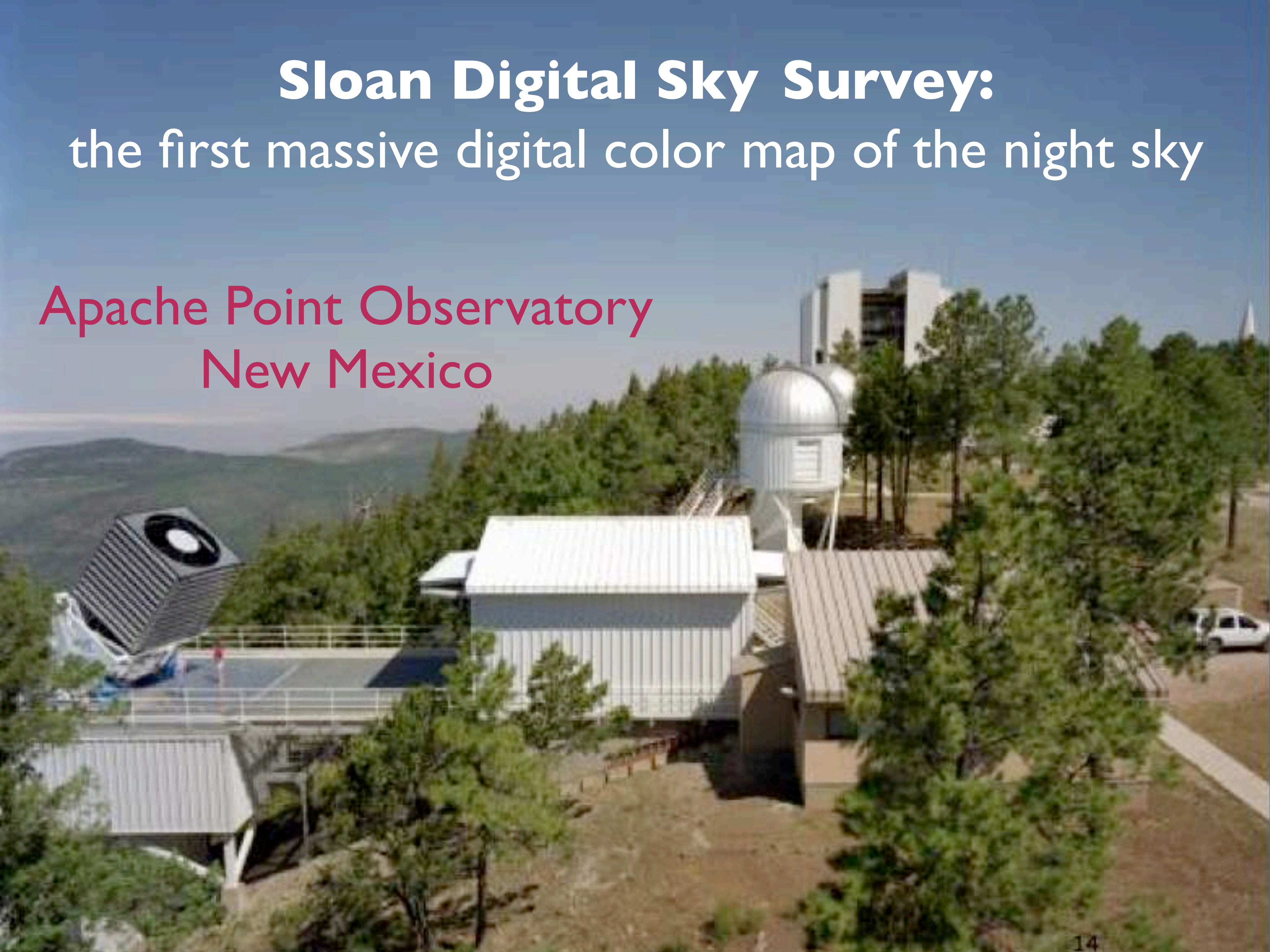




Orion: visible light   infrared light

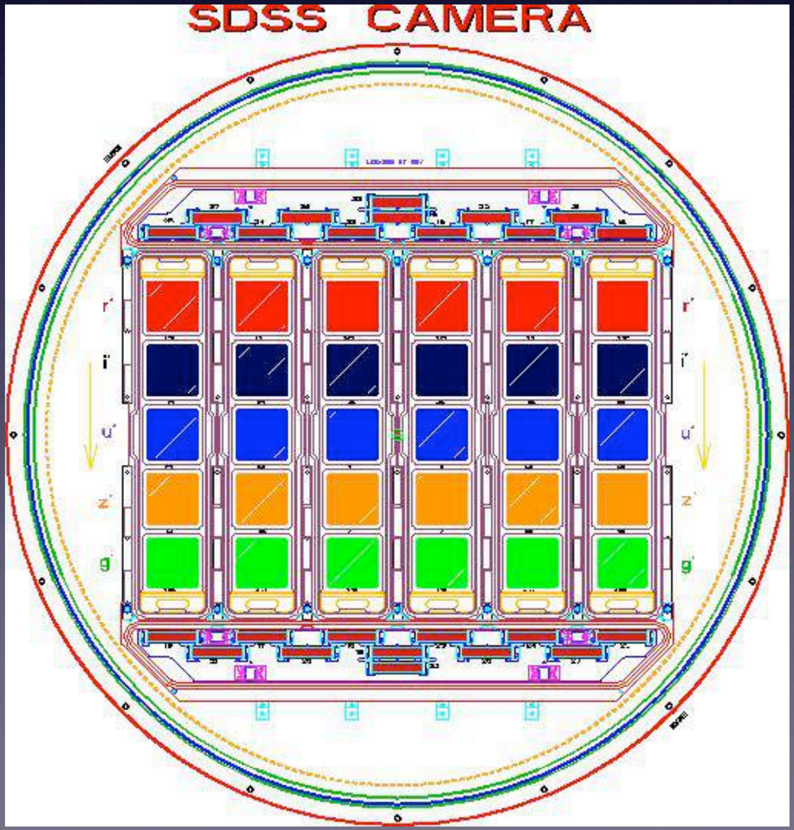
Sloan Digital Sky Survey:

the first massive digital color map of the night sky

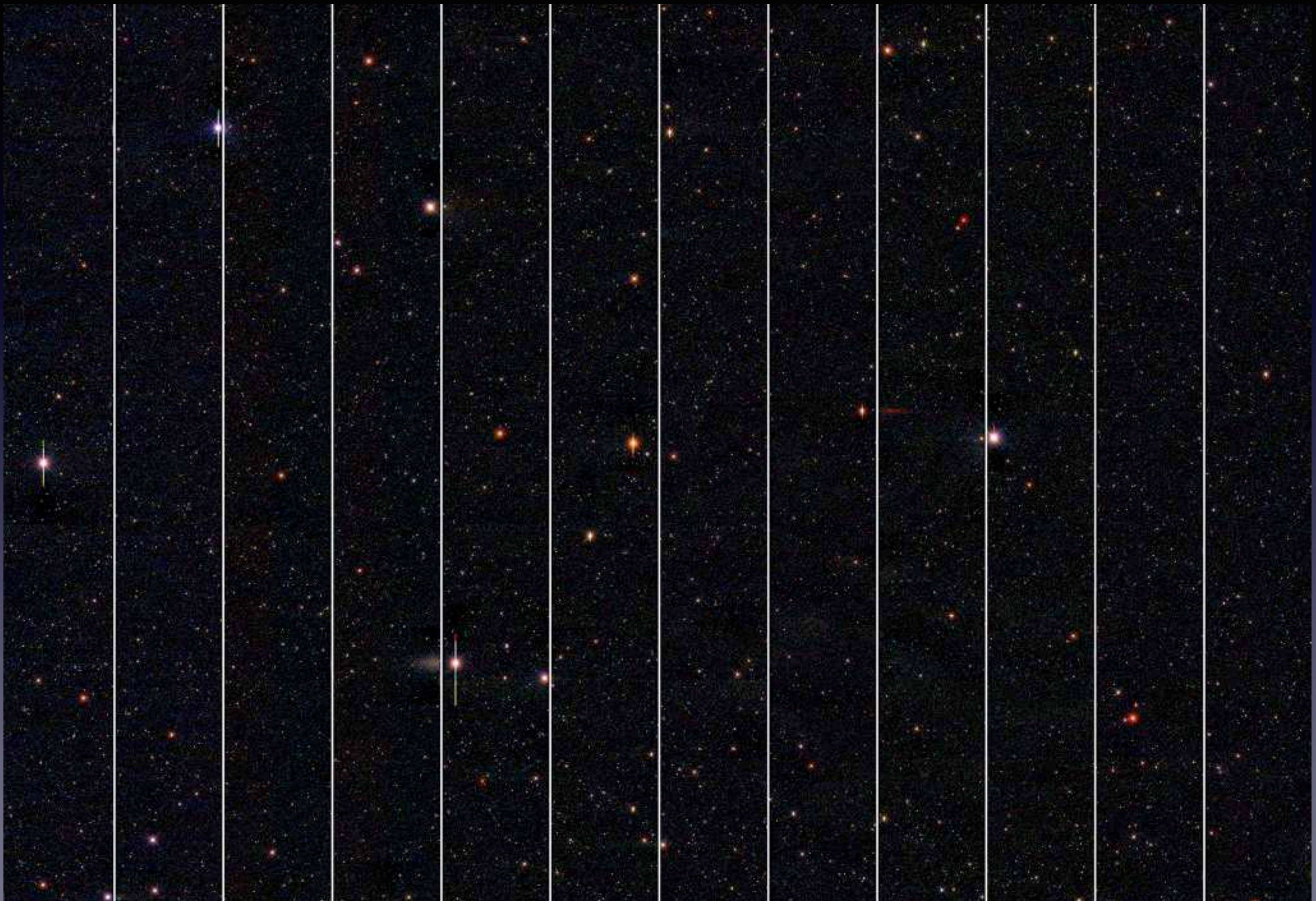


The last decade: Sloan Digital Sky Survey

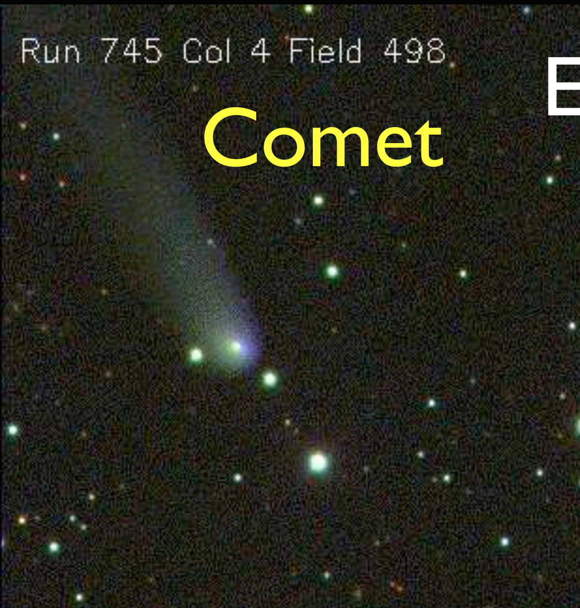
