Powered by Rainbows, Stars, and Machines: A Rapid and Inexorable AI Revolution in Galaxy Science

Joel Leja

Assistant Professor of Astronomy & Astrophysics

Institute for Computational & Data Sciences

PennState Institute for Computational and Data Sciences

PennState

images: **NASA / ESA**

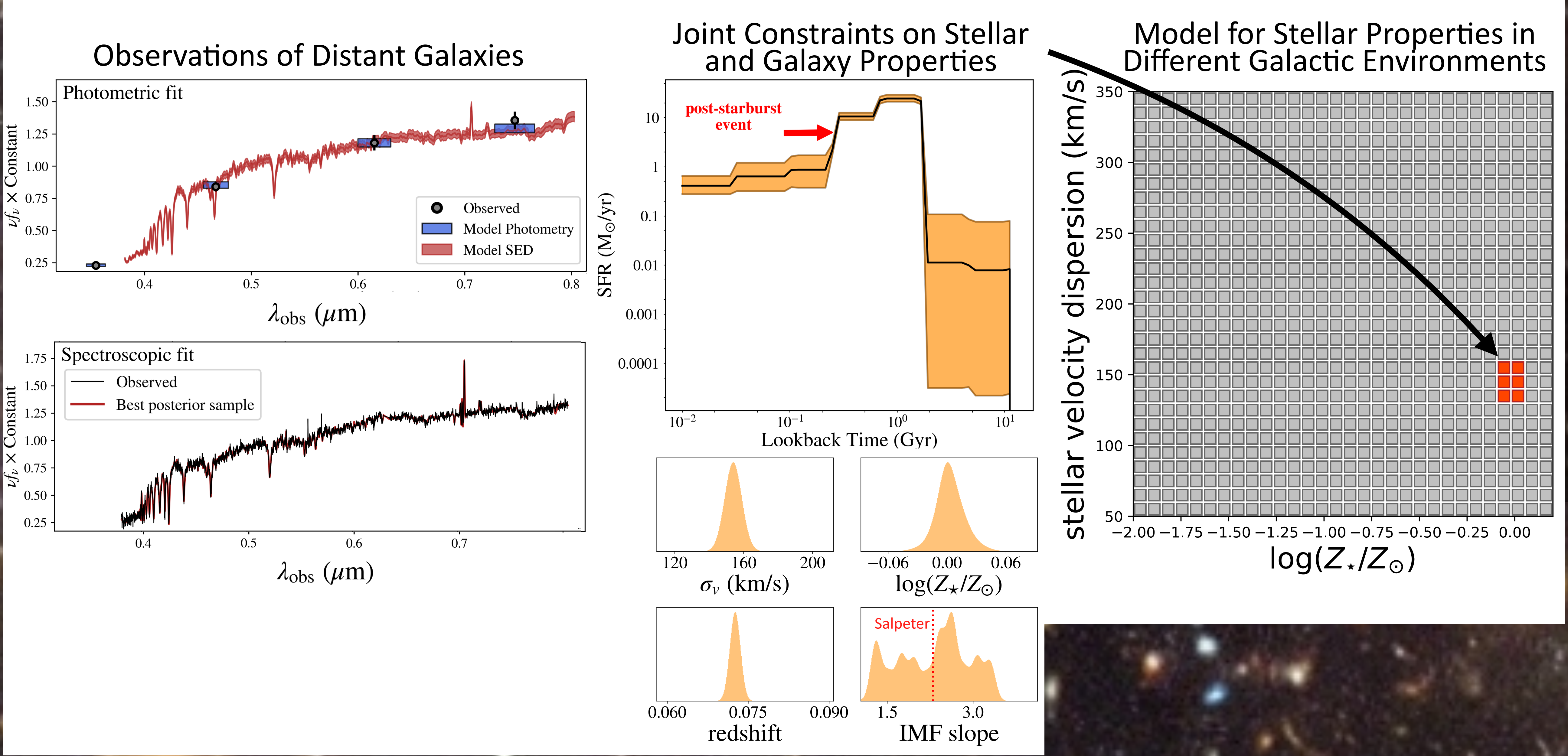
Galaxies are **cosmic ecosystems**

- host nearly all star formation
- forges for heavy elements
- processing centers for cosmic gas
- homes for black holes, transient events, planets
- trace expansion and large-scale structure of the Universe

Big Research Questions

- When and how do galaxies of all masses and types

Galaxies as Laboratories for Physics in Exotic Environments



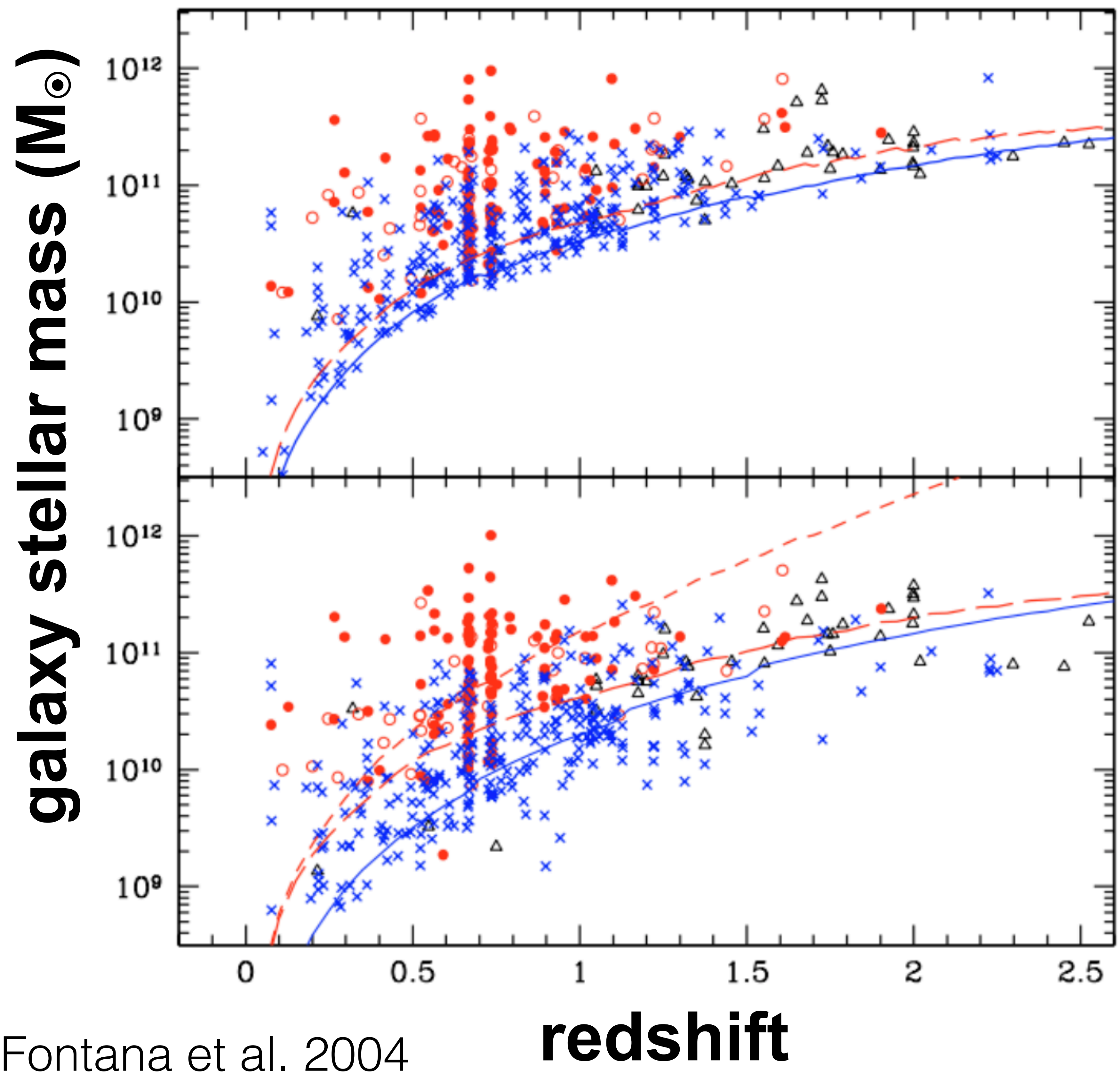
I: State of the Field

II: A New Paradigm From High-Dimensional Models

III: A Million Times Faster with Machine Learning

IV: Surveys of the Future

20 Years Ago, Observing Galaxies in the Distant Universe Was a **New Frontier**

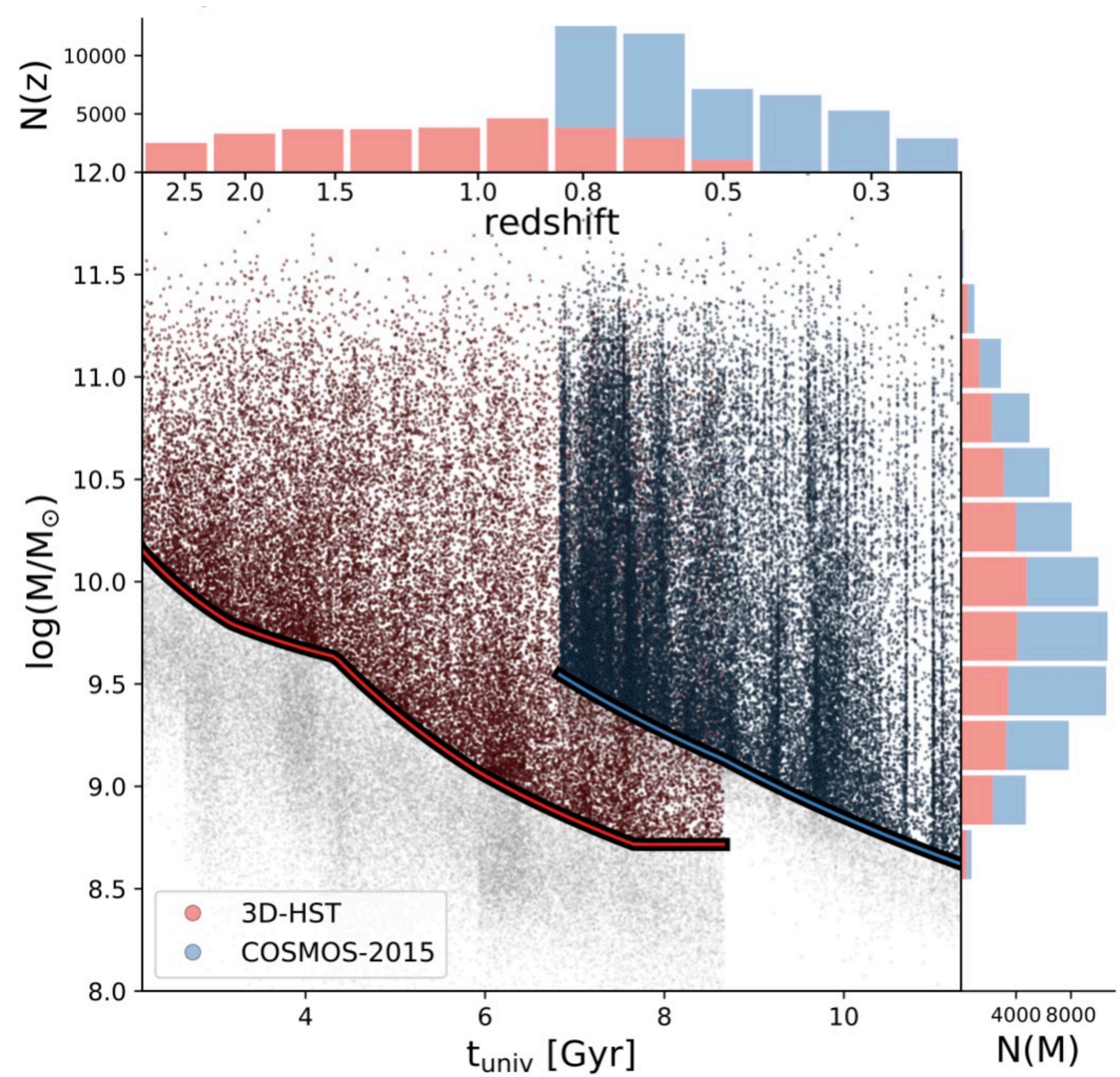


Fontana et al. 2004

- **Relatively shallow survey depths**
- **Restricted wavelength coverage** (<1μm)
- **Strong selection function**

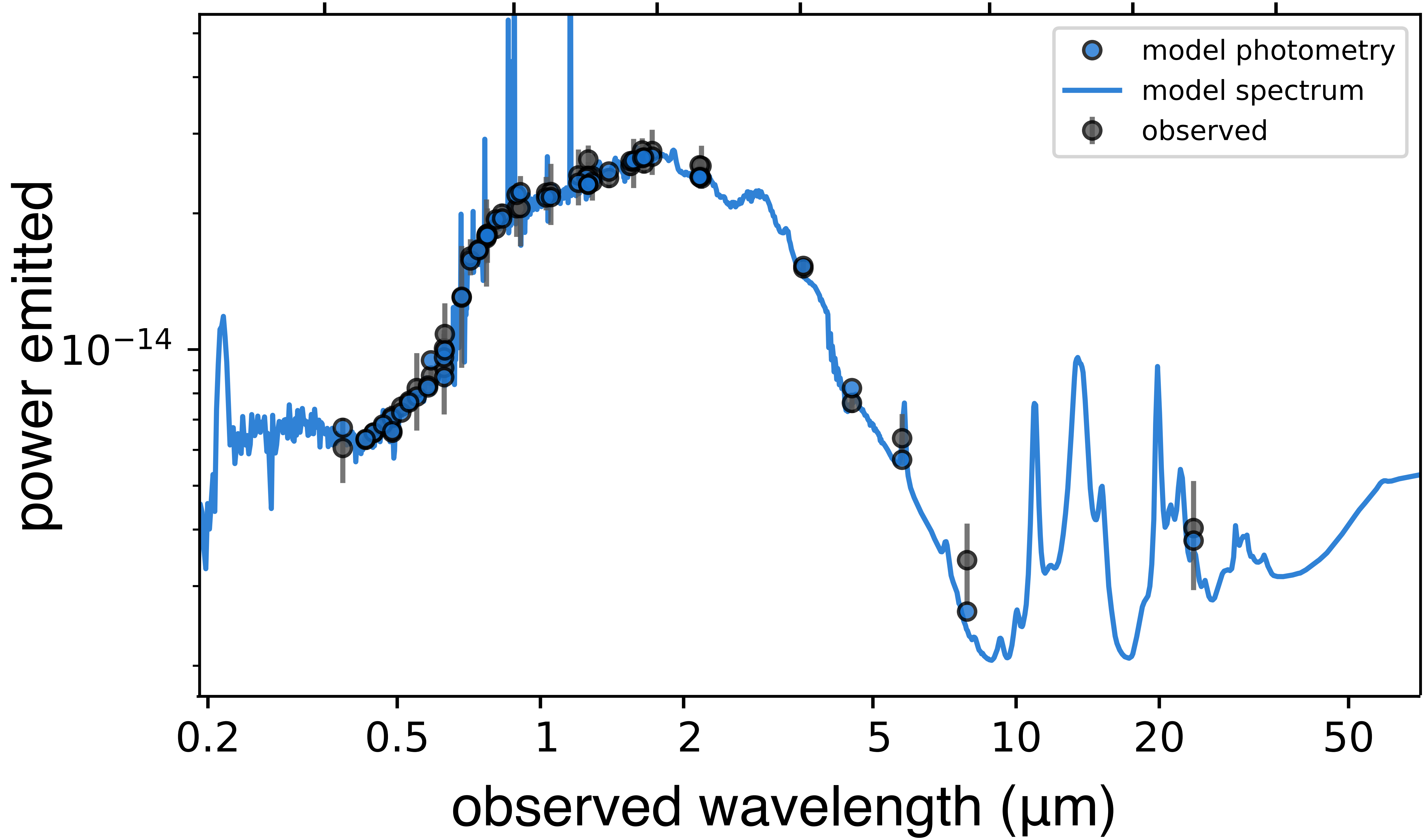
Today, The Census of Galaxies in the Universe is **Nearly Mature**

Surveys now provide **deep, complete** samples, covering ~10^5 galaxies over 85% of cosmic time



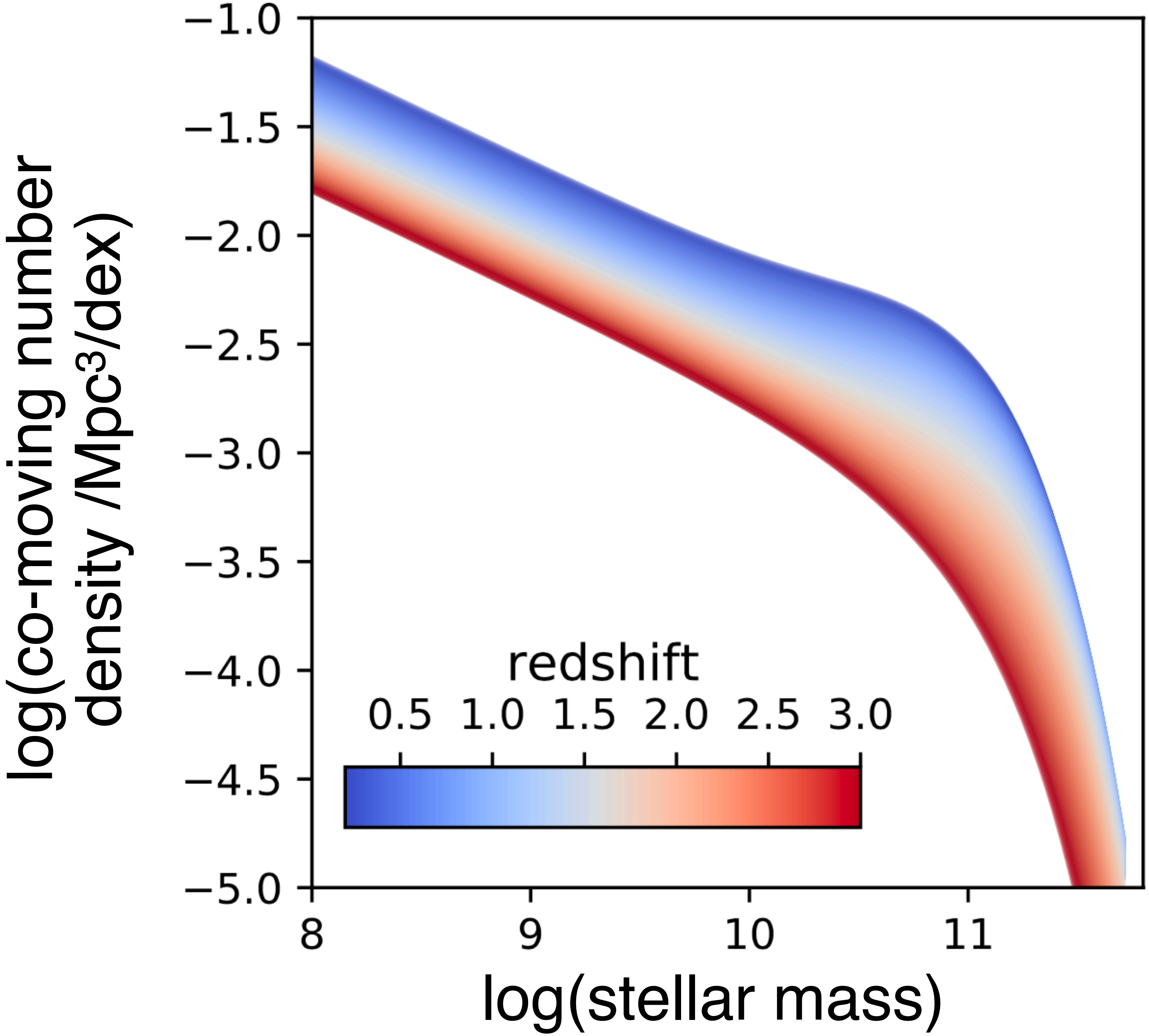
Leja et al. 2020

Galaxy surveys are multiwavelength, with up to 30-40 bands of UV-IR photometry and **well-measured** redshifts



