Initiative for the Theoretical Sciences, The CUNY Graduate Center

Adventures in the Theoretical Sciences
An informal, online summer school.

June 4 and 5
Few-body and many-body chaos
Vladimir Rosenhaus

June 11 and 12
Multi-parameter models and information geometry
Katherine Quinn

June 25 and 26
Big universe, big data: Emerging challenges in astrophysics
Viviana Acquaviva

June 18 and 19
Statistical physics and machine learning
David J. Schwab

July 2 and 3
Precision and emergence in the physics of life
William Bialek

July 9 and 10
Driven quantum systems
Vadim Oganesyan

Six lecturers will present four hours of lecture and discussion each, touching a wide range of topics. Our goal is to introduce students to the excitement of our fields, and to encourage thinking about theory as a unifying activity. We expect students to have solid backgrounds in statistical physics and quantum mechanics; more specialized topics will be introduced as needed. Our target audience overlaps advanced undergraduates and beginning graduate students in the US, and MSc students abroad.

Registration: itsatcuny.org/summerschool

Lectures from 11am-1pm (EDT)

Questions: info@itsatcuny.org

BIG UNIVERSE, BIG DATA: EMERGING CHALLENGES IN ASTROPHYSICS, PART 2

Viviana Acquaviva

CUNY / Universitat de Barcelona

vacquaviva@citytech.cuny.edu

ABOUT THE SESSION

- Feel free to use video or not
- For questions: please use the "Chat" (not private chat, not raise your hand) option
- I'll review and answer them, if I can, when I can
- To encourage participation so I'll try to answer questions from different people before multiple questions from the same person
- Approx. schedule:
  - Questions 11.05 -11.20
  - Bla bla from me 11.20-11.55
  - Break 11.55-12.05
  - More bla bla from me12.05-12.30
  - Q/A 12.30-1 (feel free to ask about more general topics)

QUESTIONS FROM YESTERDAY

- 09:35:02 From Yannis(NYC) : At what decimal point would we need the accuracy of our distances to be able to detect the change in distances in galaxies within let's say an exposure time of 10 years?
- 10:54:00 From Amirmohammad Chegeni : Can you show some cases that machine predicted wrong?
- 10:57:18 From John Vastola : For some of the galaxy-related tasks you use ML for, how long does it usually take to code up a reasonable prototype?
- 10:57:27 From John Vastola : Also, what resources would you recommend for a beginner?
- 10:57:54 From Irsad Tio Majid : what methods can be used if there so much noise who brings the number become blur/cant be see clearly?
- 10:59:14 From Leonid Pomirchi : it's supposed to be antigravity between particle and antiparticles. May it be explanation of the universe expanding?
- 11:01:38 From Marya : Could you introduce some references on applications and programing tutorials of machine learning in astrophysics.

RESOURCES

GENERAL ML (many, but these I can personally recommend!)

- https://www.coursera.org/learn/machine-learning
- https://www.coursera.org/learn/intro-to-deep-learning

PHYSICS ML:

- https://arxiv.org/abs/1803.08823 (has notebooks!)
- https://physics.bu.edu/~pankajm/MLnotebooks.html
- https://arxiv.org/abs/1903.10563

ASTRO ML:

- https://www.astroml.org/
- https://github.com/vacquaviva/MLSummerSchool_CCA19 (or anything public in my GitHub)
- Also, keep an eye out for my book 'ML for Physics & Astronomy' ;) (in 2022?)

SQUEEZING INFORMATION ABOUT GALAXIES USING BAYESIAN INFERENCE AND ML

WHAT CAN WE LEARN ABOUT GALAXIES?

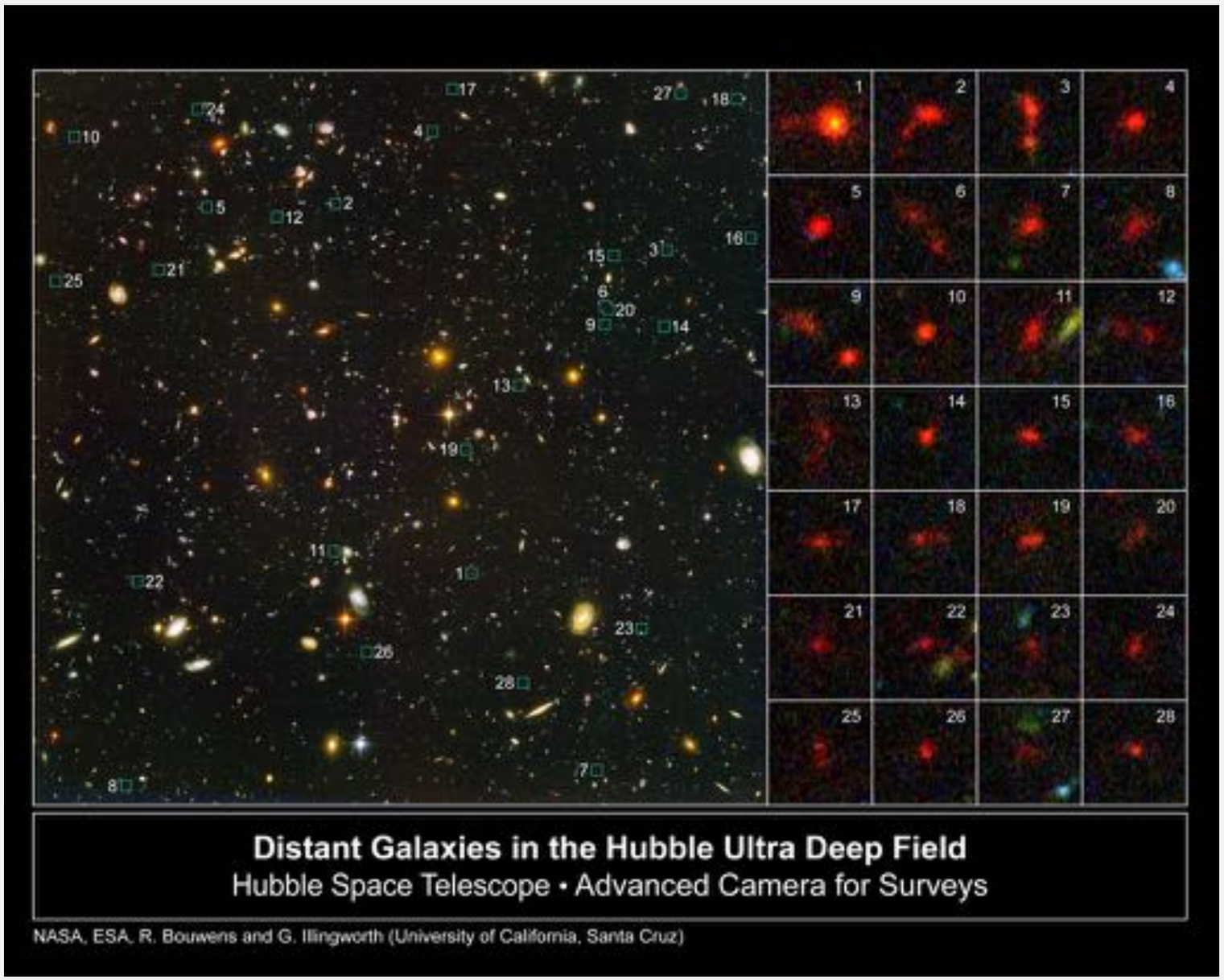
How far away is the galaxy?
How many stars (of each type) are in the galaxy?
How did the galaxy assemble its stars?
What is the galaxy's chemical composition?
What is the role (composition, geometry) of cosmic dust (the dark stuff you see in the first pic)?
How often galaxy collisions/mergers happen?
How do all these change through cosmic time?

HOW CAN WE LEARN ABOUT (DISTANT) GALAXIES?

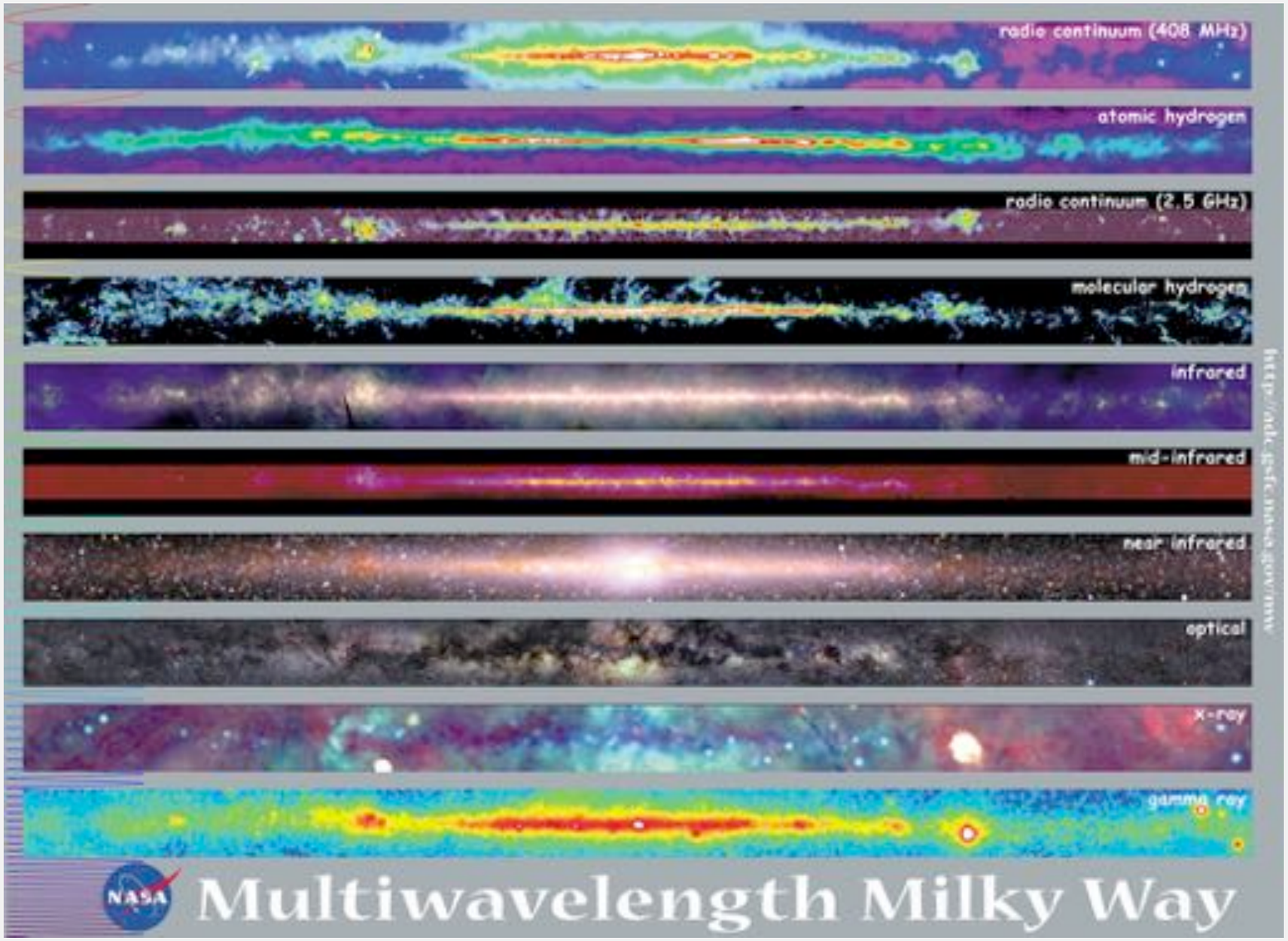
PROBLEM:

If I say "galaxy" I bet you think of something like this…

BUT IN REALITY THIS IS WHAT THEY LOOK LIKE (AND THIS IS WITH THE BEST TELESCOPE WE HAVE).

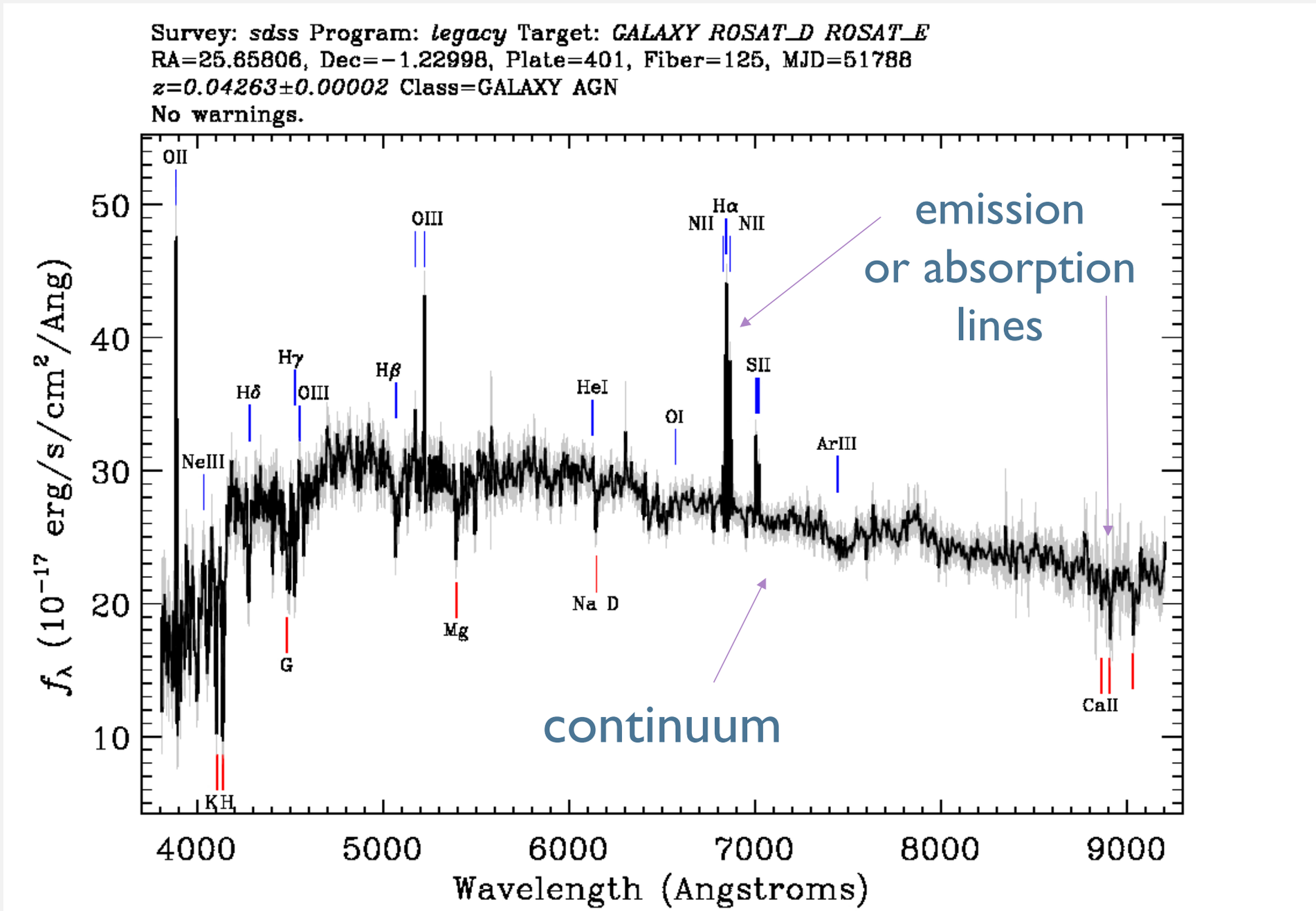


ONE THING WE CAN DO TO COLLECT MORE INFORMATION IS TO LOOK AT THEM IN DIFFERENT WAVELENGTHS.

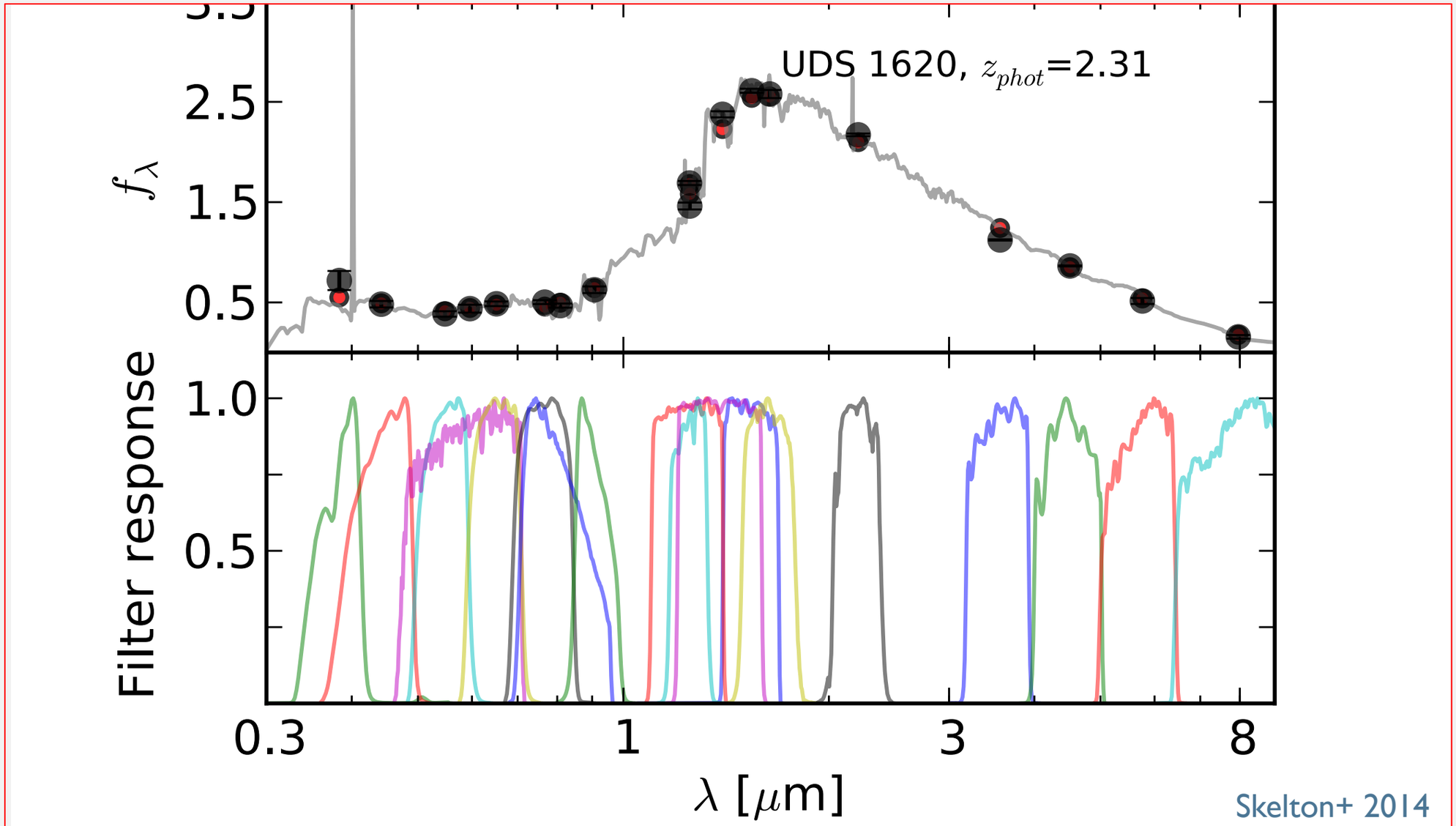


THE MILKY WAY LOOKS BEAUTIFUL AT MANY WAVELENGTHS.

