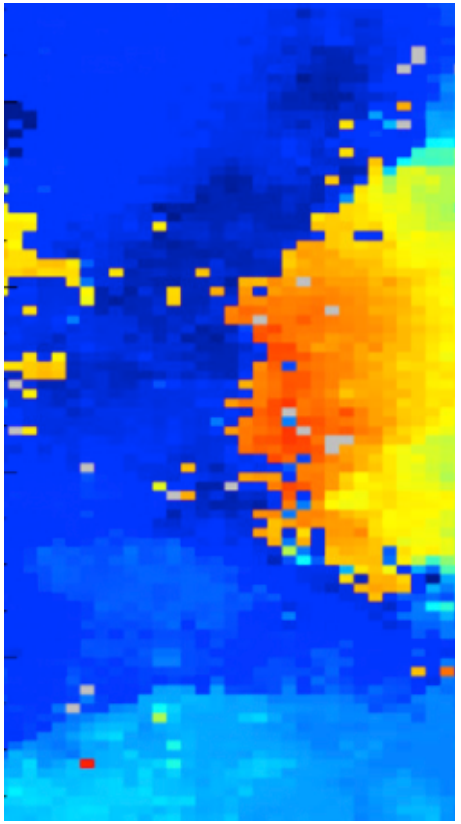
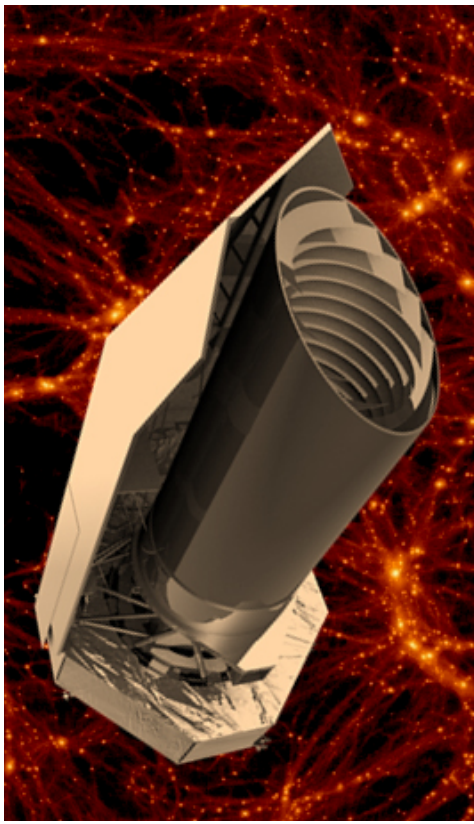
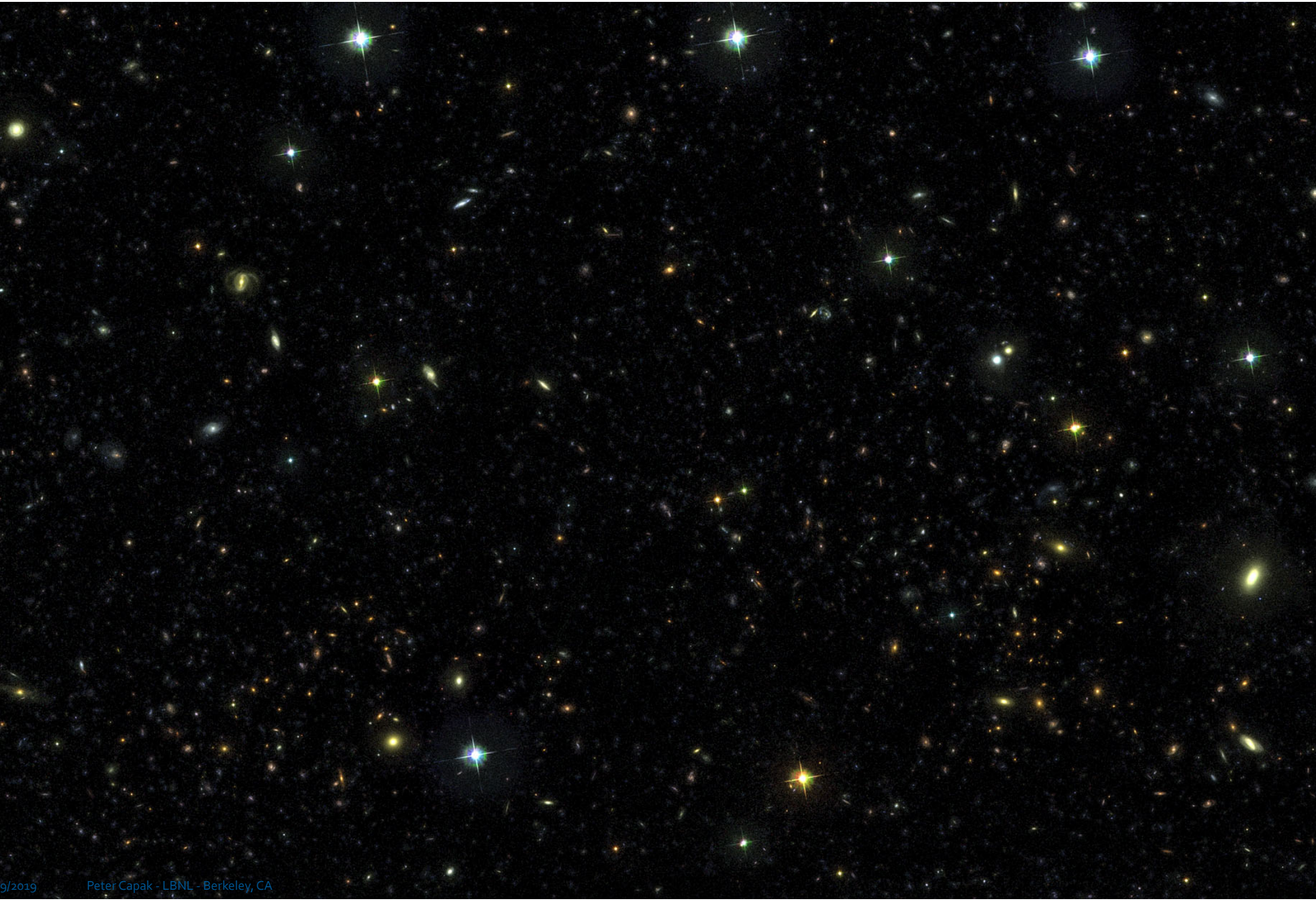
Developing a Standard Model of Galaxies for Cosmology

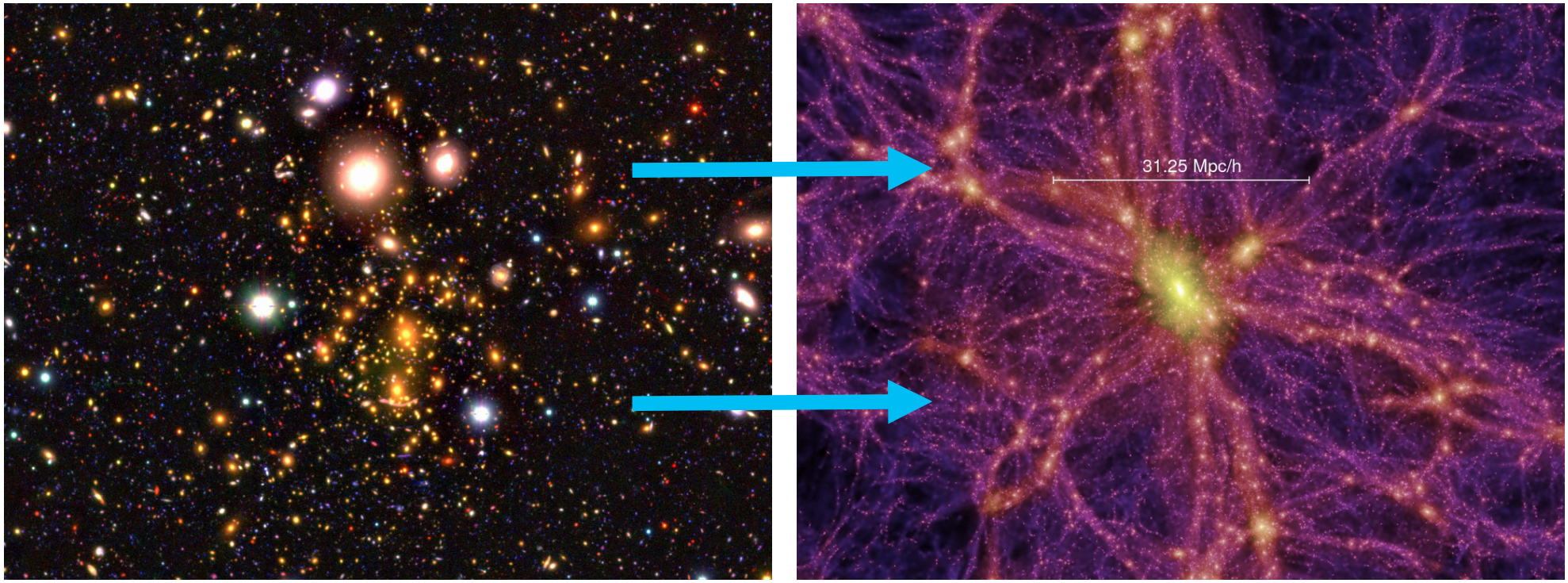
Peter L. Capak
IPAC – Caltech
Associate of the Cosmic Dawn Center







We see the galaxies but need to infer the dark matter.



Capak et al. 2004

Cosmology is like demographic studies, you need to measure average properties of complex individuals.

~29

Ve can treat the galaxy urvey problem like the US rying to make a census rom space.

Galaxies are a bias proxy for natter just as light is a bias roxy for population.

Ve don't need to nderstand why its bias, but eed to know how its bias.

Cosmology is like demographic studies, you need to measure average properties of complex individuals.

Ve don't need a precise
nodel of people to
nderstand how cities built
p.

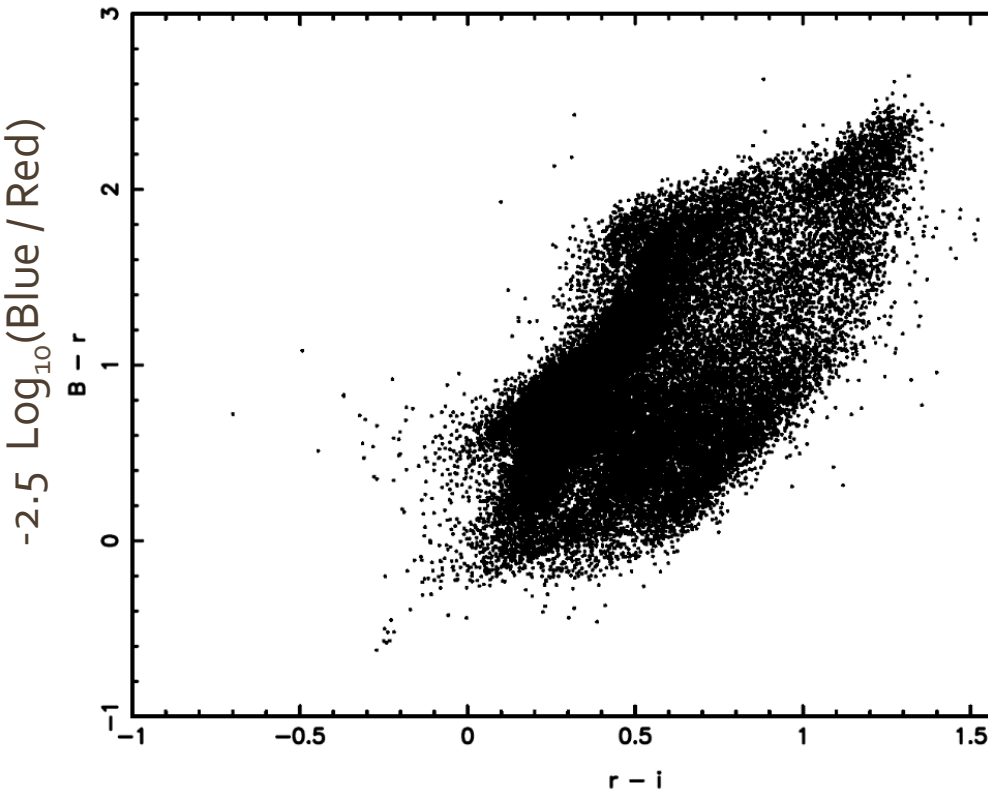
Modern cosmology needs to
neasure galaxy evolution
ut doesn't need to
nderstand it.

Ve just need to characterize
heir behavior accurately
nough.

Galaxies mostly look alike, they are strongly clustered in observed quantities.

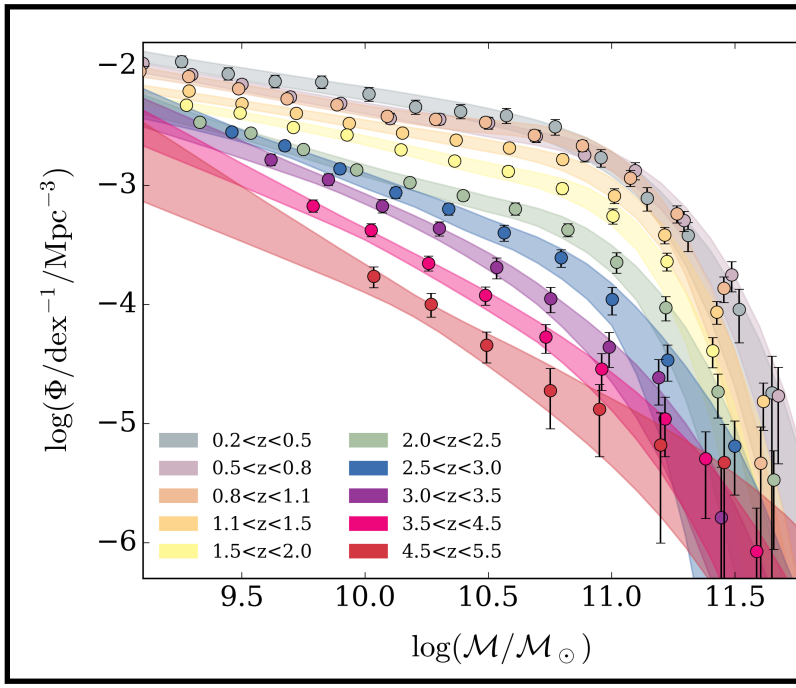
- Current surveys (CANDELS, COSMOS, HSC, DES) have samples of hundreds of thousands to millions of galaxies.
- The next generation of surveys will have billions of objects.
- Most of the galaxies are clustered in data space.
  - They mostly look alike!
- The distribution of points in data space is what contains the information on how galaxies evolve.

Example Of Galaxy Colors



When we analyze galaxy surveys we usually bin by some quantity or estimate statistical distributions.

- For example, constructing mass functions.
- What are the steps to creating a mass function?
  - Estimate redshifts to galaxies.
  - Estimate masses for the galaxies.
  - Bin the galaxies by redshift and mass.
  - Estimate incompleteness as a function of redshift and mass.
  - Estimate the space density based on the number of objects per mass and redshift bin, the area of the survey, and the completeness.
- Estimating the redshifts and masses are highly non-linear, non-gaussian operations.
- These non-linear operations are applies to a very complex high-dimensional data set with noise.
- This means it is very difficult to understand what is going on. As a result we have endless arguments over:
  - Photo-z outliers
  - Scatter in photo-z/spec-z plots
  - Representativeness of spectroscopic samples
  - Photo-z techniques
  - SED Libraries
  - And so on…….



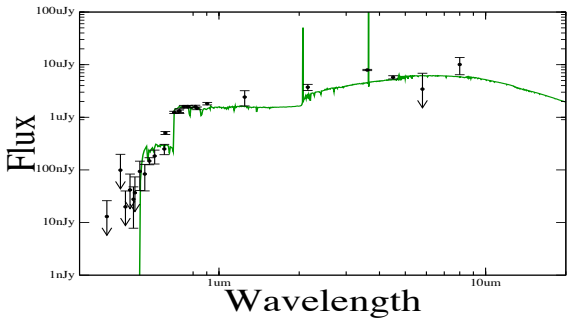
Davidzon+17:
stellar mass function

Lets re-think how we do galaxy evolution studies.

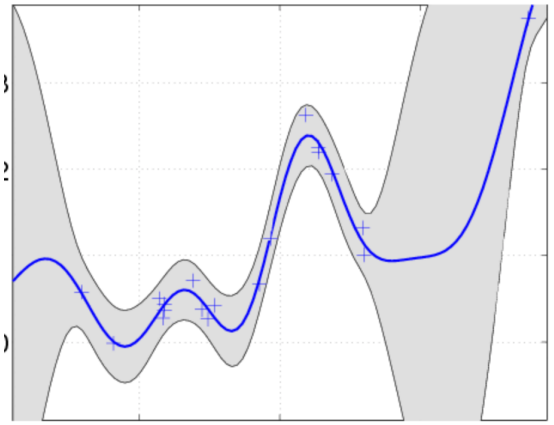
- For this analysis I will only consider photometric surveys for clarity.
- What are the steps to creating a mass function?
  - Estimate redshifts to galaxies = Use the color of a galaxy and its flux to estimate the redshift
  - Estimate masses for the galaxies = Use the color of a galaxy to estimate its mass to light ratio, then multiply by its bolometric flux.
  - Bin the galaxies by redshift and mass = Bin galaxies by color and flux.
  - Estimate incompleteness as a function of redshift and mass = Estimate completeness by color and flux.
  - Estimate the space density based on the number of objects per mass and redshift bin, the area of the survey, and the completeness = Estimate the space density based on the number of objects per color and flux bin, the area of the survey, and the completeness.
- These operations can all be done in data space with gaussian error where it would be much easier to understand what is going on.
- So why are we making this a complicated non-linear problem by fitting galaxies one at a time with spectral models?
- Because its conceptually and computationally hard to bin data in a complex high-dimensional space like an astronomical survey.

