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

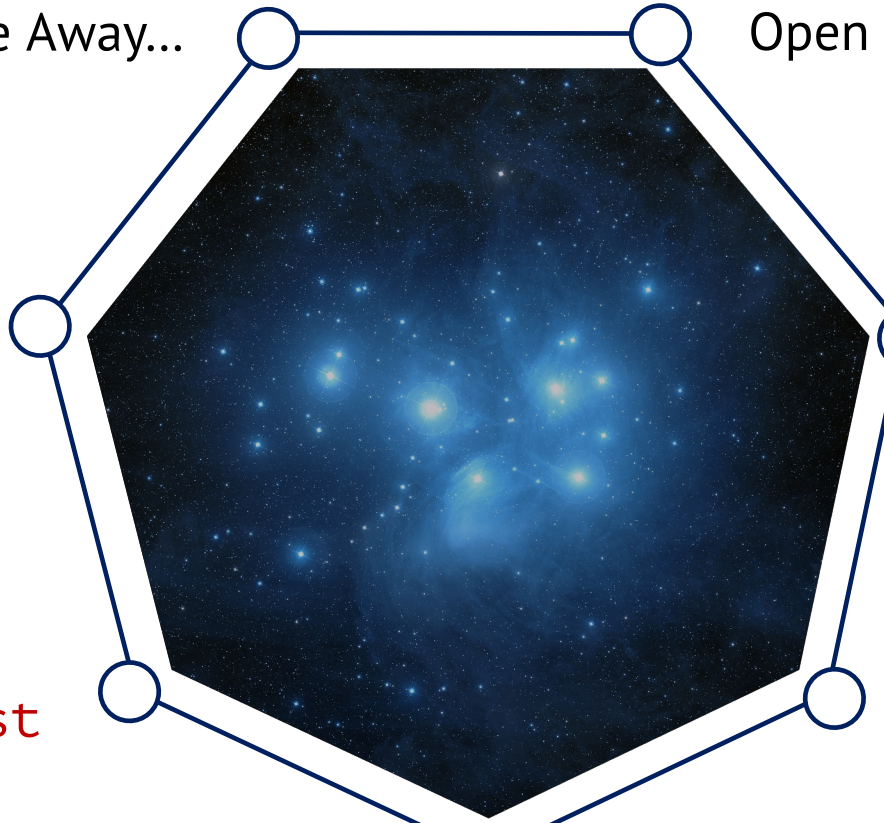
Mahmudunnobe[1] et. al. 2021

Presenter:

Bhavesh Rajpoot[2]

[1]Minerva Schools at KGI, San Francisco, CA 94103, USA
[2]Department of Physics and Astronomy, Universität Heidelberg, Germany

Outline



To Take Away…

Open Clusters

Results

Methods

`Random Forest` in Action

Crash Course on `Random Forest`

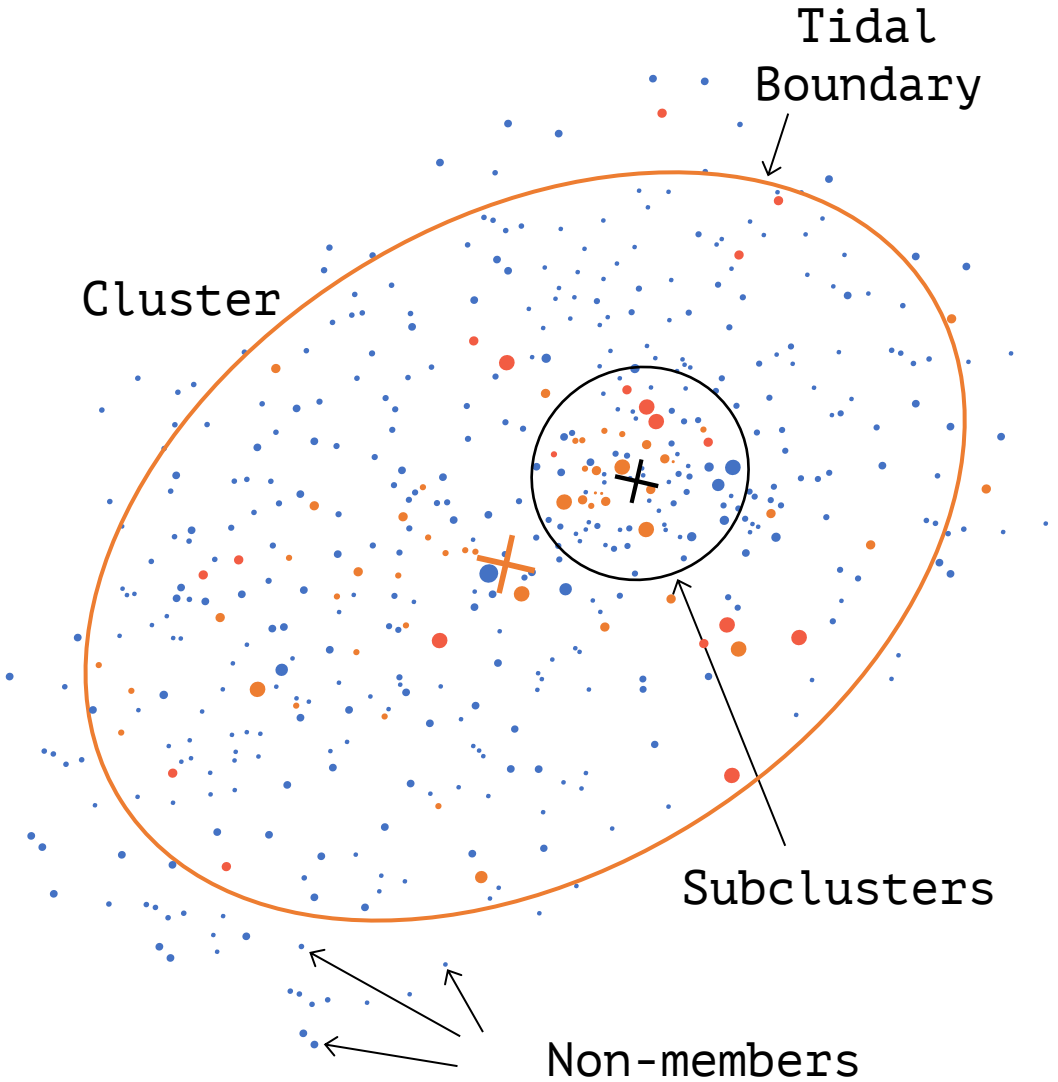
Sample Selection

Open Clusters

- Ensemble of $\sim 10^2 - 10^4$ stars
- Member properties:
  - *loosely* gravitationally bound
  - irregularly distributed

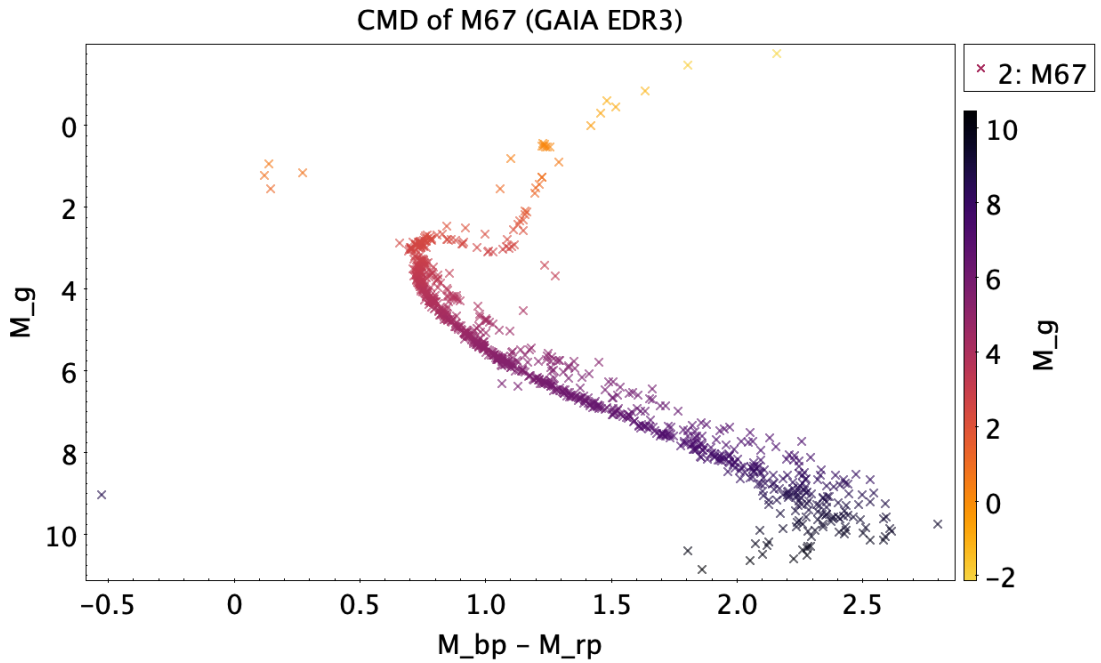
M45, NASA/ESA/Caltech

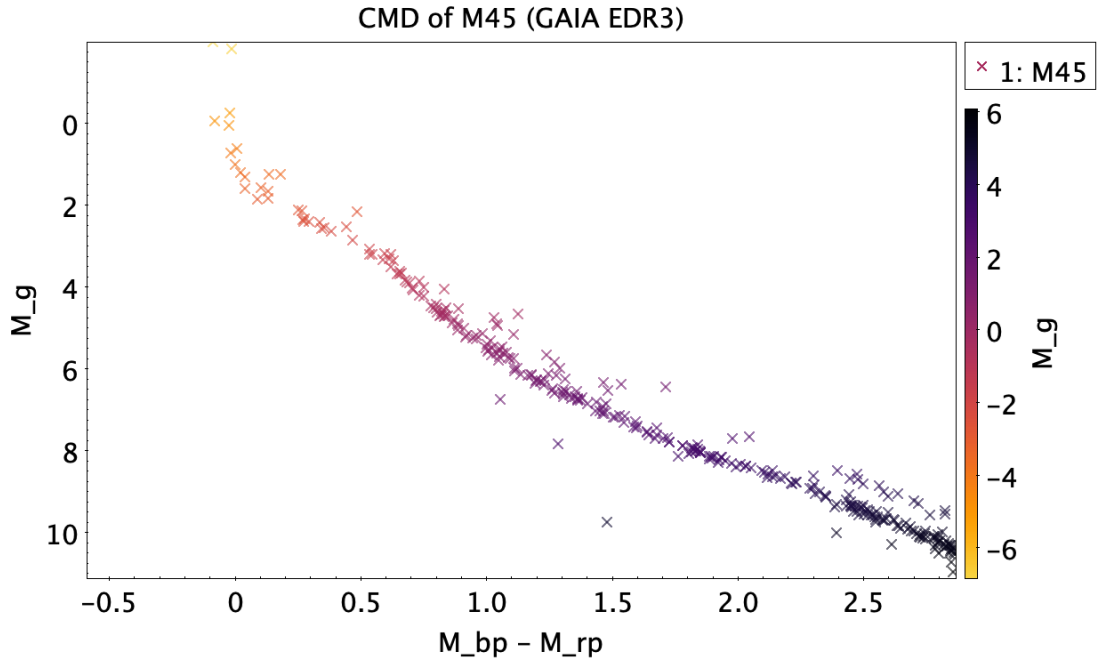
Merging Clusters in R136, NASA

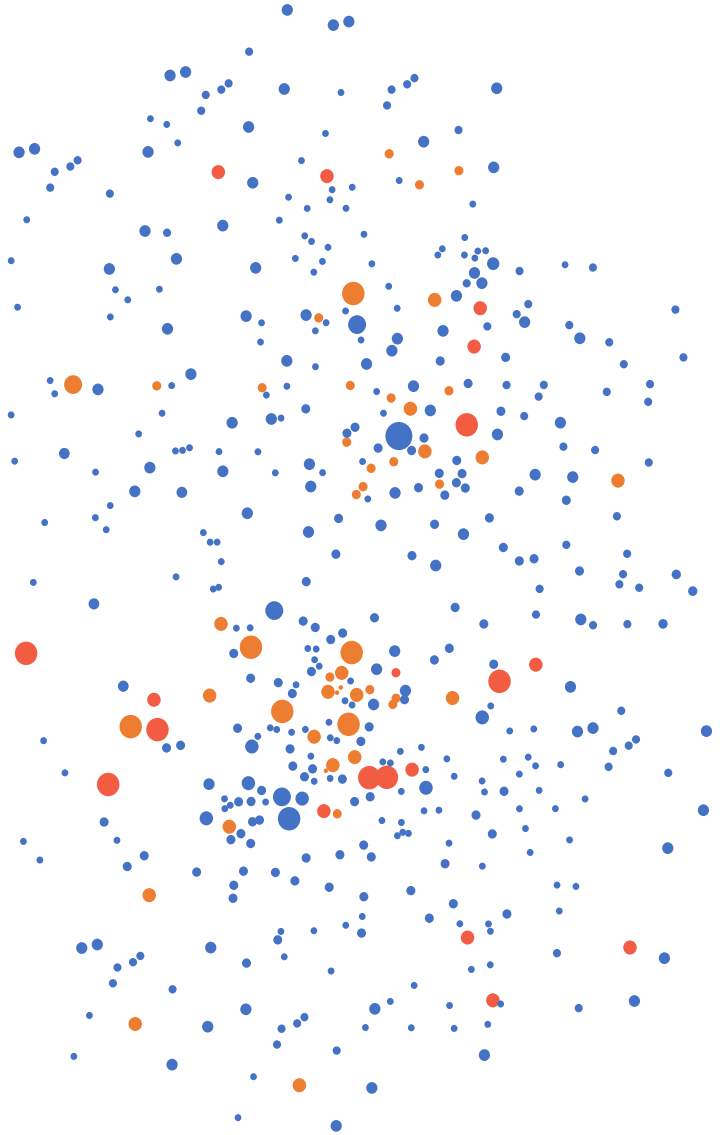


Open Clusters

- Member properties:
  - young with abundant metals
  - same ages but differ in mass
  - moves with common velocity

