

# Pragmatic Reproducible Research for Data Scientists

Biomedical Data Science Day  
February 4, 2020

Luke Rasmussen, MS  
Clinical Research Associate  
Division of Health and Biomedical Informatics  
Department of Preventive Medicine

[@lrasmus](#)

[luke.rasmussen@northwestern.edu](mailto:luke.rasmussen@northwestern.edu)

[github.com/lasmus](https://github.com/lasmus)

<https://github.com/StatTag/pragmatic-reproducible-research/tree/master/bdsd-2020>

# Acknowledgements

- This represents joint work with:
  - Dr. Leah J. Welty, Project Director
  - Abigail S. Baldridge, Biostatistician
  - Eric Whitley, Developer
- Supported in part through a Clinical Translational Sciences Award presented by National Institutes of Health to Northwestern University Clinical and Translational Sciences Institute (UL1TR001422).



<https://sites.northwestern.edu/stattag/>

<https://github.com/StatTag>

# Overview of Reproducible Research

# Introduction

Origins lie in the inconvenience of irreproducible research

“In the mid-1980s, we realized that our laboratory’s researchers often had difficulty reproducing their own computations without considerable agony.”

Schwab M, Karrenbach M, Claerbout J. “Making Scientific Computations Reproducible”. Computing in Science & Engineering 2000 2:6, 61-67

## MAKING SCIENTIFIC COMPUTATIONS REPRODUCIBLE

*verify a research paper’s computational results, readers typically have to recreate them from scratch. ReDoc is a simple software filing system for authors that lets readers easily reproduce computational results using standardized rules and commands.*

In the mid-1980s, we realized that our laboratory’s researchers often had difficulty reproducing their own computations without considerable agony. We also noticed that junior students, who typically build on the work of more advanced students, frequently spent a great deal of time and effort just to reproduce their colleagues’ computational results.

Reproducing computational research poses challenges in many environments. Indeed, the problem occurs wherever people use the traditional methods of scientific publication to describe computational research. For example, in a

cent progress in electronic publishing) they can only recompute the results by invoking the various programs exactly as the author invoked them; such information is something that is usually undocumented and difficult to reconstruct.

To address these problems, we developed ReDoc, a system for reproducing scientific computations in electronic documents. Since implementing it in the early 1990s, ReDoc has become our principal means for organizing and transferring our laboratory’s scientific computational research.

ReDocs are best defined operationally: After

# Kinds of Reproducible Research

- Computational Reproducibility
  - Have all of the information about the code, software and hardware.
  - Can re-run it on the same data set
- Empirical Reproducibility
  - Able to reproduce the non-computational experiments
  - Requires very clear methods
- Statistical Reproducibility
  - Full details about statistical tests
  - Pre-registration of studies to prevent p-value hacking

# Definition For Today

“...the ability to recompute data analytic results given an observed dataset and knowledge of the data analysis pipeline...”

Leek & Peng, PNAS February 10, 2015; 112 (6)

<https://doi.org/10.1073/pnas.1421412111>

# What Isn't Reproducible Research?

- There are many best practices and concepts that can be confused with it:

## Open Science

- Open data
- Open source
- Preregistration of studies
- Prepublication servers

Reproducible research can be disseminated openly.

## Replication

- “*...the chances other experimenters will achieve a consistent result...*” – Leek & Peng (2015)

Reproducible research can help others replicate your work.

## Tools / Software Engineering

- Unit tests
- Bash scripts
- Any specific programming language (Python, C++)
- Any specific tool (Jupyter Notebooks, Docker)

Reproducible research can be automated.

## Scientific Rigor

- Are the methods used the right ones?
- Quality of the data collected and used
- “*...invalid reports can do more harm than irreproducible reports*” – Shiffren, Borner & Stigler 2018

Reproducible research can help us evaluate our methods.

