
# Clustering for Astrophysics

## Python Meetings Season 2 Episode 1

Matthias Gerlach, 10/04/26

# Overview
## Clustering for Astrophysics

1. Astrophysical data
2. Clustering types
3. Preparing the data
4. K-Means
5. ~~Gaussian Mixture Modelling~~
6. DBSCAN / HDBSCAN
7. Final remarks

} + some examples along the way! (> 20 citations generally)

# Astrophysical Data
## what do we deal with on a daily basis?

- Lots of different types of data:

1. **Astrometry** (sky coordinates, distances/parallaxes, proper motions and radial velocities, ...)
2. **Photometry** (apparent magnitudes, colors, light curves, ...)
3. **Spectroscopy** (rotation, velocity dispersion, temperature, metallicity, ...)
4. **Imaging** (morphology, ...)
5. **N-body / Hydrodynamical Simulations** (time evolution, velocities, energies, metallicities, morphologies, hierarchical formation, ...)

# Astrophysical Data
## what do we deal with on a daily basis?

- Examples of classification/clustering problems one might encounter

1. **Astrometry:** Separate MW stars into disk, bulge and halo components
2. **Photometry:** Discover strange pulsating stars (time varying mean magnitudes, amplitudes, periods)
3. **Spectroscopy:** Classify $H\alpha$ emission lines from Be stars
4. **Imaging:** Automatically count rings in protoplanetary disks, classify galaxies by morphology
5. **N-body / Hydrodynamical Simulations:** Detect formation of dark matter subhalos

# Clustering
## is our friend

- **Definition:** any type of unsupervised algorithm that tries to separate data into groups (**clusters**) with features more similar to those of the same group than those of other groups, according to a chosen arbitrary notion of **similarity** or **distance** metric.
- **Use cases:**
  - Exploratory phase
  - Anomaly detection
  - Classification
- Whenever we suspect NATURE is organizing stuff into **groups**, but we still don't quite know what those groups are or what their characteristics are, **clustering** is useful.

# Clustering
## types

- **Distance based:** K-Means
- **Density based:** DBSCAN / HDBSCAN
- ~~**Model based:** Gaussian Mixture Modelling (GMM)~~

Won’t go into the math. Just the intuition. If you are interested in learning more, check out:

- ✦ Géron, A. (2022). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.
- ✦ Statquest’s Machine Learning Fundamentals YT course
- ✦ Fotopoulou, S. (2024). A review of unsupervised learning in astronomy

# K-Means

```
from sklearn.cluster import KMeans
```

# K-Means

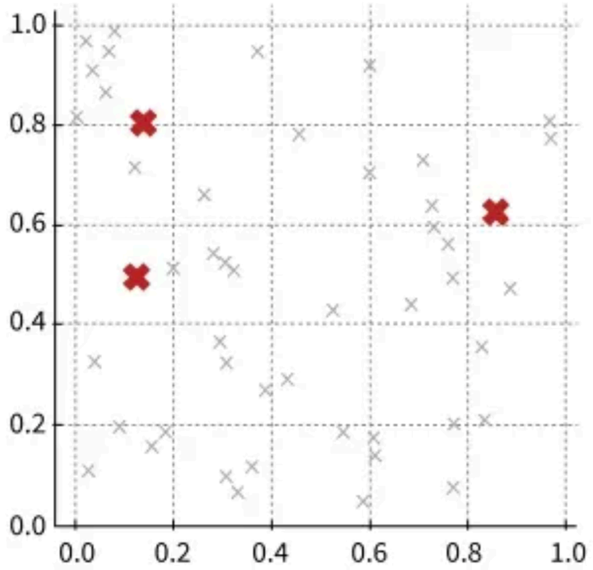
## distance based clustering

- Separates data into $k$ distinct, non-overlapping clusters by minimizing the euclidean distance between data points and their cluster centroids. It iteratively assigns data points to the nearest center and recalculates centers until convergence.

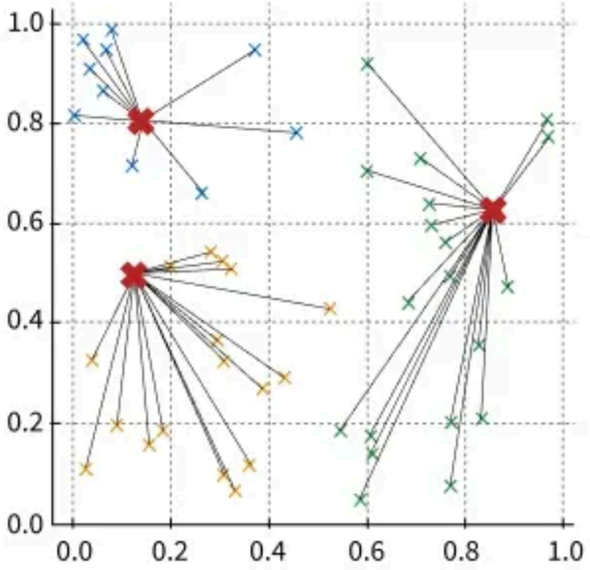
1. Initialize $k$ cluster centroids
2. Assign each datapoint to closest cluster centroid
3. Recompute cluster centroids as the mean of all data in that cluster
4. Repeat steps 2 and 3 until convergence

# K-Means

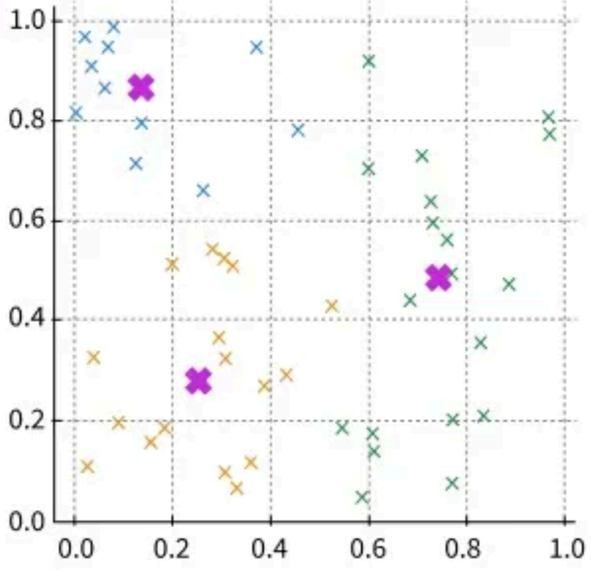
## distance based clustering



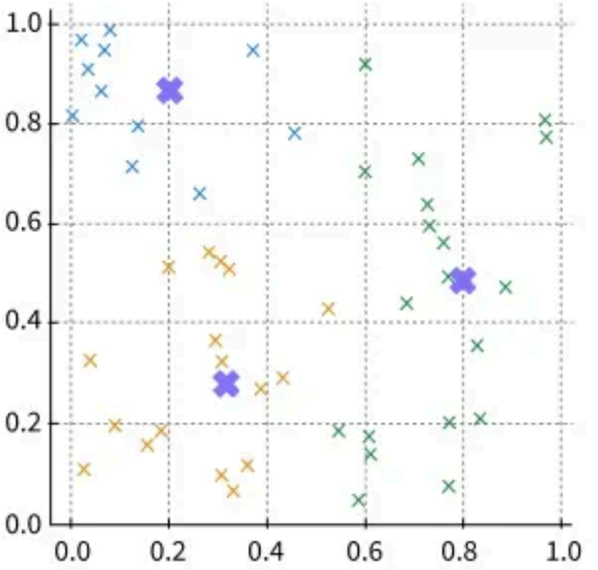

**1. Initialize centroids**




**2. Assign to cluster**




**3. Recompute centroids**

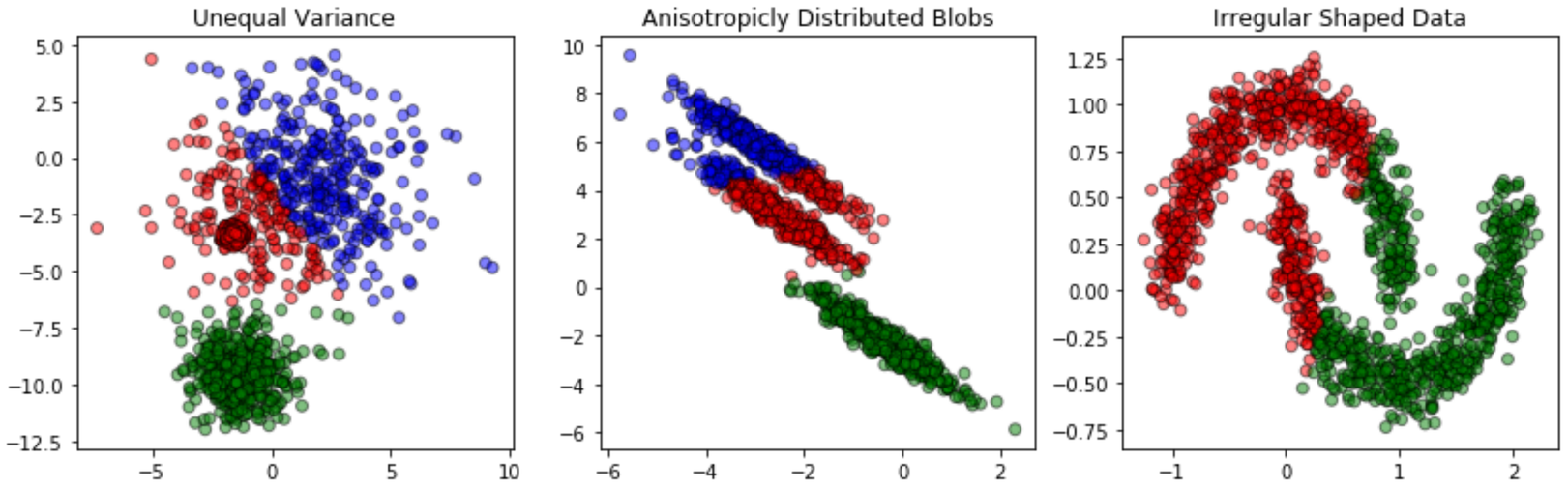



**3. Repeat until convergence**

https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/

# K-Means
## distance based clustering

- **Pros:** Intuitive, Computationally light
- **Cons:** Must manually choose number of clusters, assumes non-overlapping clusters and euclidean distance metric




https://zerowithdot.com/mistakes-with-k-means-clustering/

k-means be like:

JOURNAL ARTICLE

# Reproducible *k*-means clustering in galaxy feature data from the GAMA survey FREE

Sebastian Turner ✉, Lee S Kelvin, Ivan K Baldry, Paulo J Lisboa, Steven N Longmore, Chris A Collins, Benne W Holwerda, Andrew M Hopkins, Jochen Liske

*Monthly Notices of the Royal Astronomical Society*, Volume 482, Issue 1, January 2019, Pages 126–150, https://doi.org/10.1093/mnras/sty2690

**Published:** 05 October 2018 **Article history** ▾

**ABSTRACT**

A fundamental bimodality of galaxies in the local Universe is apparent in many of the features used to describe them. Multiple sub-populations exist within this framework, each representing galaxies following distinct evolutionary pathways. Accurately identifying and characterizing these sub-populations requires that a large number of galaxy features be analysed simultaneously. Future galaxy surveys such as LSST and Euclid will yield data volumes for which traditional approaches to galaxy classification will become unfeasible. To address this, we apply a robust *k*-means unsupervized clustering method to feature data derived from a sample of 7338 local-Universe galaxies selected from the Galaxy And Mass Assembly (GAMA) survey. This allows us to partition our sample into *k* clusters without the need for training on pre-labelled data, facilitating a full census of our high-dimensionality feature space and guarding against stochastic effects. We find that the local galaxy population natively splits into 2, 3, 5, and a maximum of six sub-populations, with each corresponding to a distinct ongoing evolutionary mechanism. Notably, the impact of the local environment appears strongly linked with the evolution of low-mass ($M_* < 10^{10}$ $M_\odot$) galaxies, with more massive systems appearing to evolve more passively from the blue cloud on to the red sequence. With a typical run time of ~3 min per value of *k* for our galaxy sample, we show how *k*-means unsupervized clustering is an ideal tool for future analysis of large extragalactic data sets, being scalable, adaptable, and providing crucial insight into the fundamental properties of the local galaxy population.

