BIOS 259:
The Art of Reproducible Science
A Hands-on Approach

Instructor: Aziz Khan <azizk@stanford.ed>

Section instructor/TA: Alvina Adimoelja <alvinaa@stanford.edu>

https://github.com/asntech/bios259-w24/

A little about instructors

Alvina Adimoelja (she/her)

PhD student, Curtis/Pritchard labs

Genetics Department

BS (Cell and Systems Biology) University of Oxford

Research interests: Cancer population genetics and evolution, genomics, spatial-omics

A fun fact: I have two cats (Dill & Sandy -> Telemachus & Andromache)

Aziz Khan (he/him)

Staff Scientist, Curtis Lab

Stanford Cancer Institute

PhD (Bioinformatics), Tsinghua University Beijing

Postdoc (Gene Regulation), University of Oslo, Norway

Research interests: Gene regulation, cancer evolution, data informatics, resource development (JASPAR, UniBind, Intervene,..)

A fun fact: I'm a trained farmer and shepherd

Twitter: @khanaziz84

A little about you!

Let’ do a **think-pair share** exercise

Use an index card and please write:

- Your name and pronouns
- Your discipline/research background
- What reproducibility can do for you?
- A fun fact about you!

A little about you!

Let’ do a **think-pair share** exercise

Now share with a colleague next:

- Your name and pronouns
- Your discipline/research background
- What reproducibility can do for you?
- A fun fact about you!

Course schedule

Date	Topic	Time	Location
02.26.2024 – Monday	Introduction to reproducibility and setting up	10:00-13:00	Alway Building M218A
02.28.2024 – Wednesday	Version Control (Git/GitHub)	10:00-13:00	M218A
03.01.2024 – Friday	Environment management (Conda, Bioconda, Mamba)	10:00-13:00	M218A
03.04.2024 – Monday	Containerization (Docker, Singularity)	10:00-13:00	M218A
03.06.2024 – Wednesday	Workflows (Snakemake, Nextflow/nf-core)	10:00-13:00	M218A
03.08.2024 – Friday	Documentation (FAIR data and open code) and wrap up	10:00-13:00	Li Ka Shing LK208

M218A: https://25live.collegenet.com/pro/stanfordsom#!/home/location/1454/details
LK208: https://25live.collegenet.com/pro/stanfordsom#!/home/location/1149/details

Office hours

Date	Instructor name	Time	Location
02.**29**.2024 – Thursday	Aziz	10:00-11:00	BMI 4022
03.**05**.2024 – Tuesday	Aziz	13:00-14:00	BMI 4022
03.**07**.2024 – Thursday	Aziz	10:00-11:00	BMI 4022

Code of conduct

Let's **work together as one community** to share, learn and shine with respect and care for everyone.

- Be a nice, regular and active learner
- Help your peers and please give everyone opportunities to speak
- If you know a concept well, you may want to skip or help a fellow classmate

Along with our course commitments, Stanford University is committed to providing a safe living and learning environment in which every person is valued and respected, inclusion is assured, and free expression and debate are encouraged.

Please visit https://intolerance.stanford.edu to submit a report.

Mode of instructions and exercises (Not graded)

- We will use active/cooperative learning approaches and also some live coding
  - Your active participation and help is very important
- To make the learning fun and interactive we will use Slido for live quizzes
  - Scan the QR and participate
- We will use **peer instructions** and clicker questions
  - Discuss with your peers and help each other
- Think-pair-share exercises
- Please do ask questions

Course level learning goals

- **Understand the importance** and causes of computational reproducibility in research
- Gain **proficiency in version control** system (Git) for collaborative code and data tracking
- Create and share **conda environments** for software dependency management
- Utilize **containerization** tools (e.g., Docker, Singularity) for portable computing environments
- Learn techniques for **automating workflows** and generating reproducible results
- Develop effective strategies for managing and **documenting data and code** to ensure reproducibility
- Implement **best practices** for transparent and reproducible project organization

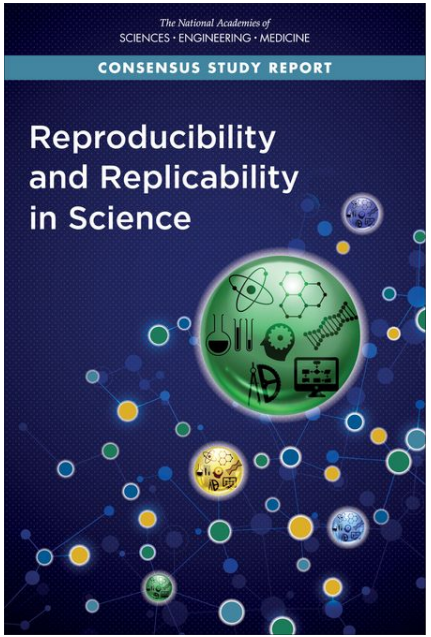
Today’s learning goals

- Explain what reproducibility is and how it differs from replicability.
- Understand why reproducibility is important in science and what are the consequences of irreproducible research
- Explain the factors and reasons for irreproducible research
- List and define tools available to facilitate levels of reproducibility

What is reproducibility?

- **Reproducibility** means *computational reproducibility* — obtaining same results using the ***same input data***, computational steps, methods, code, and conditions of analysis.

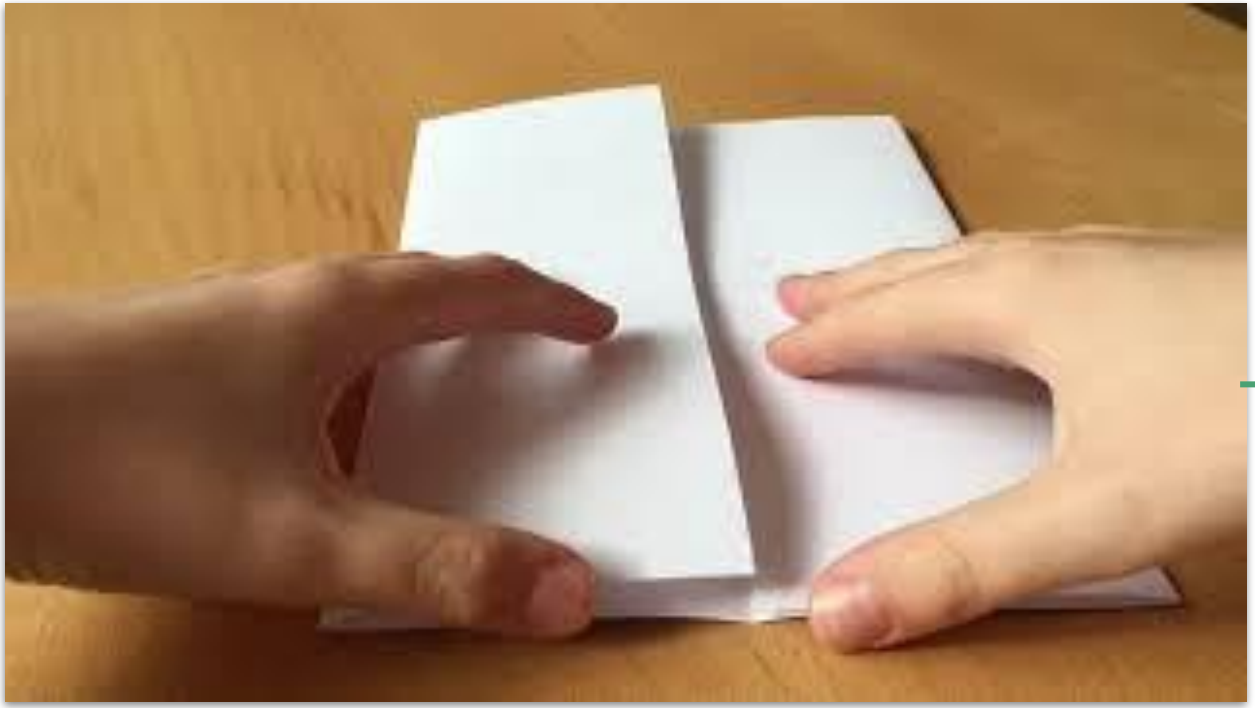
- **Replicability** means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its ***own data***.



