Mohamed bin Zayed University of Artificial Intelligence

Fall 2026, MBZUAI

CBIO7501 - On the Art of Reproducible & Open Science

Reproducibility – Lecture 1

Sharing papers, data, code and methods so that others can check, reuse and build build

Aziz Khan, PhD
Assistant Professor of Computational Biology, MBZUAI - khanlab.bio
aziz.khan@mbzuai.ac.ae

Mohamed bin Zayed University of Artificial Intelligence

Class schedule and location

Day	Session	Time	Location
Monday	Lecture	3:00 – 4:20 pm	LH2
Tuesday	Lecture	4:00 – 5:20 pm	CR1
Wednesday	Lab	5:00 – 6:50 pm	CR5

Moodle

lms.mbzuai.ac.ae, course 2743

GitHub

github.com/khan-lab/cbio7501-f26

Slack

mbzuai-all.slack.com · #cbio7501-f26

Mohamed bin Zayed University of Artificial Intelligence

Assessment and grading

Component	Weight	Criteria for credit
Attendance	5%	Attend at least 80% of classes
Participation	10%	Active engagement in class discussions and activities
Hands-on activities	45%	Complete at least 80% of the assigned hands-on tasks
Final project presentation	30%	A fully functional reproducibility project
Online assessment	10%	In-class quizzes

Mohamed bin Zayed University of Artificial Intelligence

Seven weeks, one reproducible project

Week	Lecture and lab topic
Week 1	Introduction to reproducibility and open science
Week 2	Project organisation and literate programming
Week 3	Version control with Git and GitHub
Week 4	Portable and reproducible computing environments
Week 5	Containerisation with Docker and Singularity
Week 6	Workflows and automation
Week 7	Publishing, archiving and open science practice

Mohamed bin Zayed University of Artificial Intelligence

Week 1 learning goals

Monday · reproducibility

1. Explain what reproducibility is and how it differs from replicability
2. Explain the factors and reasons behind irreproducible research
3. List and define tools and techniques that support each level of reproducibility

Tuesday · open science

4. Understand the motivation for open science
5. Explain the core FAIR data principles and open licensing
6. Use checklists and resources to make your own work more open

Mohamed bin Zayed University of Artificial Intelligence

Plan for the 90 minutes

Time	Block	What happens
0–8	Course and goals	Schedule, grading, the seven weeks
8–15	A cautionary tale	The Stapel case and publish or perish
15–27	The evidence	Surveys and large replication projects
27–35	Definitions	Reproducible, replicable, robust; a spectrum
35–45	Why it matters	Duke, the Excel COVID error, selfish reasons
45–57	Causes	Research cycle, practices, think-pair-share
57–63	ML and AI	New threats, new tools
63–68	Toolkit preview	Five pillars; tool slides optional
68–90	Activity and wrap-up	Forensic reconstruction, debrief, takeaways



Diederik Stapel

An academic star

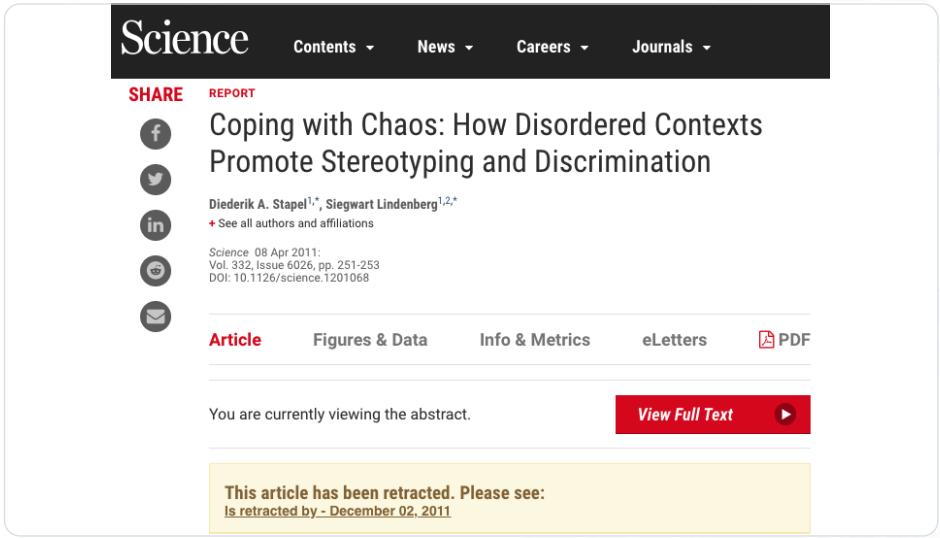
A highly successful Dutch social psychologist, internationally known for research on human behaviour, publishing in elite journals with data that often looked close to perfect.

Photo by Koos Breukel

Mohamed bin Zayed University of Artificial Intelligence

The Utrecht train station experiment





The paper argued that physical disorder in the environment, chaotic and messy surroundings, made people rely more on stereotypes, which in turn promotes discrimination.

Mohamed bin Zayed University of Artificial Intelligence

From academic star to con man



Diederik Stapel, a Dutch social psychologist, perpetrated an audacious academic fraud by making up studies that told the world what it wanted to hear about human nature. Koos Breukel for The New York Times

58

of his articles have been retracted, including the Science paper; several PhD theses were affected

The seating arrangement described at the station never existed. The data were never collected: they were fabricated.

Mohamed bin Zayed University of Artificial Intelligence

One of the reasons:

Publish or perish

What kind of academic culture lets this go on for so long?

Mohamed bin Zayed University of Artificial Intelligence

Most scientists have hit a wall trying to reproduce a result

>70%

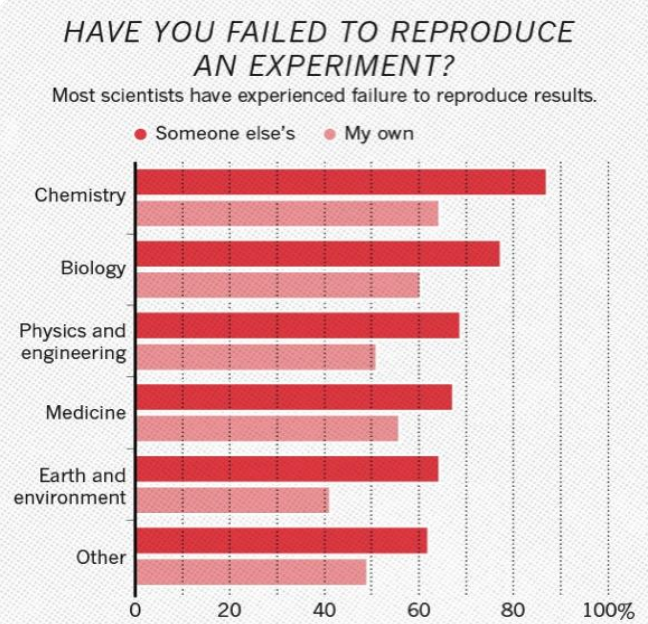
failed to reproduce another scientist's experiment

