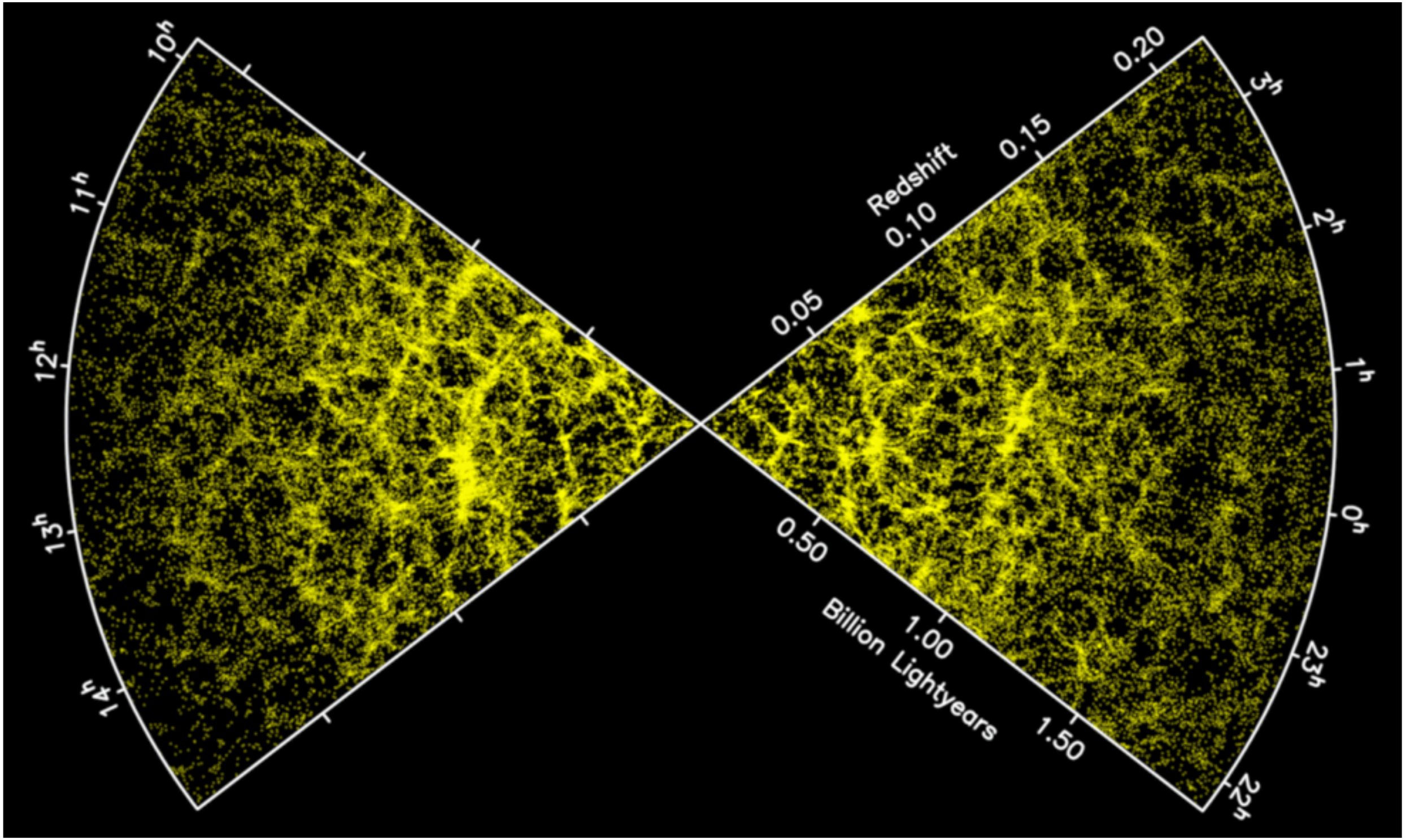
Machine Learning for photometric redshift estimation

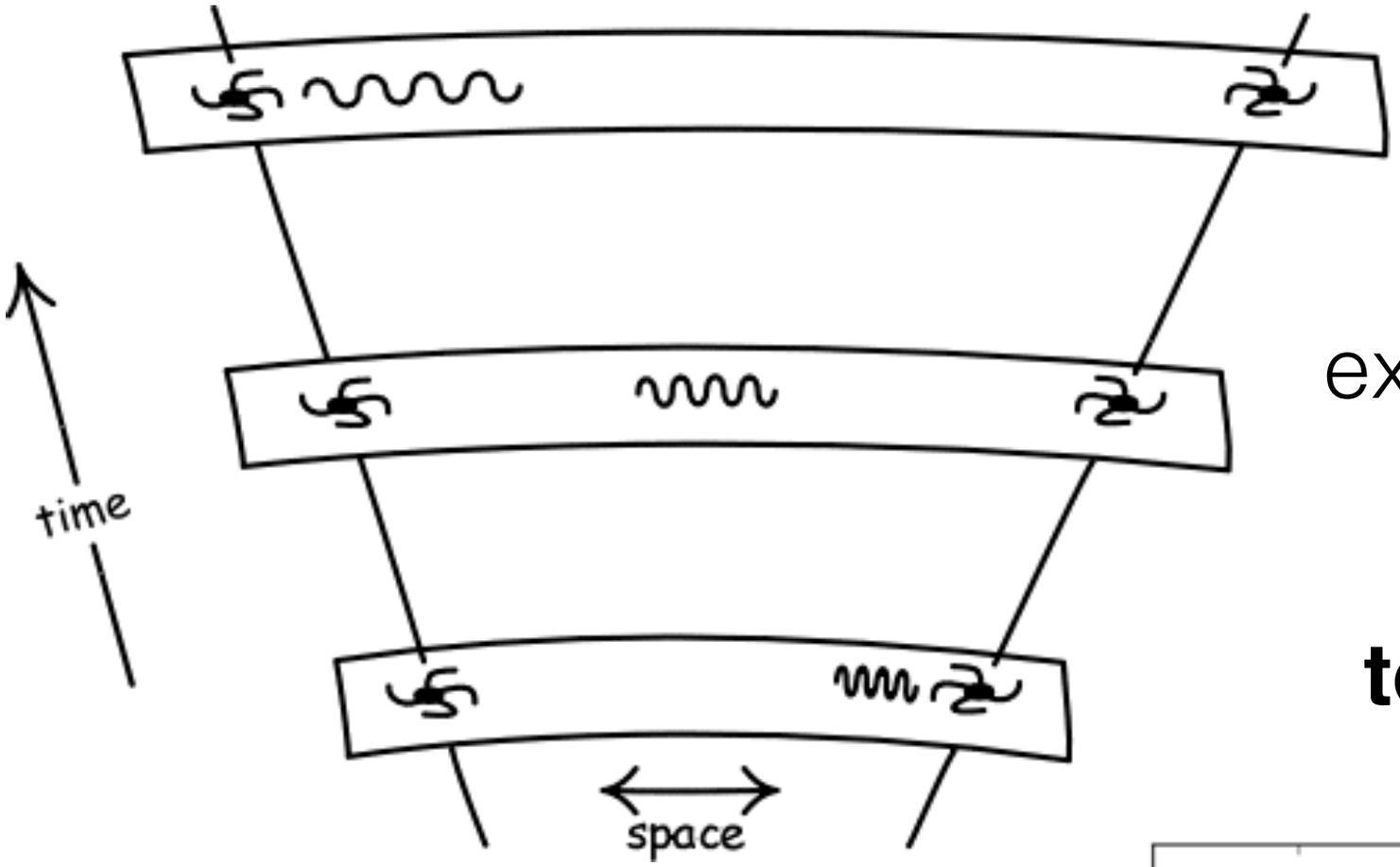
Markus Michael Rau
Ben Hoyle
Kerstin Paech
Stella Seitz

Overview

- What are photometric redshifts (photoZ)?
- Why are accurate photoZs important for cosmology?
- How can we quantify the photoZ error distribution?
- How can we avoid systematic errors?

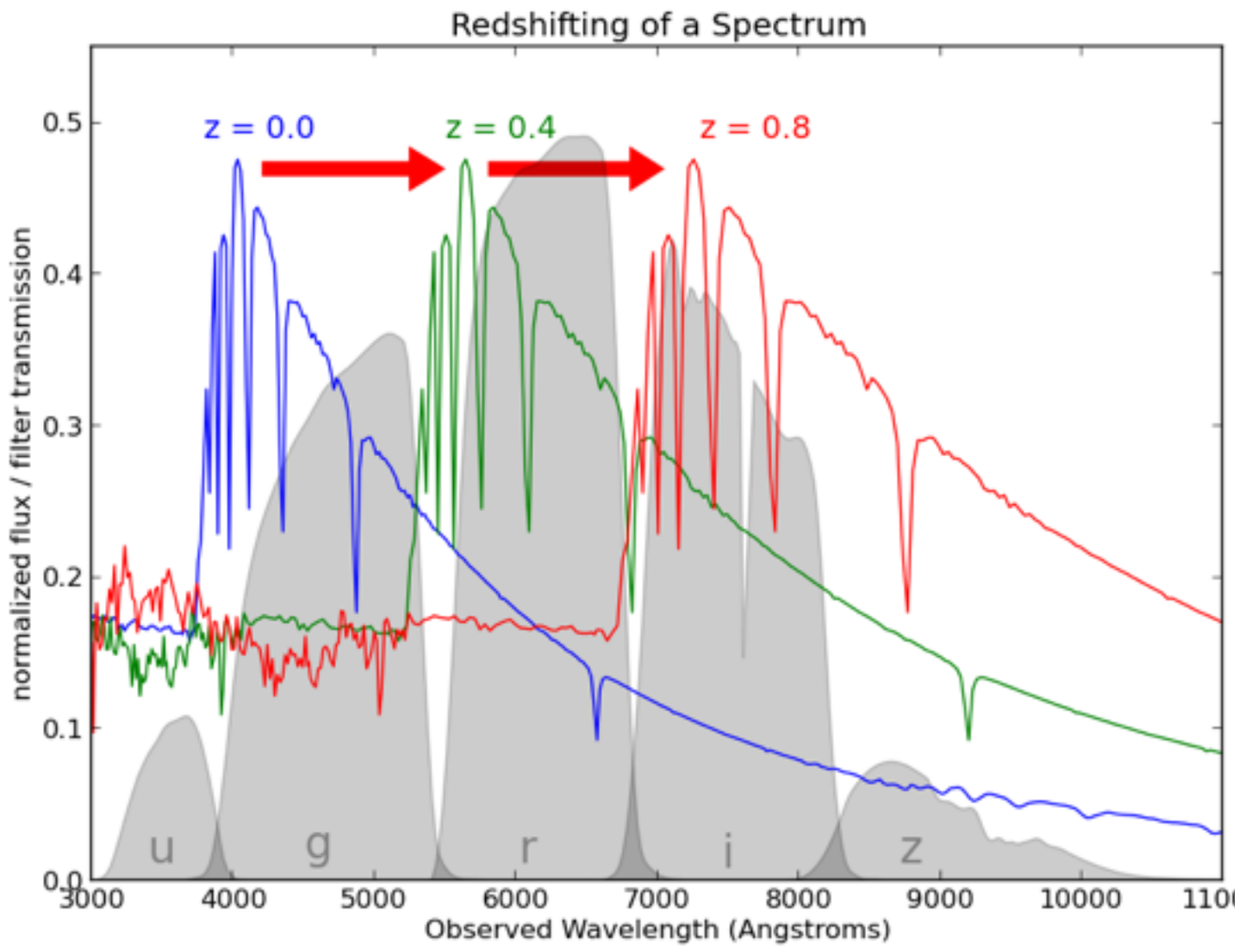
The cosmos in 3D





Distance to objects:
Background universe expands and „**stretches“ light wave and shifts spectra to longer wavelengths**

Fluxes (or magnitudes) in photometric bands are **predictors for redshift** (distance)



How can I use Machine Learning for photoZs?