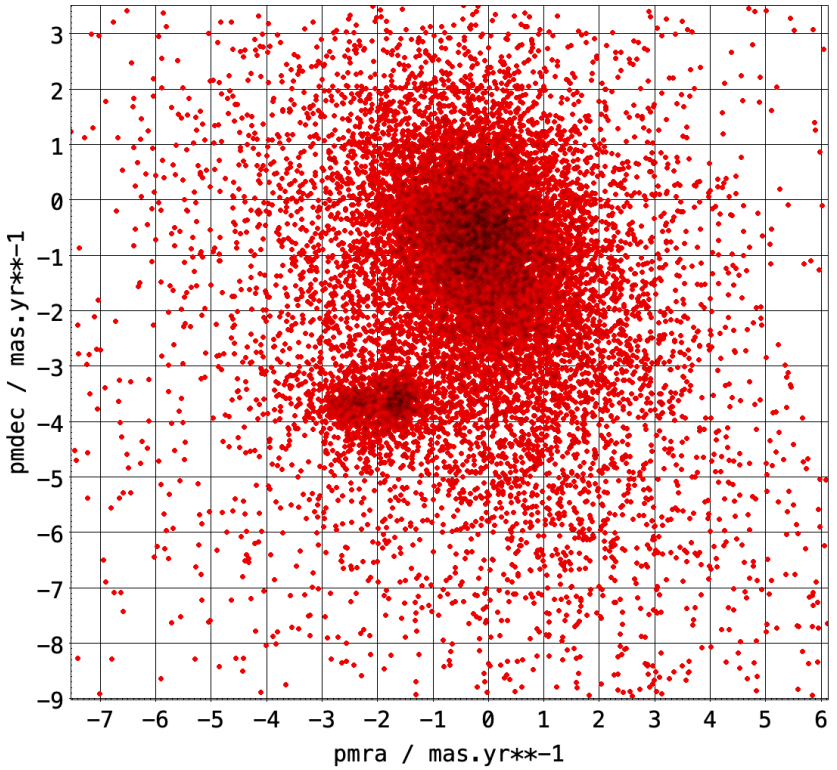




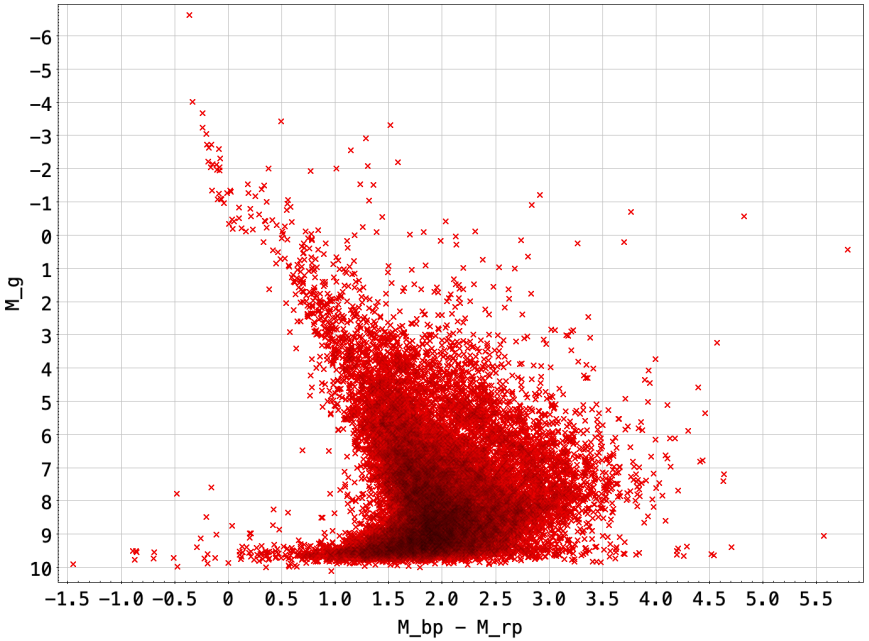


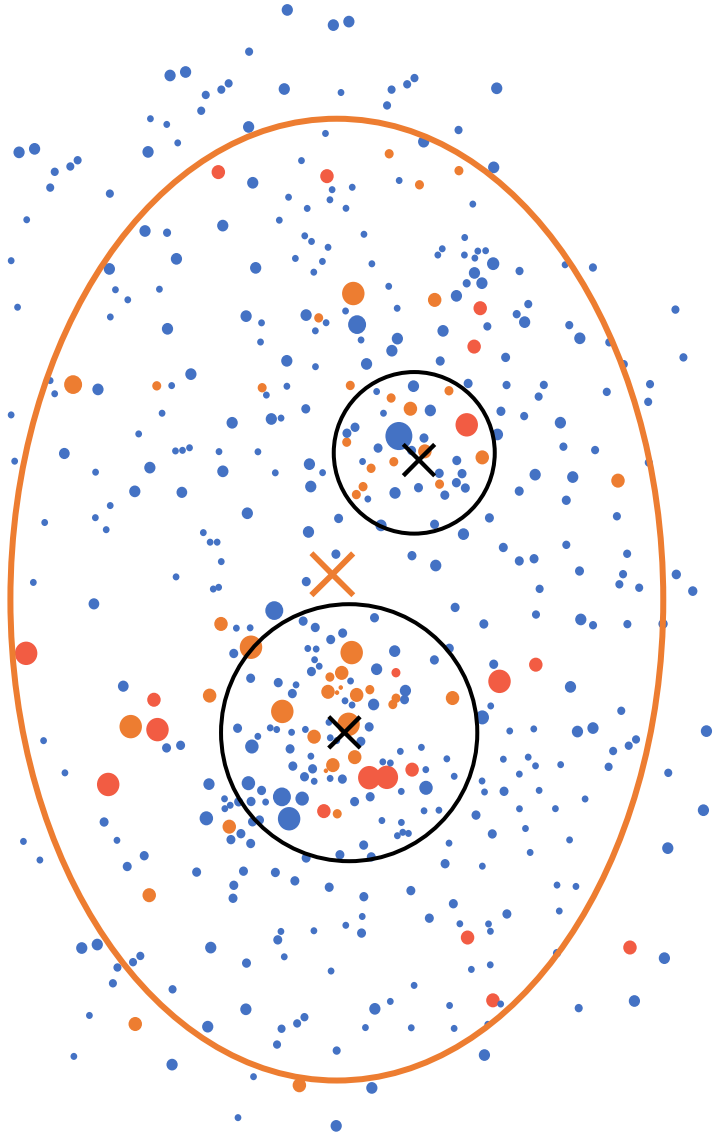
Position Space

Proper Motion Space



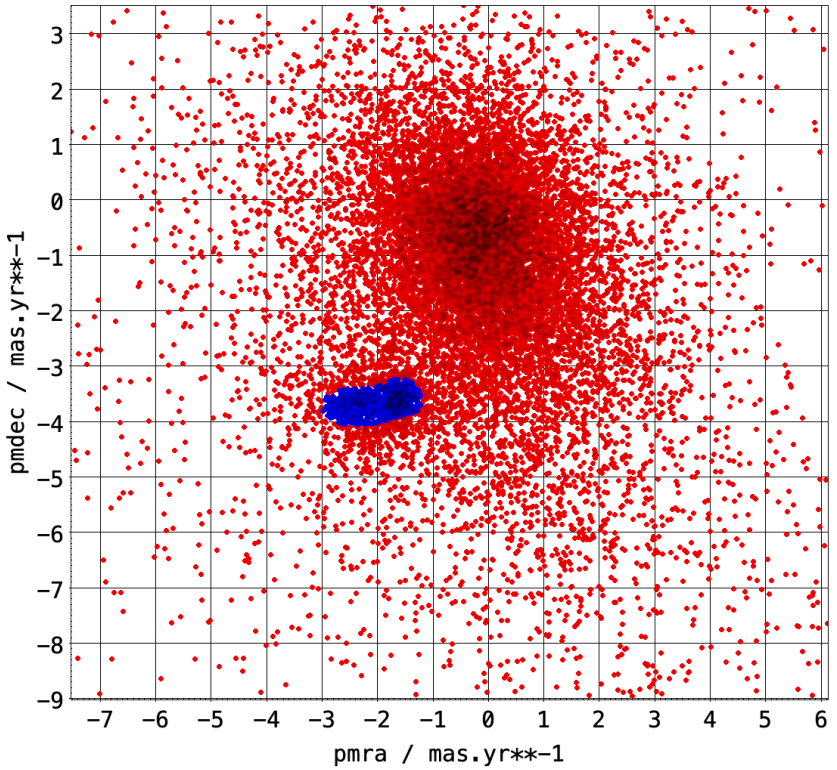
Color Magnitude Diagram



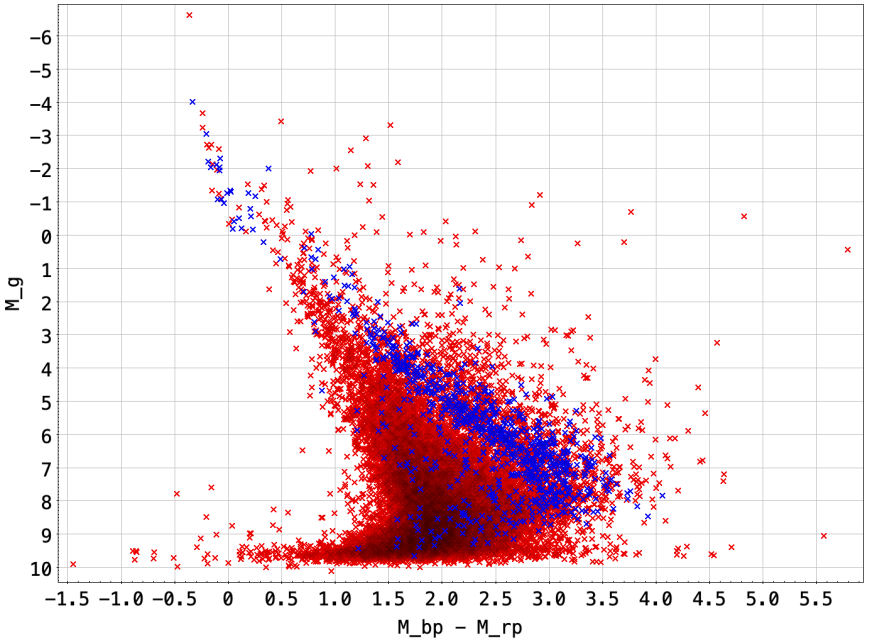


Position Space

Proper Motion Space



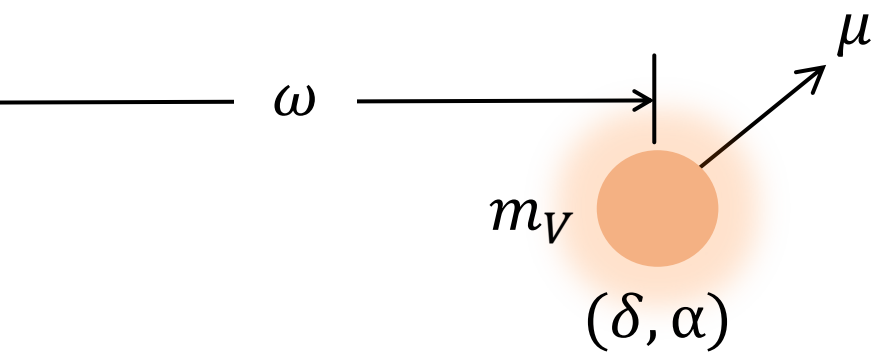
Color Magnitude Diagram

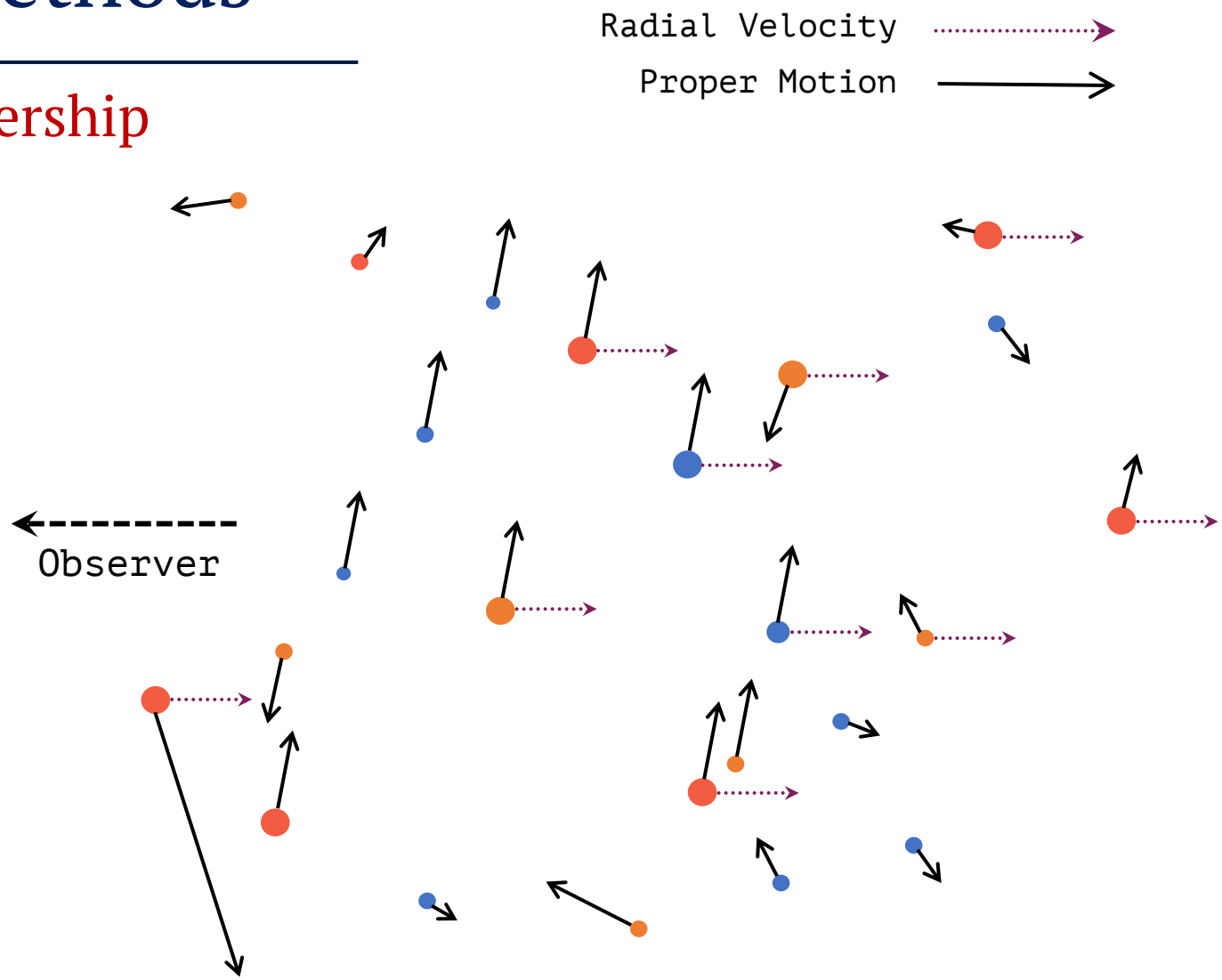


| Legend: Field Stars Members |

Methods

- Main Goal: Estimating Cluster Membership





Star Field with motion represented

Astrometric			Photometric	
Position	Proper Motion	Parallax	Magnitude	Reddening
δ, α	μ_δ, μ_α	ω	B, V, R, I	$E(B-V)$

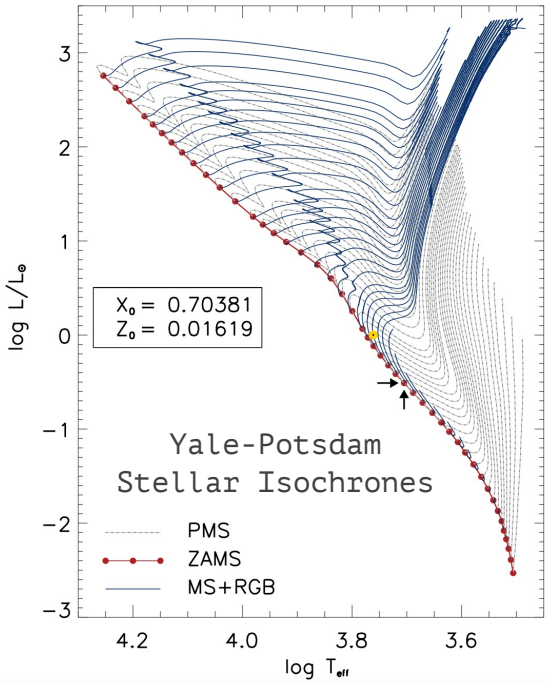
Methods

Model Bound

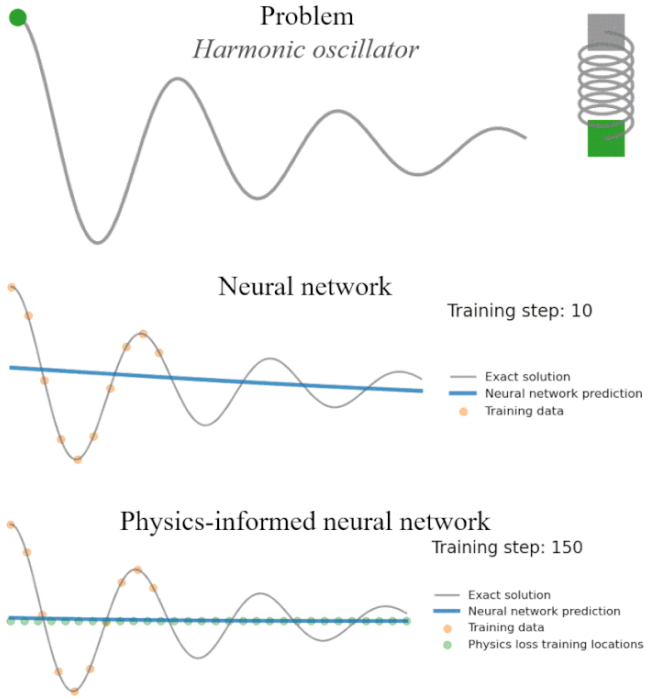
Model Adjusted

Model Free

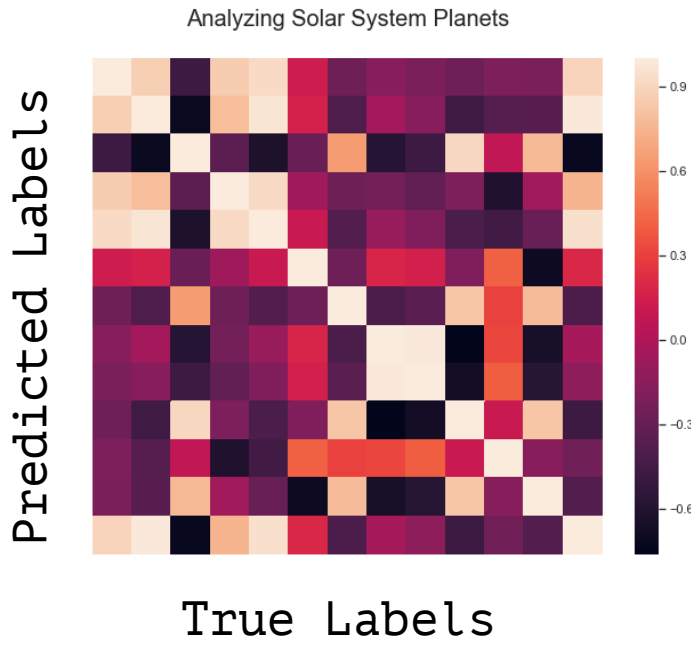
Process Driven
(Purely Theory Based)



Physics based & Data Driven



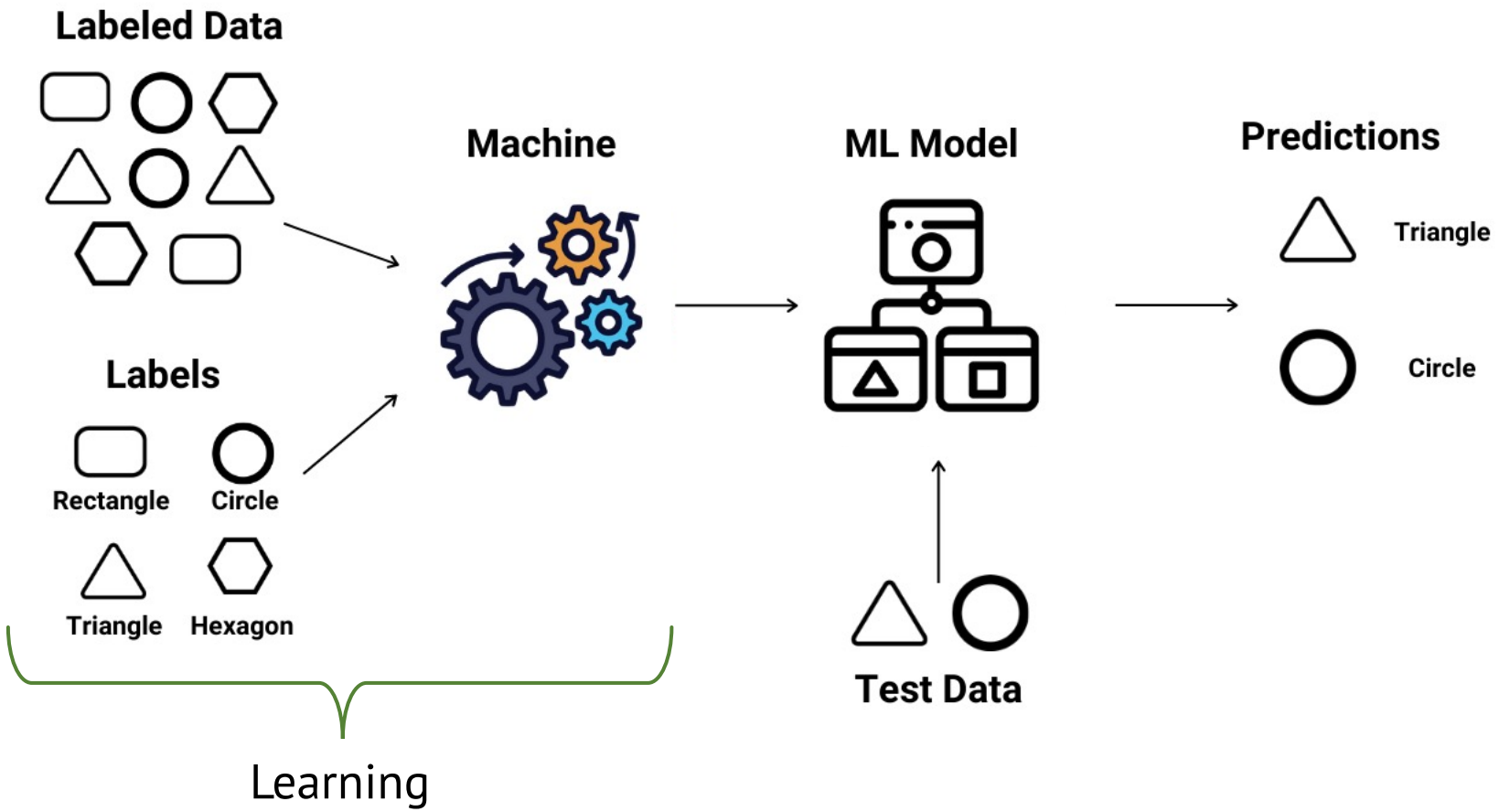
Data Driven
(Purely Data Based)



Crash course to Random Forest

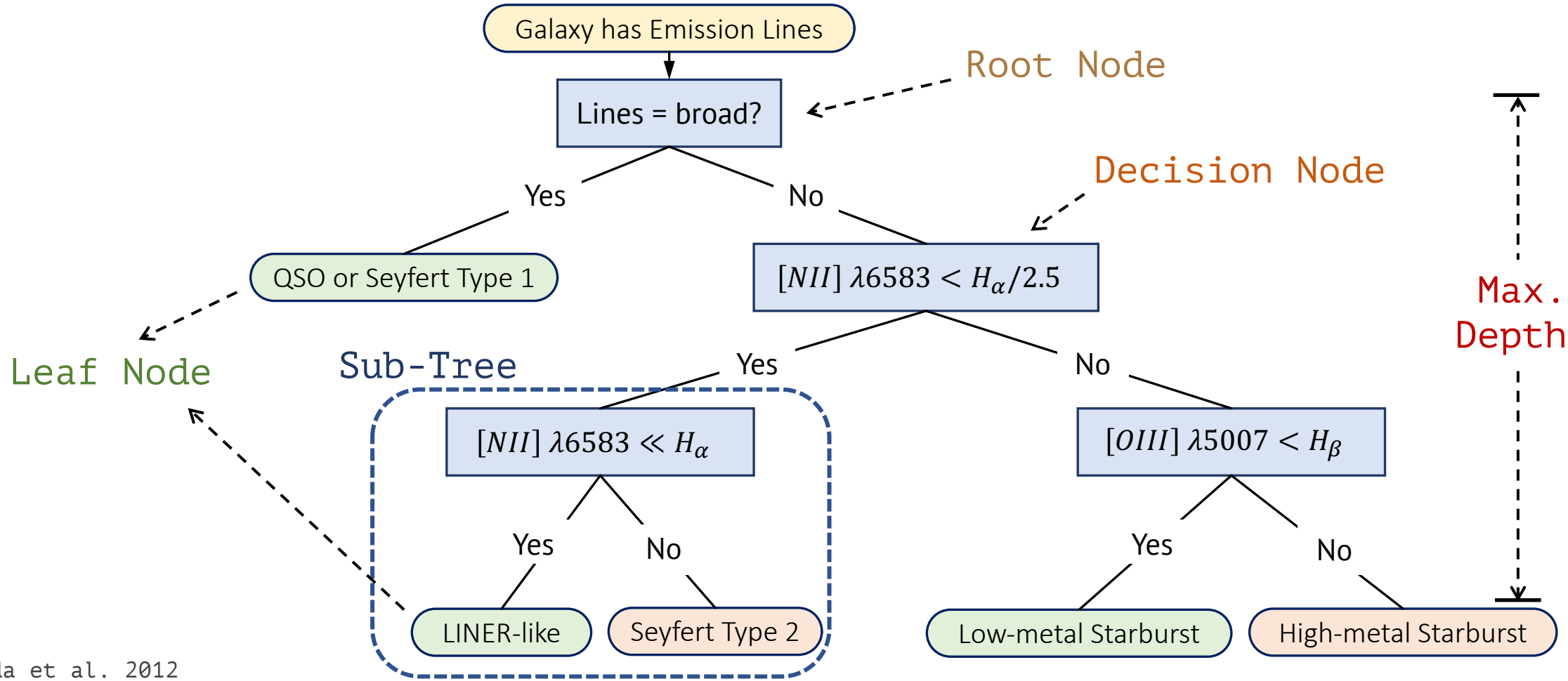
- Supervised Machine Learning

Features		Labels
No. of Sides	**Length of Sides**	**Name**
3	Equal	Triangle
6	Equal	Hexagon
5	Equal	Pentagon
4	Equal	Square
4	Unequal	Rectangle
0	-	Circle



