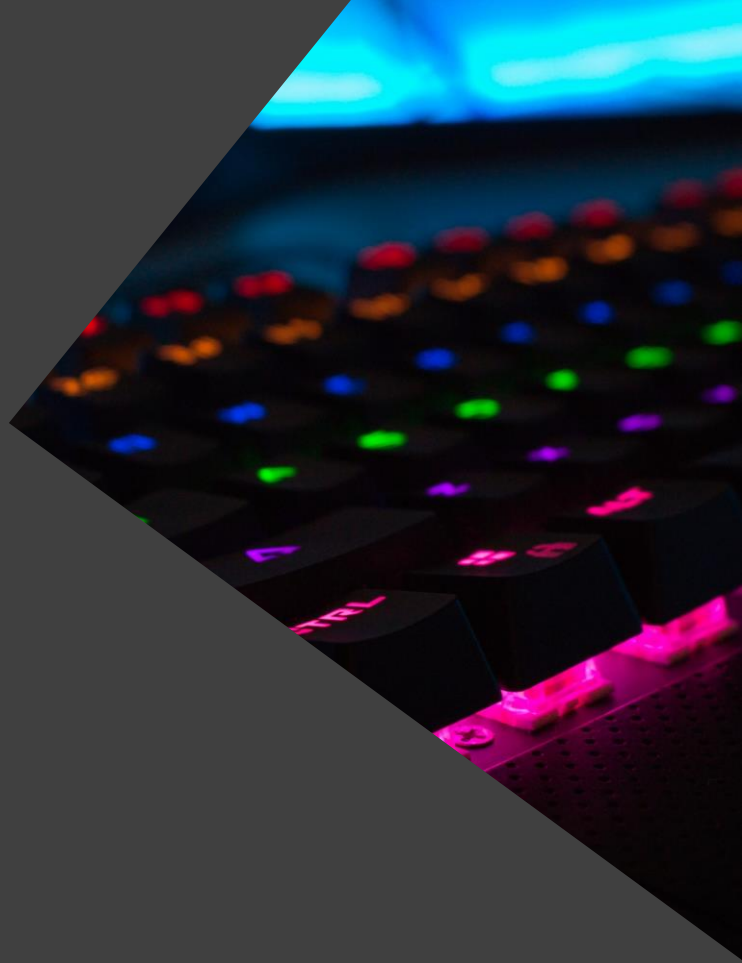
The Alan Turing Institute



The Turing Way

A how to guide for reproducible research

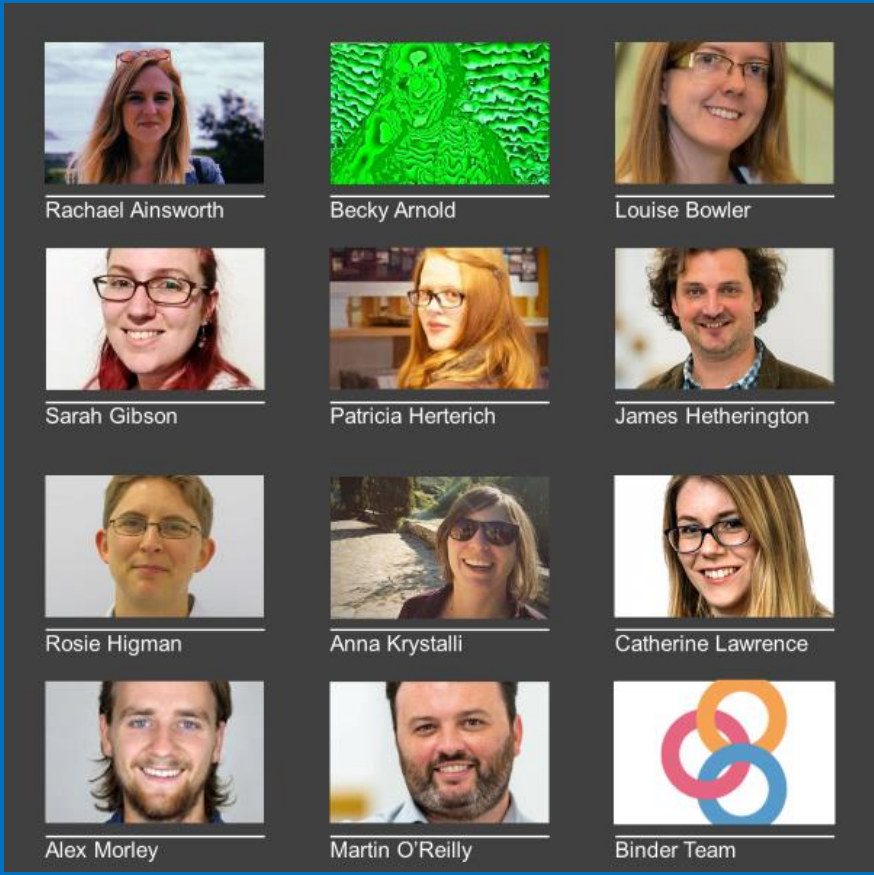
Kirstie Whitaker

#MQDataScience #TuringWay @kirstie_j

Slides available, CC-BY, at https://doi.org/10.5281/zenodo.3402510

The Turing Way is:

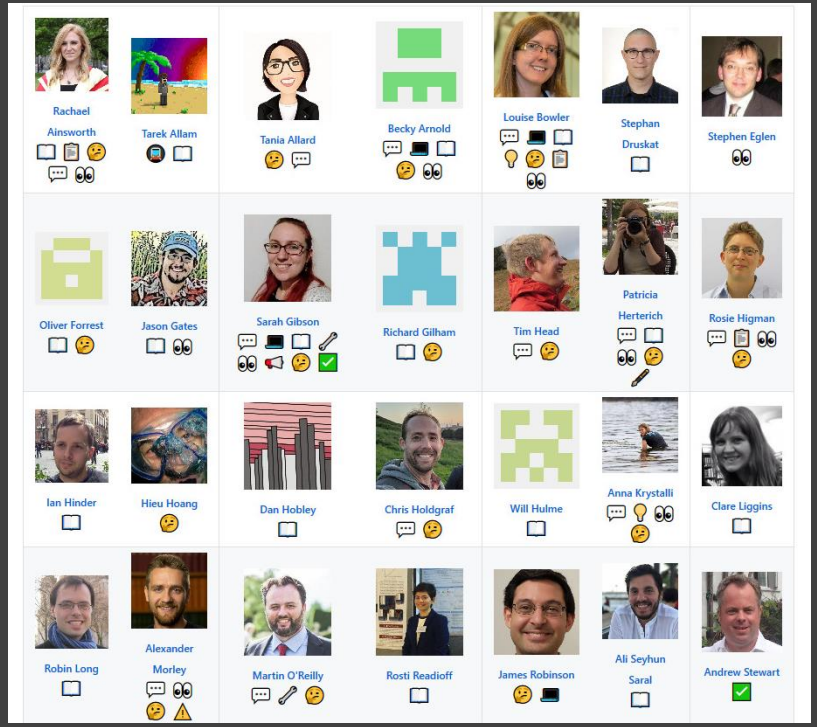
- a book
- a community
- a global collaboration
- a whole tonne of work

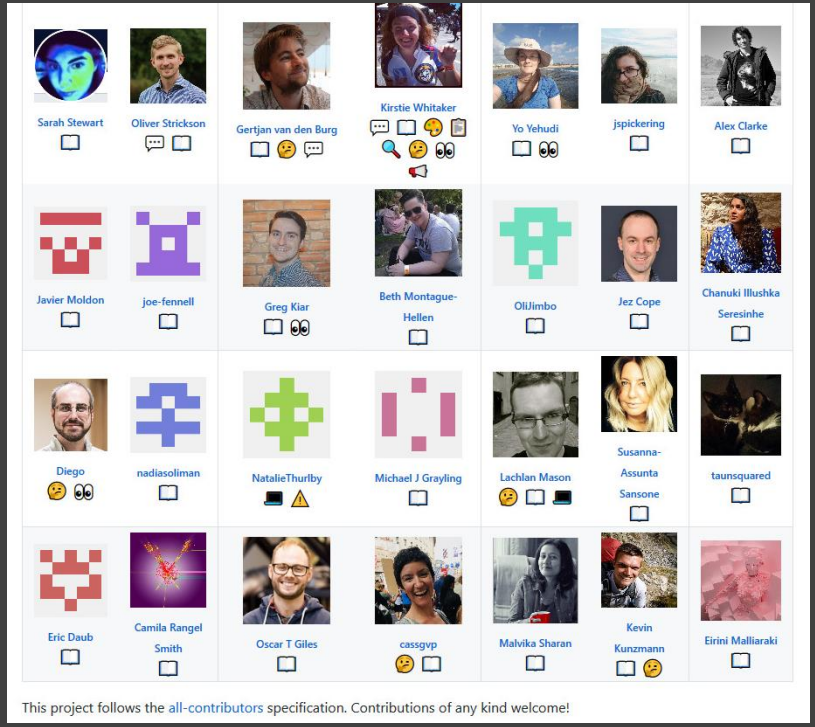


#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

Thank you to all our contributors





https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Picture credit: Chris Gorgolewski
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Academic errors have real world effects

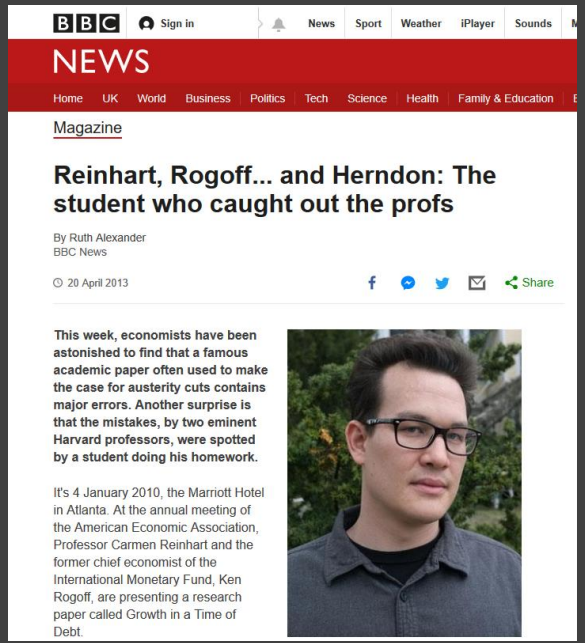
	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
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)	

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

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Academic 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

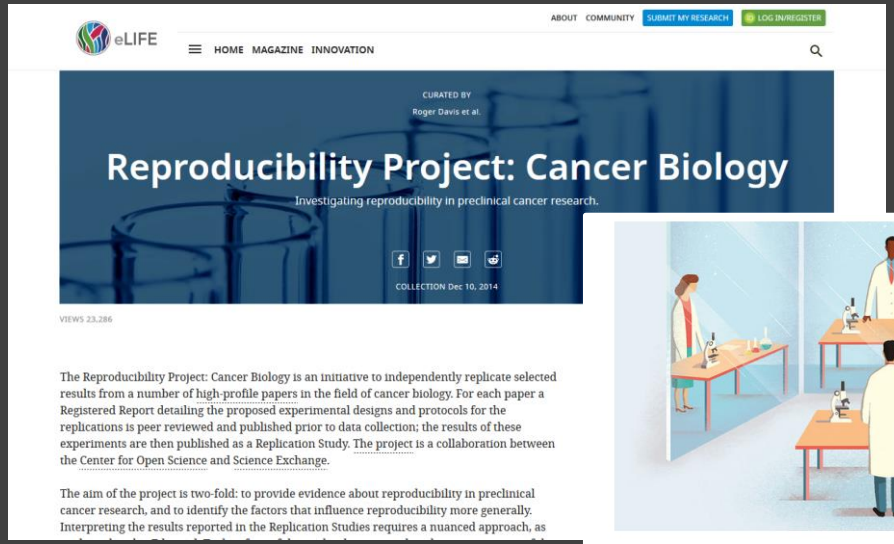
This week, economists have been astonished to find that a famous academic paper often used to make the case for austerity cuts contains major errors. Another surprise is that the mistakes, by two eminent Harvard professors, were spotted by a student doing his homework.

