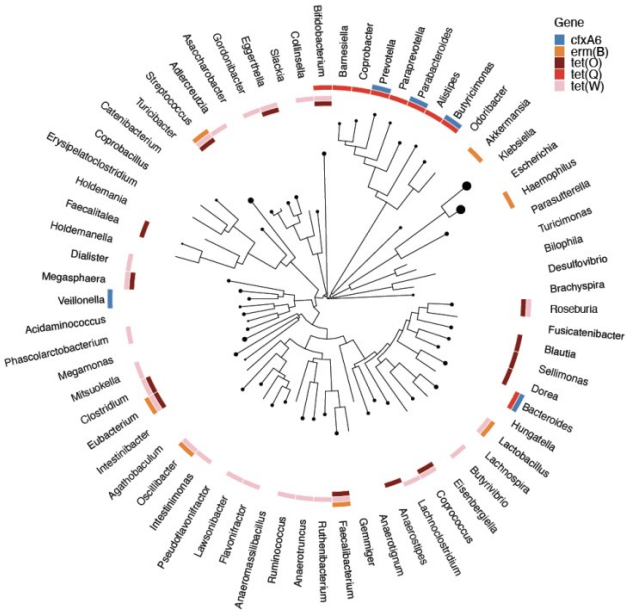
Introduction to data & code management



https://docs.google.com/document/d/1eBM4gevAKcOFw5gI6hbzgG_M_Ax4szFb0Nzg6yle338/edit?tab=t.0

HOLOGEN

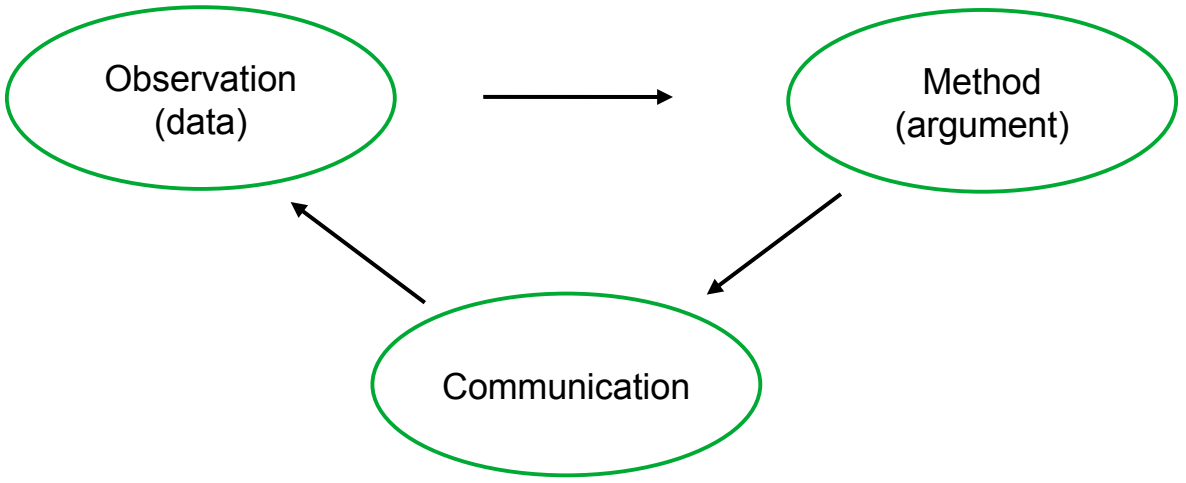
Learning goals

- 1) Elements of responsible research
- 2) Common tools for data & code management

Task: DMP

- Download EU project data management plan template from dmptuuli.fi
- Fill it tentatively for your project
- Discussion to follow

Research process & life cycle



“as open as possible,
as closed as necessary”

“What marks out modern science is not the conduct of experiments (alchemists conducted plenty of experiments), but the formation of a *critical community capable of assessing discoveries and replicating results*.” The Invention of Science: A New History of the Scientific Revolution, by David Wootton

Alchemy & algorithms: perspectives on the philosophy and history of open science

Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenoja, Mikko Tolonen

A family of alchemists at work, an engraving by Philip Galle, after a painting by Pieter Bruegel the Elder, published by Hieronymus Cock, *c*.1558.

Data Ethics

Mission and objectives

The mission of the Working Group is to work with global scholars to collaboratively establish a basic consensus for further activities and research on data ethics principles and a data ethics framework which will help CODATA advance its mission in championing global open data exchange and applications in alignment with the UNESCO Recommendations on Open Science. The specific objective is to explore the landscape and make recommendations on how CODATA activities can contribute to tackling issues of data ethics.

Some ethical considerations

- rights
- duties
- regulation
- freedom of research

Contents

Elements of responsible research (13.00 – 14.30)

- Sensitive information
- Open science, FAIR principles & reproducible research

Common tools (15.00 – 16.30)

- Version control (Git)
- Data & code sharing
- Reproducible notebooks
- Data Management Plan (DMP)

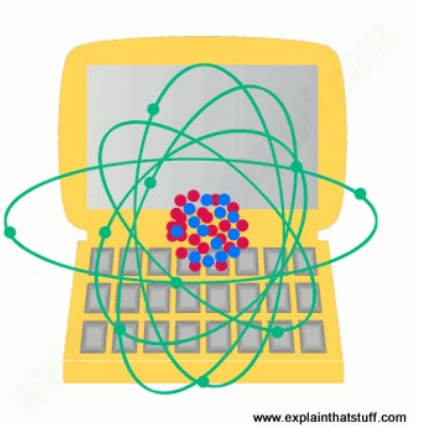
Sensitive information

Types of sensitive information?

