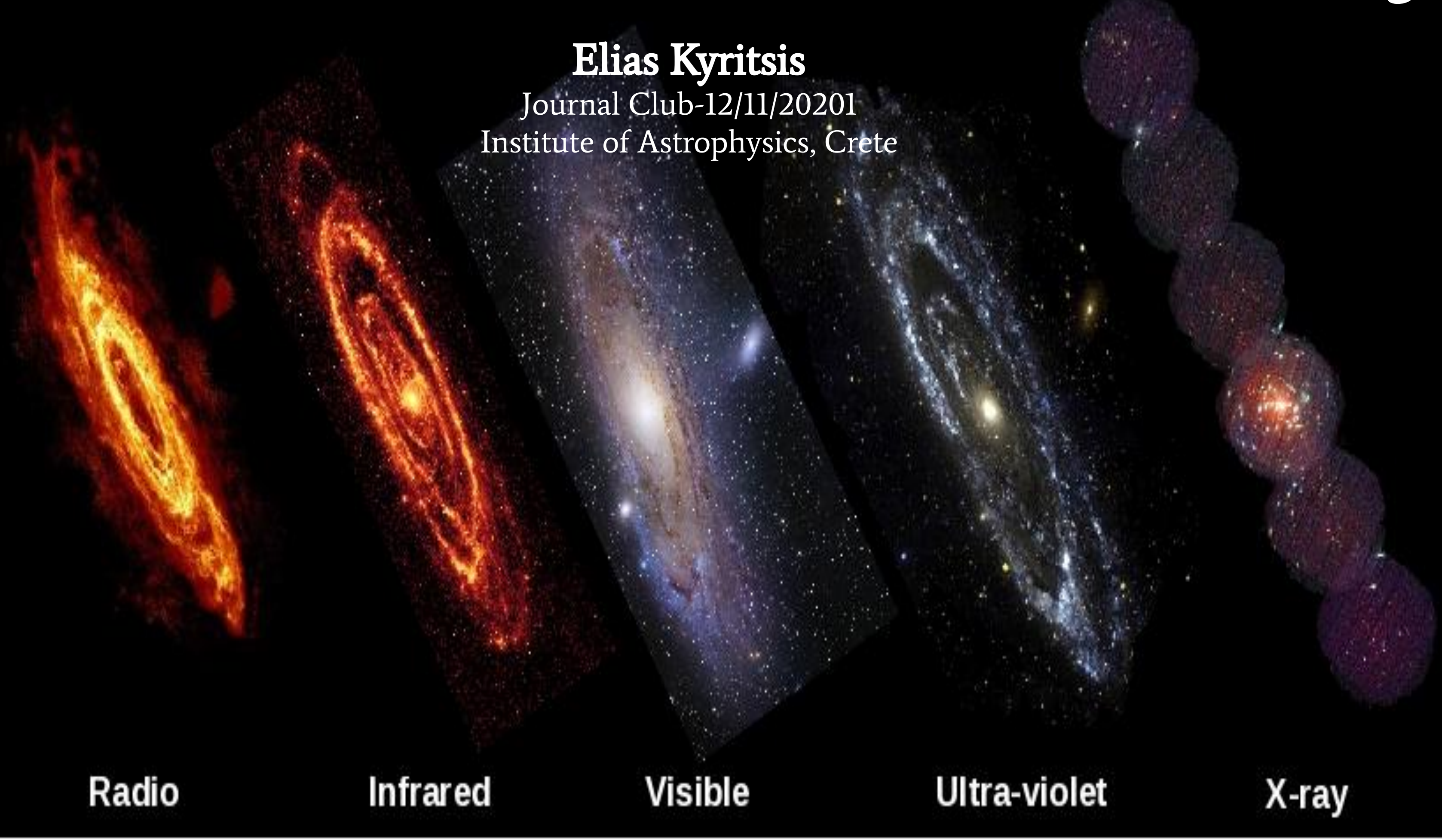
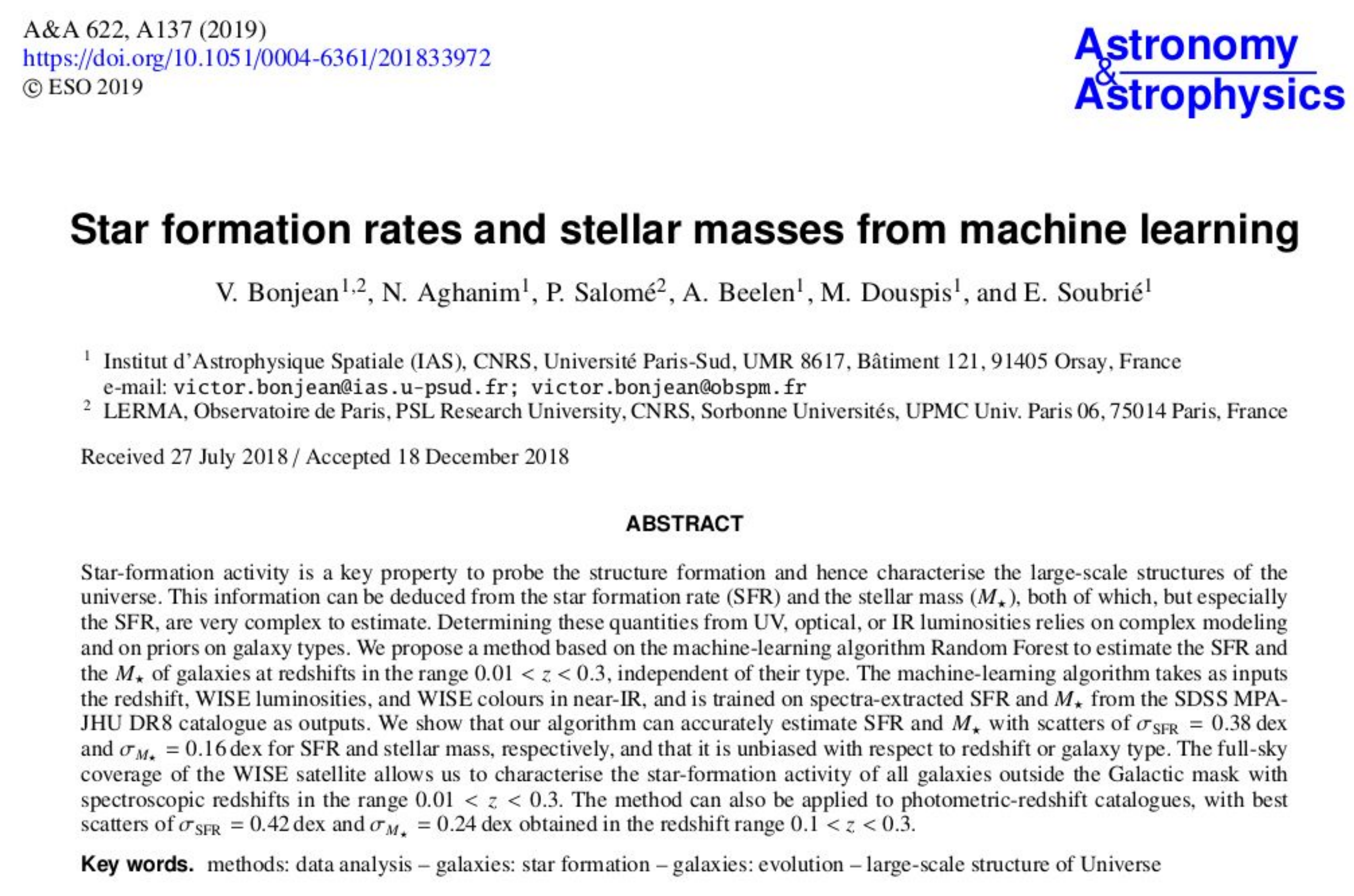
“Star formation rates and stellar masses from machine learning”

