The Alan Turing Institute

Reproducible research is impossible without software (so why don’t we reward it?)

Kirstie Whitaker

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

A beautiful example

NASA, https://flic.kr/p/tJbJf5

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Focus on the First Event Horizon Telescope Results

Shep Doeleman (EHT Director) on behalf of the EHT Collaboration

April 2019

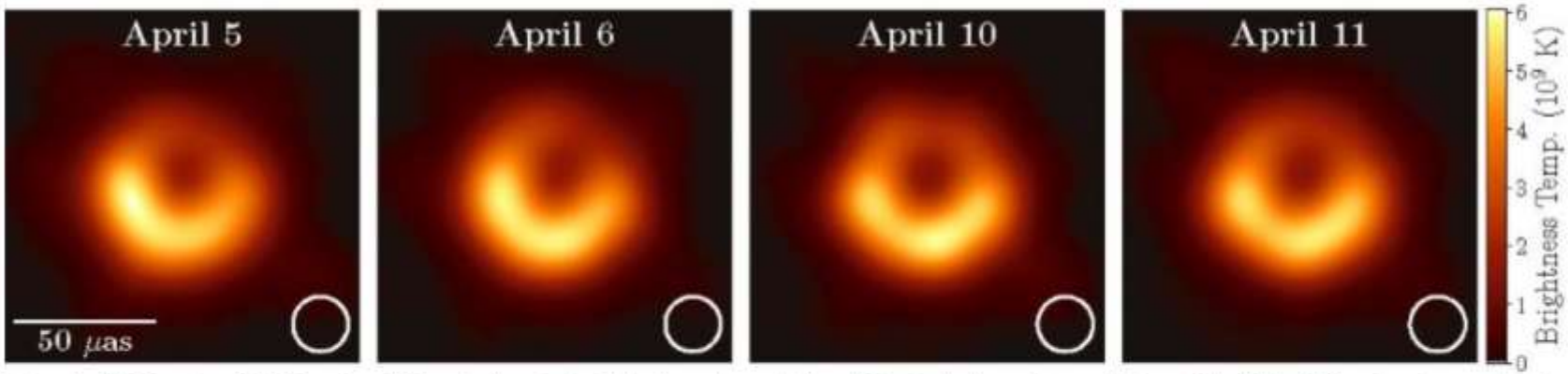
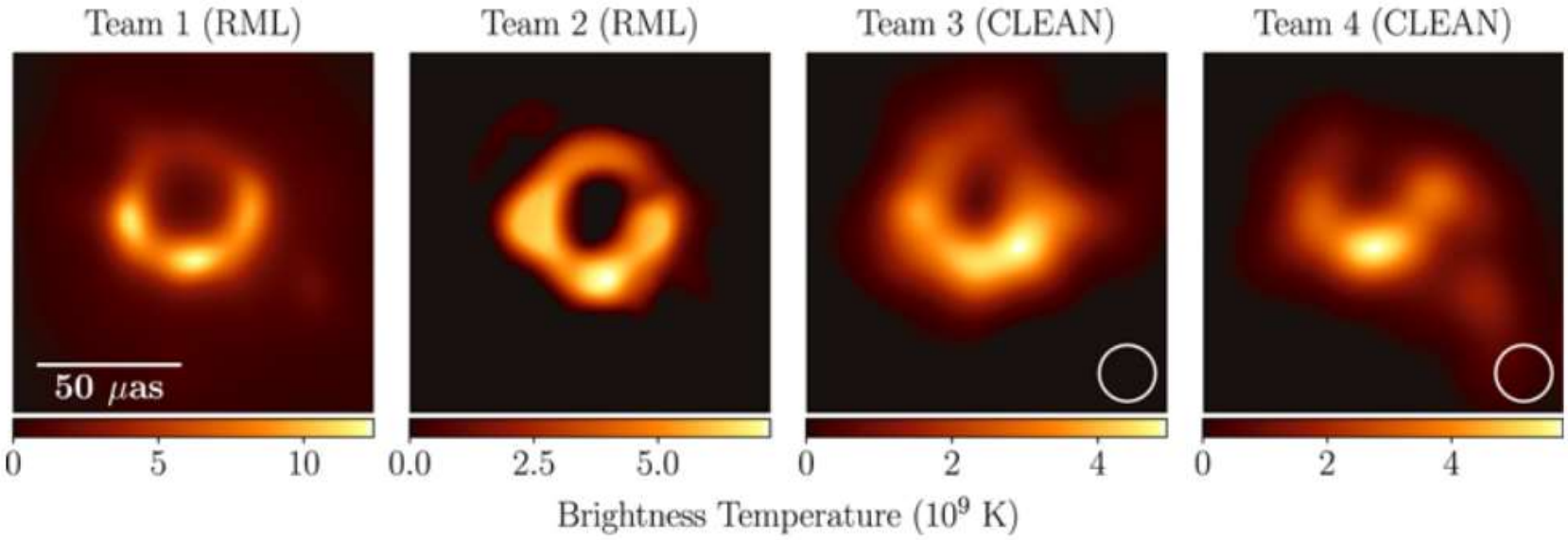


Figure 1. EHT images of M87 on four different observing nights. In each panel, the white circle shows the resolution of the EHT. All four images are dominated by a bright ring with enhanced emission in the south. From Paper IV (Figure 15).

We report the first image of a black hole.

https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



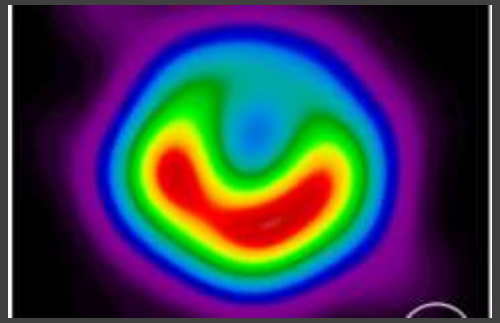
Zoom In Zoom Out Reset image size

Figure 4. The first EHT images of M87, blindly reconstructed by four independent imaging teams using an early, engineering release of data from the April 11 observations.

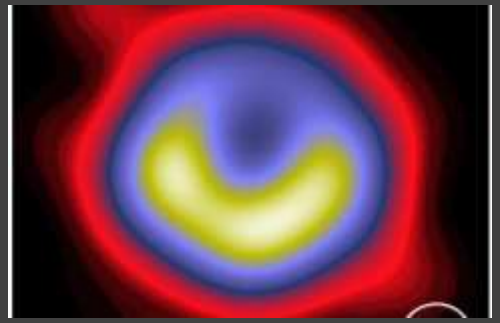
https://iopscience.iop.org/journal/2041-8205/page/Focus_on_EHT
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

https://twitter.com/sweichwald/status/1116430285342695424
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Paired_r

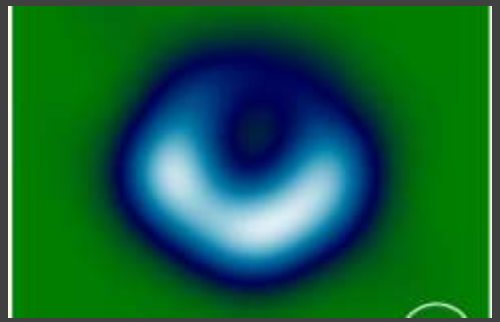


nipy_spectral



gist_stern

terrain



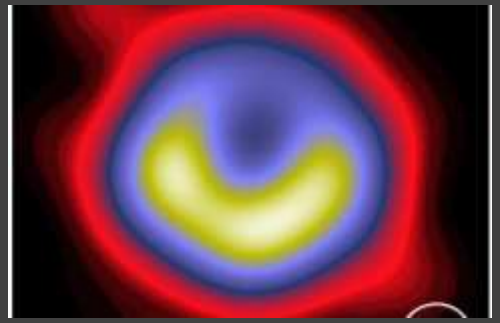
ocean

cividis

https://github.com/liamedeiros/ehtplot/blob/docs/docs/COLORMAPS.ipynb #CiteSoftware #TuringWay @kirstie_j
https://twitter.com/sweichwald/status/1116430285342695424 https://doi.org/10.5281/zenodo.2783998

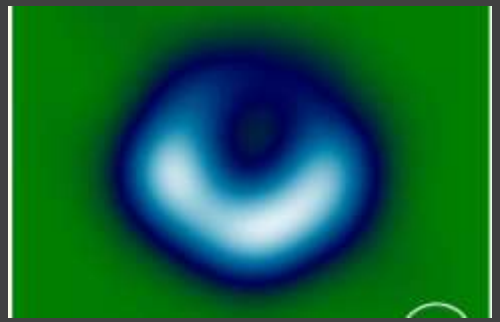
Paired_r

nipy_spectral



gist_stern

terrain



ocean

cividis

https://github.com/liamedeiros/ehtplot/blob/docs/docs/COLORMAPS.ipynb #CiteSoftware #TuringWay @kirstie_j
https://twitter.com/sweichwald/status/1116430285342695424 https://doi.org/10.5281/zenodo.2783998

- First M87 Event Horizon Telescope Results. III. Data Processing and Calibration
- A series of 6 papers published in April 2019
- Incredible long term international collaboration (200+ scientists, 60 institutes, 18 countries, 6 continents)

Software: DiFX (Deller et al. 2011), CALC, PolConvert (Martí-Vidal et al. 2016), HOPS (Whitney et al. 2004), CASA (McMullin et al. 2007), AIPS (Greisen 2003), ParselTongue (Kettenis et al. 2006), GNU Parallel (Tange 2011), GILDAS, eht-imaging (Chael et al. 2016, 2018), Numpy (van der Walt et al. 2011), Scipy (Jones et al. 2001), Pandas (McKinney 2010), Astropy (The Astropy Collaboration et al. 2013, 2018), Jupyter (Kluyver et al. 2016), Matplotlib (Hunter 2007).

doi: 10.3847/2041-8213/ab0c57
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

An introduction to me

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Picture credit: Chris Gorgolewski
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

	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

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

	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

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The humans are the hardest part of reproducibility

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

The humans are the hardest part of reproducibility and of software citation

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://the-turing-way.netlify.com/reproducibility/03/definitions.html
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The Turing Institute

https://www.turing.ac.uk/news/enigma-machine-goes-display-alan-turing-institute

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

https://www.bbc.co.uk/programmes/p0704h04
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

https://bletchleypark.org.uk
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

University network

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

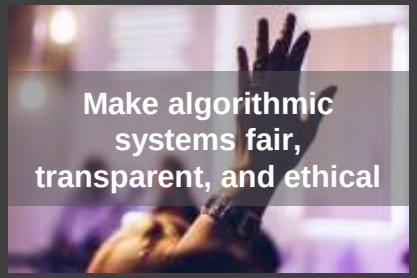
The Institute's partners and collaborators

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

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Challenges

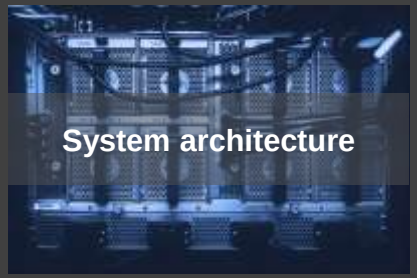
Advance data science and artificial intelligence to…

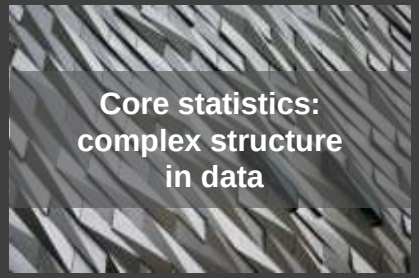


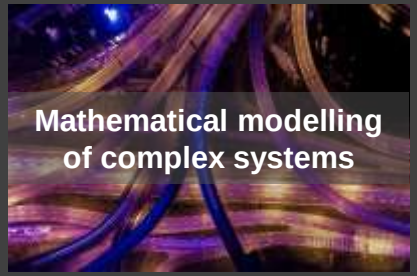
#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Core capabilities







#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Martin O’Reilly

“Make reproducible research too easy not to do.”

https://www.turing.ac.uk/people/researchers/martin-oreilly
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Martin O’Reilly

“Make reproducible research too easy not to do.