# Pragmatic Benefits

- Allows us to go back and know exactly what we did
  - 2015 Luke – “I’ll totally remember where I put that data set”
  - 2020 Luke – “AGGGH, 2015 LUKE!!!”
- Guides us to being better scientists
  - Forces us to think through parts of the process
  - Still doesn’t guarantee accuracy
- A little can go a long way
  - Not an all-or-nothing endeavor
- Benefits come with a cost
  - Willingness to change

Reproducible Research Is...

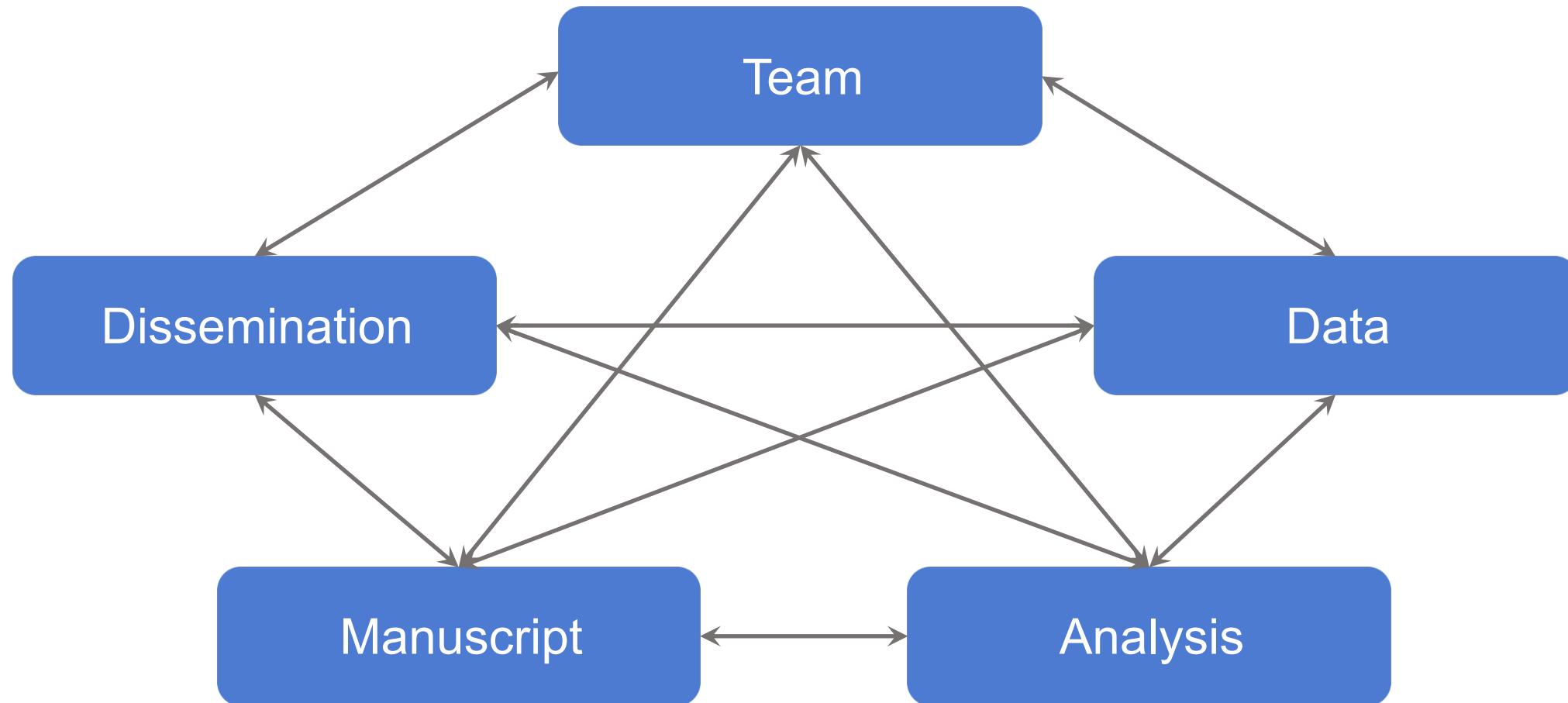


# Why Is Reproducibility Difficult?

- People
- Data
- Project-specific analysis code
- Supporting software – statistical package, compiler, interpreter
- Operating system
- Hardware

