Oskar Klein centre

Funded by the European Union

European Research Council
Established by the European Commission

pop-cosmos:

Scaleable Bayesian inference of galaxy properties under a diffusion model prior


<u>Stephen Thorp</u>


Sinan Deger


Anik Halder


Gurjeet Jagwani

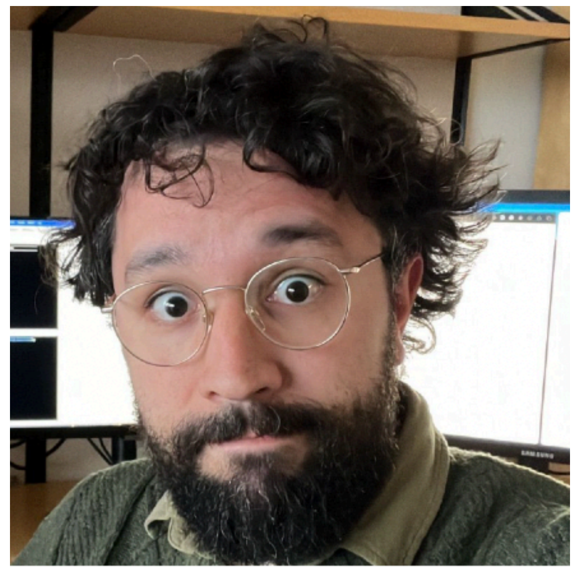

Hiranya Peiris


Madalina Tudorache


Justin Alsing


Boris Leistedt

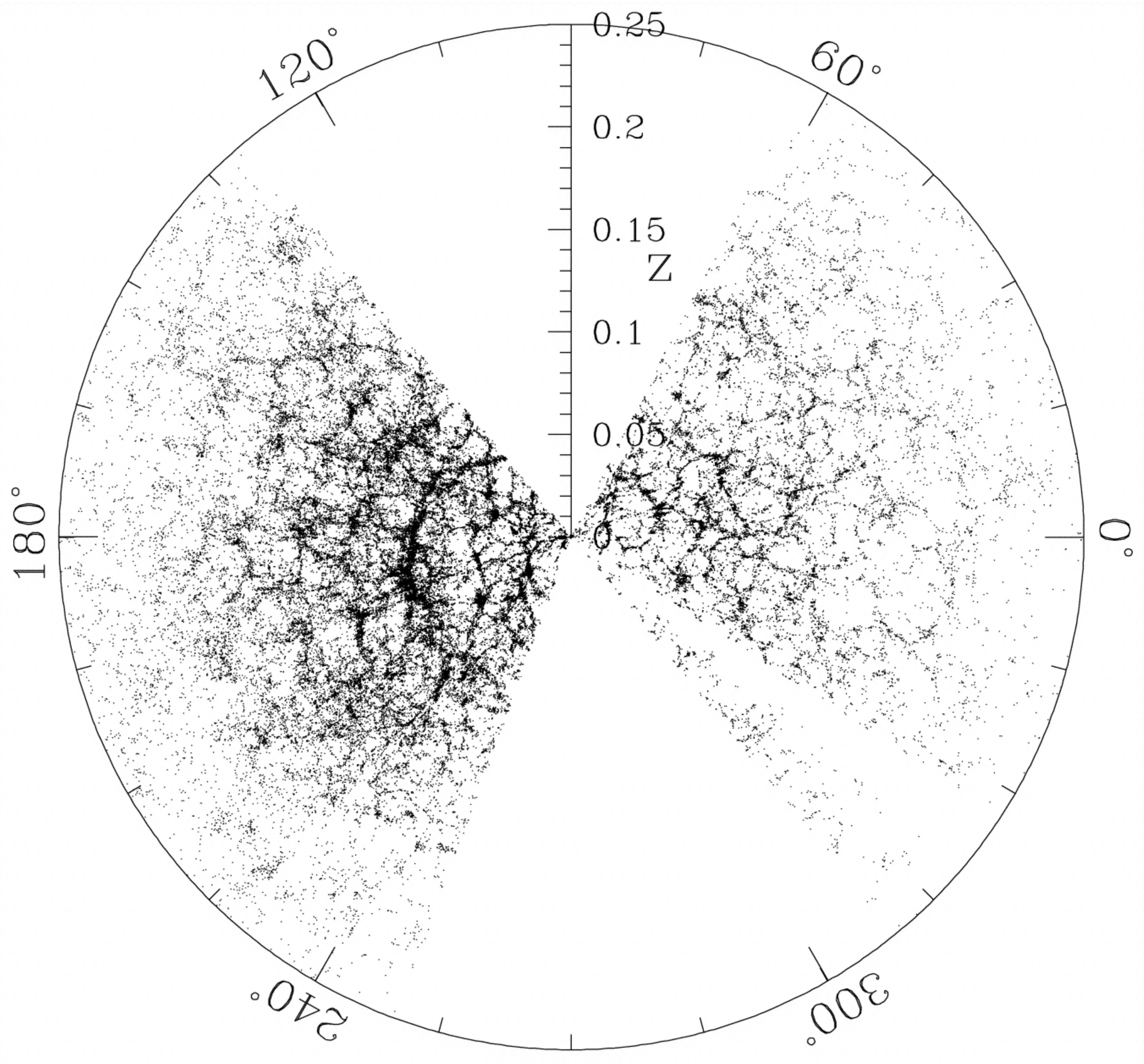

Joel Leja


Arthur Loureiro


Daniel Mortlock

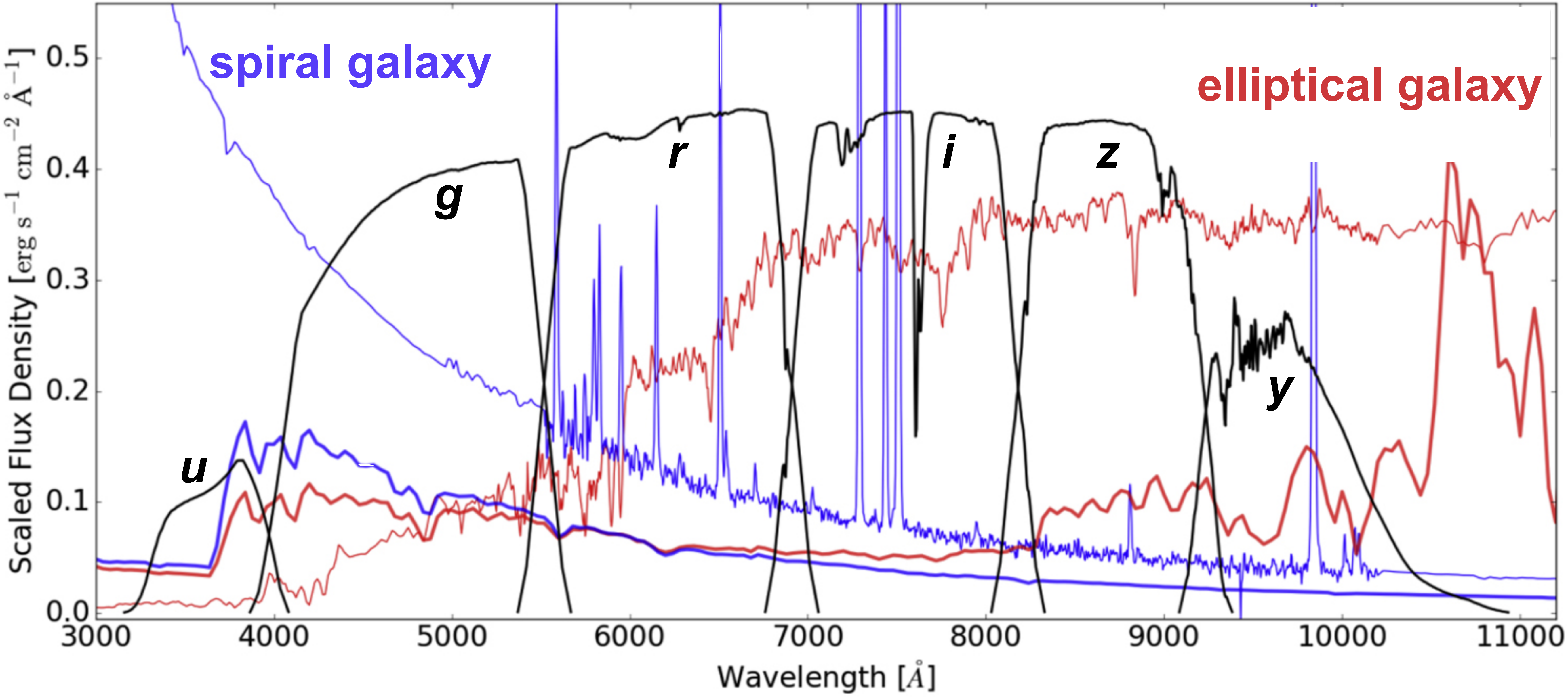
Large Scale Structure Cosmology

- Map the matter distribution in the universe using spectra or photometry of galaxies
- Galaxy clustering (positions), or weak lensing (shapes) sensitive to cosmological parameters
- Photometric surveys depend more on our ability to model galaxies well (either individually or as a population)



SDSS: *Blanton et al. (2003)*

Photometry vs. Spectroscopy



Graham et al. (2017)

Coming (very) soon…

- First-look events June 23rd: https://rubinobservatory.org
- First data June 30th!
- 20 billion galaxies; 18,000 deg^2
- Deep imaging in $ugrizy$
- Single epoch: $r \lesssim 24\,\mathrm{mag}$;
  10 year co-add: $r < 26.9\,\mathrm{mag}$
- Photometric data only; need to model galaxies with very limited information (just 6 numbers)