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.

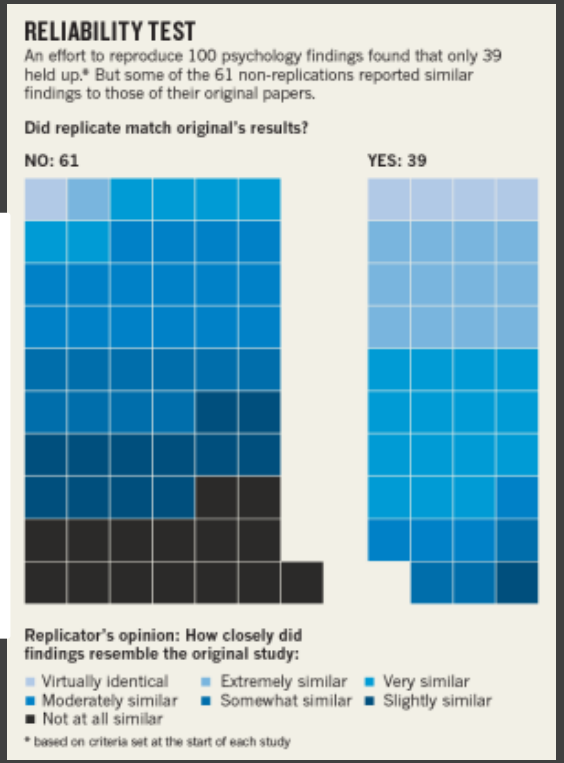
https://statmodeling.stat.columbia.edu/2013/04/16/memo-to-reinhart-and-rogoff-i-think-its-best-to-admit-your-errors-and-go-on-from-there
https://www.bbc.co.uk/news/magazine-22223190

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Explicitly replicating research is really hard



Plan to replicate 50 high-impact cancer papers shrinks to just 18



https://elifesciences.org/collections/9b1e83d1/reproducibility-project-cancer-biology
https://www.sciencemag.org/news/2018/07/plan-replicate-50-high-impact-cancer-papers-shrinks-just-18
https://www.nature.com/news/over-half-of-psychology-studies-fail-reproducibility-test-1.18248

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Fraud is not our biggest problem

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://the-turing-way.netlify.com/reproducibility/03/definitions.html
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Barriers to reproducible research

Is not considered for promotion

Held to higher standards than others

Publication bias towards novel findings

Requires additional skills

Plead the 5th

Support additional users

Takes time

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

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

The Turing Way

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

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.

A bit more background

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

https://the-turing-way.netlify.com/introduction/introduction
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

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.

A bit more background

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

https://the-turing-way.netlify.com/introduction/introduction
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

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.

A bit more background

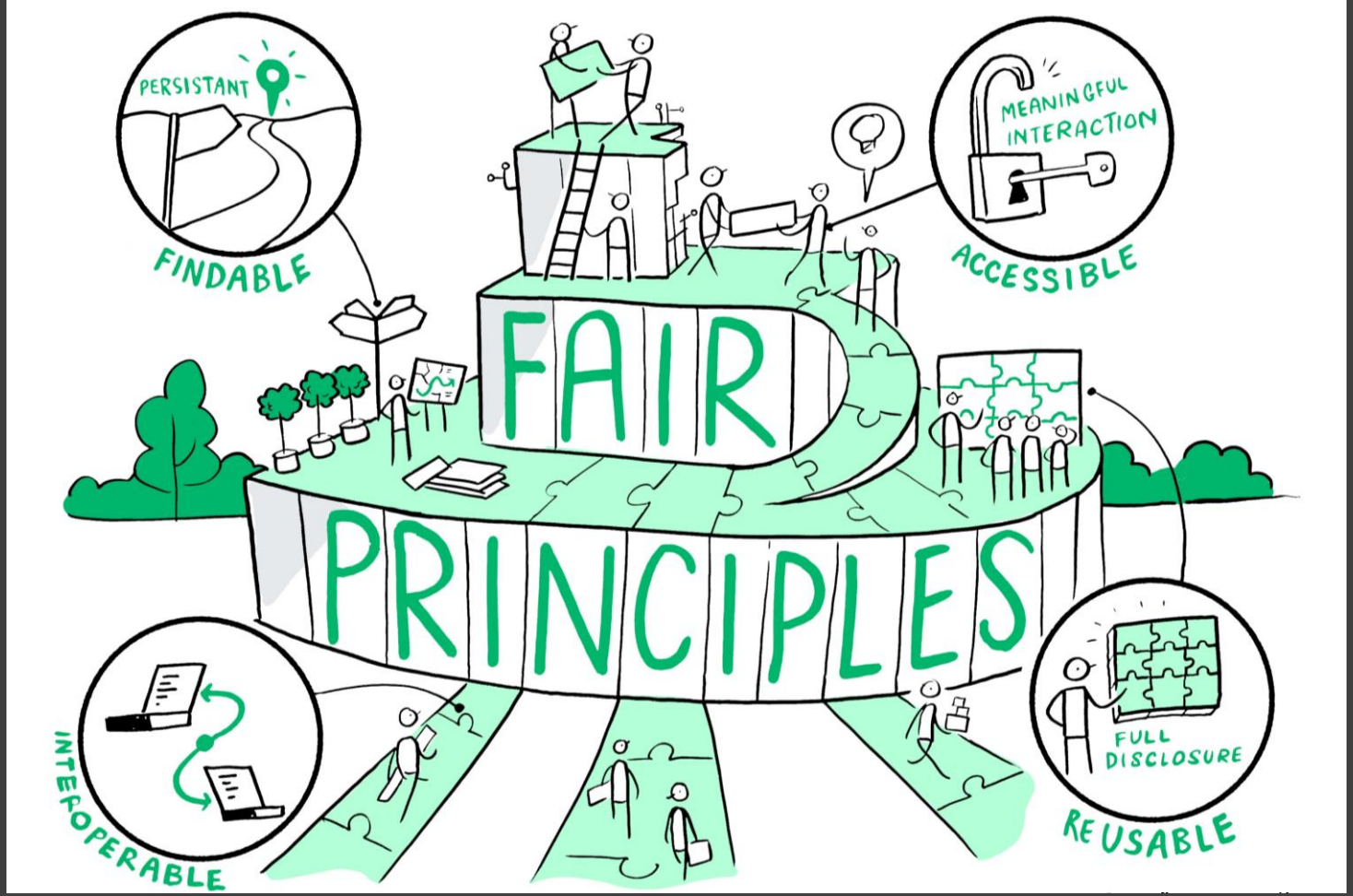
Reproducible research is necessary to ensure that scientific work can be trusted. Funders and publishers are beginning to require that publications include access to the underlying data and the analysis code. The goal is to ensure that all results can be independently verified and built upon in future 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

https://the-turing-way.netlify.com/introduction/introduction
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

FAIR….but not open

https://doi.org/10.5281/zenodo.3402510



#MQDataScience #TuringWay @kirstie_j

Wilkinson et al., Sci. Data, 2016. doi: 10.1038/sdata.2016.18

https://doi.org/10.5281/zenodo.3402510

Champion: Terry Lyons

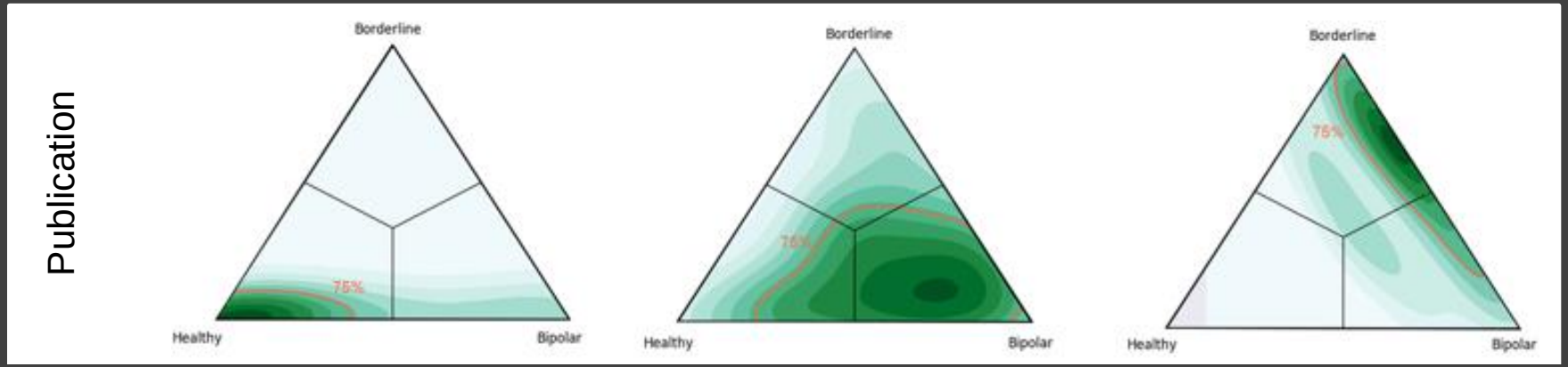
A signature-based machine learning model for bipolar disorder and borderline personality disorder

