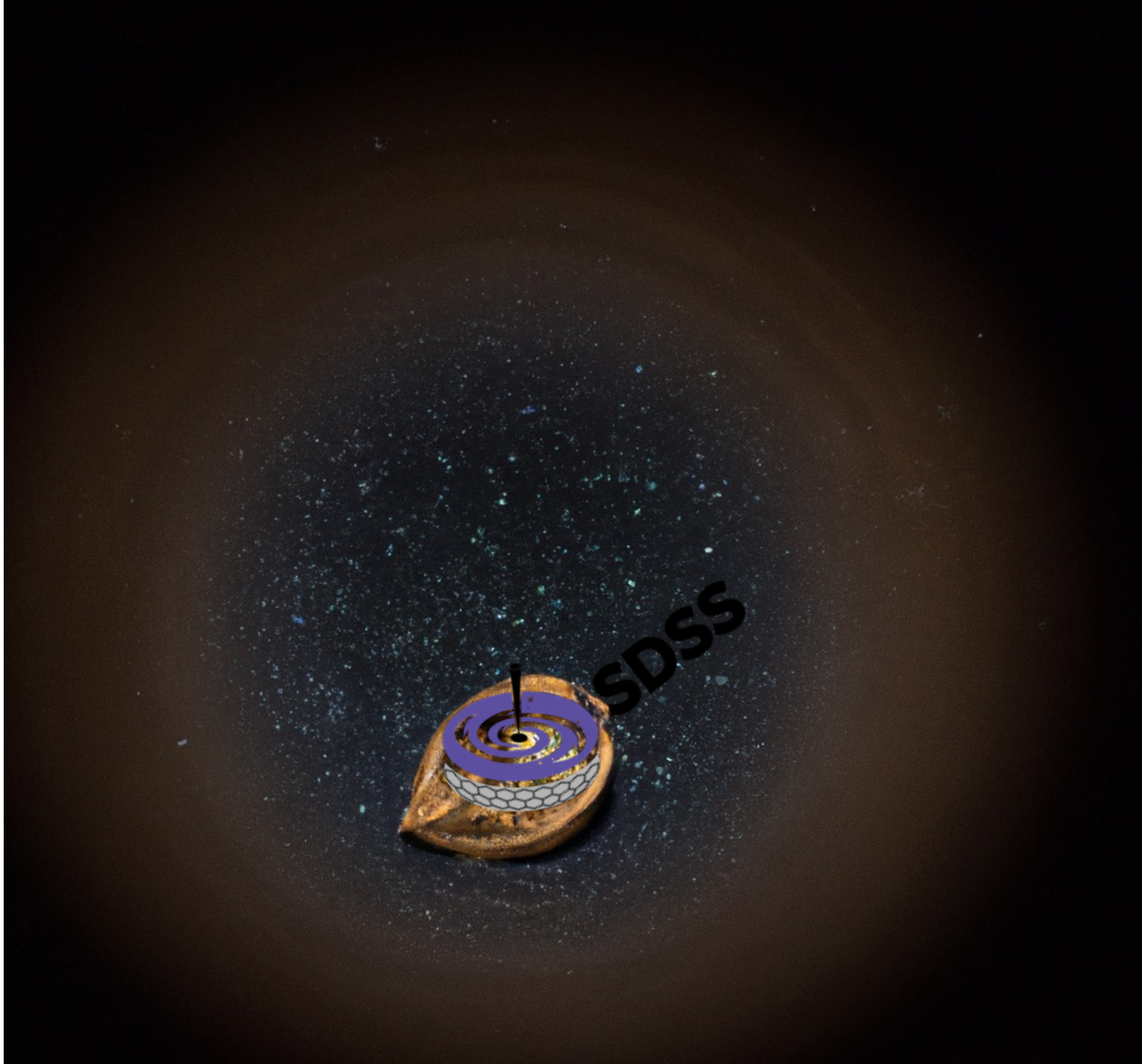
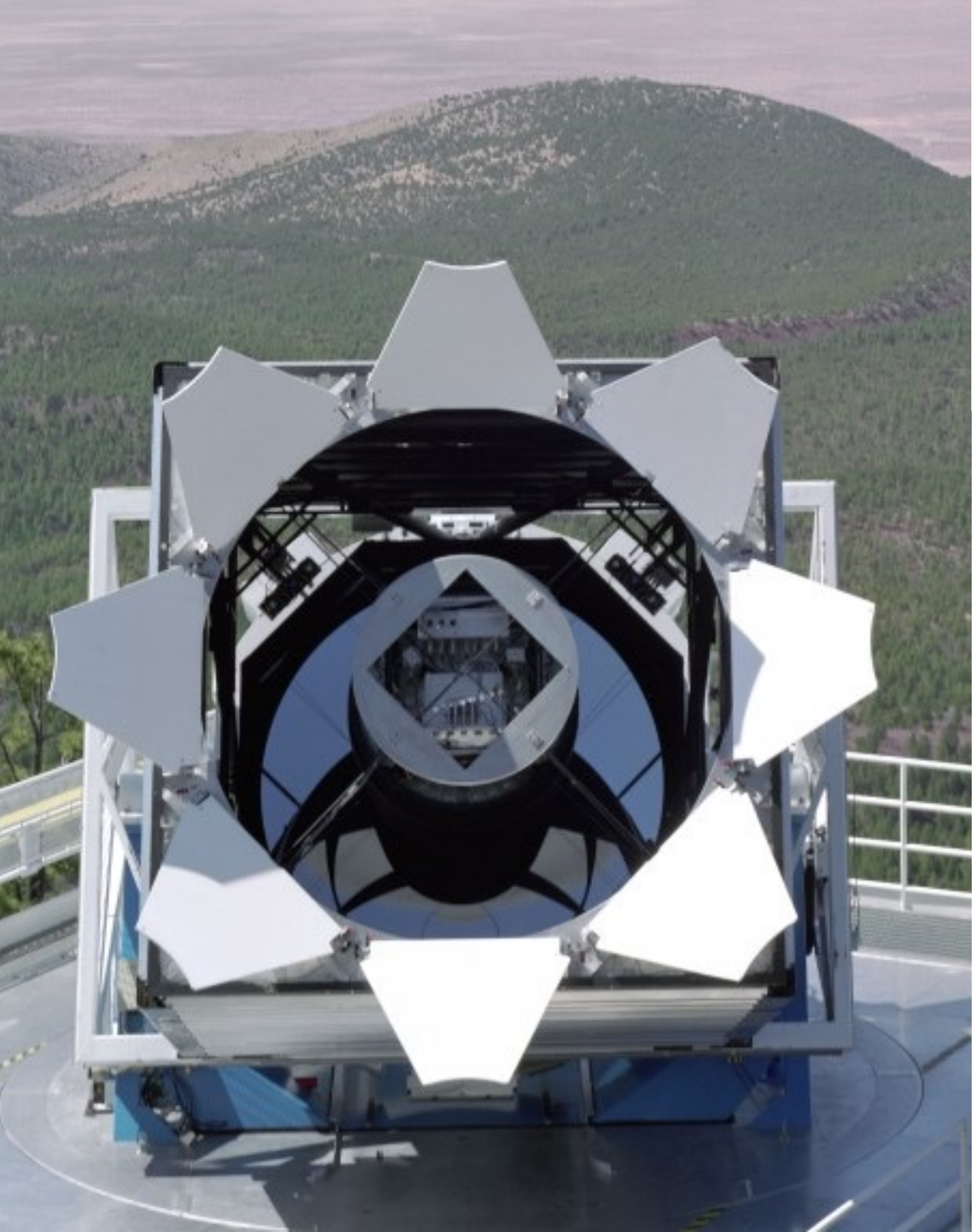
SDSS data in half a nutshell