A CHART OF THE LUMINOSITY OF A GALAXY AS A FUNCTION OF WAVELENGTH IS CALLED A SPECTRUM.

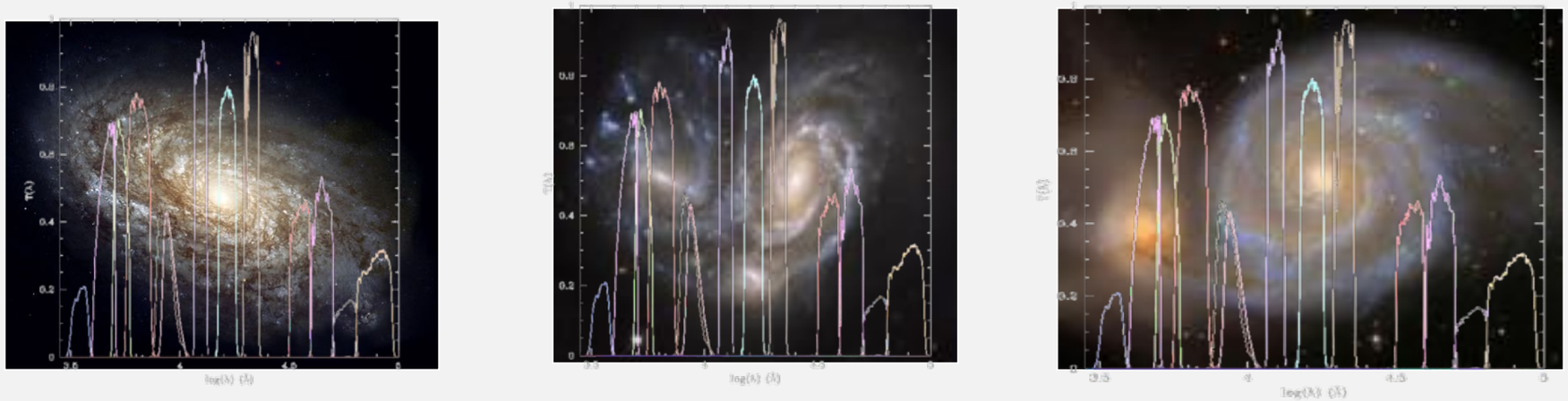


IN FACT, MANY TIMES WE DON'T HAVE A SPECTRUM, BUT A MUCH COARSER SAMPLING OF THE BRIGHTNESS, CALLED PHOTOMETRY.



Skelton+ 2014

SO IN SUMMARY….



WE HAVE: emission chart at different wavelengths, with high (spectra) or low (imaging) resolution

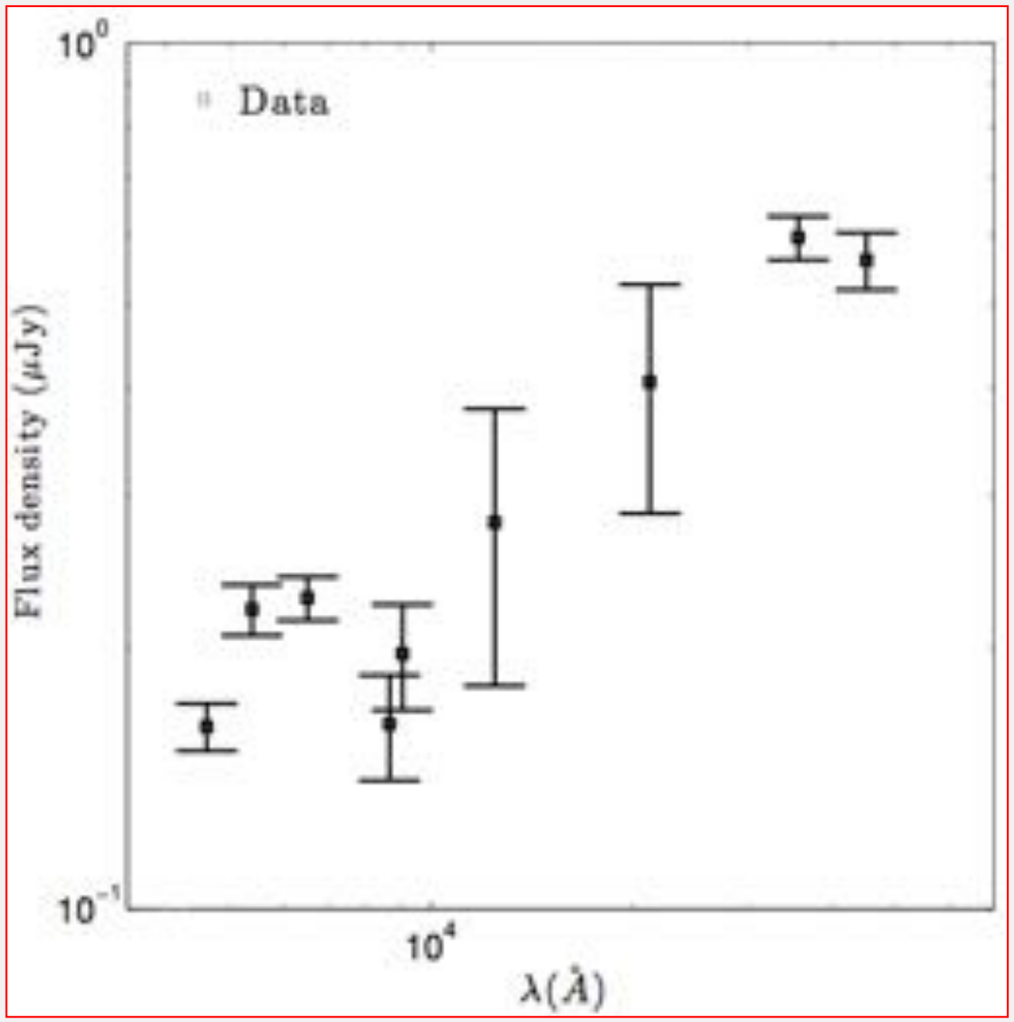
WE WANT: PHYSICAL PROPERTIES!
Stellar Mass, Star Formation History, Dust content, Chemical Enrichment History, Redshift

For a long time, the only player in town was….

SPECTRAL ENERGY DISTRIBUTION FITTING

Want:
Physical properties

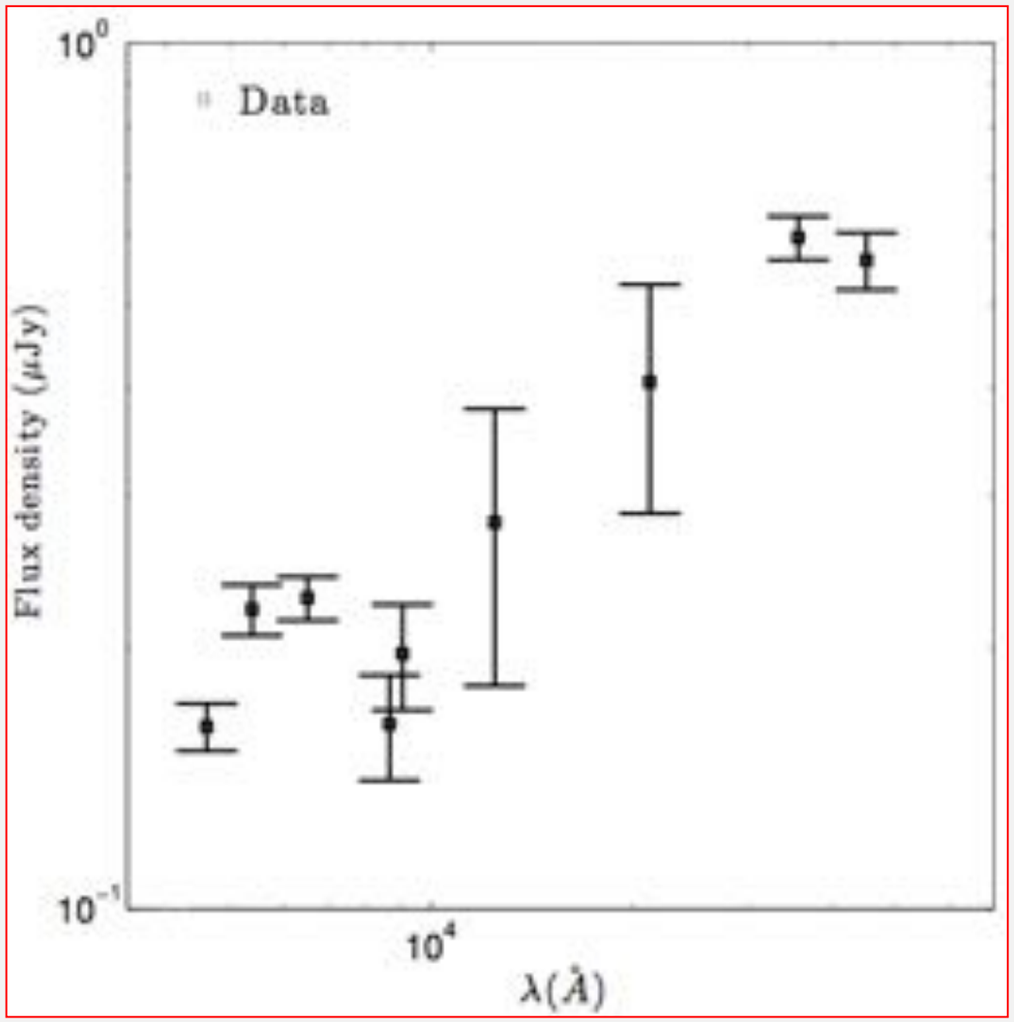
Have:
Fluxes



SPECTRAL ENERGY DISTRIBUTION FITTING

Want:
Physical properties

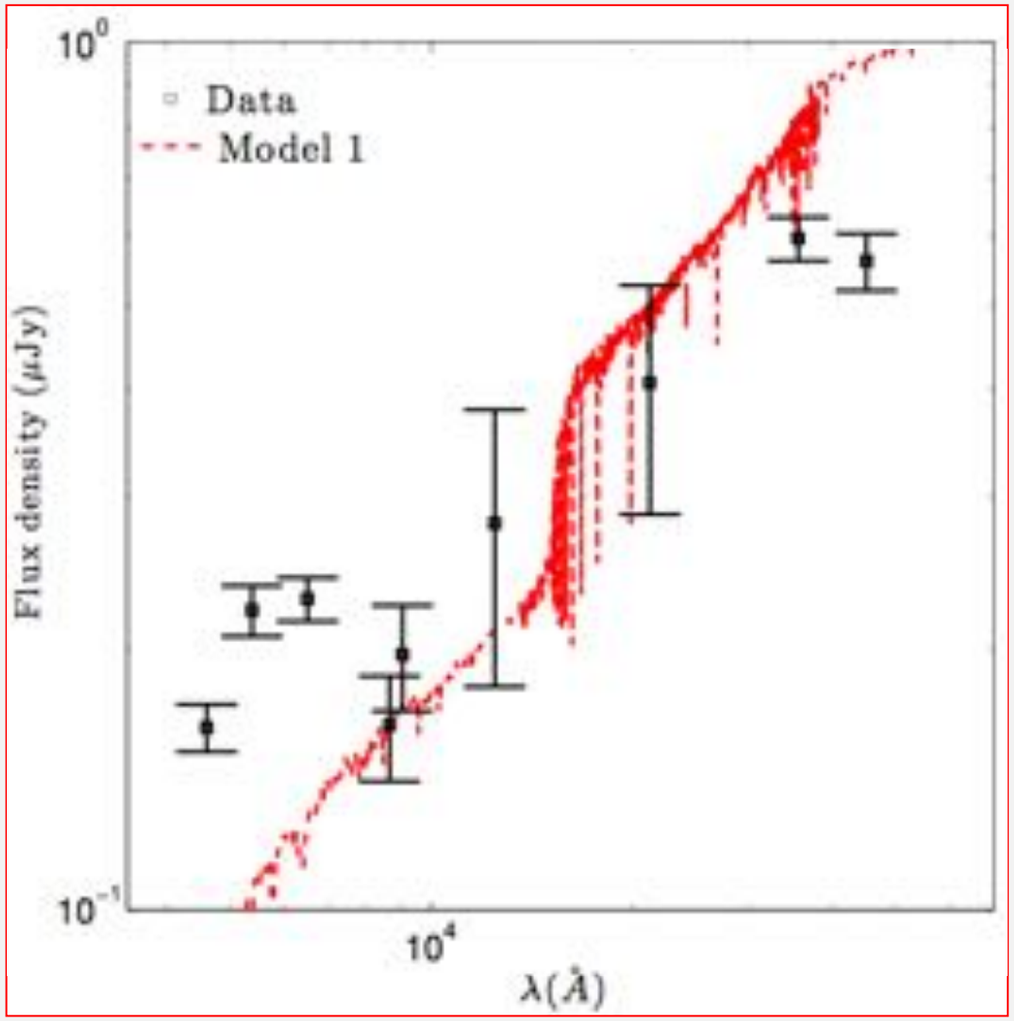
Have:
Fluxes
+
Models



SPECTRAL ENERGY DISTRIBUTION FITTING

Want:
Physical properties

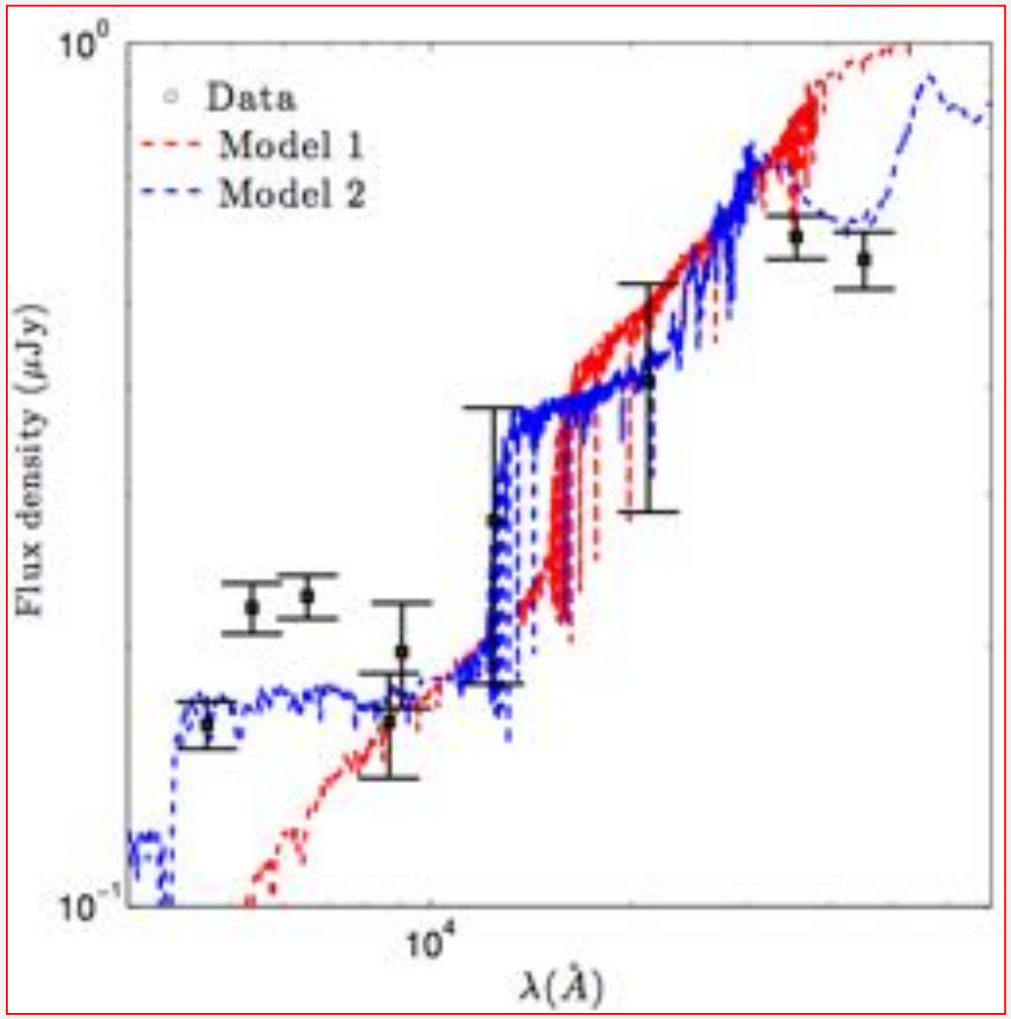
Have:
Fluxes
+
Models



SPECTRAL ENERGY DISTRIBUTION FITTING

Want:
Physical properties

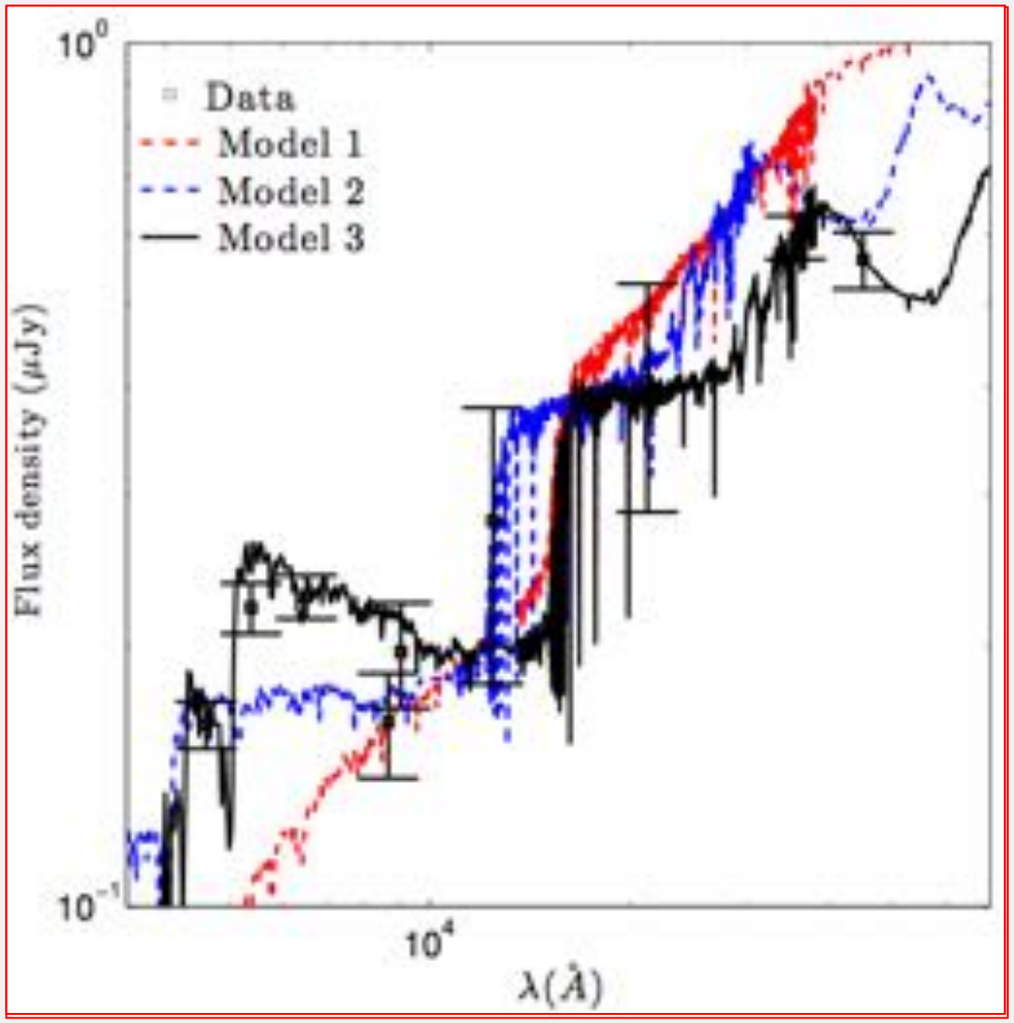
Have:
Fluxes
+
Models



SPECTRAL ENERGY DISTRIBUTION FITTING

Want:
Physical properties

Have:
Fluxes
+
Models



Properties of models are known $\Rightarrow$ infer properties of observed galaxy

Q1: How do I tell whether a model resembles the data?

