REPRODUCIBLE RESEARCH TOOLS
BIOS 692

https://mdozmorov.github.io/BIOS692/

Image credit: Bill & Melinda Gates Foundation, http://k12education.gatesfoundation.org/learning/tools-being-talked-about/

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) is 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

Introduction

- What is reproducible research?
- Why do we care?
- The cost of reproducibility
- Reproducibility in data science

WHAT IS REPRODUCIBLE RESEARCH?

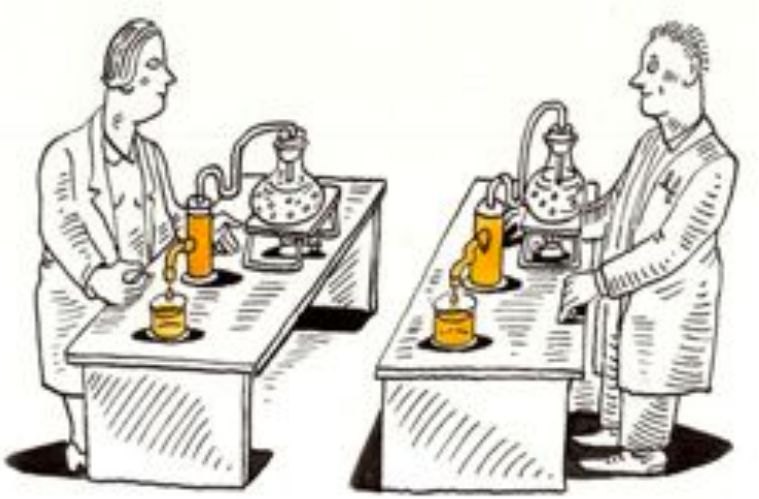


Image credit: http://blog.f1000research.com/2014/04/04/reproducibility-tweetchat-recap/

What is reproducible research?

- 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

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

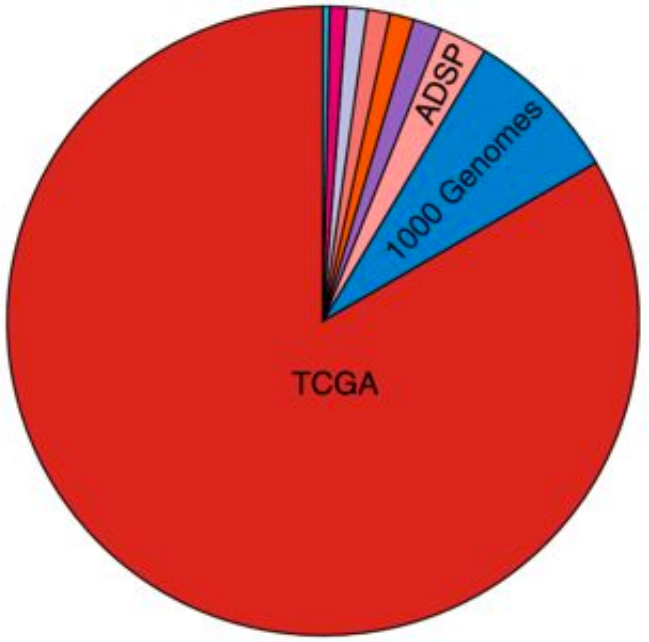


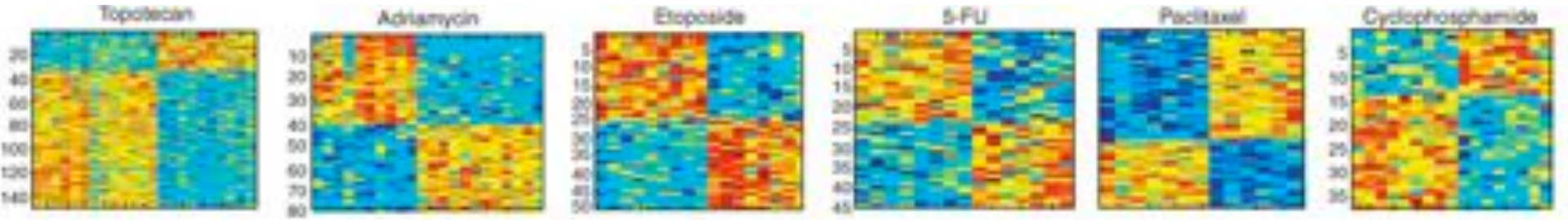
Image credit: Muir et al., "The Real Cost of Sequencing." Genome Biol. 2016
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0917-0

