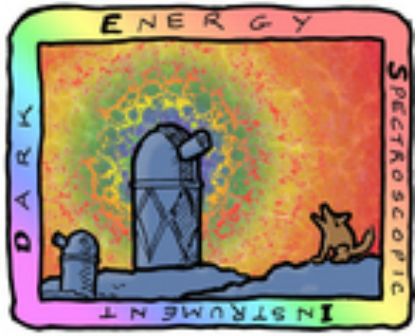


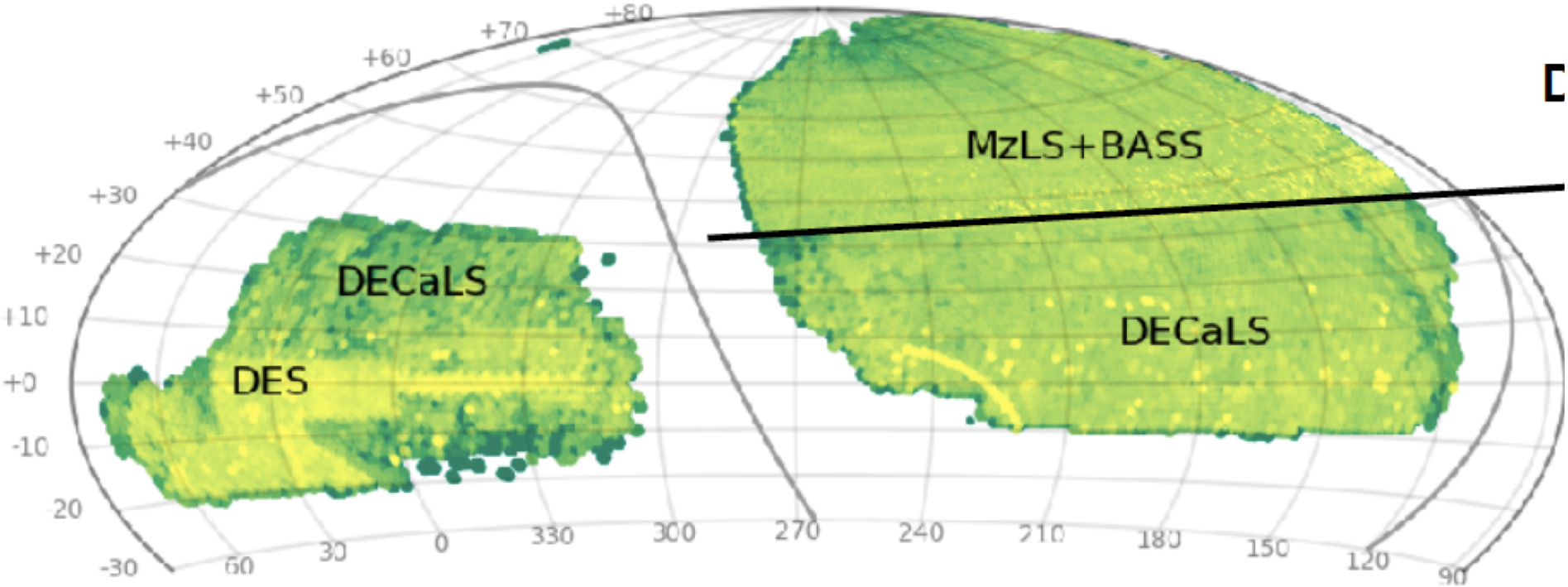
Target Selection of QSO in DESI with Random Forest

E. Burtin, C.-A. Claveau,
N. Palanque-Delabrouille, Ch.Yèche

AI telecon, November 05 ,2018

Context

- The Target Selection is based on optical imaging survey developed specially for DESI
- Three optical bands (grz) + two NIR bands (W1,W2)



