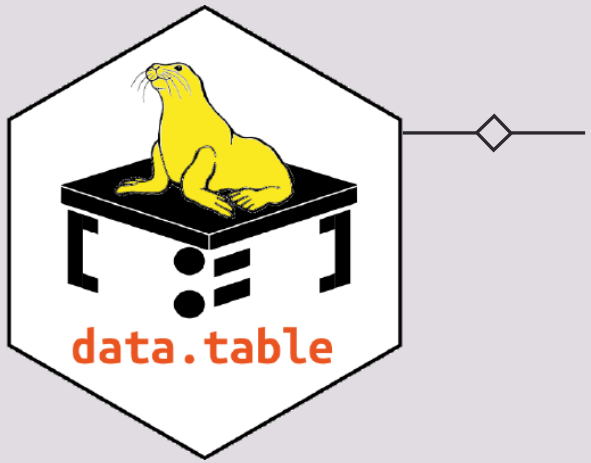
Title:

Benchmarking Performance for the data.table Package

AUTHOR: DORIS AMOAKOHENE

SUPERVISOR: *TOBY HOCKING*

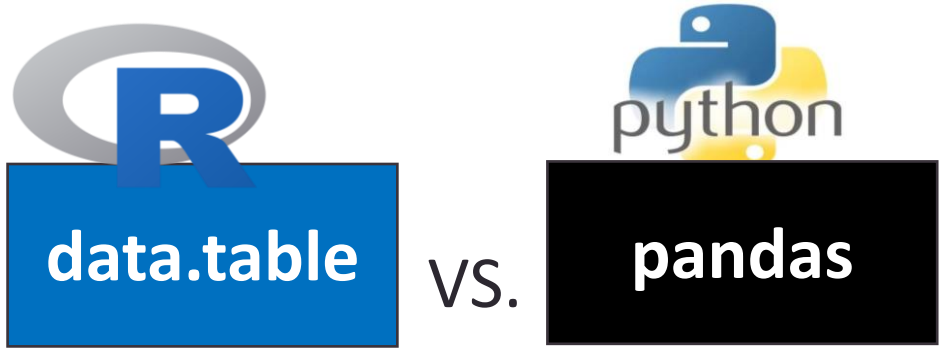
CO-WORKER: *ANIRBAN CHETIA*

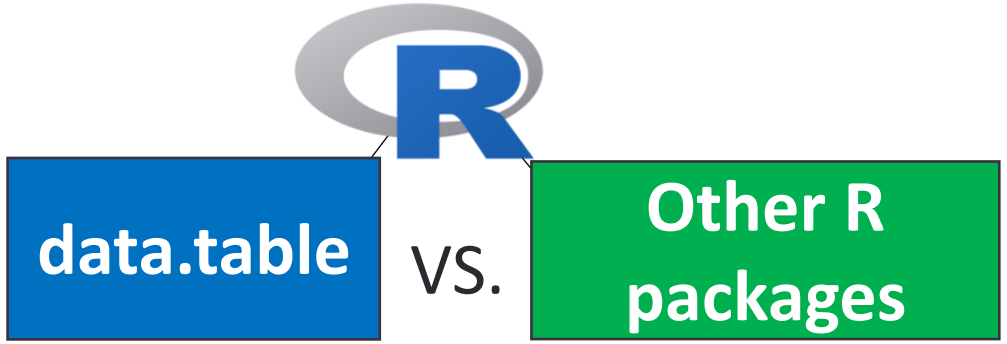


OUTLINE OF PRESENTATION

1: Introduction

- **Comparative Benchmarking**: Comparing data.table to other packages in R and python that perform similar tasks





- **Performance Testing**: We evaluate the performance of different versions of the data.table package by benchmarking their memory and time usage.

1: Introduction

- Benchmarking packages like airspeed velocity, conbench, touchstone, and pytest-benchmarks rely on a single data size N for benchmarking.

- We utilize the atime package which allows for a sequence of N values and generates a visual plot.

1: Introduction

- The **‘atime’** package in R is designed for asymptotic timing, enabling the comparison of time and memory usage across various R code as a function of the input size, denoted by ‘N’.

- It also has a built in GitHub action which was developed by Anirban Chetia, my co worker.

