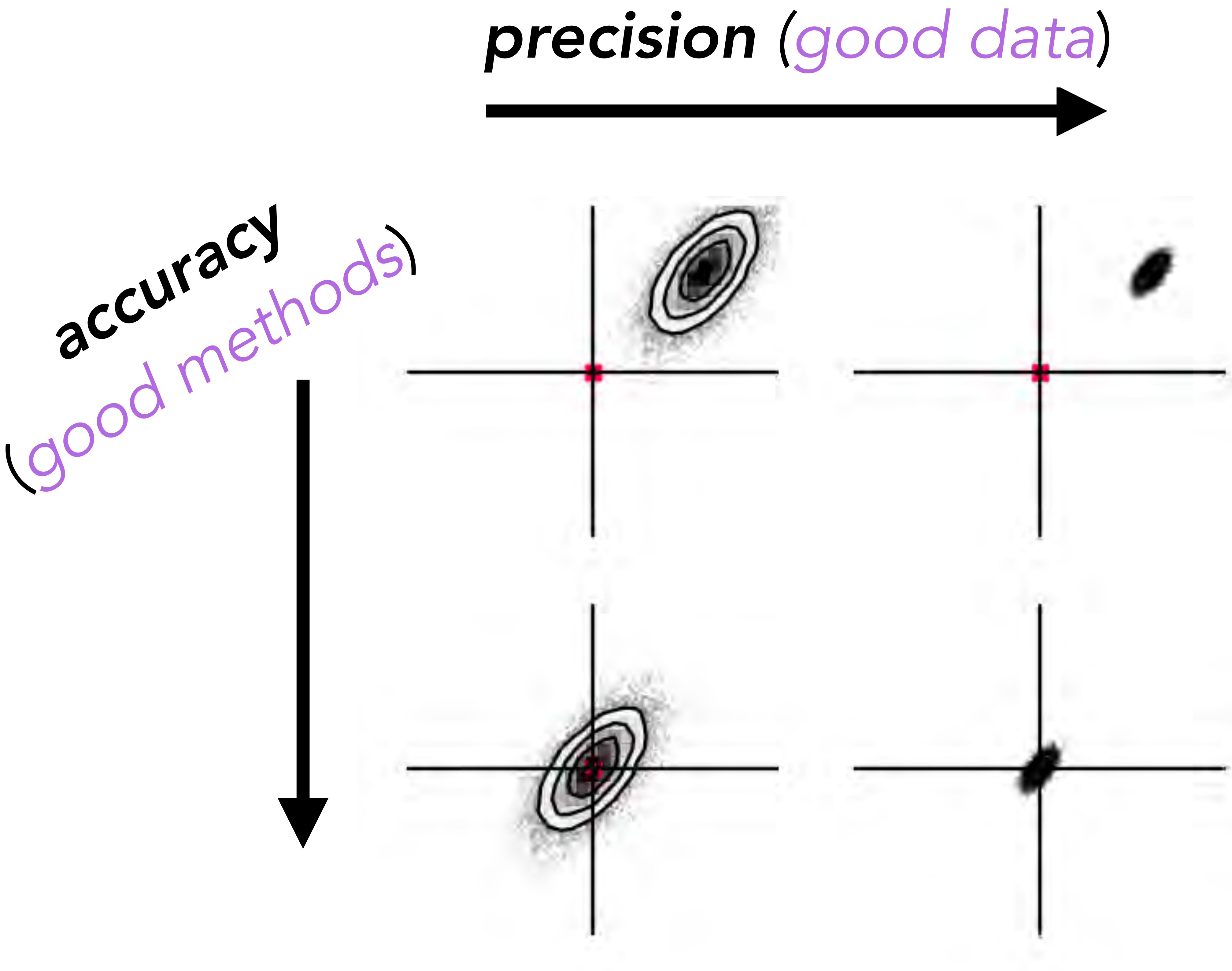
Cosmology with imaging surveys
– from **precision** to *accuracy*–

Interpretable photometric redshifts with no spectroscopy

Boris Leistedt – @ixkael, www.ixkael.com
NASA Einstein Fellow @ CCPP, New York University



Frontiers: probabilistic models & observational systematics

ROAD MAP

Cosmology with galaxy surveys

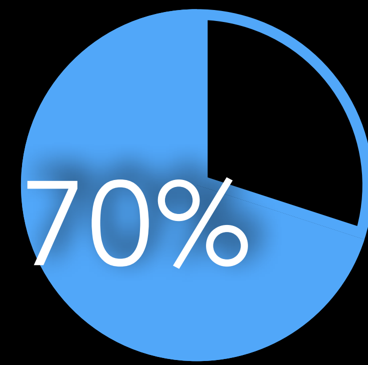
Photometric surveys and spatial systematics

Redshift distributions via hierarchical modeling

Photometric redshifts likelihood functions

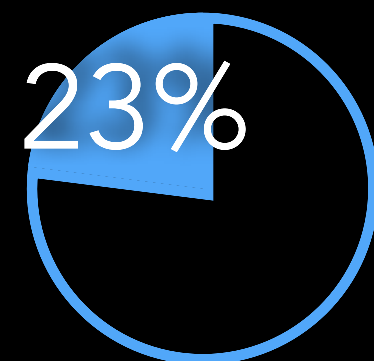
Conclusions & the future (LSST)

Energy density budget of the universe (today) and related questions



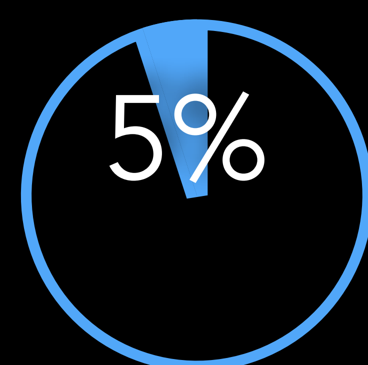
dark energy™

microscopic origin?
cosmological constant?



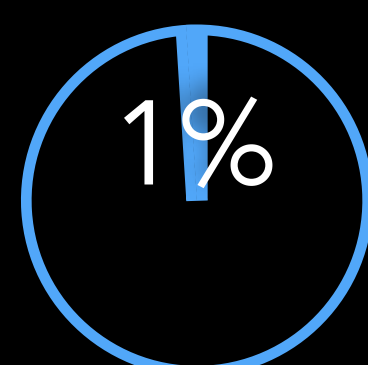
dark matter™

microscopic origin?
cosmological impact?



baryons

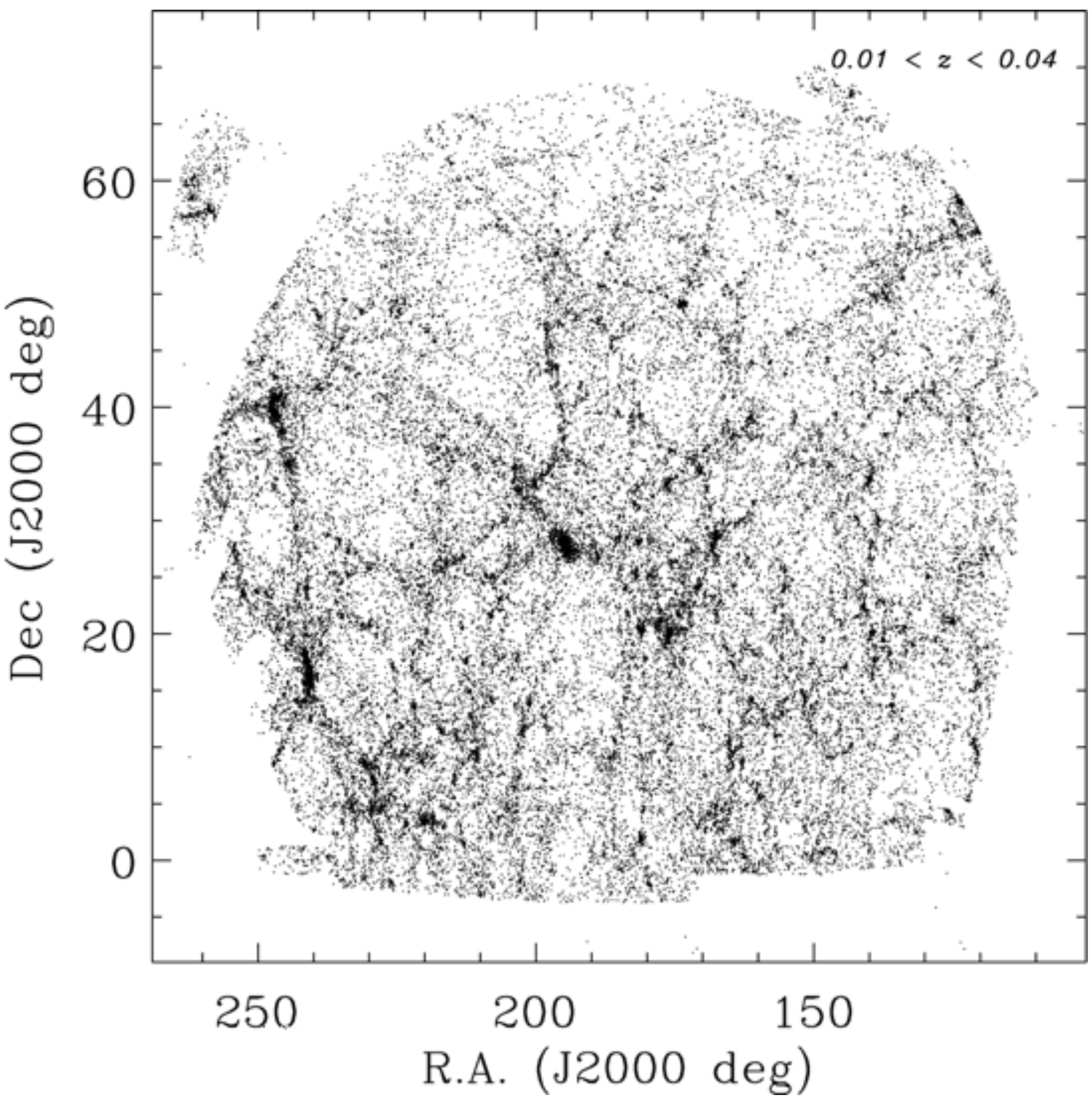
galaxy formation?
gastrophysics?



the rest

neutrinos, radiation, etc

Galaxy Surveys



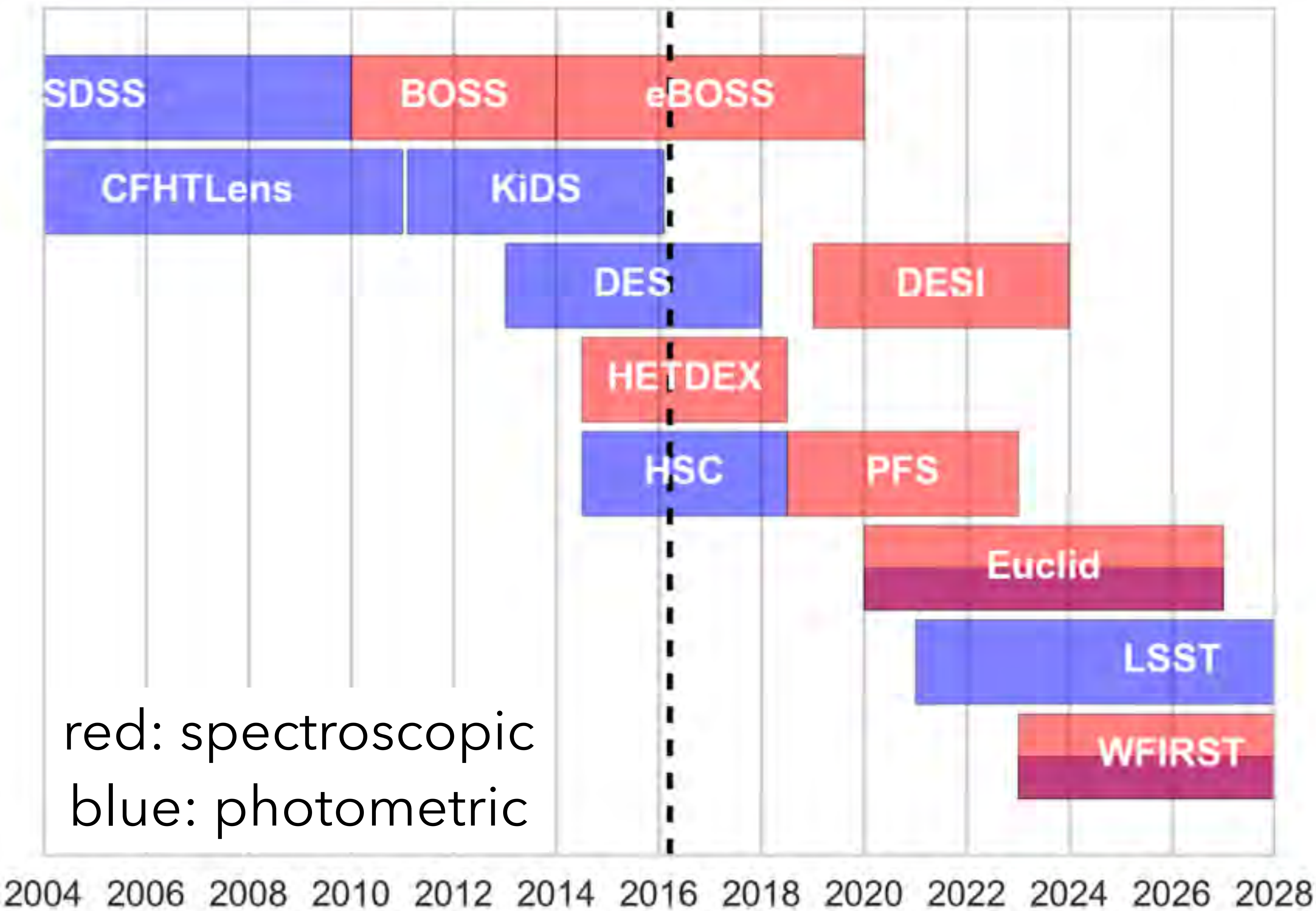
Rich space of models

early universe physics
astroparticles
gravity, etc

and observables

galaxy clustering
cosmic shear
galaxy-galaxy lensing
cross-correlations with
CMB, tSZ, etc

experimental landscape

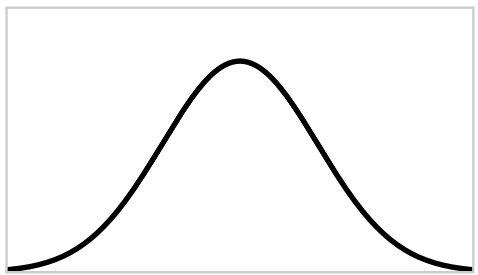


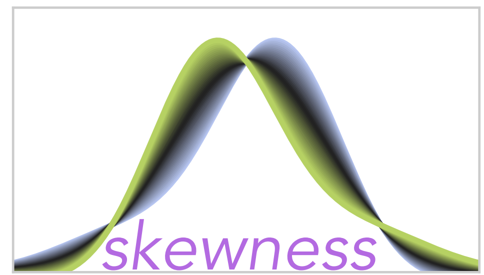
Galaxy surveys: goals

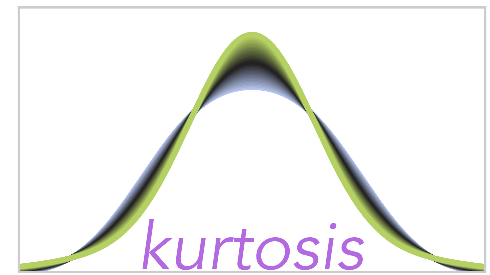
- Measure the expansion rate *H(z), growth rate f, and other cosmological parameters*
- *Constrain primordial non-Gaussianity (f_{NL}, g_{NL}, …)*
- *Constrain primordial potential / power spectrum*
- *Constrain gravity from large-scale effects*

Example: primordial non-Gaussianity

- Local type: $\Phi = \phi + f_{\mathrm{NL}}[\phi^2 - \langle\phi^2\rangle] + g_{\mathrm{NL}}[\phi^3 - 3\phi\langle\phi^2\rangle]$







- Imprinted in 3+ point correlations of the CMB and in large-scale galaxy power spectrum (2-point!)
- **<u>Current limits</u>**: $-3.1 < f_{\mathrm{NL}} < 8.5$ from CMB (Planck collaboration 2015)
  $-16 < f_{\mathrm{NL}} < 26$ from SDSS galaxies (Giannantonio+ 2014)
  $-39 < f_{\mathrm{NL}} < 23$ from SDSS quasars (Leistedt+ 2014)
- Measurement ***<u>plagued</u>*** by spatial and redshift systematics
- LSST & SphereX could detect fNL ~ 1 but will be difficult to reach!