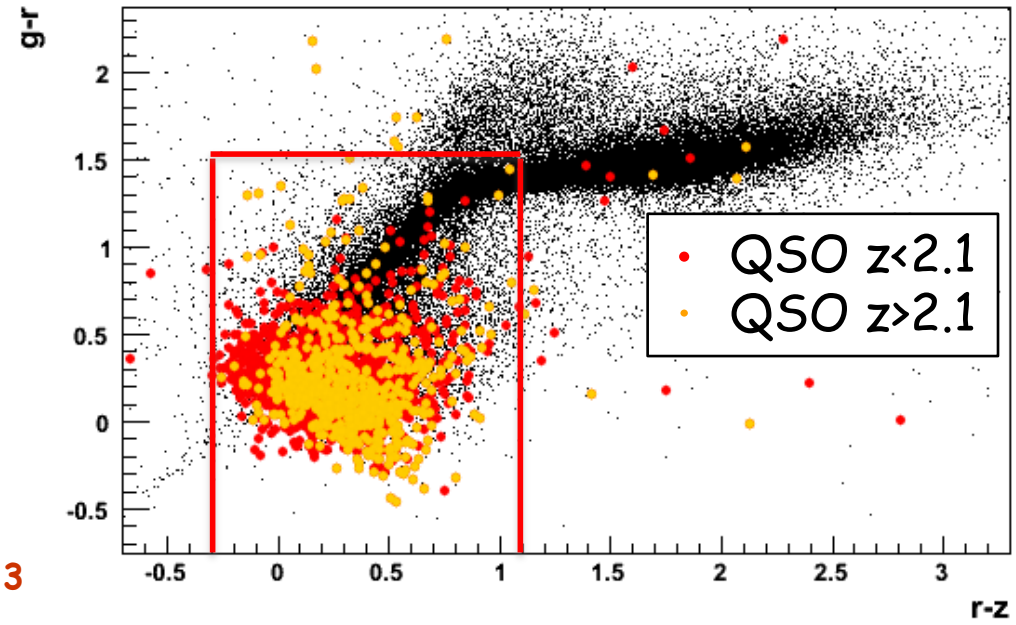
- QSO selection requires PSF objects
- Stars – QSOs confusion, specially for high-z QSOs (z>2.1)
- Density: 260 deg^{-2} targets

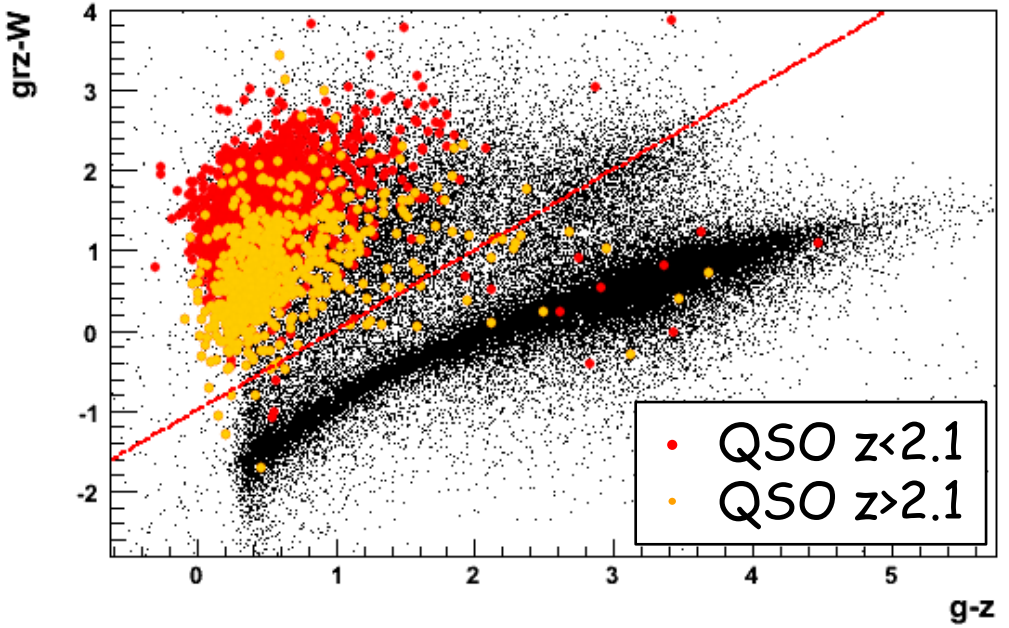
Color cut selection with DECaLS

grz flux = (gflux + 0.8*rflux + 0.5*zflux) / 2.4
W flux = 0.75*w1flux + 0.25*w2flux

