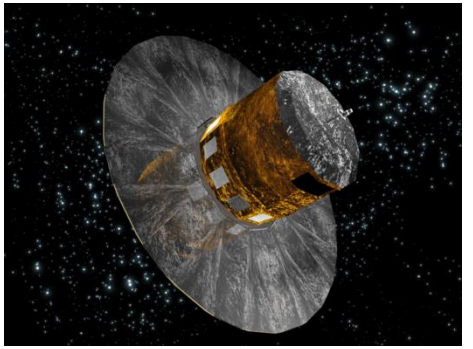
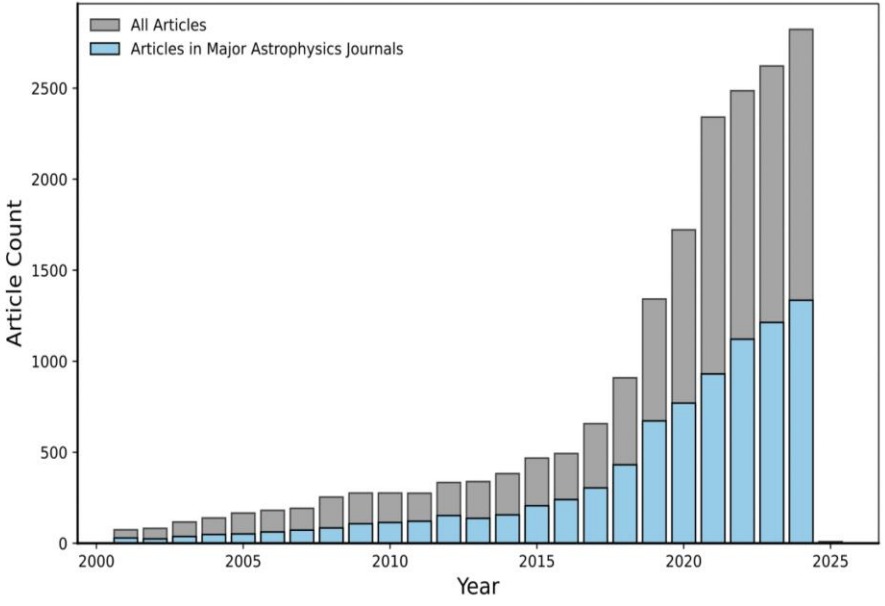
Machine Learning in Astronomy



Keywords

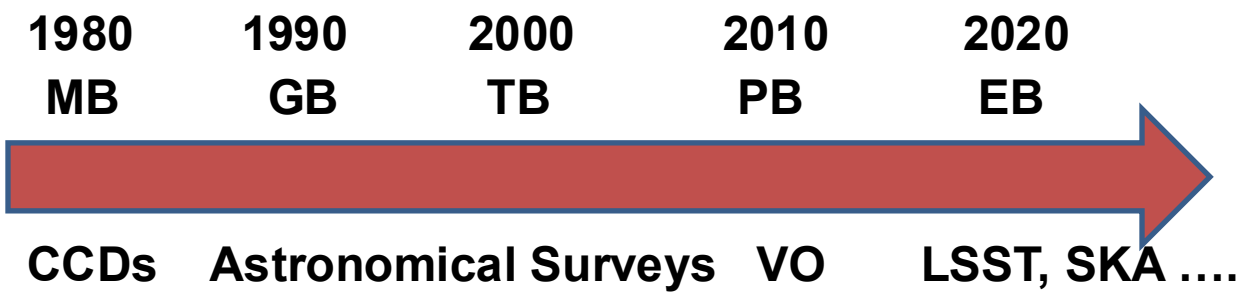
Artificial Intelligence (AI) | Machine Learning (ML) | Astronomical Techniques



Priya Hasan
Maulana Azad National Urdu University
Hyderabad,500032

Md Mahmudonobe, Mudasir Raja, Md Saifuddin,
S N Hasan
Maulana Azad National Urdu University Hyderabad,500032
Wayne State University, Michigan, USA.

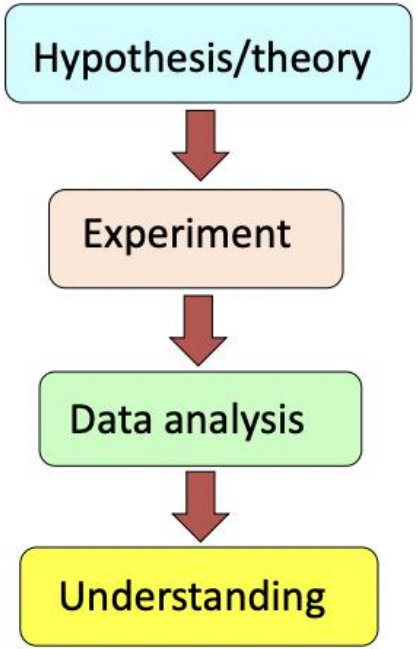
Astronomy has changed over the years: From data poor to data rich science



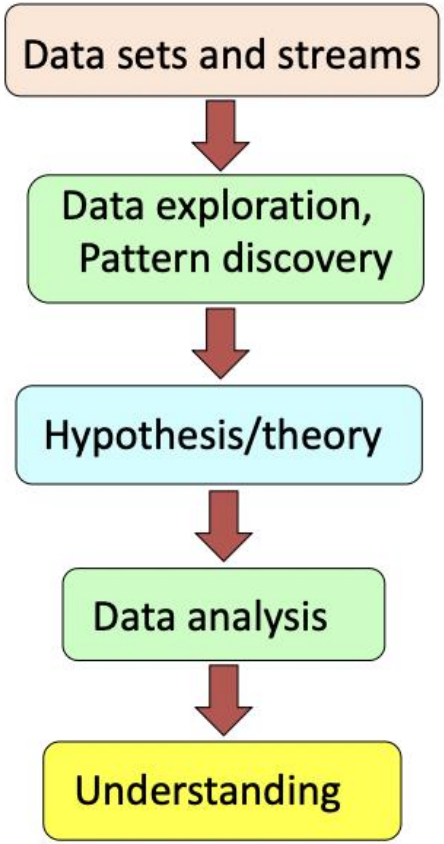
Coupled with advances in Information Technology/computing

It has become essential for astronomers to handle the big datasets and interpret the complex astrophysical problems by using sophisticated statistical techniques and technology.

Hypothesis-driven science



Data-driven science



The two approaches are complementary

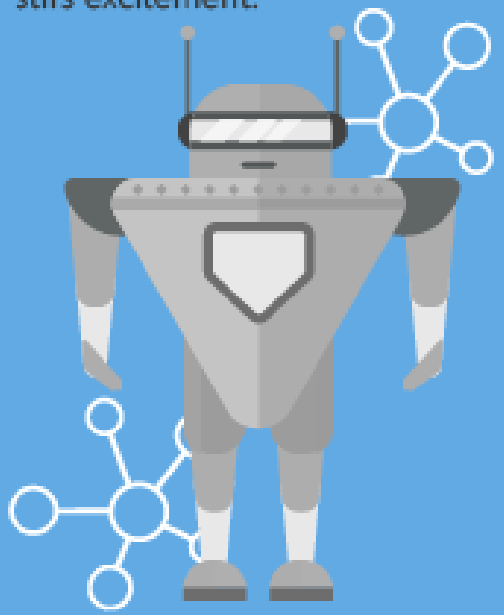
The need .....

As we collect HUGE amount of Astronomical data, it has gone beyond humans capacity to analyse without help of machines.

Training
Algorithms learn patterns from vast astronomical datasets containing billions of observations.
Classification
Systems automatically separate stars from galaxies with increasing precision.
Discovery
ML identifies patterns humans might miss, leading to new astronomical findings.

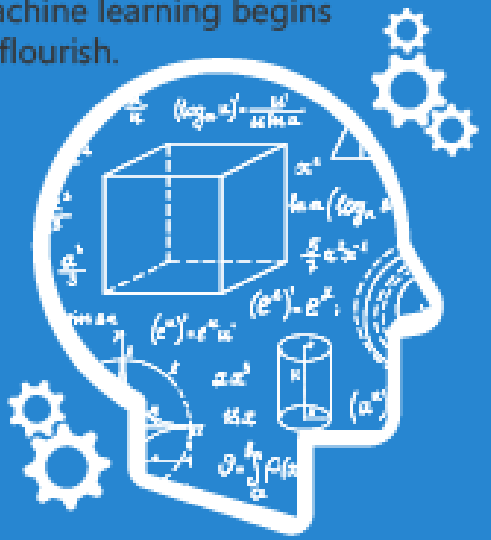
ARTIFICIAL INTELLIGENCE

Early artificial intelligence stirs excitement.



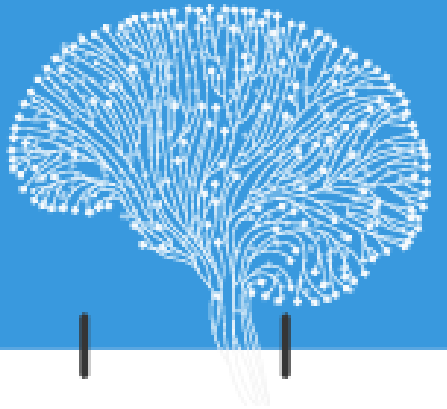
MACHINE LEARNING

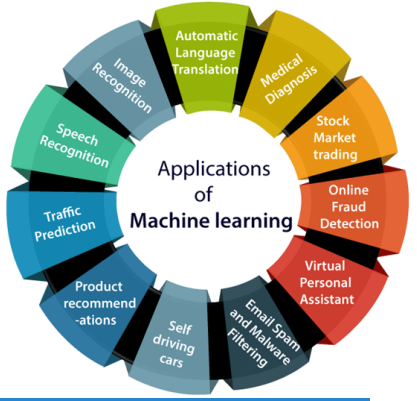
Machine learning begins to flourish.



