# Introduction to Reproducible Research via R Markdown

Lei Huang

Center for Research Informatics, University of Chicago

February 07, 2019

## Checklist

https://bit.ly/2MPikd8

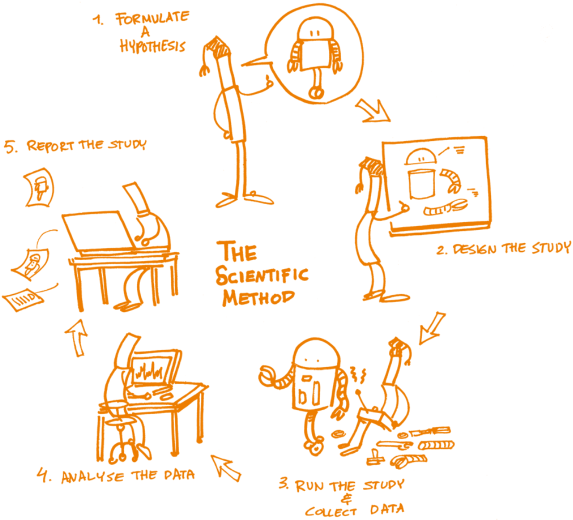
1. Latest R 3.5.2
2. Latest RStudio Desktop
3. Latest pandoc
4. Git
5. R package **packrat**
   - `install.packages("packrat")`

# Reproducibility

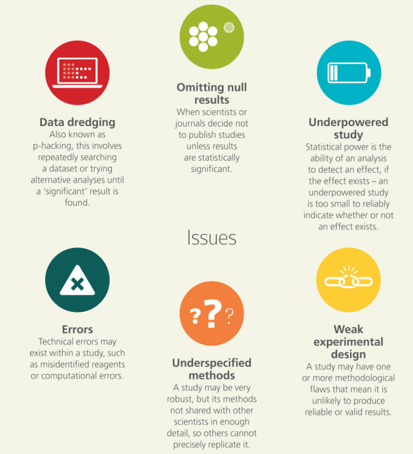
- The ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator.

  —*Goodman, Fanelli, & Ioannidis (2016)*

# The cycle of scientific research



# Reproducibility and the conduct of research



Issues

**Data dredging**
Also known as p-hacking, this involves repeatedly searching a dataset or trying alternative analyses until a 'significant' result is found.

**Omitting null results**
When scientists or journals decide not to publish studies unless results are statistically significant.

**Underpowered study**
Statistical power is the ability of an analysis to detect an effect, if the effect exists – an underpowered study is too small to reliably indicate whether or not an effect exists.

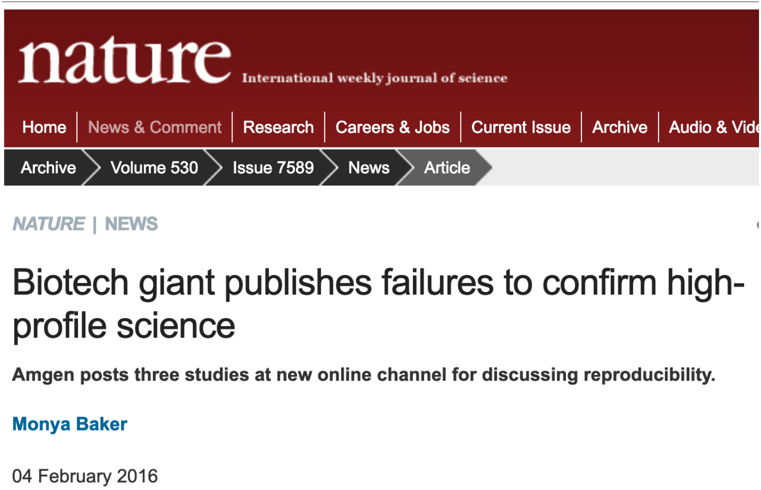
**Errors**
Technical errors may exist within a study, such as misidentified reagents or computational errors.

**Underspecified methods**
A study may be very robust, but its methods not shared with other scientists in enough detail, so others cannot precisely replicate it.

**Weak experimental design**
A study may have one or more methodological flaws that mean it is unlikely to produce reliable or valid results.

Source: Reproducibility and reliability of biomedical research: improving research practice

# Reproducible research crisis



Biotech giant publishes failures to confirm high-profile science

**Amgen posts three studies at new online channel for discussing reproducibility.**

Monya Baker

04 February 2016

# Reproducible research crisis

Studies show a very low reproducibility for articles published in scientific journals, often as low as 10-30%.