Clinical trials based on flawed and fraudulent data

Genomic signatures to guide the use of chemotherapeutics

Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan Lancaster[4] & Joseph R Nevins[1–3]

- Described drug response “gene signatures” in NCI60 cell lines
- Demonstrated these “signatures” correspond to patient-specific signatures and can be used to predict patient response to the drugs

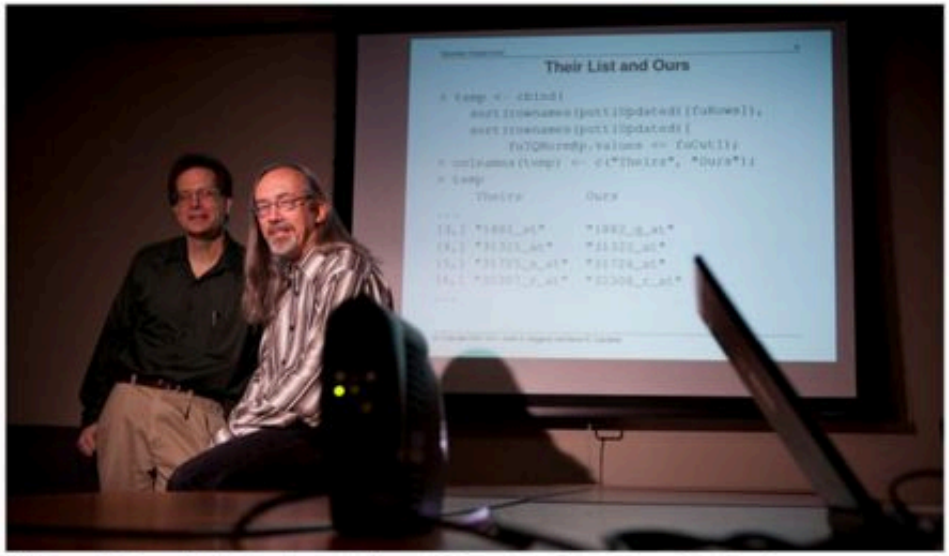


Retraction Watch "The Importance of Being Reproducible: Keith Baggerly tells the Anil Potti story" 2011, http://retractionwatch.com/2011/05/04/the-importance-of-being-reproducible-keith-baggerly-tells-the-anil-potti-story/

Bioinformatics statisticians spot errors

How Bright Promise in Cancer Testing Fell Apart

By GINA KOLATA JULY 7, 2011



Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors. Michael Stravato for The New York Times

When Juliet Jacobs found out she had lung cancer, she was terrified, but realized that her hope lay in getting the best treatment medicine could offer. So she got a second opinion, then a third. In February of 2010, she ended up at Duke University, where she entered a research study whose promise seemed stunning.

New York Times "How Bright Promise in Cancer Testing Fell Apart" 2011
http://www.nytimes.com/2011/07/08/health/research/08genes.html

“Off-by-one” error

	Published	Replicated
...		
[3,]	1881_at	1882_g_at
[4,]	31321_at	31322_at
[5,]	31725_s_at	31726_at
[6,]	32307_r_at	32308_r_at
...		