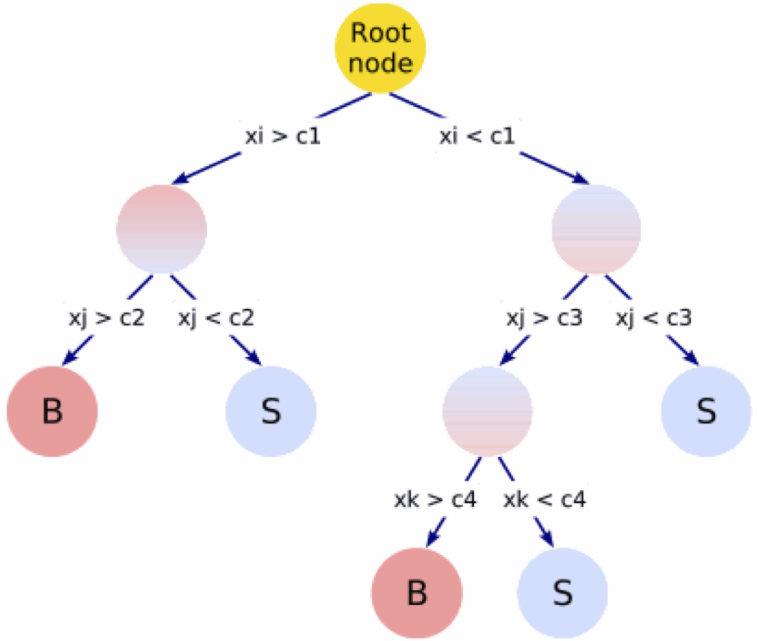
Depth: W>0 && grz>0 && 17<grz && r<22.7
Morphology: PSF && dchisq(PSF)-dchisq(SIM)>40
grz box: (g-r)<1.5 && r-z<1.1 && r-z>-.3
grz-W box: grz-W>g-z-1.0
W1/W2: snrW2>1 && snrW1>4 && W1-W2>-0.4
Main sequence rejection: ! (g-r>0.20&&abs((g-r)-1.5*(r-z)-.175)<.100 && (W1-W2<0.3 || grz-W>3))

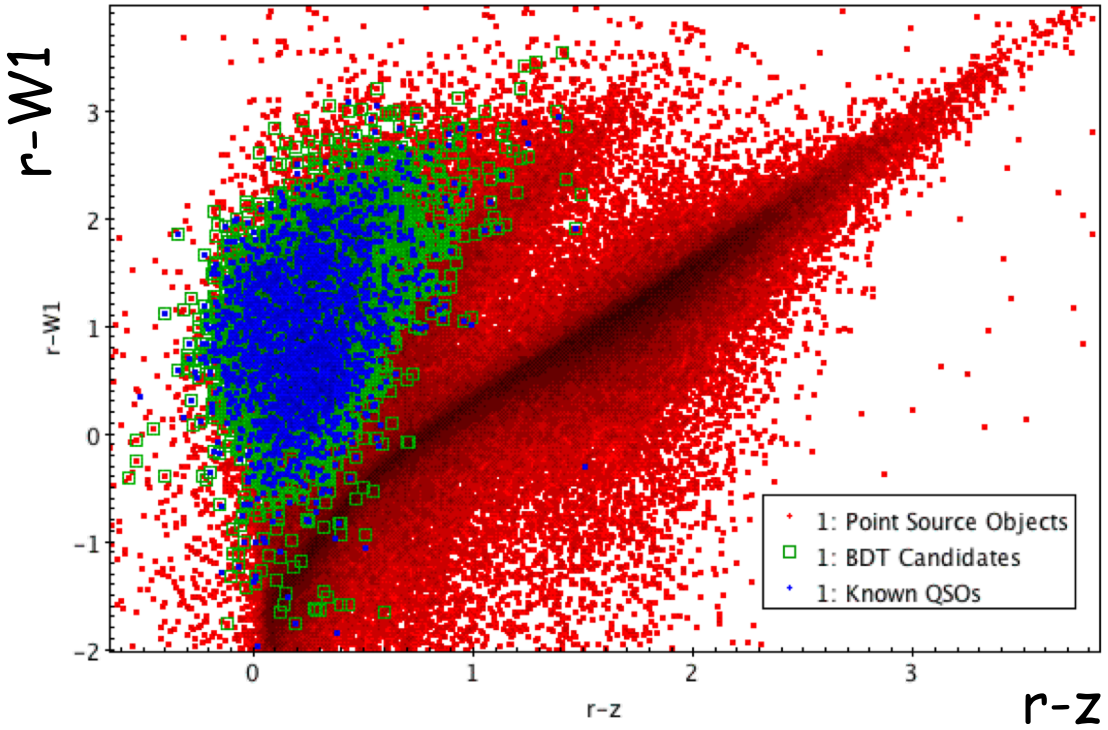
What we want to avoid!!!





