



ROMANIAN NATIONAL
INSTITUTE OF
STATISTICS



BUCHAREST
UNIVERSITY OF
ECONOMIC STUDIES



ECOLOGICAL
UNIVERSITY OF
BUCHAREST



NICOLAE TITULESCU
UNIVERSITY OF
BUCHAREST



THE UNIVERSITY OF
BUCHAREST



The 4th International Conference

New Challenges for Statistical Software - The Use of R in Official Statistics -

7-8 April 2016, Bucharest, Romania, www.r-project.ro/conference2016/

RR BY R IN OFFICIAL STATISTICS

Ciprian Alexandru
Nicoleta Caragea

R-omania Team | www.r-project.ro

Topics

8

- ☐ Research where they are going?
- ☐ Replication or Reproducible?
- ☐ Why Reproducibility?
- ☐ What is necessary for Reproducibility?
- ☐ Research Pipeline
- ☐ Research about Reproducible Research
- ☐ Missing Reproducibility
- ☐ Literate (Statistical) Programming
- ☐ Research Pipeline in R
- ☐ Development Tools
- ☐ Presentation Tools
- ☐ Basic principles
- ☐ Markdown

Research where they are going?

9

- A lot of
 - ▣ Computing
 - ▣ Data analysis
 - ▣ Data manipulation or processing
 - ▣ Big Data
 - ▣ Communicating the results
 - ▣ Reconstructing
 - ▣ **Reproducible**



Replication or Reproducible?

10

- Replication – an independent study came with same conclusion as the original study

- Reproducible
 - ▣ original data and code are studied by an independent analysts obtaining the same results of the original study
 - ▣ *data analysis* can be repeated with same results

Reproducible \neq Correctness

11

- a reproducible study could be wrong
- a wrong assumptions can be identified by an independent replication process

Why not Replication research?

12

- Lack of time
- Lack of financial resources
- Unique

Why Reproducibility?

13

- ❑ Enhancing scientific evidence
- ❑ It is the only guarantee about a study
- ❑ Ensures transparency
- ❑ Creates confidence
- ❑ Validation of the data analysis

What is necessary for Reproducibility?

14

- Publication of data sources and programs (R code)
 - ▣ analytic data, analytic code, documentation, standard tools for distribution

- Conducting studies using independent data sources

How easy can do our work reproducible?

15

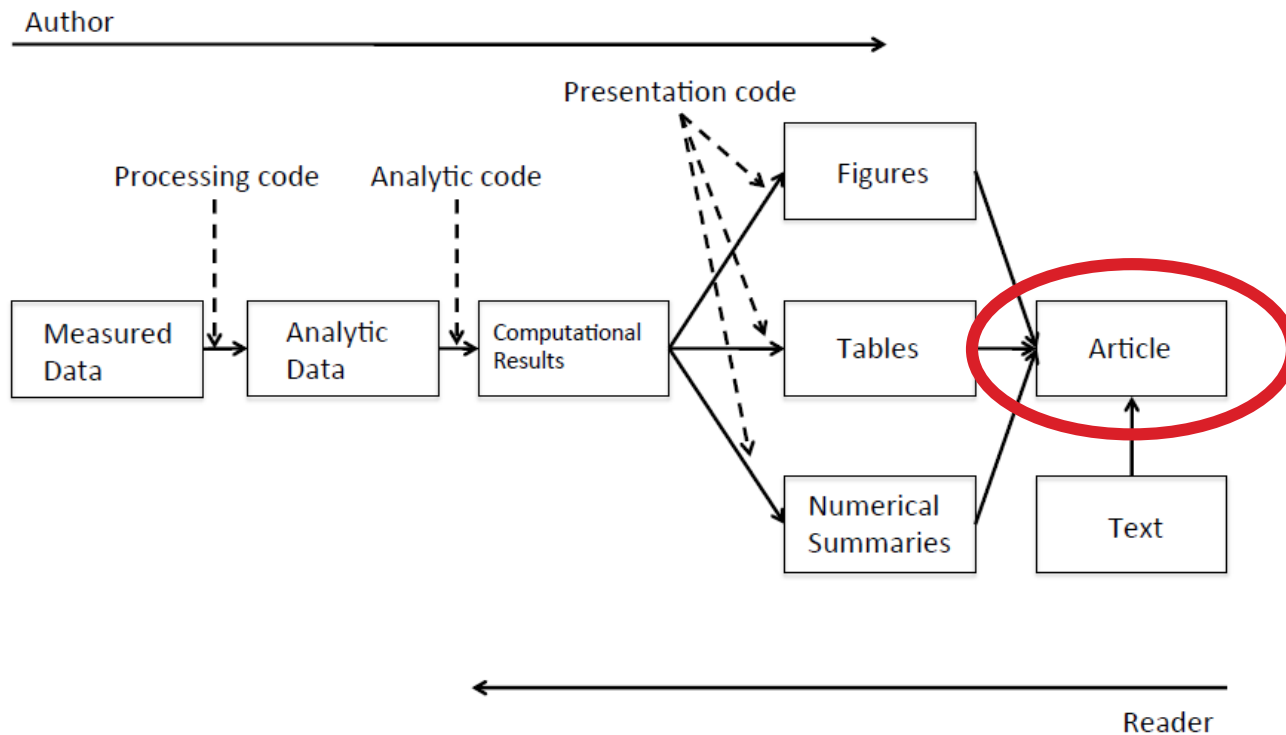
- ❑ preparation of data in a friendly format (standard)
- ❑ find a web server for uploading code & data with “almost” limitless availability
- ❑ using some tools to help readers (well known or well documented, eventually)
- ❑ making available a guidelines for future use