Francisco M. Montenegro Montes

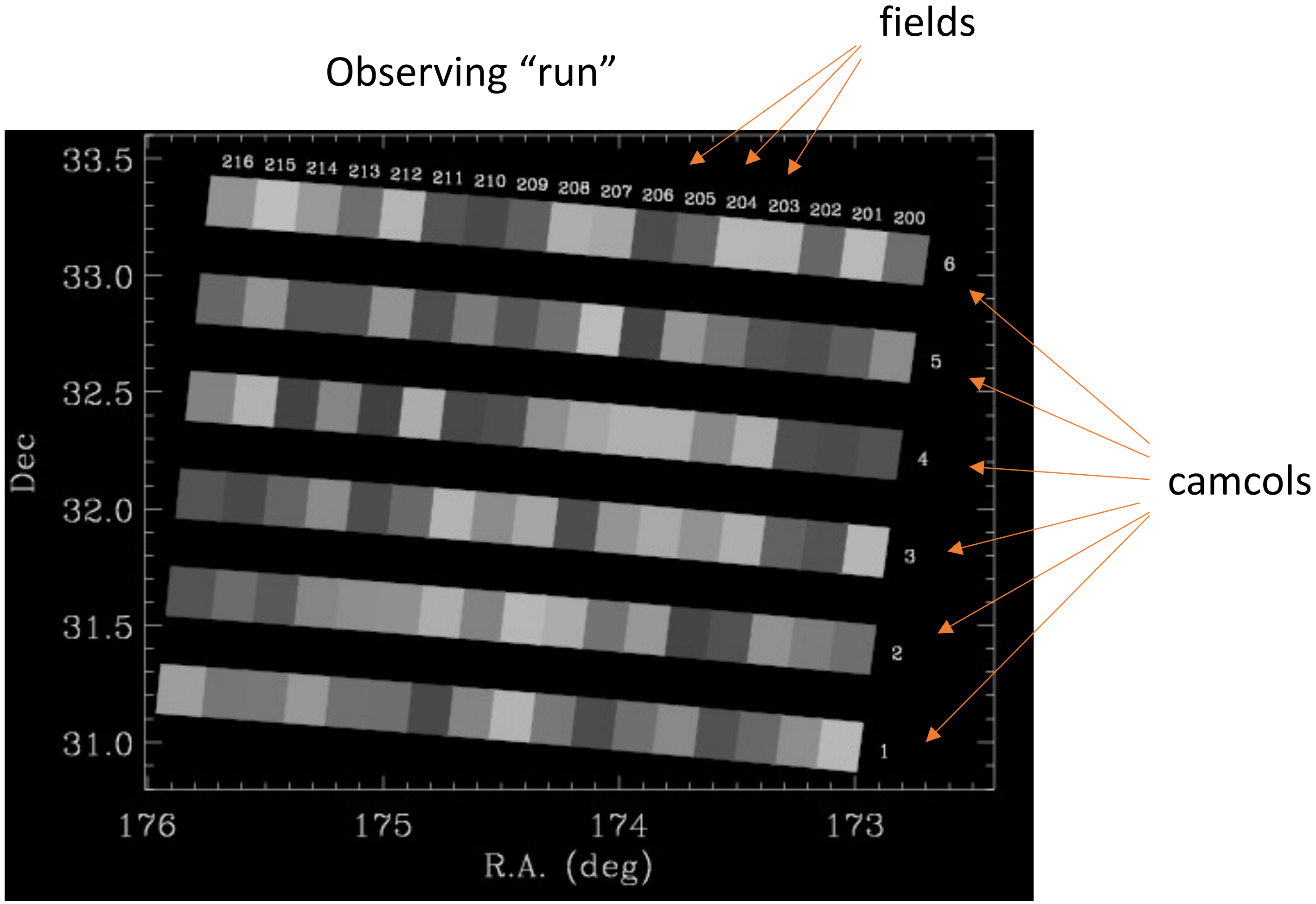
María Zambrano fellow, UCM, IPARCOS



Sloan Digital Sky Survey

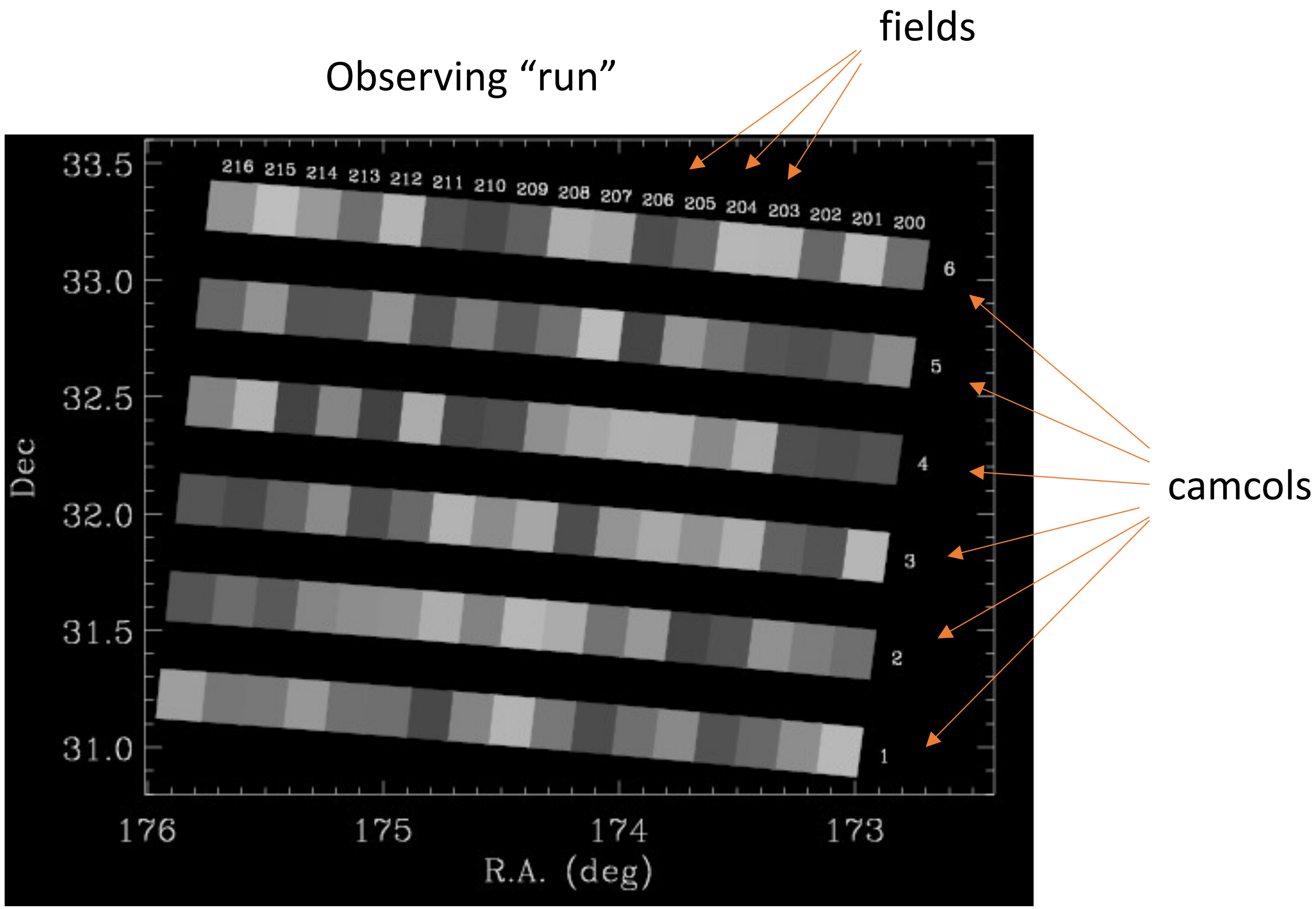
- A dedicated 2.5-m telescope to map a large fraction of the northern sky
- Collecting data since 2000. Still going on…
- Images in 5 filters (colors) and spectra for selected sources
