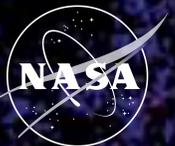


Spectroscopic Databases and Manifold Learning for Surveys of the 2020s

Dan Masters

Jet Propulsion Laboratory, California Institute of Technology

December 6, 2018



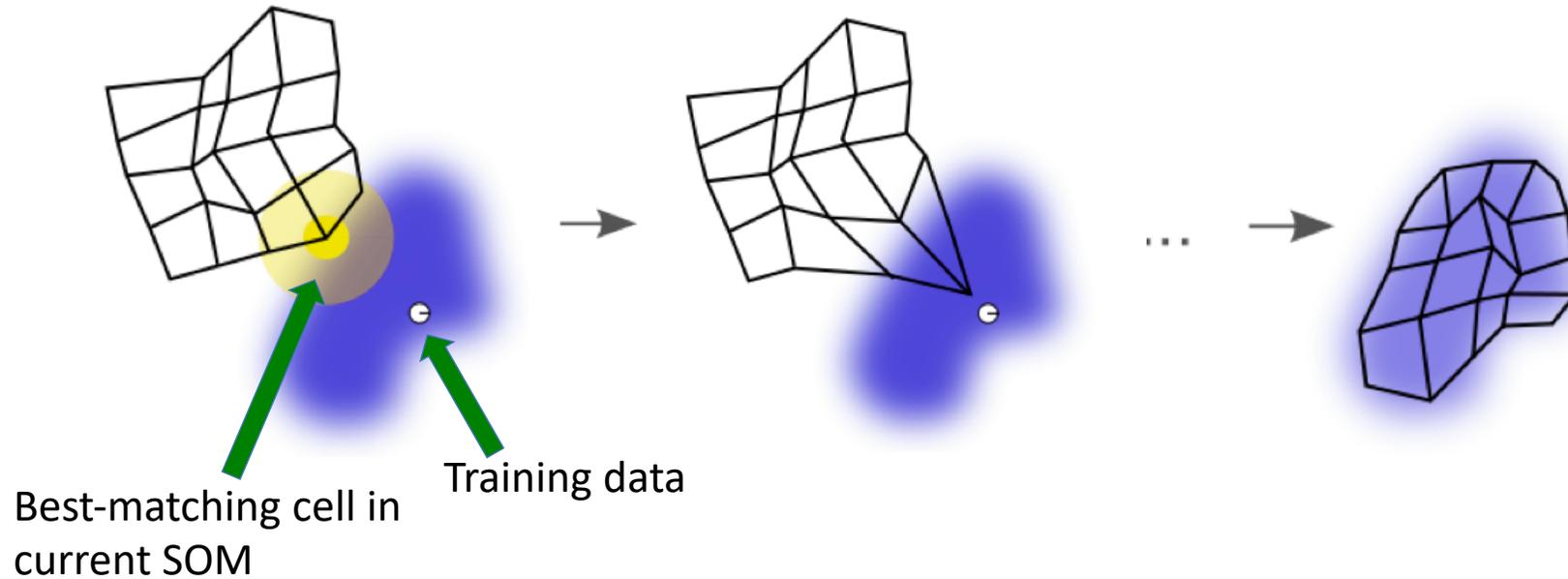
Jet Propulsion Laboratory
California Institute of Technology

© 2018 California Institute of Technology
Government sponsorship acknowledged.

Manifold learning / nonlinear dimensionality reduction (NLDR)

- Group of techniques to characterize / explore high-dimensional data and correlations in high dimensions
- Common examples include the self-organizing map (SOM), t-SNE, local linear embedding (LLE), and UMap
- Most project the high-D manifold down to a lower-D representation
- Whereas deep convolution networks try to learn a complex high-dimensional relationship between input data and output labels, NLDR just tries to unwrap the high-D data in an unsupervised way – no outputs

Training the SOM 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.

