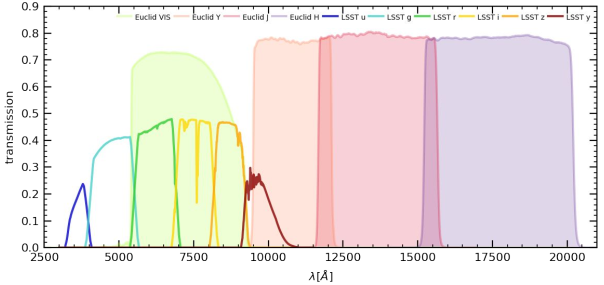
Physical properties of galaxies in Euclid and out-of-the-box objects

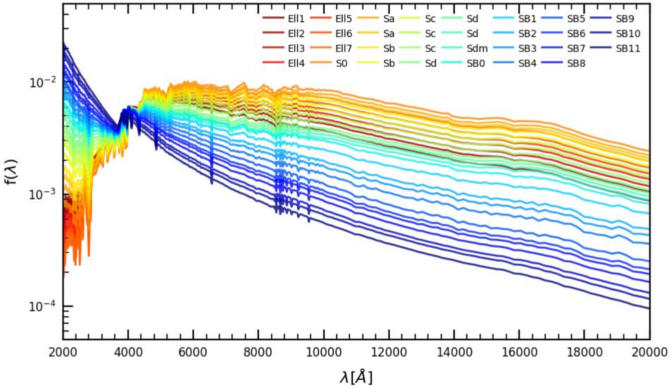
Micol Bolzonella (INAF-OAS Bologna) micol.bolzonella@inaf.it
Stefano Cavuoti (INAF-OACN Napoli)
Lars Doorenbos (University of Bern)
Lucia Pozzetti (INAF-OAS Bologna)
Andrea Enia (University of Bologna)
Laura Bisigello (University of Padova) and others from OU-PHZ and GAEV-SWG

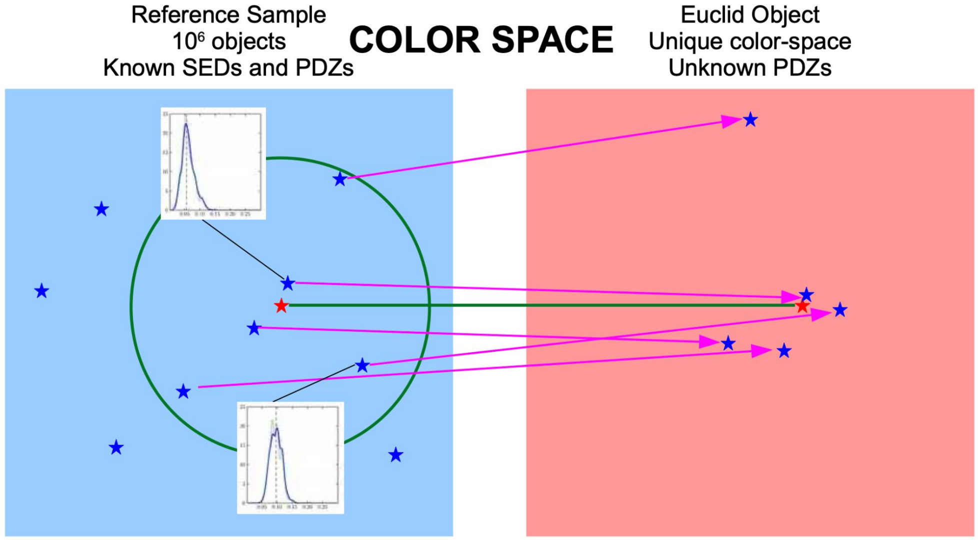
Photometric redshifts

In Euclid photometric redshifts will be computed with a 2 steps procedure:

1. **calibration**: creation of a reference sample with template fitting (**Phosphoros**) using deep observations
2. **production**: deriving photometric redshifts for the Wide sample using a nearest neighbours approach (**NNPZ**) in the colours space





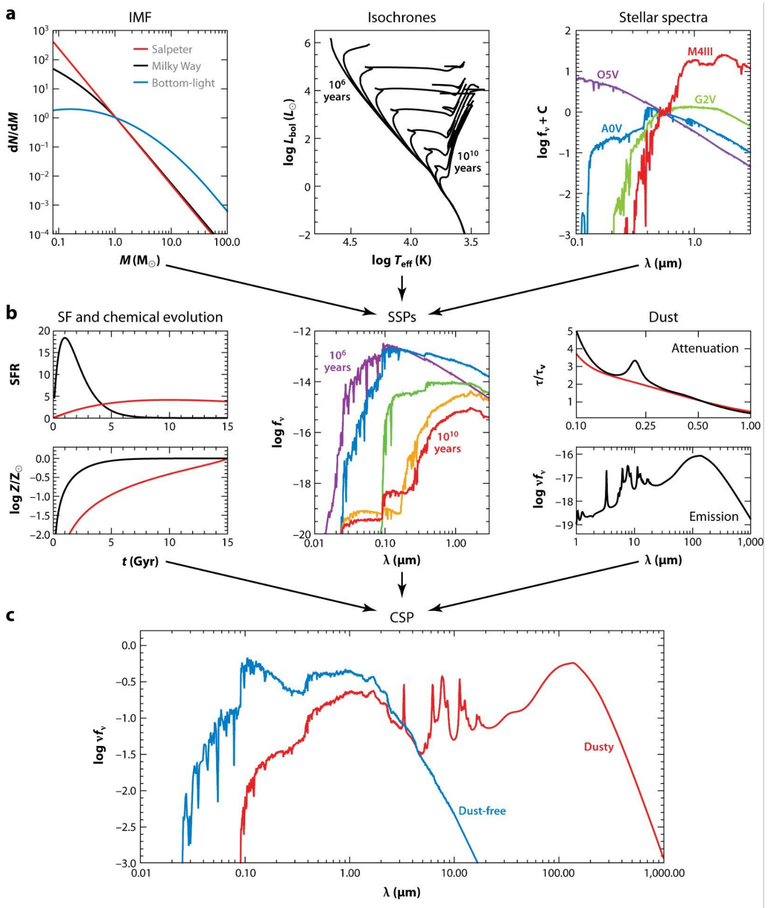


PPs and z_{phot} are not alike

Main physical properties:

