(Big) Data mining Gaia DR2 to study the Galactic open cluster population

Alfred Castro-Ginard[1]

Carme Jordi[1], Xavier Luri[1], Laia Casamiquela[2], Friedrich Anders[1], Tristan Cantat-Gaudin[1], Lola Balaguer-Núñez[1]

1: Universitat de Barcelona, Institut de Ciències del Cosmos
2: Laboratoire d'Astrophysique de Bordeaux

Context

- Trying to characterise the pre-Gaia catalogued OCs with Gaia DR2, only 1169 objects were found (from ~3000) and 60 new OCs were serendipitously detected [Cantat-Gaudin,…,ACG+2018]
- 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 1 kpc.
  - [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).

*Research highlight by A&A. Gaia image of the week: https://www.cosmos.esa.int/web/gaia/iow_20200514

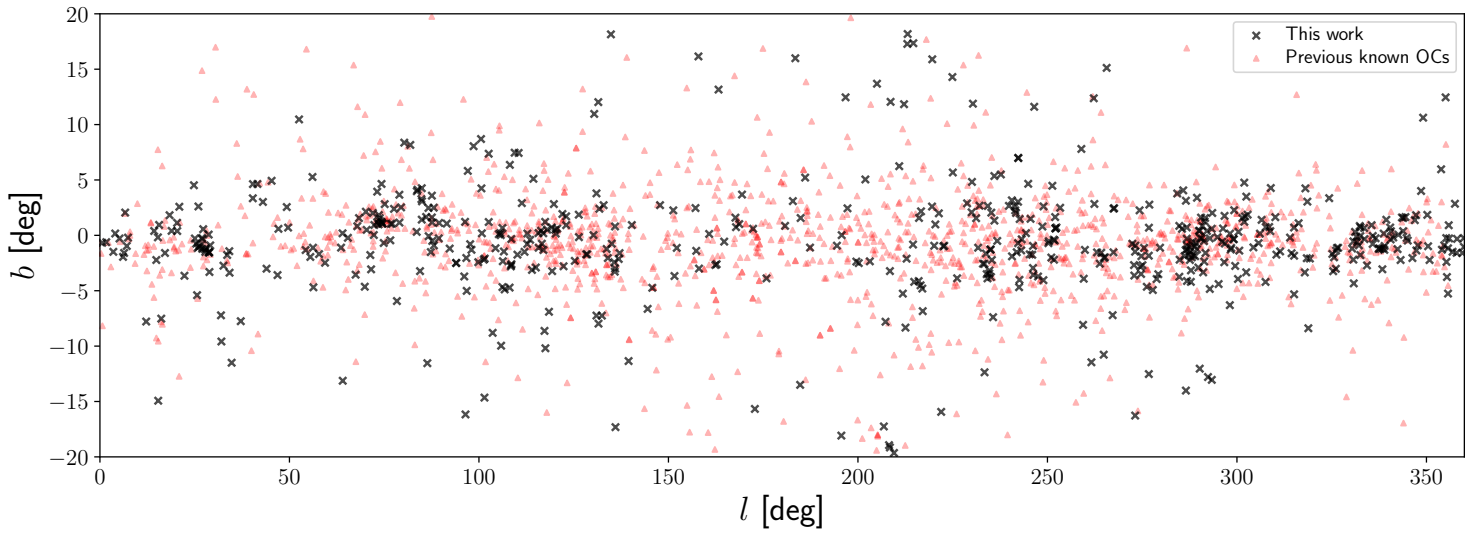
Description of the work

- Data mining methodology to automatically search for OCs in the Gaia DR2 archive
- Based on two machine learning methods:
  - Unsupervised: clustering on 5-D astrometry (position, parallax and proper motion) with DBSCAN (density based clustering algorithm).
  - Supervised: classification of over-densities by the recognition of isochrone patterns in CMDs from real OCs using an Artificial Neural Network.

