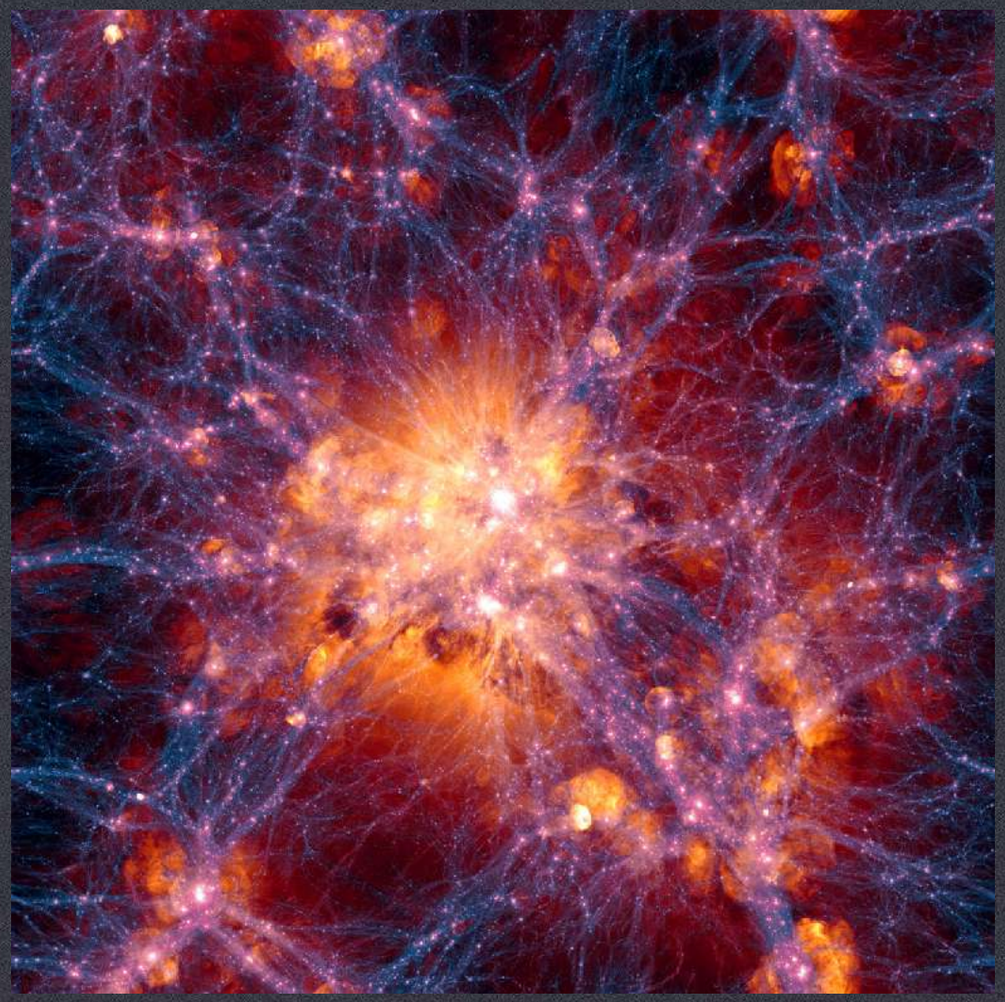


OBSERVATIONAL COSMOLOGY

AN INTRODUCTION TO THE CONNECTION BETWEEN THEORY AND DATA IN COSMOLOGY

Raul Abramo - Inst. de Física/USP

Cosmology: why we do this

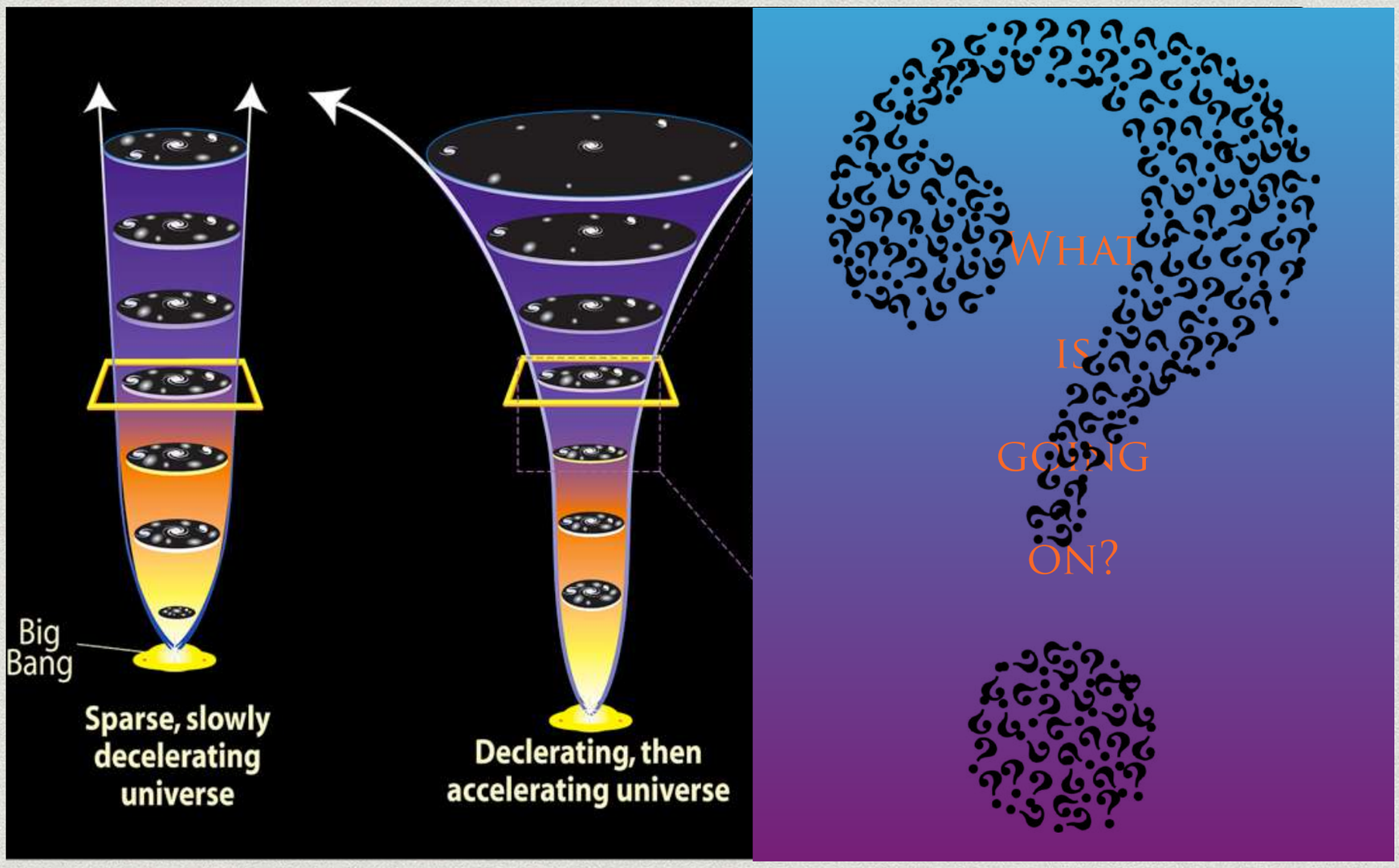


Photo: Roy Kaltschmidt. Courtesy: Lawrence Berkeley National Laboratory

Saul Perlmutter

Photo: Belinda Pratten, Australian National University

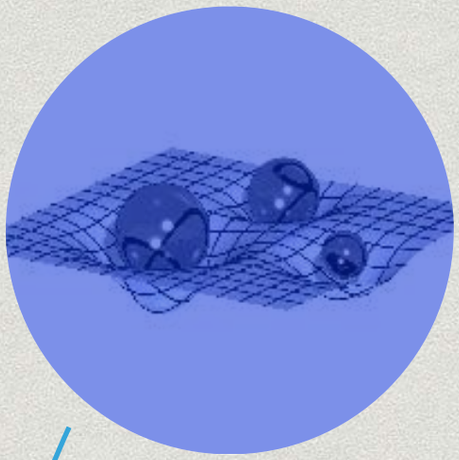
Brian P. Schmidt

Photo: Homewood Photography

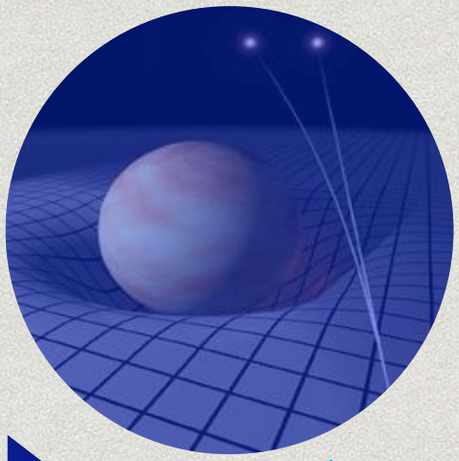
Adam G. Riess

Cosmology: why we do this

Matter curves spacetime, determines metric…



$$G_{\mu\nu} = 8\pi G T_{\mu\nu}$$



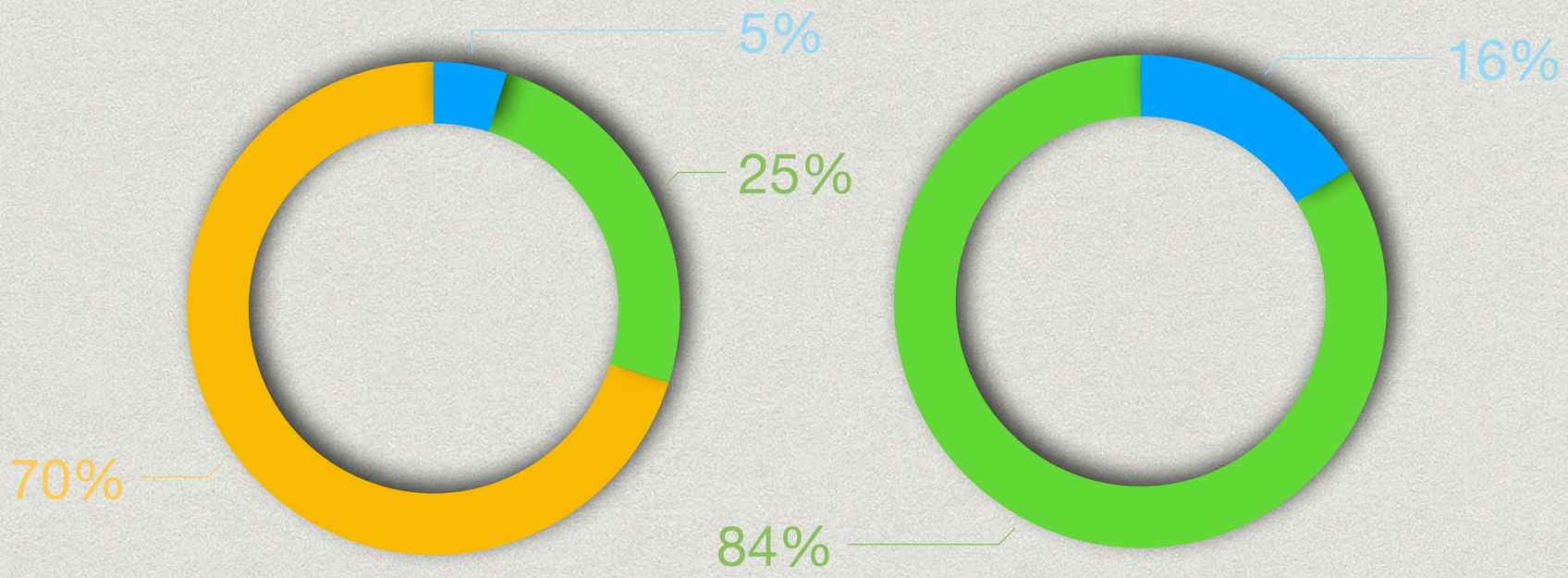
… and curvature determines how matter moves

Modified gravity

Dark energy

Only way to accelerate is to make some changes either here…… or…… … here

Cosmology: why we do this

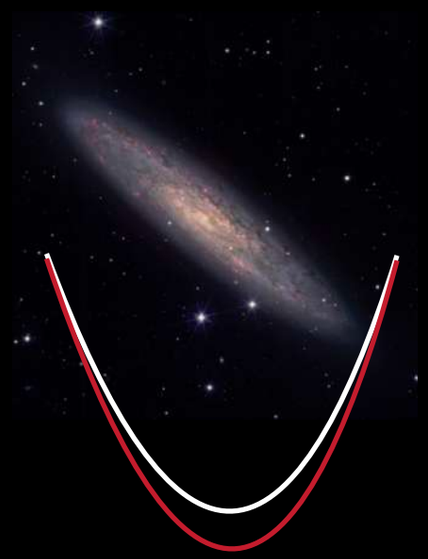


According to existing data, only a small fraction of our Universe is made of ***visible matter***: if ***dark energy*** accounts for cosmic acceleration, it makes up ~70% of the total. And, of course, there is still ***dark matter***…

Cosmic acceleration: dark energy or modified gravity?

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + 8\pi G T^{E}_{\mu\nu}$$

$$\Delta G\mu\nu + G_{\mu\nu} = 8\pi G T_{\mu\nu}$$

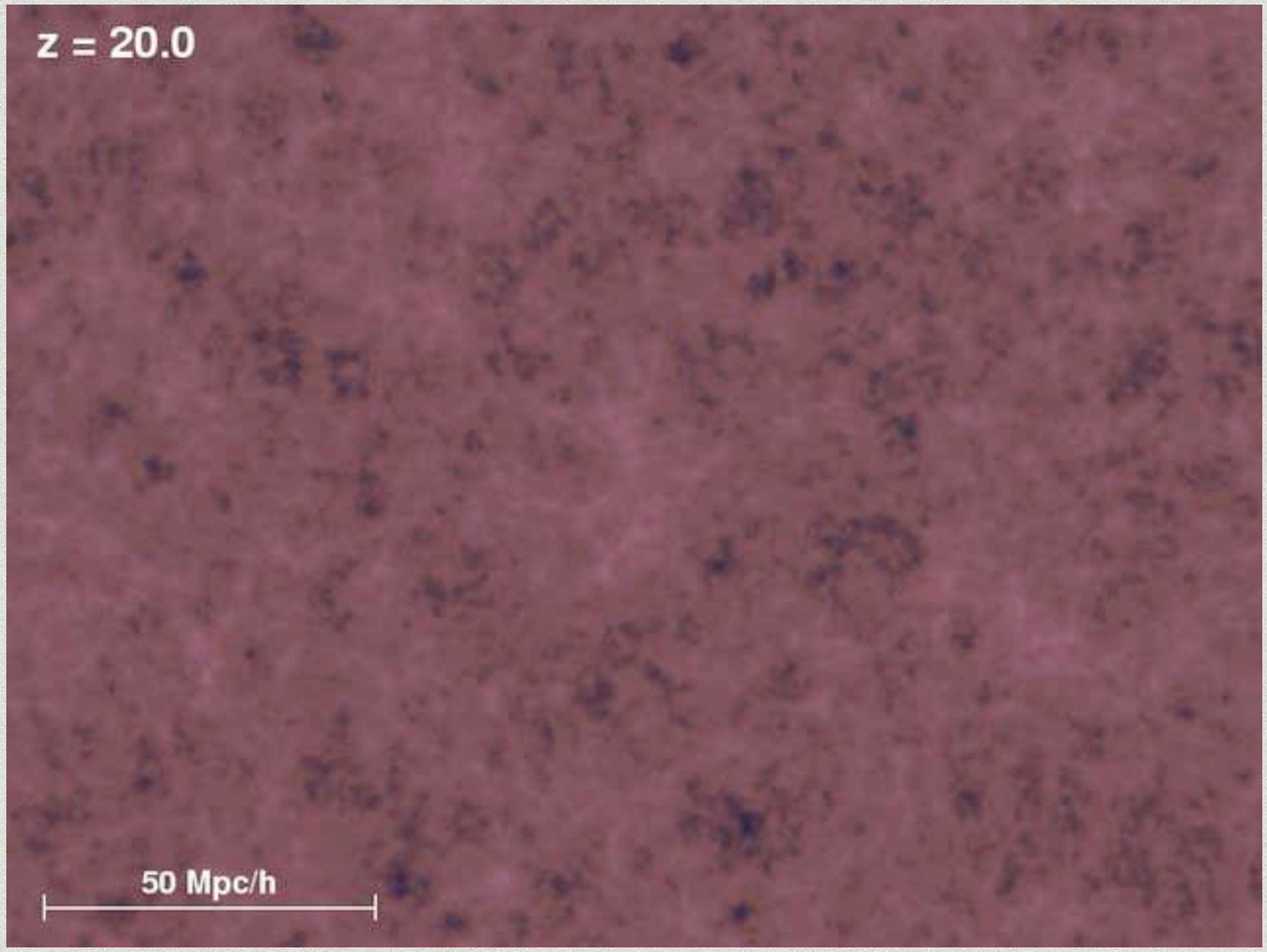


$$\nabla^2\Phi = \frac{16\pi G}{3}\delta\rho - \frac{1}{6}\delta R(f_R)$$

SAME MATTER, DIFFERENT GRAVITY

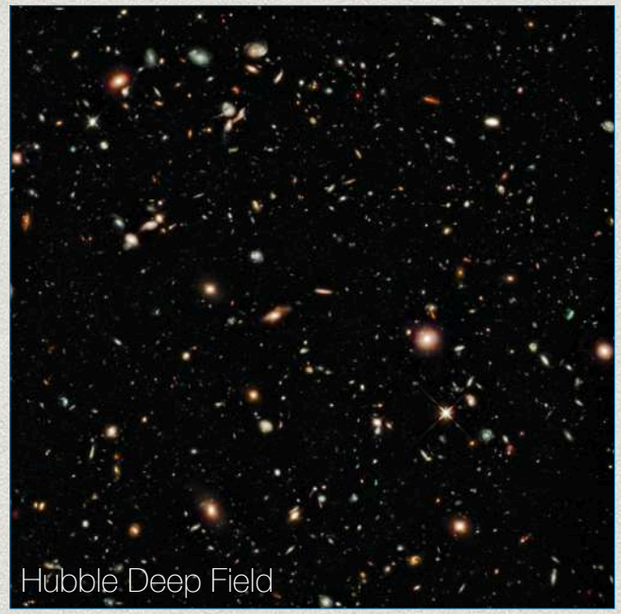
Both ***dark energy*** and ***modified gravity*** can describe the same expansion history (w/ identical Friedmann equations), but changing gravity's laws also affects the ***Poisson equation*** and the ***geodesic equations***

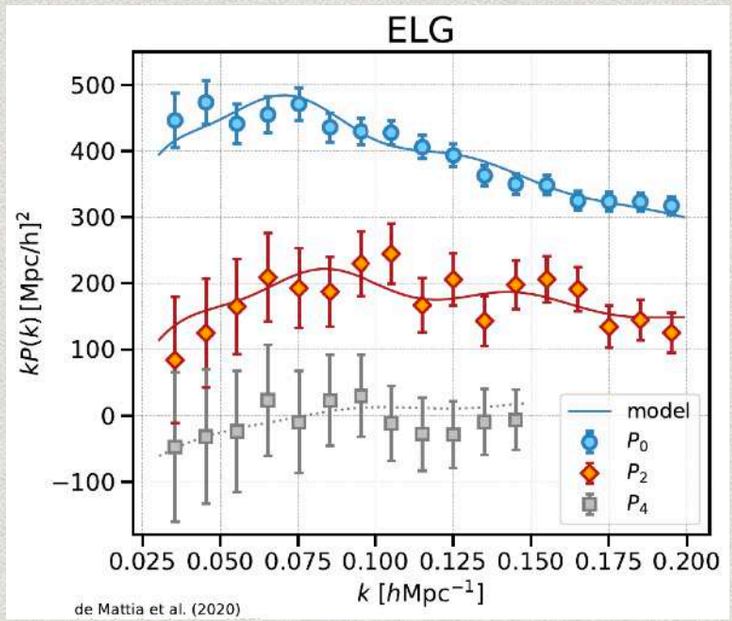
The key: large-scale structure

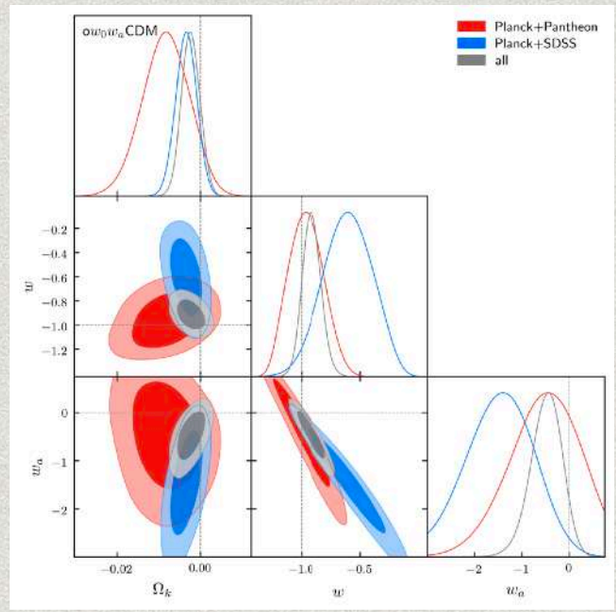


Starting from nearly Gaussian initial conditions, N-body simulations tell us that on very large scales (>100 Mpc) the Universe should look like a ***web of structures***: sheets, filaments, nodes, halos, etc.
This **competition** between **expansion** and the **gravitational interactions** of the matter clumps are the key to unlock the mysteries of dark energy, dark matter – or even modified gravity!

Cosmology:
from data to theory, and back

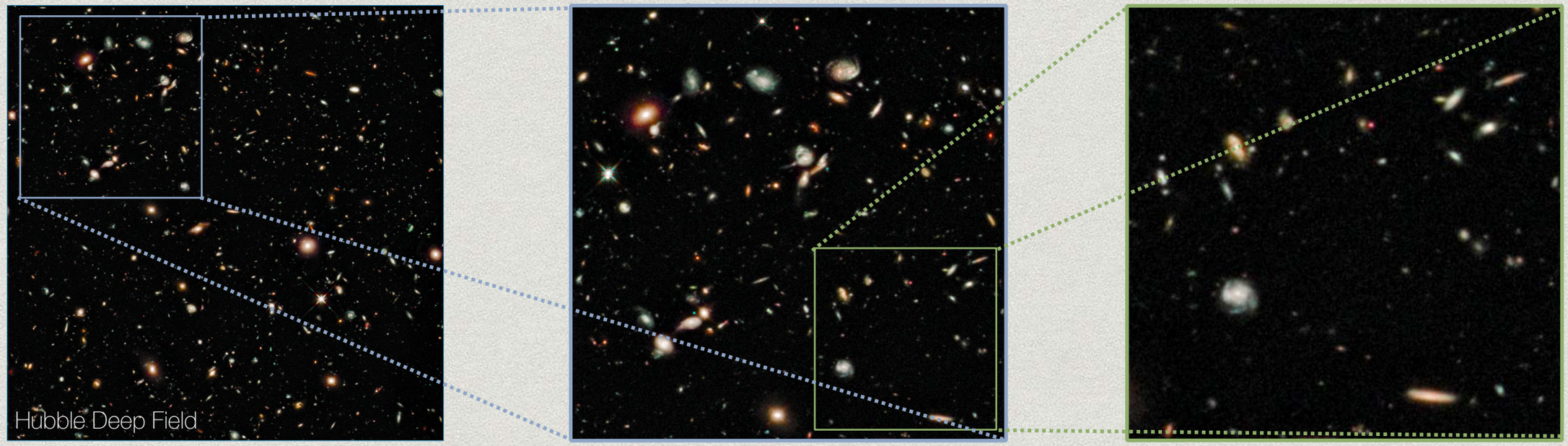






- How to get from **here** ……..to **here**……to **there**.
- How **observations** impact our **constraints**?
- How our theoretical “desires" can inform the way that we plan for **new surveys**?

The nature of astronomical data

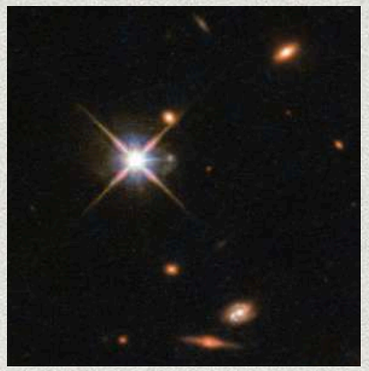


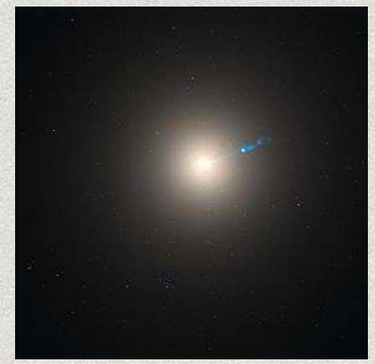
- Almost all cosmological observations start with (or involve) an **image** like this.
- Now, can you tell me **what is what**?
- What is a **star**, what is a **galaxy**, what is a **quasar**… and would you be able to estimate **how far** are the galaxies (their **redshifts**)?
- Now, we make mistakes — we all do, all the time. But how can we understand the extent to which these **errors affect our conclusions** about the underlying Physics? And how can we draw **better constraints** with the **same data sets**?

The nature of astronomical data

- This will be the main topic of these lectures: how to go from data to theory, and how theory informs the way in which we collect (and analyze) data.
- Our objects of interest are detected in a **variety of wavelengths:** microwave (e.g., CMB), radio (21cm), UV, IR, X-ray, Gamma-rays…
- The main surveys tend to focus on the **optical** range of the electromagnetic spectrum ($\sim$ 4000 Å $-$ 9000 Å).
- Increasingly, Cosmology is incorporating information from new windows: time domain (Rubin/LSST), neutrinos, gravity waves (LIGO), and probably many more to come.
- My lectures will focus on **optical surveys**, and on the **large-scale structure**.

