

Reproducible Research: What, Why, and How?

Thomas J. Leeper

Department of Political Science and Government
Aarhus University

October 28, 2014

Why?

What?

How?

Conclusion

- Reinhart and Rogoff

- Reinhart and Rogoff
- Psychology's “replication crisis”

- Reinhart and Rogoff
- Psychology's “replication crisis”
- “Most published research findings are false”

- Reinhart and Rogoff
- Psychology's “replication crisis”
- “Most published research findings are false”
- Diedrick Stapel

1 Why?

2 What?

3 How?

1 Why?

2 What?

3 How?

Why reproducible research?

- External reasons
- Internal reasons

Why?

What?

How?

Conclusion

External Reasons

External Reasons

- Philosophical perspective

External Reasons

- Philosophical perspective
- Journal requirements

External Reasons

- Philosophical perspective
- Journal requirements
- Funding agency requirements

External Reasons

- Philosophical perspective
- Journal requirements
- Funding agency requirements
- The coming revolution

Why?

What?

How?

Conclusion

Internal Reasons

Internal Reasons

- Confidence in your own work

Internal Reasons

- Confidence in your own work
- Easier workflow

Internal Reasons

- Confidence in your own work
- Easier workflow
- Easier collaboration

Why?

What?

How?

Conclusion

So what does that mean?

So what does that mean?



Open Science
@openscience



Following

"Reproducibility is collaboration with people you don't know, incl. yourself next week." –
[@philipbstark](#) #openscience



So what does that mean?

- 1 Do it for *yourself* first!
- 2 Do it for *science* second.

Why is research still irreproducible?

Why is research still irreproducible?

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

| Code | | Data |
|------|---------------------------------------|------|
| 77% | Time to document and clean up | 54% |
| 52% | Dealing with questions from users | 34% |
| 44% | Not receiving attribution | 42% |
| 40% | Possibility of patents | - |
| 34% | Legal Barriers (ie. copyright) | 41% |
| - | Time to verify release with admin | 38% |
| 30% | Potential loss of future publications | 35% |
| 30% | Competitors may get an advantage | 33% |
| 20% | Web/disk space limitations | 29% |

Why is research still irreproducible?

Why is research still irreproducible?

1 Technology

Why?

What?

How?

Conclusion



Why is research still irreproducible?

1 Technology

Why is research still irreproducible?

- 1 Technology
- 2 Individual actions

Why is research still irreproducible?

- 1 Technology
- 2 Individual actions
- 3 Collective behavior and norms

1 Why?

2 What?

3 How?

So what is reproducible research?

So what is reproducible research?

- Evolving standards and technology

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning

American Association for Public Opinion Research¹

Researchers must publish:

- 1 Research sponsor
- 2 Question wordings
- 3 Population, sampling frame, and sampling design
- 4 Sample sizes and margins of error
- 5 Dates of data collection

¹"Disclosure Standards"

American Psychological Assoc.²

- “After research results are published, **psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose**, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release...”
(8.14a)

²“Ethical Principles of Psychologists and Code of Conduct”

Assoc. for Psychological Science³

- 1 Sample sizes and exclusion criteria
- 2 Report all manipulations used
- 3 Report all outcomes analyzed

³"Submission Guidelines"

American Anthropological Association⁴

- “Anthropological researchers should seriously consider all reasonable requests for access to their data and other research materials for purposes of research. They should also make every effort to insure preservation of their fieldwork data for use by posterity.”

⁴“Code of Ethics”

CONSORT Group⁵

- “The checklist includes the 25 items selected because empirical evidence indicates that not reporting the information is associated with biased estimates of treatment effect, or because the information is essential to judge the reliability or relevance of the findings.”
- No requirement for open data or analyses

⁵“CONSORT Statement”

American Political Science Assoc.⁶

- “When statements that are challenged are based on reproducible data authors are obliged to facilitate replication.” (5.5)
- “Researchers making evidence-based knowledge claims should reference the data they used to make those claims. If these are data they themselves generated or collected, researchers should provide access to those data or explain why they cannot.” (5.6)
- “Production transparency” (6.2)
- “Analytic transparency” (6.3)

⁶“A Guide to Professional Ethics in Political Science”

European Research Council⁷

- “The European Research Council supports the basic principle of Open Access to research data. It therefore recommends to all its funded researchers that they follow best practice by retaining files of all the research data they have used during the course of their work, and that they be prepared to share this data with other researchers whenever it is not bound by copyright restrictions, by confidentiality agreements, or by contractual clauses.”

