Reproducible research tools, BIOS691

Mikhail Dozmorov

Summer 2018

Overview of the course

- We will discuss general principles for reproducible research, but will focus primarily on the practical use of relevant tools (particularly in **Linux** environment, **make**, **git**, data manipulation/graphics in **R**, and reports in **Markdown/knitr**)
- The goal is to ensure that all aspects of computational research you will do (software, data analyses, papers, presentations, posters) are integrated within reproducible framework
- Doing things properly (writing clear, documented, well-tested code) may be viewed as time consuming, but it will help you many times down the road
- Ultimately, you'll be more efficient, and your work will have greater impact

What is reproducible research?

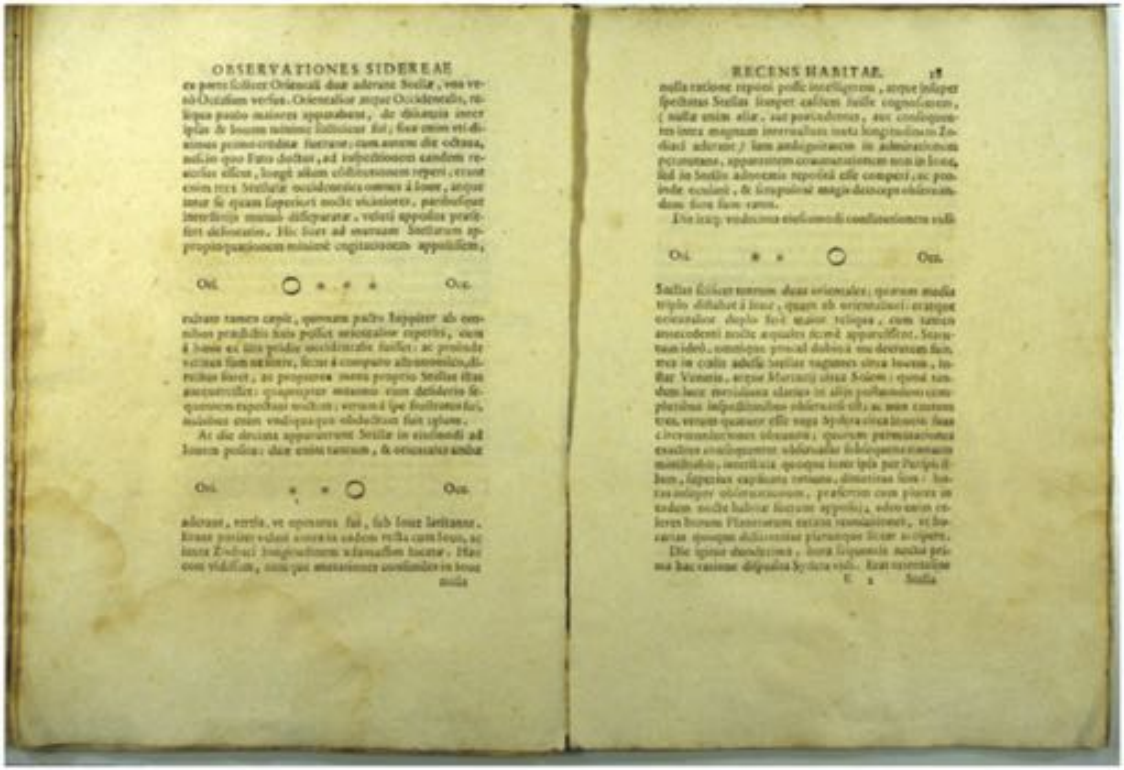
Scientific method

Reproducibility defined

- Reproducibility is the ability of an entire experiment or study to be duplicated, either by the same researcher or by someone else working independently
- Reproducibility is one of the main principles of the scientific method
- **Reproducible research is the ultimate standard for strengthening scientific evidence** by independent:
  - Investigators
  - Data
  - Methods
  - Laboratories
  - Instruments

https://en.wikipedia.org/wiki/Reproducibility

The first reproducible research: Galileo Galilei



OBSERVATIONES SIDEREAE

RECENS HABITAE.

Galileo's notes directly integrated his data (drawings of Jupiter and its moons), key metadata (timing of each observation, weather, and telescope properties), and text (descriptions of methods, analysis, and conclusions)

Two pages from Galilei's Sidereus Nuncius ("The Starry Messenger" or "The Herald of the Stars"), Venice, 1610.
http://journals.plos.org/ploscompbiol/article?id=info:doi/10.1371/journal.pcbi.1003542

Basics of reproducibility: Lab notebook

- Complete record of procedures (how), reagents (what), observations (data), and thoughts to pass on to other researchers
- Explanation of why experiments were initiated, how they were performed, and the results
- Legal document to prove patents and defend your data against accusations of fraud
- Scientific legacy in the lab

Why do we care?