Crash course to Random Forest

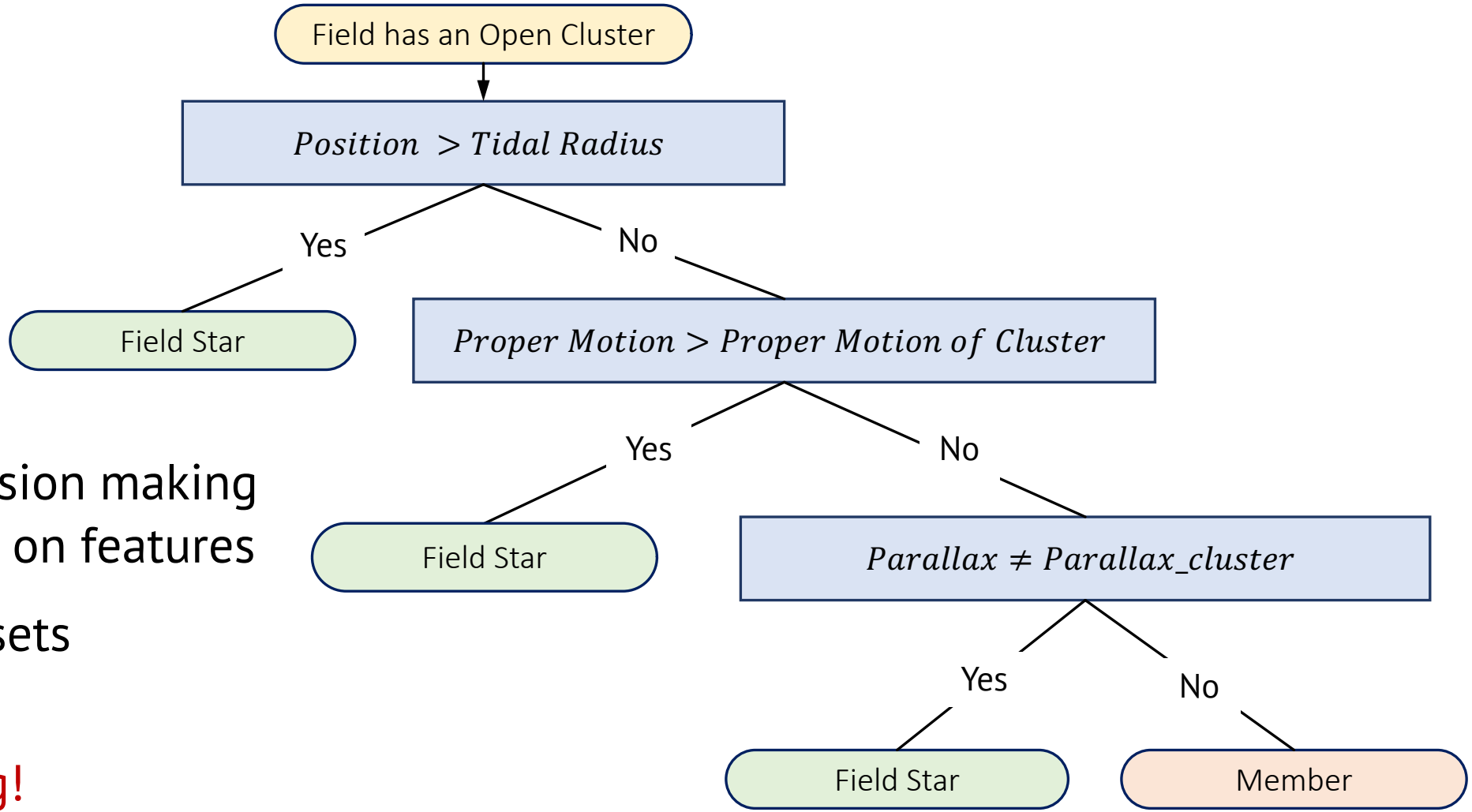
- Decision Trees



Sánchez-Almeida et al. 2012

Crash course to Random Forest

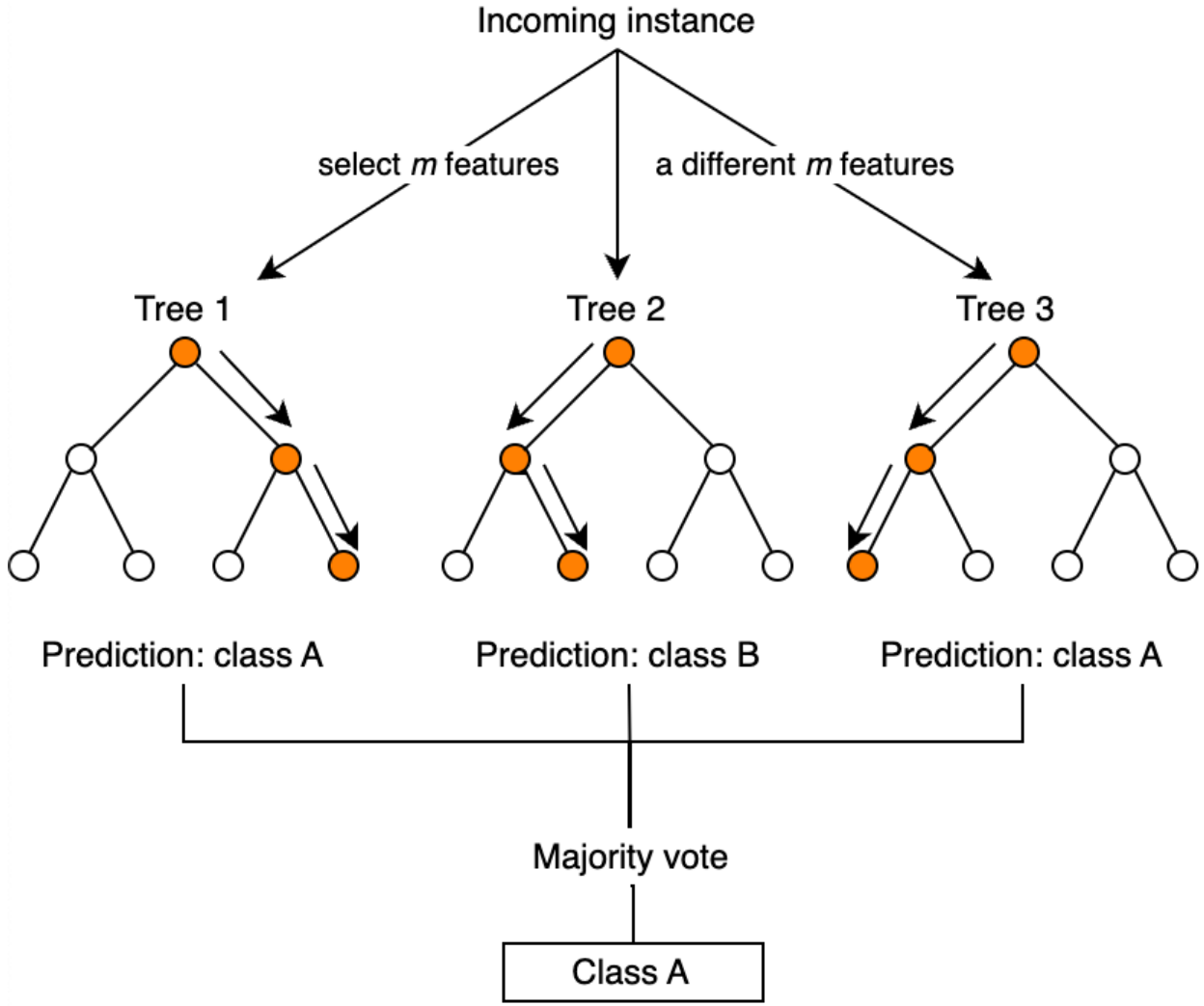
- Decision Trees



- Mirrors human decision making
- Decisions are made on features
- Handles large datasets
- Uses white box
- Prone to Overfitting!

Crash course to Random Forest

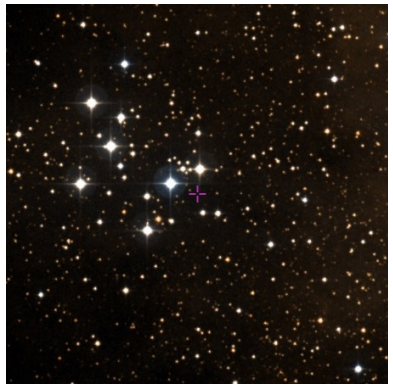
- Random Forest (RF)
  - Ensemble learning method
  - Uses '*Feature Bagging*'
  - Reduced risk of '*Overfitting*'
  - Handles large, high-dimensional dataset
  - Very short execution time
  - Strongly depends upon *Training Set*



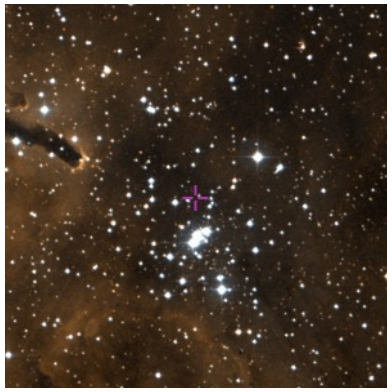
Sample Selection

NGC 2244

NGC 6913

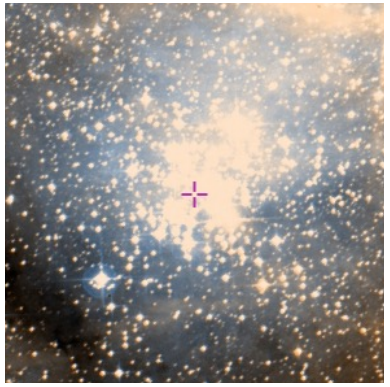


NGC 6823

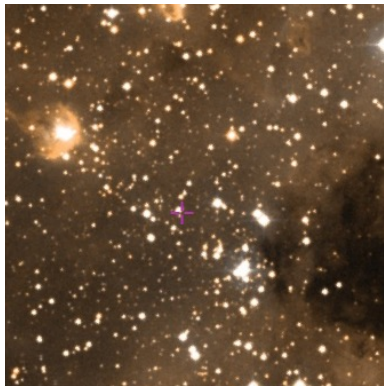


NGC 581

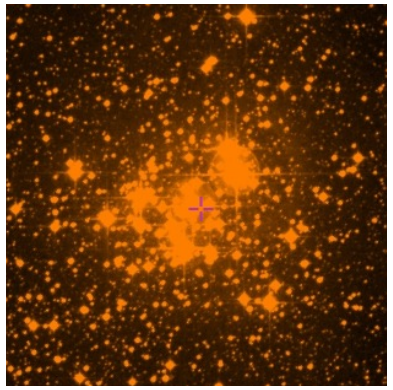
NGC 3293



NGC 1893

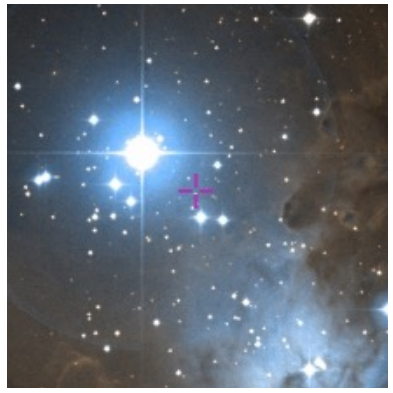


NGC 6231

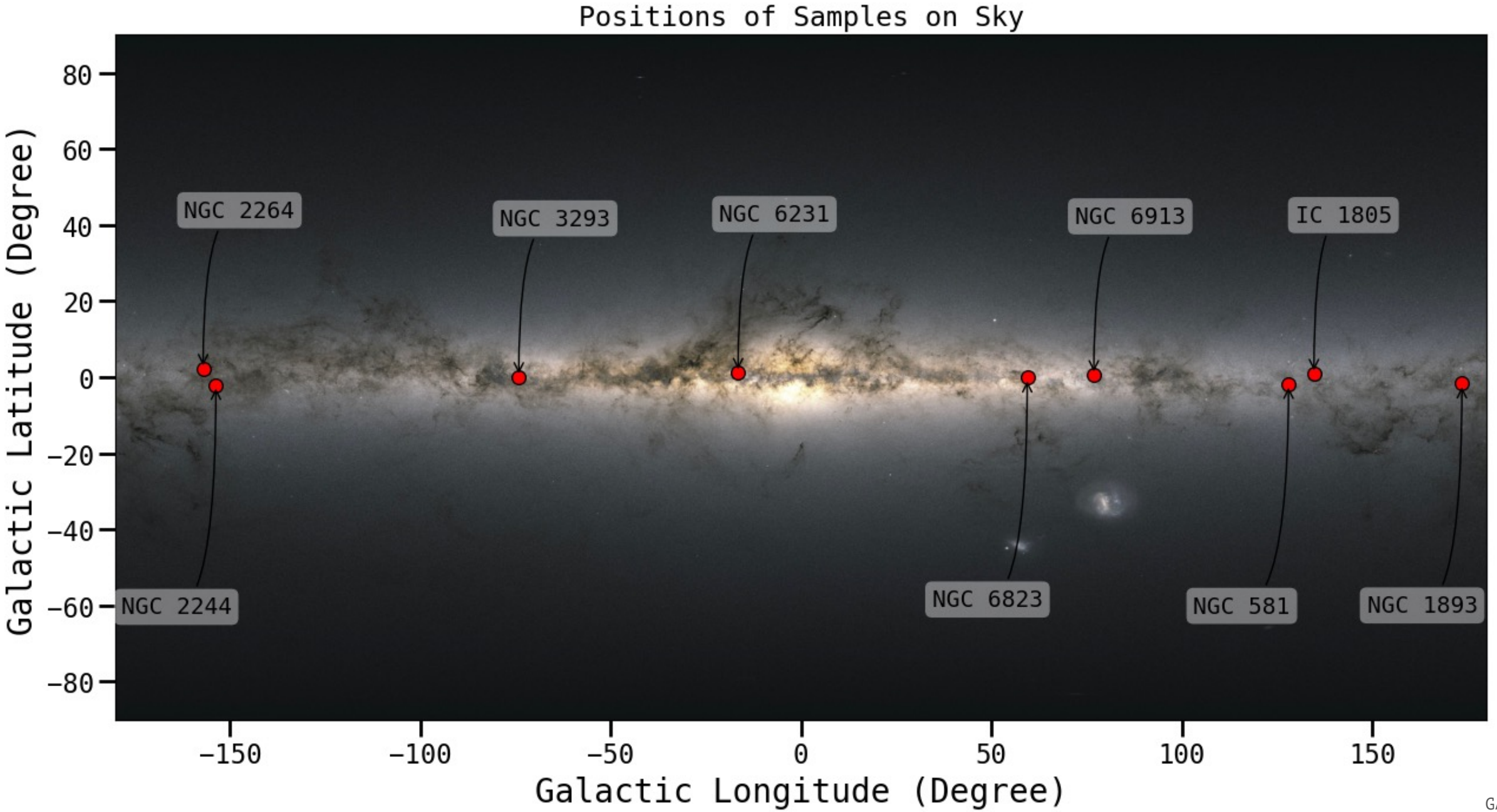


IC 1805

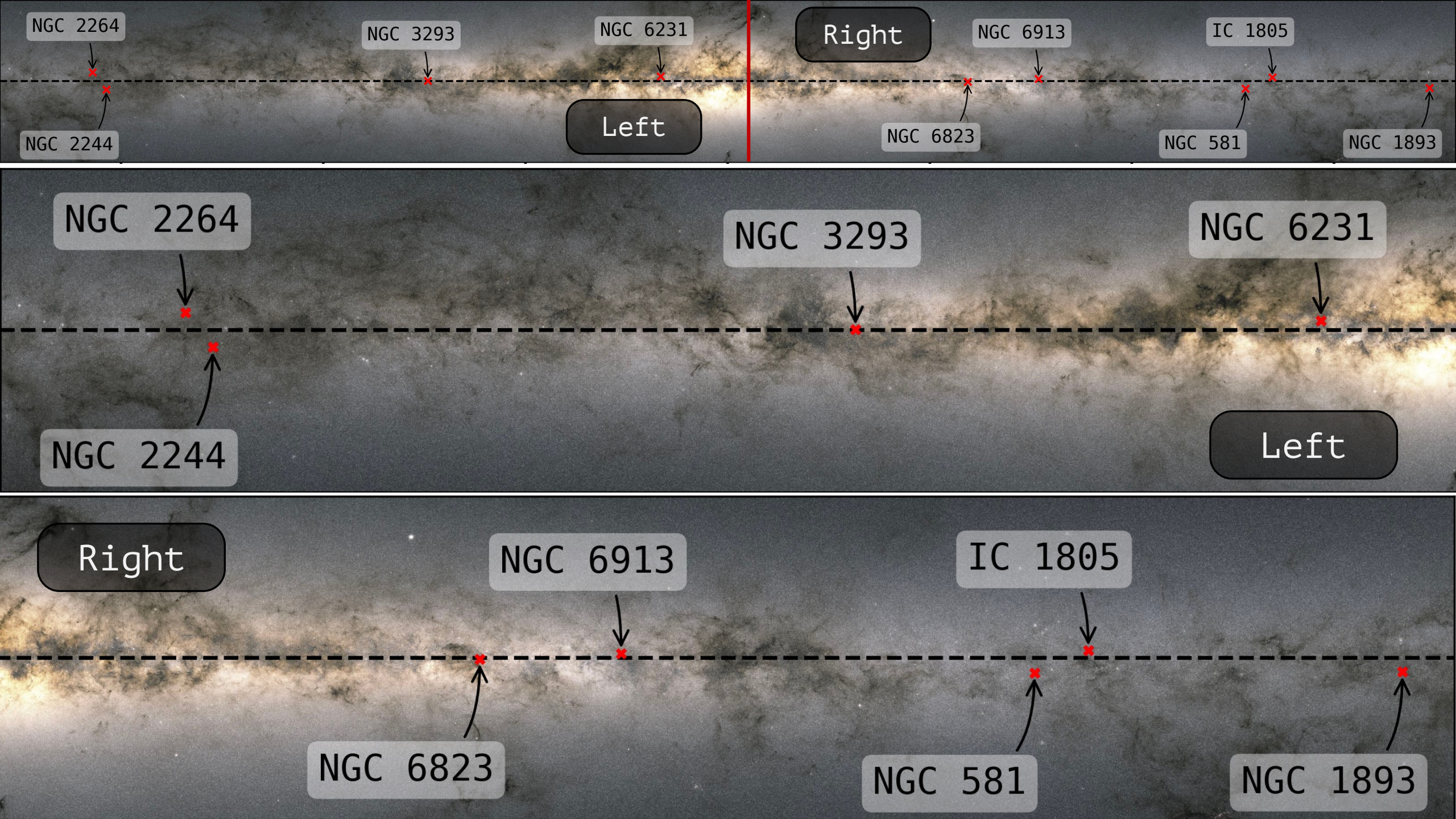
NGC 2264



DSS

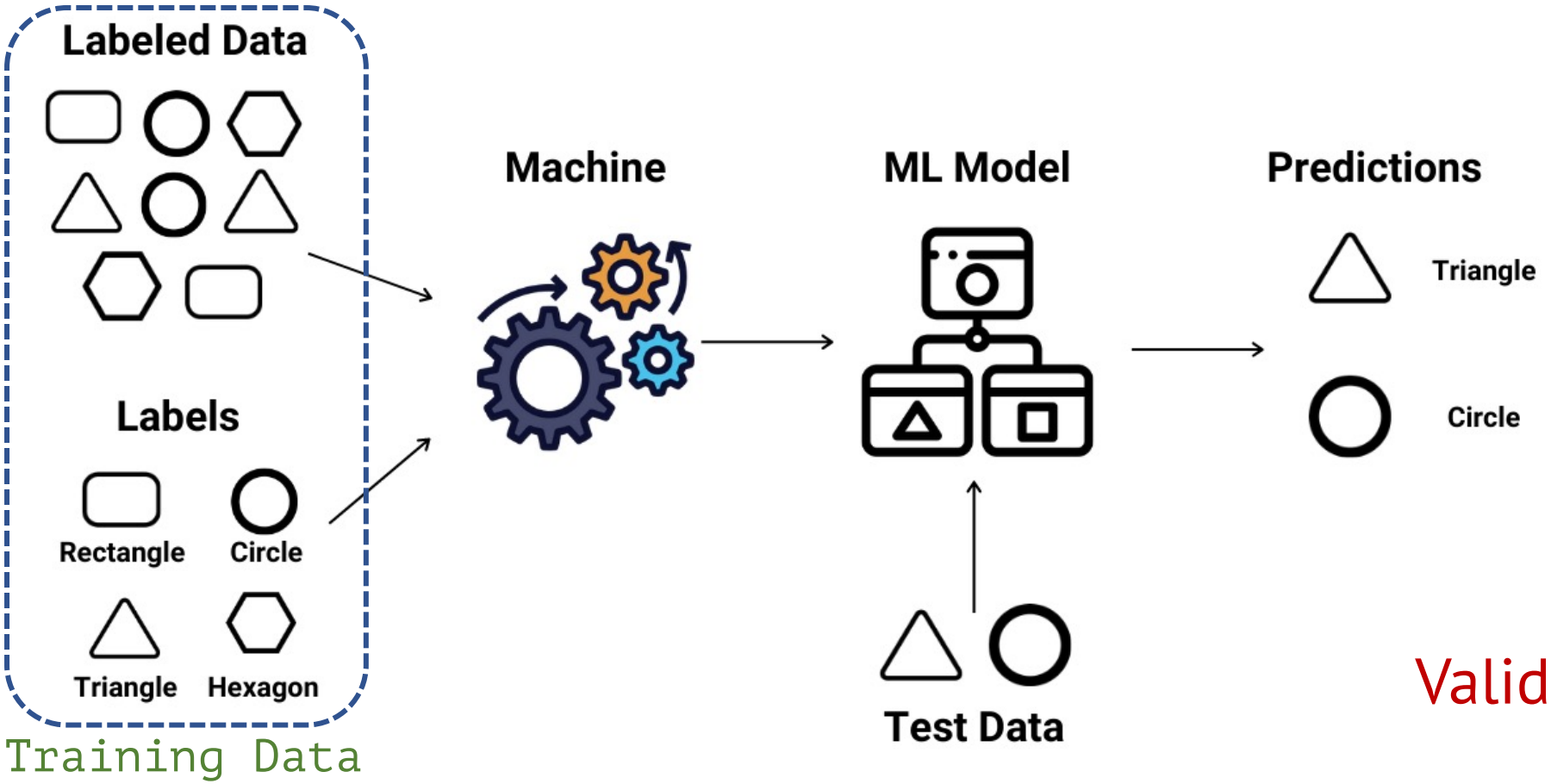

Positions of Samples on Sky
Galactic Latitude (Degree)
Galactic Longitude (Degree)
80
60
40
20
0
−20
−40
−60
−80
−150
−100
−50
0
50
100
150
NGC 2264
NGC 3293
NGC 6231
NGC 6913
IC 1805
NGC 2244
NGC 6823
NGC 581
NGC 1893

GAIA


NGC 2264
NGC 3293
NGC 6231
NGC 2244
Left
Right
NGC 6913
IC 1805
NGC 6823
NGC 581
NGC 1893
NGC 2264
NGC 3293
NGC 6231
NGC 2244
Left
Right
NGC 6913
IC 1805
NGC 6823
NGC 581
NGC 1893

Random Forest in Action

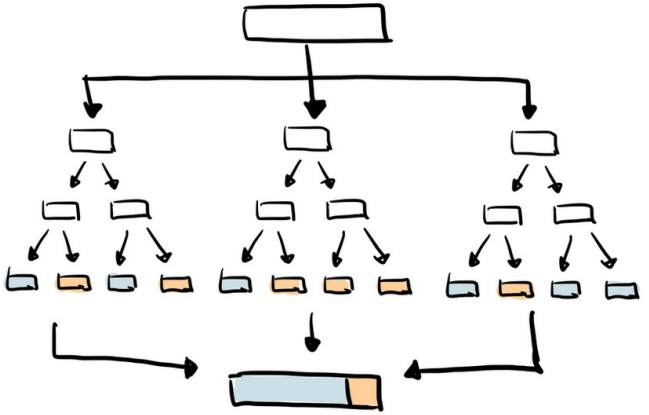
- Training + Validation + Testing



Validation?

