VERA C. RUBIN OBSERVATORY

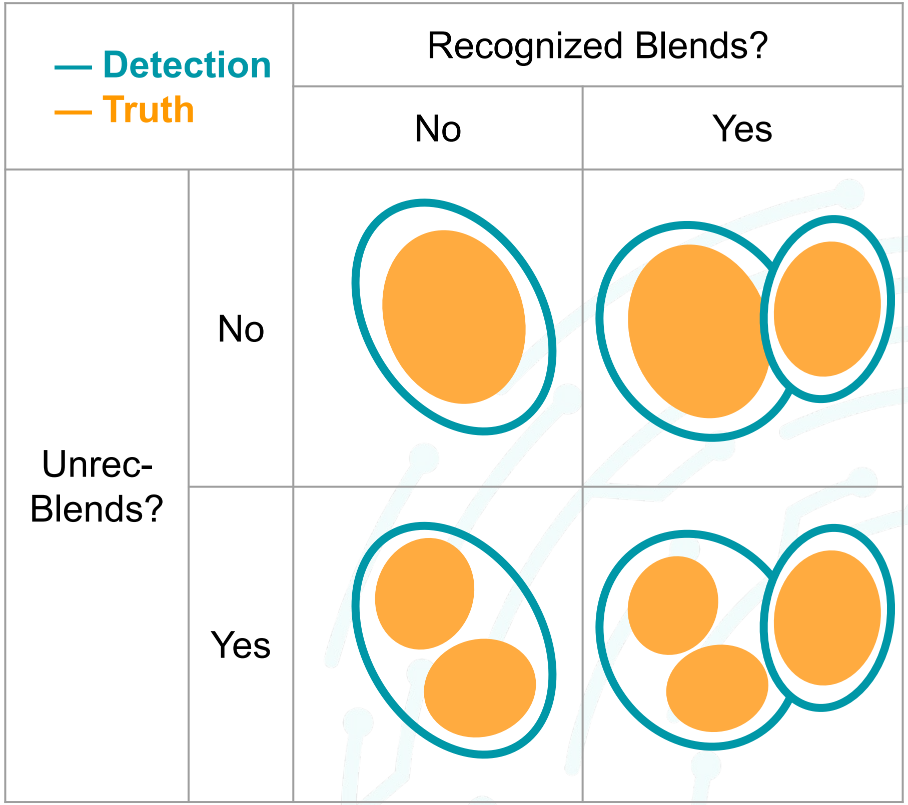
Recognizing the Unrecognized Blends

Shuang Liang
Rubin Observing Specialist

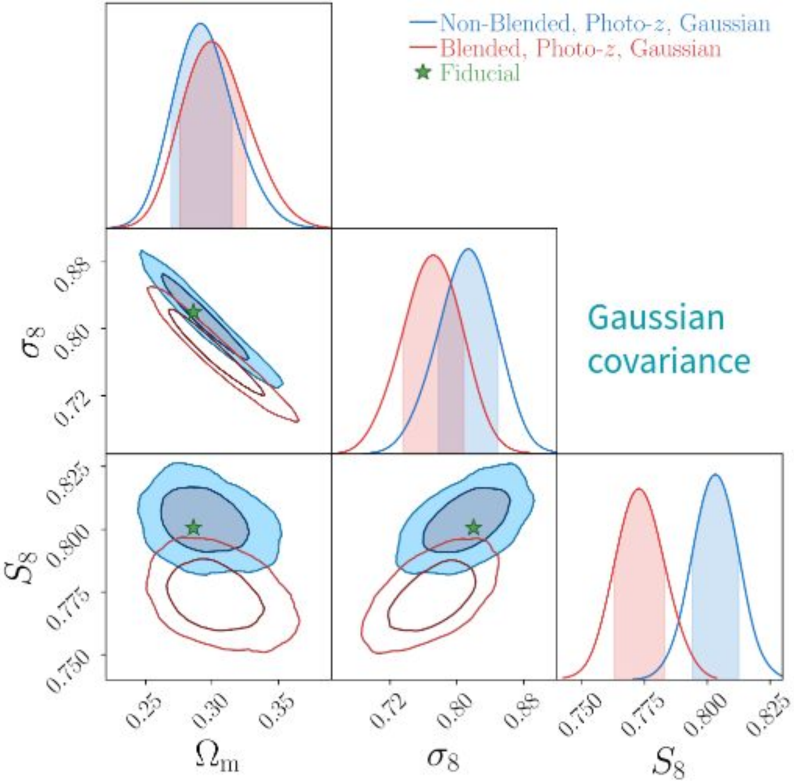
Project & Community Workshop
08 Aug 2023