It is about reducing or controlling **variation**

# It's a Complicated Process



# Pragmatic Reproducible Research Teams

# Pragmatic RR for Teams

## Why It's Tricky

- Teams are variable
  - Who is on it (even N=1)
  - What tools / workflows people like to use
- RR is a cultural change
  - Be the leader
  - Don't worry if you can't win them all over

## What You Can Do

- Have a structure and describe it
  - There is no universal “best” structure
- Write down what you do
  - You'll need this for your Methods section anyway

# Structuring Your Project

Having a central storage location and systematic file/path structure: (1) assists BCC faculty and staff who are collaborating on the same project; (2) protects against transitions in a project team; and (3) provides a long term record of work provided to investigators.

## Location:

BCC Project Folders should reside in the following location:

- <\\fsmresfiles.fsm.northwestern.edu\fsmresfiles\PrevMed\Projects\BCC\Projects>

## Project Folder Naming Convention:

Within the above folder, the Project Folder should follow this naming convention:

PILastNameFirstInitial\_REDCap Project Number

Example: For PI Jane Doe, and REDCap Project Number 123, the file name should be DoeJ\_123

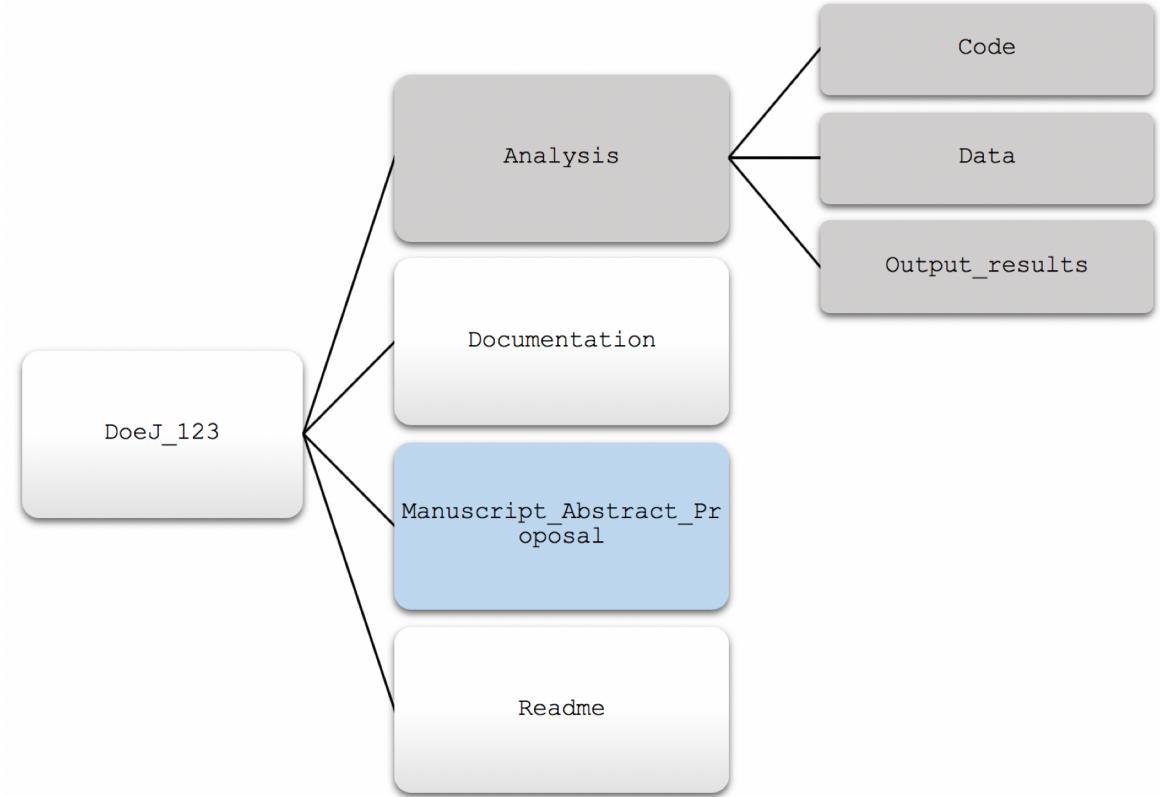
- The folder name does not include the project description - this avoids lengthy path/folder names. Project descriptions can be found in the Readme subfolder.
- Do not include leading zeros in the Project Number.

