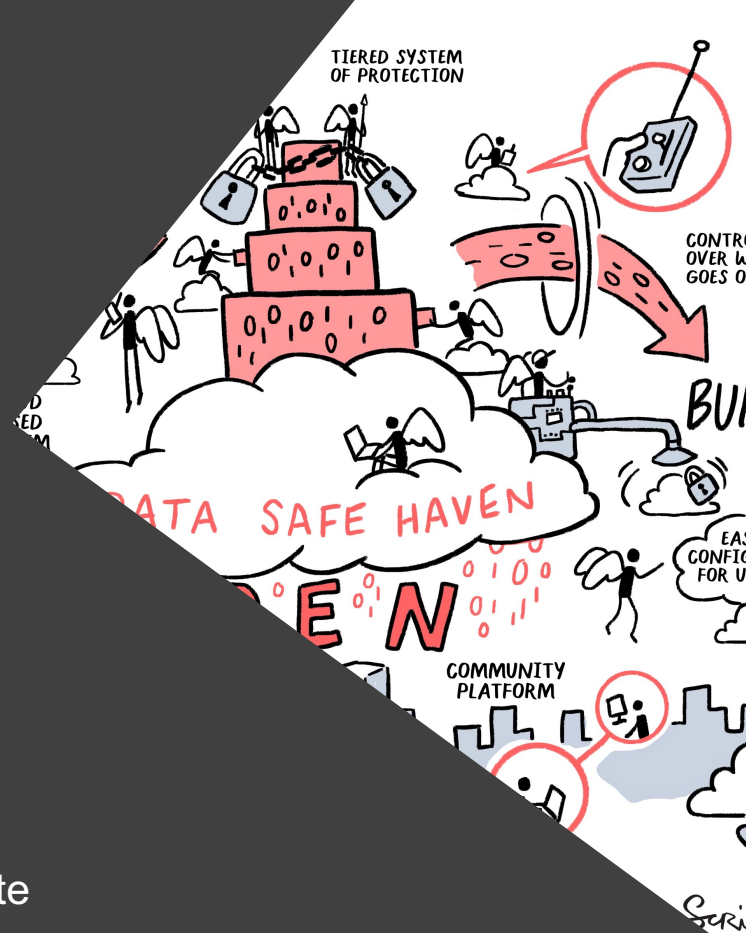


The Alan Turing Institute

Reproducible secure research environments

Martin O'Reilly

Director of Research Engineering, The Alan Turing Institute



Reproducible research

The Turing Way

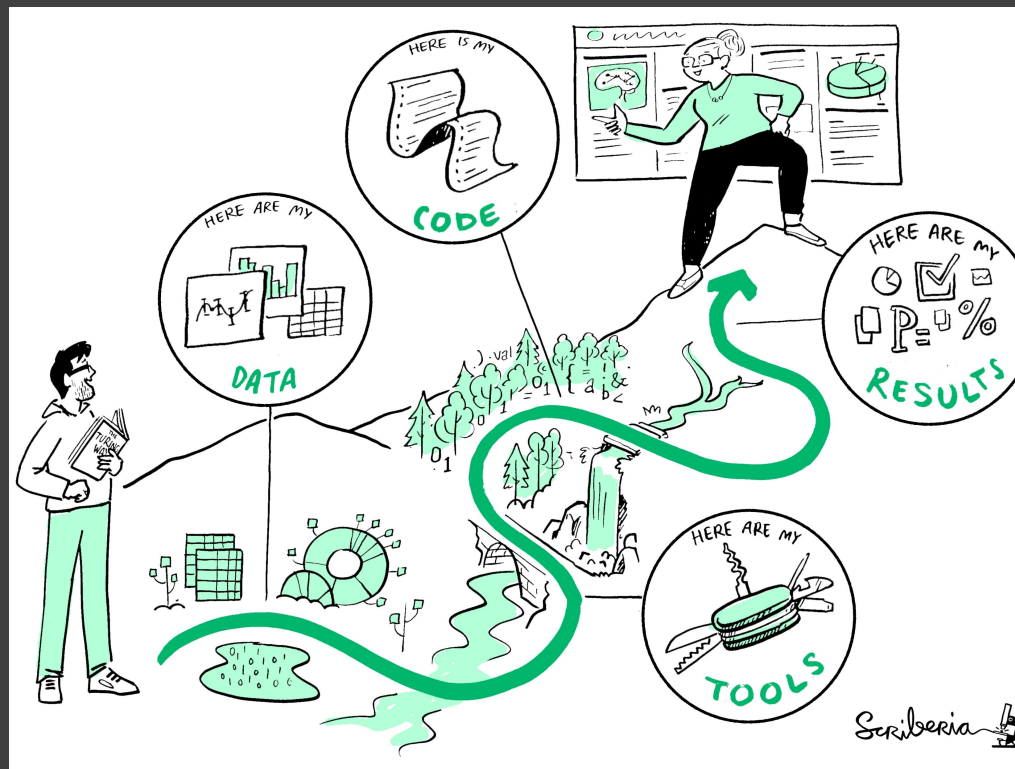
- a guide to reproducible, ethical, inclusive and collaborative data science
- open source
- community driven
- over 250 contributors
- guides for:
 - reproducible research
 - project design
 - communication
 - ethical research
 - community



