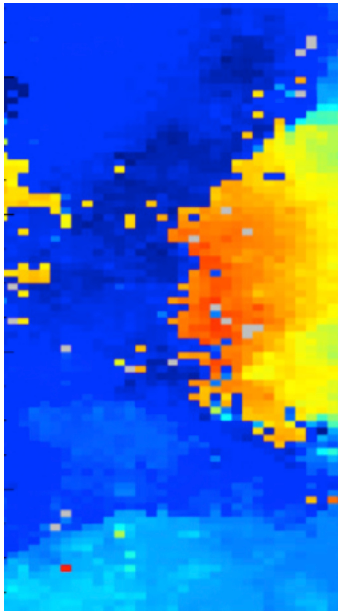
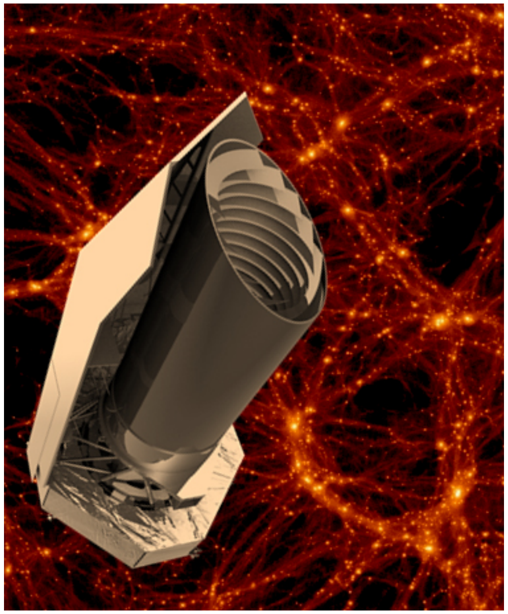
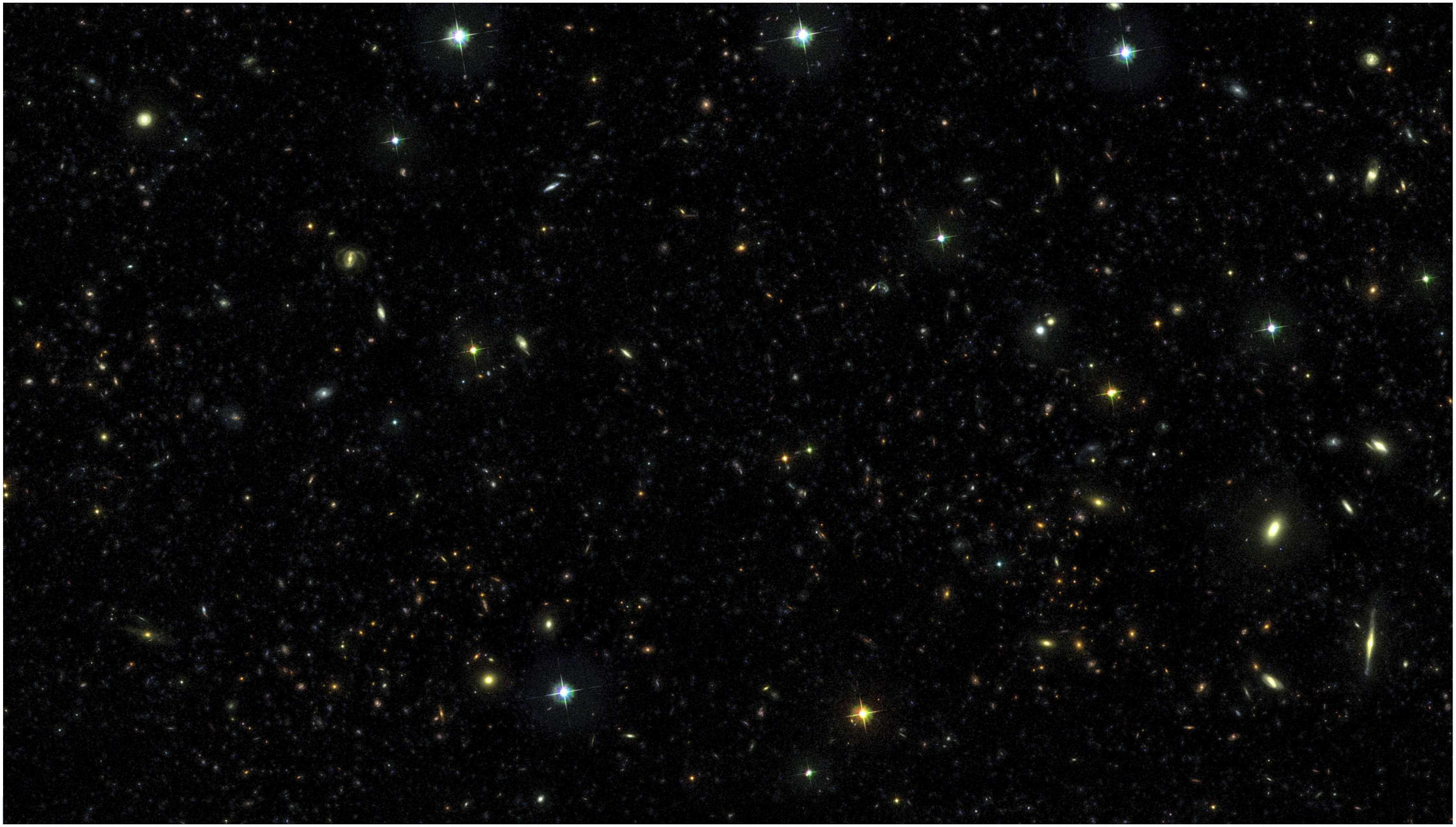
Developing a Standard Model of Galaxies for Cosmology

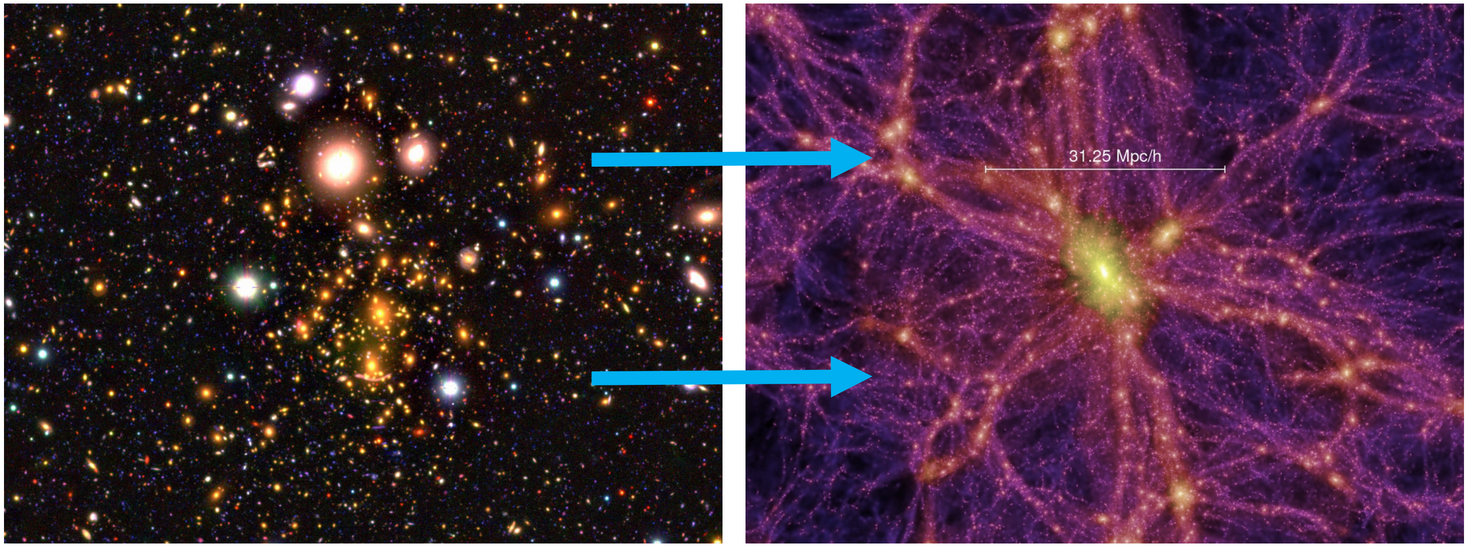
Peter L. Capak
IPAC – Caltech
Associate of the Cosmic Dawn Center







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

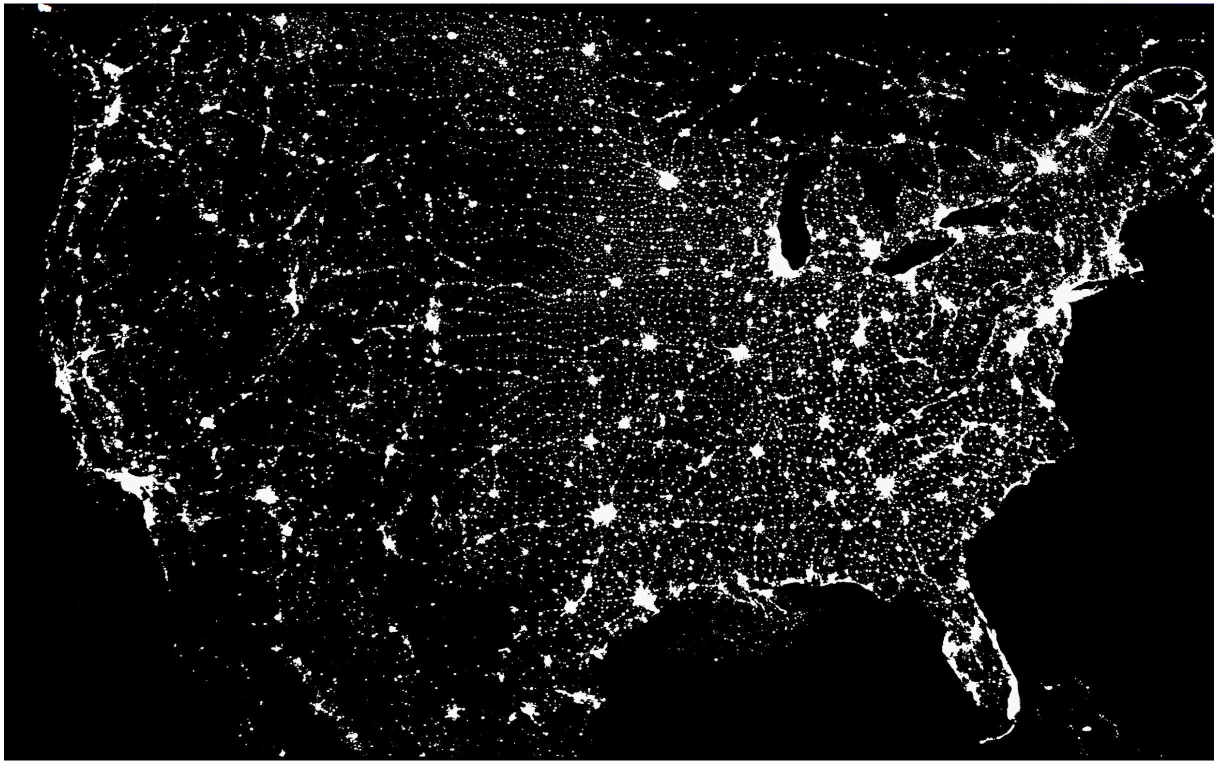


Capak et al. 2004

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

- We can treat the galaxy survey problem like the US trying to make a census from space.
- Galaxies are a bias proxy for matter just as light is a bias proxy for population.
- We don't need to understand why its bias, but need to know how its bias.