The galaxy color manifold

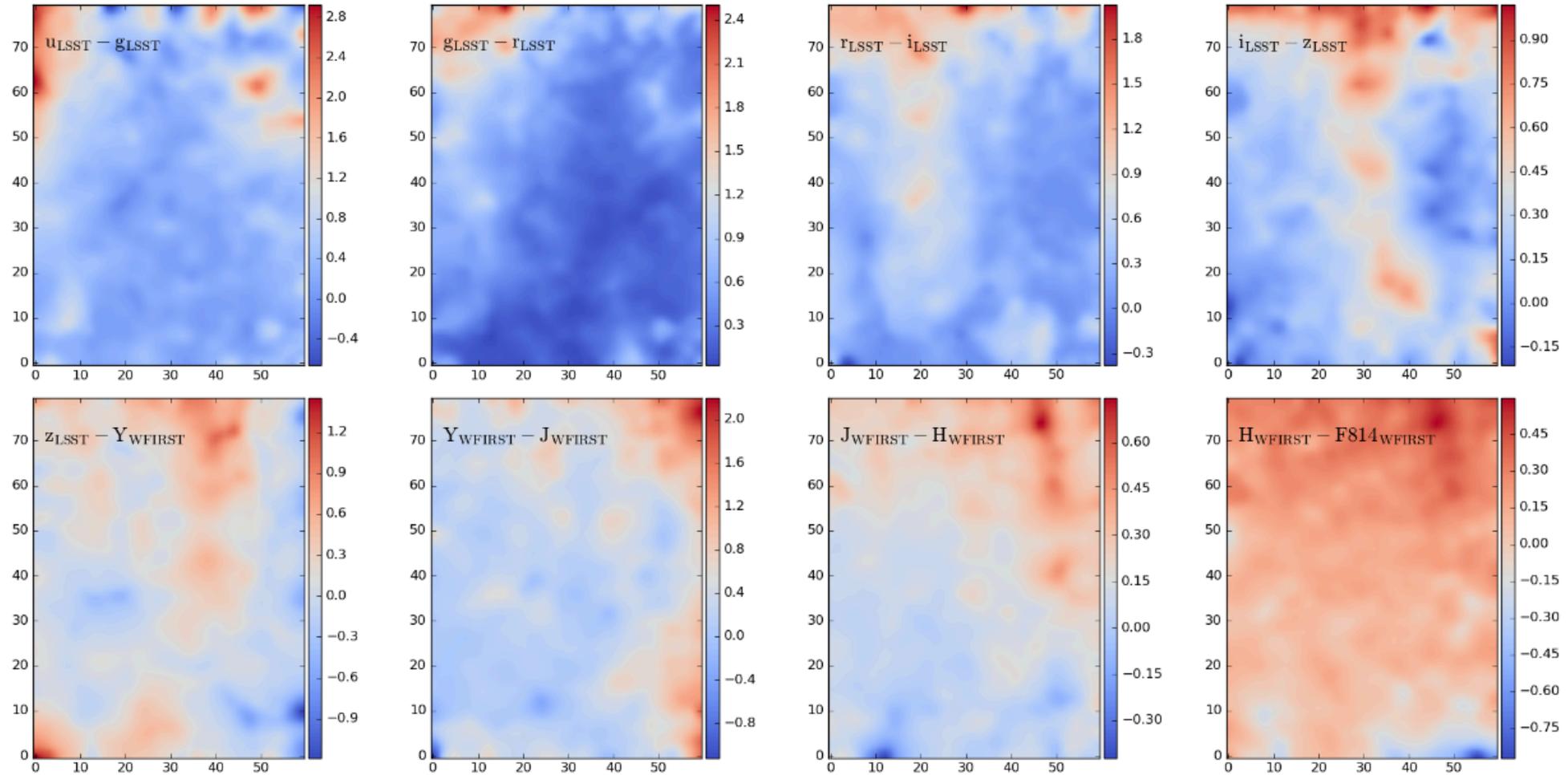