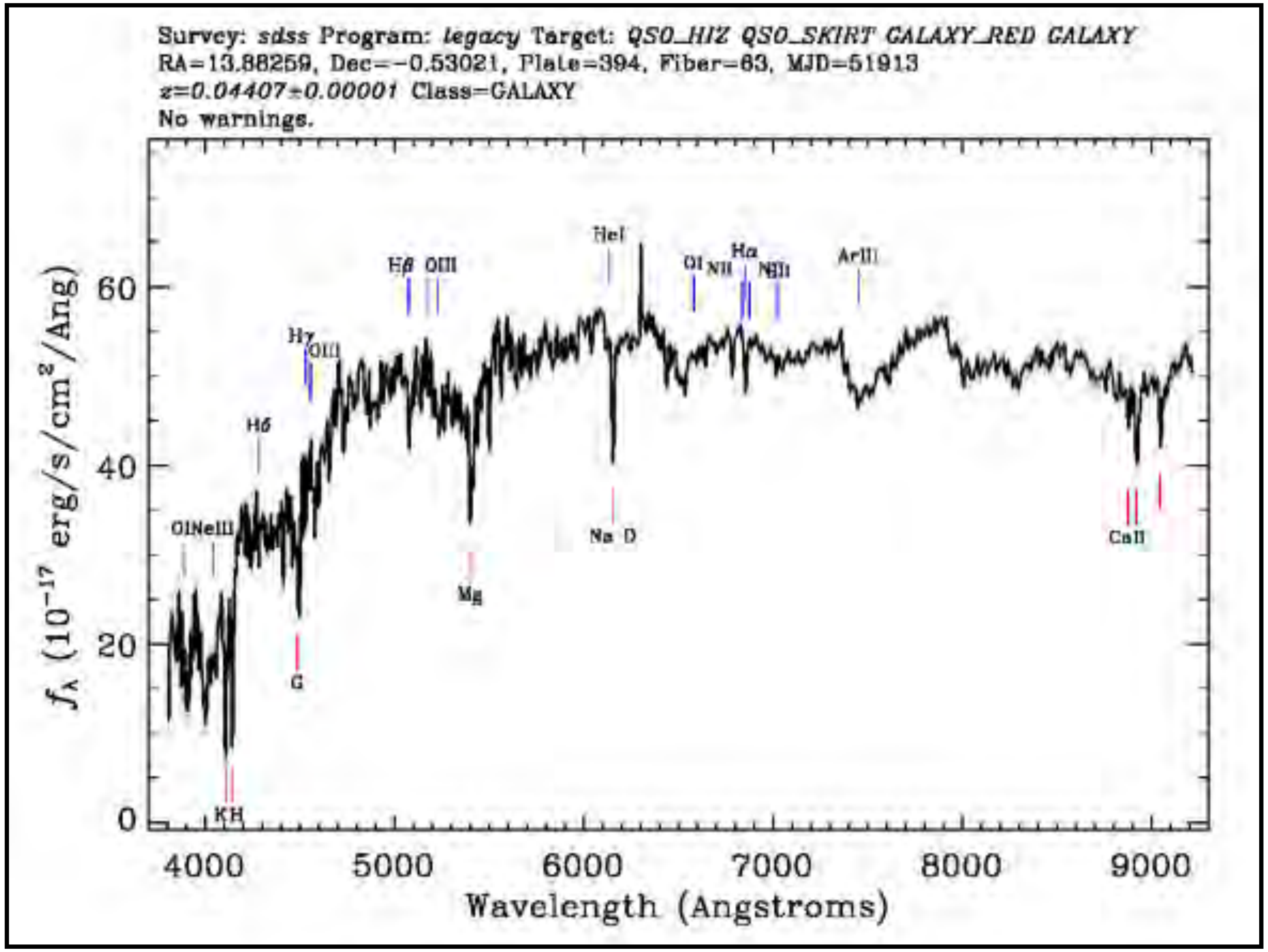
spectroscopic

vs

<u>photometric</u>

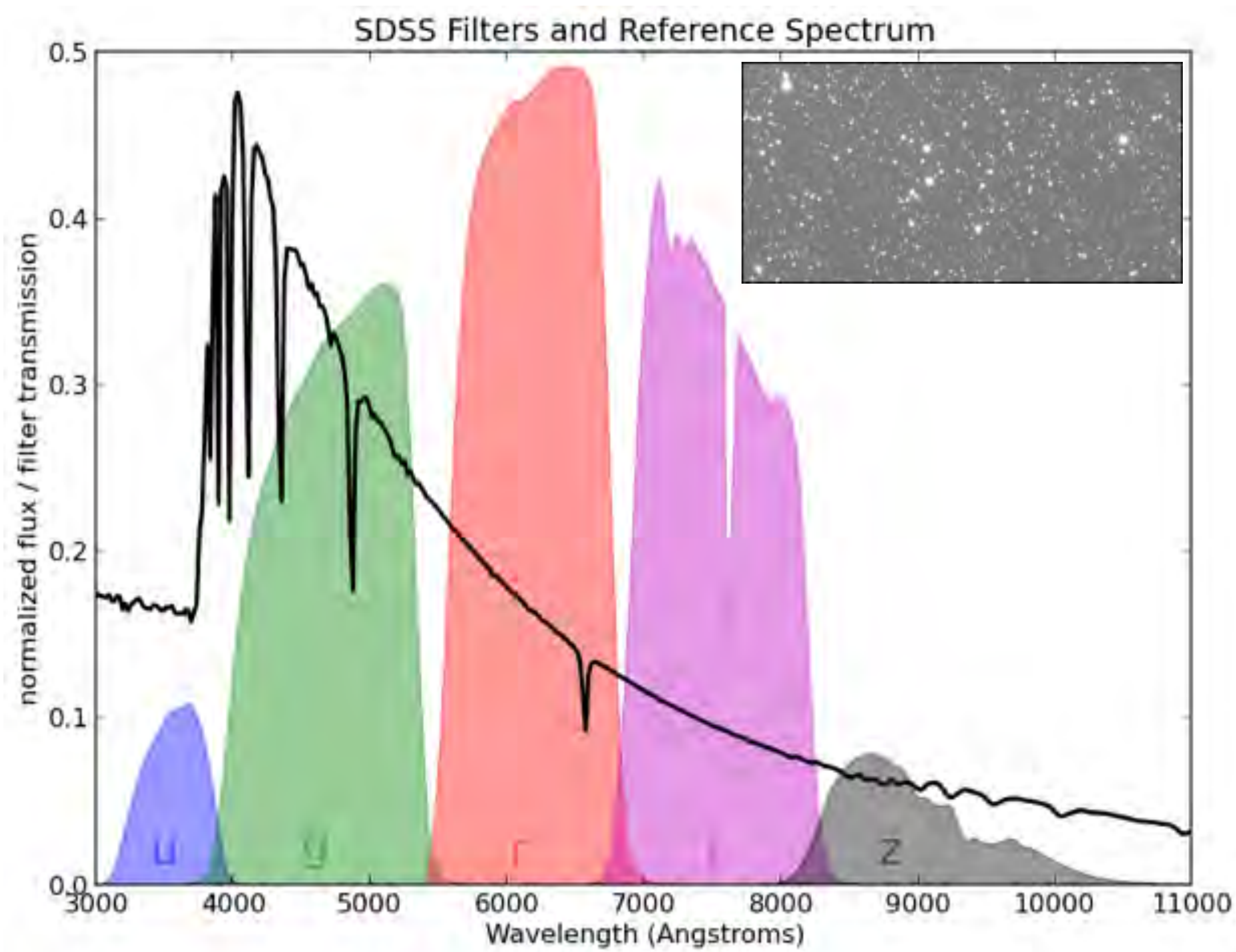
surveys

spectroscopic



- ✓ SEDs
- ✓ types
- ✓ redshifts
- ✗ shallow
- ✗ no shear

photometric



- ✓ CCD images
- ✓ deep
- ✓ shear
- ✗ no types
- ✗ no redshifts

spectroscopic

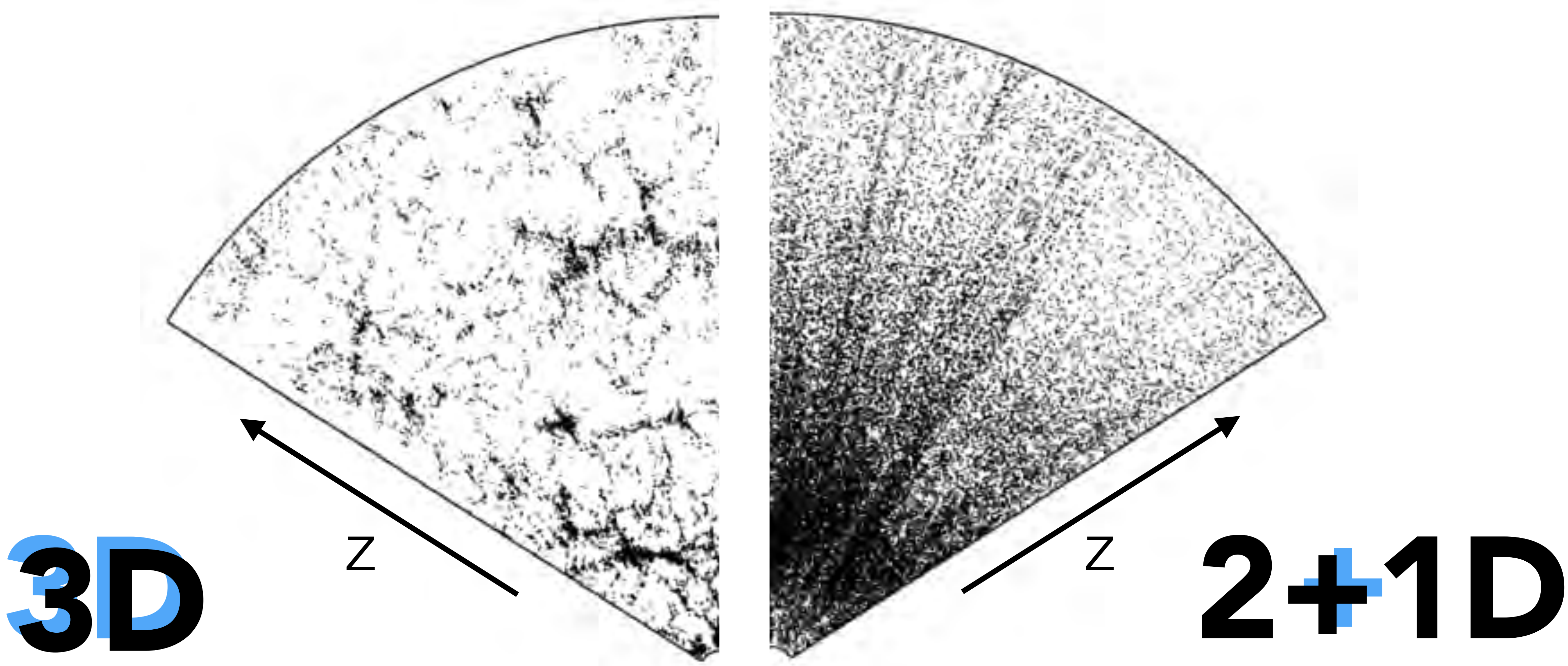
photometric

✓ types + redshifts

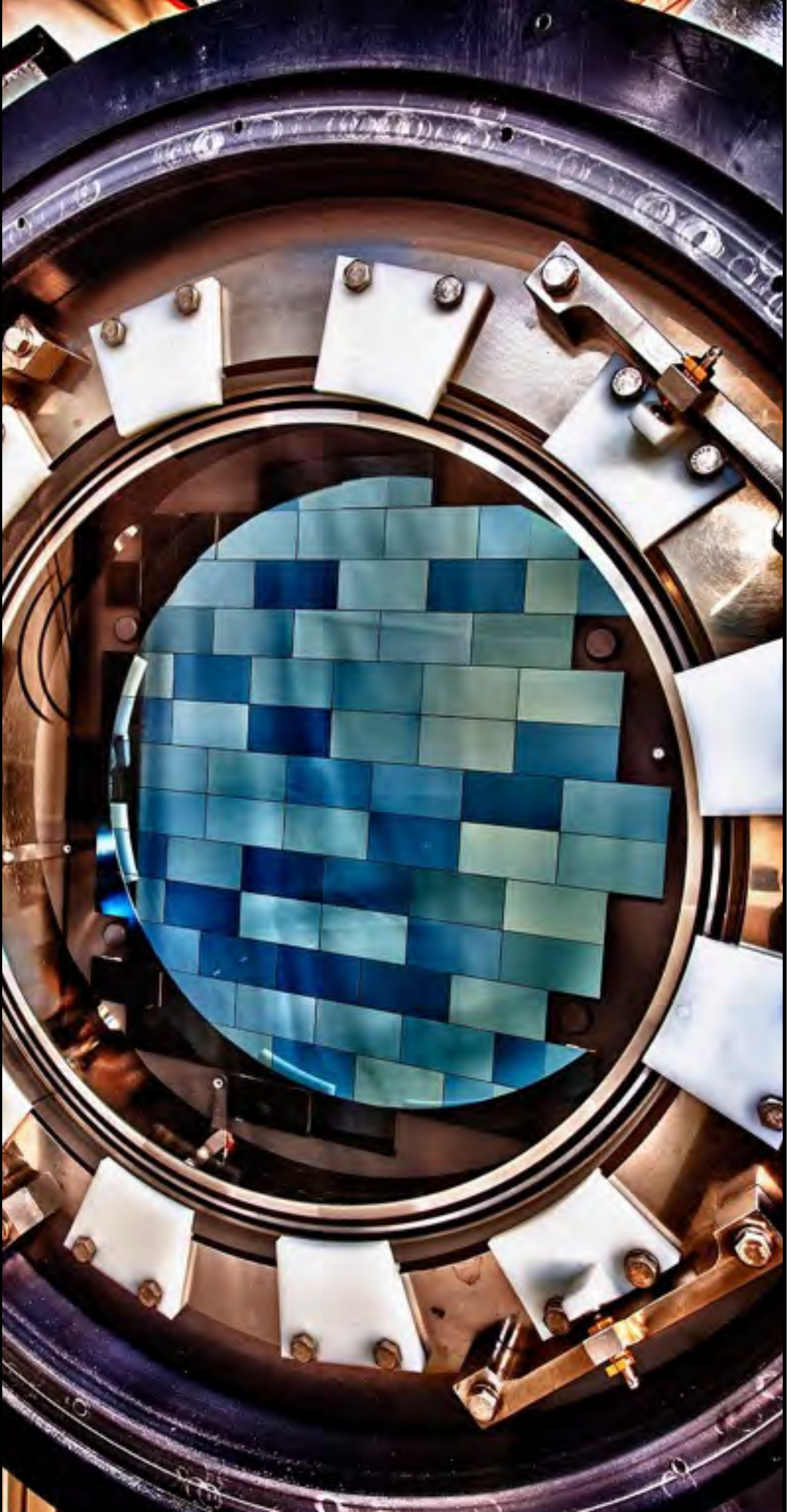
✗ no types / redshifts

✗ shallow

✓ deep



credit: Aragon-Calvo et al (2014)



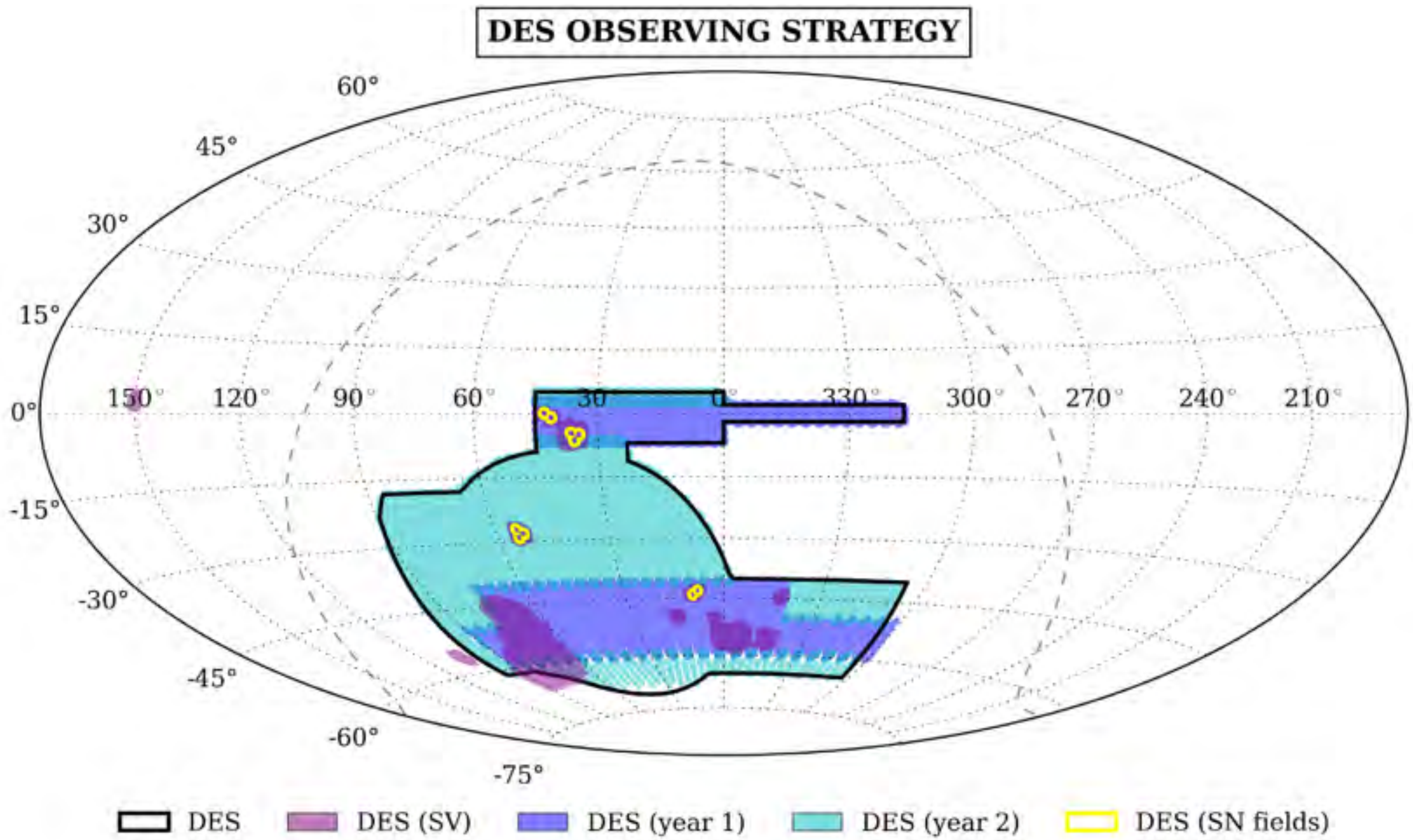
DES —
The Dark Energy Survey

3 deg^2 FOV, 570 Mpixel camera on Blanco telescope (4m, CTIO)

2013-2018, 300 collaborators

grizY bands

4 standard observational probes: galaxy clustering, galaxy lensing, supernovae, clusters



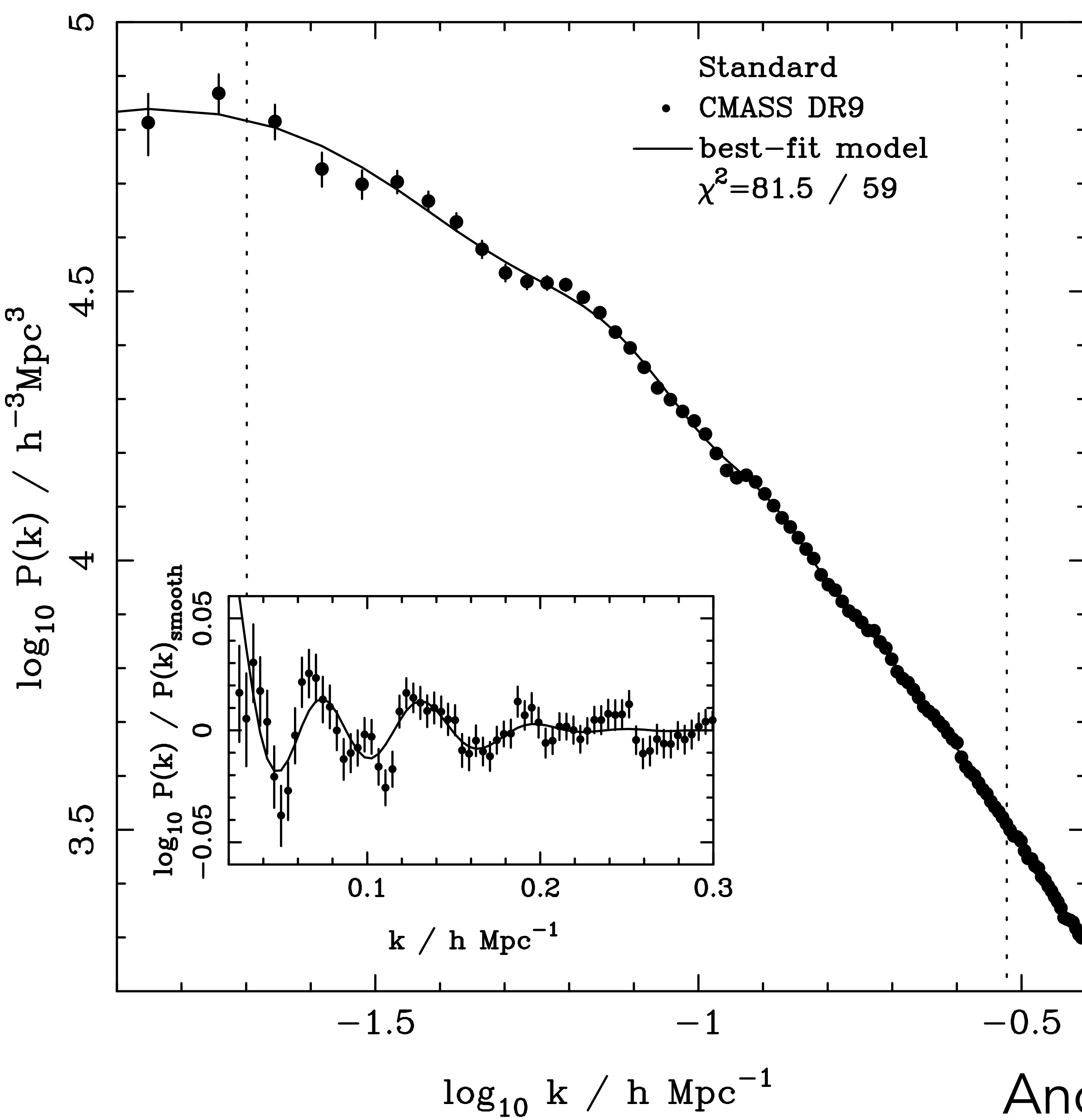
Full survey: 5000 deg^2 ***grizY*** *to 24th mag*

Science Verification: *~150 deg^2 to full depth*

Year 1: *~1500 deg^2 but shallower*

In the can *(Y2-3): 5000 deg^2 deeper than Y1*

3D matter power spectrum measured with spectroscopic survey



Anderson et al (2012)

