

Analyzing Andromeda Galaxy data using Spark

Jose Nandez

SHARCNET – University of Western Ontario

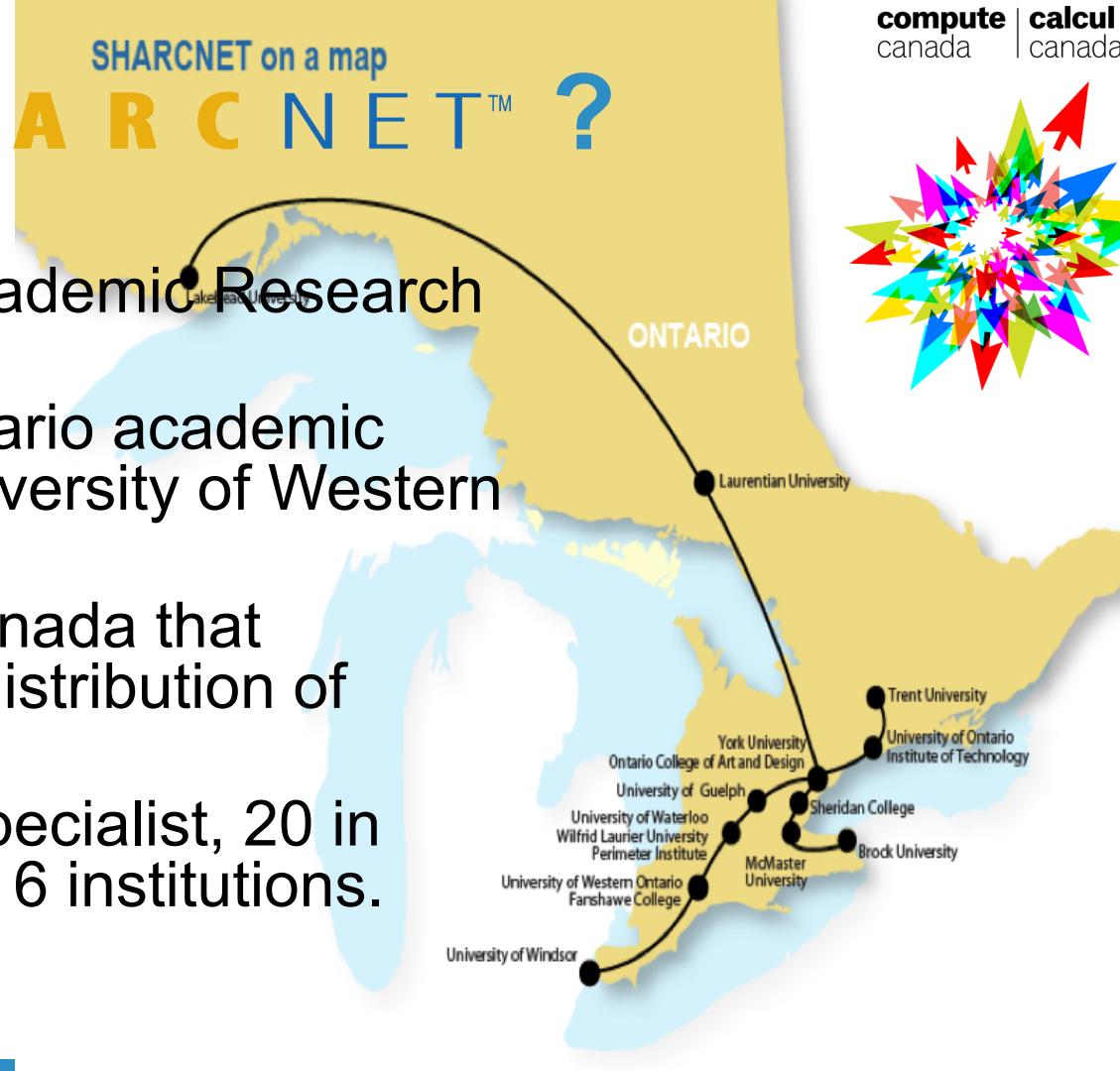
jnandez@sharcnet.ca



**SPARK
SUMMIT
EAST 2017**

What is SHARCNET™?