DEEP LEARNING

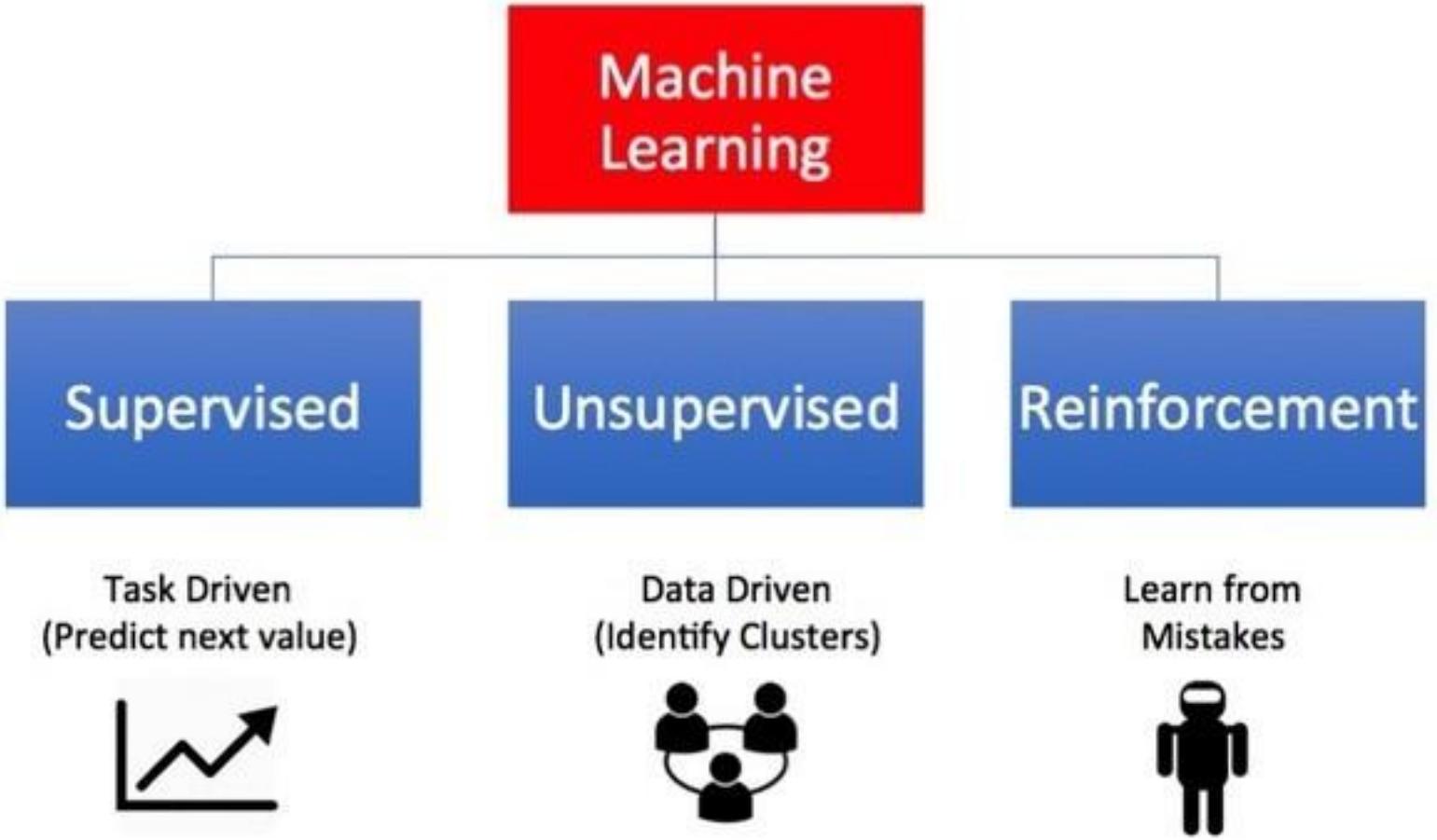
Deep learning breakthroughs drive AI boom.



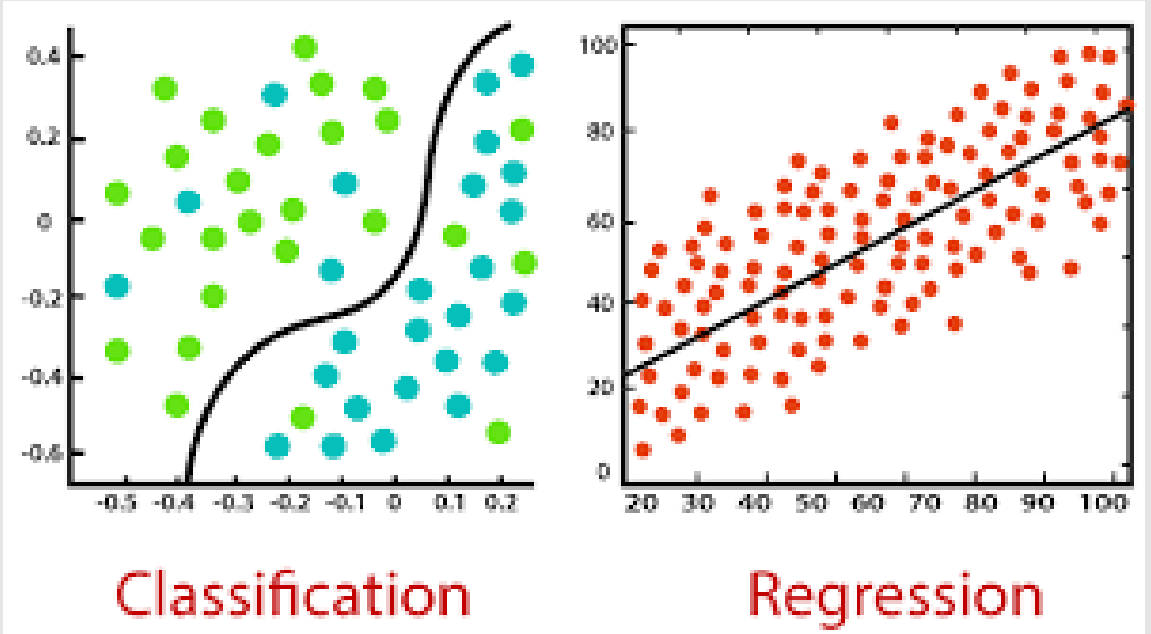


1950's 1960's 1970's 1980's 1990's 2000's 2010's

Types of Machine Learning



Supervised Learning algorithms.



Regression and Classification algorithms:

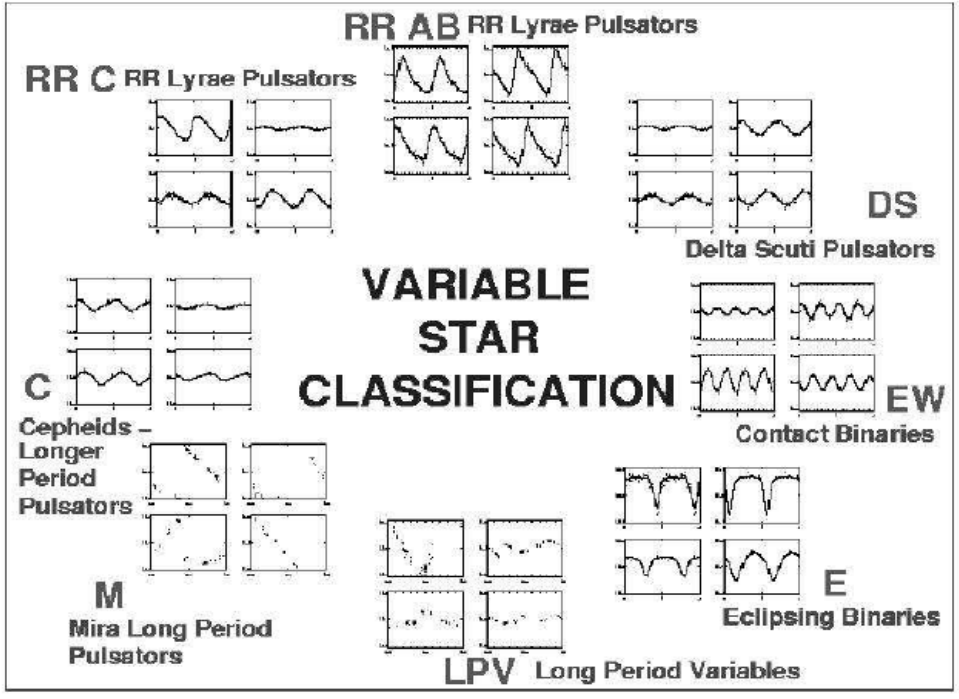
Regression algorithms are used to **predict the continuous** values such as age, redshift etc. and

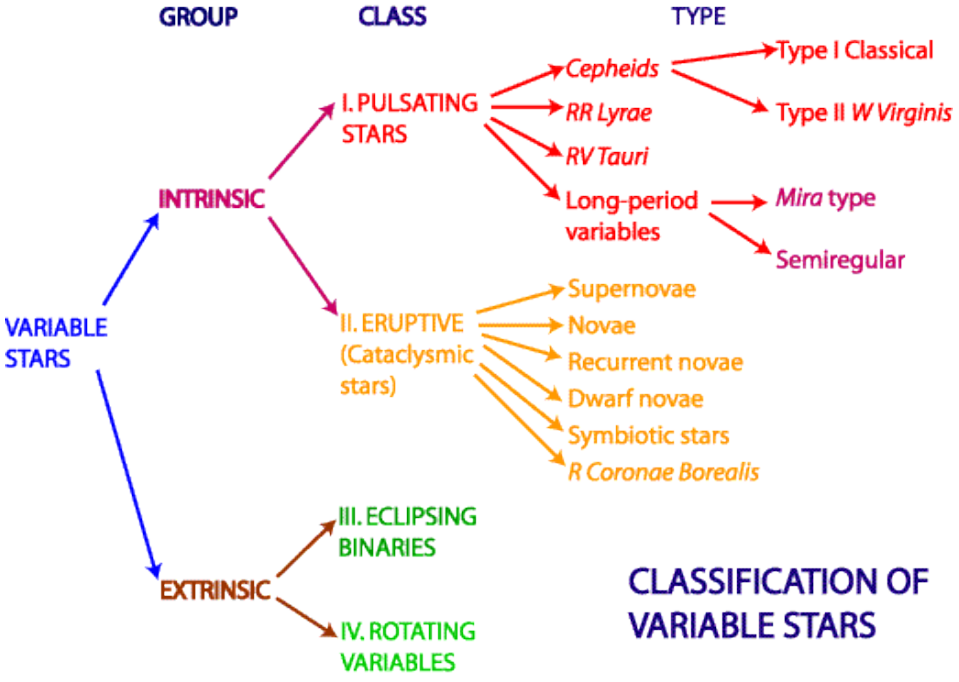
Classification algorithms are used to **predict/Classify the discrete values** such as morphological types – Spirals, Ellipticals. Members , Non-members

