Machine learning: lessons learnt with the QUBRICS survey

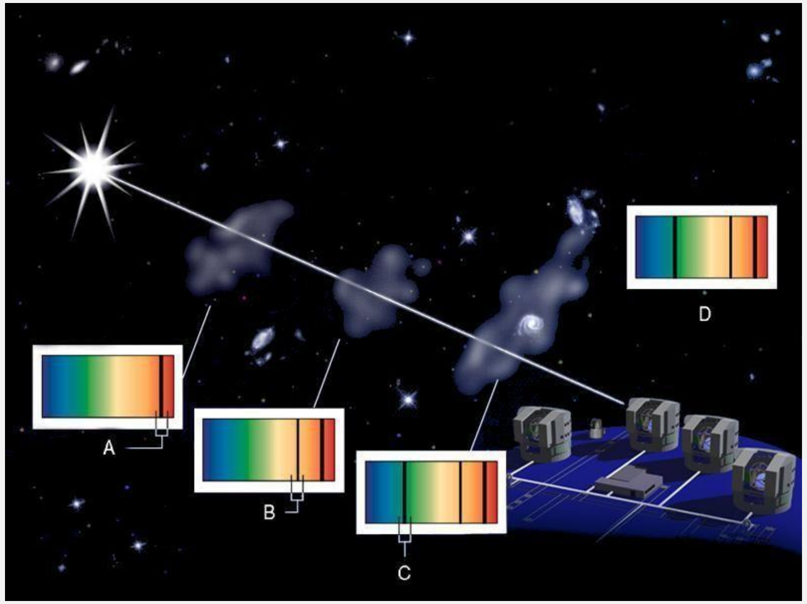
Francesco Guarneri
&
The QUBRICS team

ML-IAP 2021

18-22 October 2021

IAP, Paris and Online

Science with QSOs



- Several open issues can be tackled by exploiting QSO absorption lines:
  - primordial deuterium abundances
  - metal content of the IGM
  - variation of fundamental constants
  - epoch and responsible for reionizaiton
  - test of general relativity
- Light from QSOs is selectively absorbed by the interposing gas: spectroscopic observations are needed, but state-of-the-art facilities can’t access the best targets
- Future surveys will produce large amount of data: automatic analysis tools benefit from these large datasets and outperform classic techniques (e.g., colour selections)

The QUBRICS survey:

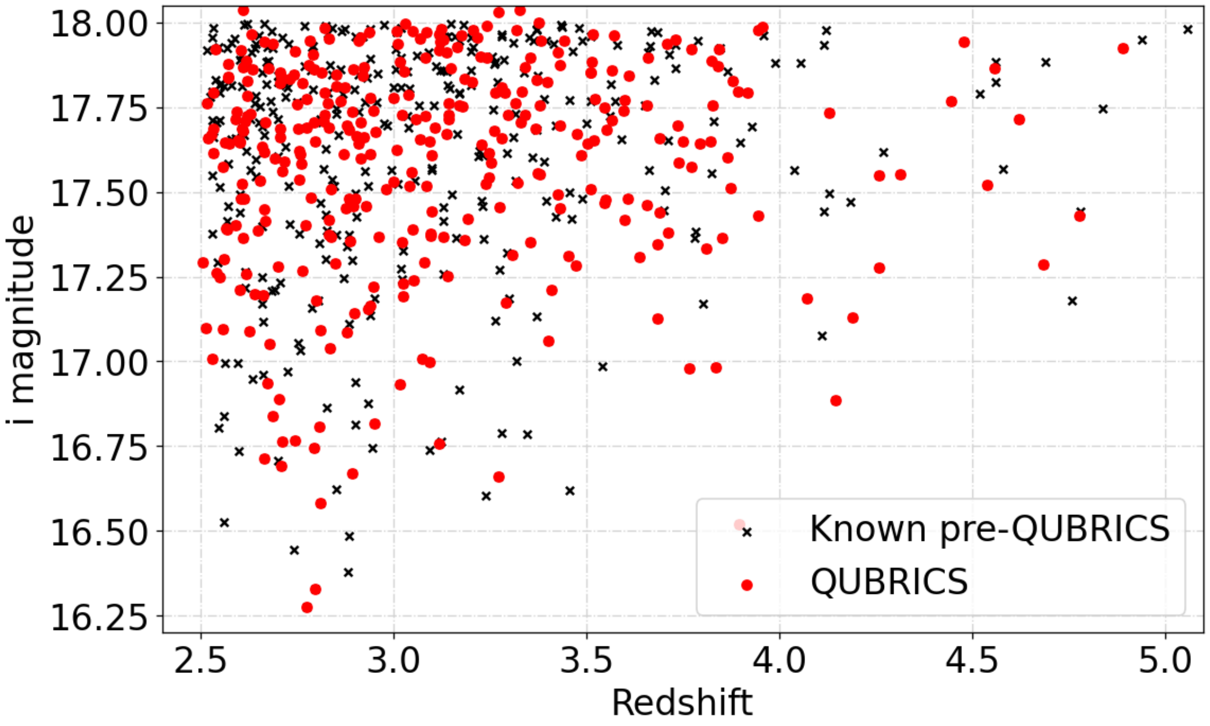
QUasars as BRIght beacons for Cosmology in the South