		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

Reproducibility

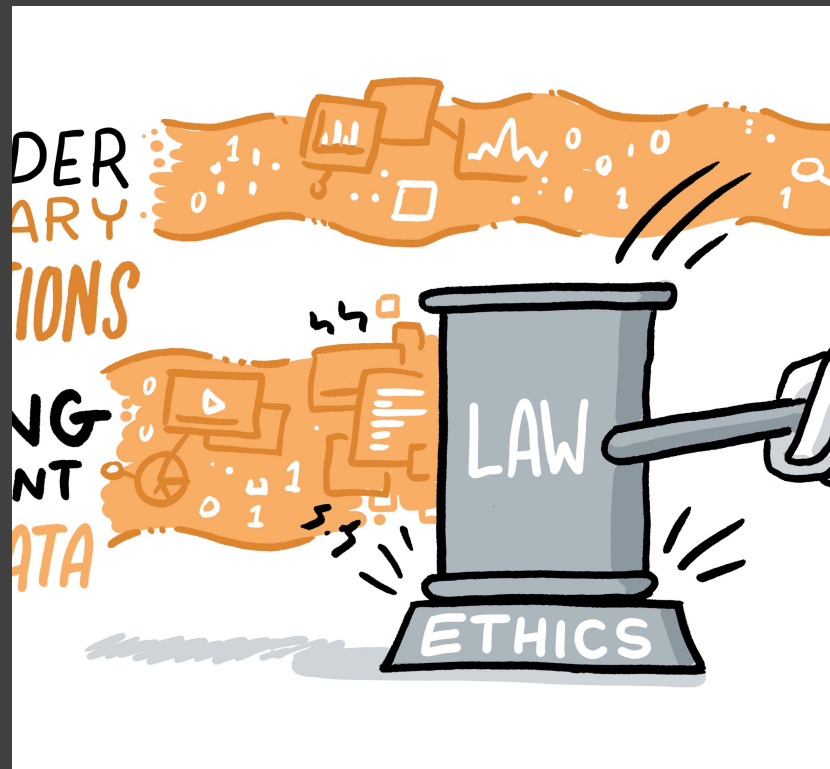
- results
- code
- data
- tools and environment
 - language version (Python, R)
 - language libraries (PyPI, CRAN)
 - operating system version
 - operating system libraries
 - other software