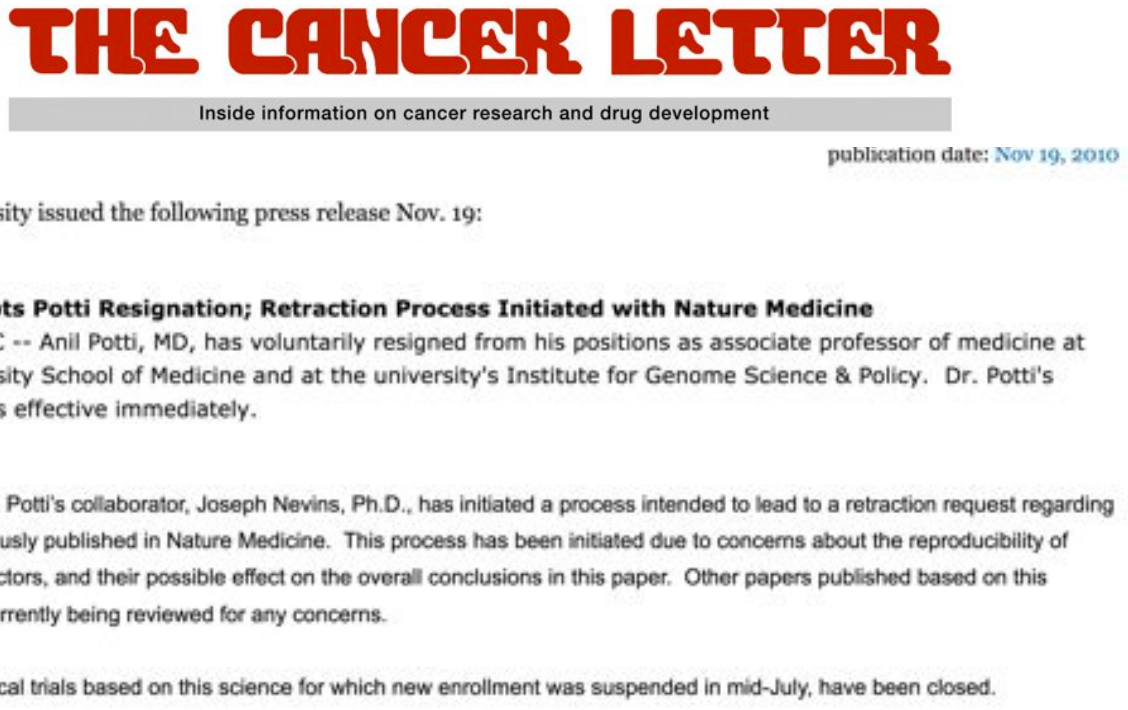
More data added

Sample	ID	Response			
1	GSM44303	RES	11	GSM9694	RES
2	GSM44304	RES	12	GSM9695	RES
3	GSM9653	RES	13	GSM9696	RES
4	GSM9653	RES	14	GSM9698	RES
5	GSM9654	RES	15	GSM9699	SEN
6	GSM9655	RES	16	GSM9701	RES
7	GSM9656	RES	17	GSM9708	RES
8	GSM9657	RES	18	GSM9708	SEN
9	GSM9658	SEN	19	GSM9709	RES
10	GSM9658	SEN	20	GSM9711	RES

RES/SEN – resistant/sensitive

Summary of the Duke case

- A total of 162 co-authors
- 40 papers
- Two-thirds are partially or completely retracted



THE CANCER LETTER

Inside information on cancer research and drug development

publication date: Nov 19, 2010

Duke University issued the following press release Nov. 19:

Duke Accepts Potti Resignation; Retraction Process Initiated with Nature Medicine

DURHAM, NC -- Anil Potti, MD, has voluntarily resigned from his positions as associate professor of medicine at Duke University School of Medicine and at the university's Institute for Genome Science & Policy. Dr. Potti's resignation is effective immediately.

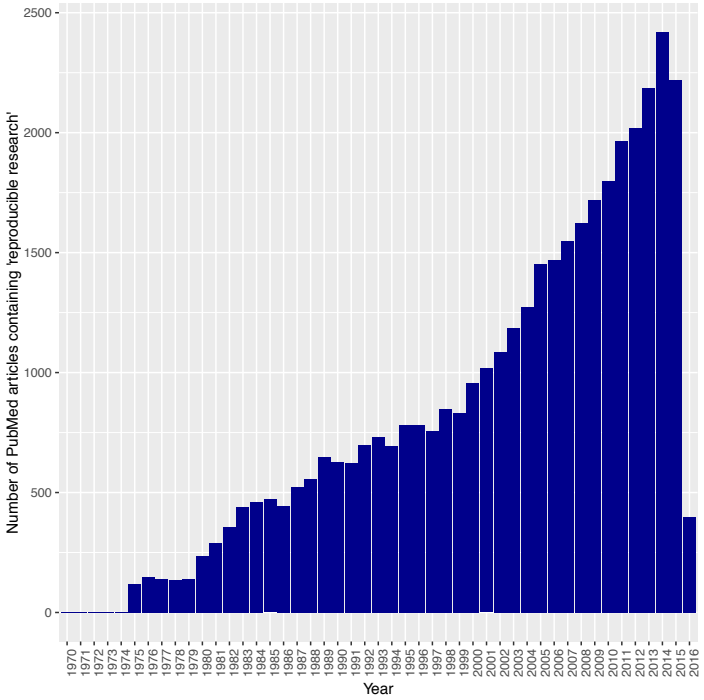
In addition, Dr. Potti's collaborator, Joseph Nevins, Ph.D., has initiated a process intended to lead to a retraction request regarding a paper previously published in Nature Medicine. This process has been initiated due to concerns about the reproducibility of reported predictors, and their possible effect on the overall conclusions in this paper. Other papers published based on this science are currently being reviewed for any concerns.