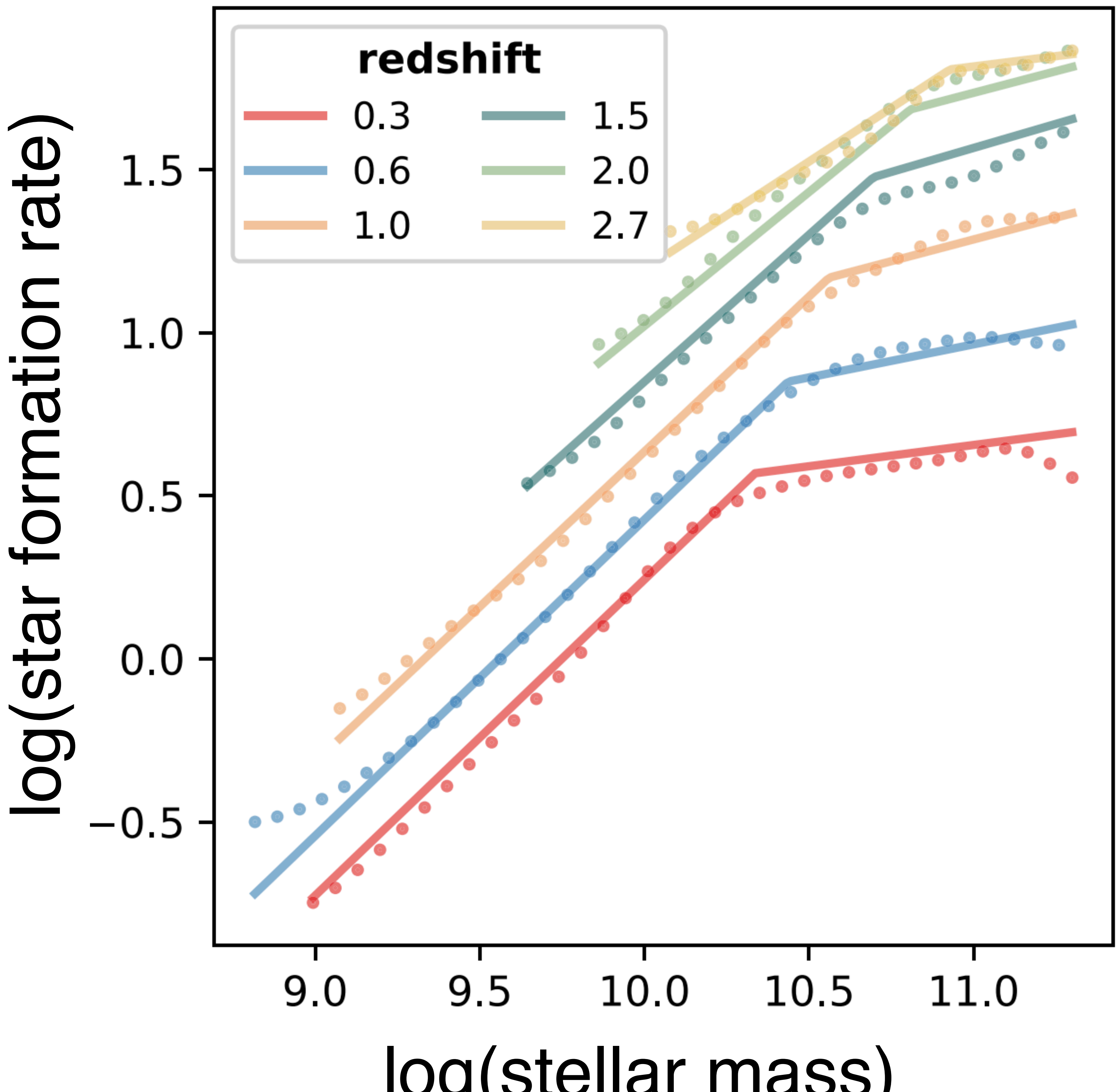
Leja et al. 2019b

Deep galaxy stellar mass functions suggest **~95% of existing stellar mass** has been surveyed over 85% of cosmic time.



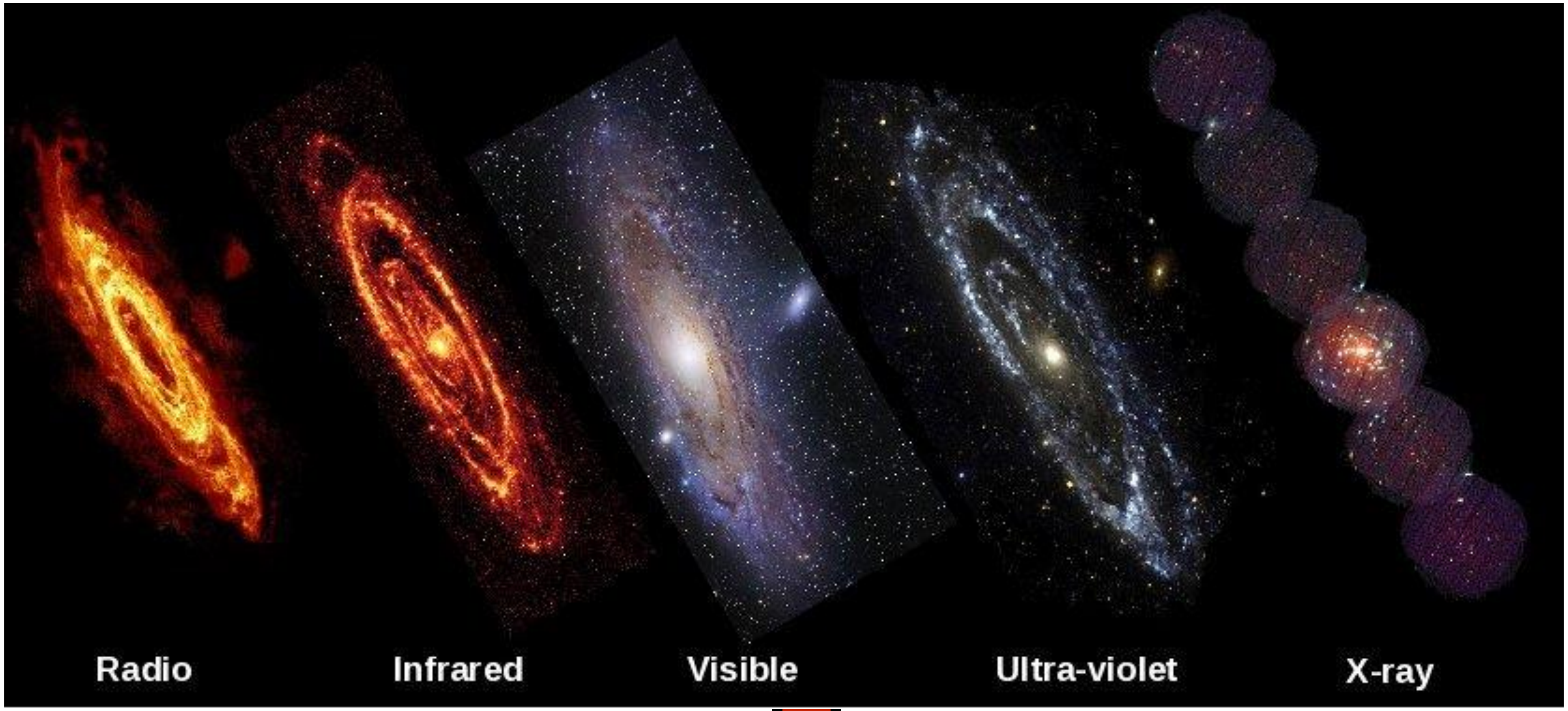
Leja et al. 2020

Ultraviolet, mid/far-infrared, and nebular emission line surveys have charted **~75-80% of star formation** over **~85% of cosmic time.**



Leja et al. 2022

The data are processed by fitting spectral energy distributions (SEDs). Take beautiful galaxy data:



The Andromeda Galaxy
Planck / NASA / ESA

… and use models to turn them into (*even more beautiful*) inferred parameters.

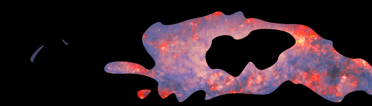
stellar mass	dust content
star formation history	chemical abundances
nebular properties	active black holes

Two Basic Ways to Infer Stellar Assembly from Observations

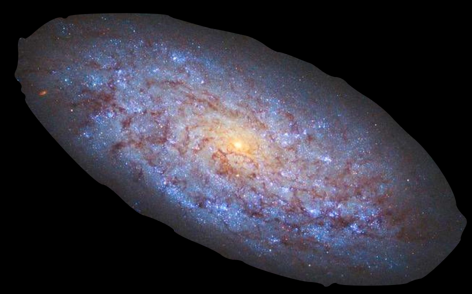
growth of
stellar mass

total instantaneous
star formation

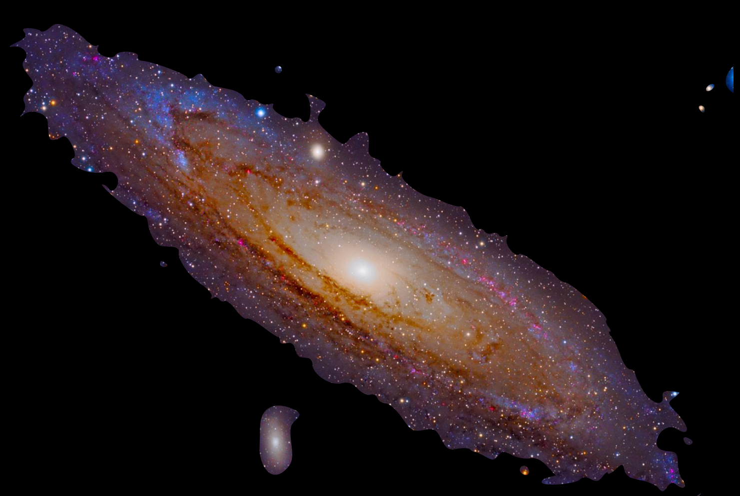
z=2

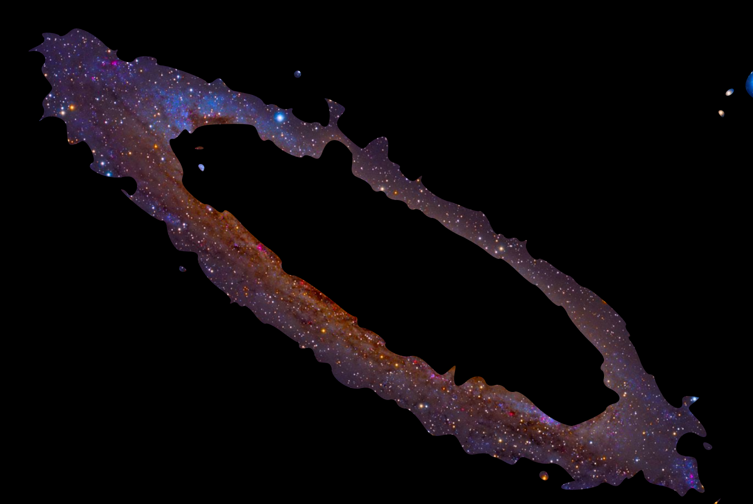


z=1



z=0

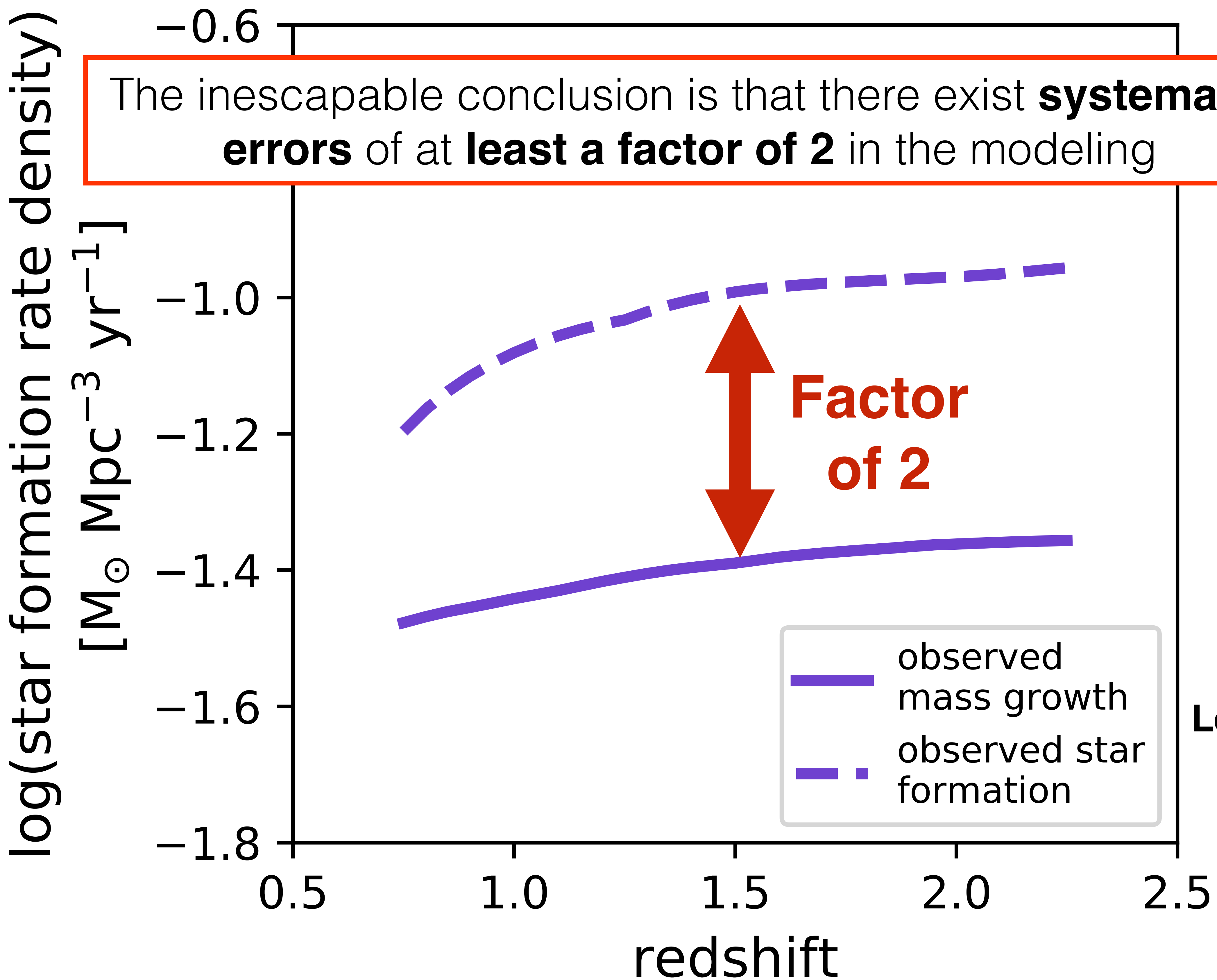




These are **integral / derivative** pairs
and typically inferred from ~independent parts of the EM spectrum

A Universe that Doesn't Add Up

The inescapable conclusion is that there exist **systematic errors** of at **least a factor of 2** in the modeling



Leja et al. 2015

also see Madau+14, Tomczak+15, Contini+16, Yu & Wang 2016, Behroozi+19

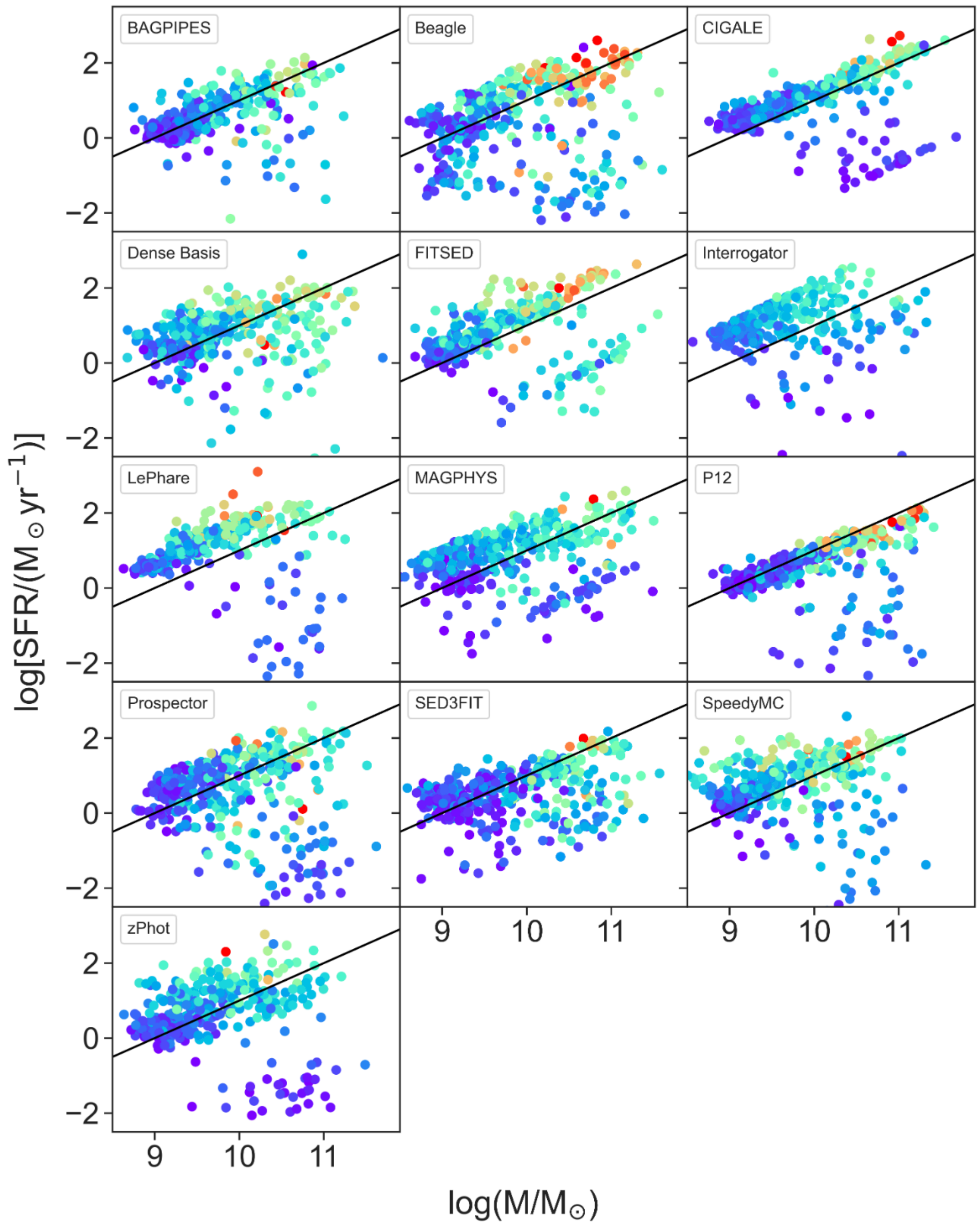
The Problem is in the Modeling

An model-fitting experiment

different galaxy SED-fitting codes applied to…

…**identical** high-quality UV-NIR HST photometry…

… produce **very different** relationships between star formation rate and stellar mass!



Pacifici et al. incl **Leja**, 2022

I: State of the Field

II: A New Paradigm From High-Dimensional Models

III: A Million Times Faster with Machine Learning

IV: Surveys of the Future

So, What Is The Path Forward?