- Human personal data
- Endangered species data
- Dual use data or methods
- IPR (academic, commercial)
- … other?

Endangered species

https://fi.wikipedia.org/wiki/Rantaruttojuuri



https://www.hsph.harvard.edu/regulatory-affairs-and-research-compliance/getting-started-2/

Definition of Sensitive Data

Sensitive data is defined as any information that is protected against unwarranted disclosure. Protection of data may be required for legal or ethical reasons, for issues pertaining to personal privacy, or for proprietary considerations. Sensitive data includes:

- Human data (e.g. health, genetic and personal information, data that may identify a person)
- Ecological data (e.g. location of endangered species or other conservation efforts)
- Confidential data (e.g. trade secrets)
- Data that is otherwise deemed sensitive

Why consider sensitivity of information?

Protecting vulnerable research subjects
Maintaining trust in research
- …?

Data protection regulations contain special processing rules and exceptions for scientific research. They are designed to support and promote research.

Varieties of sensitivity & regulation

Legal aspects
- IPR
- Rights & duties
- GDPR
- …?

Ethical aspects:
- Consent
- Good research practice
- …?

Technical aspects:
- Anonymization
- IT infrastructure
- …?

Data protection roadmap for scientific research

Taking data protection into consideration at the different phases of research and the lifespan of data.

1. **Define** the research scheme and purpose for processing personal data.
2. **Minimise** the processing of personal data.
3. **Plan** the lifespan of personal data processing, ensure the implementation of data protection principles and protect the data.
4. **Choose the basis** for processing personal data and ensure the lawfulness of processing.
5. **Implement the rights** of the data subject.
6. **Identify the roles** and responsibilities for the processing of personal data.
7. Check the basis for possible **transfers of data** outside the EU or EEA.
8. Demonstrate compliance with **data protection legislation**.
9. **Destroy, anonymize or archive** materials appropriately upon the conclusion of the study.
10. Make sure that you are familiar with data protection methods and requirements.

https://tietosuoja.fi/en/scientific-research-and-data-protection

Anonymized data

- **Anonymised data is not considered to constitute personal data**, and it is not subject to data protection regulations. (source: tietosuoja.fi)
- *Anonymous* vs. *anonymised*?
- However, ethical aspects related to **consent** remain relevant

Case example:

Journal is requiring to share data for my publication. What I should consider?

Open science, FAIR

Open research data?

Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

** A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. **Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome**. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

~2001

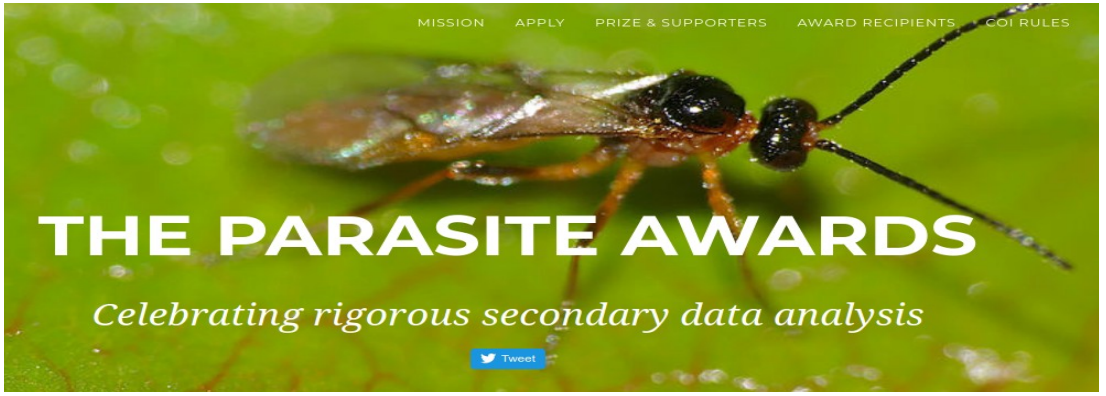
The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

A second concern - - is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."



Criticism of 'research parasites' moves NEJM in the wrong direction

By IVAN ORANSKY and ADAM MARCUS / JANUARY 26, 2016

Why open?

Element of *research quality and impact* across the *full life cycle* of research, and opportunity to advance global inclusivity in research

Why open?

Element of *research quality and impact* across the *full life cycle* of research, and opportunity to advance global inclusivity in research

Quality

- transparency
- verifiability
- trustworthiness
- integrity
- credibility
- reproducibility
- translatability into new uses

Why open?

Element of *research quality and impact* across the *full life cycle* of research, and opportunity to advance global inclusivity in research

Impact

- accessibility & broad dissemination
- reuse
- speed
- efficiency
- emergence of new innovations
- establishing priority
- receive timely feedback
- community

Why open?