Calibration data from overlapping region of spectroscopic survey and photometric survey

color-mag. space
spec.-photo.
data

Train model

KNOWN
redshift

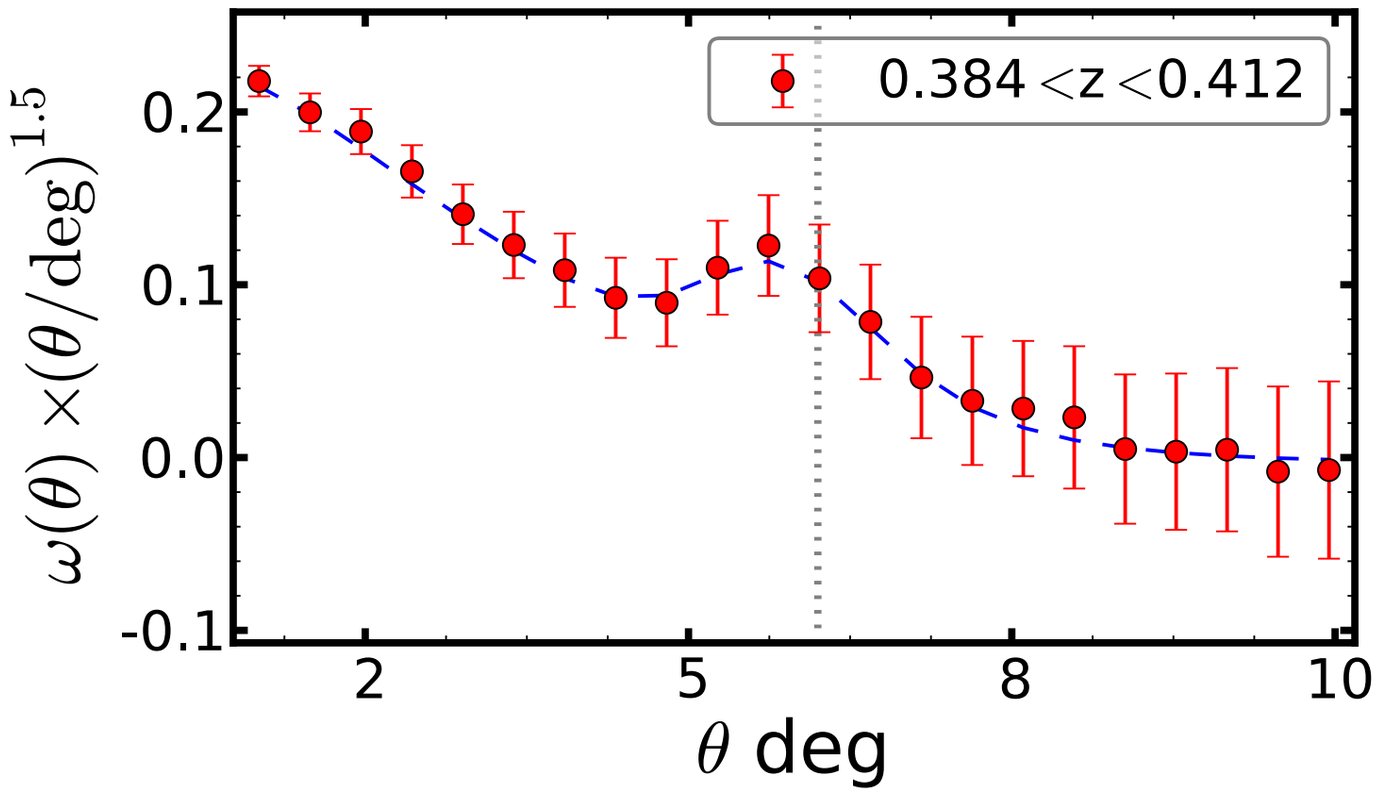
color-mag. space
only
photo. data

Apply model

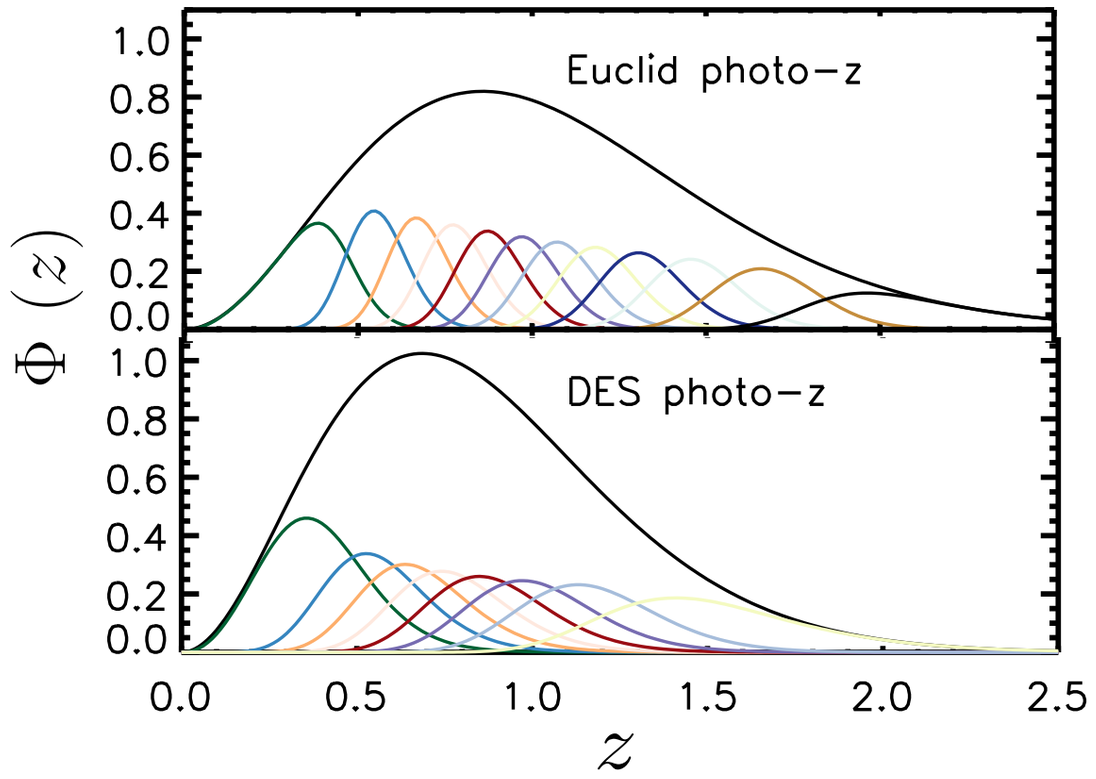
ESTIMATED
redshift
(PhotoZ)

PhotoZs can be obtained for all other galaxies of the photometric survey

Measurement



Redshift Distribution



Theory

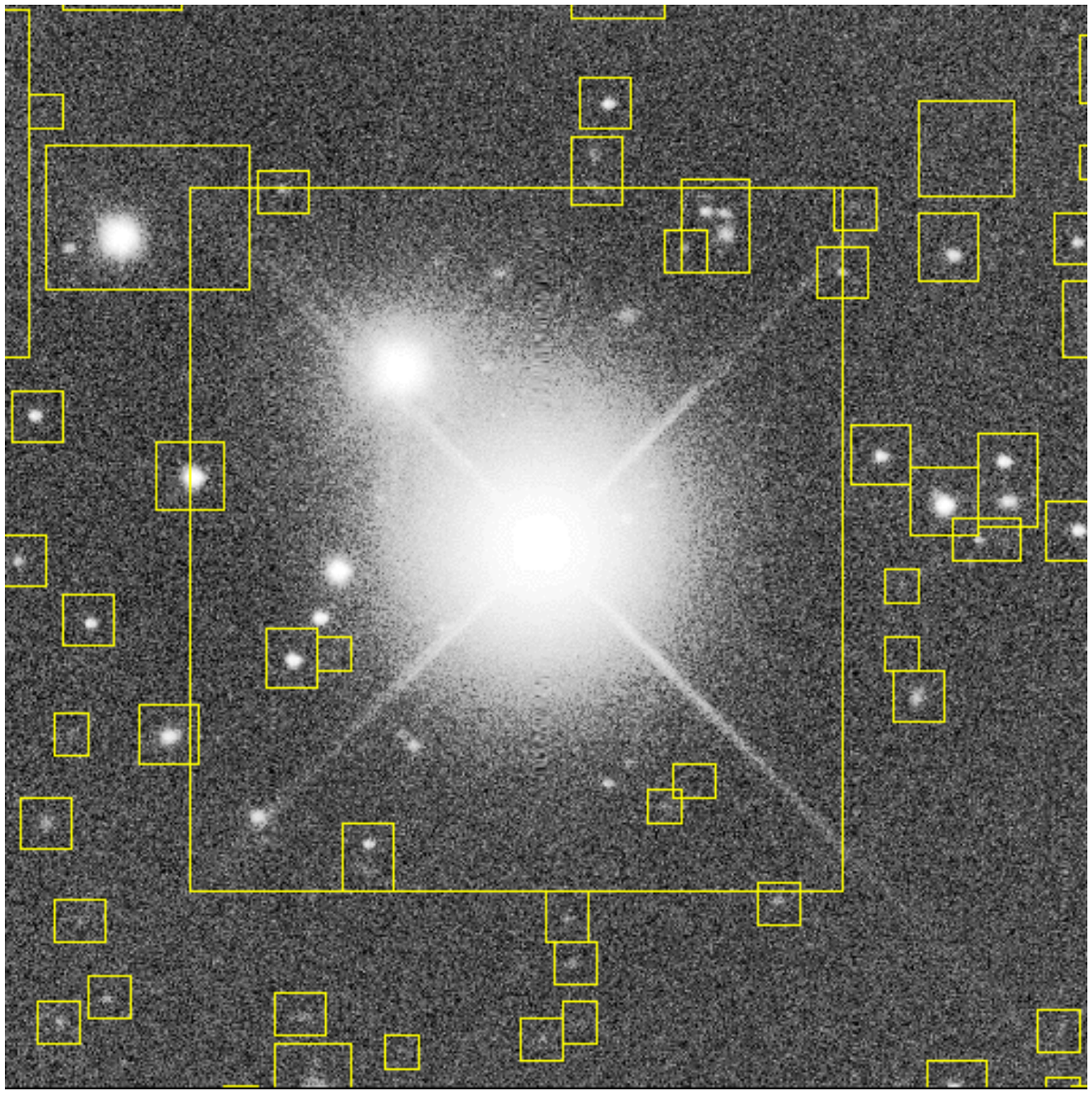
$$\Psi_\ell(k) = \int dz\, \phi(z)\, \mathcal{D}(z) j_\ell(kr(z)).$$

$$C_{\ell,\mathrm{Exact}} = \frac{1}{2\pi^2} \int 4\pi k^2 dk P(k) \Psi_\ell^2(k)$$

$$w(\theta) = \sum_{\ell \geq 0} \left(\frac{2\ell + 1}{4\pi} \right) P_\ell(cos\theta)\, C_\ell$$

Need accurate sample redshift distributions!

