OU-PHZ: overview of the requirements, algorithms and current performance

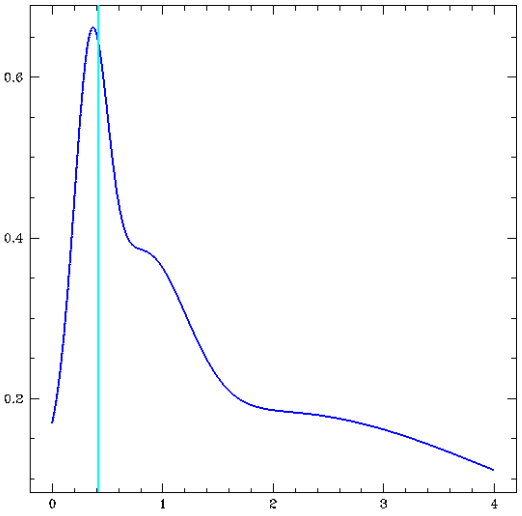
Stéphane Paltani
University of Geneva

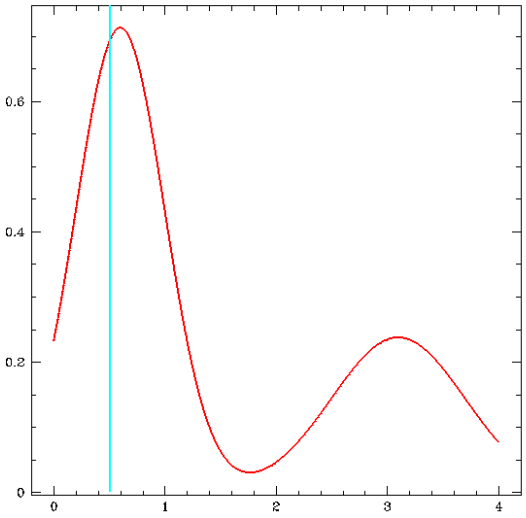
Euclid Photometric Calibration Workshop – ESAC

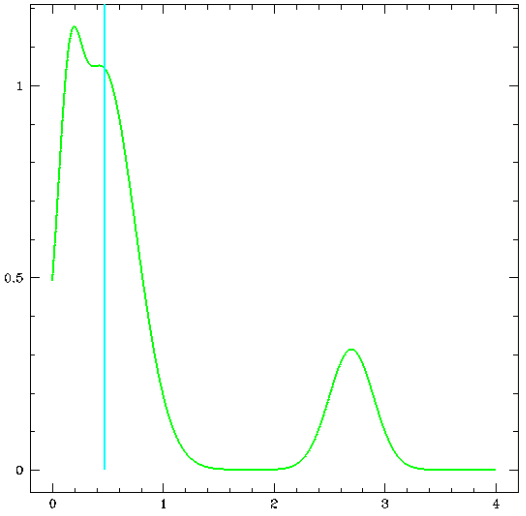
Photo-z Requirements

- Euclid is a cosmology mission. ~~Photo-z~~ Redshift requirements are defined in the Red Book for weak-lensing tomography
  - $\sigma_z \leq 0.05\ (1+z)$ (goal: 0.03) over $0.2 \leq z \leq 2.0$
  - Catastrophic failures < 10% (goal: 5%)
  - Mean redshift in each of the 10 tomographic bins known to a level $\sigma(\langle z \rangle) \leq 0.002\ (1+z)$ → Peter's talk
  - But we are requested to provide PDFs

Photo-z Requirements with PDF







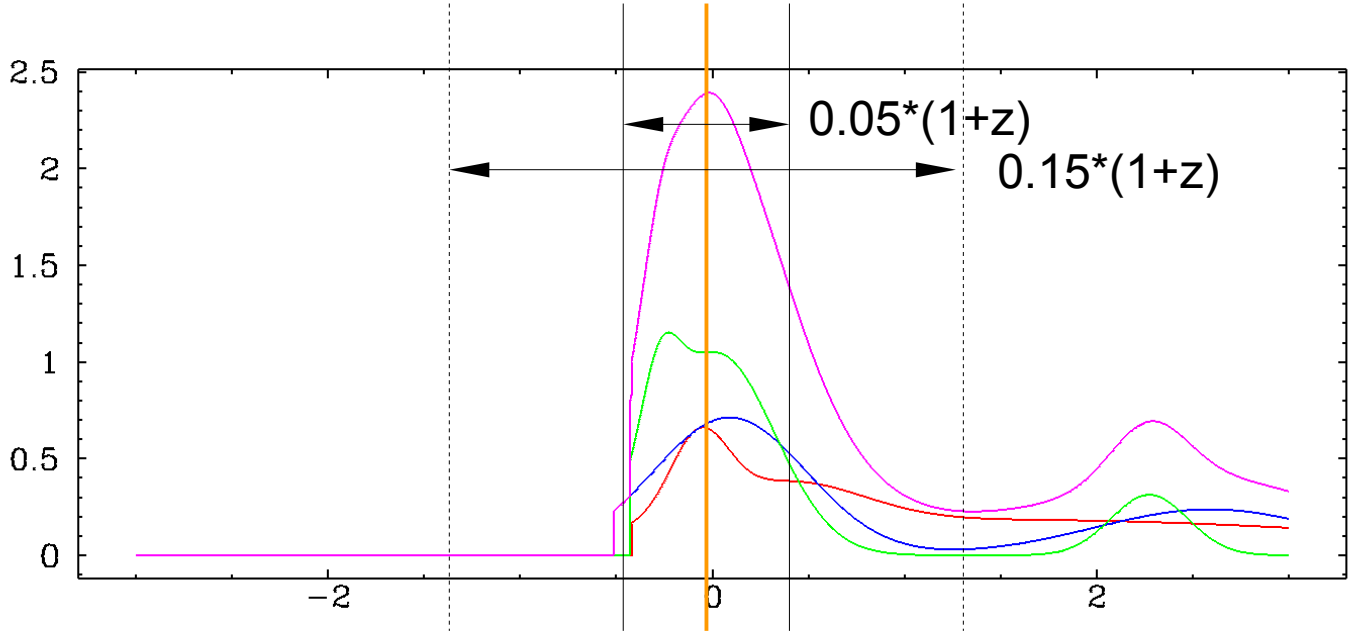
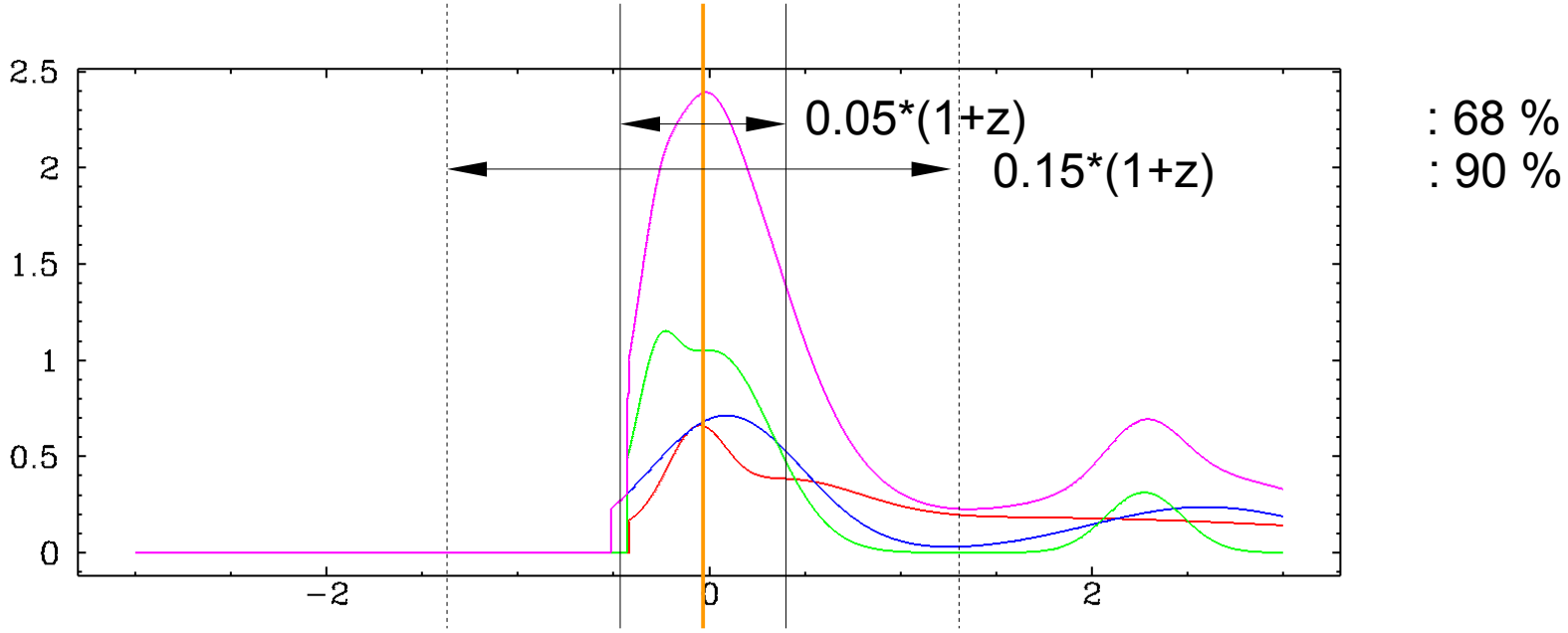


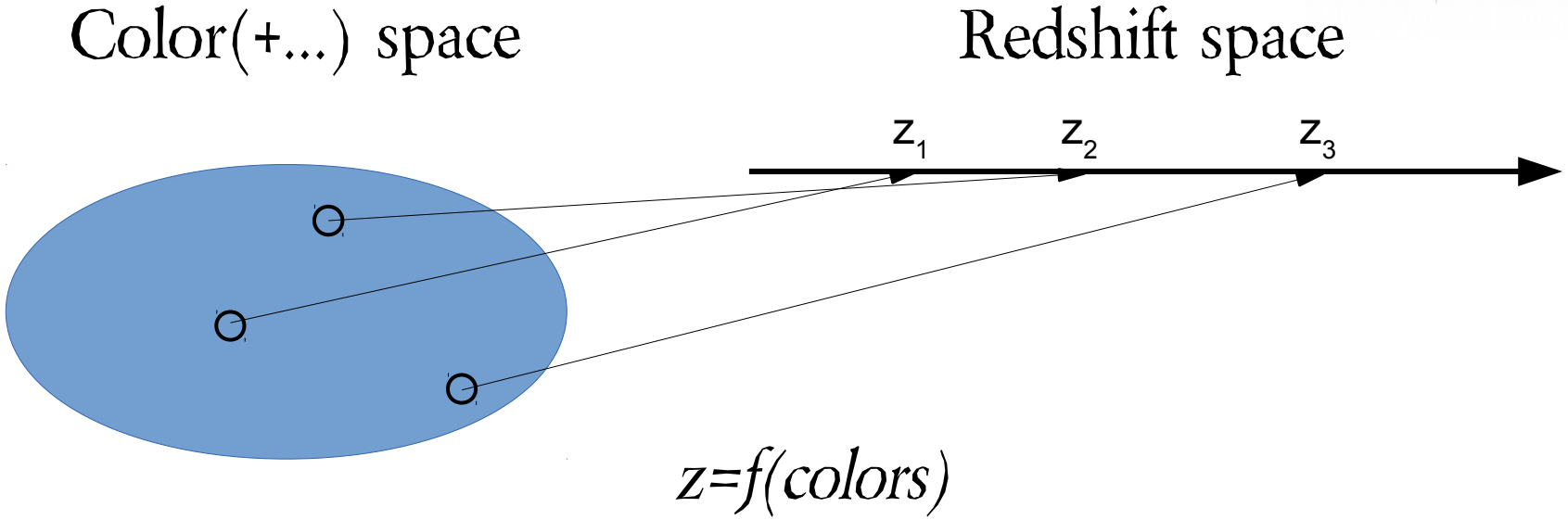
Photo-z Requirements with PDF

R-PHZ-PRD-P-010 PHZ performance

The shape of the stacked PDF for each sub-set of galaxies in the range 0.2<z<2.0 (TBD) used in the weak lensing analysis shall be such that: the integrated PDF beyond 3 sigma of the mode is <10% of the total integrated PDF over each sub-set, and the r.m.s of the pdf calculated within 3 sigma of the mode is sigma(z)< 0.05(1+z).



Photometric-Redshift Algorithms



Mapping *f* can be constructed based on prior knowledge :

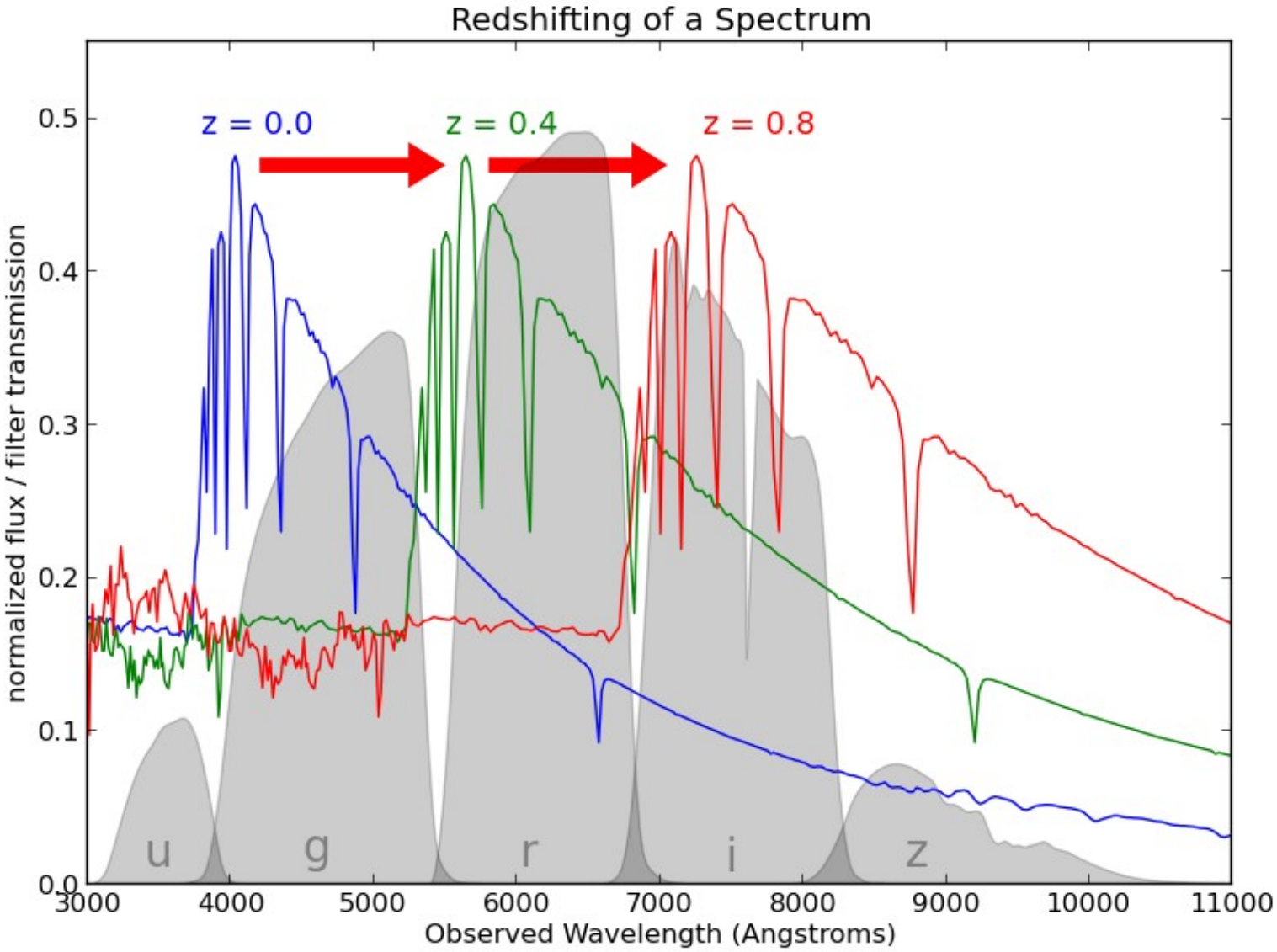
- Template-fitting: Hyper-Z, Le Phare, BpZ, Phosphoros,...