Turner et al. (2019)

Uses K-means as a galaxy classification solution for next-generation extragalactic surveys, working with 7,338 galaxies from GAMA using five features including stellar mass ($M_*$), specific star formation rate ($\mathrm{sSFR}$), color ($u - r$), half-light radius ($r_{1/2}$), and Sérsic index ($n$)

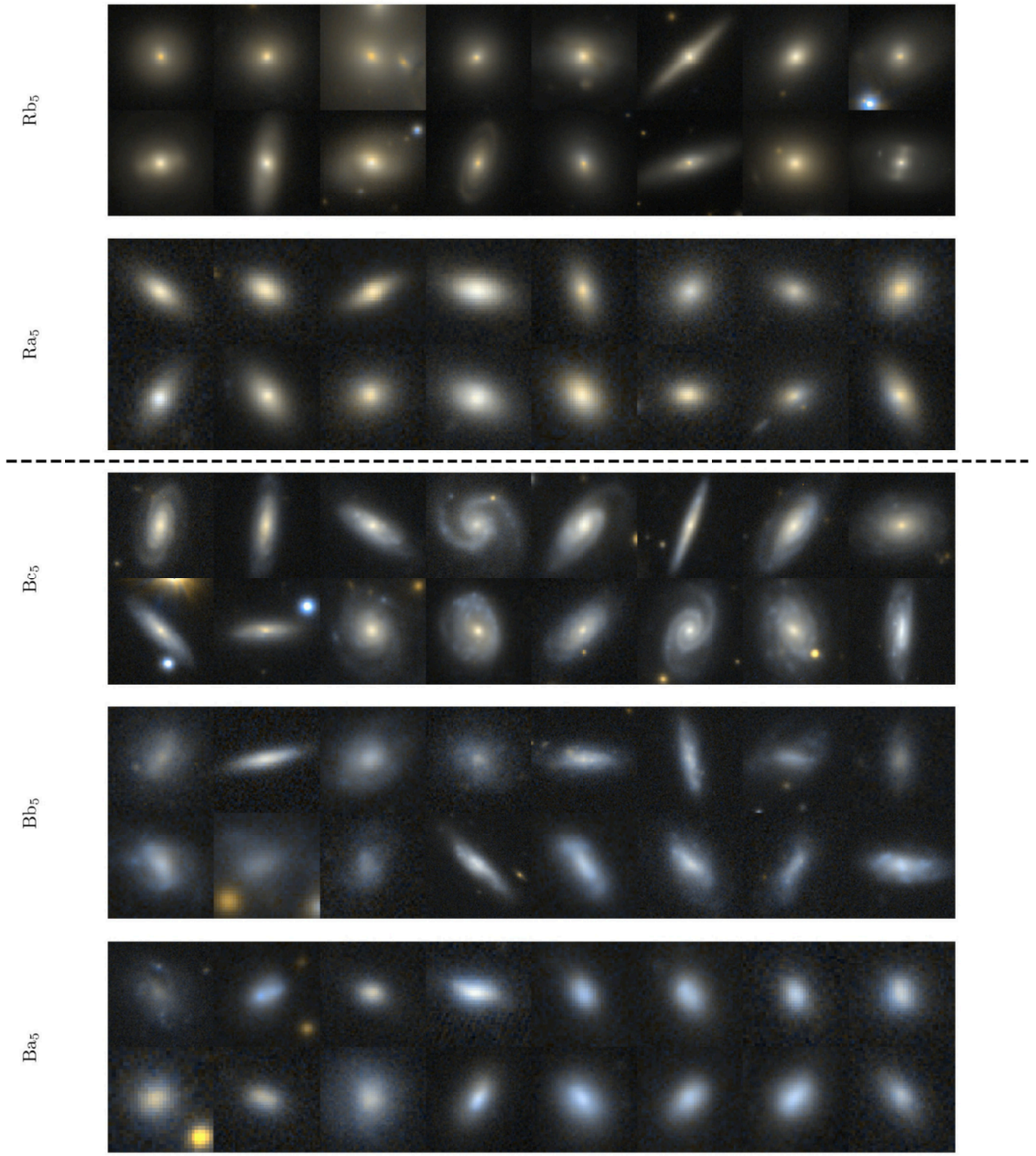



**Figure C3.** Example postage stamps of galaxies in each of the clusters in $k = 5$. The dashed black line separates the two superclusters that `k-means` finds. See Section 4.3 for discussion.

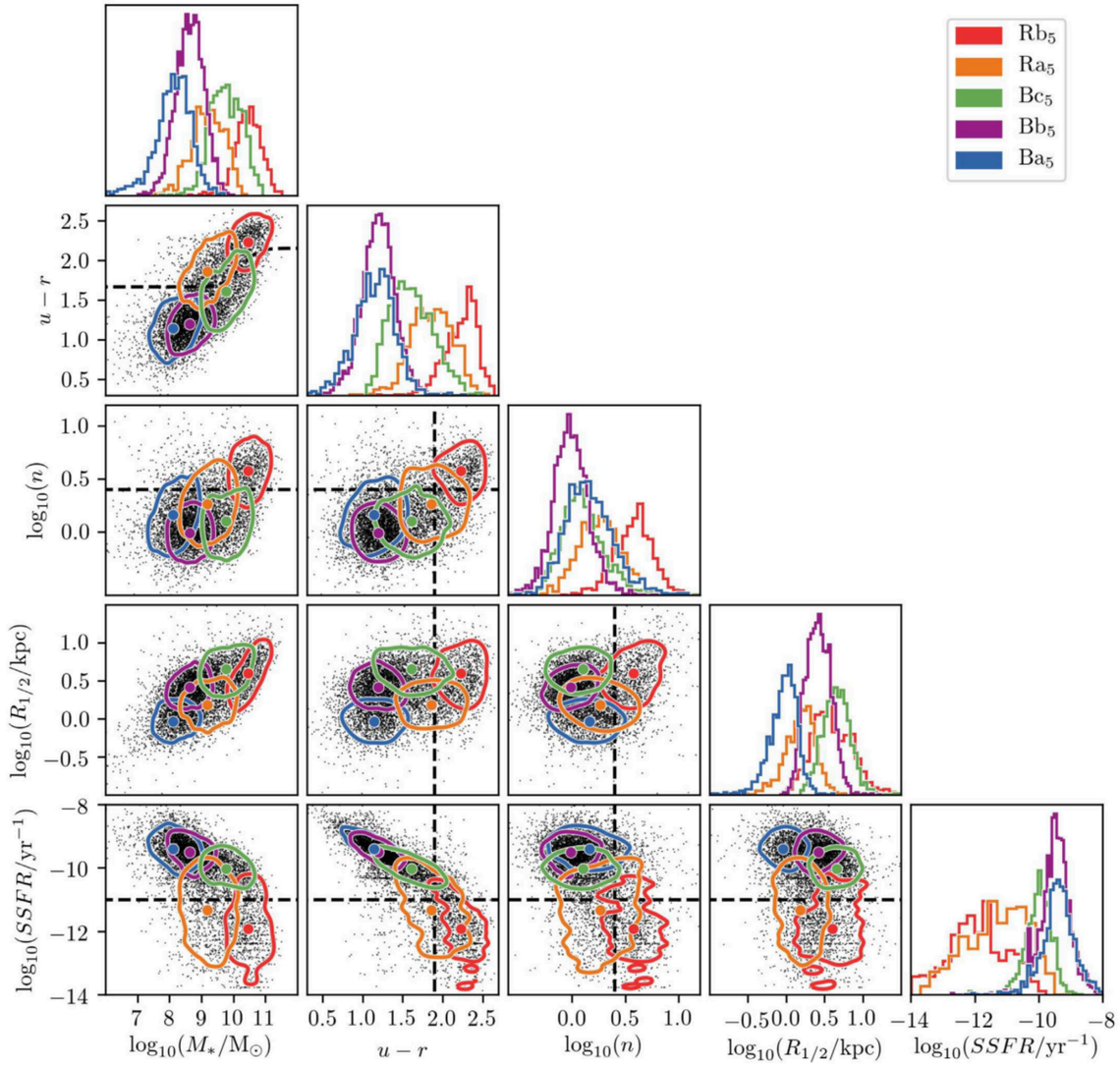



**Figure 9.** A profile of $k = 5$. Clusters are represented using coloured histograms and con tours, and their centroids are marked using filled circles of the same colour.

Turner et al. (2019)

A&A, 700, A259 (2025)
https://doi.org/10.1051/0004-6361/202555620


Astronomy & Astrophysics

# Semi-supervised classification of stars, galaxies and quasars using K-means and random-forest approaches

V. Asadi[1,⋆], H. Haghi[1,2,3,⋆], and A. H. Zonoozi[1,2]

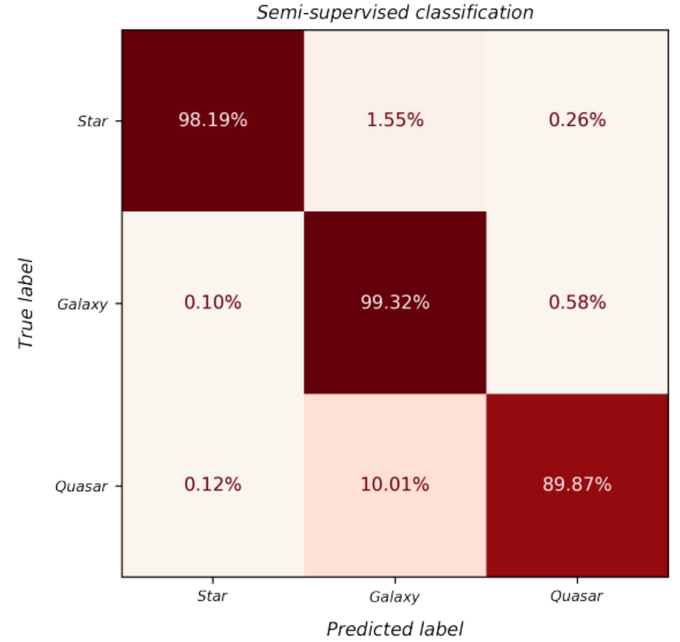
[1] Department of Physics, Institute for Advanced Studies in Basic Sciences (IASBS), PO Box 11365-9161, Zanjan, Iran
[2] Helmholtz-Institut für Strahlen-und Kernphysik (HISKP), Universität Bonn, Nussallee 14–16, 53115 Bonn, Germany
[3] School of Astronomy, Institute for Research in Fundamental Sciences (IPM), PO Box 19395-5531, Tehran, Iran