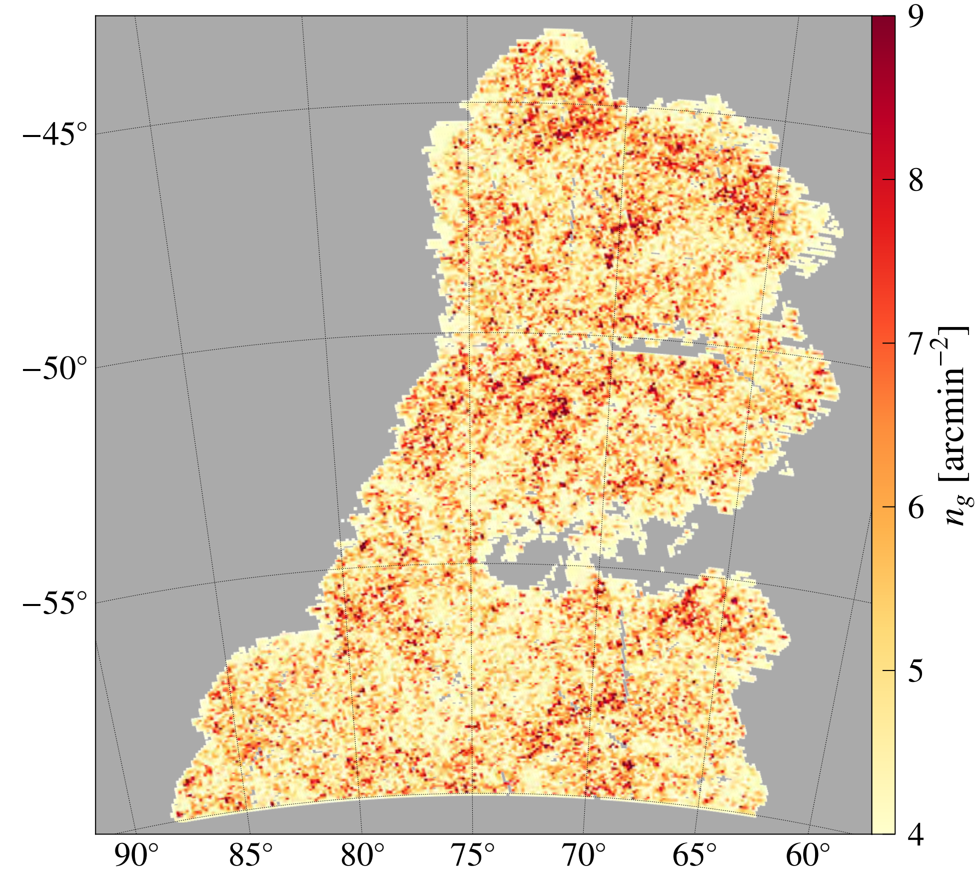
Photometric surveys

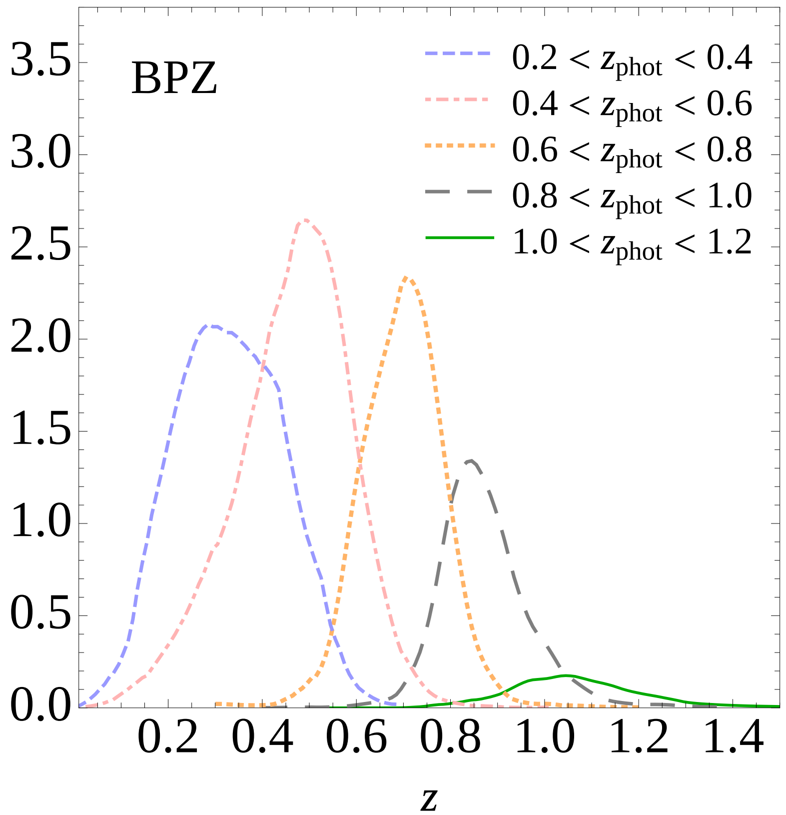
Redshifts are estimated from broad-band photometry and are **uncertain**

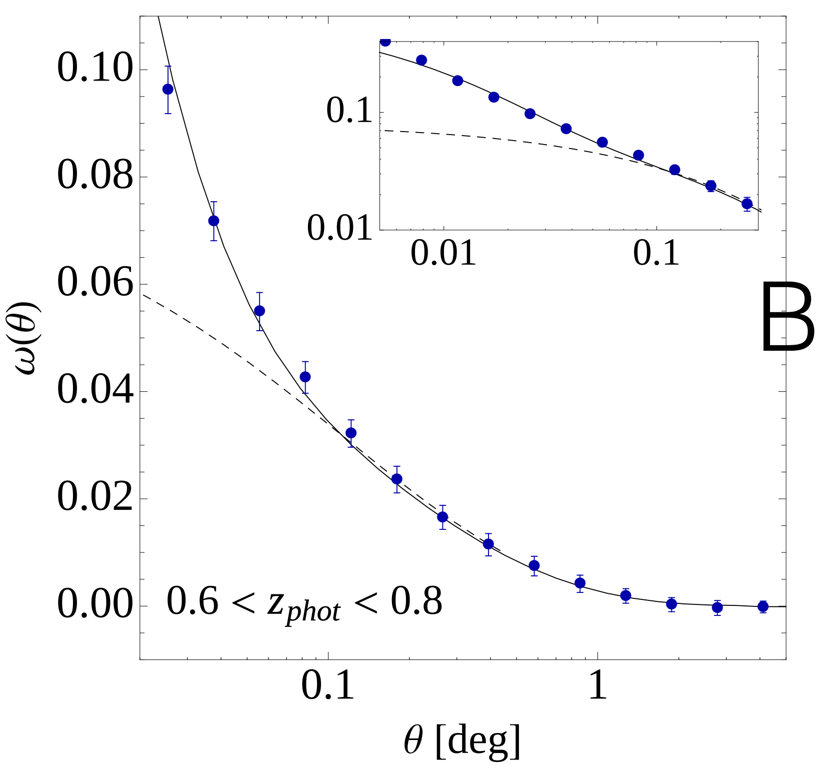
Typical approach: **tomographic analysis**
group galaxies into bins using a redshift estimate, then do 2D angular 2-point correlation analysis

DES Science Verification (SV) data

- Early data, excellent quality
- Benchmark galaxy sample: ~120 deg^2, $i_{mag} < 22.5$, cuts similar to CFHTLenS



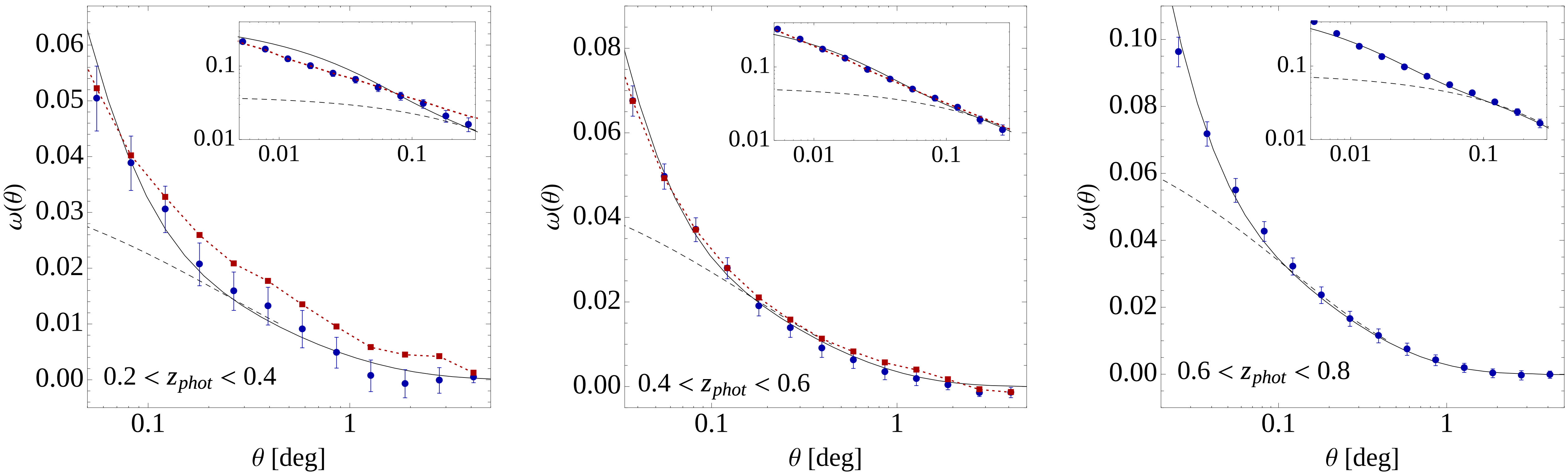


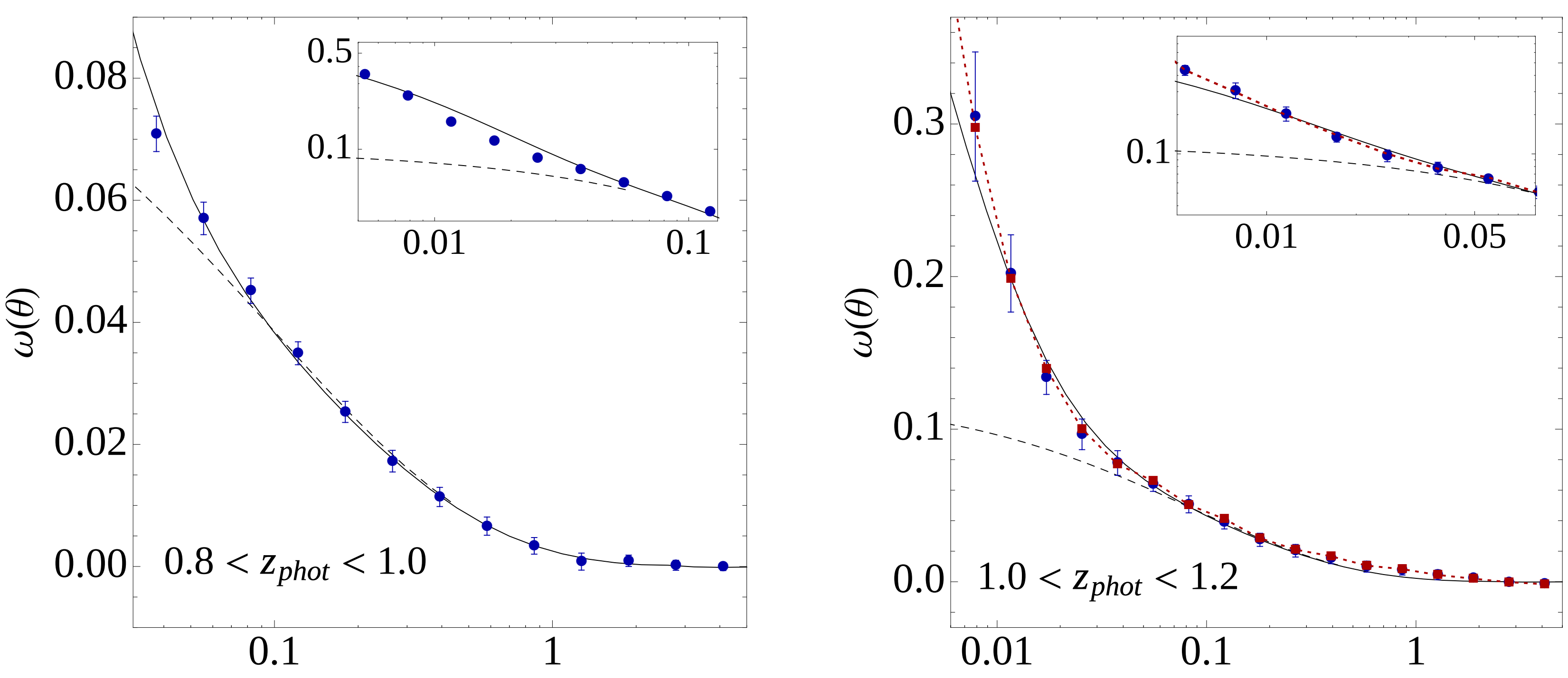


BAO with Y1: 3-4%!

DES SV angular clustering

Crocce et al (MNRAS 2015, arXiv:1507.05360)



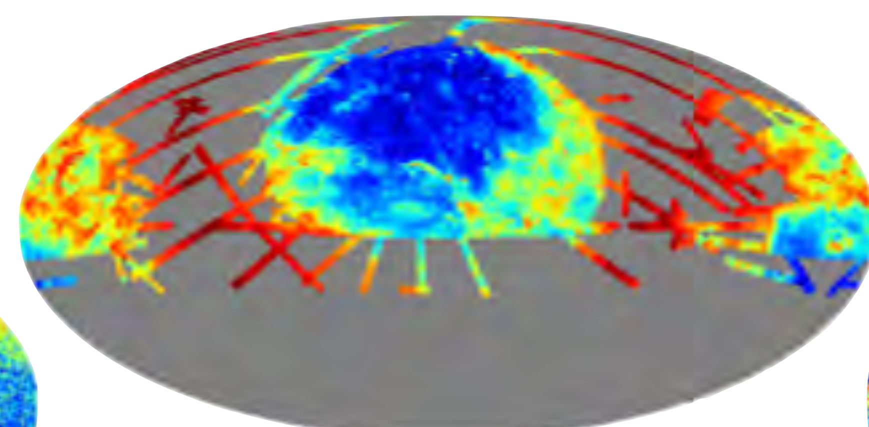


With corrections for spatial systematics

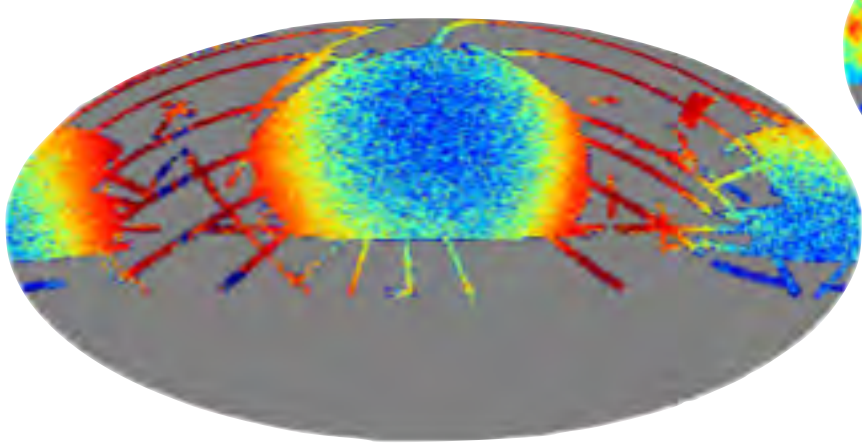
Spatial systematics?

- Anything that affects the measured galaxy properties *e.g.* dust extinction, seeing, airmass, zero points, …
- Create spatially varying depth & stellar contamination

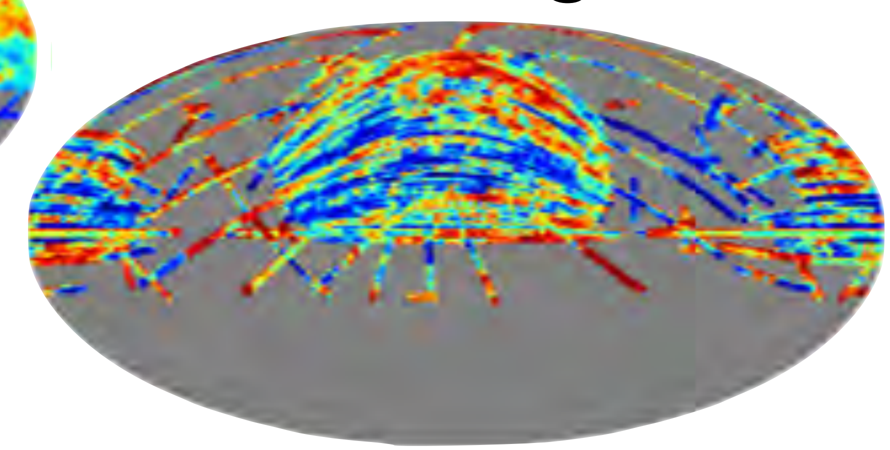
dust



stars



seeing



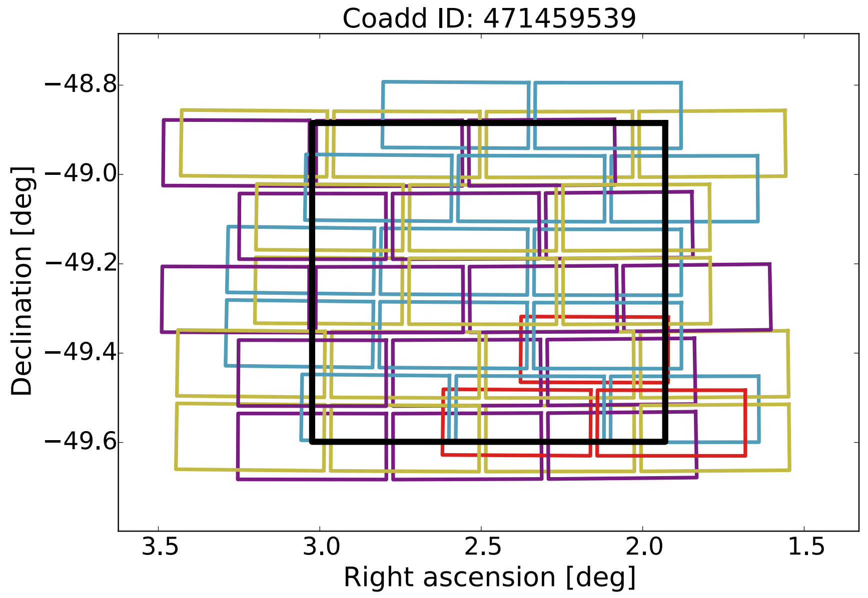
typical sources of systematics in SDSS

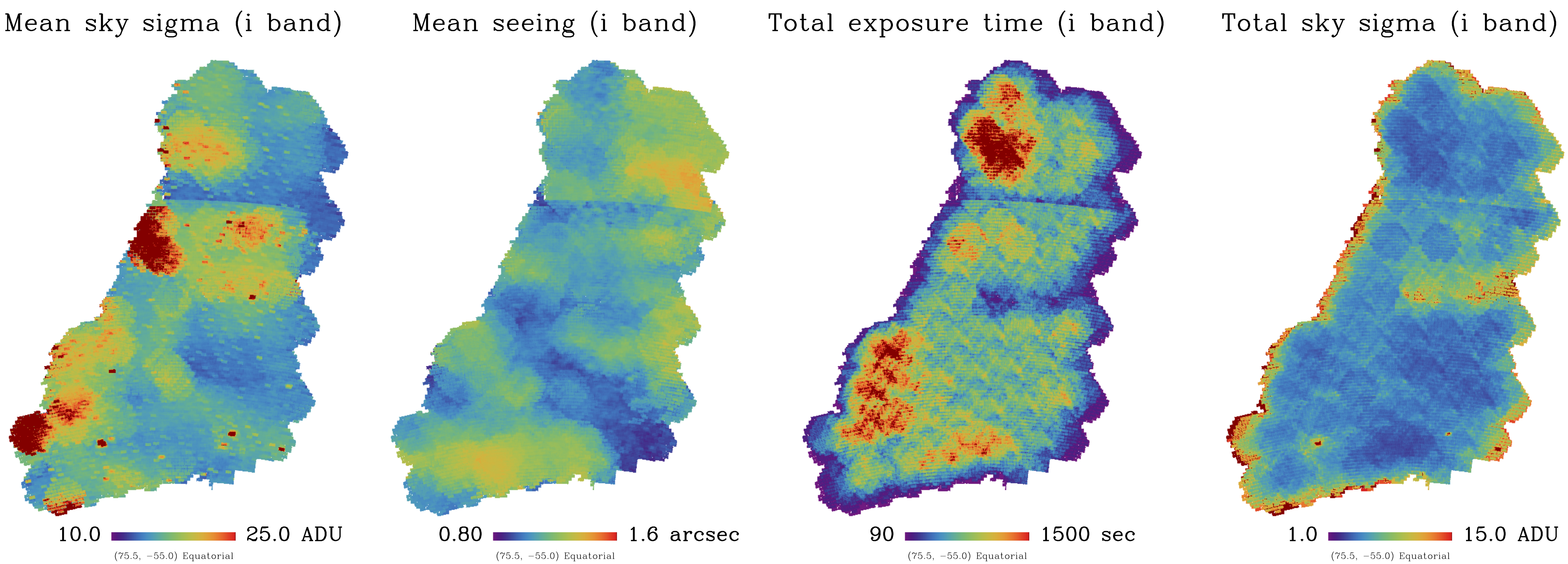
Observing conditions & systematics

Leistedt et al (ApJS 2015, arXiv:1507.05647)

Mapping & projecting image properties

Useful for null tests, systematics checks, …





Spatial systematics: state of the art

- Used to be main limiting systematic.

- **Confirmation bias** is dangerous (=fiddle with data and pipeline until results agree with expectations)
  Blinding + meaningful statistical techniques essential.

