The Alan Turing Institute

# *The Turing Way* Workshop: Reproducible, Open and FAIR Research

## Session 1

Emma Karoune & Malvika Sharan
Pronouns: she/her/hers

@turingway, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

## Emma Karoune

Historic England

The Alan Turing Institute

- **Archaeobotanist/Palaeoecologist**
    - FAIR Phytoliths project
    - Open reference collections

International Phytolith Society

- **Open Researcher & Community Manager**
    - Tools, practices and systems programme
        - Turing Way
        - DECOVID/ Turing-RSS Lab
    - SSI Fellow/ UK-Elixir FAIR data Fellow

Software Sustainability Institute

Link to Emma's SSI Fellows page

The Alan Turing Institute

Senior Researcher

Tools, Practices and Systems

The Alan Turing Institute, UK

- PhD in Bioinformatics
- Open Access & Open Source research publications
- Computational and Open Science skill training (2015-)
- Community Building in Open Science (2016-)
- Co-lead of *The Turing Way* & Open Life Science (2019-)

Open Life Science

OPEN BIOINFORMATICS FOUNDATION

Software Sustainability Institute

THE CARPENTRIES

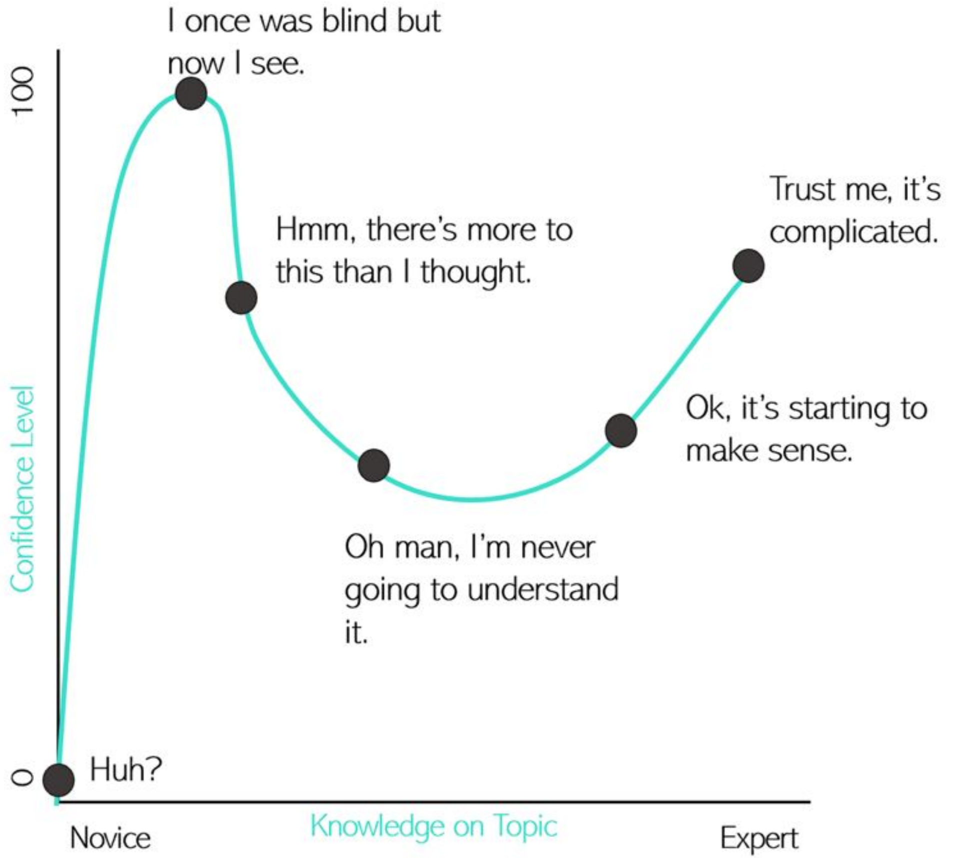
# Learning Objectives

- **Problem**: Scientific errors have real world effect
- Define what reproducible research is
- Understand what open science practices are
- Identify FAIR principles for your research
- Differentiate between FAIR and open data/research
- Learn how to implement a reproducible workflow

# Disclaimer:

You probably already know all about it!

Kaylee Somerville, The Hidden Power of Intellectual Humility - The Decision Lab. 2020. https://thedecisionlab.com/insights/society/the-hidden-power-of-intellectual-humility



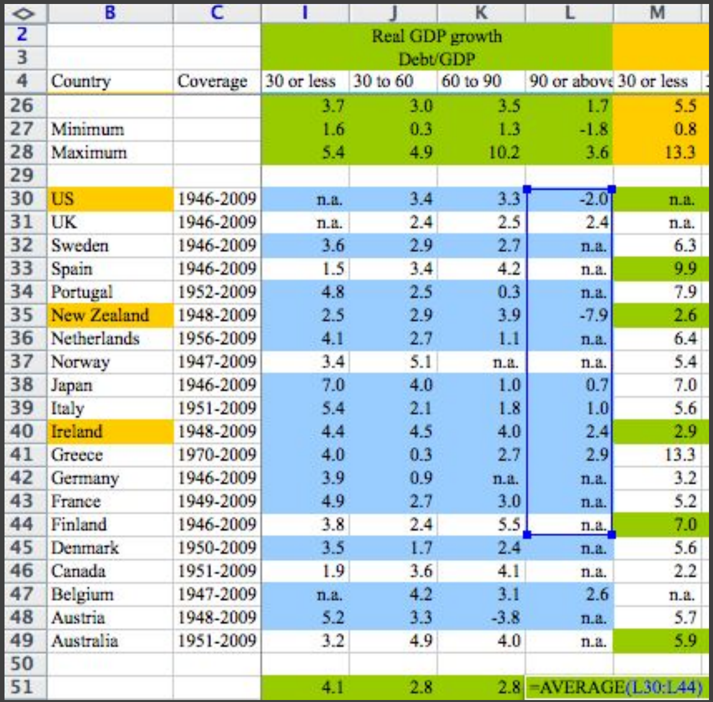
*Adapted from: Squad. (2018, December 13). Dunning-Kruger Effect: Definition, Test, Examples & Quiz. Science Terms.* **https://scienceterms.net/psychology/dunning-kruger-effect/**

# Scientific errors have real world effects

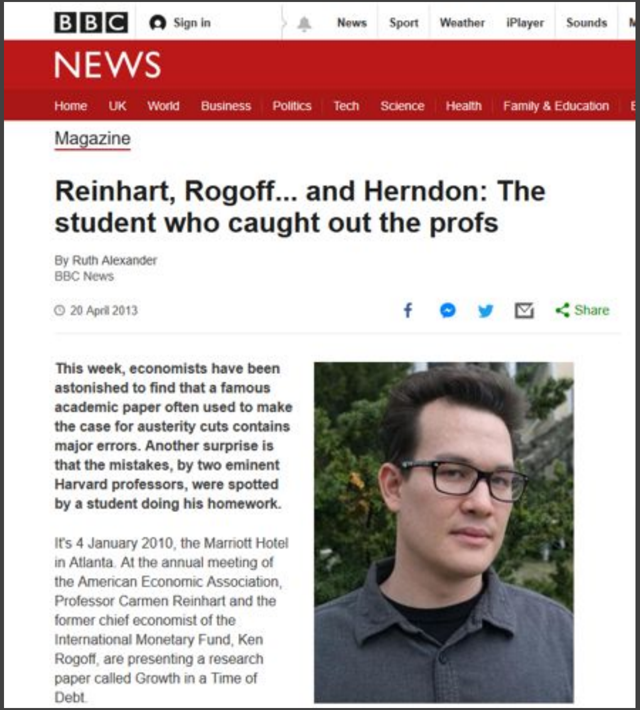
*Researchers have an obligation to consider the ethical standards (right actions) and their impact on society.*

CC-BY 4.0, https://the-turing-way.netlify.app/ethical-research/ethical-research.html, DOI: 10.5281/zenodo.6337939

# Scientific errors have real world effects



| ◇ | B | C | I | J | K | L | M |
|---|---|---|---|---|---|---|---|
| 2 | | | Real GDP growth | | | | |
| 3 | | | Debt/GDP | | | | |
| 4 | Country | Coverage | 30 or less | 30 to 60 | 60 to 90 | 90 or above | 30 or less |
| 26 | | | 3.7 | 3.0 | 3.5 | 1.7 | 5.5 |
| 27 | Minimum | | 1.6 | 0.3 | 1.3 | -1.8 | 0.8 |
| 28 | Maximum | | 5.4 | 4.9 | 10.2 | 3.6 | 13.3 |
| 29 | | | | | | | |
| 30 | US | 1946-2009 | n.a. | 3.4 | 3.3 | -2.0 | n.a. |
| 31 | UK | 1946-2009 | n.a. | 2.4 | 2.5 | 2.4 | n.a. |
| 32 | Sweden | 1946-2009 | 3.6 | 2.9 | 2.7 | n.a. | 6.3 |
| 33 | Spain | 1946-2009 | 1.5 | 3.4 | 4.2 | n.a. | 9.9 |
| 34 | Portugal | 1952-2009 | 4.8 | 2.5 | 0.3 | n.a. | 7.9 |
| 35 | New Zealand | 1948-2009 | 2.5 | 2.9 | 3.9 | -7.9 | 2.6 |
| 36 | Netherlands | 1956-2009 | 4.1 | 2.7 | 1.1 | n.a. | 6.4 |
| 37 | Norway | 1947-2009 | 3.4 | 5.1 | n.a. | n.a. | 5.4 |
| 38 | Japan | 1946-2009 | 7.0 | 4.0 | 1.0 | 0.7 | 7.0 |
| 39 | Italy | 1951-2009 | 5.4 | 2.1 | 1.8 | 1.0 | 5.6 |
| 40 | Ireland | 1948-2009 | 4.4 | 4.5 | 4.0 | 2.4 | 2.9 |
| 41 | Greece | 1970-2009 | 4.0 | 0.3 | 2.7 | 2.9 | 13.3 |
| 42 | Germany | 1946-2009 | 3.9 | 0.9 | n.a. | n.a. | 3.2 |
| 43 | France | 1949-2009 | 4.9 | 2.7 | 3.0 | n.a. | 5.2 |
| 44 | Finland | 1946-2009 | 3.8 | 2.4 | 5.5 | n.a. | 7.0 |
| 45 | Denmark | 1950-2009 | 3.5 | 1.7 | 2.4 | n.a. | 5.6 |
| 46 | Canada | 1951-2009 | 1.9 | 3.6 | 4.1 | n.a. | 2.2 |
| 47 | Belgium | 1947-2009 | n.a. | 4.2 | 3.1 | 2.6 | n.a. |
| 48 | Austria | 1948-2009 | 5.2 | 3.3 | -3.8 | n.a. | 5.7 |
| 49 | Australia | 1951-2009 | 3.2 | 4.9 | 4.0 | n.a. | 5.9 |
| 50 | | | | | | | |
| 51 | | | 4.1 | 2.8 | 2.8 | =AVERAGE(L30:L44) | |



## Reinhart, Rogoff... and Herndon: The student who caught out the profs

By Ruth Alexander
BBC News

20 April 2013

**This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.**

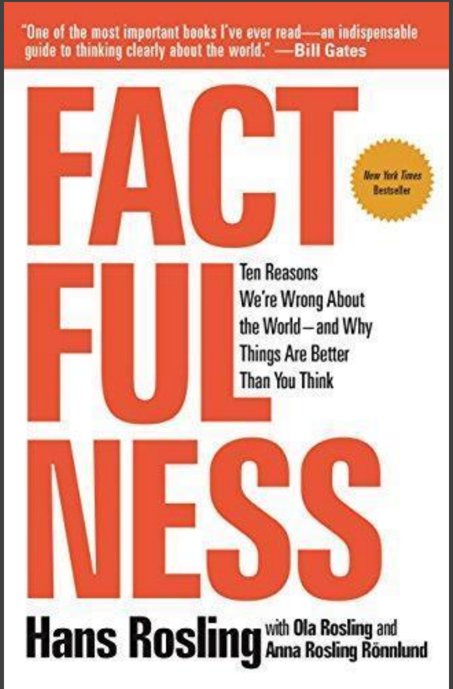
It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

https://statmodeling.stat.columbia.edu/2013/04/16/memo-to-reinhart-and-rogoff-i-think-its-best-to-admit-your-errors-and-go-on-from-there
https://www.bbc.co.uk/news/magazine-22223190

# Scientific errors have real world effects

“We will redouble our efforts to avoid such errors in the future …” - Authors

“mistaken way to examine (country specific) data provides an intellectual rationalisation for things that affect how people think about the world.”
- Daniel Hamermesh



“One of the most important books I've ever read—an indispensable guide to thinking clearly about the world.” —Bill Gates

FACT FUL NESS

New York Times Bestseller

Ten Reasons We're Wrong About the World—and Why Things Are Better Than You Think

Hans Rosling with Ola Rosling and Anna Rosling Rönnlund

https://www.gapminder.org/, Rosling, H., Rönnlund, A. R., & Rosling, O. (2018). Factfulness: Ten Reasons We're Wrong About the World--and Why Things Are Better Than You Think. https://www.bbc.co.uk/news/magazine-22223190

# Price of Popularity

Researchers in Australia used unreliable COVID-19 data and misreported that an anti-parasite drug could stop the virus from replicating in cells.

