A Study Of YSOs in the Perseus Molecular Cloud using ML Techniques

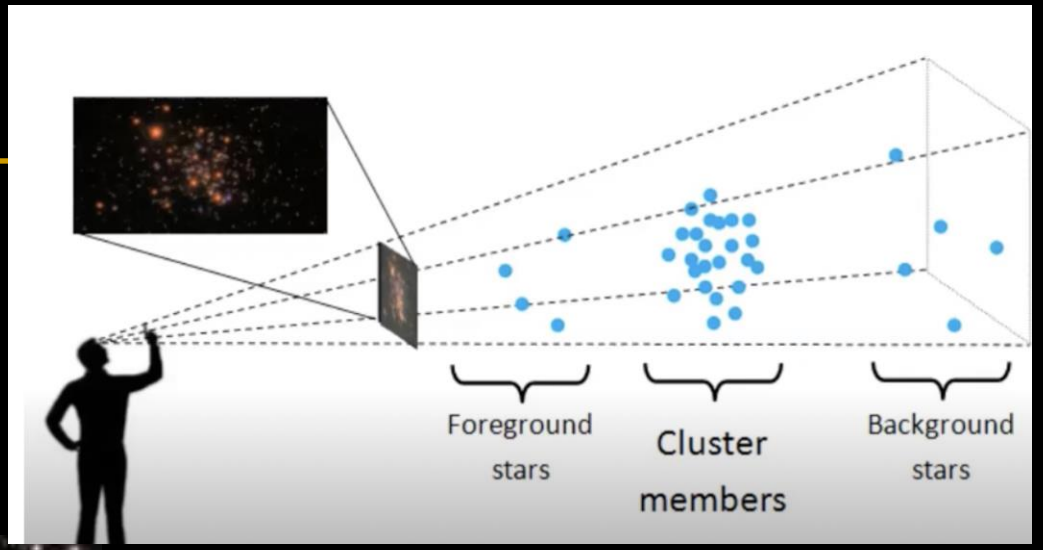
Priya Hasan
Asst Professor in Physics
Maulana Azad National Urdu University
Hyderabad
priya.hasan@gmail.com

AI/ML APPLICATIONS IN ASTRONOMY & ASTROPHYSICS
JANUARY 6 - 10, 2025,
IUCAA, INDIA

With Md Mahmudonobe, Mudasir Raja, Md Saifuddin, S N Hasan

Maulana Azad National Urdu University Hyderabad,500032
Wayne State University, USA.

The Membership Problem…



Membership?
Photometric
Kinematic
Spectroscopic

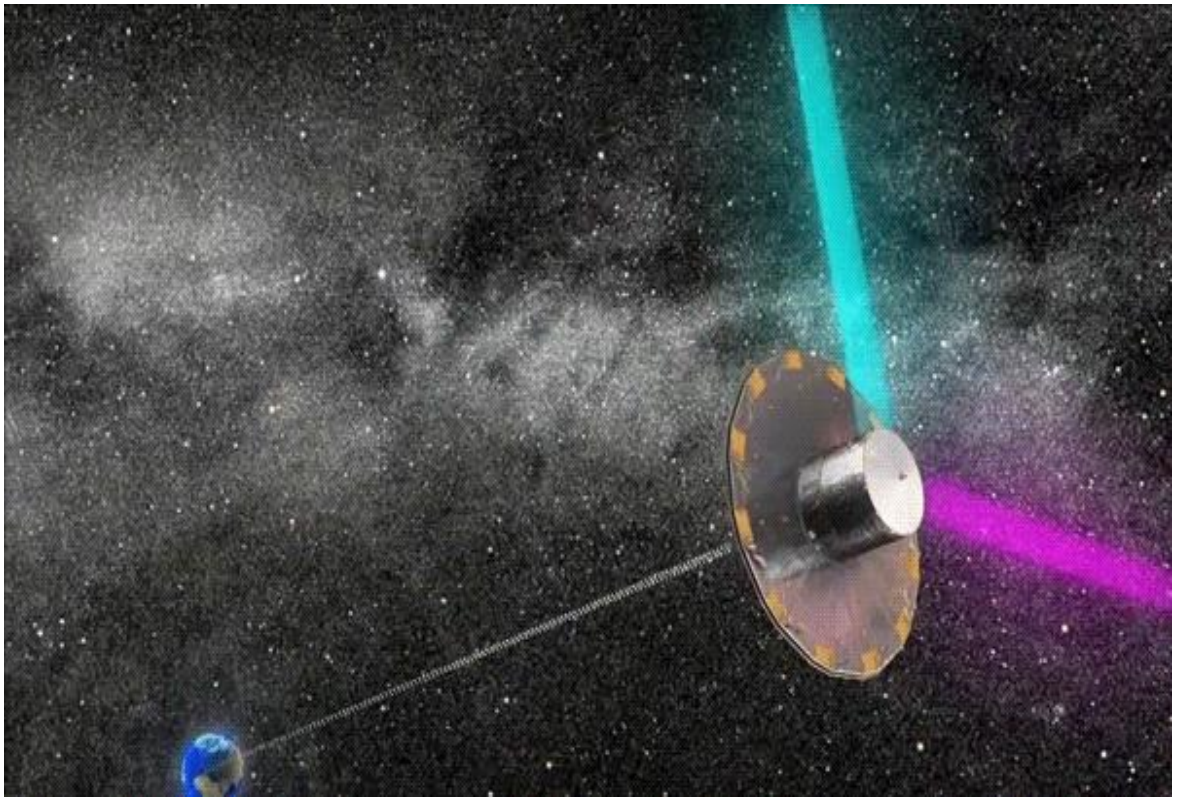
Wish list

High- precision astrometric data (positions, parallax, and proper motions) are of great importance to the studies of open clusters, because more accurate cluster members and astrophysical parameters can be obtained. YSOs identified.

GAIA: 6D revolution

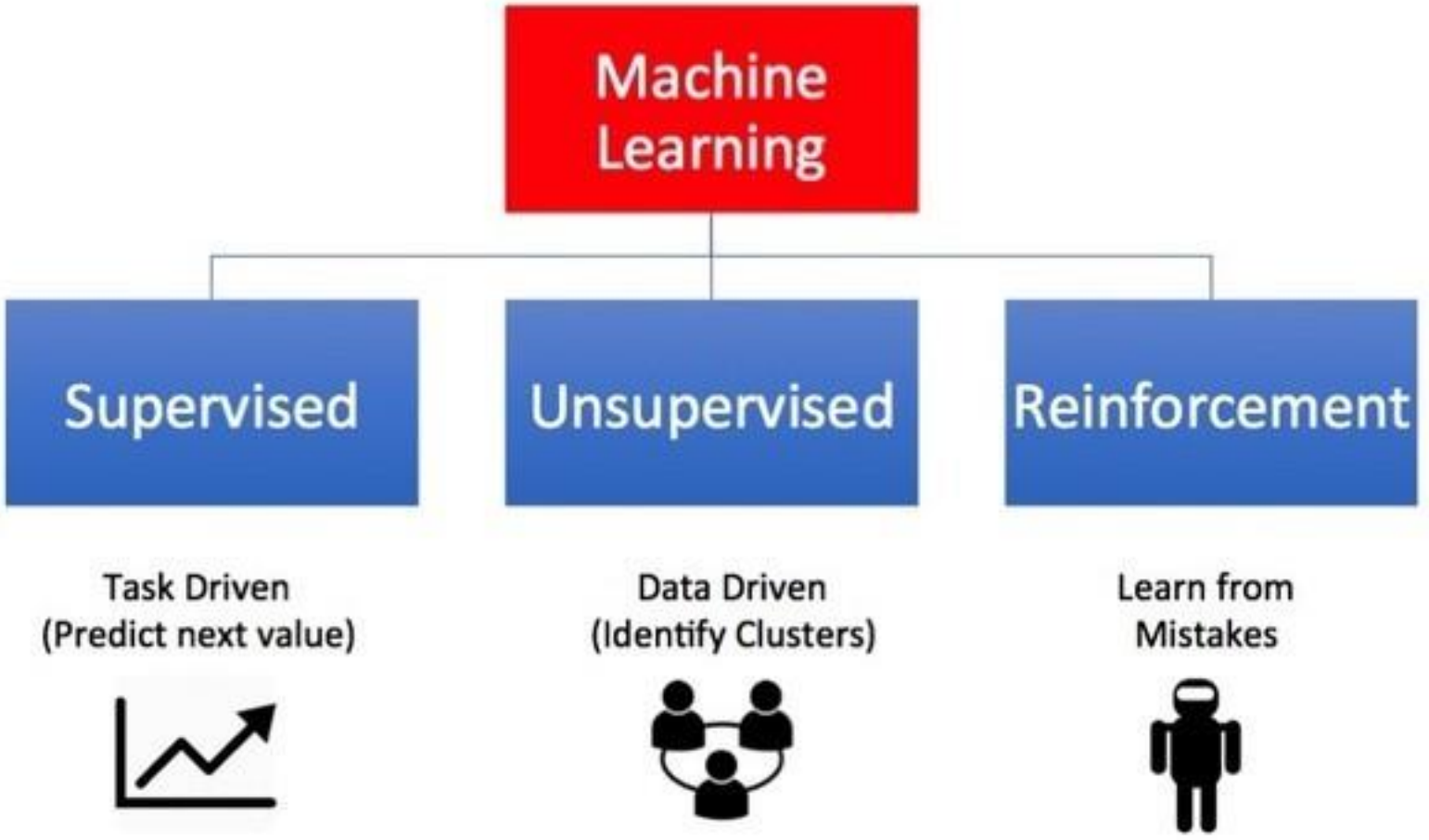
RA, Dec, parallax, RV, pmra, pmdec

Two identical, three-mirror anastigmatic (TMA) telescopes, with apertures of 1.45 m × 0.50 m pointing in directions separated by the basic angle (Γ = 106◦.5)
Accuracy of 24 microarcsec= 42 kpc,
0.06arcsec pixels



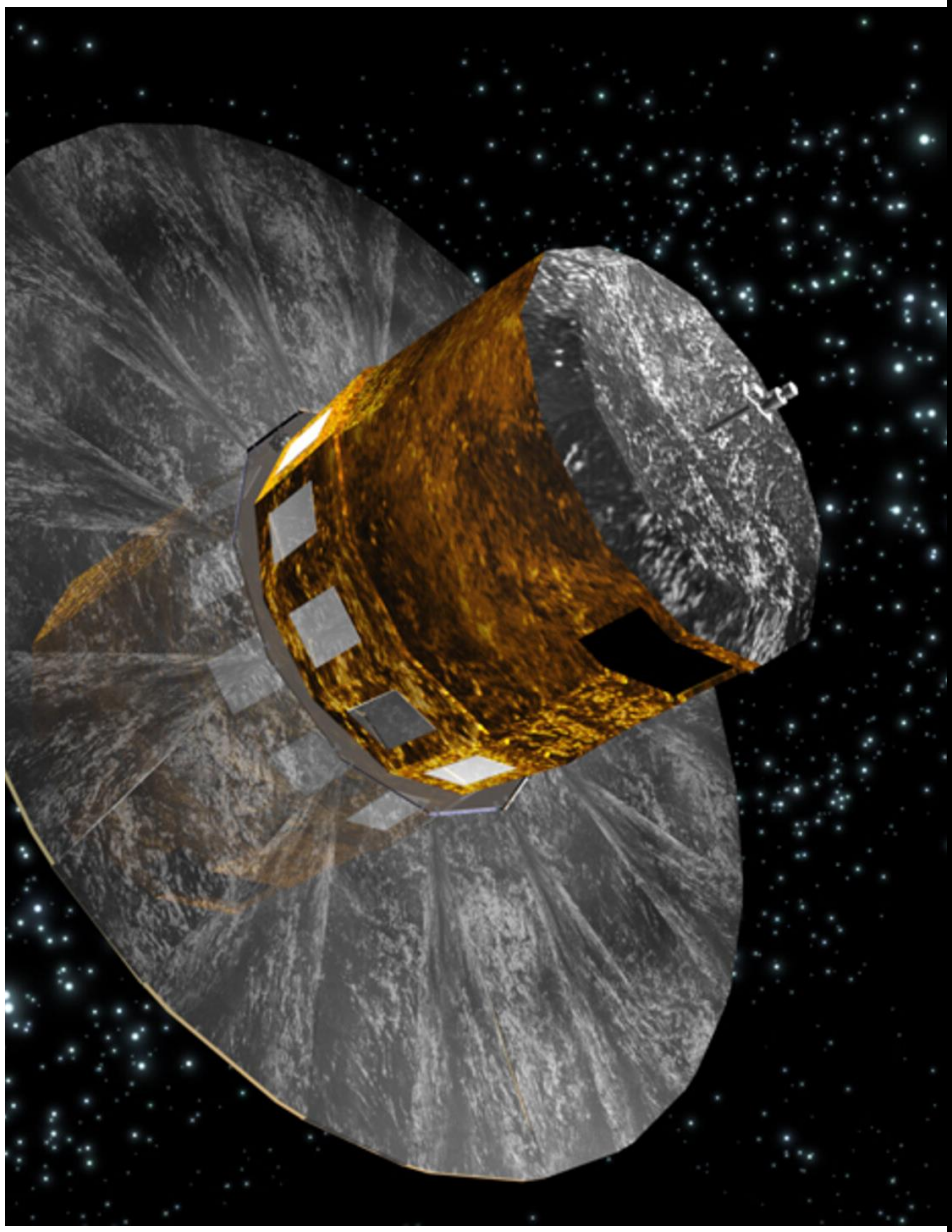
Galactic Archealogy!!! Imagine!!!

Types of Machine Learning

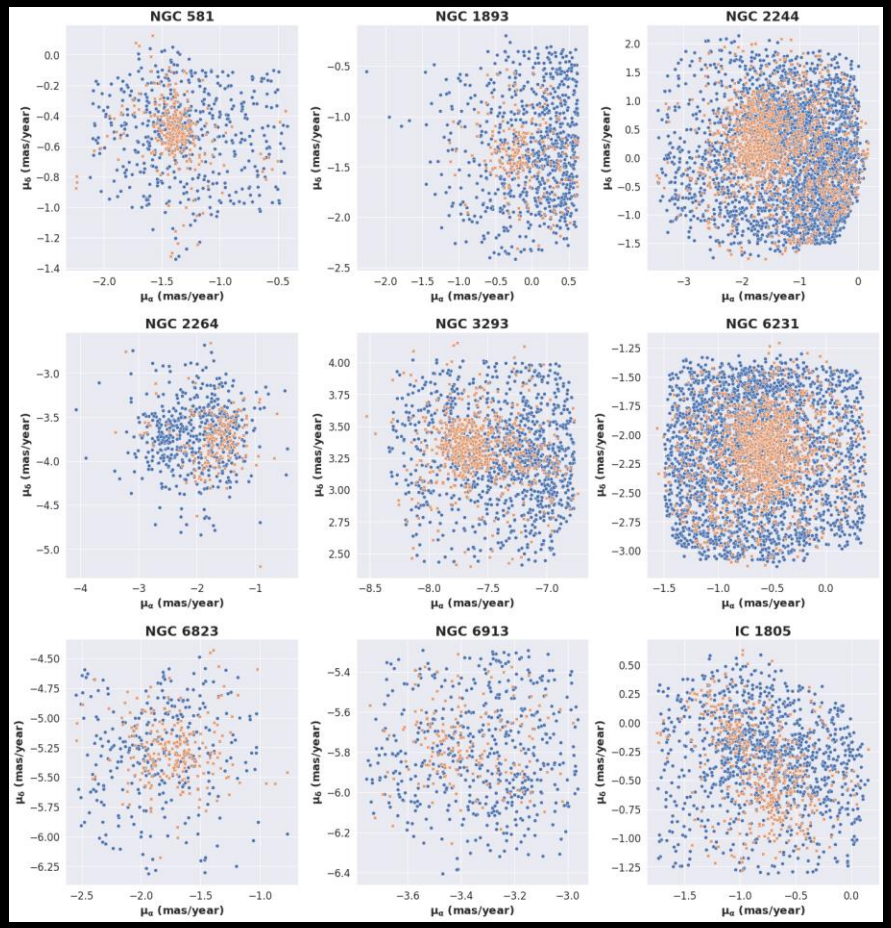


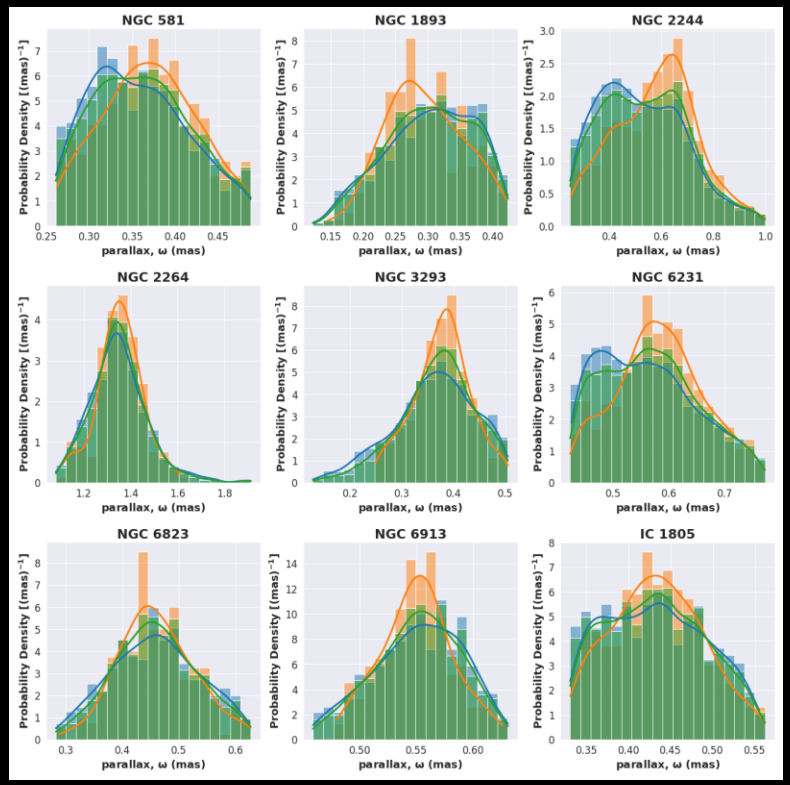
Membership of Stars in Open Clusters using Random Forest with Gaia Data