**The diagram which follows illustrates the recommended folder structure and substructure.**

**Additional folders may certainly be added as needed.**

## Date Requirements for Code/Programs/Manuscripts

Code/programs/manuscript files should include date information either inside the programs or as part of the file names. The code header should also include at the minimum the authors' names and the date last updated.



# Pragmatic Reproducible Research

## Data

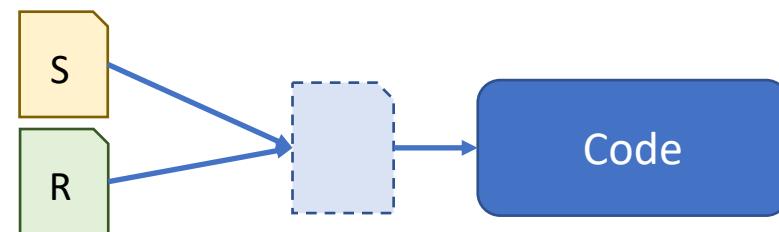
# Pragmatic RR for Data

## Why It's Tricky

- Data is variable
  - Changes - updates from EHR/EDW
- Provenance matters
  - Local file
  - Institution database server
  - External API
- Distribution and storage
  - Sensitivity of data

## What You Can Do

- Human checks
  - Document date accessed, version, etc.
- Computable checks
  - Checksum (e.g., MD5 hash)
  - Assert Your Assumptions (AYA)
- Design for missingness
  - Well-described placeholder