Do you need a biscuit?”

https://www.turing.ac.uk/people/researchers/martin-oreilly
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Martin O’Reilly

“Make reproducible research too easy not to do.

Do you need a biscuit?

If we can’t do it here, we can’t do it at all.”

https://www.turing.ac.uk/people/researchers/martin-oreilly
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The Turing Way

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

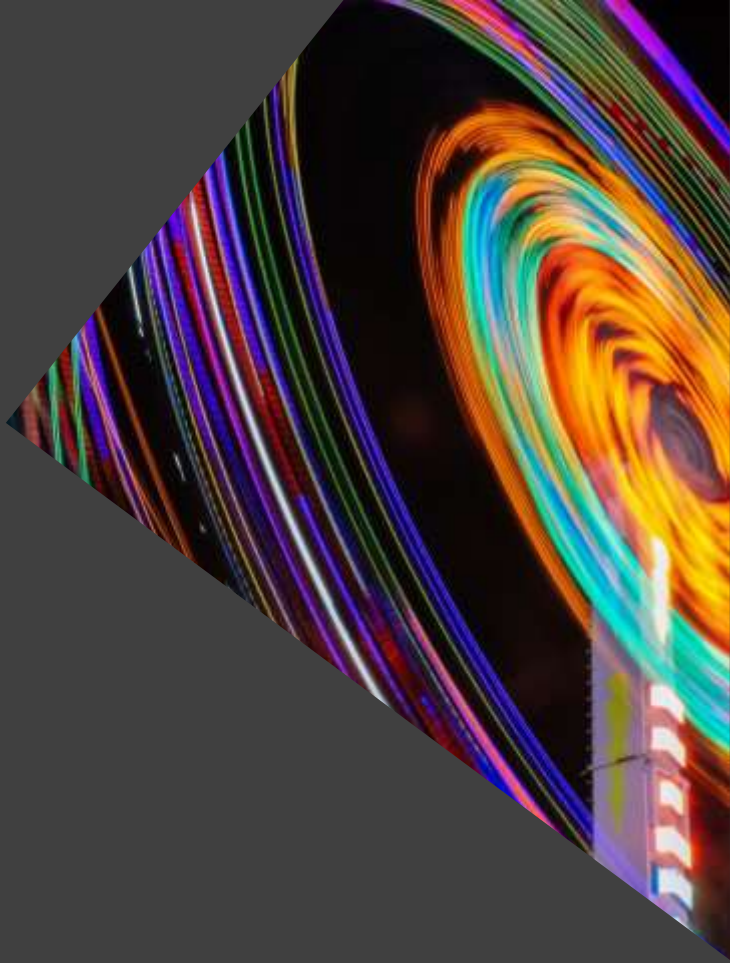
Catherine Lawrence

“We should ensure all our processes for running programmes are FAIR.

- Findable (intranet)
- Accessible (EDI)
- Interoperable across programmes and projects
- Reusable (bus factor)”

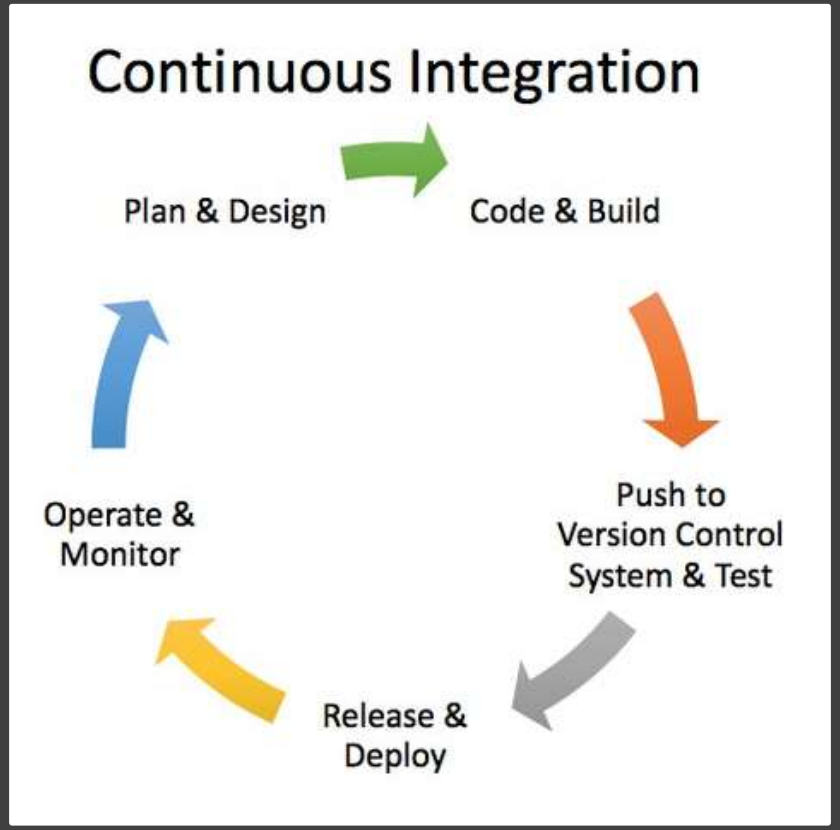
https://www.turing.ac.uk/people/business-team/catherine-lawrence
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Continuous Analysis

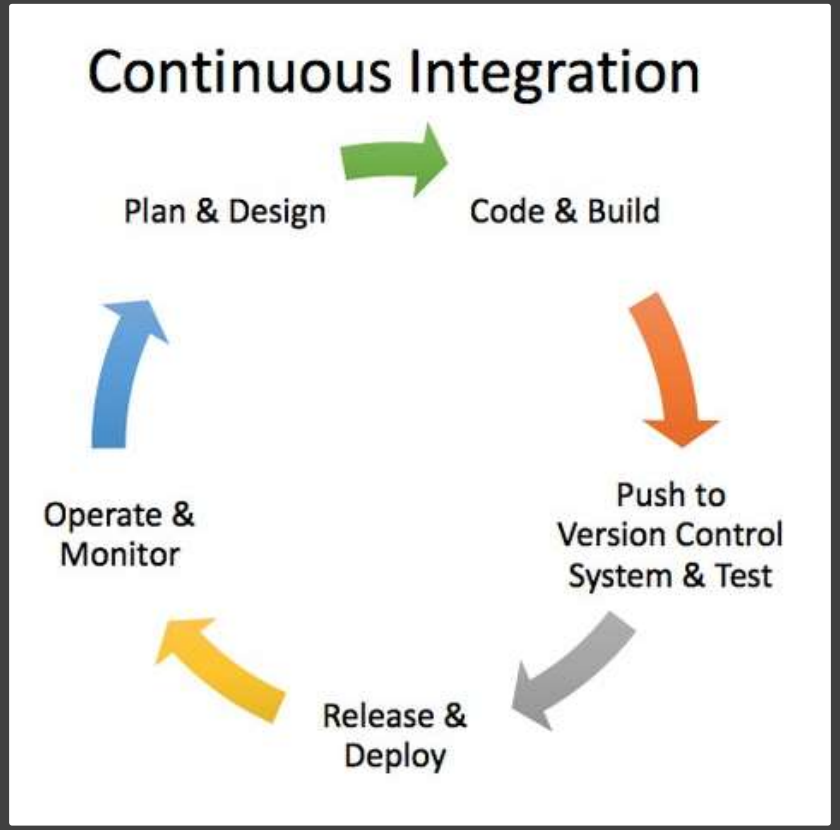


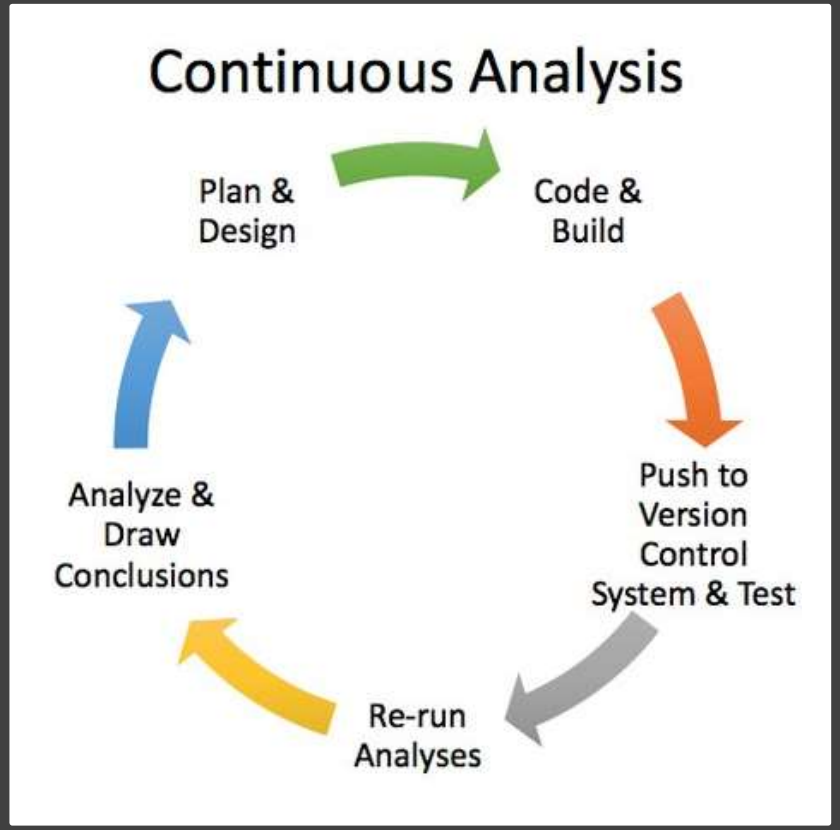
#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

