A distributed and enhanced implementation of unsupervised ANNs applied to spectrophotometry clustering in the ESA Gaia mission

EWASS 2017 — Prague

Daniel Garabato
daniel.garabato@udc.es

University of A Coruna — Department of Computer Science

June 2017

Table of contents

The Gaia Mission

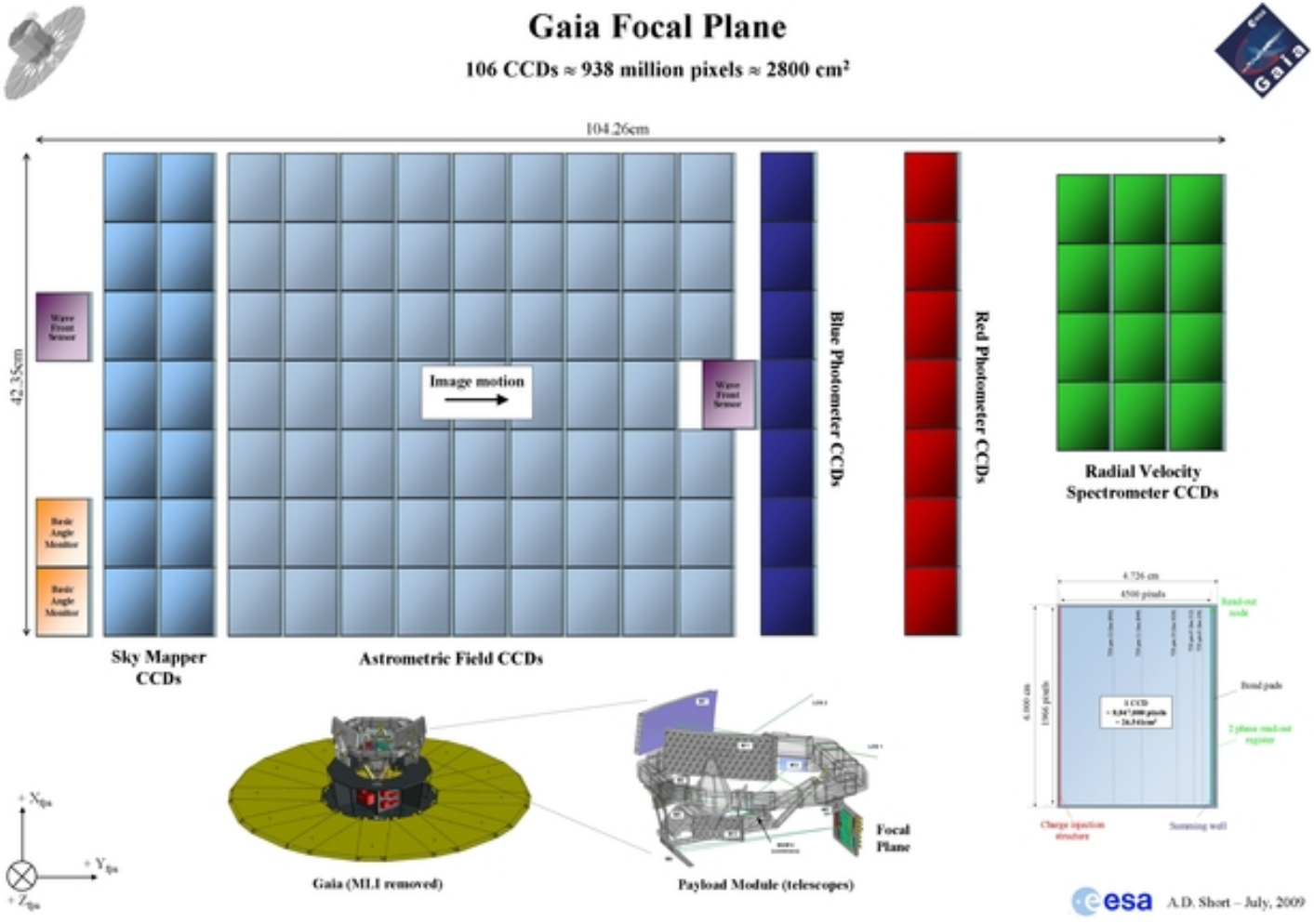
Overview

- Cornerstone mission of the European Space Agency
- Study the formation and evolution of the Milky Way
- Survey celestial bodies down to magnitude 20
- Launched in 2013 — 5 years of routine operations
- Final catalog foreseen for 2022–2023
- Up to 10^{12} observations will be measured (80 epochs)
- A PetaByte of information will be generated ($\sim$ 36 GB per day)

