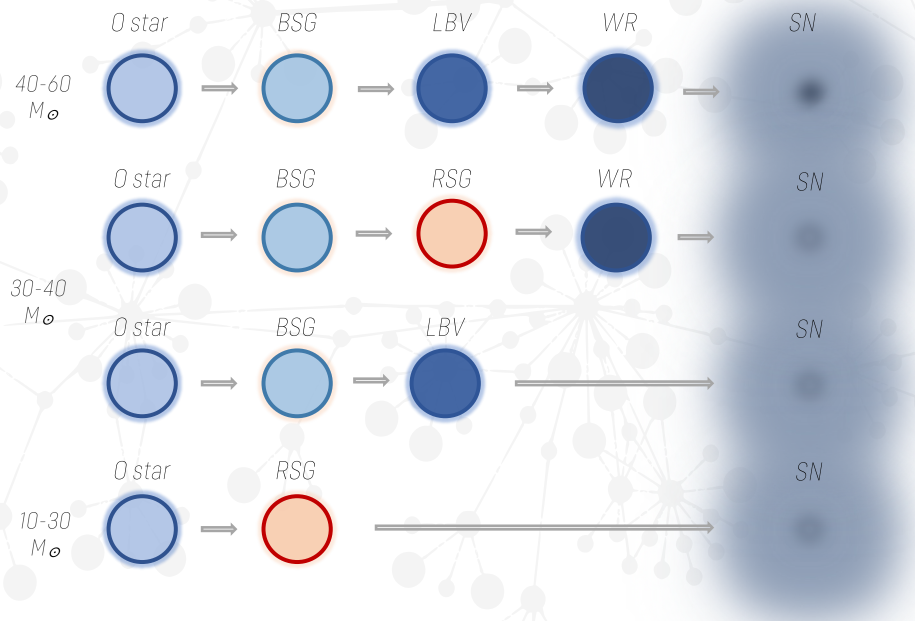
A semi-supervised "cluster-then-label" scheme for classification of evolved massive stars

Cristobal **Bordiu** – INAF OACt
and the Catania Radio Group

The science case

- Massive stars: key agents in the evolution of galaxies (chemistry, structure, dynamics)
- Post-MS evolution: short-lived, scarce, hard to detect – spectroscopic confirmation?
- Detection of new candidates is highly valuable
- JWST, LSST… automated photometric classifiers are critical to deal with the upcoming data deluge

Some (over)simplified stellar evolution tracks



The problem

We expect a continuum of evolutionary states, without perfect boundaries

- Scarce literature – supervised methods:
  - k-NN with IR colors to spot WR candidates – Morello+18
  - Coarse classification of Galactic objects (hot/cool/emission Line) – Dorn-Wallenstein+21
  - Ensemble classifier for extragalactic sources – Maravelias+22
- Reasonably good performances, but some caveats

Small datasets
Limited label reliability
Intrinsic class imbalances

→

Coarse classification schemes
Poor performance in minority classes

The challenge

Can we improve classifier performance on small/imbalanced datasets?

Semi-supervised learning – taking advantage of unlabelled data

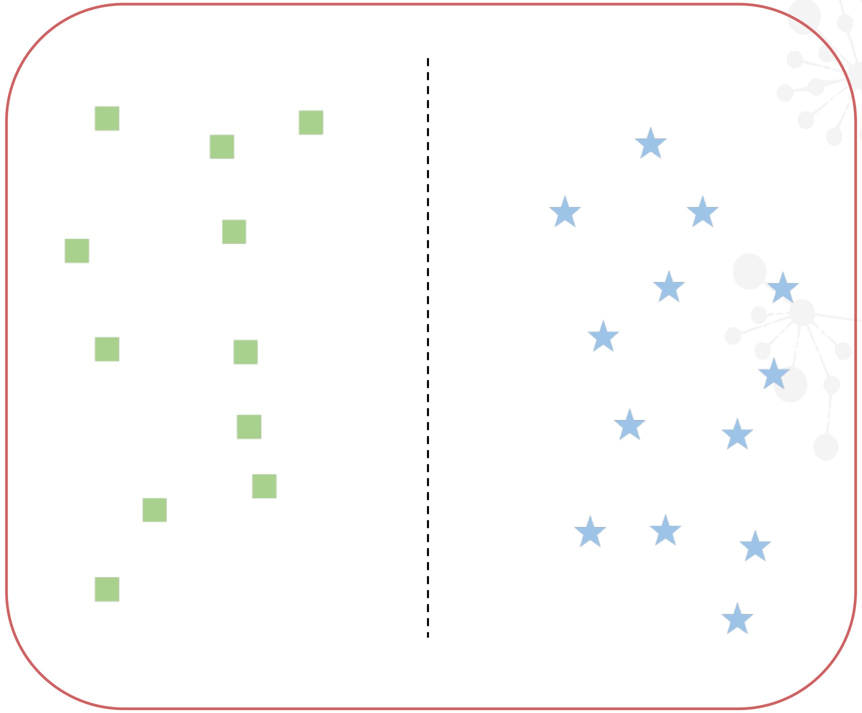
(abundant with newest observatories)