Q2: How do I pick the models?

Q3: How do I compute the uncertainties?

ONE POSSIBLE OPTION

GalMC (VA, Gawiser, Guaita 2011) was the first publicly available Markov Chain Monte Carlo algorithm for SED fitting.

https://www.drvivianaacquaviva.com/galmc.html

SpeedyMC (VA, Gawiser, Guaita 2012) was its 20,000x faster cousin, available upon request.

Q1: GALMC'S LIKELIHOOD FUNCTION

$$\mathcal{L}(\text{Age, Mass, E(B-V), SFH, z, Z}) = \alpha\, e^{-\chi^2/2} = \alpha e^{\sum_{i=1}^{n_{\text{bands}}} \frac{(\phi_i^{\text{theory}}(\text{Age, Mass, E(B-V), SFH, z, Z}) - \phi_i^{\text{obs}})^2}{2\sigma_i^2}}$$

Age, Mass, metallicity *Z*:
theoretical spectrum
from stellar populations synthesis models

Convolve with
Star Formation History
(SSP,CSF, bursty, linear, $e^{\pm\tau}$)

Add nebular emission
from young stars
(continuum + lines)

Correct for dust
(Milky Way, Calzetti)

Repeat for other stellar
populations if desired

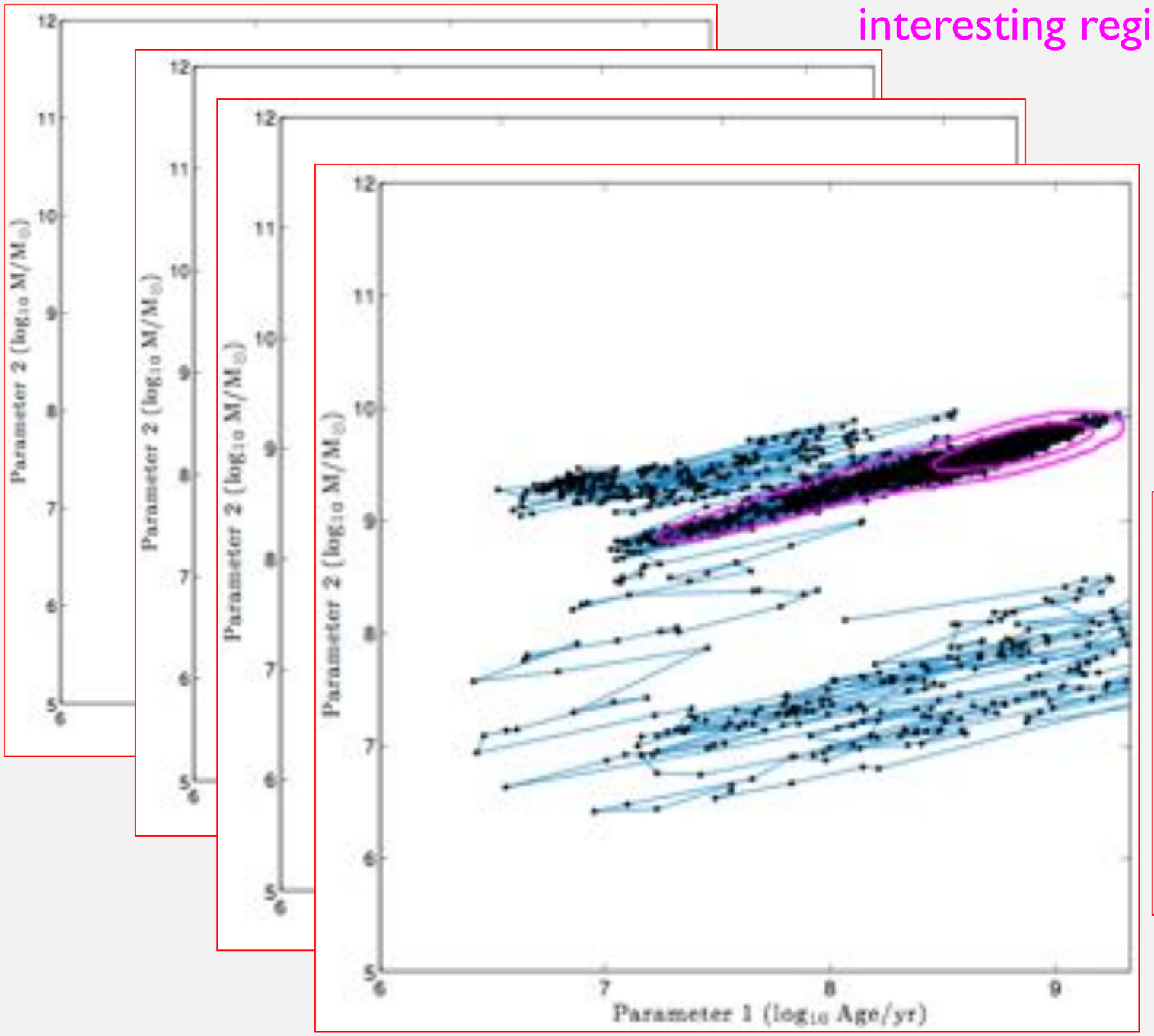
Place at right z, correct for IGM
absorption, move to obs frame, dim

Multiply by filter transmissions
and transform to flux density

Theoretical flux
entering $\mathcal{L}$

observed fluxes
and photometric error

Q2: GALMC'S PATH IN PARAMETER SPACE



THE MCMC WAY:

EFFICIENT!

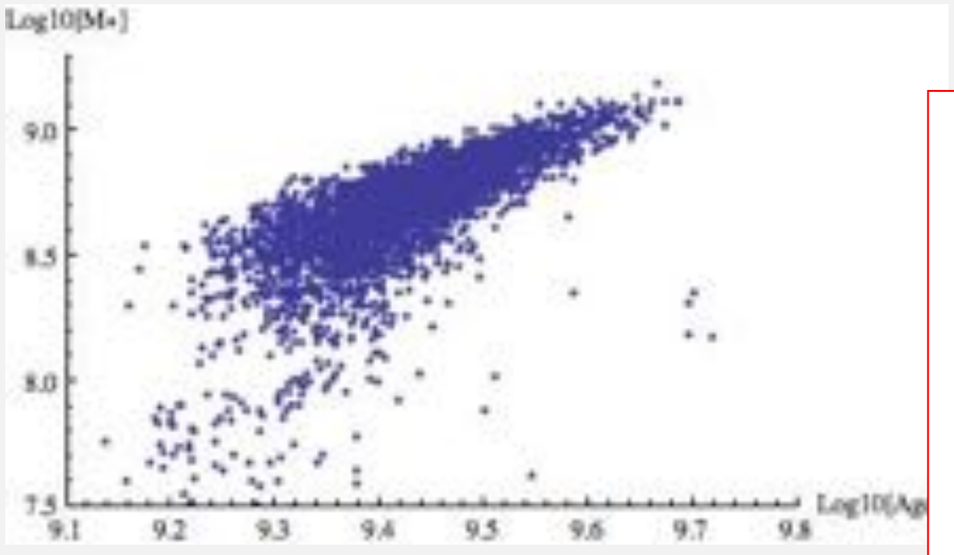
most time spent in informative region even if you don't previously know where it is

Q3: IT'S EASY TO COMPUTE UNCERTAINTIES

1. Assume Gaussian PDF and compute 68%, 95%
as 1 and 2 σ deviations from best fit: WRONG

2. Integrate PDF in many dimensions
to find contours enclosing 68%, 95% of total volume: PAINFUL

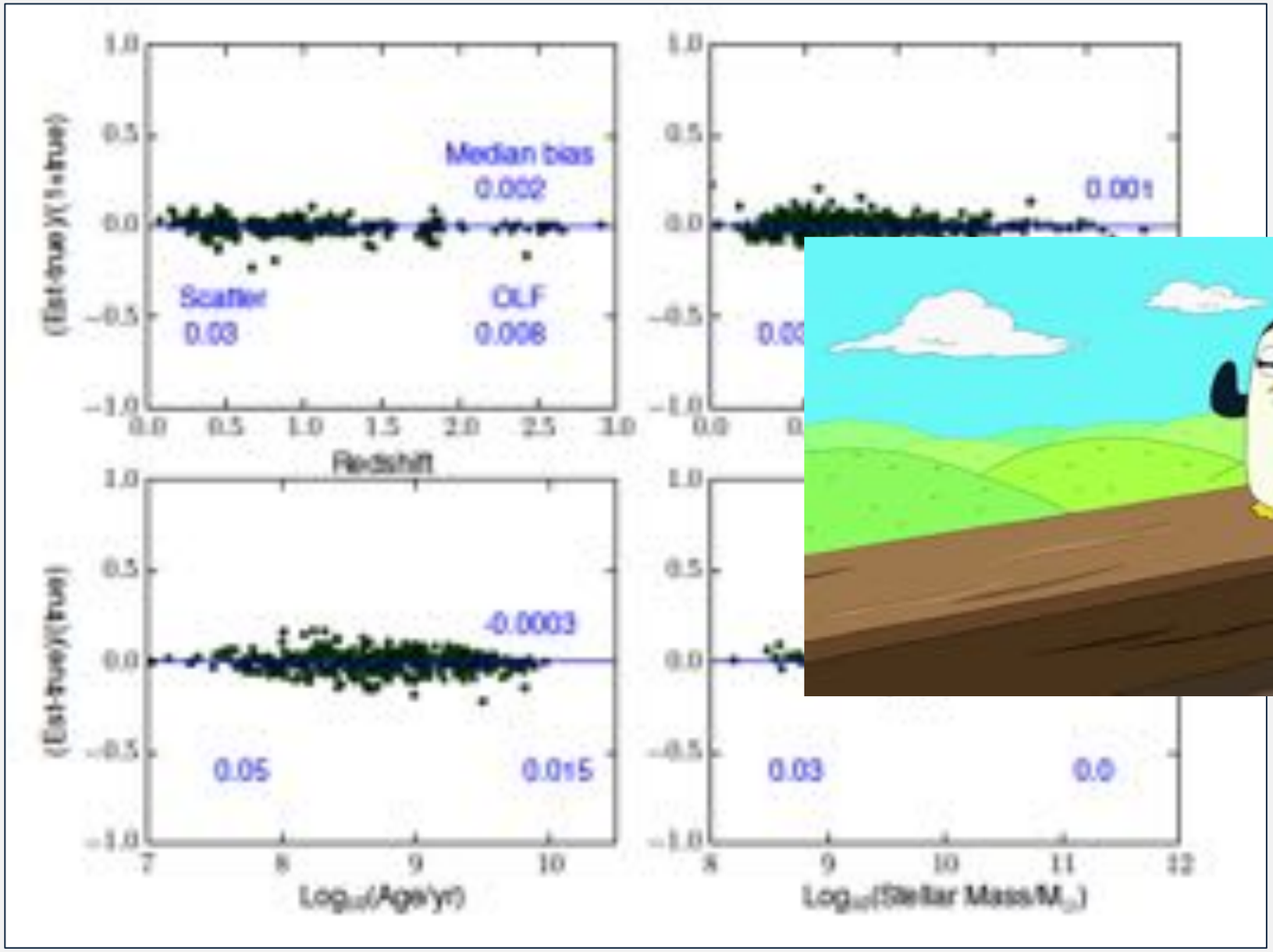
MCMC:
Density of points is Proportional to the PDF: you just have to COUNT POINTS



EASY!
no assumptions for PDF, accurate uncertainties

DOES IT WORK?

Test on mock data



FROM 2011 TO TODAY…

SINOPSIS

SatMC

Le Phare

VESPA

GalMC

SEABASs

FAST

Prospector

FADO

BEAGLE

BayeSED

BAGPIPES

MAGPHYS

CIGALE

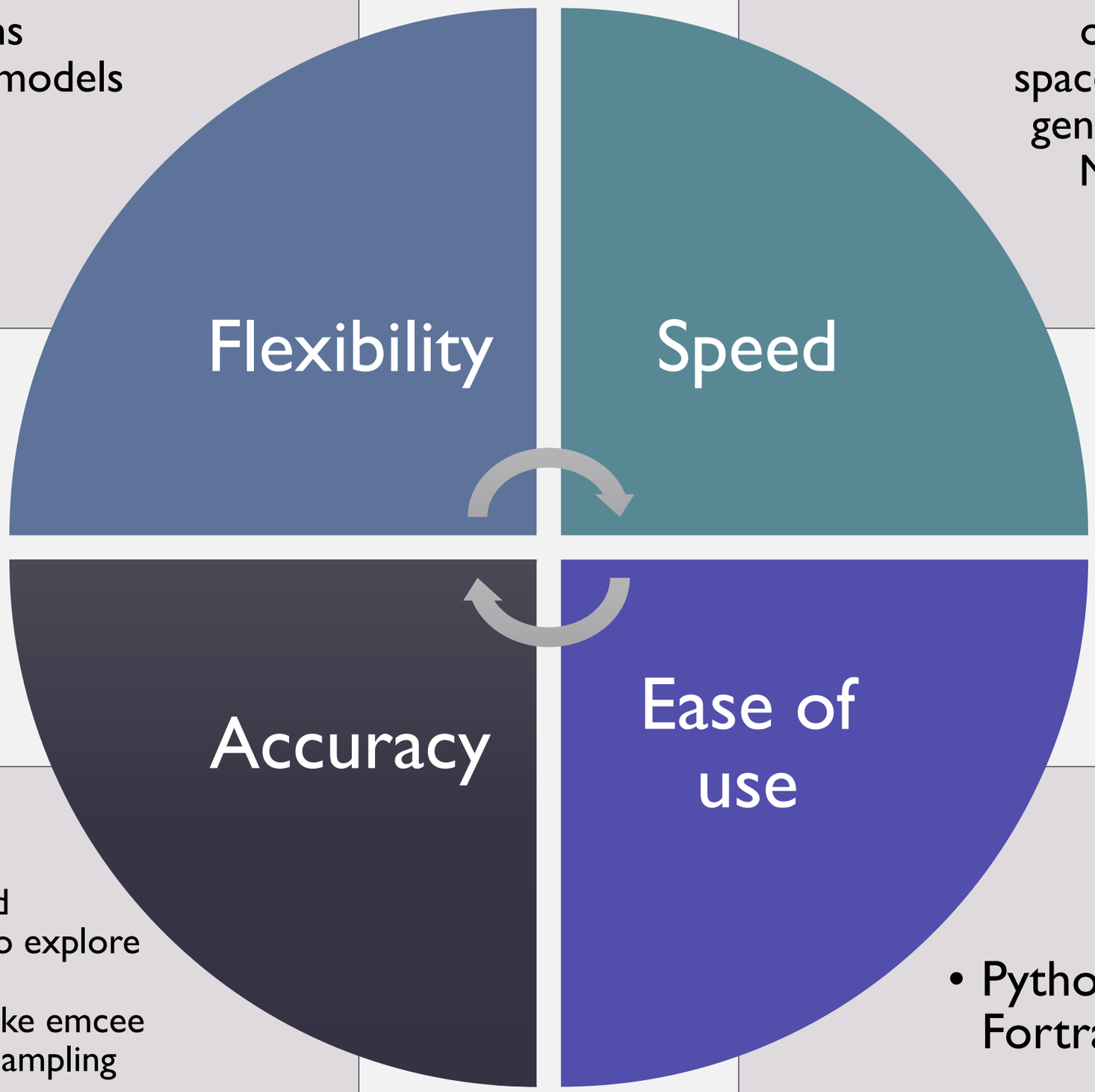
iSEDfit

Interrogator

SED FITTING: A SUCCESS STORY

- e.g. Flexible Stellar Populations Synthesis models

- Faster exploration of parameter space and model generation with MCMC, PCA

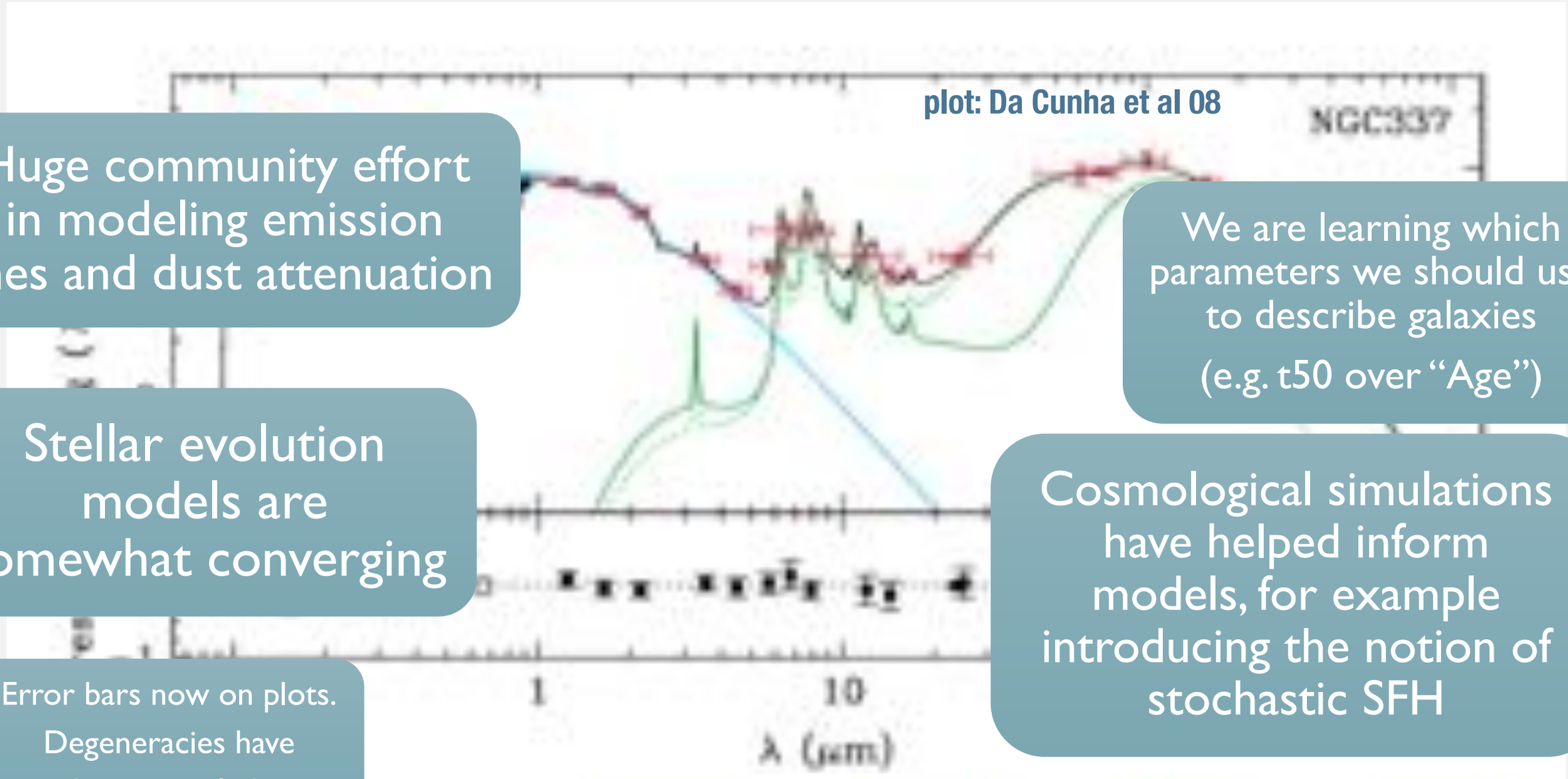


- Better suited algorithms to explore multi modal posteriors, like emcee and nested sampling

- Python vs Fortran or C

Panchromatic estimation (from ultraviolet to far infrared) FTW!