Elias Kyritsis

Journal Club-12/11/20201

Institute of Astrophysics, Crete





A&A 622, A137 (2019)
https://doi.org/10.1051/0004-6361/201833972
© ESO 2019

Astronomy & Astrophysics

Star formation rates and stellar masses from machine learning

V. Bonjean[1,2], N. Aghanim[1], P. Salomé[2], A. Beelen[1], M. Douspis[1], and E. Soubrié[1]

[1] Institut d'Astrophysique Spatiale (IAS), CNRS, Université Paris-Sud, UMR 8617, Bâtiment 121, 91405 Orsay, France
e-mail: victor.bonjean@ias.u-psud.fr; victor.bonjean@obspm.fr

[2] LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, 75014 Paris, France

Received 27 July 2018 / Accepted 18 December 2018

ABSTRACT

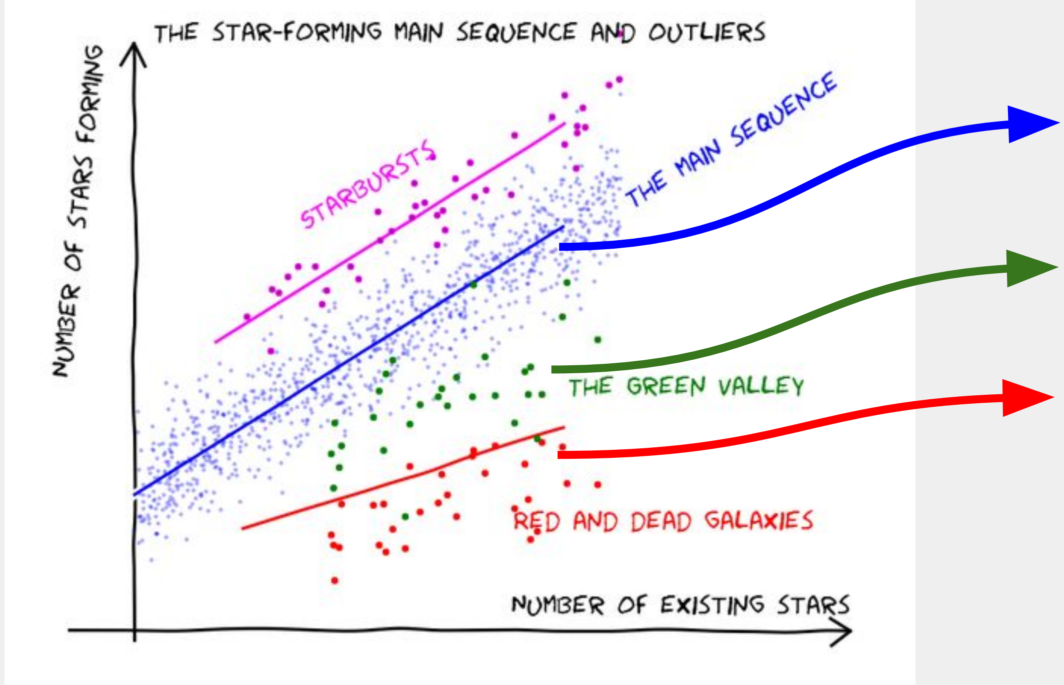
Star-formation activity is a key property to probe the structure formation and hence characterise the large-scale structures of the universe. This information can be deduced from the star formation rate (SFR) and the stellar mass ($M_\star$), both of which, but especially the SFR, are very complex to estimate. Determining these quantities from UV, optical, or IR luminosities relies on complex modeling and on priors on galaxy types. We propose a method based on the machine-learning algorithm Random Forest to estimate the SFR and the $M_\star$ of galaxies at redshifts in the range $0.01 < z < 0.3$, independent of their type. The machine-learning algorithm takes as inputs the redshift, WISE luminosities, and WISE colours in near-IR, and is trained on spectra-extracted SFR and $M_\star$ from the SDSS MPA-JHU DR8 catalogue as outputs. We show that our algorithm can accurately estimate SFR and $M_\star$ with scatters of $\sigma_{\rm SFR} = 0.38$ dex and $\sigma_{M_\star} = 0.16$ dex for SFR and stellar mass, respectively, and that it is unbiased with respect to redshift or galaxy type. The full-sky coverage of the WISE satellite allows us to characterise the star-formation activity of all galaxies outside the Galactic mask with spectroscopic redshifts in the range $0.01 < z < 0.3$. The method can also be applied to photometric-redshift catalogues, with best scatters of $\sigma_{\rm SFR} = 0.42$ dex and $\sigma_{M_\star} = 0.24$ dex obtained in the redshift range $0.1 < z < 0.3$.

Key words. methods: data analysis – galaxies: star formation – galaxies: evolution – large-scale structure of Universe

Introduction

Star-formation activity ⇨ Probe the structure formation of galaxies ⇨ Characterise the LSS of the universe

The activity of galaxy is usually defined in terms of **Star Formation Rate (SFR)** or **stellar mass ($M_\star$)**



Star forming galaxies align along a main sequence

Star-formation stops - Galaxy loose its cold gas
Transitioning stage (Green-valley)

Passive (red or dead galaxies)

Estimation of **SFR** and **$M_\star$** is crucial !

https://www.quora.com/Are-there-certain-types-of-galaxies-where-stars-are-formed-much-quicker-than-normal

Introduction

Estimation of SFR & $M_\star$ is complex !

Directly or indirectly related to observations of stars

Dependence of the star properties on wavelengths across the E/M spectrum

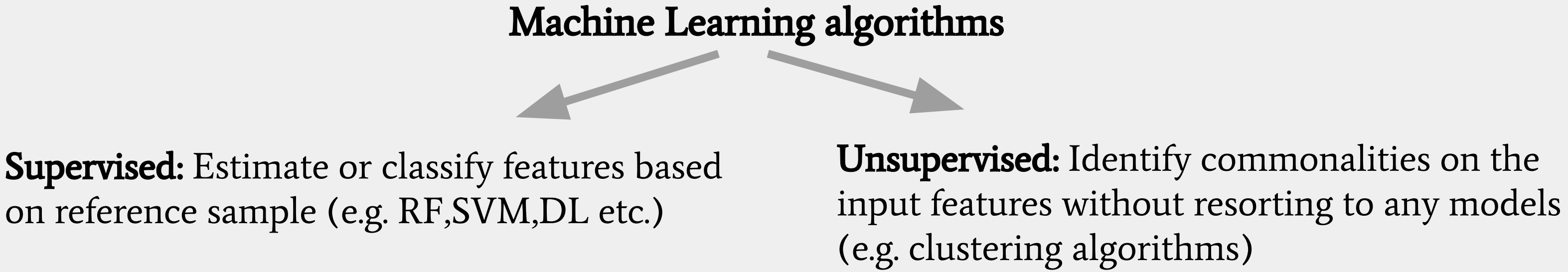
- **UV:** Traces the youngest stellar population in a galaxy (O- and B-type stars) → Directly related to SFR
- **Optical:** Traces the non-ionizing low-mass old stars → Directly related to $M_\star$
- **NIR (0.8 μm - 3 μm):** Traces the old and non-massive stars → Related to $M_\star$
- **MIR (3 μm - 70 μm):** Contribution of dust becomes predominant → Indirectly related to SFR
    - 8-12 μm: Heated small grains & PAHs → Useful to study the composition & abundance of dust
    - 20-70 μm: Thermalised dust & large grains heated by UV emission (O- and B-type stars)

Introduction

Limitations

- These relations are applicable only to galaxies with known type.
- Optical spectroscopic data are needed to estimate SFR & $M_\star$ → Costly in terms of observing time

Solution



Machine Learning algorithms

Supervised: Estimate or classify features based on reference sample (e.g. RF,SVM,DL etc.)