Stars, galaxies, quasars

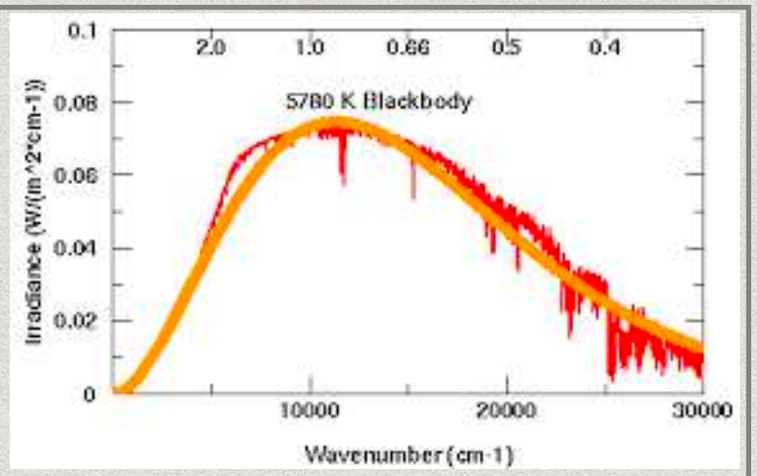


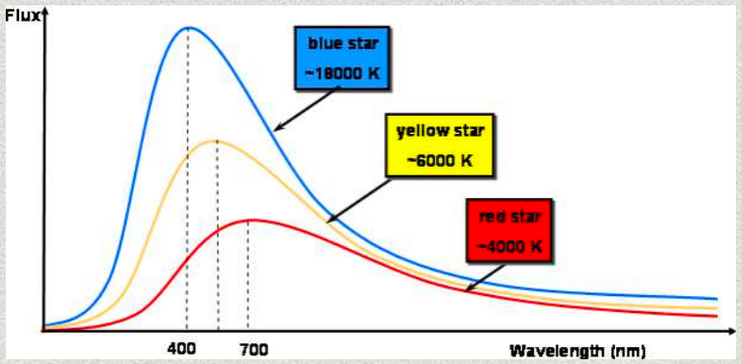


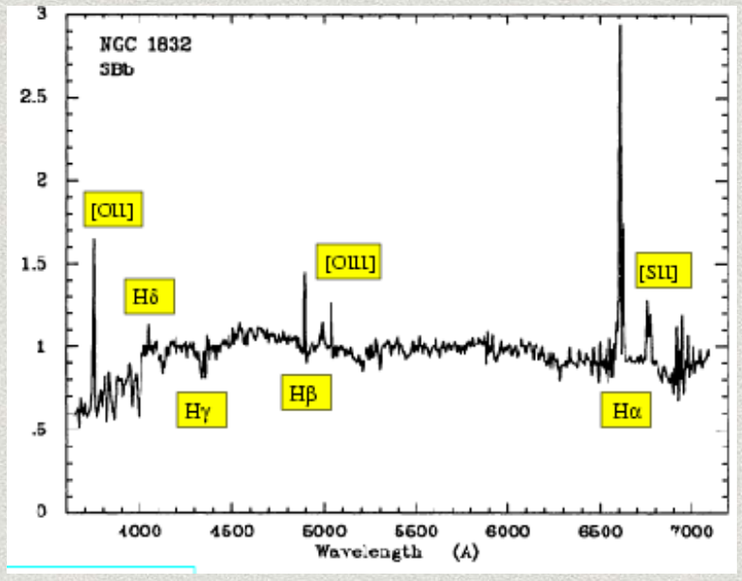
- There are many ways that we can use to tell what an object is.
- In the case of **galaxies**, their **shapes** are a start: they are often **spatially resolved** sources, presenting a spiral or elliptical form.
- However, most **faint galaxies** are not resolved, and appear in our images as **point sources**.
- **Stars** are abundant in our images, especially if we observe near the disk of the Milky Way. They cannot be resolved by even the best telescopes, so they appear as **point sources**.
- **Quasars** ("quasi-stars") are super-massive black holes at the centers of galaxies, that are often much brighter than the galaxies they inhabit, and therefore appear to us as **point sources**.
- Clearly, the **shape** of the source (its morphology) **cannot** serve as the only basis to select objects.

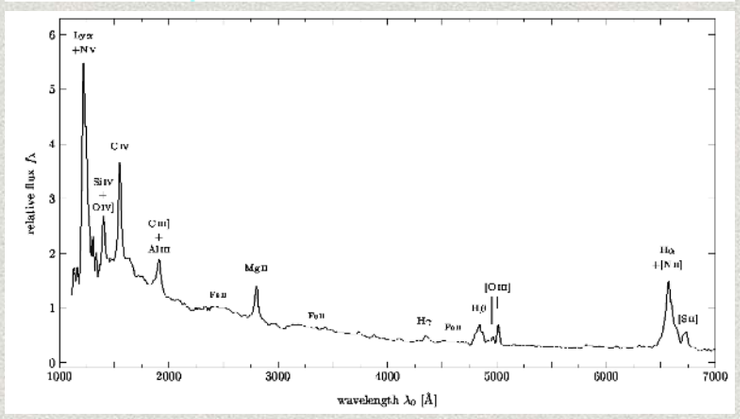
Stars, galaxies, quasars

- Stars, galaxies and quasars are completely different objects, of course: this means that they **emit light in totally different ways.**
- **Stars** are fueled by **nuclear fusion** (H→He→…→C, N, O, …), and since their gas is often in **thermal equilibrium**, the light spectrum is closer to a Blackbody spectrum.
- **Galaxies** are **composite objects** made up of millions/ billions of stars of all kinds, with a lot of gas thrown in between.
- **Quasars** are powered by the **accretion disk** around a supermassive black hole, and are much “hotter" (=bluer) than galaxies or stars. Moreover, the super-fast rotation of the disk creates a Doppler effect on the emission lines of H, He, C etc., which makes those **emission lines** very **broad**.
- All this means that the **spectral energy distributions** (SEDs), or **spectra** of these objects, allow us to clearly distinguish one from the other.



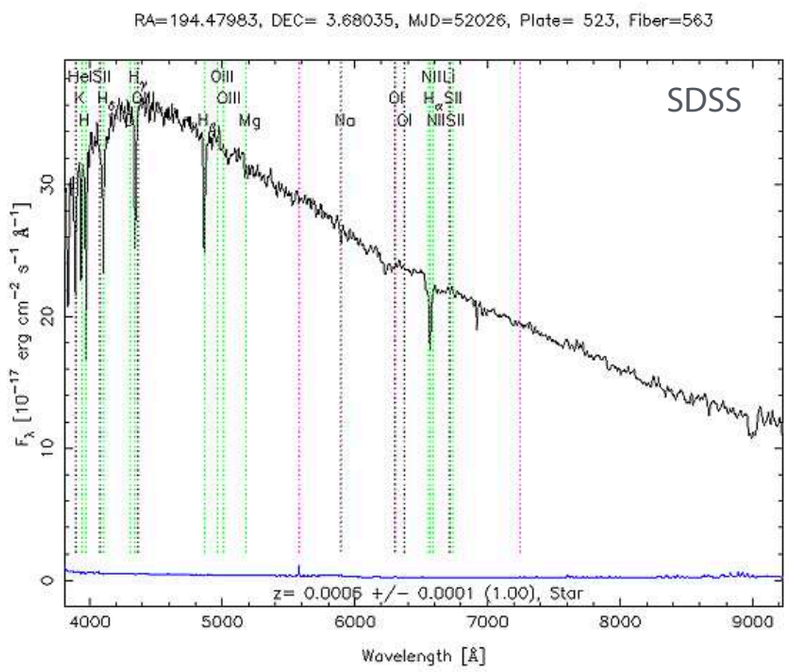




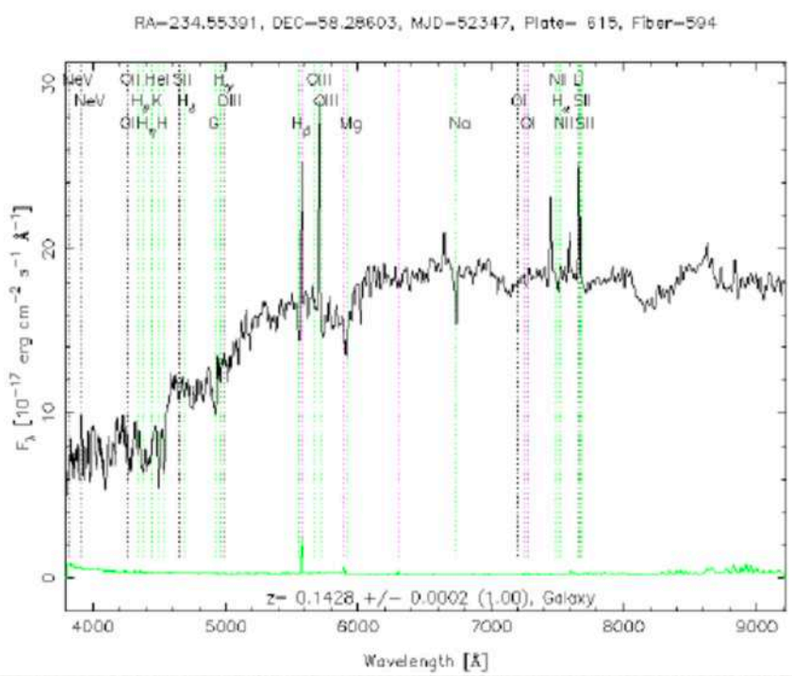


Stars, galaxies, quasars: spectra

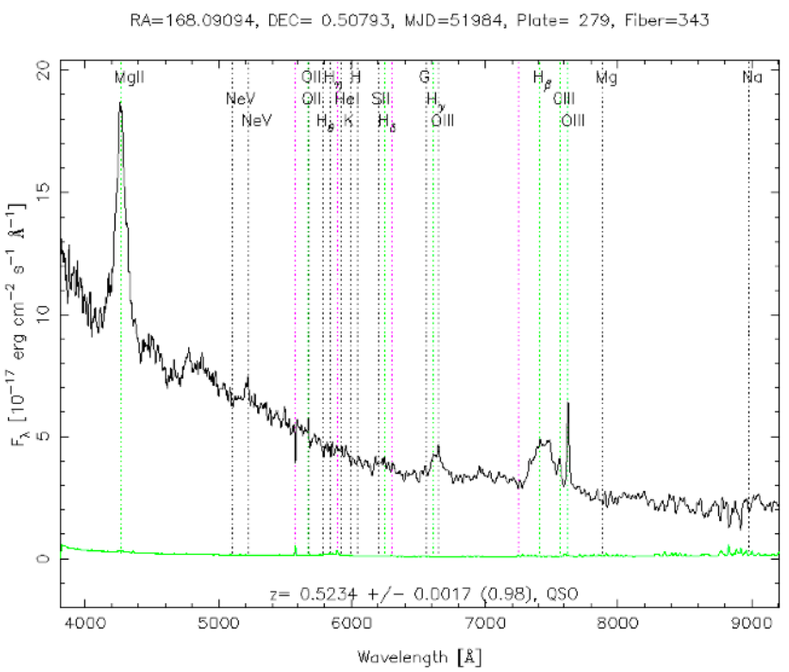
Star

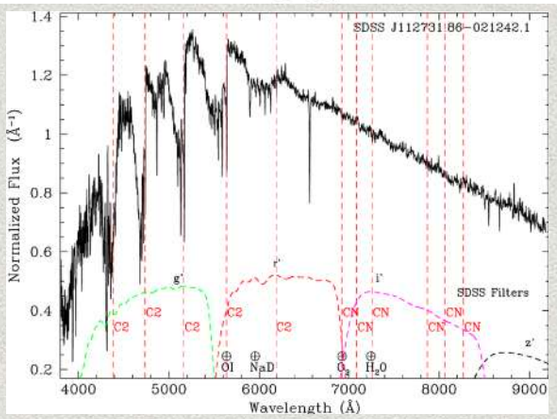


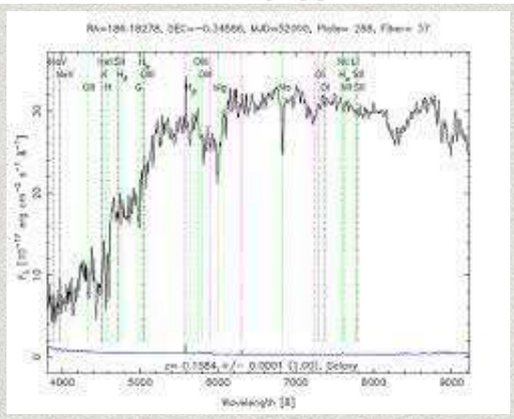
Galaxy

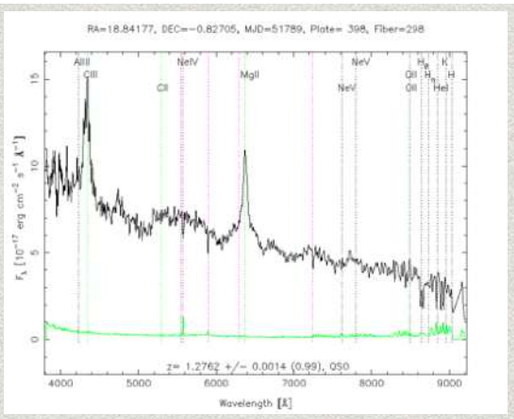


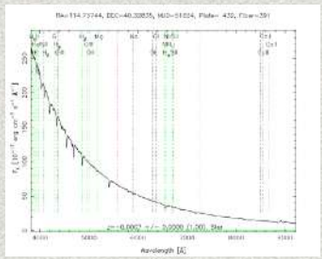
Quasar

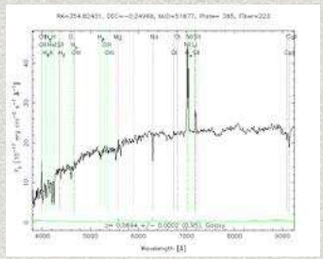


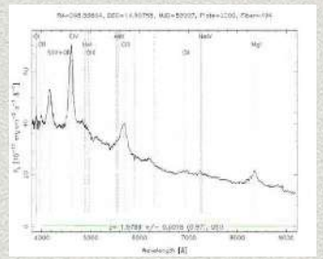




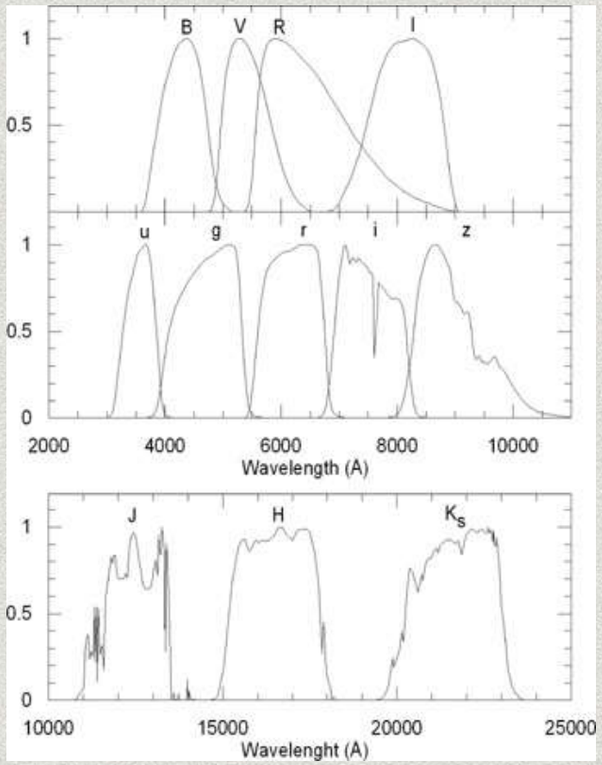




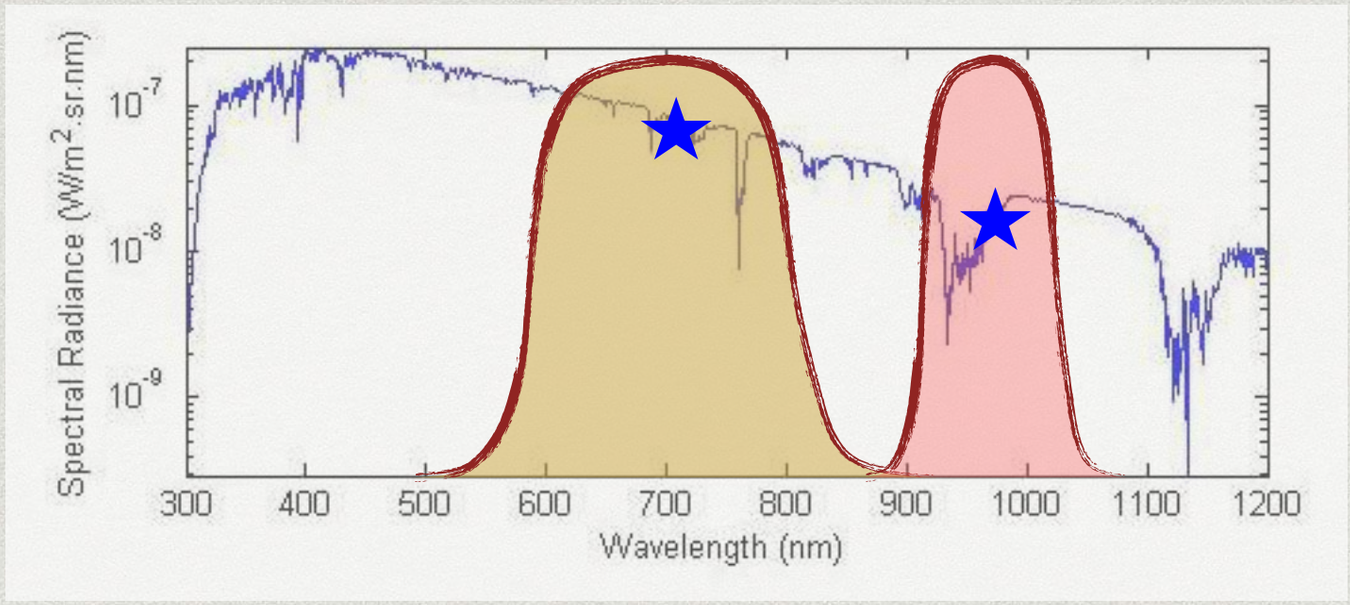




Optical bands



$$F = \int d\lambda\, T_F(\lambda) f(\lambda) \quad \text{and let's assume that} \quad \int d\lambda\, T_F(\lambda) = 1$$



- An alternative is to observe those same objects through "filters" ("passbands"), which collect the light in some spectral **window**.
- The SEDs are "averaged out" by the passbands. But these measurements in filters are **much simpler** than taking spectra.
- Let's make a simple calculation, assuming we have an SED in intervals of 1Å . If the passband has a width of 10^3 Å we have:

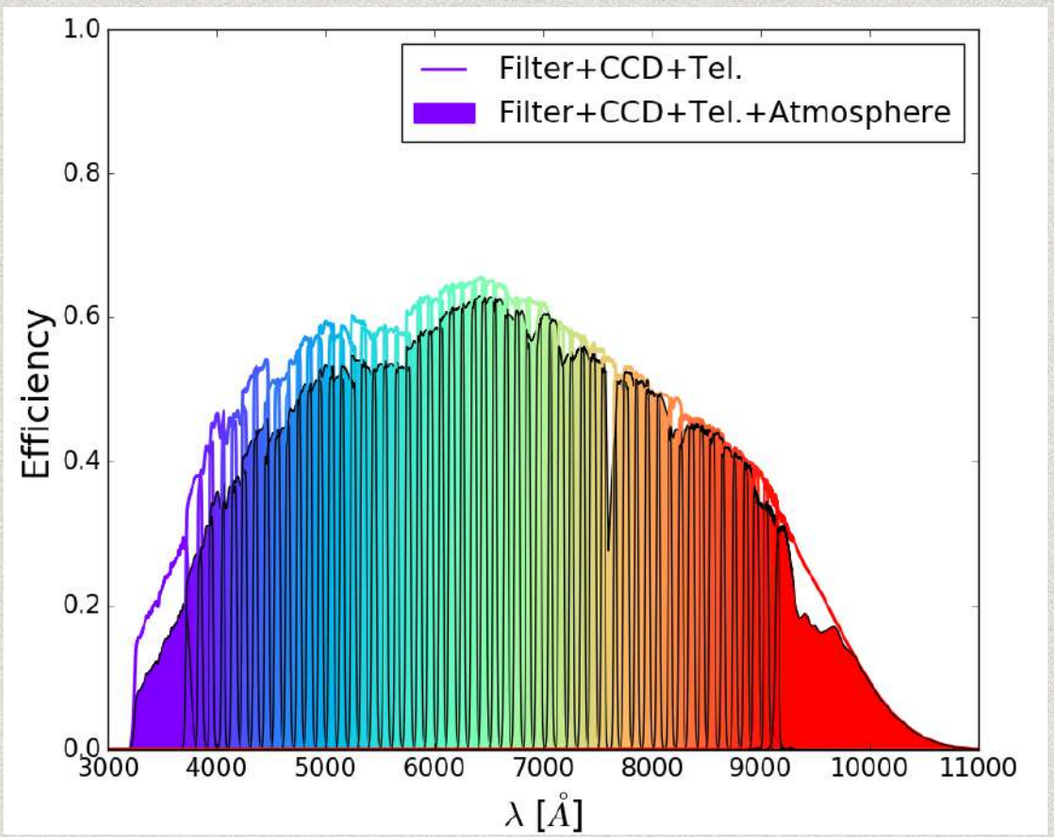
$$\sigma_F^2 = \sum \sigma_f^2 \simeq 10^3 \langle \sigma_f^2 \rangle$$

$$F = \sum f \simeq 10^3 \langle f \rangle$$

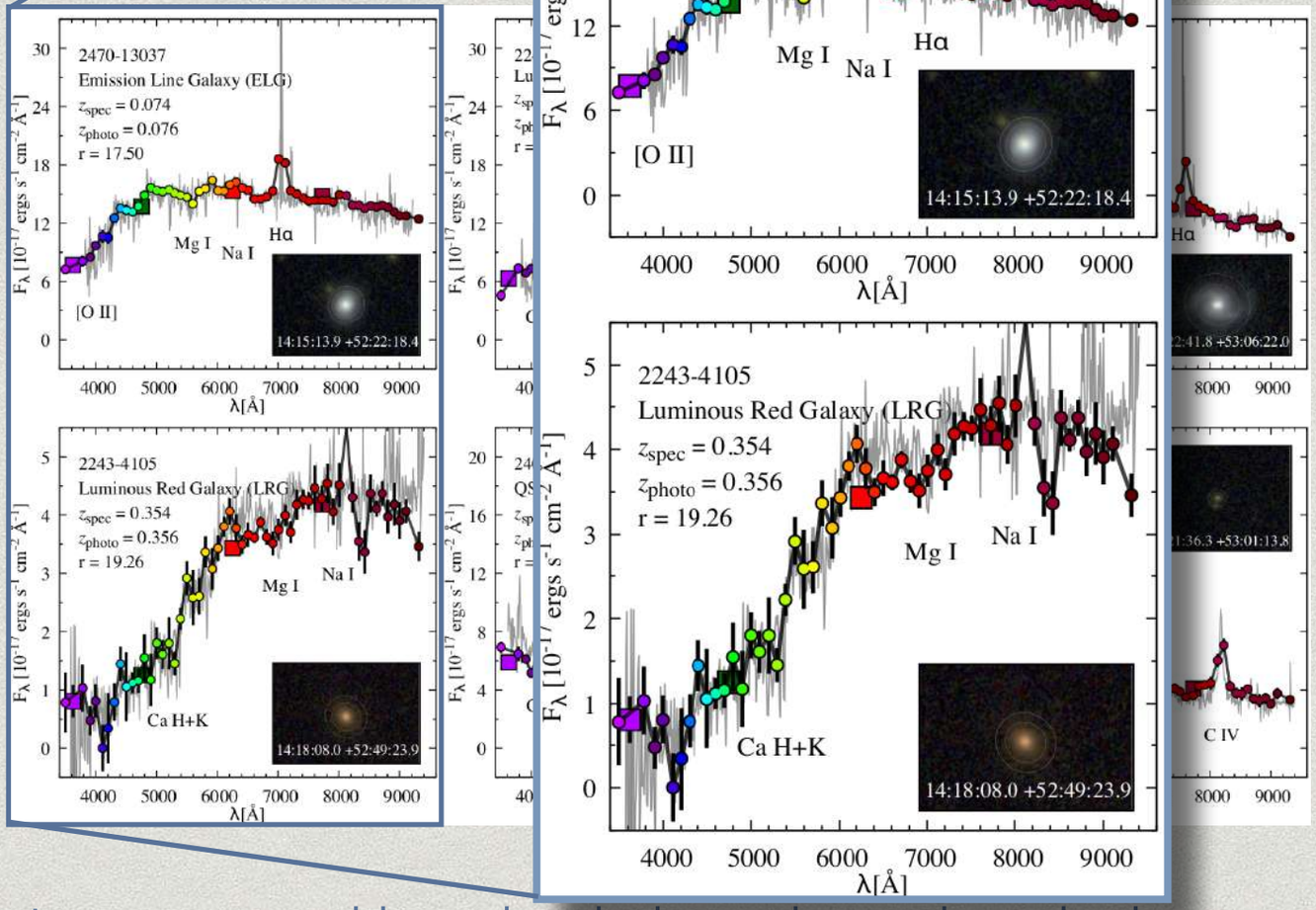
$$\Rightarrow \frac{F^2}{\sigma_F^2} \simeq 10^3 \left\langle \frac{f^2}{\sigma_f^2} \right\rangle$$, so, one **thousand** times **better SNR**!

- Of course, this comes at a price: loss of spectral resolution.