VERA C. RUBIN OBSERVATORY

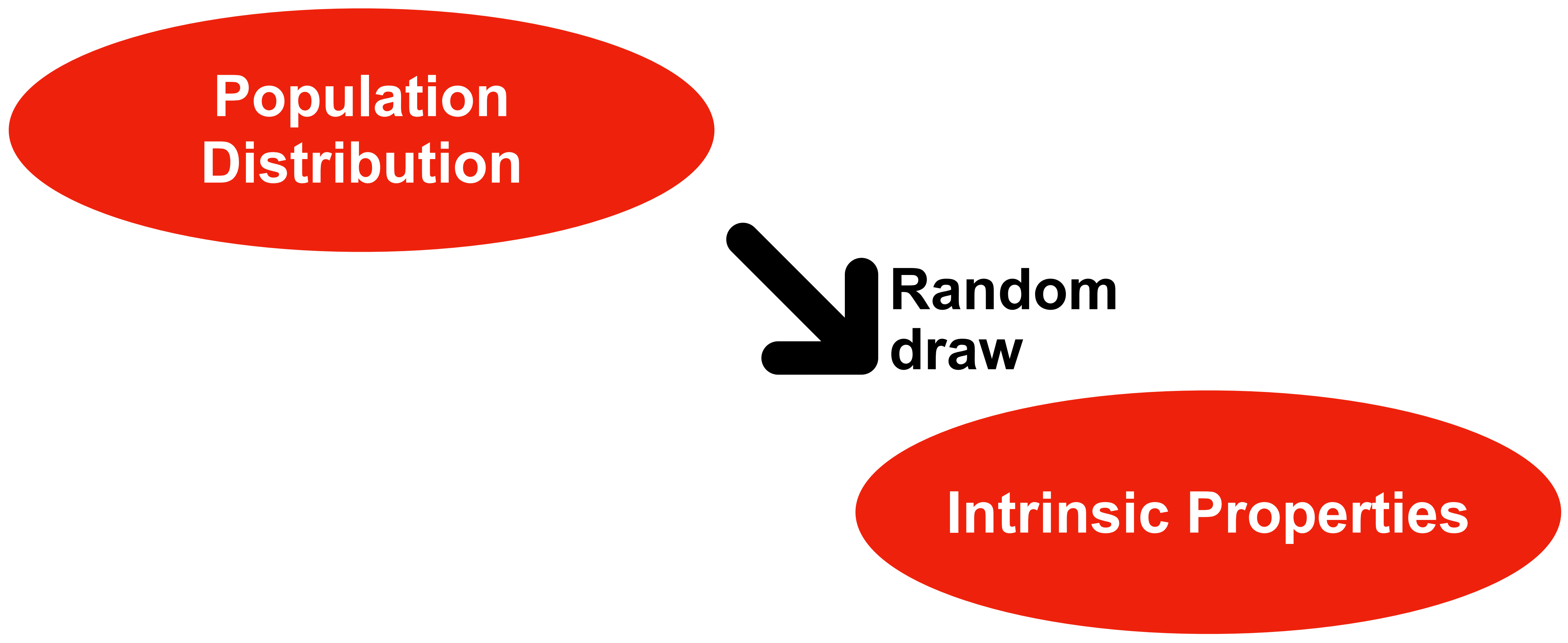
LSST Legacy Survey of Space and Time

Credit: Rubin Observatory/NOIRLab/SLAC/DOE/NSF/AURA/B. Quint

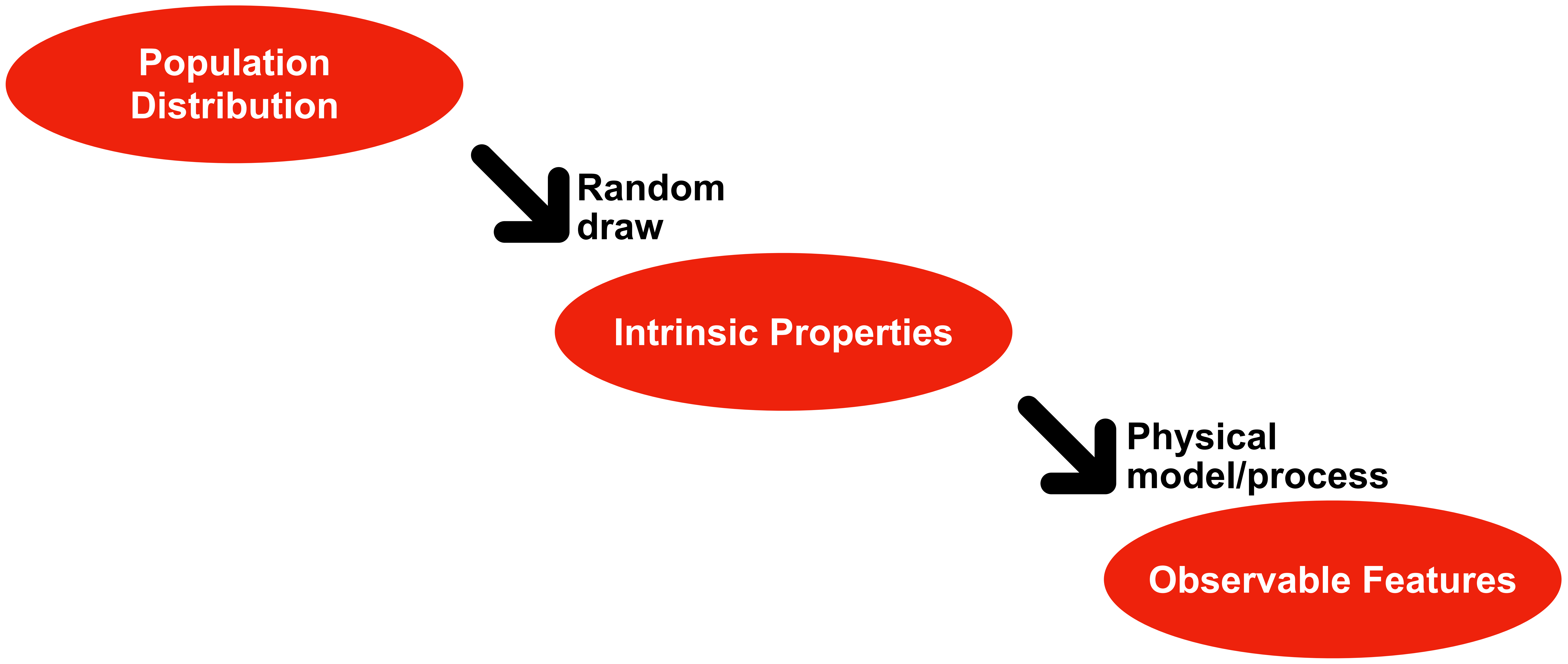
Setting up a generative model

Population Distribution

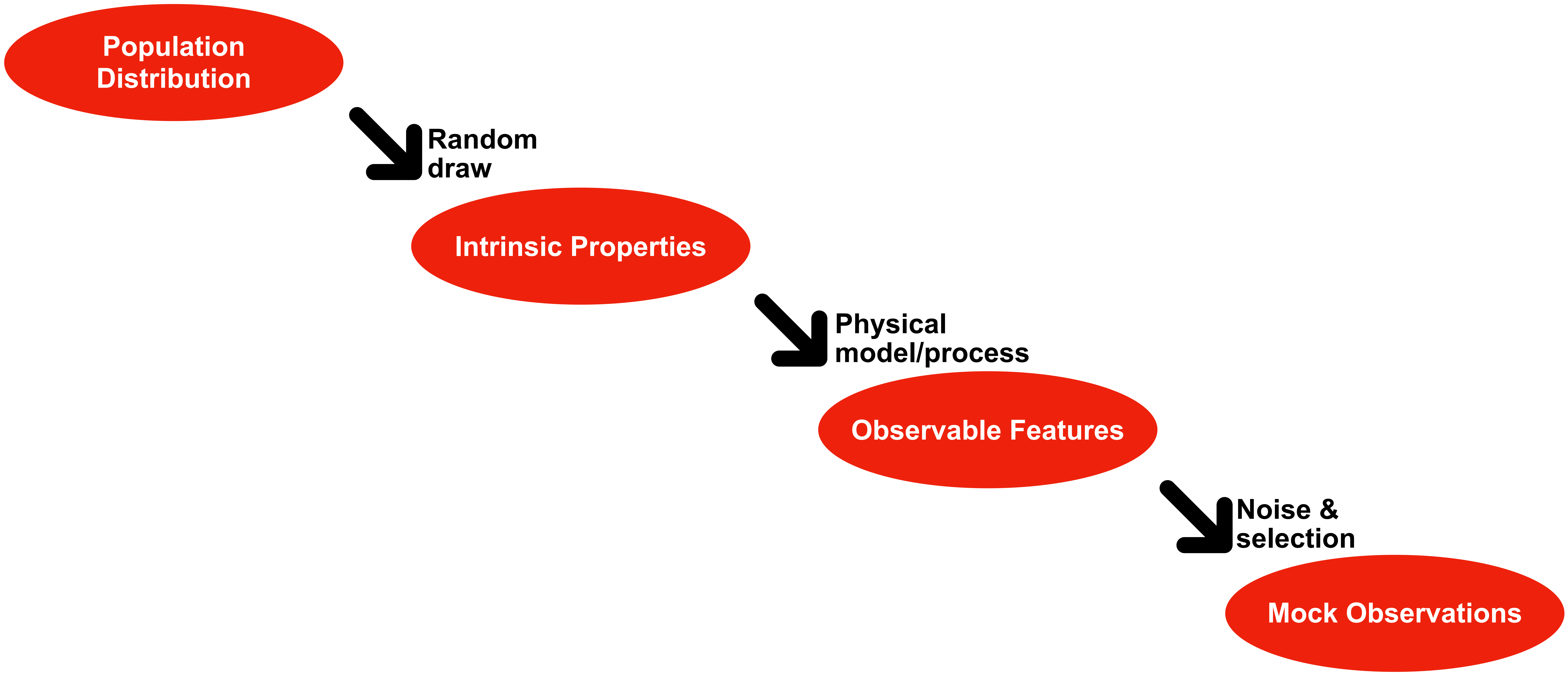
Setting up a generative model



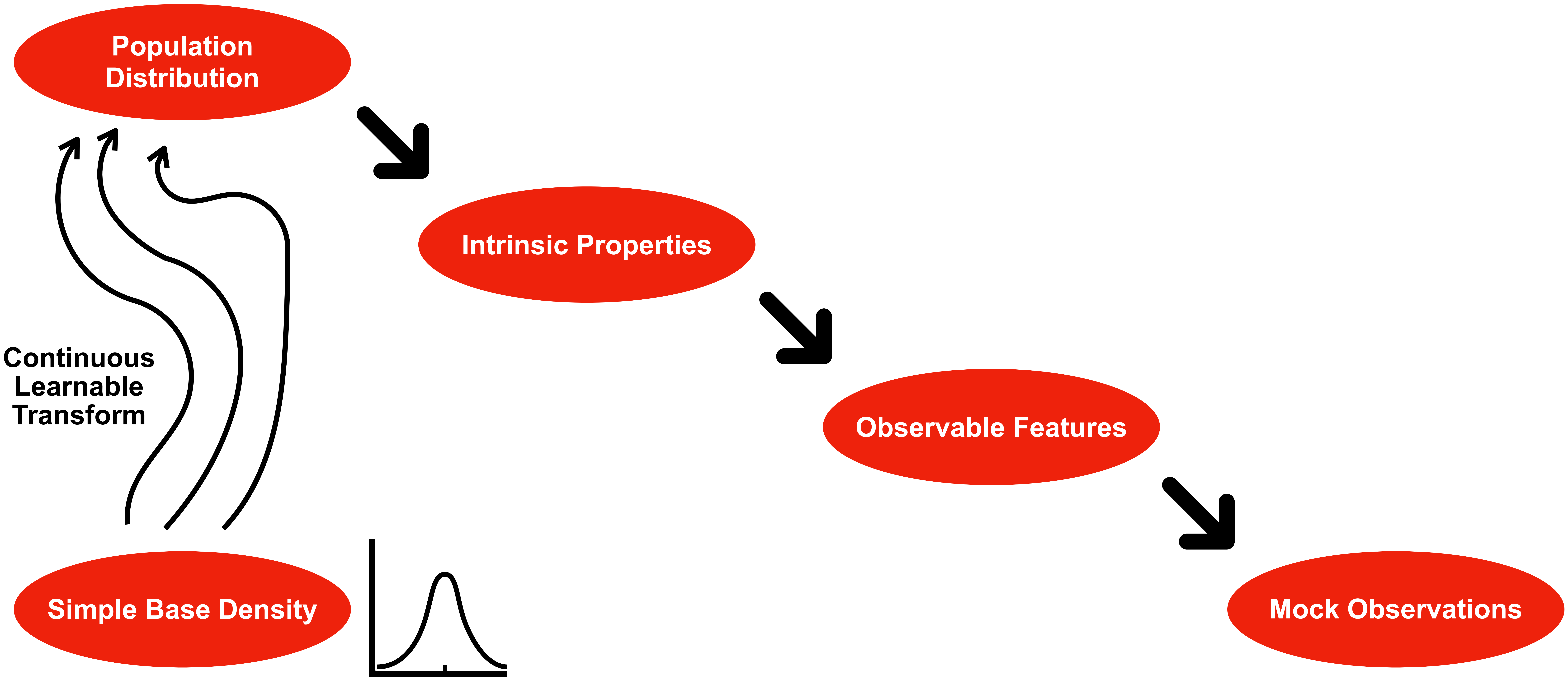
Setting up a generative model



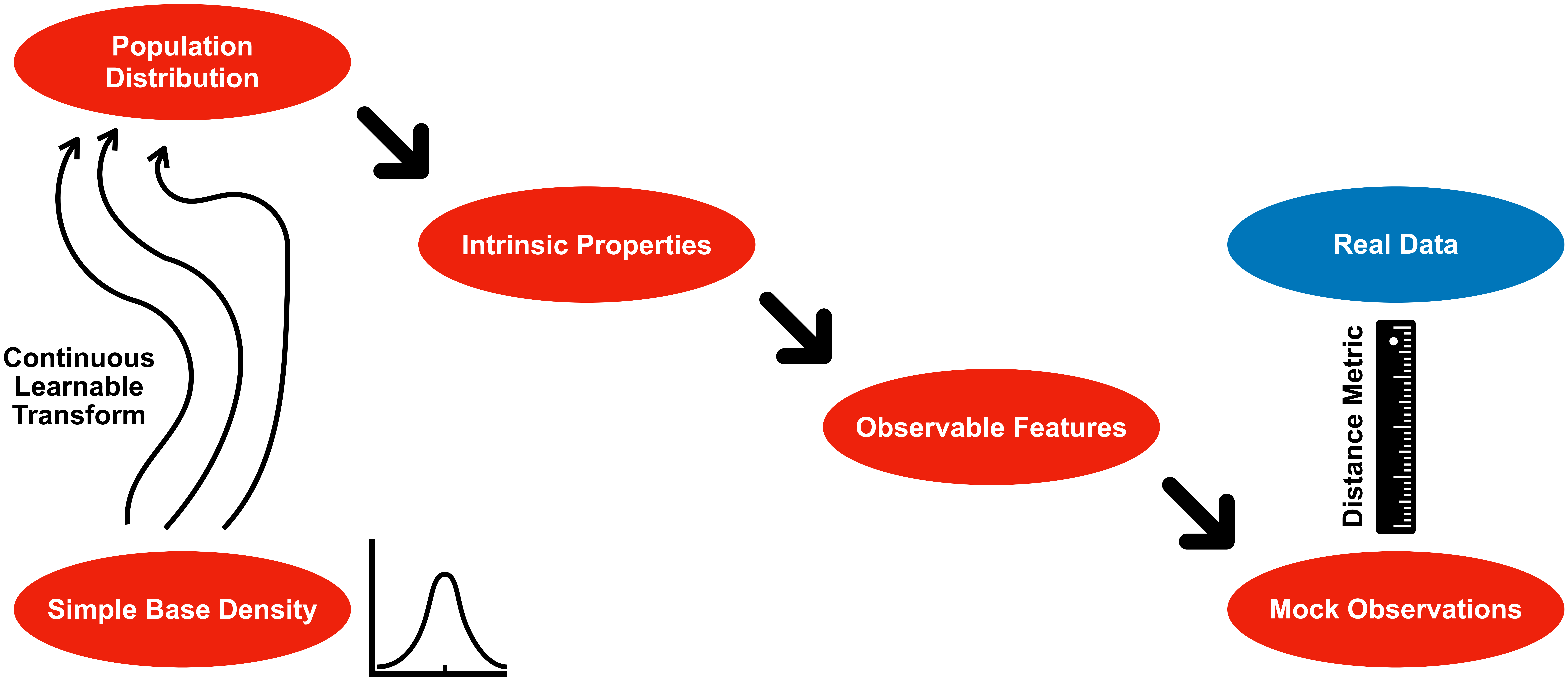
Setting up a generative model



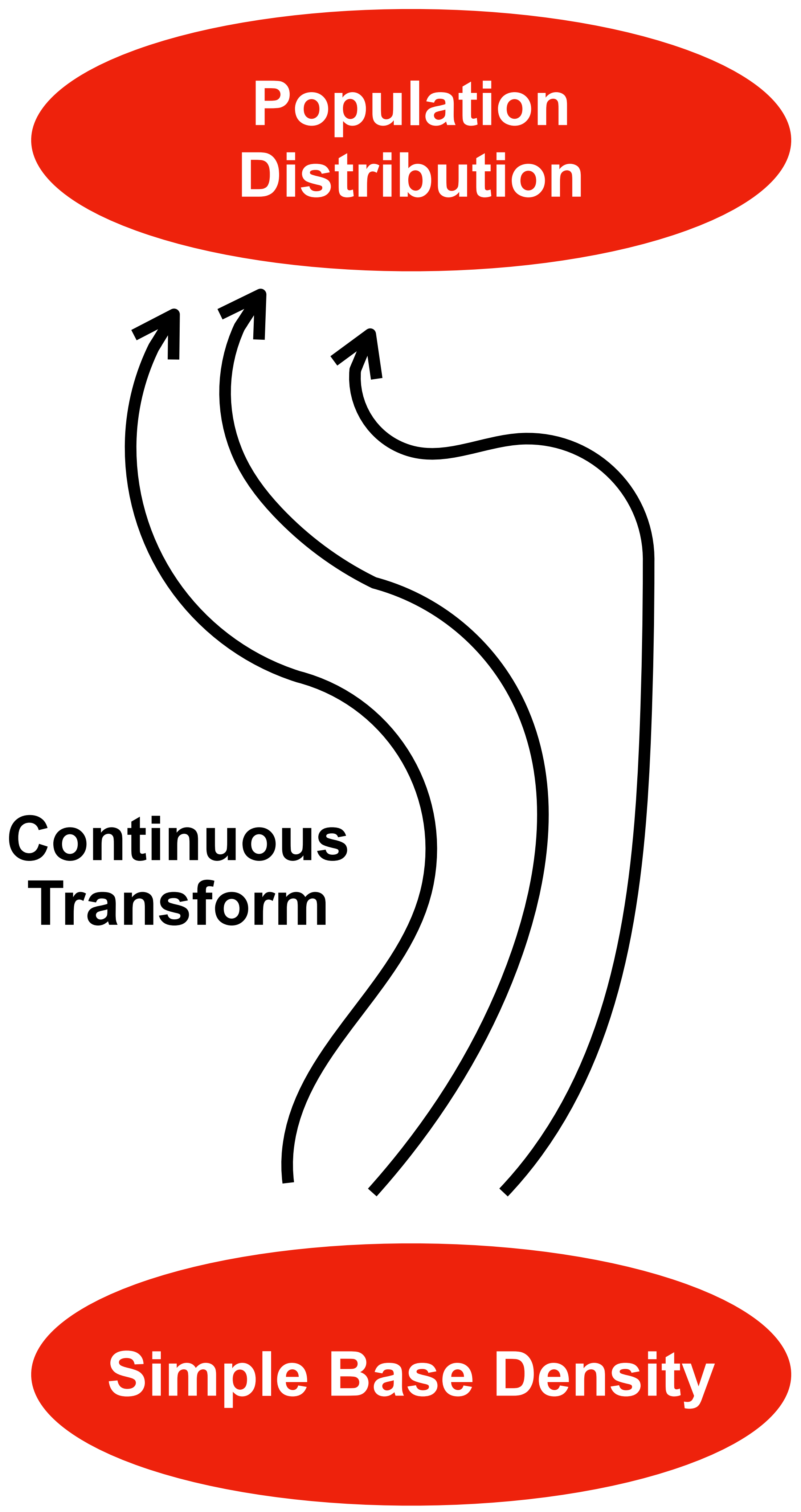
Diffusion models



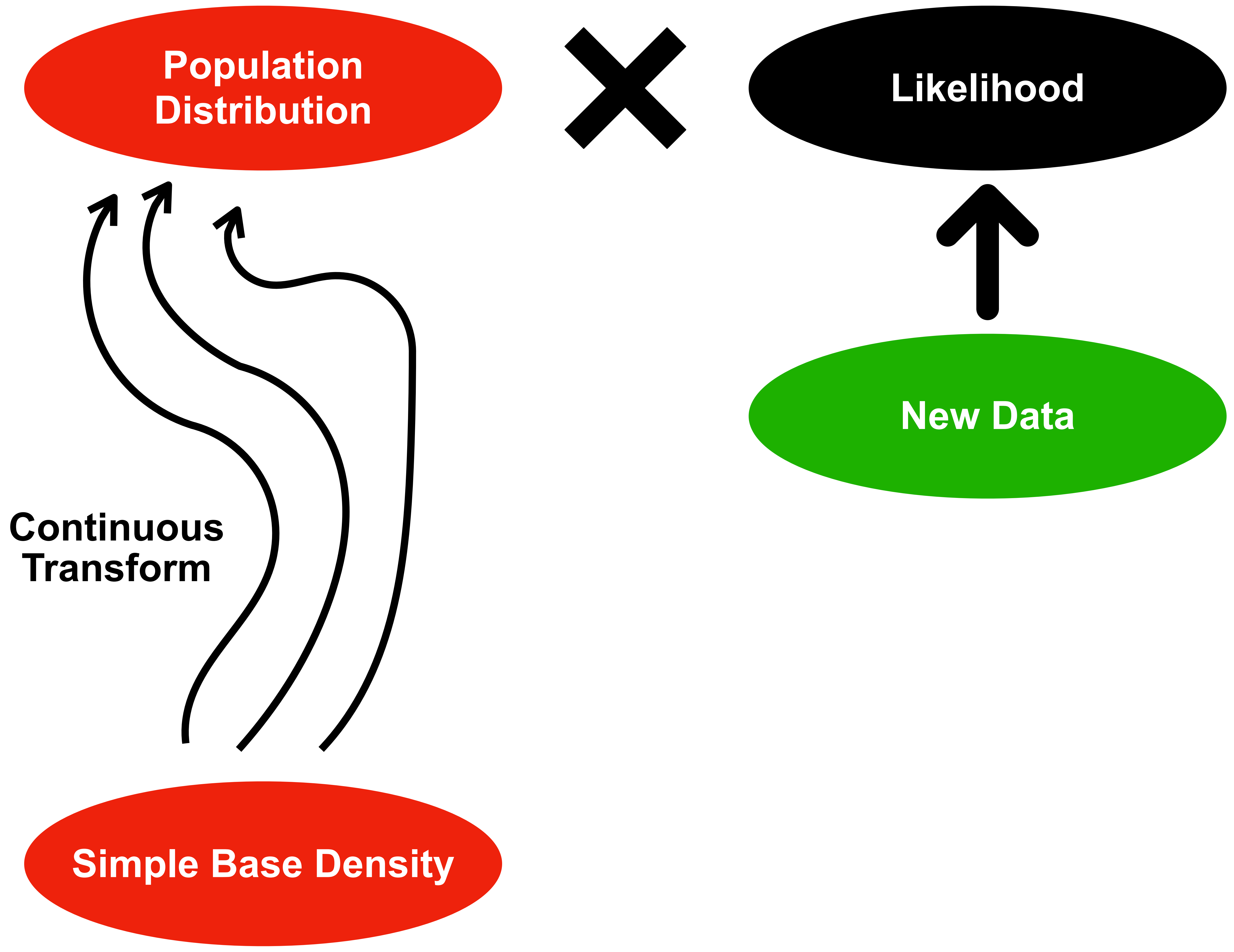
Fitting a generative model to data



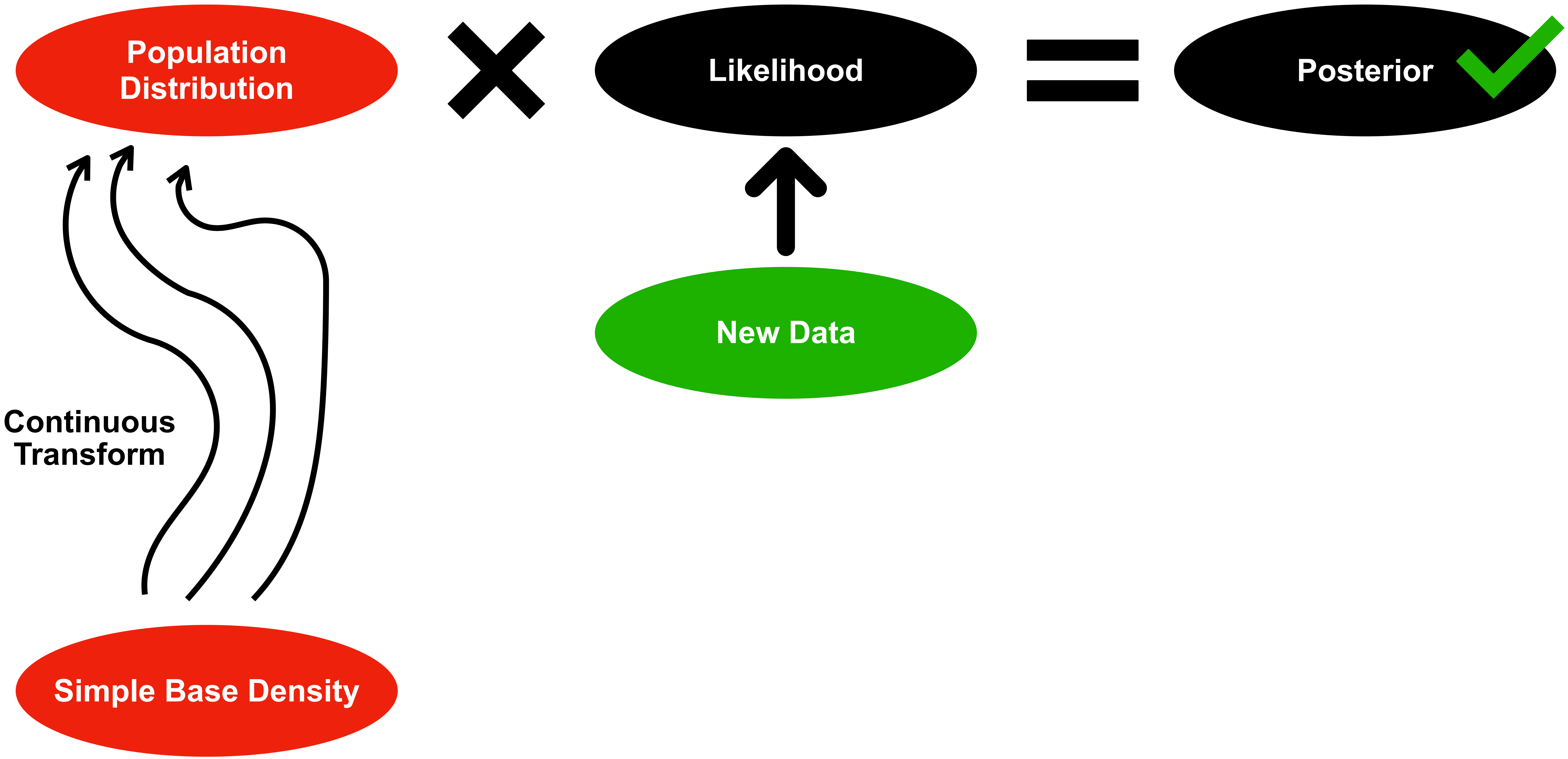
Inference under a diffusion model prior



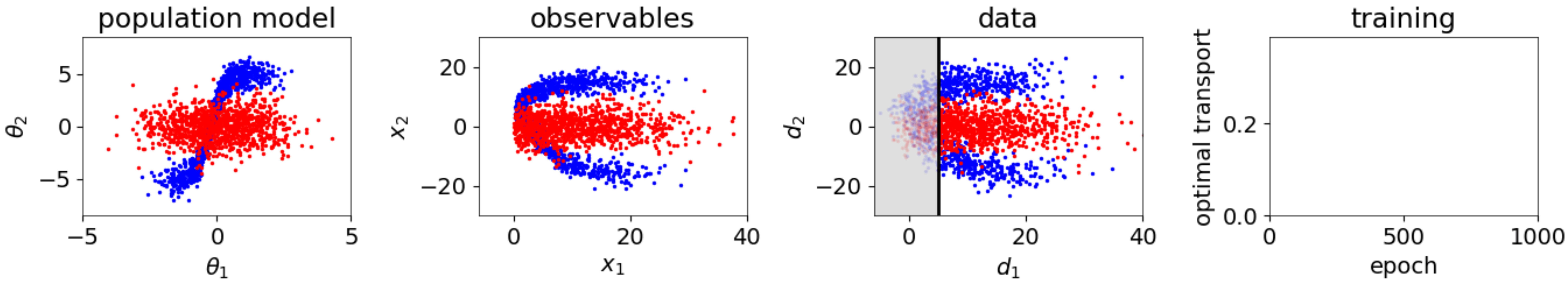
Inference under a diffusion model prior



Inference under a diffusion model prior