Recognized vs. Unrecognized Blends

- **Recognized Blends**: Overlapping objects but correctly identified, and assigned with reconstructed properties ("deblended").
  - Photo-z impacted by photometry
  - Shape estimate is noisy
  - Most deblenders assume correct detection of peak counts
- **Unrecognized Blends:** Objects overlapping too much, detected as one object.
  - Colors can be weird
  - Shape is usually wrong



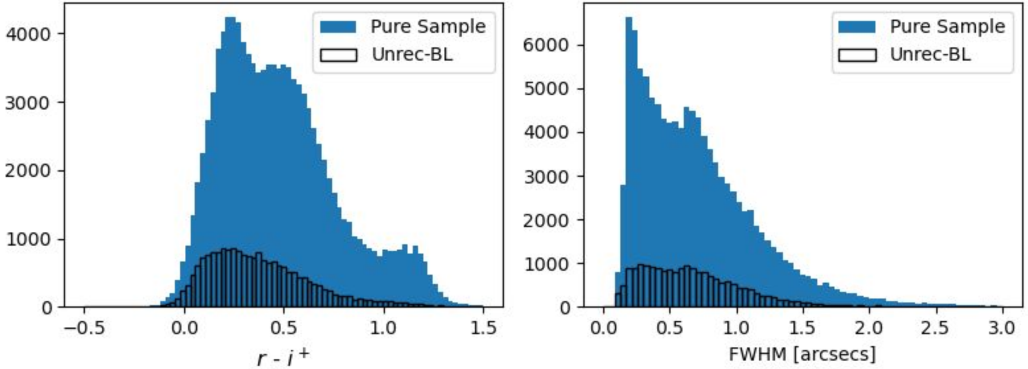
How Bad Are Unrecognized Blends?

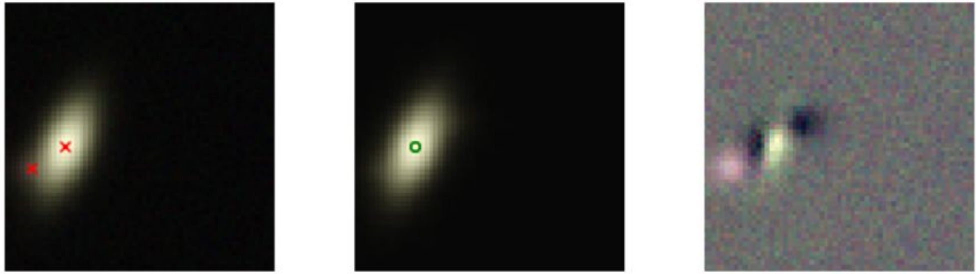


- *Dawson et al. 2016*: 14% "ambiguous" (unrecognized) blends at i~25.3
- We find: more than that, depending on the definition

- **Unrecognized blends can cause**:
  - 0.025 decrease in S_8 through cosmic shear at i<24 (*E. Nourbakhsh et al. 2022*; slides)
  - 1~2σ difference in galaxy clustering 2pt correlations on DC2 (*B. Levine et al. 2023 (in prep)*; slides)
  - 20% drop in cluster shear profile (*M. Ramel et al*, DESC Project 284; slides)