Clustering analysis

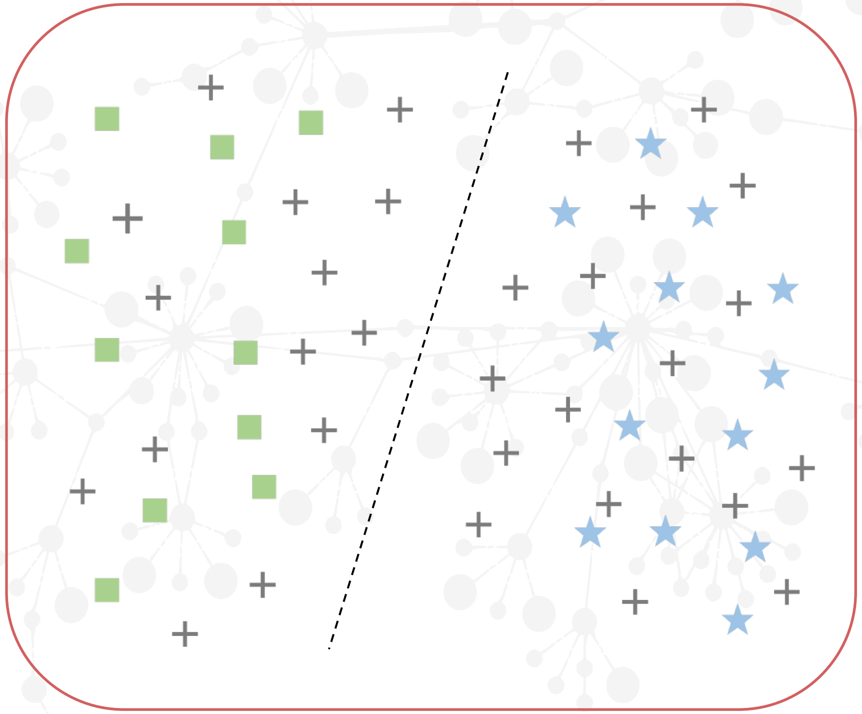
Finding partitions of the entire dataset for efficient pseudo-label generation

The challenge

Labelled data only



Labelled and unlabelled data



The goal

Investigate the performance of "cluster-then-label" methods for classification of evolved massive stars in local group galaxies

Building the dataset – samples

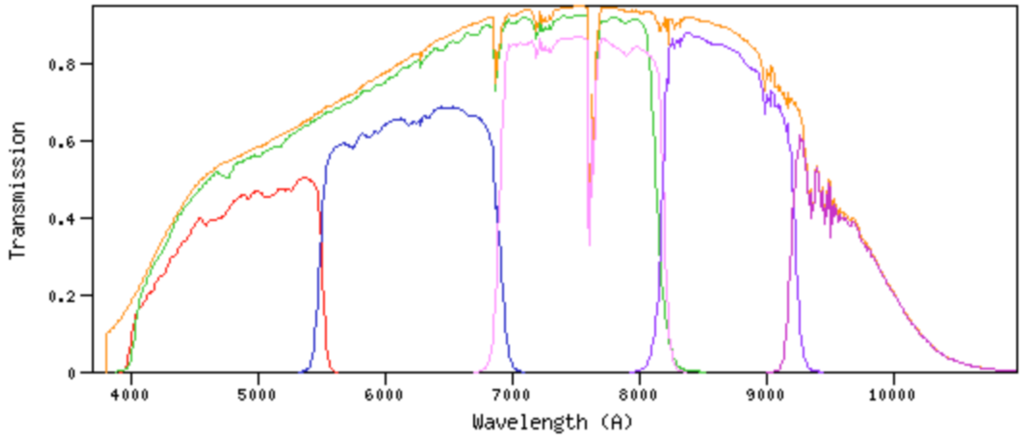
Spectroscopically **confirmed** evolved massive stars in the local group with good NIR photometry (PanSTARRS+Spitzer see e.g. Maravelias+22)