Main goal:

- Identify bright, high-redshift QSOs using data from publicly available photometric survey:
  - SkyMapper
  - Gaia
  - 2MASS
  - WISE
- Two-fold problem: first identify QSOs, then remove low-redshift objects

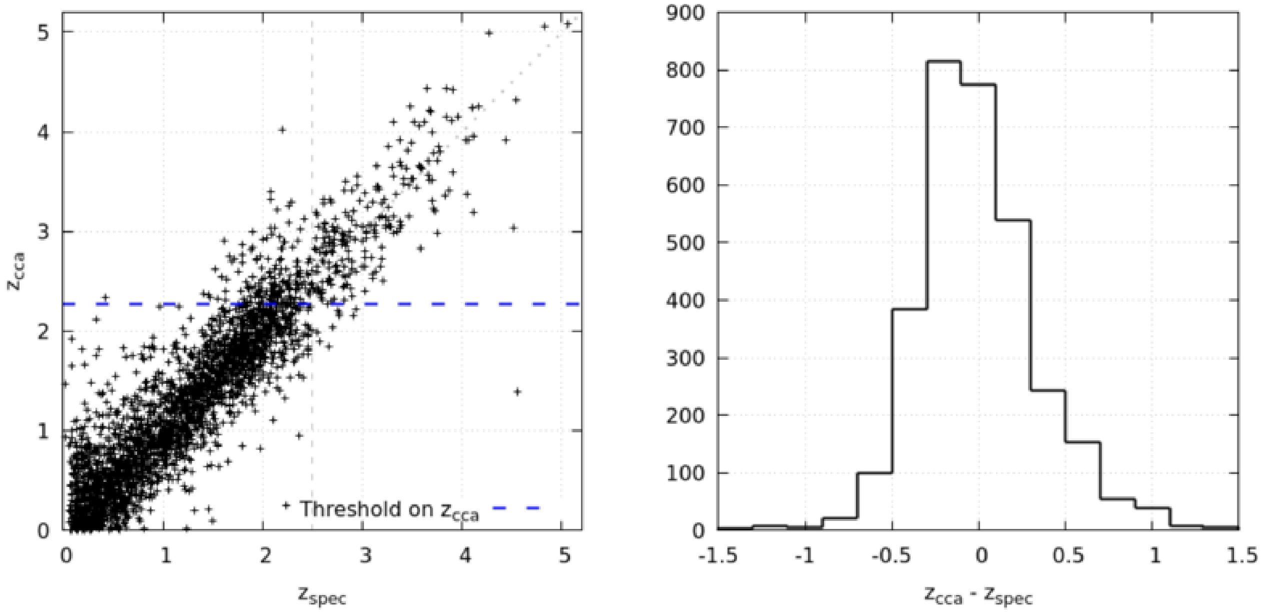
Method:

- Apply ML techniques on photometric datasets:
  - Canonical Correlation Analysis (CCA)
    - Calderone et al 2019 *ApJ* 887 268
    - Boutsia et al 2020 *ApJS* 250 26
  - Probabilistic Random Forest (PRF)
    - Guarneri et al 2021 MNRAS 506 2
- Spectroscopic follow-up to confirm the nature of high-redshift candidates