~2960



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

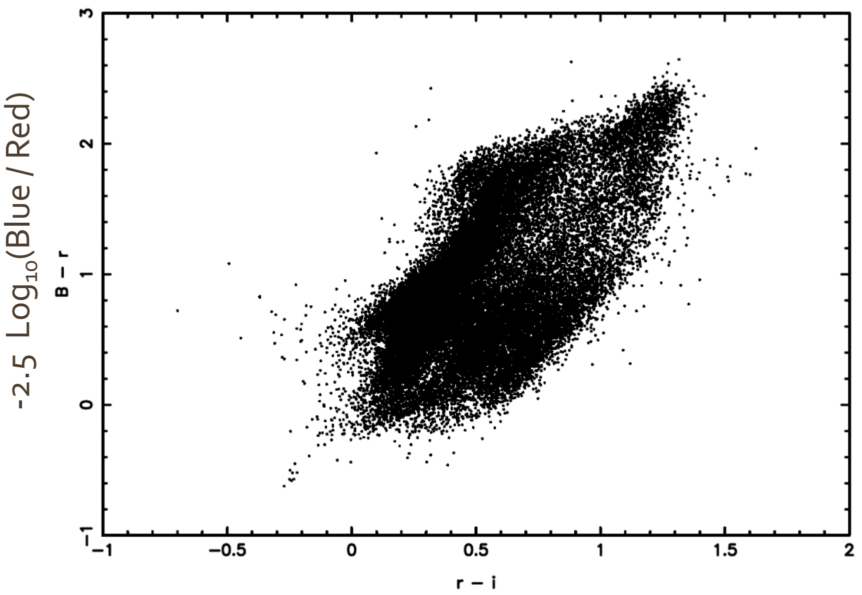
- We don't need a precise model of people to understand how cities built up.
- Modern cosmology needs to measure galaxy evolution but doesn't need to understand it.
- We just need to characterize their behavior accurately enough.

~2000

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.

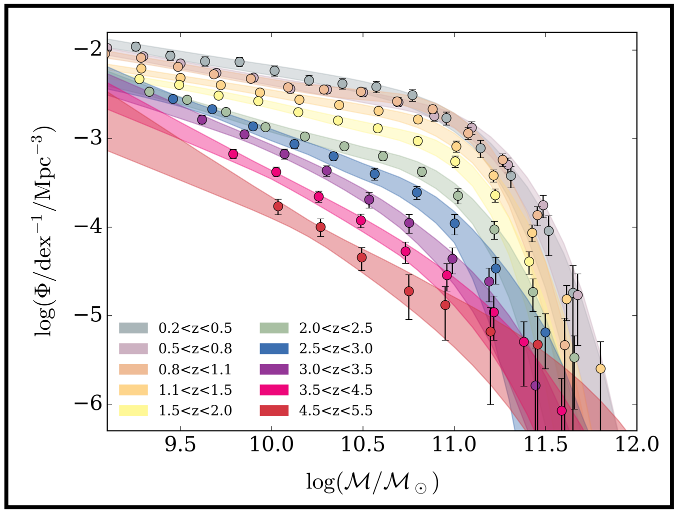
Example Of Galaxy Colors



-2.5 $\log_{10}$(Red / 0.75μm)

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…….



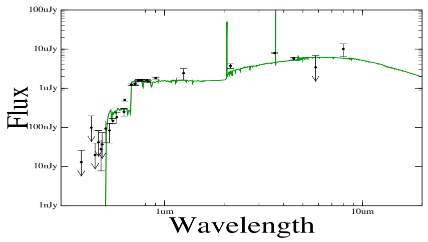
Davidzon+17:
stellar mass function

Lets re-think how we do galaxy evolution studies.

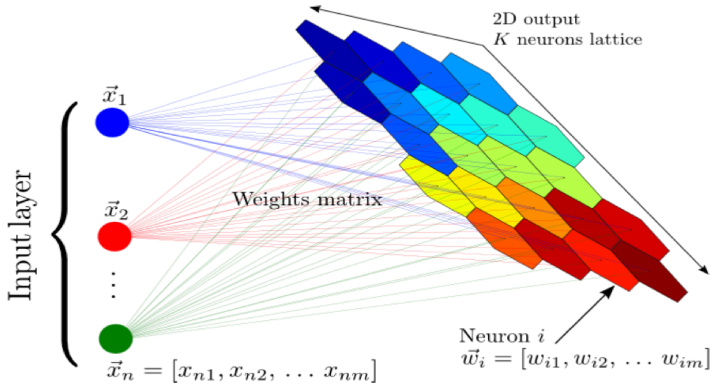
- For this analysis I will only consider photometric surveys for clarity.
- What are the steps to creating a mass function?
  - Estimate redshifts to galaxies = Use the color of a galaxy and its flux to estimate the redshift
  - Estimate masses for the galaxies = Use the color of a galaxy to estimate 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 big-data techniques

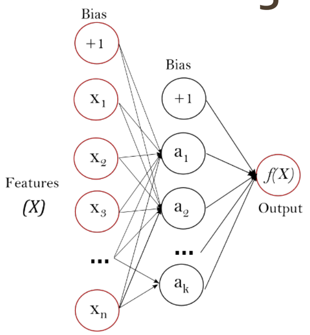
Analytic Models

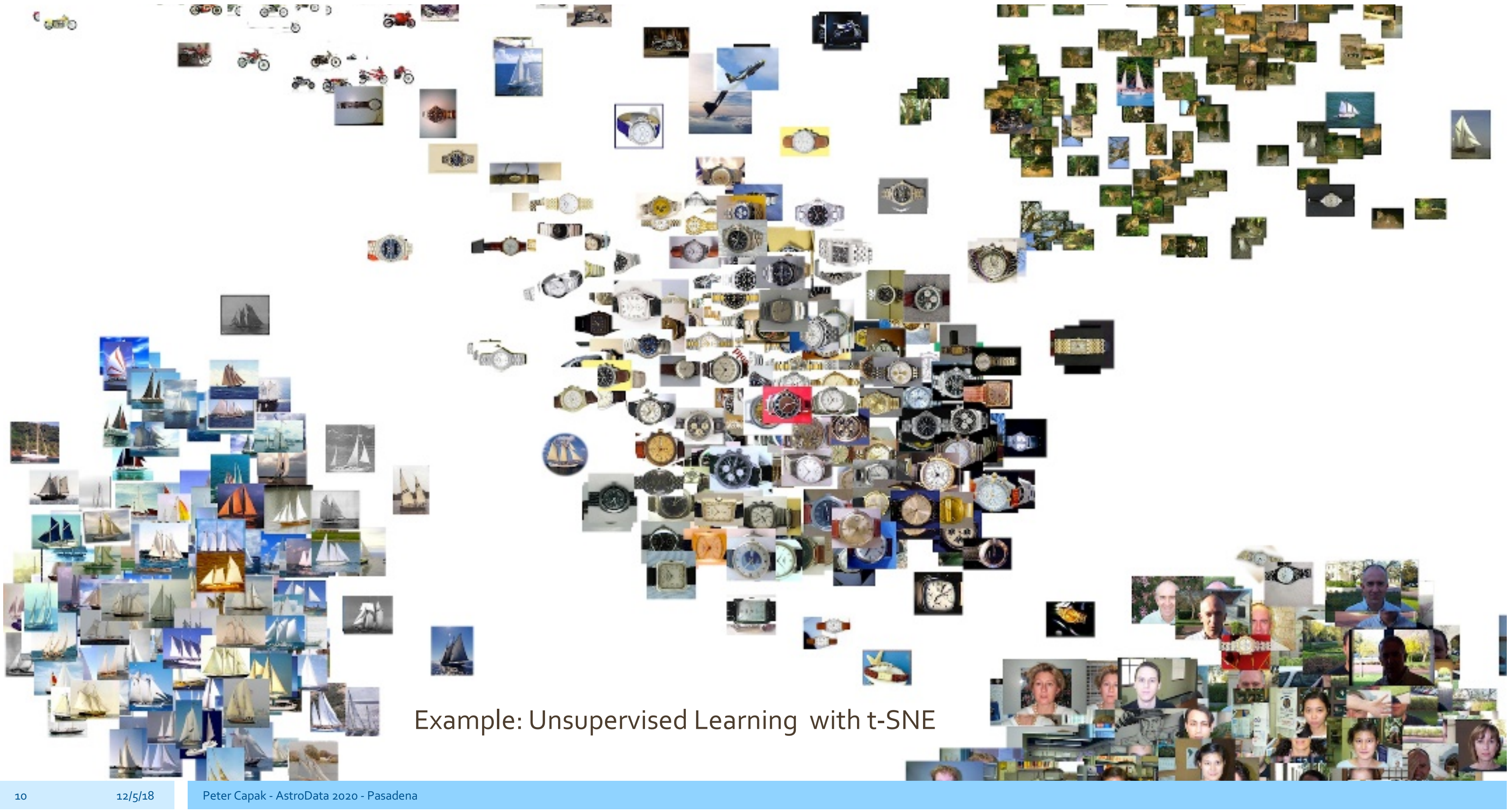


Unsupervised Learning


2D output
K neurons lattice
$\vec{x}_1$
$\vec{x}_2$
Input layer
Weights matrix
Neuron i
$\vec{w}_i = [w_{i1}, w_{i2}, \dots w_{im}]$
$\vec{x}_n = [x_{n1}, x_{n2}, \dots x_{nm}]$