Element of *research* ***quality*** *and* ***impact*** across the *full life cycle* of research, and opportunity to advance global inclusivity in research

→ *freedom* of research & *responsibility* of the researcher?

Societal impact of sharing research resources

In the first six months of the pandemic, researchers from around the world collated and curated valuable **open data** sources [1,2], as well as living reviews of **key epidemiological parameters** [3].

Real-time development of **statistical and modelling pipelines** has enabled ongoing situational awareness, such as tracking of the reproduction number [4,5].

Rapid reports have provided crucial **early analysis of virus evolution, transmission and severity** [6–8].

Interactive online apps have turned static results into flexible tools [9,10].

Open source models and data processing packages have enabled wider applications of new methodology [11,12].

These initiatives have been instrumental in supporting worldwide responses to the pandemic.

Background

Extensive national and international work, e.g.

- UNESCO recommendation on open science (2021)
- CERN open science policy
- EU's Open Science Policy (2019)
- Minimum conditions supporting research reproducibility
- Open Science and its Role in Universities: A Roadmap for Cultural Change (2018); the Association of European Research Universities
- Turning FAIR into Reality: Final report and action plan from the European Commission expert group on FAIR data (2018),
- Six Recommendations for Implementation of FAIR Practice (2020)
- EU regulation on the openness of data, e.g. the Open Data Directive.

Finnish operators actively participate in European and international work, e.g.

- European Open Science Cloud (EOSC)
- Research Data Alliance (RDA)
- International Science Council Committee on Data (CODATA)
- European research infrastructure work
- Finnish research community participates actively in dialogue where it can support the openness of research and contribute to the establishment of practical solutions in national and international cooperation.

Academy policies on open science

Open access to scientific publications

Data management and openness

Openness of research methods

Research output metadata

Data management and openness

The Academy requires that principal investigators of Academy-funded projects funded be responsible for the responsible management and opening of research data. **Research data must be made freely available as soon as possible after the research results have been published.** Sites of research must therefore provide researchers with the necessary guidance and ensure that they have access to suitable storage infrastructure. Research data shall be managed and made available following the FAIR principles (F = findable, A = accessible, I = interoperable and R = reusable).

Funded projects are requested to submit a full data management plan after the funding decision has been made. The plan must be submitted before the applicant and the supporting site of research can confirm receipt of funding.

The data shall be made open access via a national or international archive or storage service that is important for the research organisation or discipline in question. The degrees of data openness may justifiably vary, ranging from fully open to strictly confidential. The research project concerned and the publisher of the data must ensure that publishing the data will not be in breach of the Finnish Act on the Openness of Government Activities, the Finnish Data Protection Act or the Finnish Copyright Act. When making data openly available, the parties involves must also consider licensing issues.

If the research data cannot be made openly available, the metadata must be stored in a Finnish or international data finder.

The costs associated with storing and sharing research data and material are regarded as overheads for the project's host organisation, but they may also be legitimately accepted as research costs to be covered with Academy research funding.

Research Integrity guidelines and responsible science

- design, carry out and document research in a careful manner and, whenever possible, following the principles of open science
- promote the openness and further use of the data to the extent possible

Open where possible, closed if necessary

Written by milonjb. Posted in Social Sciences

National policy work for open science and research

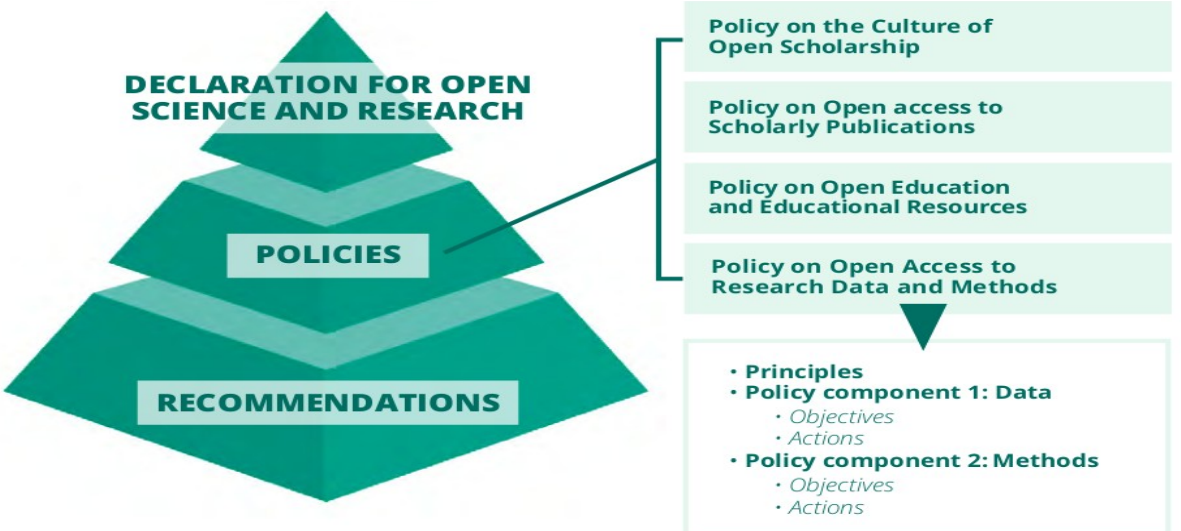
Strategic principles in national policy work

- Responsible openness
- Access to sustainable infra & services
- Researcher merit & careers

Implementation

→ Entire Finnish research community
→ Support: policies, guidelines, services, resources
→ Coordination monitors implementation & development

Guiding principle:
"as open as possible as closed as necessary"



avointiede.fi

Why not open?

