Machine learning in Astronomy and Cosmology

Ben Hoyle
University Observatory Munich, Germany
Max Plank for Extragalactic astrophysics

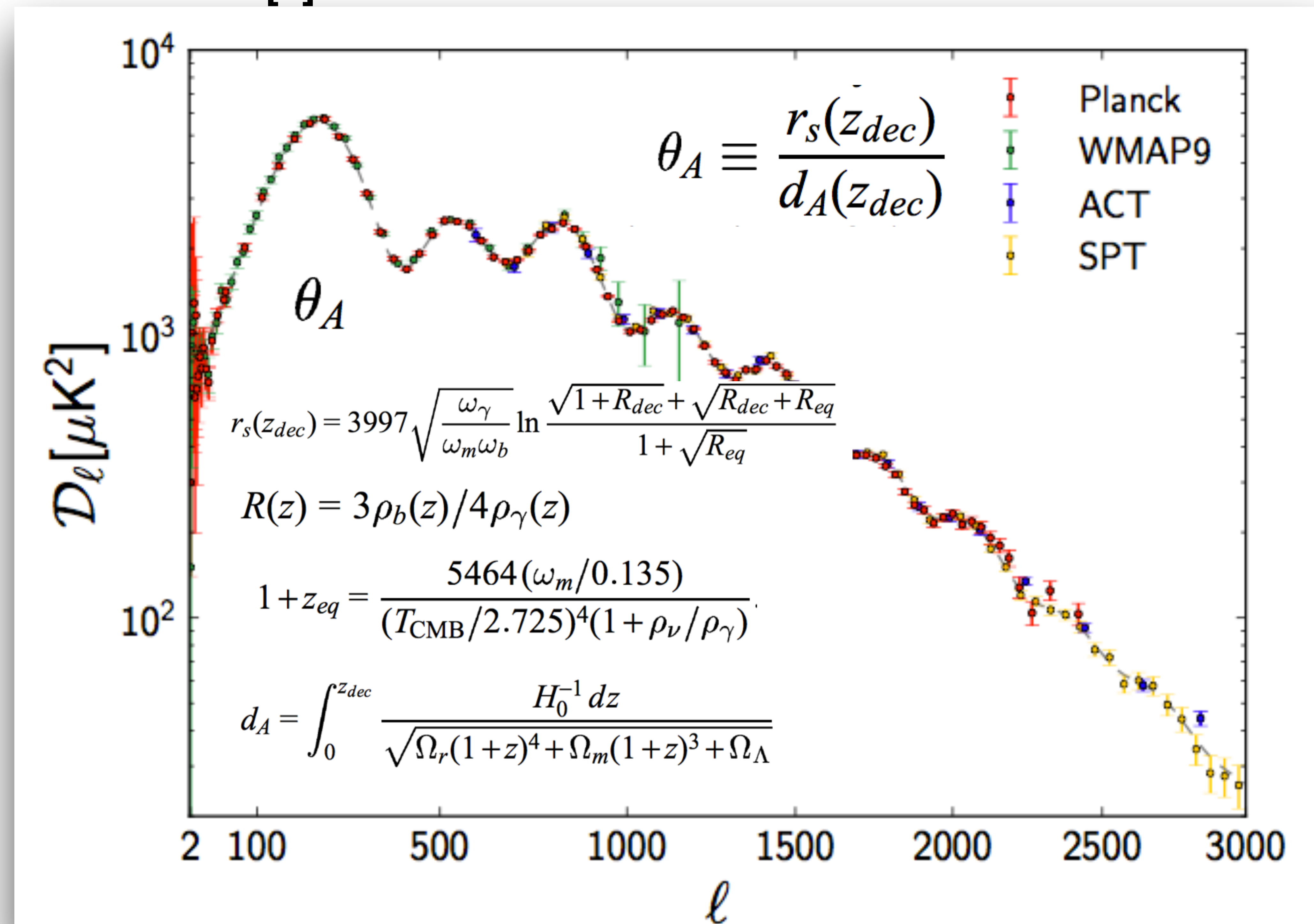
Max-Planck-Institut für extraterrestrische Physik

Collaborators: J. Wolf, R. Lohnmeyer, Suryarao Bethapudi
& Dark Energy Survey, Euclid OUPHZ

Remote talk: IIT Hyderabad, Kandi, India
& USM Munich Germany 23/11/2017

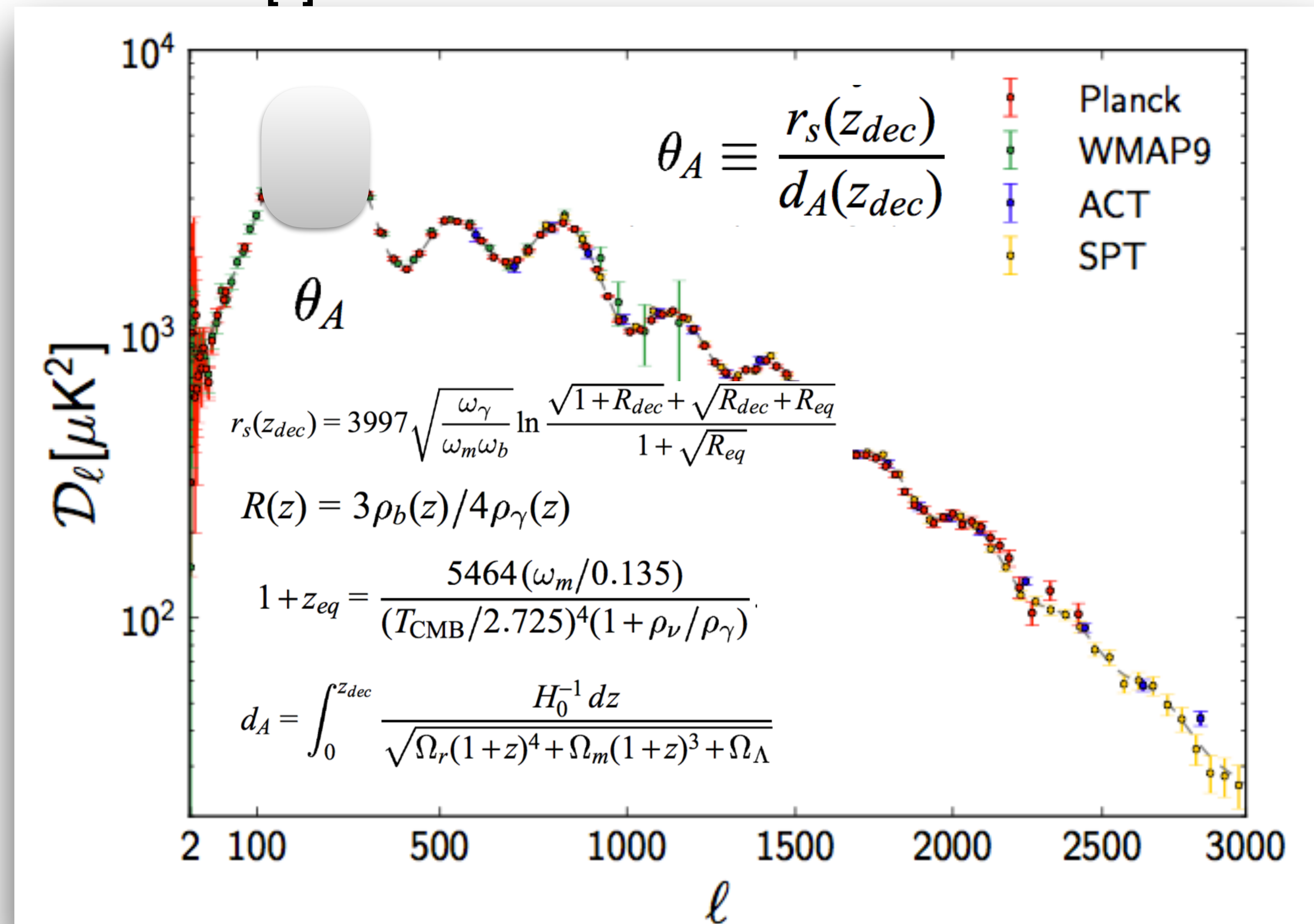
When/Why is Machine Learning suited to astrophysics/ cosmology?

When we are in a "data poor" and "model rich" regime e.g. Correlation function analysis of CMB maps, we should not use ML, rather rely on the predictive model [s].



When/Why is Machine Learning suited to astrophysics/ cosmology?

When we are in a "data poor" and "model rich" regime e.g. Correlation function analysis of CMB maps, we should not use ML, rather rely on the predictive model [s].



When/why is Machine Learning suited to astrophysics/ cosmology?

When we are in a “data poor” and “model rich” regime e.g. Correlation function analysis of CMB maps, we should not use ML, rather rely on the predictive model [s].

When we are in a “data rich” and “model poor” regime, and still want to approximate some model y=f(x); we can use machine learning to learn (or fit) an arbitrarily complex model (e.g. non-functional curves) of the data.

When/why is Machine Learning suited to astrophysics/ cosmology?

When we are in a "data poor" and "model rich" regime e.g. Correlation function analysis of CMB maps, we should not use ML, rather rely on the predictive model [s].

When we are in a "data rich" and "model poor" regime, and still want to approximate some model y=f(x); we can use machine learning to learn (or fit) an arbitrarily complex model (e.g. non-functional curves) of the data.

Cosmology is firmly in the data "rich" regime:

1) SDSS has 100 million photometrically identified objects (stars/galaxies) and 3 million spectroscopic "truth" values, for e.g. redshift, and galaxy/ stellar type

2) DES has 300 million objects with photometry, and ~400k objects with spectra

3) Gaia has >1 billion sources [stellar maps of the Milky Way]

3) Euclid with have 3 billion objects…

When/why is Machine Learning suited to astrophysics/ cosmology?

When we are in a "data poor" and "model rich" regime e.g. Correlation function analysis of CMB maps, we should not use ML, rather rely on the predictive model [s].

When we are in a "data rich" and "model poor" regime, and still want to approximate some model y=f(x); we can use machine learning to learn (or fit) an arbitrarily complex model (e.g. non-functional curves) of the data.

Cosmology is firmly in the data "rich" regime:
1) SDSS has 100 million photometrically identified objects (stars/galaxies) and spectroscopic "truth" values, for e.g. redshift, and galaxy/stellar type.

and often in the "model-poor" regime:
1) The exact mapping between galaxies observed in broad photometric bands and their redshift depends on stellar population physics, initial stellar mass functions, local environment, feedback from AGN/SNe, dust extinction,…
2) Is an object found in photometric images a faint star that is far away, or a high redshift galaxy?

Use machine learning to approximate the mapping:
redshift = f(photometric properties of training sample)
f(photometric properties of 3 billion galaxies) => photometric redshift

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

Summary/Conclusions

Why are photo-z's important?

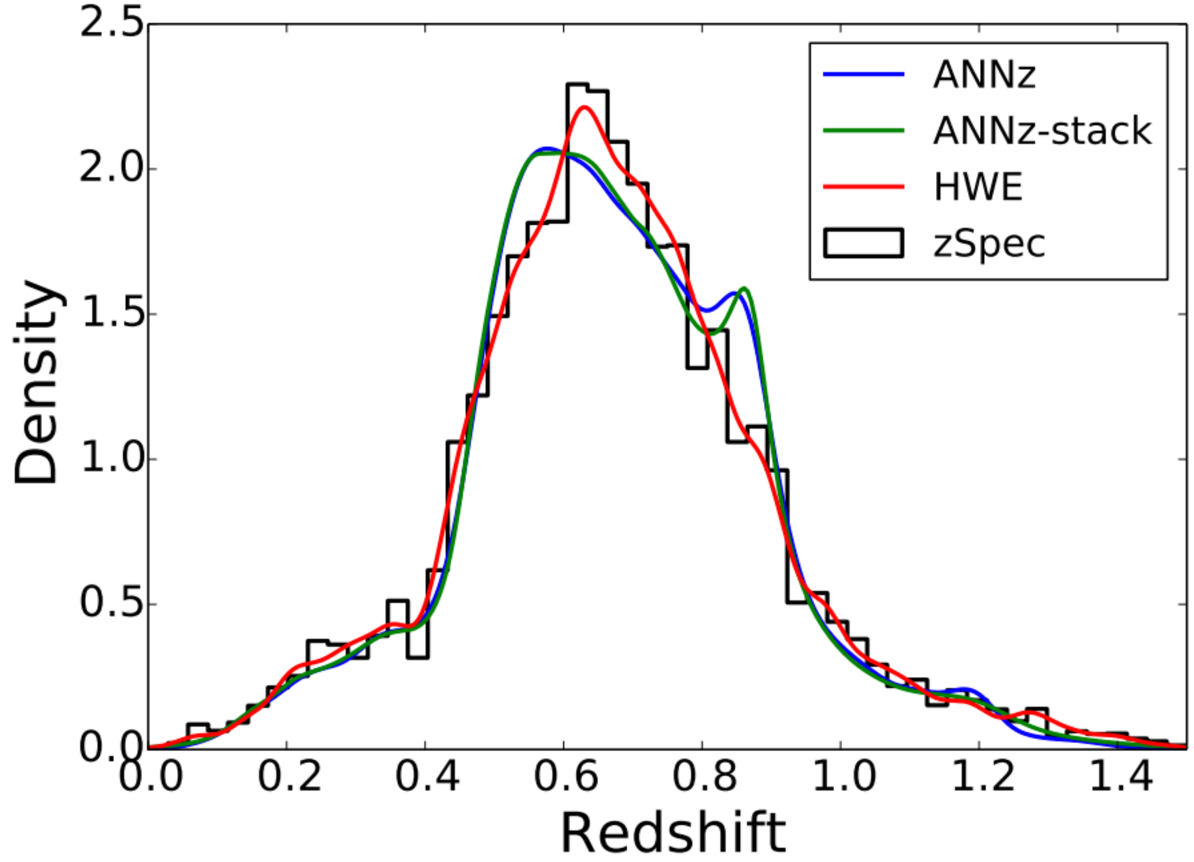


Figure 5. Sample PDF estimated using ANNz and the Highest Weight Element. The histogram shows the true spectroscopic redshift distribution.

Why are photo-z's important?

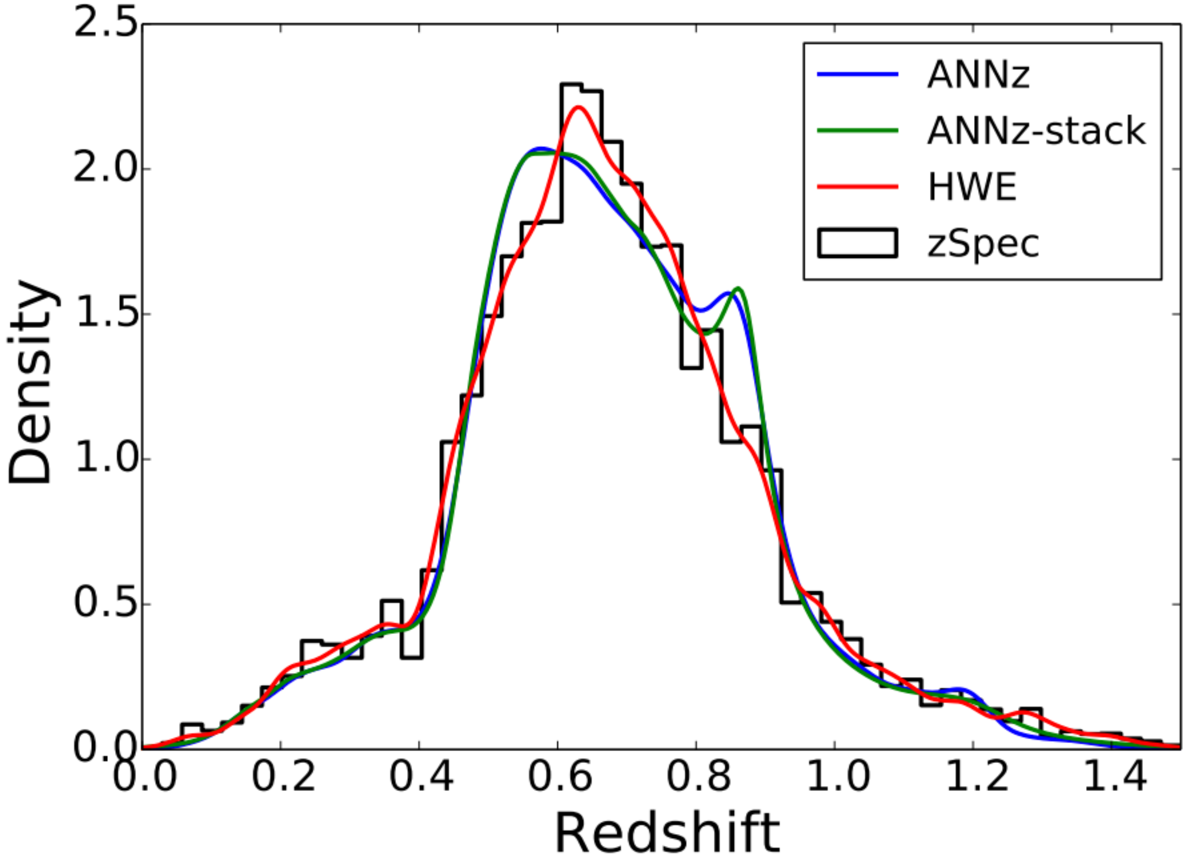


Figure 5. Sample PDF estimated using ANNz and the Highest Weight Element. The histogram shows the true spectroscopic redshift distribution.

$$Rel.Bias = \frac{C_l(z_{spec}) - C_l(z_{photo})}{C_l(z_{specz})}$$

Why are photo-z's important?

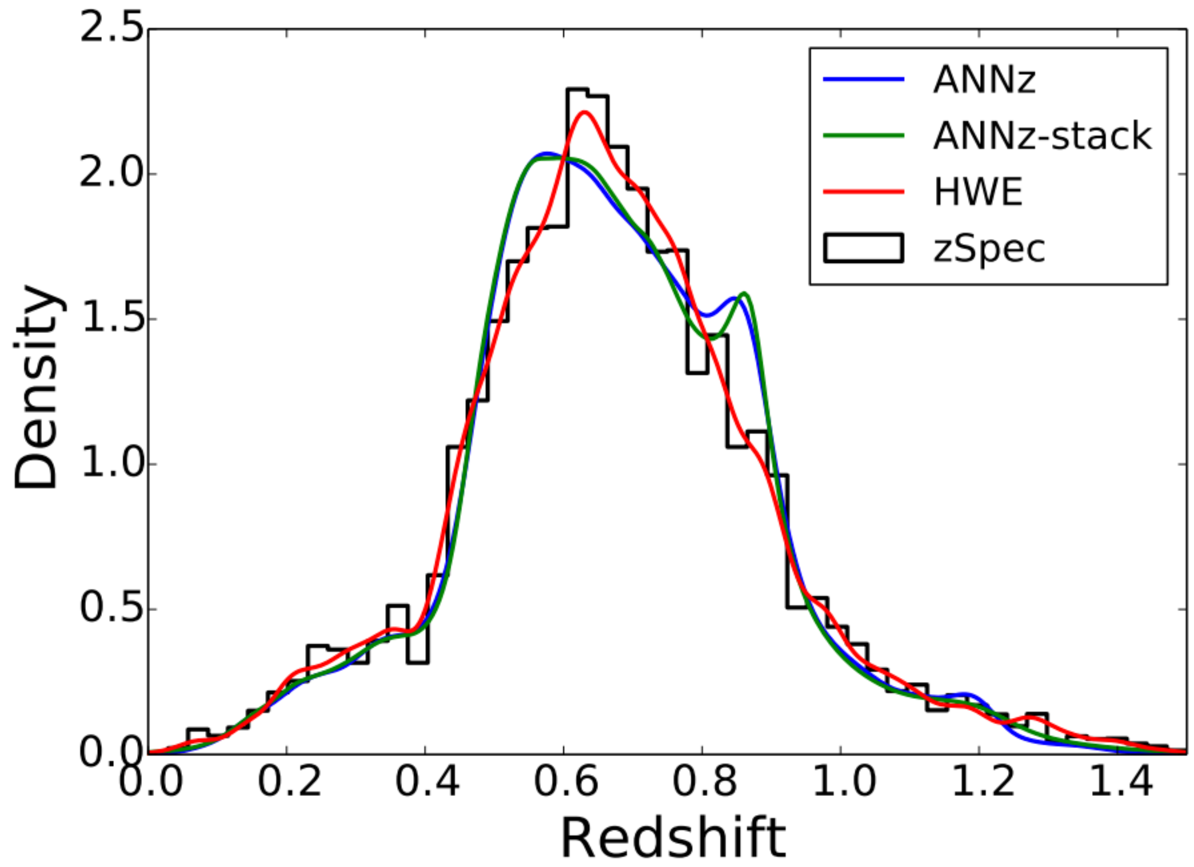


Figure 5. Sample PDF estimated using ANNz and the Highest Weight Element. The histogram shows the true spectroscopic redshift distribution.

$$Rel.Bias = \frac{C_l(z_{spec}) - C_l(z_{photo})}{C_l(z_{specz})}$$

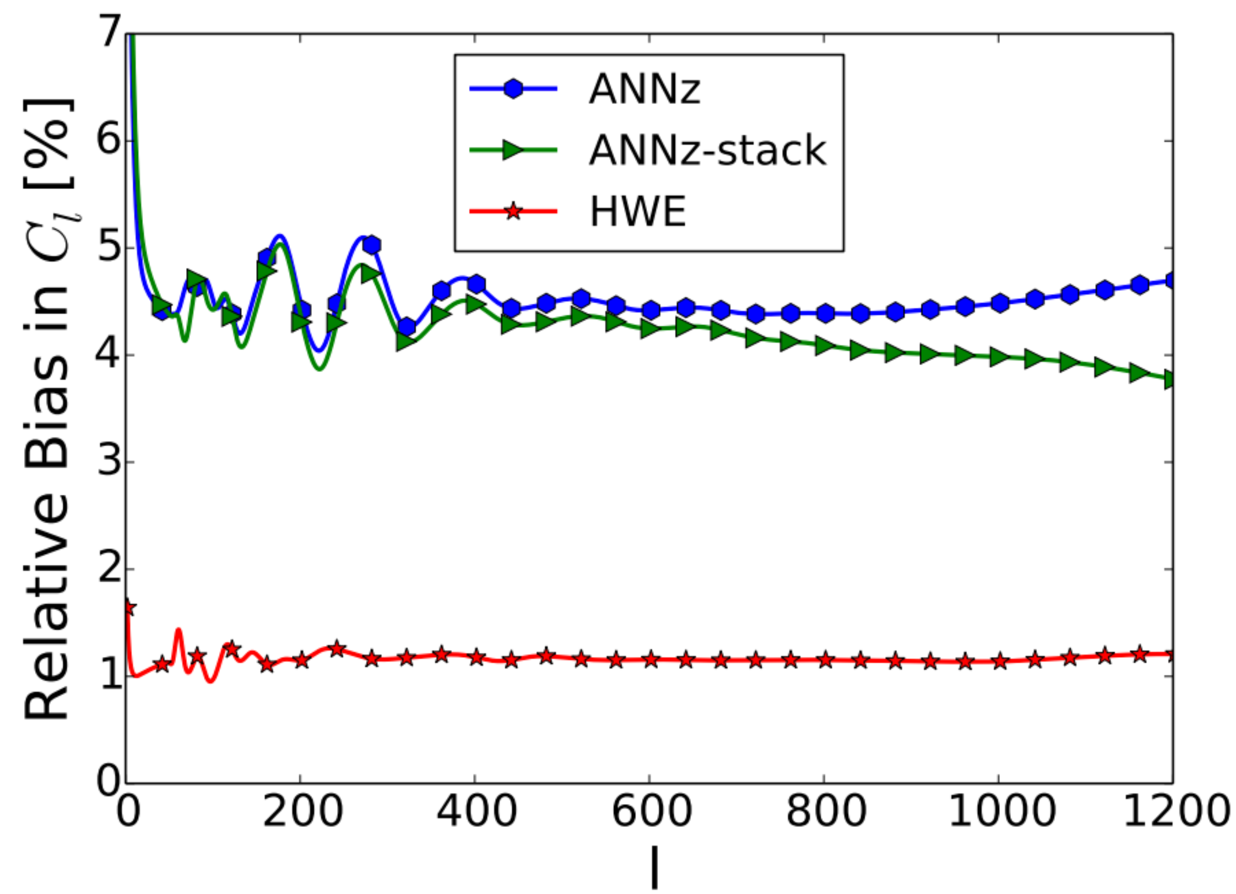


Figure 9. Bias in the angular correlation power spectrum obtained for different estimates for the sample PDF. We restrict the comparison to $\ell < 1200$.

