A needle in a haystack

A semi-supervised search for evolved stars with multiwavelength survey data and VO tools

YEAR #1 Review

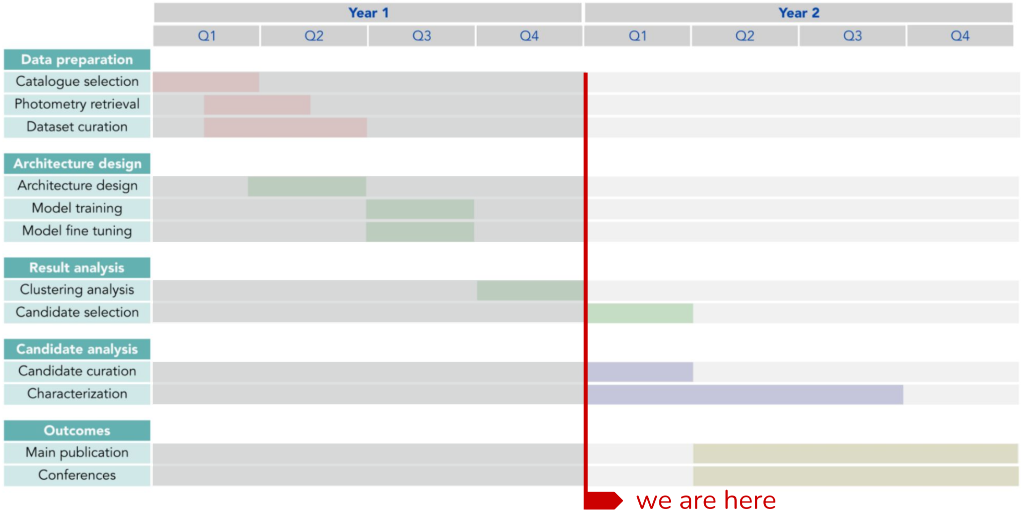
P.I : Cristobal Bordiu

(INAF Osservatorio Astrofisico di Catania)

Project scope & status (Nov' 23)

Application of **semi-supervised learning techniques** to find an optimal latent representation that allows for **clustering and labelling unknown objects** based on **multiwavelength photometry** (using a subset of **known evolved stars** as reference) aka **"Cluster-then-label"** approach.

Tentative project roadmap as submitted in the Minigrant proposal



	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Data preparation								
Catalogue selection	🟥							
Photometry retrieval	🟥	🟥						
Dataset curation	🟥	🟥						
Architecture design								
Architecture design		🟩						
Model training			🟩					
Model fine tuning			🟩					
Result analysis								
Clustering analysis				🟩				
Candidate selection					🟩			
Candidate analysis								
Candidate curation					🟪			
Characterization					🟪	🟪	🟪	
Outcomes								
Main publication						🟨	🟨	🟨
Conferences						🟨	🟨	🟨

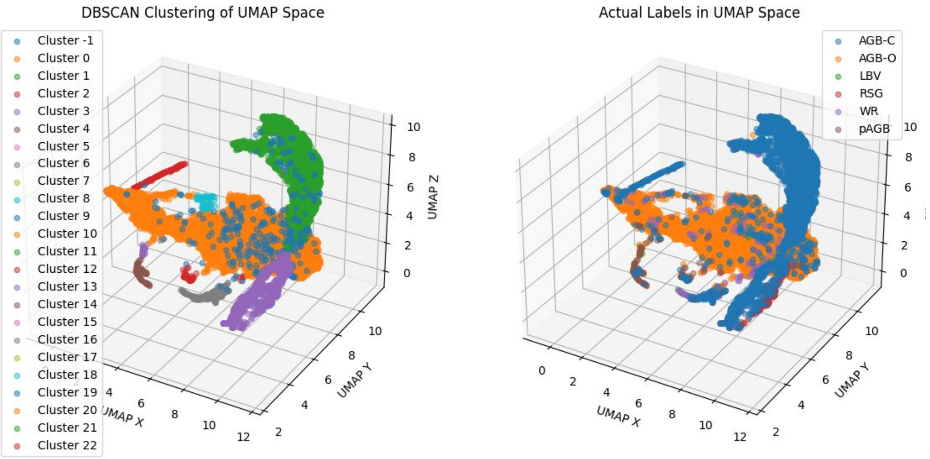
we are here

Highlights

- Data collection complete. About **11.5k sources** with complete photometry (Gaia, 2MASS, WISE) belonging to **six classes**: O-rich AGB, C-rich AGB, post-AGB, RSG, LBV, WR.

- Several **representation learning** methods tested: autoencoders (different architectures), **TSNE**, **UMAP**. **UMAP** yields the best results with a highly structured feature space that allows for the best clustering.

- **Density-based clustering** is able to recover clusters with a high degree of **purity**, required for tagging unknown objects in the cluster-then-label approach.


DBSCAN Clustering of UMAP Space

Actual Labels in UMAP Space

Left: HDBSCAN clustering of the UMAP representation of the 11.5 k sources. Right: Ground truth (real labels). Note that some source groups are split in multiple clusters, but those keep a high level of purity.

Next steps & prospects

- So far, no critical issues found

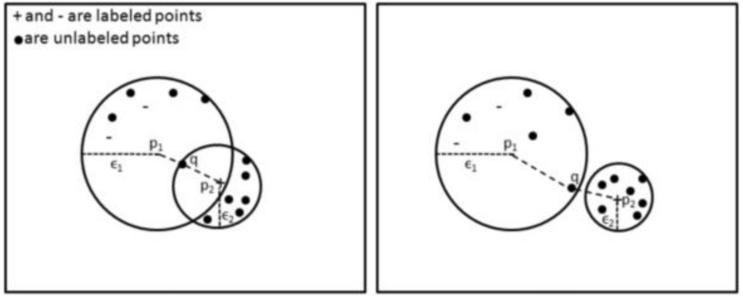
- Some preliminary results shown at **EAS 2023 (Krakow)**

- Next steps:

  1. Collect photometry of **unknown sources**. Reference dataset in Dorn-Wallenstein+2021 – Also search SIMBAD [in progress]

  2. Project known and unknown sources, apply clustering and **propagate labels**

  3. Retrieve **VO information** of the best candidates to assess classification



Cluster-then-label scheme