Supervised Learning

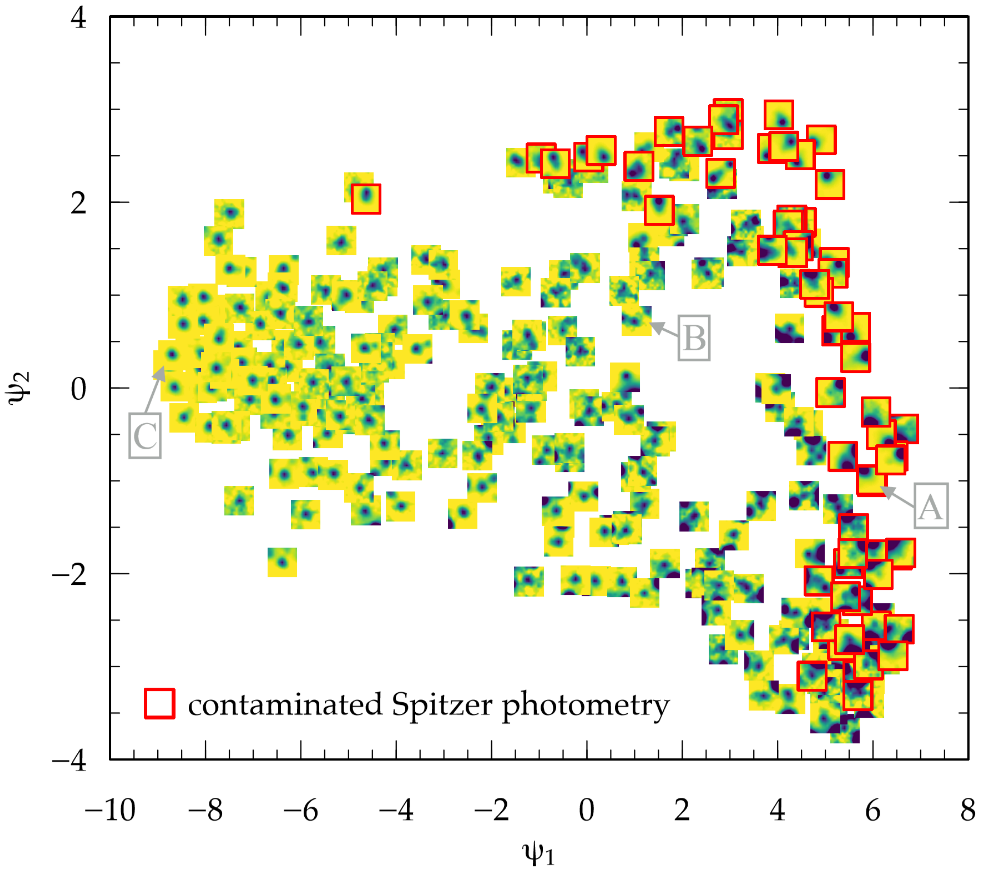




Example: Unsupervised Learning with t-SNE

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

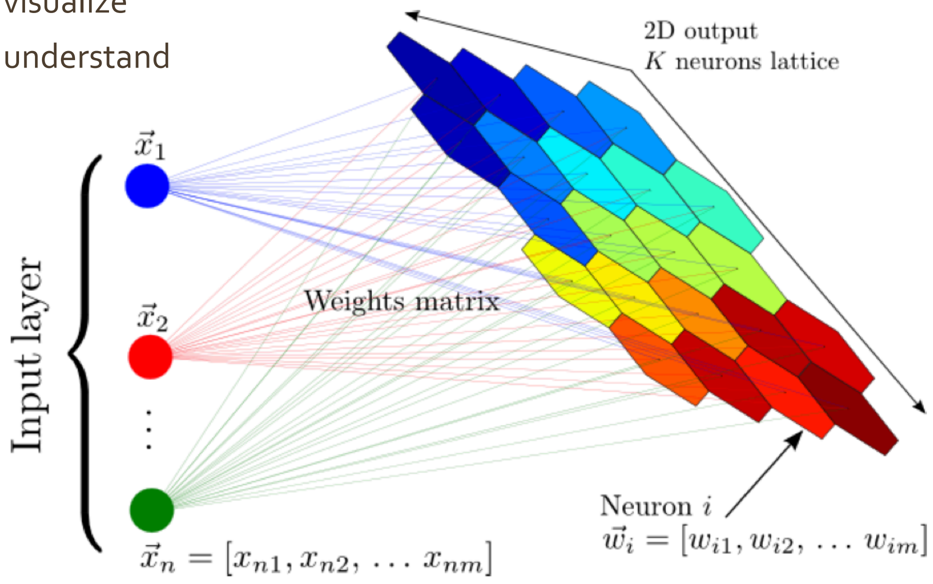
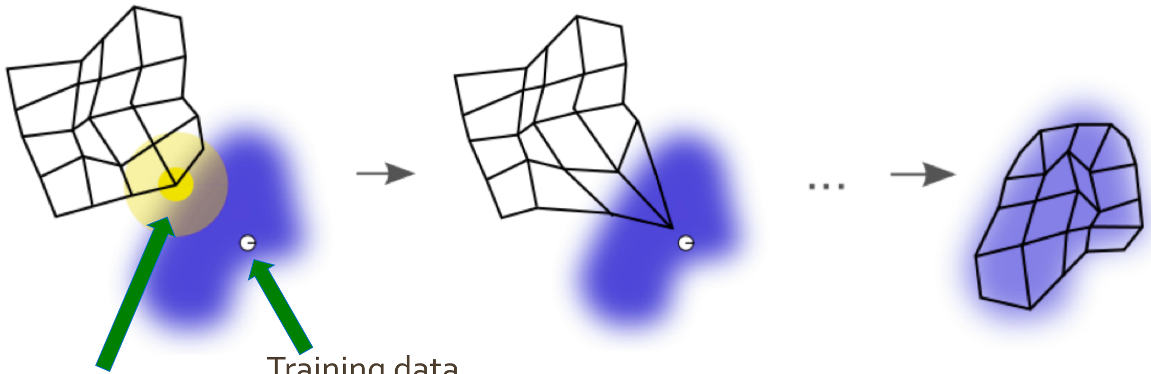


Illustration of the SOM (From Carrasco Kind & Brunner 2014)

Example: Characterizing galaxy photometry with a Self Organizing Map.

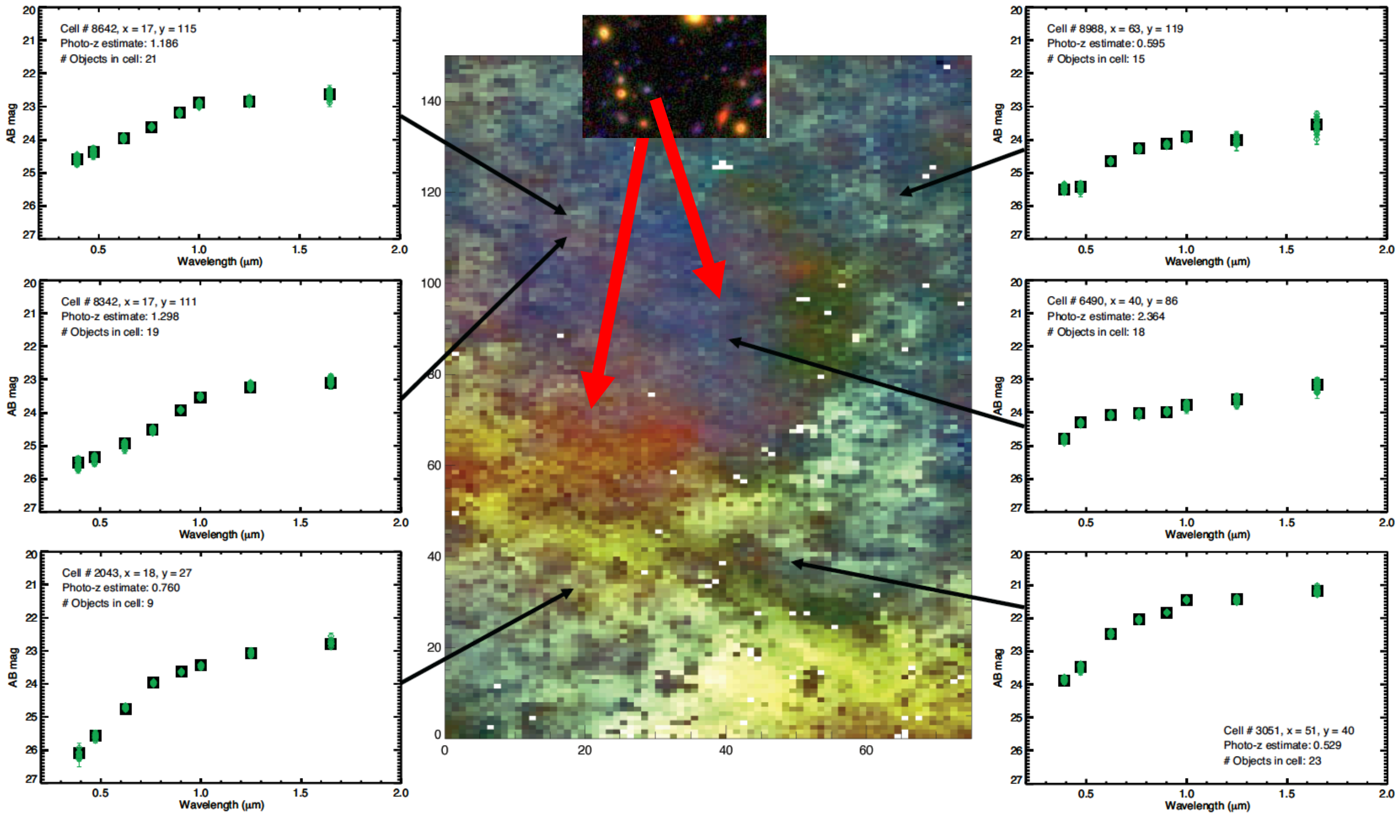


Training data

Best-matching cell in current SOM

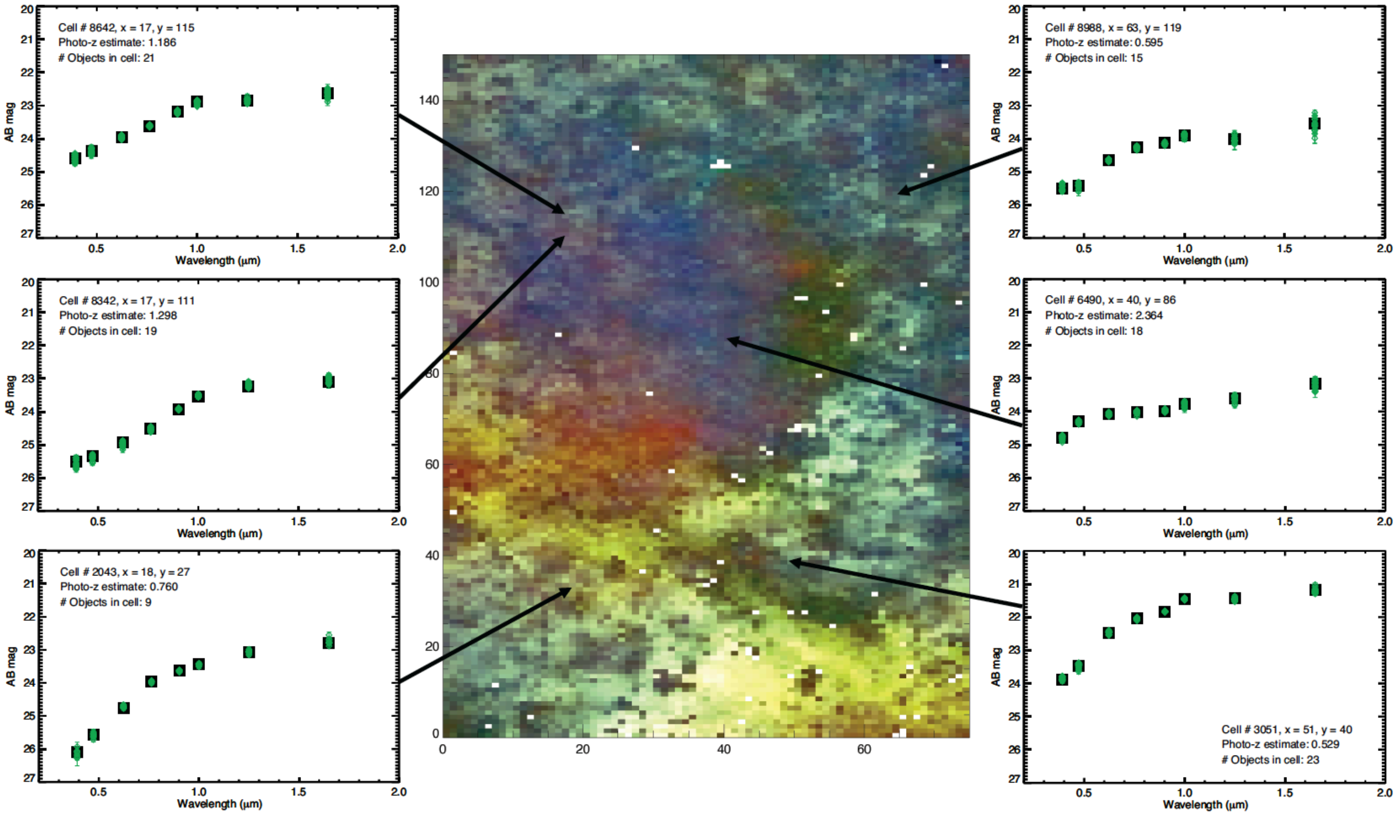
1. Initialized map is presented with training data, i.e. the colors of one galaxy from the overall sample.
2. Map moves towards training data, with the closest cells being most affected.
3. Process repeats many times with samples drawn from training set until the map approximates the data distribution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

Example: Characterizing galaxy photometry with a Self Organizing Map.



Masters, Capak et al. 2015

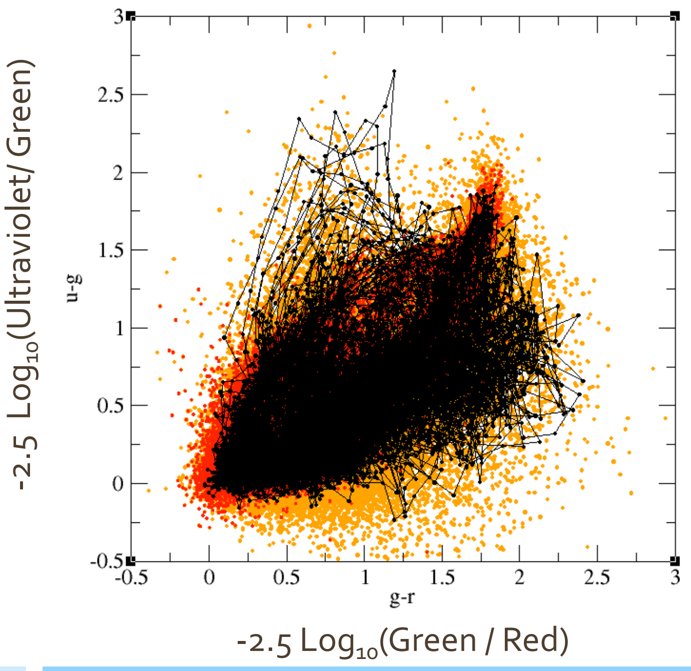
Example: Characterizing galaxy photometry with a Self Organizing Map.