- *GAIA DR2 has a very strong influence on the membership of star clusters. This is one of the most crucial parameters in studies of star clusters. In the present study, we use membership data from Cantat-Gaudin et al(2018) based on GAIA DR2 as a training set.*
- *Random Forest (RF), which is a supervised classification method, is applied to the Gaia DR2 data in this paper. We use the results from Cantat et al as our training data to find new members in a sample of nine open clusters (NGC 581, NGC 1893, IC 1805, NGC 6231, NGC 6823, NGC 3293, NGC 6913, NGC 2264, NGC 2244).*
- *The sample has clusters with ages ranging from 1.3–20 Myr, at galactocentric distance R_GC ranging from 7.3–14.5 kpc and at varying galactic latitudes l and longitudes b*



Proper Motion & Parallax plots





Validation

Divide the training data in a ratio of 30 : 70.
We made a grid with the possible range of values for important model parameters (i.e. number of trees in RF, maximum depth of a tree, minimum samples needed for a split, minimum sample for a leaf node etc)
Then we applied a randomized search 5-fold cross validation in the train subset with 100 iteration which in total builds 500 models with randomly chosen parameters from the grid and select the model which resulted in maximum precision.

Table 3: Prediction from the Random Forest Model

Cluster	Radius deg	Members before filter	Members after filter	Non-Member radius deg	Search radius deg	New Members	Precision %	Ratio of new to CG
NGC 581	0.17	306	290	0.7-0.8	0.34	525	86	1.81
NGC 1893	0.41	494	218	1.0-1.1	0.82	774	93	3.55
NGC 2244	0.67	1701	1192	1.4-1.5	1.33	3043	88	2.55
NGC 2264	0.19	186	179	1.0-1.1	0.60	514	99	2.87
NGC 3293	0.20	657	617	0.7-0.8	0.40	1089	94	1.76
NGC 6231	0.47	1580	1354	0.95-1.0	0.94	2710	92	2.00
NGC 6823	0.2	236	220	0.7-0.8	0.40	304	93	1.38
NGC 6913	0.3	170	170	0.7-0.8	0.60	536	95	3.15
IC 1805	0.33	456	430	0.7-0.8	0.66	1104	90	2.57

Eur. Phys. J. Spec. Top.
https://doi.org/10.1140/epjs/s11734-021-00205-x

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

Membership of stars in open clusters using random forest with gaia data

Md Mahmudunnobe[1], Priya Hasan[2,a], Mudasir Raja[2], and S. N. Hasan[2]

[1] Minerva Schools at KGI, San Francisco, CA 94103, USA
[2] Maulana Azad National Urdu University, Gachibowli, Hyderabad 500 032, India

Members increased by 2--3 times. Improves accuracy in determining various parameters of a star cluster ranging from distance, extinction and mass function.

The sizes revised

Likely cluster members, escaped members

Find sub-structure in velocity space as well as spatial distribution of the cluster unresolved binary sequences (NGC~6231) as well as all other possible non main-sequence members of the cluster.

Supervised Learning ?

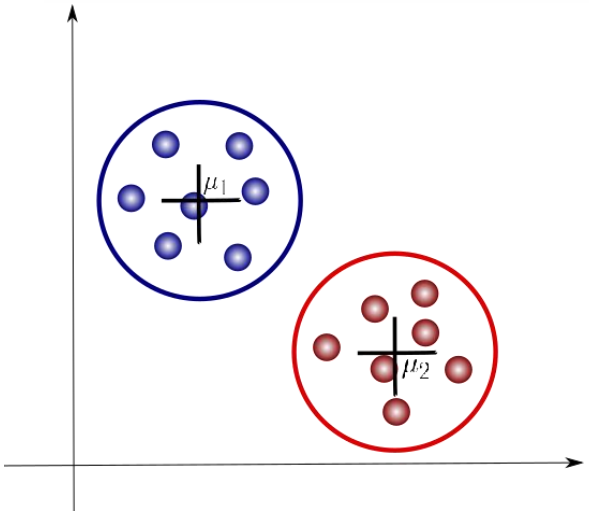
- Supervised methods (SM) (where we NEED good training data)
- Pro: It can perform better or give good accuracy or prediction even with a high number of data
- Cons: Its accuracy depends on how good the training set

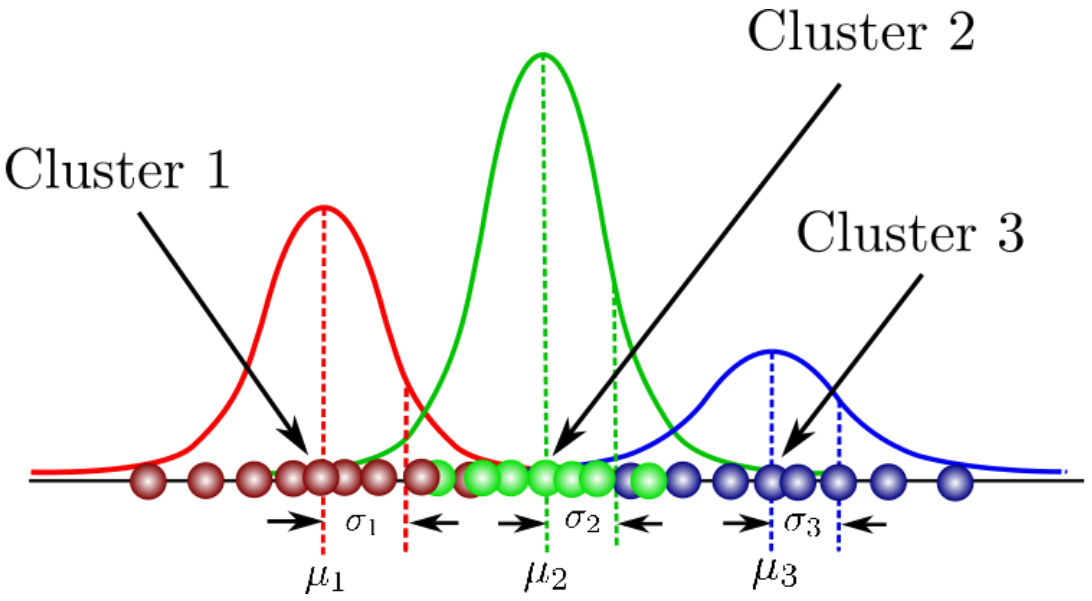
Unsupervised Learning

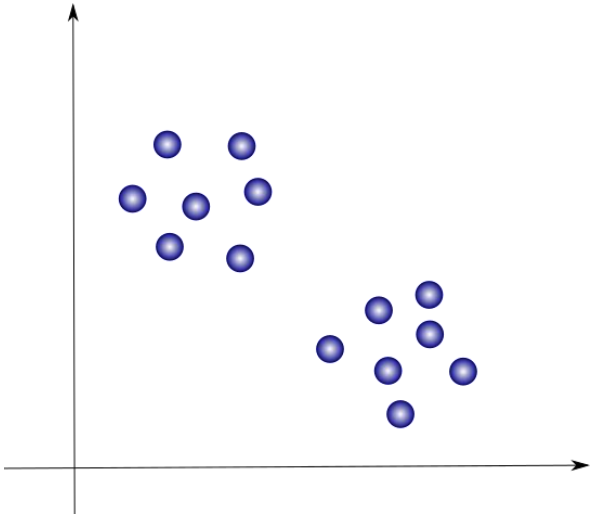
- Unsupervised method (UM)
- Which UM is better? Is there any single UM which works well for all or does it depend on the cluster?
- Which SM is better? Is there any single UM which works well for all or does it depend on the cluster?

Guassian Mixture Modelling

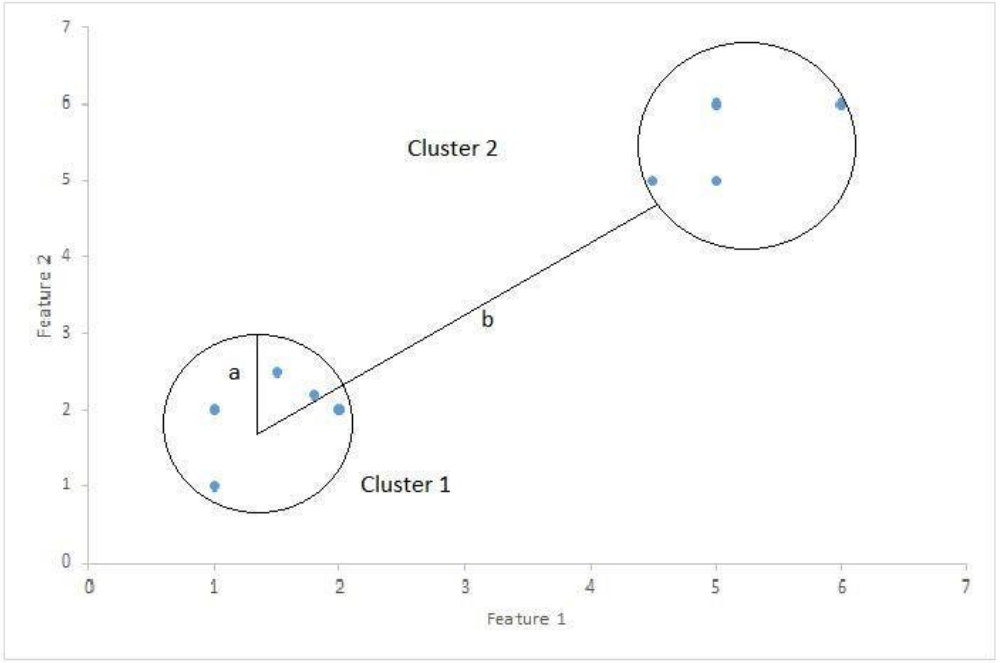
μ1 and μ2 are the centroids of each cluster







Silhouette Score



$$s = \frac{b - a}{\max(a, b)}$$

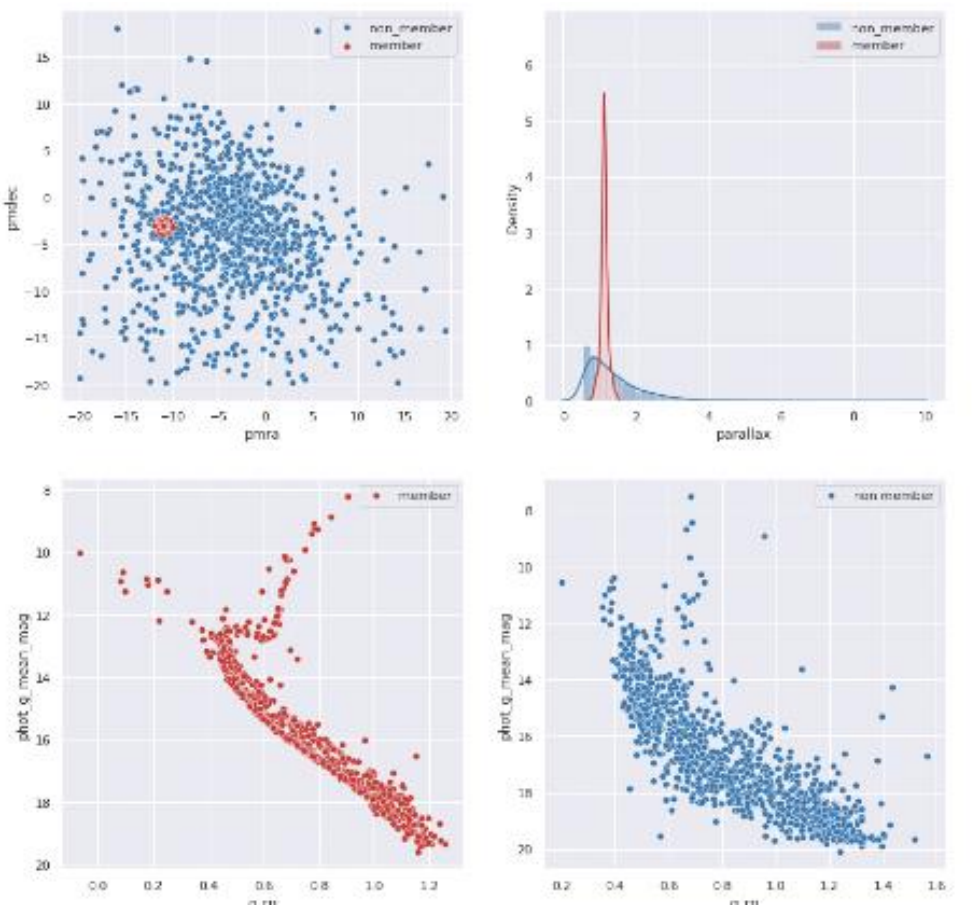
Modified Silhouette Score

$$MSS = \frac{1}{k} \sum_{i=1}^{k} \frac{(\sigma_{i,field} - \sigma_{i,member})}{\max(\sigma_{i,field}, \sigma_{i,member})}$$

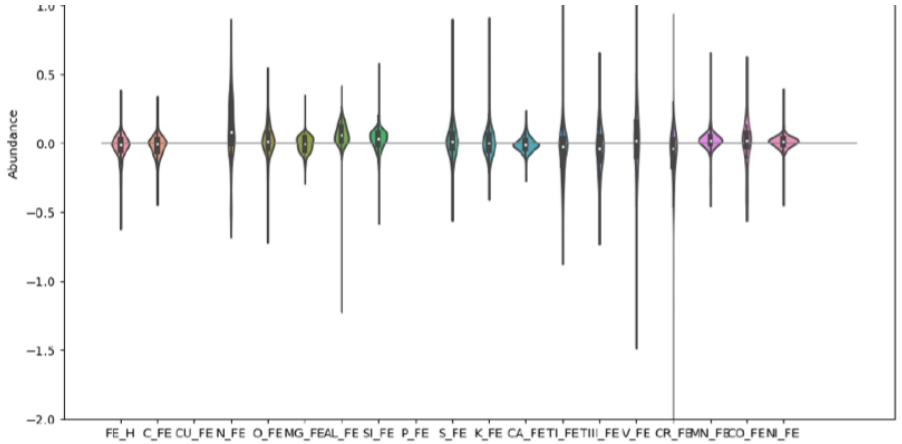
Using GMM model in Open Cluster Membership

