Open Science: Reproducibility and Interoperability in Research

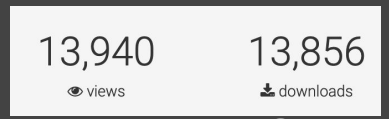
Sarah Gibson (she/her)
2i2c / JupyterHub

Acknowledgements & Links

- Malvika Sharan, *The Turing Way* Co-Lead
- *The Turing Way* community members
- Book: the-turing-way.netlify.app
- Twitter: twitter.com/turingway
- GitHub: github.com/alan-turing-institute/the-turing-way

Useful links & opportunities are listed here: https://bit.ly/turingway

Illustrations by Scriberia for The Turing Way community:
https://zenodo.org/record/3332807



- Core collaborator with *The Turing Way* (2018-)
- Research Software Engineer, Turing's Research Engineering team (2018-2021)
- Open Source Infrastructure Engineer, 2i2c (2021-)
- JupyterHub and mybinder.org team member (2019-)
- Jupyter Distinguished Contributor (2021-)
- JupyterHub Community Strategic Lead (2022-)
- Community Building in Open Source Software

2i2c

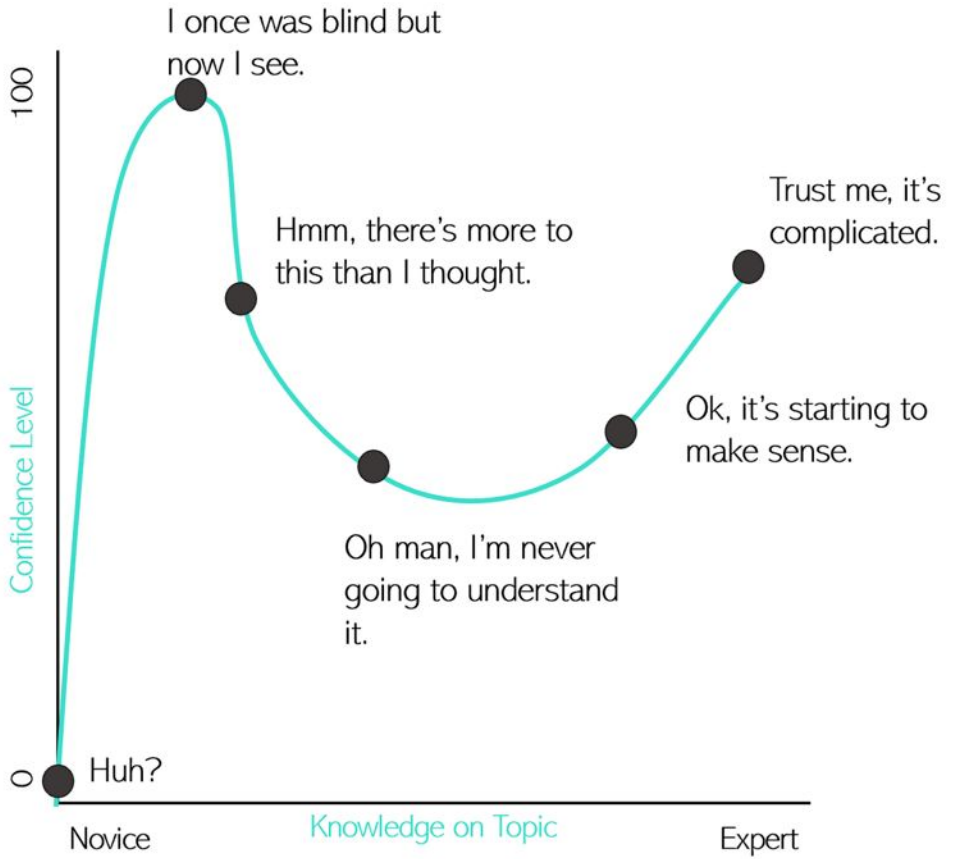
jupyterhub

Software Sustainability Institute

Open Life Science

Disclaimer:

You probably already know all about it!



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

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

Contents

1. Reproducibility in Research
2. Interoperability and Decentralisation of Systems
3. Research Infrastructure Roles
4. *The Turing Way*

Research Reproducibility

Open and reproducible research saves valuable time in verifying and building upon widely beneficial solutions.

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)	