- We now routinely **map** and **simulate** spatial systematics
  Techniques to correct clustering measurements:
  Elsner, Leistedt & Peiris: arXiv:1609.03577, 1509.08933,
  1507.05647, 1404.6530

photometric redshifts

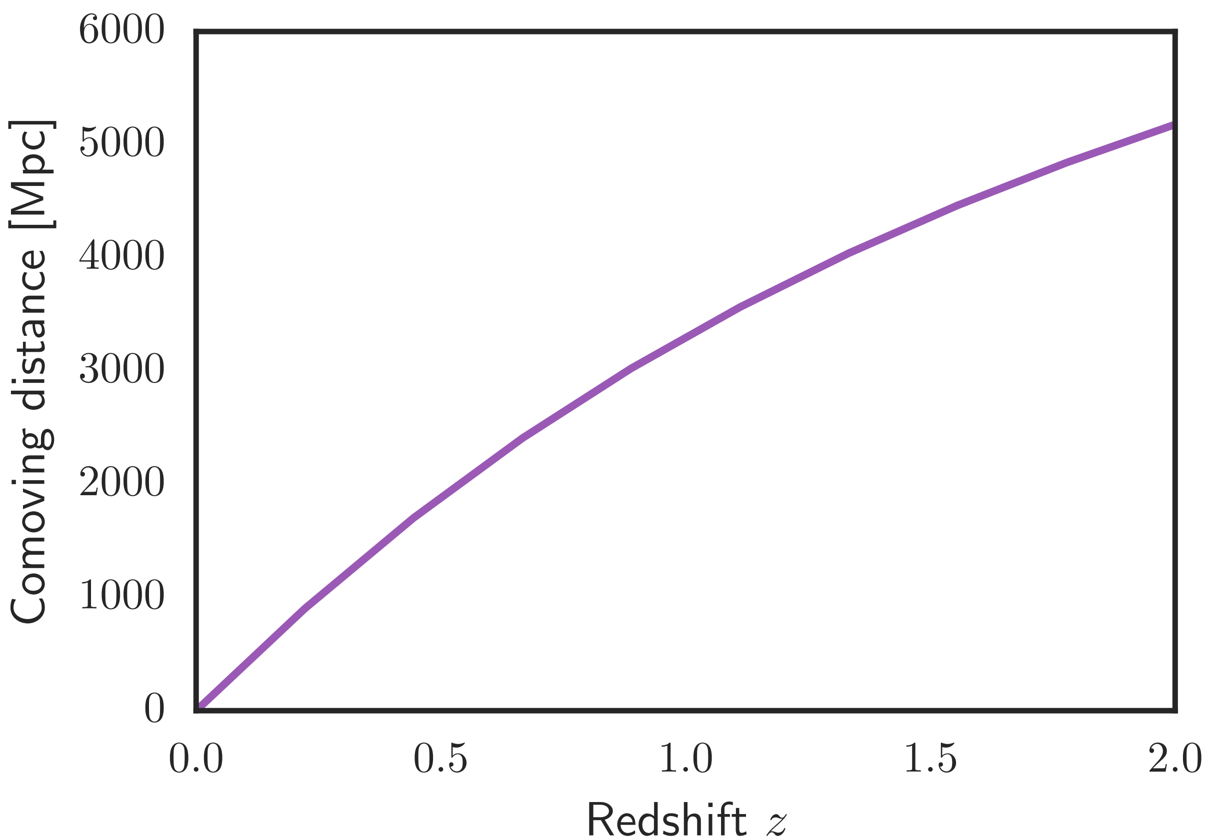
(the elefant in the room)

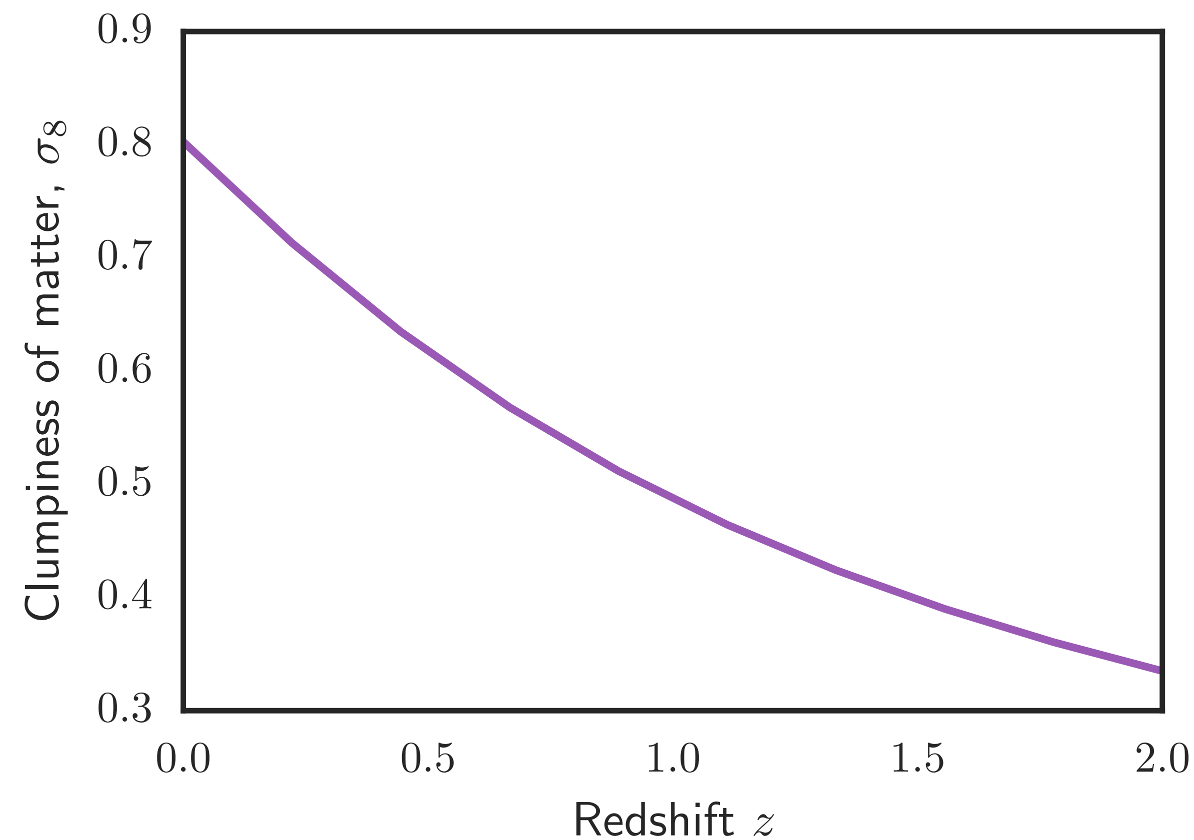
Redshift: doppler shift of electromagnetic radiation due to expansion of the universe = *indication of distance*

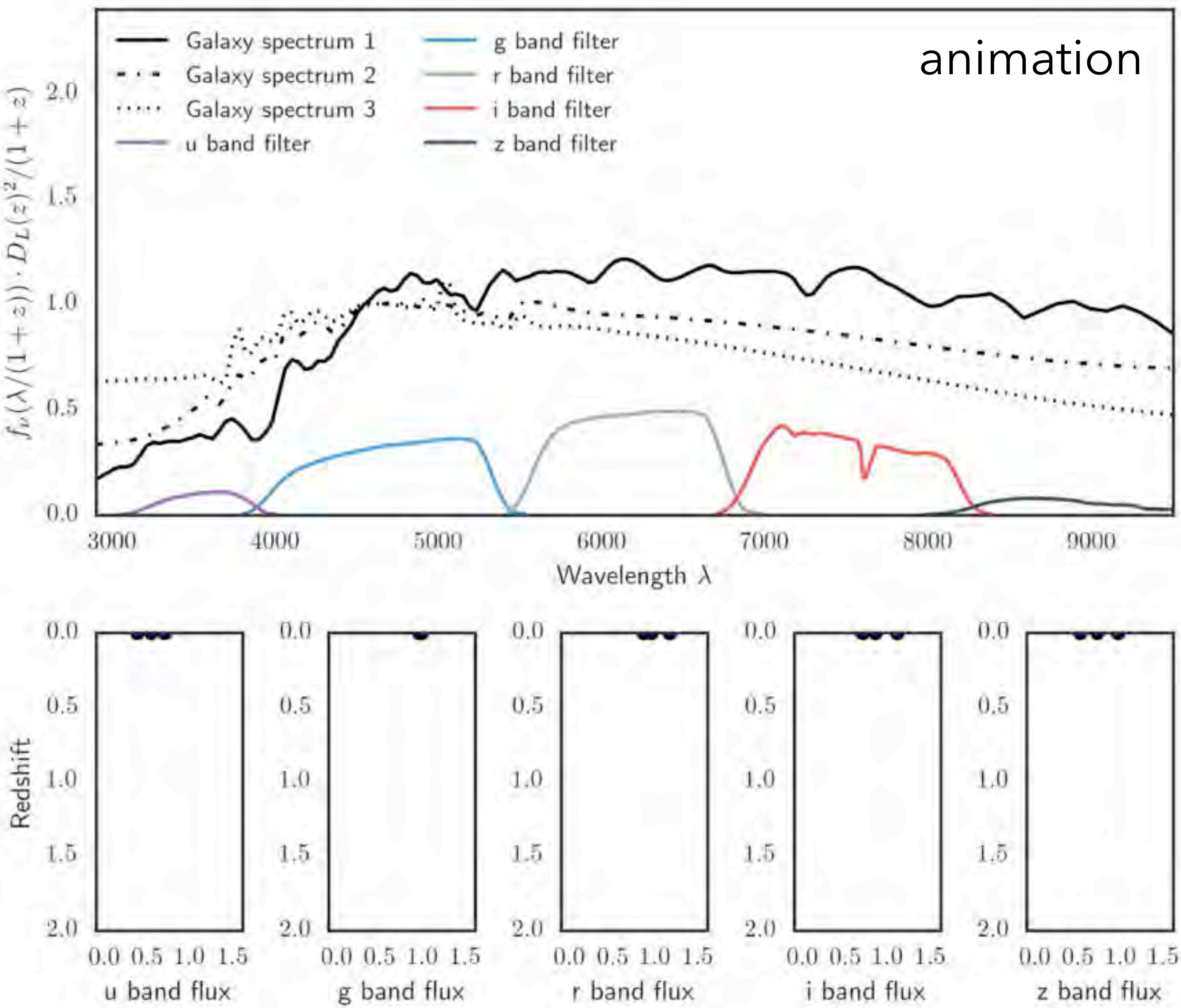
$$f_\nu(\lambda_{\rm obs}, z) = \frac{(1+z)}{4\pi D_L^2(z)} \; L_\nu\left(\frac{\lambda_{\rm obs}}{(1+z)}\right)$$

flux of redshifted object intrinsic luminosity

Critical because redshift depends on cosmology

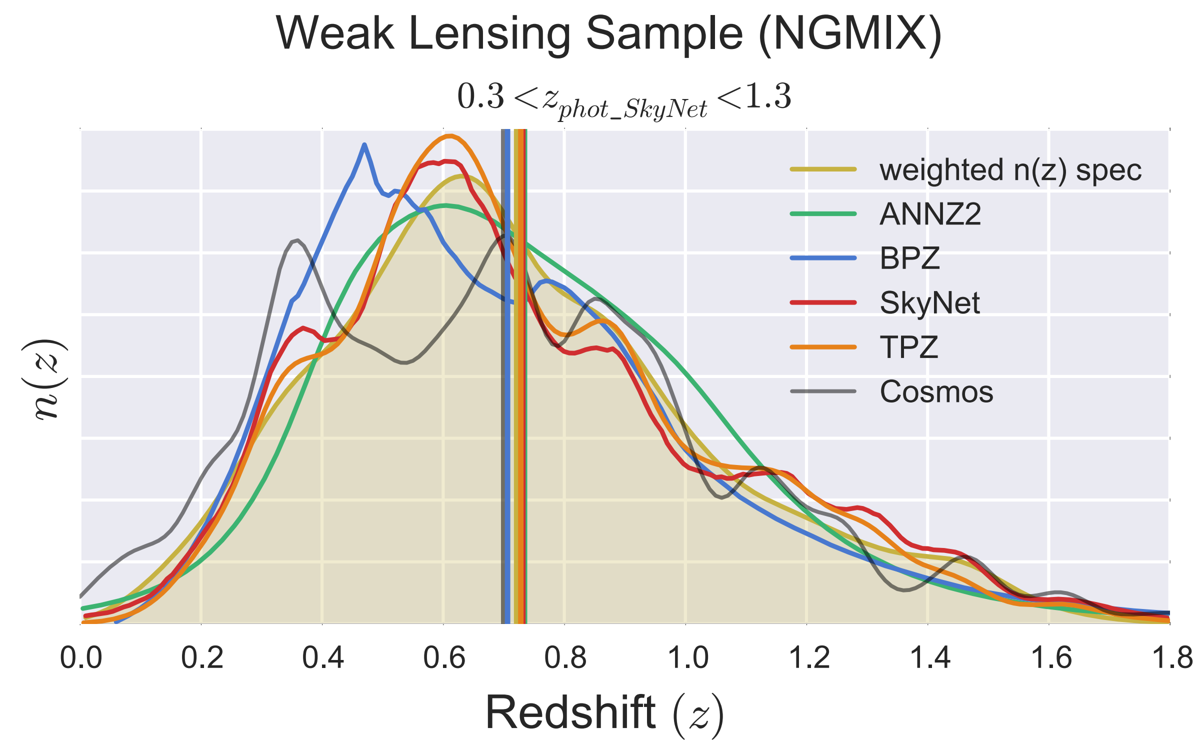


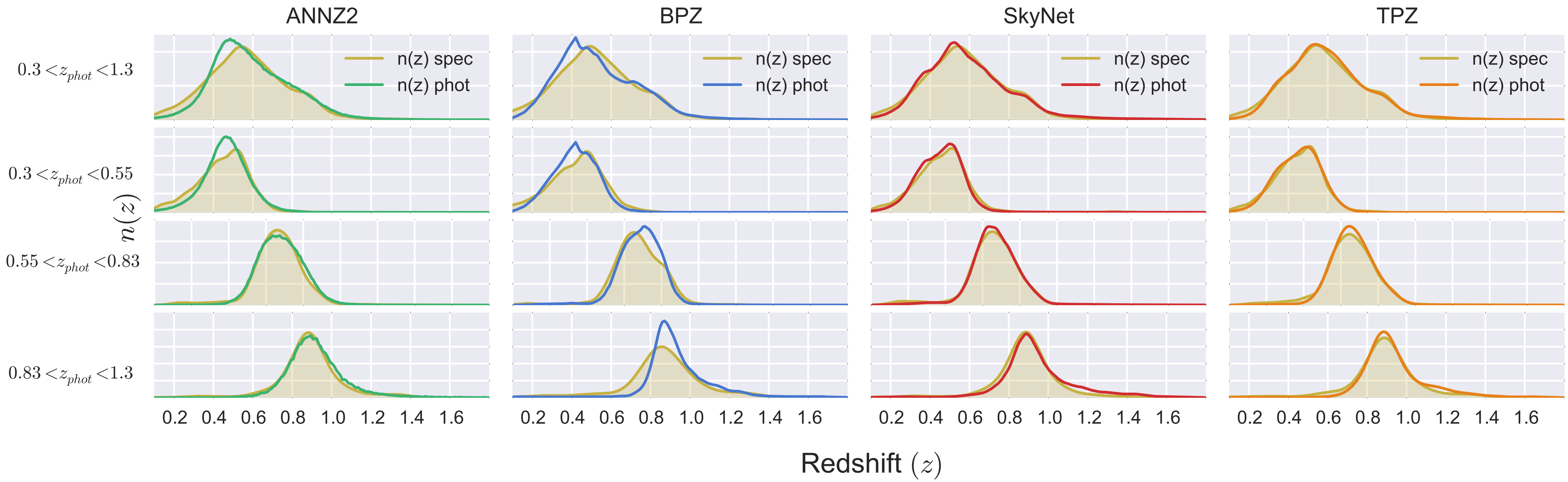




Galaxy spectrum 1
Galaxy spectrum 2
Galaxy spectrum 3
u band filter
g band filter
r band filter
i band filter
z band filter
animation
$f_\nu(\lambda/(1+z)) \cdot D_L(z)^2/(1+z)$
Wavelength λ
Redshift
u band flux
g band flux
r band flux
i band flux
z band flux

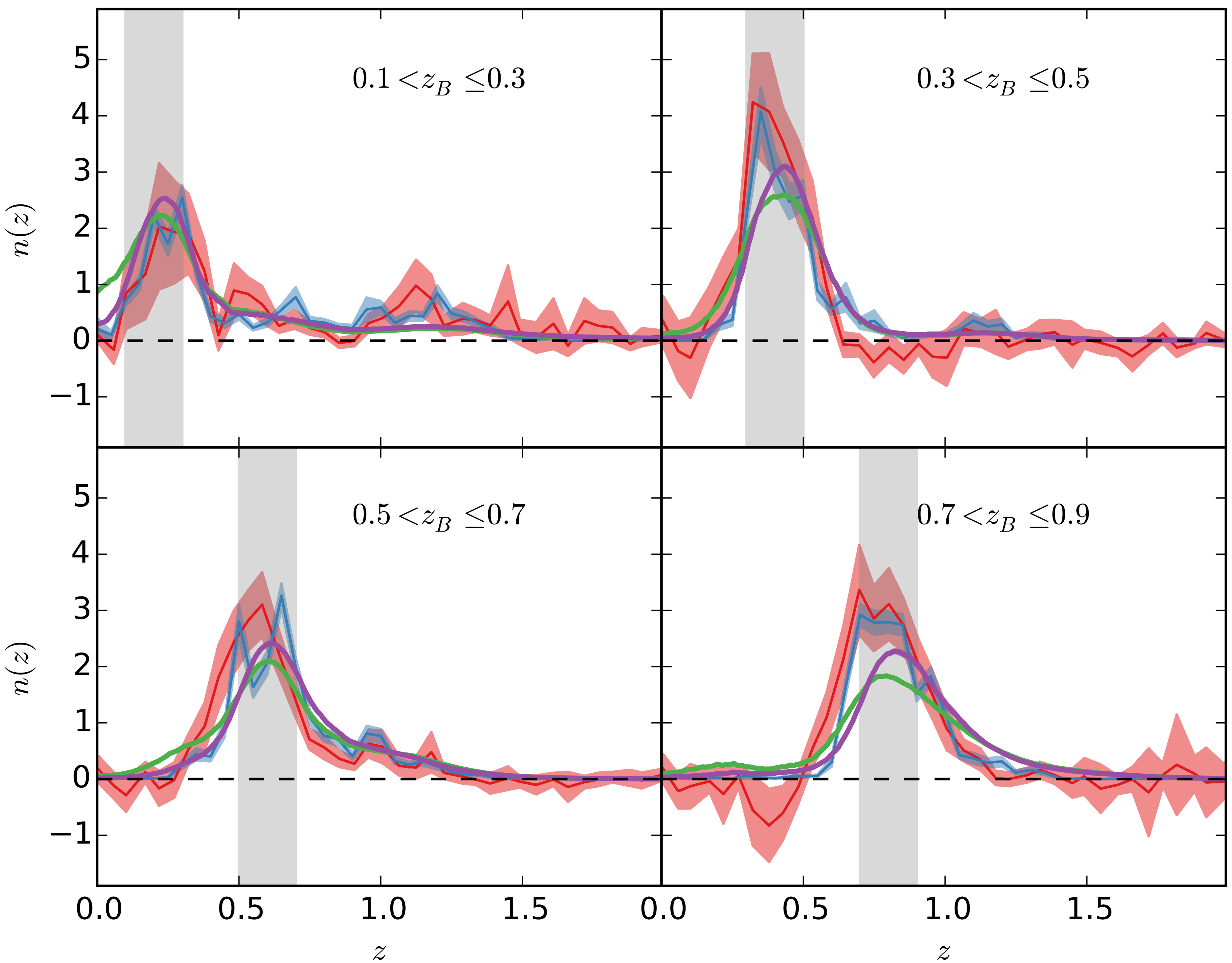
State of the art (DES)





Redshift distributions for DES SV galaxies (1507.05909)

State of the art (KIDS)



Redshift distributions for KIDS galaxies (1606.05338)

Ongoing surveys don't meet photo-z requirements

LSST requires
insanely precise photo-z's

Why is it so hard?