⁷“Open Access Guidelines for researchers funded by the ERC”

PLoS⁸

- “Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others.”

- Software created for use in publications must be open source

⁸“Editorial and Publishing Policies”

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning
- Hard to define

Why?

What?

How?

Conclusion

Irreproducibility

Irreproducibility

- Fabrication

Irreproducibility

- Fabrication
- Human error

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable
- “Available from the author”

Irreproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable
- “Available from the author (now deceased)”



Kaitlin Thaney
@kaythaney



Following

"'Reproducible research' is a redundant term. 'Irreproducible research' just used to be known as 'bullshit'." - [@fperez_org](#)
::slow clap::



RETWEETS

122

FAVORITES

61



6:11 PM - 8 May 2014

Distinguish from other concepts

Distinguish from other concepts

- *Reproducible* versus *Replicable*

Distinguish from other concepts

- *Reproducible* versus *Replicable*
- *Reproducible* versus *Automated*

Distinguish from other concepts

- *Reproducible* versus *Replicable*
- *Reproducible* versus *Automated*
- *Reproducible* versus *True*

Arrive at a definition

Stanford University's David Donoho:

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

Reproducible research
enumerates a complete set of
physical actions needed to
transforms transparent inputs
into outputs.

1 Why?

2 What?

3 How?

What makes up the ideal reproducible research product?

Past

- Data and method description
- Closed data and analysis
- Use of proprietary software
- Paywalled publications

Present

- Detailed or full protocols
- Data and analysis sharing (on request)
- Mix of proprietary and open software
- “Green” open access

Future

- Study preregistration and “outcome-blind” review
- Open lab notebooks
- Persistent, archived, open-licensed data
- Open source software
- Open peer review
- Open access publication
- Literate, reproducible output

How do you make your work more reproducible?

How do you make your work more reproducible?

Always think about your future self!

```
#DEAR FUTURE SELF,  
#  
# YOU'RE LOOKING AT THIS FILE BECAUSE  
# THE PARSE FUNCTION FINALLY BROKE.  
#  
# IT'S NOT FIXABLE. YOU HAVE TO REWRITE IT.  
# SINCERELY, PAST SELF
```

| DEAR PAST SELF, IT'S KINDA
| CREEPY HOW YOU DO THAT.

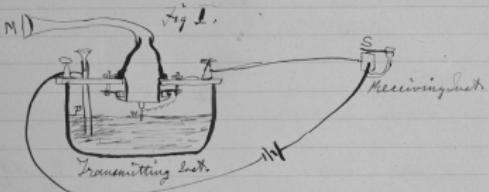
```
#ALSO, IT'S PROBABLY AT LEAST  
# 2013. DID YOU EVER TAKE  
# THAT TRIP TO ICELAND?
```

STOP JUDGING ME!



(1) Write Everything Down

40

March 10th 1876

1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece and S the armature of the Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting instrument was placed in another room and the doors of both rooms were closed.

I then shouted into M the following sentence: "Mr. Watson - come here - I want to

41

see you". To my delight he came and declared that he had heard and understood what I said.

I asked him to repeat the words - ~~He did~~ he answered "You said 'Mr. Watson - come here - I want to see you'". He then changed places and I listened at S while Mr. Watson read a few passages from a book into the mouth piece M. It was certainly the case that articulate sounds proceeded from S. The effect was loud but indistinct and muffled.