Bottlenecks:

- cultural
- technical
- financial
- continuous, iterative processual nature of methods development and application
- …?

Incentives?

- Research evaluation
- Career paths
- Support services (financial, technical, other)
- Improved citation practices

Organisations should encourage adoption and promotion of open practices in all fields of research; technical support and resources, recognition and incentives.

Open data / method / infrastructure?

Research process, maturity & life cycle

- Full workflow vs. individual methods
- Early & broad openness
- Limitations explicitly justified

Degree & type of openness

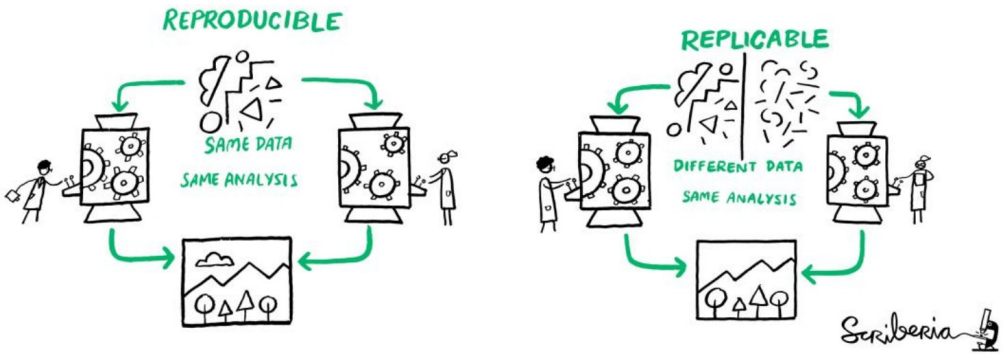
- Draft sketch → reproducible workflow → FAIR
- Transparent → Reproducible → Reusable → Interoperable
- FAIR principles

Culture and norms

- Research method as a *process*
- Standards in different fields
- IPR & research ethics

Open infrastructures

- Non-profit oriented
- Minimal costs for research
- Access can be regulated
- Open management & access policies



Version control & demo

Version control systems & tools:

- Git
- Gitlab & Github
- Continuous Integration

Discussion:

- Why version control?
- Why not version control?
- Other experiences

Recommendations for the Hologen network:

- Learn the basics of Git(hub)
- Learn to build a simple R/Python/other library
- … other topics to discuss / agree …?

Data & code sharing platforms

Data?

Code?

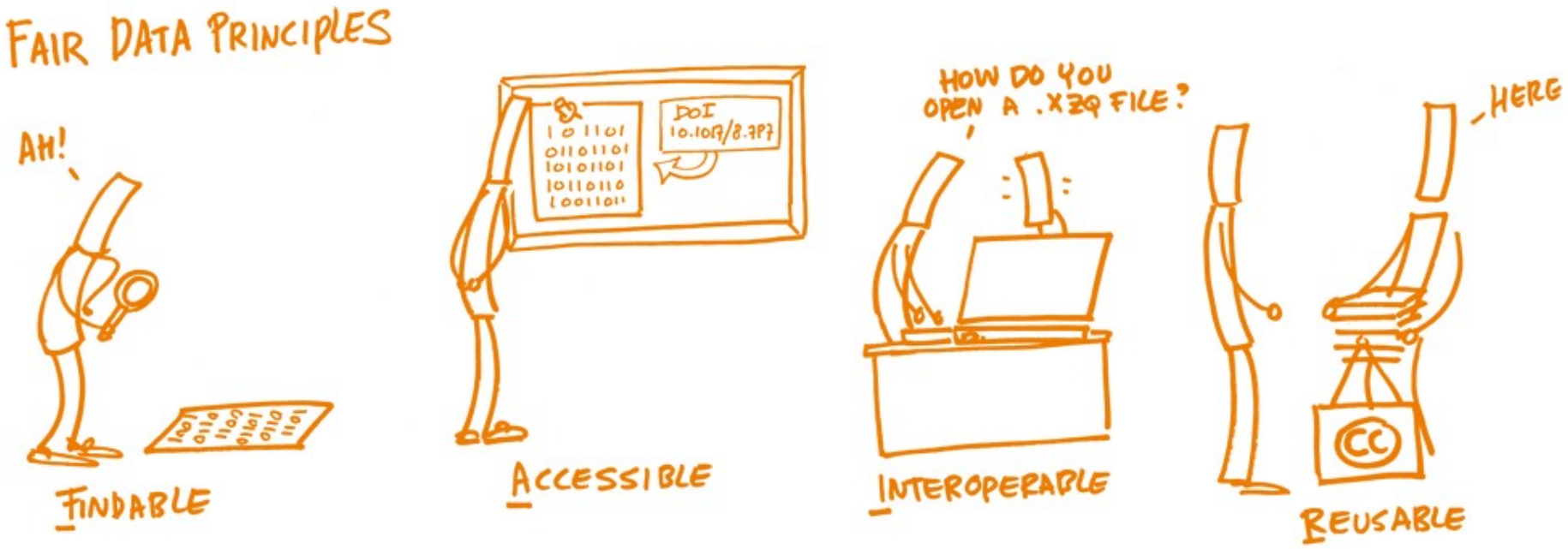
Varieties of research code

- Scripts
- Workflows
- R packages & Python modules
- Interactive apps (e.g. Shiny)
- Polished software products
- …?