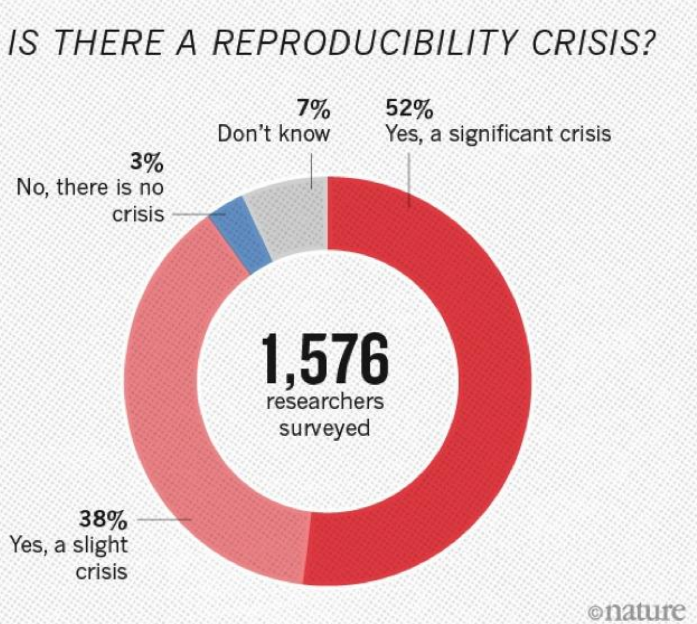
>50%

failed to reproduce one of their own

90%

said there is a 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

Mohamed bin Zayed University of Artificial Intelligence

Reproducibility Project: Psychology

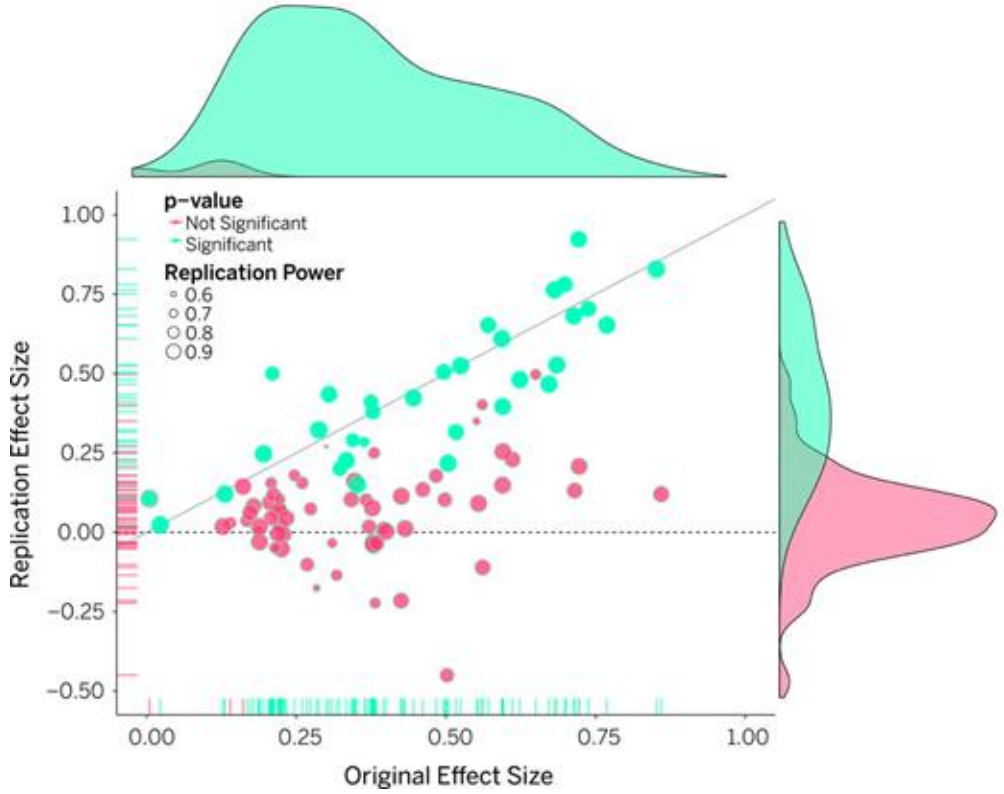
- Replication of **100 experiments** from three high-ranking psychology journals
- A large collaborative effort, 2011–2015
- Only **36%** of replications had statistically significant results

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



Mohamed bin Zayed University of Artificial Intelligence

Reproducibility Project: Cancer Biology, 2013–2021

193 → 50

experiments planned from 53 high-impact papers vs. completed, from 23 papers

85%

smaller median effect size in replications of positive results

40%

of positive effects (39/97) replicated on at least 3 of 5 criteria

None of the 193 experiments was described in enough detail to design a replication protocol from the paper alone.

Sources: Errington TM et al. eLife 10:e71601 (2021); Rodgers & Collings, eLife 10:e75830 (2021) · Lecture 1 · Reproducibility · 13

Mohamed bin Zayed University of Artificial Intelligence

The Brazilian Reproducibility Initiative, 2019–2025

- A multicentre initiative to estimate the reproducibility of Brazilian biomedical science, supported by the Serrapilheira Institute
- **56 laboratories** performed **143 replications** of 56 experiments
- Replication rates ranged from **15% to 45%**, depending on the criterion used to define success


nature

Explore content · About the journal · Publish with us · Subscribe

nature › news › article

NEWS | 25 April 2025

Huge reproducibility project fails to validate dozens of biomedical studies

Unique reproducibility effort in Brazil focuses on common methods rather than a single field — and prompts call for reform.


bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Contradictory Results

Follow th

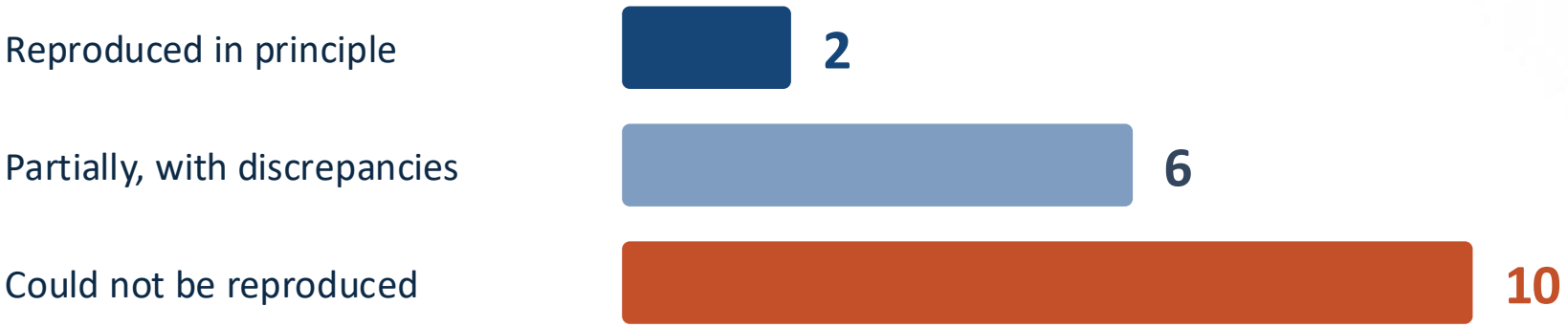
Estimating the replicability of Brazilian biomedical science

The Brazilian Reproducibility Initiative, Olavo Bohrer Amaral, Clarissa Franca Dias Carneiro,

Mohamed bin Zayed University of Artificial Intelligence

Even with the data, published microarray analyses rarely re-ran

18 Nature Genetics papers (2005–06); one table or figure each, re-analysed by two independent teams



Main cause of failure: the data were not actually available. **Main cause of discrepancies:** incomplete annotation and under-specified processing.

Mohamed bin Zayed University of Artificial Intelligence

Sharing code is necessary, but shared code mostly doesn't run

Jupyter notebooks on GitHub

24% ran without error

4% gave the same results

~864,000 executions attempted · Pimentel et al., MSR 2019

R scripts in published replication packages

74% crashed on first run

56% after automatic clean-up

9,000+ R files, Harvard Dataverse · Trisovic et al., Sci Data 2022

Most failures were mundane: missing packages, hard-coded paths, cells run out of order, absent input files.

Mohamed bin Zayed University of Artificial Intelligence

Two words that are not synonyms

Reproducibility

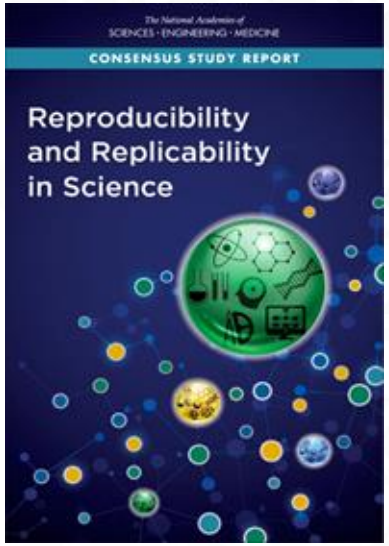
Getting consistent results when you reuse the **same data**, code, methods and conditions of analysis.

Mostly a computational question. It should be close to guaranteed.

Replicability

Getting consistent results in a new study that asks the same question and collects **its own data**.

A scientific question. Some failure is expected and informative.



http://www.nap.edu/25303

Mohamed bin Zayed University of Artificial Intelligence

Reproducibility ≠ Replicability

A rigorously conducted study, using best practices, correctly analysed and transparently reported, should be computationally reproducible, yet it may still fail to replicate.

Reproducibility is the minimum standard for good science.