Machine Learning Approach





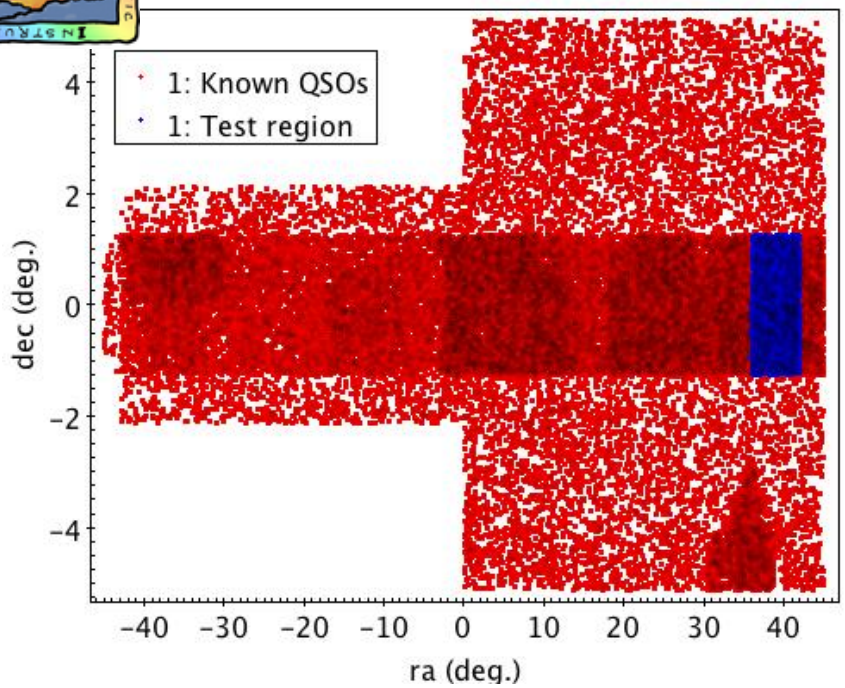
Random forest algorithm

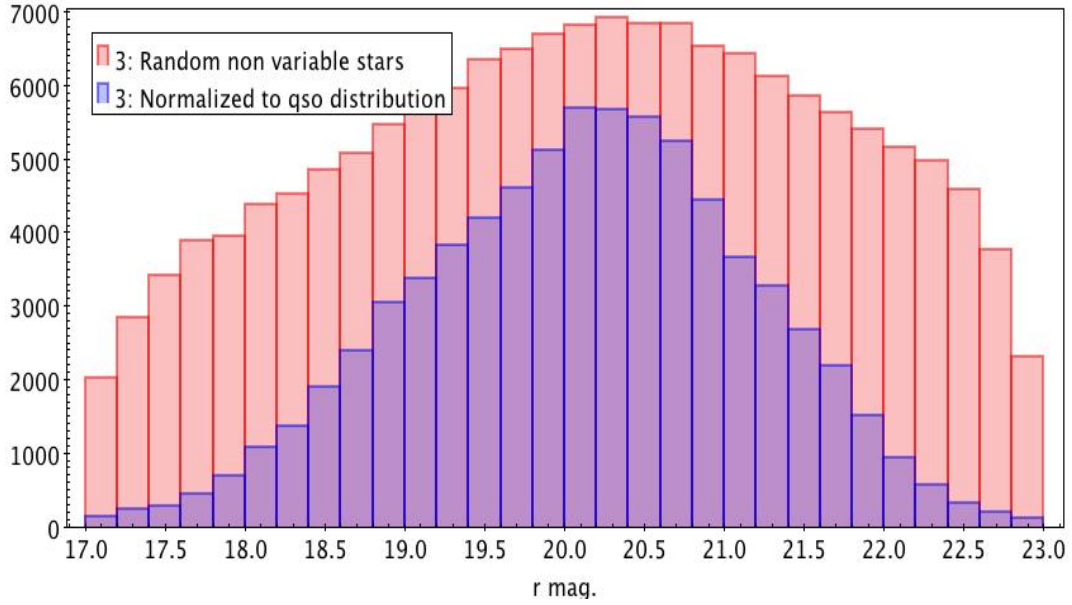
- Classification with **scikit-learn**
- Random: Subset of variables and subsample of the training sample
- Forest: Sum of the decision trees
- Training over all the known QSOs with DECaLS DR3
- Tested in stripe 82

Variables (inputs)

- All possible colors used with DECaLS (g,r,z) and WISE (W1,W2)
- a mag. reference: r