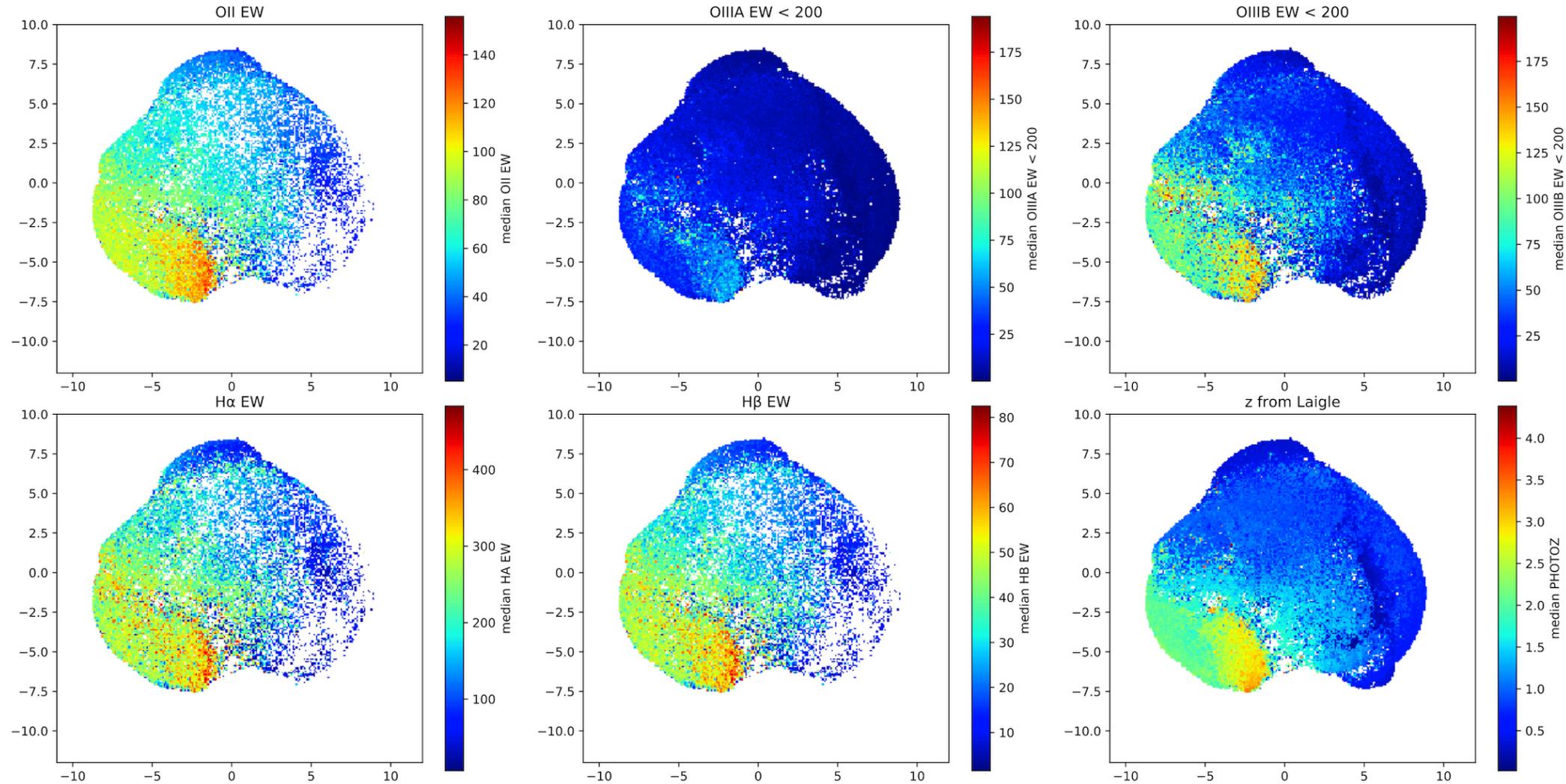
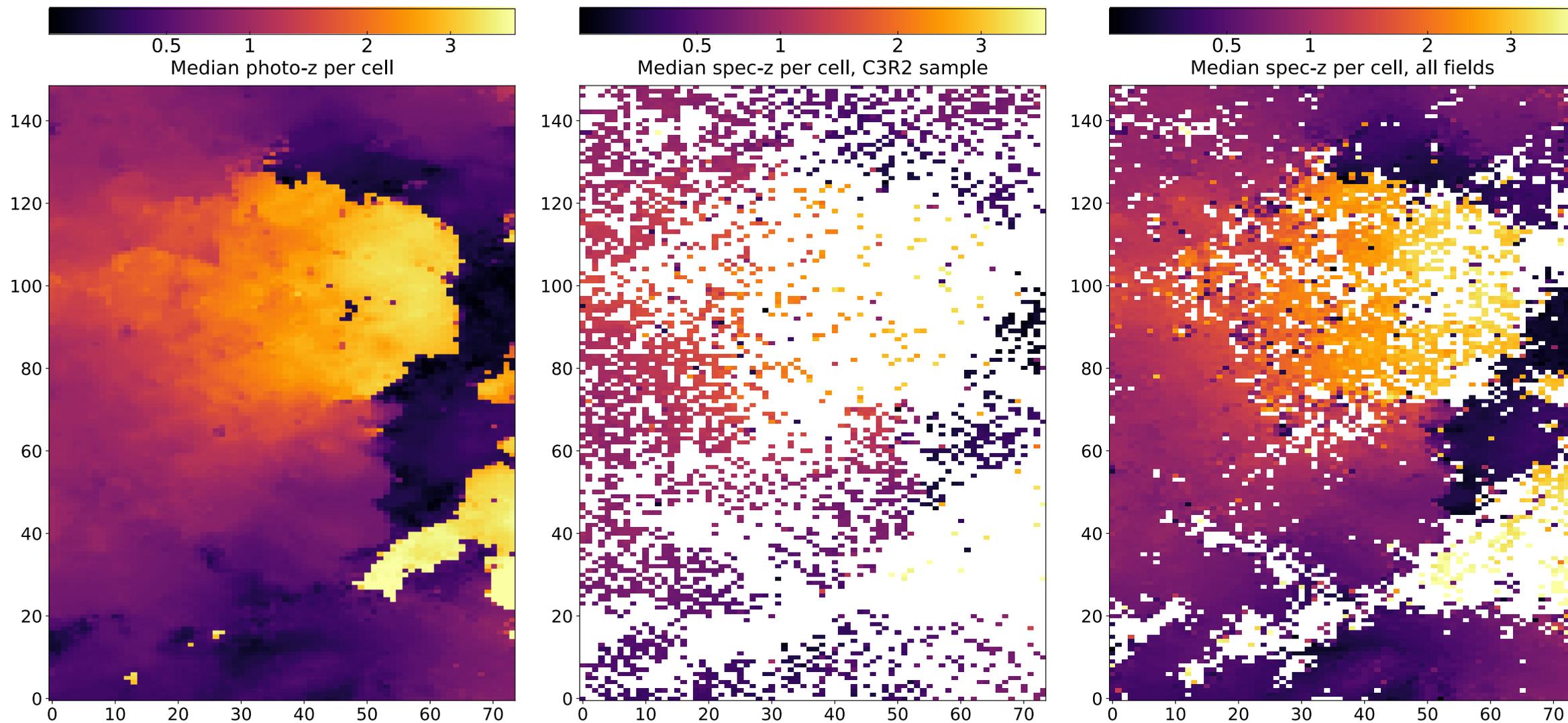


