Vicente Amado Olivo

Computational Meta-Research in Astrophysics: Optimizing Resource Allocation through Machine Learning

Problem

- Tracking the use of scientific facilities in published literature is an important metric for success
- Example: The use of archive Space Telescope data is currently identified manually in literature requiring **10** scientists, every two weeks
- A mundane task requiring skilled scientists to step away from their research
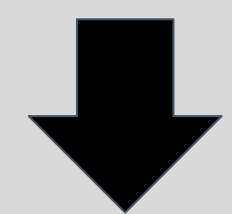
Solution

- To build an algorithm that automates the identification of the use of Space Telescopes in the stream of literature.
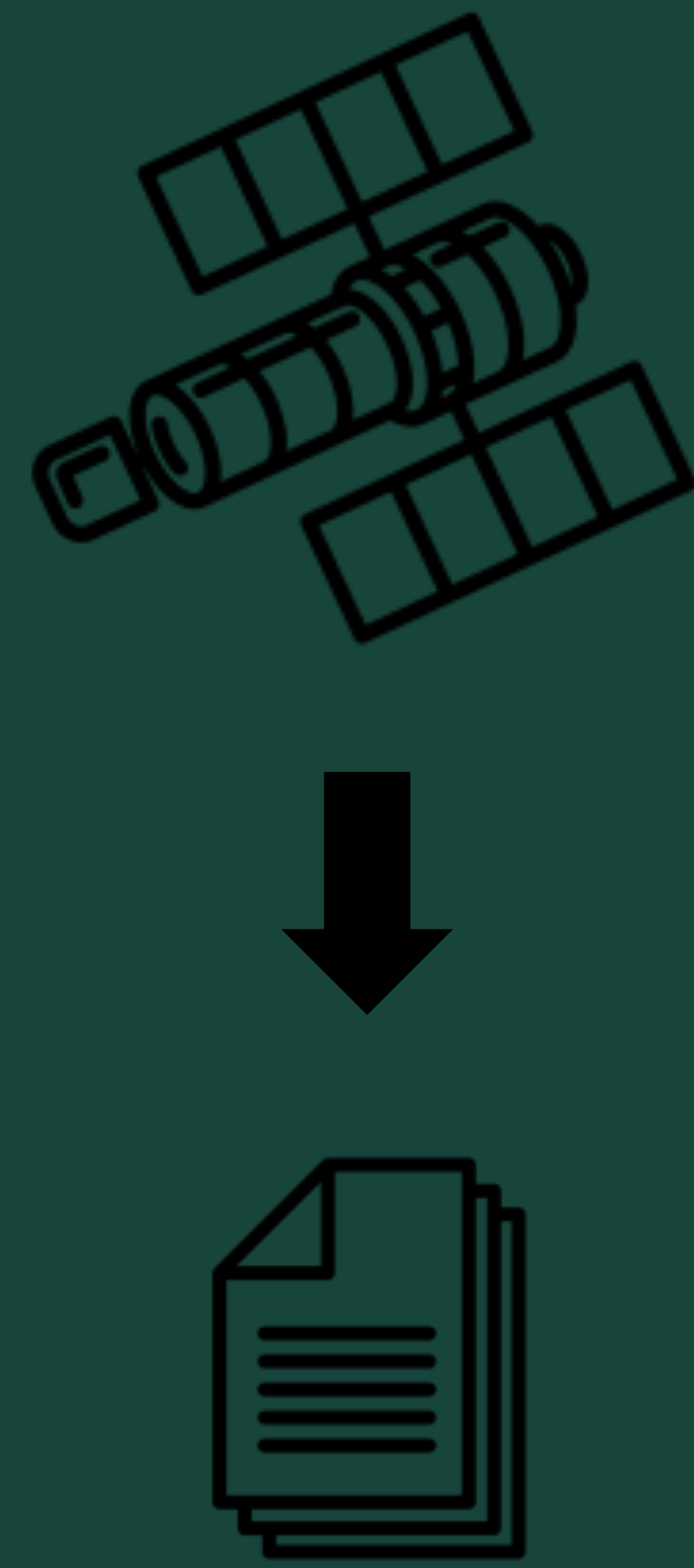
Methods