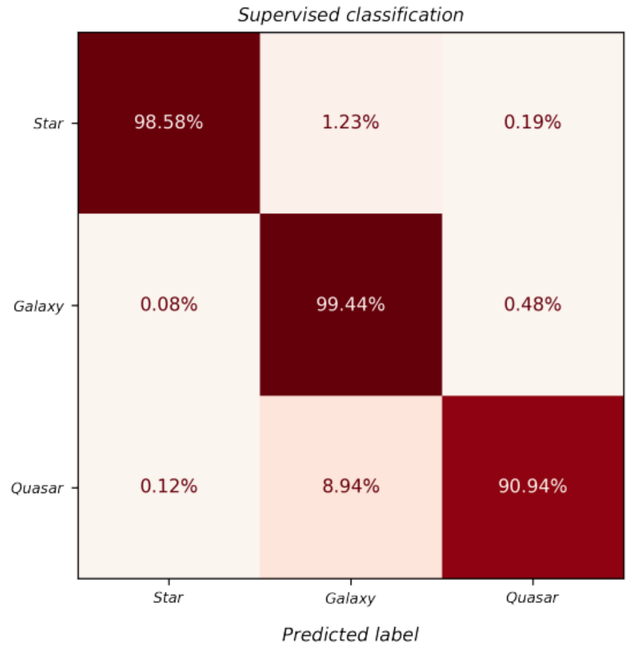
Received 21 May 2025 / Accepted 16 July 2025

**ABSTRACT**

*Context.* Classifying stars, galaxies, and quasars is essential for understanding cosmic structure and evolution; however, the vast data from modern surveys make manual classification impractical, while supervised learning methods remain constrained by the scarcity of labeled spectroscopic data.
*Aims.* We aim to develop a scalable, label-efficient method for astronomical classification by leveraging semi-supervised learning (SSL) to overcome the limitations of fully supervised approaches.
*Methods.* We propose a novel SSL framework combining K-means clustering with random forest classification. Our method partitions unlabeled data into 50 clusters, propagates labels from spectroscopically confirmed centroids to 95% of cluster members, and trains a random forest on the expanded pseudo-labeled dataset. We applied this to the CPz catalog, containing multi-survey photometric and spectroscopic data, and compared performance with a fully supervised random forest.
*Results.* Our SSL approach achieves F1 scores of 98.8%, 98.9%, and 92.0% for stars, galaxies, and quasars, respectively, closely matching the supervised method with F1 scores of 99.1%, 99.1%, and 93.1%, while outperforming traditional color-cut techniques. The method demonstrates robustness in high-dimensional feature spaces and superior label efficiency compared to prior work.
*Conclusions.* This work highlights SSL as a scalable solution for astronomical classification when labeled data is limited, though performance may be degraded in lower dimensional settings.

**Key words.** methods: data analysis – stars: general – galaxies: general – quasars: general




~48.000 objects. features: ugriz (SDSS), ZYJHKs (VISTA), W1W2 (WISE-ALLWISE)

Asadi et al. (2025)

# Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

```
from sklearn.cluster import DBSCAN
```

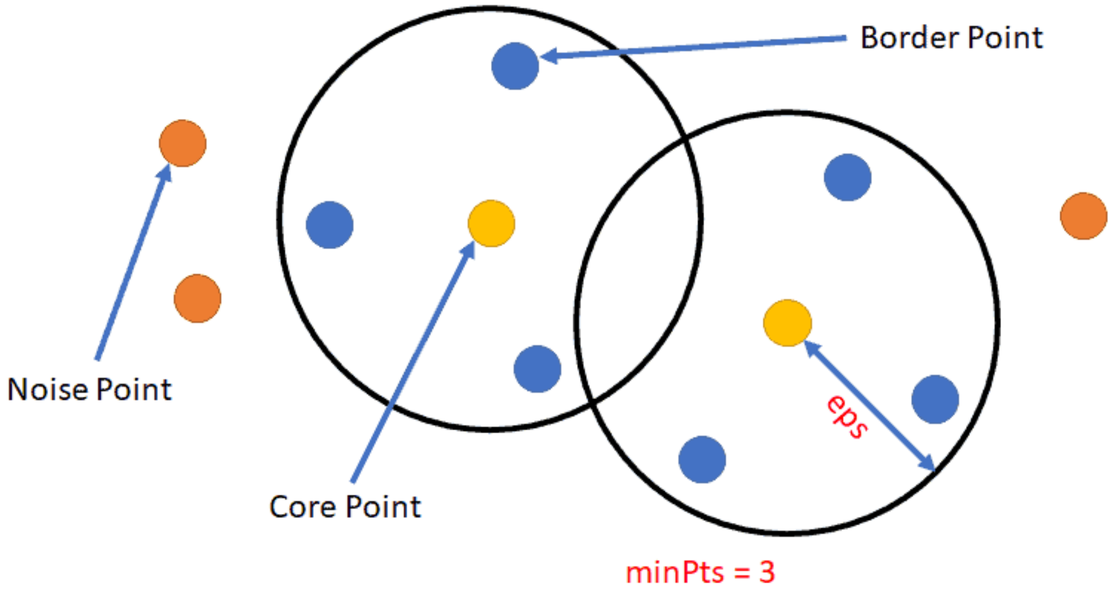
# DBSCAN

## density based clustering

- Groups together points that are closely packed (in **high density** regions) and labels isolated points as **noise**. It works by checking how many neighbors each point has within a chosen distance $\epsilon$. Dense regions become clusters, while sparse regions are ignored.

# DBSCAN
## density based clustering

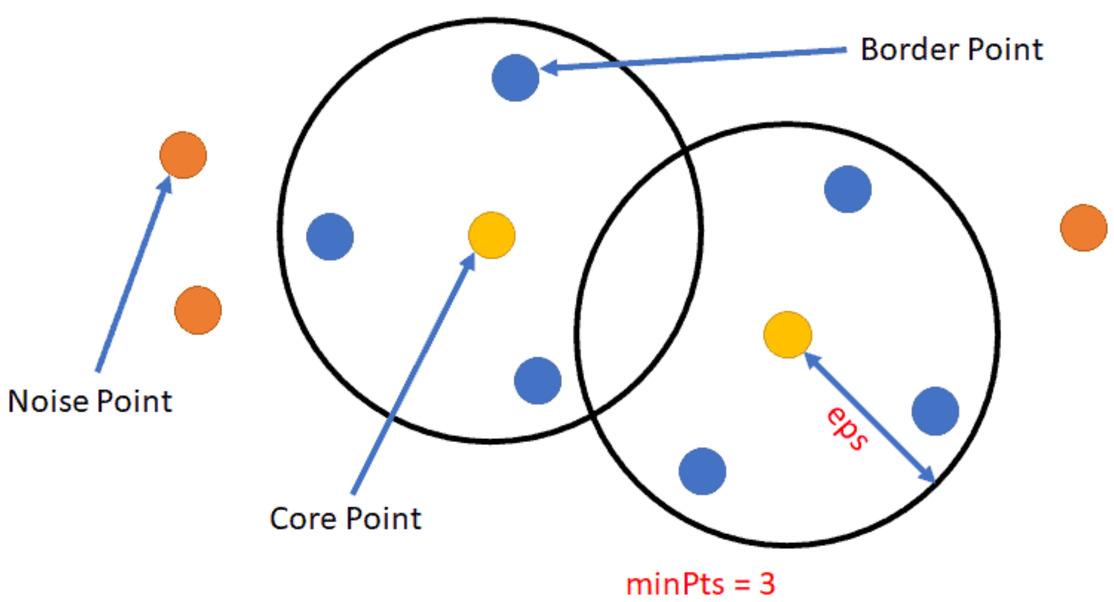
1. **Neighbors** are points that are close to each other by a distance metric $\epsilon$ (`eps`)
2. **Core points** are points with at least `minPts` **neighbors**. Points with less than `minPts` **neighbors** are **border points**.
3. **Core points** are part of the same cluster if they are **neighbors**. If not, they are from different clusters
4. **Border points** are added to each cluster if they have a **core point** as a **neighbor**.
5. **Border points** that do not have **core points** as **neighbors** are considered **noise points**.