Rau, BH et al 2015

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

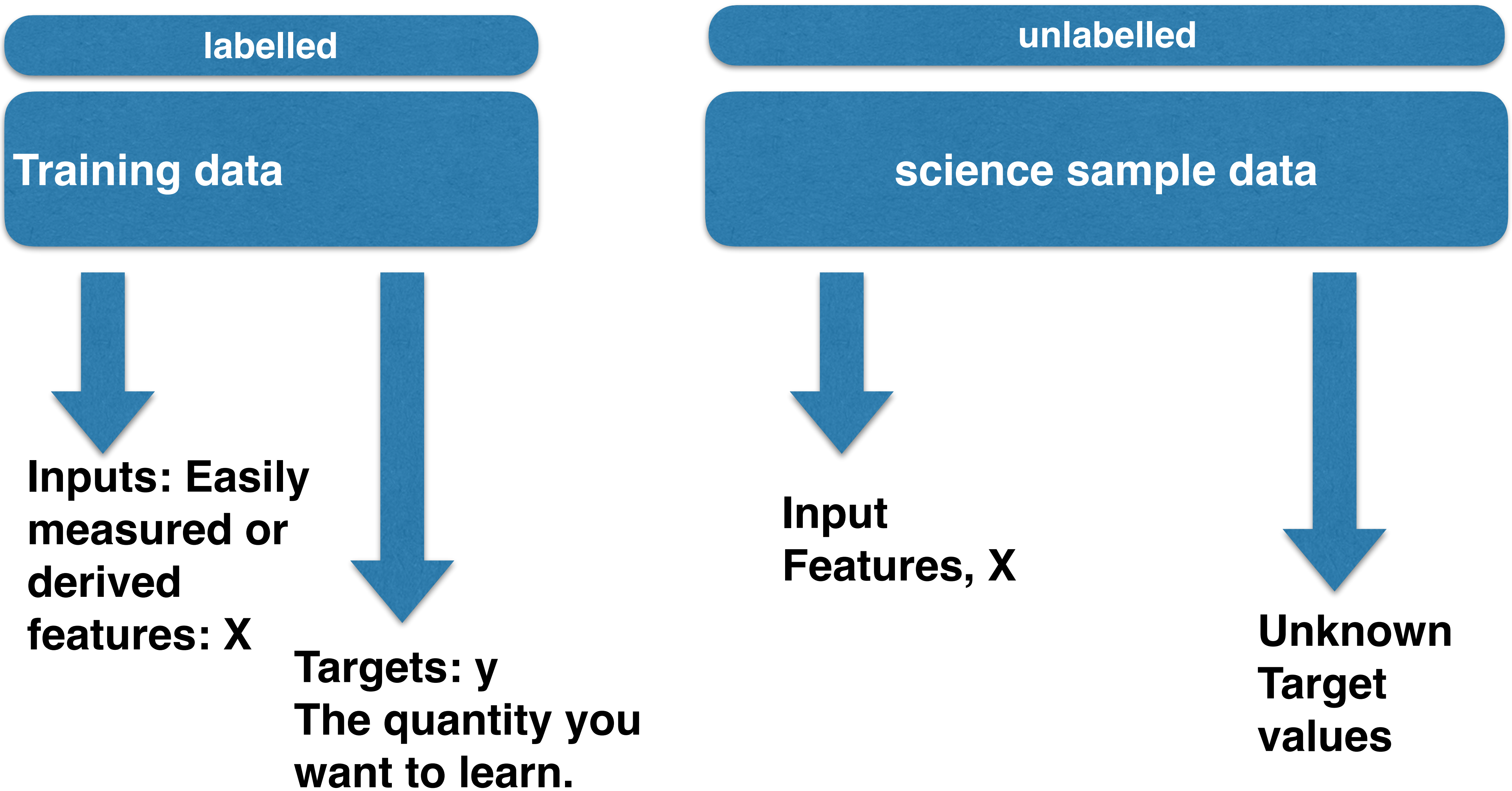
Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

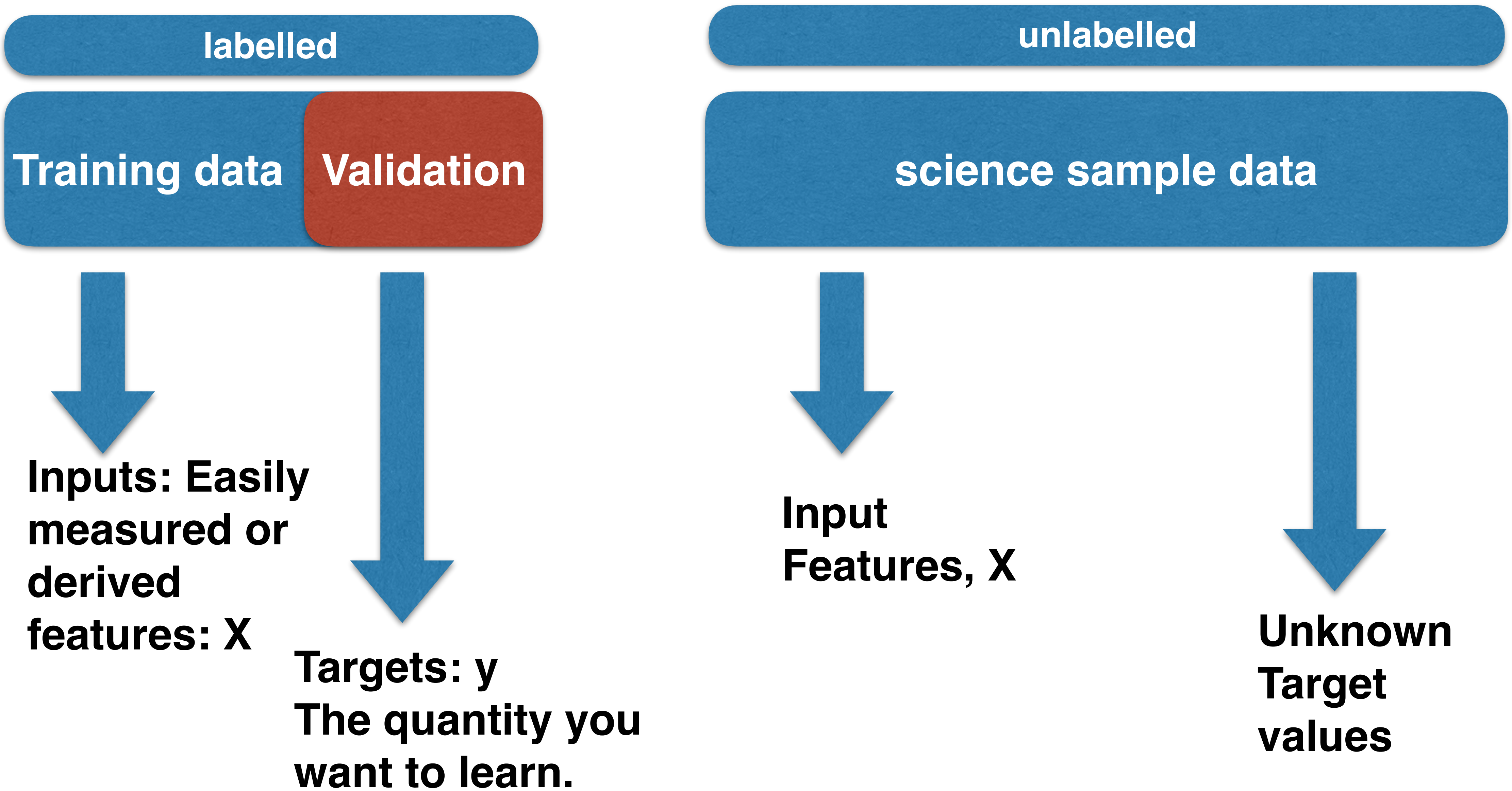
Summary/Conclusions

Supervised Machine learning framework



$$y_{train} \approx \hat{y}_{train} = f(X_{train}) \qquad \hat{y}_{sci-s} = f(X_{sci-s})$$

Supervised Machine learning framework



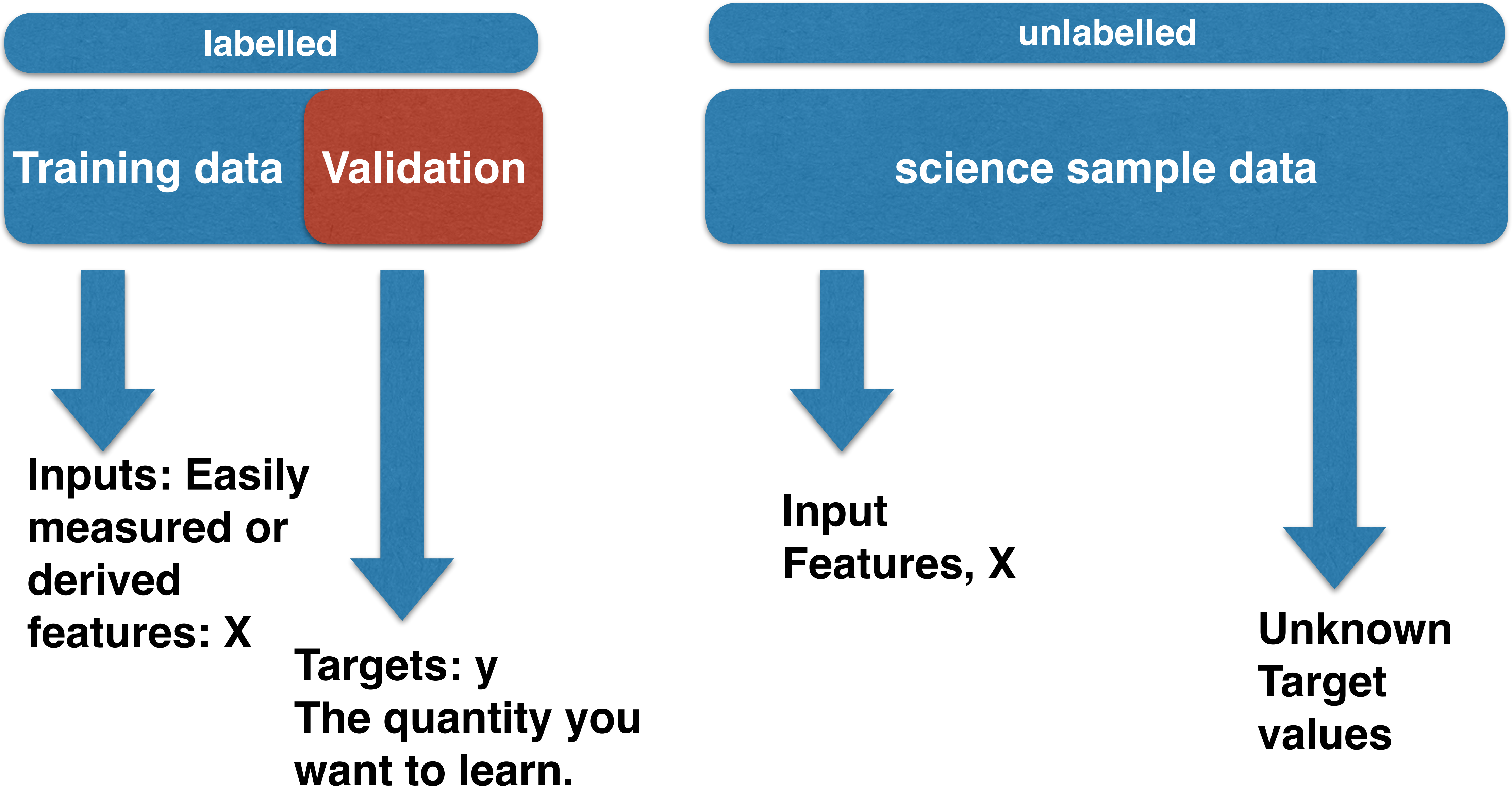
$$y_{train} \approx \hat{y}_{train} = f(X_{train})$$

$$\hat{y}_{sci-s} = f(X_{sci-s})$$

Expected Error on prediction

$$\Delta = \hat{y}_{x-val} - y_{x-val}$$

Supervised Machine learning framework



$$y_{train} \approx \hat{y}_{train} = f(X_{train}) \qquad \hat{y}_{sci-s} = f(X_{sci-s})$$

Expected Error on prediction

$$\Delta = \hat{y}_{x-val} - y_{x-val}$$

If the validation data is not representative of the science sample data, you can't use machine learning (or any analysis!) to quantify how the predictions will behave on the science sample.

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

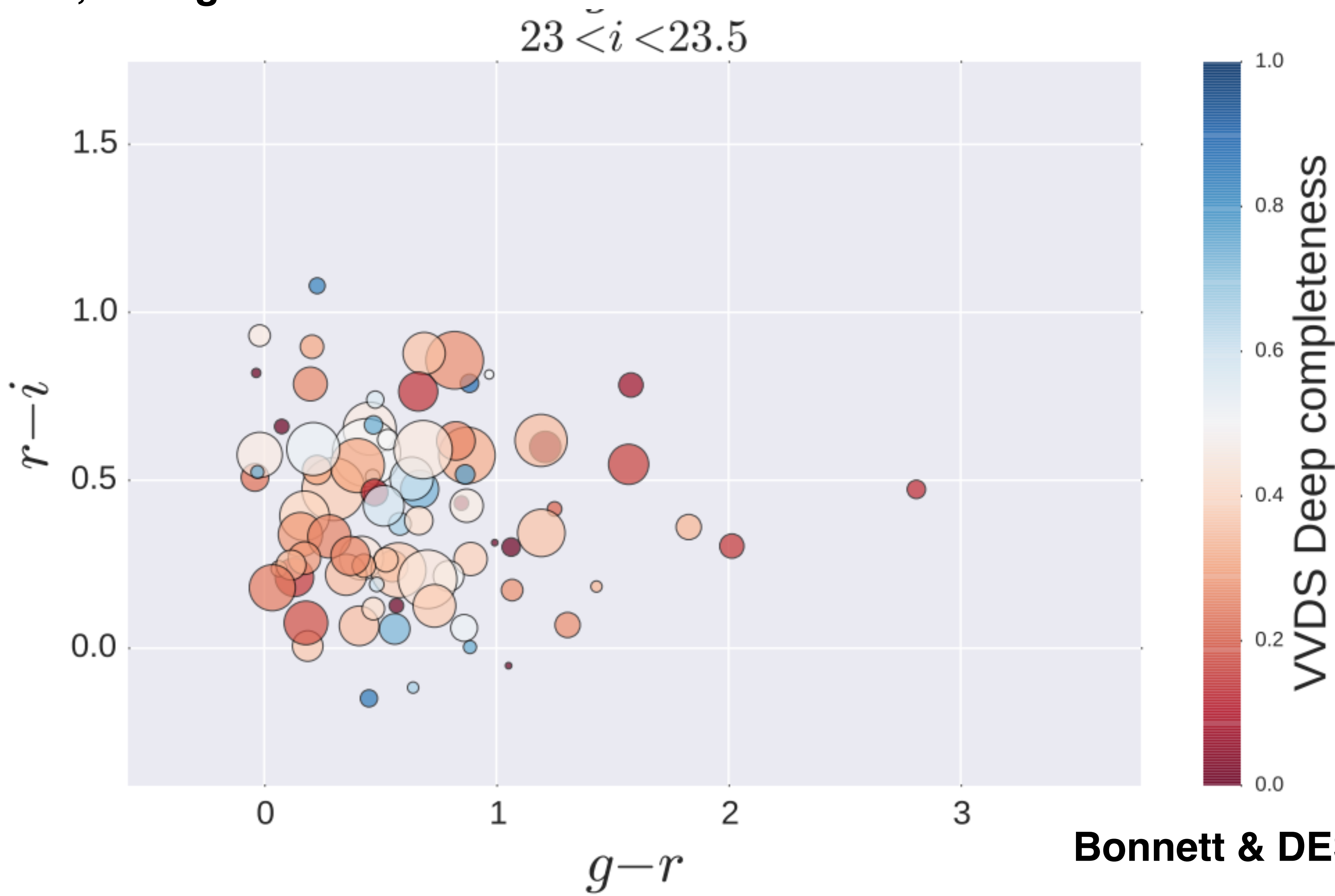
Recent, novel applications of ML

Summary/Conclusions

Photometric redshifts: current challenges

Training/validation/[test] (i.e. all labelled data) not representative of the science sample data.

- **Almost impossible/very time expensive to get spec-z measurements of high redshift, faint galaxies.**

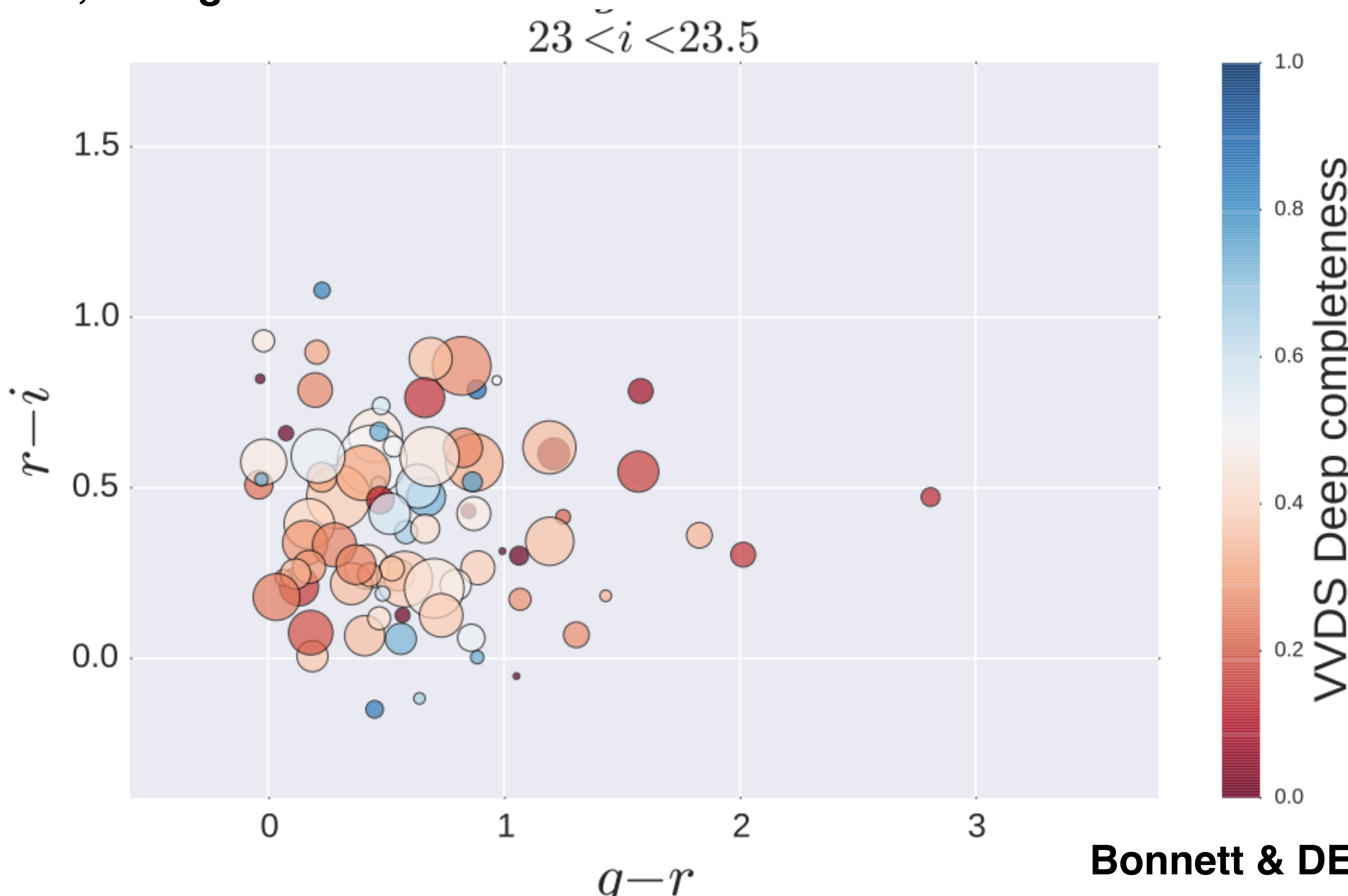


Bonnett & DES SV 2015

Photometric redshifts: current challenges

Training/validation/[test] (i.e. all labelled data) not representative of the science sample data.

- **Almost impossible/very time expensive to get spec-z measurements of high redshift, faint galaxies.**



Bonnett & DES SV 2015

This leads to incomplete labelled data (spec-z) in the input feature space

- **A covariate shift could fix this…**

Confidence flag induced label biases

The data with a confidence label (spec-z) is biased in the label direction.

We extracted 1-d spectra from simulations (known redshift), added noise. Ask DES/OzDES observers to redshift the spectra and apply a confidence flag.

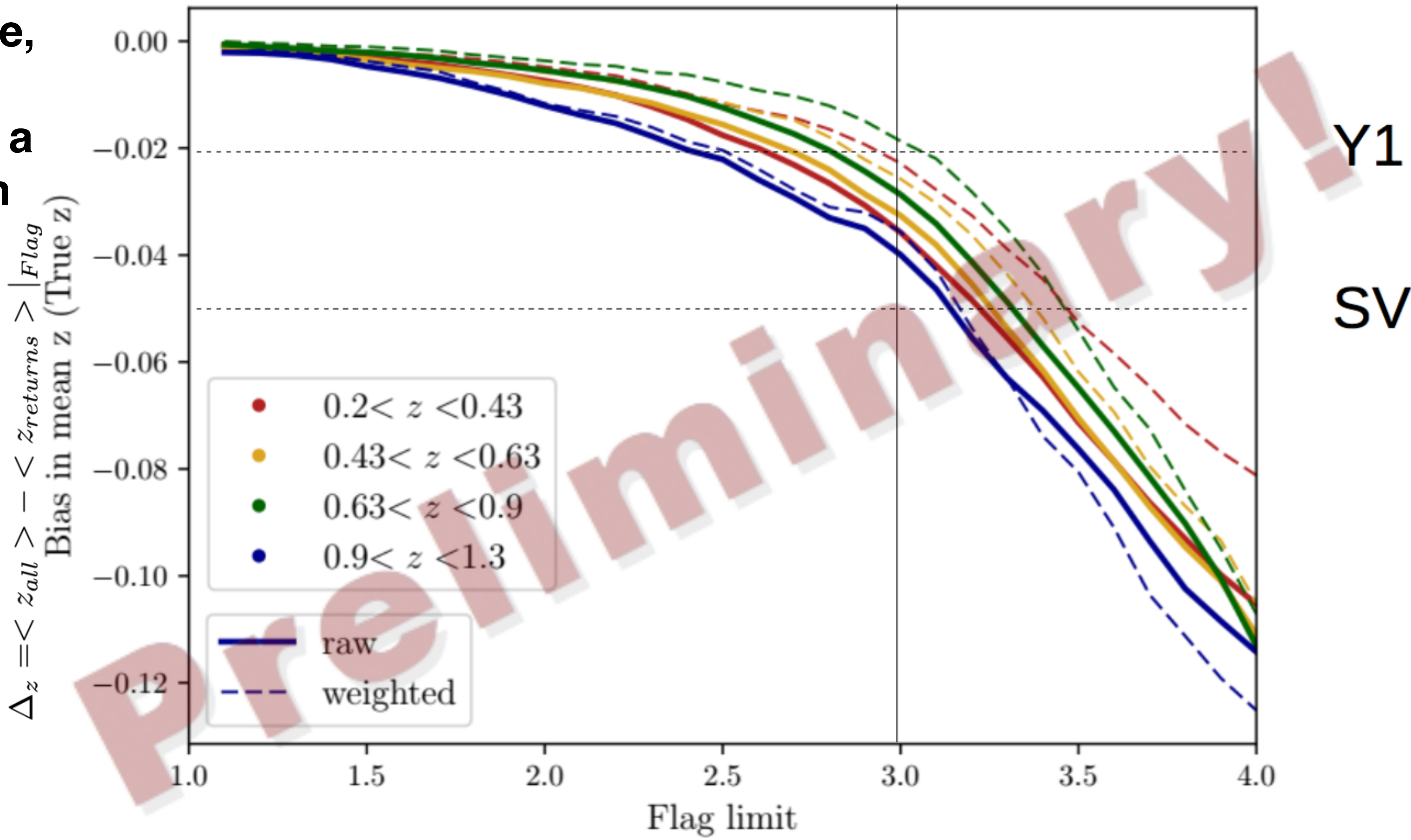
Confidence flag induced label biases

The data with a confidence label (spec-z) is biased in the label direction.

We extracted 1-d spectra from simulations (known redshift), added noise. Ask DES/OzDES observers to redshift the spectra and apply a confidence flag.