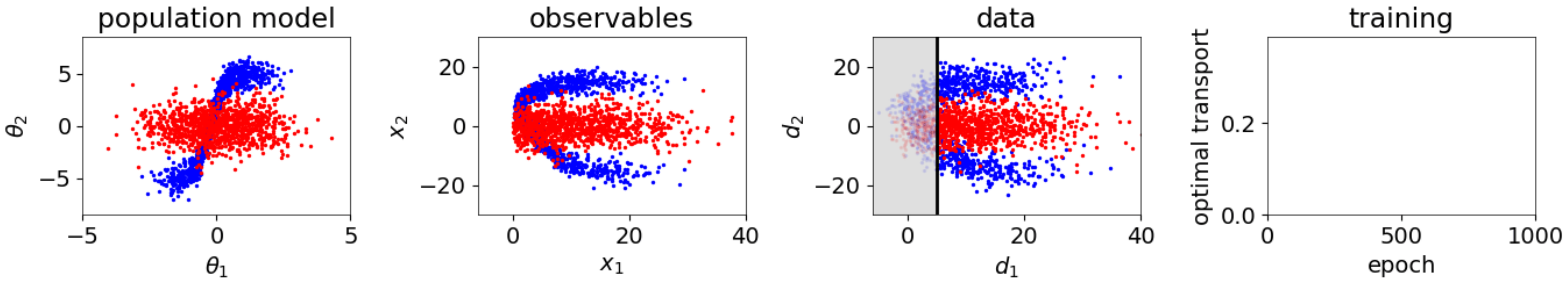
How do we forward model a galaxy catalog?



Credit: Justin Alsing

How do we forward model a galaxy catalog?

stellar population synthesis (SPS) parameters

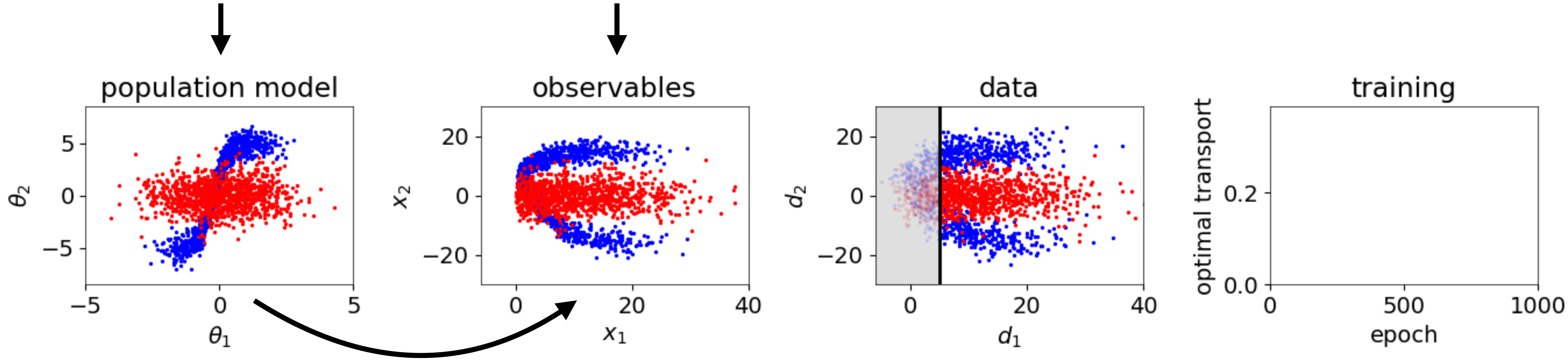


Credit: Justin Alsing

How do we forward model a galaxy catalog?

stellar population synthesis (SPS) parameters

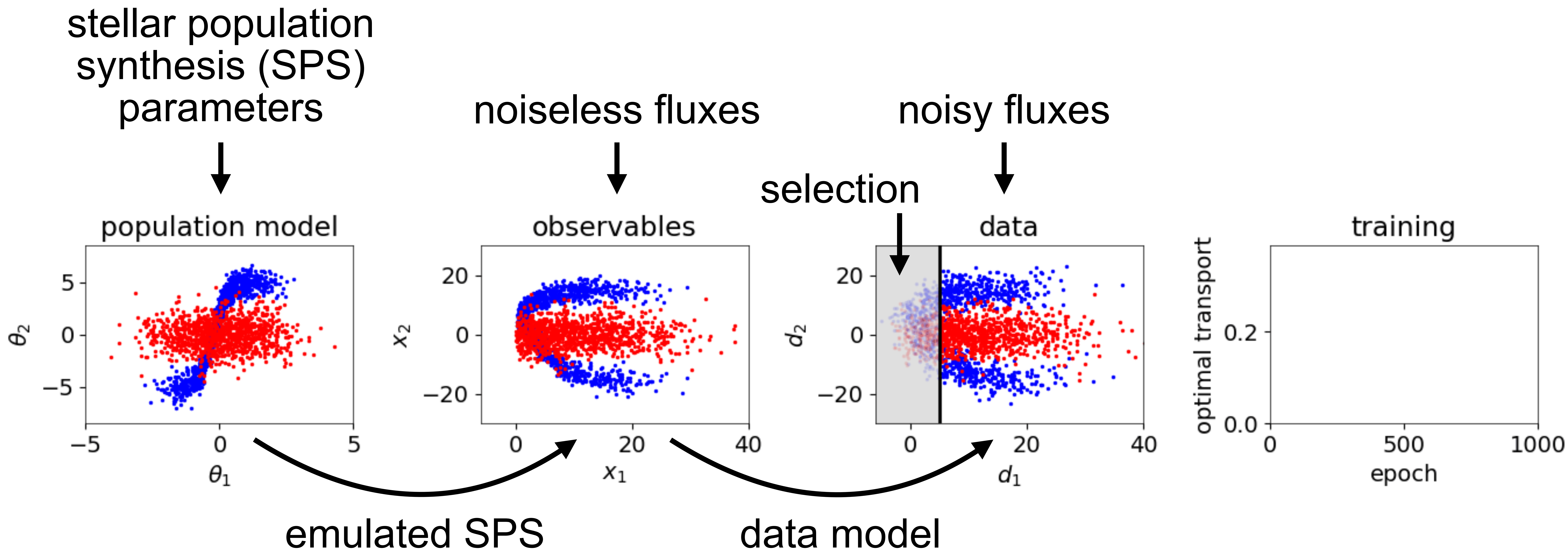
noiseless fluxes



emulated SPS

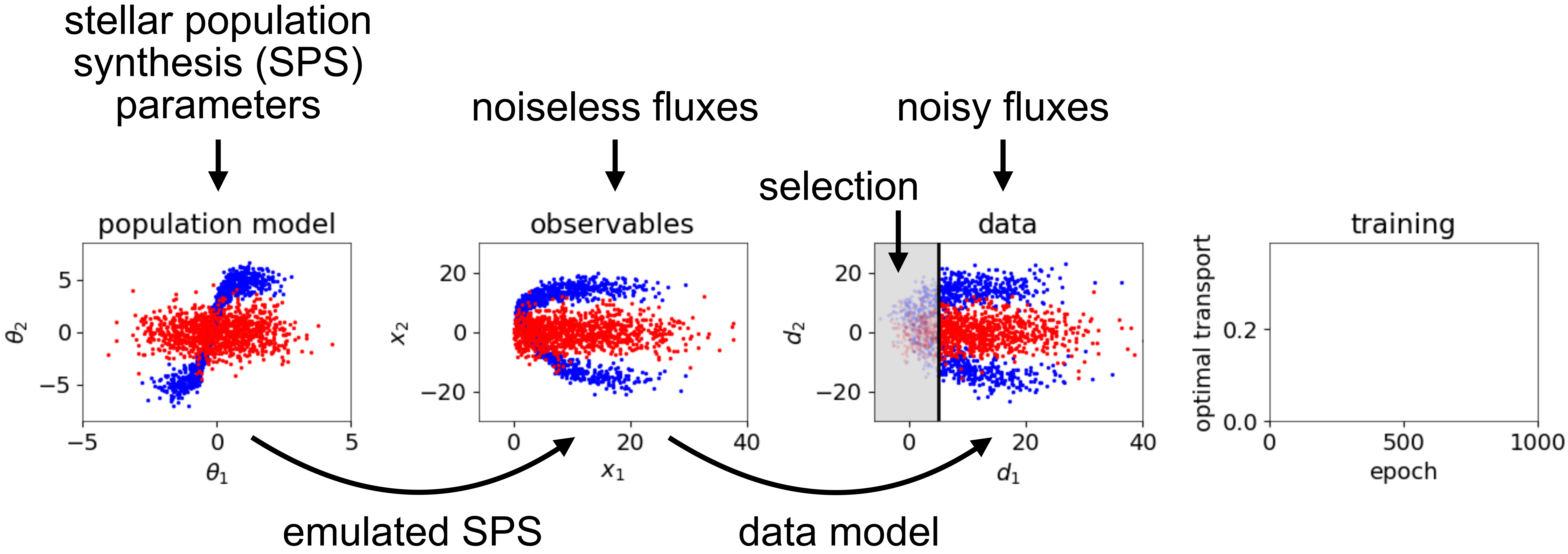
Credit: Justin Alsing

How do we forward model a galaxy catalog?



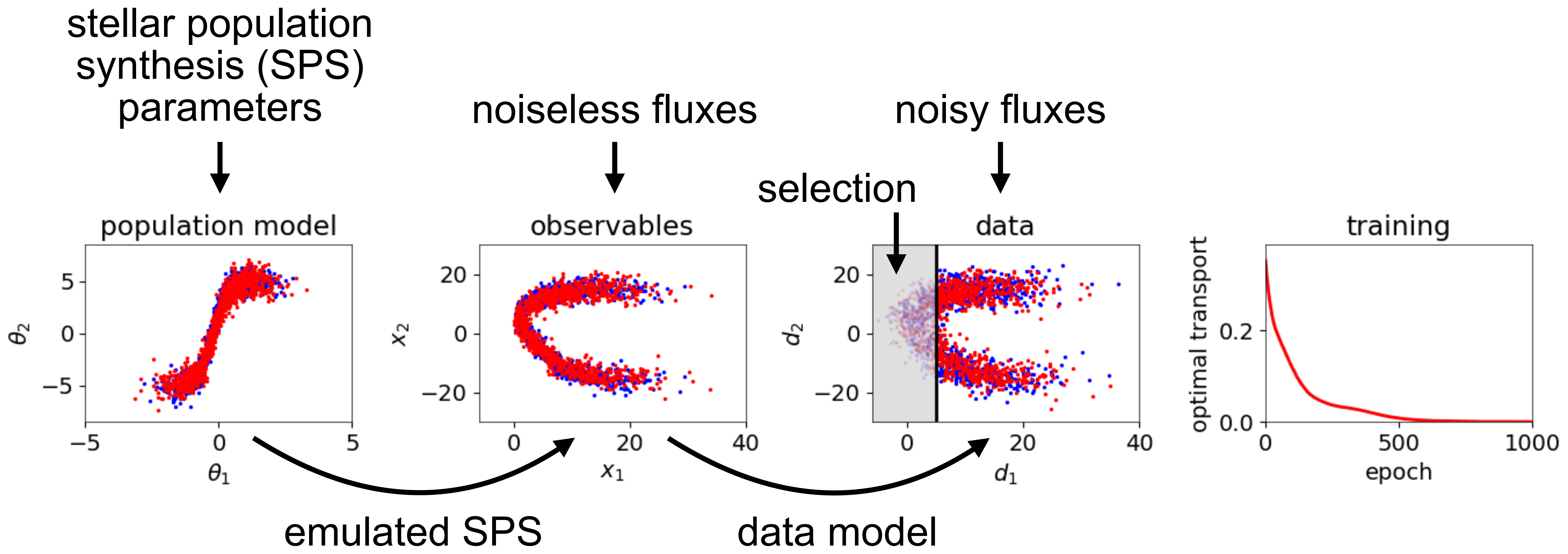
Credit: Justin Alsing

How do we forward model a galaxy catalog?



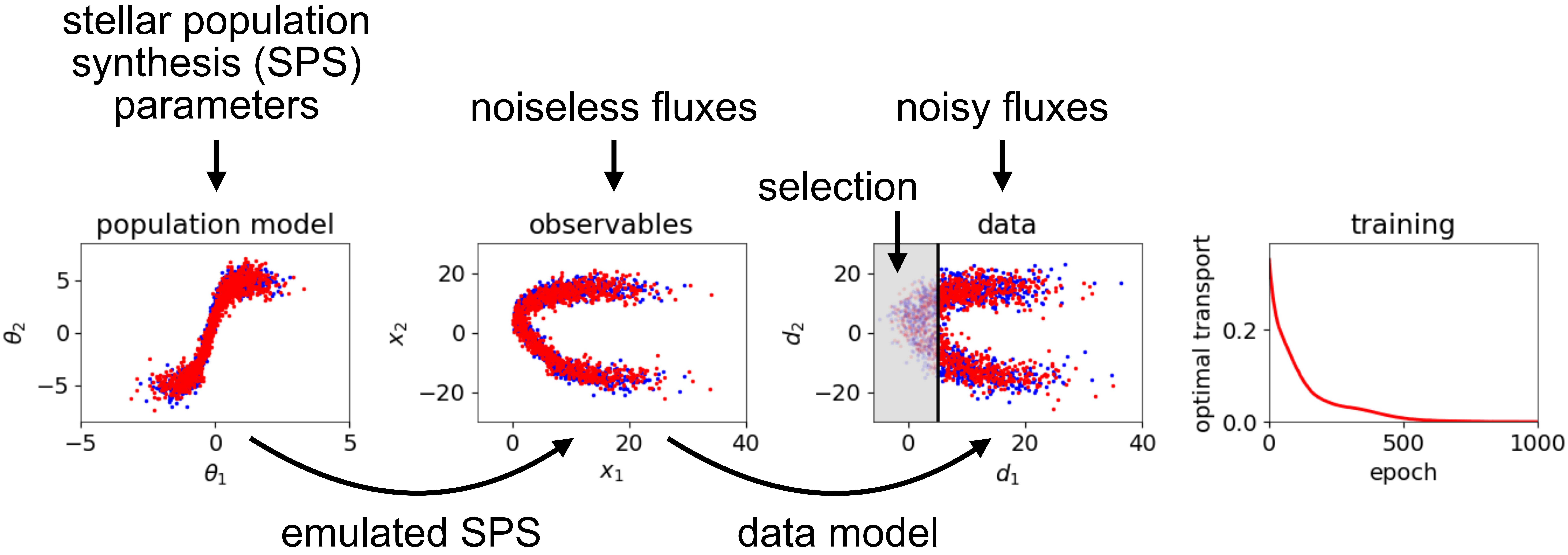
Credit: Justin Alsing

How do we forward model a galaxy catalog?



Credit: Justin Alsing

How do we forward model a galaxy catalog?