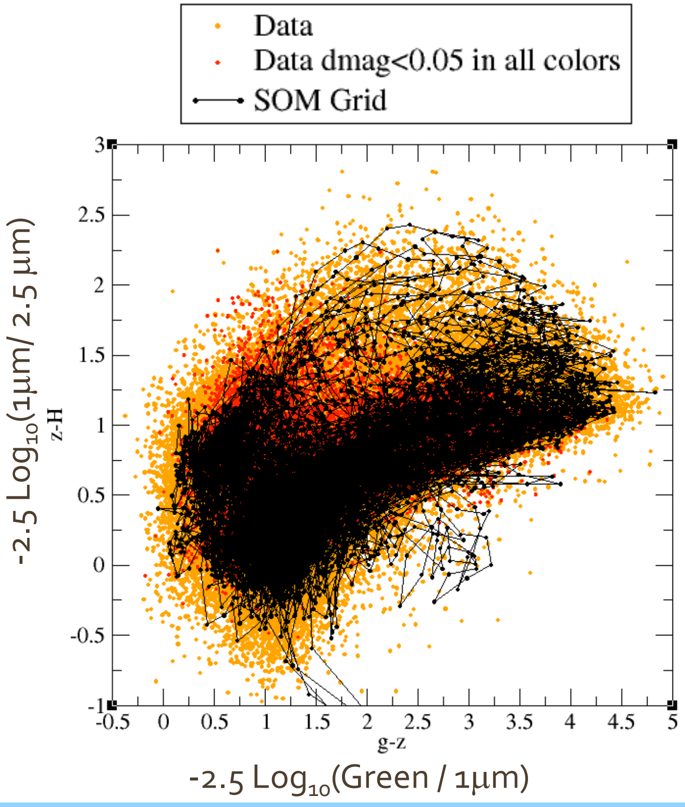


Masters, Capak et al. 2015

Example: Characterizing galaxy photometry with a Self Organizing Map.

The SOM grid in color-color plots





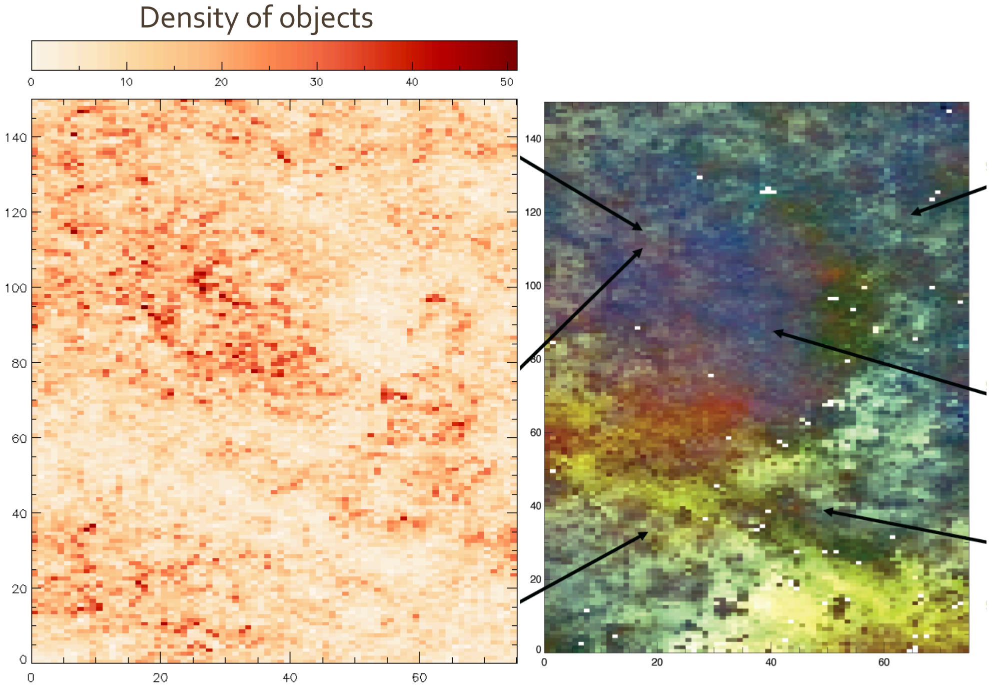
Example: Characterizing galaxy photometry with a Self Organizing Map.

- 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.

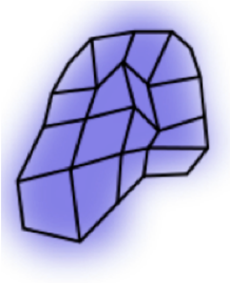
- We can now map our knowledge of the galaxy population onto that probability field.

- We could use an analytic model or a data model.

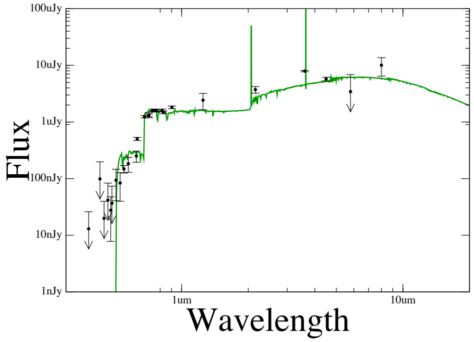


Example: Characterizing galaxy photometry with a Self Organizing Map.

- 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



+



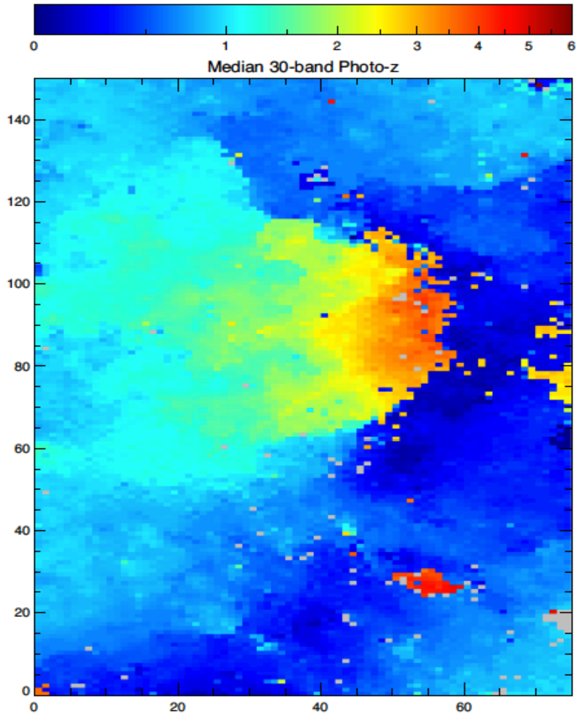
Photometric Redshifts can be calculated with a self organized map. Importantly, self organized maps tell you why certain objects are degenerate.

- Here we have colored the SOM with the median photometric redshift.

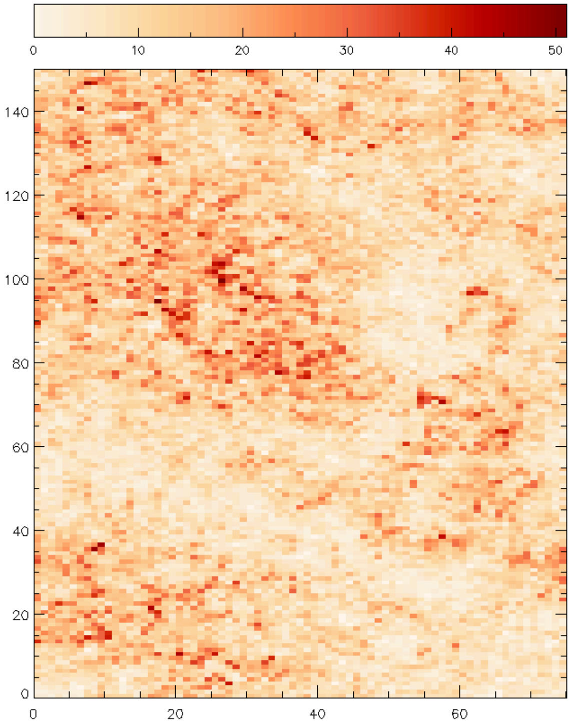
- Notice redshift varies smoothly over most of the data space. This is why photometric redshifts work.

- Notice the caustics, this is why there are degeneracies in photometric redshifts.