Stars, Galaxies

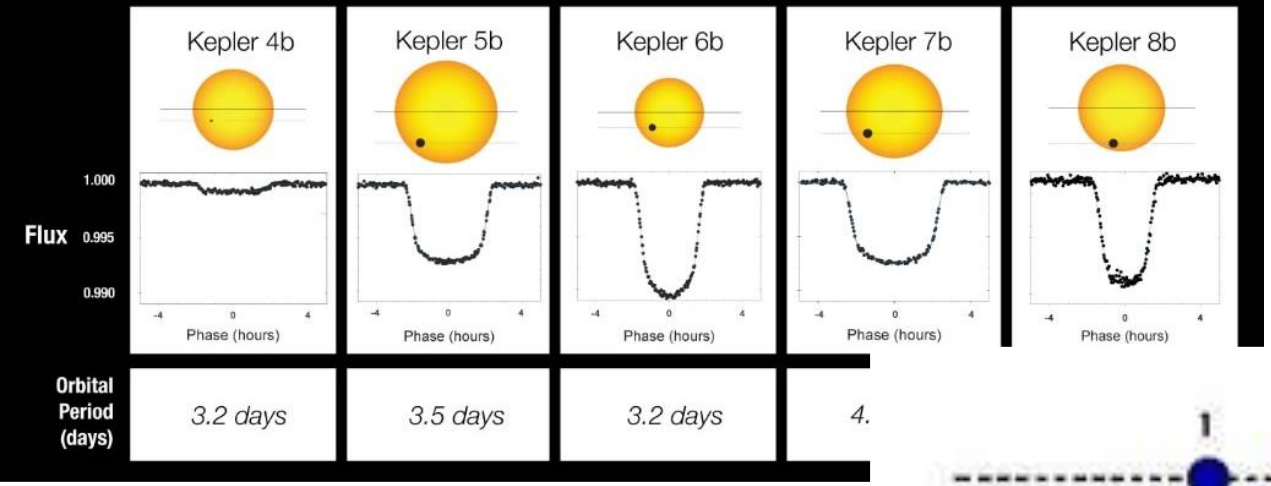
Classification of Galaxies

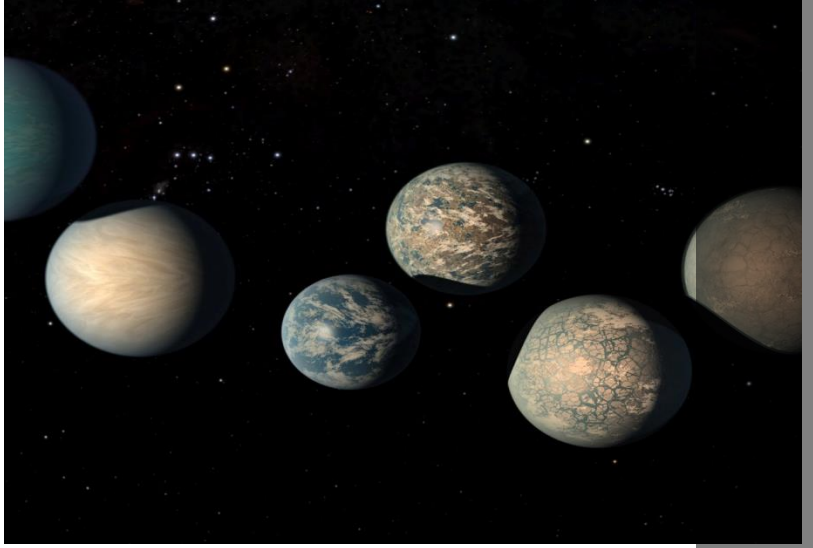
Variable stars



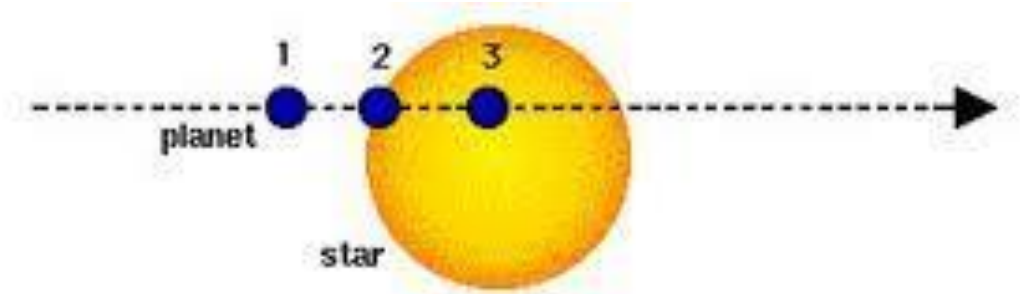


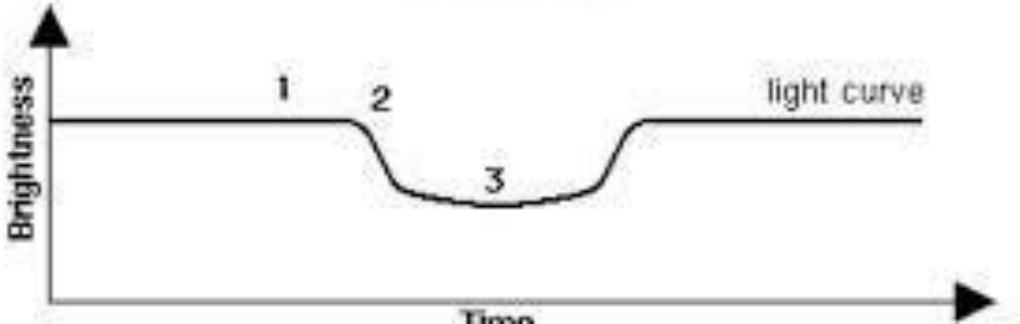
Transit Light Curves



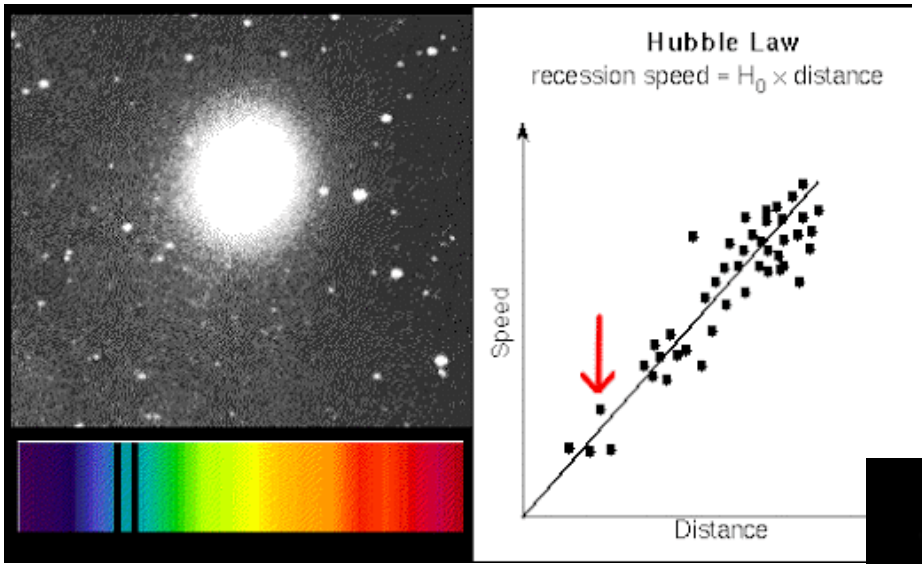


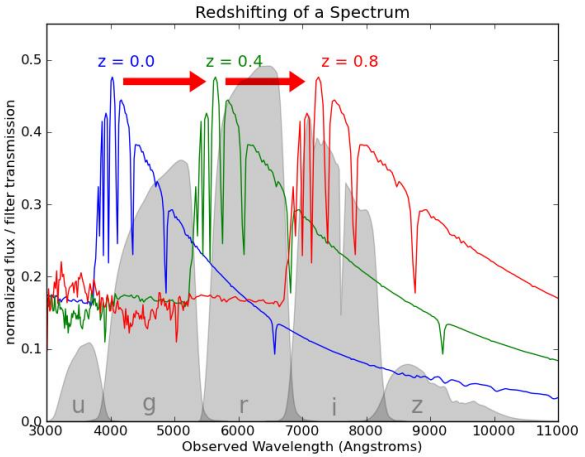
Exoplanet discoveries

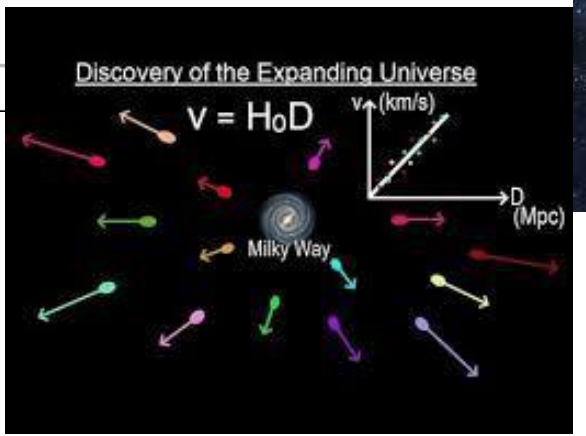




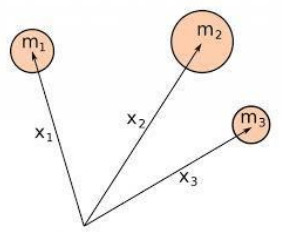
Redshift


Hubble Law
recession speed = $H_0 \times$ distance
Speed
Distance

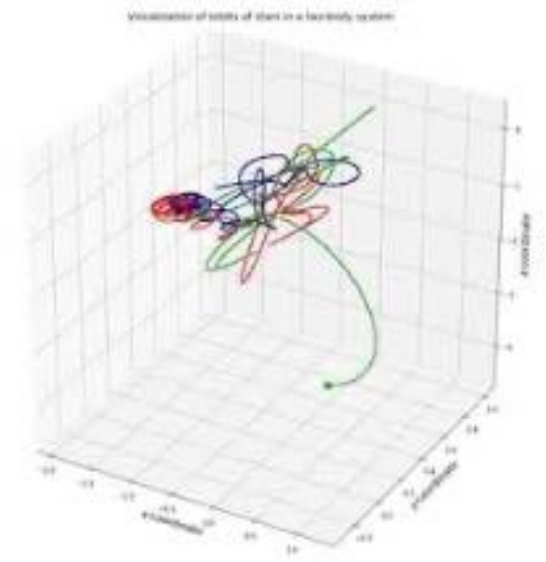

Redshifting of a Spectrum
z = 0.0
z = 0.4
z = 0.8
normalized flux / filter transmission
u
g
r
i
z
Observed Wavelength (Angstroms)


Discovery of the Expanding Universe
v = H₀D
v (km/s)
D (Mpc)
Milky Way

Three-Body Problem



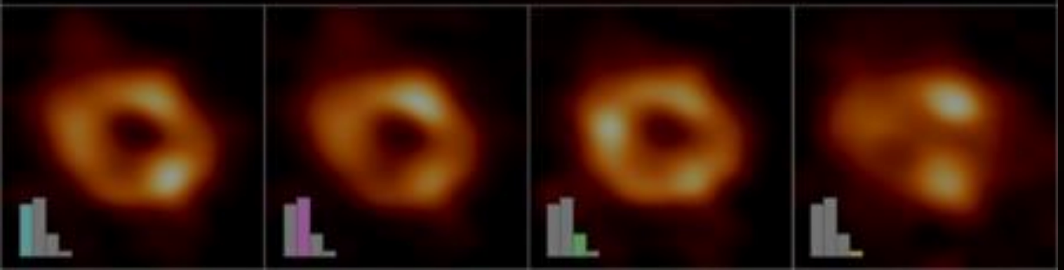
The three-body problem: The system becomes chaotic and highly unpredictable. It has no analytical solution (except for a few special cases) and its equations can only be solved numerically on a computer. They can turn abruptly from stable to unstable and vice versa.



