REG Red Española para la explotación científica de Gaia

GUASOM: VO tool for the analysis of spectrophotometric data

Marco A. Álvarez
marco.antonio.agonzalez@udc.es

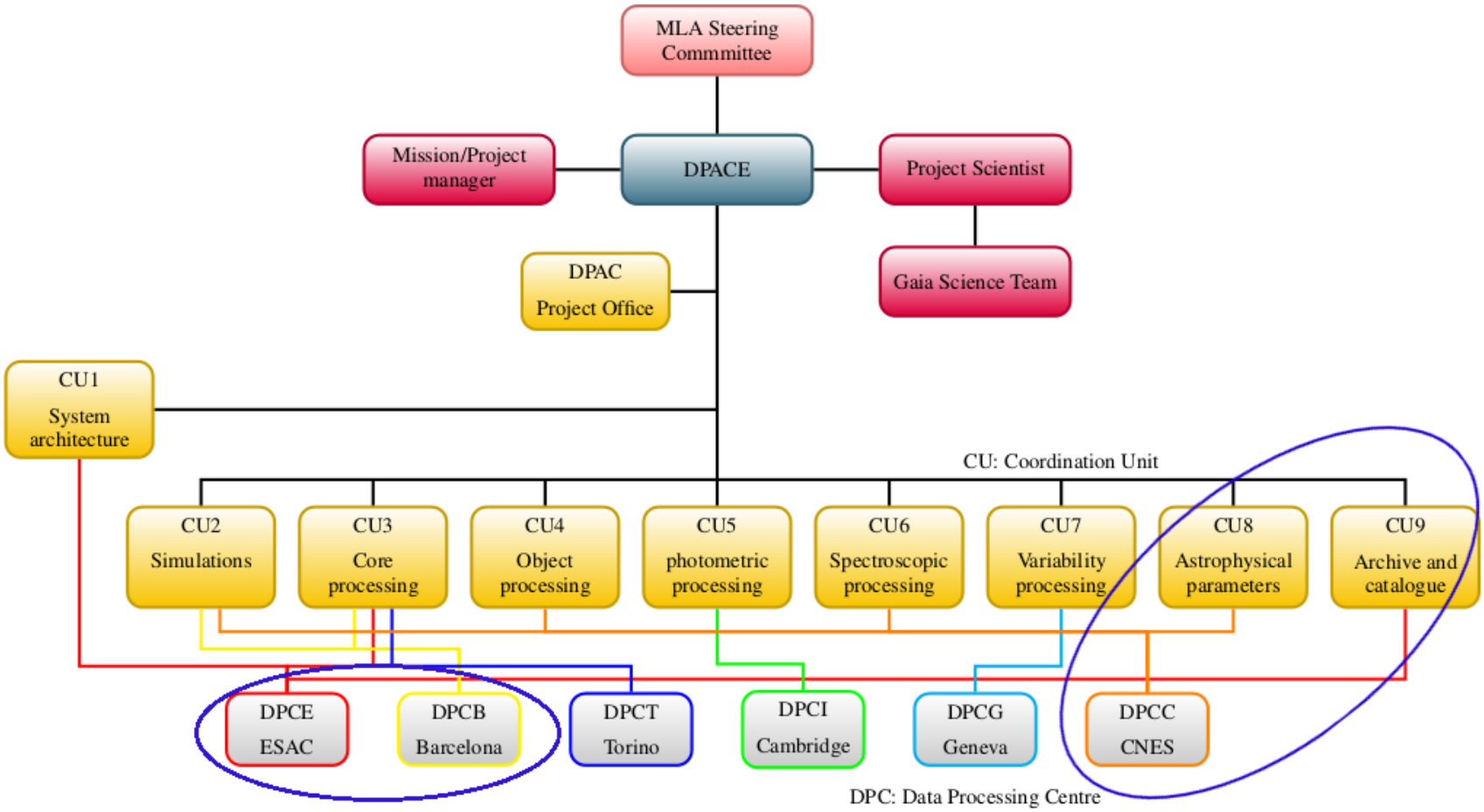
Faculty of Computer Science, A Coruña, Spain