http://www.nap.edu/25303

Our first hands-on exercise

Method and steps to create name tag



https://www.youtube.com/watch?v=ZXM-NihVwVk&t=17s

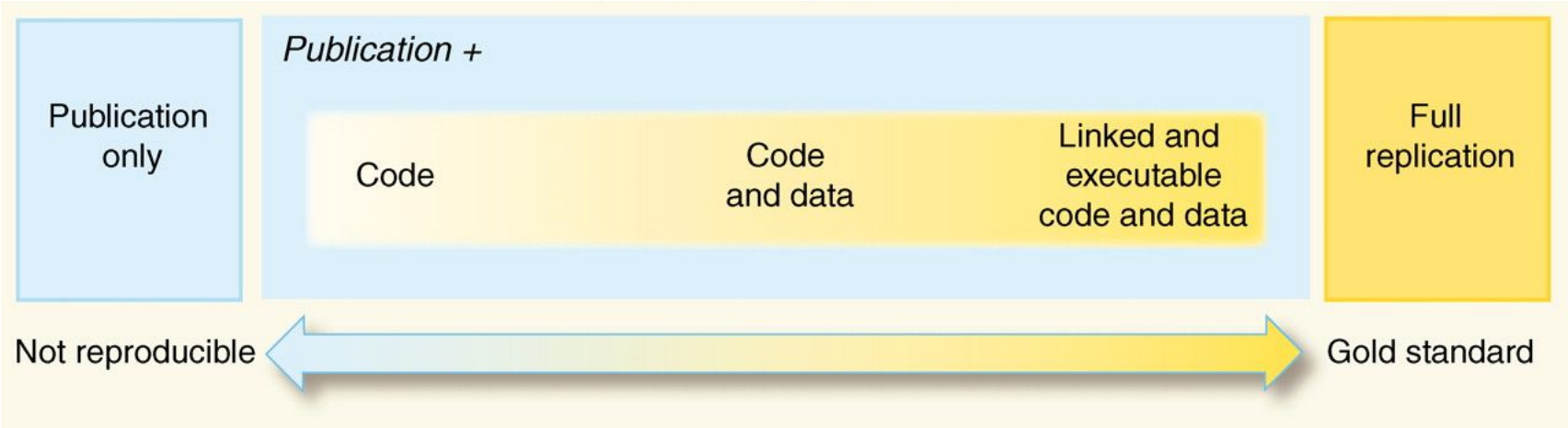
Result/Output

The *Turing Way's* definitions

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Source: The Turing Way

Reproducibility is a spectrum



Poger Peng (2021) - doi: 10.1126/science.1213847

Levels of Research Reproducibility

Trying things out in the R Console pane, saving tables and figures to named files.

Writing code in an RScript or Rmarkdown file, generating saved tables, figures, and manuscripts that can be re-run if needed.

Using Projects and {here} , sharing code, documenting with README and comments, doing code review,
and sharing code publicly on GitHub

All of the above, plus public sharing of code and data, and preserving your local package environment with {renv}

All of the above plus encapsulating the entire computing environment (R, packages, code) in a Docker image.

https://bookdown.org/pdr_higgins/rmrwr/intermediate-steps-toward-reproducibility.html

Reproducibility ≠ Replicability

A rigorously conducted study using the best practices, correctly analyzed, and transparently reported – **should be** computationally **reproducible** but it may **fail to be replicated**.

Reproducibility = least standard for good science

Reproducibility/replicability are hallmarks of good science

- Research is becoming more collaborative, transparent and reproducible
- Reproducibility is more prominent than ever due to growing dependence on data/compte -intensive research in all sciences
- There are growing international and national efforts to mandate these practices and also to develop infrastructure and tools to support reproducibility
- Universities and funders will/should have systems in place to assess the reproducibility scores/index of researchers
- A *Reproducibility Index* will influence hiring, funding, and promotions

Reproducibility good for you first then to science

1. **Track a complete history of your research**
   a. Ensure research sustainability and fair citation/acknowledgement.
   b. Provide useful insights for your own and others' work in the field.
2. **Publish validated research and avoid misinformation**
   a. Improve reproducibility rates and reduce paper retractions.
   b. Ensure accuracy and reliability of research outputs for future studies.
3. **Write your papers, thesis, and reports efficiently**
   a. Maintain easy access to results and acknowledge collaborators' contributions.
   b. Comply with high-level journal guidelines by providing necessary documentation.
4. **Get credits for your work fairly**
   a. Enable reuse and citation of research by providing reproducible components.
   b. Encourage trust and replication of research outcomes for broader impact.
5. **Ensure continuity of your work**
   a. Communicate effectively with stakeholders and potential collaborators.
   b. Enable others to build upon and reuse your research for new applications.

Five selfish reasons to work reproducibility

1. Avoid disasters
2. Easier to write papers
3. Easier to talk to reviewers
4. Continuity of your work/in the lab
5. Build your reputation

Comment | Open access | Published: 08 December 2015

Five selfish reasons to work reproducibly

Florian Markowetz

Genome Biology **16**, Article number: 274 (2015) | Cite this article

22k Accesses | **49** Citations | **474** Altmetric | Metrics

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7

Why it is important?

Ensuring replicability and reproducibility of research becomes critical when new findings have **implications for public health, well-being** and **policy decisions**.

Duke disaster

Anil Potti et al (2006) build a ML model to **predict sensitivity of individuals to chemotherapeutic drugs** using gene expression signatures

Four clinical trials started (breast and lung cancers) – 3 at Duke and 1 at Moffitt and 5th was planned.

The study was not transparent enough and data was not shared

ARTICLES • Retracted •

nature medicine

Genomic signatures to guide the use of chemotherapeutics

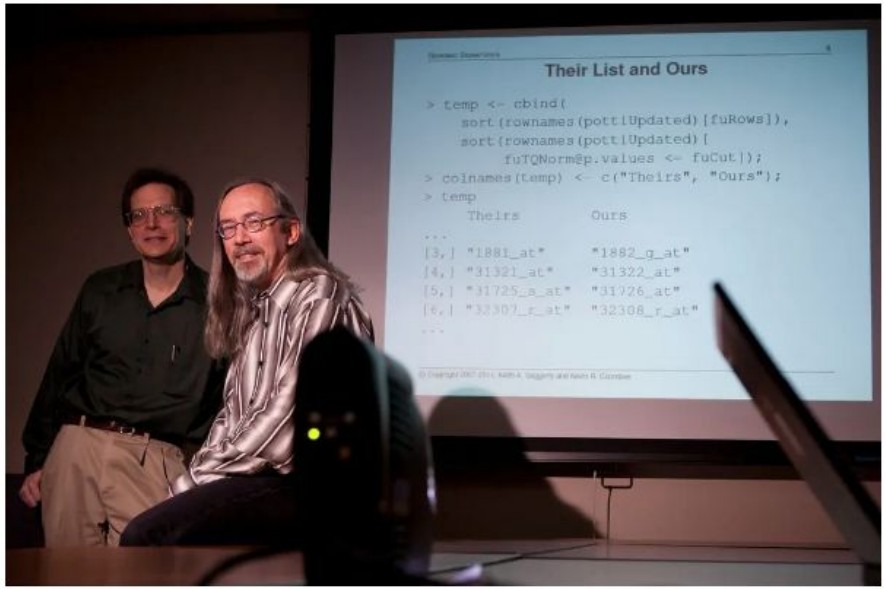
Anil Potti[1,2], Holly K Dressman[1,3], Andrea Bild[1,3], Richard F Riedel[1,2], Gina Chan[4], Robyn Sayer[4], Janiel Cragun[4], Hope Cottrill[4], Michael J Kelley[2], Rebecca Petersen[5], David Harpole[5], Jeffrey Marks[5], Andrew Berchuck[1,6], Geoffrey S Ginsburg[1,2], Phillip Febbo[1–3], Johnathan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

Using *in vitro* drug sensitivity data coupled with Affymetrix microarray data, we developed gene expression signatures that predict sensitivity to individual chemotherapeutic drugs. Each signature was validated with response data from an independent set of cell line studies. We further show that many of these signatures can accurately predict clinical response in individuals treated with these drugs. Notably, signatures developed to predict response to individual agents, when combined, could also predict response to multidrug regimens. Finally, we integrated the chemotherapy response signatures with signatures of oncogenic pathway deregulation to identify new therapeutic strategies that make use of all available drugs. The development of gene expression profiles that can predict response to commonly used cytotoxic agents provides opportunities to better use these drugs, including using them in combination with existing targeted therapies.

The New York Times

How Bright Promise in Cancer Testing Fell Apart

Share full article | 75



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

Gene names off by one and duplicated patients …

Keith Baggerly, and **Kevin Coombes**, Biostatisticians at M. D. Anderson Cancer Center found **several reproducibility issue** including **duplicated patients/cellines** (2009)

The Annals of Applied Statistics
2009, Vol. 3, No. 4, 1309–1334
DOI: 10.1214/09-AOAS291

DERIVING CHEMOSENSITIVITY FROM CELL LINES: FORENSIC BIOINFORMATICS AND REPRODUCIBLE RESEARCH IN HIGH-THROUGHPUT BIOLOGY

By Keith A. Baggerly[1] and Kevin R. Coombes[2]

Nature Medicine retraction (2011)

http://videolectures.net/cancerbioinformatics2010_baggerly_irrh/

HOW EVENTS UNFOLDED

2006

Anil Potti (pictured), a cancer geneticist at Duke University in Durham, North Carolina, and others file patent applications on the idea of using gene-expression data to predict sensitivity to cancer drugs. Potti is first author on a paper in *Nature Medicine*[1].

2007

Potti is last author on a paper in the *Journal of Clinical Oncology* (*JCO*)[2]. Duke begins three clinical trials to test Potti's predictors in patients with breast or lung cancer.

SEPTEMBER 2009

Keith Baggerly (pictured) and Kevin Coombes, statisticians at the University of Texas M. D. Anderson Cancer Centre in Houston, publish a paper in *Annals of Applied Statistics*[3] stating that they could not replicate Potti's claims. Duke suspends the trials and asks a review panel to investigate.

NOVEMBER 2009

Potti places data underlying the *JCO* paper online. Baggerly writes to Sally Kornbluth (pictured), Duke vice-dean for research, and Michael Cuffe, Duke vice-president for medical affairs, to point out differences from raw data.

DECEMBER 2009

An unredacted copy of the report by Duke's review panel, later obtained by *Nature*, shows that the panel replicated Potti's claims using his data, but were unaware that those data contained discrepancies.

JANUARY 2010

Duke restarts clinical trials.