- Digital sky survey with a 120 Megapix CCD camera
- Precise measurements for 400,000,000 objects
- Revolution in astronomy: public databases



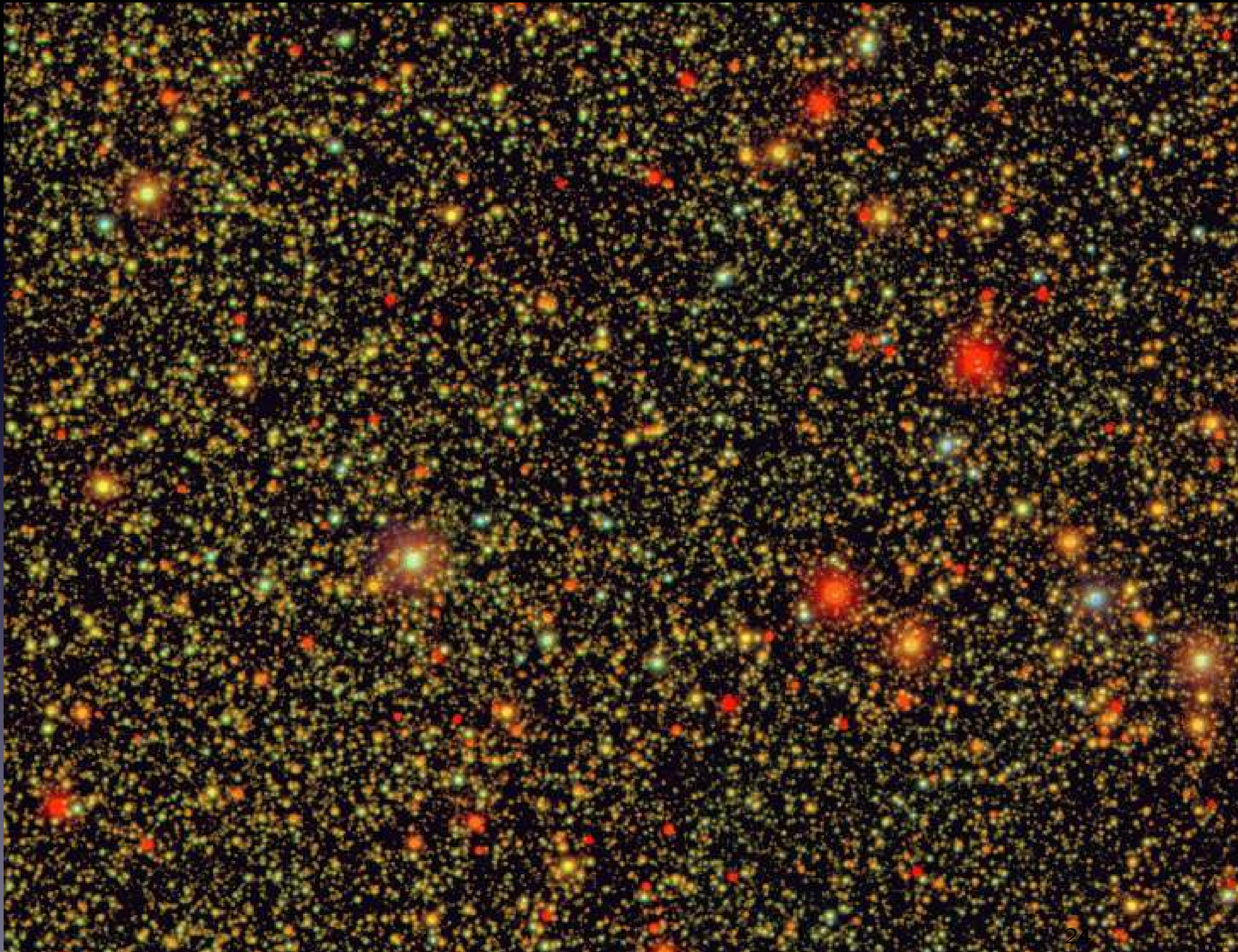
SDSS sky mapping: “drift scanning”