Proof of Principle with DR3

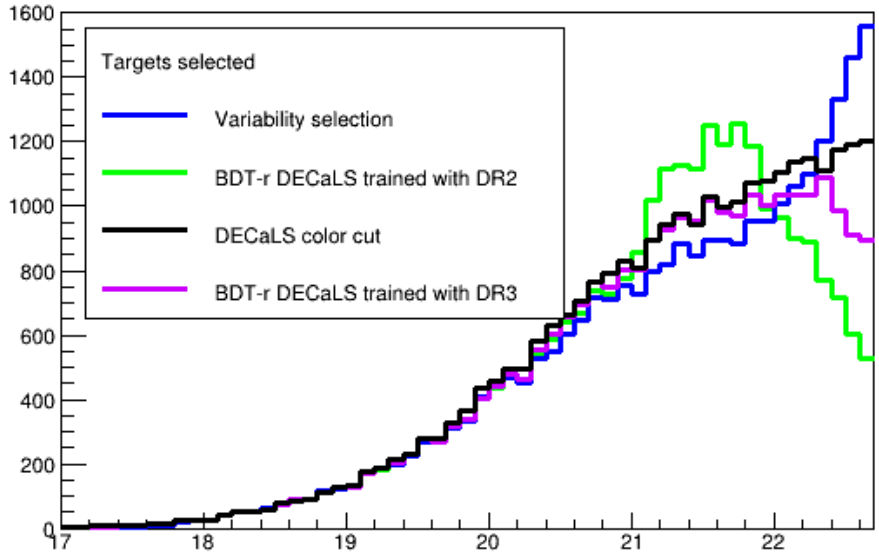


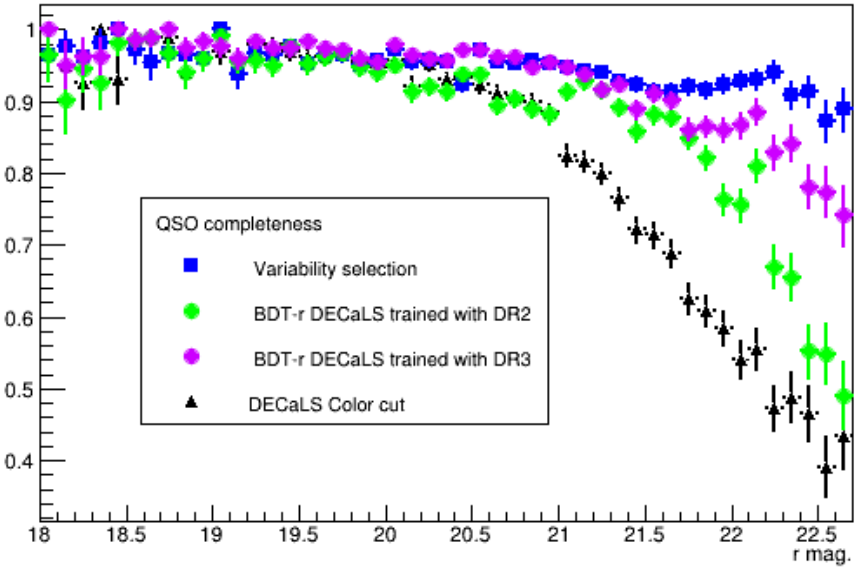


- DR3 DECaLS QSO sample:
  - Deep region of DES (fat stripe 82) : **~35000 objects**
  - Bright objects ($\sigma(r)<0.02$): **~45000 objects**
- DR3 DECaLS star sample:
  - From fat stripe 82 (known QSOs and variable objects are removed): **~75000 objects**
- To avoid bias, same r mag. distribution for stars and qsos (difference between our DR2 and DR3 RF)