Narrow bands



t al. 2021



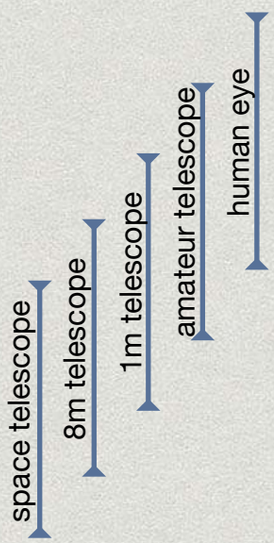
- An intermediate case, between spectroscopy and imaging in broad passbands, is when we have **narrow filters** (J-PAS, PAU, SPHEREx).
- In that case, the passbands are so narrow that we can talk about “pseudo-spectra” for the objects.
- But of course, **brighter** objects have **higher SNR**; **fainter** objects have **lower SNR**. Each telescope and survey can only go so far, and have a characteristic **magnitude limit**. And talking about magnitudes…

Fluxes and magnitudes

- The flux in a given passband is basically some number of photons, $F \leftrightarrow N_\gamma$. We can compare that number with some “standard candle” whose flux, for the same filter, is F_0 :

$$\frac{F}{F_0} \rightarrow \log_{10}\frac{F}{F_0} \rightarrow -2.5\ \log_{10}\frac{F}{F_0}$$

- This scale is called **apparent magnitude**.
- The “log scale” is similar to dB (decibels) or the Richter scale: it is calibrated in such a way that a factor of 10 in flux translates in a factor of 2.5 in magnitude, a factor of 100 in flux is a factor of 5, and so on. **Fainter** objects → **higher magnitudes**.
- Magnitudes correspond roughly to what our eyes perceive as the relationship between bright/faint objects

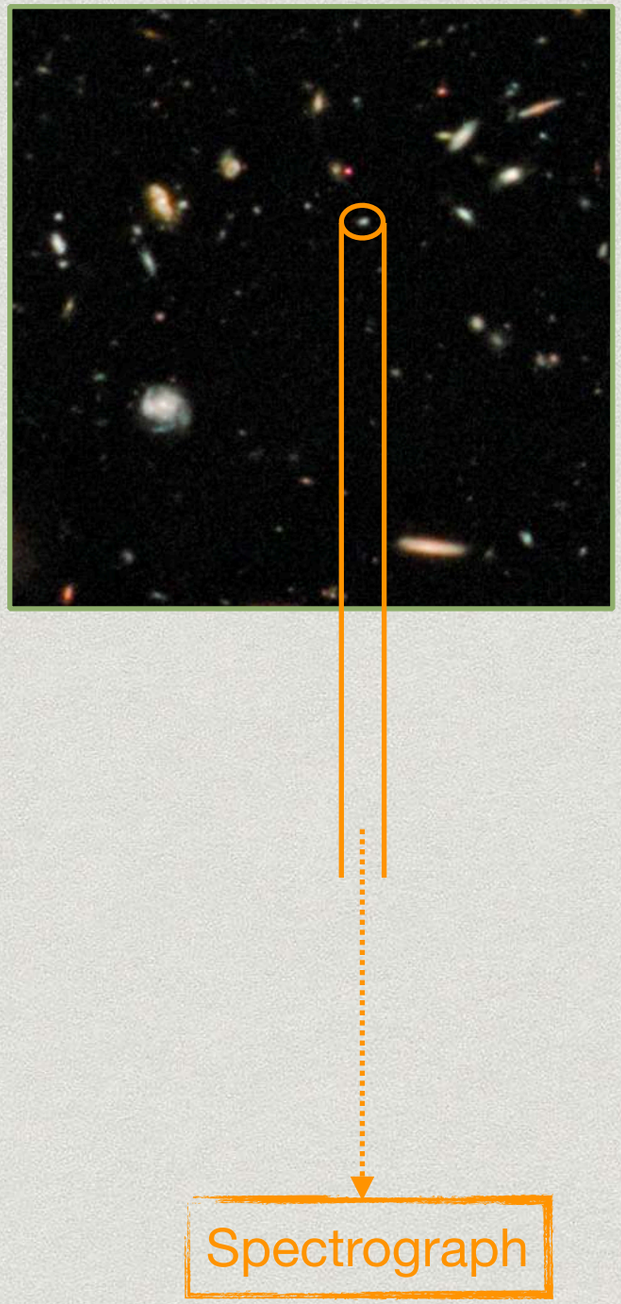


Magnitude (V)	Object
-27	Sun
-13	Full moon
-5	Venus
0	Vega* / Saturn
6	Limit for human eye on a dark night
8	Neptune
14	Pluto
20	Galaxy at z~0.5
27	Limit for an 8m telescope (on Earth)
32	Limit for the Hubble Space Telescope

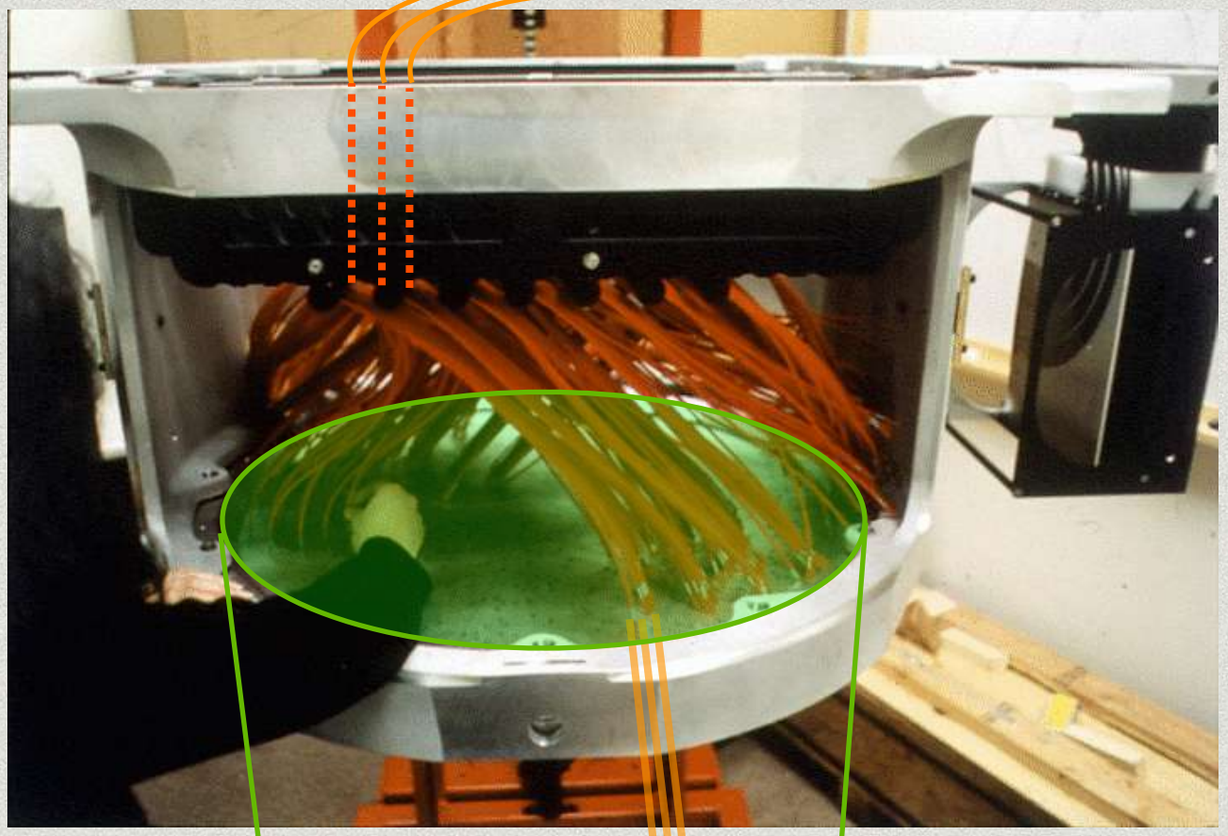
Such faint magnitudes correspond to just a **few photons/second** hitting a typical detector!

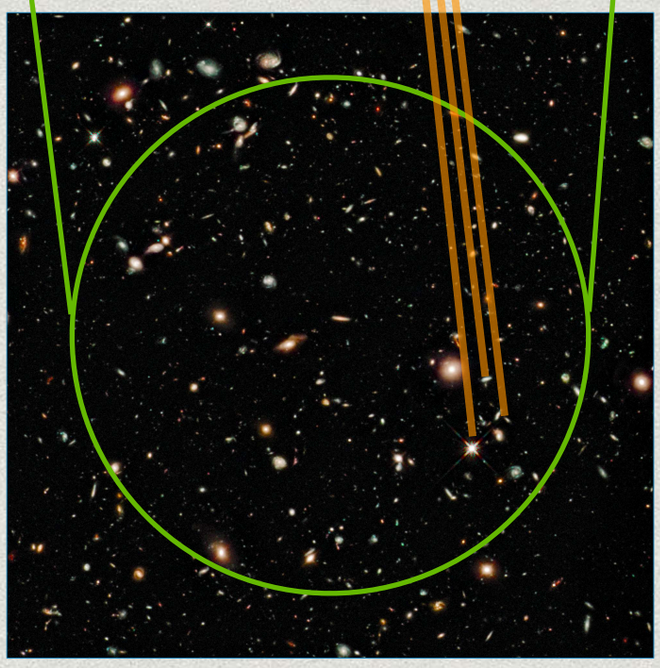
Fluxes and magnitudes

- Consider our typical image. This little square here is a tiny, **tiny** area in the sky, but you already have to **make choices**: which objects would you choose to observe with a spectrograph?
- Brighter ones or faint ones?…
- Things that look like galaxies, or other beasts?…
- What if you want to observe a quasar, but end up selecting a star by mistake — something you will only know after you take its spectrum?
- All spectrographs have to make those choices: they have to **find the targets** where they will point their instrument.
- And spectrographs are expensive instruments, that demand huge exposures (time on-target, to collect photons). Mistakes are very, very expensive…



Multi-object spectrographs

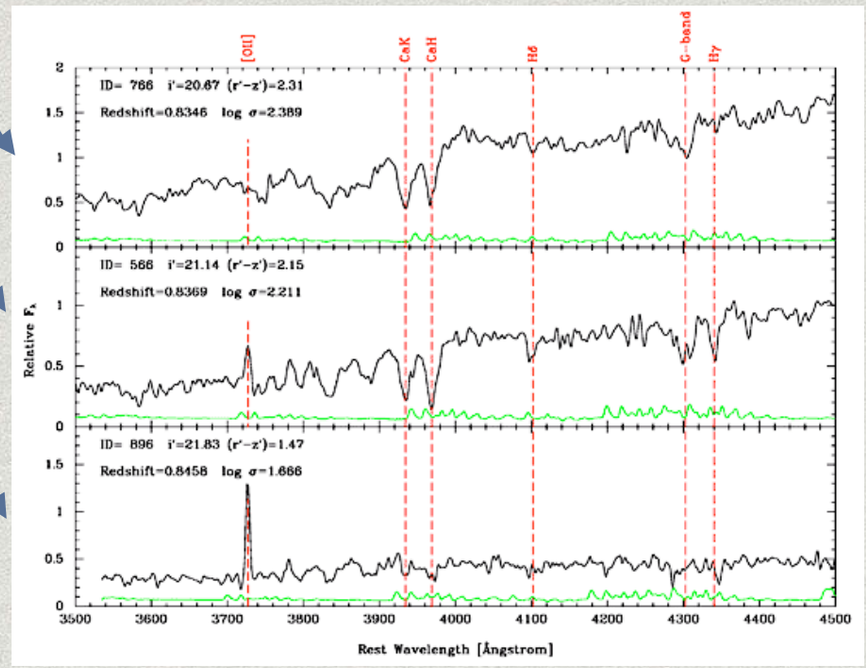




1. Select targets from imaging data

2. Place fibers on targets; point telescope and MoS; observe

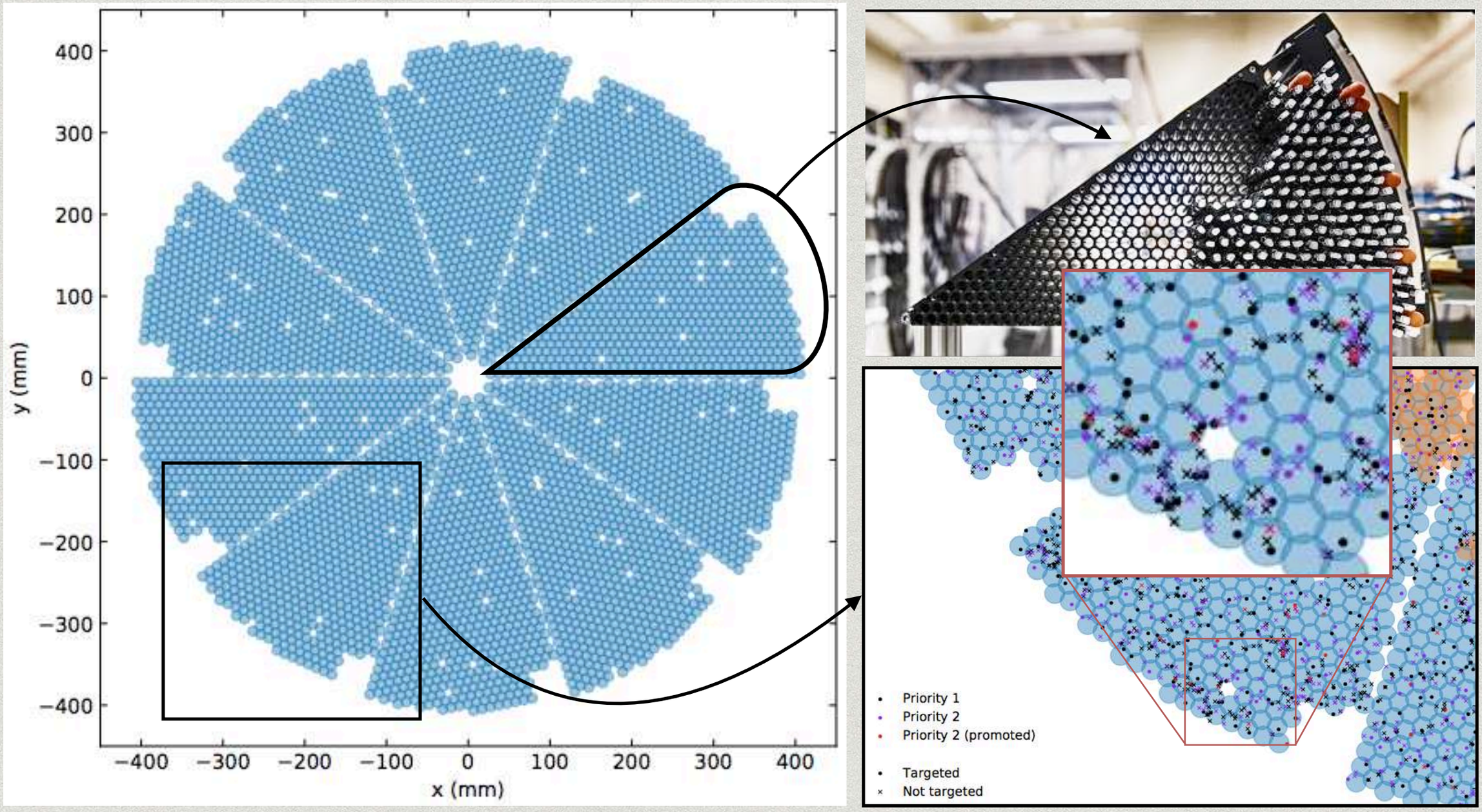
3. MoS pipes the light to CCD; acquire data; clean and reduce data



4. Play with the data!

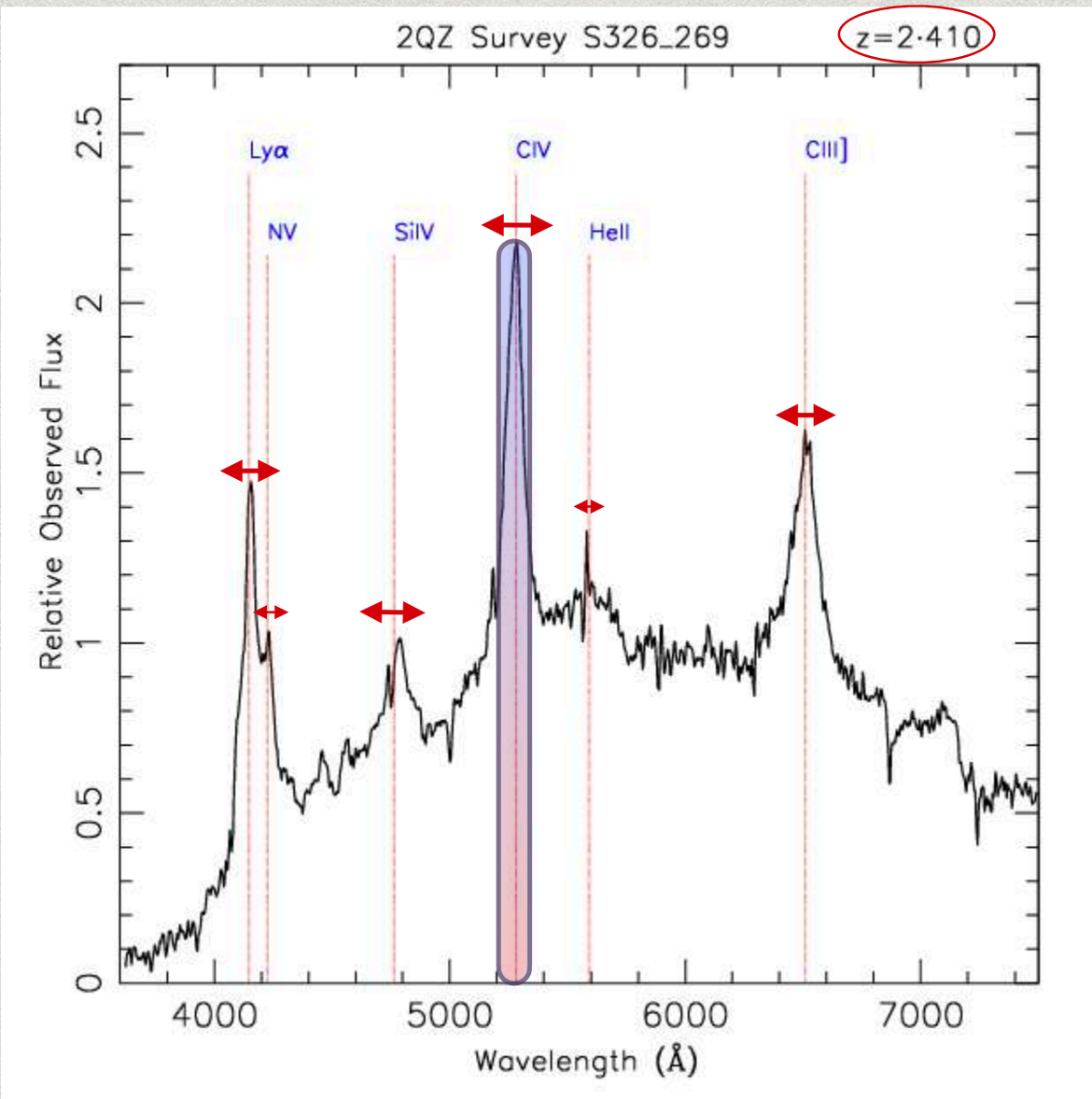
Multi-object spectrographs: Subaru's Prime Focus Spectrograph (PFS)

Multi-object Spectrographs: DESI



- Even the best MoS cannot observe **all** galaxies: they have to compromise — i.e., they must choose which targets are more interesting, and prioritize.
- E.g., down to mag(r)~23 and up to z=1 there are approx. 10.000 galaxies/deg^2 , while the top MoS manages to observe ~1/10 of that! (And that is without even counting the stars…)

From spectra to redshifts



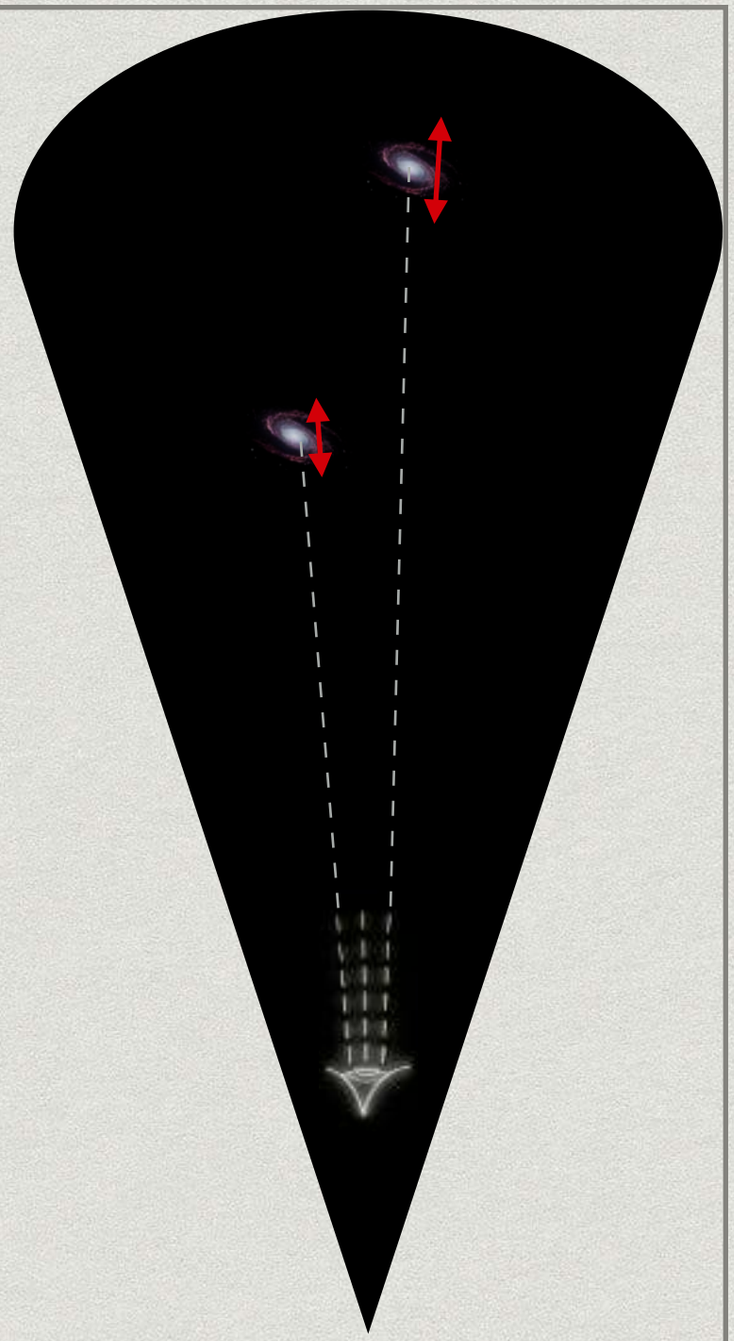
$$\frac{\lambda_{obs}^{C\,IV}}{\lambda_0^{C\,IV}} \simeq \frac{5230}{1549.01} \simeq 1 + 2.4$$

Line	λ_0 (Å)
Lyα	1215.67
N *v*	1240.15
Si *IV*+O *IV*	~1400
C *IV*	1549.01
He *II*	1640.46
C *III]*	1908.73

- The redshift of the object is obtained by fitting the positions of **several** spectral lines to their restframe values:

$$1 + z = \frac{1}{N}\sum_{i=1}^{N}\frac{\lambda^i}{\lambda_0^i} \quad , \quad \text{and the redshift error is} \quad \sigma_z^2 = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sigma_\lambda^i}{\lambda_0^i}\right)^2$$

From spectra to redshifts



- So, even with ideal spectra we have errors in the redshifts of galaxies and quasars. With typical "cosmology-grade" spectra (with resolutions of ~ 1 Å, and spectral bins with SNR $\sim 1-10$, these errors are:

  Galaxies: $\sigma_z \sim 0.0001\,(1+z)$

  Quasars: $\sigma_z \sim 0.0005\,(1+z)$

- The factor $1+z$ denotes the fact that all spectral features are broadened by the same redshift factor as the wavelengths themselves:

  $\lambda_0 \to (1+z)\lambda_0$, and $\lambda_0^1 - \lambda_0^2 = \Delta\lambda_0 \to (1+z)\Delta\lambda_0$

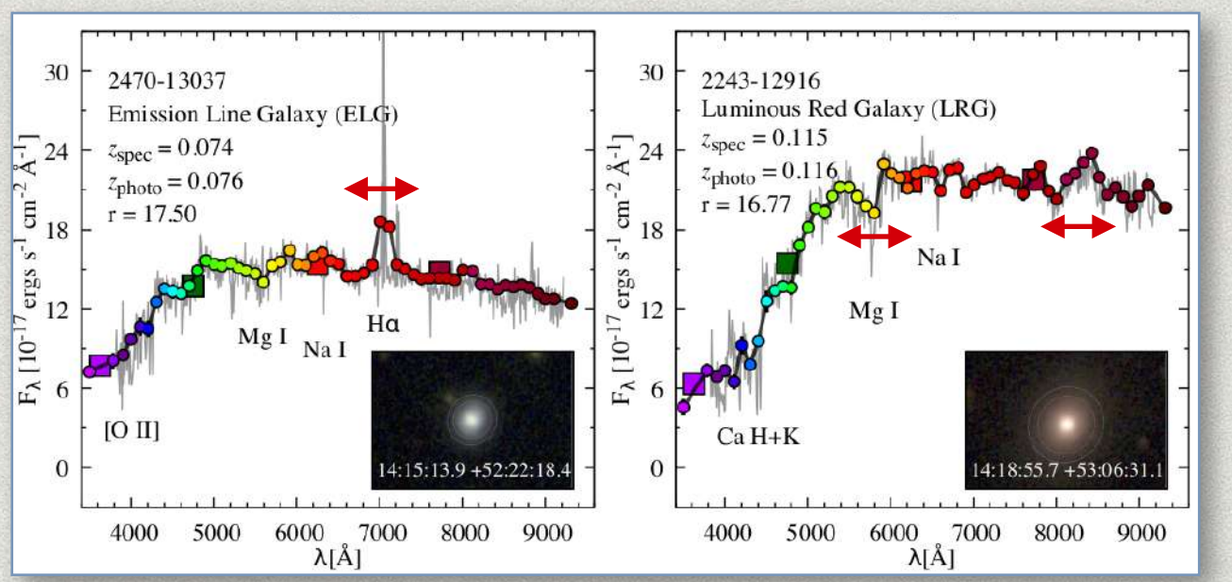
- What are the corresponding errors in the radial distances implied by these redshift errors? Remember that:

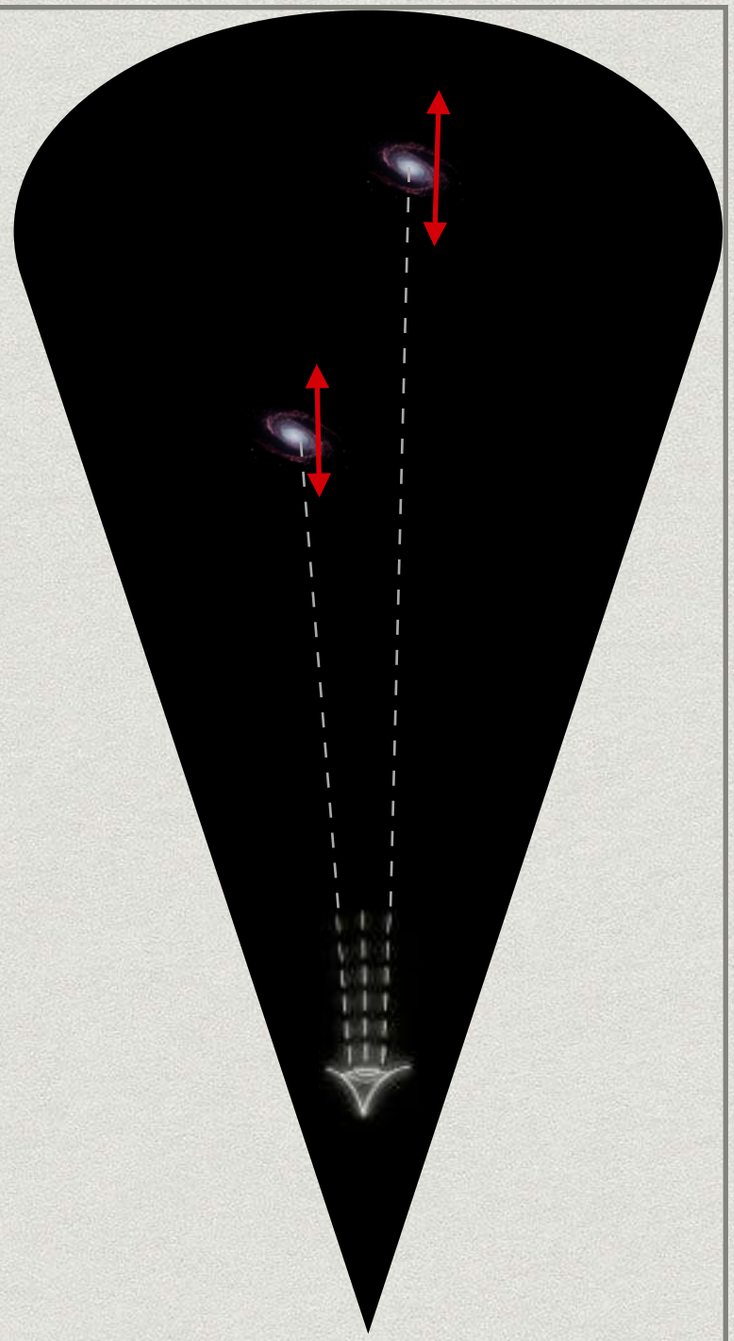
$$\chi = \int \frac{c\,dz}{H(z)} \quad \Rightarrow \quad d\chi = \frac{c\,dz}{H(z)} \quad \Rightarrow \quad \sigma_\chi = \frac{c\,\sigma_z}{H(z)}$$

Since $c/H_0 = 2997.9\,h^{-1}\,\mathrm{Mpc}$, and $H(z)$ increases with redshift (roughly as $\sim 1+z$), we get typical values of $\sigma_\chi \simeq 0.3\ h^{-1}\,\mathrm{Mpc}$ for galaxies, and $\sigma_\chi \simeq 1.5\ h^{-1}\,\mathrm{Mpc}$.

Photometric redshifts

- Often we have only fluxes in some given bands — not spectra.





- Depending on the filters, the photometric redshifts (photo-zs) can have very large errors:

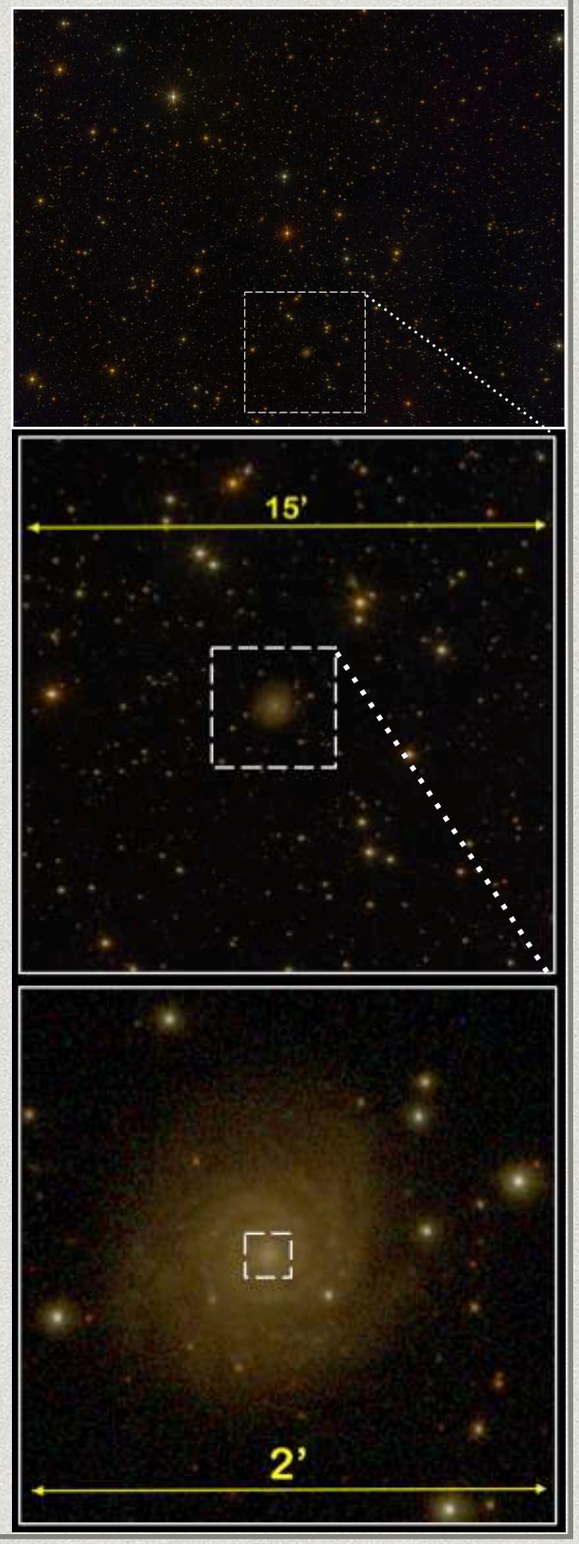
Galaxies: $\frac{\sigma_z}{1+z} \sim 0.003 - 0.1$

Quasars: $\frac{\sigma_z}{1+z} \sim 0.002 - 0.1\,(1+z)$

- This means a much larger error in the radial positions of these objects: $\sigma_\chi \simeq 10 - 100\ h^{-1}\,\mathrm{Mpc}$!

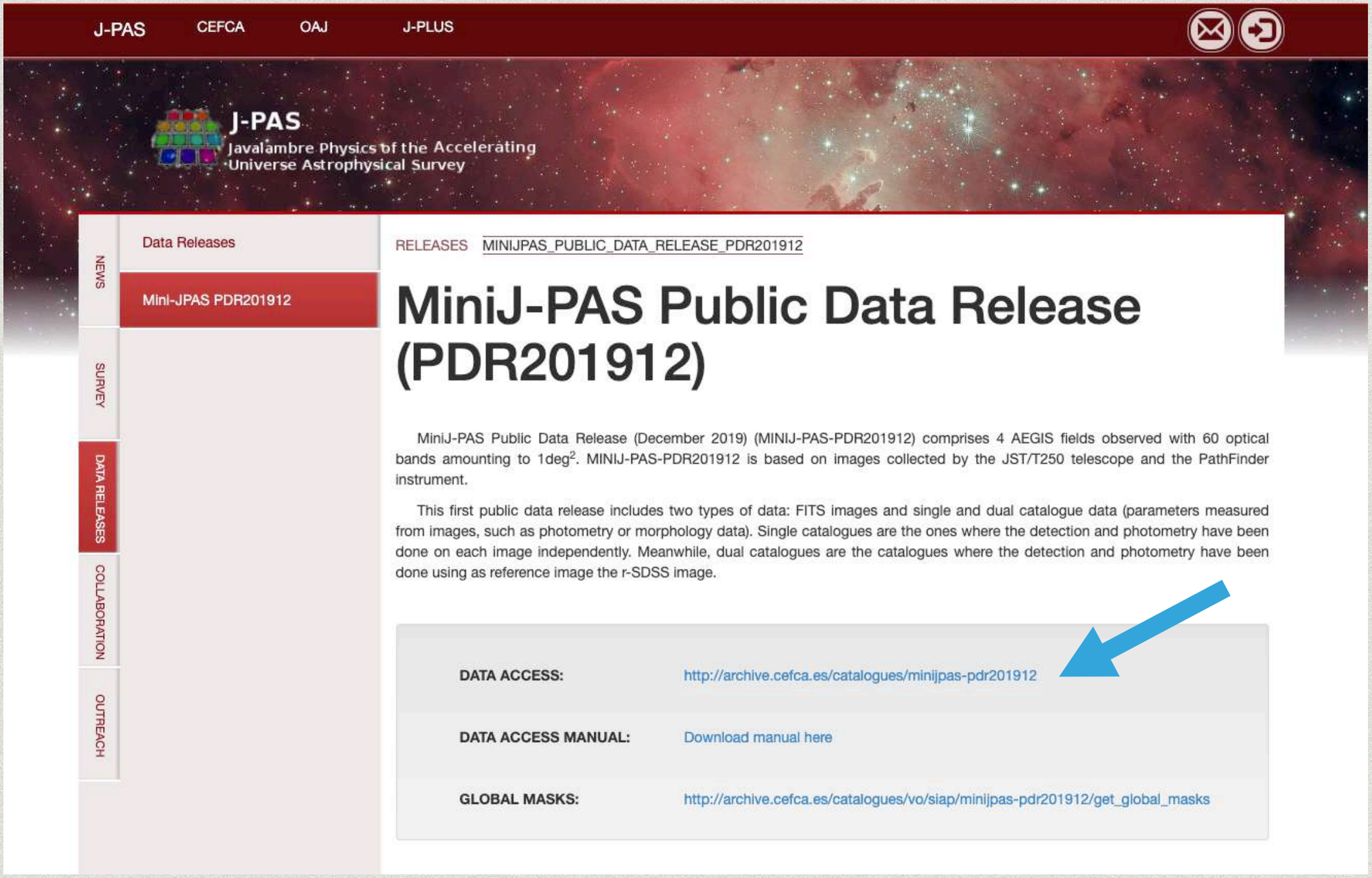
Back to basics: classification

- Even before we can even think about trying to take a spectrum, we have to make a first attempt at identifying an object in an image: i.e., we must classify the sources in our data sets into:
  - Star
  - Galaxy
  - Quasar
  - Artifact
  - Satellite/asteroid
  - Etc.
- This is far from trivial, even with narrow bands (pseudo-spectra). Here is an example directly from the best survey for this exercise — J-PAS.

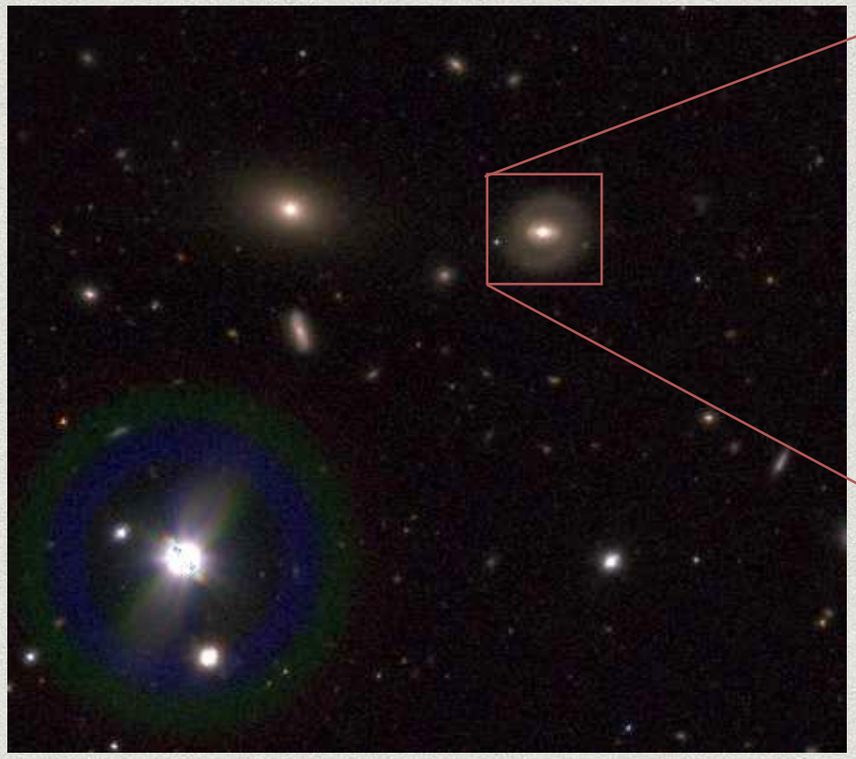


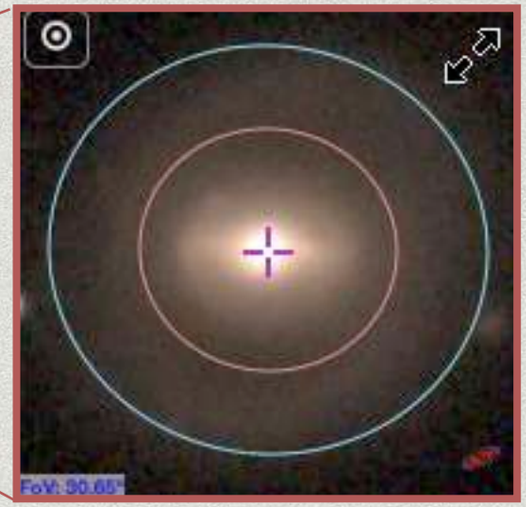
Mini-JPAS Public Data Release

http://www.j-pas.org/datareleases/minijpas_public_data_release_pdr201912

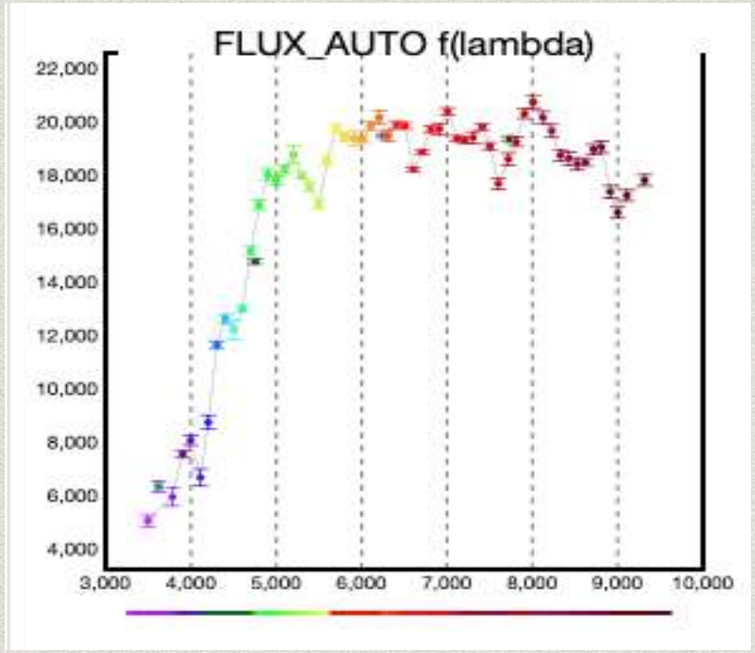


Template fitting



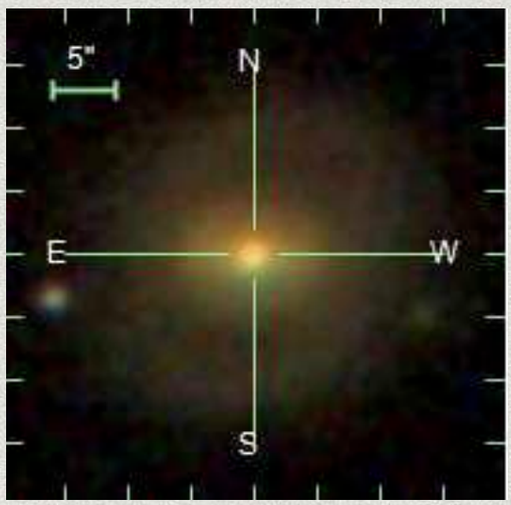


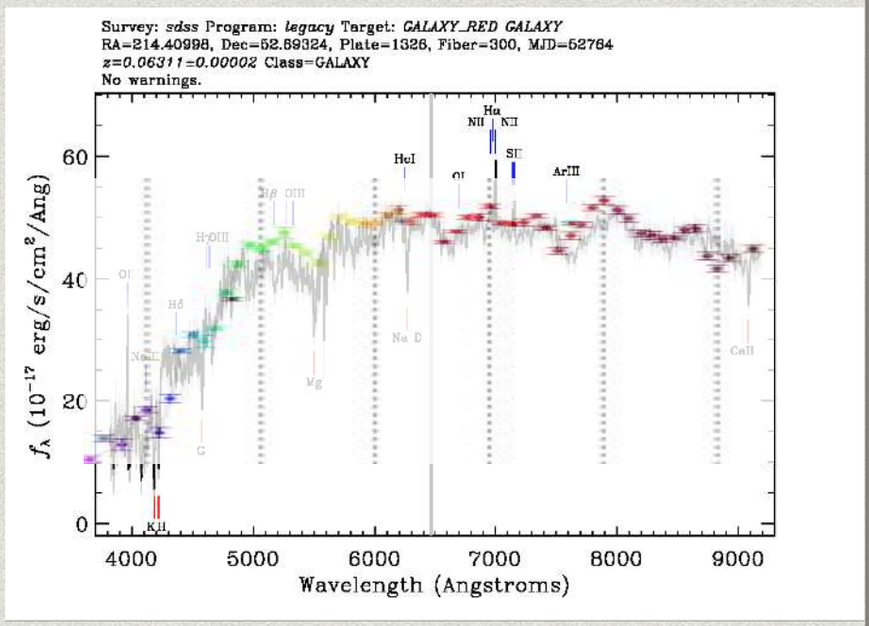
J-PAS



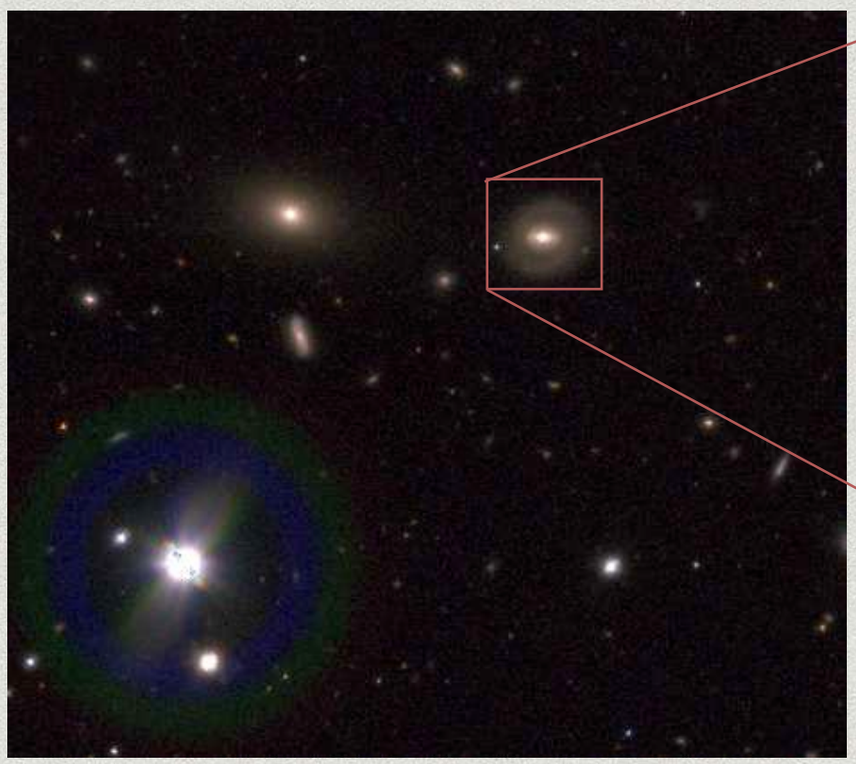
$z = 0.063$

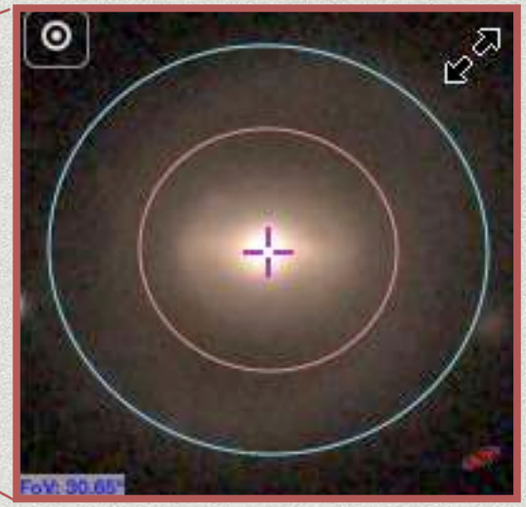
SDSS



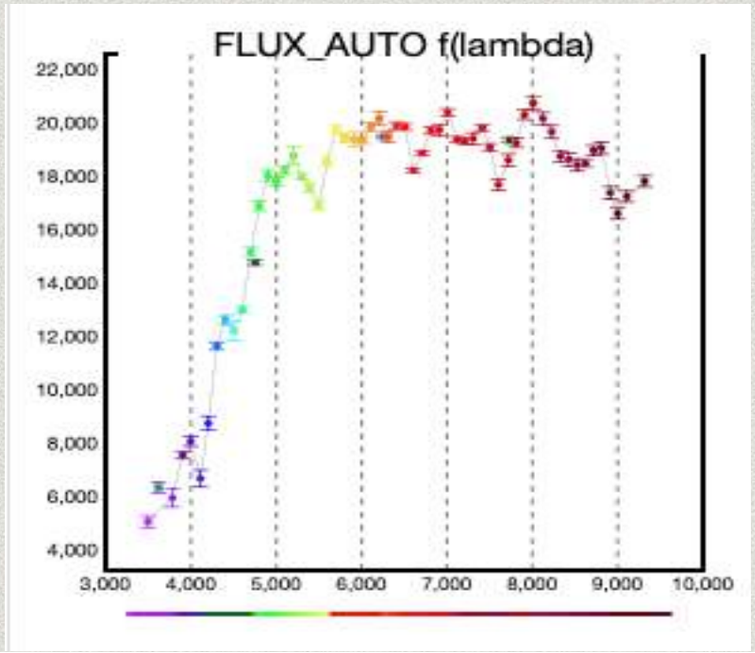


Template fitting



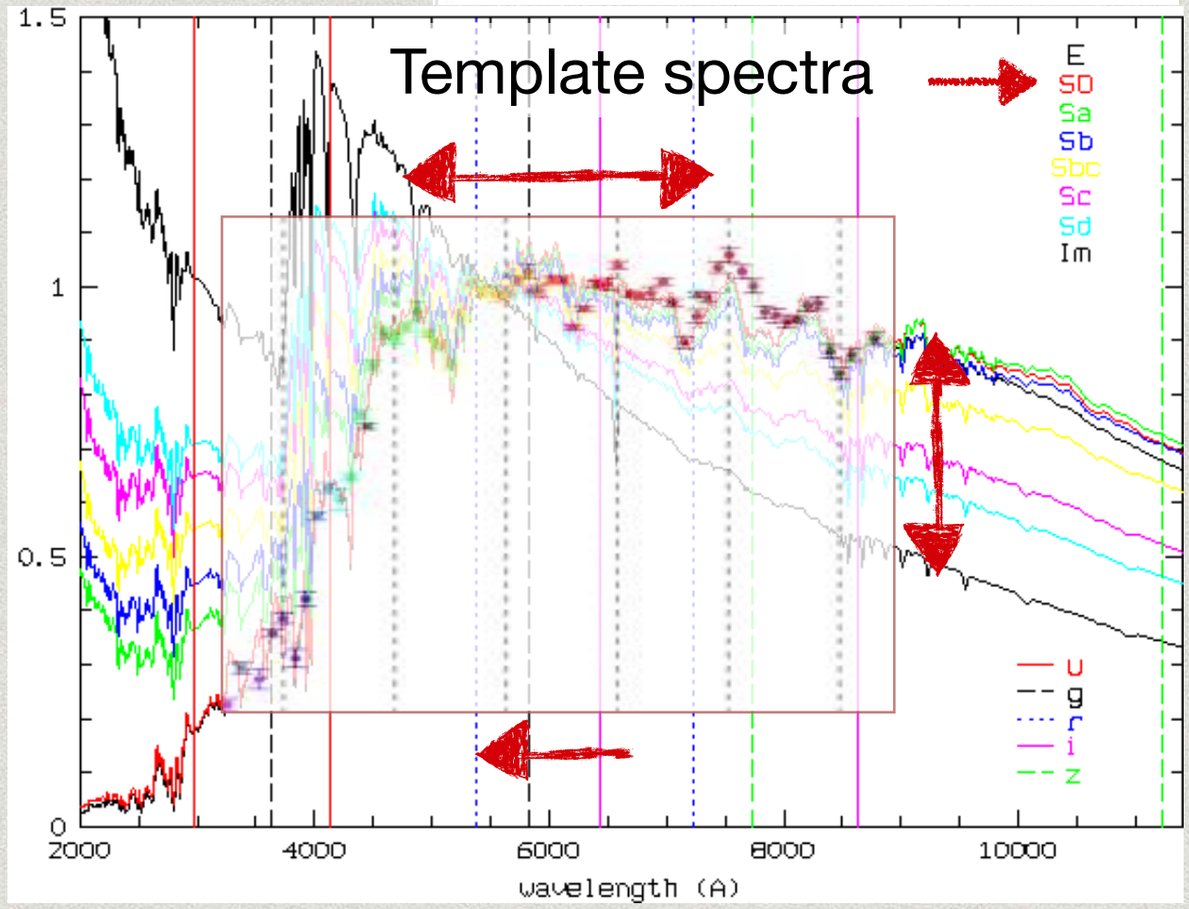


J-PAS



Many codes out there that fit pseudo-spectra to ***templates***, and try to determine the class and redshift of the objects:

- LePhare
- TPZ
- BPZ
- ANNz
- …

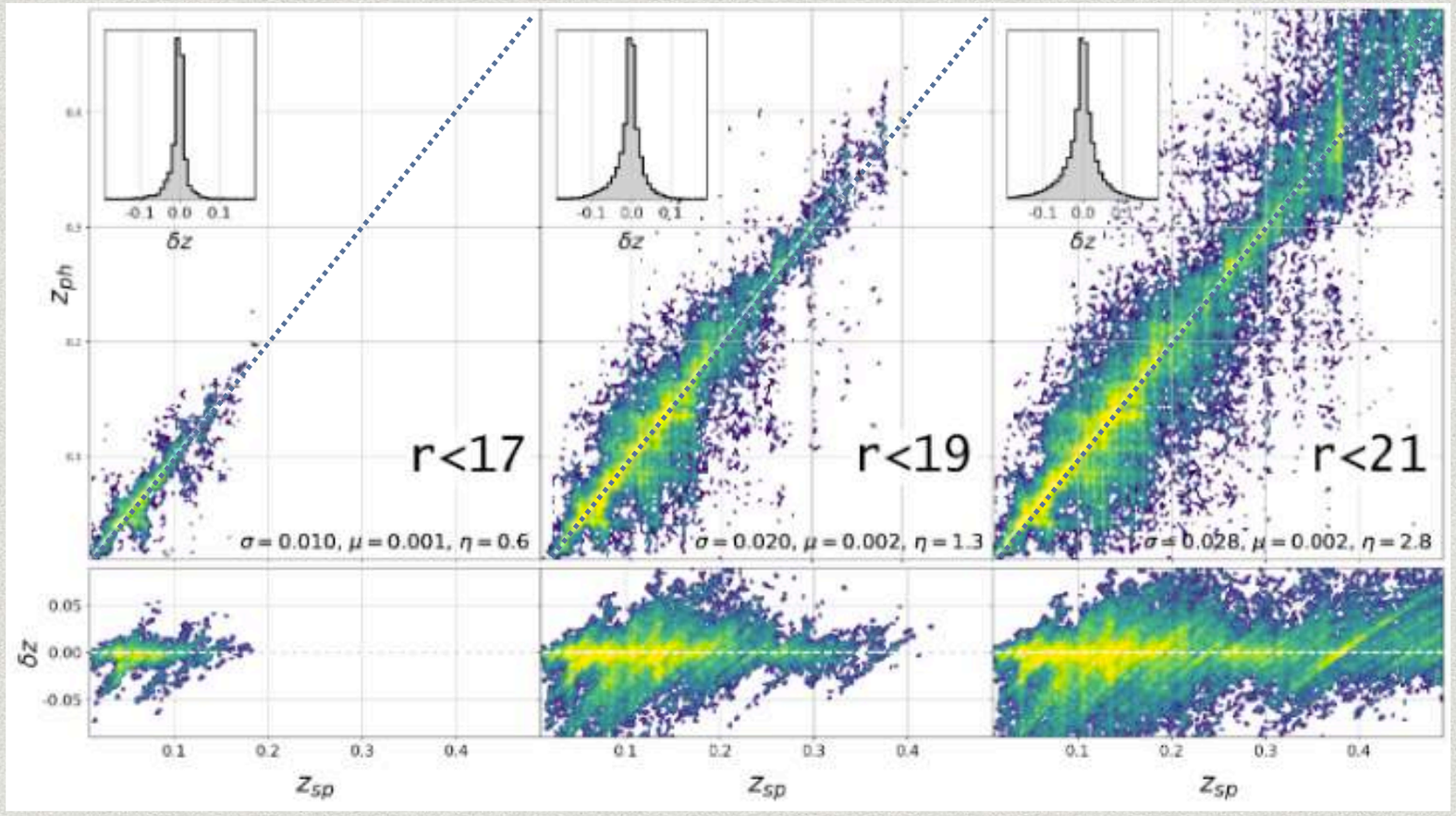


Template fitting: redshifts

- We can check how well we are estimating the redshifts by comparing with a sample of objects for which we have spectra. The plots below show:

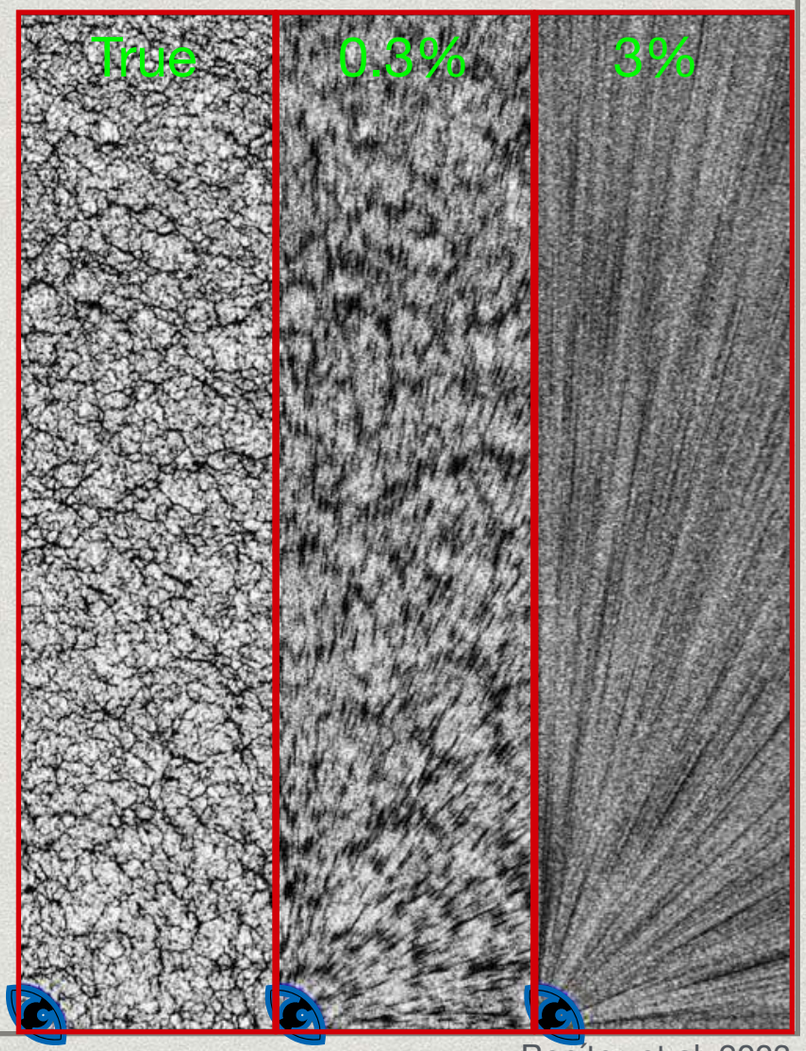
$$z_{spec}\ \text{(true)} \quad \times \quad z_{ph}\ \text{(estimated)}$$

S-PLUS



Molino et al. 2020

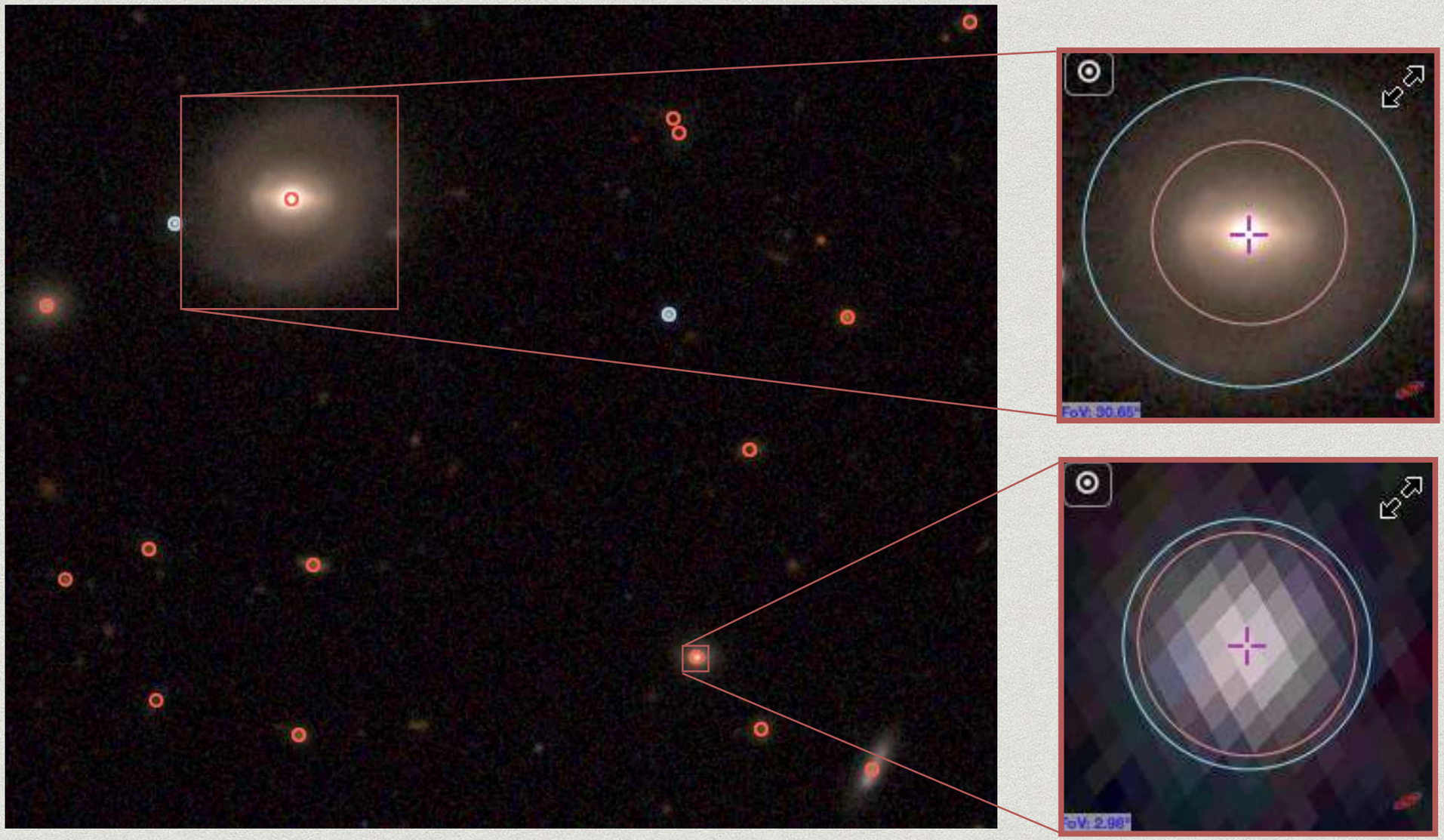
PAU/J-PAS



Benitez et al. 2003

Template fitting: classification

- But even more basic than the redshifts, we first should determine **what** an object is: star? Galaxy? Other?…

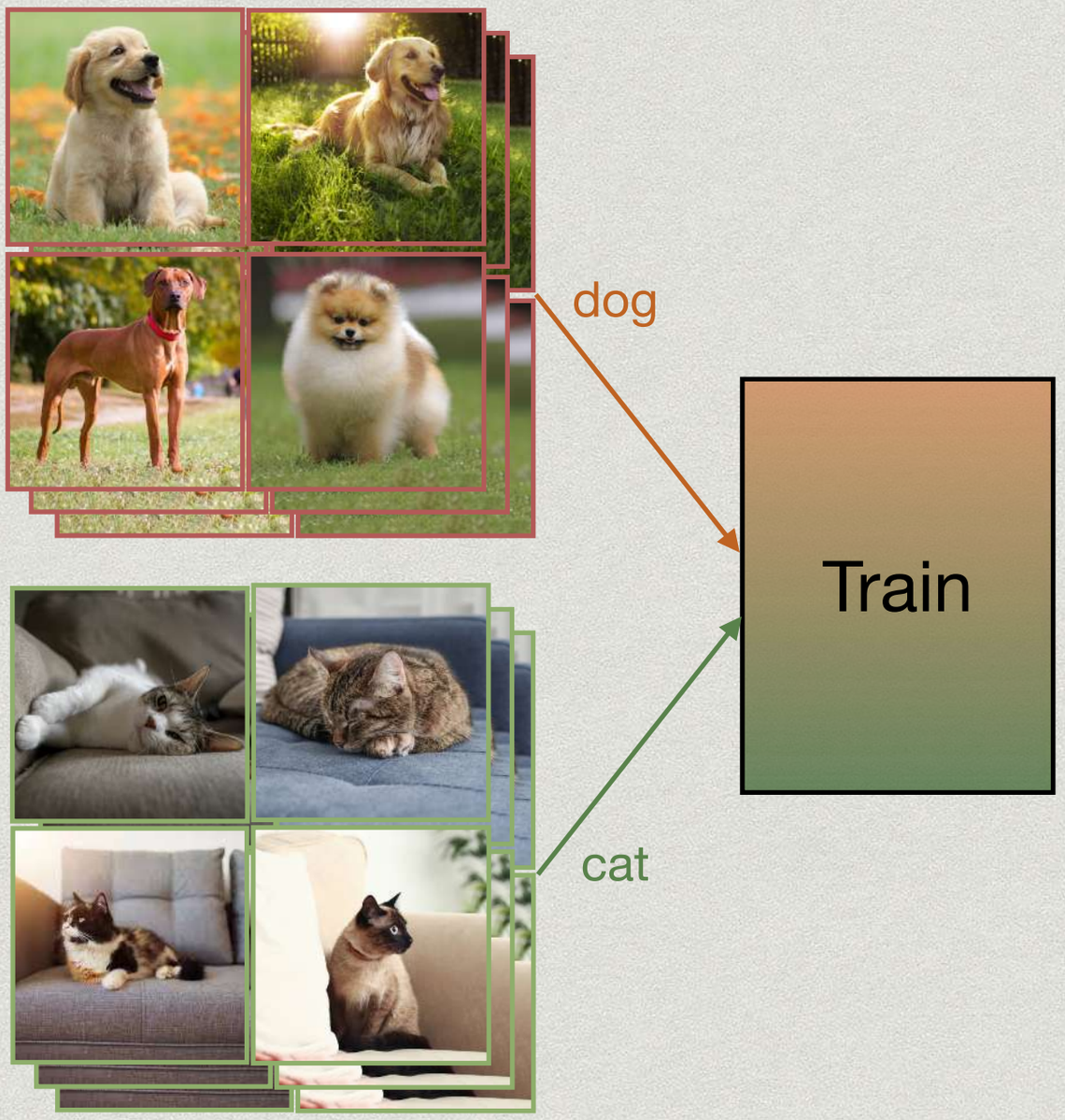


Galaxy ✓

??

Machine Learning

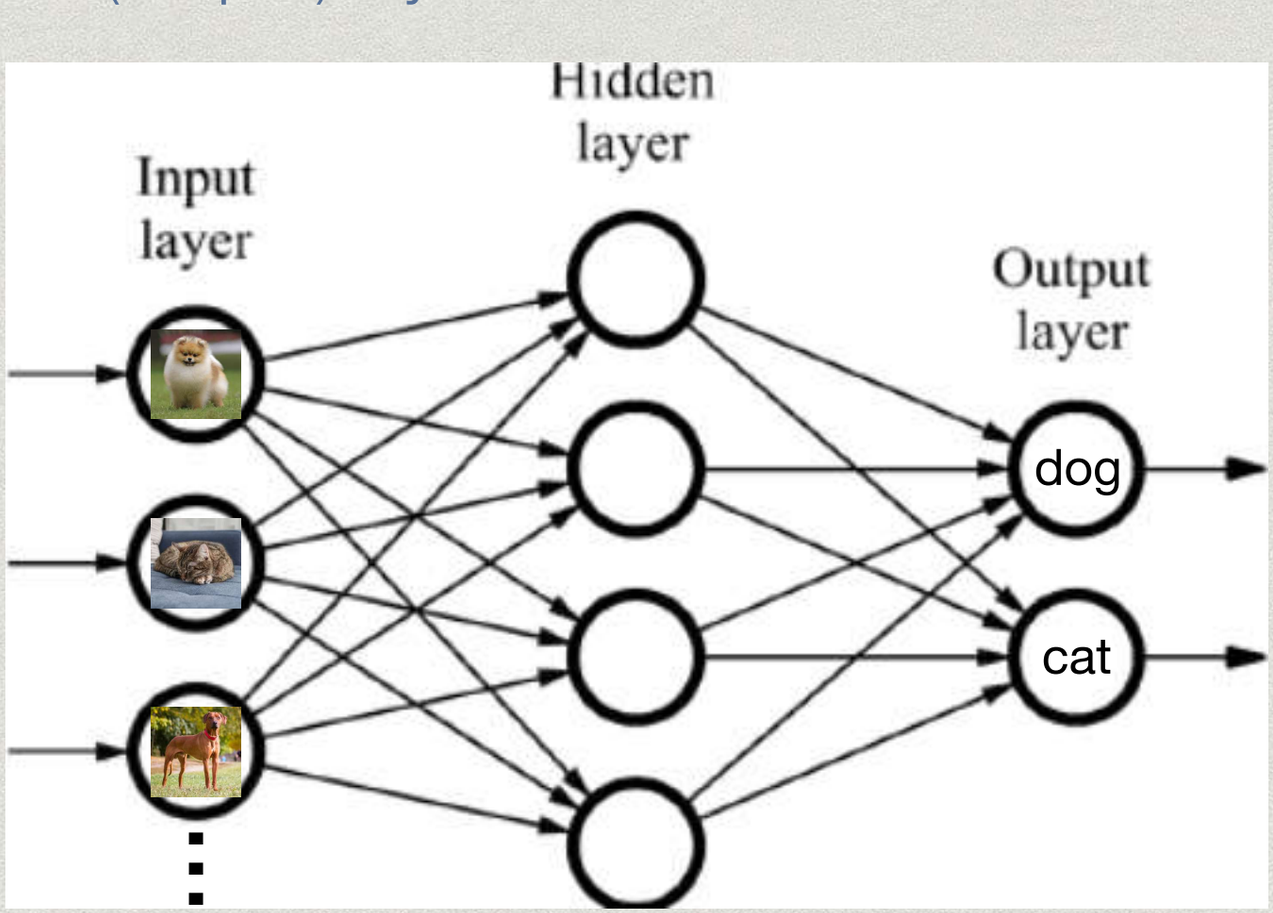
- Classification of “images" is something that all the world has been doing for a long time, in many different areas: social media, driverless cars, security, etc. etc. etc.
- The idea is to train a “machine”, using certain algorithms that can “learn" from examples.



- As we provide more and more examples, the algorithm tunes its parameters, and becomes better at the task
- This progressive tuning of the parameters (which happens through the magic of “backpropagation”) is what allow the “machine" to “learn" from the examples
- Machine learning is, therefore, not "artificial intelligence”: it is the process of learning by example.
- A number of tools are now available for anybody to use: SciKit Learn, Keras, TensorFlow,…
- New dedicated hardware (GPUs) have accelerated the learning process by huge factors.
- Often you don’t even need a computer: Google CoLab allows anyone with access to a browser to build very sophisticated Machine Learning methods.

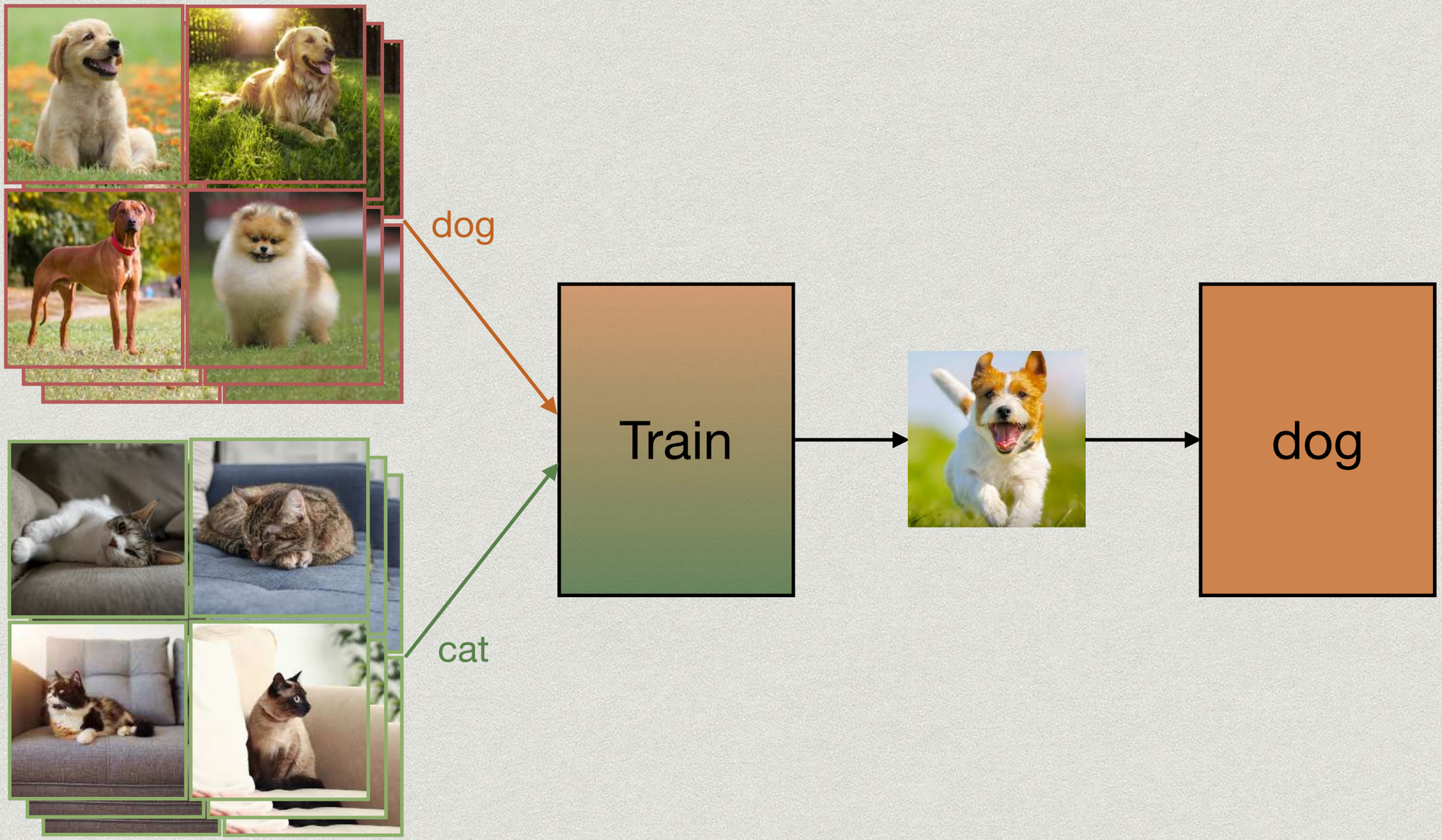
Machine Learning

- The simplest structure for a ML code is a ***neural network***
- The information is provided through the **input layer**
- That input layer feeds an **internal (hidden) layer**, which combines the input data using some **weights** in some nodes (***neurons***)
- The outputs of the nodes of the hidden layer are then combined to produce the final (output) layer