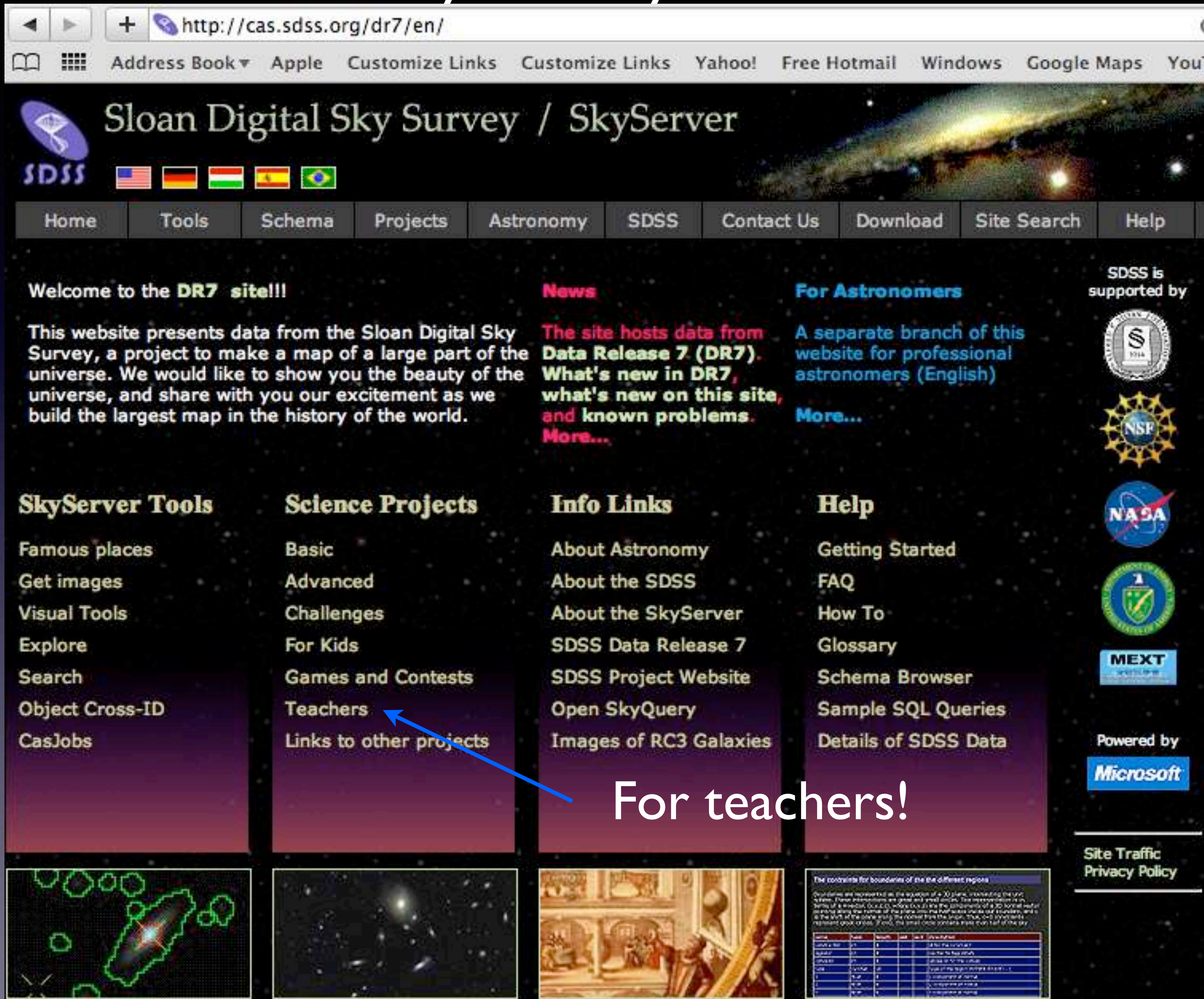
Examples of SDSS images



SDSS view along the Milky Way Disk



Astronomy "from your armchair"



For teachers!

http://cas.sdss.org/dr7/en/tools/search/sql.asp

Address Book ▾ Apple Customize Links Customize Links Yahoo! Free Hotmail Windows Google Maps YouTube

Home | Tools | Schema | Projects | Astronomy | SDSS | Contact Us | Download | Site Search | Help

available to everyone around the world

DR7 Tools

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

SQL Search

This page allows you to directly submit a SQL (Structured Query Language) query to the SDSS database server. You can modify the default query as you wish, or cut and paste a query from the SDSS Sample Queries page.

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.

Clear Query

```
-- This query does a table JOIN between the imaging (PhotoObj) and spectra
-- (SpecObj) tables and includes the necessary columns in the SELECT to upload
-- the results to the DAS (Data Archive Server) for FITS file retrieval.
SELECT TOP 10
  p.objid,p.ra,p.dec,p.u,p.g,p.r,p.i,p.z,
  p.run, p.rerun, p.camcol, p.field,
  s.specobjid, s.specClass, s.z,
  s.plate, s.mjd, s.fiberid
FROM PhotoObj AS p
  JOIN SpecObj AS s ON s.bestobjid = p.objid
WHERE
  p.u BETWEEN 0 AND 19.6
  AND g BETWEEN 0 AND 20
```