We compare the $< z_{returns} > |_{Flag}$ of the returned sample, with the $< z_{all} >$ of the requested sample, as a function of the human assigned confidence flag.

Leads: Will Hartley, Chihway Chang

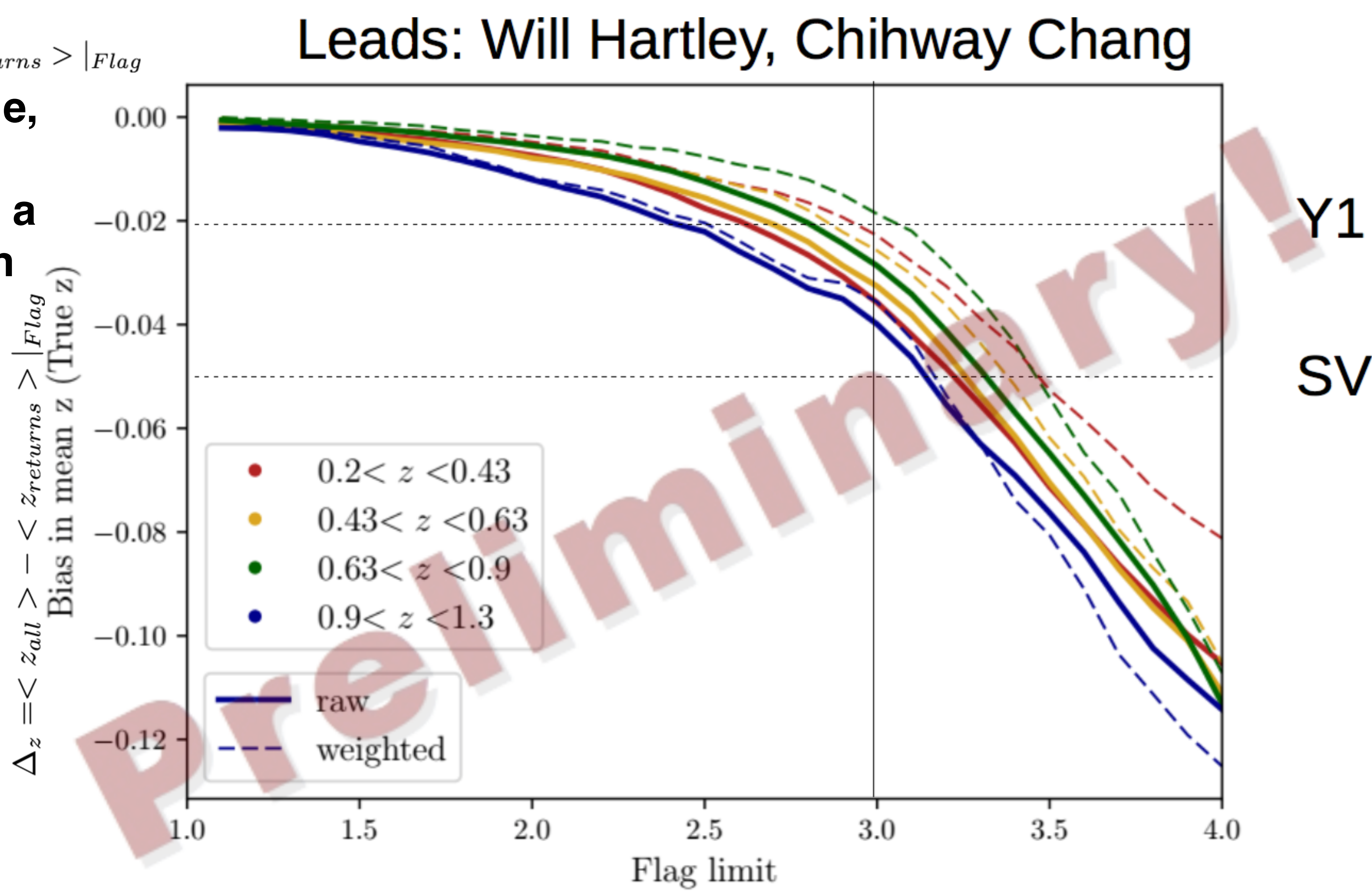


Confidence flag induced label biases

The data with a confidence label (spec-z) is biased in the label direction.

We extracted 1-d spectra from simulations (known redshift), added noise. Ask DES/ OzDES observers to redshift the spectra and apply a confidence flag.

We compare the $< z_{returns} >|_{Flag}$ of the returned sample, with the $< z_{all} >$ of the requested sample, as a function of the human assigned confidence flag.

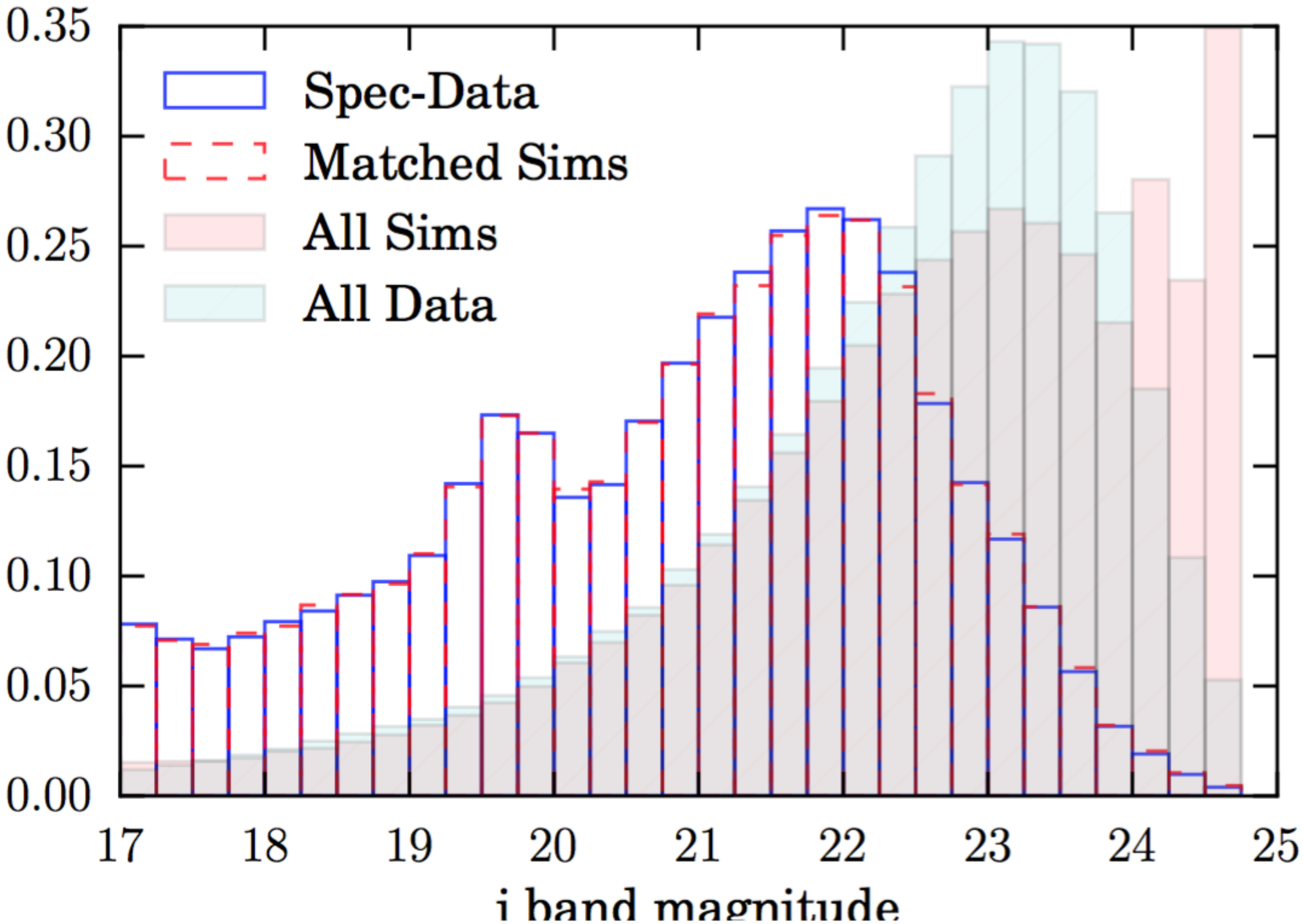


A bias Δ_z of >0.02 means that photo-z is the dominant source of systematic error in Y1 DES weak lensing analysis.

Testing the effects of these sample selection biases

Using N-body simulations, populated with galaxies we explore if any current methods can fix this covariate shift, and label bias problem.

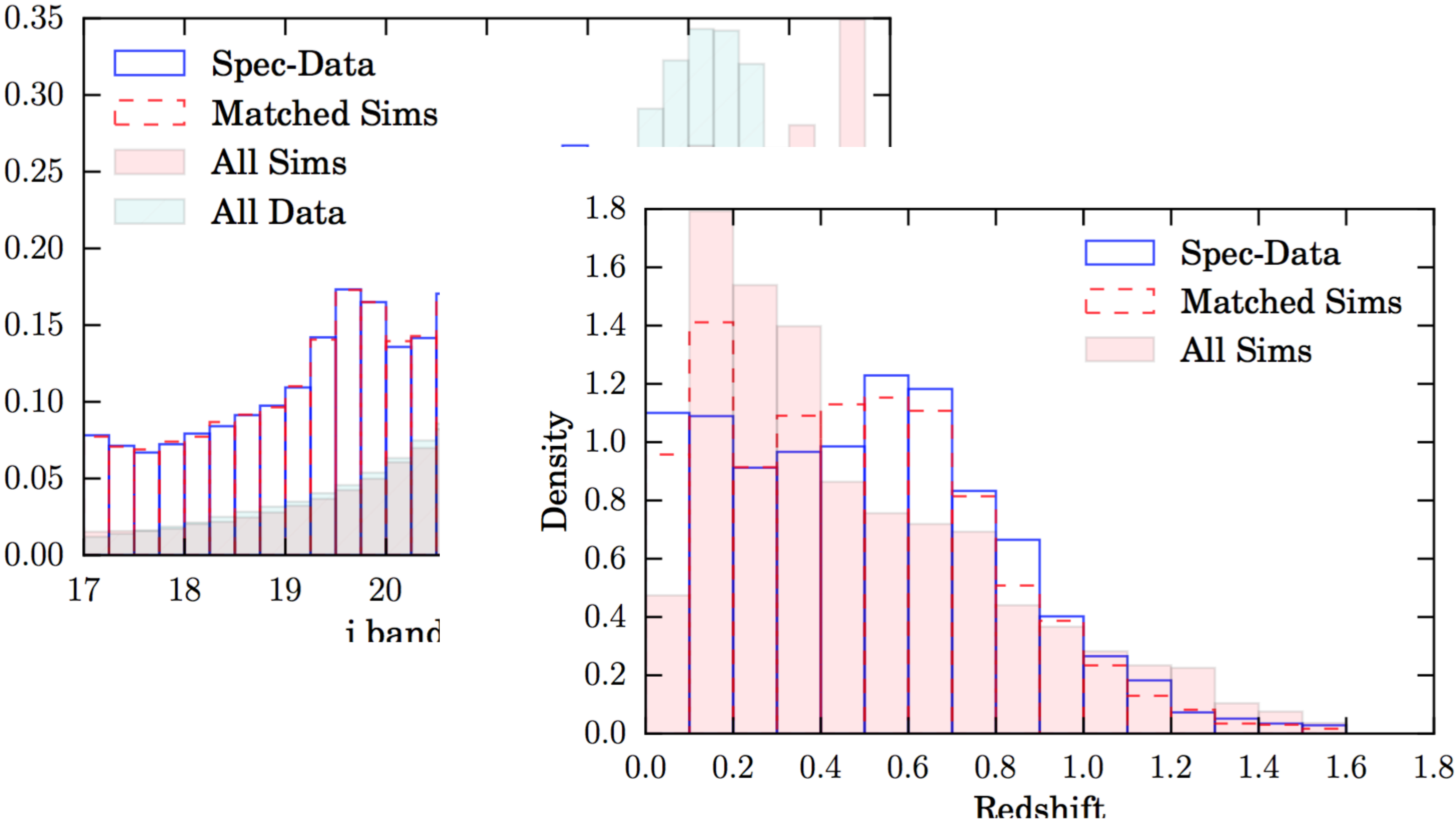
We generate "realistic" simulated spectroscopic training/validation data sets, with the view to measuring performance metrics on both the validation, and the science sample of interest.



Testing the effects of these sample selection biases

Using N-body simulations, populated with galaxies we explore if any current methods can fix this covariate shift, and label bias problem.

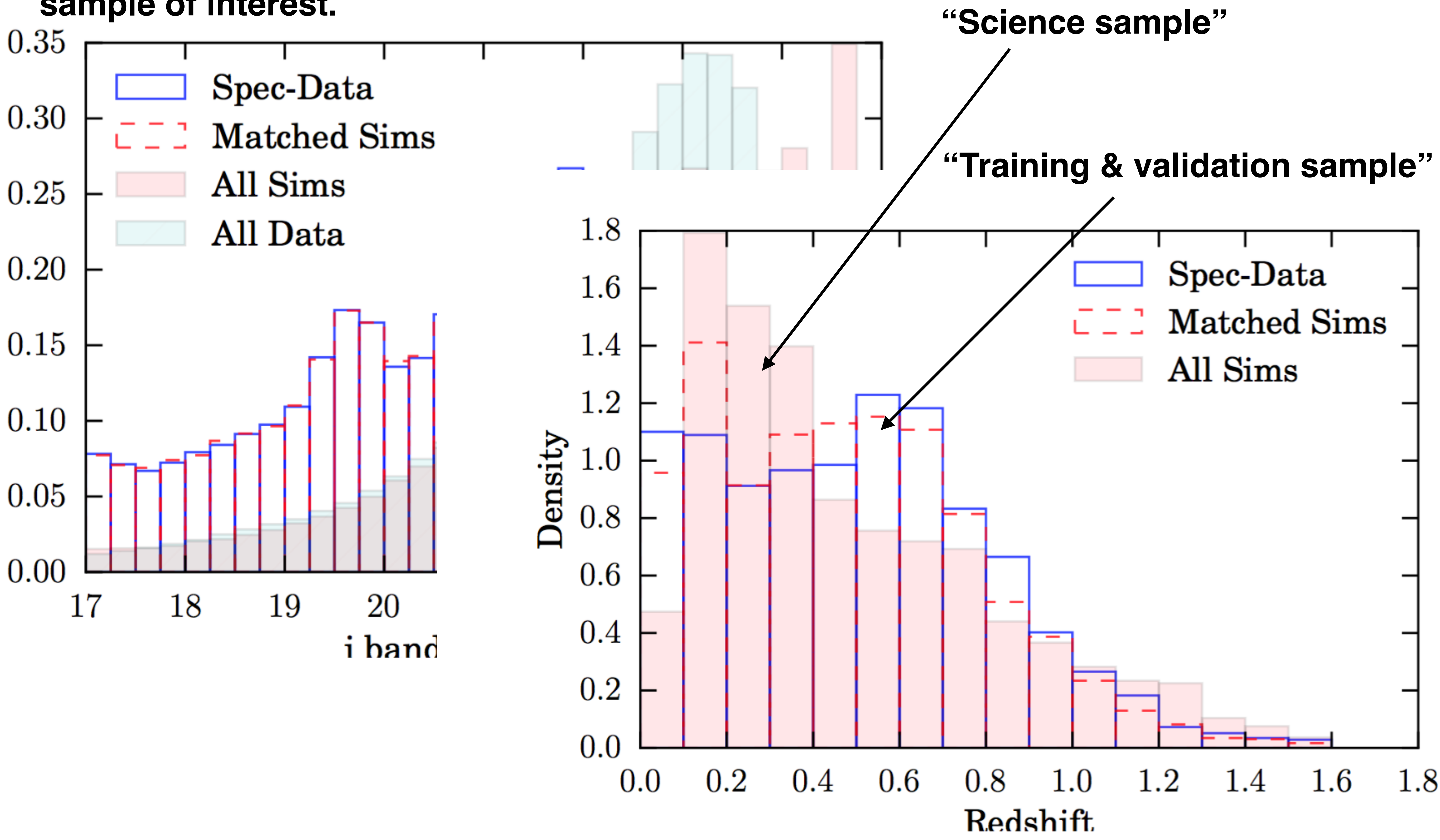
We generate “realistic” simulated spectroscopic training/validation data sets, with the view to measuring performance metrics on both the validation, and the science sample of interest.



Testing the effects of these sample selection biases

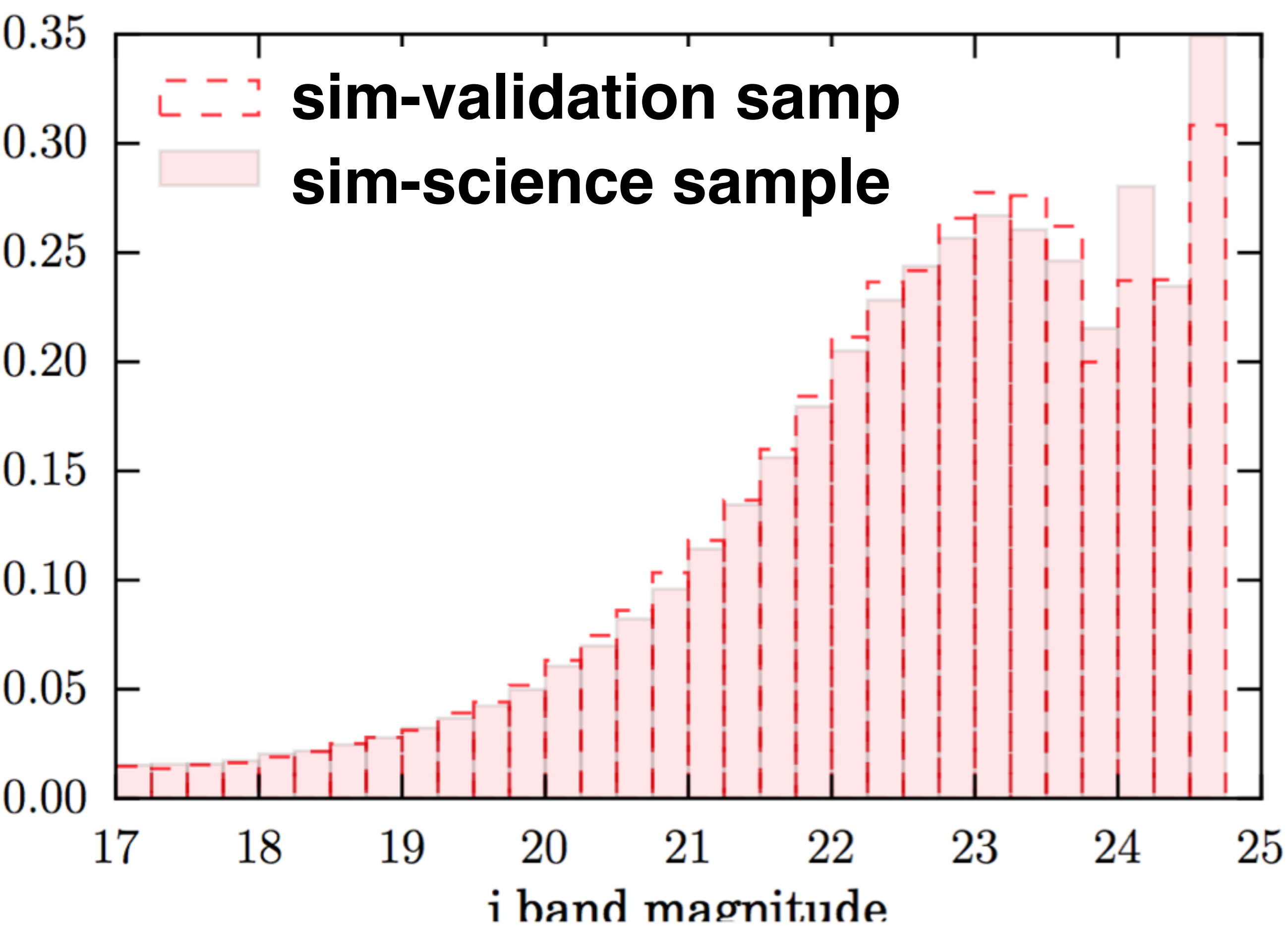
Using N-body simulations, populated with galaxies we explore if any current methods can fix this covariate shift, and label bias problem.

We generate "realistic" simulated spectroscopic training/validation data sets, with the view to measuring performance metrics on both the validation, and the science sample of interest.



Common approaches to sample selection bias

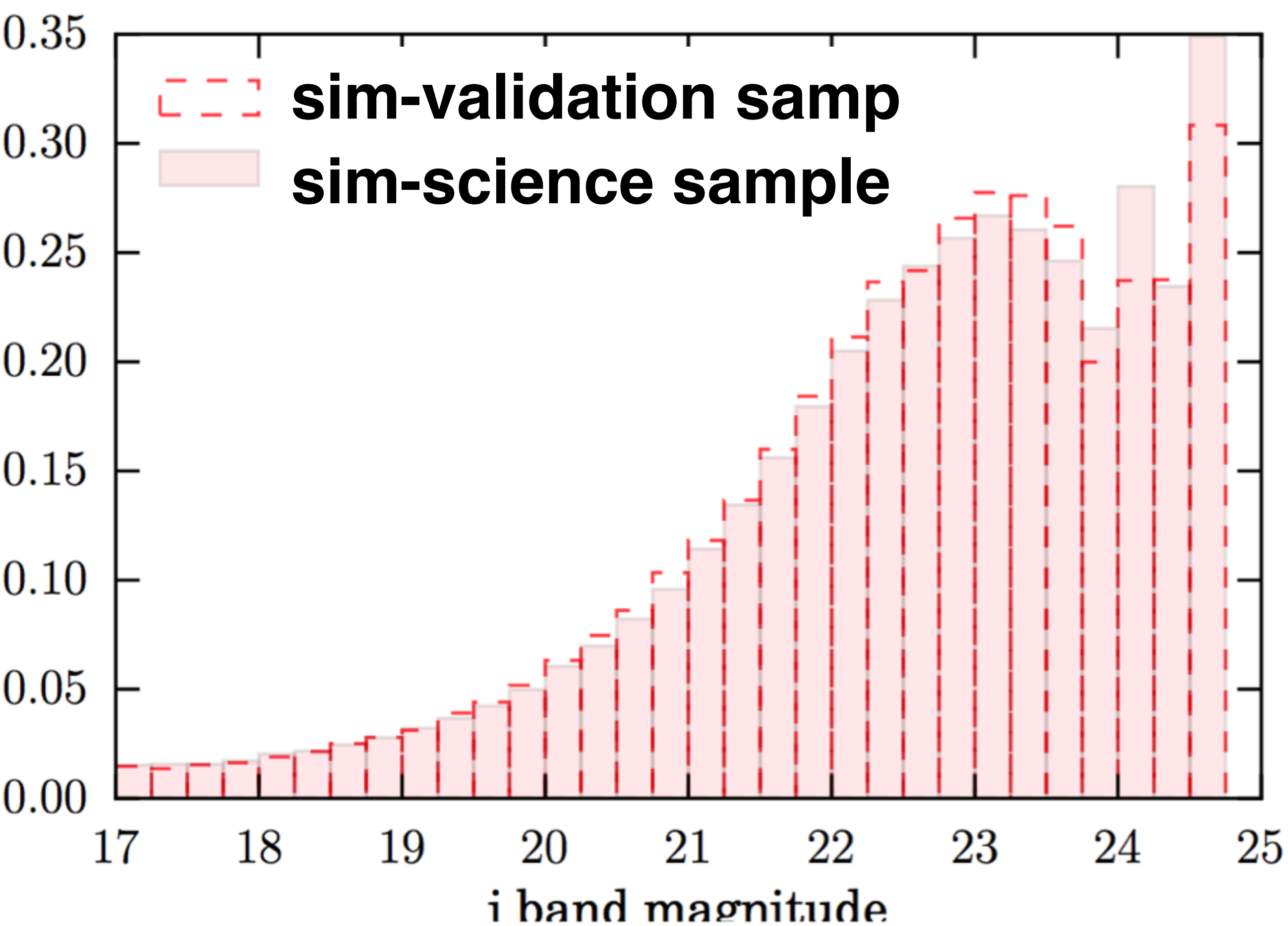
Lima et al: Reweight (using KNN) data so the input features (color-magnitude) distribution of the “simulated” validation data is that of “simulated” science sample.