Random Forest in Action

- Hyperparameter Tuning



Random Forest

Hyperparameters

- `max_depth`
- `max_features`
  (No. of features in each tree)
- `min_samples_leaf`
  (Min. samples required for a leaf node)
- `n_estimators`
  (No. of Decision Trees)

Parameters

- Position (δ, α)
- Proper Motion (μ_δ, μ_α)
- Parallax (ω)
- Magnitude (m)
- Color & Reddening

Random Forest in Action

- Hyperparameter Tuning

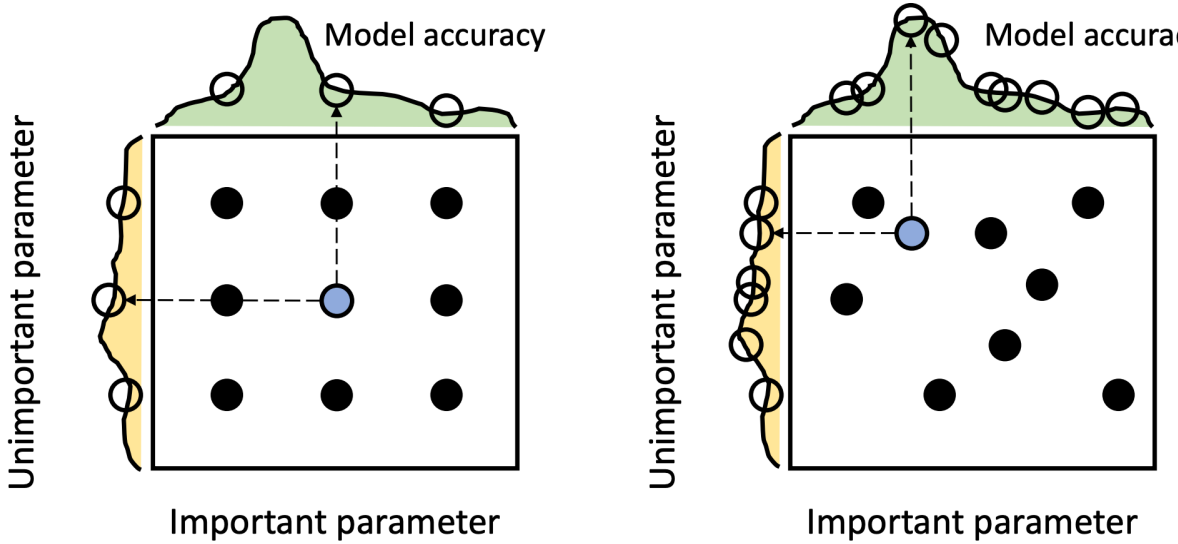
Table 2 The random selection grid with the chosen range of values for important RF model parameters

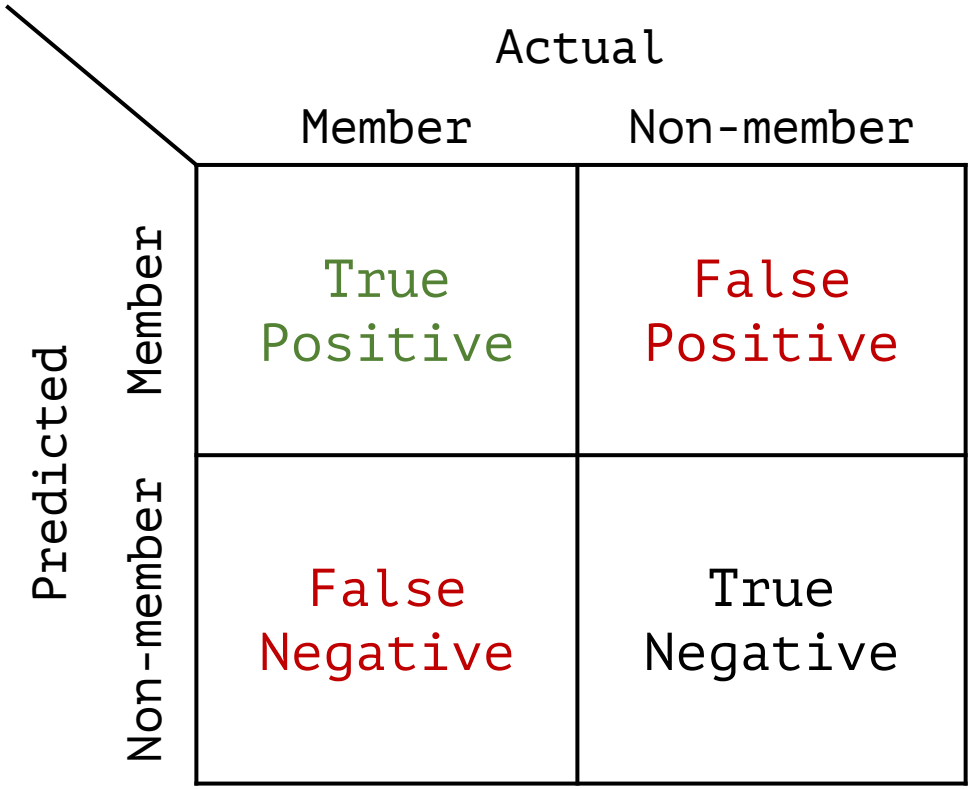
Model parameter	Chosen values to select from
bootstrap (Whether bootstrapping the samples)	True, False
ccp_alpha (Complexity parameter)	2^{-10}, 2^{-9}, 2^{-8}, 2^{-7}, 2^{-6}, 2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 0
max_depth (Maximum depth of the tree)	10, 20, 30, 40, 50, 60,70, 80, 90, 100, 110, None
max_features (Number of features in each tree)	'auto', 'sqrt'
min_samples_leaf (Minimum samples required for a leaf node)	1, 2, 4
min_samples_split (Minimum samples required to split a node)	2, 5, 10
n_estimators (Number of decision trees)	100, 200, 300, 400, 500, 600, 700, 800, 900, 1000

Mahmudunnobe et. al. 2021

Random Forest in Action

- Hyperparameter Tuning





Confusion Matrix

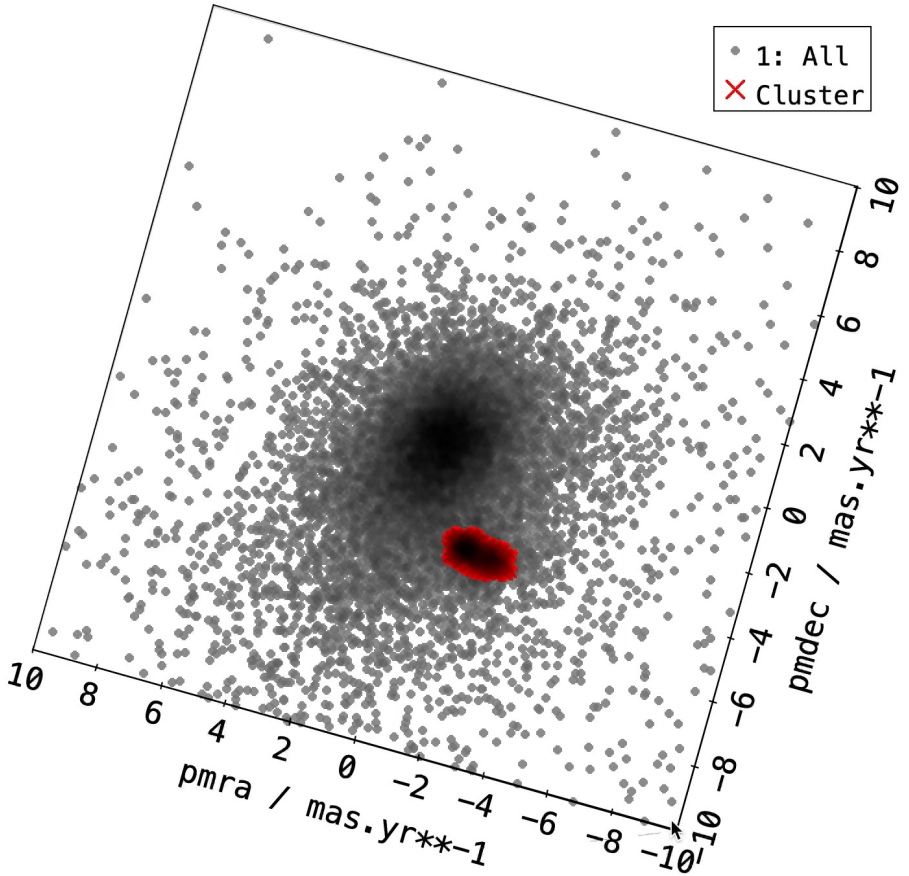
Evaluation Metric for Model Accuracy

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Random Forest in Action

Training Data

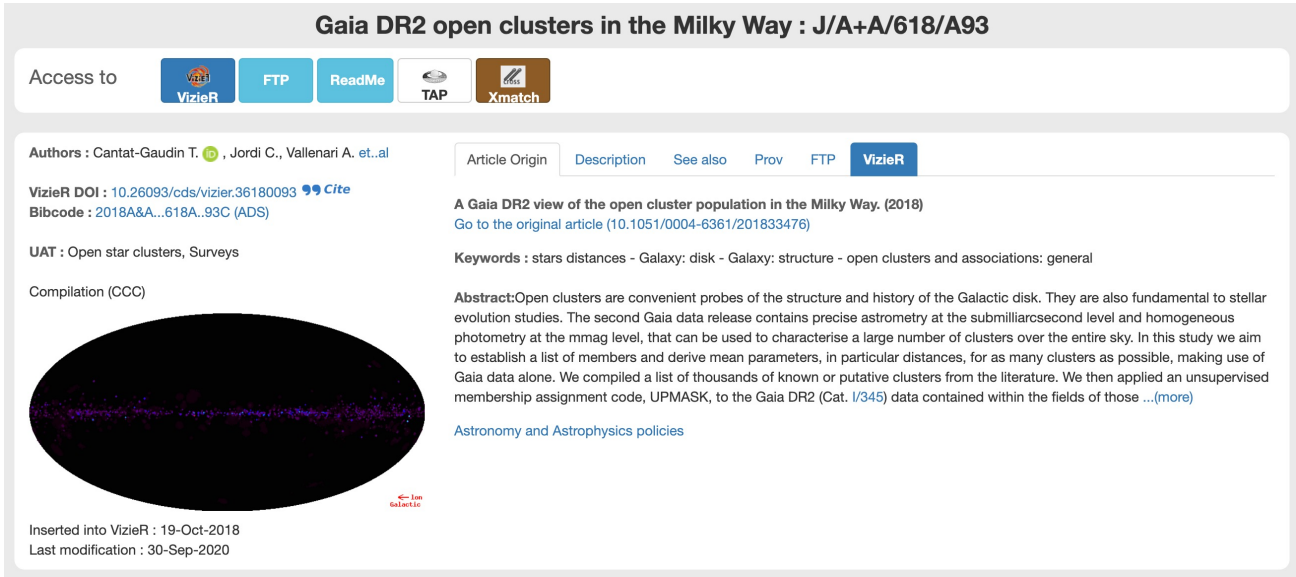
- Classified Open Cluster GAIA DR2 catalogue by Cantat-Gaudin T. et. al. 2018



Random Forest in Action

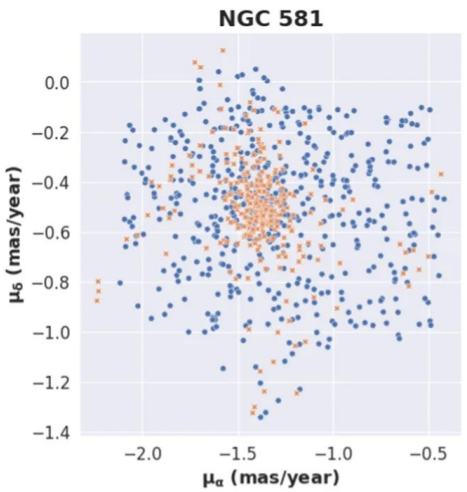
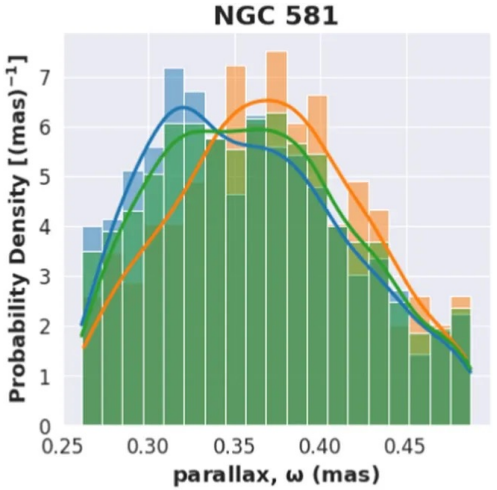
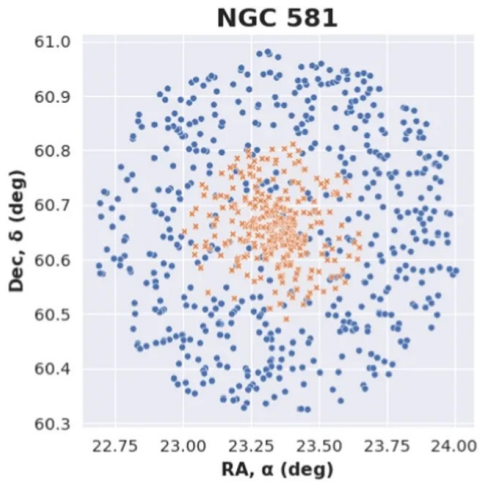
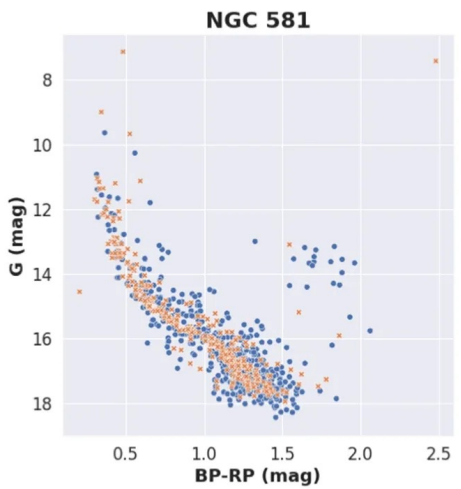
Training Data

- Classified Open Cluster GAIA DR2 catalogue by Cantat-Gaudin T. et. al. 2018
- Only using Astrometric parameters (5D)
- Training on 3D parameter space
- Stars with Membership probability $= 1$
- Members $\in$ `search_radius`
- $search\_radius = 2 \times cluster\_radius$
- Filter criterion:
  - $parallax/parallax\_error > 3$
  - $pmra\_error < 0.3$
  - $pmdec\_error < 0.3$
- Filtered members = `CG members`



Results

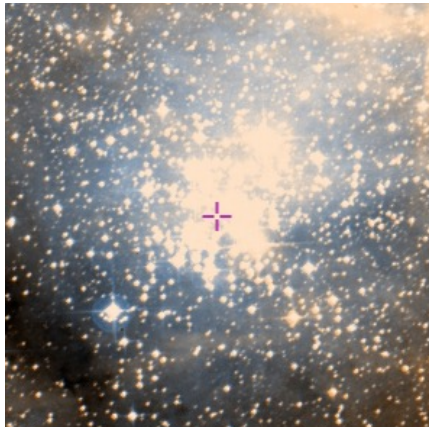
1. Distribution Plots

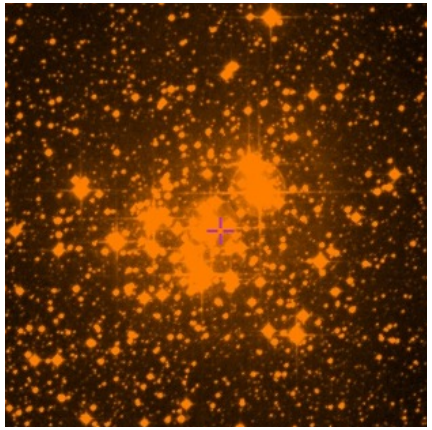
Cluster Image (DSS)	Proper Motion (μ_α, μ_δ) Distribution	Parallax (ω) Distribution	Member Distribution in (α, δ)	CMD
				

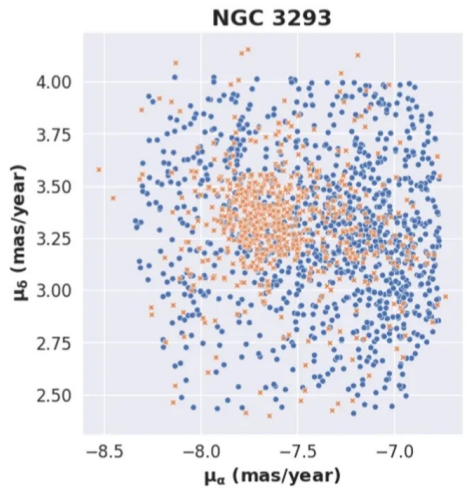
| Legend: CG Members New Members Combined |

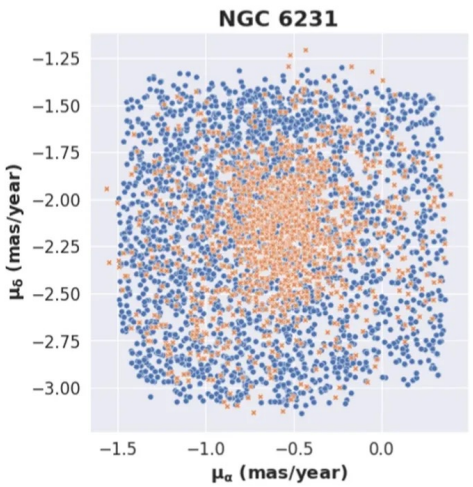
Mahmudunnobe et. al. 2021

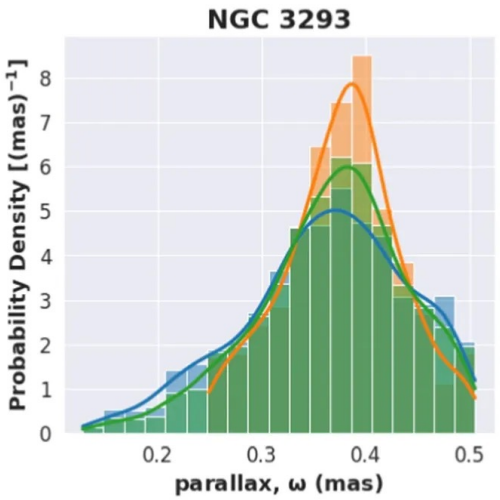
Cluster Image (DSS)	Proper Motion (μ_α, μ_δ) Distribution	Parallax (ω) Distribution	Member Distribution in (α, δ)	CMD

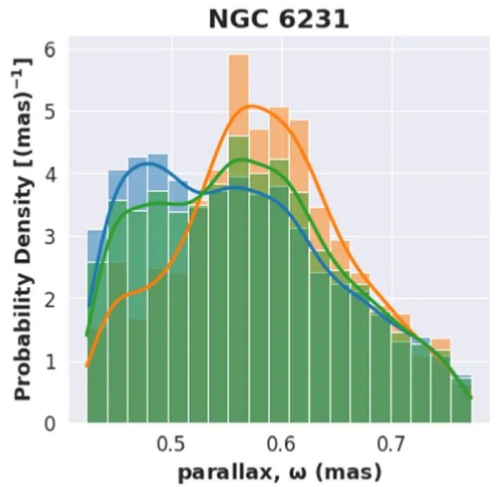


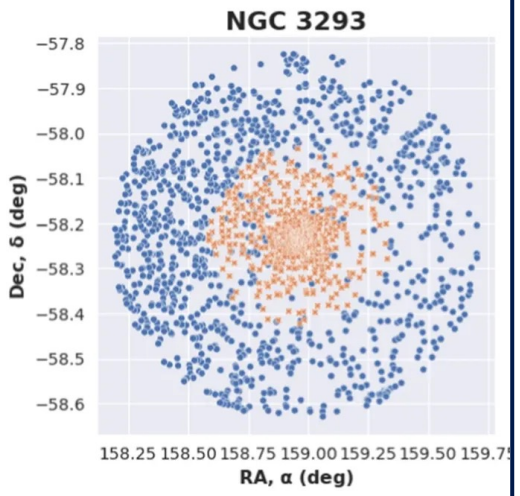


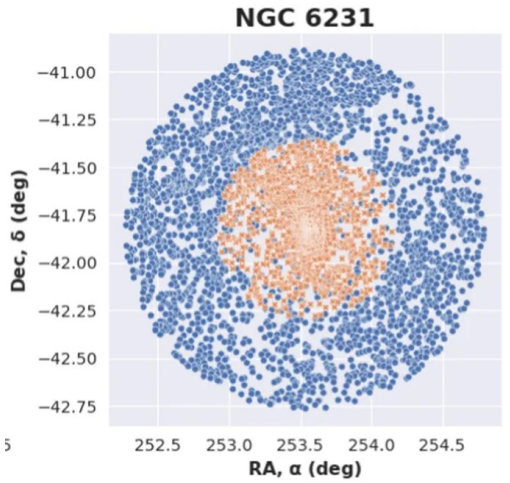


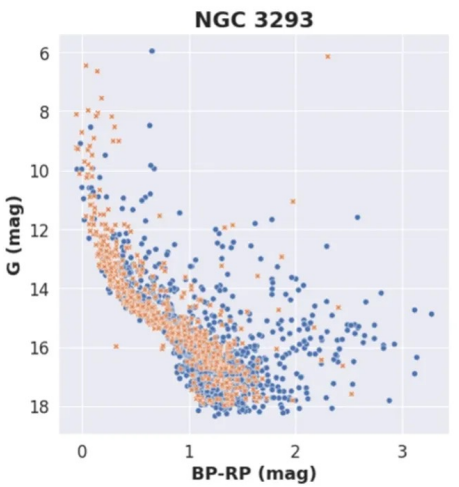


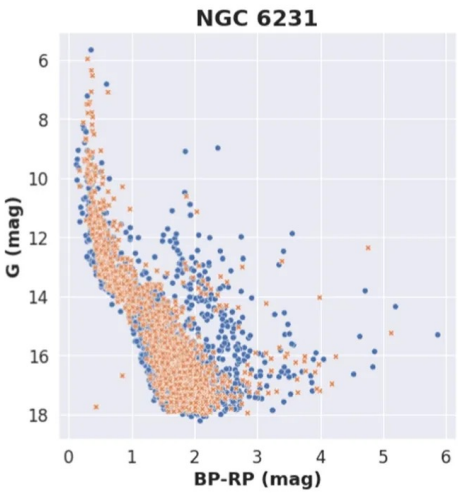








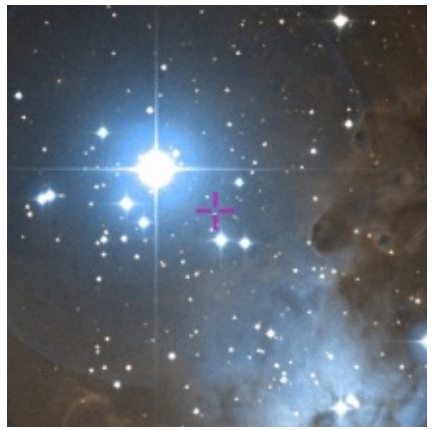


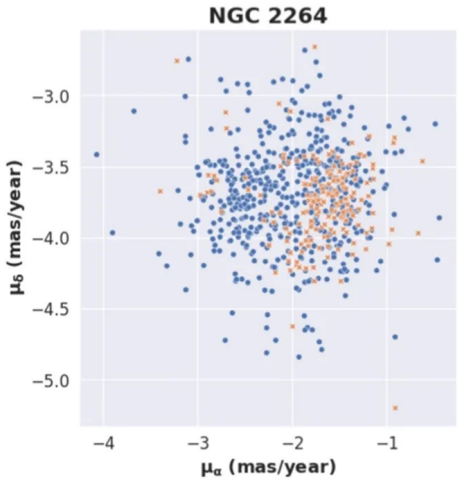


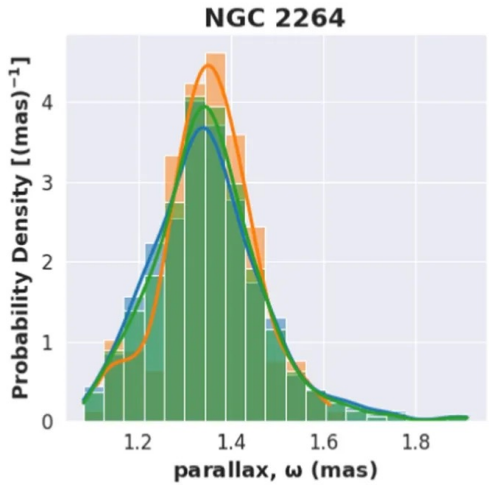
| Legend: CG Members New Members Combined |

