Star Clusters using *Gaia*

Sagar Malhotra, Alfred Castro-Ginard
(sagar@fqa.ub.edu)

Open Clusters (OCs): An Introduction

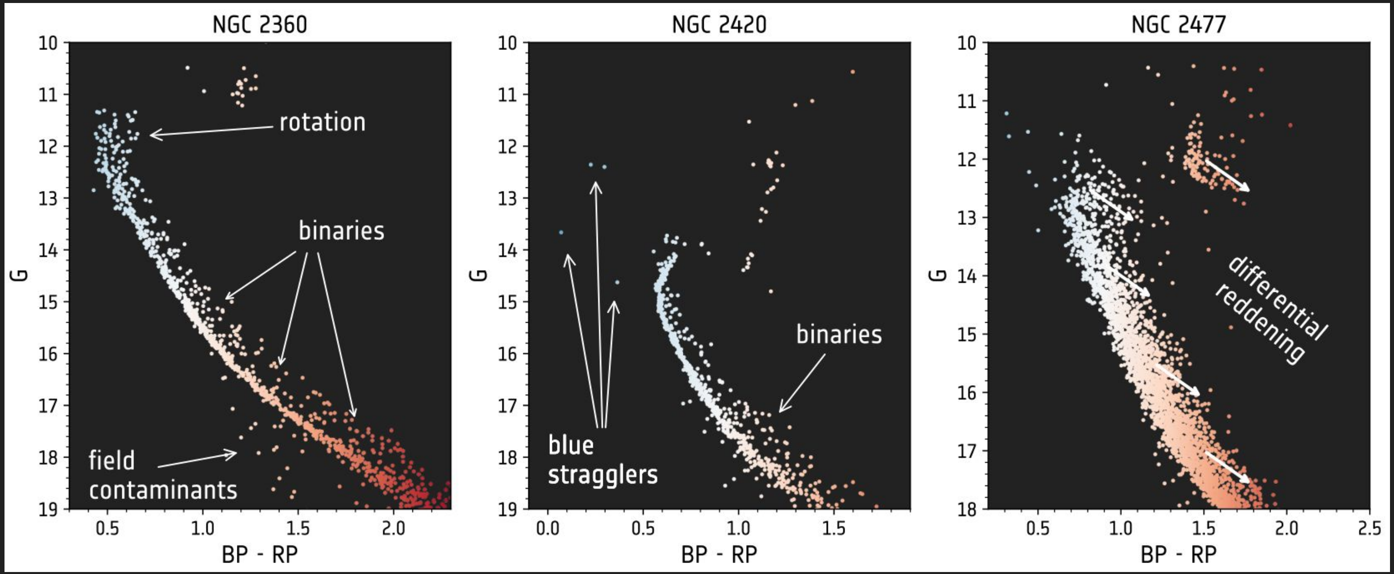
- Gravitationally bound group of coeval stars born from the same parent molecular cloud
- Most stars are believed to be born in stellar clusters/associations before dissolving into field star population
- Observations:
  - similar 3D kinematics
  - lie on a single isochrone on a CMD; similar chemical composition
- Advantages:
  - Precise distances and ages
  - tracers of the MW disk; young OCs can be used as tracers of spiral arms

Fried Lauterbach - Own work

Why *Gaia* Data?

- Exquisite photometry and astrometry: ~ 0.01 mas yr^{-1} for bright and well-behaved sources

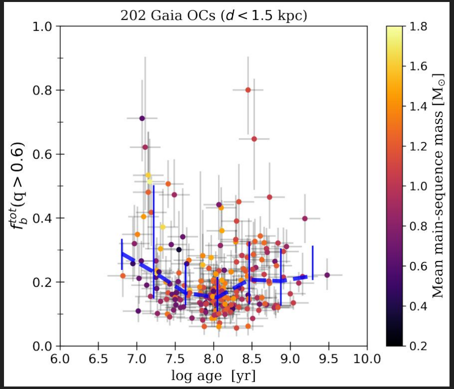
- Homogeneous parameters for a large number of sources: ~1.8 billion sources in *Gaia* DR3; over 1.4 billion sources with full astrometry