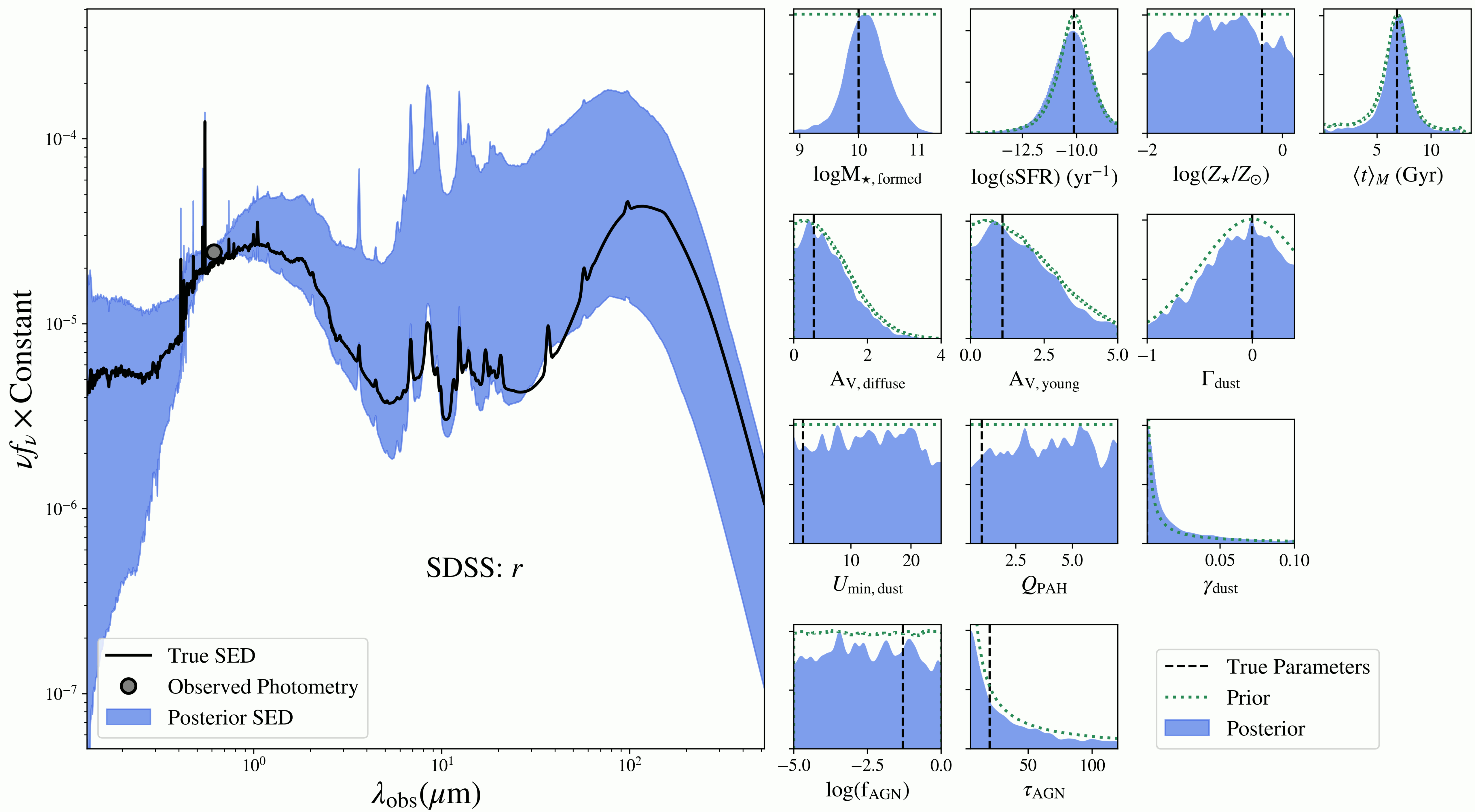
Many different types of physics affect the observations, while data give limited constraints. This forces big approximations, which create systematics.

Models must **permit more variation in physical properties**; simple models do not cover the complex+messy process of galaxy formation.

SED Parameter	approximate effect on SED

Prospector: A Bayesian Galaxy SED Fitter

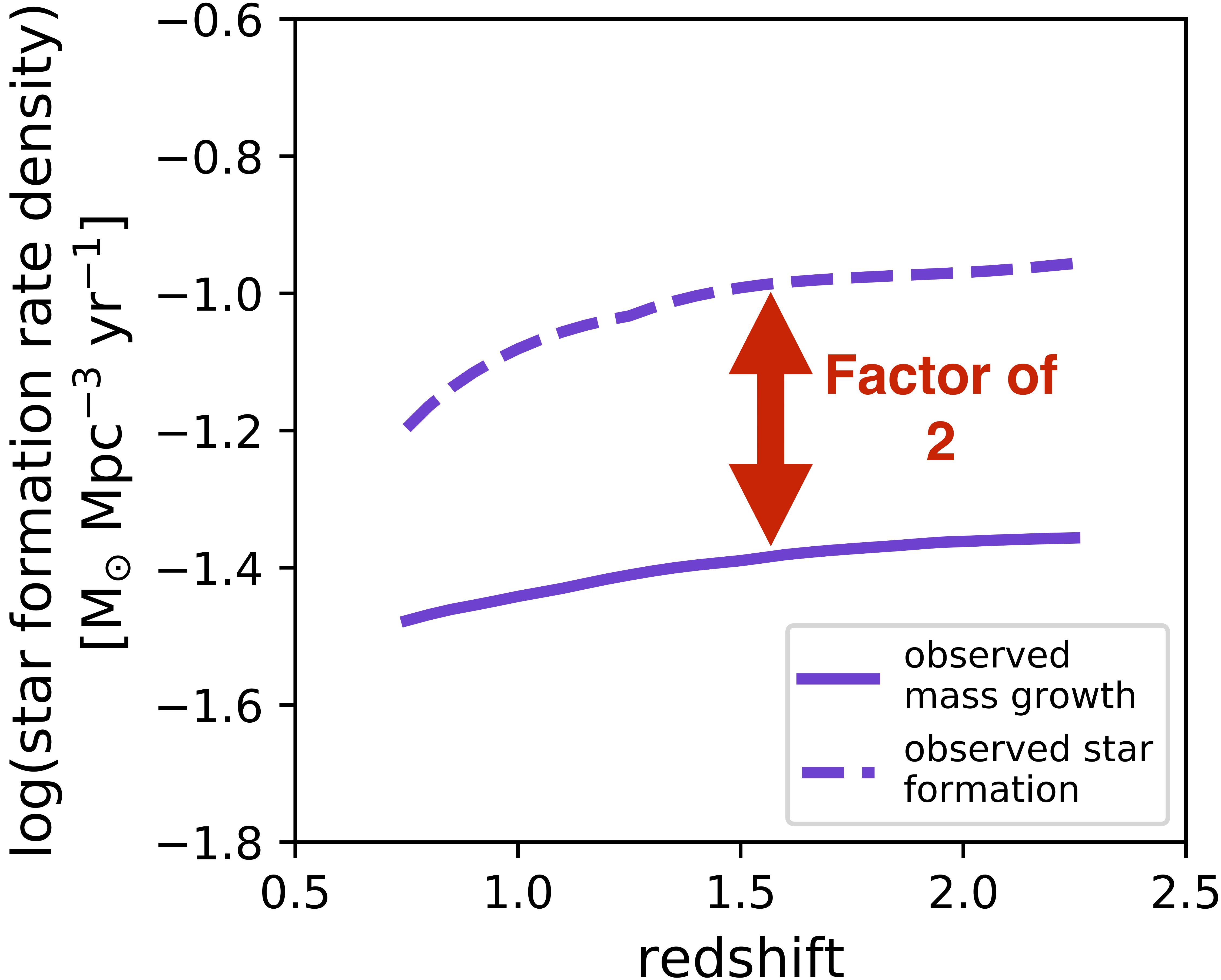
Prospector is an open-source package which fits gridless stellar populations models to galaxy observations (spectra and/or photometry). Access to 100+ parameters controlling physics in galaxies.



this fit includes 18 parameters and **nonparametric** star formation histories

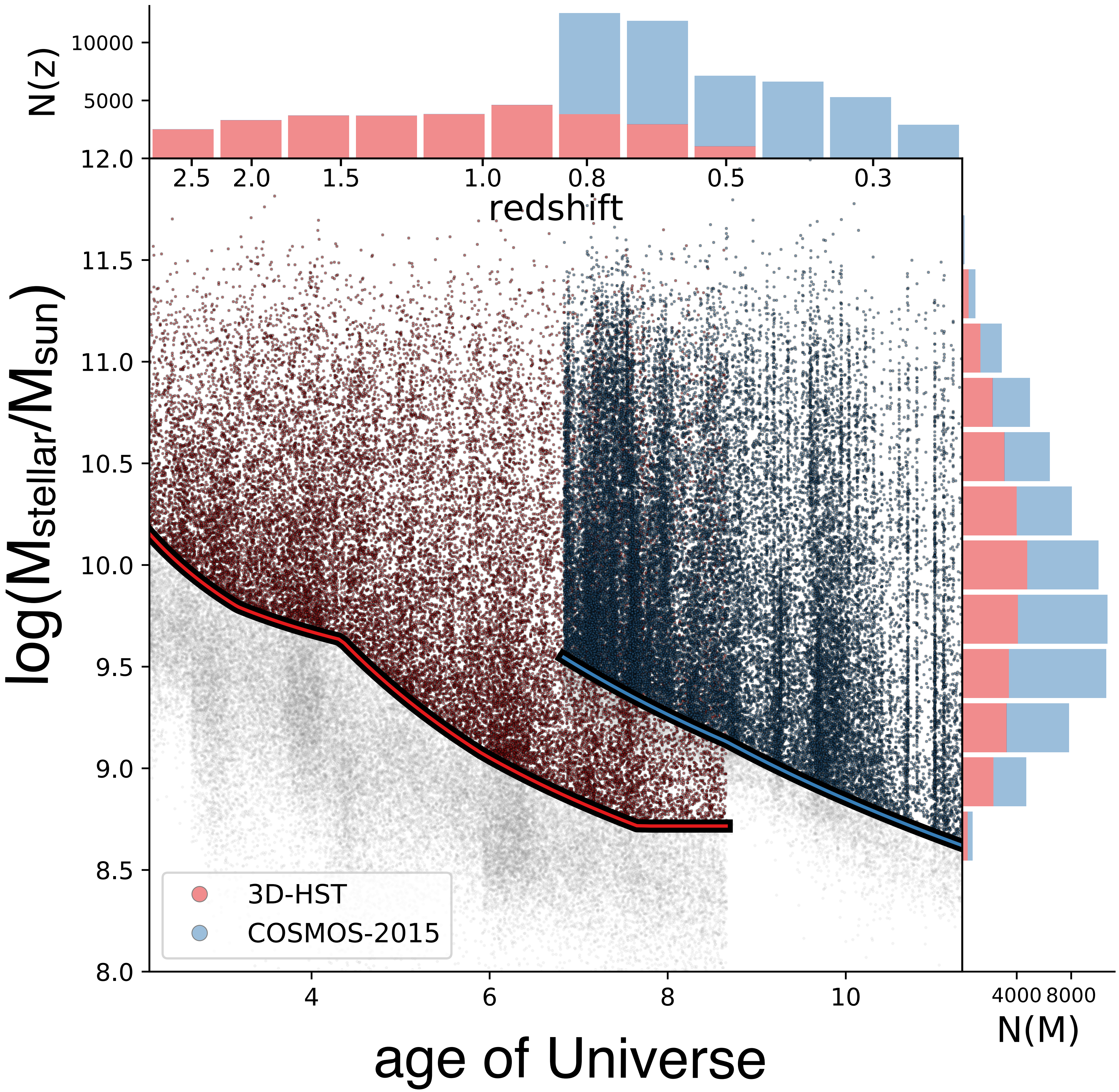
Johnson, **Leja** et al. 2021, **Leja** et al. 2017

Can We Fix the Universe with Better Models?



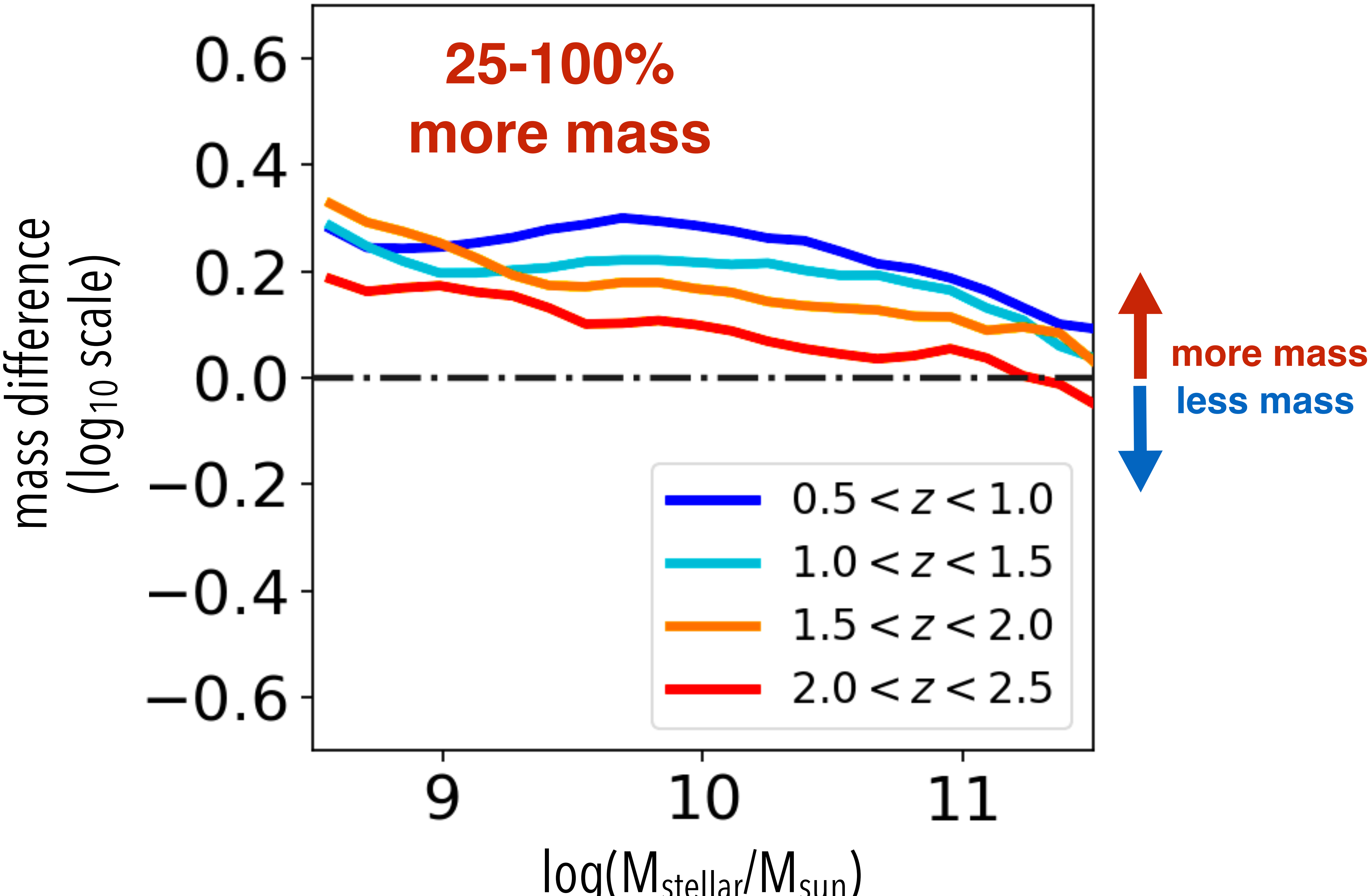
Fitting a Cosmological Sample

I fit the photometry of ~100k galaxies from two modern galaxy surveys with ***Prospector***



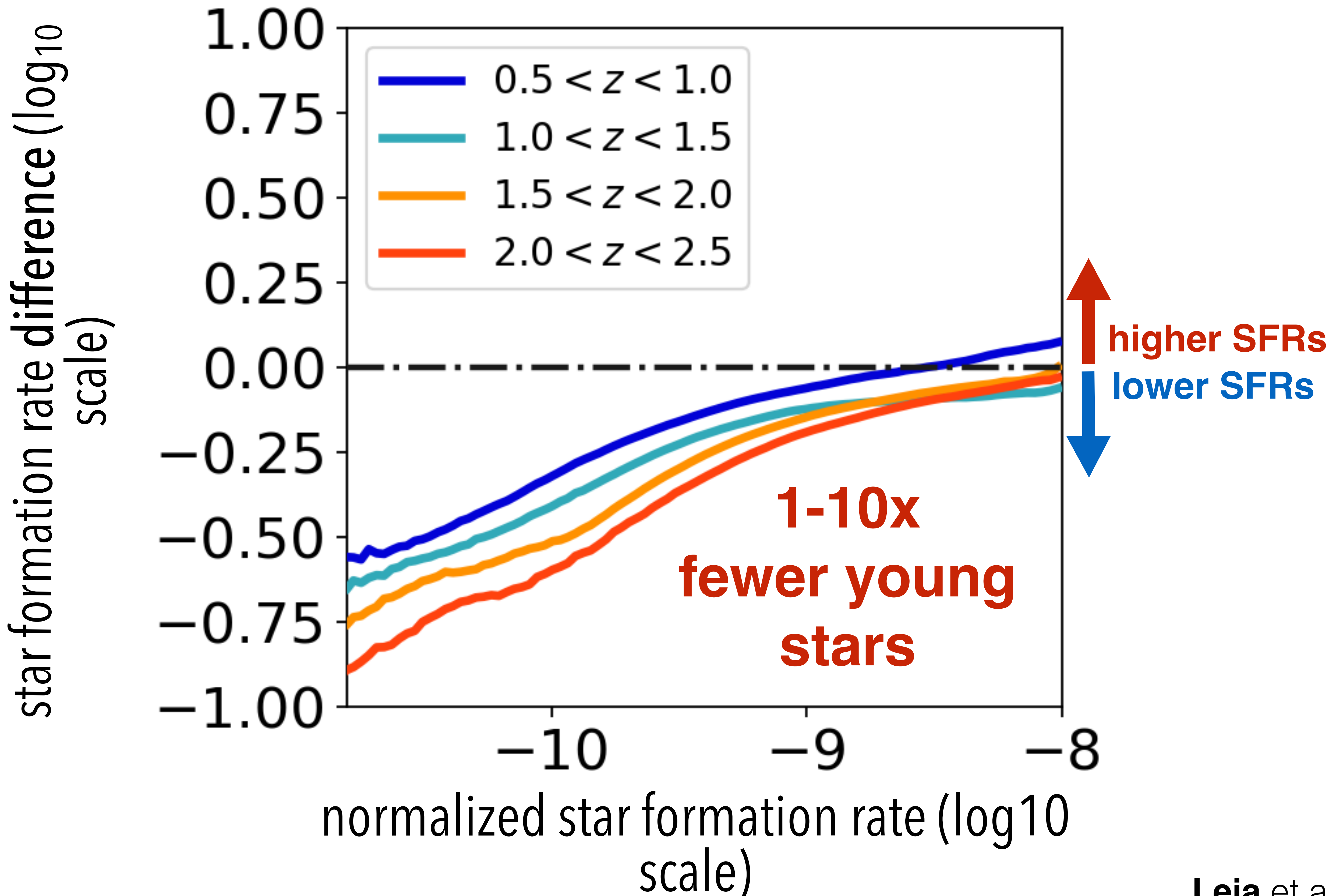
Leja et al.
2020

Surprise #1: There's a Lot More Mass in Stars



Leja et al. 2019b

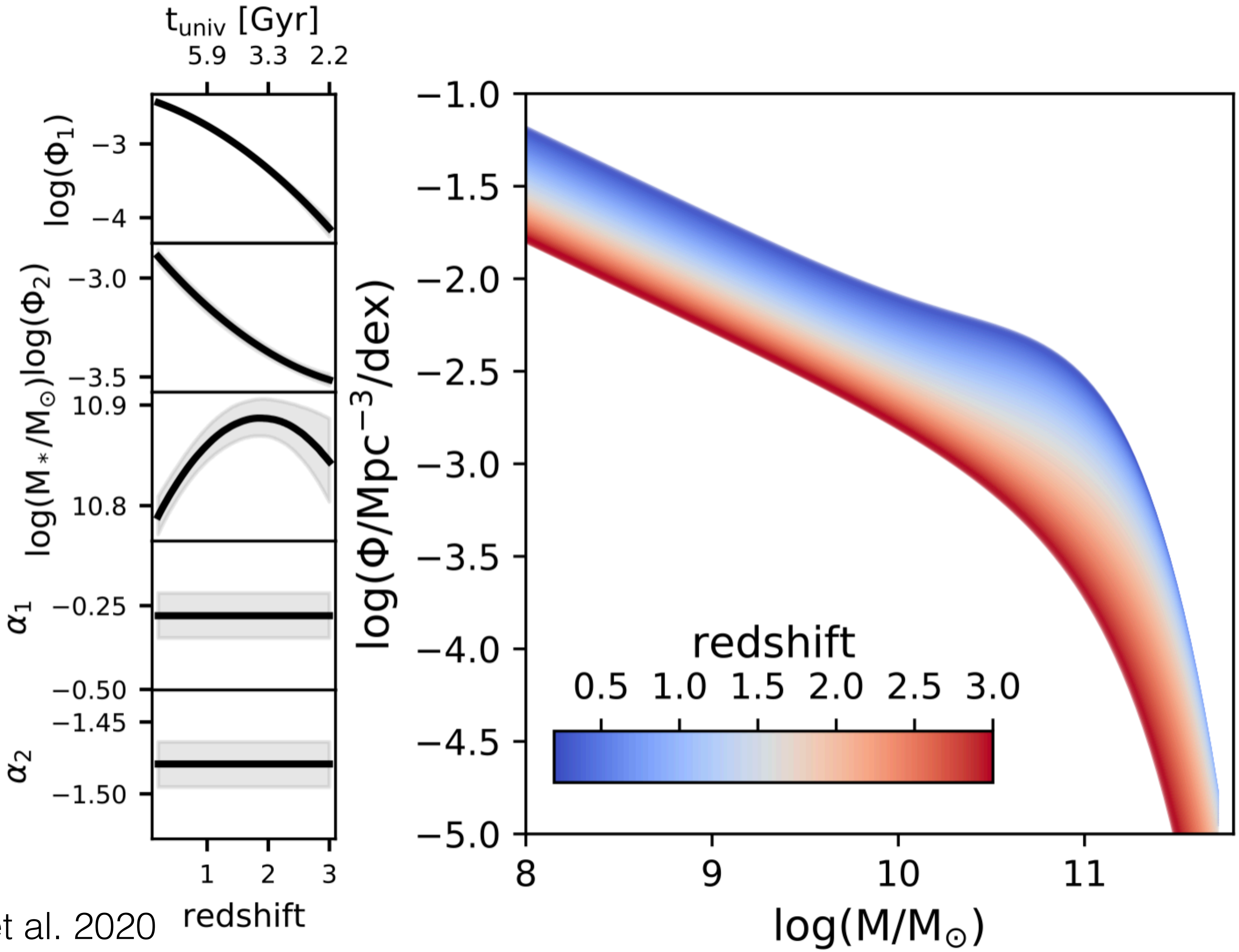
Surprise #2: There's a Lot Less Ongoing Star Formation