Data & code sharing platforms: considerations

- Open licensing
- Permanent archival (DOI)
- FAIR principles
- …?

FAIR principles to regulate the level of openness



Open licenses

Access | Use | Modify | Share

Copyright (c) 2002 JSON.org

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

The Software shall be used for Good, not Evil.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

creative commons

CC0

“No Rights Reserved”

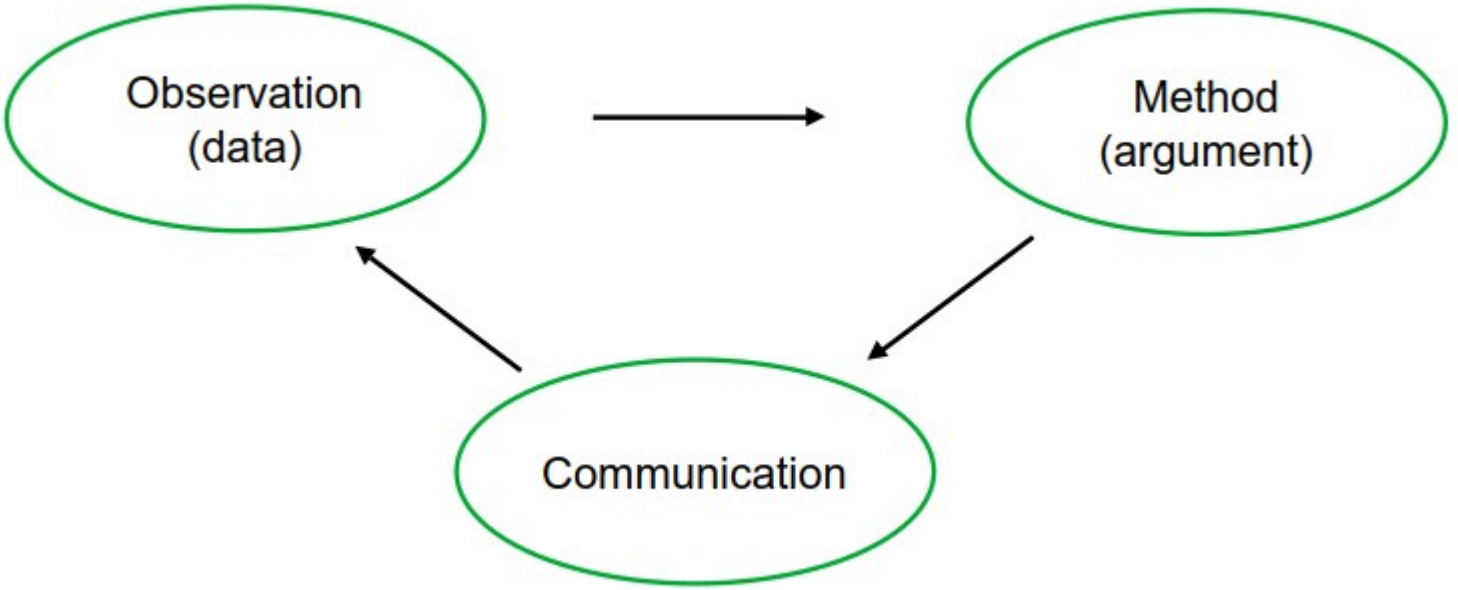
PUBLIC DOMAIN

CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law.

Open licenses

- Avoid CC-BY-NC (not for commercial use)
- NC violates the definition of “open”
- How to interpret “commercial”?
- Therefore it is often not valid as open output required by funders or organizations

