Identifying Lyα emitters by learning from post Reionization-era galaxies in the CANDELS survey

Lorenzo Napolitano (**lorenzo.napolitano@inaf.it**)

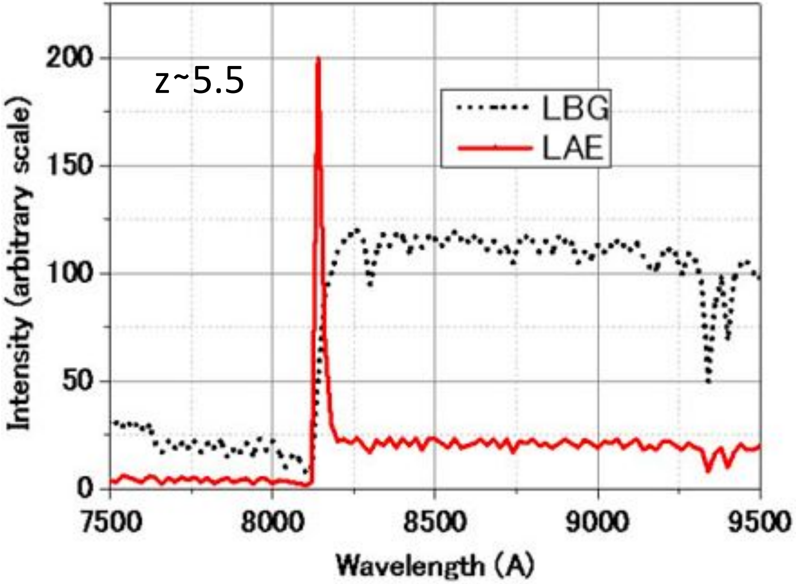
Supervisor: L. Pentericci

Collaborators: A. Calabrò, P. Santini, M. Castellano, S. Mascia, et al.

What is a Lyman-Alpha Emitter (LAE)?

- LAEs are Star Forming galaxies with strong Lyα Emission (EW > 20 Å) in their spectra.

- The UV Lyα Emission line (1216 Å restframe) is a probe to the presence of a recombination region HII.



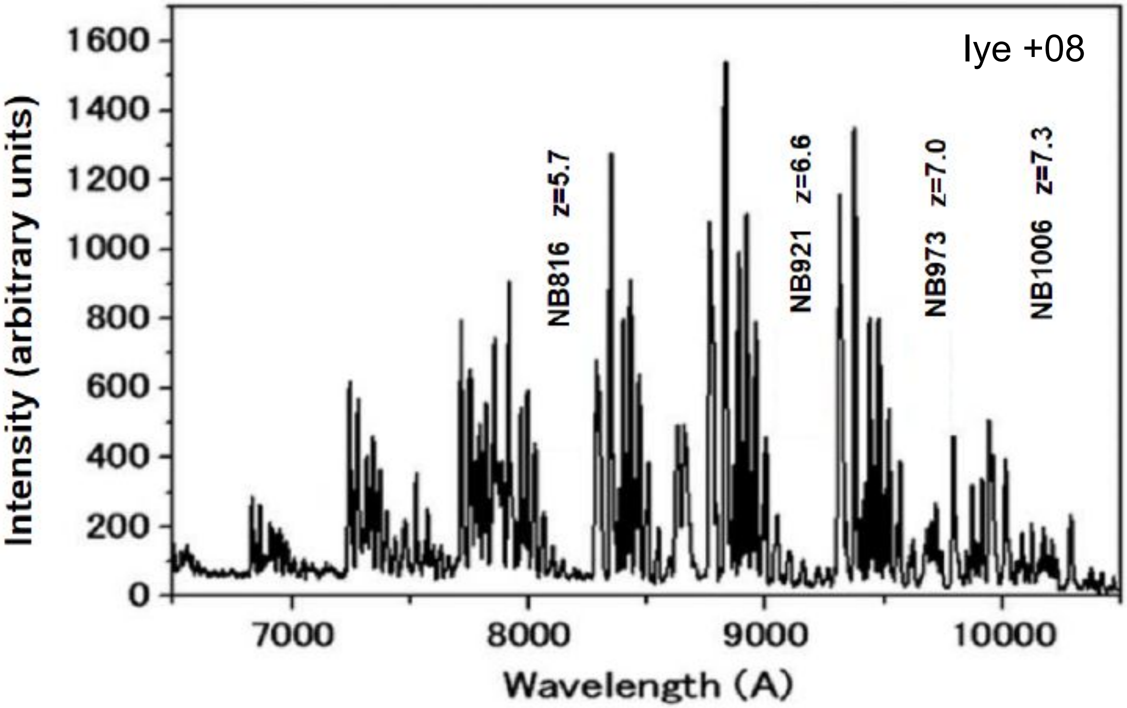
Spectroscopy is needed to confirm LAE candidates.

Candidates are often pre-selected through **Narrow Band (NB) surveys** through a colour excess (BB-NB).

$$z = \frac{\lambda_{NB}}{1216\text{Å}} - 1$$

Limitations of Narrow Band surveys

- They probe small redshift ranges (100-200 Å width), hence small cosmological volumes.
- Affected by OH atmospheric emission lines at high z.
- Severe contamination due to metal emission lines (CIV, MgII, [OII], [OIII]) of galaxies at lower redshift (Ciardullo+02, Fujita+03, Pentericci+18)
- Transient object, such as variable AGN or supernovae (Dunlop+13)



Typical Properties of LAEs

Typical physical and morphological properties of LAEs
(e.g. Ono+10, Hagen+14, Kojima+17, Paulino-Afonso+18, Ouchi+20):

- ➔ $M \sim (10^8 - 10^9)\ M_\odot$
- ➔ $SFR \sim (1\text{-}10)\ M_\odot/yr$
- ➔ $E(B\text{-}V) \sim 0 - 0.2$
- ➔ $R_e \sim 1$ kpc

***Explanation*:**
Due to its scattering nature, the N_{HI} and dust can quench the Lyα emission

Objective:

Can we distinguish LAEs just from physical and morphological properties?

Our project: CANDELS data

Physical & Morphological Data

We considered galaxies in **GOODS-South** (Merlin+21), **COSMOS** (Nayyeri+17) and **UDS** (Galametz+13).

For each galaxy we have:

(A) The **physical** properties - SED fitting Santini+22:

