DS 7347
High-Performance Computing (HPC) and Data Science
Session 14

Robert Kalescky
Adjunct Professor of Data Science
HPC Research Scientist

June 9, 2022

Research and Data Sciences Services
Office of Information Technology
Center for Research Computing
Southern Methodist University

Outline

Async Lecture: Tuesday, June 14, 2022

Session Question

Programming Best Practices

Readings and Assignments

Async Lecture: Tuesday, June 14, 2022

Async Lecture: Tuesday, June 14, 2022

- No class meeting on Tuesday, June 14, 2022
- Lecture will be recorded ahead of next Tuesday and will be posted in 2DS

Session Question

Session Question

How is reproducibility important to data science?

Programming Best Practices

Script and Parameterize Everything

- Script workflows
- Script software stack builds
- Specify software versions
- Record parameters and inputs in a separate file
- Self documenting
- Develop by running scripts rather than issuing commands directly
- Rebuild often
- Rebuild with different toolchains

Version Control Files

- Record fundamental code and data files in a version control system
- Develop on separate branches
- Commit frequently
- Commit for specific, not bulk, changes
- Use pull requests for merging

Document Procedures and Comment Code

- Document how to your execute your workflows
- Document general file and directory structures
- Comment major code functions

Use Containers and Environments

Use tools to define and isolate software environments:

- Containerize with Docker and export as needed, i.e. Singularity, etc.
- Script builds or use a tool like Spack
- Use language-specific environments
- Rebuild often
- Rebuild with different toolchains

Use Optimized Libraries

Use optimized libraries where possible:

- BLAS and LAPACK
- FFT
- Solvers
- Machine learning backends, which themselves are built on optimized libraries, e.g. DNN

Validate with Tests

- Check that outputs are correct for given inputs
- Check that performance goals are met

Automate Tasks

- Script and parameterize everything
- Use continuous integration tools, i.e. GitHub Actions
- Run correctness tests
- Run performance tests

Readings and Assignments

Readings

None

Readings and Assignments

Project

- Create a new **private** repo in GitHub using the template `https://github.com/SouthernMethodistUniversity/msds_hpc_project_template`
- Add me, `rkalescky`, to the new repo.
- Due 12:00 AM Central, Thursday, June 16, 2022