Imanol Perez Arribas, Guy Goodwin, John Geddes, Terry Lyons, Kate Saunders

https://github.com/alan-turing-institute/signatures-psychiatry
https://doi.org/10.1038/s41398-018-0334-0

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Publication



https://github.com/alan-turing-institute/signatures-psychiatry
https://doi.org/10.1038/s41398-018-0334-0

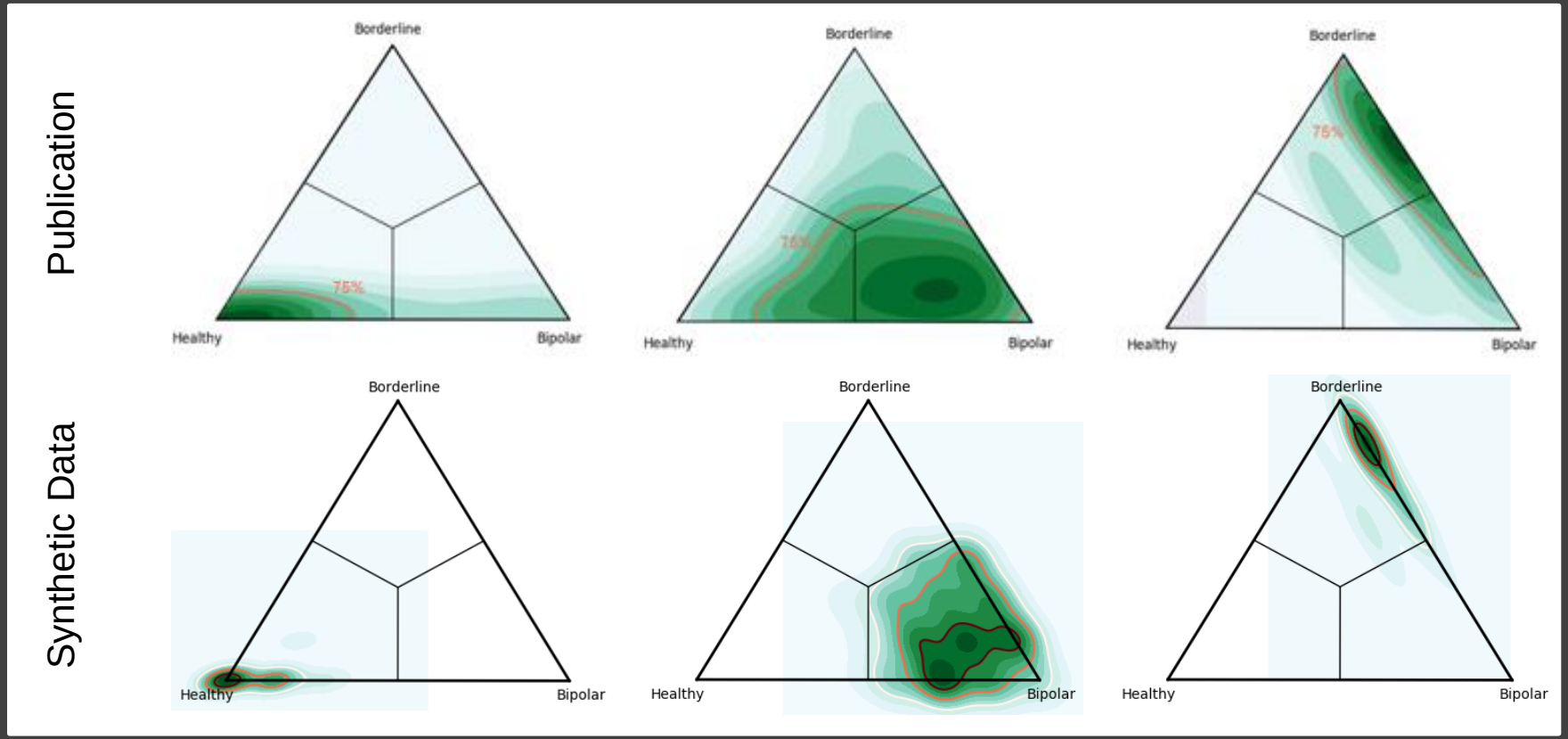
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Imanol Pérez Arribas

“We can’t share the data. The original researchers did not ask for consent to do so.

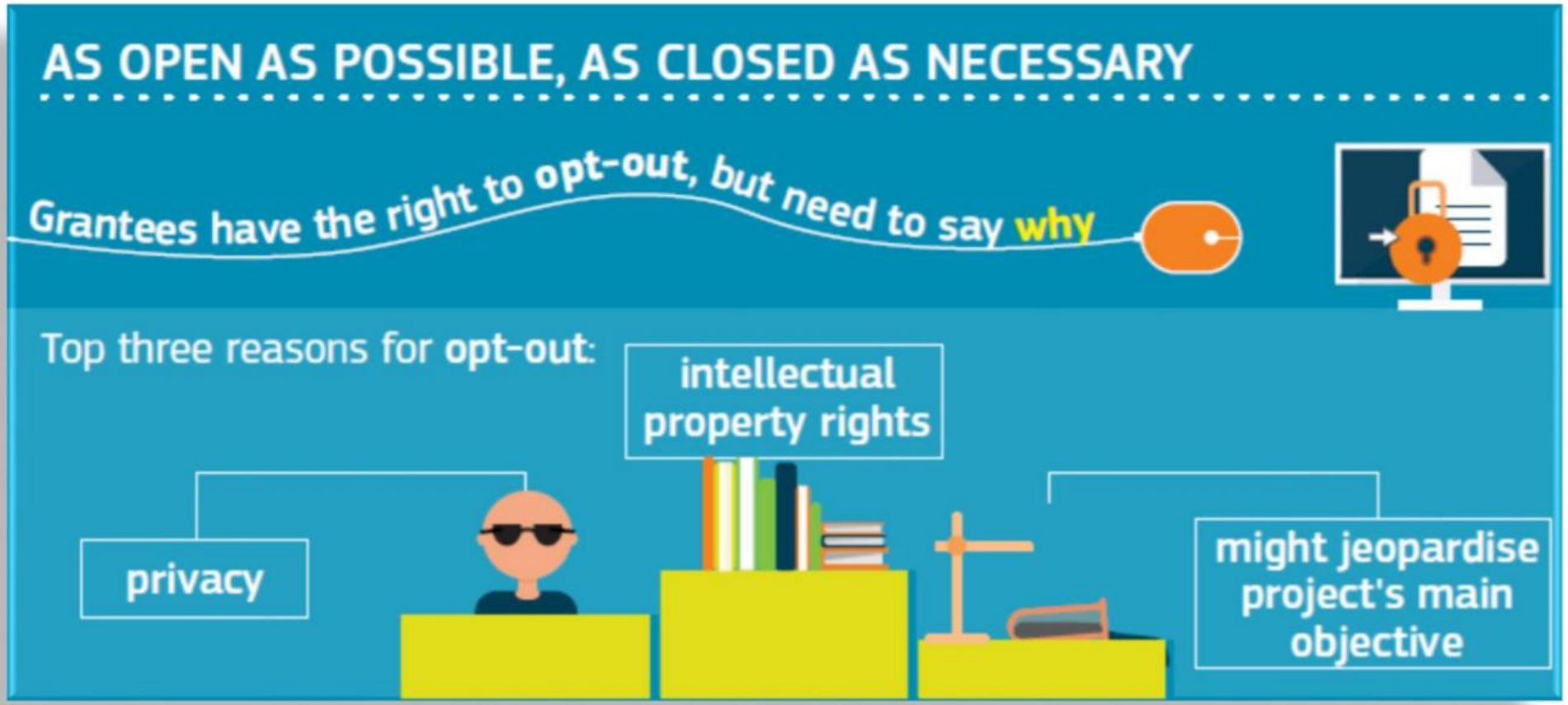
We can share simulated and synthetic data so that researchers can feel confident in applying our method to their own data.”

https://www.maths.ox.ac.uk/people/imanol.perez
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510



https://github.com/alan-turing-institute/signatures-psychiatry
https://doi.org/10.1038/s41398-018-0334-0

