Reproducible Research and Teaching in the Cloud

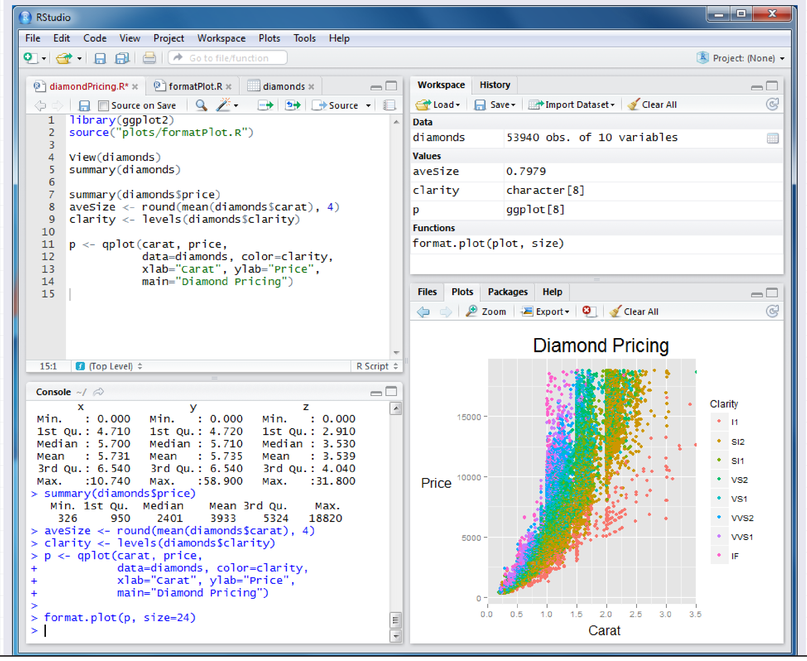
Gordon David Mosher and Thomas Girke
University of California, Riverside - Departments of Statistics and Bioinformatics

Abstract

- These days, science is increasingly supported by Data Science. Publishing of scientific data now requires technical skills often outside the realm of the researcher's expertise. This Data Science research project uses nearly twenty computer languages and file formats to move scientific data to the web, yet the end user need only enter a few commands to harness that power.
- Our research focuses on the automation of documentation tools. Existing documentation tools are connected by custom scripts to create nearly seamless paths for scientific data to reach various output formats. R is a very powerful statistical programming language and RStudio is its user friendly interface, thereby making R's power and interactive data analysis features accessible to scientists in any field. R has the ability to analyze data and create charts and graphs. Further, R will combine these results with markdown text to create complete documents ready for publication. Additional processing can be applied to create novel or specialized output formats.

** Experienced R Users - Skip to Step 2

Step 1 - Bring your markdown text, LaTeX, and data



Step 2 - Render to an R Markdown output format

Step 3 - Methods - Apply automation to publication

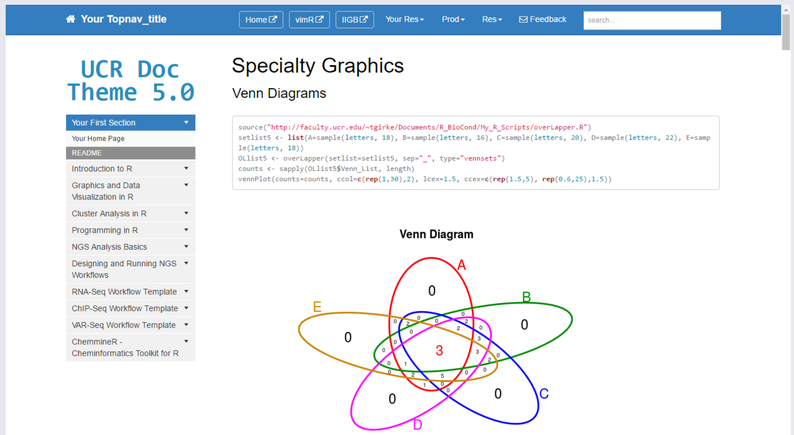
- The solution to complex publication procedures is to simplify and automate. By realizing the need to streamline the production of quality publications we have discovered several widely accepted documentation tools that can be chained together to simplify the movement of scientific data from the lab to final publication. After having selected these existing tools our method involves the writing of programs that read the coded output of each tool, manipulates it, and then writes out new code that is appropriate input for the next tool in the chain.
- The goal is to accomplish a task in seconds that previously took hours of skilled human labor. Minimizing error by simplification, and allowing a greater proportion of time and effort to be directed toward creativity and quality. The result is a finished website, pdf, or journal article in less time and requiring less technical effort.
- Our methods include support for reproducible research and teaching in the cloud. Reproducible research means that data, functions, code, and text are submitted to the tool and it will evaluate the data and publish the formatted results in the type of report selected. It also includes the ability to verify others work and expand upon it. Teaching in the cloud includes the development and automatic publication of online university courses using the same toolchains.

Conclusion

- Simplifying the process of publication for scientists results in more productivity and better quality documents. The publication of scientific data is becoming a Data Science.
- Finding tools that automate publication are easy to find, but since technology is changing so rapidly, it may be difficult to determine which toolchains to adopt. The tools we currently recommend are free to use and can be installed in Windows, Linux or OS X.
- All computations and graphs are created with the open source software `R` [1]. This poster was created using a scripted toolchain based on LaTeX, rather than a graphical layout program.

Step 4 - Add to Jekyll Doc Theme and push to GitHub

Results - Reproducible Research in the Cloud



References

[1] R Core Team.
R: A Language and Environment for Statistical Computing.
R Foundation for Statistical Computing, Vienna, Austria, 2015.

Acknowledgements

Mentoring Summer Research Internship Program (MSRIP)
CAMP - California Alliance for Minority Participation
Sponsored by the National Science Foundation

gordon.mosher@email.ucr.edu