The Alan Turing Institute

The Turing Way

A handbook for reproducible data science

Kirstie Whitaker

MRC-BSU, March 2019

Slides at https://doi.org/10.5281/zenodo.2599904

Neurohackweek 2016

Photo credit: Chris Gorgolewski

- BSc Physics
- MSc Medical Physics
- PhD Neuroscience
- Postdoc Dept Psychiatry, Cambridge
- Mozilla Fellow for Science
- Research fellow Alan Turing Institute & senior research associate Dept Psychiatry

Founding the Institute

“We will found The Alan Turing Institute to ensure Britain leads the way again in the use of big data and algorithm research”

George Osborne, Chancellor of the Exchequer
Budget Speech, March 2014

The Alan Turing Institute

EPSRC
Engineering and Physical Sciences Research Council

Network of industry, charity, government partners

Network of university members

Strategic government investment

The Institute's partners and collaborators

Lloyd's Register Foundation

intel

GCHQ

dstl

Ministry of Defence

HSBC

uclh

accenture

NATS

British Heart Foundation

Microsoft

TOYOTA mobility FOUNDATION

Office for National Statistics

GREATER LONDON AUTHORITY

CRAY

nationalgrid

SIEMENS

LLOYDS BANKING GROUP

Cystic Fibrosis Trust

ROLLS ROYCE

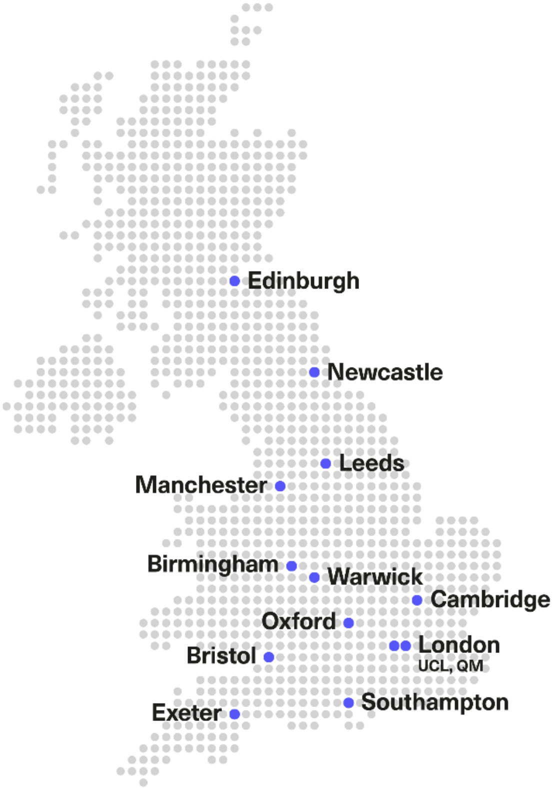
Scottish Enterprise

Nuffield Foundation

SAMSUNG

BRITISH AIRWAYS

Our university network



UNIVERSITY OF CAMBRIDGE

THE UNIVERSITY of EDINBURGH

UNIVERSITY OF OXFORD

UCL

WARWICK
THE UNIVERSITY OF WARWICK

University of BRISTOL

UNIVERSITY OF BIRMINGHAM

UNIVERSITY OF EXETER

UNIVERSITY OF LEEDS

MANCHESTER 1824
The University of Manchester

Queen Mary
University of London

Newcastle University
UK | Malaysia | Singapore

UNIVERSITY OF Southampton

Challenges

Advance data science and artificial intelligence to…

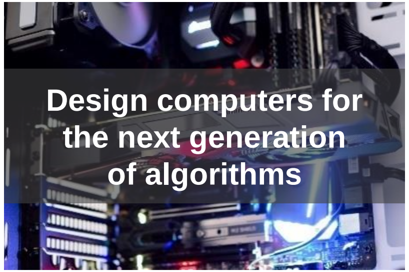
Revolutionise healthcare

Deliver safer, smarter engineering

Manage security in an insecure world

Shine a light on our economy

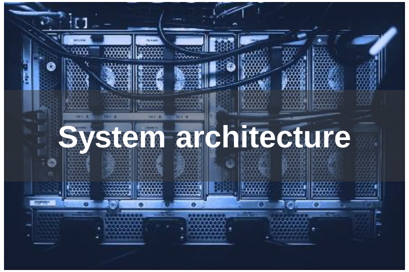
Make algorithmic systems fair, transparent, and ethical


Design computers for the next generation of algorithms

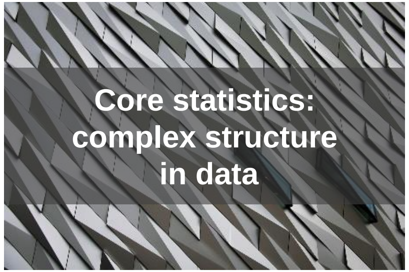
Supercharge research in science and humanities

Foster government innovation

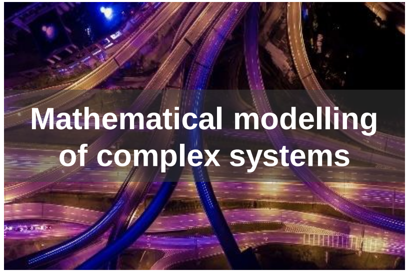
Core capabilities


System architecture

Security and robustness

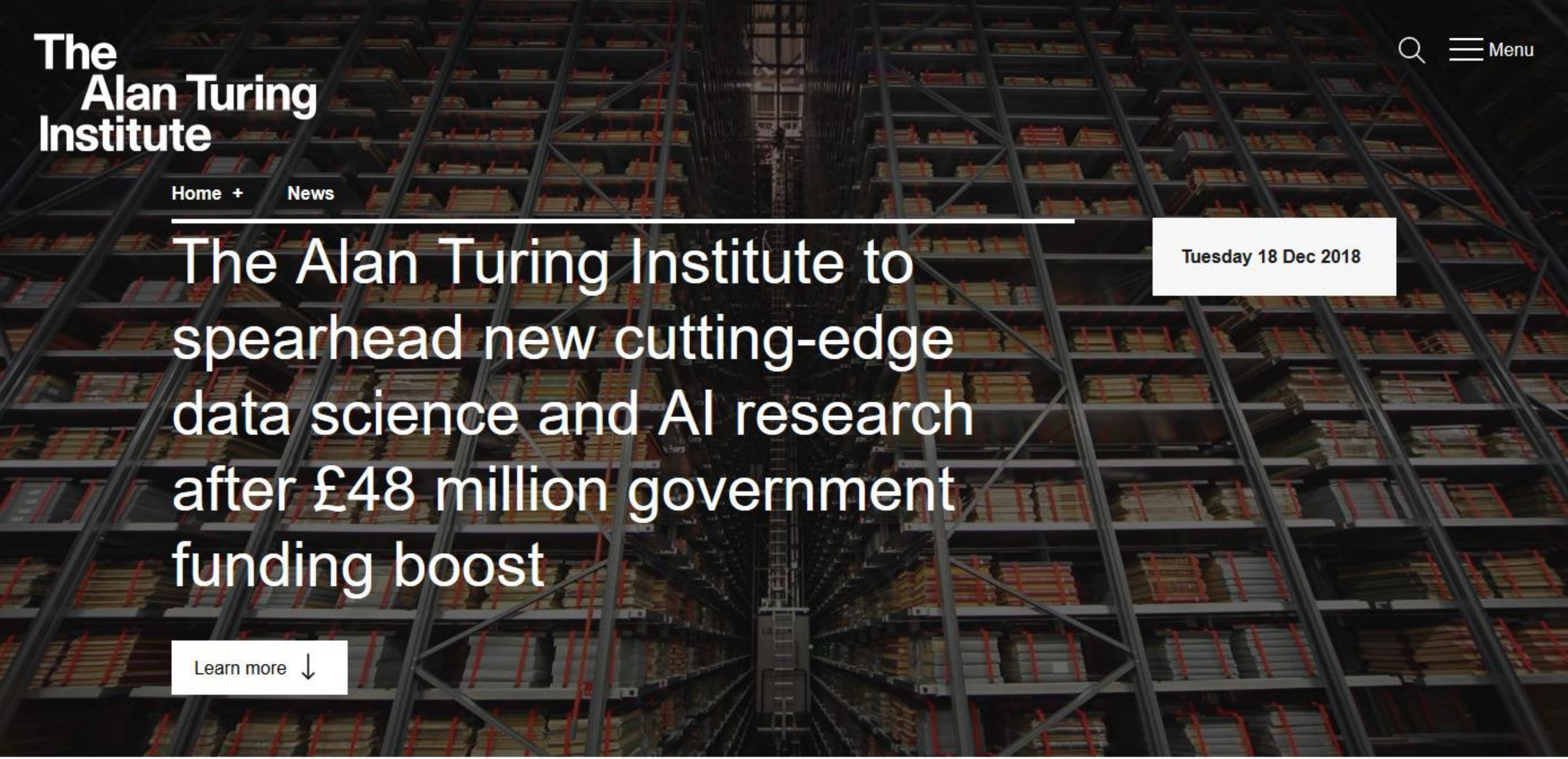

Core statistics: complex structure in data

Machine learning and artificial intelligence


Mathematical modelling of complex systems

Understanding human behaviour

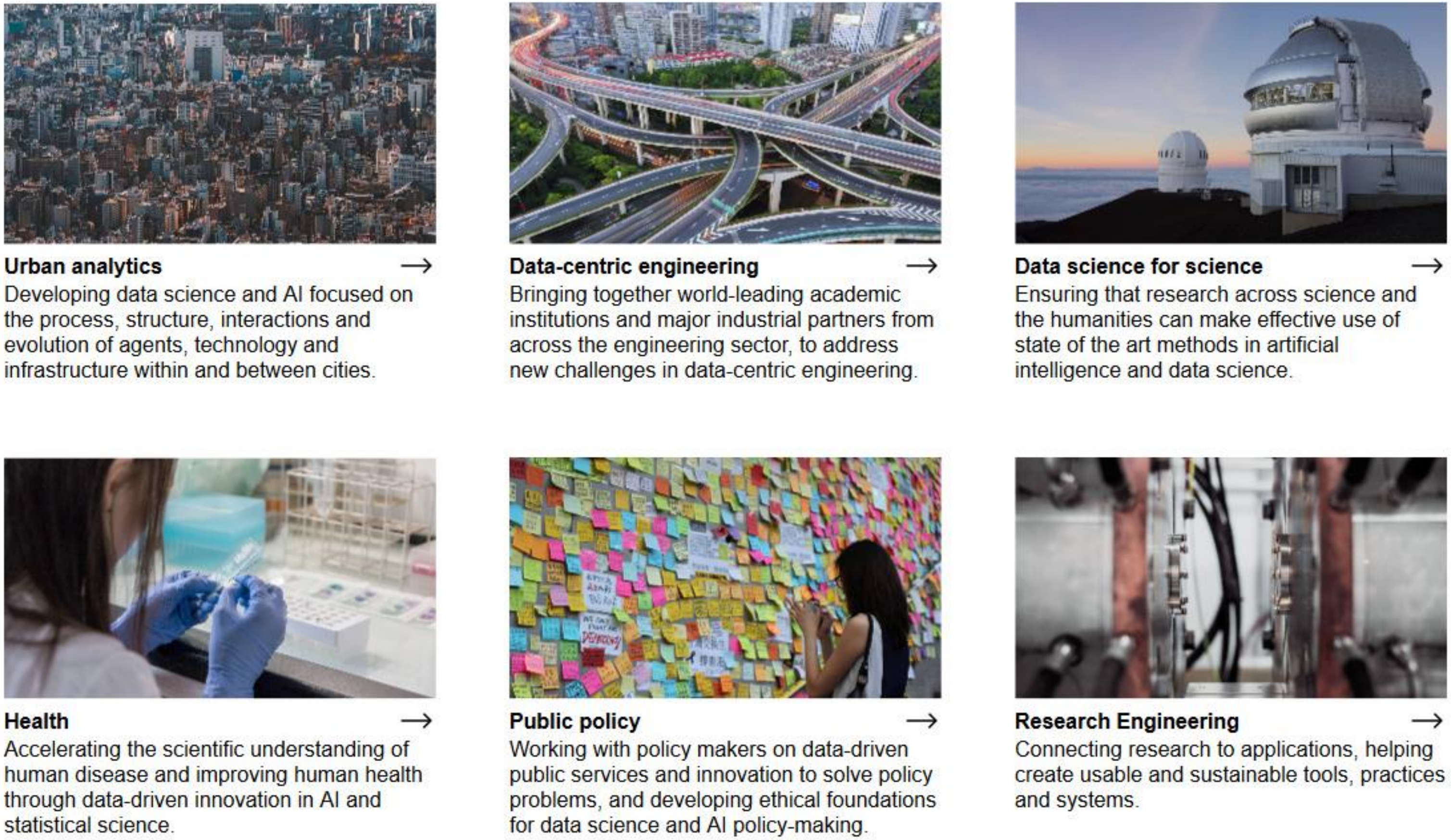
Ethics of data science and artificial intelligence



Tuesday 18 Dec 2018

The Alan Turing Institute to spearhead new cutting-edge data science and AI research after £48 million government funding boost

https://www.turing.ac.uk/news/alan-turing-institute-spearhead-new-cutting-edge-data-science-and-artificial-intelligence



