Module 1 – lesson 03

Script

This lesson will cover key reproducible principles and practices and examples of why these principles and practices are so important.

We will discuss:

* the principles of reproducibility
* standard practices that are focal points for good reproducibility
* examples of reproducible workflow practices implemented in journalism at 538.com
* how these practices help tell stories using data
* the importance of transparency in scientific research as supported at the Journal of Biostatistics
* and a great example of how these practices worked very rapidly to find a treatment for an outbreak in 2001 of a deadly strain of *E.Coli.*

A cornerstone principle of the reproducibility process and structure starts with organization. Each project should have its own file storage organization structure. Each document, code, script and product should have a specific purpose. The versions of these files should be tracked with a version control system without creating multiple copies of files.

The associated documentation with your reproducible project should be clear and well defined. There should be instructions on how to use and combine the files in your project to create each end product you want.

The components of your documentation, code, or templates that will be or might be used again in other projects should be standardized. In other words, as much boilerplate that can be created should be created to make your workflow more efficient in the future. Don’t reinvent any wheels if you can help it.

Your files, documents and code should be stored and shared in a centralized way. Cloud-based computing often provides the platform for centralized storage and sharing of your projects with your team members and external stakeholders.

Finally, a main goal of your reproducible workflow should be efficiency.

In 2013, an article was published in PLOS One Computational Biology entitled “Ten Simple Rules for Reproducible Computational Research”

While the article focused on applications in computational biology, the key principles they recommended still apply including:

* avoiding manual steps
* version control and tracking
* implementing standardized formats
* storing and tracking raw data
* organizing your output – their list recommends a hierarchical organization
* linking textual documentation to the results
* and making the work transparent by allowing public access to scripts, runs and results

When considering standard practices, think about your own work:

* What do you want to automate?
* What could you re-use?
  + code, files, formatting, graphics, logos, header, footer, boilerplate
* What should you share with your team?
* What do you find yourself doing over and over?
  + correcting or reformatting
* If you won the lottery today (and left your job), what do you need to tell your replacement so they can pick up where you left off and complete your current tasks?

Keep these questions in mind as you go through this course. The purpose of this course is to help you find the answers to these questions to improve your own workflow, teamwork and efficiency!

A good example of an organization that follows reproducible principles not only for their own analyses but also in their journalistic templates is 538.com. They write and host stories and opinion pieces covering poll analyses, politics, economics, health, popular culture, and sports. The founder, Nate Silver, and the 538 team are best known for their political polling and forecasting during the United States Presidential and related elections since 2008.

Most of their articles provide references and links to their original data sources plus details on how their figures, analyses and statistical models were developed. They also host their data, code and details behind their analyses on Github <https://github.com/fivethirtyeight/> available to the public.

We’re going to work with some of these datasets in our exercises later in this course using the “fivethirtyeight” R package <https://cran.r-project.org/web/packages/fivethirtyeight/>

It is also worth mentioning Andrew Flowers who used to work for 538.com. He gave a great presentation at the 2017 RStudio conference on how to tell stories using data. In his presentation, he highlighted the various aspects of "data journalism" and importance of workflow, data processing and transparency in analysis and communication. These are all key aspects of reproducibility. Andrew Flowers is also a contributor to the `fivethirtyeight` R package.

As mentioned in lesson 01, the Journal of Biostatistics, encourages reproducible practices by their submitting authors. When articles are accepted for publication, they are given special markings based on the level of reproducibility achieved:

* The designation of \*\*D\*\* is given if the data on which the article is based are freely available,
* A designation of \*\*C\*\* is given if the authors’ code is freely available,
* and the “gold star” designation of \*\*R\*\* is given if both data and code are available, and the Associate Editor for Reproducibility is able to use them to reproduce the results in the paper.

An example of an article given the \*\*R\*\* designation for a fully reproducible article is one published in 2009 entitled “Air pollution and health in Scotland: a multicity study.” To see the \*\*R\*\* marking, you have to download the PDF to see the print version with the R placed at the top right in a bold box.

Finally, I want to highlight a great case study which illustrates the power and importance of reproducible workflow principles. The case demonstrates how quickly a treatment can be found when applying these principles in practice. The case surrounds the outbreak of a deadly strain of E.Coli bacteria which killed 50 people in Europe in 2001.

Researchers at BGI (the Beijing Genomics Institute) worked in collaboration with the Medical Center in Hamburg-Eppendorf to rapidly sequence the genome of the pathogen. Given the severity of the outbreak, the team announced and released the genome via Twitter to the world-wide community of microbial genomicists.

A Github repository was established <https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki> to "crowdsource" analysis and research to find a treatment.

People started contributing their work in under **\*\*24 HOURS\*\*** and within **\*\*5 DAYS!!\*\*** a bacterial agent was proposed to kill the pathogen. **This case, highlights the importance of these methods and work practices not only for speed and efficiency but also for rapidly addressing problems and developing solutions that can save lives.**

Next in lesson 04 we will begin breaking down the components of reproducible templates.

Side thoughts – not part of script…..

show some examples of what is possible with reproducible documents and templates.

\* organization

\* documentation

\* faster, less re-doing of your own work or others (reinventing the wheel)

\* better teamwork, efficiency, transparency - work break out, alignment of skills sets (divide and conquer) - clearer team roles and task delegation

\* breakdown of components - those that stay the same - those that change and how often

\* documentation of workflow

\* documentation of supporting components, data, files, etc

\* avoid documentation "gaps" - avoid mistakes by novices or those unfamiliar with the project, those outside the team-members

\* documentation of code, software, platform versions - evolution over time - operating systems, operating environments can change or be incompatible or cause "unforced errors" (ok in Windows-7 but "bugs" in Windows-10)