Leja et al. 2019b

Weighing Mass with the Stellar Mass Function

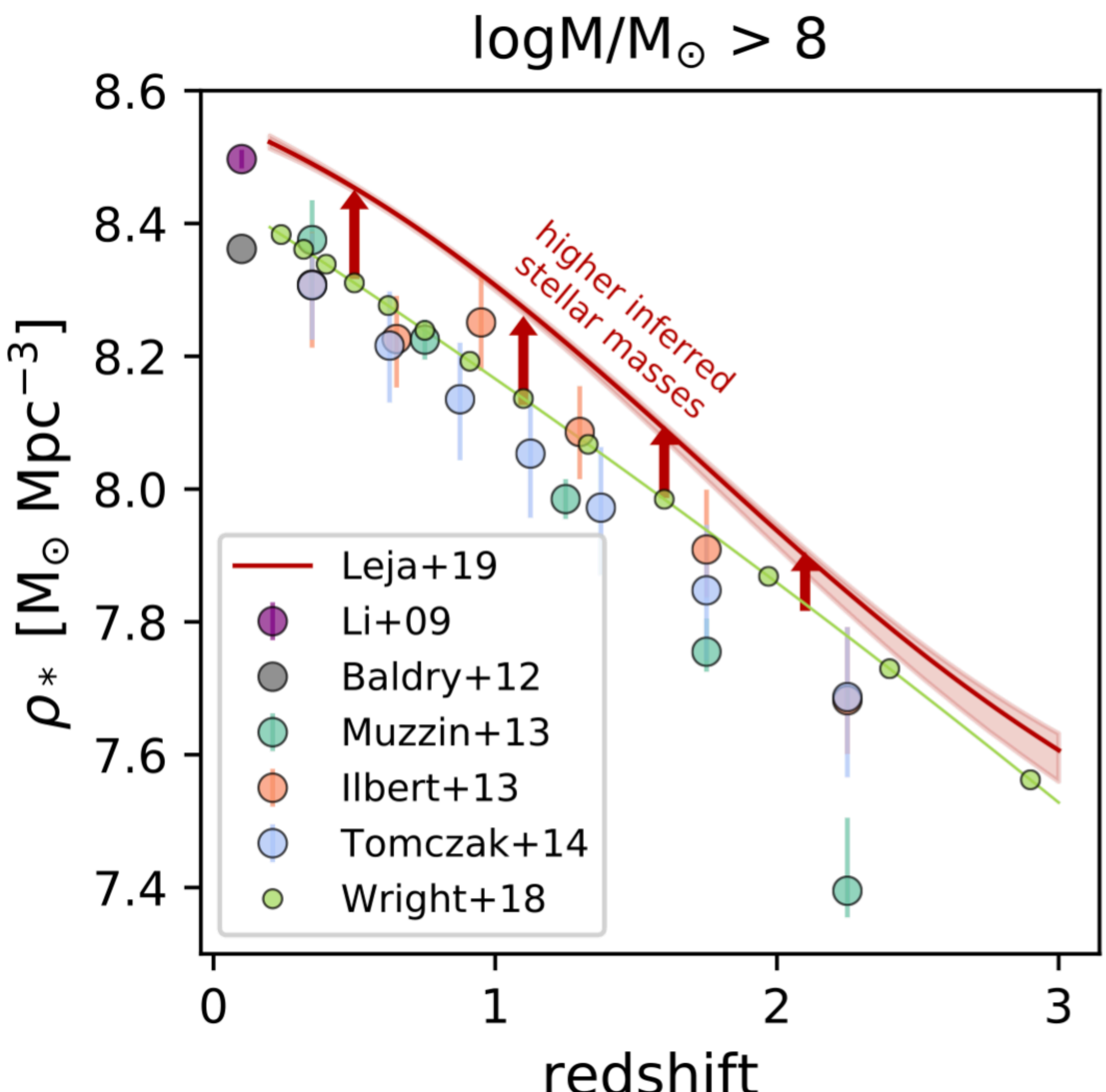
Inferred using a **Bayesian hierarchical model:** ensures smooth evolution, fit for cosmic variance directly, use full constraints on stellar mass.



Leja et al. 2020

An Older, More Evolved Universe

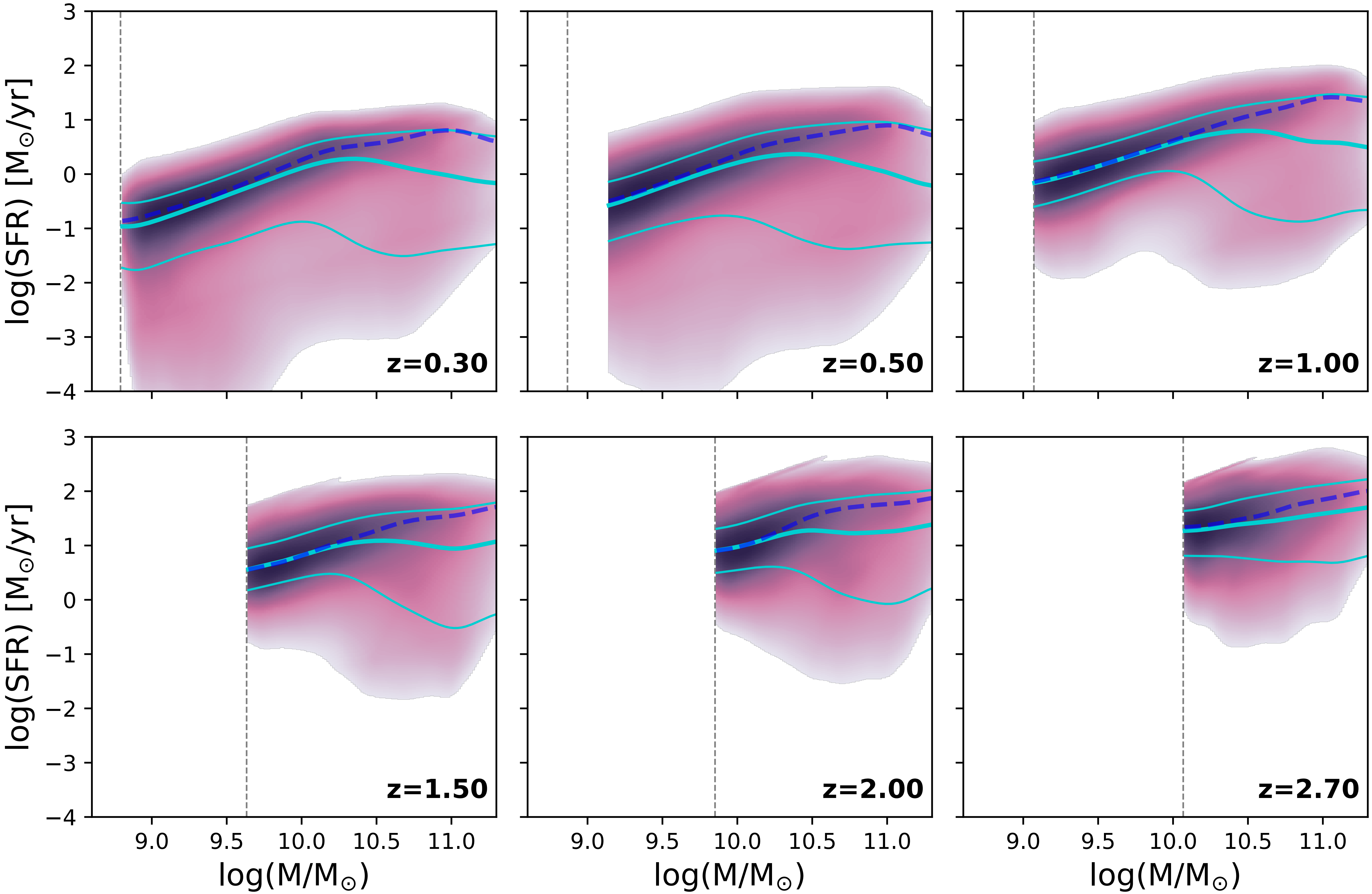
I find a higher cosmic stellar mass density by **0.1-0.2 dex** (30-60%), with the derivative maximized at **z~1.5**.



Leja et al. 2020

Learning The Star-Forming Sequence Directly

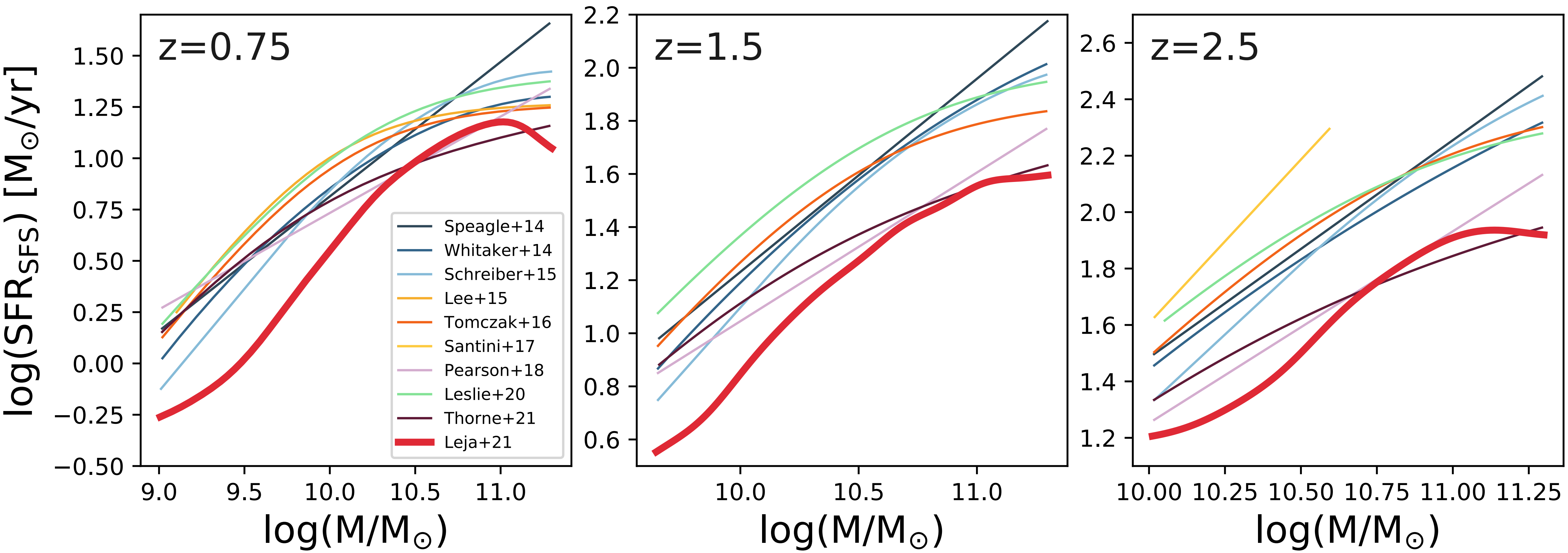
I use a **normalizing flow** to model P(mass, star formation rate, redshift) directly. I perform a novel modification to incorporate measurement errors.



Leja et al. 2022

A Novel View of the Galaxy Star-Forming Sequence

I find that galaxies are forming stars at a rate **0.2-0.5 dex** below other studies, with the offset peaking at 1.5 < z < 3.0.

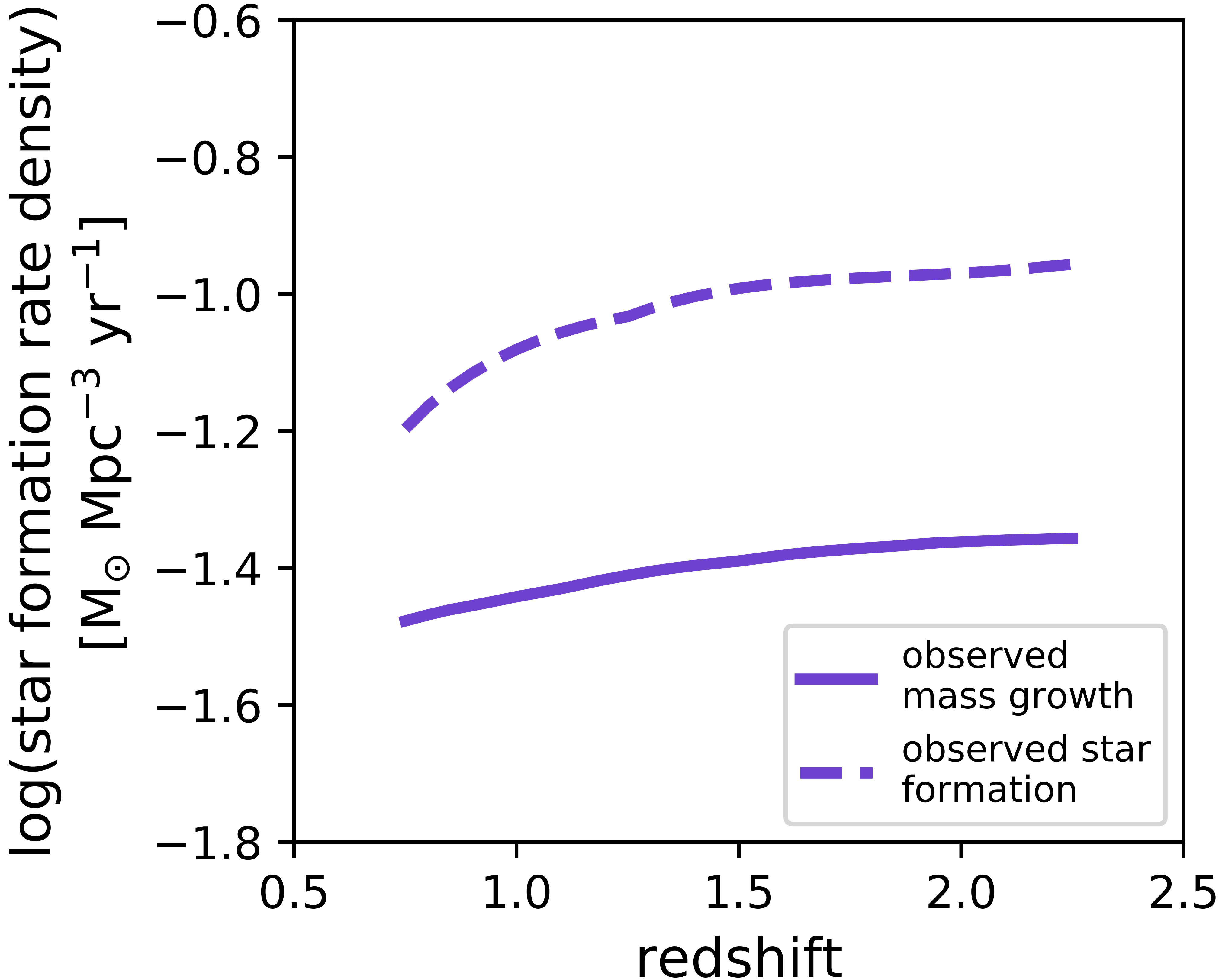


Offset caused by **higher masses** and **lower star formation rates**, a natural consequence of the more extended formation histories found by Prospector

Leja et al. 2022

The Problem is the Modeling

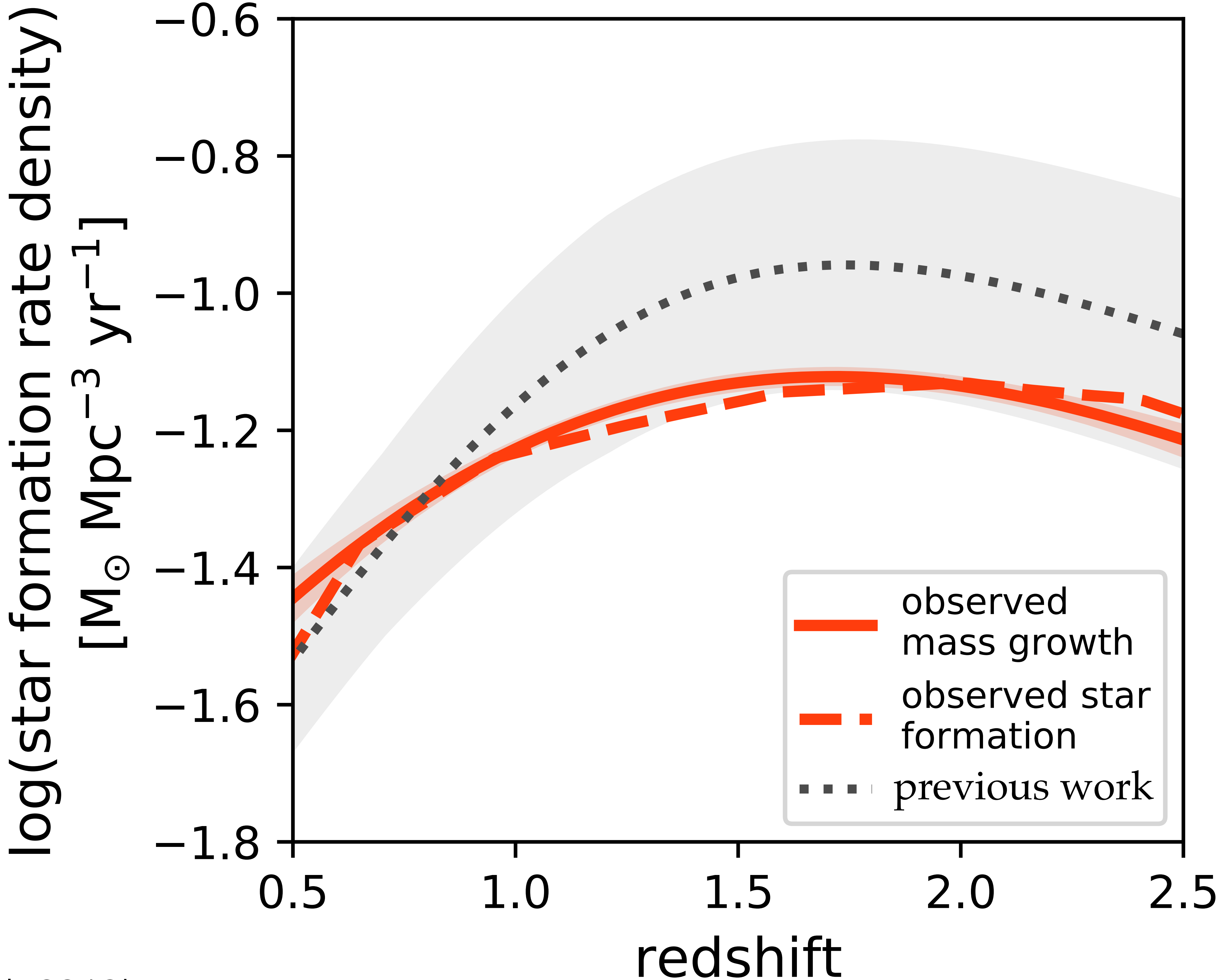
Previously, disagreement implied **systematic 2x uncertainty on the rate of galaxy assembly**.



Leja et al. 2019b

A New-found Cosmic Consensus

The high-dimensional *Prospector* modeling creates **new agreement**.
and a considerably flatter cosmic formation history!