BBC Sign in News Sport Weather iPlayer Sounds

NEWS

Home UK World Business Politics Tech Science Health Family & Education

Magazine

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

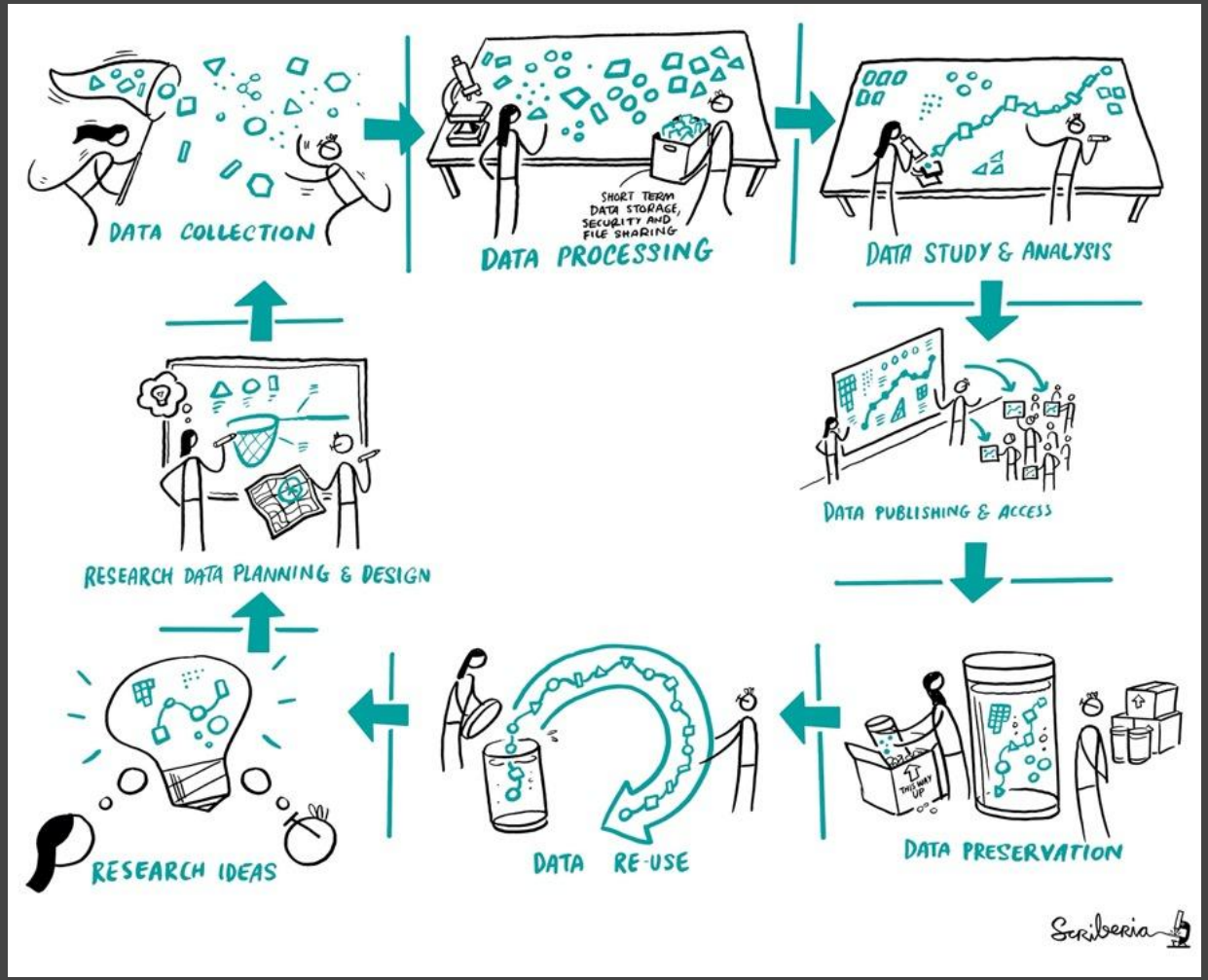
By Ruth Alexander
BBC News

20 April 2013 Share

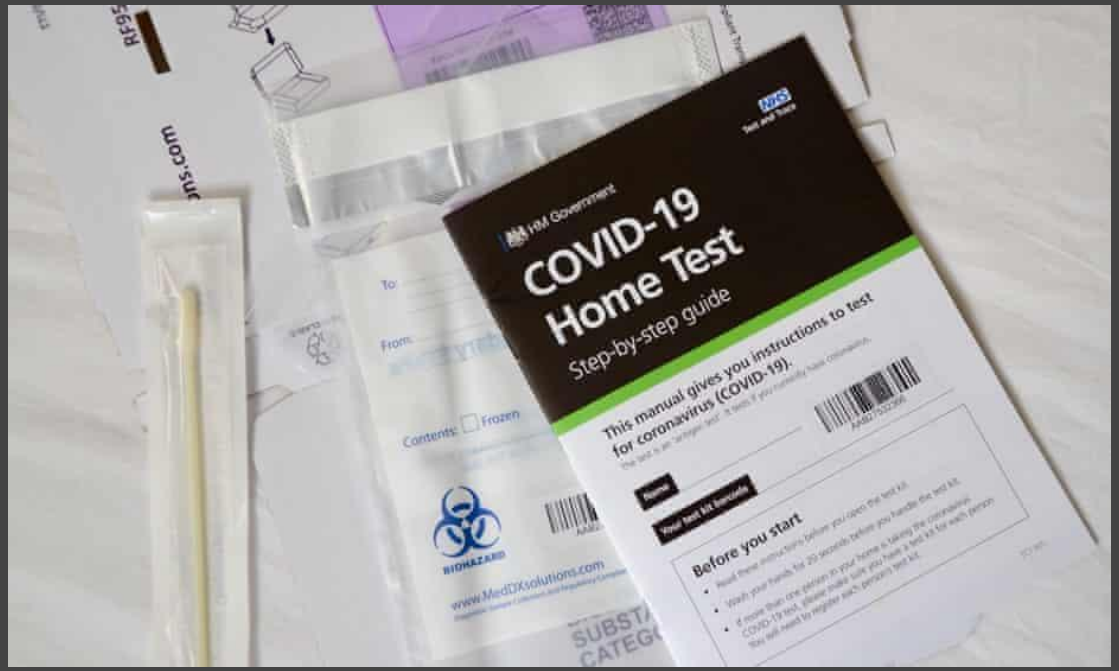
This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

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



Under-reported figures

From 25 Sept to 2 Oct

50,786

Cases initially reported by PHE

15,841

Unreported cases, missed due to IT error

8 days of incomplete data

1,980 cases per day, on average, were missed in that time

48 hours Ideal time limit for tracing contacts after positive test

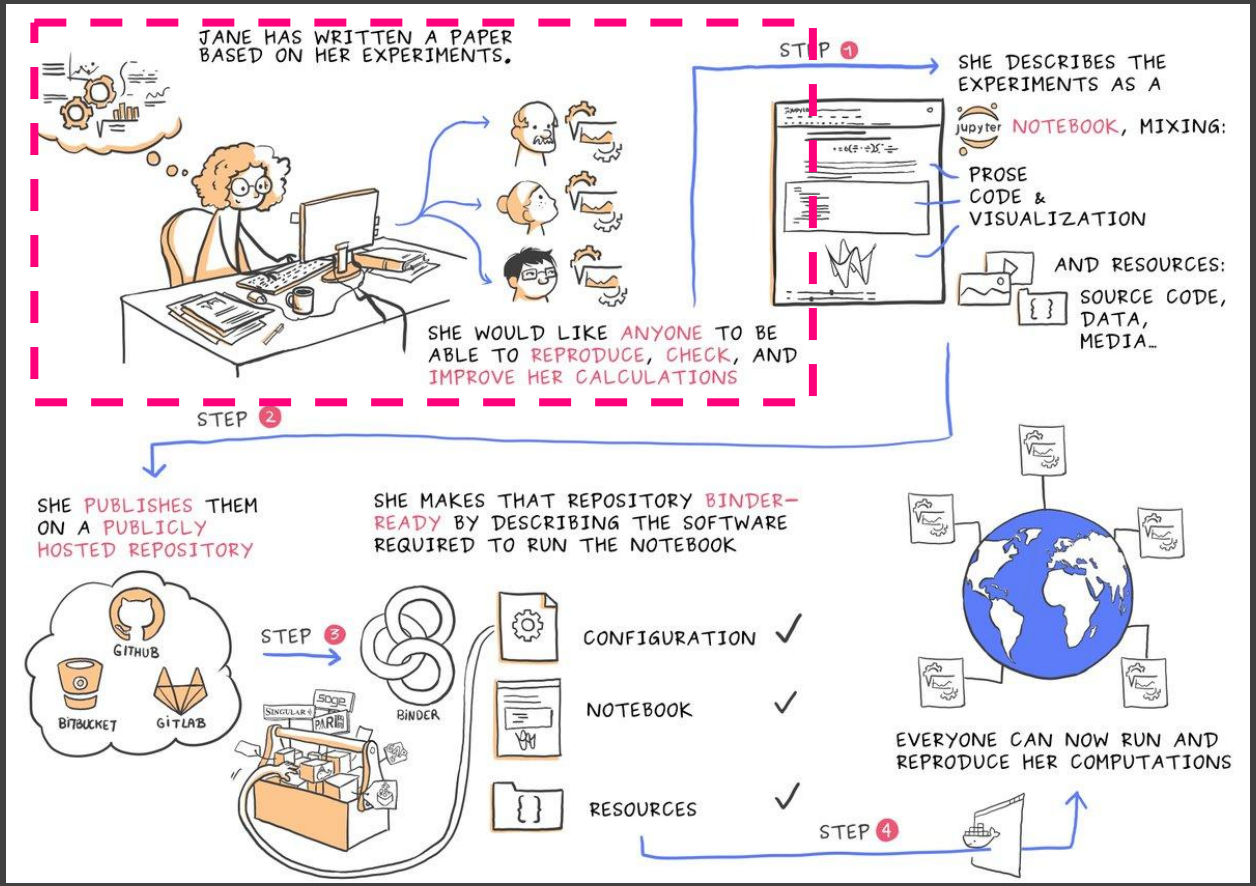
Source: PHE and gov.uk

Hern, A. (2020). Covid: how Excel may have caused loss of 16,000 test results in England. the Guardian. https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-englandhttps://www.bbc.com/news/technology-54423988