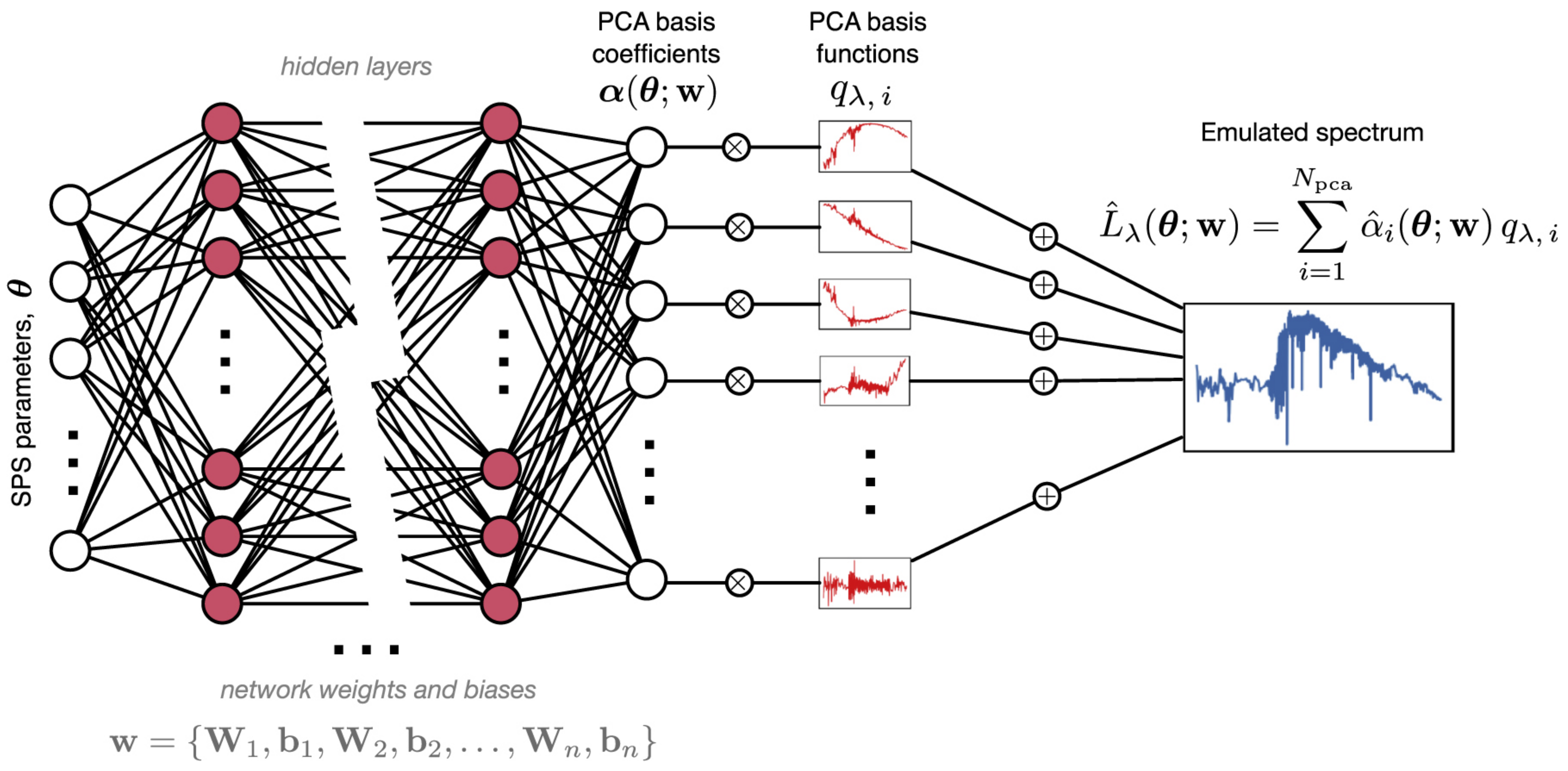


Credit: Justin Alsing

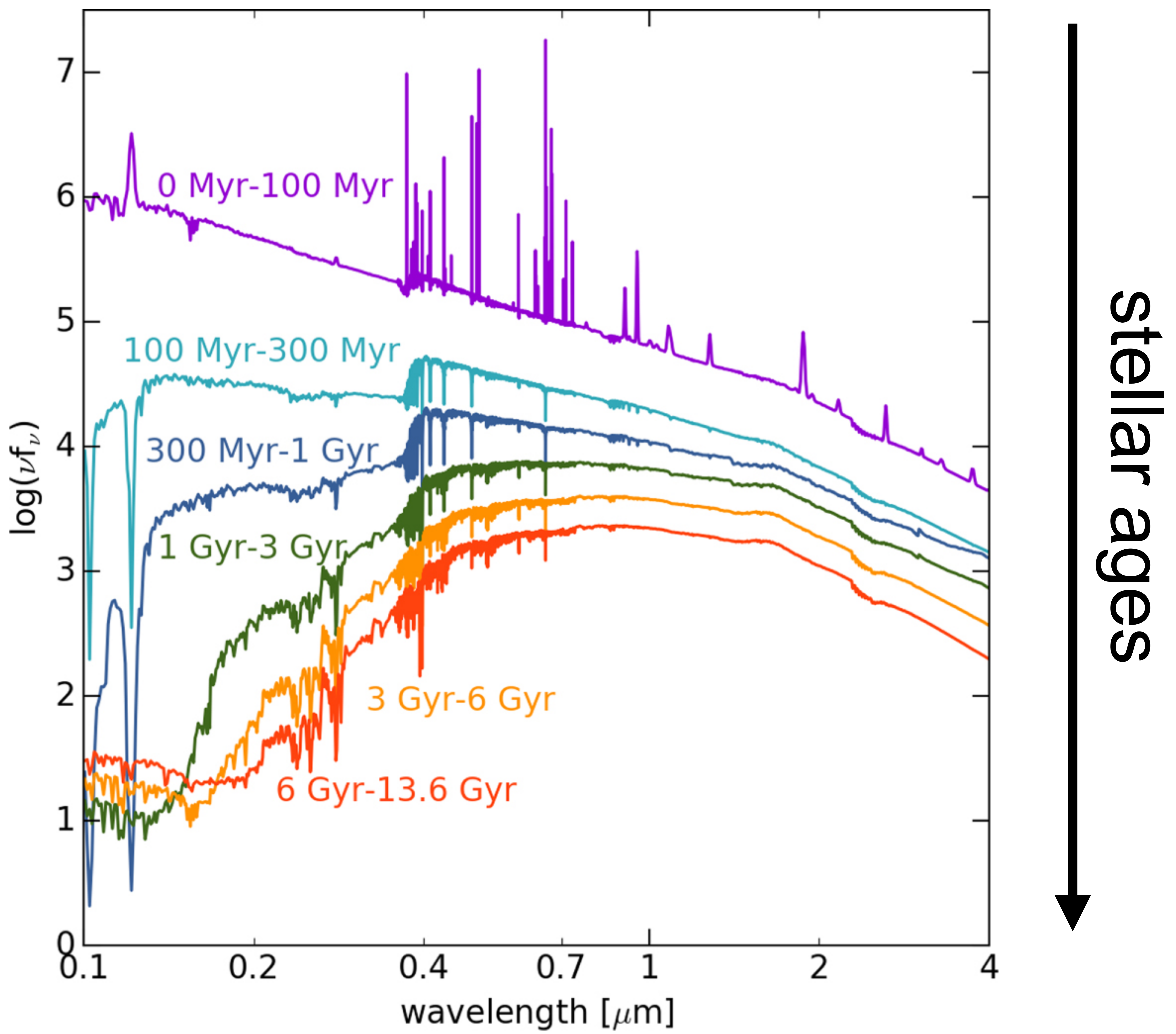
How do we represent a galaxy?

Stellar Population Synthesis (SPS)

- 16 parameters ($\approx$ Prospector-α)
- Emulated with Speculator



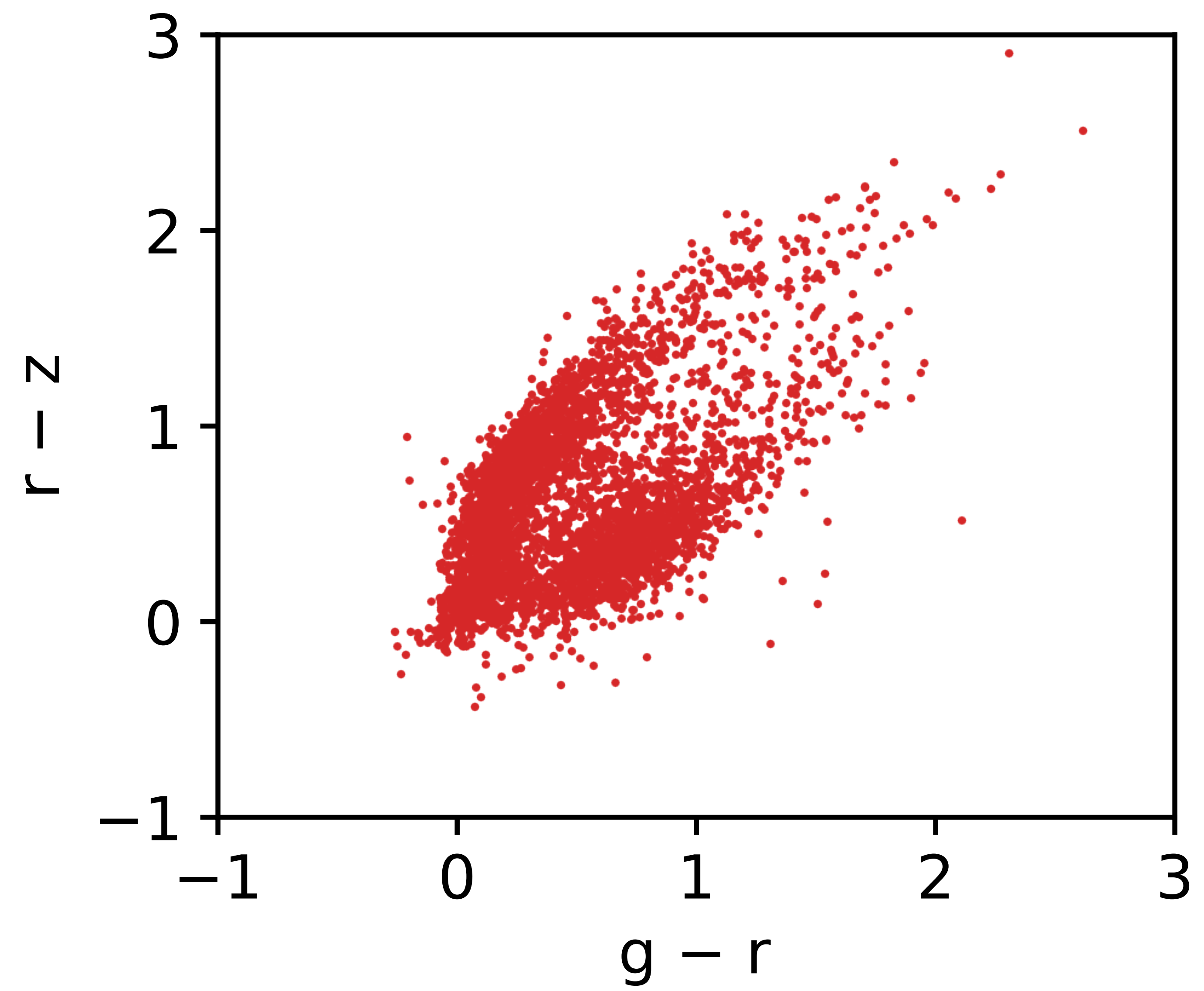
Alsing et al. (2020)



Leja et al. (2017)

What will our population model be?

A score-based diffusion model



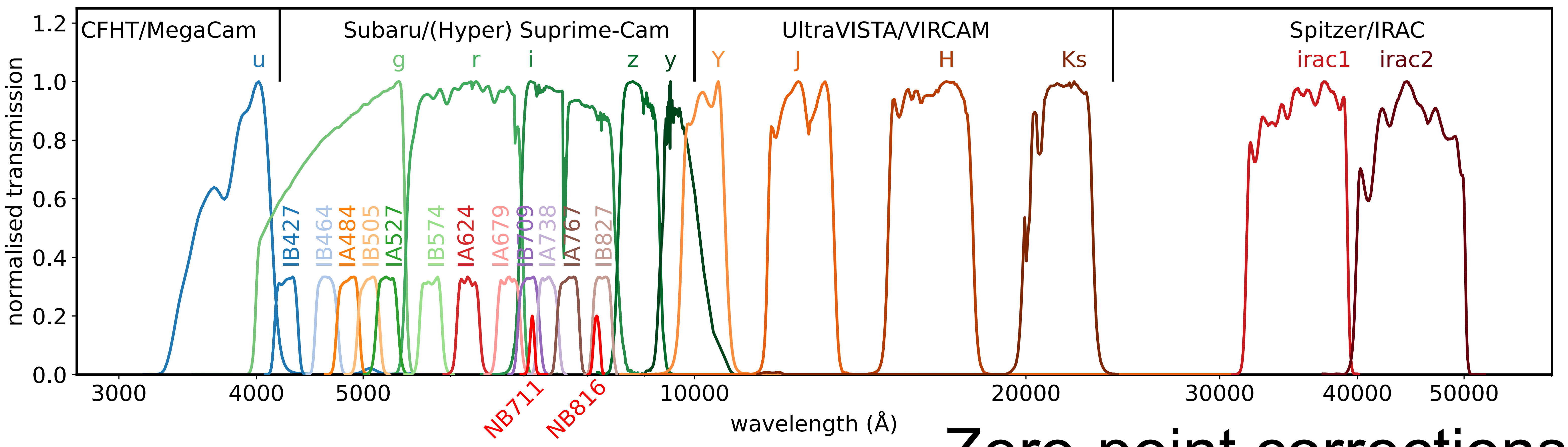
target distribution over physical parameters ($t = 0$)

differential equation

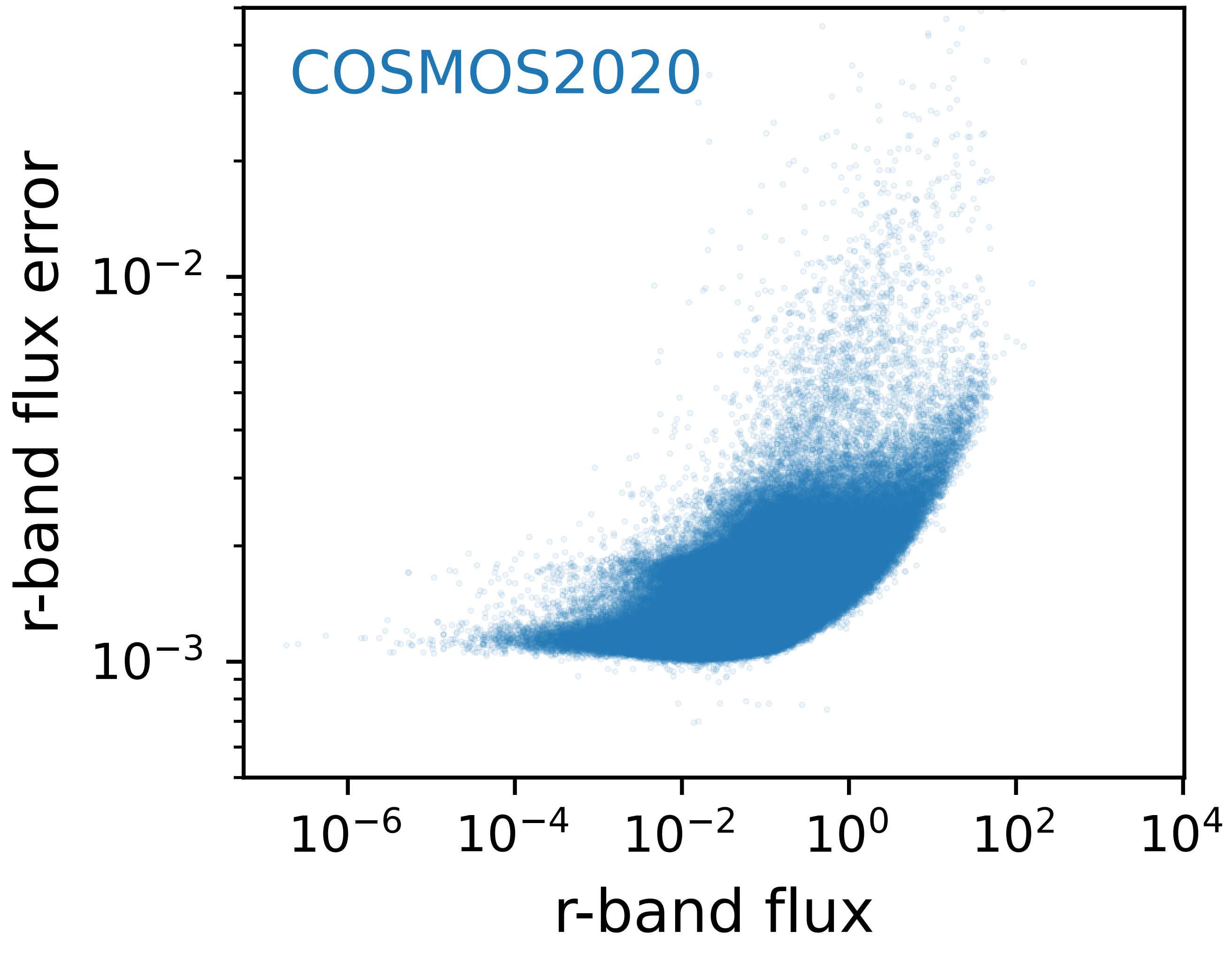
iid Gaussian noise ($t = T$)

What goes into the data model?