- redshift
- stellar mass
- Star Formation Rate
- age
- A_V
- attenuation law
- stellar metallicity
- absolute magnitudes

Physical properties are derived quantities that strongly depend on models of the electromagnetic radiation



Conroy C. 2013.
Annu. Rev. Astron. Astrophys. 51:393–455

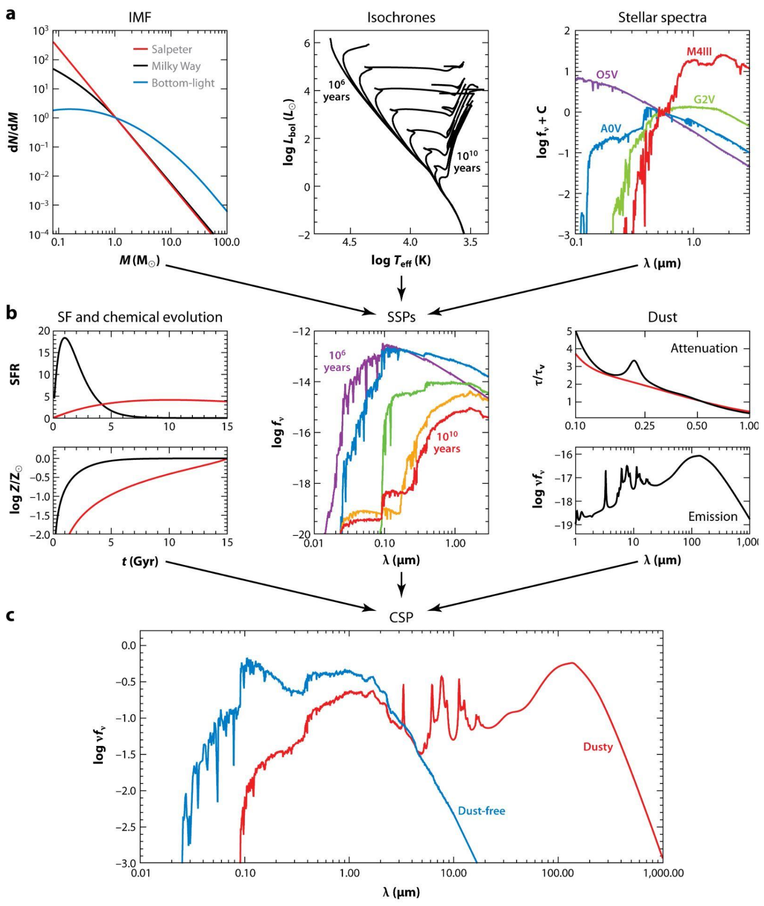
PPs and z_{phot} are not alike

Main physical properties:

- redshift
- stellar mass
- Star Formation Rate
- age
- A_V
- attenuation law
- stellar metallicity
- absolute magnitudes

Physical properties are derived quantities that strongly depend on models of the electromagnetic radiation
→ methods to derive PPs and photometric redshifts are similar, with significant differences:

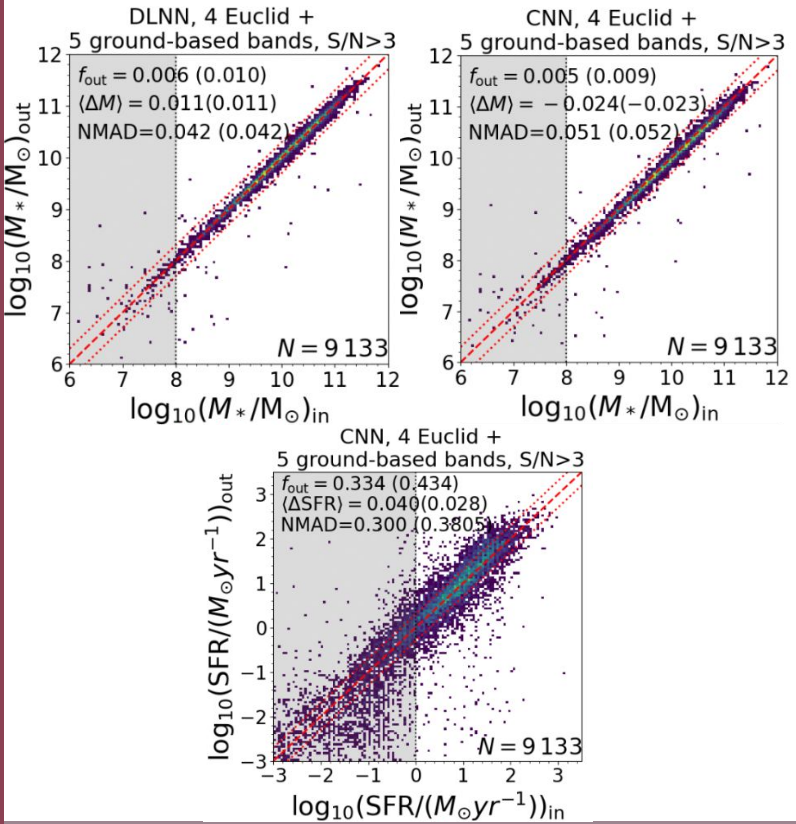
1. no ground truth → validation on simulations
2. results depend on assumed stellar population synthesis / dust models at optical/NIR λ
3. SED fitting can be considerably slower