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510



https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf

#MQDataScience #TuringWay @kirstie_j

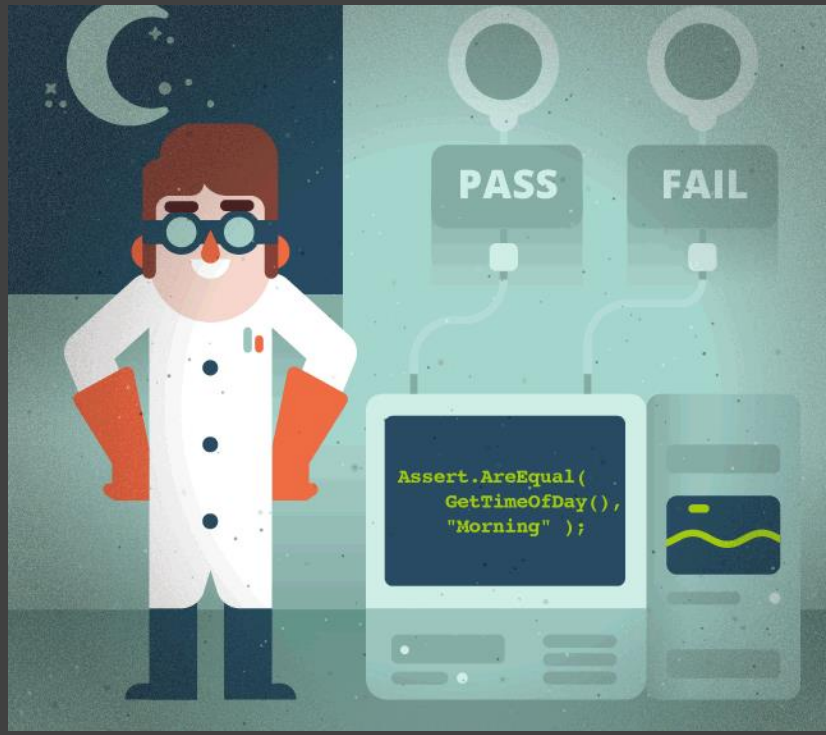
https://doi.org/10.5281/zenodo.3402510

Testing for research

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

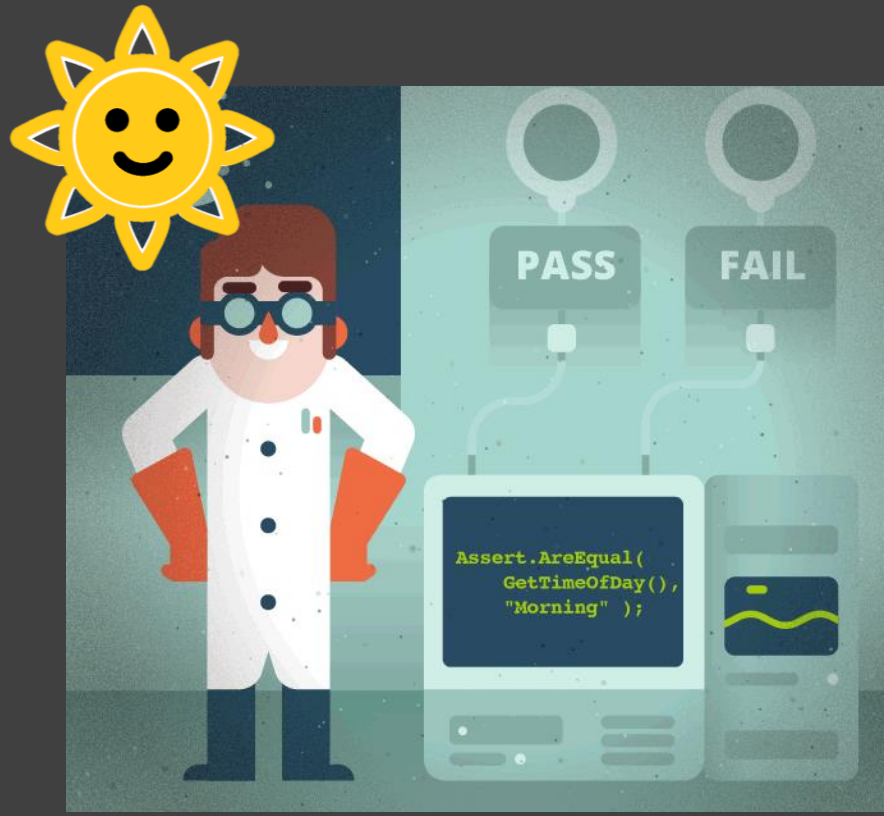
Is your code doing what you think its doing?



https://www.toptal.com/qa/how-to-write-testable-code-and-why-it-matters
https://the-turing-way.netlify.com/testing/testing.html

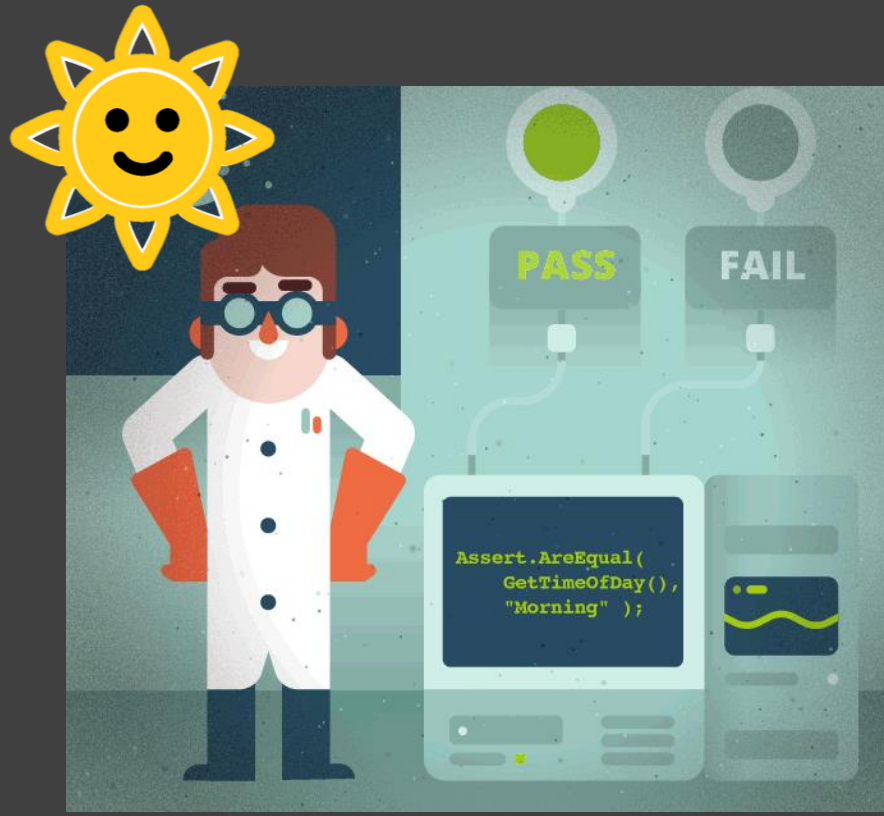
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Is your code doing what you think its doing?



Is your code doing what you think its doing?

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

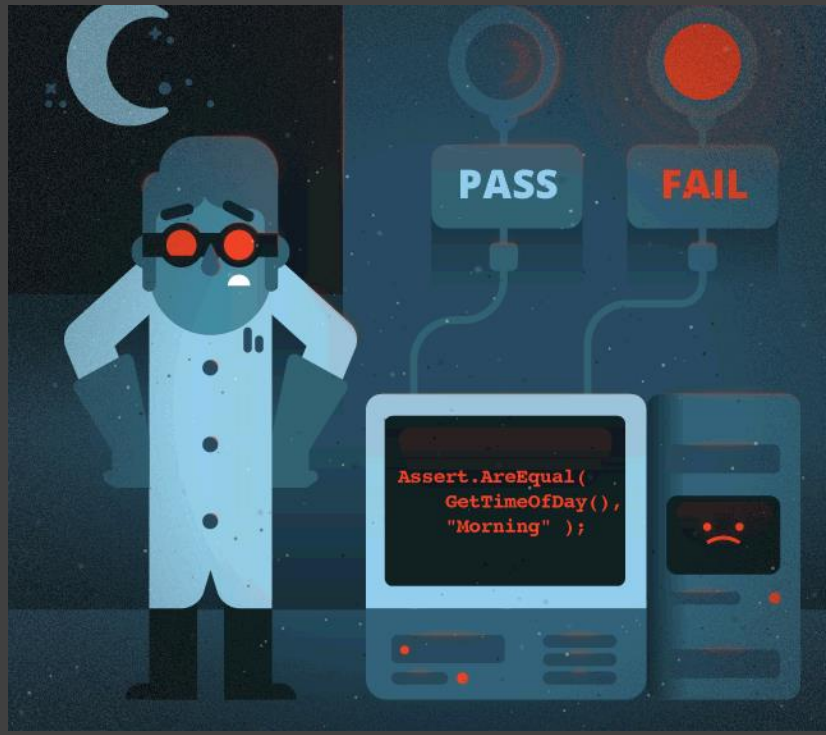


https://www.toptal.com/qa/how-to-write-testable-code-and-why-it-matters
https://the-turing-way.netlify.com/testing/testing.html

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Is your code doing what you think its doing?

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



Louise Bowler

“Add a test before you change anything.”

https://www.turing.ac.uk/people/researchers/louise-bowler
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

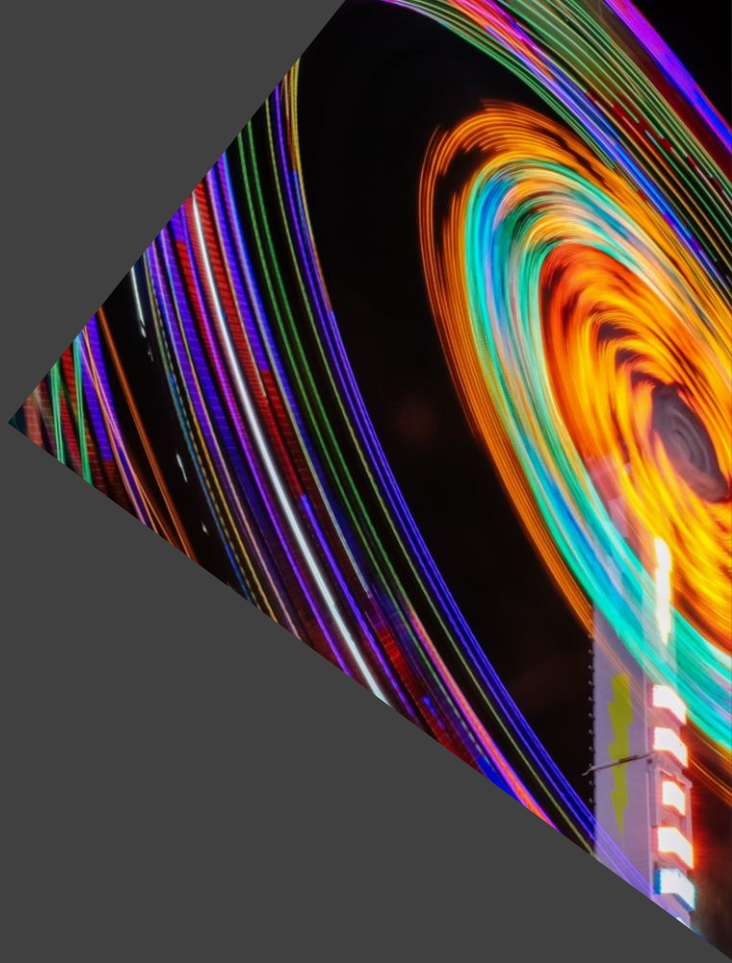
Louise Bowler

“Add a test before you change anything.