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

Data science for science →

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

Health →

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

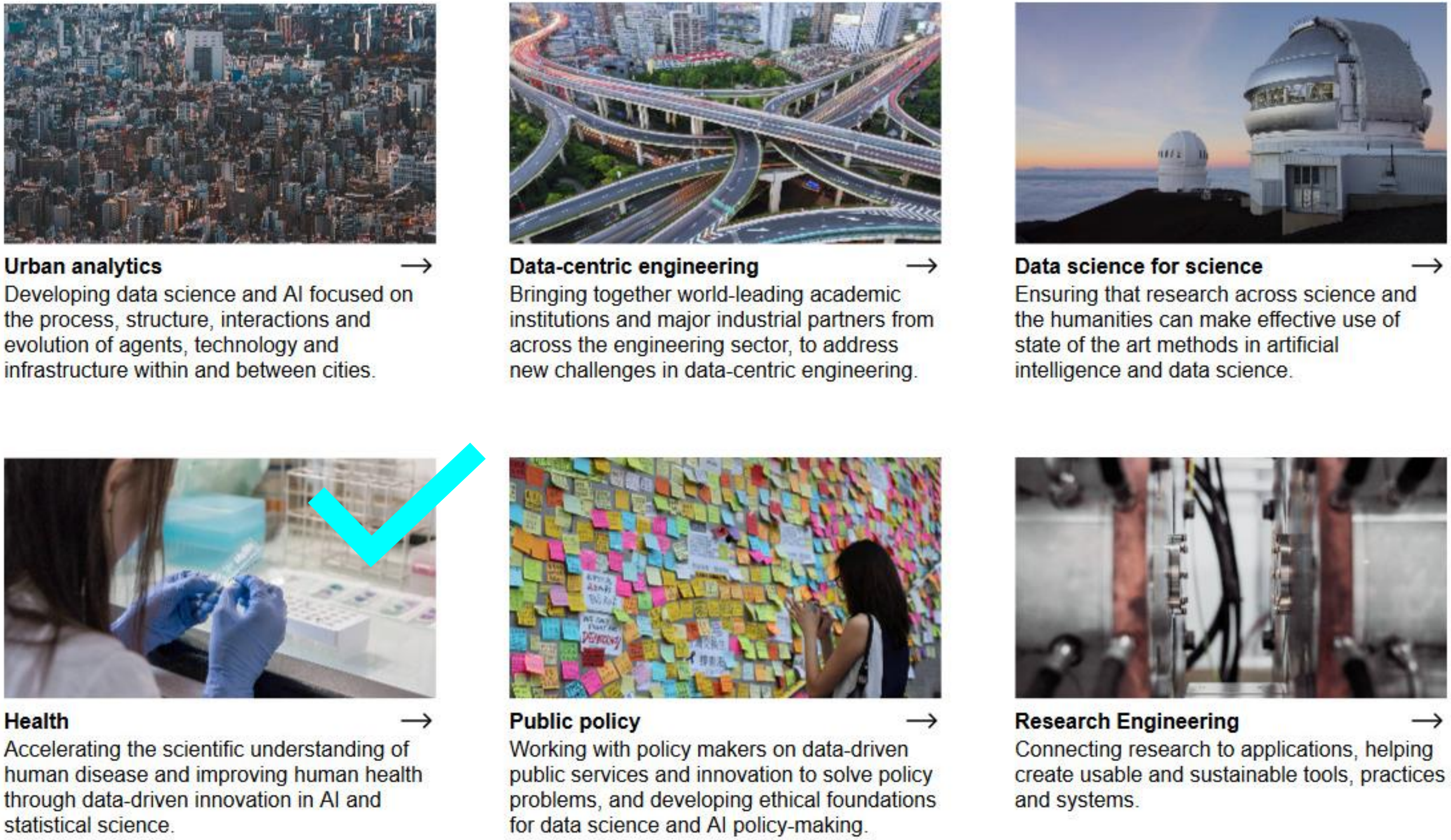
Public policy →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Research Engineering →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

https://www.turing.ac.uk/research/ai-science-engineering-health-and-government



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

Data science for science →

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

✓

Health →

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

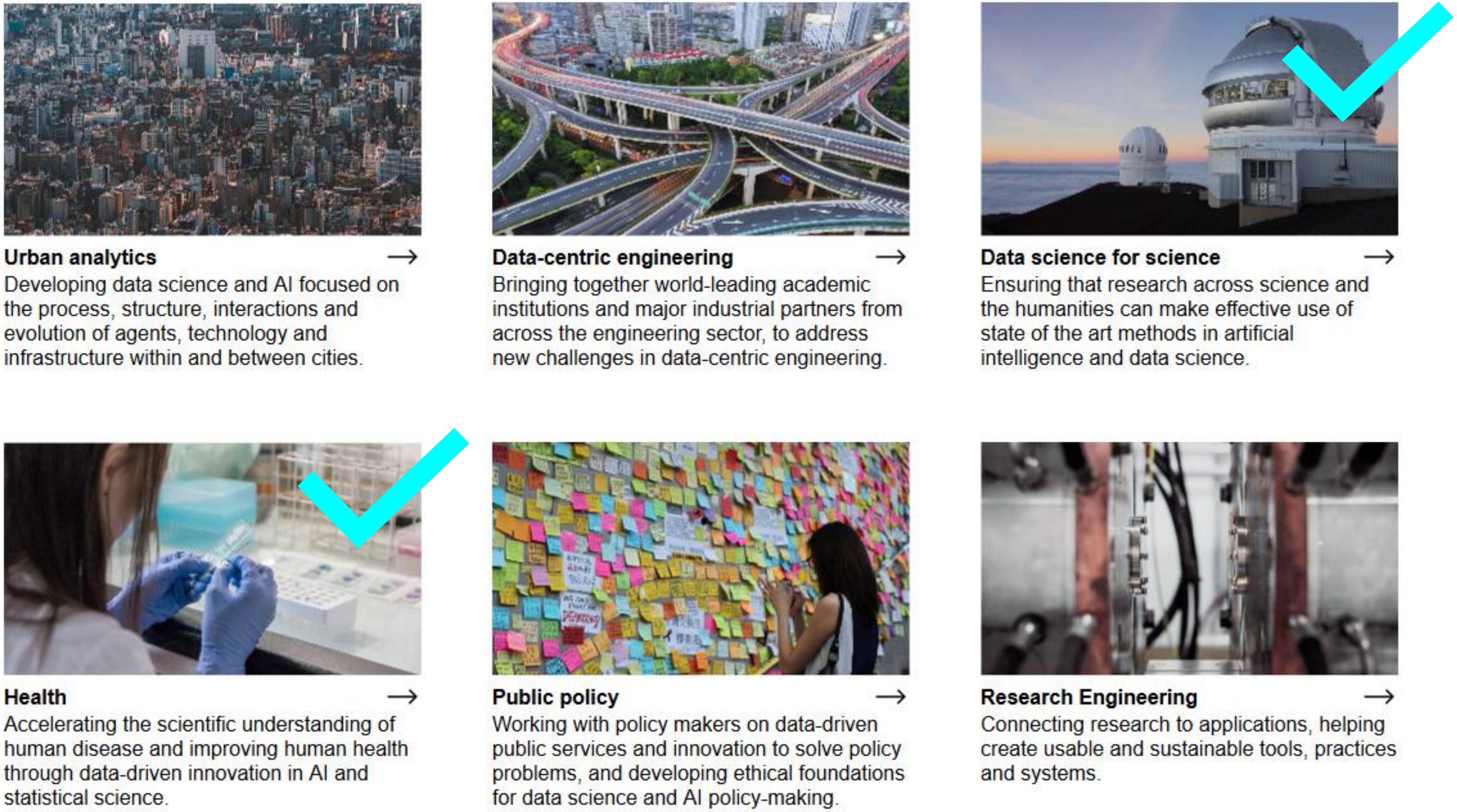
Public policy →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Research Engineering →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

https://www.turing.ac.uk/research/ai-science-engineering-health-and-government



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

Data science for science → ✓

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

Health → ✓

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

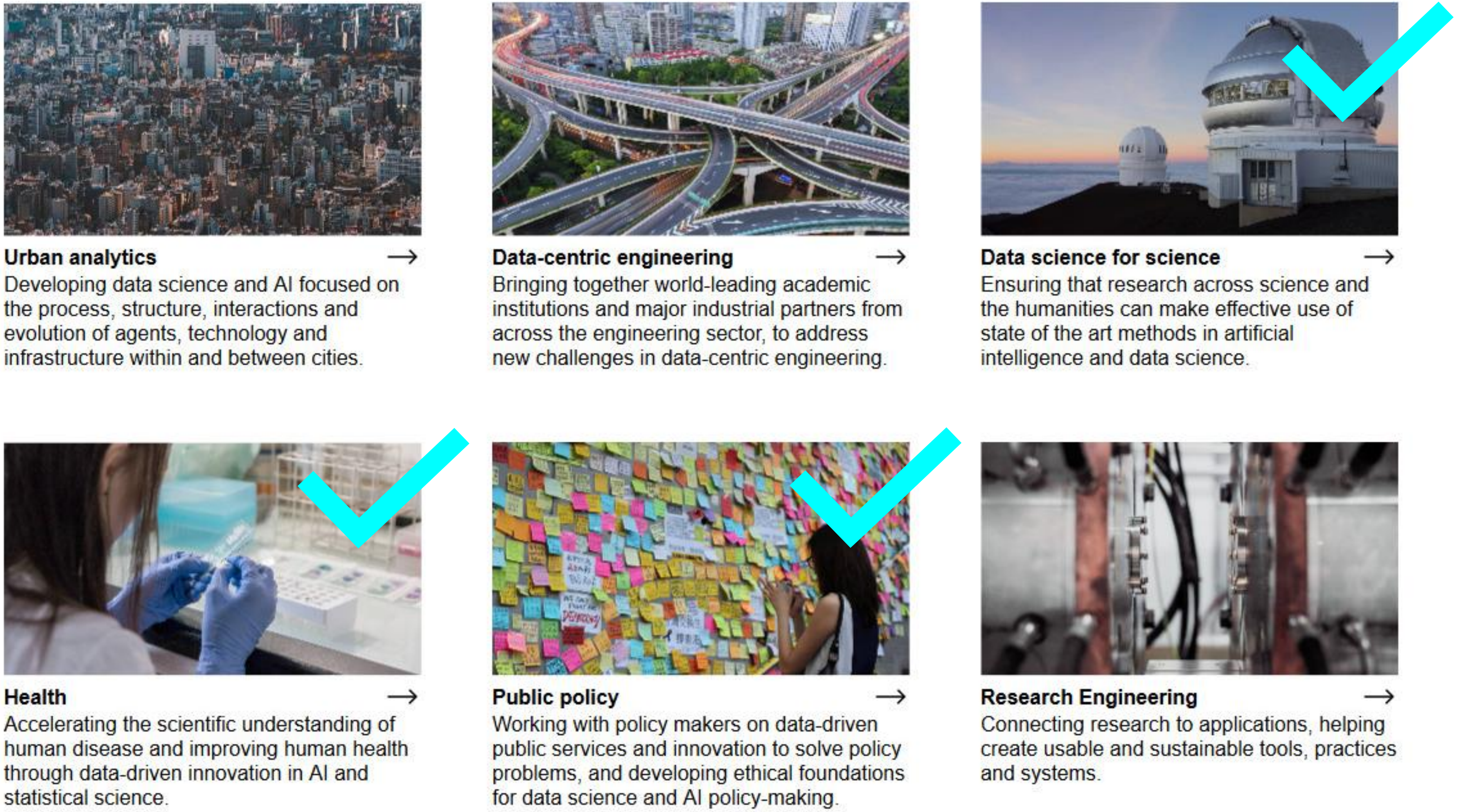
Public policy →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Research Engineering →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

https://www.turing.ac.uk/research/ai-science-engineering-health-and-government



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

✓ **Data science for science** →

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

✓ **Health** →

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

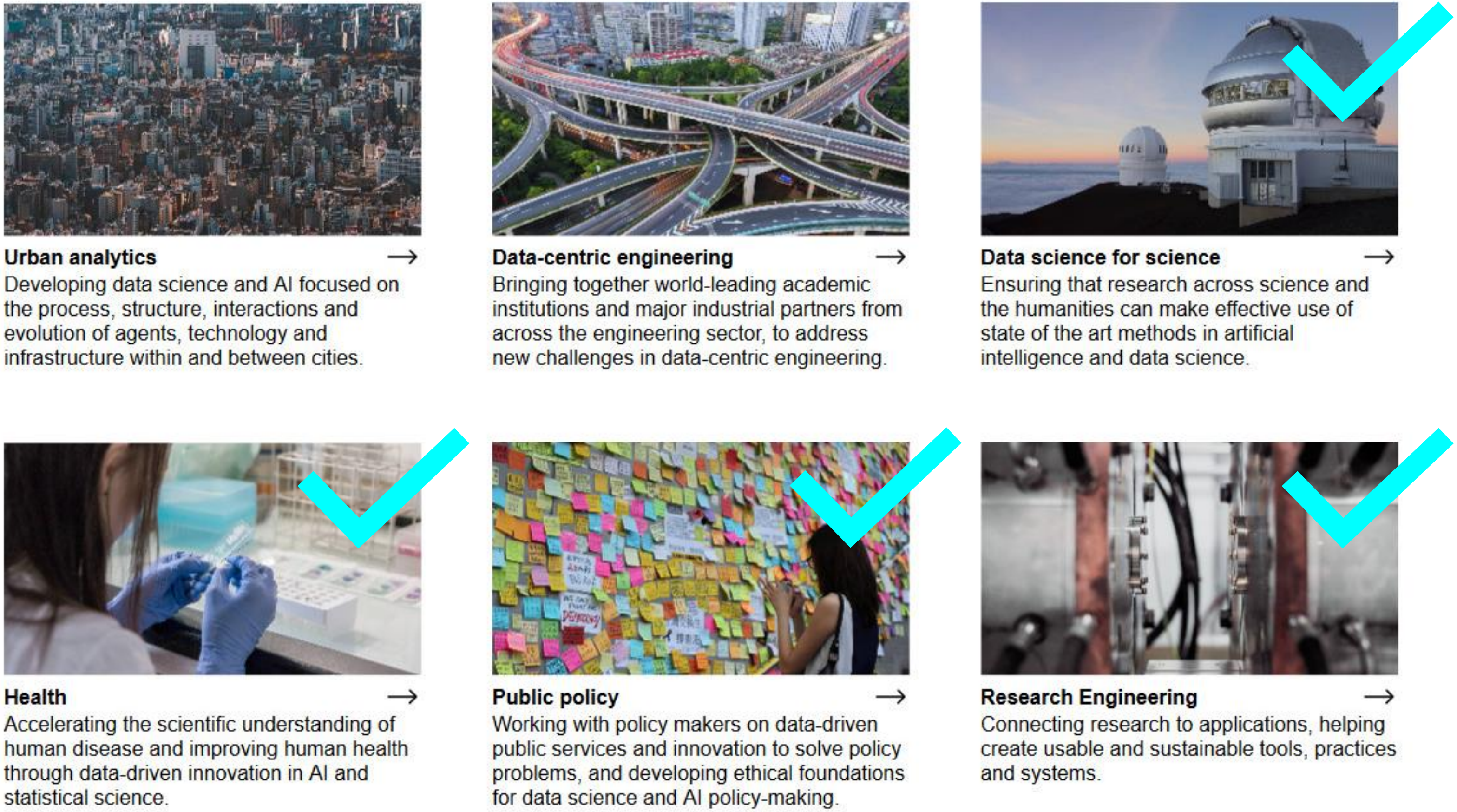
✓ **Public policy** →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Research Engineering →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