DOI: 10.5281/zenodo.7339751

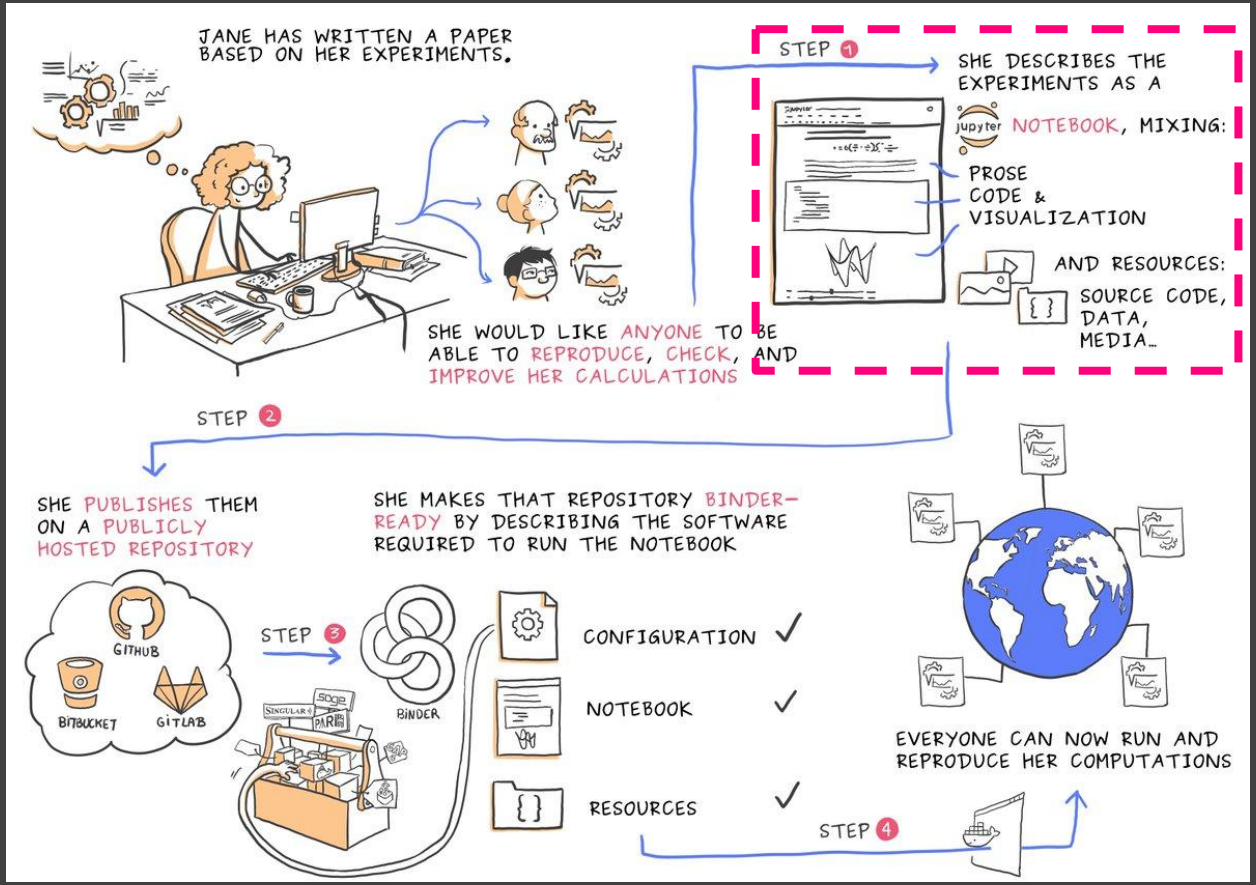
Two main dependencies of research

Project Binder for reproducing code



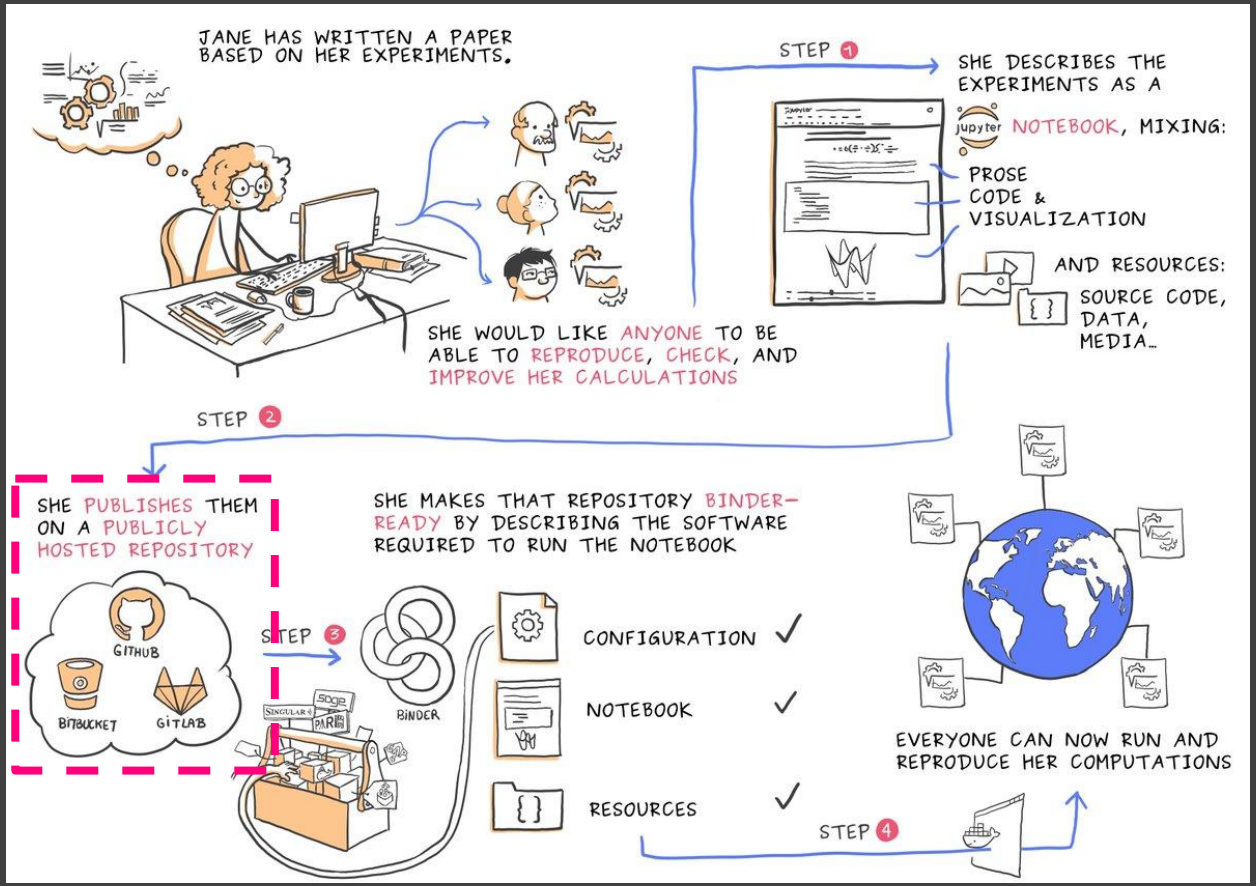
Publishing reproducible logbooks explainer comic strip, Juliette Taka and Nicolas M. Thiéry, DOI: 10.5281/zenodo.4421040