If I had read beforehand the passage given by Mr. Watson I should have recognized every word. As it was I could not make out the sense - but an occasional word here and there was quite distinct. I made out "to" and "out" and "further"; and finally the sentence "Mr. Bell Do you understand what I say? DO-YOU-UN-DER-STAND-WHAT-I-SAY" came quite clearly and intelligibly. No sound was audible when the armature S was removed.

(1) Write Everything Down

(1) Write Everything Down

- 1 Mark up your analysis files

(1) Write Everything Down

- 1 Mark up your analysis files
- 2 Write (and maintain) your research protocols

(1) Write Everything Down

- 1 Mark up your analysis files
- 2 Write (and maintain) your research protocols
- 3 Keep codebooks, questionnaires, and stimulus materials

(1) Write Everything Down

- 1 Mark up your analysis files
- 2 Write (and maintain) your research protocols
- 3 Keep codebooks, questionnaires, and stimulus materials
- 4 Try version control

(2) Get Organized

My dissertation folder

| | | | |
|---------------------------------------|--------------------------------------|---|---|
| Old Material | APSA2011 Handouts.pdf | APSA2011.aux | APSA2011.log |
| APSA2011.nav | APSA2011.out | APSA2011.pdf | APSA2011.snm |
| APSA2011.synctex.gz | APSA2011.tex | APSA2011.toc | APSR Reviews.txt |
| AQMW2011 Handouts.pdf | AQMW2011.aux | AQMW2011.log | AQMW2011.nav |
| AQMW2011.out | AQMW2011.pdf | AQMW2011.snm | AQMW2011.synctex.gz |
| AQMW2011.tex | AQMW2011.toc | Belief Importance, Content Pretest f... | Belief Importance, Content Pretest f... |
| Cengiz Erison Comments.txt | cert-noinfo.pdf | cert-repeat.pdf | cert-search.pdf |
| change-both.pdf | change-con.pdf | change-noinfo.jpg | change-noinfo.pdf |
| change-pro.pdf | change-repeat.jpg | change-repeat.pdf | change-search.jpg |
| change-search.pdf | Data Key 2011-03-20.doc | Data Key 2011-03-30.doc | Data Key 2011-05-04.doc |
| Dataverse Datafile.dta | dist.pdf | Druckman, Fein, Leeper Framing an... | Druckman, Fein, Leeper Framing an... |
| Druckman, Fein, Leeper Framing an... | Example Articles for Publication.doc | exp description.doc | fig-both1.pdf |
| fig-both4.pdf | fig-con1.pdf | fig-con4.pdf | fig-ctrl1.pdf |
| fig-ctrl4.pdf | fig-pro1.pdf | fig-prod.pdf | Framing and Biased Information Sea... |
| Framing and Biased Information Sea... | GoogleInsights-Healthcare.csv | GoogleInsights-Healthcare.xlsx | healthcare-mip.pdf |
| hovland.png | imp-noinfo.pdf | imp-repeat.pdf | imp-search.pdf |
| info-cert.pdf | Information Search Merged Data.dta | Means.xlsx | Methods Section 2011-07-29.doc |
| mip-analysis.r | mipdata.csv | mipdata-studyperiod.xlsx | MPSA2012.aux |
| MPSA2012.log | MPSA2012.nav | MPSA2012.out | MPSA2012.pdf |
| MPSA2012.snm | MPSA2012.synctex.gz | MPSA2012.tex | MPSA2012.toc |
| QR.png | R figures.r | References (partial).doc | Results Memo 2011-05-04.doc |
| Results Memo 2011-12-23.docx | Screenshot1-a.png | Screenshot-article.png | Screenshot-combined.png |
| SM Data 2011-030-30.xls | Supplemental Analysis.r | t2t3search.pdf | Tables 2011-12-23.docx |
| Tables 2011-12-28.docx | Tables.doc | Tables.xlsx | |