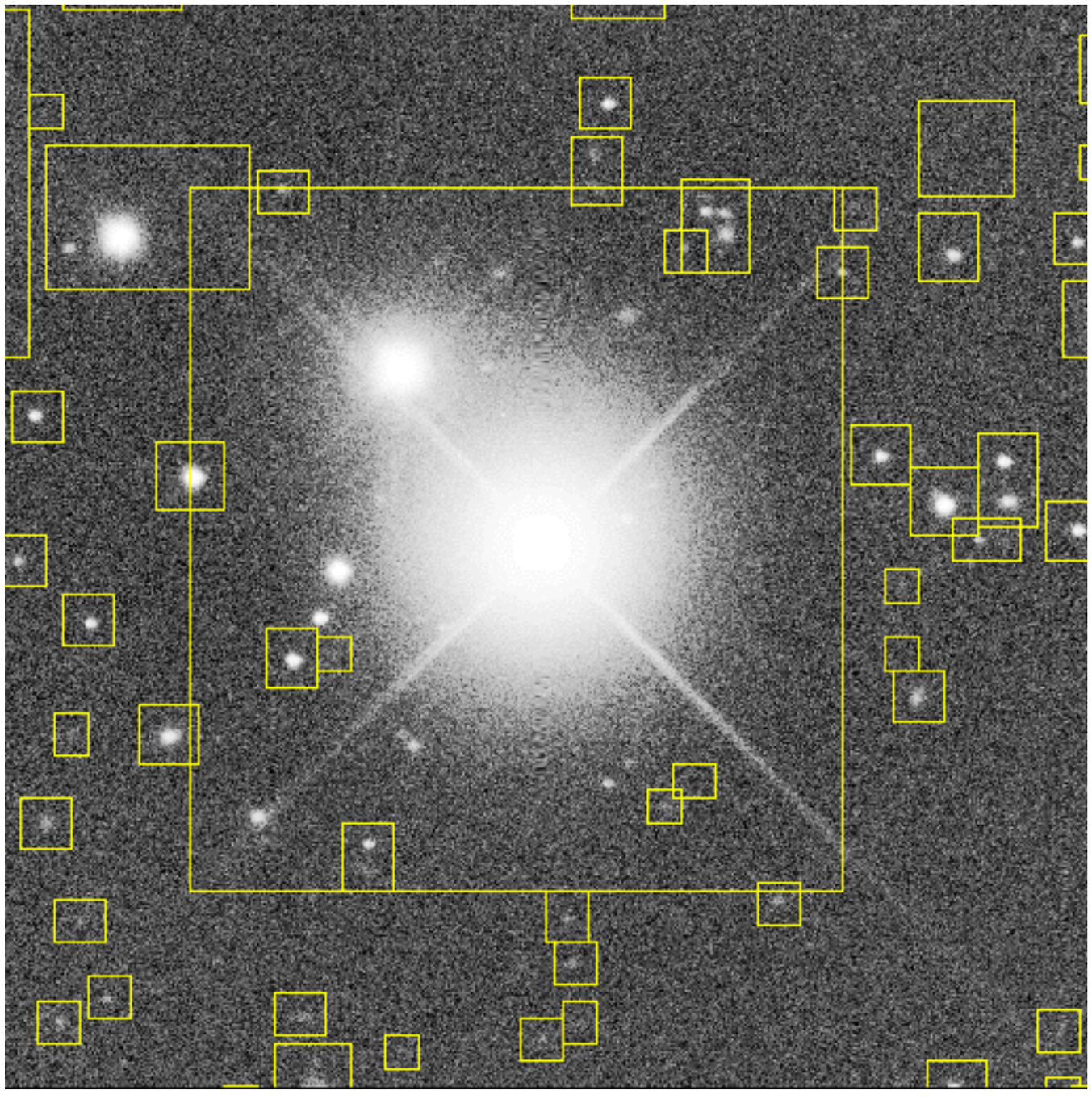
From images to photometry



Photometry is challenging!

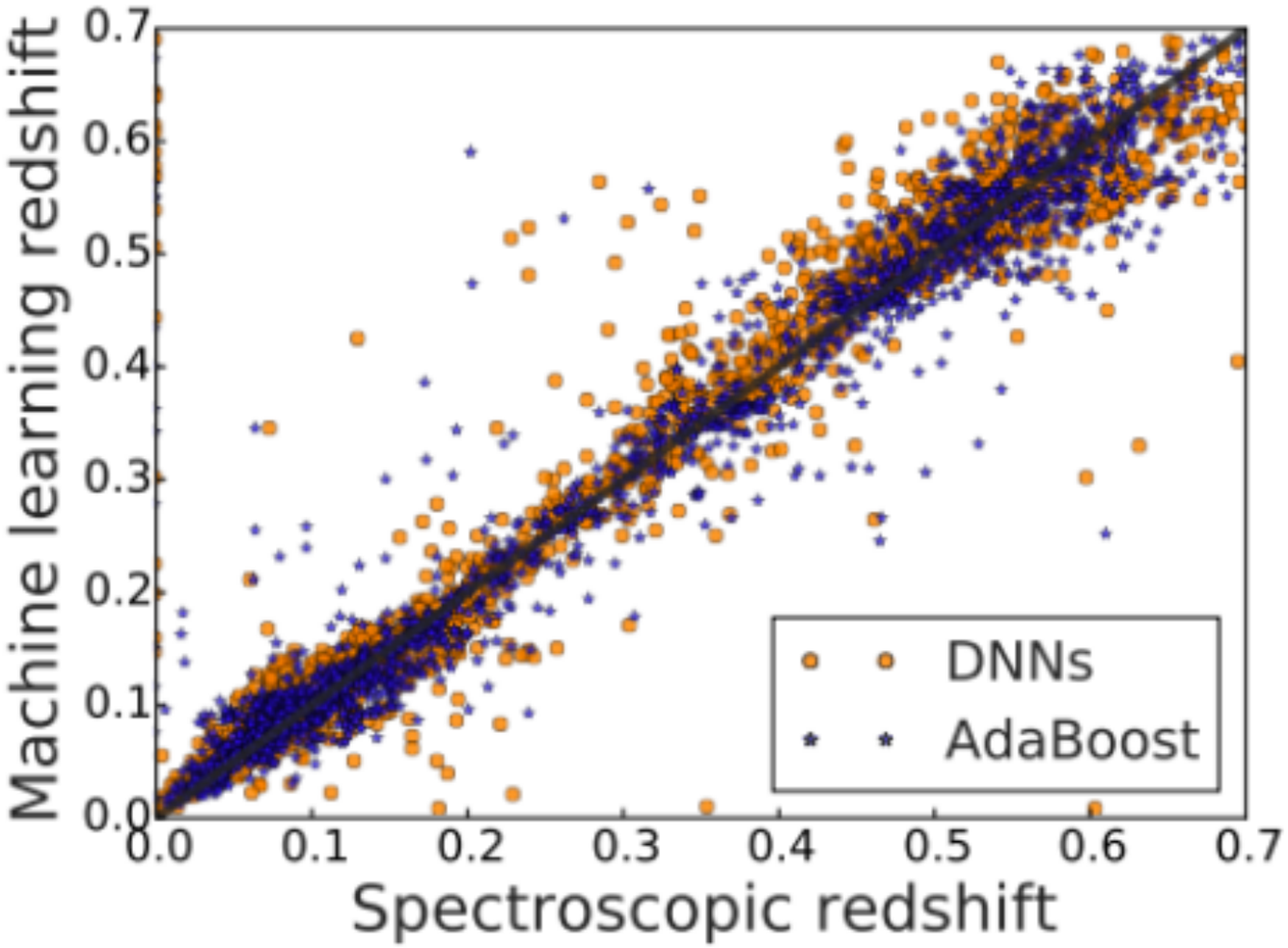
- bright objects
- crowded regions
- pixel failures
- …

From images to photometry



Photometry is challenging!