Conclusion:

a threshold between more “present time” consuming and saving for the future

Research Pipeline

16



Roger D. Peng, *Reproducible Research: Concepts and Ideas*

17

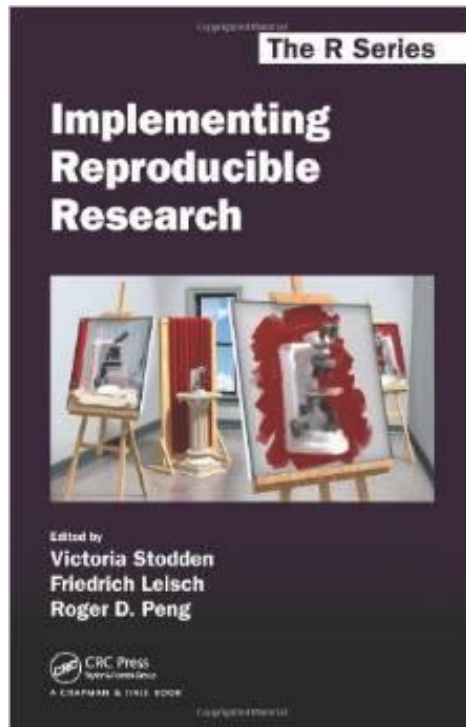


Reproducible Research in Computational Science

Computational science has led to exciting new developments, but the nature of the work has exposed limitations in our ability to evaluate published findings. Reproducibility has the potential to serve as a minimum standard for judging scientific claims when full independent replication of a study is not possible.

Research about Reproducible Research

18



Report Writing for Data Science in R



Roger D. Peng

Missing Reproducibility

19

□ Duke University



Lesson Learned

20

https://today.duke.edu/2014/02/reproducibility

SEARCH OR BROWSE ALL OF DUKE UNIVERSITY ▾

DukeTODAY

Search People, Places, Things

☒ Search Duke Today ☐ Search Duke Main

|

NEWS | OPINION | WORKING@DUKE | STUDENTS | ALUMNI

NEWS BY TOPIC ▾ | SERIES

To Teach Scientific Reproducibility, Start Young

Introducing concept to undergrads could lead to more transparency in science

February 28, 2014 | [print](#) | [f](#) | [t](#) | [g+](#) | [10](#)

ARTICLE

DURHAM, NC - The ability to duplicate an experiment and its results is a central tenet of the scientific method, but recent research has shown an alarming number of peer-reviewed papers are irreproducible.

A team of math and statistics professors has proposed a way to address one root of that problem by teaching reproducibility to aspiring scientists, using software that makes the concept feel logical rather than cumbersome.

