MACHINE LEARNING AND GAIA DR2, ON THE HUNT FOR OPEN STAR CLUSTERS

Alfred Castro-Ginard

Gaia RIA — February 18th

Universitat de Barcelona

ICCUB

IEEC
INSTITUT D'ESTUDIS
ESPACIALS
DE CATALUNYA

PRE-GAIA VIEW OF THE OC POPULATION

- Census counted with around 3000 catalogued objects compiled from heterogeneous data sources [Dias+02][Kharchenko+13][Röser+16]
- Estimated number of OCs ~10^5 [Binney&Tremaine 2008]
- Thought to be complete up to 1.8 kpc

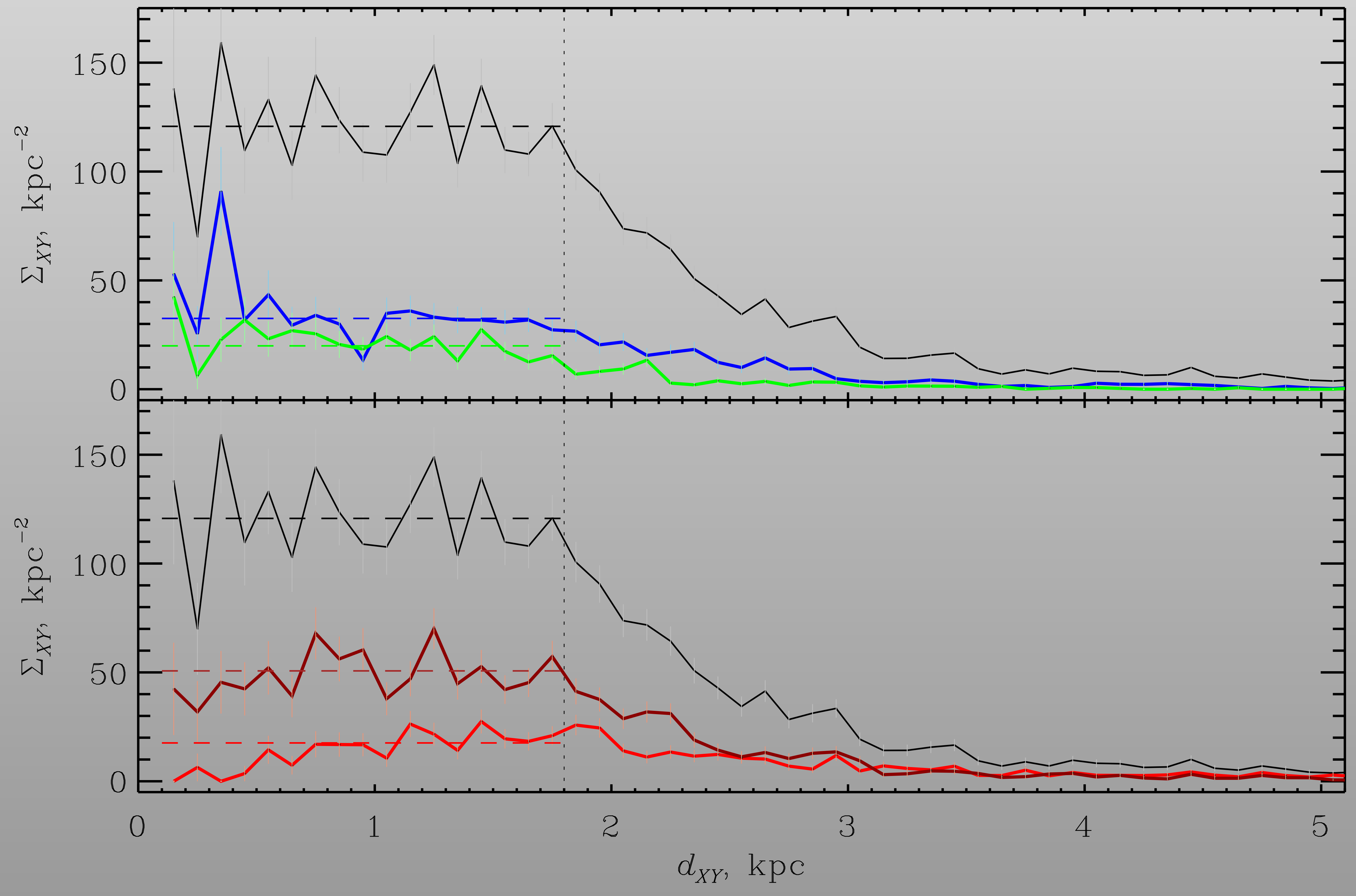


Fig. 4 from Kharchenko+13.

Space density of stellar clusters as a function of distance.

Adopted a completeness limit at 1.8 kpc

After Gaia DR2

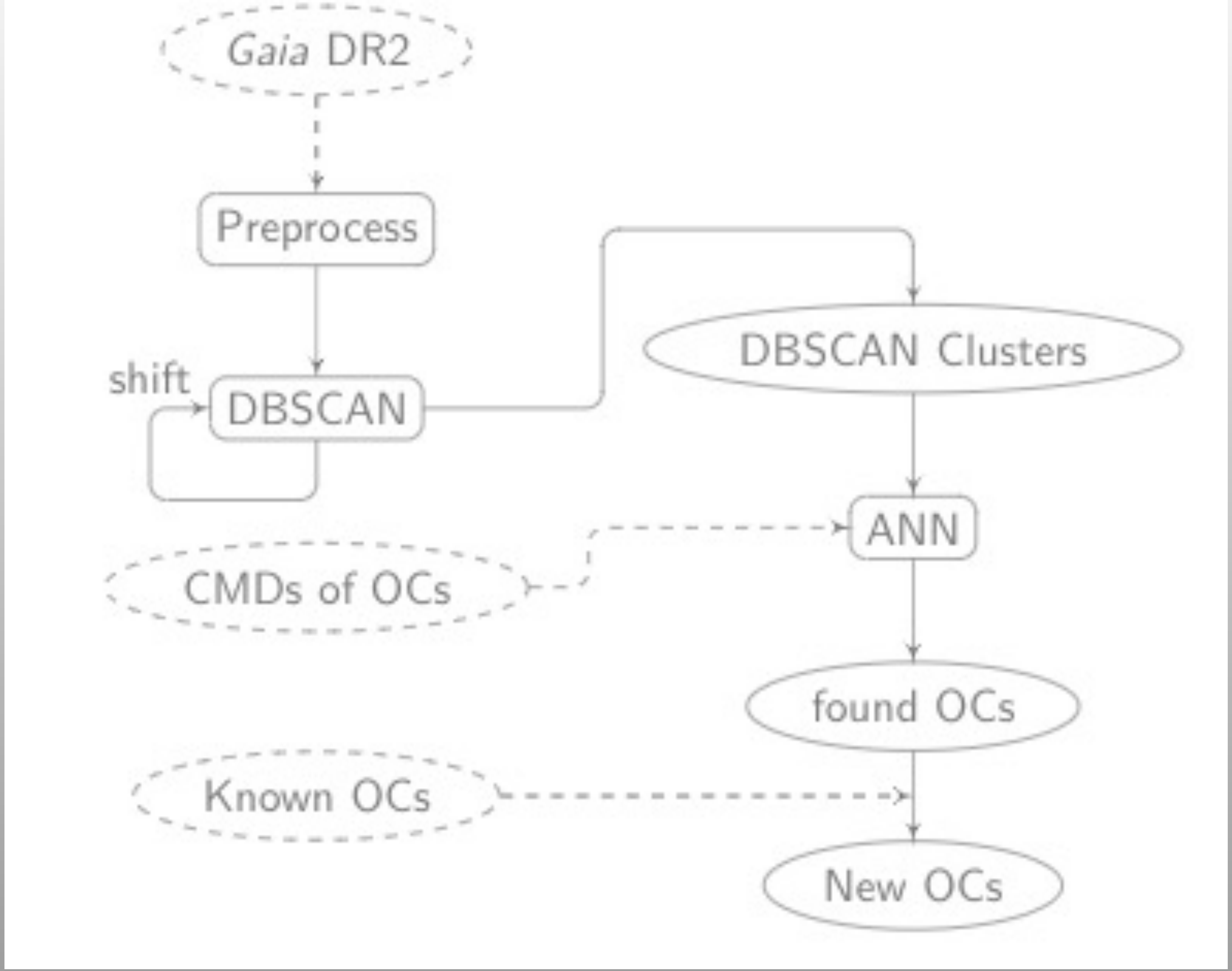
- Trying to characterise the catalogued OCs with Gaia DR2, only 1169 objects were found and 60 new OCs were serendipitously detected [Cantat-Gaudin…ACG+18]
- The remaining clusters were either discarded or not seen by Gaia (too distant, IR…)

- Dedicated studies to search for unknown OCs:
  - → [Castro-Ginard+18]: 23 new objects found with TGAS (validated with Gaia DR2), most of them located in the disc within 1kpc
  - [Cantat-Gaudin…ACG+19]: 41 new objects in the Perseus direction
  - → [Castro-Ginard+19]: 53 OCs found with Gaia DR2 in the Galactic anti-centre
  - [Sim+19]: 207 OCs by visually inspecting proper motion diagrams
  - [Liu&Pang19]: 76 high quality OCs, FoF algorithm on 5-D astrometry
  - → [Castro-Ginard+20]: 582 new OCs in the Galactic disc (Big Data)

METHOD

- Data mining methodology to automatically search OCs in the Gaia DR2 archive
- Method based on two combined machine learning algorithms:
  - Unsupervised learning: detection of over-densities in the 5-dimensional astrometric space (position, parallax and proper motions)
    - DBSCAN: density based space clustering
  - Supervised learning: classification of over-densities into real OCs or random statistical clusters
    - ANN: to detect isochrone patterns on a CMD