Or it can be discovered:

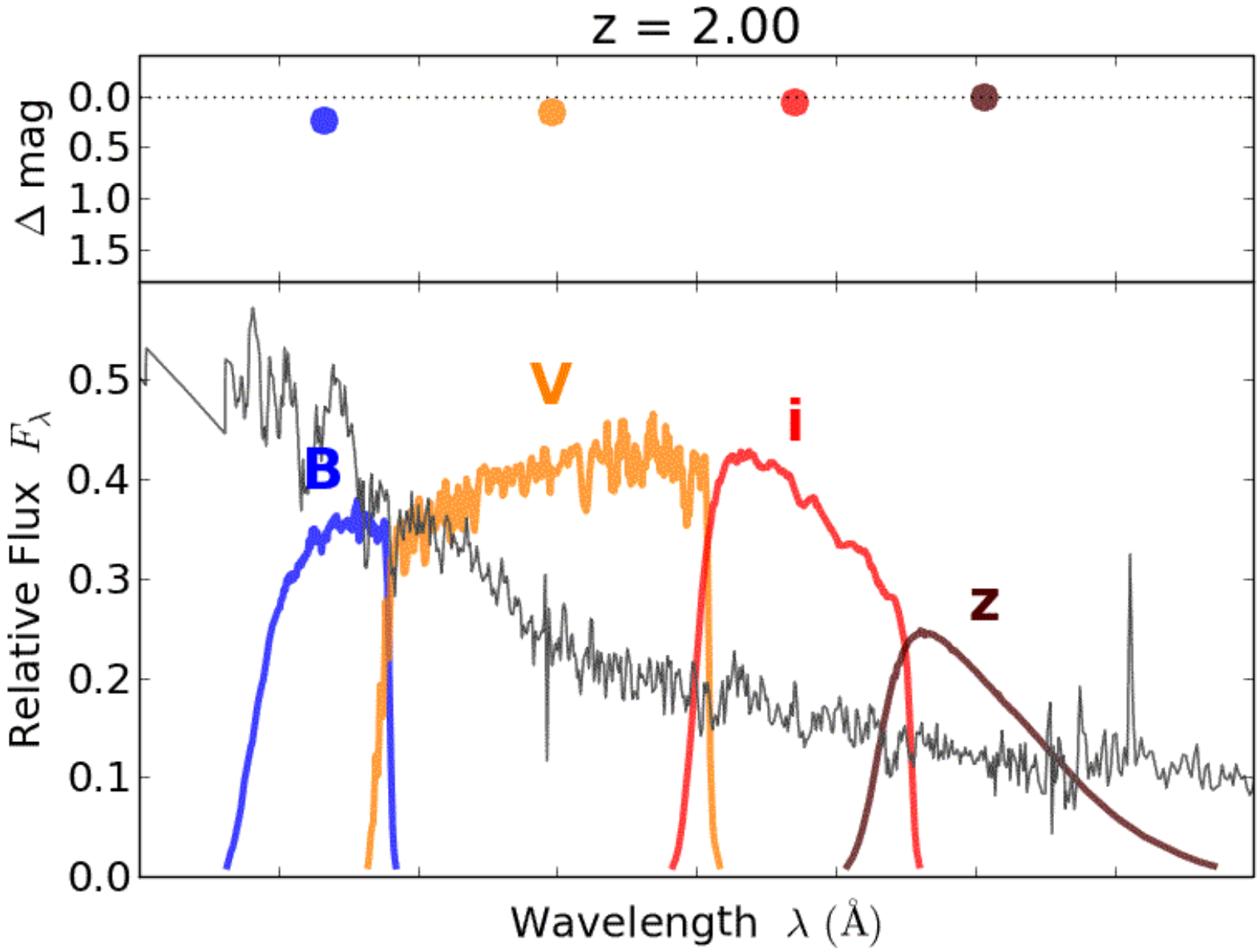
- Machine-learning: Nearest neighbors, Perceptron, Support vector regression, Random Forest, Adaboost, Gaussian Processes, …

Plus some "non-standard" approaches

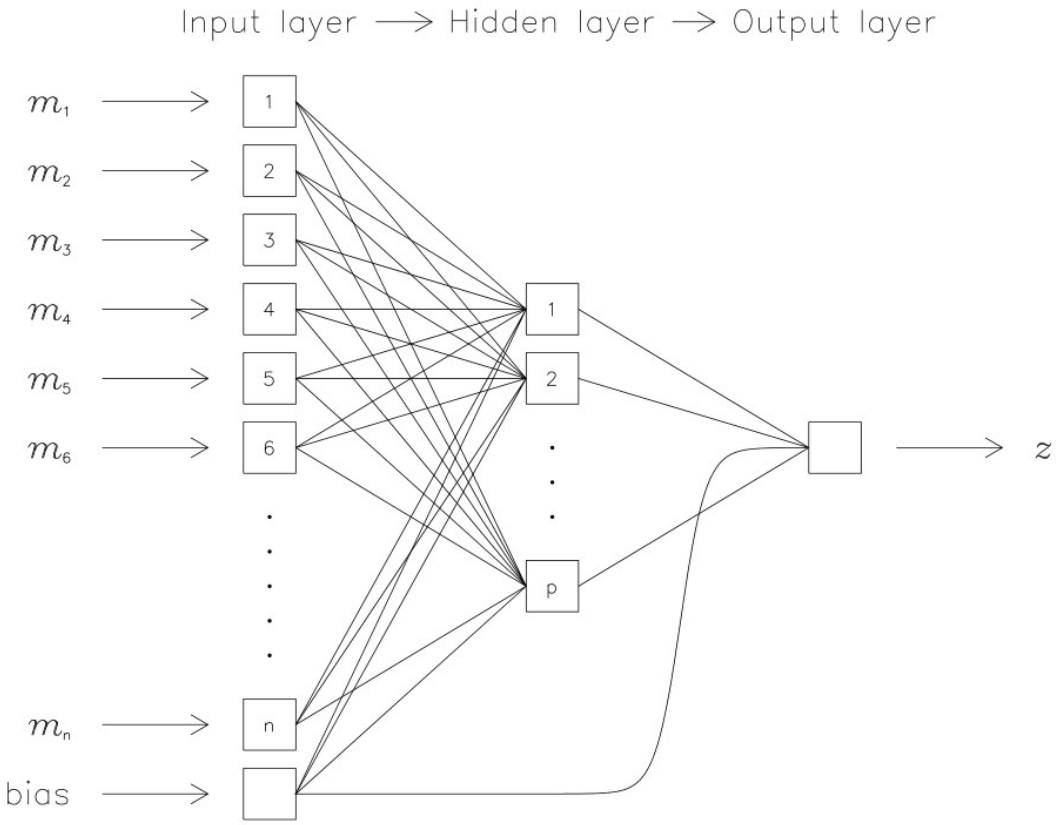
Template-Fitting Algorithms

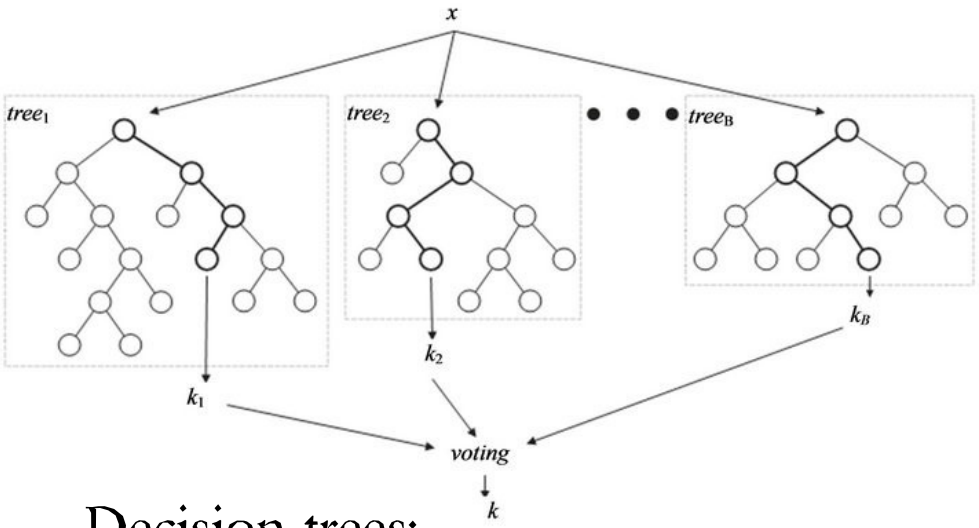


Template-Fitting Algorithms



Machine-Learning



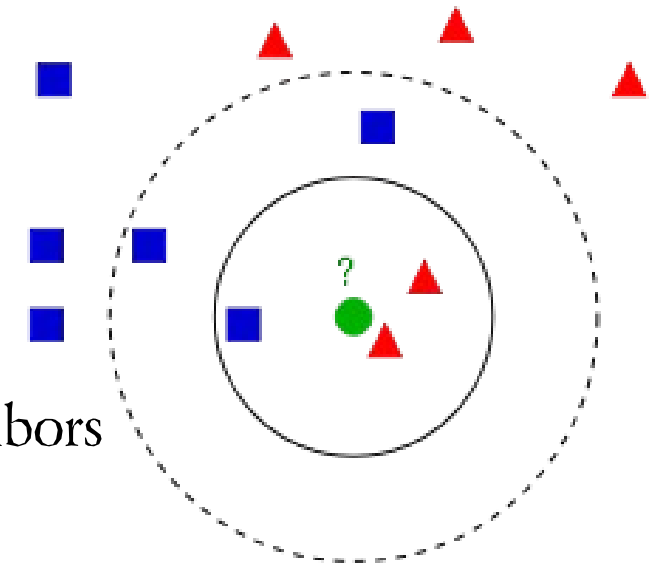


Decision trees:

Random Forest, TPZ, Adaboost, ...

Artificial neural networks:

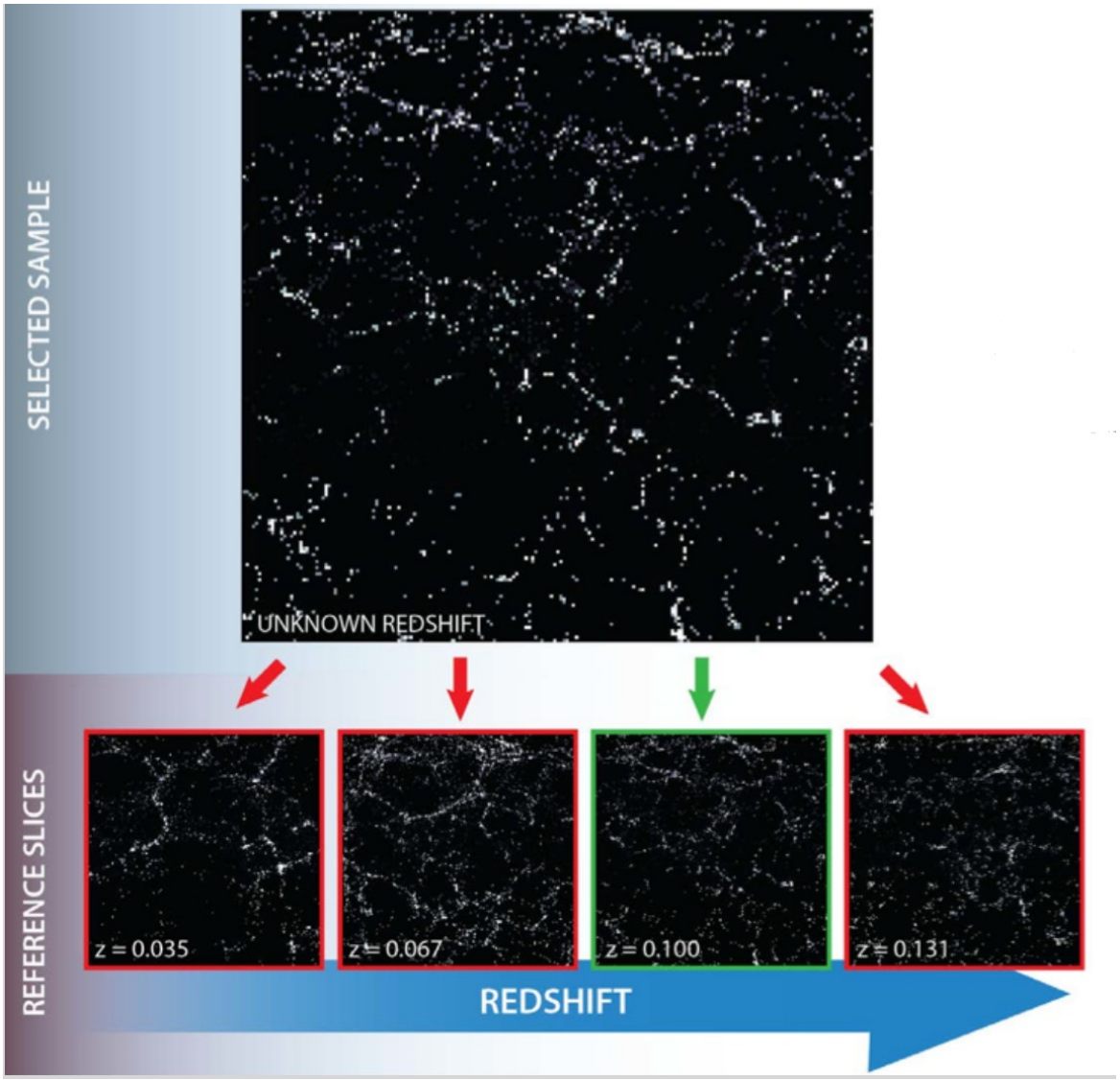
MLPQNA, ANN-z, Skynet, …



k-Nearest Neighbors

ML algorithms rely on a training set (spec-z)

A completely new approach ?