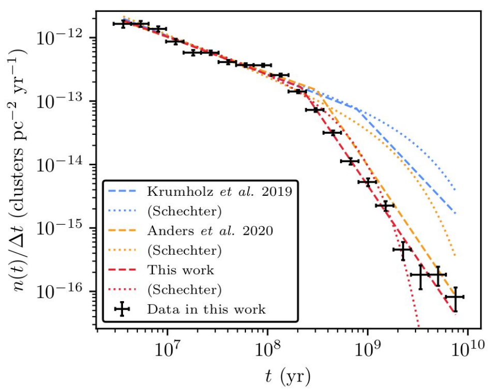
Credit:
Cantat-Gaudin & Casamiquela 2024

OCs using *Gaia*: A Paradigm Shift

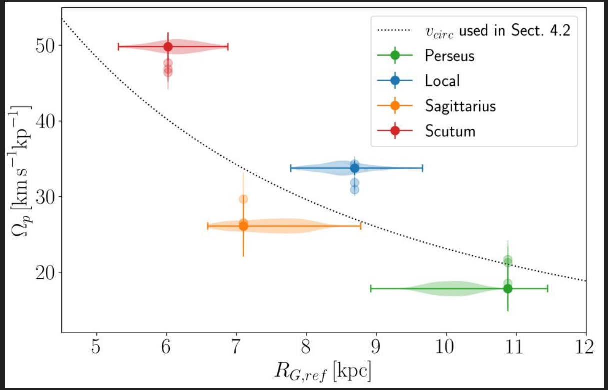
- Cluster Age Function
- Probing unresolved binaries
- Galactic Metallicity Gradient
- Tracing Milky Way Disk and Spiral Arms



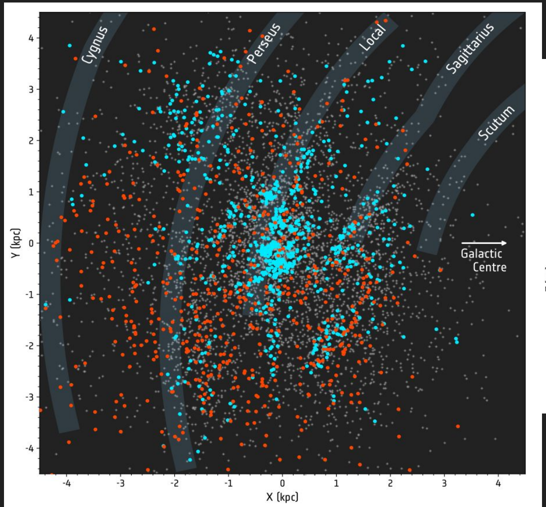
Donada et al. 2023



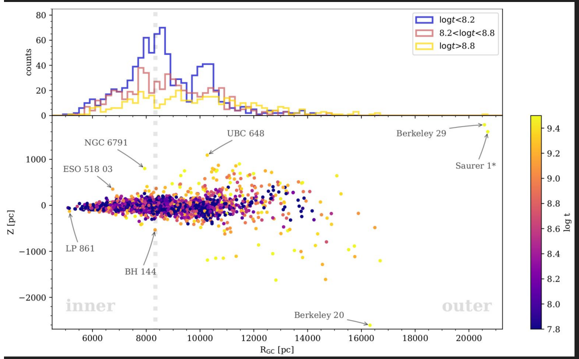
Hunt & Reffert 2024



Castro-Ginard et al. 2021



Cantat-Gaudin & Casamiquela 2024



Cantat-Gaudin et al. 2020

Detection of OCs

- OCs are expected to be "compact" objects in the position and velocity space
- Usually, we study clusters in Gaia data by detecting overdensities in the 5D astrometric parameter space (α, δ, ϖ, $\mu_{\alpha*}$, μ_{δ}) i.e. on-sky positions and proper motions
- Commonly used clustering algorithms:

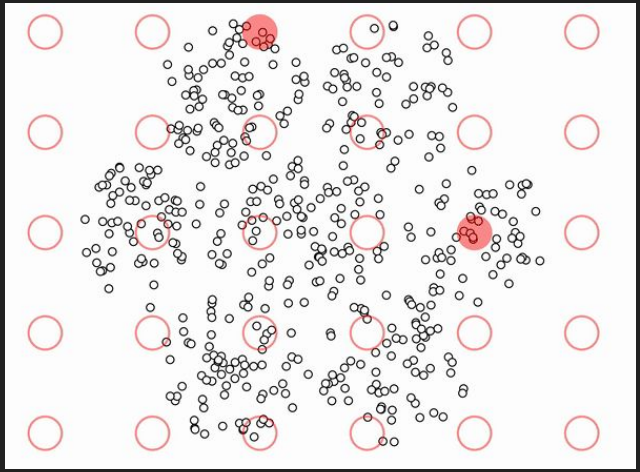
  (should work across a wide range of stellar densities and cluster sizes)

  - Unsupervised Photometric Membership Assignment in Stellar Clusters (UPMASK)
  - Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - Hierarchical-DBSCAN (HDBSCAN)
  - Ordering Points To Identify the Clustering Structure (OPTICS)
  - Gaussian Mixture Models
  - Visual Inspection
  - ….

Detection of OCs in *Gaia* Data: DBSCAN

DBSCAN uses distance between points as a proxy for the local density of an area

- Two main parameters: ϵ and m_{Pts}
  - *Core points*: if they are within the distance ϵ to at least m_{Pts}
  - *Members*: not a core point but within the distance ϵ from a core point
  - *Field (Noise)*: all other points
- m_{Pts} can be set to a fixed number (Ester et al. (1996)) or can be optimized based on the average density of stars in a field
- Can find arbitrary shaped clusters

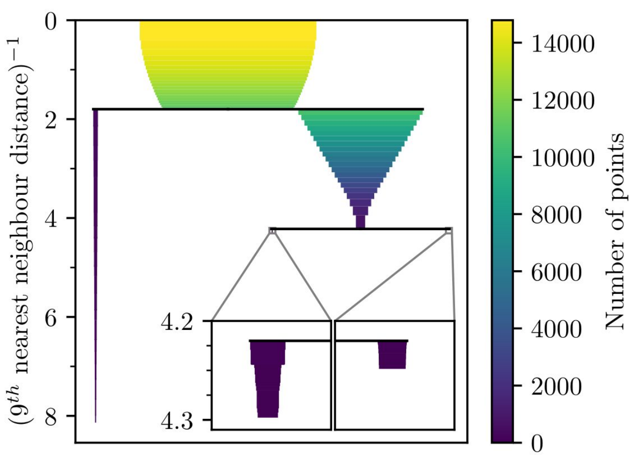


Credit: Naftali Harris (Link)