https://www.turing.ac.uk/research/ai-science-engineering-health-and-government



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

✓ **Data science for science** →

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

✓ **Health** →

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

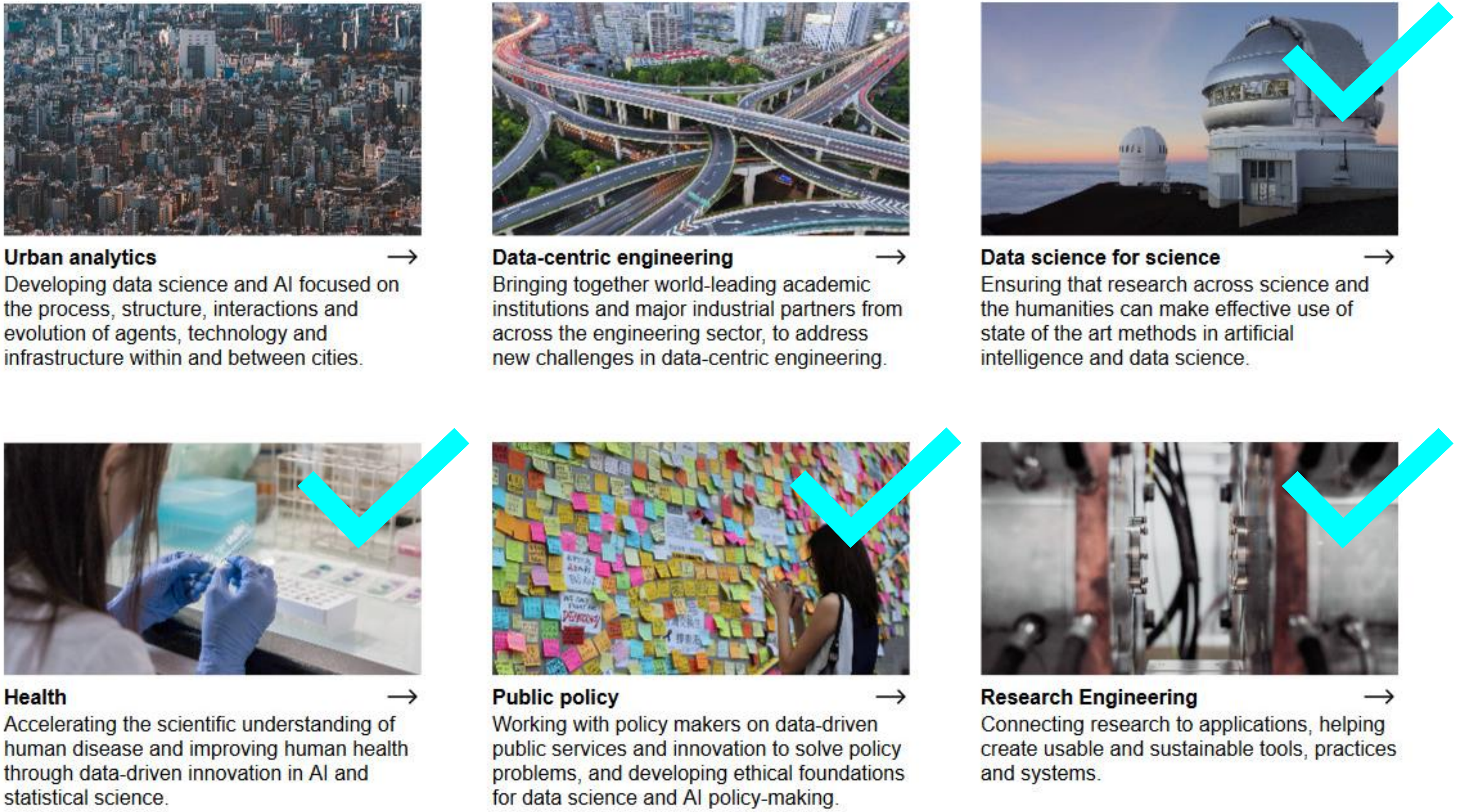
✓ **Public policy** →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

✓ **Research Engineering** →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

https://www.turing.ac.uk/research/ai-science-engineering-health-and-government



Urban analytics →

Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data-centric engineering →

Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

✓ **Data science for science** →

Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

✓ **Health** →

Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

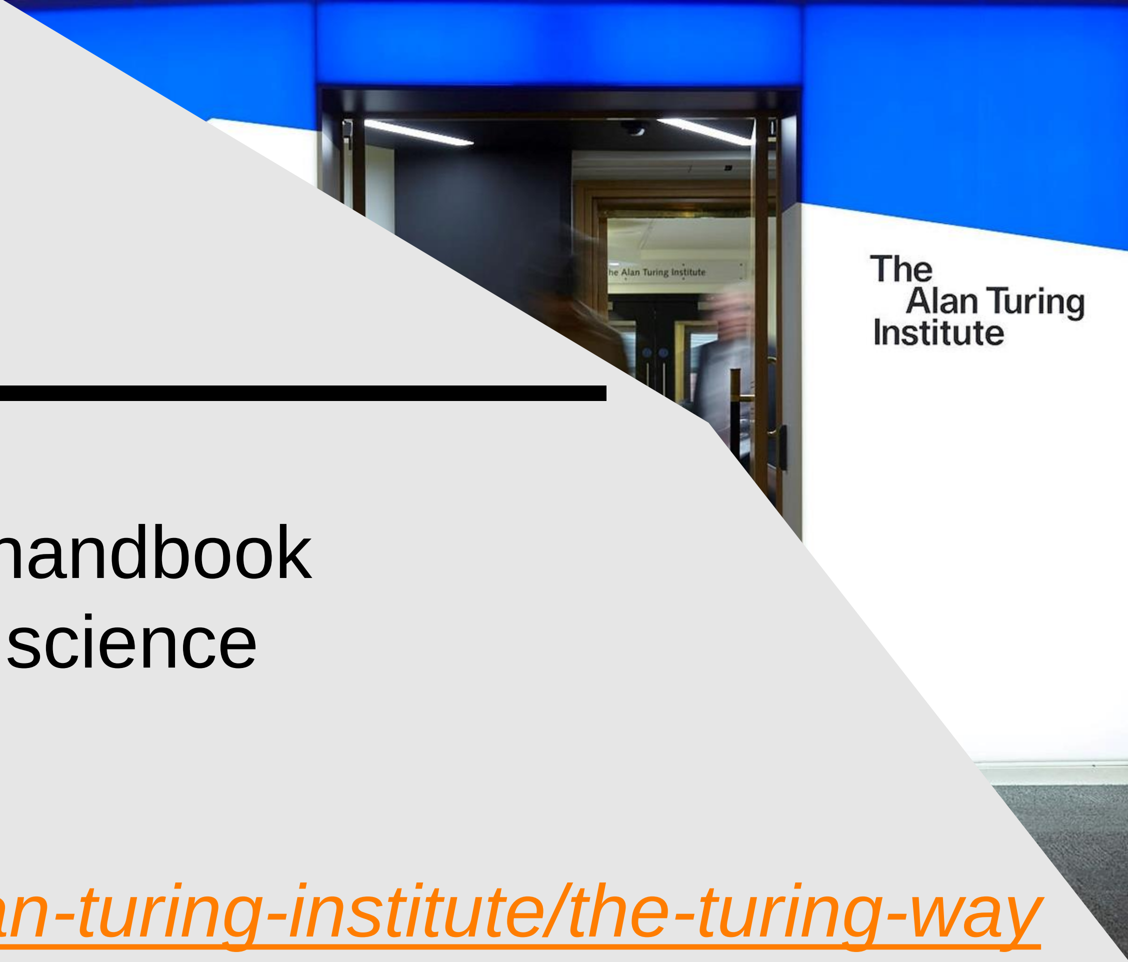
✓ **Public policy** →

Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

✓ **Research Engineering** →

Connecting research to applications, helping create usable and sustainable tools, practices and systems.

Cross cutting theme: Tools, systems and practices



The Turing Way

A lightly opinionated handbook for reproducible data science

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

What does reproducible mean?

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://dx.doi.org/10.6084/m9.figshare.7140050

Why don’t people do this already?

Is not considered for promotion

Takes time

Publication bias towards novel findings

Barriers to reproducible research

Requires additional skills

Plead the 5th

Held to higher standards than others

Support additional users

https://dx.doi.org/10.6084/m9.figshare.7140050

Why don’t people do this already?

Is not considered for promotion

Takes time

Publication bias towards novel findings

Barriers to reproducible research

Requires additional skills

Plead the 5th

Held to higher standards than others

Support additional users

https://dx.doi.org/10.6084/m9.figshare.7140050

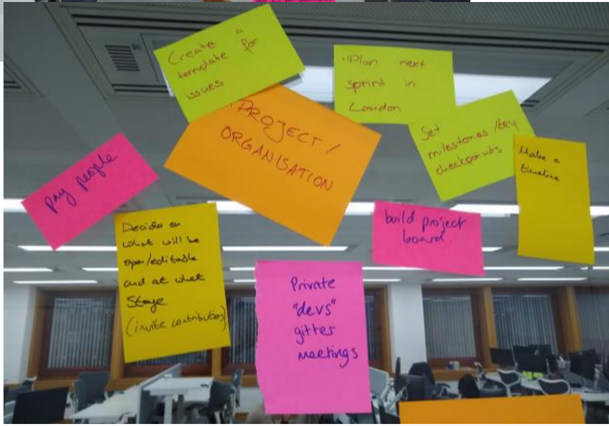
Requires additional skills

Chapters will include:

- Research data management
- Open research
- Reproducibility
- Version control with git
- Your working environment (IDE, notebooks etc)
- Capturing your compute environment
- Testing for research
- Continuous integration
- Collaborating through GitHub/GitLab

Explore options / rules for integrating external content
Decide what programming languages will focus on (and share)
Decide on chapter stage
Create a table of contents
Create checklist template
Building checklists
Build simple check-list get feedback
Building templates for tests, contents etc.
Assigning 'leaders' to each authors chapters January 2nd
Write what / why summaries for each topic
Create a chapter template
Outline the process for writing a chapter
Tailoring or checklists?

Find out what types of Azure resource will be needed for Turing Binder Hub
BinderHub
Gather all relevant documentation on JupyterHub and BinderHub
Research into linking BinderHub to Azure instances & resources
Touch base with Tim Head re Binder (tutorial on GitHub?)
Research on Binder / BinderHub (Jupyter Lab)

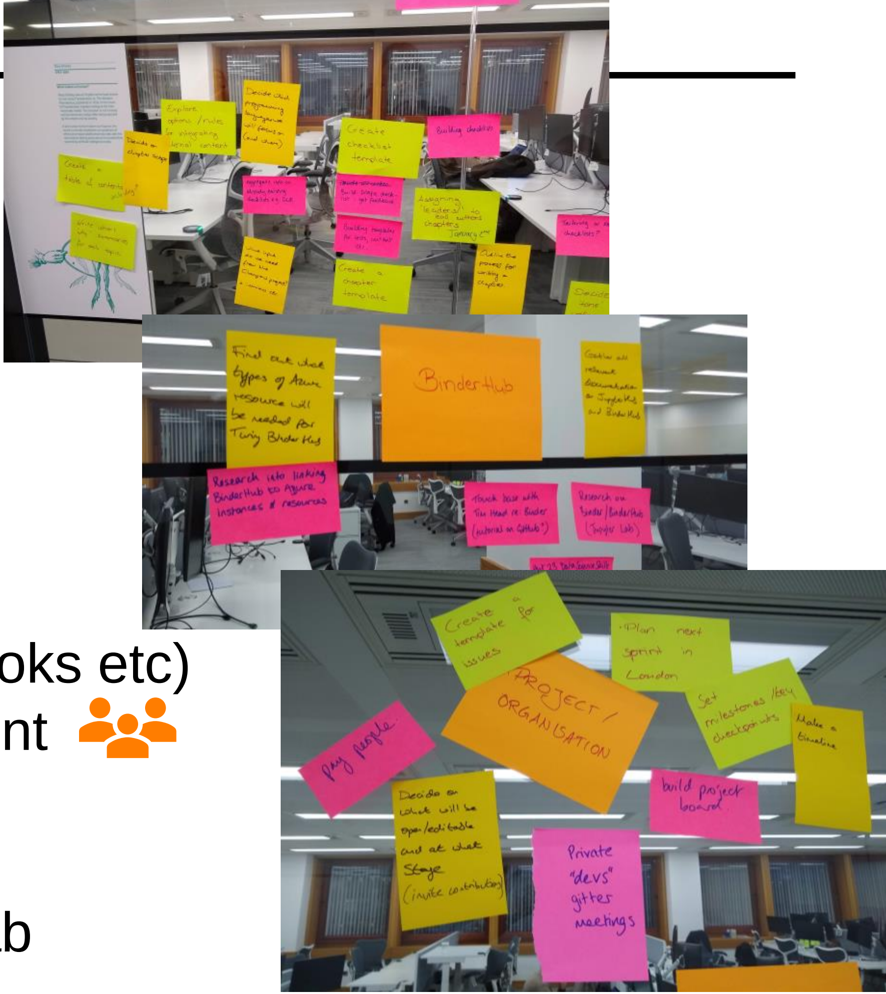

Create a template for issues
Plan next sprint in London
PROJECT / ORGANISATION
Set milestones / key checkpoints
Make a timeline
pay people
Decide on what will be open/editable and at what Stage (invite contributors)
build project board
Private "devs" gitter meetings

https://github.com/alan-turing-institute/the-turing-way/blob/master/book_skeleton.md

Requires additional skills

Chapters will include:

- Research data management
- Open research
- Reproducibility
- Version control with git
- Your working environment (IDE, notebooks etc)
- Capturing your compute environment
- Testing for research
- Continuous integration
- Collaborating through GitHub/GitLab


Decide what programming language will focus on (and why)
Explore options / rules for integrating external content
Create checklist template
Building checklists
Create table of contents
Assigning leaders to chapters
Create a chapter template
Find out what types of Azure resource will be needed for Turing BinderHub
BinderHub
Gather all relevant documentation on JupyterHub and BinderHub
Research into linking BinderHub to Azure instances & resources
Touch base with Tim Head re: Binder (tutorial on GitHub?)
Research on JupyterHub / BinderHub (Jupyter Lab)
Create a template for issues
Plan next sprint in London
PROJECT / ORGANISATION
Set milestones / key checkpoints
Make a timeline
pay people
build project board
Decide on what will be open/editable and at what stage (invite contributors)
Private "devs" gitter meetings

https://github.com/alan-turing-institute/the-turing-way/blob/master/book_skeleton.md

Built by a team….and you!

