PROJECT WORK

Aim: Data mining and visualization of AGN time-series

Deadline: Dec 23, 2023;

Extracting time-series from the LSST AGN Data Challenge

Rubin Observatory LSST

Vera C. Rubin Observatory is a brand new facility that will conduct a ten-year survey of the Southern Hemisphere sky, referred to as the Legacy Survey of Space and Time (LSST) with the goal of answering some of scientists' biggest questions about the Universe (see https://rubinobservatory.org/ for more details). Rubin Observatory LSST project will Rubin Observatory will survey the visible night sky every night for ten years, building a 500 PB database of images and a 15 PB catalog of text data describing properties of nearly 40 billion individual stars and galaxies, among which millions of active galactic nuclei (AGN).

The LSST AGN Science Collaboration has initiated in 2021 the AGN Data Challenge aiming to start planning for the AGN science with the Rubin Observatory LSST, focusing on 1) parameterization of AGN light curves, 2) AGN selection, and 3) AGN photo-z. The data challenge description can be found at https://richardsgroup.github.io/AGN_DataChallenge/, and the data sets have been made publicly available on Zenodo (https://zenodo.org/records/6878414).

AGN Data Challenge - Dataset

The dataset released in this data challenge are pulled from different source (public archive) and put together to mimic the future LSST data release catalogs as much as possible. The column names and units used for different measurements (e.g., flux) also follow that listed in the LSST Data Products Definitions Document (DPDD, https://docushare.lsst.org/docushare/dsweb/Get/LSE-163, see Section 4.3 for more details on the LSST data release catalogs, and also Ivezić et al. 2019).

The AGN Data Challenge objects included in the release dataset are drawn from two main survey fields, an extended Stripe 82 area and the XMM-LSS region, consist of stars, quasars/AGNs and galaxies. The actual class labels (when available) are: Star, Gal (for galaxy), Qso (for quasar), highZQso (for QSOs at high redshift), Agn (for AGN).

As stated, the input data were modified to comply with the DPDD standards for data release catalogs (DPDD, Section 4.3). We refer to distinct astrophysical bodies that emit light detected as “Objects” and individual instances (detection) of those objects as “Sources”. Observations from a specific point in time will appear in the Source tables in the data releases, while “co-added” (averaging/summing over time) information will appear in the *Object* tables. So-called “light curves” (brightness as a function of time) will appear in the *ForcedSource* tables (with summary statistics in the *Object* tables). Within the Data Challenge, the total number of objects (both combined) in the *Object* table is ~440,000, and total

number epochs in the *ForcedSource* table: ~5M. For more details on the data structure see Getting Started page https://github.com/RichardsGroup/AGN_DataChallenge/tree/main/getting_started and Savic et al. (2023).

Project Work

In astronomy a time-series typically represent a time-magnitude array called light-curves, in which "magnitude" can be any measure of object brightness, such as flux, luminosity, magnitude (for more details see materials from course *Practical analysis of noisy and uneven time series*). Within this project we will work with the light curves (time-series) of active galactic nuclei (AGN). AGN are among most powerful energy sources in the universe with the actively fueled super-massive black hole, residing in the center of a galaxy.

The main aim of the project work is data mining of AGN time-series within AGN Data Challenge, their characterization and visualization. The list of tasks is the following:

- Familiarize with the data set (and used data formats, e.g. parquet files), identify most relevant data structure and features.
- Write the procedure for the extraction of N-number light curves in each photometric band.
- Develop a procedure for effective/cleaver visualisation of N-number of light curves (e.g. using Datashader). Try optimizing the procedure for the large number N (note: N should be at least 100, larger numbers are preferred).
- Make a procedure which for each light curve provides basic info-sheet with the following parameters: number of points, number of gaps (think of a metric to identify gaps), mean and median sampling, maximal/minimal/mean/median magnitude, fractional variability F_var.
- Make distribution plots for all measured properties.
- *Extra task*: think of a possible metric to detect outliers.

Provide the report in a form of a jupyter notebook with all the codes and plots, supported with detailed description of all steps. Please upload the notebook on your github account.

REFERENCES and HELP:

1. AGN Data Challenge Data sets, https://zenodo.org/records/6878414
2. Info on AGN Data Challenge https://richardsgroup.github.io/AGN_DataChallenge/
3. Getting started, https://github.com/RichardsGroup/AGN_DataChallenge/tree/main/getting_started
4. Example notebook, https://github.com/cefeida42/agndc-contrib/
5. Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111, LSST: From Science Drivers to Reference Design and Anticipated Data Products, http://doi.org/10.3847/1538-4357/ab042c
6. Savic et al. 2023, ApJ, 953, 138S, The LSST AGN Data Challenge: Selection Methods, https://ui.adsabs.harvard.edu/abs/2023ApJ...953..138S/abstract