More data = more chance for errors

- High-throughput biology generates volumes of data
- Data-generating technologies are increasingly used to make clinical recommendations and treatment decisions
- A problem may be overlooked .. Published .. Get in clinical trials

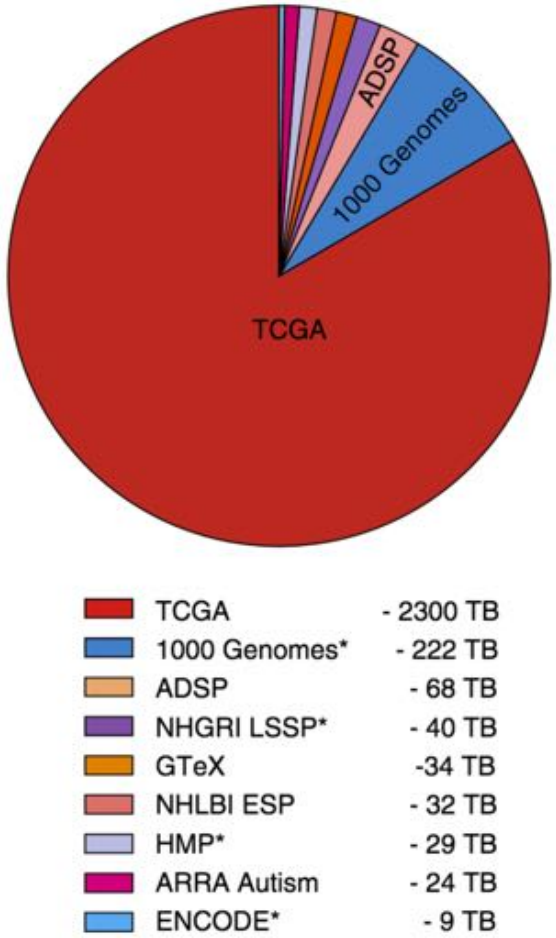
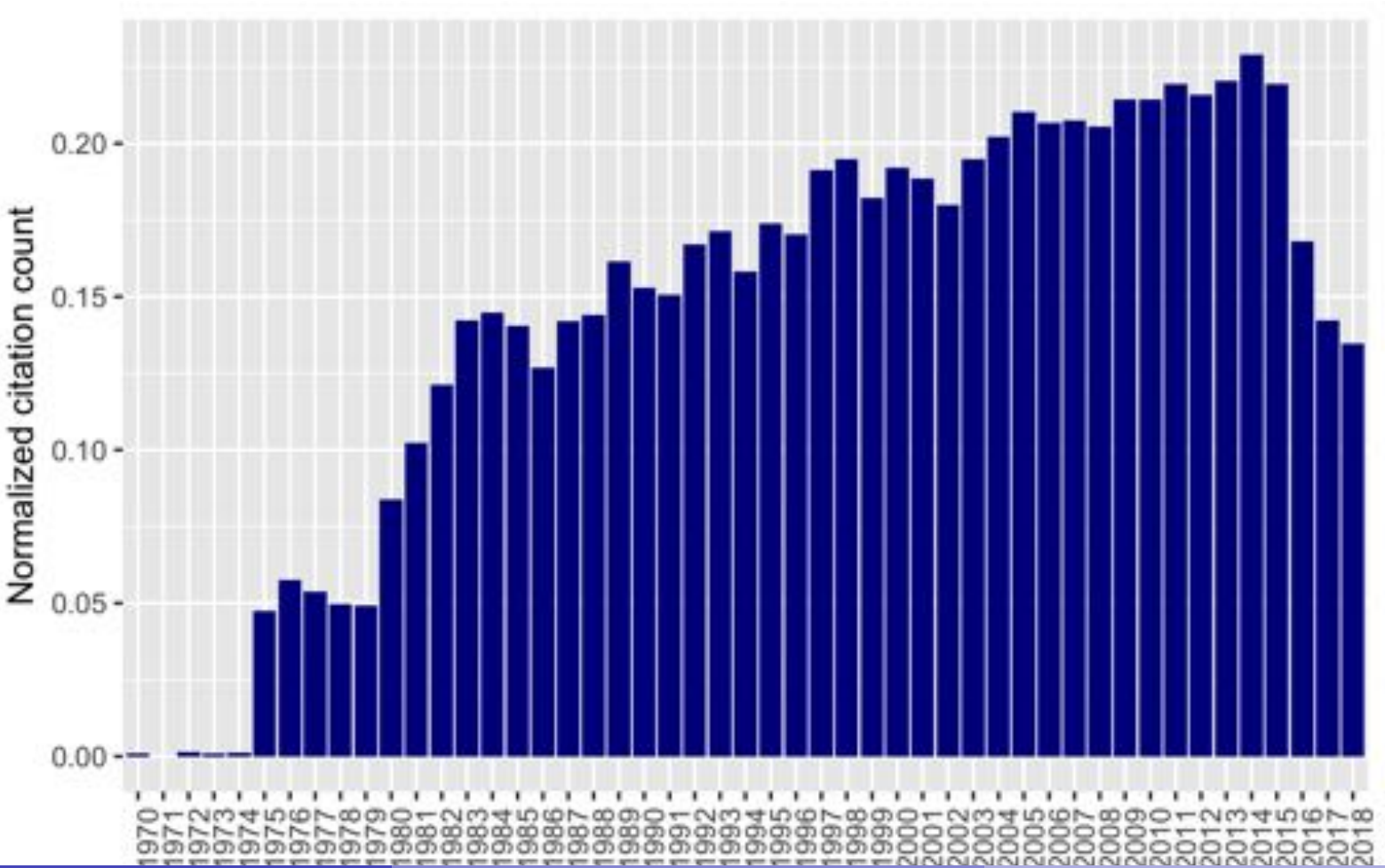