Mean Redshift



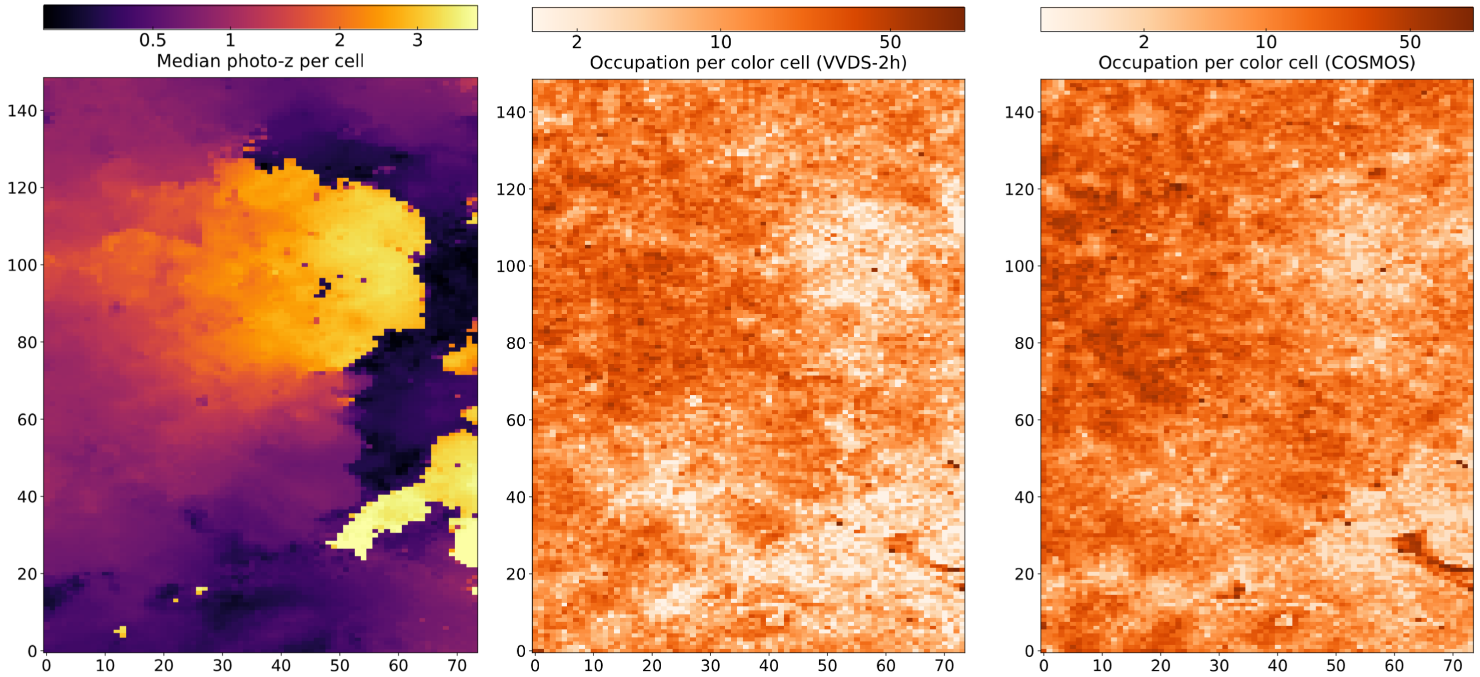
Density of objects



Masters et al. 2015, 2017

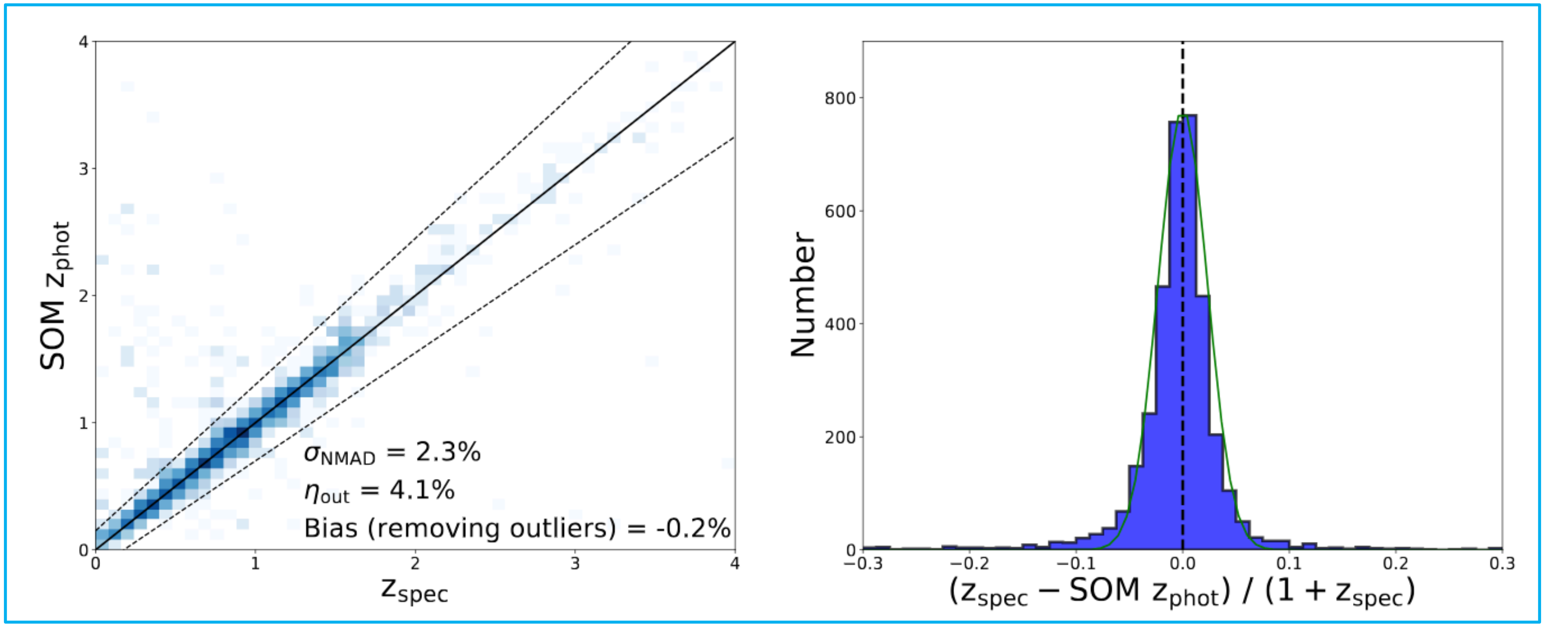
The data density field contains fundamental information about how galaxies evolve.

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



Masters et al. in prep

Redshifts estimates from color mapping are more accurate than from SED fitting



This analysis does NOT use spec-z training!

→ Method achieving unbiased performance

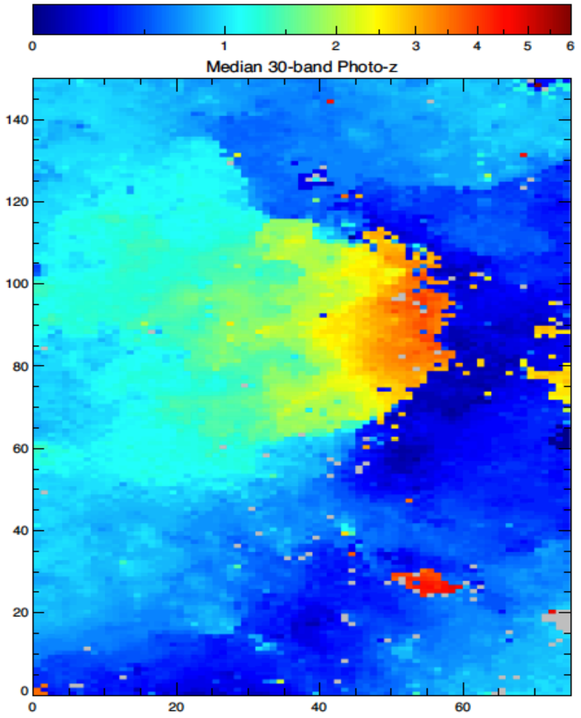
→ Outlier fraction 4.1%, scatter 2.3%, bias of -0.2%

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

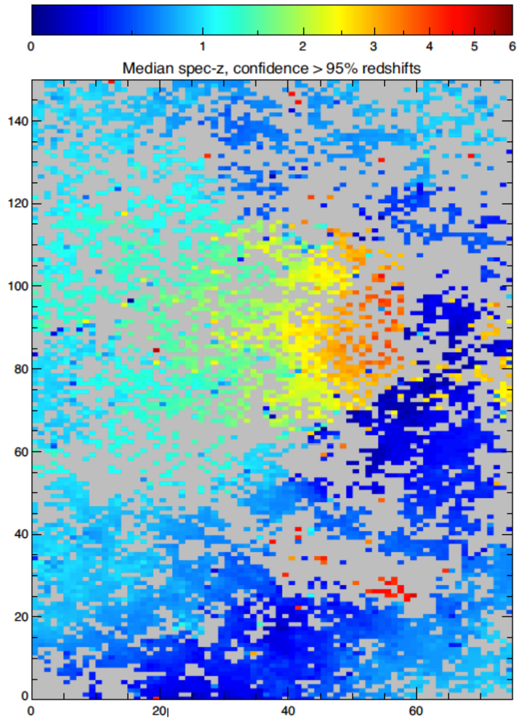
- We can compare our photo-z model to spectroscopic data.

- The agreement is good, but we do not sample a large fraction of the color space occupied by galaxies.

Photometric Redshift

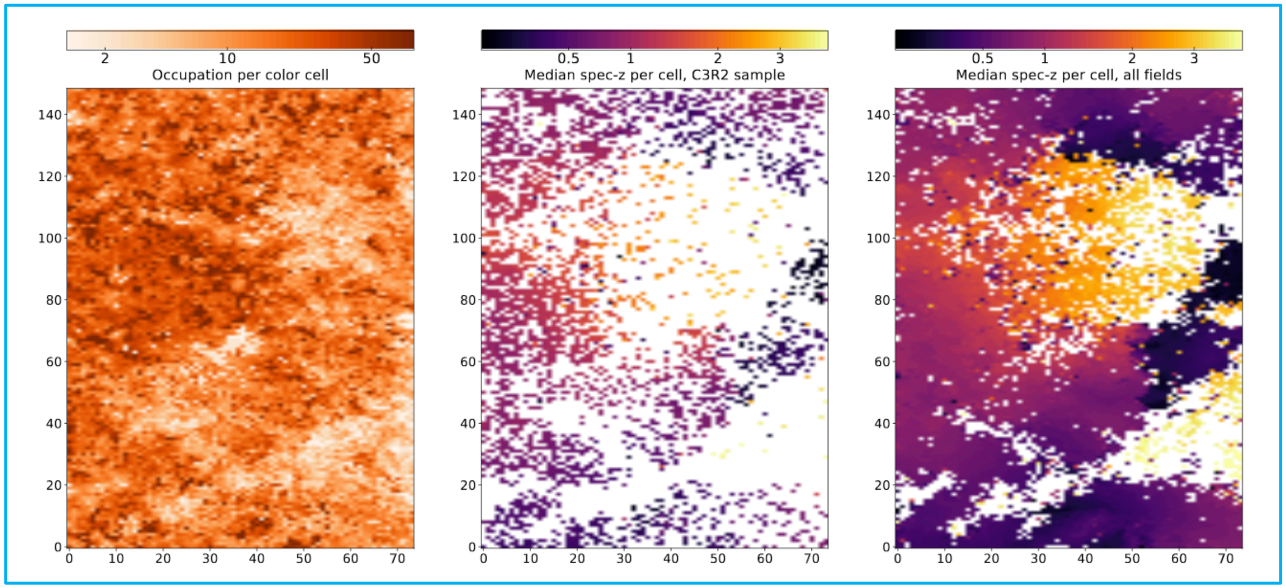


Spectroscopic Redshift



Masters et al. 2015, 2017

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

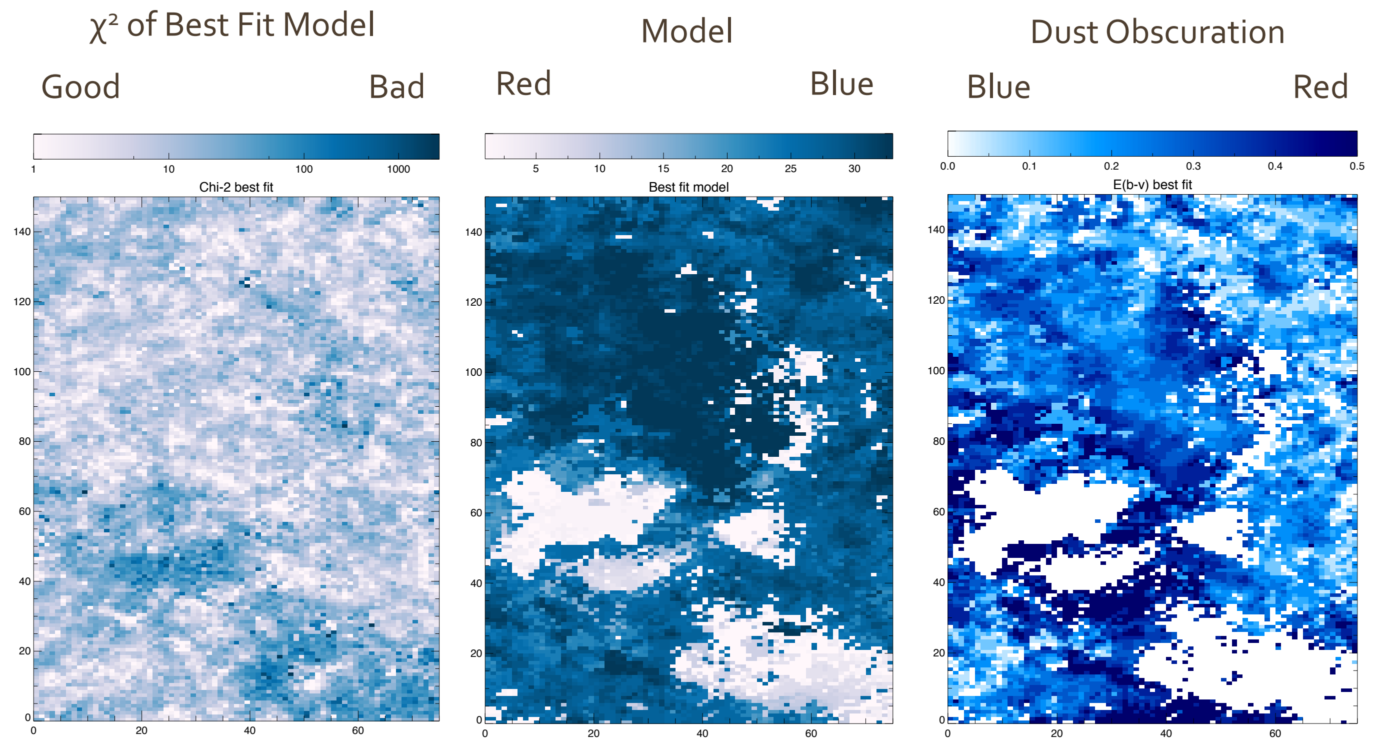


- To date C3R2-Keck covers >35% of the color space, disjoint from what was previously explored
- C3R2 has covered >75% of cells with >85% of galaxies in cell with at least 1 specz, many cells with >>1 specz
- Uncovered cells correspond to less common sources, targets of future follow-up