Millions of stellar masses and other parameters measured from CANDELS+3D-HST, Hubble Frontier Fields, ….



Huge community effort in modeling emission lines and dust attenuation

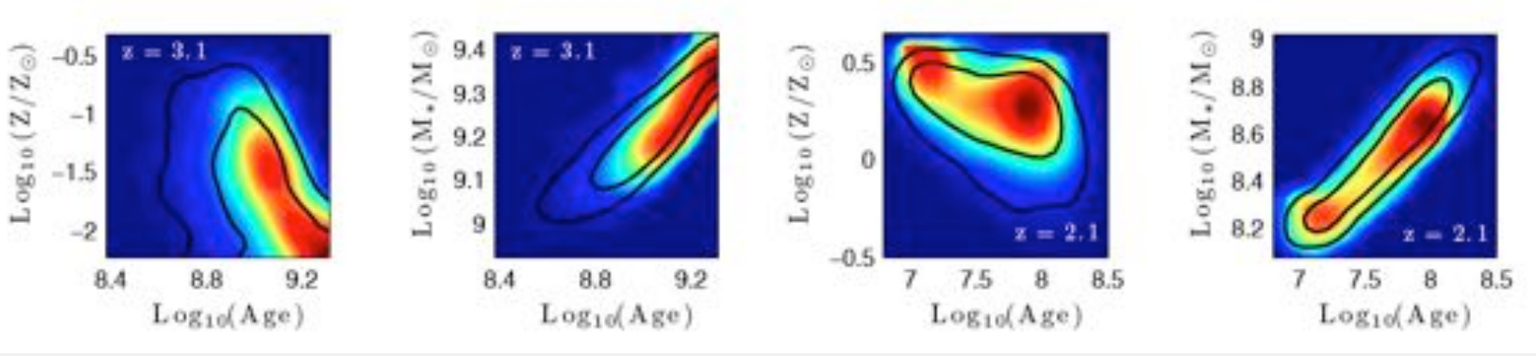
We are learning which parameters we should use to describe galaxies

(e.g. t50 over "Age")

Stellar evolution models are somewhat converging

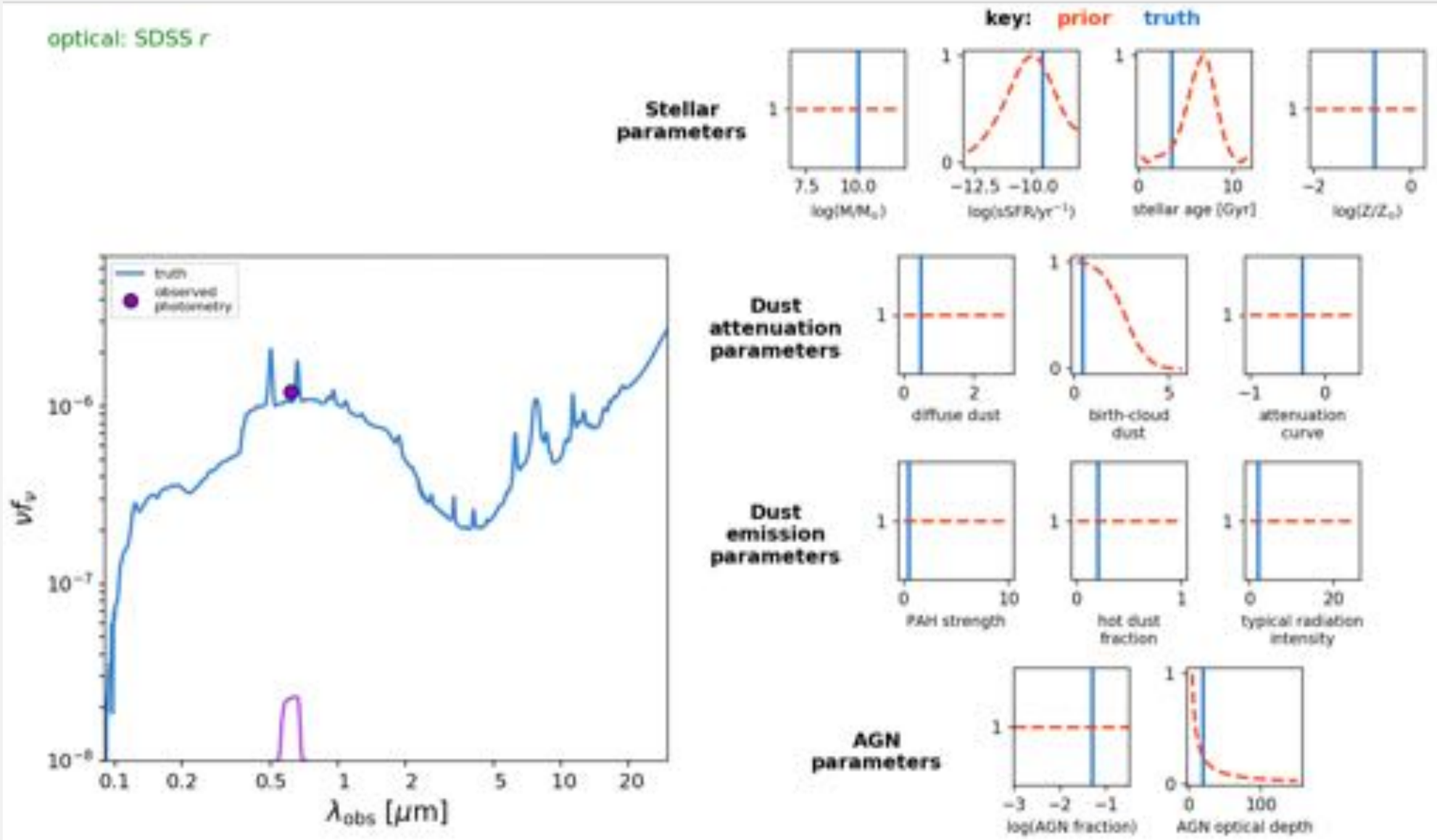
Cosmological simulations have helped inform models, for example introducing the notion of stochastic SFH

Error bars now on plots. Degeneracies have been unveiled.



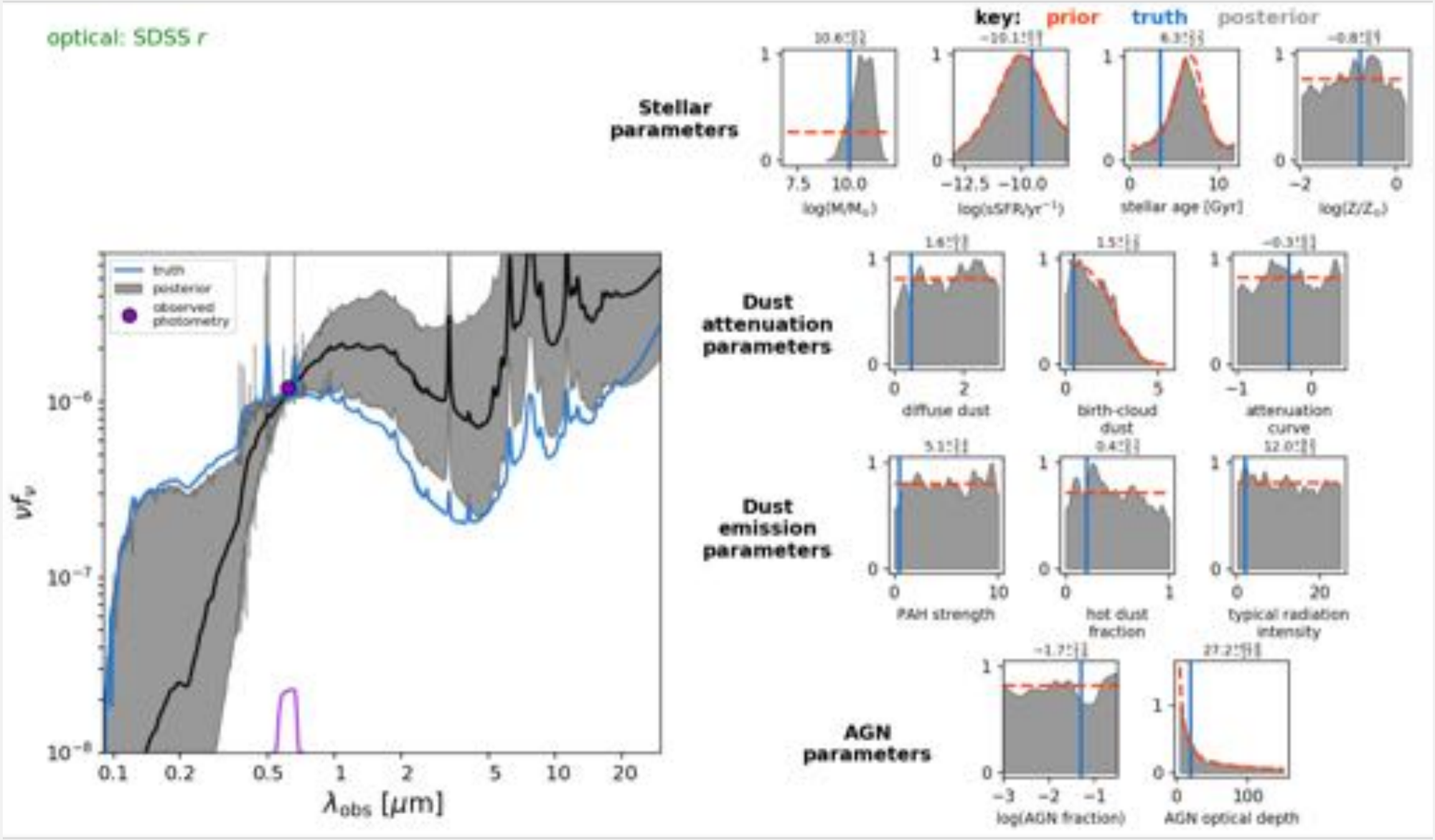
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



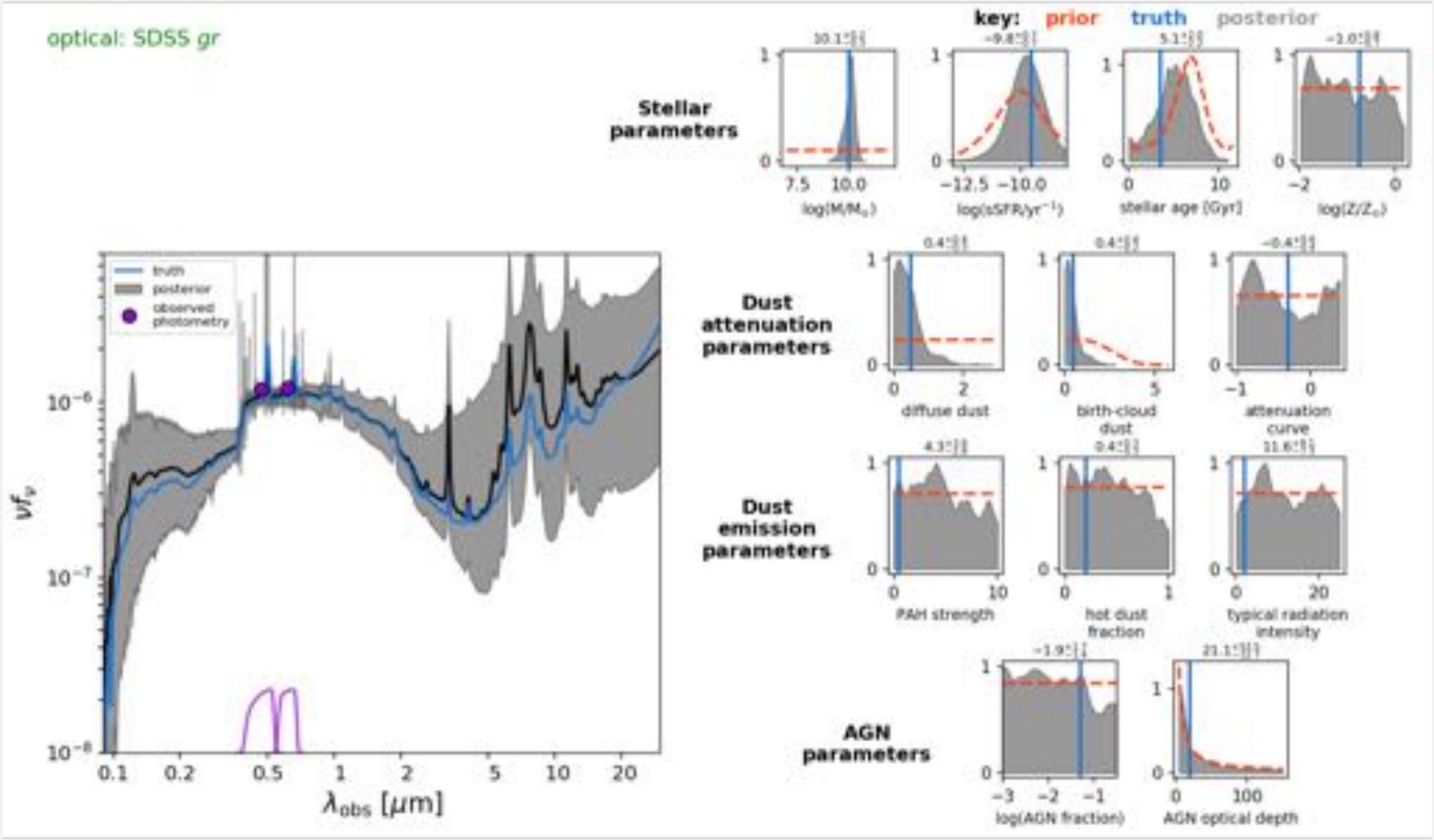
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



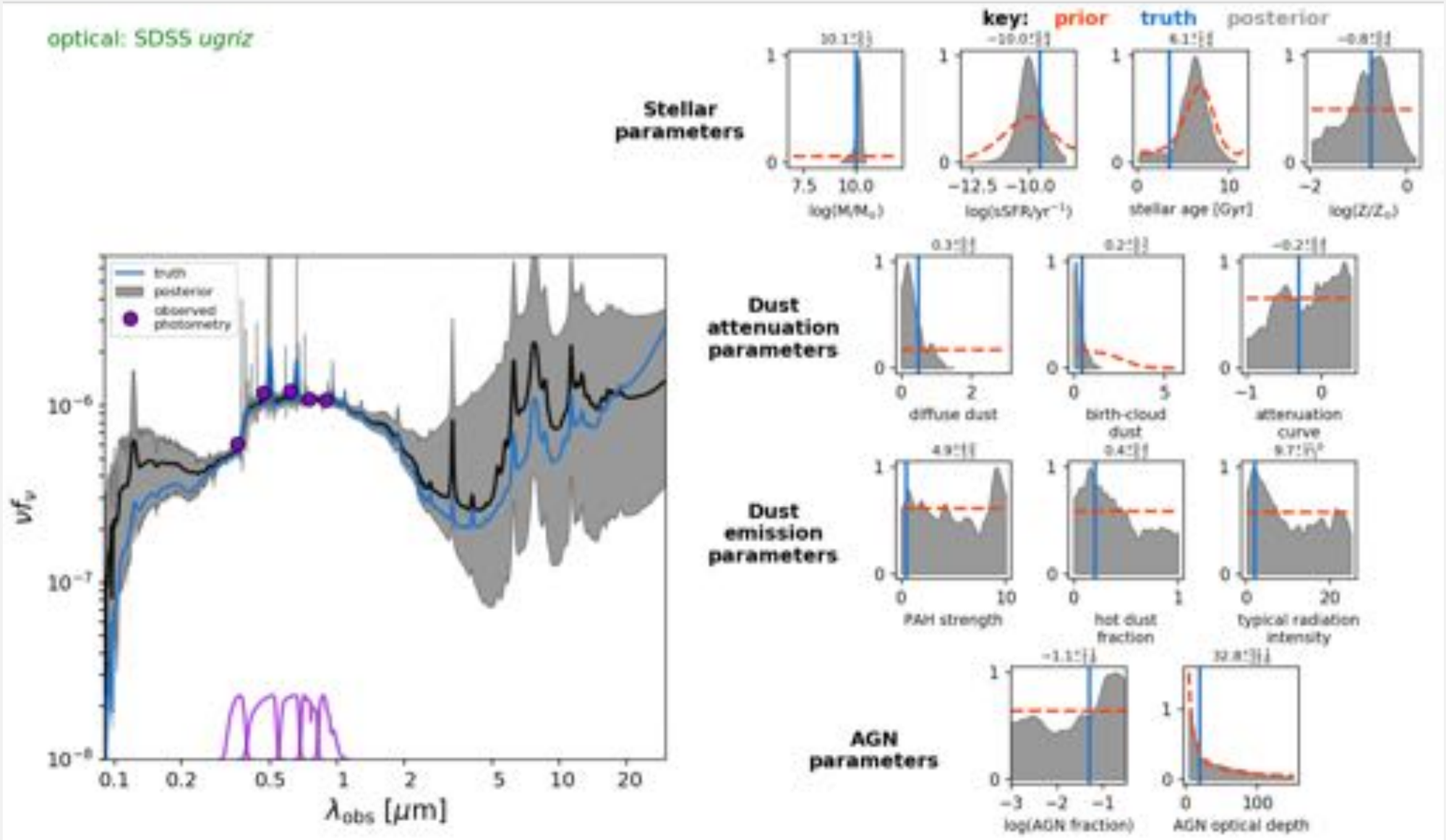
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