Researchers from Smith College, Duke University and Amherst College looked at how introductory statistics students responded to a curriculum modified to stress reproducibility. Their work is detailed in a paper published Feb. 25 in the journal [Technological Innovations in](#)

MORE INFORMATION
CONTACT: Erin Weeks PHONE: (919) 681-8057
EMAIL: erin.weeks@duke.edu

HOME PAGE FEATURES



Krzyzewski Commencement



The Value of Journals



Von der Heyden Giff



Mike Merson


<https://today.duke.edu/2014/02/reproducibility>

The 4th International Conference - New Challenges for Statistical Software - The Use of R in Official Statistics | RR by R in Official Statistics | 7-8 April 2016 | Bucharest

Lesson Learned

21

<https://sites.duke.edu/techexpo/presentation/reproducible-computational-science-challenges-and-opportunities-for-research-and-it/>



About ▾ Event information Keynote Program Vendors

Reproducible Computational Science: Challenges and opportunities for research and IT

PRESENTERS: *Hilmar Lapp, Karen Cranston, Mine Çetinkaya-Rundel, Mark Delong, Erich Huang, Dan Leehr, Darin London, Paul Magwene*

DEPARTMENTS: *Center for Genomic and Computational Biology (GCB)*

Last Tweets

- Duke IT'ers — lend us your... opinion. Yes, even if you didn't attend Tech expo this year, we'd relish your feedback.
<https://t.co/1f8Zyt8pJb>, Apr 22
- DukeTechExpo presenters: send us your slides! <http://t.co/iOX1MkIHw0>, Apr 20
- So... how'd we do? Please take a survey to help conference planners improve the event next year.
<https://t.co/1f8Zyt8pJb>, Apr 20
- RT @BrynSmith5: At @DukeTechExpo listening to the diversity panel discussion #duketecheexpo
<http://t.co/tXhNhhdFXo>, Apr 17
- RT @dukemededit: Just won a Pebble watch @DukeTechExpo #ourluckyday
<http://t.co/ZiOlfaTm76>, Apr 17

TechExpo 2015 At-a-Glance

When: Friday, April 17, from 8 AM – 4:30 PM

<https://sites.duke.edu/techexpo/presentation/reproducible-computational-science-challenges-and-opportunities-for-research-and-it/>

Reproducibility in computational research

22

- ❑ Assign a unique ID for each version of released data & code
- ❑ Use open licensing for data & code
- ❑ Workflow tracking should happen during the research process
- ❑ In your publication, include a statement describing computing environment(s) and software version(s)
- ❑ Make data, code and methods available and accessible
- ❑ Version control
- ❑ Publish data, code and methods in non-proprietary formats (if possible)
- ❑ Cite any 3rd party data and code
- ❑ Follow data and code sharing guidelines for funded research

Jane Frazier, Data Librarian, Australian National Data Service, *Research Data & Output Management*

Stodden and Miguez. Best practices for computational science: software infrastructure and environments for reproducible and extensible research.

Journal of Open Research Software, 2014. [DOI:10.5334/jors.ay]

Yale Law School Roundtable on Data and Code Sharing. Reproducible research: addressing the need for data and code sharing in computational science.

Columbia University Academic Commons, 2010. [DOI:10.1109/MCSE.2010.113]

Literate (Statistical) Programming

23

- ❑ a research paper/an article = a stream of text and code chunks
- ❑ code chunks = loads and manipulate data, computes results, presents tables, figures etc.
- ❑ literate programs produce:
 - ▣ human-readable documents
 - ▣ machine-readable documents