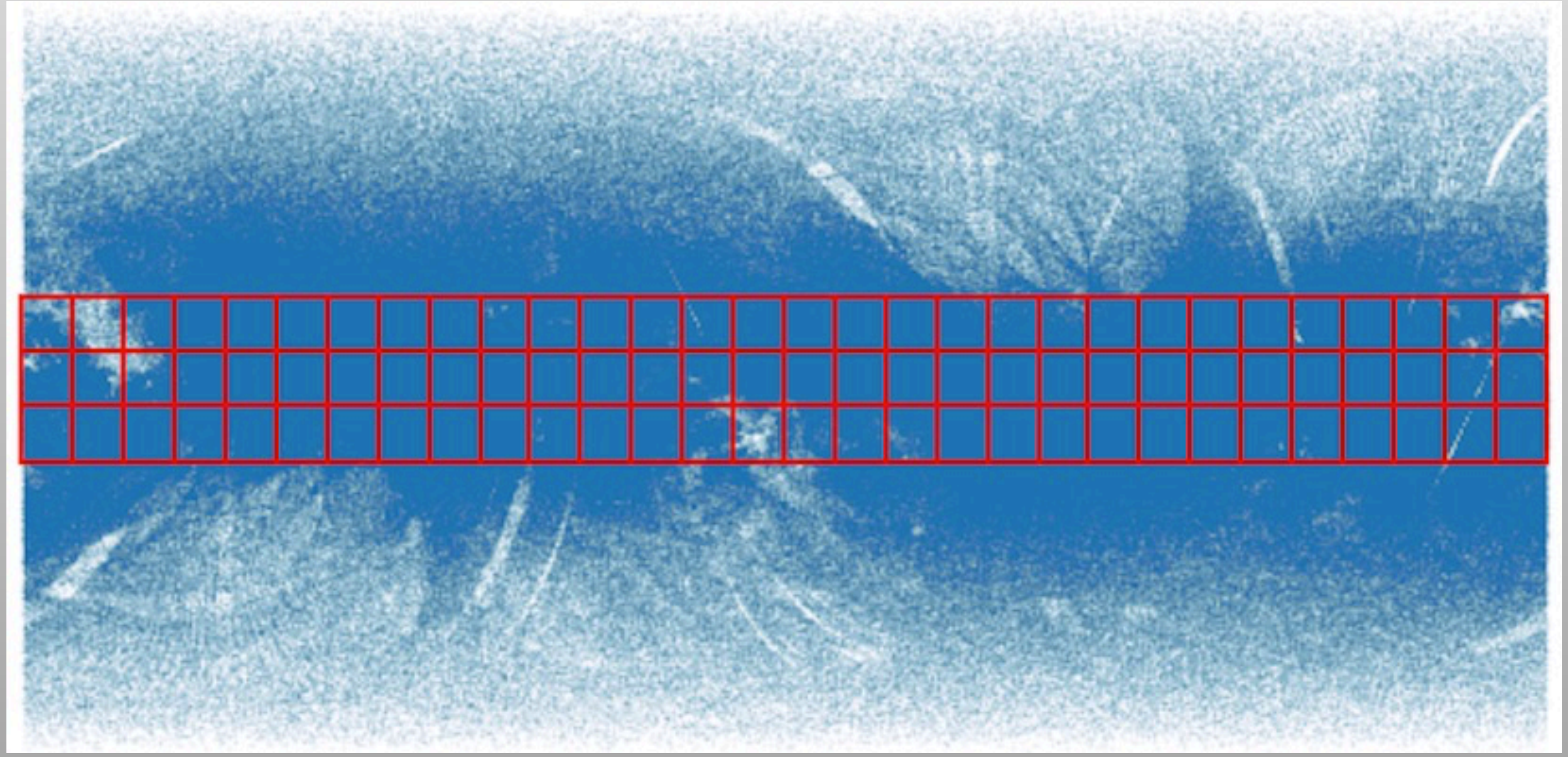
Flow Chart



Gaia DR2
Preprocess
shift
DBSCAN
DBSCAN Clusters
ANN
CMDs of OCs
found OCs
Known OCs
New OCs

PREPROCESSING

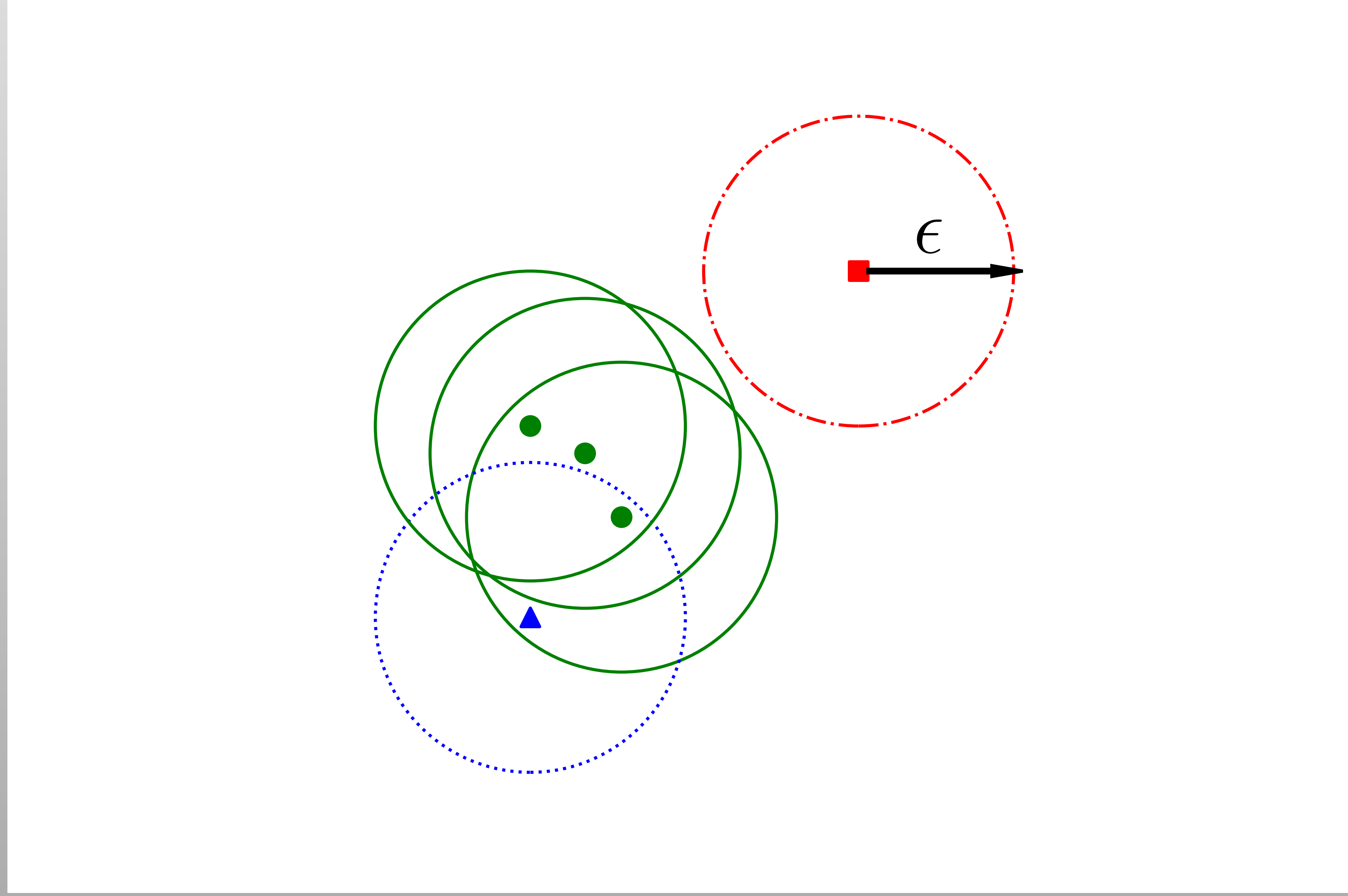
- Rejection of stars with high proper motion (>30 mas/yr) or high parallax (> 7 mas)
- Divide the area of study in rectangles of size L deg (to be optimised using simulated data). Consider only |b| < 10º (~95% of the known clusters are in this region)
- Shift rectangles to account for the clusters in the border



DBSCAN

Use a density based unsupervised algorithm to search for over-densities in the parameter space [Ester+96]

- No a priori knowledge of the number of clusters
- Finds arbitrary shaped clusters
- Need to define two parameters (e, minPts)



$$d(i,j) = \sqrt{(l_i - l_j)^2 + (b_i - b_j)^2 + (\varpi_i - \varpi_j)^2 + (\mu_{\alpha^*,i} - \mu_{\alpha^*,j})^2 + (\mu_{\delta,i} - \mu_{\delta,j})^2}$$