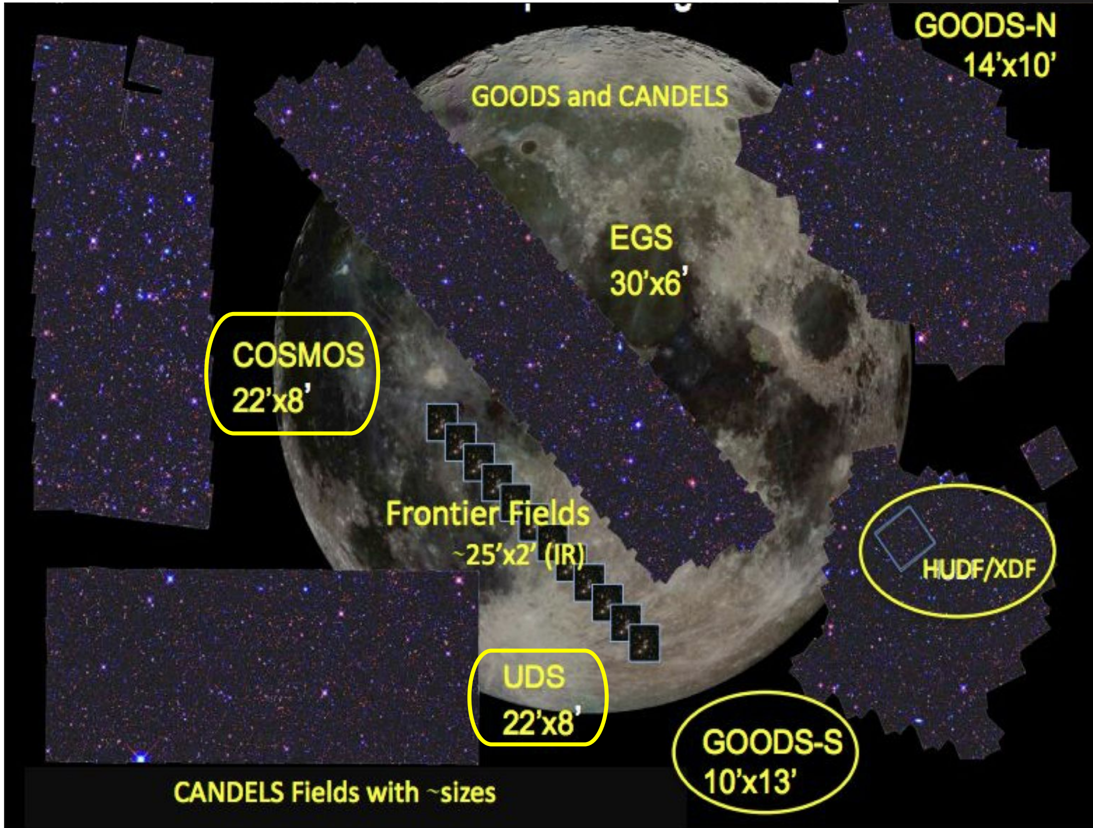
SFH(t) ~ $(t^2/\tau)\exp(-t/\tau)$, Chabrier+03 IMF

Calzetti+00 law for dust extinction

- **Stellar Mass**
- **E(B-V)**
- **SFR**
- **Metallicity**
- **Age**

(B) The **morphological** properties (van der Wel+12 fit on H_{F160W} band)

- **R_e**
- **projected axis ratio**
- **Sérsic index**



Our project: Spectroscopic Data

Spectroscopic observations available for a subset of the galaxies in CANDELS from multiple surveys. We collected the Lyα flux and equivalent width from the literature:

*In total we have the **spectroscopic information of 1578 galaxies*** in the range z ∈ [2,7.9].

*From spectroscopic data **we classify** them into **LAEs or not (NLAEs)***

Field	Galaxies	LAEs	NLAEs
GOODS-S	841	340	501
COSMOS	408	107	301
UDS	329	78	251

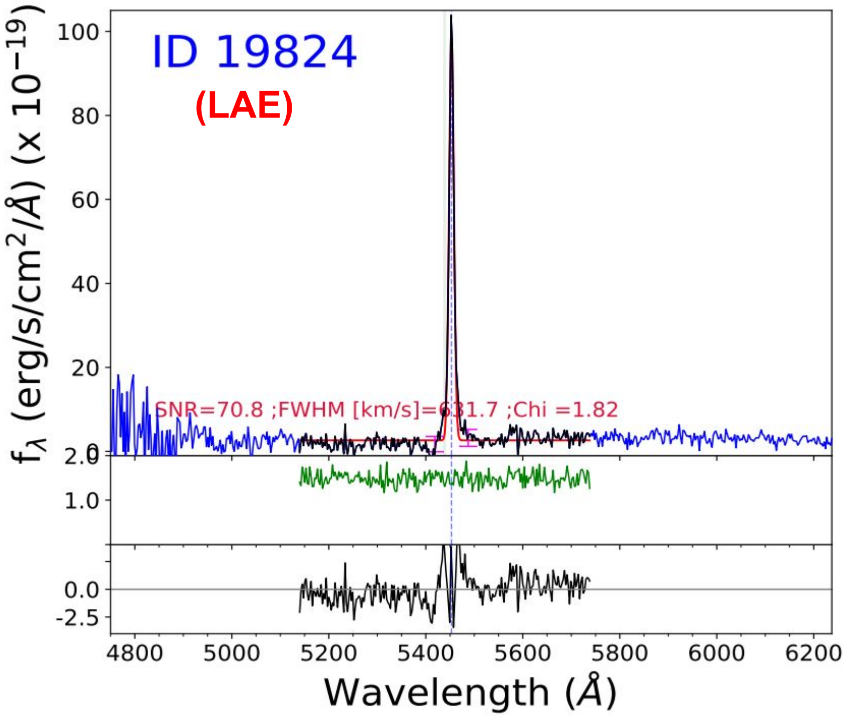
Survey	Number of galaxies	Author
VANDELS	615	Pentericci+18, Garilli+21
VUDS	162	Cassata+15, Tasca+17
MUSE-Deep/-Wide	302	Schmidt+21
CANDELS-z7	109	Pentericci+18
GMASS	20	Kurk+13
GOODS South team	144	Popesso+09, Balestra+10, Vanzella+18
DEIMOS	41	Hasinger+18
zCOSMOS-Deep	185	Lilly+07, Kashino+22

Fitting archival data:

Spectroscopic Data

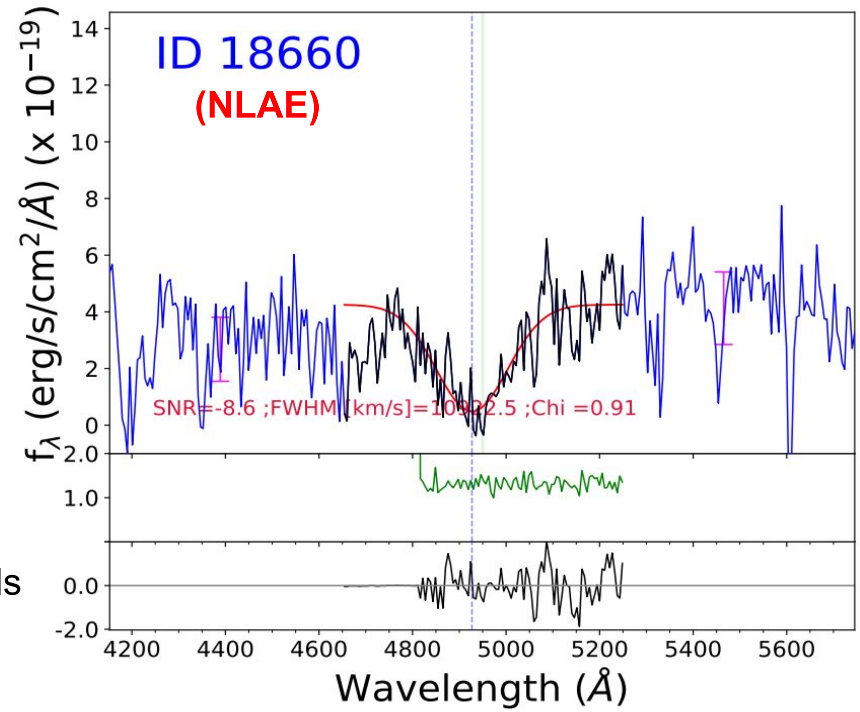
Examples of measuring Lyα Flux and EW

$$W_\lambda = \int (1 - F_\lambda / F_0) d\lambda.$$



Noise

Residuals



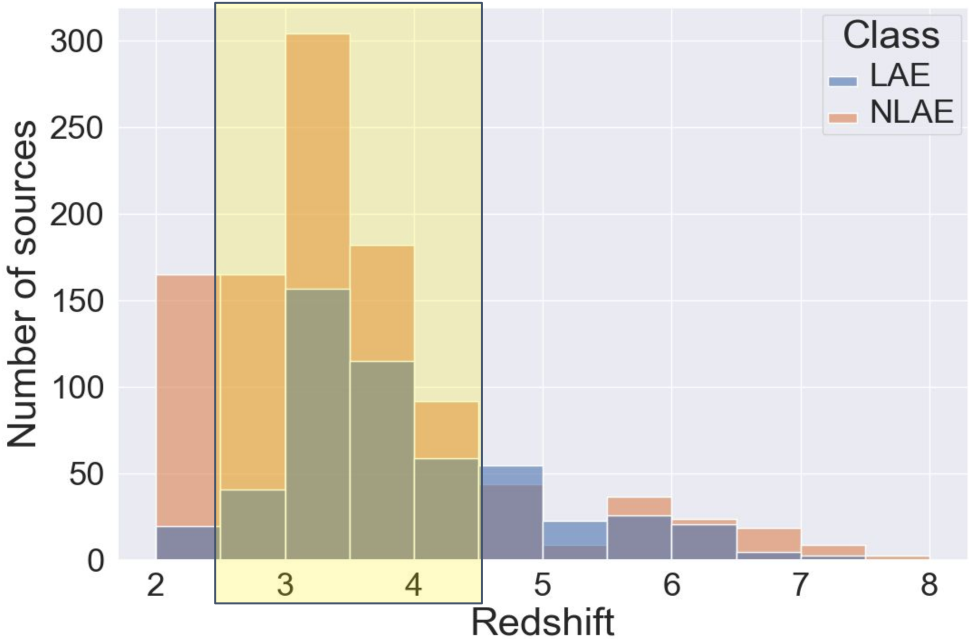
Labeled spectroscopic sample

We focused on the redshift range $z \in [2.5,4.5]$, avoiding the effect of the neutral IGM.
Our subset consists of 1115 galaxies.

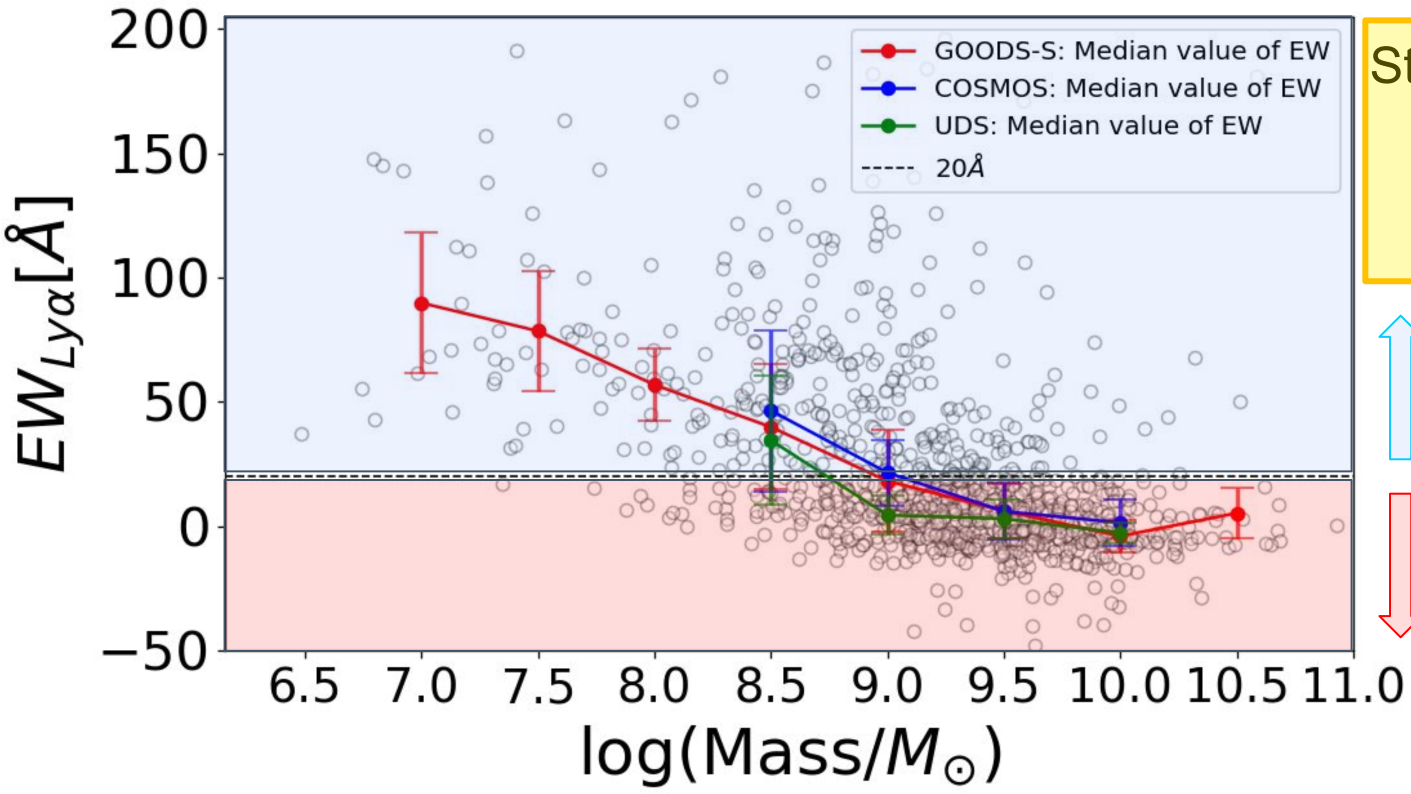
TRAINING the Machine Learning:

In the range $z \in [2.5,4.5]$:
372 LAEs and
743 NLAEs.