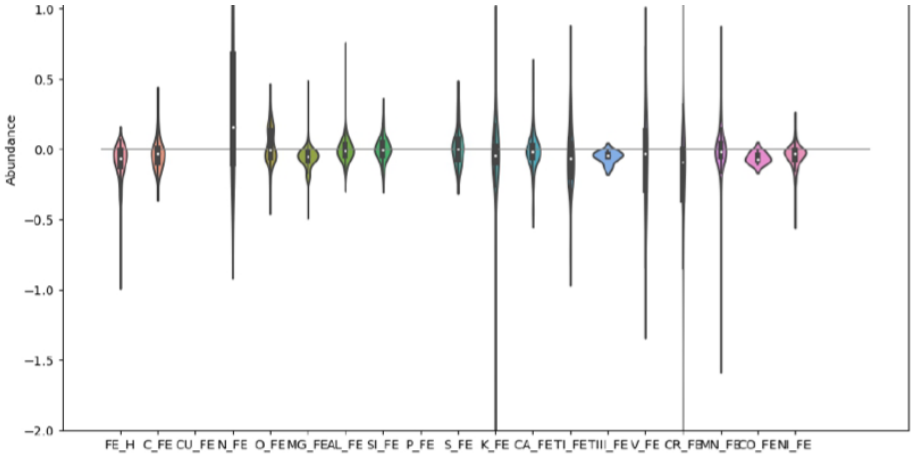
Md Mahmudunnobe, Priya Hasan, Mudasir Raja, S N Hasan 2024, Astronomy and Computing

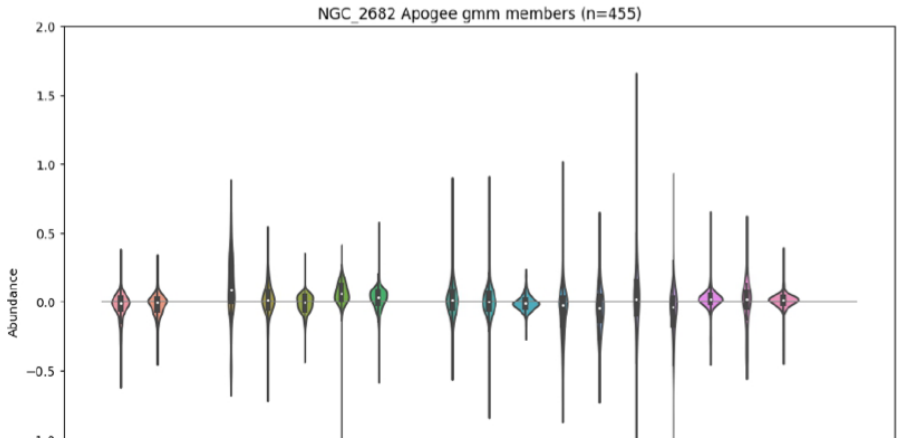
Cluster	MSS	Member GMM	Member Cantat	Ratio GMM/Cantat
NGC 2682	0.94	1390	691	2.01
NGC 752	0.93	232	240	0.97
IC 4651	0.90	875	854	1.02
NGC 2539	0.90	560	518	0.93
NGC 2099	0.90	1607	1710	0.94
NGC 581	0.87	458	152	3.01
NGC 6823	0.84	397	158	2.51
NGC 2243	0.84	484	515	0.94
IC 1805	0.81	495	136	3.63
NGC 7142	0.79	430	401	1.07
NGC 6791	0.79	1106	1654	0.67

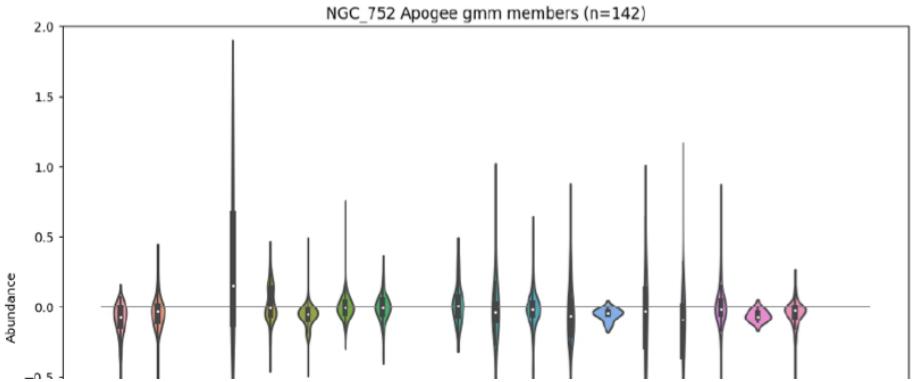


Spectroscopic Data: APOGEE and GALAH









ASteCA: M67

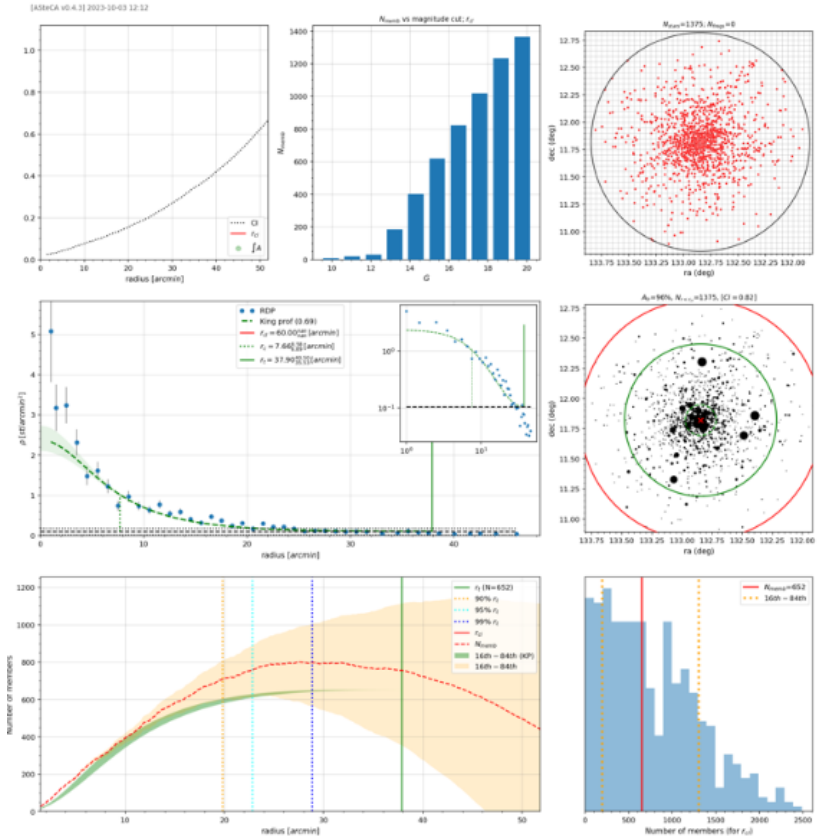


FIGURE 3.19: ASteCA plots of NGC 2682

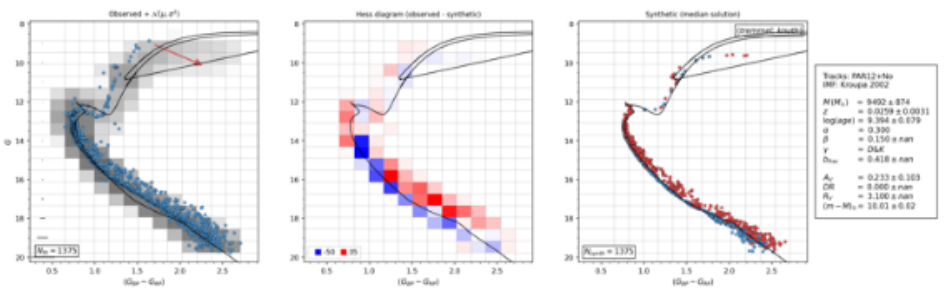
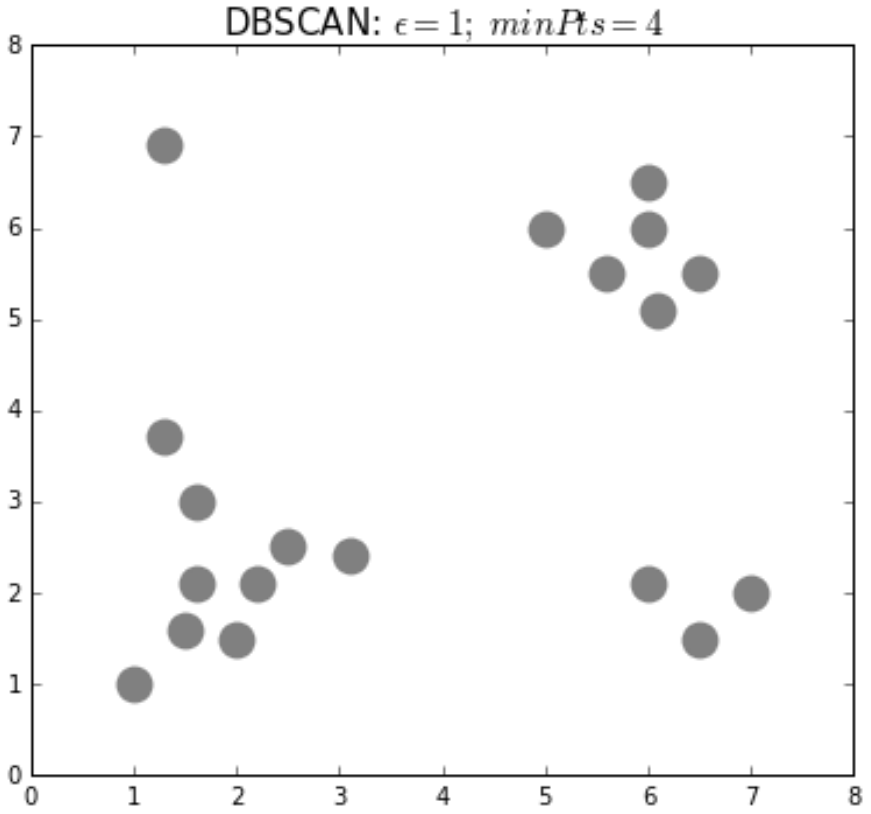
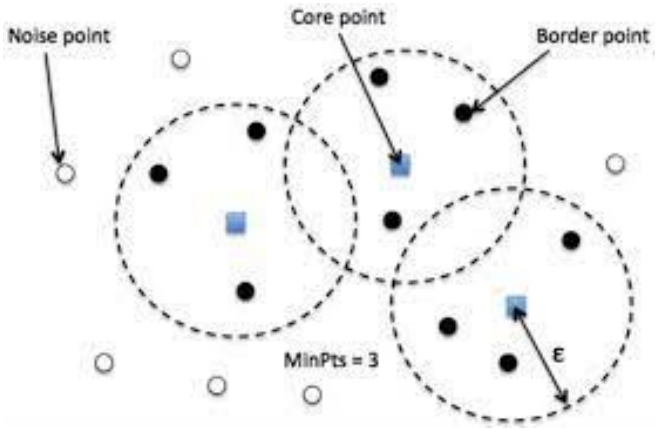


FIGURE 3.20: ASteCA CMD plots of NGC 2682

DBSCAN

Density-based spatial clustering of applications with noise





Membership determination in open clusters using DBSCAN Clustering Algorithm

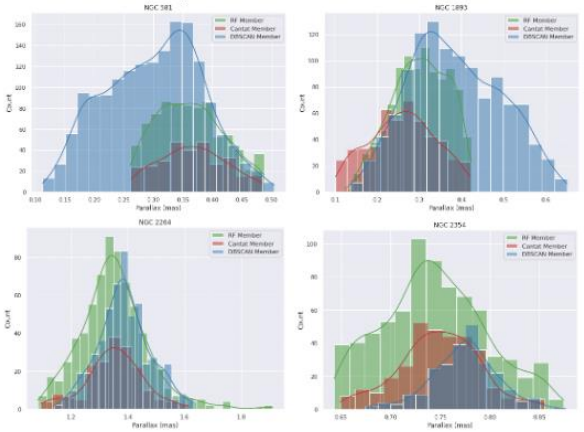
Mudasir Raja[1], Md Mahmudunnobe[2], Priya Hasan[1] and S N Hasan[1]

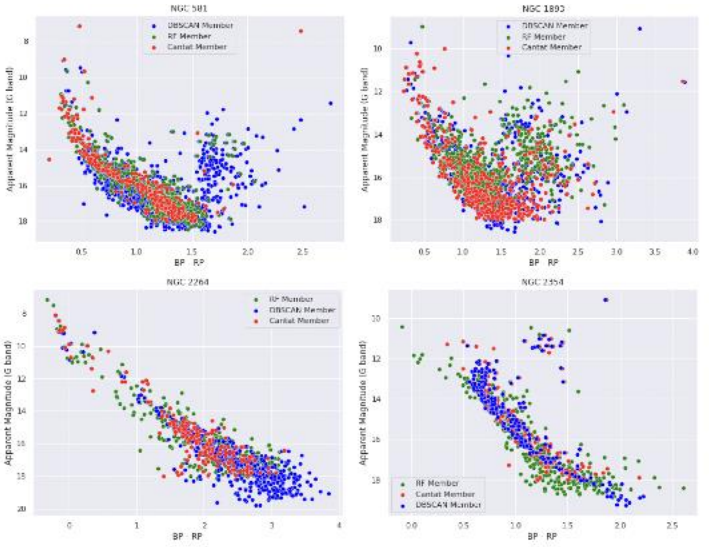
1. Maulana Azad National Urdu University Hyderabad,500032 2. Minerva University, California, USA

Table 2: Results from DBSCAN for our sample and Comparison with the results from RF

Cluster	RF Members	DBSCAN Members	Ratio of DBSCAN to RF	Parallax	Distance
					pc
NGC 581	815	1674	2.05	0.30 ± 0.08	3333
NGC 1893	992	1144	1.15	0.37±0.10	2702
NGC 2264	693	543	0.78	1.38±0.09	724
NGC 2354	747	244	0.32	0.77±0.03	1298