Figure 5. LSST and *WFIRST* colors of the trained SOM at each cell from top-left to bottom-right color-coded by: $u_{\text{LSST}} - g_{\text{LSST}}$, $g_{\text{LSST}} - r_{\text{LSST}}$, $r_{\text{LSST}} - i_{\text{LSST}}$, $i_{\text{LSST}} - z_{\text{LSST}}$, $z_{\text{LSST}} - Y_{\text{WFIRST}}$, $Y_{\text{WFIRST}} - J_{\text{WFIRST}}$, $J_{\text{WFIRST}} - H_{\text{WFIRST}}$, and $H_{\text{WFIRST}} - F814_{\text{WFIRST}}$. SOM is selected to be a mesh of 80×60 cells. The axes are arbitrary and each position on the two dimensional map points to a position in the 8 dimensional color space.

Other techniques - UMap



Model of galaxy manifold



C3R2 = Complete Calibration of the Color-Redshift Relation

Judith Cohen (Caltech) - PI of Caltech Keck C3R2 allocation

16 nights (DEIMOS + LRIS + MOSFIRE, [kicked off program in 2016A](#))

Daniel Stern (JPL) - PI of NASA Keck C3R2 allocation

10 nights (all DEIMOS; “Key Strategic Mission Support”)

Daniel Masters (JPL) – PI of NASA Keck C3R2 allocation 2018A/B

10 nights (5 each LRIS/MOSFIRE; “Key Strategic Mission Support”)