Reproducible research

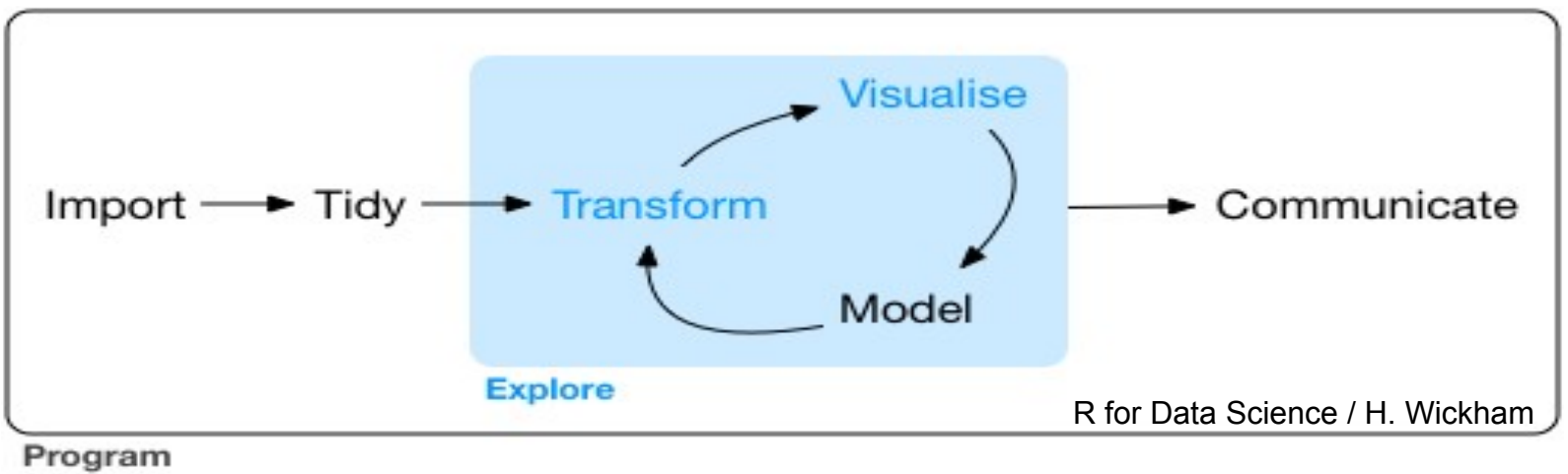


The influence of hidden researcher decisions in applied microeconomics

Nick Huntington-Klein ✉, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, Yaniv Stopnitzky

First published: 22 March 2021
https://doi.org/10.1111/ecin.12992

Researchers make hundreds of decisions about data collection, preparation, and analysis in their research. We use a many-analysts approach to measure the extent and impact of these decisions. Two published causal empirical results are replicated by seven replicators each. We find large differences in data preparation and analysis decisions, many of which would not likely be reported in a publication. No two replicators reported the same sample size. Statistical significance varied across replications, and for one of the studies the effect's sign varied as well. The standard deviation of estimates across replications was 3–4 times the mean reported standard error.



OPEN ACCESS

ESSAY

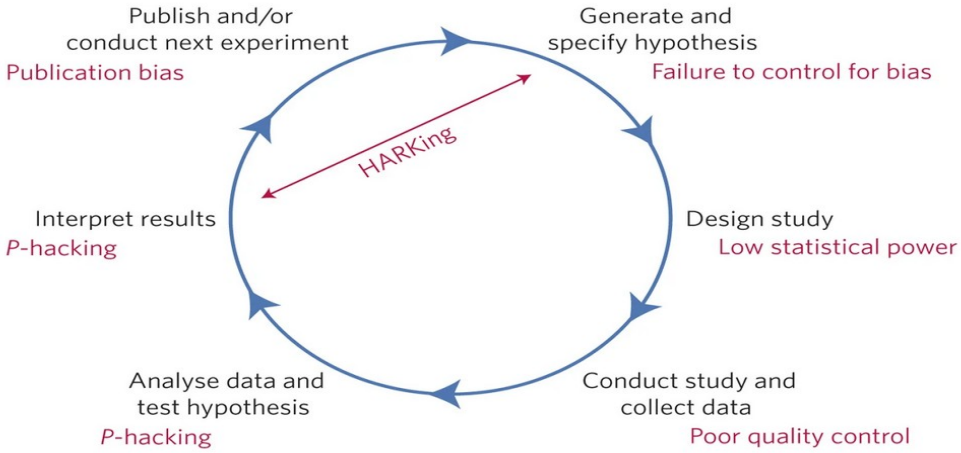
Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

Figure 1: Threats to reproducible science.

From: A manifesto for reproducible science



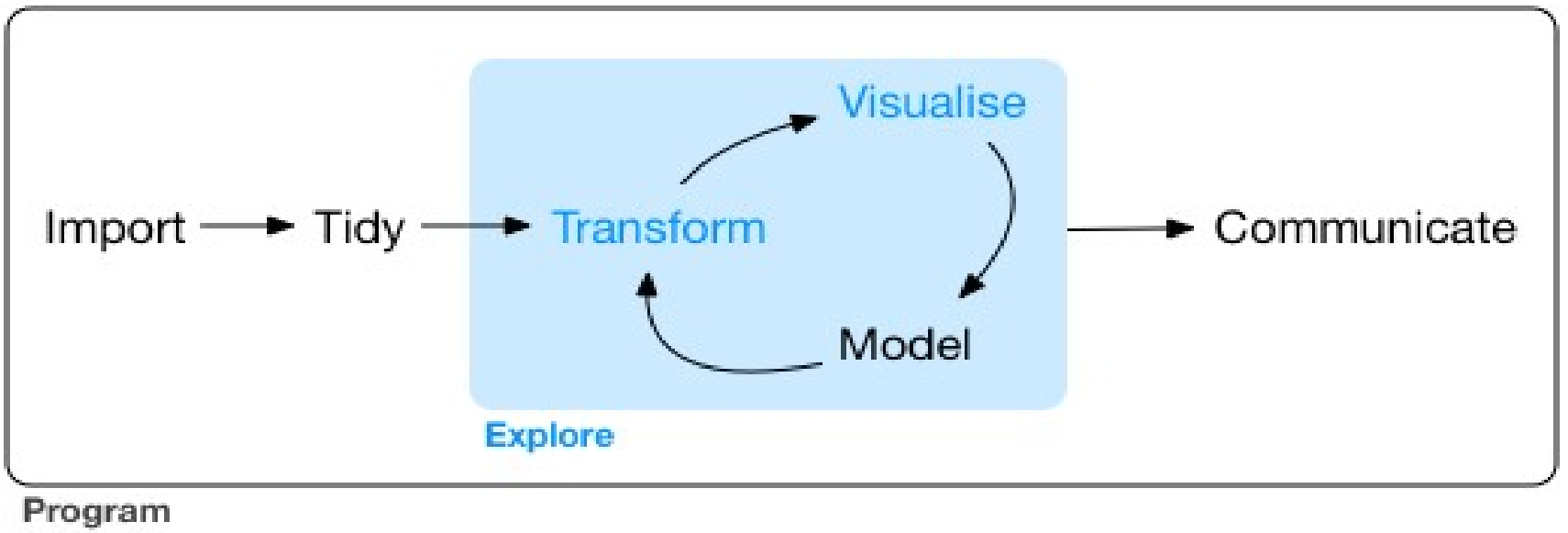
An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