- Shared Hierarchical Academic Research NETwork,
- A consortium of 18 Ontario academic institutions, lead by University of Western Ontario
- Partner of Compute Canada that oversees funding and distribution of equipment.
- Sysadmins and HPC specialist, 20 in total, distributed across 6 institutions.



What does SHARCNET do?

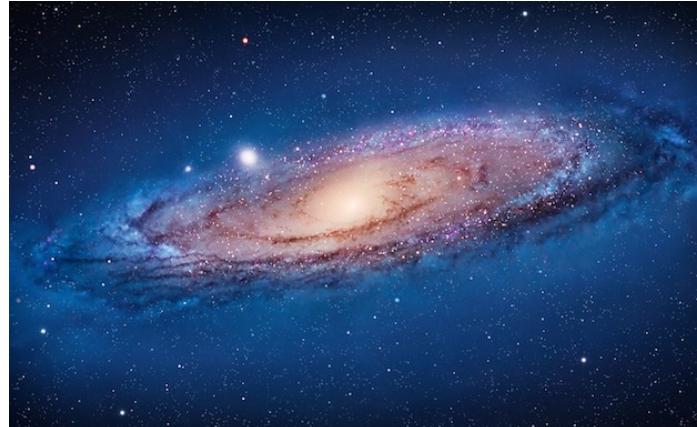
- Provides service and support to all SHARCNET researchers in High Performance Computing.
 - Researchers are part of partner universities across Ontario.
 - Starting to provide service for large data needs:
 - With storage and processing of large data sets
 - Data processing using Spark, Hadoop, etc
 - Data mining and Machine Learning

What is the Andromeda Galaxy?

- Known as M31, or Messier 31
- Spiral galaxy
- 2.5 million light-years
- Closest galaxy
- Bigger galaxy than ours

Why Andromeda galaxy?

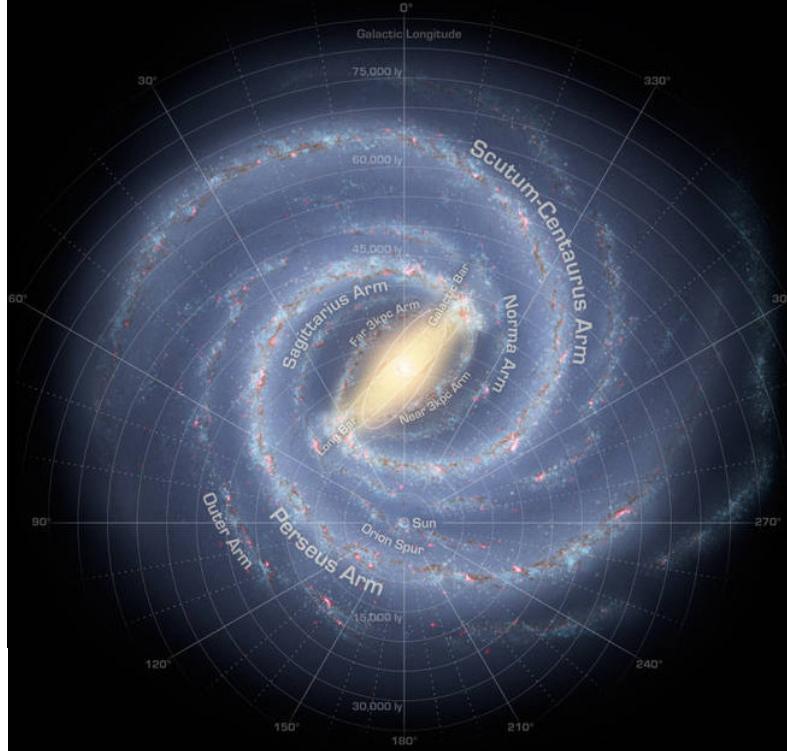
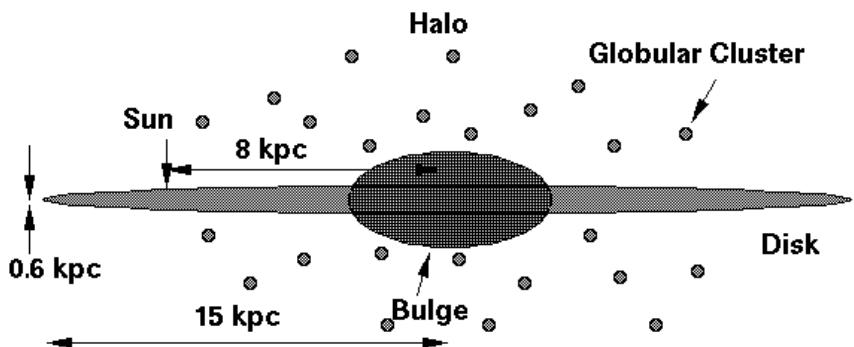
- Cool wallpaper
- t-shirts,
- Mugs ...
- Science?