*In a desperate attempt to save dying patients, doctors began justifying the drug's use against COVID-19 as the virus spread aggressively throughout Latin America.*

*Reardon, S. (2021). Flawed ivermectin preprint highlights challenges of COVID drug studies. Nature, 596, 173–174. doi: 10.1038/d41586-021-02081-w. Mega, E. R. (2020). Latin America's embrace of unproven COVID treatment hinders drug trials. Nature, 586, 481–482. doi: 10.1038/d41586-020-02958-2,* Slides under DOI: 10.5281/zenodo.6337939

# Post-publication Peer Review is Important, But …

*To consult the [experts] after an experiment is finished is often merely to ask to conduct a post mortem examination. […] can perhaps say what the experiment died of.* - Ronald Fisher

https://statmodeling.stat.columbia.edu/2013/04/16/memo-to-reinhart-and-rogoff-i-think-its-best-to-admit-your-errors-and-go-on-from-there
https://www.bbc.co.uk/news/magazine-22223190

# Avoid Errors Before the Harm Occurs

1. Mistaken research design or analysis processes
    - **Project design for open, FAIR & reproducible research**

Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. Nature, 530(7588), 27. doi: 10.1038/530027a

# Avoid Errors Before the Harm Occurs

1. Mistaken research design or analysis processes
   - **Project design for open, FAIR & reproducible research**
2. Wrong choices of tools and methods
   - **Data handling, data management, collaboration process**

Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. Nature, 530(7588), 27. doi: 10.1038/530027a

# Avoid Errors Before the Harm Occurs

1. Mistaken research design or analysis processes
   - **Project design for open, FAIR & reproducible research**
2. Wrong choices of tools and methods
   - **Data handling, data management, collaboration process**
3. Inappropriate baseline comparison
   - **Lack of technical understanding (we won’t discuss this!)**

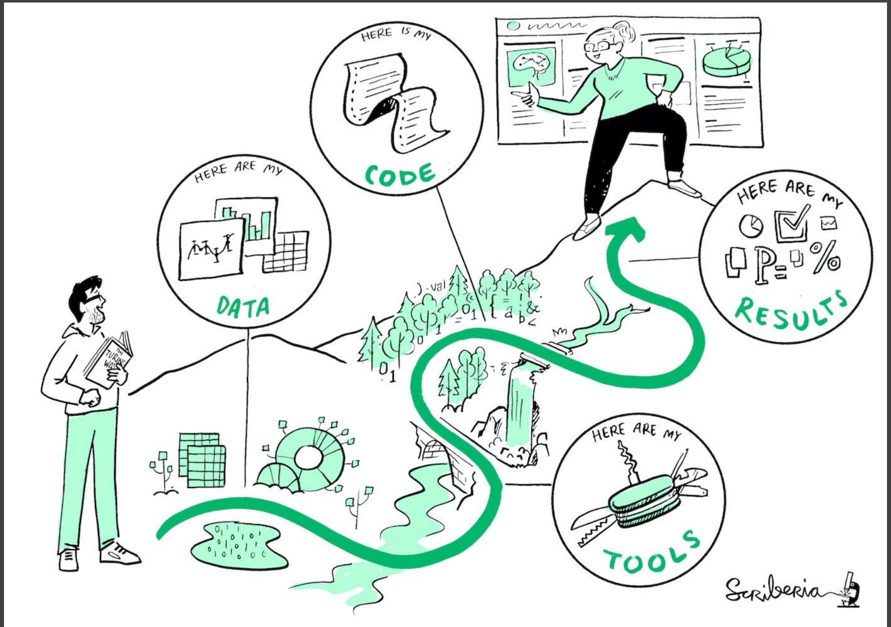
Allison, D. B., Brown, A. W., George, B. J., & Kaiser, K. A. (2016). Reproducibility: A tragedy of errors. Nature, 530(7588), 27. doi: 10.1038/530027a

## Session 1 - Reproducible, Open and FAIR research

- What is reproducible research
- What is open research
- Concerns about opening up research
- Understanding FAIR data/research
- Making your research FAIR
- Setting up repositories and working collaboratively

# Research Reproducibility

*Reproducible research save valuable time in verifying and building upon existing solutions.*

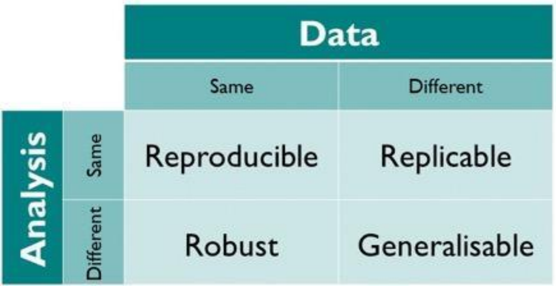


@malvikasharan, @turingway, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

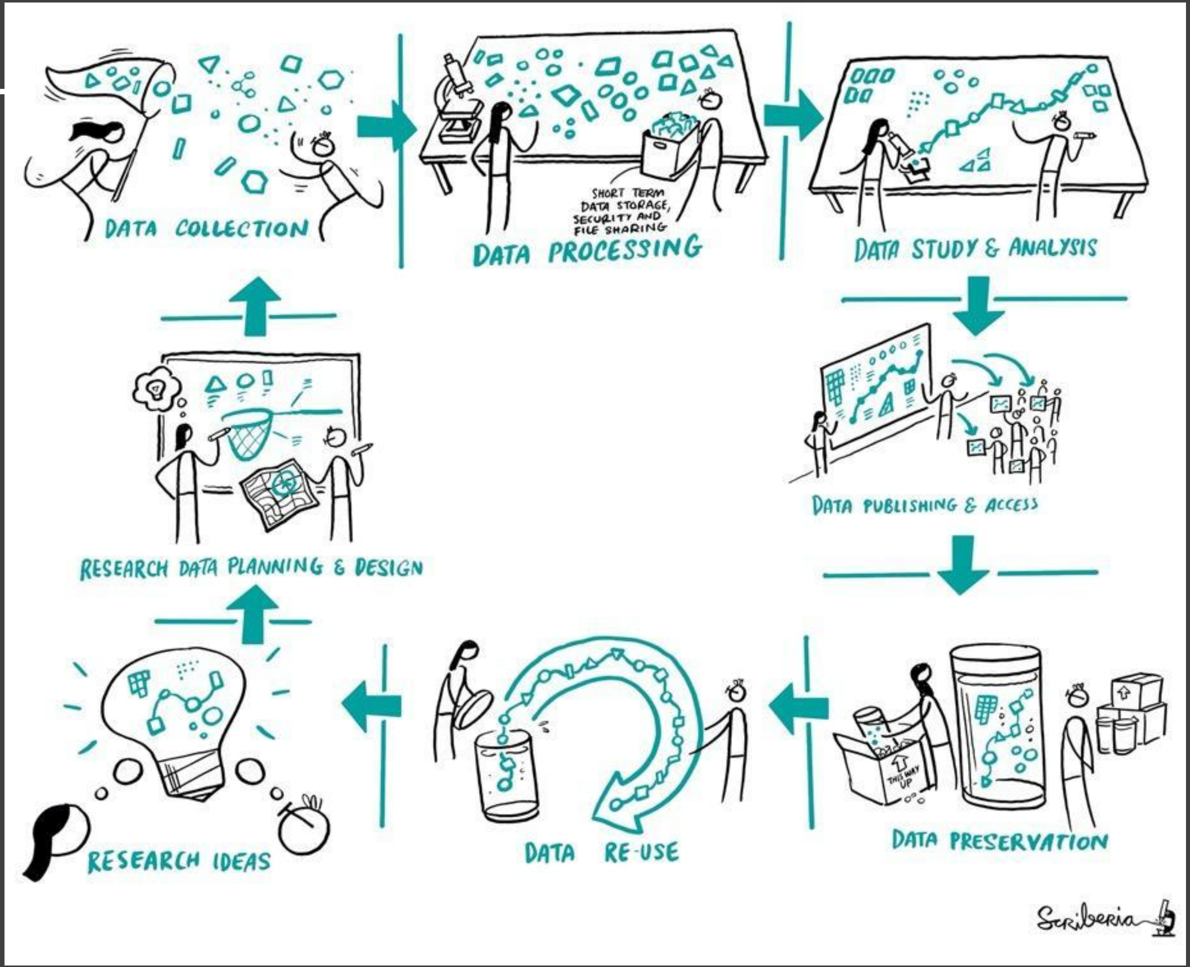
| | | Data | |
|---|---|---|---|
| | | Same | Different |
| Analysis | Same | Reproducible | Replicable |
| | Different | Robust | Generalisable |

https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

# Reproducible Research Workflows

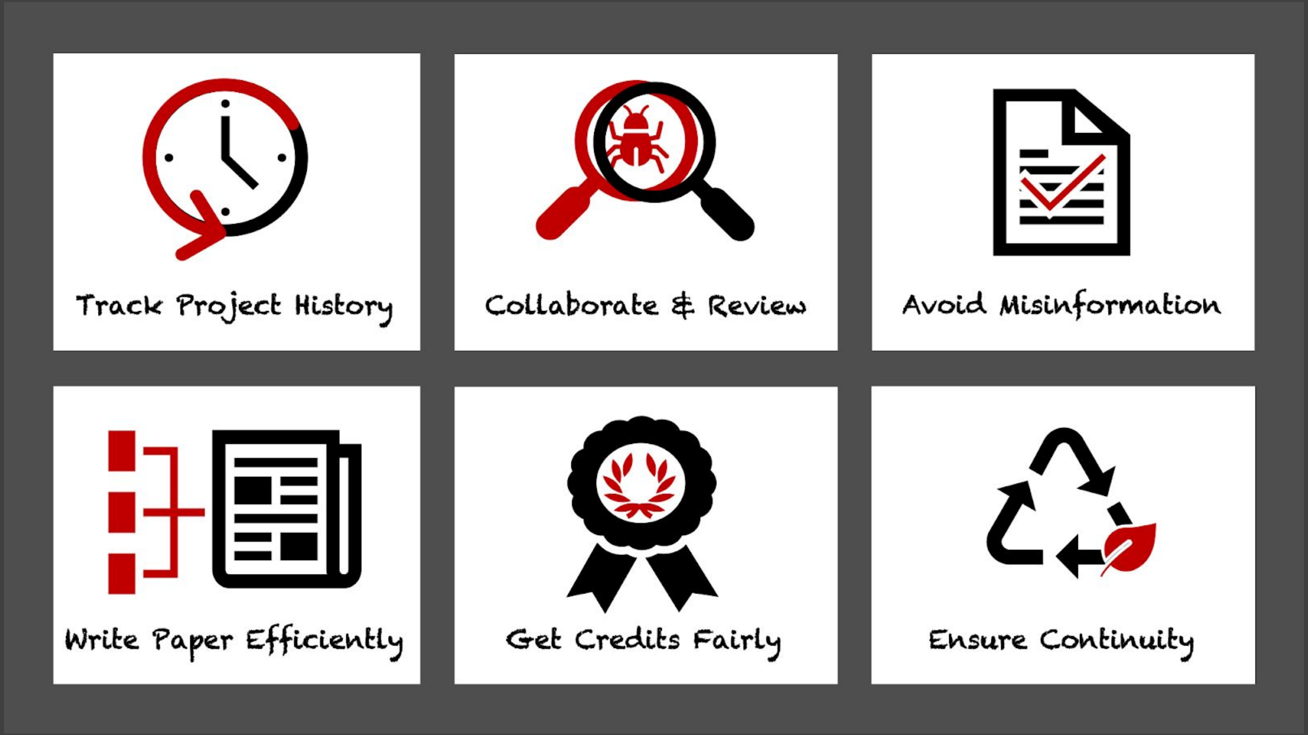


| | | Data | |
|---|---|---|---|
| | | Same | Different |
| Analysis | Same | Reproducible | Replicable |
| | Different | Robust | Generalisable |



@malvikasharan, CC-BY 4.0, CC-BY  image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# Why do reproducible research?



https://the-turing-way.netlify.app/reproducible-research/overview/overview-benefit.html, DOI: 10.5281/zenodo.6337939

Applying best practices in our research requires intention, resources, time and collaboration, which can be overwhelming.