Conroy C. 2013.
Annu. Rev. Astron. Astrophys. 51:393–455

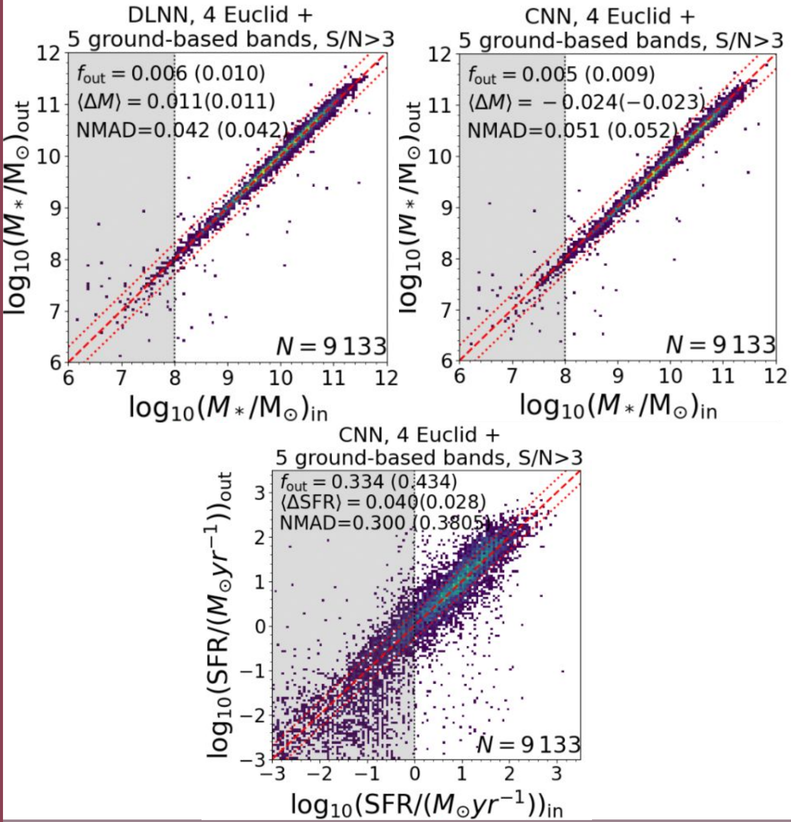
PPs with Machine Learning

Laura Bisigello+ arXiv:2209.13074: Deep Learning and Convolutional Neural Network using H-band images to recover z, M*, and SFR

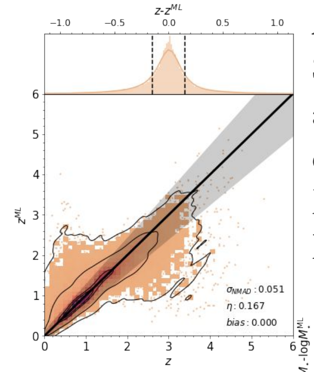


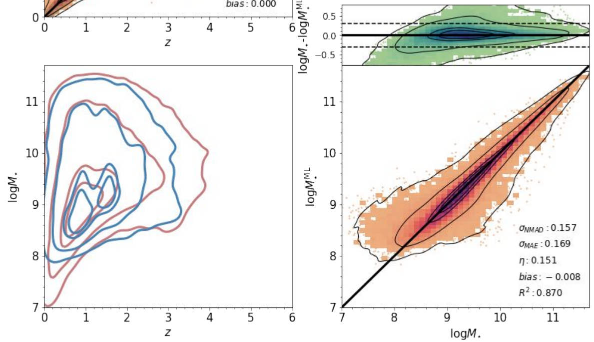
PPs with Machine Learning

Laura Bisigello+ arXiv:2209.13074: Deep Learning and Convolutional Neural Network using H-band images to recover z, M*, and SFR

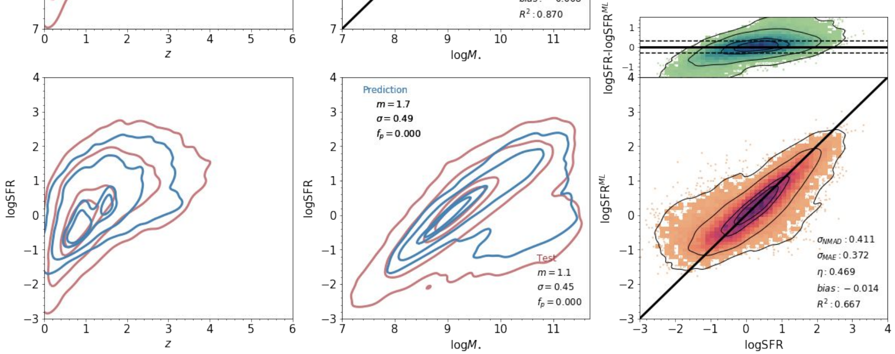


Andrea Enia (ongoing): z, M* and SFR with random forest training on a sample with SED fitting 1 mag deeper than the Wide survey [using MAMBO simulation, see also Lucia Pozzetti's talk this afternoon]





+ ongoing work by **Andrew Humphrey & Pedro Cunha** @Porto

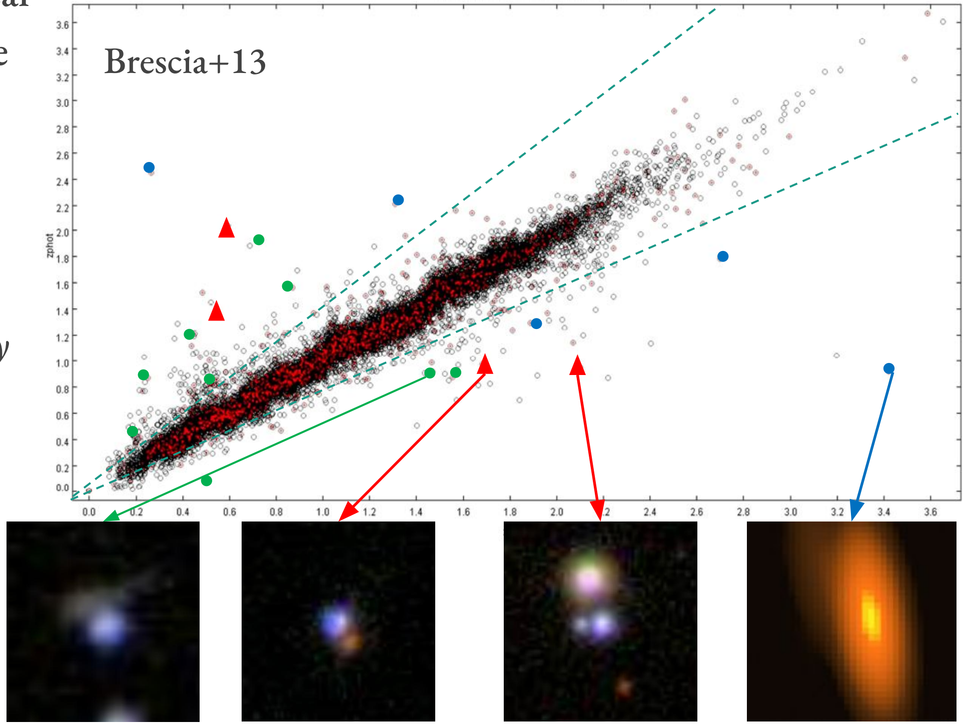


Outliers

Identifying the outliers is important to:

1. identify problems in observations / data reduction / data processing
   e.g. deblending issues when combining space and ground based observations
2. remove objects that can bias the statistical description of any interesting subsample
   fundamental to improve photometric redshifts for cosmological use
3. select weird/rare/peculiar classes of objects
   can open new space of discovery in galaxy evolution

Can we predict beforehand for which objects the redshift prediction will fail? Which properties do they have?