Physical+Morphological properties and labels
as example pairs



LAEs tend to have small stellar masses



Stellar Masses

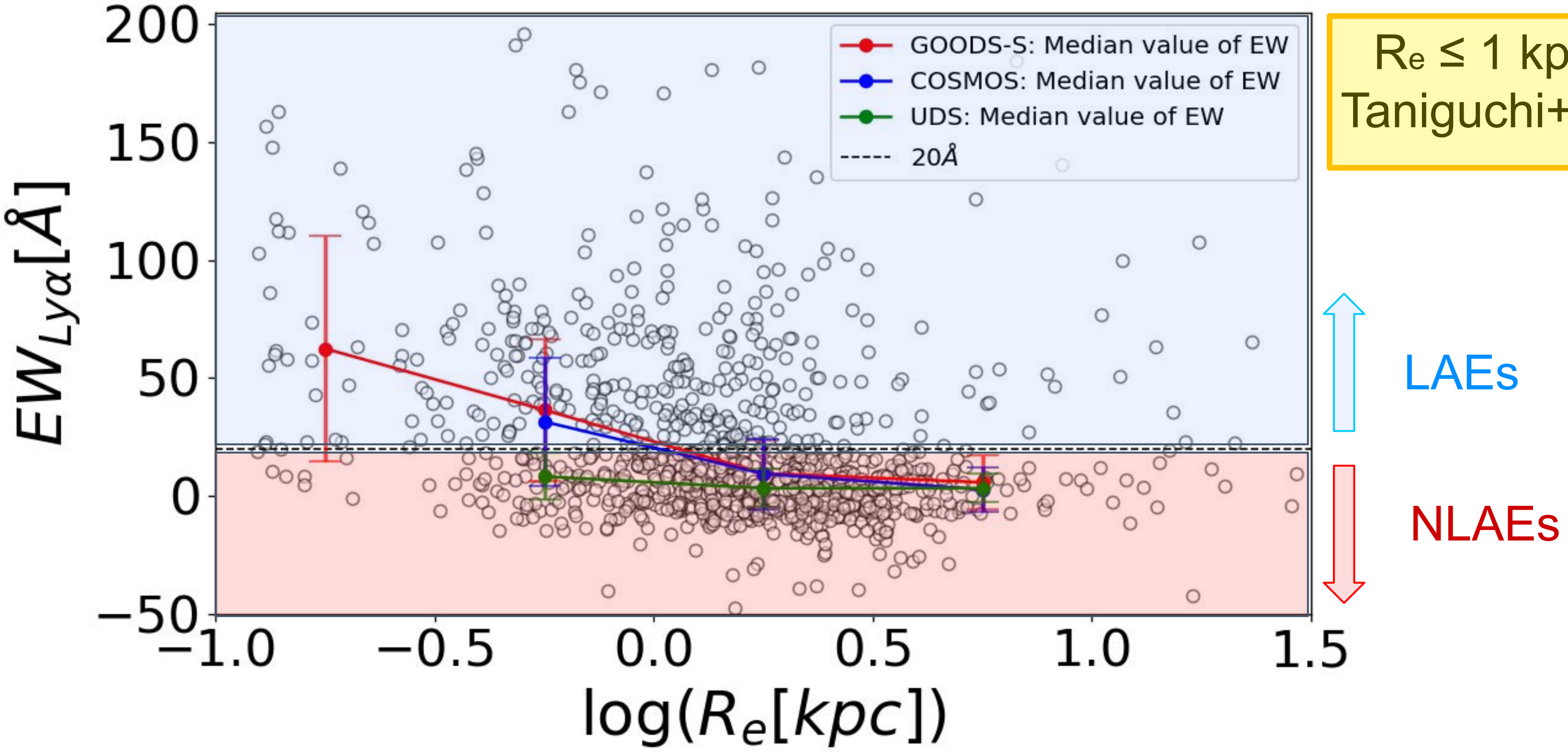
≲ 10^9 $M_\odot$

Ouchi+20

LAEs

NLAEs

LAEs tend to be compact galaxies



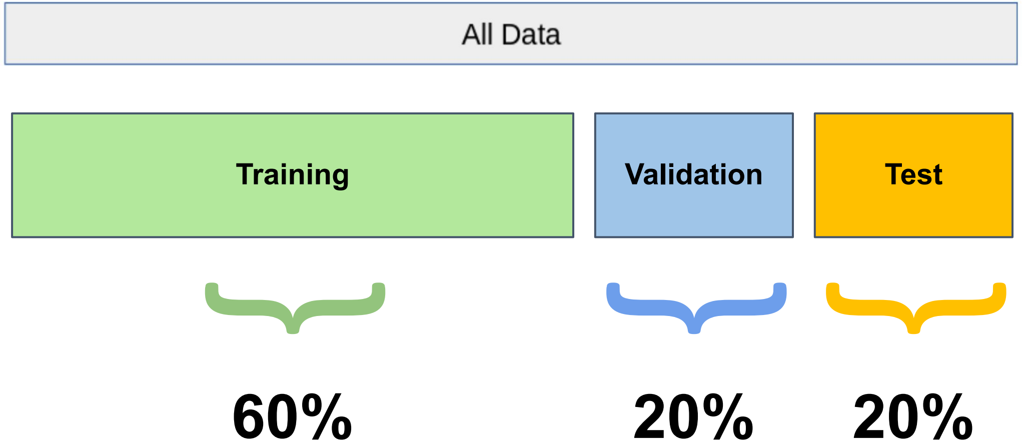
$R_e \leq 1$ kpc
Taniguchi+09

Supervised ML and Cross validation

In supervised ML we need to construct three different subsamples (**training**, **validation** and **test** sets):

- Our algorithm will know the correct labelling of the **training** set data;
- will be optimized on the **validation** set score;
- its final performances will be tested on the independent **test** set.

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$



TP - True Positives

FP - False Positives

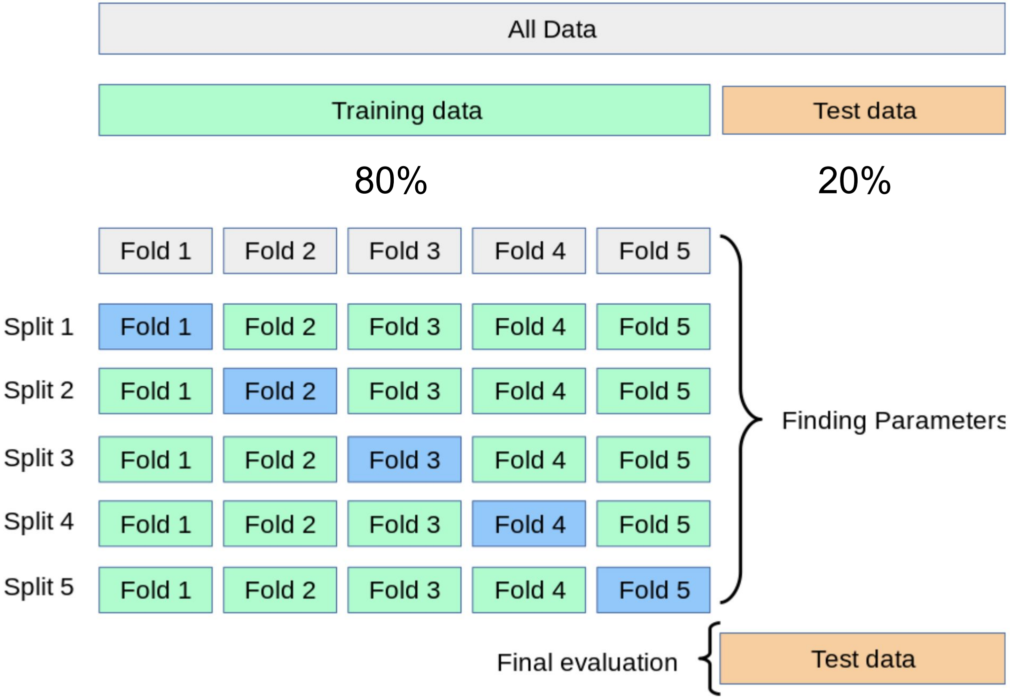
TN - True Negatives

FN - False Negatives

Supervised ML and Cross validation

We opted for a 5-fold cross validation approach:

- A model is trained using 4 of the folds as training data;
- The resulting model is validated on the remaining fold of the training set;
- The final performances are measured on the independent test set.



Supervised ML and Cross validation

In supervised ML we need to construct three different subsamples (**training**, **validation** and **test** sets):

Our algorithm will know the correct labelling of the training set data, will be optimized on the validation set score and its final perform... will be tested on the independent test set.

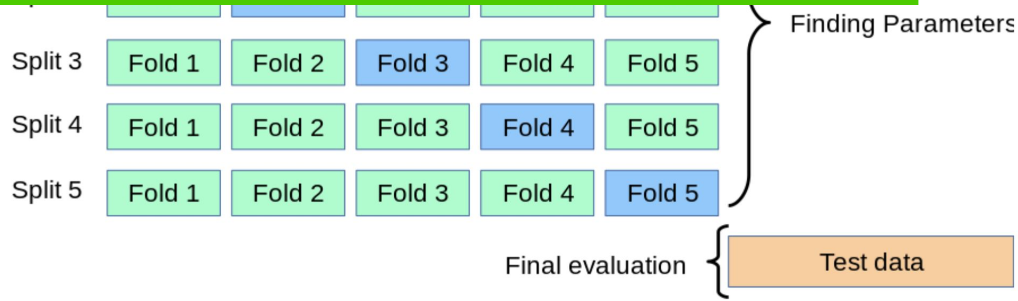
Acc ... N) ... + F

TP - Tru ... Ne
FP - False ... e N

We opted for ... on appr

- A model is trained using 4 of the folds as training data;
- the resulting model is validated on the remaining part of the data