IF I SHARE MY DATA.... PEOPLE MIGHT FIND MISTAKES
NEED TO MAKE IT OK TO BE HUMAN
SPRINGER NATURE

https://www.software.ac.uk/blog/2020-12-17-ten-arguments-against-open-science-you-can-win, DOI: 10.5281/zenodo.6337939 slido

# What are some barriers to reproducibility?

ⓘ Start presenting to display the poll results on this slide.

# Barriers to reproducible research

Held to higher standards than others

Is not considered for promotion

Publication bias towards novel findings

Requires additional skills

Plead the 5th

Support additional users

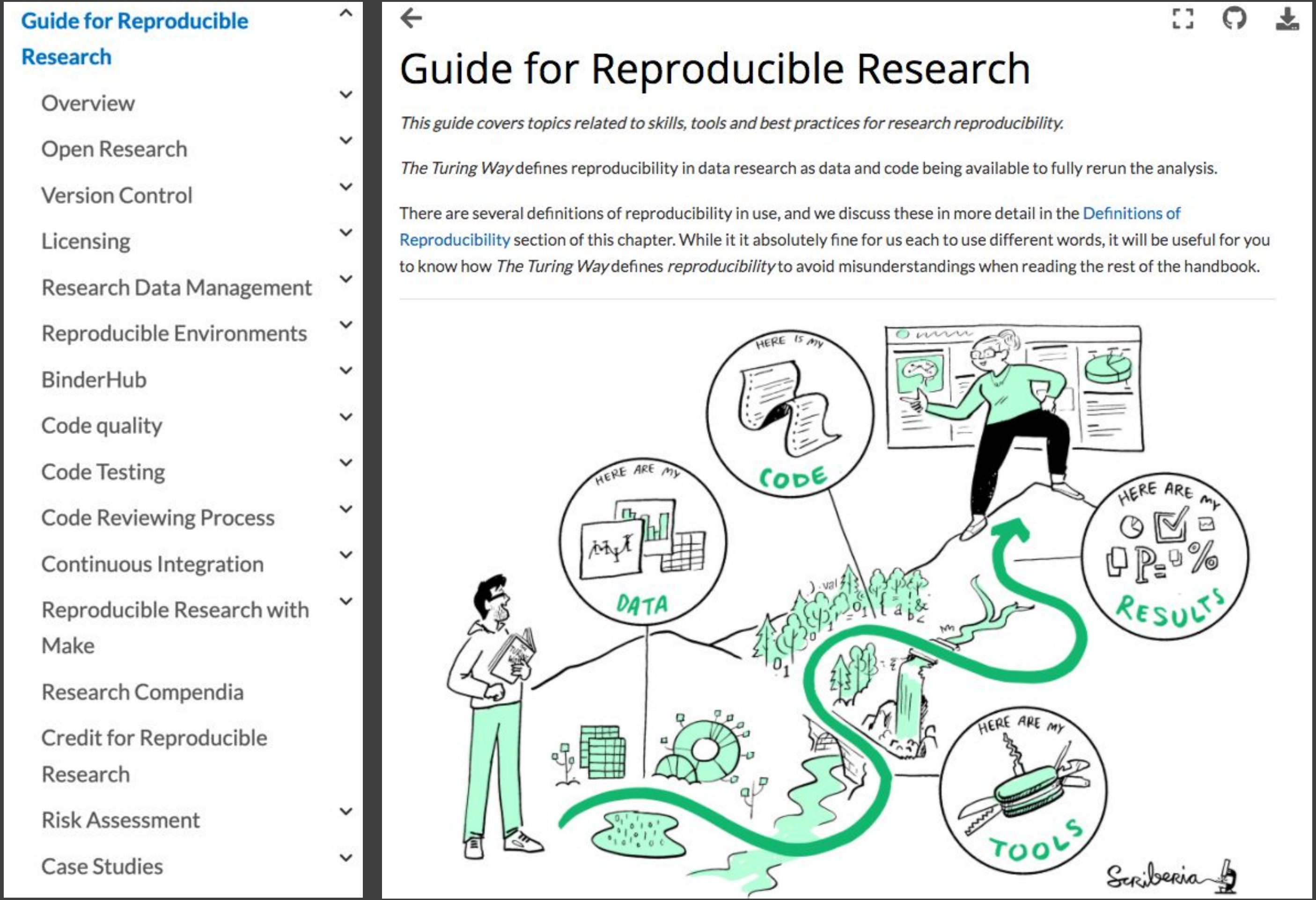
Takes time

https://doi.org/10.6084/m9.figshare.5537101

# *The Turing Way*

An Open Source project that involves and supports its **diverse community** to make data science **reproducible, ethical, collaborative and inclusive** for you.

https://github.com/alan-turing-institute/the-turing-way,
@turingway, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939



# Guide for Reproducible Research

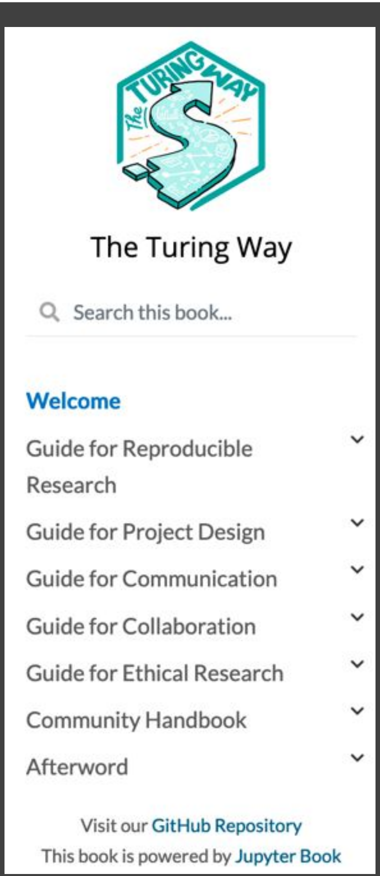
*This guide covers topics related to skills, tools and best practices for research reproducibility.*

*The Turing Way* defines reproducibility in data research as data and code being available to fully rerun the analysis.

There are several definitions of reproducibility in use, and we discuss these in more detail in the Definitions of Reproducibility section of this chapter. While it it absolutely fine for us each to use different words, it will be useful for you to know how *The Turing Way* defines *reproducibility* to avoid misunderstandings when reading the rest of the handbook.

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# A book: “Work in Progress”



The Turing Way

Search this book...

**Welcome**

Guide for Reproducible Research

Guide for Project Design

Guide for Communication

Guide for Collaboration

Guide for Ethical Research

Community Handbook

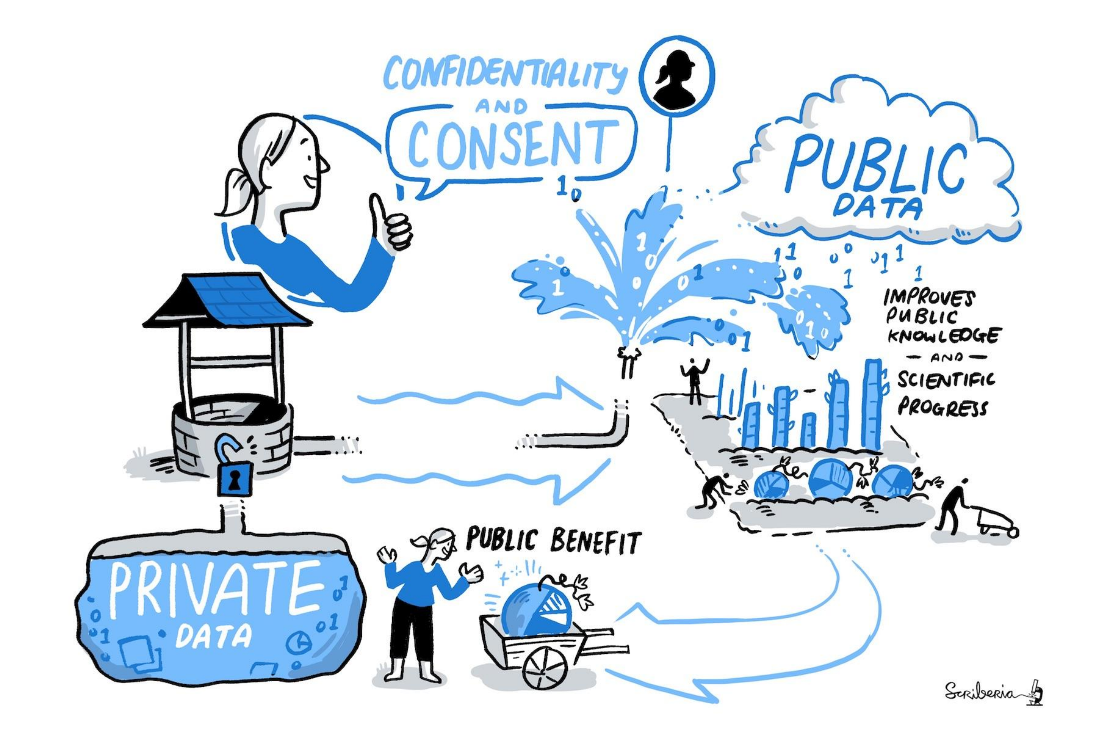
Afterword

Visit our GitHub Repository

This book is powered by Jupyter Book

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# Reproducible Research Workflows


CONFIDENTIALITY AND CONSENT
PUBLIC DATA
IMPROVES PUBLIC KNOWLEDGE — AND — SCIENTIFIC PROGRESS
PUBLIC BENEFIT
PRIVATE DATA
Scriberia

## Inner Source vs Open Source

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# Privacy and Sensitivity concerns

Is my data too sensitive to share?

I’m not sure I want others to see all my work?

Do I have permission to share my work openly?

I don’t have time or can’t see how to share my work

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# Reflection Exercise 🤔

- What are your motivations to share your work?
- What are your concerns about sharing your work?

OPEN RESEARCH
OPPORTUNITY?
SENSITIVITY?
INCLUSIVITY?
FUNDING?
CLOSED RESEARCH
Scriberia

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

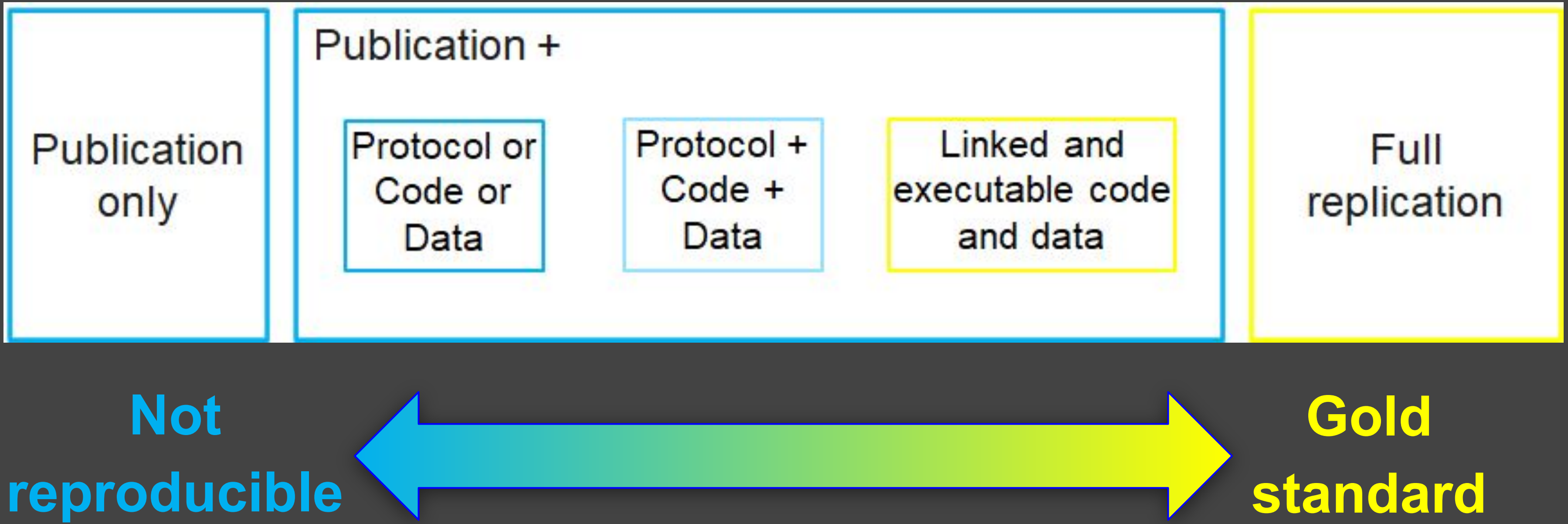
# Breakout Room: 15 Minutes

What do you need to consider to ensure that your work is reproducible?