LIA[2] UNIVERSIDADE DA CORUÑA

May 29th 2018

The Gaia Mission

Data Processing and Analysis Consortium (DPAC)



CU9 Catalogue Access

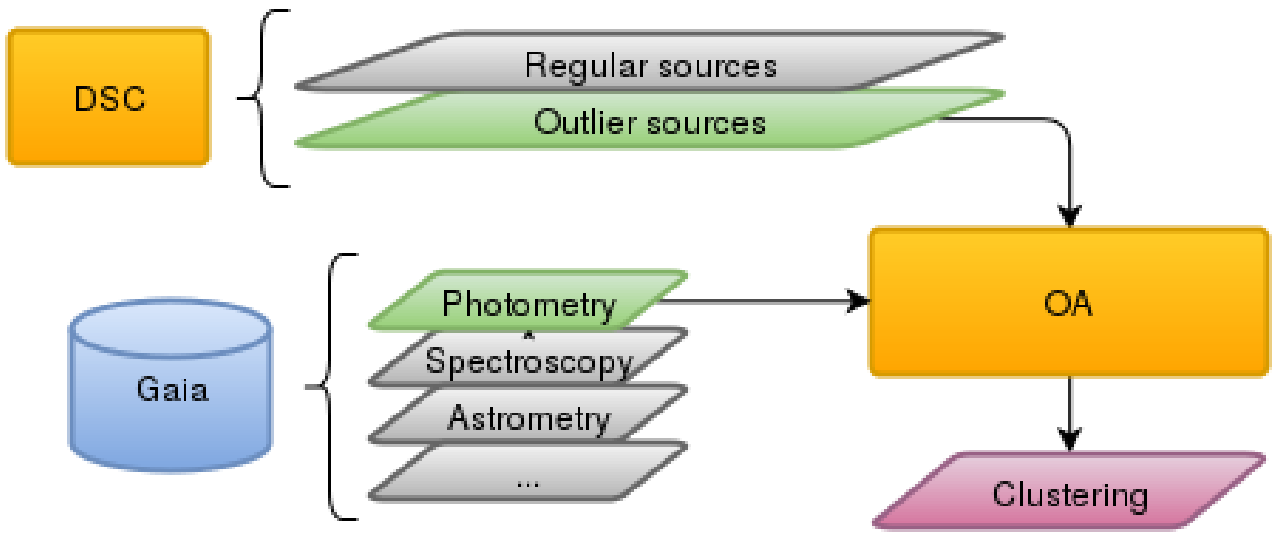
- **GWP-973 Data Mining**
  Responsible for providing a set of tools to the community in order to make the most of the Gaia Data Releases. It also provides the platforms for computer processing.
- **GWP-985 Clustering and advanced data selection for multi-D visualisation**
  The goal is to tackle the exploitation of the multi-dimensional character of the Gaia database through implementation of dimensionality reduction methods.
  Data clustering algorithms will be implemented for a multi-dimensional data selection in 2D and 3D projections.

Gaia DPAC
Data Processing & Analysis Consortium

CU8 Astrophysical Parameters

GWP-836 Source classification

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

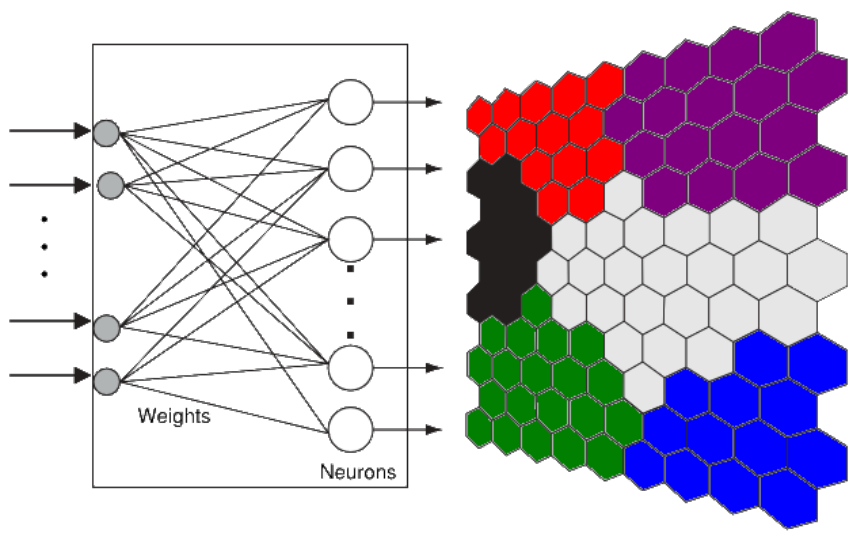


- **Outlier Analysis (OA)**

Outlier Analysis (OA)