Unsupervised: Identify commonalities on the input features without resorting to any models (e.g. clustering algorithms)

This work

Estimation of SFR & $M_\star$ independently of any complex model or any priors on galaxy types using a supervised machine learning algorithm.

Data for the construction of the training sample

WISE

All-sky survey in 4 near- & mid-infrared bands:

- W1: 3.4 μm
- W2: 4.6 μm
- W3: 12 μm
- W4: 22 μm

AllWISE Source Catalogue:

747.634.026 detected sources with accurate positions, photometry etc.

In this work: **Profile-fitting** photometry for **W1,W2,W3**

SDSS

Photometric data for the ⅓ of the sky (u,g,r,i,z) **+ Spectroscopic data** for >3 million objects

Based on SDSS data **MPA-JHU DR8** catalog was generated.

~1.8 million objects

- SFR & $M_\star$
- BPT classification for the activity of the galaxies
- spectroscopic-z ($0.1 < z < 0.3$)

Data for the construction of the training sample

Pre-processing of the data

1. Cleaning MPA-JHU catalog by removing unreliable measurements → 794.633 galaxies

2. Cross-match the MPA-JHU catalog with AllWISE SC (r=6”)
   a. Removing multiple associations → 603.293 galaxies
   b. Removing unreliable WISE magnitude measurements (~5% of the total sample)

3. Final pure training sample : **573.582 galaxies** (W1,W2,W3,$\mathrm{SFR}_{\mathrm{SDSS}}$,$M_{\star\mathrm{SDSS}}$, z)

Inputs & Outputs for the algorithm

Inputs

- z → SFR evolves with redshift
- LW1 [3.4 μm] → Proxy for the $M_{\star}$
- LW3 [12 μm] → Proxy for the SFR
- W1-W2 [3.4-4.6 μm] → Accounting for galaxy-type
- W2-W3 [4.6-12 μm] → Accounting for galaxy-type

Outputs

- $\mathrm{SFR}_{\mathrm{SDSS}}$
- $M_{\star\mathrm{SDSS}}$

Data for the construction of the training sample

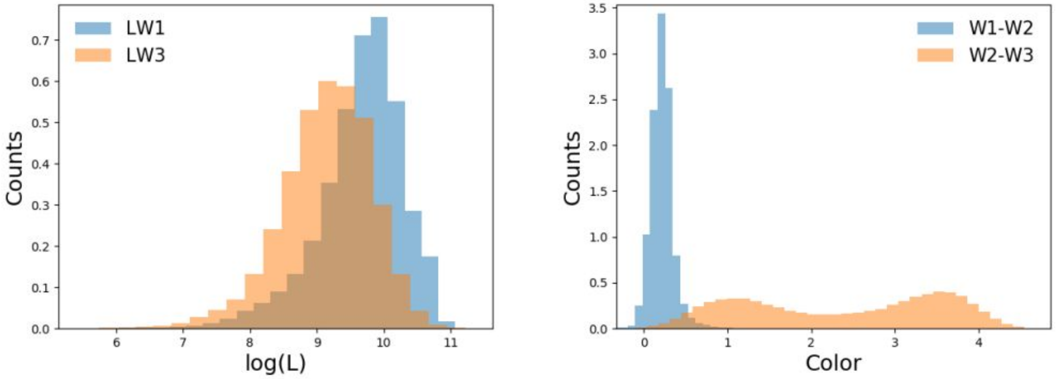


Fig. 6. Histograms showing the range of the input data (luminosities and colours) for the sources of the training catalogue.

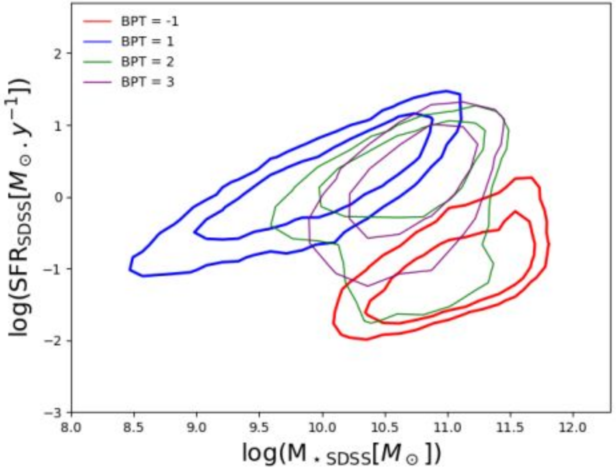
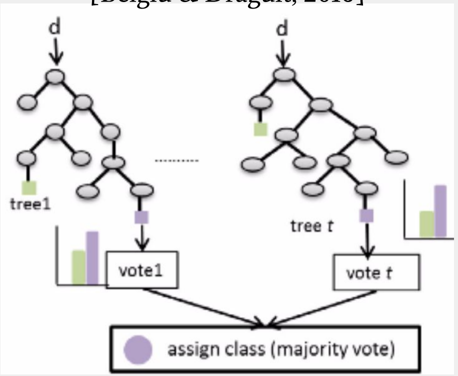