Image credit: Muir et al., "The Real Cost of Sequencing." Genome Biol. 2016

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0917-0

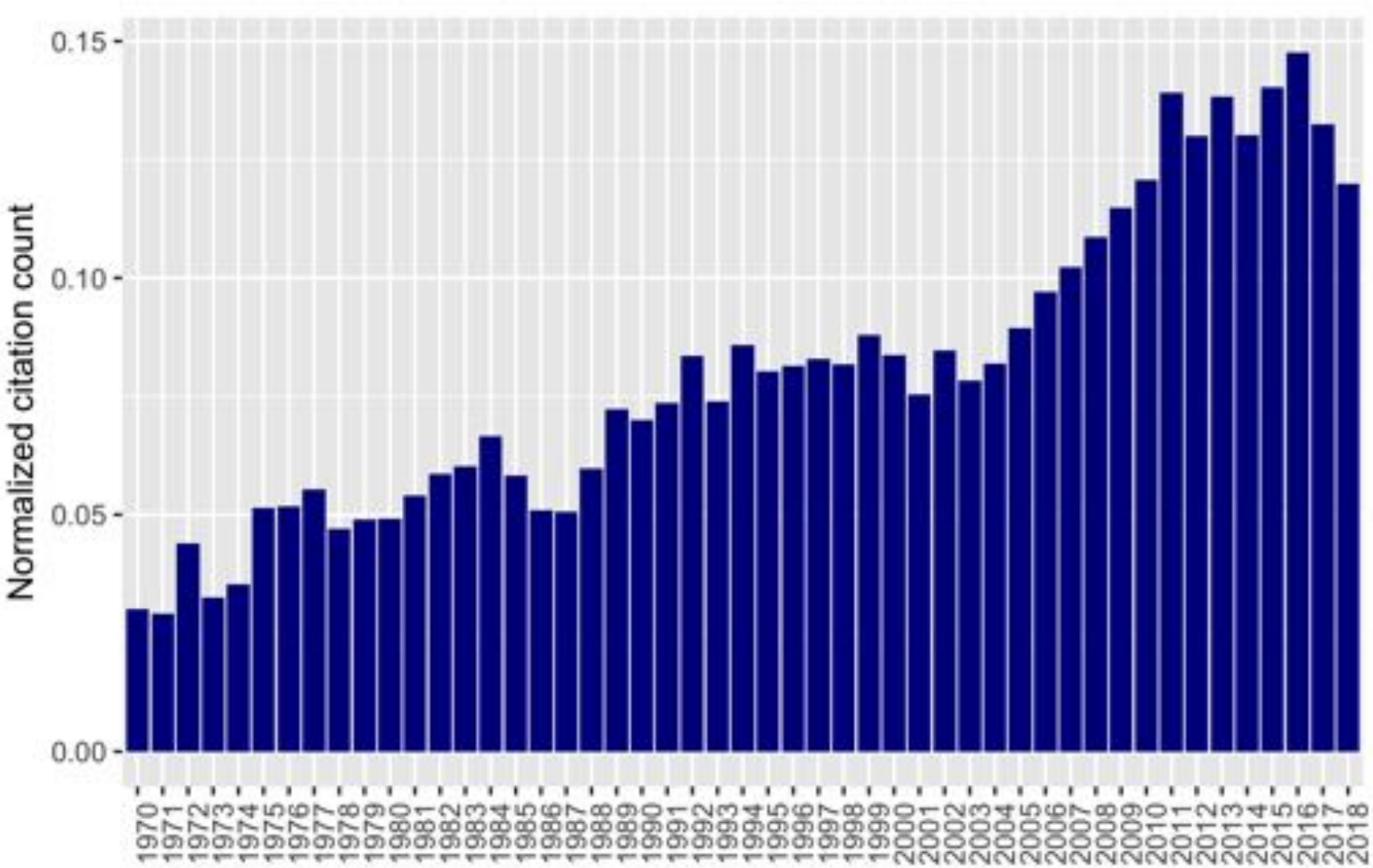
PubMed stats on "Reproducible research" vs. "Retraction"