- Huge database with images, spectra and parameters derived from their analysis. Easily accessible by the community

Observing methodology: Images



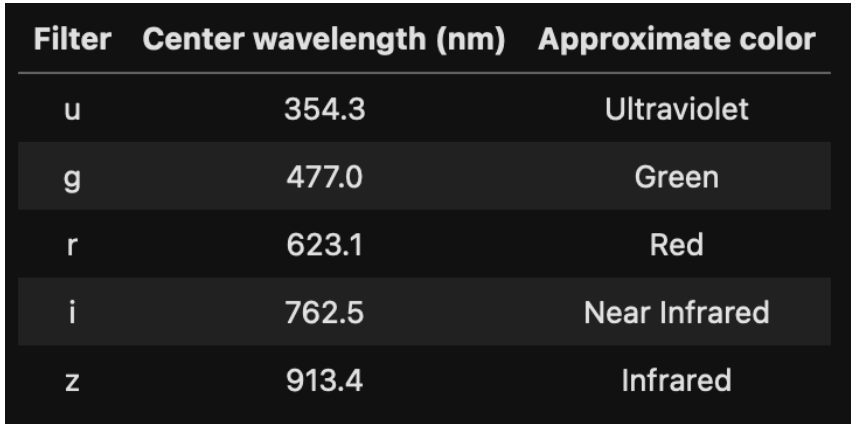
Images

- Each field is observed 5 times with a different filter. Filters are called u, g, r, I and z

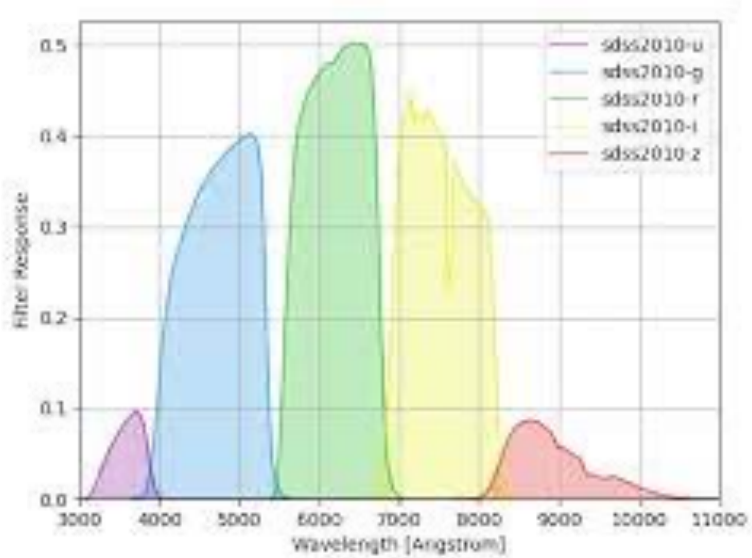

Observing "run"