Imaging Black Holes

CNN — Convolution Neural Network are class of deep learning algorithms which are quite efficient in recognizing real world objects. CNNs are the best neural nets for interpreting and understanding images.

To reconstruct an image, the EHT team developed computational imaging algorithms capable of making inferences to fill in the blanks in the data that had been gathered.



Special case…Star Clusters

Importance of studying Star Clusters

Star clusters are the basic building blocks of galaxies.

To understand the formation and evolution of galaxies it is essential to understand star cluster populations

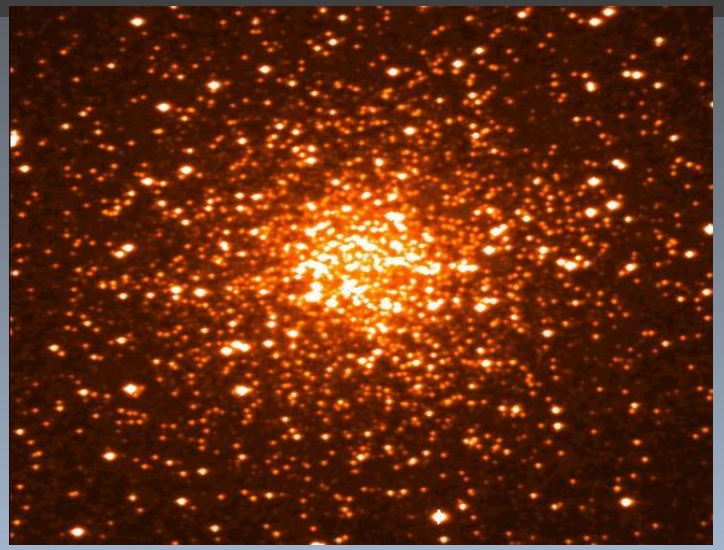
young associations

open clusters

globular clusters

Types of Clusters

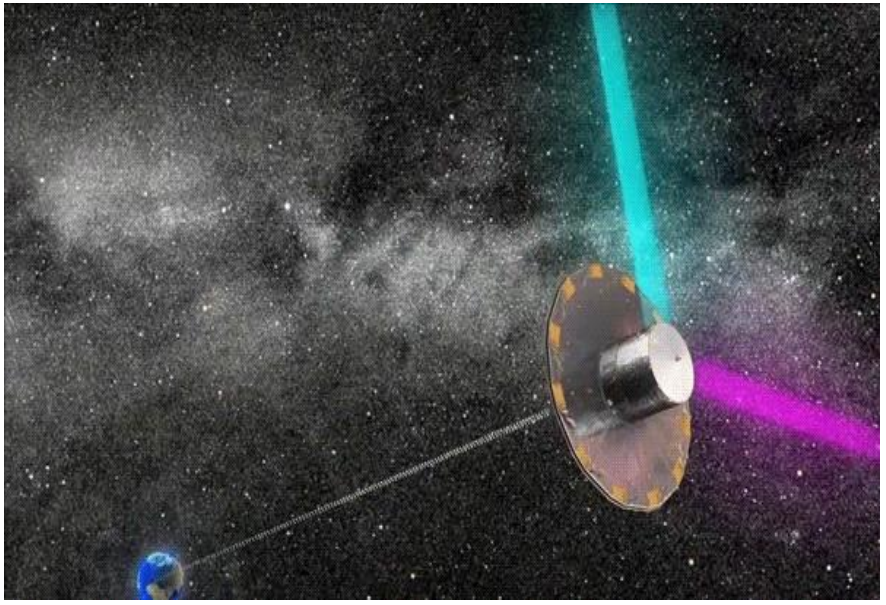
- Open Clusters
- Globular clusters
- Associations



Stars formed at the same time, distance, chemical composition…..differ only in mass.

Gaia

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



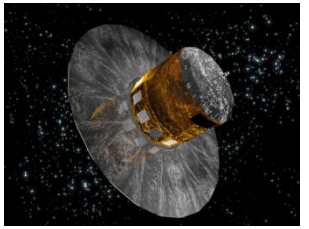
Galactic Archealogy!!! Imagine!!!

Membership?

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 is**

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



- *GAIA data is an important dataset to derive parameters of star cluster.*
- *We used GAIA DR2 to study 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*
- *We use membership data from Cantat-Gaudin et al (2018) based on GAIA DR2 as a training set.*
- *We used Random Forest (RF), which is a supervised classification method*

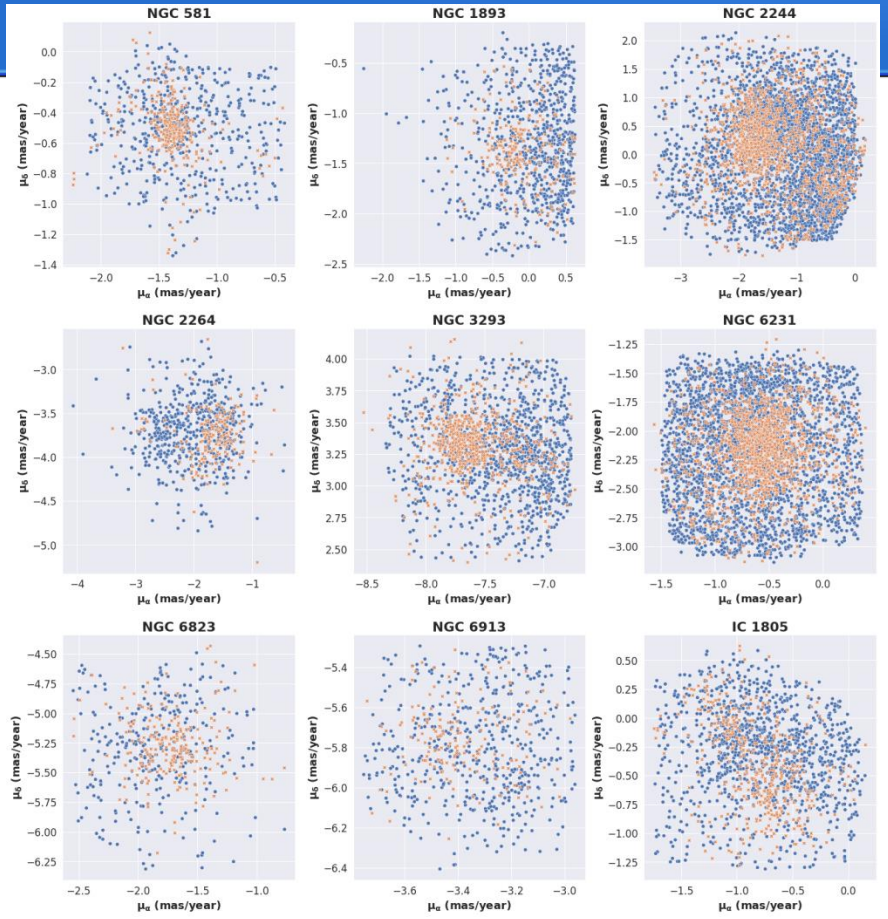
Eur. Phys. J. Spec. Top.
https://doi.org/10.1140/epjs/s11734-021-00205-x

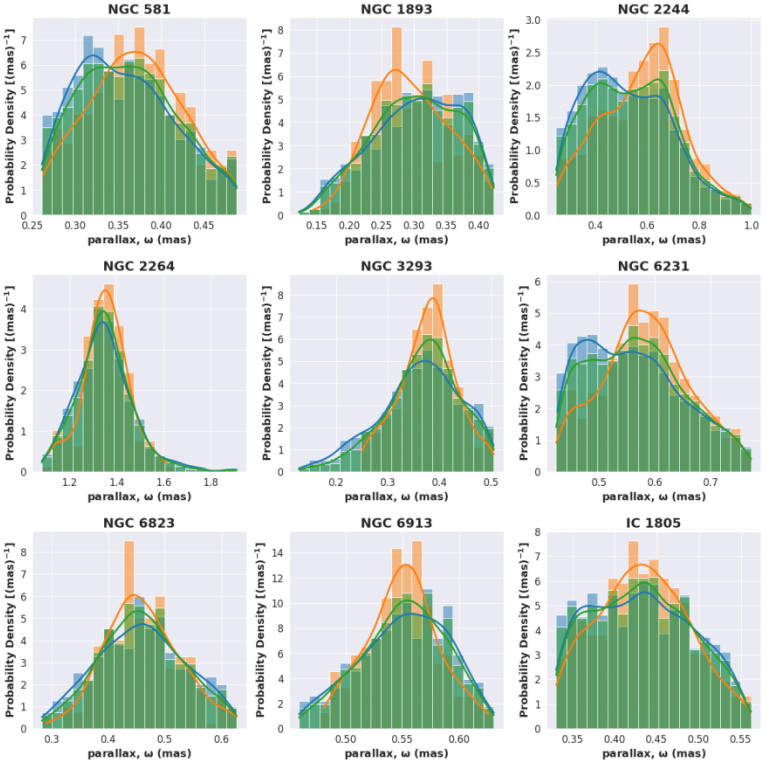
THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