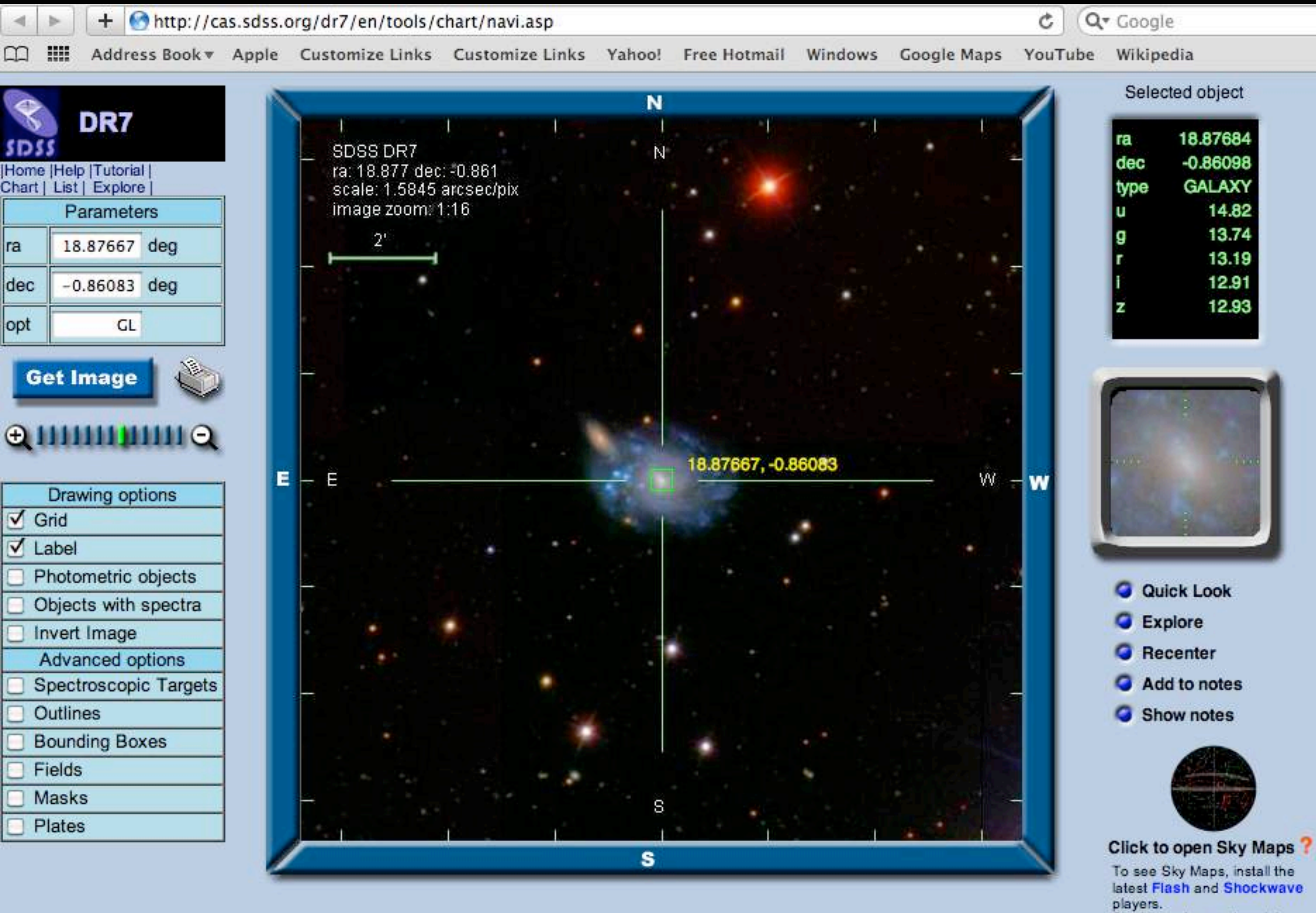
Submit | Check Syntax Only? | **Output Format** HTML XML CSV | Reset

