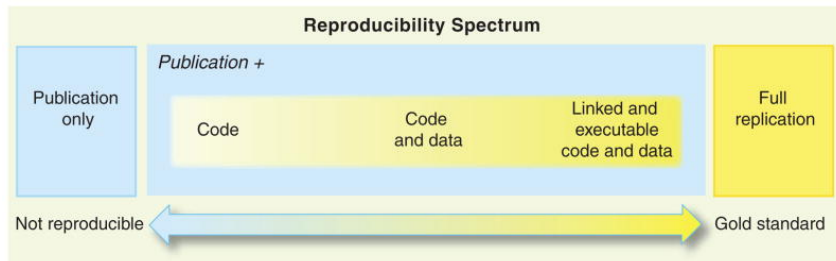


# An introduction to reproducible research for the R novice

497 / 597 Reproducible Research



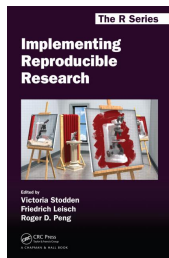
Richard Layton

Rose-Hulman Institute of Technology  
Fall 2018

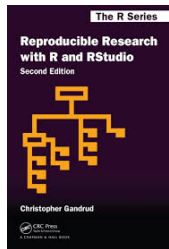
# Getting started

- ▶ Introductions
- ▶ Mystery question: *What is reproducible research?*

# Practitioners tell us:



*Research is reproducible when the data and the code used to obtain a finding are available and sufficient for an independent researcher to recreate the finding.*



- ▶ computational, data-intensive
- ▶ spans the full data, analysis, & publication workflow
- ▶ most of us have received only perfunctory training (if any)

# Events tell us:

More accountability is needed because of

- ▶ data falsification
- ▶ erroneous analysis
- ▶ misleading presentation of results



Karen EC Levy & David Merritt Johns, *When open data is a Trojan Horse: The weaponization of transparency in science and governance*, *Big Data and Society*, 2016.

# Reproduction revealed that their primary findings were false

Results were used to justify austerity policies, but the major effect disappeared after correcting for

- ▶ coding errors
- ▶ selective exclusion of available data
- ▶ unconventional weighting of summary statistics



Kenneth Rogoff & Carmen Reinhart

# Reproduction revealed that he falsified data

To obtain the results he wanted, he altered data in several ways. To date (2018) consequences include:

- ▶ clinical trials (real patients) terminated and 11 malpractice suits settled
- ▶ 18 research journal articles retracted
- ▶ Duke University must now obtain preapproval from NIH for funding changes



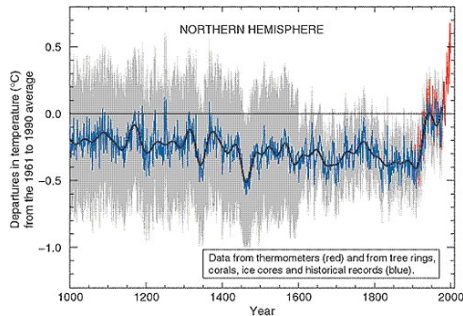
Anil Potti, formerly a cancer researcher at Duke University

## Reproduction, however, is also used to *discredit scientists*

Mann's work has withstood 15 years of scrutiny—and still holds up. But he refused to share.

*Scientists and “skeptics” are in a knife fight, and you don’t bring data to a knife fight.* — Paul Erlich

*Why should I make the data available to you, when your aim is to try and find something wrong with it?* — Phil Jones



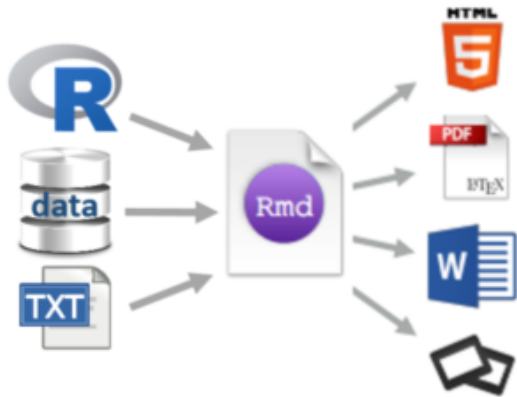
1000 years of temperature variation with uncertainties, Michael Mann

Freed Pearce, Climate change debate overheated after skeptic grasped 'hockey stick', *The Guardian*, 2010-02-09.