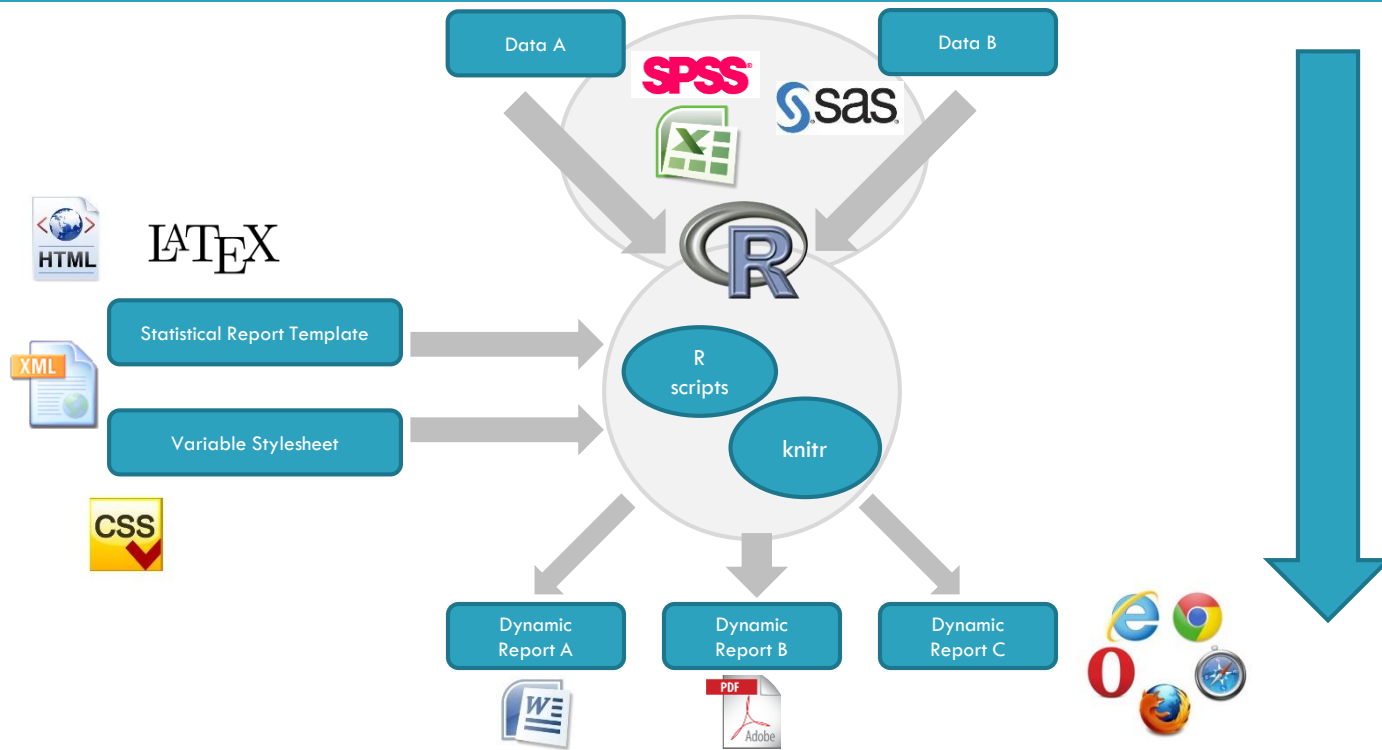
Where do we start?

24

- Using some tools
- Basic principles

Research Pipeline in R

25



Development Tools

26

- Documentation language (human readable)
 - ▣ Sweave ($\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$) [<http://leisch.userweb.mwn.de/Sweave/>]
 - ▣ knitr ($\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$, Markdown, HTML) [<http://yihui.name/knitr/>]
- Programming language (machine readable)
 - ▣ R
 - ▣ Python

Presentation Tools

27

- ❑ GitHub - powerful collaboration, code review, and code management for open source and private projects [<https://github.com/>]
- ❑ RPubS - web publishing from R [<https://rpubs.com/>]

Sweave

28

- ❑ Included in R basic installation
- ❑ Documentation language L^AT_EX
- ❑ Lacks features like caching, multiple plots, per chunk, mixing programming languages
- ❑ Not frequently updated
- ❑ Developed by Fritz Leisch, core member of R

<http://leisch.userweb.mwn.de/Sweave/>

knitr

29

- ❑ Contributed package (need installation)
- ❑ Documentation language:
 - ▣ L^AT_EX
 - ▣ Markdown
 - ▣ HTML
- ❑ Mix other programming languages
- ❑ Very popular package
- ❑ Developed by Yihui Xie while he was a graduate student at Iowa State

<http://yihui.name/knitr/>

Basic principles

30

“Golden Rule of Reproducibility: Script Everything”,
Roger D. Peng

Basic principles - Data

31

- Create an analysis folder unique for each research
- Raw data
 - ▣ Saved data - stored in your analysis folder
 - ▣ Web data – include url, description, accessed date
- Processed data
 - ▣ Named clearly (files and variable)
 - ▣ Make it tidy