The three clinical trials based on this science for which new enrollment was suspended in mid-July, have been closed.

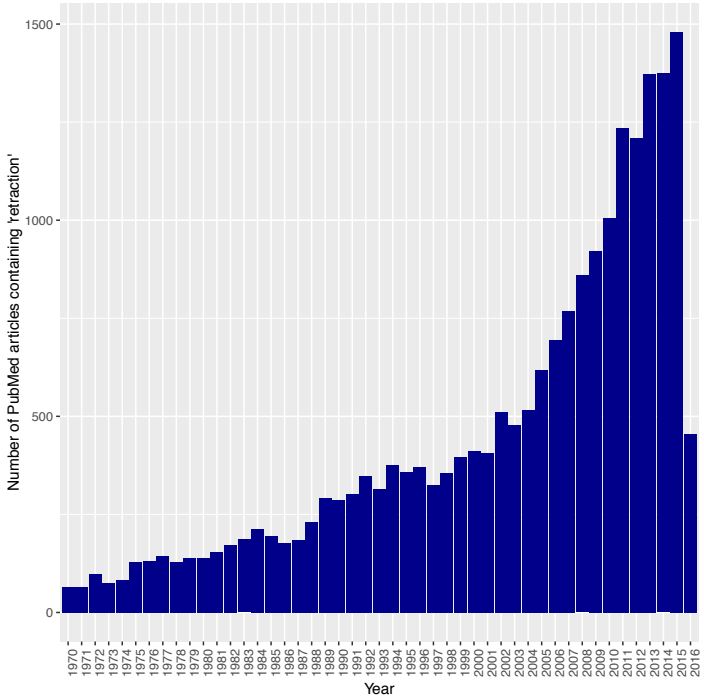
http://retractionwatch.com/2011/05/04/the-importance-of-being-reproducible-keith-baggerly-tells-the-anil-potti-story/

PubMed stats on “Reproducible research” vs. “Retraction”

“Reproducible research”



“Retraction”