Masters et al. in prep

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

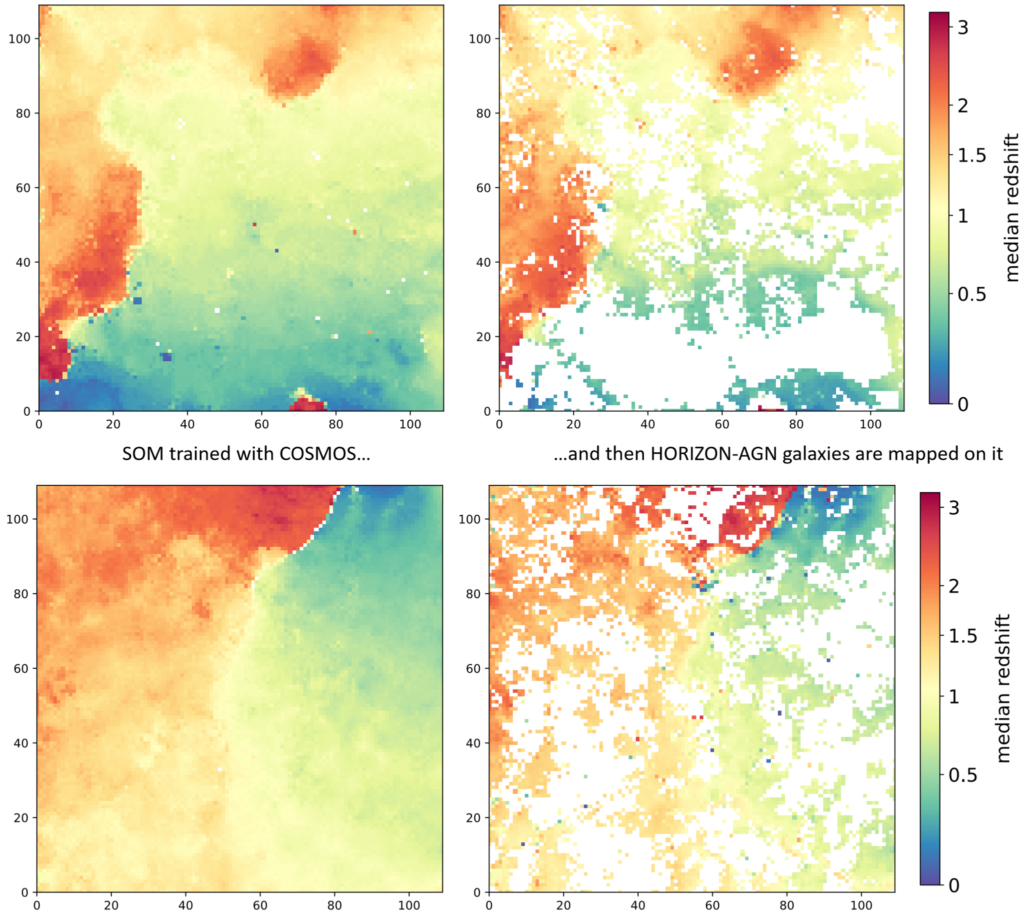
- Here we are showing the best fit model at each point on the SOM.
- It is clear there are regions where the fit is much worse.
- The SED model is not good in these regions and the parameters will be unstable.



Capak et al. in prep

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

- Do computational models match the real universe?
- Mostly, but there are large differences
- Horizon AGN:
  - Does not predict some populations of low-z galaxies observed in COSMOS.
  - Predicts z>1 galaxies that are not observed to exist.



SOM trained with COSMOS...

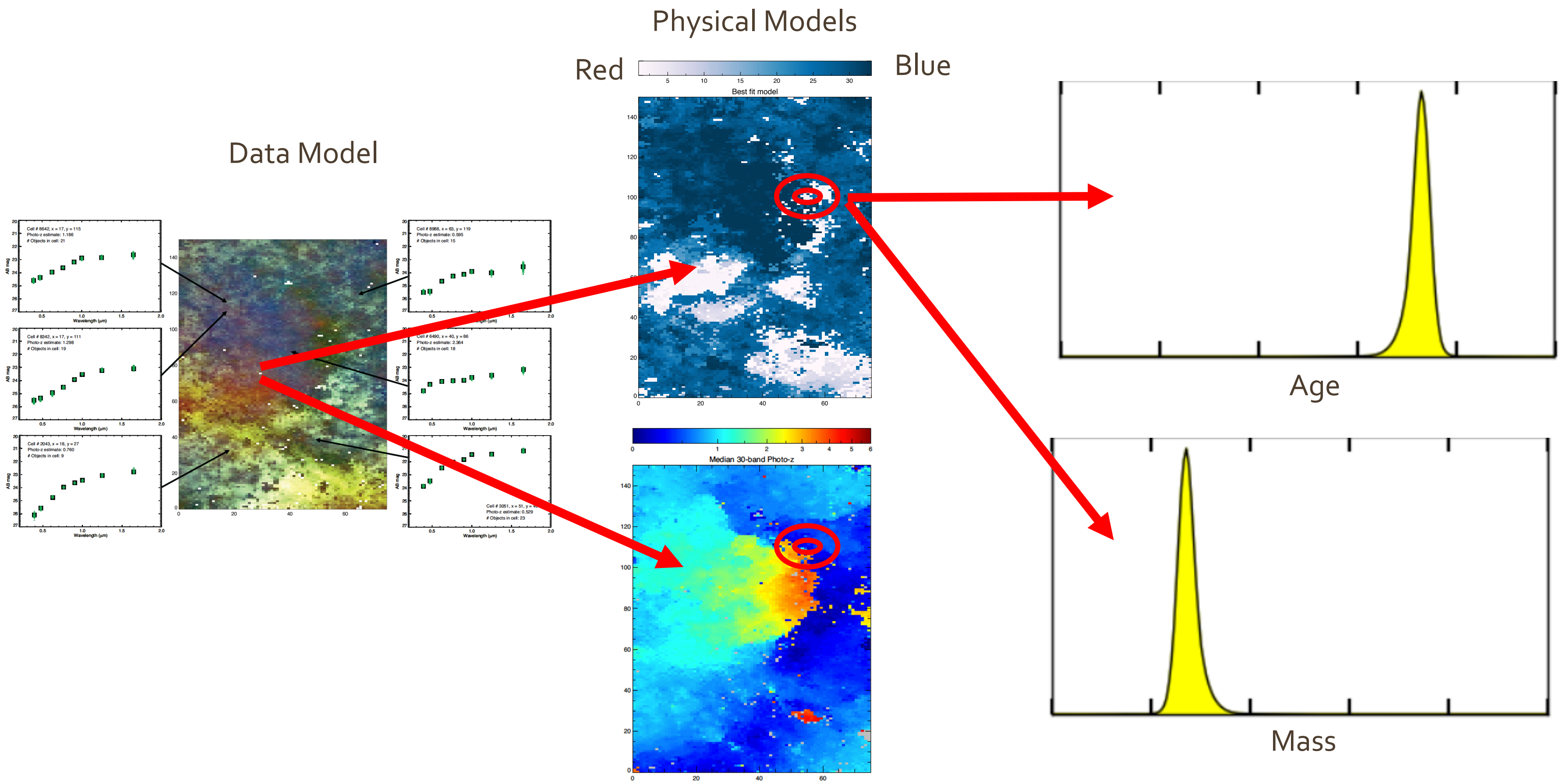
...and then HORIZON-AGN galaxies are mapped on it

SOM trained with HorizonAGN...

...and then COSMOS galaxies are mapped on it

Davidzon et al. in prep

Models can be gridded in the data space to infer the physical parameters of objects.



Physical Models

Red

Blue

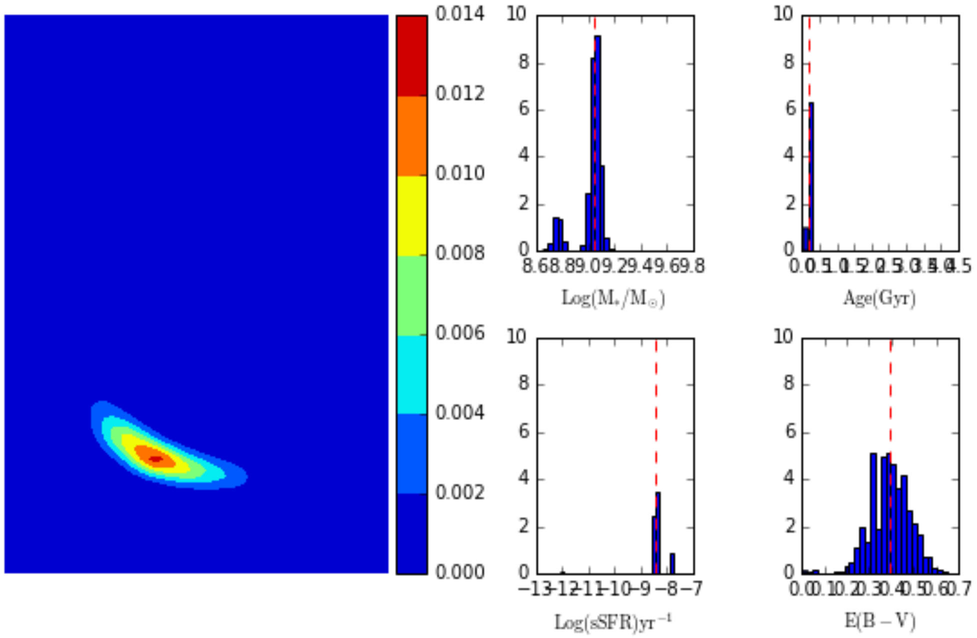
Data Model

Age

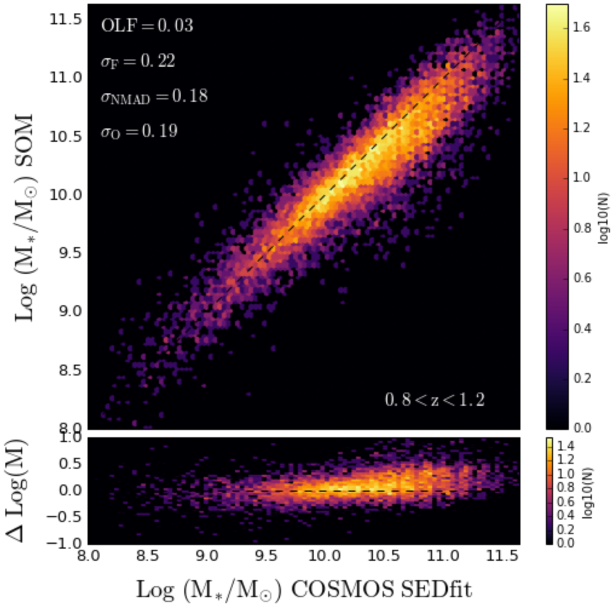
Mass

In our early tests we can recover physical parameters almost as well as full model fitting, but orders of magnitude faster.