Bayes Theorem

$$\underbrace{p(\mathrm{P}|\mathrm{D},\mathrm{M})}_{\text{posterior}} = \underbrace{p(\mathrm{D}|\mathrm{P},\mathrm{M})}_{\text{likelihood}} \times \underbrace{p(\mathrm{P}|\mathrm{M})}_{\text{prior}} \ / \ \underbrace{p(\mathrm{D}|\mathrm{M})}_{\text{evidence}}$$

Application to redshift distributions:

$$p\Big(N(z), \{z_i\}\Big|\{\mathrm{Fluxes}_i\}\Big) \propto p\Big(N(z)\Big) \prod_{i=1}^{N} p\Big(z_i\Big|N(z)\Big)\ p\Big(\mathrm{Fluxes}_i\Big|z_i\Big)$$

full posterior prior population likelihood

Photo-z (likelihood) methods:

template fitting

vs

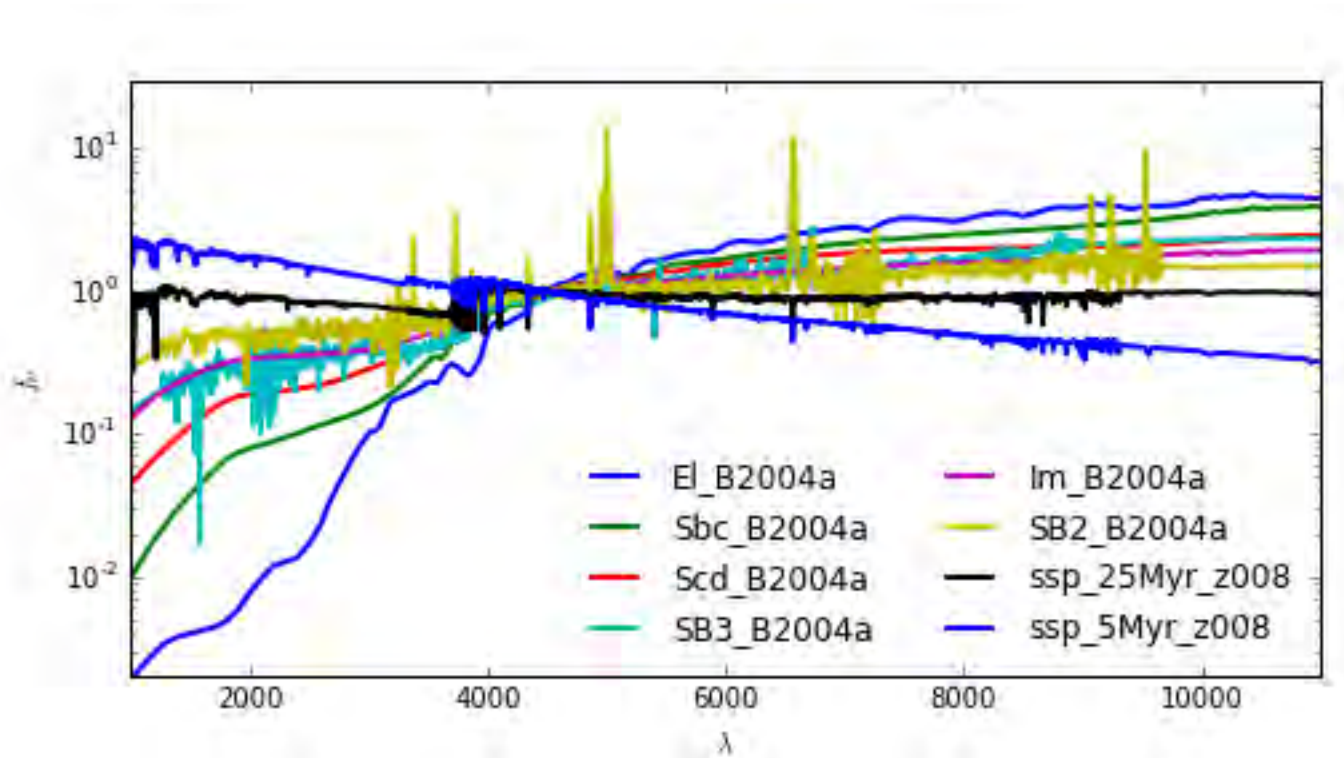
machine learning

(+new contestant: clustering redshifts)

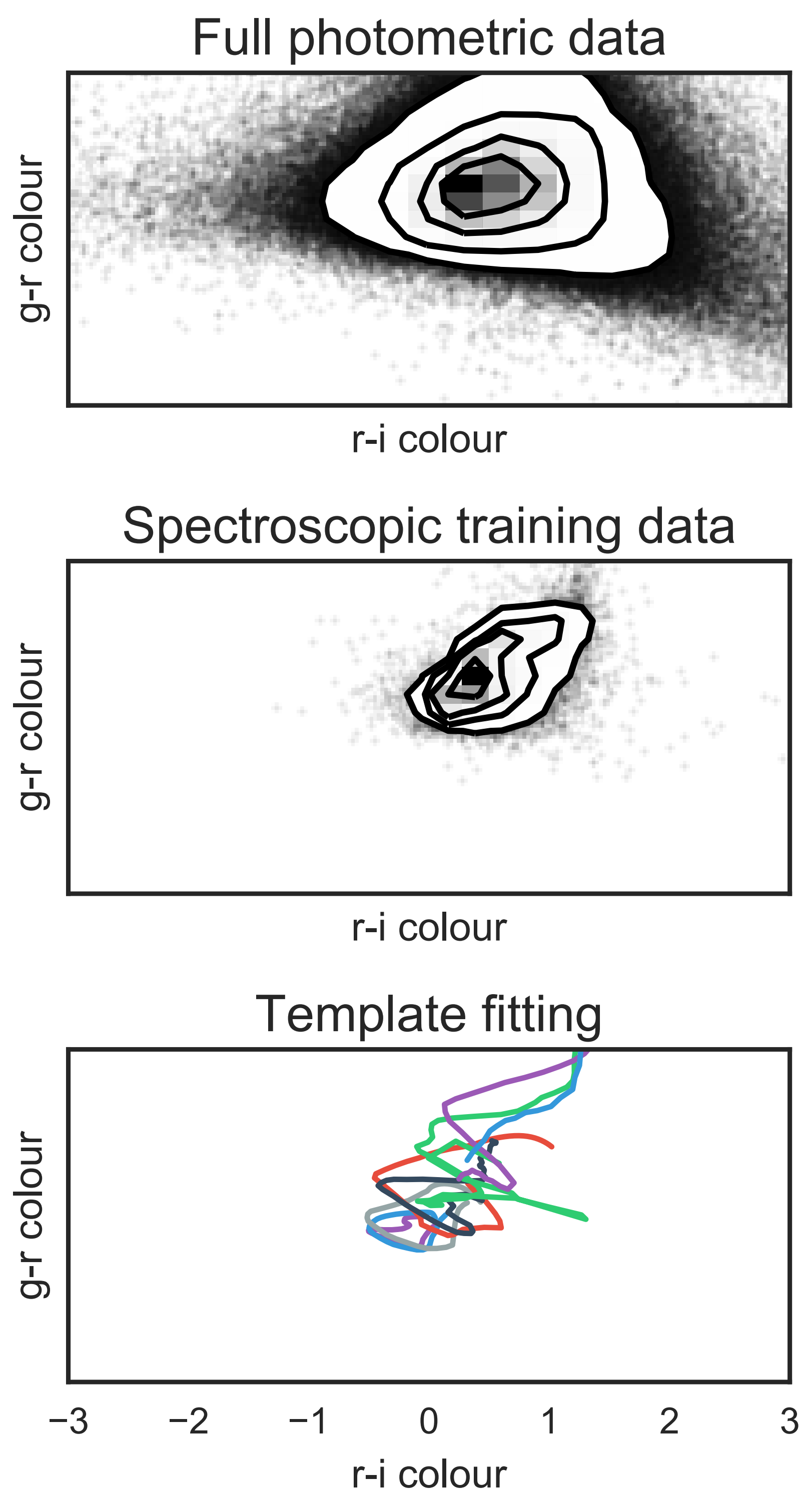
template fitting

- ✓ *physical model*
- ✓ *probabilistic*
- ✗ *need template set*
- ✗ *hard to capture data complexity*
- ✗ *sensitive to priors*

template set (CWW)



machine learning



- ✓ *captures data complexity*
- ✓ *very flexible*
- ✗ *no physical model, solves for flux=>z, cannot extrapolate*
- ✗ *not probabilistic*
- ✗ *requires representative training data*

Why is it so hard?

$$p\big(N(z), \{z_i\} \big| \{\text{Fluxes}_i\}\big) \propto p\big(N(z)\big) \prod_{i=1}^{N} p\big(z_i \big| N(z)\big)\ p\big(\text{Fluxes}_i \big| z_i\big)$$

full posterior　　prior　　population　　likelihood

- Galaxy SED models are inaccurate (high redshift, dust, star formation, variability, etc) ⇒ **likelihood is unreliable**
- Standard analyses stack redshift PDFs to obtain N(z).
  ⇒ **N(z) is biased and has no uncertainties.**
- <u>My goals</u>:
  Create & calibrate SED models and likelihood function
  Correctly infer N(z) and propagate errors into cosmology

Hierarchical inference of redshift distributions

arXiv:1602.05960

with

Daniel Mortlock

Hiranya Peiris

Hierarchical N(z) inference

`

$$p\big(N(z,t,m),\{z_i,t_i,m_i\}\big|\{\text{Fluxes}_i\}\big)$$

$$\propto\ \underbrace{p\big(N(z,t,m)\big)}_{\text{prior}}\prod_{i=1}^{N}\ \underbrace{p\big(z_i,t_i,m_i\big|N(z)\big)}_{\text{population}}\ \underbrace{p\big(\text{Fluxes}_i\big|z_i,t_i,m_i\big)}_{\text{likelihood}}$$

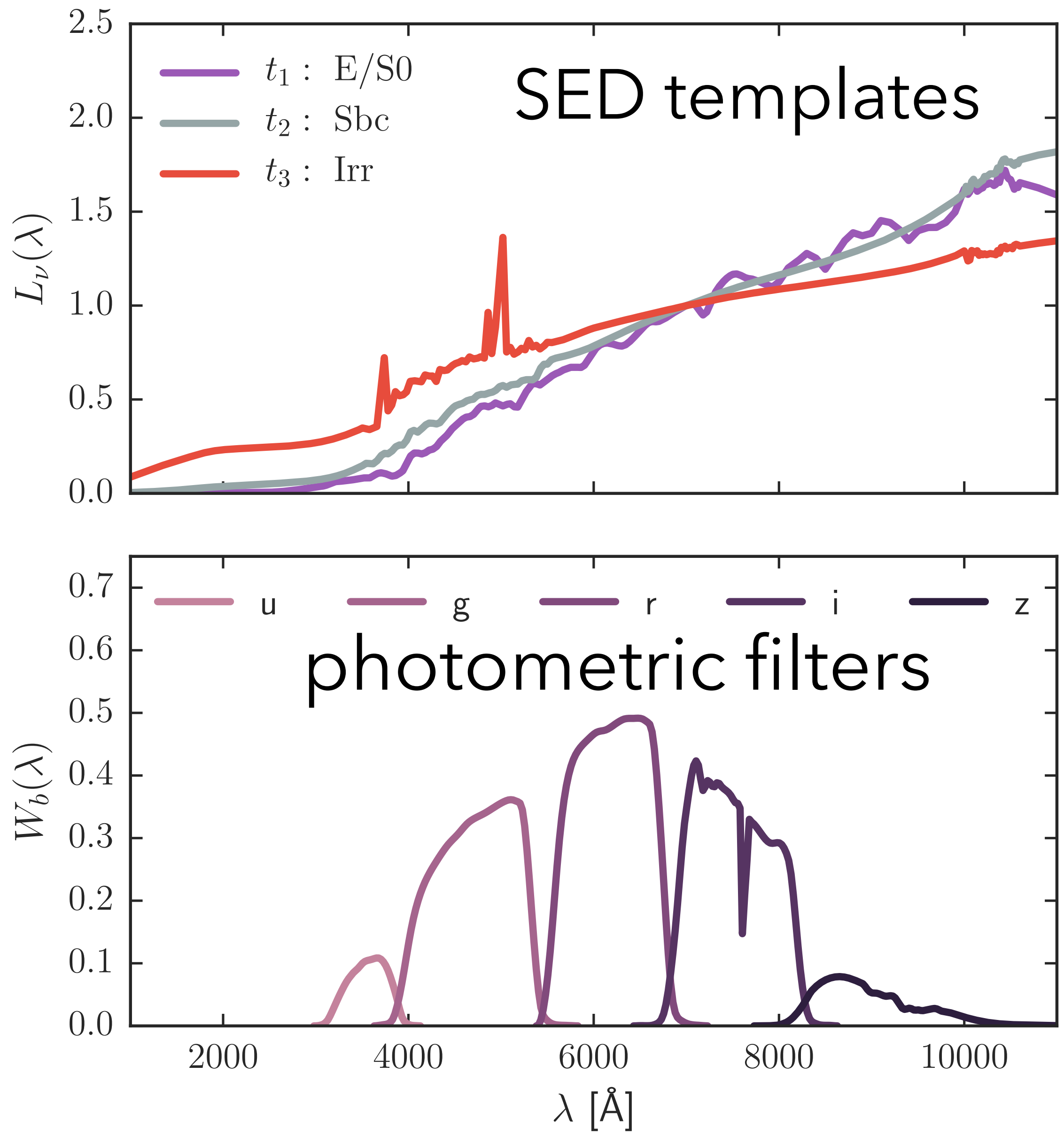
Likelihook based on SEDs, assumed to be **correct**.

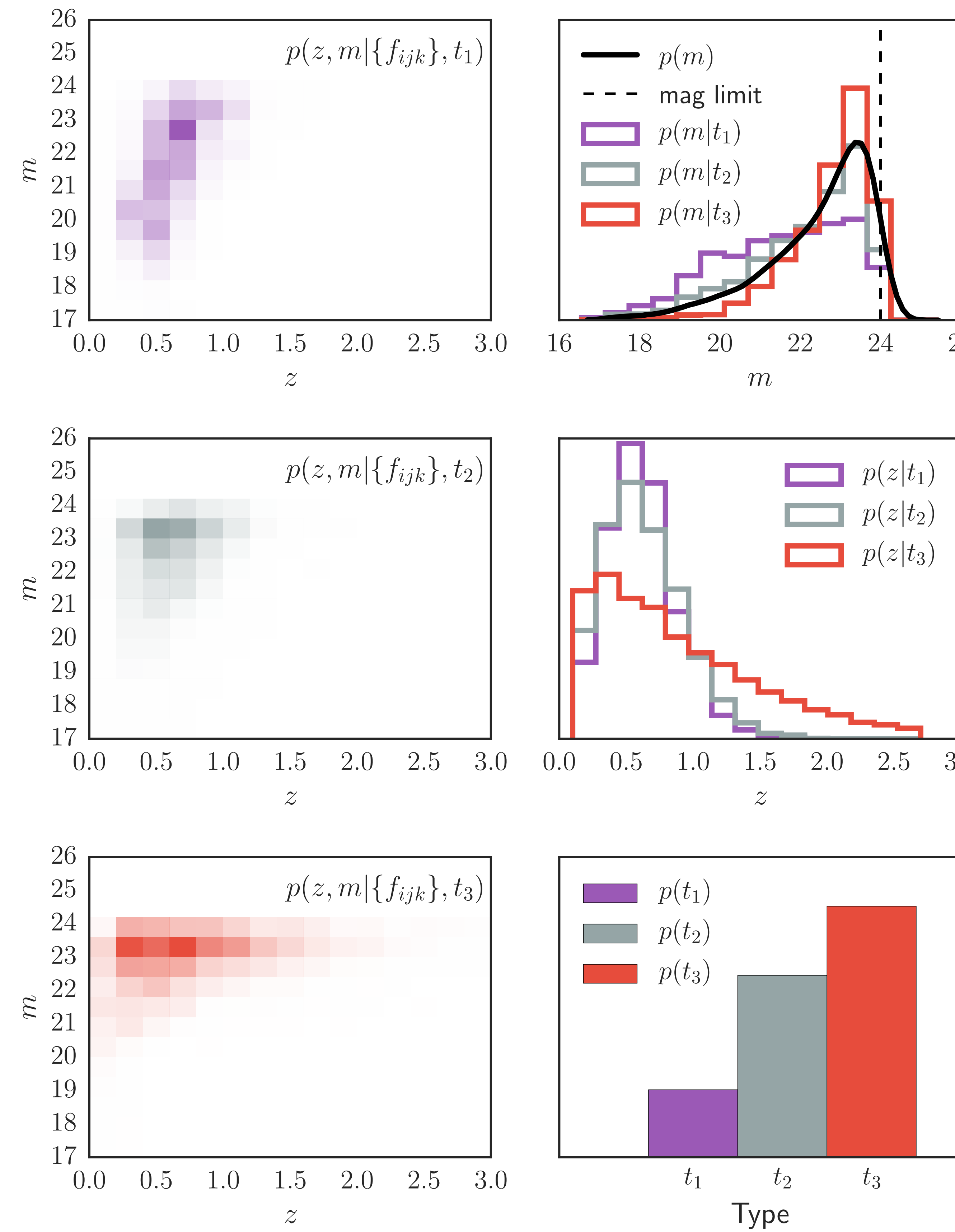
Histogram model of N*(z,t,m)* parameterized by $\{f_{ijk}\}$

Jointly infer $\{z,t,m\}_{\text{objects}}$ and $\{f_{ijk}\}$ using Gibbs sampler

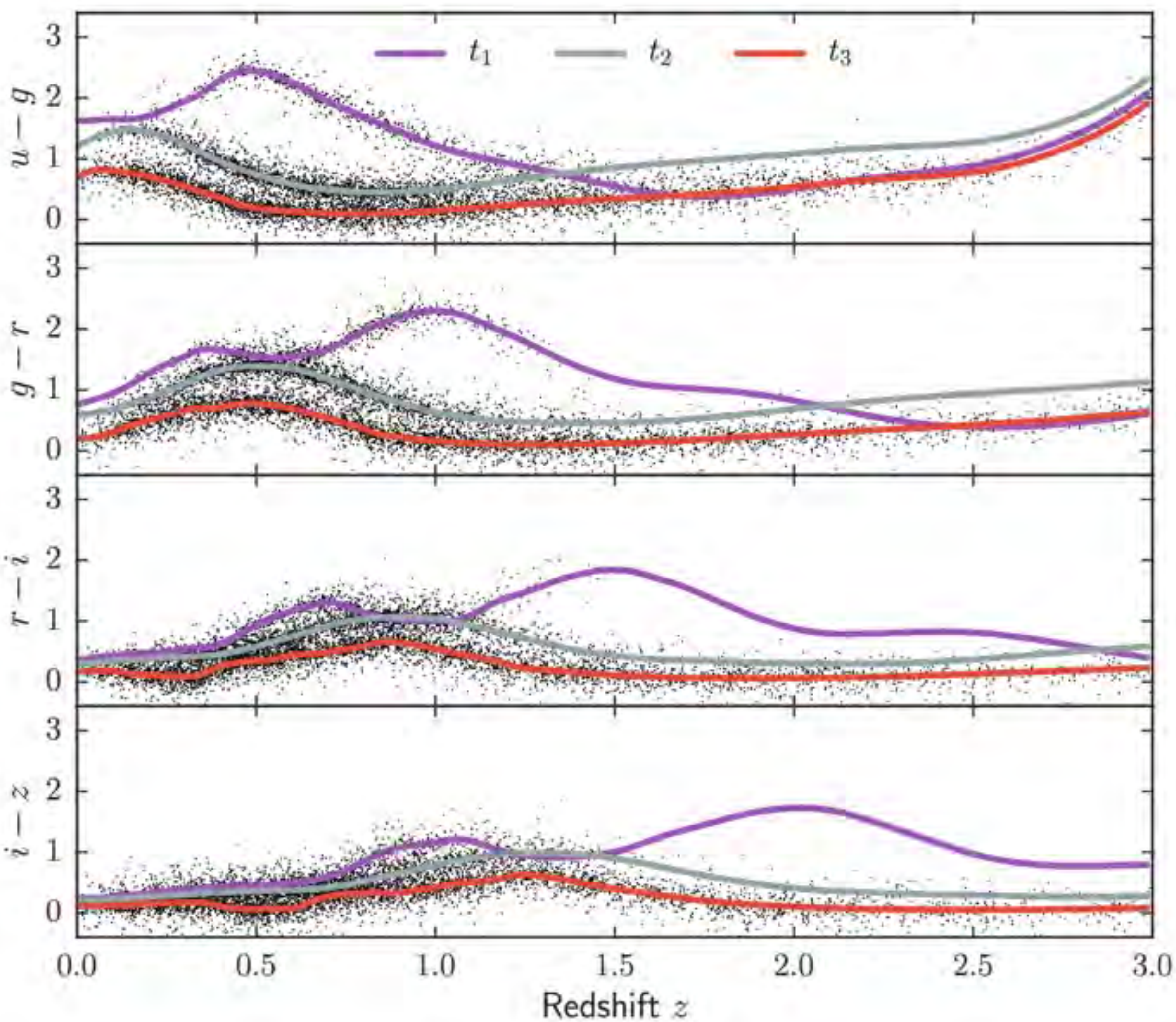
distribution N(z, t, m) of the simulation