Dave Sanders (IfA) - PI of Univ. of Hawaii Keck C3R2 allocation

6 nights (all DEIMOS) + H20

Bahram Mobasher (UC-Riverside) - PI of UC Keck C3R2 allocation

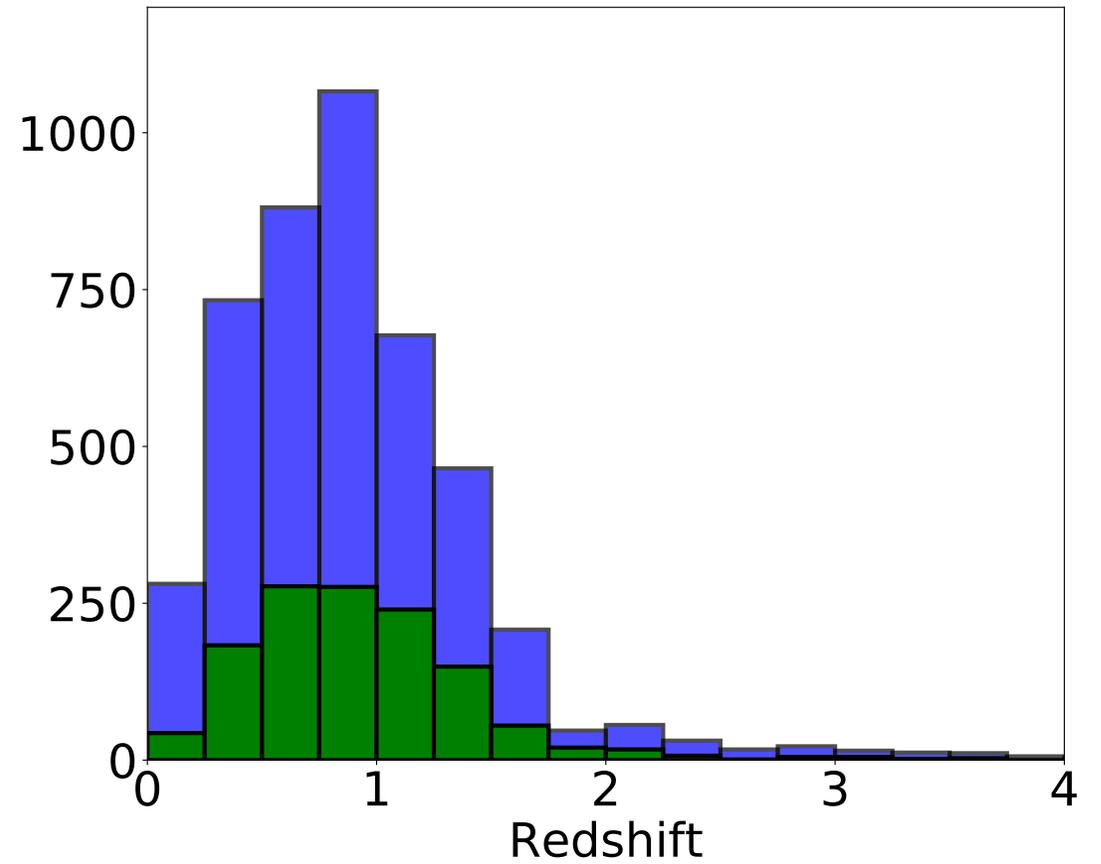
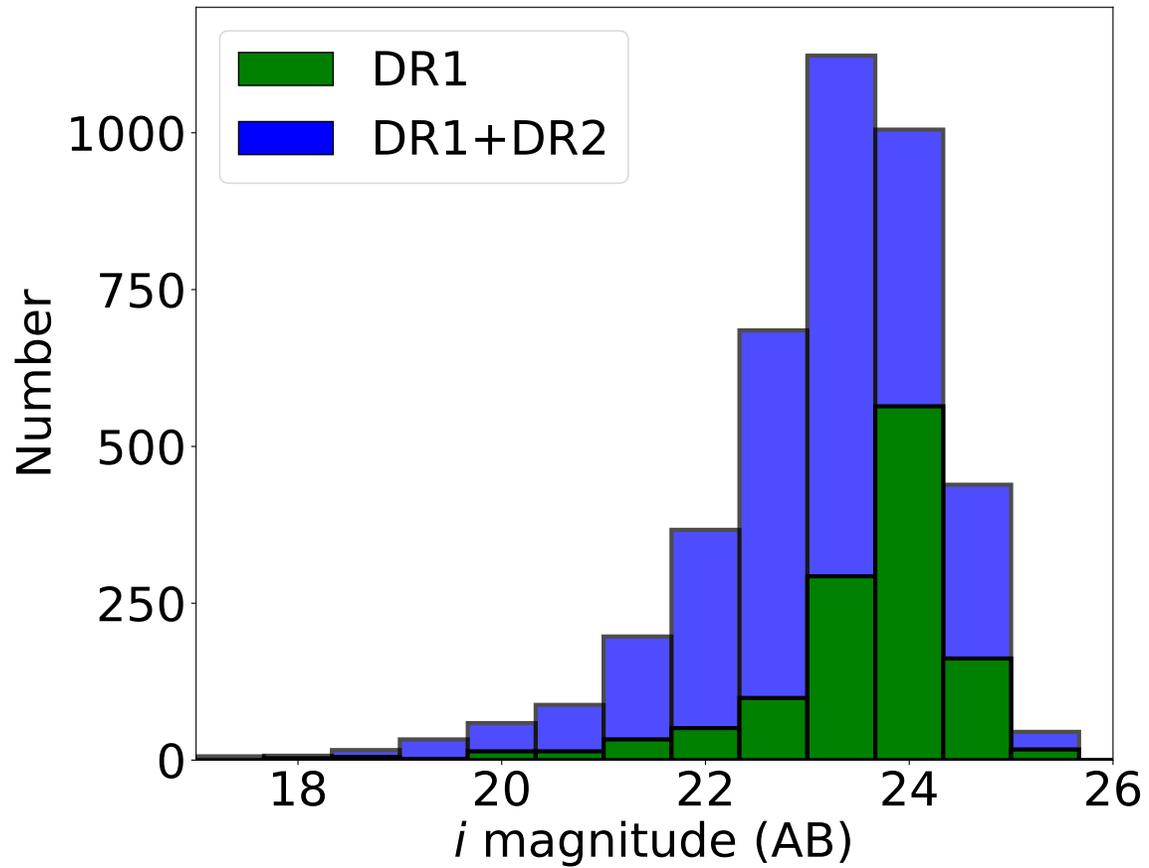
2.5 nights (all DEIMOS)

+ time allocations on VLT (PI F. Castander), MMT (PI D. Eisenstein), and GTC (PI C. Guitierrez)