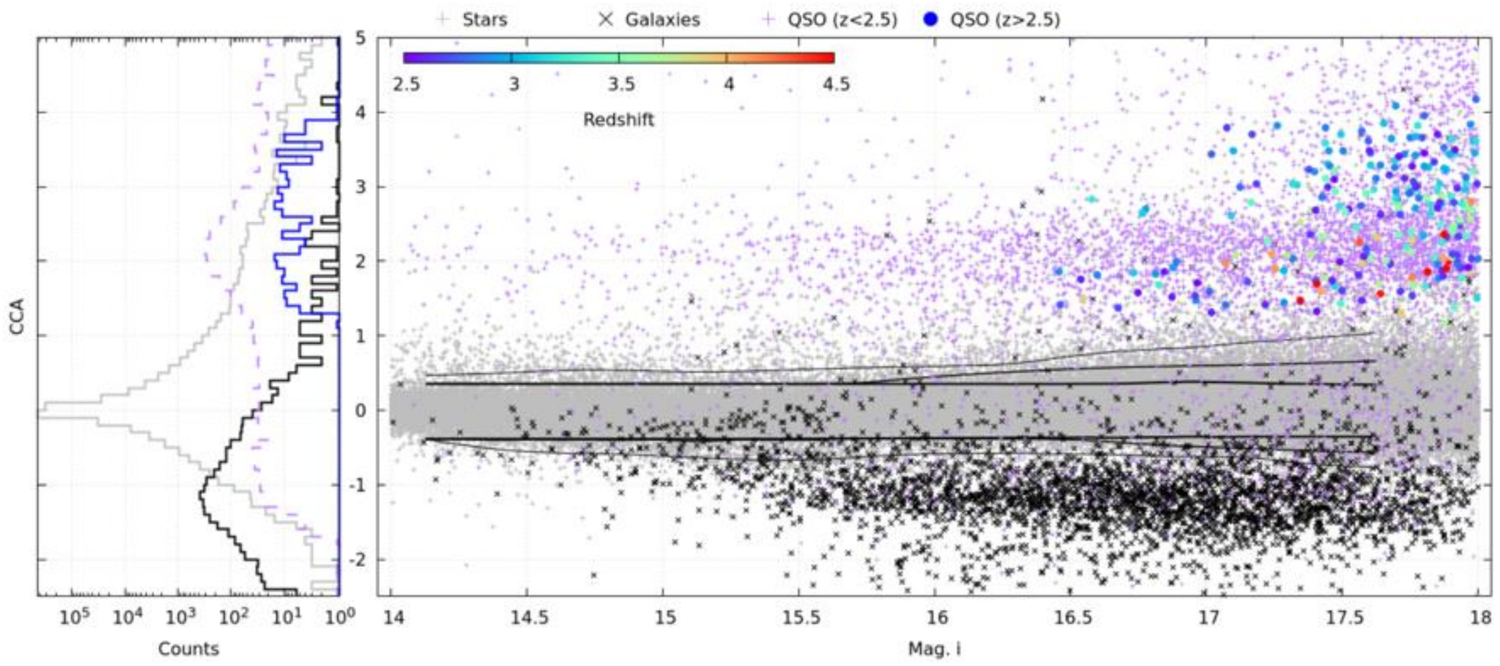


The QUBRICS survey: Canonical Correlation Analysis

(Calderone et al. 2019)