Figure 2: Parallax distribution of the four clusters

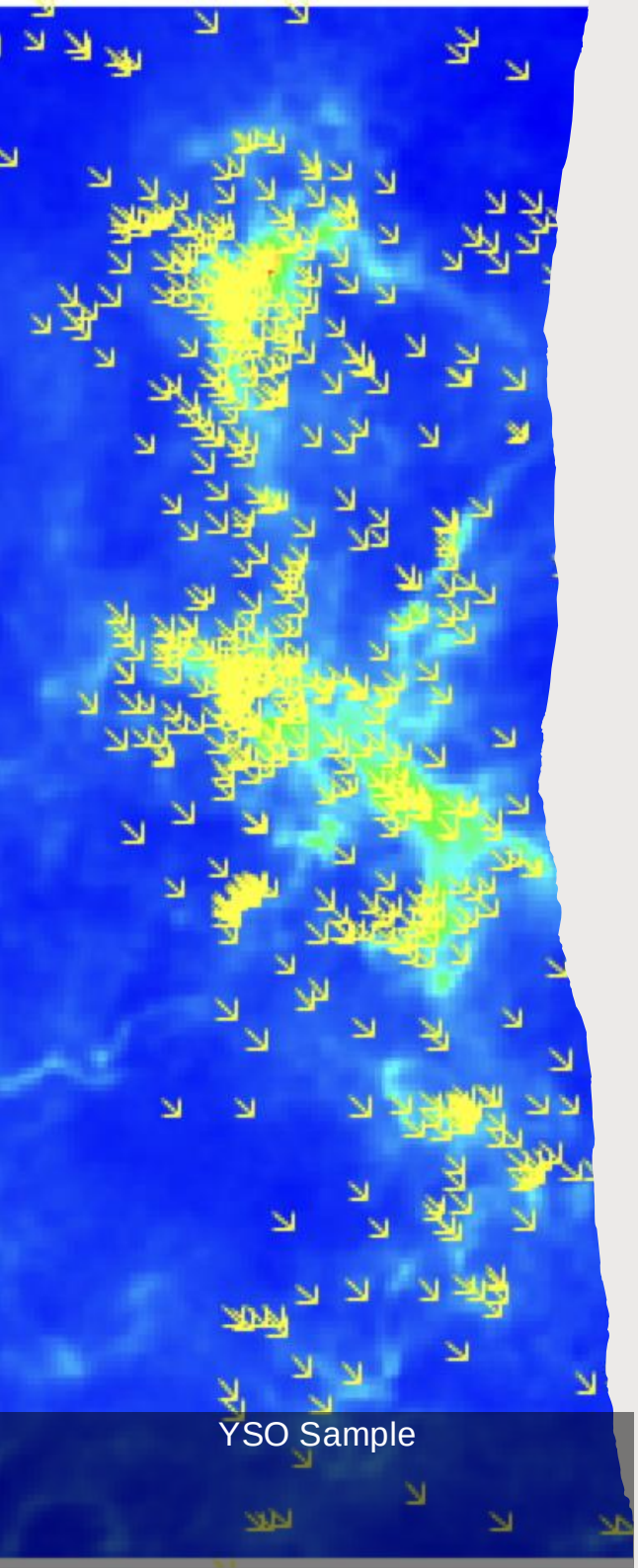




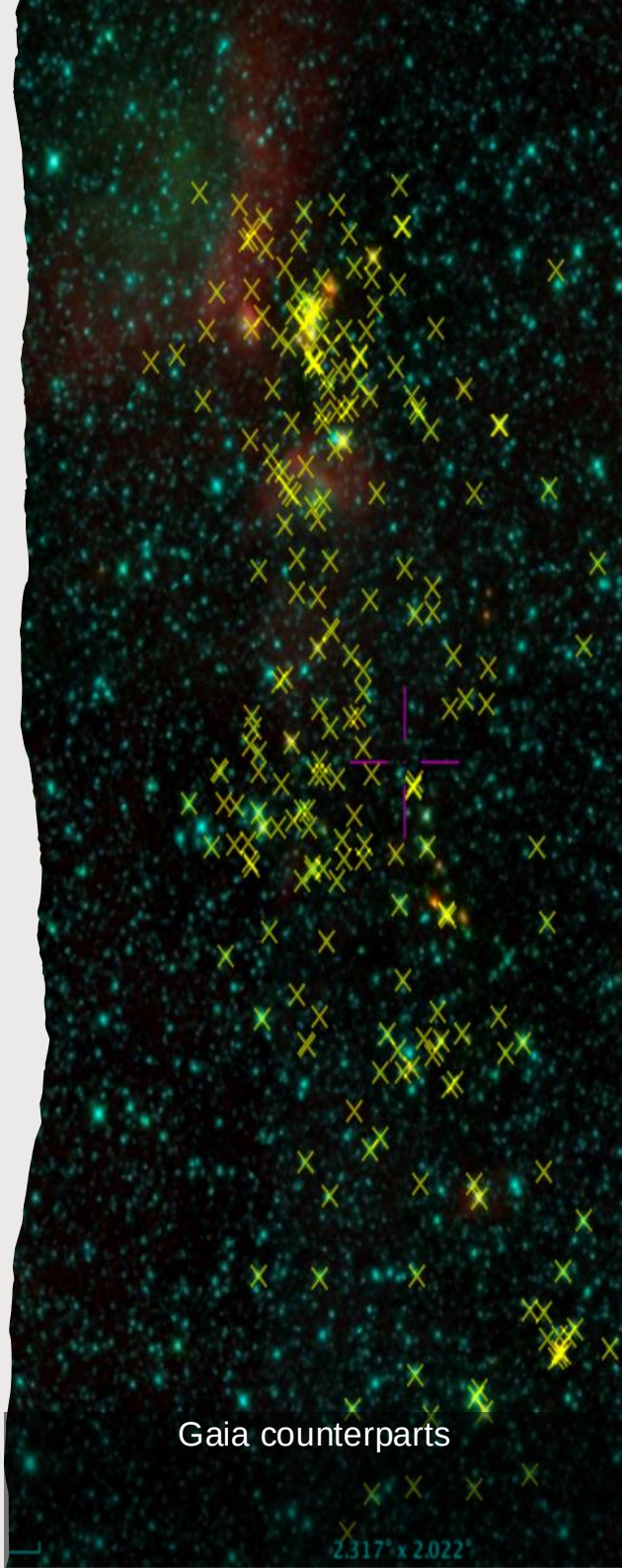
Mudasir Raja , Md Mahmudunnobe, Md Saifuddin, Priya Hasan, , S N Hasan, 2024

Kinematics and Structure in Serpens

- Use available YSO catalogs (Xray, IR….)
- Find parameters
- Download Gaia data within parameters
- DBScan to find members
- 2-3 times increase
- Repeat with OPTICS, HDBScan



YSO Sample



Gaia counterparts

J. Astrophys. Astr. (2023) 44:41
https://doi.org/10.1007/s12036-023-09945-9

STAR FORMATION

Enhanced YSO population in Serpens

PRIYA HASAN[1,*], MUDASIR RAJA[1], Md. SAIFUDDIN[1] and S. N. HASAN[2]

We compiled a sample of YSOs using IR, Xray, data

Matched with Gaia members (87)

Extracted sources with 2d radius

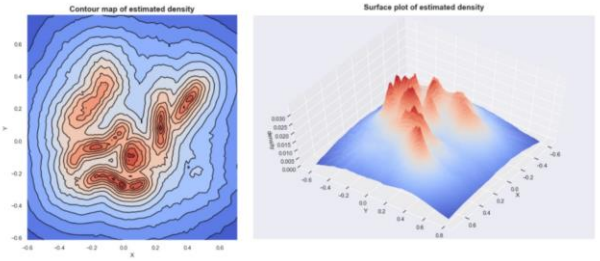
Clustering XYZ, pmRA, pmDEC with DBSCAN, OPTICS, HDBSCAN

Found 822 common YSO members in the region.

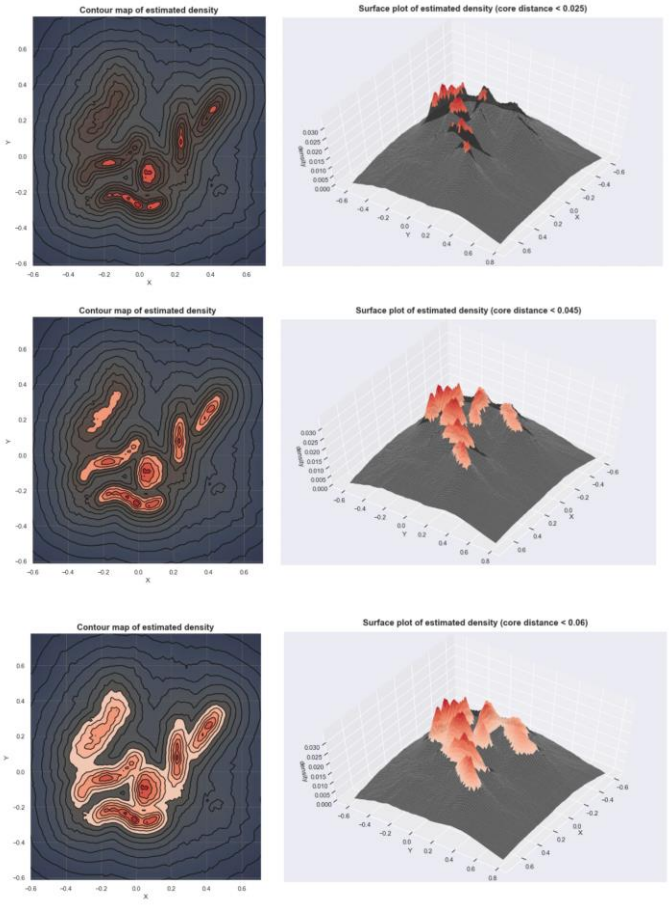
Plotted on extinction map

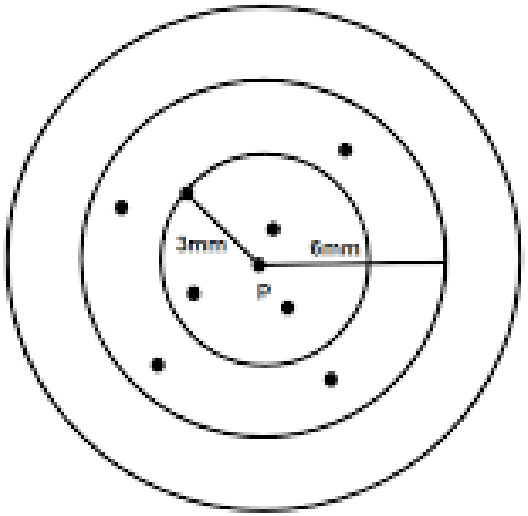
Matched with 2MASS, WISE to classify

How to select clusters?



Imagine islands on the ocean, where the sea level is the threshold and the different islands are your clusters. The land below the sea level is noise. As the sea level goes down, new islands appear and some islands combine to form bigger islands.

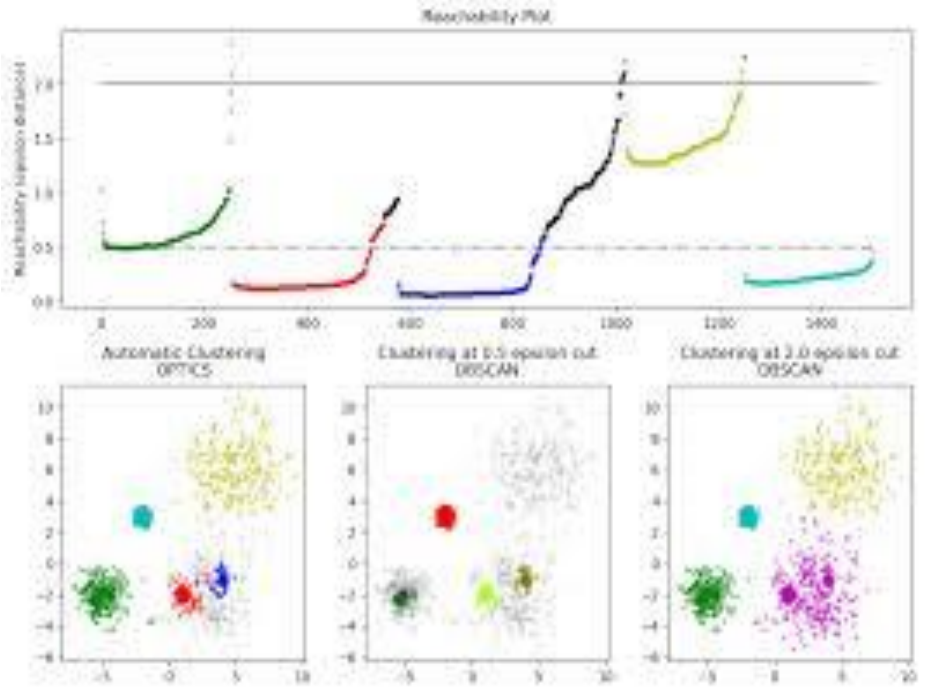




Eps = 6mm

MinPts = 5

Core_Distance(p) = 3mm

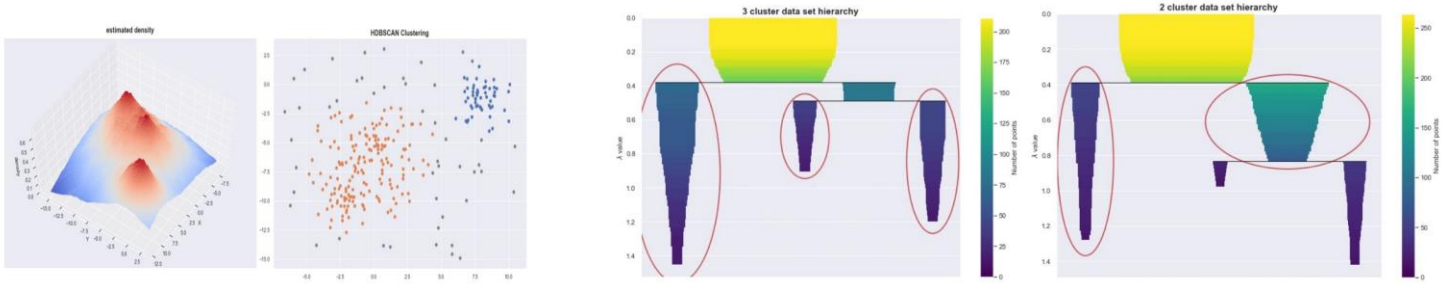


OPTICS

Ordering Points To Identify the **Clustering** Structure

Cluster Selection for varying densities?

HDBSCAN Hierarchical DBSCAN



- ✓ Arbitrarly shaped clusters
- ✓ Variable density

- We estimate densities based on core distances and form **the density landscape** (what makes these density-based)
- We can use a global threshold to **set the sea level at and identify the islands** (DBSCAN)
- We can try to decide, **are these several mountains or one mountain with multiple peaks**? (HDBSCAN)

Modus Operandi

YSO Sample (c2d Data)

Cross-Match with Gaia DR3 Data

Get parameters

Make a control sample, Query DR3 data in 9X6 deg rectangle

Use ML: DBSCAN, OPTICS, HDBSCAN to identify members

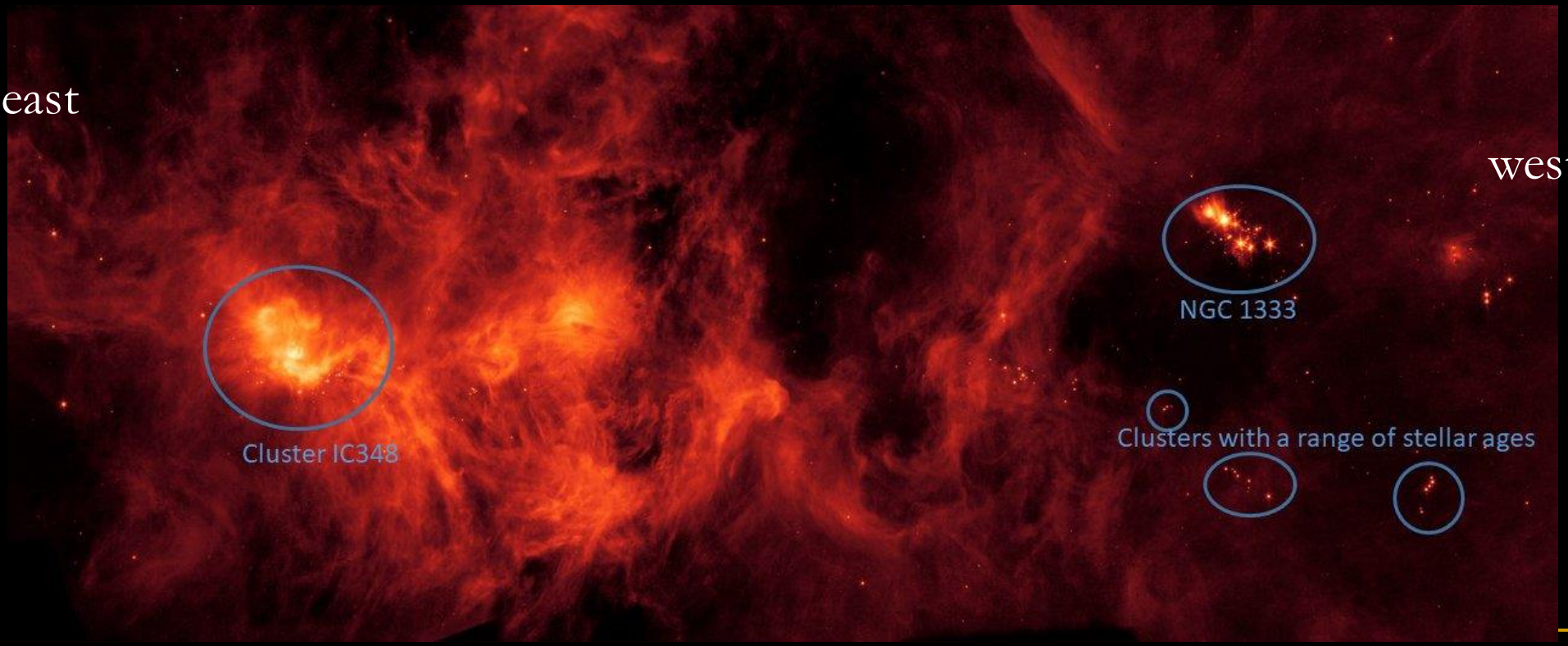
Obtain parameters

Identify YSO class using 2MASS +WISE. Identify clusters.

