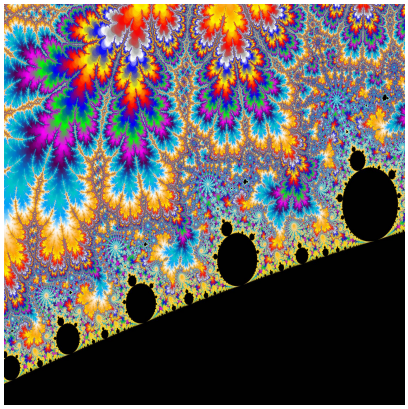

Is R the Right Toolset for eScience/Open Science? Fireside Chat with Maëlle Salmon

Description

Is R the Right Toolset for eScience/Open Science? Fireside Chat with Maëlle Salmon

Shalini Urs

Replicability Crisis to Open Science

“Let my toolset change your mindset about your dataset.” — Maëlle Salmon (paraphrasing Hans Rosling)

Many a time, a crisis leads scientists to push the envelope. The year 2015 was a turning point in research practices in Psychology. The growing concern that psychological research fails the reproducibility test, a defining feature of science, started with a study of the non-replicability of psychology findings, reported in 2015 by the Open Science Collaboration project. The results of this project led by University of Virginia psychologist Brian Nosek were shocking. When the 270 psychologists of the Open Science Collaboration tried to re-run the 100 published psychology experiments, less than half of the re-run experiments had worked. Apart from the shocking results, the study was published in the prestigious journal Science and received widespread attention from the media. The shocking revelations spurred similar studies in other disciplines, and the results were comparable or even worse. However, this study and its authors clearly stated that this is not an indictment of the research or researchers but is an indication of cultural practices in scientific communication that may be responsible for the observed results. Some call the replicability crisis an epistemic crisis. The Replicability crisis is accentuated by the mismatch between traditional scientific dissemination practices and modern computational research practices.

This crisis led to soul searching and calls for recalibrating the research practices. It became increasingly evident that we needed to move beyond traditional articles and presentations to the delivery of executable objects integrating the data and computational details, including scripts and workflows upon which the findings are based for the progress of science. The call for open science has been a significant outcome. Research data and code publication in a publicly available venue and format have become the hallmark of open science. However, technical issues and institutional barriers in coupling data and the process impede the open science movement. Diverse initiatives and studies are underway to overcome some of these barriers.

—Many a time, a crisis leads scientists to push the envelope.—

Exemplars of Open Science Initiatives: The Centre for Open Science and Whole Tale Project

The Centre for Open Science (COS) led initiative—the Transparency and Openness Promotion (TOP) Guidelines created by journals, funders, and societies to align scientific ideals with practices, currently has 5000 signatories and has been adopted by many major academic publishers, including the American Association for the Advancement of Science. TOP provides a suite of tools to guide the implementation of better, more transparent research. In addition, the COS had set up the Open Science Framework, a web-based public repository for experiment-related data and code. It is built entirely from free and open-source software.

The Whole Tale project—an NSF-funded Data Infrastructure Building Block (DIBBS), is an initiative to build a scalable, open-source, web-based, multi-user platform for reproducible research enabling the creation, publication, and execution of tales—executable research objects that capture data, code, and the complete software environment used to produce research findings. It attempts to address the barriers of coupling data with the process by connecting computational, data-intensive research efforts with the larger research process—transforming the knowledge discovery and dissemination process into one where data products are united with research articles to create “living publications” or *tales*. Brinckman et al. (2019) report on the Whole Tale environment’s design, architecture, and implementation.