- Becky Arnold
- Louise Bowler
- Sarah Gibson
- Patricia Herterich
- Rosie Higman
- Anna Krystalli
- Alex Morley
- Martin O’Reilly
- . . .

moz://a

Open Leadership Principles

Understanding
You make the work accessible and clear

Sharing
You make the work easy to adapt, reproduce, and spread

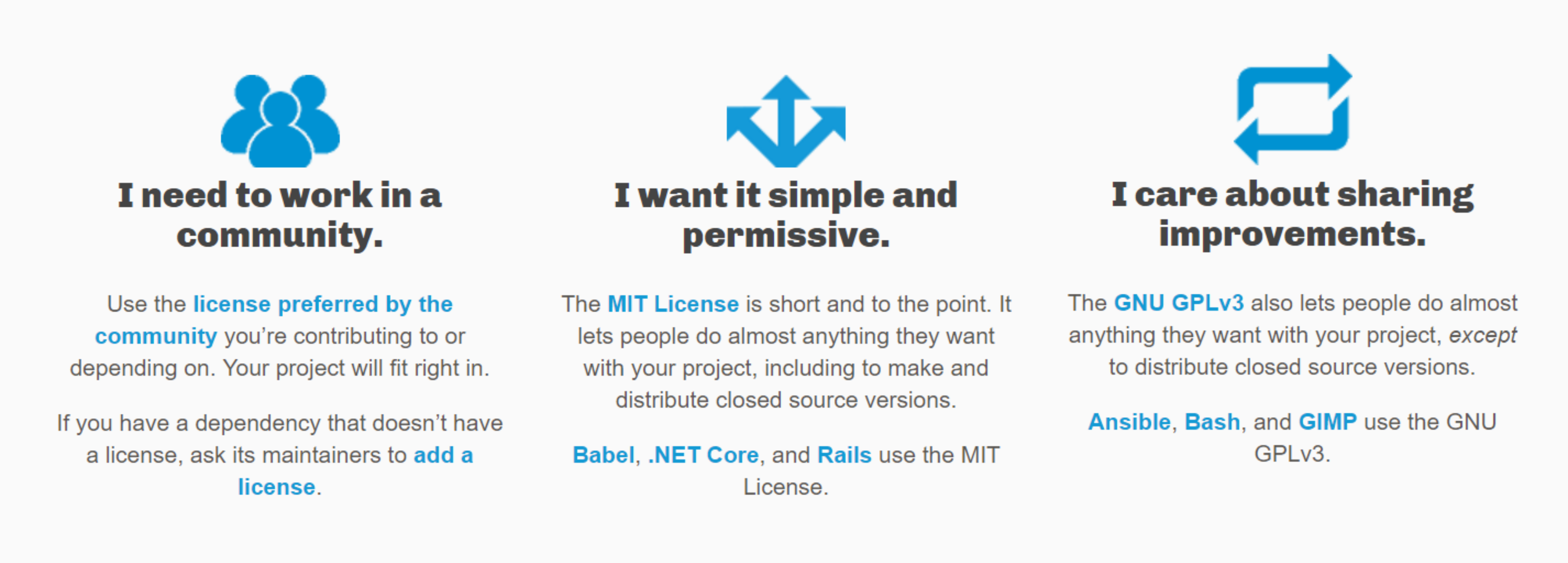
Participation & Inclusion
You build shared ownership and agency to make the work inviting and sustainable for all.

Read more
https://mozilla.github.io/olm-whitepaper

@kirstie_j
https://doi.org/10.6084/m9.figshare.7564682

moz://a

Openly licensed



I need to work in a community.

Use the **license preferred by the community** you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to **add a license**.

I want it simple and permissive.

The **MIT License** is short and to the point. It lets people do almost anything they want with your project, including to make and distribute closed source versions.

Babel, **.NET Core**, and **Rails** use the MIT License.

I care about sharing improvements.

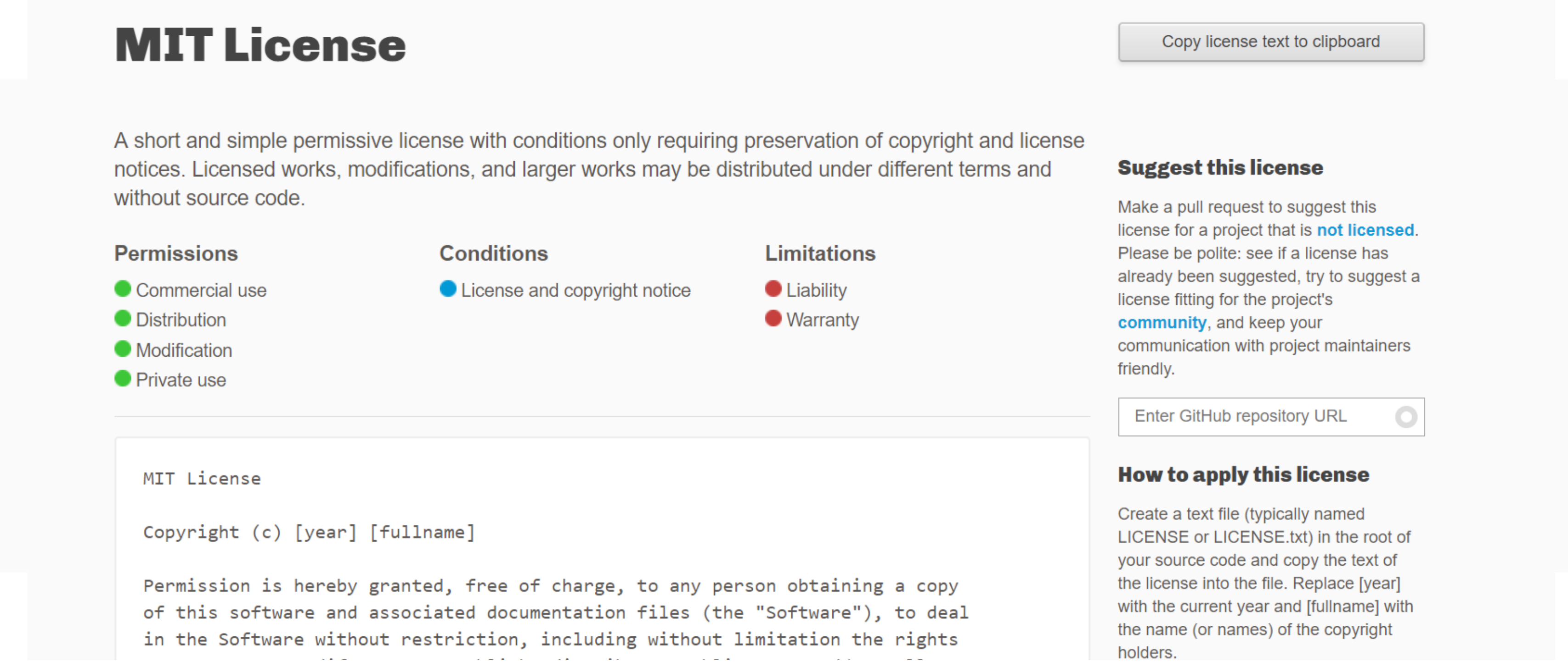
The **GNU GPLv3** also lets people do almost anything they want with your project, *except* to distribute closed source versions.

Ansible, **Bash**, and **GIMP** use the GNU GPLv3.

- CC-BY for content
- MIT for software

https://choosealicense.com

Openly licensed



MIT License

A short and simple permissive license with conditions only requiring preservation of copyright and license notices. Licensed works, modifications, and larger works may be distributed under different terms and without source code.

Permissions	Conditions	Limitations
● Commercial use	● License and copyright notice	● Liability
● Distribution		● Warranty
● Modification		
● Private use		

```
MIT License

Copyright (c) [year] [fullname]

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
```

Suggest this license

Make a pull request to suggest this license for a project that is **not licensed**. Please be polite: see if a license has already been suggested, try to suggest a license fitting for the project's **community**, and keep your communication with project maintainers friendly.

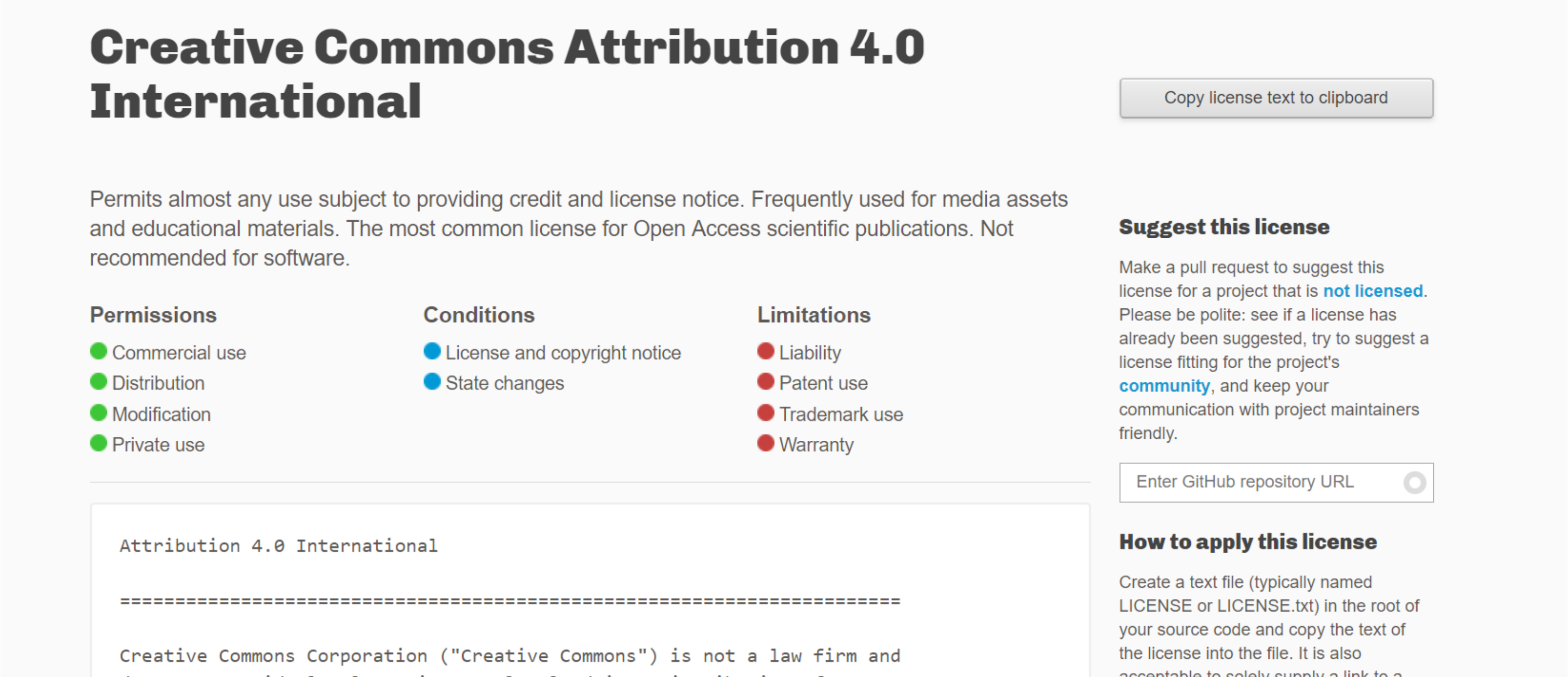
How to apply this license

Create a text file (typically named LICENSE or LICENSE.txt) in the root of your source code and copy the text of the license into the file. Replace [year] with the current year and [fullname] with the name (or names) of the copyright holders.

- CC-BY for content
- MIT for software

https://choosealicense.com/licenses/mit

Openly licensed



Creative Commons Attribution 4.0 International

Permits almost any use subject to providing credit and license notice. Frequently used for media assets and educational materials. The most common license for Open Access scientific publications. Not recommended for software.

Permissions	Conditions	Limitations
🟢 Commercial use	🔵 License and copyright notice	🔴 Liability
🟢 Distribution	🔵 State changes	🔴 Patent use
🟢 Modification		🔴 Trademark use
🟢 Private use		🔴 Warranty

```
Attribution 4.0 International

=======================================================================

Creative Commons Corporation ("Creative Commons") is not a law firm and
```

Suggest this license

Make a pull request to suggest this license for a project that is **not licensed**. Please be polite: see if a license has already been suggested, try to suggest a license fitting for the project's **community**, and keep your communication with project maintainers friendly.