gaia

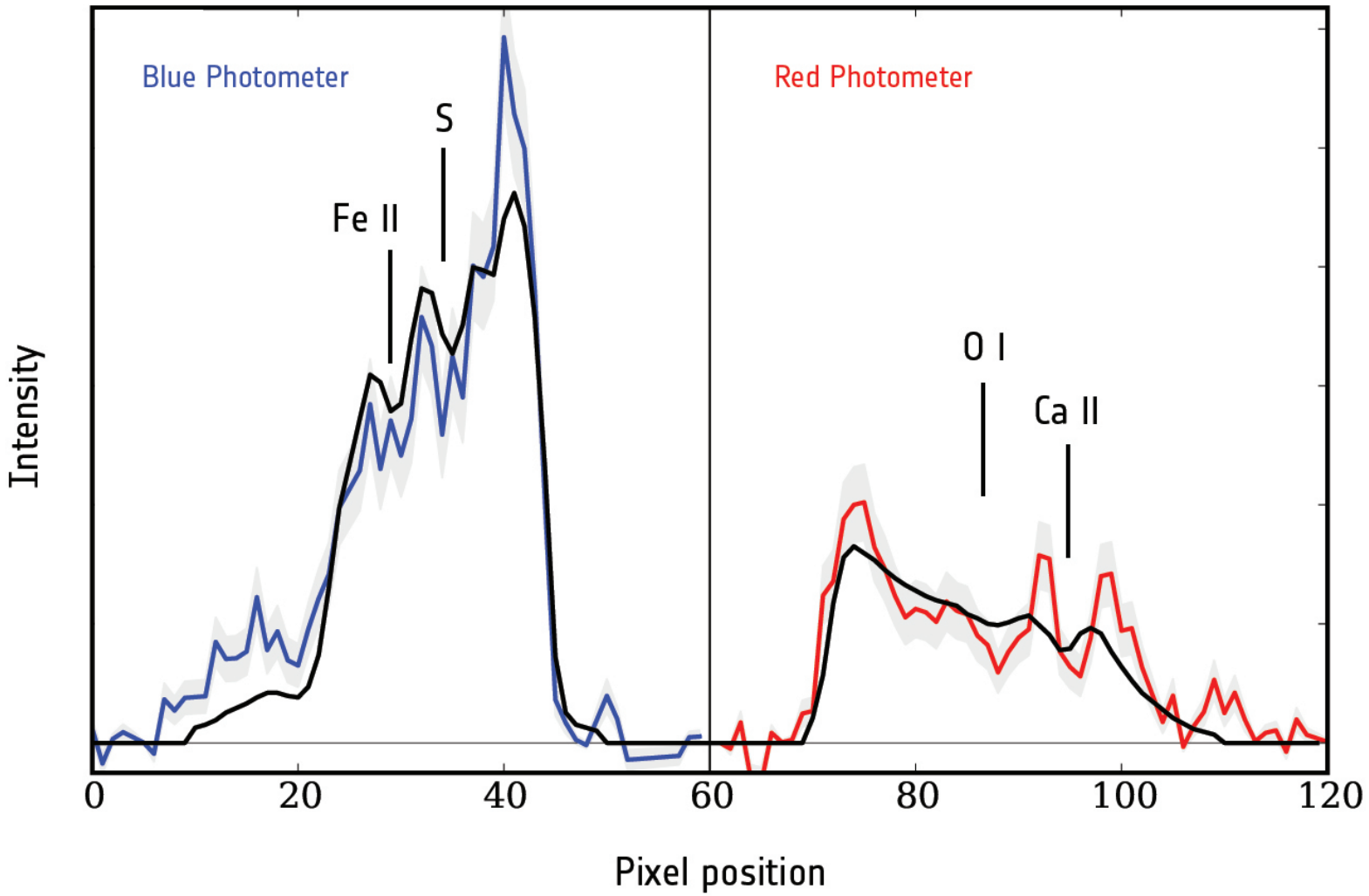
The Gaia Mission

The data — Focal Plane



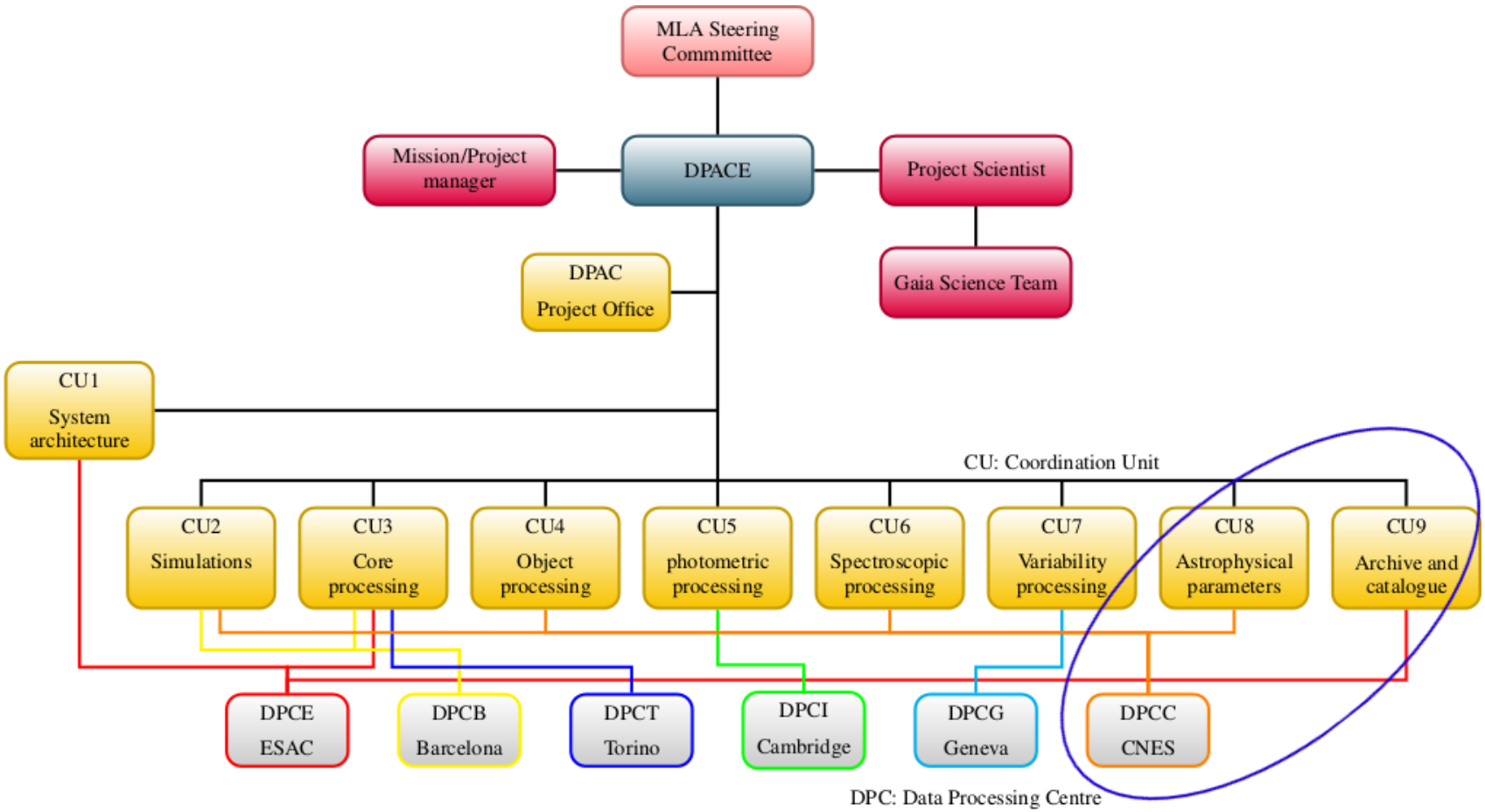
The Gaia Mission