Mohamed bin Zayed University of Artificial Intelligence

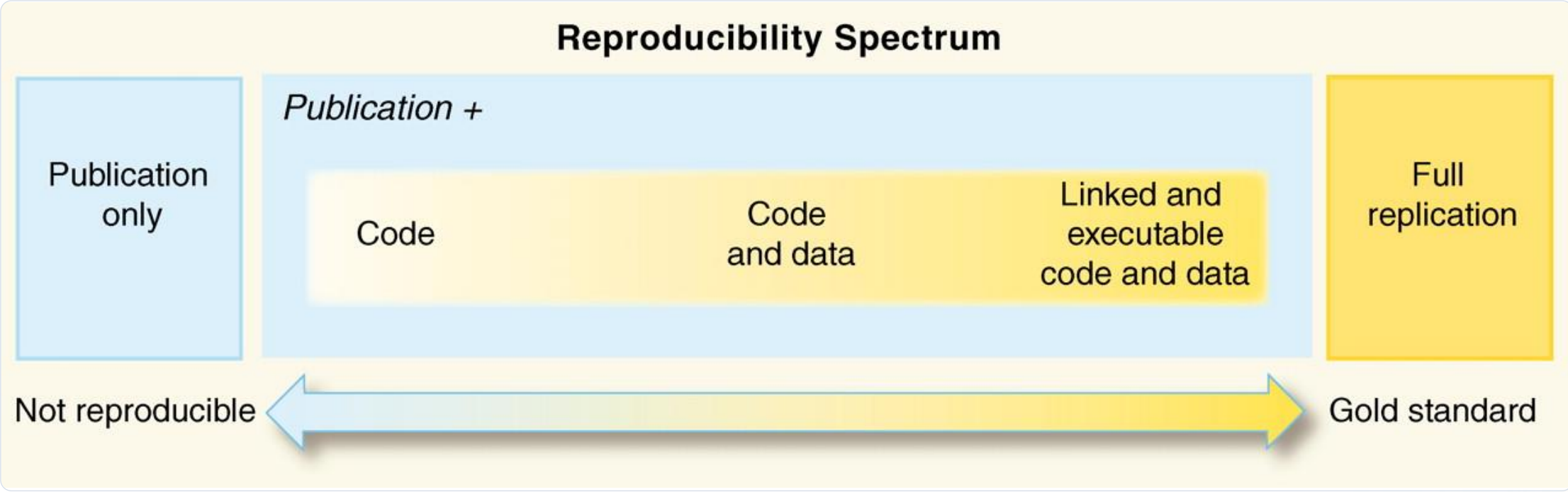
Four ways a result can hold up

Vary the data, the analysis, or both. A trustworthy finding survives all four.

	Same analysis	Different analysis
Same data	**Reproducible** rerun my code on my data	**Robust** other reasonable pipelines agree
Different data	**Replicable** same method, new cohort	**Generalisable** holds beyond the original setting

Mohamed bin Zayed University of Artificial Intelligence

Reproducibility is a spectrum



From “not reproducible” to the gold standard of full replication: each step adds code, data, or executable links between them.

Mohamed bin Zayed University of Artificial Intelligence

Fields have used these words differently, so define them when you write

Goodman, Fanelli & Ioannidis (2016)	What it asks
Methods reproducibility	Is there enough detail to repeat exactly what was done?
Results reproducibility	Does an independent repeat of the study give the same results?
Inferential reproducibility	Would others draw the same conclusions from the evidence?

Computing's ACM badges once used “reproducible” and “replicable” the other way round, before aligning with the National Academies in 2020. In this course we use the National Academies' meanings.

Mohamed bin Zayed University of Artificial Intelligence

Your most important collaborator is you, six months from now — and they don't answer email.

Reproducibility is first a gift to your own future work, and only then to everyone else.

Mohamed bin Zayed University of Artificial Intelligence

Reproducible science: why should we bother?



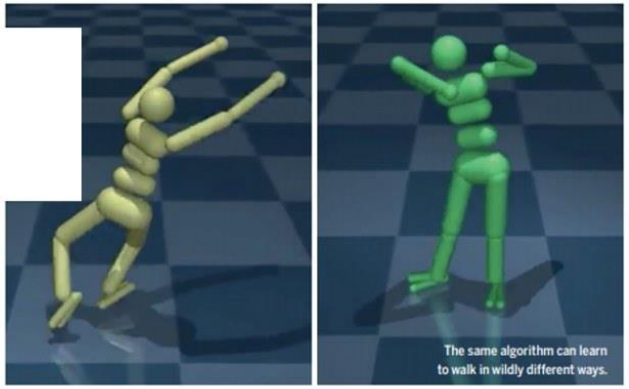
Why Most Published Research Findings Are False

John P. A. Ioannidis

nature REVIEWS DRUG DISCOVERY

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange & Khusru Asadullah



COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

BMJ

BMJ 2014;348:g3725 doi: 10.1136/bmj.g3725 (Published 13 June 2014)

Evidence based medicine: a movement in crisis?

Trisha Greenhalgh and colleagues argue that, although evidence based medicine has had many benefits, it has also had some negative unintended consequences. They offer a preliminary agenda for the movement's renaissance, refocusing on providing useable evidence that can be combined with context and professional expertise so that individual patients get optimal treatment

Power failure: why small sample size undermines the reliability of neuroscience

Katherine S. Button[1,2], John P. A. Ioannidis[3], Claire Mokrysz[1], Brian A. Nosek[4], Jonathan Flint[5], Emma S. J. Robinson[6] and Marcus R. Munafò[1]

Mohamed bin Zayed University of Artificial Intelligence

Five selfish reasons to work reproducibly

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

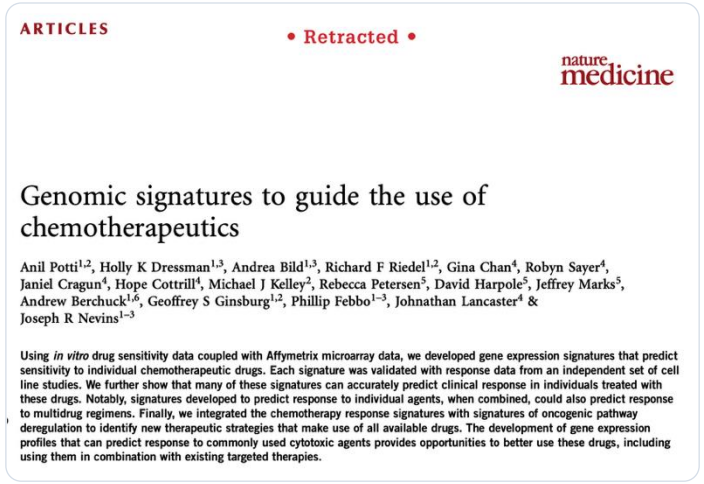
Genome Biology **16**, Article number: 274 (2015) | Cite this article

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

Reproducibility helps you first, then science and society.

Mohamed bin Zayed University of Artificial Intelligence

Duke: gene names off by one, duplicated patients



ARTICLES • Retracted •

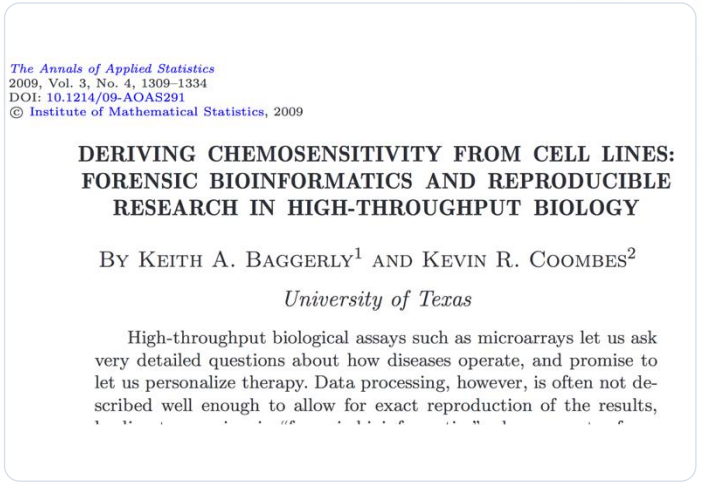
nature medicine

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.

2006 · Potti et al. build a model predicting chemotherapy sensitivity from gene-expression signatures. Four clinical trials start; a fifth is planned.



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

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

BY KEITH A. BAGGERLY[1] AND KEVIN R. COOMBES[2]

University of Texas

High-throughput biological assays such as microarrays let us ask very detailed questions about how diseases operate, and promise to let us personalize therapy. Data processing, however, is often not described well enough to allow for exact reproduction of the results,

2009 · Baggerly & Coombes (MD Anderson) find off-by-one gene lists, swapped labels and duplicated patients and cell lines.



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

2011 · Nature Medicine retraction; trials halted. The study was not transparent and data were not shared.

Mohamed bin Zayed University of Artificial Intelligence

How events unfolded

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.

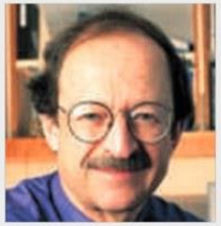
JANUARY 2010

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.