# Pragmatic Reproducible Research

## Analysis

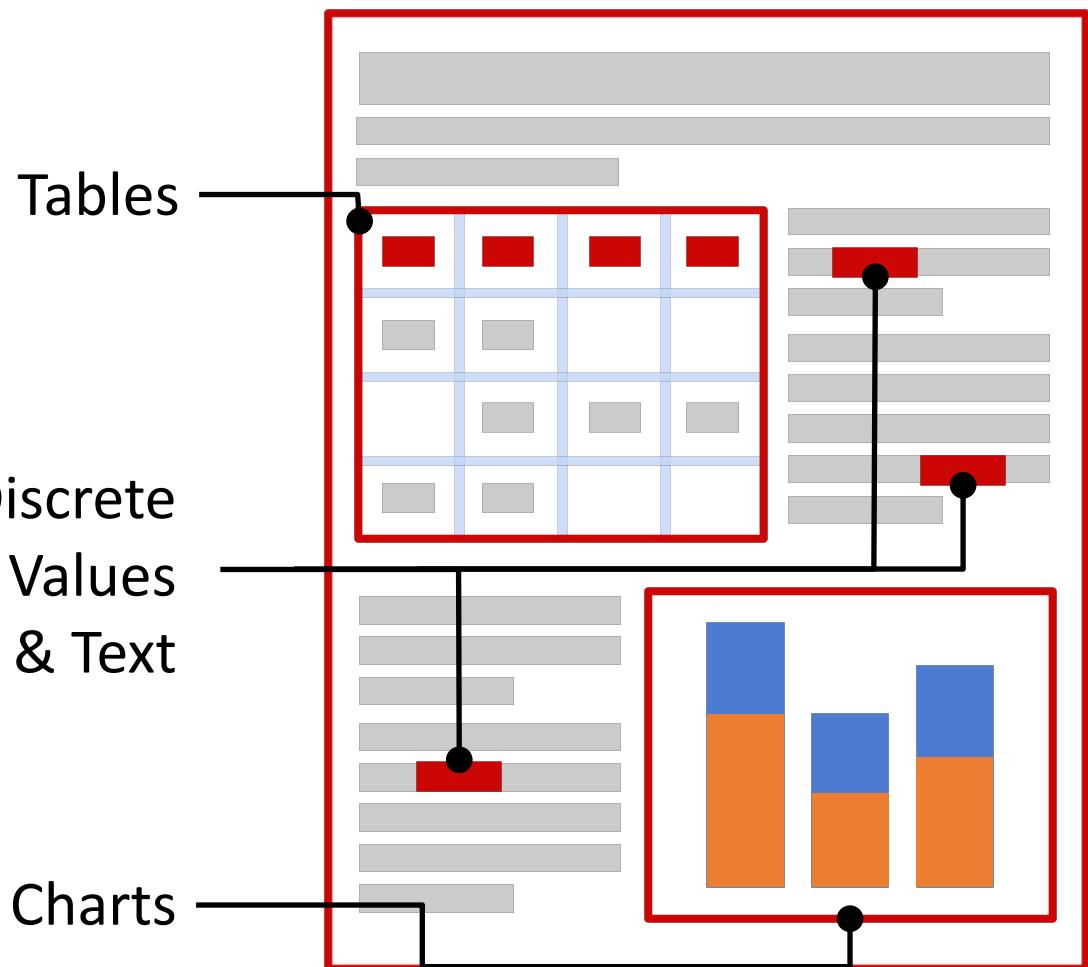
# Lab Notebook Platforms

“Weave” code chunks and text into a dynamic document

jupyter &



These tools provide an interactive environment that allows you to run multiple languages in a single session – with several forms of visual output



# Lab Notebook Platforms

## Acts as a Process Pipeline

- Automates & consolidates discrete processes
- Easy to review – highly visual

## Shows Provenance

- Trace your work
- See unsuccessful paths
- Trace the ultimate right path

## Supports Exploration and Creativity

- Interactive – feedback for faster, easier work
- Doesn't mandate “perfection” – keep *everything* - that's part of the value



# Other Pragmatic Lab Notebook Platforms

The screenshot shows a Microsoft Word document titled "NorthwesternSetup.md". The document contains two code snippets. The first snippet is under the heading "## Objectives" and discusses FHIR server constraints. The second snippet is under "## Build Notes" and details the build process for a .NET Core application, including commands like "dotnet add", "dotnet publish", and "dotnet target". A rich text editor toolbar is visible at the top, and the Microsoft Word ribbon is at the bottom.