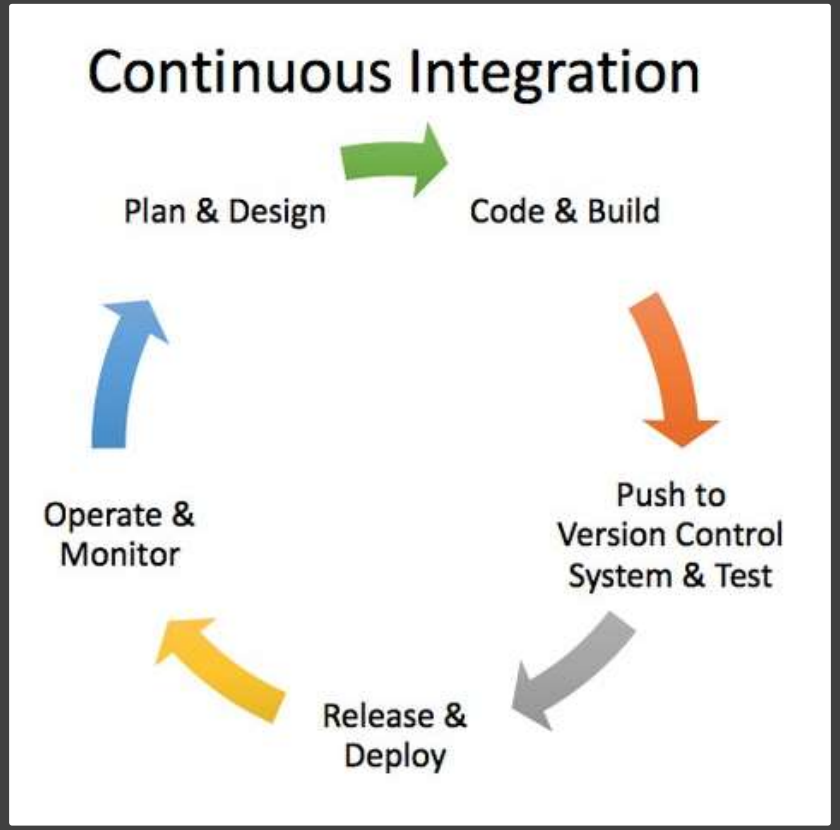


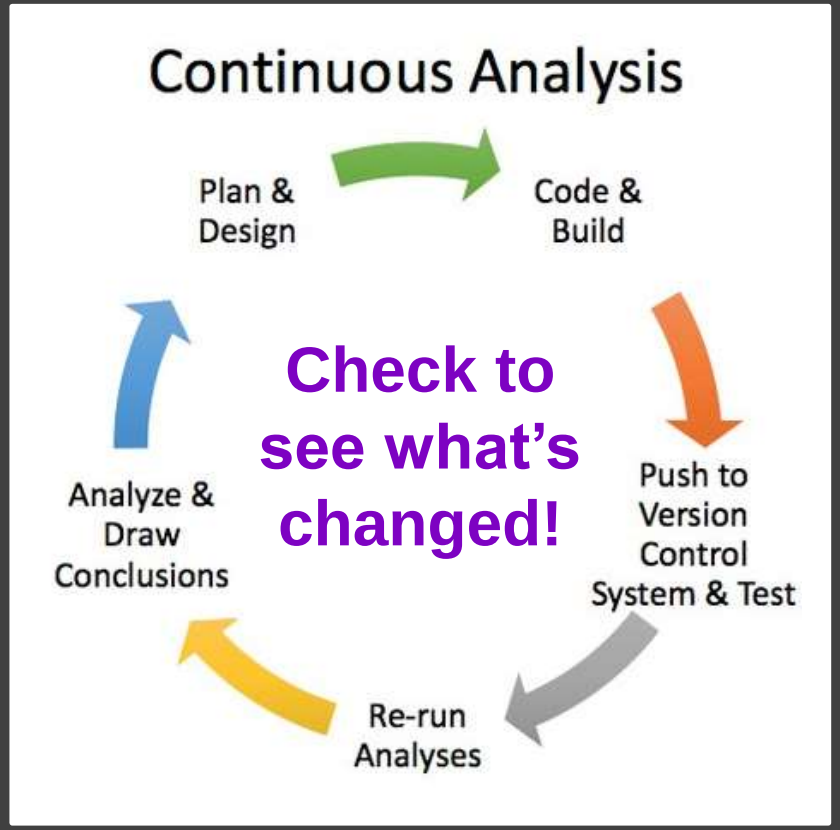
https://elifesciences.org/labs/e623676c/reproducibility-automated
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



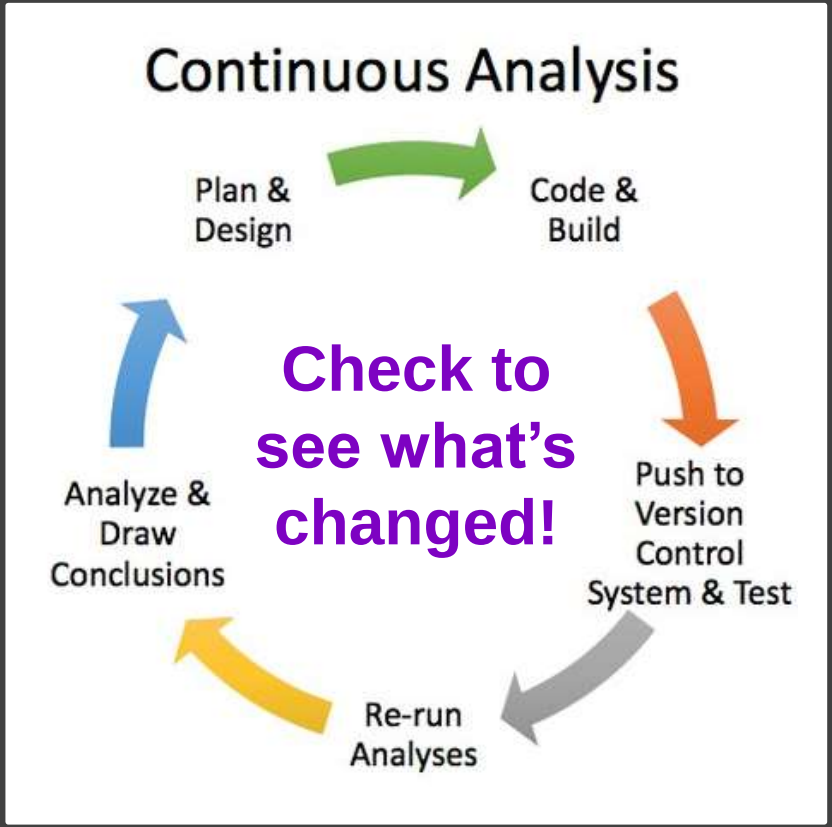


https://elifesciences.org/labs/e623676c/reproducibility-automated
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

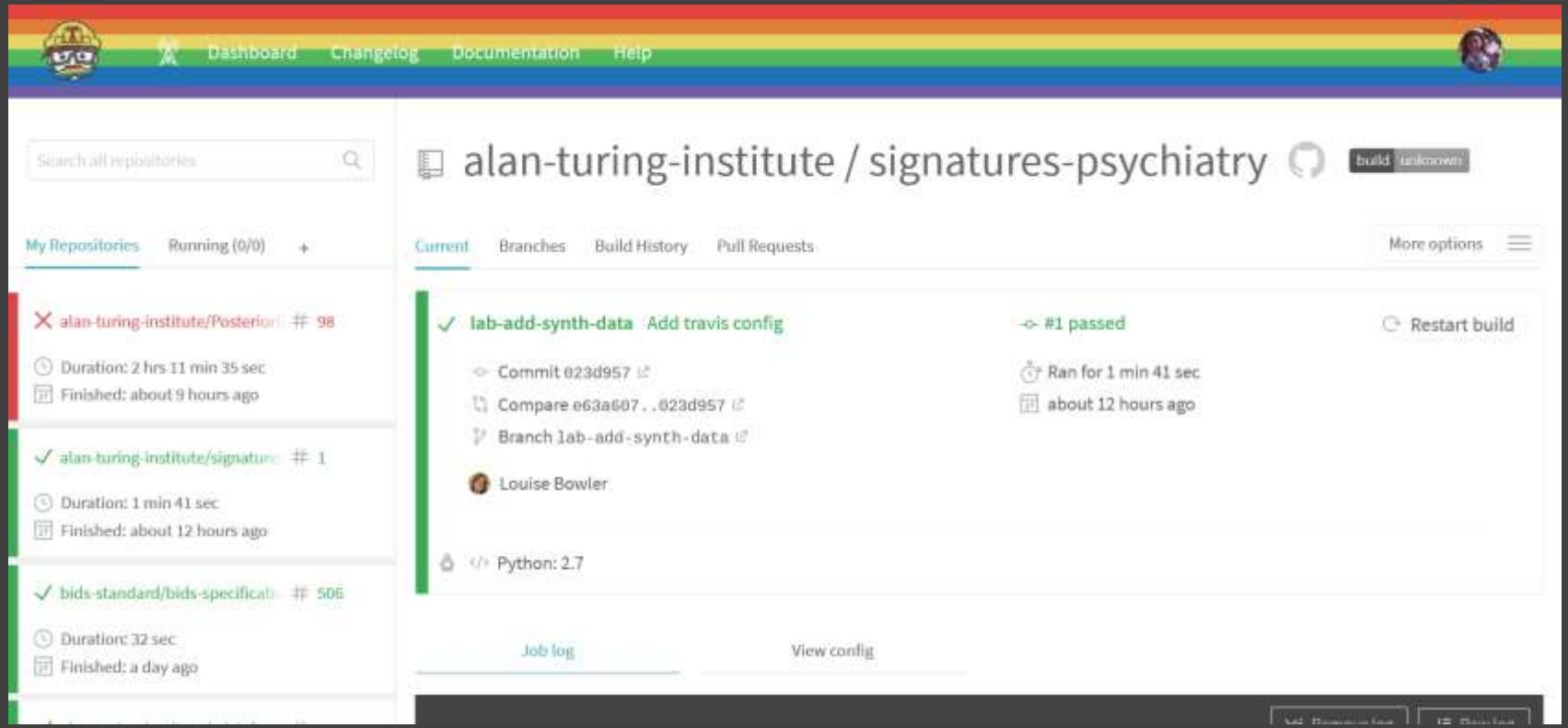




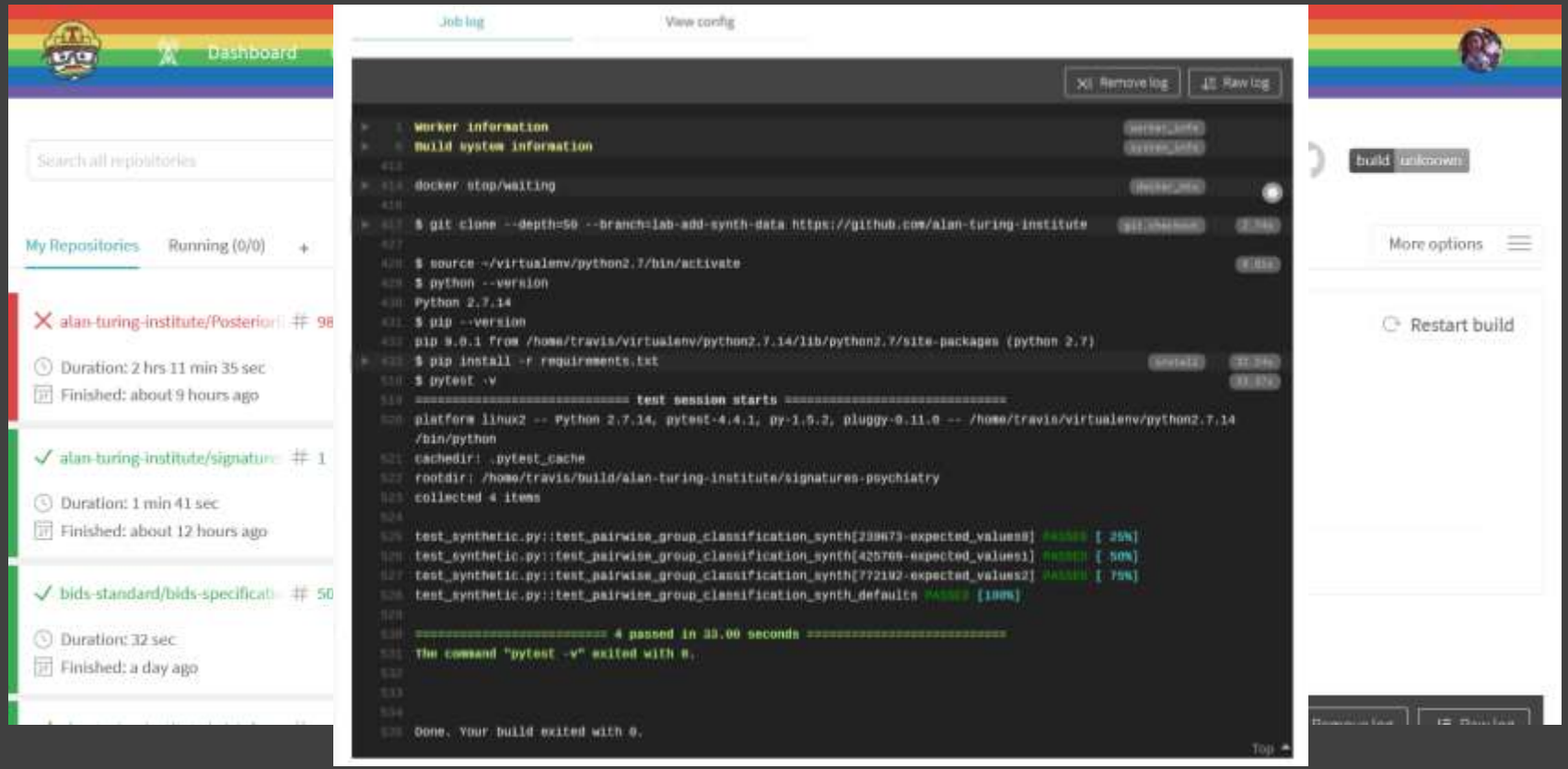
https://elifesciences.org/labs/e623676c/reproducibility-automated
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



https://elifesciences.org/labs/e623676c/reproducibility-automated
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



https://github.com/alan-turing-institute/signatures-psychiatry
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