40 | NATURE | VOL 469 | 13 JANUARY 2011

Mohamed bin Zayed University of Artificial Intelligence

16,000 COVID-19 cases went unreported in England

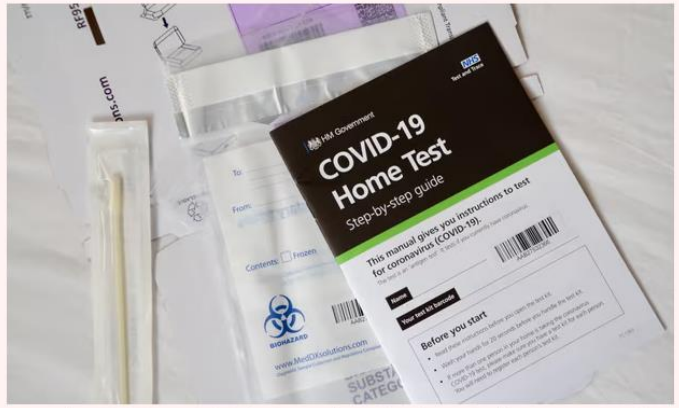
Analysis

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

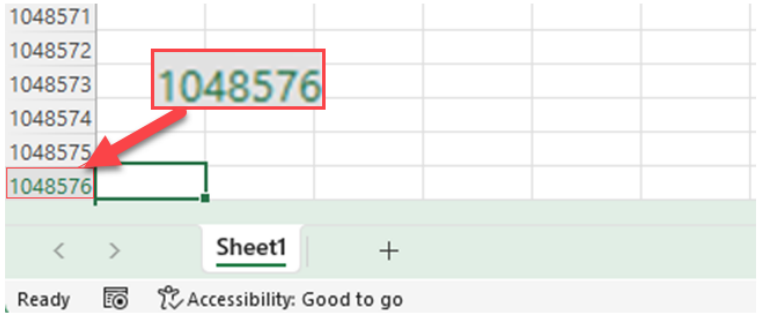


More than 50,000 potentially infectious people may have been missed by contact tracers after 15,841 positive tests were left off the daily figures. Photograph: Simon Leigh/Alamy

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.

- Test results were collated in an old Excel file format
- Older versions of Excel hold only **65,536** rows; Microsoft 365 holds **1,048,576**
- Rows past the limit were silently dropped, so contacts were not traced

All Excel versions from 2007 onwards, including Microsoft 365, can handle up to 1,048,576 rows per worksheet. But, in the old Excel versions like Excel 2003 and before, they could only handle up to 65,536 rows.



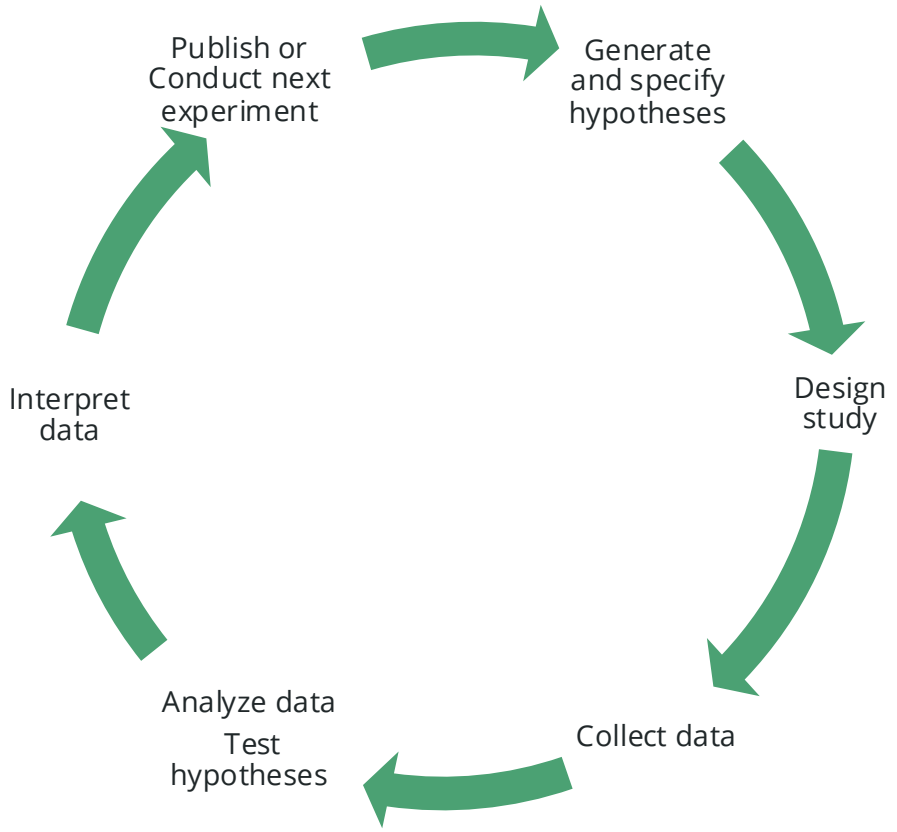
Mohamed bin Zayed University of Artificial Intelligence

Do you have a story to share with us?

A result you could not reproduce, a lost file, a version that changed everything

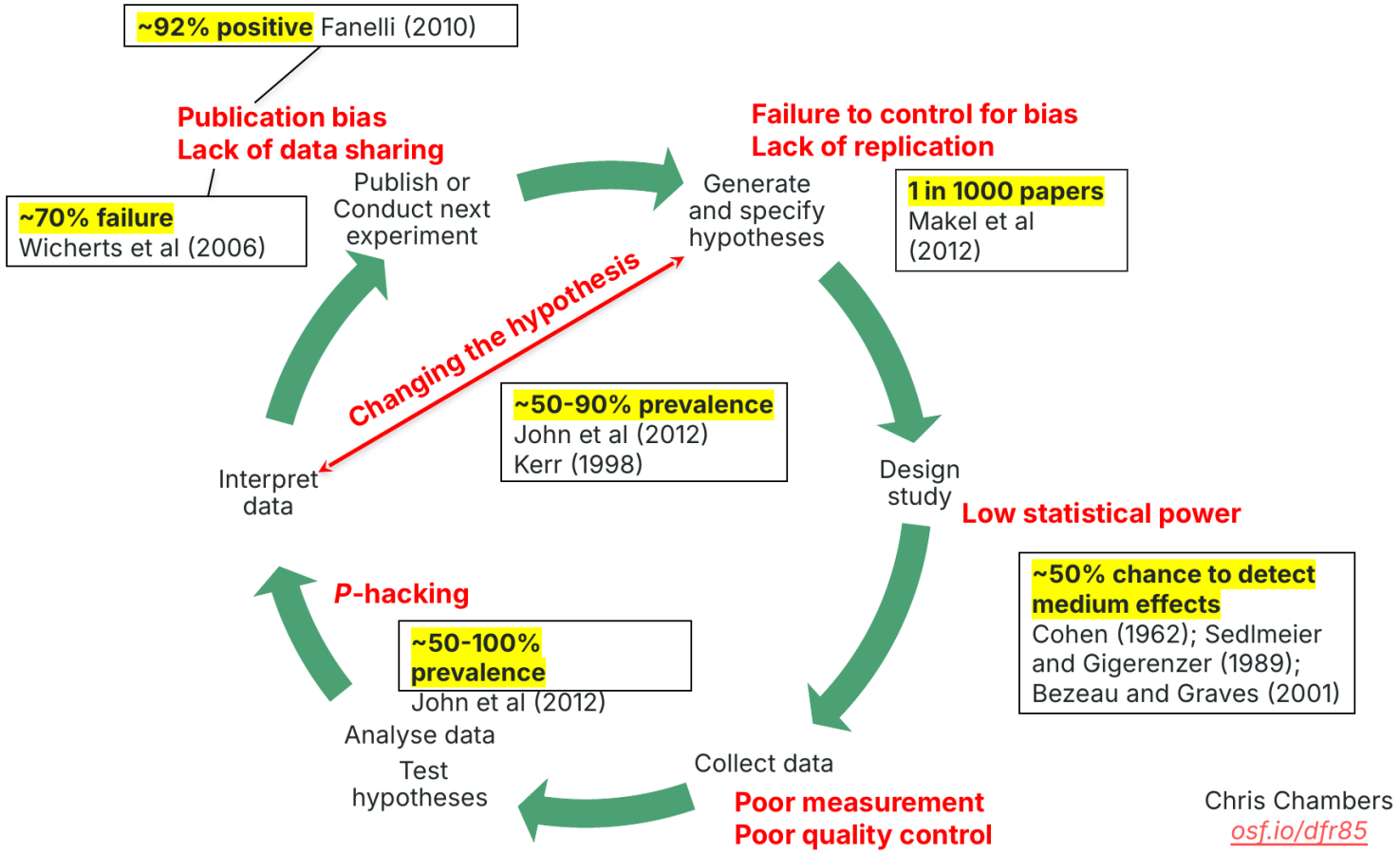
Mohamed bin Zayed University of Artificial Intelligence

The research cycle



Mohamed bin Zayed University of Artificial Intelligence

The research cycle: where it goes wrong



Mohamed bin Zayed University of Artificial Intelligence

Questionable practices, and what counters them

Practice	What it looks like	Counter-measure
p-hacking	Trying analyses until something crosses $p < 0.05$	Pre-specified analysis plan
Forking paths	Many defensible choices, one reported outcome	Multiverse or sensitivity analysis
HARKing	Presenting a post-hoc finding as the original hypothesis	Preregistration; label exploratory work
Low power	Small n, so positives are often false or inflated	Power analysis; collaboration
Publication bias	Null results stay in the drawer	Registered Reports; preprints

Mohamed bin Zayed University of Artificial Intelligence

Why the same data and code give different answers

Missing pieces

Code never shared, raw data not deposited, deposited, a step done by hand in a spreadsheet

Software drift

Unpinned package versions; a default parameter that changed between releases

Moving references

Genome build (hg19 vs hg38), annotation release, database snapshot not recorded

Hidden state

Notebook cells run out of order; variables left over in memory

Machine-specific

Hard-coded paths, OS-dependent sorting, GPU or thread non-determinism

Unseeded randomness

Sampling, clustering, train/test splits, UMAP layouts without a fixed seed

Mohamed bin Zayed University of Artificial Intelligence

Who is to blame? Responsibility is shared

Researchers

Uphold rigour: document methods, data and code and make them open. Train students in rigour and reproducibility.