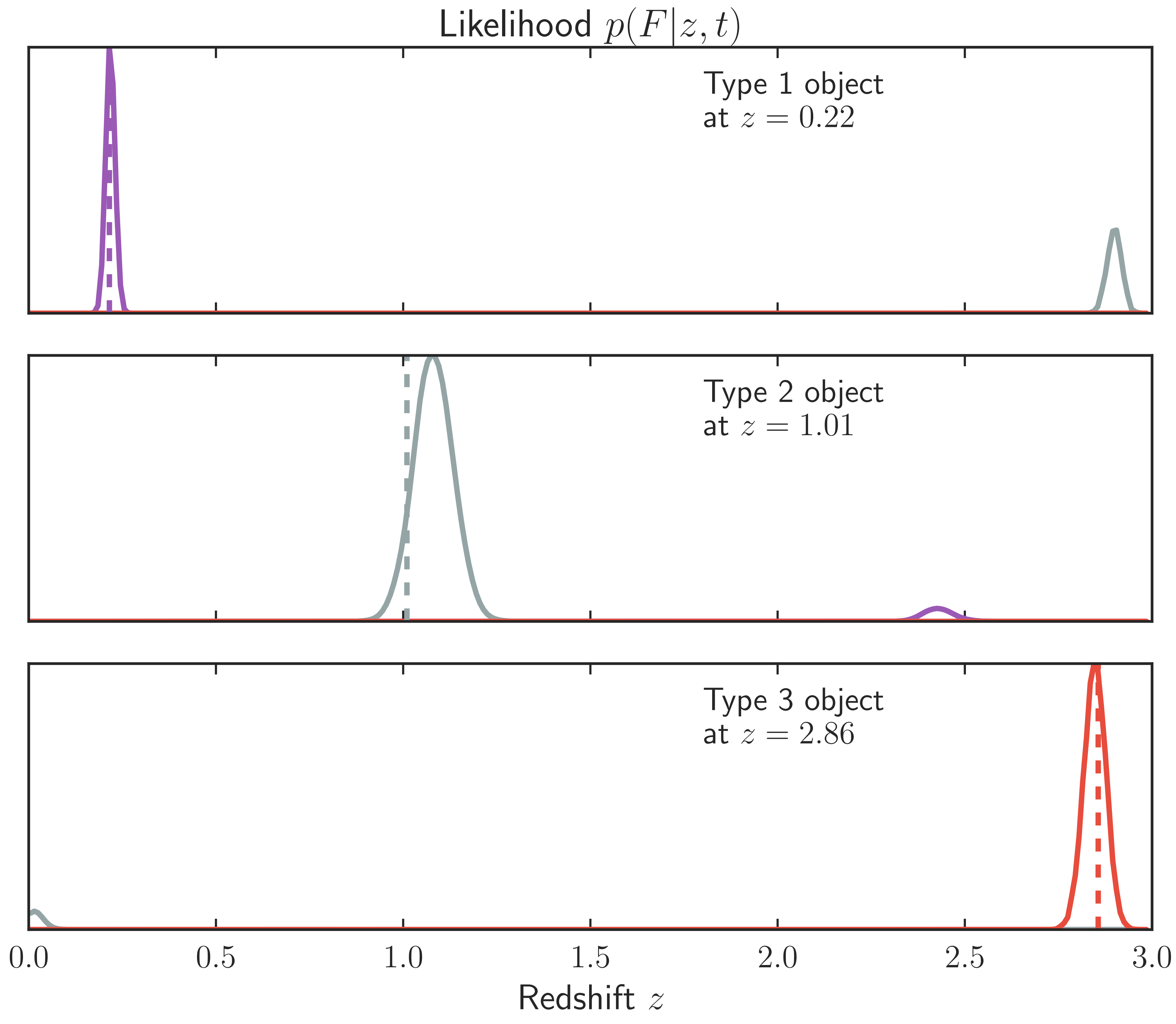
- *3 templates*
- *ugriz filters*
- *realistic distributions*





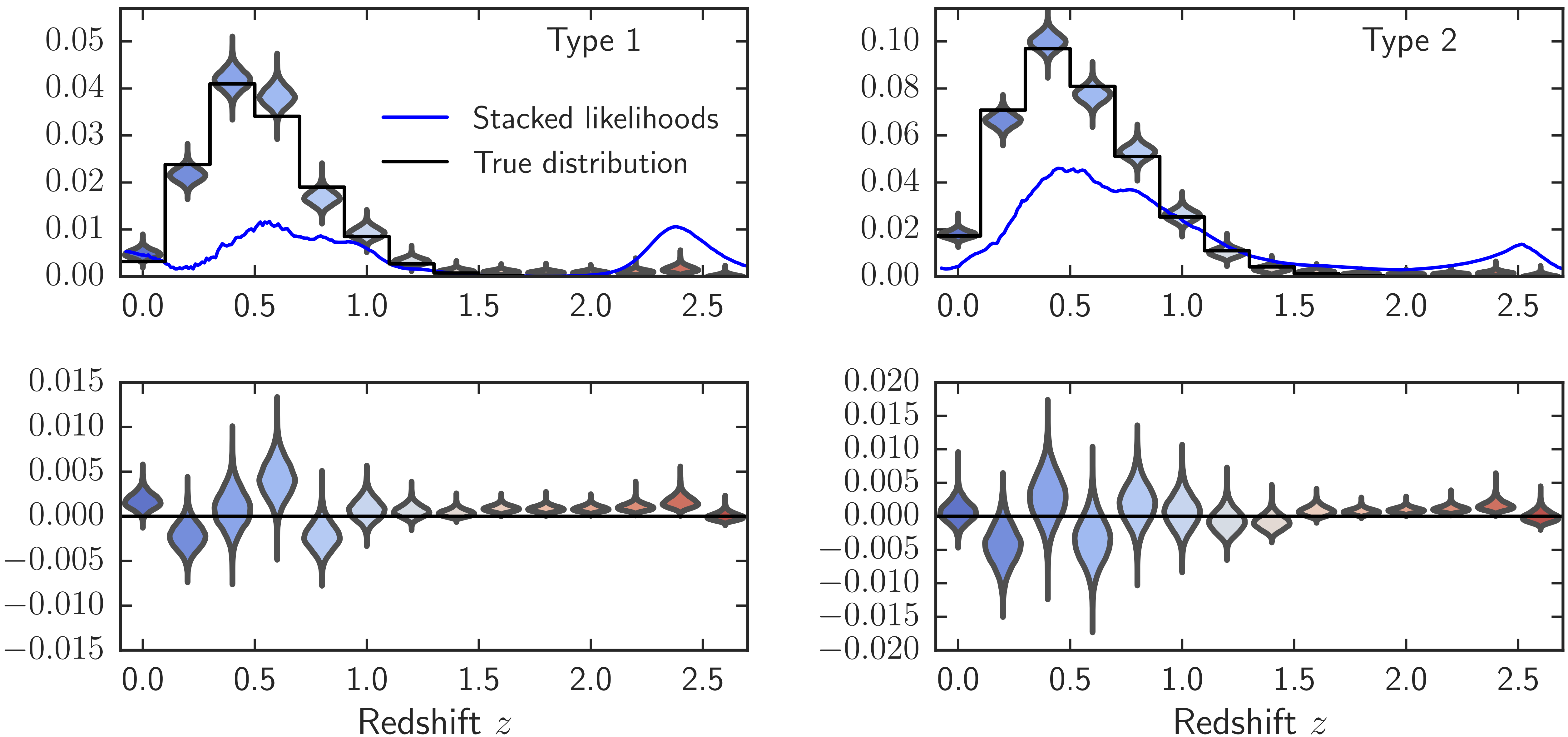
Simulated colors of 10^4 galaxies



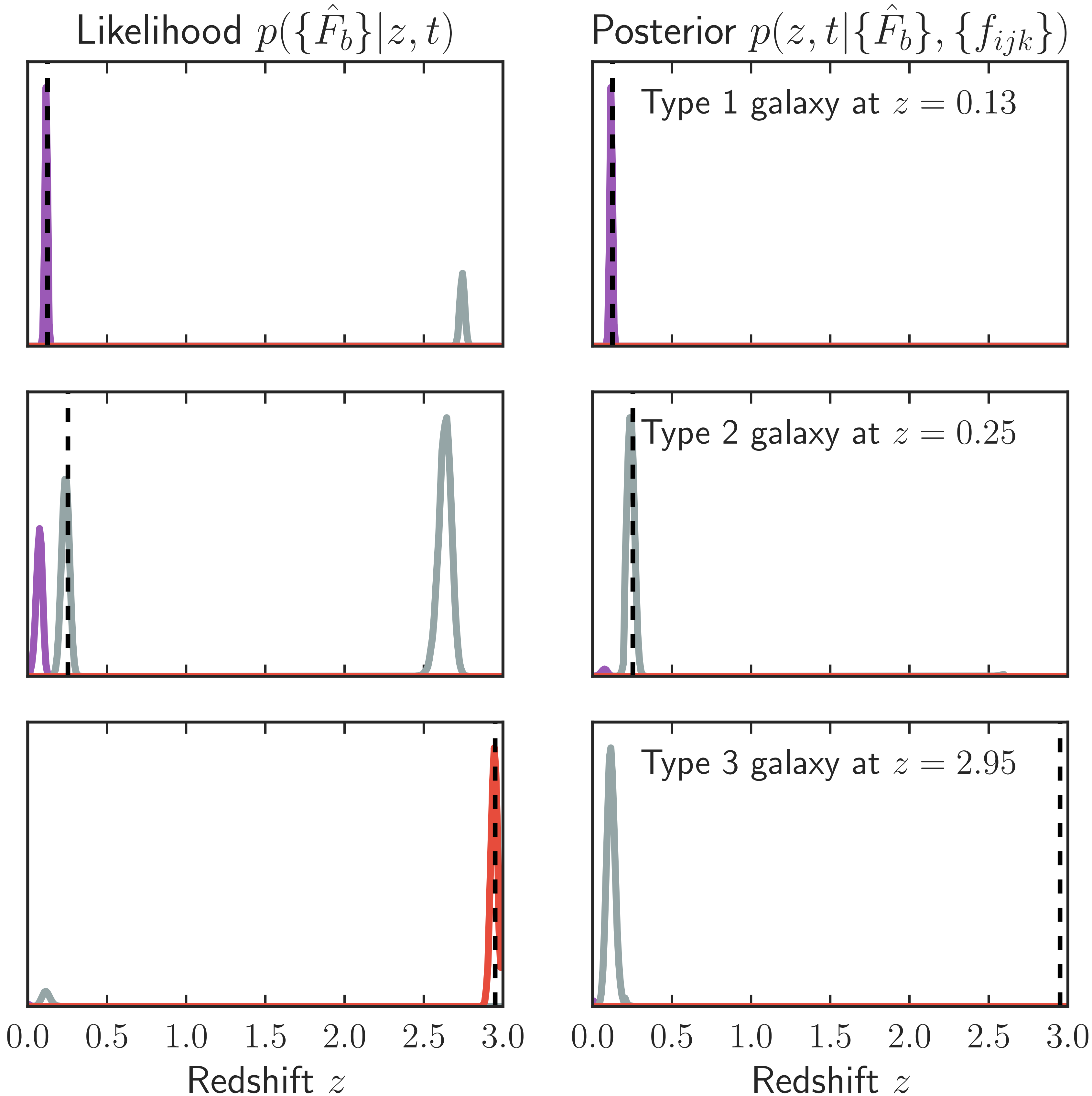


Likelihood $p(F|z,t)$
Type 1 object
at $z = 0.22$
Type 2 object
at $z = 1.01$
Type 3 object
at $z = 2.86$
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Redshift z

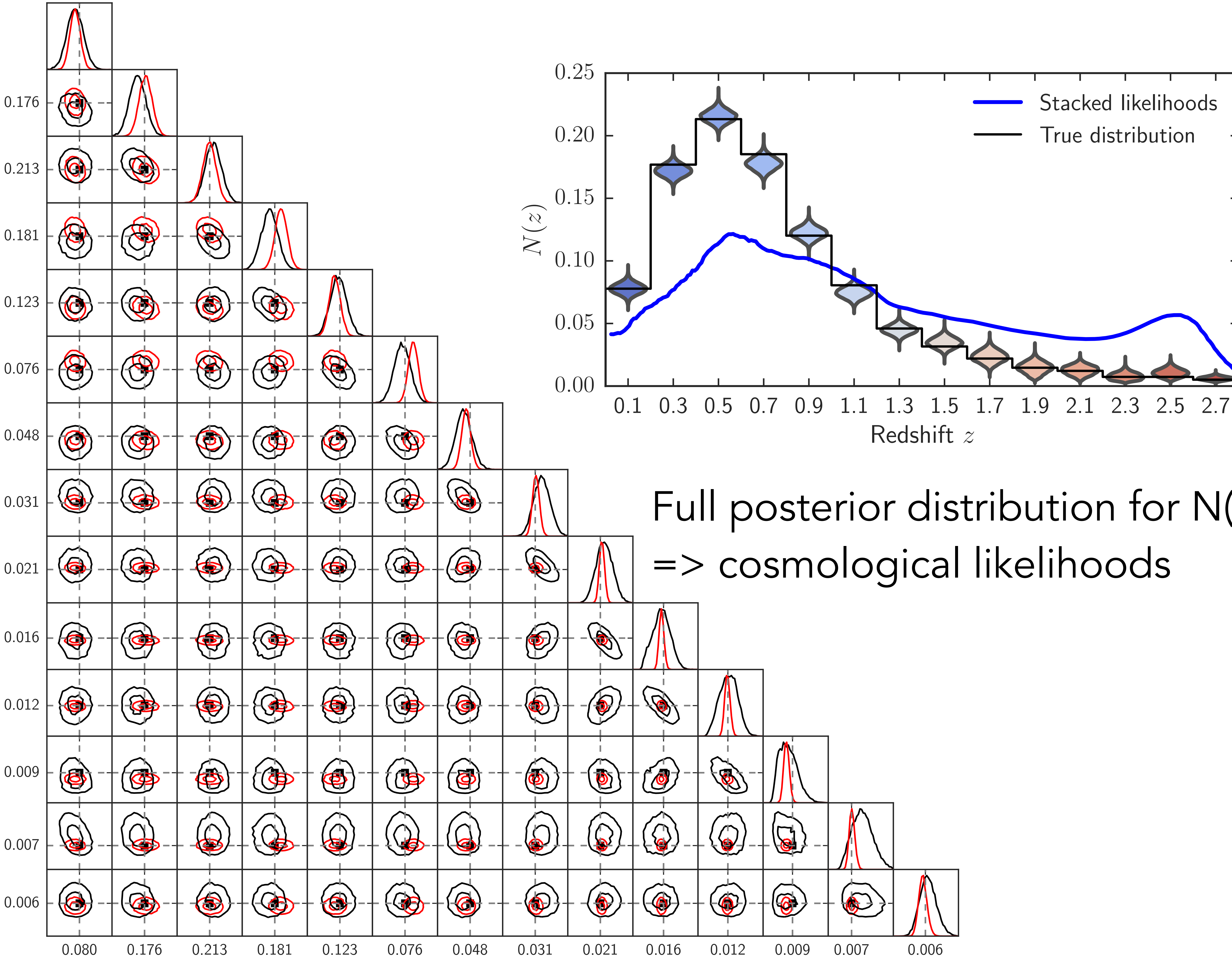
after inference: samples of the full posterior distributions



Redshift distributions correctly recovered
despite strong degeneracies



Likelihood $p(\{\hat{F}_b\}|z,t)$
Posterior $p(z,t|\{\hat{F}_b\},\{f_{ijk}\})$
Type 1 galaxy at $z = 0.13$
Type 2 galaxy at $z = 0.25$
Type 3 galaxy at $z = 2.95$
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Redshift z
Redshift z



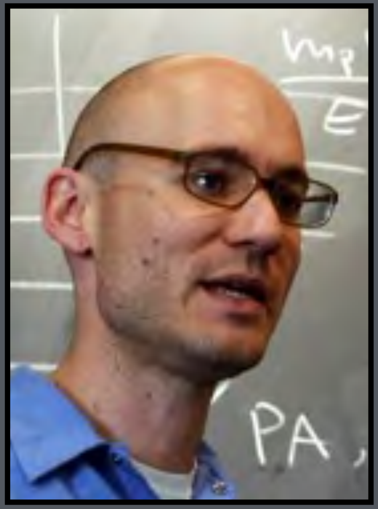
Full posterior distribution for N(z)
=> cosmological likelihoods

- [x] Hierarchical probabilistic inference of N(z)
- [x] N(z) uncertainties captured and can be propagated into cosmological likelihoods

However, real likelihoods are incorrect/biased!

Data-driven, interpretable photometric redshifts trained on heterogeneous and unrepresentative data

with David Hogg (NYU)



Will <u>**never**</u> have representative spectroscopic data

Galaxy SED models are not precise enough

Only deep spectroscopic & many-band surveys available

True PDFs needed with data <u>**and**</u> model uncertainties

Machine learning constrained by physics of the problem?

Idea

Target set: photometric survey
Training set: many-band or spectroscopic set
= deeper, heterogeneous version of target

No complete physical model for galaxy spectra
=> construct spectra compatible with training set

$$p\big(z\big|\text{Fluxes}\big) = \sum_j \; p\big(\text{Fluxes}\big|\text{Fluxes}(t_j, z)\big) \; p(z|t_j) \; p(t_j)$$

Classical template fitting *(e.g., BPZ, EAZY, ZEBRA, etc)*

- Use a small set of fixed templates based on low-redshift bright spectra or physical models

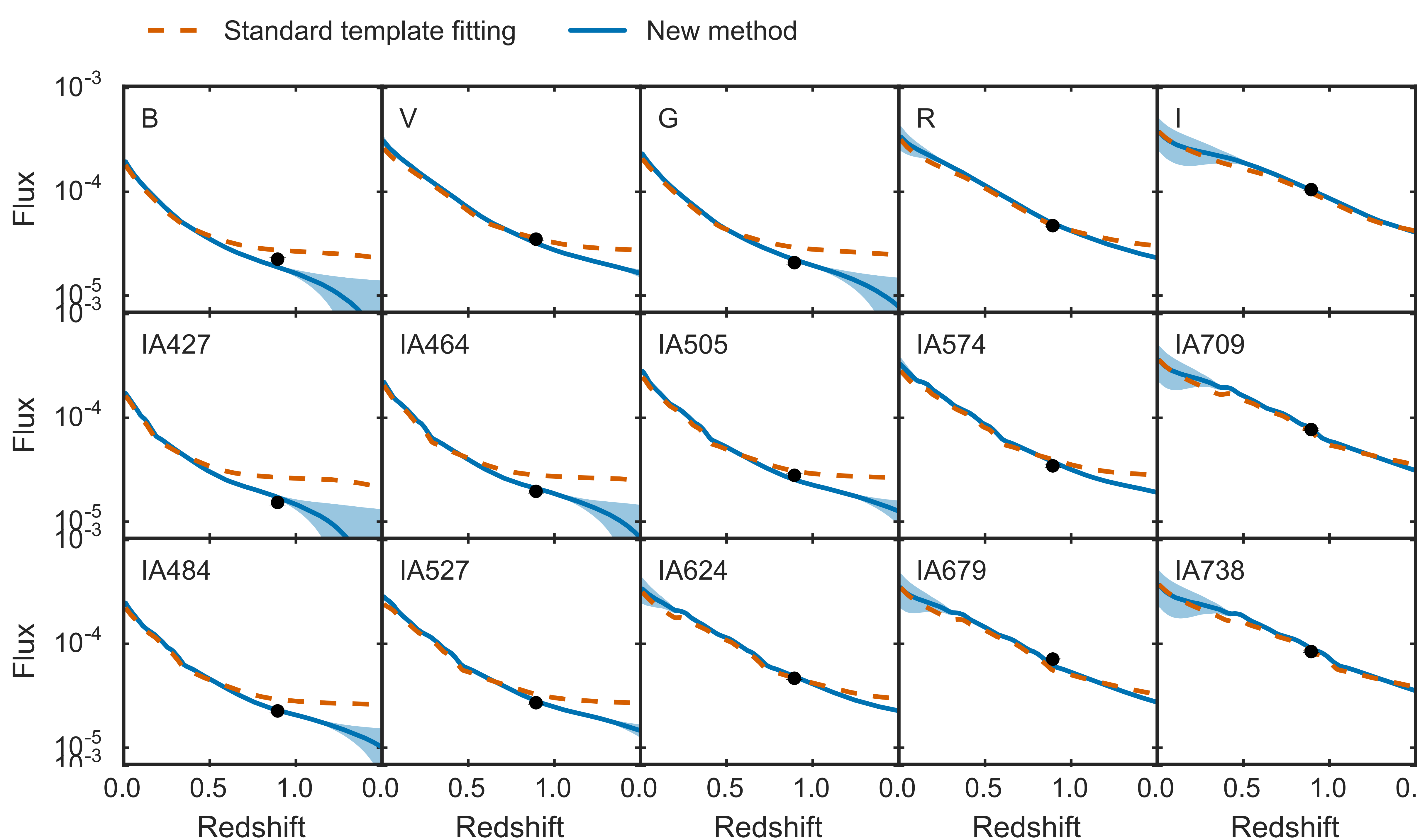
New data driven approach 1 *(work in progress…)*

- Forward model a probabilistic system of templates and priors, to be constrained from the training data.

New data driven approach 2 *(Leistedt & Hogg, arXiv:1612.00847)*

- Construct one probabilistic template per training galaxy. Pairwize comparison of target galaxies (redshifts unknown) with training galaxies (redshift known or constrained)

Example of a model per training galaxy



New method: DELIGHT^{TM}

Leistedt & Hogg (arXiv:1612.00847) – github.com/ixkael/Delight

Concept: implicitly fitting and redshifting SEDs to each training galaxy for pairwise comparison with target galaxies = *machine learning* + *template fitting*

Probabilistic, physical, and data driven
Interpretable model & PDFs. Flexibility via parameters.