There are several types of machine learning and big-data techniques

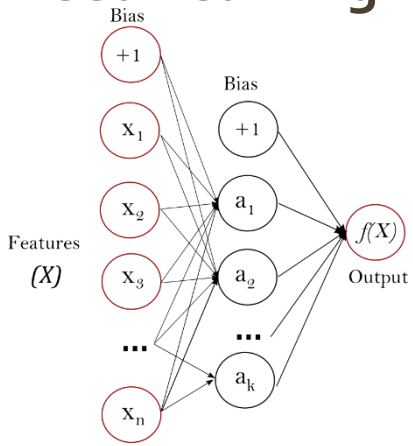
Analytic Models



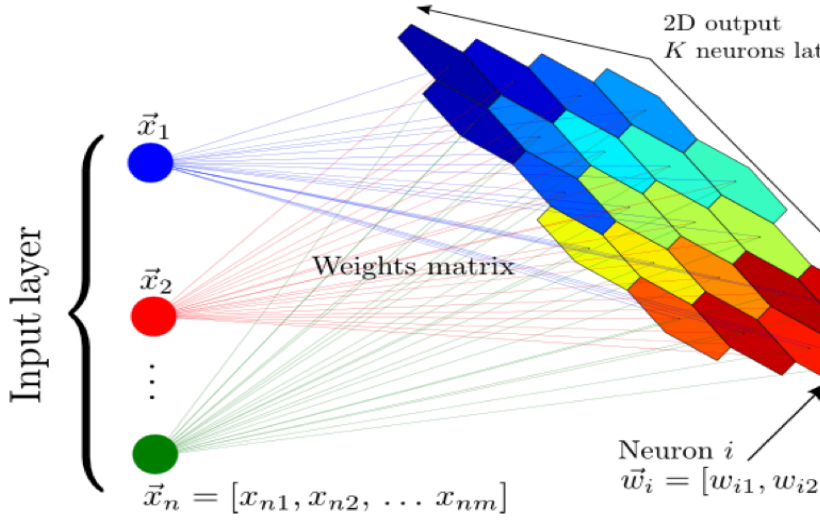
Non-linear Regression



Supervised Learning



Unsupervised Learning

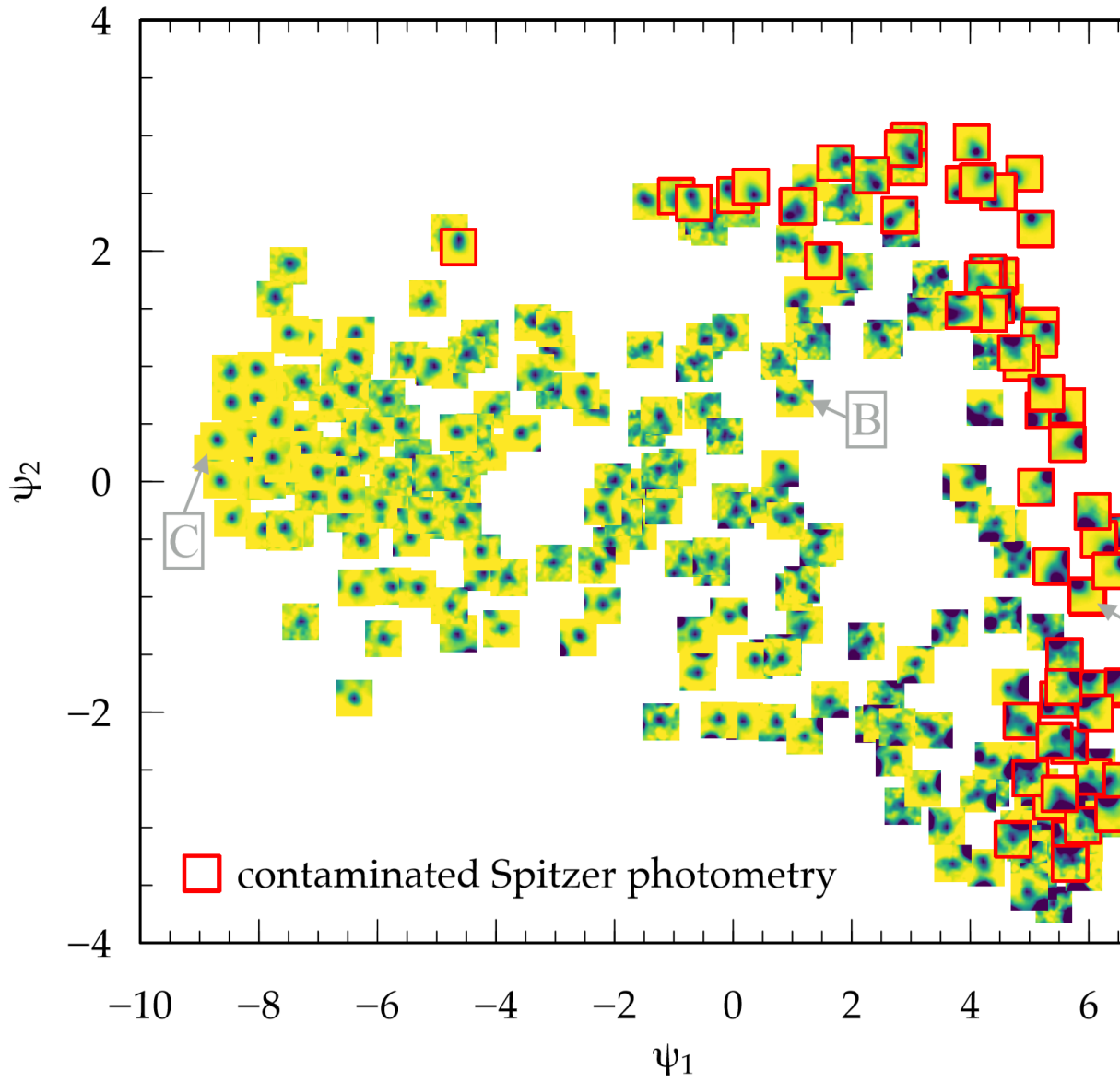


Example: Unsupervised Learning with t-SNE

Machine learning can be used to quantitatively sort astronomical data

These are spitzer postage stamps sorted with a t-SNE.

This analysis is being used to quantify the likelihood photometry is affected by bad photometry.



Example: Characterizing galaxy photometry with a Self Organizing Map.

- We adopt a widely-used technique known as the Self-Organizing Map (SOM), or Kohonen Map
- Easy to visualize
- Easy to understand

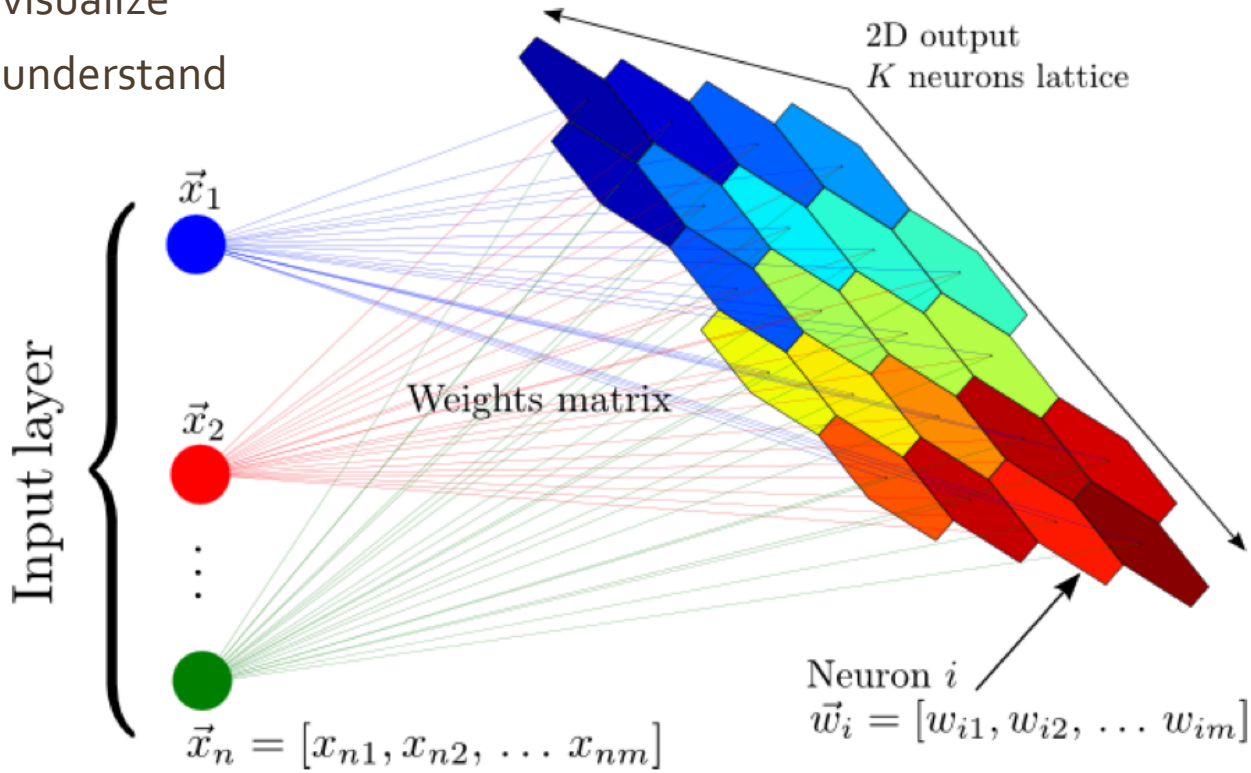
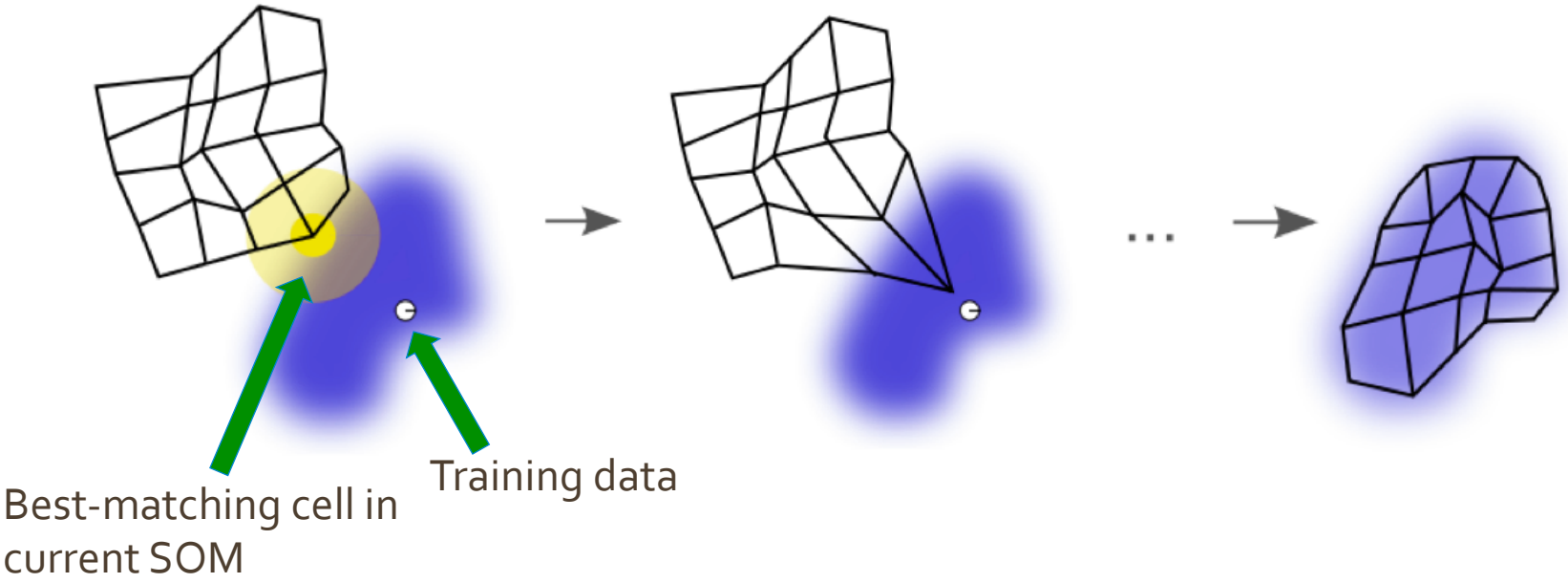


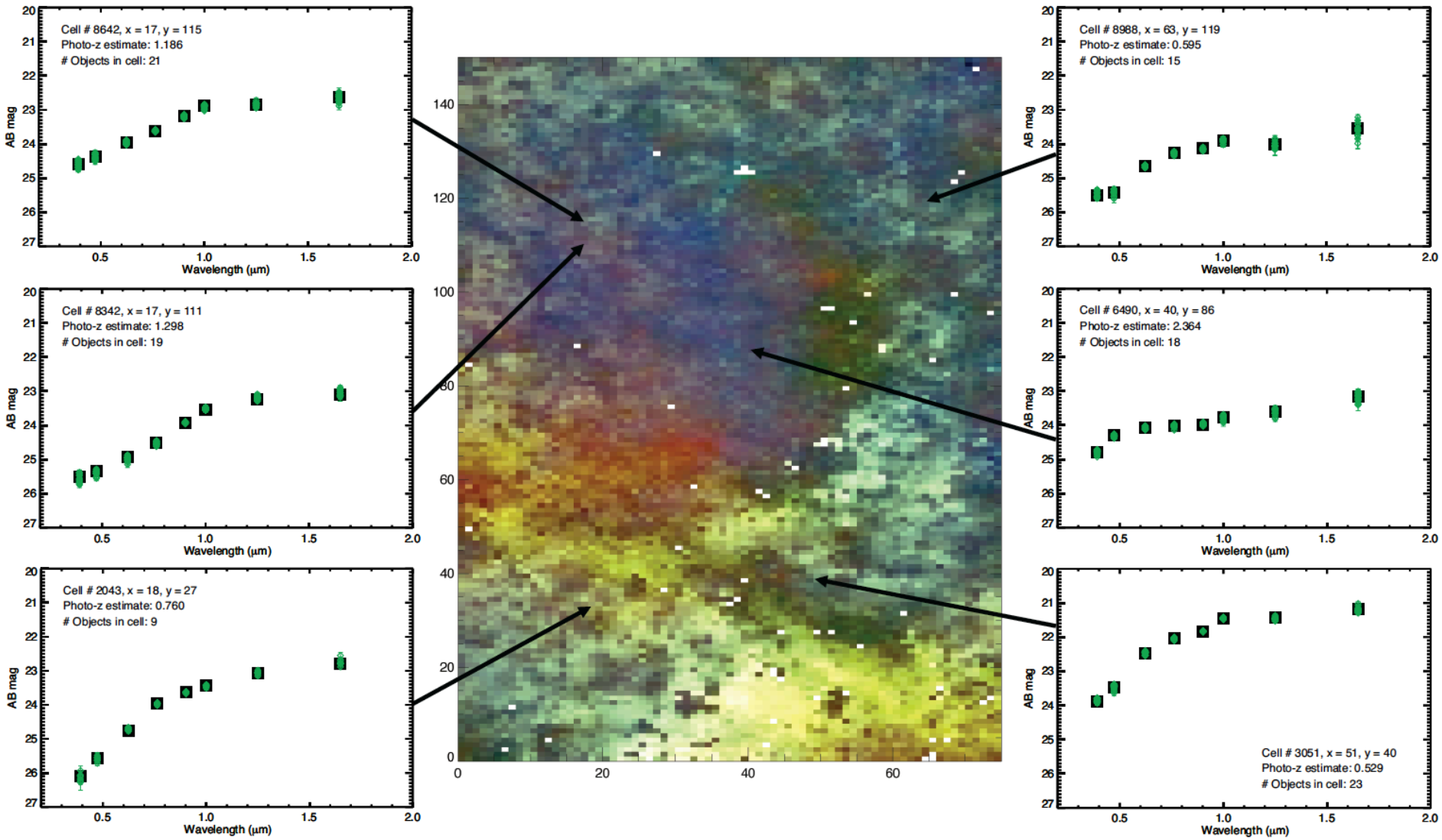
Illustration of the SOM (From Carrasco Kind & Brunner 2014)

Example: Characterizing galaxy photometry with a Self Organizing Map.



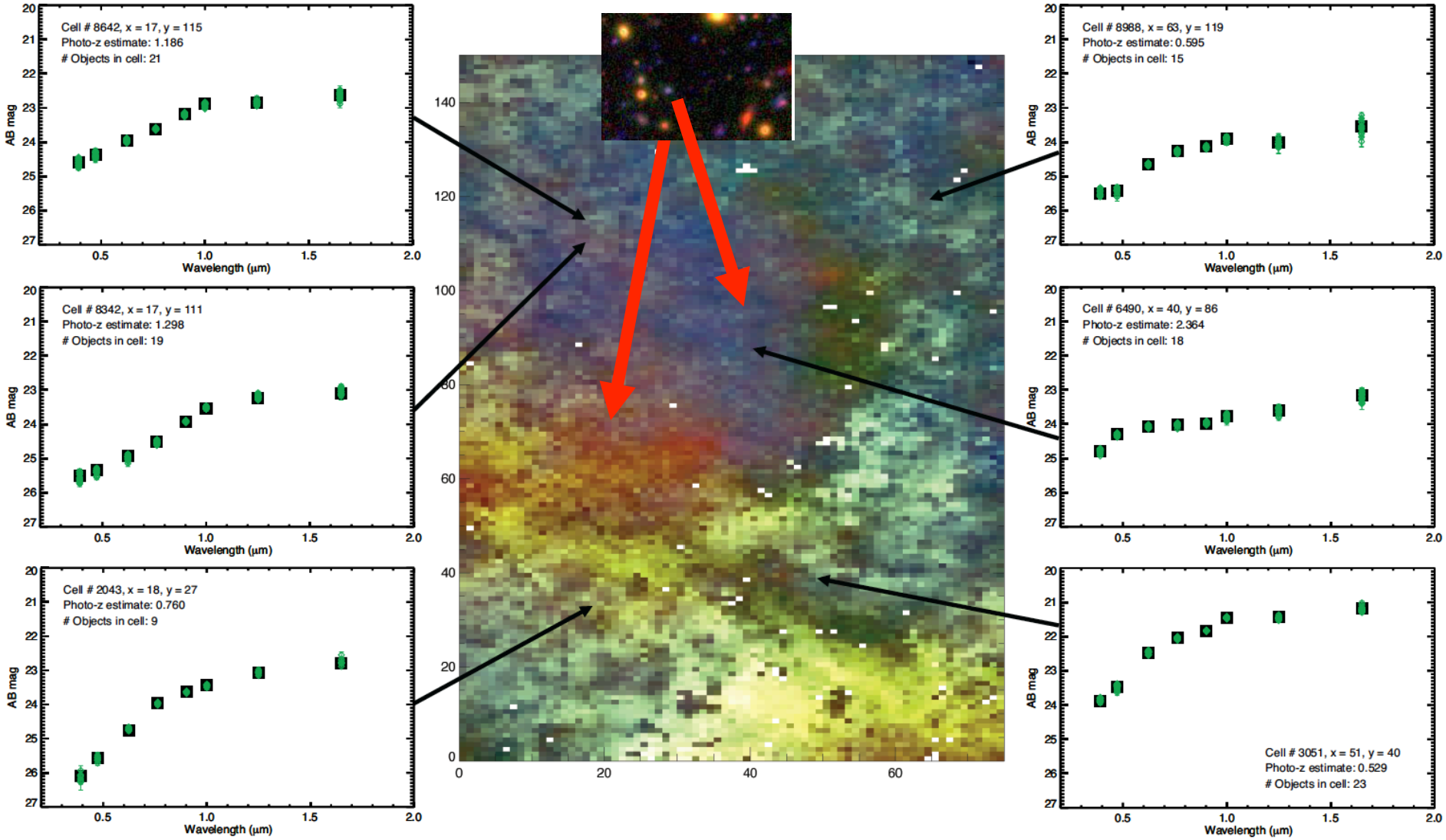
1. Initialized map is presented with training data, i.e. the colors of one galaxy from the overall sample.
2. Map moves towards training data, with the closest cells being most affected.
3. Process repeats many times with samples drawn from training set until the map approximates the data distribution well.

ple: Characterizing galaxy photometry with a Self Organizing Map



Masters, Capak et al. 2015

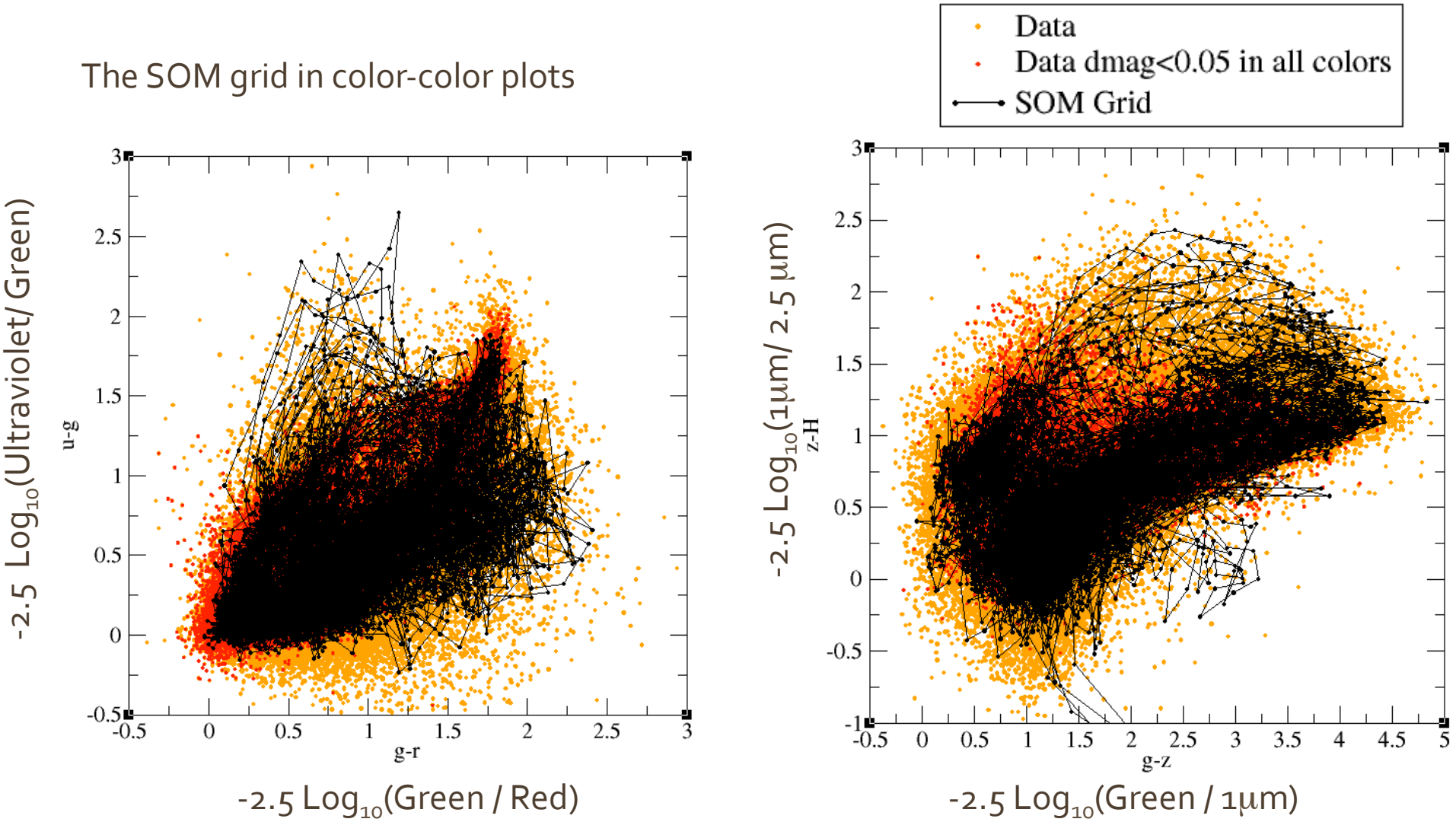
ple: Characterizing galaxy photometry with a Self Organizing Map