Institutions

Set policies for transparent research and oversee them. Recognise reproducible work in hiring and promotion.

Funding agencies

Mandate open code and data and set clear expectations. Fund and reward reproducible, open researchers.

Journals

Ensure code and data are available, in line with funders' mandates. Publish replications and negative results.

Scientific societies and the public

Demand transparency and accountability. Advocate for policies that promote open, reproducible science.

Responsibility begins with us, the researchers.

Mohamed bin Zayed University of Artificial Intelligence

Think, pair, share

Think · 2 min

On your own, write down one or two major factors that hurt reproducibility.

Pair · 3 min

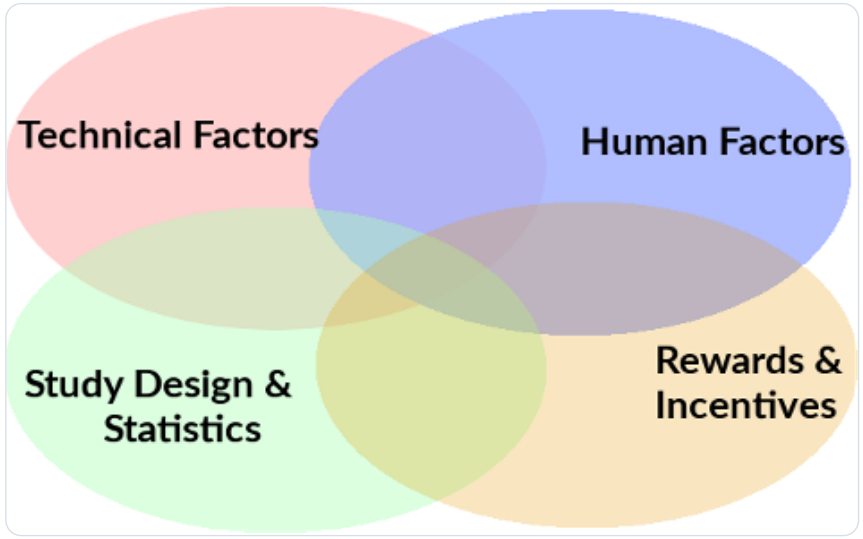
With a neighbour, discuss how each problem could be fixed, and by whom.

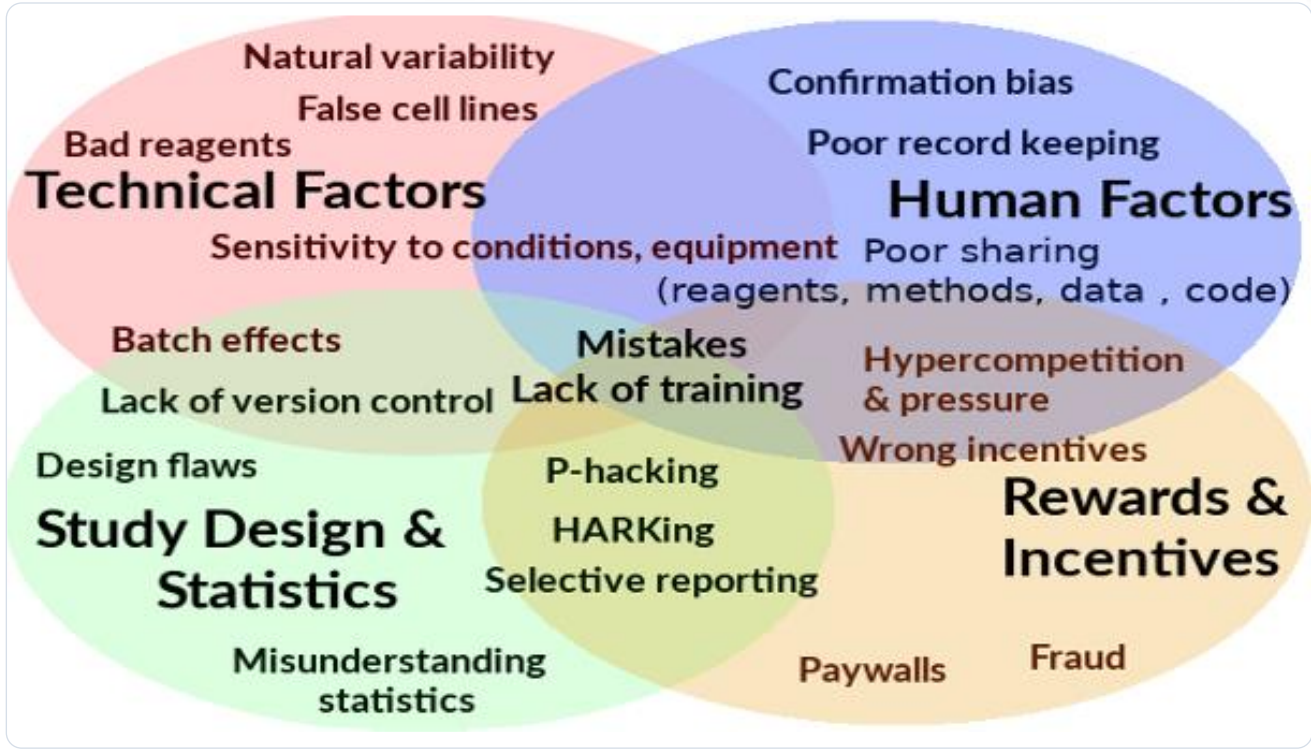
Share · 3 min

Pairs report back; we sort the answers into the four groups on the next slide.

Mohamed bin Zayed University of Artificial Intelligence

Factors that affect reproducibility





Mohamed bin Zayed University of Artificial Intelligence

Machine learning and AI have a reproducibility crisis too

Key challenges

- Code, data and trained models not shared
- Undocumented hyperparameters and preprocessing steps
- Dependence on hardware and compute: GPU type, random seeds
- Benchmark cherry-picking


Feature

ILLUSTRATION BY LORIA VIOLTA

IS AI LEADING TO A REPRODUCIBILITY CRISIS IN SCIENCE?

Scientists worry that ill-informed use of artificial intelligence is driving a deluge of unreliable or useless research. By Philip Ball

Mohamed bin Zayed University of Artificial Intelligence

How AI threatens reproducibility

Speed

A rapid pace of publication limits the depth of peer review, and AI is now used in peer review itself.

Closed models

Proprietary data and models prevent independent verification.

Benchmark chasing

Overfitting to benchmarks without real generalisation.

Compute barriers

Large models need resources few groups have, so replication is out of reach.

Generated content

AI tools may produce unverifiable or hallucinated text, code and citations.

Mohamed bin Zayed University of Artificial Intelligence

Hidden prompts aimed at AI peer reviewers



The Guardian

Int

News | Opinion | Sport | Culture | Lifestyle

World US politics UK Climate crisis Middle East Ukraine Environment Science Global development Football Tech Business Obituaries

Artificial intelligence (AI)

This article is more than **2 months old**

Scientists reportedly hiding AI text prompts in academic papers to receive positive peer reviews

Research papers found carrying hidden white text giving instructions not to highlight negatives as concern grows over use of large language models for peer review



arXiv > cs > arXiv:2507.06185

Search...

Help | Adv

Computer Science > Computers and Society

[Submitted on 8 Jul 2025]

Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review

Zhicheng Lin

Mohamed bin Zayed University of Artificial Intelligence

AI as a tool for reproducibility

Experiment tracking

Automatic logging of parameters, runs and results with MLflow or Weights & Biases.

Environments and workflows

AI assistance in writing Dockerfiles, Snakemake and Nextflow pipelines.

Code quality

AI-assisted linting, documentation and unit tests.

Error detection

AI-driven spotting of anomalies and inconsistencies in data and results.