Detecting Unrecognized Blends with ML



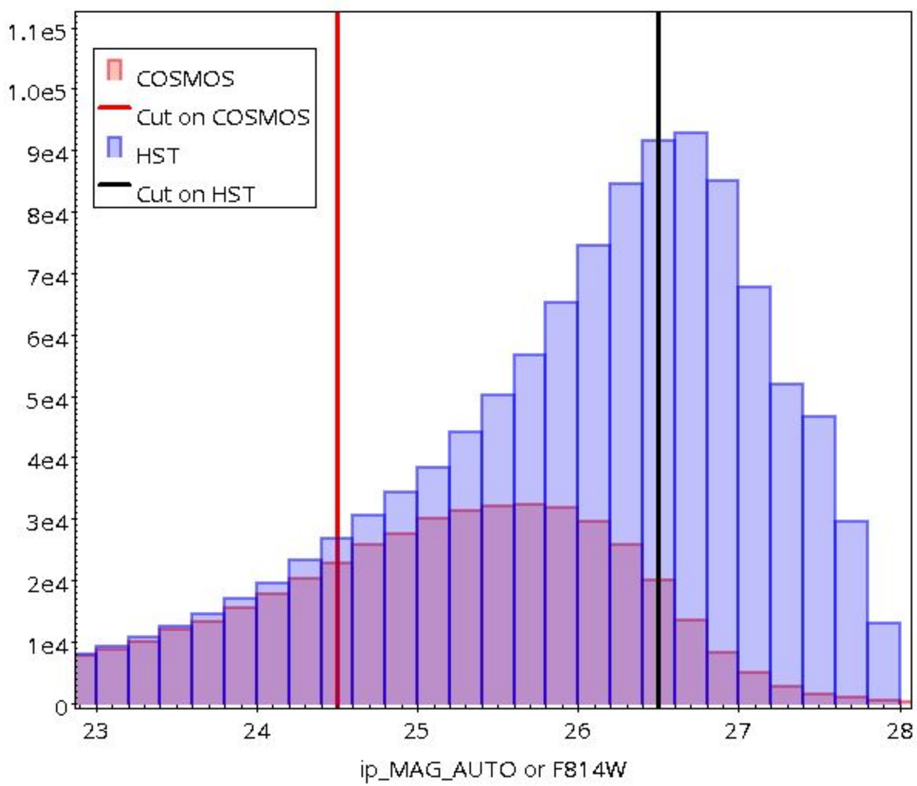


- **Catalog-based (this talk):**
  - Use Machine Learning to capture unique color/morphology combination of blends
  - Fast, but heavily affected by image processing pipeline

- **Pixel-level detection:**
  - Use Machine Learning to detect color gradient within an object (dipole-like model residual)
  - Computation-intensive

Image: S. Kamath 2020 (Ph.D. Thesis)

Catalog-based Detection: Data Set



Ground-based observation: The COSMOS catalog (Laigle et at. 2016)

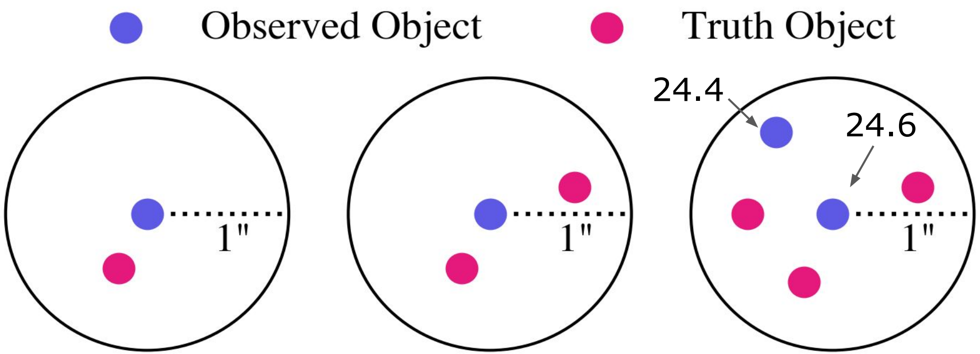
- Precise photo-z from 30 band photometry
- Depth and PSF comparable to LSST gold sample
- No shape measurement, only size

"Truth" catalog: HST coverage of the COSMOS field (Koekemoer et al. 2007)

- Awesome resolution
- Deep, but not deep enough for <2 mag blends → pushes to a <24.5 mag cut on COSMOS