➢ 2. Visualizing the atime result of various functions with data.table

```
atime::atime(
N=10^seq(1,20),
setup={
    ...
},
"data.table::fwrite" = {
    data.table::fwrite()
  },
"pandas::to_csv" = {
    reticulate::py_run_string()
    }
  )
```

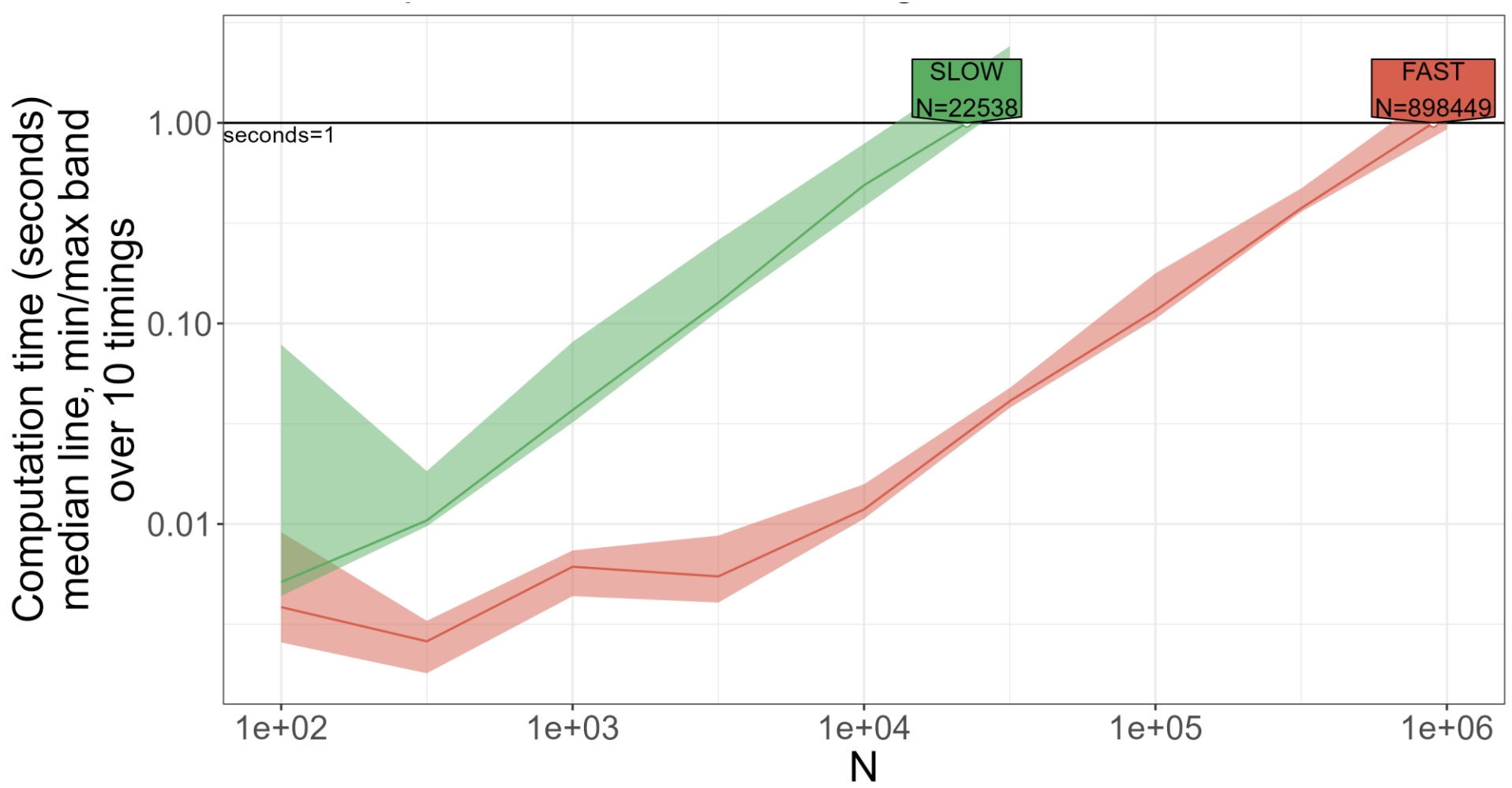
atime::atime function for comparative benchmarking