Project Binder for reproducing code



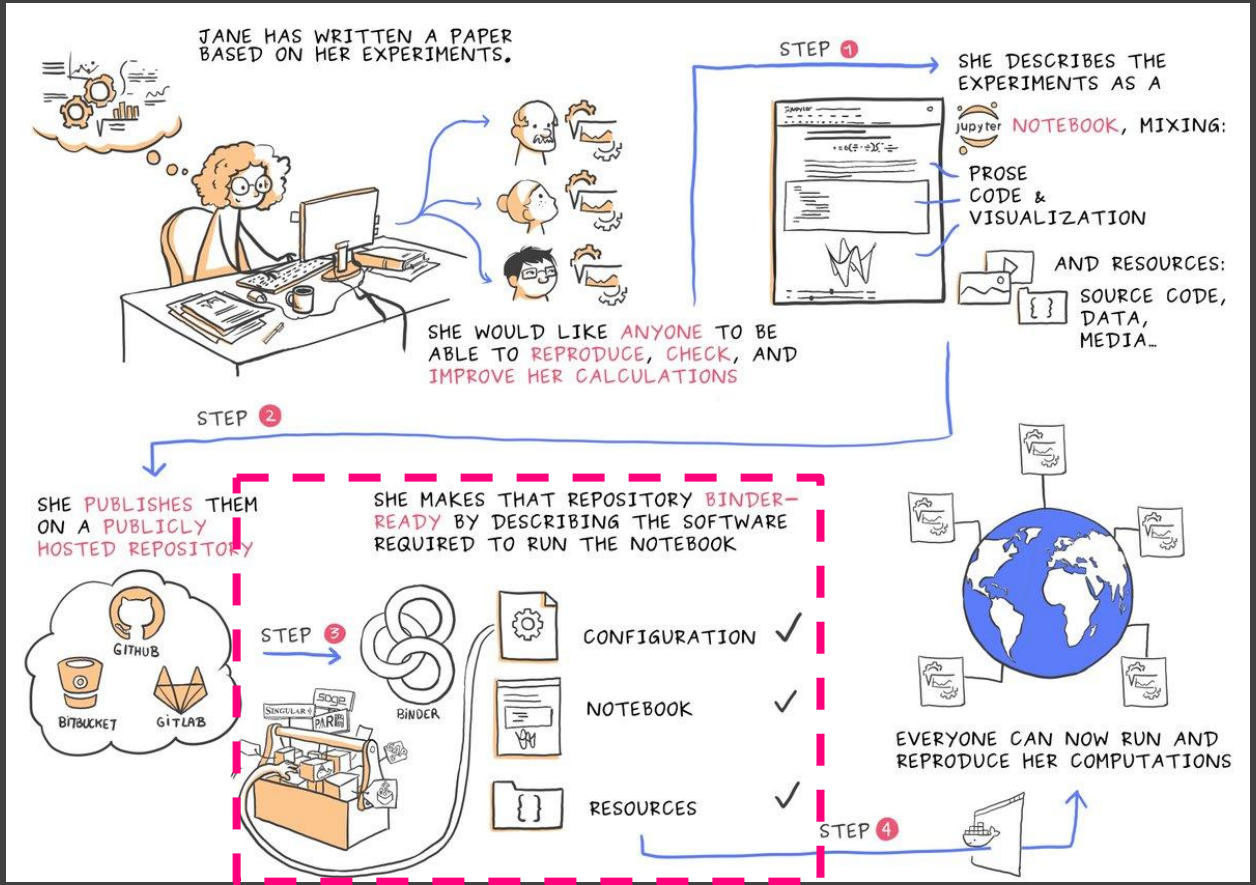
Publishing reproducible logbooks explainer comic strip, Juliette Taka and Nicolas M. Thiéry, DOI: 10.5281/zenodo.4421040

Project Binder for reproducing code



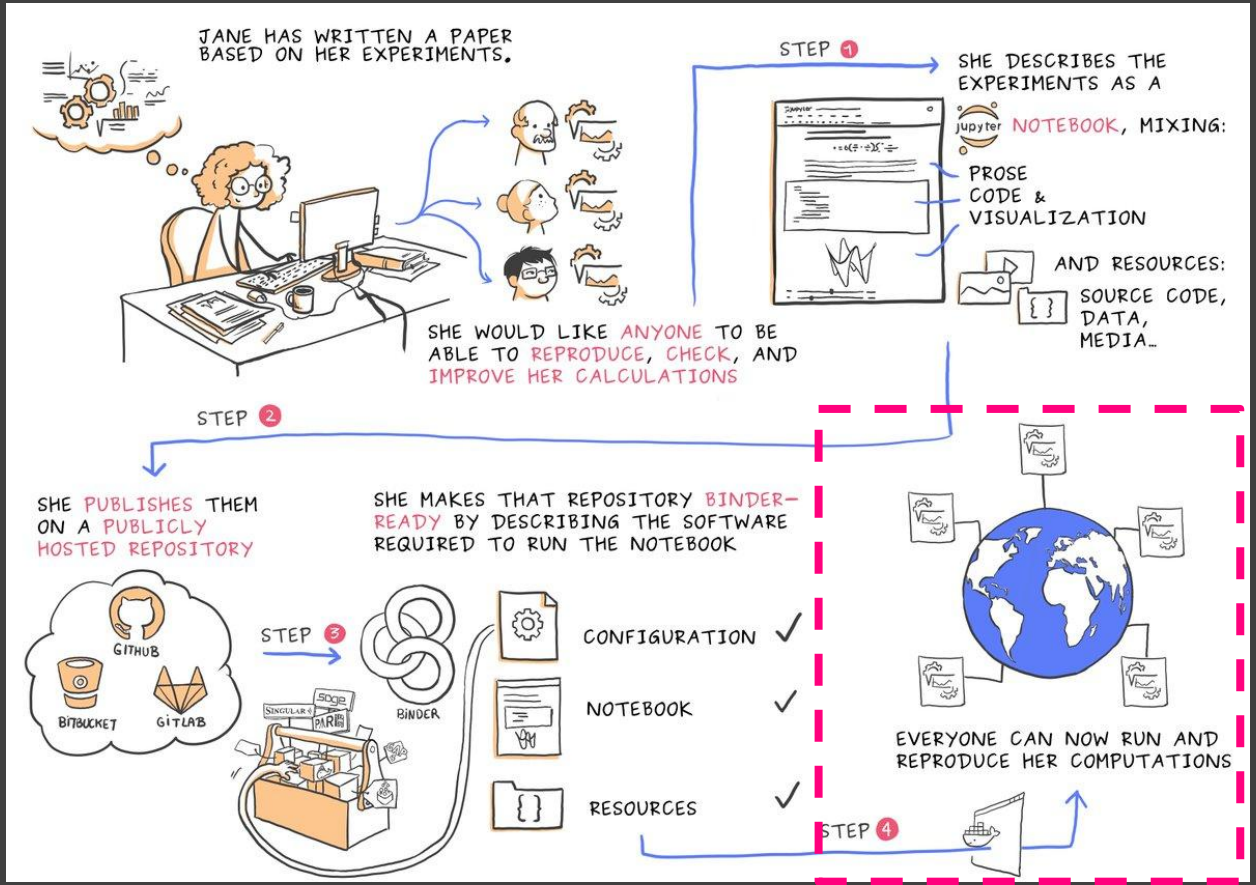
Publishing reproducible logbooks explainer comic strip, Juliette Taka and Nicolas M. Thiéry, DOI: 10.5281/zenodo.4421040

Project Binder for reproducing code



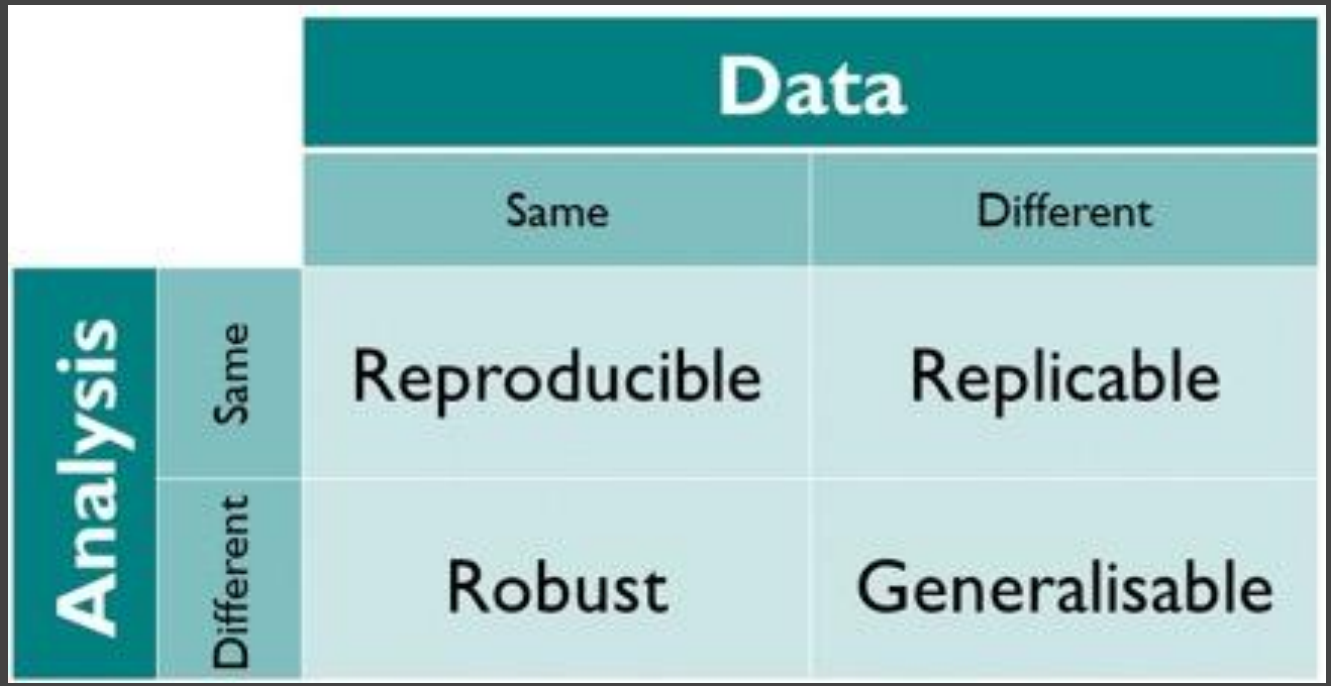
Publishing reproducible logbooks explainer comic strip, Juliette Taka and Nicolas M. Thiéry, DOI: 10.5281/zenodo.4421040