Spec-z: A compilation of C3R2 + zCOSMOS + DEIMOS + VUDS; ~20,000 matched spectra

Catalog Matching: A Hidden Pitfall

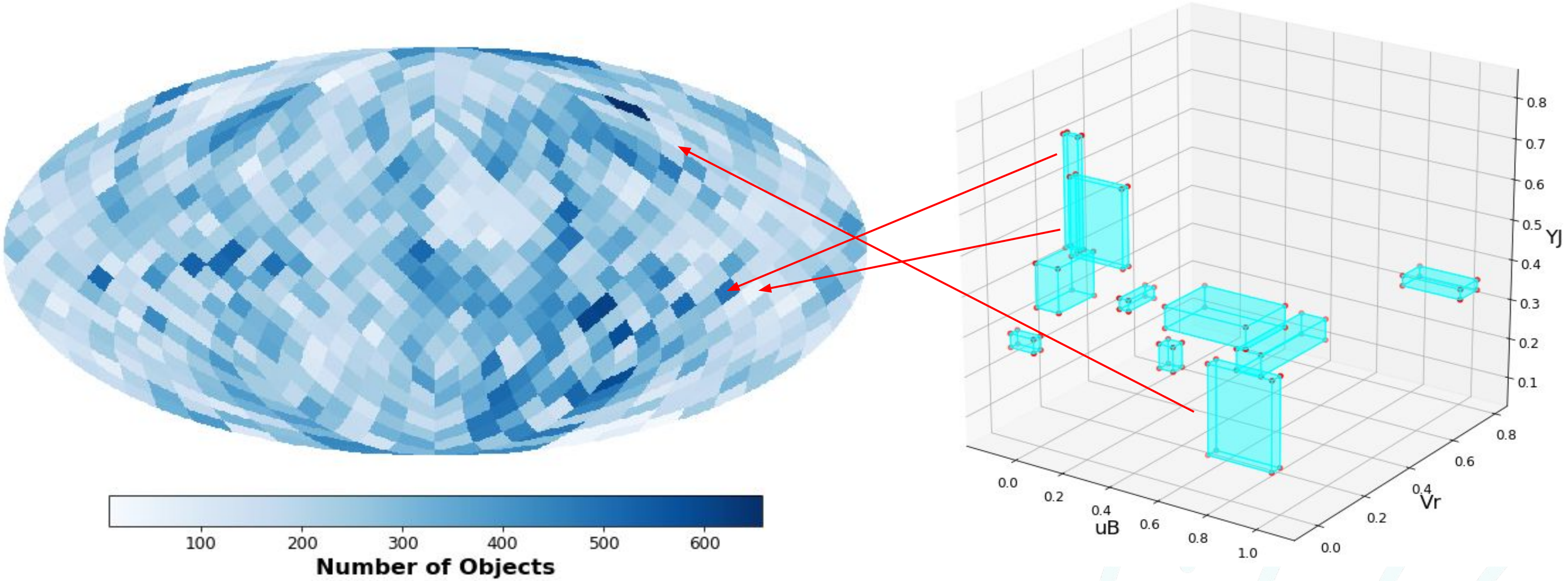


ANY selection applied to the ground-based catalog pre-matching can lead to mis-labeling of unrecognized blends! The selection includes (but not limited to):

- Magnitude / SNR cut
- Star/galaxy separation
- Photometry quality flags: saturation, bleeding, cosmic-ray, satellites, bad pixel/column, truncation, …

