HARVARD UNIVERSITY
Information Technology
UNIVERSITY RESEARCH COMPUTING

Best practices and Learned Lessons in Refactoring Researcher-Developed Statistical R Packages

Naeem Khoshnevis

Mahmood M. Shad

FAS Research Computing

Harvard University

October 11, 2022

About me

- Ph.D. in Geophysics (Numerical Analysis)
- MSc. in Computer Science (Deep Reinforcement Learning)
- RSE (Statistical Software)

RSE

- The Four Pillars of Research Software Engineering *
- Ten Reasons to Be a Research Software Engineer **
- Software papers and the software itself are getting more credit
- I enjoy what I am doing, but RSE can be challenging

*. Cohen, J., Katz, D. S., Barker, M., Hong, N. C., Haines, R., & Jay, C. (2020). The four pillars of research software engineering. IEEE Software, 38(1), 97-105.

**. C. Cannam, D. Gorissen, J. Hetherington, C. Johnston, S. Hettrick, and M. Woodbridge. Ten reasons to be a research software engineer. https://www.software.ac.uk/blog/ 2013-08-23-ten-reasons-be-research-software-engineer. Accessed 02-Oct-2022.

RSE – Researcher Collaboration

HARVARD UNIVERSITY
Information Technology
UNIVERSITY RESEARCH COMPUTING

Software Eng. *Data Structure* *Algorithm* *MPI* *DevOps* **HPC** *CPU intensive* *Schedulers (SLURM)* **Big Data** *Memory intensive* *Singularity* *Docker* **Cloud Computing** *Numerical Analyses* *Machine Learning* **Research** *Statistical Analyses*

Knows what can be done.

RSE

The biggest challenge is bridging this gap.

Knows what Needs to be done.

Researcher

Domain Knowledge *Securing Funding* **Science** *State of the Art* *Publication* **HPC** *Writing Grants* *Schedulers (SLURM)* **Cloud Computing** *Spinning up a container*

- *Prior knowledge on the domain*
- *Reading scientific papers*
- *Attending weekly lab meetings*

- *Data structure and algorithm*
- *Familiar with version control systems*
- *Can contribute to an ongoing Software Dev.*

**** extreme cases ****

If the RSE meets the researcher at the other end

An RSE fully understands the research, knows what needs to be done.

Computer Scientist + domain research background

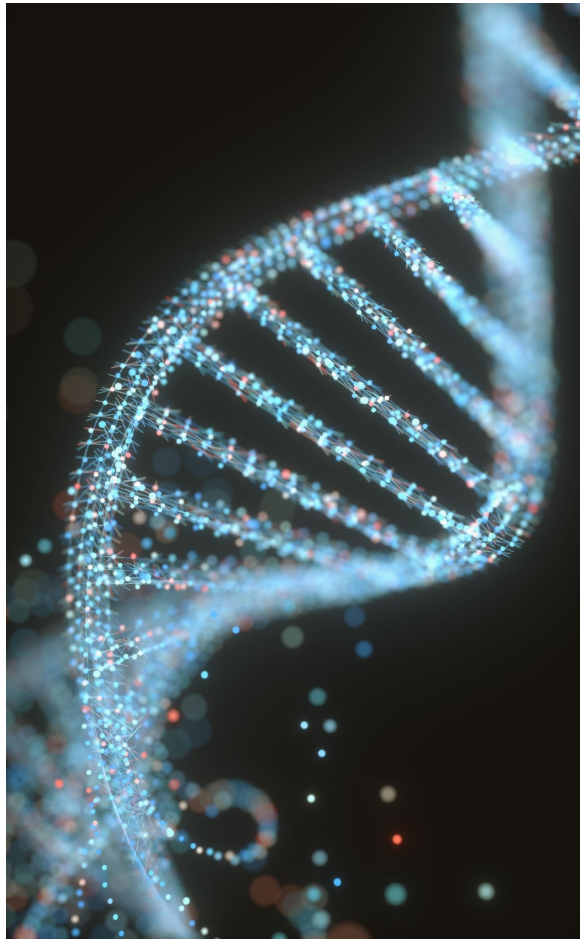
Computational Scientists

I need to convert this function into GPU or scale it out.

If the Researcher meets the RSE at the other end

Research Status

(Done or Ongoing)



The research is Done

- There is a CPU version of the code, and they need a GPU version.
- The code is there but needs to be converted into a package (or standard software).
- The code is developed, but it is not efficient.

Advantages

- You have an idea about the project, inputs, and outputs, interpreting the results.
- There are running examples, most probably with different data sizes.

Disadvantages

- The core developer may not be around.
- The code was developed for one or two papers.
- There might be a lack of quick interest in the final product.

The research is Ongoing

- You are part of the team and can steer the development process.

Advantages

- The lab members use the developed code, and you get feedback about the performance.
- With frequent interaction with researchers, you close the knowledge gap.
- Researchers have an idea and implement a part of it.
- They know (sort of) when we can call the project done.

Disadvantages

- They may start and contribute some inefficient code.
- They are testing some hypotheses. It may or may not work.