Here is a partial list:

- The biotech company Amgen had a team of about 100 scientists trying to reproduce the findings of 53 “landmark” articles in cancer research published by reputable labs in top journals.
  Only 6 of the 53 studies were reproduced (about 10%).
- Scientists at the pharmaceutical company, Bayer, examined 67 target-validation projects in oncology, women’s health, and cardiovascular medicine. Published results were reproduced in only
  14 out of 67 projects (about 21%).
- The project, PsychFileDrawer, dedicated to replication of published articles in experimental psychology, shows a
  replication rate 3 out of 9 (33%) so far.

# Reproducible research crisis

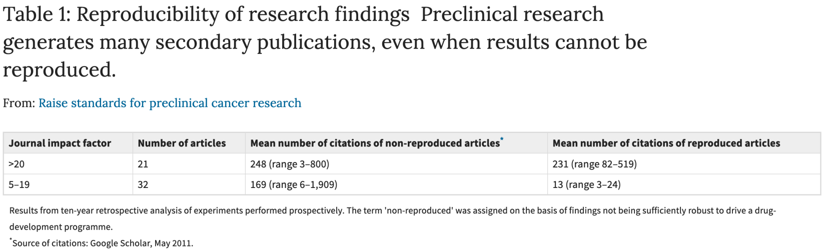


Table 1: Reproducibility of research findings  Preclinical research generates many secondary publications, even when results cannot be reproduced.

From: Raise standards for preclinical cancer research

| Journal impact factor | Number of articles | Mean number of citations of non-reproduced articles* | Mean number of citations of reproduced articles |
|---|---|---|---|
| >20 | 21 | 248 (range 3–800) | 231 (range 82–519) |
| 5–19 | 32 | 169 (range 6–1,909) | 13 (range 3–24) |

Results from ten-year retrospective analysis of experiments performed prospectively. The term 'non-reproduced' was assigned on the basis of findings not being sufficiently robust to drive a drug-development programme.

*Source of citations: Google Scholar, May 2011.

# How to make your work reproducible

- Commit to do it
- Keep track of things, perhaps with a version control tool
- Use software whose operation can be coded/automated
- Don't save output
- Save data in non-proprietary formats

# Some suggestions for conducting reproducible research

- Document everything in the analysis appropriately
- Don't do things by hand
  - Clean up spreadsheet data
  - Edit tables or figures
  - Download data from web sites by clicking the links
  - . . .
- Be careful when using GUI-based data analysis software

# Some suggestions for conducting reproducible research

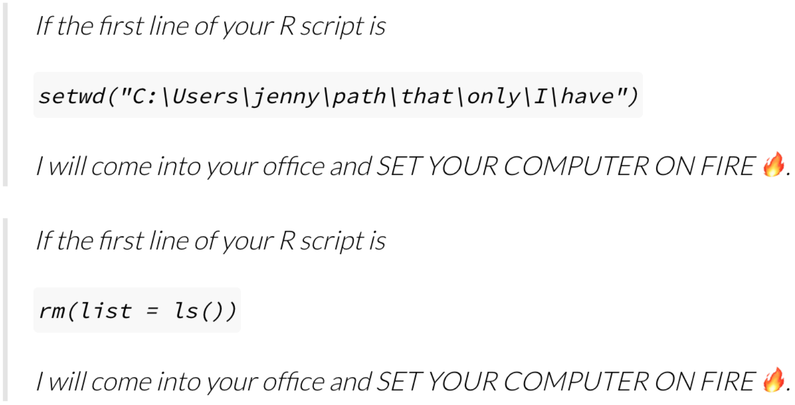
- Avoid saving data analysis output (tables, figure, processed data, etc.) except perhaps temporaritly for efficiency purposes. Try regenerate them on the fly with the same setting in previous analysis
- Save the data + code that generated the output, rather than the output itself
- Intermediate files can be saved as long as there is clear documentation of how they were created

# Some suggestions for conducting reproducible research in R

Have you ever started your R code with the following lines?

```
setwd("/Users/me/my_project_path")
rm(list = ls())
```

# Some suggestions for conducting reproducible research in R



> *If the first line of your R script is*
>
> ```
> setwd("C:\Users\jenny\path\that\only\I\have")
> ```
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.*

> *If the first line of your R script is*
>
> ```
> rm(list = ls())
> ```
>
> *I will come into your office and SET YOUR COMPUTER ON FIRE 🔥.*

—Jenny Byran

# Some suggestions for conducting reproducible research in R

- Use **here** package

```
install.packages("here")

library(here)
here("data", "file_i_want.csv")
```