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

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

- Run the analysis from start to finish as you're developing
- 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://www.youtube.com/watch?v=3GwjfUFyY6M
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Collaborating on GitHub/GitLab
6. Research Data Management
7. Reproducible Environments
8. Testing
9. Reviewing
10. Continous Integration
11. Reproducible Research with Make
12. Risk Assessment

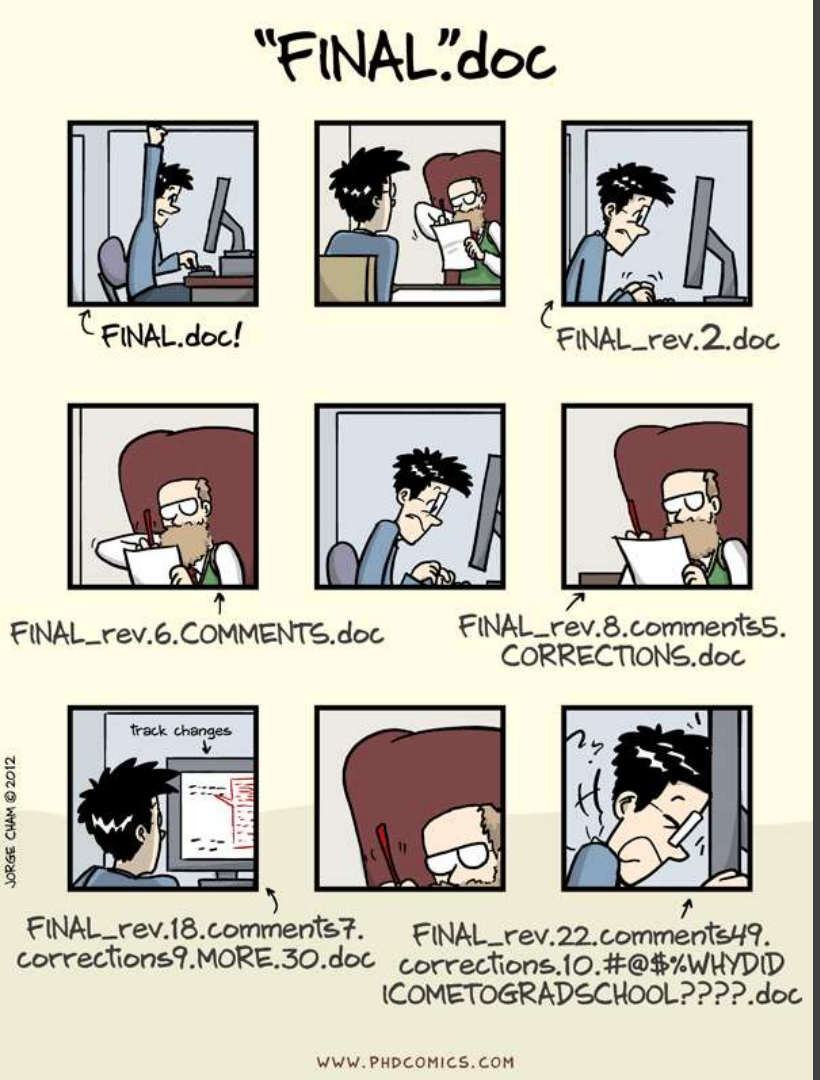
Continuous integration

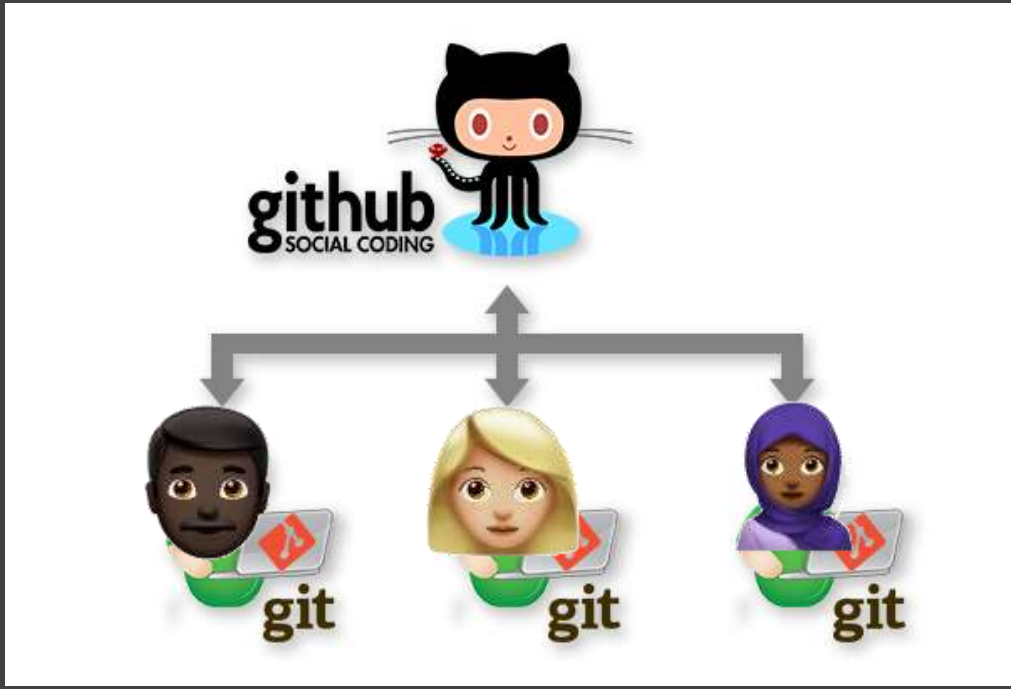
Prerequisite	Importance	Notes
Experience with the command line	Necessary	A tutorial on working via the command line can be found here
Version control	Necessary	See the chapter on this for more information
Testing	Very helpful	See the chapter on this for more information
Reproducible computational environments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

Table of contents

- Summary
- How this will help you/ why this is useful
  - What are continuous delivery and continuous deployment?
- What is Travis and how does it work?
- Setting up continuous integration with Travis
  - Basic steps

https://the-turing-way.netlify.com/continuous_integration/continuous_integration.html
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998





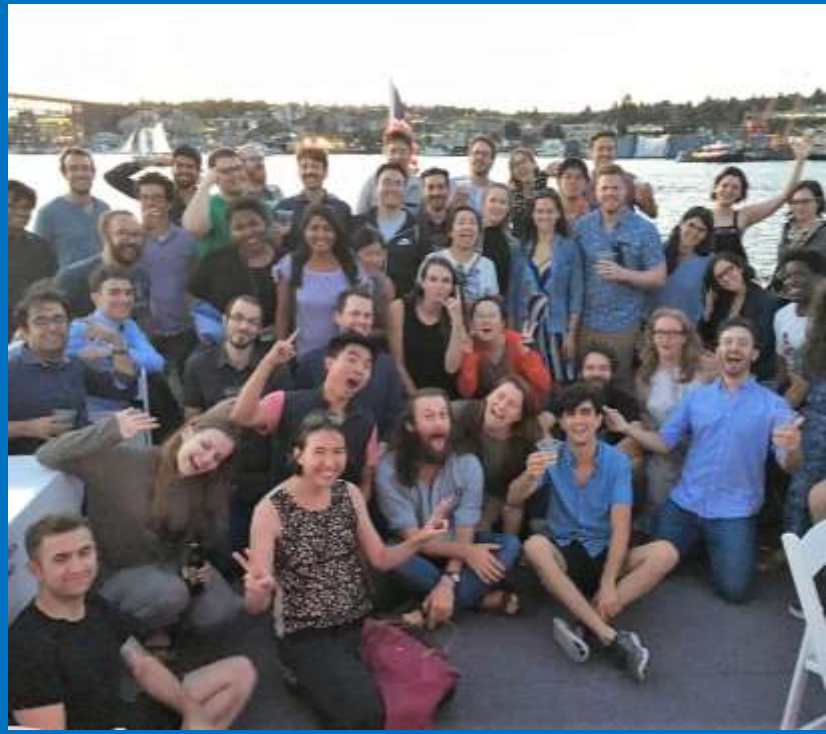
http://phdcomics.com/comics/archive.php?comicid=1531

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Neurohackademy

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



https://neurohackademy.org
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

https://www.youtube.com/watch?v=hSsjxbRxgqY
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Workshops & trainings

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

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

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Rosie Higman

“There’s no point in running events when you’re only preaching to the choir. We need to show researchers the selfish reasons to follow our recommendations.”

https://rosiehigman.wordpress.com
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

https://www.software.ac.uk/cw19
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

A Good Checklist

- ✓ Adds value
- ✓ Modular
- ✓ Customisable
- ✓ Guides & encourages communication

Checklist Manifesto

- **Codify best practice:** distil and collate community knowledge.
- **Level the team:** Spread responsibility and level authority.
- **Create awareness:** Bring focus to the routine, prepare for the unexpected.
- **Bring teams together:** Act of reviewing fosters feeling of teamwork and shared ownership.