A manifesto for reproducible science

Marcus R. Munafò ✉, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis

Nature Human Behaviour **1**, Article number: 0021 (2017) | Cite this article

Reproducible research



R for Data Science / H. Wickham

PLOS | MEDICINE

Browse | For Authors | Abou

OPEN ACCESS

ESSAY

898,944
VIEWS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

How to Make More Published Research True

John P. A. Ioannidis

Published: October 21, 2014 • DOI: 10.1371/journal.pmed.1001747

PLOS | MEDICINE

Browse | For Authors | Abou

OPEN ACCESS

ESSAY

898,944
VIEWS

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: Aug 30, 2005 • DOI: 10.1371/journal.pmed.0020124

BROWSE PUBLISH ABOUT

PLOS MEDICINE

OPEN ACCESS

CORRESPONDENCE

Why Most Published Research Findings Are False: Problems in the Analysis

Steven Goodman, Sander Greenland

Published: April 24, 2007 • https://doi.org/10.1371/journal.pmed.0040168

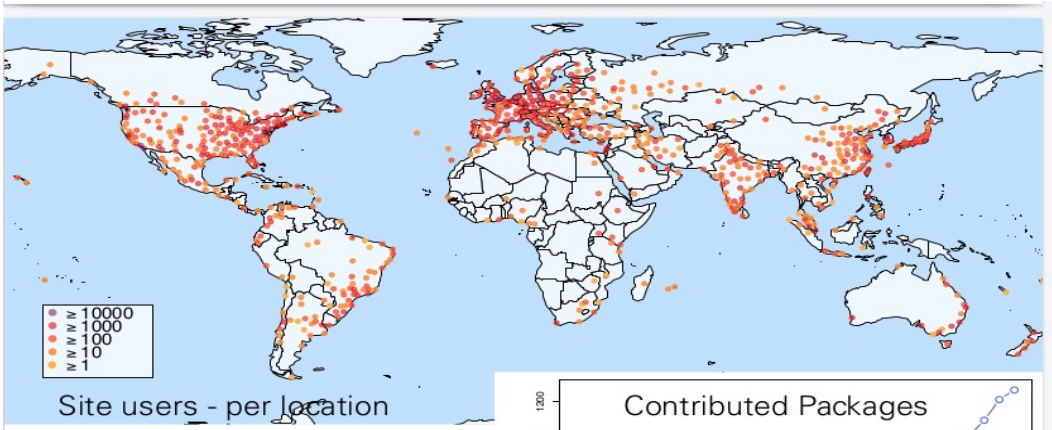
"*I have begun to think that no one ought to publish biometric results, without lodging a well arranged and well bound manuscript copy of all his data, in some place whereit should be accessible, under reasonable restrictions, to those who desire to verify his work.*"

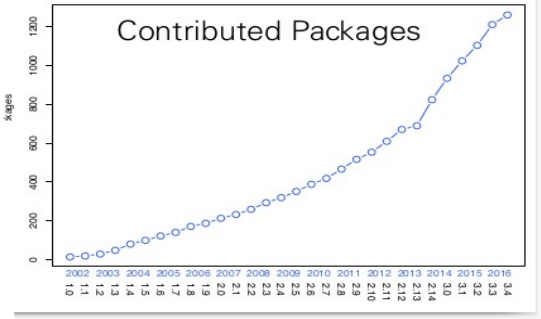
Francis Galton (1901), *Biometrika* 1:1, pp. 7-10.

Alchemy & algorithms: perspectives on the philosophy and history of open science

Leo Lahti, Filipe da Silva, Markus Petteri Laine, Viivi Lähteenoja, Mikko Tolonen

A family of alchemists at work, an engraving by Philip Galle, after a painting by Pieter Bruegel the Elder, published by Hieronymus Cock, *c.*1558.





