Aurélie Névéol, LIMSI-CNRS

Reproducibility in Research:
Insight from experiments in Computer Science and beyond

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
Campus Universitaire Bât. 508 - F-91405 Orsay Cedex
www.limsi.fr

Outline

What are the challenges in reproducibility?

What do we gain by aiming for reproducibility?

How can reproducibility be achieved?

Replicability, reproducibility, repeatability

Are these terms equivalent?

A definition:

- "Independently running a research experiment and yielding the same results on each iteration"

↳ Reproducibility is the essence of science

Reasons to work reproducibly

Reproducibility…

- Helps avoid disaster… and move science forward
- Makes it easier to publish papers
- Helps you get your point across
- Enables continuity of your work
- Helps build your reputation, e.g. attracts more citations

Piwowar HA, Day RS, Fridsma DB. Sharing detailed research data is associated with increased citation rate. PLoS One. 2007 Mar 21;2(3):e308.

Markowetz F. Five selfish reasons to work reproducibly. Genome Biol. 2015 Dec 8;16:274. .

Challenges in Reproducibility

Reports of a reproducibility crisis in many disciplines

Discipline	Failed to reproduce others' experiment	Failed to reproduce own experiment
Chemistry	90%	60%
Biology	80%	60%
Physics and engineering	70%	50%
Medicine	70%	60%
Earth and environment science	60%	40%
Other	60%	50%

Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016 May 25;533(7604):452-4.

How is this possible?

Data is often unavailable

- e.g. medical data due to confidentiality
- Software due to commercial strategy
- Seemingly insufficient details are left out of protocols

Reporting bias

- Space limitation in papers (e.g. conference papers in computer science)
- Novelty is valued more than reproducibility

Learning from reproducibility (or lack thereof)

The tale of the Zigglebottom tagger

Variability lies in…

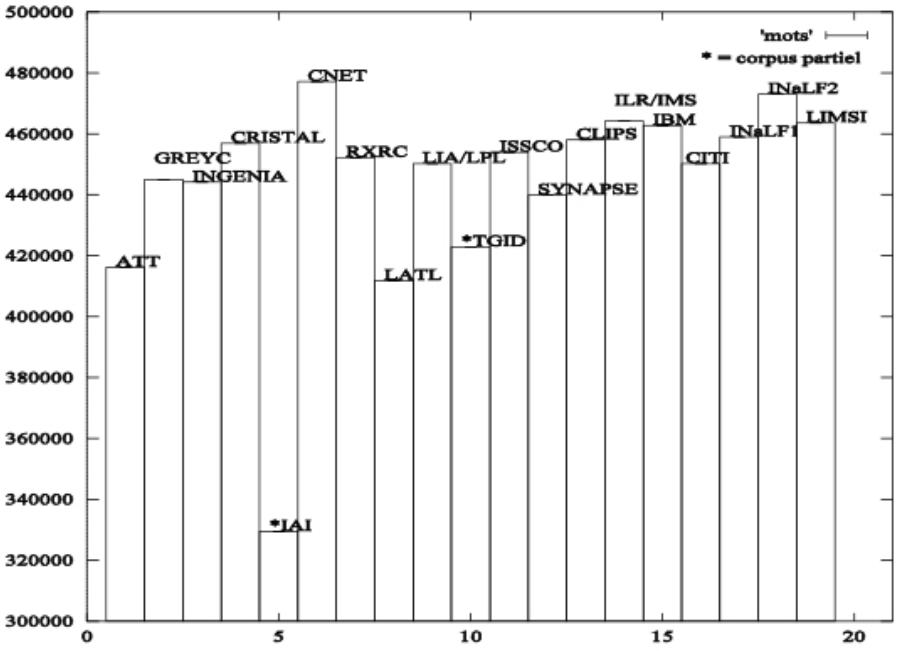
- Pre-processing (what is being pre-processed?)
  - Tokenization
  - Stop-word lists
  - "Data cleaning", e.g. normalization of case, diacritics
- Software versions, system variations
- Parameters, including training/test split