JUNE 2010

Harold Varmus (pictured), director of the National Cancer Institute in Bethesda, Maryland, asks the Institute of Medicine to review Duke's trials.

JULY 2010

The Cancer Letter reveals that Potti made false claims about his CV. Trials are suspended and an investigation begins.

NOVEMBER 2010

JCO paper is retracted. Duke closes the trials permanently. Potti resigns.

DECEMBER 2010

Institute of Medicine study begins, but will now focus more generally on criteria for genomics predictor.

JANUARY 2011

Nature Medicine paper is retracted.

Be cautious with Excel

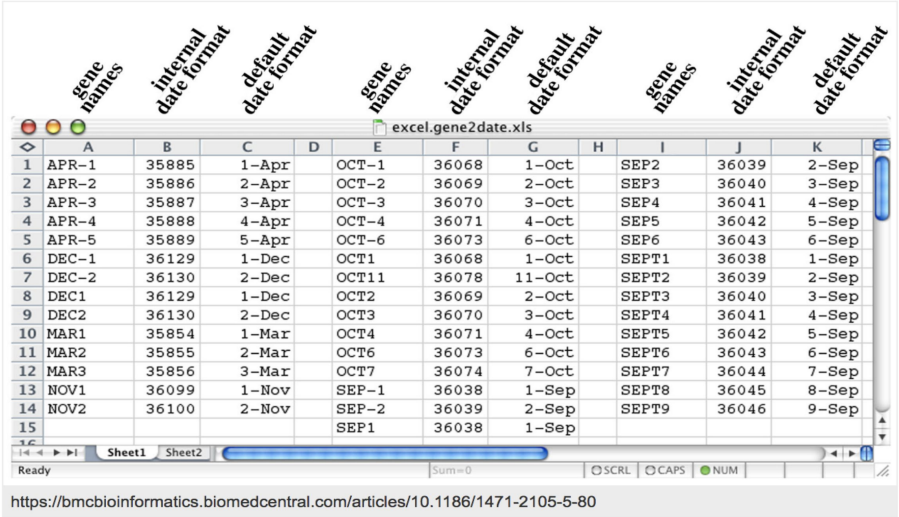
Comment | Open Access | Published: 23 August 2016

Gene name errors are widespread in the scientific literature

Mark Ziemann, Yotam Eren & Assam El-Osta

Genome Biology **17**, Article number: 177 (2016) | Cite this article

115k Accesses | **38** Citations | **2375** Altmetric | Metrics



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr		OCT-1	36068	1-Oct		SEP2	36039	2-Sep
2	APR-2	35886	2-Apr		OCT-2	36069	2-Oct		SEP3	36040	3-Sep
3	APR-3	35887	3-Apr		OCT-3	36070	3-Oct		SEP4	36041	4-Sep
4	APR-4	35888	4-Apr		OCT-4	36071	4-Oct		SEP5	36042	5-Sep
5	APR-5	35889	5-Apr		OCT-6	36073	6-Oct		SEP6	36043	6-Sep
6	DEC-1	36129	1-Dec		OCT1	36068	1-Oct		SEPT1	36038	1-Sep
7	DEC-2	36130	2-Dec		OCT11	36078	11-Oct		SEPT2	36039	2-Sep
8	DEC1	36129	1-Dec		OCT2	36069	2-Oct		SEPT3	36040	3-Sep
9	DEC2	36130	2-Dec		OCT3	36070	3-Oct		SEPT4	36041	4-Sep
10	MAR1	35854	1-Mar		OCT4	36071	4-Oct		SEPT5	36042	5-Sep
11	MAR2	35855	2-Mar		OCT6	36073	6-Oct		SEPT6	36043	6-Sep
12	MAR3	35856	3-Mar		OCT7	36074	7-Oct		SEPT7	36044	7-Sep
13	NOV1	36099	1-Nov		SEP-1	36038	1-Sep		SEPT8	36045	8-Sep
14	NOV2	36100	2-Nov		SEP-2	36039	2-Sep		SEPT9	36046	9-Sep
15					SEP1	36038	1-Sep				

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

SCIENCE / TECH / MICROSOFT

Scientists rename human genes to stop Microsoft Excel from misreading them as dates

Illustration by Alex Castro / The Verge

/ Sometimes it's easier to rewrite genetics than update Excel

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 5:44 AM PDT

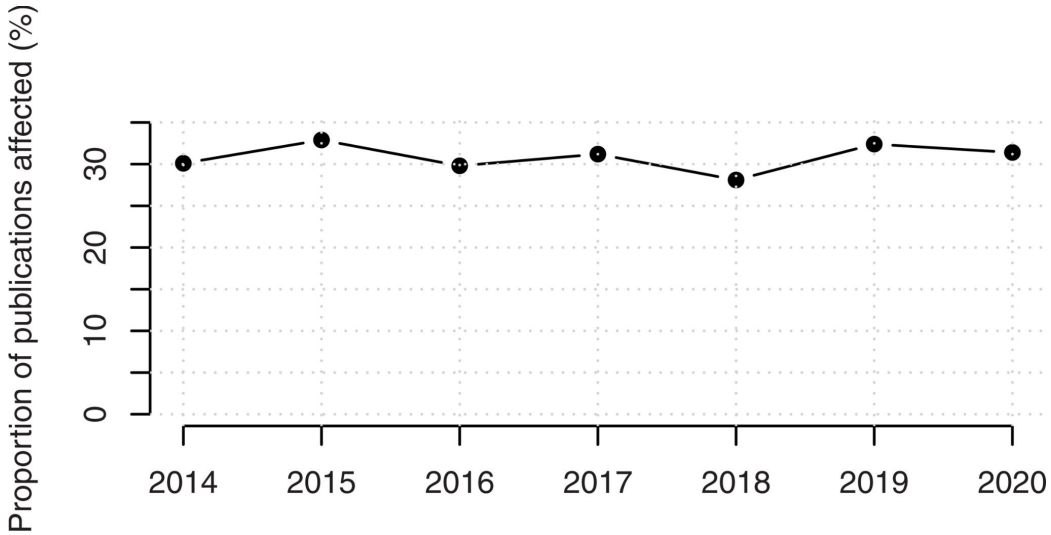
0 Comments (0 New)

If you buy something from a Verge link, Vox Media may earn a commission. See our ethics statement.

https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates

Gene name errors is still a problem

Scanning PubMed articles with supplementary Excel gene lists showed 30.9% (3,436/11,117) gene name errors.

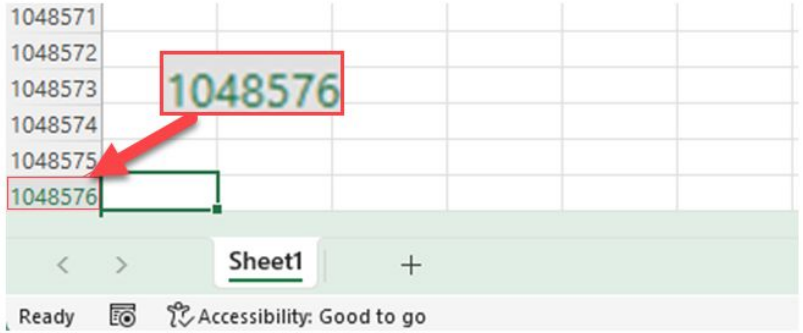


https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008984

https://github.com/jennybc/scary-excel-stories

16,000 COVID cases went unreported in England

- Excel has Row and Column Capacity limits
- Microsoft 365 Excel, the maximum number of rows that it can handle is 1,048,576
- Older versions can handle 65,536
- The maximum number of columns are 16,384



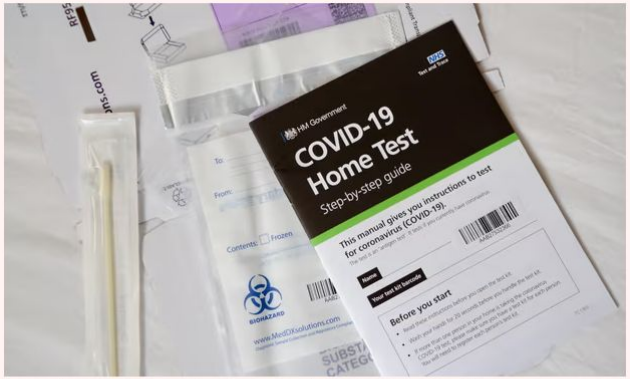
Analysis

Covid: how Excel may have caused loss of 16,000 test results in England

Alex Hern
UK technology editor

Public Health England data error blamed on limitations of Microsoft spreadsheet

- Coronavirus - latest updates
- See all our coronavirus coverage



More than 50,000 potentially infectious people may have been missed by contact tracers after 15,841 positive tests were left off the daily figures. Photograph: Simon Leigh/Alamy

A million-row limit on Microsoft's Excel spreadsheet software may have led to Public Health England misplacing nearly 16,000 Covid test results, it is understood.