DBSCAN — DETERMINATION OF EPSILON

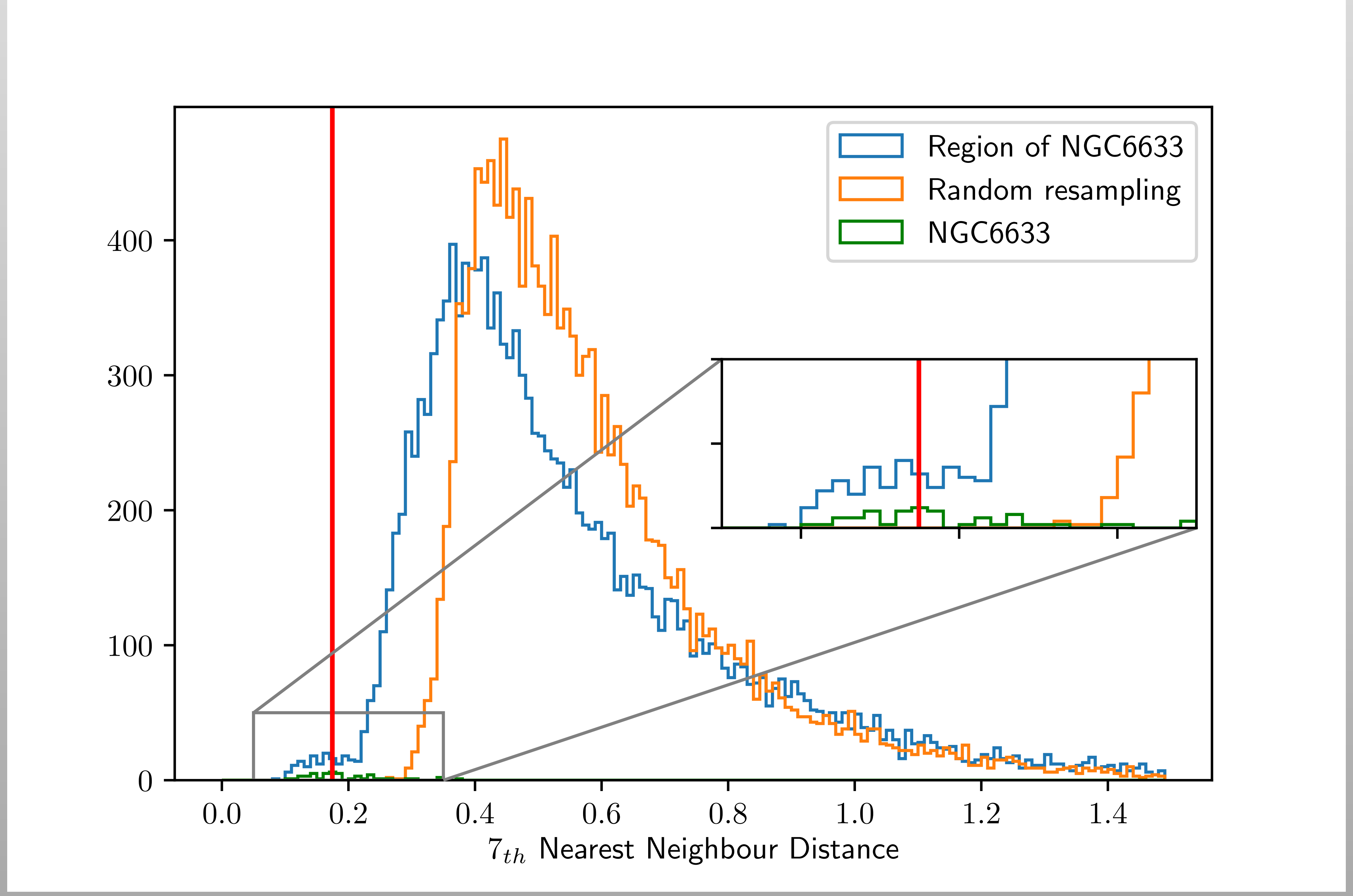
Leave minPts to be optimised using simulated data (together with L)

For the determination of e:

- Distance between the kth nearest neighbours in a cluster should be smaller than the distance between stars belonging to the field

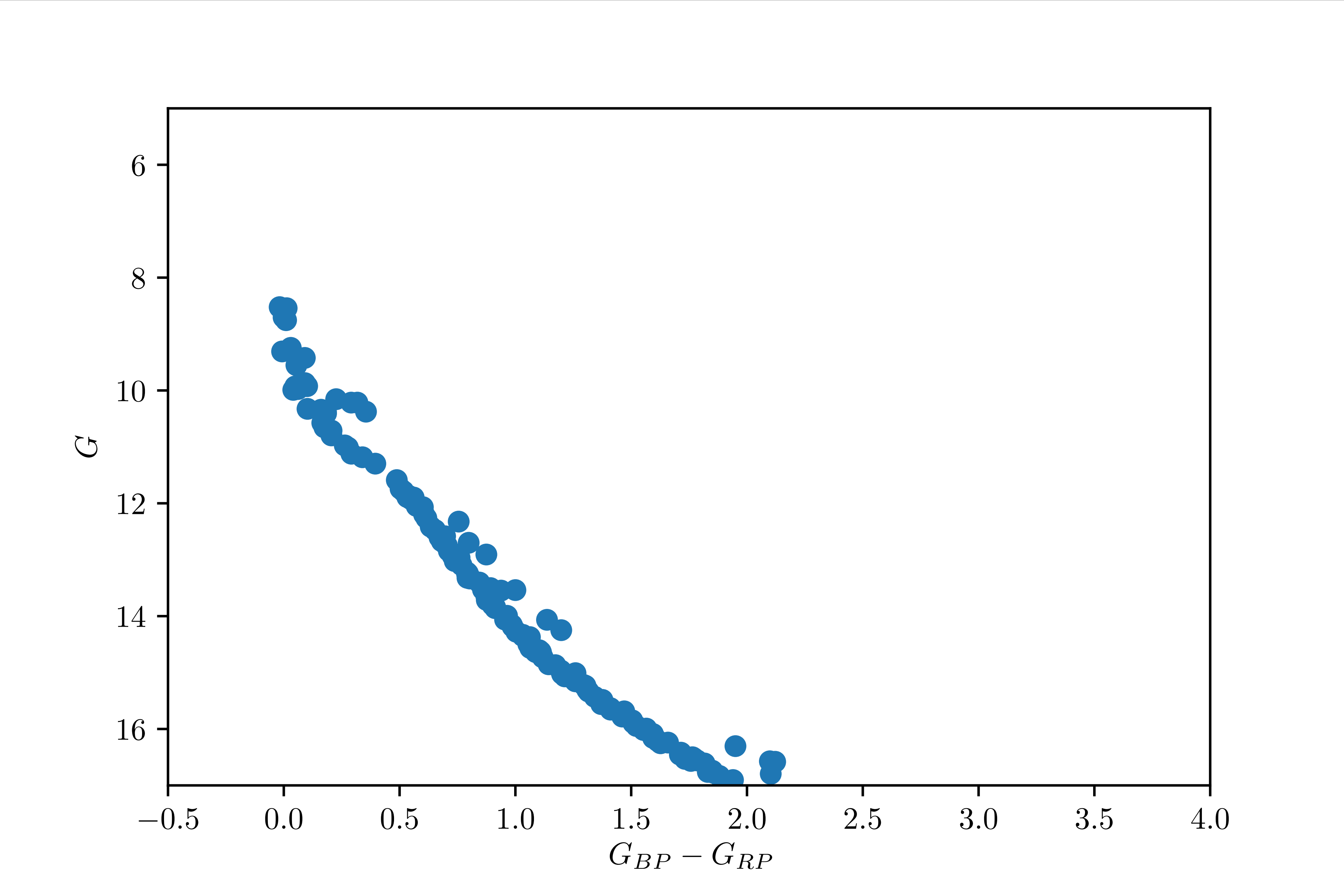
- Compute e as:

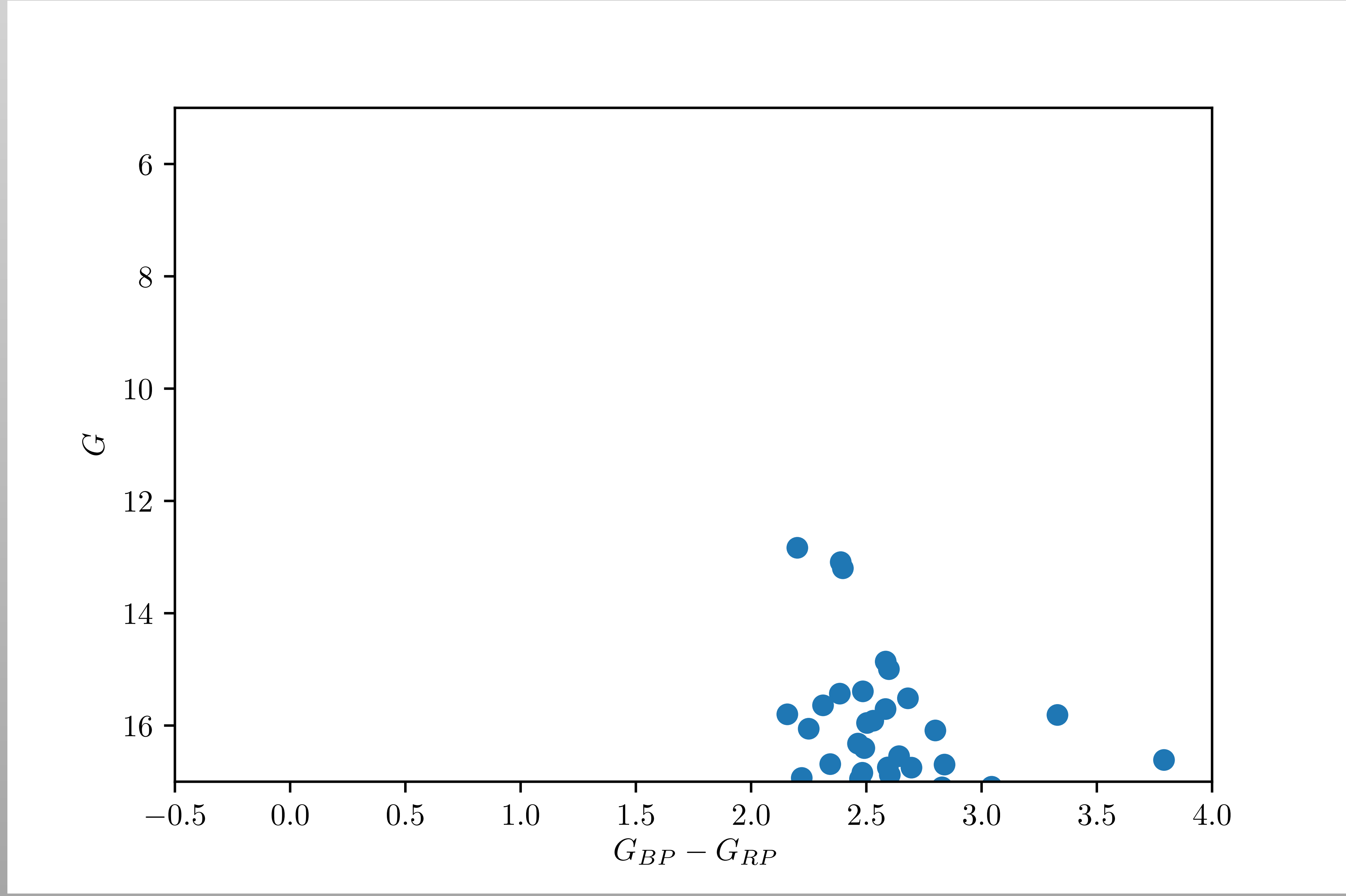
$$\epsilon = (\epsilon_{kNN} + \epsilon_{rand})/2$$