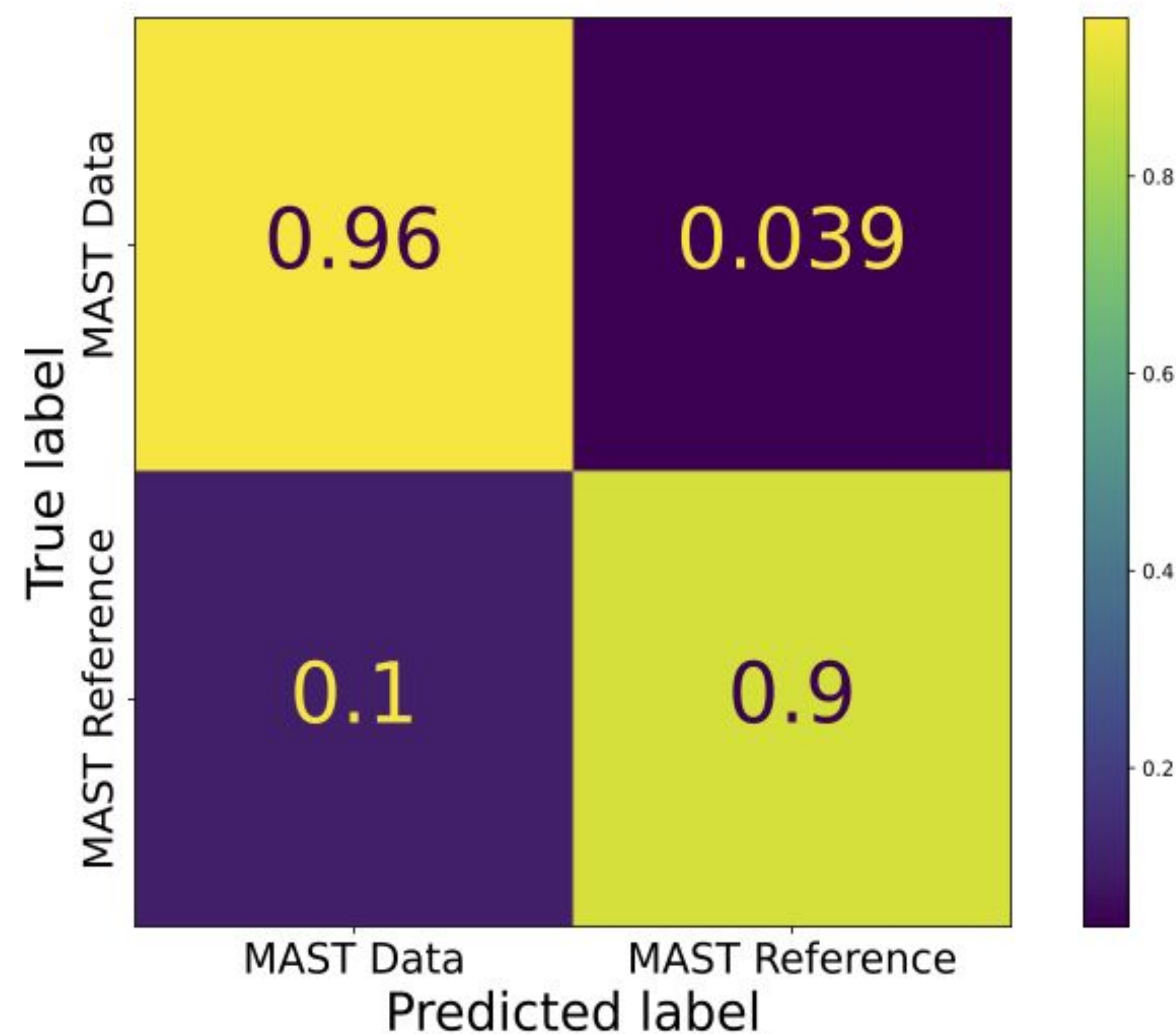Collection of an initial **14,808** full-text astronomy publications
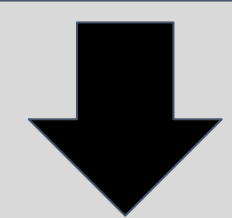
Conversion of full-text to machine-understable format, testing two different techniques

Classification of machine-understandable documents using three different classifiers

The results show that a focused approach on only parts of the text, such as sentences with predetermined keywords, performs better than full-text algorithms. The best algorithm combination is TF-IDF with a Support Vector Machine classifier with an accuracy **0.93**.

CMSE