Basic principles - Figures

32

- Exploratory figures
 - ▣ In general include all data set
 - ▣ Explore your data until they told you everything
 - ▣ They are not pretty, but very useful

- Final figures
 - ▣ Use the most representative figures
 - ▣ Axes/Labels/Colors are set to make the figures clear and easy understandable
 - ▣ Multiple panels is preferred

Basic principles – R code

33

- Raw / unused scripts
 - ▣ Small comments is recommended (our memory is not a HDD)
 - ▣ Use multiple versions
 - ▣ Keep all intermediary analysis, until you are sure that they are useless

- Final scripts
 - ▣ Clear commented (small in row, large for sections)
 - ▣ Include processing details
 - ▣ Keep only relevant analysis, that are included in final paper

- R markdown files
 - ▣ Used for generate reproducible reports
 - ▣ Integrates Text and R code

Basic principles – Text

34

- ❑ README files
 - ▣ Optional for R markdown
 - ▣ Include all instructions for analysis

- ❑ Text of analysis / report
 - ▣ Title, introduction, methods, results and conclusions
 - ▣ Include only relevant analysis regarding the conclusions
 - ▣ References are important

Coding Standards for R

35

- ❑ Use text files and editor
- ❑ Indent the code, 4 spaces minimum, or 8 (better)
- ❑ Width of the code – limits to 80 columns
- ❑ Functions – keep it small

Markdown

36

Italics

This text will appear italicized!

Bold

****This text will appear bold!****

Headings

This is a secondary heading

This is a tertiary heading

Markdown

37

Unordered Lists

- first item in list
- second item in list
- third item in list

Ordered Lists

1. first item in list
2. second item in list
3. third item in list

Markdown

38

Links

[Johns Hopkins Bloomberg School of Public Health] (<http://www.jhsph.edu/>)

[Download R] (<http://www.r-project.org/>)

[RStudio] (<http://www.rstudio.com/>)

Markdown

39

Advanced Linking

I spend so much time reading [R bloggers][1] and [Simply Statistics][2]!

[1]: <http://www.r-bloggers.com/> "R bloggers"

[2]: <http://simplystatistics.org/> "Simply Statistics"

Markdown

40

Newlines require a double space after the end of a line.

First line

Second line

First line Second line

First line

Second line

Markdown

41

- ❑ The Official Markdown Documentation
[<http://daringfireball.net/projects/markdown/basics>]
- ❑ Github's Markdown Guide
[<https://help.github.com/articles/github-flavored-markdown/>]

Romanian Social Data Archive

42

□ www.roda.ro



Reproducible Research - Economics

43



The diagram illustrates the combination of two resources to create reproducible research papers. On the left, the *RePEc* logo is shown with a blue glow effect. This is followed by a plus sign (+). To the right of the plus sign is the RODA logo, which consists of a green square containing a stylized orange and green leaf-like shape, with the text "RODA" above it and "ARHIVA ROMÂNĂ DE DATE SOCIALE" below it. To the right of the RODA logo is an equals sign (=). Finally, on the far right, the text "Reproducible Research Papers" is displayed.

$$\text{RePEc} + \text{RODA} = \text{Reproducible Research Papers}$$

GitHub

44

□ <https://github.com/>

GitHub

References

45

- ❑ The Real Reason Reproducible Research is Important [<http://simplystatistics.org/2014/06/06/the-real-reason-reproducible-research-is-important/>]
- ❑ Reproducible Data Analysis [<http://www.biomedicale.univ-paris5.fr/SpikeOMatic/ReproducibleDataAnalysis/ReproducibleDataAnalysis.html>]
- ❑ Get your R education going with GitHub [<http://www.r-bloggers.com/get-your-r-education-going-with-github/>]
- ❑ Everyone loves R markdown and Github; stories from the R Summit, day two [<http://www.r-bloggers.com/everyone-loves-r-markdown-and-github-stories-from-the-r-summit-day-two/>]
- ❑ Introducing the Reproducible R Toolkit and the checkpoint package [<http://www.r-bloggers.com/introducing-the-reproducible-r-toolkit-and-the-checkpoint-package/>]

Thank you!

46



Ciprian Alexandru
alexciopro@yahoo.com

https://github.com/alexciopro/iCMS2015_Keynote