Selecting outliers

Stefano Cavuoti & Lars Doorenbos

Approaches:

- Unsupervised → k-nearest neighbours (kNN)
  - Find unusual objects in the data
  - See if the outliers found match failure cases for downstream models
- Supervised → Random forest regressor
  - Find unusual objects in the data w.r.t. a given redshift/physical parameter predictor
  - Directly learn where a downstream model will fail

Tests on semi-analytic lightcone GAEA [De Lucia, Fontanot, Hirschmann] with Euclid (I_E, Y_E, J_E, H_E) & LSST (u, g, r, i, z, y) bands. Features = magnitudes and colours

We divide outliers into two types:

- Tail outliers
  - At least in 5th percentile (high or low) for at least 1 feature
- Inner outliers
  - "Normal" values for each individual feature

Unsupervised outliers

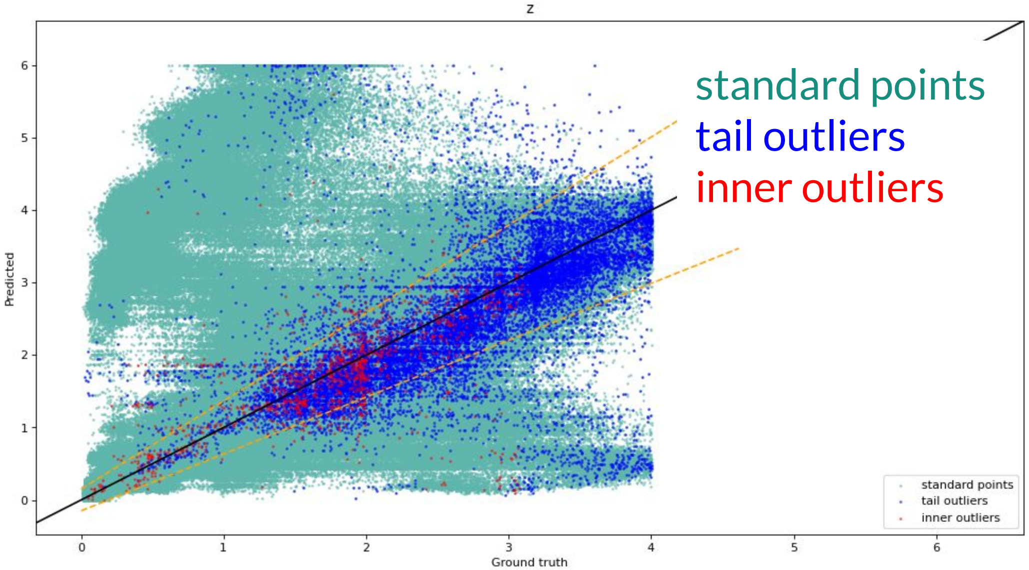
Flag objects as outliers when their kNN (k=20) distance > 5σ:

→ Much higher percentage of failure cases (~28%) in outliers than in sample (~10%)

→ Not many objects flagged: 17 000 out to ~4 million objects

→ lowering the threshold of what constitutes an outlier, more outliers are found, but the sample is less pure

→ Outlier sample not large and/or pure enough to influence global statistics when removed



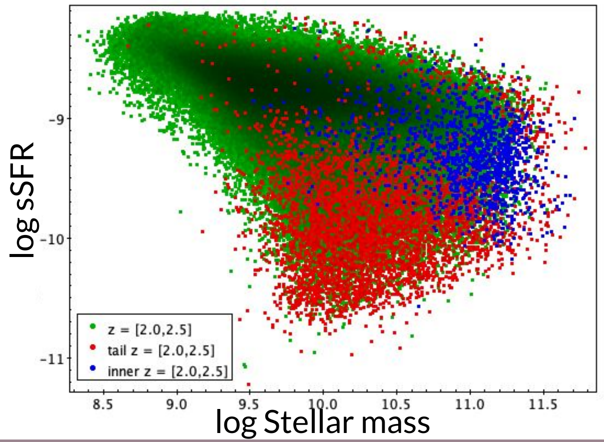
standard points
tail outliers
inner outliers

Physical Properties of outliers

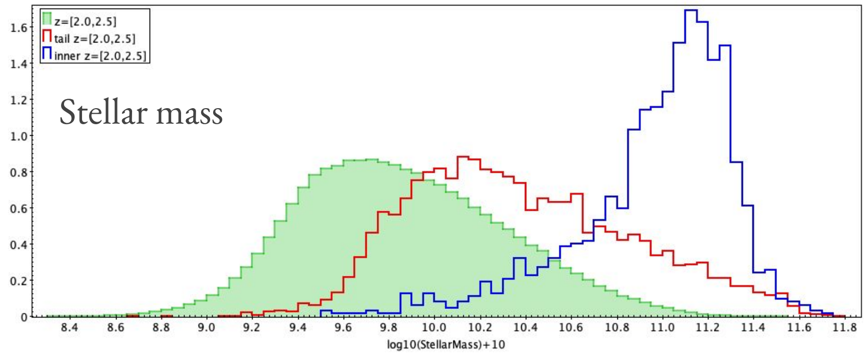
Outliers are typically

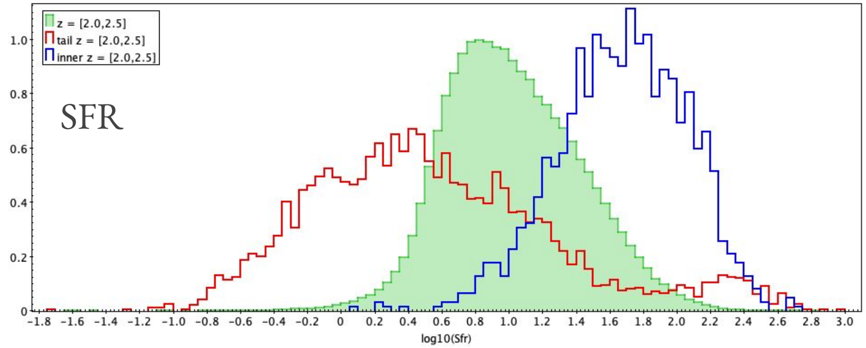
- more massive
- less star forming for tail outliers
- more star forming for inner outliers

→ selection of outliers to filter possible rare and interesting galaxies?