Project Binder for reproducing code



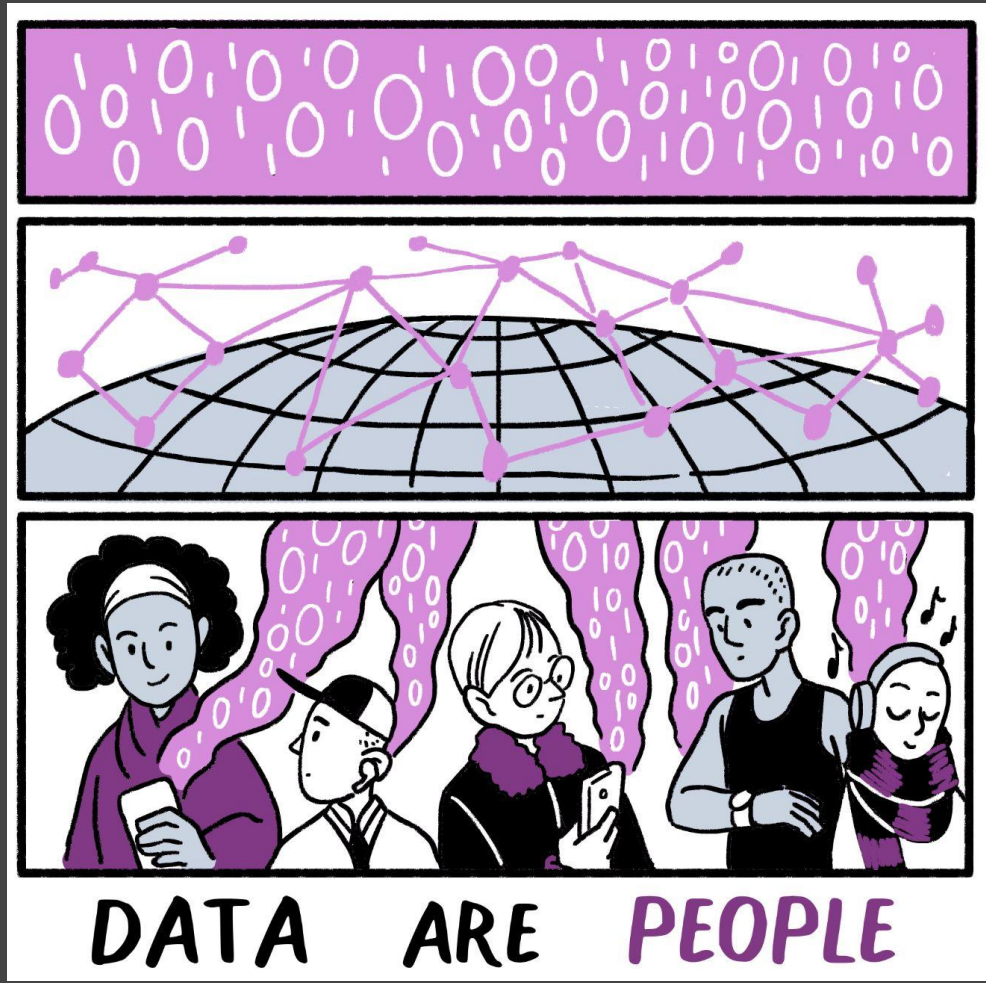
Publishing reproducible logbooks explainer comic strip, Juliette Taka and Nicolas M. Thiéry, DOI: 10.5281/zenodo.4421040

Definitions



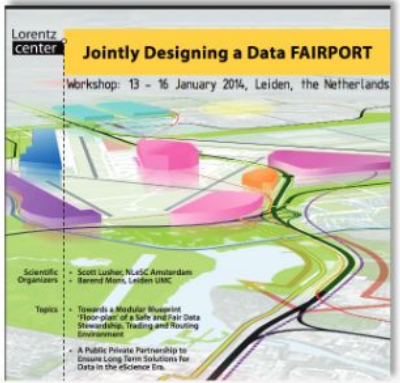
Synthetic data

- A minimalist dataset
- Anonymised / pseudonymised
- Allows validation of results and building upon the work done



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
    - open principles should applied when you can
    - NEVER for private, confidential or sensitive data
- Always apply FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable)



2014

2016

SCIENTIFIC DATA

SPRINGER NATURE

OPEN

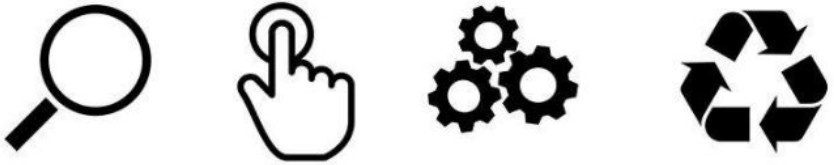
SUBJECT CATEGORIES
» Research data
» Publication characteristics

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 **A**ccessible **I**nteroperable **R**eusable



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: *You would not buy food with no labels!*

Annotation makes it easier to find important things

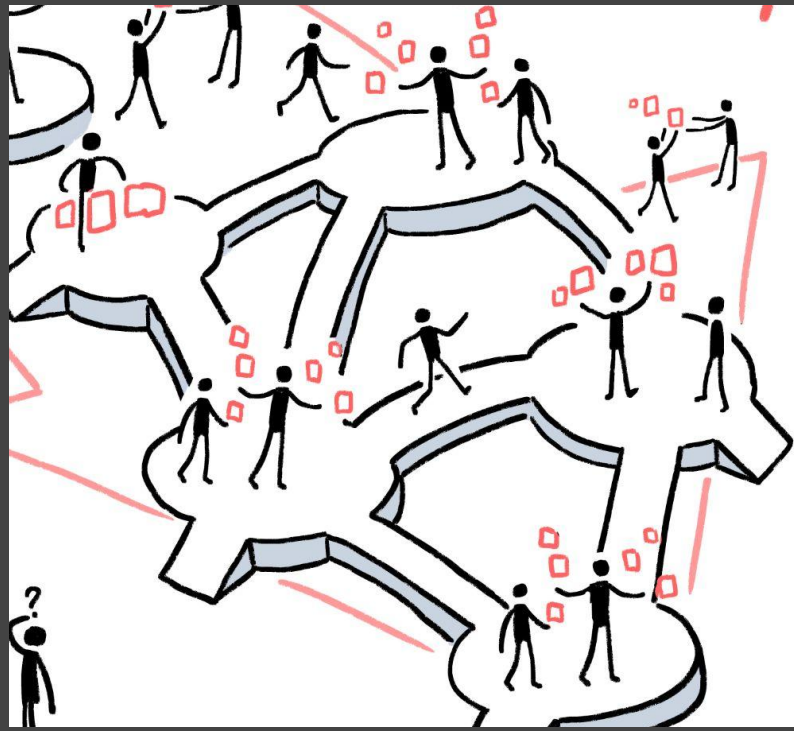
FAIR doesn't require data to be open, but needs Metadata information along with detailed research process.