Andromeda Galaxy in Science

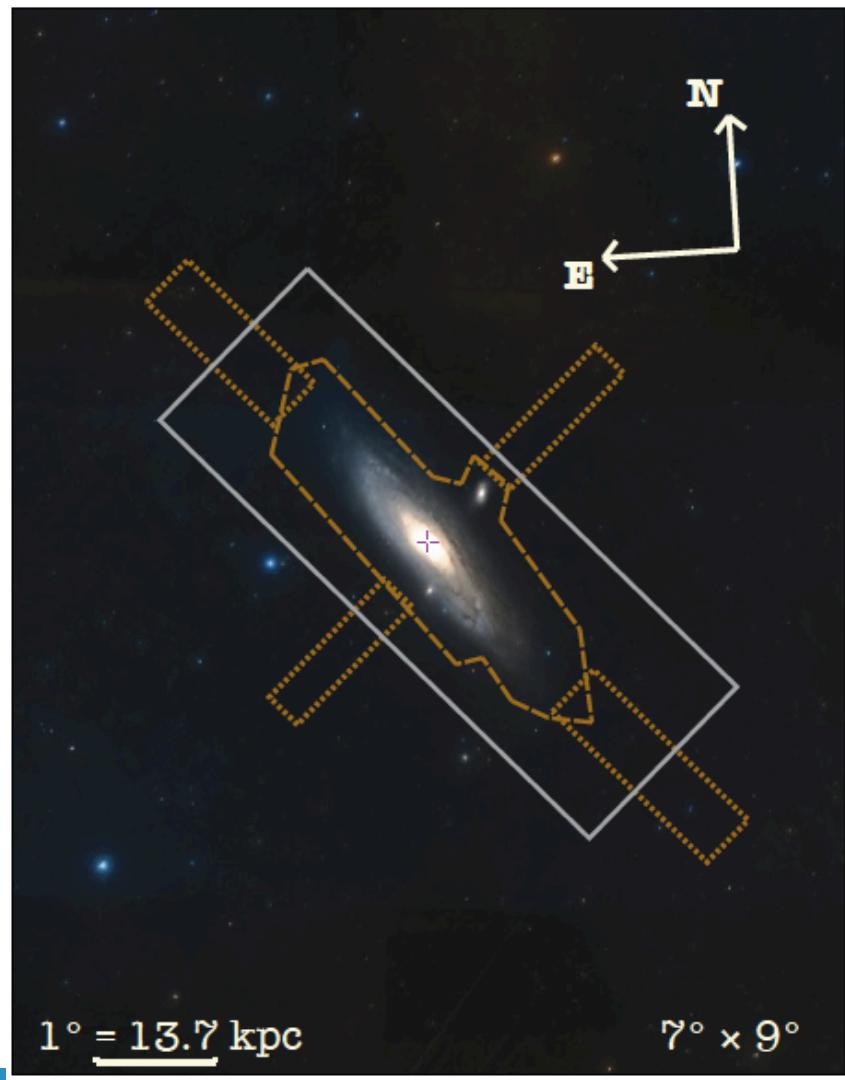
- It has a ~ trillion stars
- 2.5 times longer than our galaxy
- Thought to have merged with another galaxy
- It contains about 26 known black holes
- It can be used as a galaxy laboratory for extragalactic astronomy
- Our galaxy will collide with it (in about 4 billion years)