Brad Keyes, Mann retirement: Analysis, reax, *Climate Sceptic*, 2016-05-08.

Jeff Leek, De-weaponizing reproducibility, 2015-03-13.

Our focus is on explicitly linking the report, paper, or talk to the data and scripts that generate the findings



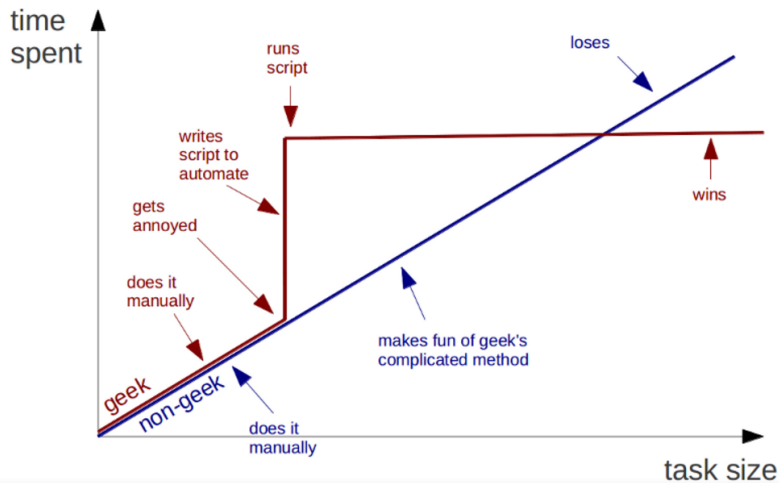
Changes are automatically updated and embedded in the output document.

Cut and paste no more!

R Markdown (.Rmd) files are the central link



# The primary (and probably most important) beneficiary is ... you



- ▶ reproducible for your future self
- ▶ faster updating of results
- ▶ faster resumption of work after hiatus

# Steps towards reproducibility

- ▶ Write scripts (avoid manual copy, paste, mouse-clicks)
- ▶ Plan the organization and naming scheme for files
- ▶ Strive for simplicity & readability
- ▶ Write for reusability & testability
- ▶ Agree on a workflow for collaborating before starting a manuscript
- ▶ DRY (don't repeat yourself)
- ▶ Link files explicitly
- ▶ Use version control
- ▶ Plan data management
- ▶ License your software
- ▶ Manage package dependencies

Karl Broman, [Initial steps toward reproducible research](#).

Jenny Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy Teal, and Greg Wilson, [Good enough practices for scientific computing](#), 2016-01.

# Steps towards reproducibility: in this course

- ▶ Write scripts (avoid manual copy, paste, mouse-clicks)
- ▶ Plan the organization and naming scheme for files
- ▶ Strive for simplicity & readability
- ▶ Write for reusability & testability
- ▶ Agree on a workflow for collaborating before starting a manuscript
- ▶ DRY (don't repeat yourself)
- ▶ Link files explicitly
- ▶ Use version control
- ▶ Plan data management
- ▶ License your software
- ▶ Manage package dependencies

Karl Broman, [Initial steps toward reproducible research](#).

Jenny Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, Tracy Teal, and Greg Wilson, [Good enough practices for scientific computing](#), 2016-01.

# Start installing the software today

Installation instructions are on the website



for data carpentry, analysis, and data graphics



our interface to all the software



for local version control

# How the course is organized

- ▶ Course materials are reproducible  
<https://github.com/DSR-RHIT/me497-reproducible-research>
- ▶ [Syllabus](#) for policies and procedures
- ▶ [Calendar](#) for agenda, assignments, and due dates