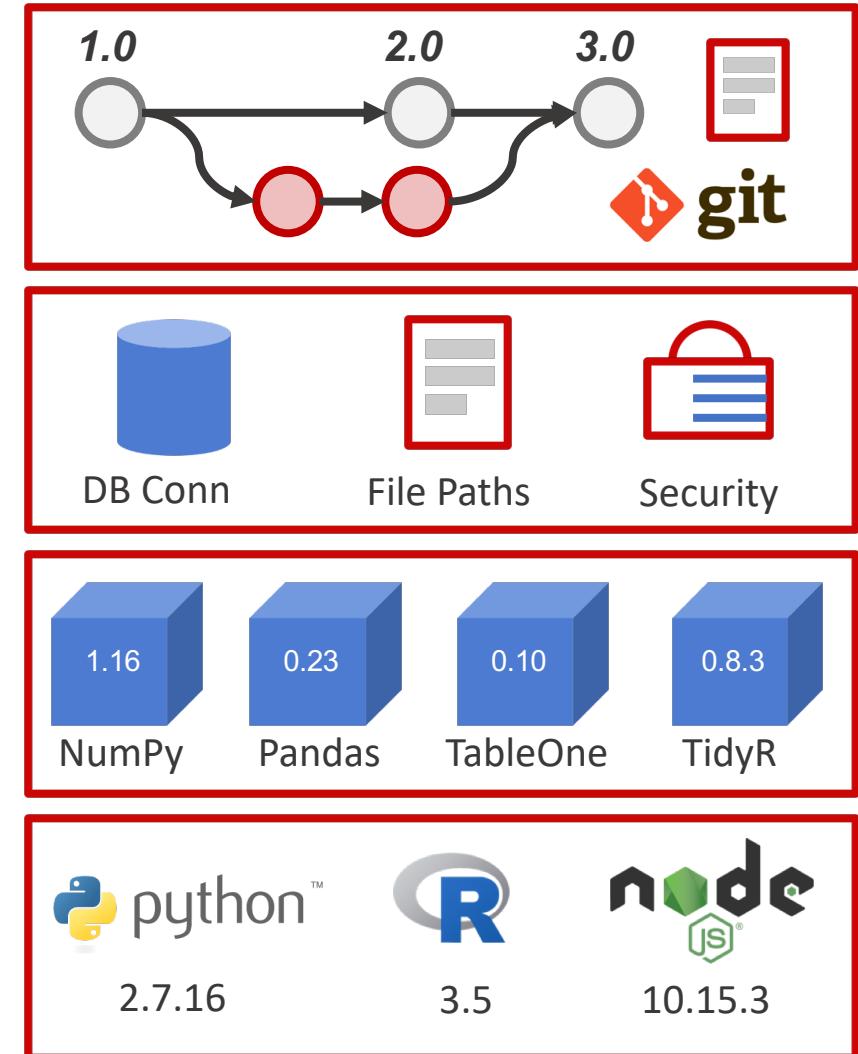
Microsoft Word

The screenshot shows a Windows Notepad window titled "Hello.txt - Notepad". It contains the text "Hello from Notepad!". The Notepad interface includes a menu bar (File, Edit, Format, View, Help) and status bar (Ln 1, Col 20, 100%, Windows (CRLF), UTF-8).

- Having the discipline to document
- Your environment
  - What you did
  - What you expected
  - What you got

# Analysis

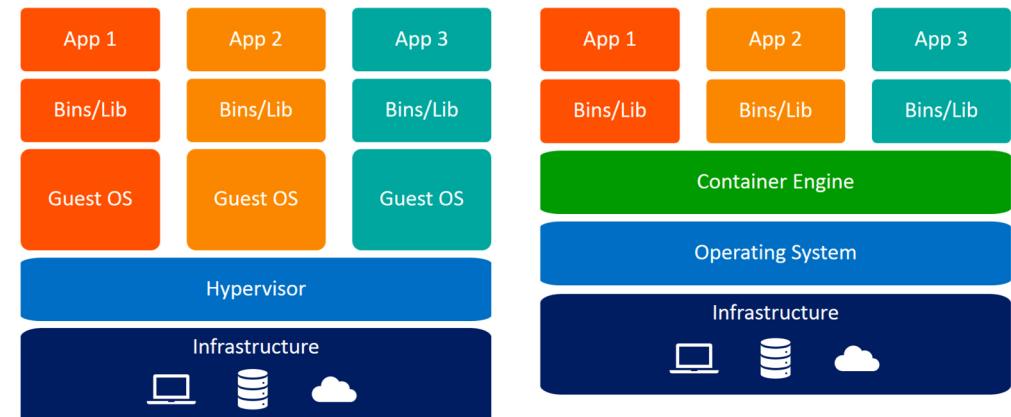
- Think of your work as a series of **variables**
  - + Source code revision
  - + System / application / user configuration
  - + Module / package versions
  - + Language / framework versions
- The variables are **combined** to create a single overall analysis pipeline
- A change to even a *single* variable can influence reproducibility