Alternative:
Apply a Convolutional Neural Network directly to the image cutouts produced by traditional tools



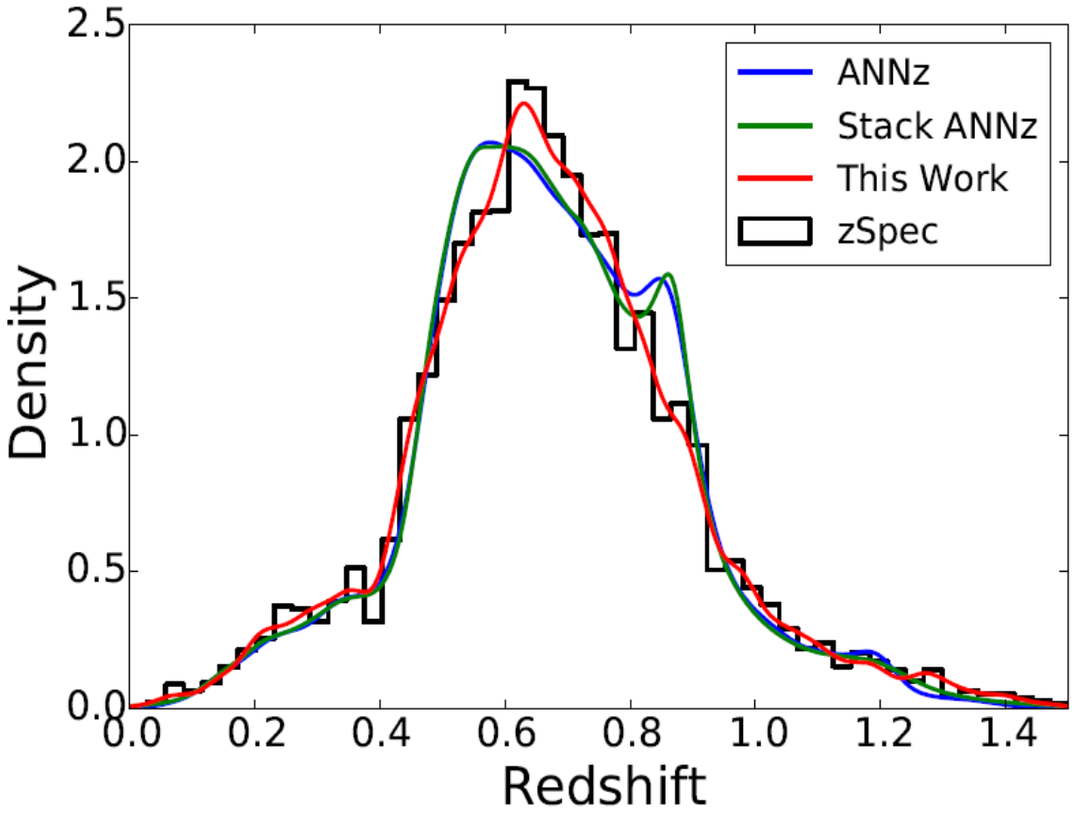
Hoyle 2015
arXiv:1504.07255

Consistent with state-of-the art „conventional“ method

In the following we use traditional photometry as inputs.

Photometry provides incomplete information about redshift → **Uncertainty** in photometric redshift

Sample distribution from different methods



Rau et al. 2015

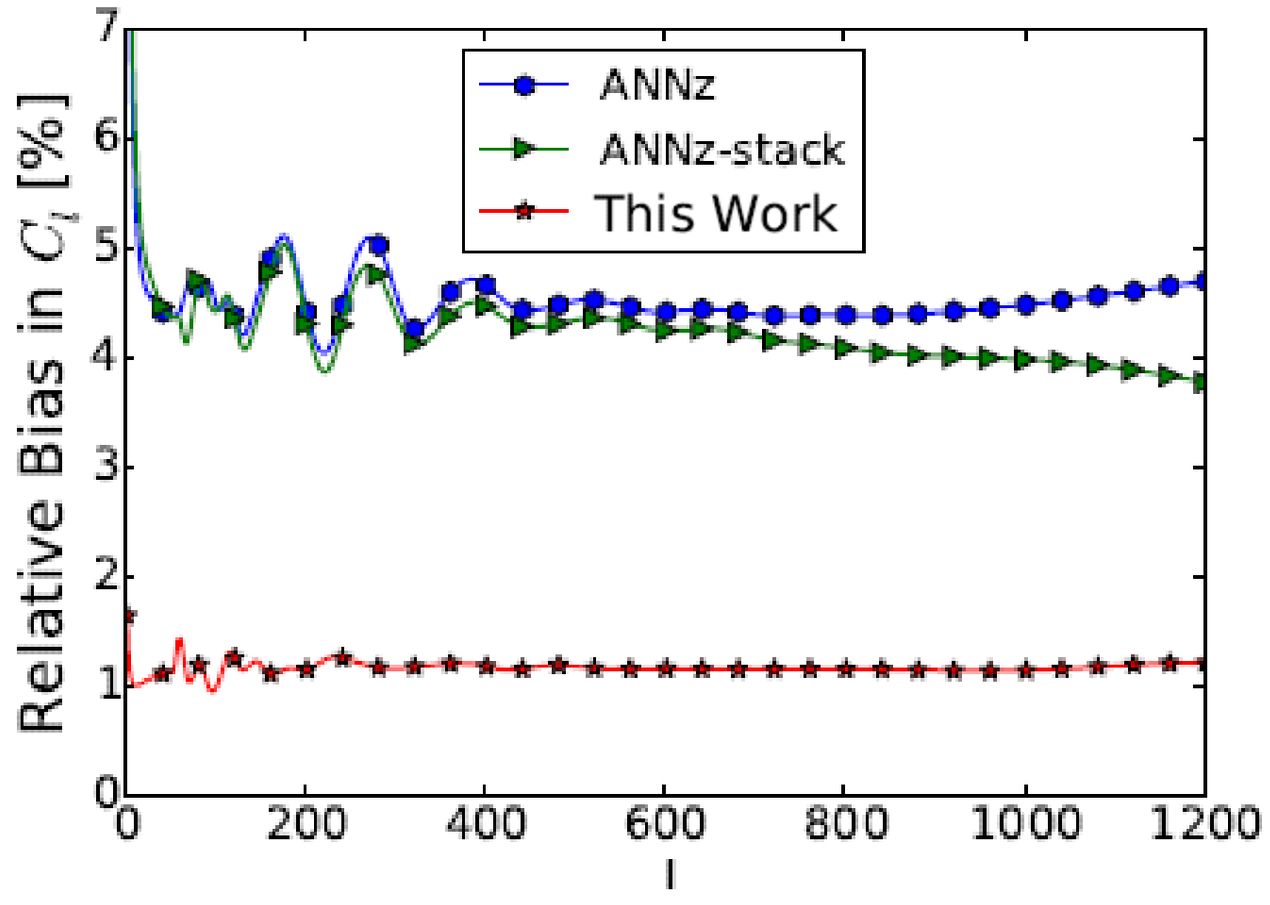
arXiv:1503.08215

$$\text{Relative Bias}\,(C_\ell) = \frac{C_\ell^{\text{phot}} - C_\ell^{\text{spec}}}{C_\ell^{\text{spec}}}$$

Traditional ML methods (e.g. ANNz) determine a single ‚best fit' prediction

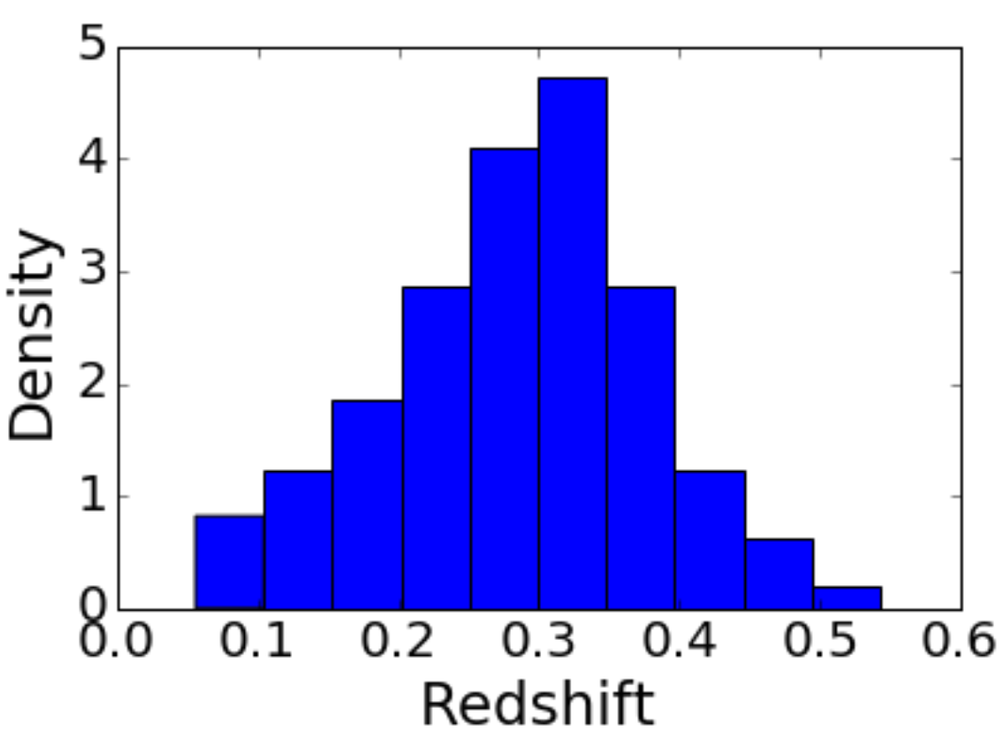
This can lead to **biased** observables

We need to estimate the error distribution!



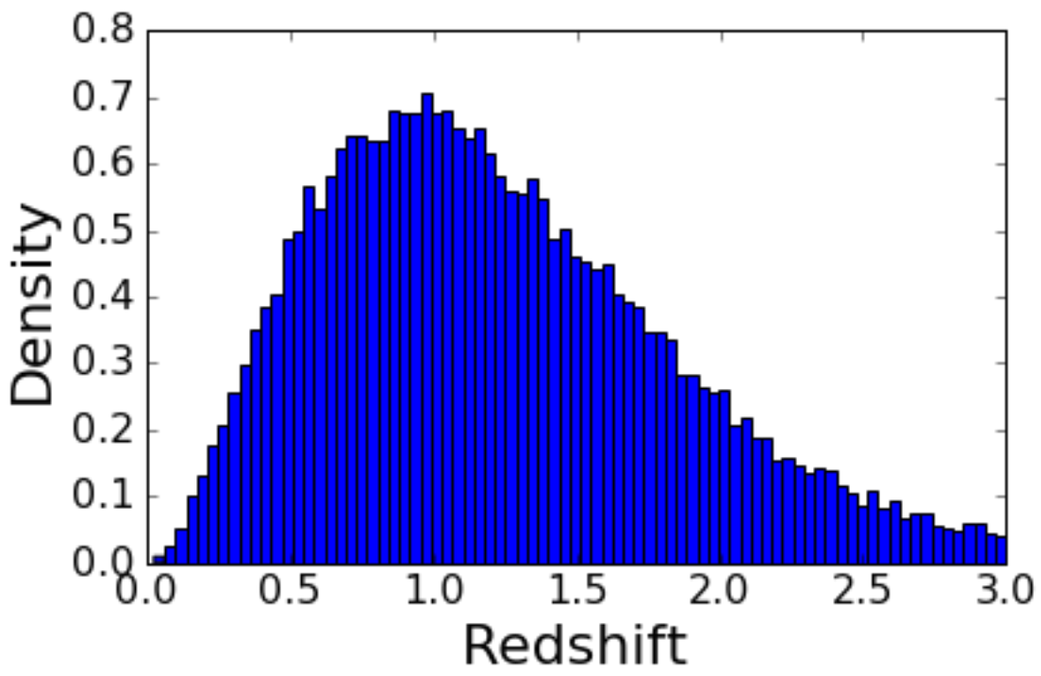
How can we estimate the photoZ error distribution?