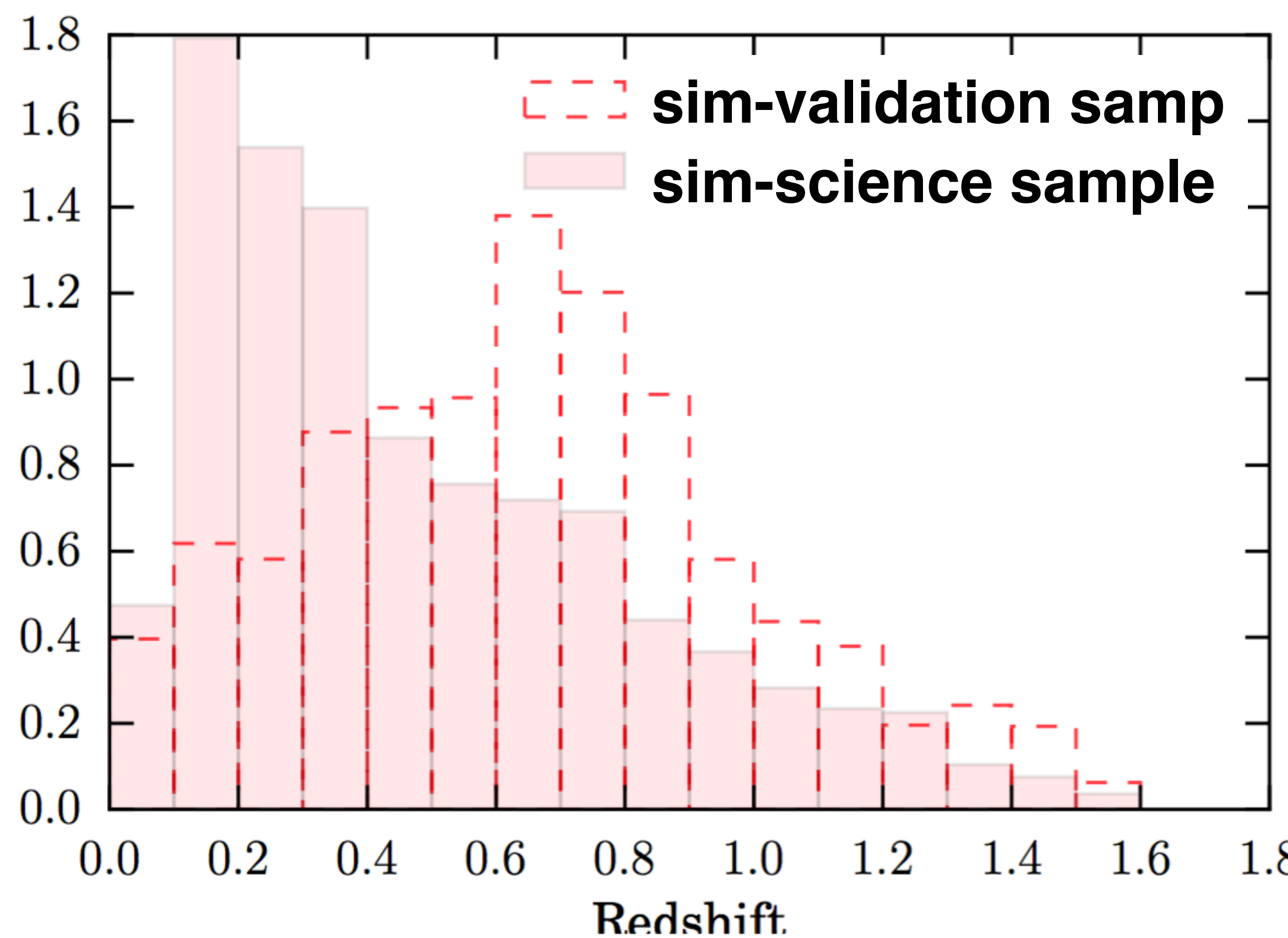


Hope this re-weighting captures any redshift difference between validation and science sample.

Common approaches to sample selection bias

Lima et al: Reweight (using KNN) data so the input features (color-magnitude) distribution of the "simulated" validation data is that of "simulated" science sample.

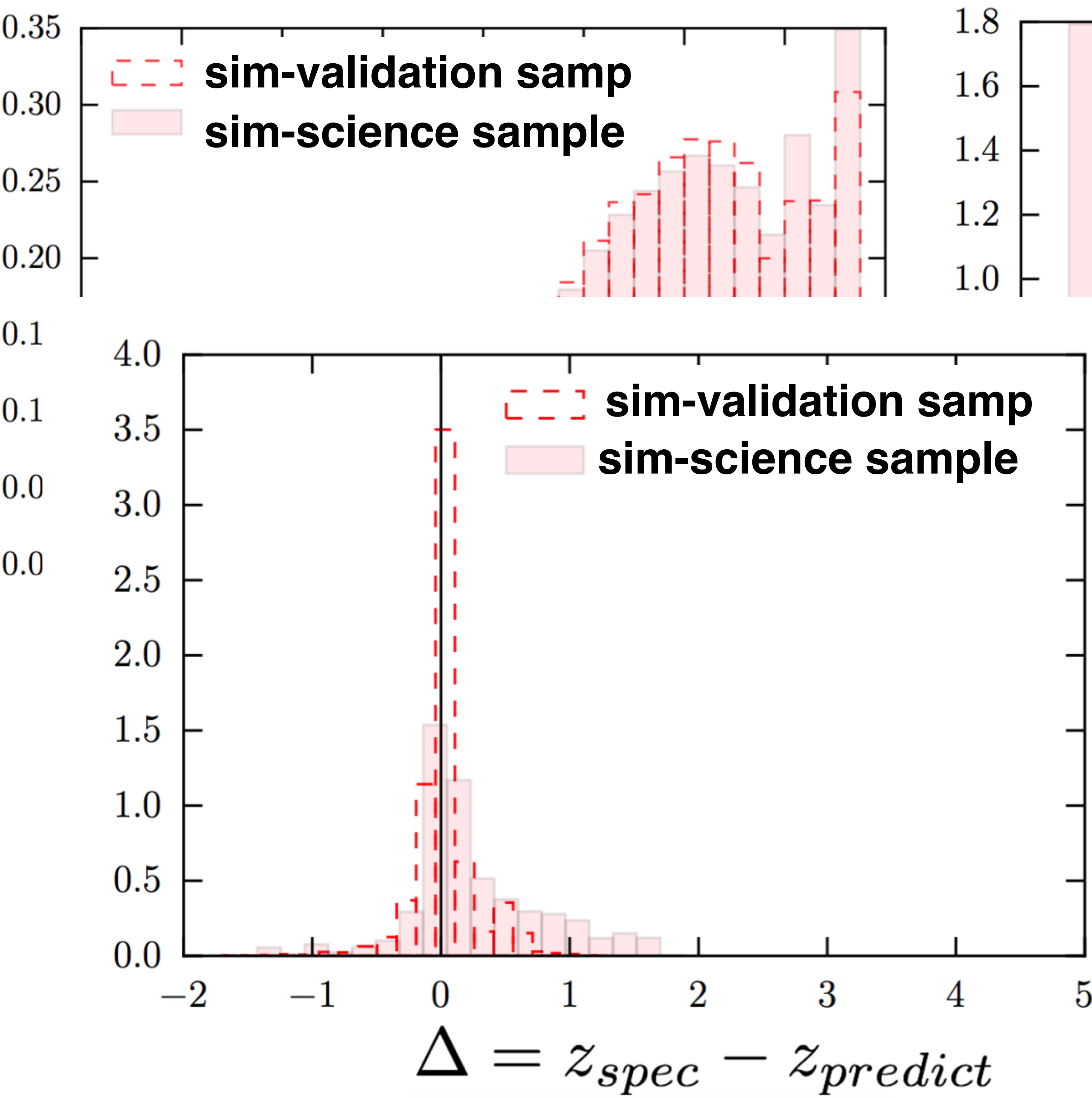


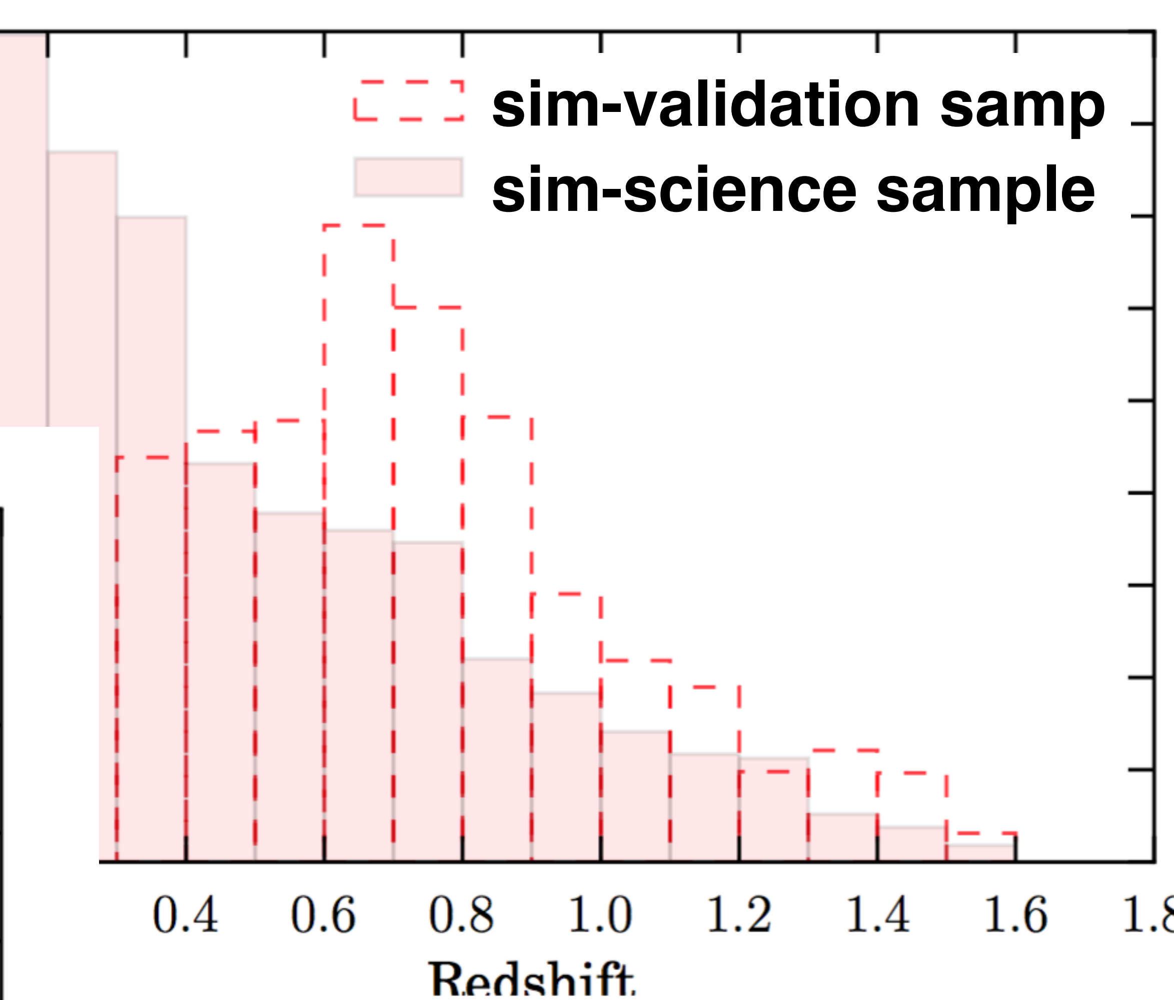


Hope this re-weighting captures any redshift difference between validation and science sample.

Common approaches to sample selection bias

Lima et al: Reweight (using KNN) data so the input features (color-magnitude) distribution of the “simulated” validation data is that of “simulated” science sample.

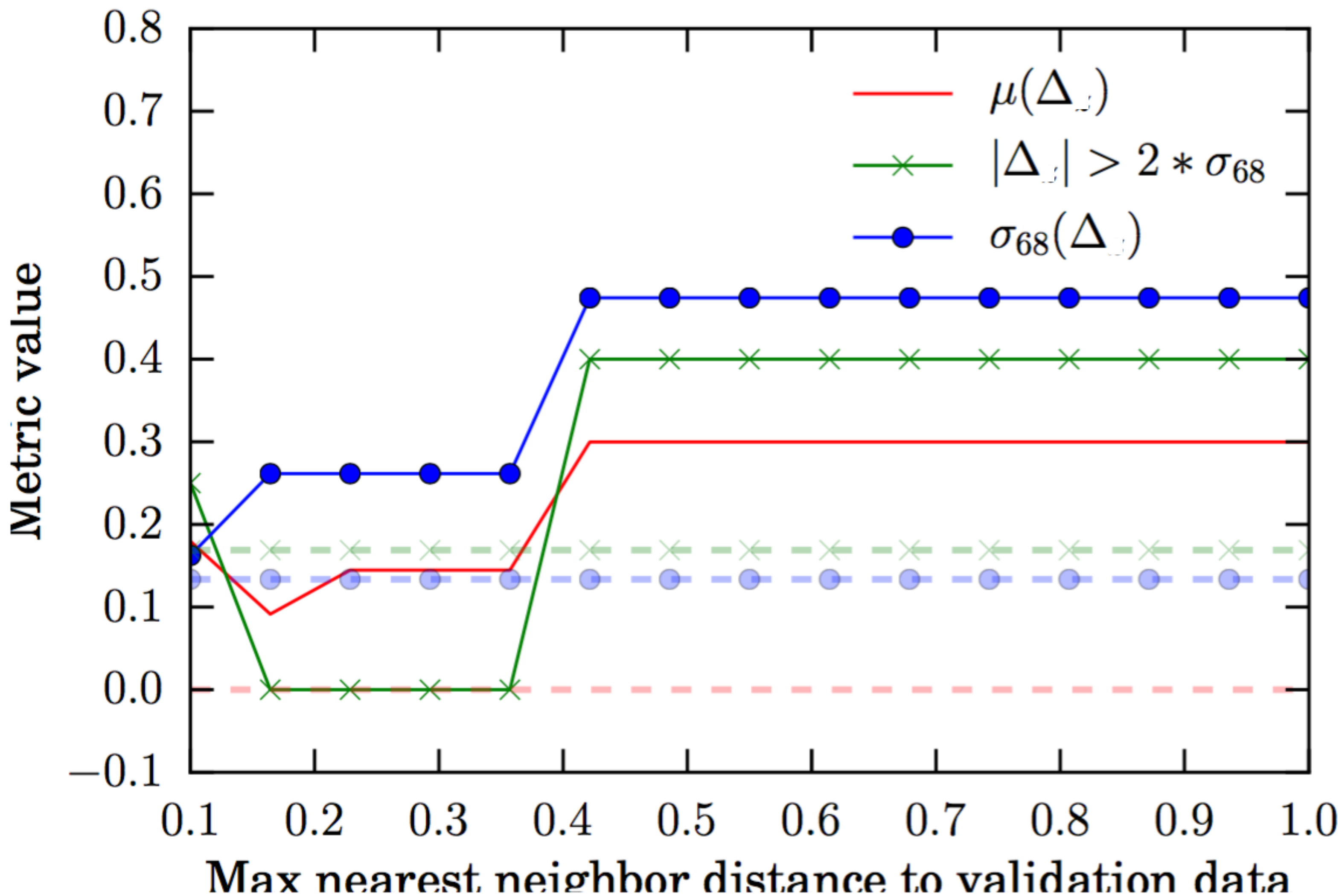




Hope this re-weighting captures any redshift difference between validation and science sample.

Common approaches to sample selection bias

Data culling: Remove science sample like data, that is not “close by” in KNN space to the “simulated” training/validation data.



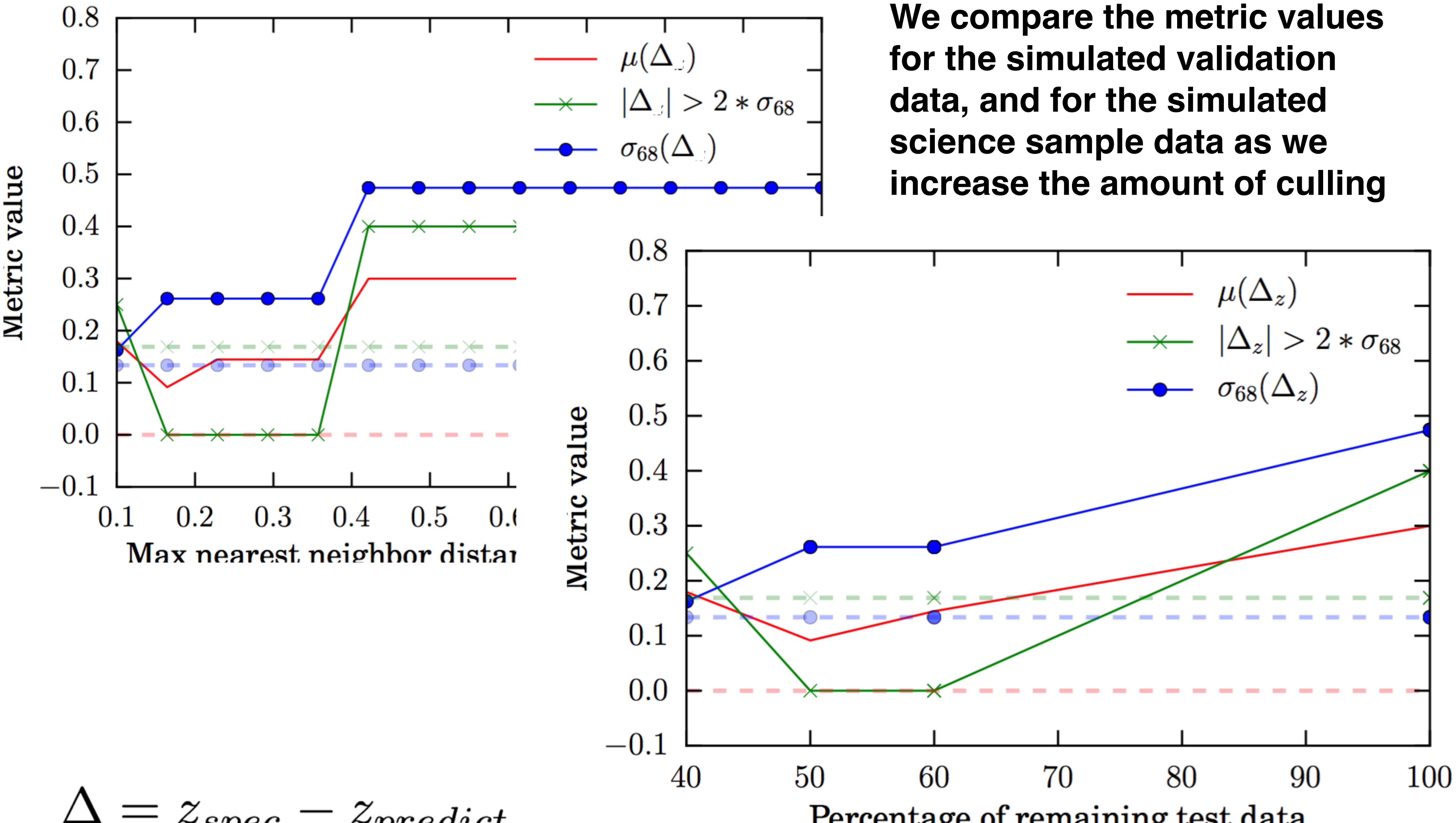
We compare the metric values for the simulated validation data, and for the simulated science sample data as we increase the amount of culling

$$\Delta = z_{spec} - z_{predict}$$

Common approaches to sample selection bias

Data culling: Remove science sample like data, that is not "close by" in KNN space to the "simulated" training/validation data.

We compare the metric values for the simulated validation data, and for the simulated science sample data as we increase the amount of culling



$$\Delta = z_{spec} - z_{predict}$$

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

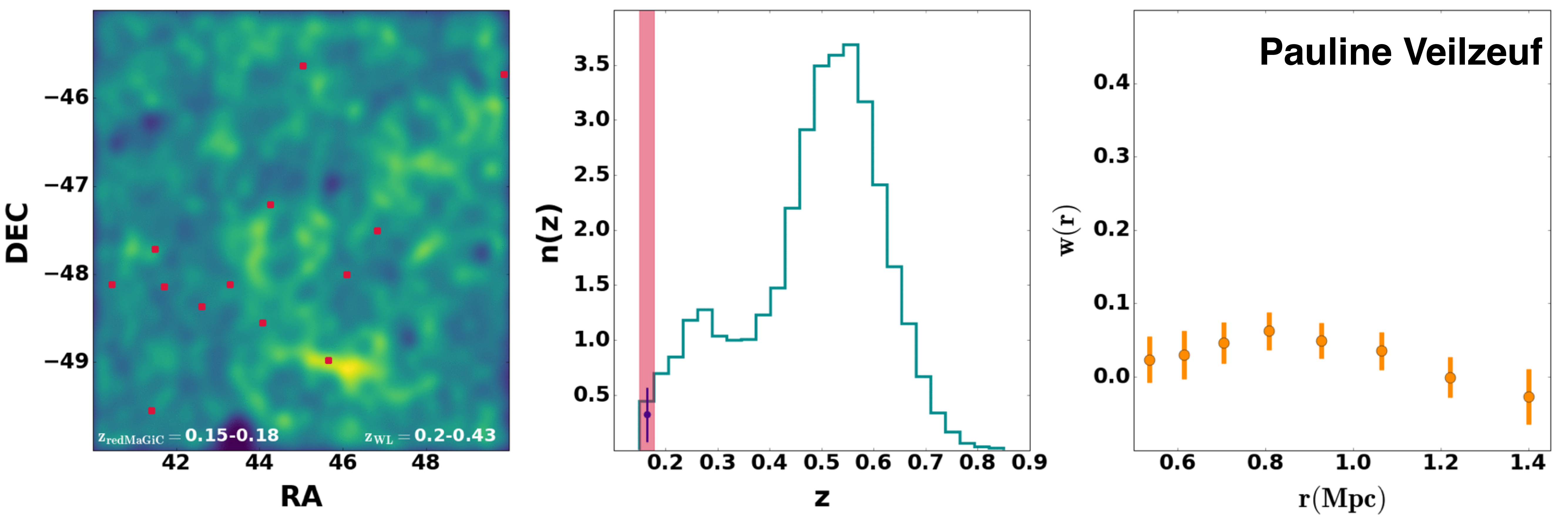
Summary/Conclusions

Overcoming this problem in the Dark Energy Survey Y1

Method 1:

Replace spec-z targets with COSMOS 30-band photometric redshifts, which for DES purposes are as accurate as spec-z, but don’t have redshift selection effects.

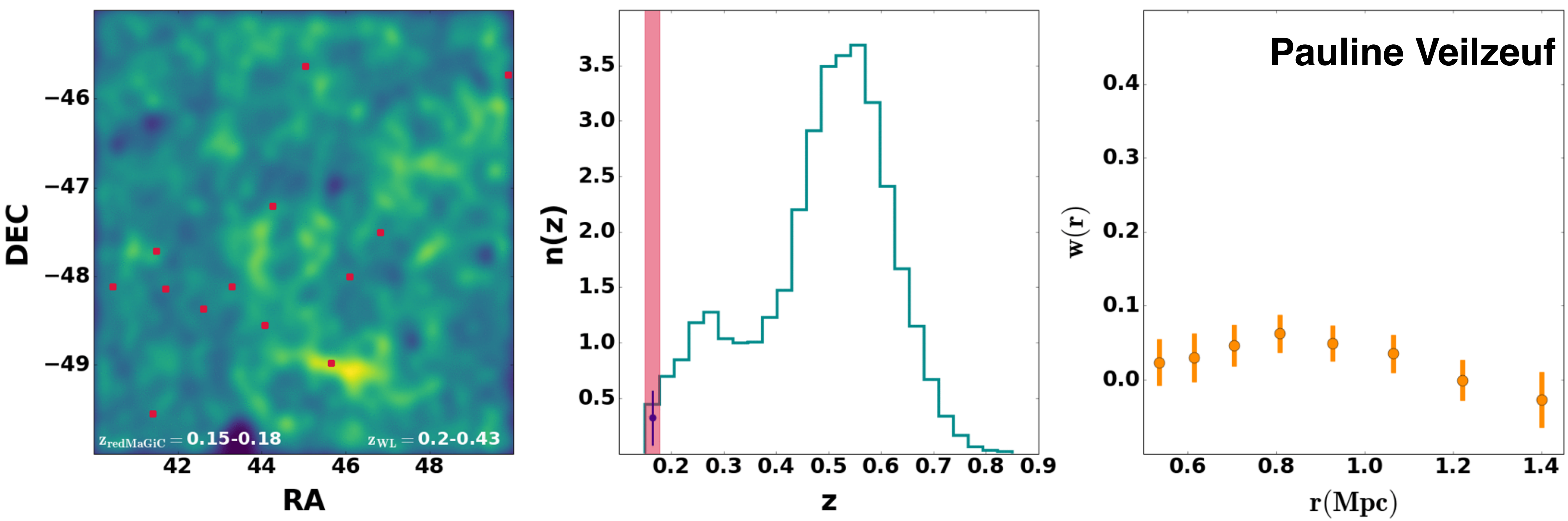
This induces new, but tractable problems.



