Best Practices for Reproducible Research

Internship Abstract

Biostatistics Lab at NASA's Johnson Space Center

Gabriella Novak

Summer 2020

Introduction Over recent decades, scientific research has become increasingly dependent on computational assistance to record, store, and analyze data. The continued incorporation of technology introduces the need to repeat these analyses and arrive at consistent, reproducible results. Reproducibility refers to this ability to recreate computational results –including tables, figures, and values reported within text– given input data, code, and sufficient documentation. Reproducible results are verifiable, increasing trust between the researchers who produced them and the reviewers, funding agencies, and readers who rely on such results to make informed decisions. Availability and documentation of previously developed computational tools, methods, and data sets save time, effort, and funds by reducing redundant work in the scientific community and opens the possibility for improvement and application in other projects by other researchers or reuse by the original researcher. Quality documentation aids in effective communication about the research to funding agencies and supervisors and avoids embarrassing and possibly detrimental situations of knowledge loss. This ability to share work and understand it in the absence of the original investigator is vital for continuity and extension of the research. Hallmarks of reproducible research – organization, documentation, and automation throughout the research process– improve quality of scholarship and reduce errors that can widely impact the scientific community.

Methods I reviewed the available literature on reproducible research and synthesized the standards and guidance provided by the National Academies of Sciences, Engineering, and Medicine and several prominent researchers into a primary and supplementary report. The former contains environment-independent reproducible research recommendations which are applicable regardless of software environment. The latter provides similar guidance specifically to R and RStudio users. It extends the primary report's recommendations with implementations in a software environment with excellent reproducibility capabilities and established usage in government, industry, and academic research. The supplement additionally has an accompanying R package, called `SKReproTools`. I developed this package for internal NASA use. It combines the action of several existing R packages with novel code to produce a user-friendly reproducible workflow while abstracting away the technical minutiae that can be time and expertise prohibitive for some researchers.

Results The primary environment-independent recommendations document defines and motivates reproducibility, addresses common challenges and primary risks, and established Golden Rules of Reproducibility. The bulk of the report provides guidance on various aspects of increasing reproducibility in research organized into three broad categories: organization, documentation, and automation. The recommendations given in the document do not suggest a drastic and immediate overhaul of established procedure, which can be prohibitively frustrating and hamper productivity. Rather, the advice is modular so research teams can apply techniques in combinations that befit their needs and goals. The R and RStudio supplement contains a discussion of R as an analysis software and a brief introduction to the software environments on which an R-based workflow relies heavily. The remainder of the document pairs the guidance in the primary report with specific implementations in R through the RStudio IDE. The structure matches that of the environment-independent document in order to facilitate ease of cross referencing. The supplement

concludes with a worked example tutorial which demonstrates the creation of a minimal reproducible project and highlights the features and functionality of `SKReproTools`.

Conclusion Guidance on best practices for reproducible research will improve the quality of scholarship and streamline the workflow of the Biostatistics, biomedical, and environmental science labs at Johnson Space Center. The techniques detailed in these reports will reduce errors that could lead to inappropriate conclusions and endanger the reputation, funding, and suitability for publication of a research project. In addition to my work articulating reproducible research guidelines, I gained valuable insight during my time as an intern. Not only did I have the opportunity to gain further experience with literature review, communication of complex topics to non-experts, and R package development but I was able to interact with NASA statisticians and learn about their work. This is particularly important for statistics as it is a wide umbrella containing applicability in a wide range of environments. The experiences I have had during my internship have allowed me to better inform my educational and career trajectory for the future.

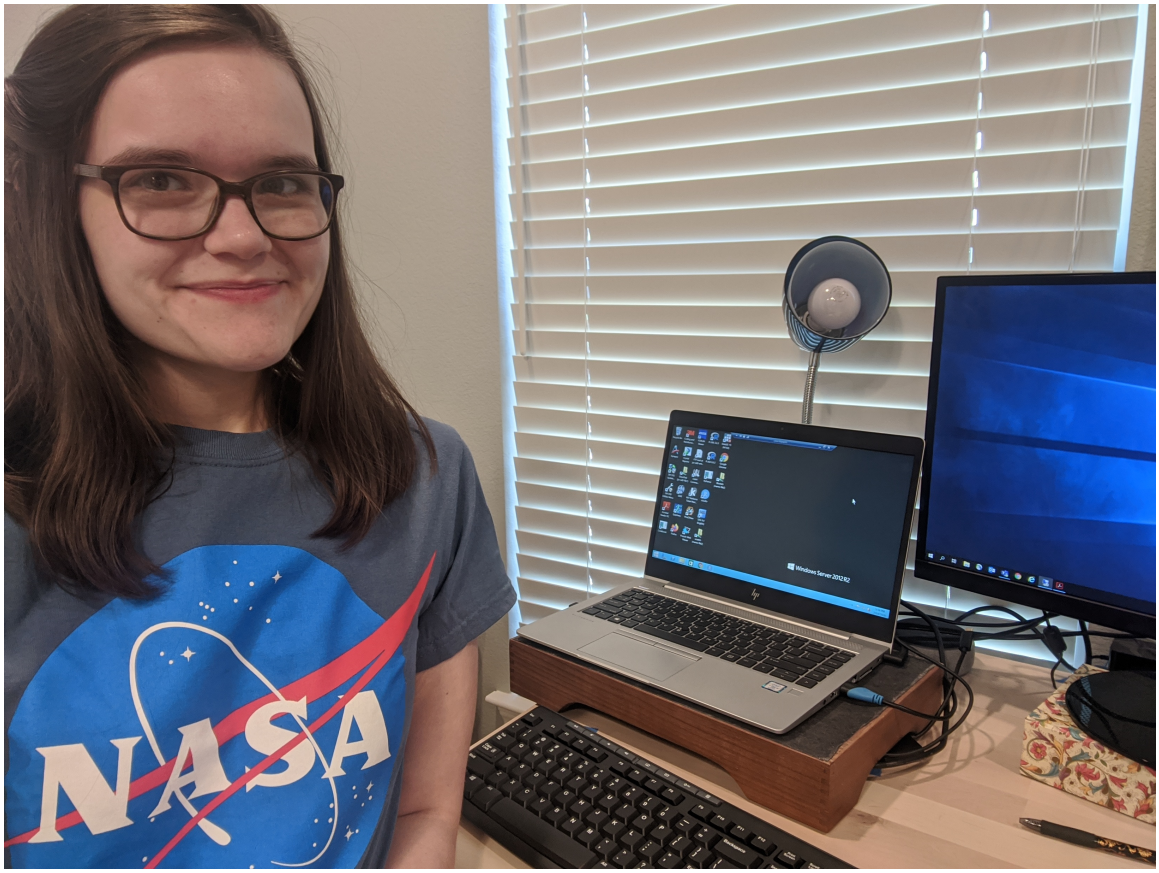

Figure 1: In summer 2020, JSC hosted its first virtual intern class. Here, I am posing with my telework station.