Metadata: "data descriptors" that facilitate cataloguing data and data discovery

Adapted from talk by Philippe Rocca-Serra (2020)

Interoperability and Decentralisation of Systems

The backbone of collaborative research



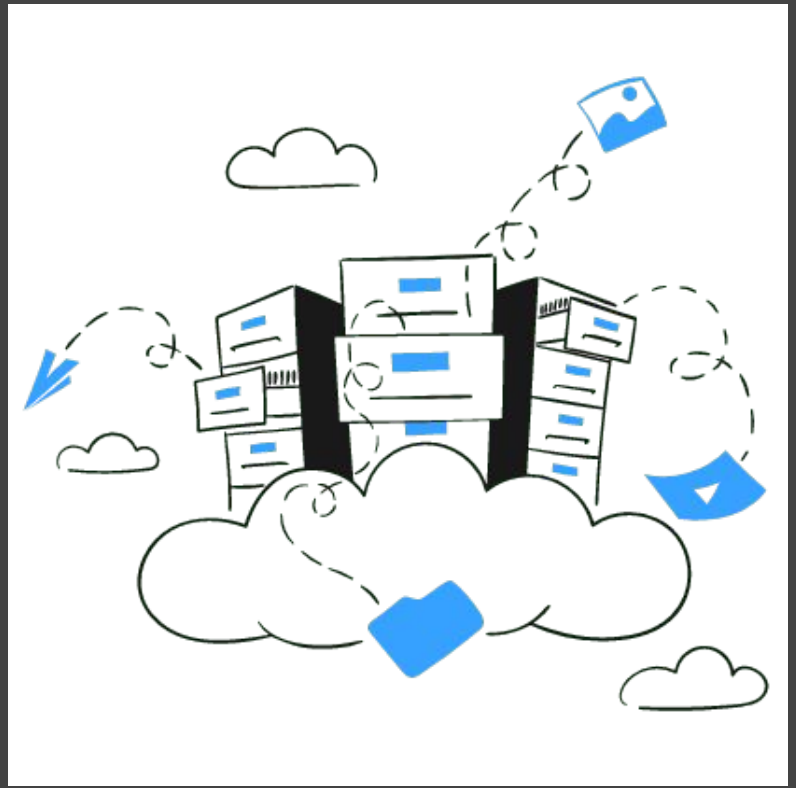
Our requirements for computational research have changed

We are beyond capabilities of local machines

- Ever larger datasets
- More computational resources required to process and analyse data

Require dedicated platforms

- High performance computing
- Cloud



Who governs the infrastructure we use?

- Designing, deploying, maintaining, operating infrastructure for large-scale collaborative research requires a specific skill set
    - Often not present or incentivised on research project teams
- Rely on private companies for this infrastructure
- Motivated by financial stakeholders - not by public interest
- Locks the research in to using a specific service or vendor
- Erodes public trust

Controversial £360m NHS England data platform 'lined up' for Trump backer's firm

Patients will have no say over records going to Palantir, the software giant run by billionaire Republican backer

Prof. James Hetherington @jamespjh@mastodon.me.uk
@jamespjh

For those of us work on safe use of NHS data to save lives through research, this is a disaster for public trust. Research w personal data must be on nonprofit open public infrastructure. Let's work to fix this. @HDR_UK @turinginst There are alternatives github.com/alan-turing-in...

Ungoed-Thomas, J. (2022) Controversial £360m NHS England data platform 'lined up' for Trump backer's firm. The Guardian https://www.theguardian.com/society/2022/nov/13/controversial-360m-nhs-england-data-platform-lined-up-for-trump-backers-firm

Unburdening researchers from infrastructure management

2i2c

Our mission is to make research and education more **impactful**, **accessible**, and **delightful** by developing, operating, and supporting infrastructure for interactive computing.

- Experts in cloud-based infrastructure operations using open source technologies
  - Specifically the Jupyter stack
- Provide managed services to scientific research and education communities globally
- Refine common scientific workflows in the cloud and support communities implementing them

Maintaining researcher ownership of infrastructure

2i2c

Healthcare equivalents?

- Turing Data Safe Haven
- …

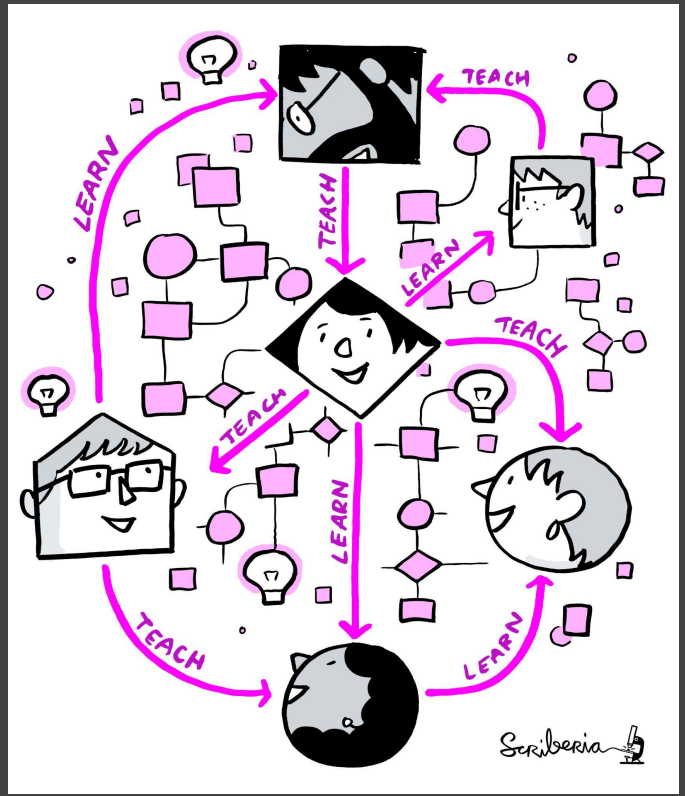
- Not for profit - our stakeholders are the communities we serve
- All developments are contributed back upstream to the OSS projects we rely upon
- Right to Replicate - no vendor lock-in, can take your hub and go at any time

(Definitions of MUST, MUST NOT, SHOULD, MAY, etc are defined in RFC 2119)

User Code and Data	**May** be Open Source	We encourage adopting and producing open source code and data, but this is up to the user. e.g., licenses for user content/code
User Environment	**Should** be Open Source	Strong preference for open source tools only, although in some cases user needs may override this. e.g., Python, R, PyData stack.
2i2c Infrastructure	**Must** be Open Source	Strong commitment to using only open source software. e.g., JupyterHub, Kubernetes, Postgresql
Cloud Provider Infrastructure	**Must** be Portable	See this blog post for more information.

Research Infrastructure Roles

Software is only 10% of the problem - the rest is people



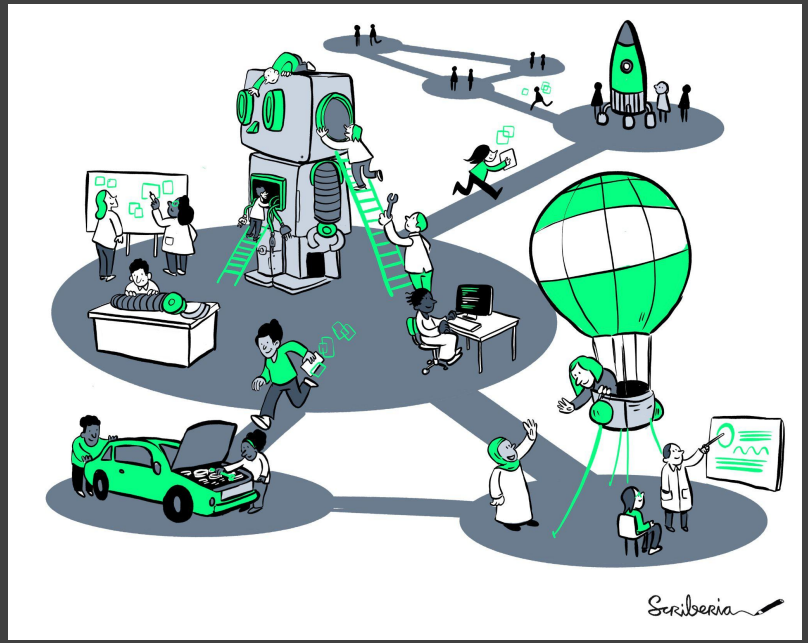
What other support do you need to be open?