Reproducibility for sensitive data

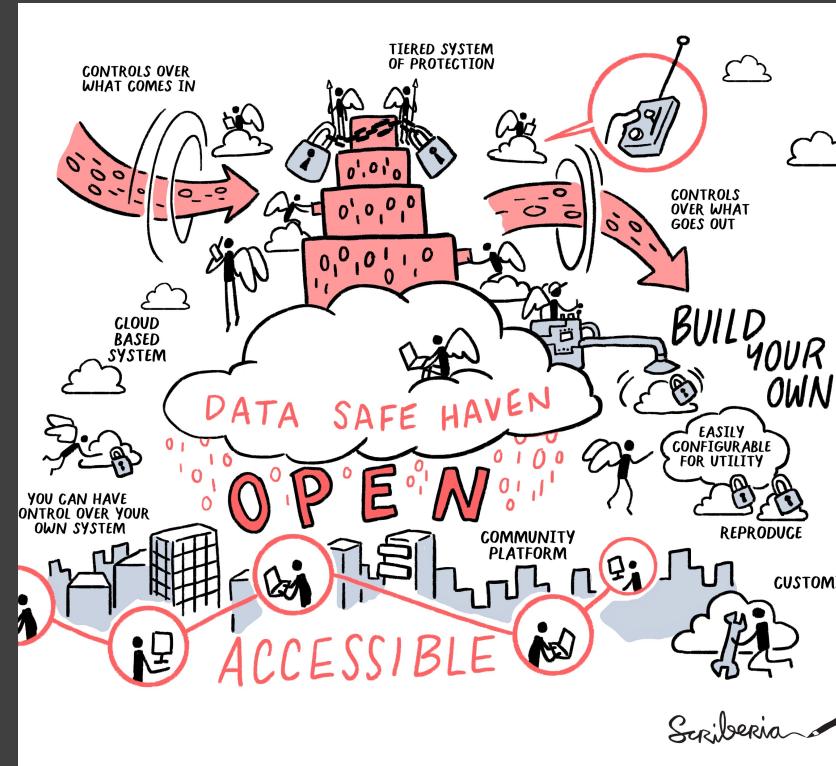
Ethical reproducibility

- consent for data is often narrow
- need to balance benefit vs risk of harm
- communicate benefit of reproducing research to data subjects
- reduce risk by:
 - supporting reproducibility in secure research environments
 - making less sensitive versions of data more widely accessible



Turing Data Safe Haven

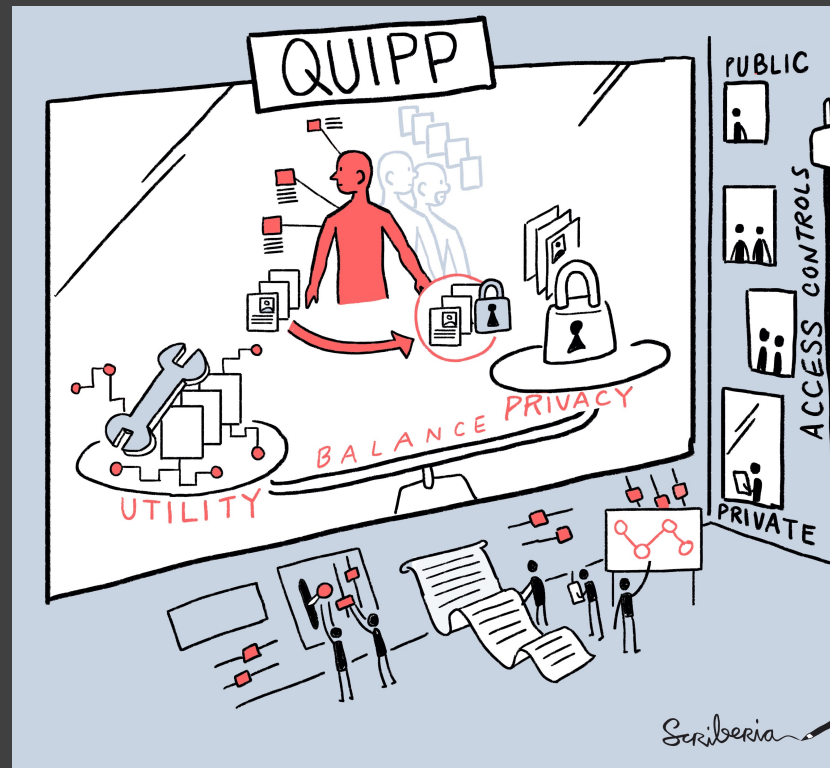
- Secure, scalable **cloud-based** compute environment
- Full suite of **data science software** and tools
- **Reproducible** software defined infrastructure
- Strong **isolation** between multiple projects
- Framework for agreeing **data sensitivity**