Why Andromeda?



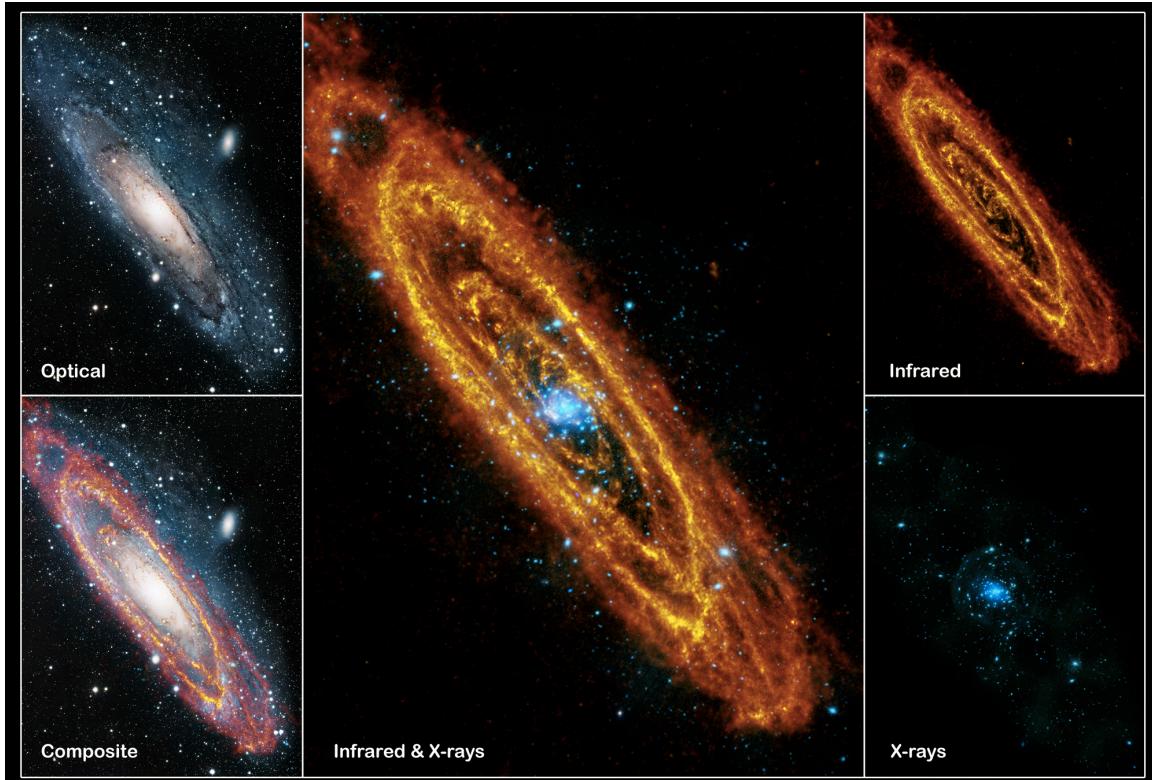
Particularly...

- It has been recognized the extension of Andromeda.
- The area shows the extension of the galaxy, further than thought before.
- M. Rafiei Ravandi et al 2016.