Pedersen T. 2008. Empiricism is not a matter of faith. Computational Linguistics:*34(3):465-470*

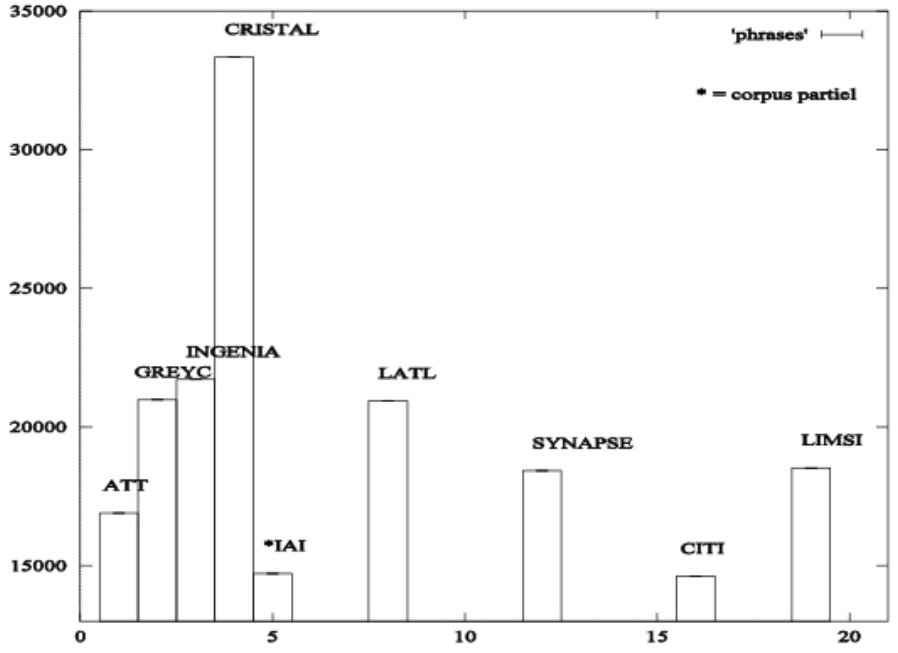
Fokkens A, Van Erp M, Postma M, Pedersen T, Vossen P, Freire N. 2013. Offspring from Reproduction Problems: What Replication Failure Teaches Us. Proc *ACL:* 1691-1701

Variability on corpus: GRACE

Counting « words »



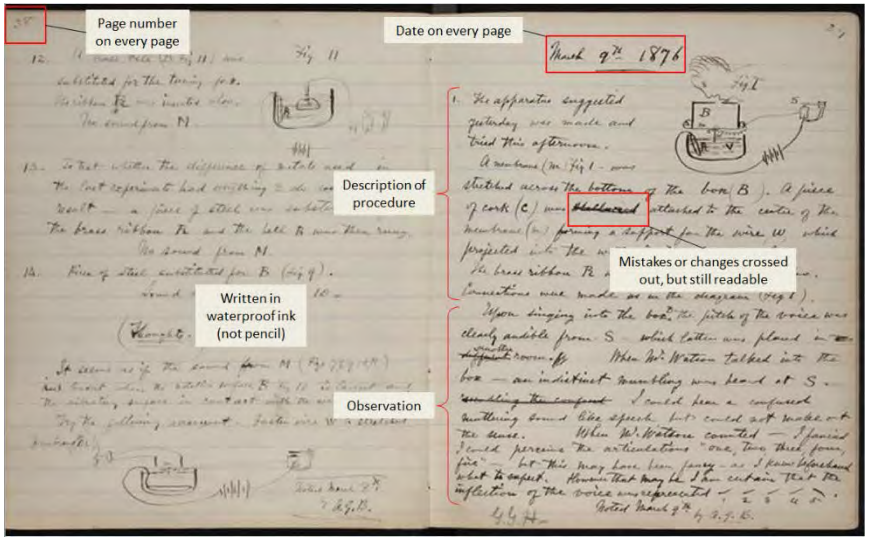
Counting « sentences »



Standardization and Documentation

- Standardized components, procedures, workflows
- Documenting complete system set-up across entire provenance chain

How to do this – efficiently?