Particularly if you’re just going to tidy up your code before sharing it.”

https://www.turing.ac.uk/people/researchers/louise-bowler
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

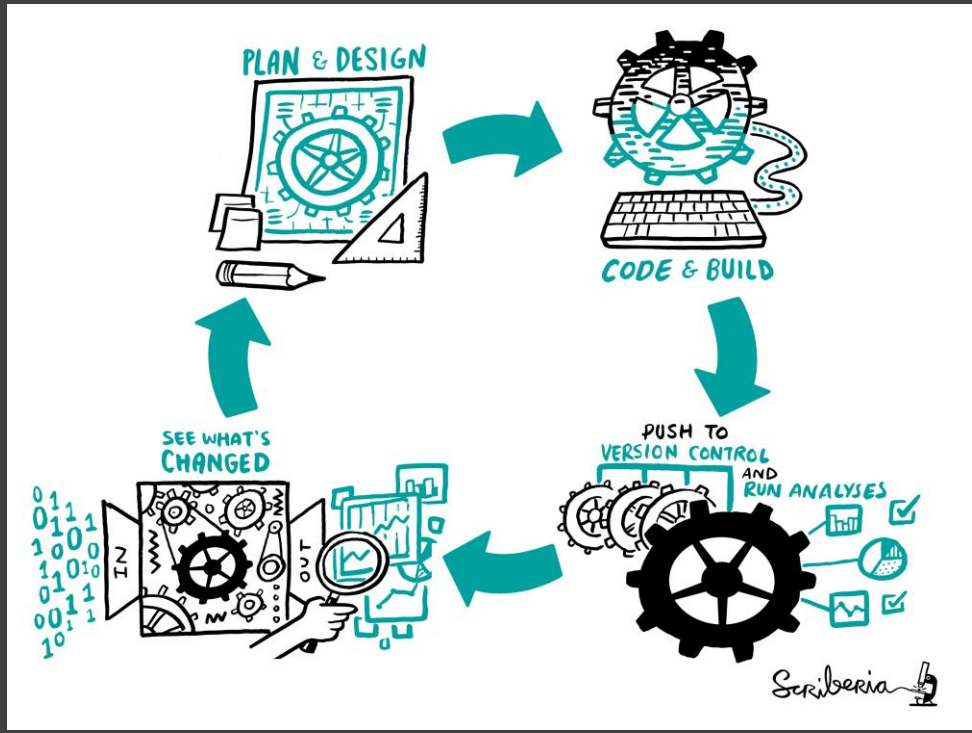
Continuous Analysis



#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

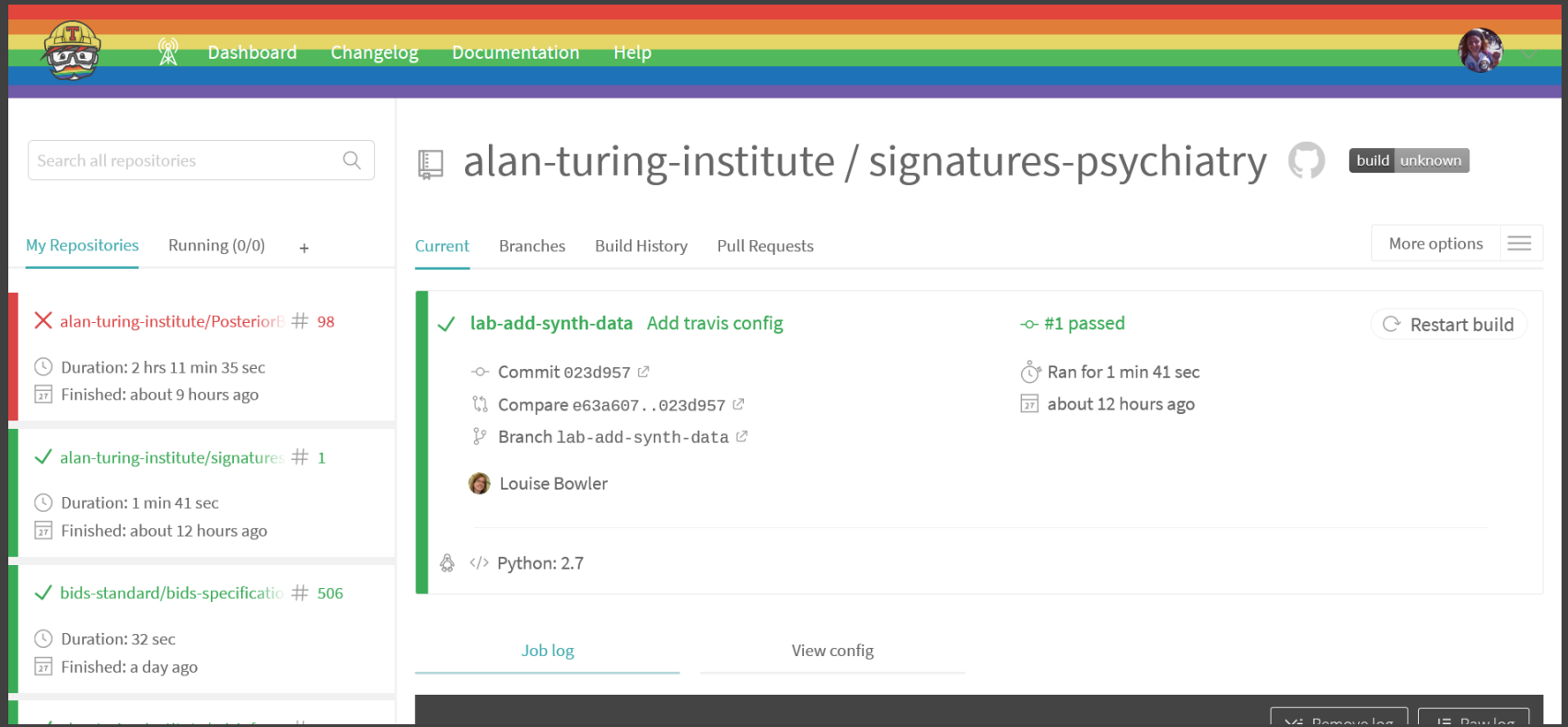
- Plan and design your experiment
- Write down those steps in code
- Push to version control and run the analyses
  - Traditionally done on the cloud, but the important part is that **all steps** are run **every time**
- Test to see what's changed



https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html

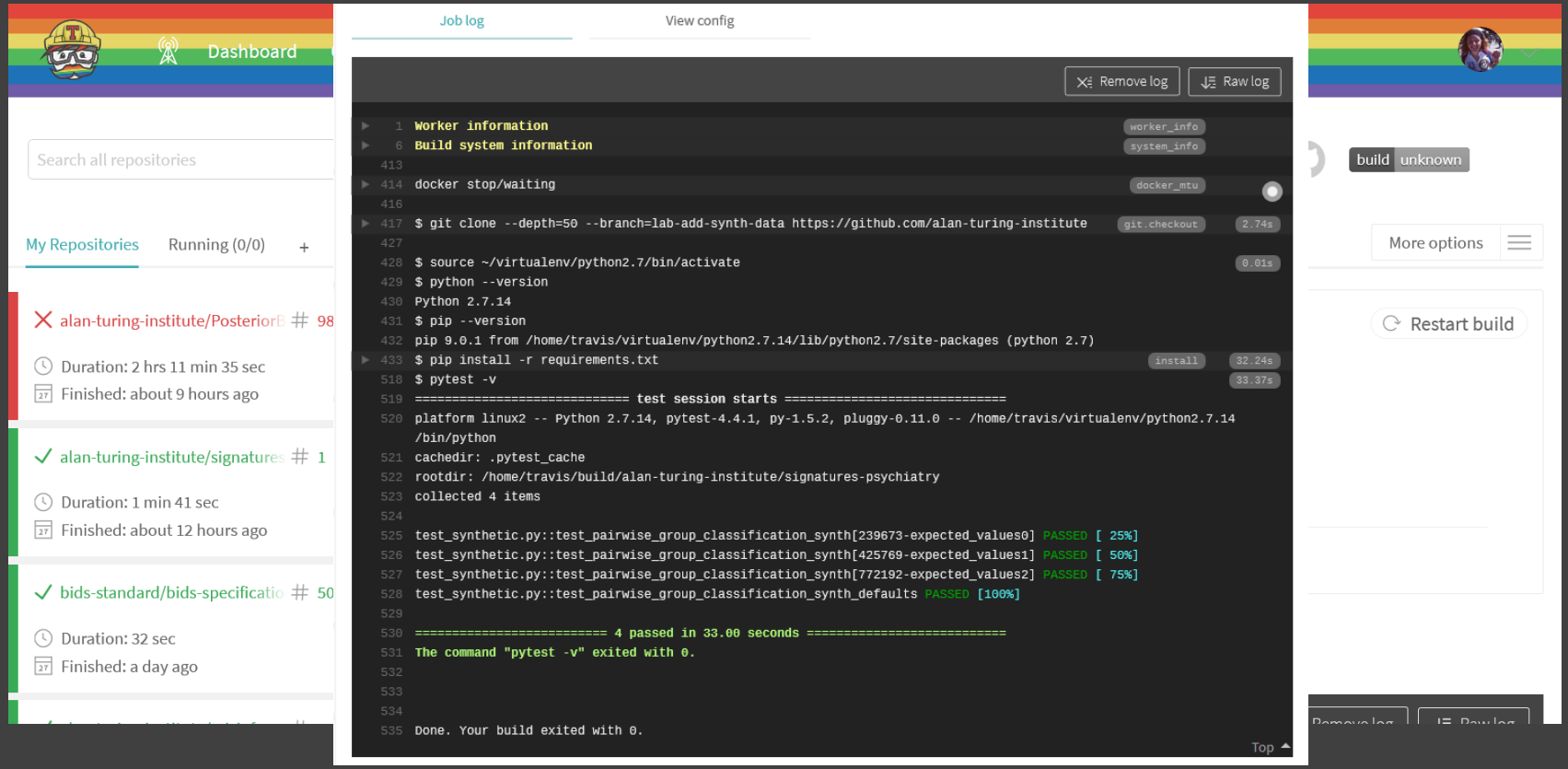
#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510



https://github.com/alan-turing-institute/signatures-psychiatry
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

https://the-turing-way.netlify.com/
continuous_integration/continuous_integration.html

