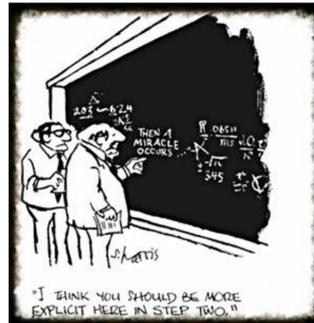
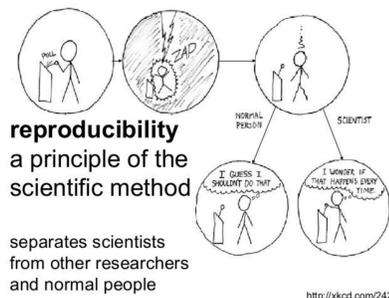


Reasons, Challenges and Some Tools for Doing Reproducible Research in Transportation Research

What Is Reproducible Research (RR)?

"The term **reproducible research** refers to the idea that the ultimate product of academic research is the paper along with the laboratory notebooks and the full computational environment used to produce the results in the paper such as the code, data, etc. that can be used to reproduce the results and create new work based on the research." (Wikipedia; accessed on the 6th of January 2020)

We are at a crossroads where how academic work should be conducted, documented and disseminated, and RR is an unstoppable and inevitable future. This paper aims to **introduce basic elements of RR to researchers in Transportation Research and facilitate this transition in the Transportation Research community.**



http://xkcd.com/242/

Source: statsmakemery.com

Why Bother?

"Computation-based science publication is currently a doubtful enterprise because there is not enough support for identifying and rooting out sources of error in computational work"

— David L. Donoho (Donoho 2010)

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

— John Claerbout, paraphrased in Buckheit and Donoho (1995).

The primary target of RR is computational analysis, which is deeply rooted in numerous disciplines thanks to the rapid advancement of computing technologies and the indispensable role of computing in the modern science and engineering, including Transportation Research.

By developing and relying more and more on computational algorithms we hope that the computational part of our work can be readily implemented by ourselves, easily disseminated to and precisely reproduced by other researchers, and conveniently modified, extended or enhanced in future studies. Ironically, what is happening could not be further away from the original goal, and many computational analyses become a black box as the algorithms used become ever larger and more complex, more data-thirsty, and more intelligent, which makes even precisely documenting an algorithm itself often challenging. The gap between what was actually done and what is described in the published paper is often so unfillable that reproducing the analysis upon which main findings are based is extremely difficult.

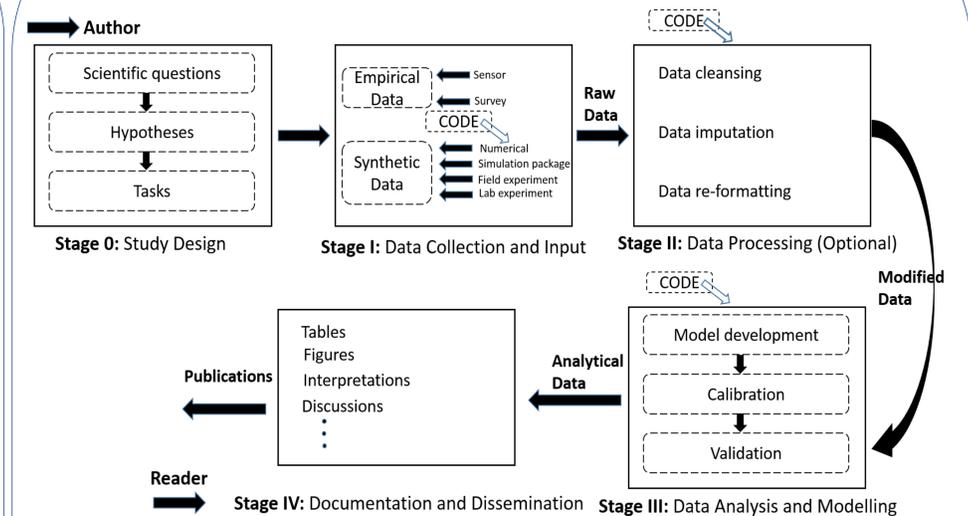
Main Benefits of RR

- Main benefits of RR to you as an author
 - ✓ Greater impact of your research. Making your research reproducible can increase your reputation as a researcher, and help attract more citations of your work.
 - ✓ Improved work and work habits, and improved teamwork. Making your research reproducible can make you more efficient and help you minimize the chance of error.
- Main benefits of RR to your readers
 - ✓ Improved quality and trustworthiness of the research they are reading.
 - ✓ Easy reproducibility of the analysis.
 - ✓ Easy extension of the analysis for further investigation on a related topic. This is also true to the author(s).
 - ✓ Improved productivity.
- Main benefits of RR to the society
 - ✓ Stewardship of public goods. RR has an irreplaceable role in curbing and preventing academic misconducts, frauds, and scandals.
 - ✓ Increased public access to public goods.
 - ✓ Better quality of the education system.

How to Conduct Reproducible Research

- The idea of RR is akin to the idea of literate programming proposed in 1984 by computer scientist and mathematician Donald E. Knuth for a better documentation of programs by considering programs to be works of literature
- A literate program is a document that is a mixture of code chunks/segments (sequences of commands in some programming language, e.g., R) and text chunks/segments (description of the problem, the code, and the results) and is written, formatted, and organized to be read by humans rather than a computer.
- Although literate programming has never gained a large following, when coupled with other tools for testing and validating code, it provides a powerful mechanism for conveying descriptions, carrying out reproducible data analysis, and enhancing readability of the final document. **This is the essential idea that dynamic document borrows from literate programming.**
- **Dynamic document is the cornerstone of RR**, consisting of three essential parts: data, source codes, and narratives. A dynamic document is an ordered composition of code chunks and text chunks that describe and discuss a problem and its solution, and can be regarded as a source document from which the published static document can be generated. The content of the dynamic document is dynamically generated, e.g., figures, tables, and etc. are generated by executing the code chunks through tangle, and inserted into the document through weave. Thus, **the contents including figures, tables, etc. can be updated on the fly each time a view of the document is generated.**
- To obtain a dynamic document, we need a **computing language** for doing the analysis, and a **documentation language** for narratives.

A Reproducible Procedure for Transportation Research



Tools for Doing Reproducible Research

- Data storage: public folder in Dropbox, GitHub, or similar service providers.
- Data analysis: **R**, **Markdown**, **bookdown**
- Figures, tables and equations: **R**, **Markdown**, and **bookdown**
- Documentation: **R**, **R Markdown**, and **bookdown**
- Collaboration & feedback: Git and GitHub



Challenges and a way for moving forward

- Technical challenges
 - ✓ "dependency hell": Because of constant changes of operational systems and software, a computational analysis that is original reproducible may become irreproducible later on
 - ✓ Researchers face significant barriers in learning these tools and approaches which are not part of their typical curriculum
 - Cultural or behavioral challenges: reluctance to publish the code; lack of requirements or incentive
- Note:** this article was written entirely in a reproducible way using Rstudio and RMarkdown. The source code for generating the exact document can be downloaded from the webpage:
<http://www.connectedandautonomoustransport.com/reproducible-research.html>

Acknowledgements

This research was partially funded by the Australian Research Council (ARC) through Dr. Zuduo Zheng's Discovery Early Career Researcher Award (DE160100449).