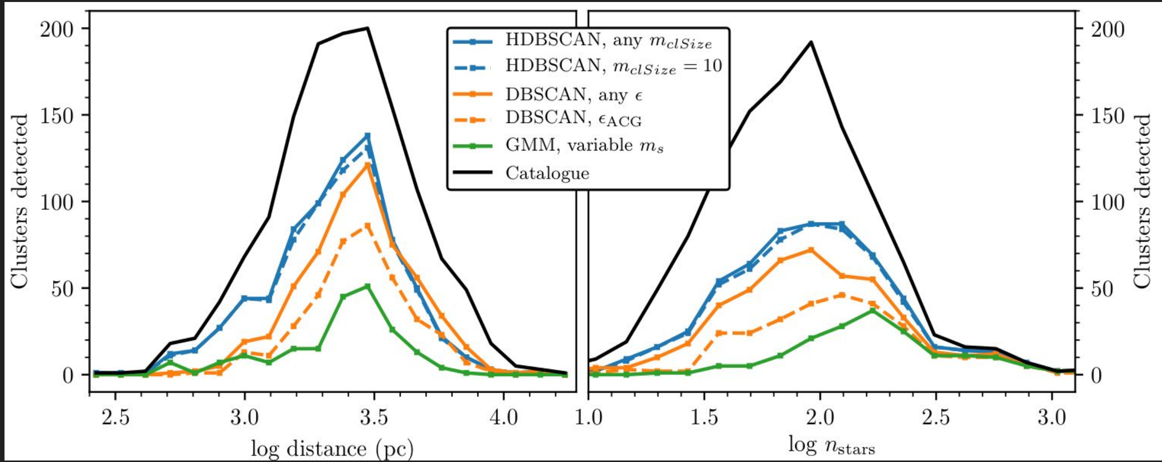
HDBSCAN: The most effective algorithm*

- ϵ replaced by m_{ClSize}

*lots of false positives



Hunt & Reffert 2021 Fig. 3



Hunt & Reffert 2021 Fig. 5

Algorithm	Reported OC candidates [a]	Fraction with CST > 3σ	Total crossmatches [b]	Mean runtime (mins) [c]
DBSCAN (ACG)	1518–1538	58.9%–59.6%	382	1.19 (1 repeat)–10.3 (30 repeats)
DBSCAN (model)	5212–51920	22.4%–2.1%	593	0.885
HDSBCAN	1196–49693	82.0%–5.2%	756	2.36
GMM	314–2465	60.5%–20.5%	213	21.9 (m_s = 800) 47.0 (variable m_s)

Hunt & Reffert 2021 Tab. 4

Logistics

Runtime

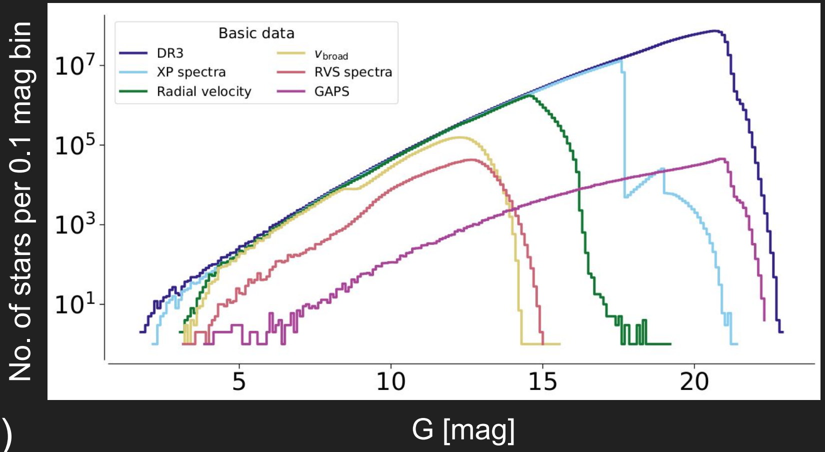
DBSCAN [Castro-Ginard et al. 2022]

(PyCOMPSs + *dislib* library [Tejedor et al. 2015, Álvarez Cid-Fuentes et al. 2019])

- Used total of 144 cores (3 nodes) of MareNostrum
- Each application of DBSCAN for the whole Galactic disk ranging from 12 to 27h depending on (L, m_{Pts}) pair [**G_{thresh} <= 18**]

HDBSCAN [Hunt & Reffert 2023]

- Parameters used (m_{ClSize} ϵ {10, 20, 40, 80}, m_{Pts} = 10)
- 8 days of runtime on a machine with a 48 core Intel(R) Xeon(R) E5-2650 CPU with 48 GB of RAM
- RAM-limited due to memory usage
- **G_{thresh} <= 21**



Gaia Collaboration, Vallenari et al. 2022

Expectations from *Gaia* DR4

(66 months of data)