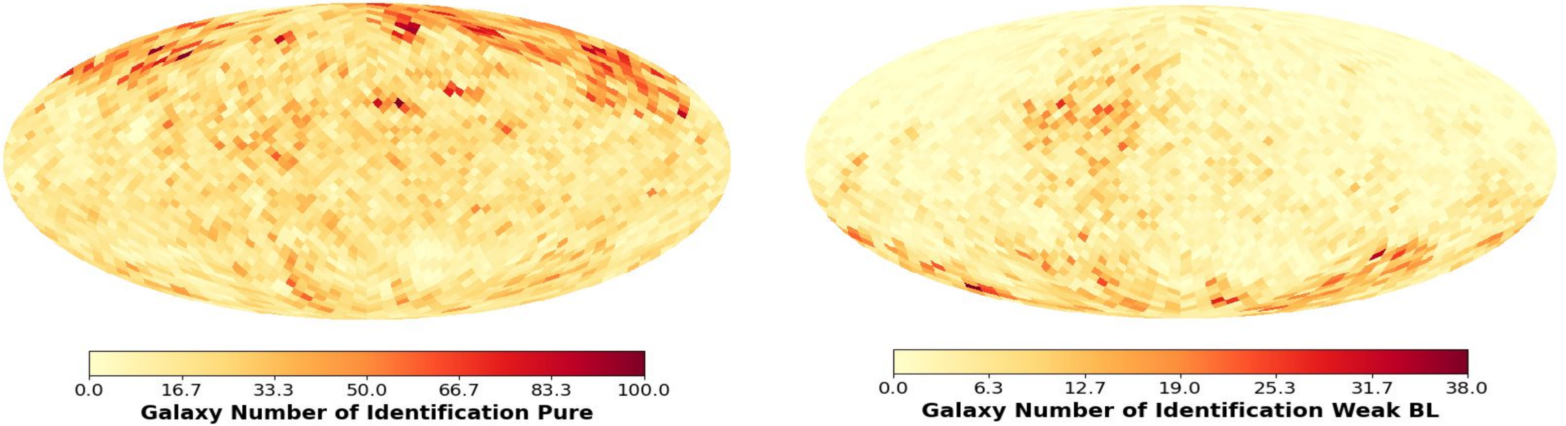
Takeaway: Match first, then select!

SOM: A Map from N-d Features to 2-d



A self-organizing map (SOM) is an unsupervised neural network designed to represent a high-dimensional data set with a low-dimension (usually 2-d) map. SOM is **topology-preserving**, meaning galaxies with similar high-dimensional features are close to each other in the 2-d map.