Individual Galaxy



STACK

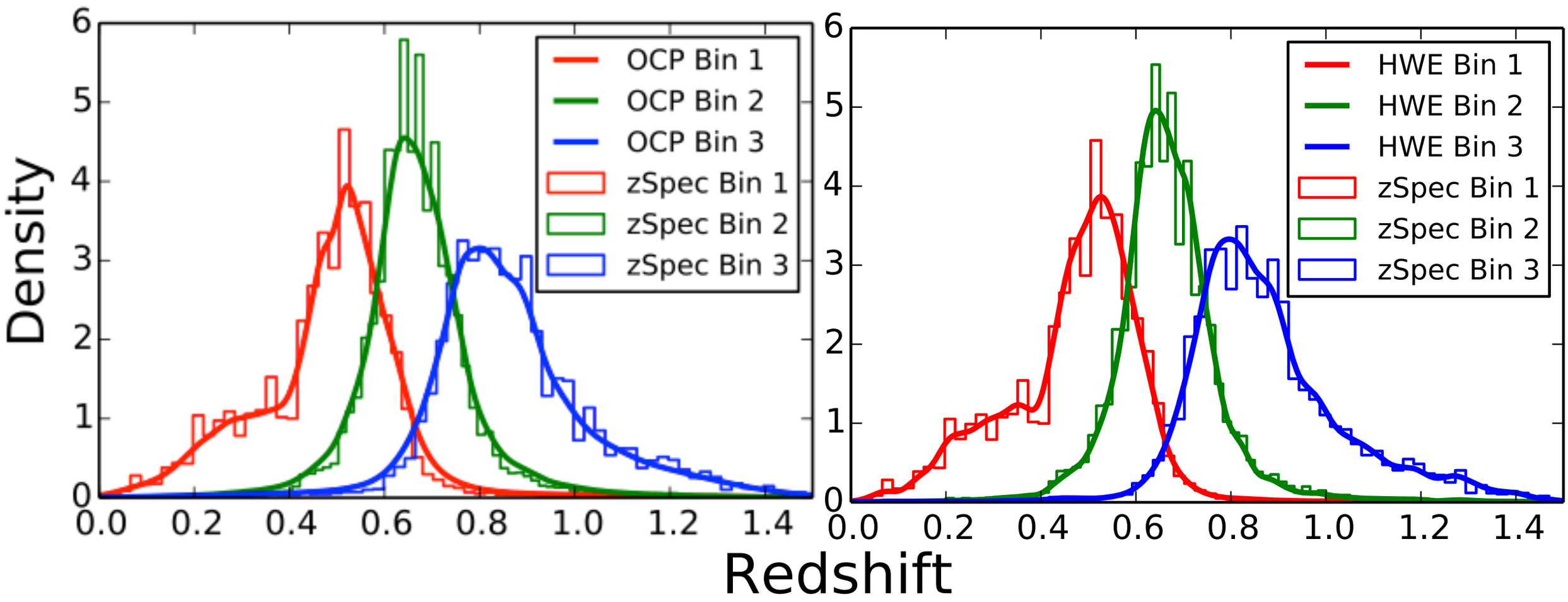
Galaxy Sample



Basic idea: View regression as a classification problem.
The probabilities for class membership
determine the height of the redshift bin.

Problem: Computationally and storage **expensive**

Alternative: The Highest Weight Element (HWE)

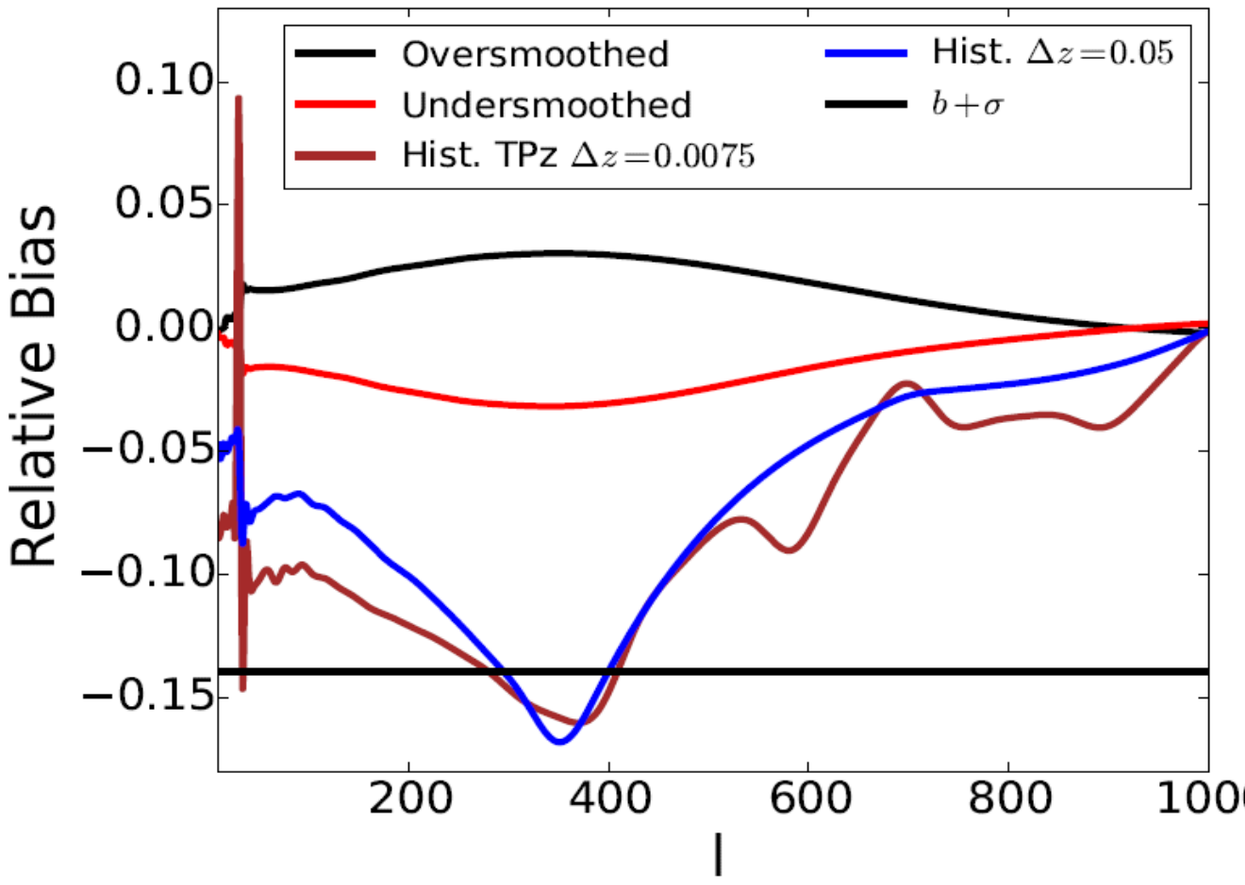


Idea: Find nearest neighbor in color-magnitude space

Remember: We are primarily interested in the **sample PDF**

The HWE is extremely **fast** to compute and extremely **storage efficient**.

Bandwidth selection



Density estimate evaluated on spectroscopic calibration data!

Large samples difficult to obtain (here 17000 objects)

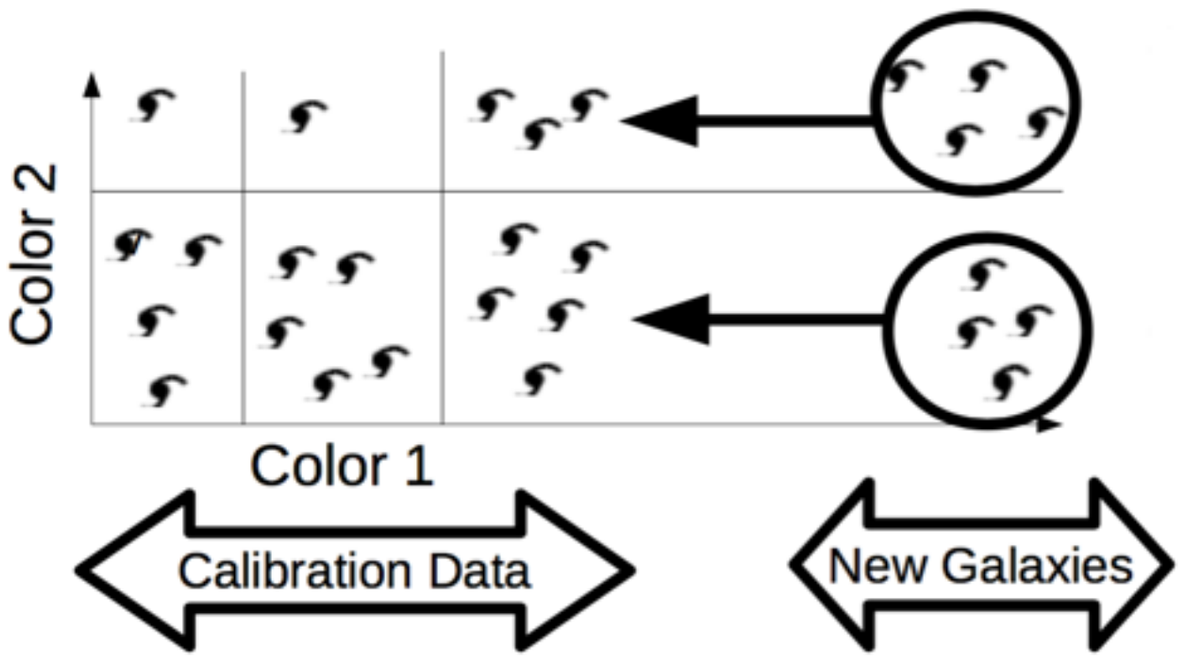
Bandwidth selection
with Silvermans „rule of thumb“
bw = 0.031
Oversmoothed = 1.2*bw
Undersmoothed = 0.8*bw