- "Cluster-z": Use the spatial distribution of galaxies to infer their redshift distribution
- Plan agreed with SWG-WL is to use Cluster-z for validation
- SDC-CH started to look into a public code: The-wiZZ (C. Morrison)
- → Vivien

Template-Fitting or Machine-Learning ?

- Overall, it is difficult to clearly select one over the other, so I have provided this decision tree:

- What kind of scientist are you?

  - An astrophysicist: Use TF
  - A cosmologist: Use ML
  - Both/don't know: You have to listen to my talk

Template-Fitting Advantages

- ✔ Based on astrophysical knowledge; the better the knowledge, the better the algorithms
- ✔ Any physical process that is understood can be modeled explicitly (e.g., see Audrey's talk)
- ✔ Constructs naturally a likelihood, and can be turned into a fully Bayesian approach
- ✔ Can cope with infomative priors in a very natural way, e.g. luminosity function, cosmological volume

Template-Fitting Disadvantages

- ✗ Knowledge of the sky is imperfect and incomplete
- ✗ No easy guideline regarding the number of templates, so there is a trade-off between catastrophic outliers (fewer templates) and degeneracies (more templates)
- ✗ Computation intensive, especially if one includes the whole variety of bells and whistles
- ✗ Cannot easily cope with additional features (galaxy shape, etc. ; but is it useful ?)
- ✗ Link between photometry and galaxy properties not clear (e.g., aperture effects)

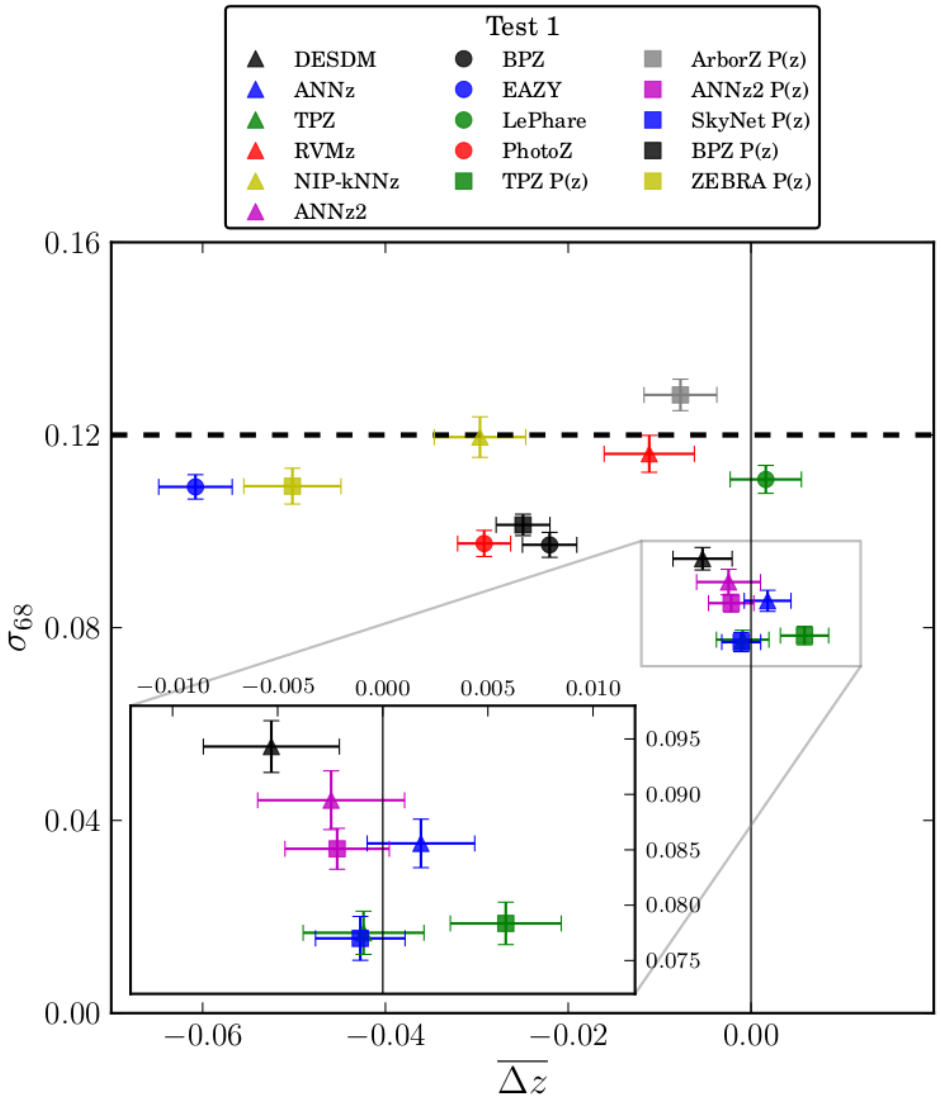
Machine-Learning Advantages

- ✔ A competitive ML algorithm can be written from scratch in 2 hours, and many algorithms can be tested in 10 more minutes
- ✔ No need to understand the astrophysics or to model anything
- ✔ Can easily incorporate additional features; can use simultaneously several types of photometry; good ML algorithms can do it without loss of stability
- ✔ A sound ML algorithm will be optimal where training set is "good"
- ✔ Not very demanding computationally, except some training phases
- ✔ Not linked to galaxy properties, so photometry does not really matter

Machine-Learning Disadvantages

- ✗ The ML algorithm is only as good as the training set
- ✗ A good training set is difficult to build
- ✗ ML algorithms have "hidden priors" in the selection of the training set
- ✗ Many/all algorithms cannot produce naturally a PDF
- ✗ No easy guideline regarding the model complexity ; it can be tested, but only globally, so it is prone to overfitting and underfitting, at least locally
- ✗ Extrapolations might occur

But is ML better ?



DES (Sánchez et al. 2014)

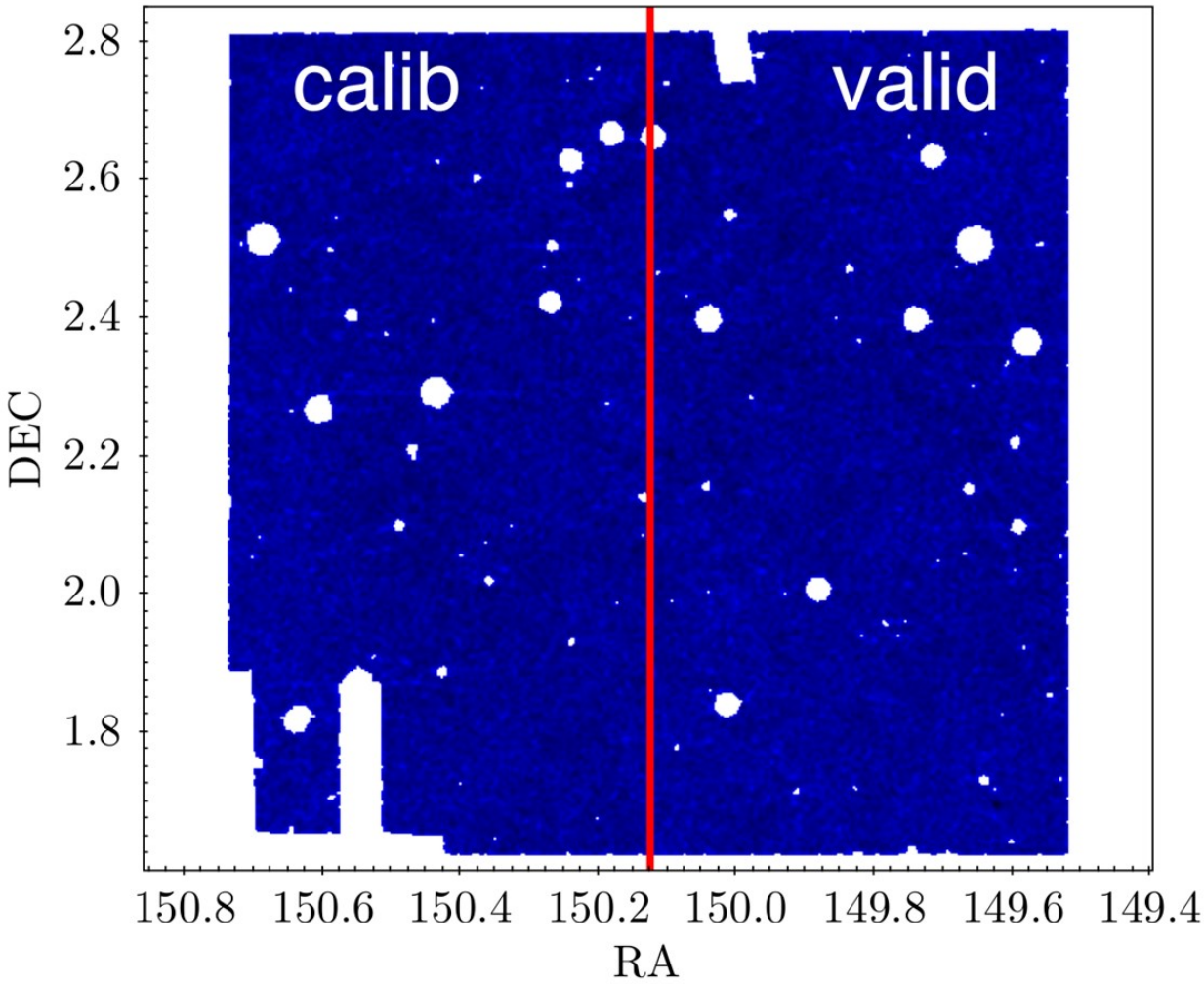
It is often perceived that ML algorithms are superior. Maybe…

But :

- There is a "sprinter" effect
- One needs to consider the fact that the training set and the test set come from the same population (at least use the weighting of Lima et al. 2008)
- All developers of algorithms who are co-authors of the DES paper develop ML algorithms…

Data Challenge 2

- DES and Ultra-VISTA data on the COSMOS field
  - Processed through OU-EXT+OU-MER
  - Simulates depth of the Euclid survey
- Significant set of spec-z, 29'964 validation spec-z's