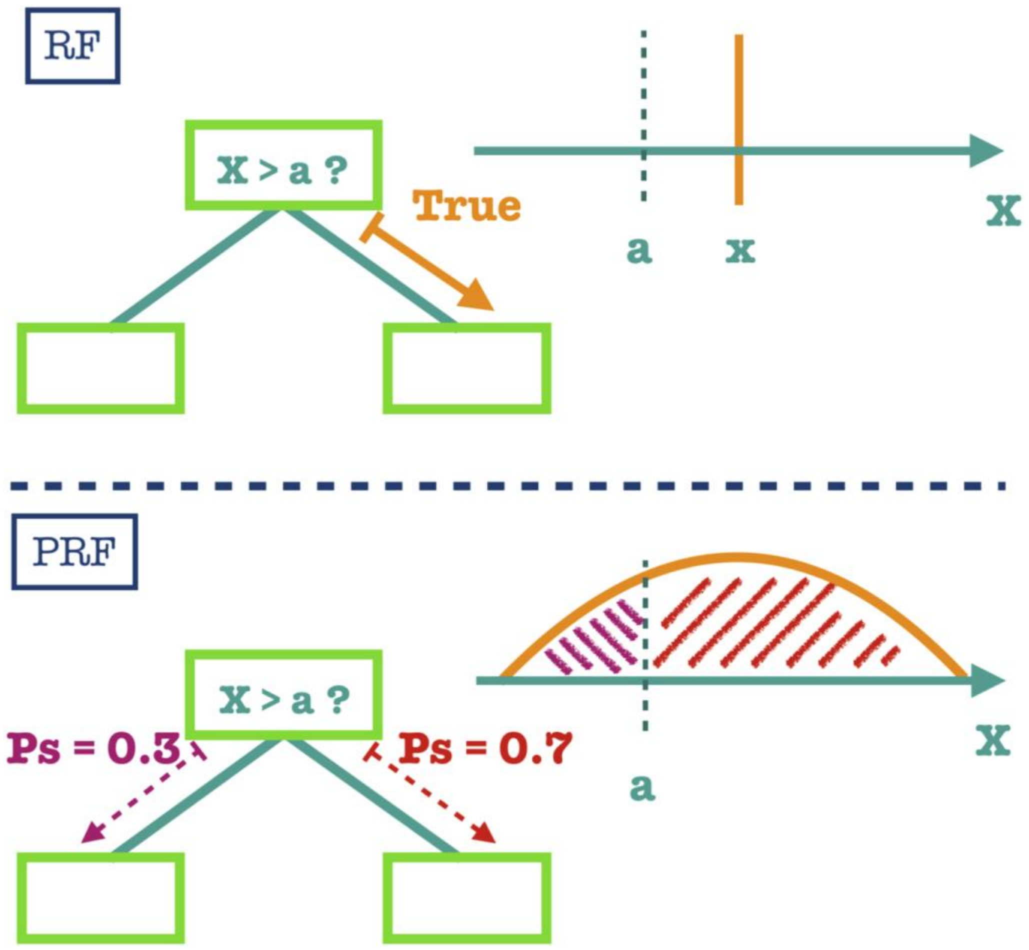


- High dimension selection process based on linear combination of colours
- Used for classification and regression
- Measurement uncertainties are not included in the model, and missing data can't be dealt with



The QUBRICS survey: Probabilistic Random Forest (Reis et al. 2019)