Masters, Capak et al. 2015

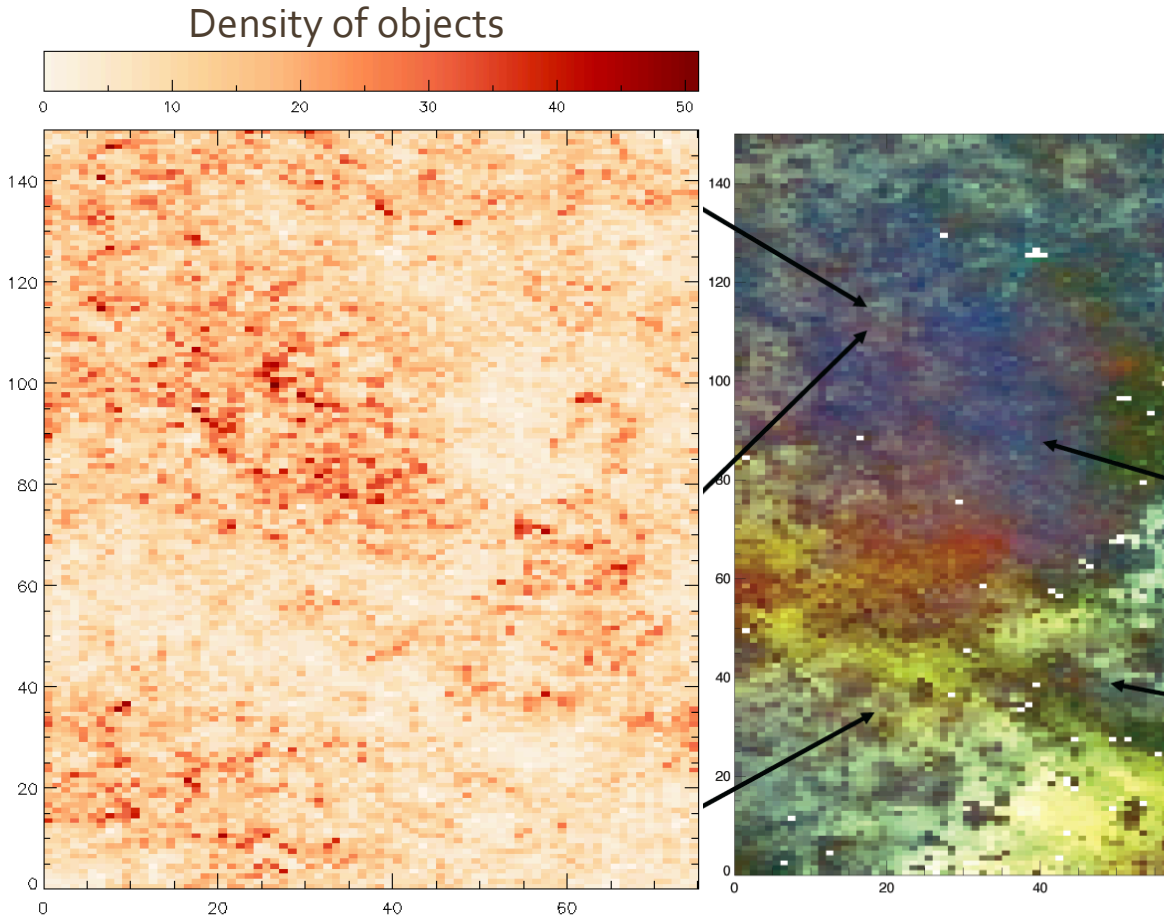
mple: Characterizing galaxy photometry with a Self Organizing M

The SOM grid in color-color plots



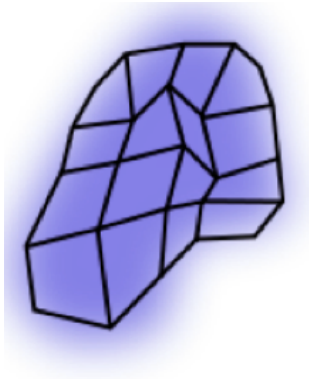
ample: Characterizing galaxy photometry with a Self Organizing Ma

- The SOM provides a map of the complex high-dimensional data space!
- The SOM parameterizes the large number of data points into a probability density field.
- We can now map our knowledge of the galaxy population onto that probability field.
- We could use an analytic model or a data model.

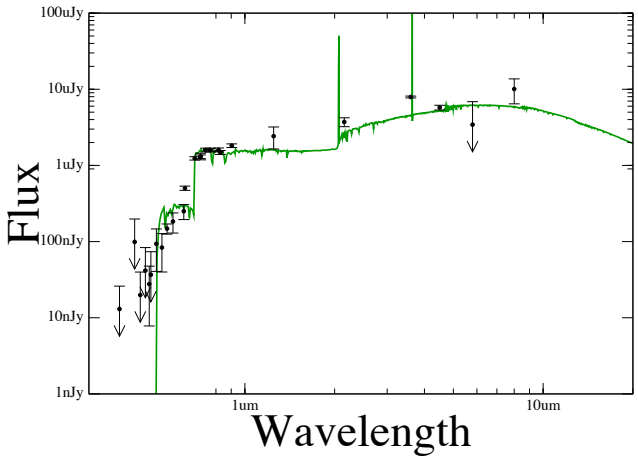


mple: Characterizing galaxy photometry with a Self Organizing M

- SOM provides a map of the data space
- Parameterizes the data into a probability density field
- A model provides a way to map that probability density field to a physical parameter
- Could be an analytic model or a data model



+



Photometric Redshifts can be calculated with a self organized map. Importantly, self organized maps tell you why certain objects are degenerate.

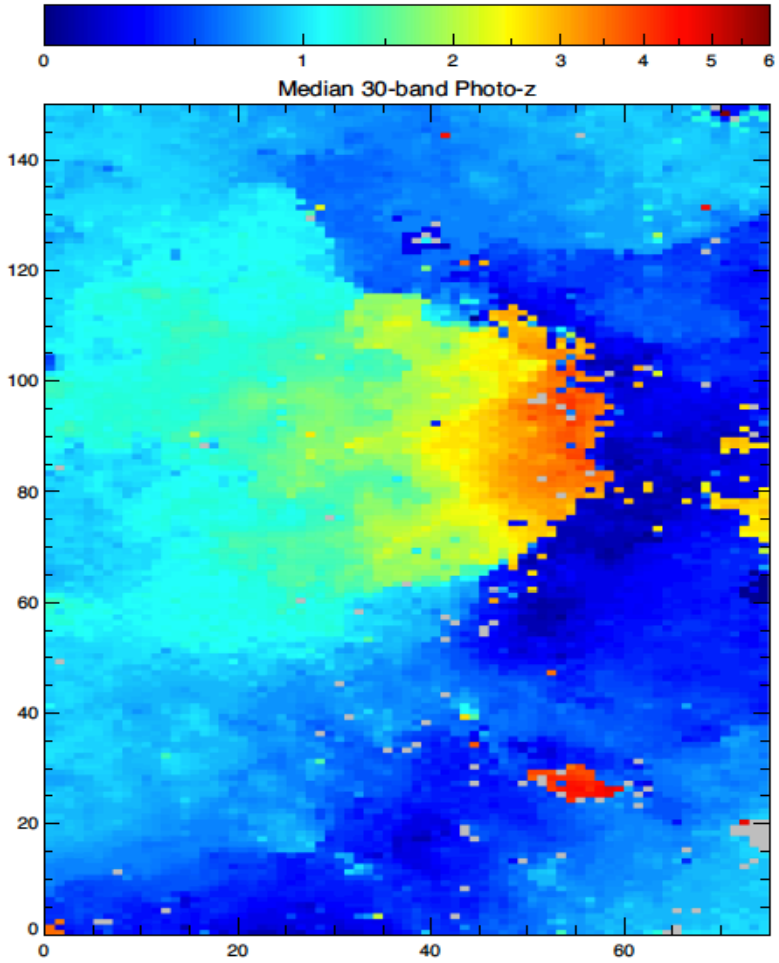
re we have colored the SOM
th the median photometric
dshift.

tice redshift varies smoothly
er most of the data space. This
why photometric redshifts work.

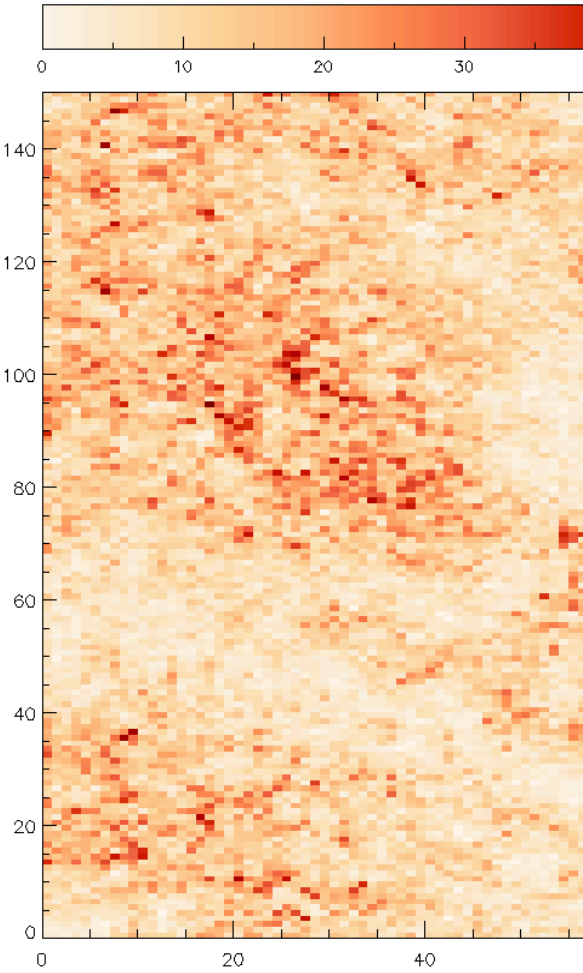
tice the caustics, this is why
ere are degeneracies in
otometric redshifts.

al. 2015, 2017

Mean Redshift

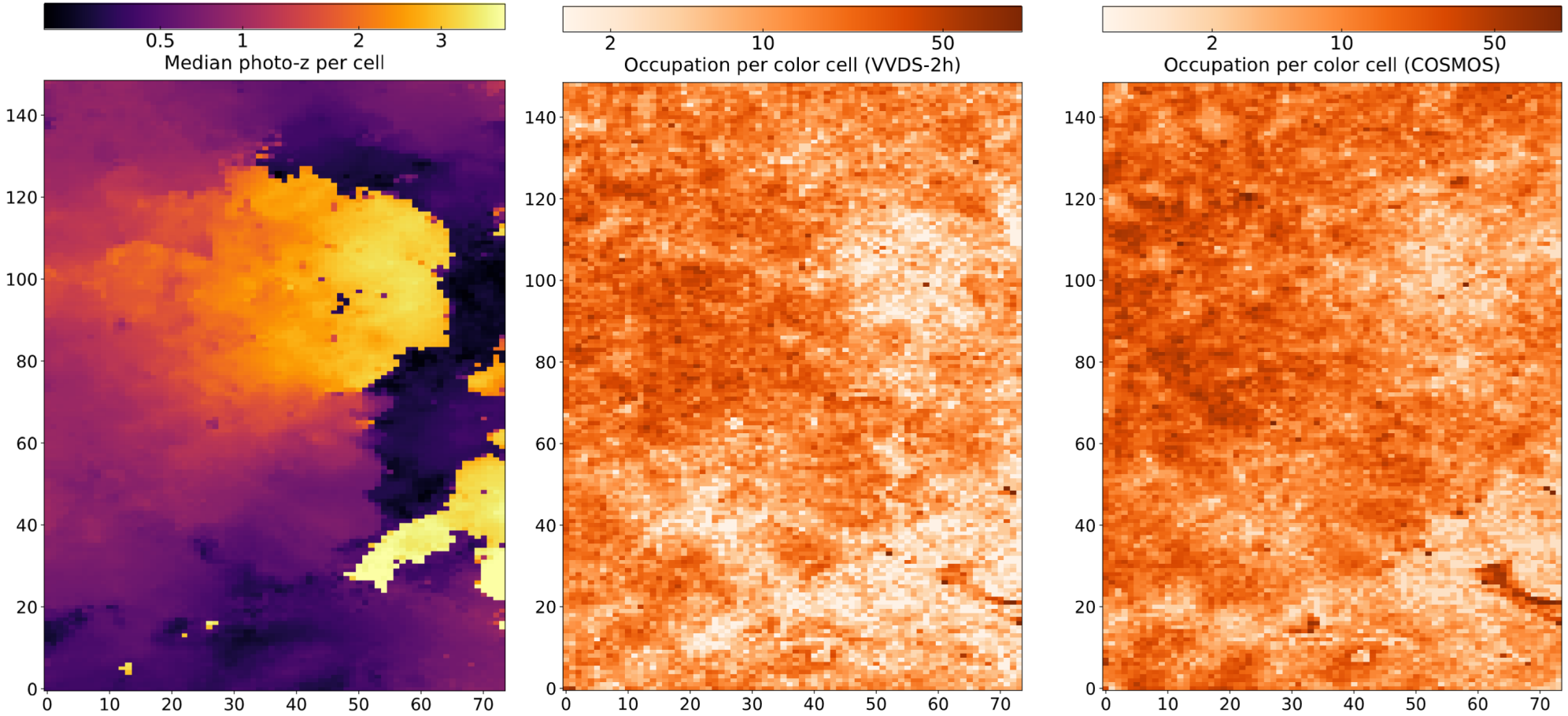


Density of objects

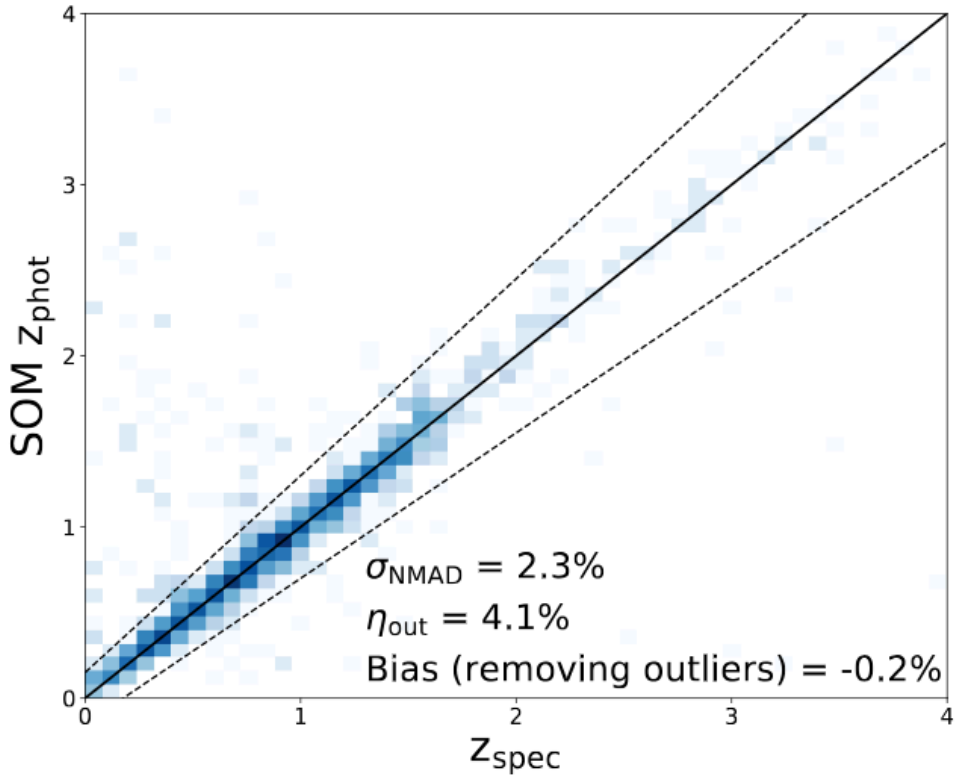


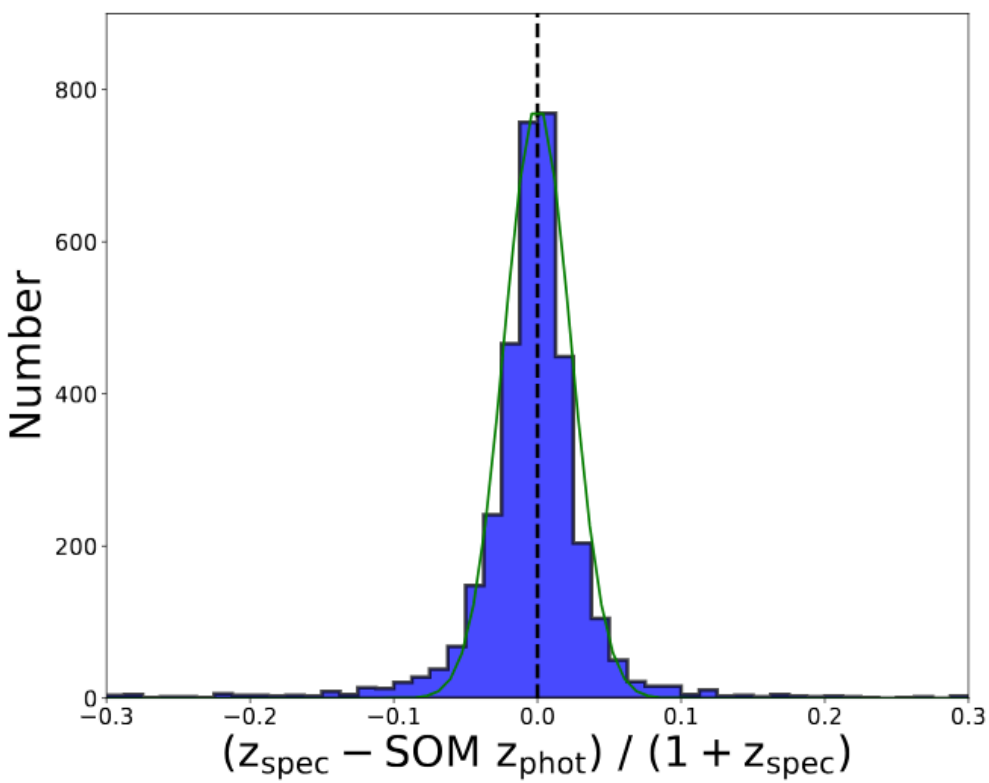
The data density field contains fundamental information about how galaxies evolve.

e density field contains information on the space density of objects because the color is strongly rrelated with redshift. This measures cosmic variance between fields empirically.



Redshifts estimates from color mapping are more accurate than from SED fitting because they use ensemble information





This analysis does NOT use spec-z training!