HARVARD UNIVERSITY
Information Technology
UNIVERSITY RESEARCH COMPUTING

Managing Expectations and Building Effective Communication!

Best Practices (before starting)

Deliverables

- Have a clear understanding of the final product.
  - ✓ Refactored Software runs on HPC
  - ✓ User and Developers Documentation

Call it Done!?

- Have a checklist for when we can call it done.
  - ✓ Software successfully process XYZ data for 1 ~ n cores on the HPC.
  - ✓ Do we need to do verifications test? Who is doing that?

No Ambiguities

- Eliminate ambiguous words from your communication.
  - ✓ Standard Software
  - ✓ Optimizing Software
  - ✓ Cloud Ready
  - ✓ Refactoring

Best Practices (before starting)

Test Data

- Ask for at least three data sets.
  - ✓ **Small**: Get results in a fraction of a second
  - ✓ **Medium**: Can challenge a personal computer
  - ✓ **Big**: Can challenge HPC systems

Avoid Assumptions

- People make assumptions.
  - ✓ I tried to run the code, but I have received an error.
  - ✓ I want to be able to run the code on HPC.

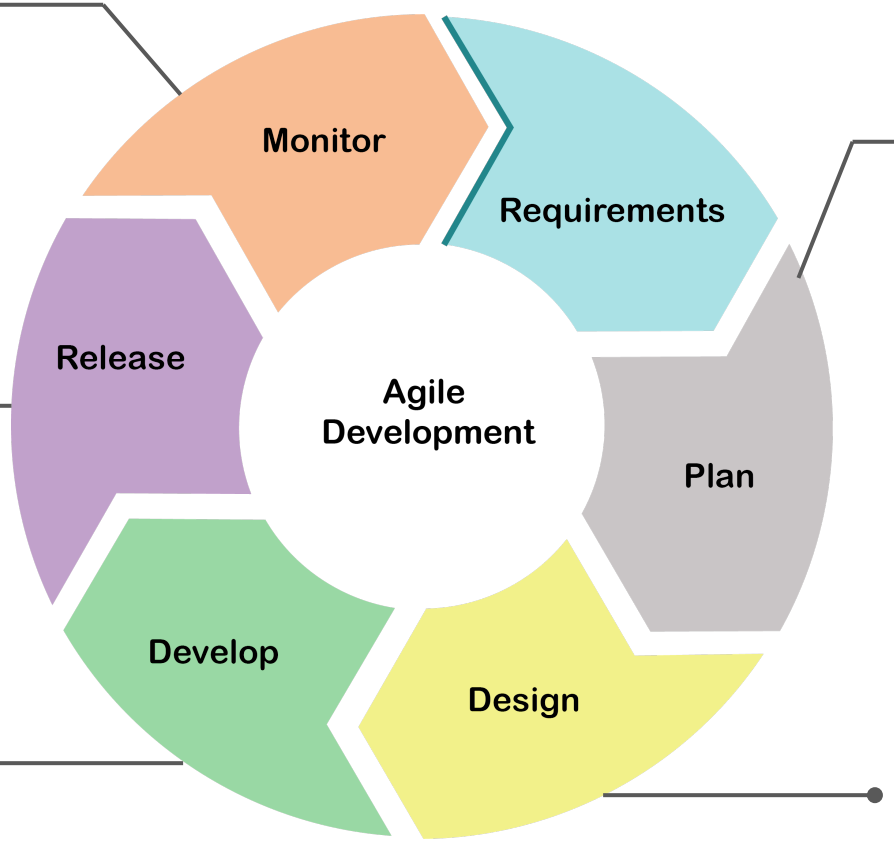
Keep Old Code

- Have convention to keep old implementation.
  - ✓ For debugging purposes (verification process).
  - ✓ Researchers may think old implementation was better.

Project Life Cycle (during project)

HARVARD UNIVERSITY
Information Technology
UNIVERSITY RESEARCH COMPUTING

- ✓ Comprehensive logging
- ✓ Help them to visualize the final product although it is not fully completed.
- ✓ Also, User's and Developer's Doc



- ✓ Plan on releasing a minimum package or code that works.
- ✓ Use placeholders for features that are not implemented.

https://saigontechnology.com/blog/agile-vs-waterfall-in-software-development

Notes on RSE as a Career

- ❑ You should expect to get involved in the research alongside the researchers.
  - ✓ How will your next year look like? You cannot keep reading papers for each project.
- ❑ Not all citations are the same.
  - ✓ Not sure how citations on different topics may help your career.
- ❑ Switching between projects is not as easy (and fun) as it seems.
  - ✓ How do you want to introduce yourself?
    - CPU or GPU, R or Python, Numerical or Statistical Analyses
    - Finite Element or Finite Difference.
- ❑ Teams are small and each RSE focus on one topic, possibly you should expect less collaboration.
- ❑ RSE is a developing field. It does not have well-defined framework. Each project has a unique situation.
- ❑ Although it should be agile (bi-weekly sprints), it is possible to experience a delay in getting feedback from the researchers.

Thank You!