Reproducible research



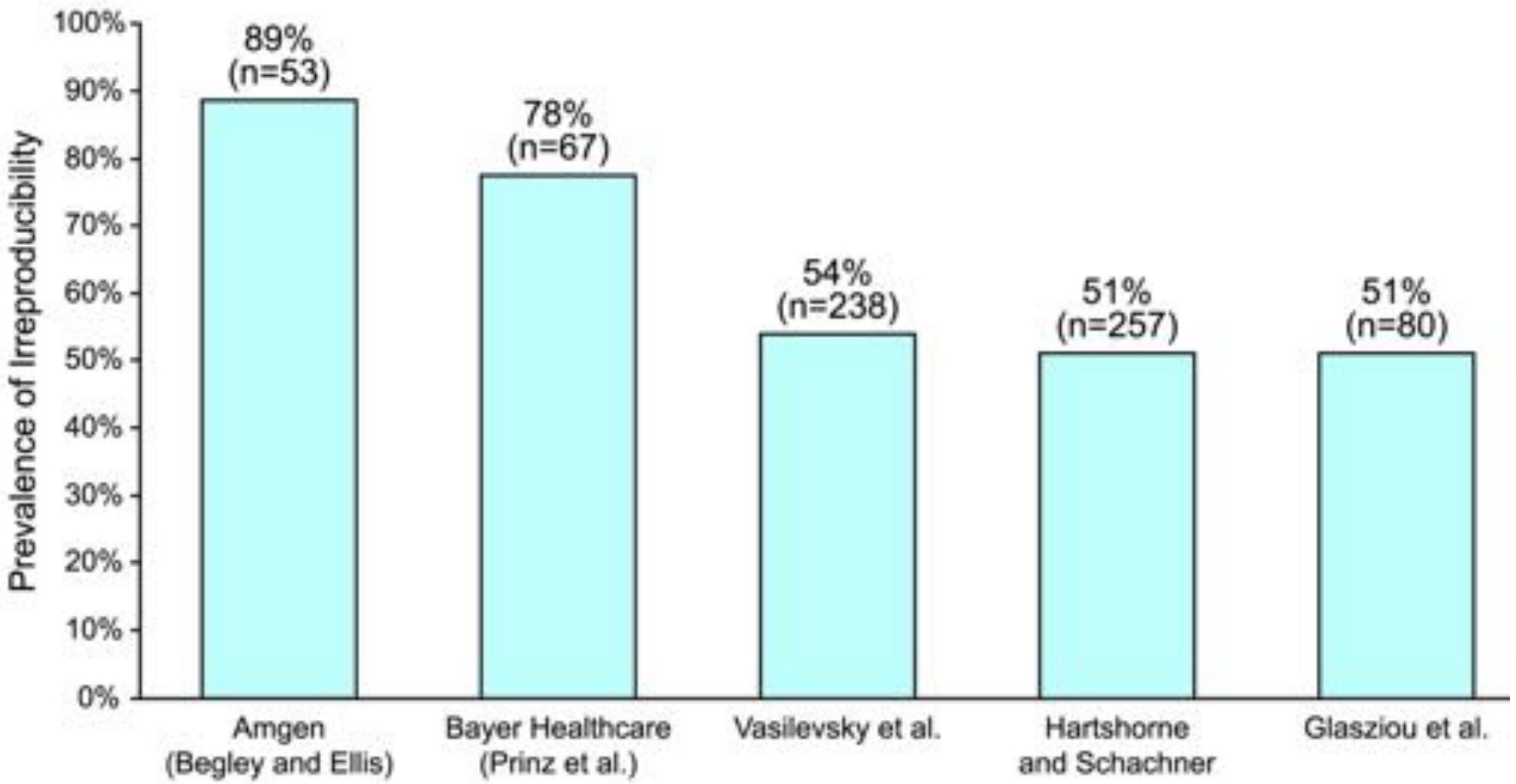
PubMed stats on "Reproducible research" vs. "Retraction"

Retraction



The cost of reproducibility

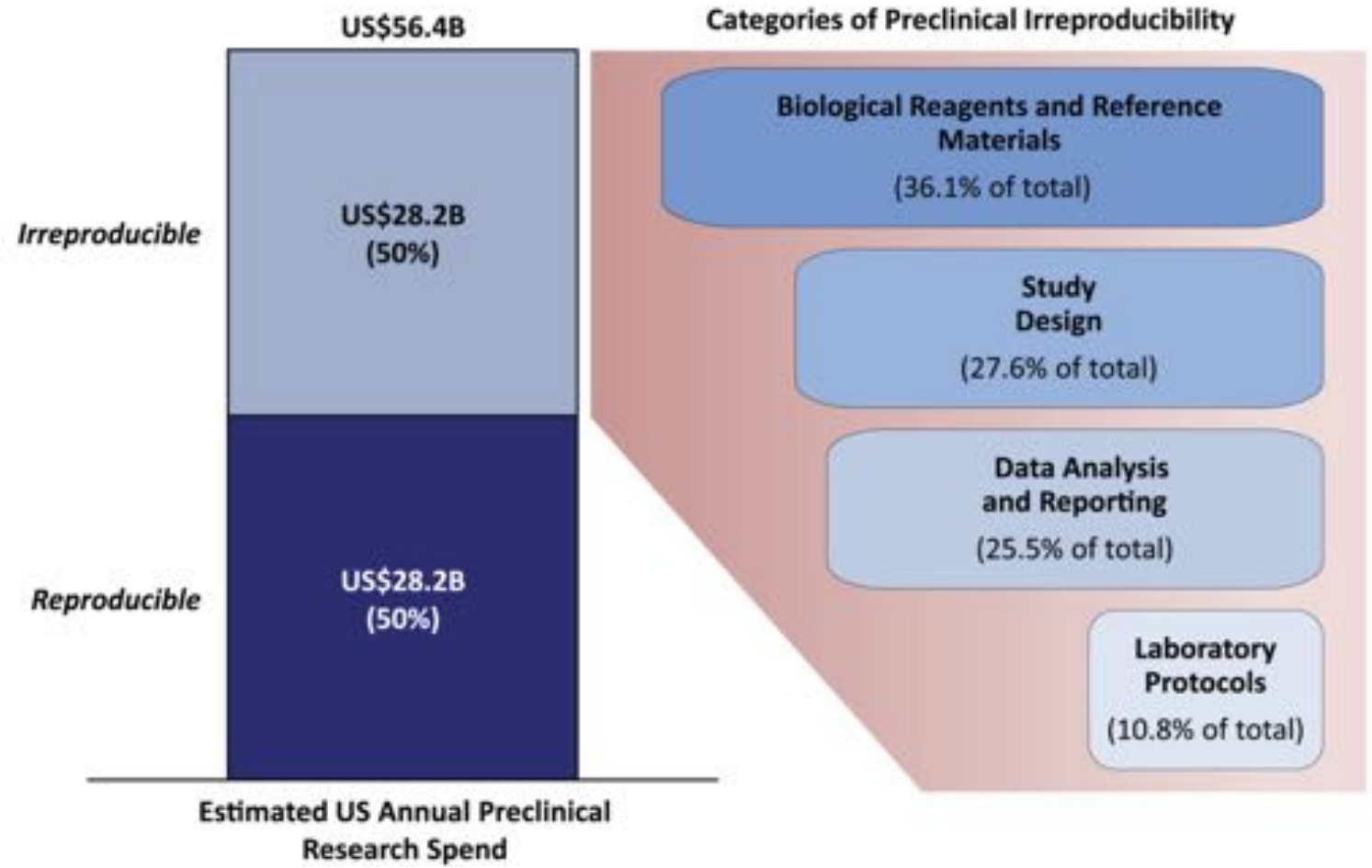
Irreproducibility ranges from 51% to 89%