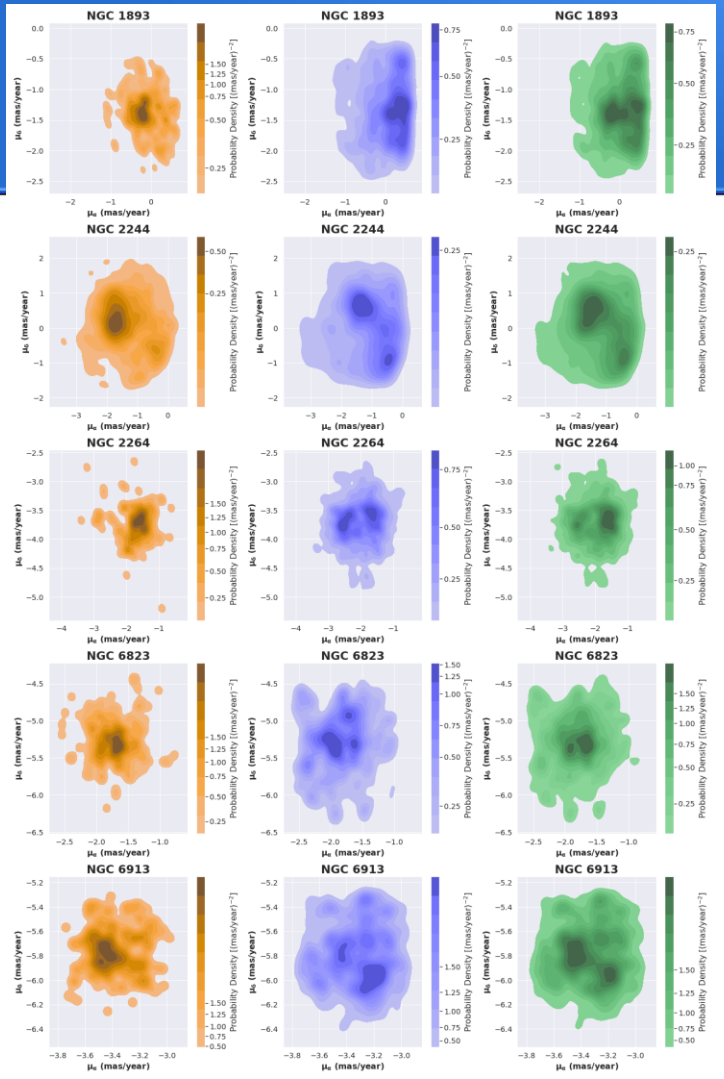
Membership of stars in open clusters using random forest with gaia data

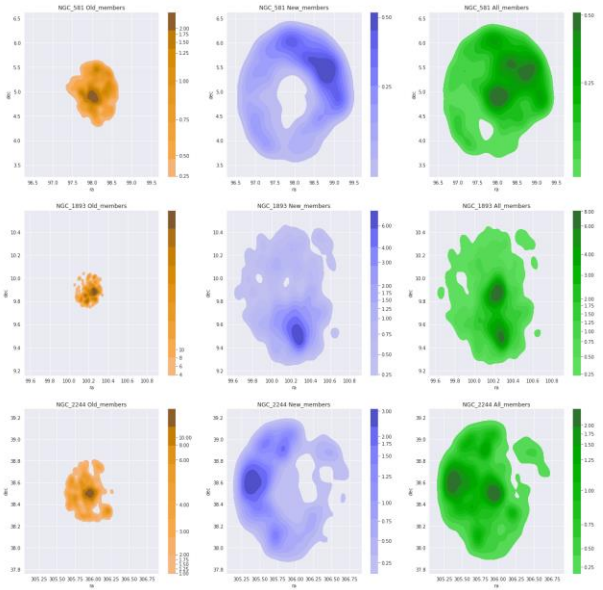
Proper Motion & Parallax plots

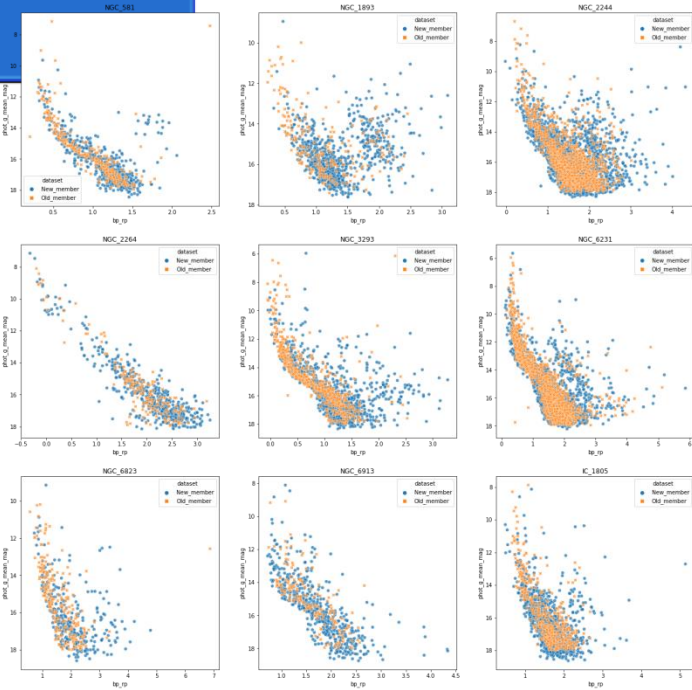




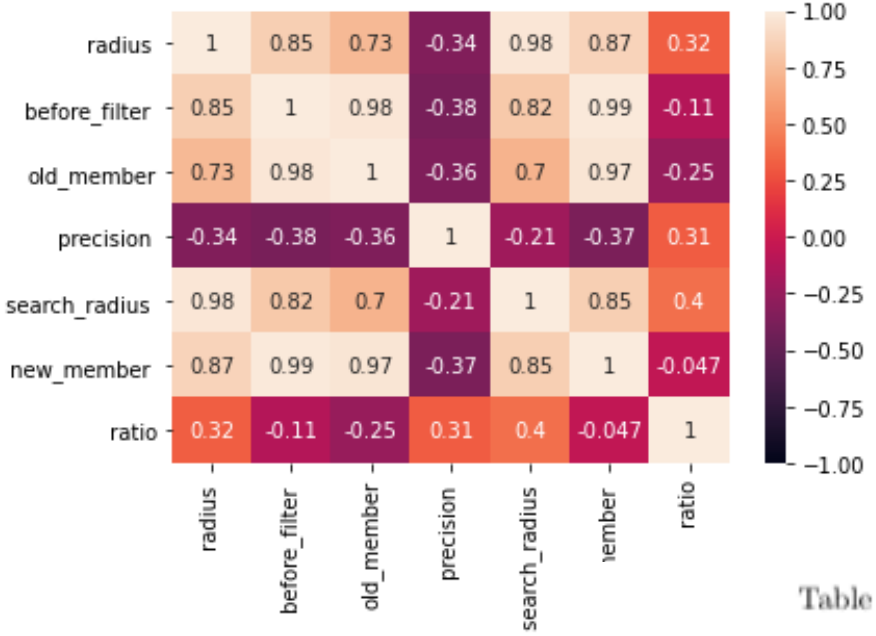
Plots, Color Magnitude Diagrams







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

Results

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

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

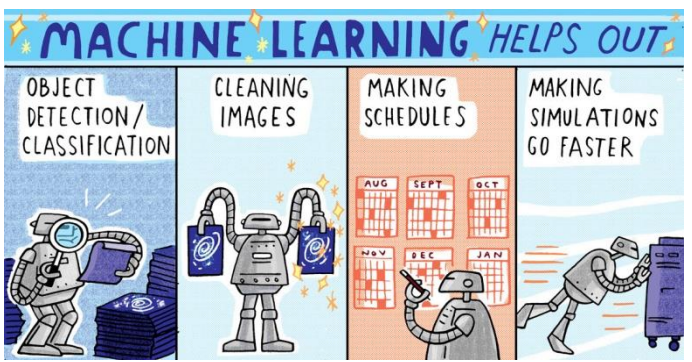
Regular Article

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.



Unsupervised Learning

- Unsupervised method (UM)

- Which UM is better? Is there any single UM which works well for all clusters or does it depends on the cluster?

Types of clustering algorithms

Centroid-based — uses Euclidean distance to assign every point to the nearest cluster center. Example: K-means

Connectivity-based — assumes that nearby objects (data points) are more related than far away objects. Example: Hierarchical Agglomerative Clustering (HAC).

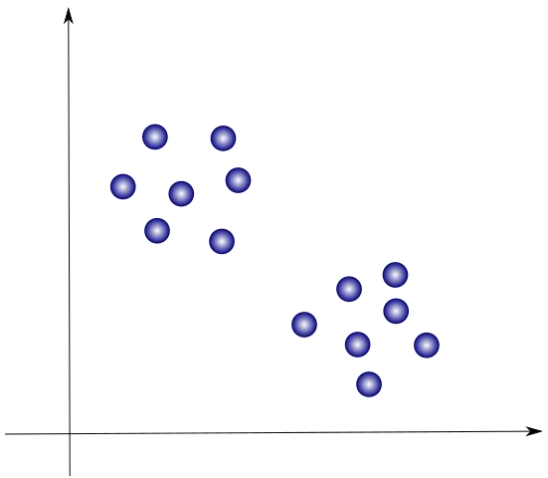
Density-based — defines clusters as dense regions of space separated by low-density regions. They are good at finding arbitrarily shaped clusters. Example: Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

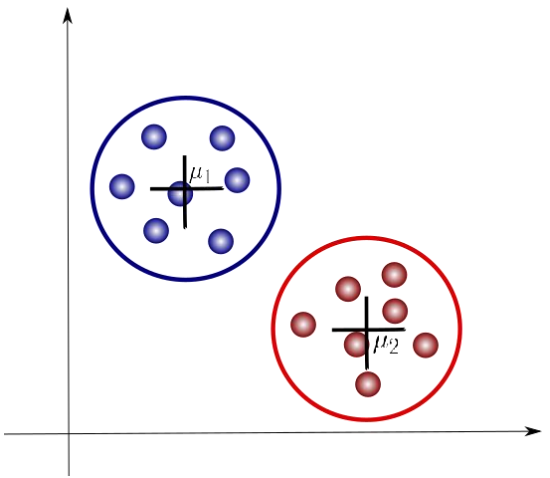
Distribution-based — assumes the existence of a specified number of distributions within the data. Each distribution with its own mean (μ) and variance (σ^2) / covariance (Cov). Example: Gaussian Mixture Models (GMM).