R for Open Science

Andy Wills (2019) draws parallels between the emergence of the open-source software movement and the open science movement. According to him, the reproducibility problem was not with analysis or the data but with the research practices. Since psychology experiments primarily involved computer-based testing, the unavailability of the testing program’s source code was perhaps the reason for irreproducibility. One could not audit research properly without access to the source code on which the experiment was based—the testing software, the raw data, and the analysis scripts. The Free and Open Source Software community perhaps could show how science could be open and transparent. Just as computing has its advocates of open-source software, psychology and other sciences started gaining advocates for Open Science. According to Wills, an excellent reason for using R is that all analyses take the form of scripts. Thus if the analysis is done entirely in R, a complete, reproducible record of the analysis path is already created. Anyone with an internet connection can download R and reproduce the

analysis using the script. In other words, we can easily achieve the goal of open, reproducible science with R.

Tina Amirtha (2014) says that the rise of R language is bringing open source to science. R is crossing over from just calculating statistics to scientific experimentation and bringing hacker culture. Thanks to R, the "open science" advocates are succeeding at getting science to go open source. A growing number of researchers have joined the R development community to create new libraries that branch away from statistical analysis and into parsing the ever-increasing quantity of scientific articles and data that find their way online. Moreover, it could change the way we do science in a significant way.

Lortie (2017) proposed R as a natural bridge between data and open science and a powerful ally in promoting transparent, reproducible science.

R: Origin and transition to a Research Software

Open-source R is the statistical programming language data experts worldwide use for everything from mapping broad social and marketing trends online to developing financial and scientific models that help drive our economies and scholarship.

The R language, closely modeled on the S Language for Statistical Computing (Bell Labs) in the mid-1970s, was first implemented in the early 1990s by Robert Gentleman and Ross Ihaka at the University of Auckland. They established R as an open-source project in 1995, and since 1997 the R project has been managed by the R Core Group. In February 2000 came the first release of R. The R Foundation provided the official public structure for the R Community. In addition, the R Foundation ensures the financial stability of the R-project and holds and administers the copyright of R software and its documentation. R's capabilities are extended through user-created *packages*, which offer statistical techniques, graphical devices, import/export, reporting, etc. These packages and their easy installation and use have driven their widespread adoption in data science. Researchers also use the packaging system to organize research data and code and systematically report files for sharing and archiving.

Even before the replicability crisis, R was being touted as an alternative to MATLAB, SAS., and SPSS packages which were the mainstay of researchers across sciences and social sciences. *R lets experts quickly, easily interpret and interact with and visualize data.*

When you see powerful analytics, statistics, workflows, and visualizations used by data scientists and researchers, the chances are that the R language is behind them. As of June 2022, R ranks 16th in the TIOBE index, a measure of programming language popularity. R had peaked at eighth place in August 2020. Marwick et al. (2018). investigate how the R programming language software packages' structure and tooling are used to produce research compendia in various disciplines. Using real-world examples, they show how researchers can improve the reproducibility of their work using research compendia based on R packages and related tools.

Maëlle Salmon on the Power of R, Open Source Community, and rOpenSci

Listen to this episode of InfoFire, wherein I am in conversation with Dr. Maëlle Salmon, who describes

herself as Aficionada a R, R(esearch) Software Engineer & Blogger and works part-time as Research Software Engineer at rOpenSci, in addition to being an Associate editor for rOpenSci's Software Peer Review—a peer-review system of R packages. We chat about a host of topics and trivia around the R software, her passion for R, her experience with the R Open Source Community, and her two projects—CHAI(Cardiovascular Health effects of Air pollution in Telangana, India) and Health Impact Assessment (HIA) of cycling network expansions in seven European cities, and R-Ladies Global and Blogging experience.

Our conversation started with Salmon telling us her views on open science, R language, and rOpenSci, a not-for-profit organization. rOpenSci aims to transform science through open data, software & reproducibility. It aims to create technical and social **infrastructure** in the form of carefully vetted, staff, and community-contributed R software tools that lower barriers to working with scientific data sources on the web.