The data — Blue & Red photometer spectra

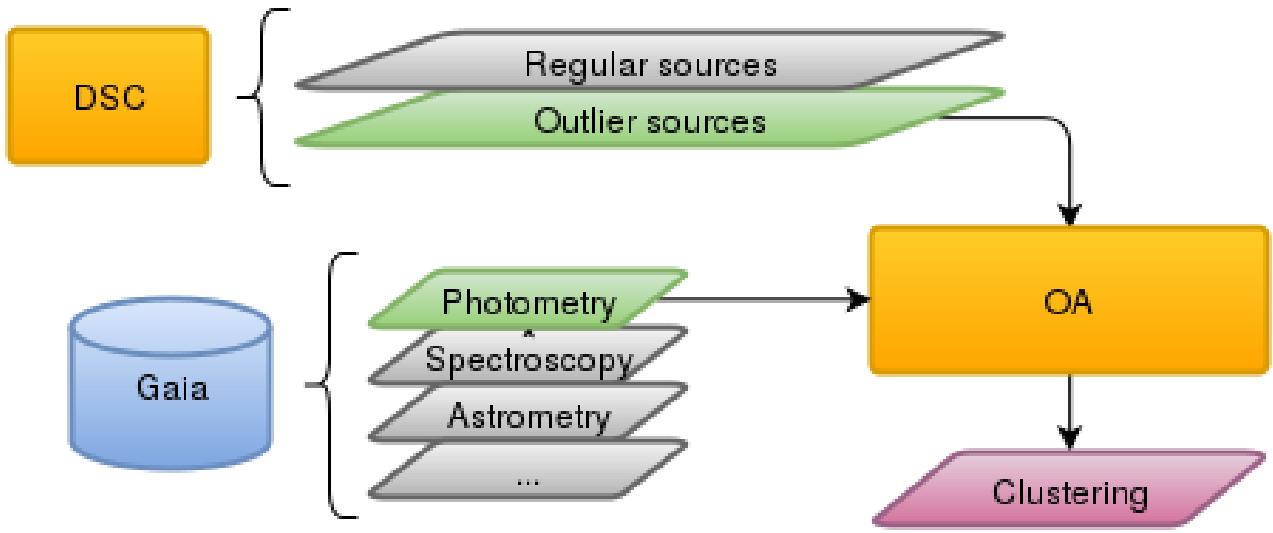


The Gaia Mission

Data Processing and Analysis Consortium (DPAC)



- Main classifiers:
  - Discrete Source Classifier (DSC)
  - Object Cluster Analysis (OCA)

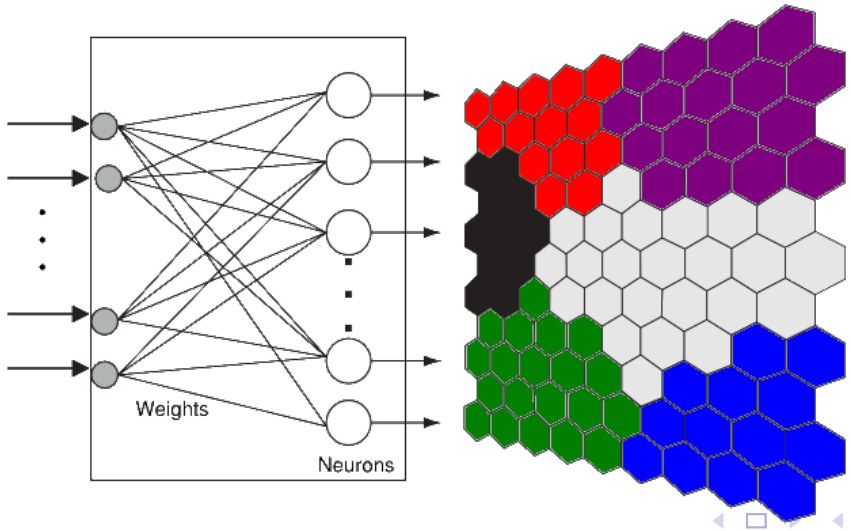


- **Outlier Analysis (OA)**

Outlier Analysis (OA)