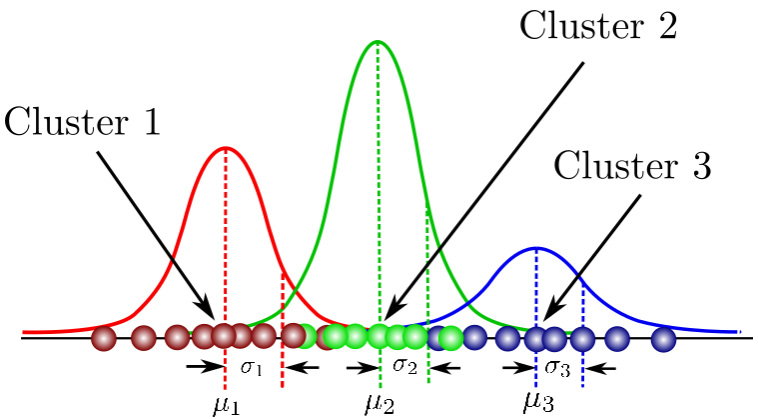
Clustering Algorithm: Gaussian Mixture Modelling/DBSCAN

- Does it work for all clusters?
- Field/Cluster ratio?

Guassian Mixture Modelling

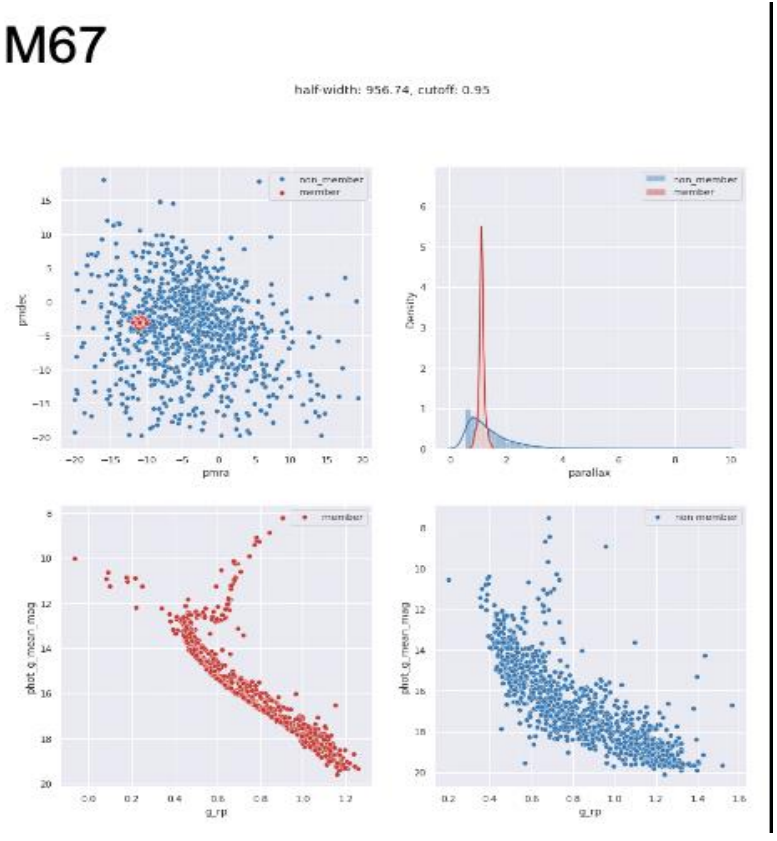






μ1 and μ2 are the centroids of each cluster

Using GMM model in Open Cluster Membership



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
NGC 2141	0.59	284	831	0.34
NGC 1893	0.51	592	169	3.50

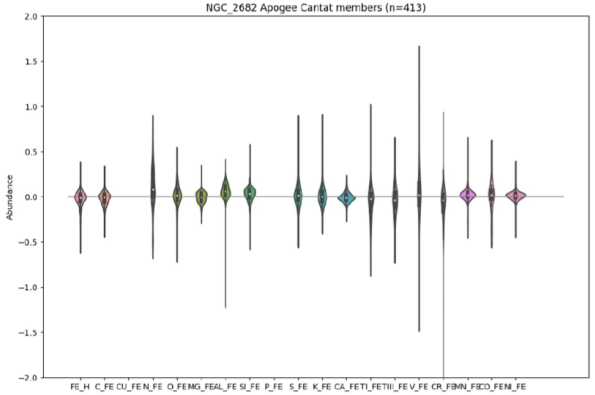
Md Mahmudunnobe, Priya Hasan, Mudasir Raja, S N Hasan, in prep

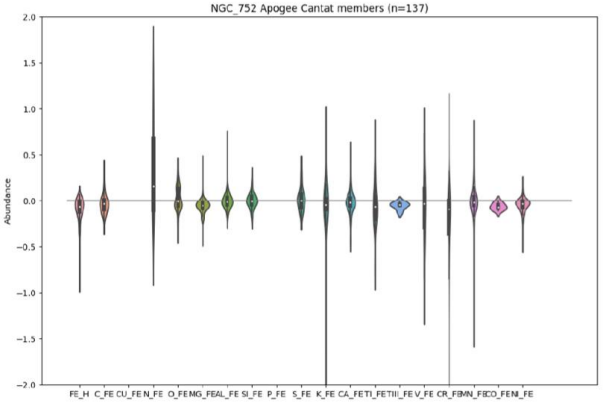
Modified Silhouette Score

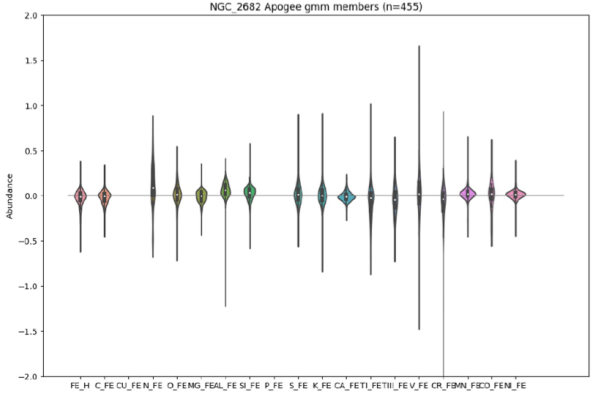
$$s = \frac{b - a}{\max(a, b)}$$

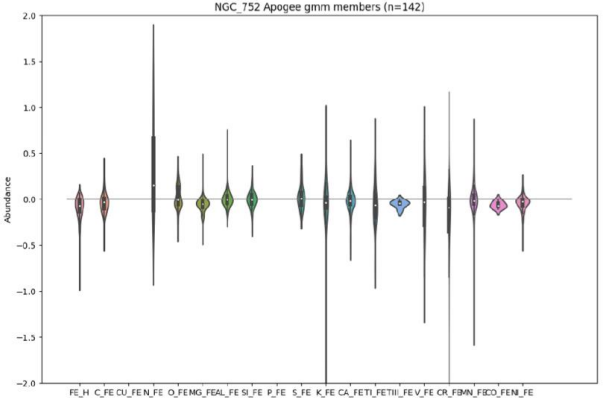
$$MSS = \frac{1}{k} \sum_{i=1}^{k} \frac{(\sigma_{i,field} - \sigma_{i,member})}{\max(\sigma_{i,field}, \sigma_{i,member})}$$