fields

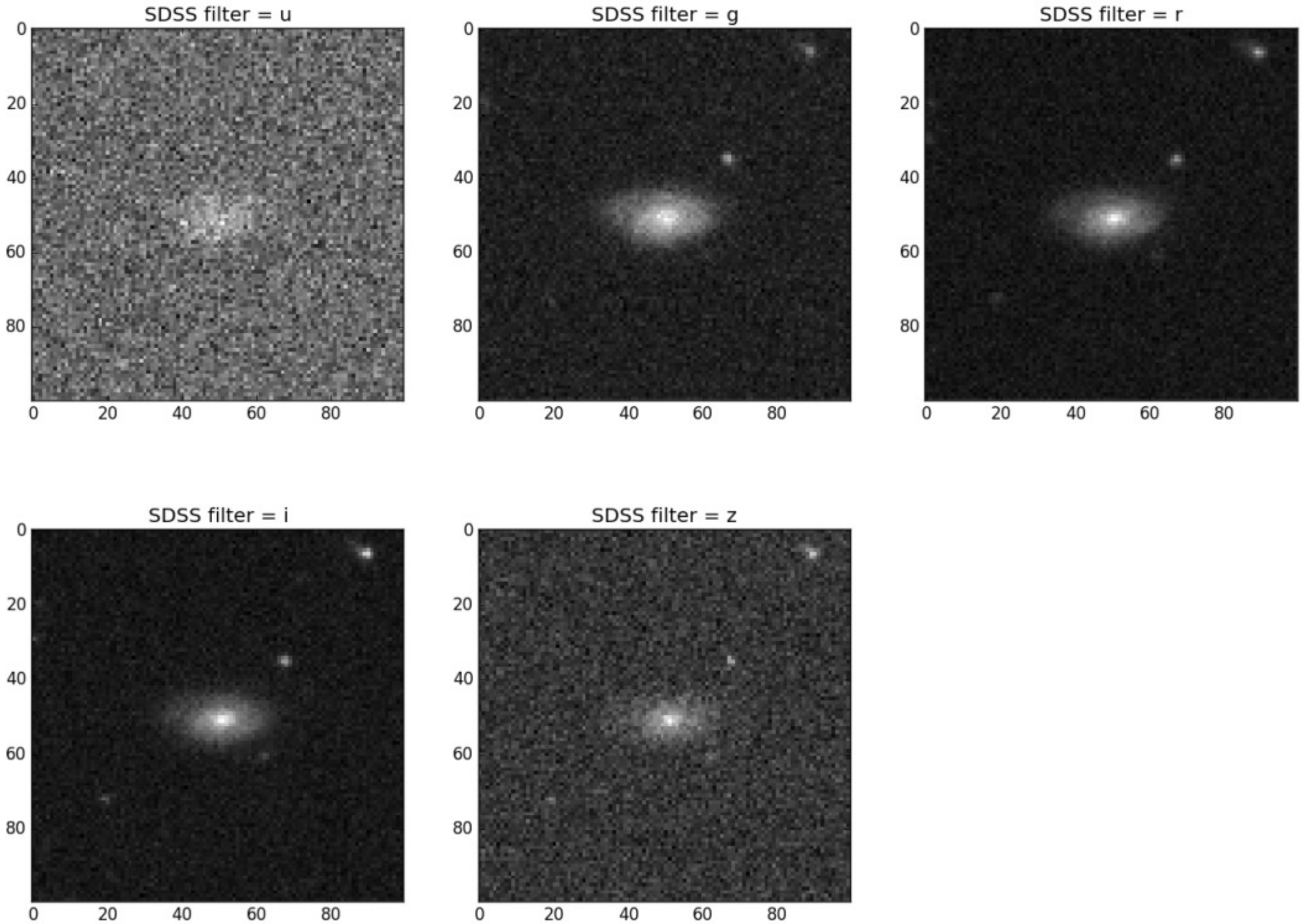
camcols

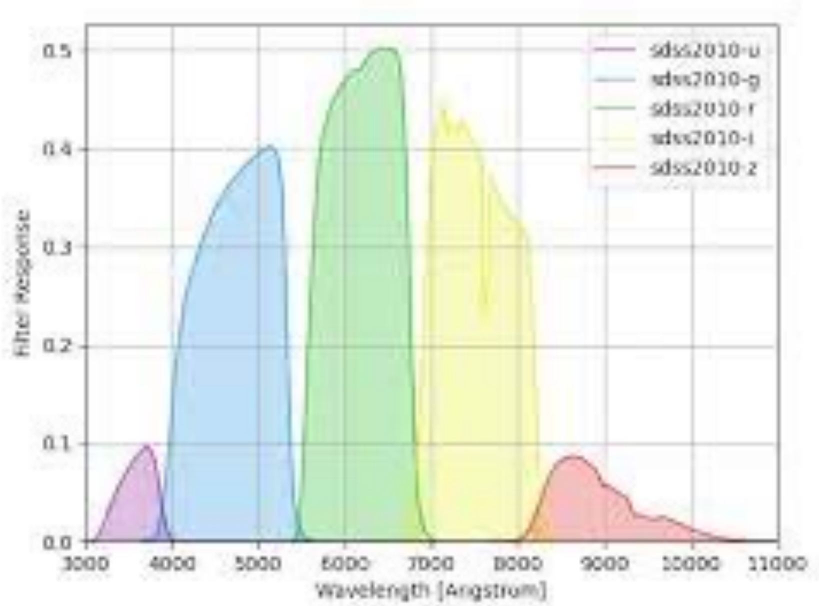


Filter	Center wavelength (nm)	Approximate color
u	354.3	Ultraviolet
g	477.0	Green
r	623.1	Red
i	762.5	Near Infrared
z	913.4	Infrared