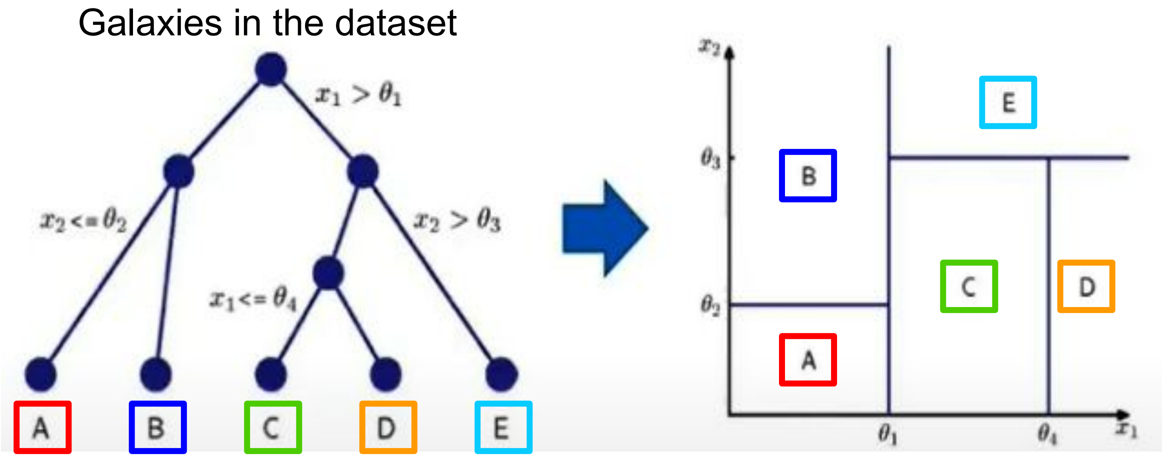
This have the advantage to increase the number of samples which can be used for learning the model.
It is of key importance when applying ML to small datasets, like in our case.



Decision Tree classifier

It is a binary recursive partition of the features' space. The goal is to find the optimal partition so that different classes are segregated in different hyper-rectangles.

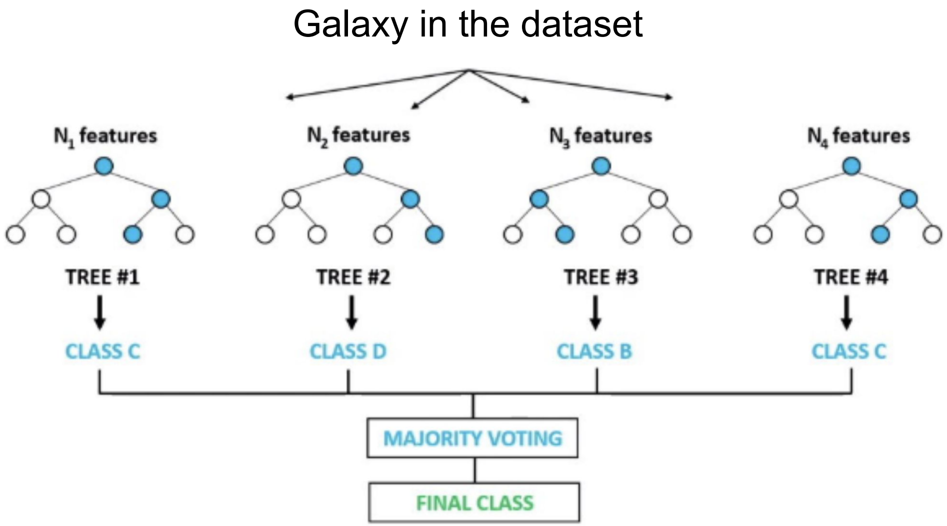
It is transparent but it suffers from the overfitting problem.



Random Forest classifier

Breiman+01 introduced this ensemble learning classifier that combines multiple decision trees to improve classification performance.

- Each tree in the forest is slightly different from the others.
- During prediction each tree votes. The final prediction is the majority vote of all the trees.

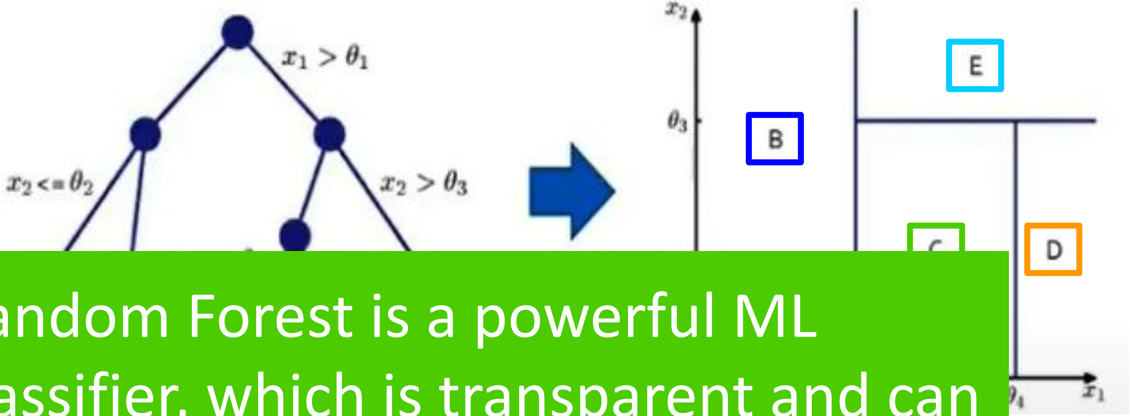


Decision Tree classifier

It is a binary recursive partition of the features' space. The goal is to find the optimal partition so that different classes are segregated in different hyper-rectangles

It is trans the overfitti

Galaxies in the dataset

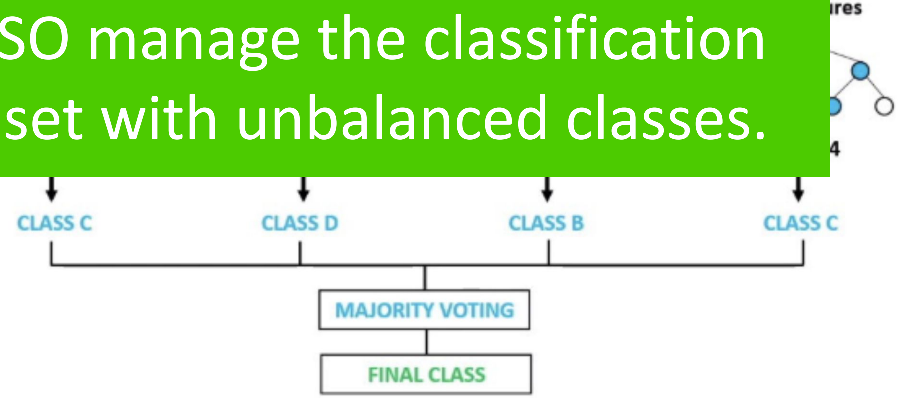


Ra ss

Breima ble l classifier ecisio improve cla nce.

Random Forest is a powerful ML classifier, which is transparent and can be successfully applied to predict never seen data.
It can ALSO manage the classification of a dataset with unbalanced classes.

- Each tree in the forest is slightly different from the others.
- During prediction each tree votes. The final prediction is the majority vote of all the trees.