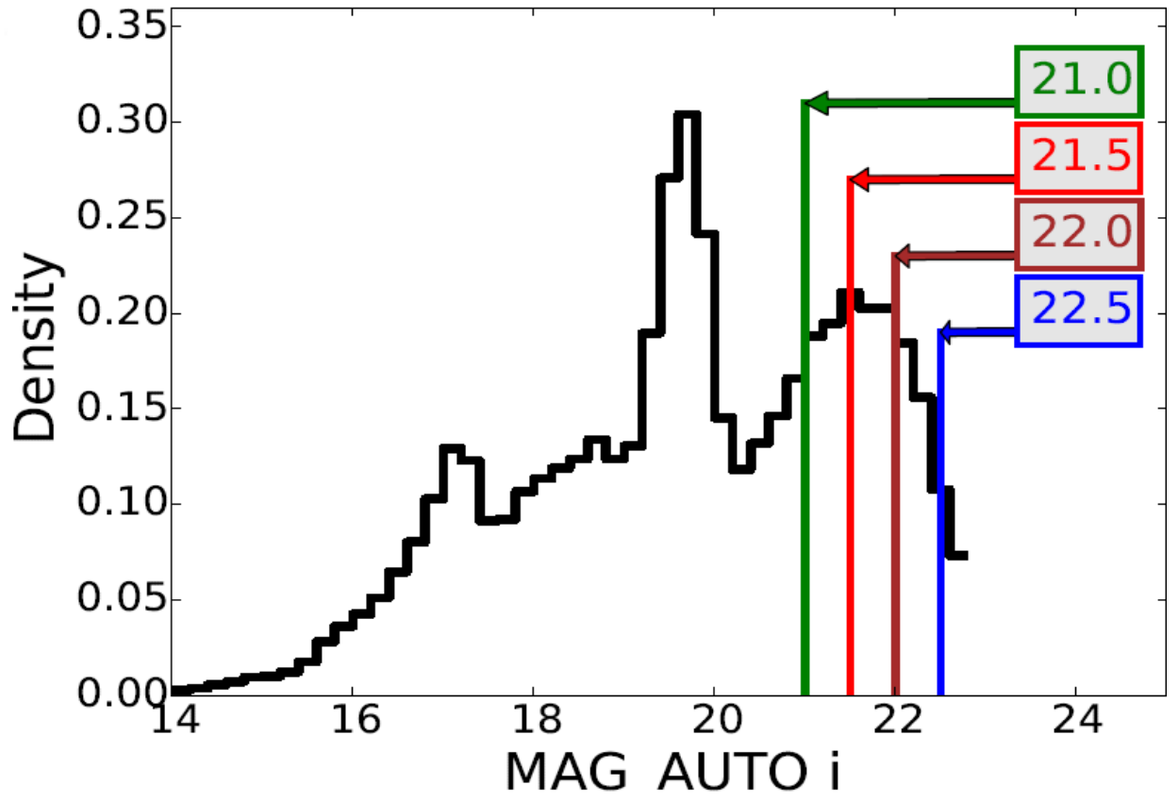
Bandwidth selection alone can create a bias at 1-2% level

