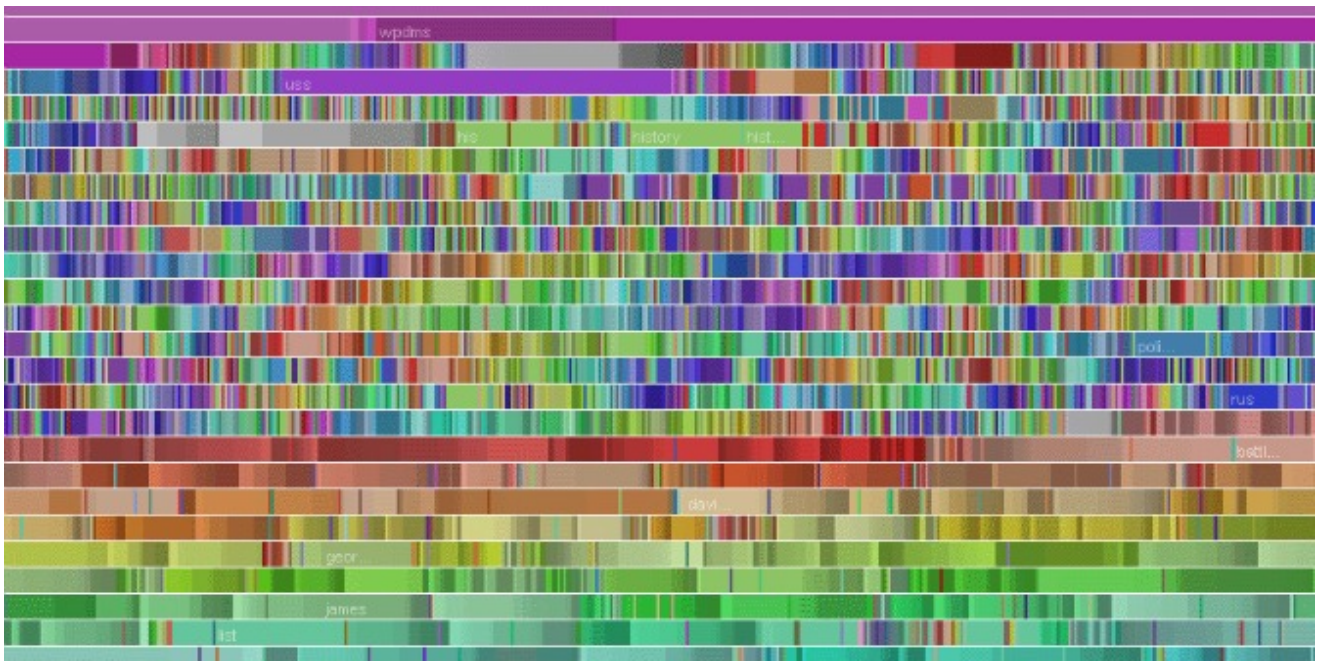
Reproducible computing with rctrack: Software package addresses fundamental scientific challenges of Big Data era.

blogs.lse.ac.uk/impactofsocialsciences/2014/07/07/reproducible-computing-rctrack-big-data-challenge/

7/7/2014

Published descriptions of data sets and analysis procedures are helpful ways to ensure scientific results are reproducible. Unfortunately the collection and provision of this information is often provided by researchers in retrospect and can be fraught with uncertainty. The only solution to this problem is to computationally collect and archive data files, code files, result files, and other details while the data analysis is being performed. **Stan Pounds** *highlights the release of rctrack, a software package that automatically enables this file collection for analyses performed with the open-source R programming language, thereby minimizing the burden of collecting and archiving details.*

Recent advances in computing and other technologies have empowered researchers in many disciplines to rapidly and economically collect very large data sets on their systems of interest. Biologists can use new biotechnologies to quickly and economically determine the sequence and abundance of the DNA and RNA for almost every gene. Also, mobile devices allow researchers to immediately capture data from almost any geographic location and instantly forward it to a centralized database for subsequent statistical analysis and interpretation. We are now able to measure our systems of interest at a level of detail that was unimaginable just a few years ago. IBM estimates that 90% of the world's data has been created in the last 2 years alone, and the rate at which we can collect data continues to grow rapidly. Welcome to the era of big data, with all of its opportunities and challenges!