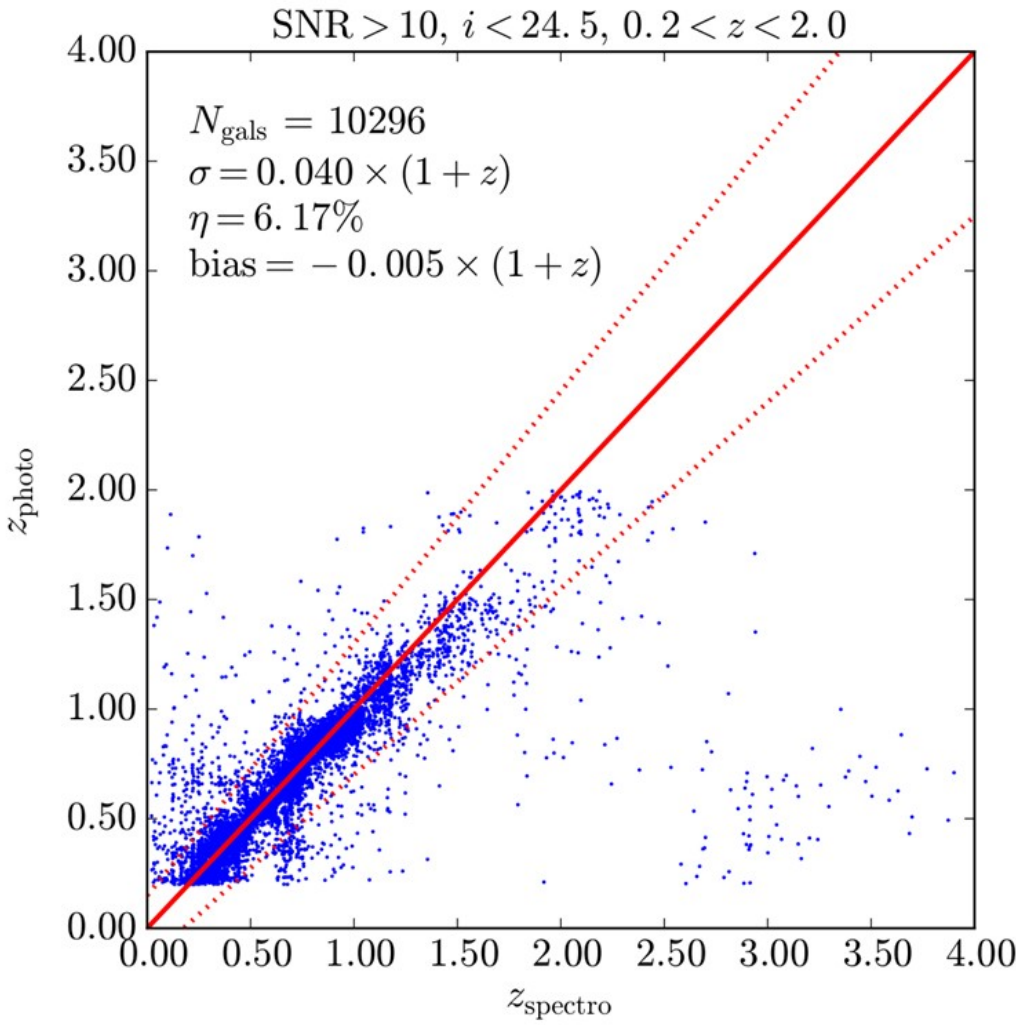


Data Challenge 2 Results

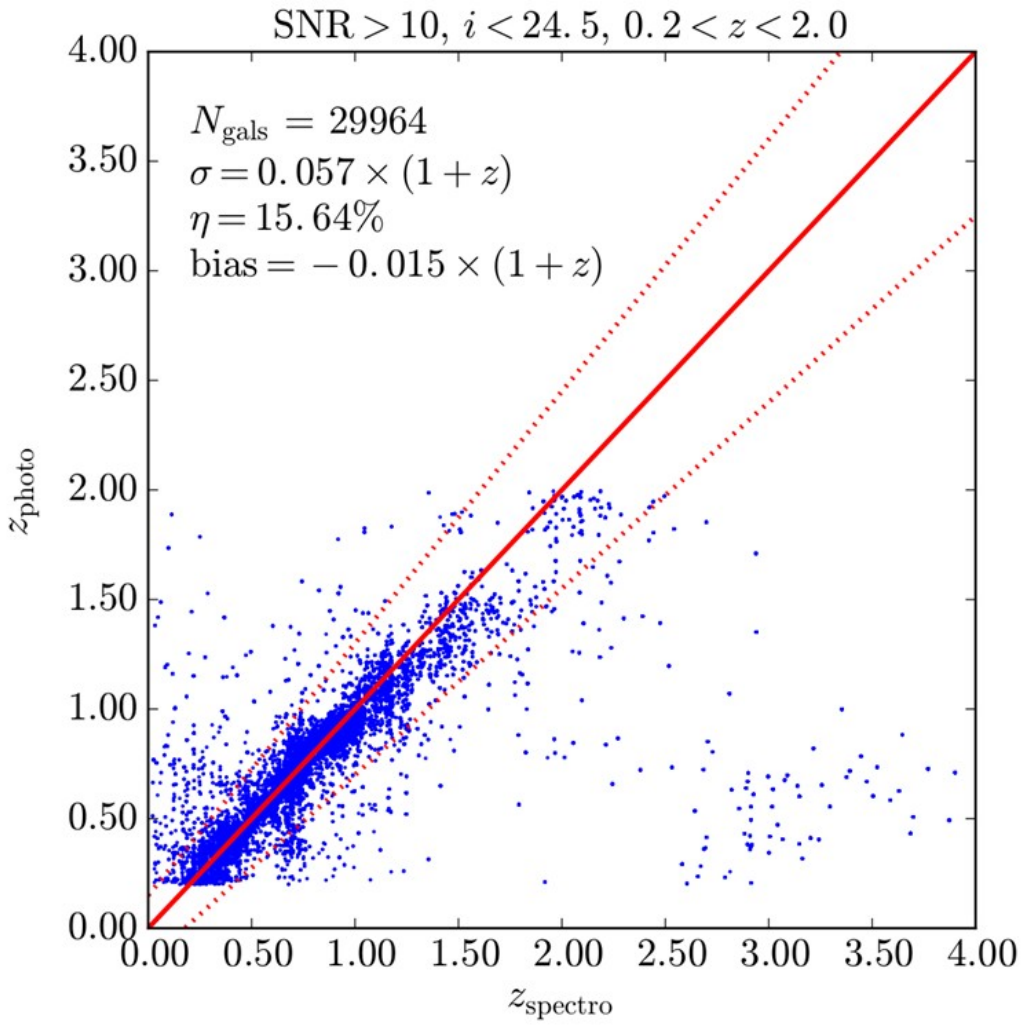
- Results weighted by the density of spec-z's in colors and magnitude (Lima et al. 2008)
- No method can meet the requirements: $\sigma_z/(1+z)<0.05$, OF<10%
- Note the "used fraction"!

Method	Type	$\sigma_z/(1+z)$	Outlier fraction	Used fraction
MLPQNA	ML-NN	0.057	11.99	0.60
AdaBoost	ML-DT	0.068	21.97	1.00
Le Phare	TF	0.070	17.49	0.85
ANNz	ML-NN	0.077	21.77	0.94
SOM+RF	ML-DT	0.064	18.92	0.78
Color prior+Le Phare	ML-*k*NN, TF	0.057	15.60	0.94

Data Challenge 2 Results

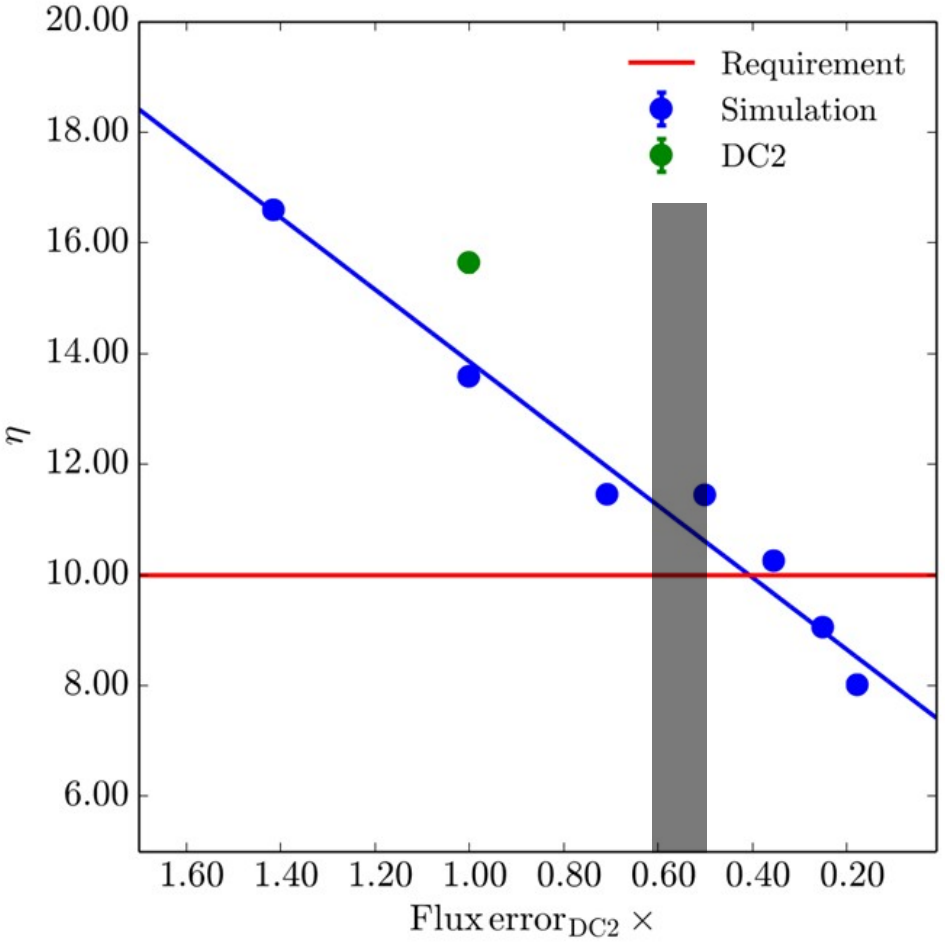


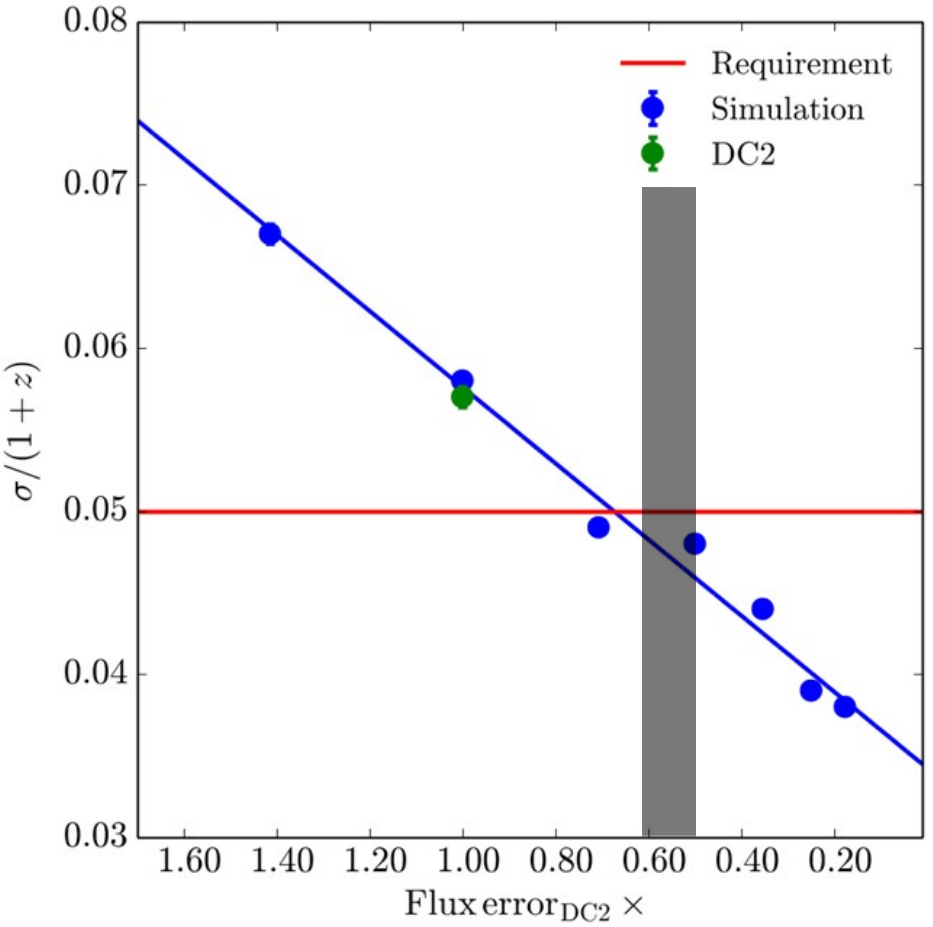
Unweighted



Weighted

DC2 and Photometric Depth





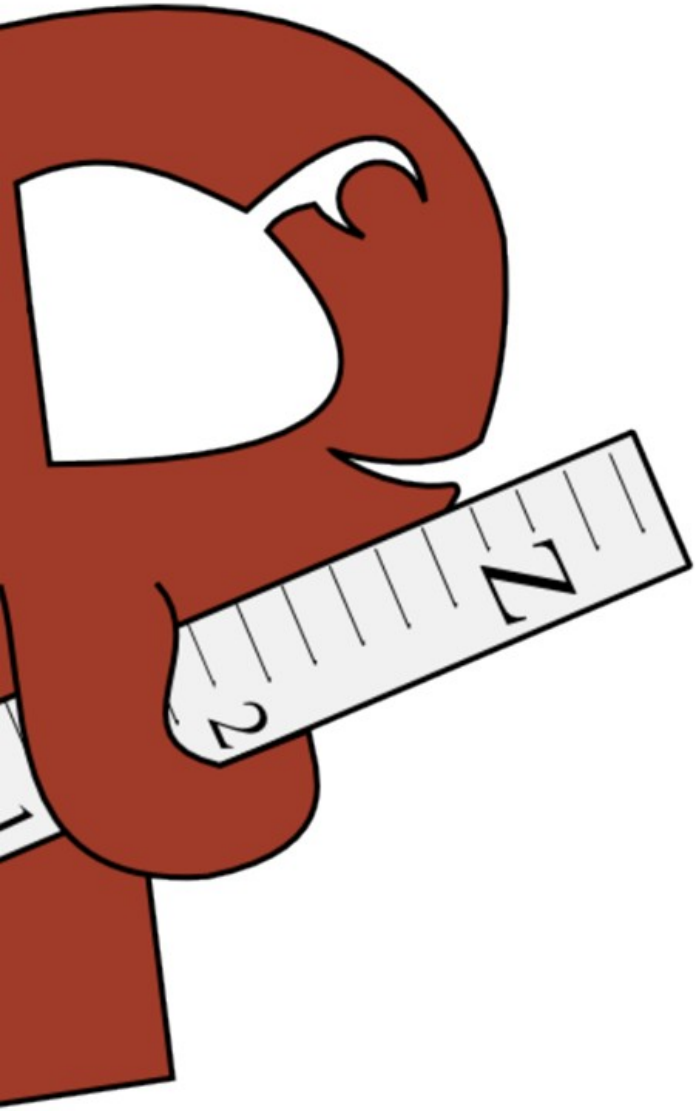
- DES is shallower than expected
- Requirements would be met with Red Book depths

→ P. Capak

How to improve Template-Fitting ?