M31 (438 sources)	M33 (449 sources)
RSG – **64%**	RSG – **51%**
BSG – **15%**	BSG – **22 %**
YSG – **13%**	YSG – **19%**
WR – **3%**	WR – **4%**
B[e]SG – **3%**	B[e]SG – **1%**
LBV – **2%**	LBV – **3%**

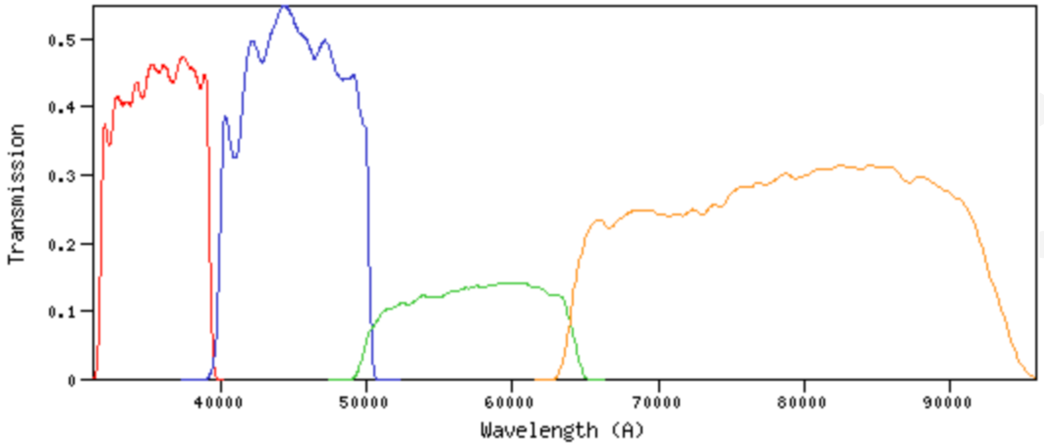
Preprocessing: catalogue crossmatching*, outlier removal, foreground object removal, photometry quality assessment, spectral type mapping*

Building the dataset – features

Pan-STARRS



Spitzer/IRAC



g-r r-i i-z z-y y-[3.6] [3.6]-[4.5] [4.5-5.8] [5.8-8.0]

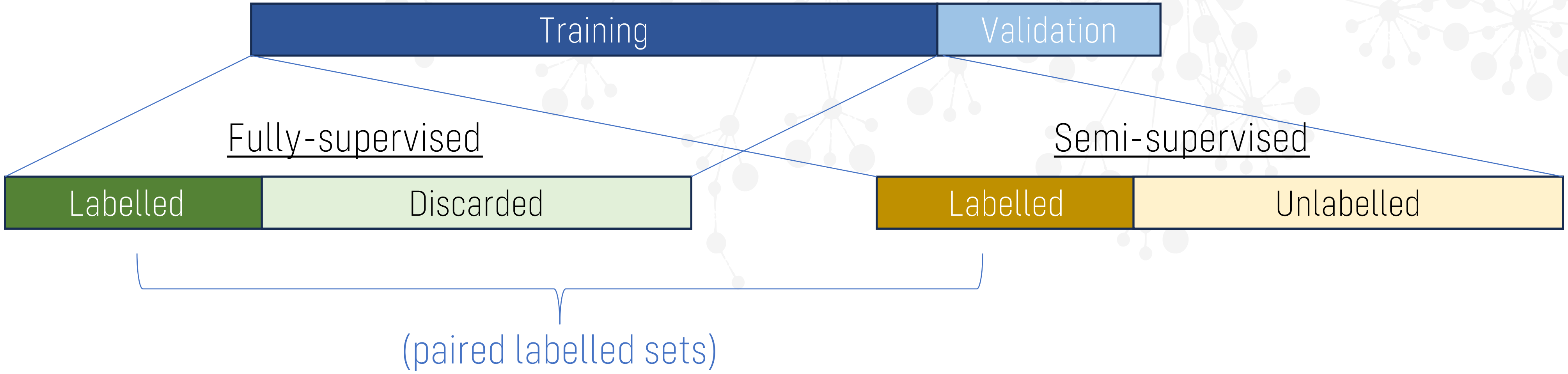
Likely affected by extinction

Aperture photometry on star position (no detection)