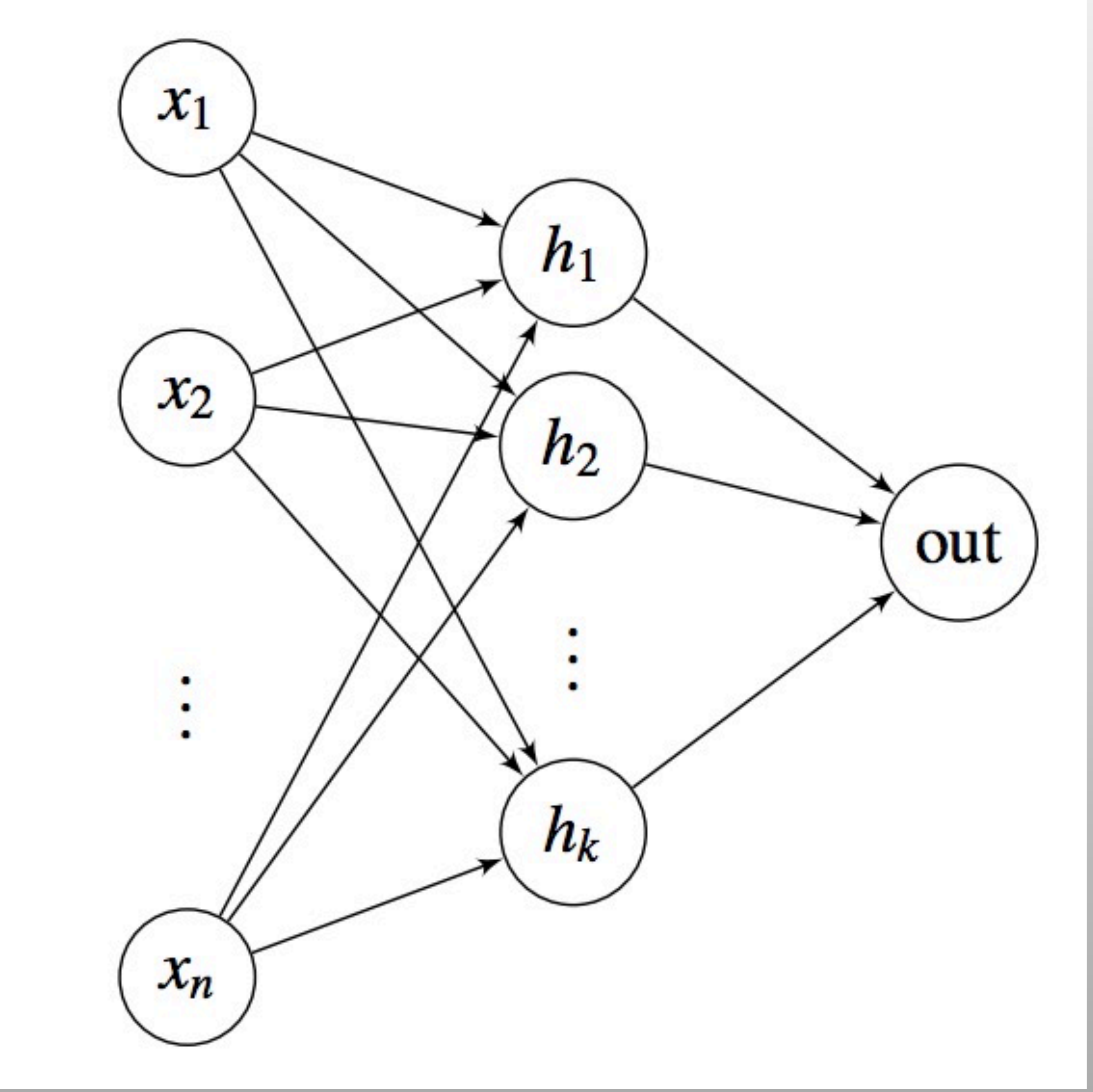
Real OCs vs. Statistical Clusters

- DBSCAN finds statistical clusters that may or may not correspond to physical OCs
- Distinguish between them using Gaia photometry: stars in an OC follow an isochrone in a CMD





ANN — Introducing physical meaning



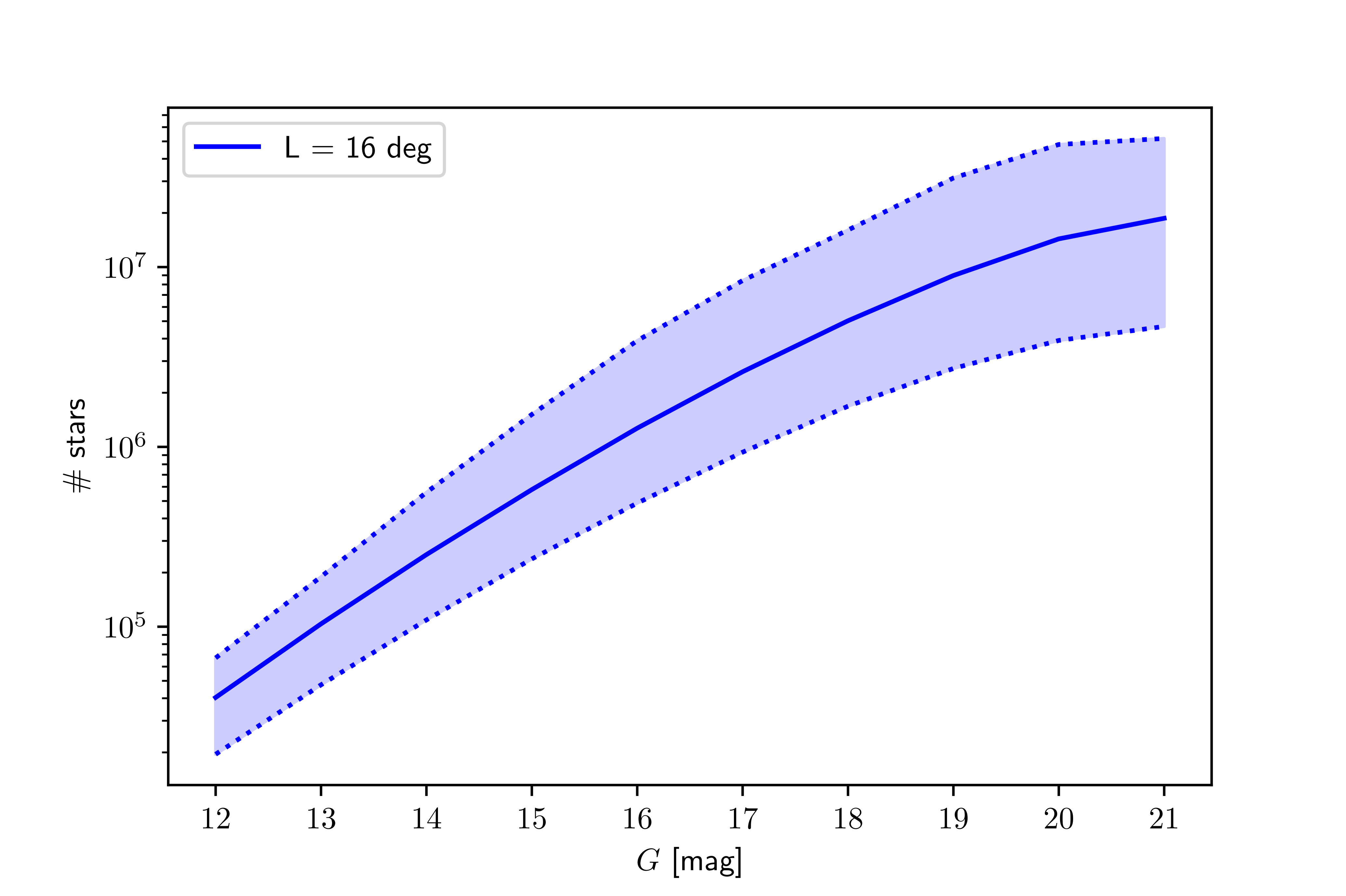
- Need labeled data to train the network
- Train on CMD to recognise the isochrone pattern
  - For the 1229 OCs in [Cantat-Gaudin… ACG+18] plus data augmentation
  - Random field stars selected from the same region

Training - test splits set to 67%-33%
Classify 90.27% of the cases to the right class