Spectroscopic Data: APOGEE and GALAH









ASteCA: M67

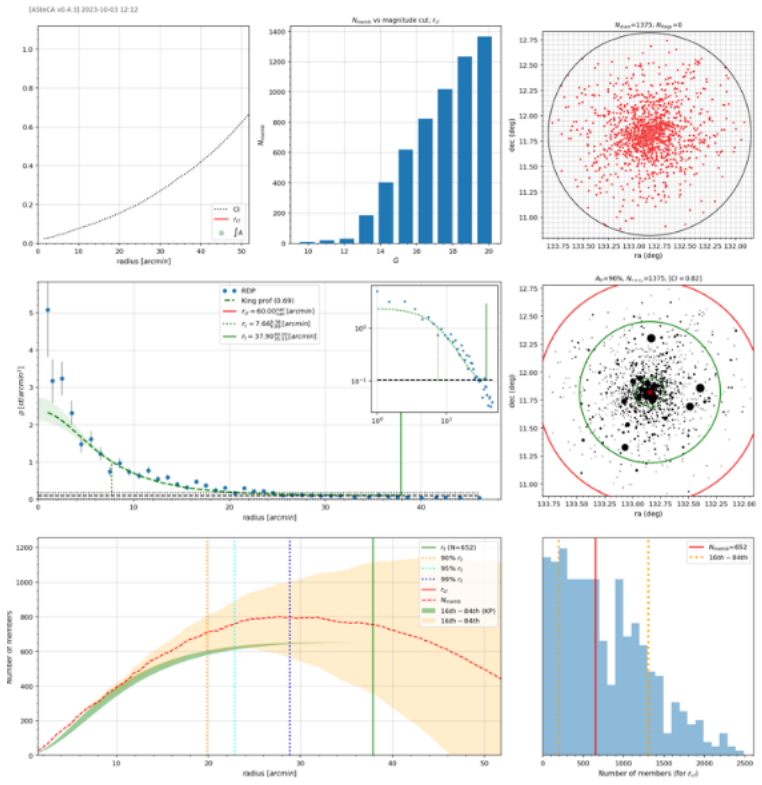


FIGURE 3.19: ASteCA plots of NGC 2682

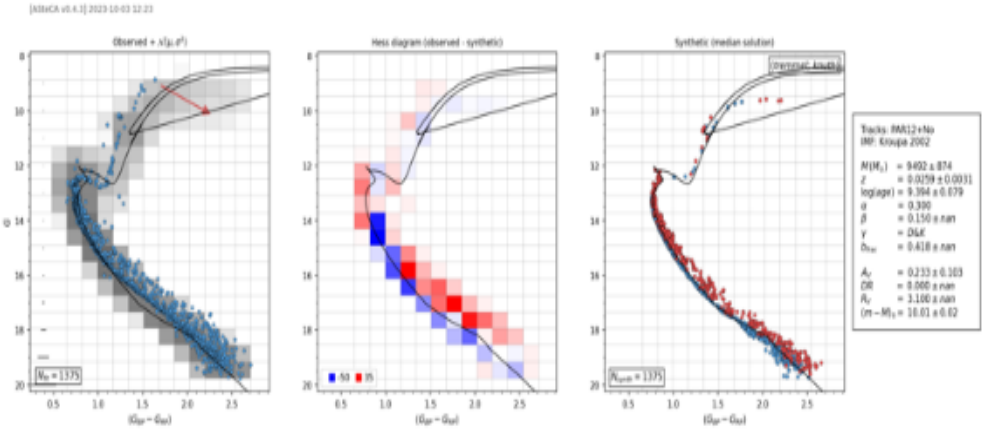


FIGURE 3.20: ASteCA CMD plots of NGC 2682

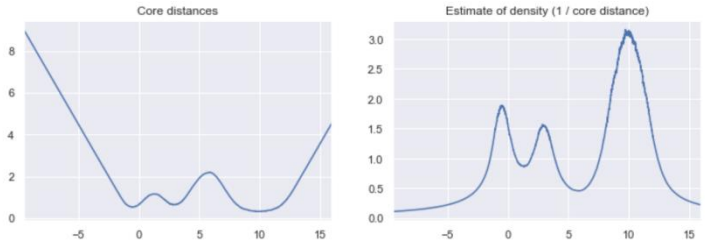
Results

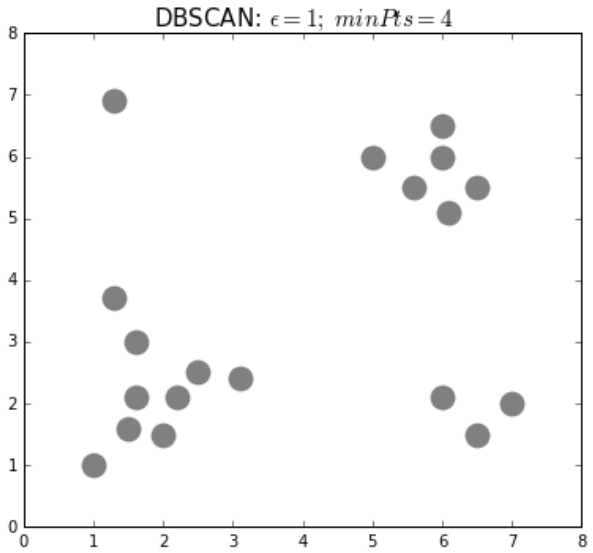
The sample of stars in clusters can be increased by a large factor, almost 1–2 times. This improves our accuracy in determining various parameters. The sizes of the studied clusters also increased with the increase in membership and we can study the outer regions of clusters.

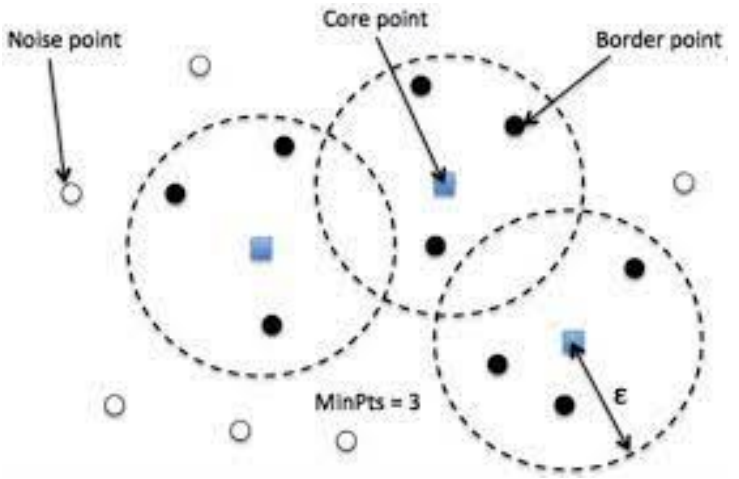
• As we have not used photometric data while estimating membership, we can identify variables, premain sequence stars (NGC 1893), as well as all other possible non main-sequence members of the cluster.

DBSCAN

Density-based spatial clustering of applications with noise







Membership determination in open clusters using DBSCAN Clustering Algorithm

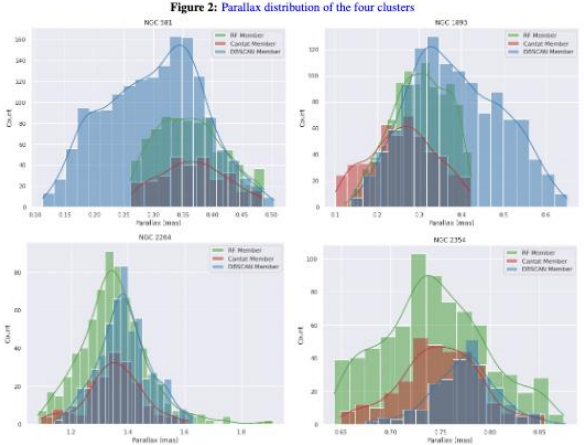
Mudasir Raja[1], Md Mahmudunnobe[2], Priya Hasan[1] and S N Hasan[1]

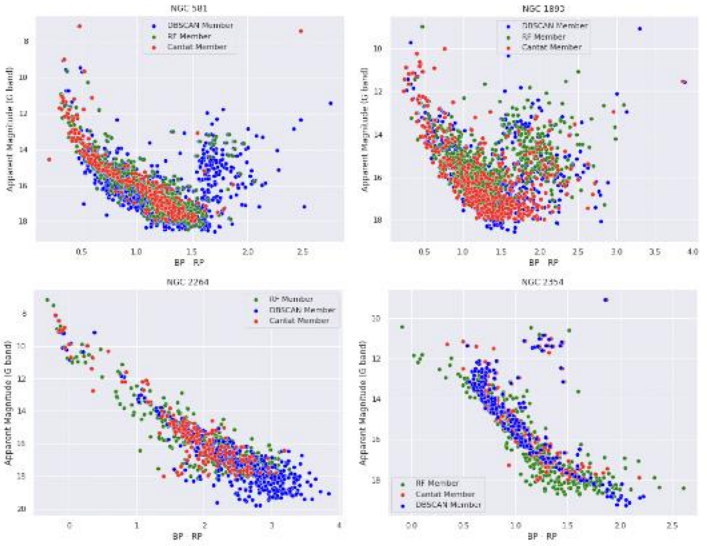
1. Maulana Azad National Urdu University Hyderabad,500032 2. Minerva University, California, USA

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, Priya Hasan, , S N Hasan, in prep

M67 (NGC2682)

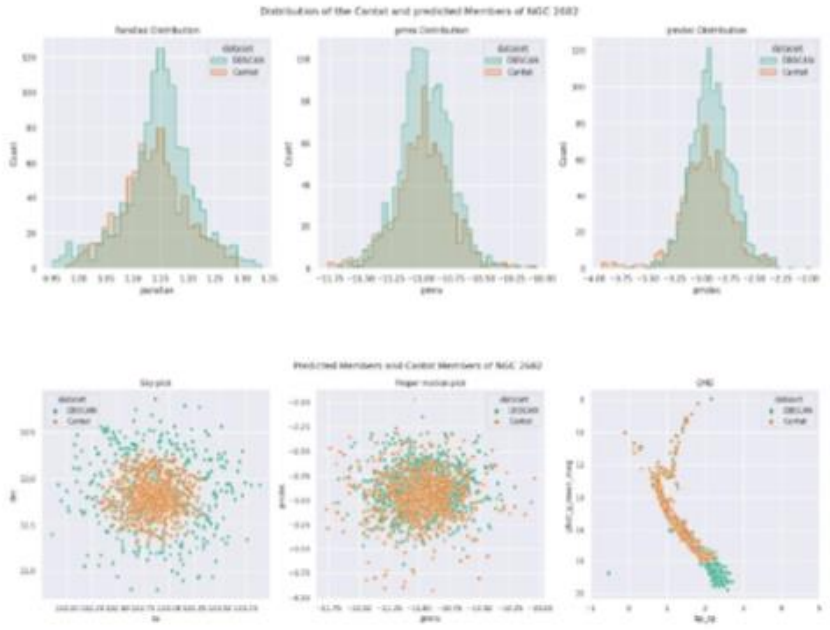


FIGURE 4.3: Revised members of NGC 2682 CG(orange) and DBSCAN (green).

APOGEE/GALAH (M67)

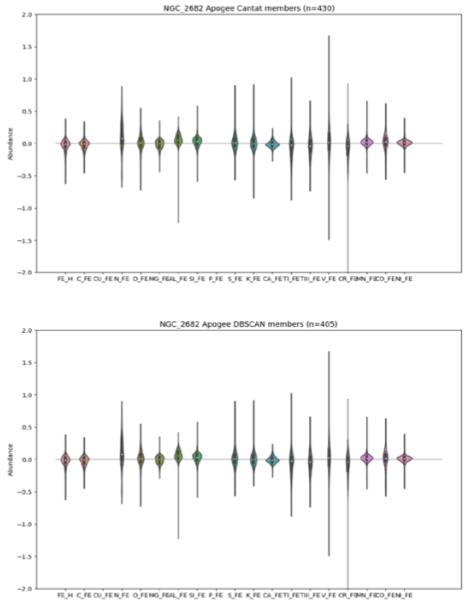


FIGURE 4.12: Chemical abundances of members from APOGEE for NGC2682 (a) Upper plot (Cantat-Gaudin et al., 2018) (b) Our results