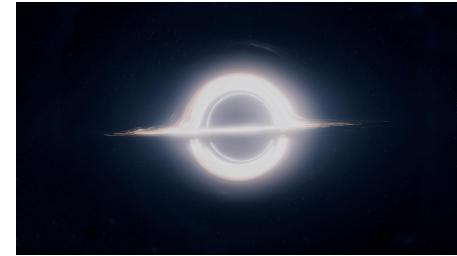
Extended Andromeda

- They were taken from Spitzer-IRAC which is an Infrared telescope.
- It has 426,529 new sources.
- Extends observations for disc and halo.



Classification of these objects

- Do all these sources (426,529) are part of Andromeda?
- Are they all known from previous catalogs?
- What type of object (such as Black holes, galaxies, etc) are those new sources?
- What can we learn from these new objects?

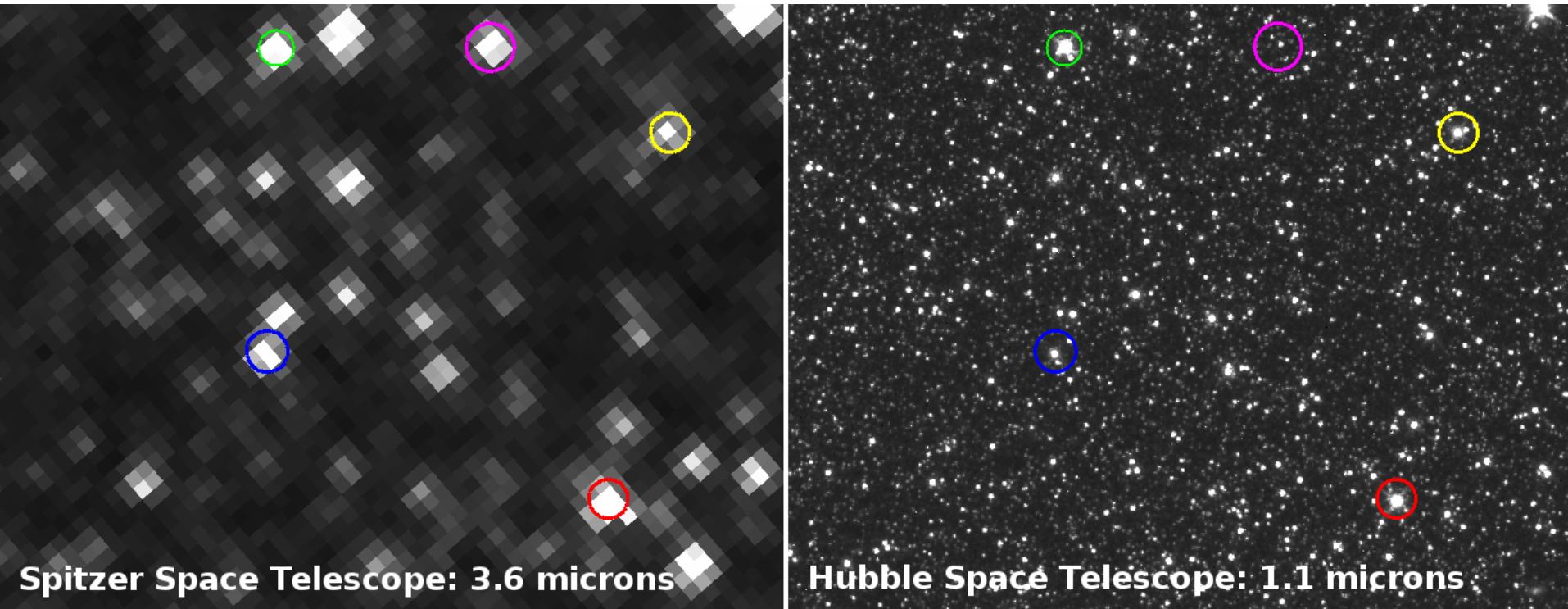


Which catalogs?

- Astronomical databases :
 - SIMBAD (39,022)
 - NED (126,862)
 - MAST (118,854,914)
- Sources only around M31, sources are in different wavelengths (IR, Optical, UV)
- Then compare them with the observed objects.