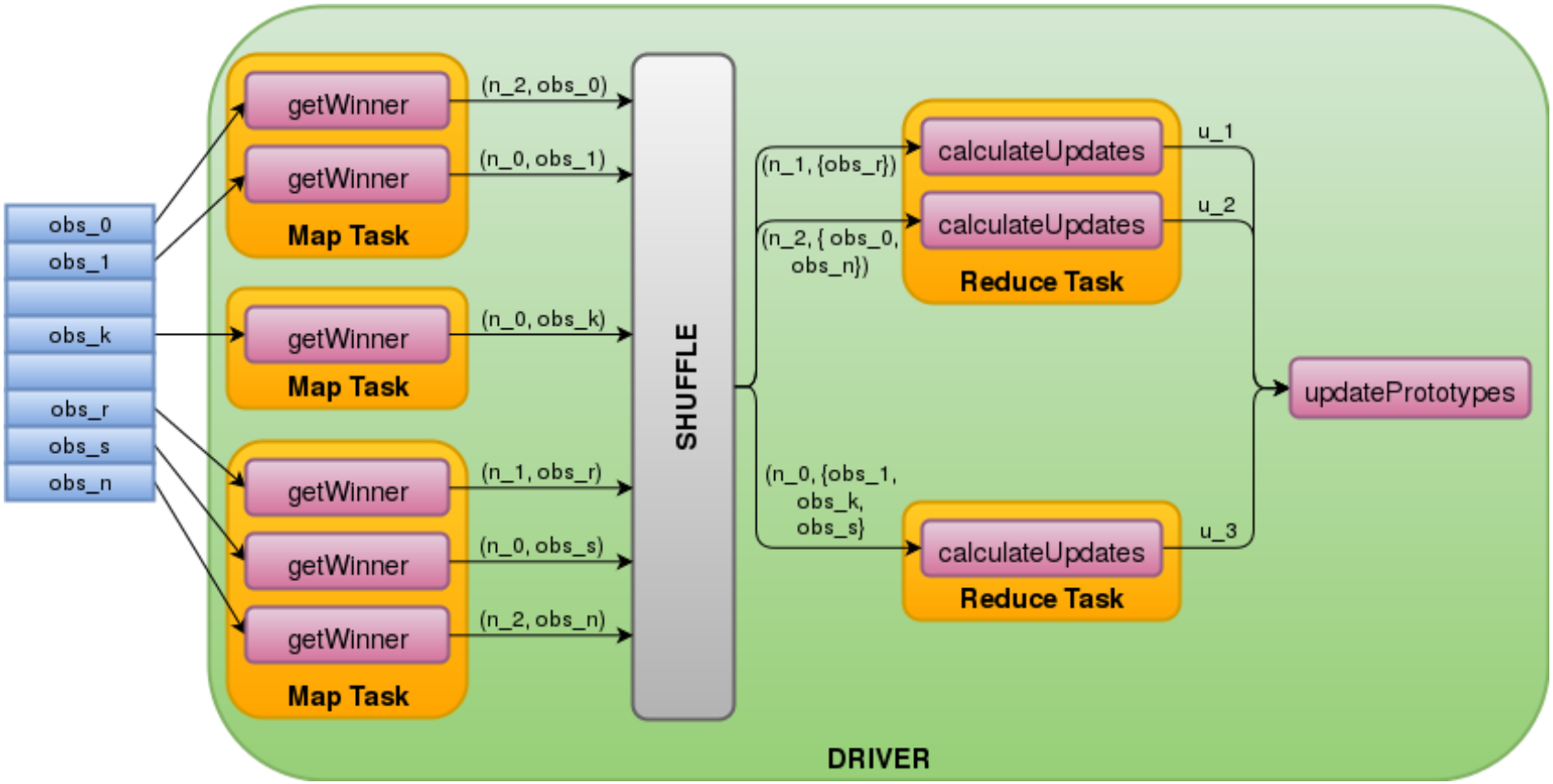
Overview

- Analyze outlier sources:
  - Misclassified sources
  - Damaged sources or artifacts
  - Sources whose nature is unkwnown
- Clustering — Self-Organized Maps (SOM):
  - **Unsupervised learning**: Group sources by their nature
  - Reduce the high dimensionality of the data
  - Distributed computing ($\sim$ 1 PB) $\rightarrow$ Batch SOM
  - Optimization $\rightarrow$ Fast SOM