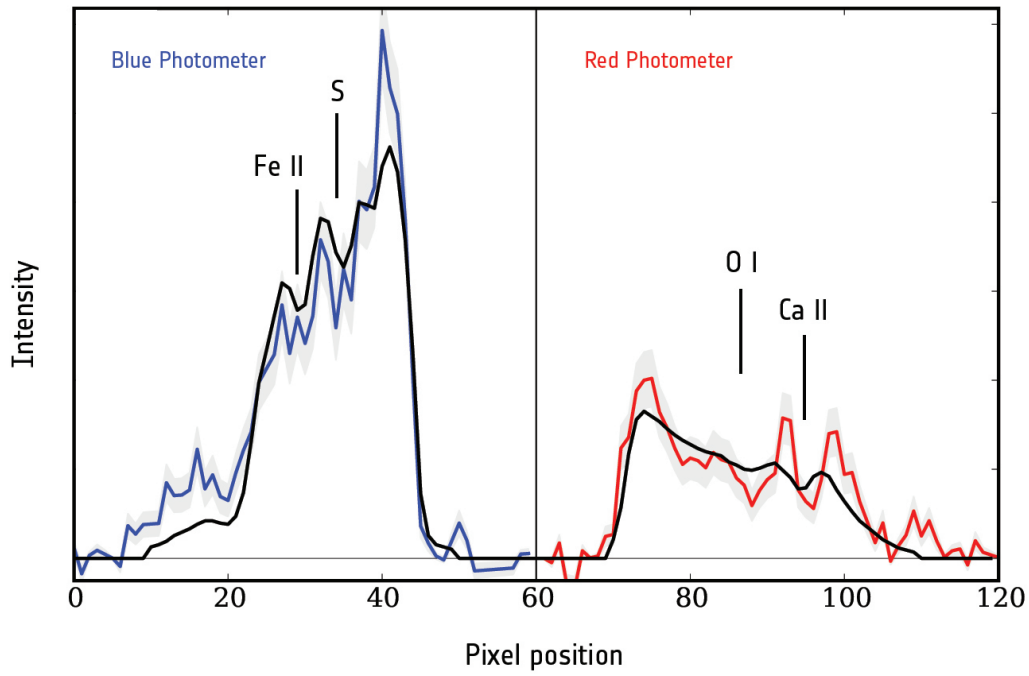
Overview

- Analyze outlier sources:
  - Misclassified sources
  - Damaged sources or artifacts
  - Sources whose nature is unknown
- Clustering | Self-Organized Maps (SOM):
  - **Unsupervised learning**: Group sources by their nature
  - Reduce the high dimensionality of the data
  - Distributed computing ($\sim$ 100 TB) ■ Batch SOM
  - Optimization ■ Fast SOM



Outlier Analysis (OA)