So far, no Big Data involved

DBSCAN IN PARALLEL

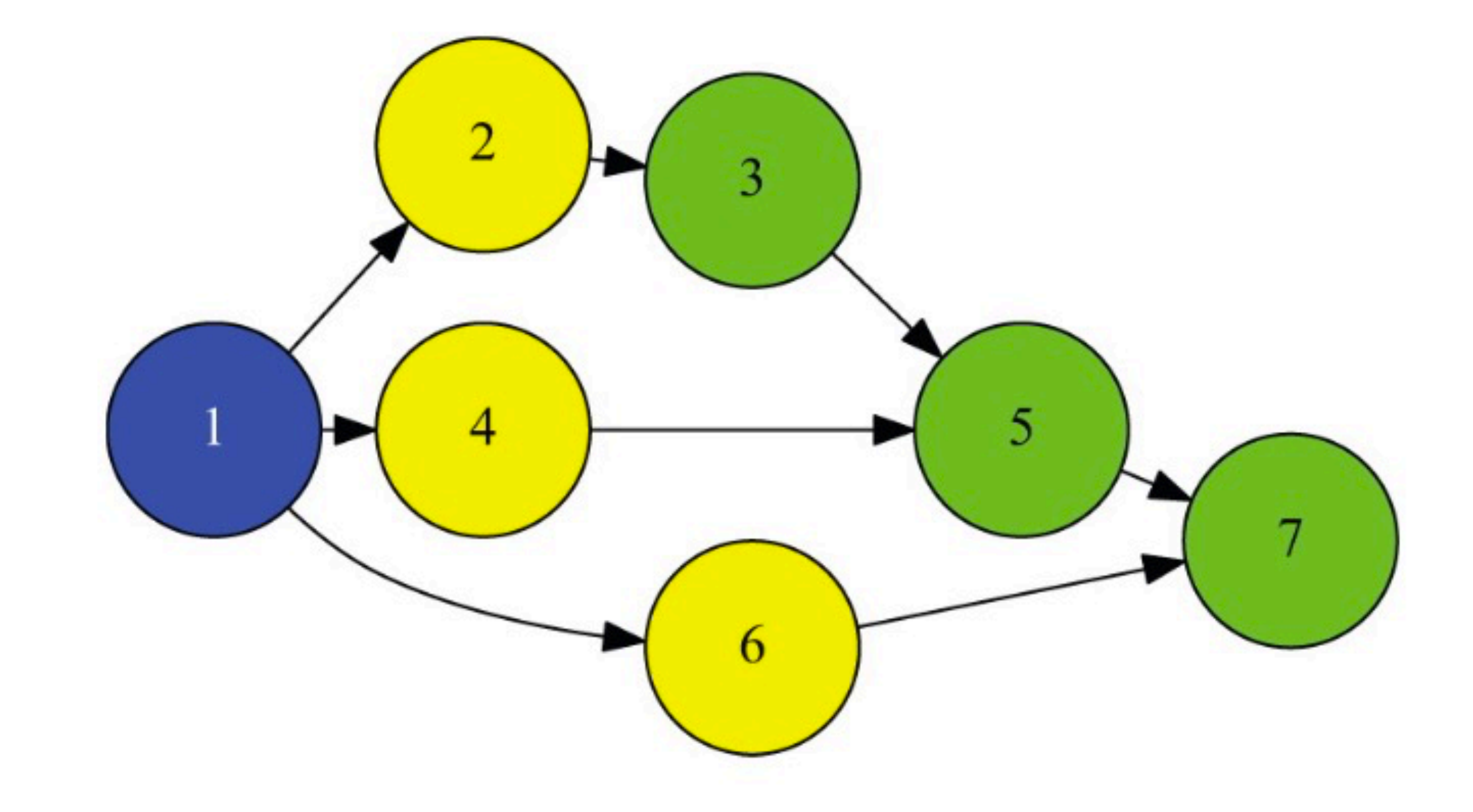
Run DBSCAN to the whole Galactic disc up to magnitude G = 17 (10^8 stars) [Castro-Ginard+20]



- DBSCAN on $\sim 10^7$ stars in each box (large enough data)
- Two level parallelization
  - Computation of each box in parallel
  - Parallelization of the DBSCAN algorithm if needed

DBSCAN IN PARALLEL

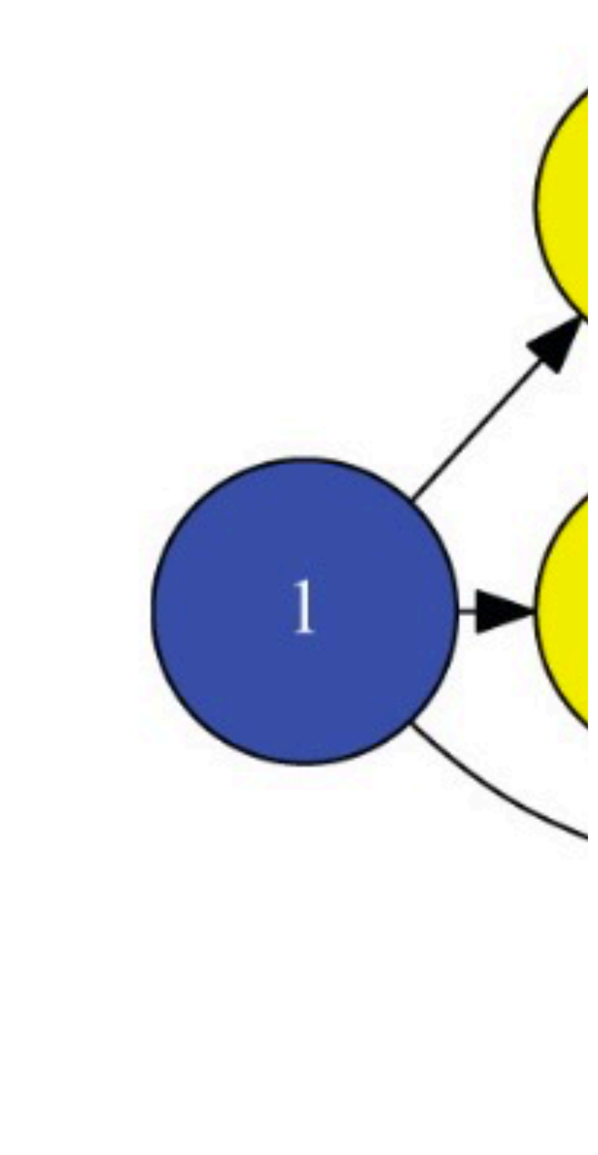
Use PyCOMPSs framework [Tejedor+15]

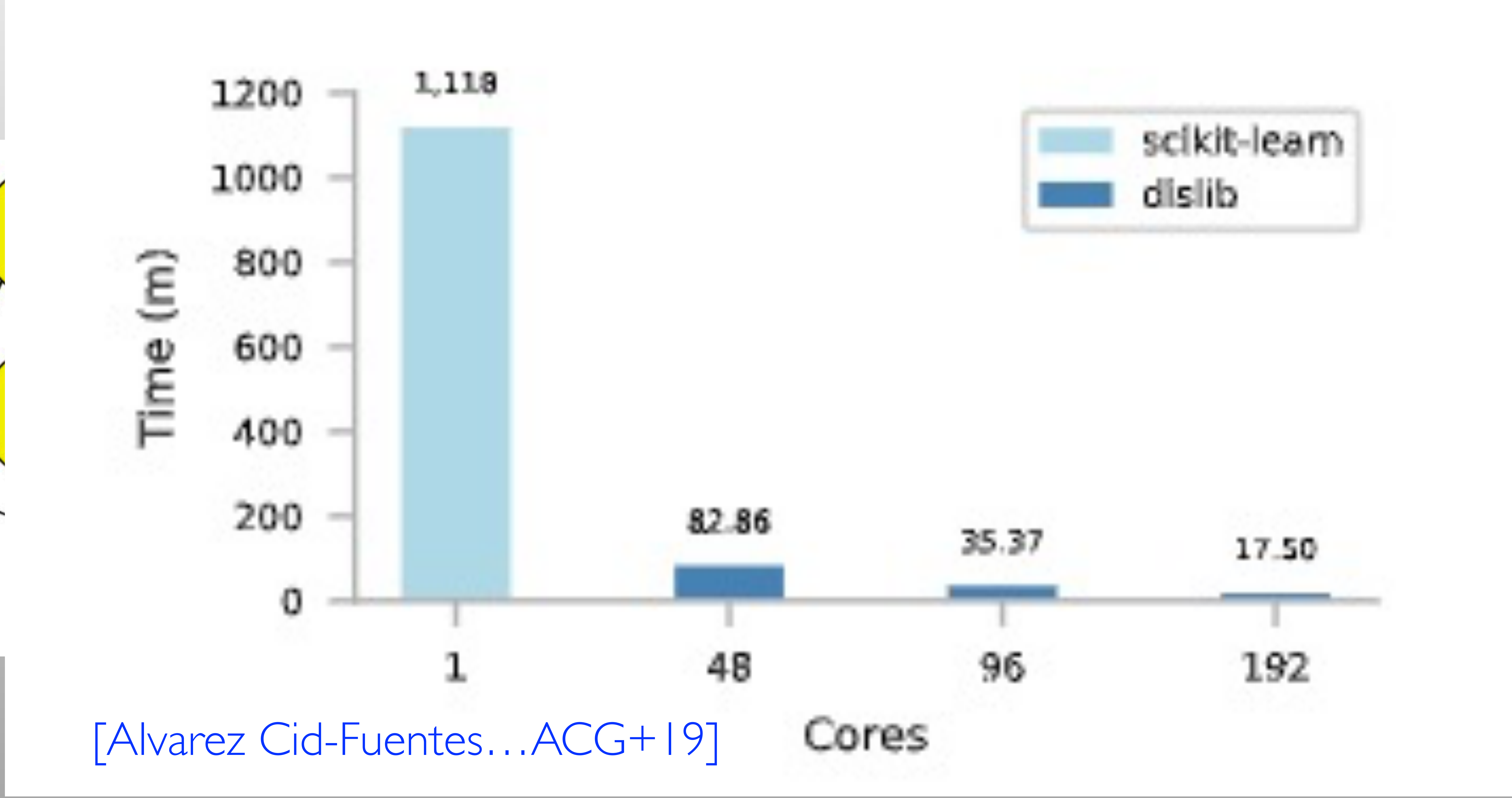


- Exploit parallelism of applications at task level
- Task — decorated python function
- Builds a task graph taking into account data dependencies
- Schedule and execute application in the distributed environment based on the graph

DBSCAN IN PARALLEL

Use PyCOMPSs framework [Tejedor+15]