Neutrinos travel faster than light anomaly

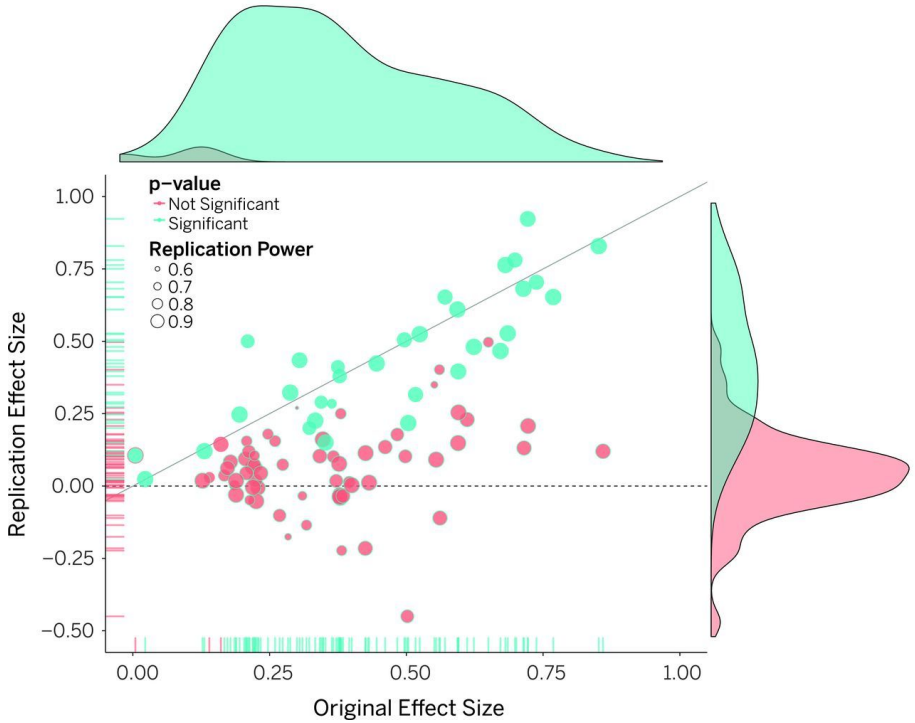
- The 2011 OPERA experiment mistakenly observed **neutrinos appearing to travel faster than light**.
- This **violates special relativity**, a cornerstone of the modern understanding of physics for over a century.
- It took repeated efforts before they found it was due **a fiber-optic cable attached improperly**

OPERA (Oscillation Project with Emulsion-tRacking Apparatus)

https://en.wikipedia.org/wiki/2011_OPERA_faster-than-light_neutrino_anomaly

Reproducibility Project in Psychology

- Replication of **100 experiments** reported in papers published in 2008 in three high-ranking psychology journals.
- A large effort started in 2011 and published their findings in 2015
- **36% of replications had statistically significant results**



RESEARCH ARTICLE

Estimating the reproducibility of psychological science

OPEN SCIENCE COLLABORATION Authors Info & Affiliations

SCIENCE · 28 Aug 2015 · Vol 349, Issue 6251 · DOI: 10.1126/science.aac4716

Reproducibility Project in Cancer Biology

- An 8-year effort to replicate experiments from high-impact cancer biology papers published **between 2010 and 2012**.
- 200 individuals contributed in some way to complete this project with a ~$2 million grant from Arnold Ventures
- Started with **193 experiments** from **53 papers**
- ~46% of effects replicated successfully on more criteria than they failed

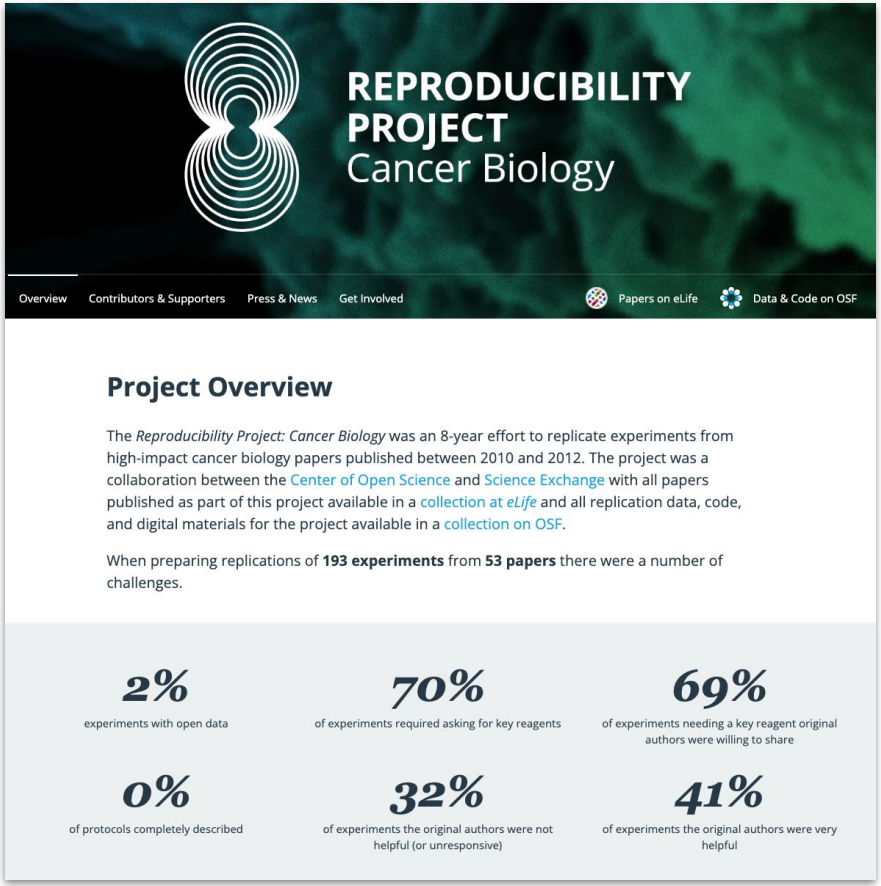
CENTER FOR
OPEN SCIENCE

ScienceExchange

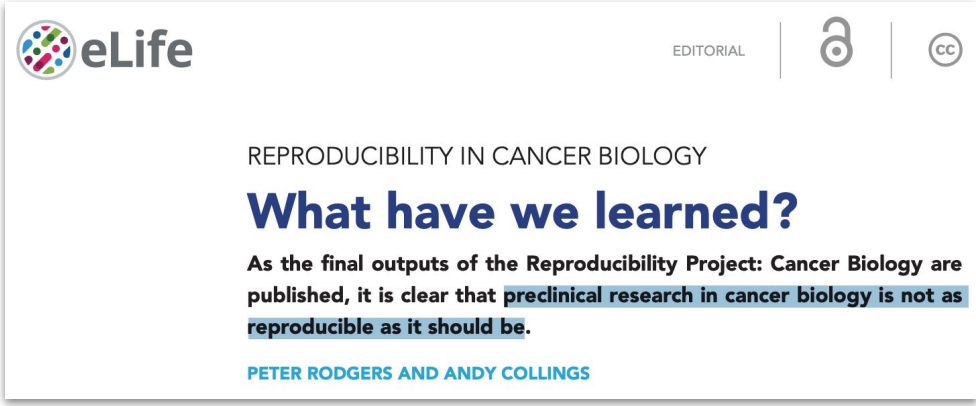
https://www.cos.io/rpcb; https://doi.org/10.7554/eLife.67995

Initial challenges for replicating 193 experiments from 53 papers:

Due to a number of challenges only **50 replication experiments from 23 of the original papers** were completed



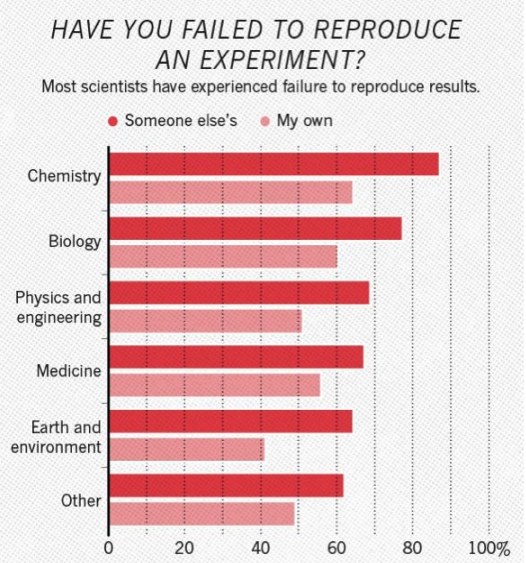
“Preclinical research in cancer biology is not as reproducible as it should be”

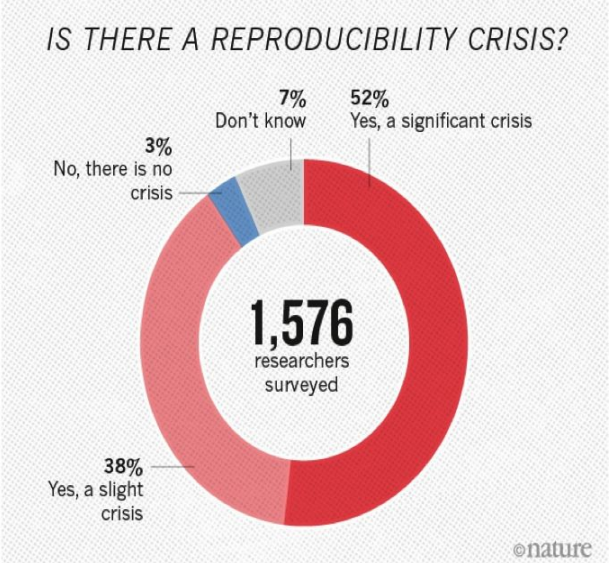


https://www.cos.io/rpcb; https://doi.org/10.7554/eLife.67995

Nature’s 2016 survey on reproducibility - 1,576 responders

- > 70% of researchers have tried and failed to reproduce another scientist's experiments
- > 50% failed to reproduce their own experiments
- 90% said, significant or slight “reproducibility crisis”





Number of respondents from each discipline:
Biology **703**, Chemistry **106**, Earth and environmental **95**, Medicine **203**, Physics and engineering **236**, Other **233**

©nature

https://www.nature.com/articles/533452a

Who is responsible?