- Restart a new R session

# Some suggestions for conducting reproducible research in R

- Preserve the package enviroment
  - provide sessionInfo()
  - use R package version management , e.g. **packrat** package from RStudio and **checkpoint** package from Microsoft R
  - use R workflow packages such as **ProjectTemplate**, **workflowr**, and **drake**
  - Containerize the enviroment with **VirtualBox** and **Docker**

# Some suggestions for conducting reproducible research in R

- Set your seed
  - Random number generators generate pseudo-random numbers based on an initial seed (in R use the `set.seed()` function)
  - Whenever you generate random numbers for non-trivial purpose, **alway set the seed**

# What is Markdown

- A lightweight markup language with plain text formatting syntax, created by John Gruber and Aaron Swartz
- Allows one to focus on writing as opposed to formatting
- Used to convert text to HTML (and other formats)
- More information at https://daringfireball.net/projects/markdown/

# What is Markdown

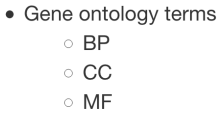
### HTML code

```
<body>
<h4>Gene ontology terms</h4>
  <ul>
    <li>BP</li>
    <li>CC</li>
    <li>MF</li>
  </ul>
</body>
```

### Markdown code

```
Gene ontology terms
  * BP
  * CC
  * MF
```

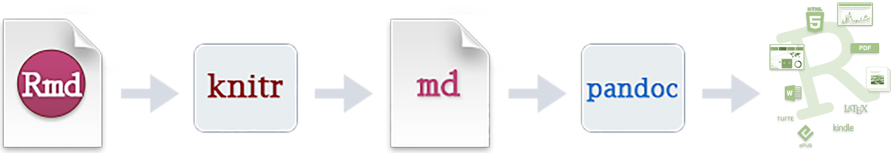
### Markdown rendering



- Gene ontology terms
    - BP
    - CC
    - MF

# What is R Markdown

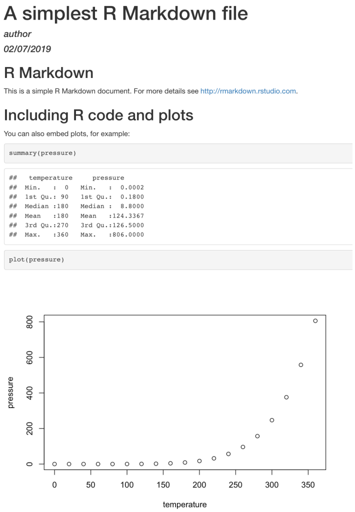
- An extension to Markdown by integrating R code (and other programming languages) and Markdown text.
- R code is evaluated as part of the processing of the markdown
- The output from R code is inserted into markdown document
- Generate high quality reports that can be shared with others
- A core tool in **literate statistical progamming**



# What is R Markdown

## Code for a R Markdown file:

~~~
---
title: "A simplest R Markdown file"
author: "author"
date: "02/07/2019"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is a simple R Markdown document. For more details see
<http://rmarkdown.rstudio.com>.

## Including R code and plots

You can also embed plots, for example:

```{r pressure, echo=TRUE}
summary(pressure)
plot(pressure)
```
~~~



A simplest R Markdown file

*author*

*02/07/2019*

R Markdown

This is a simple R Markdown document. For more details see http://rmarkdown.rstudio.com.

Including R code and plots

You can also embed plots, for example:

```
summary(pressure)
```

```
##   temperature     pressure
##  Min.   :  0   Min.   :  0.0002
##  1st Qu.: 90   1st Qu.:  0.1800
##  Median :180   Median :  8.8000
##  Mean   :180   Mean   :124.3367
##  3rd Qu.:270   3rd Qu.:126.5000
##  Max.   :360   Max.   :806.0000
```

```
plot(pressure)
```

# What is knitr

- An R package written by Yihui Xie https://cran.r-project.org/web/packages/knitr/index.html
- Support R Markdown, LaTex, and HTML as documentaiton languages
- Can output PDF, HTML, MS Word and other format files
- Integrated with RStudio IDE, but can be used independently from command line in R enviroment

# What is knitr good for

- Software documentation
- Short/medium-length technical documents
- Tutorial, homework, exam
- Reports (especially those need to be updated periodically)
- Data preprocessing documents or summaries
- Books (with **bookdown**) and blogs (with **blogdown**), etc.

# What is knitr NOT good for

- Time-consuming computations
- Documents that require precise formatting

# Resources for reproducible research

- R Markdown The Definitive Guide
- Rstudio Rmarkdown cheatsheet
- Happy Git and GitHub for the useR