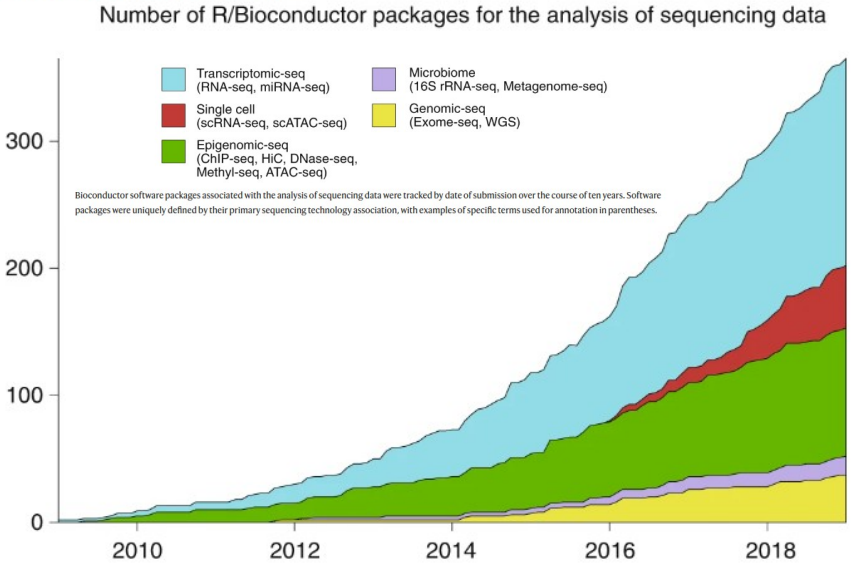
World's largest bioinformatics project:

- 2300+ packages
- Tens of thousands of users & publications
- Multi-omics integration

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Fig. 1: Number of Bioconductor packages for the analysis of high-throughput sequencing data over ten years.

From: Orchestrating single-cell analysis with Bioconductor



nature methods

Explore content ▾ About the journal ▾ Publish with us ▾

nature › nature methods › perspectives › article

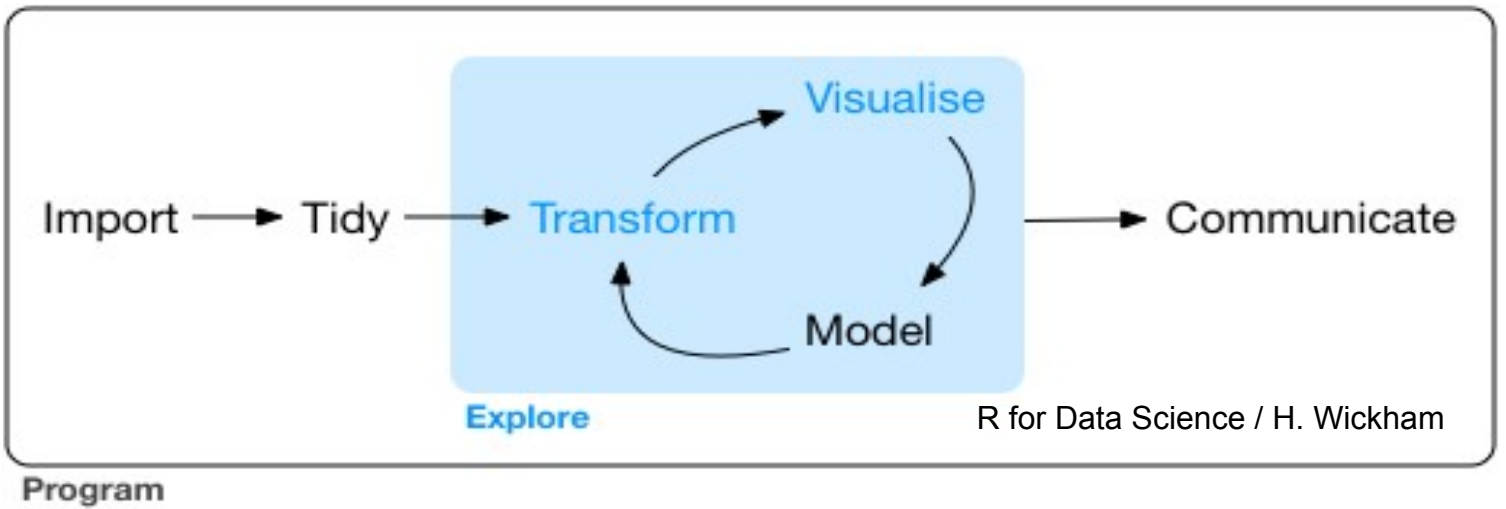
Perspective | Published: 02 December 2019

Orchestrating single-cell analysis with Bioconductor

Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger,

Reproducible notebooks / demo

- Jupyter notebooks
- Rmarkdown → Quarto (quarto.org)
- … other…?



R for Data Science / H. Wickham

Data Management Plan (DMP)

- Overview of a typical plan
- Collaboratively develop your own tentative plan (team work)

Best (or good enough) practices

Data & code

- Reproducibility
- Supporting standard libraries
- Version control & semantic versioning
- Permanent DOI

Publications:

- Data availability statement
- Code availability statement

Projects:

- Data management plan

Learning goals revisited

Elements of responsible research

- Sensitive information
- Open science, FAIR principles & reproducible research

Common tools for data & code management

- Version control (Git)
- Data & code sharing platforms
- Reprocibility
- Data Management Plan (DMP)