*The R Project for Statistical Computing*

Google Summer of Code Proposal

# CI Optimization for R Package Performance Testing

Sagnik Mandal

January 25, 2025 v2.0

# Contents

# 1. Project Info

**Project title:** Optimizing a performance testing workflow by reusing minified R package versions between CI runs

**Short Title:** CI Optimization for R Package Performance Testing

**Idea Page:** Idea Description on the R GSoC Wiki

# 2. Bio

I am Sagnik Mandal, a sophomore pursuing an Integrated Dual Degree in Materials Science and Technology at IIT (BHU), Varanasi. I have been a self-taught programmer since over 6 years now, and have also contributed to various open source projects, and used to maintain a couple of packages on the Arch Linux User Repository (AUR).

I first learnt about R, when I was following a online course on "Dealing with materials data : collection, analysis and interpretation", this course discussed data analysis using R for materials science. I have also been working under a professor at my institute on a project that involves data analysis and machine learning for material science applications.

I have used Github Actions in some of my personal projects, last year I applied to Joplin for GSoC to work on a similar project where I had planned to use Github Actions to automatically fetch build Joplin Plugins, along with preventing malicious code from being executed. While the project was shortlisted by Joplin, I couldn't make it to the final list of selected students, but this experience has helped me understand the working of Github Actions.

Other information about me can be found on my resume.

## 3. Contact Information

**Name:** Sagnik Mandal

**Postal Address:** 503, Satish Dhawan Hostel, IIT (BHU), Varanasi, Uttar Pradesh, India, 221005 (Timezone: UTC+5.5°)

**Telephone(s):** +91-7470989815°

**Email(s):** sagnik.mandal.mst23@iitbhu.ac.in°, acriticalcynic@outlook.com°

**Other communications channels:** Google Meet, Zoom°, Discord°, WhatsApp°

## 4. Affiliation

**Institution:** Indian Institute of Technology (Banaras Hindu University), Varanasi, India

**Program:** B.Tech. & M.Tech. (Integrated Dual Degree) in Materials Science and Technology

**Stage of Completion:** Sophomore, expected graduation in 2028

**Contact to Verify:** Dr. Chandan Upadhyay° (*email: cupadhyay.mst@iitbhu.ac.in°*)

## 5. Schedule Conflicts

I have my Summer Break from May 10 - July 10, 2025, so for majority of the project duration, I will be able to dedicate 35 hours a week to the project. I have already setup my build environment and through the tests I have also tackled parts of the project.

The first and the last week of the project will be a bit tight for me, as I will be travelling back home and then back to college, but I will try to make up for the lost time by working extra hours during the rest of the project duration.

## 6. Mentors

**Evaluating Mentor:** Anirban Chetia (anirban166) (*email: ac4743@nau.edu°*)

**Co-Mentor(s)**: Toby Dylan Hocking (tdhock) (*email: toby.hocking@r-project.org°*)

**Contact with Mentors:** I have been in touch with both Anirban and Toby over Email and Github.

# 7. Coding Plan and Methods

## 7.1. Project Scope

The project will involve these three repositories:

1. data.table
   - `.ci/atime/tests.R` - This file contains benchmark tests for various data.table functions, including performance regression tests across different versions. Currently there are about 15 test cases, and for each test case, the workflow installs multiple versions of data.table before running the tests. In total, the workflow builds and **installs approximately 28 different versions of data.table** for each run.
   - `.github/workflows/performance-tests.yml` - This workflow triggers the `Autocomment-atime-results` action to run performance tests on pull requests.

2. Autocomment-atime-results
   - `action.yml` - This is the main file that defines the GitHub Action. It sets up the R environment, installs required packages, and runs the tests defined in .ci/atime/tests.R using the atime package. After running tests, it logs execution time and comments on the PR with performance results.

3. atime
   - `R/versions.R` - This file contains the actual functions which build and install different versions of the package. The functions include `atime_versions_install` (installs different git versions of a package with modified names), `pkg.edit.default` (modifies package files to enable installation of multiple versions), and `atime_versions_exprs` (creates benchmark expressions with appropriate package references).