Integrate Photometric Error Against SOM

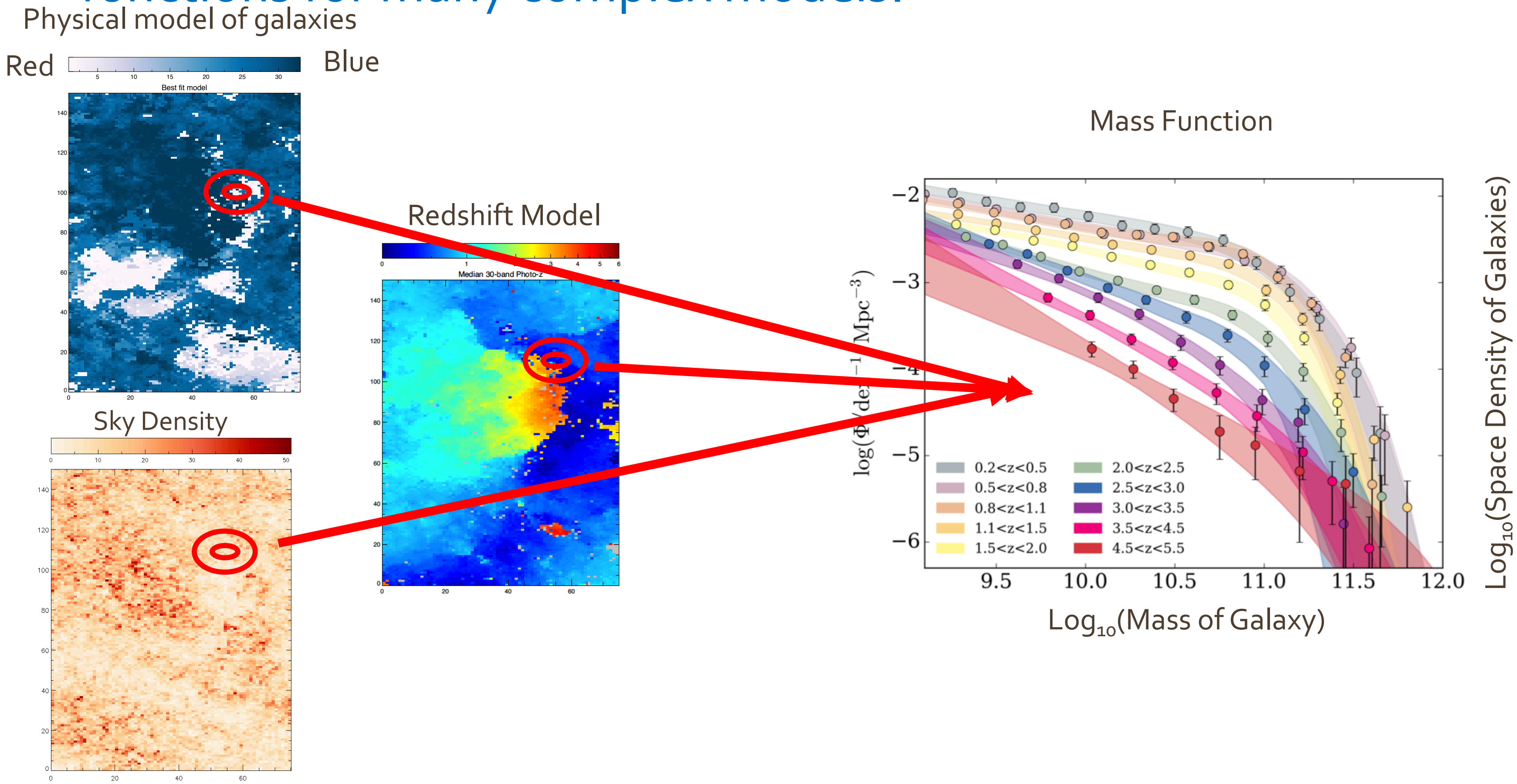


Comparison to SED Fits



Hemmati et al. in prep

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



Physical model of galaxies

Red Blue

Redshift Model

Sky Density

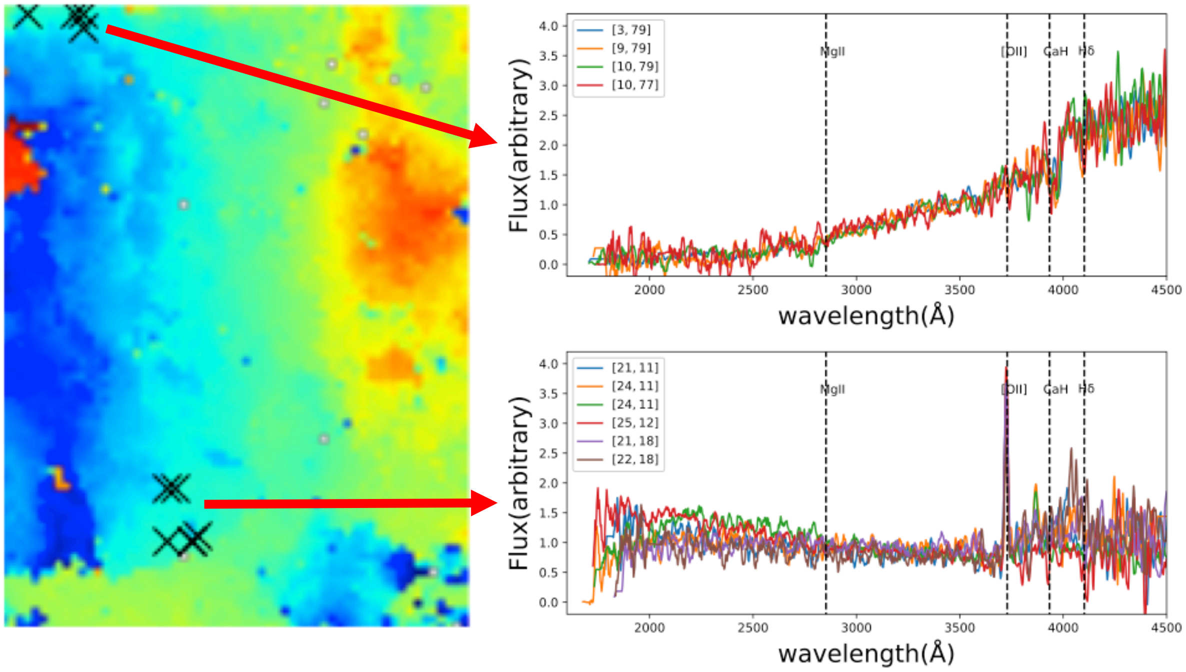
Mass Function

Log$_{10}$(Mass of Galaxy)

Log$_{10}$(Space Density of Galaxies)

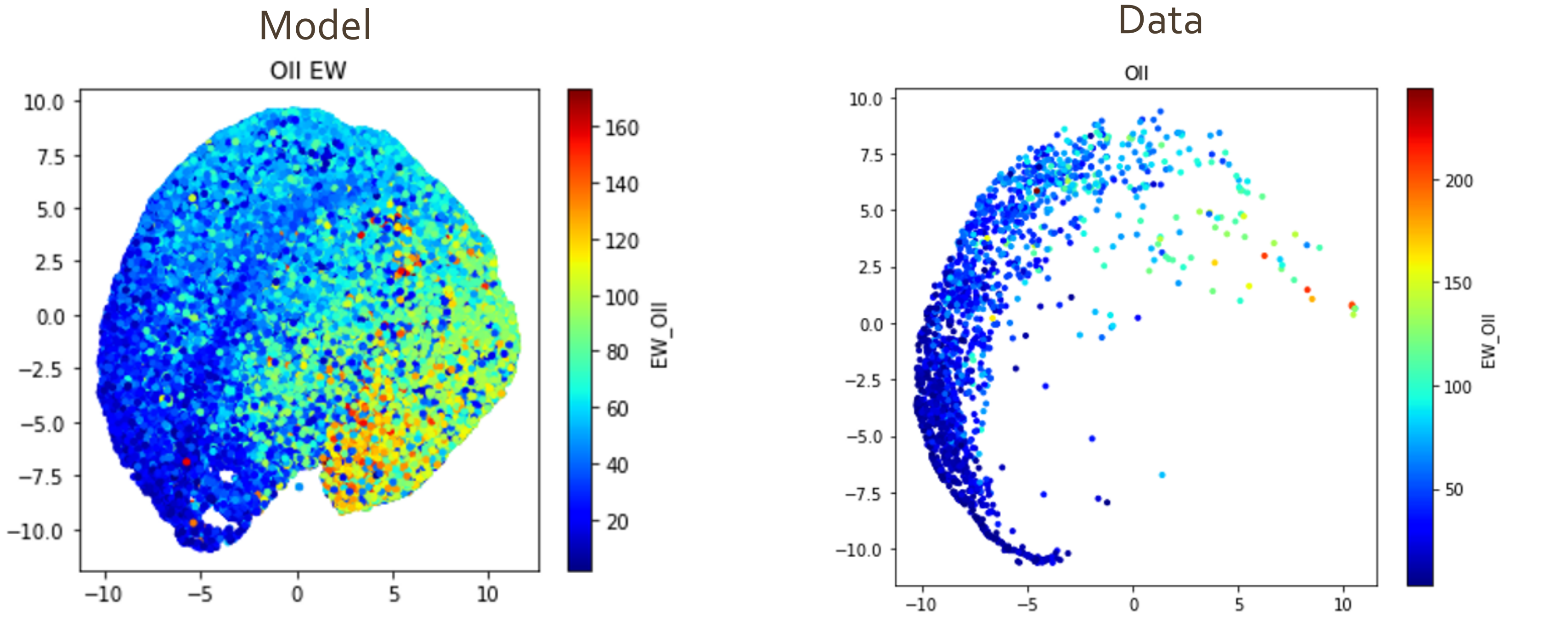
We can also map the detailed properties of galaxies across data space.

- The color is strongly correlated with the high-resolution spectra
- Typically sensitive to 100Å equivalent width variations



Hemmati et al. 2019

OII Equivalent width can be predicted from photometry.



Model

Data

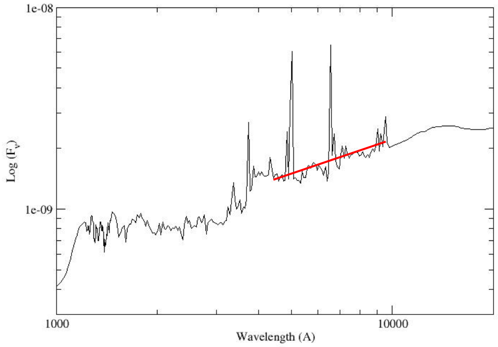
We can measure other emission lines too

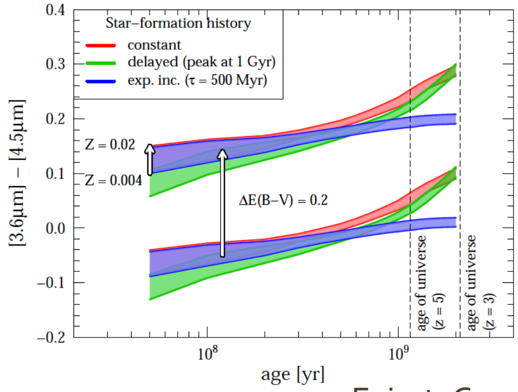
- Current emission line estimates are very model dependent

- However, they can be measured statistically from the photometry

- Just need to choose estimator carefully to account for galaxy physics

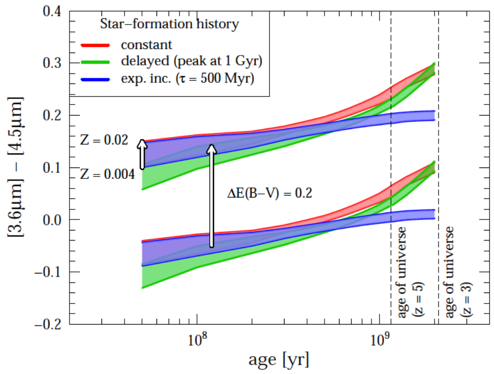
- Started with a case study at (z>5) because there is no other way to measure lines there

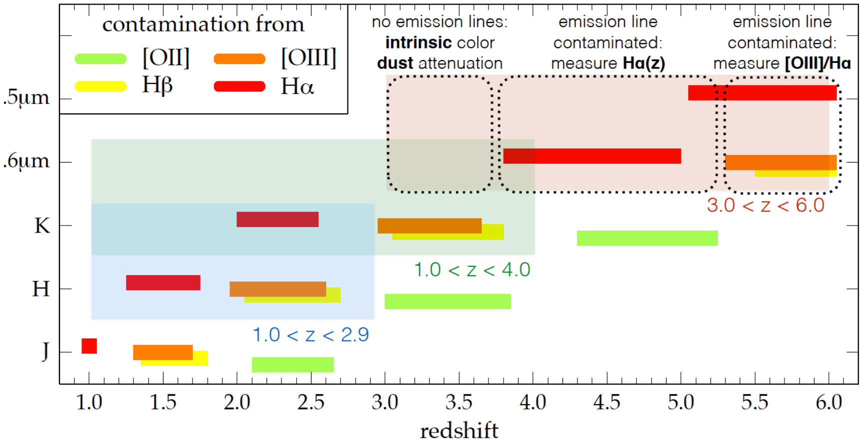




Faisst, Capak et al. 2016a

We can estimate rest-frame optical line emission for cosmology planning and JWST targeting.