GitHub issue templates as checklists for Open Reproducible Research

- **Library of customisable templates for common tasks** + infrastructure for domain specific variations
- **Ability to programmatically create domain/task specific issue sets**
- **Open for contribution** ***Community ownership and sense of value imperative!***

Part of the Turing Way project - https://github.com/alan-turing-institute/the-turing-way

@annakrystalli

https://checklib.github.io/checklib
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Anna Krystalli

“Checklists are a great way to make it really easy for busy people to do reproducible research. They can catch easily forgotten steps.”

https://alexmorley.me
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Anna Krystalli

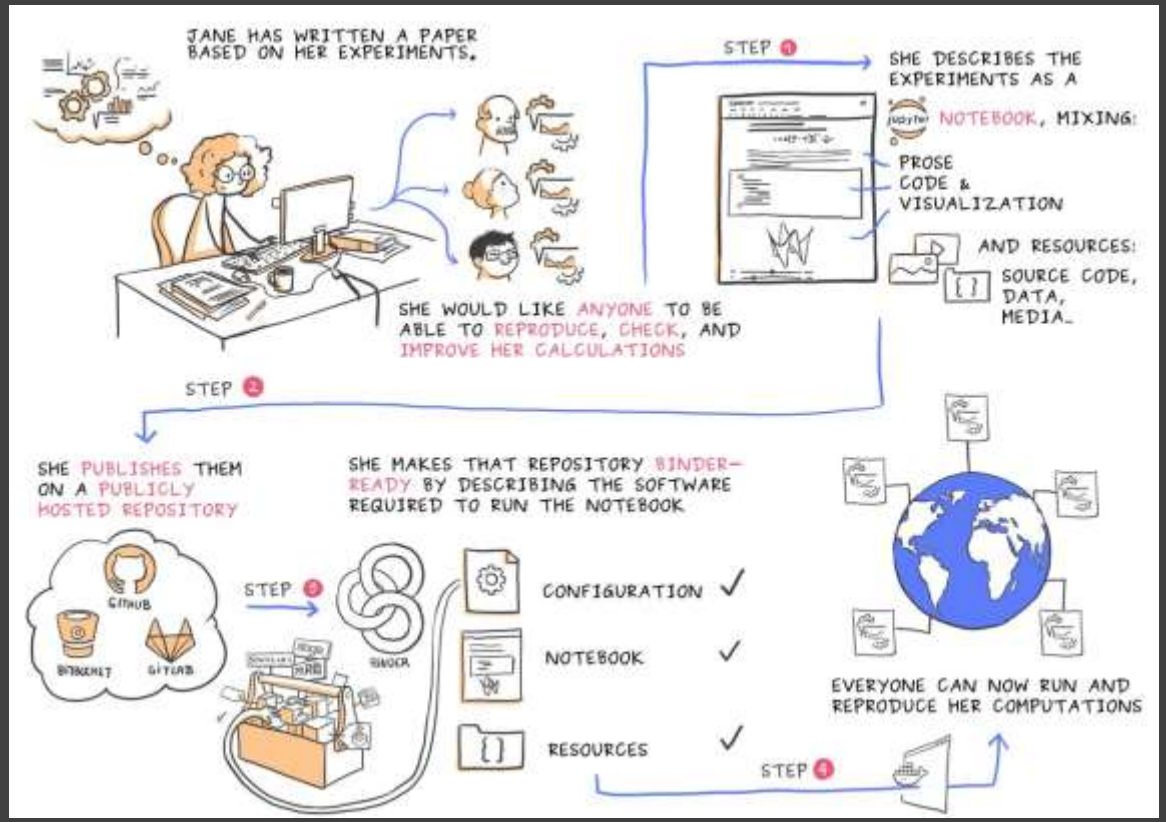
“Checklists are a great way to make it really easy for busy people to do reproducible research. They can catch easily forgotten steps…… like citing software!”

https://alexmorley.me
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

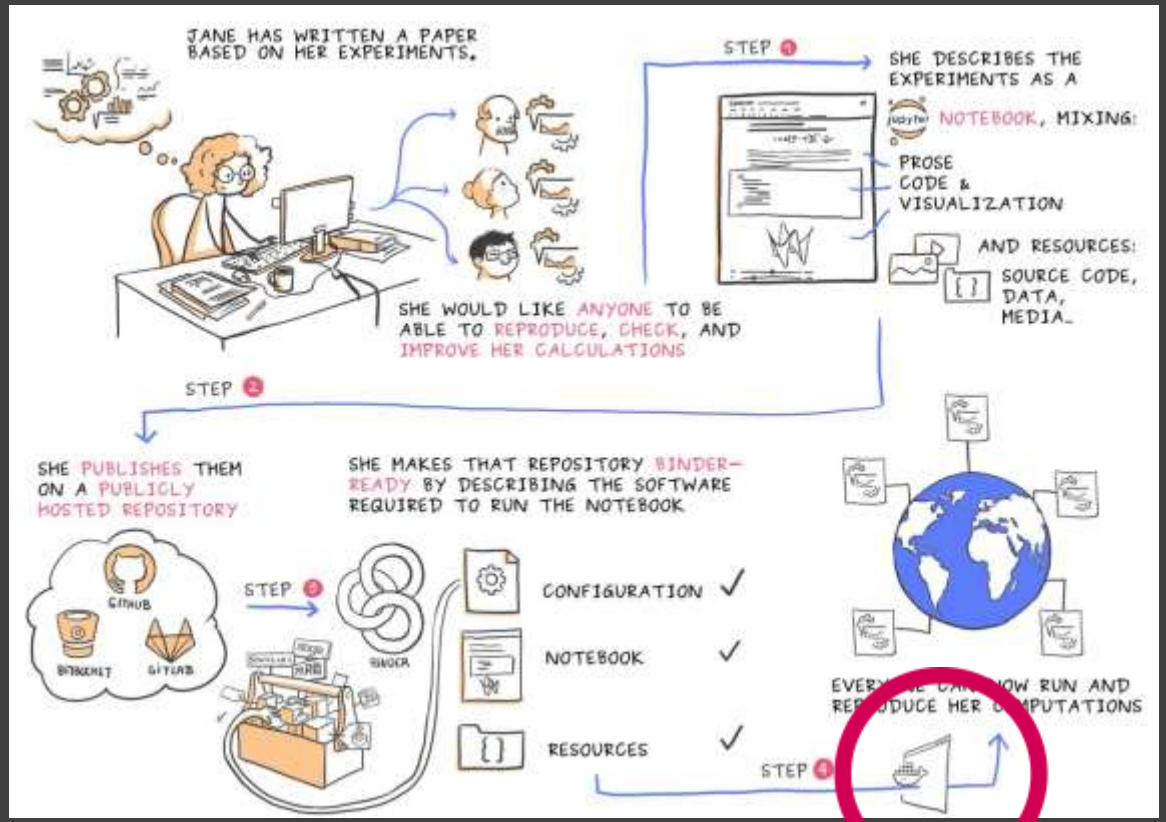
Turing Way & Binder

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

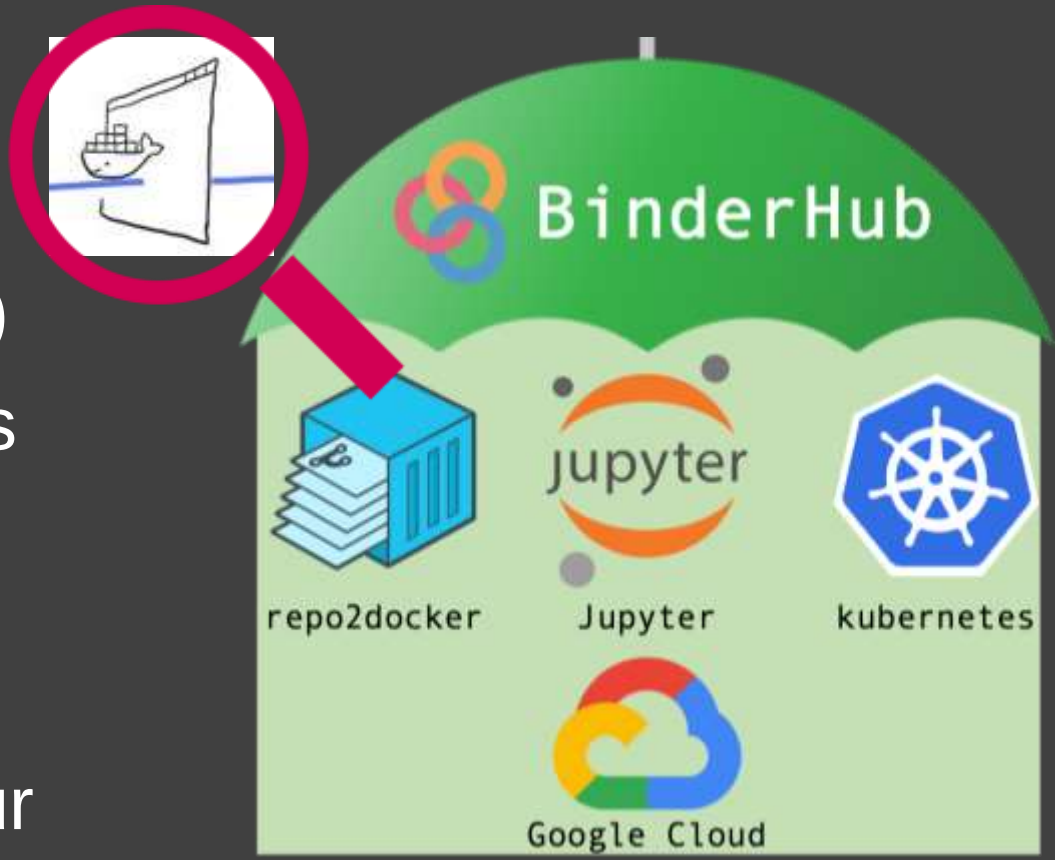


Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



Courtesy of Juliette Taka: https://twitter.com/mybinderteam/status/1082556317842264064
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

- Coordinate cloud computing resources with Kubernetes (k8s)
- Make it easy for users to access with a JupyterHub
- Set up the environment from your GitHub repository

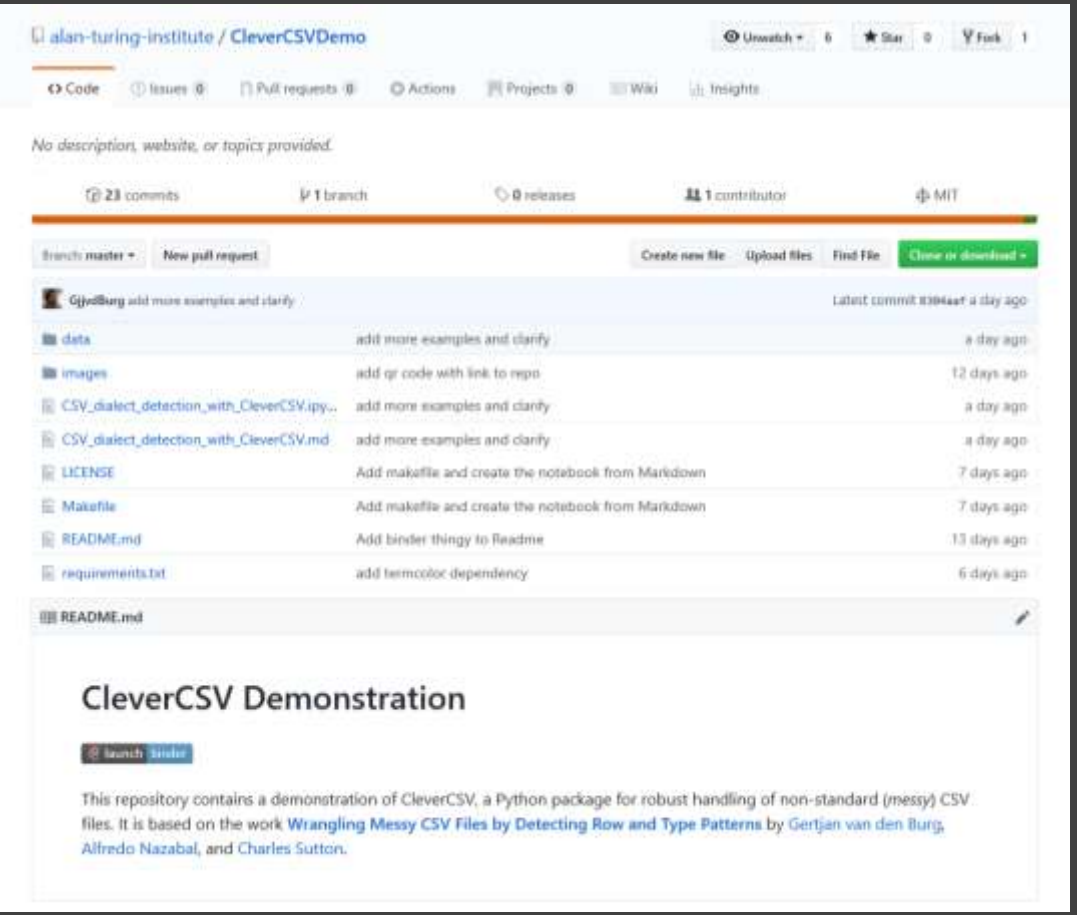


https://binderhub.readthedocs.io
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Gertjan van den Burg