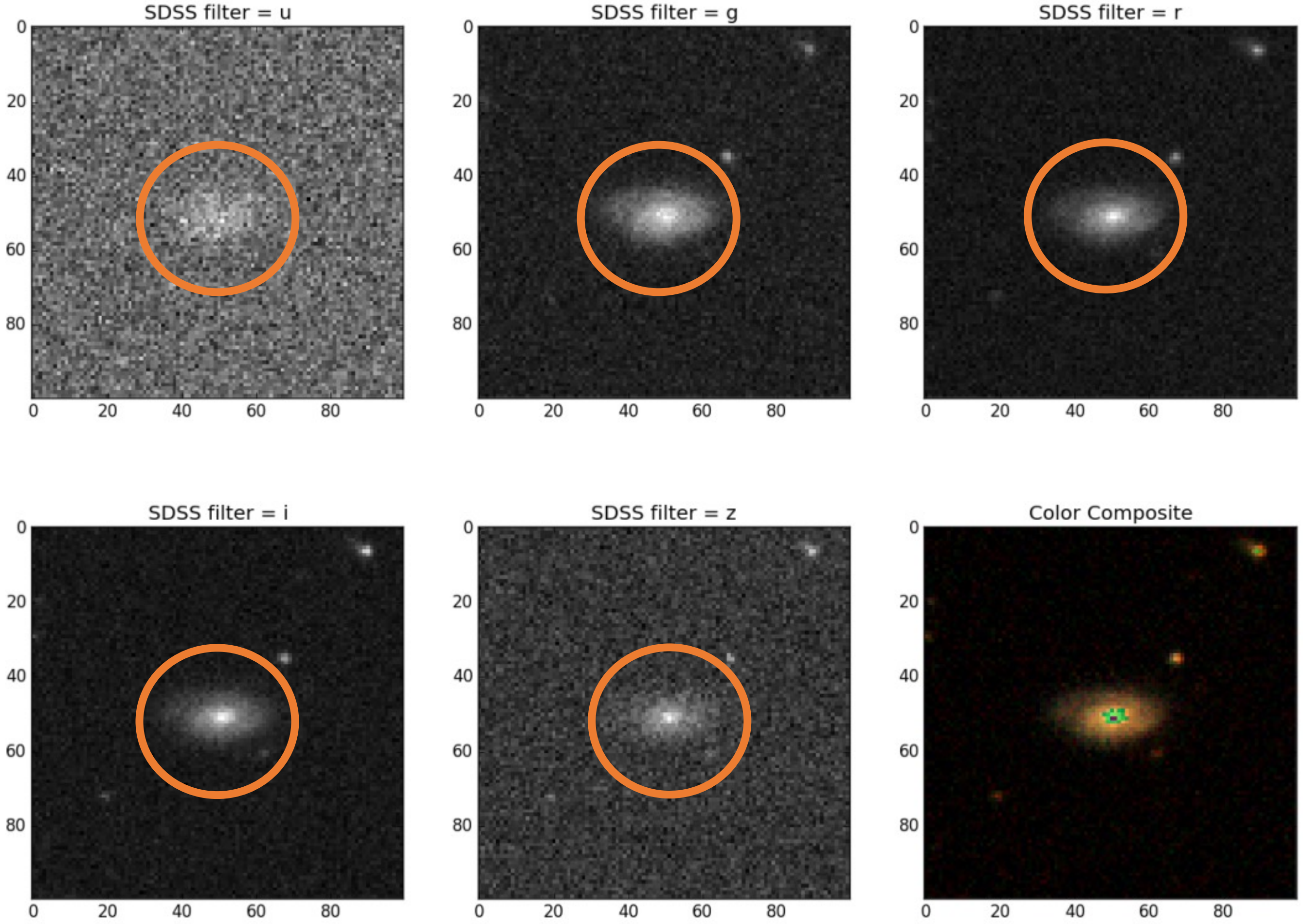


Images


SDSS filter = u
SDSS filter = g
SDSS filter = r
SDSS filter = i
SDSS filter = z
0
20
40
60
80


sdss2010-u
sdss2010-g
sdss2010-r
sdss2010-i
sdss2010-z
Filter Response
0.0
0.1
0.2
0.3
0.4
0.5
Wavelength [Angstrom]
3000
4000
5000
6000
7000
8000
9000
10000
11000

Images



Count the photons inside the circles and measure the "flux" of each object.

Astronomers normally express these in a logarithmic scale: magnitudes

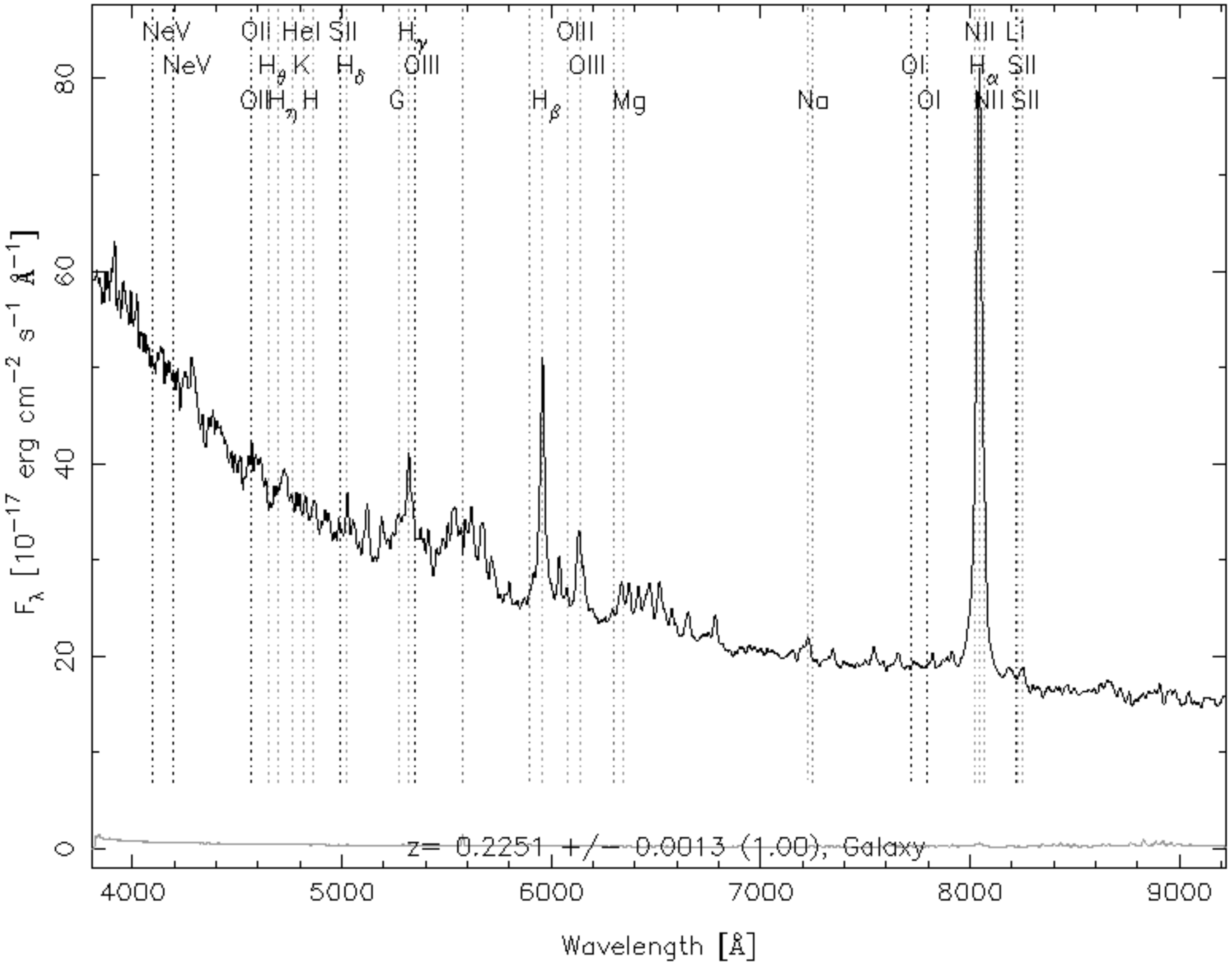
$$\boxed{m = m_0 - 2.5\, log_{10} {}^{F}\!/_{F_0}}$$