Overcoming this problem in the Dark Energy Survey Y1

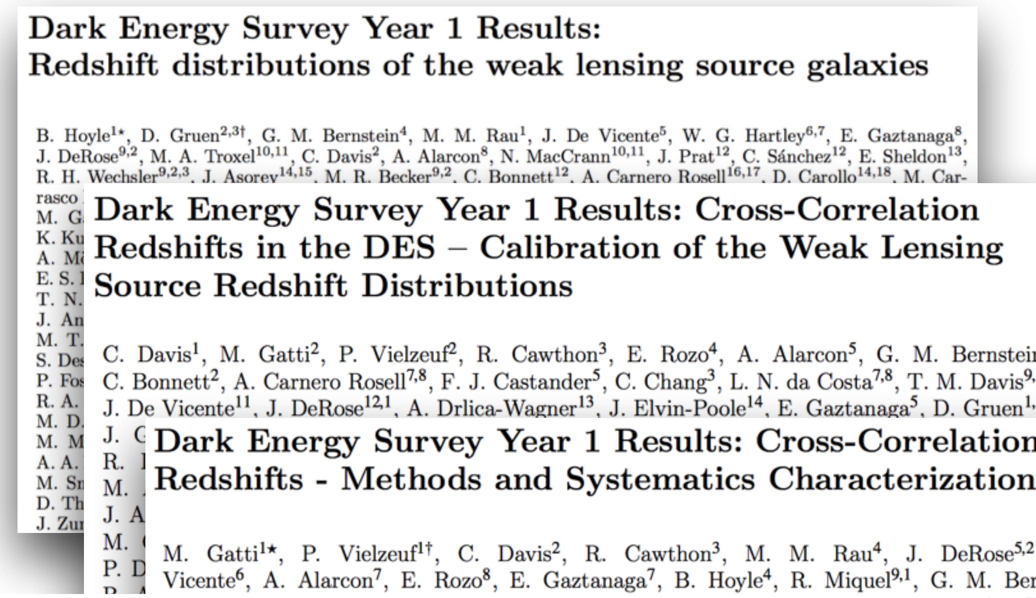
Method 1:
Replace spec-z targets with COSMOS 30-band photometric redshifts, which for DES purposes are as accurate as spec-z, but don't have redshift selection effects.
This induces new, but tractable problems.

Method 2:
The clustering redshift approach:
only need complete samples across the sky, not "representative".

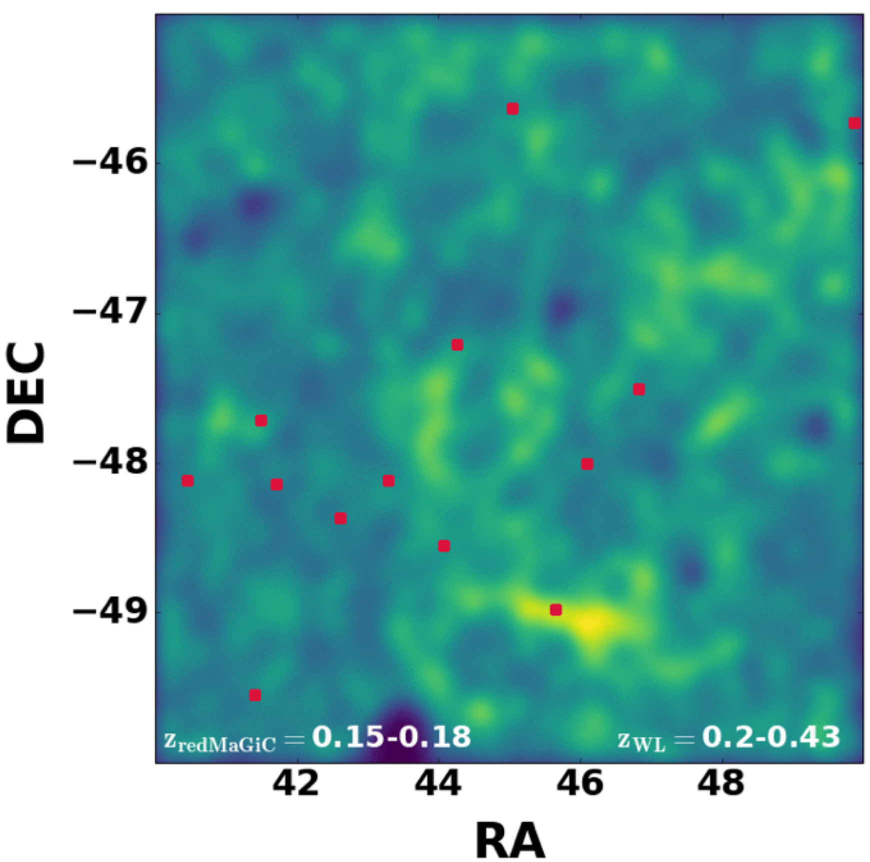


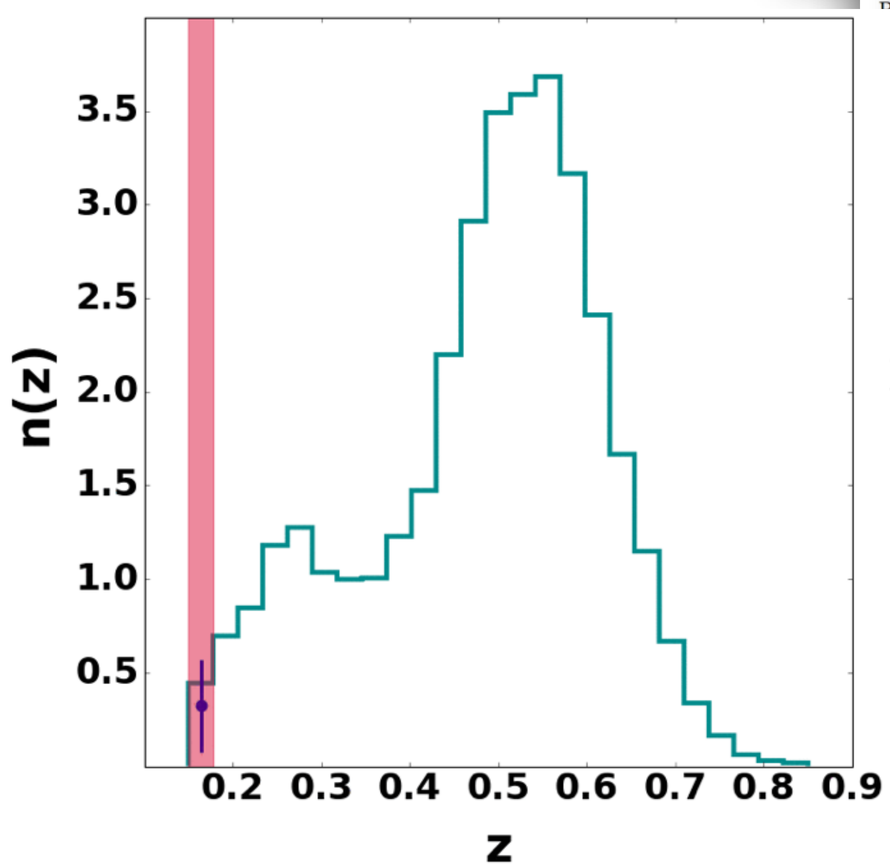
Overcoming this problem in the Dark Energy Survey Y1

Method 1:
Replace spec-z targets with COSMOS 30-band photometric redshifts, which for DES purposes are as accurate as spec-z, but don't have redshift selection effects.
This induces new, but tractable problems.

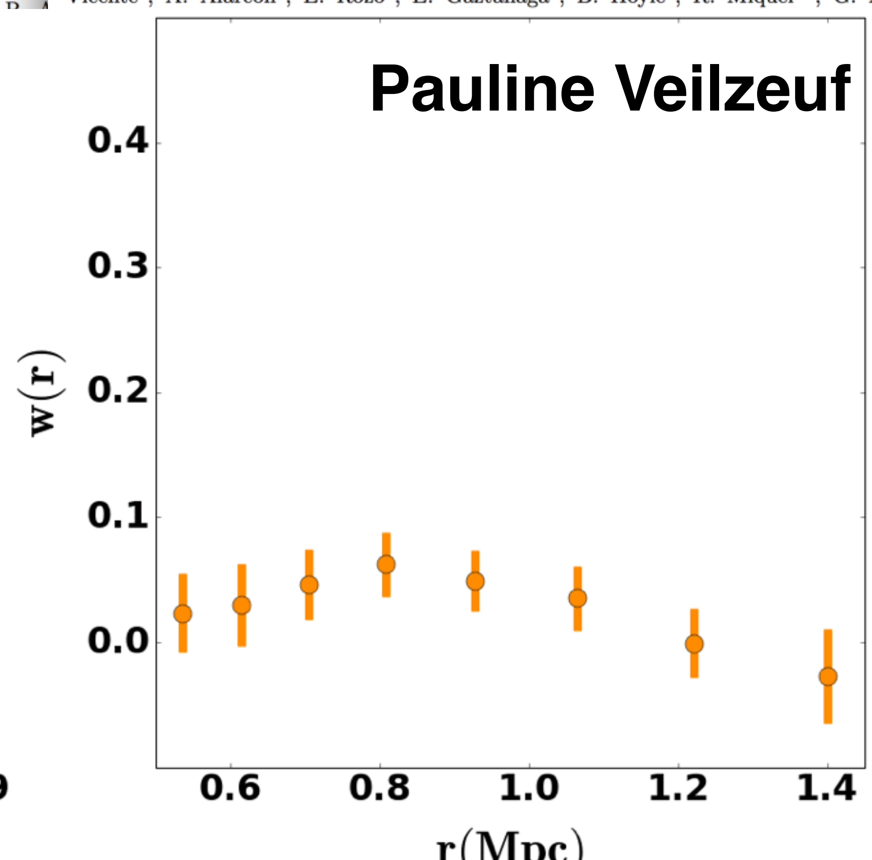


Method 2:
The clustering redshift approach:
only need complete samples across the sky, not "representative".





Pauline Veilzeuf



Validating photo-z distribution in Y1 Dark Energy Survey

Value	Bin 1	Bin 2	Bin 3	Bin 4
z^{PZ} range	0.20–0.43	0.43–0.63	0.63–0.90	0.90-1.30
COSMOS final Δz^i, tomographic uncertainty	$+0.001 \pm 0.020$	-0.014 ± 0.021	$+0.008 \pm 0.018$	-0.057 ± 0.022
WZ final Δz^i	$+0.008 \pm 0.026$	-0.031 ± 0.017	-0.010 ± 0.014	—
Combined final Δz^i	$+0.004 \pm 0.0$			022

Δ_z and it's uncertainty

Δ_z = <z_true> - <z-photz>

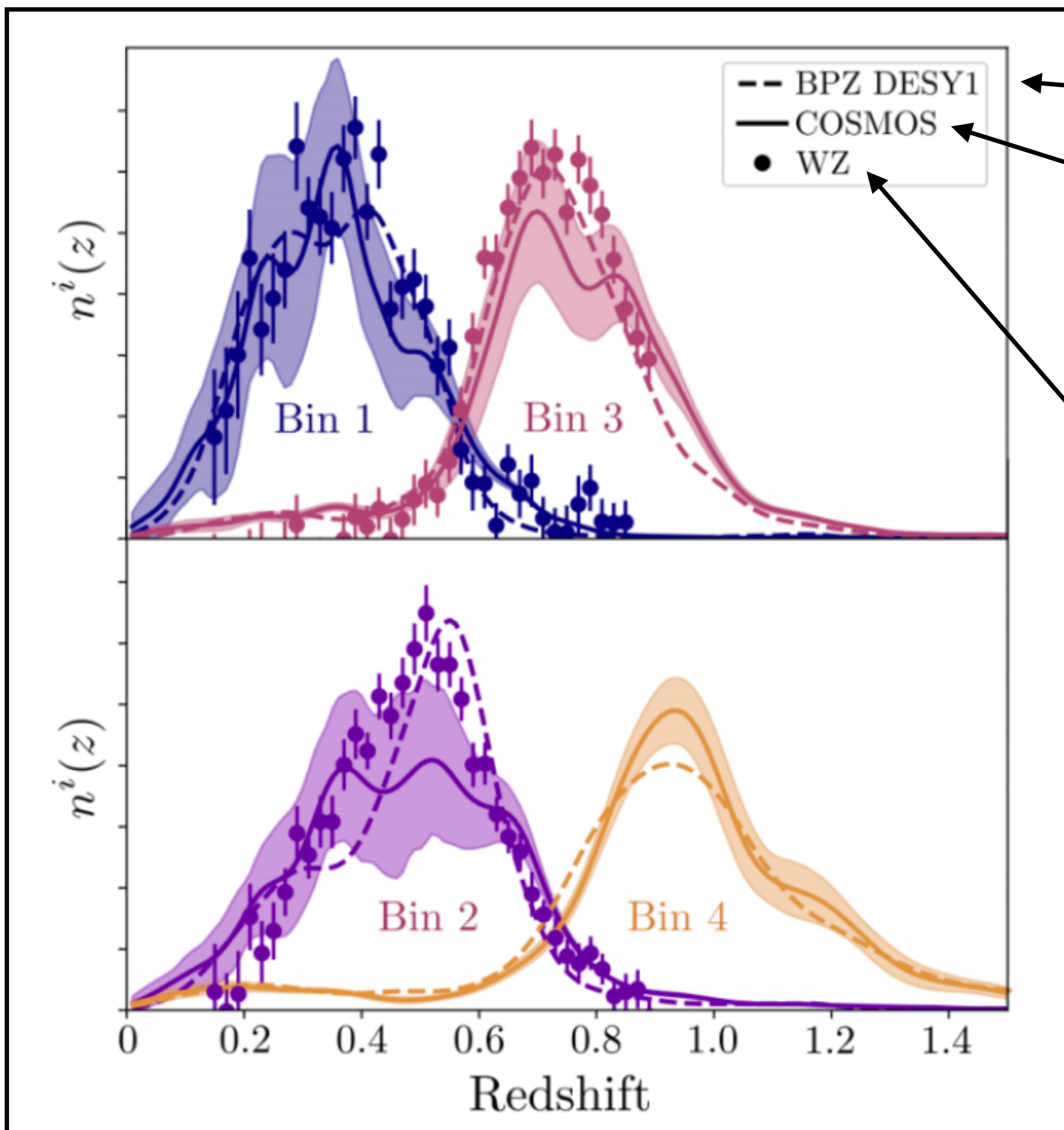


Photo-z predictions

Method 1:
Color-redshift mapping using
30 band photo-z [cosmic variance]

Method 2:
Estimation of dndz of a sample
using the clustering technique
(i.e, cross correlate with a sample
of objects with known redshifts)

Hoyle, Grün & DES et al 2017

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

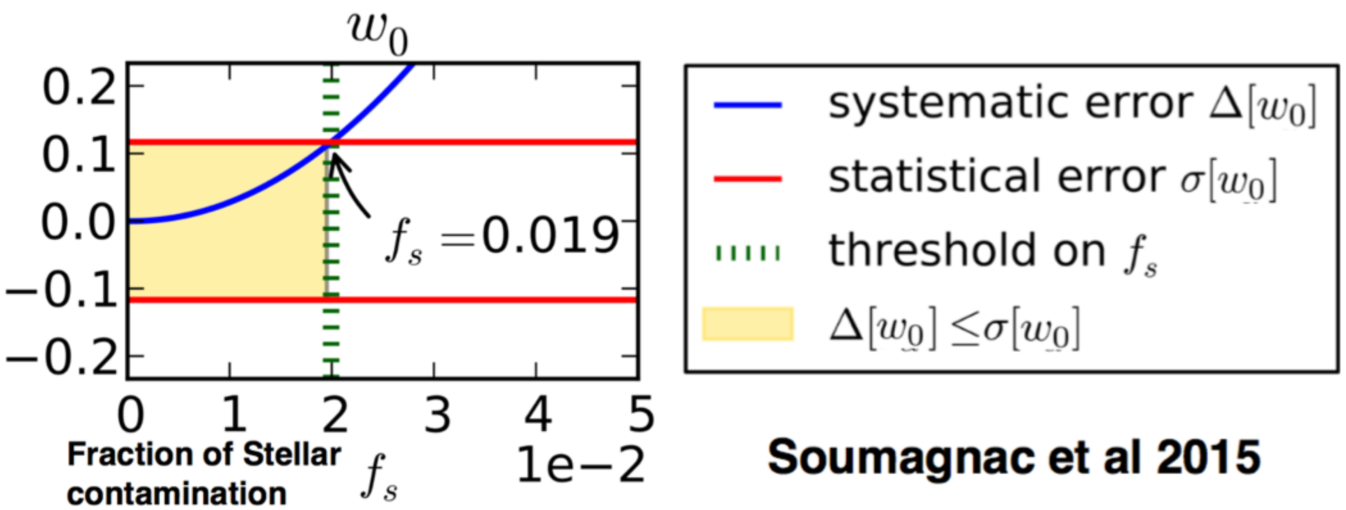
Other common applications of ML

Recent, novel applications of ML

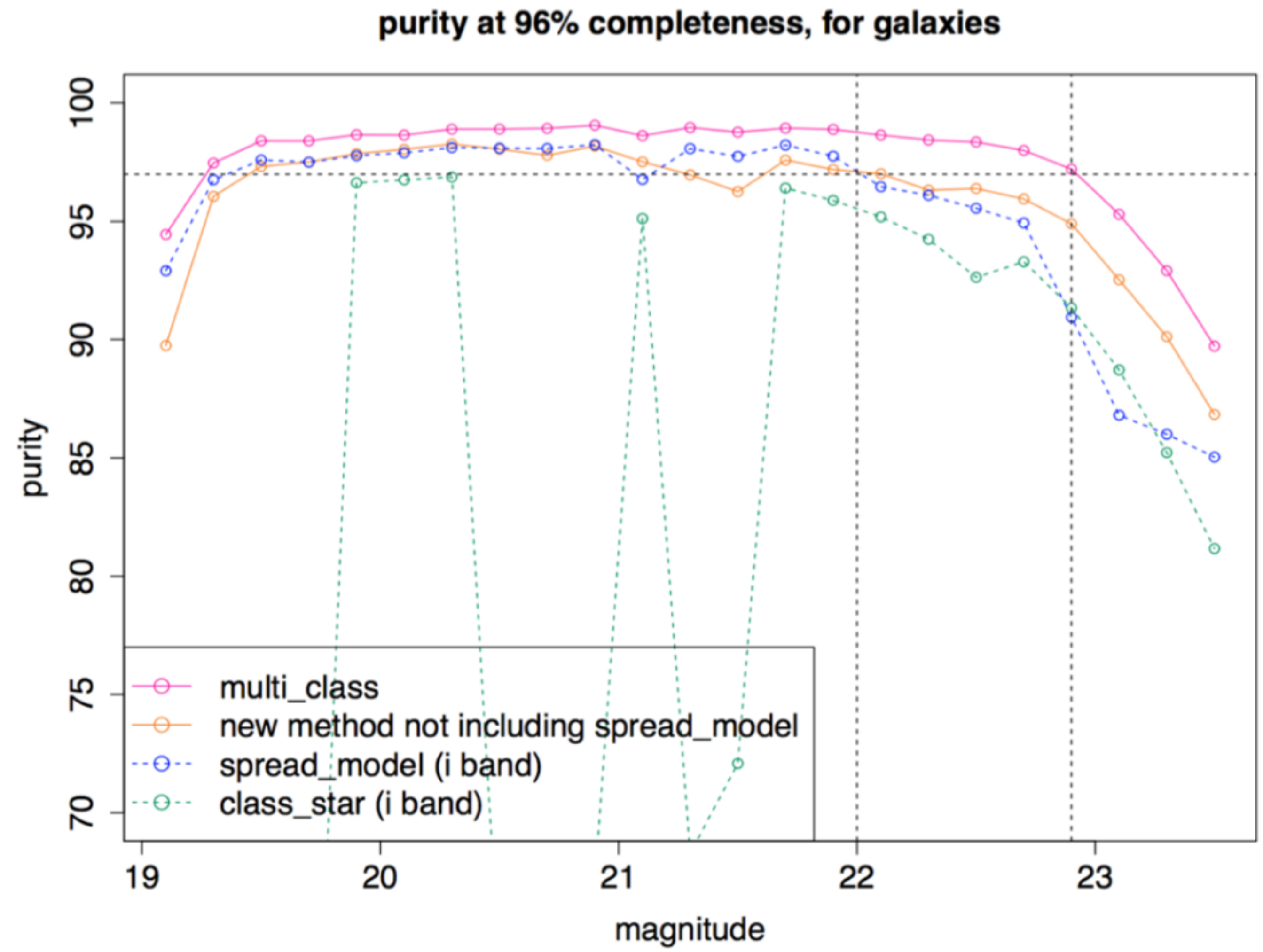
Summary/Conclusions

Star Galaxy separation

Given an image of the night sky, is an object a star in our galaxy, or a far away galaxy?
Improvement in star-galaxy classification leads to reduced errors in cosmological analysis e.g. DES SV analysis:

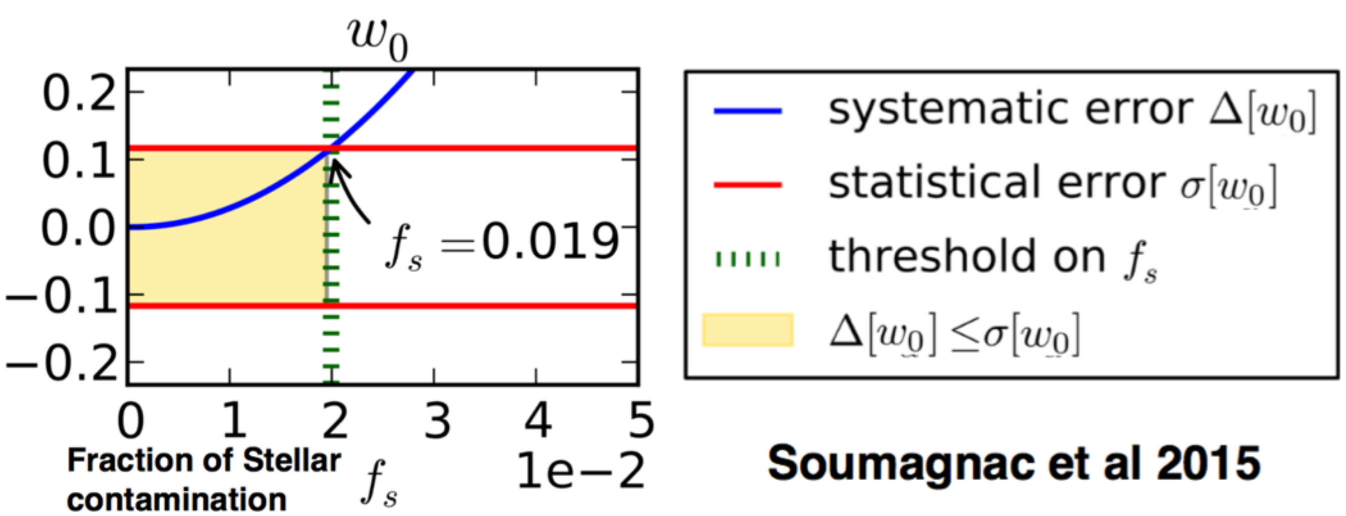


Soumagnac et al 2015

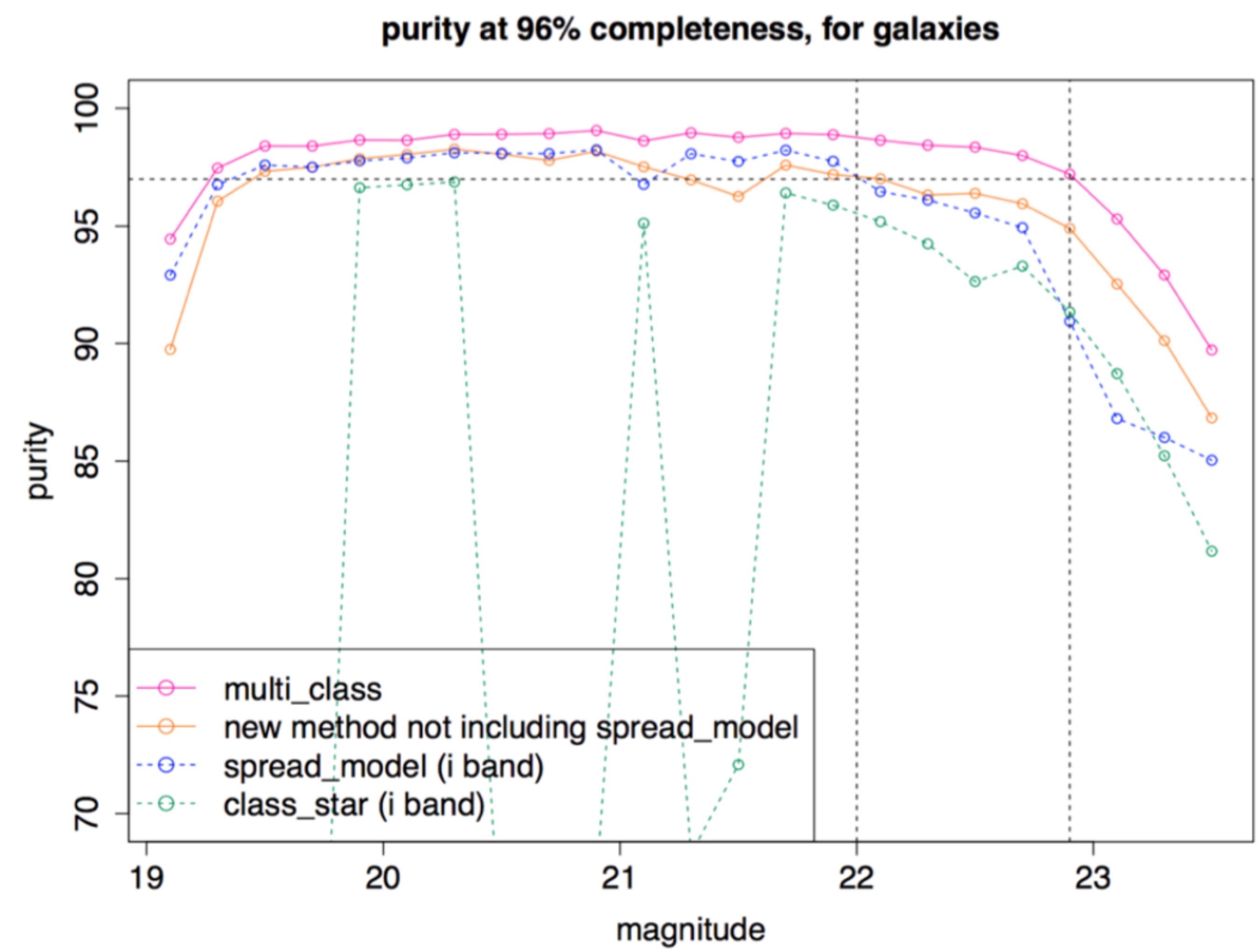


Star Galaxy separation

Given an image of the night sky, is an object a star in our galaxy, or a far away galaxy?
Improvement in star-galaxy classification leads to reduced errors in cosmological analysis e.g. DES SV analysis:



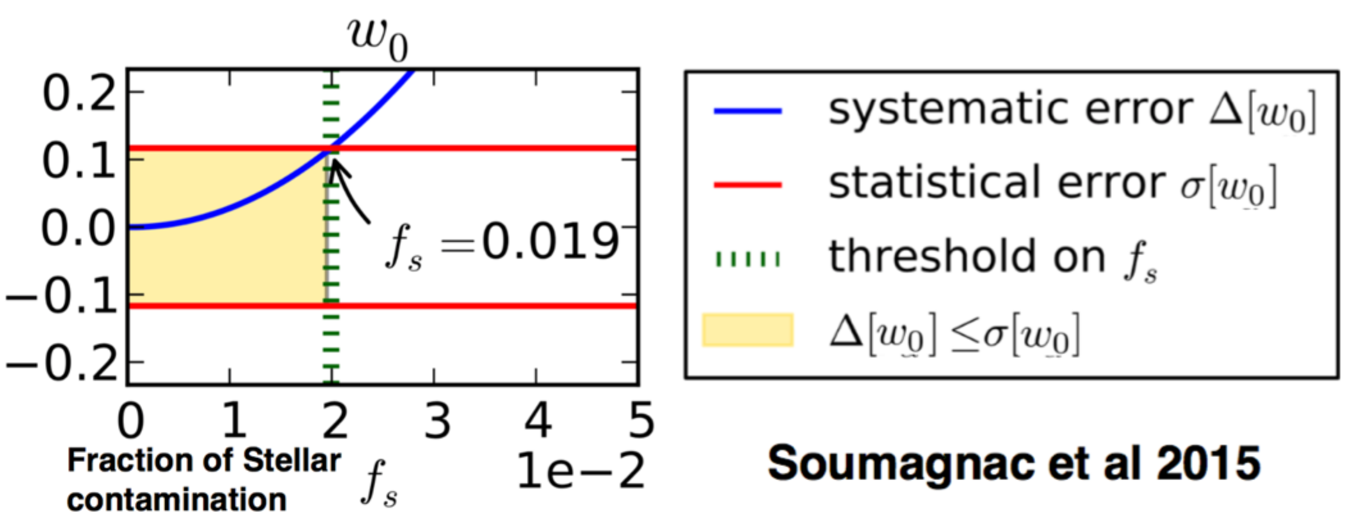
Soumagnac et al 2015



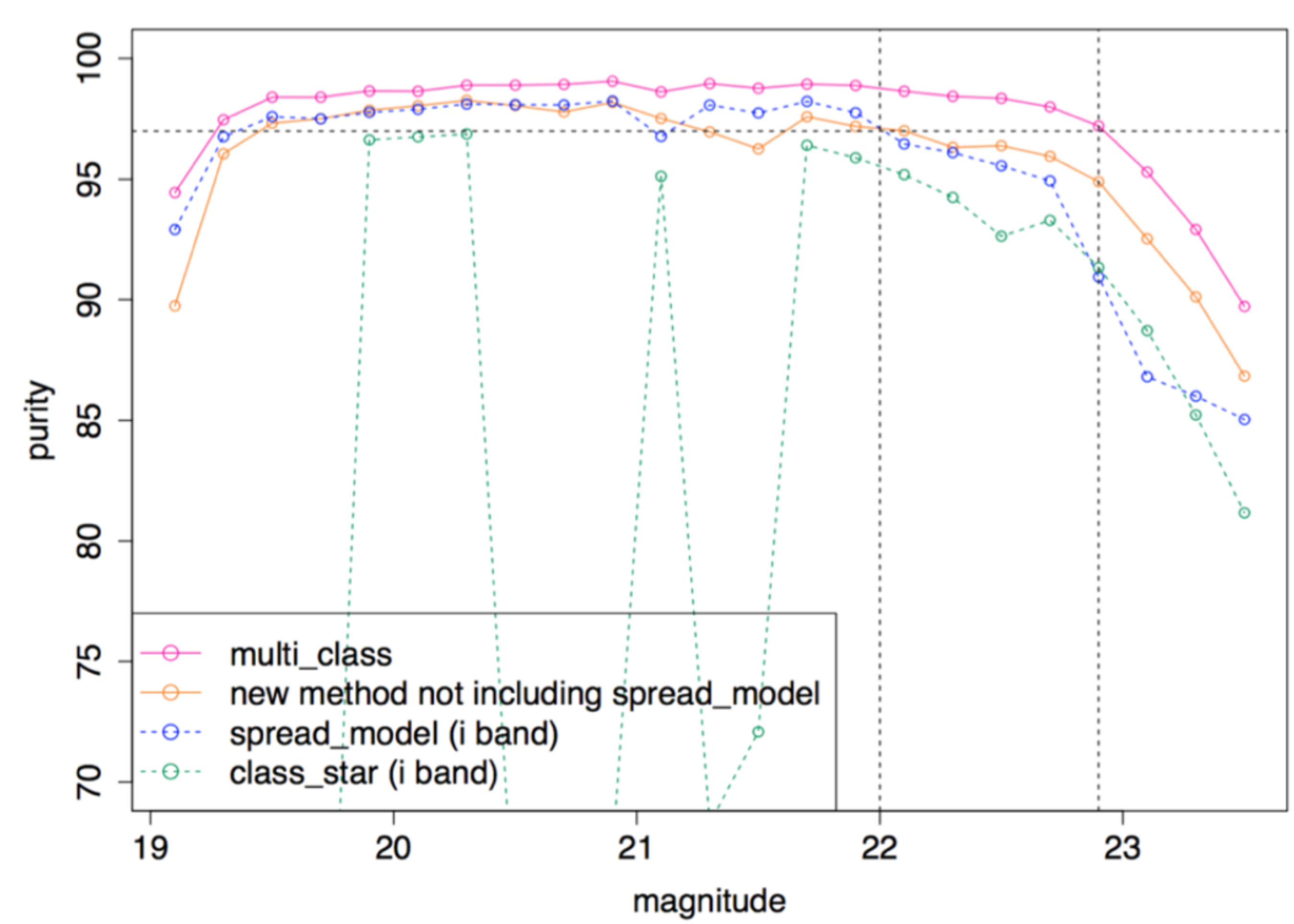
In Y1 we face a similar problem as before labelled data is biased!

Star Galaxy separation

Given an image of the night sky, is an object a star in our galaxy, or a far away galaxy? Improvement in star-galaxy classification leads to reduced errors in cosmological analysis e.g. DES SV analysis:



Soumagnac et al 2015



In Y1 we face a similar problem as before labelled data is biased!

Moving towards higher order measurements of the predicted signal. e.g. does the number density of stars increase as one approaches the LMC / our Galaxy disk (Nacho Sevilla, BH, DES et al in prep)

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

Summary/Conclusions

Convolutional Neural Networks

Galaxy Zoo: A massive program to train members of the public to visually inspect 1 Million galaxies more than 50 times each

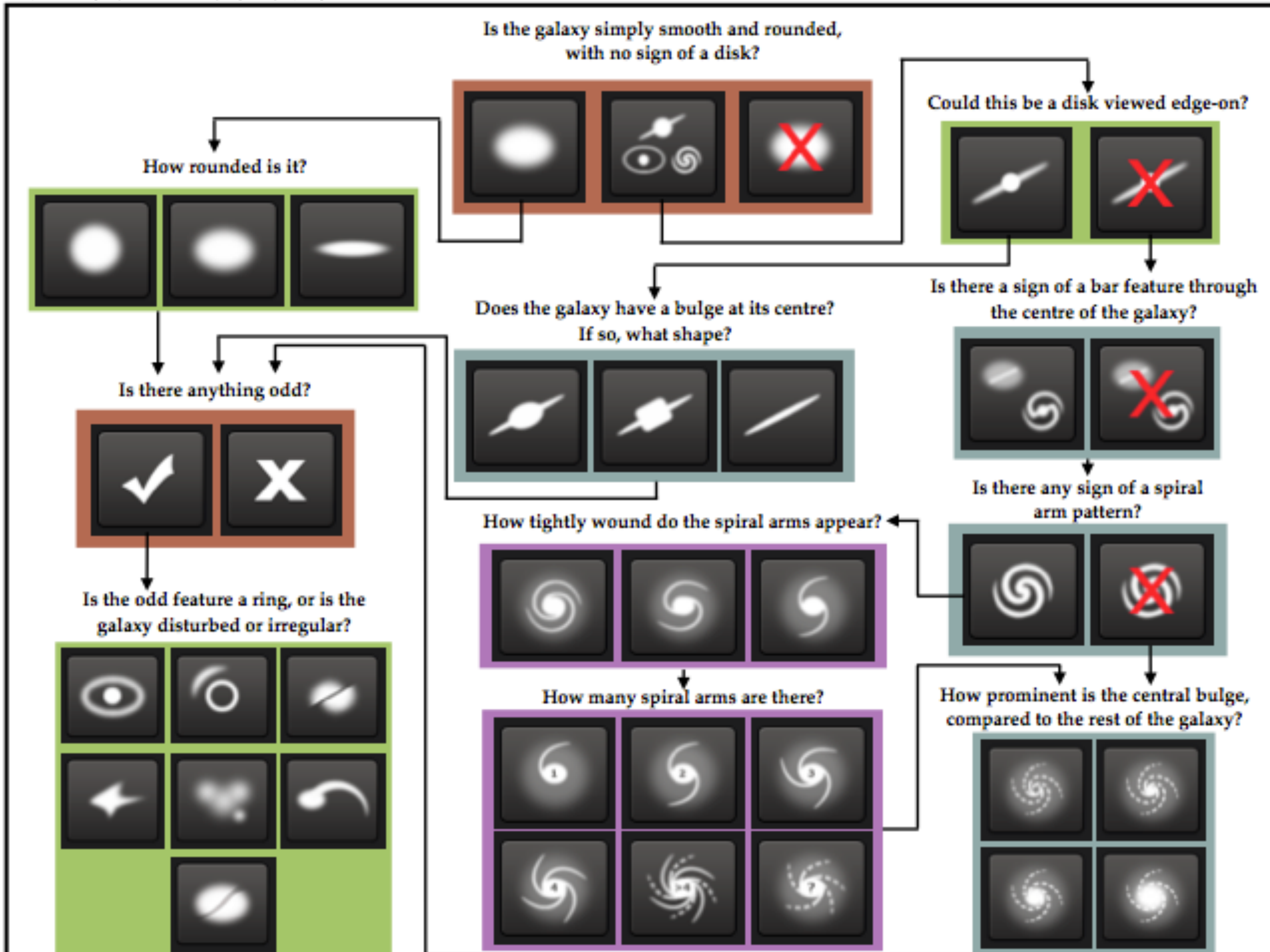


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Willet et al 2013

Convolutional Neural Networks

Galaxy Zoo: A massive program to train members of the public to visually inspect 1 Million galaxies more than 50 times each

Kaggle-contest:
use ML to reproduce the classifications of humans.

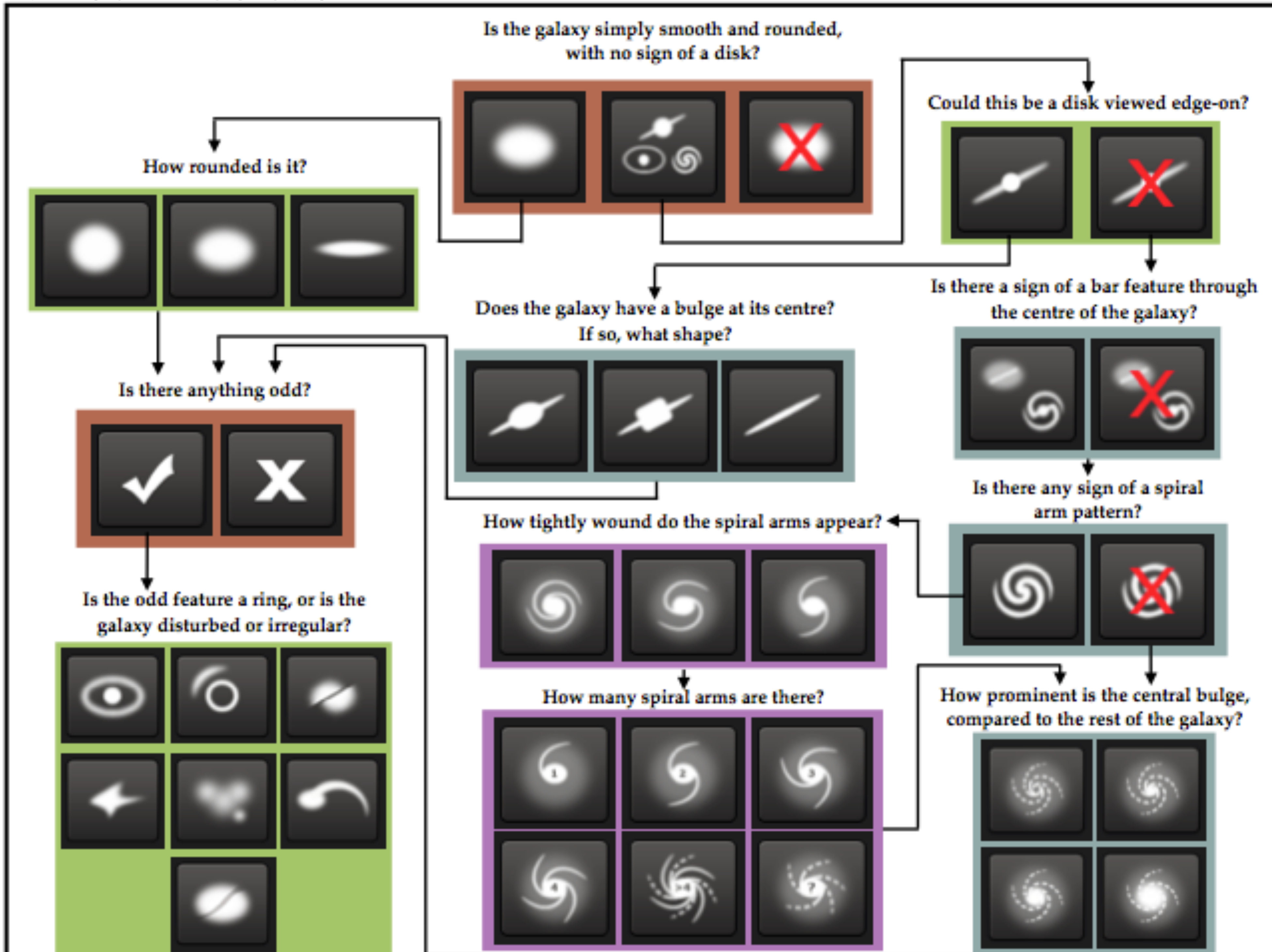


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge **Willet et al 2013**

Convolutional Neural Networks

Galaxy Zoo: A massive program to train members of the public to visually inspect 1 Million galaxies more than 50 times each

Kaggle-contest: use ML to reproduce the classifications of humans.

Could apply results to the 100's million of galaxies and repeat for new surveys

First application of Deep ML with 2d-CovNets in Astrophysics (Dieleman et al 2015)

