
The Turing Way

A handbook for reproducible research

Kirstie Whitaker

Collaborations Workshop Demo, 2 April 2019

Slides at <https://doi.org/10.5281/zenodo.2621280>





Neurohackweek 2016

Photo credit: Chris Gorgolewski

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



Our university network



THE UNIVERSITY
of EDINBURGH



UNIVERSITY OF
BIRMINGHAM



UNIVERSITY OF LEEDS



The University of Manchester

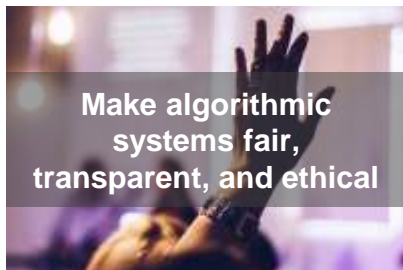


Newcastle
University
UK | Malaysia | Singapore

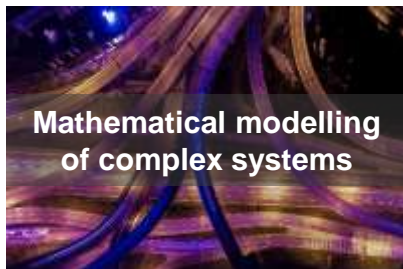
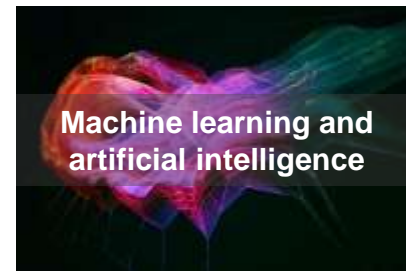
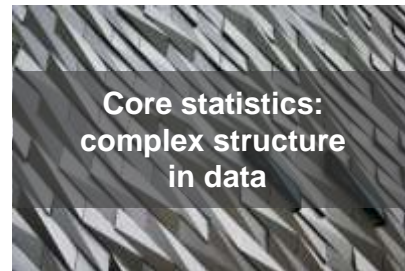


Challenges

Advance data science and artificial intelligence to...



Core capabilities



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

Tuesday 18 Dec 2018

Learn more ↓

<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.

Cross cutting theme: Tools, systems and practices



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.

Fraud is not our biggest problem



SPRINGER NATURE

Errors in the literature 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-rogooff-i-think-its-best-to-admit-your-errors-and-go-on-from-there>

<https://www.bbc.co.uk/news/magazine-22223190>

Errors in the literature 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 NEWS Sign in News Sport Weather iPlayer Sounds

Home UK World Business Politics Tech Science Health Family & Education

Magazine

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

By Ruth Alexander BBC News

© 20 April 2013

f t e 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>

Explicitly replicating research is very (very) hard



The Reproducibility Project: Cancer Biology is an initiative to independently replicate selected results from a number of high-profile papers in the field of cancer biology. For each paper a Registered Report detailing the proposed experimental designs and protocols for the experiments is peer reviewed and published prior to data collection; the results of these experiments are then published as a Replication Study. The project is a collaboration between the Center for Open Science and Science Exchange.

The aim of the project is two-fold: to provide evidence about reproducibility in preclinical cancer research, and to identify the factors that influence reproducibility more generally. Interpreting the results reported in the Replication Studies requires a nuanced approach, as explained in this Editorial. To date four of the studies have reproduced important parts of the original papers; four of the studies have reproduced parts of the original papers but also contain results that could not be interpreted or are not consistent with some parts of the original paper; two of the studies could not be interpreted; and two studies did not reproduce the parts of the original papers that they attempted to reproduce.

COLLECTION

RELIABILITY TEST
PLAN
Rep who ant
Cym
DEC
RELI
Rep
Get
Don
PLAN



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

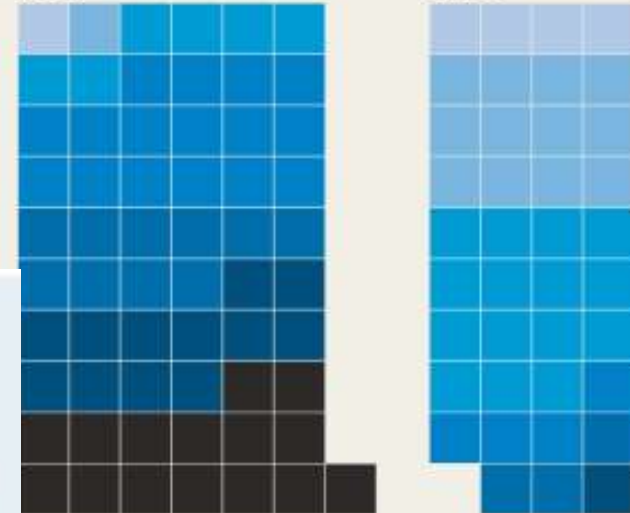
RELIABILITY TEST

An effort to reproduce 100 psychology findings found that only 39 held up.* But some of the 61 non-replications reported similar findings to those of their original papers.

Did replicate match original's results?

NO: 61

YES: 39



Replicator's opinion: How closely did findings resemble the original study:

- Virtually identical
- Extremely similar
- Very similar
- Moderately similar
- Somewhat similar
- Slightly similar
- Not at all similar

* based on criteria set at the start of each study

<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>

Tools, Practices and Systems

- Focus on real cross-project needs
 - Driven by 'researcher pain points'.
- We will not make things just because we think they're interesting.
 - Usefulness to applied researchers is key.



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

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

Support additional users

Held to higher standards
than others

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

Support additional users

Held to higher standards
than others



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



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



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
Stanford
team lead
data, 📦



Zach Sailer
Project Jupyter
team blue
📦, 🗺️, 💬



Erik Sundell
Sandvik CODE
team blue
📦, 🗺️



Carol Willing
Project Jupyter
team red
Python, Community

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


What is Jupyter Book?