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

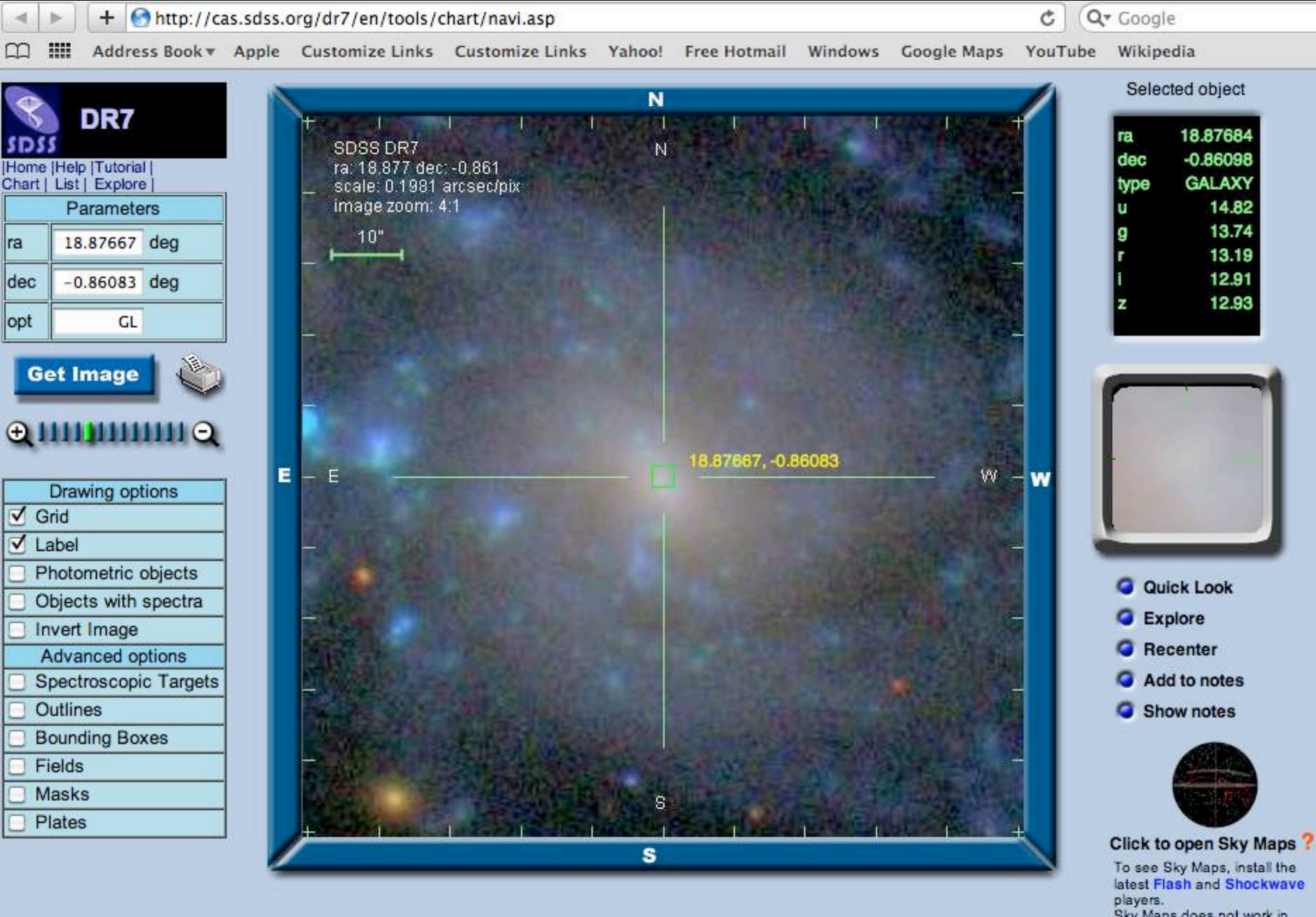
For an introduction to the Structured Query Language (SQL), please see the Searching for Data How-To tutorial. In particular, please read the Optimizing Queries section.

The inclusion of the imaging and spectro columns for DAS upload in your query (as in the default query on this page) will ensure that when you press **Submit**, the appropriate button(s) are displayed on the query results page to allow you to upload the necessary information to the DAS to retrieve the FITS file data corresponding to your CAS query. The imaging columns needed for upload to the DAS are *run*, *rerun*, *camcol*, and *field*. The spectroscopic columns needed are *plate*, *mjd*, *fiberid*, and optionally *sprerun* (the latter requires a join with the PlateX table).

“Navigation” around the sky...



"Navigation" around the sky: zoom in, zoom out...



Additional, more detailed, information...

http://cas.sdss.org/dr7/en/tools/explore/obj.asp?ra=18.87667&dec=-0.86083

Address Book ▾ Apple Customize Links Customize Links Yahoo! Free Hotmail Windows Google

DR7 SDSS

Explore Home

Search by
ObjId
Ra,dec
5-part SDSS
Plate-MJD-Fiber
SpecObjId

Summary

PhotoObj
PhotoTag
More Observations
Field
Frame
PhotoZ
Neighbors
Finding chart
Navigate
FITS

SpecObj
All Spectra
SpecLine
SpecLineIndex
XCredShift
ELredShift
Spectrum
Plate
FITS

NED search
SIMBAD search
AKARI FIS
AKARI IRC
ADS search

Notes
Save in Notes
Show Notes

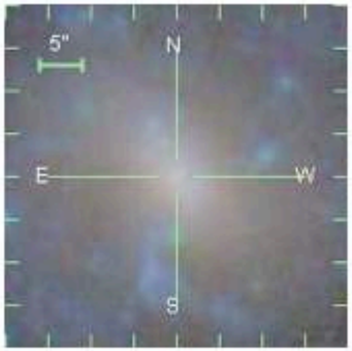
Print

SDSS J011530.44-005139.5

GALAXY ra=18.87683906, dec=-0.86097998, ObjId = 587731511532060697

Column names link to glossary entries. Move mouse over a column name to get its units.

mode	PRIMARY
status	TARGET PRIMARY OK_STRIPE OK_SCANLINE PSEGMENT RESOLVED OK_RUN GOOD SET
flags	DEBLEND_DEGENERATE BAD_MOVING_FIT BINNED1 INTERP COSMIC_RAY NOPETRO CHILD
PrimTarget	TARGET_GALAXY TARGET_GALAXY_RED TARGET_QSO_CAP
SecTarget	



u	g	r	i	z
14.82	13.74	13.19	12.91	12.93

err_u	err_g	err_r	err_i	err_z
0.01	0.00	0.00	0.00	0.00

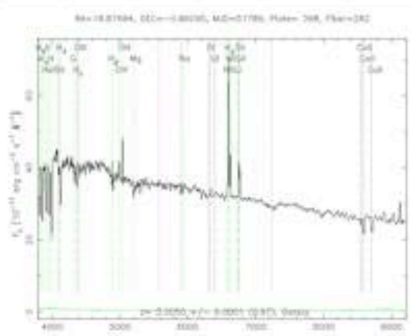
run	rerun	camcol	field	obj	rowc	colc
2738	40	1	44	25	972.5	1786.6