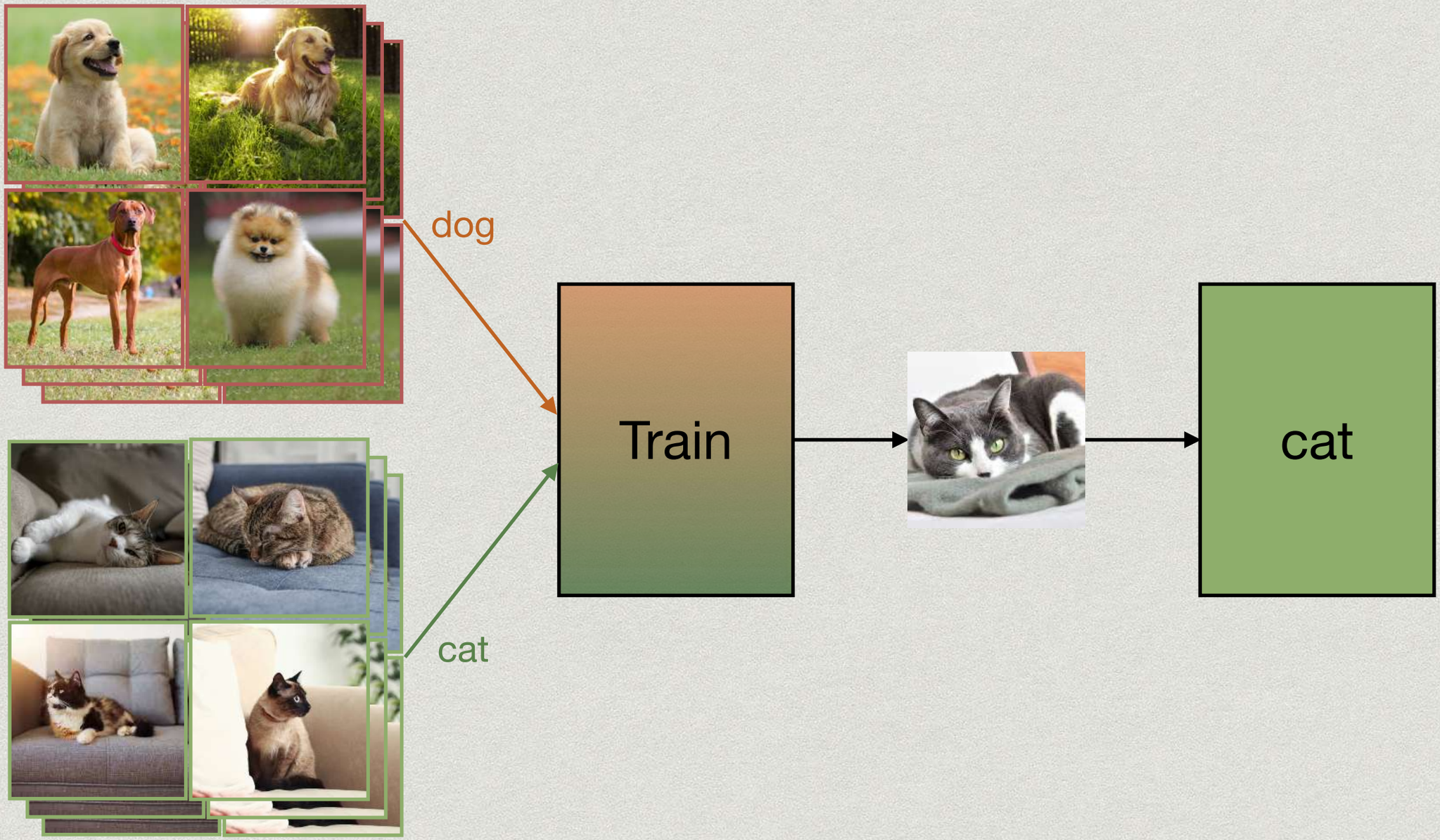
Machine Learning

- After the “machine" has “learned” from all our examples, we start **testing** it on new data
- If this new data looks like the data we used in the training process, then the results are typically very positive



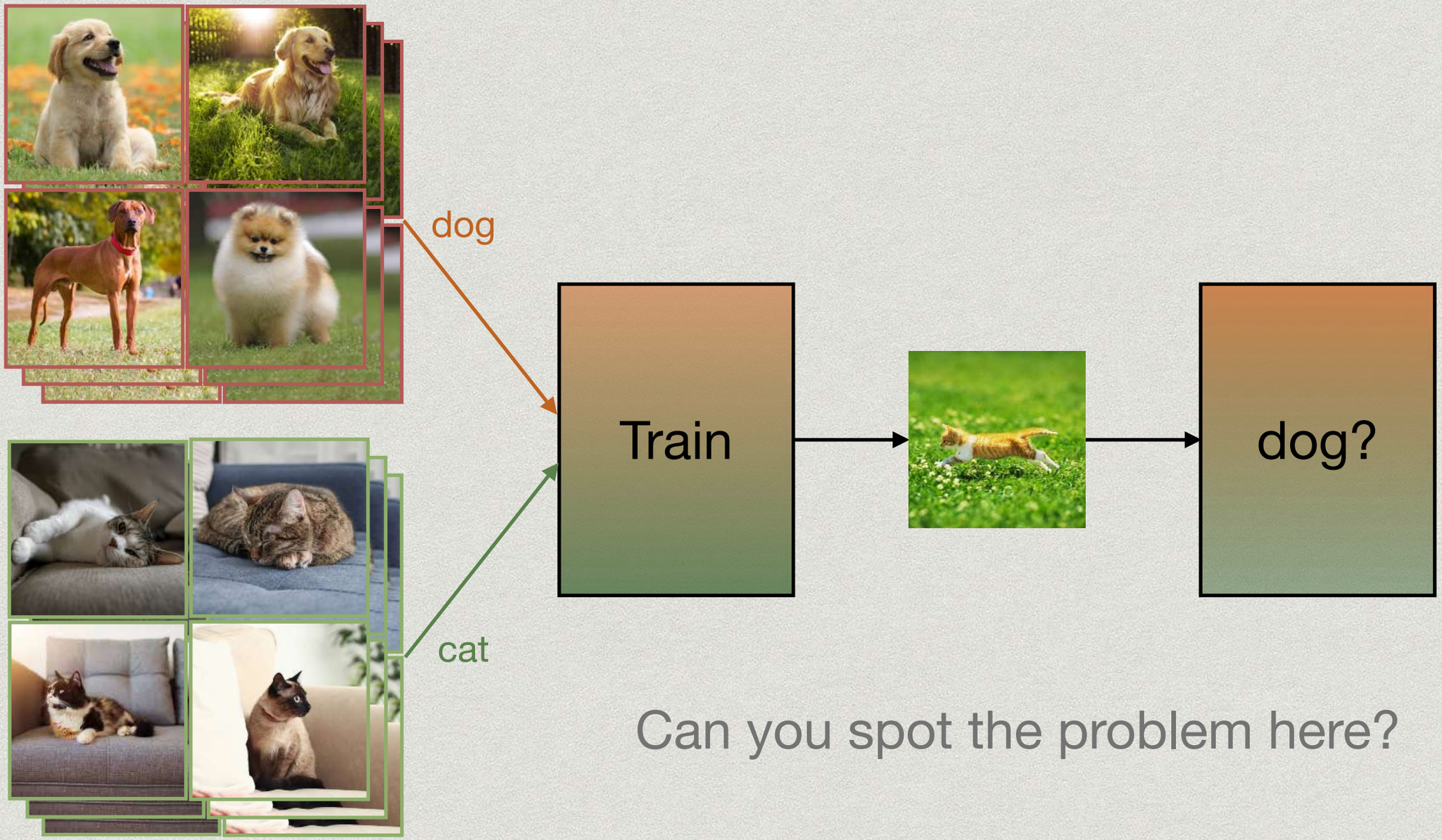
Machine Learning

- After the “machine" has “learned” from all our examples, we start **testing** it on new data
- If this new data looks like the data we used in the training process, then the results are typically very positive



Machine Learning

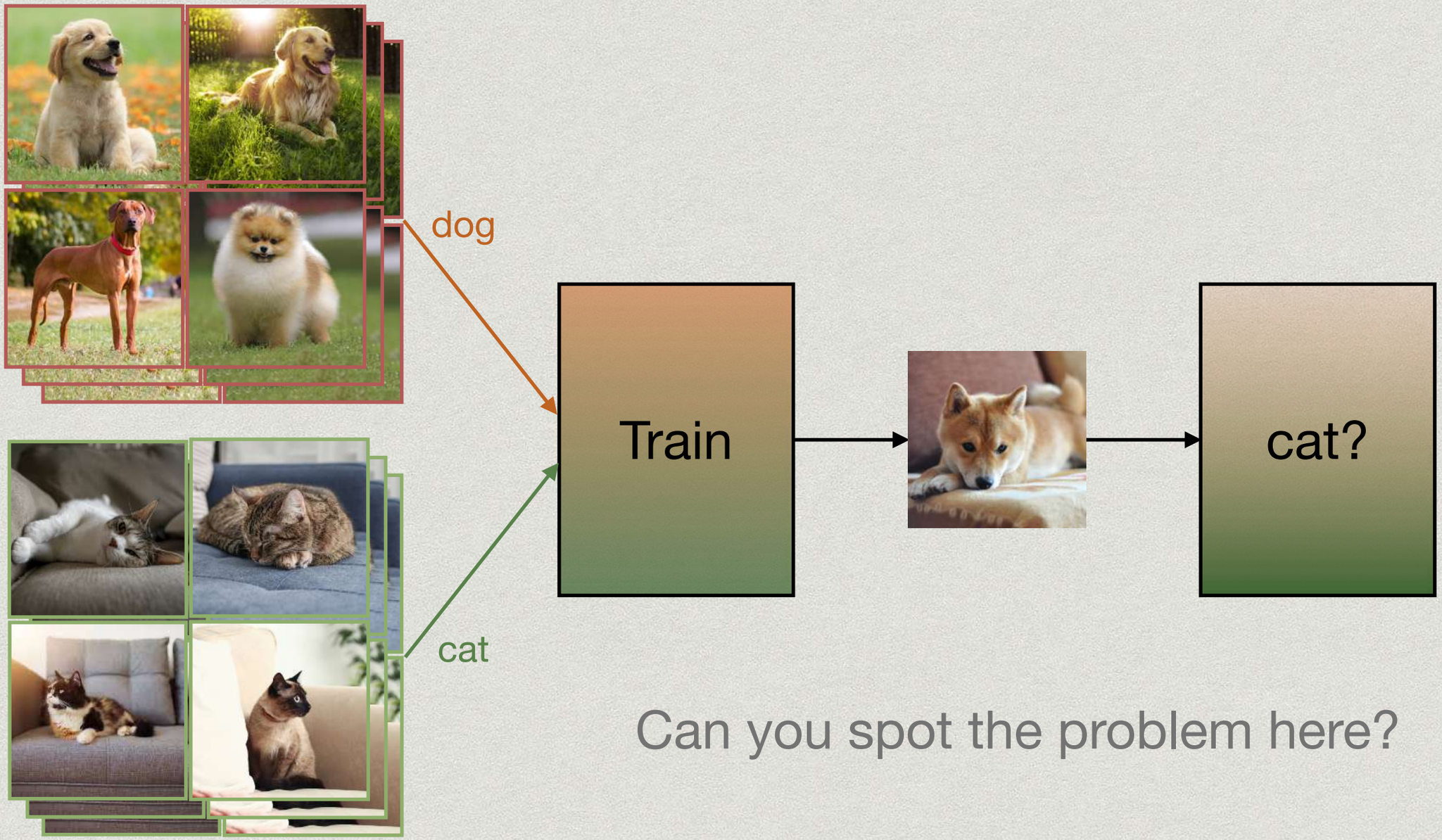
- However, sometimes we use an example (or instance) to test the ML method and we get a wrong result .
- Often these examples reveal underlying issues with our training sets!



Can you spot the problem here?

Machine Learning

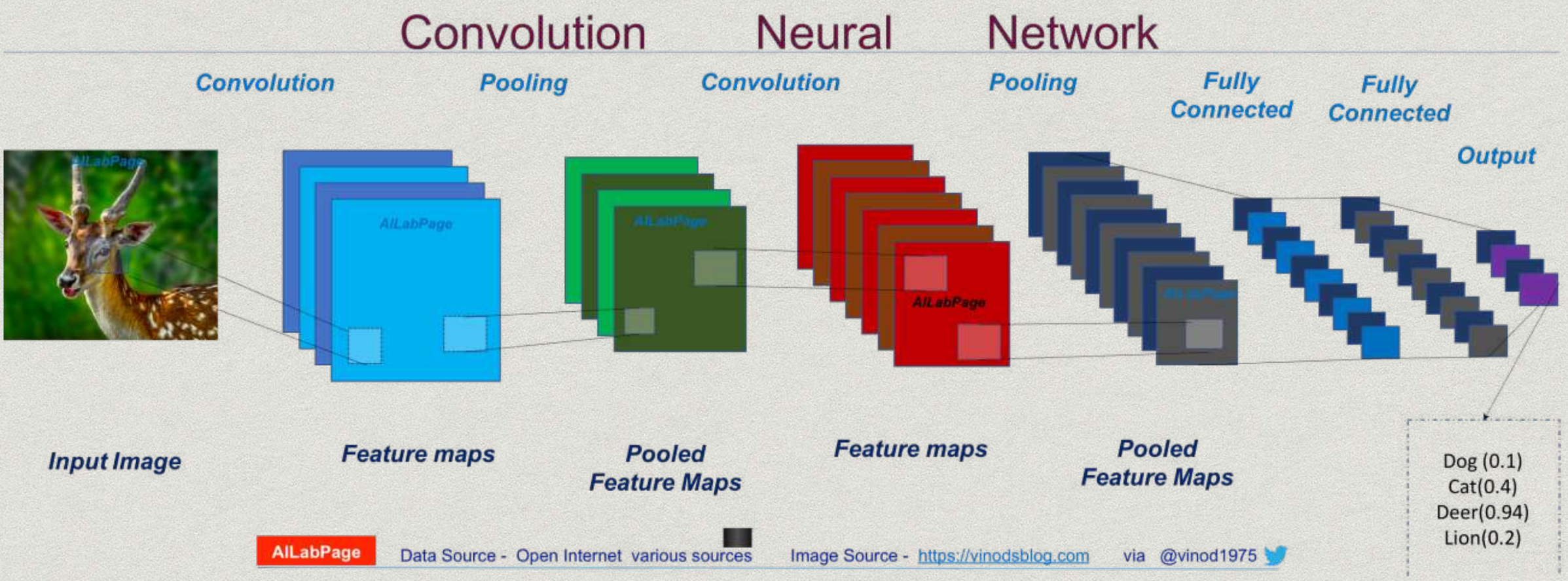
- However, sometimes we use an example (or instance) to test the ML method and we get a wrong result .
- Often these examples reveal underlying issues with our training sets!



Can you spot the problem here?

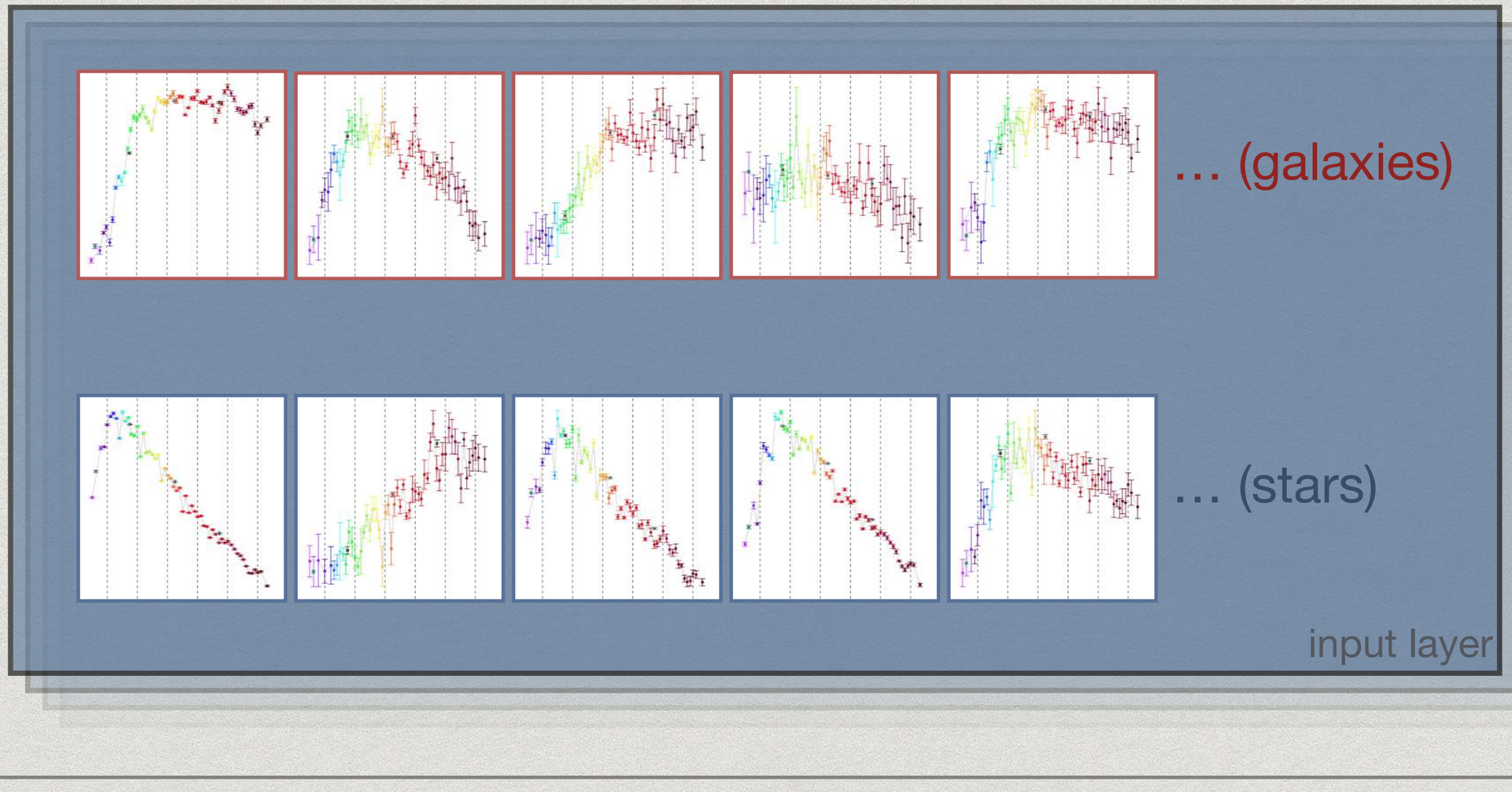
Convolutional Neural Networks

- Some of the more popular (and more powerful) ML methods attempt to detect multiple features at the same time, by using convolutional filters, pooling, dropout layers and other techniques
- These are called convolutional neural networks — and their extension, deep networks



Machine Learning in Astrophysics

- The most basic information that we can feed to a ML code are the **fluxes** in the filters
- Other information can also be passed on (morphology, variability, etc.)



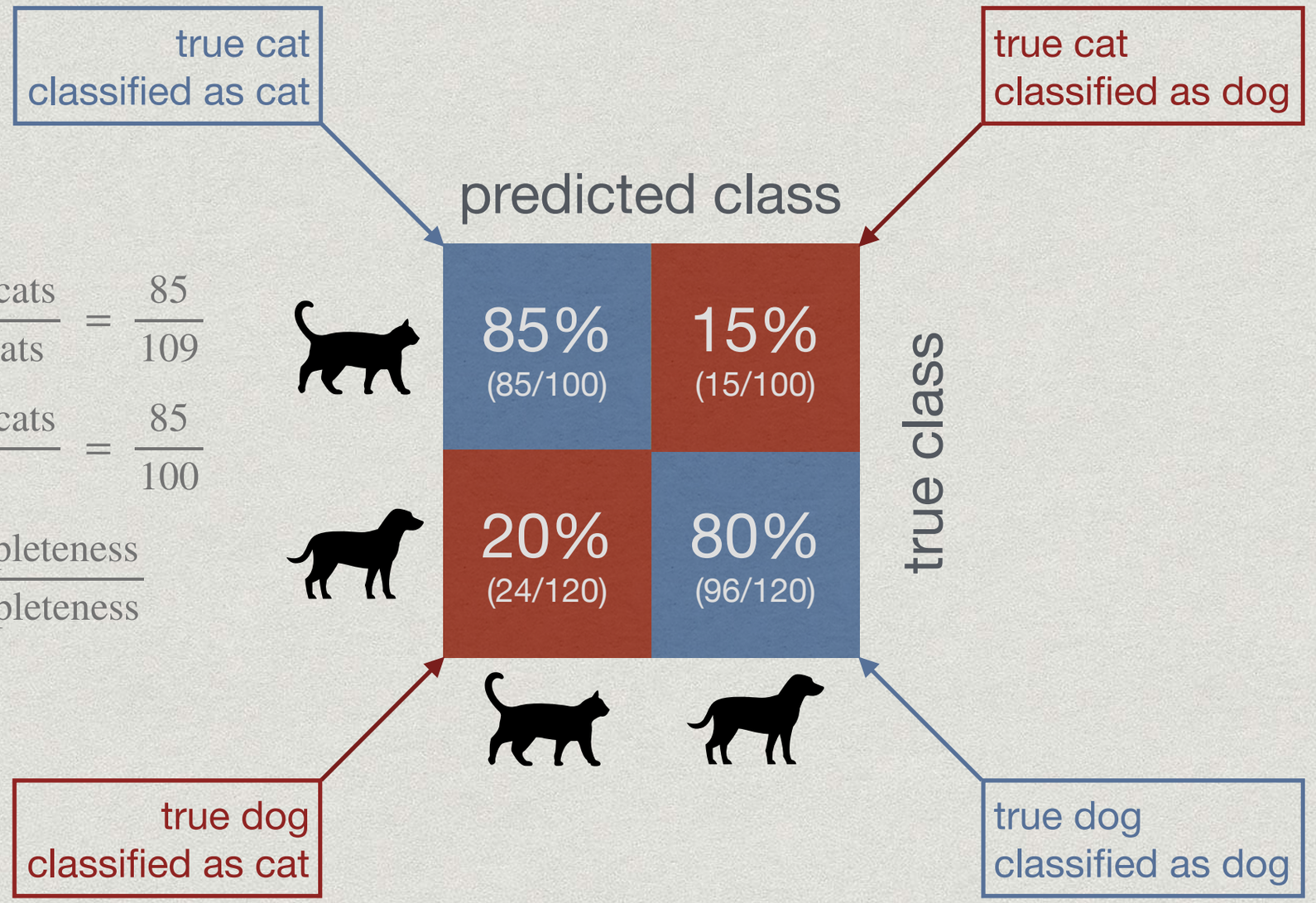
Certainty and confusion

- The success and failure of the methods can be visualized in terms of **confusion matrices: true v. predicted classes** for a **test sample**.
- These matrices allow us to monitor the **performance** of the ML classifier.

$$\text{precision/purity} = \frac{\text{predicted true cats}}{\text{all predicted cats}} = \frac{85}{109}$$

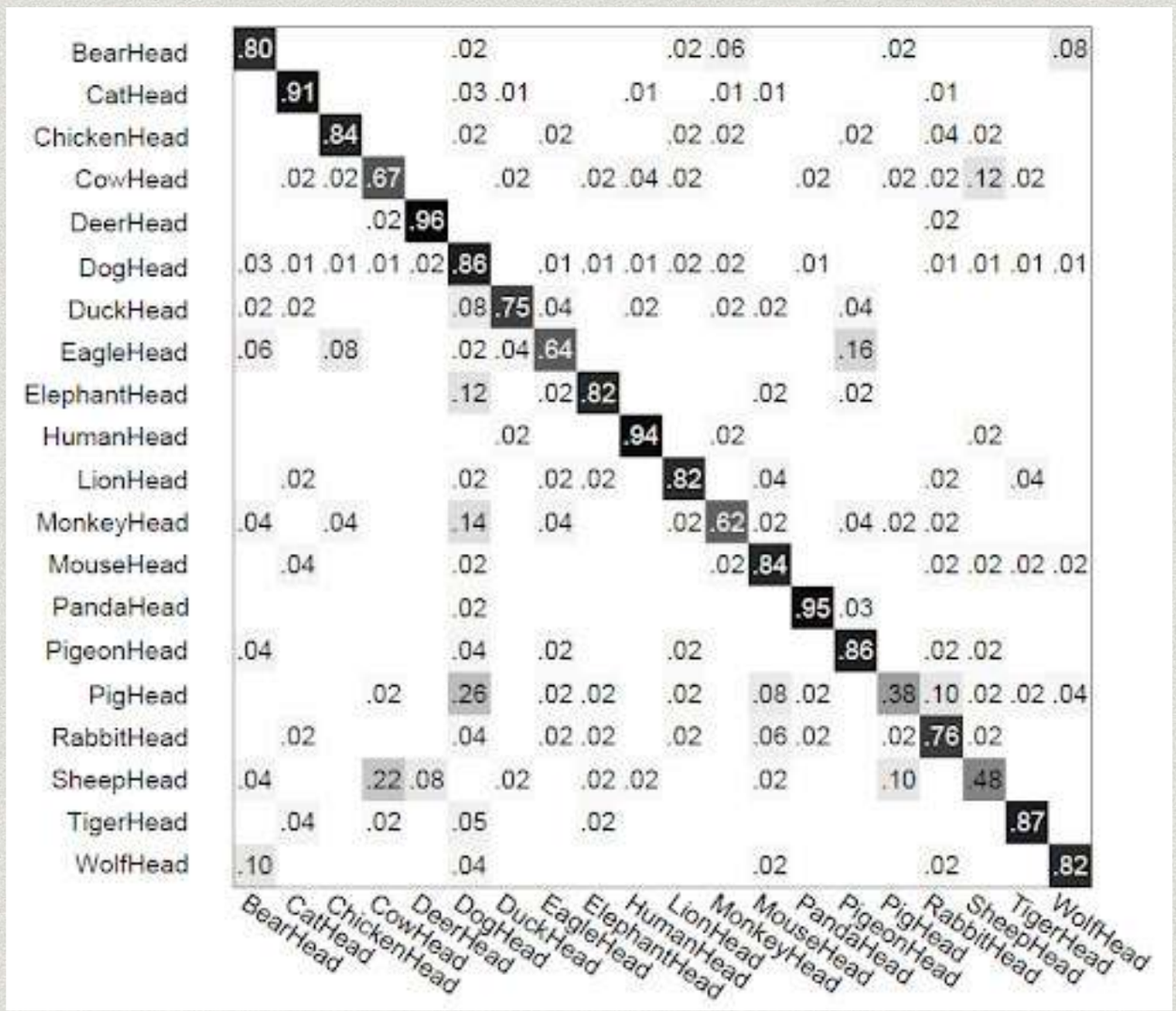
$$\text{recall/completeness} = \frac{\text{predicted true cats}}{\text{all true cats}} = \frac{85}{100}$$