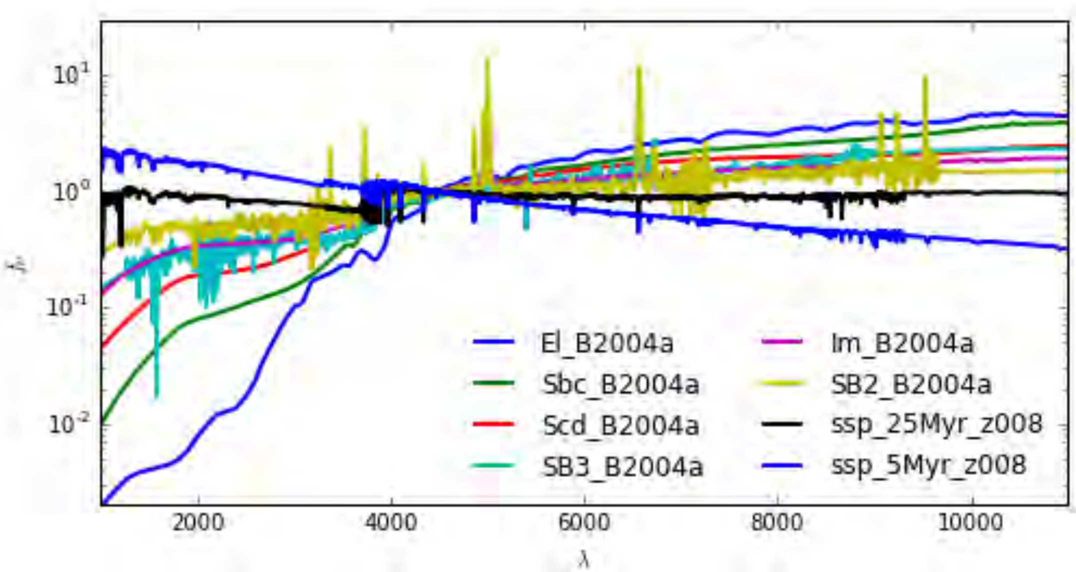
Use much more data than existing methods: heterogeneous combination of spectroscopic or deeper photometric data

Fast to (re-)train/apply. No need to store tabulated PDFs.

SED model

How to quickly construct SED model and make predictions?

$$p(\underbrace{\{\hat{F}'_b\}|z'}_{\text{target}}, t_i) = p(\{\hat{F}_b\}|z', \underbrace{z, \{\hat{F}_b\}}_{\text{training}})$$

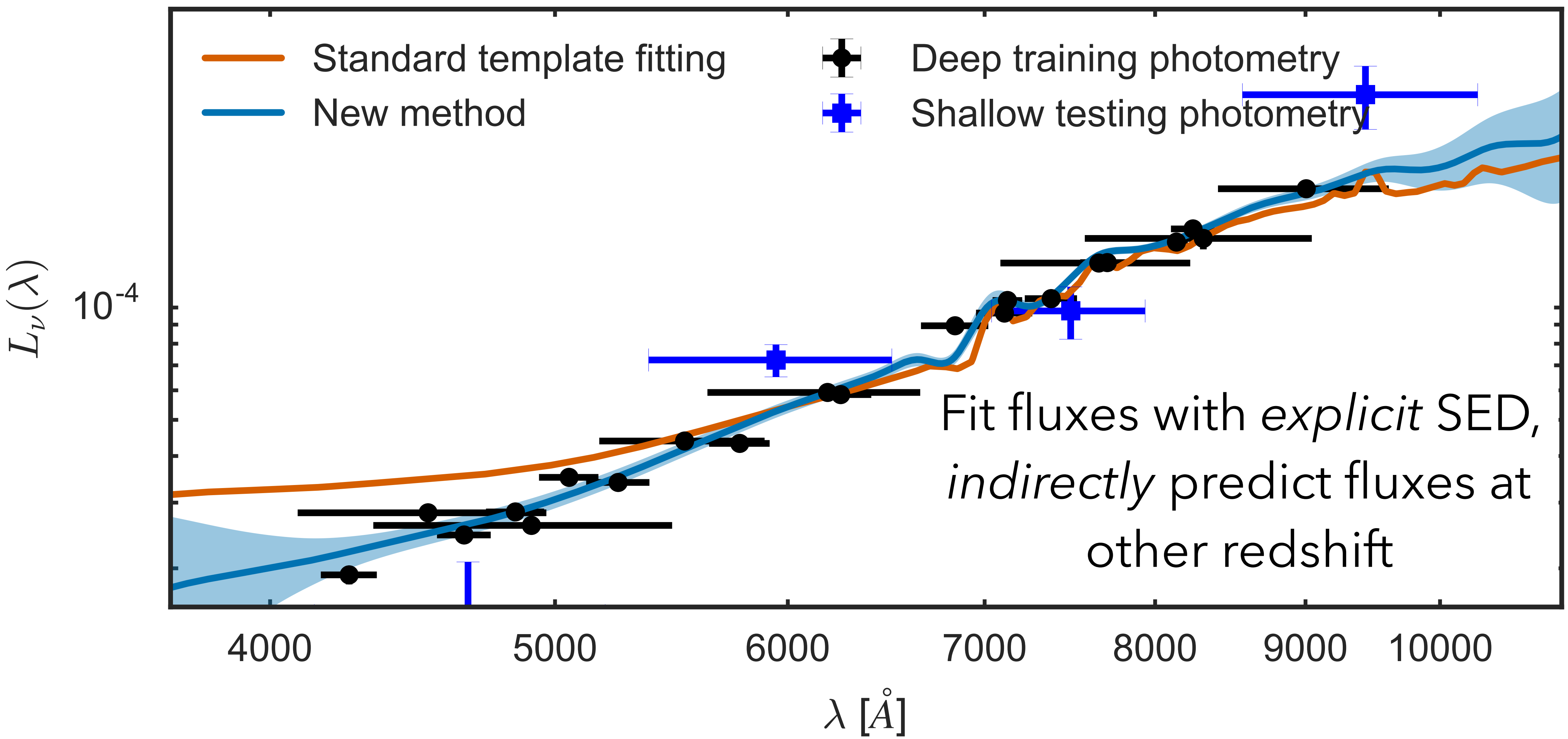


SED model:

$$L_\nu(\lambda) = \underbrace{\sum_k \alpha_k T_\nu^k(\lambda)}_{\text{templates}} + \underbrace{R_\nu(\lambda)}_{\text{residuals}}$$

The ~~crazy~~ intractable way

Explore all SEDs compatible with training galaxy (noisy fluxes + spec-z) via MCMC

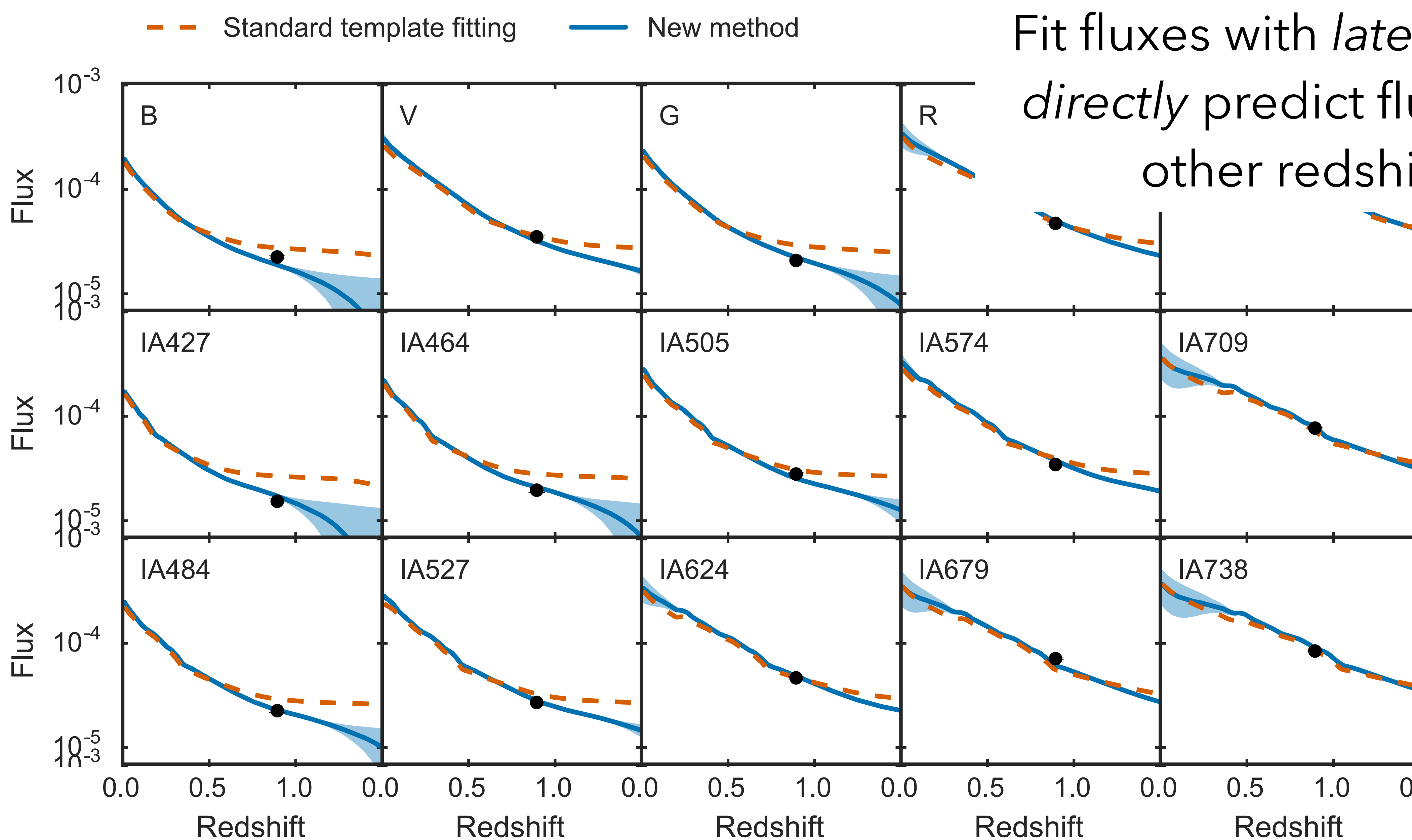


The ~~elegant~~ efficient way

Directly fit for training galaxy in flux-redshift space

+ force the fit to correspond to underlying SEDs

Fit fluxes with *latent* SED, *directly* predict fluxes at other redshift



<introduction to Gaussian Processes>

Gaussian processes

$$f \sim \mathcal{GP} \iff p(f(\vec{x}), f(\vec{x}')) \text{ is Gaussian } \forall \vec{x}, \vec{x}'$$

characterized by mean and kernel

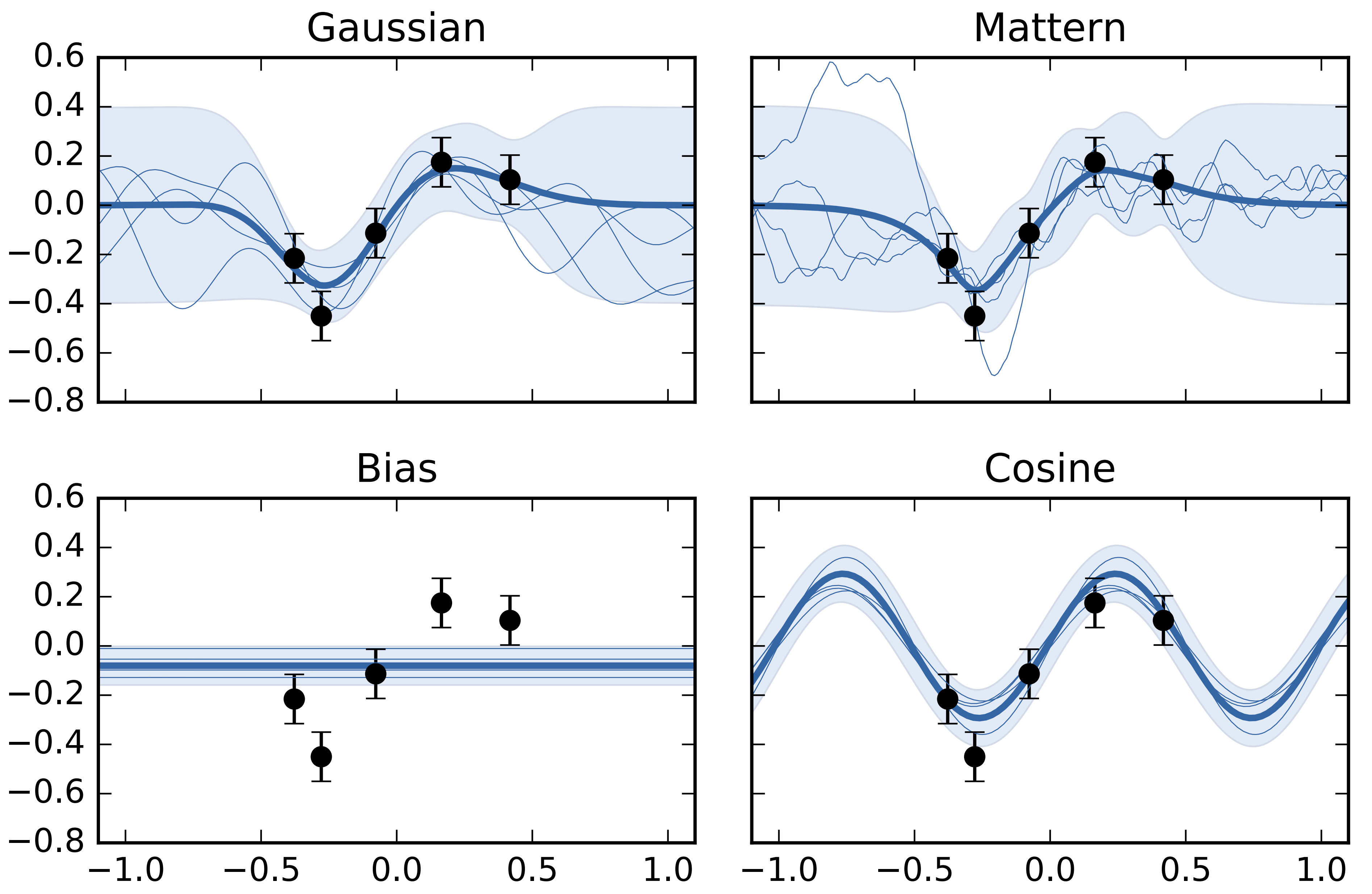
$$\begin{aligned} m(\vec{x}) &= \mathbb{E}[f(\vec{x})] \\ k(\vec{x}, \vec{x}') &= \mathbb{E}[(f(\vec{x}) - m(\vec{x}))(f(\vec{x}') - m(\vec{x}'))] \end{aligned}$$

for Gaussian likelihood, posterior/predictions **tractable**
see Rasmussen & Williams (2006)

Fitting with GPs = using priors over functions

Modelling correlated signal and/or noise

Choice of kernel is key (captures correlations)



</introduction

to Gaussian Processes>

Photo-z gaussian process

if SED model is: $L_\nu(\lambda) \sim \mathcal{GP}\Big(\underbrace{\sum_k \alpha_k T_\nu^k(\lambda)}_{\text{templates}},\ \underbrace{k(\lambda, \lambda')}_{\text{residuals}}\Big)$

then the fluxes: $F(b, z) \sim \mathcal{GP}\Big(\underbrace{\mu^F(b, z),\ k^F(b, b', z, z')}_{\text{mean flux and covariance}}\Big)$

GP with physical mean function **and residuals**

Fitting and predicting photometric fluxes while capturing the physics of redshifts

Analytically tractable under simple assumptions

Photo-z gaussian process (proof)

Redshifted galaxy SED

$$f_\nu(\lambda_{\rm obs}, z) = \frac{(1+z)}{4\pi D_L^2(z)}\ L_\nu\left(\frac{\lambda_{\rm obs}}{(1+z)}\right)$$

Photometric fluxes

$$F(b,z) = \frac{\int_0^\infty f_\nu(\lambda, z)\ W_b(\lambda)\ {\rm d}\lambda/\lambda}{\int_0^\infty g^{\rm AB}\ W_b(\lambda)\ {\rm d}\lambda/\lambda}$$

SED model

$$L_\nu(\lambda) = \underbrace{\sum_k \alpha_k T_\nu^k(\lambda)}_{\text{templates}} + \underbrace{R_\nu(\lambda)}_{\text{residuals}}$$

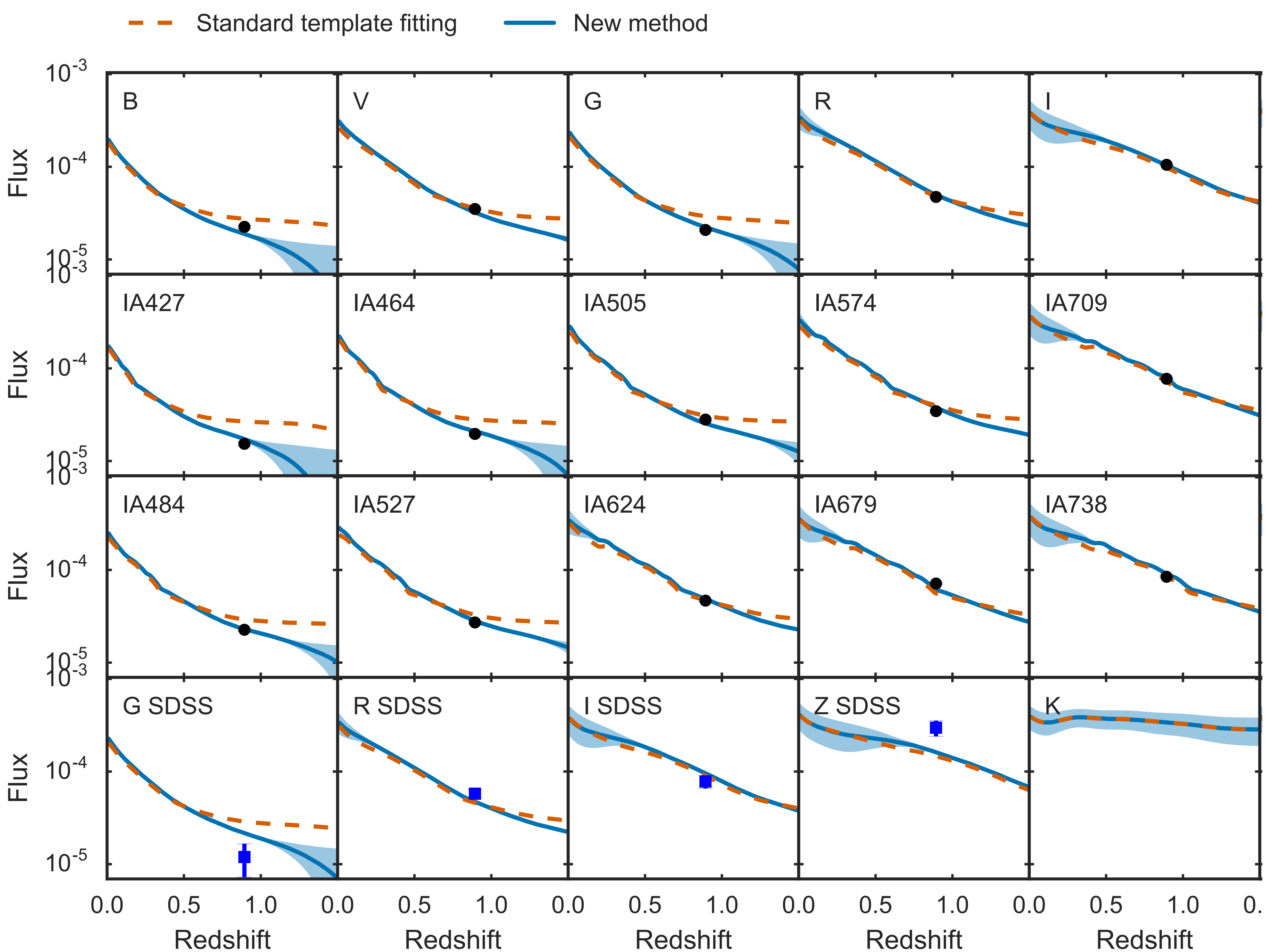
Photo-z gaussian process (proof)

$$R_\nu(\lambda) \sim \mathcal{GP}\Big(0,\ k(\lambda, \lambda')\Big)$$

$$\Longrightarrow \qquad L_\nu(\lambda) \sim \mathcal{GP}\Big(\sum_k \alpha_k T_\nu^k(\lambda),\ k(\lambda, \lambda')\Big)$$