*Build an online book with
Jupyter Notebooks and Markdown*




jupyter.org/jupyter-book





← TOGGLE SIDEBAR

 Interact

Introduction

Search

- 1. Data Science
- 2. Causality and Experiments
- 3. Programming in Python
- 4. Data Types
- 5. Sequences
- 6. Tables
- 7. Visualization
- 8. Functions and Tables
- 9. Randomness
- 10. Sampling and Empirical Distributions**
 - 10.1 Empirical Distributions
 - 10.2 Sampling from a Population
 - 10.3 Empirical Distribution of a Statistic
- 11. Testing Hypotheses

Sampling and Empirical Distributions

An important part of data science consists of making conclusions based on the data in random samples. In order to correctly interpret their results, data scientists have to first understand exactly what random samples are.

In this chapter we will take a more careful look at sampling, with special attention to the properties of large random samples.

Let's start by drawing some samples. Our examples are based on the `top_movies.csv` data set.

```
top1 = Table.read_table(path_data + 'top_movies.csv')
top2 = top1.with_column('Row Index', np.arange(top1.num_rows))
top = top2.move_to_start('Row Index')

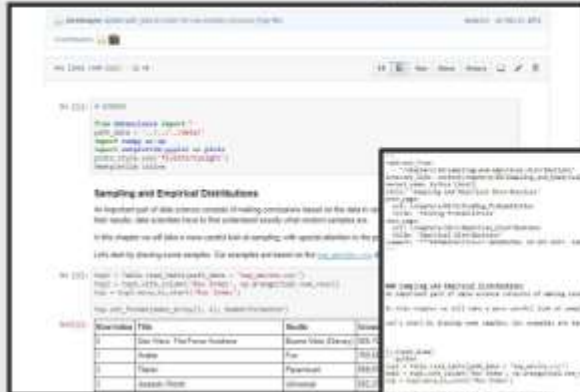
top.set_format(make_array(3, 4), NumberFormatter)
```

Row Index	Title	Studio	Gross	Gross (Adjusted)	Year
0	Star Wars: The Force Awakens	Buena Vista (Disney)	906,723,418	906,723,400	2015
1	Avatar	Fox	760,507,625	846,120,800	2009
2	Titanic	Paramount	658,672,302	1,178,627,900	1997
3	Jurassic World	Universal	652,270,625	687,728,000	2015

ON THIS PAGE

- SAMPLING AND EMPIRICAL DISTRIBUTIONS**
- SAMPLING ROWS OF A TABLE
- DETERMINISTIC SAMPLING
- K-RANDOM SAMPLING
- SCHEMES
- A STATISTICAL SAMPLE
- RANDOM SAMPLES DRAWN WITH OR WITHOUT REPLACEMENT

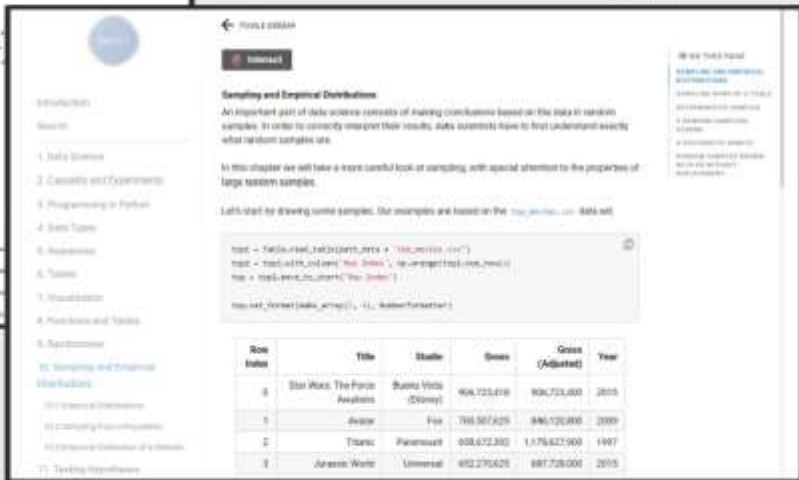
mypage.ipynb



mypage.md



mypage.html



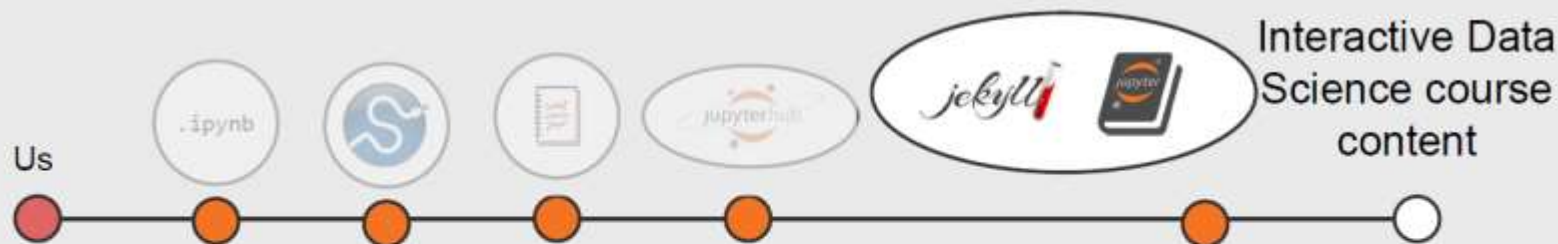
Interactive buttons let readers explore

- A page built from a notebook gets an interact button
- Use JupyterHub/Binder to spawn a Jupyter server
- `git pull` the underlying notebook for the page
- Initialize an interactive environment



In summary

Jupyter Book builds on tools in the Jupyter ecosystem to create interactive, beautiful books.



jupyter.org/jupyter-book



The Turing Way

A lightly opinionated handbook
for reproducible data science

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

The Turing Way

A lightly opinionated **handbook**
for reproducible data science

[https://github.com/alan-turing-institute/
the-turing-way-book](https://github.com/alan-turing-institute/the-turing-way-book)

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

Requires additional skills

Chapters will include:








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

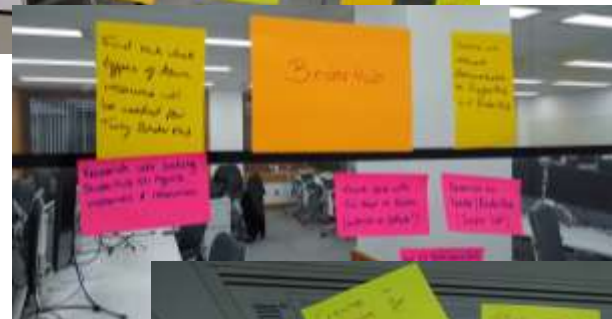


https://github.com/alan-turing-institute/the-turing-way/blob/master/book_skeleton.md

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
















Built by a team....and you!