Alexander Graham Bell's Notebook, March 9 1876
https://commons.wikimedia.org/wiki/File:Alexander_Graham_Bell's_notebook,_March_9,_1876.PNG

knitr

IPython
Interactive Computing

git

docker

The Shared Task Model

Primary goal is to provide a forum for direct comparison of approaches

- Availability of shared material
- Specific definition of a “task”
- Corpora and annotations, split into training, development and test sets
- Evaluation metrics and scripts

Shared Tasks Examples

Information Retrieval and information extraction

- MUC, TREC, CLEF, CTCIR

Computational Linguistics

- Semeval, GRACE, EASY, DEFT

Translation

- WMT

BioNLP, curation

- i2b2, BioCreAtive, BioASQ

The PRIMAD[1] model: which attributes can we “prime”?

Defining Types of Reproducibility

- Data
  - Parameters
  - Input data
- Platform
- Implementation
- Method
- Research Objective
- Actors

What do we gain by priming one or the other?

[1] Juliana Freire, Norbert Fuhr, and Andreas Rauber. Reproducibility of Data-Oriented Experiments in eScience. Dagstuhl Reports, 6(1), 2016.

Types of Reproducibility and Gains

Label	Data		Platform / Stack	Implementation	Method	Research Objective	Actor	Gain
	Parameters	Raw Data						
Repeat	-	-	-	-	-	-		Determinism
Param. Sweep	**x**	-	-	-	-	-		Robustness / Sensitivity
Generalize	(x)	**x**	-	-	-	-		Applicability across different settings
Port	-	-	**x**	-	-	-		Portability across platforms, flexibility
Re-code	-	-	(x)	**x**	-	-		Correctness of implementation, flexibility, adoption, efficiency
Validate	(x)	(x)	(x)	(x)	**x**	-		Correctness of hypothesis, validation via different approach
Re-use	-	-	-	-	-	**x**		Apply code in different settings, Re-purpose
Independent *x* *(orthogonal)*							**x**	Sufficiency of information, independent verification

Levels of reproducibility (in computer science)

1. Availability: the system and data it was tested on must be available (or there must be sufficient detail available to reconstruct the system and dataset).
2. Builds: the code must build.
3. Runs: the built code must run.
4. Evaluation: it must be possible to run on the same data

and measure the output using the same implementation of the same scoring metric.

Neurology

PLOS ONE | Articles | For Authors | About Us | Search | advanced search

OPEN ACCESS PEER-REVIEWED

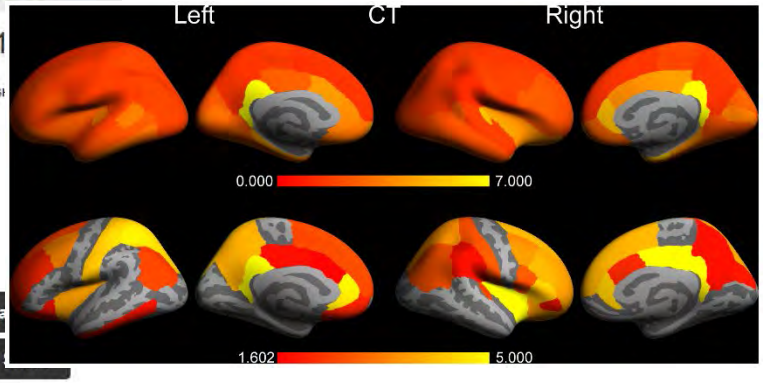
68,919 VIEWS | 10 CITATIONS | 124 SAVES

RESEARCH ARTICLE

The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements

Ed H. B. M. Gronenschild, Petra Habets, Heidi I. L. Jacobs, Ron Mengelers, Nico Rozendaal, Jim van Os, Machteld Marcelis