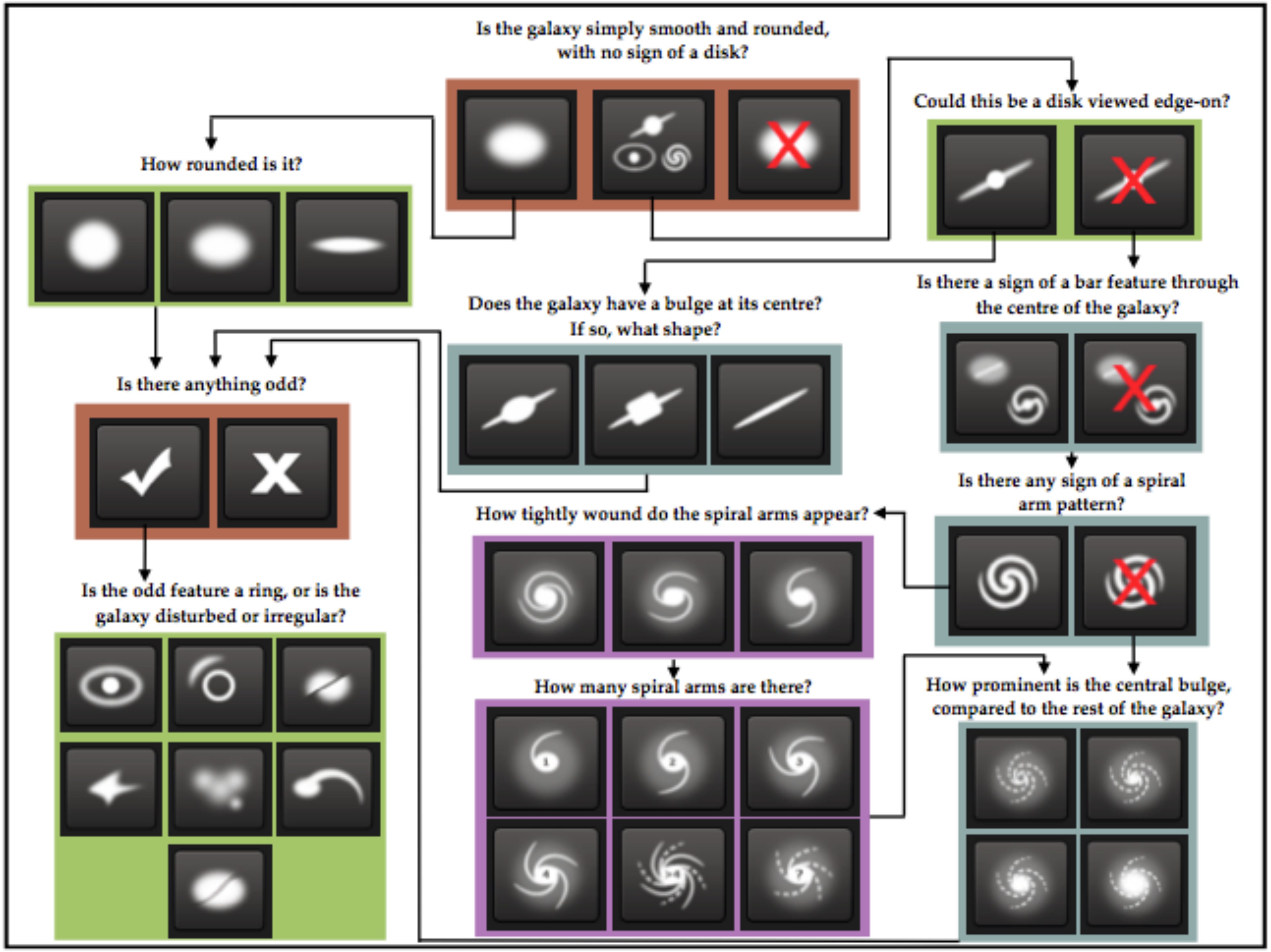


Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

Willet et al 2013

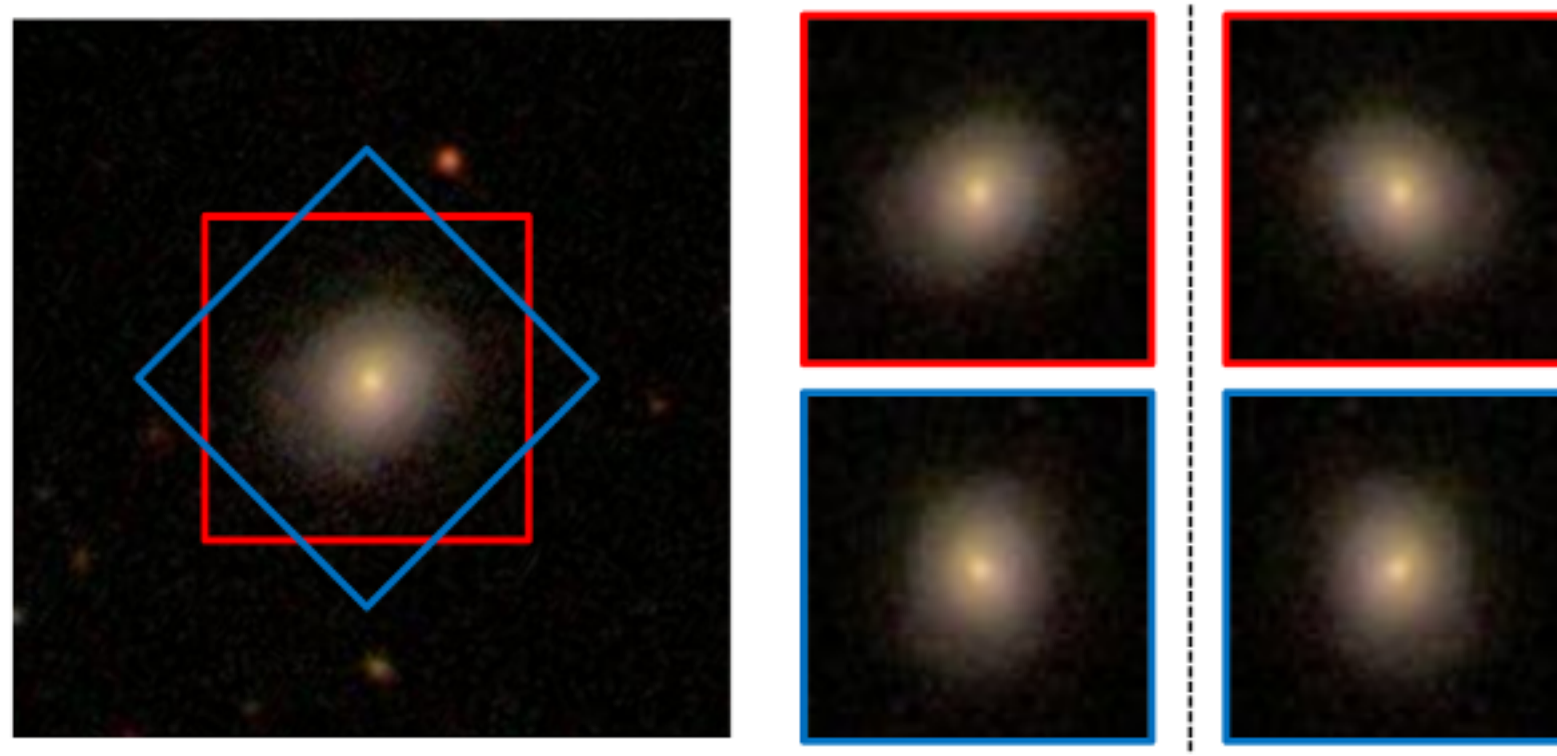
CNNs for Galaxy Zoo

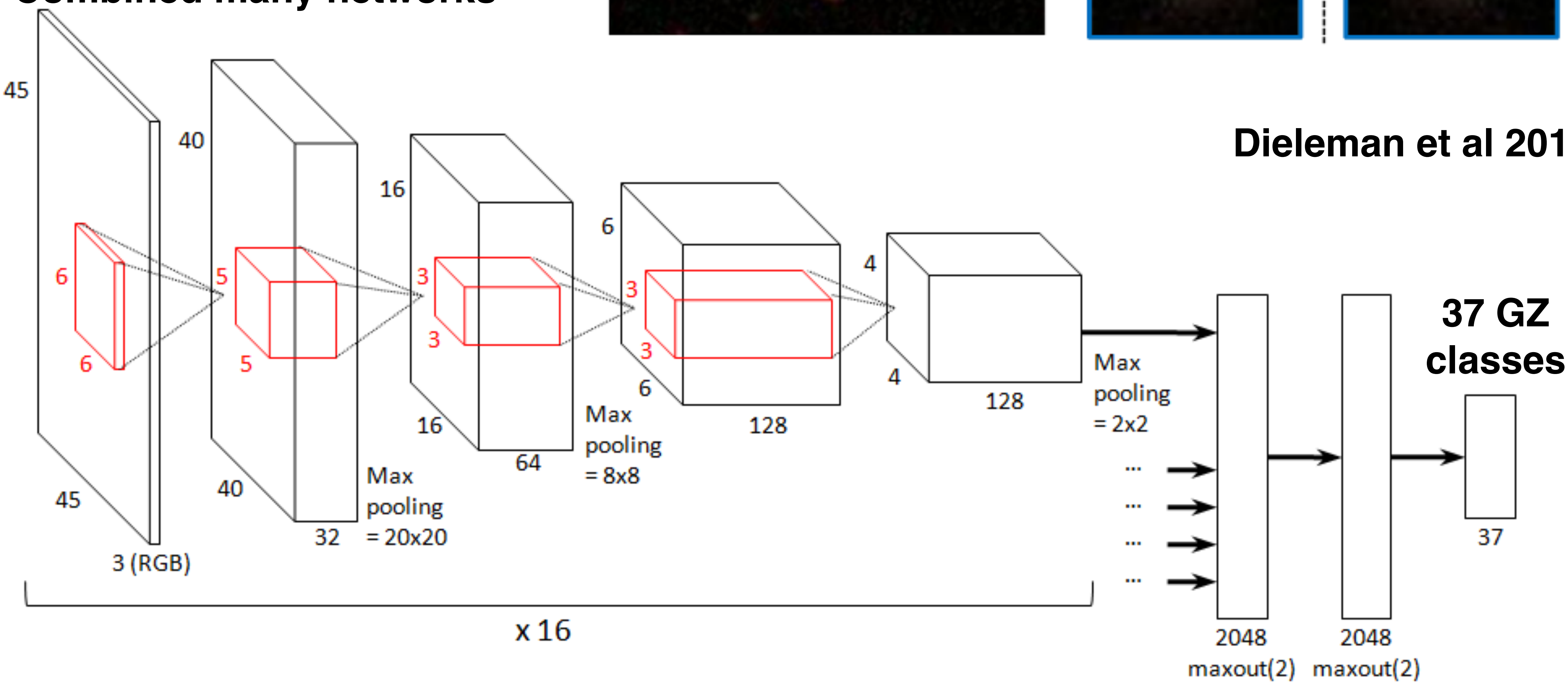
Extract centre of image
=> the galaxy,
rescaled to 45x45 pixels

Data augmentation

Dropout/Max pooling

Combined many networks





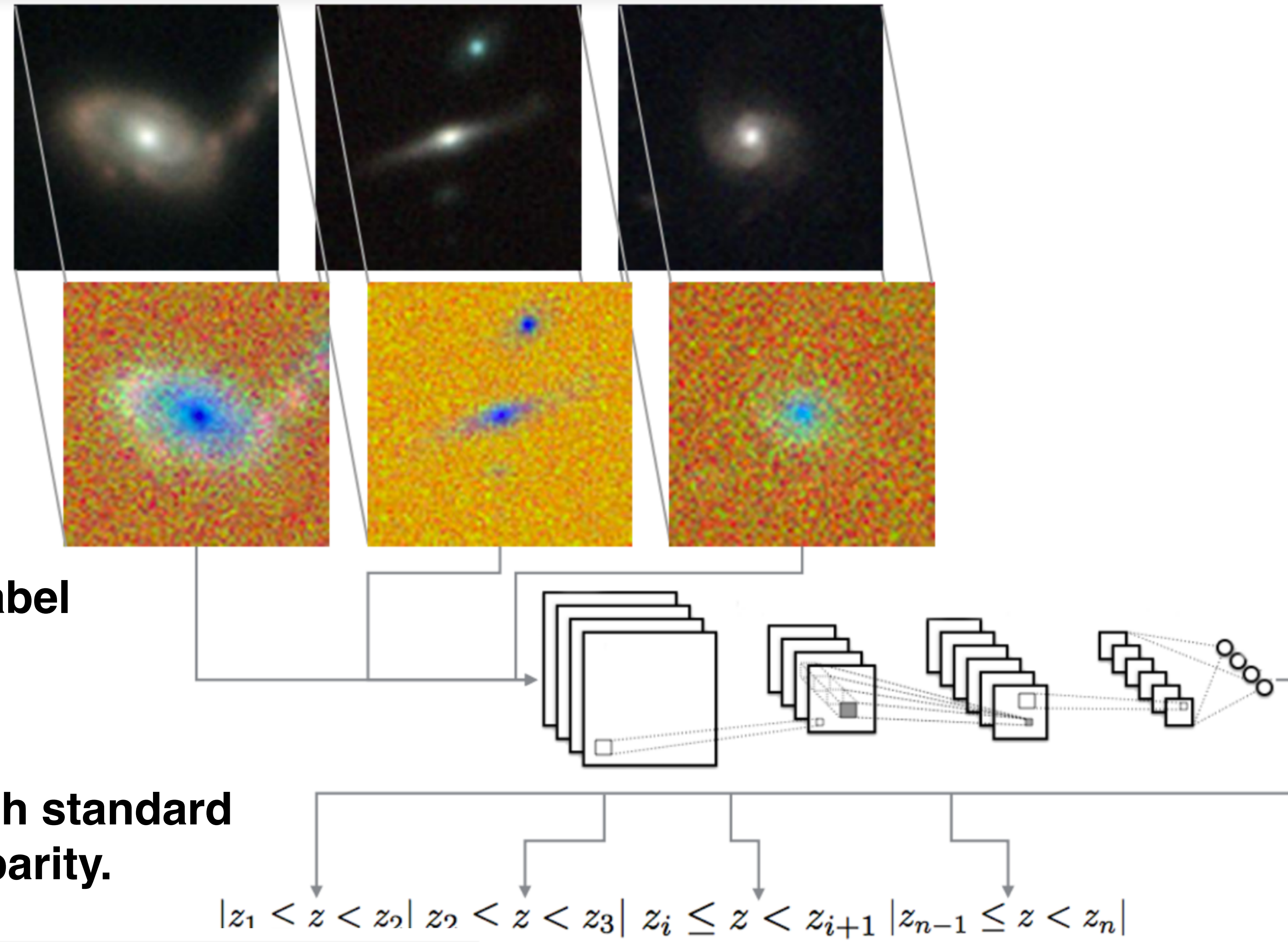
Dieleman et al 2015

http://benanne.github.io/2014/04/05/galaxy-zoo.html

CNNs for redshift estimates

. arXiv:1504.07255 [pdf, other]
Measuring photometric redshifts using galaxy images and Deep Neural Networks
Ben Hoyle

Inputs: galaxy image
->
ImageNet architecture
->
Targets: spec-z



everything about biased label data is still a problem

Compared performance with standard ML algorithms, and found parity.

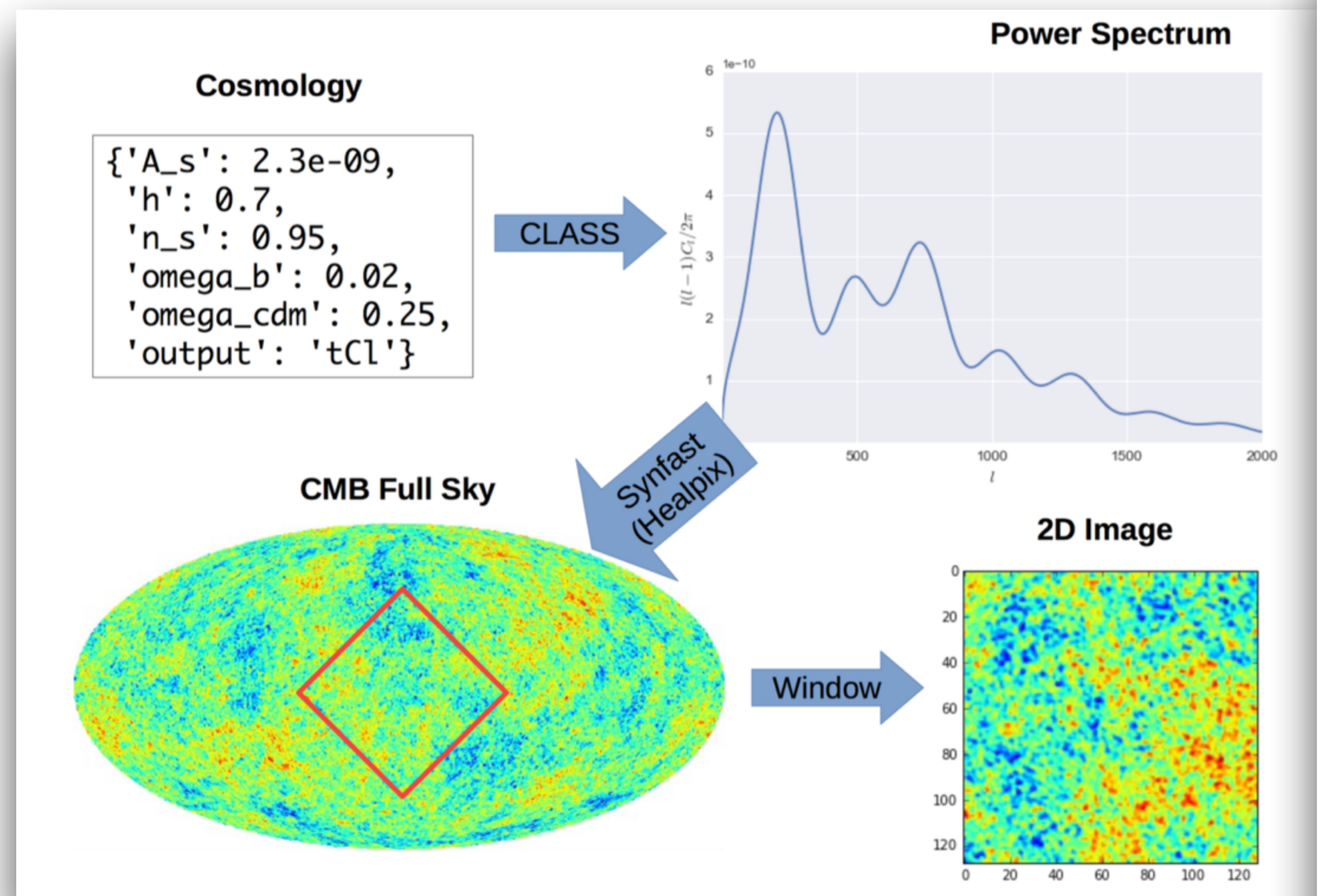
MLA	μ	σ_{68}	σ_{95}	$\lvert\Delta\ /(1+z_{spec})\rvert > 0.15$
DNNs	0.00	0.030	0.10	1.71%
AdaBoost	−0.001	0.030	0.10	1.56%

$$\Delta = z_{spec} - z_{predict}$$

CNNs for Cosmic Microwave Background radiation

Is there information in the CMB that is not contained in Cls? E.g. Higher order moments, such as non-Gaussianities.

Measuring Cosmological Parameters from Simulated CMB Images with Convolutional Neural Networks



2D CNN Configuration
input (128×128)
Conv2D (3×3) - 16 Conv2D (3×3) - 16
maxpool (2×2)
Conv2D (3×3) - 32 Conv2D (3×3) - 32
maxpool (2×2)
Conv2D (3×3) - 64 Conv2D (3×3) - 64
maxpool (2×2)
Conv2D (3×3) - 128 Conv2D (3×3) - 128
maxpool (2×2)
FC - 256
FC - 128
FC - 1 / FC - 2

1D CNN Configuration
input (16384)
Conv1D $(4, \mathit{Stride}\ 4)$ - 128 Conv1D $(4, \mathit{Stride}\ 4)$ - 128
maxpool (4)
Conv1D $(4, \mathit{Stride}\ 4)$ - 256 Conv1D $(4, \mathit{Stride}\ 4)$ - 256
maxpool (4)
FC - 256
FC - 128
FC - 1 / FC - 2

	ΔA_s	$\Delta\Omega_{CDM}$	$\Delta A_s^{(single)}$
PolSpice correlation function	$1.45 \cdot 10^{-10}$	0.025	$3.3 \cdot 10^{-11}$
2D CNN	$1.68 \cdot 10^{-10}$	0.0357	$7.19 \cdot 10^{-11}$
1D CNN	$1.91 \cdot 10^{-10}$	0.0437	-

Robert Lohmeyer Master thesis 2017

Supervisor BH

A random sample of CNN papers

Spectral classification using convolutional neural networks
https://arxiv.org › cs
by P Hála - 2014 - Cited by 2 - Related articles
Dec 29, 2014 - This thesis is about training a **convolutional neural network** (ConvNet) to ... neural networks and deep learning methods in **astrophysics**.

Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks

Yashar D. Hezaveh, Laurence Perreault Levasseur, Philip J. Marshall

arXiv:1704.02744 [pdf, other]
Finding strong lenses in CFHTLS using convolutional neural networks
Colin Jacobs, Karl Glazebrook, Thomas Collett, Anupreeta More, Christopher McCarthy
Comments: 16 pages, 8 figures. Accepted by MNRAS
Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Astrophysics of Galaxies (astro-p

A Convolutional Neural Network For Cosmic String Detection in CMB Temperature Maps

Razvan Ciuca, Oscar F. Hernández, Michael Wolman

(Submitted on 29 Aug 2017)

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

Summary/Conclusions

Generative Adversarial Networks (GANs)

Generative:
Deep ML NN1: Input (random noise) vector -> output something / image

Adversarial:
Deep ML NN2: distinguish examples of training data examples from non-training data, e.g. that obtained from NN1

Networks:
Deep ML Convolution Neural Networks.

As training proceeds, NN1 generates more and more realistic "examples" from a random noise vector, and NN2 get better and better at distinguishing training data, from everything else, e.g that generated by NN1.

The problem with GANs:
Mode collapse. Difficult learning —> Wasserstein GAN.
https://arxiv.org/abs/1701.07875

https://github.com/bobchennan/Wasserstein-GAN-Keras/blob/master/mnist_wacgan.py
https://raw.githubusercontent.com/farizrahman4u/keras-contrib/master/examples/improved_wgan.py

Recent GAN applications

GANs to peer within a galaxy image: sub PSF properties of galaxies. Schawinski et al 2017

GANs produce one realisation of what the input galaxy could look like.
http://space.ml/supp/GalaxyGAN.html

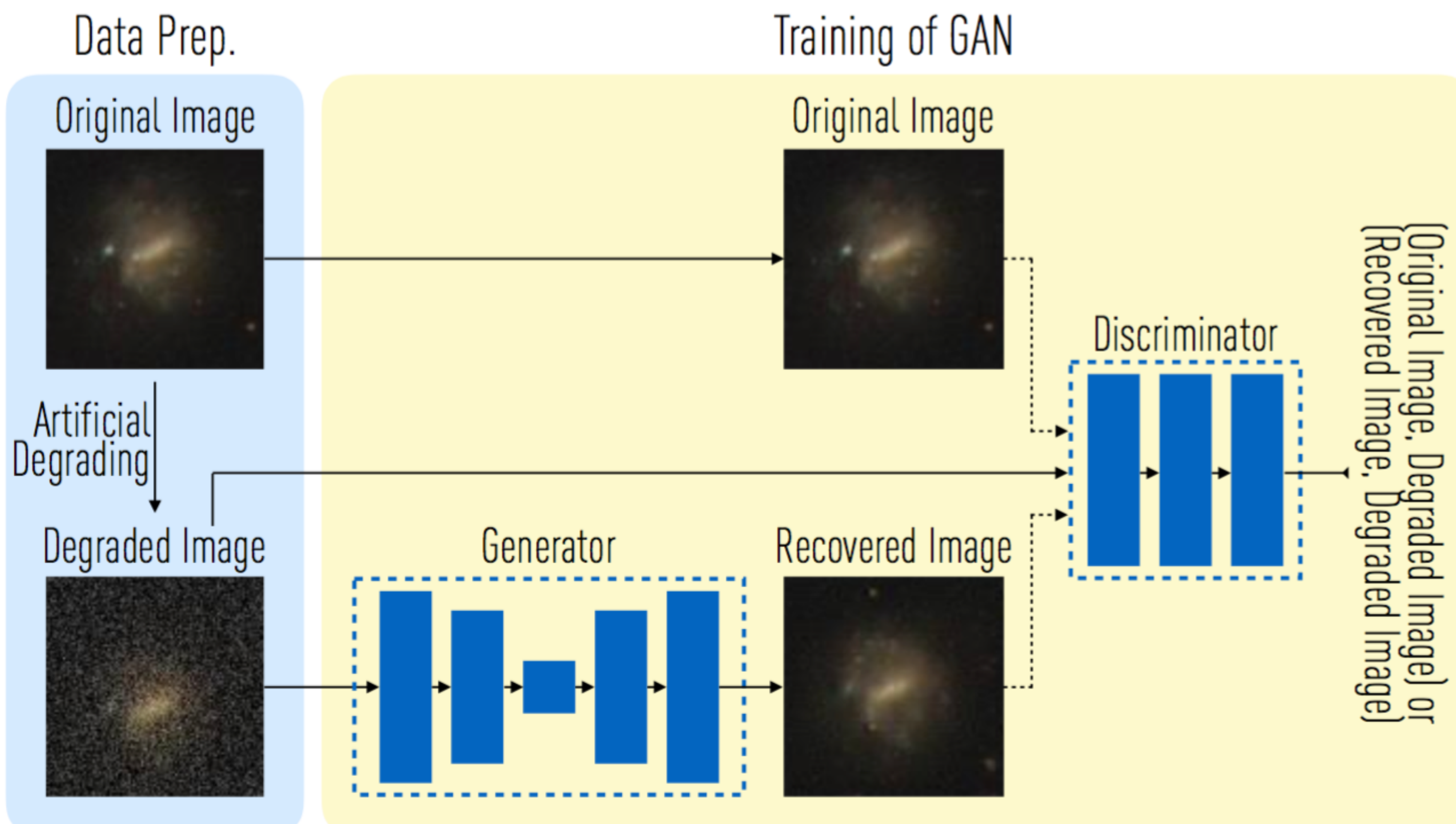
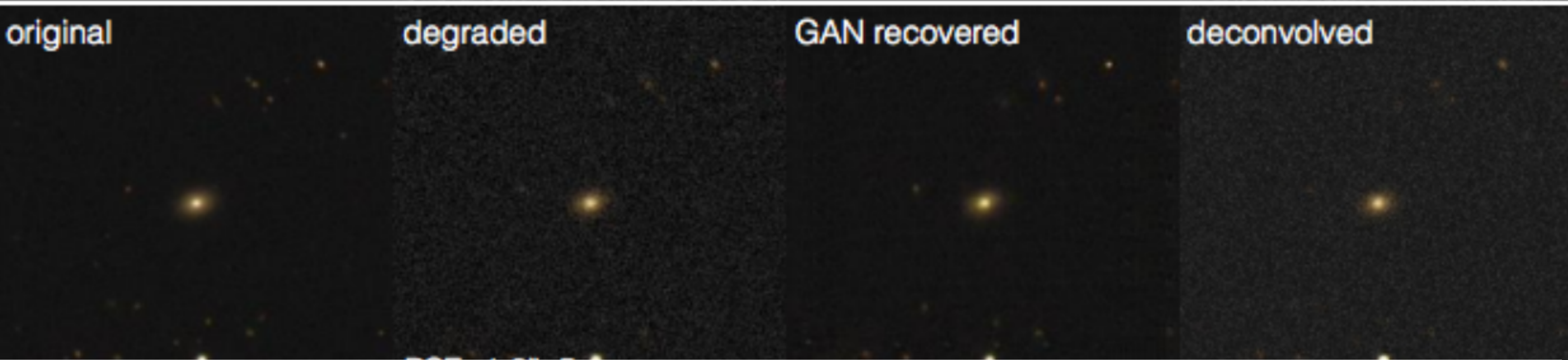


Figure 1. Schematic illustration of the training process of our method. The input is a set of original images. From these we automatically generate degraded images, and train a Generative Adversarial Network. In the testing phase, only the generator will be used to recover images.



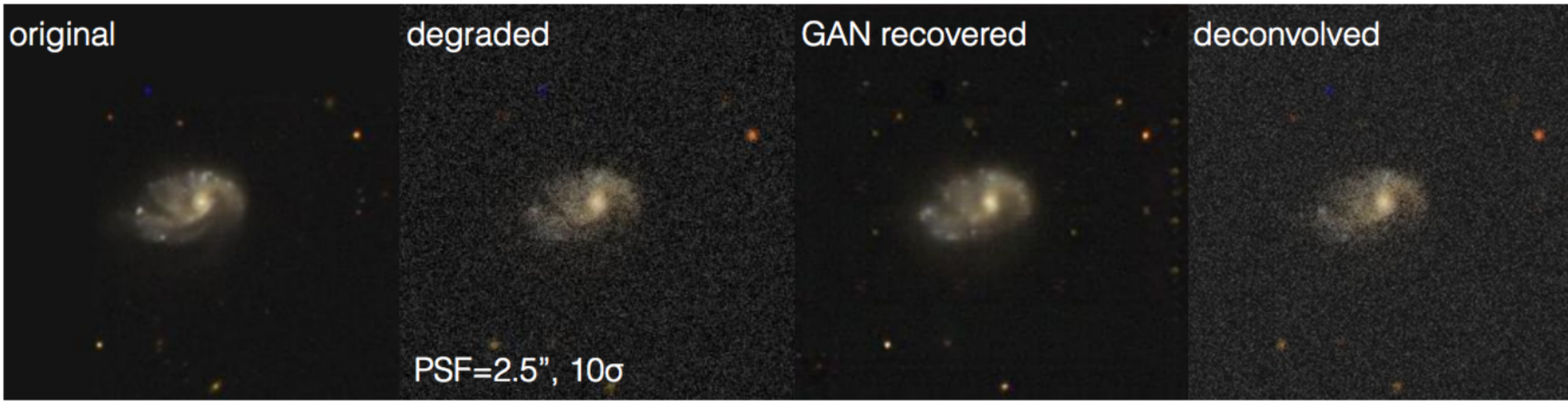
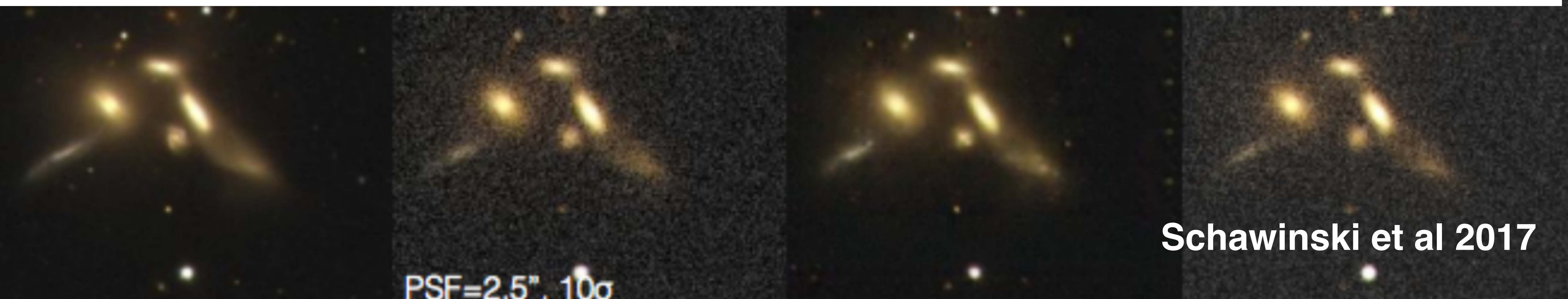


Figure 2. We show the results obtained for one example galaxy. From left to right: the original SDSS image, the degraded image with a worse PSF and higher noise level (indicating the PSF and noise level used), the image as recovered by the GAN, and for comparison, the result of a deconvolution. This figure visually illustrates the GAN's ability to recover features which conventional deconvolutions cannot.