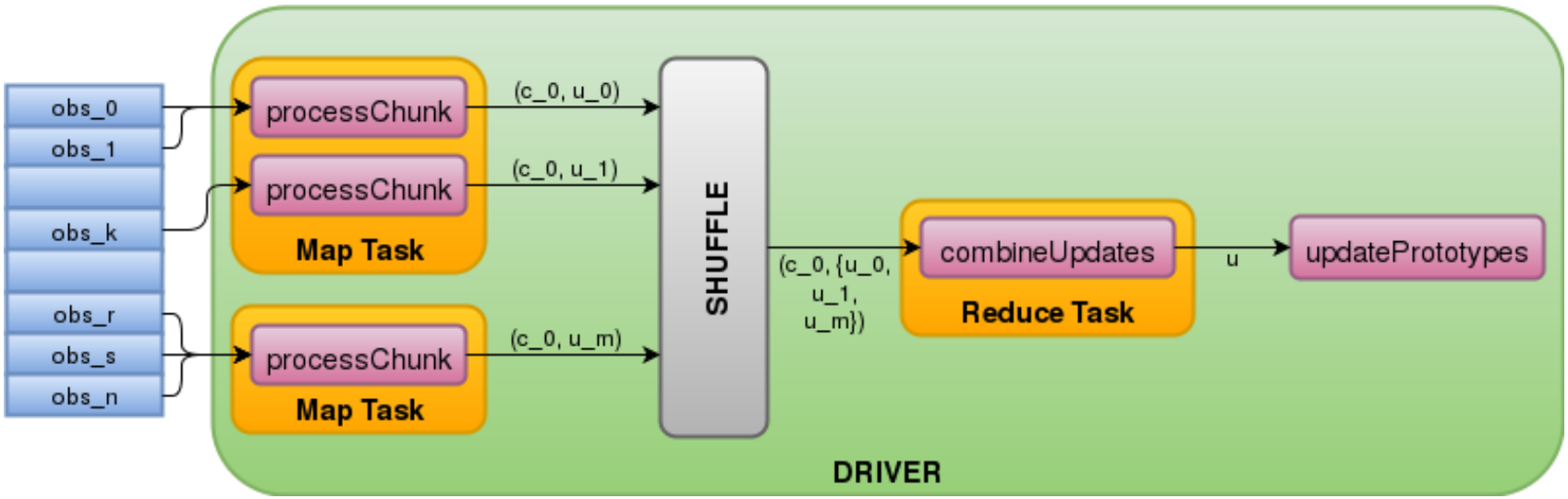
A distributed SOM

Apache Hadoop — Apache Spark



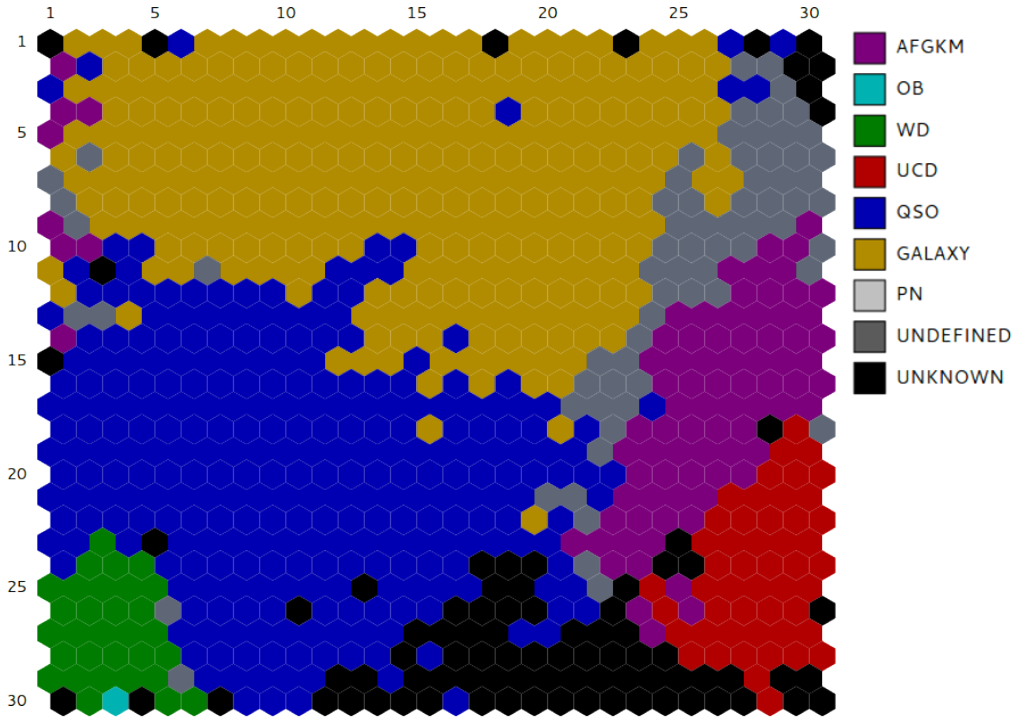
A distributed SOM

SAGA framework (CNES)



Results

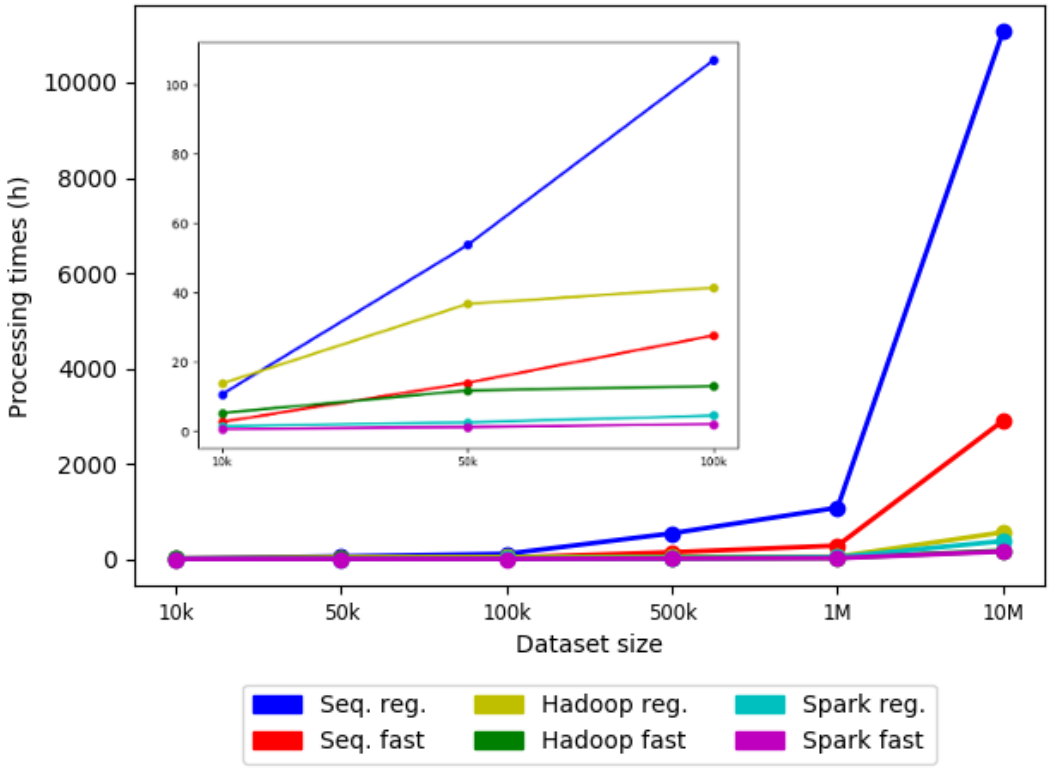
Clustering performance



To check if the SOM is working correctly, we train a map with a well known data set, such the SDSS with 10125 objects, and we use a labeling process in order to idenfify the clusters.

Results

Execution times



- SAGA framework:
  - Introduces a $\sim 10\,\%$ of overhead with respect to a pure Hadoop due to the framework's management tasks
  - OA is expected to be executed in a couple of weeks using CNES hardware to process 100M sources

Conclusions & Future work

- A powerful tool for unsupervised classification of outliers has been developed
- This algorithm is very useful to identify and classify "weird celestial objects", with special interest in those sources whose nature is unkwnown
- Such an algorithm is scalable and it can be applied to huge volumes of data
- The execution times were considerably reduced by using the Fast SOM approach
- We expect that the execution times can be even better using GPU computing techniques