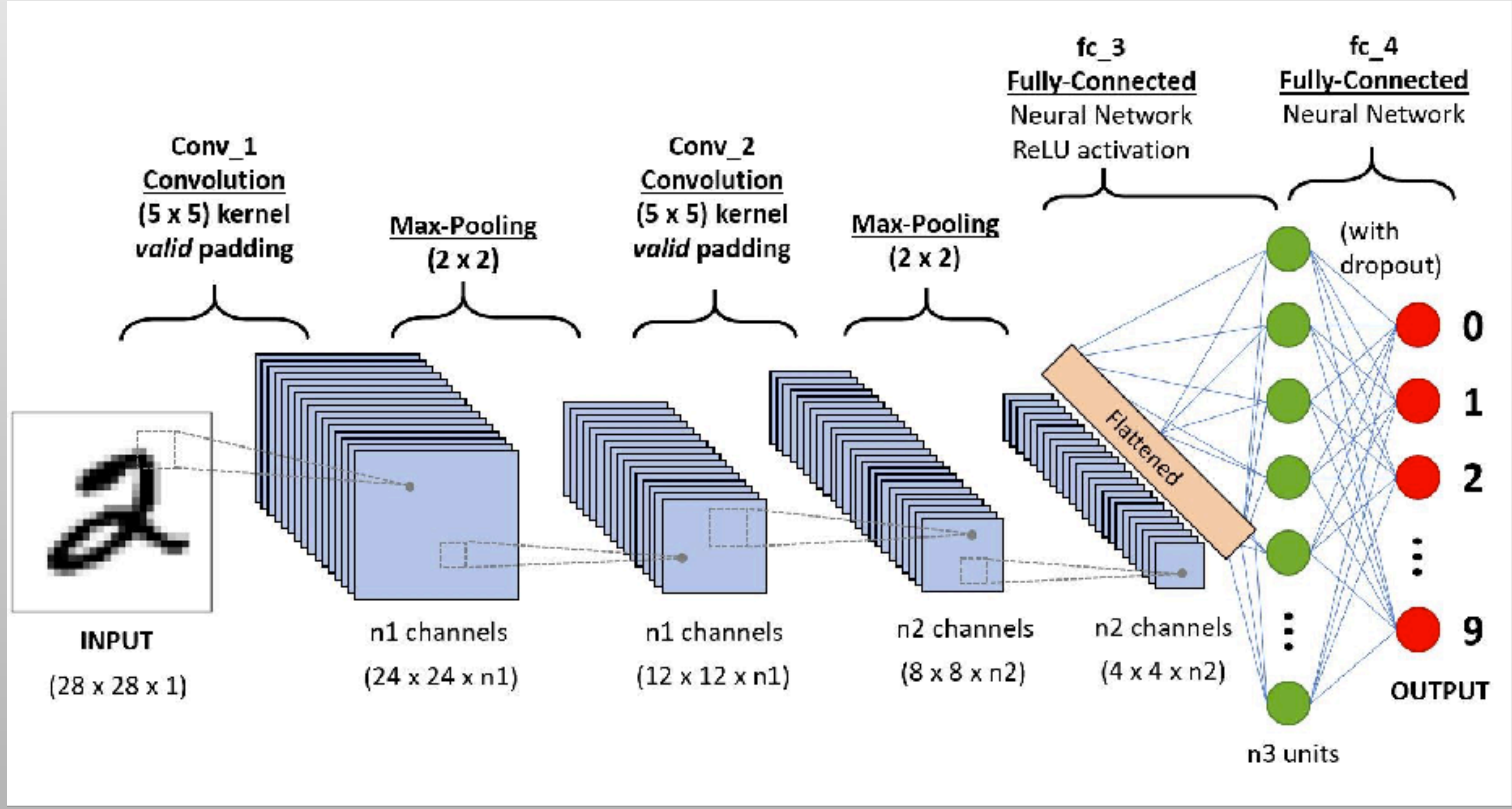


[Alvarez Cid-Fuentes…ACG+19]

plications at

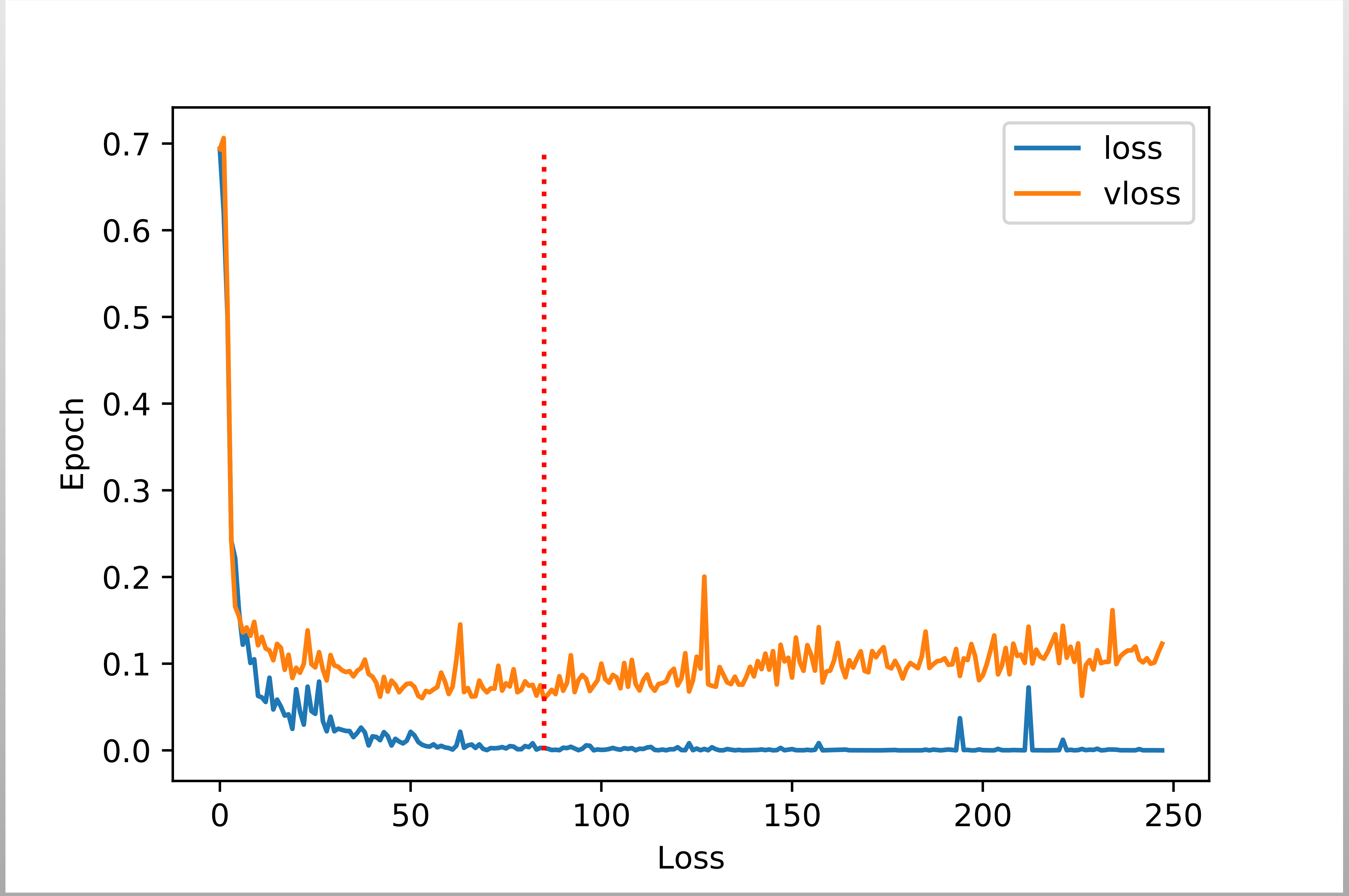
on function
g into
cies
pplication in
nent based

Deep Learning for OC recognition



- Introduction of deep learning architecture for a more robust classification (feature extraction)
- Training with real OCs + simulated data (isochrones from Padova [Bressan+12]) — ~20.000 samples
- Training in two steps
  - Minimise validation loss
  - Minimise false positives from [Castro-Ginard+19]

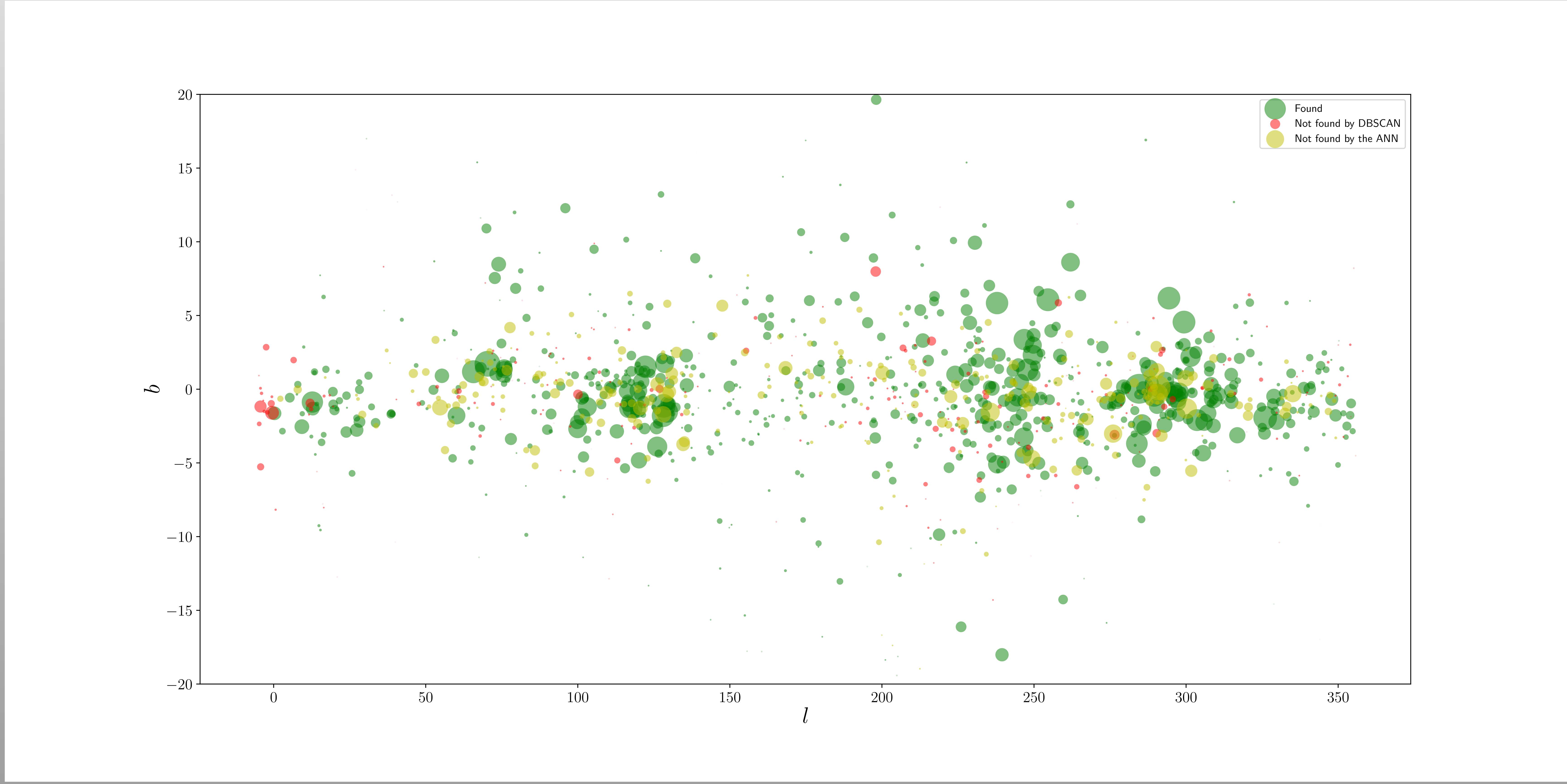
DEEP LEARNING FOR OC RECOGNITION



- Introduction of deep learning architecture for a more robust classification (feature extraction)
- Training with real OCs + simulated data (isochrones from Padova [Bressan+12]) — ~20.000 samples
- Training in two steps
  - Minimise validation loss
  - Minimise false positives from [Castro-Ginard+19]

Method Limitations

- Detection limited to the most compact cluster in a given *LxL* area.
- Only clusters with well-defined CMD.