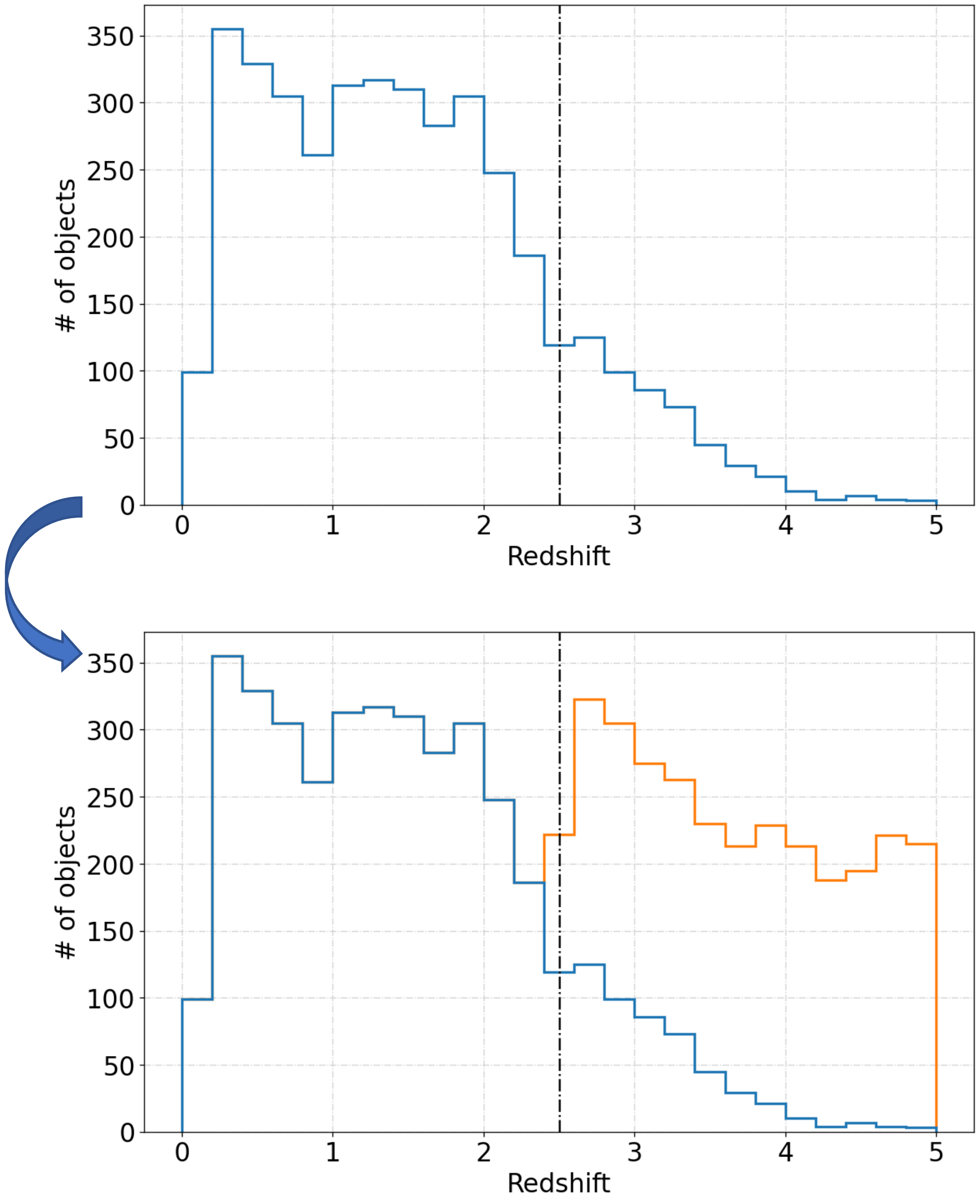
- Generalization of the original Random Forest (RF) to account for measurement uncertainties
- In the PRF each feature is a probability distribution function: this improves performances and considers errors as variance of the distribution
- Naturally handles missing data!



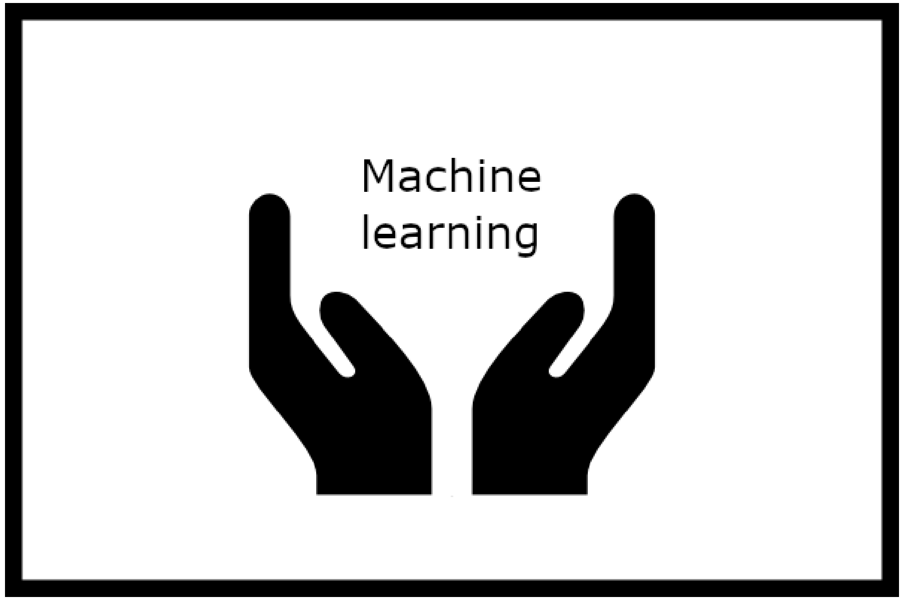
Reis et al. 2019 - arxiv:1811.05994

Good training produces good predictions

- Very few high redshift QSOs with respect to those at low and intermediate: training dataset is unbalanced
- Currently two possible solutions:
  - over/under-sampling techniques
  - synthetic data generation
- Simple oversampling strategy: draw multiple copies of objects in the minority class



Some care is required!