$$u = u_0 - 2.5\, log_{10} {}^{F_u}\!/_{F_{u,0}}$$

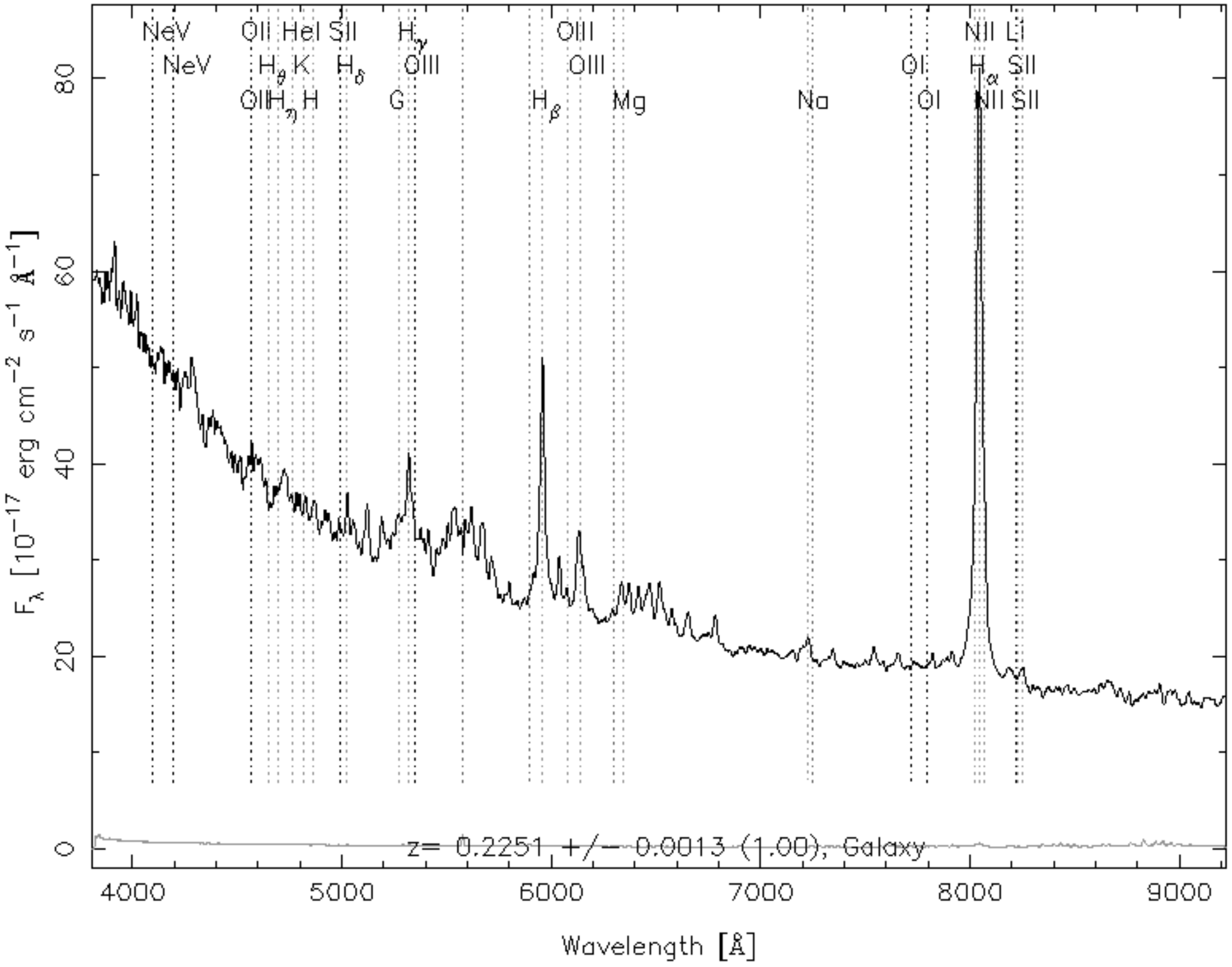
$$g = g_0 - 2.5\, log_{10} {}^{F_g}\!/_{F_{g,0}}$$

...

Spectra. How do you get it?

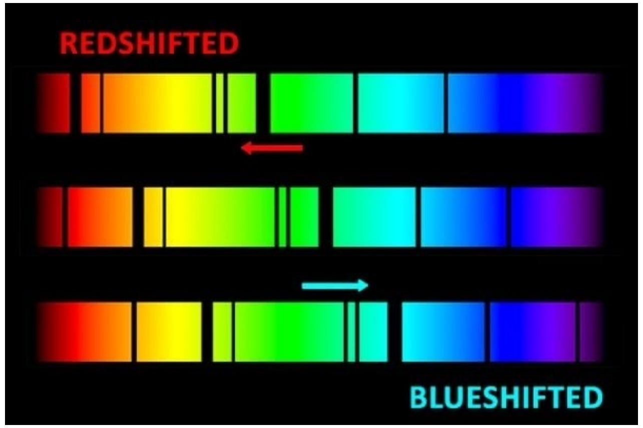


Spectra



A wealth of information is coming from the spectrum:

- Emission and absorption lines produced by atomic transitions associated with the object -> Elements present, physical conditions, etc…
- Shape of the continuum emission
- **Redshift** : displacement of the emission lines due to movement in the line of vision or due to the expansion rate of the universe

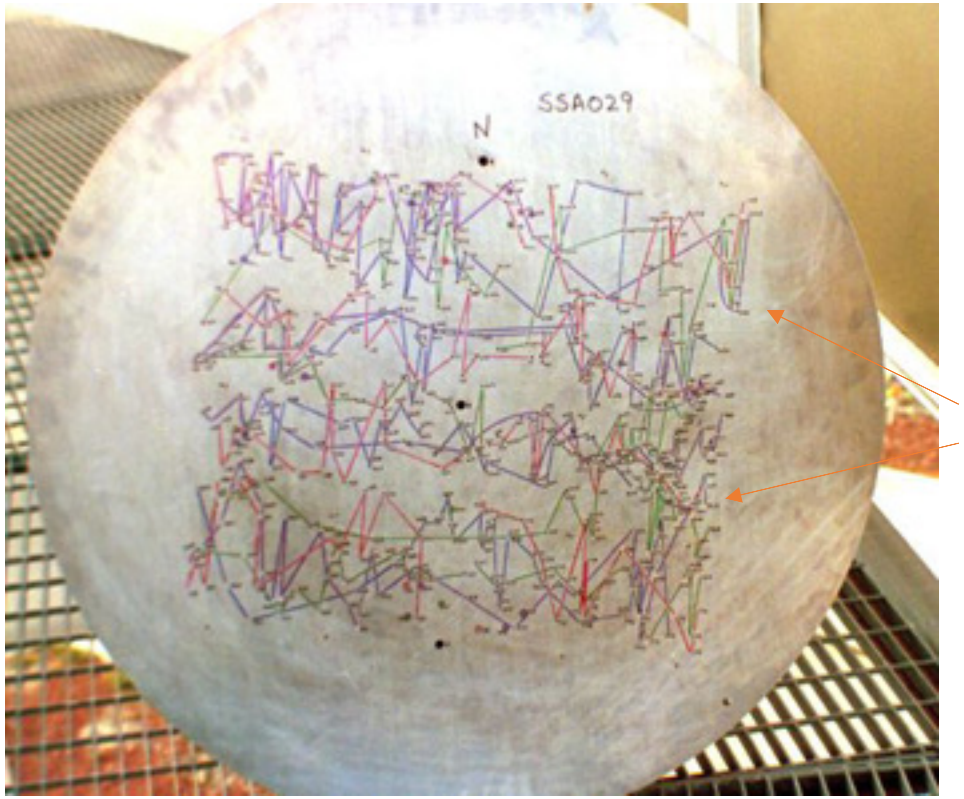


Spectra are great but expensive…

- With all this information we can reliably know what our source is: Star, Galaxy, QSO, etc…
- Let’s take spectra of everything!