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



<https://github.com/alan-turing-institute/the-turing-way/blob/master/contributors.md>

The emoji key to celebrate our contributors

 <p>Becky Arnold</p> <p>🗨️ 💻 📖 😞 👁️</p>	 <p>Louise Bowler</p> <p>🗨️ 💻 📖 💡 😞 📄 👁️</p>	 <p>Jason M. Gates</p> <p>📖 👁️</p>	 <p>sgibson91</p> <p>🗨️ 💻 📖 🔧 👁️ 📣 😞 ✅</p>	 <p>Tim Head</p> <p>🗨️ 😞</p>	 <p>Patricia Herterich</p> <p>🗨️ 📖 👁️ 😞 🖋️</p>	 <p>Rosie Higman</p> <p>🗨️ 📄 👁️ 😞</p>
 <p>Anna Krystalli</p> <p>🗨️ 💡 👁️ 😞</p>	 <p>Alexander Morley</p> <p>🗨️ 👁️ 😞 ⚠️</p>	 <p>Martin O'Reilly</p> <p>🗨️ 🔧 😞</p>	 <p>Oliver Strickson</p> <p>🗨️ 📖</p>	 <p>Kirstie Whitaker</p> <p>🗨️ 📖 😞 📄 🔍 😞 👁️ 📣</p>	 <p>Chris Holdgraf</p> <p>🗨️ 😞</p>	

<https://github.com/alan-turing-institute/the-turing-way/blob/master/README.md>

Open Leadership Principles



Understanding

You make the work accessible and clear

Read more

<https://mozilla.github.io/olm-whitepaper>



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.

[@kirstie_j](#)

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

Thank you

Lets discuss how we can work together!

kwhitaker@turing.ac.uk t: @kirstie_j g: @KirstieJane

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

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

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

tinyletter.com/TuringWay

doi: 10.5281/zenodo.2621280

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

Held to higher standards than others

*Make
reproducibility,
“too easy
not to do”*

*Share the
responsibility
of
reproducibility*
























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!