Fig. 2. SFR and $M_\star$ provided in the SDSS catalogue. Each contour represents a galaxy population as defined by their position on the BPT diagram. Red contours represent passive galaxies, blue contours star-forming galaxies, green contours the galaxies from the green valley, and purple contours the AGNs.

Random Forest Algorithm

[Belgiu & Draguit, 2016]



- How does it work ?
    - RF is a collection of a large number of decision trees
    - Randomly-selected data subsets of the initial dataset
    - Randomly-selected subsets of the features in each node of each decision tree

Each tree in the forest suggests a class → Majority vote → Final prediction

Optimization of the RF

Training sample: 60% + Validation sample: 20% + Test sample: 20%

Varying M (number of trees) & d_{max} (max depth) → RF score

RF score = $100 \times R^2$ where $R^2 = 1 - \sigma_{res}^2 / \sigma^2$

σ_{res}^2 : residual sum of squares

σ^2 : variance of the output distribution

Final optimized values: $M=40$, $d_{max} = 12$ → RF score = 84.5 %

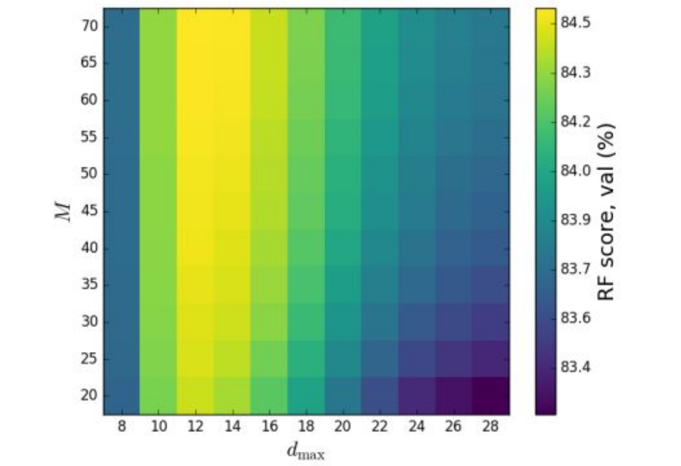
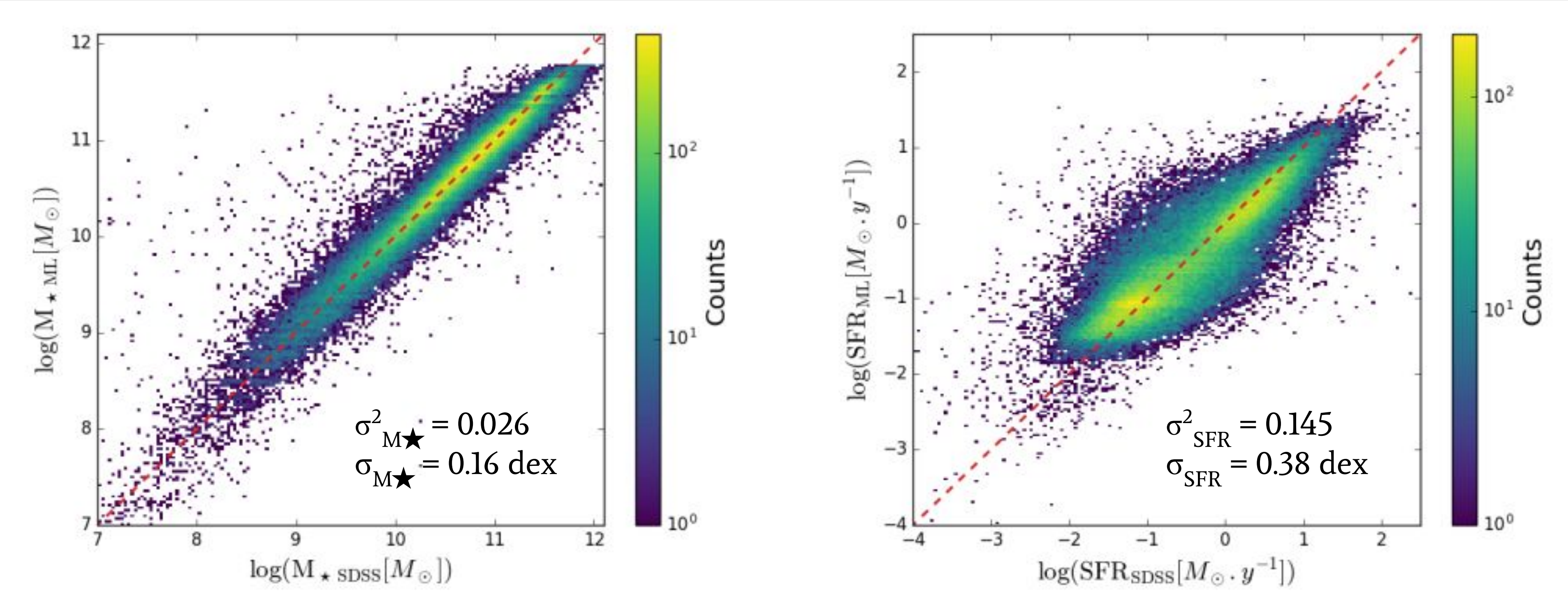