Schawinski et al 2017

Recent GAN applications

GANs to peer within a galaxy image: sub PSF properties of galaxies. Schawinski et al 2017

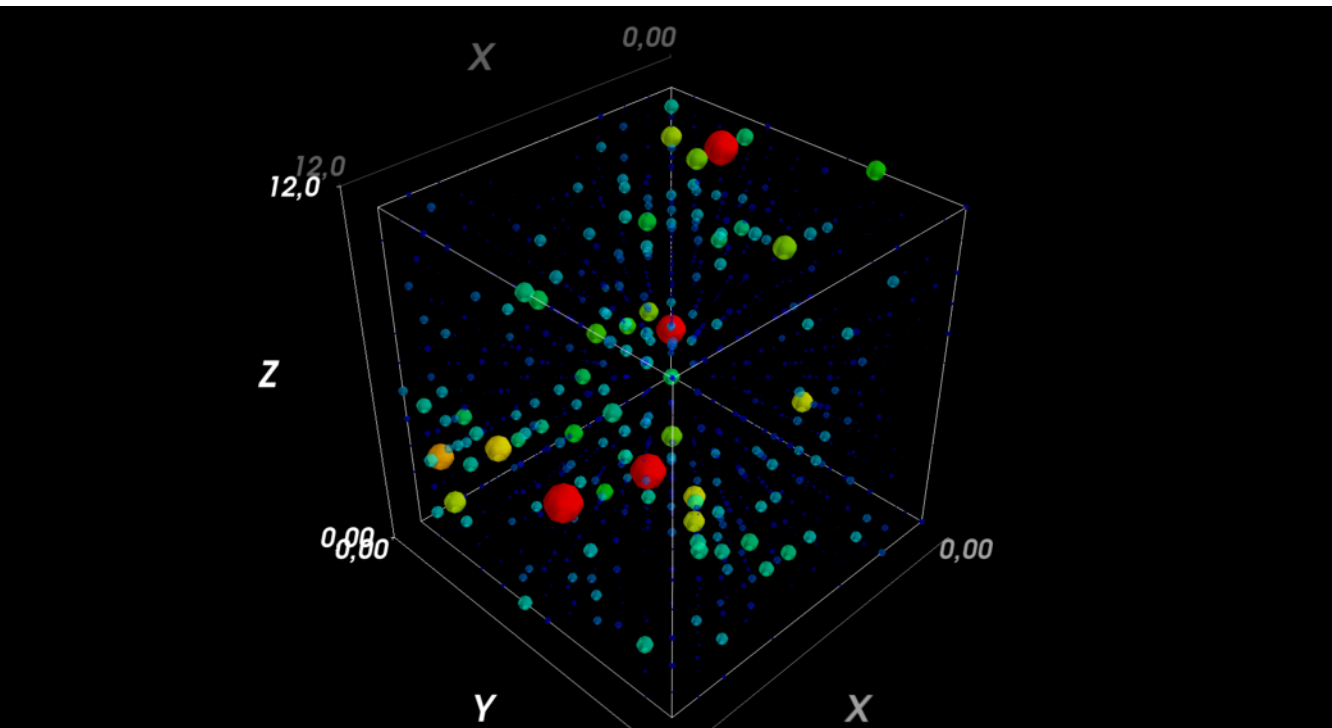
GANs produce one realisation of what the input galaxy could look like.
http://space.ml/supp/GalaxyGAN.html

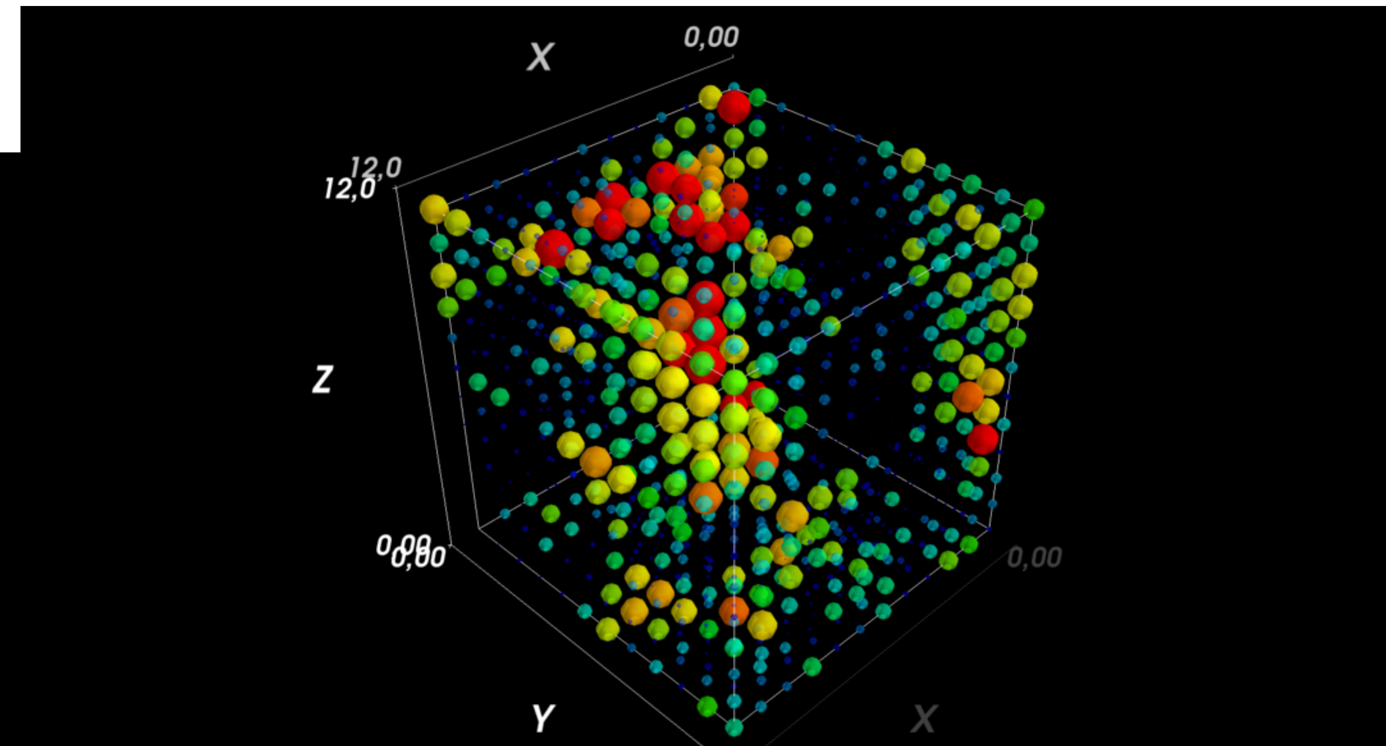
Getting "labels" for the science sample data one cares about, is very challenging.

Again, move towards higher order measurements of the predicted signal:
E.g. does gas predicted to exist in some part of the galaxy/disk give off radiation which can be observed in other bands?

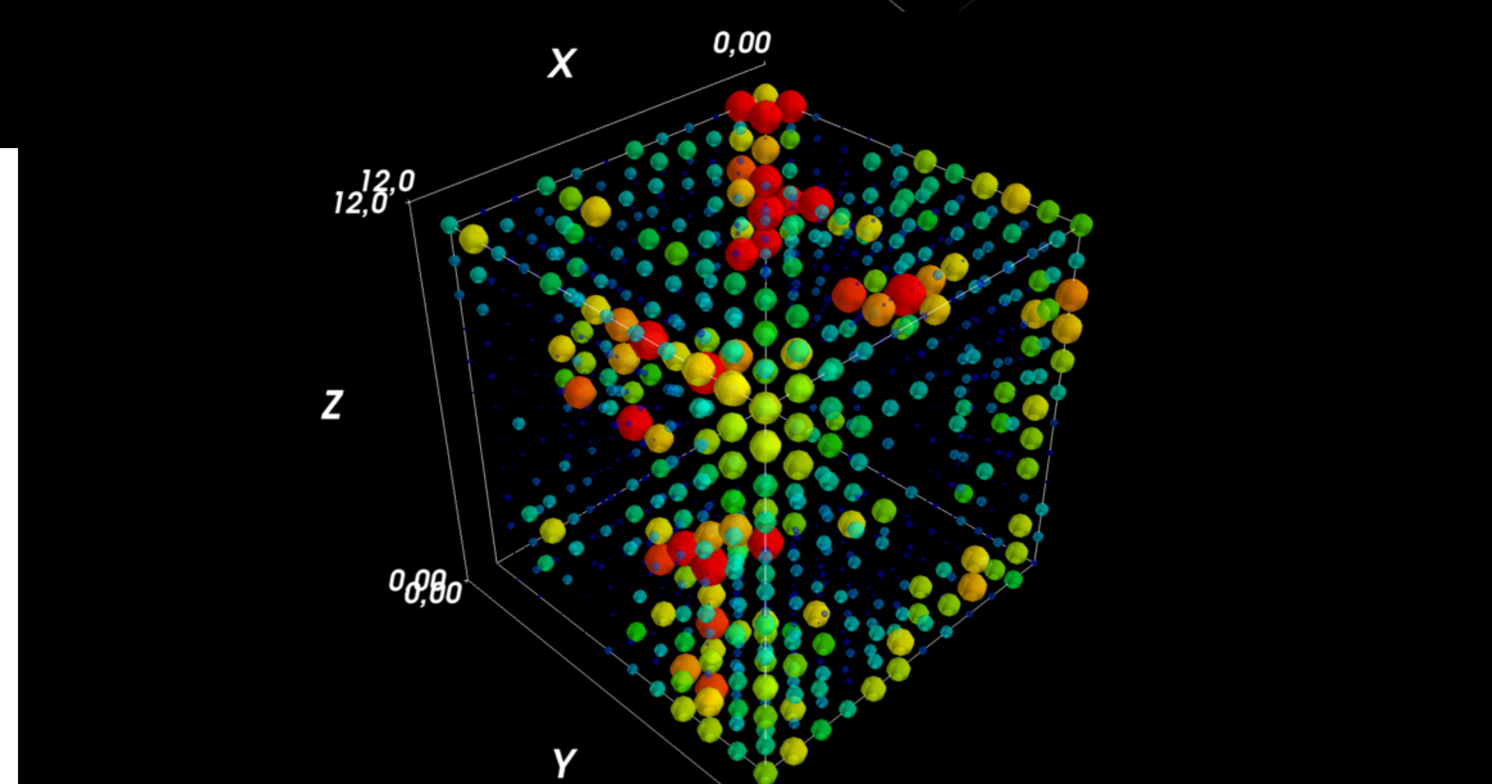
GANs to generate a realisation of a Dark-Matter N-body simulation.

In essence we replace a very computationally expensive Nbody simulation code, like Gadget, with a Deep 3-d CovNet
—ongoing work with Julien Wolf

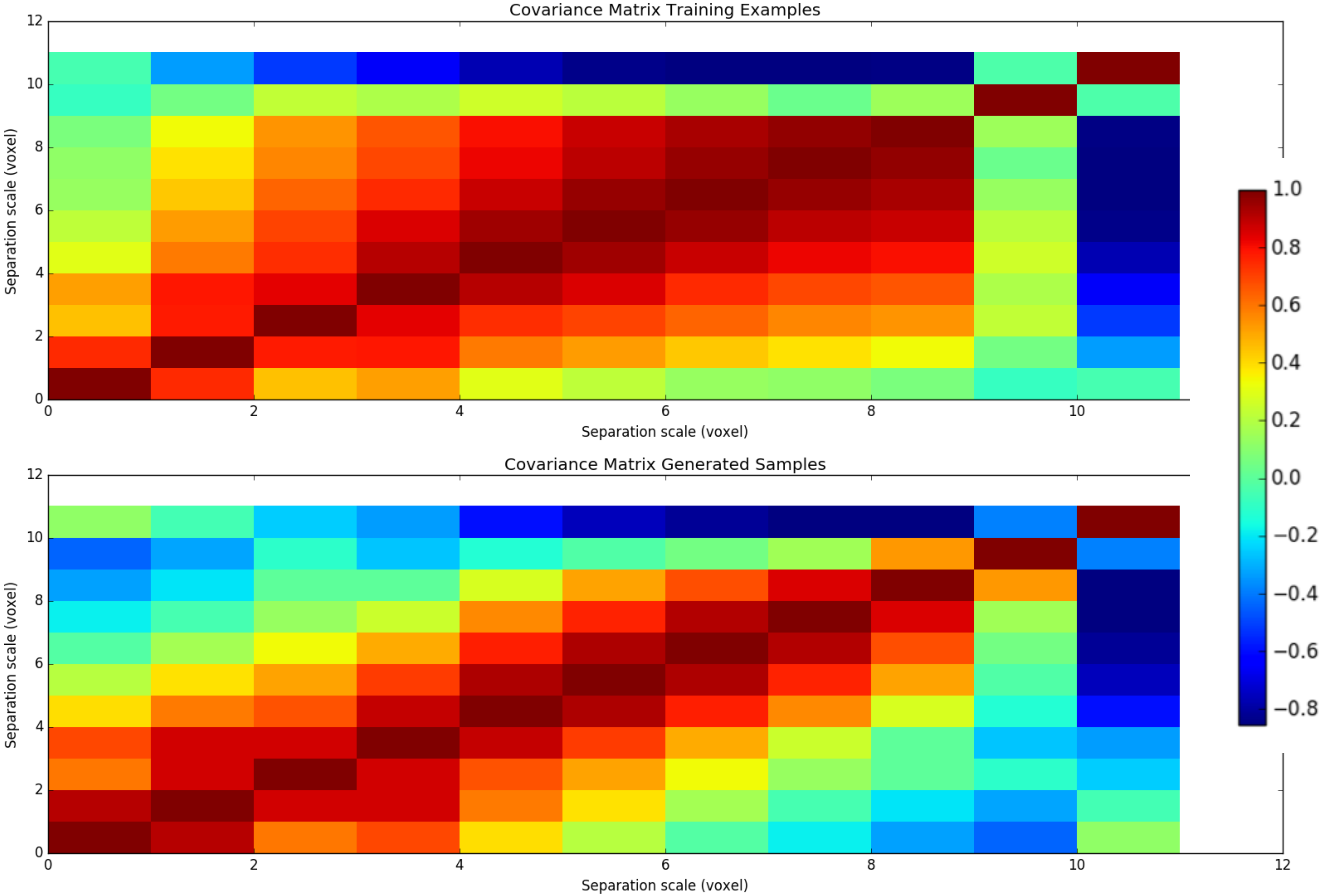




If we want to measure covariance matrices for correlation functions to estimate BAOs, we have to call Gadget many 100's - 1000s of times.



Julien Wolf (USM) Master Student



Julien Wolf (USM) Master Student

Overview

Photometric redshifts for cosmology

Machine learning workflow

The biggest problem for ML in cosmology:
Unrepresentative labelled data

Dealing with unrepresentative labelled data

Other common applications of ML

Recent, novel applications of ML

Summary/Conclusions

New Algorithms for ML / applied to astrophysics

Random forests / Decision tree based methods — with MINT (He et al 2013) feature selection.

Algorithm Novelty:

Grow a decision tree, but rather than randomly selecting from the input features (X), we can use both the "shape of X on the science sample" and the shape of X in the labelled data, as a guide to selecting which features the tree should choose. Mutual information defines the correlations (or "shapes").

Applicable if we have many 1000's of input features, which may be correlated, and the labelled data may have different input feature correlations from the unlabelled data.

Suryarao has working code on git-hub, and some very nice preliminary results on test data. We will move to real-world data soon.

Work in advanced progress with Suryarao Bethapudi.

Summary/Conclusions

Accessing new / existing data

Cosmology is in the realm of "big data"; 100's millions/ billions of galaxies are being observed: SDSS/DES/LSST/Euclid/LOFAR/SKA. Millions have target values.

Many possibilities of applying machine learning in new and interesting ways.

Some cosmological analysis is in a state of crisis:

Unrepresentative labelled data means we need new ideas, and potentially new algorithms.

Higher order measurements of predictions is one way to proceed.

Cutting edge algorithms being implemented in astrophysics/cosmology

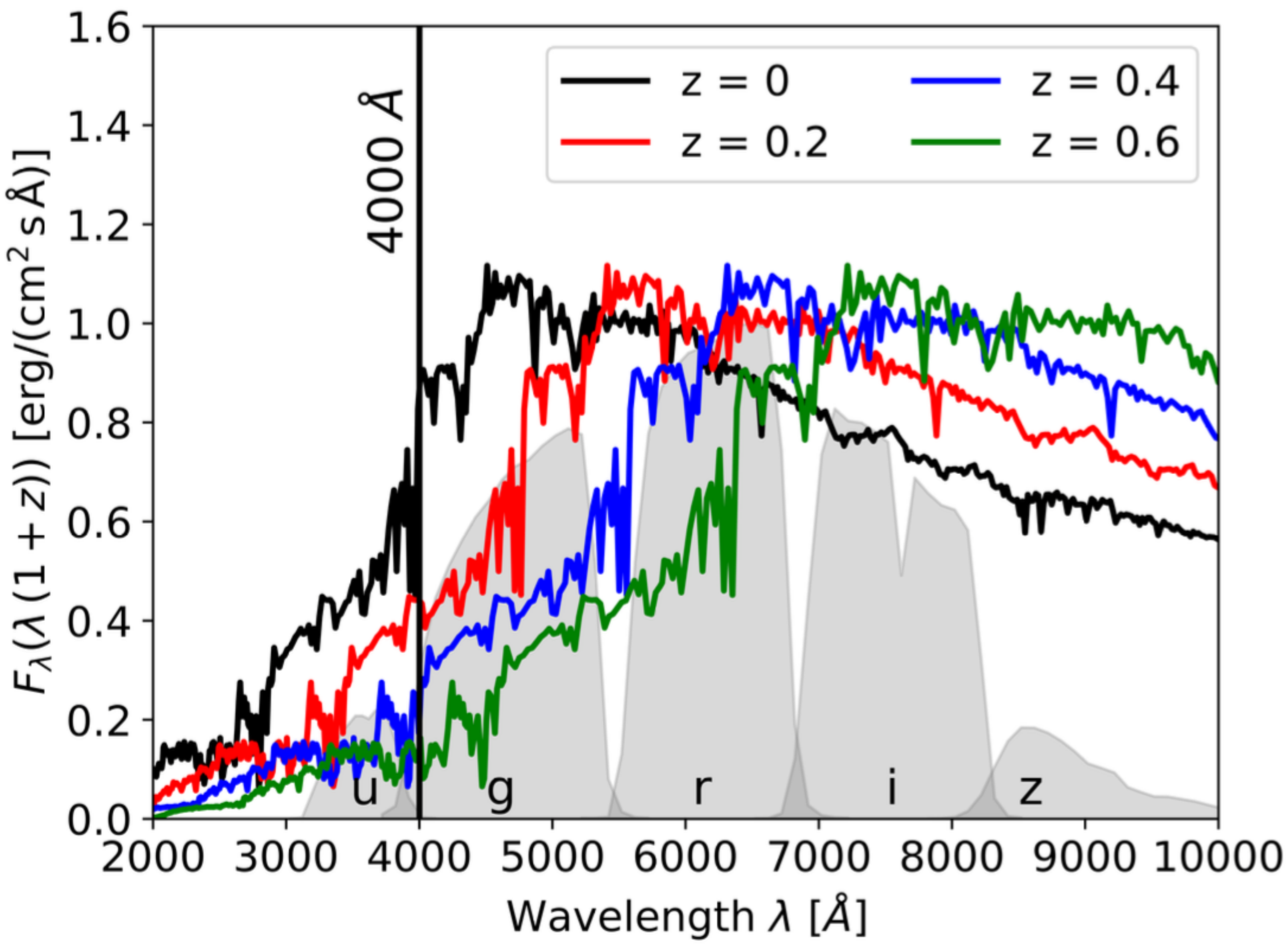
Deep ML: CNNs / GANs.

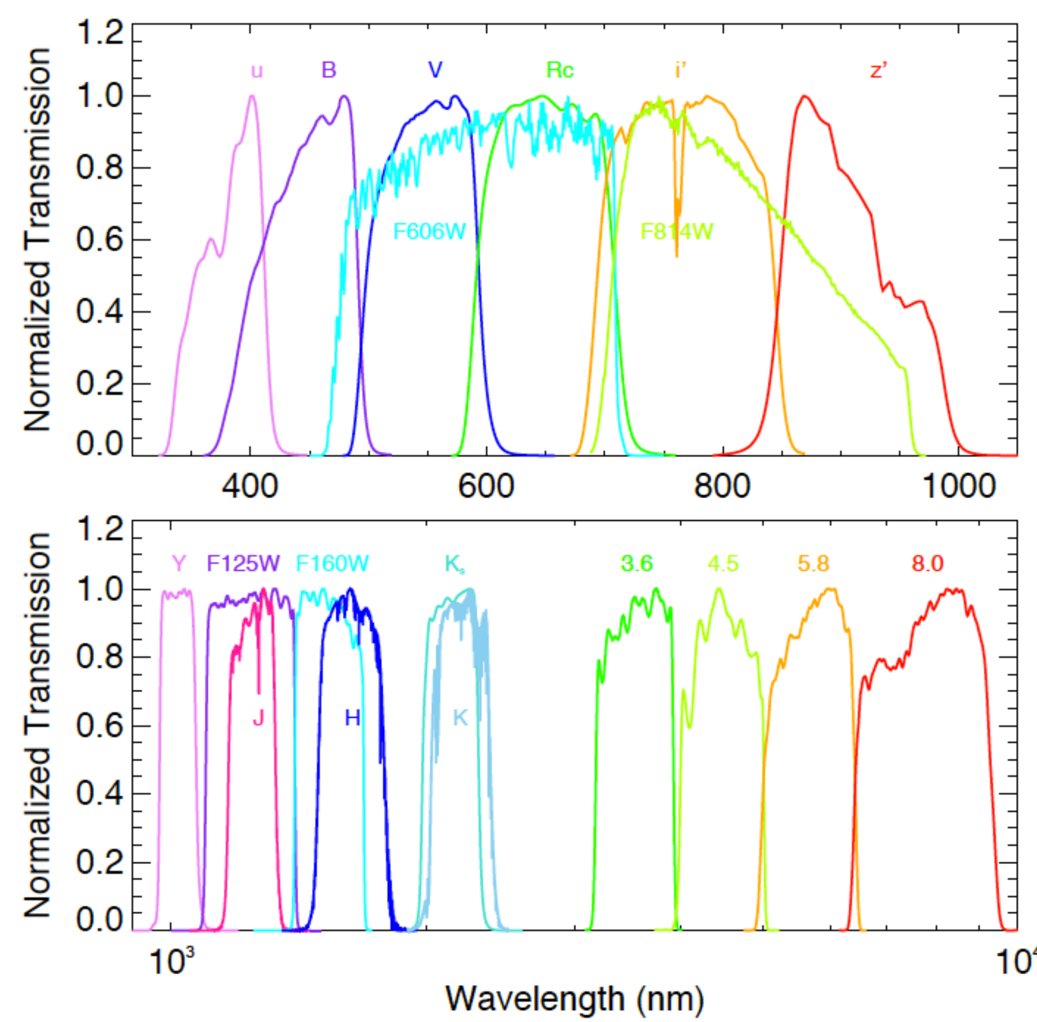
New algorithms being developed for ML, and ML in astrophysics/cosmology.

Photometric and spectroscopic redshifts

A spectrograph has a high wavelength resolution, allowing the ID of absorption/ emission lines, each with a “fingerprint”. Compare to the wavelength of these fingerprints measured in the lab, and lambda shift = redshift. — spec-z is expensive.

If instead we measure the spectrum in broader photometric filters, we convolve the true spectrum with the filter, and get one measurement per filter. One needs strong absorption features. — photo-z is cheap





Markus Rau 2017 Phd Thesis