2.0<z<2.5



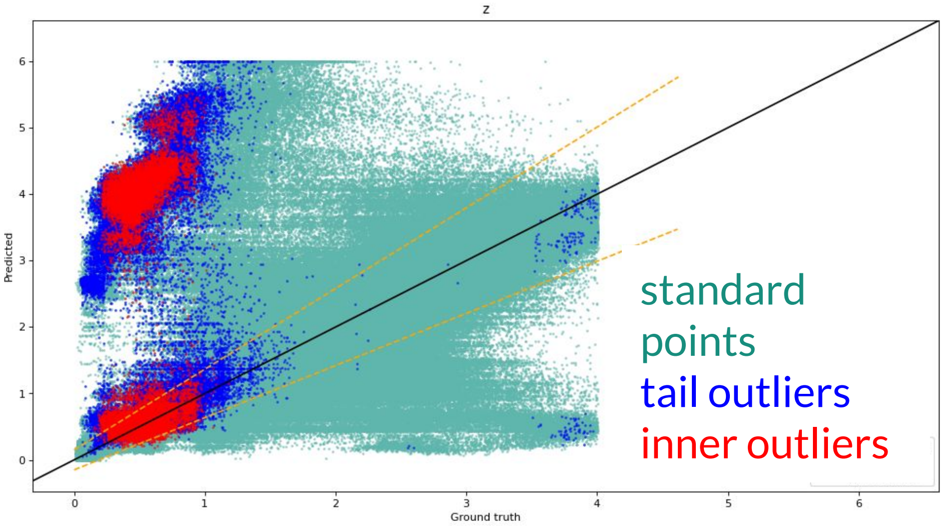


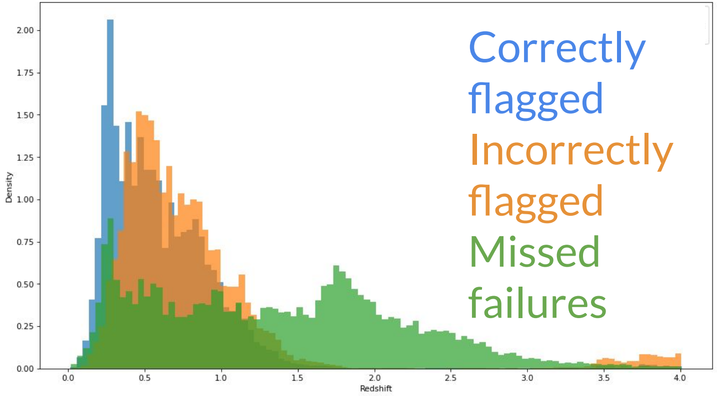
standard points
tail outliers
inner outliers

Supervised outliers

Random forest regressor → objects flagged as outliers when the predicted residual > 2

- ~65 000/4E6 objects flagged as outliers ~70% of them truly photoz outliers
- The RF flags outliers at low redshift and a few at the high end
- The part from 1.5 to 3.5 none of the failures are found
- Lowering the outlier threshold definition, more outliers are found, but the sample is less pure (with a threshold = 1 → 250 000/4E6 objects selected, 60% of them outliers)
- Removing them the overall statistics of photoz improves (amount depending on threshold)





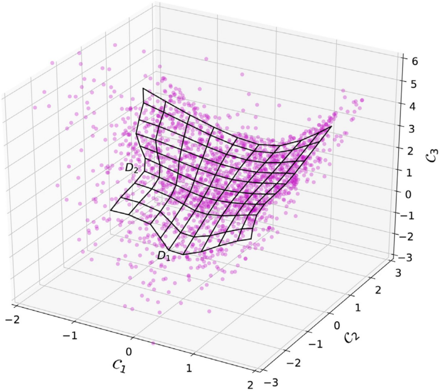
Self Organising Maps

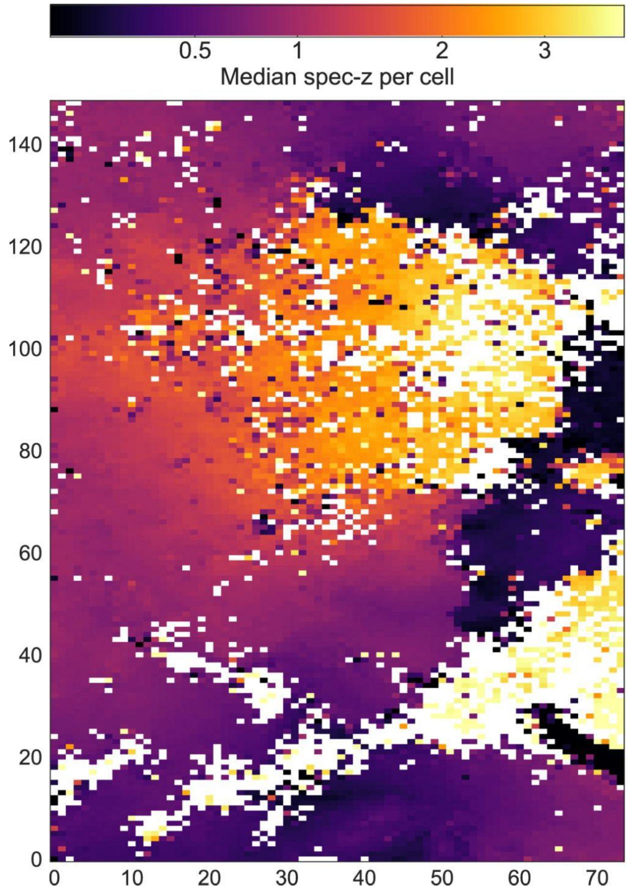
Unsupervised ML, training phase to create a compressed, discretised, lower dimensional representation of the training set.