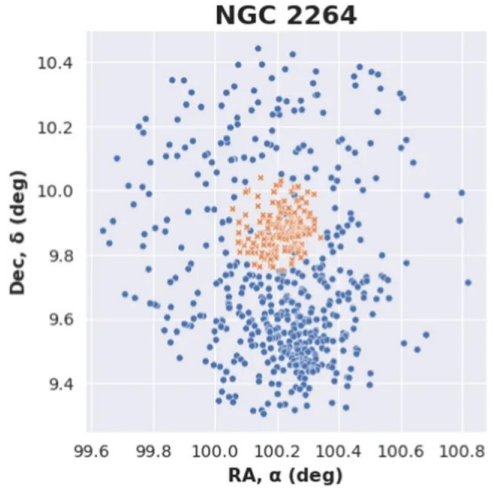
Mahmudunnobe et. al. 2021

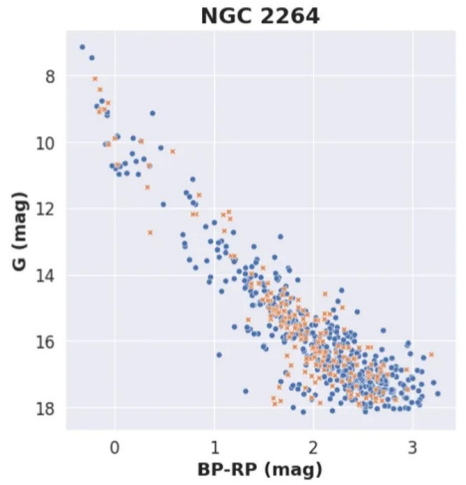
Cluster Image (DSS)	Proper Motion (μ_α, μ_δ) Distribution	Parallax (ω) Distribution	Member Distribution in (α, δ)	CMD

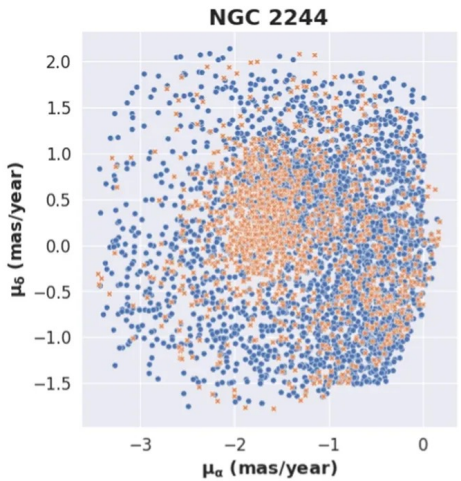


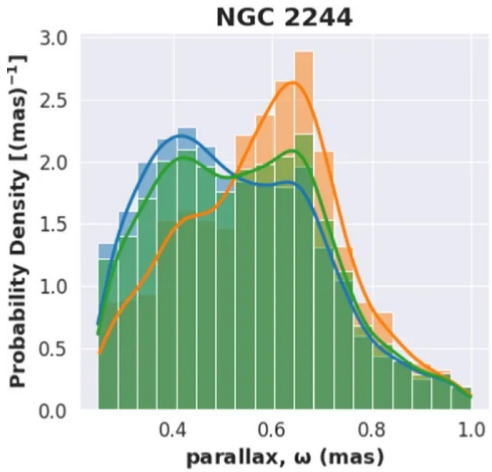


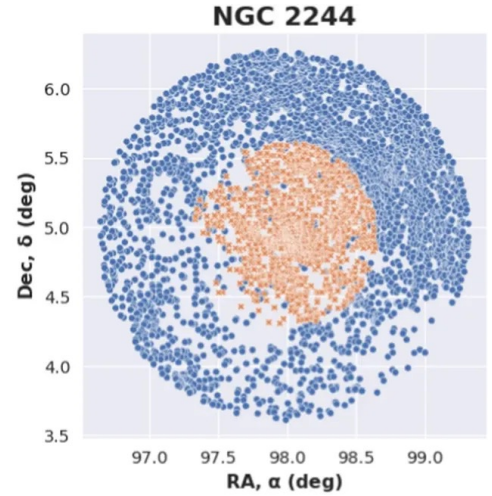


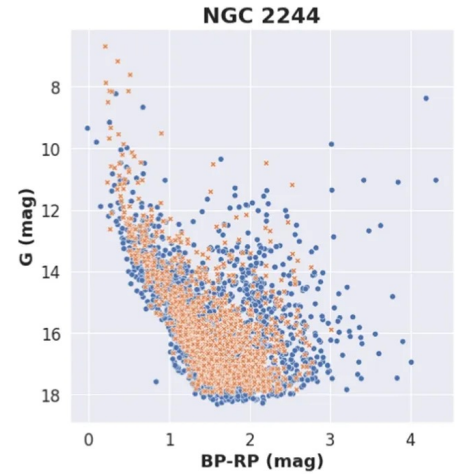










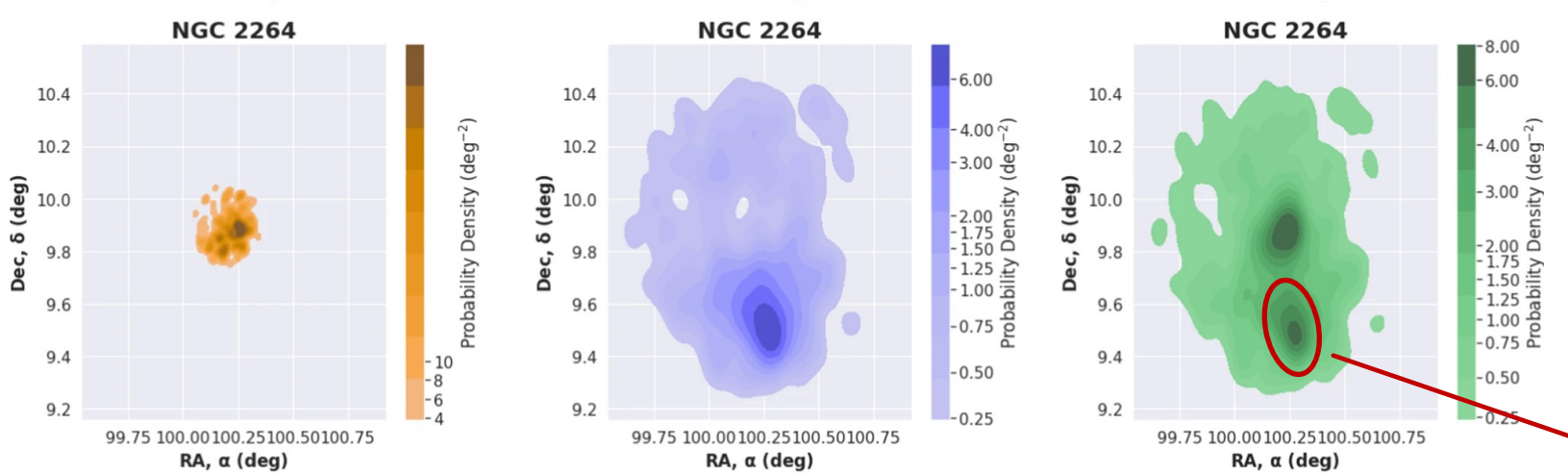


| Legend: CG Members | New Members | Combined |

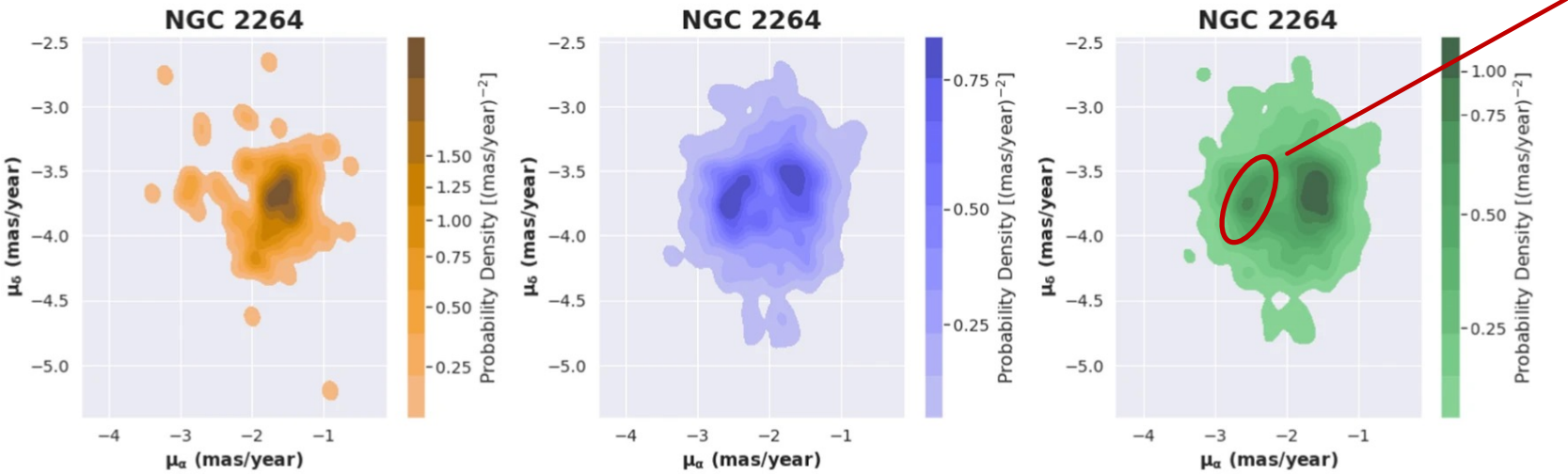
Mahmudunnobe et. al. 2021

2. KDE Plots – Case 1: NGC 2264

Position Space



Proper Motion Space



Bimodalities in both (α, δ) & (μ_α, μ_δ) space

Substructures

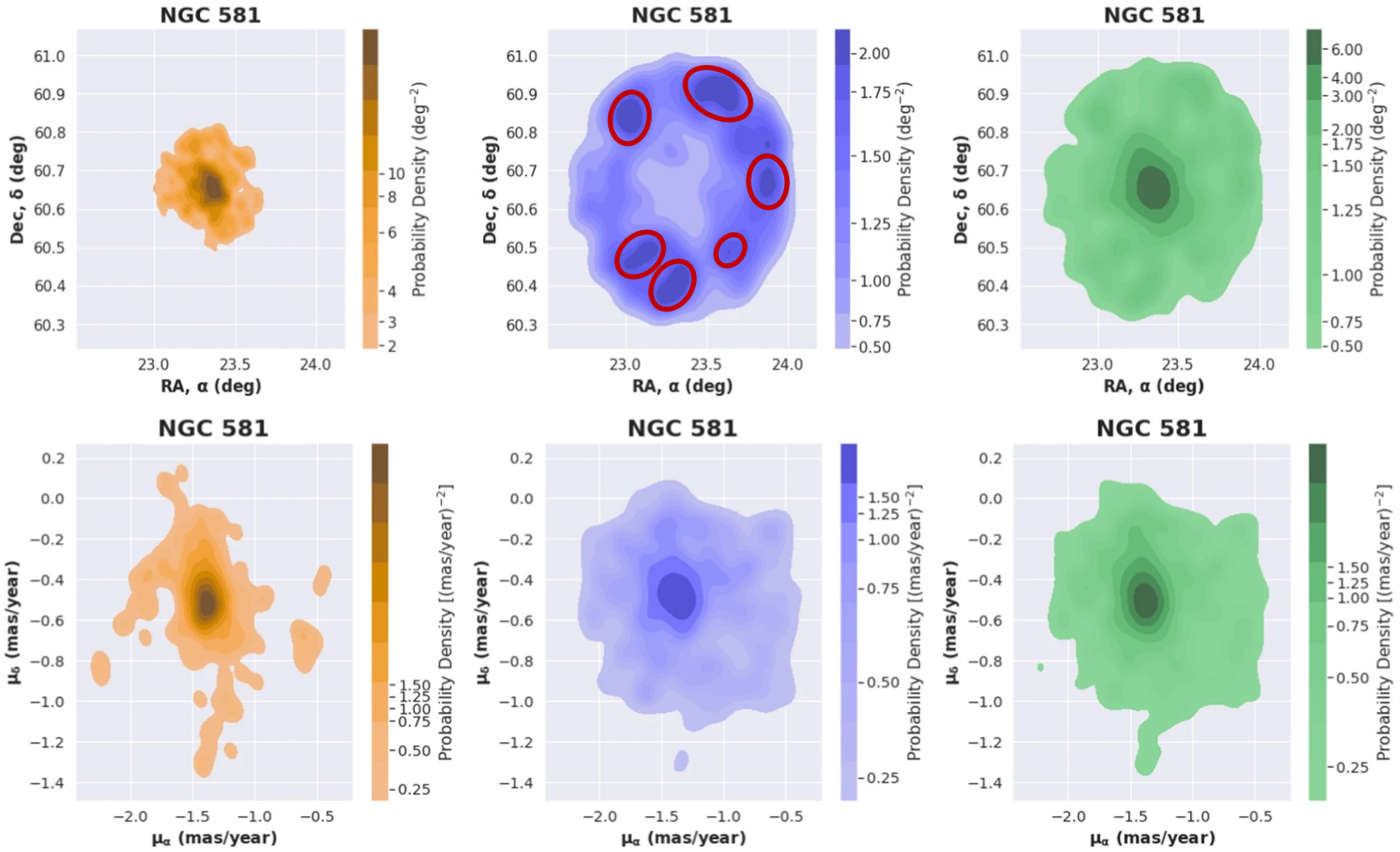
| Legend: CG Members | New Members | Combined |

Mahmudunnobe et. al. 2021

2. KDE Plots – Case 2: NGC 581

Position Space

Proper Motion Space



Bimodalities only in (α, δ) space

↓

Asymmetric Stellar Distribution

| Legend: CG Members | New Members | Combined |

Mahmudunnobe et. al. 2021

3. Prediction Table & Correlation Matrix

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

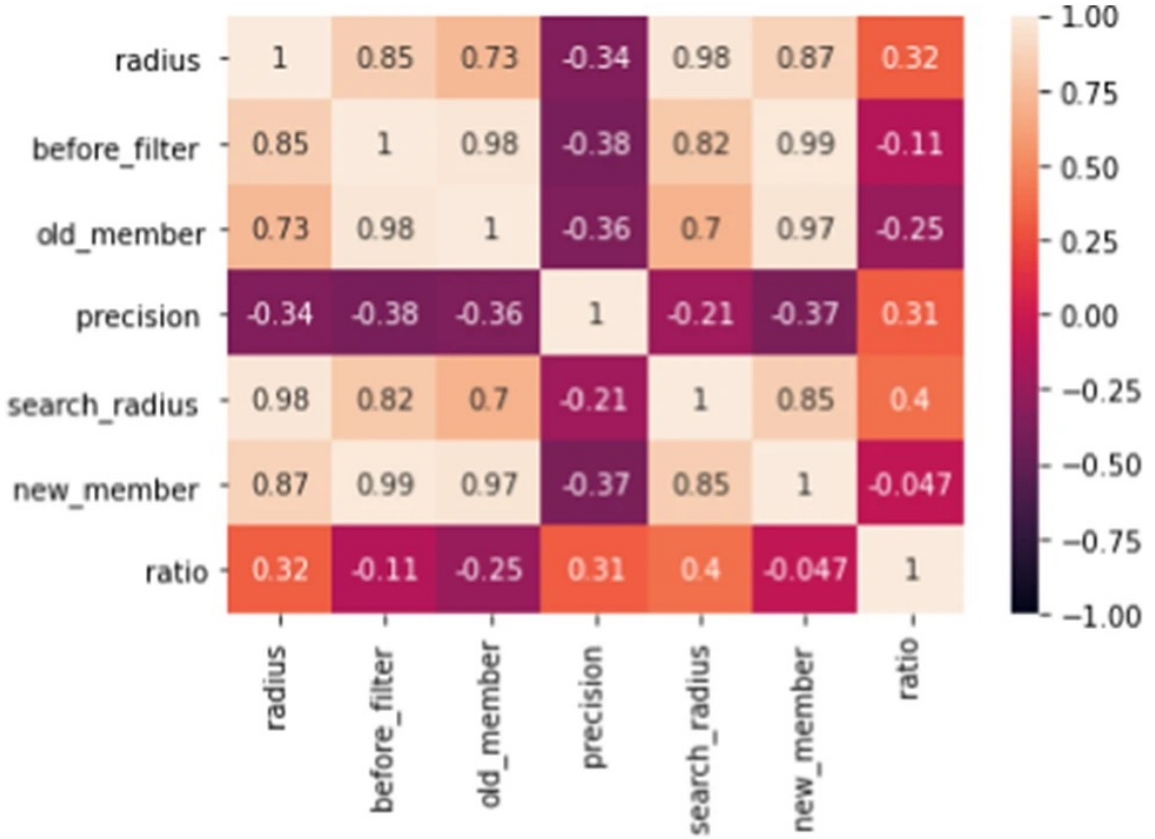
Parameters given to RF

Predictions by RF

3. Prediction Table & Correlation Matrix

Key takeaways:

- Precesion
  - *mildly anti-correlated* to:
    - search_radius: -0.21
    - old_member: -0.36
    - new_member: -0.37
  - *correlated* to ratio of new to CG: 0.31
- Ratio of new to CG
  - *anti-correlated* to old_member: -0.25
  - *correlated* to search_radius: 0.4
- *Tight correlation* b/w old_member & new_member: 0.97



4. Result Table

Table 4 Cluster and sub-clump centers and half radius, $R50$.

Cluster	CG Members Center (α, δ)	$R50$	Combined members Center (α, δ)	$R50$
NGC 1893	(80.72, 33.44)	0.17	(80.75, 33.47) (80.86, 33.96)	0.61
NGC 2244	(98.02, 4.90)	0.37	(98.03, 4.90) (98.74, 5.40) (98.37, 5.55)	0.90
NGC 2264	(100.26, 9.88)	0.08	(100.27, 9.49) (100.23, 9.87)	0.33
NGC 6913	(305.95, 38.50)	0.17	(305.94, 38.50) (305.39, 39.60)	0.45
NGC 581	(23.36, 60.66)	0.11	(23.35, 60.65)	0.27
NGC 3293	(158.96, − 58.24)	0.11	(158.96, − 58.24)	0.34
NGC 6231	(253.56, − 41.83)	0.24	(253.56, − 41.82)	0.66
NGC 6823	(295.79, 23.29)	0.10	(295.79, 23.30)	0.24
IC 1805	(38.22, 61.47)	0.27	(38.22, 61.51)	0.60

Clusters with sub-clumps found

Clusters with no sub-clumps found

Old properties

New properties

To Take Away...

- RF is a highly suitable method
- No. of Stars increased by 2-3 times:
  - improved accuracy for the determination of physical parameters
- Identification of Substructures
- Identification of non-main sequence members:
  - variables
  - pre-main sequence stars
  - unresolved binary sequences

Future Challenges

- Testing other Supervised ML methods
- Testing with GAIA EDR3

Thank you for your patience

Questions?

References

- M. Mahmudunnobe, P. Hasan, M. Raja, and S. N. Hasan, “Membership of stars in open clusters using random forest with gaia data,” *Eur. Phys. J. Spec. Top.*, vol. 230, no. 10, pp. 2177-2191, Sep. 2021, doi: 10.1140/epjs/s11734-021-00205-x.
- X. Gao, “A Machine-learning-based Investigation of the Open Cluster M67,” ApJ, vol. 869, no. 1, p. 9, Dec. 2018, doi: 10.3847/1538-4357/aae8dd.
- Cantat-Gaudin, T., “A Gaia DR2 view of the open cluster population in the Milky Way”, *Astronomy and Astrophysics*, vol. 618, 2018. doi:10.1051/0004-6361/201833476.
- G. Javakhishvili, V. Kukhianidze, M. Todua, and R. Inasaridze, “A method of open cluster membership determination,” A&A, vol. 447, no. 3, pp. 915-919, Mar. 2006, doi: 10.1051/0004-6361:20040297.
- J. Cabrera-Cano and E. J. Alfaro, “A non-parametric approach to the membership problem in open clusters.,” *Astronomy and Astrophysics*, vol. 235, p. 94, Aug. 1990.
- A. Krone-Martins and A. Moitinho, “UPMASK: unsupervised photometric membership assignment in stellar clusters,” *Astronomy & Astrophysics*, Volume 561, id.A57, 12 pp., vol. 561, p. A57, Jan. 2014, doi: 10.1051/0004-6361/201321143.