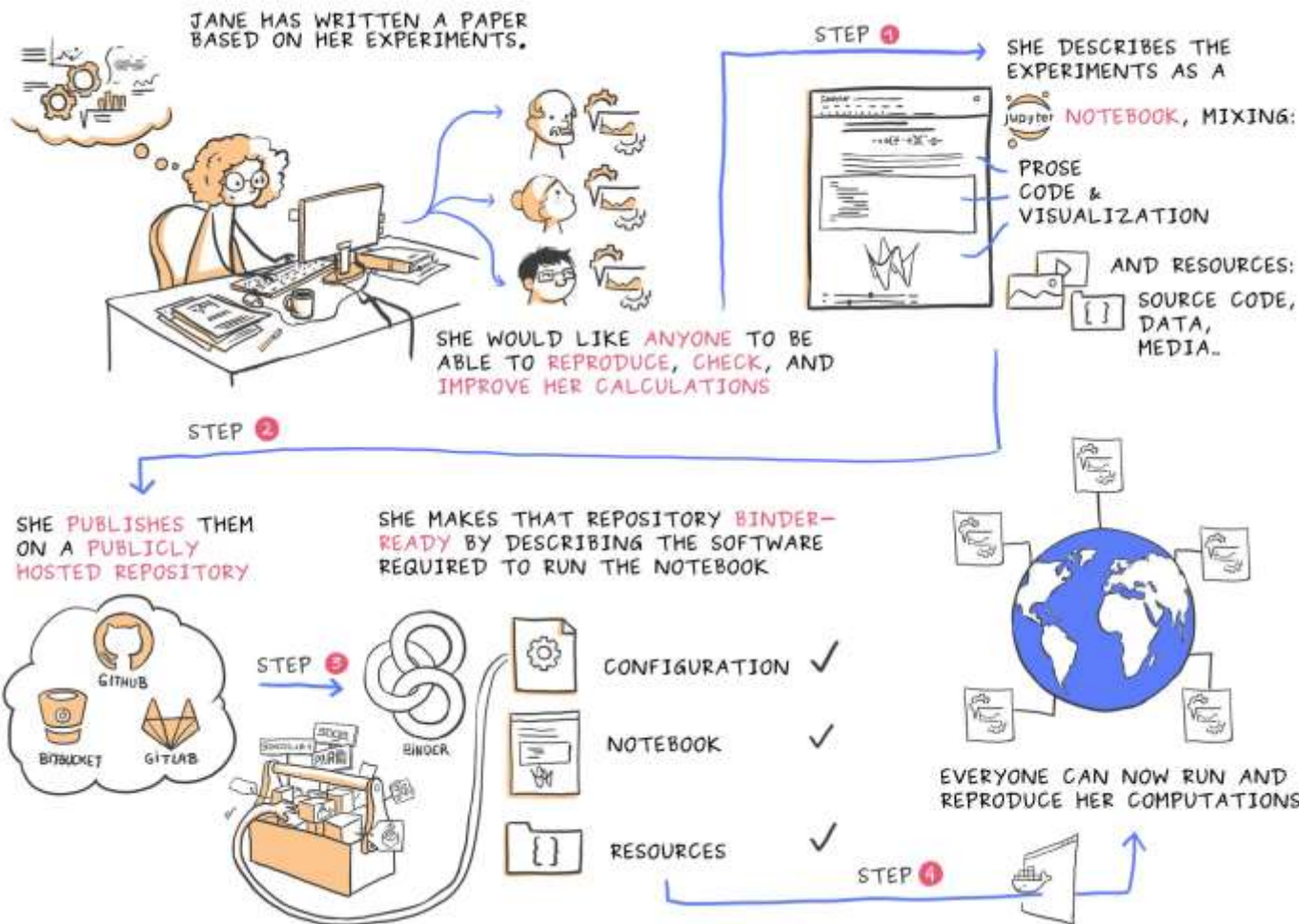
Binder Team

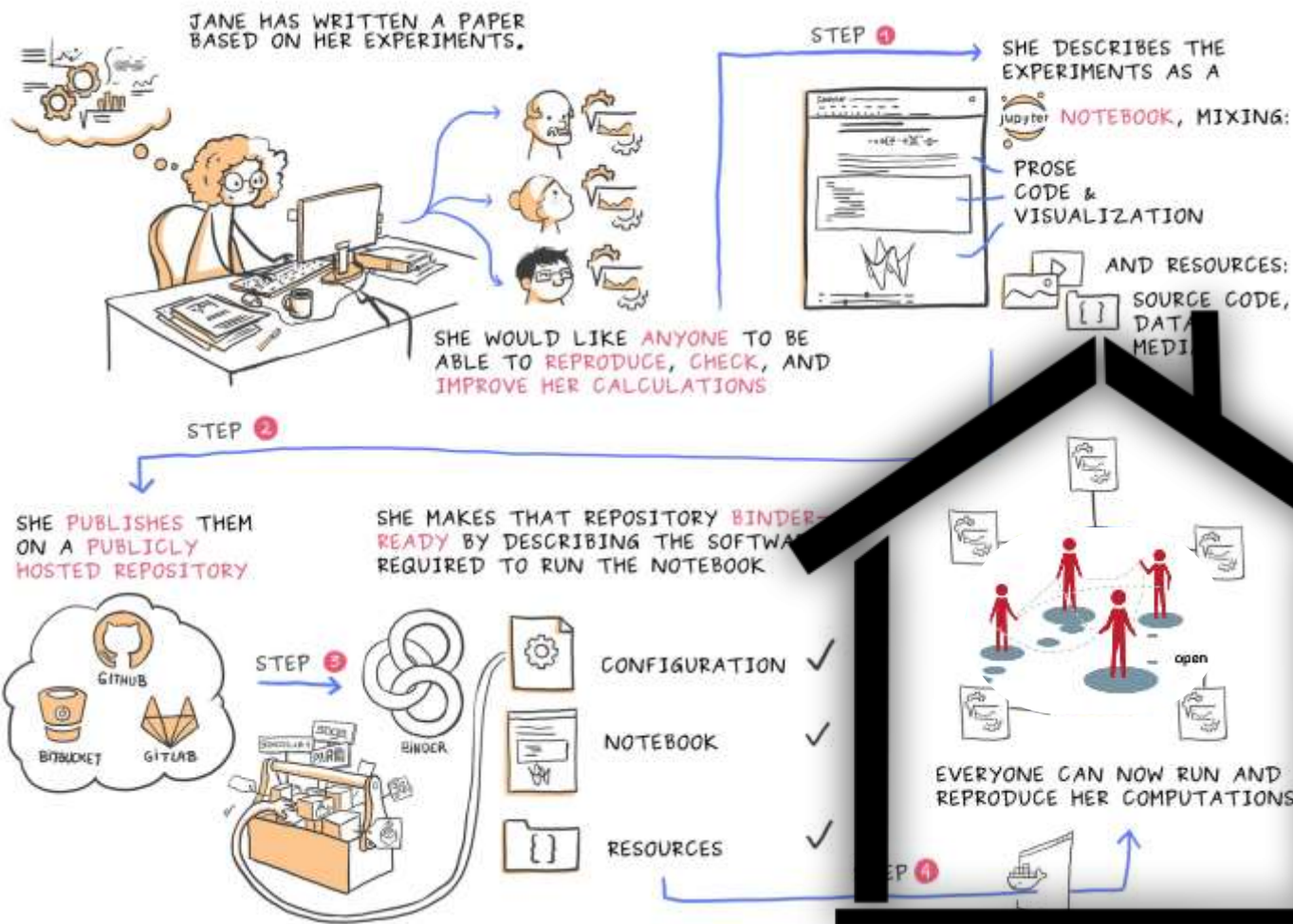
Binder's governance and team structure is defined in the [Binder Project Governance](#) page. Below we list the current team members of Binder.

(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>





Courtesy of Juliette Belin: <https://twitter.com/JulietteTaka/status/1062735653929000960>

sgibson91 / magprop

Watch 0 Unstar 1 Fork 0

Code Issues 2 Pull requests 1 Projects 1 Insights Settings

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

Edit

python27 mcmc astrophysics gamma-ray-astronomy gamma-ray-burst modeling emcee Manage topics

33 commits 7 branches 0 releases 1 contributor MIT

Tree: ff527ae769 New pull request

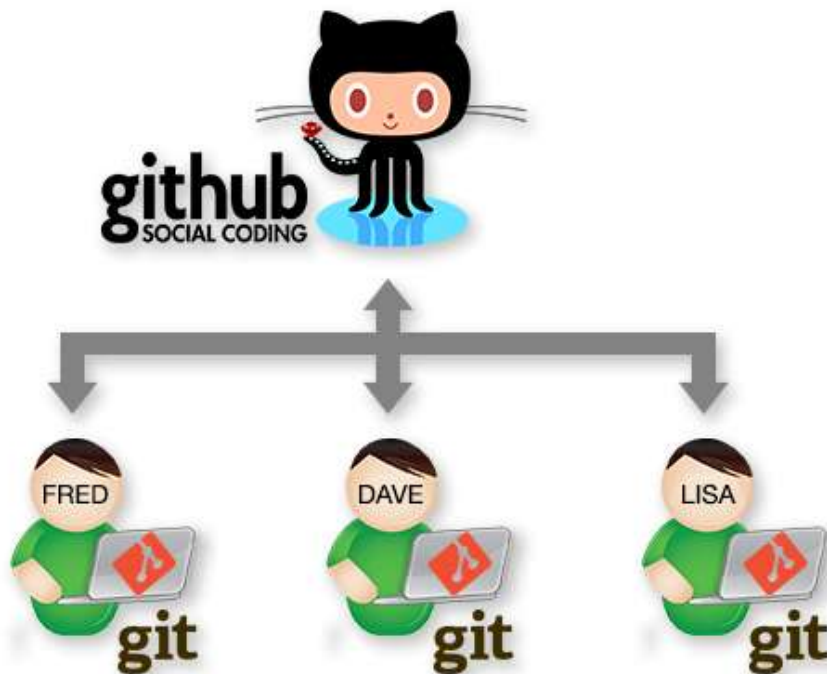
Create new file Upload files Find file Clone or download

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

Version control



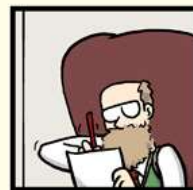
"FINAL".doc



FINAL.doc!