Each neuron in the map is initialised randomly, than for each input data the best matching unit is found in the map, minimising the distance between the properties in the observed object and the ones held by the neuron.

At the end, similar data points will be mapped close to each other.

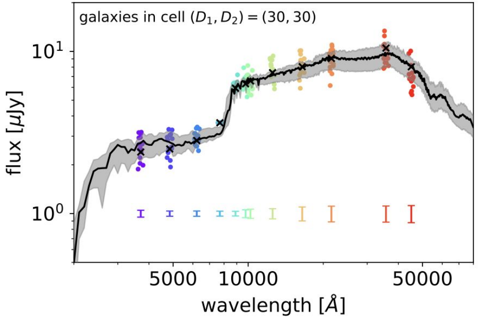
See Davidzon+19
Masters+15
Stanford+21



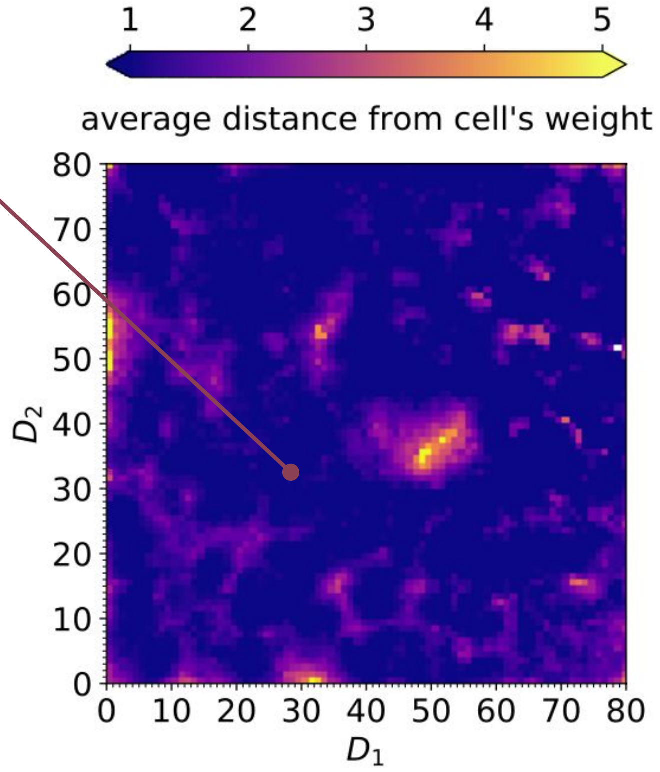


Self Organising Maps

Normal cells are populated by similar objects with a small scatter of SEDs



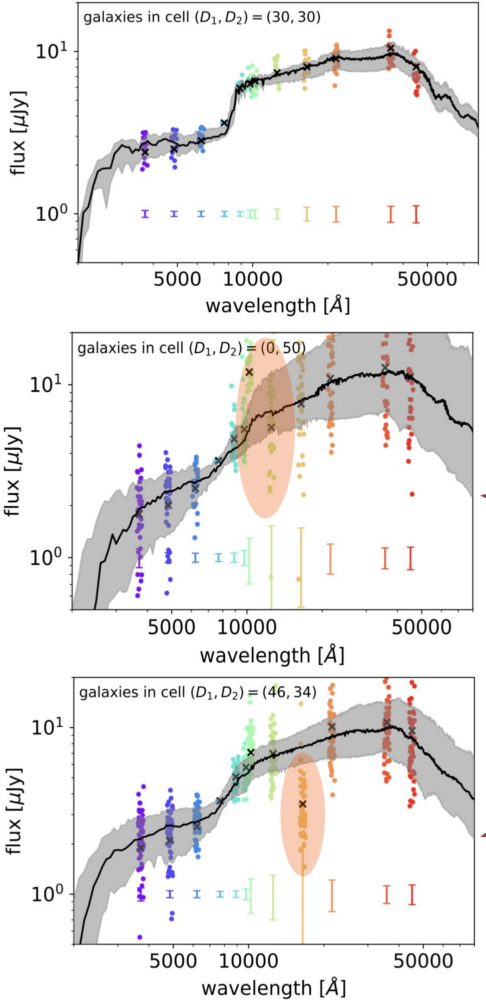
Davidzon+22



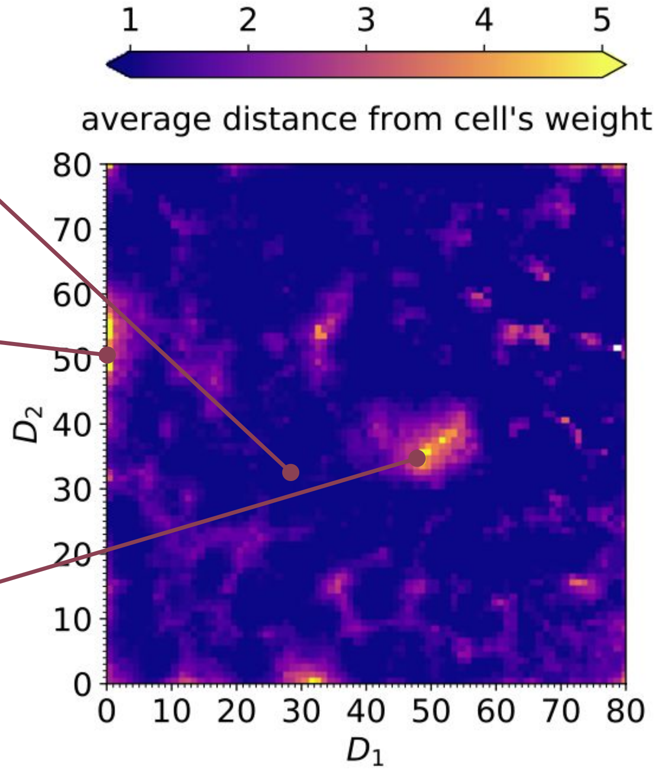
Self Organising Maps

Normal cells are populated by similar objects with a small scatter of SEDs

large intra-cell distance → diagnostic for peculiar SEDs.