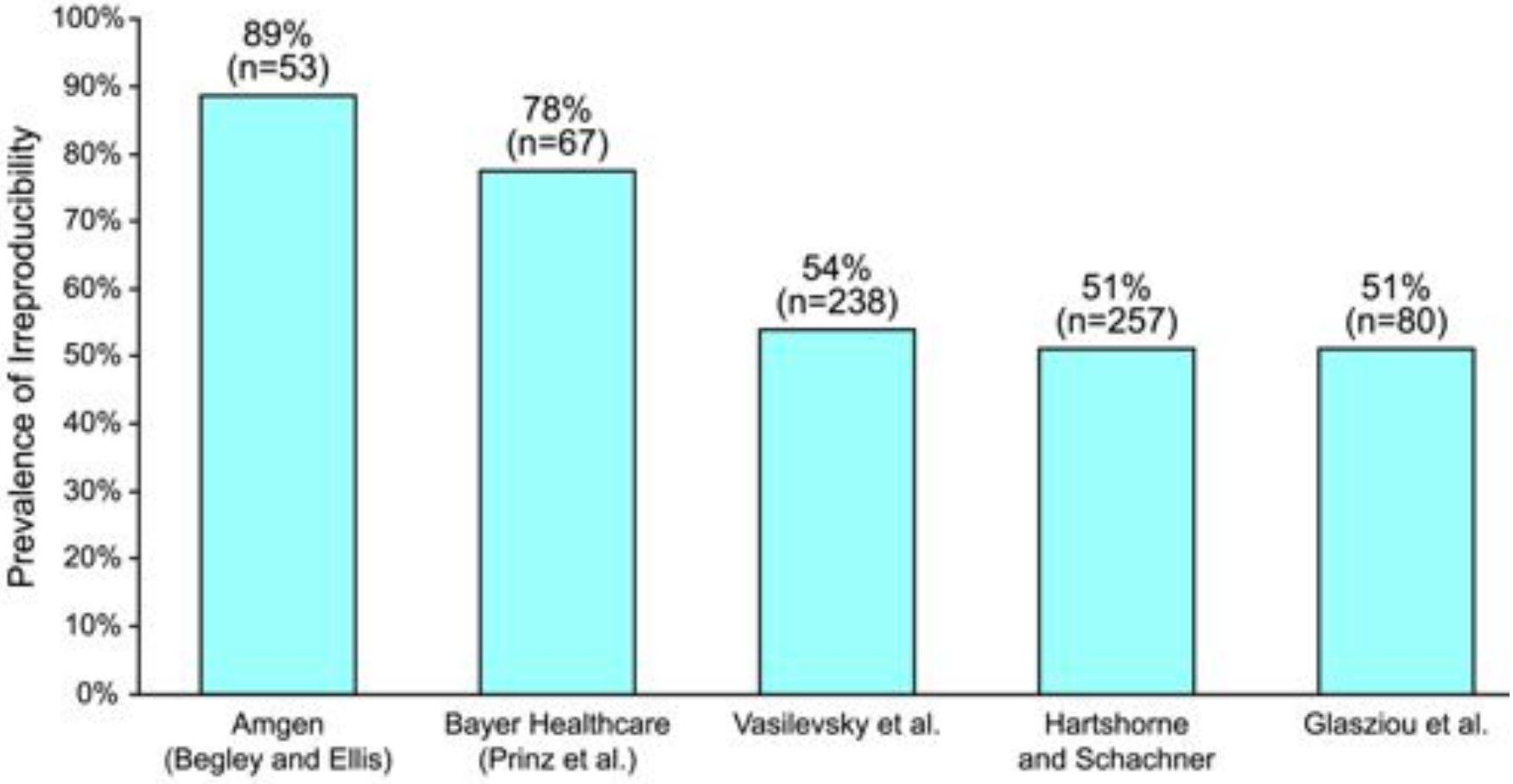


Number of publications per year, from 1970 through April 2016

THE COST OF REPRODUCIBILITY

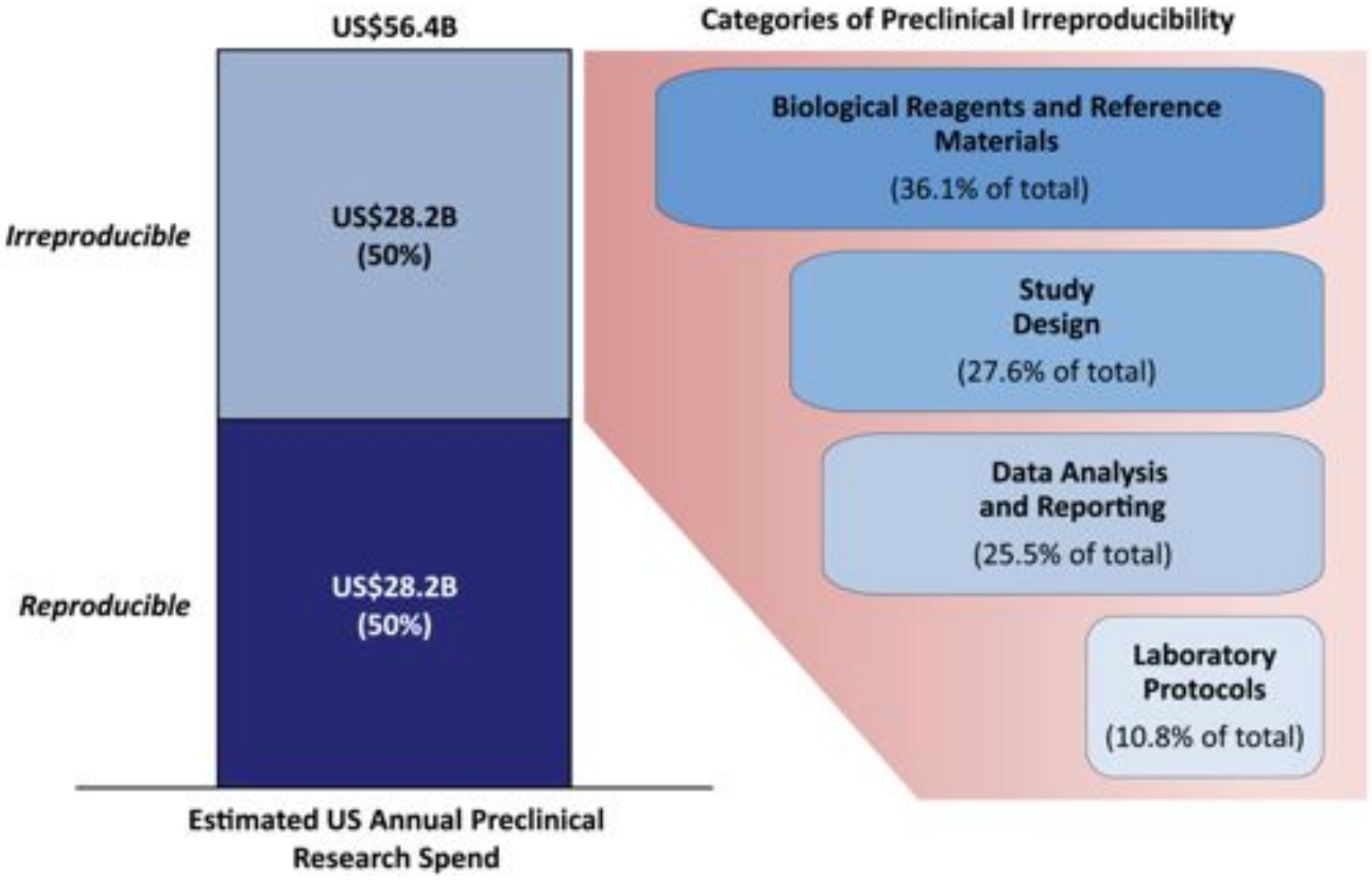
Image credit: The Cost of Living

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

REPRODUCIBILITY IN DATA SCIENCE