Mohamed bin Zayed University of Artificial Intelligence

When should you 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

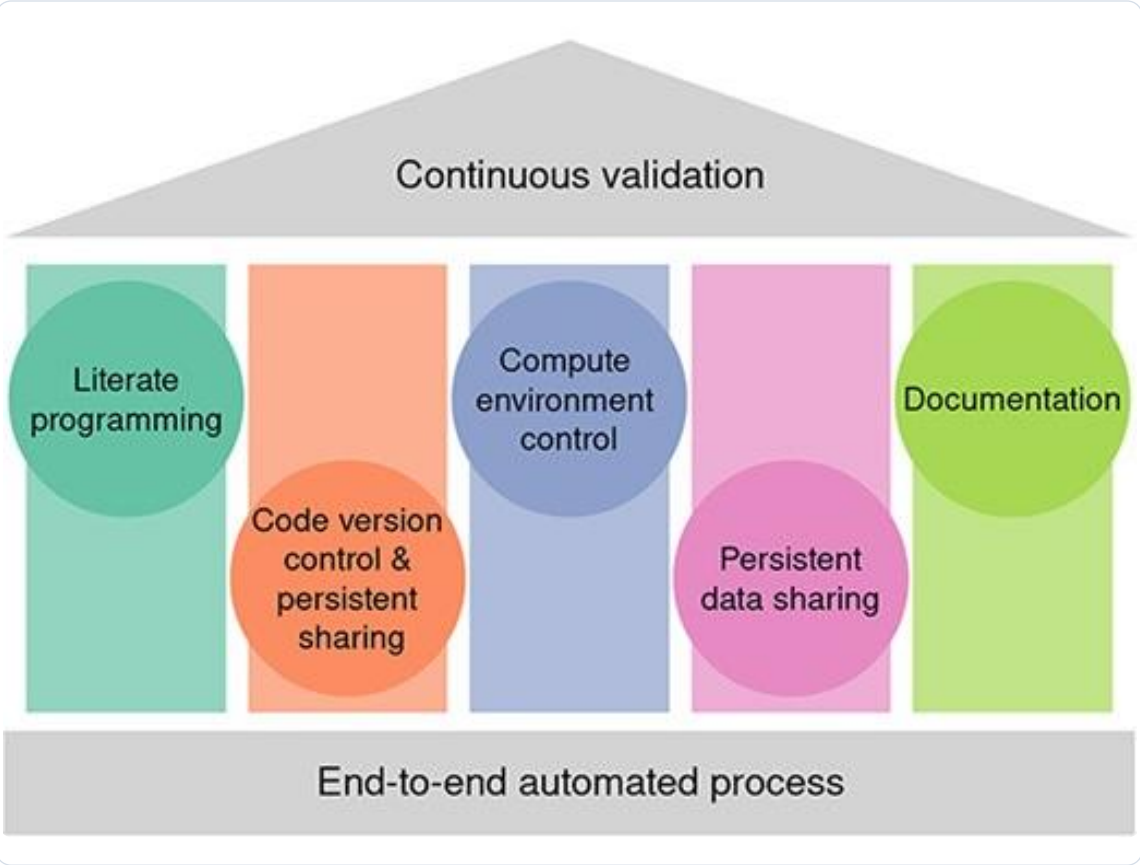
When you co-author a paper

When you review a paper

All the time

Mohamed bin Zayed University of Artificial Intelligence

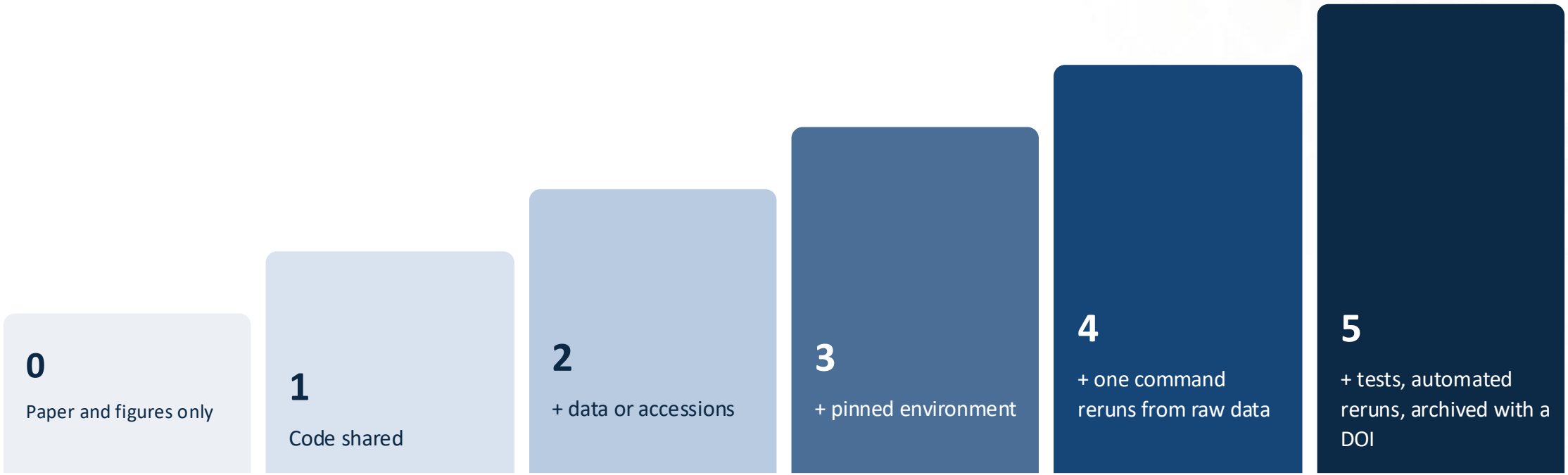
Five pillars of computational reproducibility



Pillar	Taught in
Literate programming	Week 2
Code version control and sharing	Week 3
Compute environment control	Weeks 4–5
Persistent data sharing	Week 7
Documentation	Every week
End-to-end automation	Week 6

Mohamed bin Zayed University of Artificial Intelligence

The reproducibility ladder: climb one rung at a time



Mohamed bin Zayed University of Artificial Intelligence

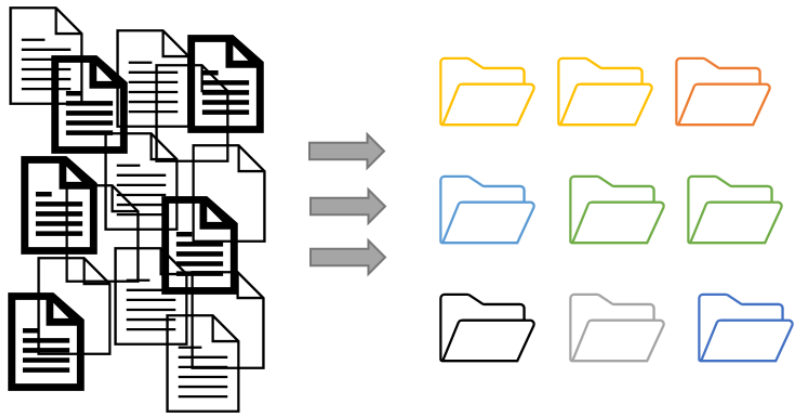
Reproducibility starts on day one

It starts when you create the first folder for the project:

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

Not when your PI asks you to put the code in a repository because a journal or funder requires it.

✗ Bad	✓ Good
`Myabstract.docx`	`2020-06-08_abstract-for-sla.docx`
`Joe's Filenames Use Spaces and Punctuation.xlsx`	`Joes-filenames-are-getting-better.xlsx`
`figure 1.png`	`Fig01_scatterplot-talk-length-vs-interest.png`
`fig 2.png`	`Fig02_histogram-talk-attendance.png`
`JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt`	`1986-01-28_raw-data-from-challenger-o-rings.txt`



messy/unstructured

organised folder structure

Mohamed bin Zayed University of Artificial Intelligence

A predictable project layout

```
my-project/
├── README.md
├── data/
│   ├── raw/          # read-only
│   └── processed/    # made by code
├── code/
├── envs/             # lockfiles
├── results/          # regenerable
├── docs/
└── Snakefile         # one entry point
```

Raw data is sacred

Never edit it. Everything else is derived and can be deleted.

Relative paths only

No /home/aziz/… in code. The project runs wherever it is copied. copied.

README first

What this is, how to install, how to run, where outputs go.

Mohamed bin Zayed University of Artificial Intelligence

Version control: a lab notebook for code

Commit small, commit often

One logical change per commit, with a message that says why.

Keep big data out of git

Ignore FASTQ and BAM files; record accessions and checksums instead.

Tag what you publish

Tag the exact commit behind each submission; archive it with a DOI.