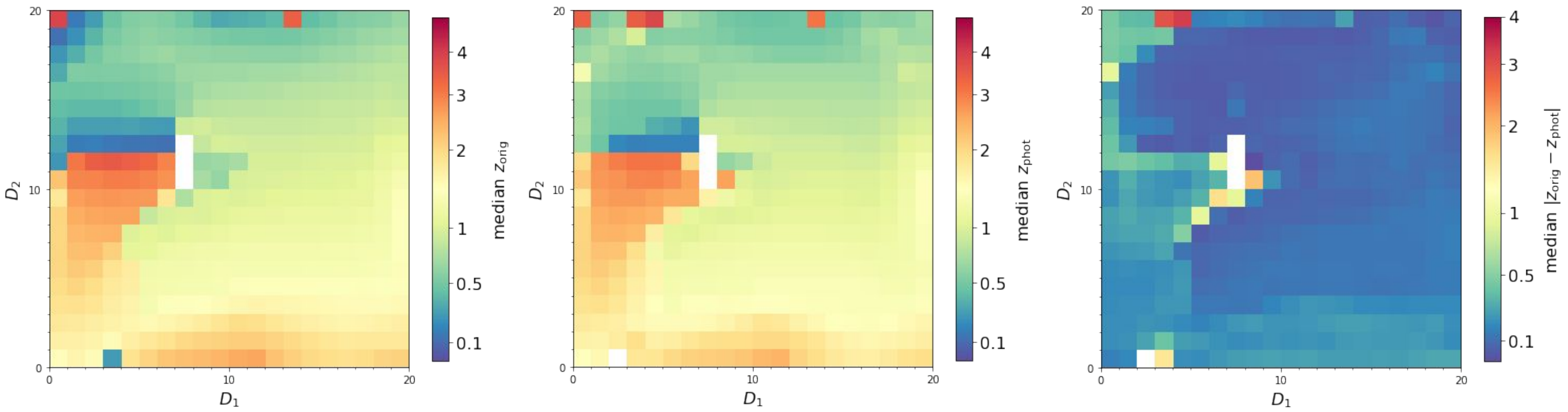


Davidzon+22



GAEA SOM

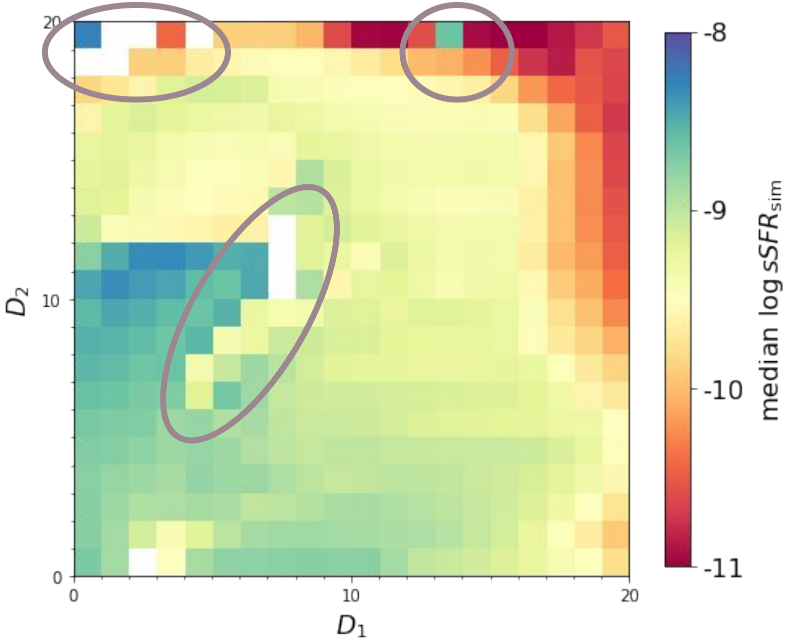
Basic 20x20 map, training on Euclid + LSST intrinsic colours → discontinuities and caustics in the parameter space divide galaxies with similar colours but different redshifts.



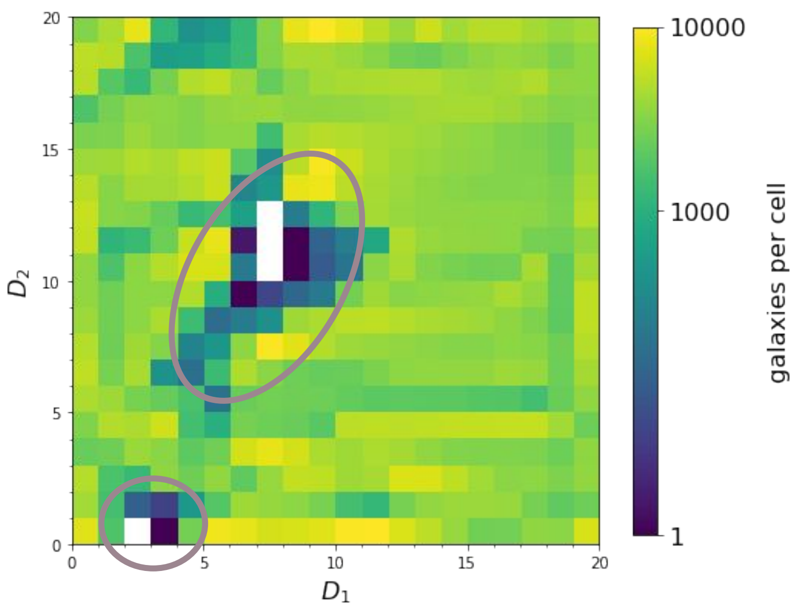
Self Organising Maps

Potentially interesting objects:

Cells with neighbours with extremely different labels, discontinuities



Poorly populated cells



Physical Properties & Outliers

Machine Learning methods are now the new standard:

- accelerate the computation
- use complex relations between input features
- for Physical Properties they should be tested carefully (and multi-dimensional PDF have to be provided)

ML can also used to flag normal classes of objects and outliers:

- well behaved objects can be used to
  - study the evolution of their properties using clean samples
  - determine stacked SEDs of galaxies → build a base of templates
- out-of-the box objects can be useful to
  - augment their representation in training samples
  - direct them to tailored treatments, that would be cumbersome for large samples
  - analyse them with dedicated follow-up observations

Many galaxies

Wide Survey (15000 deg^2)