- Research object: (1) Data, (2) code, (3) workflow, (4) documentation
- Practices: Version control, licensing, data management, communication, collaboration, reusability, long-term archiving
- Instructions: (i) Each room is assigned a research object. (ii) Nominate a notetaker in your room. (iii) Select 1 or 2 practices. (iv) Discuss at what stages of research cycle are these practices applied. (v) Report!

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# Reproducible Research Spectrum



Publication only

Publication +

Protocol or Code or Data

Protocol + Code + Data

Linked and executable code and data

Full replication

**Not reproducible**

**Gold standard**

*Adapted from Peng 2011*

*https://www.science.org/doi/abs/10.1126/science.1213847*

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# Reproducibility: Where should we start?

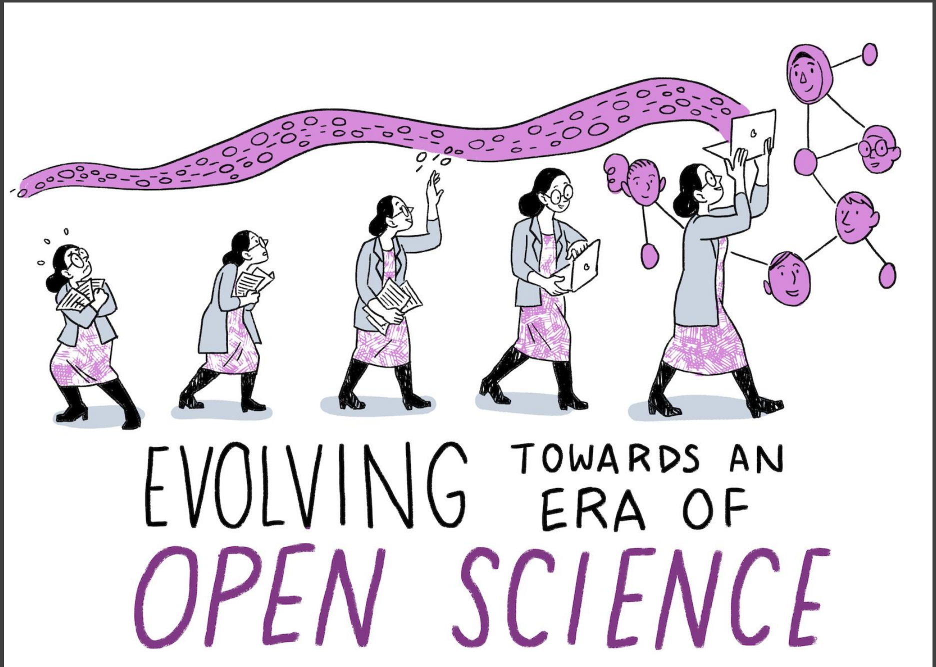
Reproducible research workflow:

1. Collaboration
- Using collaborative, version controlled and open ways to work with others.

2. Transparency
- Clear documentation of methods, data, code - openly shared research compendium.

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

# **Open Science** to enable Collaboration and Transparency



- Open Science practices remove barriers from sharing and using scientific resources at all stages of research.
- Make research findings accessible to all rather than keeping them locked away (for example, behind a paywall).

@turingway, CC-BY 4.0, https://www.fosteropenscience.eu/, DOI: 10.5281/zenodo.4609987

Open Science aims to transform research by making it more transparent → accessible → reliable → reproducible → reusable → collaborative → beneficial to society.

To achieve this openness in research, we need to make each element of the research process ...

- *publicly available,*
- *with permission to view, use, modify and distribute, and*
- *description for how one can collaborate.*

# Open Science is an umbrella term for open research practices

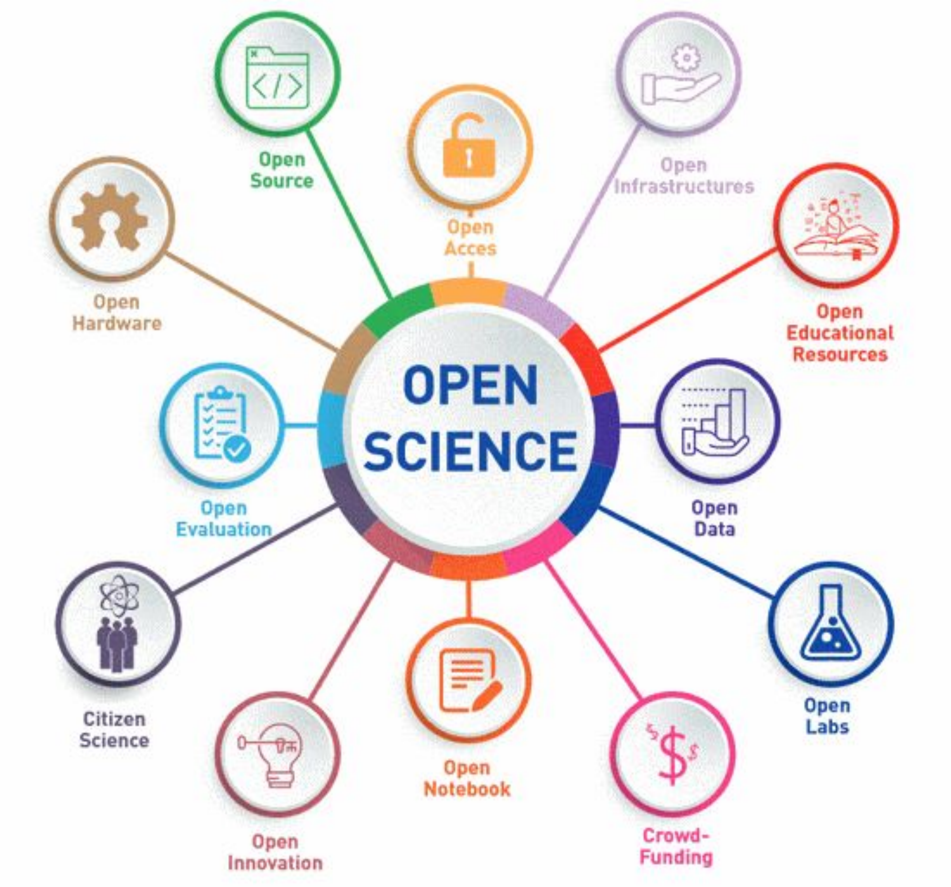
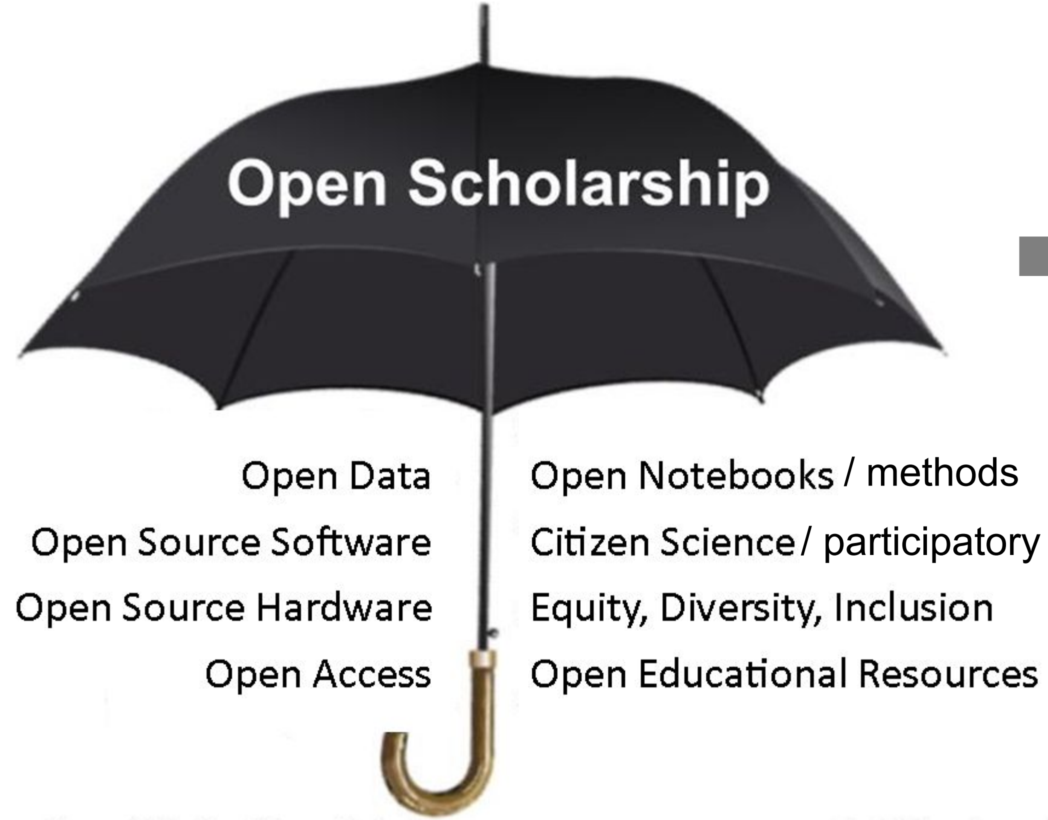


Image from UNESCO Open Science brochure, available under the CC-BY 4.0 license, DOI: 10.5281/zenodo.6337939

# What are some barriers we should work to remove?



- Open Data
- Open Source Software
- Open Source Hardware
- Open Access
- Open Notebooks / methods
- Citizen Science / participatory
- Equity, Diversity, Inclusion
- Open Educational Resources

Transparency →
Reproducibility →
Research Quality →
Sustainability

+

Collaboration →
Inclusive Research →
Equity and Diversity →
Global Accessibility

/.meetup.com/Berlin-Open-Science-Meetup/

Robin Champieux and Danielle Robinson

@turingway, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

Open Science

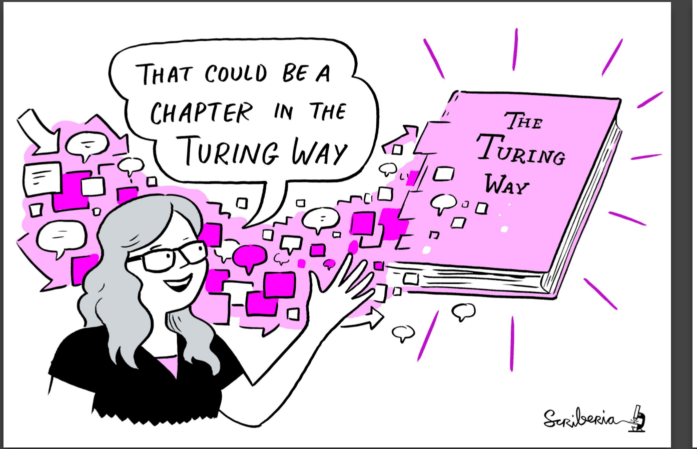
Open Research

Open Scholarship

We are going to use these terms interchangeably today to cover important considerations for you!