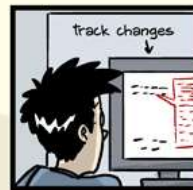
FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



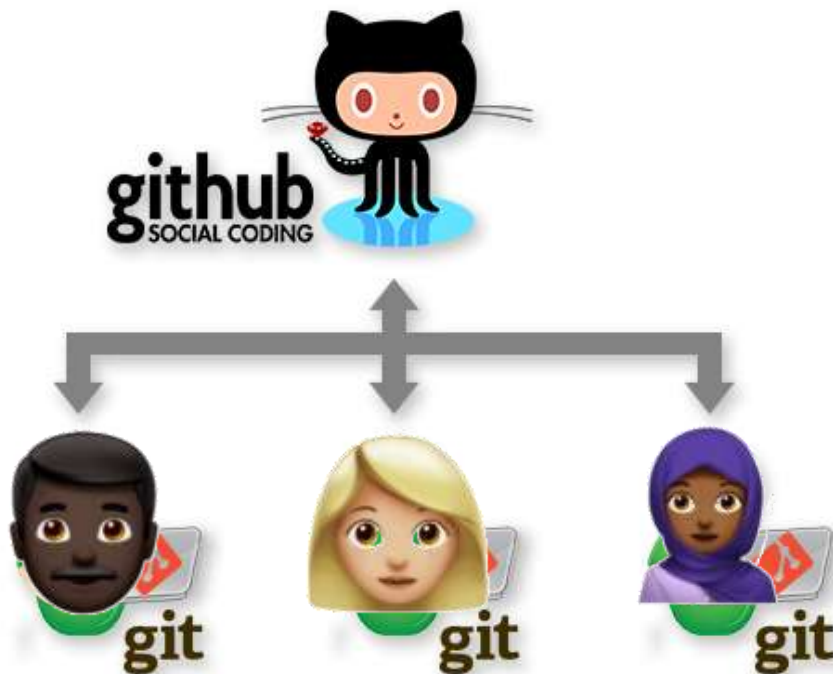
FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.##\$%WHYDID
ICOMETOGRADSCHOOL?????do



Version control



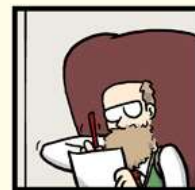
"FINAL".doc



FINAL.doc!



FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10.##\$%WHYDID
ICOMETOGRADSCHOOL?????.do



JORGE CHAN © 2012

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)

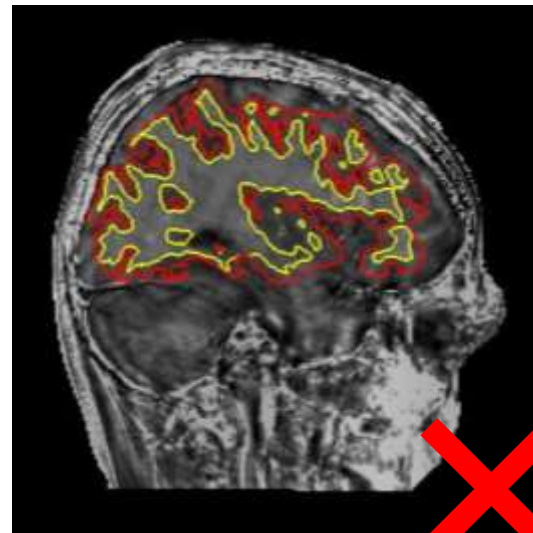
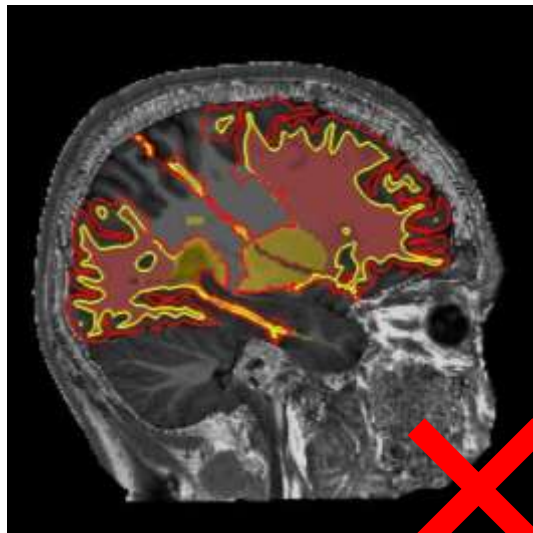
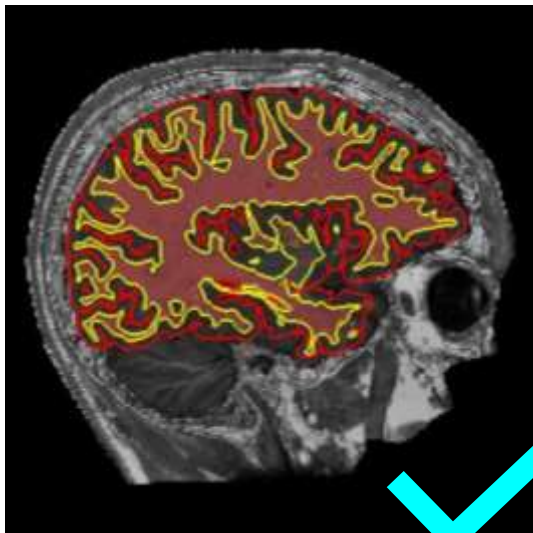
Is your code doing what you think it's doing? Does $2 + 2 = 4$?



```
Assert.AreEqual(  
  GetTimeOfDay(),  
  "Morning" )
```


Testing (aka making explicit sanity checks)