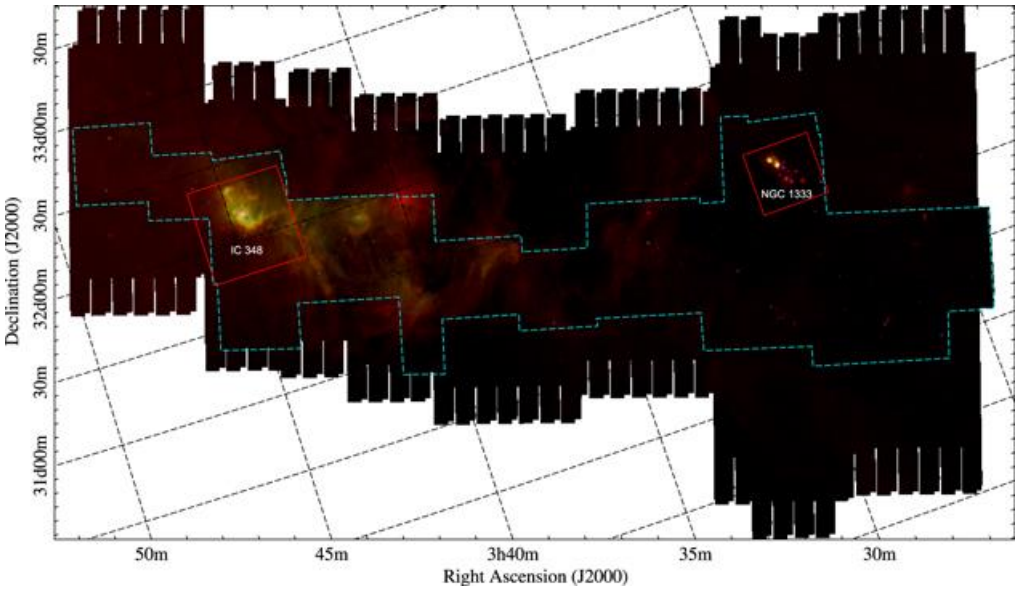
Find SFR

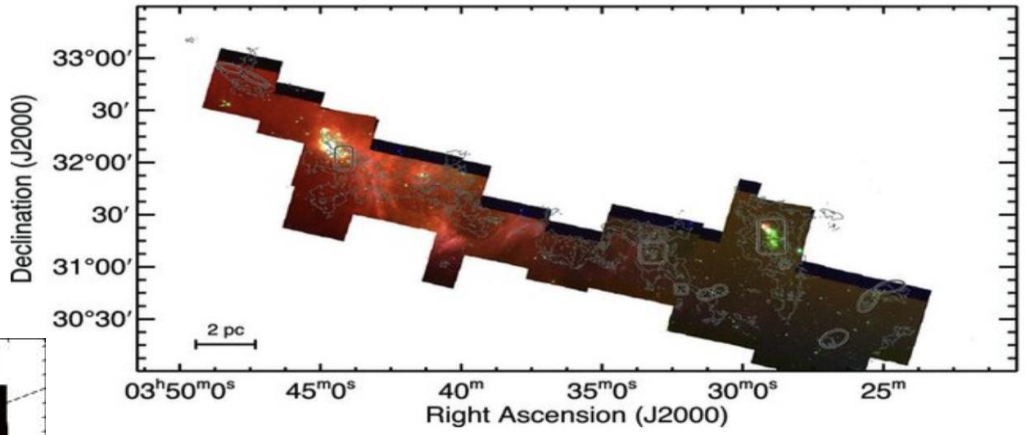
Hasan et al., 2023

The Perseus molecular cloud represents an ideal target for studying the fundamental properties of young stars and their environment since the complex is sufficiently nearby Consisting of an elongated chain of dark clouds, Perseus spans over an area of 7° × 3° in the plane of the sky. The most prominent substructures are Barnard 5 (B5) and IC 348, at the eastern edge and Barnard 1 (B1), NGC 1333, L1448, L1451, and L1455 at the western edge of the complex.

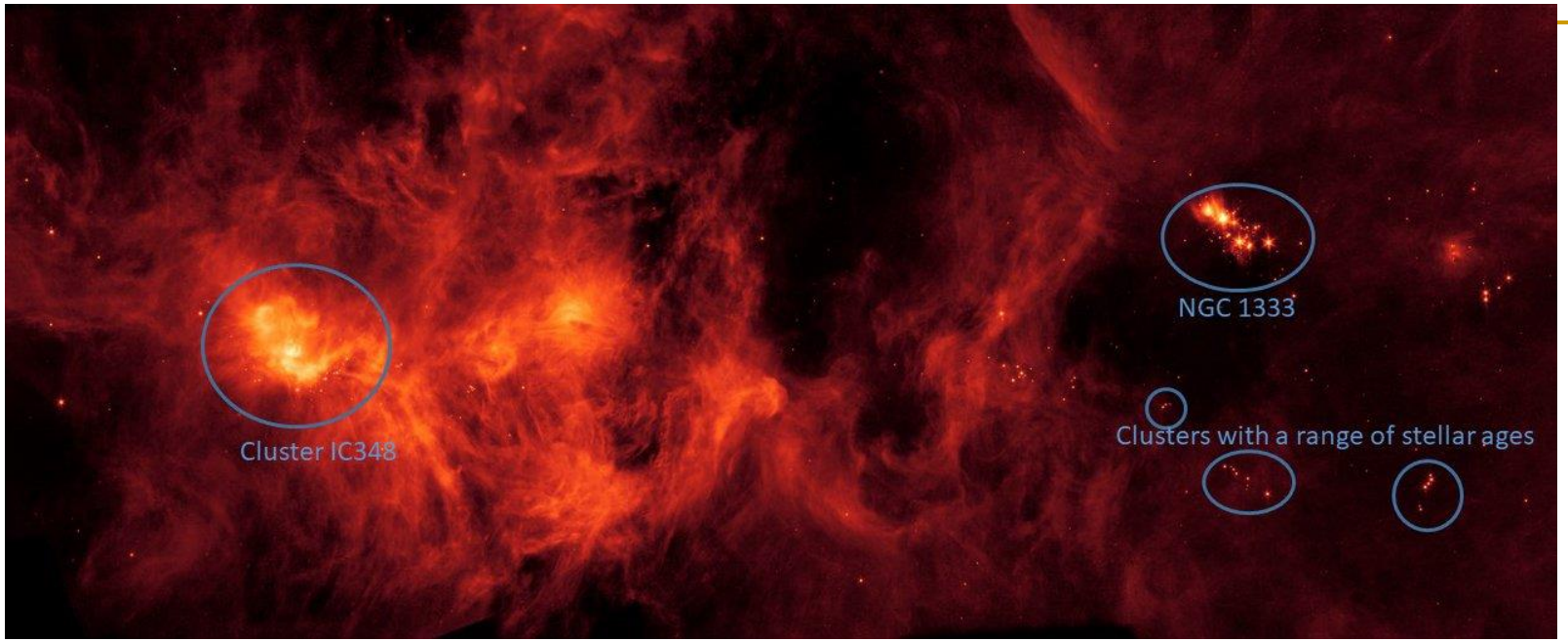


c2D (Young et al 2015)

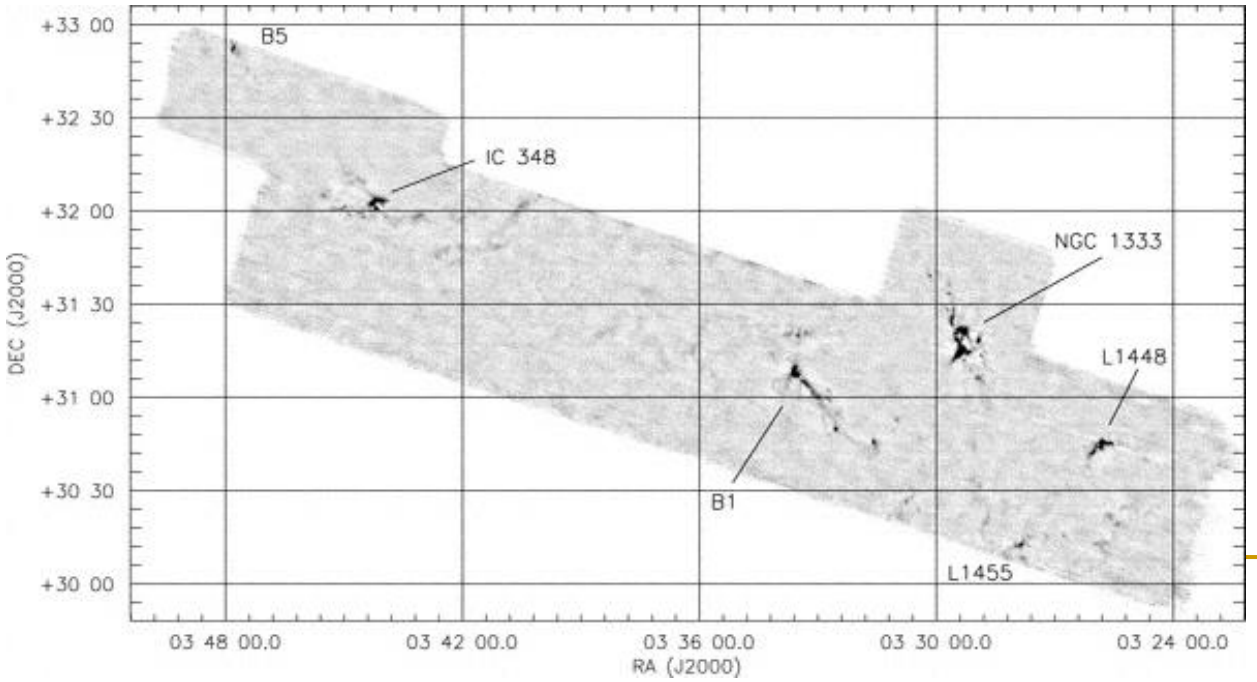




Spitzer IRAC (color) image of the c2d coverage of the Perseus cloud made from 3.6, 4.5, and 8.0 μm images of the region (Evans et al. 2009, Young et al 2015). Sources: 369