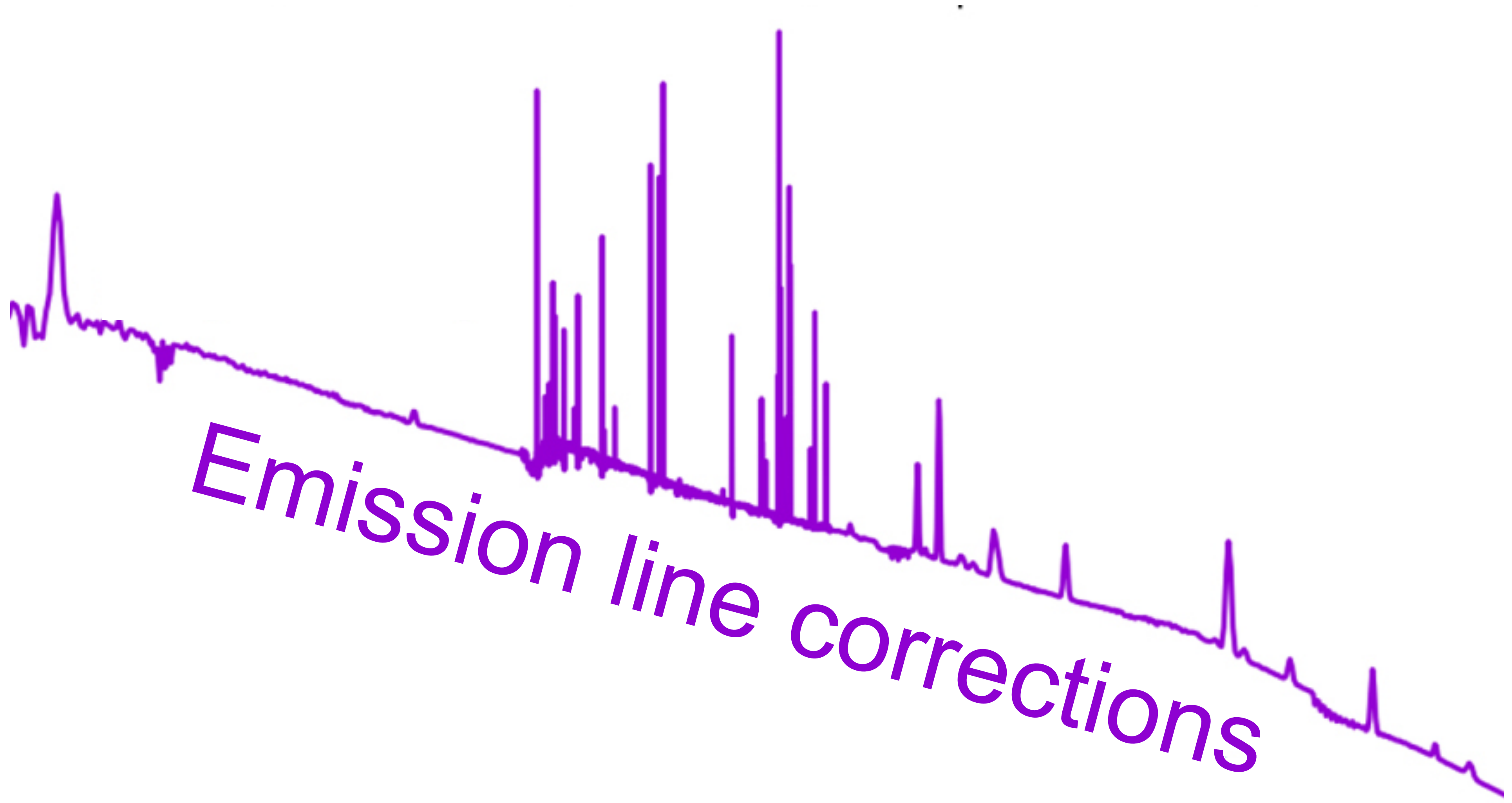
Uncertainty and calibration



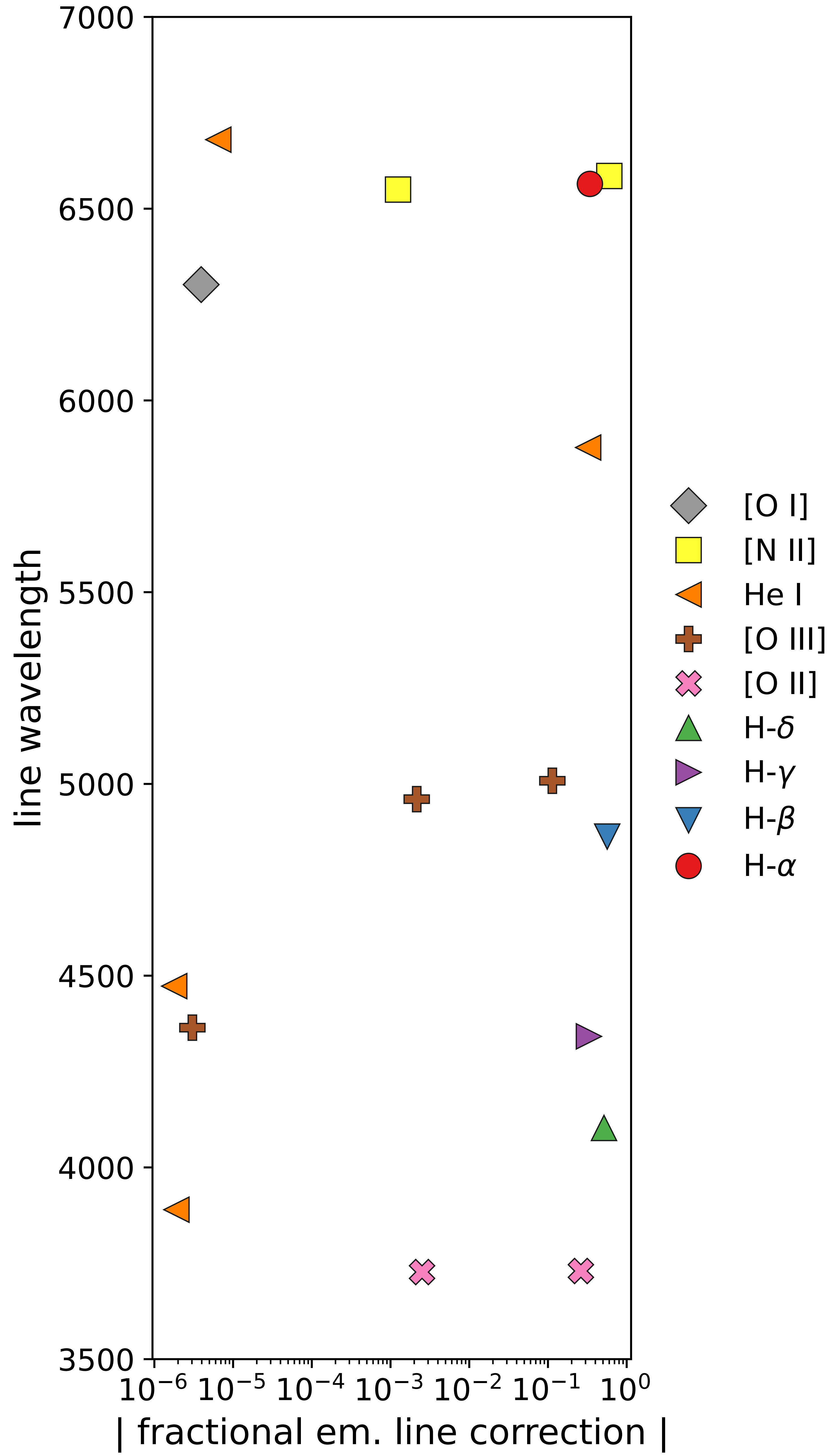
Zero-point corrections

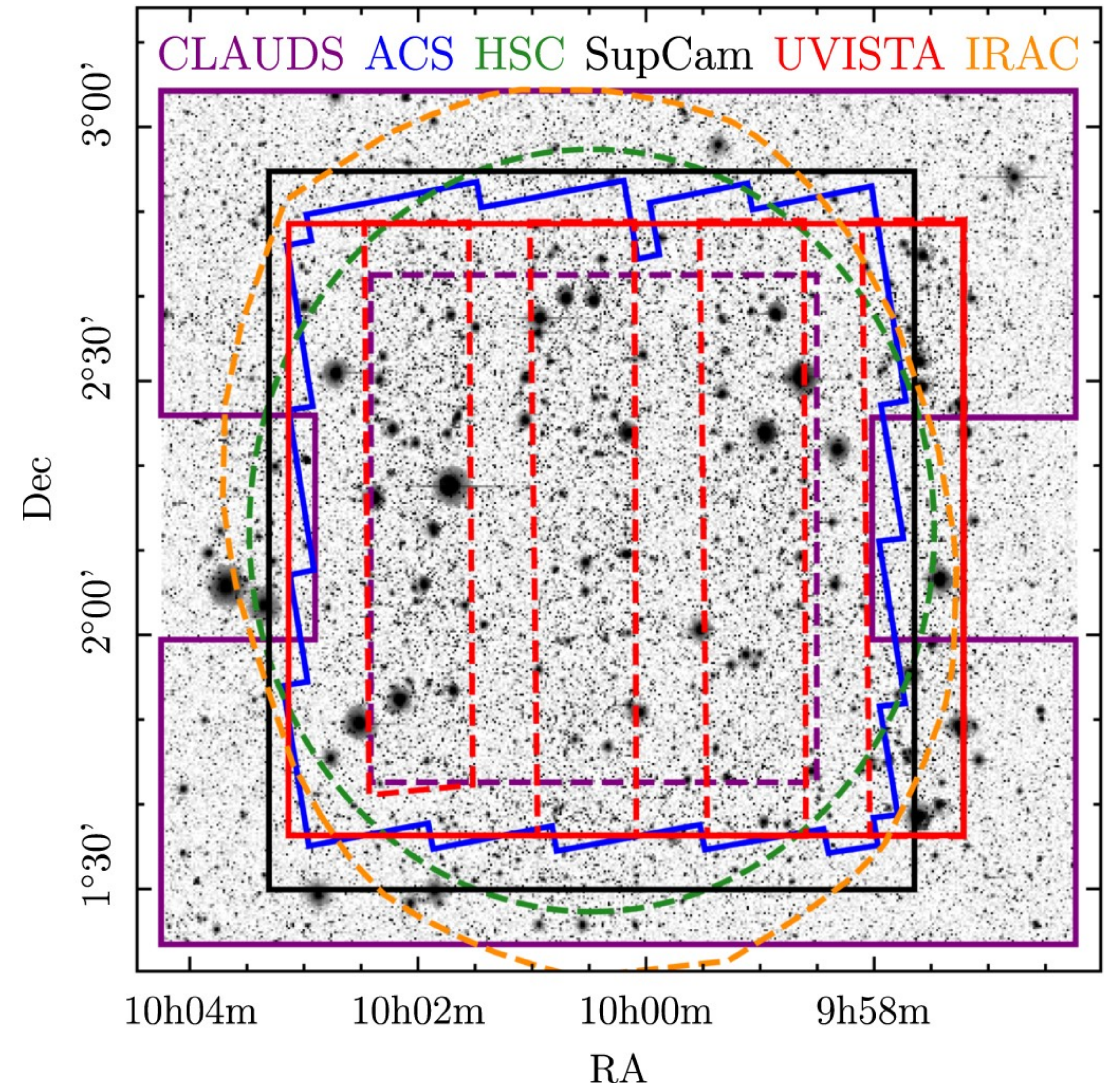


Uncertainty model



Emission line corrections

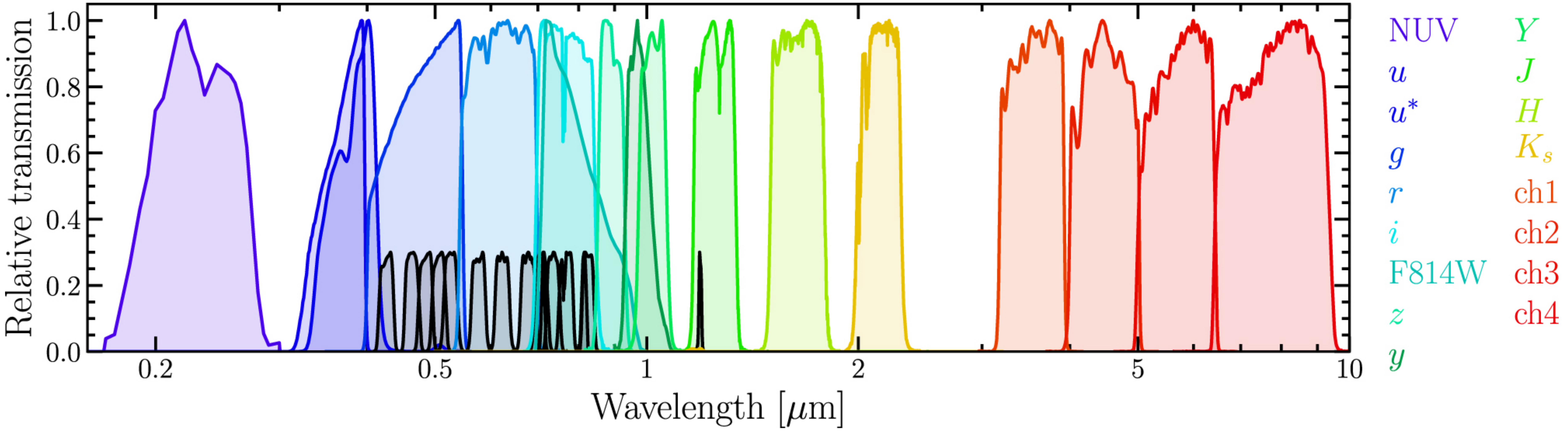




What data will we use?

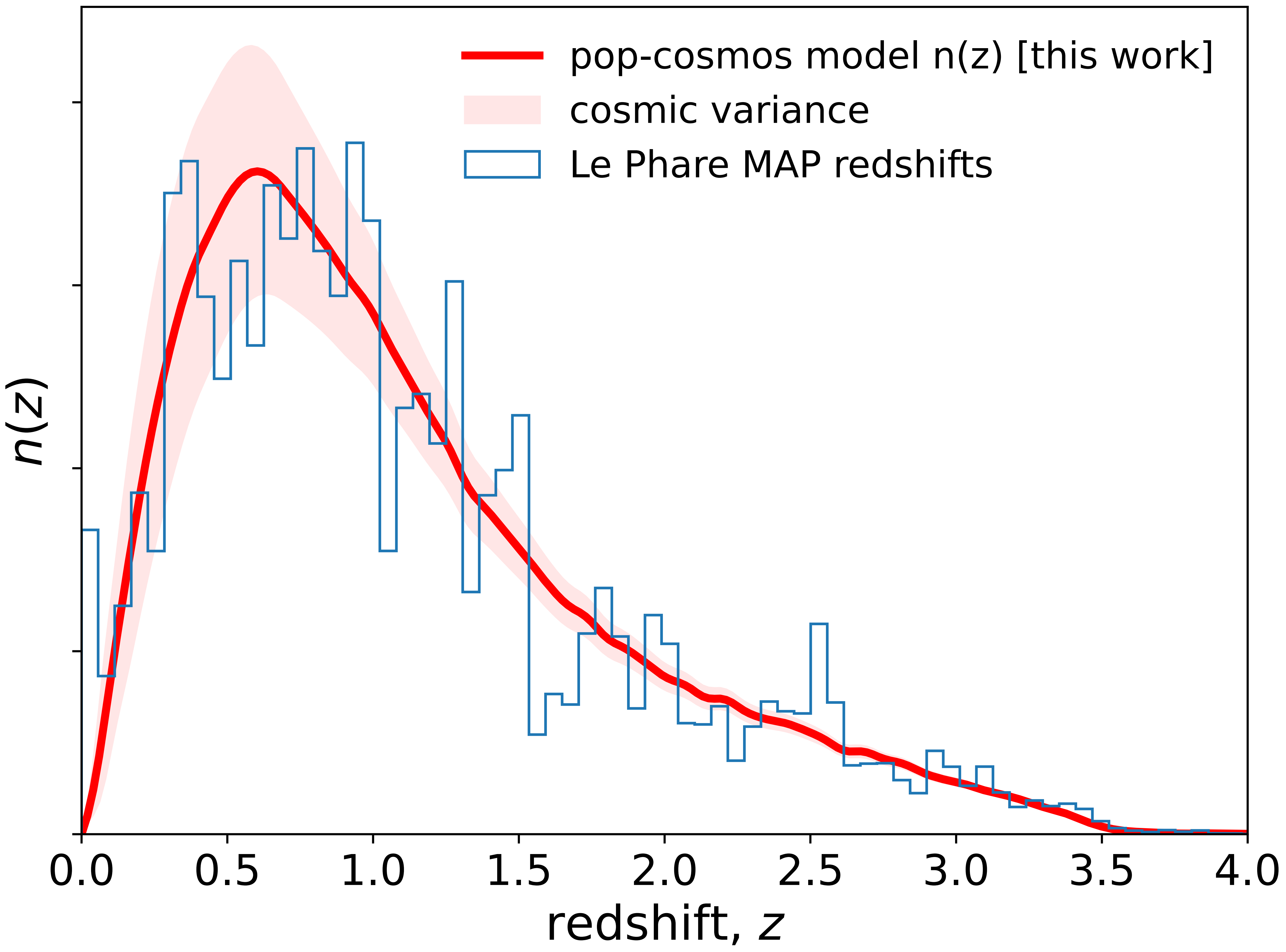
COSMOS2020 (Weaver et al. 2022)

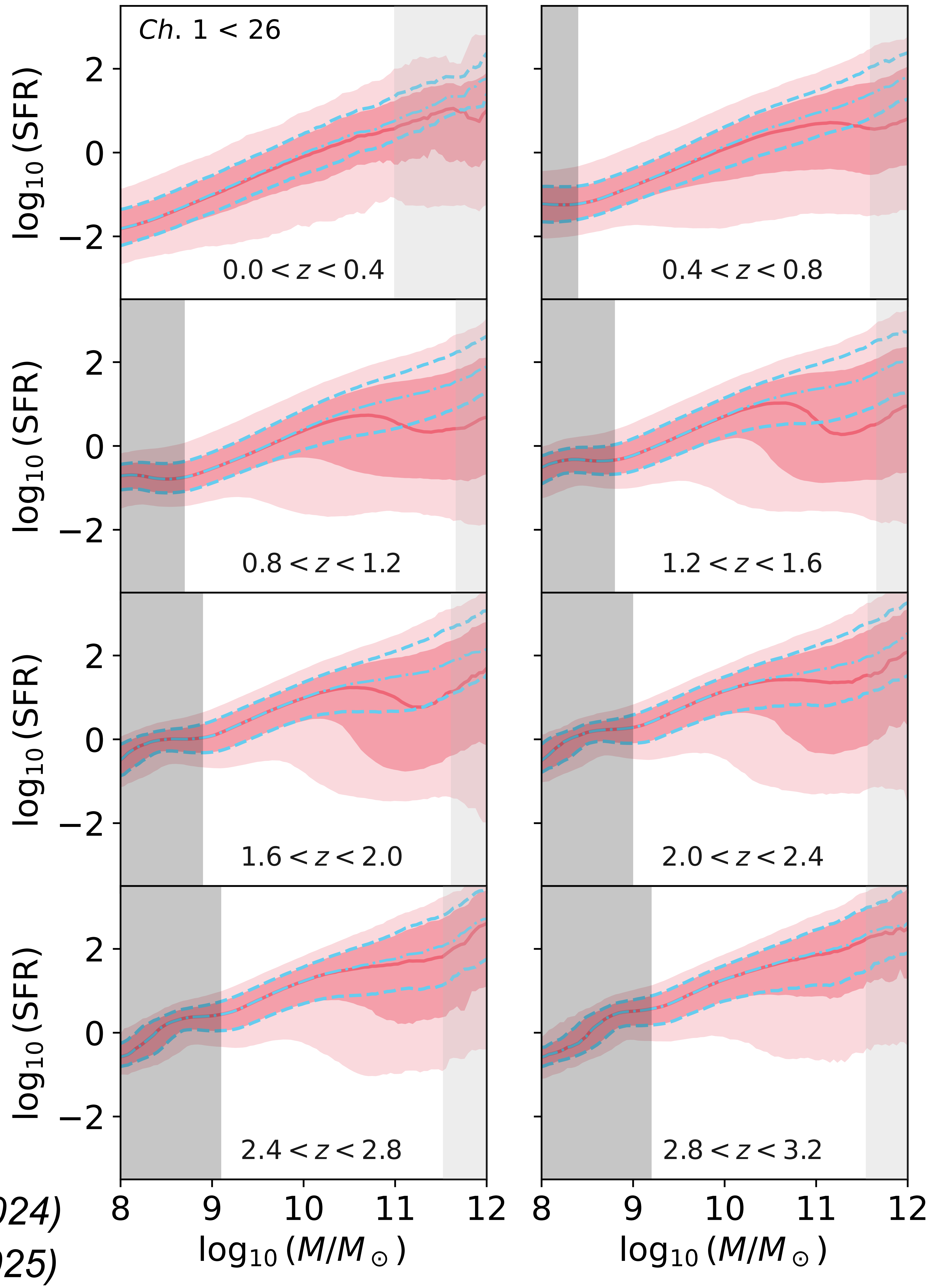
- Has $\approx 420{,}000$ galaxies with *Spitzer* IRAC *Ch. 1* < 26
- Wide and narrow bands
- Coverage from near-UV to mid-IR



Trained Model

Population-level predictions…





Alsing et al. (2024)