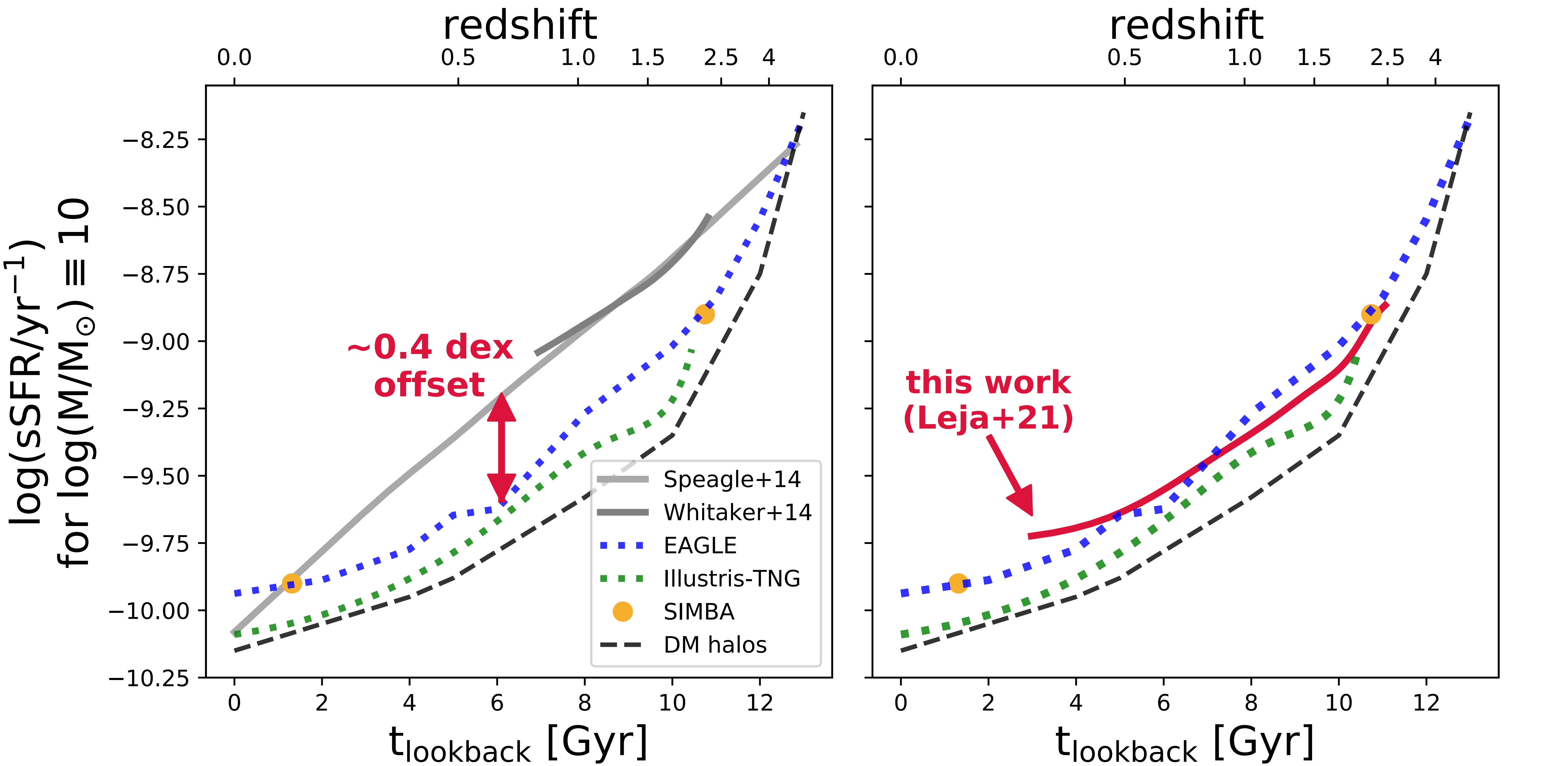


Leja et al. 2019b

This Solves A Long-Standing Disagreement With Simulations

To match observed SFRs, previously simulations needed to invoke **exotic forms of feedback** to decouple accretion and star formation (e.g. Mitchell+14).

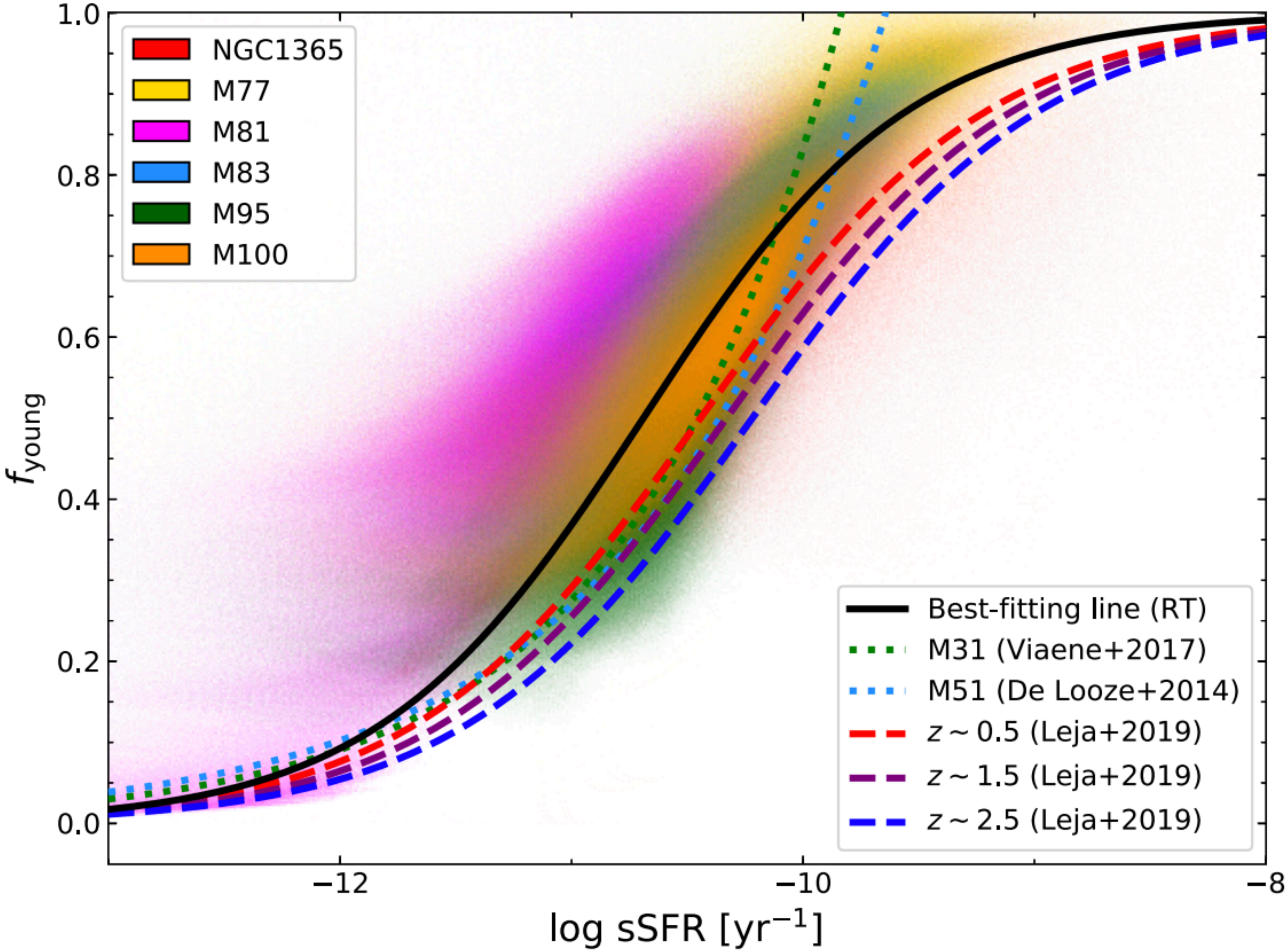
This is *no longer necessary*.



Leja et al. 2022

Agreement with Full Radiative Transfer Models

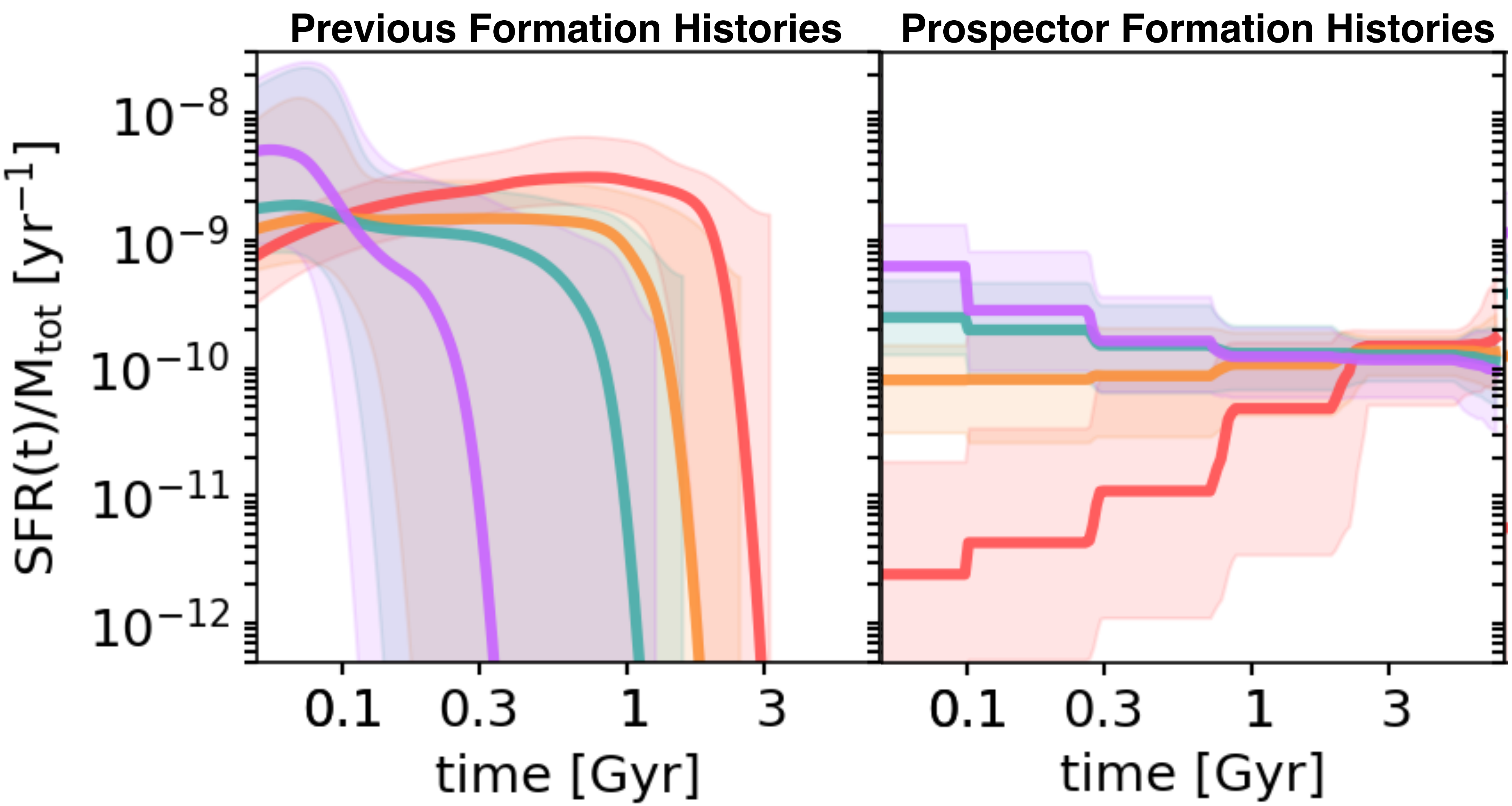
Full, detailed radiative transfer models of nearby star-forming galaxies (e.g. Andromeda) agree with surprising new Prospector estimates of star formation rate — 'old' stars power much of the **UV** and **IR** emission!



Nersesian et al. 2019

Are The New Formation Histories Reasonable?

We show average **star formation histories** based on position relative to normal star-forming galaxies (higher, equal, less, quiescent)



New histories consistent with evolution of mass function, while classic fits imply there should be no galaxies ~3 Gyr ago ($t_{universe}$=7 Gyr)

Leja et al. 2019b

I: State of the Field

II: A New Paradigm From High-Dimensional Models

III: A Million Times Faster with Machine Learning

IV: Surveys of the Future

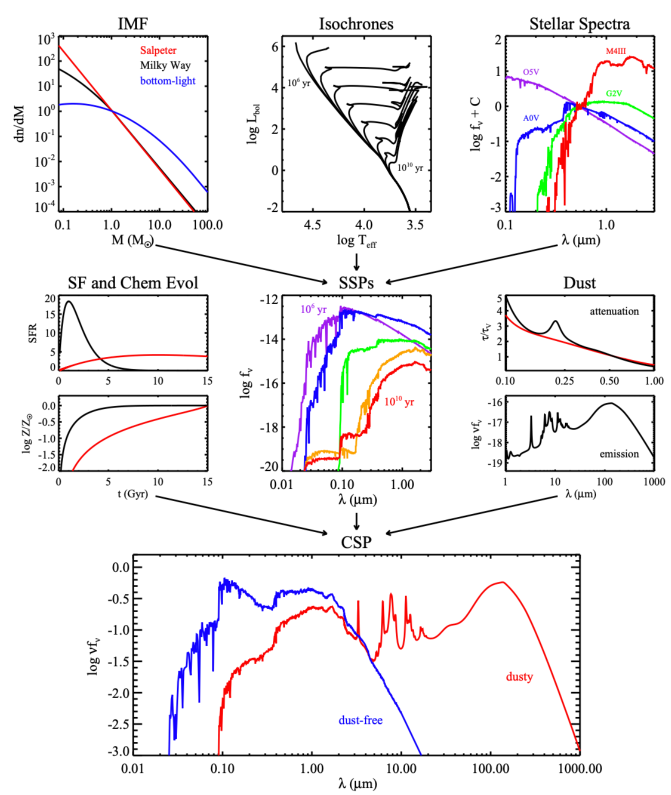
The Pressing Need for Additional Speed

High-dimensional models suffer from the curse of dimensionality. This means each stellar pop model must be generated **on-the-fly** — a compute-intensive task.

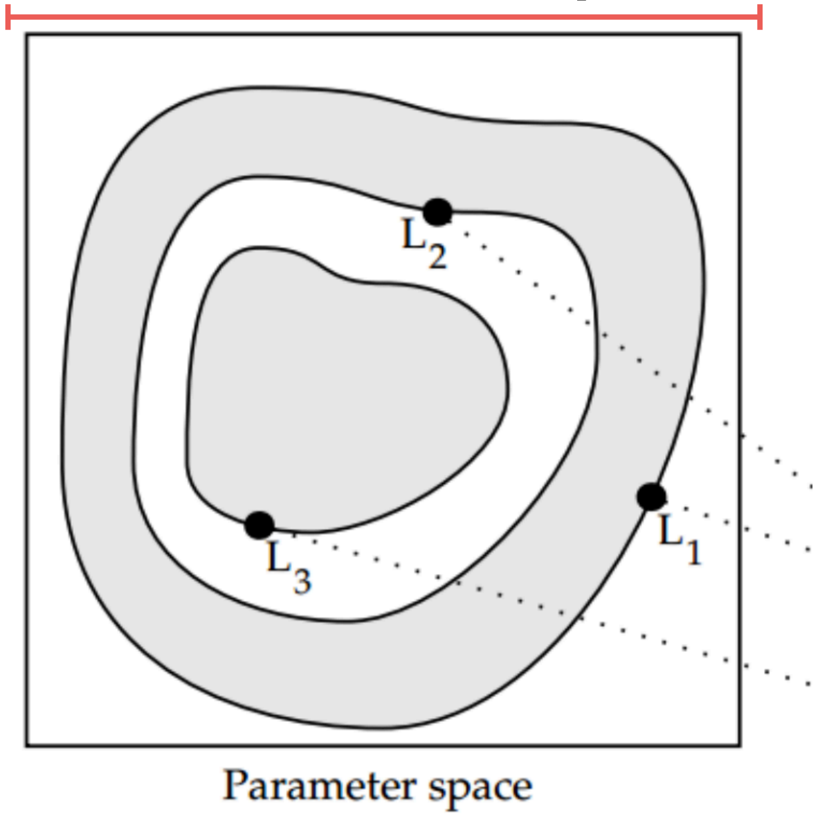
What is driving the computational budget?

galaxy observations

0.05s / model



~1 million models per fit



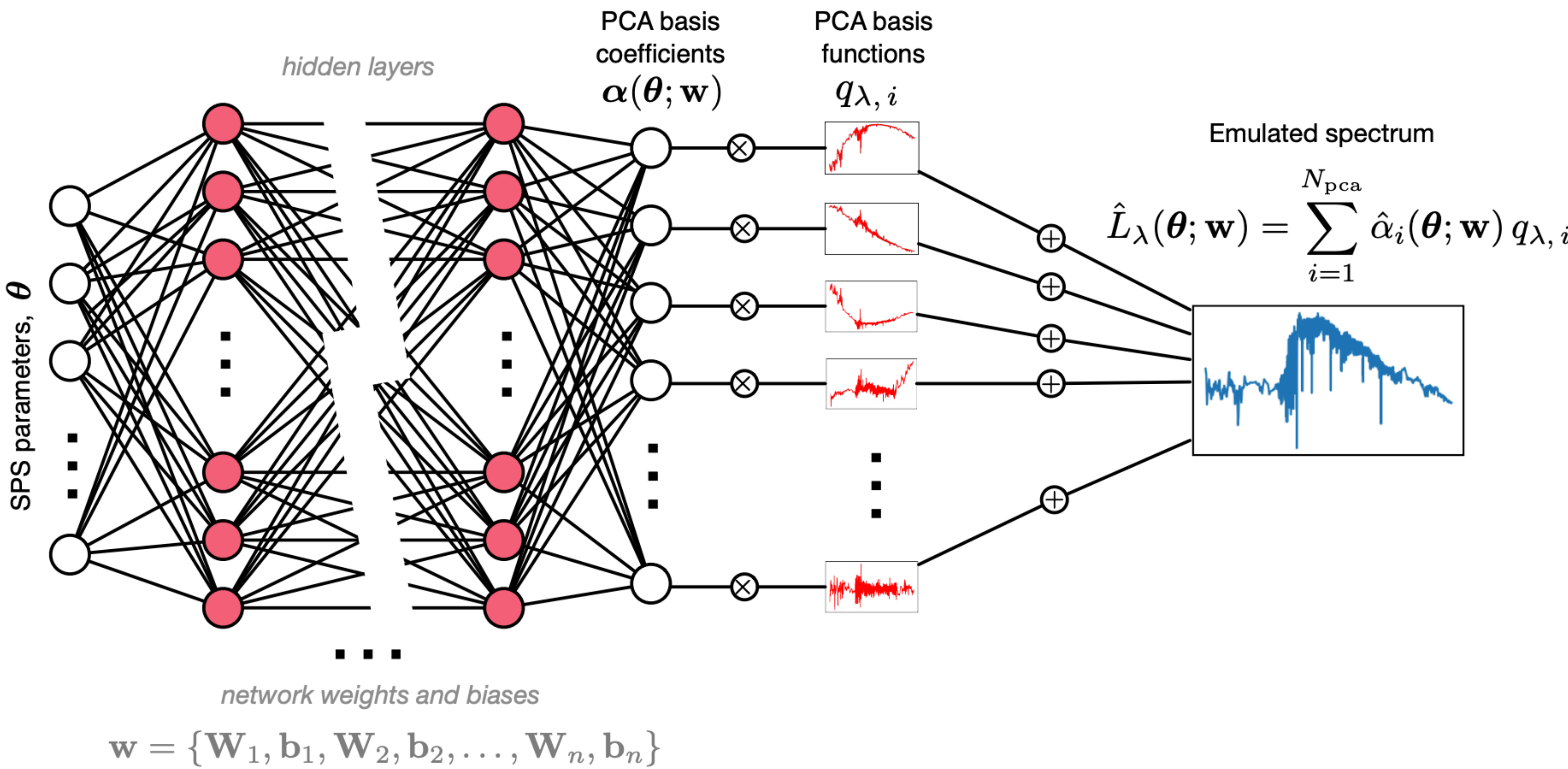
= 14 hours/ object

solutions

This means it takes **several million CPU-hours** to analyze a typical deep extragalactic field (~10^5 galaxies)

Neural Net Emulation: A Promising Solution

Training a neural net emulator to replicate stellar population synthesis outputs reduces model generation time by **~1000 (10^4 on a GPU)**