- Community manager
- Project manager
- Research software engineer
- Research application manager
- Programme manager
- Data steward
- Data wrangler
- Research infrastructure developer
- Partnerships development manager
- Grant application manager
- Community engagement lead
- Librarian
- Communications manager
- Events manager
- Participatory research manager

jupyterhub

OUTREACHY

There is more to contributing than software engineering



Applying ‘best practices’ for reproducible and collaborative research requires intention, resources, time and training *which can be overwhelming*.

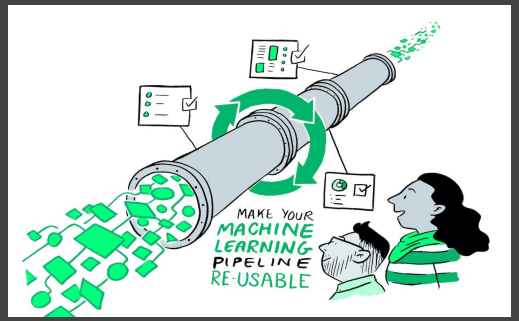
Blog post: https://www.software.ac.uk/blog/2020-12-17-ten-arguments-against-open-science-you-can-win

The Turing Way

An open source, community-led guide on Data Science.

*We involve and support a **diverse community** to make research **reproducible, ethical, open, and inclusive** for everyone.*

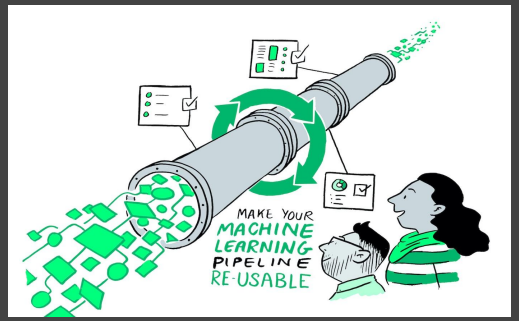
The Turing Way Guide



Reproducibility

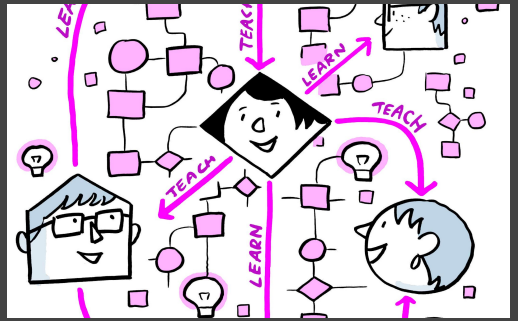
@drsarahlgibson, @turingway, CC-BY 4.0, DOI: 10.5281/zenodo.7339751

The Turing Way Guides

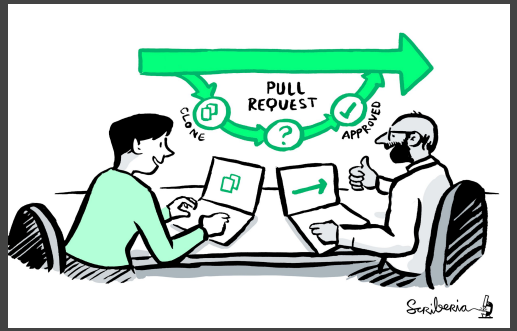


Reproducibility

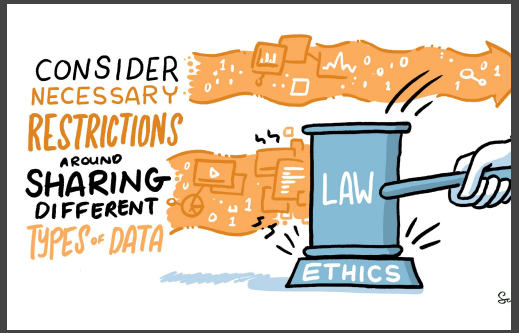
Project design



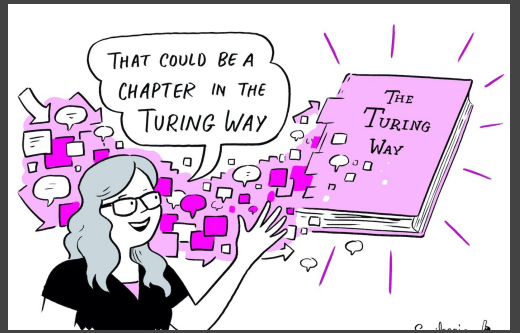
Communication



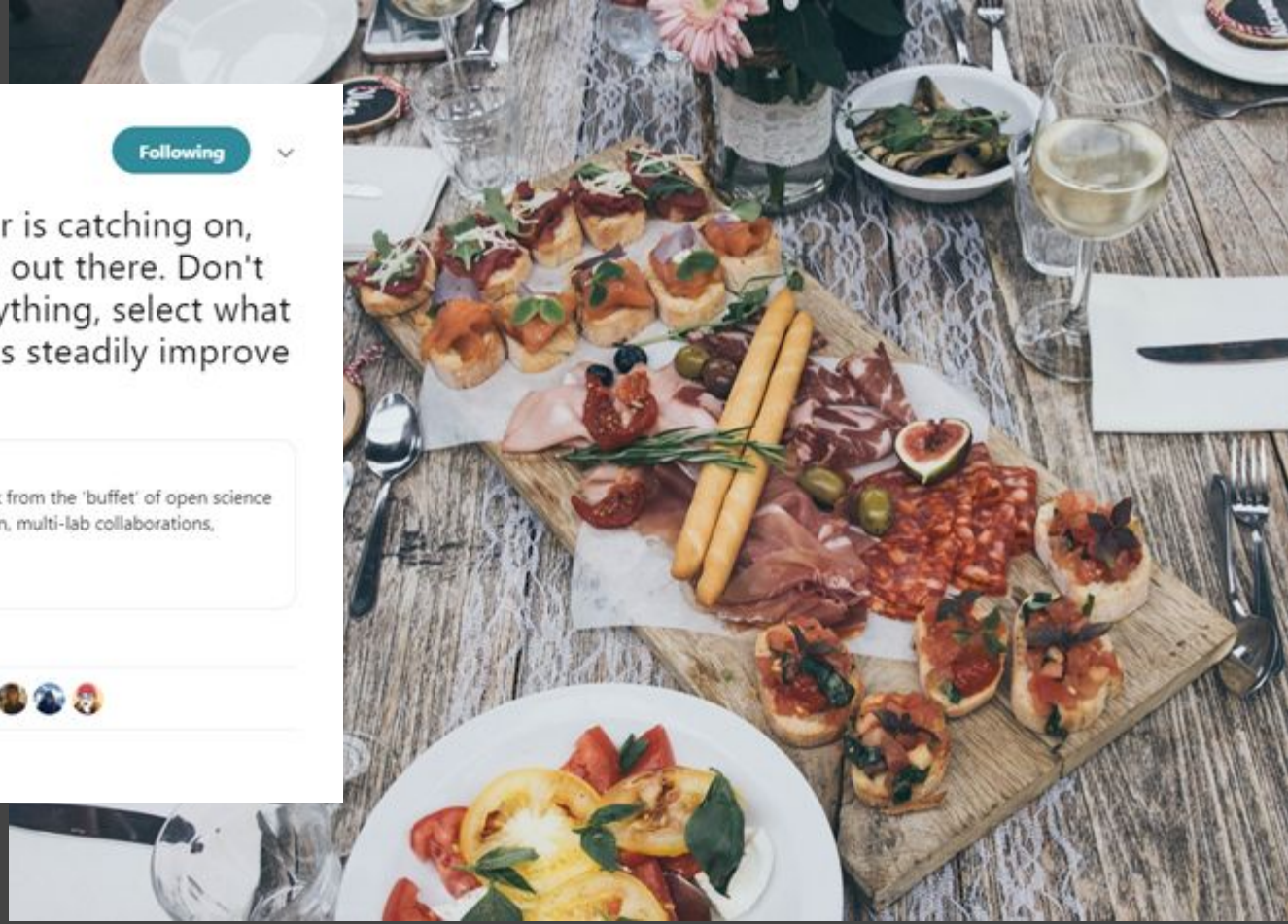
Collaboration



Ethical research



Community handbook



Dr. Christina Bergmann
@chbergma

Following

So glad the buffet metaphor is catching on, there are so many solutions out there. Don't try to stuff yourself on everything, select what works for this study and let's steadily improve our fields... #openscience