data size to vary over

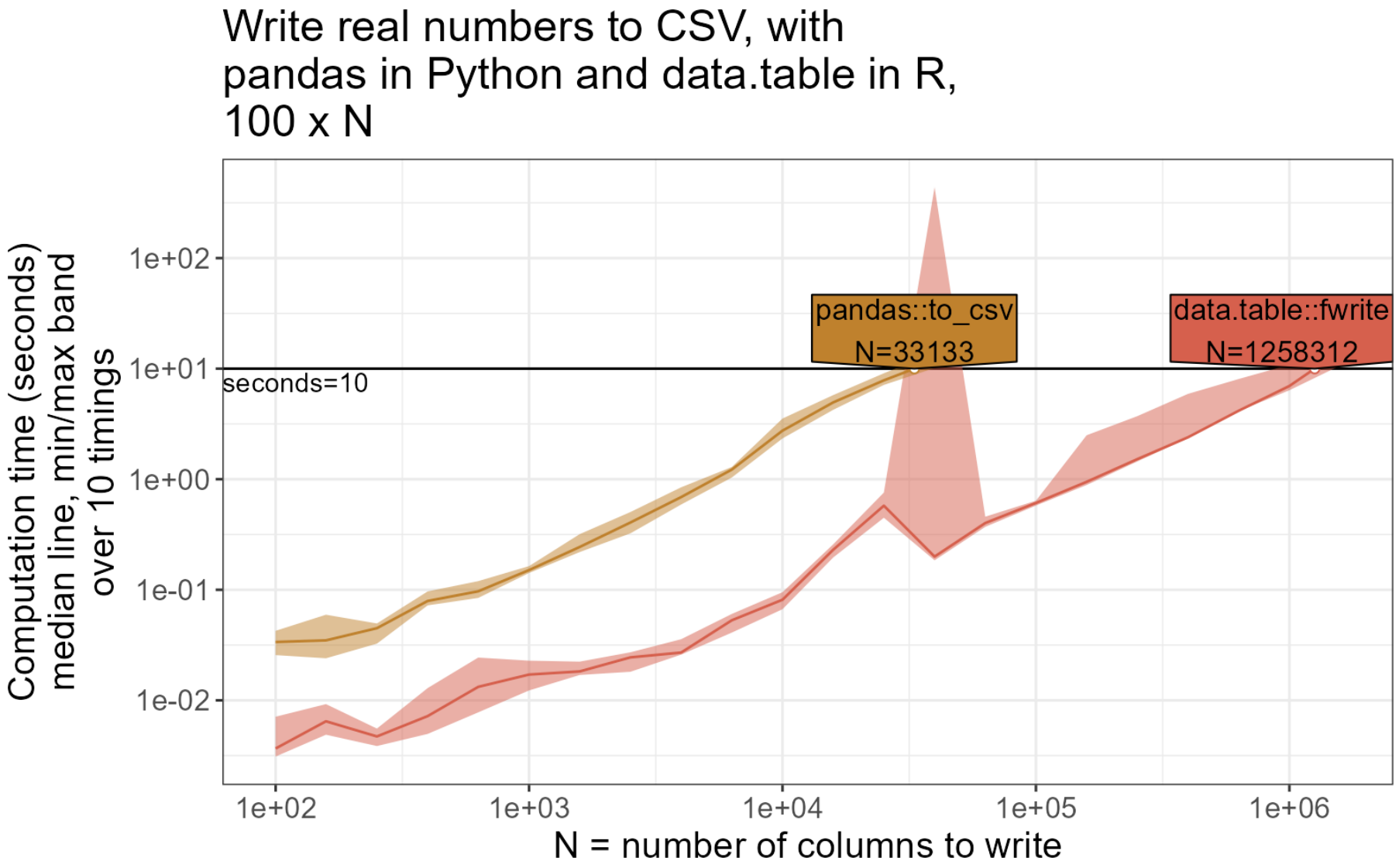
expression to evaluate for every data size, before timings.

named list of expressions to time.