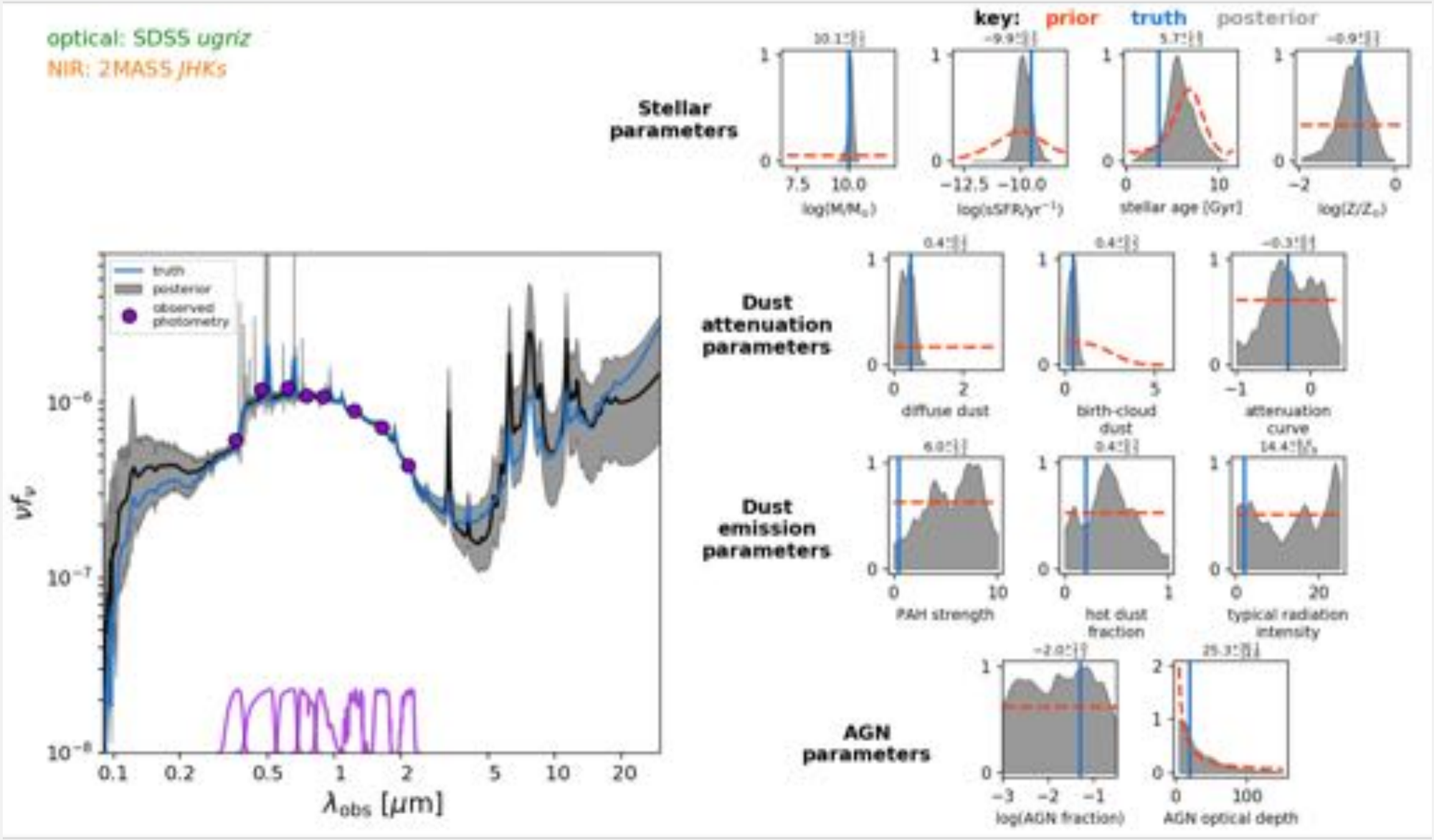
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



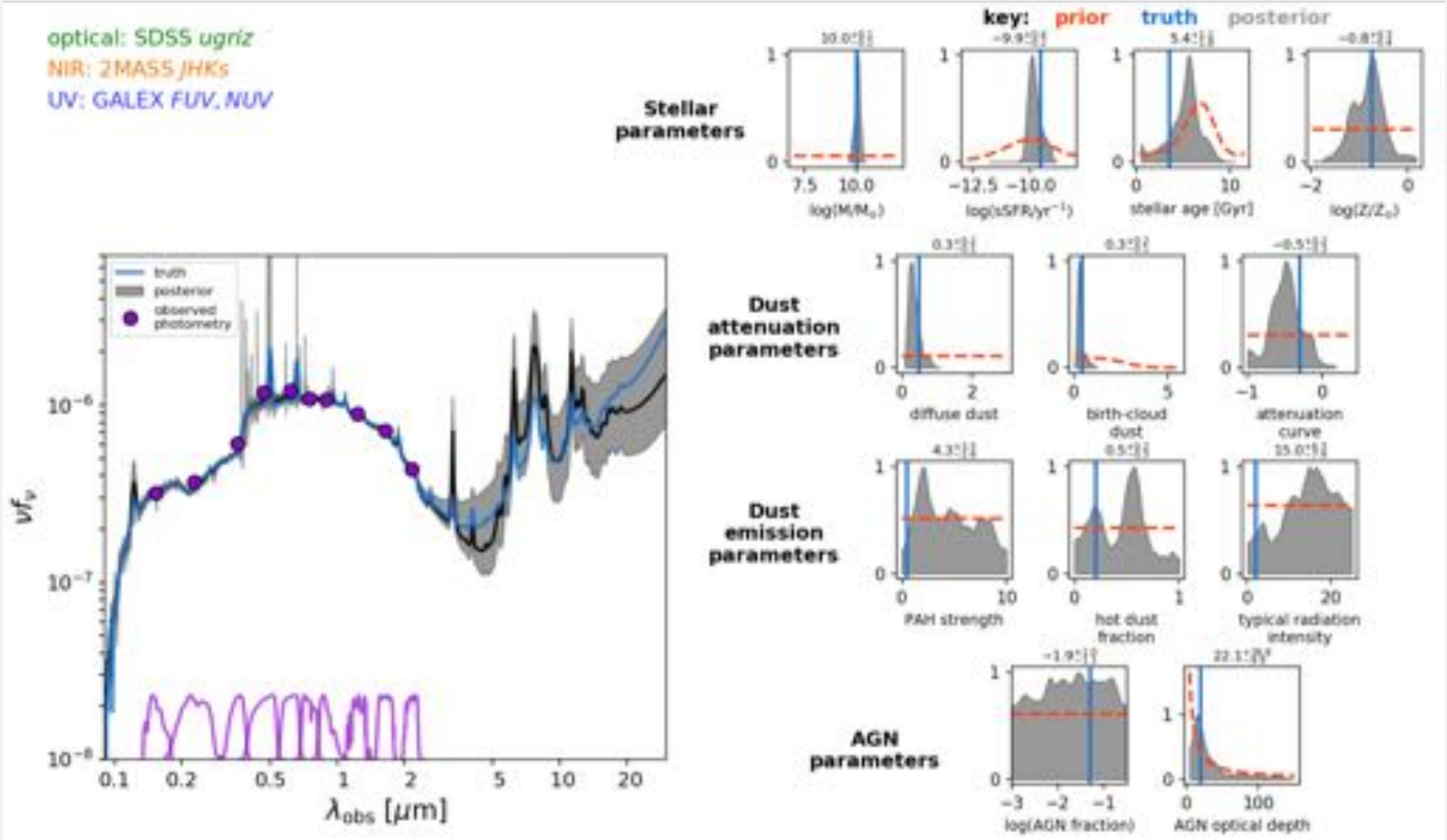
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



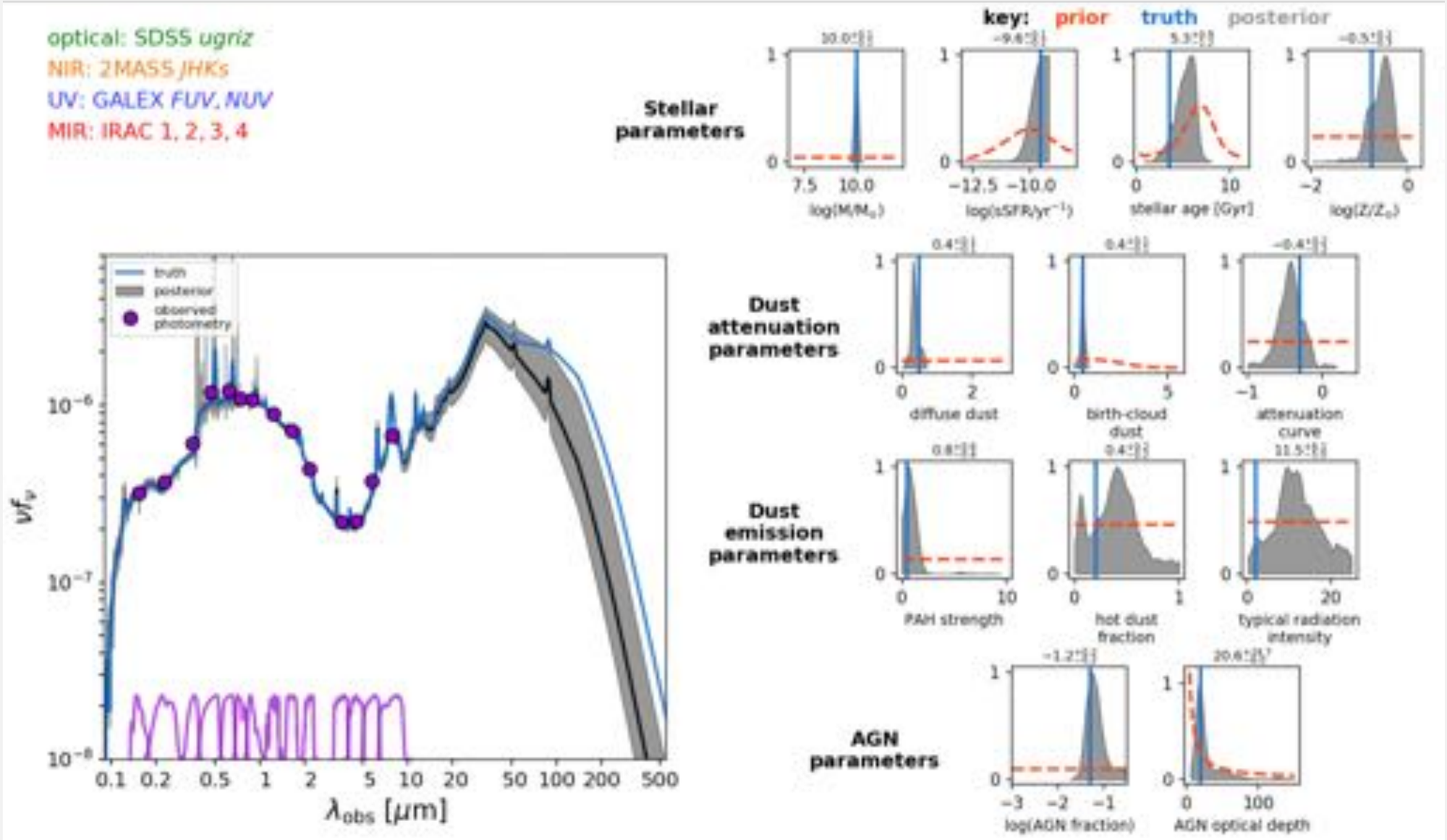
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



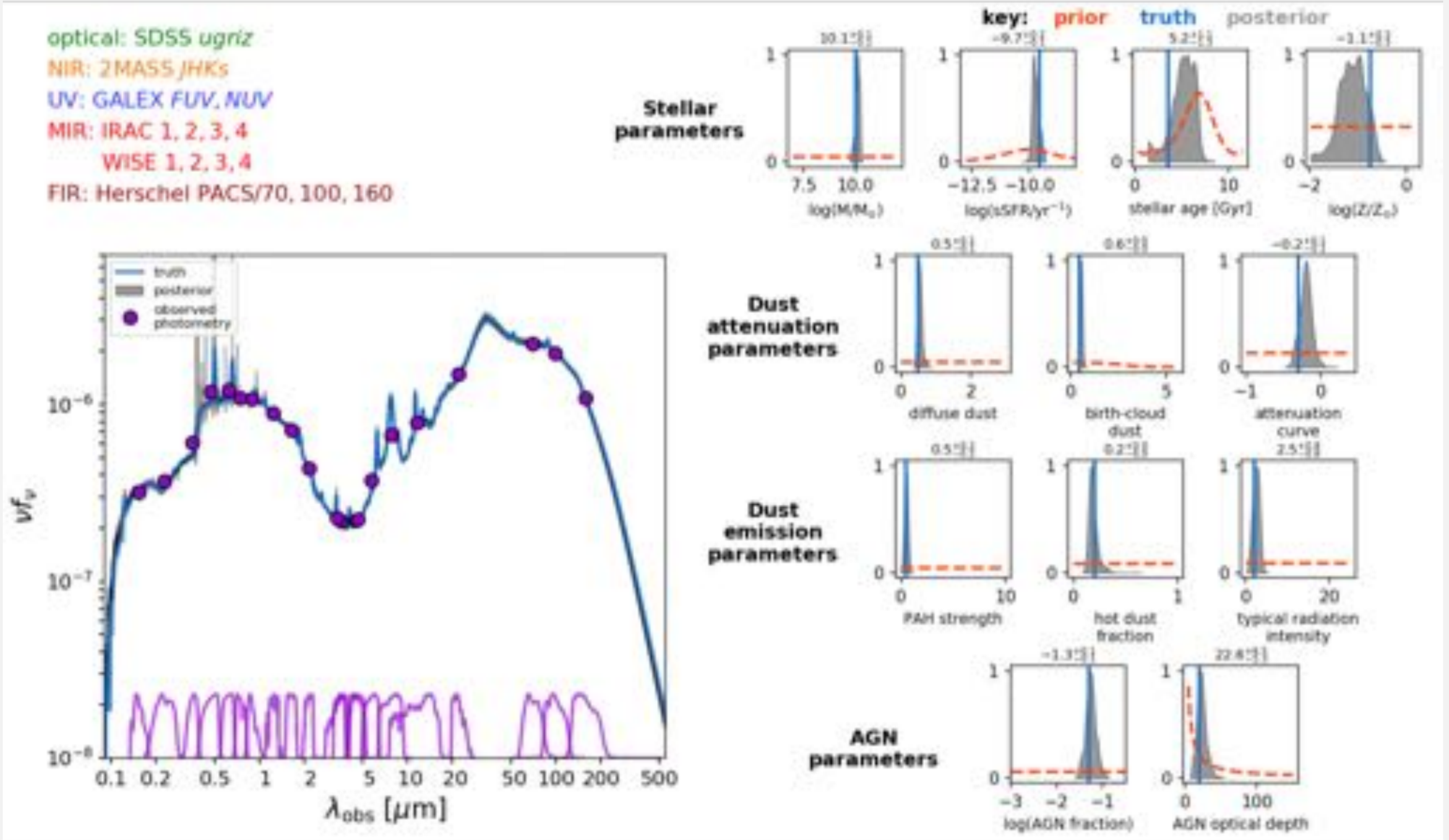
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML

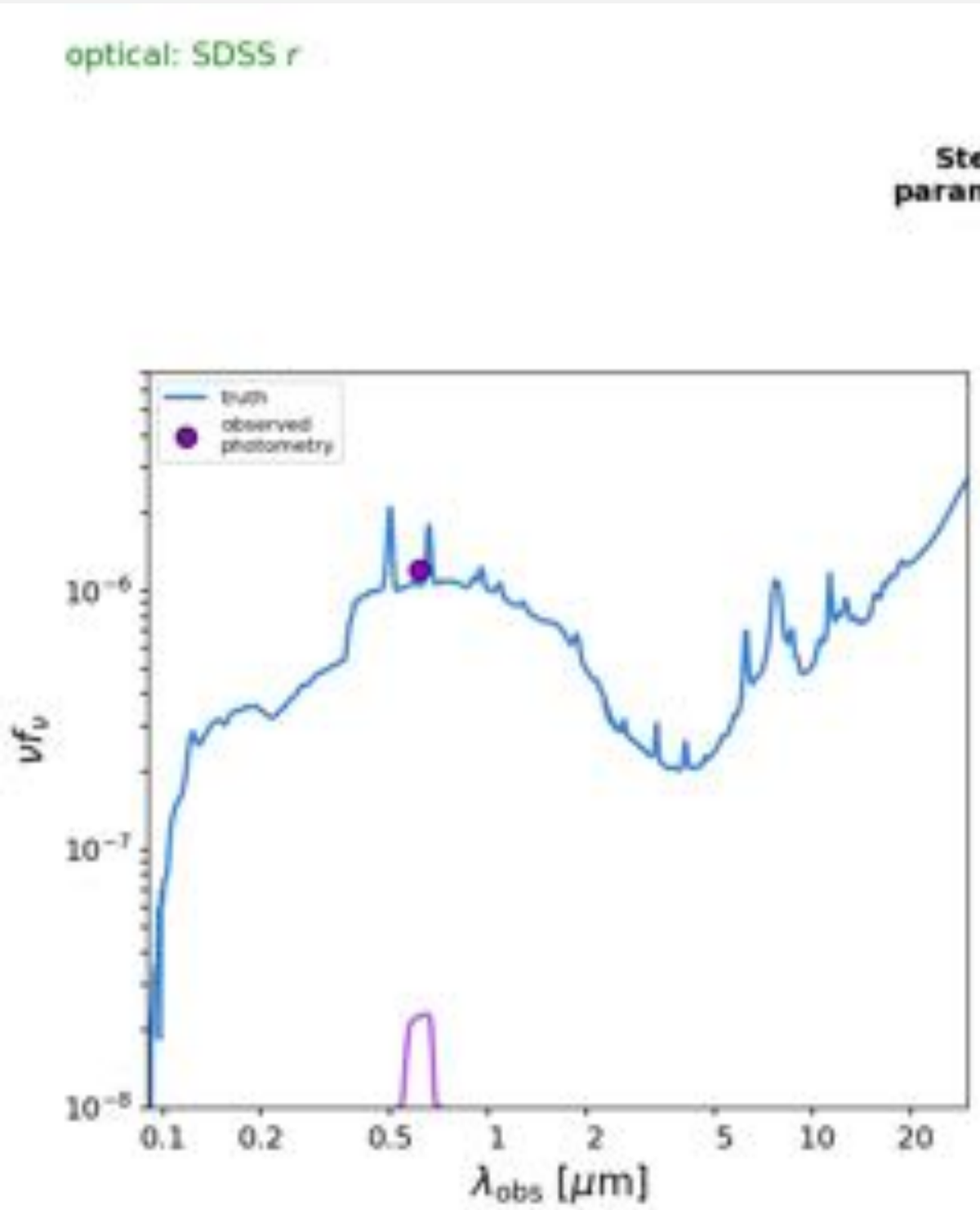


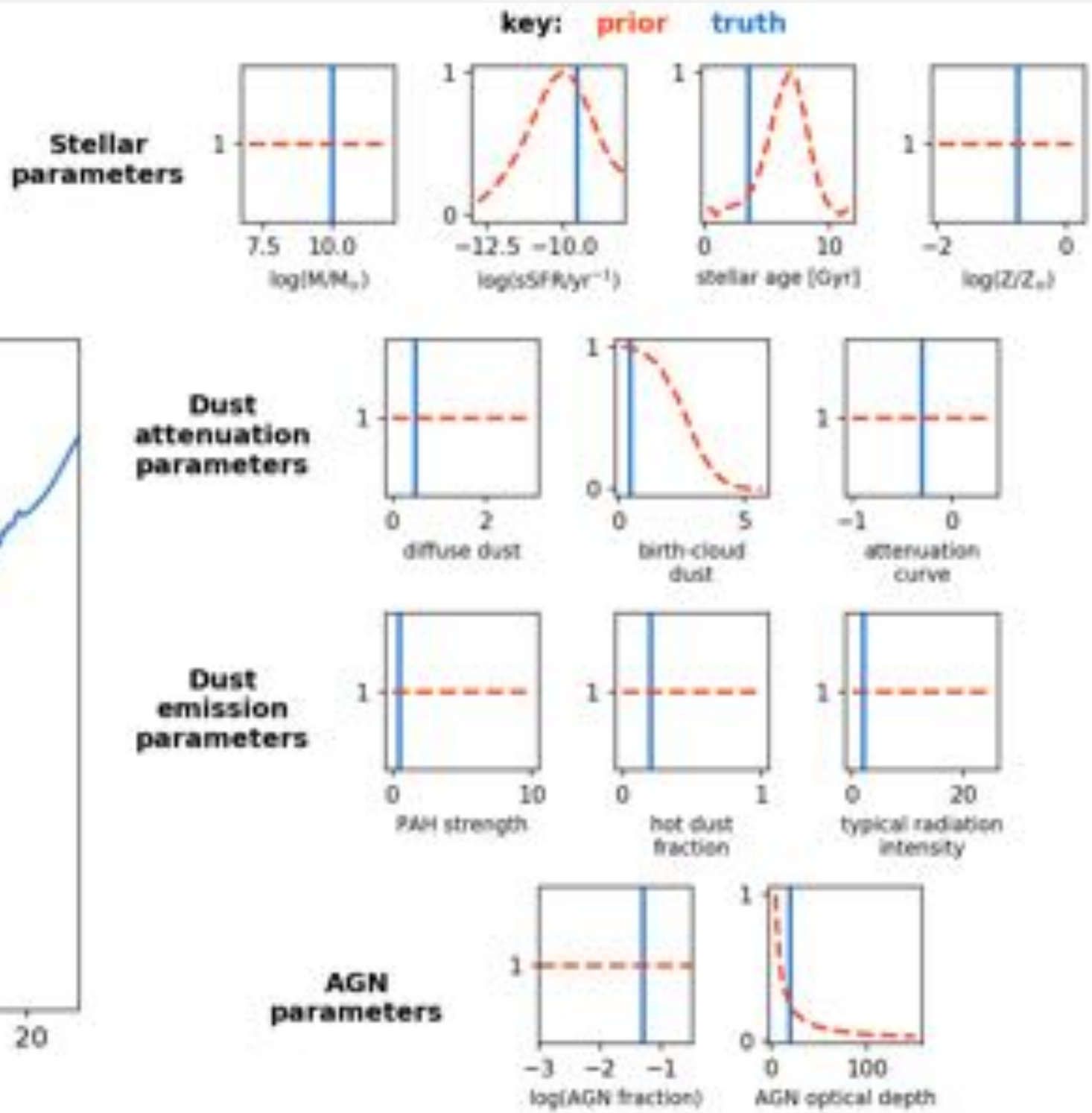
PROSPECTOR

HTTPS://PROSPECT.READTHEDOCS.IO/EN/LATEST/DEMO.HTML



now in real time



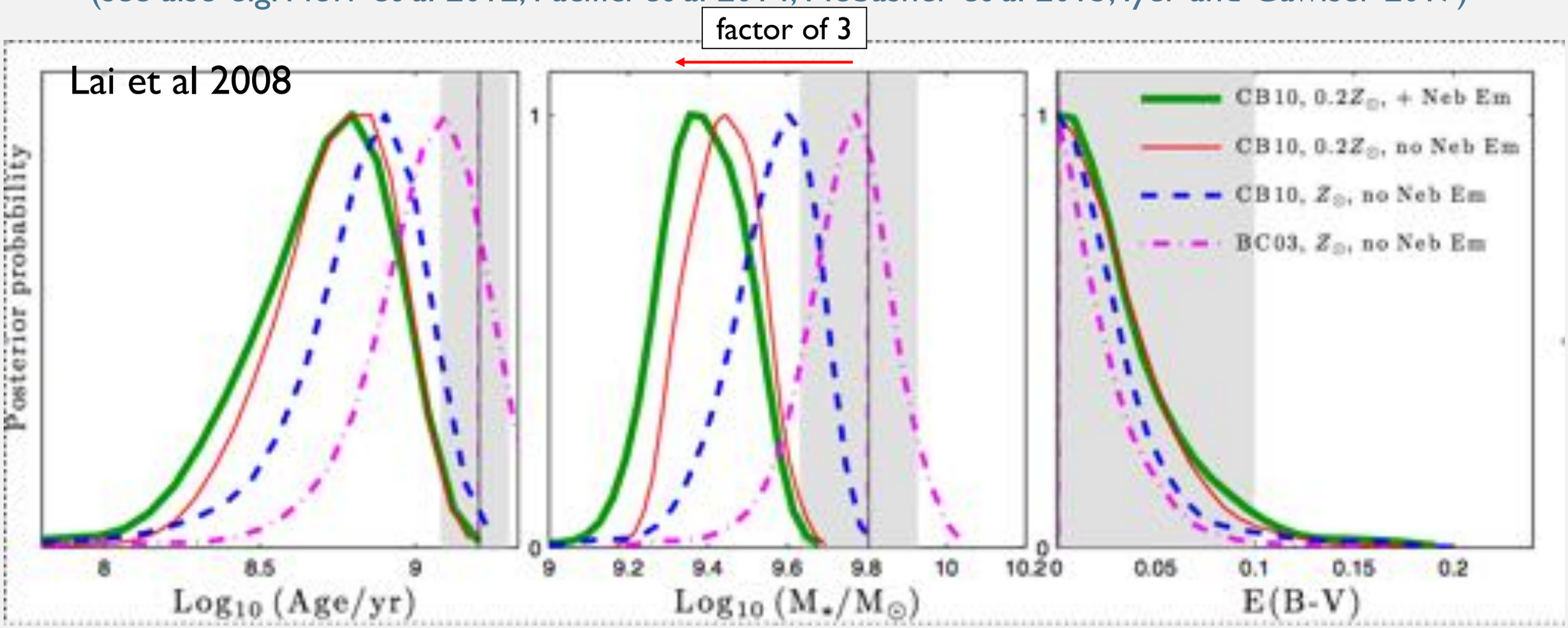


SED FITTING: A FAILURE STORY

ASSUMPTIONS MADE IN SED FITTING COUNT, OFTEN MORE THAN PHOTOMETRIC ERRORS ☹

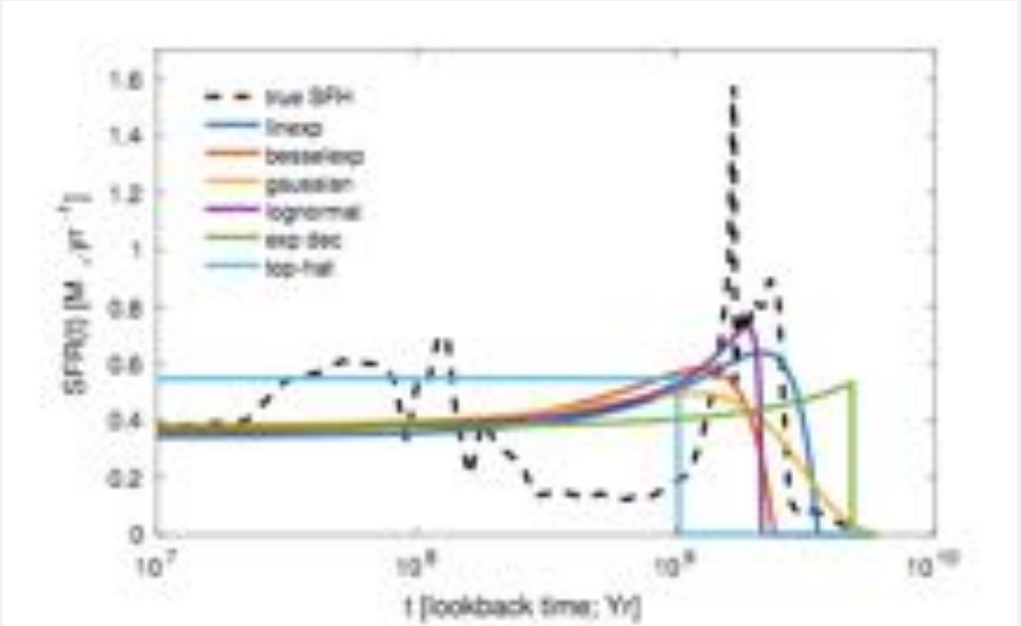
Plot from VA et al ApJ 737, 2011; LAE galaxies at z = 3
(see also e.g. Pforr et al 2012, Pacifici et al 2014, Mobasher et al 2015, Iyer and Gawiser 2017)

Lai et al 2008



THIS MATTERS BECAUSE IT MEANS THAT TAKING BETTER DATA WILL NOT HELP

SIMPLISTIC ASSUMPTIONS INDUCE BIASES IN INFERRED GALAXY PROPERTIES :(



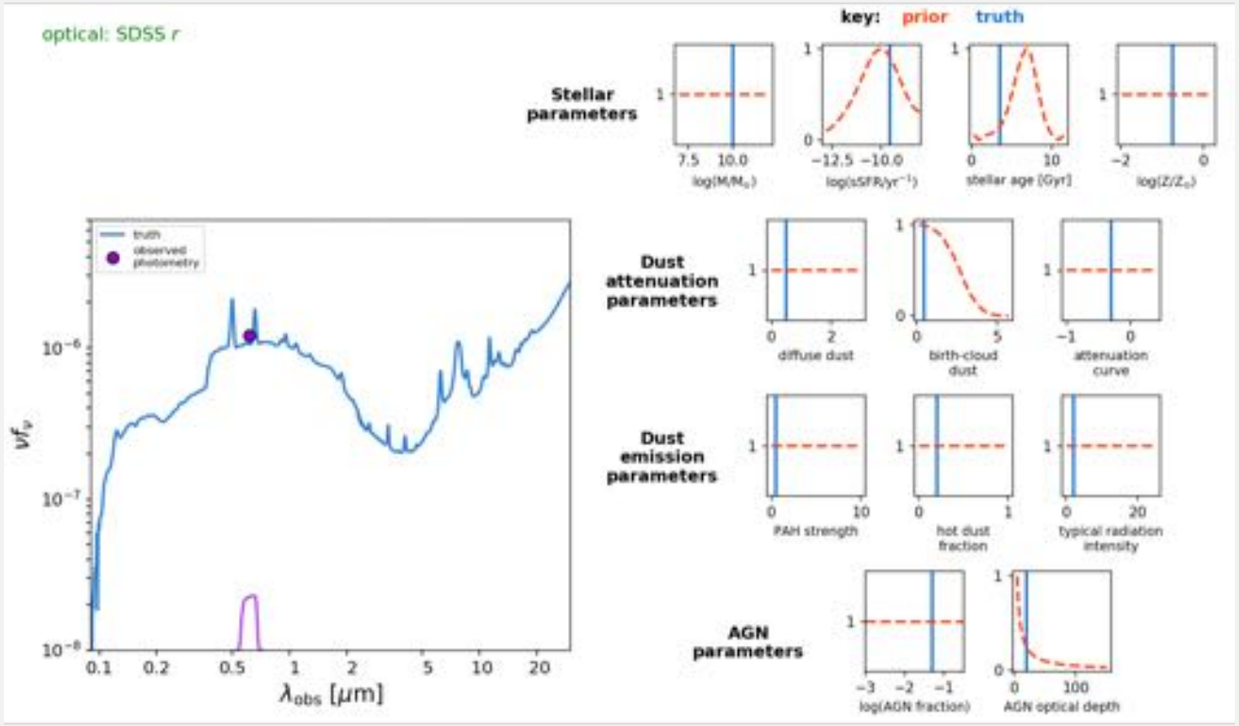
We need to simplify models so that they don't have too many parameters.
This causes a bias toward young stellar populations
(outshining)

TABLE 3
BIAS IN THE ESTIMATION OF PHYSICAL QUANTITIES DUE TO DIFFERENT SFH PARAMETRIZATIONS AT $z \sim 1$. [NO DUST OR NOISE]

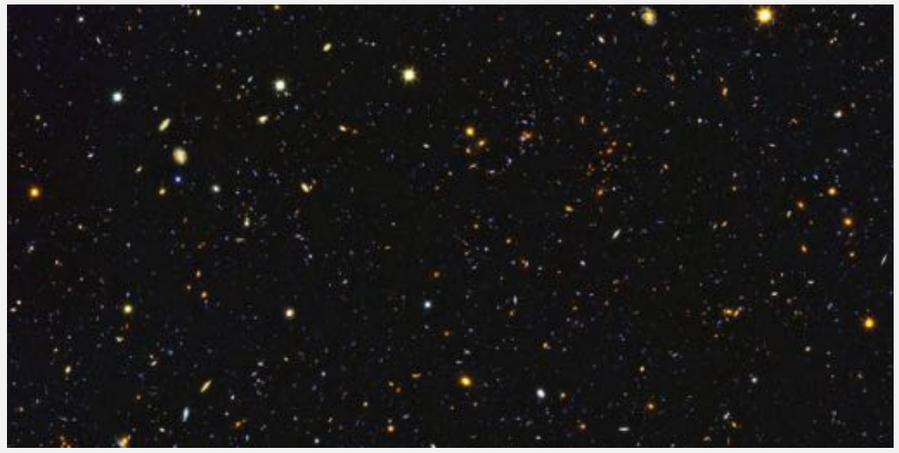
	M_*	SFR_{100}	SFR_{inst}	t_{90}	t_{50}	t_{10}	Age
CSF	-14%	5%	4%	-24%	-13%	-19%	-43%
tophat	-17%	2%	-11%	-27%	-24%	-41%	-59%
exponential	-20%	-2%	-7%	-21%	-34%	-55%	-70%
Linexp	-18%	-1%	-7%	-18%	-28%	-39%	-50%
Gaussian	-16%	-1%	-7%	-20%	-31%	-27%	-16%
lognormal	-14%	2%	-3%	-16%	-25%	-33%	-26%
Besselexp	-19%	-1%	-7%	-21%	-34%	-42%	-43%
Dense Basis	-6%	4%	1%	-4%	-4%	-22%	-29%

(Iyer & Gawiser 2017; also Maraston '10, Mobasher et al '15, Pacifici et al 13)

TOO MUCH TIME!



x



15k galaxies

~ 0.5 million CPU hours ~ 57 years

For 30 billion galaxies: ~ 100 million years

Initiative for the Theoretical Sciences, The CUNY Graduate Center

Adventures in the Theoretical Sciences
An informal, online summer school.

June 4 and 5
Few-body and many-body chaos
Vladimir Rosenhaus

June 11 and 12
Multi-parameter models and information geometry
Katherine Quinn

June 25 and 26
Big universe, big data: Emerging challenges in astrophysics
Viviana Acquaviva

June 18 and 19
Statistical physics and machine learning
David J. Schwab

July 2 and 3
Precision and emergence in the physics of life
William Bialek

July 9 and 10
Driven quantum systems
Vadim Oganesyan

Six lecturers will present four hours of lecture and discussion each, touching a wide range of topics. Our goal is to introduce students to the excitement of our fields, and to encourage thinking about theory as a unifying activity. We expect students to have solid backgrounds in statistical physics and quantum mechanics; more specialized topics will be introduced as needed. Our target audience overlaps advanced undergraduates and beginning graduate students in the US, and MSc students abroad.

Registration: itsatcuny.org/summerschool

Lectures from 11am-1pm (EDT)