Article | About the Authors | Metrics | Comments | Related Content | Download | Print | Show Figures



- Abstract
- Introduction
- Materials and Methods
- Results
- Discussion
- Supporting Information
- Acknowledgments
- Author Contributions
- References

Reader Comments (5)

Figures

Abstract

FreeSurfer is a popular software package to measure cortical thickness and volume of neuroanatomical structures. However, little if any is known about measurement reliability across various data processing conditions. Using a set of 30 anatomical T1-weighted 3T MRI scans, we investigated the effects of data processing variables such as FreeSurfer version (v4.3.1, v4.5.0, and v5.0.0), workstation (Macintosh and Hewlett-Packard), and Macintosh operating system version (OSX 10.5 and OSX 10.6). Significant differences were revealed between FreeSurfer version v5.0.0 and the two earlier versions. These differences were on average 8.8±6.6% (range 1.3–64.0%) (volume) and 2.8±1.3% (1.1–7.7%) (cortical thickness). About a factor two smaller differences were detected between Macintosh and Hewlett-Packard workstations and between OSX 10.5 and OSX 10.6. The observed differences are similar in magnitude as effect sizes reported in accuracy evaluations and neurodegenerative studies.

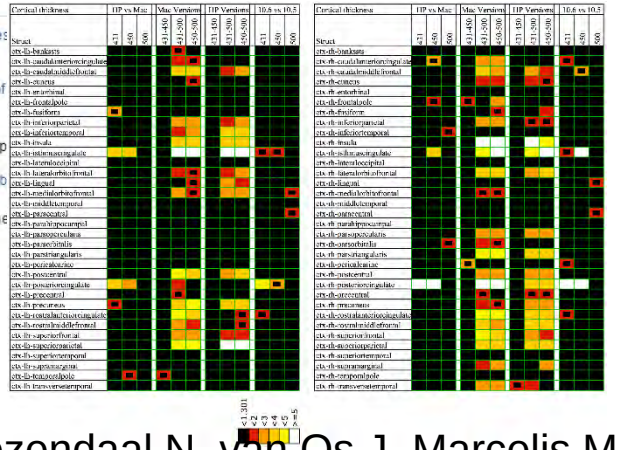
The main conclusion is that in the context of an ongoing study, users are discouraged to update to a new major release of either FreeSurfer or operating system or to switch to a different type of workstation without repeating the analysis; results thus give a quantitative support to successive recommendations stated by FreeSurfer developers over the years. Moreover, in view of the large and significant cross-version differences, it is concluded that formal assessment of the accuracy of FreeSurfer is desirable.

Comments

In praise of progres
Posted by GedR

Media Coverage of Article
Posted by PLoS_ONE_Group

Comments made b authors
Posted by EdGrone



Gronenschild EH, Habets P, Jacobs HI, Mengelers R, Rozendaal N, van Os J, Marcelis M. The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements. PLoS One. 2012;7(6):e38234.

Bioinformatics

Obtain workflows from MyExperiments.org

- March 2015: almost 2.700 WFs (approx. 300-400/year)
- Focus on Taverna 2 WFs: 1.443 WFs

Try to re-execute the workflows

- Majority of workflows fails
- Only 23.6 % are successfully executed

(correctness of results not checked yet)