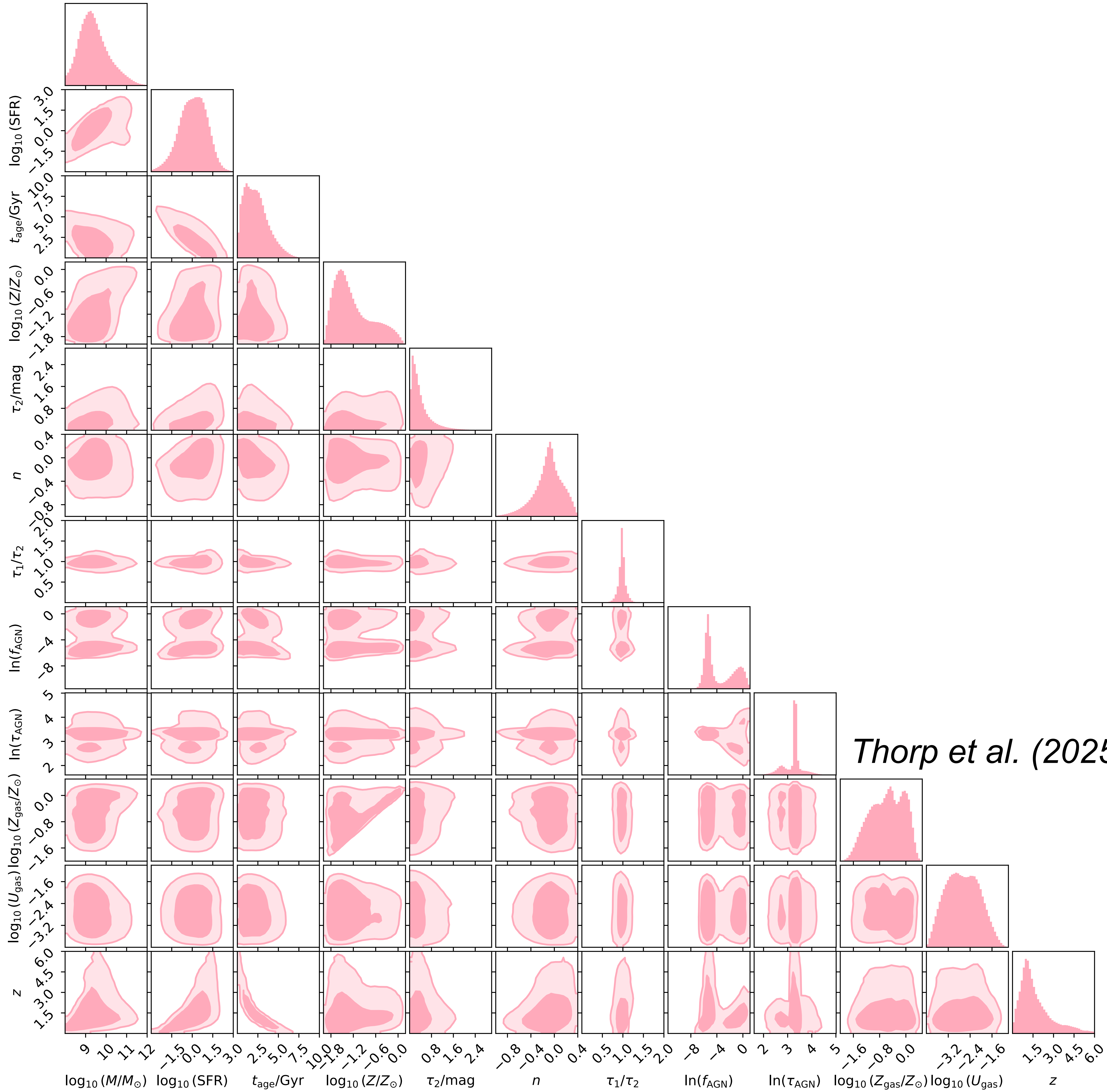
Thorp et al. (2025)

Trained Model

As a prior…

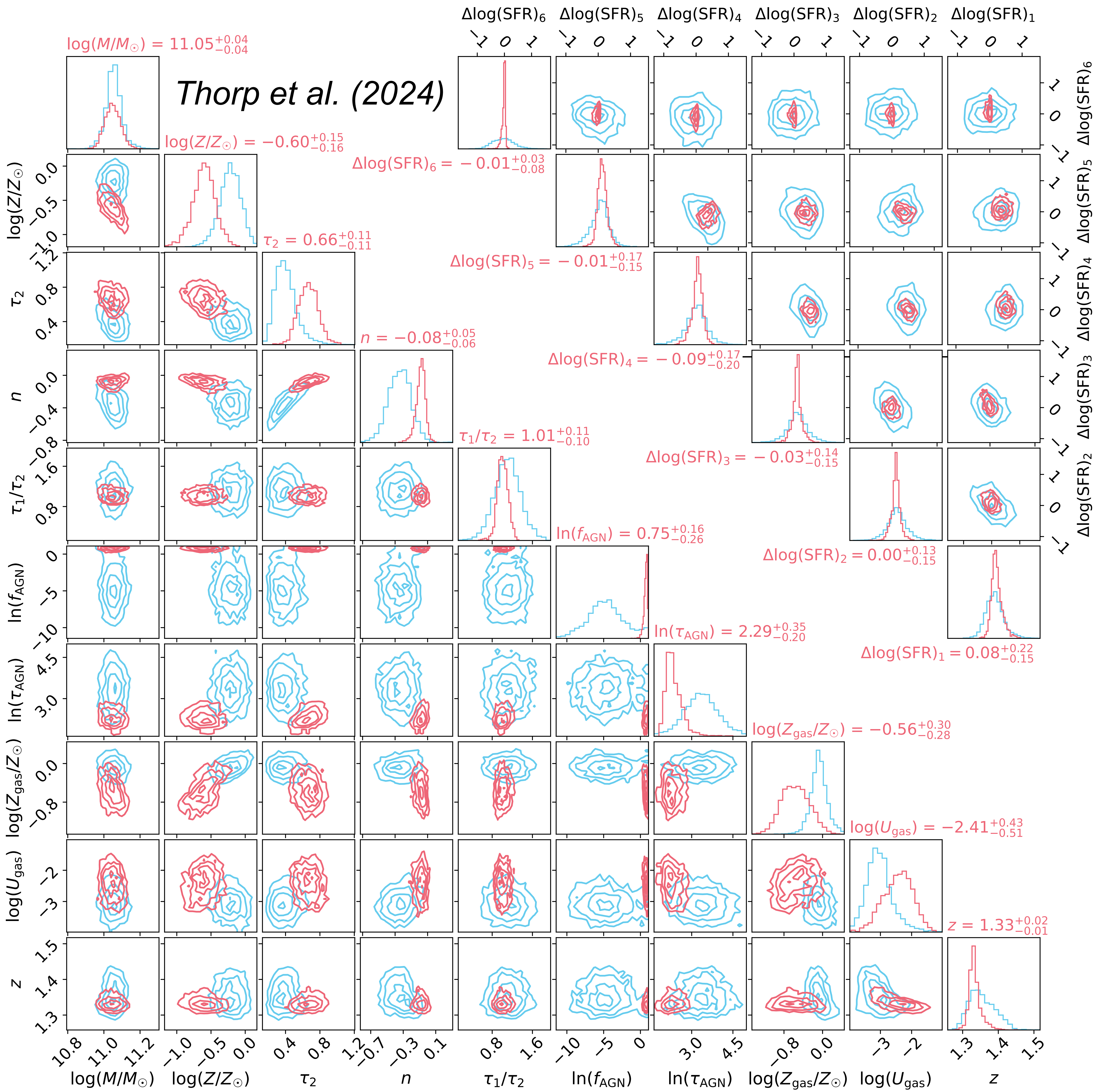
- The learned population distribution over SPS parameters can be used for downstream inference for individual galaxies

- Why? Some science applications need redshift/ parameter inference for *specific galaxies*, not just population-level constraints

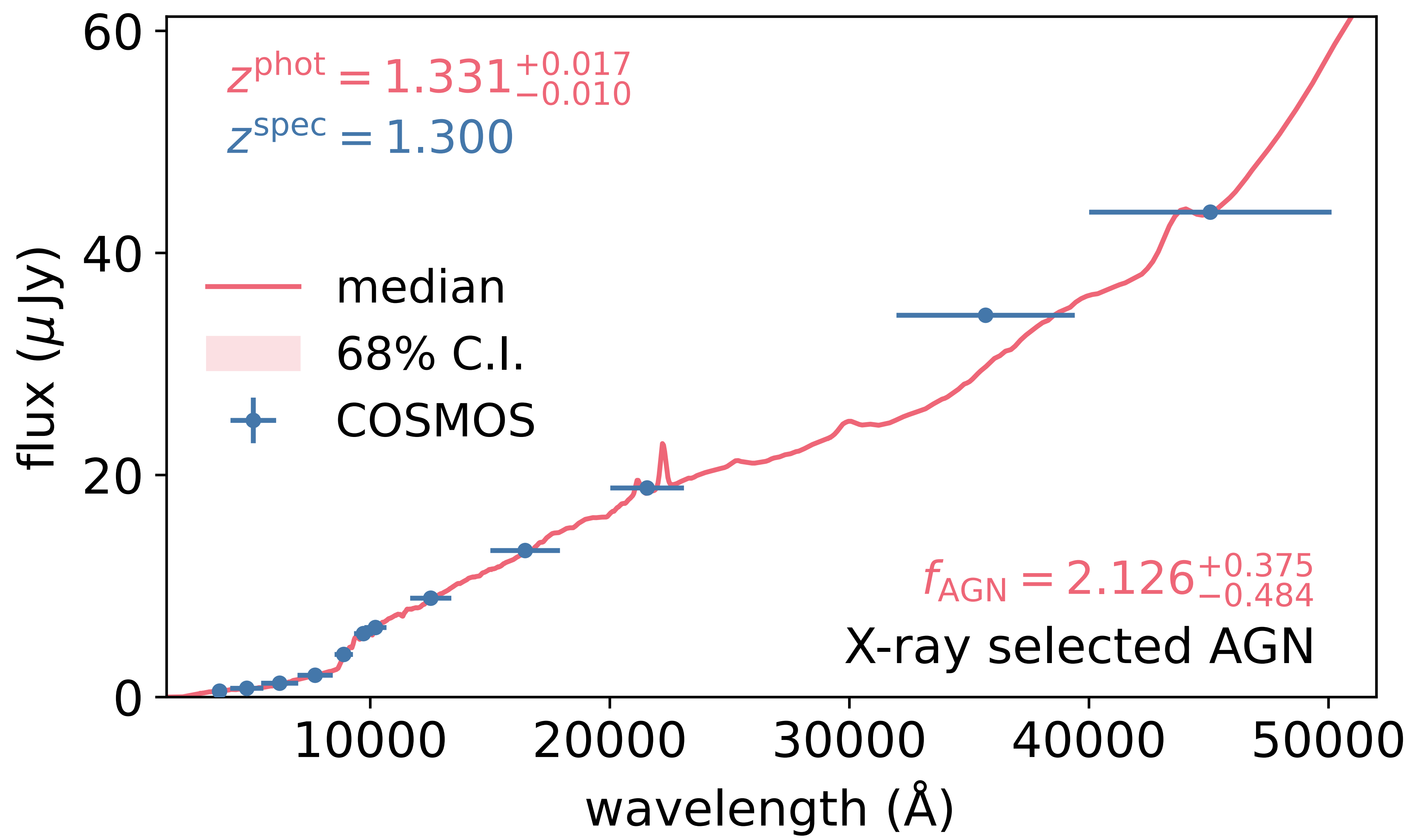


Thorp et al. (2025)

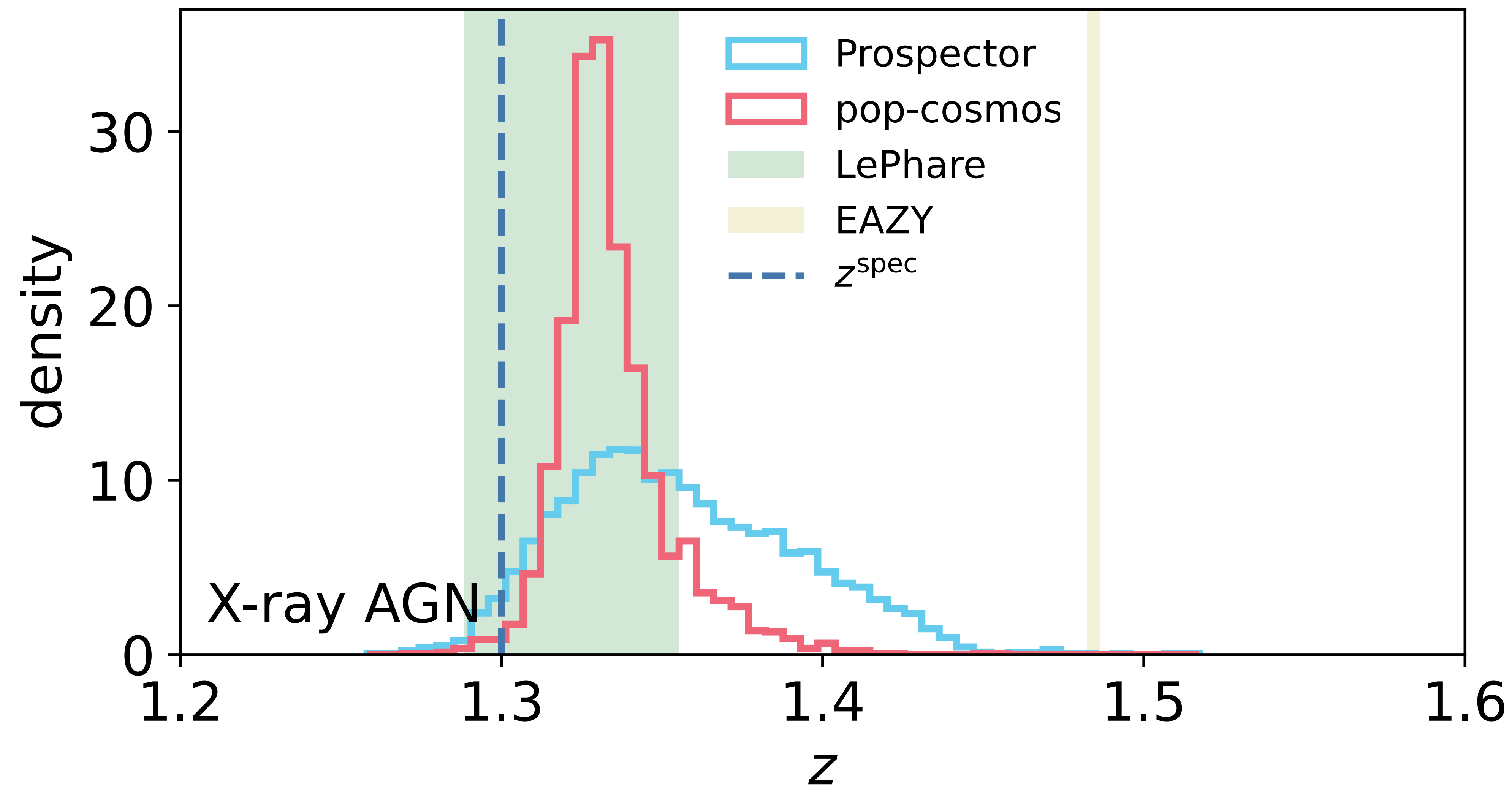
SED Fitting



Thorp et al. (2024)
log(M/M⊙) = 11.05 +0.04 −0.04
log(Z/Z⊙) = −0.60 +0.15 −0.16
Δlog(SFR)6 = −0.01 +0.03 −0.08
τ2 = 0.66 +0.11 −0.11
Δlog(SFR)5 = −0.01 +0.17 −0.15
n = −0.08 +0.05 −0.06
Δlog(SFR)4 = −0.09 +0.17 −0.20
τ1/τ2 = 1.01 +0.11 −0.10
Δlog(SFR)3 = −0.03 +0.14 −0.15
ln(fAGN) = 0.75 +0.16 −0.26
Δlog(SFR)2 = 0.00 +0.13 −0.15
ln(τAGN) = 2.29 +0.35 −0.20
Δlog(SFR)1 = 0.08 +0.22 −0.15
log(Zgas/Z⊙) = −0.56 +0.30 −0.28
log(Ugas) = −2.41 +0.43 −0.51
z = 1.33 +0.02 −0.01
log(M/M⊙)
log(Z/Z⊙)
τ2
n
τ1/τ2
ln(fAGN)
ln(τAGN)
log(Zgas/Z⊙)
log(Ugas)
z
Δlog(SFR)6
Δlog(SFR)5
Δlog(SFR)4
Δlog(SFR)3
Δlog(SFR)2
Δlog(SFR)1



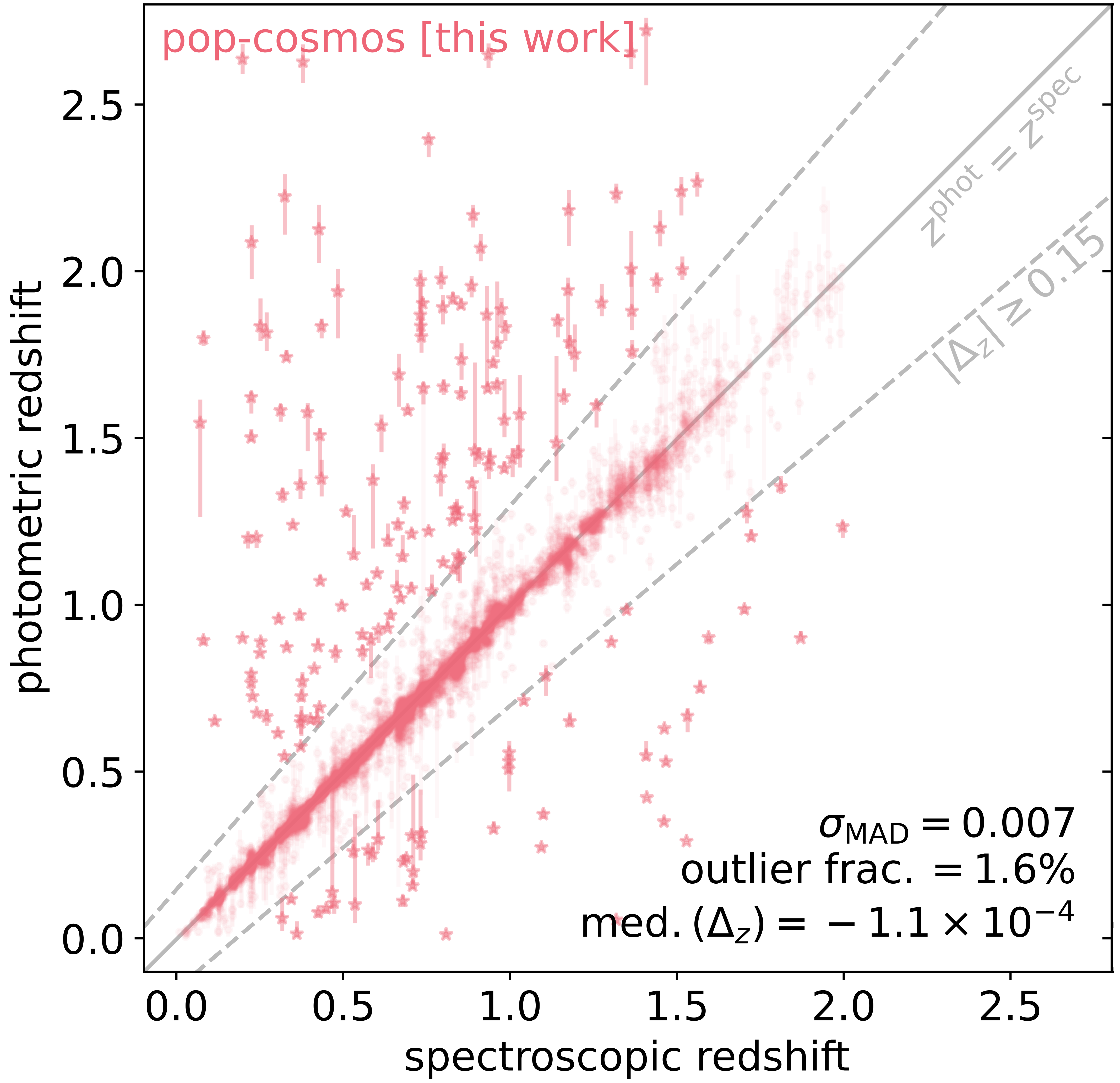
zphot = 1.331 +0.017 −0.010
zspec = 1.300
median
68% C.I.
COSMOS
fAGN = 2.126 +0.375 −0.484
X-ray selected AGN
flux (μJy)
wavelength (Å)
0
20
40
60
10000
20000
30000
40000
50000

