3rd URSI AT-AP-RASC, Gran Canaria, 29 May – 3 June 2022

Artificially Intelligent Pipeline Sequencer

Brian M Kirk*(1,2), Urvashi Rau(1), and Ramyaa(2)
(1) National Radio Astronomy Observatory, New Mexico, USA e-mail: bkirk@nrao.edu; rurvashi@nrao.edu
(2) New Mexico Institute of Mining and Technology, New Mexico, USA; e-mail: ramyaa.ramyaa@nmt.edu

Data taken with a radio interferometer must pass through a sequence of calibration and analysis stages in order to produce images that an astronomer can use to study astrophysical phenomena. Processing stages include data editing, calibration of the instrument hardware, and image reconstruction through deconvolution. Steps within each of these stages may be assembled in a variety of sequences and combinations, referred to as 'recipes'. Any single end-to-end data analysis recipe may be thought of as one possible path through all possible configurations. Overall, there is a predictable sequence of stages, but each stage may contain several optional paths depending on the characteristics of the outputs from the previous stage, and if certain conditions are not satisfied some stages (or steps) are revisited as subsequent steps, thus forming loops.

Current automated pipelines essentially hard-code a "few", most-used paths through this entire maze, with a small set of heuristics-based decision points. While pipelines do remove a significant amount of burden from end users by processing 80% of the data volume, they have yet to capture the full complexity of the problem and achieve true science ready results for most projects (i.e. outputs that require no further customized tuning and which may directly be used for astrophysical interpretation). Most current pipeline executions still need to be reviewed by a trained Data Analyst for quality assurance as there is often no objective set of metrics to define correctness that applies in all situations. Data Analysts may need to optimize individual steps or repeat entire sequences after the pipeline has executed before admitting data to enter the archive for use.

The full configuration space of the software tools is quite vast. In addition to sequencing stages and steps, most steps also need to decide parameters based on characteristics of the input data passed to that method. The process of creating a complete data analysis recipe is the subject of training workshops where students and scientists are taught the principles of each analysis step, how to run and configure software that implements each step, and how to decide the precise data-driven sequence of steps (and their parameters) required to optimally analyse a given dataset. It is this learning process, *without specific instruction*, that we are attempting to automate.

Our research looks into changing methodologies from a prescribed sequence of instructions to a sequence generated from the input data itself. This approach requires a software framework where instructions can be learned. We're starting this research with model based Reinforcement Learning for planning and Genetic Algorithms for optimization. These techniques have the ability to evaluate actions with respect to global results and optimize the preceding calibration and imaging steps accordingly. When scaled, such frameworks are expected to learn across the breadth of our existing problem space and generalize the knowledge for inference in our environment of numerically-continuous states. In comparison to a hardcoded recipe that needs to be tuned by a Data Analyst after execution, our vision is that this generalized knowledge will produce an individualized recipe optimal to the dataset provided.

We have set out a roadmap of how we are building and testing these methods. To start, we have identified simplified scenarios we can simulate that reflect the larger problems we want to address. With these simulations, we have begun to test the applicability of the new methodologies to our problem set while keeping interpretation of results clear; interpretability is a fundamentally important characteristic in our approach. We will report on our research roadmap and current progress.