Leonard Freedman, Iain Cockburn, and Timothy Simcoe, "The Economics of Reproducibility in Preclinical Research." PLOS Biol 2015

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165

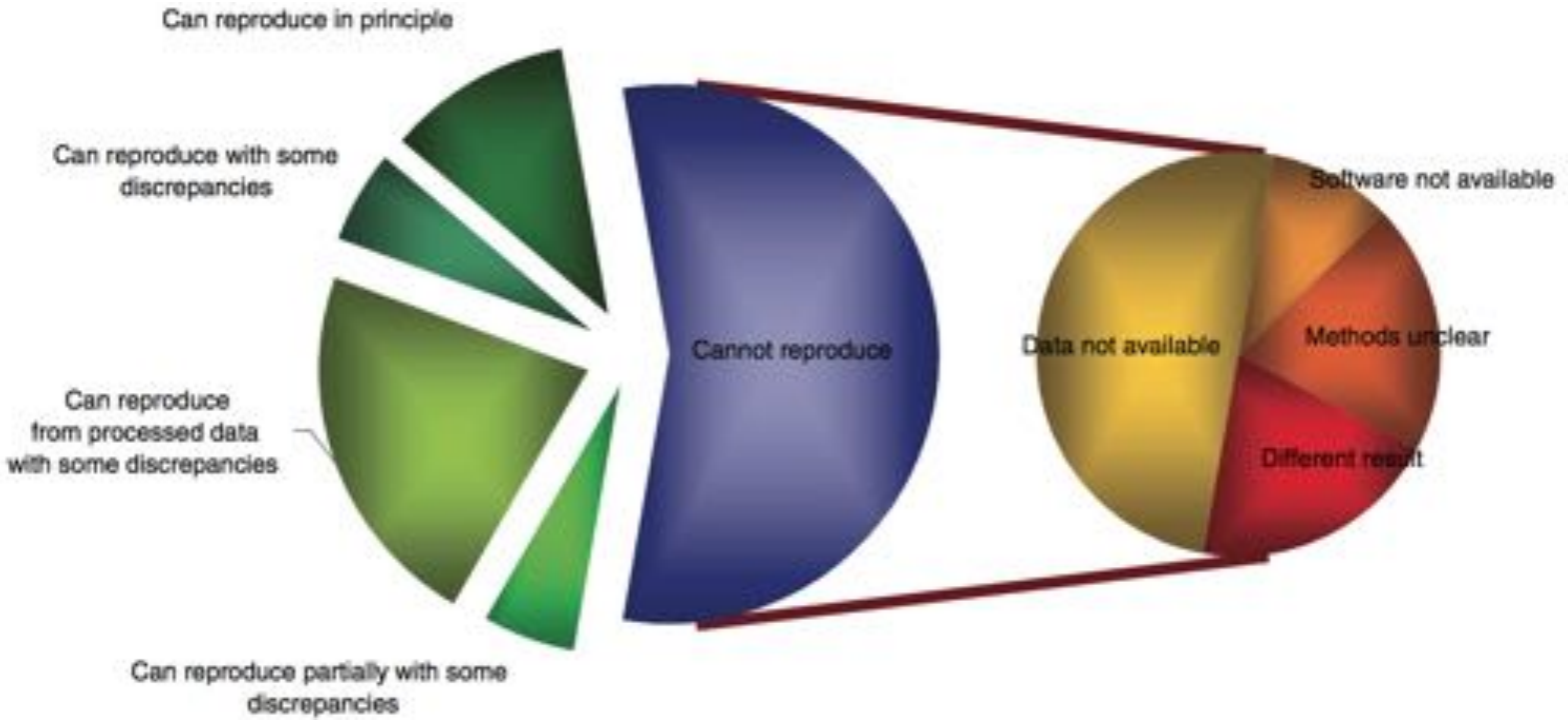
Cost of irreproducibility



Leonard Freedman, Iain Cockburn, and Timothy Simcoe, "The Economics of Reproducibility in Preclinical Research." PLOS Biol 2015