ASteCA: M67

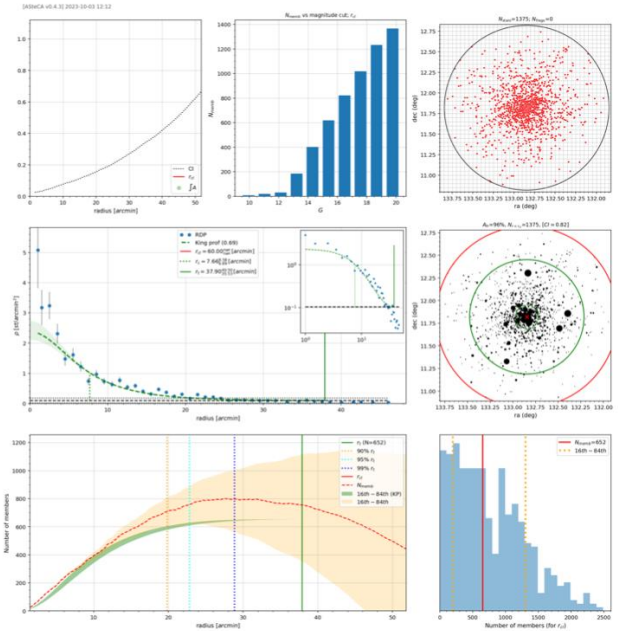


FIGURE 4.22: ASteCA plots of NGC 2682

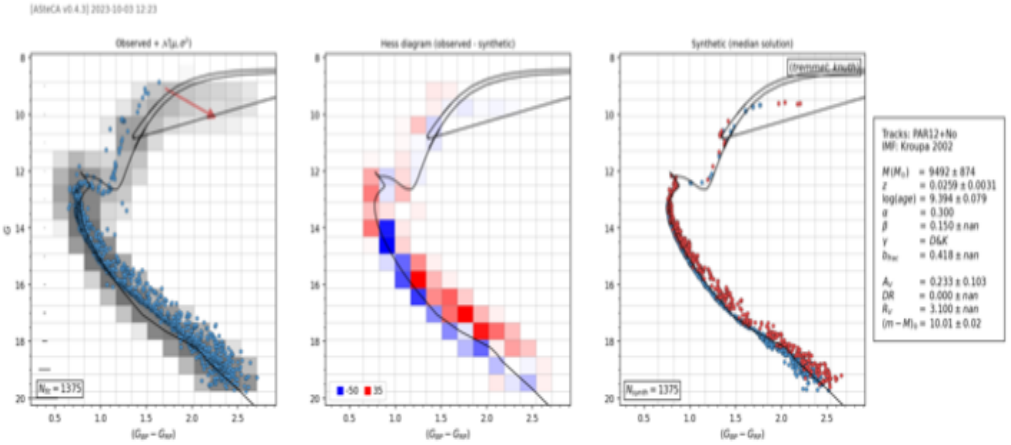


FIGURE 4.23: ASteCA CMD plots of NGC 2682

Results

Cluster	R_{cl} (arc min)	ASteCA log(age)	ASteCA d	Cantat log(age)	Cantat d
NGC 2682	64.65	9.39	1000	9.63	889
NGC 2244	83.55	7.34	1061	7.1	1478
NGC 3293	44.51	7.4	1459	7.01	2710
NGC 6913	46.17	7.8	1318	7.34	1608
NGC 7142	116.49	9.6	2238	9.49	2406
NGC 6231	77.17	7.45	1032	7.14	1475
NGC 2243	22.95	9.89	3311	9.64	3719

TABLE 4.4: ASteCA Parameters vs parameters from (Cantat-Gaudin and Anders, 2020) using DBSCAN

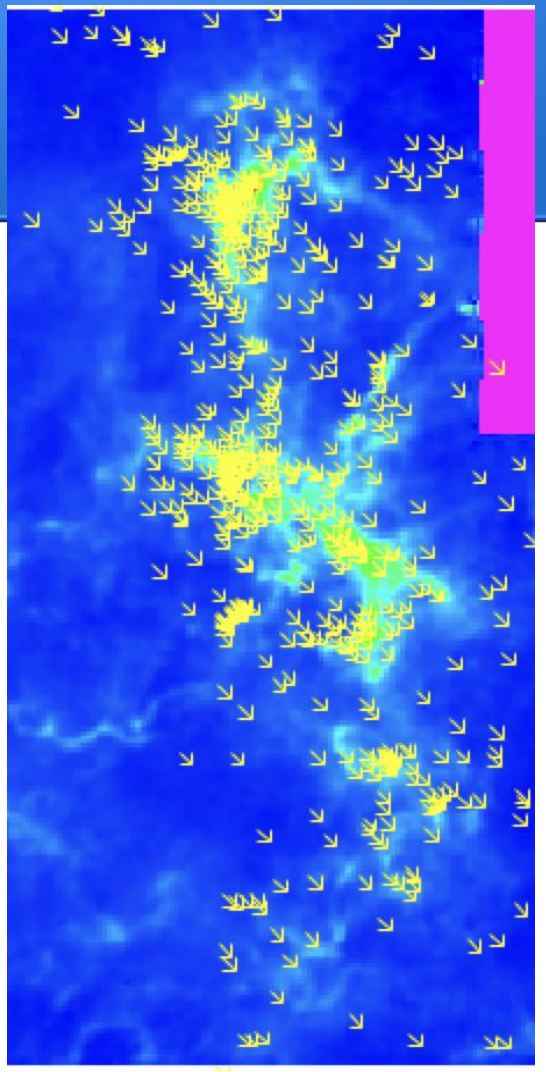
.

Kinematics and Structure in Serpens

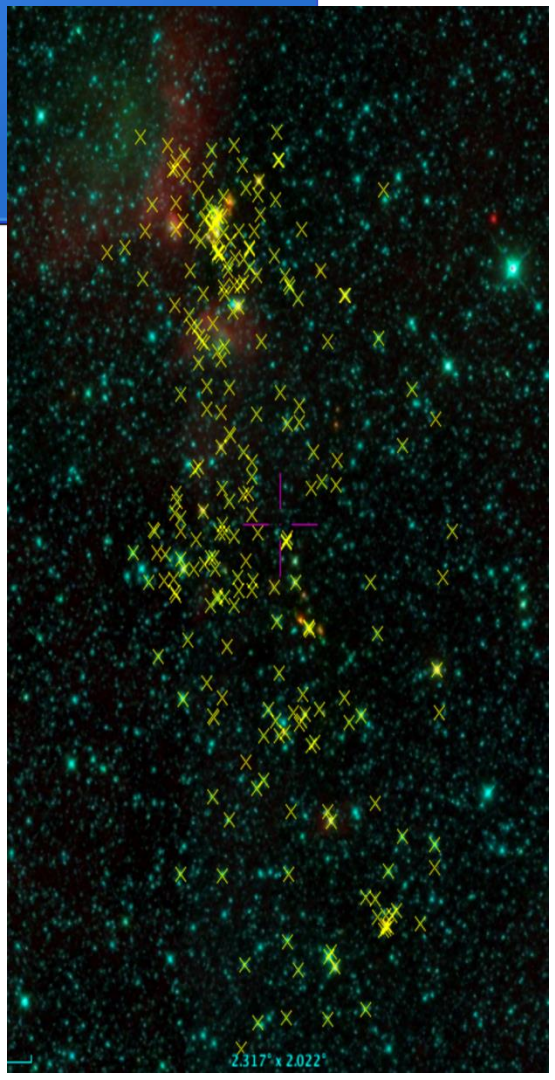
- Use available YSO catalogs (Xray, IR….)

- Find parameters
- Download Gaia data within parameters

- DBScan to find members
- 2-3 times increase
- Repeat with OPTICS, HDBScan



YSO Sample



Gaia counterparts

Modus Operandi

YSO Sample (IR,Xray….)

Cross-Match with Gaia EDR3 Data

Get parameters

Make a control sample, Query EDR3 data in 2deg radius

Use ML: DBSCAN, OPTICS, HDBSCAN to identify members

Plot on Extinction map

Obtain parameters

Identify YSO class using 2MASS +WISE

Gaia Sample (1196)

- Query 2 d radius data, RUWE < 1.4, RPlx > 10,
- 5 > pmRA > 1
- 11.6 < pmDE < -6
- 4.2 > Plx > 0
- 1196 stars
- RV= -5. 08 km/s (66 stars)
- XYZ, pmRA, pmDE

Results

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

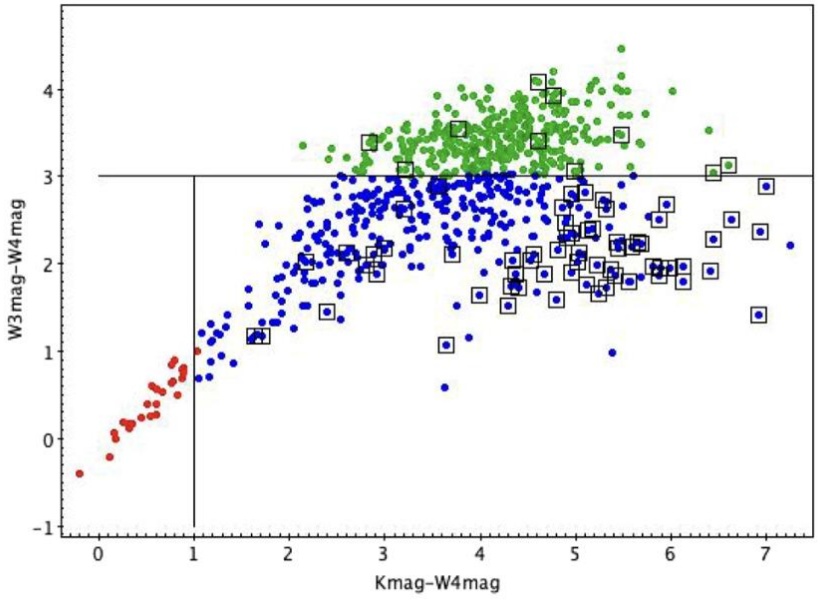


Figure 4. YSO classification using 2MASS and WISE: The class II stars are in green, class III in blue and photospheres in red. The stars from the control sample are the black squares.

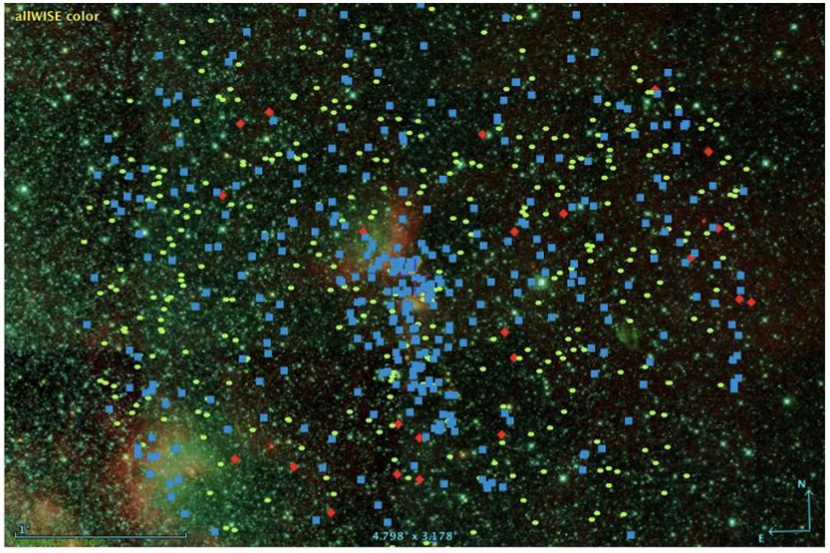


Figure 5. YSO distribution using 2MASS and WISE colors, where class II is green ovals, class III is blue squares and photospheres is red rhombuses.

Issues in Machine Learning

- What algorithms can approximate functions well and when
- How does the number of members in training setsinfluence accuracy
- Problem representation / feature extraction
- Intention/independent learning
- Integrating learning with systems
- What are the theoretical limits of learnability
- Transfer learning
- Continuous learning

Thank you