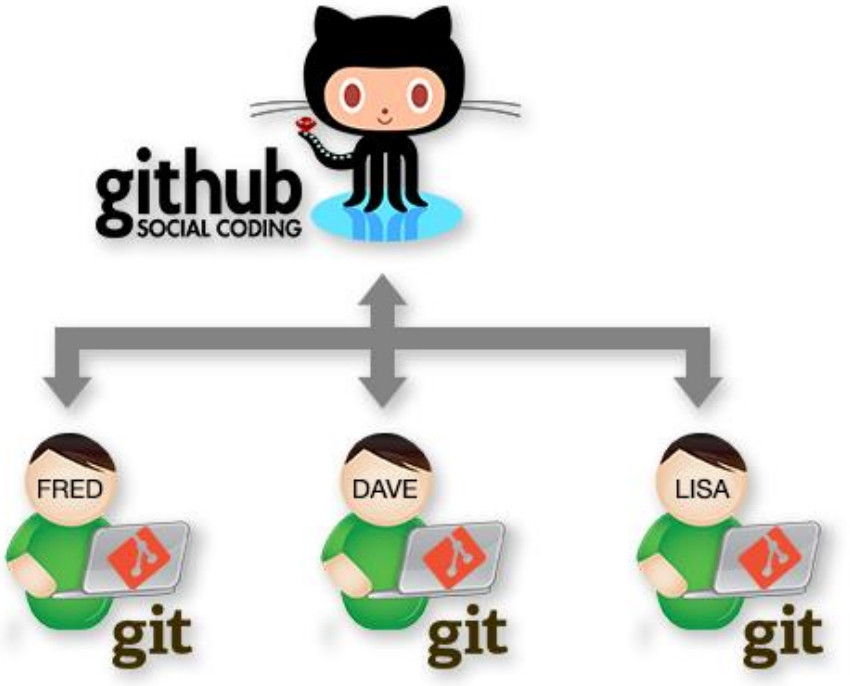
How to apply this license

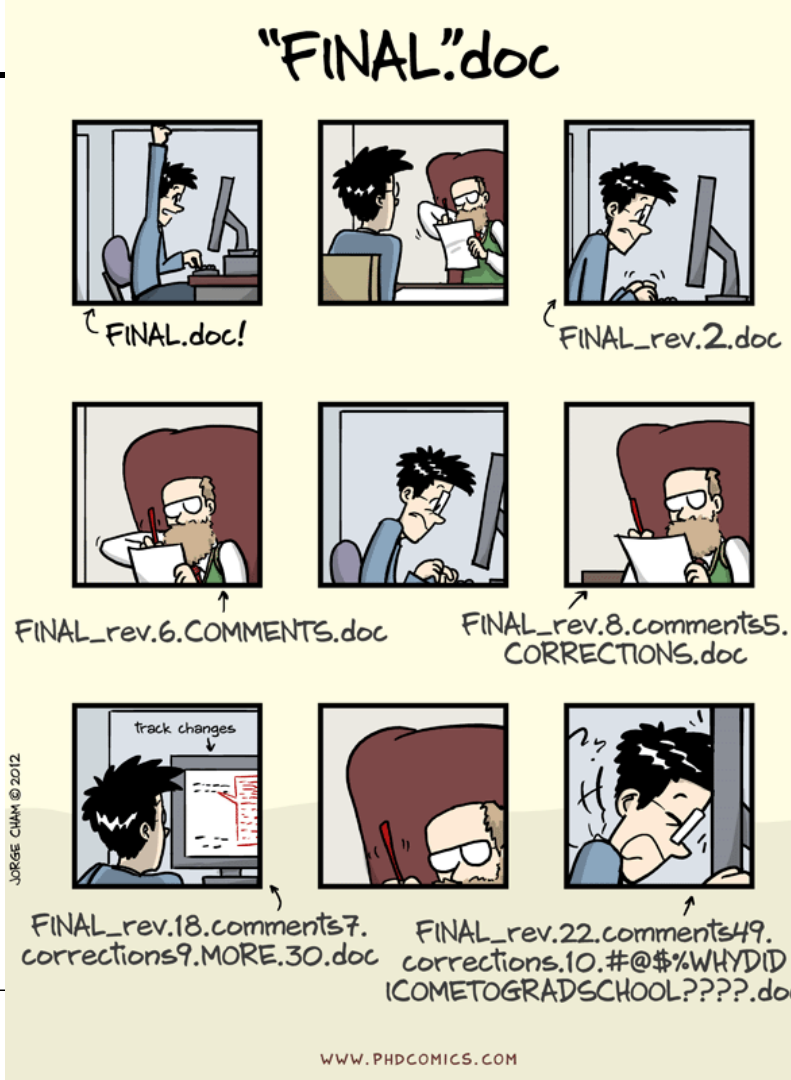
Create a text file (typically named LICENSE or LICENSE.txt) in the root of your source code and copy the text of the license into the file. It is also acceptable to solely supply a link to a

- CC-BY for content
- MIT for software

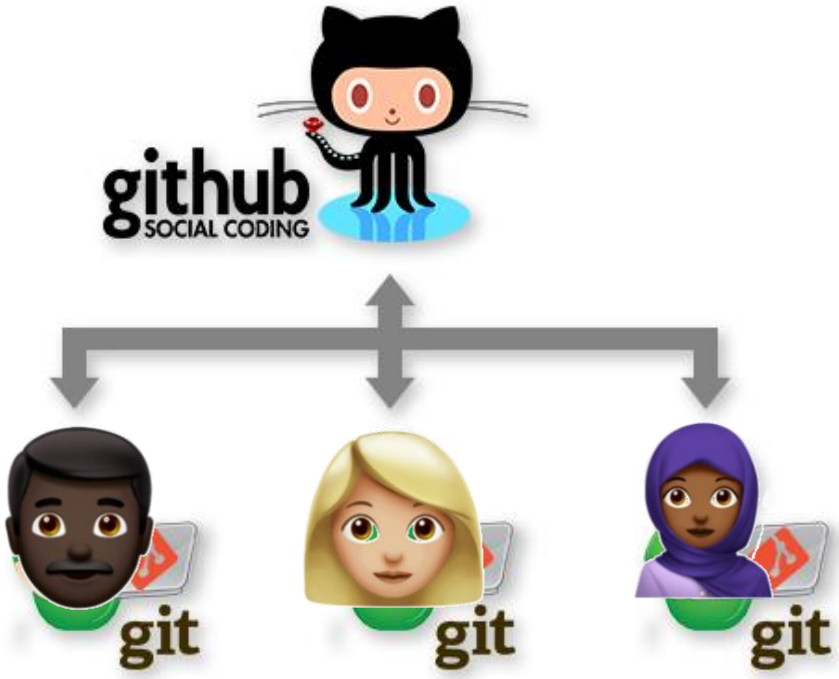
https://choosealicense.com/licenses/cc-by-4.0

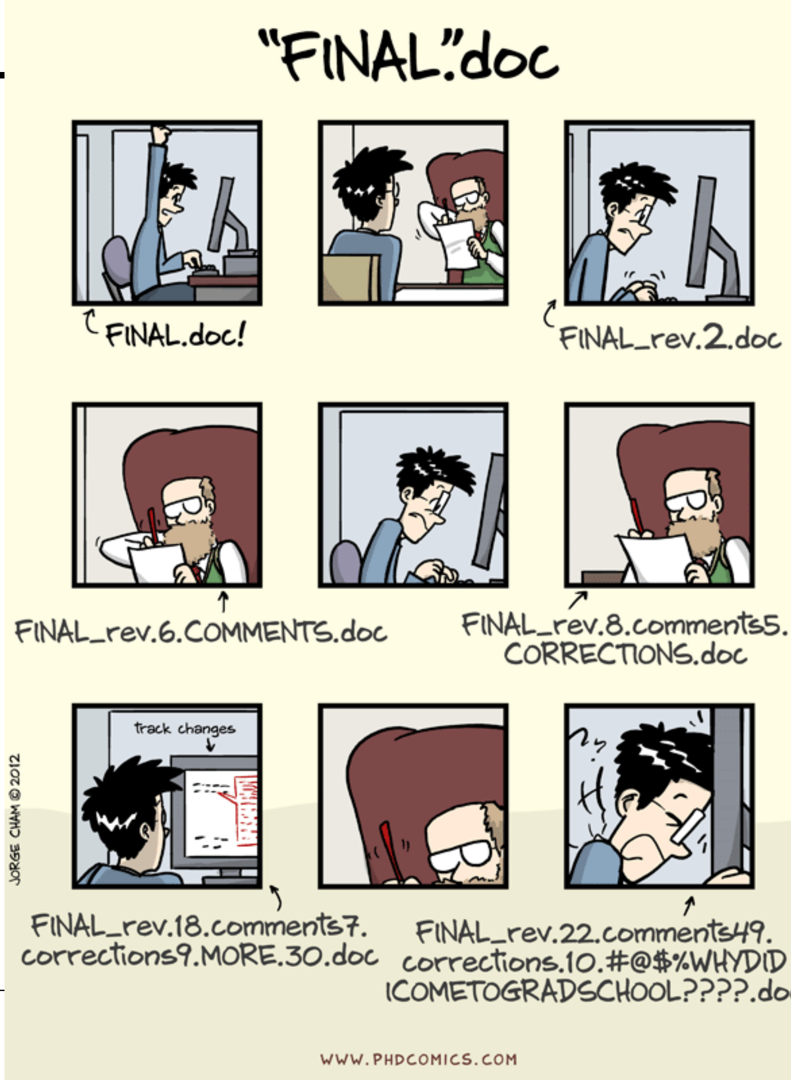
Version control





Version control





Testing (aka making explicit sanity checks)

Is your code doing what you think it’s doing? Does 2 + 2 = 4?

Testing (aka making explicit sanity checks)

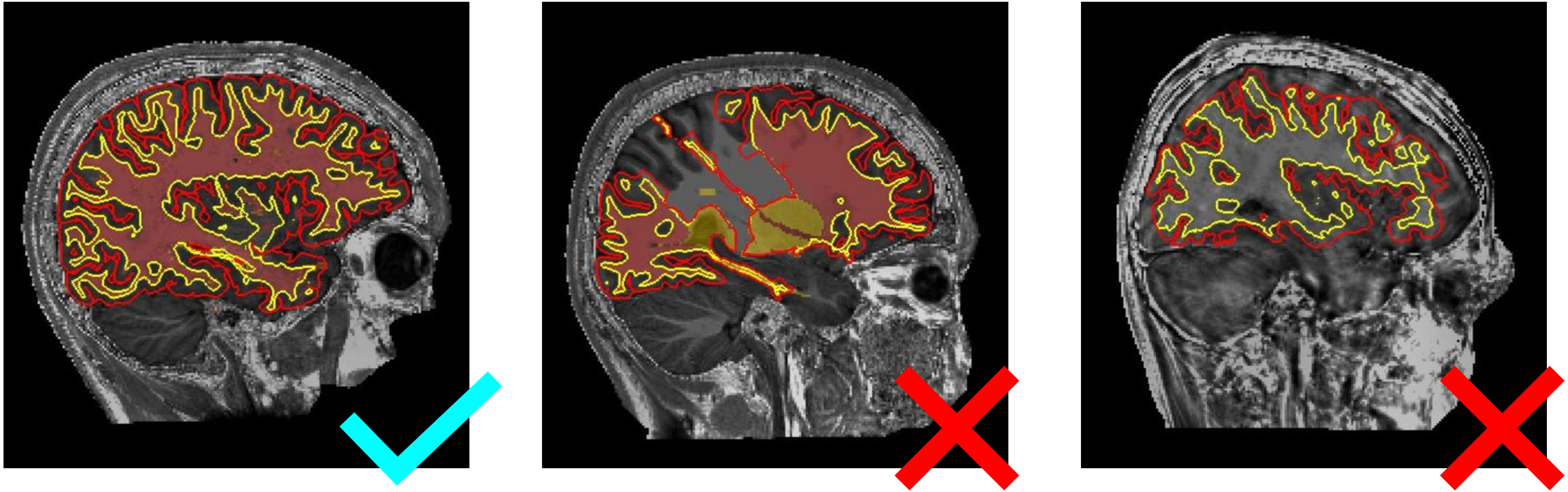
Is your code doing what you think it's doing? Does 2 + 2 = 4?

```
Assert.AreEqual(
          GetTimeOfDay(),
          “Morning” )
```

https://www.toptal.com/qa/how-to-write-testable-code-and-why-it-matters

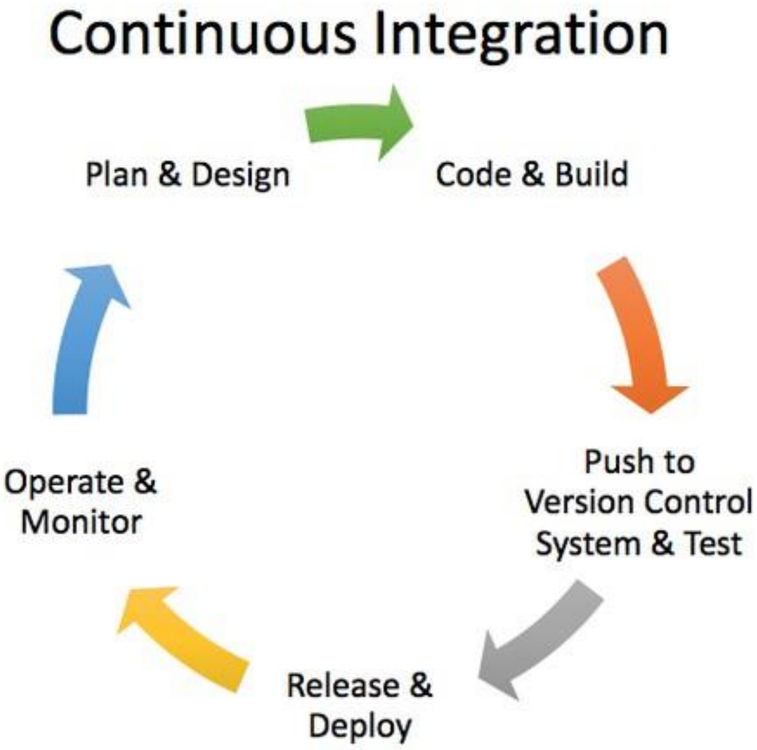
Testing (aka making explicit sanity checks)

Is your code doing what you think it's doing? Does 2 + 2 = 4?



