Reproducibility Report for ACM SIGMOD 2022 Paper: "DataPrism: Exposing Disconnect between Data and Systems"

ABDUL WASAY, Intel Labs, USA

We verified all trends presented in the DataPrism paper, i.e., DataPrism requires the least number of interventions to identify data problems that lead to system failure compared to other state-of-the-art tools. The code to reproduce experiments from the paper is well-documented and easy to use.

1 INTRODUCTION

We successfully reproduced all experiments from DataPrism [1]. The authors made all artifacts available, along with clear instructions on how to obtain all experimental results presented in the paper. In this document, we summarize the reproducibility process.

2 SUBMISSION

The authors provided a code repository, instructions, and data sources to run and reproduce graphs from the paper:

- Code repository: https://github.com/sainyam/DataPrism
- Instructions: https://github.com/sainyam/DataPrism/blob/main/README.md
- Data: https://drive.google.com/file/d/1syQhwIRwdWJBqqT0mJQWfN9LGsZOylnV/

The instructions are easy to follow. The scripts are self-contained and produce Figures 6, 7, 8, and 9 from the paper. These Figures are sufficient to verify the paper's claim.

3 HARDWARE AND SOFTWARE ENVIRONMENT

The paper's contributions are at an algorithmic level, and as such, it does not rely on a specific hardware environment. The paper does not provide any detail about the hardware used. All software packages needed are provided as a requirements.txt file that can be used to create a Conda environment identical to that used by the authors. Table 1 summarizes the hardware we used to reproduce the results:

Table 1. Hardware & Software environment

	Repro Review
CPU	Intel(R) Xeon(R) Platinum 8272CL
cores	2
GHz	2593 MHz
RAM	8GB
Storage	HDD

4 REPRODUCIBILITY EVALUATION

4.1 Process

The reproducibility process went smoothly. At first, there were some issues with dependencies and experimental scripts, but the authors fixed those issues promptly. Once we had access to the fixed codebase, reproducing results took less than 24 hours.

4.2 Results

All experiments from the paper were reproduced (Figures 6, 7, 8, and 9). Overall, these figures. The key trends in Figures 6, 7, and 8 that we verify are: (i) DataPrism requires the least number of interventions compared with BugDoc, Anchors, and GrpTest, (ii) Anchors require the maximum number of interventions, and (iii) GrpTest is the second best but does not work for all cases. Figure 9 is an experiment that studies the effect of the threshold parameter on the number of interventions required. The trend in the reproduced figure matches the one presented in the paper.

5 SUMMARY

Overall, we were able to reproduce the findings of Dataprism. The code base is easy to use and comes with clear instructions. We appreciate the authors' cooperation during the reproducibility process.

REFERENCES

[1] Sainyam Galhotra, Anna Fariha, Raoni Lourenço, Juliana Freire, Alexandra Meliou, and Divesh Srivastava. 2022. DataPrism: Exposing Disconnect between Data and Systems. In *Proceedings of the 2022 International Conference on Management of Data* (Philadelphia, PA, USA). 15 pages.