The **responsibility** is **shared among various stakeholders** in research ecosystem.

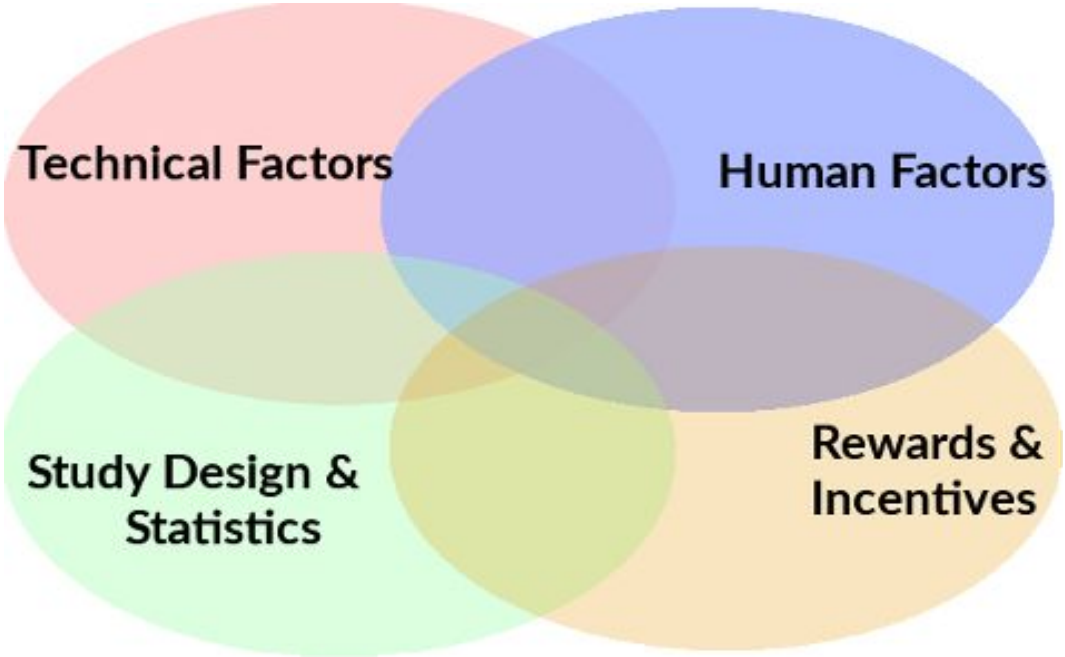
1. **Researchers**
   - Uphold scientific rigor – **document** methods, data, code, and make those **open access**
   - **Train students** about research rigor and reproducibility
2. **Institutions**
   - Establish **policies and guidelines** for transparent research practices and oversee those
   - Develop tools to assess a researchers **reproducibility score**: help in hiring and promotions
3. **Funding agencies**
   - **Mandate** open code/data and set expectations and requirements for grantees
   - Funding **incentives for reproducible** and open researchers
4. **Journals**
   - Ensure code/data are available (compliance with **funder's mandate**)
   - Publish **replication** studies and **negative results**
5. **Scientific societies/public**
   - **Demand** for transparency and accountability in research
   - **Advocate for policies** and initiatives that promote open science and reproducibility

Think-pair share exercise

Take 1-2 minutes to think and write on paper:

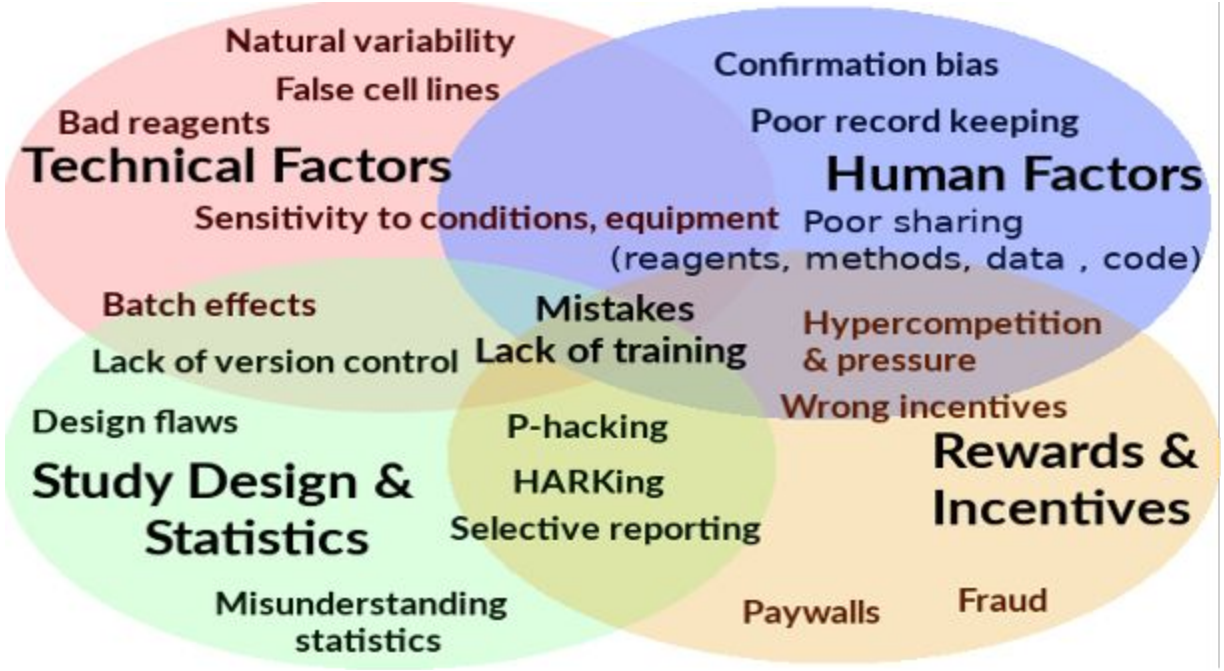
- Name 1-2 major factor(s) impacting reproducibility?
- How we can fix those problems?

Factors impacting reproducibility



https://doi.org/10.6084/m9.figshare.8267630.v1

Factors impacting reproducibility



https://doi.org/10.6084/m9.figshare.8267630.v1

Easy to fix factors behind irreproducible research

- **Not enough documentation** on how experiment is conducted and data is generated
- **Data** used to generate original results **unavailable**
- **Software** used to generate original results **unavailable**
- Difficult to recreate **software environment** (libraries, **versions**) used to generate original results
- **Difficult** to rerun the **computational steps**

https://eglerean.github.io/reproducible-research/01-motivation/

How to ensure computational reproducibility?

- Better **organization of projects** and record software versions and dependencies
- Git **version control** for your code
- Jupyter/R **Notebooks** for documentation
- **Automate** your scripts/pipeline (Nextflow, Snakemake, GNU make ..)
- **Containers** (docker, singularity, biocontainers)

Organize your projects well

Your should think about reproducibility:

When you **create the first folder** with project name**:**

```
> cd projects
> mkdir <project_name>
```

Not when your PI tells you to put the code in a repository **because the Journal and/or the funding agency mandates it**.

Organize your projects well

- A good starting point is to keep all files associated with a project in a **single folder**
- **Different projects** should have **separate folders**
- Use **consistent and informative directory structure**
- Add a **README file** to describe the project and instructions on reproducing the results
- Use **.gitignore**
- Your mileage may vary: it's **not a one-size-fits-all**

```
project_name/
├── README.md
├── data/
│   ├── README.md
│   └── sub-folder/
├── processed_data/
├── manuscript/
├── results/
├── src/
│   ├── LICENSE
│   ├── requirements.txt
│   └── ...
└── doc/
    ├── index.rst
    └── ...
```

https://eglerean.github.io/reproducible-research/

Use version control for code and data

Git is an **open source distributed version control system**:

- It tracks changes in files, revert to previous versions, compare changes, and collaborate with others
- It enables multiple users to work on the same project simultaneously, coordinating changes and resolving conflicts
- Git repositories serve as backups for your project, ensuring that your work is not lost even if your local machine fails

https://git-scm.com/

Version control and data sharing helps to avoid disasters

BBC

Home News Sport Business Innovation Culture Travel Earth Video Live

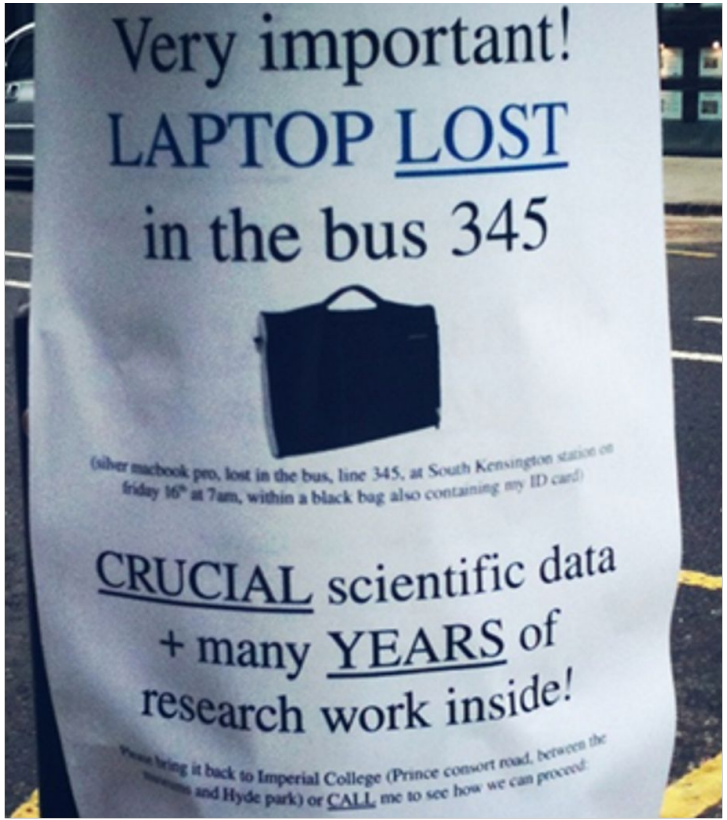
South Africa student fights to keep thesis during robbery

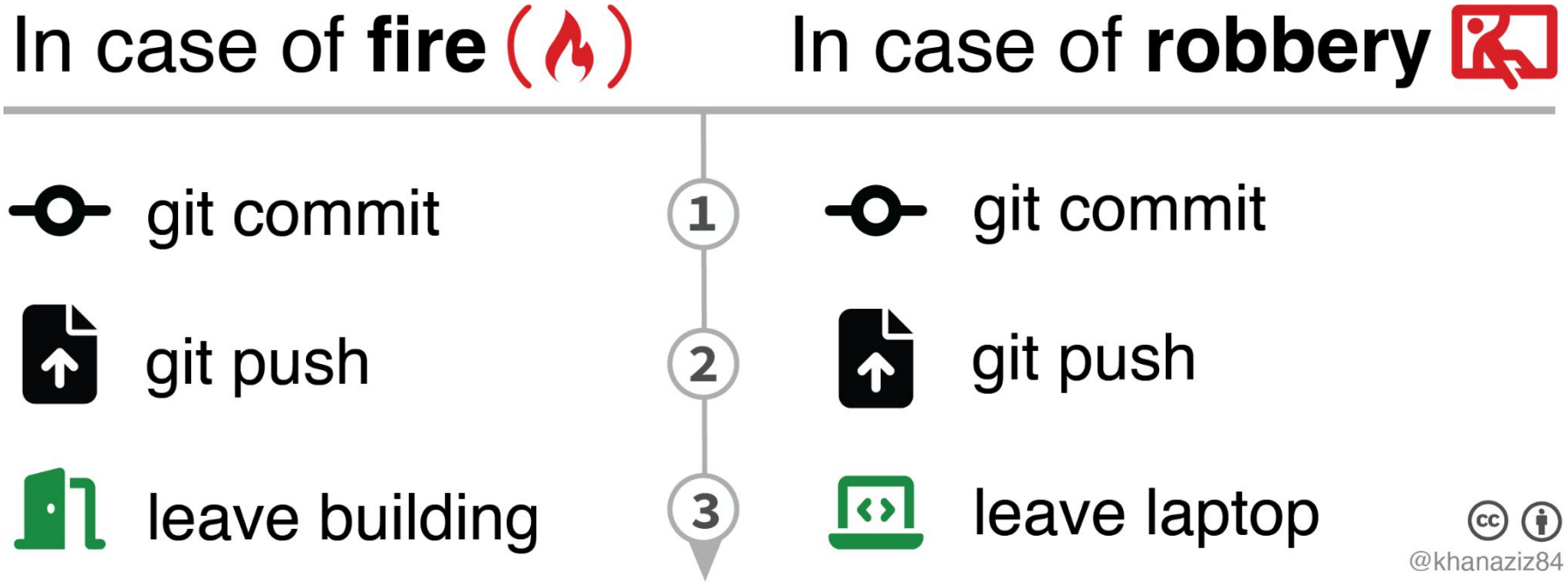
13 September 2017

Share

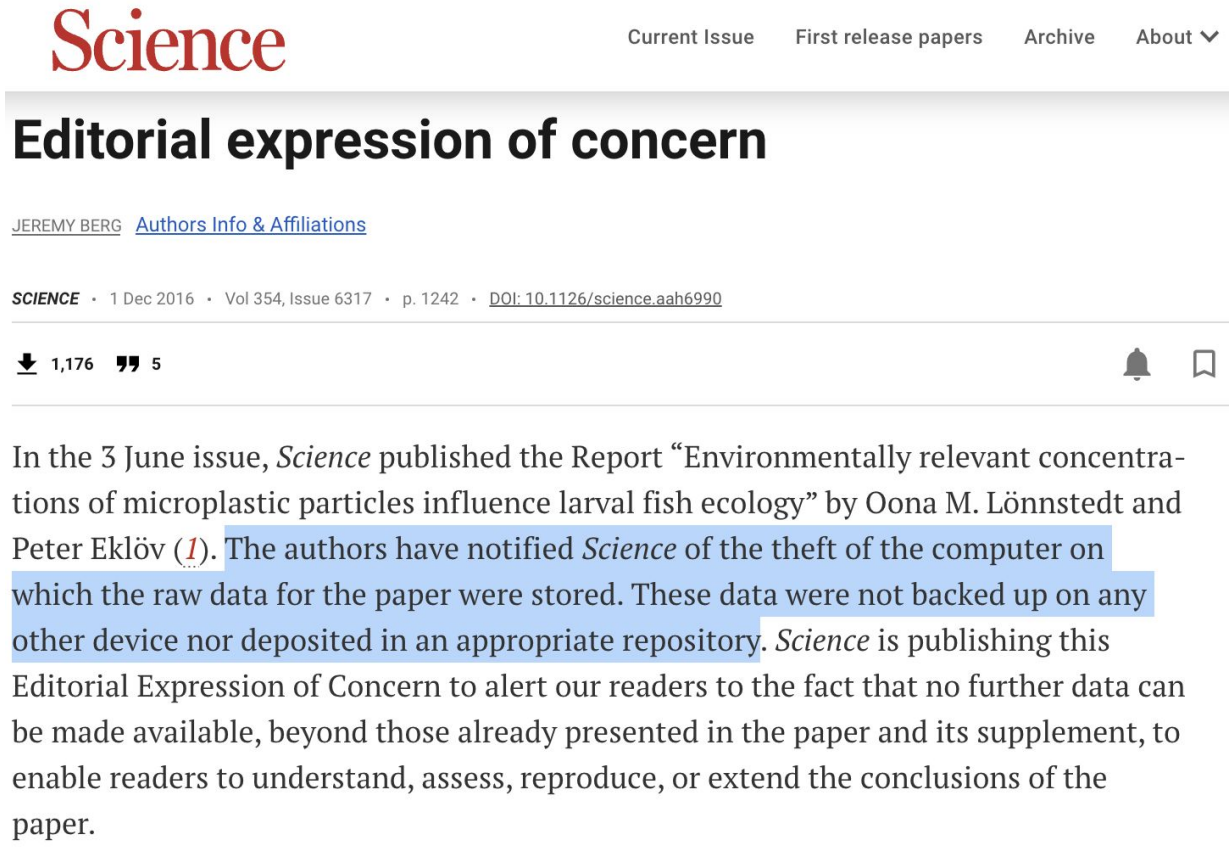
Ms Ntuli was attacked as she returned from work

The prospect of losing the only copy of her master's thesis during a robbery was just too much for one South African student to bear.




In case of fire
In case of robbery
git commit
1
git commit
git push
2
git push
leave building
3
leave laptop
@khanaziz84

"The authors have notified Science of the **theft of the computer** on which the **raw data for the paper** were stored. These data were **not backed up** on any other device nor **deposited in an appropriate repository**."



Science

Current Issue First release papers Archive About

Editorial expression of concern

JEREMY BERG Authors Info & Affiliations

SCIENCE • 1 Dec 2016 • Vol 354, Issue 6317 • p. 1242 • DOI: 10.1126/science.aah6990

1,176 5

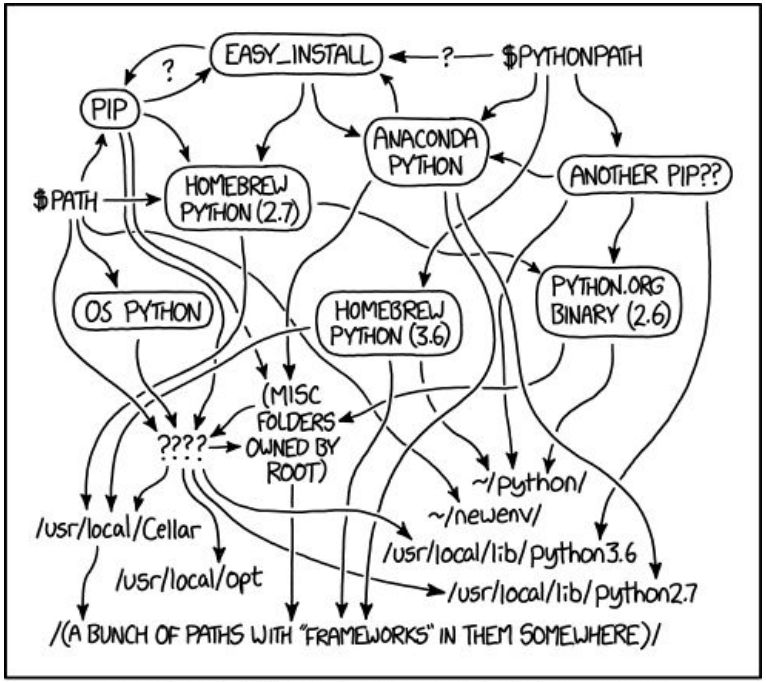
In the 3 June issue, *Science* published the Report "Environmentally relevant concentrations of microplastic particles influence larval fish ecology" by Oona M. Lönnstedt and Peter Eklöv (*1*). The authors have notified *Science* of the theft of the computer on which the raw data for the paper were stored. These data were not backed up on any other device nor deposited in an appropriate repository. *Science* is publishing this Editorial Expression of Concern to alert our readers to the fact that no further data can be made available, beyond those already presented in the paper and its supplement, to enable readers to understand, assess, reproduce, or extend the conclusions of the paper.

https://www.science.org/doi/10.1126/science.aah6990

Track software version and avoid dependency hell

Our codes often depend on other codes that in turn depend on other codes ...