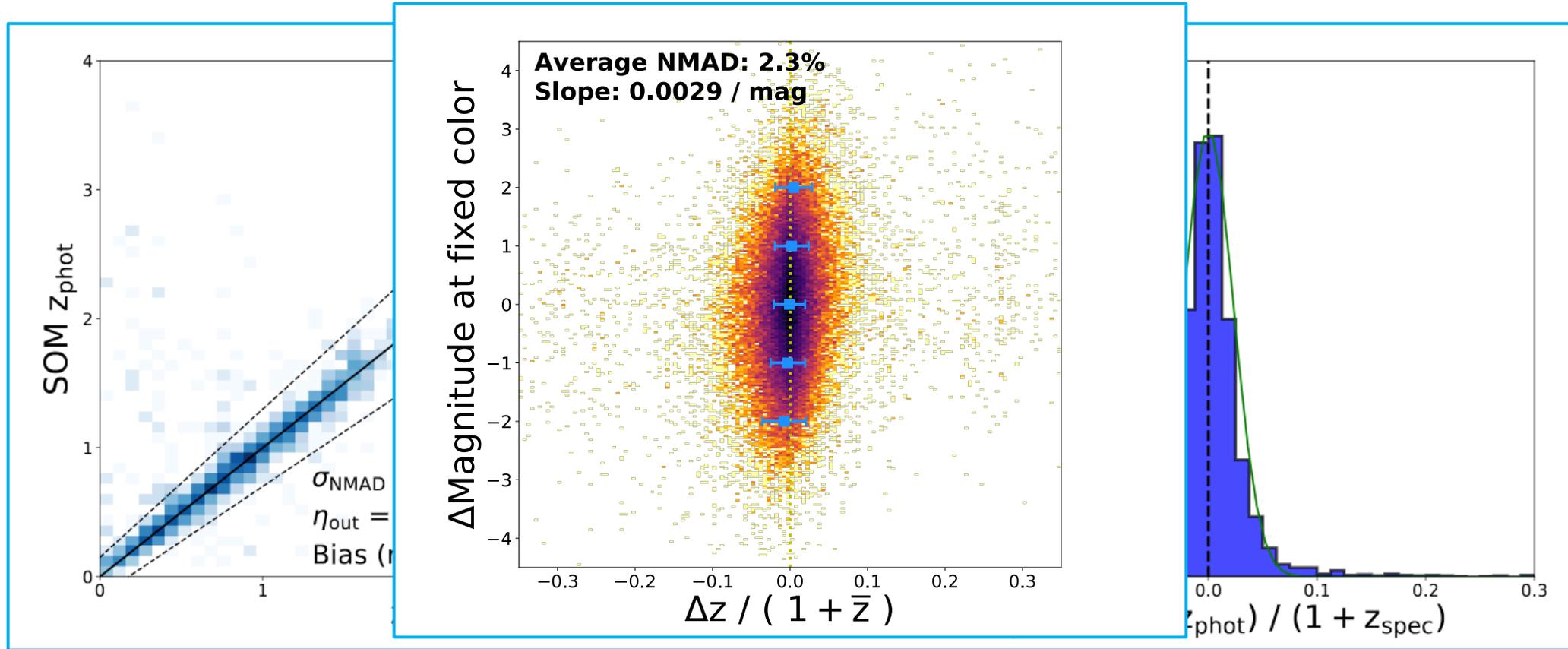
-Sample drawn from 6 fields totaling $\sim 6 \text{ deg}^2$

Additional Collaborators: Peter Capak, S. Adam Stanford, Nina Hernitschek, Francisco Castander, Sotiria Fotopoulou, Audrey Galametz, Iary Davidzon, Stephane Paltani, Jason Rhodes, Alessandro Rettura, Istvan Szapudi, and the Euclid Organization Unit – Photometric Redshifts (OU-PHZ) team

C3R2-Keck stats through DR2 (2016A-2017A)

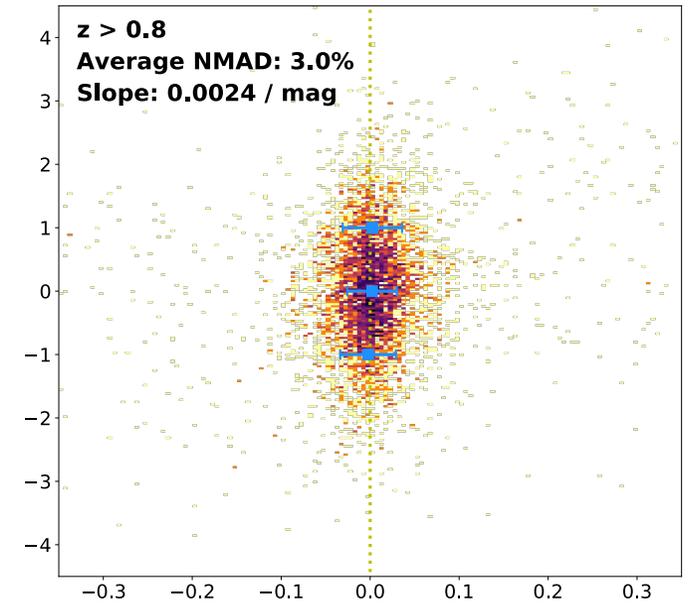
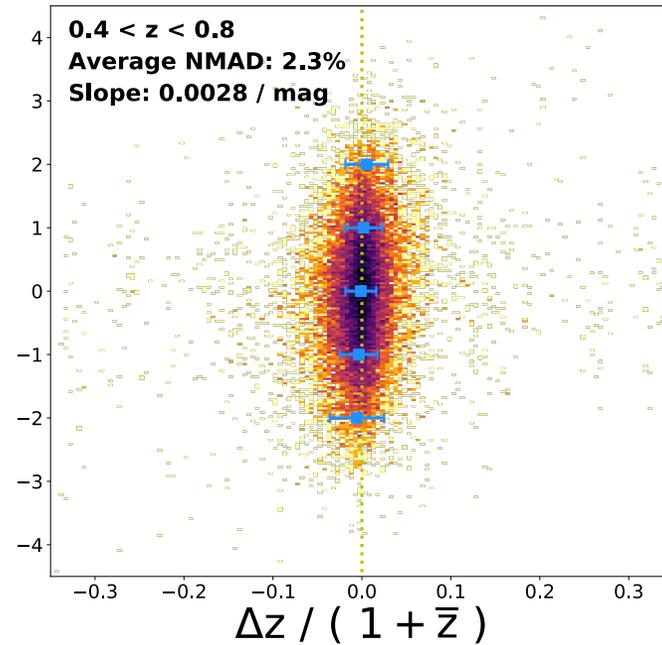
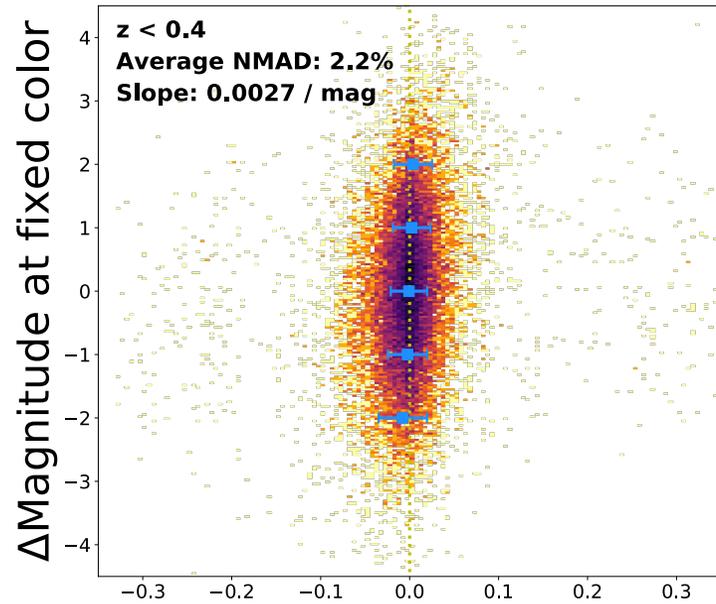