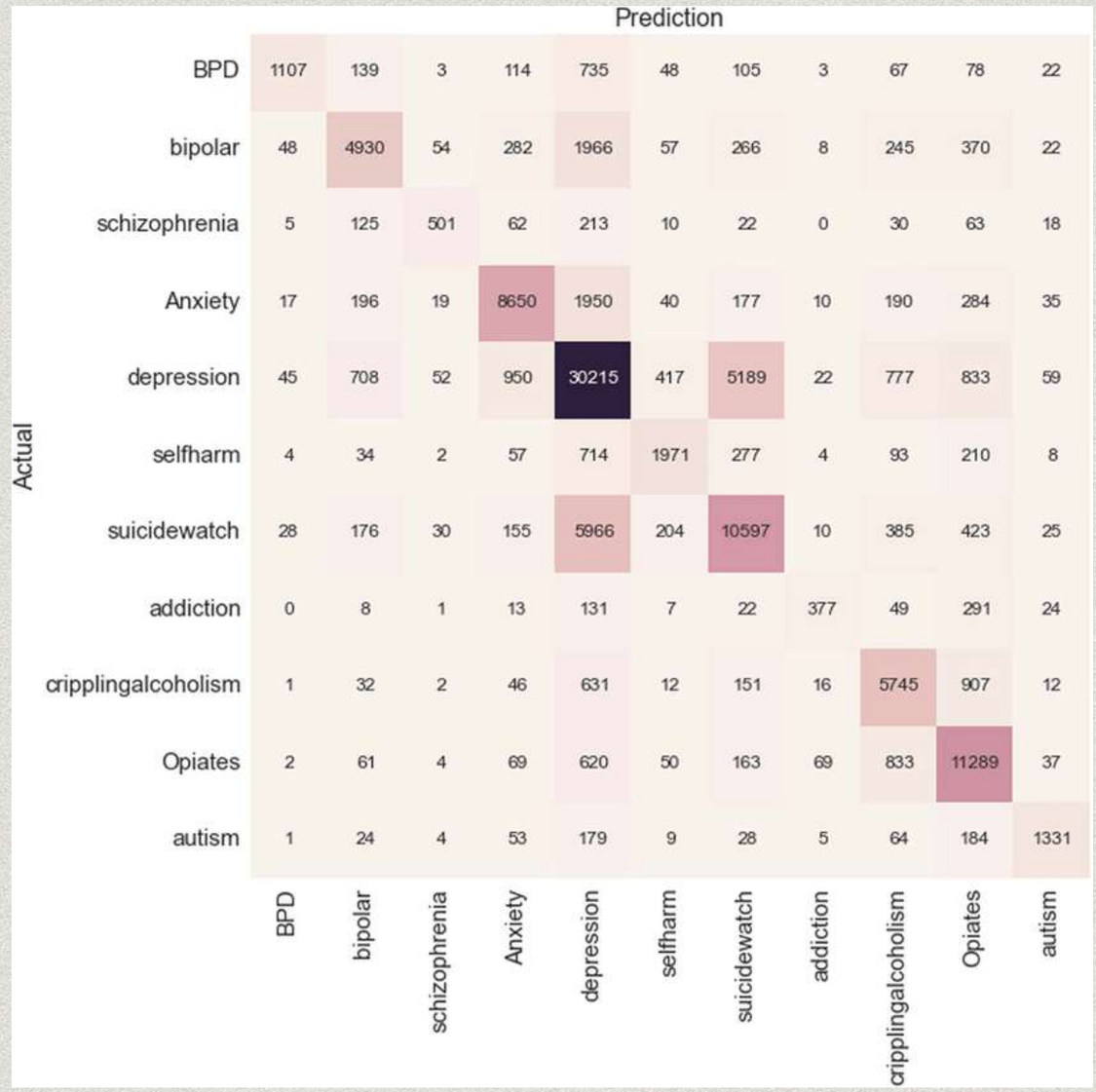
$$\text{F1 score} = 2\frac{\text{purity} \times \text{completeness}}{\text{purity} + \text{completeness}}$$



Certainty and confusion

- We often have **many classes**, with large **confusion matrices**

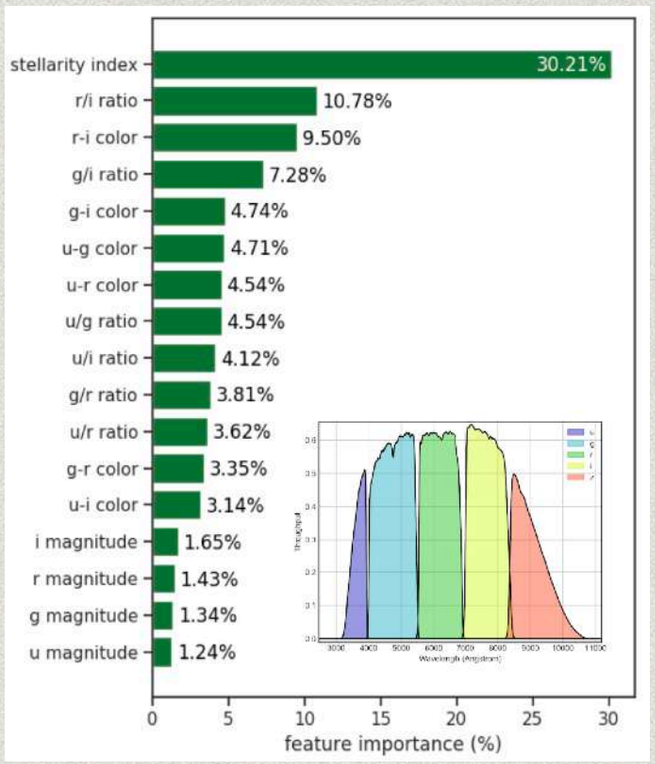


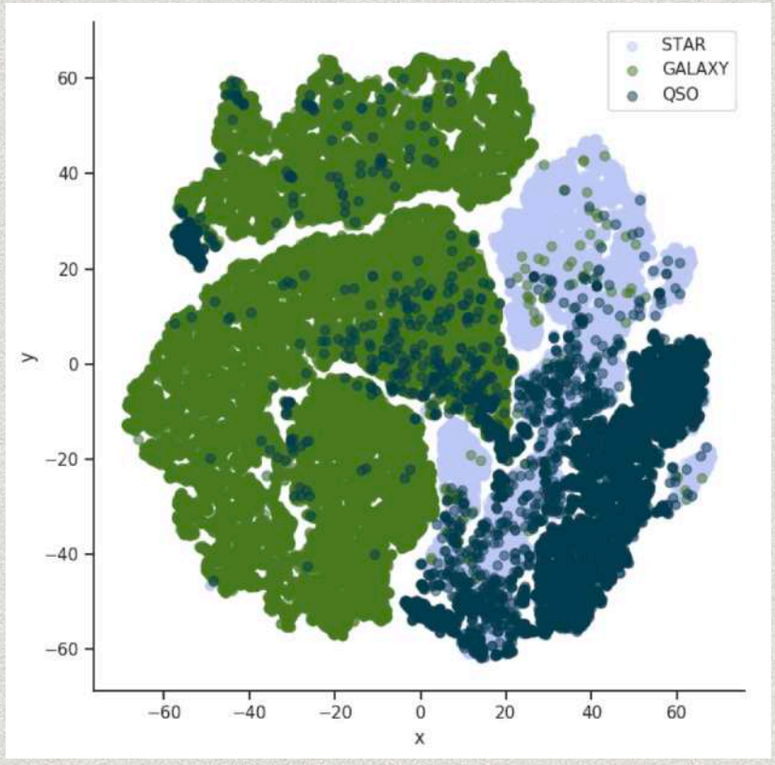


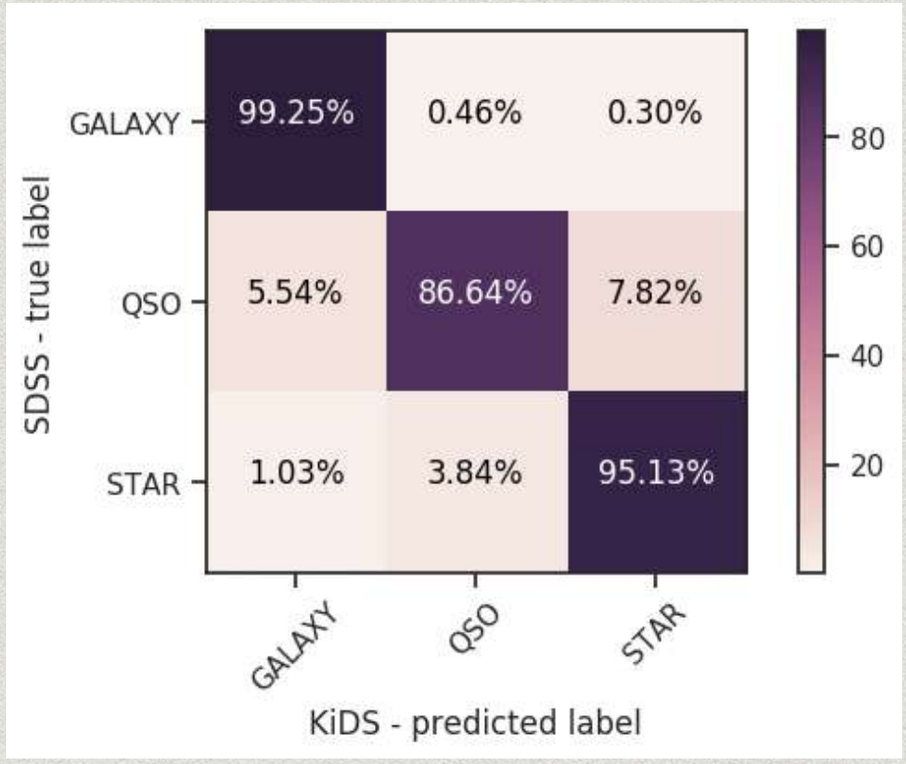
Certainty and confusion

- The idea of ML classifiers is being applied to the star/galaxy/quasar separation in astrophysical surveys

KIDS survey, $r < 22$ → 3.4 Million objects



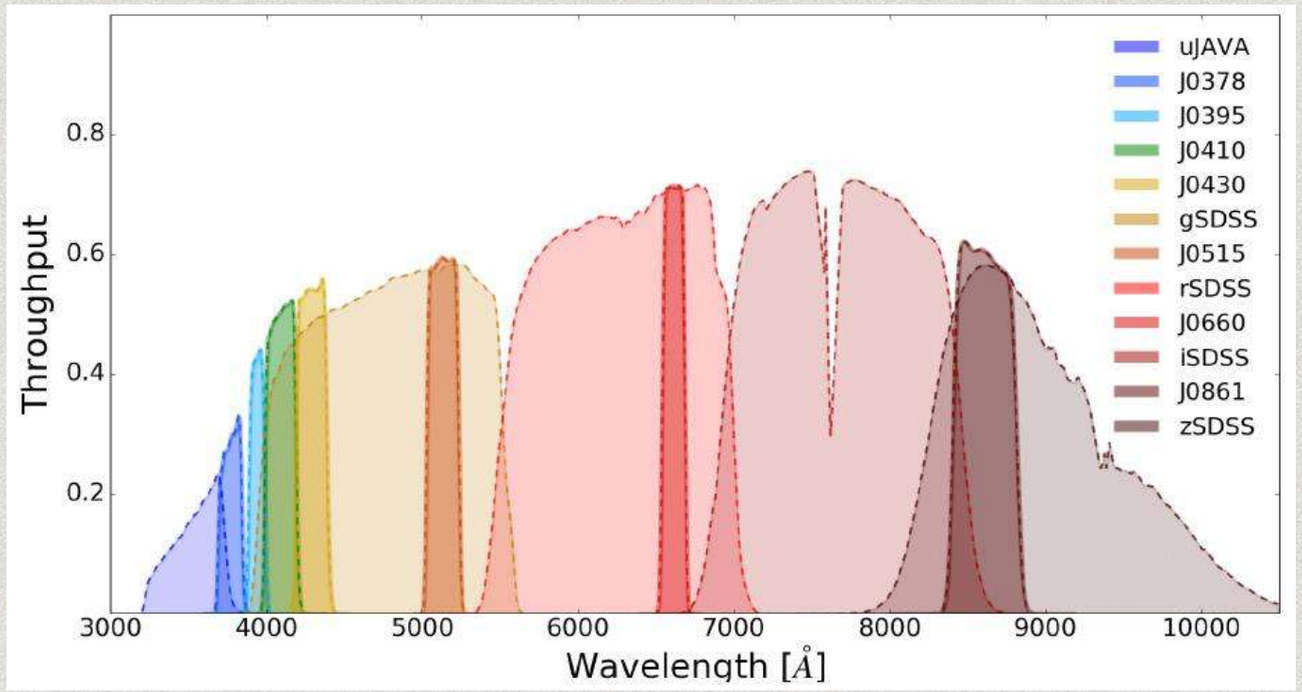


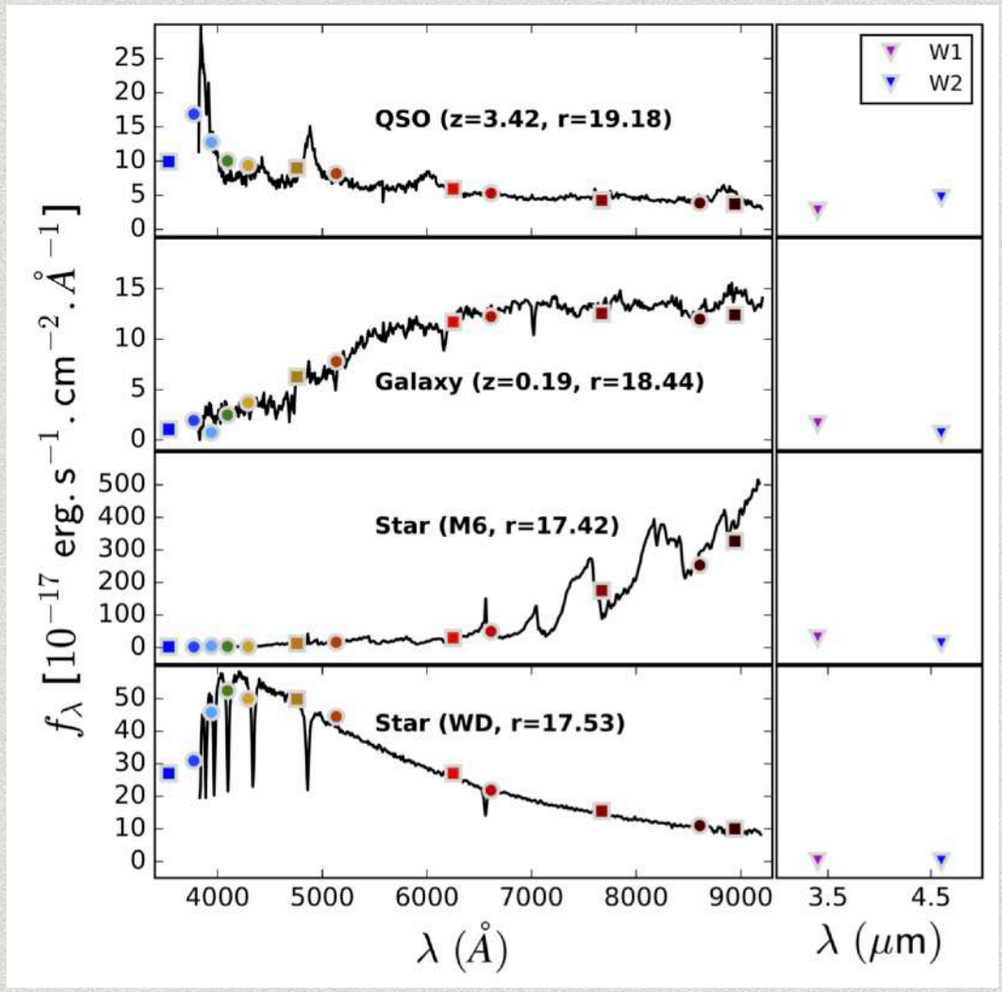


Nakoneczny et al. 2019 (KIDS)

ML and pseudo-spectra

- With more filters, and narrower bands, it becomes easier to classify objects — e.g., J-PLUS & S-PLUS

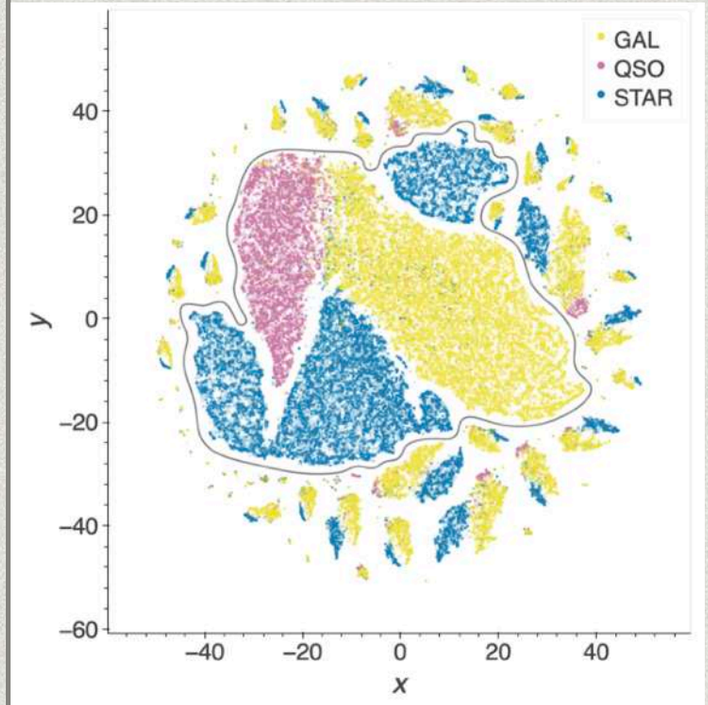


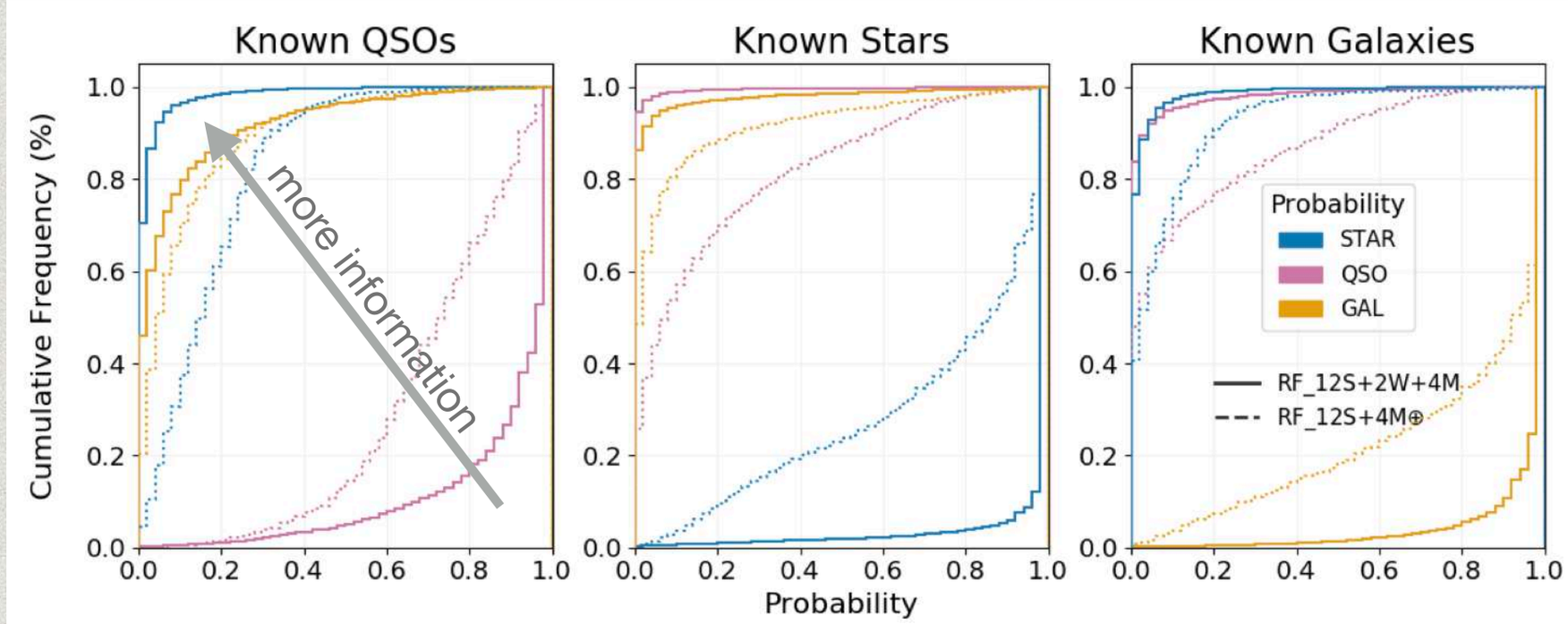


Cenarro et al. 2017 (J-PLUS); Mendes de Oliveira et al. 2019 (S-PLUS)

ML and pseudo-spectra

- With more filters, and narrower bands, it becomes easier to classify objects

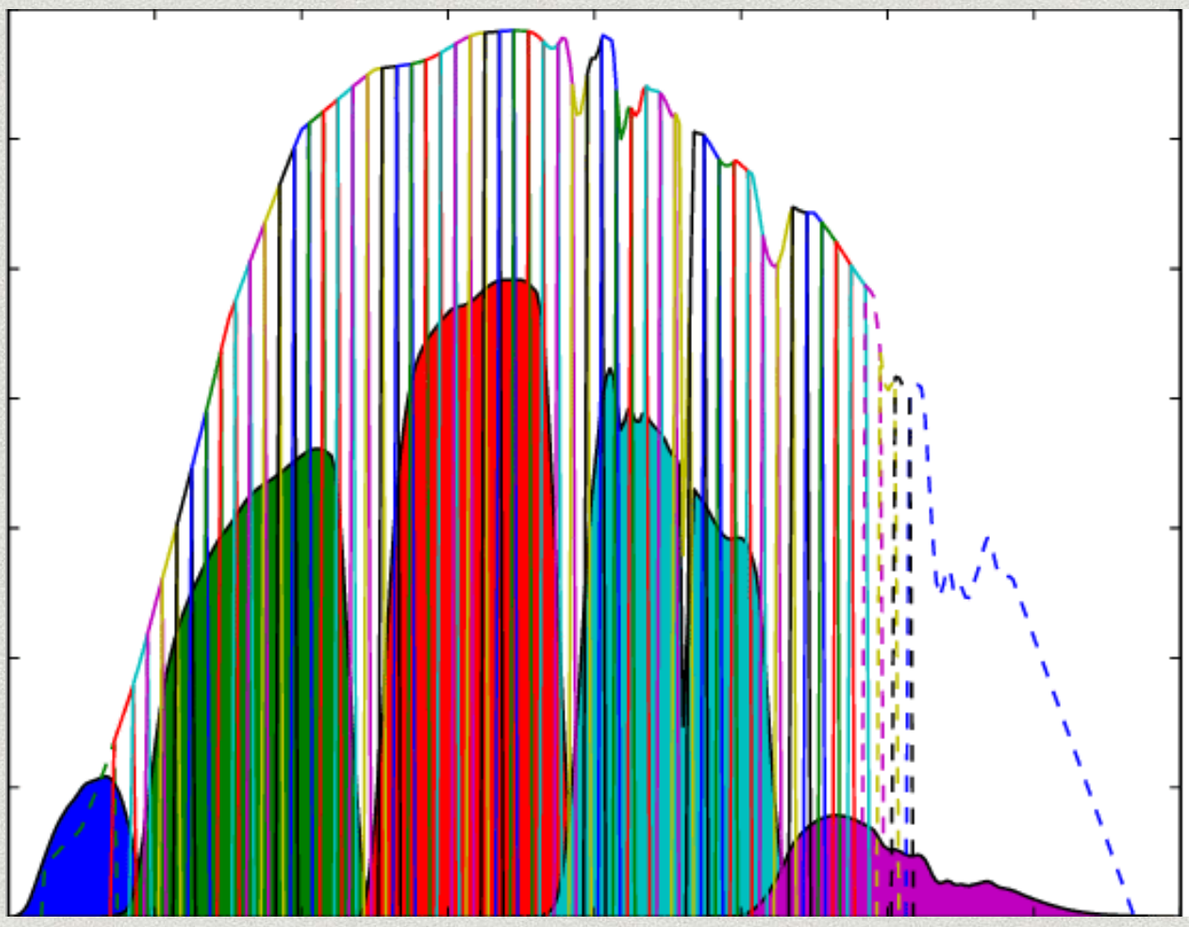


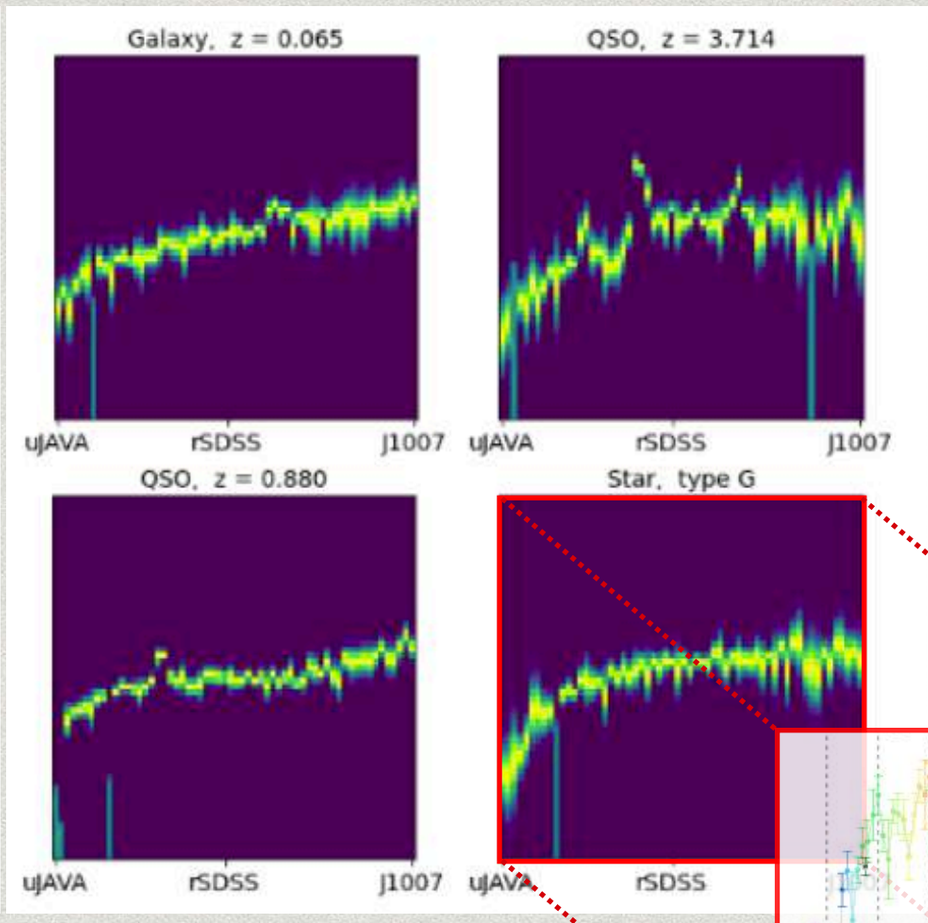


Nakazono et al. 2021 (S-PLUS)

ML and pseudo-spectra

- J-PAS: ultimate survey with “pseudo-spectra”
- Data with errors is like an image!