fiberMag_r	petroMag_r	devMag_r	expMag_r	psfMag_r	modelMag_r
17.56	12.97	13.14	13.19	18.16	13.19

extinction_r	petroRad_r	parentId	nChild
0.11	106.724	587731511532060693	0

SpecObjID = 112249473974927360

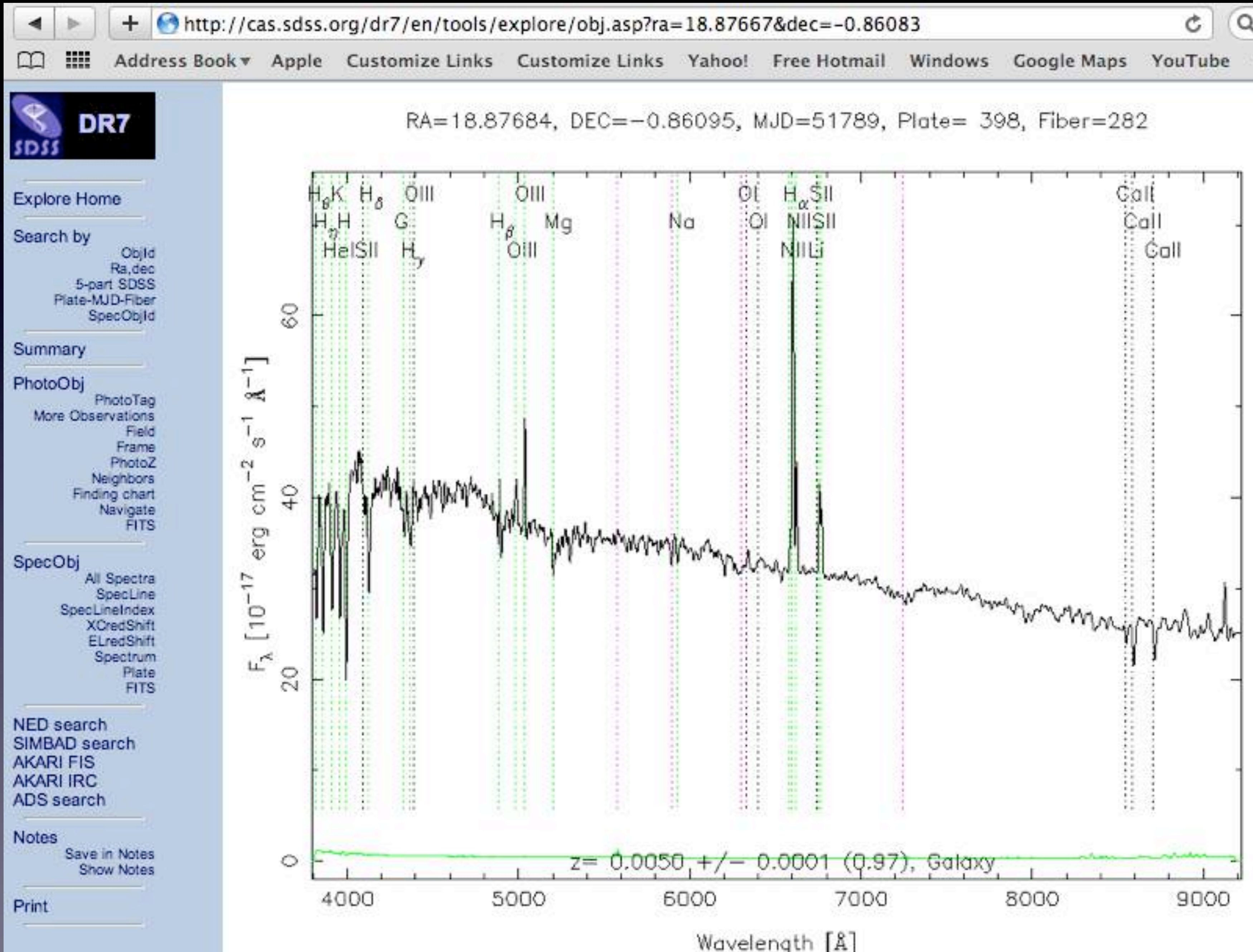
plate	mjd	fiberId	z	zErr	zConf	specClass	ra	dec	fiberMag_r	objId
398	51789	282	0.005	0.00006	0.969081	GALAXY	18.87684	-0.86095	17.53	587731511532060697