→ Method achieving unbiased performance

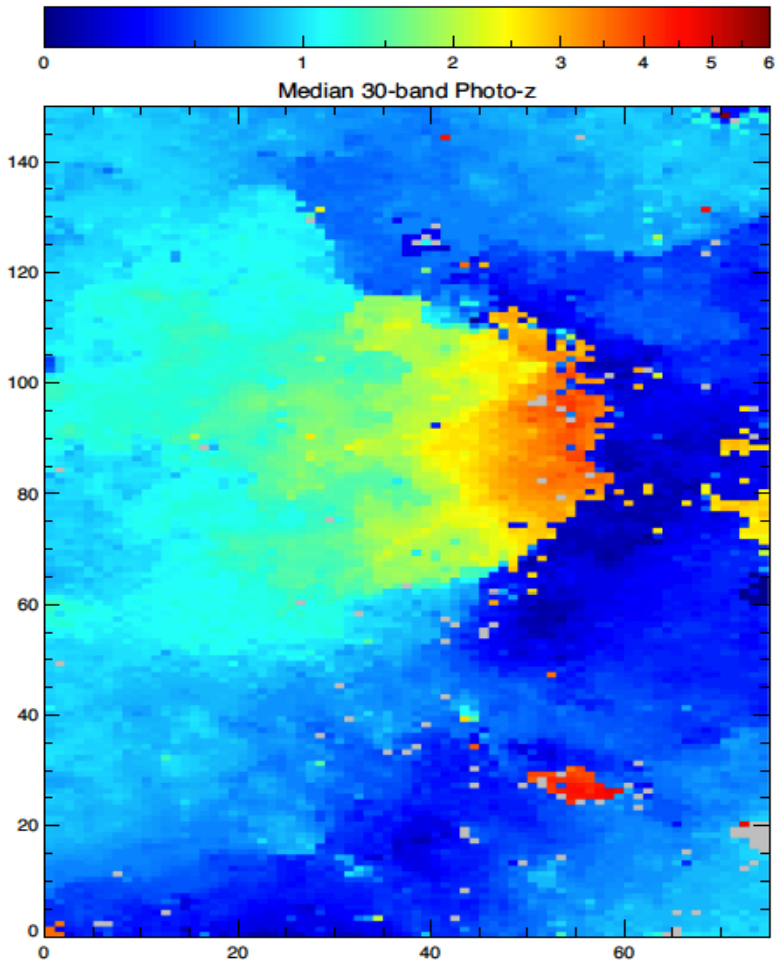
→ Outlier fraction 4.1%, scatter 2.3%, bias of -0.2%

A data map like the SOM quantitatively tells you how well you are sampling the underlying galaxy population.

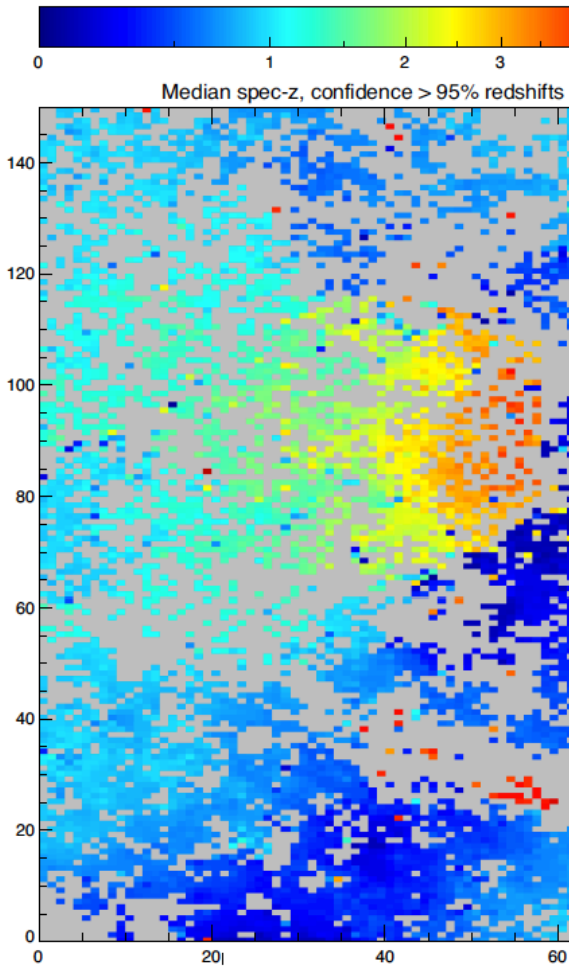
e can compare our photo-z
odel to spectroscopic data.

e agreement is good, but
rrent spectroscopic samples do
t cover a large fraction of the
or space occupied by galaxies.

Photometric Redshift

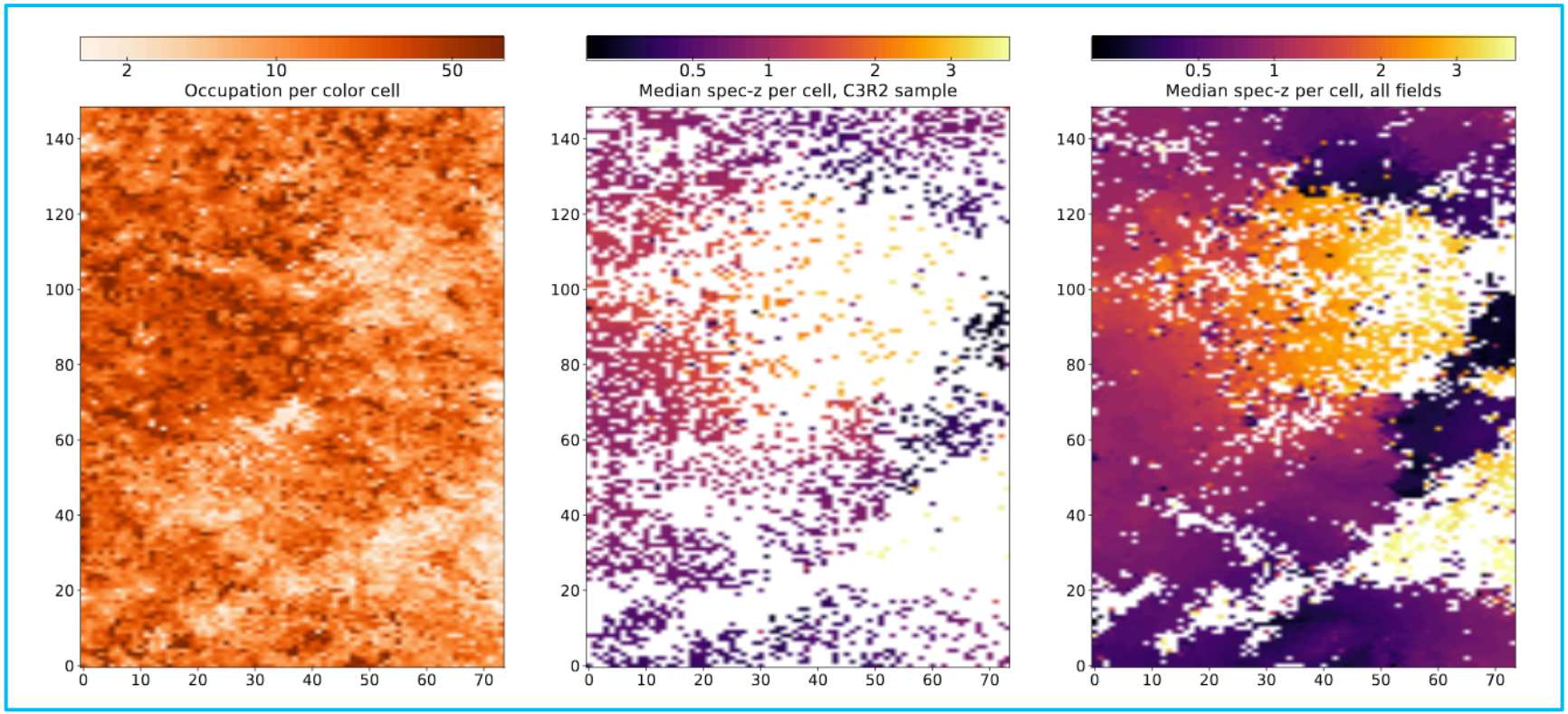


Spectroscopic Redshift



al. 2015, 2017

We are undertaking the C3R2 survey to fully sample the color space with spectroscopy



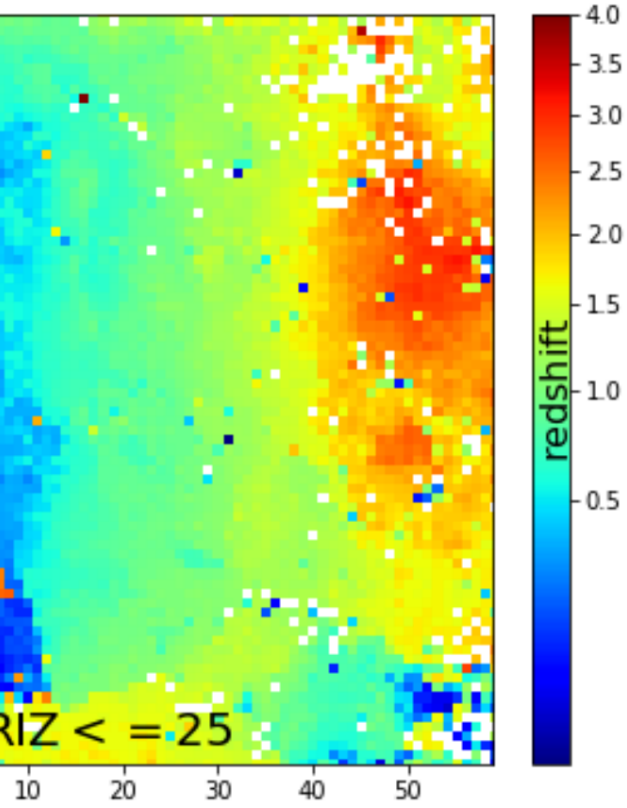
- To date C3R2-Keck covers >35% of the color space, disjoint from what was previously explored
- C3R2 has covered >75% of cells with >85% of galaxies in cell with at least 1 specz, many cells with >>1 specz
- Uncovered cells correspond to less common sources, targets of future follow-up

t al. 2019 (under review)

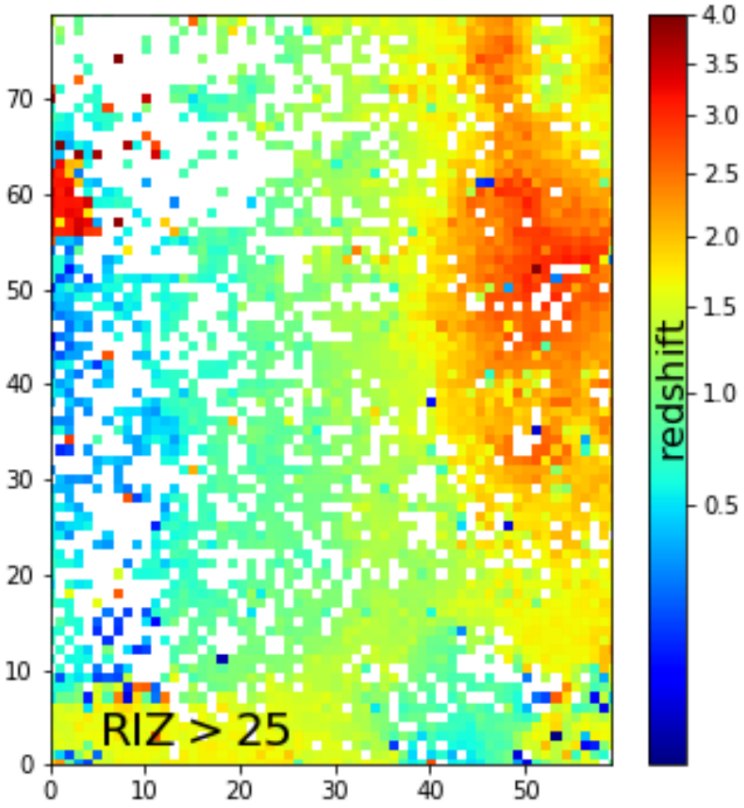
The color space does not strongly evolve with object brightness so we do not need to observe many faint objects.

A comparison of Euclid and WFIRST lensing samples

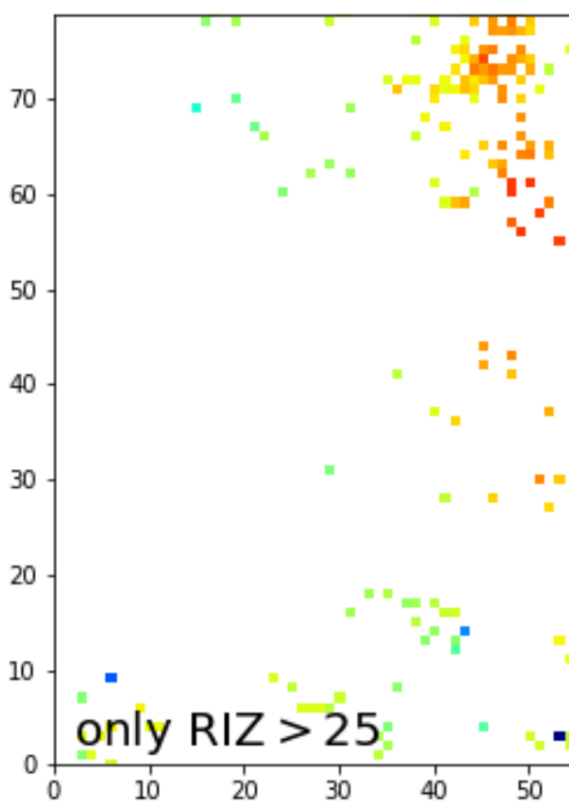
Bright



Faint



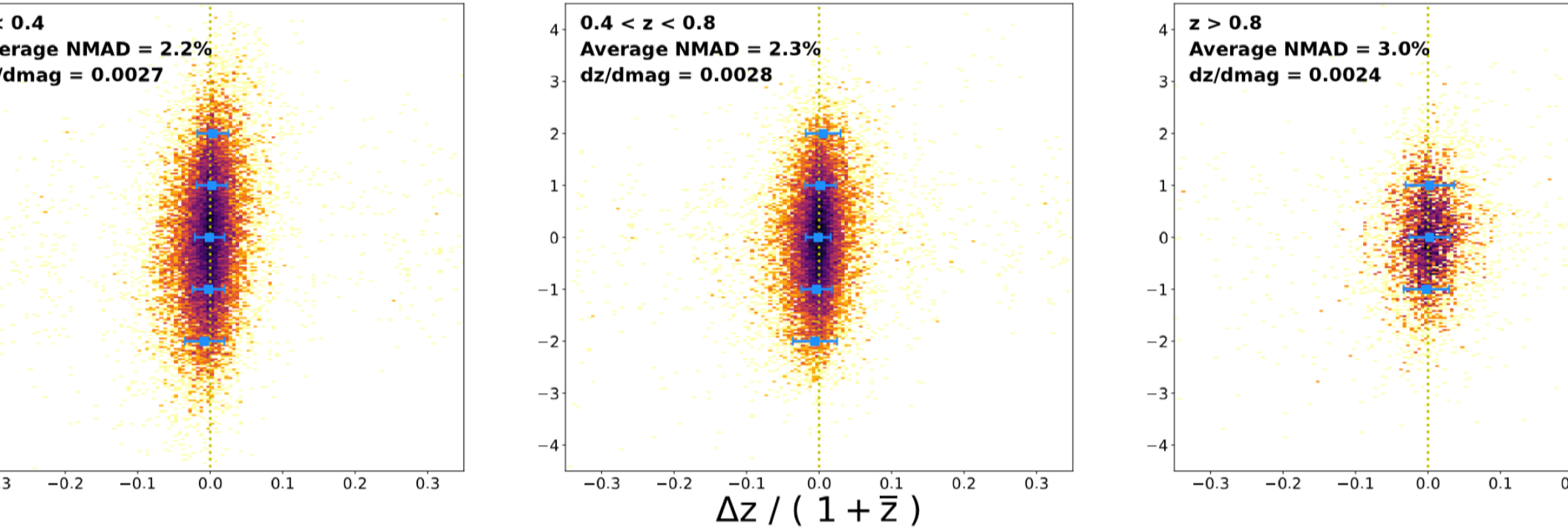
Unique to faint s



et al. 2019

There is a systematic effect with magnitude, but it is well controlled and present in simulations.

A comparison of spectroscopic redshifts at fixed color but varying magnitude



This same effect is present in simulated data at similar levels

Our present thinking is it's a noise or selection bias effect (Speagle et al. in prep)

t al. 2019 (under review)

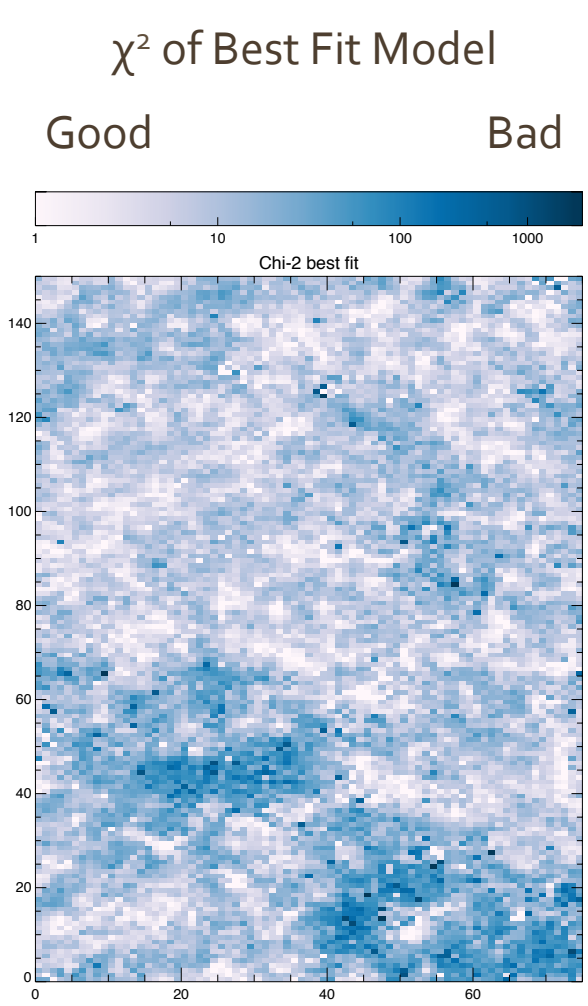
The data map can quantitatively tell you how well a model maps your data.

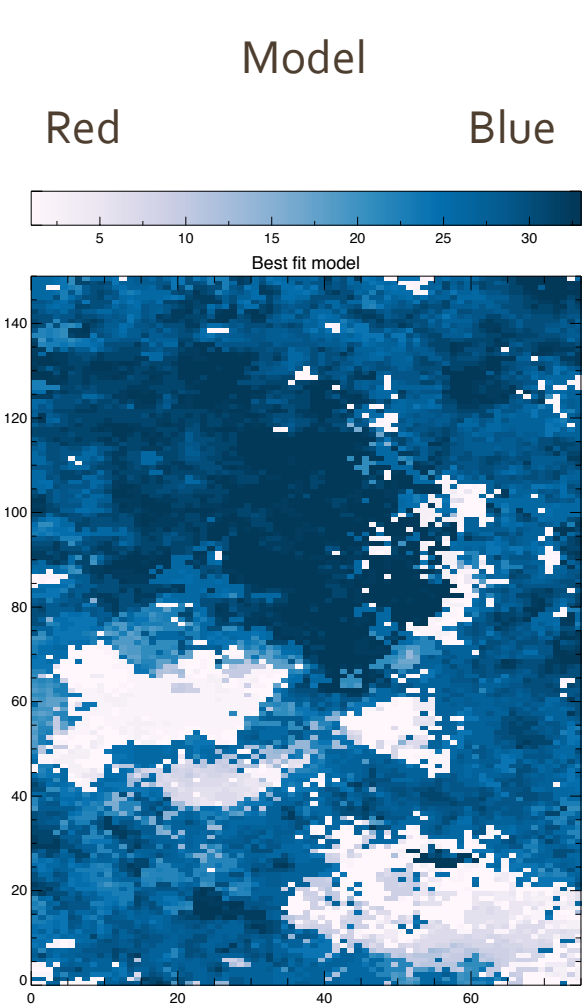
e we are showing
best fit model at
point on the
M.