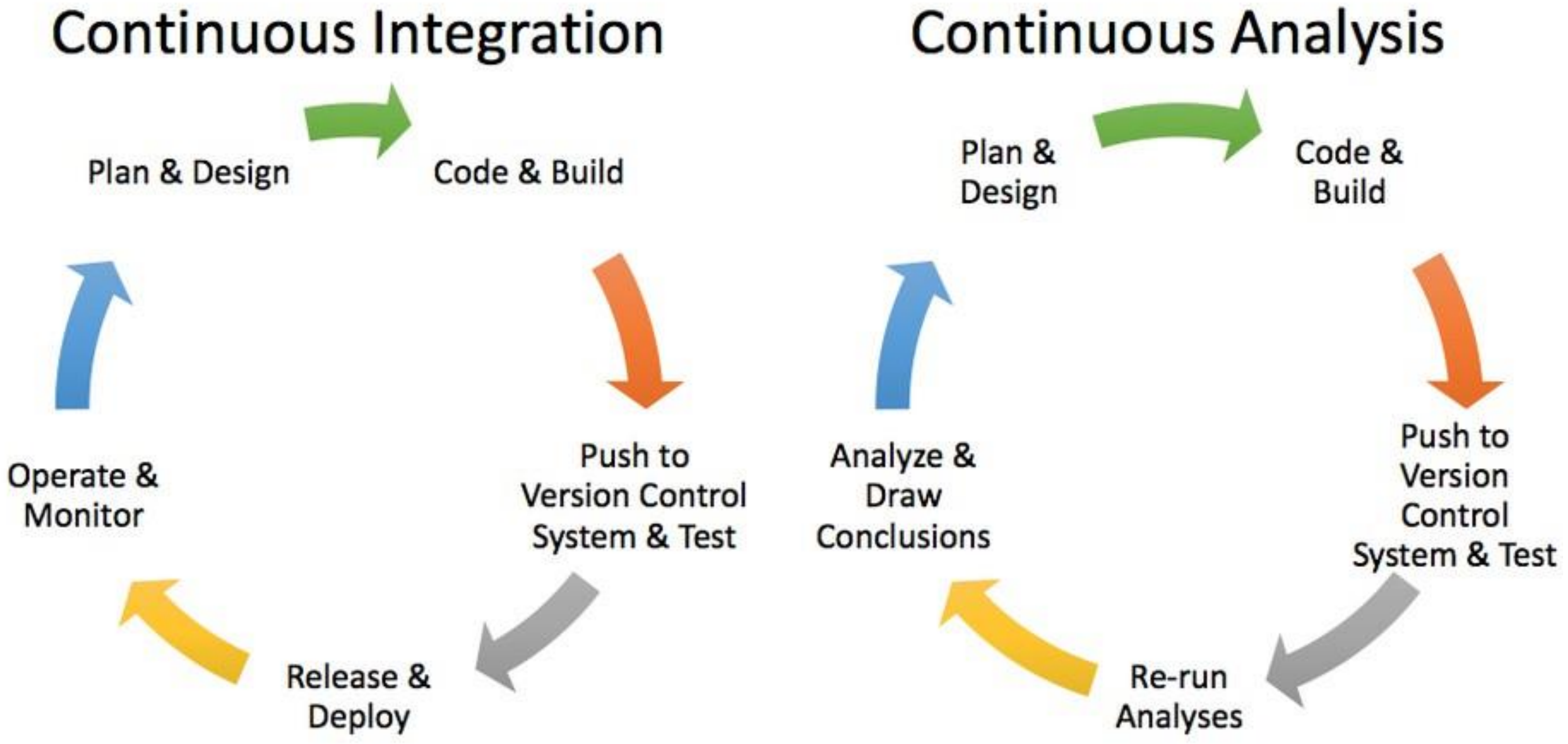
A very simple check: Is total brain volume within an expected range?

Continuous integration for research



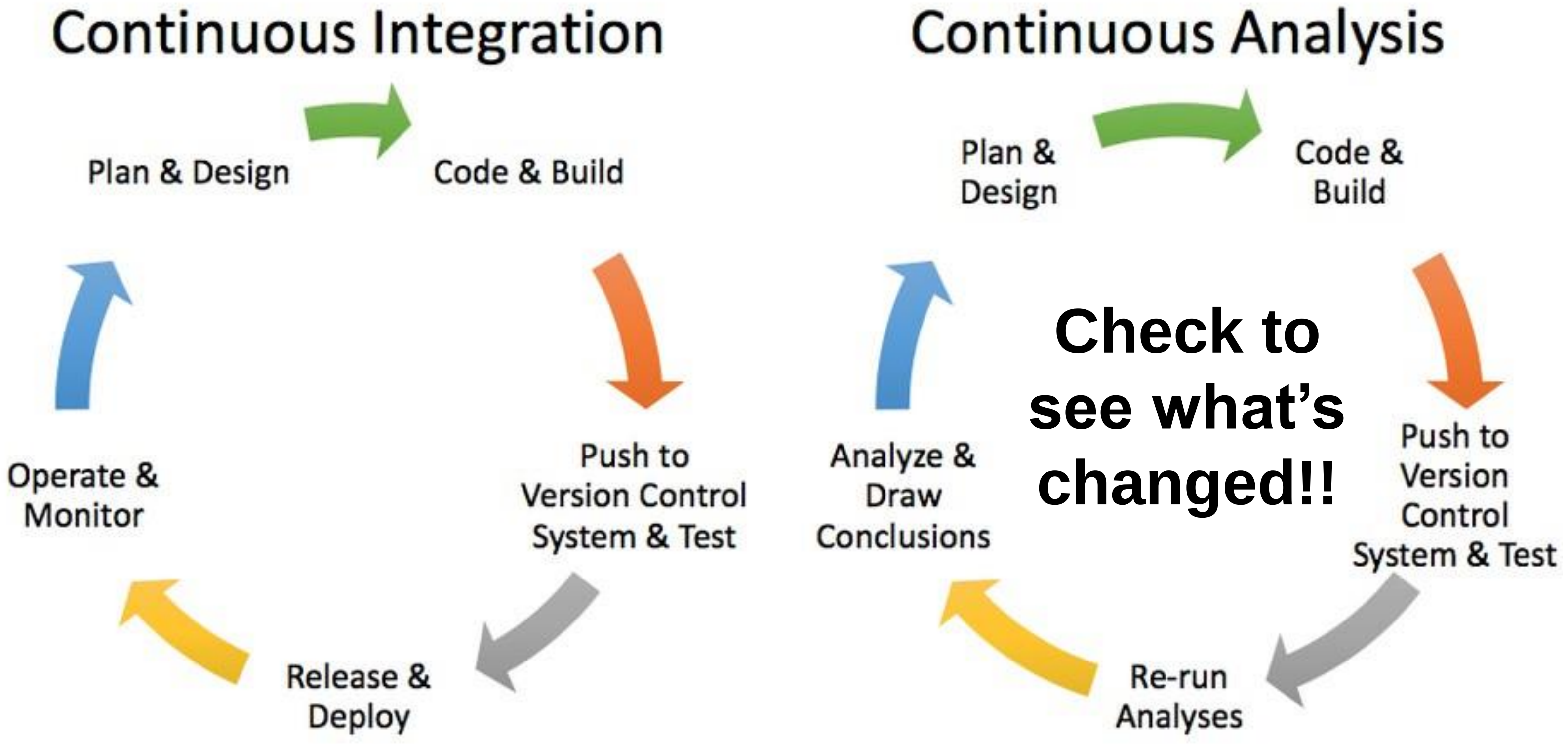
https://elifesciences.org/labs/e623676c/reproducibility-automated

Continuous integration for research



https://elifesciences.org/labs/e623676c/reproducibility-automated

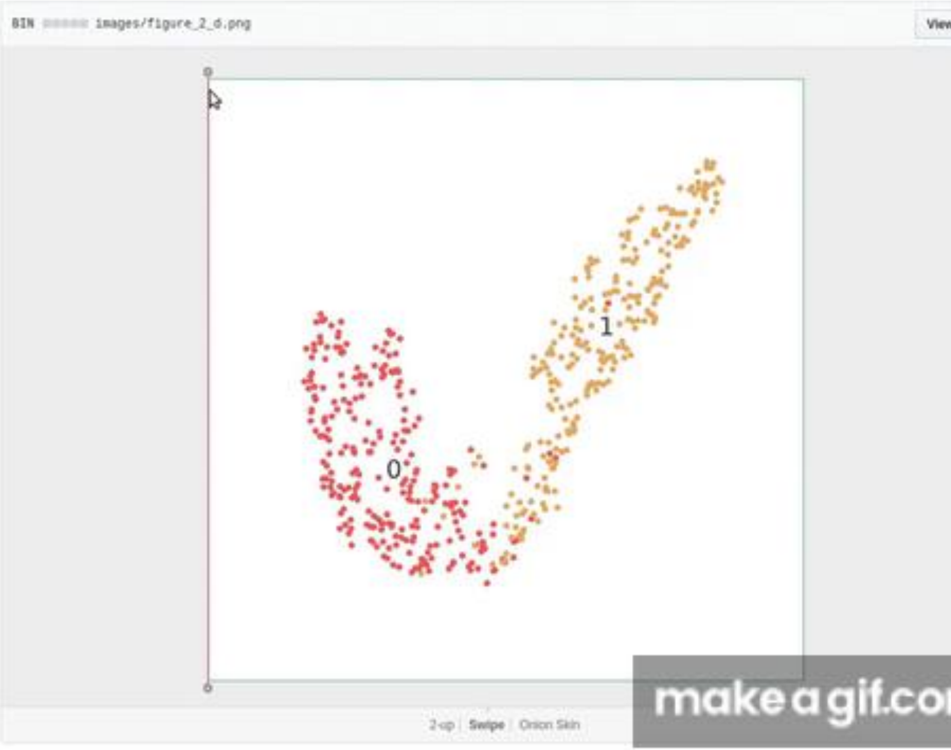
Continuous integration for research

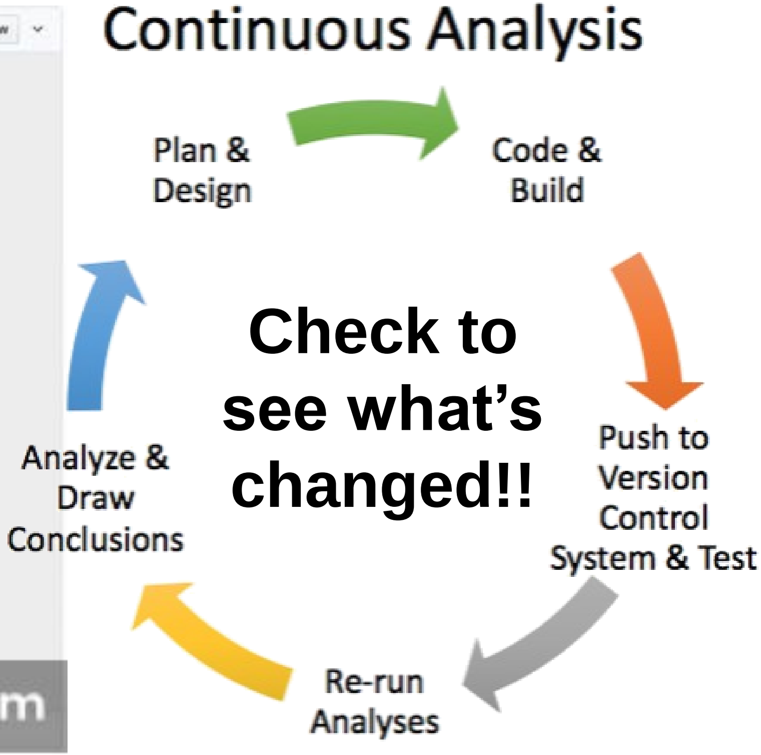


Check to see what’s changed!!

https://elifesciences.org/labs/e623676c/reproducibility-automated

Continuous integration for research




Continuous Analysis
Plan & Design
Code & Build
Push to Version Control System & Test
Re-run Analyses
Analyze & Draw Conclusions

Check to see what’s changed!!

https://elifesciences.org/labs/e623676c/reproducibility-automated

Held to higher standards than others

Make reproducibility, "too easy not to do"

Share the responsibility of reproducibility

https://github.com/alan-turing-institute/the-turing-way/blob/master/book_skeleton.md

Checklists for researcher, PI and admin team

- Researcher
  - Version control
  - Capturing compute environment
  - Writing and running the code

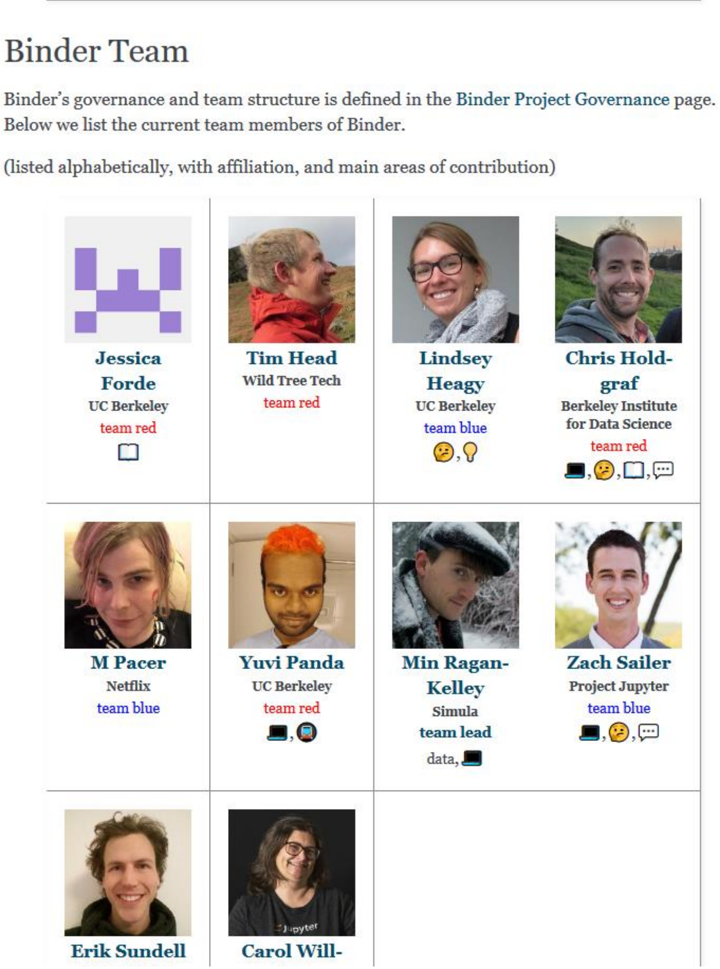
- PI
  - Results presented are those from the final run of the analysis
  - Check that another researcher can run the code

- Admin
  - Version control
  - Data and code archive
  - Open access publication

https://github.com/alan-turing-institute/the-turing-way/blob/master/book_skeleton.md

Interactive checks

- Binder to the rescue!
- Repo2docker: capture the compute environment and builds a container
- Send to cloud resources
- Open a link in a browser and run the code!