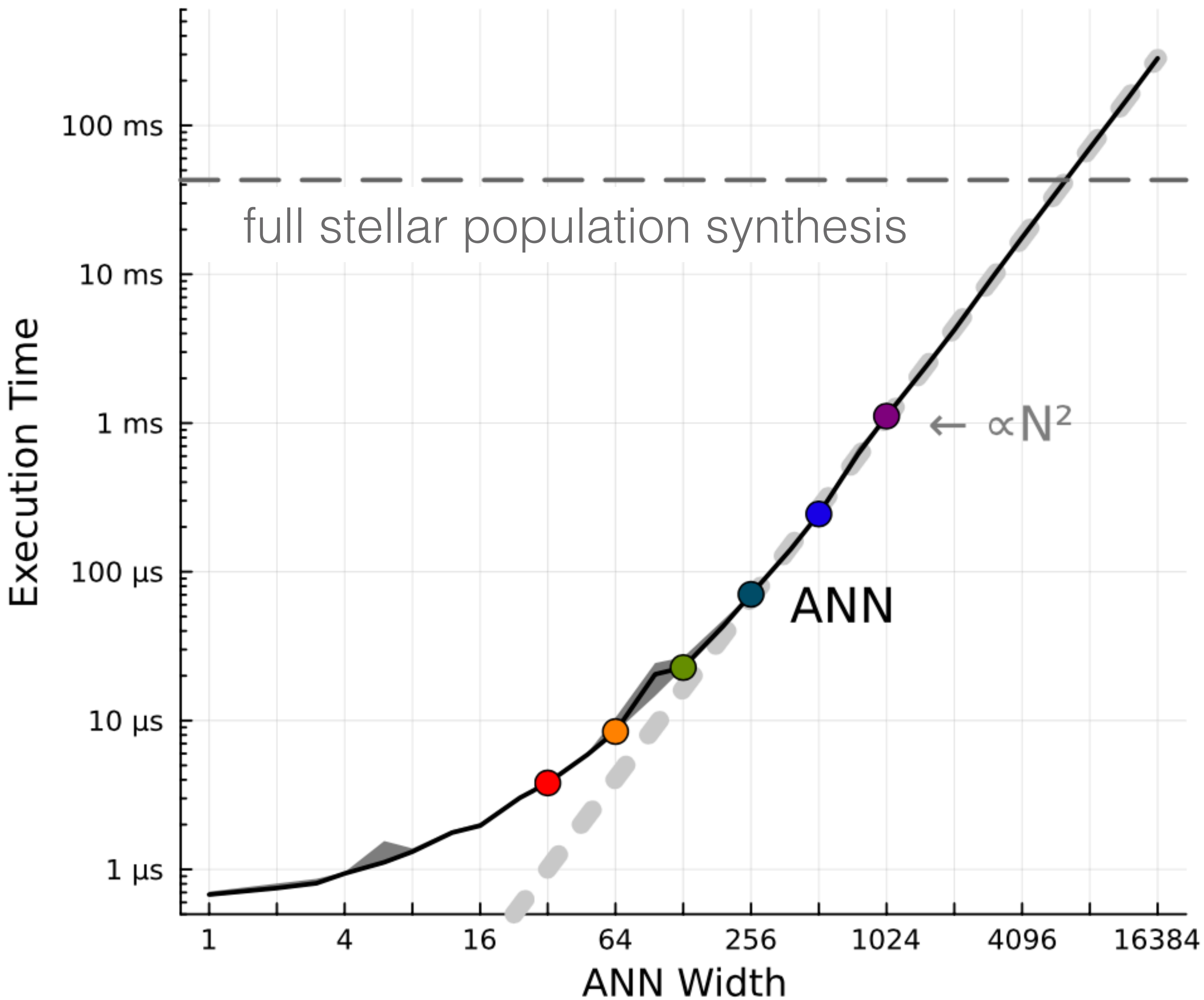


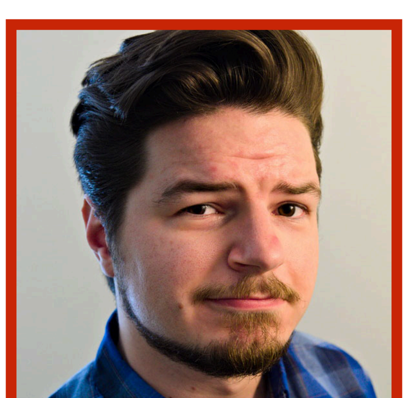
Alsing, Peiris, **Leja** et al. 2020

Optimizing Neural Net Emulators for Inference

Larger neural networks produce more accurate fluxes at the cost of increased execution time (**<0.01 mag error** with large-but-slow networks)

So: how accurate do they need to be?

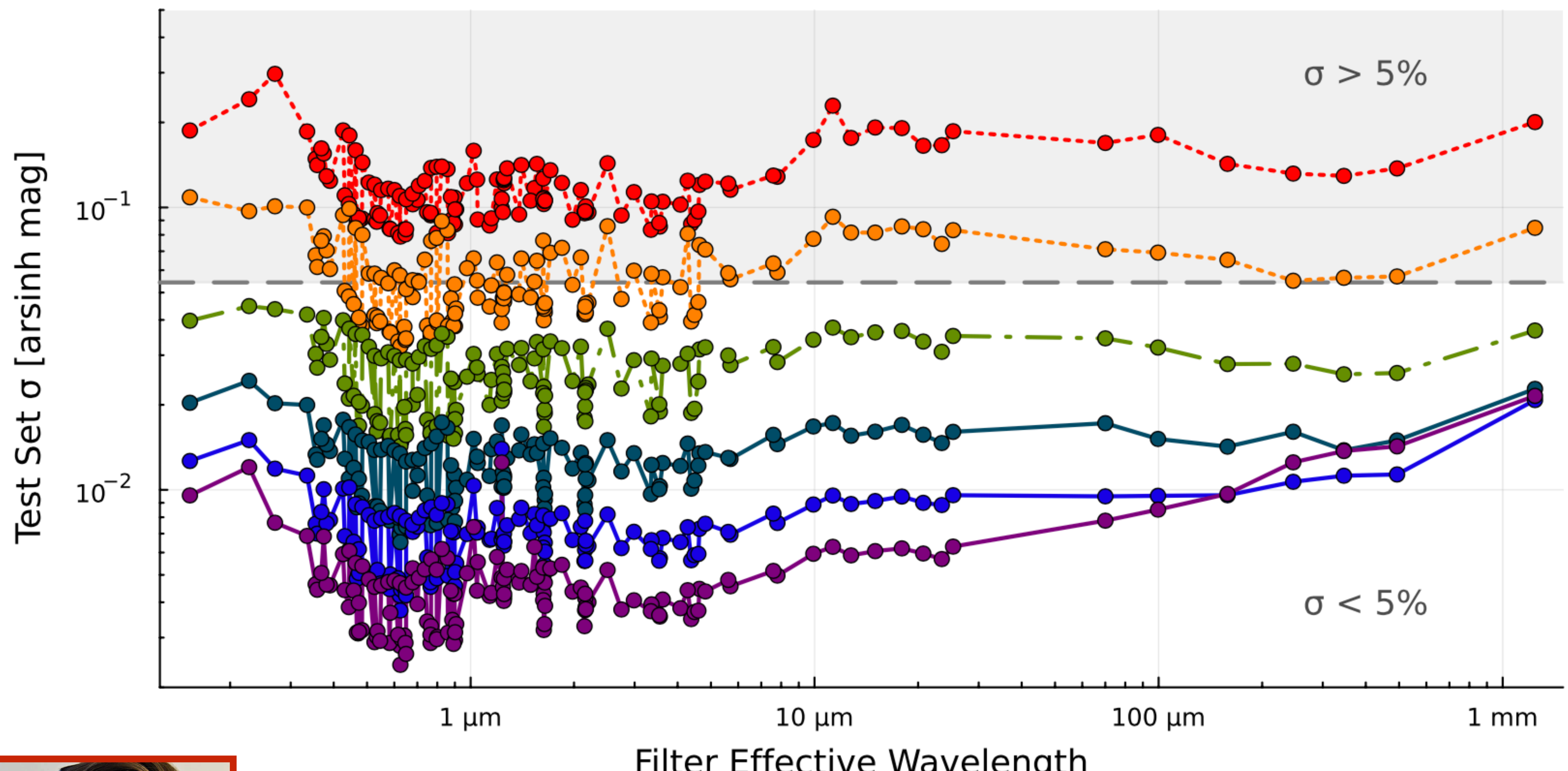




Mathews, **Leja** et al., ApJ submitted

`parrot`: As Simple as Possible, but No Simpler

We've built a neural net emulator ('parrot') for Prospector, emulating ~137 photometric bands. We built this many times with differing levels of accuracy to test parameter recovery.

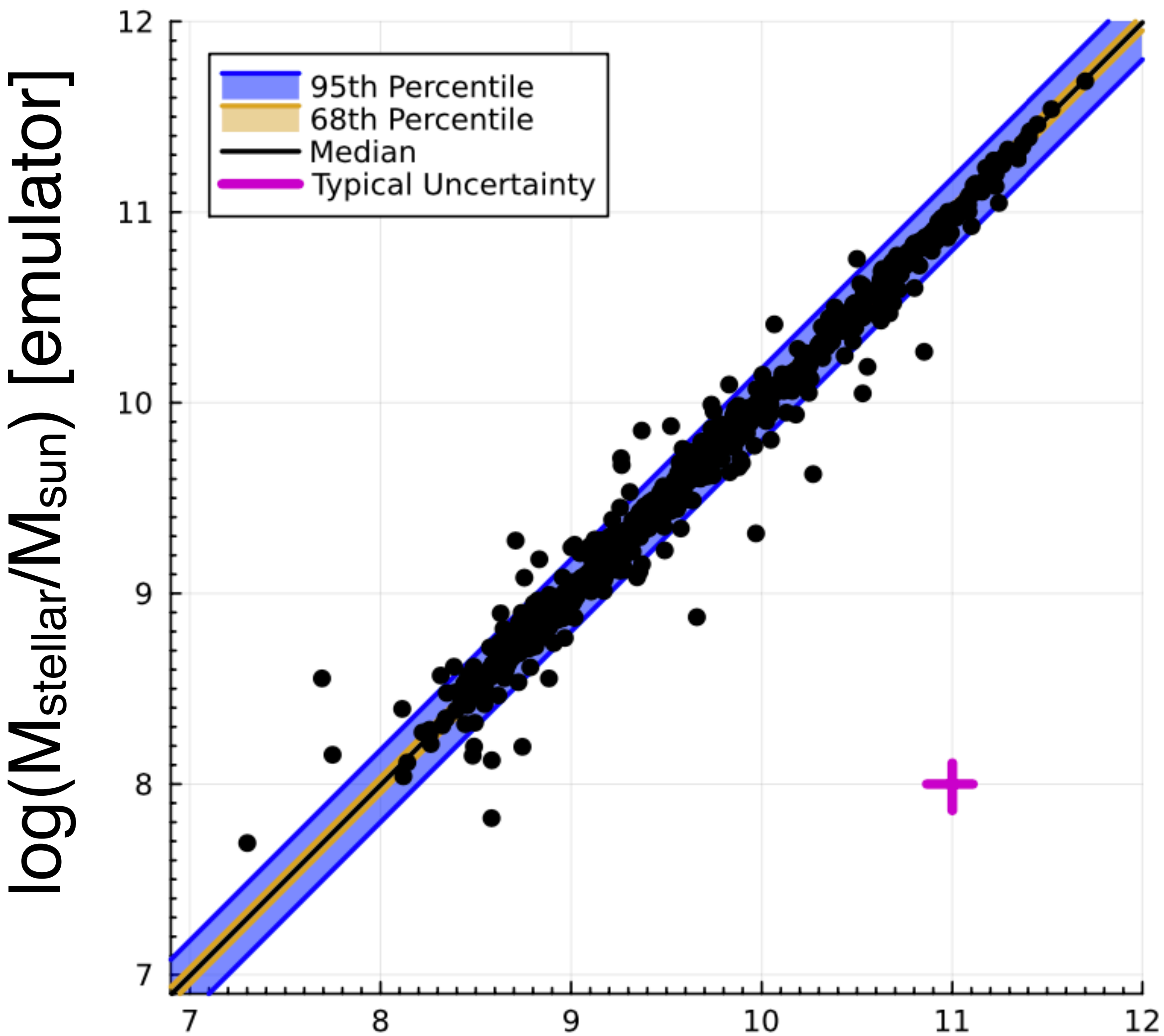


Mathews, **Leja** et al., ApJ submitted

A Sufficiently Accurate Network is Indistinguishable from Full Pop Synthesis

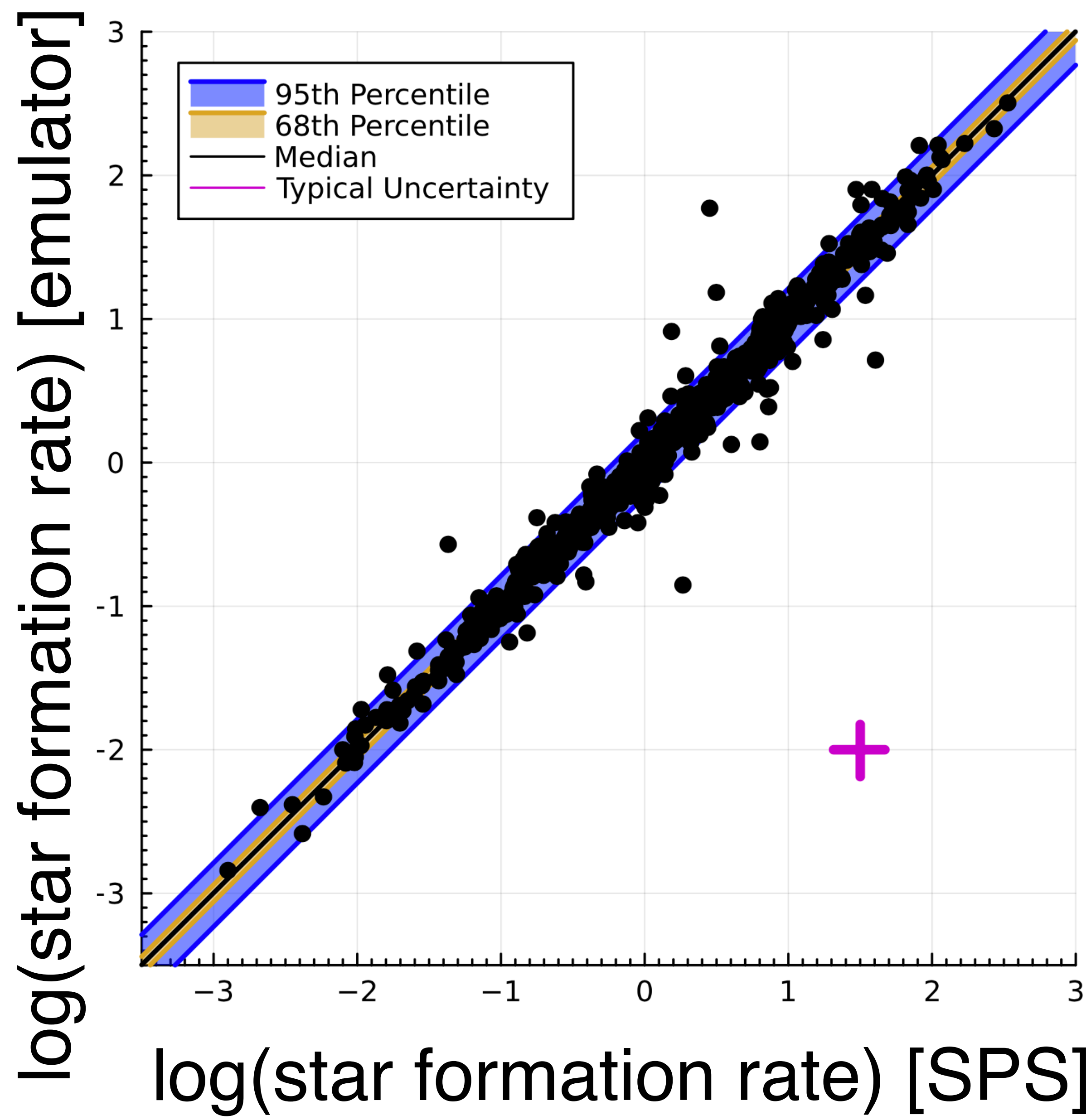
~1000 randomly selected galaxies from a typical deep extragalactic survey.

Stellar mass shows no bias and ~0.1 dex dispersion



log($M_{stellar}$/M_{sun}) [SPS]

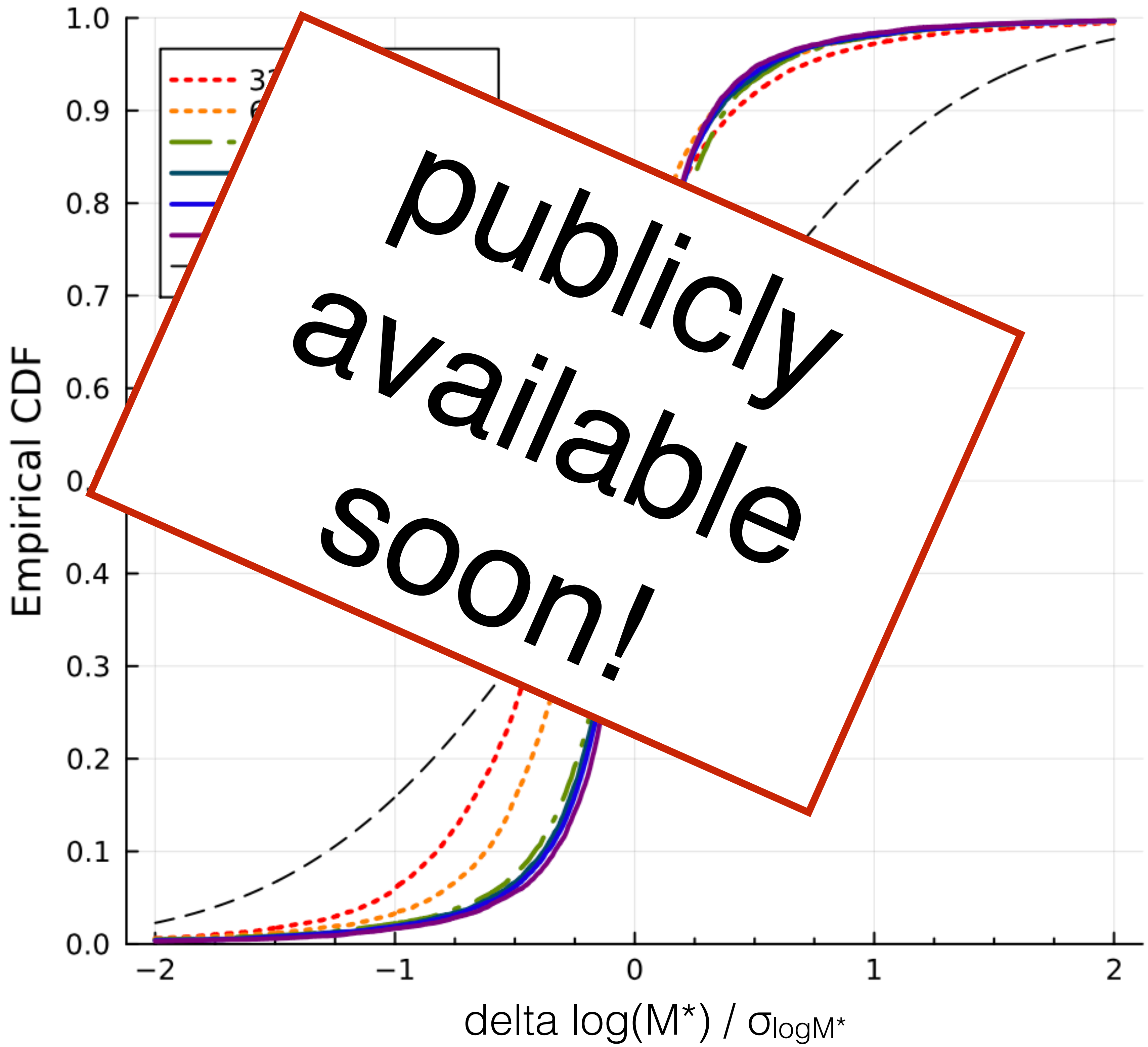
Star formation rate shows no bias and ~0.1 dex dispersion



Mathews, **Leja** et al., ApJ submitted

The Quest for the Simplest Sufficient Neural Network

We find that networks need to be 2-5x more precise than the typical observational uncertainty. For 5% errors in observed flux, this corresponds to a 128-node network with ~100μs execution time (vs ~50000μs normal calculation!)



