Reproducible Research @ Coursera

RK

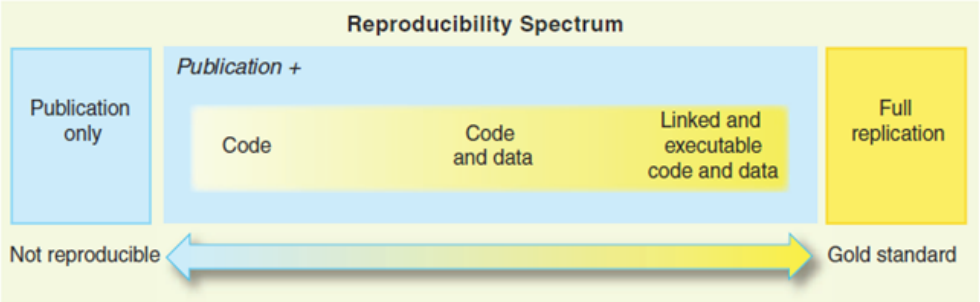
May 11, 2014

Context

I am a big fan of Literate programming. Seeing a course being offered on Coursera picked my interest. The whole lecture series comprises 4 lectures, each spanning an hour each. So, spending 4 to 5 hours on something that I had already learnt felt like a waste of time. However I realized that mind plays tricks on us and always gives us an illusion of mastery over something just because we are familiar with the topic. True learning happens only when we recall things from memory at regular intervals and take some tests along the way. Hence, I immersed myself for about 4 hours listening to the videos and attempting the tests. At the end of it, I am glad that I spent my time on Coursera. This document will contain some of my learnings from all the four lectures.

Week 1

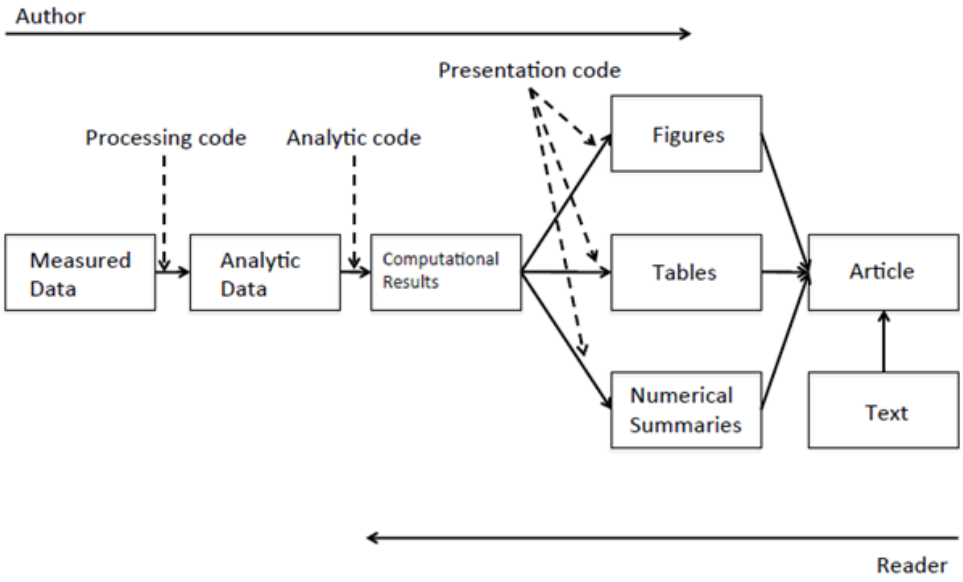
- When you communicate your analysis, it is becoming more important that others can reproduce your research.
- RR is about communicating what you have exactly to others
- Replication is important when you want to have a strong verification mechanism for your finding
- What's Reproducible Research ? Make analytic data and code available so that others can reproduce your analysis
- The analysis is becoming very complex. Analyzing algo means understanding the code and unless there is an RR document, it becomes difficult to understand the results
- Science mag has a nice article, titled, Reproducible Research in Computational Science written by Roger Peng



The above illustration is used for making a case for RR that can act as a bridge between full replication of results and no replication of results. Agreed the reproducibility is not checked with other datasets,

but at least the availability of code and the data used is a good starting point for someone trying to replicate the results of a academic paper or research.

- A video highlighting the need for RR - Deception at Duke: Fraud in cancer care?. The data had been manipulated by Dr. Potti at Duke to prove a theory correct. Duke university is seeing to it that all the papers written by Dr.Potti retracted. If the data and the code were distributed to everyone, this problem would not have arisen. This is a strong case for making RR mandatory for every researcher.
- What are the needs of a RR ?
  - Analytic data is made available
  - Analytic code is available
  - Documentation of code
  - Standard means of distribution
- The advantage of RR is that a reader gets access to the research pipeline



- Literate statistical analysis is a stream of text and code. Literate programs can be weaved to produce human-readable documents and tangled together to produce machine-readable documents
- Literate programming uses documentation language and programming language
- `Sweave` uses LaTeXas the documentation language and `R` as the programming language
- `knitr` uses `R` as the programming language and has a wide variety of documentation languages such as LaTeX, `html` and `Rmd`.
- Golden rule :Script everything. Whatever you do as an analysis, try to program it. If you want to understand EM algorithm, after going through the math, code it and see how it works on a sample dataset. Create a RR document so that is keeps a running log of what you have learnt. Also it will serve as a cue for retrieval practice.
- The dataset that you need depends on the kind of analysis that you want to do, i.e depending on whether the analysis is Descriptive or Exploratory or Causal or Inferential or Predictive or Mechanistic or combination of analysis.

Week 2

- John Gruber - creator of Markdown
- Markdown - simplified version of "markup" languages
- R Markdown - coupling of code with markdown
- Standard procedure : `.Rmd` is converted to `.md` which is then converted to `.html`
- `.Rmd` can be converted to slides via `slidify` package
- Literate programming - Original idea from Donald Knuth
- Literate programs are weaved to produce human readable documents and tangled to produce machine readable documents
- Literate programming needs a documentation language and a programming language
- Dependencies are not checked explicitly in `knitr`. This means you need to be careful when you are using variables from a code chunk that is cached.
- Never knew that the new trend to submitting programming assignments for online course is via posting github repository and the relevant SHA1.

Week 3

- Don't do any analysis by hand as it becomes data cleaning an irreproducible process. Things done by hand cannot be precisely documented.
- Version control slows you down but it is better in the long run.
- Don't save any output till the very end.
- There is difference between *Replication* and *Reproducibility*
- RR solves the following problems : Transparency, Data availability, Software availability, Improved transfer of knowledge. What RR doesn't guarantee is validity/correctness of the analysis
- An analysis can be reproducible and can be wrong.
- RR comes at the downstream of the scientific dissemination process
- Journal of biostatistics : The editor stamps the paper with C, D and R depending on whether the author has submitted code, the dataset and RR document
- Cochrane collaboration - Meta analysis model that is used for evidence based analysis

Week 4

- `cacher` package can be used by authors to create cache packages from data analysis for distribution.This package provides tools for caching statistical analyses in key-value databases which can subsequently be distributed over the web.
- Readers can use `cacher` package to inspect others data analysis
- A case study that shows how RR is useful in scientific research and the way it provides a healthy discussion.
- A video lecture, titled "The Importance of Reproducible Research in High-Throughput Biology: Case Studies in Forensic Bioinformatics" Dr.Keith Baggerly who was the first person to critique Dr.Potti's work. In the lecture, Dr.Baggerly makes a strong case for reproducible research.