Spectra are great but expensive…

- With all this information we can reliably know what our source is: Star, Galaxy, QSO, etc…
- Let’s take spectra of everything!



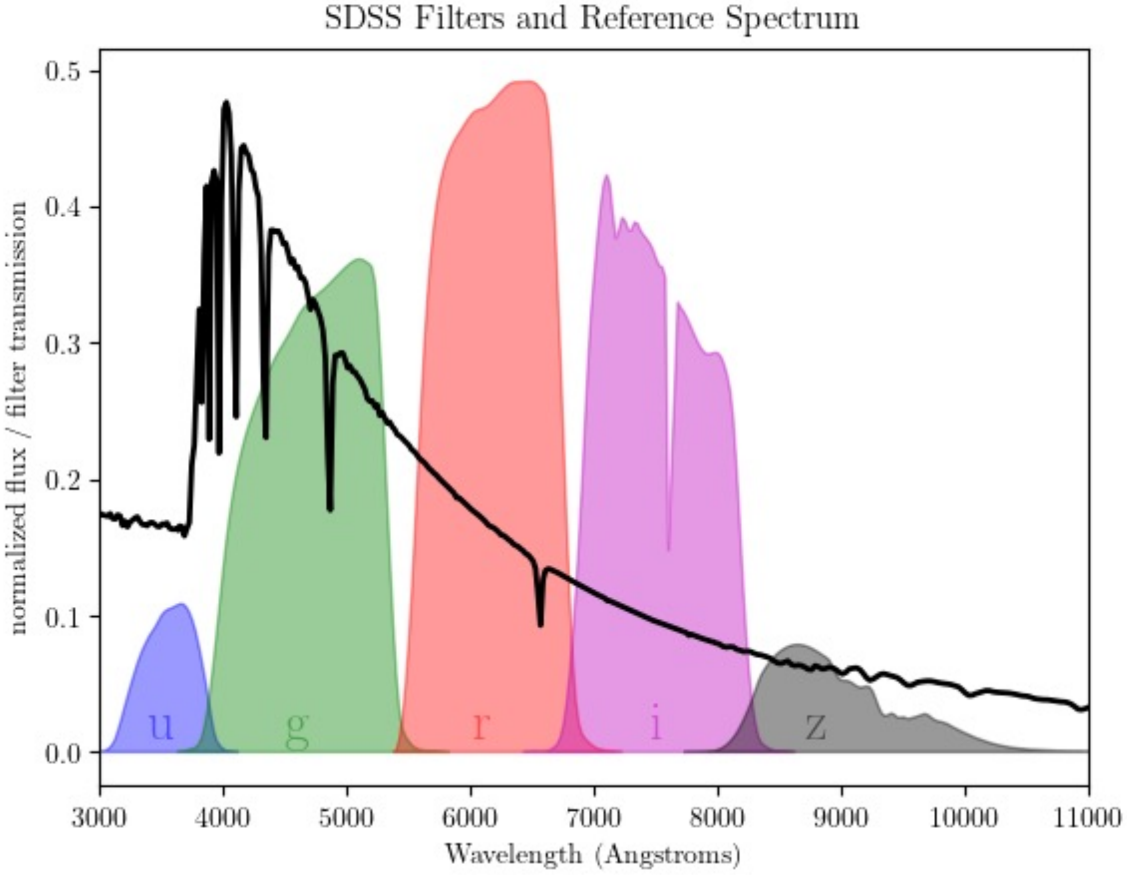
fibers

SDSS ”plate”

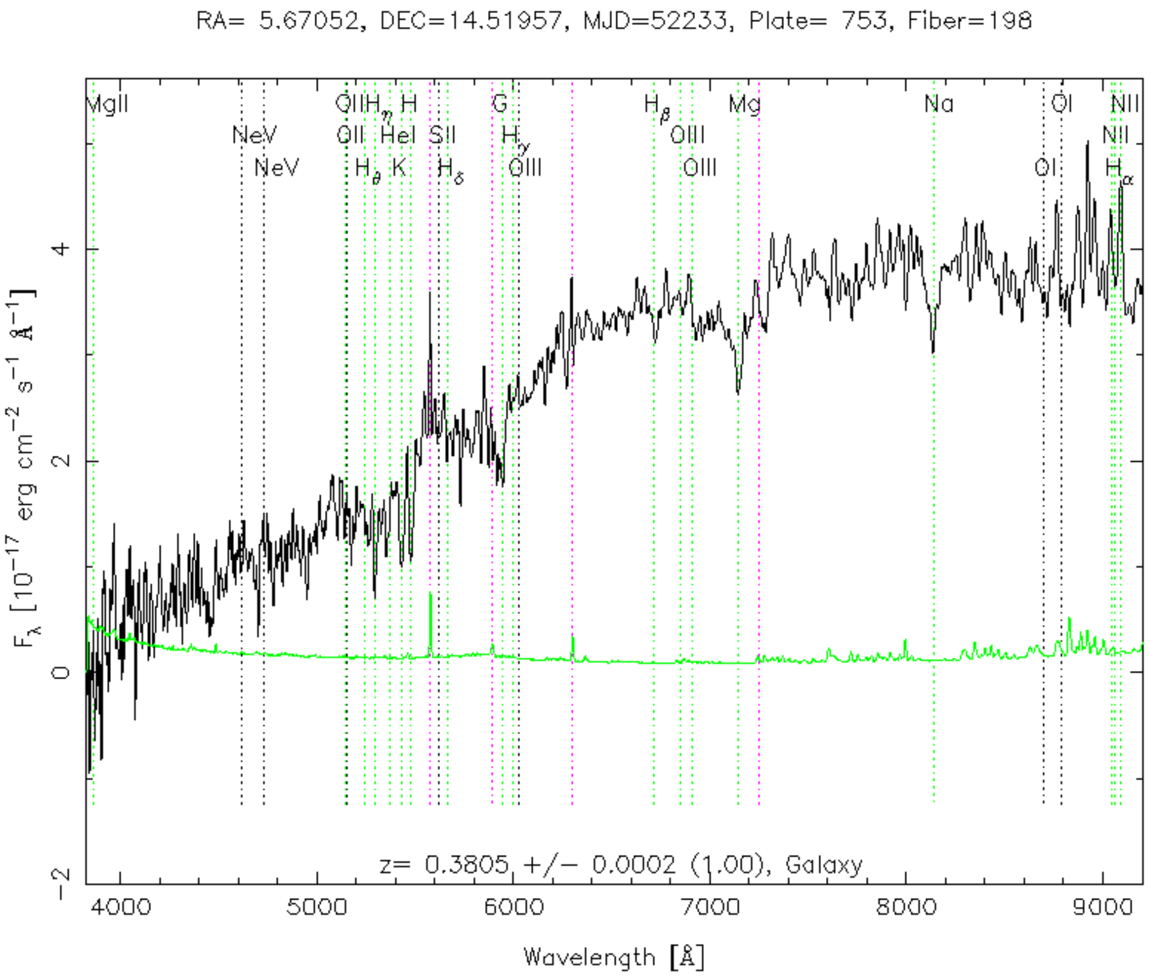
- Spectra take 10x longer time to take than images.