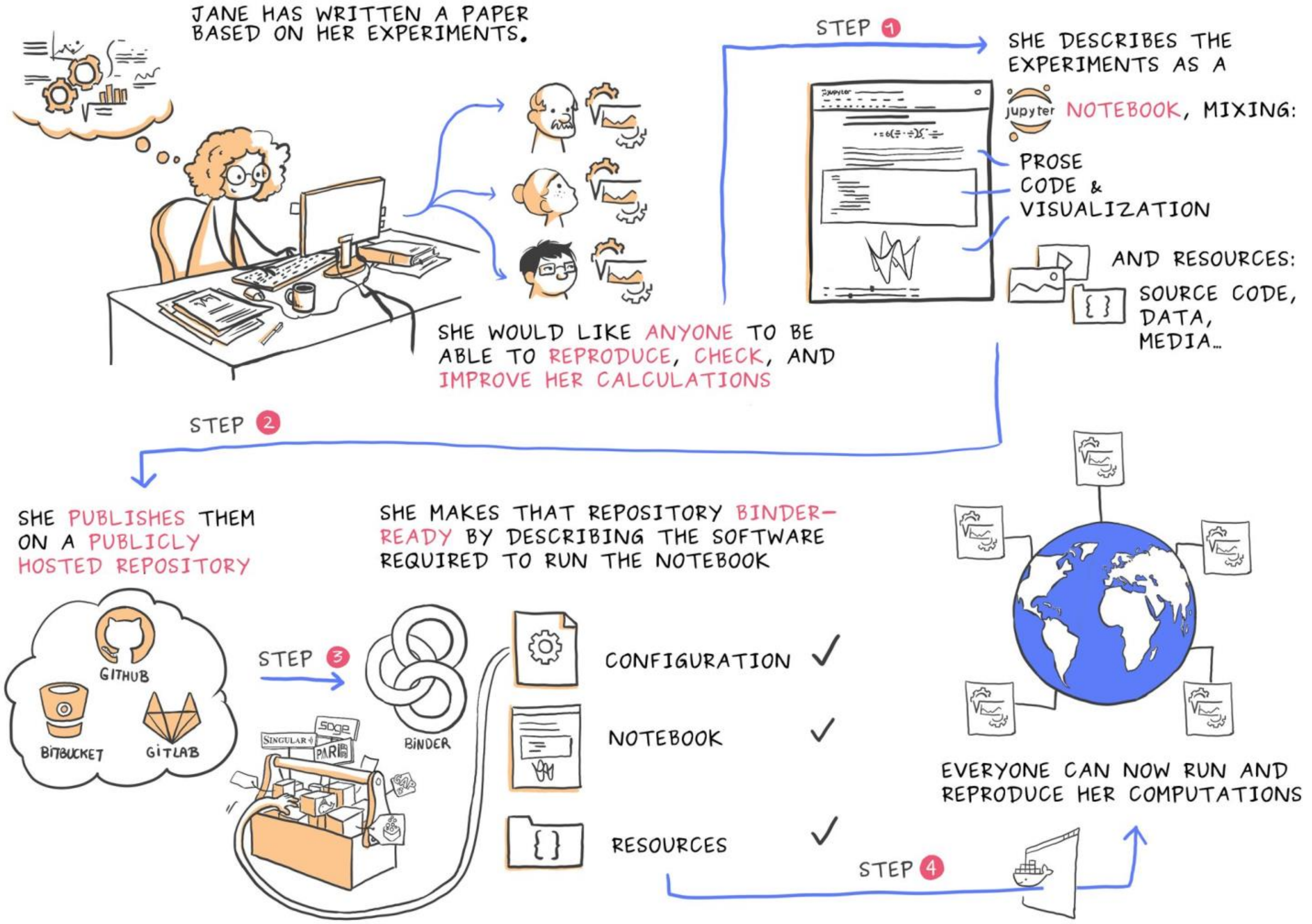
Binder Team

Binder's governance and team structure is defined in the Binder Project Governance page. Below we list the current team members of Binder.

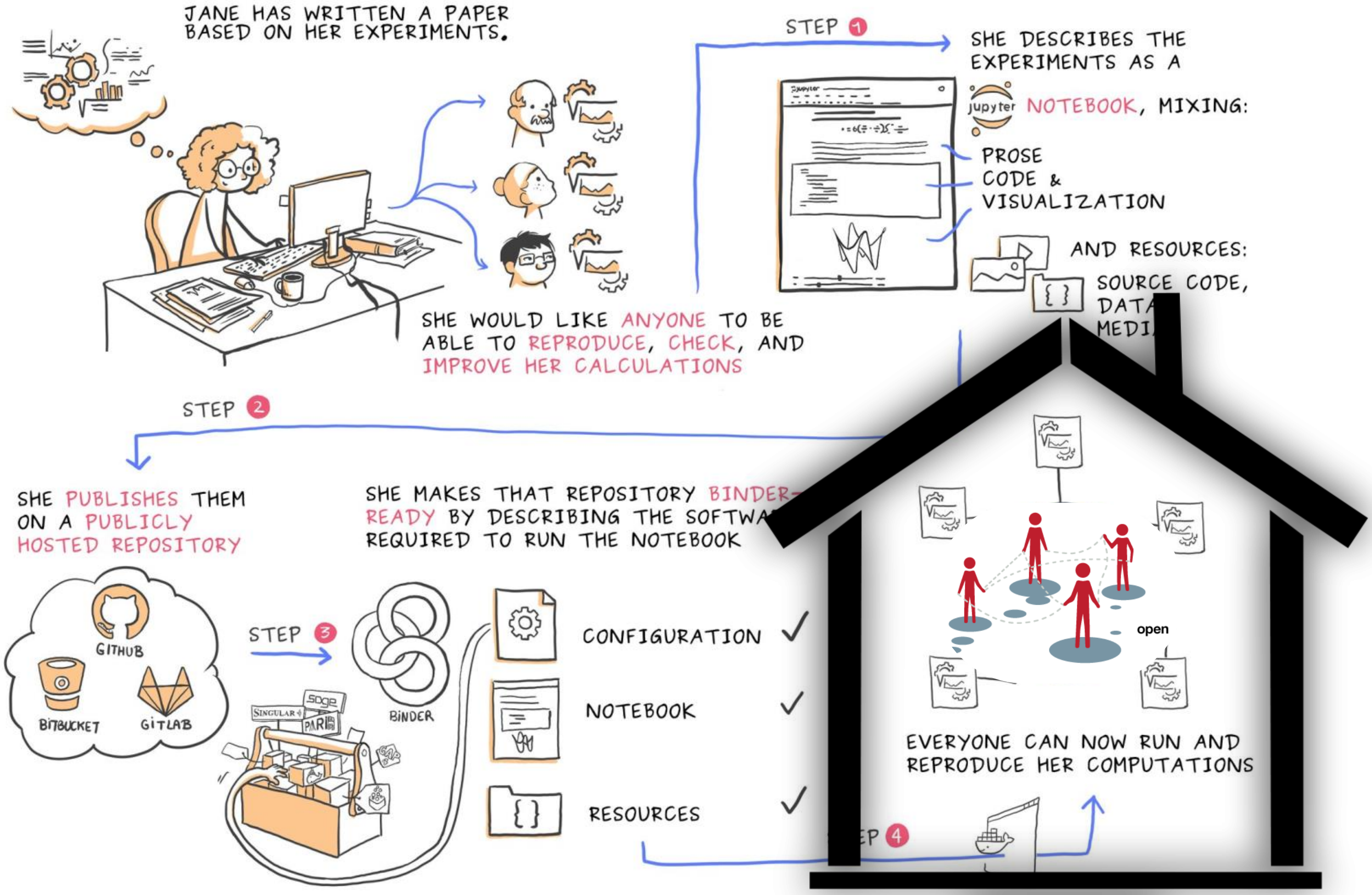
(listed alphabetically, with affiliation, and main areas of contribution)

Jessica Forde UC Berkeley team red 📖	Tim Head Wild Tree Tech team red	Lindsey Heagy UC Berkeley team blue 😕,💡	Chris Holdgraf Berkeley Institute for Data Science team red 💻,😕,📖,💬
M Pacer Netflix team blue	Yuvi Panda UC Berkeley team red 💻,🚇	Min Ragan-Kelley Simula team lead data,💻	Zach Sailer Project Jupyter team blue 💻,😕,💬
Erik Sundell	Carol Will-		

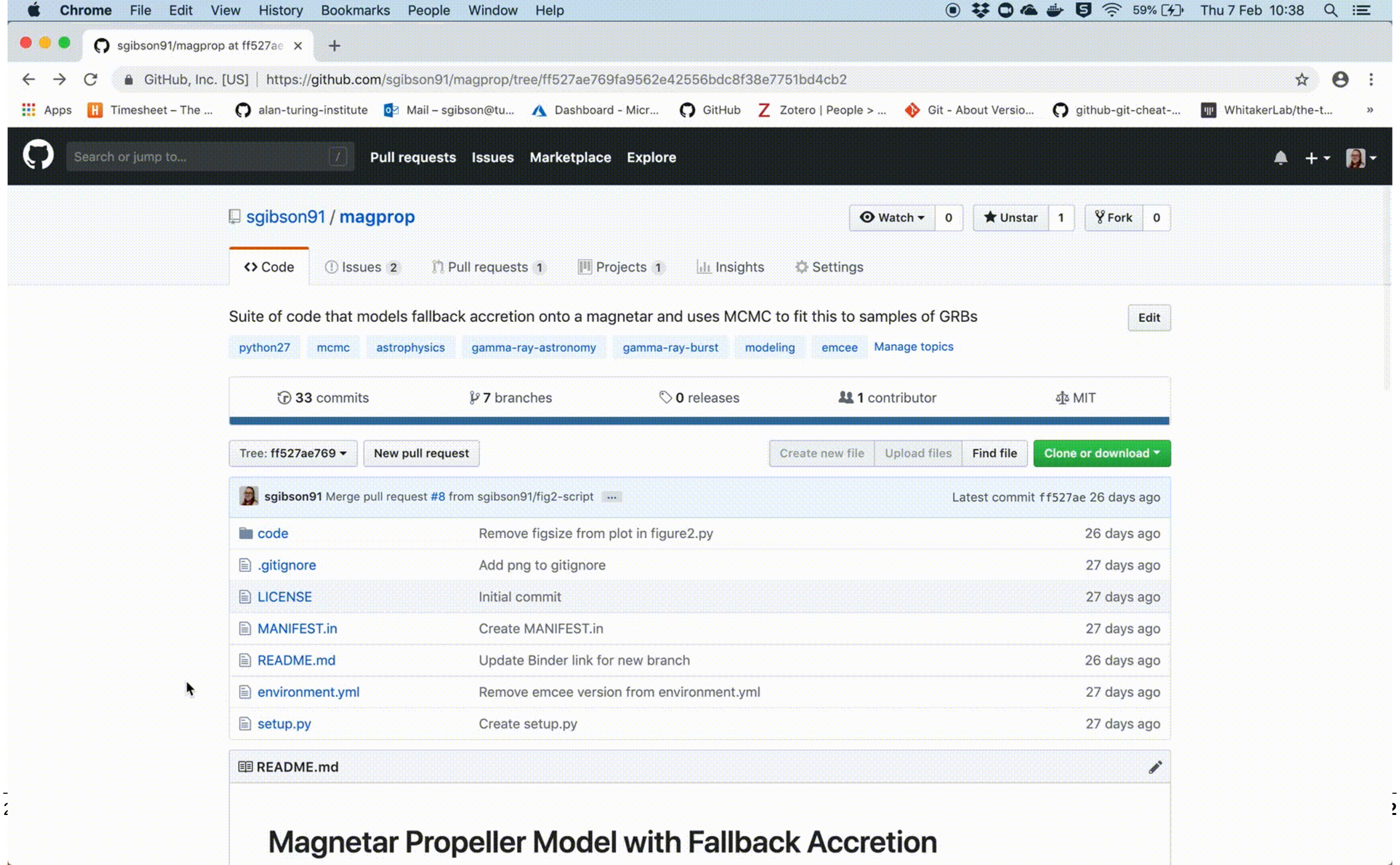
https://jupyterhub-team-compass.readthedocs.io/en/latest/team.html#binder-team


JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA,
MEDIA...
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWARE REQUIRED TO RUN THE NOTEBOOK
BINDER
SINGULAR
sage
PARI
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
STEP 4
EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS

Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960


JANE HAS WRITTEN A PAPER BASED ON HER EXPERIMENTS.
SHE WOULD LIKE ANYONE TO BE ABLE TO REPRODUCE, CHECK, AND IMPROVE HER CALCULATIONS
STEP 1
SHE DESCRIBES THE EXPERIMENTS AS A
jupyter NOTEBOOK, MIXING:
PROSE
CODE &
VISUALIZATION
AND RESOURCES:
SOURCE CODE,
DATA
MEDI
STEP 2
SHE PUBLISHES THEM ON A PUBLICLY HOSTED REPOSITORY
GITHUB
BITBUCKET
GITLAB
STEP 3
SHE MAKES THAT REPOSITORY BINDER-READY BY DESCRIBING THE SOFTWA REQUIRED TO RUN THE NOTEBOOK
BINDER
SINGULAR
sage
PARI
CONFIGURATION ✓
NOTEBOOK ✓
RESOURCES ✓
EP 4
open
EVERYONE CAN NOW RUN AND REPRODUCE HER COMPUTATIONS

Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960



Suite of code that models fallback accretion onto a magnetar and uses MCMC to fit this to samples of GRBs

sgibson91 Merge pull request #8 from sgibson91/fig2-script ... Latest commit ff527ae 26 days ago

code	Remove figsize from plot in figure2.py	26 days ago
.gitignore	Add png to gitignore	27 days ago
LICENSE	Initial commit	27 days ago
MANIFEST.in	Create MANIFEST.in	27 days ago
README.md	Update Binder link for new branch	26 days ago
environment.yml	Remove emcee version from environment.yml	27 days ago
setup.py	Create setup.py	27 days ago

README.md

Magnetar Propeller Model with Fallback Accretion

Thank you

(Lets now discuss how we can work together!)