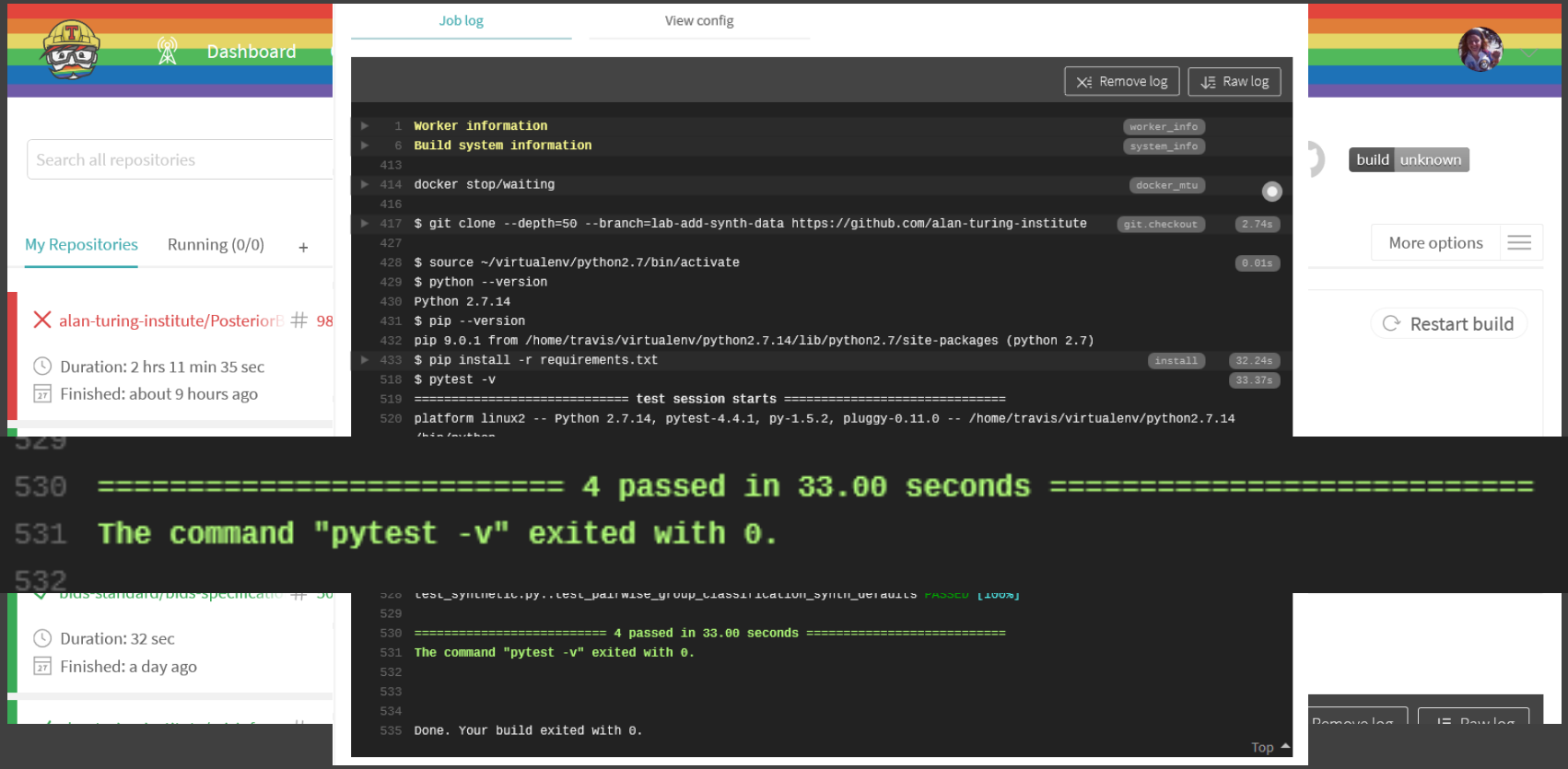


https://github.com/alan-turing-institute/signatures-psychiatry

#MQDataScience #TuringWay @kirstie_j

https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html

https://doi.org/10.5281/zenodo.3402510



https://github.com/alan-turing-institute/signatures-psychiatry

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html

- Run the analysis from start to finish as you work
- Many times tests will fail as expected: you're developing the analysis!
- Sometimes tests will fail unexpectedly
- CI makes you be explicit about what has changed

https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html

https://doi.org/10.5281/zenodo.3402510

Welcome to the Turing Way

The Turing Way is a lightly opinionated guide to reproducible data science.

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.

A bit more background

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

https://the-turing-way.netlify.com/introduction/introduction
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Becky Arnold

“There are a lot of things you need to know before you can jump into continuous integration.

Version control is a prerequisite for pretty much everything.”

https://software.ac.uk/about/fellows/becky-arnold
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

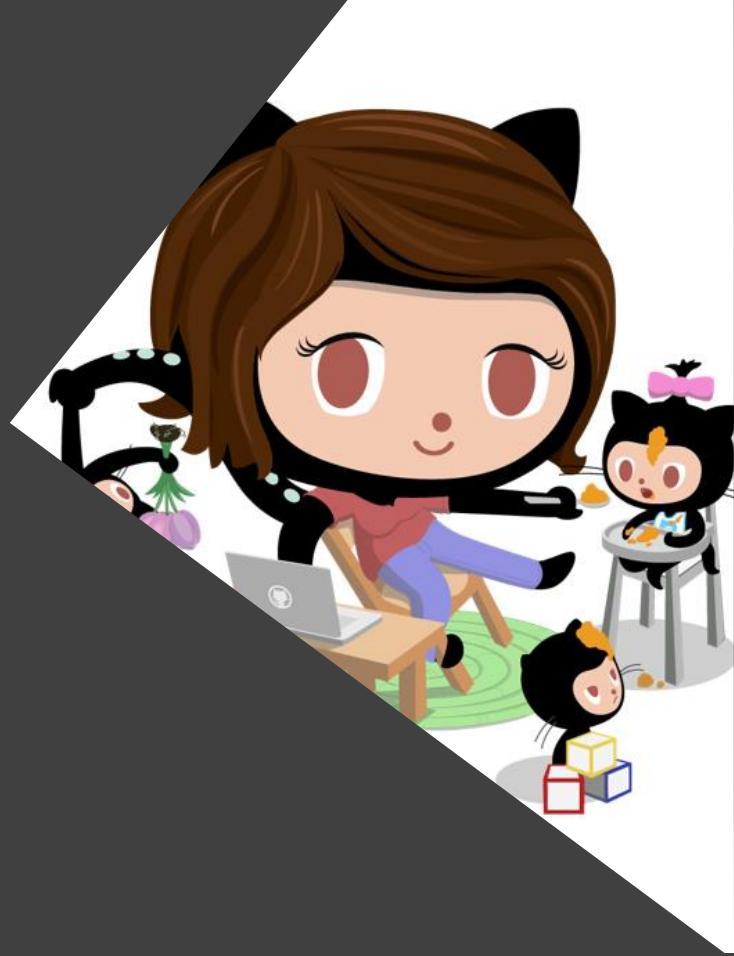
Continuous integration

Prerequisite	Importance	Notes
Experience with the command line	Necessary	A tutorial on working via the command line can be found here
Version control	Necessary	See the chapter on this for more information
Testing	Very helpful	See the chapter on this for more information
Reproducible computational environments	Necessary	See the chapter on this for more information, particularly the sections on YAML files and containers

Table of contents

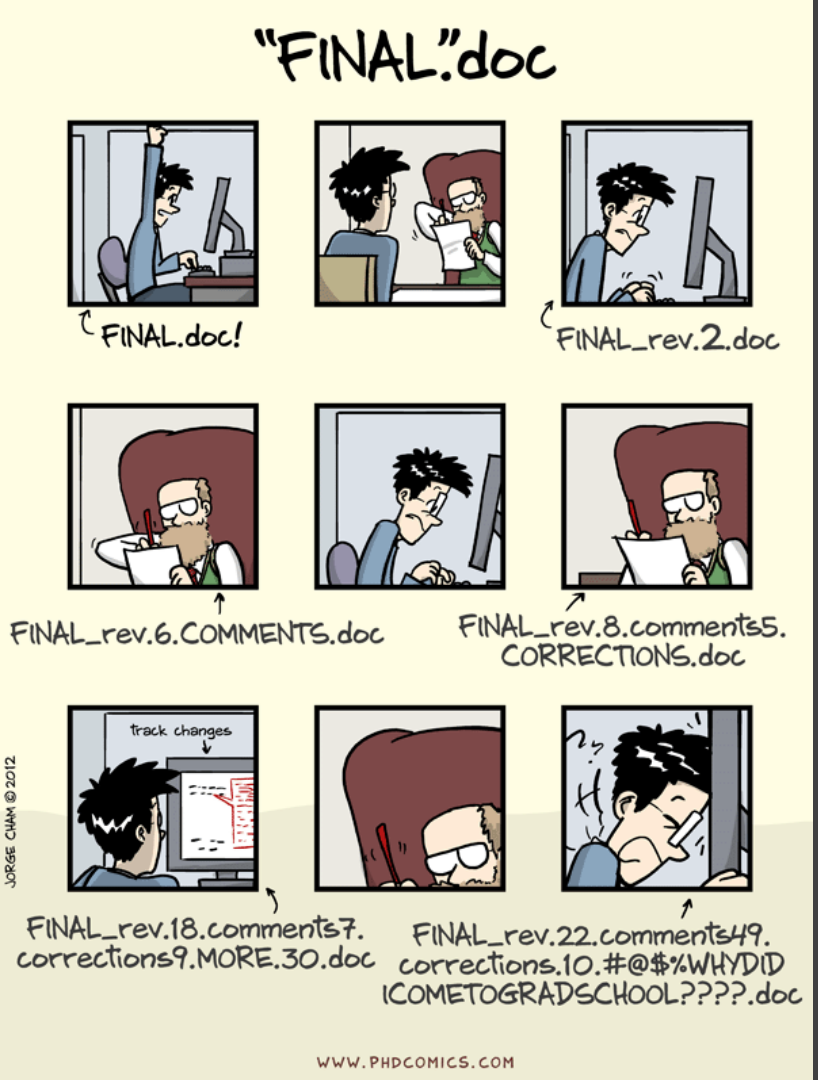
https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