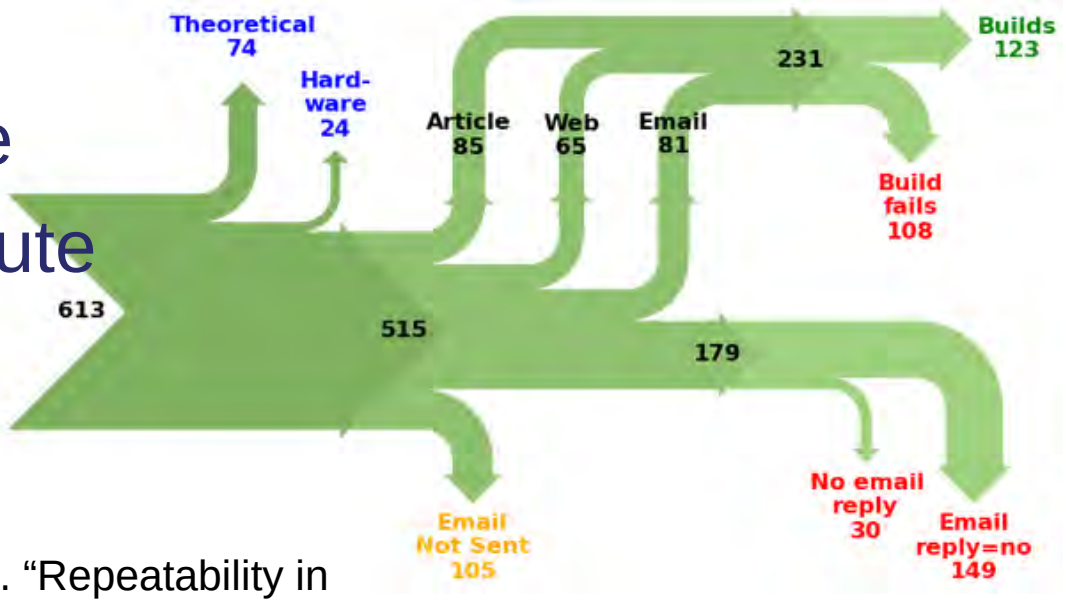
Rudolf Mayer, Andreas Rauber, “A Quantitative Study on the Re-executability of Publicly Shared Scientific Workflows”, 11th IEEE Intl. Conference on e-Science, 2015.

Computer Science

613 papers in 8 ACM conferences

Process

- download paper and classify
- search for a link to code (paper, web, email twice)
- download code
- build and execute



Christian Collberg and Todd Proebsting. “Repeatability in Computer Systems Research,” CACM 59(3):62-69.2016

Biomedical Natural Language Processing

Reproducibility track at CLEF eHealth 2016

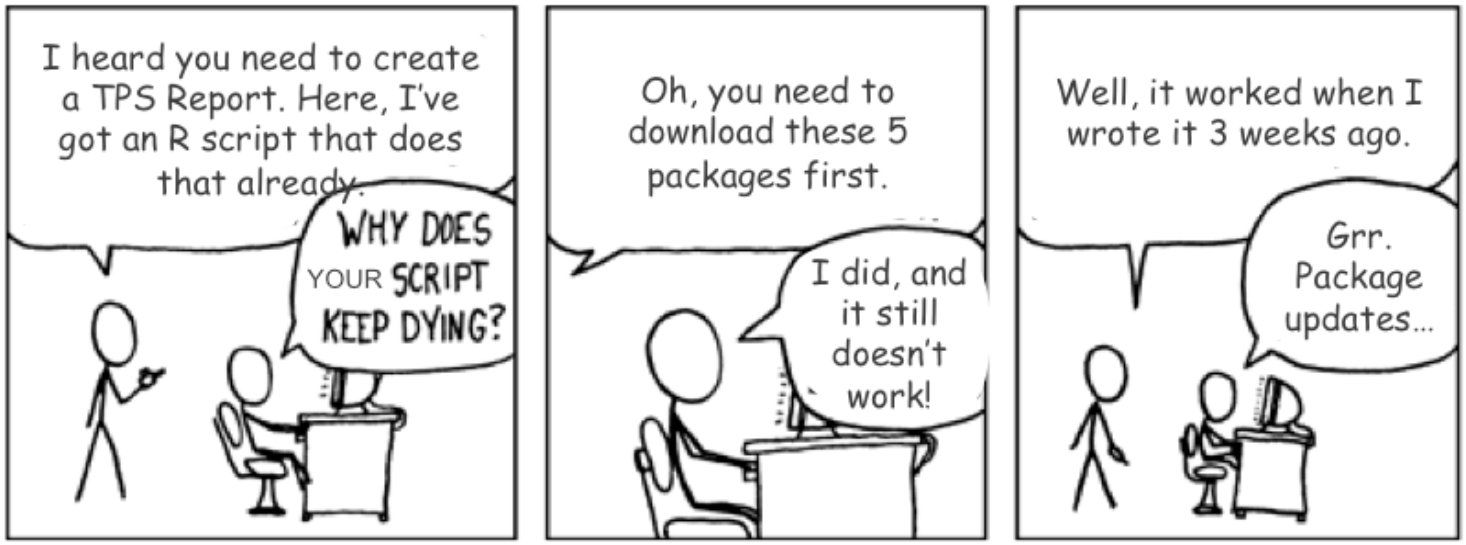
- An automatic coding task
- 4 analysts aim to reproduce participants runs