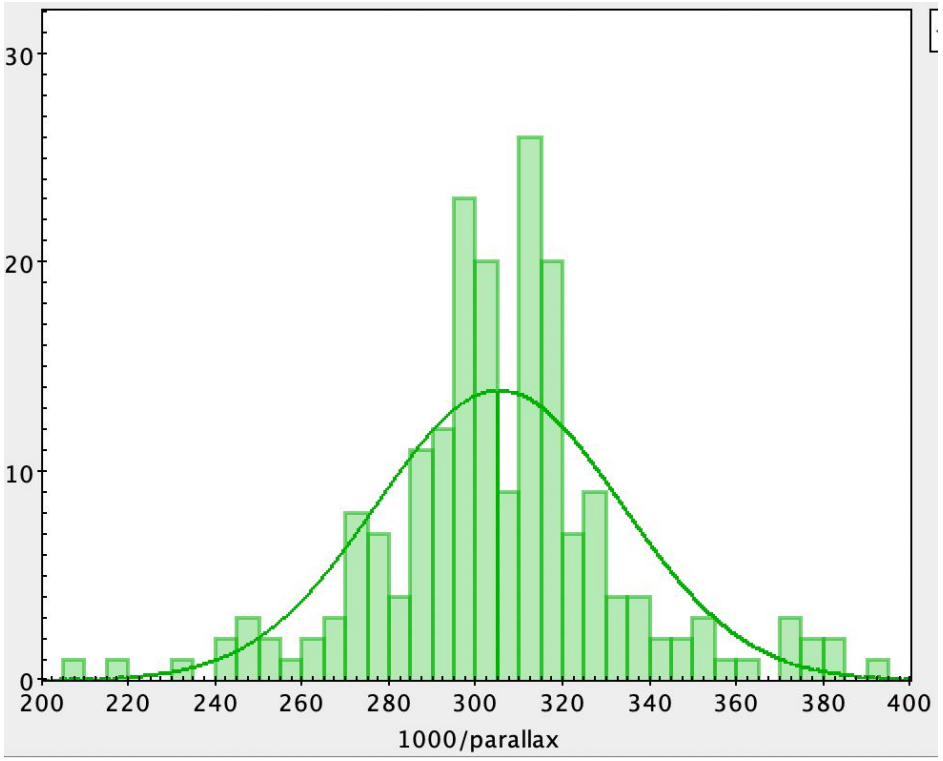


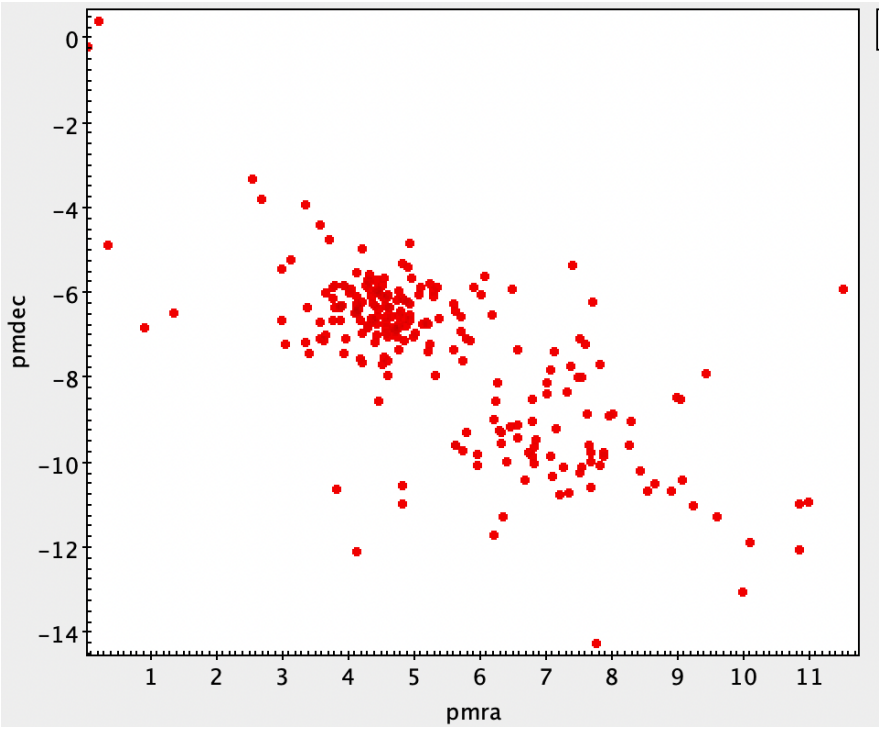
Spitzer Image



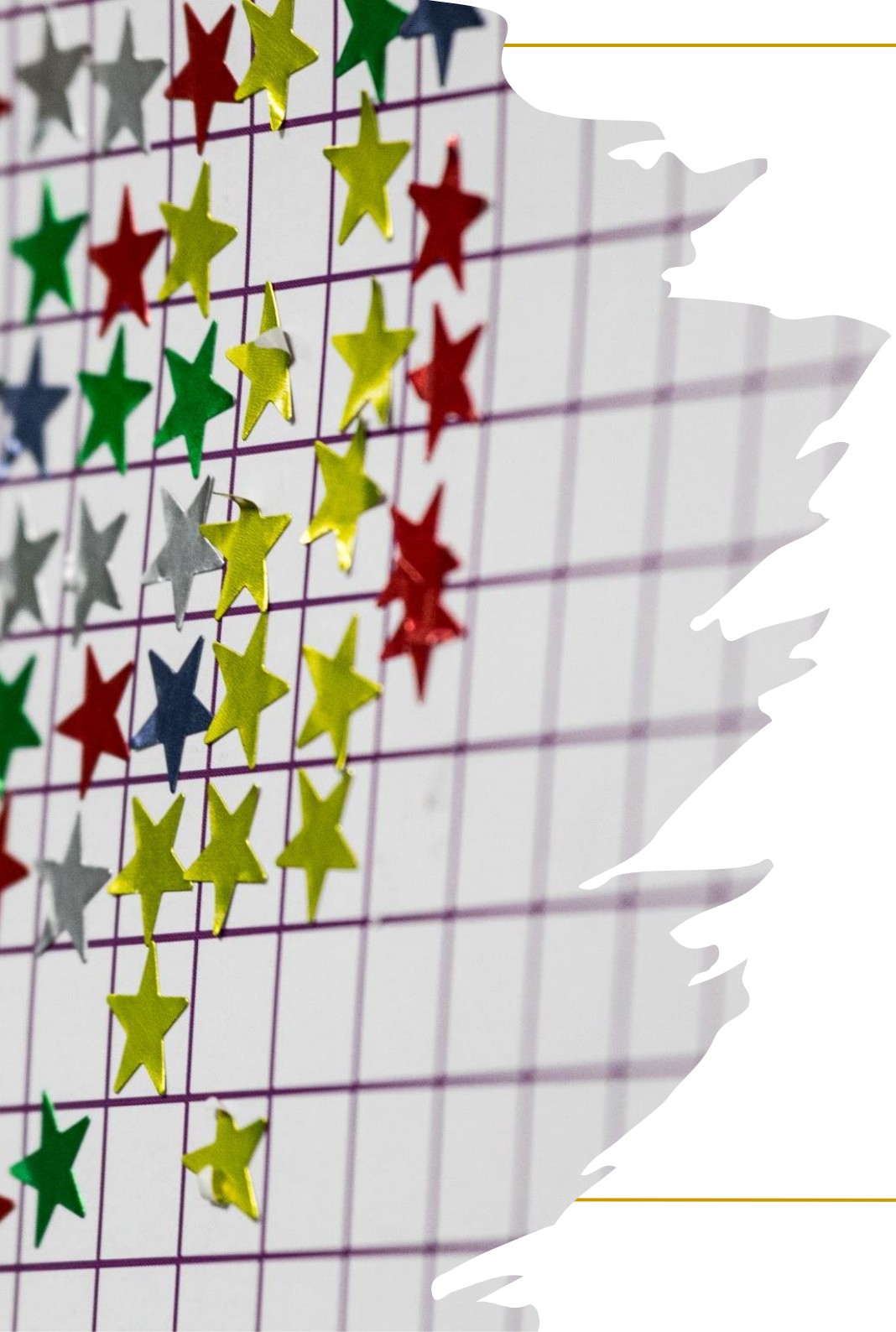
Bolocam 1.1 mm map of the Perseus Molecular Cloud, which covers 7.5 deg(143 pc^2 at a distance of 250 pc (34000 data points in total).

Match: c2d+Gaia DR3=252





Mean:	305.1004
Standard Deviation:	28.296259



Gaia Query Sample (1196)

- Query 9X 6 deg,
- RUWE < 1.4, RPlx > 3,
- 9.4 > pmRA > 2
- -4 < pmDE < -12
- 4.2 > Plx > 0
- 3123 stars (464111 stars without filters)
- XYZ, pmRA, pmDE

Machine Learning: Unsupervised Techniques
Using Clustering to find new members

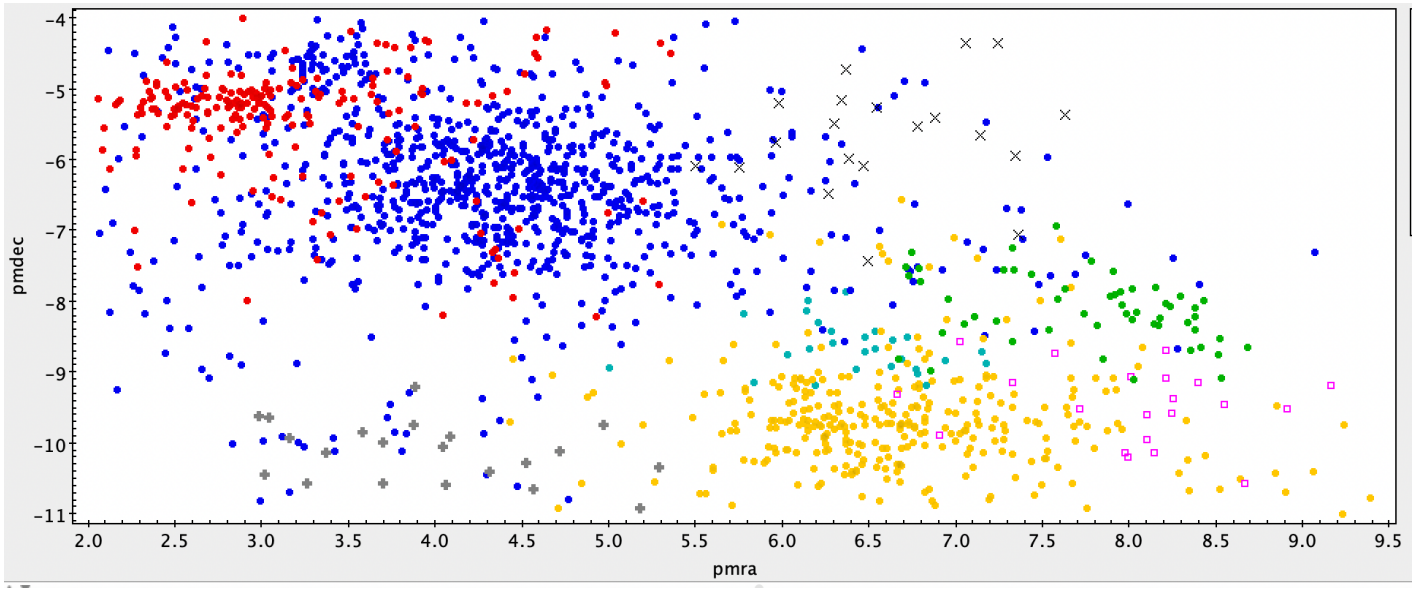
Density-based clustering

- Independent of shape and number of clusters
- Estimate the densities
- Pick regions of high density
- Combine points in these selected regions

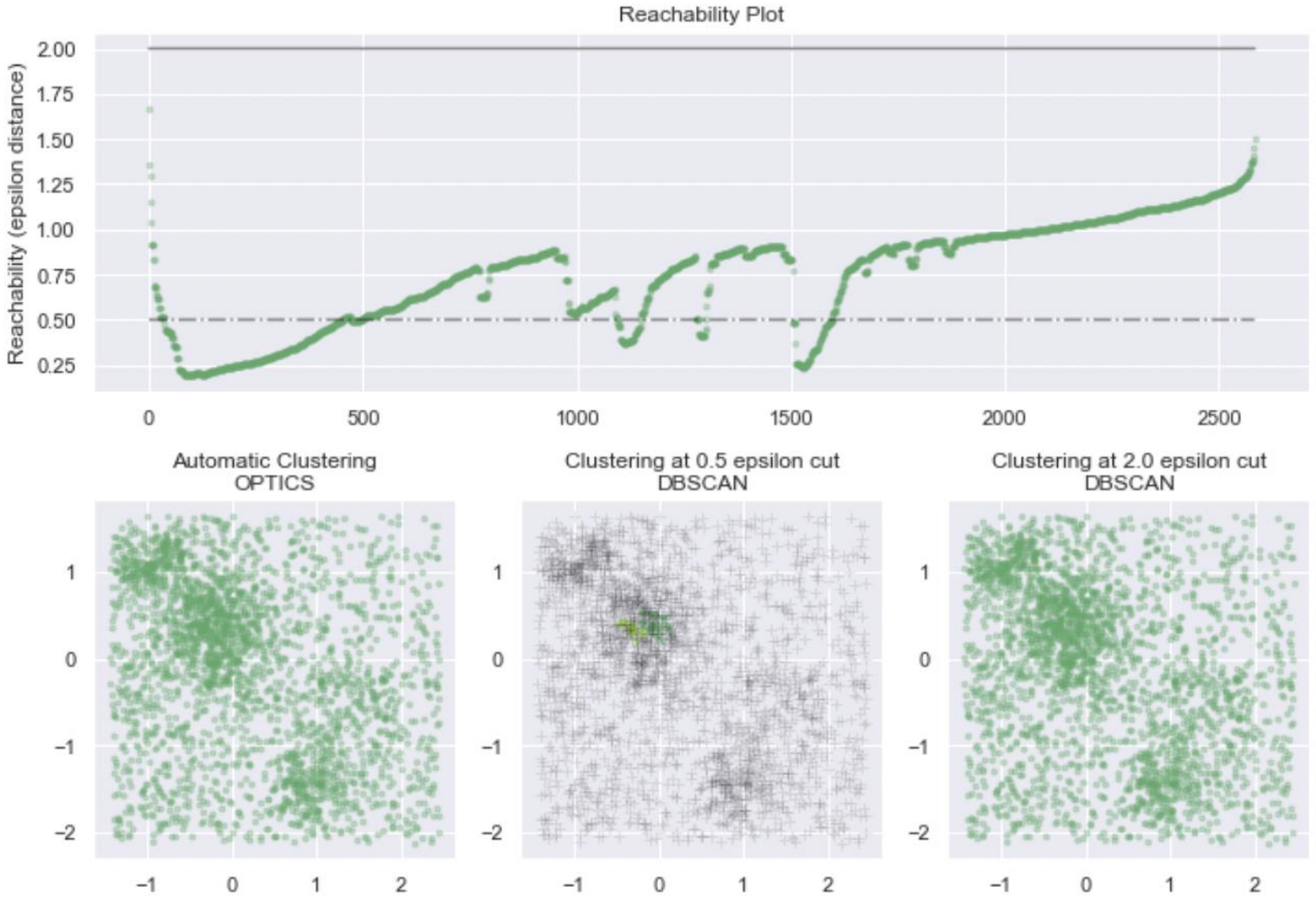
DBSCAN OPTICS HDBSCAN

DBSCAN (3123 points)

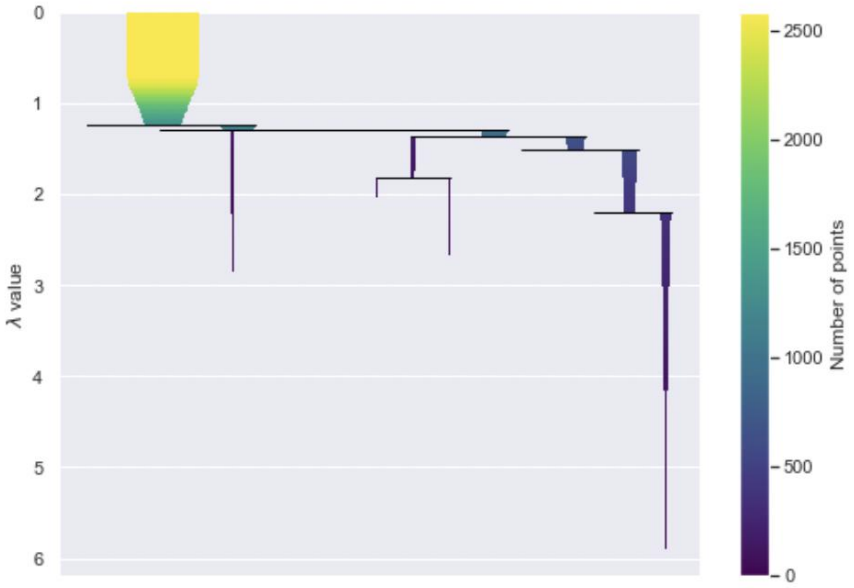
-1	1015
2	889
3	305
4	1
6	53
1	27
0	26
5	21
7	21
8	20
9	8