> **Priya Silverstein** @priyasilverst
> Lastly, @MicheleNuijten wrapping up: take your pick from the 'buffet' of open science practices from transparency, statistics, preregistration, multi-lab collaborations, attending @improvingpsych meeting, etc!
> Show this thread

11:11 AM - 9 Mar 2019

10 Retweets **28** Likes

1 10 28

https://twitter.com/chbergma/status/1104338904646385665; @turingway, DOI: 10.5281/zenodo.4609987

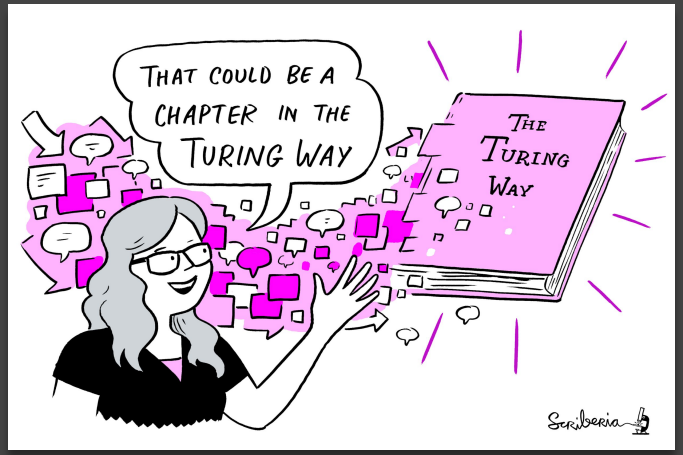
Book:
the- turing-way.netlify.app/

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

Twitter:
twitter.com/turingway

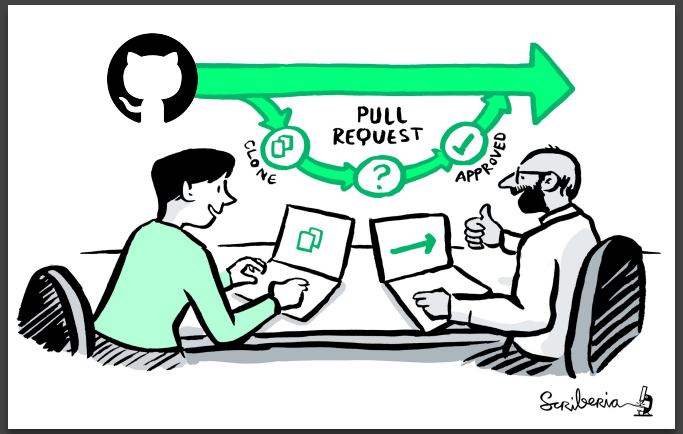
Email:
theturingway@gmail.com

A Book



A Community

An Open Source Project



A Culture of Collaboration

Pathways for Collaboration

Develop & share

Maintain & improve

Share resources

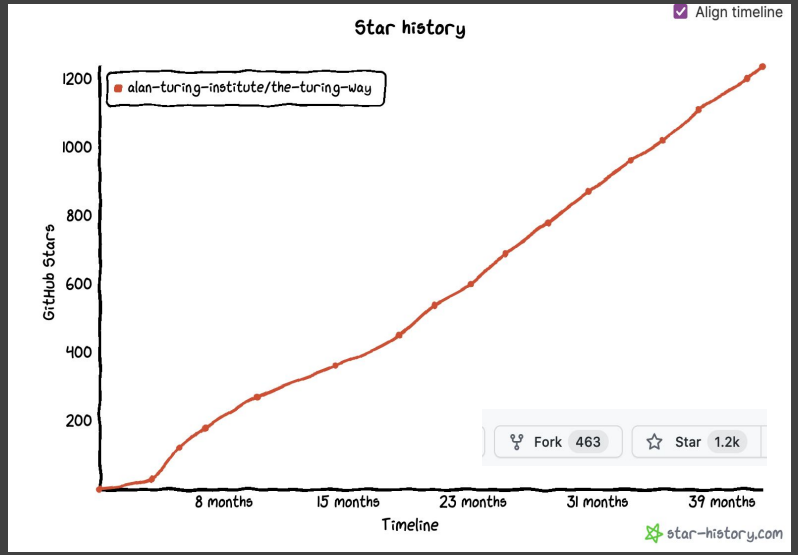
Review and update

Make it global

Share best practices

Project and community growth

- 3 years, >250 Live Chapters
- Community resources, events, guidance, templates, training
- 400+ direct GitHub contributors and thousands of users



https://zenodo.org/record/3332807

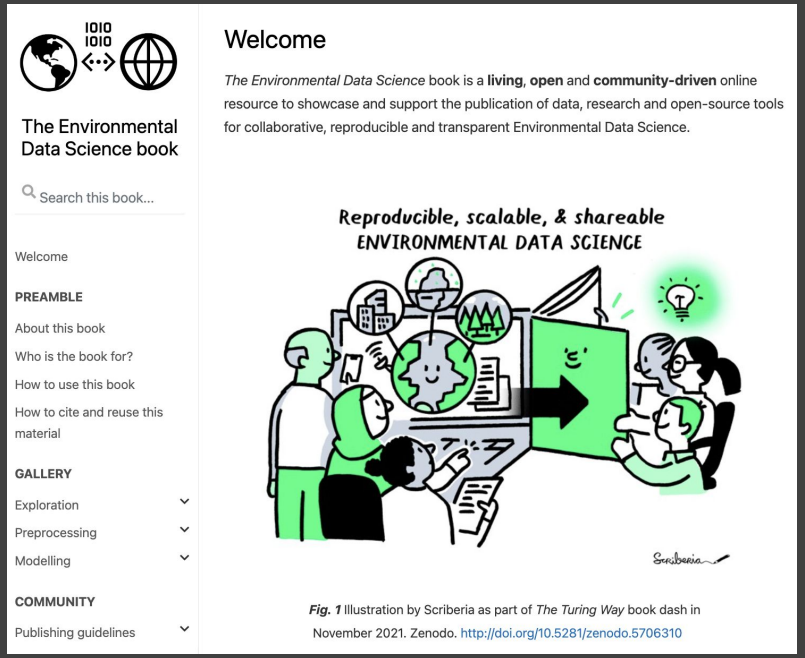
zenodo

13,940 views

13,856 downloads

Open Science in other domains: Environmental Data Science Book

- Supports publication of data, research, and open source tools for collaborative, reproducible, and transparent environmental science
- Community-based model replicated from *The Turing Way*



Take home messages

- Build a community that appreciates the value of decentralisation
- Redeployable, transferable, interoperable, reproducible
- What does a Healthcare Data Science book based on *The Turing Way* look like?

Acknowledgements & Links

- Malvika Sharan, *The Turing Way* Co-Lead
- *The Turing Way* community members
- Book: the-turing-way.netlify.app
- Twitter: twitter.com/turingway
- GitHub: github.com/alan-turing-institute/the-turing-way

Useful links & opportunities are listed here: https://bit.ly/turingway

Illustrations by Scriberia for The Turing Way community:
https://zenodo.org/record/3332807