(2) Get Organized

- 1 Use a folder structure than can be shared

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
 - RawData.csv
 - CleanData.csv
 - Codebook.txt
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
 - GatherAndMerge.R
 - DataCleaning.R
 - Descriptives.R
 - Regression.R
 - Figures.R
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
 - Distributions.png
 - MarginalEffects.png
 - PredictedValues.png
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
 - Descriptives.tex
 - Regression.tex
 - MarginalEffects.tex
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
 - Draft.tex
 - References.bib
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
 - Slides.tex
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
 - Protocol.tex
 - StimulusMaterials.pdf
 - Questionnaire.txt
- README

Cataloging Information **DATA & ANALYSIS** Comments (0) Versions

Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Select all files

Total Number of Files: 10 Total Downloads: 48 Downloads of Files in This Version: 44

Codebook

| | | |
|---|--|------------------|
| <input type="checkbox"/> Study 1 (Lab) Key.docx application/octet-stream - 21 KB - 7 downloads |  Download | Study 1 data key |
| <input type="checkbox"/> Study 2 (Exit Poll) Key.docx application/octet-stream - 16 KB - 3 downloads |  Download | Study 2 data key |

Data

| | | |
|---|--|--|
| <input type="checkbox"/> Study 1 (Lab) Data.tab Tab Delimited - 54 KB - 18 downloads + analyses |  Download as...  Access Analysis + Subsetting | Study 1 Data  View Data Citation [+] |
| <input type="checkbox"/> Study 2 (Exit Poll) Data.tab Tab Delimited - 42 KB - 4 downloads + analyses |  Download as...  Access Analysis + Subsetting | Study 2 Data  View Data Citation [+] |

Experimental Materials

| | | |
|--|--|-------------------------------|
| <input type="checkbox"/> Con Pretreatment Articles.doc MS Word - 28 KB - 1 download |  Download | Con Pretreatment Articles |
| <input type="checkbox"/> Control Pretreatment Articles.doc MS Word - 23 KB - 1 download |  Download | Control Pretreatment Articles |
| <input type="checkbox"/> Pro Pretreatment Articles.doc MS Word - 28 KB - 2 downloads |  Download | Pro Pretreatment Articles |
| <input type="checkbox"/> Questionnaire.doc MS Word - 97 KB - 4 downloads |  Download | Questionnaire |