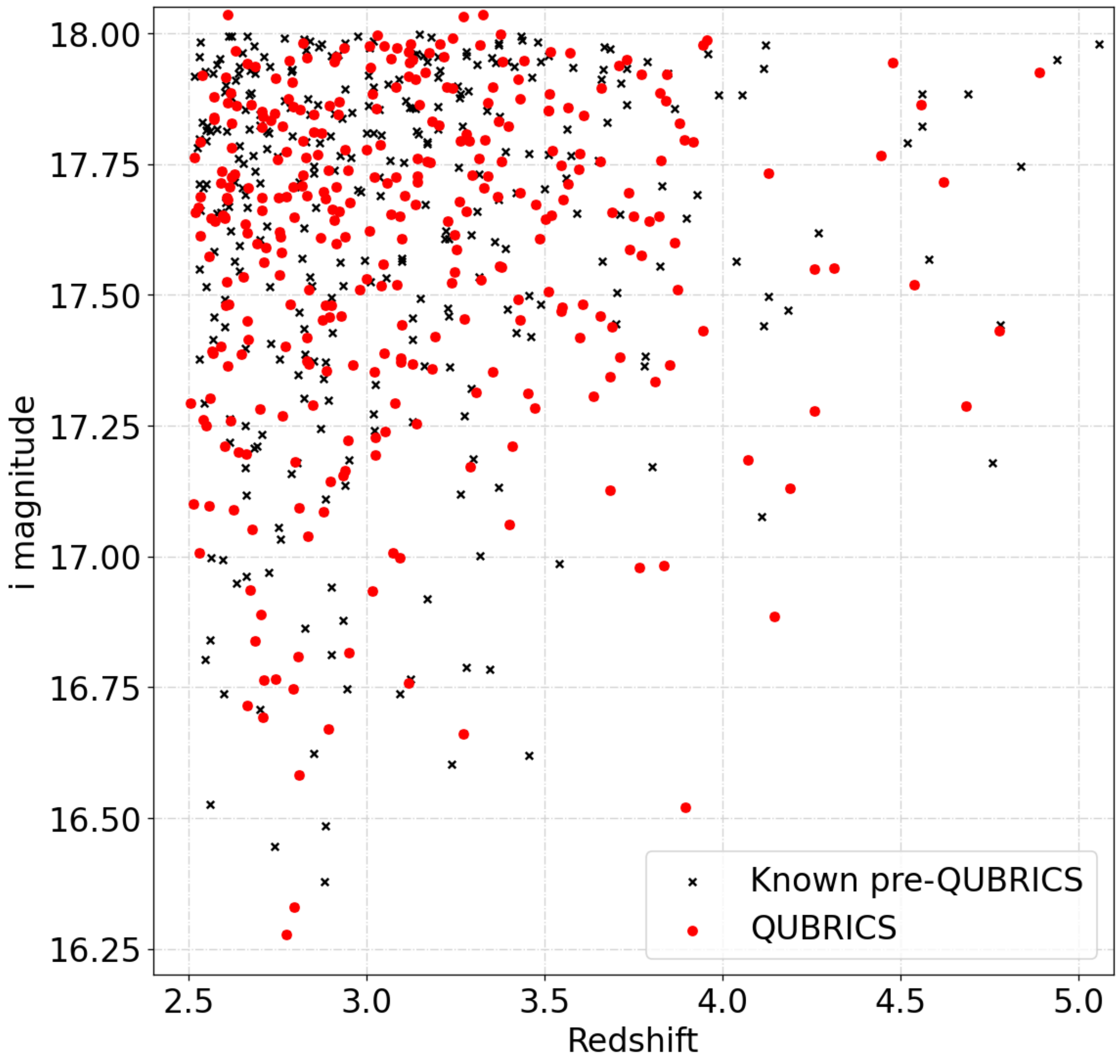


Working on QUBRICS has highlighted some peculiarities of machine learning:

- ML should not be treated as a “black box”: trying to understand the selection method is beneficial
- Results to good to be true need some attention:
    - Different models have unique strength and weaknesses: comparing gives useful insights
- ML complements well classic techniques (e.g., SED fitting or pre-processing)
    - Combining different techniques requires even more care!
- Physics behind the problem should not be ignored:
    - Feature selection
    - Identification of non-physical results