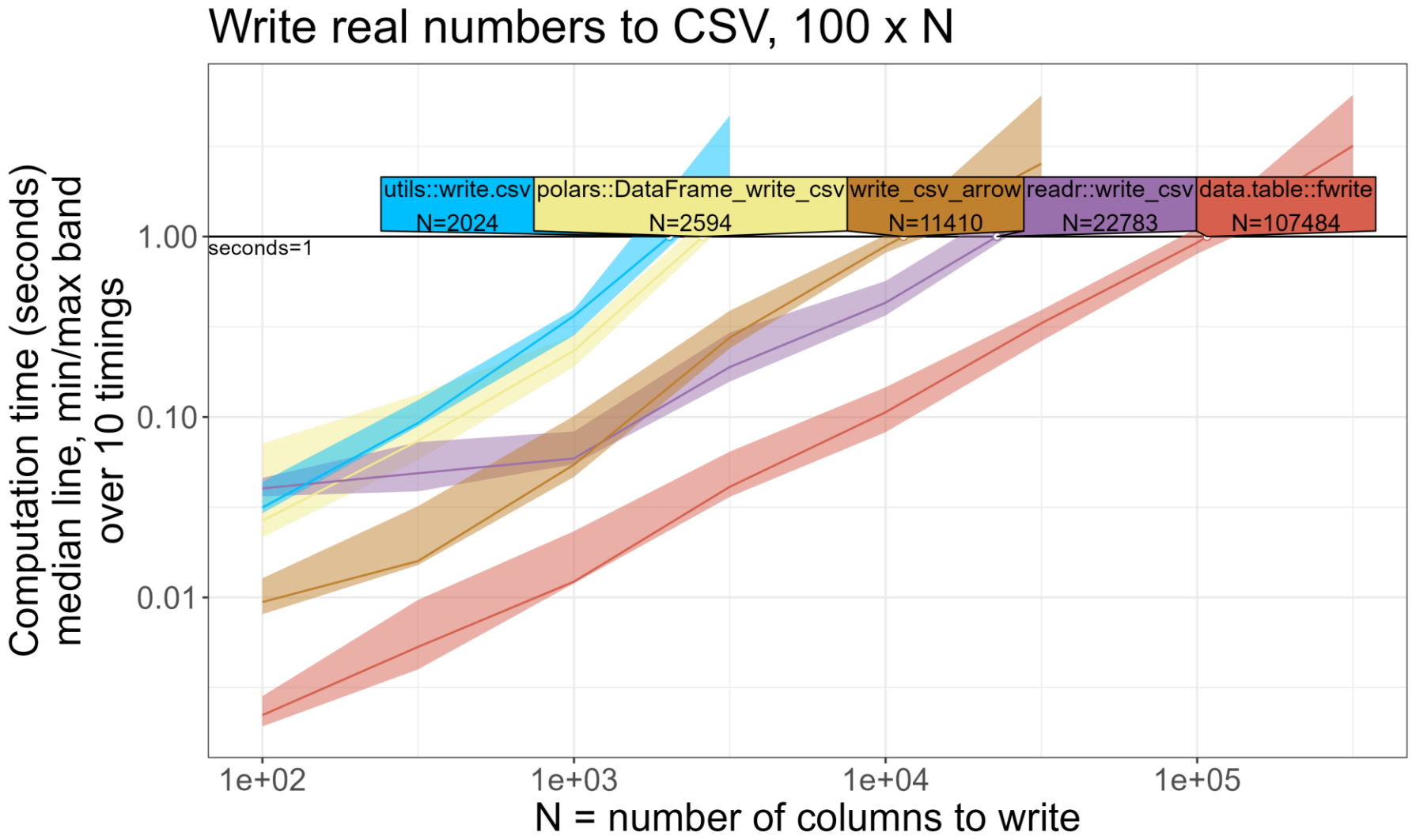
2: Comparing the performance of two functions to illustrate fast and slow execution