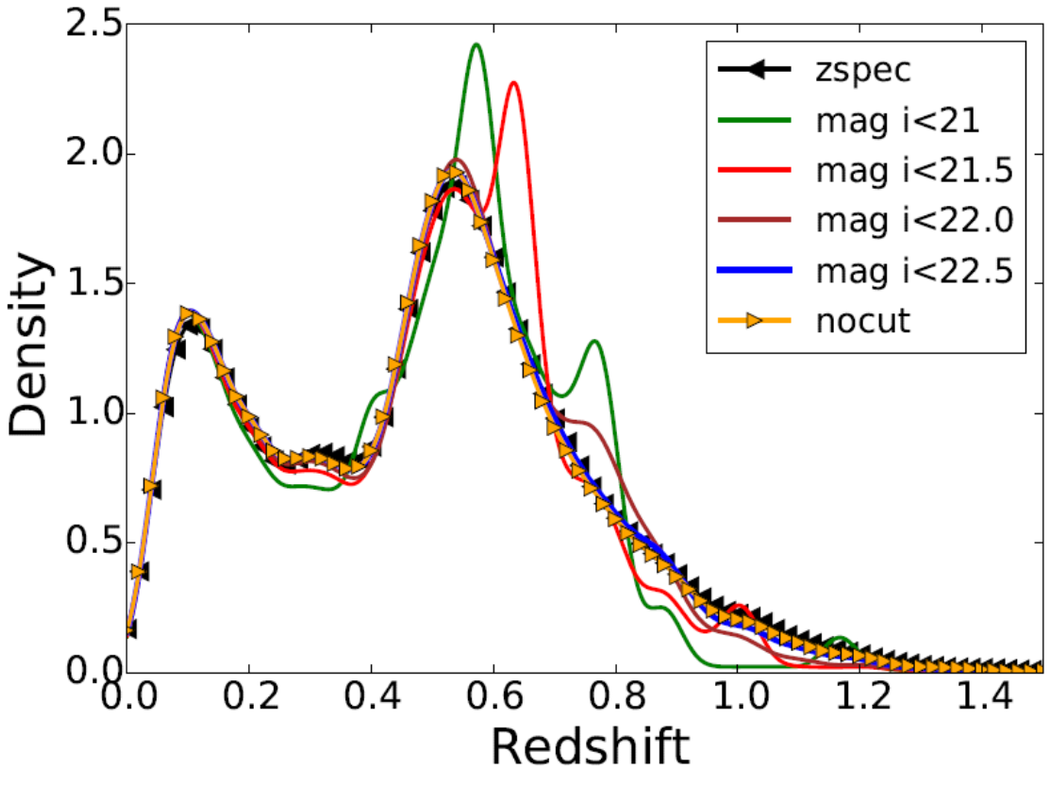
Higher if lensing weights are included

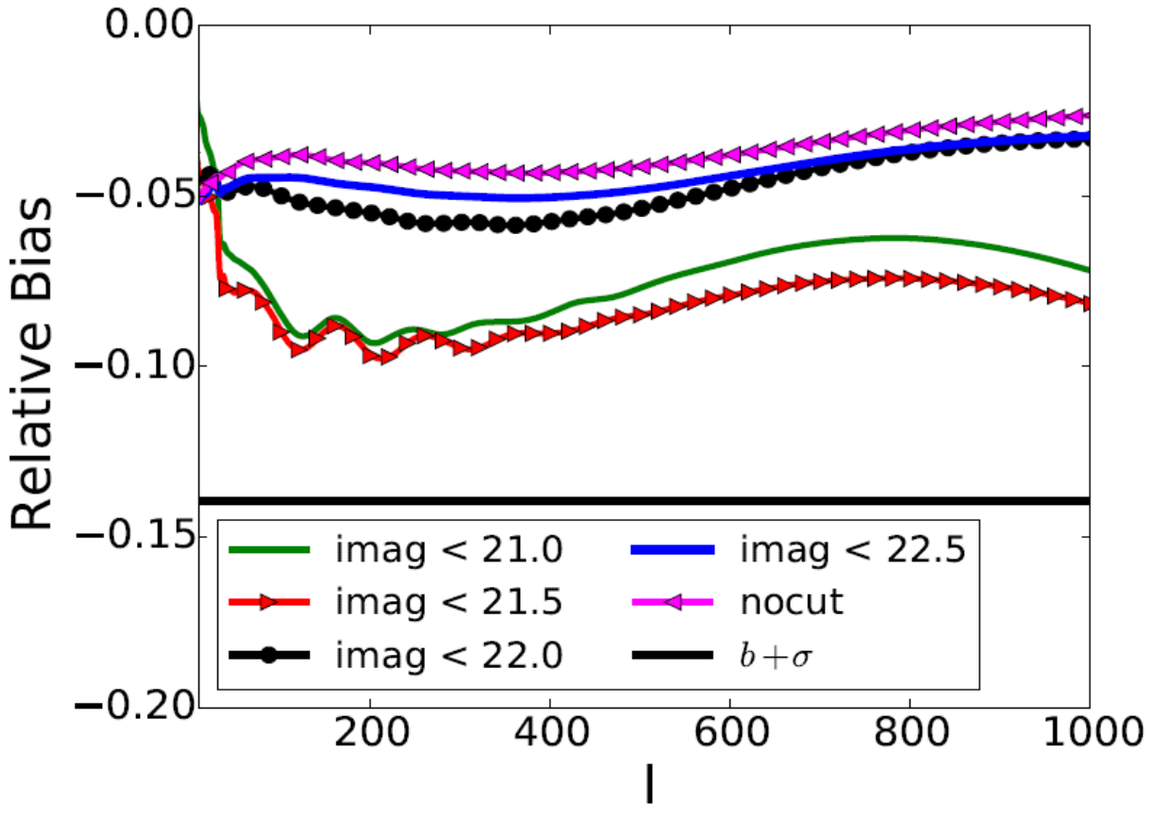
Sample selection bias

apply magnitude cuts to the calibration sample

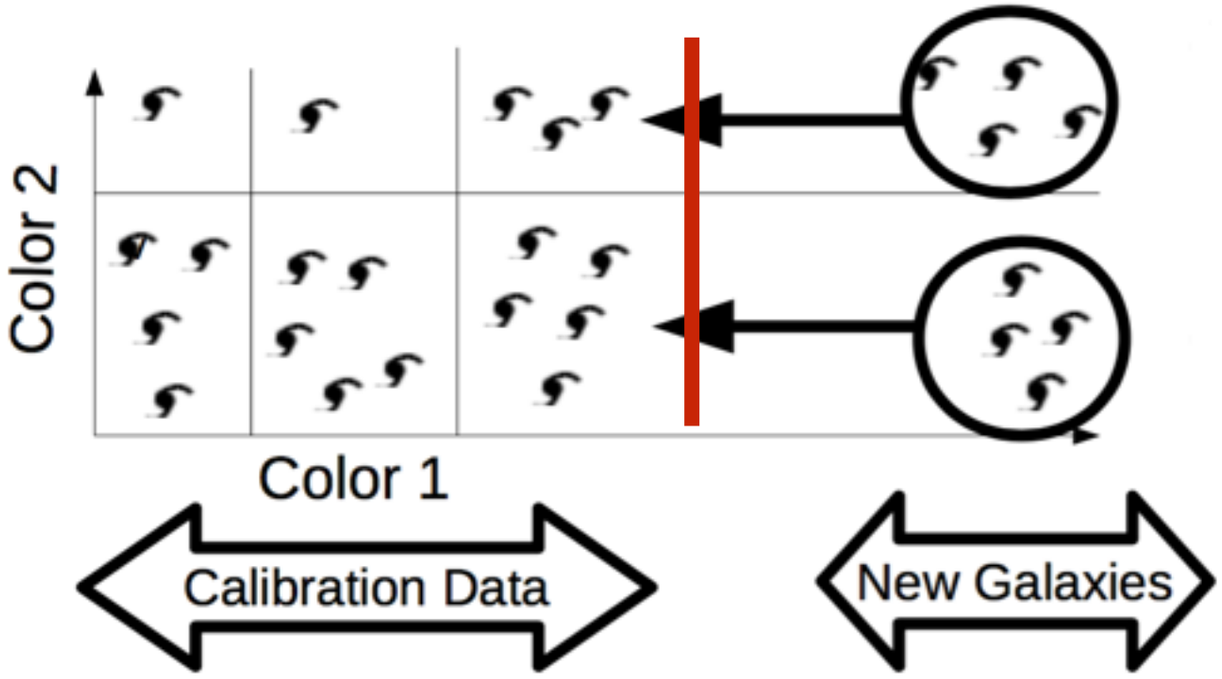






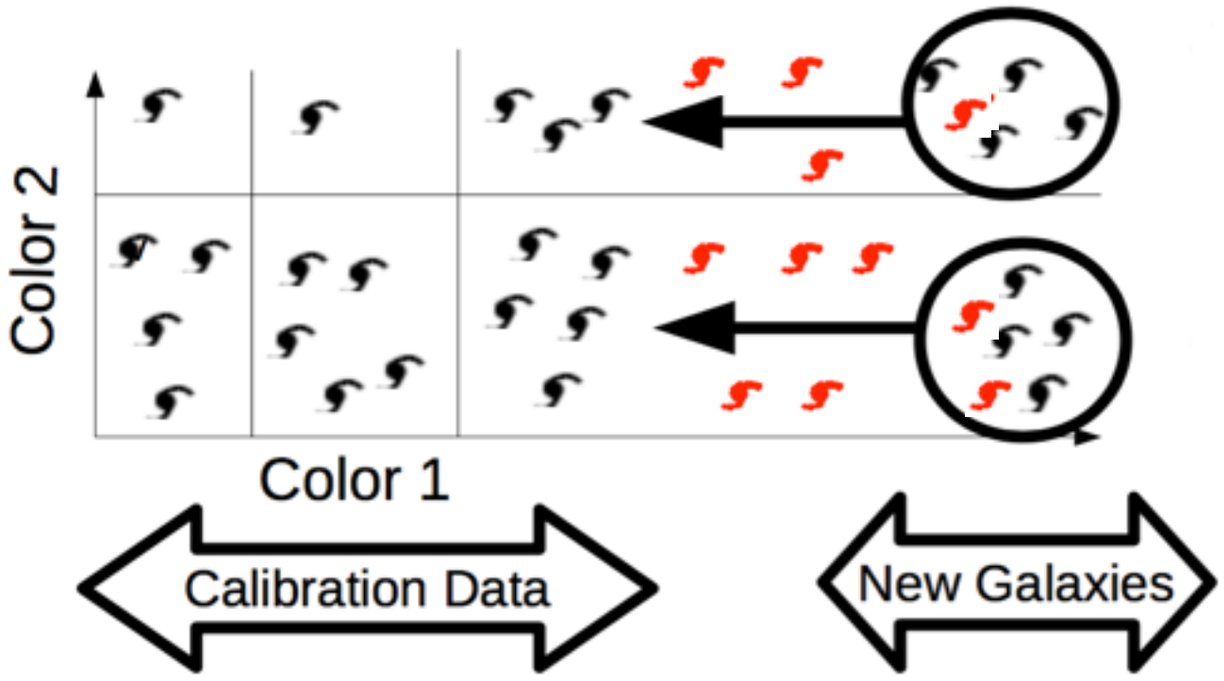


How do we deal with this problem?



Remove galaxies with incomplete spectral calibration

We loose data



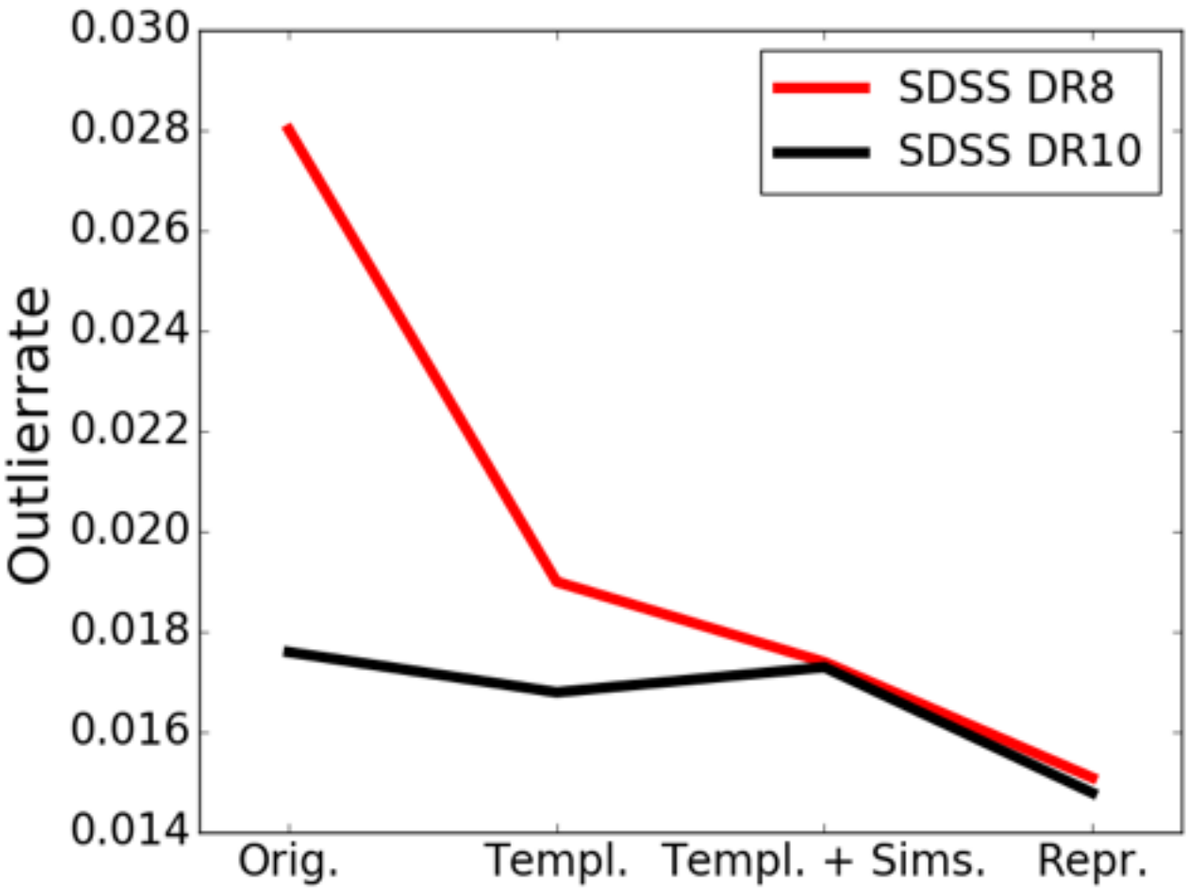
Augment training set: put galaxies artificially at a higher redshift