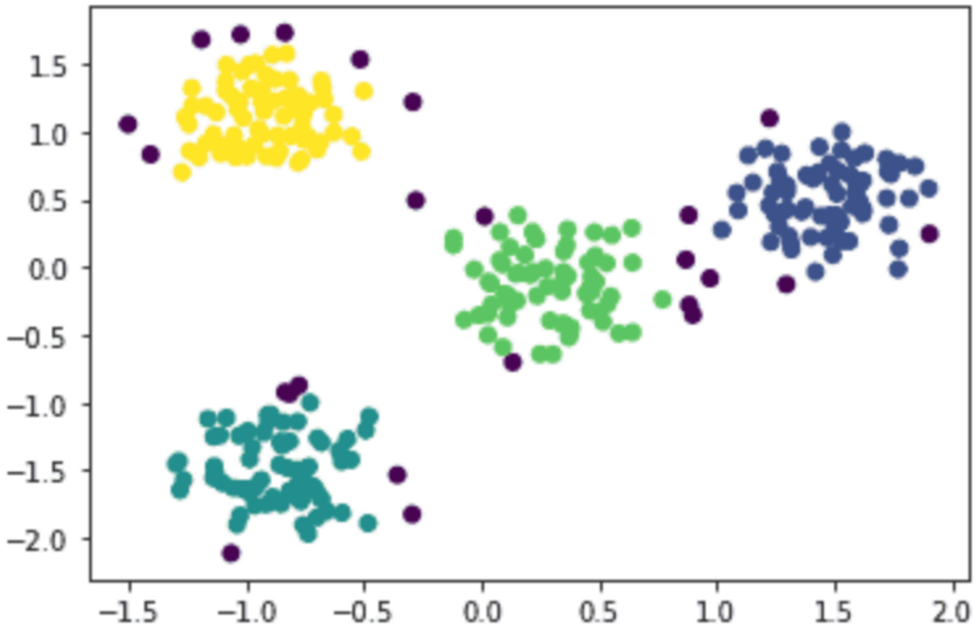



https://machinelearninggeek.com/dbscan-clustering/

# DBSCAN

## density based clustering



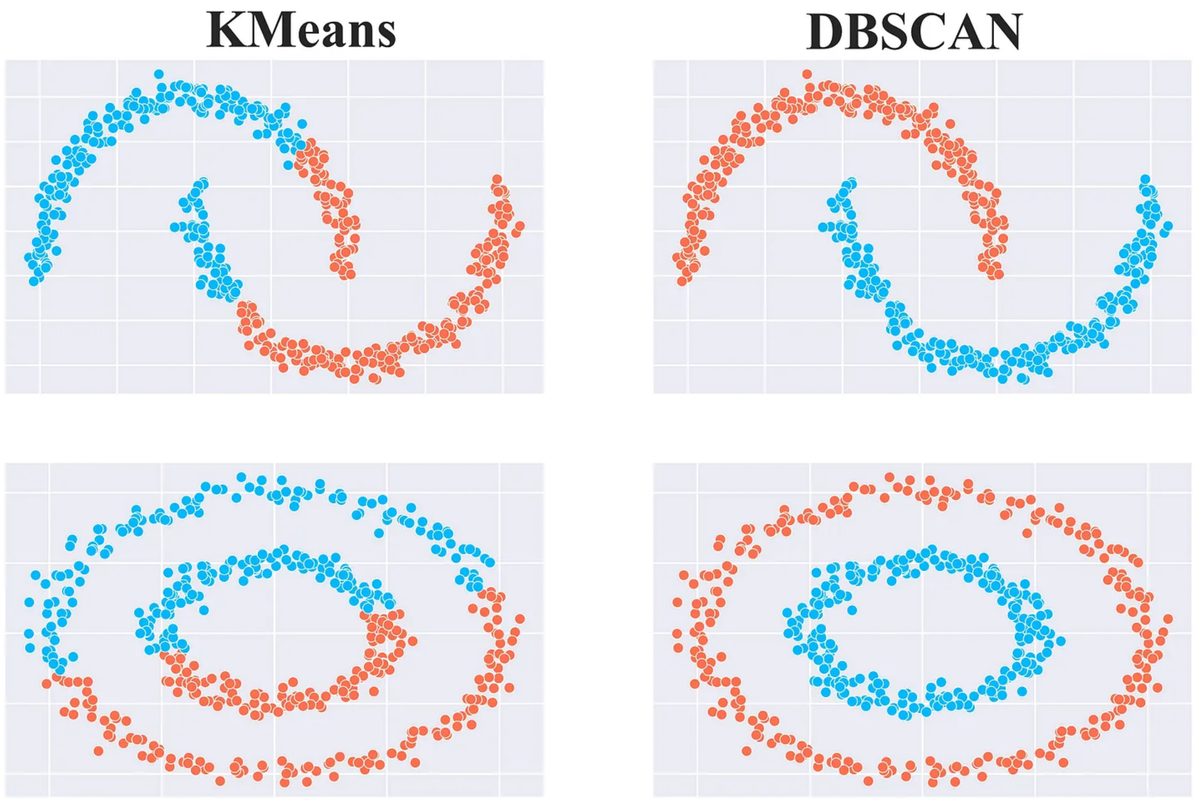




https://machinelearninggeek.com/dbscan-clustering/

# DBSCAN

## density based clustering

- **Pros:** Finds optimal number of clusters automatically. Basically solves all problems with K-Means
- **Cons:** Computationally heavy for large datasets. Still cannot deal with clusters with variable density. Sensible to parameter tuning.




https://blog.dailydoseofds.com/p/the-limitations-of-dbscan-clustering

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **481,** 1624–1630 (2018)
Advance Access publication 2018 September 4

doi:10.1093/mnras/sty2361

# TSARDI: a Machine Learning data rejection algorithm for transiting exoplanet light curves

D. Mislis,★ S. Pyrzas and K. A. Alsubai

*Qatar Environment and Energy Research Institute (QEERI), Hamad Bin Khalifa University (HBKU), Qatar Foundation, PO Box 5825, Doha, Qatar*

**ABSTRACT**

We present TSARDI, an efficient rejection algorithm designed to improve the transit detection efficiency in data collected by large-scale surveys. TSARDI is based on the Machine Learning clustering algorithm DBSCAN, and its purpose is to serve as a robust and adaptable filter aiming to identify unwanted noise points left over from data detrending processes. TSARDI is an unsupervised method, which can treat each light curve individually; there is no need of previous knowledge of any other field light curves. We conduct a simulated transit search by injecting planets on real data obtained by the QES project and show that TSARDI leads to an overall transit detection efficiency increase of ∼11 per cent, compared to results obtained from the same sample, but using a standard sigma-clip algorithm. For the brighter end of our sample (host star magnitude <12), TSARDI achieves a detection efficiency of ∼80 per cent of injected planets. While our algorithm has been developed primarily for the field of exoplanets, it is easily adaptable and extendable for use in any time series.

**Key words:** methods: data analysis – techniques: photometric – Planetary Systems.

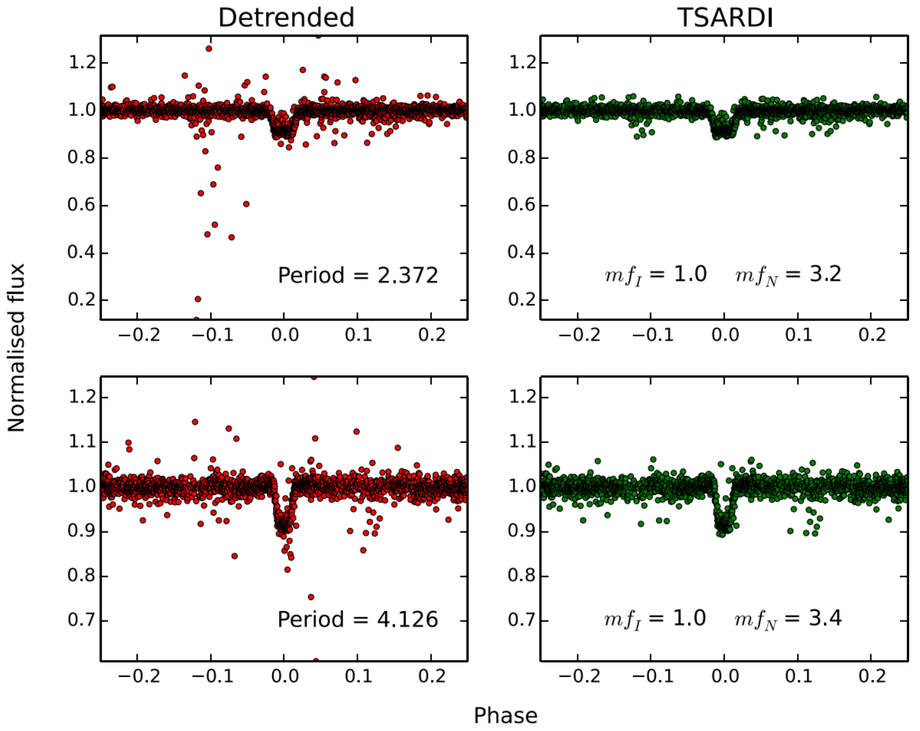



**Figure 8.** Two examples of transits of a $2\,R_J$ object (with orbital periods indicated), yielding depths of 0.088 and 0.132 (top and bottom, respectively). We show the detrended light curve on the left-hand panels; and the final TSARDI- filtered light curves on the right-hand panels. We also indicate the corresponding ($mf_I$, $mf_N$) values.

Mislis et al. (2018)

A&A 618, A59 (2018)
https://doi.org/10.1051/0004-6361/201833390


Astronomy & Astrophysics

# A new method for unveiling open clusters in *Gaia*

## New nearby open clusters confirmed by DR2

A. Castro-Ginard[1], C. Jordi[1], X. Luri[1], F. Julbe[1], M. Morvan[1,2], L. Balaguer-Núñez[1], and T. Cantat-Gaudin[1]

[1] Dept. Fisica Quantica i Astrofisica, Institut de Ciències del Cosmos (ICCUB), Universitat de Barcelona (IEEC-UB), Martí Franquès 1, 08028 Barcelona, Spain
e-mail: acastro@fqa.ub.edu
[2] Mines Saint-Etienne, Institut Henri Fayol, 42023 Saint-Etienne, France

Received 8 May 2018 / Accepted 11 June 2018

**ABSTRACT**

*Context.* The publication of the *Gaia* Data Release 2 (*Gaia* DR2) opens a new era in astronomy. It includes precise astrometric data (positions, proper motions, and parallaxes) for more than 1.3 billion sources, mostly stars. To analyse such a vast amount of new data, the use of data-mining techniques and machine-learning algorithms is mandatory.
*Aims.* A great example of the application of such techniques and algorithms is the search for open clusters (OCs), groups of stars that were born and move together, located in the disc. Our aim is to develop a method to automatically explore the data space, requiring minimal manual intervention.
*Methods.* We explore the performance of a density-based clustering algorithm, DBSCAN, to find clusters in the data together with a supervised learning method such as an artificial neural network (ANN) to automatically distinguish between real OCs and statistical clusters.
*Results.* The development and implementation of this method in a five-dimensional space ($l$, $b$, $\varpi$, $\mu_{\alpha^*}$, $\mu_\delta$) with the Tycho-Gaia Astrometric Solution (TGAS) data, and a posterior validation using *Gaia* DR2 data, lead to the proposal of a set of new nearby OCs.
*Conclusions.* We have developed a method to find OCs in astrometric data, designed to be applied to the full *Gaia* DR2 archive.

**Key words.** surveys – open clusters and associations: general – astrometry – methods: data analysis

**DBSCAN + ANN**

**Appendix A: Colour-magnitude diagrams of the identified open clusters**

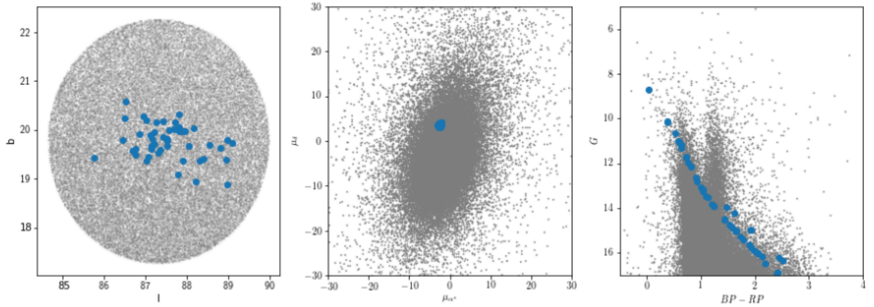


**Fig. A.1.** Member stars (blue) together with field stars (grey) for UBC1 in ($l, b$) (*left panel*) and in proper motion space (*middle panel*). The shows the sequence of the identified members (outlining an empirical isochrone) (*right panel*).