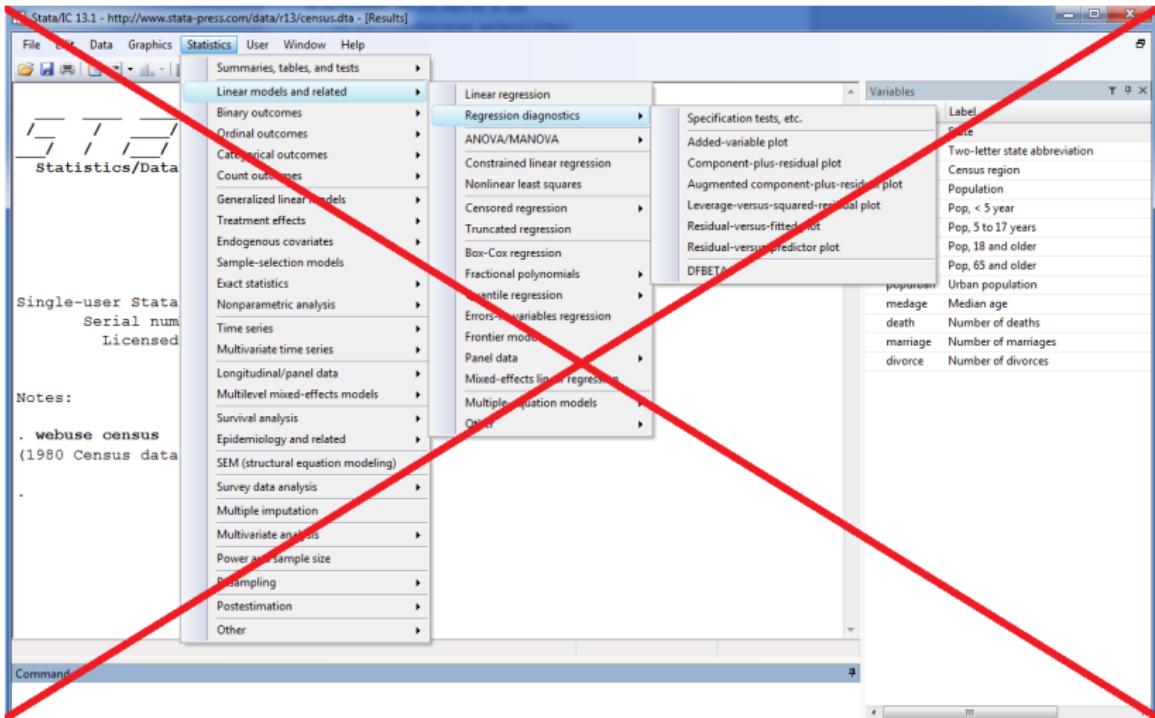
Replication code

| | | |
|---|--|--------------------------|
| <input type="checkbox"/> Study 1 (Lab).do Stata Syntax - 5 KB - 3 downloads |  Download | Study 1 Replication code |
| <input type="checkbox"/> Study 2 (Exit Poll).do Stata Syntax - 2 KB - 1 download |  Download | Study 2 Replication code |

(2) Get Organized

- 1 Use a folder structure than can be shared

- 2 Never use absolute file paths in code



(3) Abandon Point-and-Click

- 1 Don't clean data by hand
- 2 Use scripts rather than menus for graphics
- 3 Record your OS and software (and their versions)

(4) Publicly Archive Your Research

- 1 Use persistent, public archives, not your website or “on request”

Where do you archive your research?

- Dataverse Network
- Data Dryad
- figshare

(4) Publicly Archive Your Research

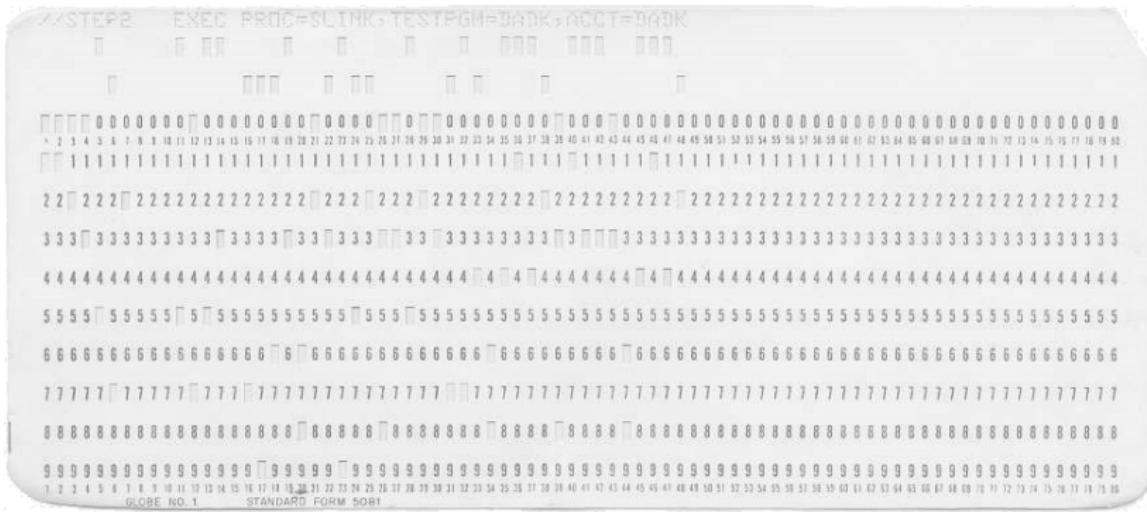
- 1 Use persistent, public archives, not your website or “on request”
- 2 Use Simple, Structured, and Semantic open file formats

Why?

What?

How?