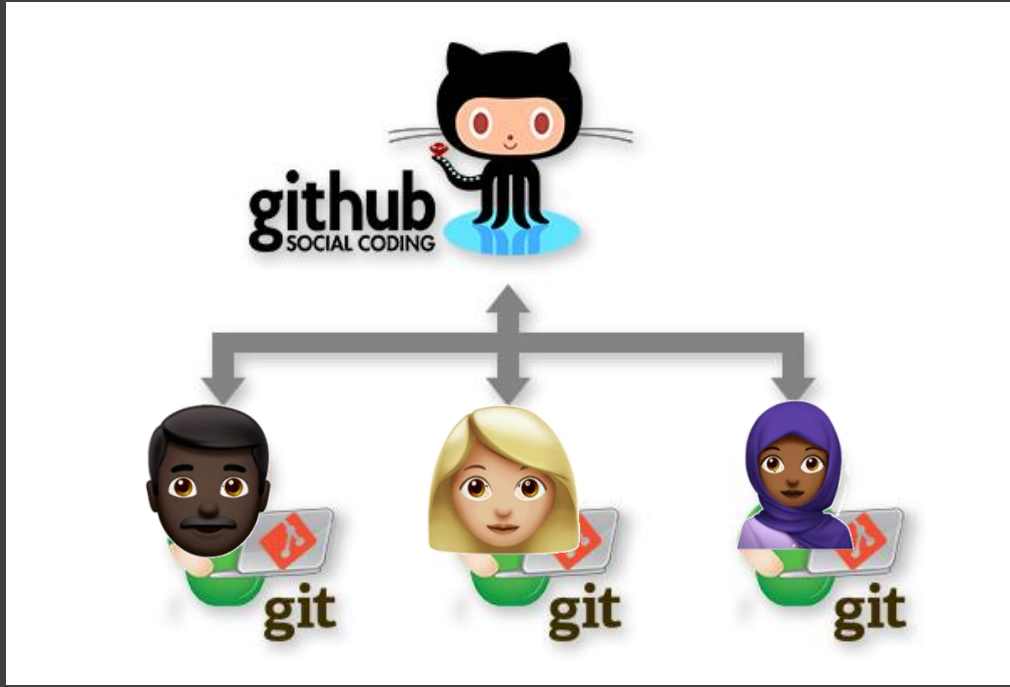
Version control



#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510





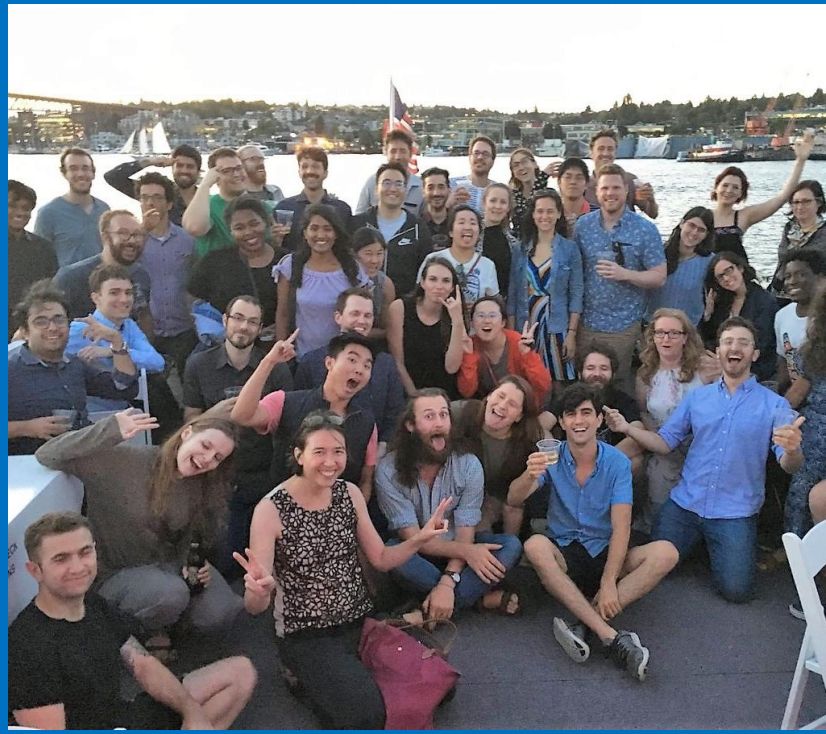
http://phdcomics.com/comics/archive.php?comicid=1531

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

Neurohackademy

“Every hackathon should have a gong that you can ring when you complete your first pull request.”

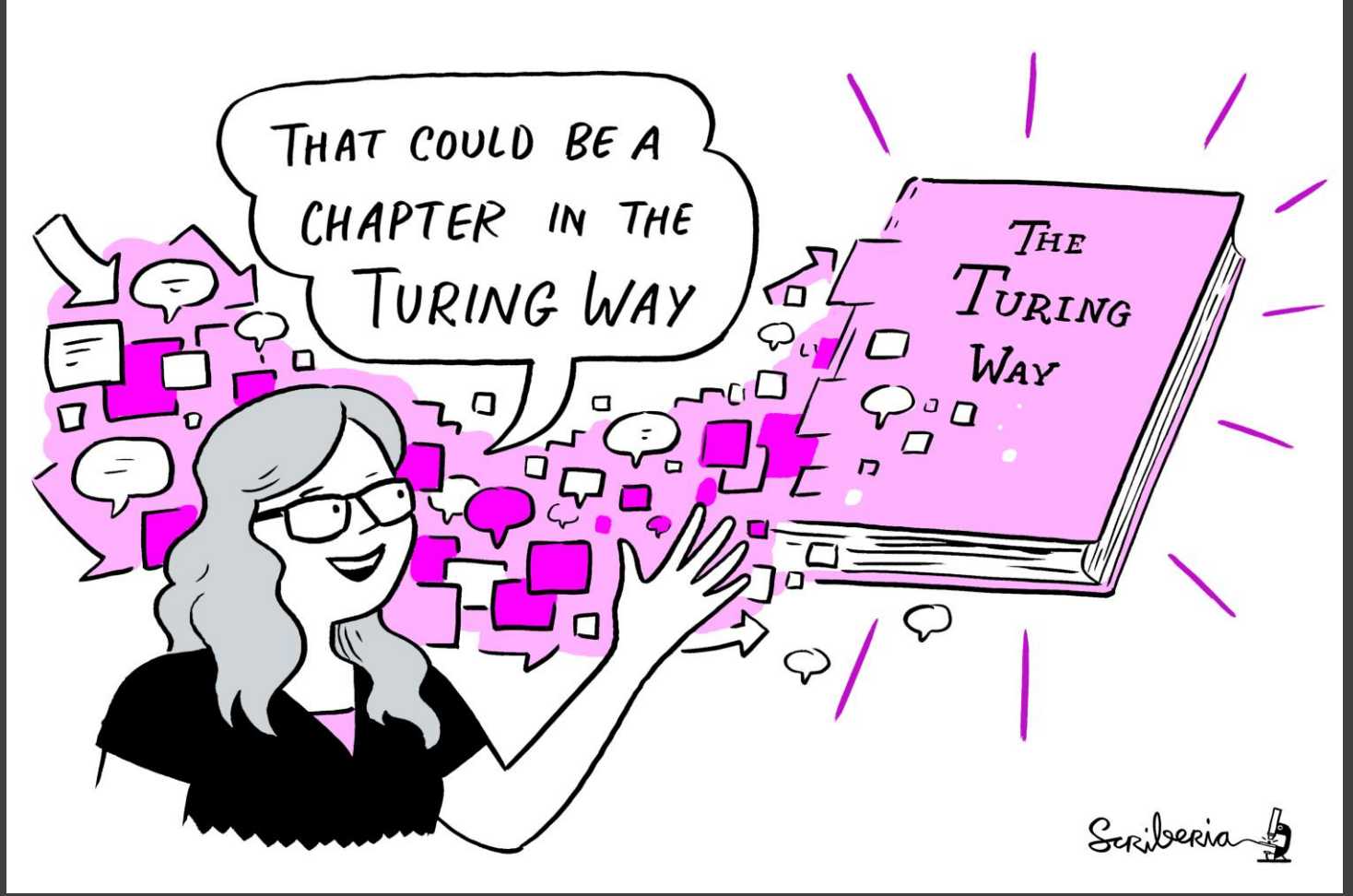


https://neurohackademy.org
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

A global collaboration

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510



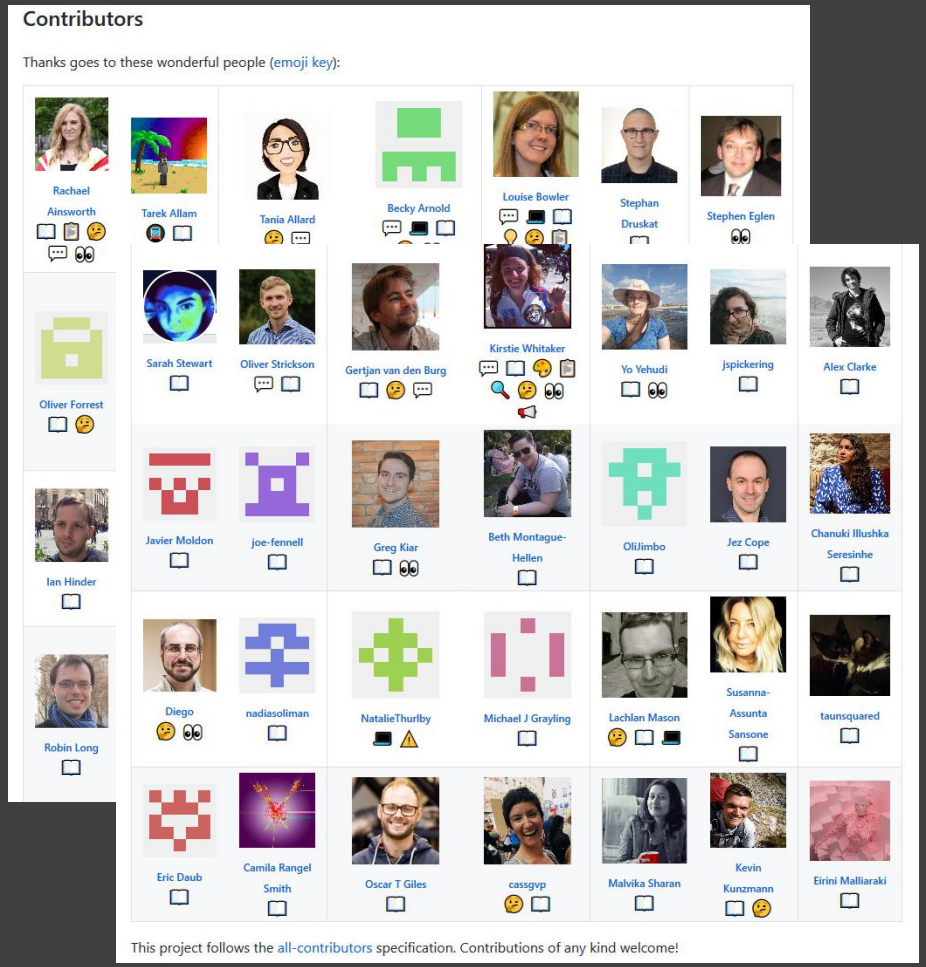
#MQDataScience #TuringWay @kirstie_j

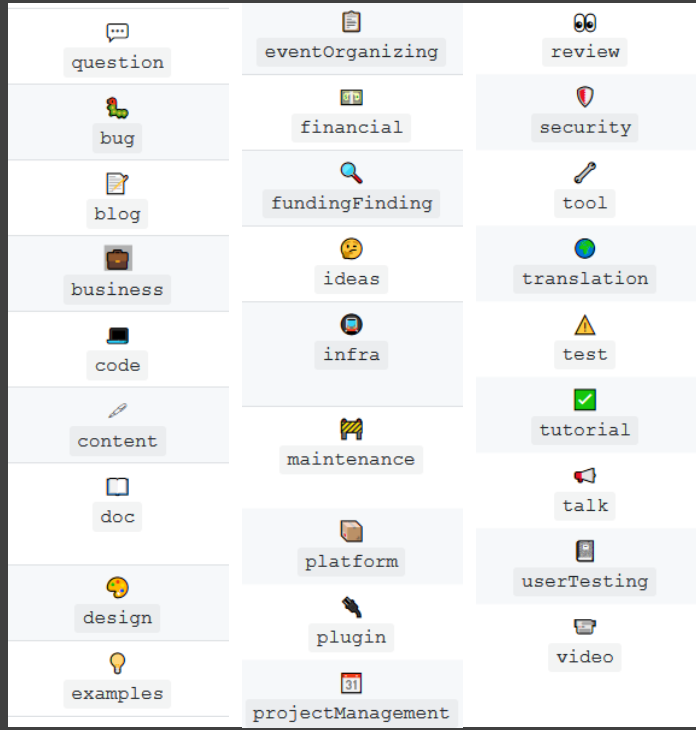
https://doi.org/10.5281/zenodo.3402510

Patricia Herterich

“What really sets The Turing Way apart is HOW we’re writing the book. The focus on community, the commitment to transparency and working open right from the beginning is an exciting (and terrifying) new way of working.”