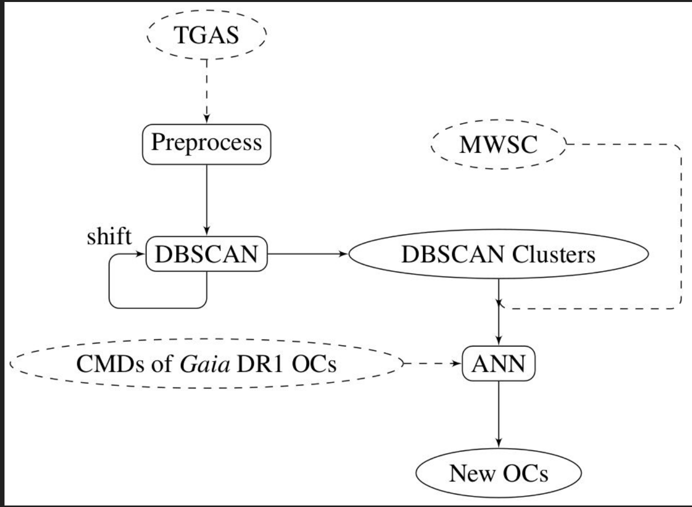
- Better precision of the photometry and astrometry; better characterization and detection of OCs, particularly at larger distances
- Epoch data of photometry and astrometry for each source, which will significantly increase the size of the data
- Number of Sources: similar to *Gaia* DR3 with a few new detections

Road Ahead

- Improvement in cluster detection:
  - Better membership probabilities
  - Including tidal tails
  - Dealing with time series data
- Open Cluster Selection Function

Summary

- ~ 730M sources in *Gaia* DR3 selected for detecting OCs
- No more than 20M sources in one field (file size ~ 3GB) for applying HDBSCAN (one field: 1 HEALPix level 5 pixel + stars from neighboring 8 pixels)
- How SPACIOUS helps?
    - Availability of the whole Gaia dataset avoids the need to launch thousands of queries to the Gaia Archive
    - Flexibility of allocating memory, number of executors
    - Parallelisation across different sections of the sky wherein iterative tasks such as applying HDBSCAN can be easily scaled over large data



Castro-Ginard et al. 2018 Fig. 1

Let's try an example with HDBSCAN ====> link to the example usage notebook

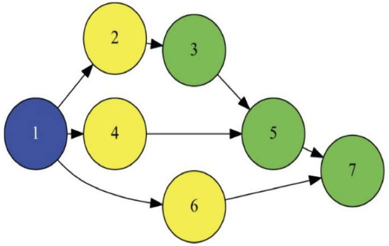
Backup Slides

dislib | distributed computing library

dislib github repo

Method: DBSCAN in parallel

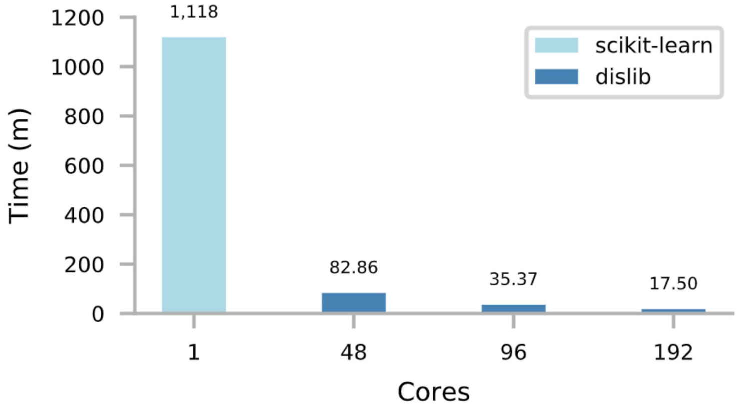
Use PyCOMPSs framework + *dislib* library [Tejedor+15,Álvarez Cid-Fuentes…**ACG**+19]



- Exploit parallelism of applications at task level
- Task — decorated python function
- Builds a task graph taking into account data dependencies
- Schedule and execute application in the distributed environment based of the graph

x

Credit: A. Castro-Ginard



Álvarez Cid-Fuentes et al. 2019