- Improve knowledge of the astrophysics, through deep fields, to generate better templates and better priors
- Understand better the properties of emission lines, galactic absorption, intergalactic absorption, intrinsic reddening, and treat them correctly
  → Audrey's talk
- Bayesian approach removes the issue regarding the number of templates (but put more constraints on the knowledge of the priors)
- Tricks:
  - Zero-point corrections; more complex model?
  - Template adaptation
  - Correct treatment of upper limits
  - Marginalization of the scale factor
  - …

Phosphoros

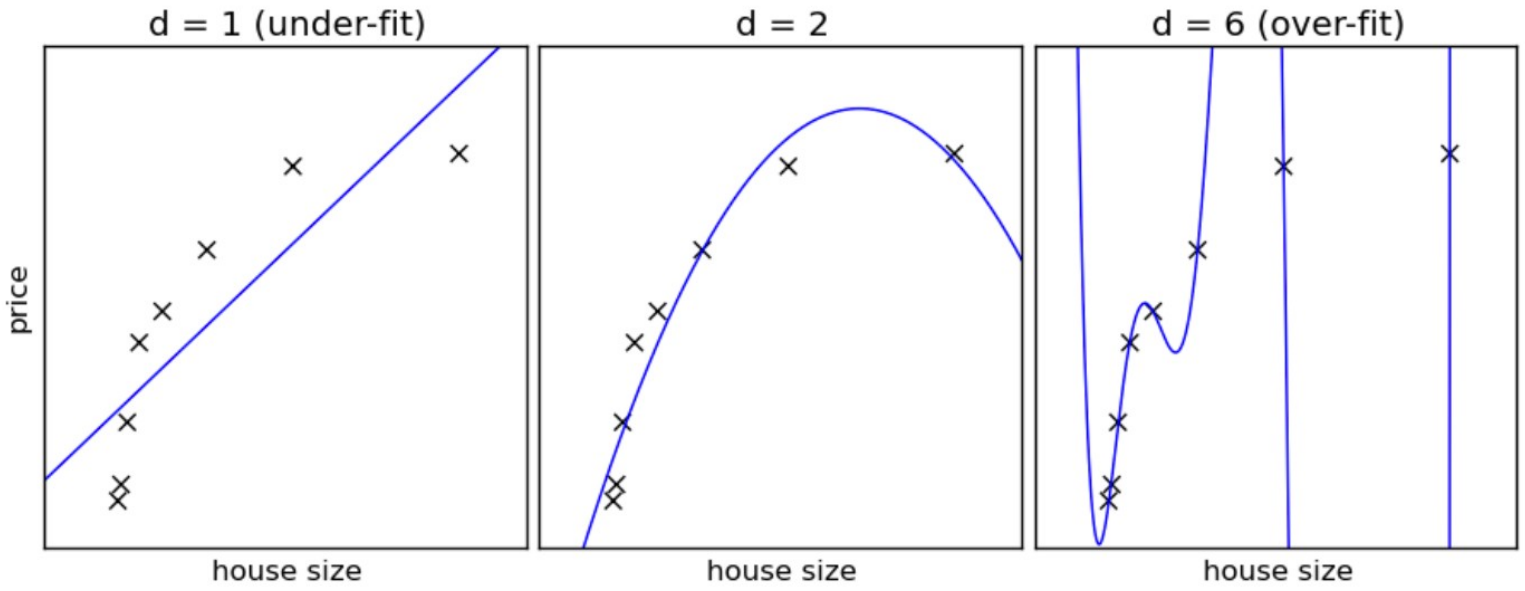


- Phosphoros is the template-fitting code developed for Euclid by the Swiss SDC
- Computationally extremely efficient (C++)
- Implements most of the features found in other TF codes
  - Zero-point correction
  - (Luminosity) priors
  - IGM (several choices)
  - Upper limits
- More features are being implemented
- Fully Bayesian: marginalization
- Phosphoros will be used for physical parameters

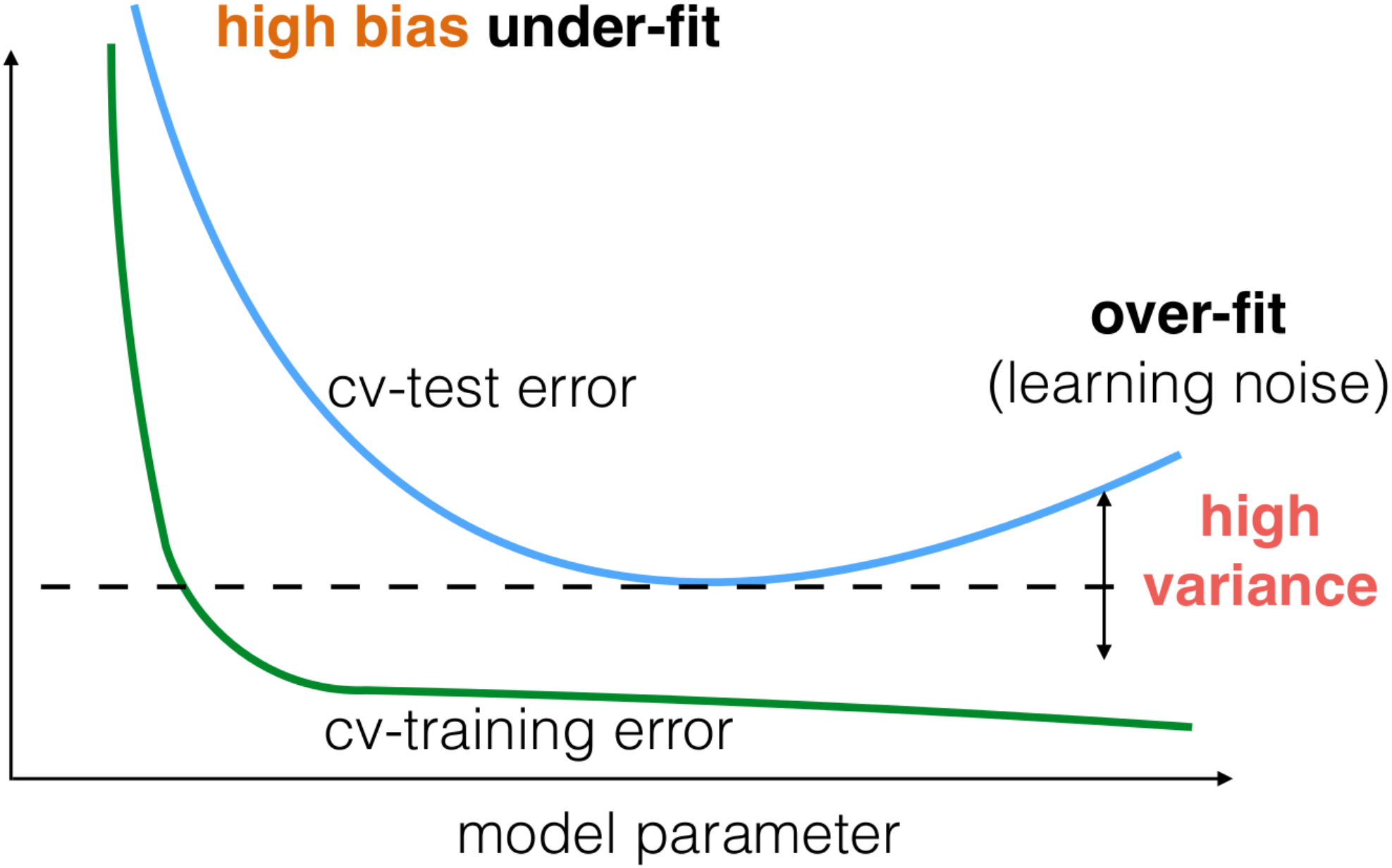
- Phosphoros 0.5 released to OU-PHZ members, will be public after validation

How to improve Machine-Learning ?

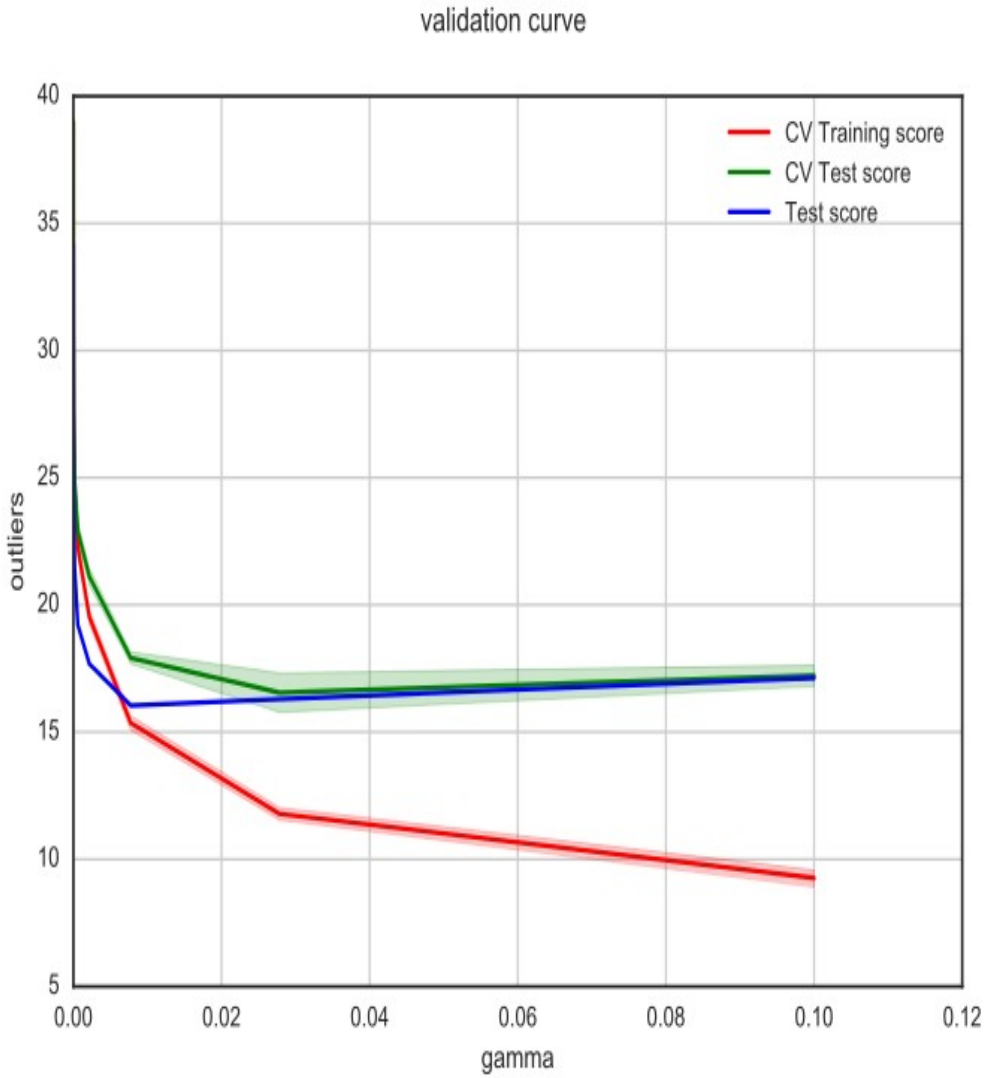
- Additional features (object size, different kinds of photometry)? Maybe, but little success (so far)
- Machine-learning requires a model complexity:
  - Too small → underfitting
  - Too large → overfitting
- Training set must be fully representative → extrapolation



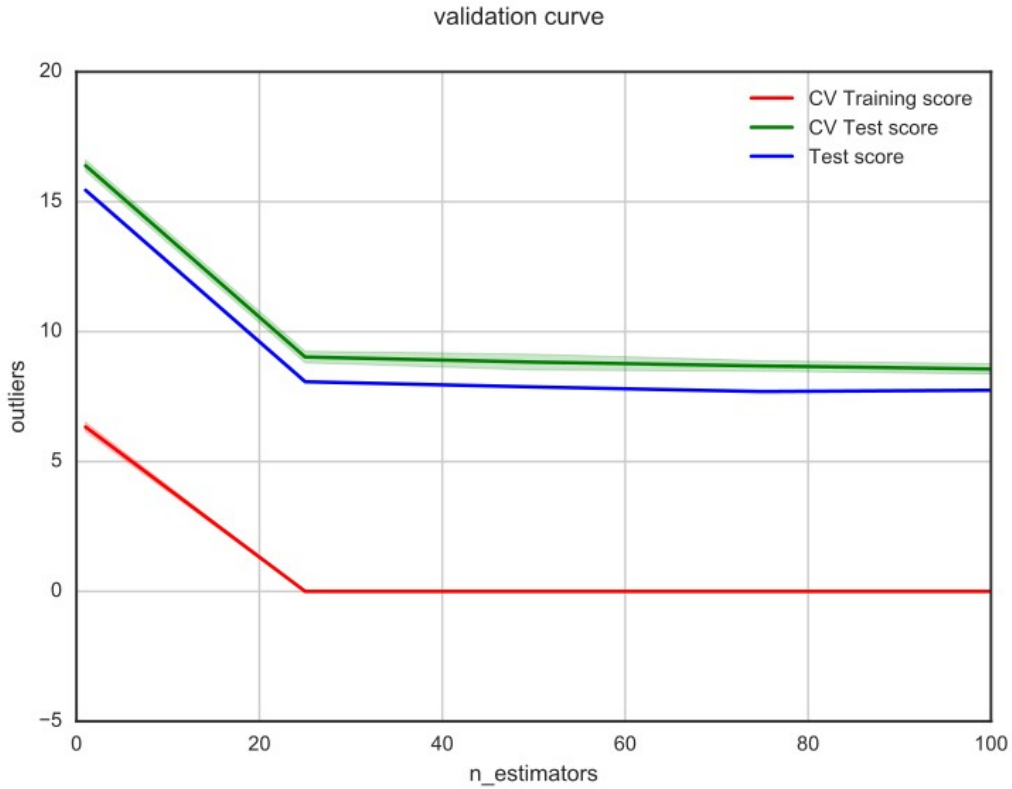
Tuning Model Complexity



Cross-Validation Curves

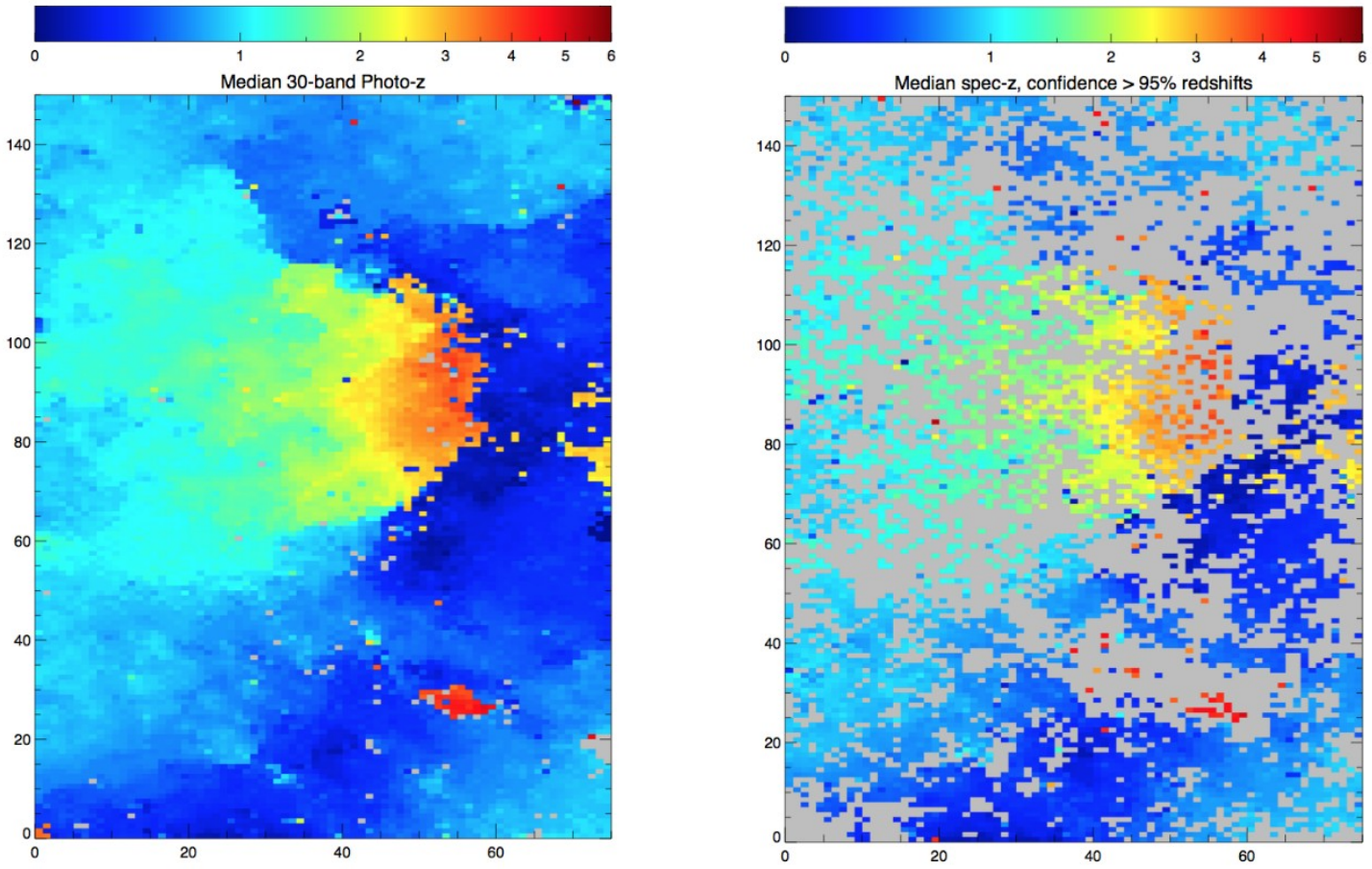


SVR



AdaBoost

How to avoid extrapolation ?