How hard could it be?



We defined 2" or 2 arcsec as a good match. Arcsec = $1/3600^\circ$, angular measurement, not linear measurement (such as miles/km).

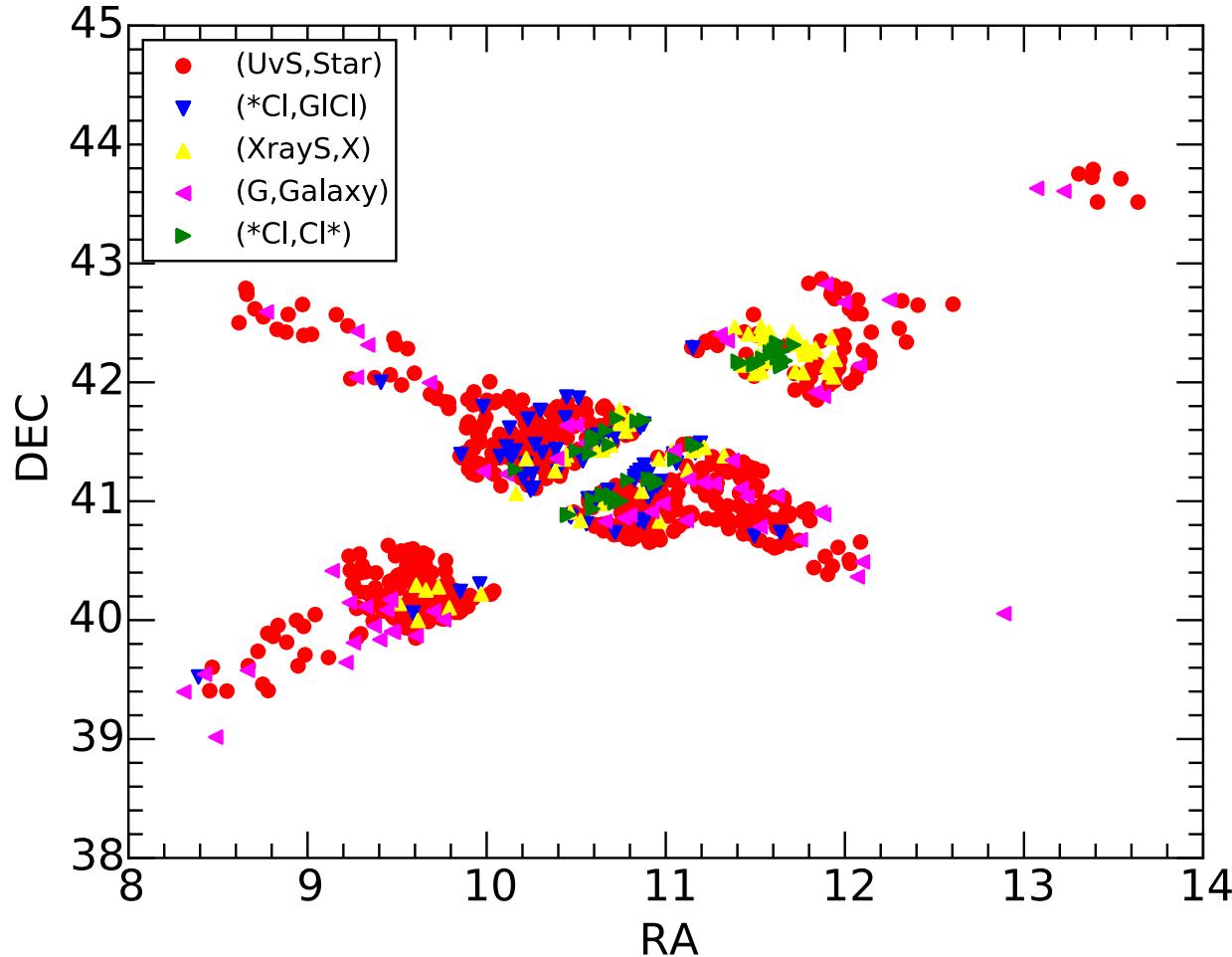


key – value =
 $((RA, DEC), Observations) \rightarrow$
join(), groupByKey(),
filter(), map(), sortByKey()