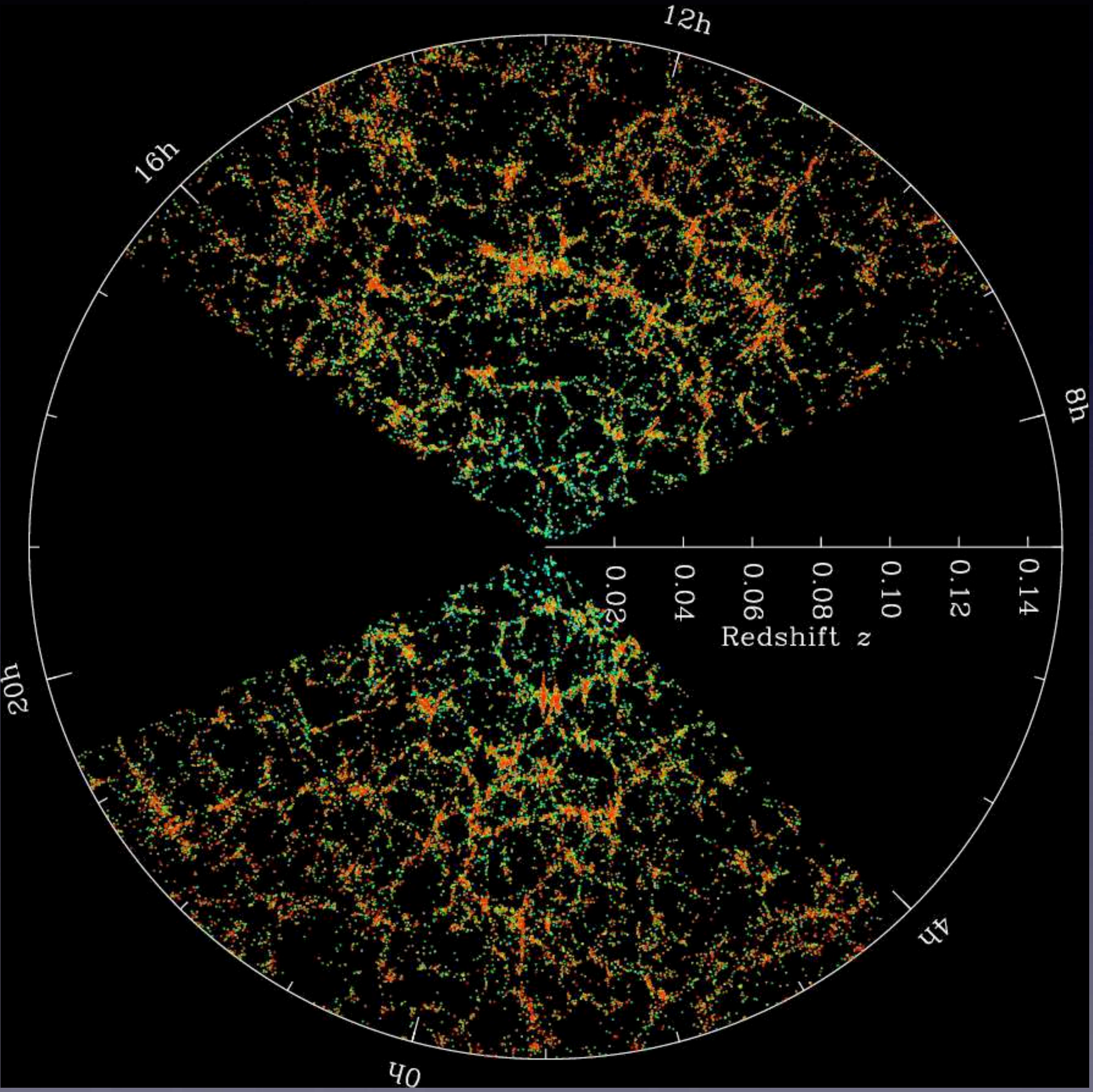
zStatus	XCORR_EMLINE
zWarning	OK
PrimTarget	TARGET_GALAXY TARGET_GALAXY_RED
SecTarget	
eClass	0.095797
emZ	0.006
emConf	0.874995
xcZ	0.005
xcConf	0.969081

Cross-identifications

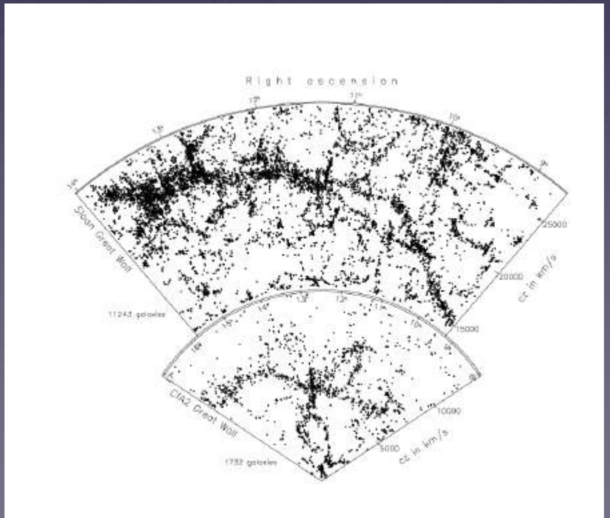
For example, spectra (here: a Seyfert [active] galaxy)



The spatial distribution of SDSS galaxies



Left: every dot is one SDSS galaxy
Note inhomogeneous distribution!
Details of this distribution contain information about the structure formation in the Universe
Below: the so-called "SDSS Great Wall"



"Ask Not What Data You Need To Do Your Science, Ask What Science You Can Do With Your Data."

The era of surveys...

- Standard: "What data do I have to collect to (dis)prove a hypothesis"?
- Data-driven: "What theories can I test given the data I already have?"

1) Introduction

Why do astronomers need Big Data:

- to make sky maps of stars
- to make sky maps of galaxies
- to search for rare objects
- to search for objects that change with time
  (either brightness or position)

Until recently the state of the art was exemplified by SDSS survey.

The next-generation Large Synoptic Survey Telescope will start in about 3 years, will survey the sky for 10 years, and obtain an equivalent of SDSS (30 TByte) every clear night!

1) Introduction

- Large Synoptic Survey Telescope: Big Data!
- astroML


BREAK
TIME !!