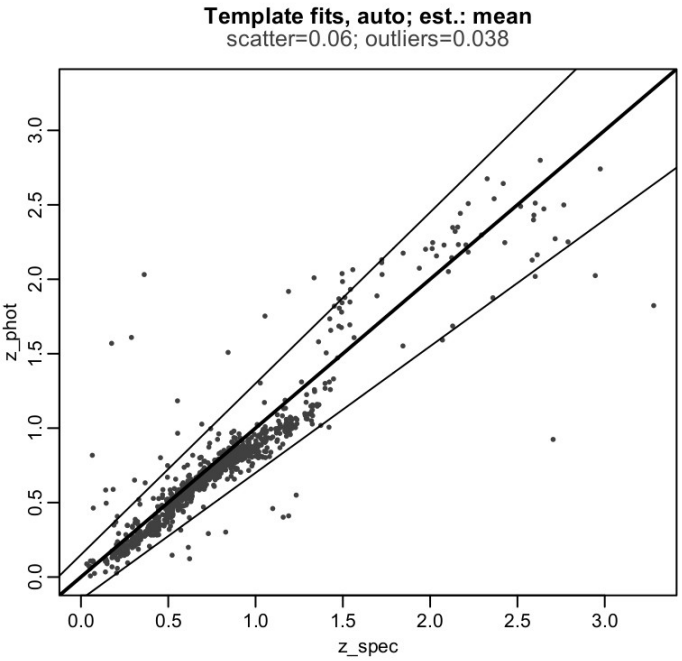
We need to cover the color-space of galaxies with spec-z, similarly to the bias calibration (Masters et al. 2015; → Peter's talk)

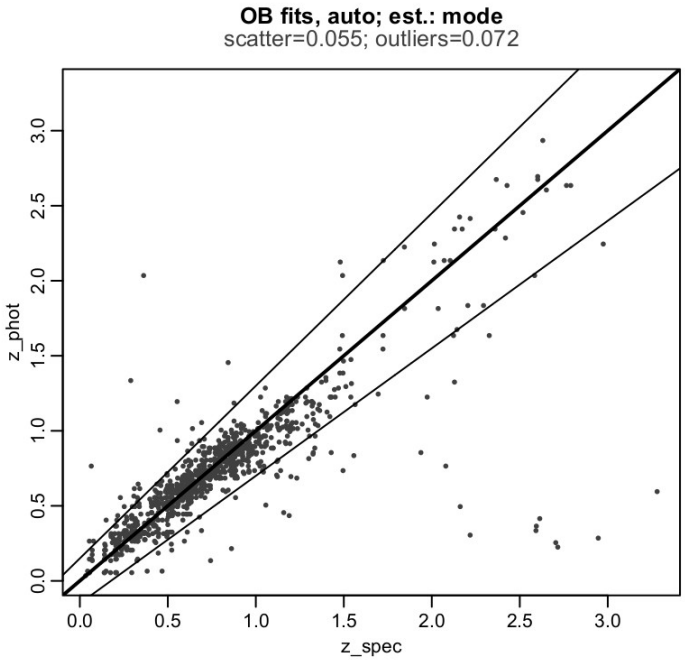
How many spectra are needed ? Presumably depends on the algorithm

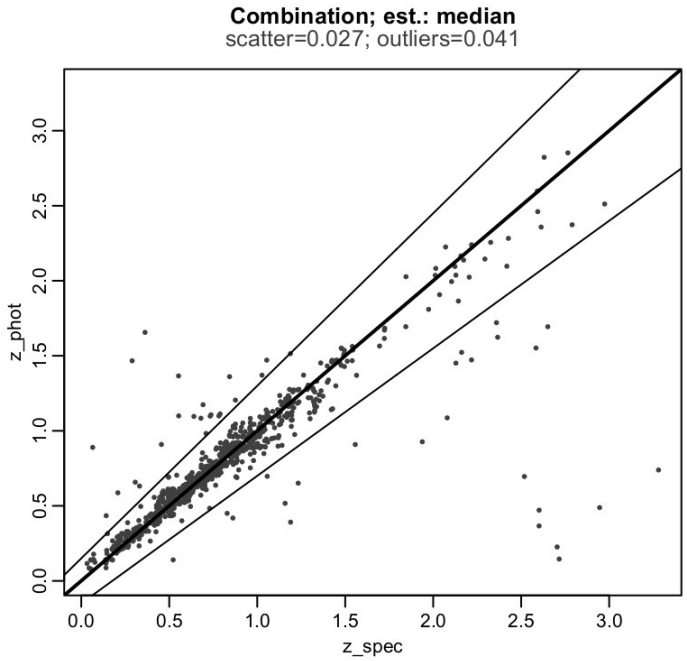
What ELSE can be Improved ?

- Different algorithms have their own strengths and weaknesses
  - Select or combine the different estimates improves the photo-z's (Hildebrandt et al. 2010; Dahlen et al. 2013)
- Some objects are not well behaved
  - Identify them, and remove them from the WL sample
  - Use different mapping for different classes of objects

Photo-Z Combination





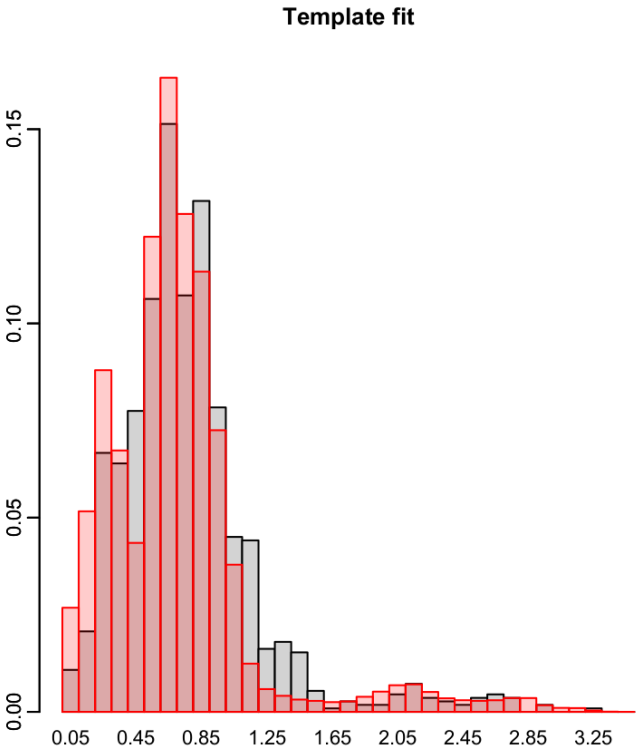


Le Phare TPZ Combination

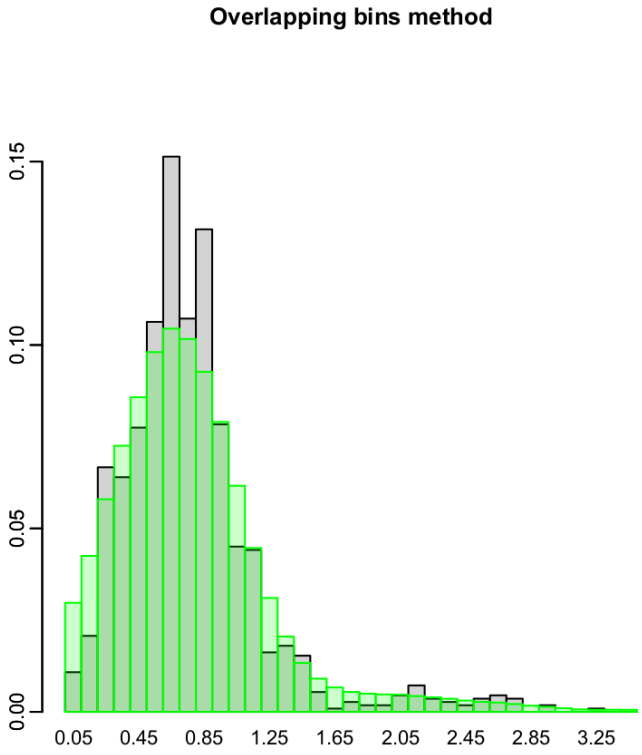
Classifier-based combination (Random Forest)

Süveges et al. in prep.

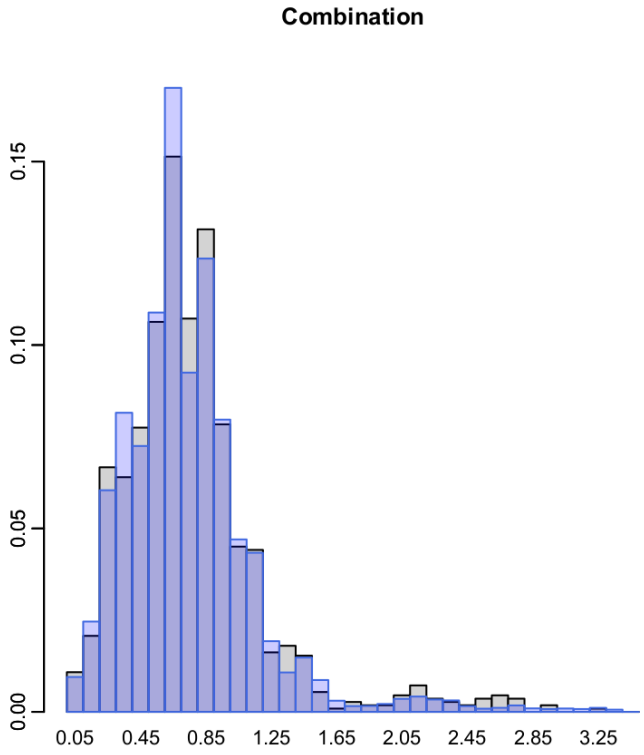
N(z) Reconstruction



Le Phare



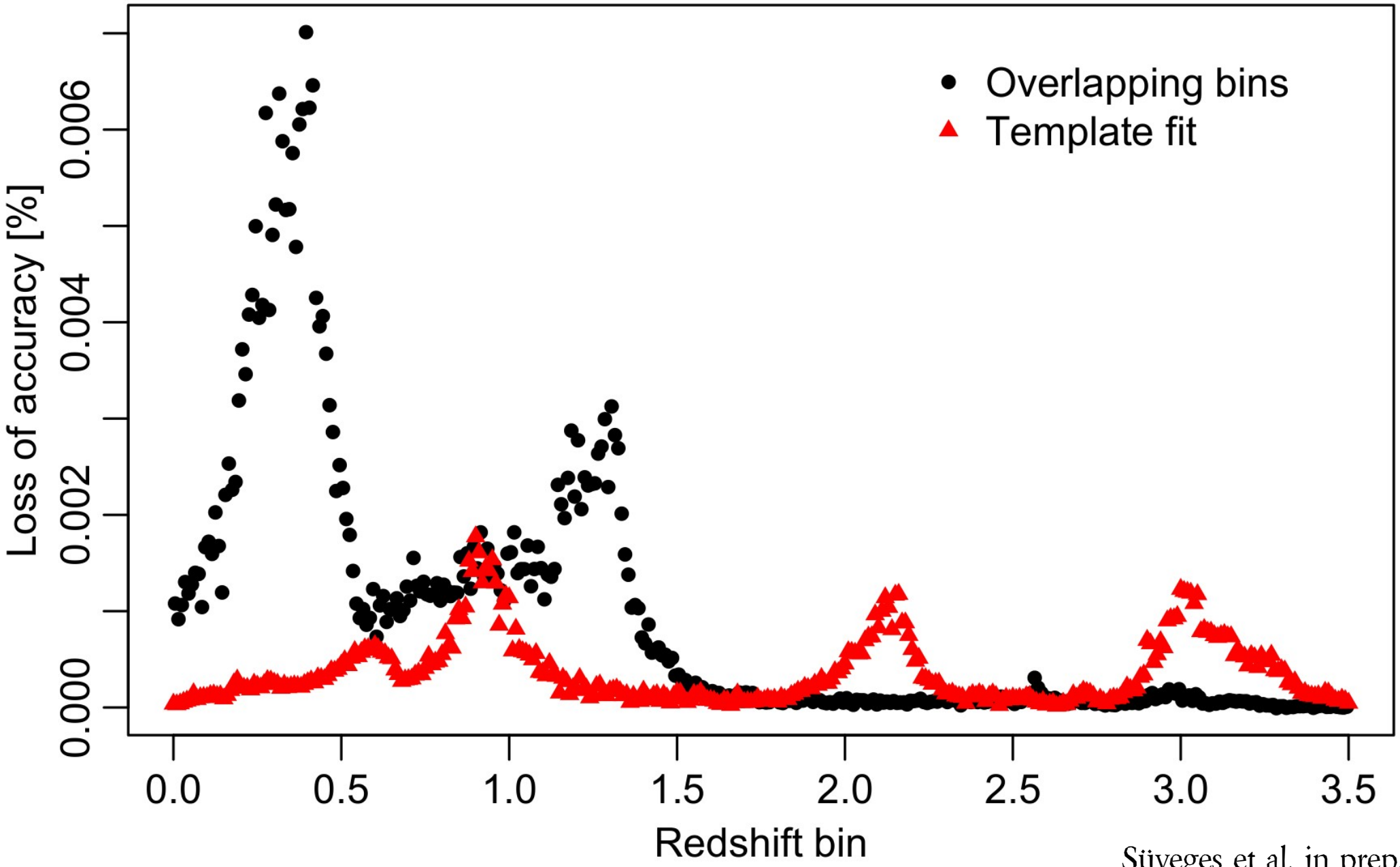
TPZ



Combination

Süveges et al. in prep.