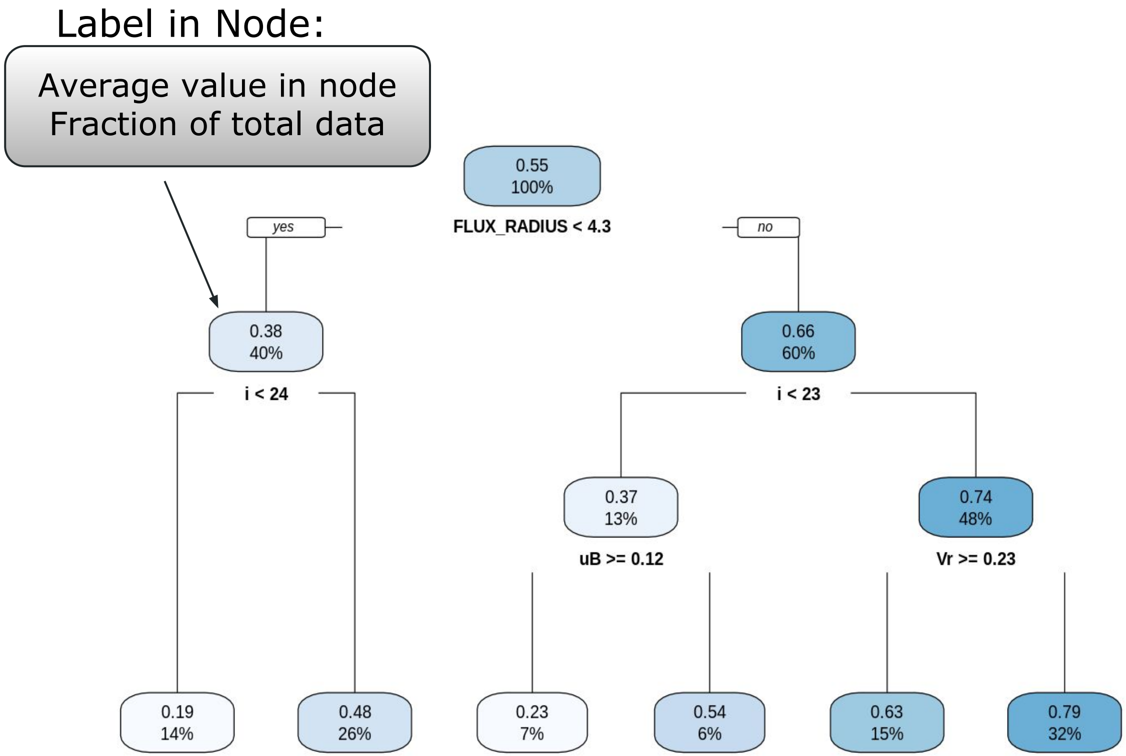
Detecting Unrecognized Blends with SOM



Using Features (*uB*, *BV*, *Vr*, ri^{+}, $i^{+}z^{++}$, $\boldsymbol{i^{+}}$, FWHM, $z^{++}Y$, *YJ*, *JH*) for training
Using Features (*uB*, *BV*, *Vr*, ri^{+}, $i^{+}z^{++}$, $\boldsymbol{i^{+}}$, FWHM) for detecting unrec-bl

Pure sample and Unrec-BL occupy very different regions in SOM!
Unique features detected! 🎉🎉🎉

Detecting Unrec-BL with Random Forest



Using Features (uB, BV, Vr, ri^{+}, $i^{+}z^{++}$, $\boldsymbol{i}^{+}$, FWHM)

The training sample is labeled with:

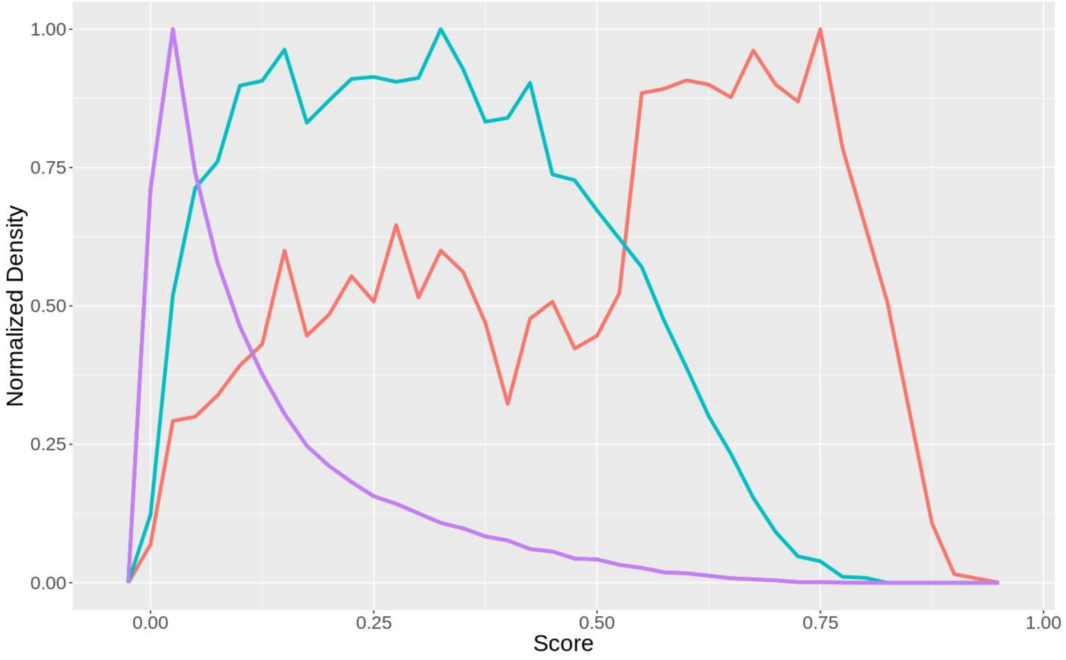
0 = Pure
1 = Unrecognized Blend

Each tree gets a subsample of training data and feature.