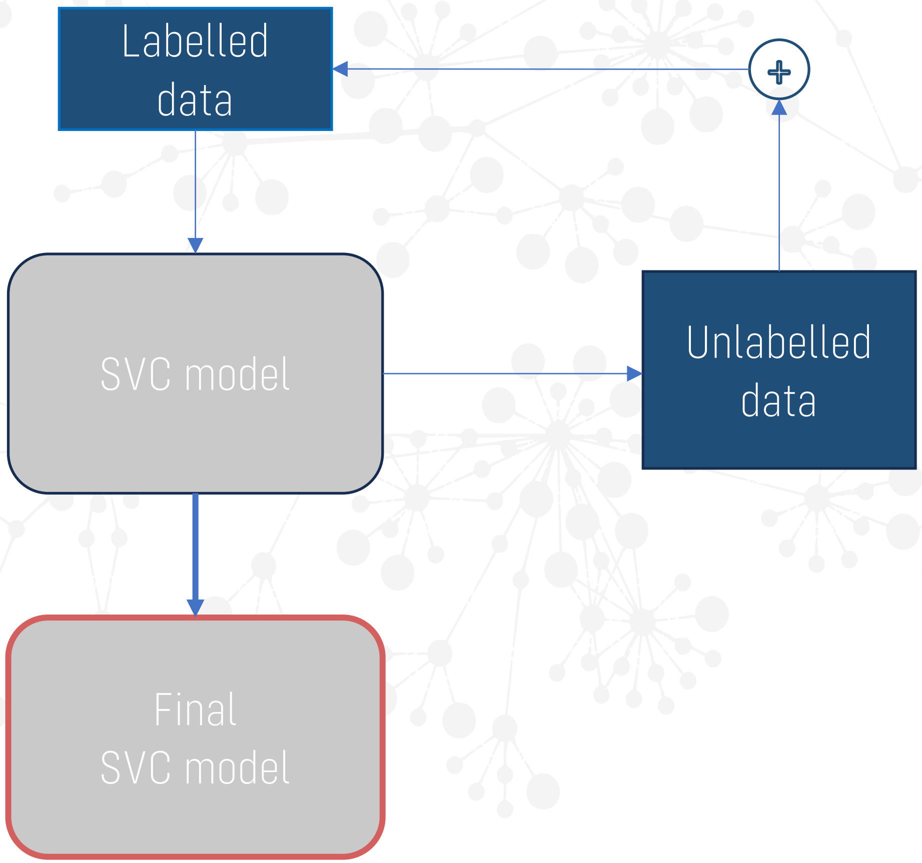
Experimental setup

Benchmark to compare method performance for different % of labelled data.

- Baseline model: supervised SVC ('rbf' kernel)
- Unsupervised methods: self-training SVC; DBSCAN+SVC; S3DB+SVC
- No resampling, no imputation
- K-fold (k=5) cross-validation (M31 data)
- Generalisation test (M33 data)



Semi-supervised method I: self-training



Self training

L labelled data, U unlabelled data

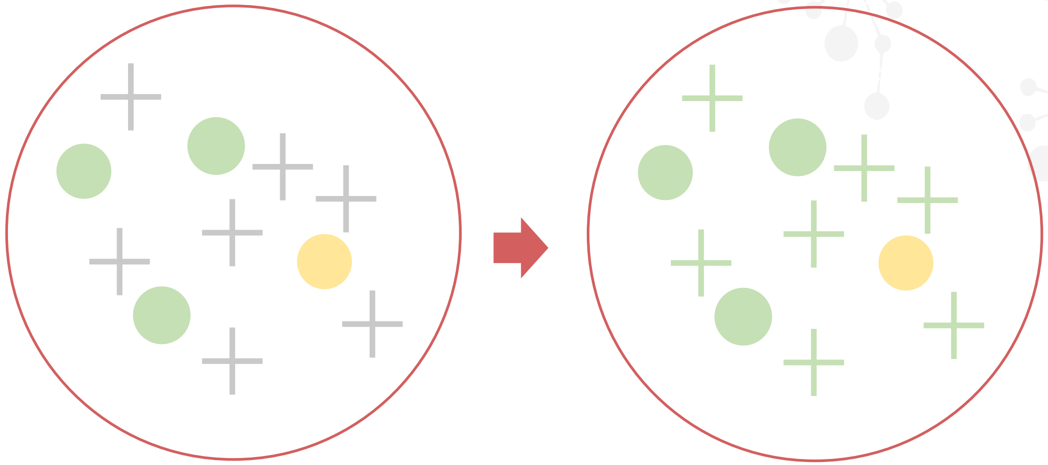
1. Train the classifier (SVC) on L
2. Predict (pseudo-)labels on U.
3. Retrain the classifier on L+U to obtain final model