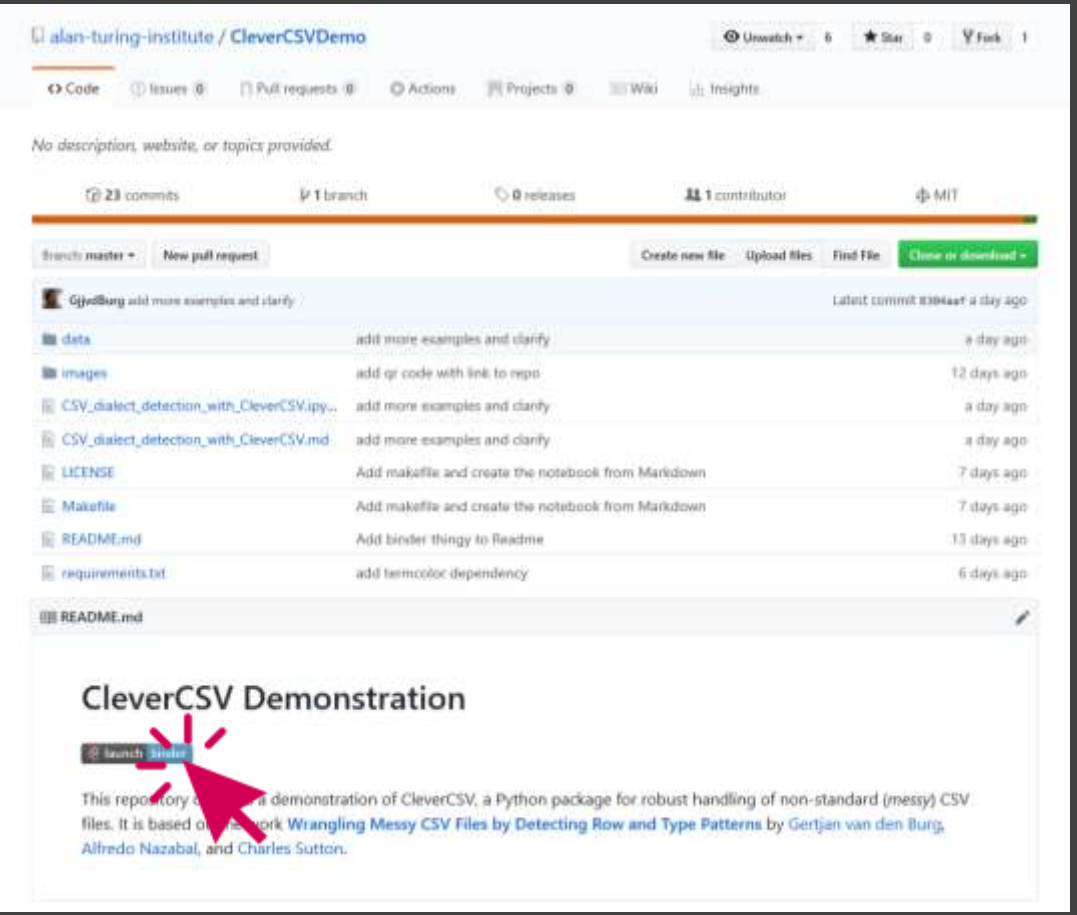
“The fun part of data science is the modelling. Being able to read in information from a csv file should not be the hardest part.”

https://gertjanvandenburg.com
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



– https://github.com/alan-turing-institute/CleverCSVDemo

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



– https://github.com/alan-turing-institute/CleverCSVDemo

– “Wrangling Messy CSV Files by Detecting Row and Type Patterns” arXiv:1811.11242

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

jupyter CSV_dialect_detection_with_CleverCSV Last Checkpoint: 16 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

CSV dialect detection with CleverCSV

Author: Gertjan van den Burg

In this note we'll show some examples of using CleverCSV, a package for handling messy CSV files. We'll start with a motivating example and then show some other files where CleverCSV shines. CleverCSV was developed as part of a research project on automating data wrangling. It achieves an accuracy of 97% on over 9300 real-world CSV files and improves the accuracy on messy files by 21% over standard tools.

Handy links:

- Paper on arXiv
- CleverCSV on GitHub
- CleverCSV on PyPI
- Reproducible Research Repo

IMDB Movie data

Alice is a data scientist who would like to analyse the movie ratings on IMDB for movies of different genres. She found a dataset shared by a user on Kaggle that contains information of over 14,000 movies. Great!

The data is stored in a CSV file, which is a very common data format for sharing tabular data. The first few lines of the file look like this:

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

IMDB Movie data

Alice is a data scientist who would like to analyse the movie ratings on IMDB for movies of different genres. She found a dataset shared by a user on Kaggle that contains information of over 14,000 movies. Great!

The data is stored in a CSV file, which is a very common data format for sharing tabular data. The first few lines of the file look like this:

```
fn,tid,title,wordsInTitle,url,imdbRating,ratingCount,duration,year,type,nrOfWins,nrOfNominations,nrOfPhotos,nrOf
NewsArticles,nrOfUserReviews,nrOfGenre,Action,Adult,Adventure,Animation,Biography,Comedy,Crime,Documentary,Drama
,Family,Fantasy,FilmNoir,GameShow,History,Horror,Music,Musical,Mystery,News,RealityTV,Romance,SciFi,Short,Sport,
TalkShow,Thriller,War,Western
titles01/tt0012349,tt0012349,Der Vagabund und das Kind (1921),der vagabund und das kind,http://www.imdb.com/titl
e/tt0012349/,8.4,40550,3240,1921,video.movie,1,0,19,96,85,3,0,0,0,0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0
titles01/tt0015864,tt0015864,Goldrausch (1925),goldrausch,http://www.imdb.com/title/tt0015864/,8.3,45319,5700,19
25,video.movie,2,1,35,110,122,3,0,0,1,0,0,1,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
titles01/tt0017136,tt0017136,Metropolis (1927),metropolis,http://www.imdb.com/title/tt0017136/,8.4,81007,9180,19
27,video.movie,3,4,67,428,376,2,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0
titles01/tt0017925,tt0017925,Der General (1926),der general,http://www.imdb.com/title/tt0017925/,8.3,37521,6420,
1926,video.movie,1,1,53,123,219,3,1,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
titles01/tt0021749,tt0021749,Lichter der Großstadt (1931),lichter der gro stadt,http://www.imdb.com/title/tt0021
749/,8.7,70057,5220,1931,video.movie,2,0,38,187,186,3,0,0,0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0
```

Seems pretty standard, let's load it with Pandas!

In [1]:
```
%xmode Minimal
```

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

In [1]:
```
%xmode Minimal
import pandas as pd
df = pd.read_csv('./data/imdb.csv')
```

```
Exception reporting mode: Minimal

ParserError: Error tokenizing data. C error: Expected 44 fields in line 66, saw 46
```

Oh, that doesn't work. Maybe there's something wrong with the file? Let's try opening it with the Python CSV reader:

In [2]:
```
import csv
with open('./data/imdb.csv', 'r', newline='') as fid:
    dialect = csv.Sniffer().sniff(fid.read())
    print("Detected delimiter = %r, quotechar = %r" % (dialect.delimiter, dialect.quotechar))
    fid.seek(0)
    reader = csv.reader(fid, dialect=dialect)
    rows = list(reader)

print("Loaded %i rows." % len(rows))
```

```
Detected delimiter = ' ', quotechar = "'"
Loaded 13928 rows.
```

Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

It turns out that on the 65th line of the file, there's a movie with the title `Dr. Seltsam\, oder wie ich lernte\, die Bombe zu lieben (1964)` (the German version of Dr. Strangelove). The title has commas in it, that are escaped using the `\` character! Why are CSV files so hard? 😀

CleverCSV to the rescue!

CleverCSV detects the dialect of CSV files much more accurately than existing approaches, and it is therefore robust against these kinds of format variations. It even has a wrapper that works with DataFrames!

In [3]:
```
from ccsv.wrappers import csv2df

df = csv2df('./data/imdb.csv')
df
```

Out[3]:

	fn	tid	title	wordsInTitle	url	imdbRating	ratingCount	duration	year	type	...	News
0	titles01/tt0012349	tt0012349	Der Vagabund und das Kind (1921)	der vagabund und das kind	http://www.imdb.com/title/tt0012349/	8.4	40550.0	3240.0	1921.0	video.movie	...	0
1	titles01/tt0015864	tt0015864	Goldrausch (1925)	goldrausch	http://www.imdb.com/title/tt0015864/	8.3	45319.0	5700.0	1925.0	video.movie	...	0
2	titles01/tt0017136	tt0017136	Metropolis (1927)	metropolis	http://www.imdb.com/title/tt0017136/	8.4	81007.0	9180.0	1927.0	video.movie	...	0
3	titles01/tt0017925	tt0017925	Der General (1926)	der general	http://www.imdb.com/title/tt0017925/	8.3	37521.0	6420.0	1926.0	video.movie	...	0
			Lichter der	lichter der gro	http://www.imdb.com							

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

Huh, that's strange, Python thinks the *space* is the delimiter and loads 13928 rows, but the file should contain 14,762 rows according to the documentation. What's going on here?

It turns out that on the 65th line of the file, there's a movie with the title `Dr. Seltsam\, oder wie ich lernte\, die Bombe zu lieben` (German version of Dr. Strangelove). The title has commas in it, that are escaped using the `\` character! Why are CSV files so hard? 😐

CleverCSV to the rescue!

CleverCSV detects the dialect of CSV files much more accurately than existing approaches, and it is therefore robust against these kinds of fo
even has a wrapper that works with DataFrames!

In [3]:
```
from ccsv.wrappers import csv2df

df = csv2df('./data/imdb.csv')
df
```

Out[3]:

	fn	tid	title	wordsInTitle	url	imdbRating	ratingCount	duration	year	type	...	News
0	titles01/tt0012349	tt0012349	Der Vagabund und das Kind (1921)	der vagabund und das kind	http://www.imdb.com/title/tt0012349/	8.4	40550.0	3240.0	1921.0	video.movie	...	0
1	titles01/tt0015864	tt0015864	Goldrausch (1925)	goldrausch	http://www.imdb.com/title/tt0015864/	8.3	45319.0	5700.0	1925.0	video.movie	...	0
2	titles01/tt0017136	tt0017136	Metropolis (1927)	metropolis	http://www.imdb.com/title/tt0017136/	8.4	81007.0	9180.0	1927.0	video.movie	...	0
3	titles01/tt0017925	tt0017925	Der General (1926)	der general	http://www.imdb.com/title/tt0017925/	8.3	37521.0	6420.0	1926.0	video.movie	...	0
			Lichter der	lichter der gro	http://www.imdb.com							

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

```
df
```

			Episode 2005)	episode	/title/tt0672466/							
14757	titles04/index.html.9992	tt0675644	"Playhouse 90" The Miracle Worker (TV Episode ...	playhouse the miracle worker tv episode	http://www.imdb.com/title/tt0675644/	7.3	8.0	5400.0	1957.0	video.episode	...	0
14758	titles04/index.html.9994	tt0679222	"Private Screenings" Robert Mitchum and Jane R...	private screenings robert mitchum and jane rus...	http://www.imdb.com/title/tt0679222/	7.0	20.0	3600.0	1996.0	video.episode	...	0
14759	titles04/index.html.9995	tt0680064	"Providence" All the King's Men (TV Episode 2002)	providence all the king s men tv episode	http://www.imdb.com/title/tt0680064/	NaN	NaN	3600.0	2002.0	video.episode	...	0
14760	titles04/index.html.9997	tt0681024	"QI" Adam (TV Episode 2003)	qi adam tv episode	http://www.imdb.com/title/tt0681024/	7.6	89.0	1800.0	2003.0	video.episode	...	0

14761 rows × 44 columns

Hooray! 🎉

How does it work? CleverCSV searches the space of all possible dialects of a file, and computes a *data consistency measure* that quantifies how much the resulting table "looks like real data". The consistency measure combines patterns of row lengths in the parsing result and the data type of the resulting cells. This mimicks how a human would identify the dialect. If you're wondering why this problem is hard, it's because every dialect will give you *some* table, but not necessarily the correct one. More details can be found in the paper.