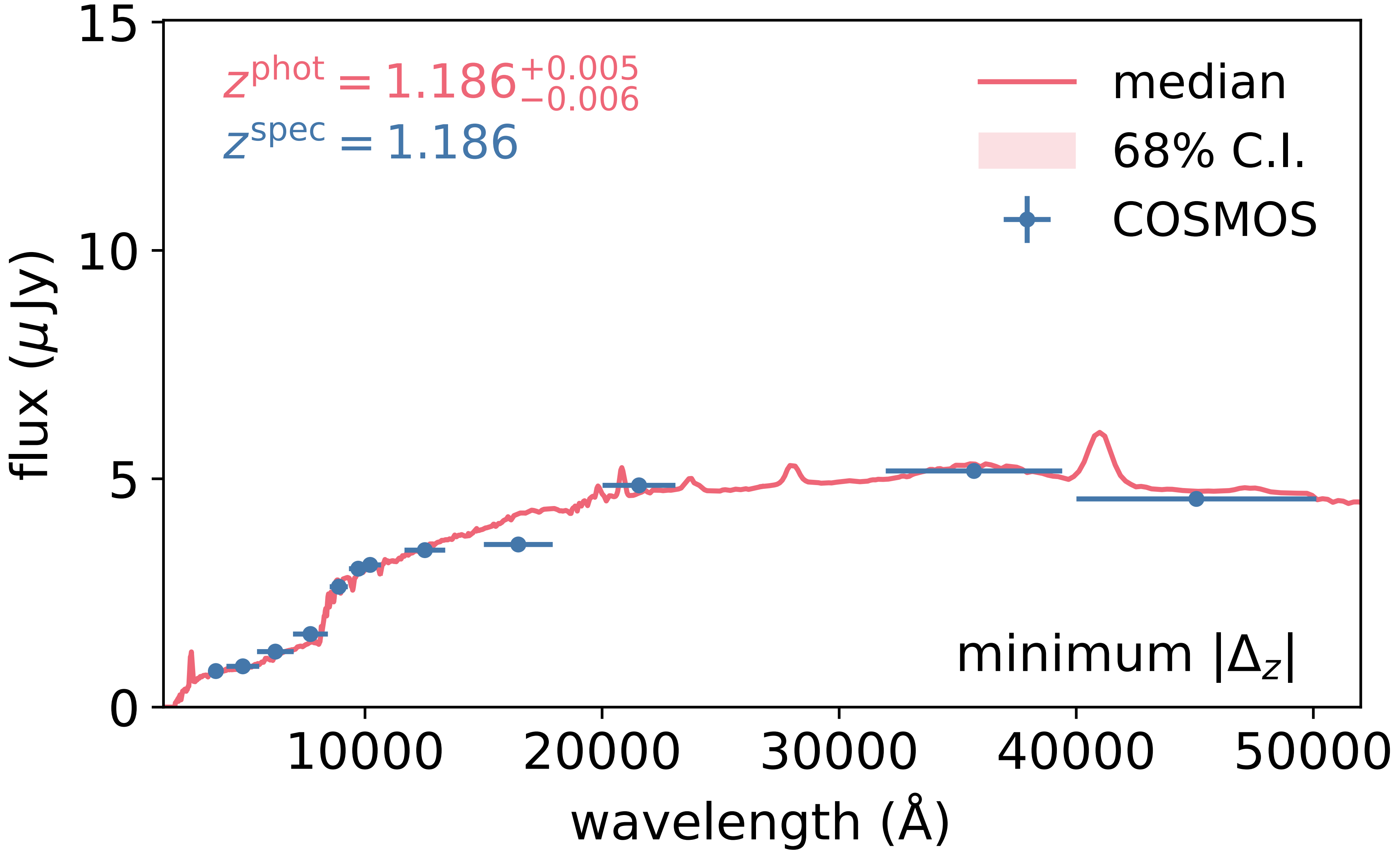


Prospector
pop-cosmos
LePhare
EAZY
zspec
X-ray AGN
density
z
0
10
20
30
1.2
1.3
1.4
1.5
1.6

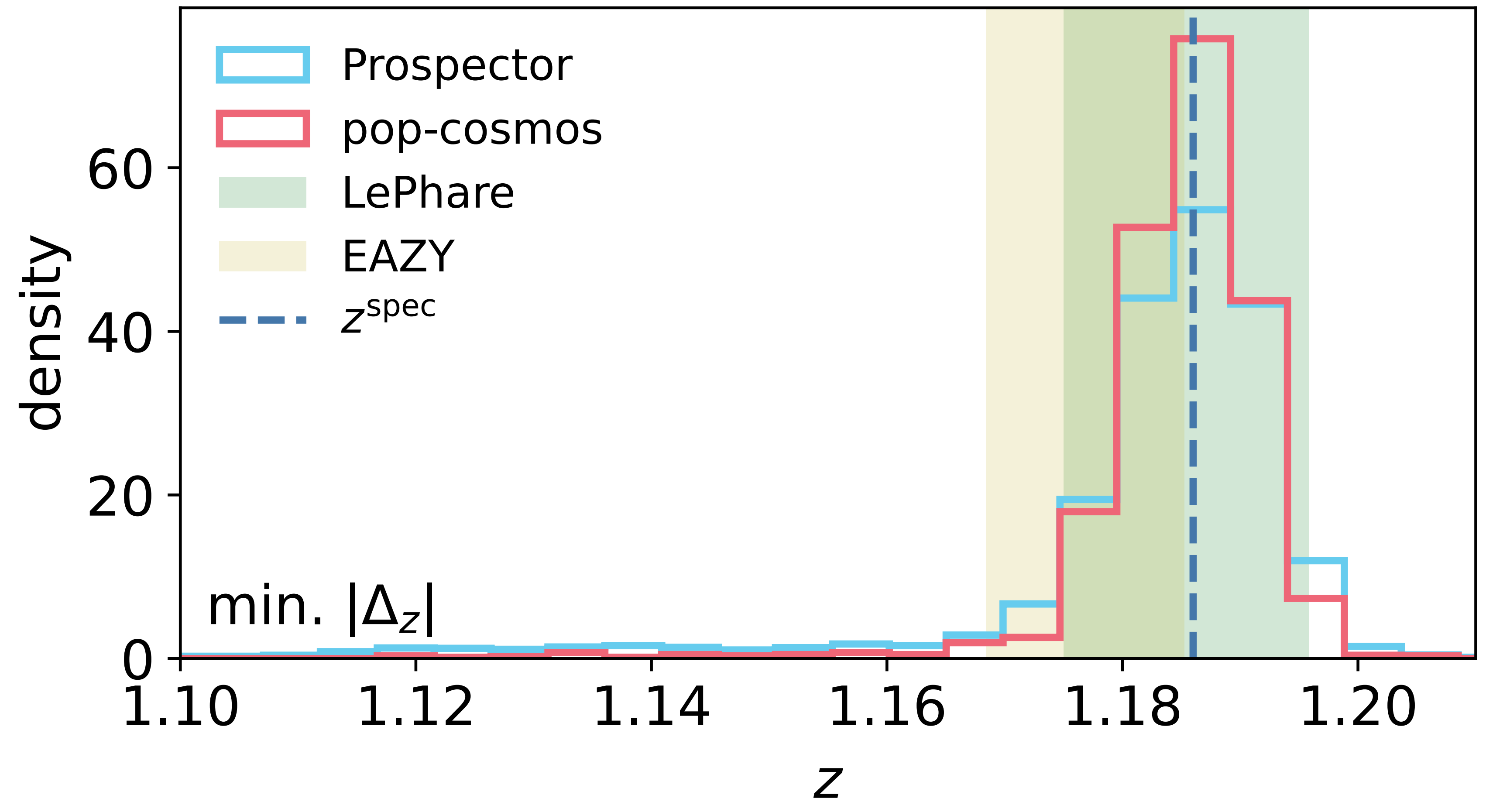
SED Fitting

Redshift inference compared to spectroscopy…



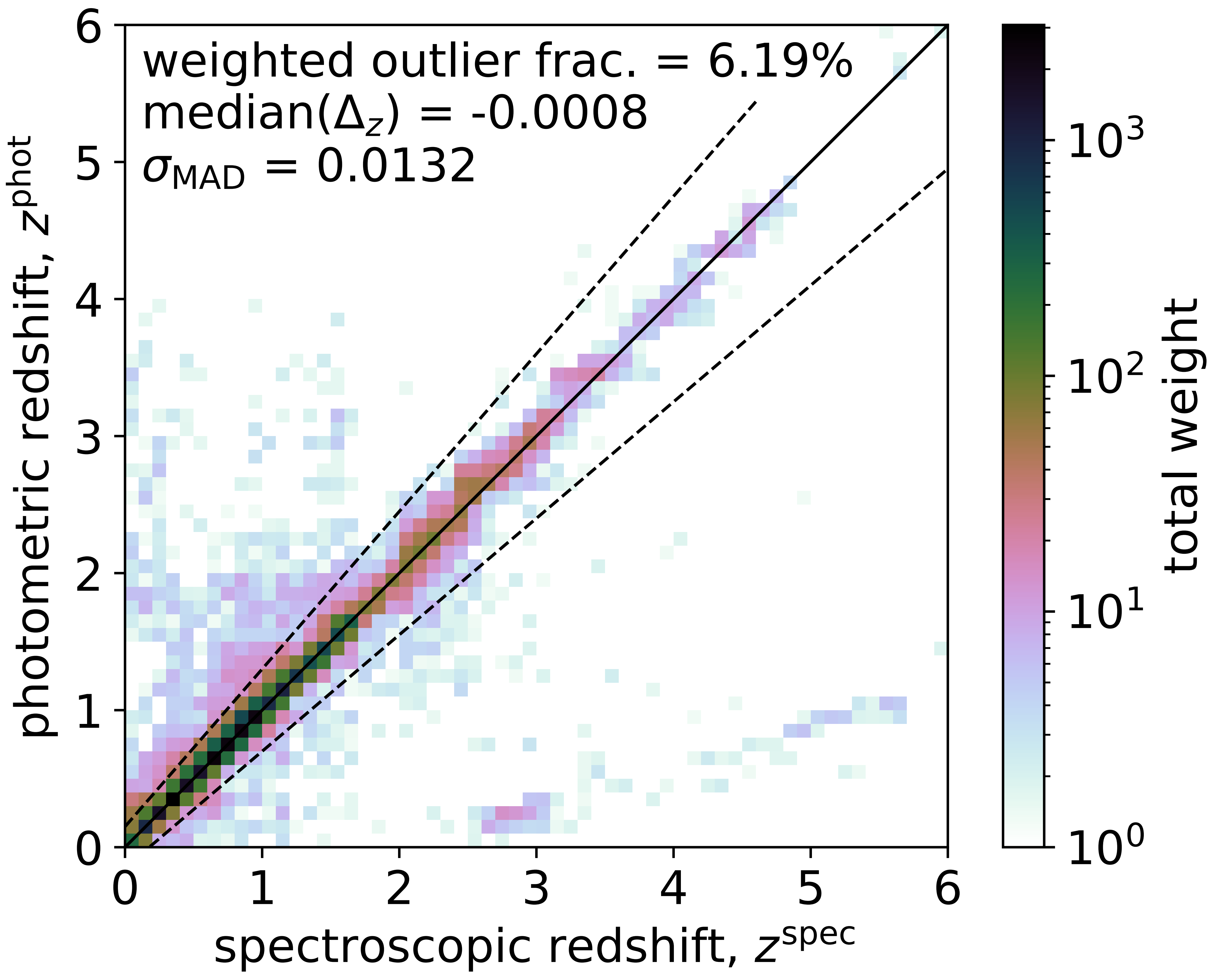


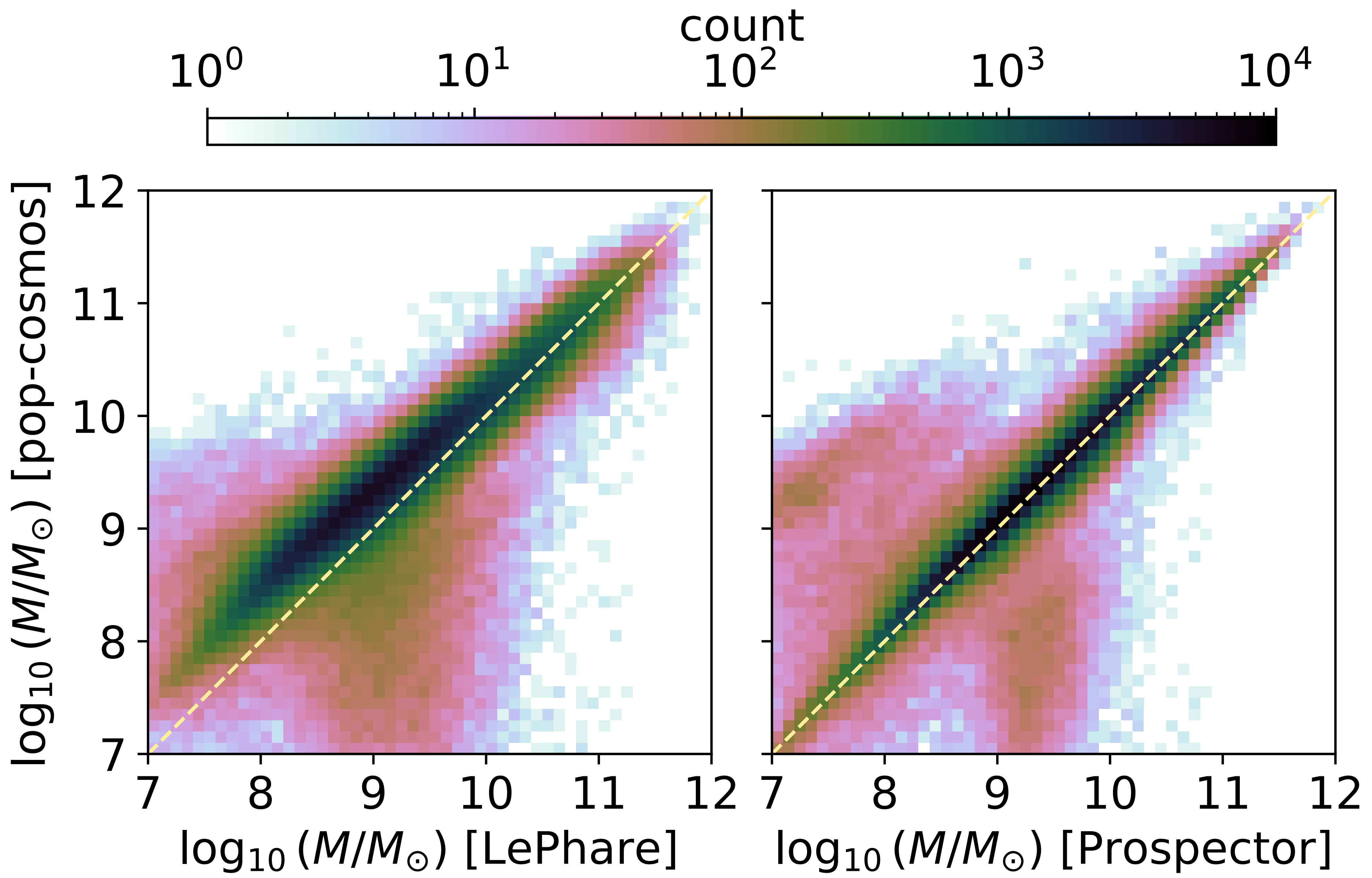
Thorp et al. (2024)



SED Fitting

Redshift and stellar mass inference…



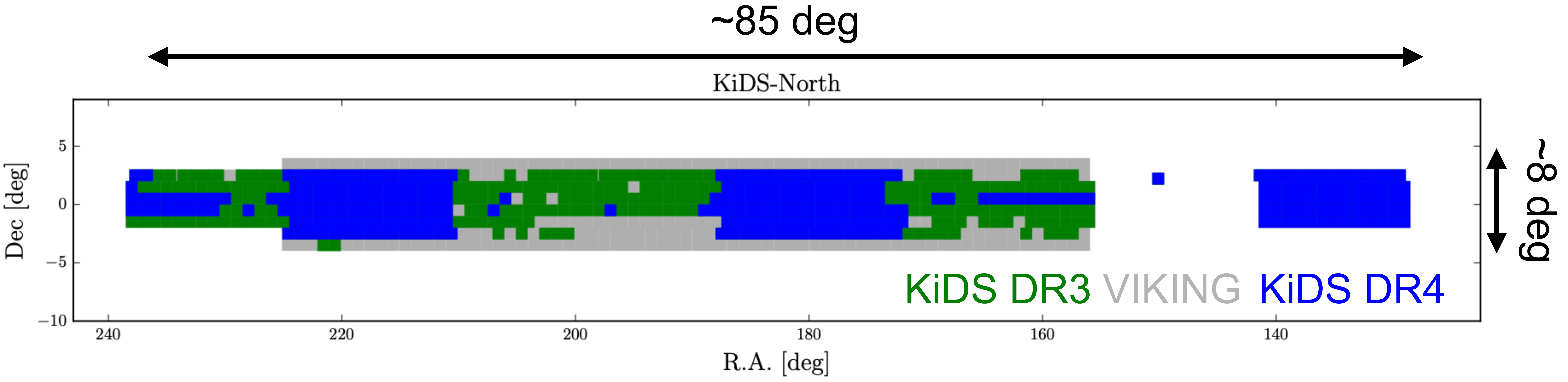


Thorp et al. (2025)

Transferring to other surveys?