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002165

Microarray reproducibility



- 18 Nat. Genet. microarray experiments
- Less than 50% reproducible
- Problems:
  - Missing data (38%)
  - Missing software/hardware details (50%)

Sequencing reproducibility

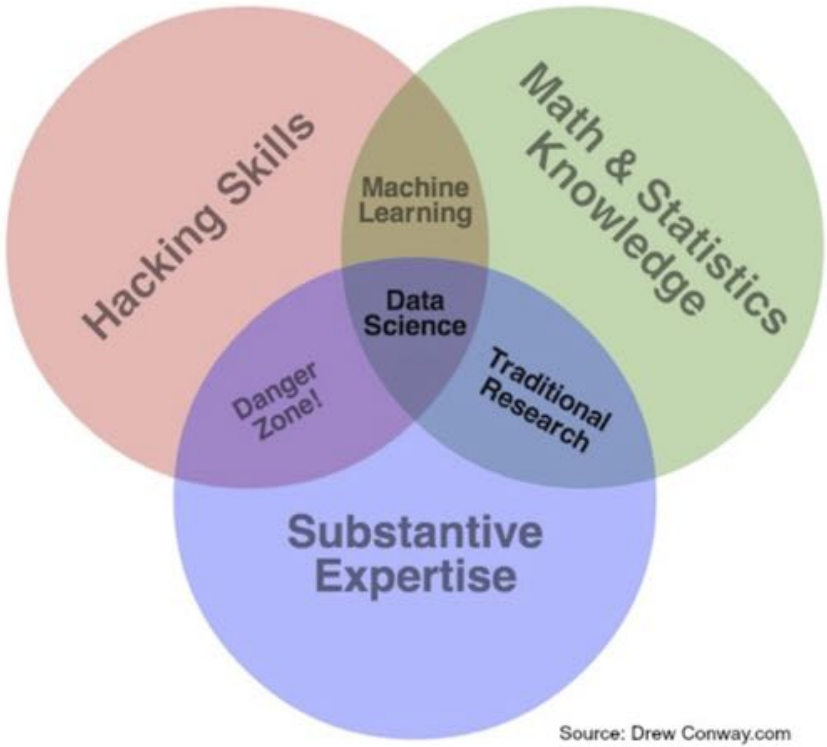
- NGS: run-of-the-mill variant calling (align, process, call variants):
  - 299 articles published in 2011 citing the 1000 Genomes project pilot publication
  - Only 19 were NGS studies with similar design
  - Only 10 used tools recommended by 1000G.
  - Only 4 used full 1000G workflow (realignment & quality score recalibration).