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

jupyter CSV_dialect_detection_with_CleverCSV Last Checkpoint: 16 minutes ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

CSV dialect detection with CleverCSV

Author: Gertjan van den Burg

In this note we'll show some examples of using CleverCSV, a package for handling messy CSV files. We'll start with a motivating example and then show some other files where CleverCSV shines. CleverCSV was developed as part of a research project on automating data wrangling. It achieves an accuracy of 97% on over 9300 real-world CSV files and improves the accuracy on messy files by 21% over standard tools.

Handy links:

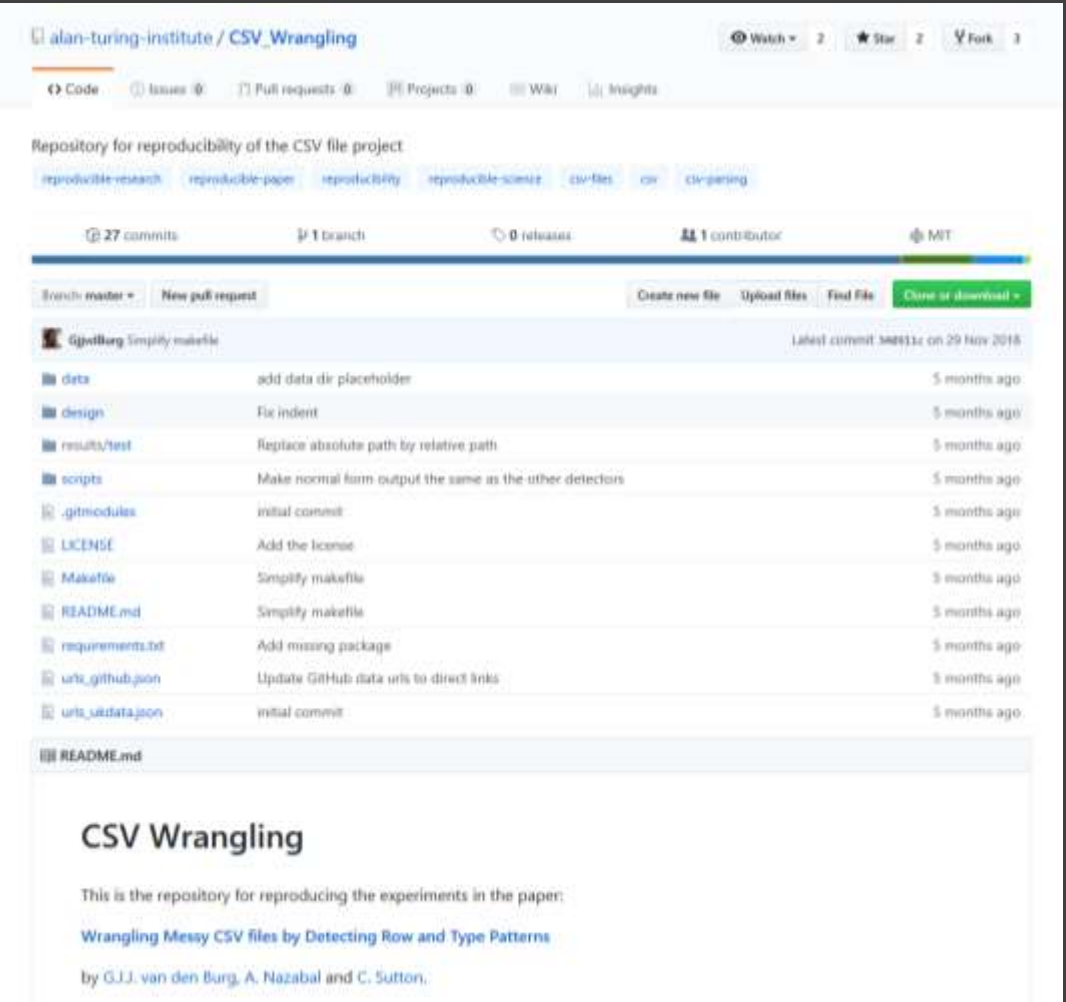
- Paper on arXiv
- CleverCSV on GitHub
- CleverCSV on PyPI
- Reproducible Research Repo

IMDB Movie data

Alice is a data scientist who would like to analyse the movie ratings on IMDB for movies of different genres. She found a dataset shared by a user on Kaggle that contains information of over 14,000 movies. Great!

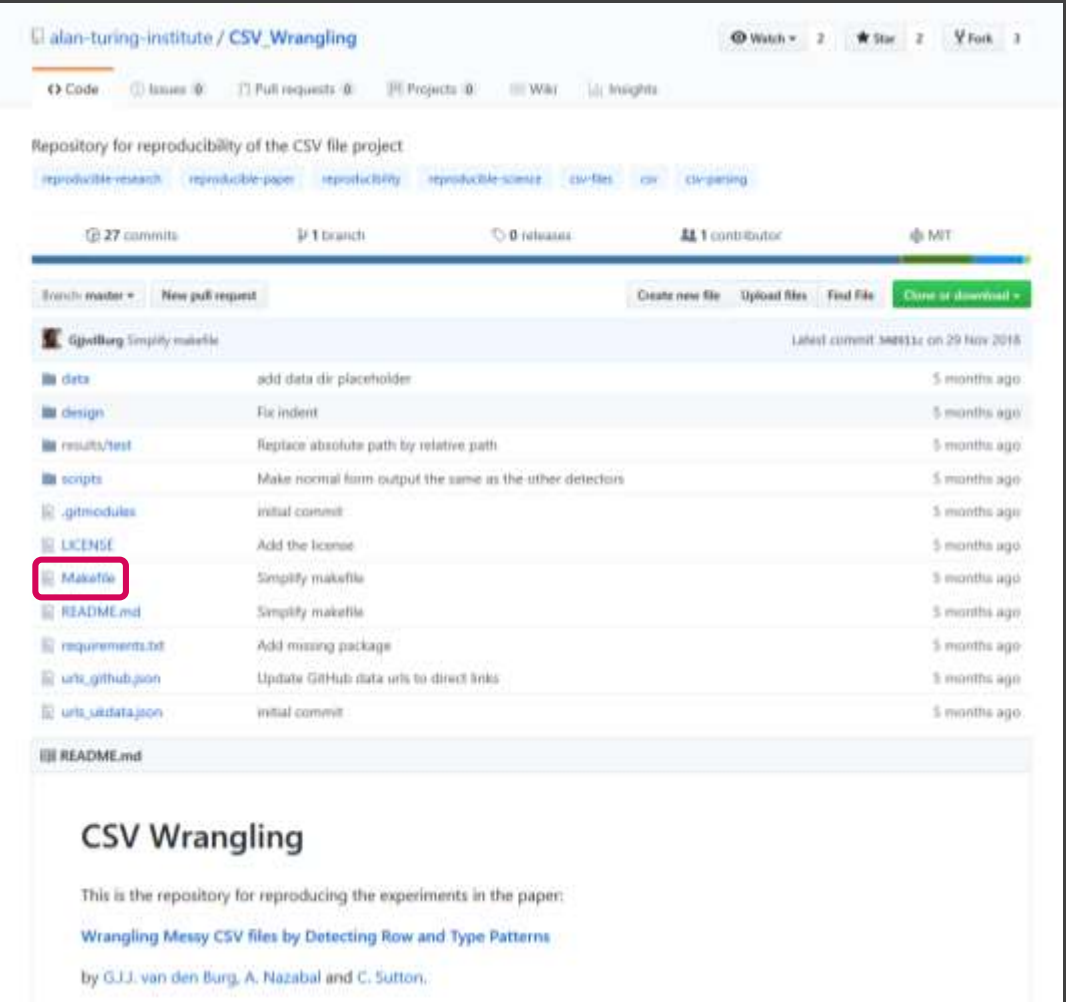
The data is stored in a CSV file, which is a very common data format for sharing tabular data. The first few lines of the file look like this:

https://github.com/alan-turing-institute/CleverCSVDemo
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



- https://github.com/alan-turing-institute/CSV_Wrangling
- “Wrangling Messy CSV Files by Detecting Row and Type Patterns” arXiv:1811.11242

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



- https://github.com/alan-turing-institute/CSV_Wrangling
- “Wrangling Messy CSV Files by Detecting Row and Type Patterns” arXiv:1811.11242

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The Turing Way

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Reproducible Environments
6. Testing
7. Reviewing
8. Continous Integration
9. Research Data Management
10. Reproducible Research with Make

What is Make

Make is a build automation tool. It uses a configuration file called a Makefile that contains the *rules* for what to build. Make builds *targets* using *recipes*. Targets can optionally have *prerequisites*. Prerequisites can be files on your computer or other targets. Make determines what to build based on the dependency tree of the targets and prerequisites (technically, this is a directed acyclic graph). It uses the *modification time* of prerequisites to update targets only when needed.

Why use Make for Reproducible Research?

There are several reasons why Make is a good tool to use for reproducible research:

1. Make is available on many platforms
2. Make is easy to learn
3. Makefiles are text files, which makes them easy share and keep in version control.
4. Many people are already familiar with Make
5. Using Make doesn't exclude using other tools such as Travis, Docker, etc.

Learn Make by Example

One of the things that might scare people off from using Make is that existing Makefiles can seem daunting and it may seem difficult to tailor to your own needs. In this hands-on tutorial we will

https://the-turing-way.netlify.com/make/make.html
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The Turing Way

1. Introduction
2. Reproducibility
3. Open Research
4. Version Control
5. Reproducible Environments
6. Testing
7. Reviewing
8. Continous Integration
9. Research Data Management
10. Reproducible Research with Make

What is Make

Make is a build automation tool. It uses a configuration file called a Makefile that contains the *rules* for what to build. Make builds *targets* using *recipes*. Targets can optionally have *prerequisites*. Prerequisites can be files on your computer or other targets. Make determines what to build

tree of the targets and prerequisites (technically, this is a directed

nodification time of prerequisites to update targets only when needed.

ducible Research?

why Make is a good tool to use for reproducible research:

any platforms

which makes them easy share and keep in version control.

ly familiar with Make

clude using other tools such as Travis, Docker, etc.

No chapter on citing software….

Learn Make by Example

One of the things that might scare people off from using Make is that existing Makefiles can seem daunting and it may seem difficult to tailor to your own needs. In this hands-on tutorial we will

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

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

- Check analysis on my phone
- Share the responsibility with busy PIs
- Requires version control, capturing environment and new build for each change

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Software as infrastructure

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Project Jupyter

- A community of people and an ecosystem of open tools and standards for interactive computing.
- Empower people to use other open tools.
- Slides by Chris Holdgraf (thank you!)