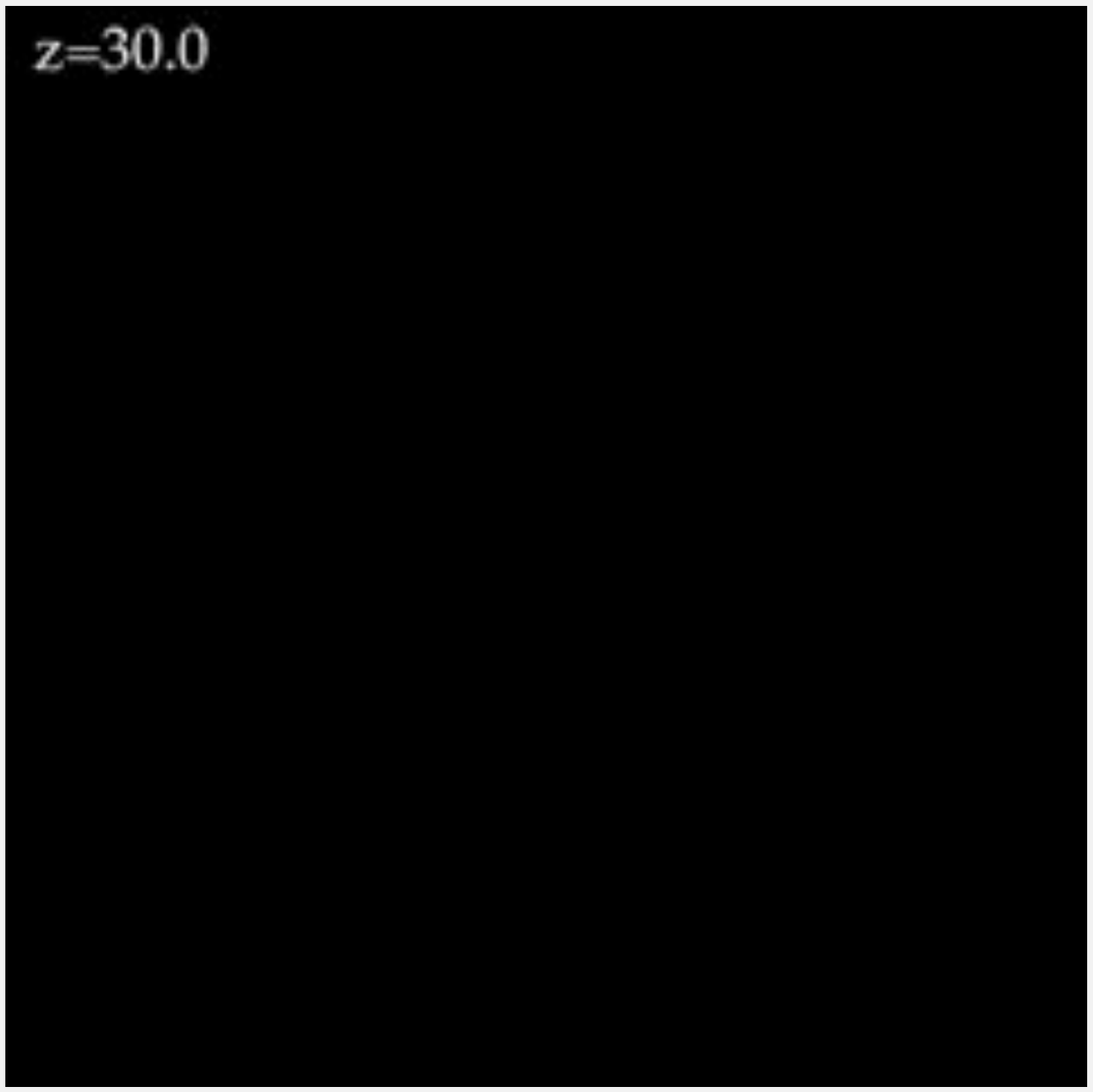
Questions: info@itsatcuny.org

BEYOND SED FITTING

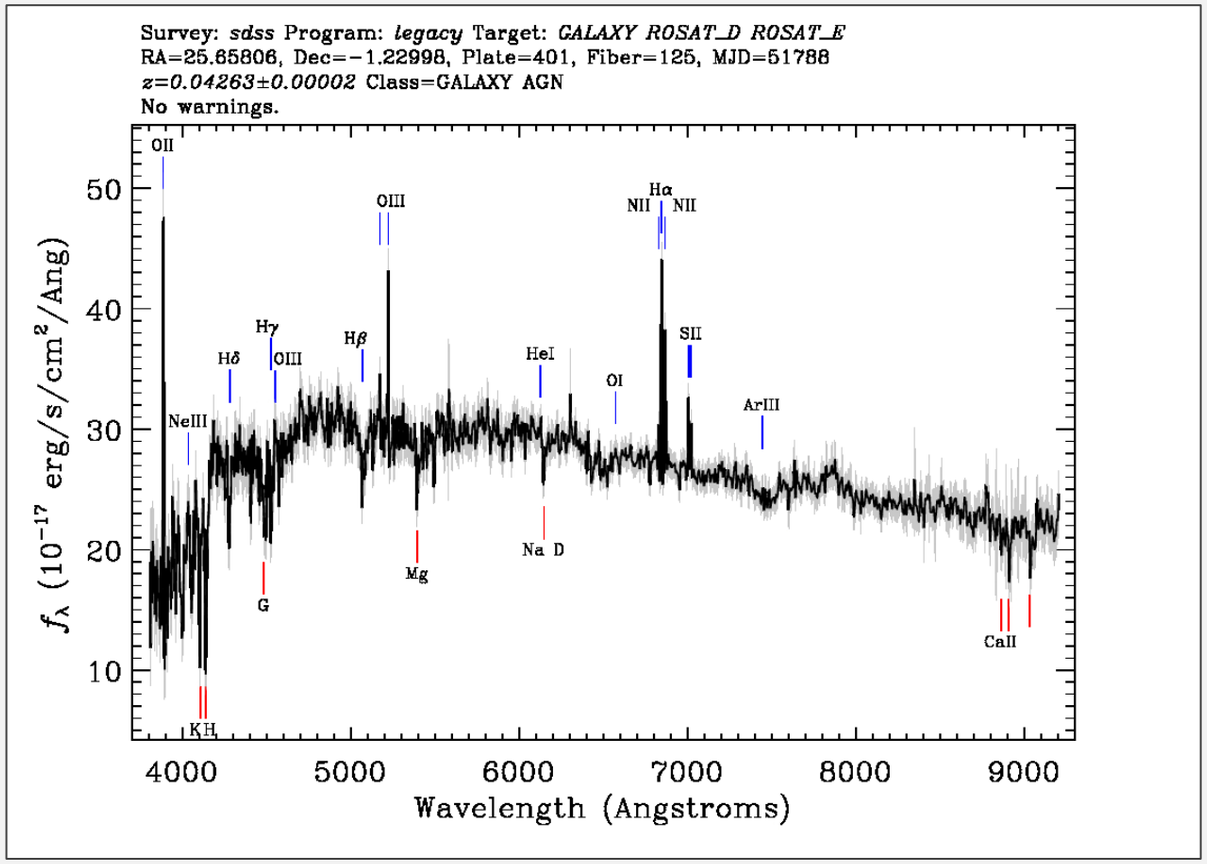
DETERMINING GALAXY STAR FORMATION HISTORIES FROM THEIR SPECTRA USING ML

Project with Chris Lovell, *MNRAS*, Volume 490, Issue 4, Dec 2019

what we want to know



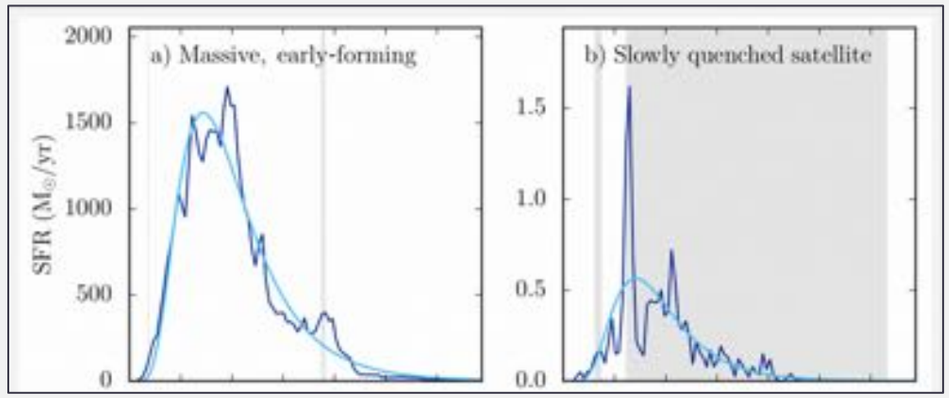
what we can observe



Video Credit: Phil Hopkins

MACHINE LEARNING TO MEASURE SFHS

plot by Diemer+ 17



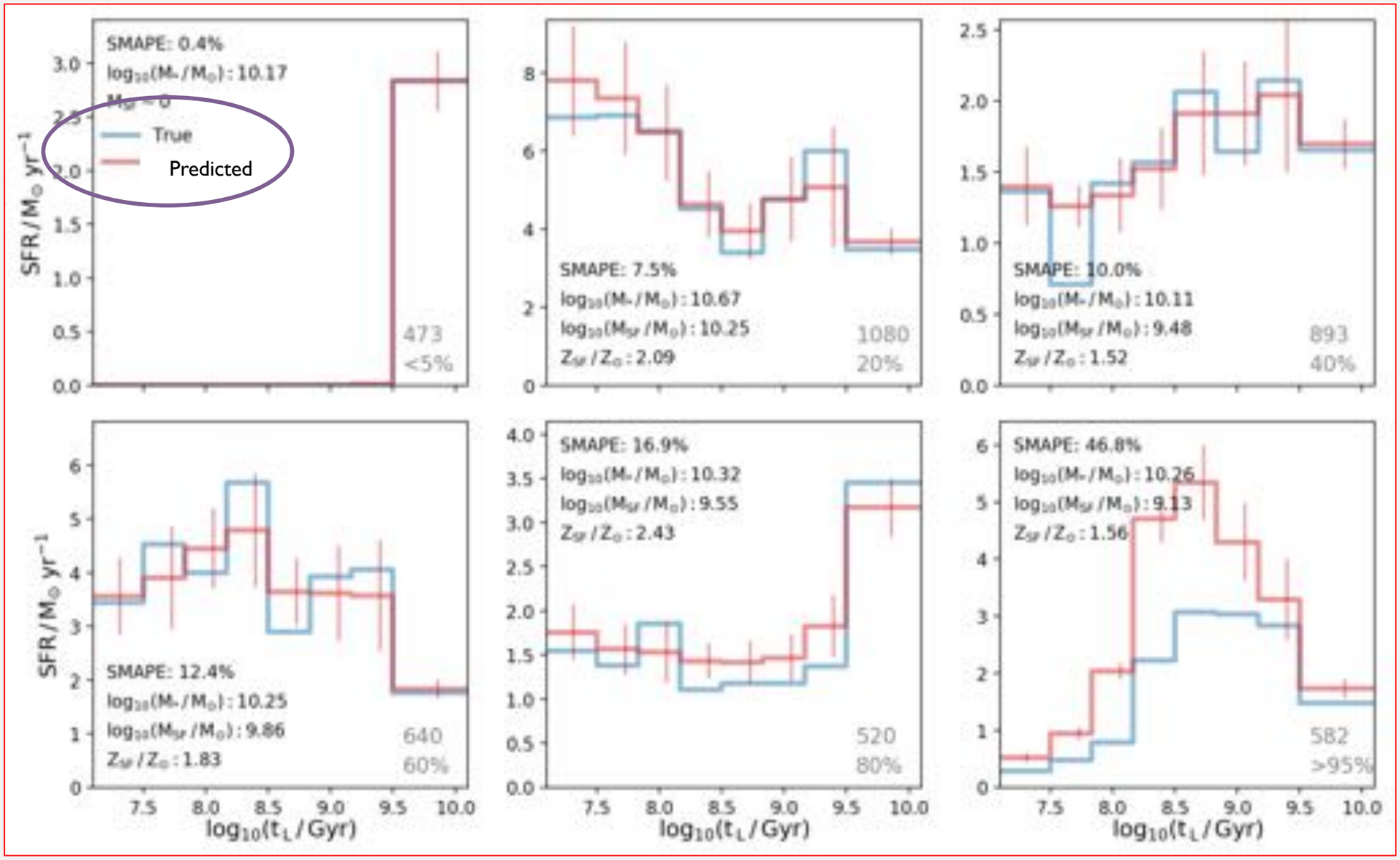
- Take star formation histories of galaxies from Illustris and EAGLE (state-of-the-art sims)
- Generate a learning set catalog of galaxies with realistic spectra
- Teach a CNN the connection between spectra (observed) and star formation history (inferred)
- Test within the same simulation and across various simulations to evaluate generalization properties

WHY MACHINE LEARNING?

- We don't need to specify a model for SFH
- We can hope to avoid the outshining bias
- We learn from the object AND the population

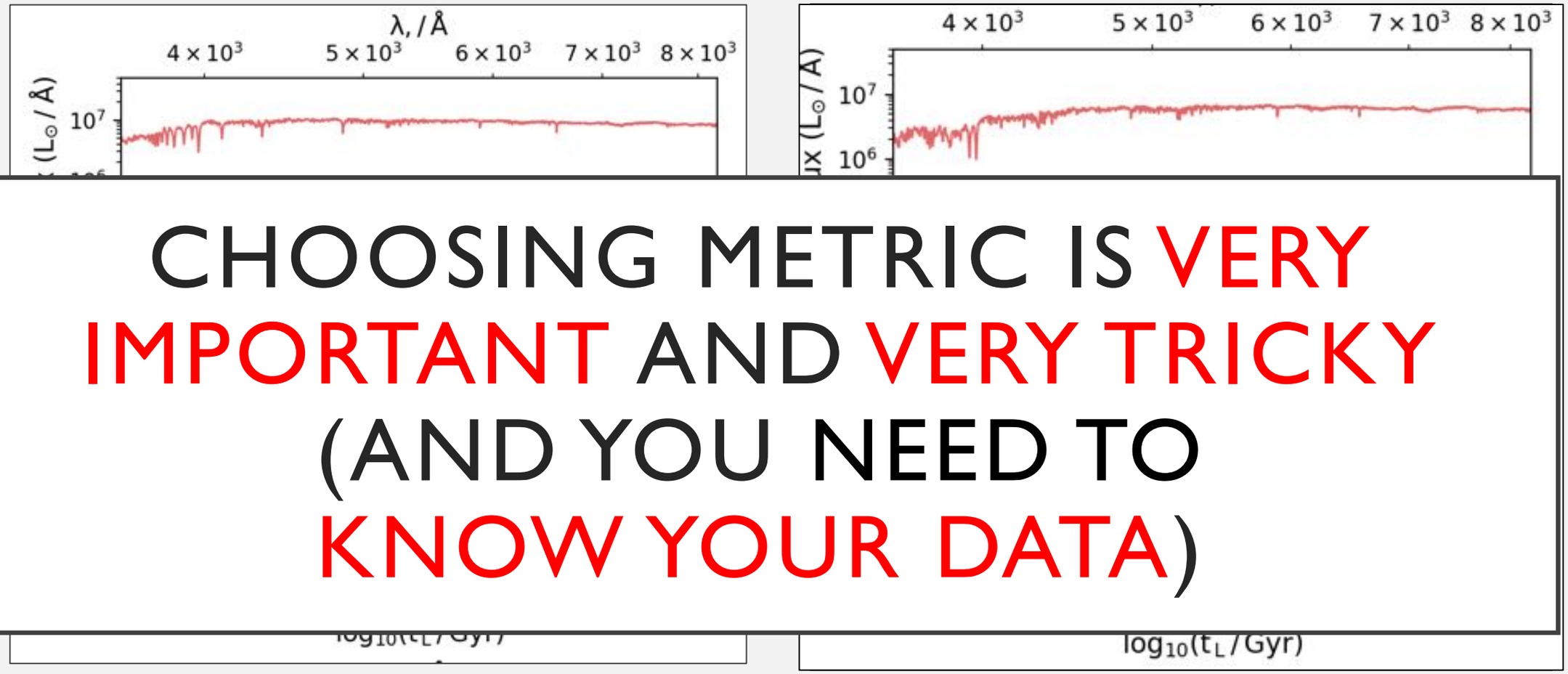
- Less transparent generalization properties
- Dependent on reliability of simulations (typically no ground truth!)
- Training on data only possible for limited samples (e.g. nearby galaxies?)

ML FOR SFH – SIX EXAMPLES



- Example fits in the 5th, 20th, 40th, 60th, 80th and 95th percentile of the performance metric distribution.

EVALUATING MODELS



CHOOSING METRIC IS VERY IMPORTANT AND VERY TRICKY (AND YOU NEED TO KNOW YOUR DATA)

Initial metric: Median Absolute Error

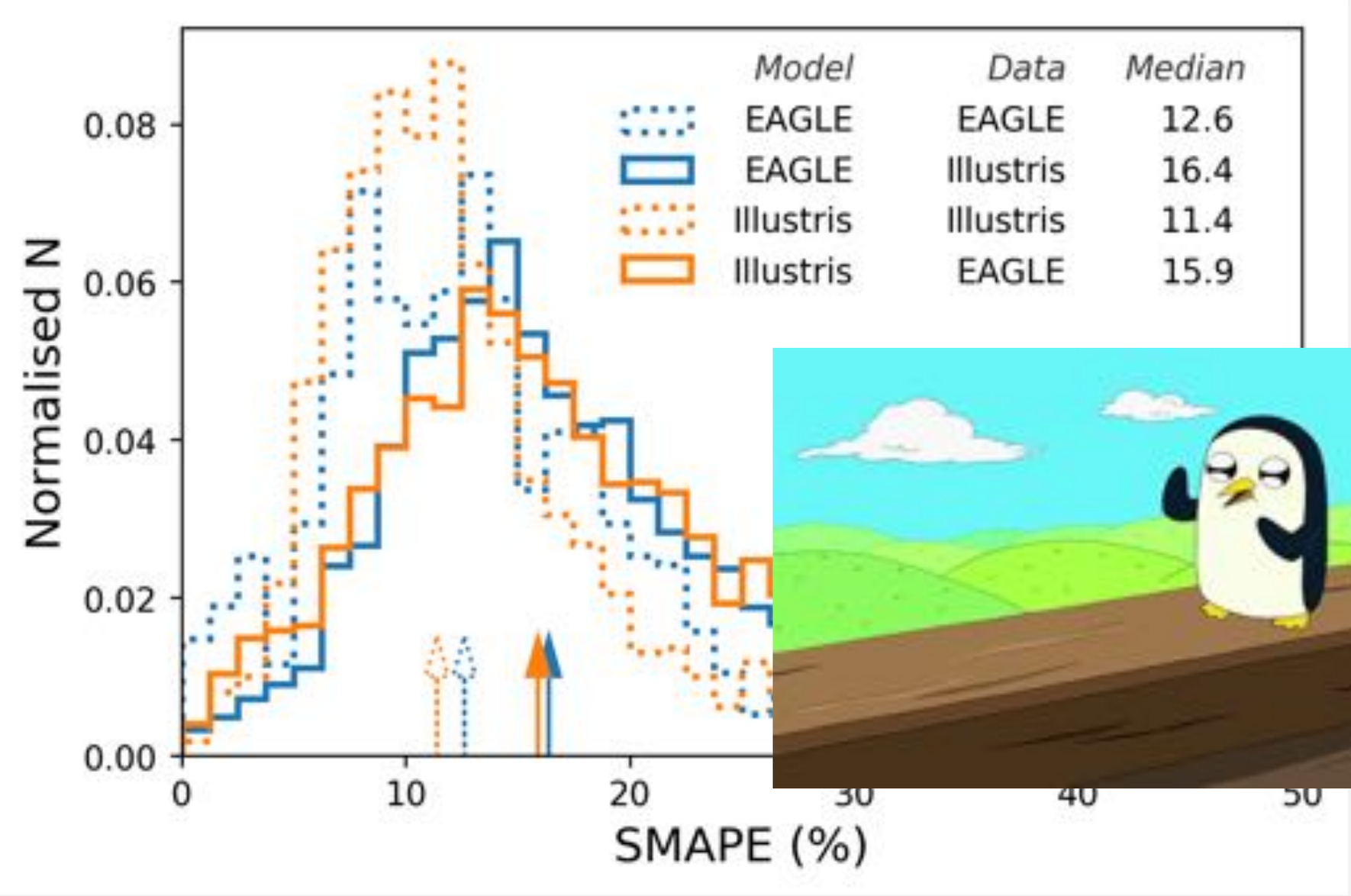
Second iteration: Median Absolute Percentage Error

Third iteration: SMAPE

Fourth iteration: SMAPE · output response

CHECKING FOR GENERALIZATION

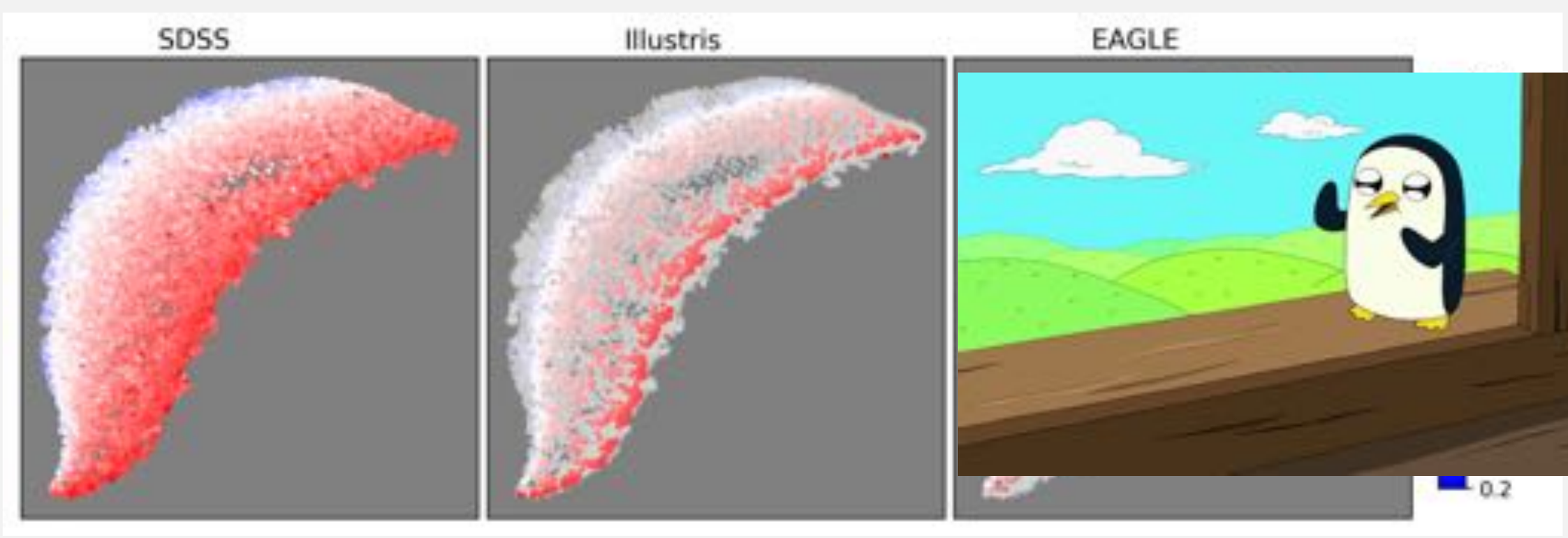
1. Let's look at what happens when we train on one simulation and test on the other.