The data | Blue & Red photometer spectra

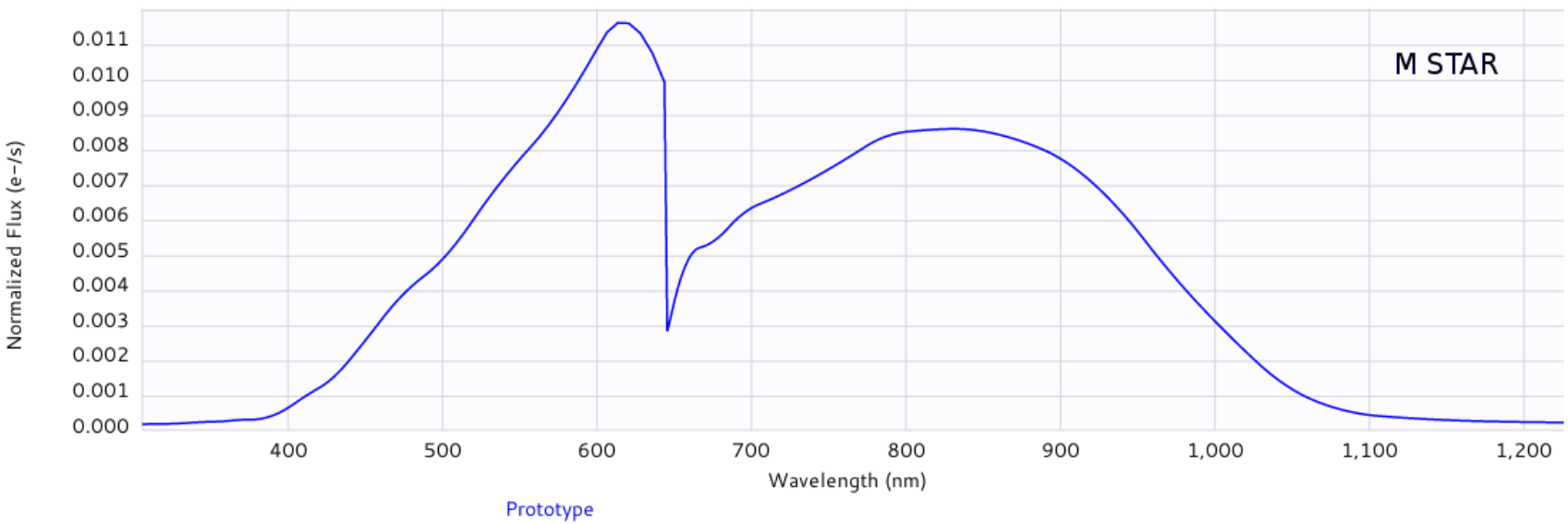


- Integrated in the SAGA framework (CNES)
- OA was successfully executed in the last validation cycle with approximately 23 million sources

Outlier Analysis (OA)

Main concepts

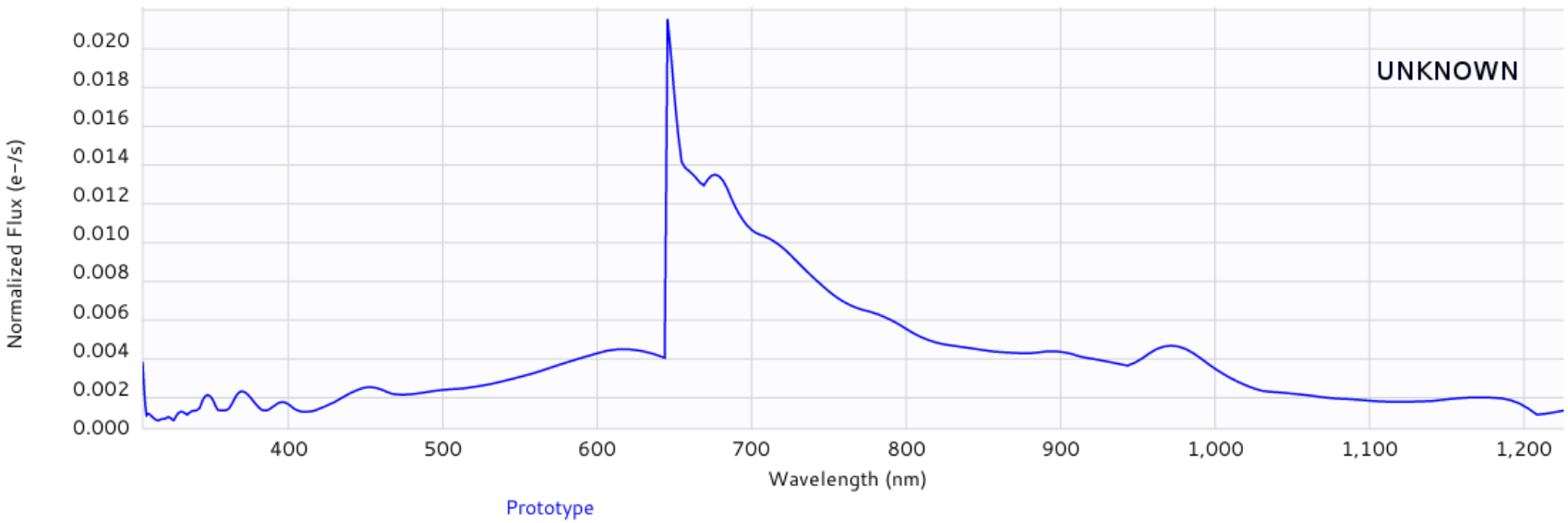
- Neurons are the main components
- Each neuron represents sets of objects of same nature
- **Prototype** is the representative BP/RP spectra of each neuron