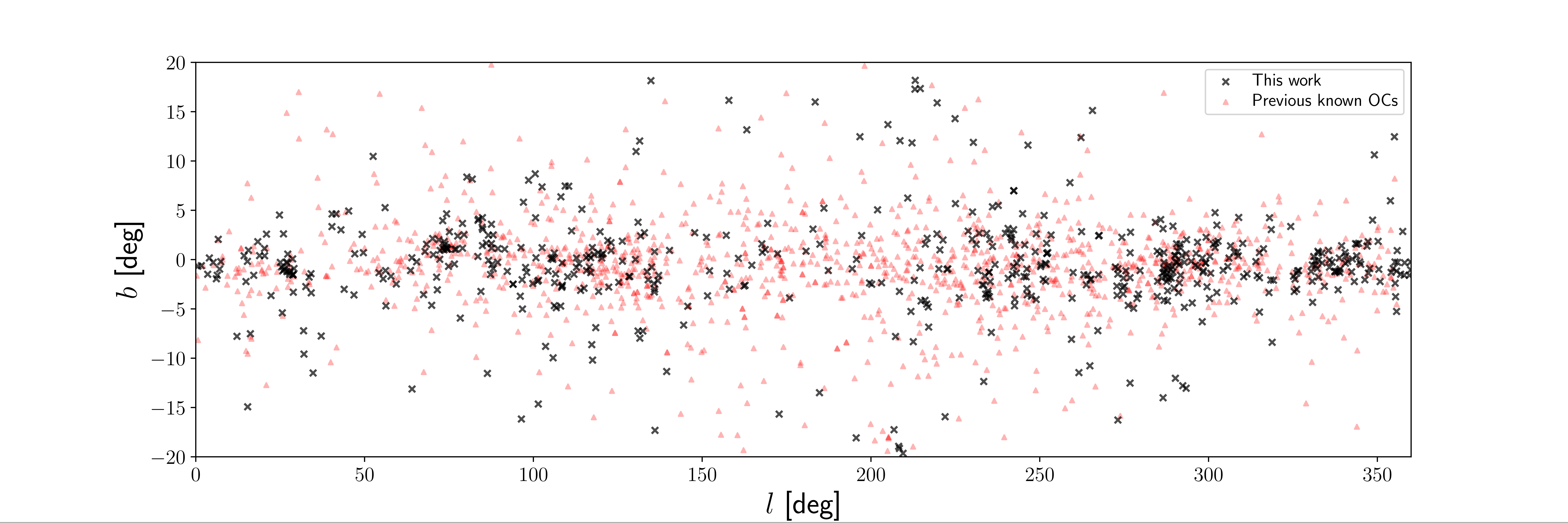


Some Results

- More than 650 UBC clusters.

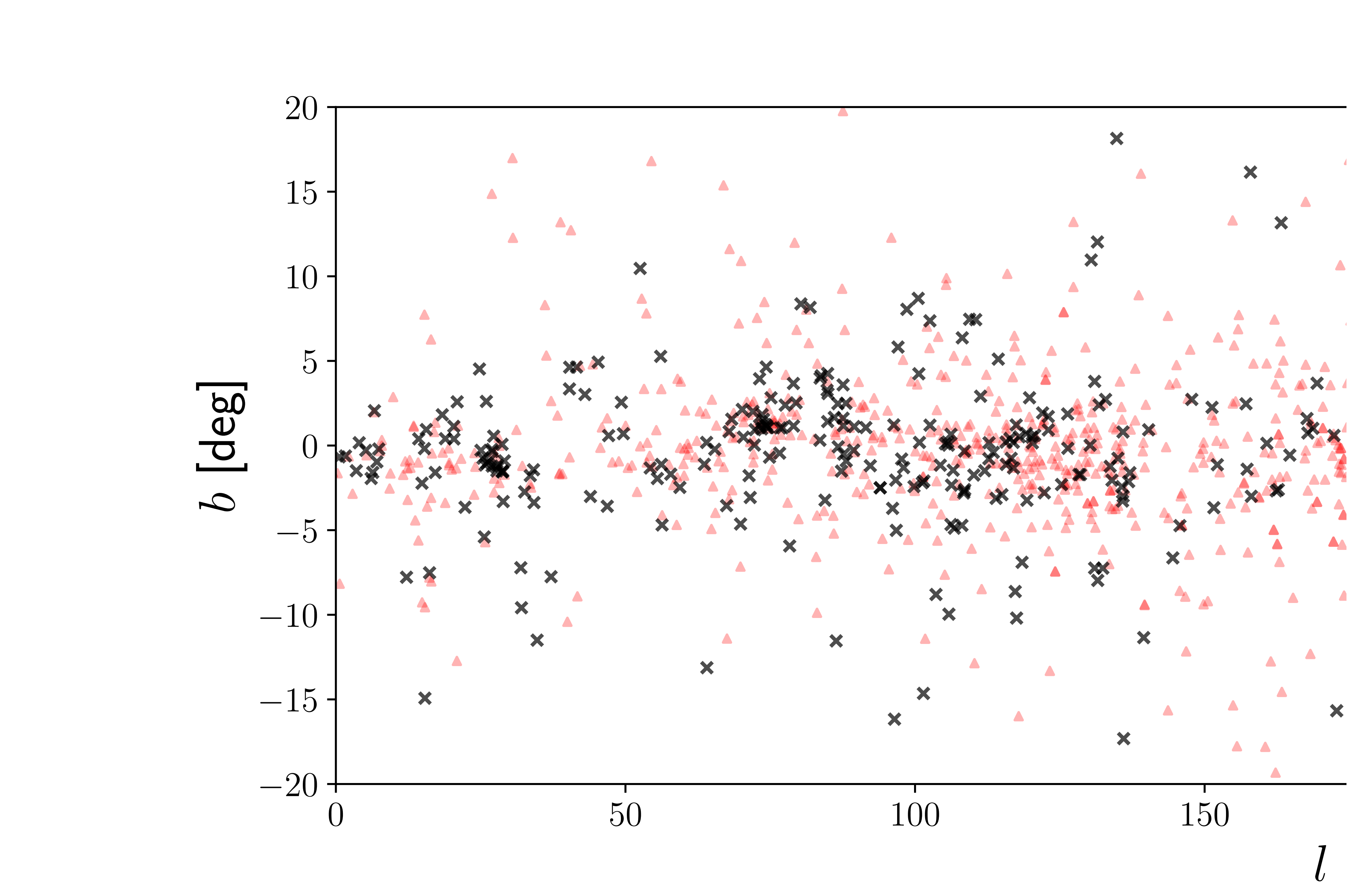
Some Results

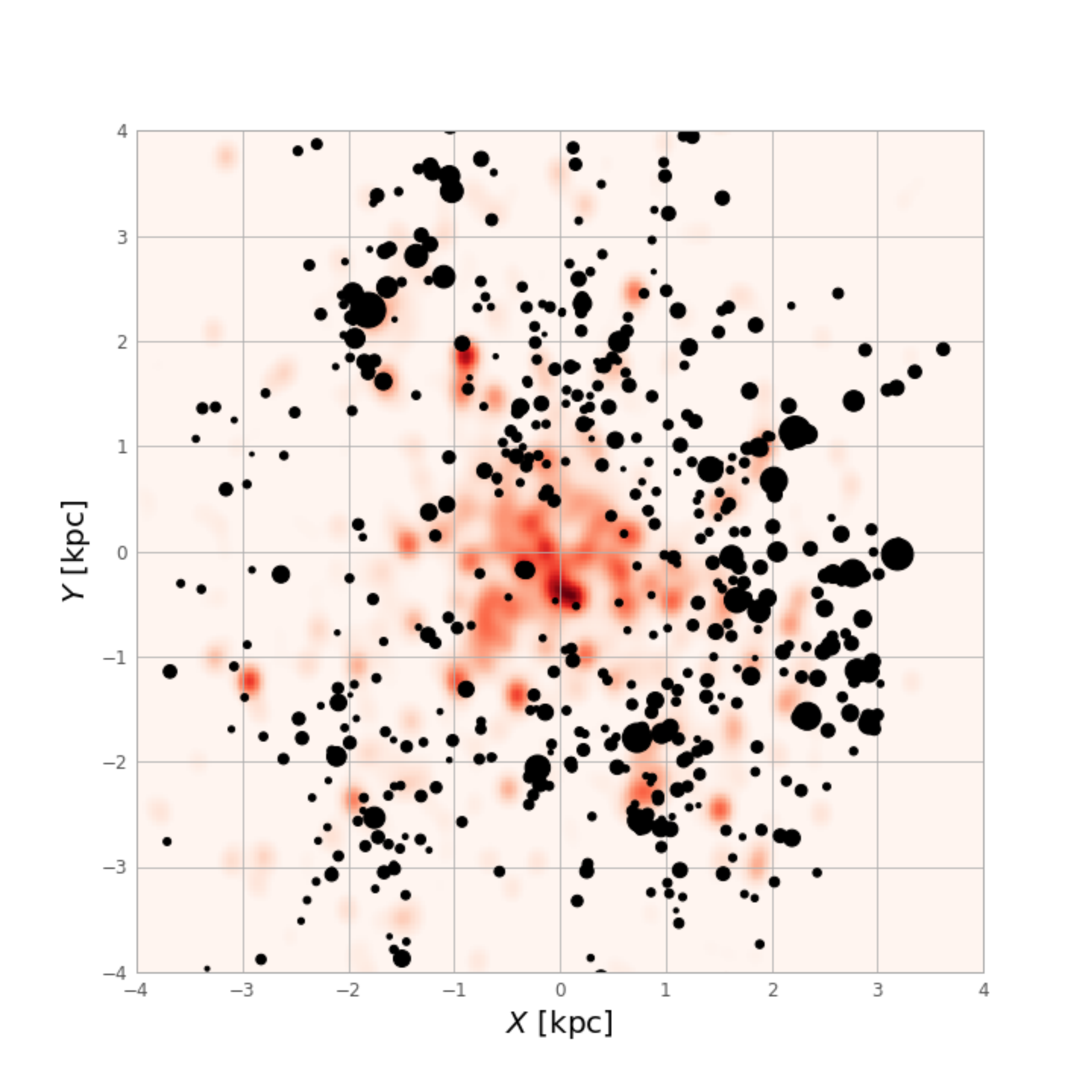
- More than 650 UBC clusters.