CHECKING FOR GENERALIZATION

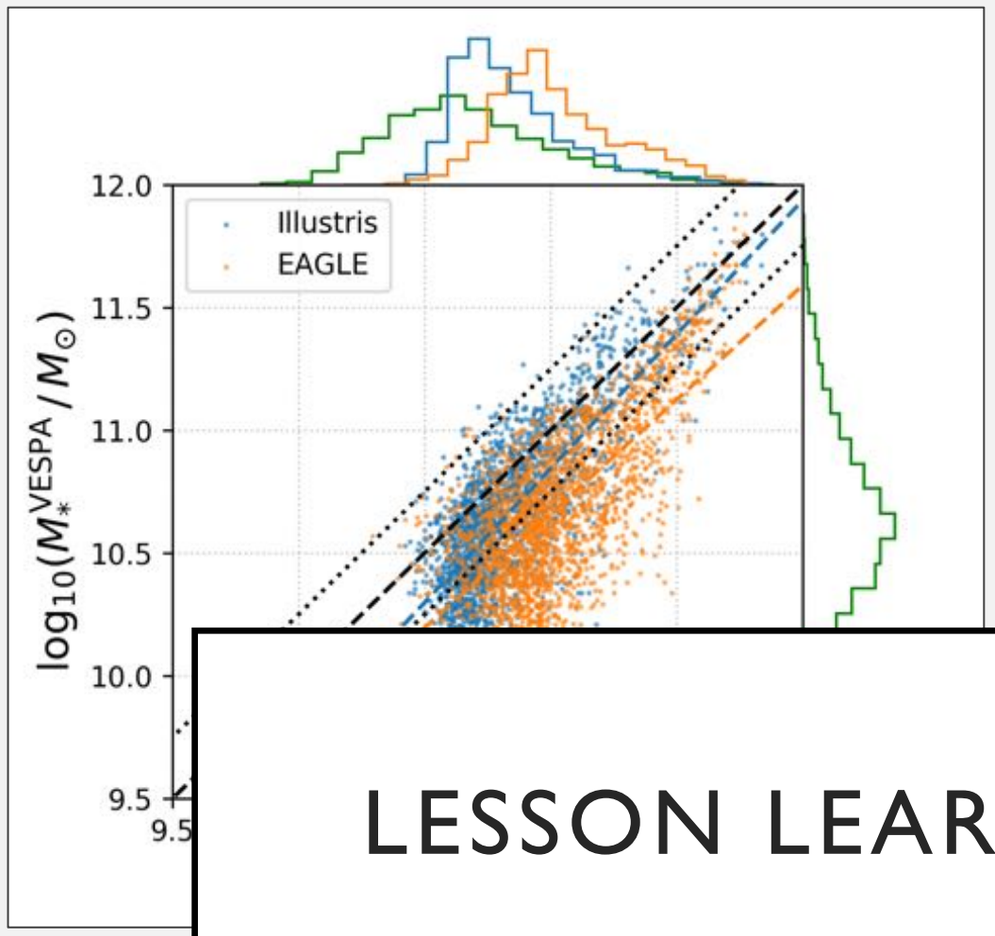
2. Are the spectral shapes of real and simulated galaxies similar?
And how do their distances compare to different sims?

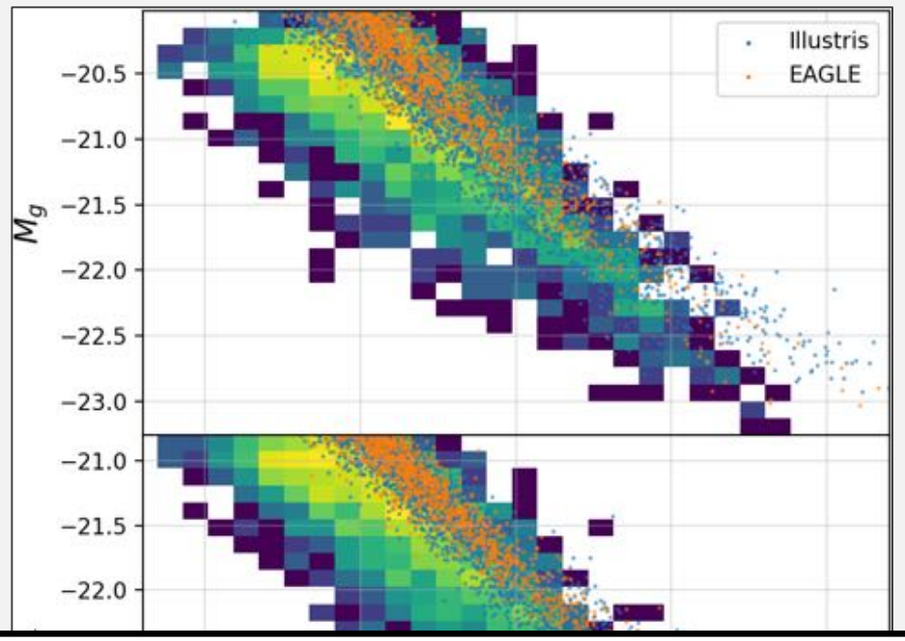
t-SNE project data to 2D keeping
"distance" as similarity measure



CHECKING FOR GENERALIZATION

3. If I get the SFHs right, I should get the total stellar masses right.
   Is it true?



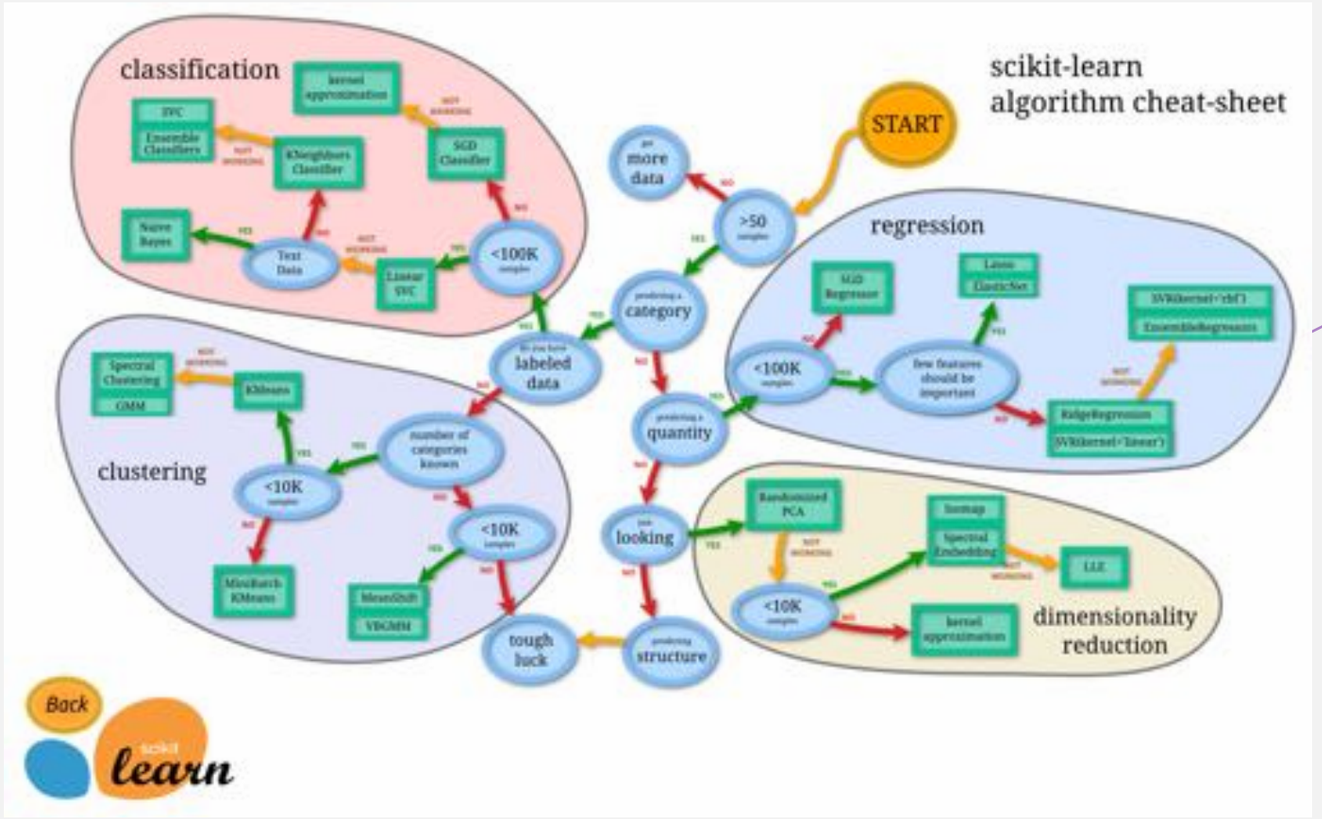


LESSON LEARNED: TEST TEST TEST.

SCIENTIFIC METHOD RULES!

NEXT STEPS FOR ML

IN ASTRONOMY AND BEYOND



VALIDATION/
ADAPTATION

INTERPRETABILITY

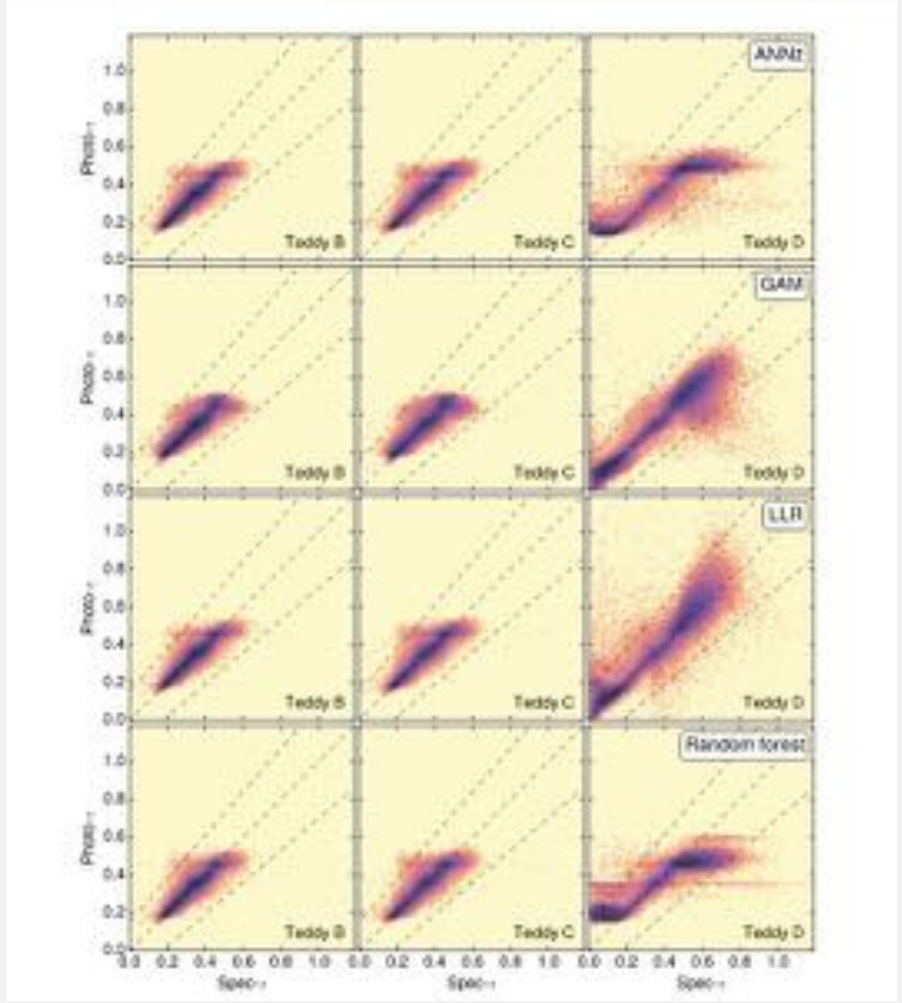
VALIDATION CHALLENGES

1. How can we devise tests when ground truth is not available?

2. What happens when the training and application domain are different?

1. How can we compensate for differences?
2. How can we assess the expected performance?

(note: everything algorithm dependent!)



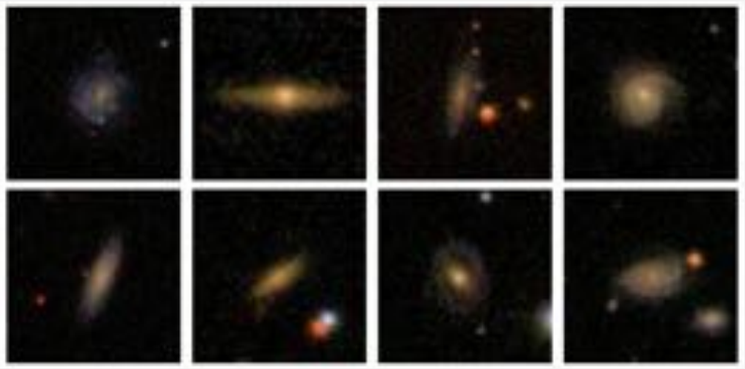
Beck et al 2016

TRANSFER LEARNING

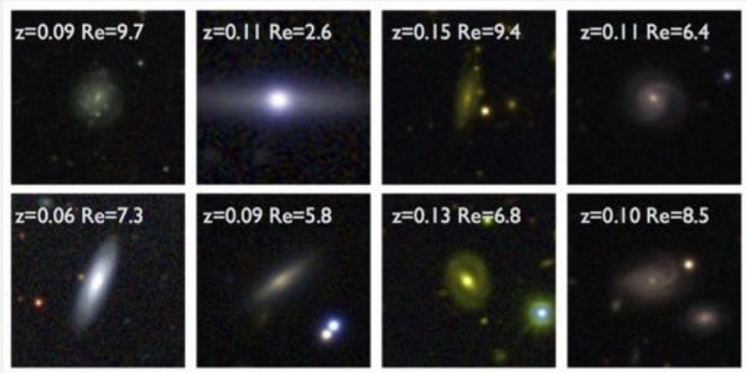
Large training sets are hard to come by +
most models rely on deep neural networks which are very expensive to train

Can domain adaptation be used for TRANSFER LEARNING (recycling knowledge?)

Dominguez-Sanchez et al 2018



SDSS galaxies



DES galaxies
(different depth, PSF, seeing)

Significant improvement in accuracy (90 → 95% if domain adaptation step is included);
Training set size improvement of one order of magnitude for given performance.

INTERPRETABILITY / EXPLAINABILITY

Molnar 2016, Interpretable ML book

LIME (Locally Interpretable Model-Agnostic Explanations, Ribeiro et al 2016) uses surrogate simpler models to gain insights on the decision making of an ”opaque” algorithm such as a neural network

SHAP (SHapley Additive exPlanations, Lundberg et al 2018) uses game theory to understand the interactions between different features (players)

Recourse Analysis (Ustun et al 2018), aims at understanding what change in features would cause a decision to be reversed

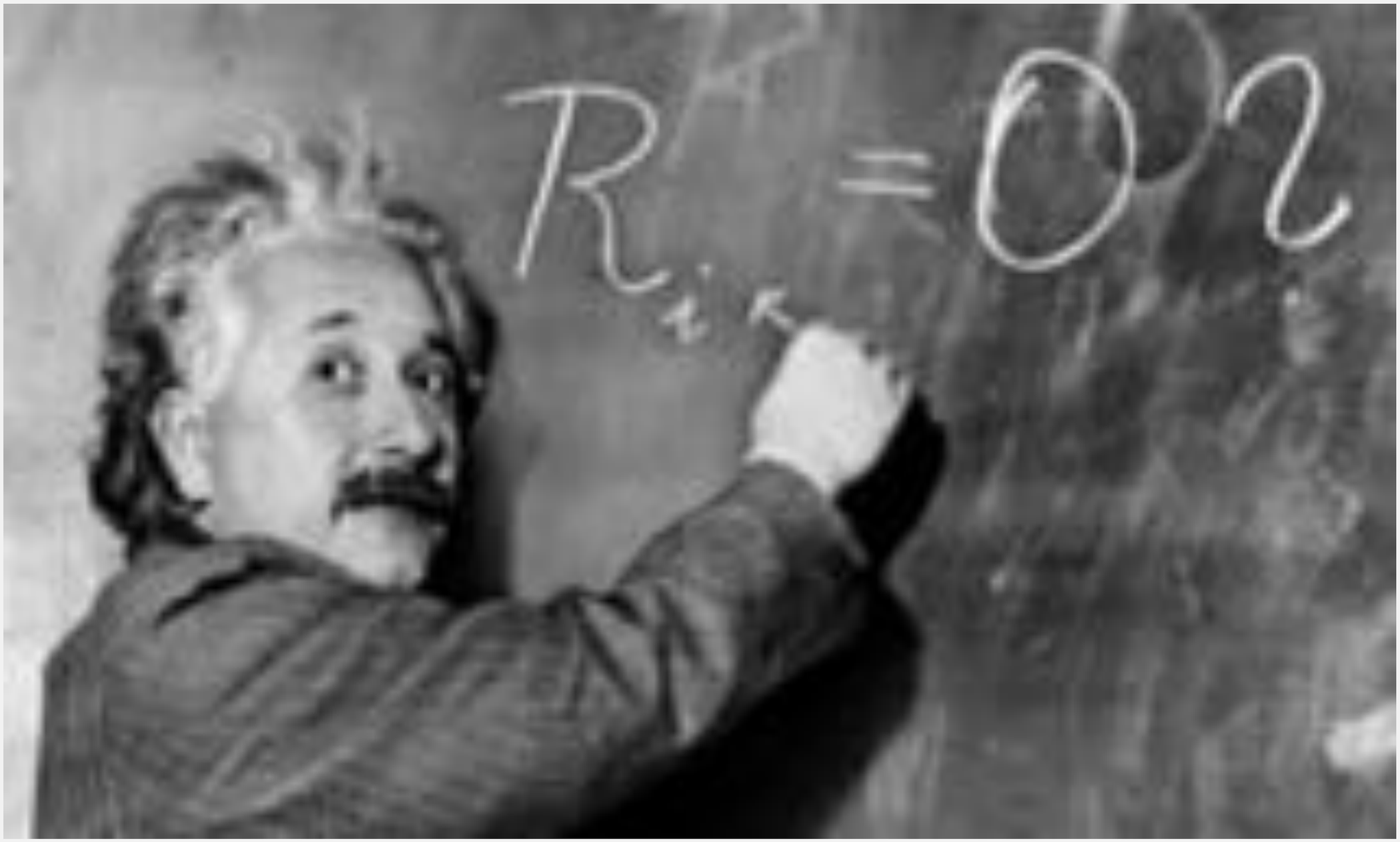
However:

Science has much higher standards for explainability

WHAT DOES IT MEAN

TO BE AN (ASTRO)PHYSICIST

TODAY?



PROBABLY SOMETHING
DIFFERENT FROM THIS.

SOME FINAL THOUGHTS

Data analysis in Astronomy is changing radically because we now have better, bigger, wider, and new data.

This is good and it creates new (exciting) responsibilities. This is my wish list.

As scientists, we need to have a broad understanding of computational tools, to commit to the scientific method more seriously than ever, and to hold to a high standard for reproducibility and sharing.

As mentors, we need to encourage cross-disciplinary collaborations and be aware of the ever-changing job market.

As reviewers/funding agencies, we need to be open to higher-risk, higher-reward tools.

It's not easy but I think it's worth it.

THANK YOU!