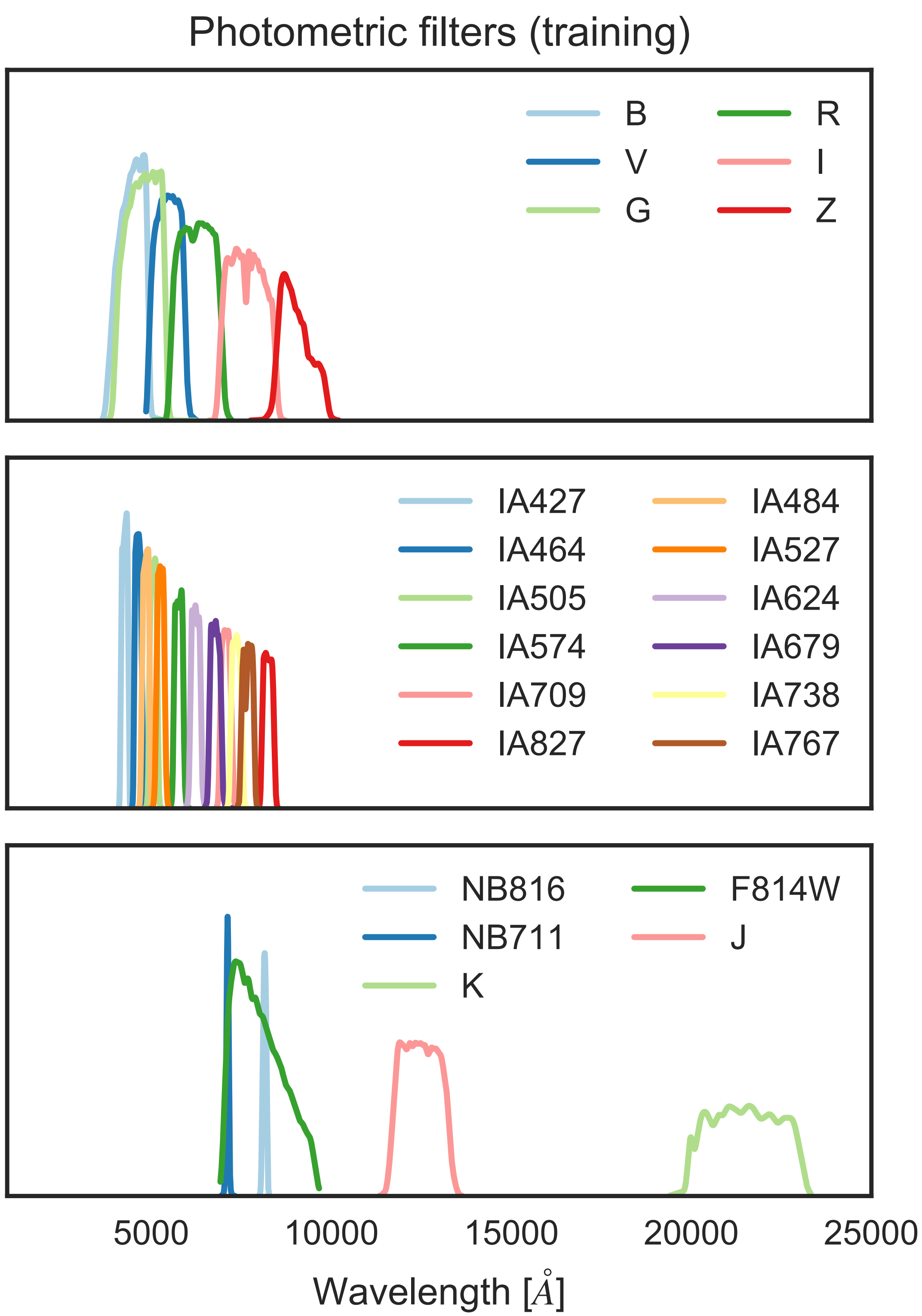
$$F(b, z) = \frac{(1+z)}{4\pi D_L^2(z) C_b} \int_0^\infty L_\nu\left(\frac{\lambda}{(1+z)}\right)\ W_b(\lambda)\ \mathrm{d}\lambda/\lambda$$

$$\Longrightarrow \qquad F(b, z) \sim \mathcal{GP}\Big(\mu^F(b, z),\ k^F(b, b', z, z')\Big)$$



Standard template fitting
New method
Flux
Redshift
B
V
G
R
I
IA427
IA464
IA505
IA574
IA709
IA484
IA527
IA624
IA679
IA738
G SDSS
R SDSS
I SDSS
Z SDSS
K

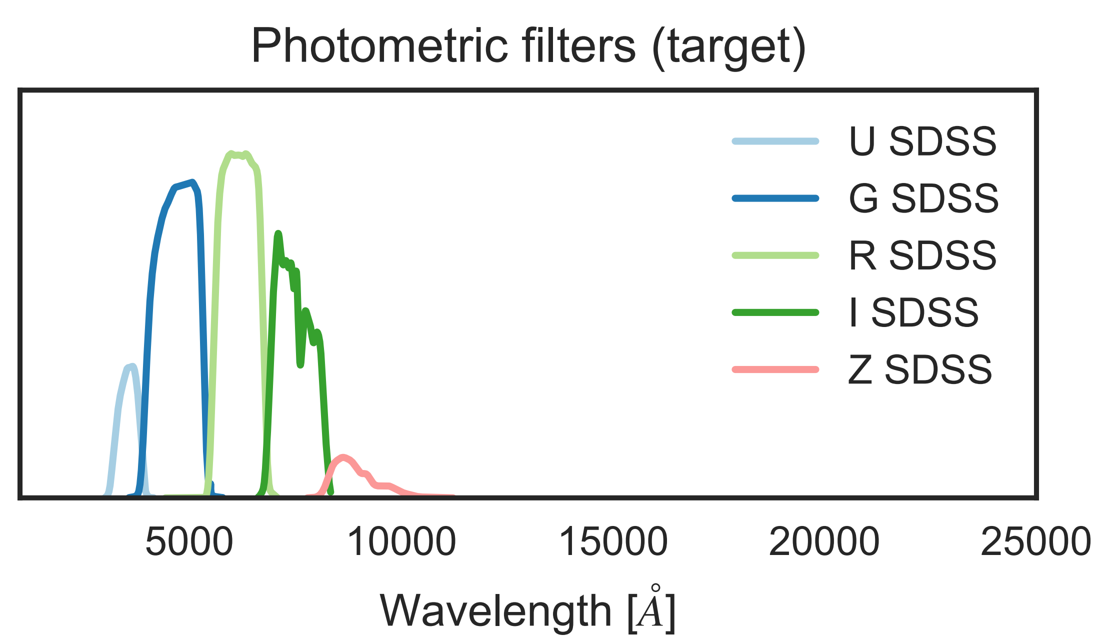
G10 / COSMOS data

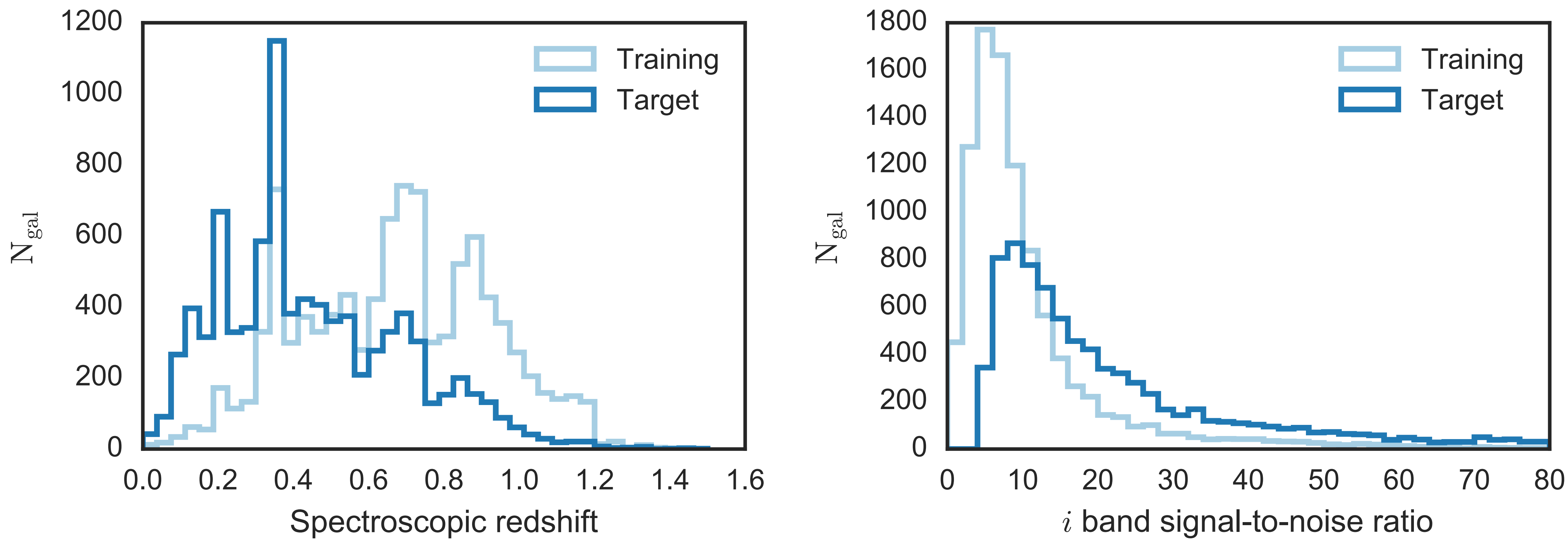


training: deep SUBARU/HST bands **with spectroscopic redshifts**

target: *ugriz* SDSS bands

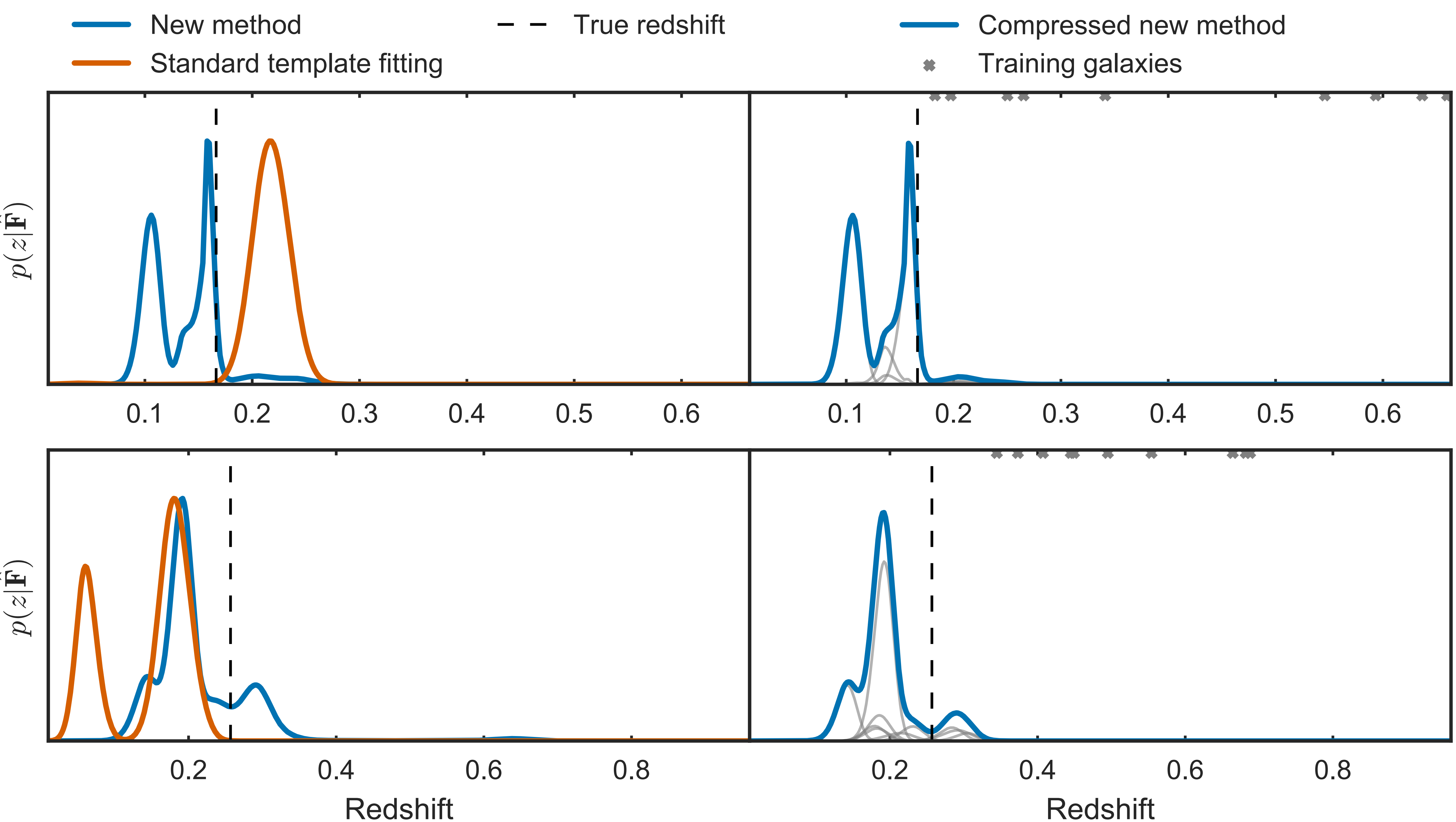
training/target: 10k/10k objects

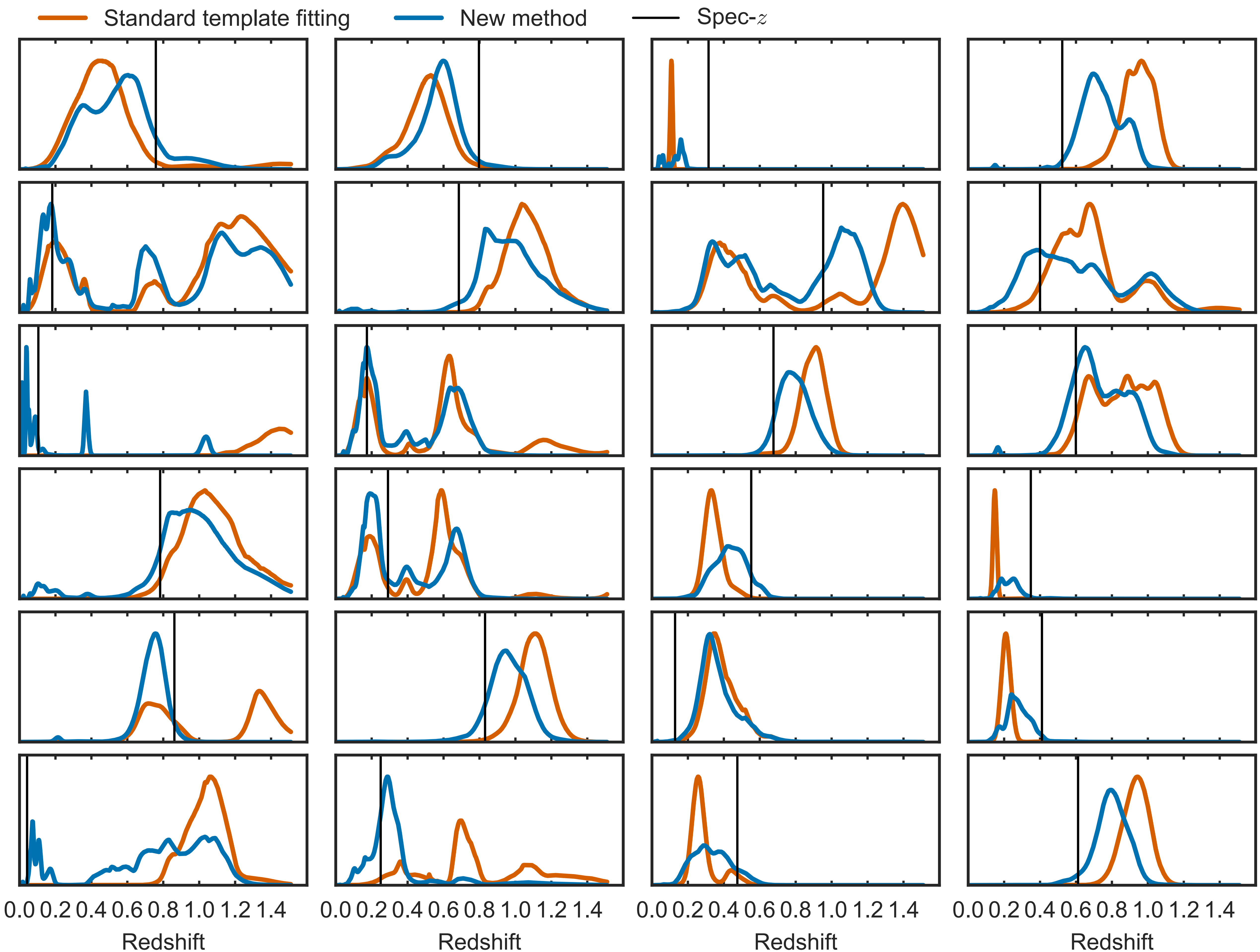




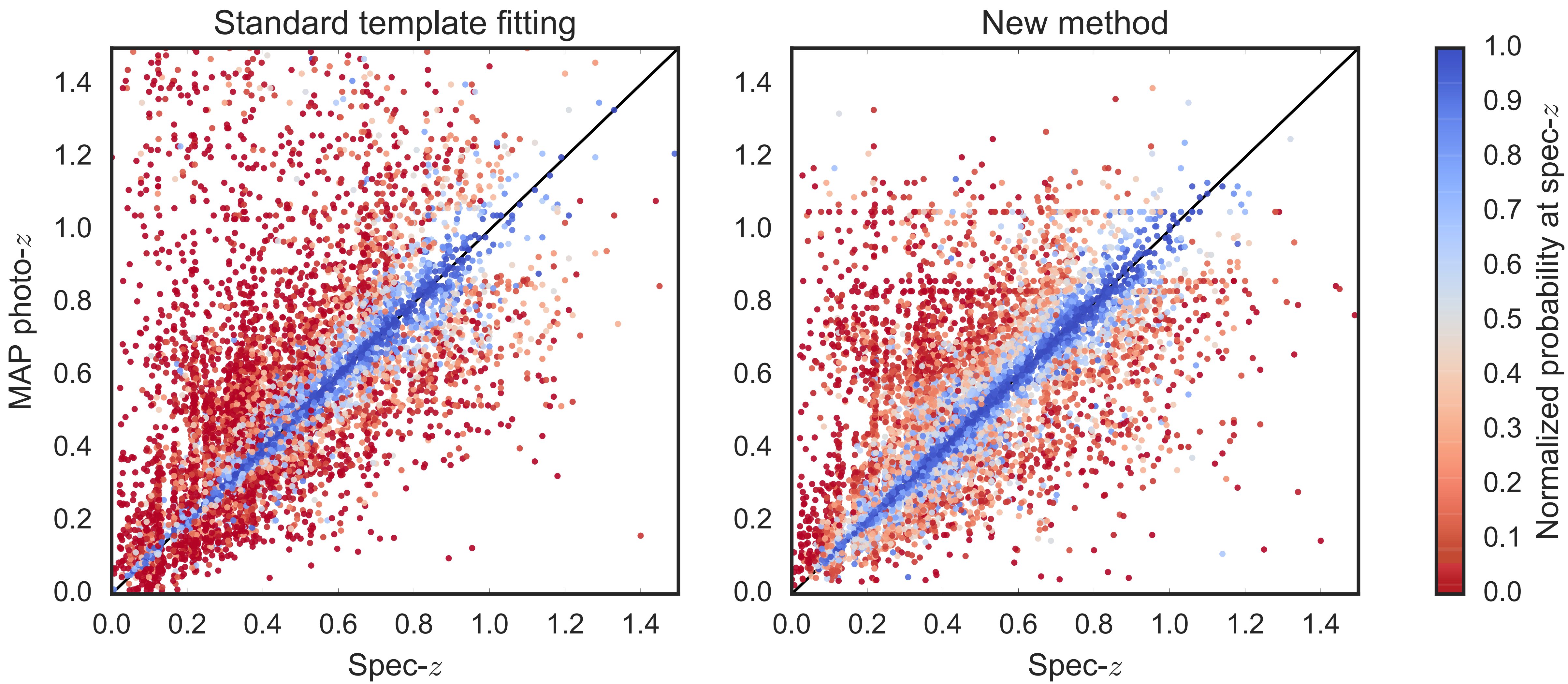
unrepresentative training set with different bands & noise

a closer look at two PDFs…





Standard template fitting
New method
Spec-z
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4
Redshift



7 fixed templates ⇒ 10,000 probabilistic templates

(a system of types) (one per training galaxy)

Improvement, but more data/flexibility required.

Not exploiting low-dimensionality of galaxy types.

Conclusions

Imaging surveys

diverse science: fundamental physics, astrophysics

systematics limited — require exquisite photo-z's

***DELIGHT* — GITHUB.COM/IXKAEL/DELIGHT**

data-driven method with physics & machine learning

delivers accurate, interpretable redshifts probabilities

What's next?

fit SED templates and luminosity functions, calibrate photo-z likelihood without spectroscopic redshifts

LSST DESC
Dark Energy Science Collaboration

20 billion galaxies

17 billion stars

7 trillion sources detected
in single epochs

30 trillion forced photometry

10 million alerts per nigh