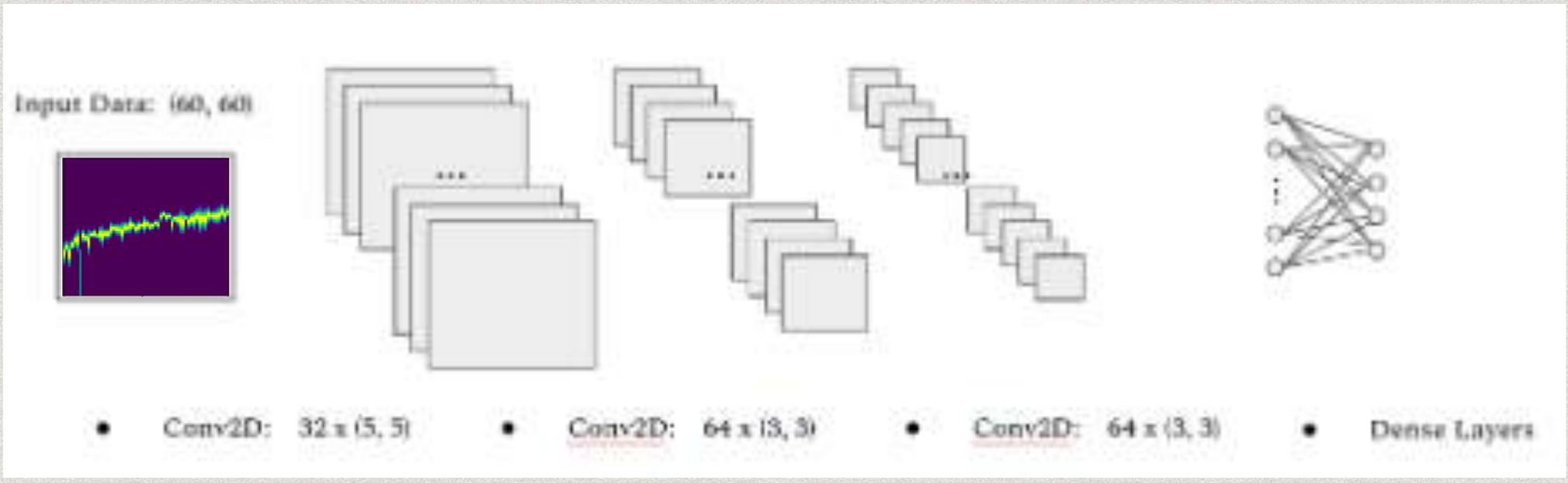




Rodrigues, Abramo et al. 2021 (J-PAS; to appear)

ML and pseudo-spectra

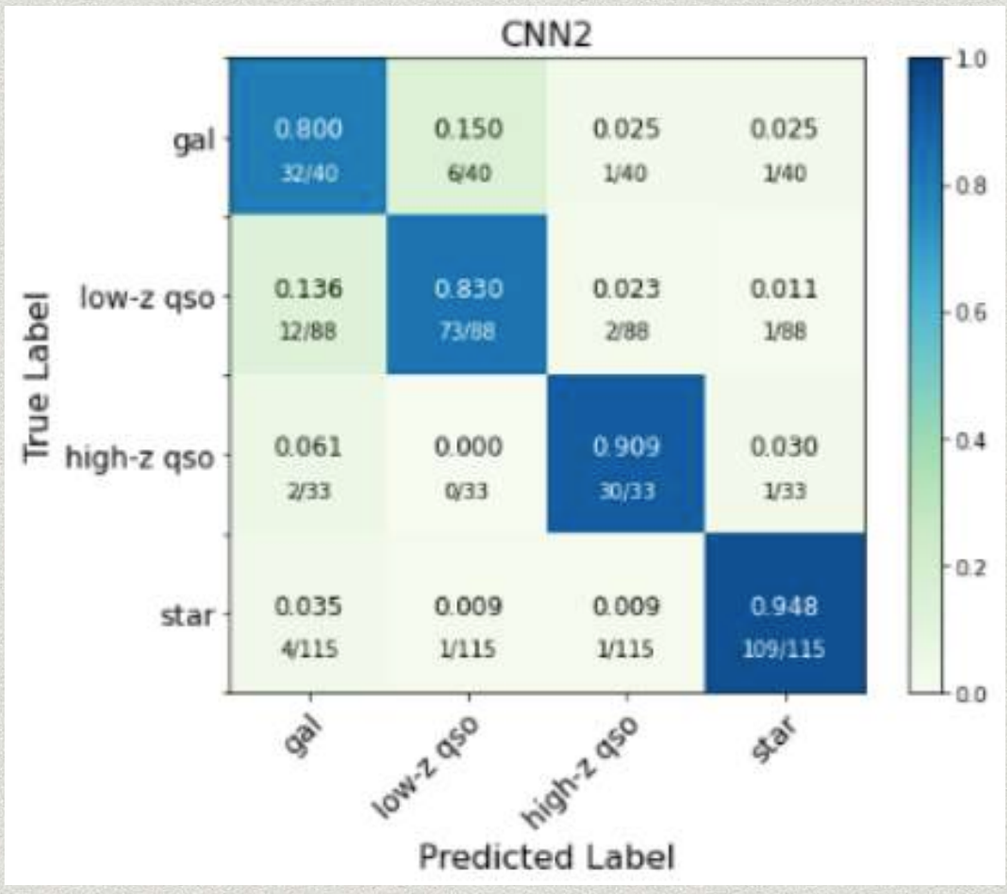
- J-PAS: ultimate survey with “pseudo-spectra”
- Data with errors is like an image!
- Classify stars, galaxies, low-z and high-z quasars



Rodrigues, Abramo et al. 2021 (J-PAS; to appear)

ML and pseudo-spectra

- J-PAS: ultimate survey with “pseudo-spectra”
- Data with errors is like an image!
- Classify stars, galaxies, low-z and high-z quasars

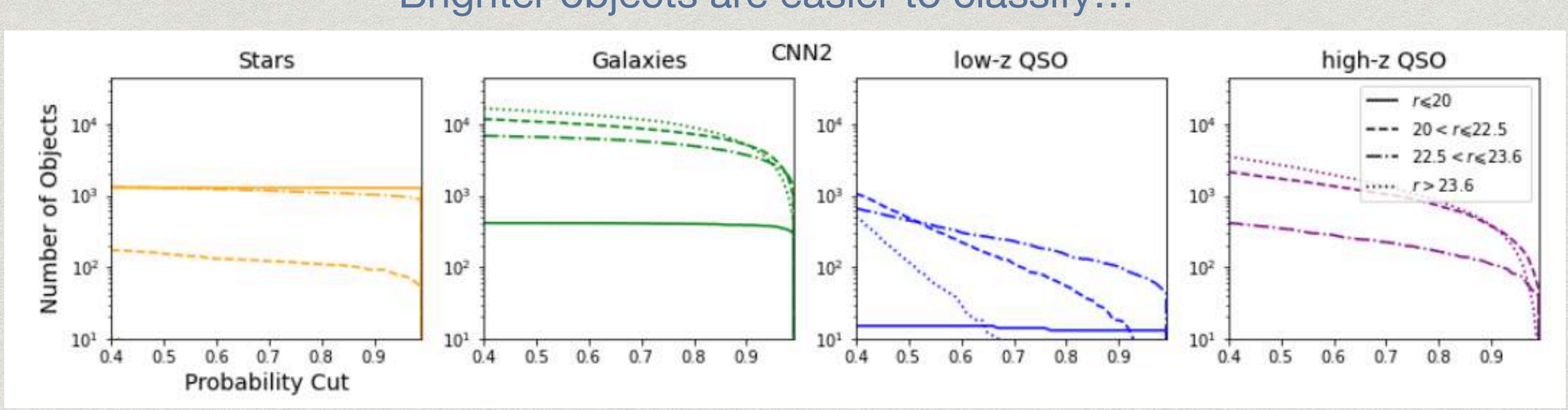


Rodrigues, Abramo et al. 2021 (J-PAS; to appear)

ML and pseudo-spectra

- J-PAS: ultimate survey with “pseudo-spectra”
- Data with errors is like an image!
- Classify stars, galaxies, low-z and high-z quasars

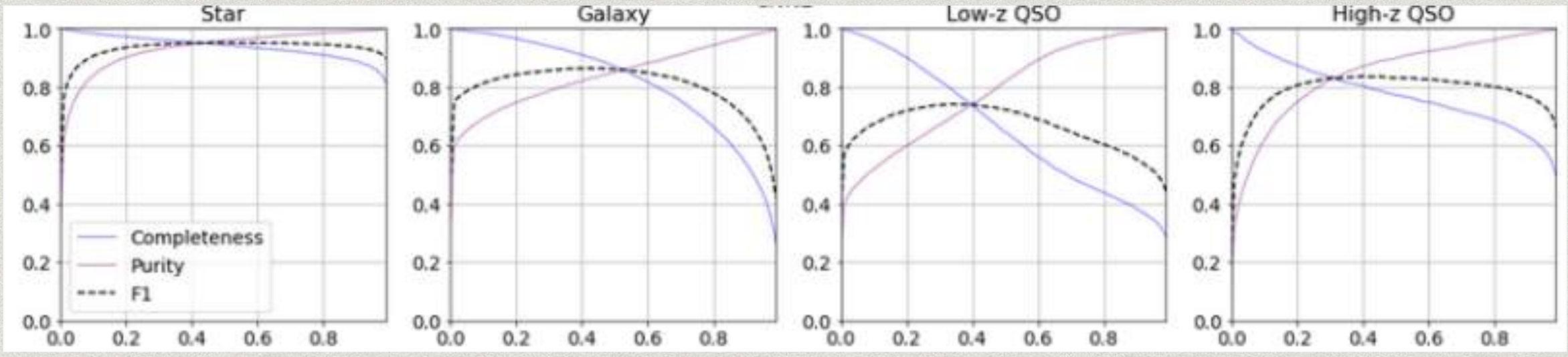
Brighter objects are easier to classify…



Rodrigues, Abramo et al. 2021 (J-PAS; to appear)

ML and data science

- But, wait a minute here… ML methods are like a “black box”, their outputs are not under any kind of control… how do we know that the "**confidence**" that they give us corresponds to any kind of **probability**, in the same sense that, e.g., of a χ^2 ?
- There is **no such thing as a “ML probability”** — there is only a “confidence”, a “score”, a … what ??…

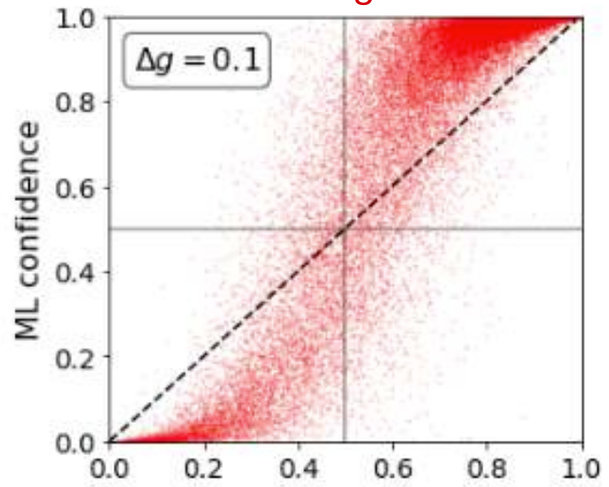
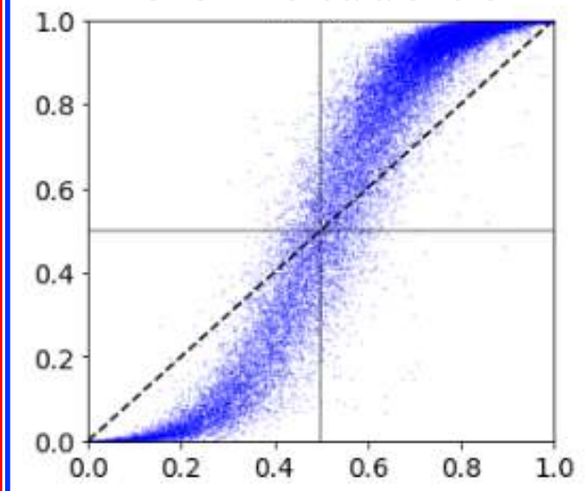
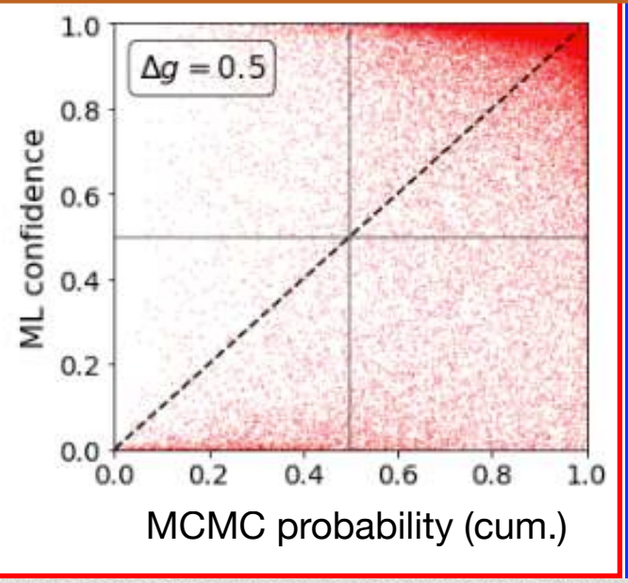
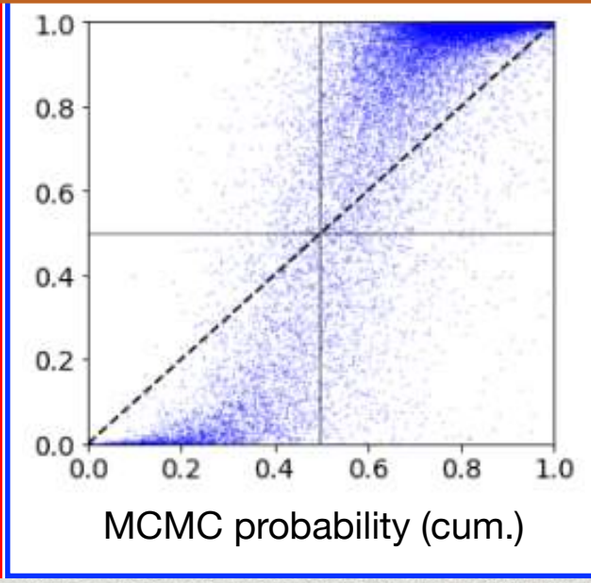


ML confidence

Rodrigues, Abramo et al. 2021 (J-PAS; to appear)

ML and data science

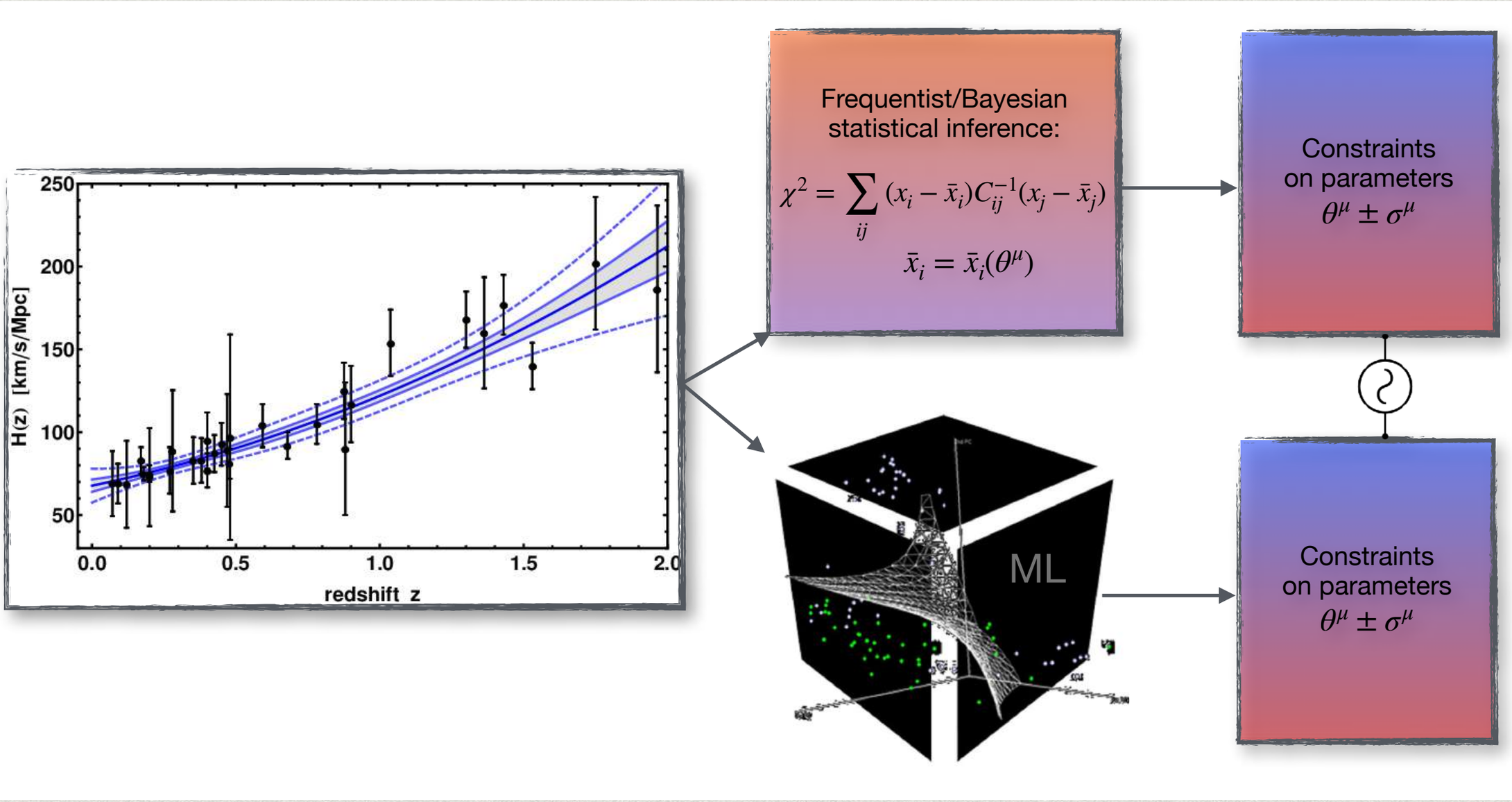
- We recently showed that the **only way** to establish a correspondence between the ML confidence and a frequentist/Bayesian probability, is by including the data errors in the ML method

	NOT including data errors	INCLUDING data errors
All data errors with nearly the same SNR		
Data errors with varying SNR		

Rodrigues, Abramo & Hirata, 2021

ML and data science

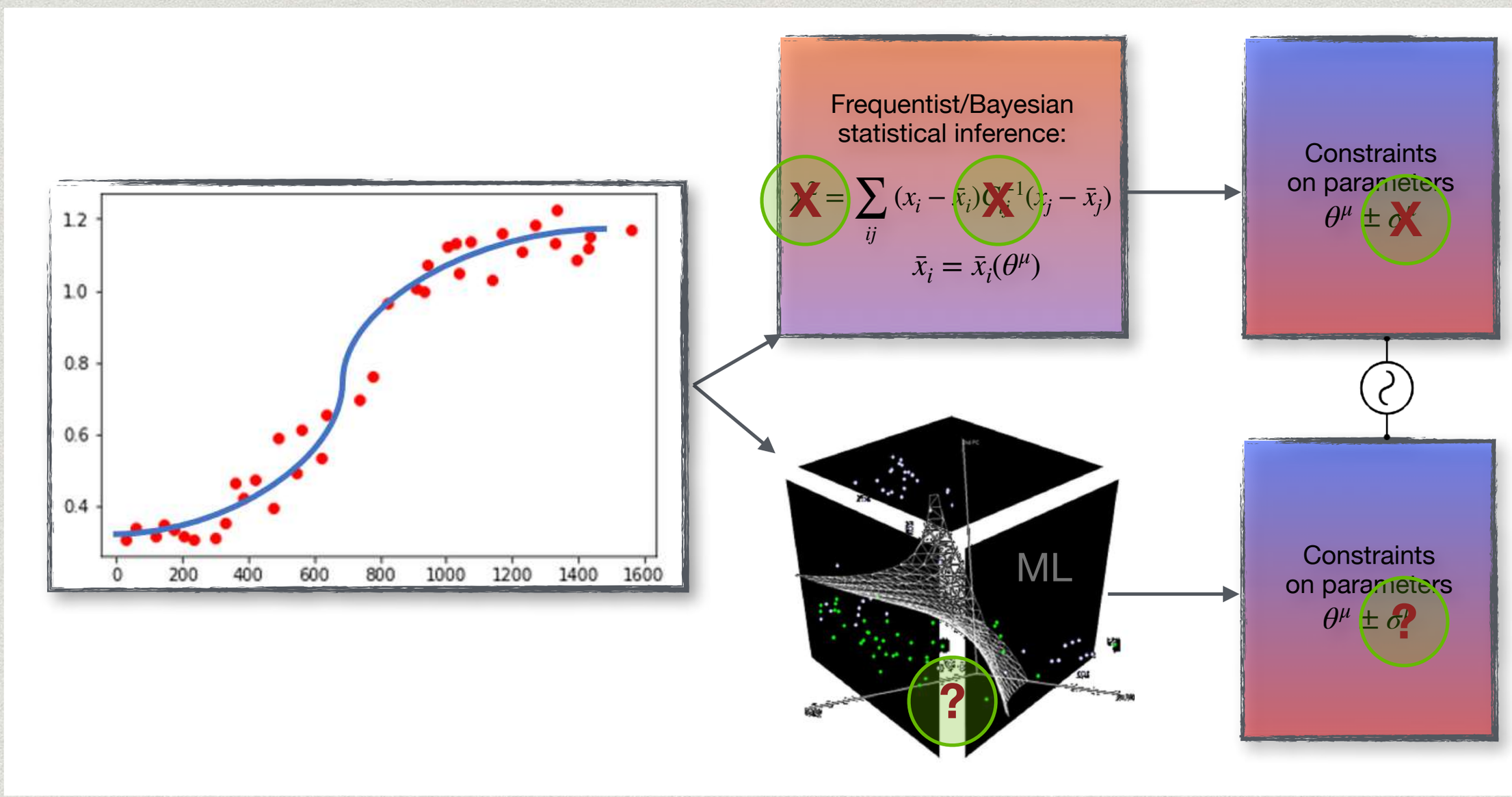
- Therefore, by providing the ML methods the information about ***data errors*** we are able to start bridging the gap between the ML "black box" and ***statistical inference***.



Rodrigues, Abramo & Hirata, 2021

ML and data science

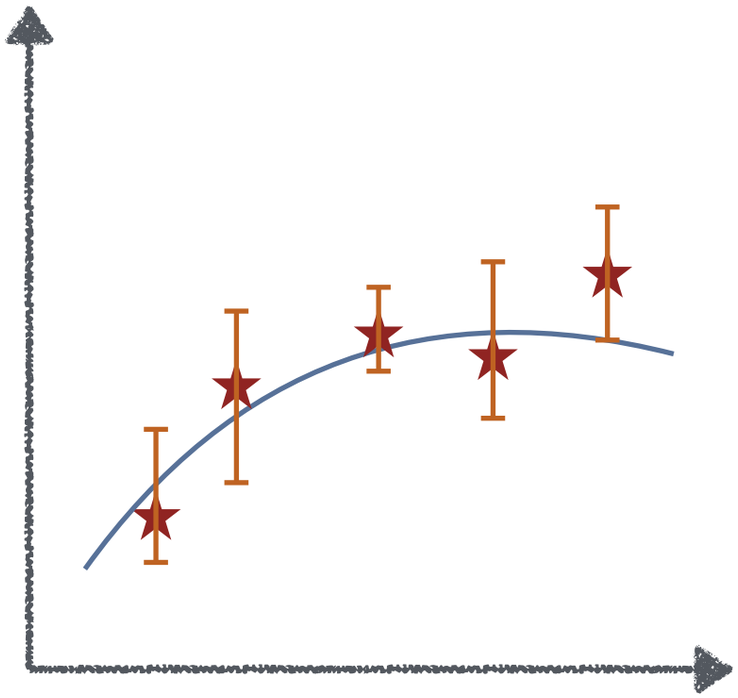
- But **why** is this so important? If we want to use ML in our “inference pipeline”, we must be able to assess the **accuracy of our predictions** — the **power of our constraints**. And without a “probability" assigned to the ML output , it is impossible to determine that accuracy!



Rodrigues, Abramo & Hirata, 2021

ML and data science

- But **why** is this so important? If we want to use ML in our “inference pipeline”, we must be able to assess the **accuracy of our predictions** — the **power of our constraints**. And without a “probability" assigned to the ML output , it is impossible to determine that accuracy!



- Ultimately, all parameter constraints must flow from the **measurement errors**.
- Other factors (e.g., ***priors***) may also play a role — but at the basis of the statistical inference are the errors in input data.
- So, we have come full circle: in the end, Frequentist, Bayesian and ML tools have to rely on the same information, and they are able to give us similar outputs — constraint on cosmological parameters.

Rodrigues, Abramo & Hirata, 2021