Predicting computational reproducibility of scientific pipelines using collaborative filtering

Soudabeh Barghi, Lalet Scaria, Tristan Glatard
Department of Computer Science and Software Engineering
Concordia University, Montreal, Quebec, Canada
first.last@concordia.ca
* These authors have contributed equally

***Abstract*—**

I. Introduction

Computational reproducibility, the property through which computational results can be recomputed over time and space [**?**], has become a critical component of scientific methodology with the rise of the reproducibility crisis in several domains [**?**]. Among the factors hampering computational reproducibility, infrastructural characteristics such as the operating system play an important role. In neurosciences, our primary field of interest, several studies have shown the effect of the operating system on computational results. However, conducting such reproducibility studies at scale is cumbersome due to the execution time of data analysis pipelines, which easily exceeds a few hours.

In this paper we investigate approximate methods to predict the reproducibility of a computational analysis from the first files that it produces. Our main intuition is that reproducibility errors are caused by a reduced number of factors that originate in the analysis pipeline and input data.

II. Method

A. Collaborative filtering using ALS

B. Sampling the Training Set

III. Dataset

IV. Results

V. Discussion

VI. Conclusion

Acknowledgment