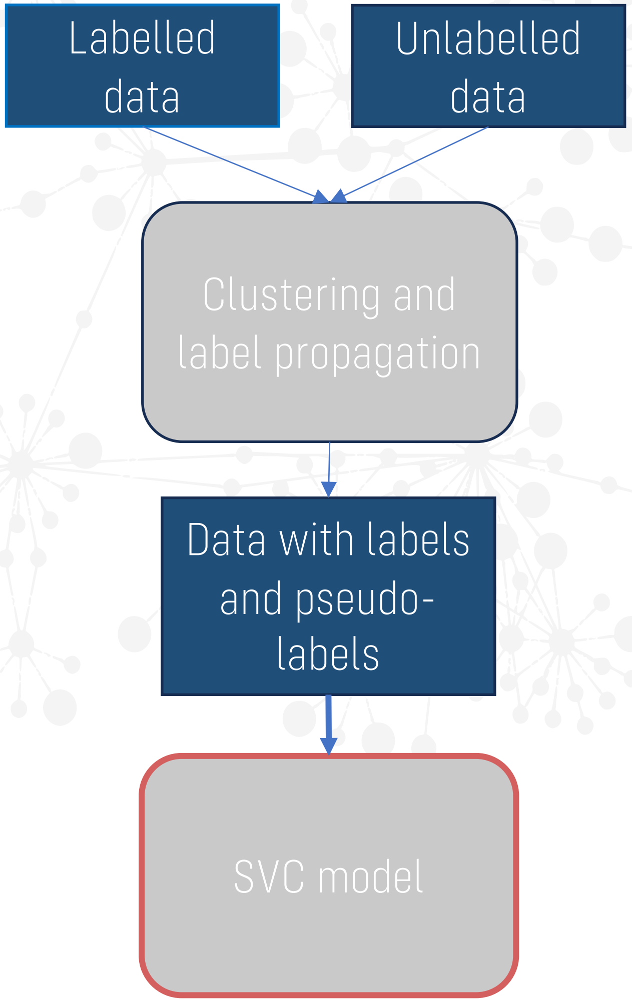
Semi-supervised method II: DBSCAN + SVC

DBSCAN

Density based clustering

1. Cluster L and U together
2. Tune DBSCAN for cluster purity (small clusters)
3. Assign pseudo-labels to U by intra-cluster majority voting
4. Train the classifier on L+U