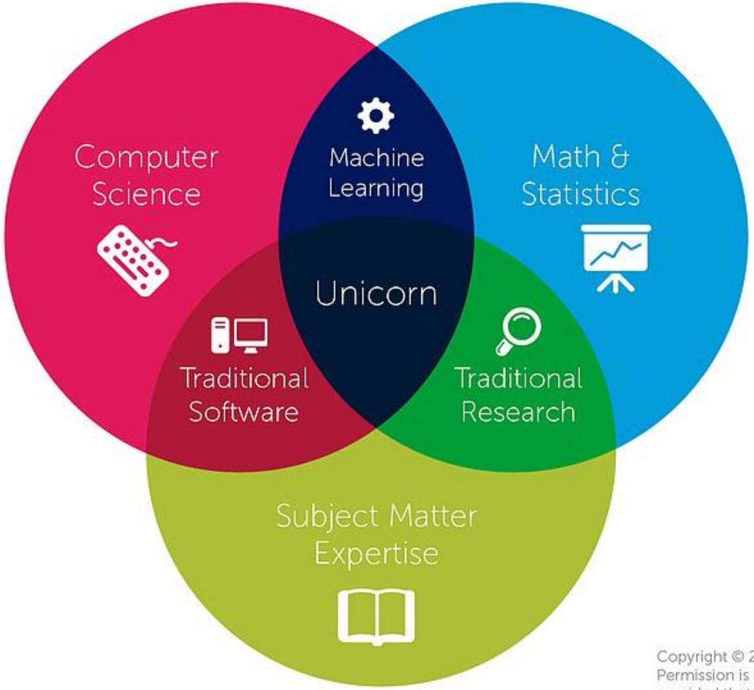
kwhitaker@turing.ac.uk
@kirstie_j
github.com/alan-turing-institute/the-turing-way
gitter.im/alan-turing-institute/the-turing-way

doi: 10.5281/zenodo.2599904

Building a culture of collaborative science

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

The Data Science Unicorn

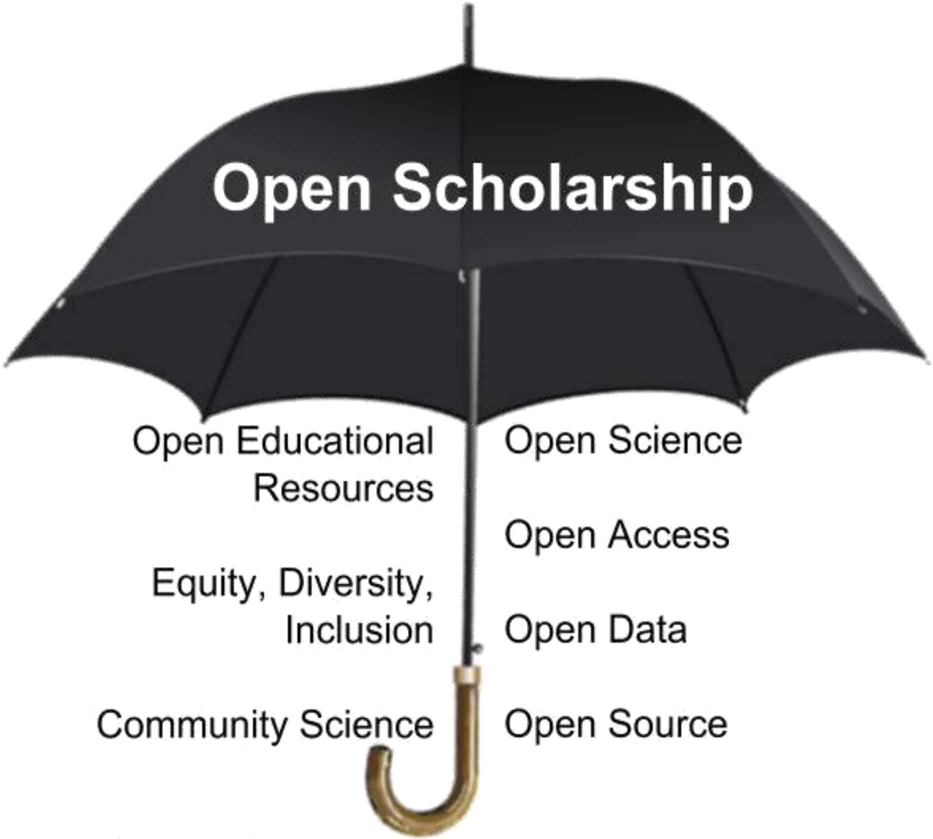


https://www.luther.edu/computer-science/data-science-major/why-study

How can we incentivise team science?

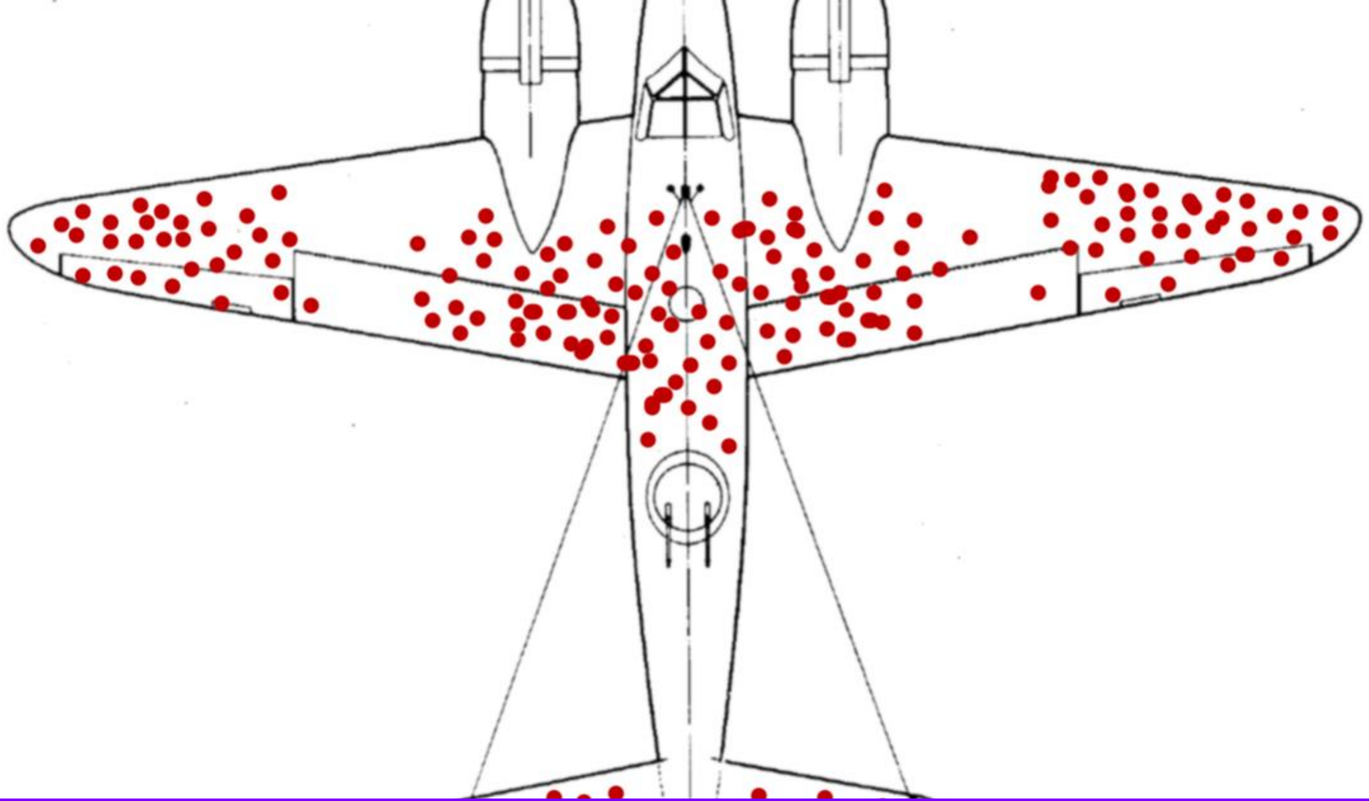
https://neurohackademy.org/apply

Open is so much more than reproducible



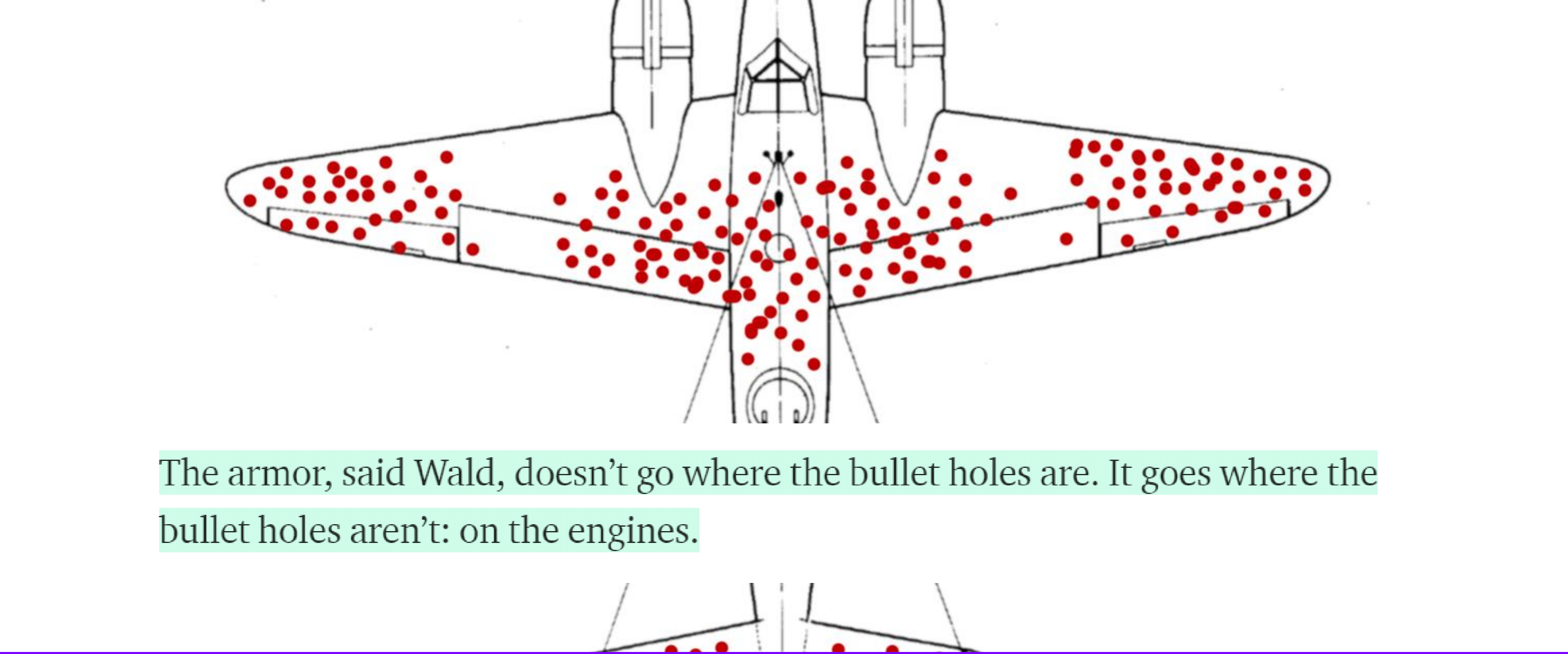
Adapted from: https://www.meetup.com/Berlin-Open-Science-Meetup/

Robin Champieux and Danielle Robinson



https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

https://doi.org/10.6084/m9.figshare.7564682

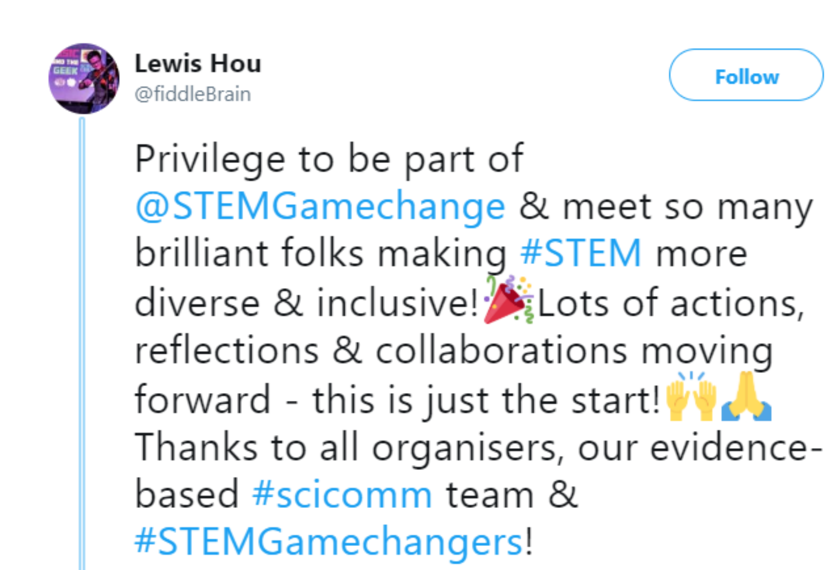


The armor, said Wald, doesn’t go where the bullet holes are. It goes where the bullet holes aren’t: on the engines.

https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

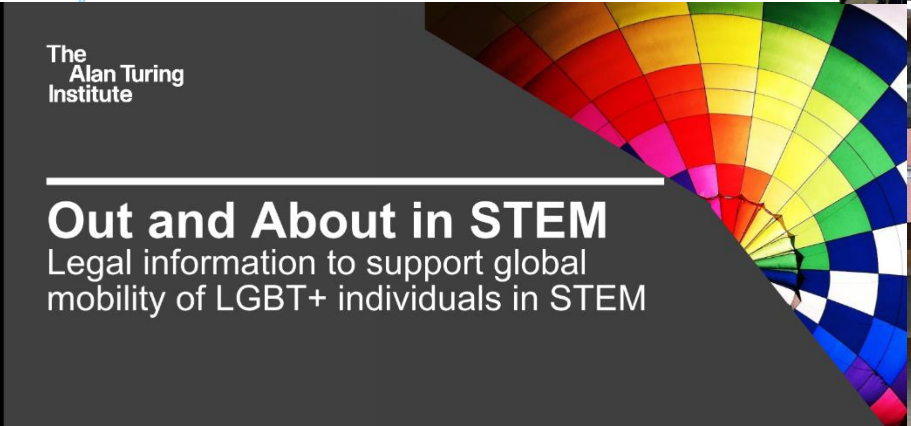
https://doi.org/10.6084/m9.figshare.7564682

moz://a



Lewis Hou
@fiddleBrain

Privilege to be part of @STEMGamechange & meet so many brilliant folks making #STEM more diverse & inclusive!🎉Lots of actions, reflections & collaborations moving forward - this is just the start!🙌🙏 Thanks to all organisers, our evidence-based #scicomm team & #STEMGamechangers!



The
Alan Turing
Institute

Out and About in STEM

Legal information to support global mobility of LGBT+ individuals in STEM

https://stemgamechangers.github.io

Data science at scale

ttps://www.flickr.com/photos/mozfest/22455631157/in/album-72157658649418943

Thank you!

The Alan Turing Institute

UNIVERSITY OF CAMBRIDGE

moz://a

Please come and join us!

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

gitter.im/alan-turing-institute/the-turing-way

@kirstie_j, @whitakerlab

doi: 10.6084/m9.figshare.7649156

ttps://www.flickr.com/photos/mozfest/22455631157/in/album-72157658649418943