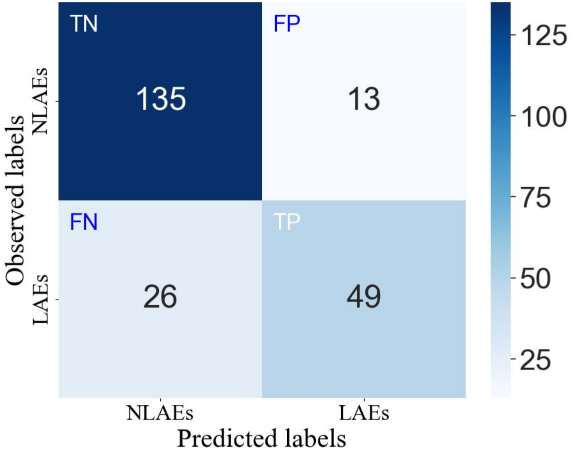


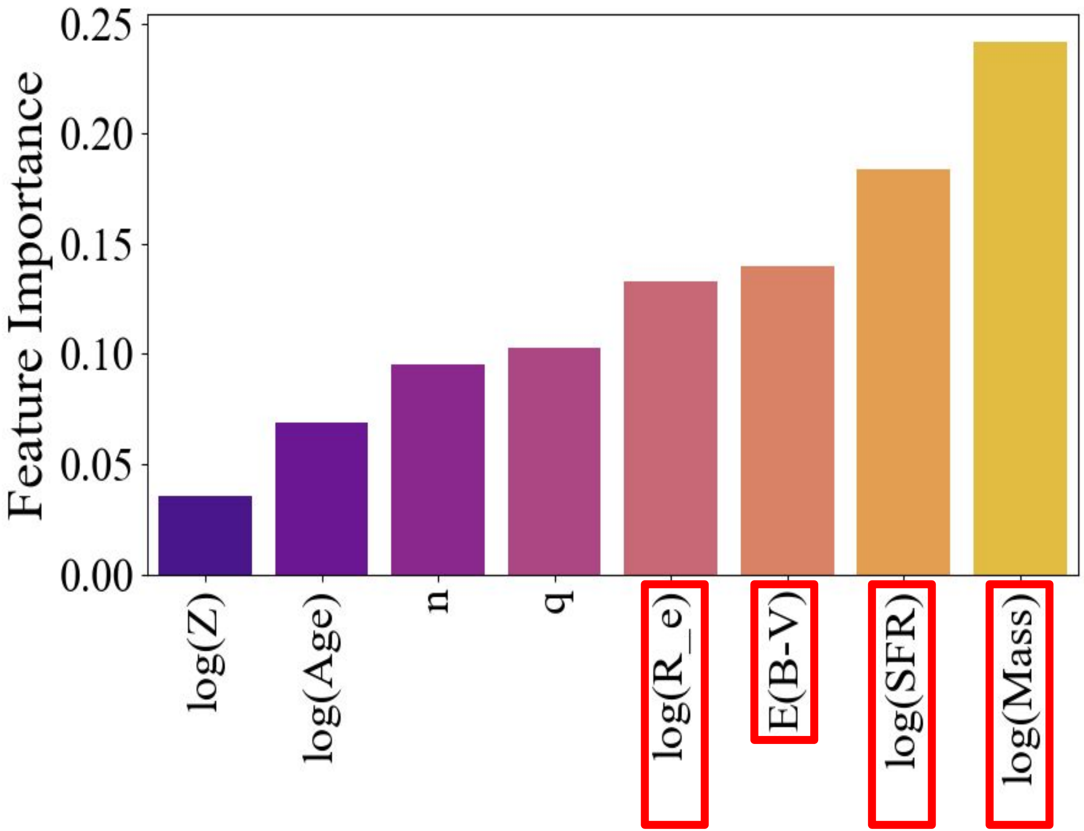
Random Forest Classifier

Optimal hyper-parameters:
n_estimators = 500
max_features = 3
max_depth = 20

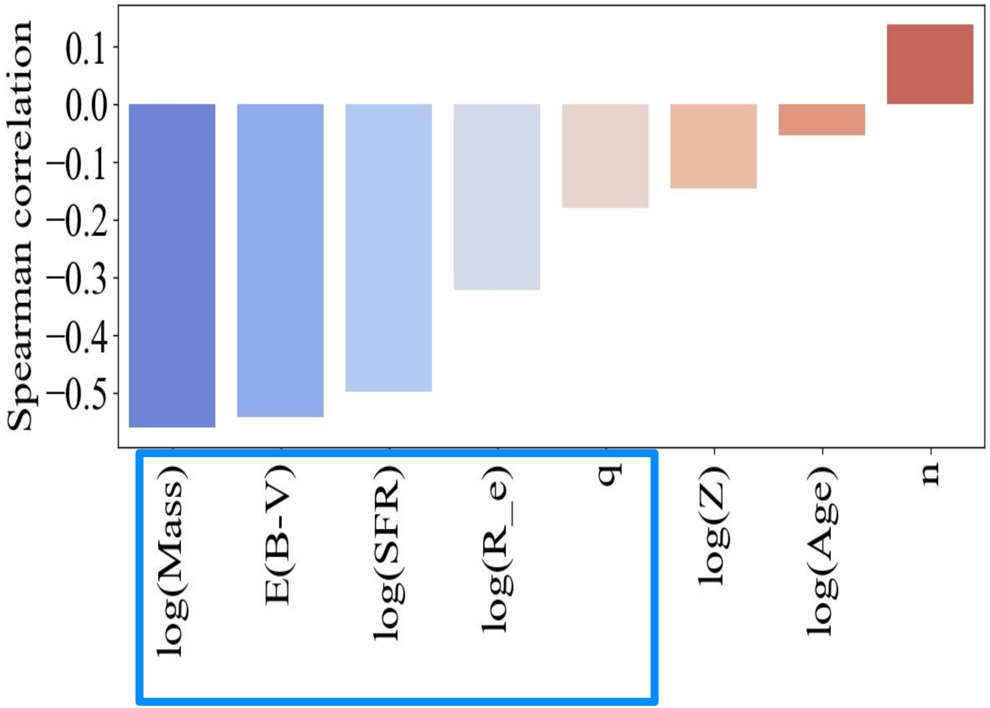
Cross Validation accuracy: 79.4 ± 3.6%

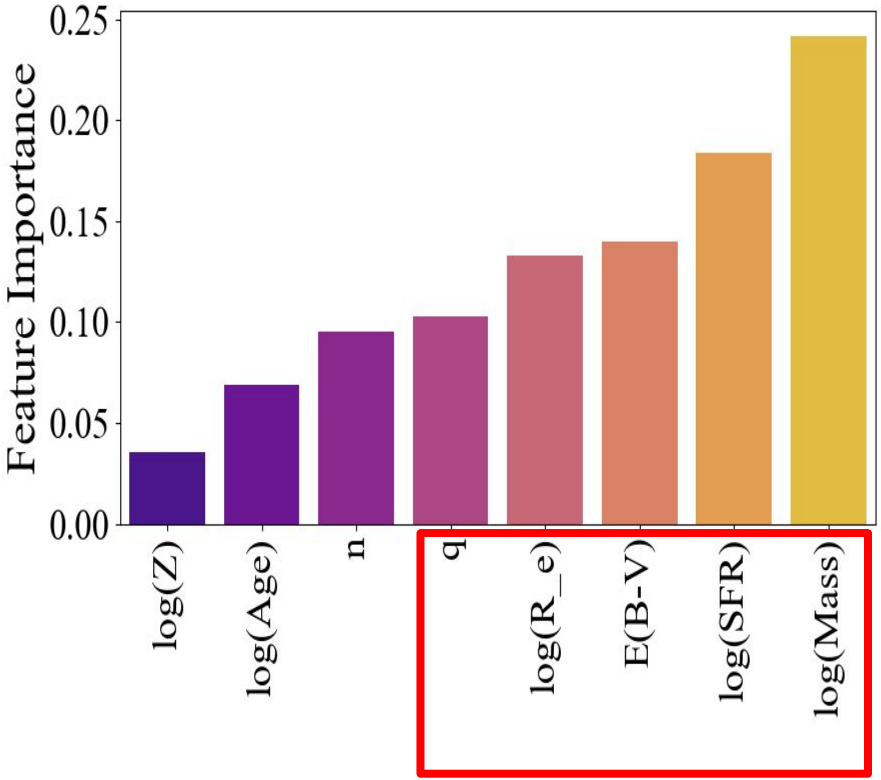
Test accuracy: 79.7 ± 2.1%



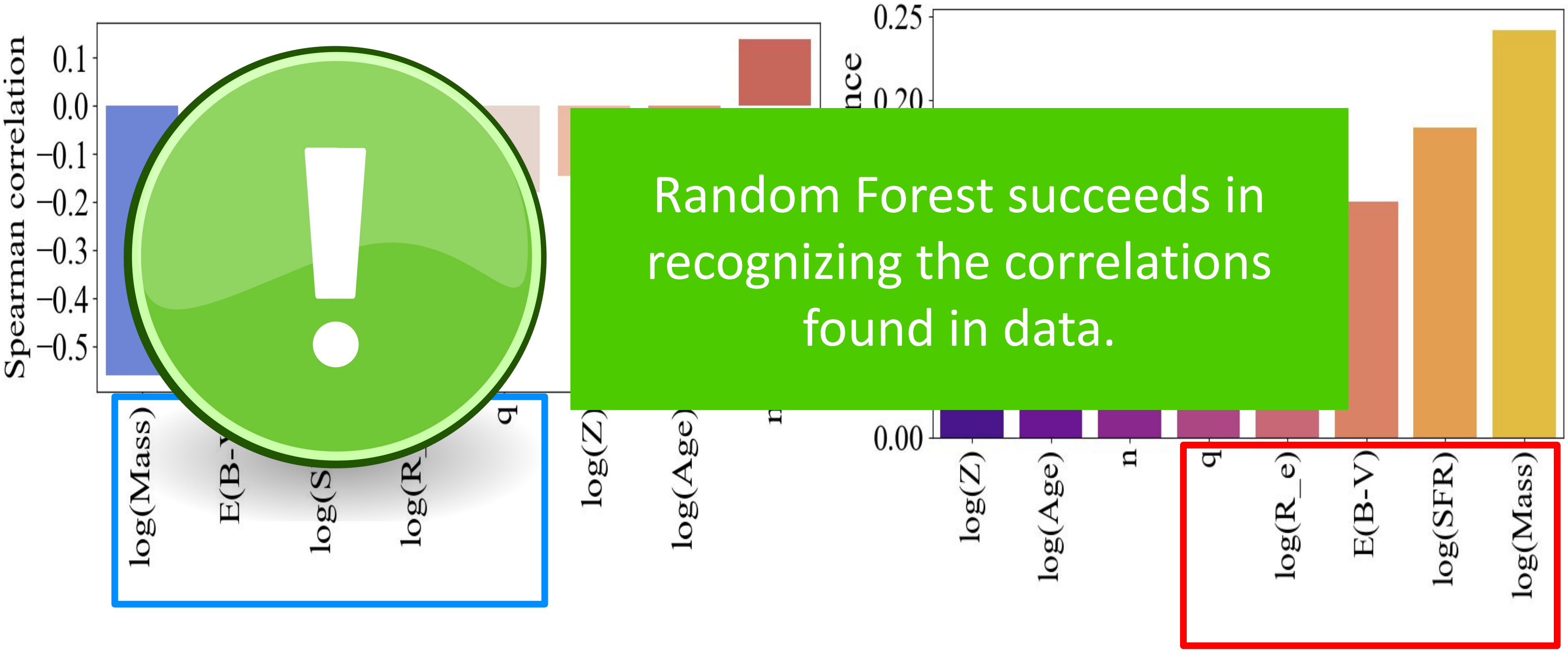


Random Forest Classifier





Random Forest Classifier



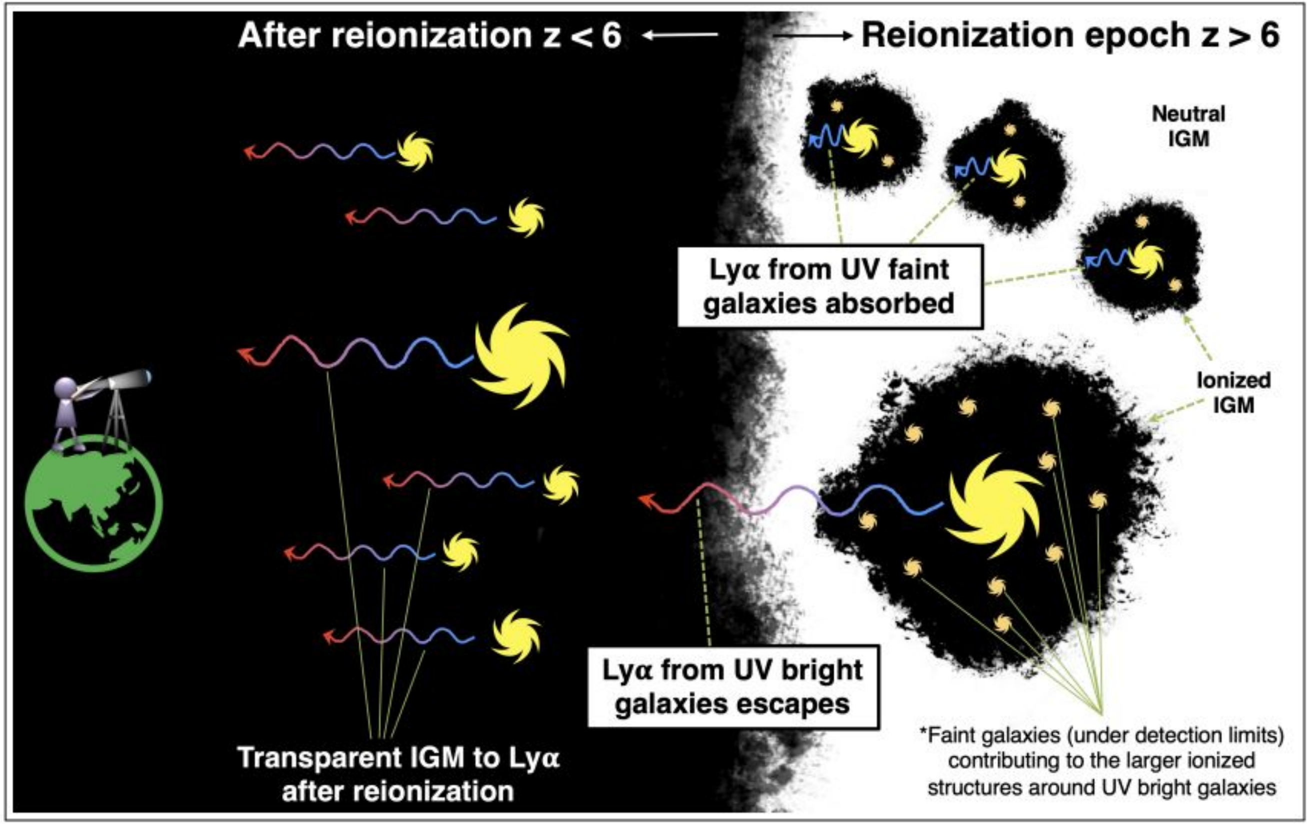
Random Forest succeeds in recognizing the correlations found in data.

Possible applications: topology of Reionization - LAEs clustering

By drawing informed predictions on LAE candidates, we can plan successful "blind" spectroscopic surveys.

In turn this will open the possibility to systematically study the spatial overdensities of confirmed LAEs, probing the spatial distribution scenario of the ionized gas during the Epoch of Reionization.