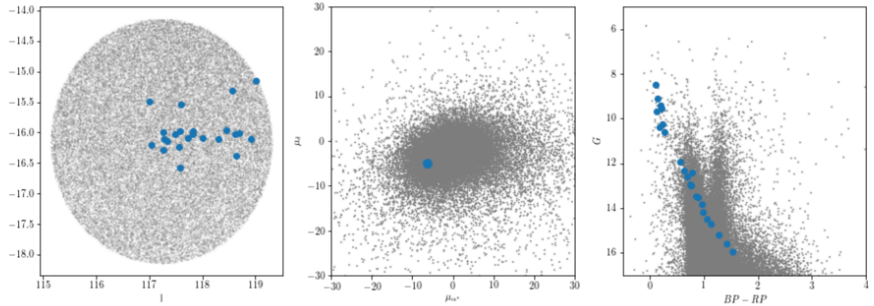


**Fig. A.2.** As in Fig. A.1 but for UBC2.

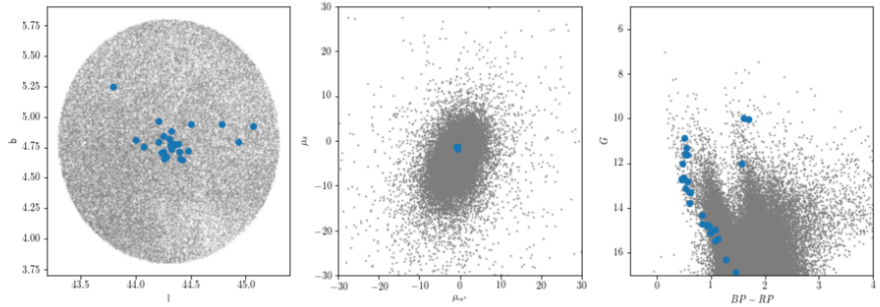


**Fig. A.3.** As in Fig. A.1 but for UBC3.

A. Castro-Ginard et al. (2018)

# Hierarchical DBSCAN (HDBSCAN)

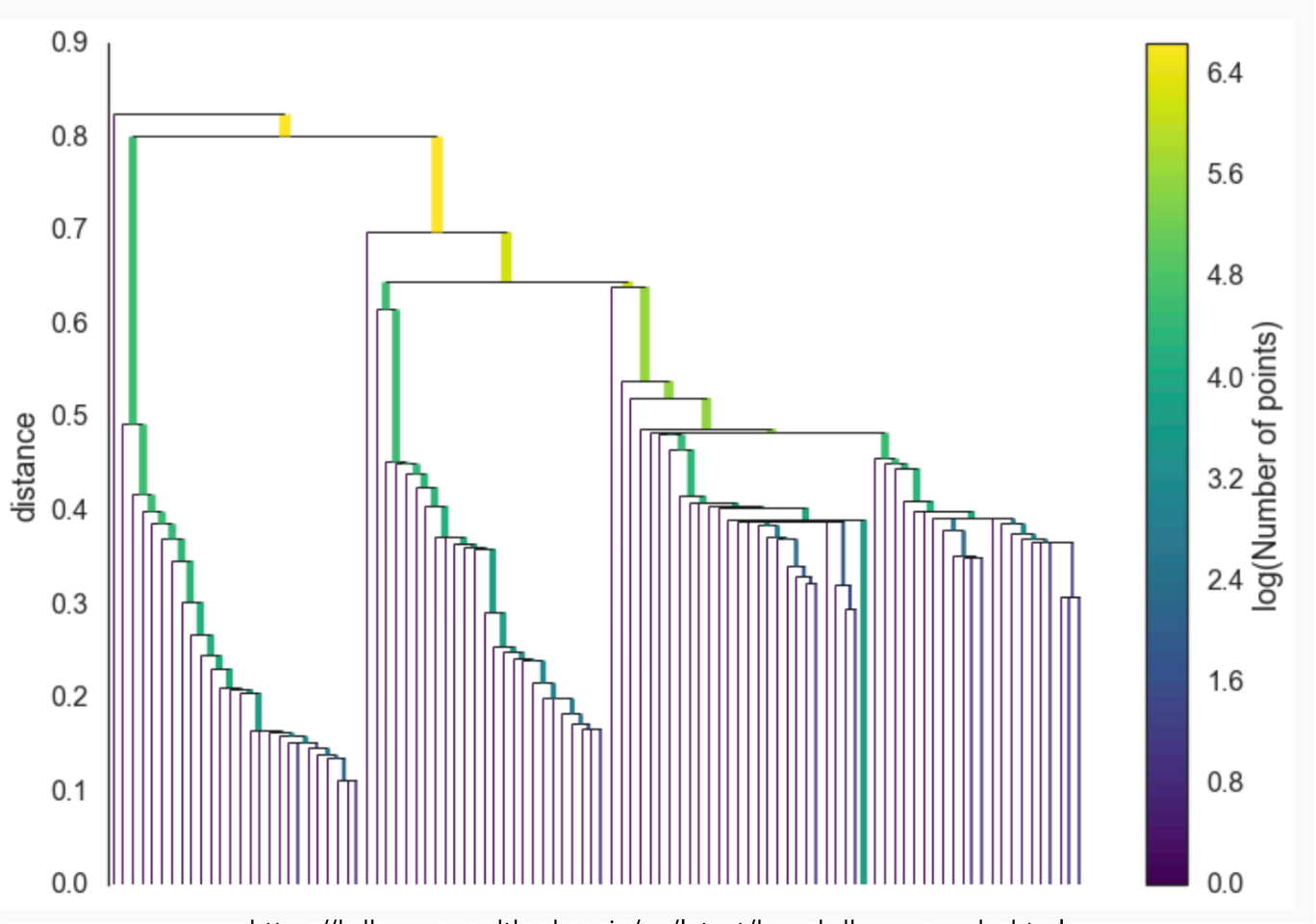
```
import hdbscan
```

# HDBSCAN

## density based clustering

Build upon DBSCAN but far more complex. In summary:

1. Run DBSCAN multiple times while **varying the density threshold** $\epsilon$. As we decrease density, clusters are **split up** into multiple clusters. We do this until every point is a cluster. We construct a **hierarchical tree** as shown.

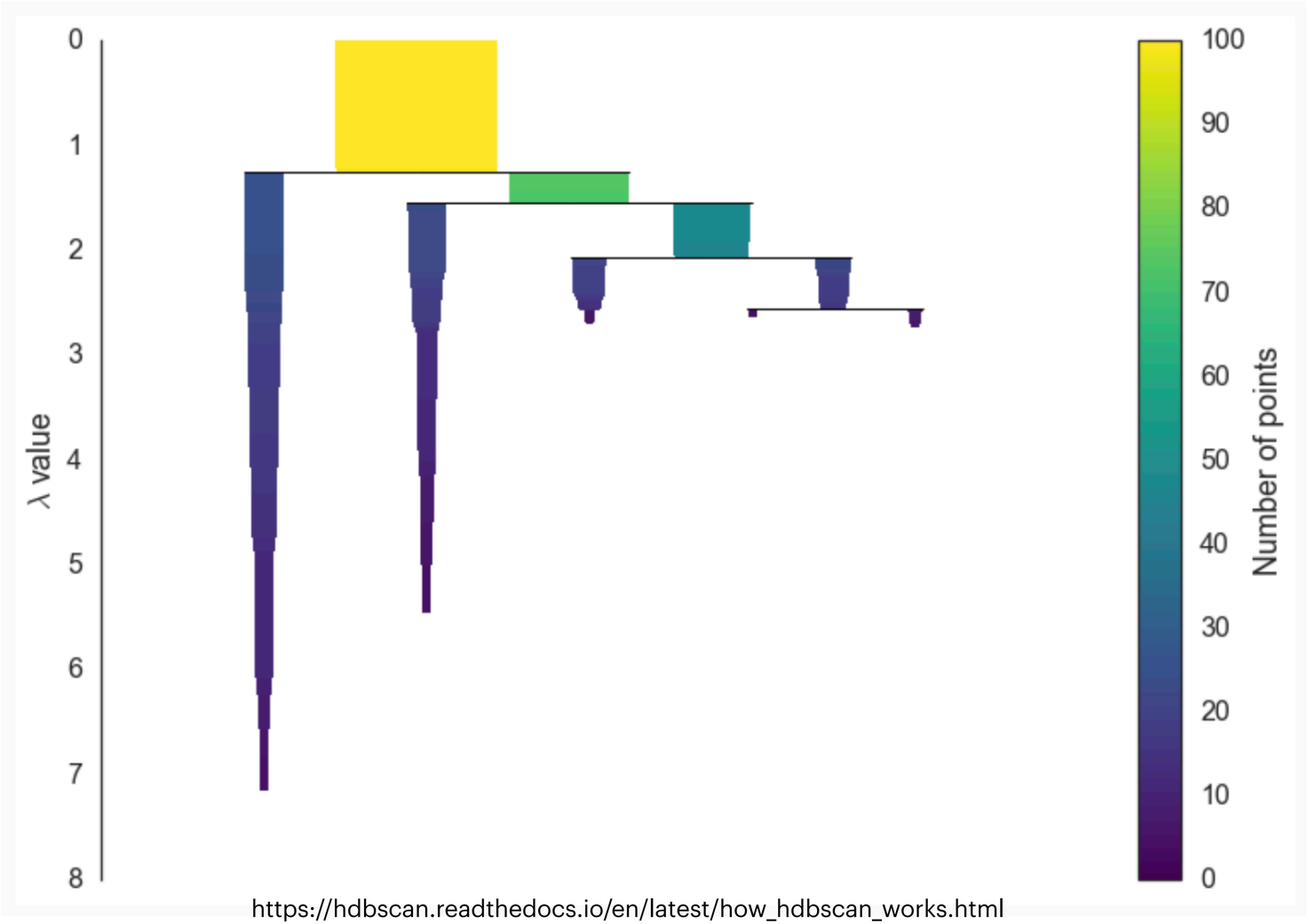



https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

# HDBSCAN

## density based clustering

2. We transform the hierarchical tree into a **condensed hierarchical tree** by considering that a cluster survived the density change ($\epsilon$) if it lost less than `min_cluster_size` points. If not, we consider that the cluster actually **split up** into two separate clusters.