# Automating Reproducible Analyses

- Scripted execution
  - Guarantees (and documents) the same commands are run
  - Long-running processes... go get a cup of [coffee | tea]!
  - Running on my laptop today vs. 5 years from now

- Virtual machines and containers
  - Capture the software environment



<https://www.bmc.com/blogs/containers-vs-virtual-machines/>

# Pragmatic Reproducible Research

## Manuscripts

# Dynamic Documents: R Markdown

```
 NHANES Example.Rmd x
  ABC Knit
1 ---
2 title: 'Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014'
3 geometry: margin=.75in
4 output: html_document
5 sansfont: calibri Light
6 fontsize: 11pt
7
8 ---
9
10 <style type="text/css">
11 h1.title {
12   font-size: 11pt;
13   font-weight: bold;
14 }
15 </style>
16
17 ```{r setup, include=FALSE}
18 knitr::opts_chunk$set(echo = FALSE)
19 setwd("R:/NUCATS/NUCATS_Shared/BERDShared/Analysis Manager/Data and Programs/R")
20 analysis<-read.csv("Analysis.csv")
21 attach(analysis)
22 library(tableone)
```

## Association of Education with Anthropometrics in US Adults: National Health and Nutrition Examination Study 2013-2014

**Introduction:** Education level has been shown to be associated with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

**Methods:** This study included adult ( $\geq 30$  years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were assessed by questionnaire and anthropometric measurements were taken by study personnel. Education was dichotomized based on matriculation to postsecondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate linear regression.

Do you have non-technical collaborators who are willing to work this way?  
Think of a clinical expert working on a phenotype algorithm paper.

```
30
31 **Methods:** This study included adult ( $\geq 30$  years) participants from the
32 Examination Study (NHANES). Education and demographic information were assessed
33 measurements were taken by study personnel. Education was dichotomized based on
34 matriculation to postsecondary education. Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous
35 data. We examined the association between BMI and education level using multivariate linear regression.
36
37
38
39
40
41
```

(Beta: -0.19, 95% CI: -0.79 to 0.41).

**Table 1.** Association of Education with Participant Characteristics among 2013-2014 NHANES Participants

	No Postsecondary	Postsecondary	p
n	2159	2649	
Gender = Male (%)	1079 (50.0)	1208 (45.6)	0.003
Race (%)			<0.001
Mexican American	456 (21.1)	173 (6.5)	
Non-Hispanic Asia	161 (7.5)	411 (15.5)	

# Dynamic Documents and Track Changes

I create a dynamic document, generate the Word file and send it to collaborators.

They send back this

I have two choices:

1. Continue in Word, and loose the dynamic nature of the document.
2. Re-enter all of their changes in my source file.

## Association of Education with-and Anthropometrics in US Adults: Results from the National Health and Nutrition Examination Study 2013-2014

**Introduction:** Education level ~~may be has been shown to be associated~~ with body mass index (BMI). Gender, race, marital status and metabolic characteristics may alter this association.

**Methods:** ~~We studied~~This study included adult ( $\geq 30$  years) participants from the 2013-2014 National Health and Nutrition Examination Study (NHANES). Education and demographic information were ~~self-reported assessed by questionnaire~~ and anthropometric measurements were taken by ~~trained~~ study personnel. Education was dichotomized based ~~on matriculation to~~ post-secondary education (yes/no). Associations were estimated using T-tests or Wilcoxon rank sum tests for continuous data and chi-squared tests for categorical data. We examined the association between BMI and education level using multivariate linear regression.