- SDSS could only take a limited number of them per “plate”

But still...



Vega, Blue star

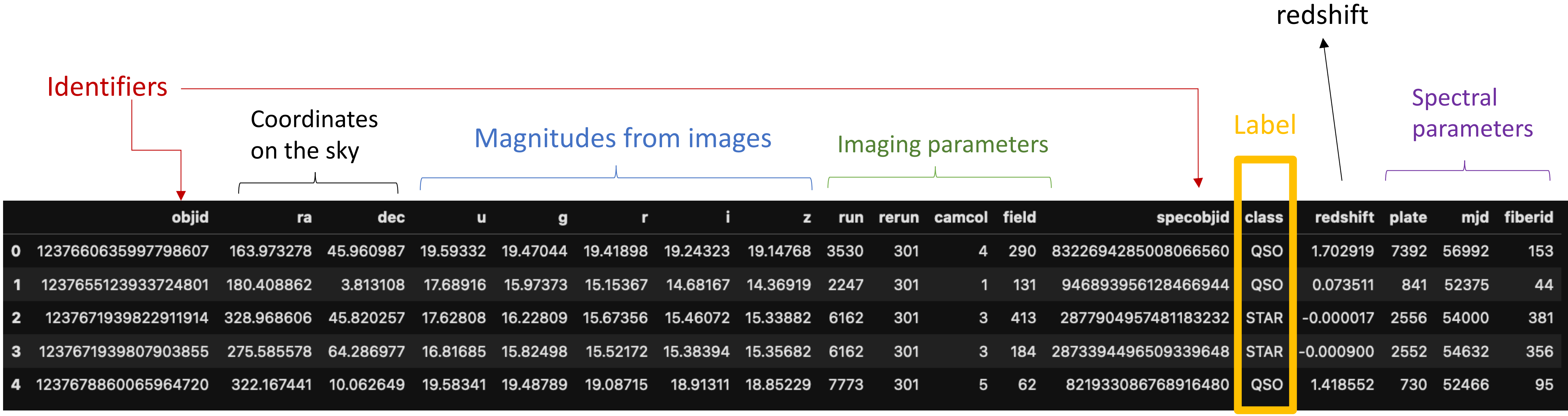


Some Luminous Red galaxy

ML can help us classifying objects without spectra

Our dataset

- 20000 objects catalogued by the SDSS survey for which spectra have been obtained
- We have information from the photometry (images) and about their spectra
- Labels are the "type of object" we could infer from the spectrum



redshift

Identifiers

Coordinates on the sky

Magnitudes from images

Imaging parameters

Label

Spectral parameters

	objid	ra	dec	u	g	r	i	z	run	rerun	camcol	field	specobjid	class	redshift	plate	mjd	fiberid
0	1237660635997798607	163.973278	45.960987	19.59332	19.47044	19.41898	19.24323	19.14768	3530	301	4	290	8322694285008066560	QSO	1.702919	7392	56992	153
1	1237655123933724801	180.408862	3.813108	17.68916	15.97373	15.15367	14.68167	14.36919	2247	301	1	131	946893956128466944	QSO	0.073511	841	52375	44
2	1237671939822911914	328.968606	45.820257	17.62808	16.22809	15.67356	15.46072	15.33882	6162	301	3	413	2877904957481183232	STAR	-0.000017	2556	54000	381
3	1237671939807903855	275.585578	64.286977	16.81685	15.82498	15.52172	15.38394	15.35682	6162	301	3	184	2873394496509339648	STAR	-0.000900	2552	54632	356
4	1237678860065964720	322.167441	10.062649	19.58341	19.48789	19.08715	18.91311	18.85229	7773	301	5	62	821933086768916480	QSO	1.418552	730	52466	95

3 categories: STAR, GALAXY, QSO

Our exercise:

- Let’s train a machine-learning model that can predict the labels (class for each object) from the other parameters we have in our table

- Let’s measure the performance of our classifier

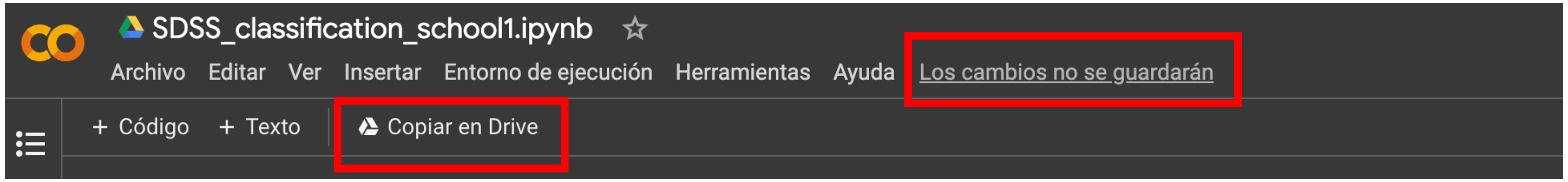
- Let’s reflect about the result and see if we can do better

Let’s get started!

- First go to this notebook in colab

https://colab.research.google.com/drive/1sMAIQ-j3OwheuWCODmPD8PDoWKEbq3bc

- You have read access only, so in order to introduce changes you should copy it to your local Google drive



- Get the data file from indico: `SDSS_20k.csv` and place it in your **home folder** in your Google drive

- Once you open your local copy of the notebook, you can run it and make changes. You should be able to read the data file directly from the notebook.