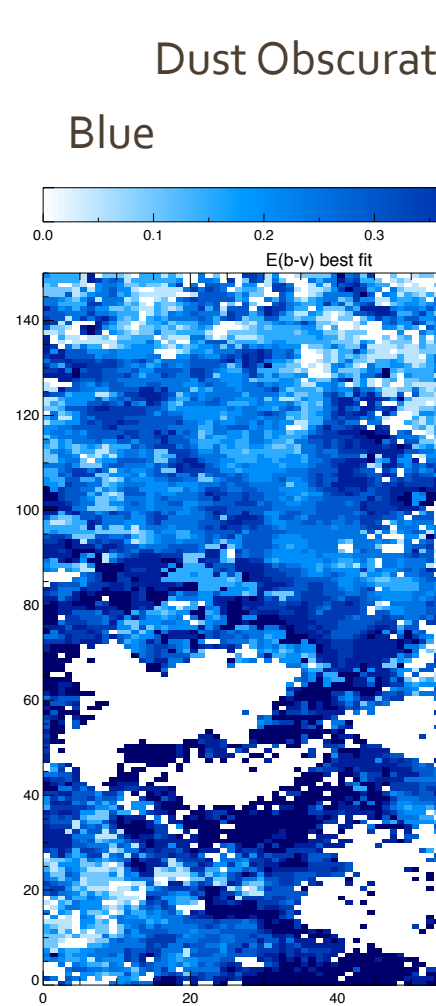
clear there are
ons where the fit
uch worse.

SED model is not
d in these regions
the parameters
be unstable.

l. in prep







data map can
titatively tell you how
a model maps your data.

omputational models match the
universe?

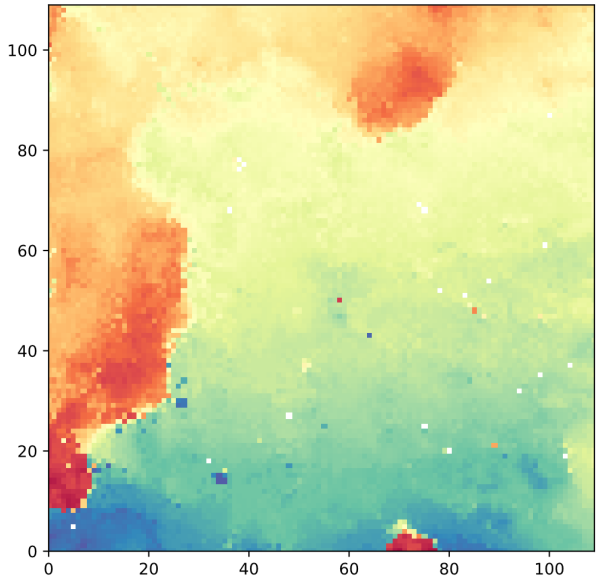
tly, but there are large differences

zon AGN:

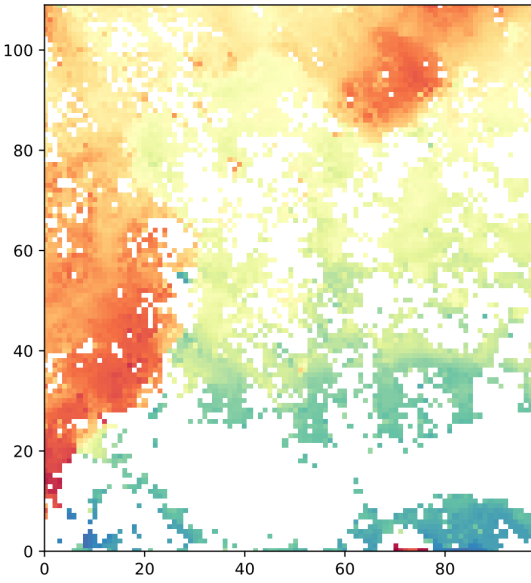
es not predict some populations of low-z
axies observed in COSMOS.

dicts z>1 galaxies that are not observed to
st.

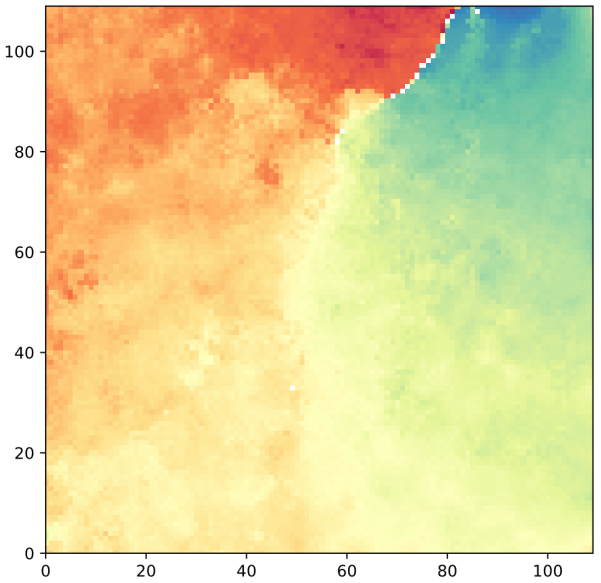
et al. in prep



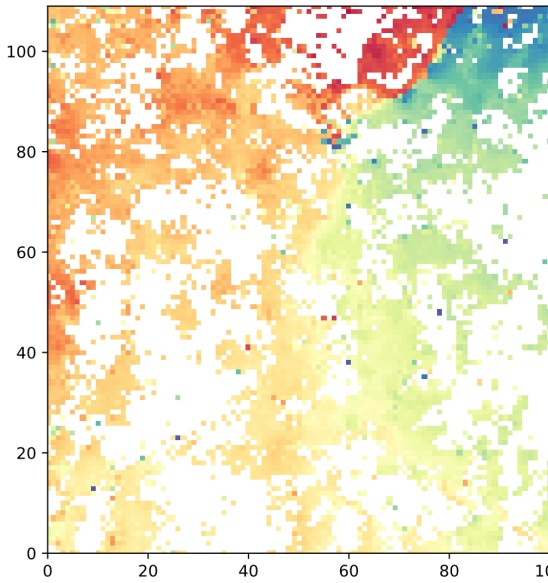
SOM trained with COSMOS...



...and then HORIZON-AGN galaxies a



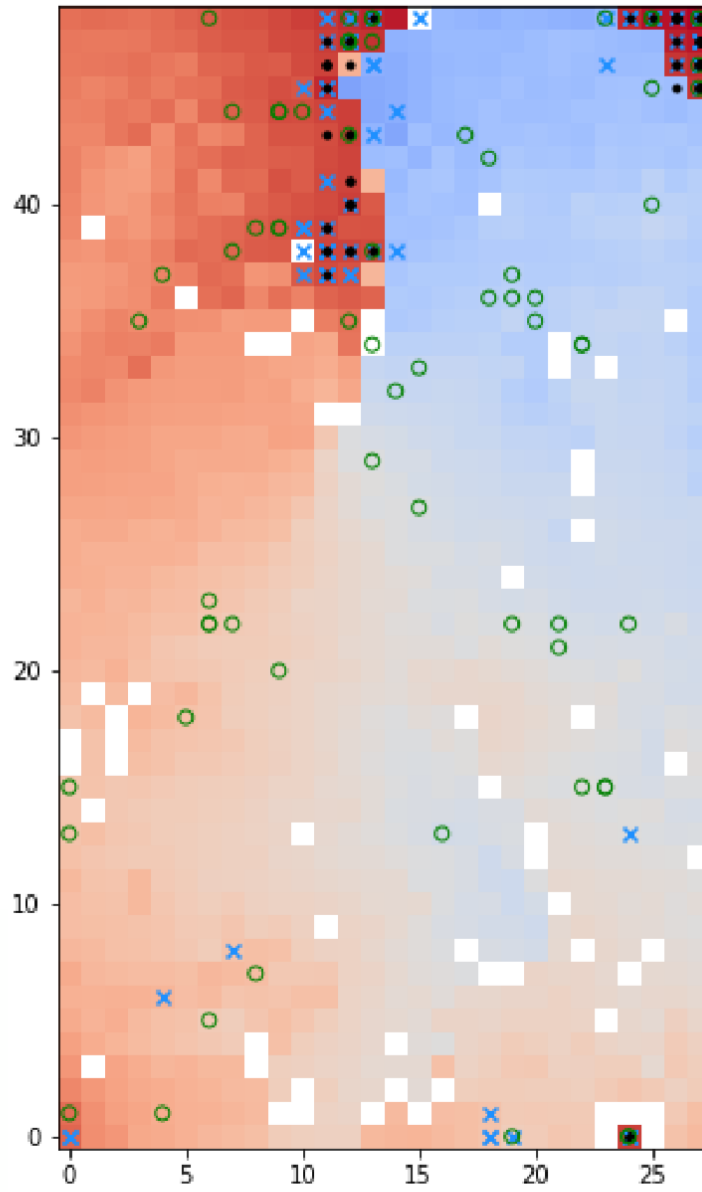
SOM trained with HorizonAGN...



...and then COSMOS galaxies are mapp

The data map can quantitatively tell you which galaxies scatter into and out of your selection function.

- Compare intrinsic colors with noisy data from different possible surveys (DES, HSC, Euclid)
- We know the density of different observed types of objects
- Can self-consistently test the completeness and contamination in the high-z sample



● intrinsic colors

○ DES+Euclid

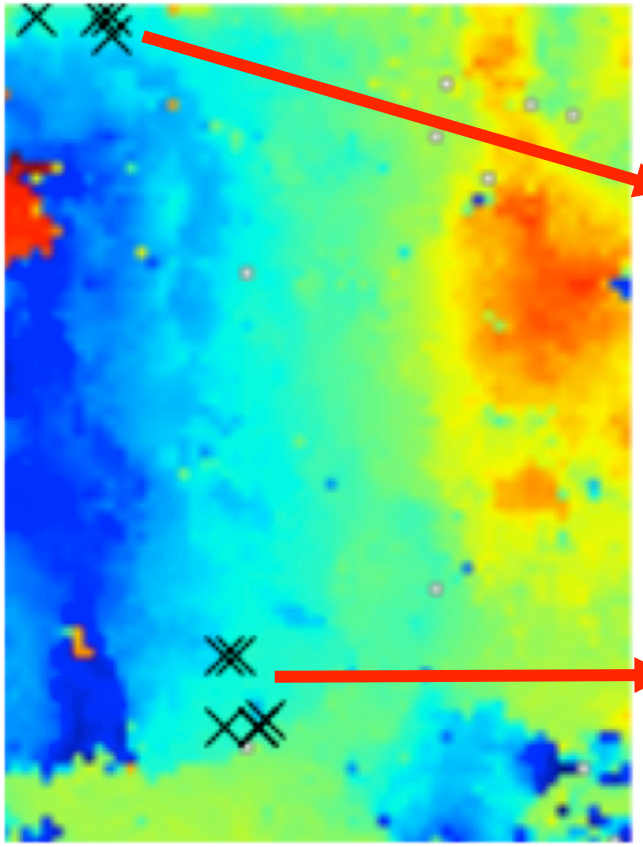
x HSC+EuclidDeep

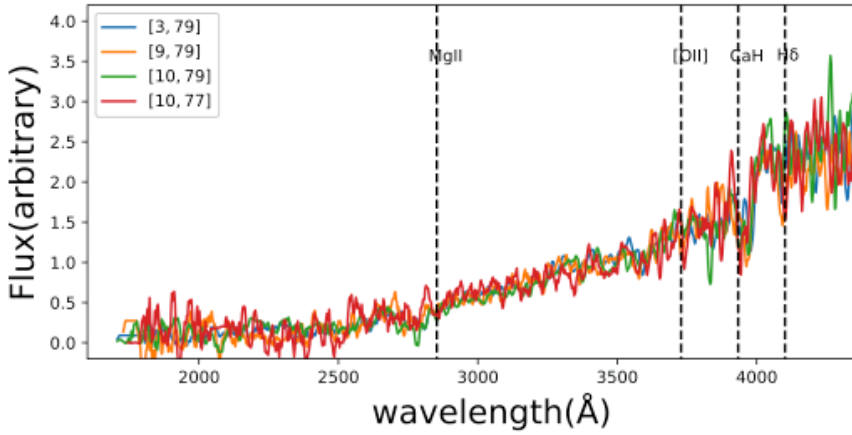
et al. in prep

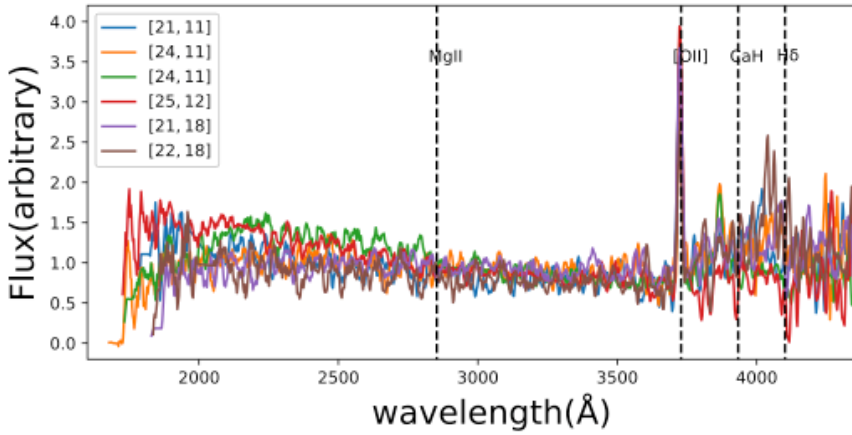
We can also map the detailed properties of galaxies across data space.

The color is strongly correlated with the high-resolution spectra

Typically sensitive to 100Å equivalent width variations



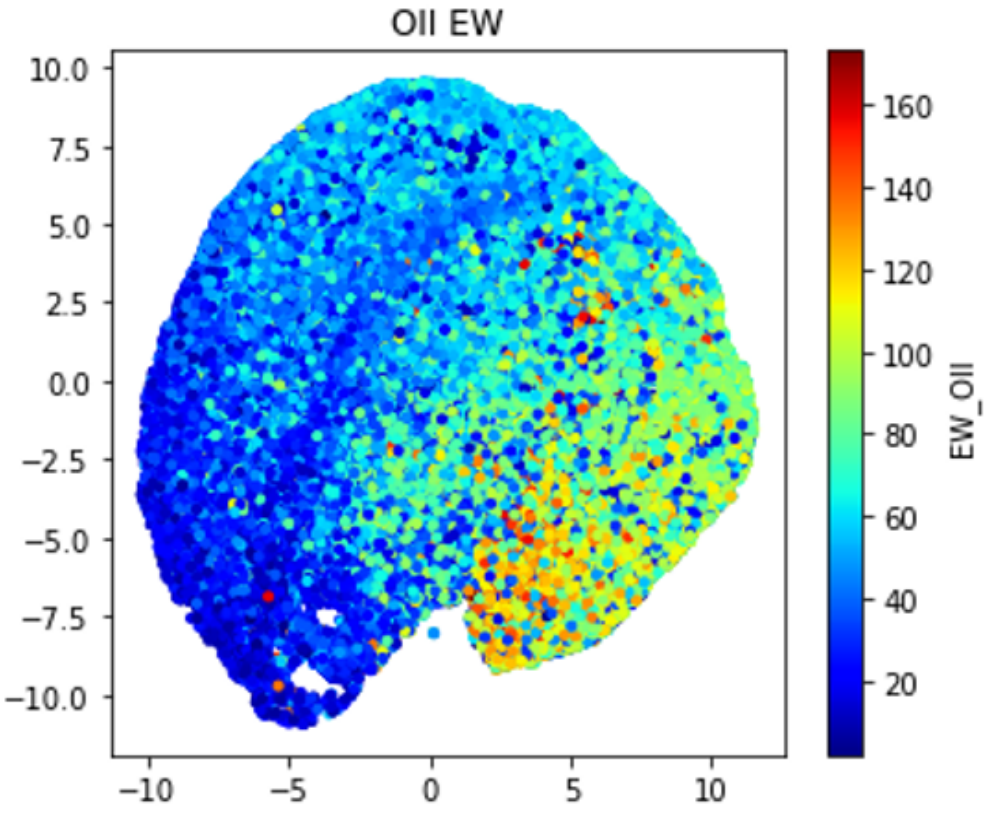




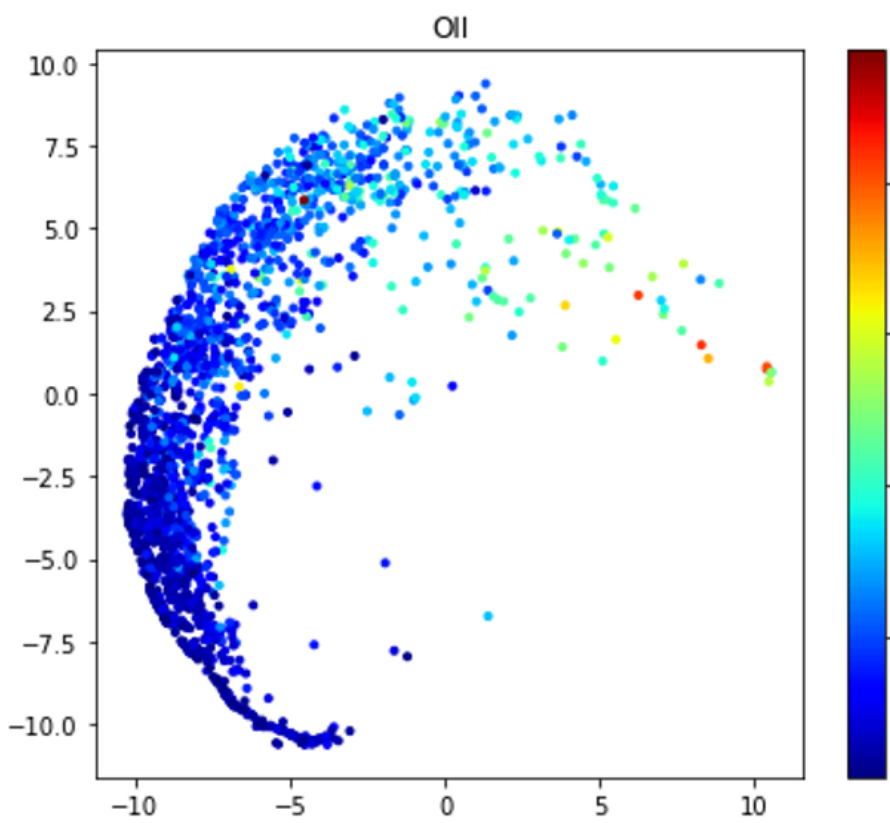
ti et al. 2019

OII Equivalent width can be predicted from photometry.

Model

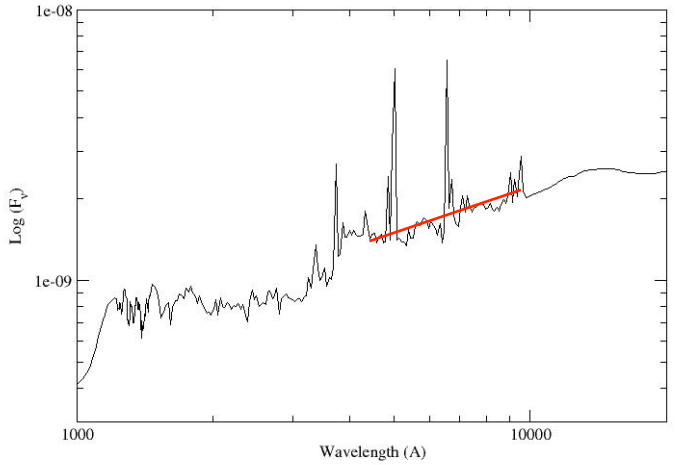


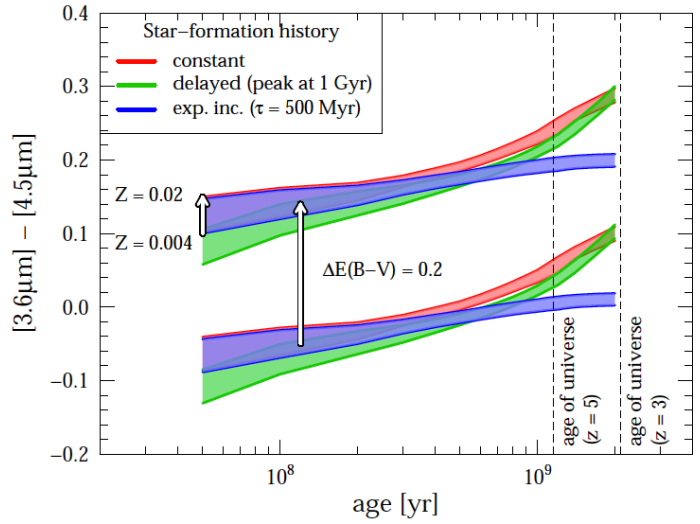
Data



We can measure other emission lines from photometry too.