Results with Stripe 82

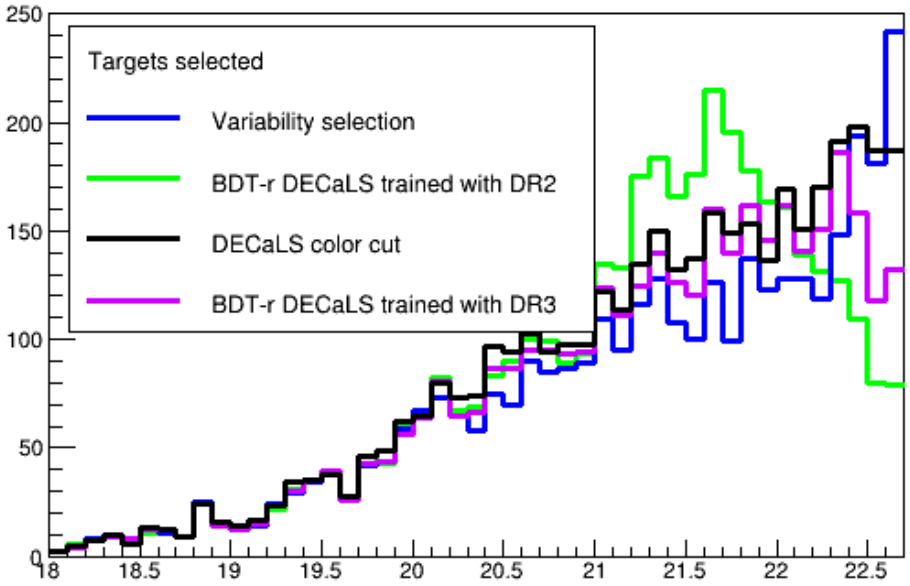
- Test over the stripe 82, ra>0
- Same budget for the four selections: ~240 deg^2 objects
- Comparison of the Color Cut selection (black triangles) and the RF (purple circles)
- Gain in completeness for faint objects r>22