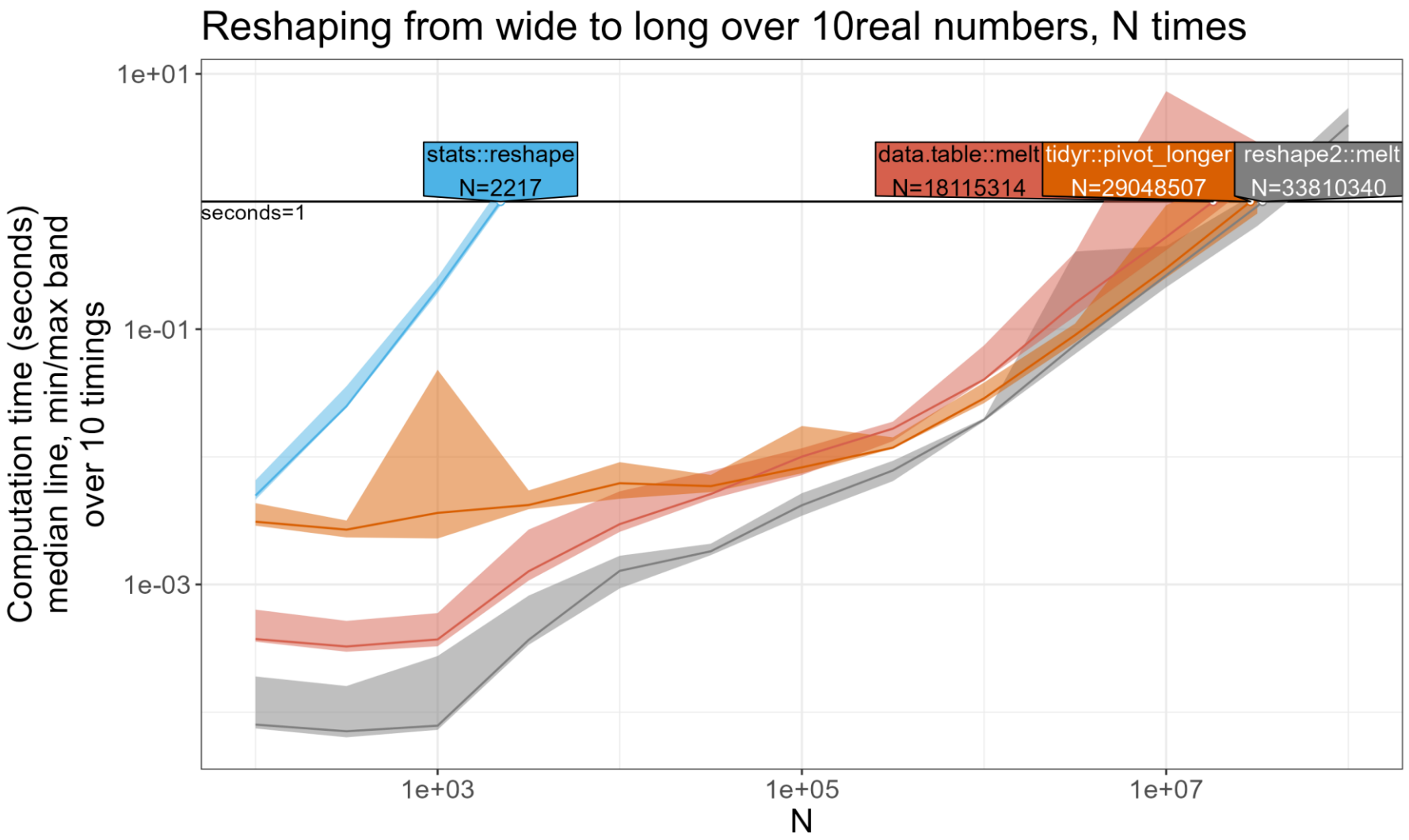
➢ 2. Benchmarking data.table::fwrite against pandas::to_csv



2: Benchmarking data.table::fwrite against other R packages (readr , polars, utils) for csv writing



2: Benchmarking data.table::melt against other R packages (reshape, tidyr, reshape2) for wide-to-long data reshaping



Data.table isn't always fastest, which is why we use benchmarking to identify optimization opportunities and improve performance

