A Reproducible Research Pipeline

Using Git and Data Version Control (dvc)

Lukas Erhard

2025-02-03

University of Stuttgart, CSS Lab

Problem 1

How to write code collaboratively

Having multiple people work on code

Being able to work on code with multiple people has major advantages

- The research is way faster
- The code has less bugs
- It keeps the research reproducible

Having multiple people work on code

Being able to work on code with multiple people has major advantages

- The research is way faster
- The code has less bugs
- It keeps the research reproducible

BUT WHAT ABOUT PYTHON VS R?????

Problem 2

Keeping track of the data

aka “oh no, my data is too big for git”

Introduction to dvc as storage

TODO: What is dvc?

Problem 3

Ensure the order of execution

aka “Why are my results from today different from yesterday?”

Using the dvc DAG

TODO: Explain dvc.yaml and the DAG