https://rd-alliance.org/users/patricia-herterich
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510



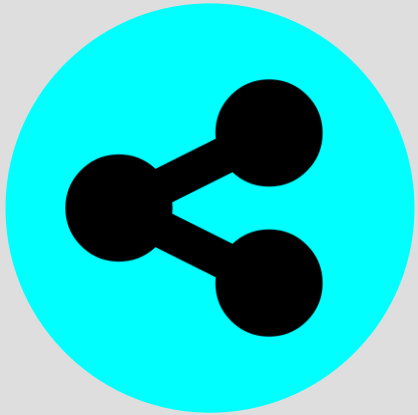


https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

https://github.com/alan-turing-institute/the-turing-way
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

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

moz://a

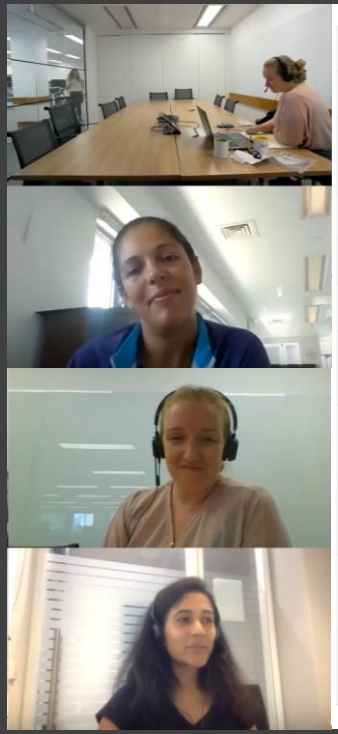
#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Join us at an online “Collaboration Café”

– 2 hours every week

– Zoom call, connect from anywhere

– “Shut up and write” or discuss and plan in breakout rooms



Time	Activity
Start	Welcome, code of conduct review
5 mins	Introductions and personal goal setting
15 mins	1st pomodoro session
35 mins	Break
40 mins	2nd pomodoro session
1hr 0 mins	Break
1hr 5 mins	3rd pomodoro session
1hr 25 mins	Break
1hr 30 mins	Open discussion: celebrations, reflections and future directions
2hr 0 mins	Close

https://www.youtube.com/watch?v=I0z7OEbBzes

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

https://github.com/alan-turing-institute/the-turing-way/blob/master/project_management/online-collaboration-cafe.md

Metrics for success

- 20 new chapters
- 100 authors
- 200 contributors
- 1000 mailing list subscribers
- 50 first pull requests
- 20 new contributors to other open source projects

https://github.com/alan-turing-institute/the-turing-way/blob/master/project_management/tps-funding-application-20190429.md

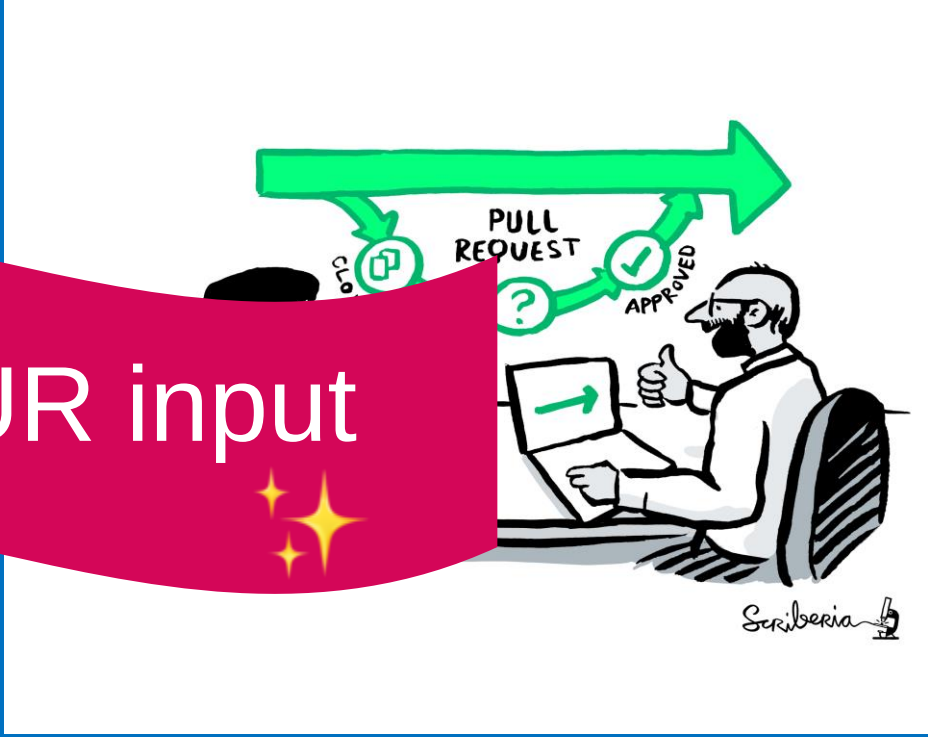
#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

Metrics for success

- 20 new chapters
- 100 autho
- 200 contri
- 1000 mail
- 50 first pull requests
- 20 new contributors to other open source projects

and YOUR input

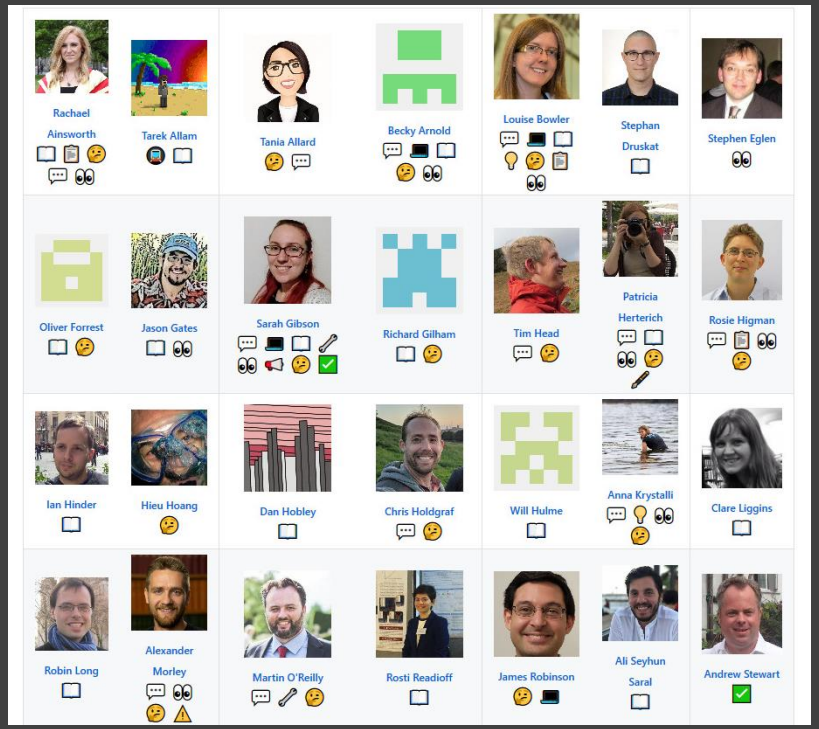


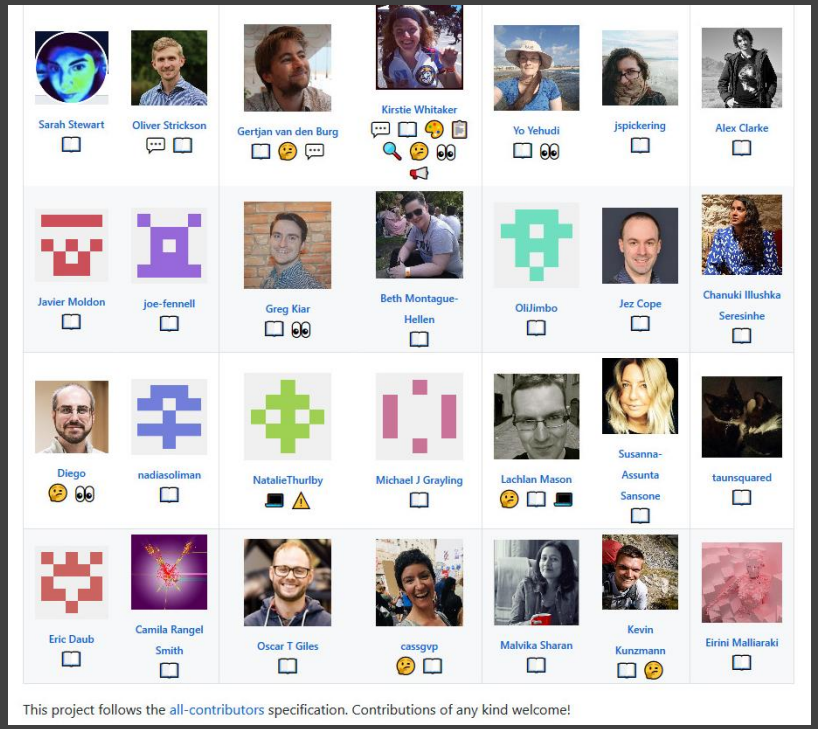
https://github.com/alan-turing-institute/the-turing-way/blob/master/project_management/tps-funding-application-20190429.md

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510

Thank you to current (& future) contributors





https://github.com/alan-turing-institute/the-turing-way#contributors
https://allcontributors.org/docs/en/emoji-key

#MQDataScience #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.3402510

Thank you

- Book: https://the-turing-way.netlify.com
- Newsletter: https://tinyletter.com/TuringWay
- GitHub: https://github.com/alan-turing-institute/the-turing-way
- Chat: https://gitter.im/alan-turing-institute/the-turing-way
- Next Collaboration Café: 18 September at 7pm UK time

- This work was supported by The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Tools, Practices and Systems" theme within that grant, and by The Alan Turing Institute under the EPSRC grant EP/N510129/1

- Unsplash photos by Adolfo Felix, James Pond, Jose Alejandro Cuffia, Kinson Leung, Mateo Vrbnjak, Mimi thian, Omar Albeik, Perry Grone, Toa Heftiba, Tomasz Frankows, Wilmer Martinez
- Noun Project icons by Aybige, Luis Prado, Edward Boatman, Becris, Rose Alice Design, Hyemm.work
- Original artwork by Scriberia: https://doi.org/10.5281/zenodo.3332807

#MQDataScience #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.3402510