References

- L. Li, Z. Shao, Z.-Z. Li, J. Yu, J. Zhong, and L. Chen, “Modeling Unresolved Binaries of Open Clusters in the Color-Magnitude Diagram. I. Method and Application of NGC 3532,” ApJ, vol. 901, no. 1, p. 49, Sep. 2020, doi: 10.3847/1538-4357/abaef3.
- W. L. Sanders, “An improved method for computing membership probabilities in open clusters.,” Astronomy and Astrophysics, Vol. 14, p. 226-232 (1971), vol. 14, p. 226, Sep. 1971.
- F. Spada, P. Demarque, Y.-C. Kim, T. S. Boyajian, and J. M. Brewer, “The Yale-Potsdam Stellar Isochrones (YaPSI),” ApJ, vol. 838, no. 2, p. 161, Apr. 2017, doi: 10.3847/1538-4357/aa661d.
- J. L. Zhao and Y. P. He, “An improved method for membership determination of stellar clusters with proper motions with different accuracies.,” Astronomy and Astrophysics, vol. 237, p. 54, Oct. 1990.
- Gaia Collaboration, “The Gaia mission”, *Astronomy and Astrophysics*, vol. 595, 2016. doi:10.1051/0004-6361/201629272.
- “Seeing Statistics.” https://seeing-statistics.com/issue4/ “Single Plot — mw-plot 0.9.0 documentation.” https://milkyway-plot.readthedocs.io/en/latest/matplotlib_single.html

Appendix

Earlier Methods

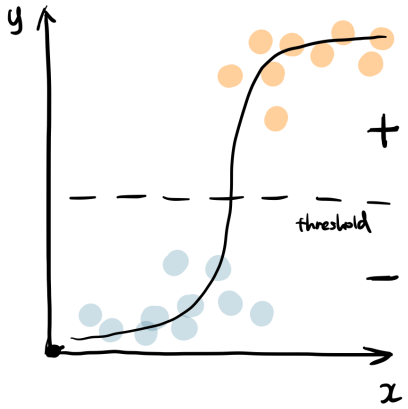
- Main Goal: Estimating Cluster Membership

- Parametric: Vasilevskis-Sanders method
  - Type: probabilistic model solved by Max. Likelihood method
  - Model: based on the distribution of Proper Motion dispersion
  - Failure case:
    - significant internal motion in a cluster or its rotation
    - small cluster member-to-field star ratio

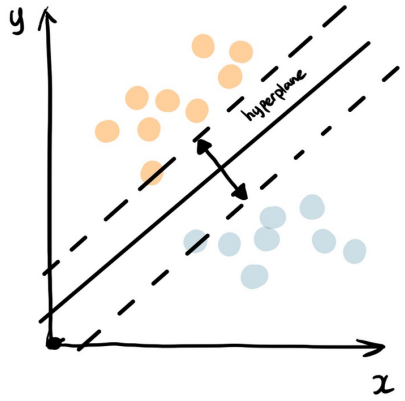
- Non-Parametric model
  - Type: probabilistic model solved by Discriminant Analysis
  - Model: based on distribution-free techniques

- Model-Free methods: Data-driven techniques (ML)

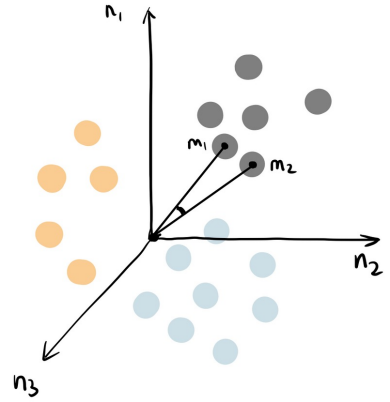
Supervised Machine Learning Techniques



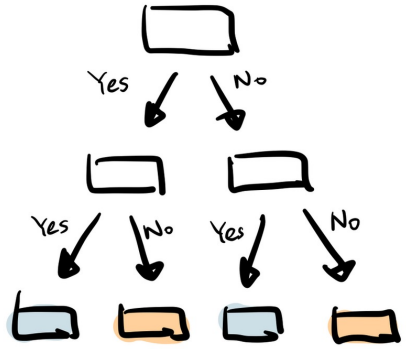
Logistic Regression



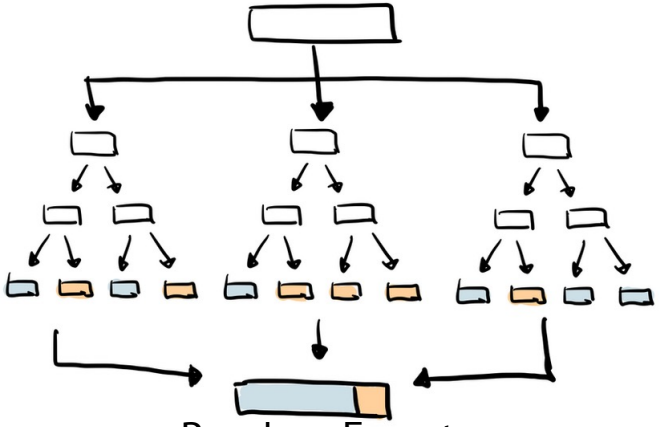
Support Vector Machine



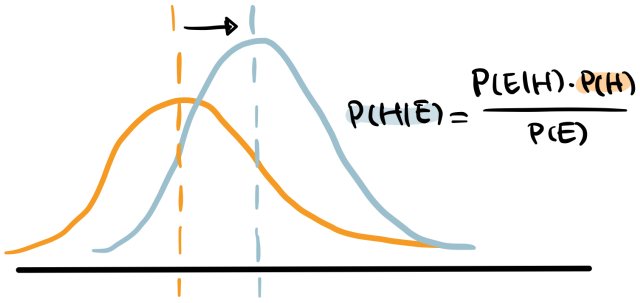
K-Nearest Neighbors



Decision Trees

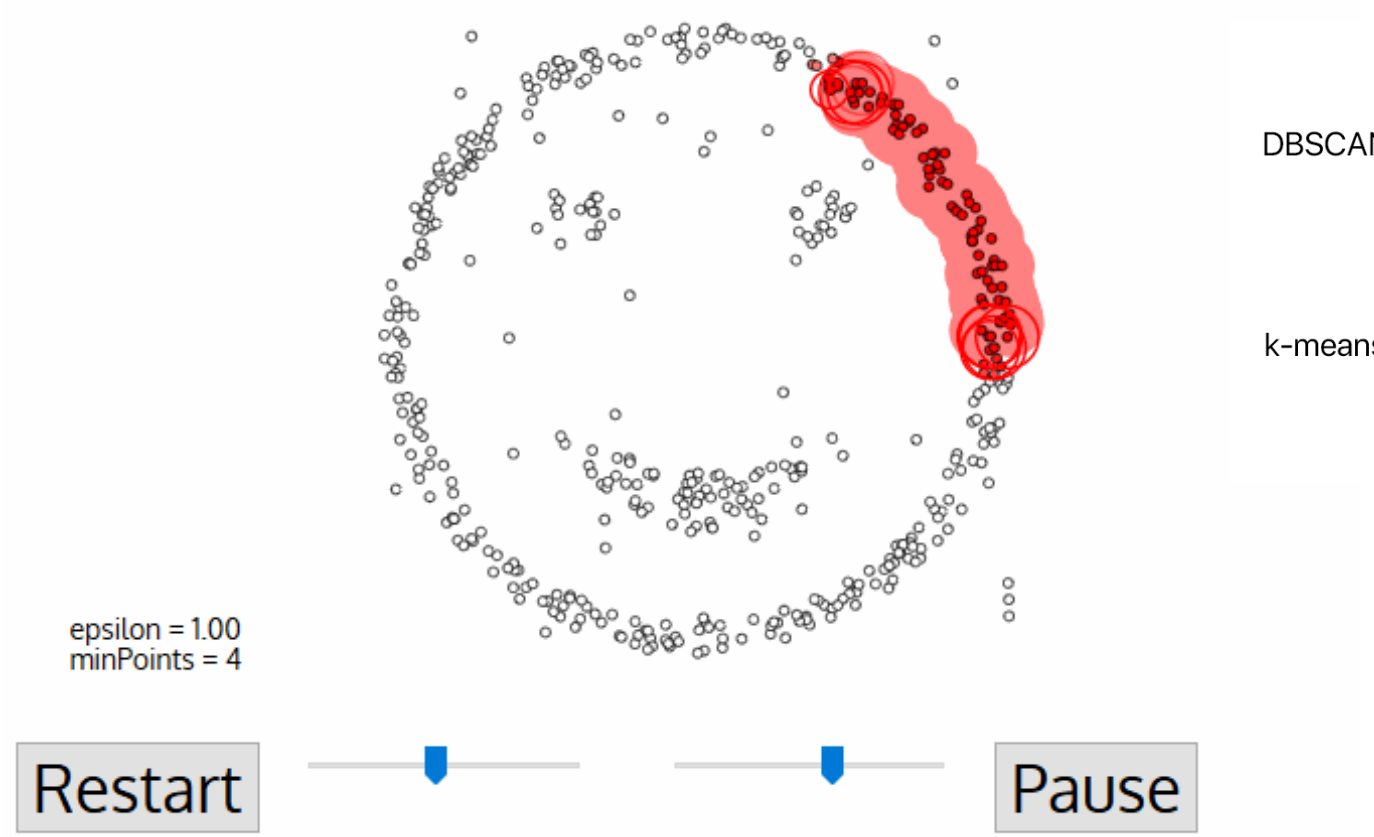


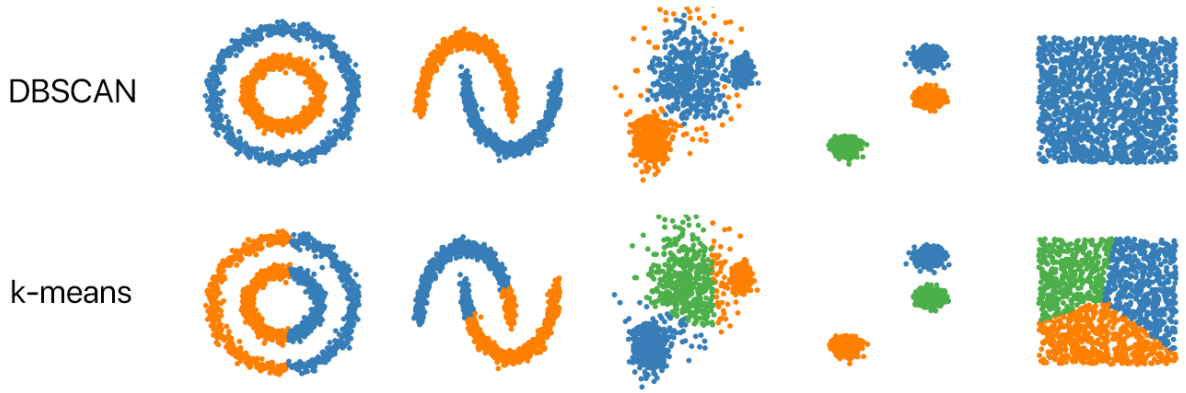
Random Forest



Naïve Bayes

DBSCAN (Density-based Spatial Clustering of Applications with Noise)

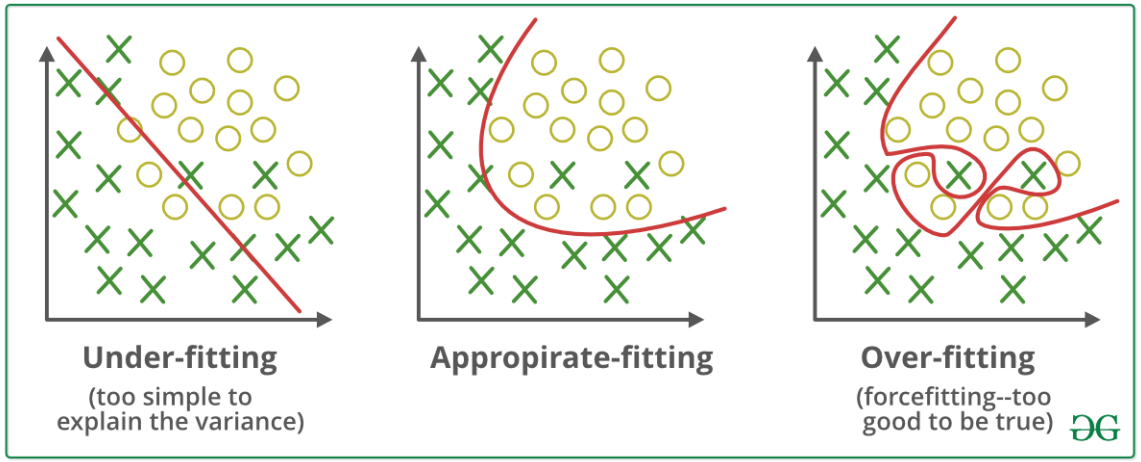


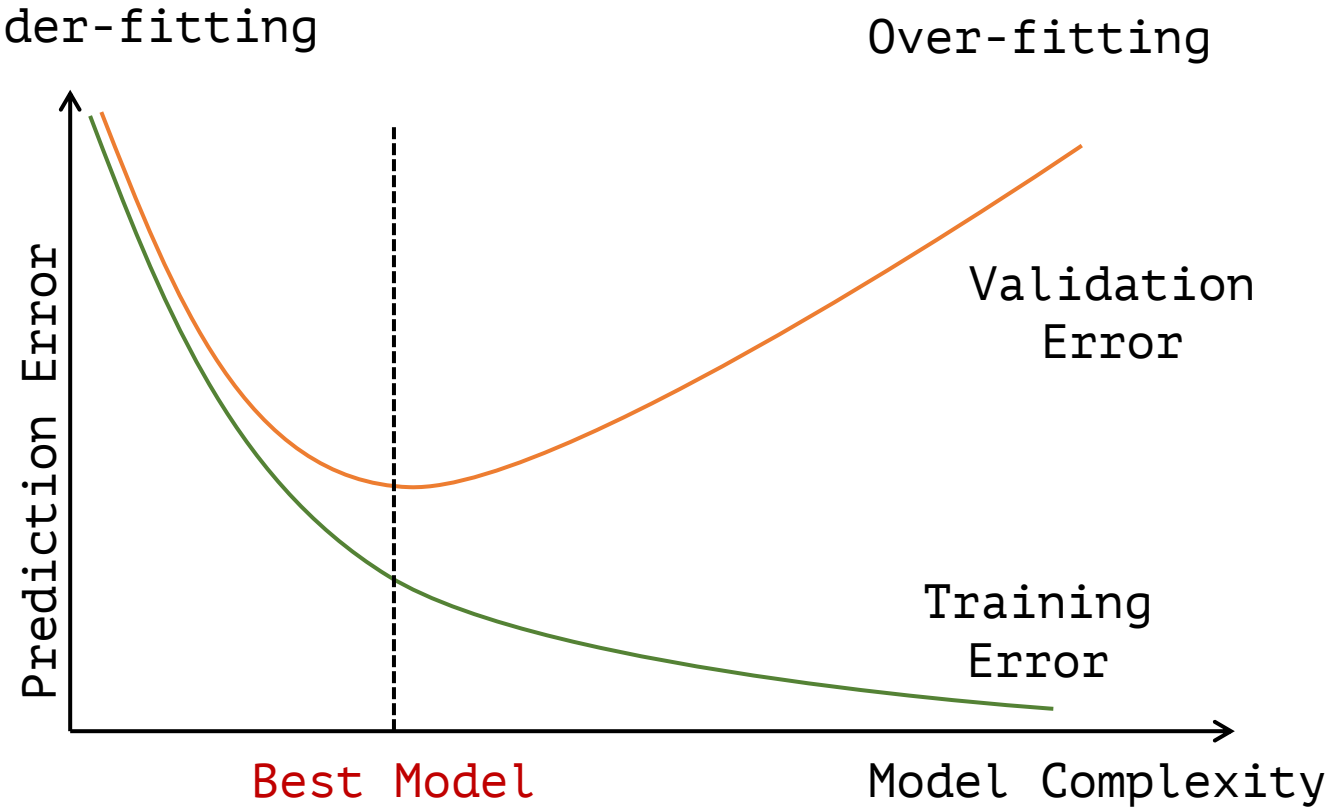


Random Forest in Action

- Training + Validation + Testing

evaluation of the ML model to ensure its accuracy

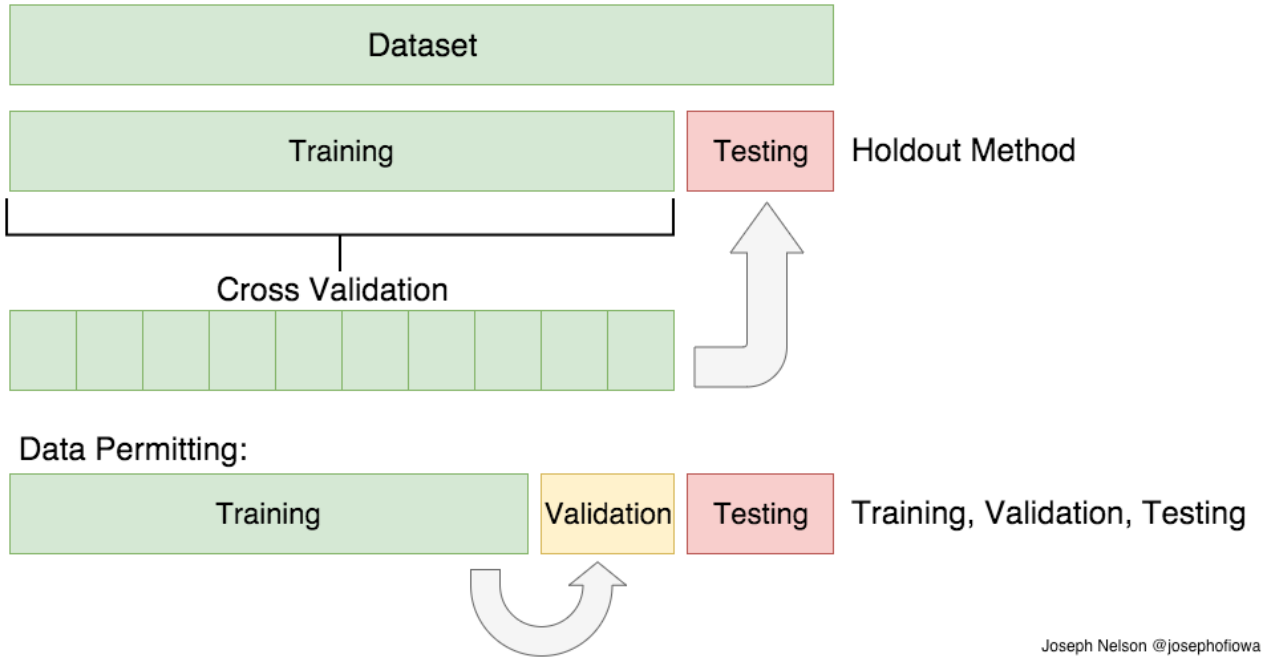


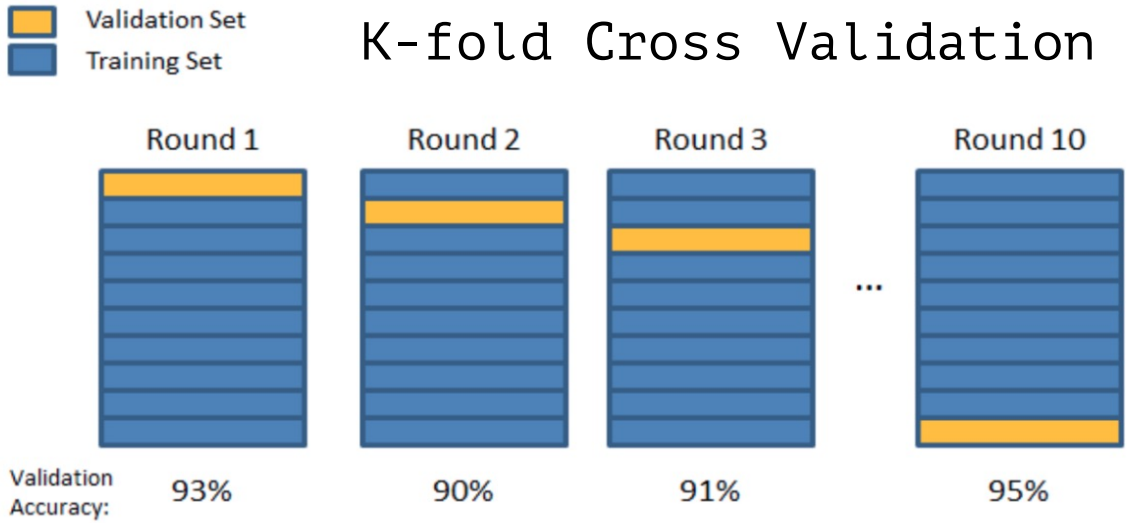


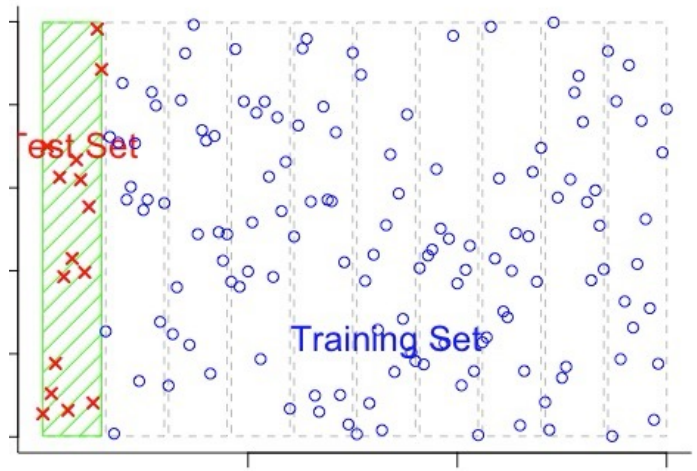
Random Forest in Action

- Training + Validation + Testing

How to apply it to your dataset?