Mathews, **Leja** et al., ApJ submitted

But Wait: How Fast is Fast Enough?

With neural net emulators, we can generate models ~500x faster than standard stellar population synthesis, and achieve ~10 minute/object fits.

Not good enough: at this speed it will take **~800 million CPU-hours** to fit all of LSST!

New Workflow

galaxy observations

Simulation-based Inference using Normalizing Flows (~1s)

= ~1 second/ object

solutions

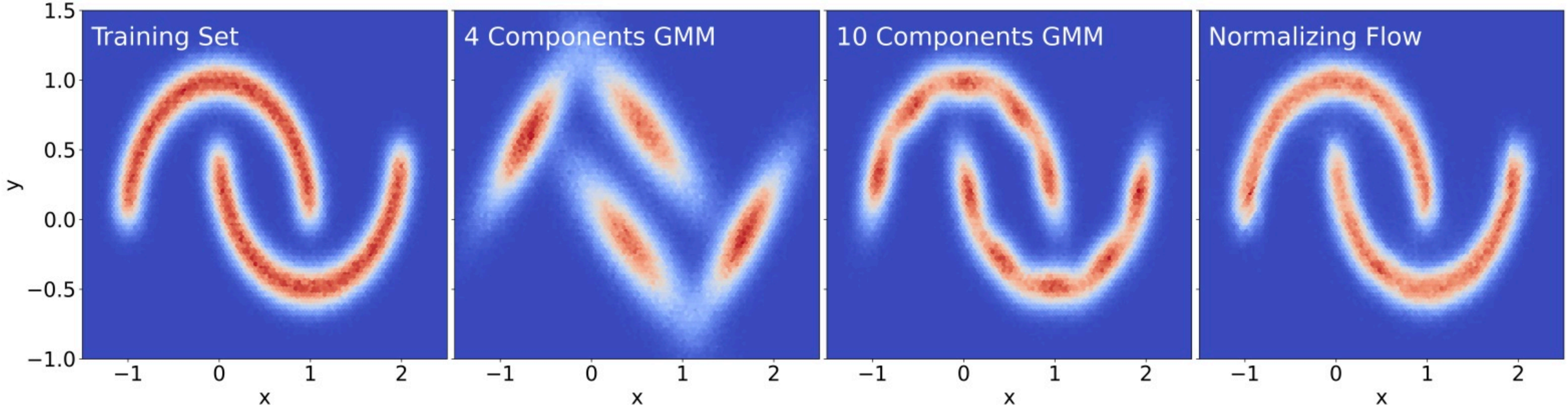
e.g., Wang, **Leja**, et al. 2022; see also Hahn & Melchior 2022

Simulation-Based Inference: Inside the Black Box

A type of Bayesian neural network: i.e., machine learning with real uncertainties!

This is done by simulating your data, *plus noise*, and the 'truth', many times, and **learning the direct transformation from noisy data to Bayesian posteriors.**

Specifically, use ***normalizing flows*** to learn the transformation from an N-dimensional Gaussian to an **arbitrary N-dimensional PDF**

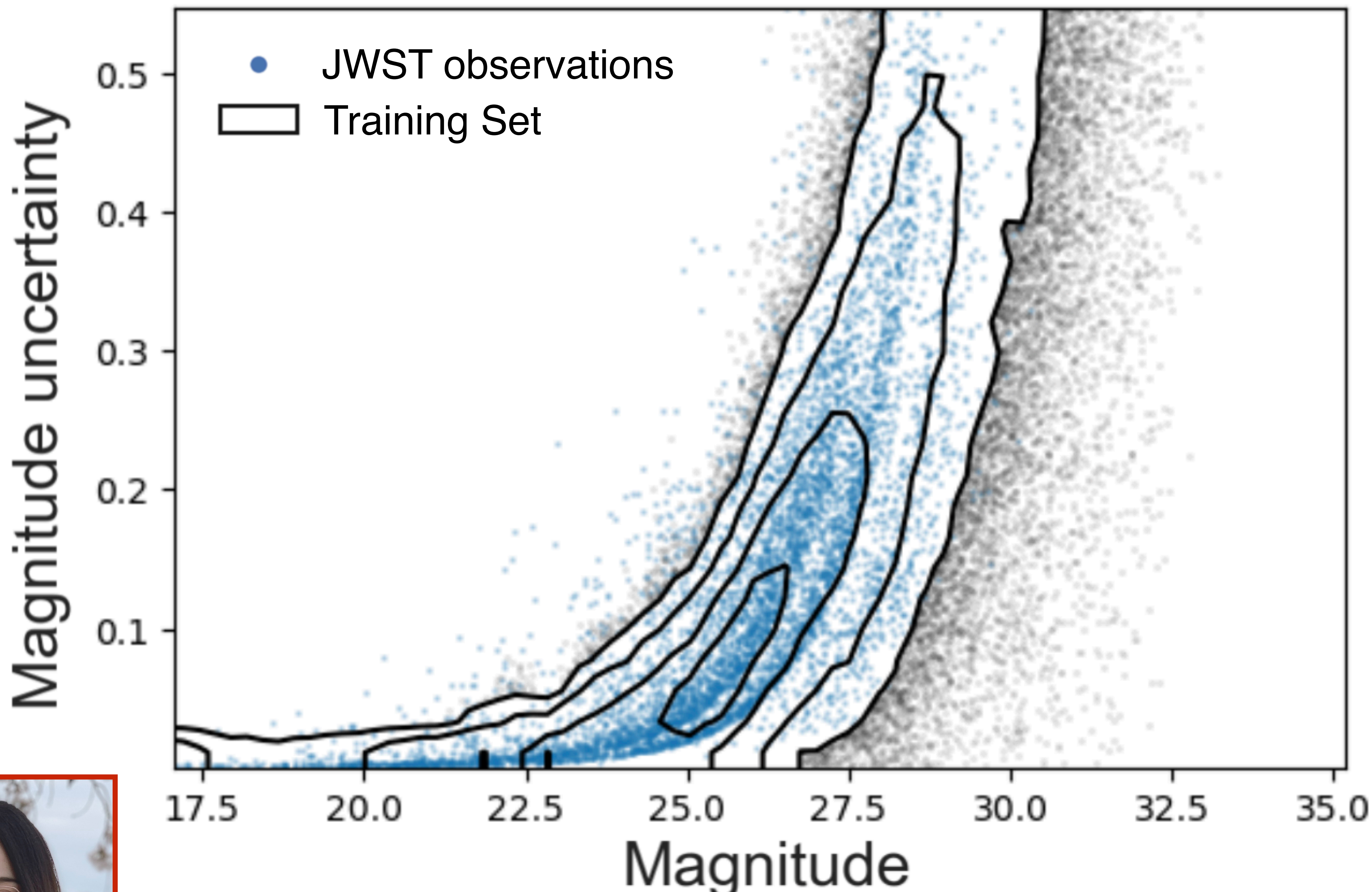


Ting & Weinberg 2021

input: (noisy) galaxy observations
output: P(z, star formation rate, stellar mass,…)

A Key Challenge: Astronomical Data is Often Weird

Everything must match the simulation — so no missing bands, no masked pixels, and no strong variations in noise patterns. Any objects with such properties are unfittable.

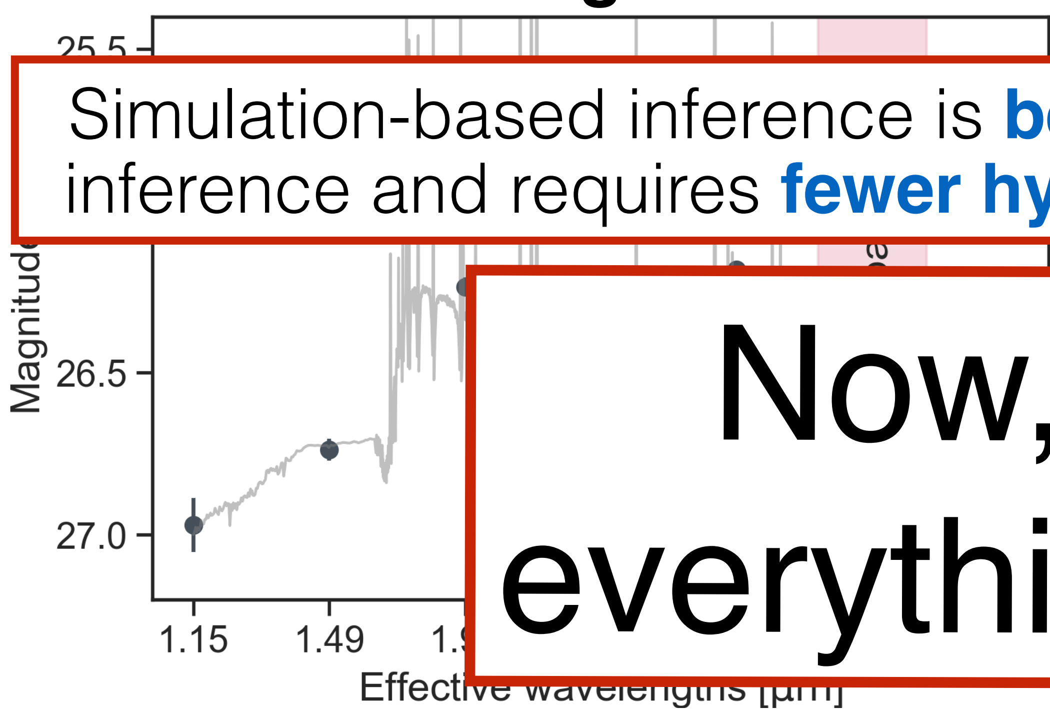


Wang, **Leja**, et al. 2022, NeuRIPS

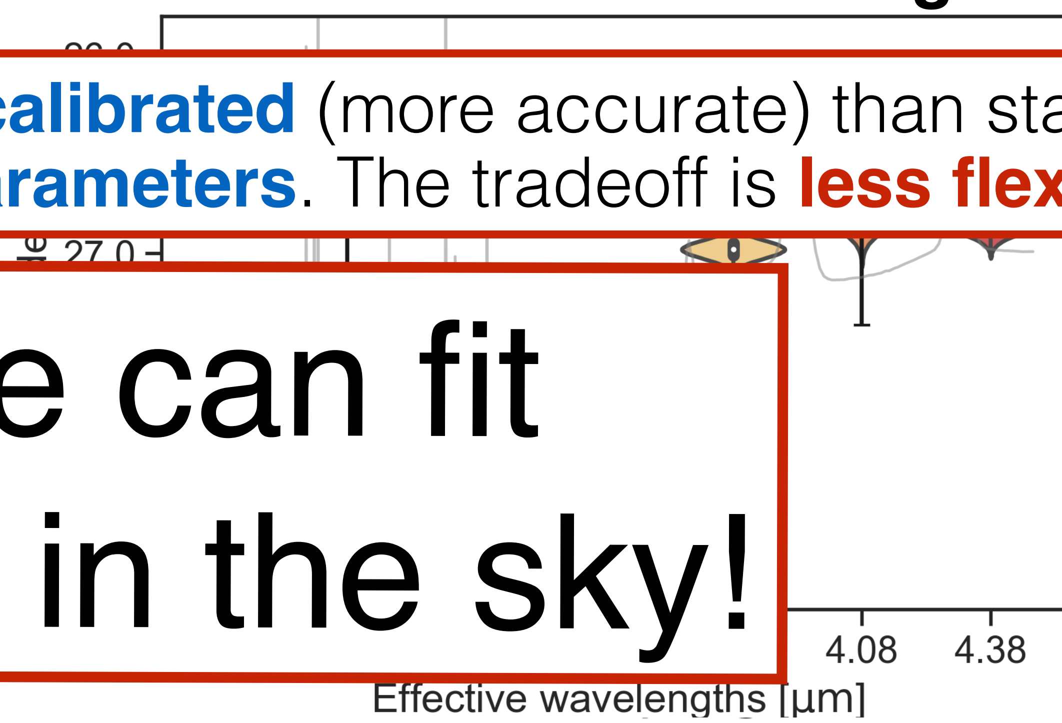
Retooling the Machinery to Fit Unusual Data

We've pioneered a technique to **overcome these limitations**; using internal Monte Carlo simulations and nearest-neighbor searches, we can now apply to objects with **missing data** or **out-of-distribution noise**.

Missing bands

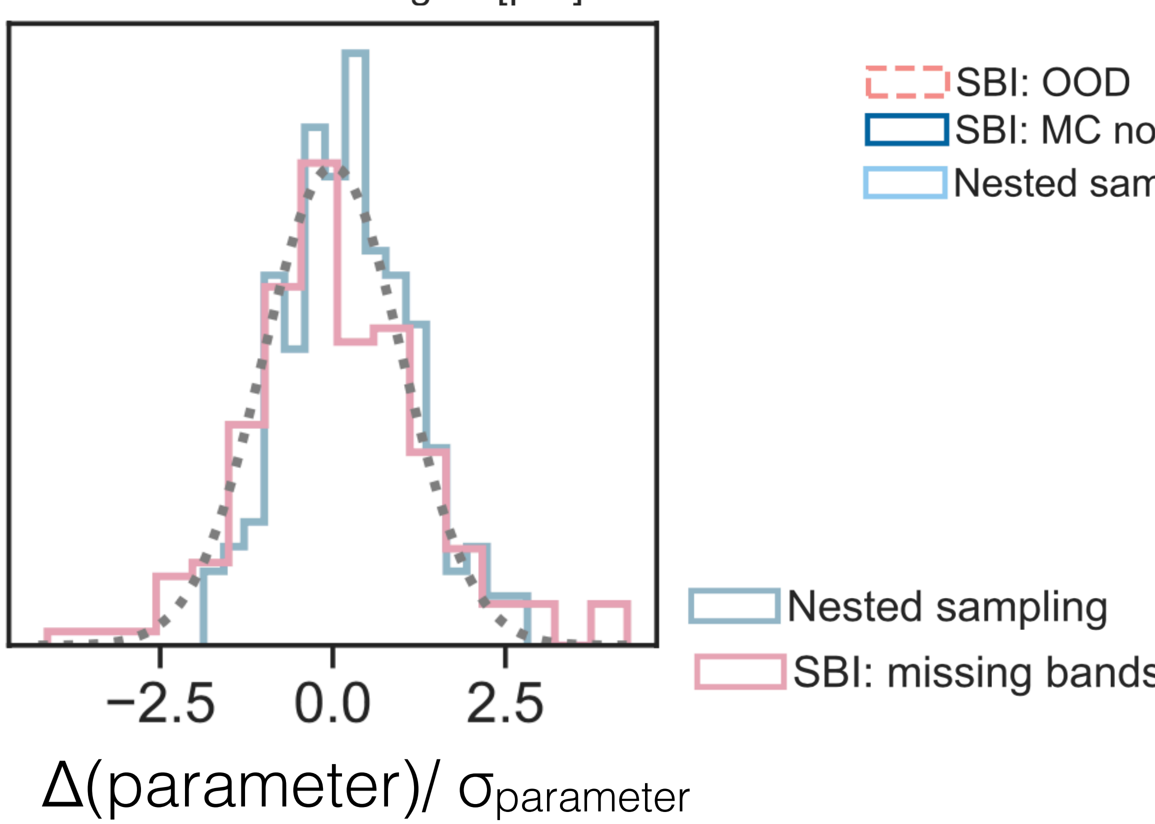


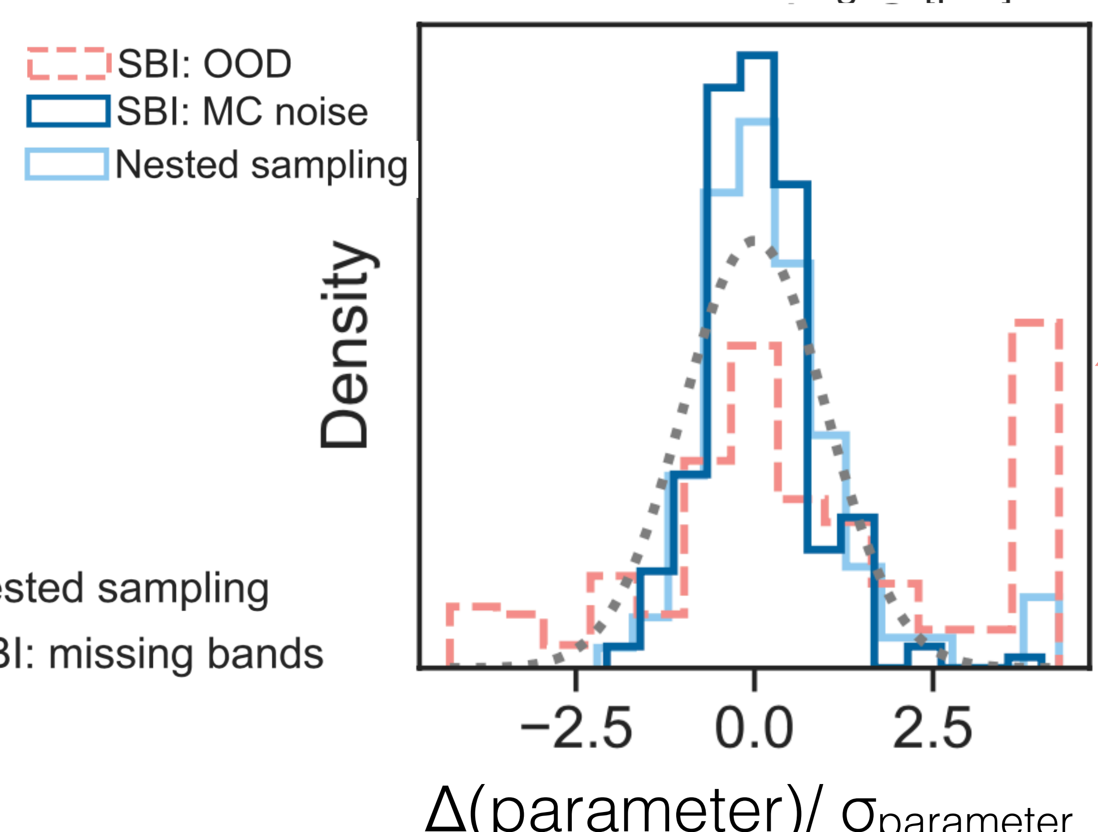
Noise outside of training set



Simulation-based inference is **better calibrated** (more accurate) than standard inference and requires **fewer hyperparameters**. The tradeoff is **less flexibility.**

Now, we can fit everything in the sky!





standard algorithm

Wang, **Leja**, et al. 2022, NeuRIPS; Wang, **Leja**+ ApJ in prep

Time required to analyze 50k galaxy SEDs

2019: **1.5 million CPU-hours** on **Harvard's brand-new computing cluster**

2021: A **couple of weeks on a laptop** with a neural net emulator

2023: A **couple of hours on a laptop** with simulation-based inference (and **more accurate**)