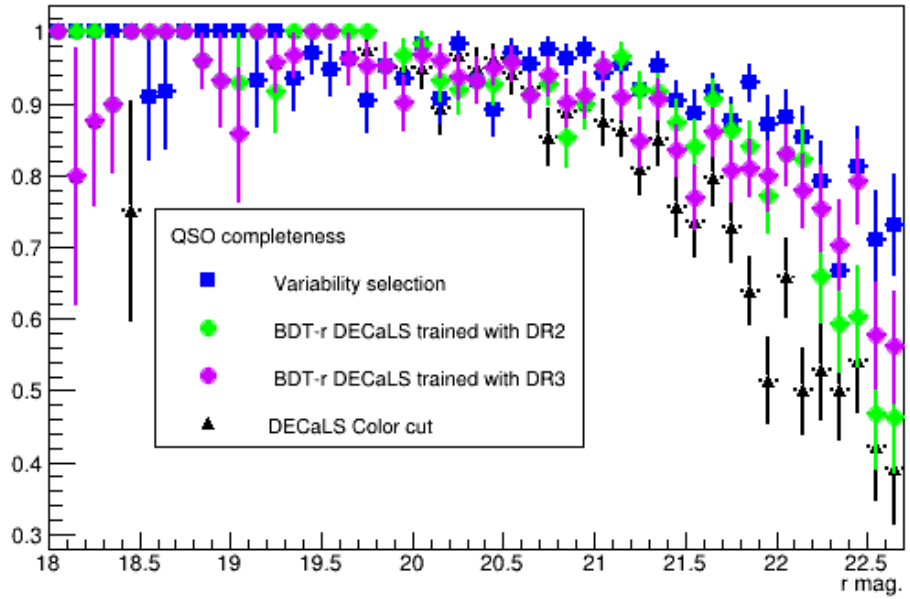




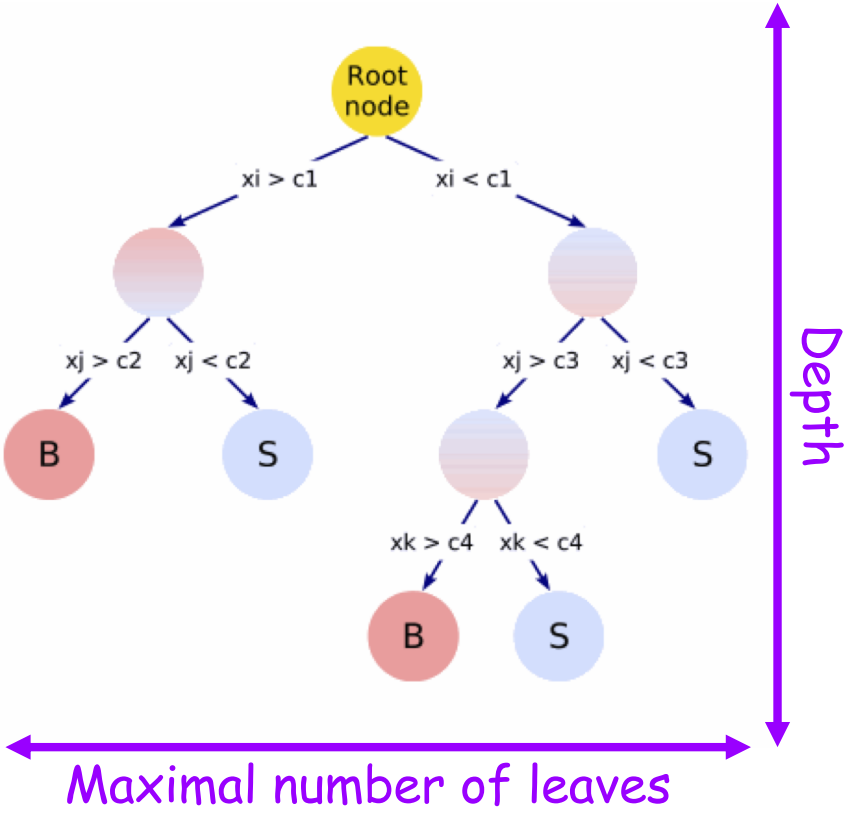
Results with test region in Stripe 82

- Test over the test region of stripe 82, 36<ra<42
- Same budget : ~265 deg^{-2}
- Comparison of the Color Cut selection (black triangles) and the RF (purple circles)
- Same trends: Gain in completeness for faint objects r>22…





Optimization of the RF Training



Parameters to avoid overtraining and oversized RF

- Number of trees in the forest
- The maximum depth of the tree
- The maximum number of leaves

Scikit-learn implementation

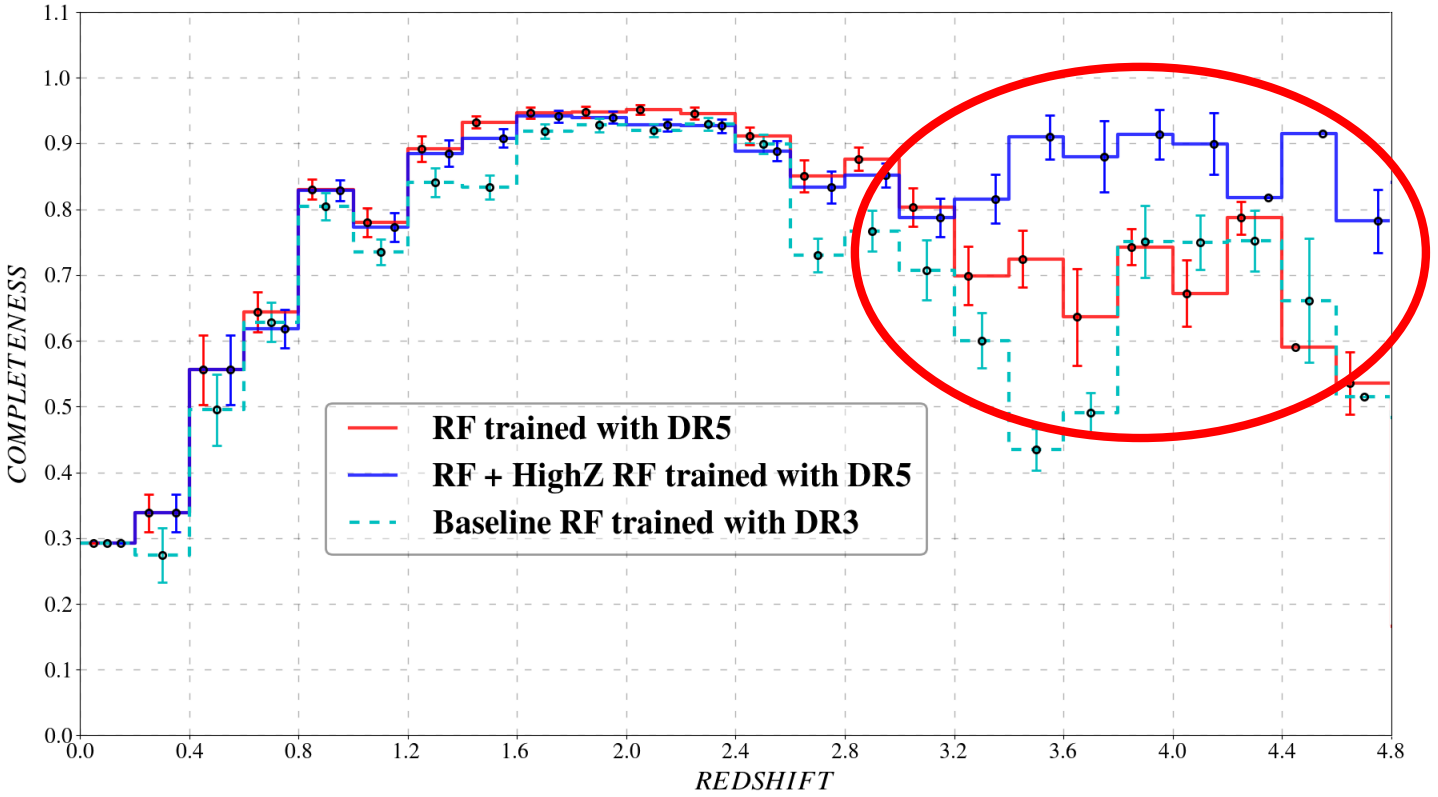
- **RandomForestClassifier**
- n_estimators
- max_depth
- max_leaf_nodes
- As a result the last node (leaf) is not pure in Signal/Background