➢ 3. Visualizing the atime result for performance regression

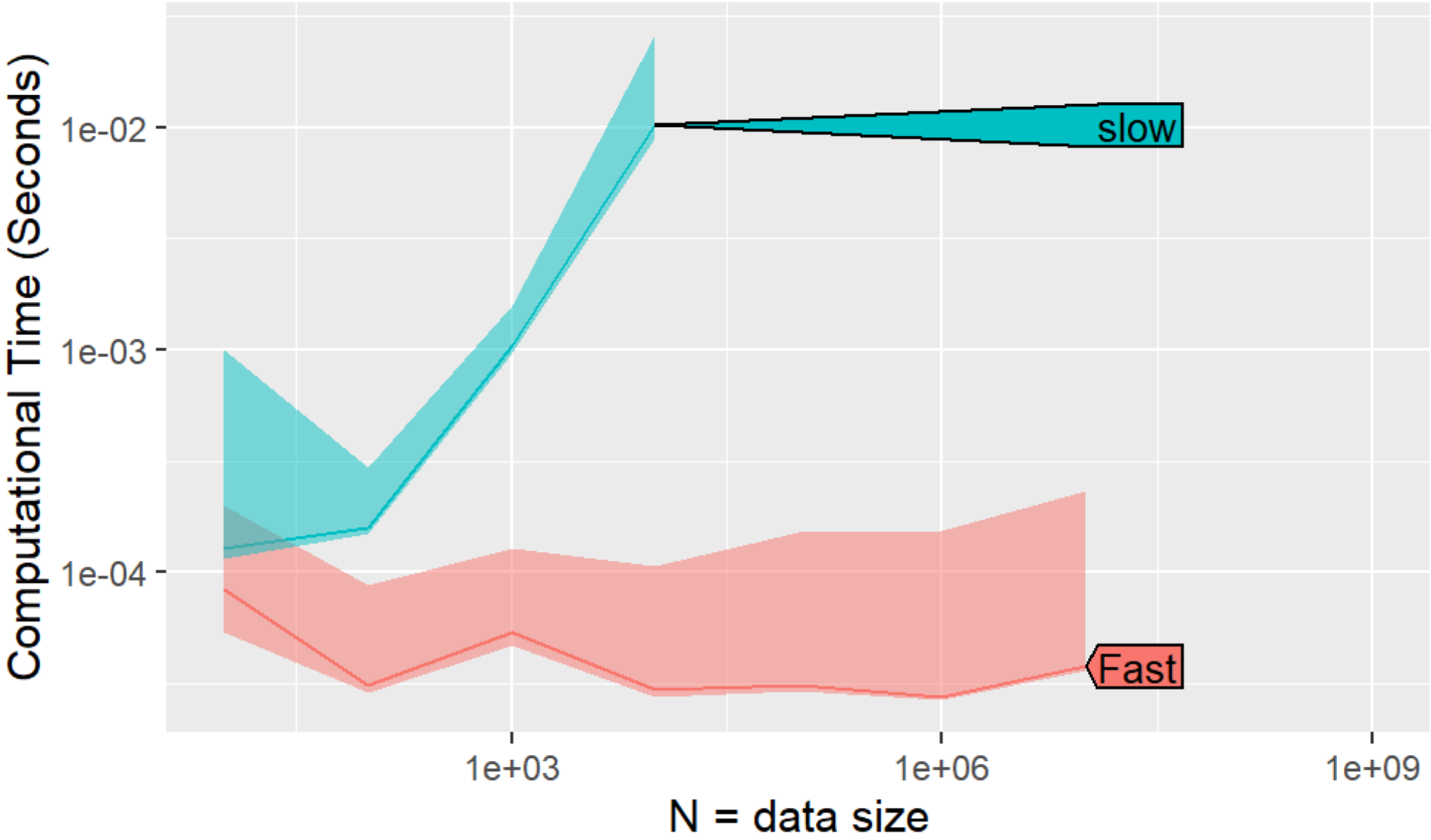
```
atime::atime_versions(
pkg.path = "~/data.table",
pkg.edit.fun = pkg.edit.fun,
N = 10^seq(1,20)
setup = {
...
},
expr=data.table:::`[.data.table`(...),
  "slow"="15f0598b9828d3af2eb8ddc9b38e0356f42afe4f",
  "fast"="6f360be0b2a6cf425f6df751ca9a99ec5d35ed93"
  )
```

atime_versions, atime function for performance testing over different version of an R package