C3R2 – Results from SOM method



- Compare spectroscopic fraction of outliers (1% of galaxies) to SOM position ($\eta_{\text{out}} = 2.3\%$)
- Illustrates weak (and measurable) secondary dependence of redshift on magnitude *at fixed color* in the LSST+Euclid color space

Remarkably stable relationship of $dmag/dz$ at fixed color



Spectroscopic Databases: Requirements

- We will have hundreds of thousands of deep galaxy spectra in the mid-2020s
- Careful vetting necessary for calibration sample
- Database that can easily ingest new spectroscopy (e.g., from grisms)
- Machine learning-based redshifts may prove critical
- Huge task - how do we get there?

What happens to the manifold when we go deeper, as with WFIRST?

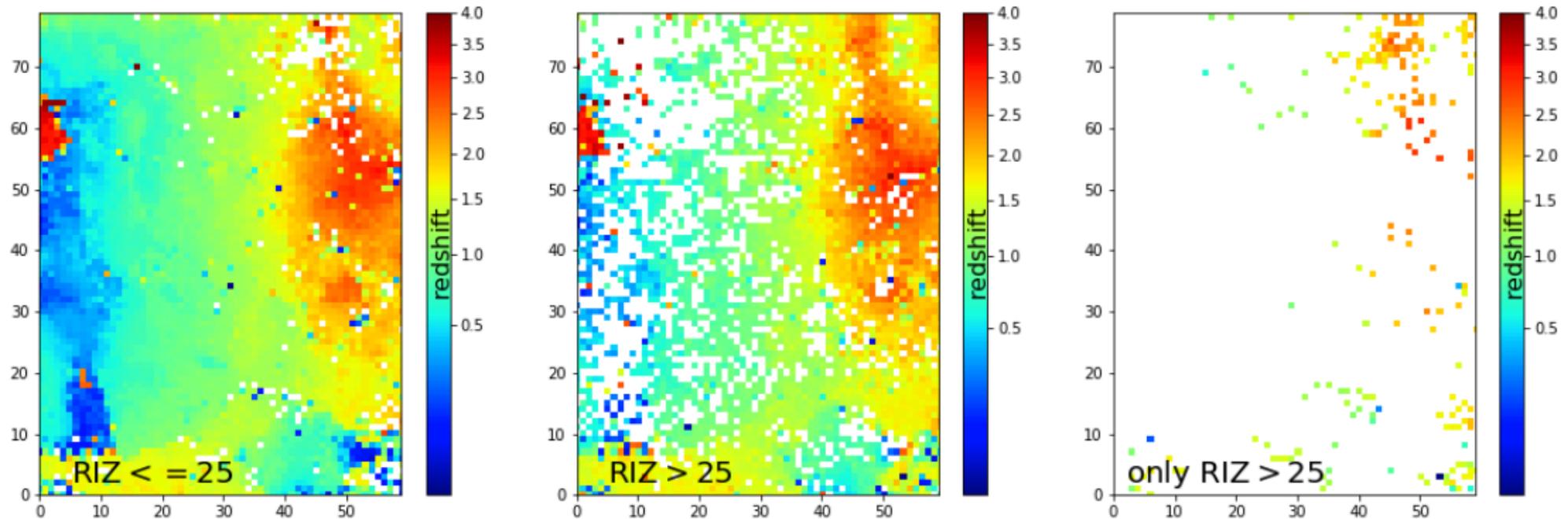
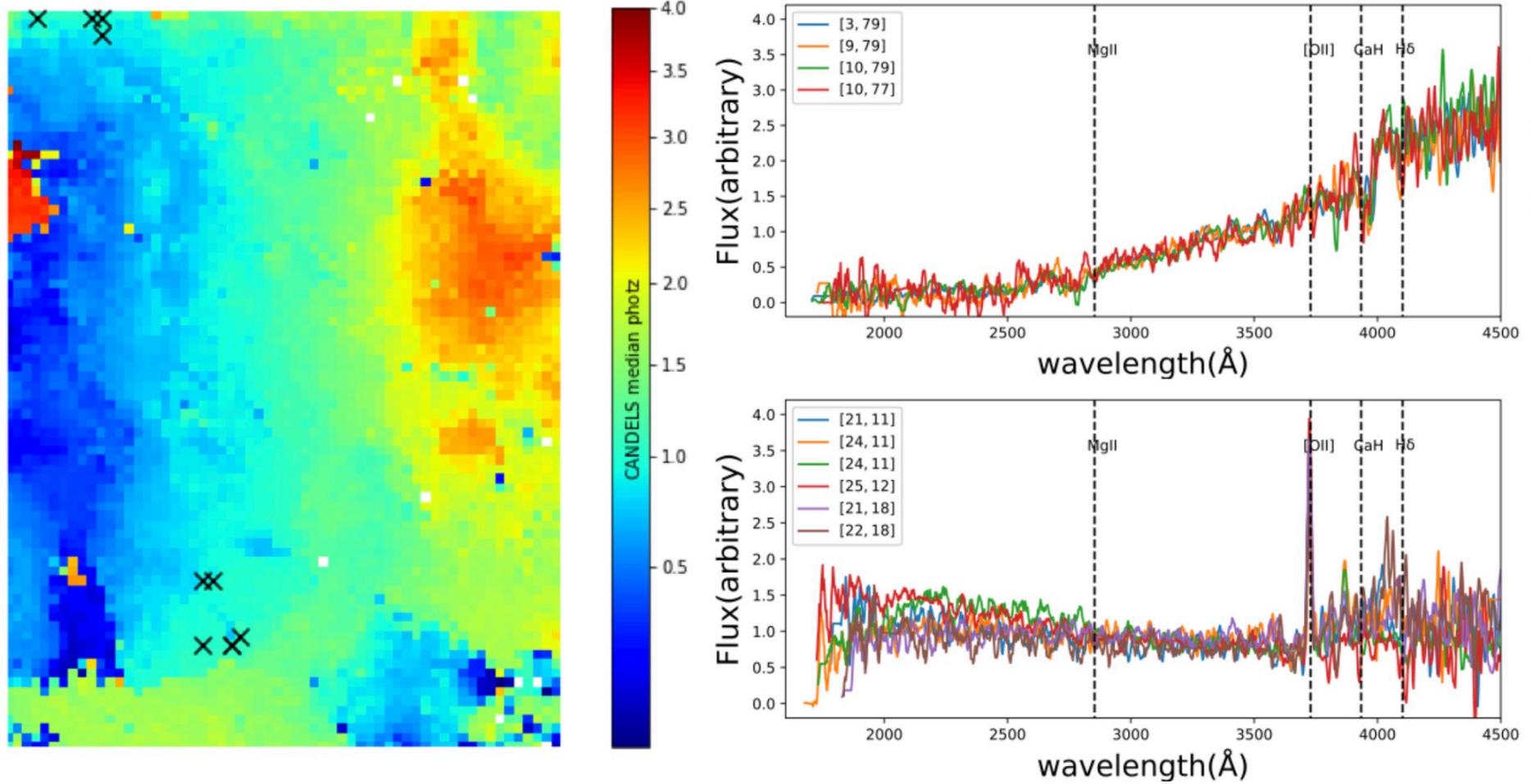


Figure 10. Bright ($r_{\text{IZ}} < 25$; *Euclid* depth) and faint ($r_{\text{IZ}} > 25$) galaxies in *WFIRST* lensing sample are mapped to the SOM color coded by median redshifts (shown on left and middle panels). More than 95% of the SOM cells contain at least one bright galaxy, $\sim 71\%$ of the SOM cells contain at least one faint object, and only $\sim 4\%$ of cells contain only faint galaxies (right panel).

Galaxies at fixed observable (e.g. color) are spectrally very similar



Galaxies are not unique

- The manifold of galaxy observables is finite.
- We can measure it really well with large surveys.
- Continuity constraints could then allow us to build a dynamic picture of galaxy growth
 - Individual galaxies can be thought of as moving along the manifold.
- What could we learn from this?

Measure the high-dimensional manifold. Then what?

- We have a well-defined target for simulations
- What if we find (as is common) that the simulations produce unphysical galaxies, or can't produce certain real galaxies?
- Is there a way to systematically search for the simulation parameters that produce the observed universe?
- What have we learned about galaxies at the end?
- Manifold mapping can definitely enable very rapid physical parameter estimation.

Models – can we match them to the data?