Image credit: Bill & Melinda Gates Foundation

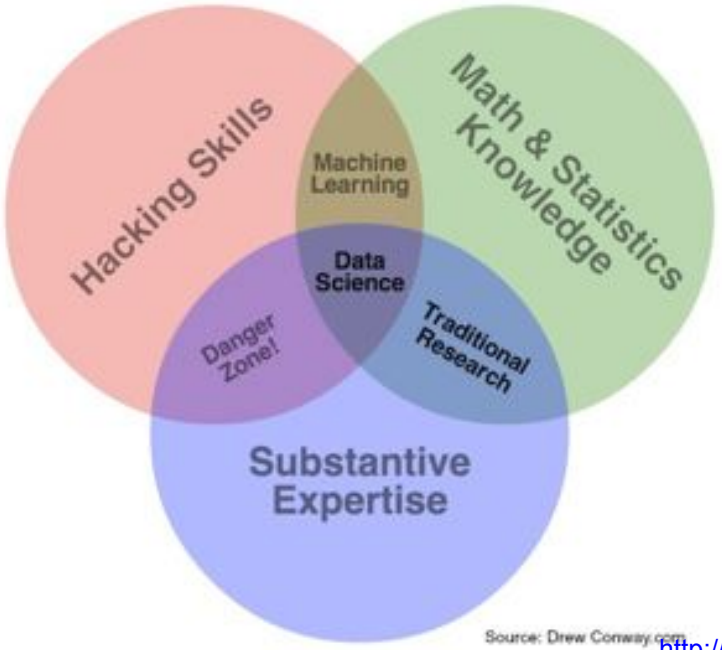
Reproducibility in data science

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.jblumenstock.com/teaching/course=infx573

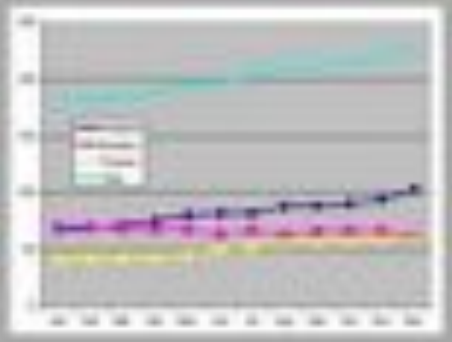
http://bigdata-madesimple.com/what-everybody-ought-to-know-about-data-scientist/