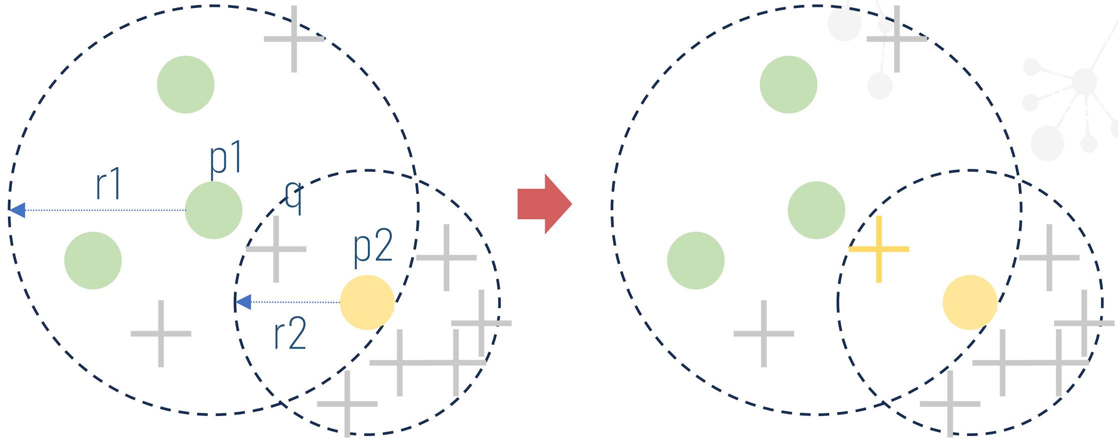
Semi-supervised method III: S^{3}DB+SVC

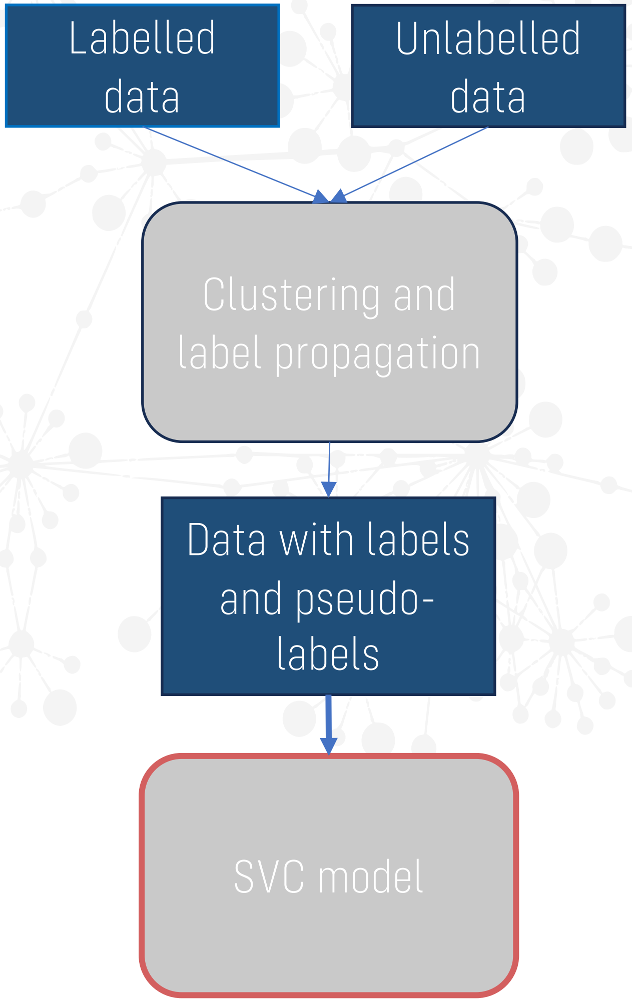
S^{3}DB (semi-supervised seeded density-based)

A semi-supervised version of OPTICS (Peikari+18)