Each node maximizes the split of 0s and 1s

Each object in validation sample will get a score in a "leaf" from each tree, to be averaged out (voting) among trees for a final score

Score Histogram on Validation Sample



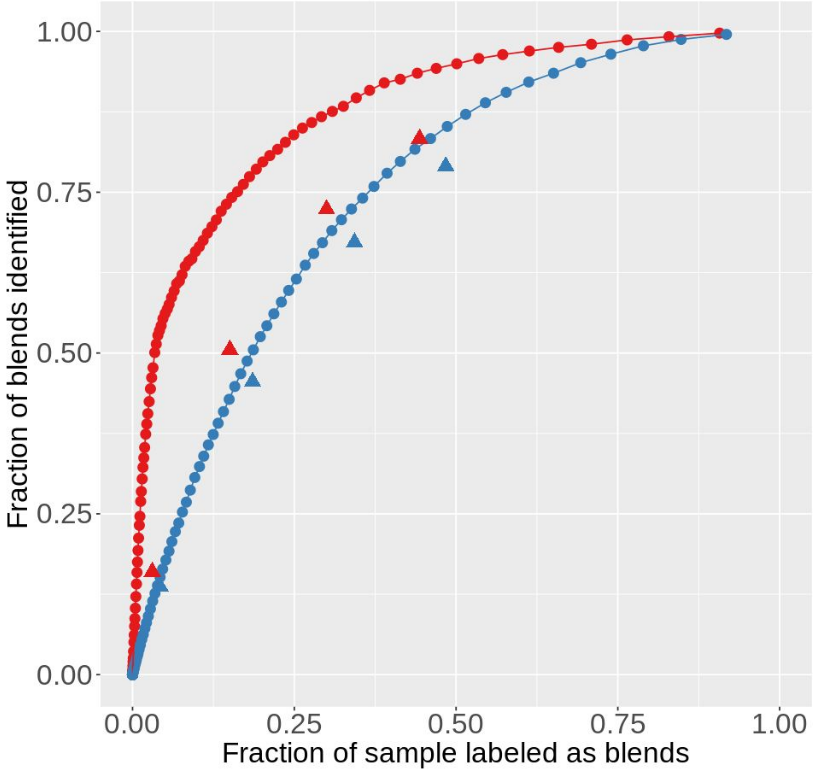
— pure

— weak blends, $|mag\_i_{obs} - mag\_i_{truth}| < 2$

— strong blends, $|mag\_i_{obs} - mag\_i_{truth}| < 1$

Pure sample and Unrec-BL getting very different scores!
Unique features detected! 🎉🎉🎉

Results From Both ML algorithms



At a depth of $i<24.5$, we can identify the majority of unrec-bl at very small cost!

●● : using Random Forest
▲▲: using Self-organizing Map

▲●: strong blends, $|mag\_i_{obs} - mag\_i_{truth}| < 1$

▲●: weak blends, $|mag\_i_{obs} - mag\_i_{truth}| < 2$

P. Adari et al. 2023 (in prep)

So You Think Your ML Algorithm Works Better

But what is better?

- ML is only as good as your training sample
- How to best define "unrecognized blends"?
  - Counting number of sources in a matching group (e.g. FoF)
  - Checking for overlapping of isophotes (e.g. Friendly)
  - Flux-weighted statistics (e.g. "purity" from J. Sanchez+2021)

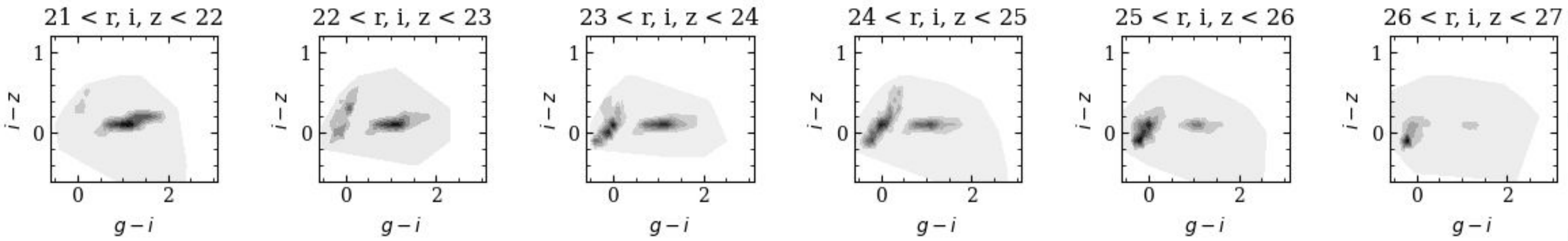
What kind of unrec-bl causes more trouble? Those with larger "bias" of:

- Shape measurement: what is the "true shape"?
- Photo-z estimation: what is the "true redshift"?
- What is a good non-one-to-one metric? (lost-and-found)
- Is the ultimate metric – cosmology inference – the only good metric?