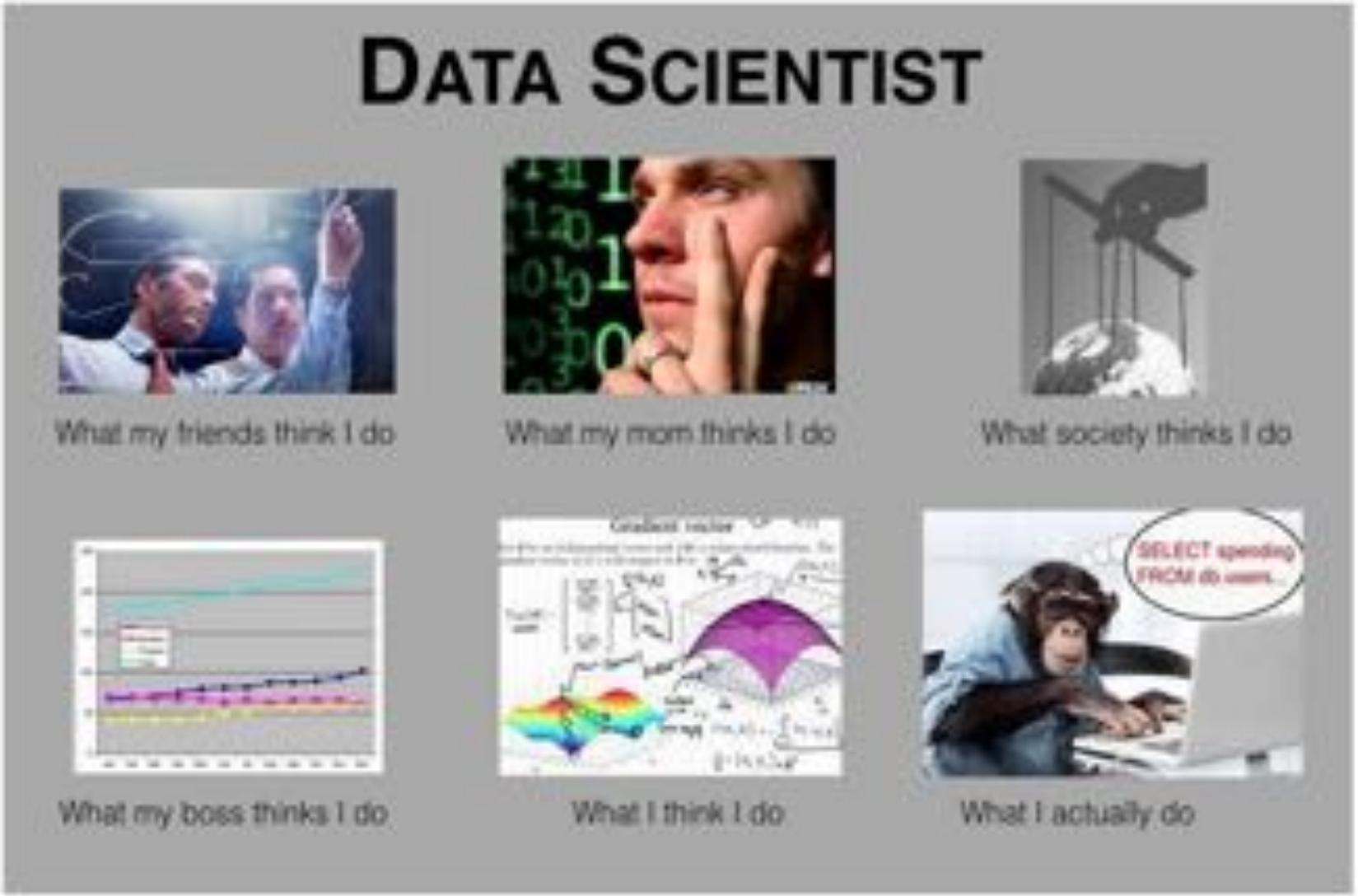
Nekrutenko, Anton, and James Taylor. "Next-Generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility." Nature Reviews. Genetics 13, no. 9 (2012): 667–72. https://doi.org/10.1038/nrg3305.

REPRODUCIBILITY IN DATA SCIENCE

Reproducibility in data science

- A data scientist is often referred to as someone who knows more statistics than a computer scientist and more computer science than a statistician. *Joshua Blumenstock*
- Data Scientist = statistician + programmer + coach + storyteller + artist. *Shlomo Aragmon*





http://www.sharpsightlabs.com/machine-learning-prerequisite-isnt-math/

Steps in reproducible research

The most important is the mindset, when starting, that the end product will be reproducible. *Keith Baggerly*

- Points to consider before starting a project:
  - Experimental design
  - Data generation
  - Data analysis
  - Results interpretation
  - Dissemination of results

Common approach: write report around results

Point and click approach

- Use MS Excel for data entry/cleaning/preparation, and possibly statistical analysis

Problems

- With point-and-click, there's no way to record/save the steps that generated the (copy/pasted) results
- Data files are kept separately from the analysis code, and from reports
- After modifications of one of the files, it becomes unclear which version corresponds exactly to the reported results
- Every time something changes, you have to regenerate the figures/results/reports by hand – very time consuming

Zeeberg BR et al. "Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics" BMC Bioinformatics 2004

http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-80

Better approach: write report that generates results

- The report is automated via code
- Data is attached to the well-documented code
- History of any changes should be preserved

The final report should be self-sufficient and reproducible with a single command

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.

