Dataiku x ALMA Challenge - Brief

Thanks for your interest in the Dataiku x ALMA Challenge!

The **ALMA Observatory** is a nonprofit organization looking at the oldest and coldest things in the Universe from a high altitude plateau in Chile.

As Dataiku wanted to provide users with an opportunity to grow their skills, in a typical collaborative manner, ALMA tasked us with helping them automatize quality tests performed on their observations, before processing them for scientific discoveries (like... the black hole image).

Read on for more info on the challenge, data, and the plan to tackle it together.

Document outline:

Objectives

1. Develop your Dataiku & data science skills, collaboratively!
2. Learn from fellow users, the Dataiku Neurons, and in-house data scientists.
3. Make an impact on an exciting, for-good initiative.

Project Description: Flux Measurement Quality Assurance for ALMA Calibrators

Everything we know about the Universe, how old it is, how far are the nearest or most distant galaxies, what is the temperature of our Sun, or other stars… **every piece of information we have comes from the light we receive from the sky**. Light coming from stars, entire galaxies, quasi-stellar objects (“quasars”), and other elements that have been traveling for millions of years across the Universe to get to us. And the light we receive is even brighter than what our eyes can see.

There are more types of light (traveling at different frequencies) than the visible light humans can perceive. That is why we have so many telescopes in the world. To have different sets of eyes specialized to see different types of light. So no piece of information coming from the Universe is left behind. **One of these types of light travels in the form of radio waves**, coming from the coldest parts of the Universe, and galaxies, where stars and planets are being born.

Here is where the **ALMA Observatory** comes in. Observing light in the form of radio waves, to answer questions about our cosmic origins: how do stars like our Sun are born? How are planets like our own formed, and can that happen in other places of the galaxy? Or other galaxies? Can molecules essential to life on Earth be produced in other parts of the Universe? Where? And how?

The ALMA Observatory is a nonprofit organization using Dataiku DSS since 2017. The Observatory is an aperture synthesis telescope consisting of 66 antennas located on the

Chajnantor Plateau, 5,000 meters altitude in northern Chile, and it is recognized as "**the most complex astronomical observatory ever built on Earth**" (more information). ALMA operates over a broad range of observing frequencies of light between 80 and 900 GHz (for more context, here's a recommended talk from Data Analyst Ignacio Toledo).

But independently of the kind of astronomical observatory and frequency studied, all observations of astronomical sources in the sky require the use of instruments to collect and record the information in a digital format. **The information recorded in this way will need to be processed (or "reduced" in astronomy jargon)** in order to provide data that can be used for scientific research and analysis.

The data processing (or reduction) is a process that consumes human and computing time: in the case of ALMA, the processing time can be in the order of days or even weeks in some extreme cases.

Sadly, **the instruments used to make the observations and collect the information from the sky are not free from sporadic problems and bugs**: hardware components might fail, software systems might crash or be buggy, power supplies temporarily fail, etc. Furthermore, even when all the instruments are working OK, the weather conditions can also negatively affect the data acquisition process.

If these problems are not detected before the data reductions start, a double negative effect arises:

1. A significant amount of processing resources are wasted,
2. The observation might not be repeated (for example, the object to be observed is no longer in the right position on the sky, or the required conditions are never met again).

To minimize the probability of this happening, **a quality check named Quality Assurance level Zero (QA0) is performed before sending the data to be processed.** To do this, the observatory uses a huge number of "metadata", or contextual information, related to the observations (or executions) that can be analysed in a short time lapse - in the order of minutes to hours. While certain algorithms help provide recommendations, most of the process is handled manually by an Astronomer on Duty (AoD) who analyzes the metadata with certain criteria to determine the QA of the data: Pass, Semipass, or Fail.

This challenge is about using the contextual information, or metadata available, to find out if ML or AI methods could be used to automatise as much as possible the QA0 decision. This would result in freeing up AoD time and making the decision in the minimum time possible, so that all resources are leveraged to produce useful observations.

About the Data

Because ALMA has several observing modes with their own particularities, this challenge will focus on a subset of observations that has the advantage of being both homogeneous and publicly available: **the observation of Flux Calibrator Sources.**

As mentioned above, all data obtained by the observatory must be reduced. Part of this process is related to the "calibration" of the information acquired: the data is systematically affected by the characteristics of the instruments being used and the observing conditions at the time, and the calibration process takes care of removing all these effects so the data acquired at ALMA and at a particular time can be later compared with data acquire in other observatories and time periods.

One of the calibrations deals with the "absolute" brightness (or flux) of the astronomical sources observed. When ALMA (or any other observatory) measures the brightness or flux of a source, the brightness is measured in arbitrary units. To compare observations, sources that have a known brightness are used, such as quasi-stellar objects ("quasars") and Solar System members.

Dedicated observations are performed on these objects for this purpose. They are compiled in the ALMA Calibrator Source Catalogue (SC), which constitutes the basis of the project conducted with the Neurons.

DSS Project

The ALMA staff prepared a DSS project with all the data made available for the challenge.