Overall, results can be reproduced, but…

- Replication is not easy
- No analyst was able to replicate every run
- Documentation shortcomings reported

Névéol A, Cohen KB, Grouin C, Robert A. Replicability of Research in Biomedical Natural Language Processing: a pilot evaluation for a coding task. Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, LOUHI. 2016

More BioNLP



Source: a parody of xkcd

Studied 2 R libraries

- Needed to contact authors to use sucessfully
- Produced extra documentation and test cases

Cohen K, Xia J, Roeder C, Hunter L. 2016. Reproducibility in Natural Language Processing: A Case Study of two R Libraries for Mining PubMed/MEDLINE. LREC/4REAL workshop

Take Home message: Aim at achieving reproducibility

At different levels

- Re-run, ask others to re-run
- (Re-implement)
- (Port to different platforms)
- Test on different data,
  vary parameters (and report!)

If something is not reproducible ->
investigate!
(you might be onto something)

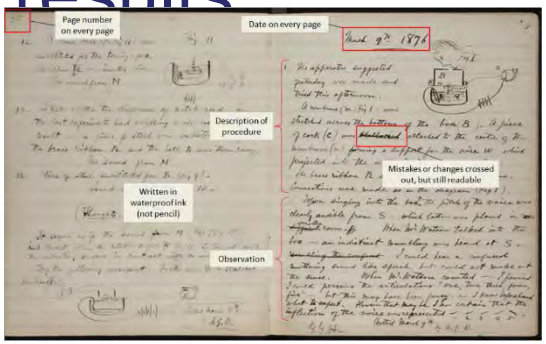
Aim for better procedures and documentation

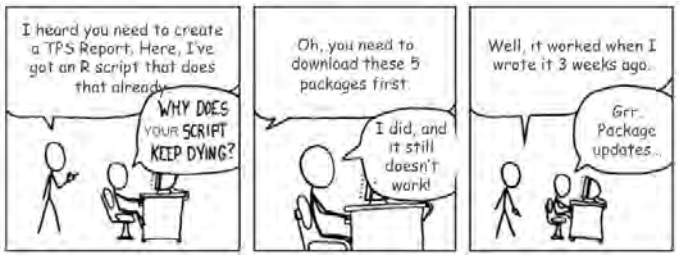
Plan your research procedure

- Define a protocol
- Have a data management plan

Document, document, document

- the research process, environment, interim results





Acknowledgements

- Andreas Rauber (Vienna University of Technology)
- Kevin B. Cohen (University of Colorado)
- Cyril Grouin (LIMSI-CNRS), Aude Robert (INSERM/CépiDC)
- Patrick Paroubek and Pierre Zweigenbaum (LIMSI-CNRS)

CABeRneT ANR-13-JS02-0009-01

CLEF initiative

Methods in Research on Research

A presentation delivered at the

first MiRoR training event
October 19-21, 2016
Ghent, Belgium

This project has received funding from the EU Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement #676207