~~4808 participants, We found that 2649 (55.1%) of the 4808 participants~~  
~~ost-secondary education. Participants with some post-secondary~~  
~~-secondary education was associated with significantly~~  
~~lower BMI (Beta: -0.08 to -0.21), although A~~  
~~fter adjusting for gender, race, age, marital~~  
~~status and total cholesterol, the association was no longer statistically~~  
~~secondary education was no longer significantly associated with BMI (Beta: 0.09 to 0.41).~~

Table 1. Association of Education with Participant Characteristics among  $n = 4808$  NHANES Participants

	No Postsecondary	Postsecondary	p
Gender = Male (%)	2159	2649	
Race (%)	1079 (50.0)	1208 (45.6)	0.003
Mexican American	456 (21.1)	173 (6.5)	<0.001

Leah J Welty A few seconds ago  
Leah, the spacing in this table is too far apart. Can you please make it single-spaced? And make the middle two columns wider?

Leah J Welty  
Formatted: Font: Bold

Leah J Welty  
Formatted: Font: Bold

Leah J Welty  
Formatted: Font: Bold

# Alternative Approach: StatTag

- StatTag creates a link between a statistical code file and a Word document
- Embeds output (values, tables, figures, verbatim) in document
- Can work separately on the code and the Word document but retain link
- Software agnostic: connects Stata, SAS or R code and Word document
- Can connect multiple code files (of different types) to the same document



<https://sites.northwestern.edu/stattag/>

# Conclusion

# Continue the Momentum

“The most important tool is the *mindset*, when starting, that the end product will be reproducible.”

– *Keith Baggerly*

There are a lot of opinions,  
but there is no wrong way

Identify what could change,  
and how to control for it

Start somewhere, but **start**

# Thank You!

<https://github.com/StatTag/pragmatic-reproducible-research/tree/master/bdsd-2020>

# Resources

- Karl Broman – Initial Steps Toward Reproducible Research (<https://kbroman.org/steps2rr/>)
- Victoria Stodden – Enabling Reproducible Research: Open Licensing for Scientific Innovation (March 3, 2009). International Journal of Communications Law and Policy, Forthcoming. Available at SSRN: <https://ssrn.com/abstract=1362040>
- Carl Boettiger - An introduction to Docker for reproducible research, with examples from the R environment. (2015) ACM SIGOPS Operating Systems Review, Special Issue on Repeatability and Sharing of Experimental Artifacts. 49(1), 71-79. DOI: 10.1145/2723872.2723882. arXiv:1410.0846 [cs.SE]
- Jeffrey Leek & Richard Peng - Opinion: Reproducible research can still be wrong: Adopting a prevention approach. PNAS February 10, 2015 112 (6) 1645-1646; <https://doi.org/10.1073/pnas.1421412111>
- Suzanne Bakken - The journey to transparency, reproducibility, and replicability. JAMIA, 26(3) 185-187. <https://doi.org/10.1093/jamia/ocz007>
- Leslie McIntosh et al., Repeat: a framework to assess empirical reproducibility in biomedical research. BMC Medical Research Methodology. 2017 17:143. <https://doi.org/10.1186/s12874-017-0377-6>
- Enrico Coiera et al. Does health informatics have a replication crisis? Journal of the American Medical Informatics Association, Volume 25, Issue 8, 1 August 2018, Pages 963–968, <https://doi.org/10.1093/jamia/ocy028>

# Resources

- Anaconda (<https://www.anaconda.com/distribution/>)
- Code Ocean (<https://codeocean.com/>)
- Docker (<https://www.docker.com/>)
- Dryad (<https://datadryad.org/>)
- Jupyter Notebook (<https://jupyter.org>)
- NLM UMLS (<https://uts.nlm.nih.gov>)
- Natural Language Toolkit (<https://www.nltk.org/>)
- Python (<https://www.python.org>)
- R (<https://www.r-project.org>)
- R Markdown (<https://rmarkdown.rstudio.com>)
- StatTag (<http://stattag.org>)