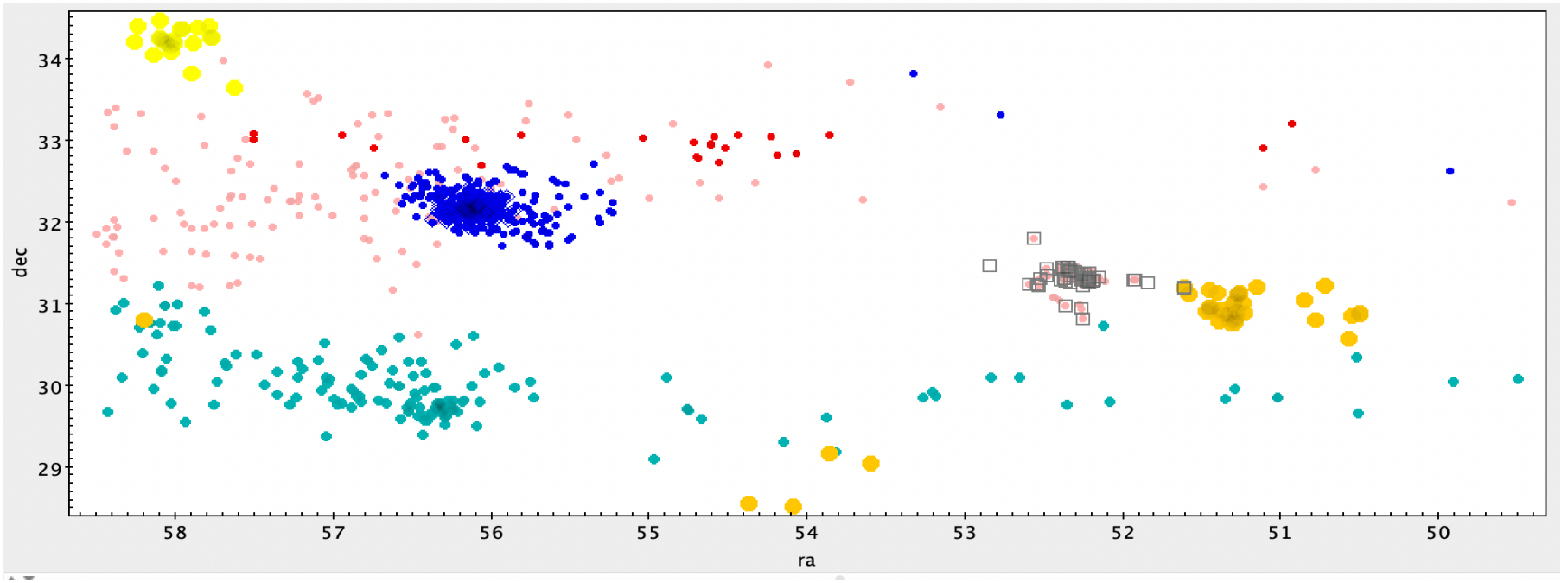
Optics

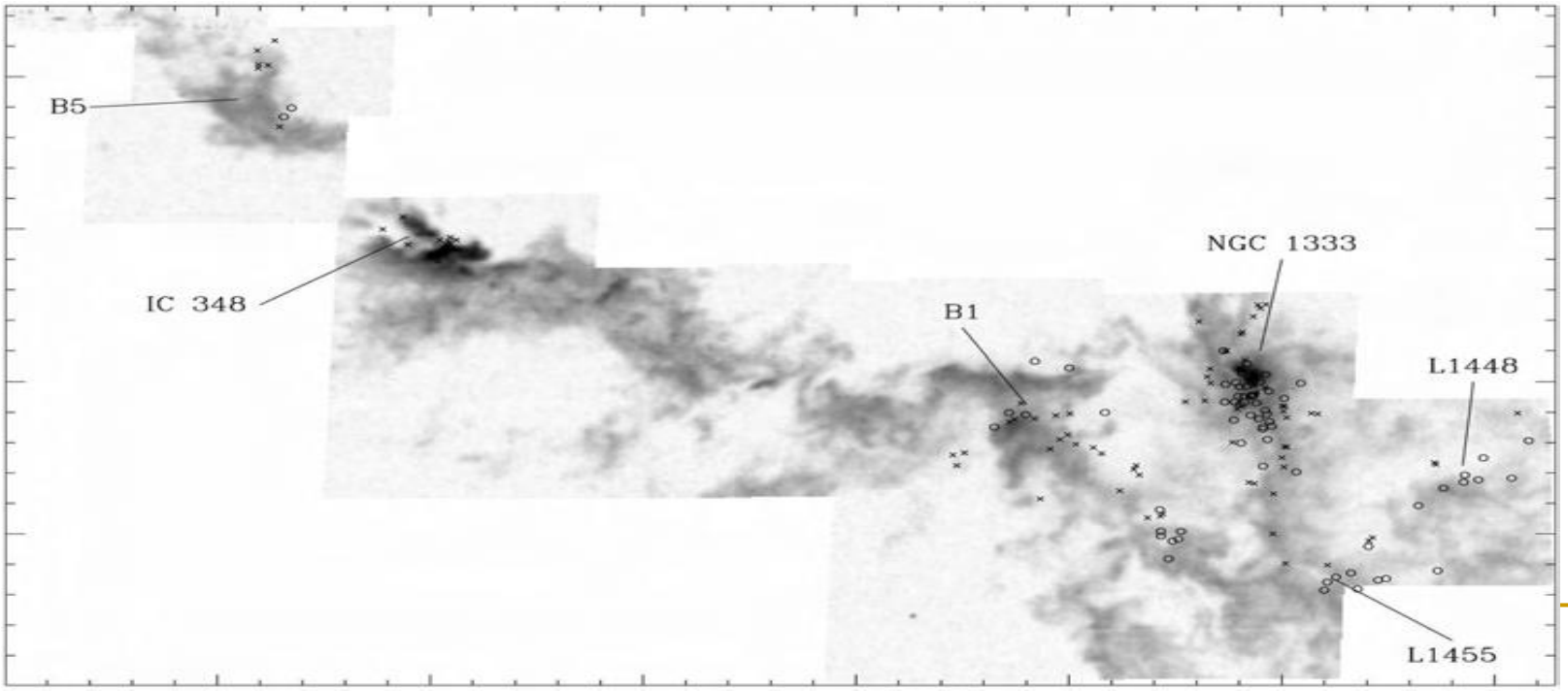

Reachability Plot
Reachability (epsilon distance)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0
500
1000
1500
2000
2500
Automatic Clustering
OPTICS
Clustering at 0.5 epsilon cut
DBSCAN
Clustering at 2.0 epsilon cut
DBSCAN
1
0
−1
−2
−1
0
1
2

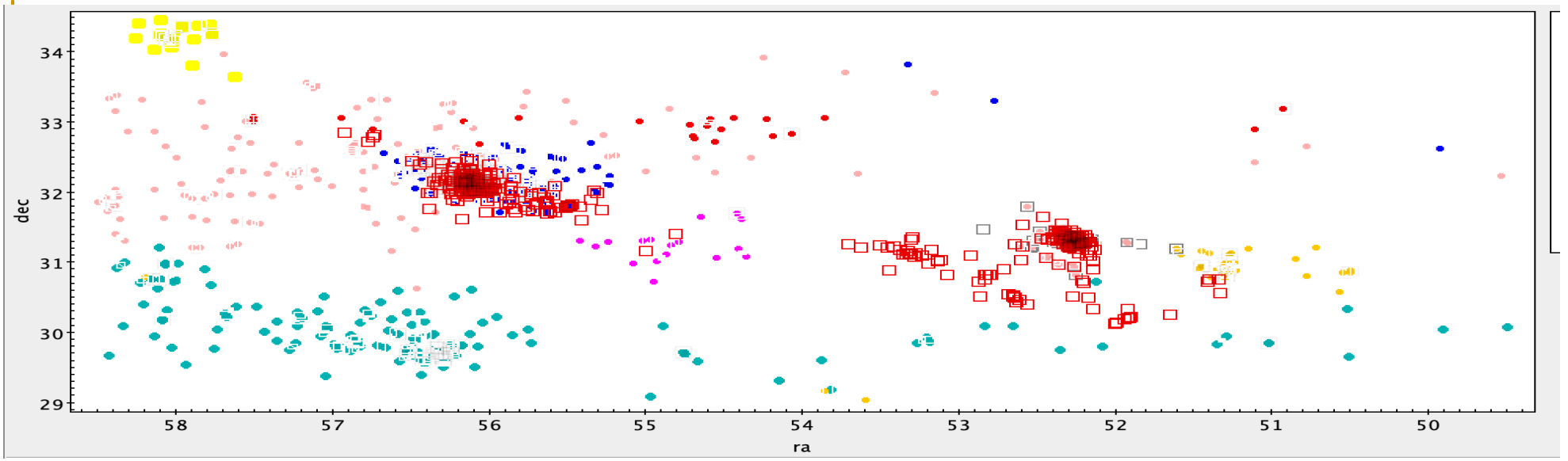
HDBSCAN

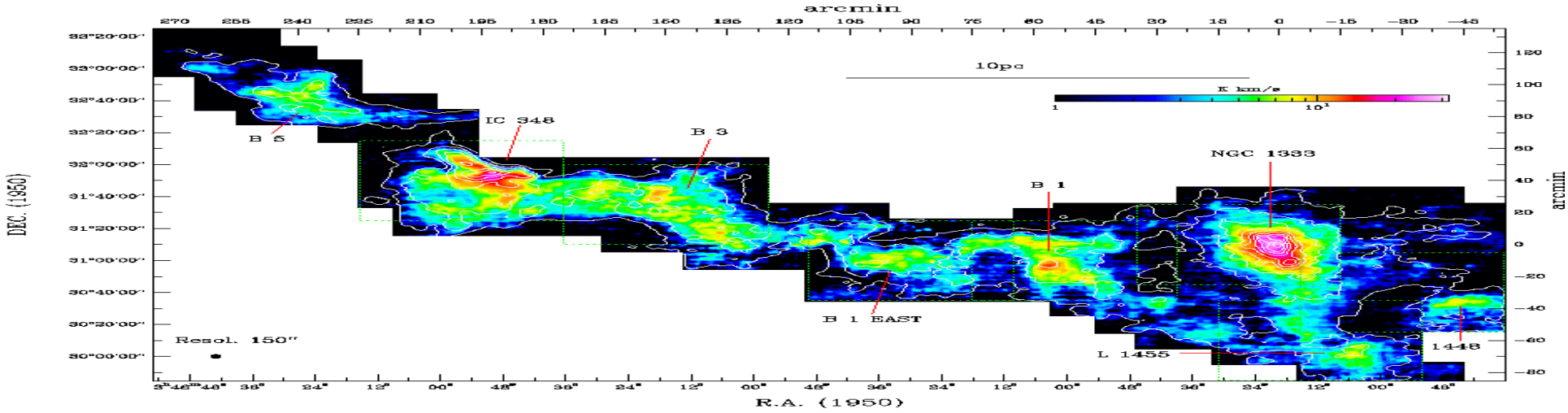


- hd0 14
- hd1 1761
- hd2 18
- hd3 148
- hd4 38
- Hd5 202
- Hd6 21
- Hd7 25
- Hd8 357


dec
ra
34
33
32
31
30
29
58
57
56
55
54
53
52
51
50

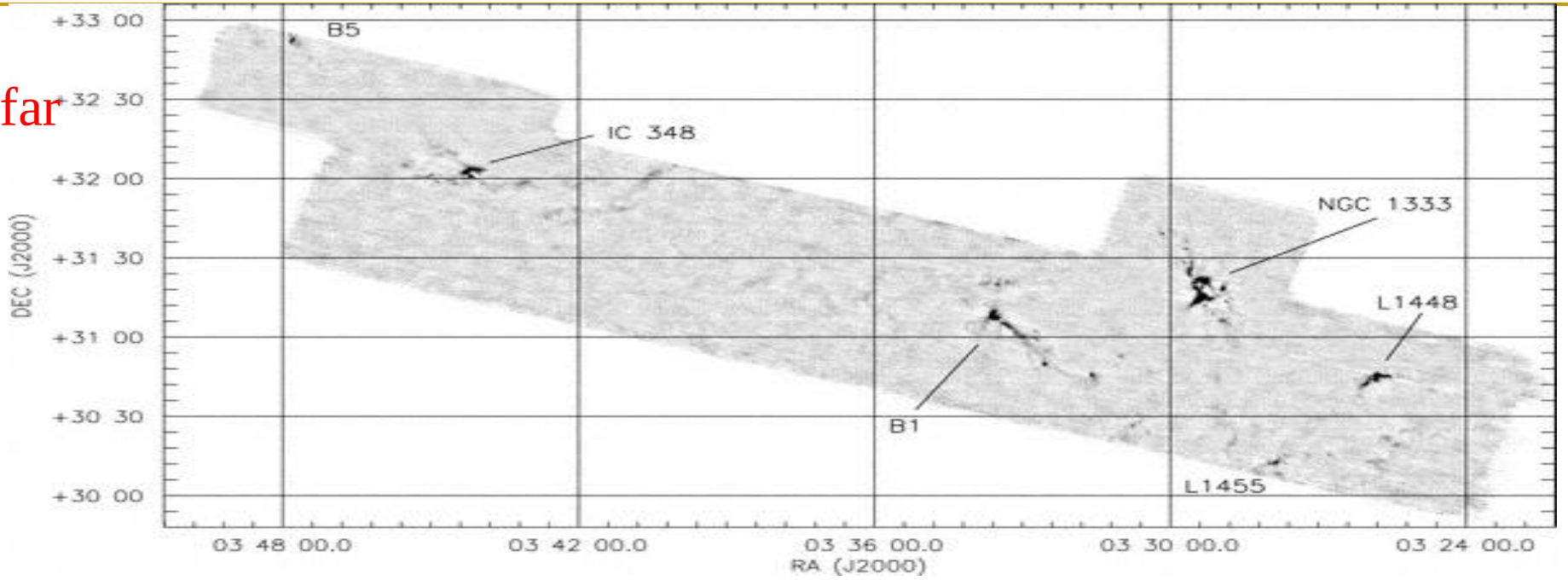

B5
IC 348
NGC 1333
B1
L1448
L1455



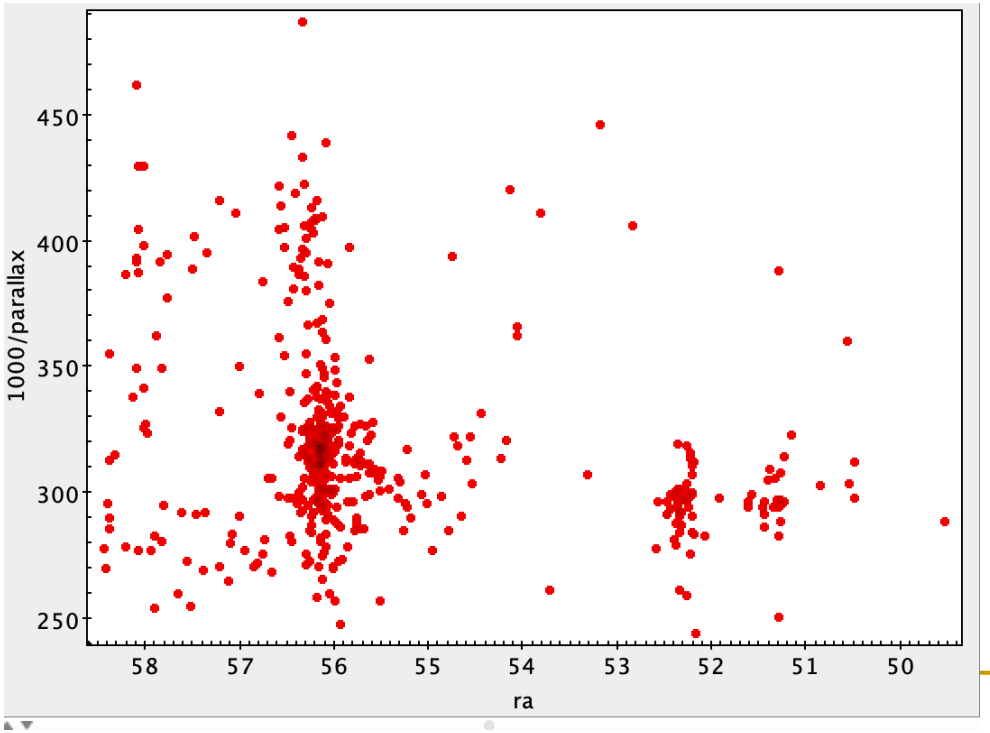


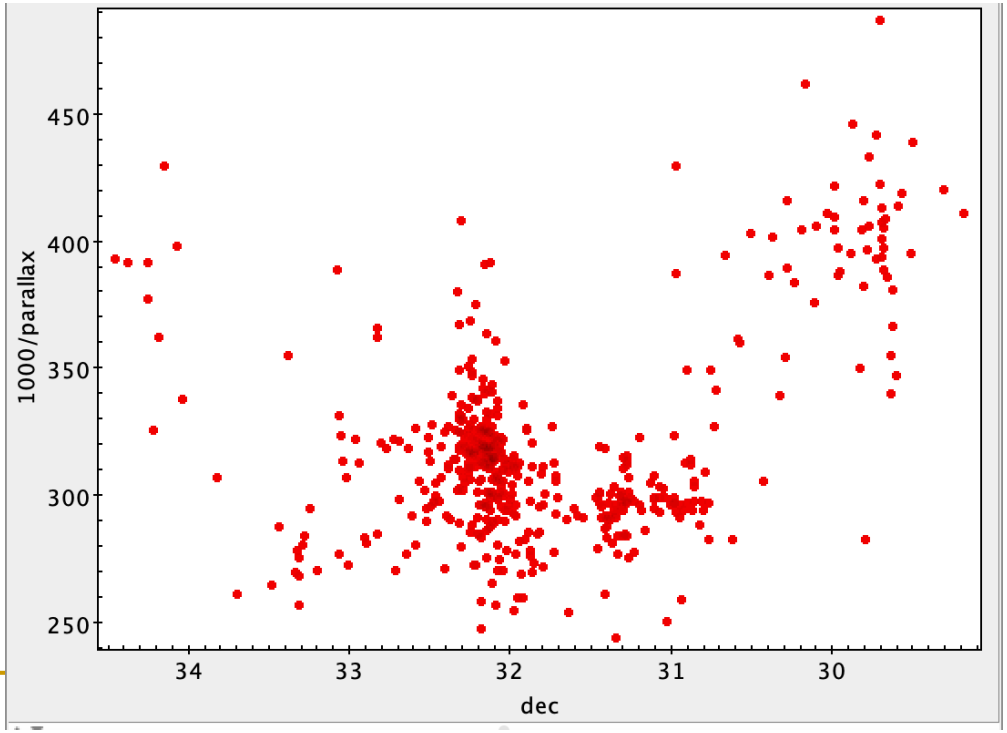
212 match with c2d (red boxes)