Conclusion



1996 Codebook.pdf - Adobe Reader

File Edit View Window Help 23 / 38 101% Tools Fill & Sign Comment

Open

you describe your opinion of (INSERT ITEM: ROTATE ITEMS 8.-I. AND g.-I.) as very favorable, mostly favorable, mostly UNfavorable, or very unfavorable? (INTERVIEWERS: PROBE TO DISTINGUISH BETWEEN "NEVER HEARD OF" AND "CAN'T RATE")

| | | Very Favor- able | Mostly Favor- able | Mostly Unfavor- able | Very Unfavor- able | Never Heard of | Can't Rate |
|-------|--|------------------------|--------------------------|----------------------------|--------------------------|----------------------|---------------|
| (115) | a. Network television news (1-96) | 1 | 2 | 3 | 4 | 5 | 6 |
| (116) | b. Local TV news (1-96) | 1 | 2 | 3 | 4 | 5 | 5 |
| (117) | c. The daily newspaper you are most familiar with (1-96) | 1 | 2 | 3 | 4 | 5 | 6 |
| (118) | d. Congress (1-96) | 1 | 2 | 3 | 4 | 5 | 6 |
| (119) | e. Tobacco companies (7-94) | 1 | 2 | 3 | 4 | 5 | 6 |
| (120) | f. Labor unions (2-96) | 1 | 2 | 3 | 4 | 5 | 6 |
| (121) | g. Bill Clinton (2-96) | 1 | 2 | 3 | 4 | 5 | 6 |
| (122) | h. Hillary Clinton (2-96) | 1 | 2 | 3 | 4 | 5 | 5 |
| (123) | i. Bob Dole (2-96) | 1 | 2 | 3 | 4 | 5 | 6 |

(4) Publicly Archive Your Research

- 1 Use persistent, public archives, not your website or “on request”
- 2 Use Simple, Structured, and Semantic open file formats
- 3 Be explicit about data licensing

How to license data?



Attribution
CC BY



Attribution-NoDerivs
CC BY-ND



Attribution-NonCommercial-ShareAlike
CC BY-NC-SA



Attribution-ShareAlike
CC BY-SA



Attribution-NonCommercial
CC BY-NC

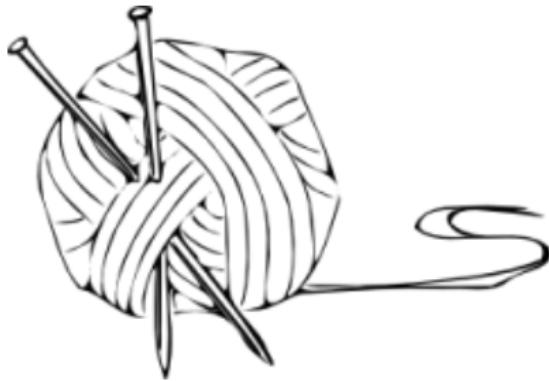


Attribution-NonCommercial-NoDerivs
CC BY-NC-ND

(4) Publicly Archive Your Research

- 1 Use persistent, public archives, not your website or “on request”
- 2 Use Simple, Structured, and Semantic open file formats
- 3 Be explicit about data licensing
- 4 Create useful metadata

(5) Learn Literate Programming



Learn
to knit
after
lunch!

Where to go next?

- rOpenSci
- “Challenges in Irreproducible Research”
- Karl Broman’s resources
- 2011 “Reproducible Research” conference slides
- “Six steps to a Better Relationship with Your Future Self.”
- “Ten Simple Rules for Reproducible Computational Research.”
- *Reproducible Research with R and RStudio.*
- Software Carpentry
- Johns Hopkins Data Science Certificate on Coursera

People to follow?

- @victoriastodden
- @carlystrasser
- @l_peer
- @OSFramework and @BrianNosek
- @RetractionWatch
- @UCBITSS
- @OpenScience

Reproducibility isn't everything

- Data archiving and data citation
- Open protocols and materials
- Methodological transparency
- Free and open-source software (FOSS)
- Open access

1 Why?

2 What?

3 How?

In the end...

- Be reproducible *for you*
- Science will benefit as a result