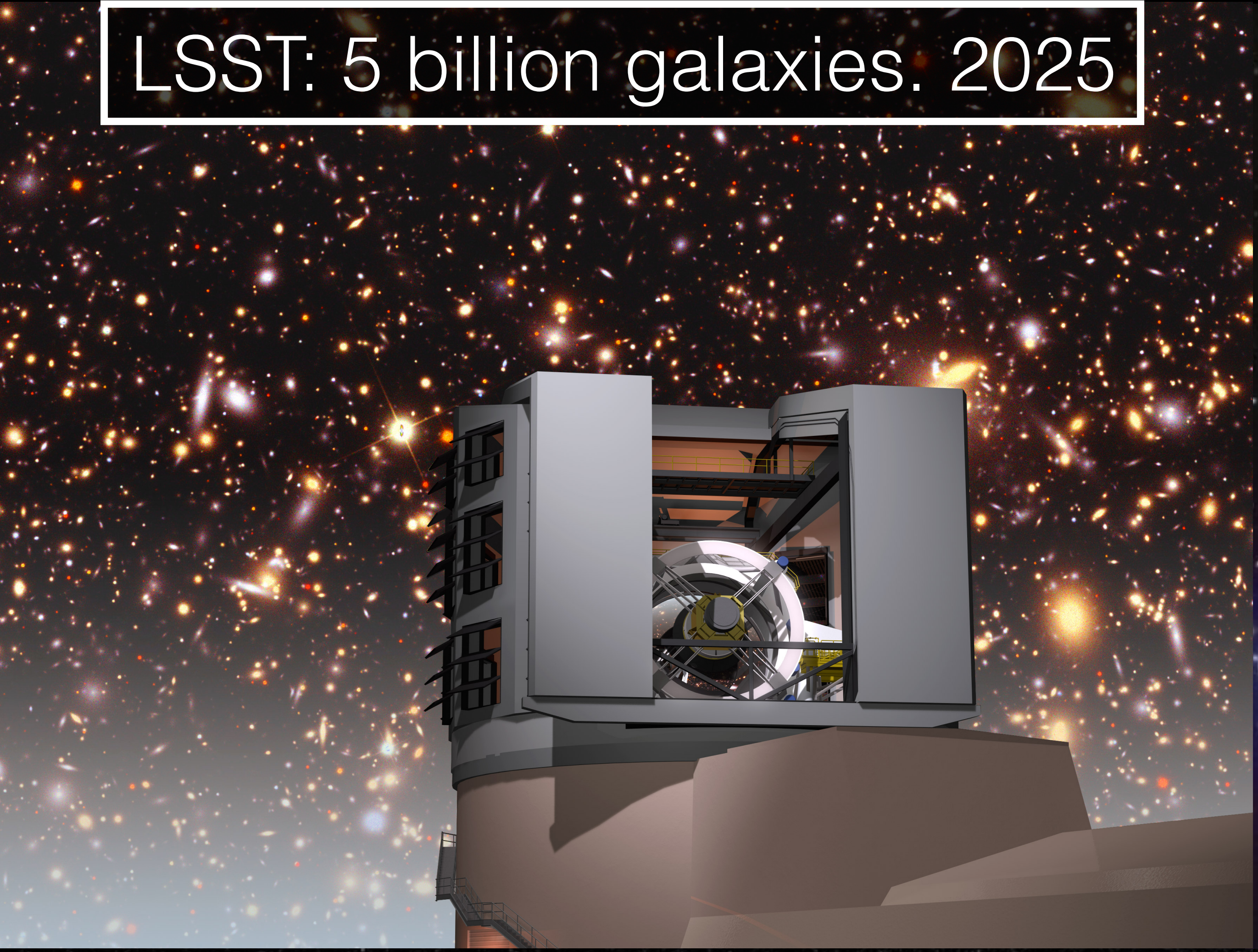
This rapid rate of increase opens up **new eigenvectors for scientific modeling!**

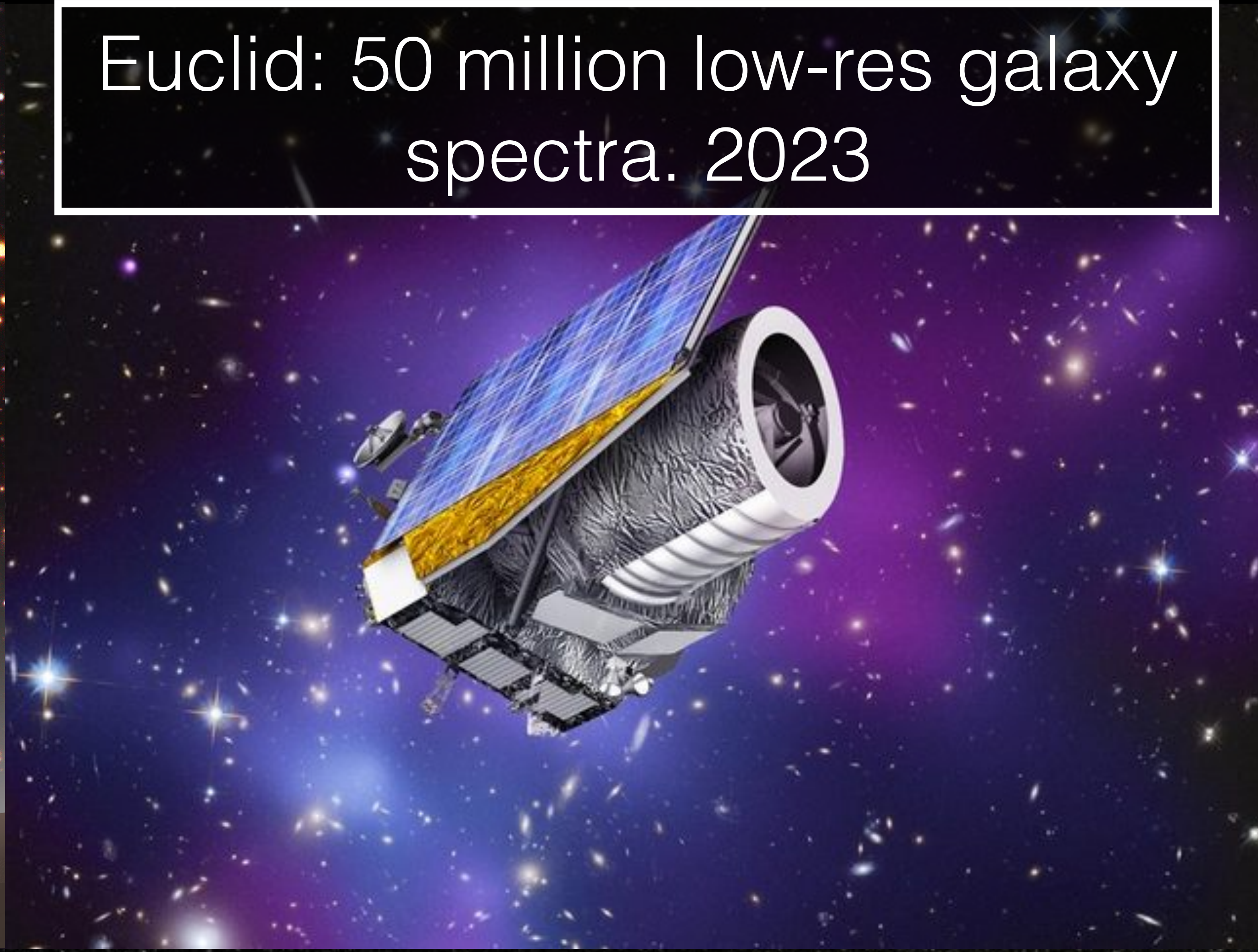
I: State of the Field

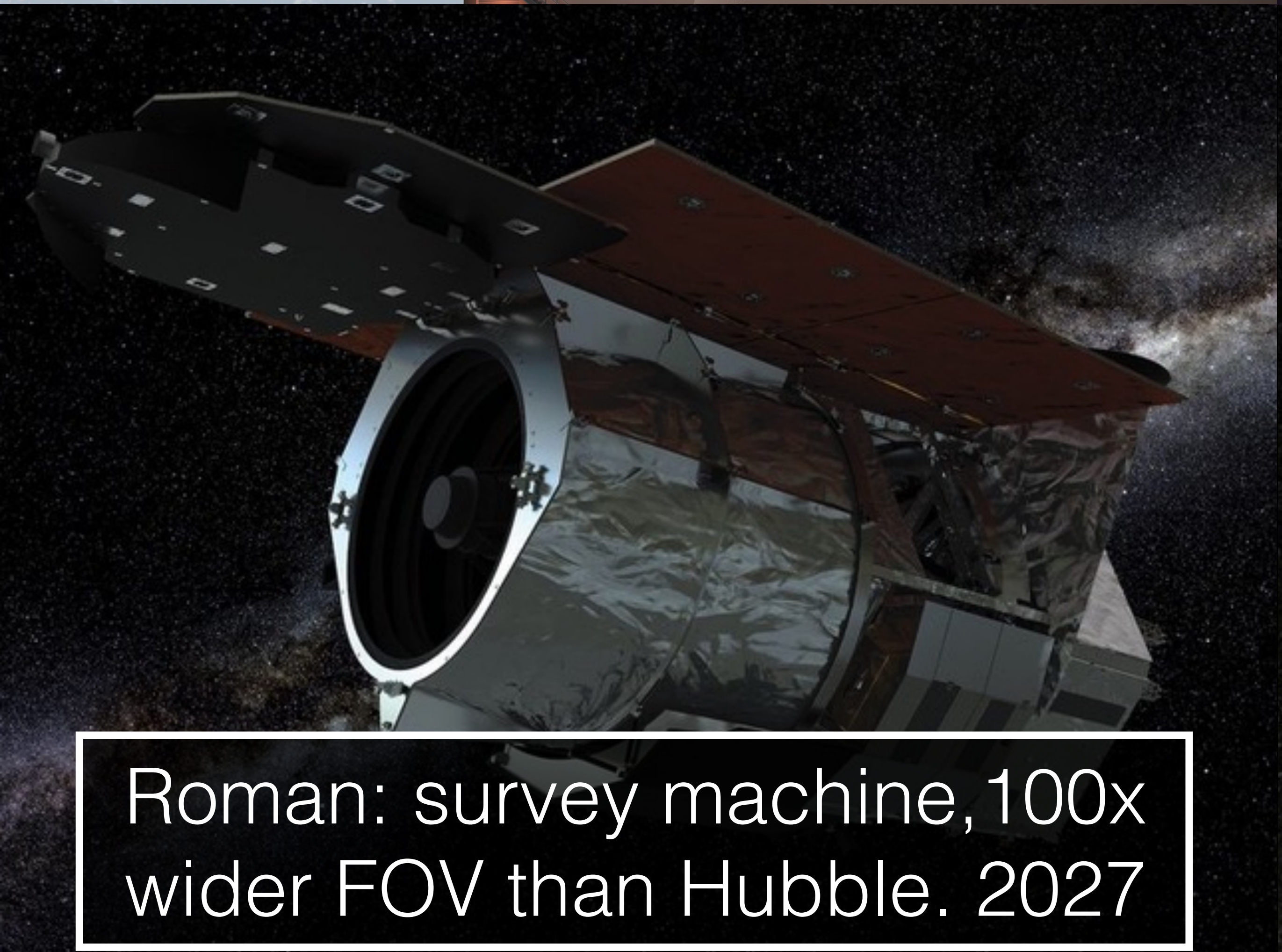
II: A New Paradigm From High-Dimensional Models

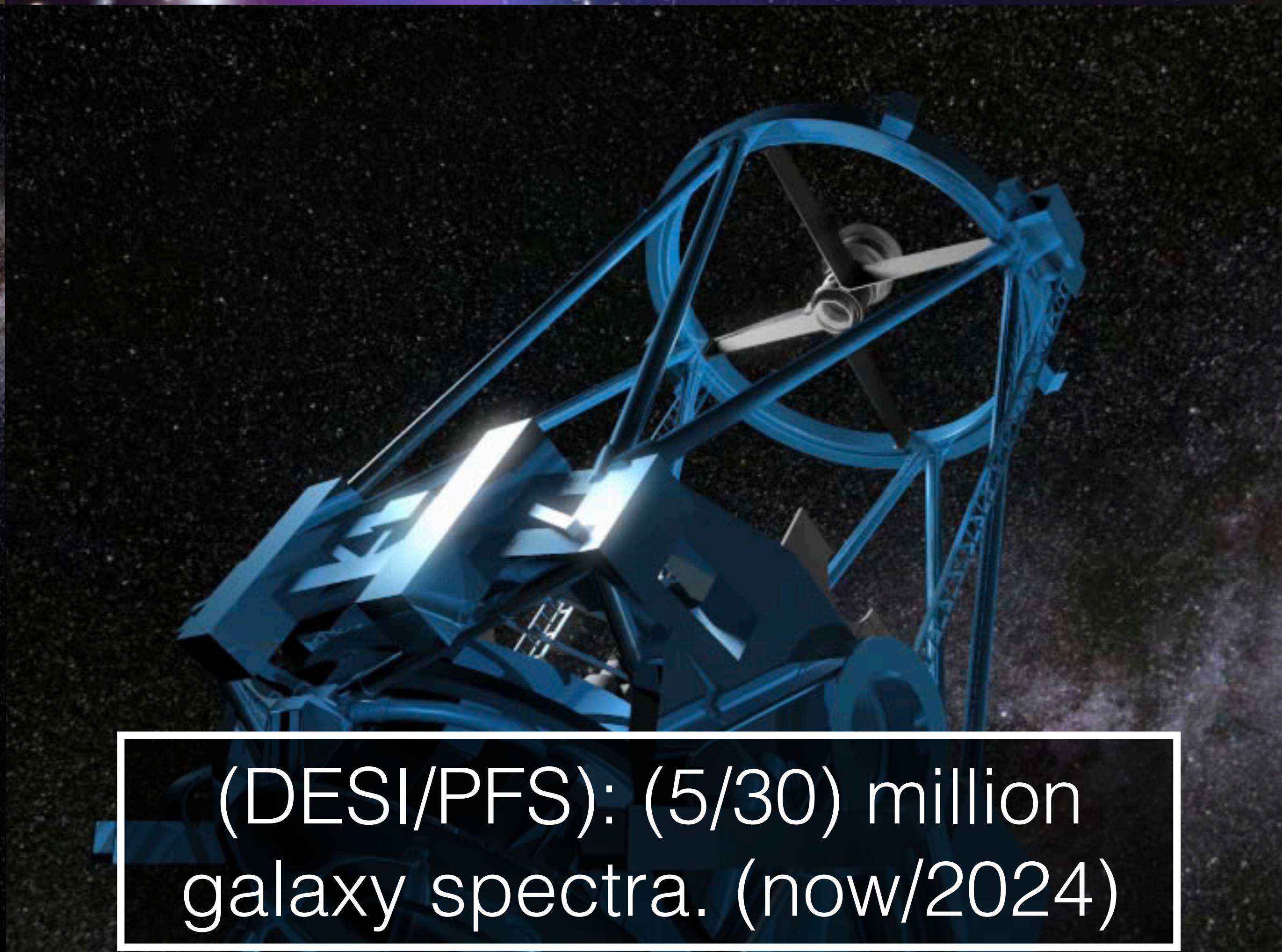
III: A Million Times Faster with Machine Learning

IV: Surveys of the Future


LSST: 5 billion galaxies. 2025


Euclid: 50 million low-res galaxy spectra. 2023


Roman: survey machine, 100x wider FOV than Hubble. 2027


(DESI/PFS): (5/30) million galaxy spectra. (now/2024)

Formation of First Stars & Galaxies with JWST

Leading the galaxy modeling for UNCOVER, the **deepest extragalactic observations in JWST's first Cycle,** designed to find **first stars/galaxies**. First results coming in weeks!

Wang,
Leja,et al.
in prep

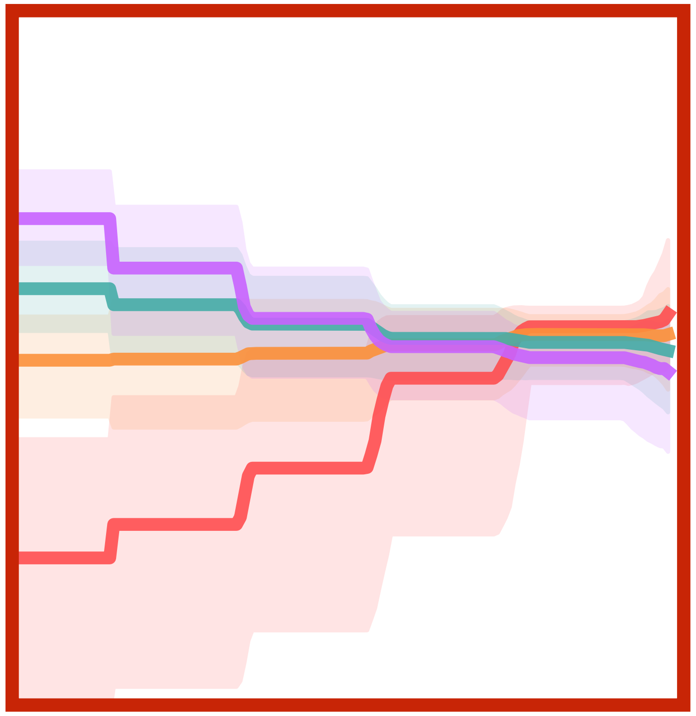
<u>The Future: Jointly Modeling *All* Galaxies Across Cosmic Time</u>

For example, LSST+Roman+Euclid overlap will provide 0.3-2um imaging plus 1-2um spectra for **500 million galaxies**

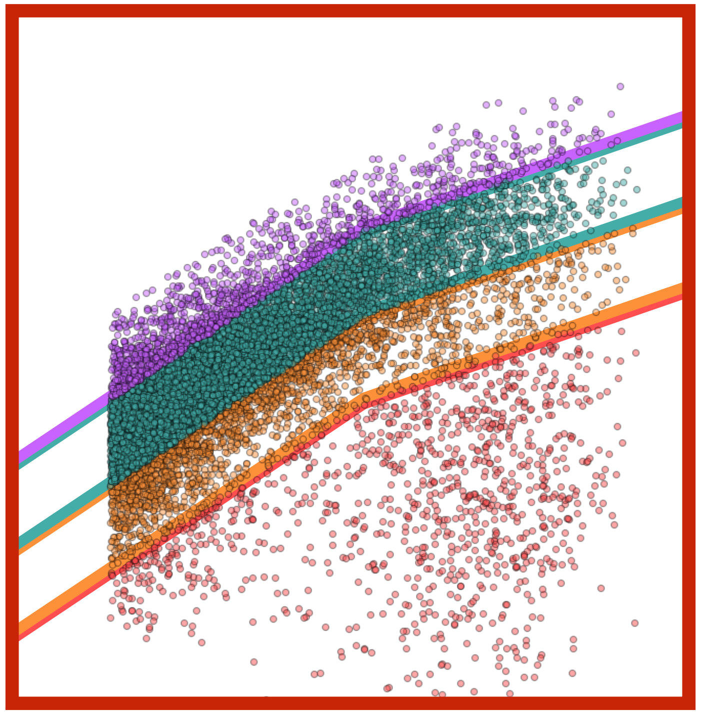
Future opportunity to model the **entire population simultaneously**

- ✳not just rare objects, but *populations* of them!
- ✳chart mass build-up of *every type of galaxy* with a **self-consistent hierarchical model**
- ✳simultaneously incorporate novel **spatially resolved data**

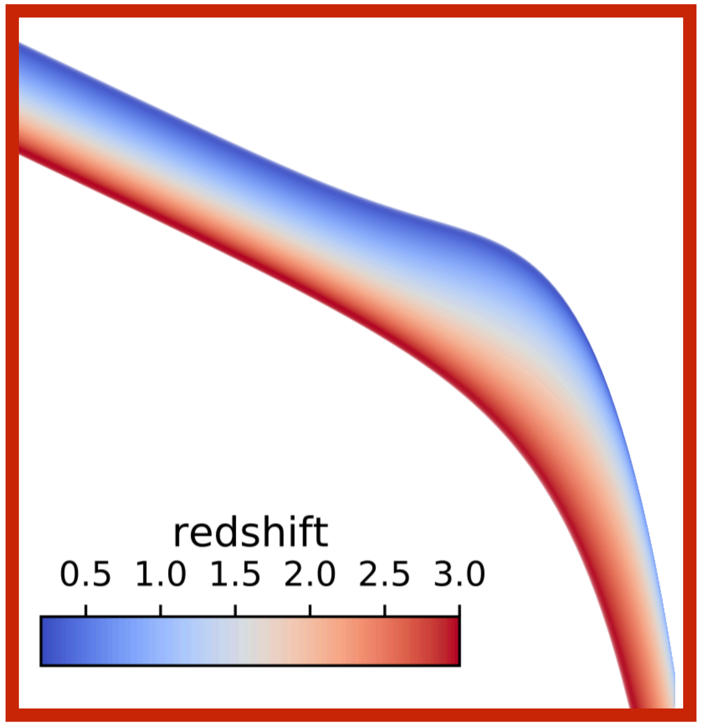
star formation history(M,z)



star formation rates(M,z)



stellar mass function (z)



Takeaways

With my collaborators, I have built a new discovery engine, Prospector, which is producing **new insights** into galaxy evolution. These new high-dimensional models already solve several **long-standing problems** in galaxy assembly.

Coupling to sophisticated machine-learning techniques permit **new generations of models** that are:

- 10^5-10^6 times faster (fit the whole sky!)
- much higher dimensionality (consider everything at once!)
- much wider in scope of physics (what else is possible?)

Ongoing and near-future projects include:

- **stellar modeling in galaxies** (e.g. IMF, abundance patterns, isochrones, rare phases of evolution; SDSS/Keck!)
- **training neural nets** to perform **ultra-fast fitting** (e.g. emulators, SBI)
- **building new inference tools** for exquisite new data (e.g. spatially resolved galaxies, fast photoionization models)
- **galaxy population modeling** with next-gen surveys (e.g. LSST, PFS, Roman — let's fit everything at once!)