Book: https://the-turing-way.netlify.app/welcome, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939

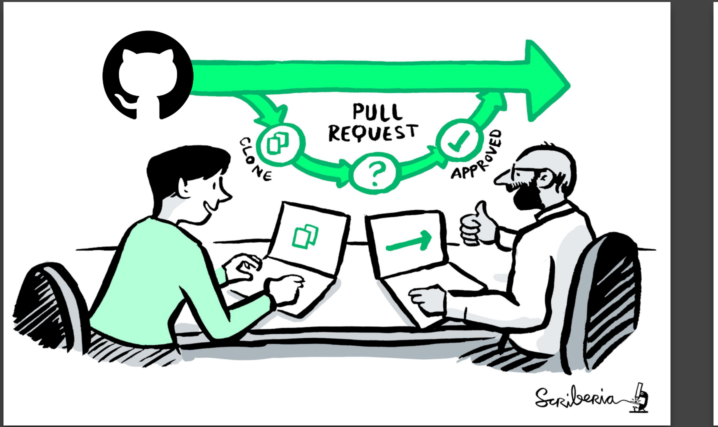
## A Book



## A Community

**Book:**
the- turing-way.netlify.app/

## An Open Source Project



## A Culture of Collaboration

There's more to collaboration than you might think!
Different teams
Common goal
Exchange of knowledge
Production
Treat each other kindly
Build diverse teams
Code of conduct
Inclusive workspace
Explicitely open for contribution
Scriberia

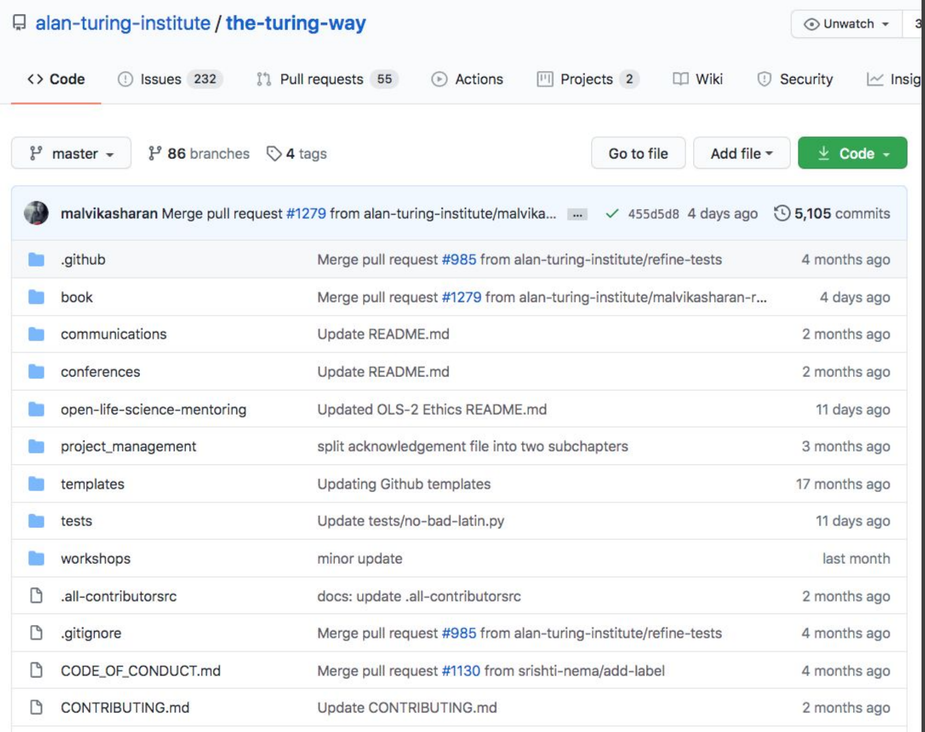
**GitHub:**
github.com/alan-turing-institute/the- turing-way

**Twitter:**
twitter.com/turingway

**Email:**
theturingway@gmail.com

# A Collaborative, Version Controlled and Open Project

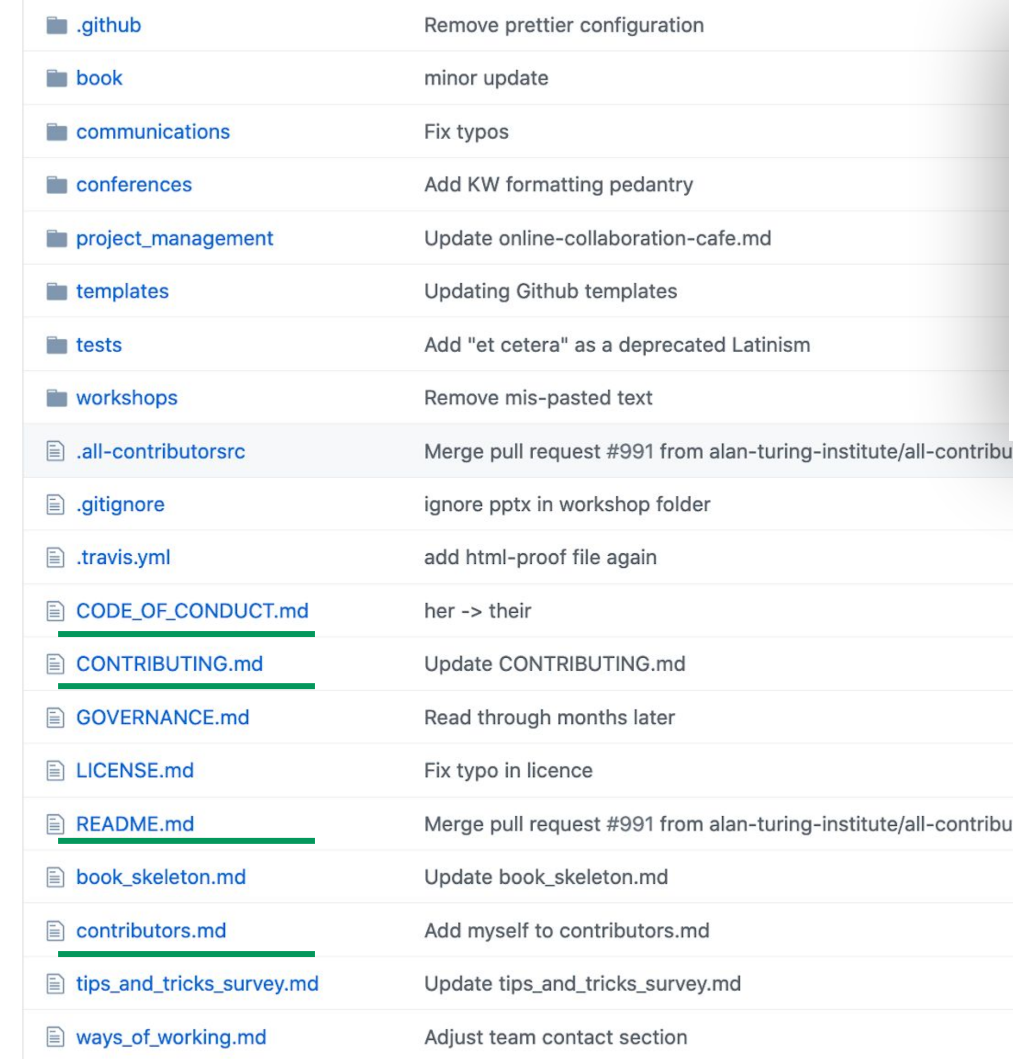
- Shared online
- Designed for open collaboration
- Hosted on GitHub with history and versions
- Described with open license
- Community oriented



malvikasharan Merge pull request #1279 from alan-turing-institute/malvika... ✓ 455d5d8 4 days ago 5,105 commits

| | | |
|---|---|---|
| .github | Merge pull request #985 from alan-turing-institute/refine-tests | 4 months ago |
| book | Merge pull request #1279 from alan-turing-institute/malvikasharan-r... | 4 days ago |
| communications | Update README.md | 2 months ago |
| conferences | Update README.md | 2 months ago |
| open-life-science-mentoring | Updated OLS-2 Ethics README.md | 11 days ago |
| project_management | split acknowledgement file into two subchapters | 3 months ago |
| templates | Updating Github templates | 17 months ago |
| tests | Update tests/no-bad-latin.py | 11 days ago |
| workshops | minor update | last month |
| .all-contributorsrc | docs: update .all-contributorsrc | 2 months ago |
| .gitignore | Merge pull request #985 from alan-turing-institute/refine-tests | 4 months ago |
| CODE_OF_CONDUCT.md | Merge pull request #1130 from srishti-nema/add-label | 4 months ago |
| CONTRIBUTING.md | Update CONTRIBUTING.md | 2 months ago |

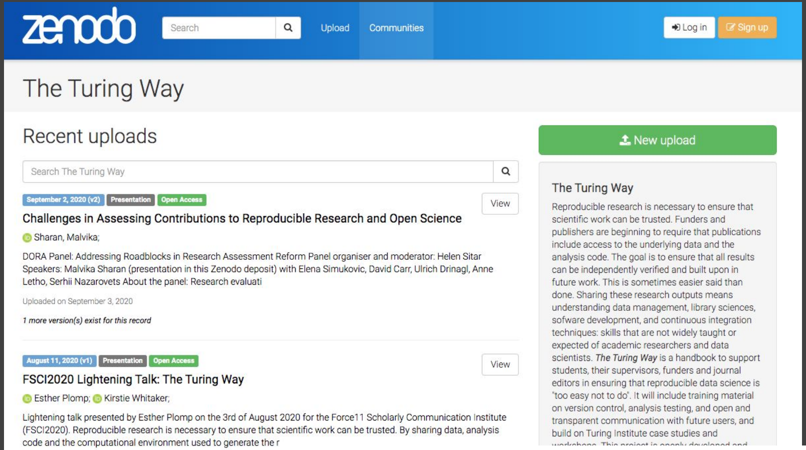
https://github.com/alan-turing-institute/the-turing-way/, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.6337939



| | | |
|---|---|---|
| .github | Remove prettier configuration | |
| book | minor update | |
| communications | Fix typos | |
| conferences | Add KW formatting pedantry | |
| project_management | Update online-collaboration-cafe.md | |
| templates | Updating Github templates | |
| tests | Add "et cetera" as a deprecated Latinism | |
| workshops | Remove mis-pasted text | |
| .all-contributorsrc | Merge pull request #991 from alan-turing-institute/all-contributors/a... | 5 days ago |
| .gitignore | ignore pptx in workshop folder | 9 months ago |
| .travis.yml | add html-proof file again | last month |
| CODE_OF_CONDUCT.md | her -> their | 6 months ago |
| CONTRIBUTING.md | Update CONTRIBUTING.md | 2 months ago |
| GOVERNANCE.md | Read through months later | 5 months ago |
| LICENSE.md | Fix typo in licence | 2 months ago |
| README.md | Merge pull request #991 from alan-turing-institute/all-contributors/a... | 5 days ago |
| book_skeleton.md | Update book_skeleton.md | 13 months ago |
| contributors.md | Add myself to contributors.md | 11 months ago |
| tips_and_tricks_survey.md | Update tips_and_tricks_survey.md | 14 months ago |
| ways_of_working.md | Adjust team contact section | 5 months ago |

# An Open Science project

- everyone can freely read, reuse, distribute, modify and help develop
- the project belongs to the research community (CC-BY license)
- Builds in collaboration with other projects



The Turing Way

Recent uploads

September 2, 2020 (v2) Presentation Open Access

Challenges in Assessing Contributions to Reproducible Research and Open Science

Sharan, Malvika;

DORA Panel: Addressing Roadblocks in Research Assessment Reform Panel organiser and moderator: Helen Sitar Speakers: Malvika Sharan (presentation in this Zenodo deposit) with Elena Simukovic, David Carr, Ulrich Drinagl, Anne Letho, Serhii Nazarovets About the panel: Research evaluati

Uploaded on September 3, 2020

*1 more version(s) exist for this record*

August 11, 2020 (v1) Presentation Open Access

FSCI2020 Lightening Talk: The Turing Way

Esther Plomp; Kirstie Whitaker;

Lightening talk presented by Esther Plomp on the 3rd of August 2020 for the Force11 Scholarly Communication Institute (FSCI2020). Reproducible research is necessary to ensure that scientific work can be trusted. By sharing data, analysis code and the computational environment used to generate the r

Uploaded on August 11, 2020

August 6, 2020 (v1) Journal article Open Access

The Turing Way workshop onBoost your reproducibility with Binder

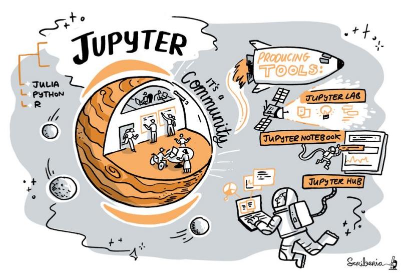
Sharan, Malvika;

This workshop was organized with the UKDRI team. Title: The Turing Way workshop on Boost your r Binder Date: 11 June, 2020 13:00 - 17:00 (GMT) Instructors: Kirstie Whitaker, Sarah Gibson, Malvika S https://hackmd.io/@malvikasharan/BinderJune2020 Agenda:

Uploaded on August 6, 2020

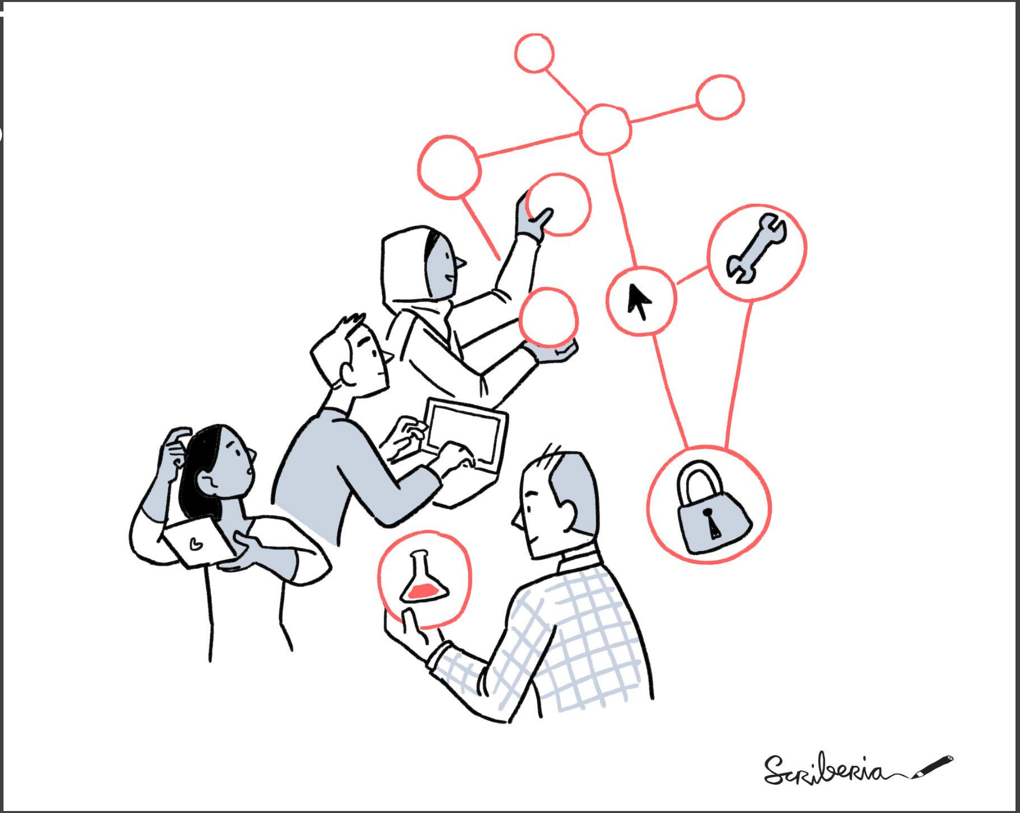
The Turing Way

Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future work. This is sometimes easier said than done. Sharing these research outputs means understanding data management, library sciences, sofware development, and continuous integration techniques: skills that are not widely taught or expected of academic researchers and data scientists. *The Turing Way* is a handbook to support students, their supervisors, funders and journal editors in ensuring that reproducible data science is "too easy not to do". It will include training material on version control, analysis testing, and open and transparent communication with future users, and build on Turing Institute case studies and



@turingway, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

# 🌻 Open Science: Where should we start?

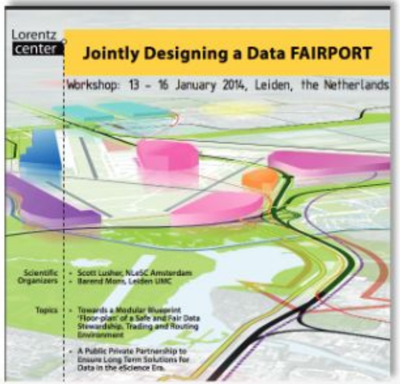


CC-BY 4.0, CC-BY  image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# FAIR principles

- Findable,
- Accessible
- Interoperable
- Reusable

CC-BY 4.0, CC-BY  image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939



Lorentz center

Jointly Designing a Data FAIRPORT

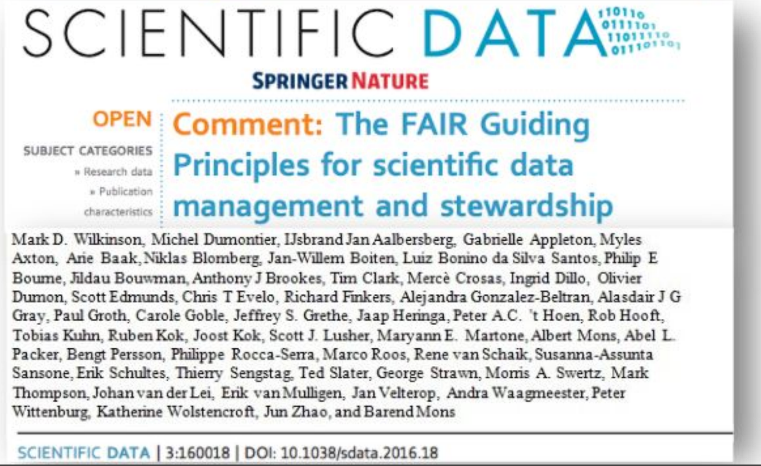
Workshop: 13 - 16 January 2014, Leiden, the Netherlands

Scientific Organizers
- Scott Lusher, NLeSC Amsterdam
- Barend Mons, Leiden UMC

Topics
- Towards a Modular Blueprint 'Floor-plan' of a Safe and Fair Data Stewardship, Trading and Routing Environment
- A Public Private Partnership to Ensure Long Term Solutions for Data in the eScience Era.

2014

2016



SCIENTIFIC DATA

SPRINGER NATURE

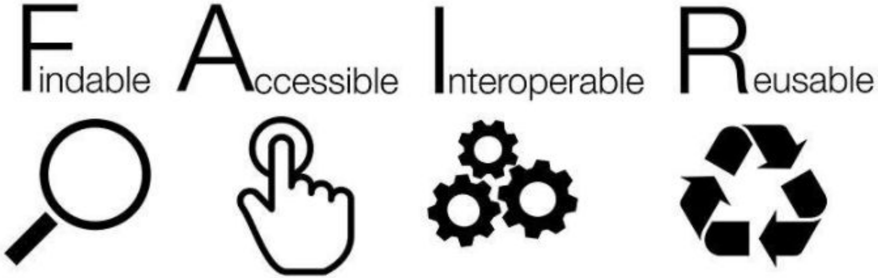
OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons

SCIENTIFIC DATA | 3:160018 | DOI: 10.1038/sdata.2016.18



Findable Accessible Interoperable Reusable

A set of principles to enhance the value of all digital resources

**Developed** and **endorsed** by *researchers, service providers, publishers, funding agencies and industry partners*

FAIR principles from **Wilkinson *et al*. (2016)**
**DOI: 10.1038/sdata.2016.18**

# FAIR data analogy

Annotation makes it easier to find important things

You would not buy food with no labels!

Labels make different foods easier to find and access in stores, combine with other foods (interoperable) and use in different ways.

Adapted from talk by Philippe Rocca-Serra (2020)

# What is the meaning of FAIR data ?

## F = Findable

*in an **online*** data catalogue / archive / portal

findable by **humans** and by **machines**

- ENA for DNA sequences
- GBif and OBIS for biodiversity data
- BioImage Archive for images of biological material
- Zenodo as a a general-purpose open-access repository

**Standardised** and **rich** discovery **Metadata** explaining:

☑ **Who**: is the **author** / **contact person** for questions

☑ **How**: were the data created --> **procedures** / **protocols**

☑ **How**: to **access** the data, consider **licenses**

☑ **What**: **keywords** describe the data

☑ **What**: **parameters** were measured, species & **geography** covered

☑ **When**: were the **data** and **updates** created

CC-BY 4.0, CC-BY image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# What is the meaning of FAIR data ?

## A = Accessible Data & Metadata

*from catalogue/archive/portal*

*via **machine to machine** and **human interfaces***

- **Web interfaces** for human searches & downloads
- **APIs** for searching & accessing
- Clear **instructions** for access (download, request access,.)
- **Keeping metadata** when data is deleted
- **Metadata update** when updating data / information
- **All data levels** should be archived: raw data is the most important and at a minimum must be provided

CC-BY 4.0, CC-BY image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# What is the meaning of F A I R data ?

## I = Interoperable

*readable & understandable by* ***humans / code*** *:*

❑ **Community-accepted** data formats & file types
  ○ open (non-proprietary)
  ○ sustainable (think in 10 years from now)

❑ **Clear**, **controlled vocabulary** for data & metadata
  ○ describing all relevant terms/values/units
  ○ specific ⟶ data/metadata “dictionary”

❑ Your data should be **standalone**, packaged up with
  ○ all necessary information and files to allow the data to be understood by anyone at any time

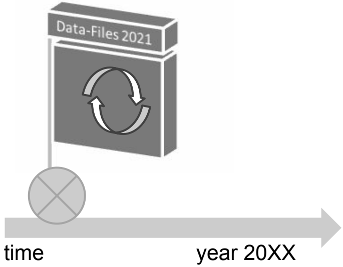
❑ **Readable** by code:
  ○ machine readable descriptions of data: files and format

CC-BY 4.0, CC-BY image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# What is the meaning of F A I R data ?

## R = Re-usable

Know *how I can* ***trust, repeat, re-analyse***, ***re-use*** the data.
Necessary to provide:



- ❑ Data **usage licence** --> full terms & conditions
- ❑ Data **provenance** --> metadata and information on:
  - ○ every data life-cycle stage
  - ○ documentation / protocols / references
  - ○ link to accompanying data and publications
  - ○ instruments & software used
- ❑ **Relationship** between the different levels of data you provide is documented:

  **raw**--> **quality controlled** -->**processed**-->**published**

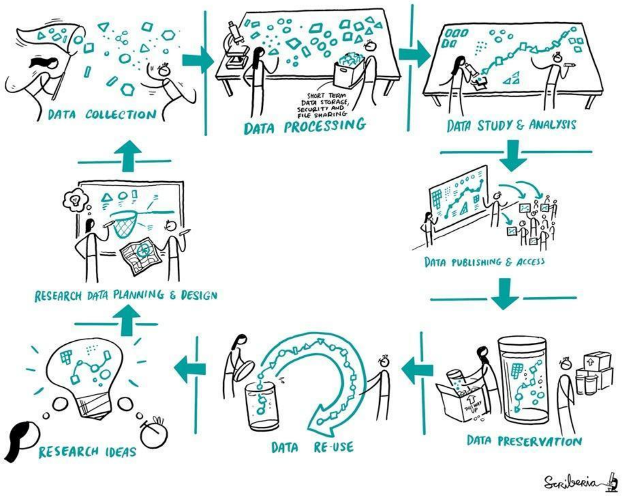
CC-BY 4.0, CC-BY image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# What is a data life-cycle?
# Data provenance

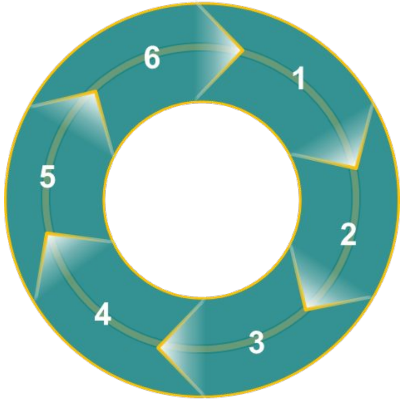
## Data Life-Cycle:

*- covers the entire period of time over which data exists*

*- encompasses all the stages: **first Capture** → **data re-use***



1. **Sample acquisition** → raw data:
   sample preparation, experimental settings / parameters, raw data acquisition
2. Data **quality control**: checking and updating of collected data
   documentation of QC procedures
3. Data **processing** & **analysis**: guided by scientific question
   documentation of processing steps, analysis methodology
4. **Archiving** & **publication**: data are placed in an online catalogue
   discovery metadata, provenance metadata, provenance files, references and links
5. Data **dissemination** / **integration**: adding data to well known portals, brokers
6. Data **reuse**: only possible with sufficient provenance information!



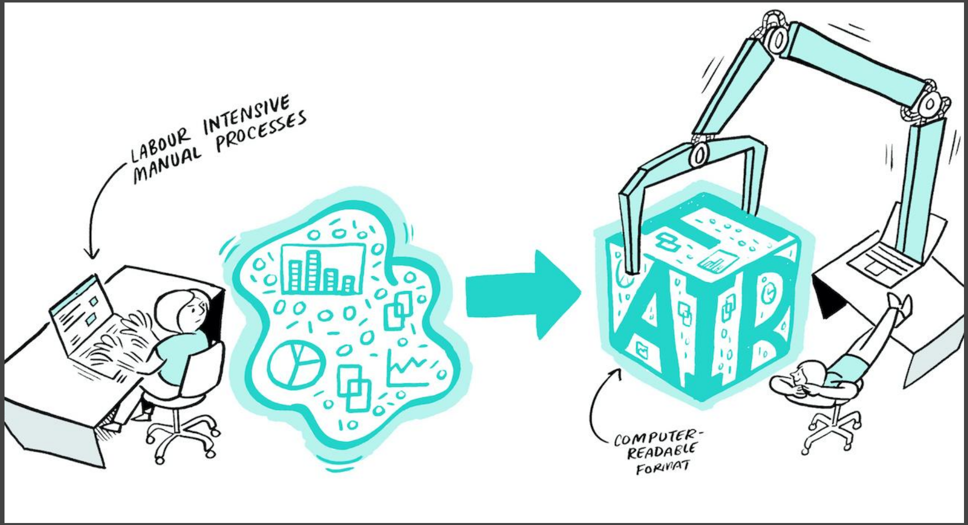
CC-BY 4.0, CC-BY  image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# Breakout Room: FAIR Reflection

*What part of your research lifecycle/outputs can be made open?*

*What you will have to do in your project to maintain "this" aspect of FAIR.*

- Room 1: Findable
- Room 2: Accessible
- Room 3: Interoperable
- Room 4: Reusable



CC-BY 4.0, CC-BY  image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# Reproducible research doesn’t always mean open

- Reproducibility can be facilitated by open, but open is a choice
- Reproducibility needs to be considered at all stages
- As open as possible, as closed as necessary
  - Open principles should be applied when you can
  - NEVER for private, confidential or sensitive data
- Always apply FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)

CC-BY 4.0, CC-BY image by *The Turing Way* and Scriberia, , DOI: 10.5281/zenodo.6337939

# Reproducibility

- Is my code **correct**?
- Can others **read and test** it?
- Is my **workflow robust**?
- Have I provided **guidance**?
- Is my work **citable**?