## 7.2. Broad Tasks

1. **Package Minification**: Identify areas for optimization in the package installation process. Partially implemented in the tests, this script will extract the package tarball, remove unnecessary files and directories, and install the package using `R CMD INSTALL`.
2. **Updating Atime to use cached packages**: Modify the `atime` package to use the cached packages. This will involve updating the `atime_versions_install` function to check for the presence of the cached package and install it if found.
3. **Artifact Caching & Retrieval Workflow**: Update the `Autocomment-atime-results` workflow to check for the availability of the cached packages. If found, download and install them directly; if not, rebuild the minified versions and upload them as artifacts for future runs.
4. **Support for PRs from Forks**: Adapt the workflow to securely handle PRs from forks. This will ensure repository secrets don't get leaked due to malicious actors, as well as allowing us to test all PRs, rather than just those from the maintainers.
5. **Testing and Documentation**: Test the workflow with different versions of data.table and document the gains in terms of time and resources saved. Also, document the workflow to make it easier for contributors to understand and use.

## 7.3. Documentation

# 8. Timeline

**Total Hours:** 175 (35 hours/week × 5 weeks) **Project Duration:** June 2 - July 14, 2025

## 8.1. Pre-GSoC Period ( $CURRENT_DATE - May 8, 2025 )

- Repository setup with mirrored repositories for the project:
  - `data.table` with historical branches
  - `Autocomment-atime-results` fork
  - Local `atime` development environment
- Initial minification script prototype.
- Audit current resource use and CI runtimes [done partially]

## 8.2. Community Bonding Period ( May 8 - June 1, 2025 )

- Won't be available for the first week of the community bonding period, as I will be travelling back home.
- Discuss project expectations and goals with mentors, and community members. Finalize the project plan and timeline.
- Start working on the project to get a headstart.

## 8.3. Week 1: Core Minification System (June 2-8)

**Objective:** Implement reliable package minification

- Finalize file exclusion list through empirical size analysis of 10 historical versions
- Develop versioned artifact naming convention

## 8.4. Week 2: Caching Integration (June 9-15)

**Objective:** Connect atime to cached artifacts

- Modify `atime_versions_install` to:
  - Check for cached packages before source build
  - Implement SHA-256 verification for cached packages
- Implement cache fallback mechanism

## 8.5. Coding Period Week 3 (June 16 - June 22)

- TODO: Primary Focus: `Updating Atime to use cached packages` & `Artifact Caching & Retrieval`

## 8.6. Coding Period Week 4 (June 23 - June 29)

- TODO: Primary Focus: `Workflow Integration` & `Support for PRs from Forks`

## 8.7. Coding Period Week 5 (June 30 - July 6)

- TODO: Primary Focus: `Testing and Documentation`

## 8.8. Final Week (July 7 - July 14)

# 9. Management of Coding Project

## 9.1. Communication Strategy

- Bi-weekly sync calls with mentors (Google Meet/Zoom etc.) to track progress and discuss blockers.
- Regular updates will be added on a weekly blog which I will maintain. This will also act as the report for the project.

## 9.2. Testing

- Add unit tests for the minification script and the caching workflow.
- Test the workflow with different versions of data.table to ensure that the cached packages are being used correctly.

# 10. Test Submissions

**EASY**: Script to Minify R Package°

- Wrote a script that is agnostic to the package name and version, and can be used to minify any R package tarball. It also installs the minified package using `R CMD INSTALL`, to ensure that the package is installable.

**MEDIUM**: Github Action to Minify R Packages and Upload as Artifact°

- Created a GitHub Action that reads the package name and version from the issue description, checks if the package is already minified, and if not, minifies the package and uploads it as an artifact. The action also installs the minified package using `R CMD INSTALL`.
- It then comments on the issue with package size details, and time taken to minify the package.

**HARD**: Supporting PRs from Forks in Autocomment-atime-results°

- Modified the `Autocomment-atime-results` workflow to support PRs from forks. The workflow is divided into two parts: one that runs the tests and uploads the results as artifacts, and another that downloads the results and comments on the PR with the results. The workflow has been tested to work with PRs from forks and non-forks.
- Then cloned the data.table repository, with all its historical branches required for the atime results, and tested the workflow to work with PRs from forks and non-forks.