Some Results

- More than 650 UBC clusters.



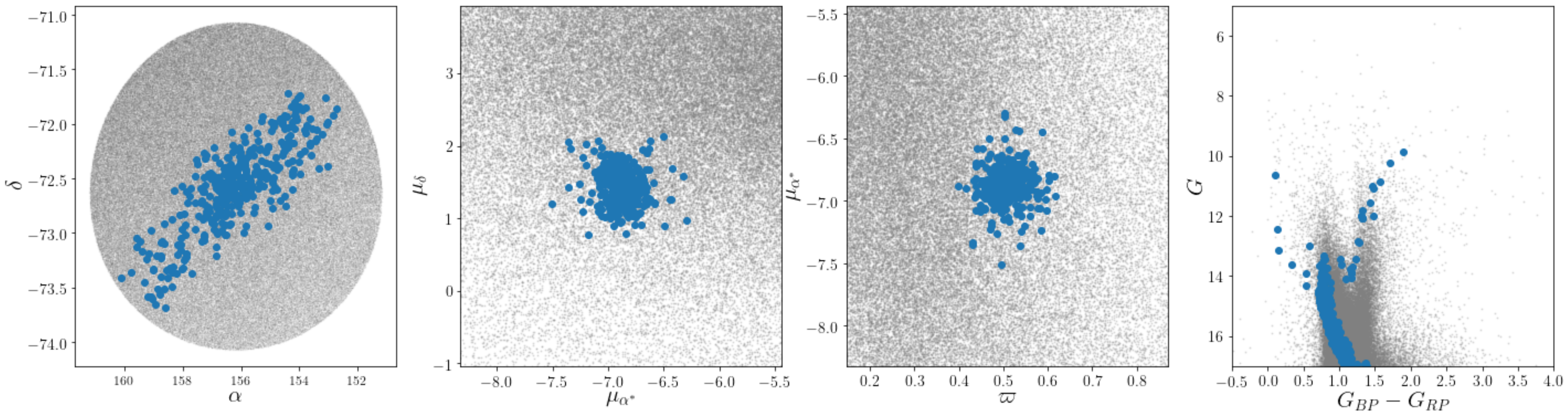


Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.

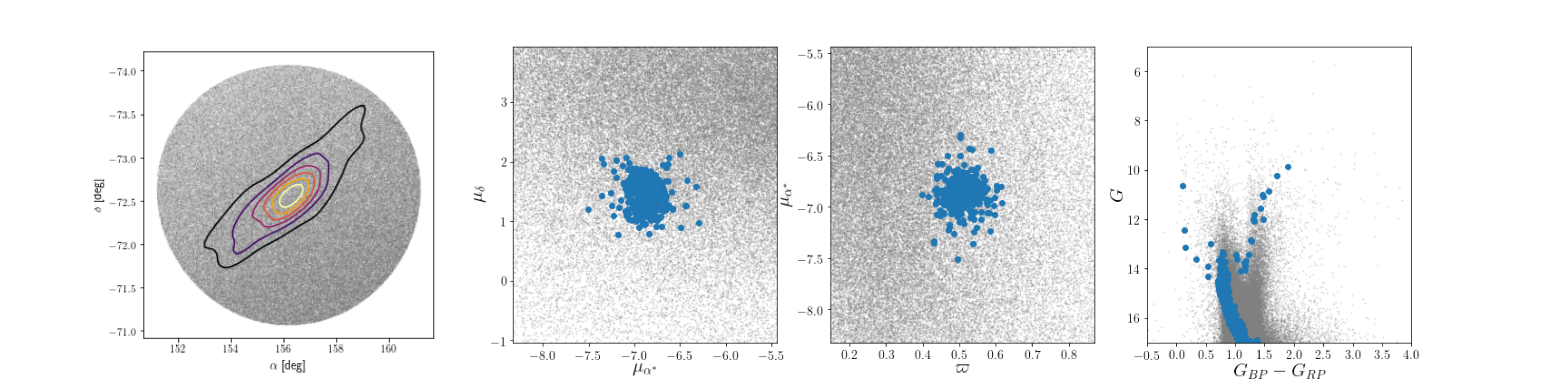
Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.



Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.

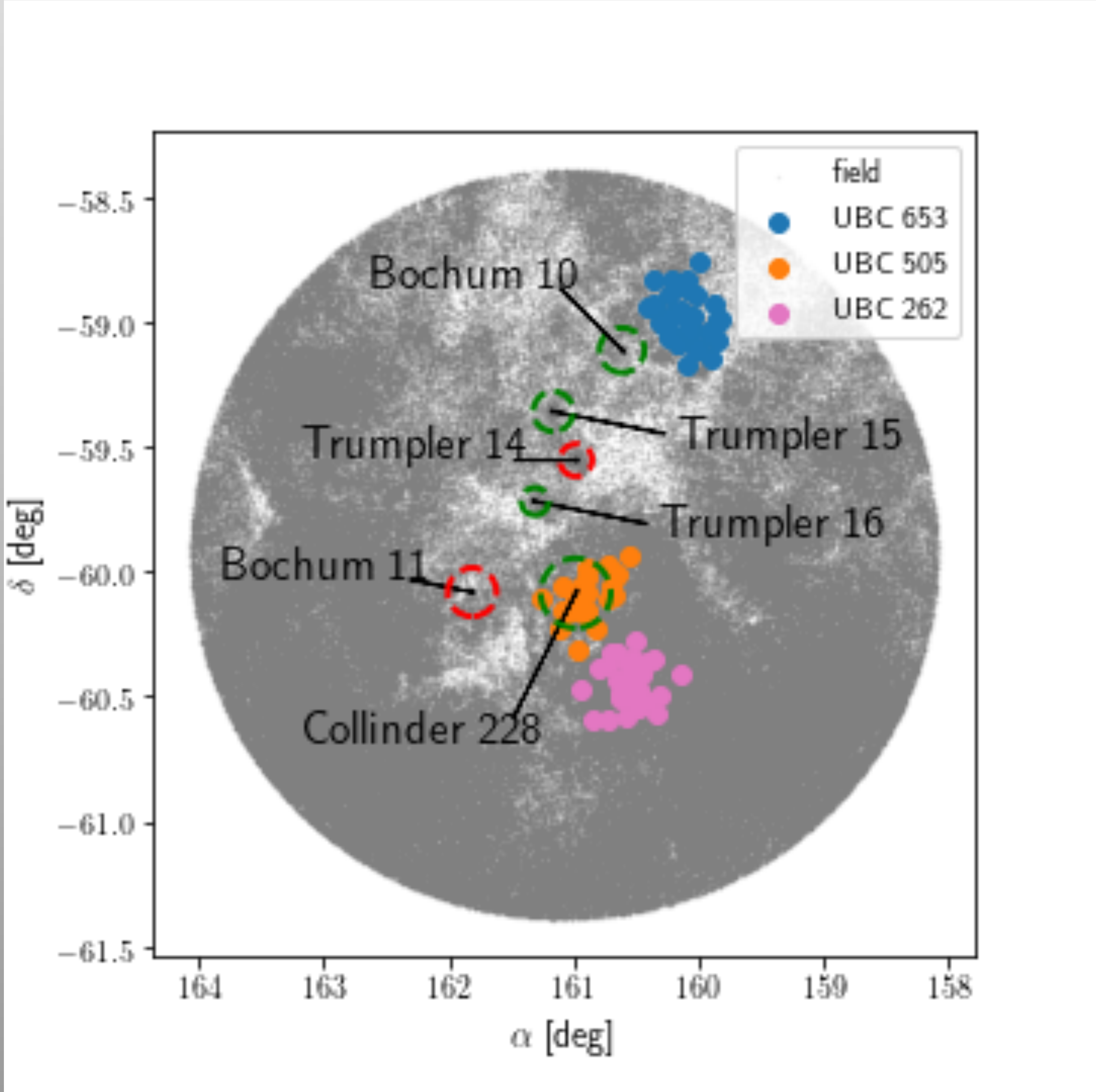


Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.
- Can detect substructure in richer regions.

Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.
- Can detect substructure in richer regions.



Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.
- Can detect substructure in richer regions.
- Still detecting clusters at 1-2 kpc.

Some Results

- More than 650 UBC clusters.
- Can detect some features of individual clusters.
- Can detect substructure in richer regions.
- Still detecting clusters at 1-2 kpc.