Kilo Degree Survey (KiDS; Kuijken et al. 2019)

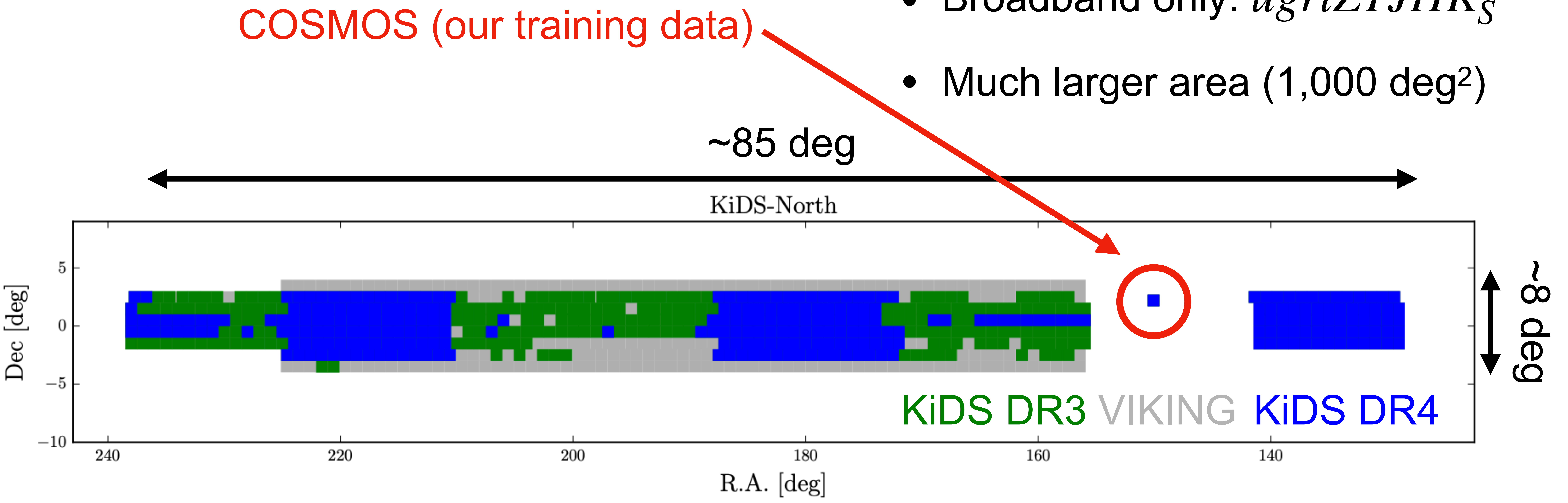
- Has $\approx 70{,}000{,}000$ galaxies
- Broadband only: $ugriZYJHK_S$
- Much larger area (1,000 deg^2)



Transferring to other surveys?

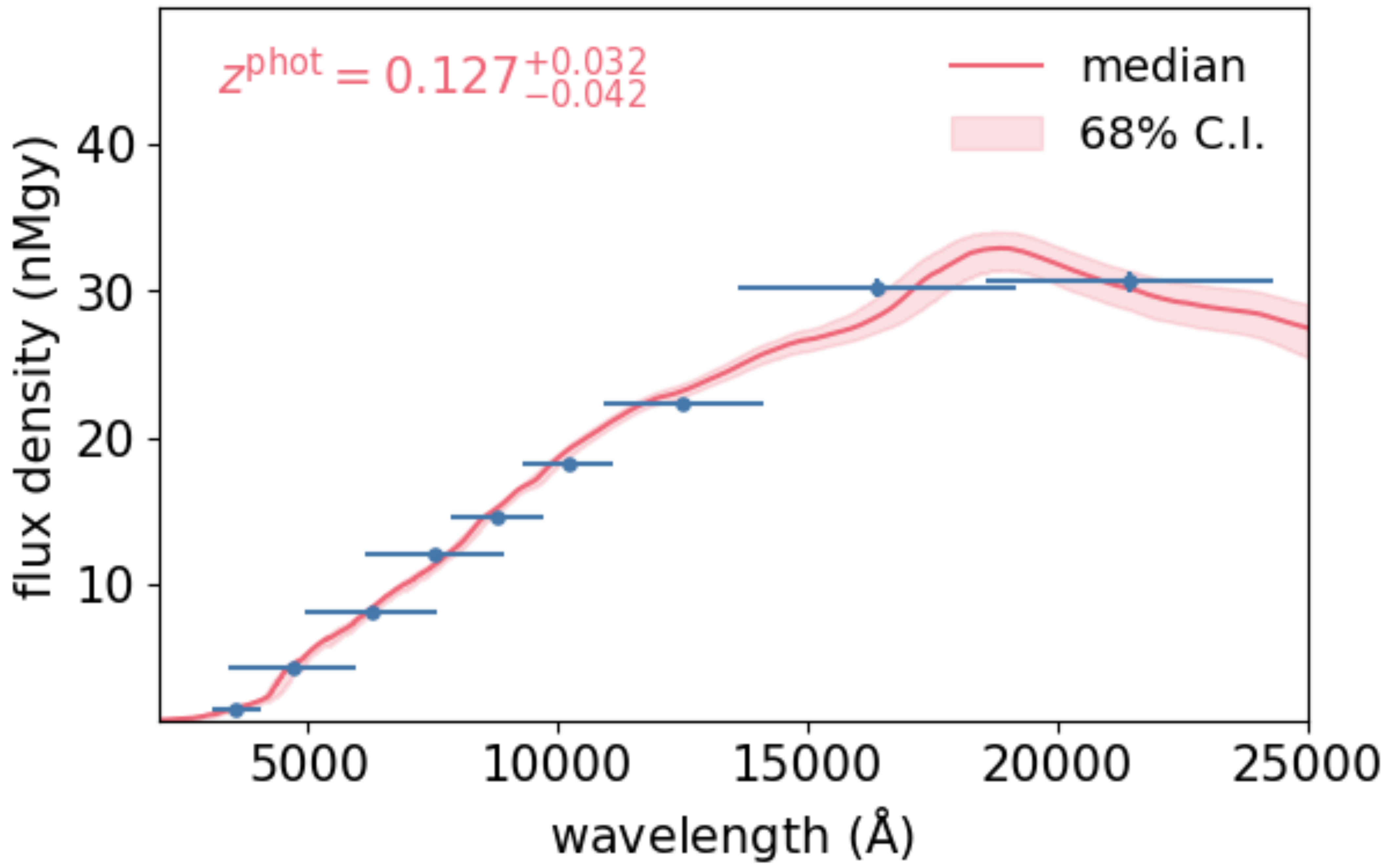
Kilo Degree Survey (KiDS; Kuijken et al. 2019)

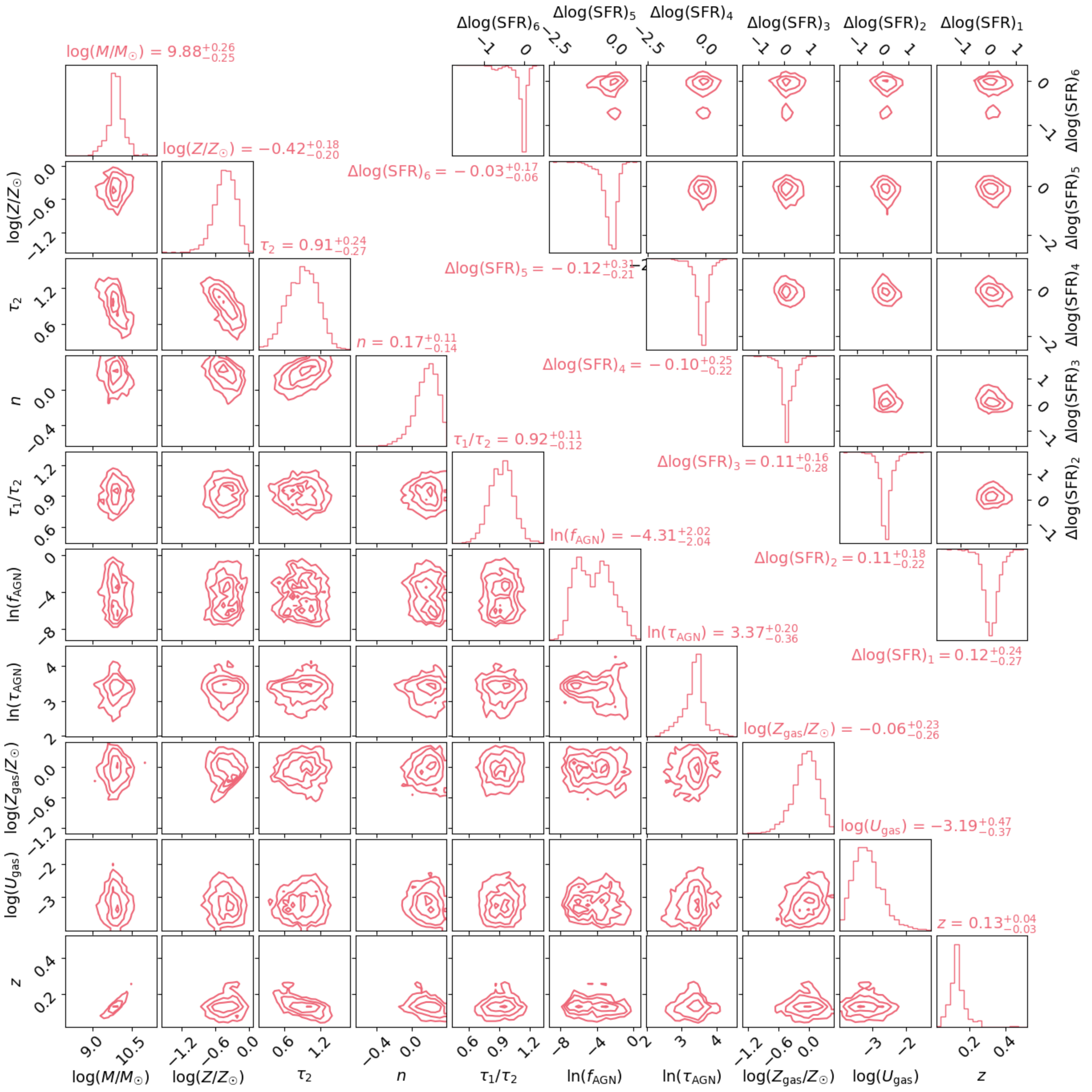
- Has $\approx 70{,}000{,}000$ galaxies
- Broadband only: $ugriZYJHK_S$
- Much larger area (1,000 deg^2)



SED Fitting

Kilo Degree Survey data…

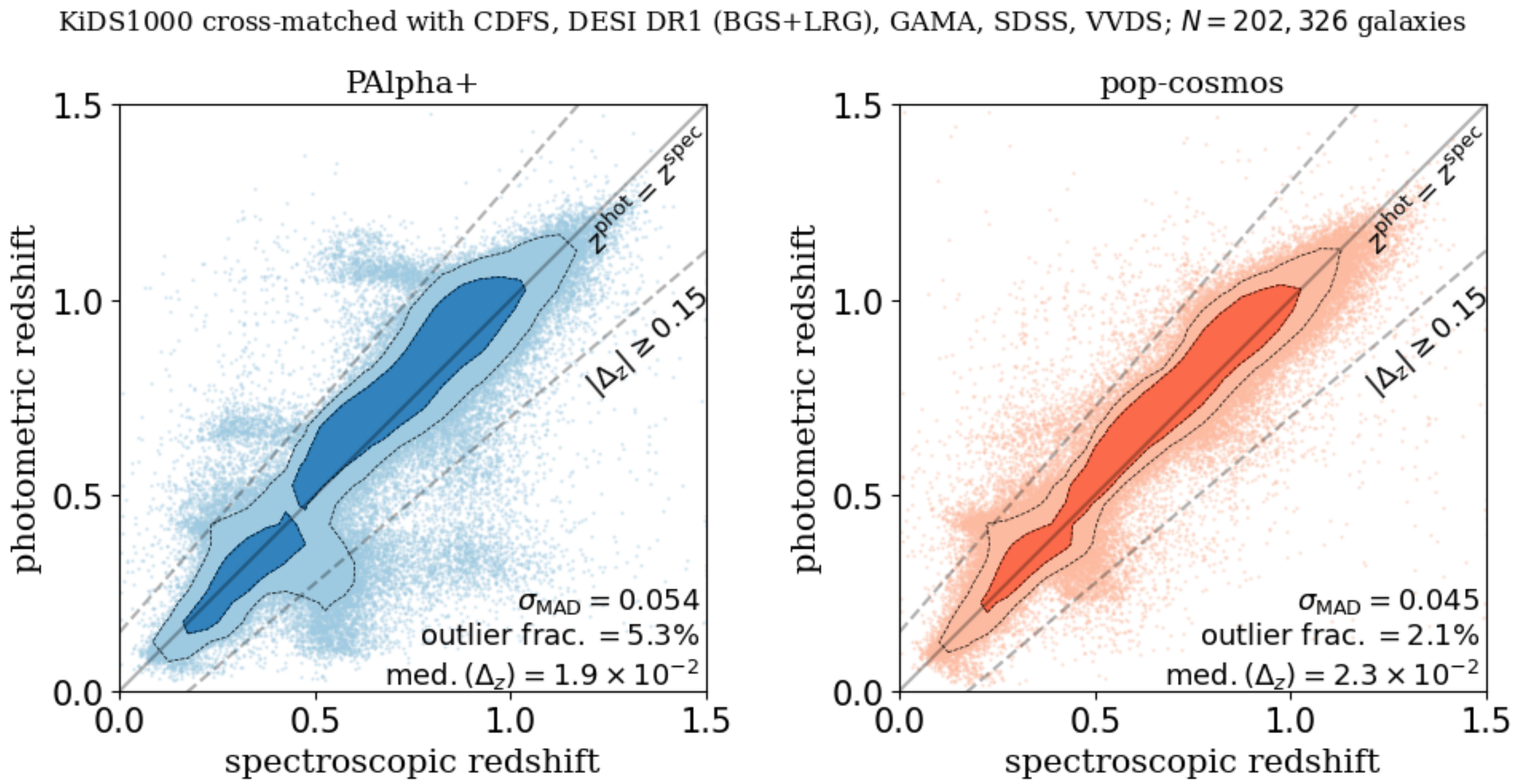




Halder et al. (in prep.)

SED Fitting

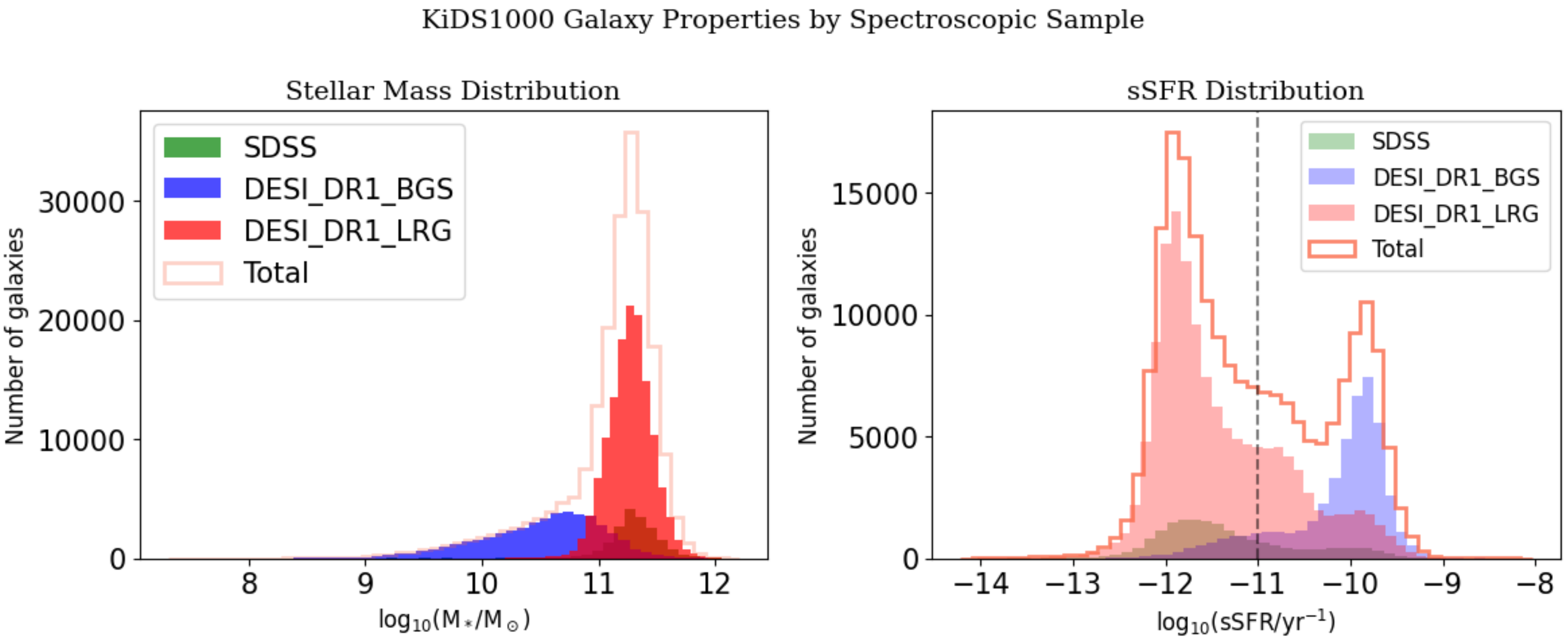
Kilo Degree Survey data…



Halder et al. (in prep.)

SED Fitting

Kilo Degree Survey data…



Halder et al. (in prep.)

Scaleability

- Currently using emulated SPS + diffusion prior + affine-invariant sampler
- Gets full MCMC chains with throughput of ~7 GPU-sec per galaxy (or ~0.7 GPU-sec per galaxy with a simpler analytic prior)
- Good enough for Stage III surveys; could be faster for Stage IV
- Bottleneck is flow/diffusion model log probability; room for improvement there, e.g. distillation/consistency models, ANPE, bespoke solvers (a cool idea from Luigi's highlight talk on Monday!)

Summary

- Data-driven prior over galaxy physical parameters, using a score-based diffusion model
- Train on the deepest, highest-fidelity data that we have
- Use the model as a prior over physical parameters when doing inference for new surveys
- Scaling to ~few millions of galaxies possible with emulation, parallelised MCMC, GPUs
- Bigger data with ANPE, distillation, bespoke solvers?

arXiv 2506.12122