far

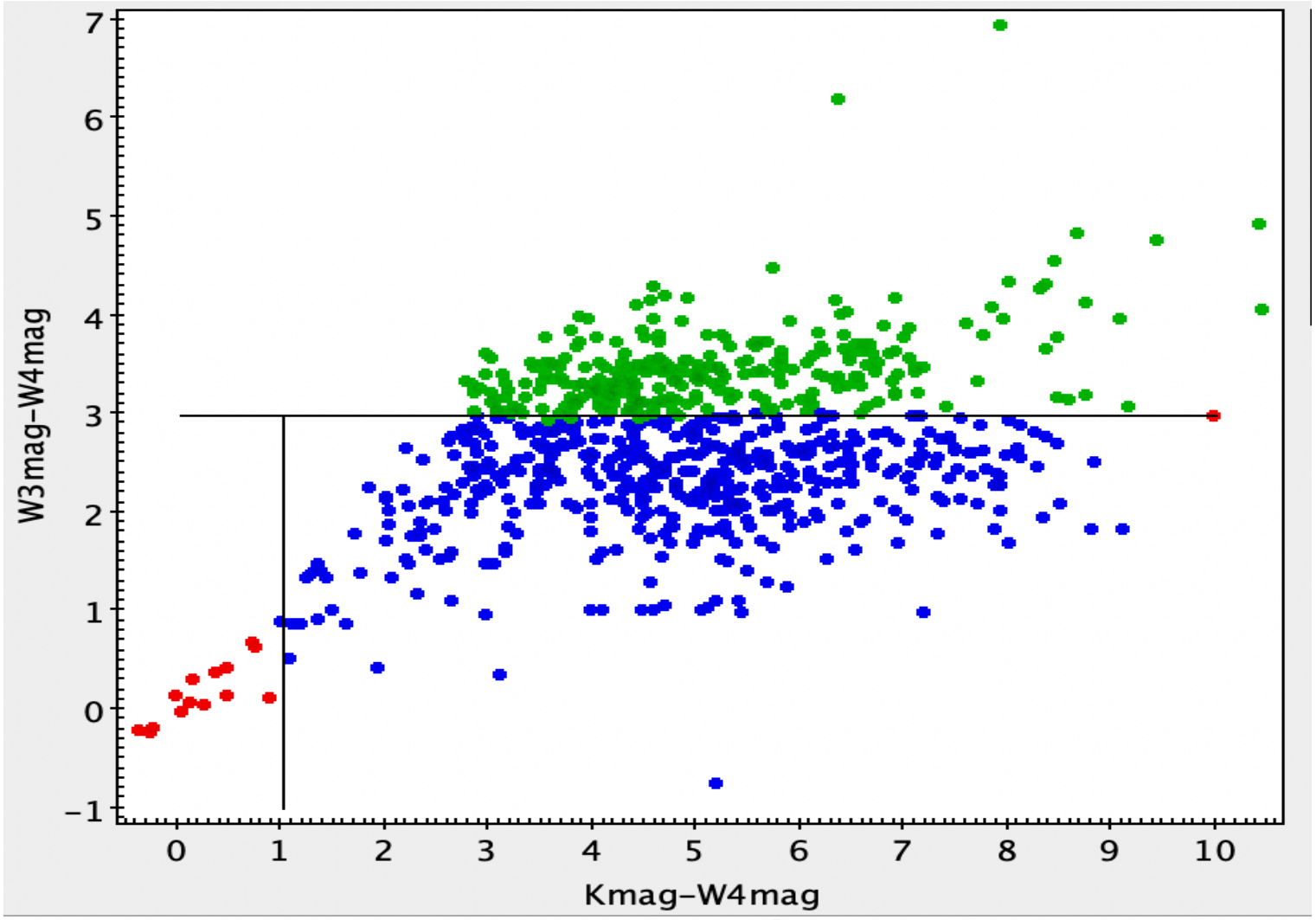

B5
IC 348
NGC 1333
L1448
B1
L1455
+33 00
+32 30
+32 00
+31 30
+31 00
+30 30
+30 00
DEC (J2000)
03 48 00.0
03 42 00.0
03 36 00.0
03 30 00.0
03 24 00.0
RA (J2000)

near


1000/parallax
450
400
350
300
250
58
57
56
55
54
53
52
51
50
ra


1000/parallax
450
400
350
300
250
34
33
32
31
30
dec

Xmatch with 2MASS, WISE (759 sources)



SFR

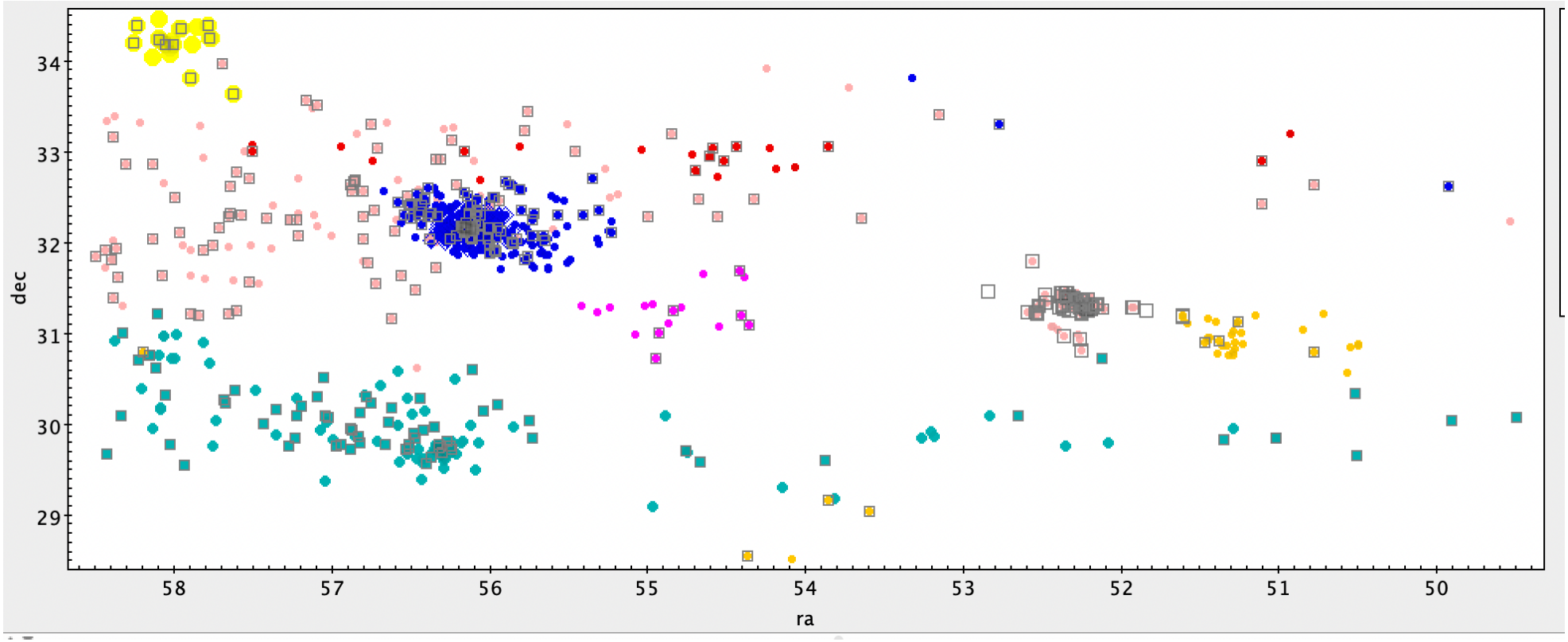
The number of YSOs in a given evolutionary stage in a star-forming region, Ni is converted into a SFRi using

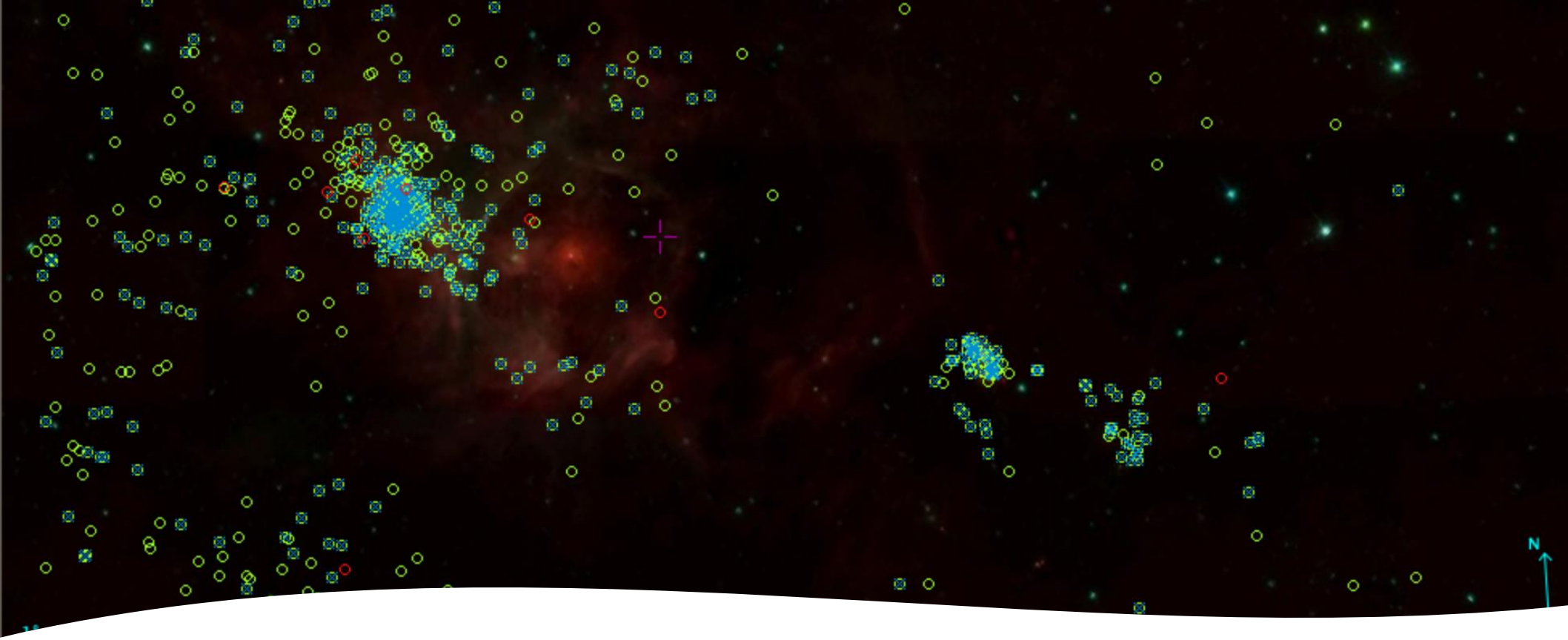
$$SFR_i = N_i \langle M \rangle / \tau_i$$

where the index i runs over the different protostellar classes (i=0, I, F, II, and III for objects in Class 0, Class I, Class F, Class II, and Class III, respectively), $\langle M \rangle$ is the mean protostellar mass present in the region, and τ_i is the YSOs typical lifetime in each class.

Distribution of CII sources (278 sources)

AG=1.86 \pm 1.22



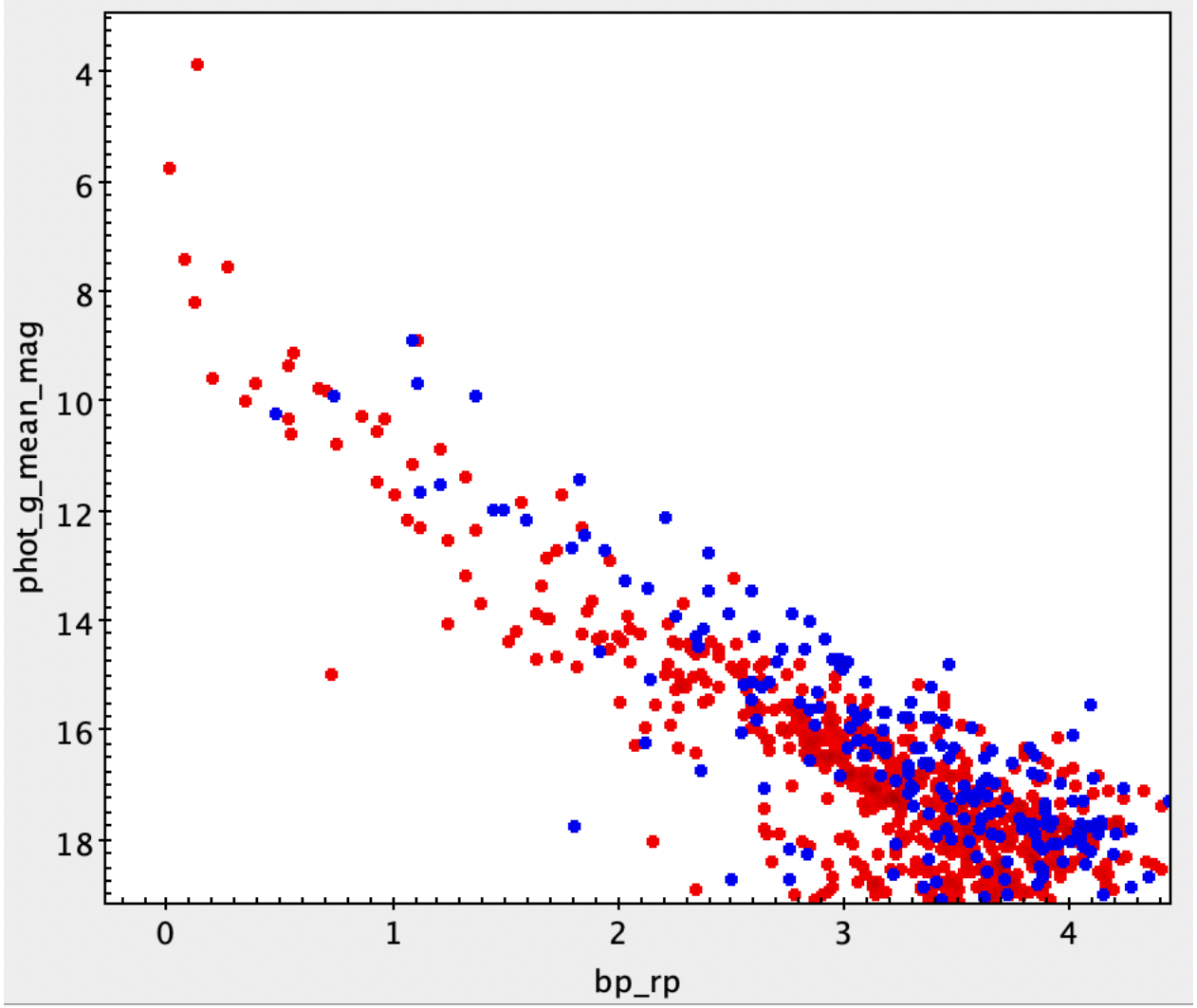


CII sources, SFR from star counts

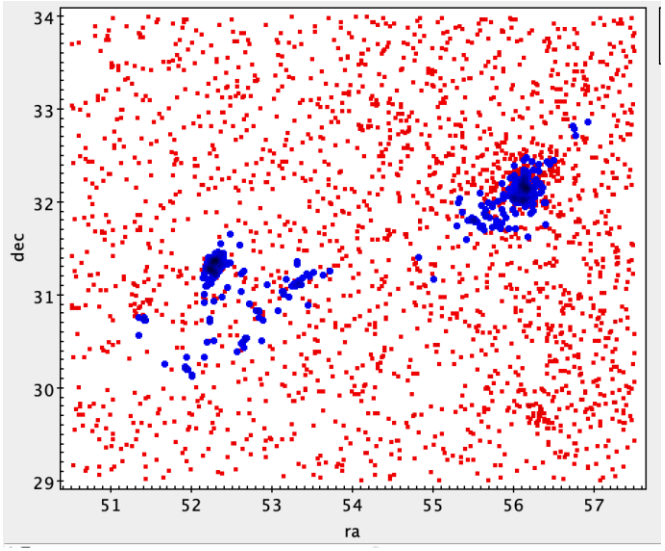
SFRi = Ni⟨M⟩/τi, for class II protostars, τII is often taken to be ≈ 2 Myr. The mean mass is usually to be ⟨M⟩ = 0.5 M⊙ which corresponds to the mean mass in a Milky-Way like IMF.

- High SFH at the eastern side near IC 348 and lesser at the western side near NGC 1333

CMD



Blue: c2d
Red: Our sample



Results

We compiled a sample of YSOs using c2d

Matched with Gaia members (252)

Extracted sources with 2d rad

Clustering XYZ,pmRA,pmDE with DBSCAN, OPTICS, HDBSCAN

Got 809 YSO members. Identified clusters.

Matched with 2MASS, WISE to classify

Found SFR…..Hasan, 2025 (in prep)

Thank You!