1. Cluster L and U together
2. Tune S^{3}DB and assign pseudo-labels to U by reachability
3. Train the classifier on L+U

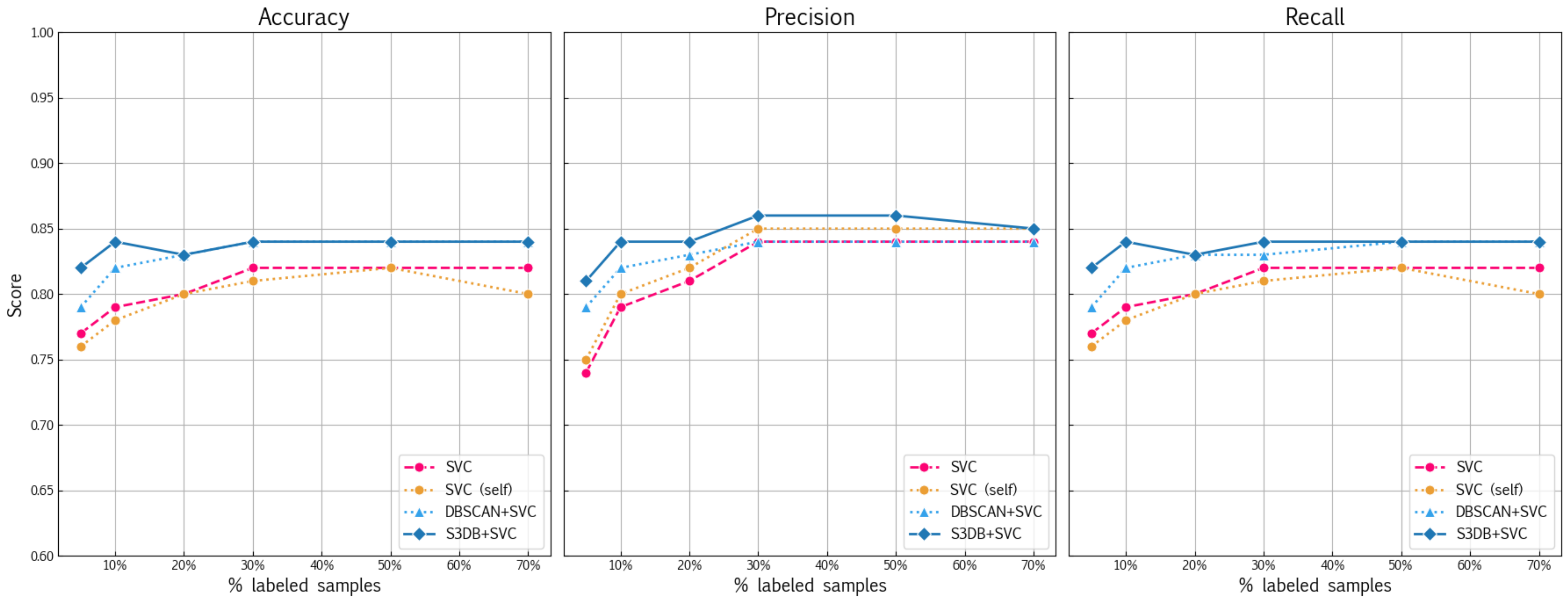
$$d_{ij} = \min(r_i, \|p_i - q_j\|_2)$$





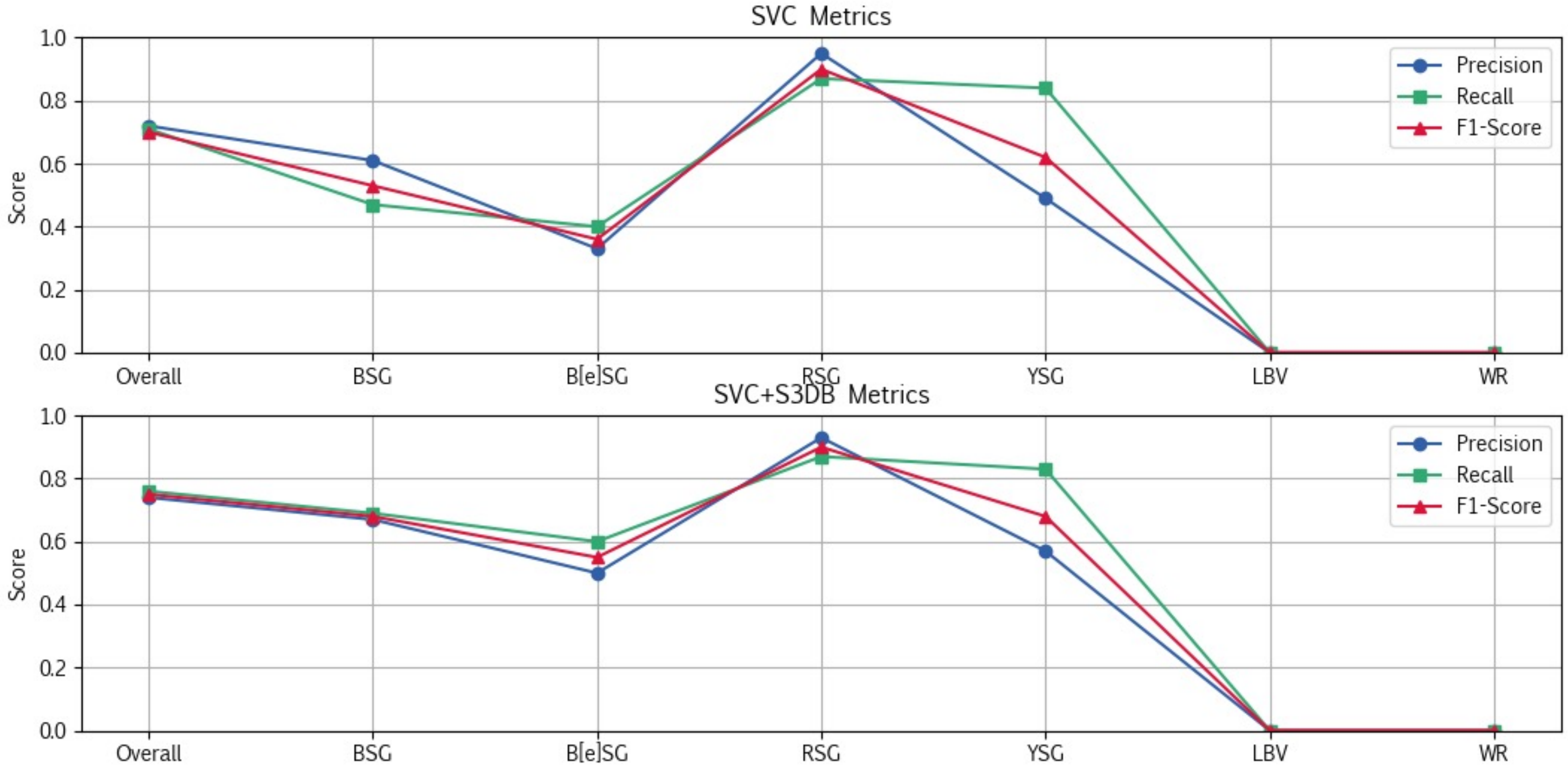
Results

Clustering-based methods improve all the metrics particularly in the low-labelled data regime (▲4-7%)