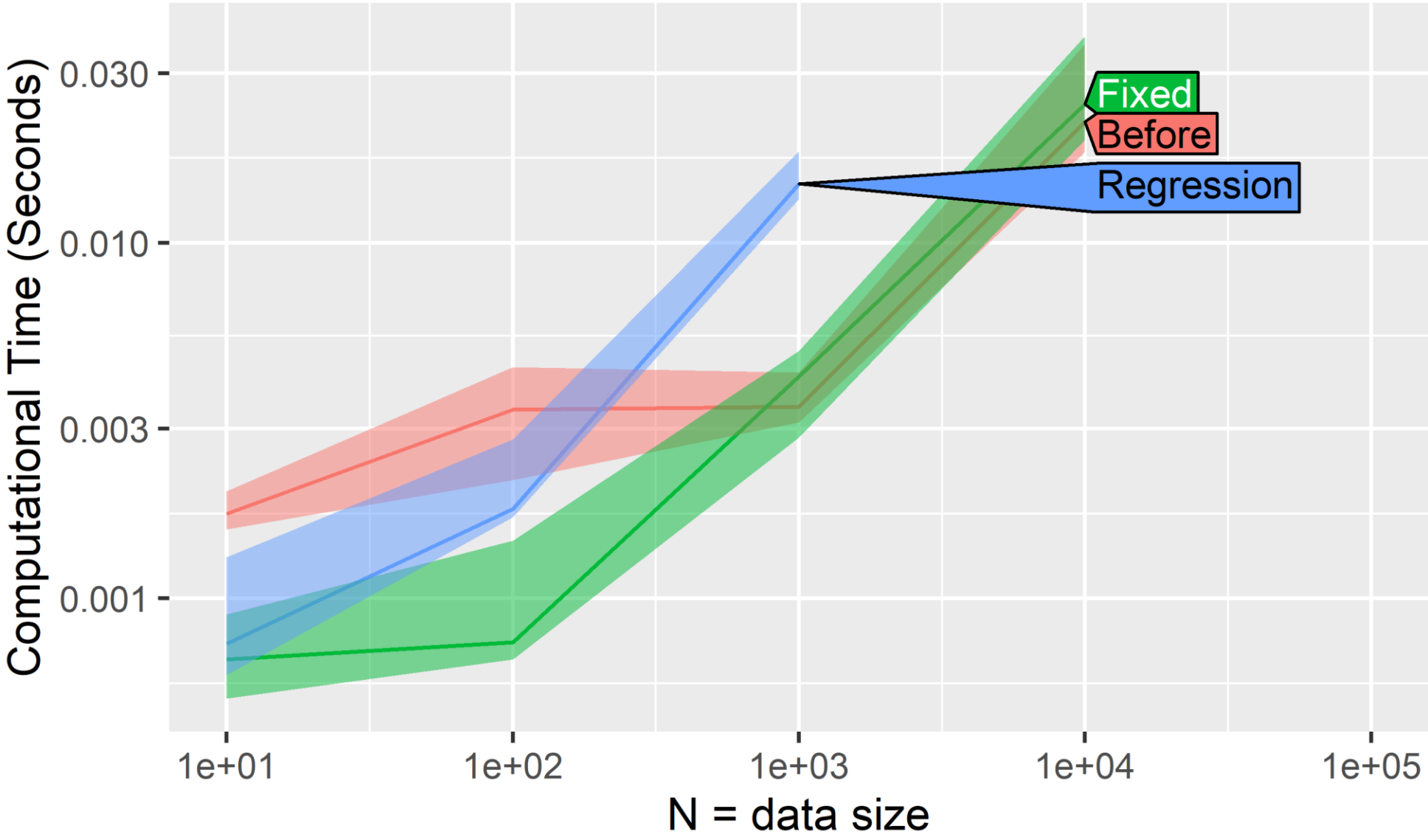
Path to git clone of repo containing R package(data.table).