Safer data

- statistical disclosure control
- synthetic data
- learn a model of the data in a privacy-preserving manner
- generate non-sensitive data from the model
- privacy measures can be non-intuitive (e.g. differential privacy)
- privacy-utility trade-off for real-world data not clearly understood
- Turing projects are looking at this for health and government data

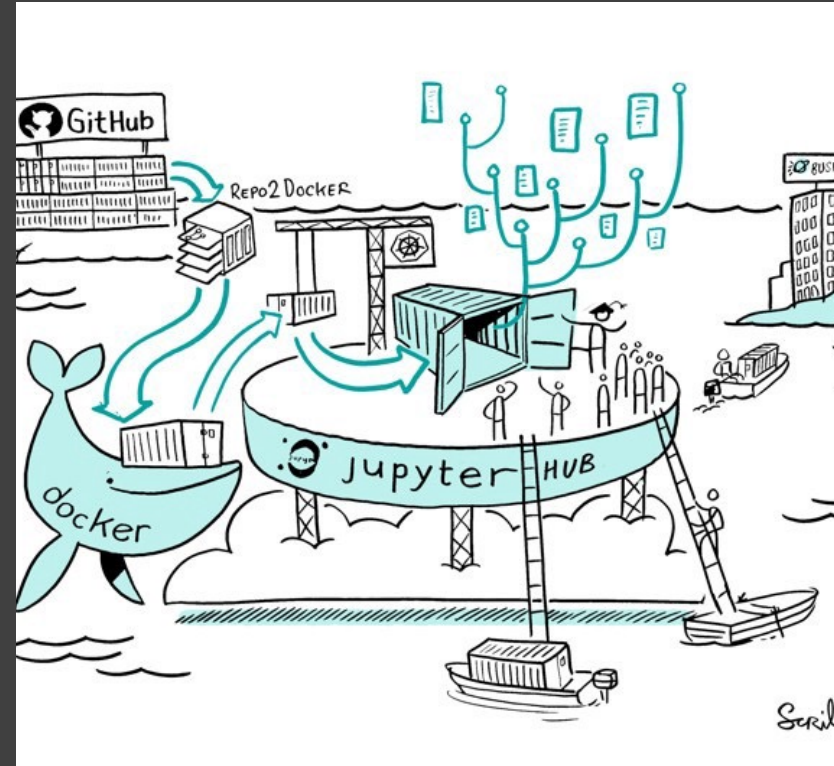
<https://doi.org/10.6084/m9.figshare.14748117>



<https://www.turing.ac.uk/research/research-projects/quipp-quantifying-utility-and-preserving-privacy-synthetic-data-sets>

Tools and environment

- software defined infrastructure provides reproducible technical security
- **but** secure administration of an independent Safe Haven is non-trivial
- version controlled virtual machine images with data science tooling
- secure access to Python and R package repositories
- support for virtual environments for R and Python
- looking at secure support for Docker



Data and code

- read-only data shares for input data
- databases with temporal versioning support (PostgreSQL + MS SQL)
- collaborative code development with version control via internal GitLab
- looking at support for versioning file-based data
- looking at more automated code ingress and results egress



Questions?

Martin O'Reilly | moreilly@turing.ac.uk | [@martinoreilly](https://twitter.com/martinoreilly)

<https://doi.org/10.6084/m9.figshare.14748117>