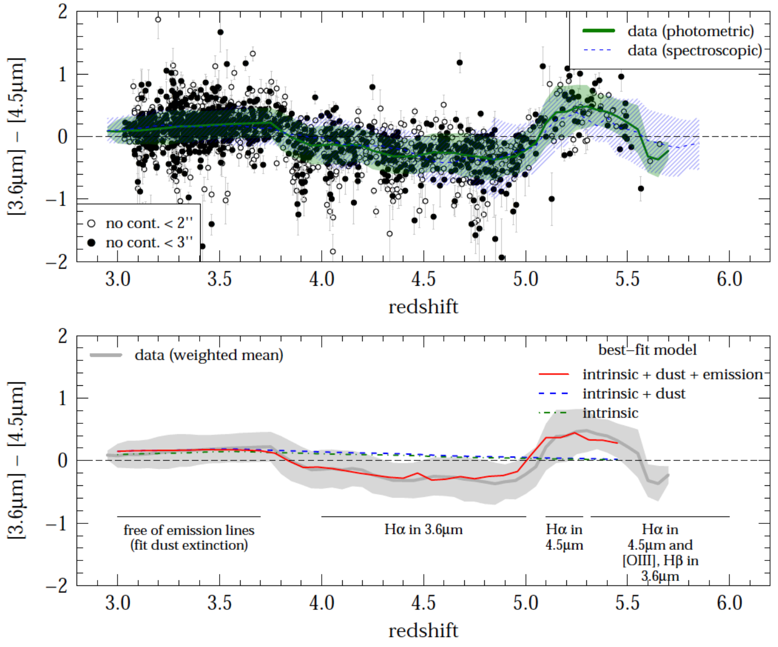


- Can measure Hα EW (and other lines) from photometry if redshift is know
- Redshift from color manifold (SOM)
- Create a forward model of the galaxy population including lines in colors

Faisst, Capak et al. 2016a

We can estimate rest-frame optical line emission for cosmology planning and JWST targeting.

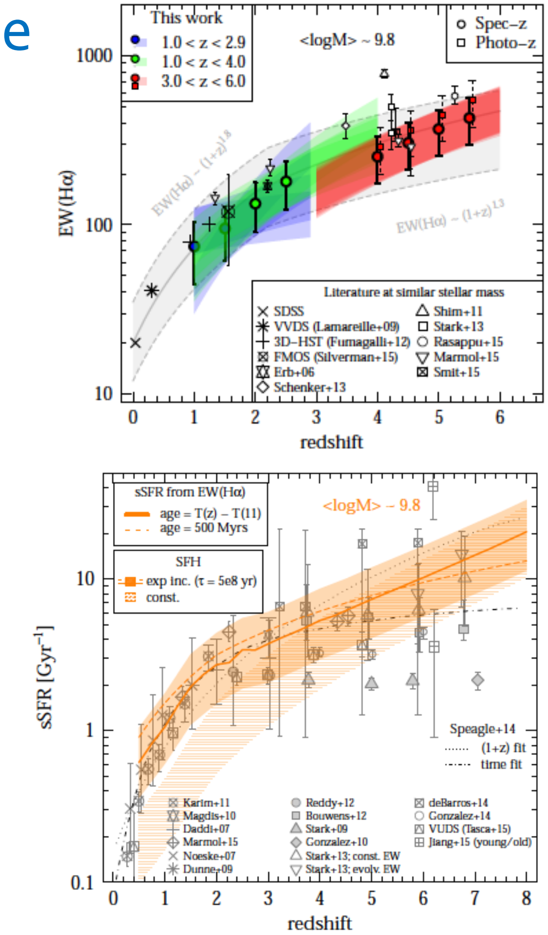
- Can clearly see signatures in raw photometry
- Both in spec-z and photo-z sample
- Fit our forward model to the data to extract line EW



Faisst, Capak et al. 2016a

Photometric line distributions estimates agree with 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

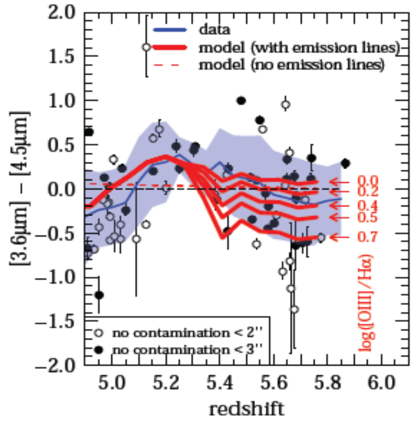


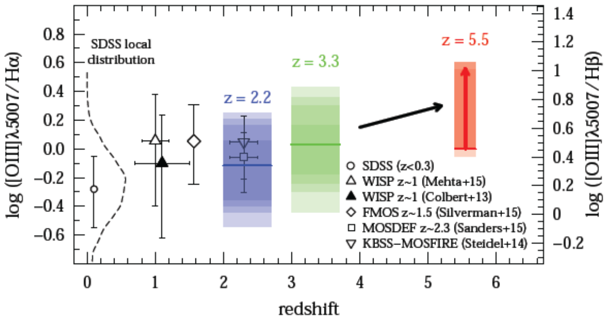
Faisst, Capak et al. 2016a

Can estimate line ratio's at z>4 for JWST targeting

- O[III]/Hα is also measured by our forward model

- Consistent with other estimates and measurements



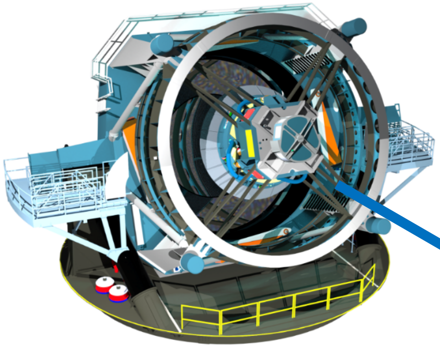


Faisst, Capak et al. 2016a

We can statistically combine data sets

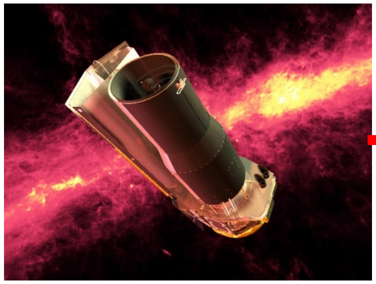
- Generate model with one data set
- Expand it to higher dimensions with another

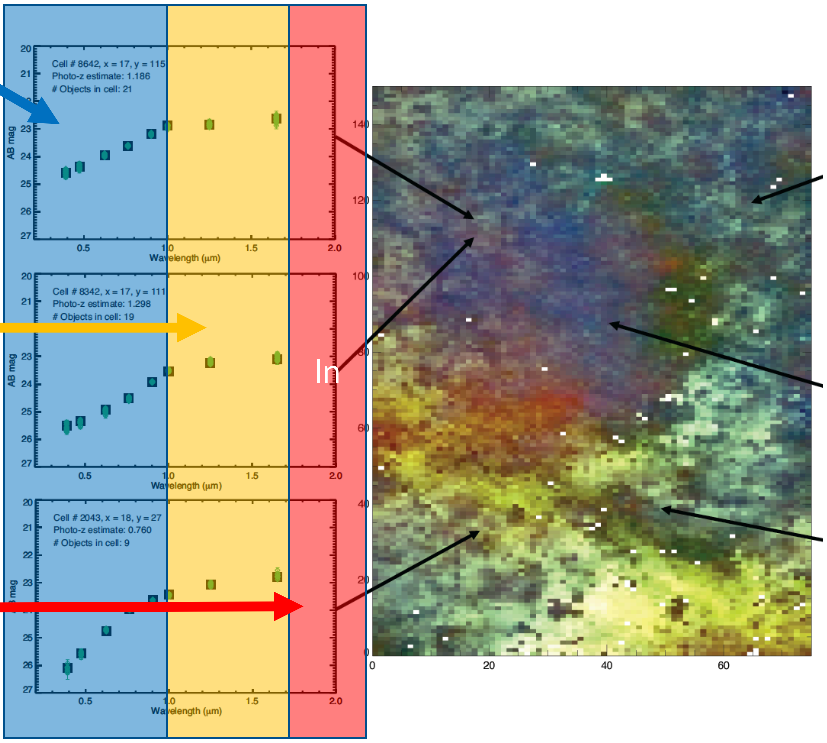
LSST



WFIRST

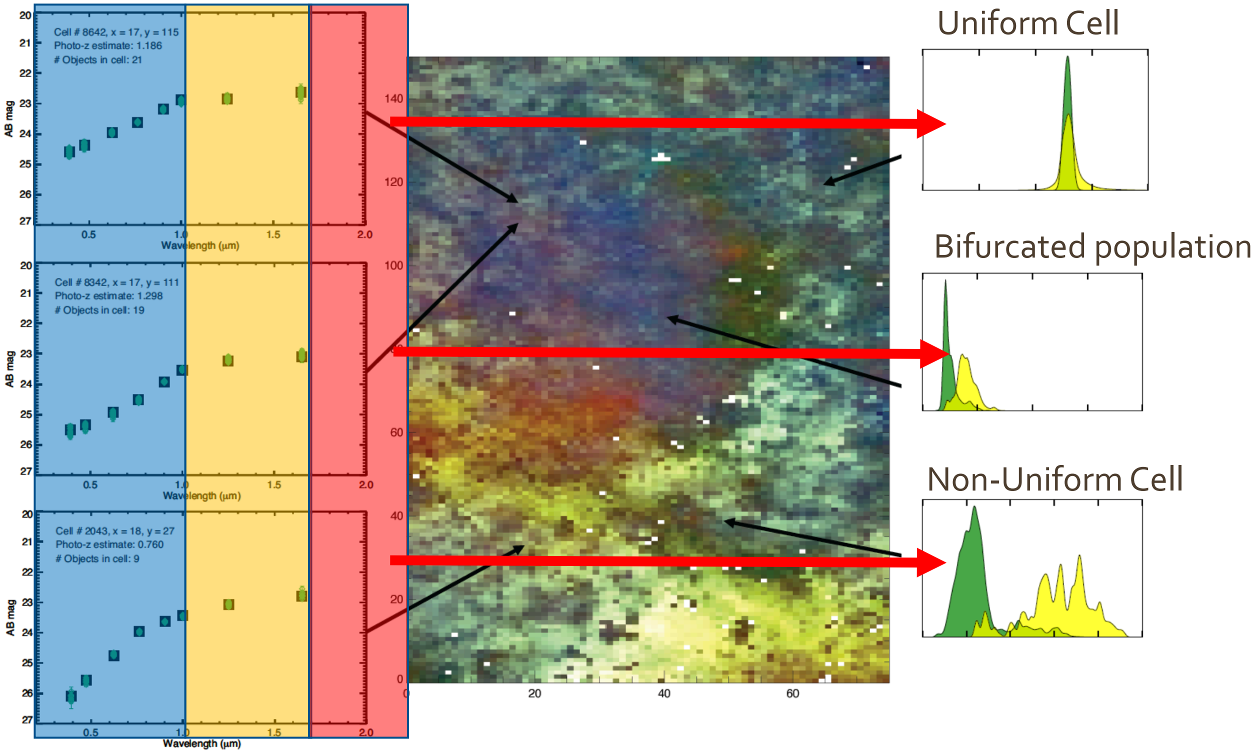
Spitzer





This allows us to combine knowledge in a standard model

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



Uniform Cell

Bifurcated population

Non-Uniform Cell

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.