function called to edit package before installation

3: setDT extremely slow for very wide input #5427



3: groupby with dogroups (R expression) performance regression #4200



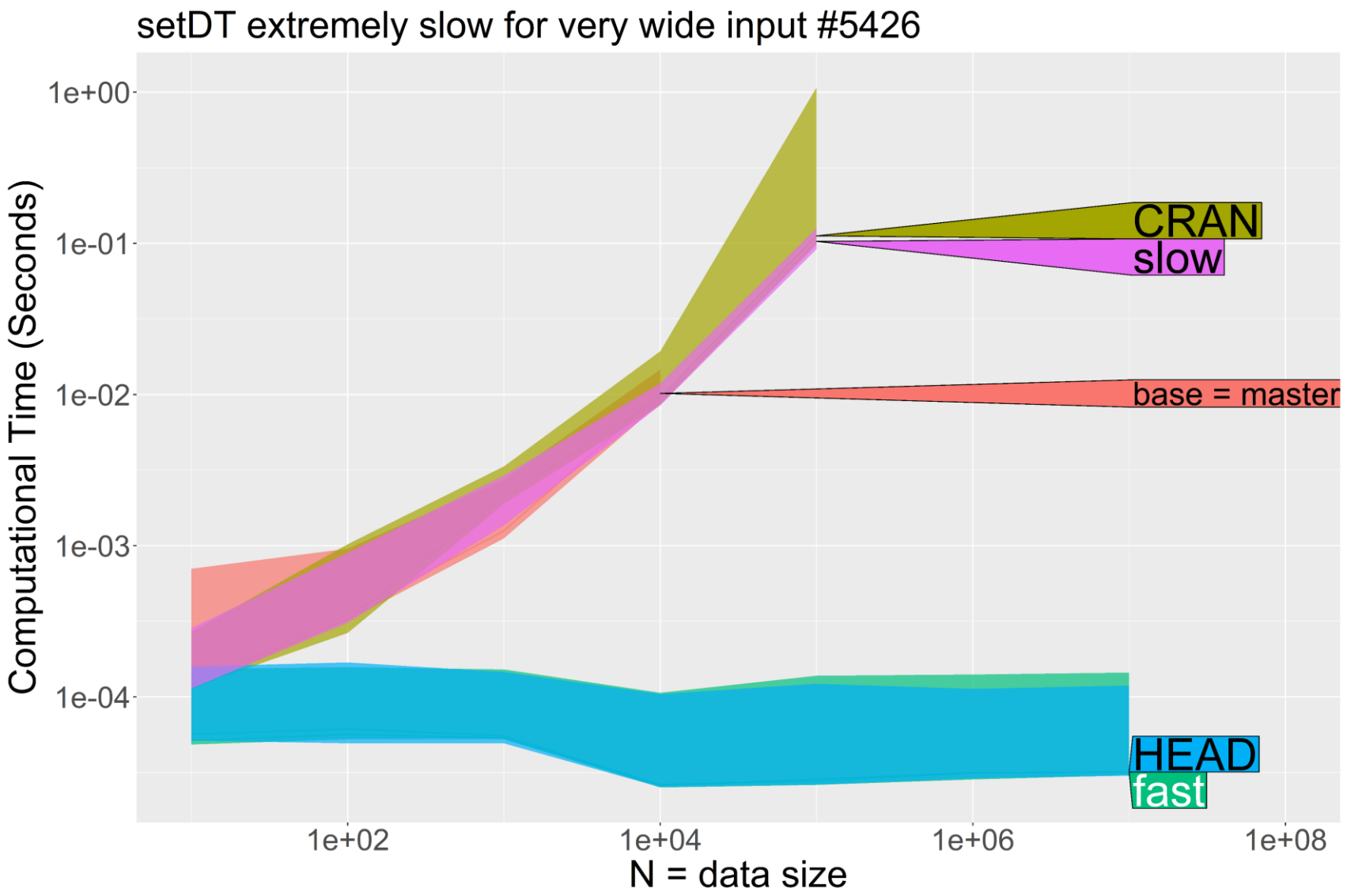
3: GitHub Action

- **A GitHub Action** is an automated workflow that can be set up in a GitHub repository to perform various tasks such as building, testing, and deploying software

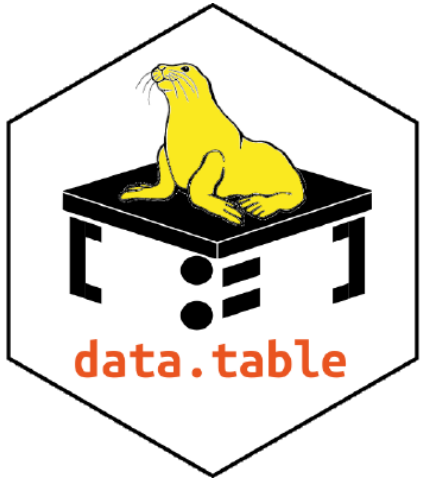
- It helps to identify and address any performance issues, ensuring that the package is performing well.

- Our GitHub Action for R packages runs **atime::atime_pkg** and **comments the generated results on pull requests to help identify potential performance regressions** introduced from the incoming changes.

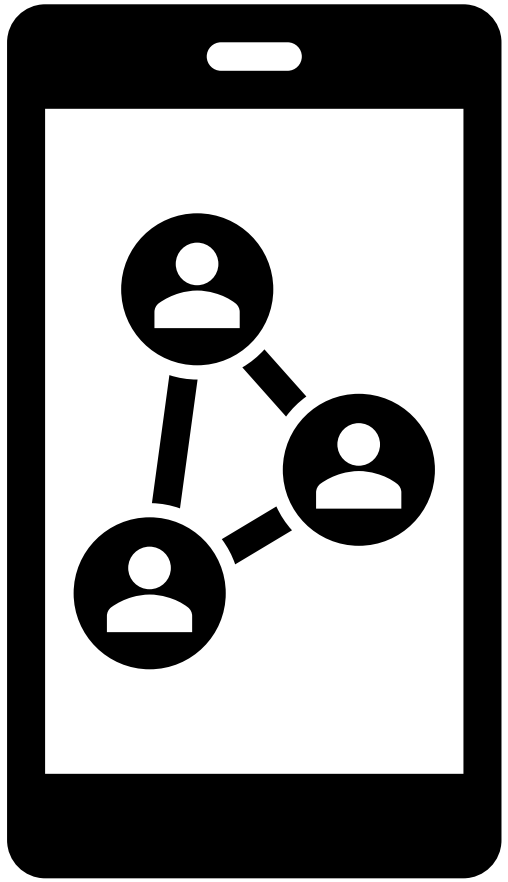
3: Head is Fast, improved speed relative to slow/CRAN/base



5: Conclusion



- data.table is an efficient package for data manipulation.
- atime package proves to be exceptionally useful for conducting comparative benchmarking and performance testing.



GitHub:DorisAmoakohene

LinkedIn: Doris Amoakohene

Email: @daa464@nau.edu