Salmon shared how rOpenSci develops R toolsets to support open and reproducible science. These toolsets first help access the shared data and support reproducibility. For example, when you cannot share the data in a paper because it is patients' data. rOpenSci develops tools for accessing data online. There are packages with scripts to access these datasets and orchestrate workflow in a way that will save your time and save computer time.

Salmon also shared her incredible experience of community-driven learning and review at rOpenSci when the first R package she submitted to the community. According to Salmon, even if the R Community is not perfect (like any other open source community), it is very welcoming, and one is not made to feel inferior. She discovered that, most often than not, the limitations are not with the R Language but with one's skills and experience. For example, when she wrote and submitted her first R package to access open-air quality data, which (per the policy and process of rOpenSci) was reviewed like a journal article but openly, she felt lucky to have found her niche. The community's kindness and helpfulness made her feel welcome and stay. Furthermore, she discovered that the open-source culture improved the toolsets.

When asked to give specific examples of R toolsets that she believes are better than similar ones, Salmon picked the **magick R package**, the **image processing tool in R** bound to the ImageMagick library: the most comprehensive open-source image processing library available. She talked about how one can play with the package as it supports many standard formats such as png, jpeg, tiff, pdf, etc., and enables different manipulations types such as rotate, scale, crop, trim, flip, blur, etc.

Another R tool Salmon picks as an exemplar is the **Targets** package, a pipeline tool to maintain a reproducible workflow. According to Salmon, it is time-efficient as Targets skip costly runtime for tasks that are already up to date, orchestrate the necessary computation with implicit parallel computing, and abstract files as R objects. In addition, an up-to-date Targets pipeline is tangible evidence that the output aligns with the code and data, substantiating trust in the results. Finally, she adds that it is also quite remarkable because when you use this package, you get other visualization of your pipeline.

"Let my toolset change your mindset about your dataset."

When asked about her well-known quote, "**Let my toolset change your mindset about your dataset," which is an adaption of the very famous quote of** Hans Rosling: "Let My Dataset Change Your Mindset," she had this to say. For example, one can collect all this data about *Salmonella and not do anything about it; that is cool. However, applying some steps to this data, if you can, for example, detect outbreaks and look for causes of these outbreaks and prevent further outbreaks,* which will change your mindset (a la John Snow, who is famous for his investigations into the causes of the 19th-century cholera epidemics, and is also known as the father of (modern) epidemiology. So toolsets help us in doing stuff on data like this and change your mindset about your dataset.

Young and passionate Maëlle Salmon's contributions to the R and R community are commendable. She has made 150 repositories and has 4,153 contributions to her credit in the last year alone.

References

Amirtha, T (2014, March 28). How The Rise Of The "R" Computer Language Is Bringing Open Source To Science. *Fast Company*. Retrieved from https://www.fastcompany.com/3028381/how-the-rise-of-the-r-computer-language-is-bringing-open-source-to-science

Brinckman, A., Chard, K., Gaffney, N., Hategan, M., Jones, M. B., Kowalik, K., … & Turner, K. (2019). Computing environments for reproducibility: Capturing the "Whole Tale." *Future Generation Computer Systems*, *94*, 854-867.

Hicks, D. J. (2021). Open science, the replication crisis, and environmental public health. *Accountability in Research*, 1-29.

Lortie, C. J. (2017). Open sesame: R for data science is open science. *Ideas in Ecology and Evolution*, *10* (1).

Marwick, B., Boettiger, C., & Mullen, L. (2018). Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, *72*(1), 80-88.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349* (6251), aac4716.

Wills, A (2019, February 19). Open Science, Open Source and R. *Linux Journal*. Retrieved from https://www.linuxjournal.com/content/open-science-open-source-and-r

Cite this article in APA as: Urs, S. (2022, June 15). *Is R the right toolset for eScience/open science? Fireside chat with Maëlle Salmon*. Information Matters, Vol. 2, Issue 6. https://informationmatters.org/2022/06/is-r-the-right-toolset-for-escience-open-science-fireside-chat-with-maelle-salmon/