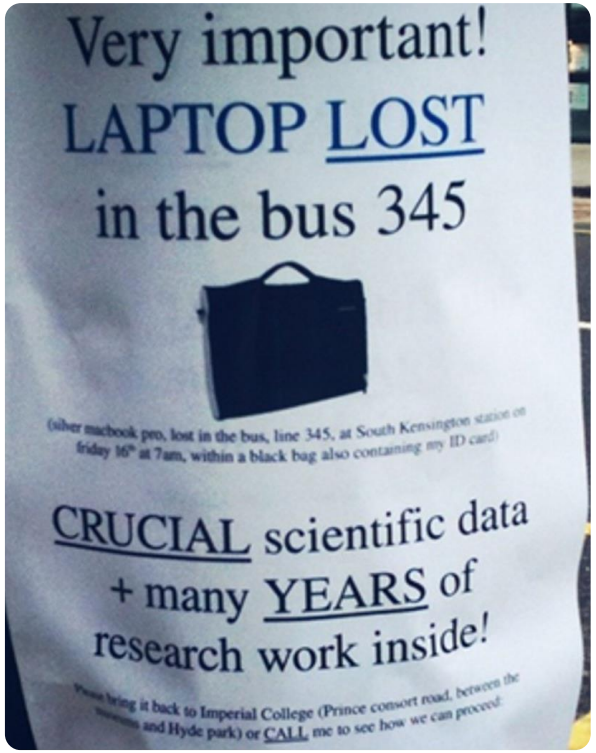
```
git init
git add code/align.smk
git commit -m "Use hg38; drop
  ENCODE blacklist regions"

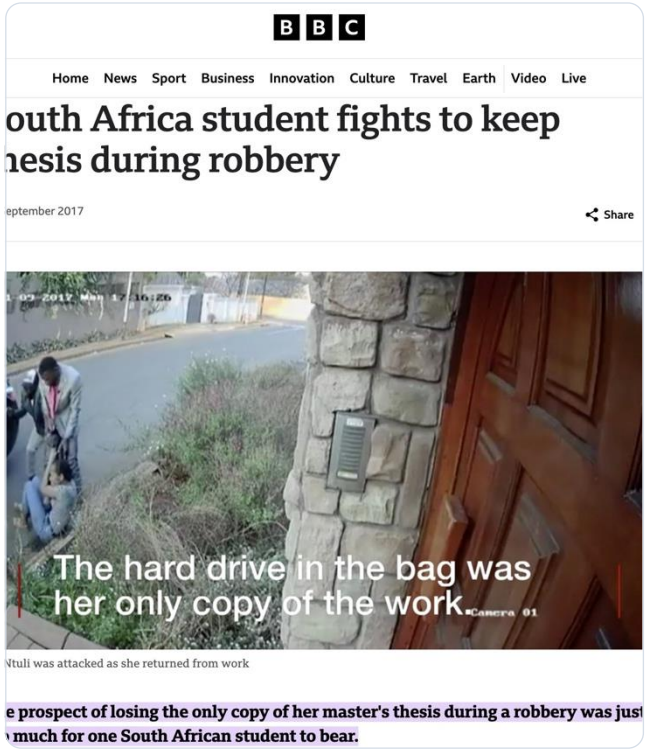
git tag -a v1.0-submission
git push --tags
```

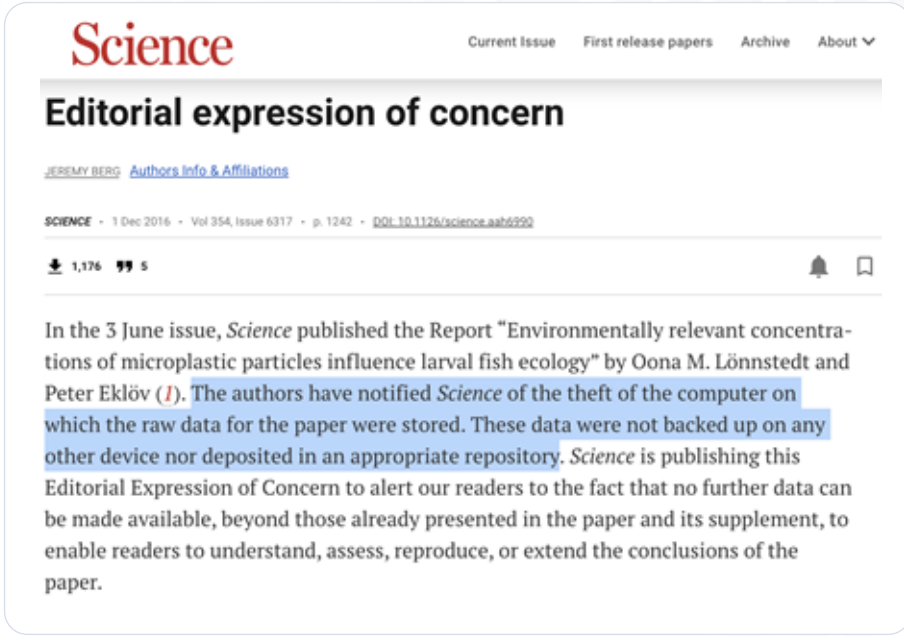
GitHub + Zenodo: each release gets a citable DOI

Mohamed bin Zayed University of Artificial Intelligence

Version control and archiving prevent disasters



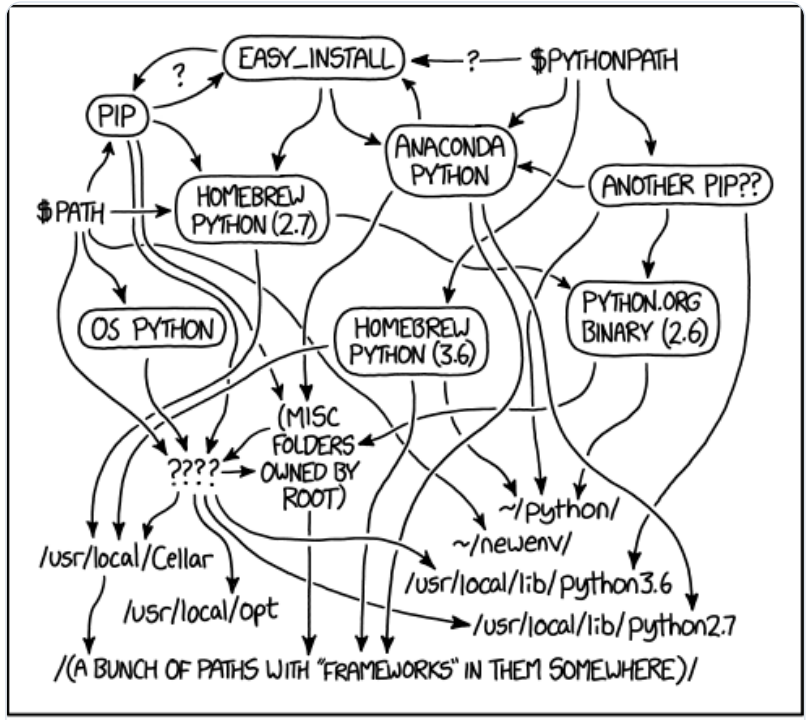




*"The authors have notified Science of the **theft of the computer** on which **the raw data for the paper were stored**. These data were **not backed up** on any other device nor deposited in an appropriate repository."*

Mohamed bin Zayed University of Artificial Intelligence

Track software versions and avoid dependency hell



MY PYTHON ENVIRONMENT HAS BECOME SO DEGRADED THAT MY LAPTOP HAS BEEN DECLARED A SUPERFUND SITE.

Our code depends on other code, which depends on other code… We control our code, but how do we control its dependencies?

On an HPC cluster, pin module versions:

```
# not reproducible
$ module load cuda
# reproducible
$ module load cuda/11.4.2
# see what exists
$ module avail cuda
```

Conda, Bioconda, Mamba, pip, pixi and virtualenv install, record and isolate exact versions per project.