DATA SCIENTIST

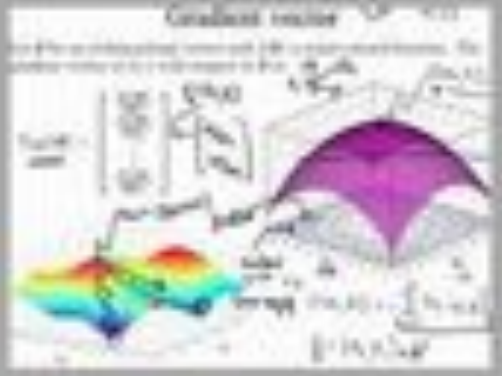
What my friends think I do

What my mom thinks I do

What society thinks I do



What my boss thinks I do



What I think I do

What I actually do

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

- 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

The importance of software

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

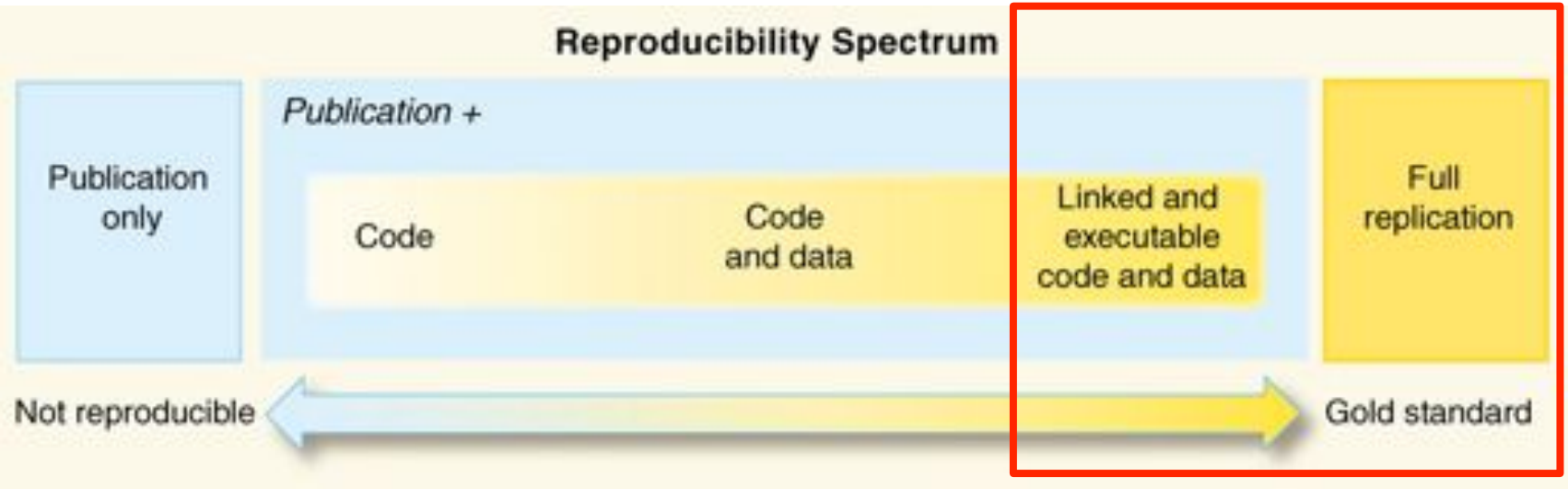


Image credit: Roger Peng "Reproducible Research in Computational Science" *Science* 2011
http://science.sciencemag.org/content/334/6060/1226

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?

Six degrees of reproducibility

- **5:** 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.)

- **4:** Easy reproducibility, but require some proprietary source packages (MATLAB, SAS, etc.)