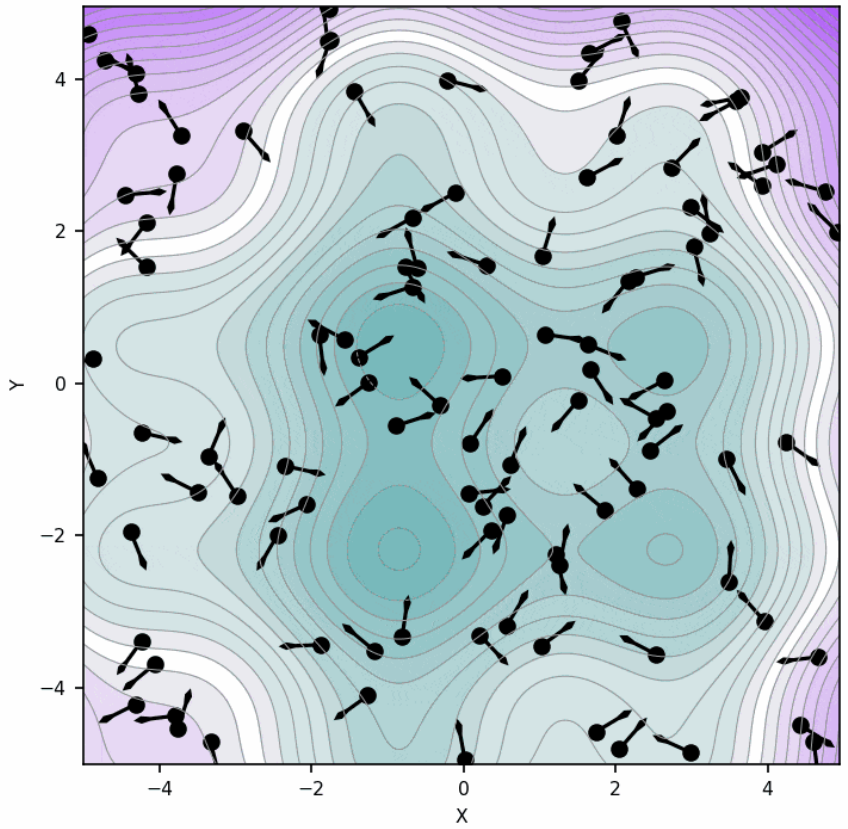


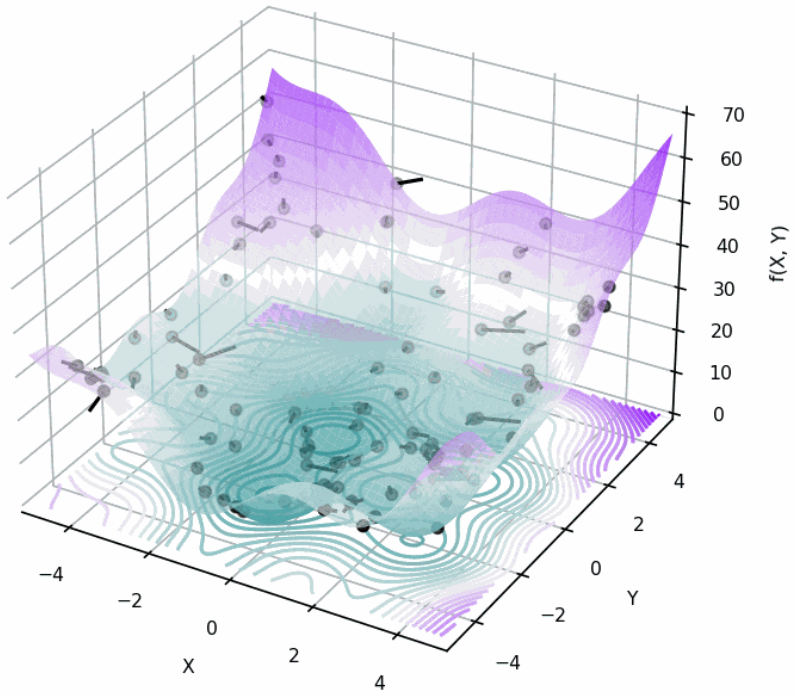




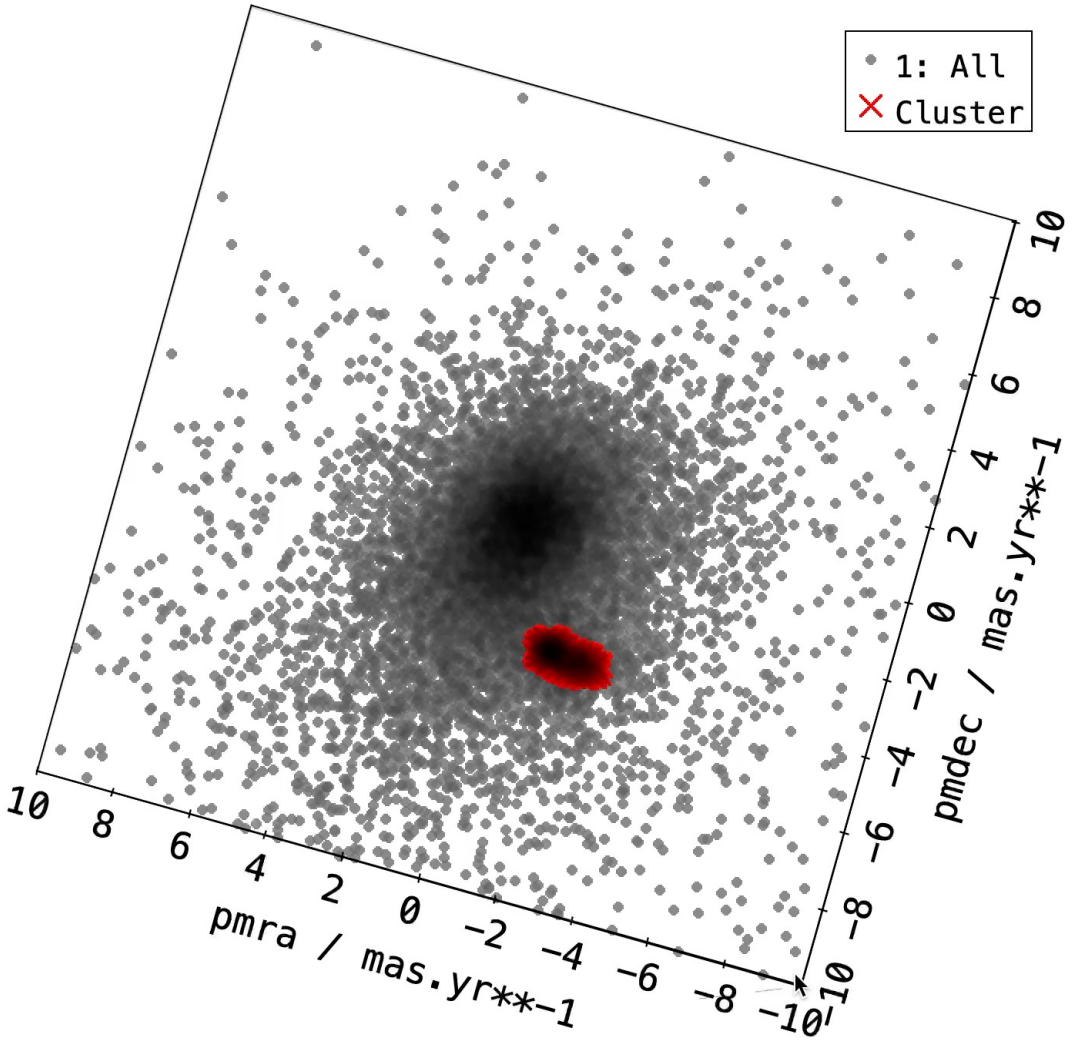
Hyperparameter Tuning

[1/100] w:0.800 - c_1:2.000 - c_2:2.000

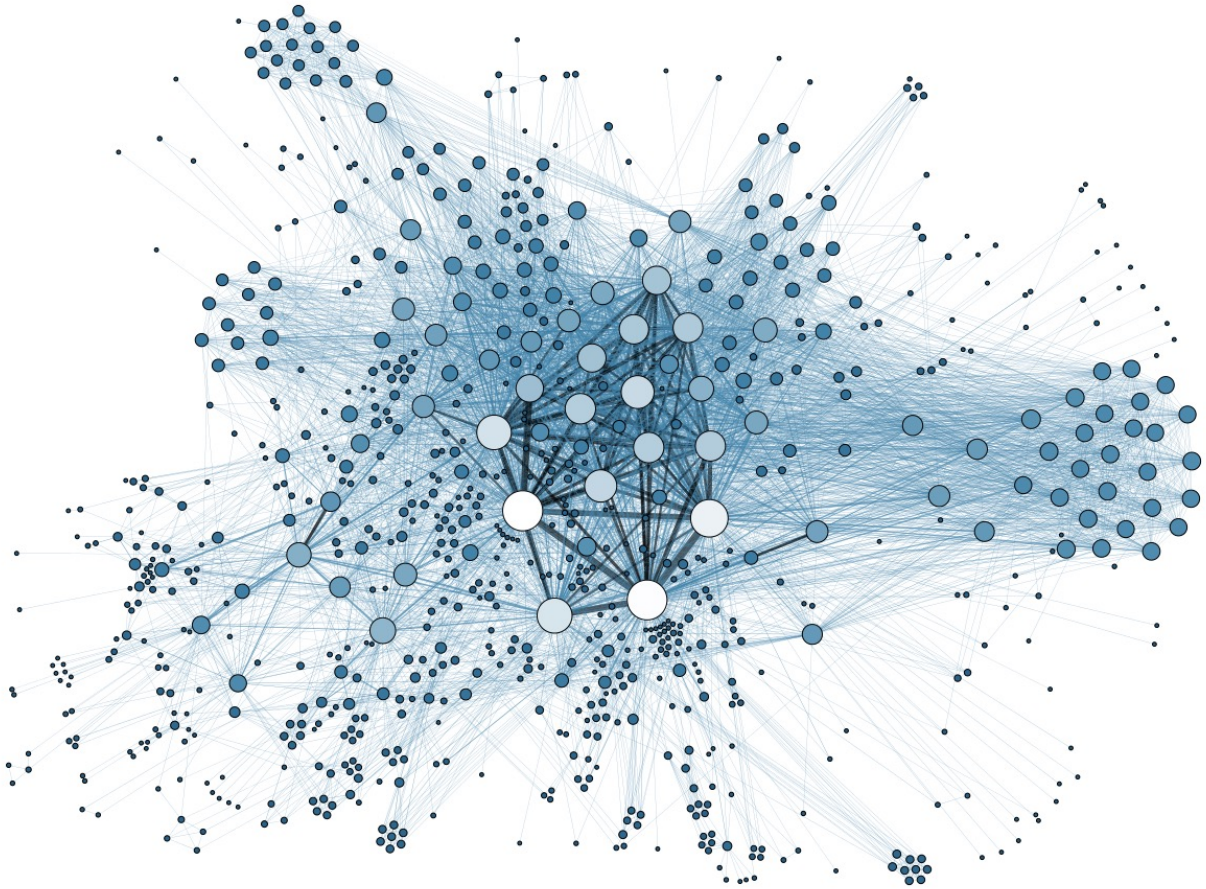




3D parameter space for Model Training

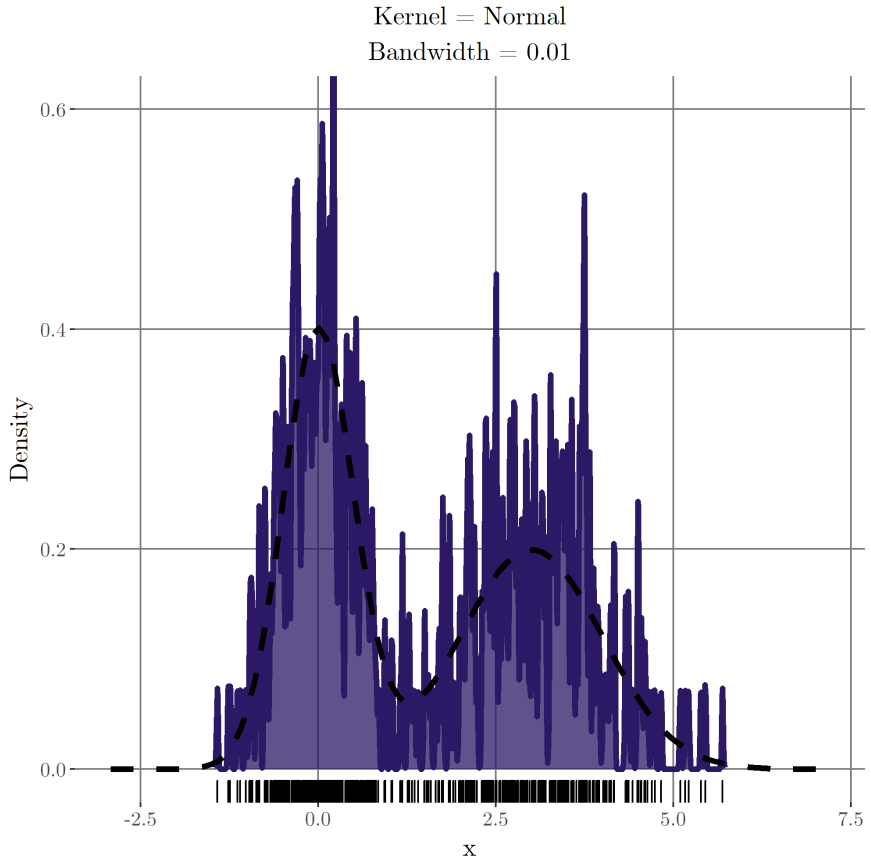


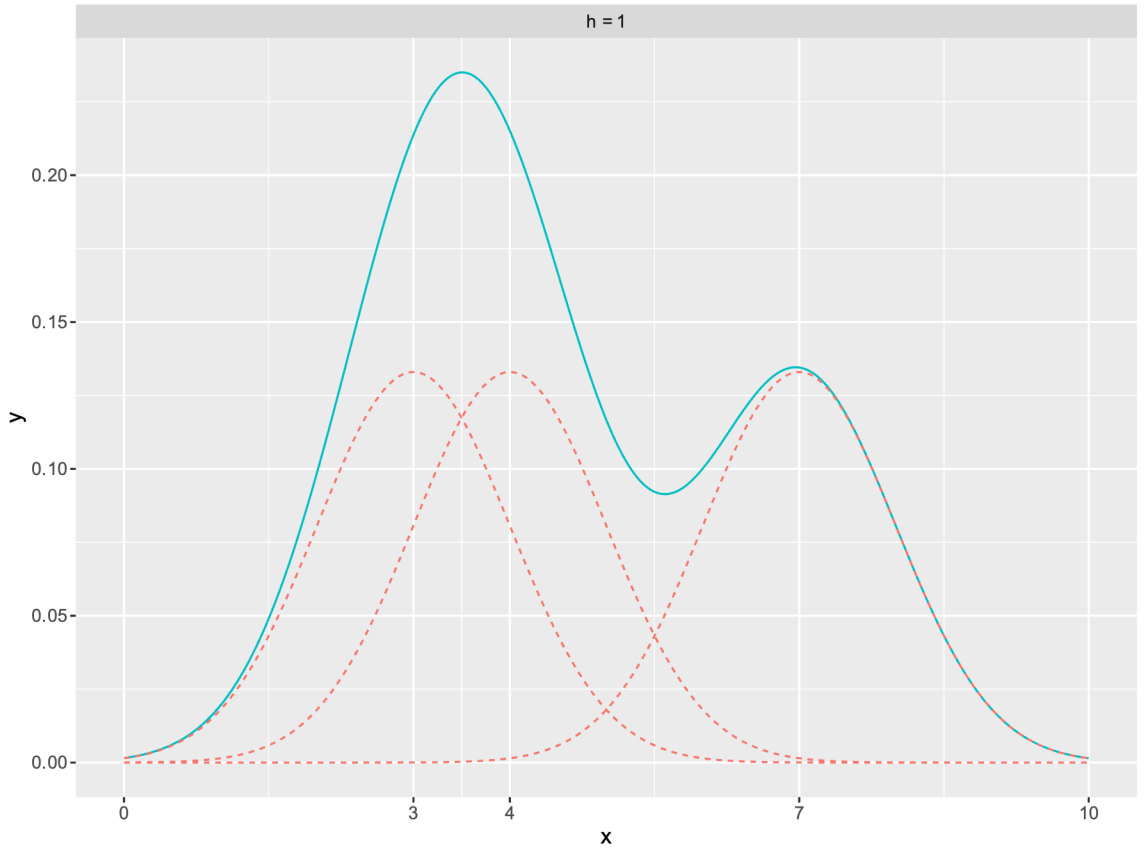
High-Dimensional parameter space for Model Training

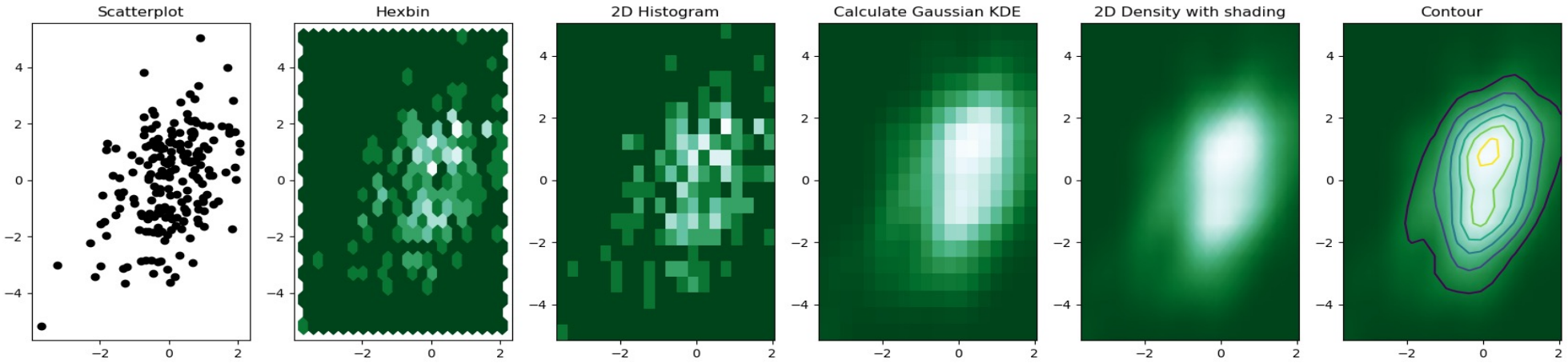


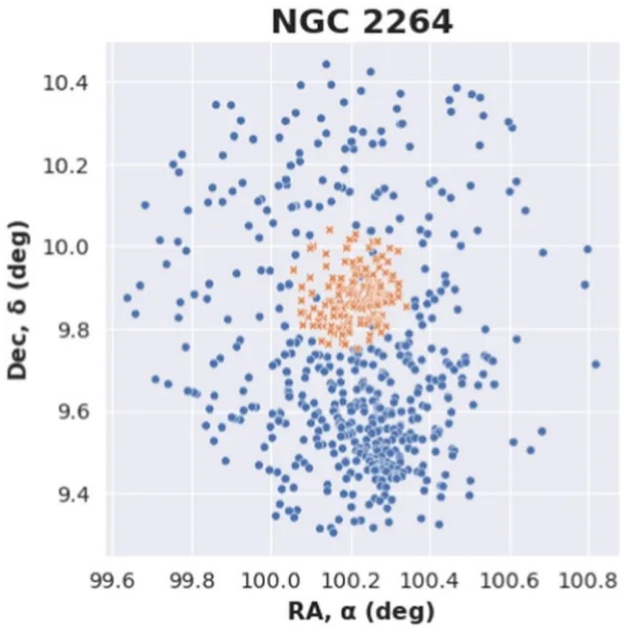
Grandjean, Martin (2014)

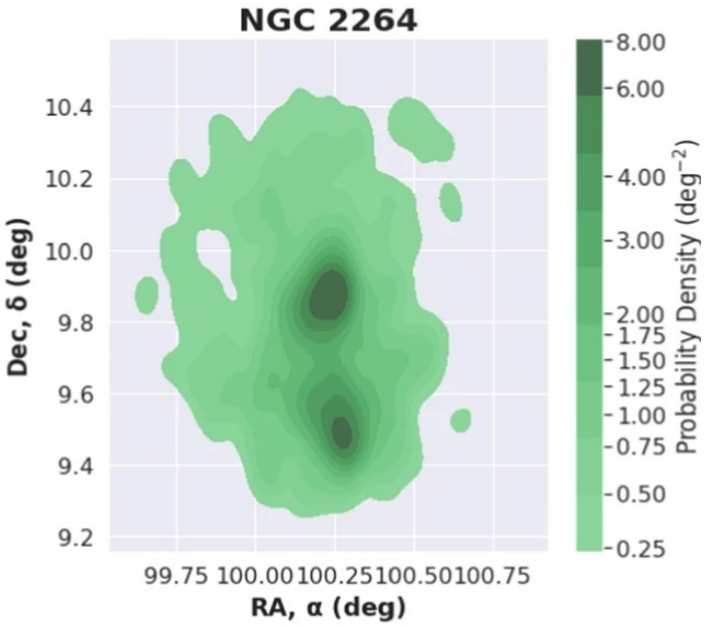
KDE Plots - Introduction











Mahmudunnobe et. al. 2021

Member Probability in Random Forest

Membership Probability estimated by Random Forest:

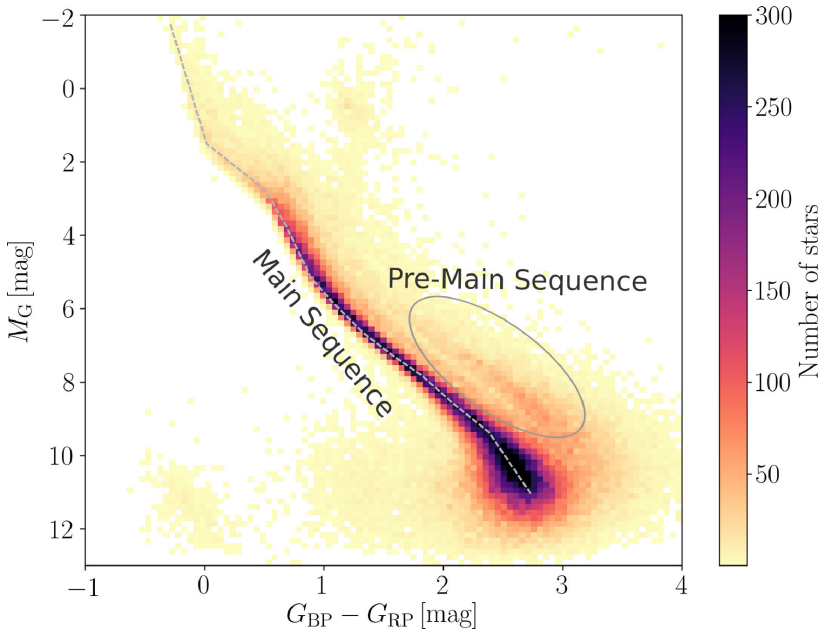
$$P_{RF} = \frac{N_c}{N}$$

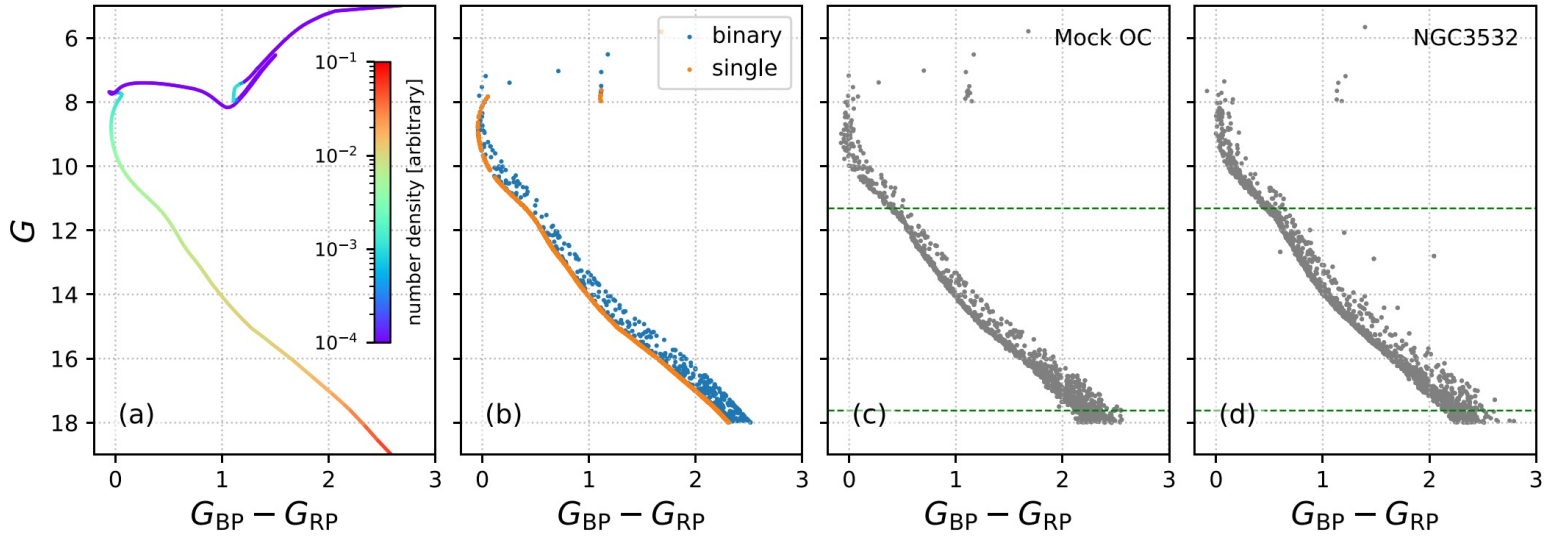
where,

N_c = no. of decision trees which classified a star as a cluster member

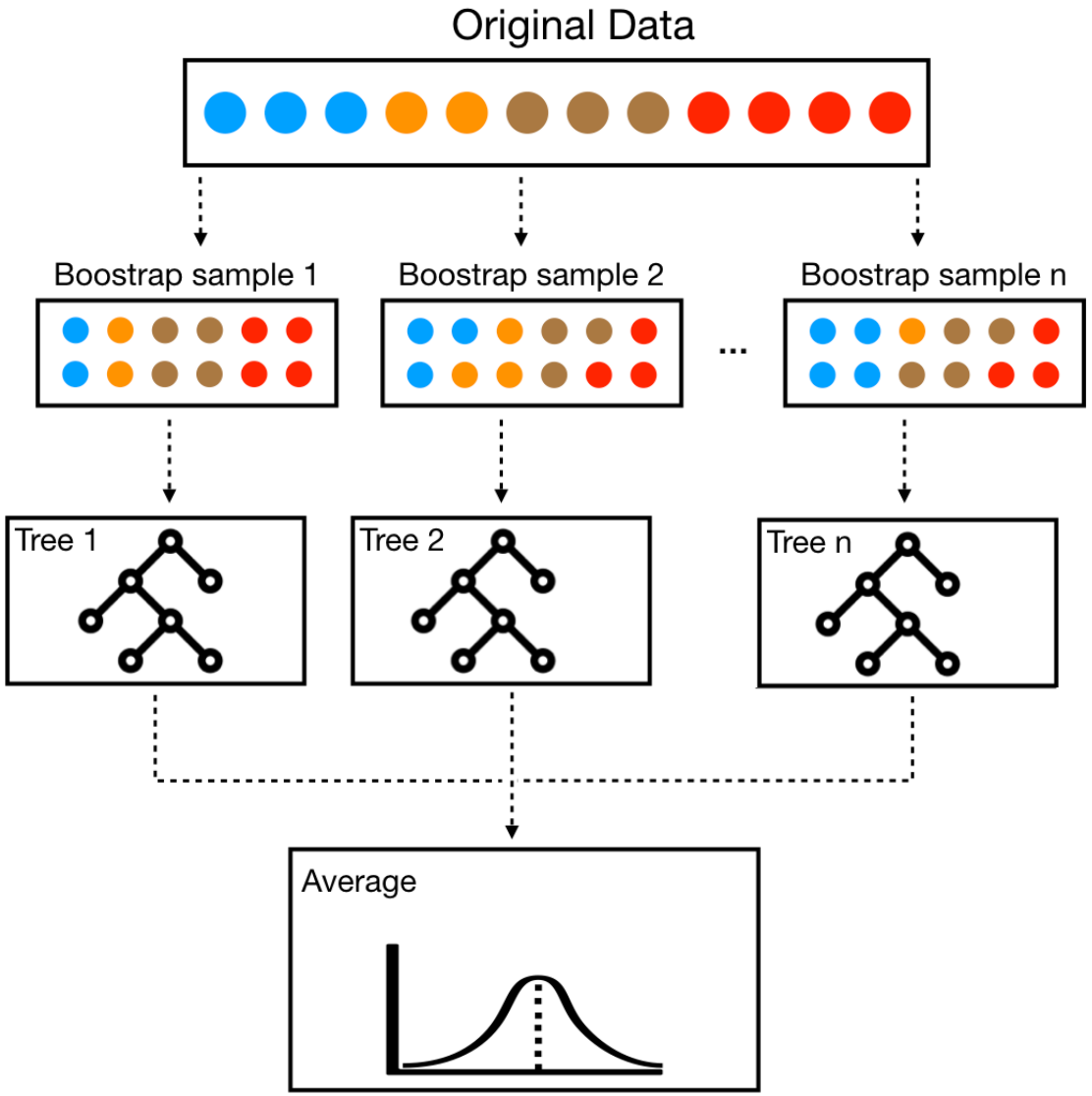
N = Total no. of decision trees

CMD

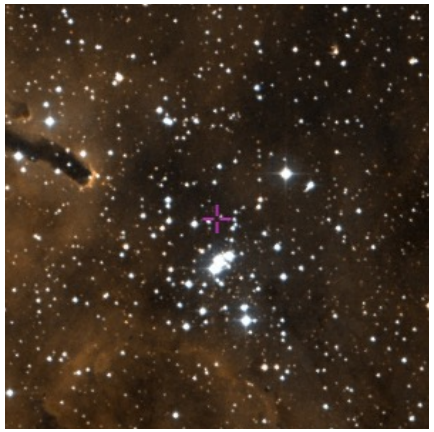

Pre-Main Sequence
Main Sequence
M_G [mag]
$G_{BP} - G_{RP}$ [mag]
Number of stars

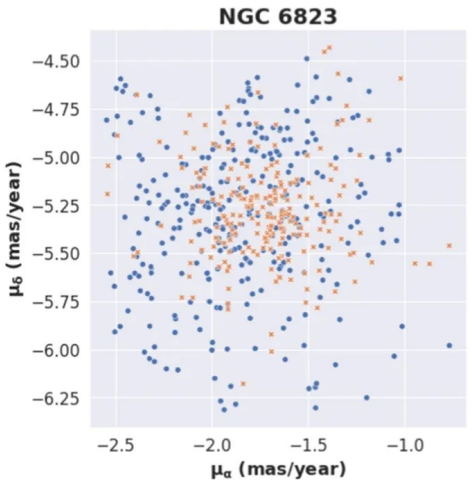

(a)
(b)
(c)
(d)
binary
single
Mock OC
NGC3532
number density [arbitrary]
G
$G_{BP} - G_{RP}$

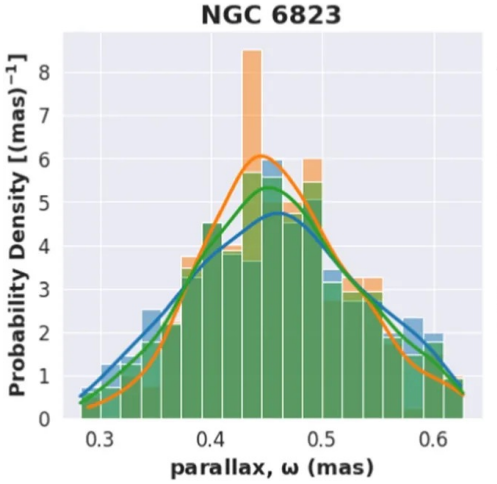
Bagging

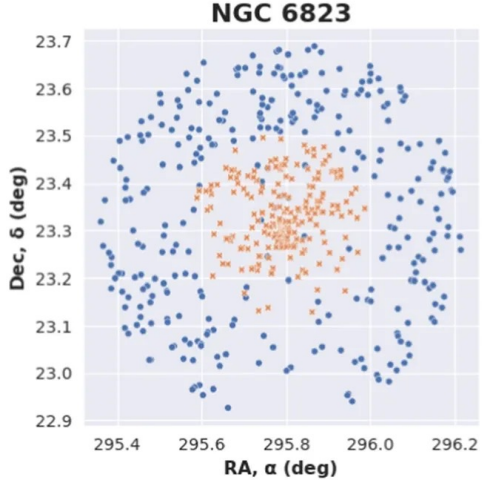

Original Data
Boostrap sample 1
Boostrap sample 2
...
Boostrap sample n
Tree 1
Tree 2
Tree n
Average

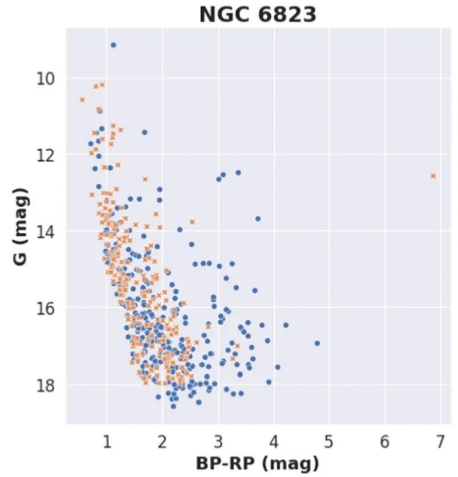
Cluster Image (DSS)	Proper Motion (μ_α, μ_δ) Distribution	Parallax (ω) Distribution	Member Distribution in (α, δ)	CMD

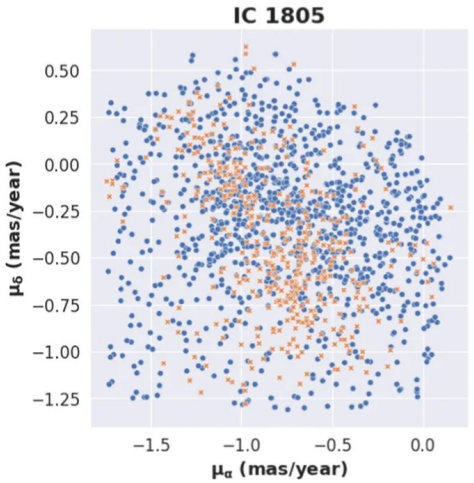


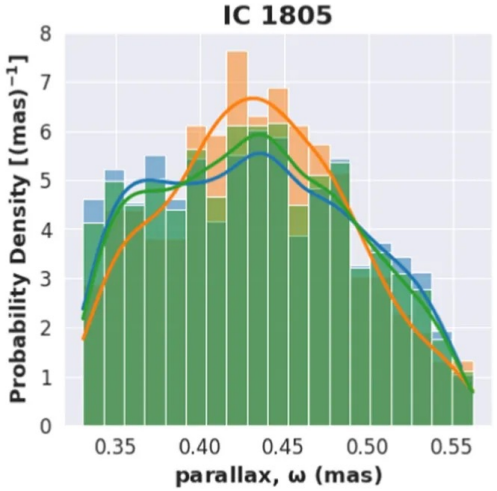


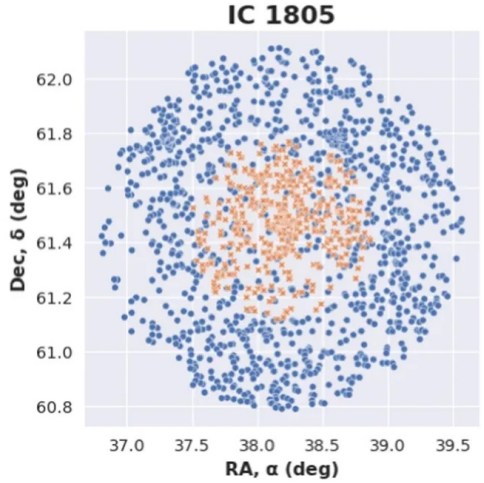


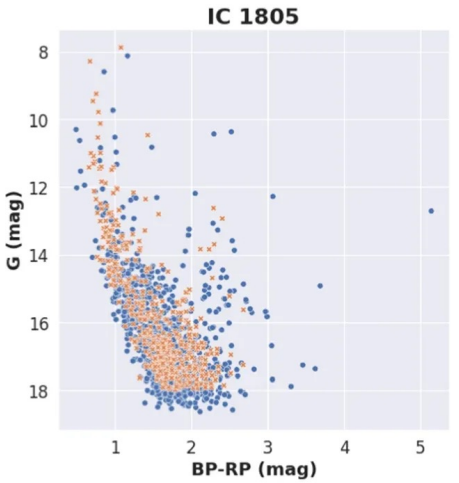








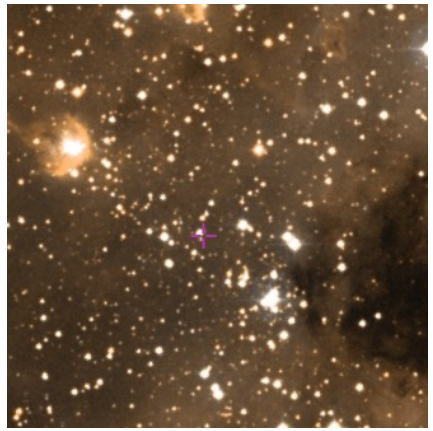


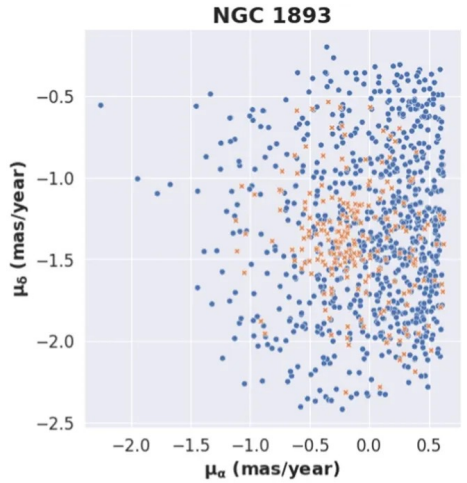


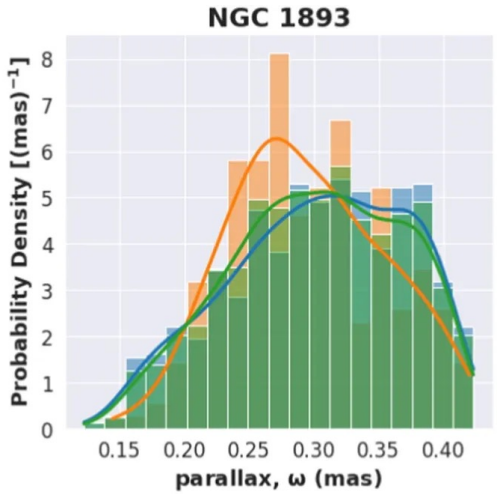
| Legend: CG Members New Members Combined |

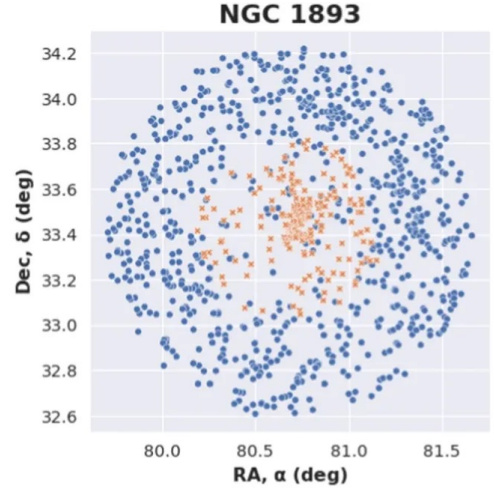
Mahmudunnobe et. al. 2021

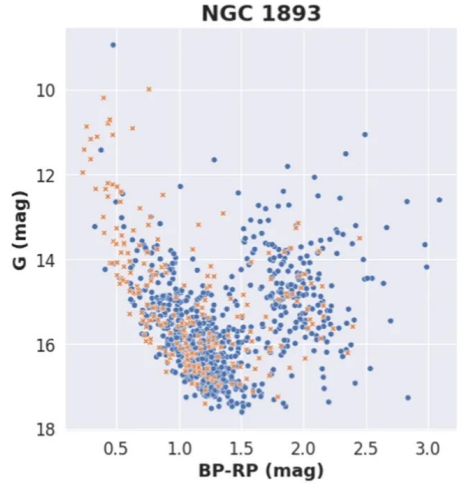
Cluster Image (DSS)	Proper Motion (μ_α, μ_δ) Distribution	Parallax (ω) Distribution	Member Distribution in (α, δ)	CMD

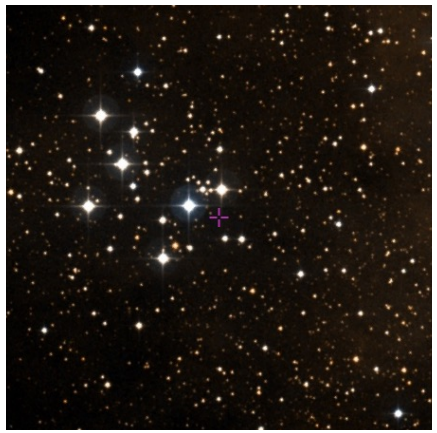


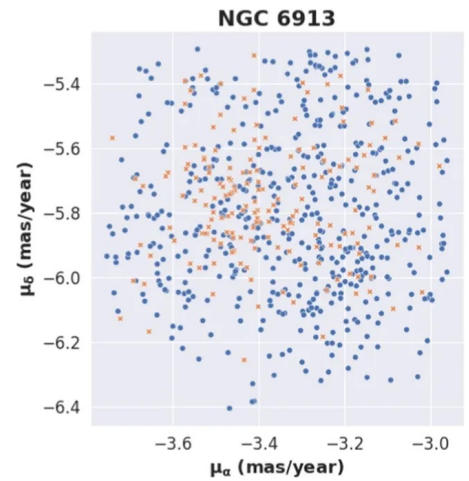


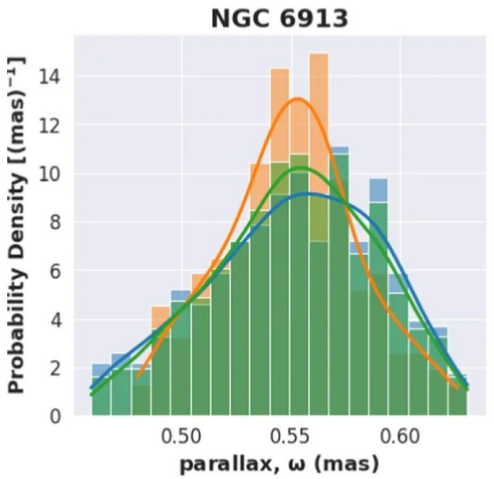


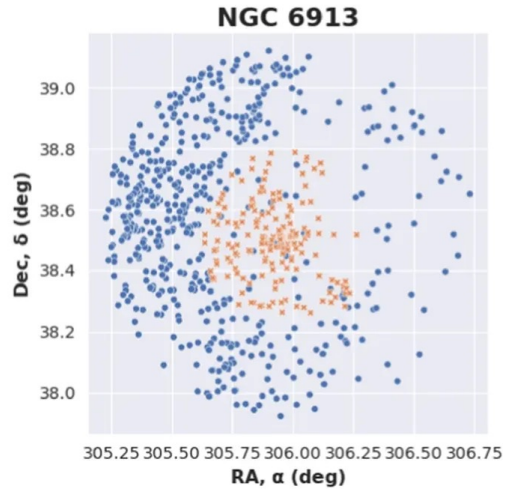


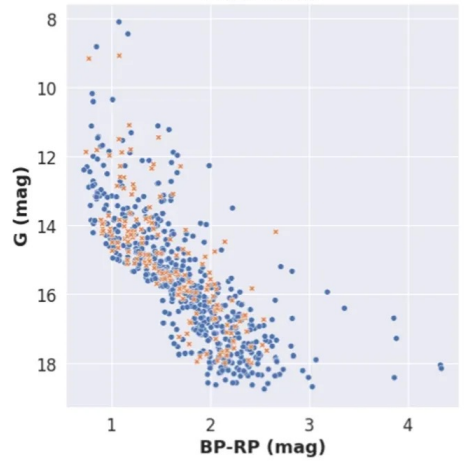












| Legend: CG Members | New Members | Combined |

Mahmudunnobe et. al. 2021