We can control our code but how can we control dependencies?



Conda, Biocoda, Mamba, pip, Lmod, Virtualenv ...

- Installing a **specific set of dependencies**, possibly with well defined versions
- **Recording the versions** for all dependencies
- **Isolate environments** on your computer for projects that have conflicting dependencies
- Isolate environments on computers with many users
- Using **different Python/R versions** per project

Load environment modules with versions

A not reproducible way:

```
> module load samtools
```

A reproducible way:

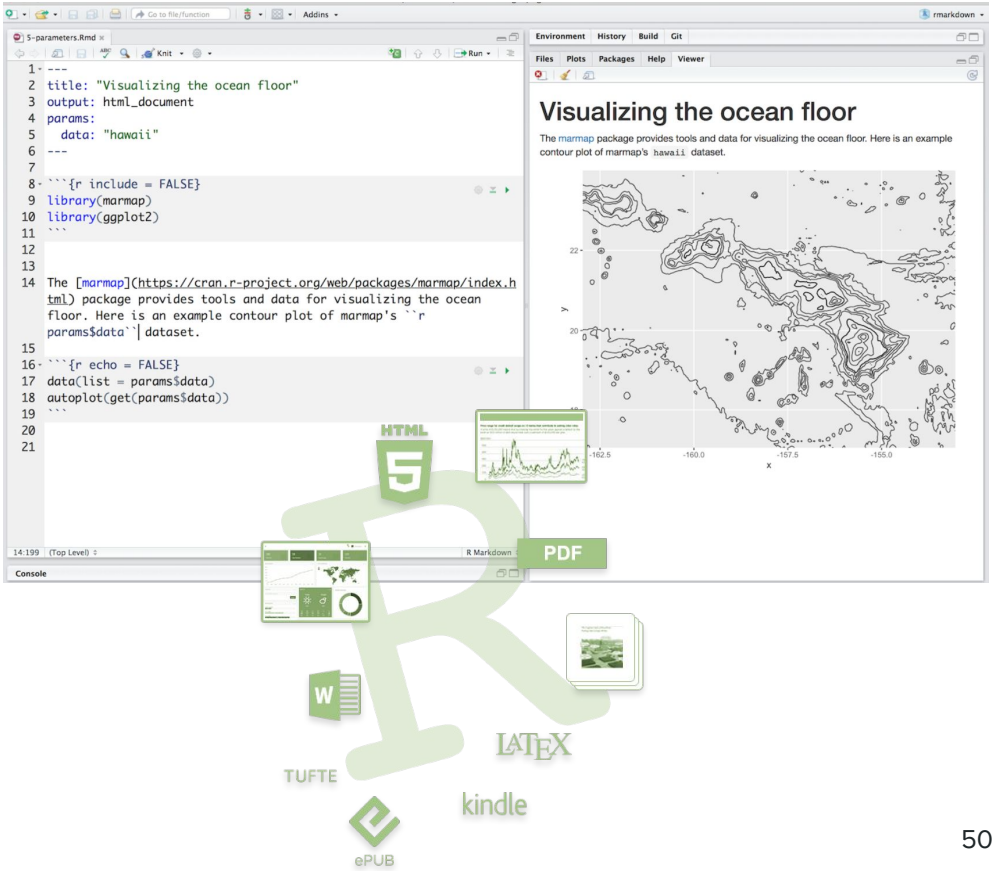
```
> module load samtools/1.15
```

Find the latest version:

```
> module avail samtools
```

Do literate programming: Document using R Markdown

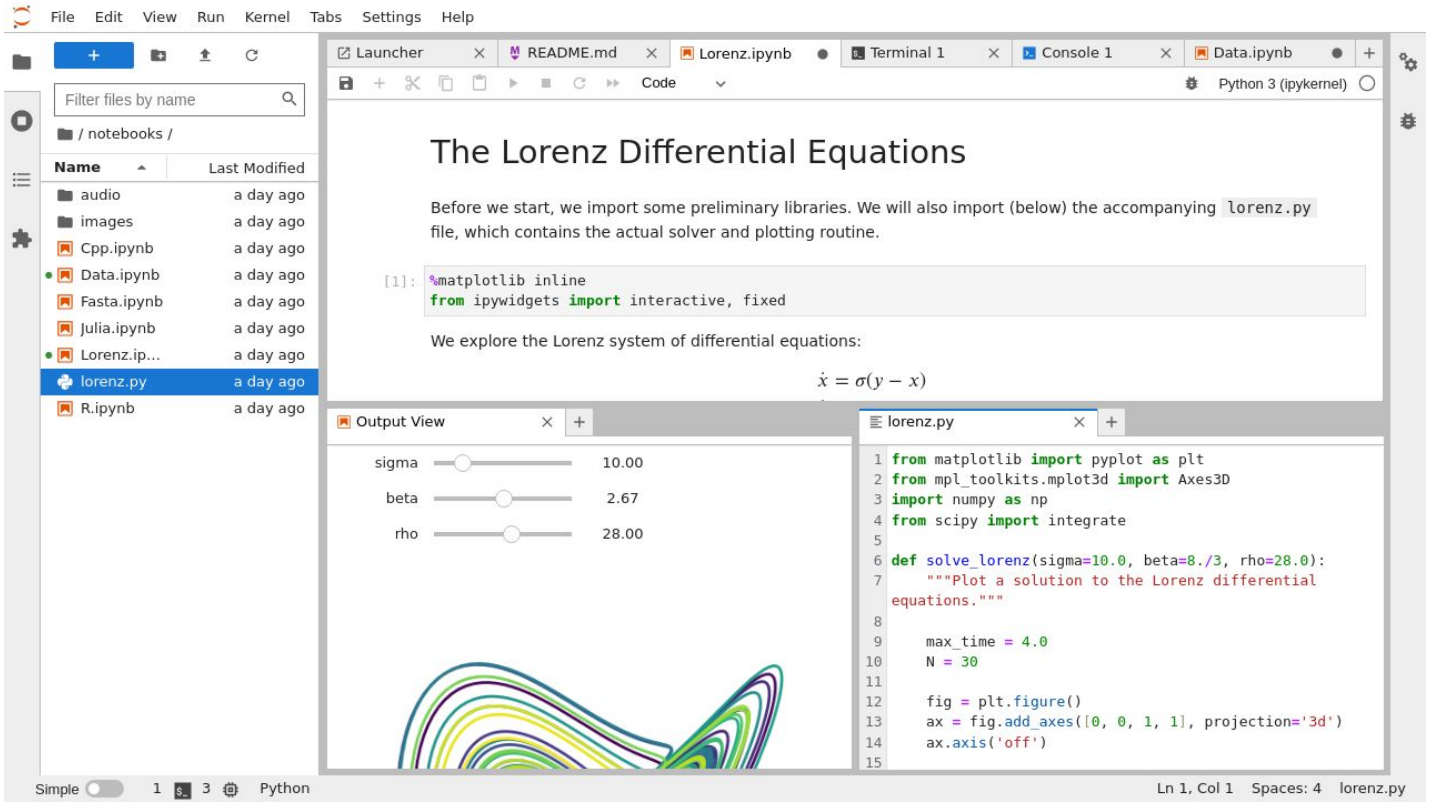
- R Markdown documents are fully reproducible
- Use notebooks to put together narrative text and code to produce figures
- You can use multiple languages including R, Python, and SQL
- Export the outputs in different readable formats