Outlier Analysis (OA)

Main concepts

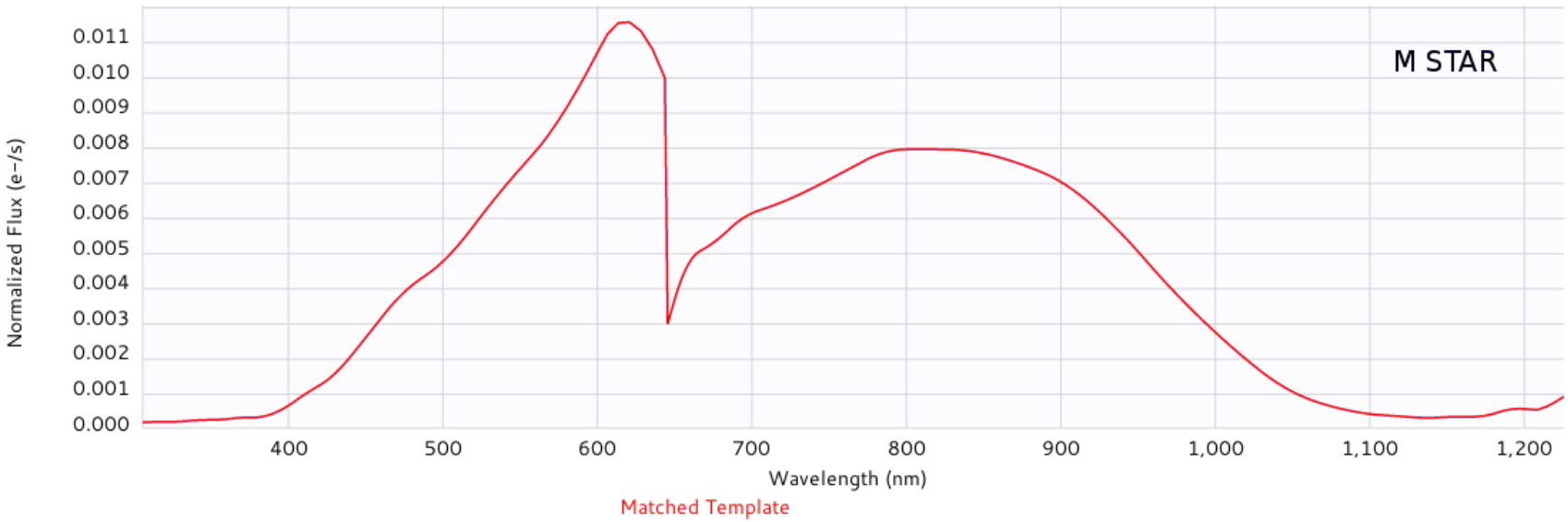
- Neurons are the main components
- Each neuron represents sets of objects of same nature
- **Prototype** is the representative BP/RP spectra of each neuron



Outlier Analysis (OA)