*Reproducibility should ensure higher scientific standards allowing others to test and reuse your work ...*

# Open Source

- Is my code **freely available**?
- Can others **modify and share** it?
- Is my **workflow reusable**?
- Have I provided **permission**?
- Is my work **open for collaboration**?

*… and Open Source should allow anyone to reuse, report errors, fix issues, build on and collaborate*

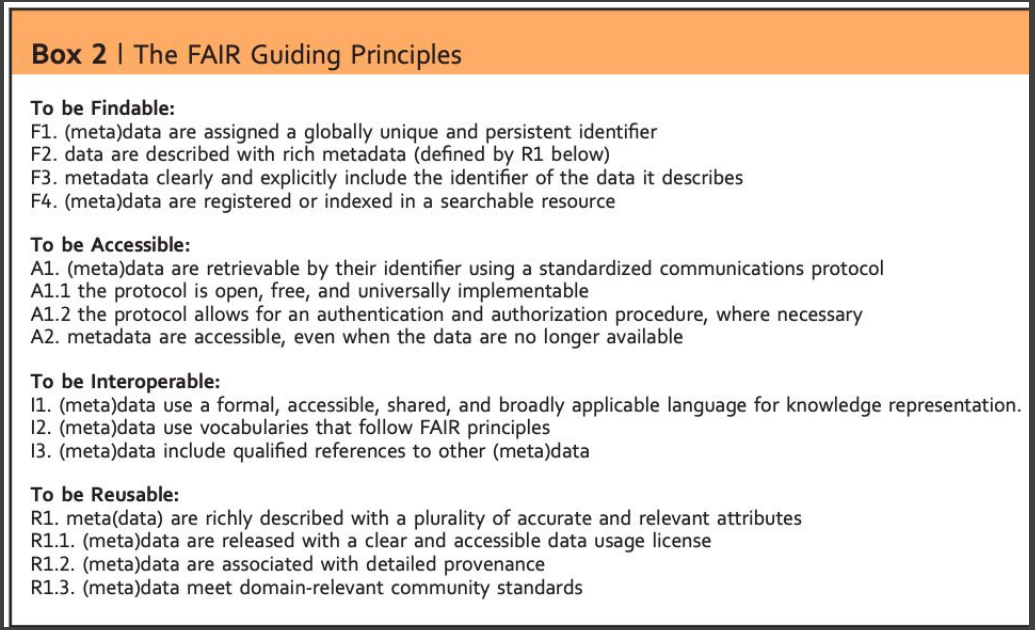
# FAIR doesn’t need to be open

- FAIR does not require data to be open

- FAIR requires open metadata

- Detailed information about research/data should be open

- FAIR applies open standards for interoperability



**Box 2 | The FAIR Guiding Principles**

**To be Findable:**
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

**To be Accessible:**
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

**To be Interoperable:**
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

**To be Reusable:**
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Metadata: information about the “data descriptors” that facilitate cataloguing data and data discovery

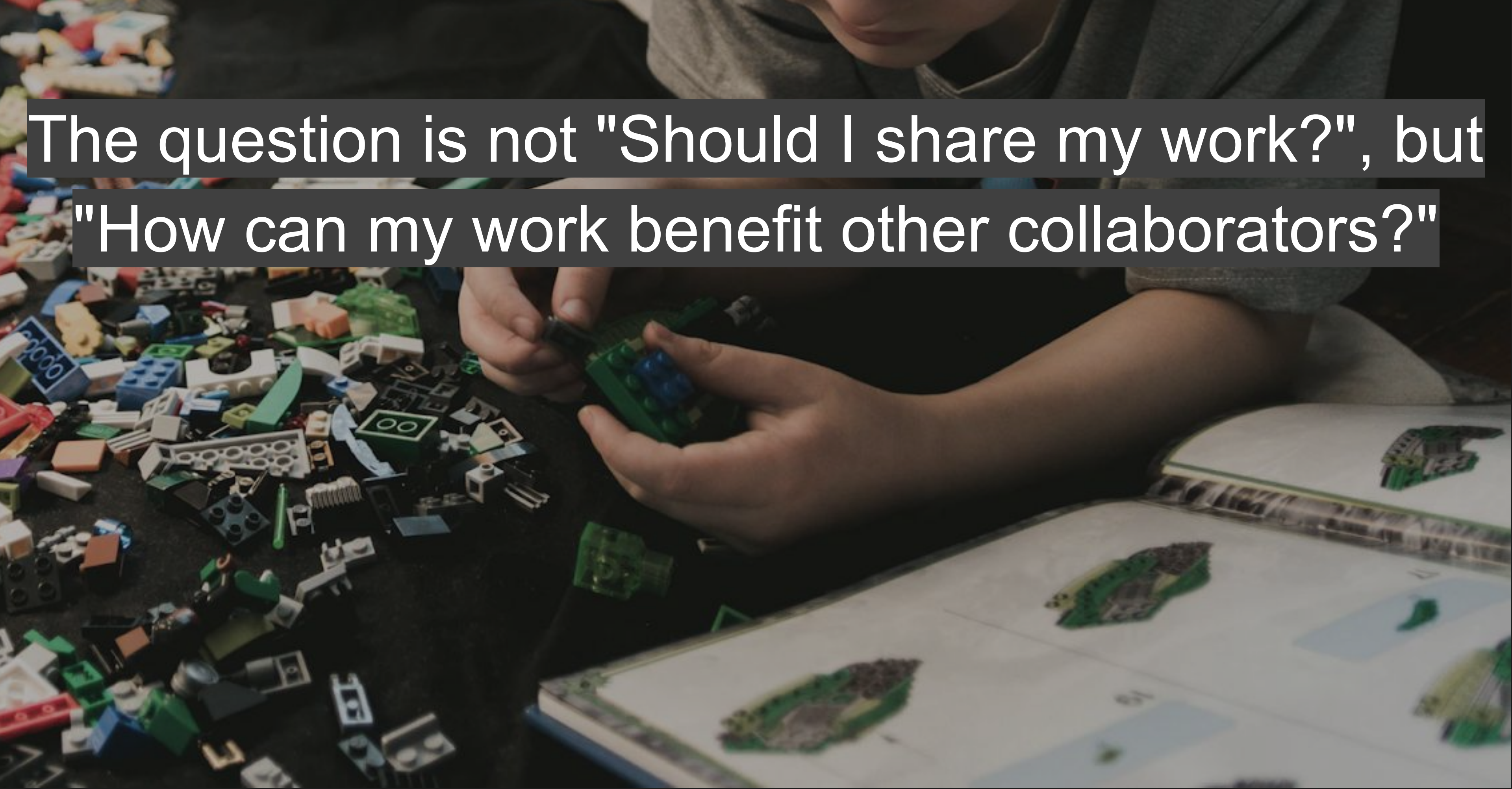
# Is this Reproducible, Open or FAIR?

- TARO is a PhD researcher in a health data science team
- Collaborates with people in their team of engineers and data scientists
- They have written Python code for a commonly used dataset
- Their supervisor suggested them to publish their work online
- Created a public repository to share their code, data and documentation
- Sent out an email to their team members to use their code

Image by *The Turing Way* and Scriberia, available under the CC-BY 4.0 license, DOI: 10.5281/zenodo.6337939 slido

# Is this Reproducible, Open or FAIR?

ⓘ Start presenting to display the poll results on this slide.



The question is not "Should I share my work?", but "How can my work benefit other collaborators?"

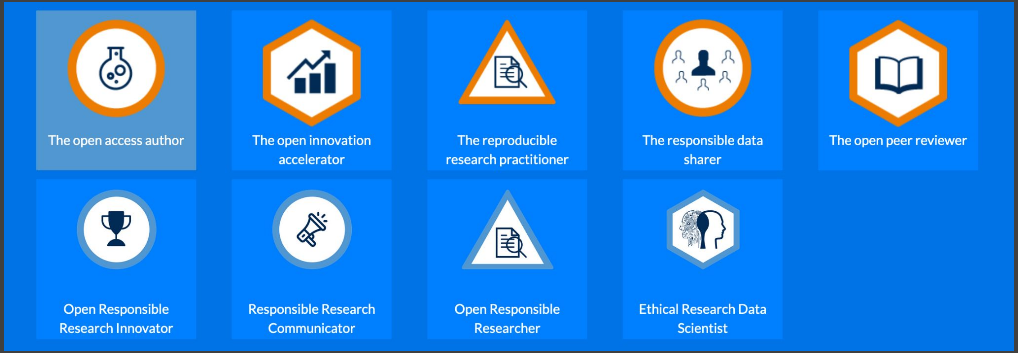
Image on Unsplash by @iankosky, @openlifesci, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

# You are your number one collaborator!

Good practices benefit (future) you!

Illustration by The Ludic Group LLP from Kirstie Whitaker's keynote presentation at Scientific Data in 2017. Used under a CC-BY 4.0 license. DOI: 10.6084/m9.figshare.5577340.v1, @malvikasharan, @turingway, DOI: 10.5281/zenodo.6337939

# Open Science path is self defined!



The open access author

The open innovation accelerator

The reproducible research practitioner

The responsible data sharer

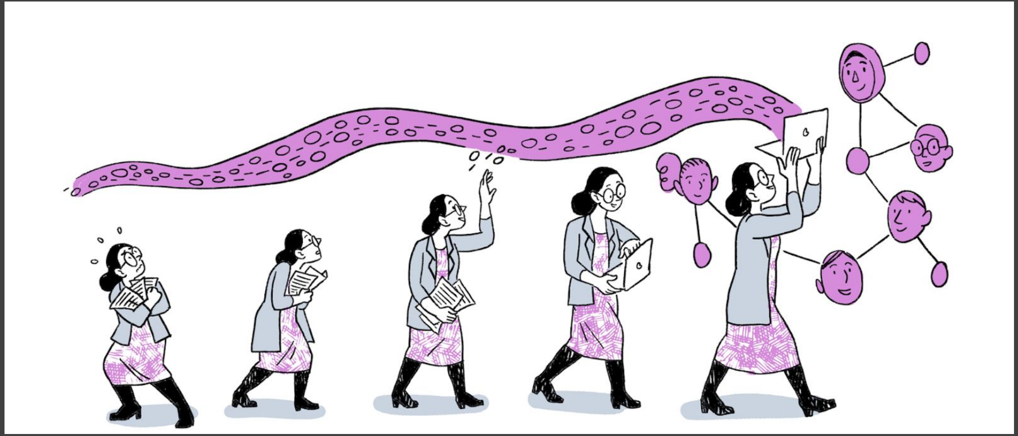
The open peer reviewer

Open Responsible Research Innovator

Responsible Research Communicator

Open Responsible Researcher

Ethical Research Data Scientist



- Reproducibility is essential.
- Open Science involves many concepts - take the steps that are possible.
- FAIR provides guidance for actionable steps that make your work reproducible.
- *Ask for feedback and help whenever you can.*

CC-BY image by The Turing Way and Scriberia, @turingway, https://www.fosteropenscience.eu/, DOI: 10.5281/zenodo.6337939

# Reproducible & Open Science Projects

Online Repository

License/Permission

Documentation

Version Control

Dependencies and data

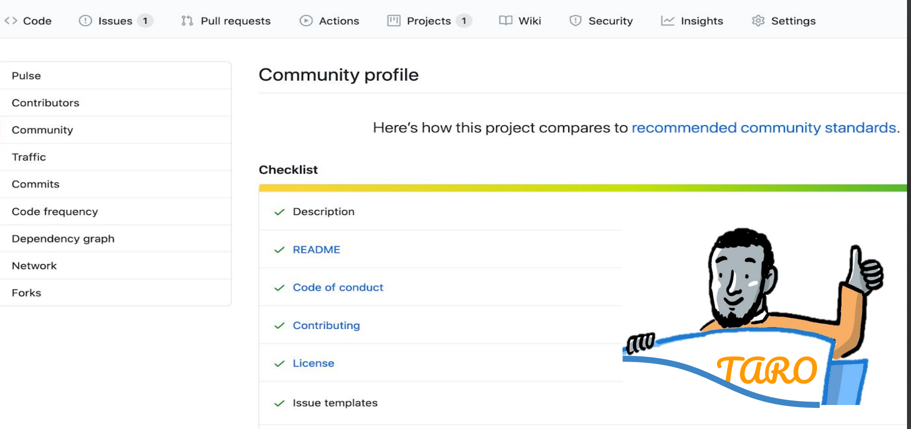
Review & Verify

Report/Fix Bugs

Release & Cite

@openlifesci, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

# Ensure Use and Reuse



Community profile

Here's how this project compares to recommended community standards.