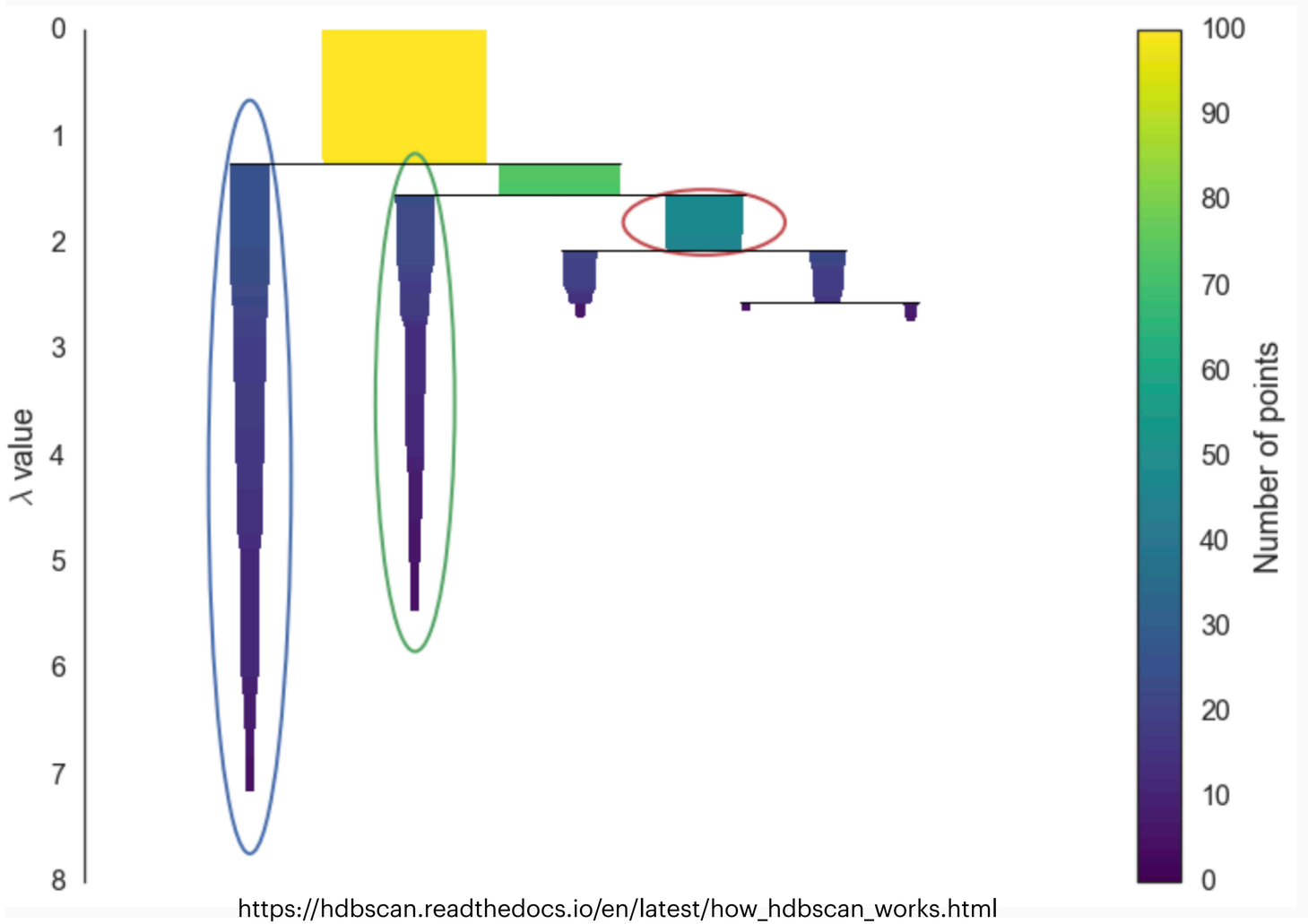



https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

# HDBSCAN

## density based clustering

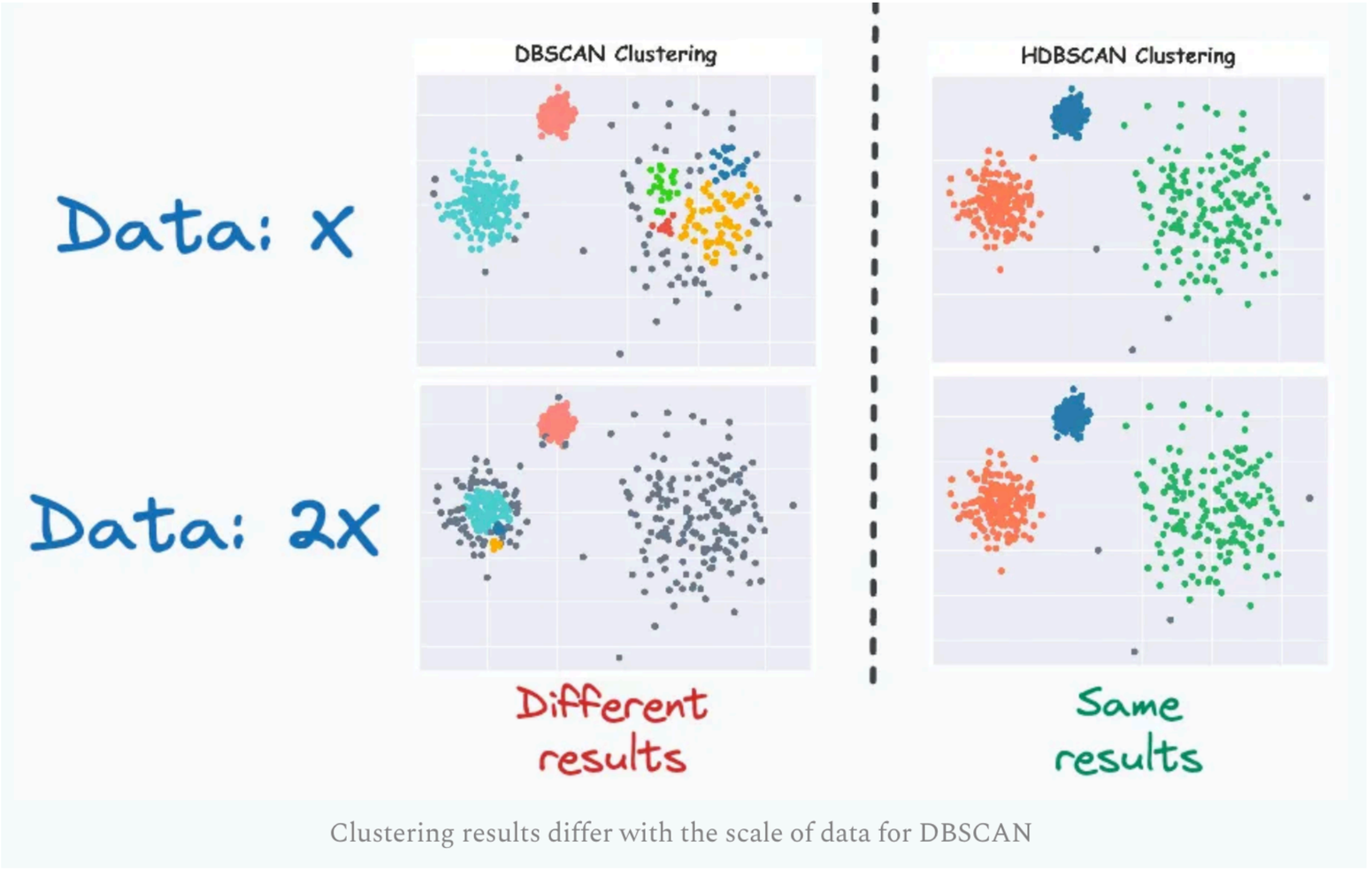
3. We keep only the most **stable clusters** across densities by integrating the areas of clusters in the condensed hierarchical tree. This way we can keep clusters with different densities!




https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

# HDBSCAN

## density based clustering




Clustering results differ with the scale of data for DBSCAN

https://blog.dailydoseofds.com/p/the-limitations-of-dbscan-clustering

A&A 633, A154 (2020)
https://doi.org/10.1051/0004-6361/201936648


Astronomy & Astrophysics

# Unsupervised star, galaxy, QSO classification

## Application of HDBSCAN★

C. H. A. Logan[1] and S. Fotopoulou[2]

[1] H. H. Wills Physics Laboratory, University of Bristol, Bristol, UK
e-mail: crispin.logan@bristol.ac.uk
[2] Centre for Extragalactic Astronomy, Department of Physics, Durham University, Durham DH1 3LE, UK
e-mail: sotiria.fotopoulou@durham.ac.uk

Received 6 September 2019 / Accepted 12 November 2019

**ABSTRACT**

*Context.* Classification will be an important first step for upcoming surveys aimed at detecting billions of new sources, such as LSST and Euclid, as well as DESI, 4MOST, and MOONS. The application of traditional methods of model fitting and colour-colour selections will face significant computational constraints, while machine-learning methods offer a viable approach to tackle datasets of that volume.
*Aims.* While supervised learning methods can prove very useful for classification tasks, the creation of representative and accurate training sets is a task that consumes a great deal of resources and time. We present a viable alternative using an unsupervised machine learning method to separate stars, galaxies and QSOs using photometric data.
*Methods.* The heart of our work uses Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) to find the star, galaxy, and QSO clusters in a multidimensional colour space. We optimized the hyperparameters and input attributes of three separate HDBSCAN runs, each to select a particular object class and, thus, treat the output of each separate run as a binary classifier. We subsequently consolidated the output to give our final classifications, optimized on the basis of their F1 scores. We explored the use of Random Forest and PCA as part of the pre-processing stage for feature selection and dimensionality reduction.
*Results.* Using our dataset of ~50 000 spectroscopically labelled objects we obtain F1 scores of 98.9, 98.9, and 93.13 respectively for star, galaxy, and QSO selection using our unsupervised learning method. We find that careful attribute selection is a vital part of accurate classification with HDBSCAN. We applied our classification to a subset of the SDSS spectroscopic catalogue and demonstrated the potential of our approach in correcting misclassified spectra useful for DESI and 4MOST. Finally, we created a multiwavelength catalogue of 2.7 million sources using the KiDS, VIKING, and ALLWISE surveys and published corresponding classifications and photometric redshifts.

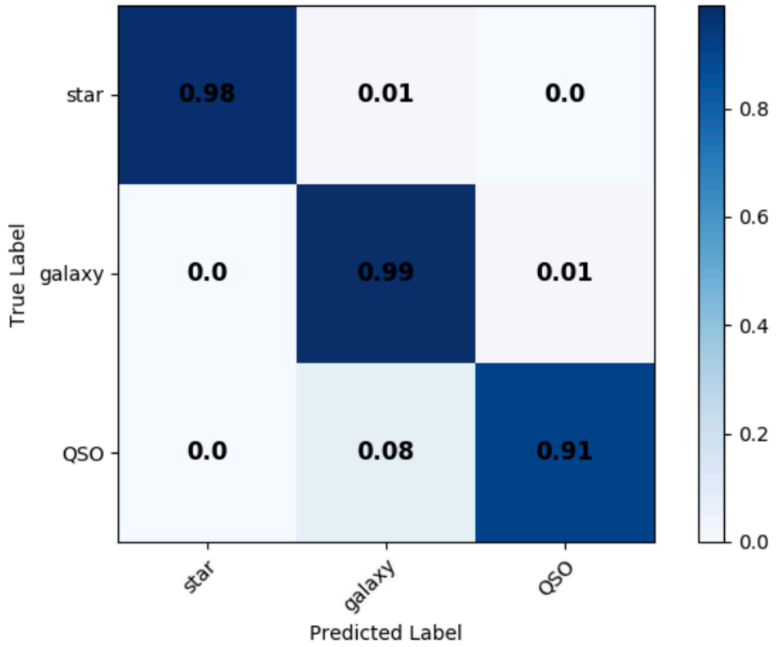
**Key words.** stars: general – galaxies: general – galaxies: active – methods: data analysis – surveys

**Table 2.** Top 10 attributes from the output of RF.

| STAR | GAL | QSO | ALL |
|---|---|---|---|
| $J_3 - W1$ | $K - Y_3$ | $z - u_3$ | $K - Y_3$ |
| $K - J_3$ | $K - J_3$ | $i - u_3$ | $K - J_3$ |
| $Y_3 - W1$ | $K - Z_3$ | $r - g_3$ | $K - H_3$ |
| $K - H_3$ | $K - H_3$ | $u_3 - z_3$ | $J_3 - W1$ |
| $J_3 - K_3$ | $J_3 - K_3$ | $u_3 - i_3$ | $J_3 - K_3$ |
| $H_3 - W1$ | $Y_3 - K_3$ | $Y - u_3$ | $Y_3 - W1$ |
| $K - Y_3$ | $J_3 - W1$ | $u - z$ | $H_3 - W1$ |
| $H_3 - K_3$ | $Y_3 - W1$ | $z - g_3$ | $H_3 - K_3$ |
| $Y_3 - W2$ | $J - K$ | $r - u_3$ | $J - K$ |
| $J - K$ | $H_3 - K_3$ | $u - i$ | $Y_3 - K_3$ |

**Notes.** The first three columns are the top 10 attributes for when the labels were binary for STAR/non-STAR, GAL/non-GAL, QSO/non-QSO and the fourth column is for the multi-label setup.