Fig. 3. Percentage score of the RF results on the validation sample as a function of the RF parameters M and d_{max} (M being the number of trees and d_{max} the maximum depth). Setting $M = 40$ and $d_{max} = 12$ is enough in our case to optimise the RF.

Results



$\sigma^2_{M\bigstar} = 0.026$
$\sigma_{M\bigstar} = 0.16$ dex

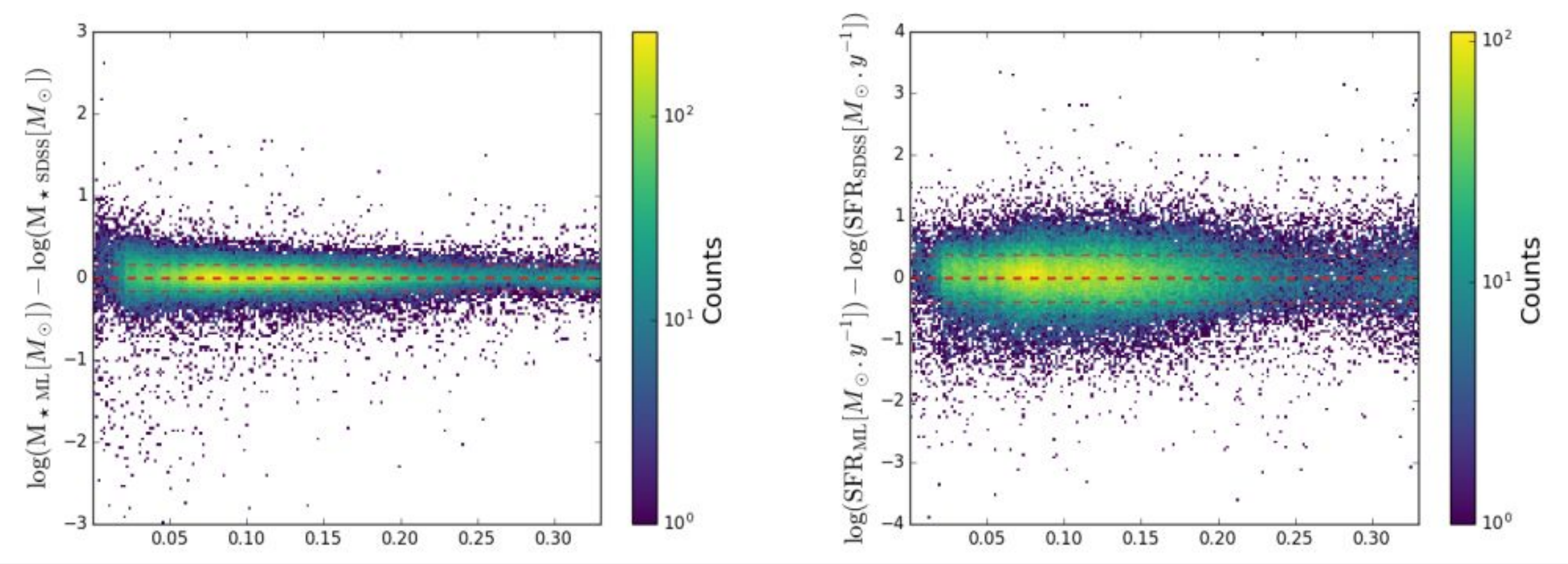
$\sigma^2_{SFR} = 0.145$
$\sigma_{SFR} = 0.38$ dex

Overall good agreement between reference values & predictions → RF is well trained

Results

Chasing the biases

Does the z-dependence of SFR & $M_\star$ induce potential biases ?