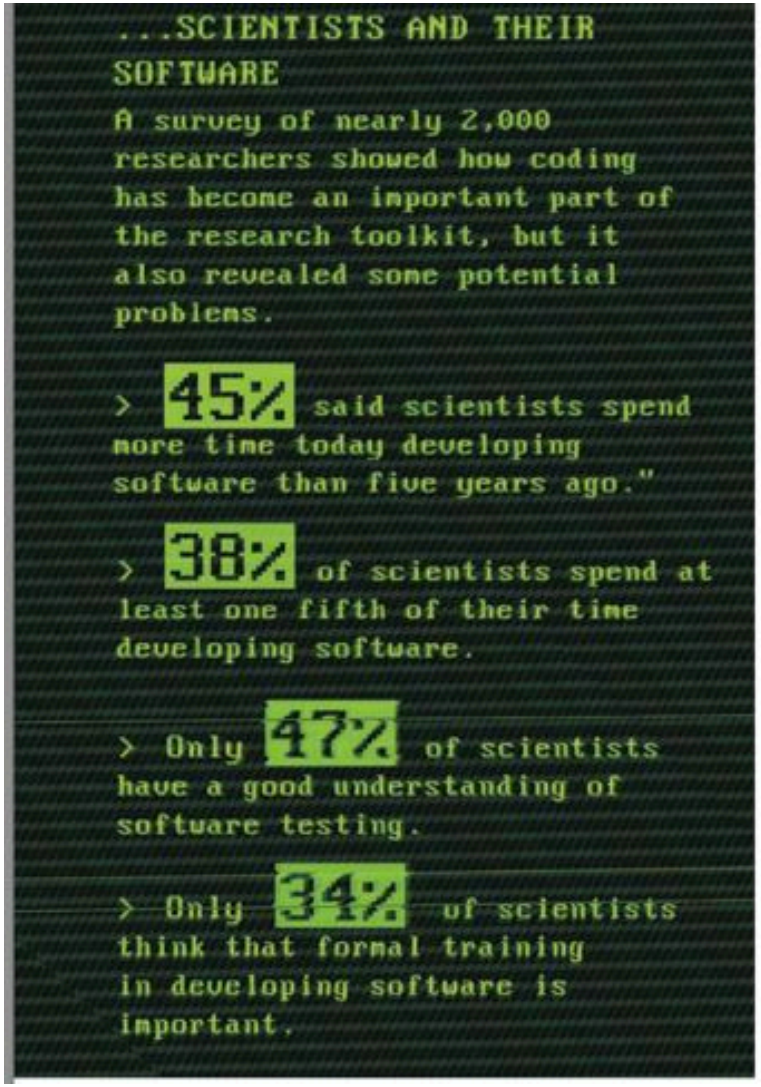
- **3:** Reproducibility requires considerable effort

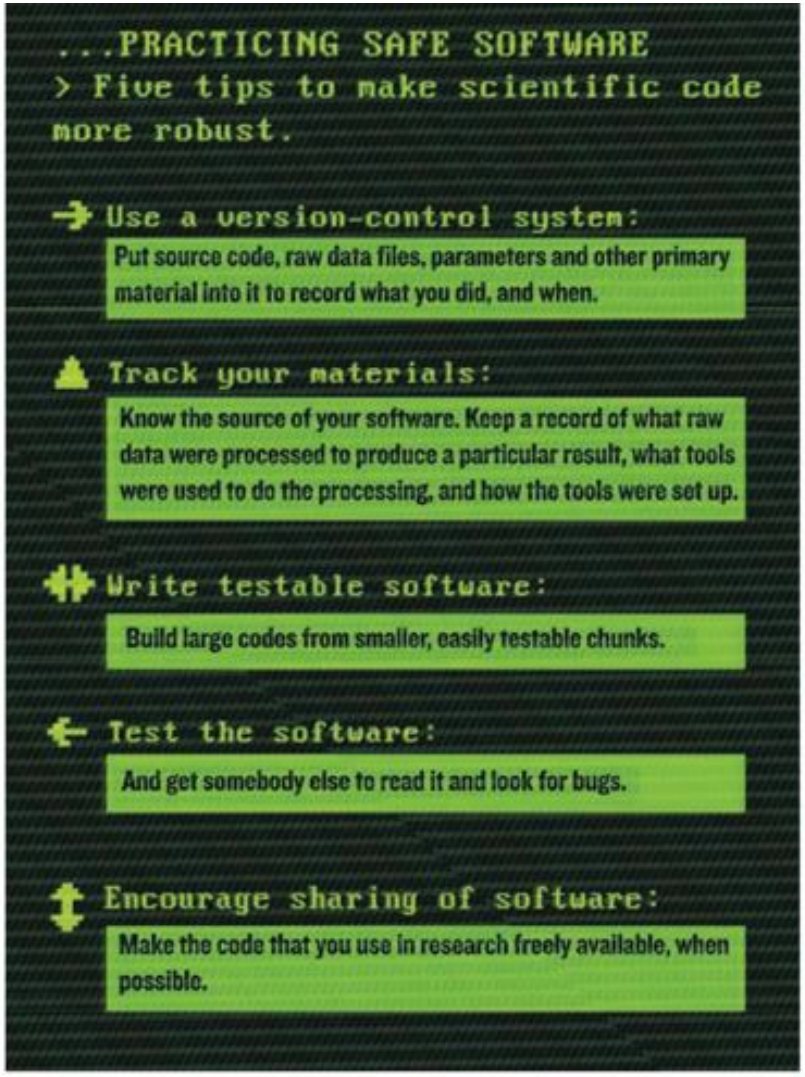
- **2:** Reproducibility requires extreme effort

- **1:** The results cannot seem to be reproduced

- **0:** The results cannot be reproduced

Scientific computing





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