I_E (SNR=10) < 25 or H_E (SNR=5) < 24

↓

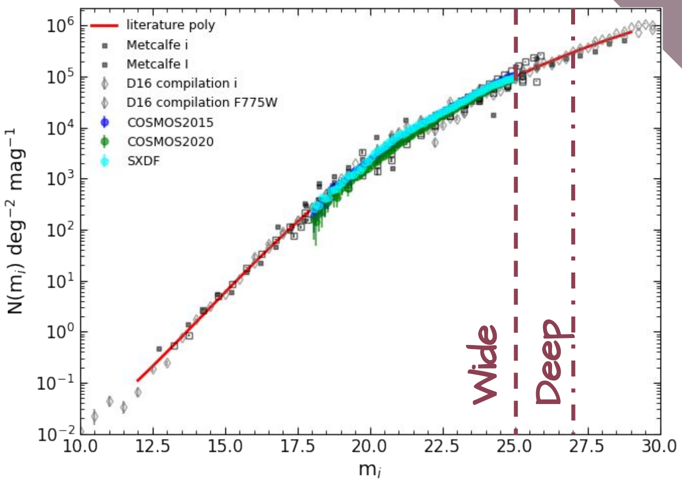
1.5 × 10^9 galaxies

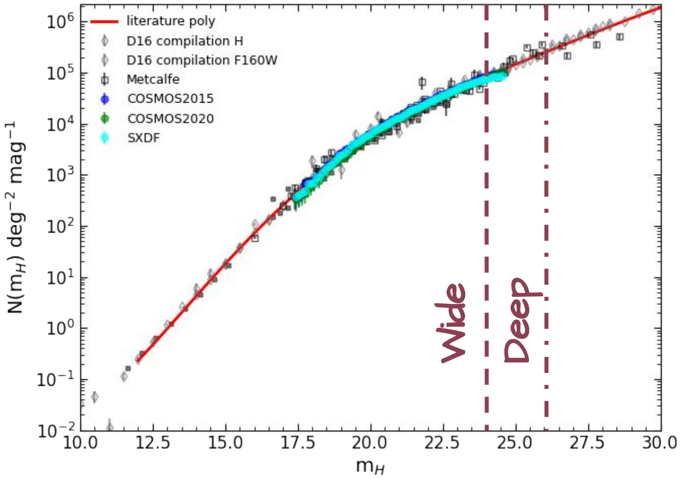
Deep Survey (53 deg^2)

I_E < 27 or H_E < 26

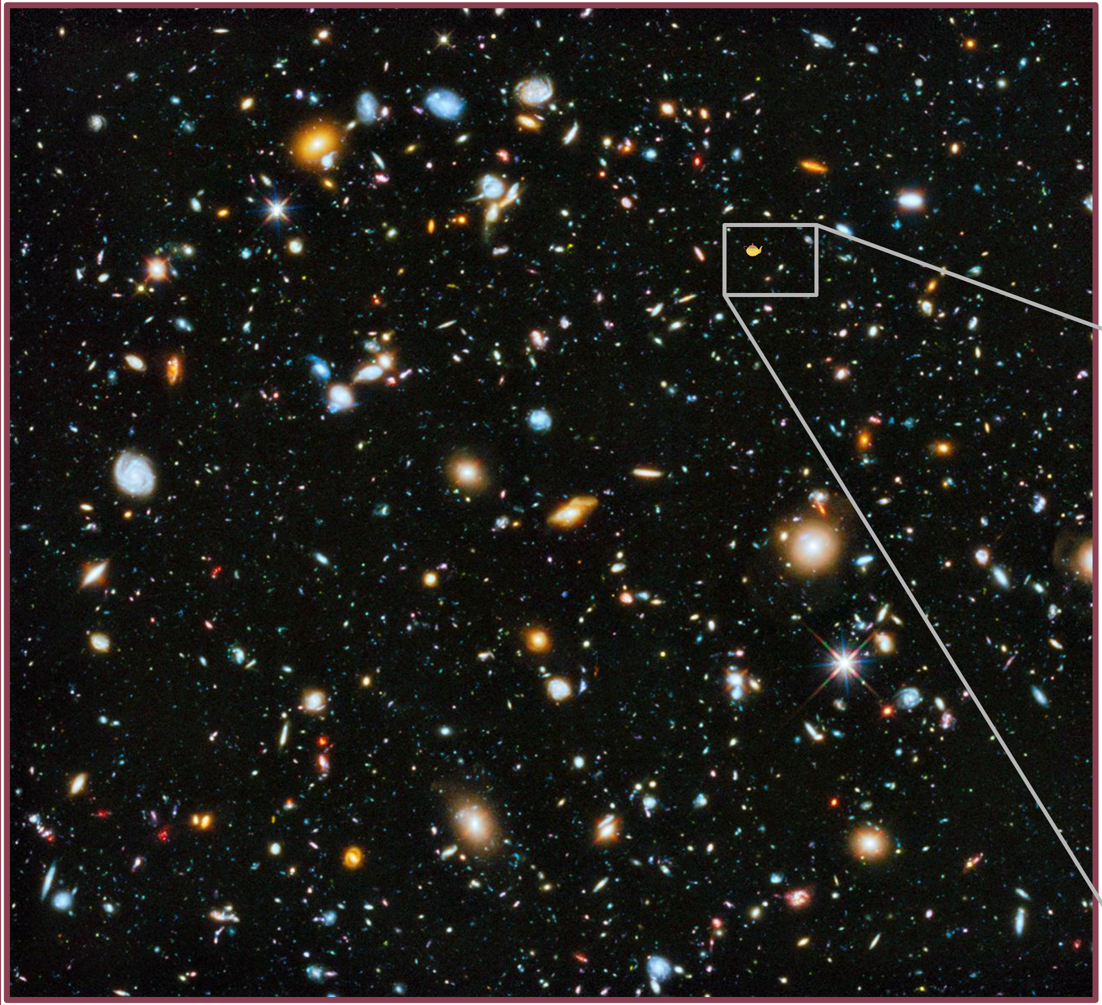
↓

16 × 10^6 galaxies





Teapots in the sky



Q1: 50 deg^2: 5% outliers → 250 000 objects at $H_E<24$ / $I_E<25$

DR1: 2500 deg^2 : 5% outliers → ~10 million objects at $H_E<24$ / $I_E<25$