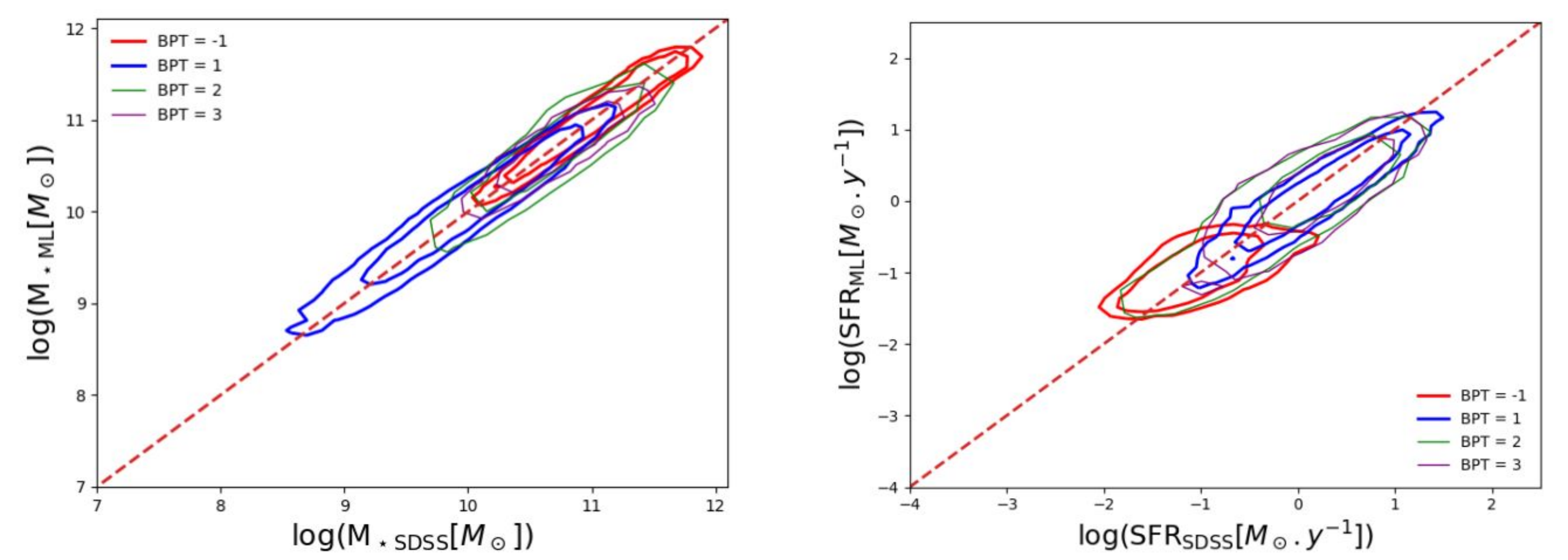


In general, no obvious bias on redshift is observed

Results

Chasing the biases

Do the different galaxy types induce potential biases ?

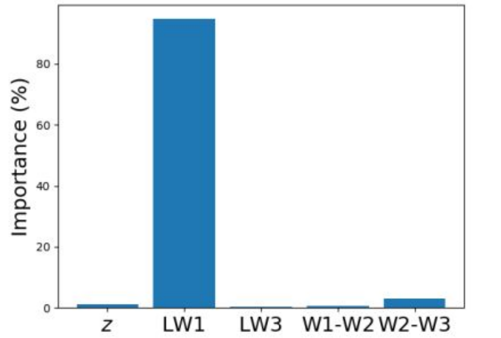


RF performs equally well regardless the galaxy-type

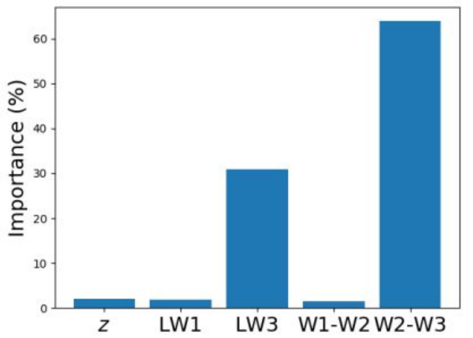
Results

Feature importance

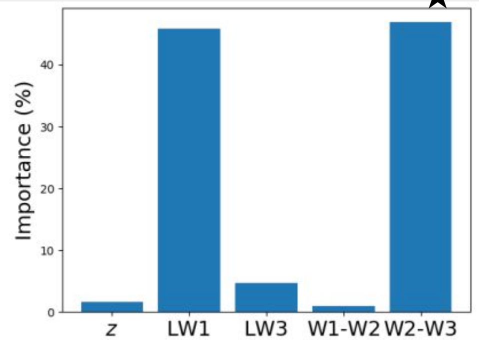
RF trained only to estimate $M_\star$



RF trained only to estimate SFR



RF trained only to estimate both SFR & $M_\star$



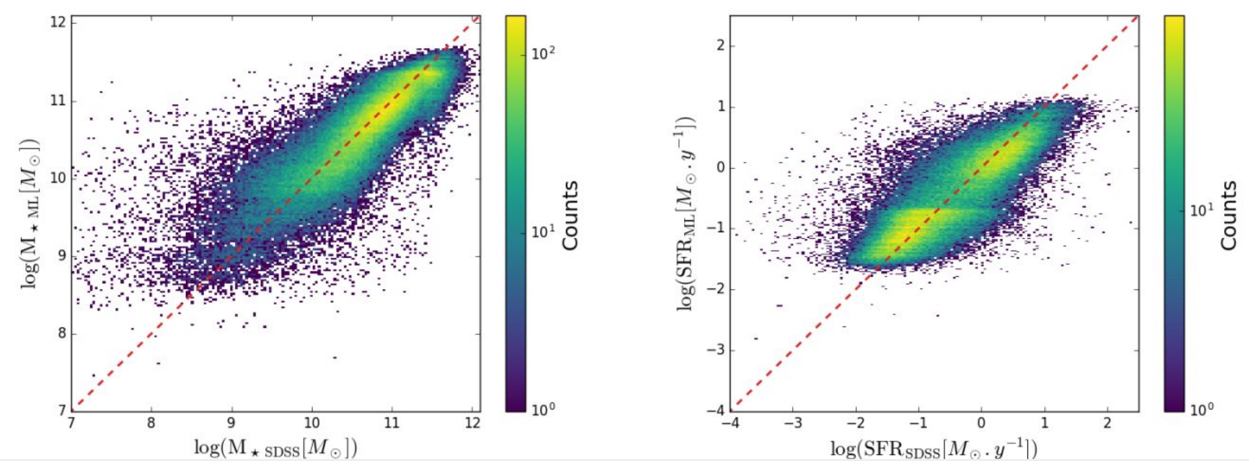
- Redshift is not important (only on the dependence on distance, as it is also hidden in LW1 & LW2).
- W1-W2 is not important, whatever the output.
- The important features follow the expected behavior.