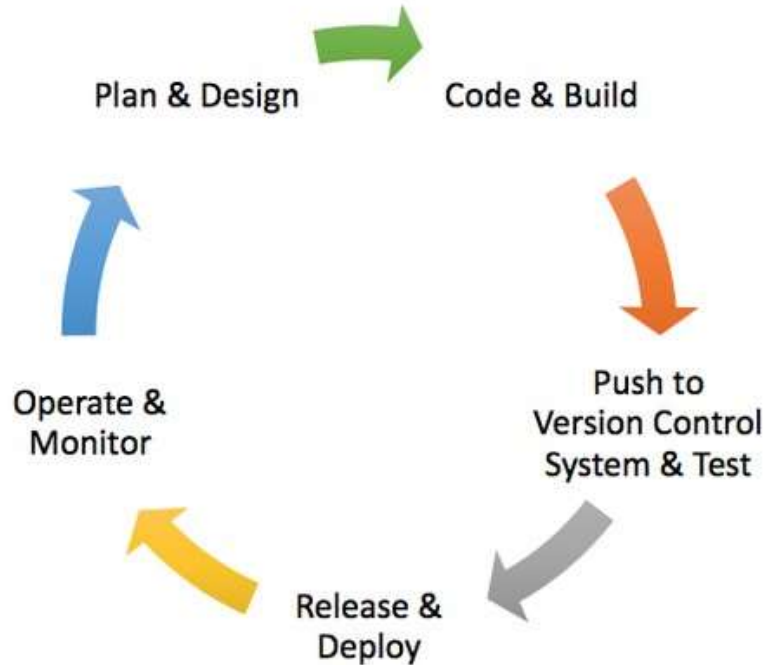
Is your code doing what you think it's doing? Does $2 + 2 = 4$?



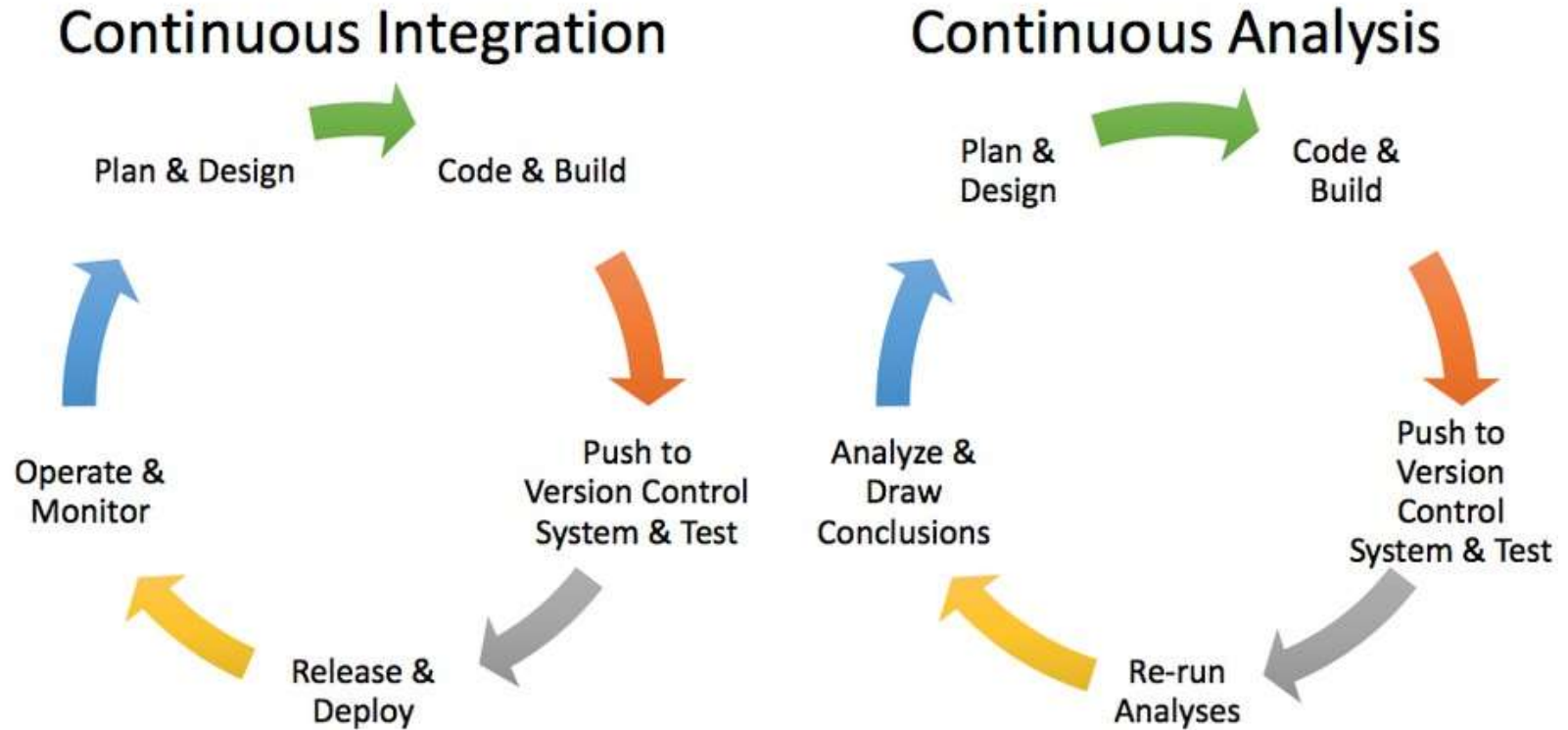
A very simple check: Is total brain volume within an expected range?