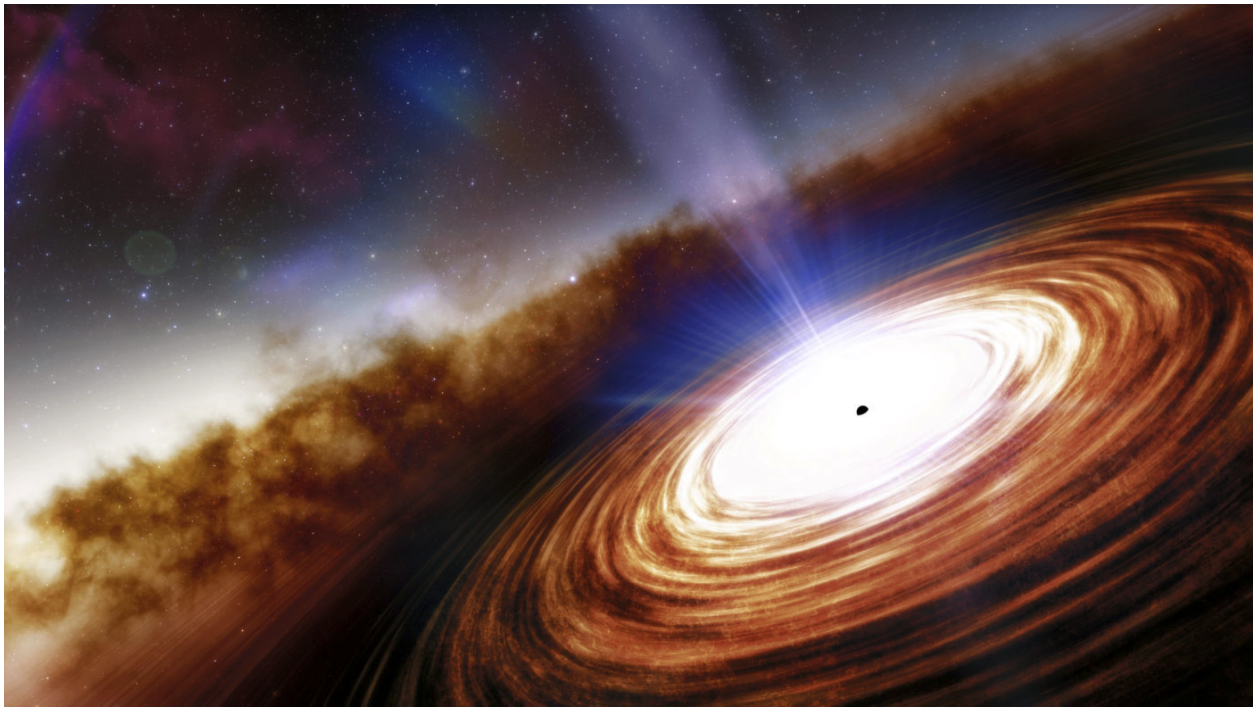
The notebook named 'ALMA SC Challenge_ Description' contains information on what data is stored in the Science Calibrator Database, how ALMA observes, and what kind of metadata or contextual information is available after an observation is done.

A recommended starting point is the Source Catalogue summary table (SC_SUMMARY_TABLE), which gathers information on the observed quasar sources since 2017. Ideally, all observations for which QA0 decision is deemed "pass" should be ingested into the Source Catalogue ("true").

Recommended Process

1. Understand the structure of an Execution (or observation),
2. Identify the metadata of each Execution and how it is related to the observed sources,

3. Find out the most relevant parameters or variables determining the quality of the Execution. Successful Executions are eventually ingested into the Calibrator Source Catalogue.



Participants

- **You & 10 fellow Dataiku users** from around the world - all levels of data science experience are welcome, we all learn from each other!
- **Dataiku Neurons**, including Ignacio Toledo (Data Analyst, ALMA Observatory) who enabled this project.
- **Dataiku data scientists:** Matthieu Scordia (Data Science Lead, APAC) & Darien Mitchell-Tontar (Data Scientist, Denver).
- **ALMA Observatory staff members**: Rosita Hormann (Software Engineer), Jorge Garcia (Science Archive Content Manager), Celia Verdugo (Data Analyst).

Collaboration Tools

- **Dataiku DSS!** Link and credentials to access the shared instance will be provided during the kickoff event (April 15th).

- **Slack**: you will be invited to a dedicated Slack channel for questions and real-time interactions.

Timeline

- **End of June -** Kickoff event to get to know fellow participants, discover the project and data, ask questions, and start rolling up our sleeves :-)
- **First week of July** - Choose a segment to focus on in smaller groups, based on your interests and skills.
- **Every 2 weeks** - Bi-weekly syncs to share updates on progress between groups, insights you've discovered, and next steps.
- **End of August June** - Target end for the Challenge!

Involvement

Participation in the challenge is voluntary and depends on your interest & bandwidth.

A minimum of one hour per week on average is recommended to check in on progress and make a contribution to advance the project. Should life happen and your overall participation will be adversely affected, just let us know :-)

We look forward to tackling this challenge together!