Results

Can we train a RF model without any redshift information ?

The need of z to compute SFR & $M_\star$ is very restrictive (spec-z : hard to obtain, photo-z: not always precise)

Train the RF only with: W1,W2,W1-W2 and W2-W3

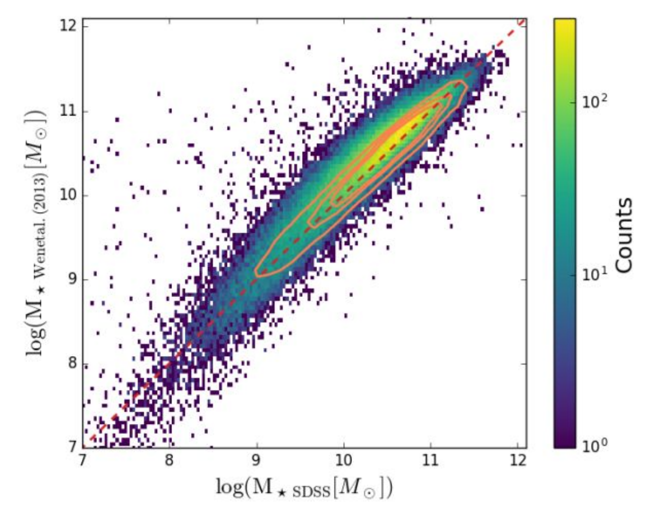


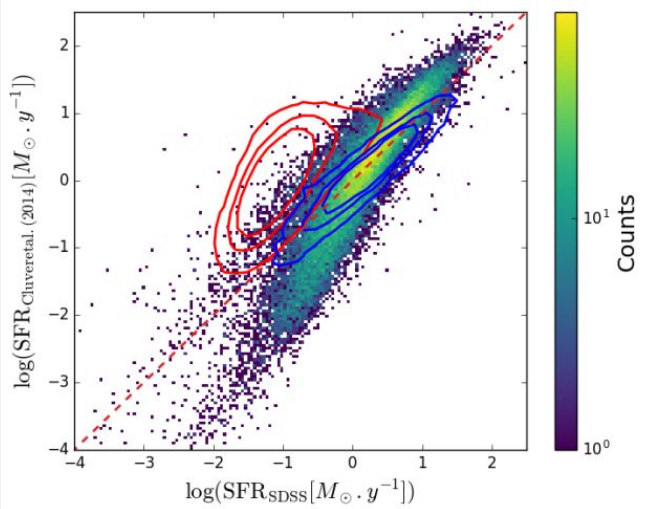
The accuracy of the method is highly degraded

Redshift information is needed !

Results

Comparison with the standard methods for the calculation of SFR & M$_\star$





$$\log\left(M_{\star \mathrm{Wen}}\right) = 1.12 \times \log\left(\mathrm{LW1}\right) - 0.04,$$

$\sigma_{\mathrm{M}\star\mathrm{Wen}}$= 0.23 dex & $\sigma_{\mathrm{M}\star\mathrm{ML}}$ = 0.16 dex

5 features are more effective than 1 !

$$\log\left(\mathrm{SFR}_{\mathrm{Cluver}}\right) = 1.13 \times \log\left(\mathrm{LW3}\right) - 10.24.$$

$\sigma_{\mathrm{SFRCluver}}$= 0.47 dex & σ_{SFRML} = 0.30 dex (active galaxies)
$\sigma_{\mathrm{SFRCluver}}$= 0.49 dex & σ_{SFRML} = 0.38 dex (passive galaxies)

5 features are more effective than 1 !
A linear relation, such as Cluver, is not effective in passive galaxies.

Take home message

- Development of a new method, based on machine learning, for the the estimation of SFR & $M_\bigstar$. Features needed : LW1, LW2, W1-W2 , W2-W3 and z .

- The method performs well (typical scatters), independently of galaxy-type, and it is unbiased with respect to redshift range $0.1 < z < 0.3$.

- The method can be adapted to any catalog of value-added SFR & $M_\bigstar$ cross-matched with AllWISE.