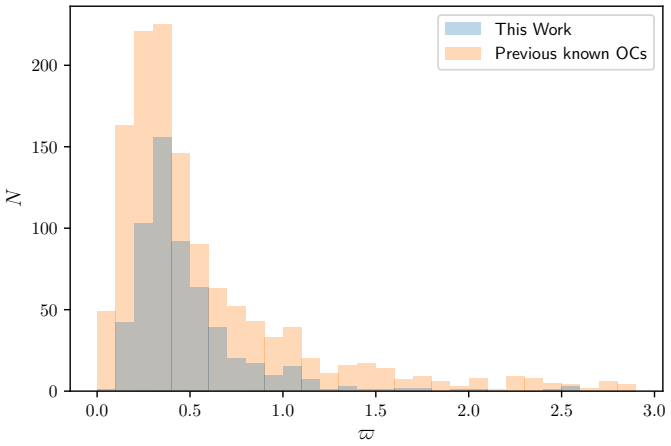
- Big Data problem: Galactic disc (|b| < 20 deg), up to G = 17 (10^8 stars).
  - Parallelization of DBSCAN based on data dependencies (PyCOMPSs, [Tejedor+15], [Álvarez Cid-Fuentes,…,ACG+19]) -> Barcelona SuperComputing Center (MareNostrum).
  - Deep Learning ANN for more robust feature extraction (isochrones in CMD).

Results

- More than 650 detected OCs (UBC clusters).

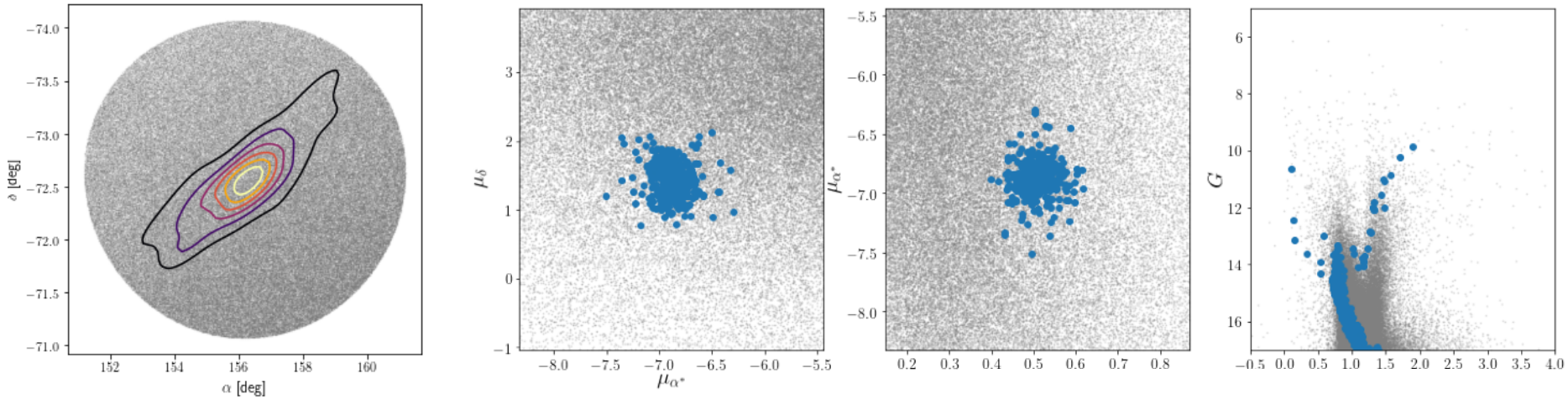


- Still detecting clusters at 1-2 kpc (census was thought to be complete).

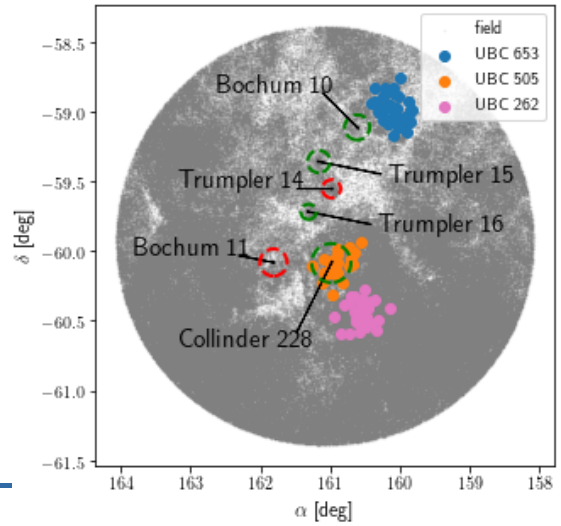


Results

- Can detect features of individual clusters (UBC 274).



- Can detect substructure in rich regions (Carina nebula).



Impact and prospects for future

- Science enabling catalogue.
- Use ANN to compute ages, absorptions and distances for OC population [Cantat-Gaudin,Anders,ACG+20].
- Re-analysis of the Milky Way's cluster age function [Anders,…,ACG+20].
- Open cluster view of Milky Way's spiral structure [Castro-Ginard+,in preparation].