Continuous integration for research

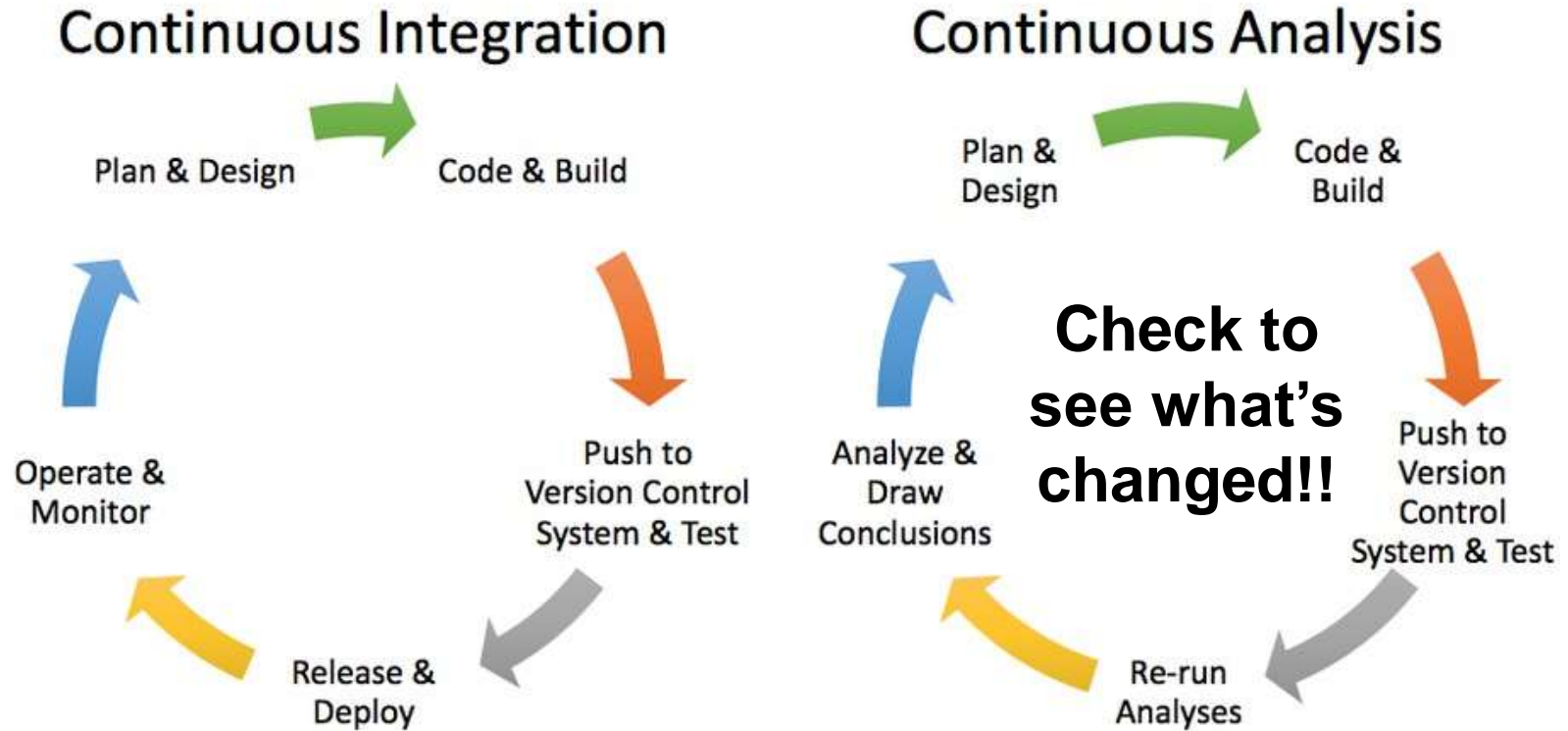
Continuous Integration



Continuous integration for research



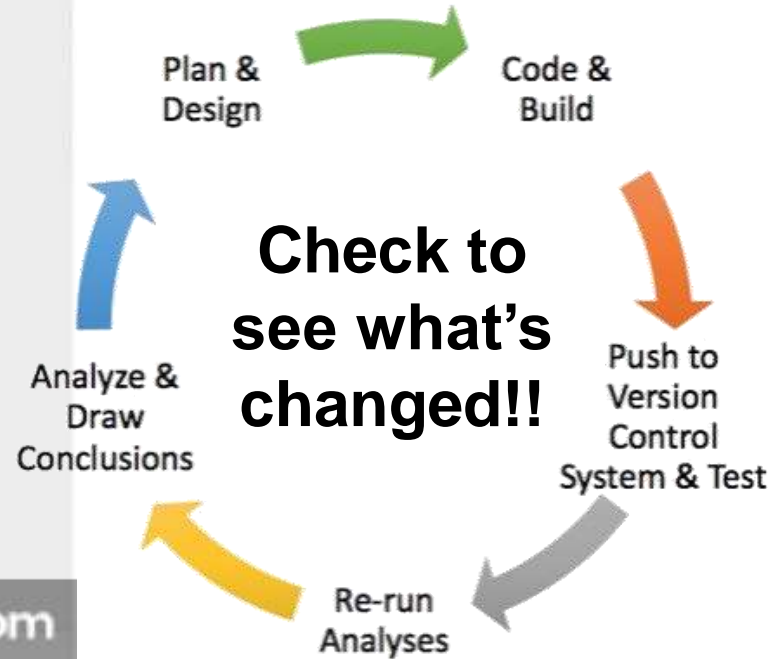
Continuous integration for research



Continuous integration for research



Continuous Analysis

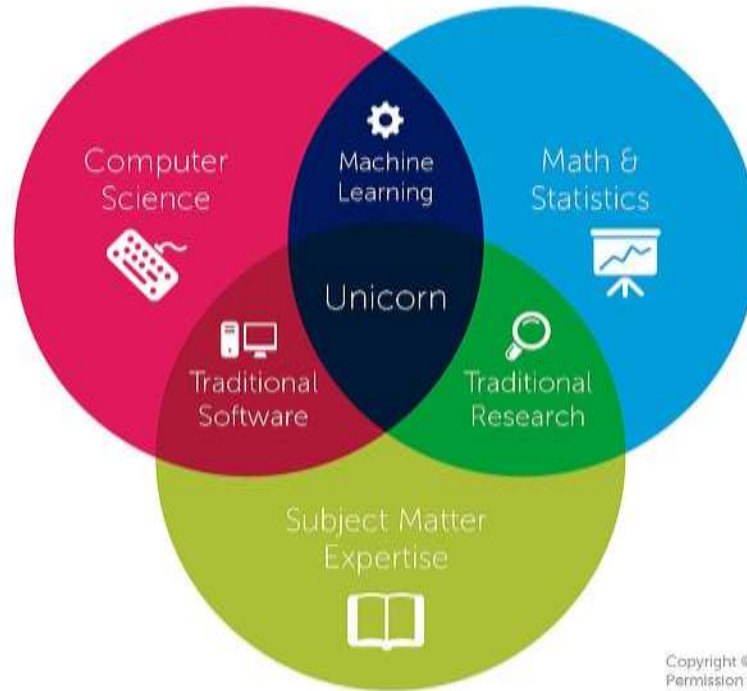




Building a culture of collaborative science

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

The Data Science Unicorn



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

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