Feature Importance



Süveges et al. in prep.

Problematic Sources

- Sources that have different nature need a different mapping
  - Stars, AGN, QSOs, ???
  - Including them would require to increase the model complexity, but they are rare, so difficult to train
  - But AGN and QSO's are important for Legacy Science

- Alternatively, we can try to identify them using all possible parameters (features): eROSITA, WISE, Galex, morphology
  - Use a supervised classifier (human or machine-learning)
  - Define the mapping for each class

Photo-z with Human Decision Tree

- Algorithm for optimal photo-z reconstruction for X-ray sources (Salvato et al. 2009)
- Uses non-photometric data
- Note: it's a decision tree!

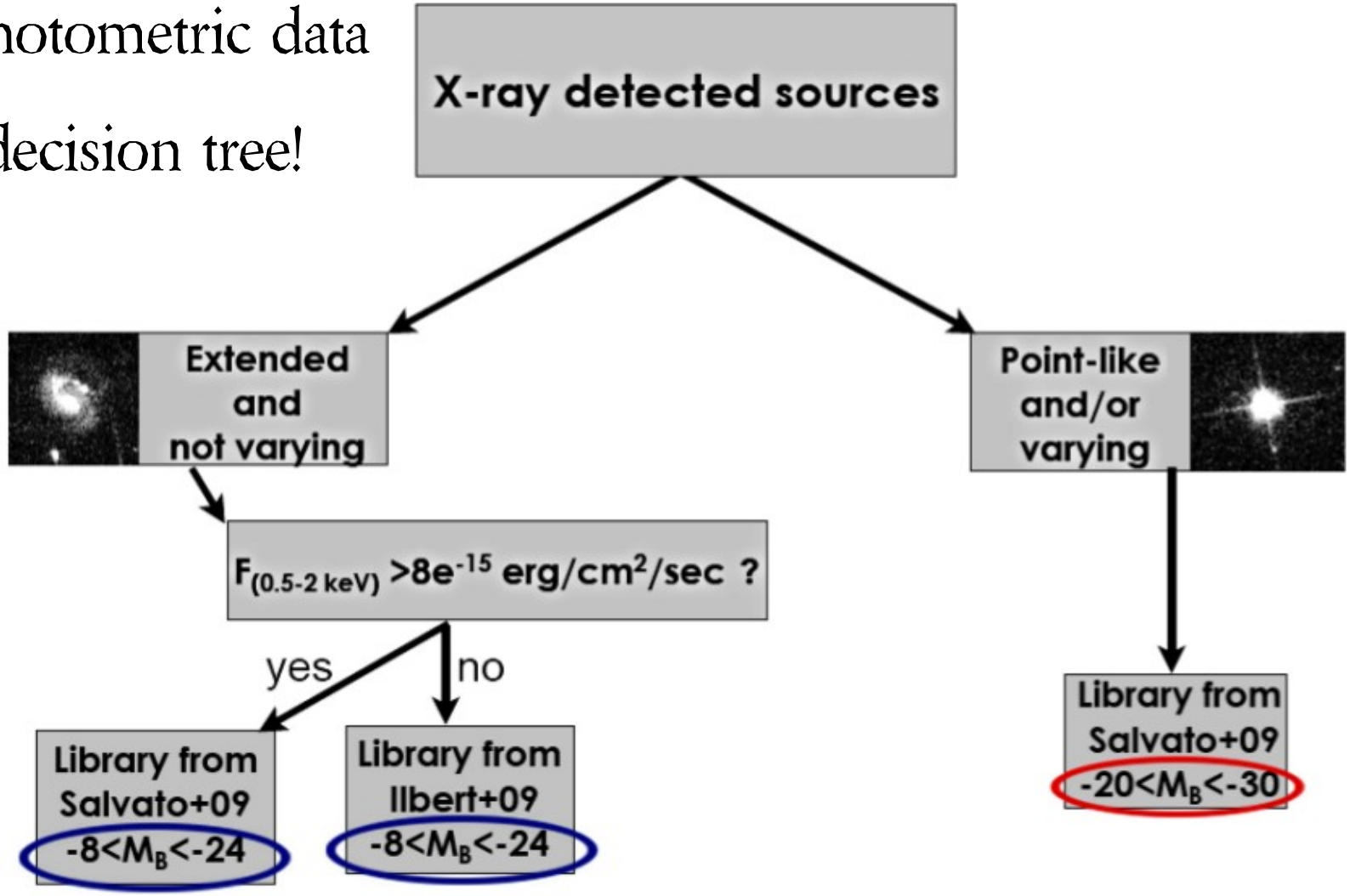
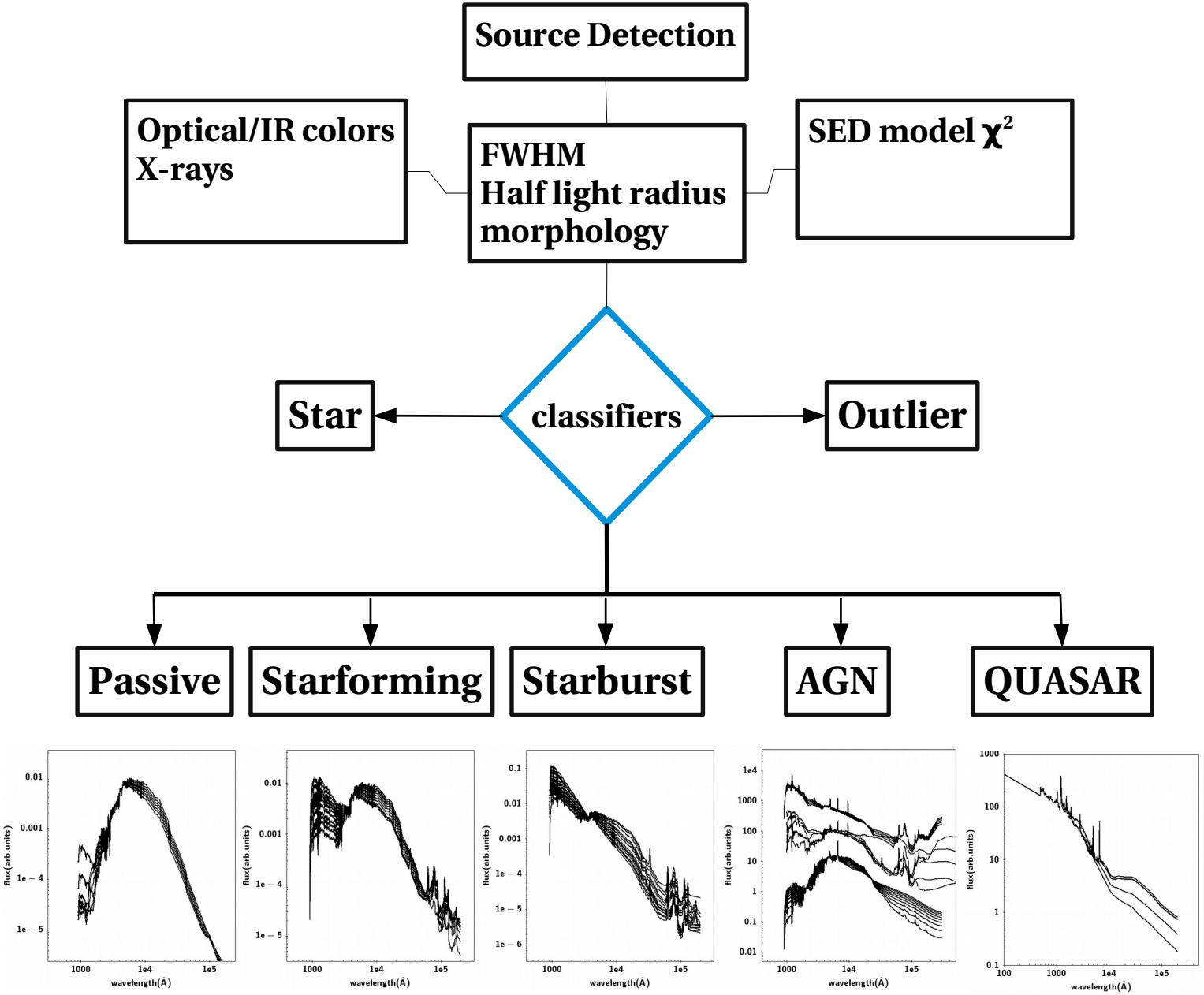
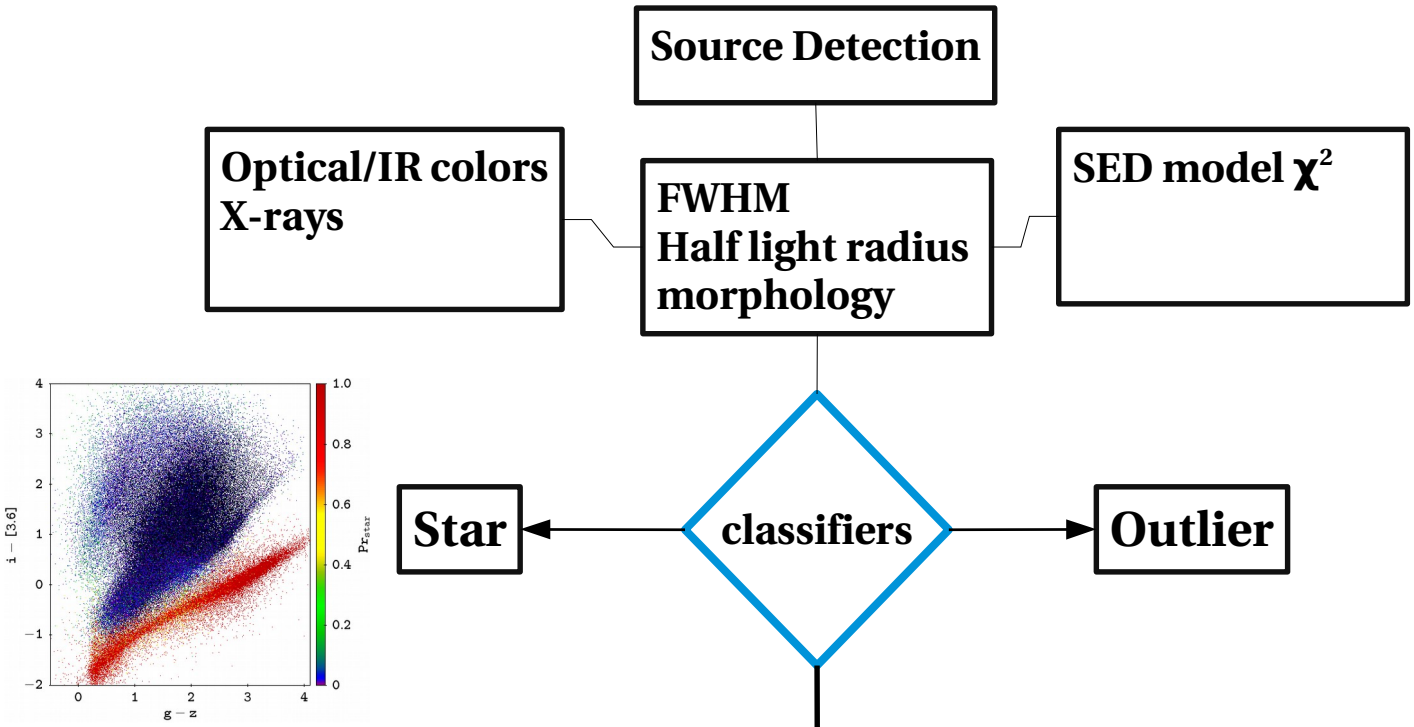