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

The science is the code

An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The ***actual scholarship*** *is the complete software development environment and the complete set of instructions which generated the figures.*

Buckheit and Donoho
(paraphrasing John Claerbout)
WaveLab and Reproducible Research, 1995

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Vicky Steeves (joinmastodon.org) @VickySteeves · May 9
Buckheit and Donoho quote, DRINK! #reproducibility #csvconf

3 1 10

Kyle Cranmer @KyleCranmer · May 9
Lol... just now @fperez_org

2 7

Karthik Ram
@_inundata
Following

Replying to @KyleCranmer @VickySteeves @fperez_org

A talk about reproducible research is not scholarship itself. It is merely advertising of the Buckheit and Donoho quote.

10:17 PM - 9 May 2019 from Southbank, Melbourne

6 Retweets 13 Likes

3 6 13

code

outational science in a scientific
e scholarship itself, it is merely
arship. The ***actual scholarship*** *is*
development environment and the
tions which generated the figures.

uckheit and Donoho
hrasing John Claerbout)
d Reproducible Research, 1995

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

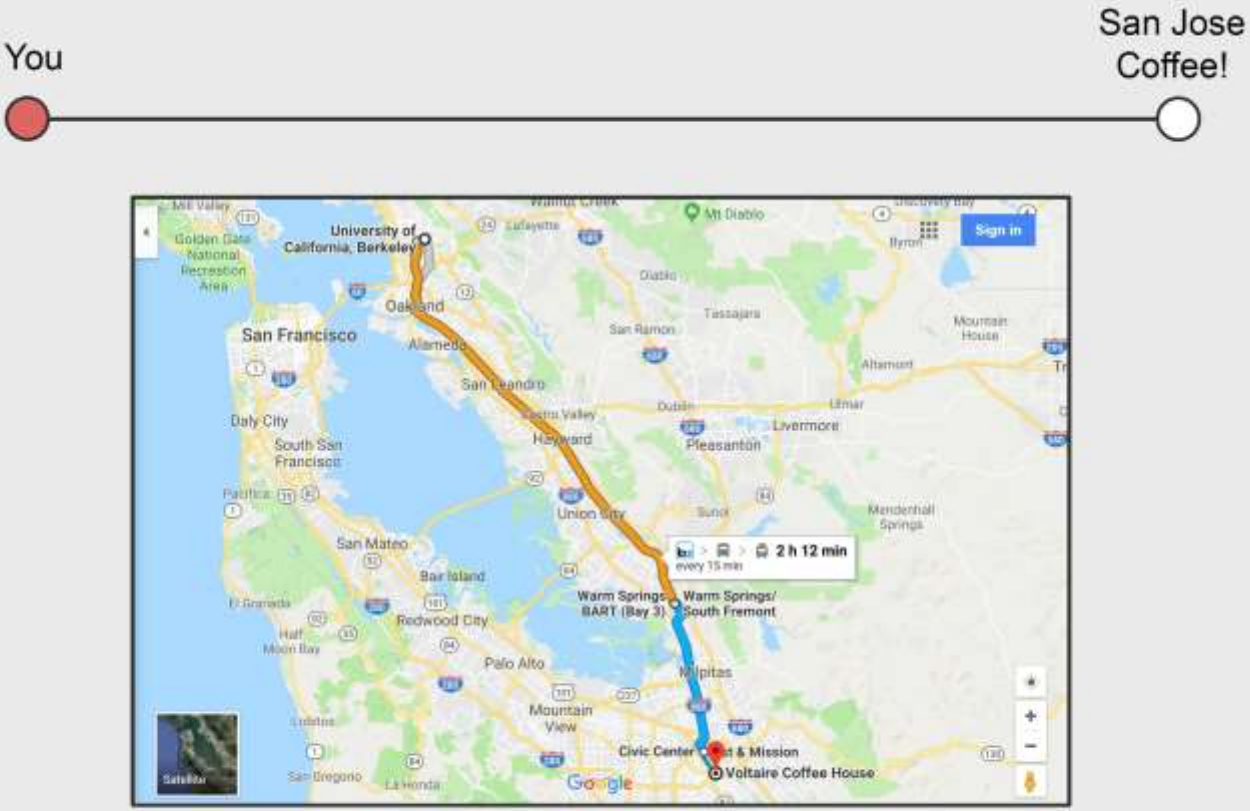
https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

One option: walk there by myself

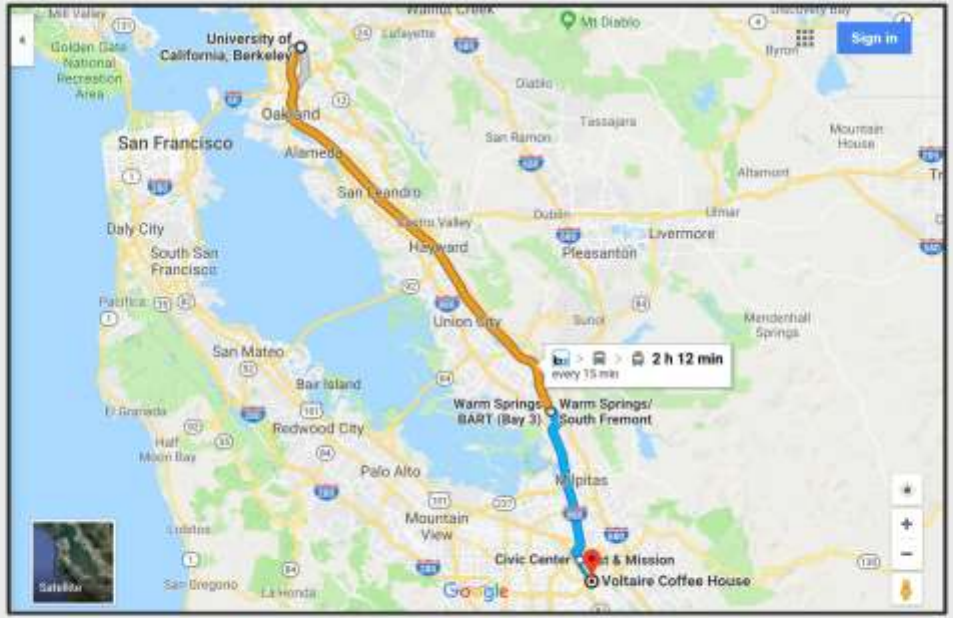
Another option: pay somebody to drive me

***My favorite option**: use public infrastructure*

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998



https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

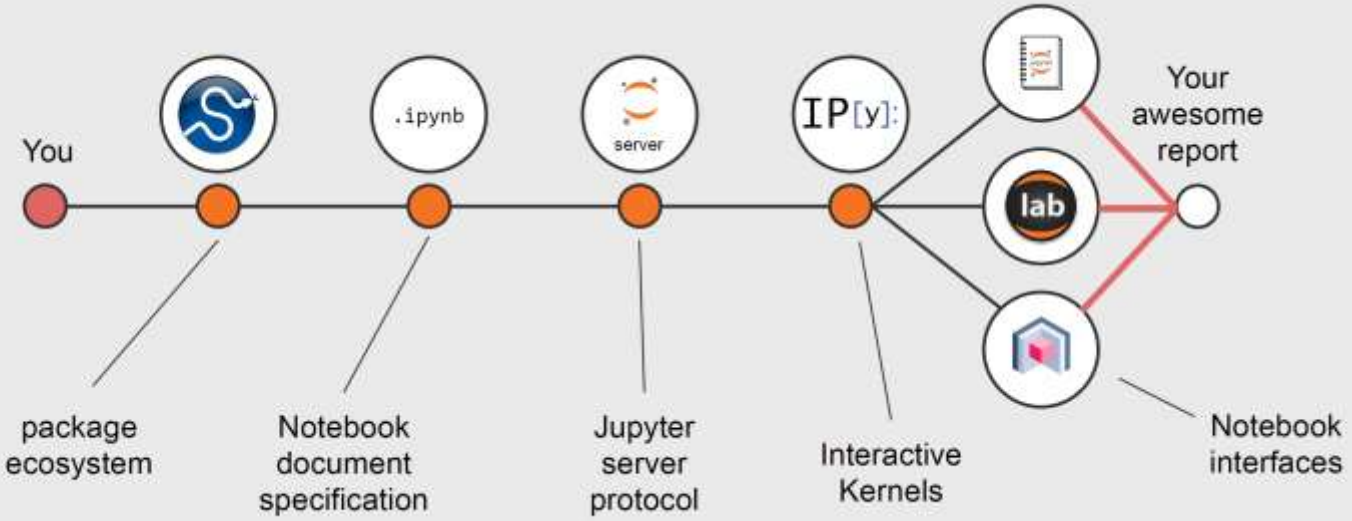
https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Public infrastructure gets us closer to our goal.

It makes the last mile shorter.

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Jupyter shortens the last mile by creating and leveraging public infrastructure



https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Public infrastructure gets us closer to our goal.

It makes the last mile shorter.

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Public infrastructure gets us closer to our goal.

It makes the last mile shorter.

https://doi.org/10.5281/zenodo.2747640
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Paying taxes is good

- We benefit from shared resources
- Some are so fundamental that we take them for granted
- We need them to get our jobs done

Tax summary description	Description of PESA source (See PESA Table 5.2)	Public Sector Expenditure (£bn)	%
Welfare	'Social Protection' excluding state pensions	174.4	23.8
Health	Health	145.8	19.9
State Pensions	Within 'Social Protection' [1]	93.8	12.8
Education	Education	87.8	12.0
National Debt interest	Within General Public Services, but shown in more detail in table 5.2	44.5	6.1
Defence	Defence	38.7	5.3
Public Order & Safety	Public Order & Safety	31.6	4.3
Transport	Economic Affairs, without Business and Industry but shown in more detail in table 5.2	31.2	4.3
Business & Industry	Economic Affairs, without Transport	21.4	2.9
Government Administration	Captured under General Public Services, but shown in more detail in table 5.2	15.2	2.1
Environment	Environment protection	11.4	1.6
Culture (e.g. sports, libraries, museums)	Recreation, Culture & Religion	11.8	1.6
Housing and utilities (e.g. street lights)	Housing & Community Amenities	12.1	1.6
Overseas Aid	Captured under General Public Services, but shown in more detail in table 5.2	8.8	1.2
UK Contributions to EU budget	EU Transactions	5.4	0.7

https://www.gov.uk/government/publications/how-public-spending-was-calculated-in-your-tax-summary/how-public-spending-was-calculated-in-your-tax-summary

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Academi-coin

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

- Citations are academic currency (whether they should be or not!)
- They’re the best way we have to endorse good work.
- We should be citing the software we use.

#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

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
#CiteSoftware #TuringWay @kirstie_j
https://doi.org/10.5281/zenodo.2783998

Next steps

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

The humans are the hardest part of reproducibility and of software citation

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

How can we change researchers' behaviour?

- **Handbook**, a place to capture knowledge easily, no excuse that they didn't know to/how
- **Checklists**, for researchers, PIs, funders and business team members
- **Technology**, to make it easy to cite the work
- **Case studies**, to show that it can be done
- **Community**, to advocate for change

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

It takes a village

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Rachael Ainsworth

Becky Arnold

Louise Bowler

Sarah Gibson

Patricia Herterich

James Hetherington

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Rosie Higman

Anna Krystalli

Catherine Lawrence

Alex Morley

Martin O’Reilly

Binder Team

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998

Thank you

- https://the-turing-way.netlify.com
- https://tinyletter.com/TuringWay
- https://github.com/alan-turing-institute/the-turing-way
- https://gitter.im/alan-turing-institute/the-turing-way
- Unsplash photos by Freddy Castro, James Pond, Kinson Leung, Mateo Vrbnjak, Mimi Thian, Omar Albeik, Perry Grone, Toa Heftiba, Tomasz Frankows, Jonathan Brinkhorst, Eric Weber
- Noun Project icons by Aybige, Luis Prado, Edward Boatman, Becris, Rose Alice Design, Hyemm.work

#CiteSoftware #TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.2783998