A visualization created by IBM of Wikipedia edits. Image credit: Fernanda B. Viégas (Wikimedia CC BY 2.0)

The challenges of storing, organizing, maintaining, and securing big data are not merely technical nuisances. Left inadequately addressed, these technical problems can rapidly metastasize into interpretational challenges that threaten to undermine reproducibility, the principle by which science establishes its own legitimacy. Reproducibility

has long been the criterion used to confirm the scientific validity of a new discovery or theory. A novel scientific discovery or theory is confirmed as valid when supportive evidence is consistently obtained across multiple studies by several investigators working independently. Conversely, novel theories or discoveries are debunked as invalid when other studies consistently produce evidence to the contrary. The primary findings of research studies involving big data are the computationally obtained results of statistical analyses. However, the published analysis results of some studies involving big data cannot be internally recapitulated from that study's own published data; thus, the concept of reproducibility across studies becomes meaningless. If the computational analysis results of a study cannot be internally recapitulated, then it is unclear exactly what data were included in the analysis and what computational procedures were used to obtain the published result. Thus, it is also unclear whether the published result was obtained using appropriate data and statistically rigorous methods, exactly what the result of the published study really is, and whether or how other investigators should attempt to further evaluate that result. These uncertainties make it impossible for science to use reproducibility as the criterion to determine the validity of such published research results.

Big data research results are often not internally reproducible because their analysis necessarily involves a very large number of complex data and sophisticated software files that continuously change with updates, corrections, and other modifications. Most data-analysis software does not automatically collect and archive all of these files and other information needed to adequately document the analysis so that it can be exactly recapitulated at a later time. Thus, investigators must manually maintain complete and accurate records of a seemingly endless list of details about exactly which data files, software files, and analysis options were used to generate a particular set of results. Subsequently, the published descriptions of the data sets and analysis procedures are written from memory and frequently do not completely and accurately represent how the analysis that generated the reported results was actually performed.

The only solution to this problem is to computationally collect and archive data files, code files, result files, and other details while the data analysis is being performed. The recently published rctrack package does exactly that for analyses performed with the freely available, open-source R statistical computing programming language that is widely used by statisticians, engineers, and scientists in many disciplines. R is an ideal platform on which to perform and disseminate publicly transparent and fully reproducible big data analyses because the software and thousands of application-specific add-on packages are already freely available. This removes many financial and technical barriers to the development and execution of complex data-analysis programs.

Image credit: R Foundation (GNU License)

The rctrack package minimizes the programming burden of collecting and archiving details to adequately document an analysis so that it is fully reproducible. After the package has been installed, the data-analysis program needs to include only three additional statements: one statement to make the rctrack package available for use, one to activate detail-collection procedures, and one to discontinue detail-collection procedures and generate a read-only archive folder that includes data files, code files, result files, and other details needed to reproduce the analysis. Some users may want to specify options that control the archiving to avoid excessive redundancy, which wastes storage space. No other modifications to the data-analysis program are necessary because rctrack embeds the detail-collection operations into the functions that read or write files or perform other tasks that must be documented for reproducible computing. Thus, the detail collection is quietly performed in the background as the data-analysis program is running.

The rctrack package can produce an archive that provides a complete and fully transparent record of the data analysis to facilitate a thorough peer review. By providing the rctrack archive as supplementary materials of a

scientific publication, authors will allow their peers to reproduce the results and thereby confirm exactly what data and methods were used to generate those results. The merits and limitations of the data and the analysis methods can then be openly discussed, critiqued, and debated to improve the interpretation of the results and the rigor of future research.

The rctrack archive can also be used as a practical example of how to complete a similar analysis of other data sets. It can serve as a template for others to modify, so that they can more readily perform a similar analysis of their own data. This strategy may be used within a research group to increase their productivity by making it easy for a larger number of members to perform that type of analysis. Also, published rctrack archives can be used by peer researchers to perform a similar analysis of their own data to more rapidly complete their independent replication studies that further support or question the validity of the published result. In this way, reproducibility can resume its crucial role in determining the legitimacy of scientific discoveries.

The rctrack package was developed by author Stan Pounds and Zhifa Liu. For more on its development and design, the full paper can be found at BMC Bioinformatics – doi: 10.1186/1471-2105-15-138

Note: This article gives the views of the authors, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

About the Author

Dr. Stan Pounds *is Director of the Division of Biostatistics for Computational Biology and Bioinformatics at St. Jude Children's Research Hospital in Memphis, TN. His group develops and applies innovative computational methods to perform rigorous statistical analysis of genomic and imaging data to advance our biological understanding of childhood cancer and improve its treatment.*