Mohamed bin Zayed University of Artificial Intelligence

Freeze the software, not just the code

Package environments

conda / mamba, Python venv, R renv

Pins: language packages and tool versions

`environment.yml` · `renv.lock`

Containers

Docker on laptops and cloud; Apptainer on HPC clusters

Pins: the whole operating-system user space

`Dockerfile` · `image@sha256`

Functional builds

Guix, Nix

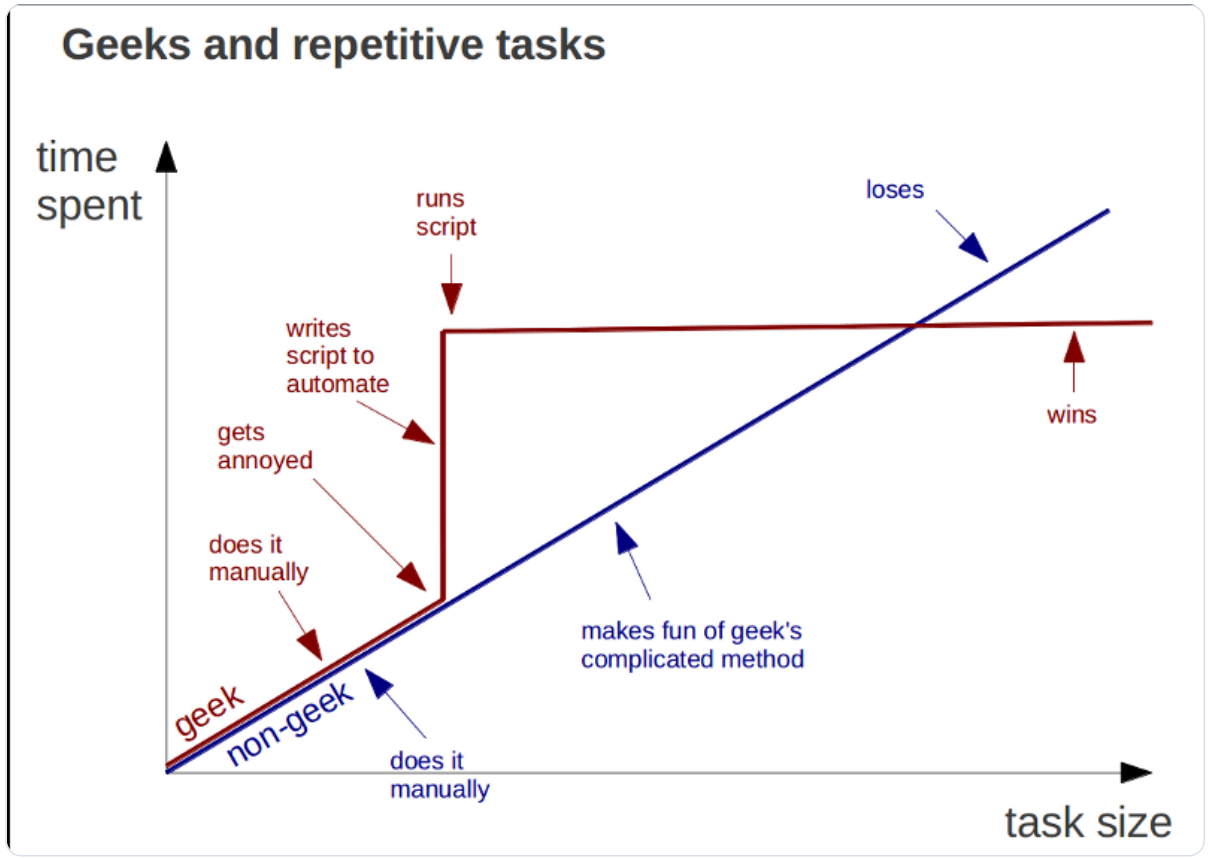
Pins: every dependency down to the compiler

`manifest.scm` · `flake.lock`

Minimum standard: exact versions in a lockfile, and the session info printed at the end of every analysis.

Mohamed bin Zayed University of Artificial Intelligence

The best documentation is automation

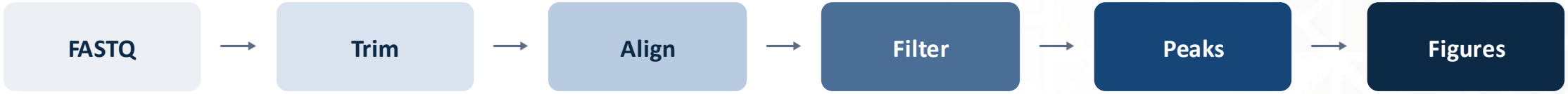


- A repetitive simple task: put it in a shell script
- Good old GNU make
- Modern workflow managers: Snakemake, Nextflow, WDL

A pipeline that runs end to end is a methods section that cannot be out of date.

Mohamed bin Zayed University of Artificial Intelligence

Workflow managers turn steps into a rerunnable graph



Declares dependencies

Each rule names its inputs, outputs and environment: the methods section, executable.

Reruns only what changed

Change a filter threshold and only the downstream steps run again.

Scales without edits

The same file runs on a laptop, a SLURM cluster or the cloud.

Snakemake (Python-like) · Nextflow and the nf-core community pipelines

Mohamed bin Zayed University of Artificial Intelligence

Literate documents: text, code and results together

Quarto, R Markdown and Jupyter put the narrative next to the code that produced each number and figure.

The risk is hidden state: a notebook that only works in the order you happened to run it.

House rules

1. Restart and run all before you share or commit
2. Render reports from source; never edit the output
3. Heavy compute in the workflow; notebooks for figures
4. Imports and parameters at the top
5. End with session info

Mohamed bin Zayed University of Artificial Intelligence

Record where every input came from, and which version

Input	Record at minimum
Sequencing data	Repository accession (GEO, SRA, ENA, EGA) and sample IDs
Reference genome	Build and source, e.g. GRCh38 from a named provider; patch level
Annotation	Release number, e.g. a specific GENCODE or Ensembl release
Databases	Version or download date (JASPAR, dbSNP, gnomAD, MSigDB…)
Downloaded files	Checksum (md5 or sha256) stored next to the file

Best of all: make the download itself a workflow step, so provenance is code rather than a note.

Mohamed bin Zayed University of Artificial Intelligence

Same code, same data, different answer?

Set and report seeds

Subsampling, clustering, UMAP or t-SNE, permutation tests, train/test splits and model training.

Expect small numerical drift

Multi-threading and GPUs can change the order of floating-point sums, so last digits can differ.

Define “the same” in advance

Identical files, identical within a tolerance, or the same conclusion?

```
# R
set.seed(20261025)

# Python
rng = np.random.default_rng(
    20261025)
torch.manual_seed(20261025)
```

Then check stability across several seeds: one lucky seed is not a result.

Mohamed bin Zayed University of Artificial Intelligence

Ten simple rules for reproducible computational research

1. Track how every result was produced
2. Avoid manual data manipulation
3. Archive exact versions of all programs
4. Version-control your scripts
5. Keep intermediate results, in standard formats
6. Record the random seeds
7. Store the raw data behind every plot
8. Produce output in inspectable layers
9. Link statements in the text to results
10. Share scripts, runs and results openly

Mohamed bin Zayed University of Artificial Intelligence

Forensic reconstruction

A TYPICAL METHODS PARAGRAPH

"Reads were trimmed and aligned to the human genome using Bowtie2 with default parameters. Duplicates were removed and peaks were called with MACS2 ($p < 0.01$). Differential binding between conditions was assessed with DESeq2, and regions with FDR < 0.05 were considered significant. Motif enrichment was performed using HOMER."

1 · Alone, 8 min

List every decision you would need to rerun this. Circle each one the paragraph leaves unstated.

2 · In pairs, 8 min

Draw the analysis as boxes and arrows. On each arrow, write what is missing to make that step rerunnable.

3 · In pairs, 4 min

Place the paper on the ladder (0–5). Write the one fix that would help a reader most.

Mohamed bin Zayed University of Artificial Intelligence

Four sentences, twenty-plus hidden decisions

Data

Accessions; replicates per condition; input or control samples; read length; paired or single end

References

Genome build; which contigs kept; blacklist regions; annotation release; HOMER genome and motif library version

Parameters

Trimming tool and adapters; MAPQ filter; duplicate tool; MACS2 mode and genome size; consensus peak rule; count method; DESeq2 design; HOMER background and region size

Software

A version for each of four named tools and the unnamed ones (trimmer, dedup, counting); what “default” meant in that version

A shared workflow file would have answered every one of these questions without a single extra sentence of prose.

Mohamed bin Zayed University of Artificial Intelligence

Five things to remember

1. Reproducibility (same data) should be near-certain; replicability (new data) is the scientific test.
2. Across fields, many published results do not hold up, and effects shrink when repeated.
3. Most computational failures are mundane: missing files, versions, paths and seeds.
4. Prose cannot carry a pipeline; code, environments and workflows can.
5. Climb the ladder one rung at a time, starting with your next analysis.

Mohamed bin Zayed University of Artificial Intelligence

Reading list

Core

National Academies. *Reproducibility and Replicability in Science* (2019)

Sandve GK et al. Ten simple rules for reproducible computational research. *PLoS Comput Biol* 9:e1003285 (2013)

Wilson G et al. Good enough practices in scientific computing. *PLoS Comput Biol* 13:e1005510 (2017)

Markowetz F. Five selfish reasons to work reproducibly. *Genome Biol* 16:274 (2015)

Ziemann M et al. The five pillars of computational reproducibility. *Brief Bioinform* 24:bbad375 (2023)

Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 533:452–454 (2016)

Further

Noble WS. A quick guide to organizing computational biology projects. *PLoS Comput Biol* 5:e1000424 (2009)

Wilson G et al. Best practices for scientific computing. *PLoS Biol* 12:e1001745 (2014)

Alnasir JJ. Fifteen quick tips for success with HPC. *PLoS Comput Biol* 17:e1009207 (2021)

Peng RD. Reproducible research in computational science. *Science* 334:1226 (2011)

Open Science Collaboration. *Science* 349:aac4716 (2015); Errington TM et al. *eLife* 10:e71601 (2021)

Baggerly KA & Coombes KR. *Ann Appl Stat* 3:1309 (2009); Ioannidis JPA et al. *Nat Genet* 41:149 (2009)

Mohamed bin Zayed University of Artificial Intelligence

Open science: making it possible for others to check

Homework!

Pick one recent paper in your area. Try to find its raw data, its code and its software versions. Bring a one-line verdict on paper: where does it sit on the ladder, and what stopped you going higher?