Checklist

- ✓ Description
- ✓ README
- ✓ Code of conduct
- ✓ Contributing
- ✓ License
- ✓ Issue templates

- Create a project repository
- Create a README file with information on their scripts
- Add an open license in their repository for reuse
- Make it easy to test
- Add a minimum guideline for reporting errors
- Release citable versions of documentation/code/data

@openlifesci, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

# READMEs for Open and Collaborative Projects

**Motivation:**

Learn how to communicate your project effectively.

**Method**:

Write clear description of the project in README file.

# Why do READMEs matter?

# What is a README?

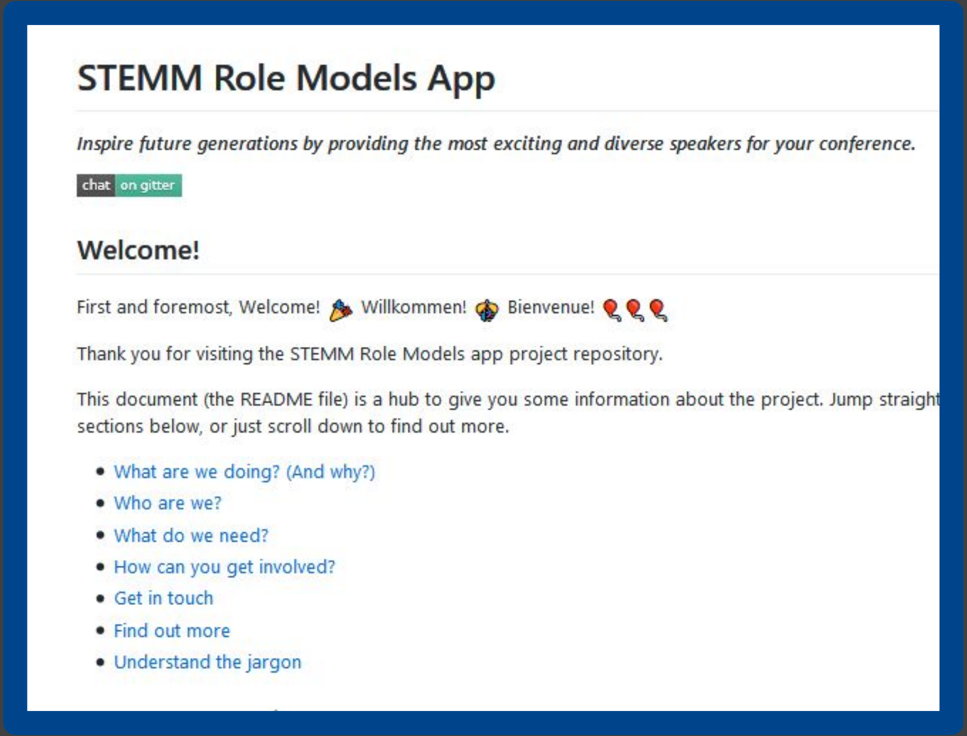
- Found in the root directory of your repository
- In ALL CAPS, a request for all to “read me!”
- First stop for your collaborators

*Could also be: website landing page, list of bullet points*

# In your README, show:

- what you're doing, what’s your motivation
- what makes your project special and exciting
- who your collaborators are, how do they contributors
- where are the key resources, where to find them

# A closer look at a README



## STEMM Role Models App

*Inspire future generations by providing the most exciting and diverse speakers for your conference.*

chat on gitter

### Welcome!

First and foremost, Welcome! 🎉 Willkommen! 🎊 Bienvenue! 🎈🎈🎈

Thank you for visiting the STEMM Role Models app project repository.

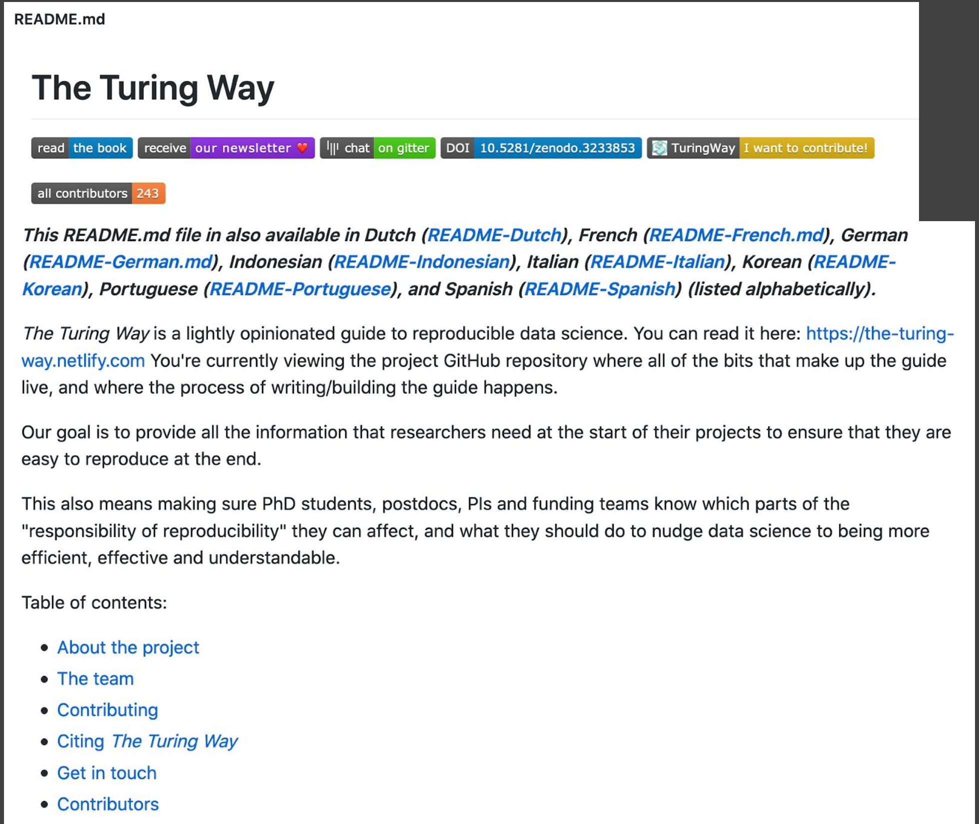
This document (the README file) is a hub to give you some information about the project. Jump straight sections below, or just scroll down to find out more.

- What are we doing? (And why?)
- Who are we?
- What do we need?
- How can you get involved?
- Get in touch
- Find out more
- Understand the jargon

README example: STEMM Role Models App

- Welcome message!
- Project description & vision
- How to:
  - Test and verify
  - Fix errors
  - use (license)
  - Get involved
  - Report issues

# README = Project Documentation



README.md

## The Turing Way

*This README.md file in also available in Dutch (README-Dutch), French (README-French.md), German (README-German.md), Indonesian (README-Indonesian), Italian (README-Italian), Korean (README-Korean), Portuguese (README-Portuguese), and Spanish (README-Spanish) (listed alphabetically).*

*The Turing Way* is a lightly opinionated guide to reproducible data science. You can read it here: https://the-turing-way.netlify.com You're currently viewing the project GitHub repository where all of the bits that make up the guide live, and where the process of writing/building the guide happens.

Our goal is to provide all the information that researchers need at the start of their projects to ensure that they are easy to reproduce at the end.

This also means making sure PhD students, postdocs, PIs and funding teams know which parts of the "responsibility of reproducibility" they can affect, and what they should do to nudge data science to being more efficient, effective and understandable.

Table of contents:

- About the project
- The team
- Contributing
- Citing *The Turing Way*
- Get in touch
- Contributors

- project description and features
- installation instructions
- tutorials and requirements
- how to run associated tests
- list of authors/contributors
- contact information
- links to related material

# Assignment: Create a project repository

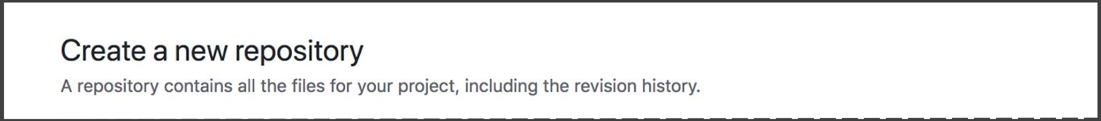
Add top-level files: README and LICENSE

- README to communicate about your work
- Also try to add a License -  to allow others to use, modify, build upon your work
  - We will cover licenses in the next session in more detail.

Use The Turing Way chapter for README to guide your assignment

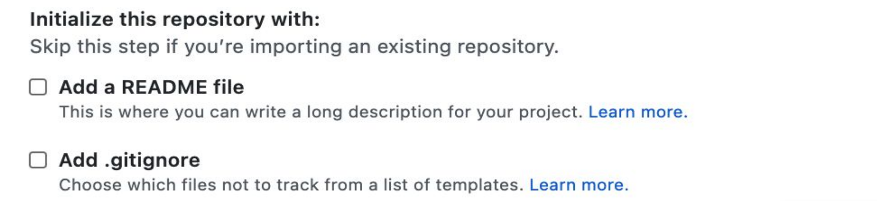
https://the-turing-way.netlify.app/project-design/project-repo/project-repo-readme.html

# Create a GitHub Repository with README file



Create a new repository

A repository contains all the files for your project, including the revision history.



**Initialize this repository with:**

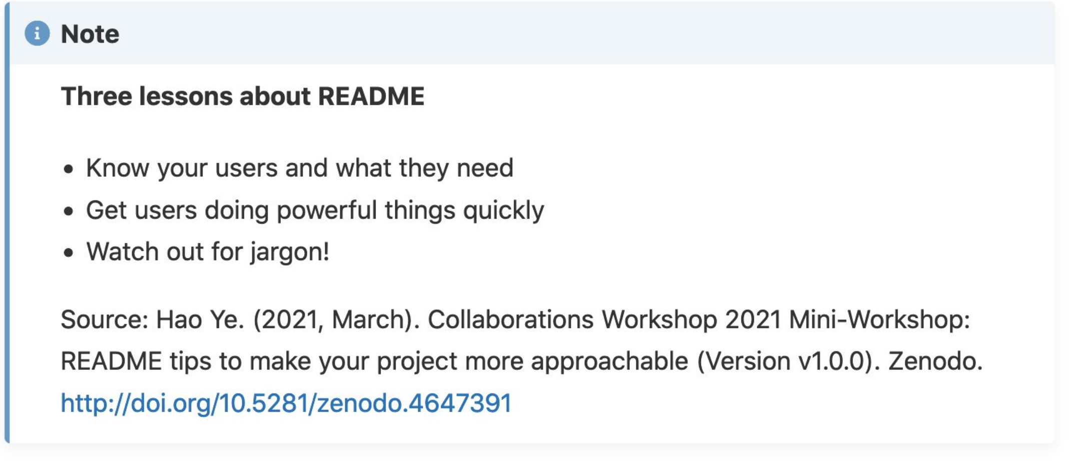
Skip this step if you're importing an existing repository.

☐ **Add a README file**

This is where you can write a long description for your project. Learn more.

☐ **Add .gitignore**

Choose which files not to track from a list of templates. Learn more.



**Note**

**Three lessons about README**

- Know your users and what they need
- Get users doing powerful things quickly
- Watch out for jargon!

Source: Hao Ye. (2021, March). Collaborations Workshop 2021 Mini-Workshop: README tips to make your project more approachable (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.4647391

# End of Part 1: Further Reading and Examples

- *The Turing Way* README
- Open Life Science README
- Purple Booth’s README Template
- Thoughtbot’s Blog on How to Write a Good README
- Matias Singer’s curated List of Awesome READMES

@openlifesci, CC-BY 4.0, The Turing Way, DOI: 10.5281/zenodo.6337939

# Assignment: Create a project repository

Add top-level files: README and LICENSE

- README to communicate about your work
- Also try to add a License -  to allow others to use, modify, build upon your work
  - We will cover licenses in the next session in more detail.

Use The Turing Way chapter for README to guide your assignment

https://the-turing-way.netlify.app/project-design/project-repo/project-repo-readme.html

The Alan Turing Institute

## End of Session 1

Several slides in this presentation uses *The Turing Way* and Open Life Science contents shared under CC-BY 4.0 License.