(b)




(d)

Logan & Fotopoulou (2019)

The Astronomical Journal, 158:122 (16pp), 2019 September https://doi.org/10.3847/1538-3881/ab339a


# Untangling the Galaxy. I. Local Structure and Star Formation History of the Milky Way

Marina Kounkel and Kevin Covey
Department of Physics and Astronomy, Western Washington University, 516 High Street, Bellingham, WA 98225, USA; marina.kounkel@wwu.edu
*Received 2019 June 26; revised 2019 July 10; accepted 2019 July 11; published 2019 August 23*

## Abstract

*Gaia* DR2 provides unprecedented precision in measurements of the distance and kinematics of stars in the solar neighborhood. Through applying unsupervised machine learning on DR2's 5D data set (3D position + 2D velocity), we identify a number of clusters, associations, and comoving groups within 1 kpc and $|b| < 30^{\circ}$ (many of which have not been previously known). We estimate their ages with the precision of ~0.15 dex. Many of these groups appear to be filamentary or string-like, oriented in parallel to the Galactic plane, and some span hundreds of parsec in length. Most of these string lack a central cluster, indicating tha rather than the result of tidal stripping or dynamical processing. The your the Local Arm. The older ones appear to be remnants of several other a traced by dust and gas. The velocity dispersion measured from the ense age, suggesting a timescale for dynamical heating of ~300 Myr. This ti which the population of strings begins to decline, while the populatic increase, suggesting that dynamical processes are disrupting the weakl individual clusters to be identified at the oldest ages. These data shed a n a large-scale cloud collapse.

*Unified Astronomy Thesaurus concepts:* Milky Way dynamics (1051); (1608); Stellar associations (1582); Star clusters (1567); Stellar ages (1

*Supporting material:* interactive figures, machine-readable tables



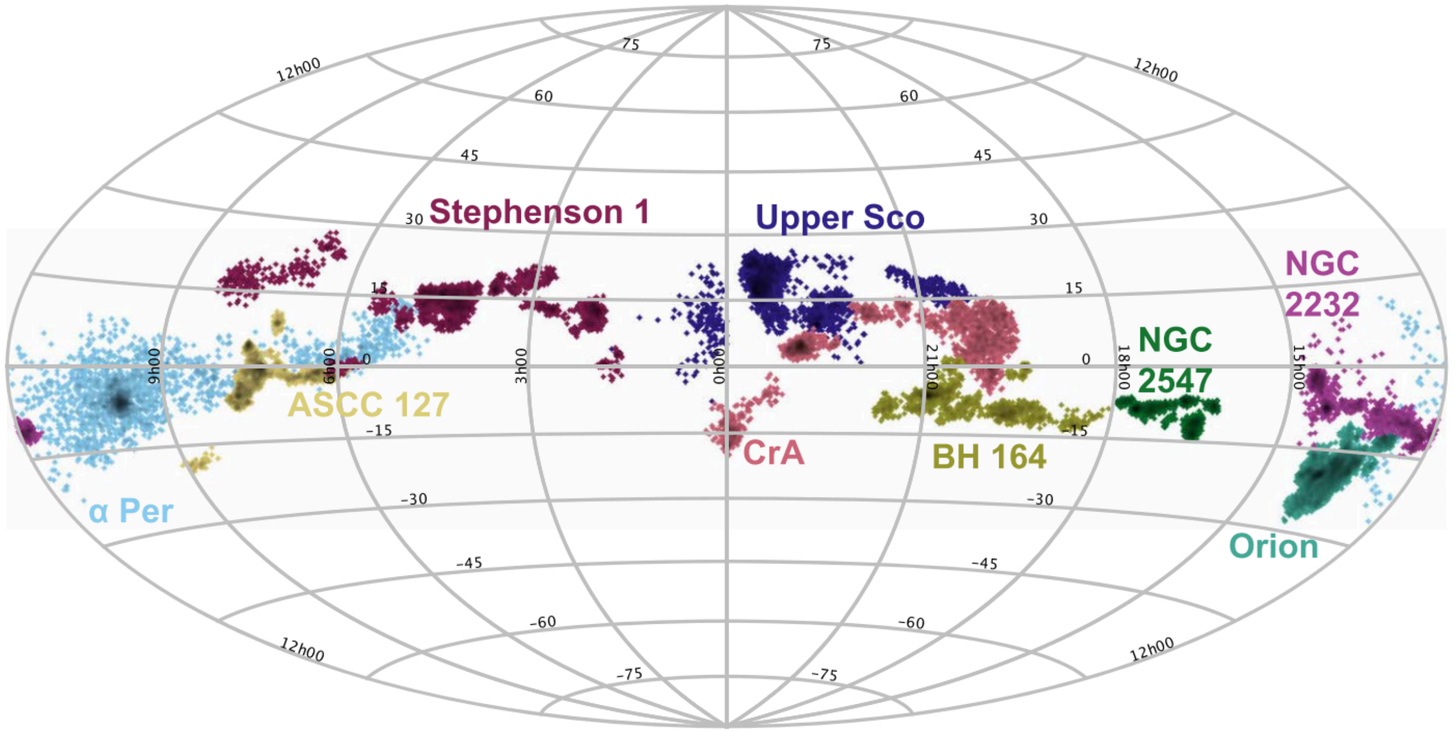



**Figure 10.** Some examples of notable structures, plotted in the galactic coordinates. See the interactive version of Figure 8 for full projections of all of the individual strings.

Kounkel & Covey (2019)

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **540,** 264–288 (2025) https://doi.org/10.1093/mnras/staf705
Advance Access publication 2025 May 1

# Nearby stellar substructures in the Galactic halo from DESI Milky Way Survey Year 1 Data Release

Bokyoung Kim [1★] Sergey E. Koposov [1,2] Ting S. Li [3] Sophia Lilleengen [4] Andrew P. Cooper [5,6] Andreia Carrillo [4,7] Monica Valluri [8] Alexander H. Riley [4] Jiwon Jesse Han [9] Jessica Nicole Aguilar,[10] Steven Ahlen [11] Leandro Beraldo e Silva [8,12,13] Davide Bianchi [14] David Brooks,[15] Amanda Byström [1] Todd Claybaugh,[10] Shaun Cole [4] Kyle Dawson,[16] Axel de la Macorra [17] Jaime Forero-Romero [18,19] Oleg Y. Gnedin [8] Satya Gontcho A Gontcho [10] Gaston Gutierrez,[20] Julien Guy [10] Klaus Honscheid,[21,22] Robert Kehoe,[23] Namitha Kizhuprakkat,[5,6] Martin Landriau [10] Laurent Le Guillou [24] Michael Levi [10] Gustavo E. Medina [3] Aaron Meisner [25] Ramon Miquel,[26,27] John Moustakas [28] Claire Poppett,[10,29] Francisco Prada [30] Graziano Rossi,[31] Eusebio Sánchez [32] Michael Schubnell,[33] Ray Sharples [4,34] David Sprayberry,[25] José Arturo Trelles Hernández,[17] Benjamin Alan Weaver[25] and Hu Zou [35]

*Affiliations are listed at the end of the paper*

Accepted 2025 April 28. Received 2025 April 8; in original form 2024 October 2

**ABSTRACT**

We report five nearby ($d_{\rm helio} < 5$ kpc) stellar substructures in the Galactic halo from a subset of 138 661 stars in the Dark Energy Spectroscopic Instrument (DESI) Milky Way Survey Year 1 Data Release. With an unsupervised clustering algorithm, HDBSCAN*, these substructures are independently identified in Integrals of Motion ($E_{\rm tot}$, $L_z$, $\log J_r$, $\log J_z$) space and Galactocentric cylindrical velocity space ($V_R$, $V_\phi$, $V_z$). We associate all identified clusters with known nearby substructures (Helmi streams, M18-Cand10/MMH-1, Sequoia, Antaeus, and ED-2) previously reported in various studies. With metallicities precisely measured by DESI, we confirm that the Helmi streams, M18-Cand10, and ED-2 are chemically distinct from local halo stars. We have characterized the chemodynamic properties of each dynamic group, including their metallicity dispersions, to associate them with their progenitor types (globular cluster or dwarf galaxy). Our approach for searching substructures with HDBSCAN* reliably detects real substructures in the Galactic halo, suggesting that applying the same method can lead to the discovery of new substructures in future DESI data. With more stars from future DESI data releases and improved astrometry from the upcoming *Gaia* Data Release 4, we will have a more detailed blueprint of the Galactic halo, offering a significant improvement in our understanding of the formation and evolutionary history of the Milky Way Galaxy.

**Key words:** surveys – stars: abundances – Galaxy: halo – Galaxy: kinematics and dynamics – (Galaxy:) solar neighbourhood.

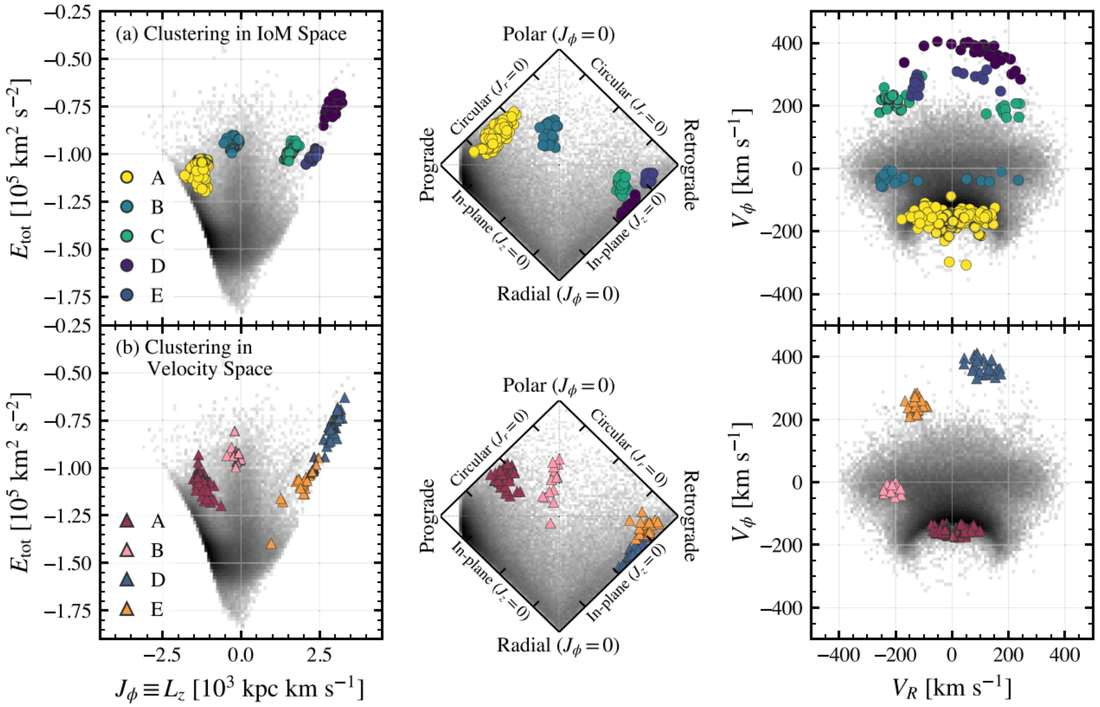



**Figure 6.** Distribution of the substructures identified by HDBSCAN* in $E_{\rm tot} - L_z$ space (left panels), the action diamond (middle panels), and $V_R - V_\phi$ space (right panels). All stars in the halo subset are shown in the grey log-scaled 2D histogram, and coloured points represent clustered groups identified in the corresponding kinematic space. In the middle panels, the horizontal axis corresponds to $J_\phi/J_{\rm tot}$, while the vertical axis indicates $(J_z - J_r)/J_{\rm tot}$, where $J_{\rm tot} = J_r + |J_\phi| + J_z$.