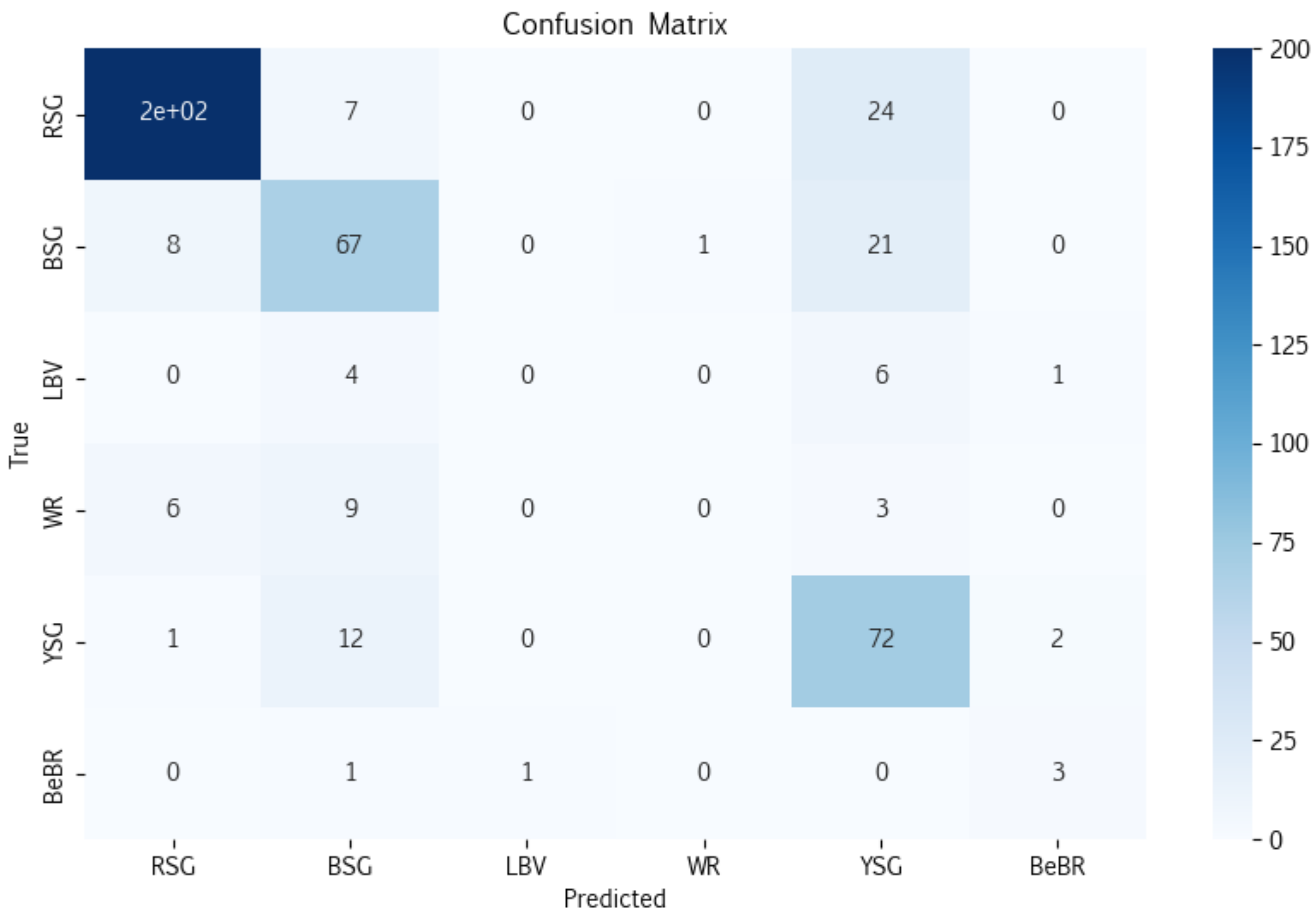


Test on unseen data (M33)

Slightly worse performance (systematics? Extinction?) – still, S^{3}DB method generalises better. Improvement of 5-10% in minority classes



What is the classifier getting wrong?



Conclusions and next steps / WIP

- First results promising: S^3DB offers good performance with fewer labelled data points, margin for improvement

- What now?
  - Investigate dataset dependency
  - Investigate classifier dependency
  - Investigate feature sensitivity
  - Better data > better models – clean outliers, include new features
  - Application to Galactic objects (distance influence!)

Take home message

Unlabelled data contains valuable information that can improve classification performance even in small/imbalanced datasets

Thanks for your attention!

Questions?

cristobal.bordiu@inaf.it

This research is funded by INAF mini-grant *"A needle in a haystack: a semi-supervised search of evolved massive stars"* under F. 0. 1.05.12.04.02