Continuing Efforts

Try a deeper catalog of:

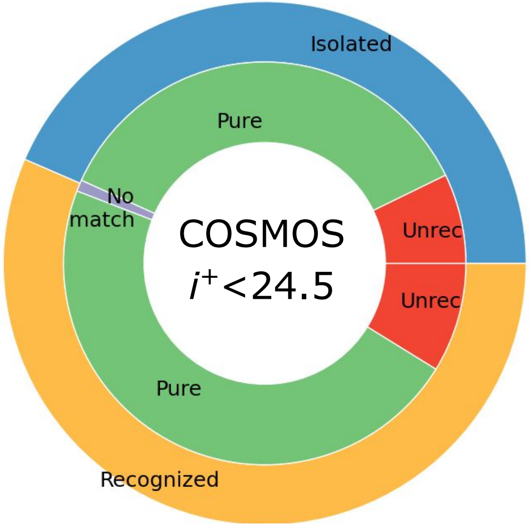
- Real data? – No truth catalog as deep as 25.3+2=27.3 mag yet
- Simulation? – Can we trust the color distribution?
  - Ongoing: use SYNTHETIC (git) for an LSST full depth simulation

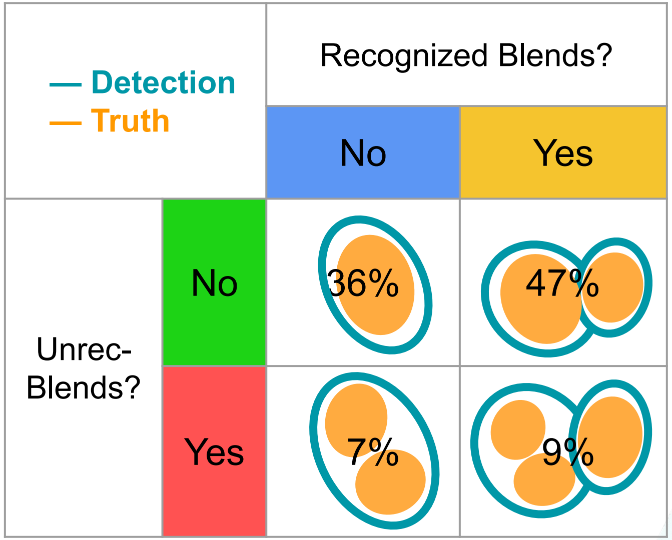


Unrec-BL ***x*** shear response:

- Metadetect: too much trouble; we might not be there yet
- BFD: seems possible; need more thoughts
- FPSF: **Amazingly, it is possible to consider the selection (removal) of unrec-bl with an analytical shear response pipeline!** (Only for SOM)

(backup) Catalog Matching: Some Numbers





	Total	Recognized	Isolated	Pure	Weak BL	Strong BL
Num. Sources	138849	78457	60392	116471	22378	5326
Pure Fraction	83.0%	83.1%	82.8%	100%	0%	0%
Num. Spectra Match	17935	12592	5343	16542	1393	318
Num. Photo-z Outliers	480	399	81	383	97	34
Outlier Fraction	2.68%	3.17%	1.52%	2.32%	6.96%	10.7%

Takeaways:

- **9% is more than 7%!**
- **Unrec-BLs have high photo-z outlier rate**

We define two types of unrecognized blends:

Weak: $|mag\_i_{obs} - mag\_i_{truth}| < 2$

Strong: $|mag\_i_{obs} - mag\_i_{truth}| < 1$

Figures and Tables on this page show weak blends unless specified.