Kim et al. (2025)

The Astrophysical Journal, 976:161 (10pp), 2024 December 1 https://doi.org/10.3847/1538-4357/ad8352

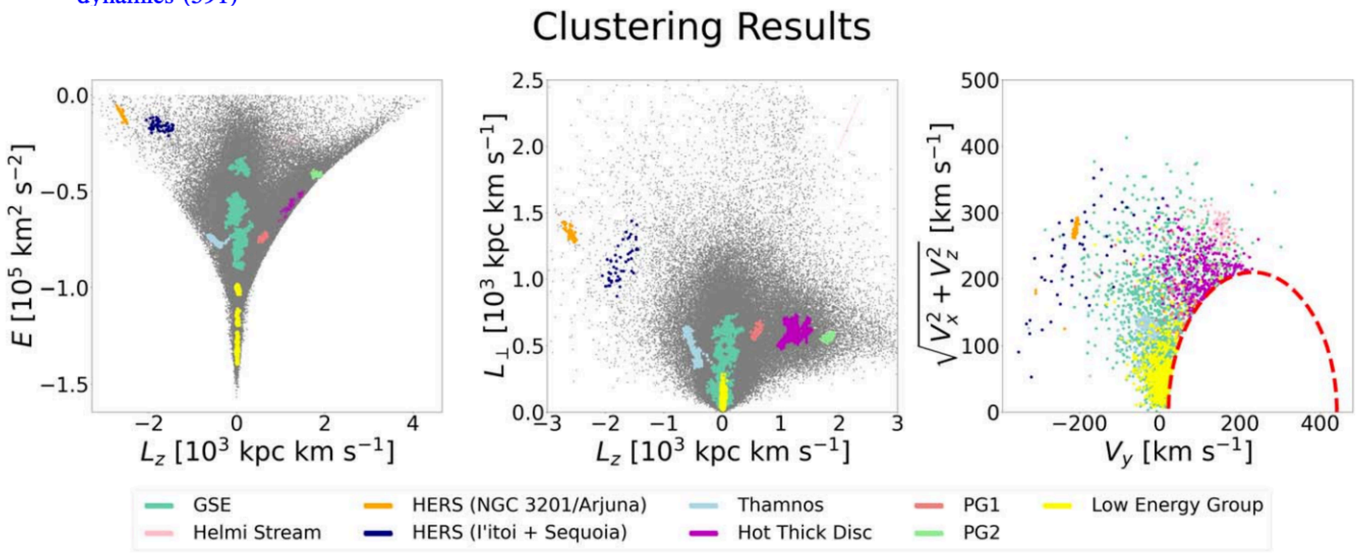

# Exploration of Halo Substructures in Integrals-of-motion Space with Gaia Data Release 3

Haoyang Liu, Cuihua Du, Dashuang Ye, Jian Zhang, and Mingji Deng
College of Astronomy and Space Sciences, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China; ducuihua@ucas.ac.cn
*Received 2024 August 23; revised 2024 September 16; accepted 2024 October 2; published 2024 November 20*

## Abstract

Using kinematic data from the Gaia Data Release 3 catalog, along with metallicity estimates robustly derived from Gaia BP/RP spectra, we have explored the Galactic stellar halo in search of both known and potentially new substructures. By applying the Hierarchical Density-Based Spatial Clustering of Applications with Noise clustering algorithm in integrals-of-motion space (i.e., $E$, $L_z$, and $L_\perp = \sqrt{L_x^2 + L_y^2}$), we identified five previously known substructures: Gaia-Sausage-Enceladus (GSE), Helmi streams, I'itoi and Sequoia, and the hot thick disk. We additionally found NGC 3201 and NGC 5139 in this work, and NGC 3201 shares similar distributions in phase space and metallicities to Arjuna, which possibly implies that they have the same origin. Three newly discovered substructures are Prograde Substructure 1 (PG1), Prograde Substructure 2 (PG2), and the Low Energy Group. PG1, with a higher $V_\phi$ than typical GSE member stars, is considered as either a low-eccentricity and metal-rich part of GSE or part of the metal-poor disk. PG2, sharing kinematic similarities with Aleph, is thought to be its relatively highly eccentric component or the mixture of Aleph and the disk. The Low Energy Group, whose metal-poor component of metallicity distribution function has a mean value [M/H] $\sim -1.29$ (compared to that of Heracles [M/H] $\sim -1.26$), may have associations with Heracles.

*Unified Astronomy Thesaurus concepts:* Galaxy stellar halos (598); Galaxy kinematics (602); Galaxy dynamics (591)




Liu et al. (2024)

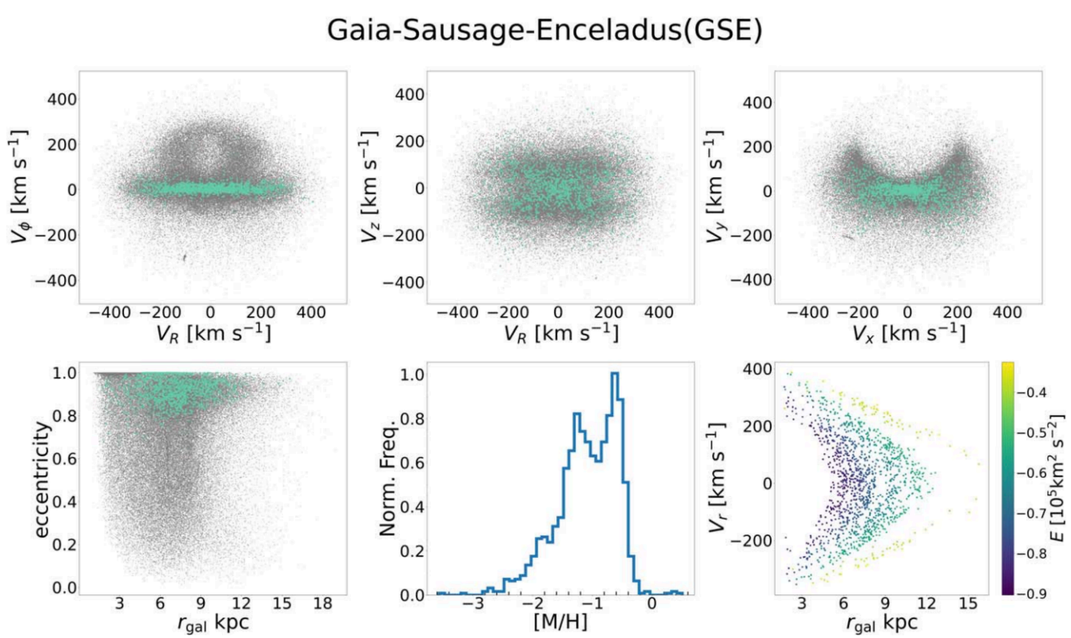



**Figure 2.** Top row: the velocity distributions in the $V_R$–$V_\phi$, $V_R$–$V_x$, and $V_x$–$V_y$ spaces, where the sausage shape is visible in the $V_R$–$v_\phi$ space. Bottom row: the eccentricity distribution along the galactocentric distance $r_{gal}$ and metallicity distribution of GSE with two noticeable peaks that peak around −1.25 and −0.7, respectively, where the latter peak is related to eccentric Splash stars. The chevron-like shape in the $V_r$–$r_{gal}$ space is shown and color-coded by energy.

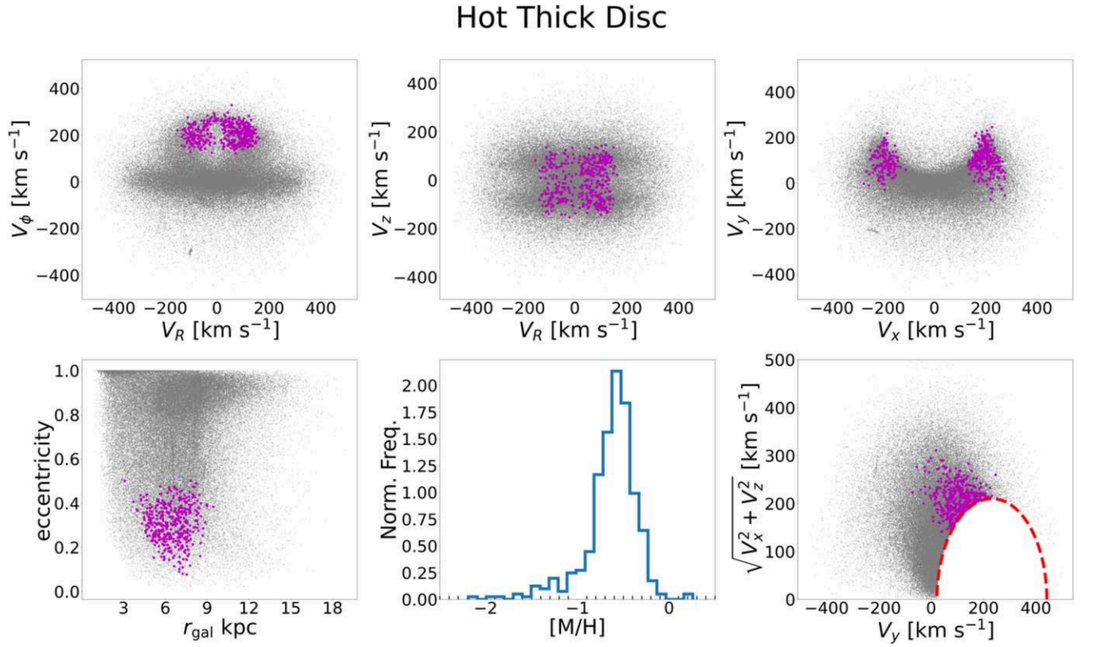



**Figure 6.** Top row: the velocity distributions in the $V_R$–$V_\phi$, $V_R$–$V_x$, and $V_x$–$V_y$ spaces. Bottom row: the eccentricity distributions along the galactocentric distance $r_{gal}$, the metallicity distributions of the hot thick disk, and the Toomre diagram. The hot thick disk is situated at the ends of the banana-shaped arc in the the $V_X$–$V_y$ plane. The close attachment to the curve where $|V - V_{LSR}| = 210$ km s$^{-1}$ is also visible in the bottom right panel.

# Final Remarks

# Key Takeaways
## about clustering

- Clustering is **cool** and **useful** in many contexts.
- These and many other clustering algorithms exist.
- Semi-supervised learning has become the most popular way to use clustering in recent years.
- The most critical step is not the algorithm itself, but the choice and scaling of **features**. Most definitely the first time you use clustering, your results will be ~~BAD~~. **Your brain** needs to have physical intuition: understanding which properties best distinguish different objects is key to obtaining meaningful results.

# THANK YOU!