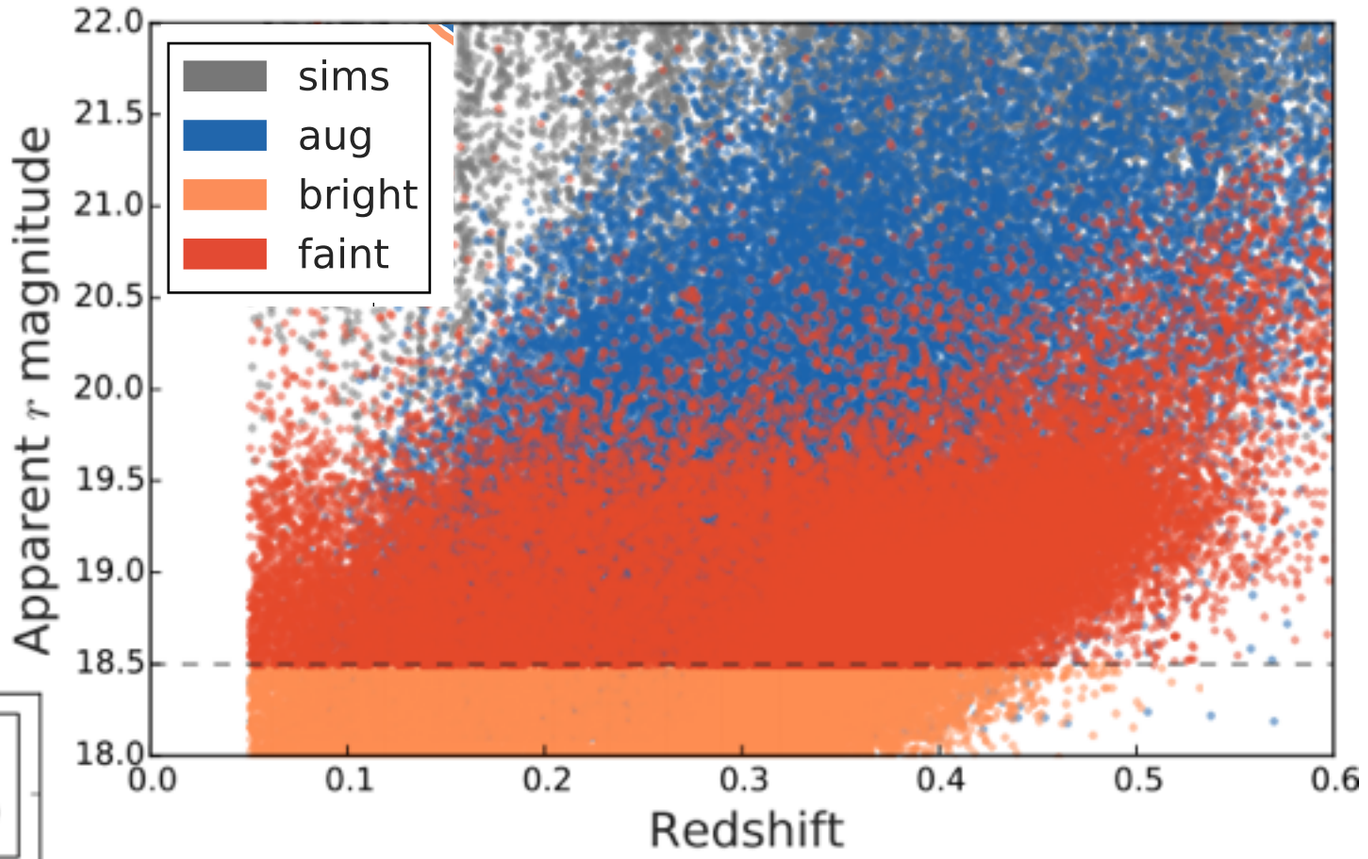
Model guided Extrapolation

Data augmentation

Add artificial galaxies to the calibration data

Combine model fitting (templates) and simulations with ML





Hoyle et al. 2015

arXiv:1501.06759

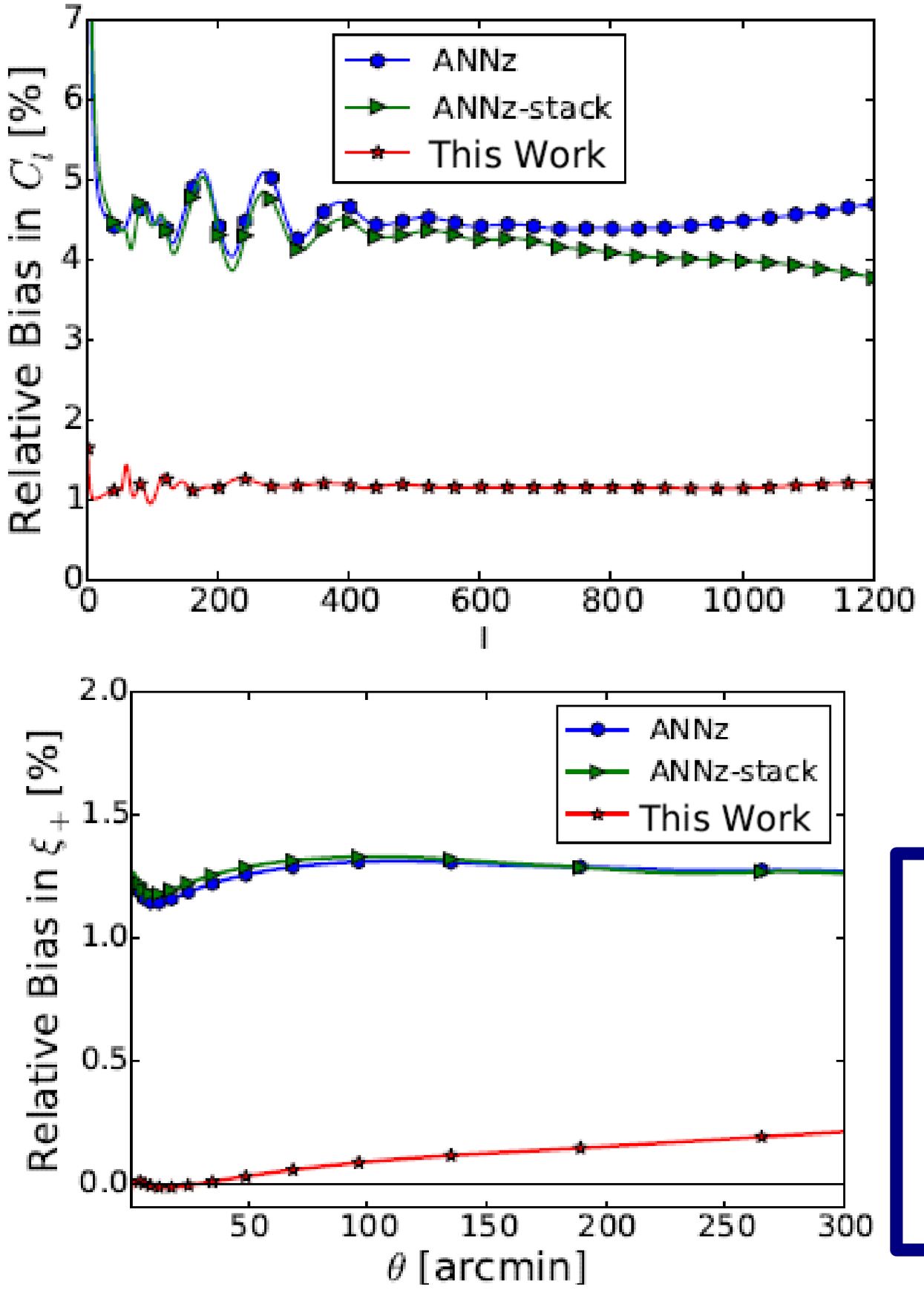
Extrapolate into the unknown!

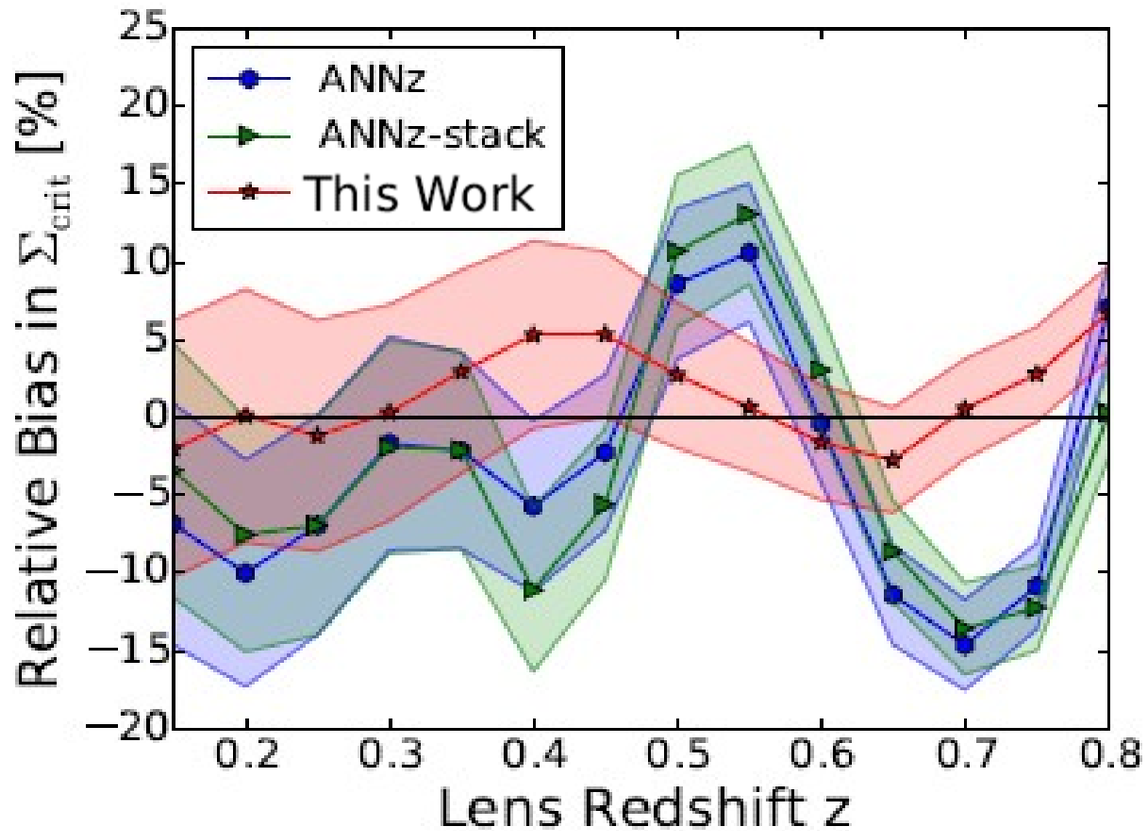
Outlierrate $\mathrm{frac}\left(\left|\frac{z_{\mathrm{phot}} - z_{\mathrm{spec}}}{1 + z_{\mathrm{spec}}}\right| > 0.15\right)$

Summary

- Inaccurate photoZ can **bias cosmological observables**
- We can accurately and efficiently estimate **redshift distributions** with representative calibration data
- We are challenged by various sources of systematic error:
  - **Missing spectra** for faint objects
  - Bandwidth selection
  - Variations between fields

  These effects are **not understood** well enough!



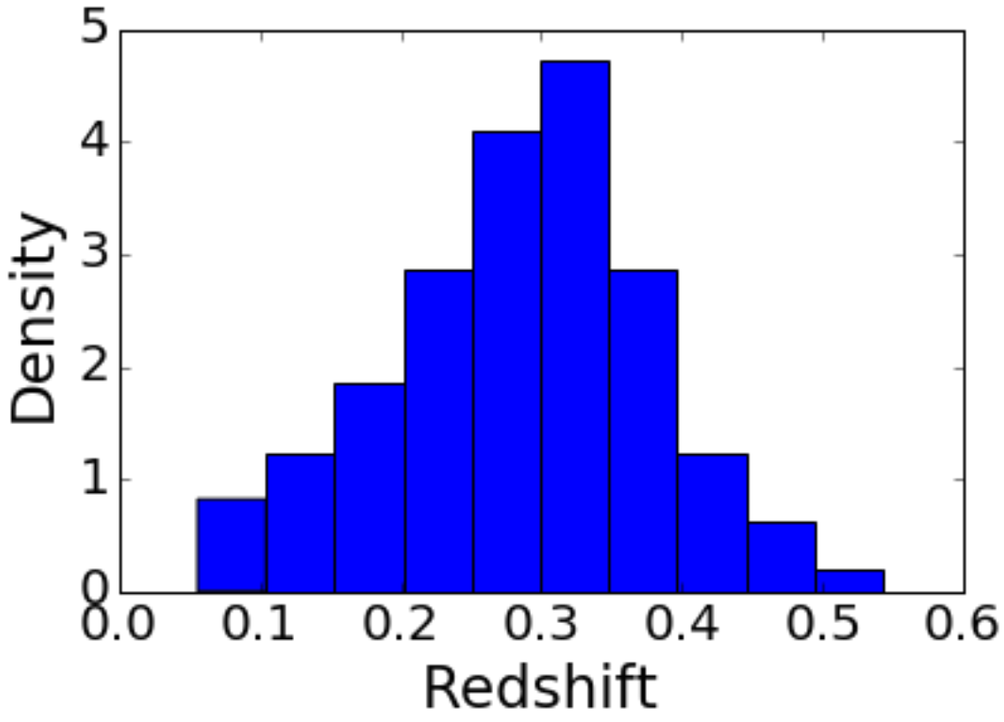


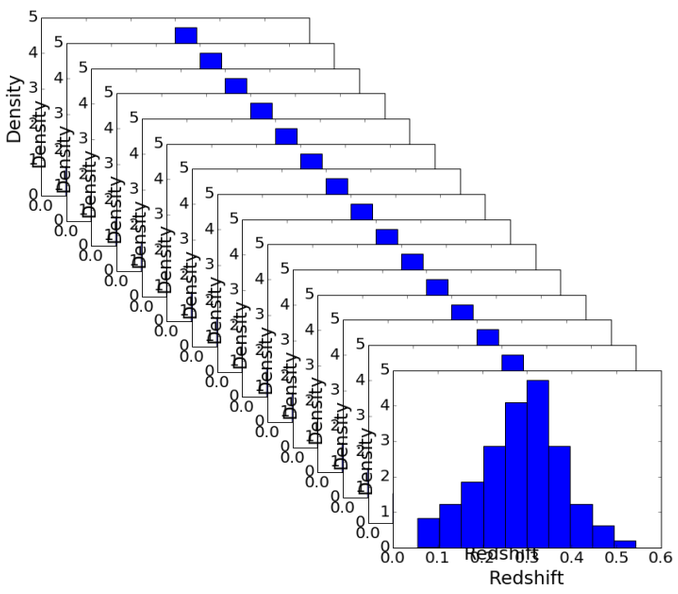
$$Relative\ Bias(X) = \frac{X_{phot} - X_{spec}}{X_{spec}}$$

Insufficient treatment of PhotoZ error distribution biases cosmological observables

Classify binwise probability of redshift bin membership equals bin height

Individual galaxy distribution





Add up distributions of all objects

Sample distribution