Benefits

- Reduce by a factor 3 the size of the RF
- Faster computation of the probability
- Better control of the overtraining

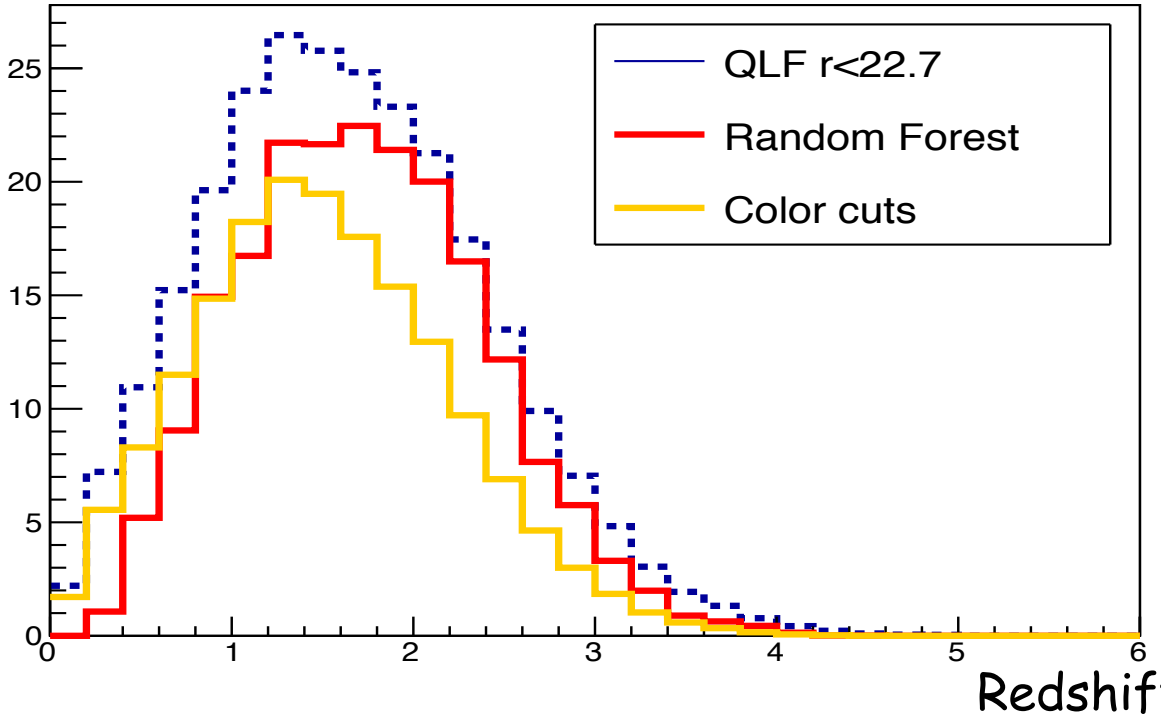
Latest developments

- Two different RF:
  - RF trained for all redshifts, 245 deg^{-2}
  - RF trained for z> ~3.5, 15 deg^{-2}
- Different training for DR3, DR5, DR7
  - Very stable results (probability cut slightly tuned)



Gain in completeness at high-z

Conclusions



- Very simple to implement in scikit-learn
- Own persistency of the RF in *desitarget* code
- Very stable as a function of the various imaging releases (DR5, DR7…)
- Easy to tune to achieve the target budget
- Much efficient that the color cut approach specially for high-z QSOs