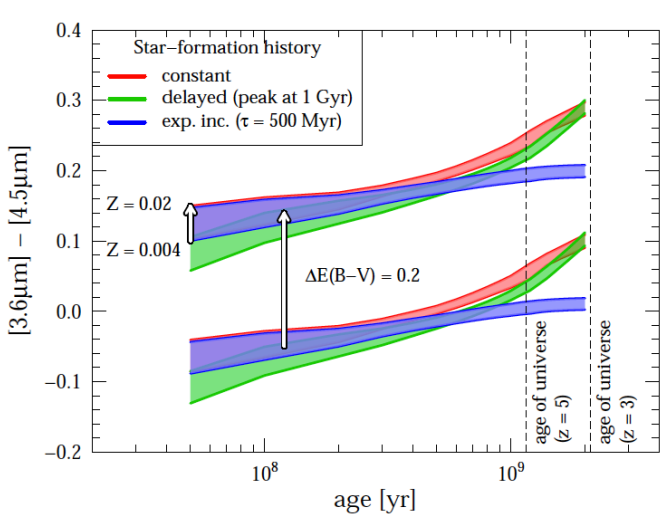
- Current emission line estimates are very model dependent
- However, they can be measured statistically from the photometry
- Just need to choose estimator carefully to account for galaxy physics
- Started with a case study at (z>5) because there is no other way to measure lines there

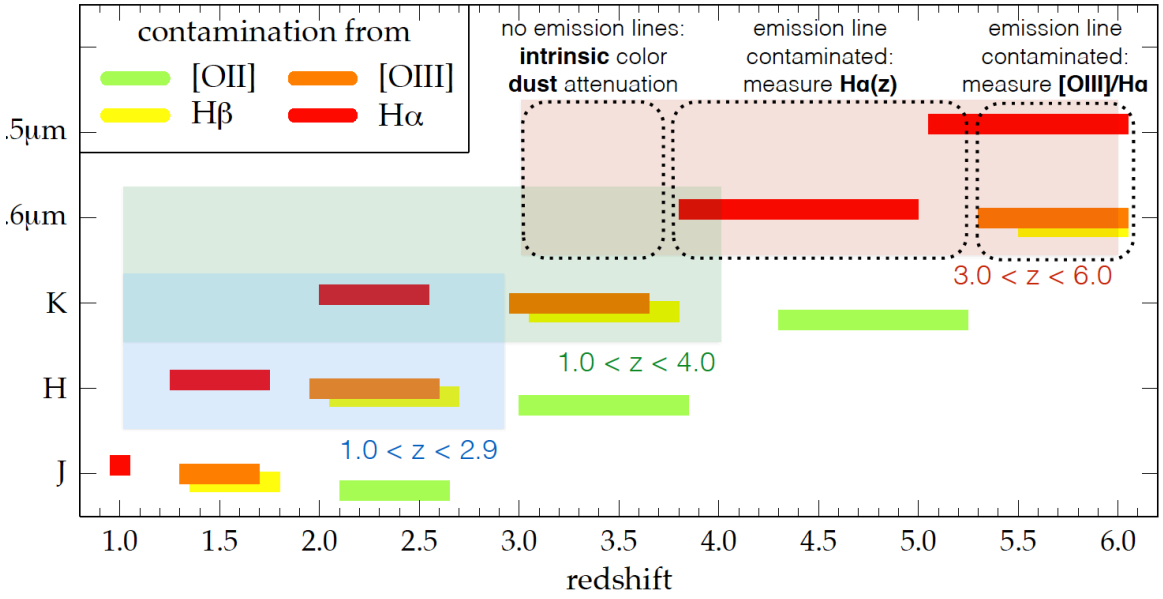




Faisst, Capak et al.

One use is to estimate rest-frame optical line emission for cosmology planning and JWST targeting.



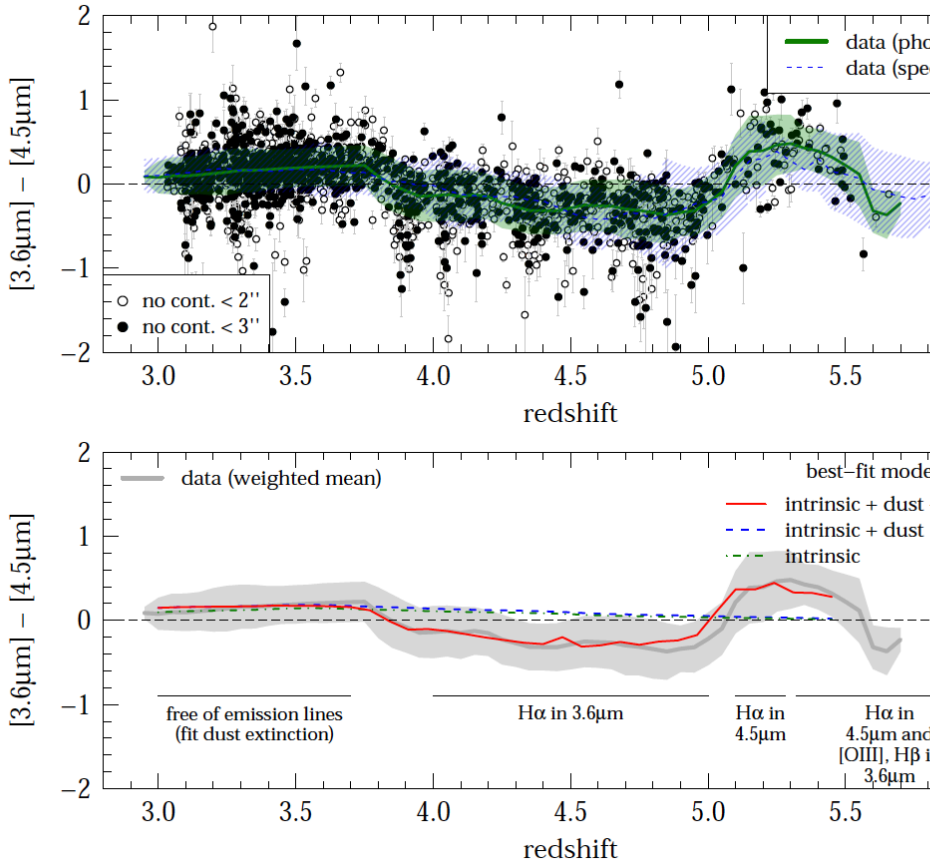


- Can measure Hα EW (and other lines) from photometry if redshift is know
- Redshift from color manifold (SOM)
- Create a forward model of the galaxy population including lines in colors

Faisst, Capak et

One use is to estimate rest-frame optical line emission for cosmology planning and JWST targeting.

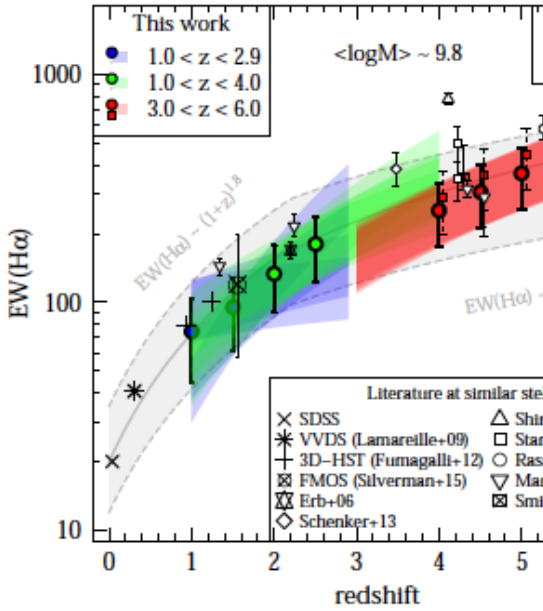
- Can clearly see signatures in raw photometry
- Both in spec-z and photo-z sample
- Fit our forward model to the data to extract line EW

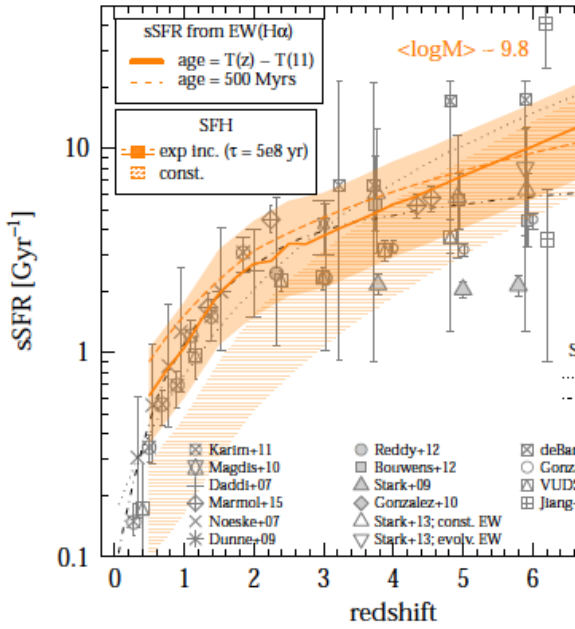


Faisst, Capak et

notometric line distributions estimates agree th spectroscopic ones.

- Estimate Hα EW distribution agrees with direct measurements
- Evolution at high-z is consistent with model fitting results
- Derived physical properties (specific star formation rate) also agree

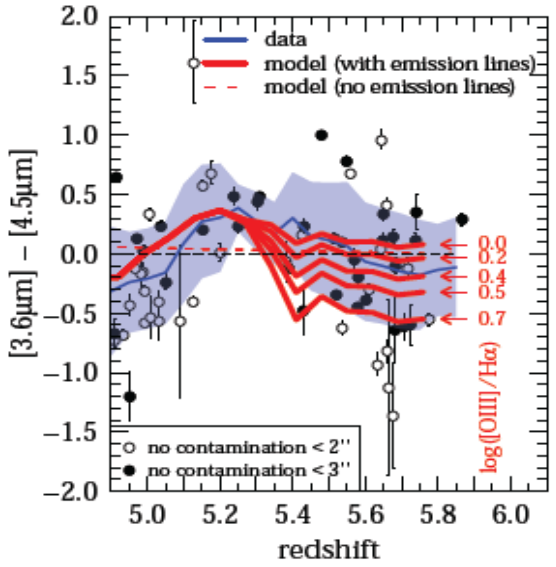


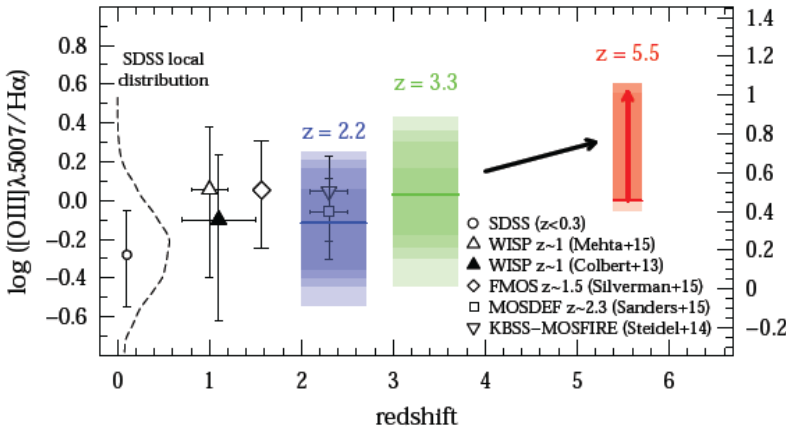


Faisst, Capak et al. 2016a

There are several lines we can estimate for JWST targeting.

- O[III]/Hα is also measured by our forward model
- Consistent with other estimates and measurements





Faisst, Capak et al. 2016a

have successfully estimated Hα for individual galaxies at 3.9<z<
se estimates can be used to target JWST spectroscopy

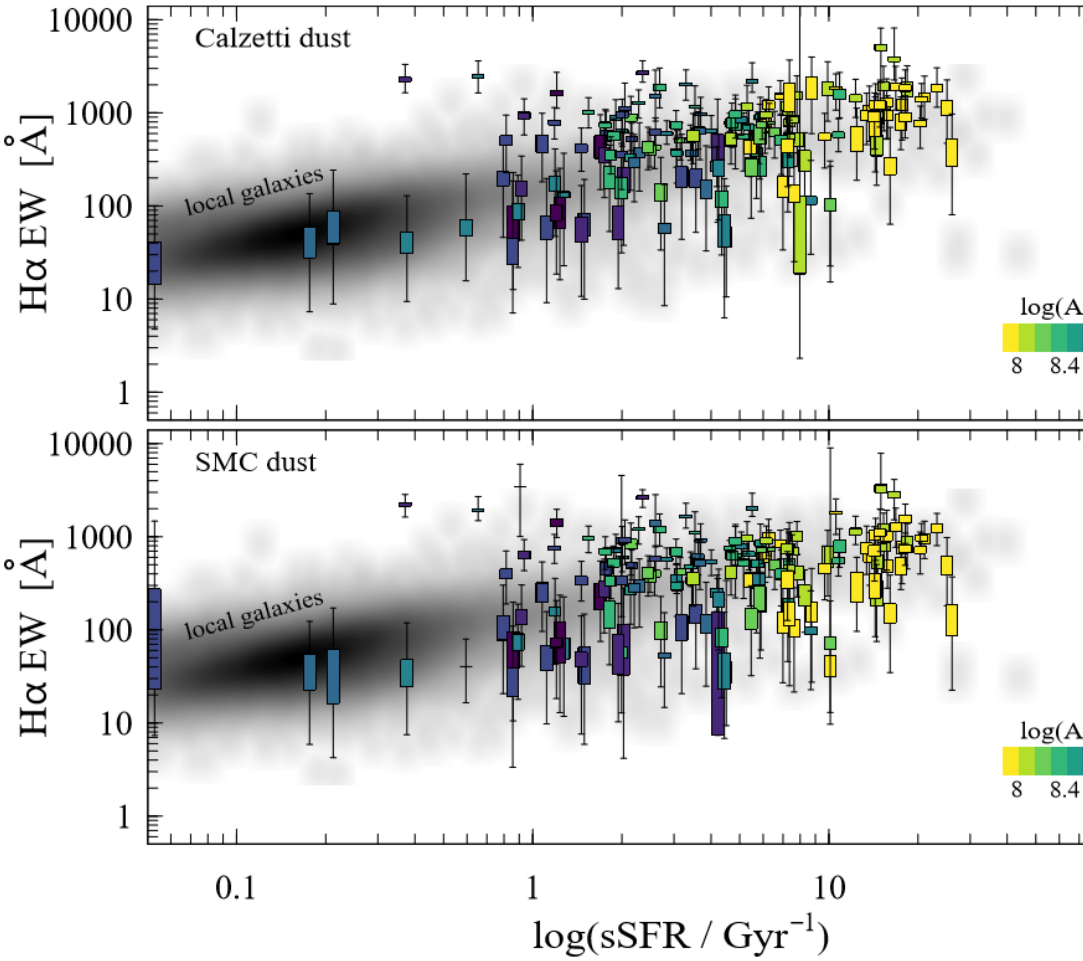
g just the 3.6-4.5um color + UV slope we estimate the Ha equivalent width

model dependent than full model fit, but itive to assumptions about dust properties

- th extinction curve and relative line to continuum inction
- ese can be calibrated with JWST eventually

ribution of recovered values consistent what is seen in SDSS

sSFR galaxies have higher equivalent hs and younger ages as expected



l. in prep

an independently measure the main ence and see possible indications of ching at z~4

s an indicator of prompt (<10Myr) old star formation

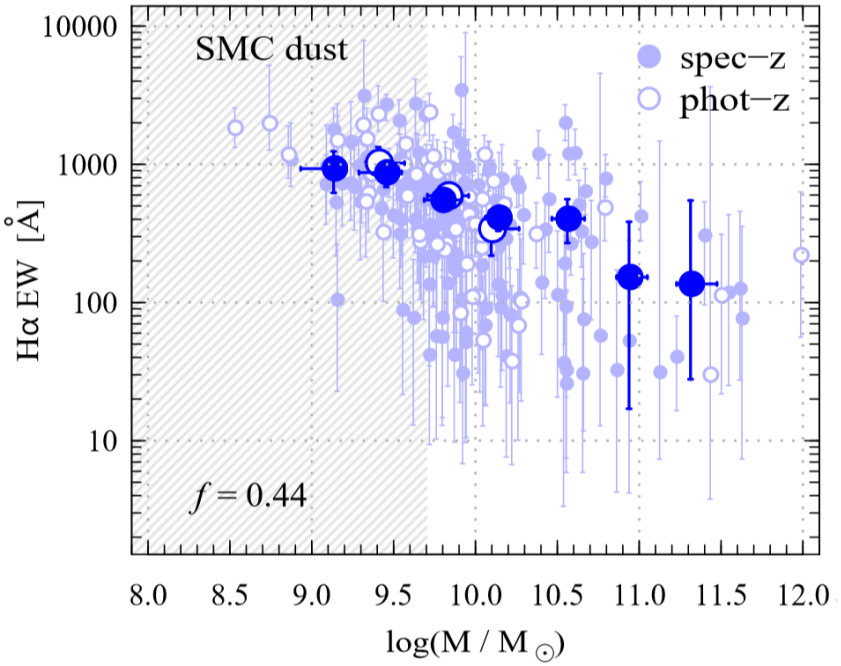
s an indicator of the SFR over ~100Myr

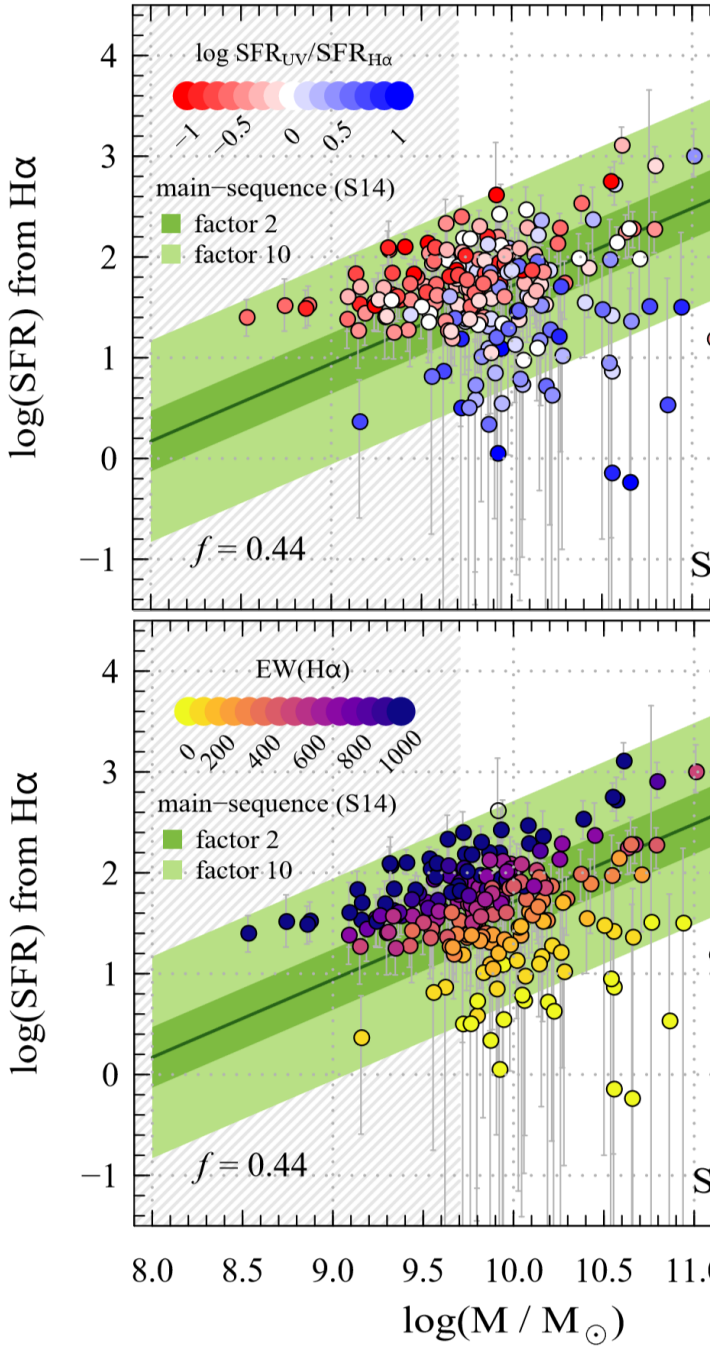
he ratio of Hα to UV is an age indicator

scatter in the Hα n sequence is larger n the UV one cating bursts of star nation at z~4

gradient in the UV/ SFR indicates cycling oss the MS and/or nching

. in prep





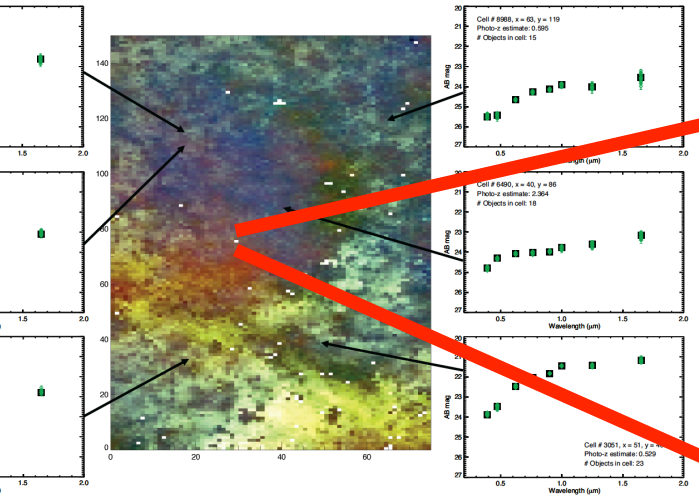
ls can be gridded in the data space to infer the physical parameters of ob

