Mentor: Juan-Carlos Maureira LSSDS 2017

CatBuilder: from mosaic images to light curves

This project aims at building an small pipeline to transform a set of DECam mosaic images of the same field (taken in different moments) into a set of light curves of each object identified in the field. For this, classical tools will be used: imcopy (or similar) to split mosaic into CCDs, crblaster for cosmic rays removal, sextractor for source extraction and uncalibrated catalog building, and scamp for calibrated catalog building. Once catalogs have been indexed (in a database for example), an object match procedure shall be done in order to obtain for each object a list of tuples (observation date, calibrated magnitud), that also should be indexed for posterior query.

Milestones:

1. Identify the individual procedures involved in the pipeline.
2. Identify best input/output of data required to glue the procedures together.
3. Design an orchestration algorithm to pipeline data through each procedure (fostering parallelism and distribution of tasks).
4. Implement the procedures according the orchestration requirements.
5. Implement the orchestration algorithm in python/bash or any other language suitable for orchestrating jobs in a HPC system.
6. Test the implemented pipeline with a single mosaic image (until calibrated catalog extraction) and evaluate performance.
7. Run a set of mosaic images and evaluate the performance of the whole process.
8. Plot a couple of light curves from your database.

Resources:

Literature:

- Sextractor, CRBlaster and Scamp manuals.
- LDAC catalog format (Sextractor format)
- FITS specs
- SLURM manual (srun, sbatch)
- Python and Bash manuals
- Subprocess, multiprocessing, Spark, SQLite3, astropy, Promise/Future, pyslurm, Celery, etc.

- Building Data Pipelines with Python. Understanding Pipeline Frameworks, Workflow Automation, and Python Toolsets.

Data sets:

- **Dataset will be available in the NLHPC system used for developing and testing the pipeline.**

Software:

- Python, Sextractor, CRBlaster, Scamp.

Computing:

- **NLHPC levque cluster will be available the whole week for developing this project.**

Mentor: Mauricio Cerda LSSDS 2017

Breast Cancer image processing

In clinical conditions, breast samples are treated with specific compounds (HER2), that mark with colors cancer related molecules.

A known pipeline is to start from a large 2D image (30 GB per image) to manually look for cells (blue center), surrounded by HER2 marks (brown), and count the number of cancer positive cells (blue center surrounded by brown) vs total cells [2]. As the % of cancer associated cells increases also the cancer "category" (0 to 3).

Milestones:
- Pre-process images selecting non-blank regions
- Computing Regions of Interest (ROIs) associated to cells
- Test supervised approach for classification of ROIs using SVM and CNN with a pre-trained database
- Compare proposed automatic algorithms with clinical available information (known category classification)

Resources:

Literature:
- Arch Pathol Lab Med. 2007;131(1):18-43.

Data sets:
- IN PREPARATION.

Software:
- Python (PIL, matplotlib), imageMagick, ndpiTools, caffe, libsvm or similar

Computing:
- HPC for parallel image processing, GPUs for caffe training.

Mentor: Ashish Mahabal LSSDS 2017

Light Curve Classification

Sky surveys repeatedly observe large swaths of the sky in one or more filters. With time, the time series of observed fluxes for individual astronomical sources (called light curves) accumulate. Provided here are light curves over several years from the Catalina Real-time Transient Survey for a set of ~50,000 variable objects showing periodicity. There are over a dozen classes. The main goal is to separate the light curves by class. Often statistical features are used do so. Linked below are a couple of sites that list many such features. A harder problem is to design features that separate specific classes. These could be variations on existing features to pick nuanced variations in certain classes. Project participants will be expected to be able to separate the classes. For extra recognition they can try designing features.

Some caveats: Using the brightness of objects as a feature is not a good idea as it reflects more on the (in)completeness of the sample rather than an intrinsic property. Similarly, location/position of objects, when available may not be good 'features' (though astronomers do use distance from the plane of the Galaxy as an indicator based on prior knowledge for many classes – we do not have such "external" information provided here). Since all these sources are periodic, and periods can be a good indicator of certain classes, computing periods is time consuming and fraught with errors, so be extra careful if you are going in that direction.

The classes are not balanced i.e. different classes have different number of representatives. Feel free to ignore classes with only a few hundred sources (ideal Machine Learning training sets are large – larger the better).

Milestones:

Separate at least two classes (and then five classes)
*Design features not listed in the resources below to separate at least two classes better

Resources:

- http://nirgun.caltech.edu:8000/scripts/description.html
- http://isadoranun.github.io/tsfeat/FeaturesDocumentation.html

Literature and Data sets:

Drake et al. 2014 set (`2014,ApJS,213,9 – arXiv:1405.4290`)

Software:

- Python libraries: numpy, sklearn, pandas, matplotlib; jupyter notebooks
- R?

Computing:

- No special needs

Mentor: Rodrigo Assar LSSDS 2017

Breast Cancer PAM50 Panel for genomic study

Microarray expression provides a comparative measure of patient vs healthy subject differences per gene (-infinity to infinity values). Typical microarray provides information about 20.000 variables. The problem is to identify variables with significant values.

A know pipeline is to start from microarray, filter-out variables using for instance a t-test, and then to make k-means with resulting variables. In [1], 50 variables were identified with k=5. These groups were associated with breast cancer sub-types (and later with specific treatments).

Milestones:

- Reproduce PAM50 analysis (US dataset)
- Change the analysis to decision trees / random forest
- For 80 new Chilean patients assign the correct group and compare with available clinical information (directly the known subtype)
- Determine if PAM50 US profiles groups corresponds to Chilean profiles, thus infer if Chilean patients have different genetic behavior or not.

Resources:

Literature:

- [1] J Clin Oncol. 2009 Mar 10; 27(8): 1160–1167.

Data sets:

- IN PREPARATION

Software:

- R, python

Computing:

- Desktop PC / HPC

Mentors: Amelia Bayo & Michael Kuhn LSSDS 2017

Galactic Archeology: the Chemical Evolution of Stellar Populations

Elements heavier than hydrogen and helium (metals) are produced by nuclear reactions in stars and released into the interstellar medium when stars die. New generations of stars form from gas in the interstellar medium, which has been enriched by earlier generations of stars. Low-mass stars have long lives on the main sequence, during which their chemical abundances at the surface remain mostly unchanged, so they can be used to investigate chemical enrichment history of the Galaxy.

Jofré et al. (2017) used a phylogenetic tree to classify stars in the stellar neighborhood based on metal abundance, which is a novel approach to this problem for astronomy. This project aims to confirm these results using other statistical methods and extend the classification to a set of thousands of nearby stars.

Milestones:

- Reproduce the classification of 22 solar twins from Jofré et al. (2017)
- Apply other exploratory methods (e.g., PCA, *k* means, mixture models)
- Apply methods to a set of 2000 stars from Hinkel et al. (2014)
- Examine Galactic kinematics of groups using the Gaia TGAS catalog

Resources:

Literature:

- Jofré, P., Das, P., Bertranpetit, J., & Foley, R. 2017, MNRAS, 467, 1140
- Hinkel, N. R., Timmes, F. X., Young, P. A., et al. 2014, AJ, 148, 54

Data sets:

Sample of 22 FGK type stars with HARPS spectroscopy (Jofré et al. 2017)
Metal abundances of 2000+ stars in the Hypatia Catalog (Hinkel et al. 2014)
TGAS catalog

Software:

- R (or Python)
- MEGA

Computing:

- Desktop PC or laptop