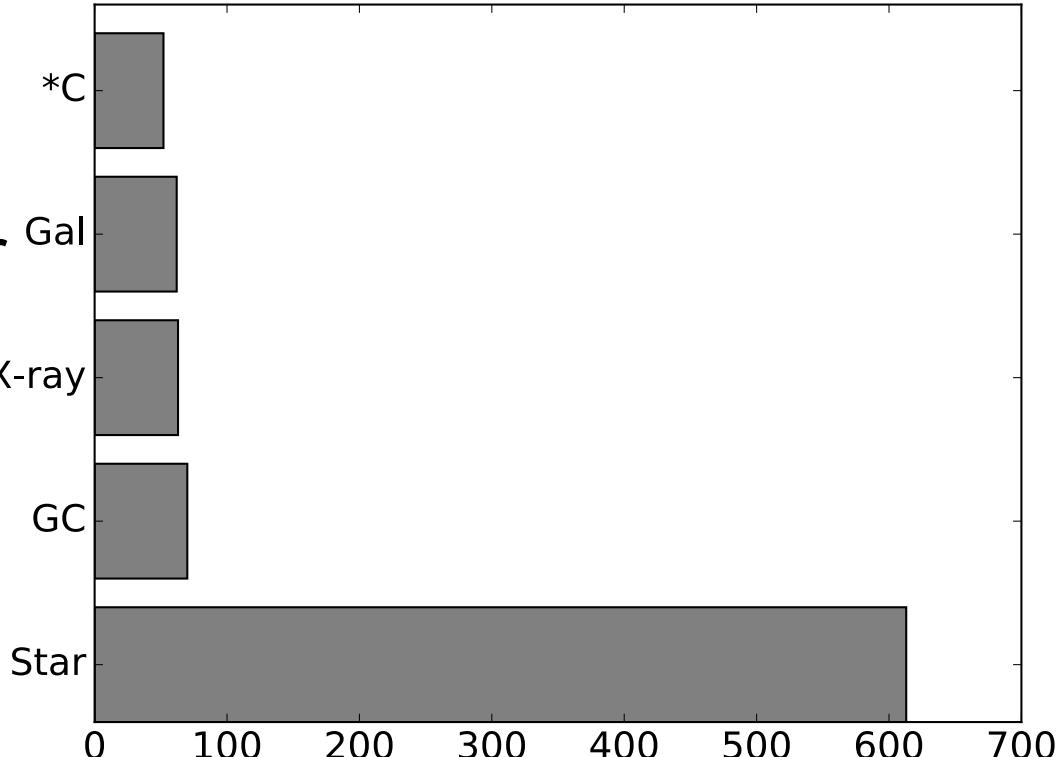
Classification of the IRAC Catalog

NED+SIMBAD+IRAC



Counts?

- 613 Stars
- 70 Globular Cluster
- 63 X-rays sources
- 62 Galaxies
- 52 Star clusters
- Total known sources: 1,391



And the rest?

- They are not part of SIMBAD, NED or MAST
- What about other catalog?
- Can we classify them?
- Can we use machine learning?

Conclusions

- MAST has a higher resolution than IRAC-catalog, SIMBAD and NED.
- Only 1,391 known sources from a matched between NED + SIMBAD + IRAC-catalog.
- The rest could be classified using ML using the known object features in order to give a classification.
- We need more data for a better classification.

Thank You!

Collaborator: Prof. Pauline Barmby, Department of Physics,
University of Western Ontario

Photos:

Mainly from NASA, ESO, EarthSky, MacOS.



**SPARK
SUMMIT
EAST 2017**