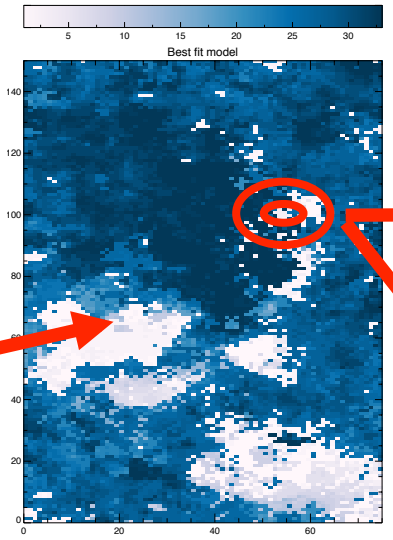
Physical Models

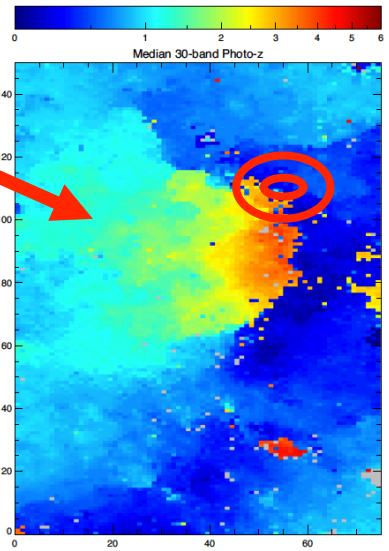
Red

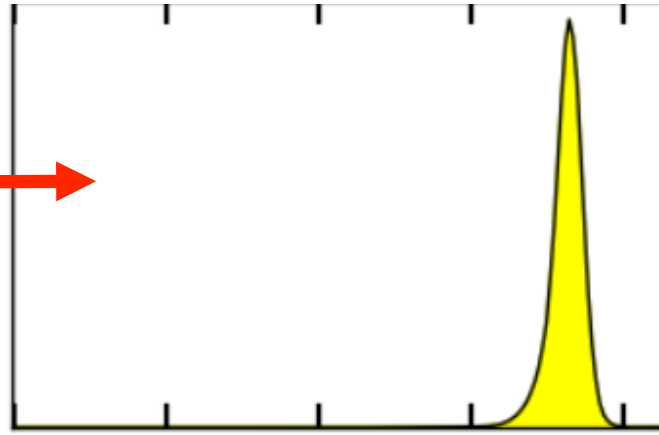
Blue

Data Model

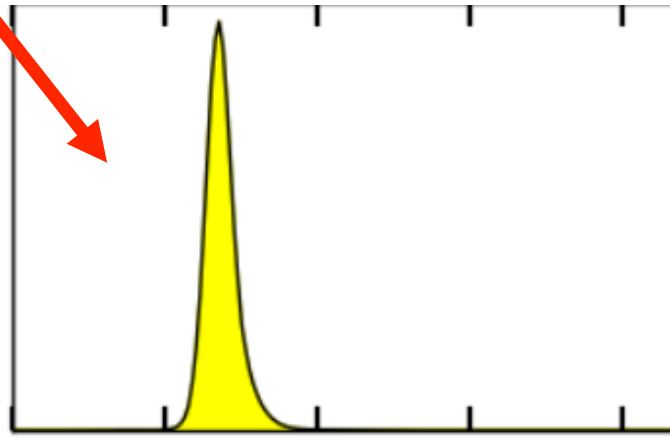








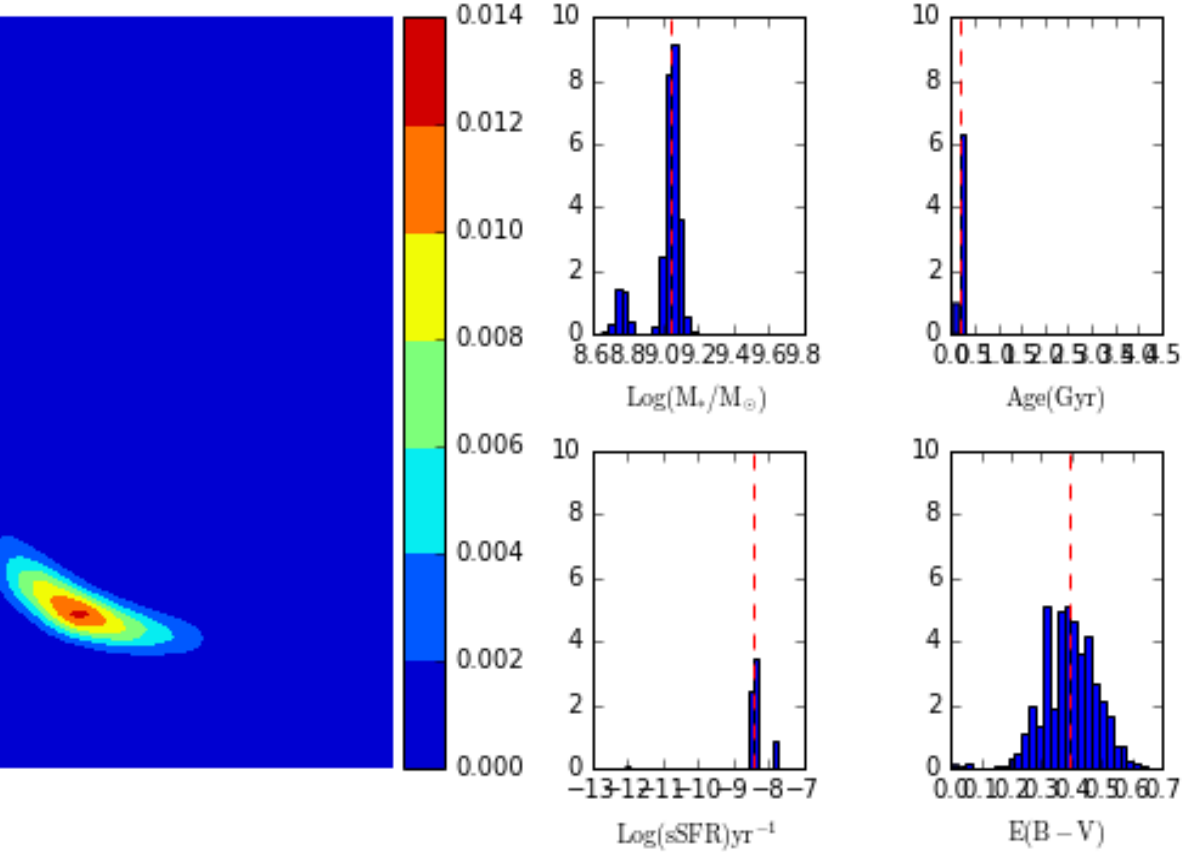
Age



Mass

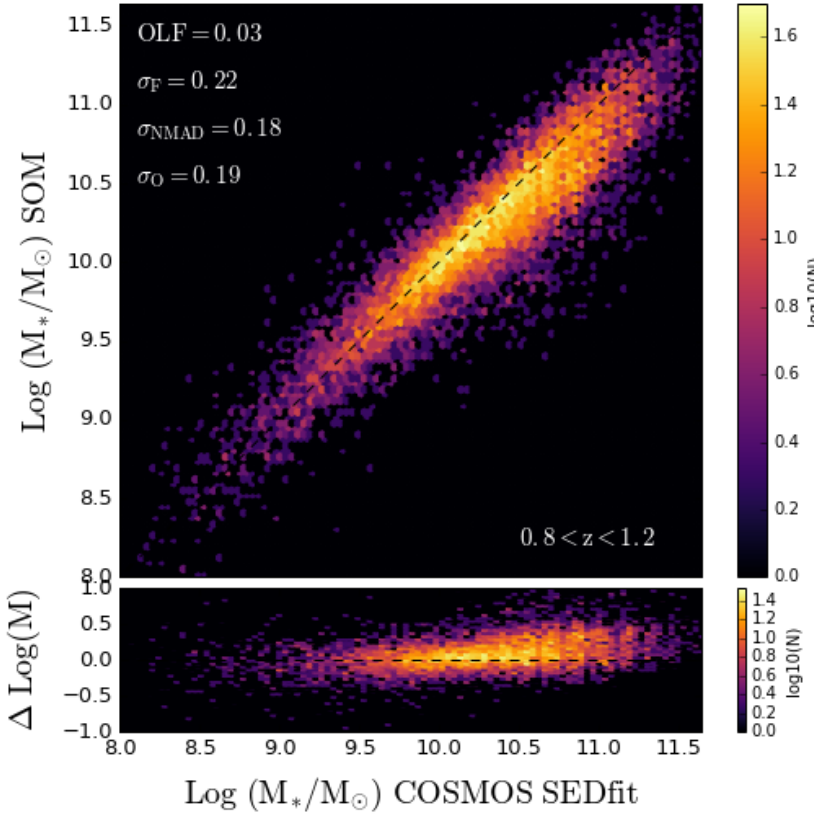
In our early tests we can recover physical parameters almo as well as full model fitting, but orders of magnitude faste

ntegrate Photometric Error Against SOM



nati et al. in prep

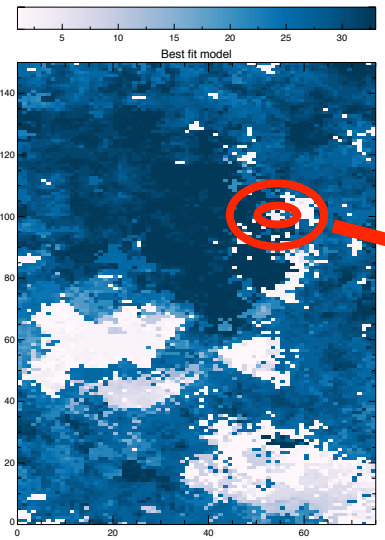
Comparison to SED Fits



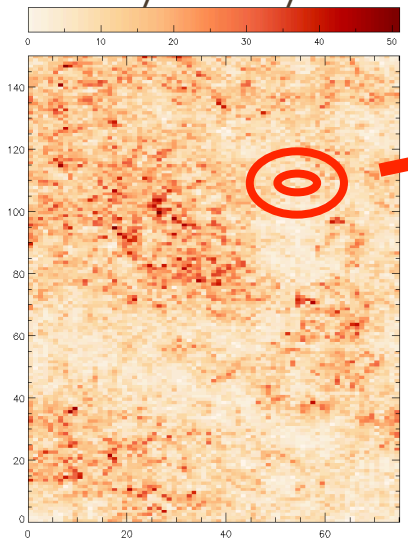
This means we can now go directly from d space to mass functions for many complex models.

Physical model of galaxies

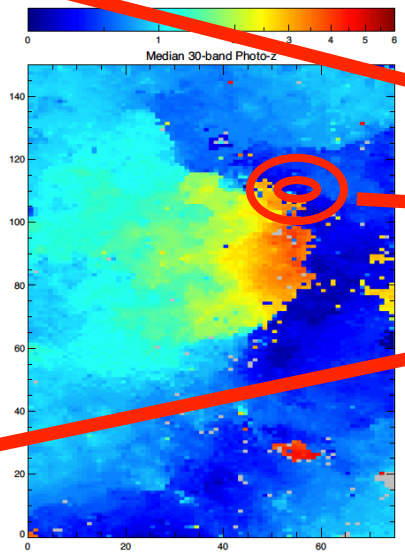
ed Blue



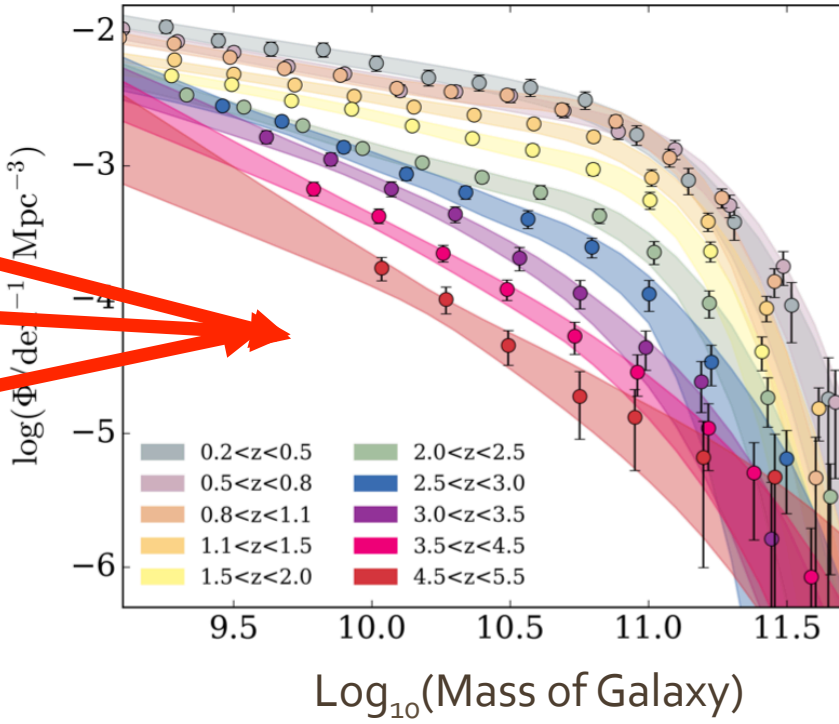
Sky Density



Redshift Model

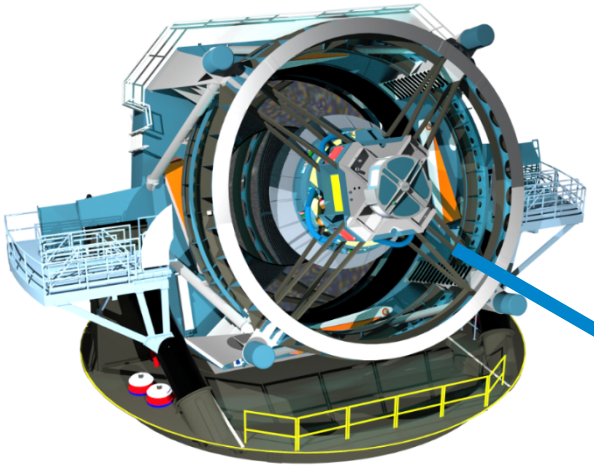


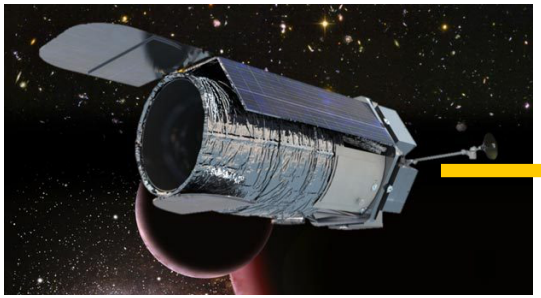
Mass Function

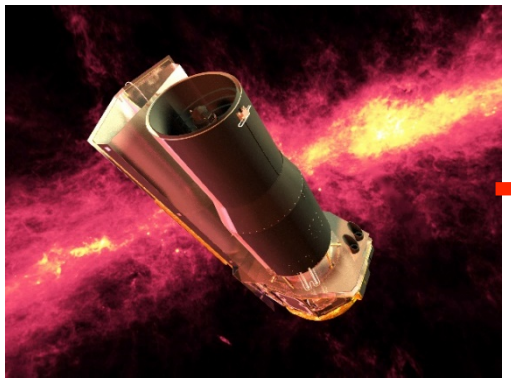


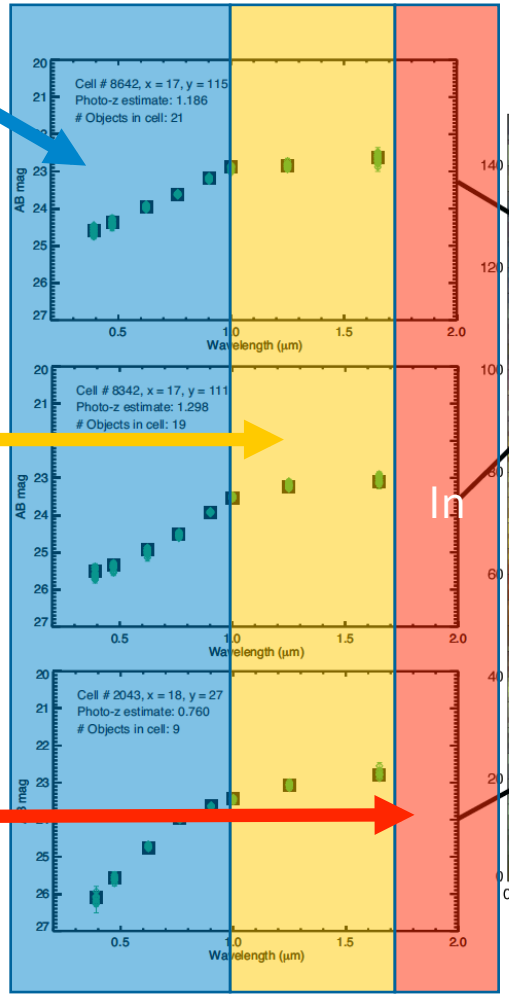
It is possible to statistically combine data s

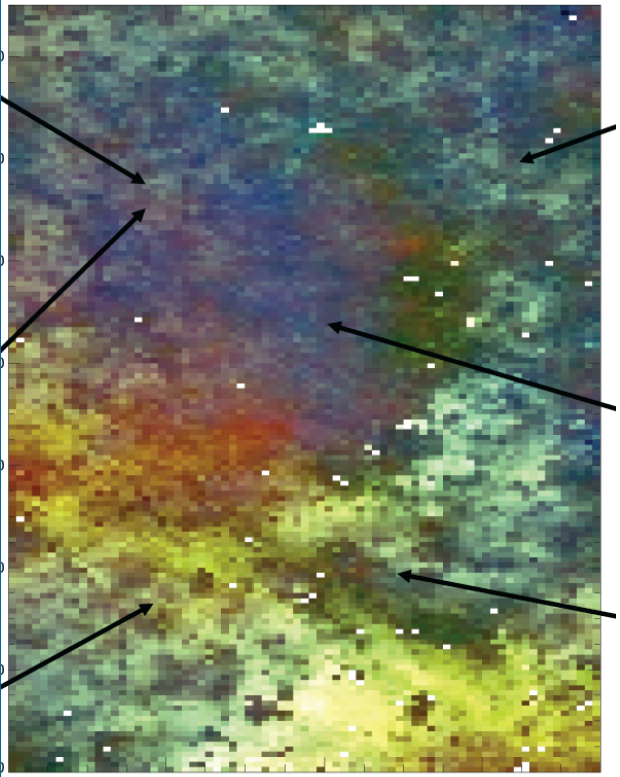
- Generate model with one data set
- Expand it to higher dimensions with another