Do literate programming: Document with Jupyter Notebook

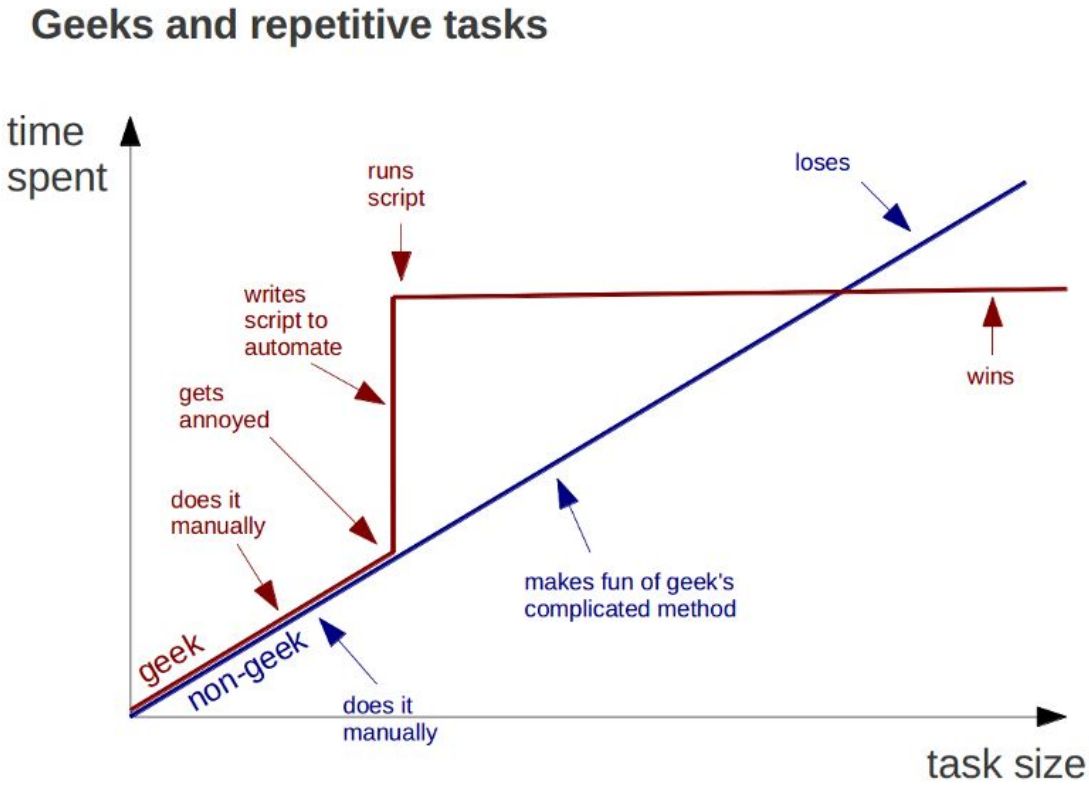
Jupyter is like R Markdown but for Python

Jupyter.org



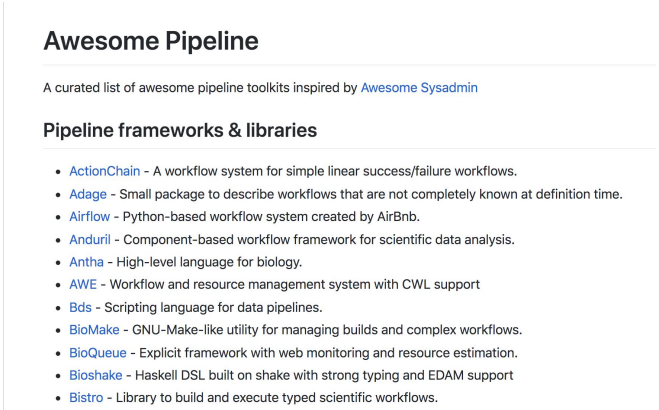
Automation helps reproducibility and saves time

The best documentation is automation



Tools for automation of computational steps

- If you have a repetitive simple task, put them in to a shell
- Good old GNU make
- More recent snakemake, nextflow, WDL etc.



Awesome Pipeline

A curated list of awesome pipeline toolkits inspired by Awesome Sysadmin

Pipeline frameworks & libraries

- ActionChain - A workflow system for simple linear success/failure workflows.
- Adage - Small package to describe workflows that are not completely known at definition time.
- Airflow - Python-based workflow system created by AirBnb.
- Anduril - Component-based workflow framework for scientific data analysis.
- Antha - High-level language for biology.
- AWE - Workflow and resource management system with CWL support
- Bds - Scripting language for data pipelines.
- BioMake - GNU-Make-like utility for managing builds and complex workflows.
- BioQueue - Explicit framework with web monitoring and resource estimation.
- Bioshake - Haskell DSL built on shake with strong typing and EDAM support
- Bistro - Library to build and execute typed scientific workflows.

https://github.com/pditommaso/awesome-pipeline

nextflow

Snakemake

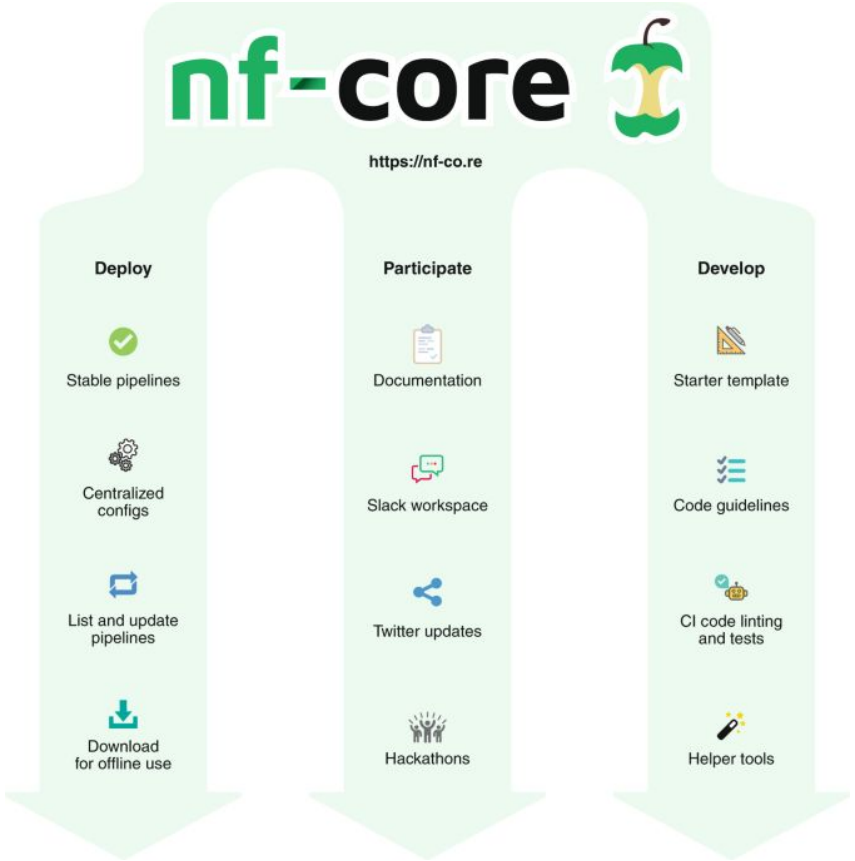
{wdl}

Toil

GNU Make

nf-core

- Data-type specific **Nextflow** pipelines
- Use best practices
- Highly reproducible
- Version controlled
- Widely used
- Active and growing community



nextflow

docker OR OR CONDA

Scalability and reproducibility via containerized environments

Containers can be **built** to bundle all the necessary ingredients (**data, code, environment**).

- **Lightweight**: Share the host OS kernel, reducing overhead
- **Portable**: Can run on any platform with Docker installed
- **Scalable**: Easily deploy and scale applications with Docker containers
- **Consistency**: Ensure consistent environments across development, testing, and production

- Popular container implementations are **Docker** and **Singularity**

Live quiz

When you should worry about reproducibility?

- Before you start the project
- While you write the code and do the analysis
- When you write the paper as lead author
- When you co-author a paper
- When you review a paper
- All the time

Adapted from Dr. Markowetz

slido

When you should worry about reproducibility?

ⓘ Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

A few key principles to ensure reproducibility: summary

- Make projects tiddy and give meaningful names to files
- Use end-to-end containerized workflows for key analysis
- Use R/Python Notebooks for downstream analysis and visualization using version control environments per project
- Open licensing should be used for code and data where possible
- Data must be available and accessible through FAIR principles
- Protocols and methods must be available and accessible
- All 3rd party data and software should be cited
- Comply with funding agency and institutional requirements

“Reproducibility is like brushing your teeth. Once you learn it, it becomes a habit.”

Irakli Loladze, Bryan College of Health Sciences in Lincoln, Nebraska

Reading material

1. Reproducibility and Replicability in Science (2019) by NAS https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science
2. Wilson G, et al. (2017) Good enough practices in scientific computing. PLOS Computational Biology 13(6). https://doi.org/10.1371/journal.pcbi.1005510
3. Noble WS (2009) A Quick Guide to Organizing Computational Biology Projects. PLOS Computational Biology 5(7). https://doi.org/10.1371/journal.pcbi.1000424
4. Wilson G et al. (2014) Best Practices for Scientific Computing. PLOS Biology 12(1): e1001745. https://doi.org/10.1371/journal.pbio.1001745
5. Sandve GK et al. (2013) Ten Simple Rules for Reproducible Computational Research. PLOS Computational Biology 9(10). https://doi.org/10.1371/journal.pcbi.1003285
6. Alnasir JJ (2021) Fifteen quick tips for success with HPC, i.e., responsibly BASHing that Linux cluster. PLOS Computational Biology 17(8). https://doi.org/10.1371/journal.pcbi.1009207
7. Ziemann M, Poulain P, Bora A (2023). The five pillars of computational reproducibility: bioinformatics and beyond. Brief Bioinform.;24(6) doi:10.1093/bib/bbad375
8. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7

Resources

1. https://github.com/mkrapp/cookiecutter-reproducible-science
2. https://eglerean.github.io/reproducible-research/
3.