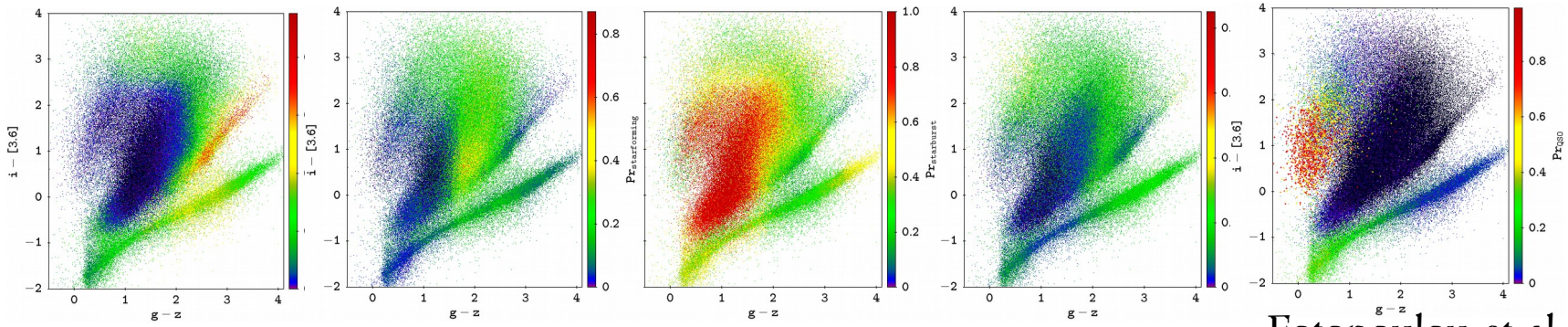
Photo-z with Random Forest



Fotopoulou et al. in prep

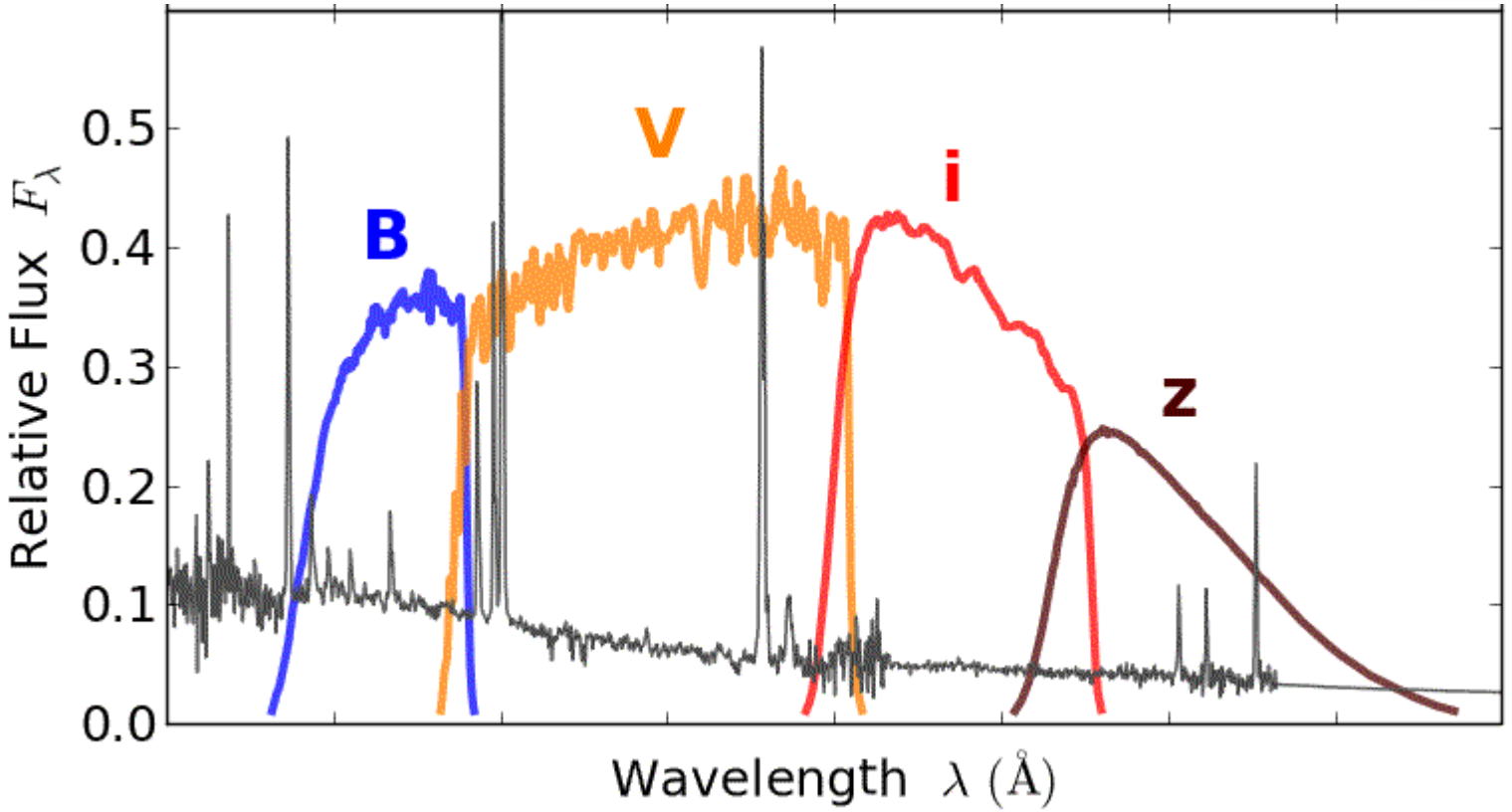
Photo-z with Random Forest





Fotopoulou et al. in prep.

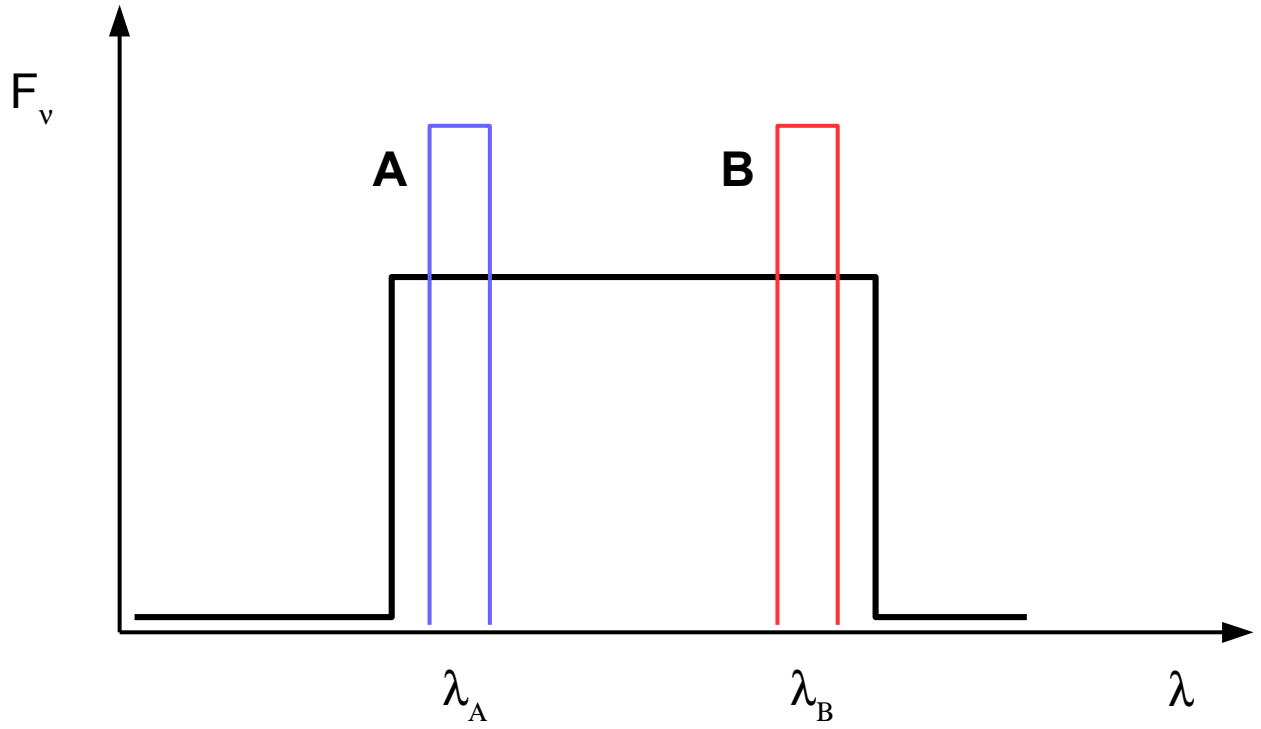
How do we determine fluxes ?



$$F_\nu = \frac{\int_\lambda f_\lambda T(\lambda) d\lambda}{\int_\lambda T(\lambda) \frac{c}{\lambda^2} d\lambda}$$

F_ν, the energy flux, determines the energy received per unit of time, in erg s^{-1} cm^{-2} Hz^{-1}

Energy Fluxes and Photon Fluxes



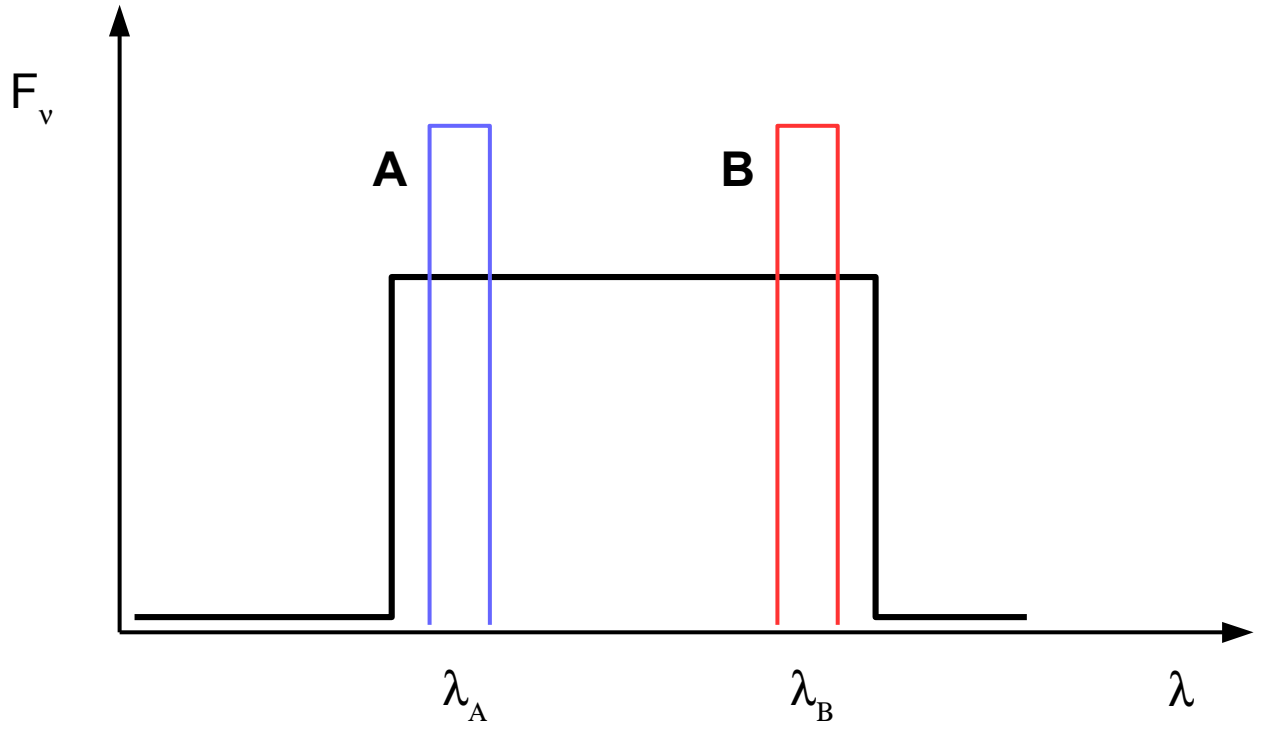
Let's assume sources A and B are monochromatic sources which have the same Energy Flux, and define $\phi_\nu = f_\nu/h\nu$

$$h\nu_A \Phi_\nu(A) = F_\nu(A) = F_\nu(B) = h\nu_B \Phi_\nu(B),$$

with Φ_ν the Photon Flux in photons s^{-1} cm^{-2} Hz^{-1}

So, we have : $\Phi_\nu(A) < \Phi_\nu(B)$

Calibration in Photon Flux ?

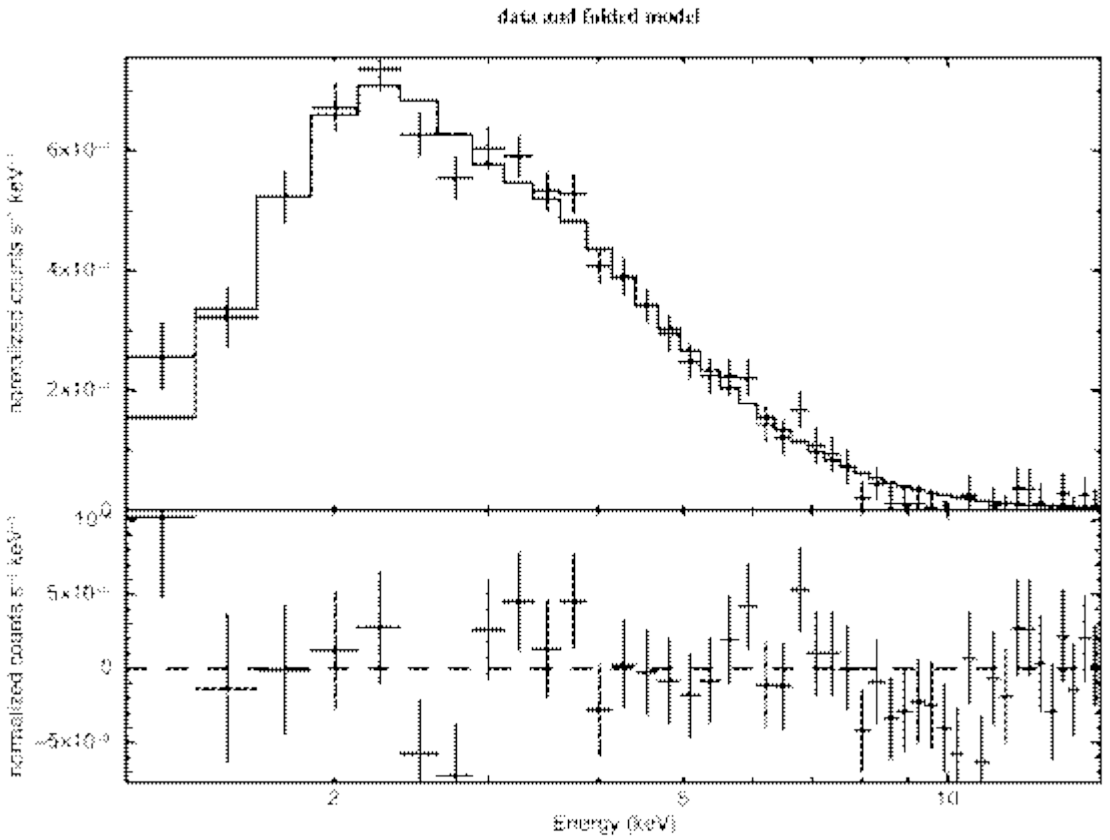


For a CCD counting device, the counts are proportional to the Photon Flux, and not to the Energy Flux

If one calibrates the Photon Flux instead of the Energy Flux, which is probably trivial, we can get rid of this bias.
Both TF and ML algorithms can benefit from this improvement

Color-dependent calibration (i.e. using a color term) can alleviate the problem, but not entirely (and is it done?)

Beyond fluxes



In X-ray astronomy, each observation comes with its own response. The source spectral properties are then obtained by forward fitting an emission model through the response to the count rates

In fact, one should not even subtract the background in order to preserve fully correct statistical behavior !

Conclusions

- Core photo-z algorithms are mature, little to gain here
  - Completely new approaches?
- Improvements can be obtained from:
  - Improving astrophysical knowledge
  - Adding new features
  - Tuning the model complexity
  - Tuning the training set
  - Combining different methods
  - Use distinct mapping for different kinds of objects
- Can we gain something from the calibration? Photon flux? Response per object?