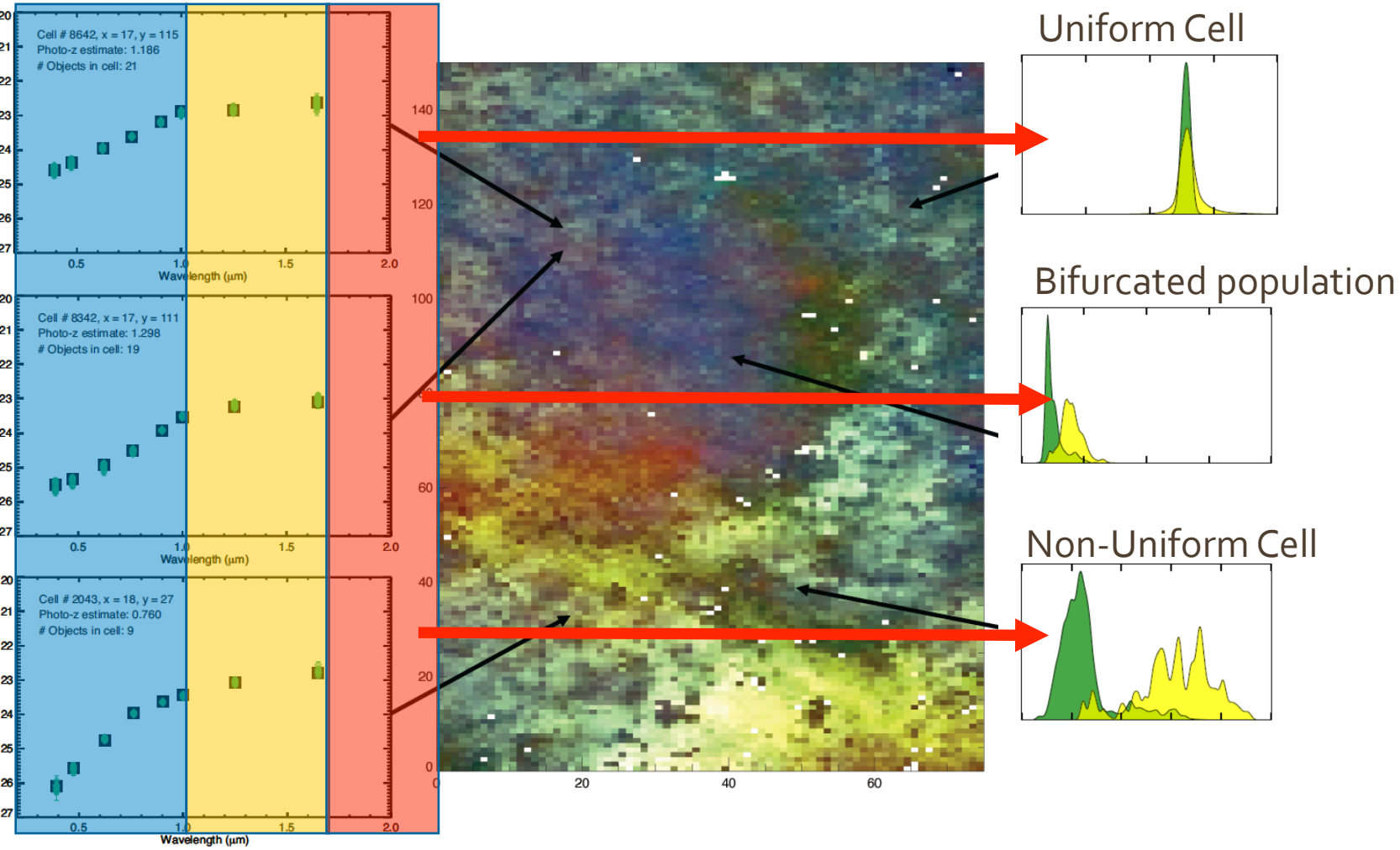






This allows us to combine knowledge in a standard model

- Generate model with one data set
- Expand it to higher dimensions with another



Conclusions

- We should be projecting our models into data space. We are currently doing the opposite which makes it difficult to understand systematics.
- By working in data space you also gain much more information because you can average similar objects.
- Working in data space is also computationally much more efficient because you sample at the native information density rather than the model density.
- Mapping data space allows one to combine statistical information from multiple data sources in a coherent way.
- Analysis could be combined into a "Standard Model" of galaxies.

The 25 deg² COSMIC Dawn Survey

laboration between the
WN Center, IAP, IPAC, and
waii centered on Euclid

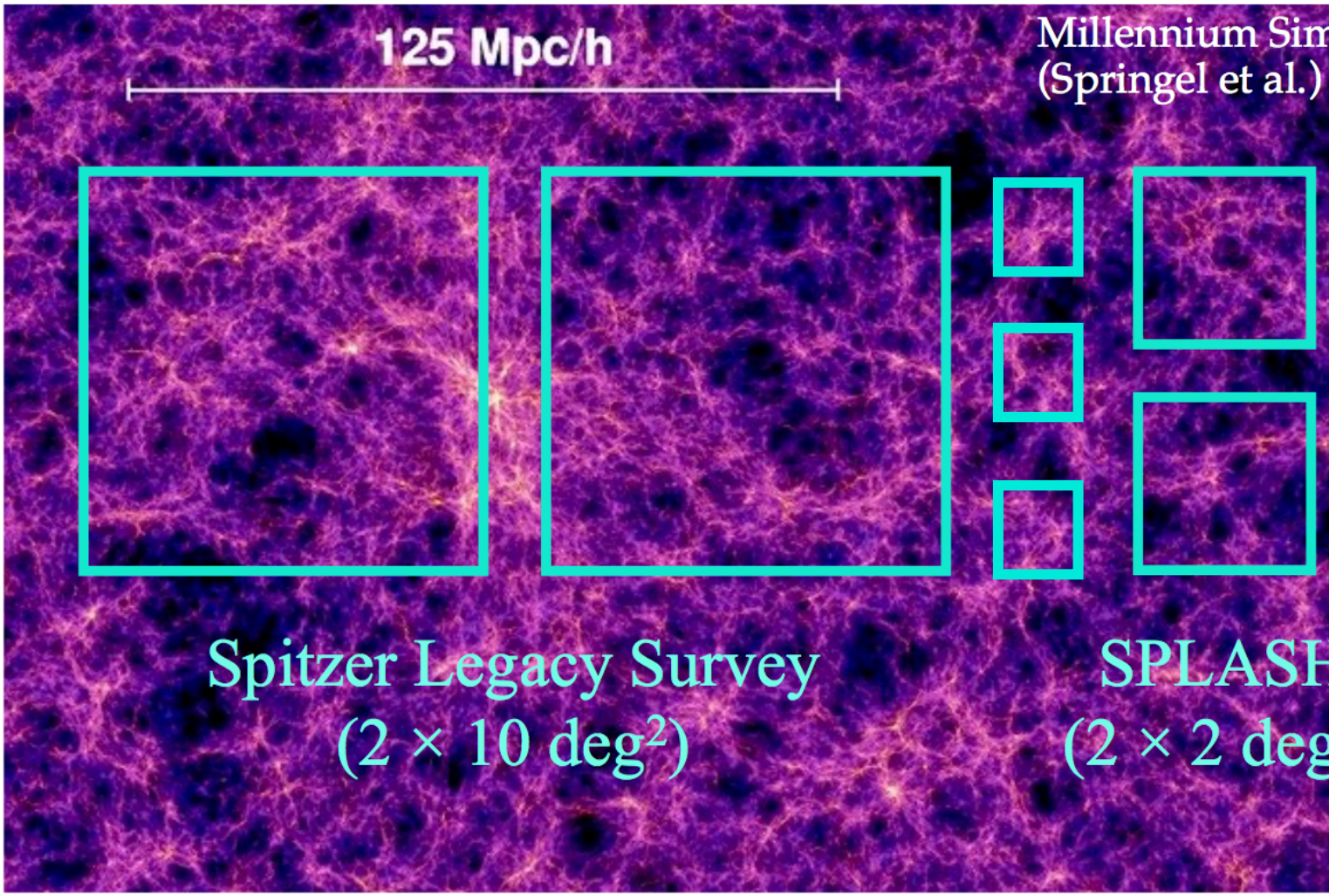
per-Suprime-Cam
w-2021)+ Spitzer (now) +
clid (2023)

rted two surveys to get
required data

PLASH
- 2 x 2 deg²
  - COSMOS, SXDS

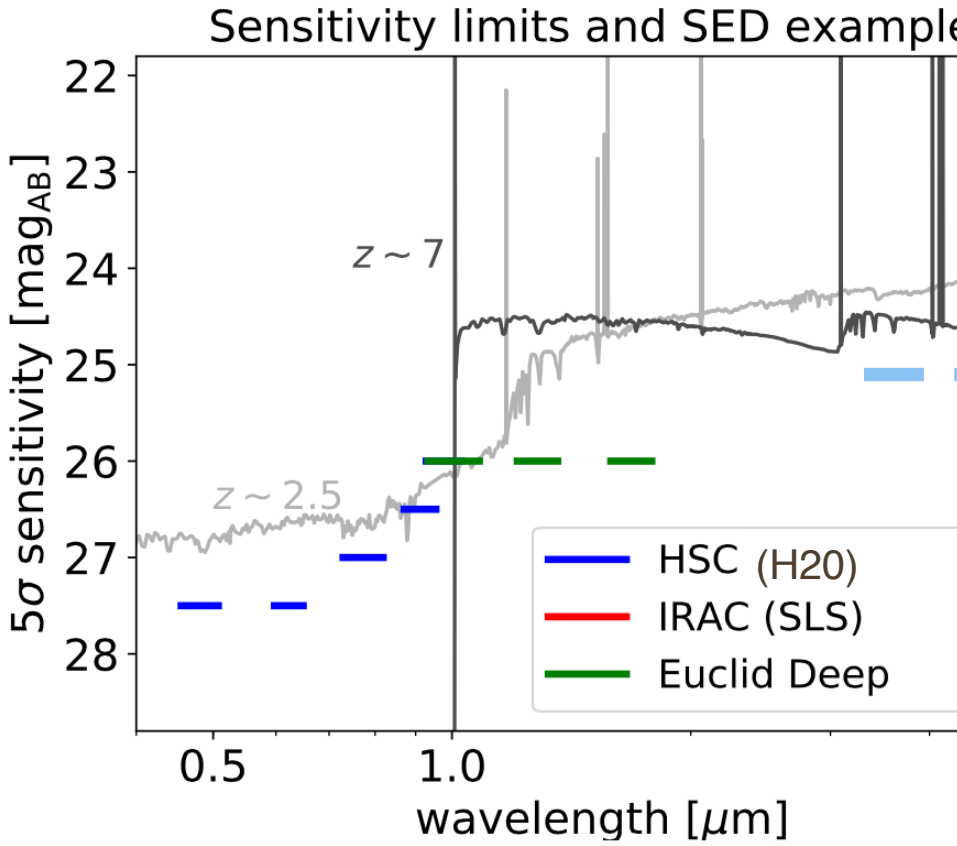
awaii-Two-O (H20)
- 2 x 10 deg²
  - NEP, CDFS

o using existing fields
- 3 x 0.5 deg²
- HDF-N, EGS, VVDS-2h



Cosmic Dawn Survey

- Centered on Spitzer + Euclid fields
  - Depth designed around best practices in current HST Fields (see Wed talks)
- >12,000h (~1.3 years) of Spitzer Time
  - CDFS, NEP, 2h per pixel 20 deg^2 = 6,000h
  - SPLASH 7h per pixel 3.6 deg^2 = 2,500h
  - Archival data in HDFN, EGS, others >3,500h
- Hyper-Suprime Cam getting optical imaging in g,r,I,z,y
  - SPLASH = HSC SSP Ultra-Deep
  - CDFS, NEP being done by H20 program in Hawaii
    - 30 nights HSC
    - 10 nights Keck
    - Hawaii Intensive Program
  - Other fields in archive
- We have been waiting for HSC to get deep enough!

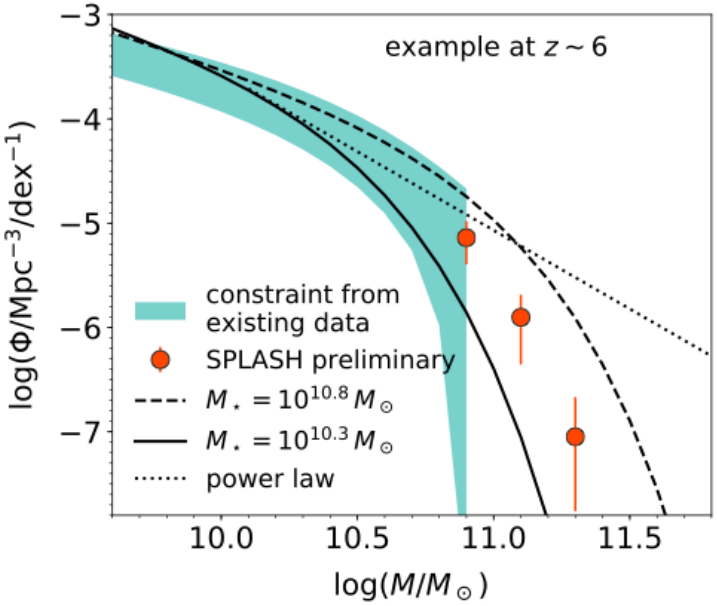


20 sq deg (H20 = PI D. Sanders; SLS = PI

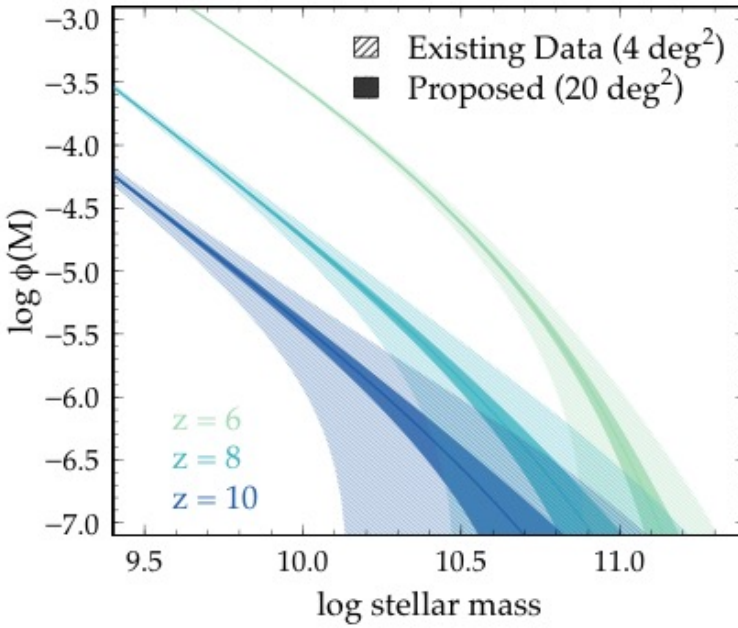
H20: High-Mass Galaxies at z>5

- When and how do the first massive galaxies form?
- Existing CANDELS data provides no constraint above M* ($10^{10.4}$) at z>5
- SPLASH + SLS will provide meaningful constraints to z~10
  - 3.6 square degrees to 25.5 AB
  - 20 square degrees to 24.6 AB

Current →



H20 →



Davidzon et al. in prep

SPLASH: Constraints Or Halo Mass

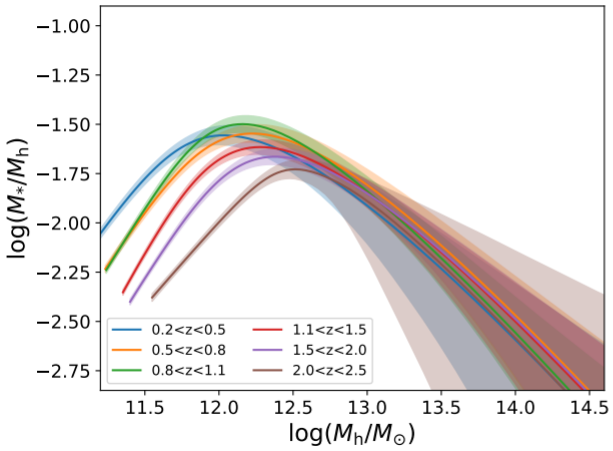
nstraints on the halo-mass to stellar mass io up to z~6 with HSC + Spitzer data

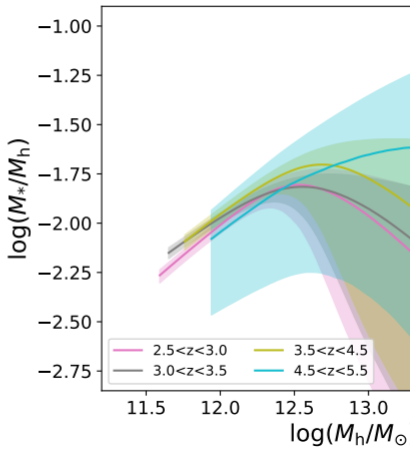
ing two techniques
Abundance matching and clustering

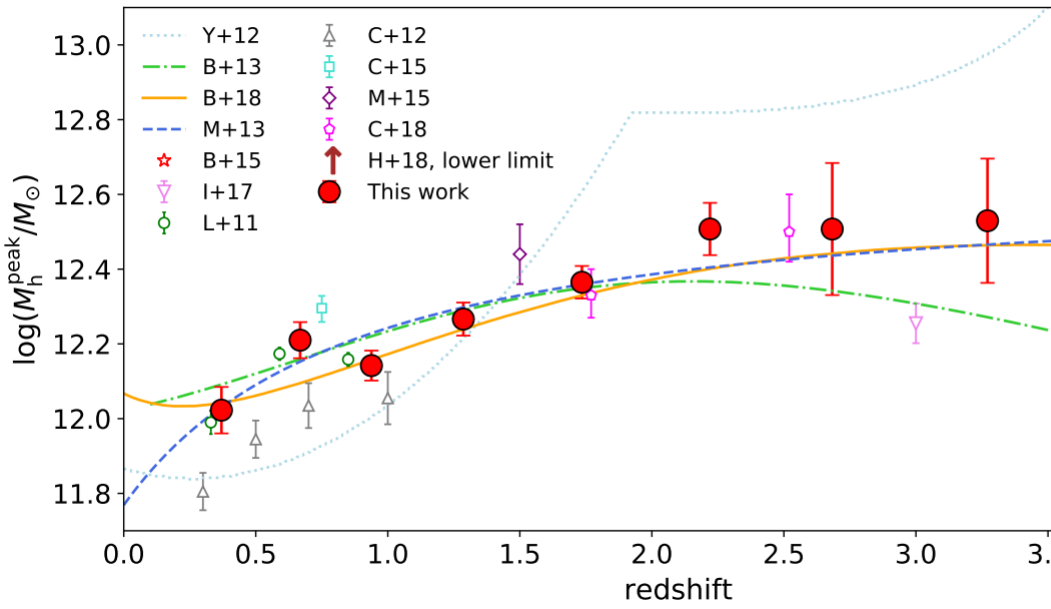
undance matching results published in grand et al. 2018

e rise in peak efficiency out to z~4

SMOS + SXDS analysis underway



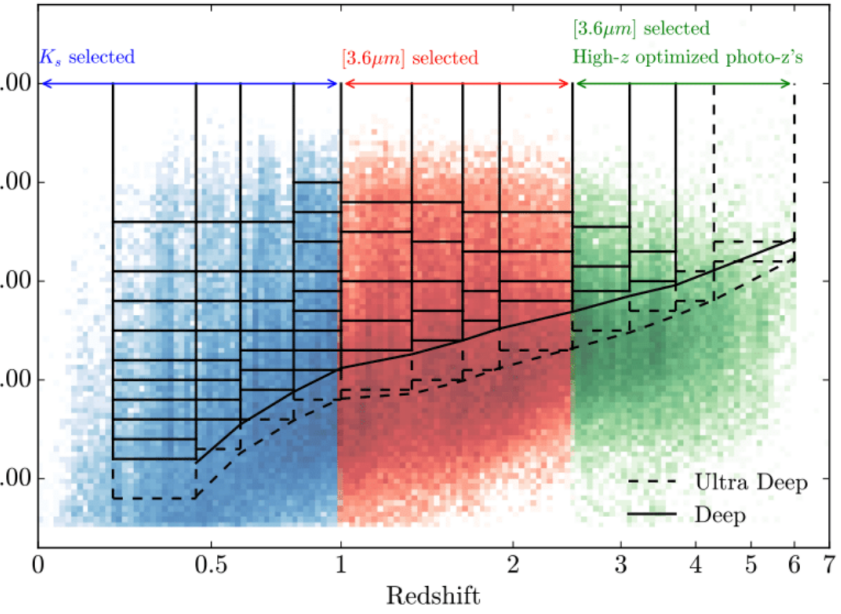




Legrand et al. 2018

SPLASH: Constraints On Halo Mass

- Clustering results in progress
- Similar to AM but start to see drop at z>4
- COSMOS + SXDS analysis underway



on/Davidzon et al. in prep