The current state of QUBRICS

- Good success rate, but there is room for improvements: synthetic data are being tested to improve performances
- Main contaminants are low redshift QSOs: galaxies and stars are reliably removed from the candidate sample



Conclusions and future perspectives

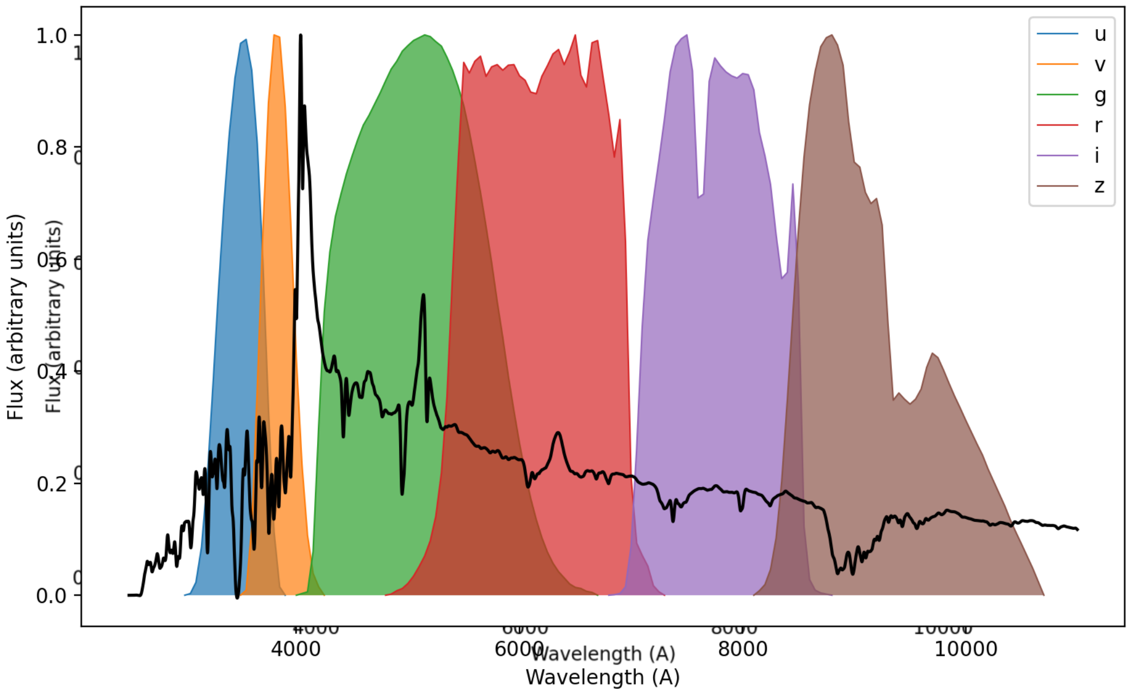
What we've learnt so far…

- Machine learning enables efficient and reliable selection of QSO targets
- Appropriate training sets are crucial: significantly improves performances
- Machine learning should be complemented with previous knowledge, and not used as black box

Moving forward…

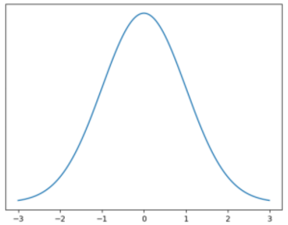
- Improve synthetic data generation:
  - Better modelling of synthetic spectra
  - More selection techniques (e.g., XGBoost)
- New datasets, aiming at higher redshift:
  - Pan-STARRS
  - DES

Synthetic data: a possible solution?



Synthetic magnitude from synthetic spectra

+ Easy to generate in large quantities
+ Can be tailored to a specific class of objects
- Need proper calibration
- Difficult to reproduce some QSO properties or events unrelated to QSO physics (e.g., variability or bad weather)



u_psf Float32	v_psf Float32	g_psf Float32	r_psf Float32	i_psf Float32	z_psf Float32
16.0683	15.9957	15.6528	15.6666	15.7353	15.8188
16.3609	16.2027	15.9034	15.7321	15.3772	15.7216
16.2037	16.1539	15.7506	15.6444	15.2661	15.4215
16.2325	16.091	15.967	15.8516	15.7475	15.8026

PRF – QSO Selection
(Guarneri et al. 2021)