J. B. Buckheit and D. L. Donoho. (1995)

http://www-stat.stanford.edu/~donoho/Reports/1995/wavelab.pdf

Software + data = reproducible report

- Distribute fully automated report with data and code

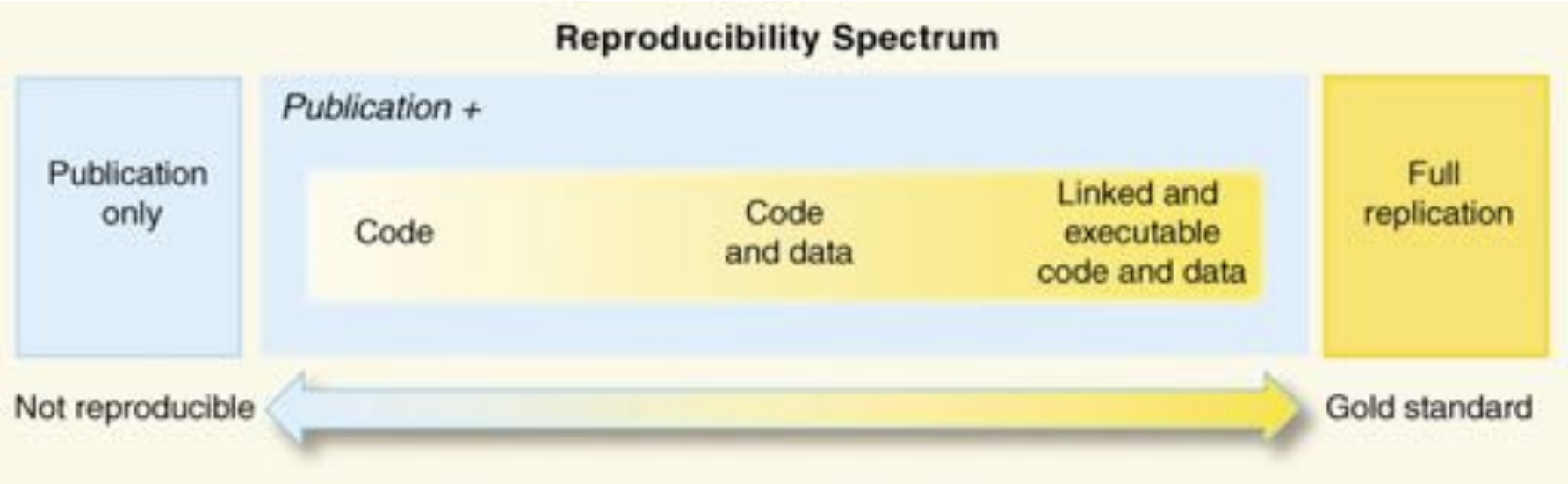


Image credit: Roger Peng "Reproducible Research in Computational Science" Science 2011

http://science.sciencemag.org/content/334/6060/1226

Quality of reproducible report

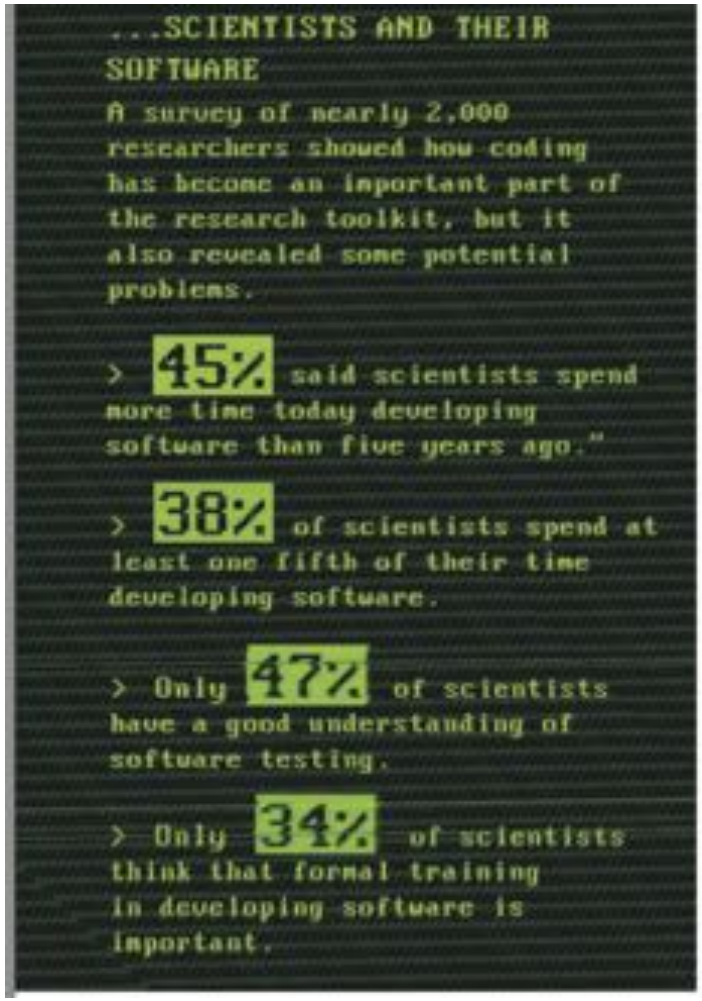
- Are the tables and figures reproducible from the code and data?
- Does the code actually do what you think it does?
- In addition to what was done, is it clear why it was done? (e.g., how parameter settings were chosen?)
- Is your code scalable to accommodate more data/methods?

"Data/code is available upon request" or "Data/code is available at the lab's website" are completely unacceptable in the 21st century.

Six degrees of reproducibility

1. The results cannot be reproduced
2. The results cannot seem to be reproduced
3. Reproducibility requires extreme effort
4. Reproducibility requires considerable effort
5. Easy reproducibility, but require some proprietary source packages (MATLAB, SAS, etc.)
6. The results can be easily reproduced by an independent researcher with at most 15 min of user effort, requiring only standard, freely available tools (C compiler, R, Python, etc.)

Scientific computing



http://www.nature.com/news/2010/101013/full/467775a.html

Summary: Rules of reproducible research

- For every result, keep track of how it was produced
- Avoid manual data manipulation steps
- Archive the exact versions of all external programs used
- Version control all custom scripts
- Record all intermediate results, when possible, in standardized formats
- For analyses that include randomness, note underlying random seeds
- Always store raw data behind plots
- Generate hierarchical analysis output, allowing layers of increasing detail to be inspected
- Connect textual statements to underlying results
- Provide public access to scripts, runs, and results

Sandve, Geir Kjetil, Anton Nekrutenko, James Taylor, and Eivind Hovig. "Ten Simple Rules for Reproducible Computational Research." Edited by Philip E. Bourne. PLoS Computational Biology 9, no. 10 (October 24, 2013): e1003285. https://doi.org/10.1371/journal.pcbi.1003285.