Main concepts

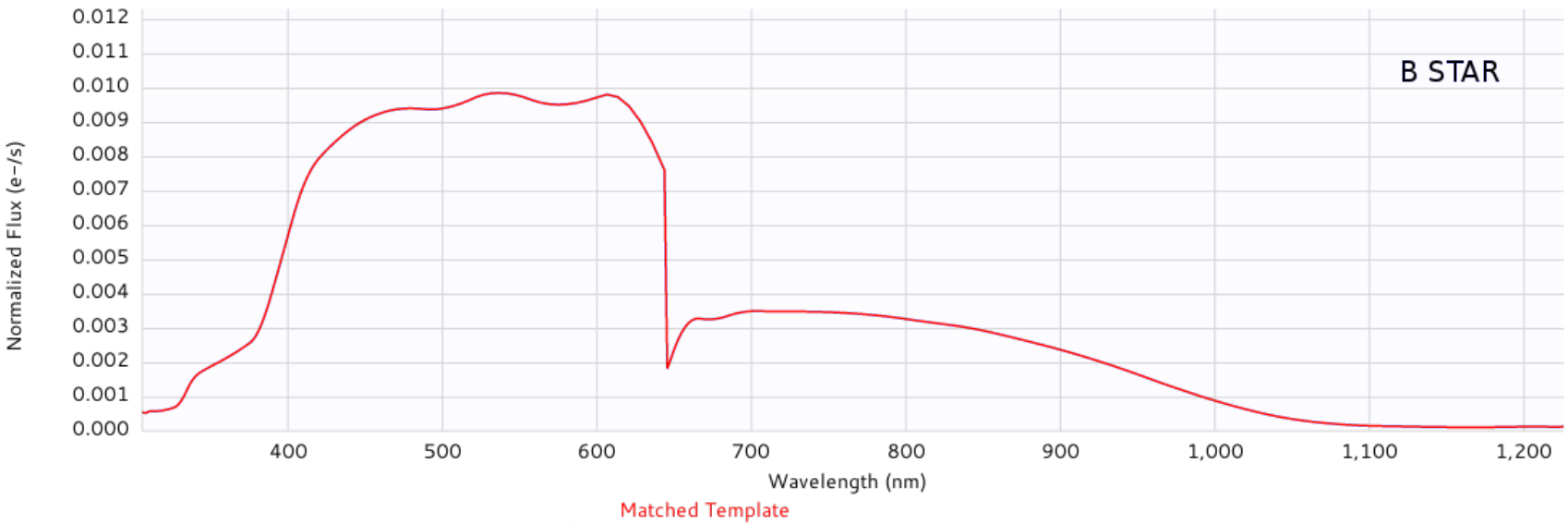
- Less distance between neurons means more similarity
- More distant neurons represent weirdest objects
- **Matched Template** is the most similar library template, which represents types of objects of the real world. Obtained directly from real observations or synthetic data



Outlier Analysis (OA)

Main concepts

- Less distance between neurons means more similarity
- More distant neurons represent weirdest objects
- **Matched Template** is the most similar library template, which represents types of objects of the real world. Obtained directly from real observations or synthetic data



Conclusions

- Our tool has demonstrated to be useful for detection of outliers, with special interest in those sources whose nature is unknown
- The visualization tool allows us to understand the inner knowledge in the SOMs
- The set of representations of a SOM are useful for a better understanding of different aspects of the data
- 3D visualization is useful to identify easily different regions of the map and to combine multiple representations in one
- Extra features like cross-match and SAMP communications gives our tool a flexible way to connect with other applications, which is extremely useful

Publications

Marco A. Álvarez, Carlos Dafonte, Daniel Garabato, and Minia Manteiga.

Analysis and Knowledge Discovery by Means of Self-Organizing Maps for Gaia Data Releases, pages 137–144.

Springer International Publishing, 2016.

Daniel Garabato, Carlos Dafonte, Marco A. Álvarez, and Minia Manteiga.

Distributed unsupervised clustering for outlier analysis in the biggest milky way survey: Esa gaia mission.

In *Ubiquitous Computing and Ambient Intelligence*, pages 840–852. Springer International Publishing, 2017.

Daniel Garabato, Carlos Dafonte, Minia Manteiga, Diego Fustes, Marco A. Álvarez, and Bernardino Arcay.

A distributed learning algorithm for self-organizing maps intended for outlier analysis in the GAIA – ESA mission.

pages 895–901. IFSA - EUSFLAT 2015, 2015.