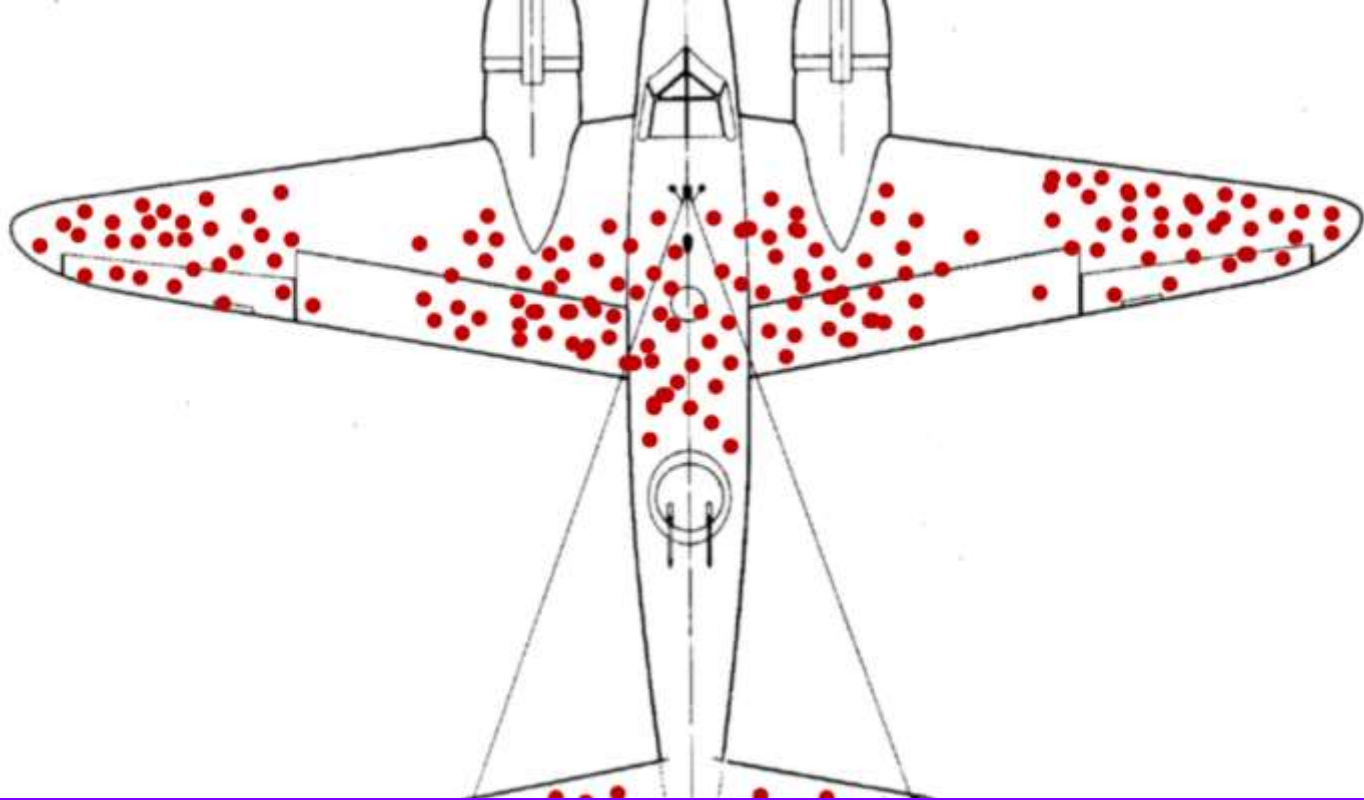


*How can we incentivise
team science?*

02/04/2019
<https://neurohackademy.org/apply>

Open is so much more than reproducible

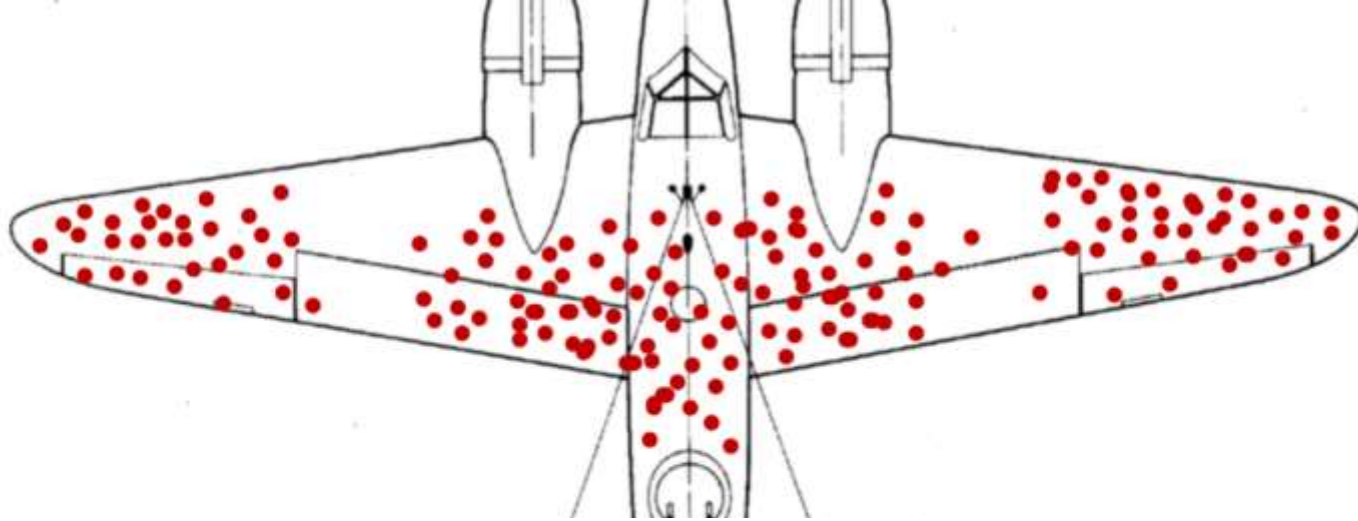




<https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d>

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

moz://a



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

Follow

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



<https://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](https://doi.org/10.6084/m9.figshare.7649156)