Jung+21

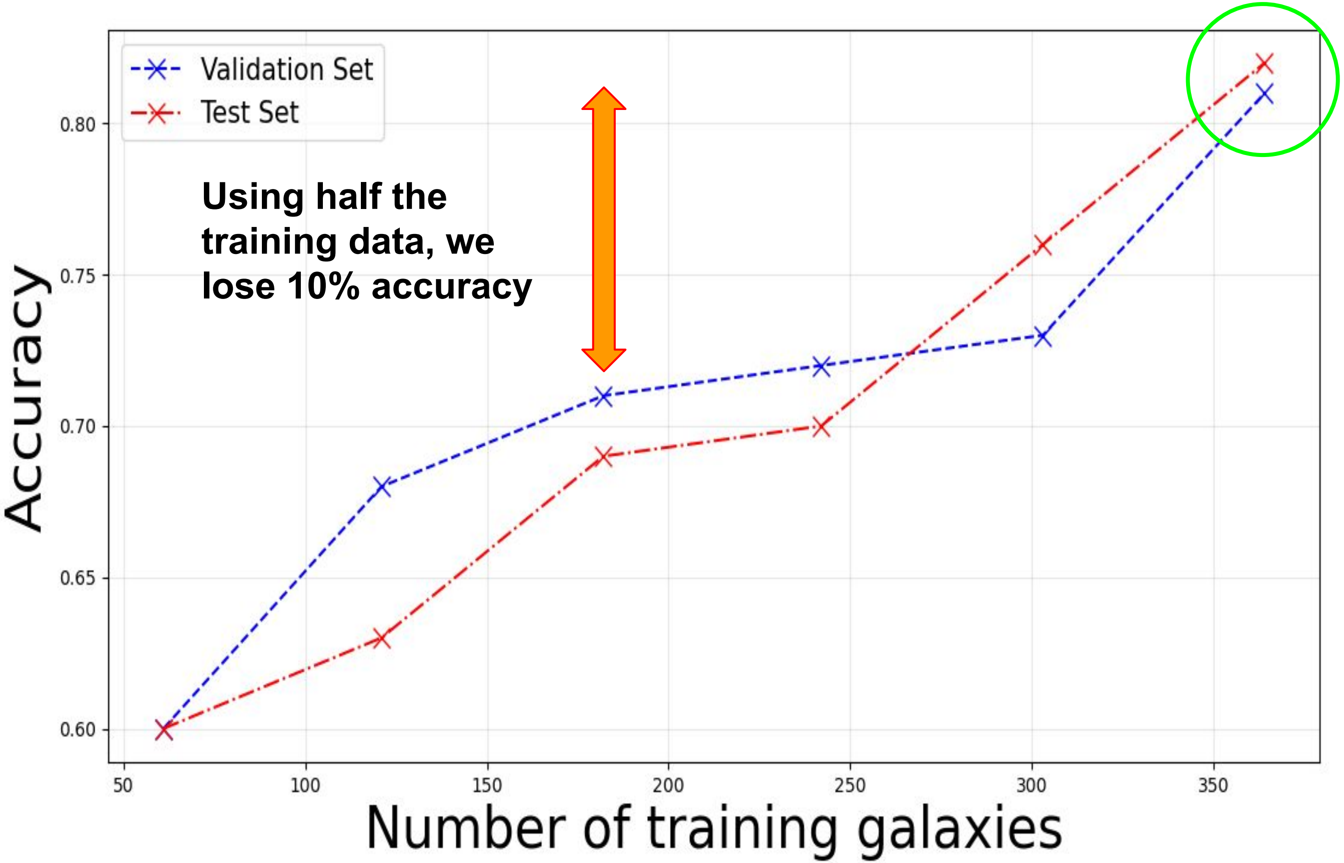


“Probing the Epoch of Reionization with the first galaxies"

Supervisor: Prof. Laura Pentericci (Sapienza)
Co-Supervisor: Dr. Marco Castellano (INAF-OAR)

Thank you for your attention

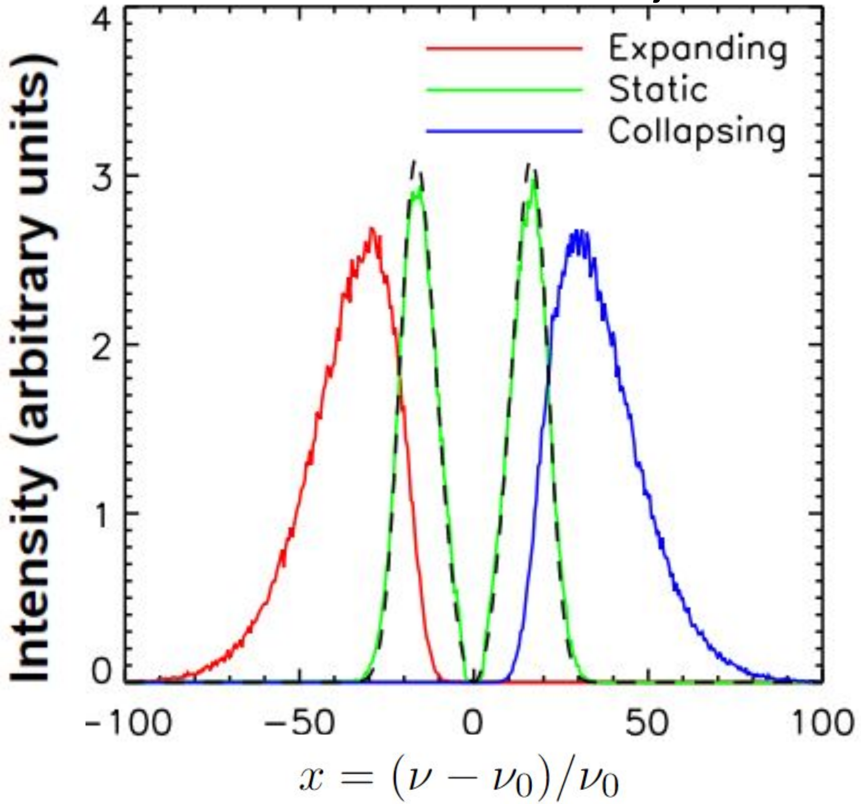
ML Performances and the training sample



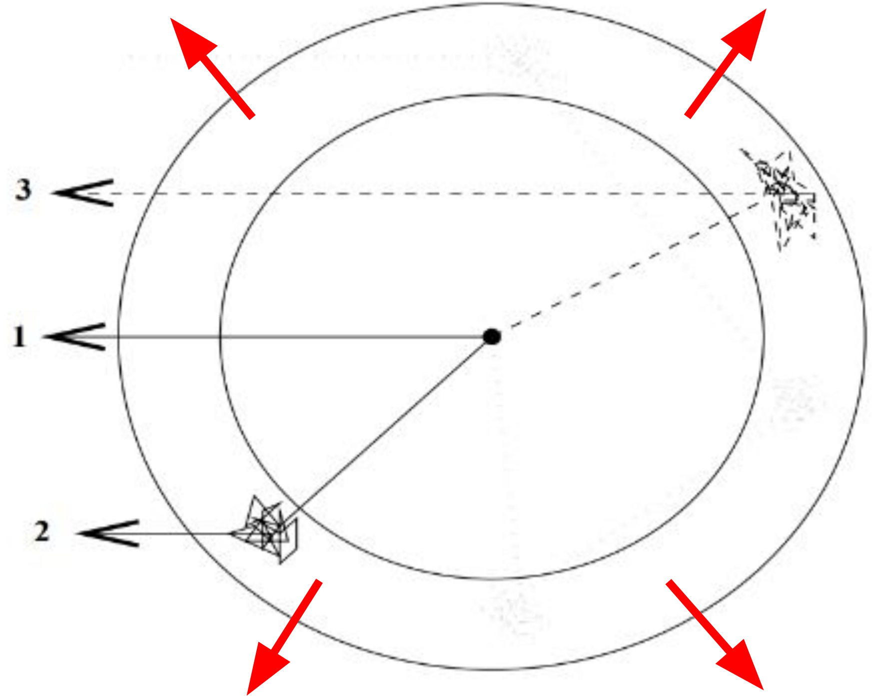
This performances are the best we could achieve given the current number of galaxies with a spectroscopic follow-up.

Lyα Radiative Transfer in the ISM

Dijkstra +17



Verhamme et al. +06



Parameters: $N_{\rm HI},\ v_{\rm shell},\ \tau_{\rm dust}$

Topology of Reionization - LAEs clustering

21-cm signal expected from SKA in the late 2020s **galaxies Lyα signal map obtained from simulations**