

---

# The Turing Way

Reproducible, Ethical, Collaborative Data Science

Kirstie Whitaker

Turing Health Conference, January 2019

Slides at <https://doi.org/10.6084/m9.figshare.7819442>



Neurohackweek 2016

Photo credit: Chris Gorgolewski

# 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 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



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.

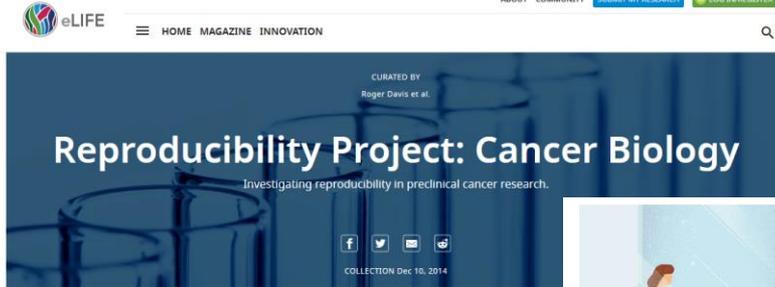


It's 4 January 2010, the Marriott Hotel in Atlanta. At the annual meeting of the American Economic Association, Professor Carmen Reinhart and the former chief economist of the International Monetary Fund, Ken Rogoff, are presenting a research paper called Growth in a Time of Debt.

<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



VIEWS 23,286

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 replications 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

ABOUT COMMUNITY [SUBMIT MY RESEARCH](#) [LOG IN/REGISTER](#)



DAVIDE BONAZZ

REL  
CANC  
Rep  
wh  
anc  
Cyn  
INSIG  
BIOC  
Rep  
Get  
Dirk  
INSIG

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

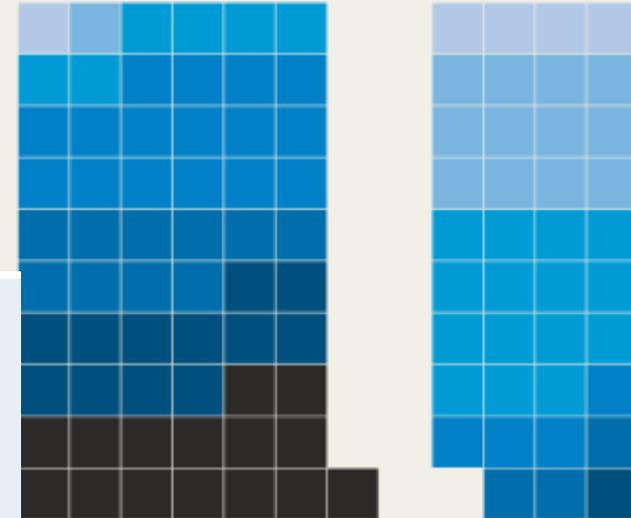
## RELIABILITY TEST

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>

---

# Fraud is not our biggest problem



**SPRINGER NATURE**

# 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>

---

# 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

<https://doi.org/10.6084/m9.figshare.7140050>

[github.com/alan-turing-institute/the-turing-way/blob/master/chapters/open\\_research.md](https://github.com/alan-turing-institute/the-turing-way/blob/master/chapters/open_research.md)

# 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
- . . .



---

# 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

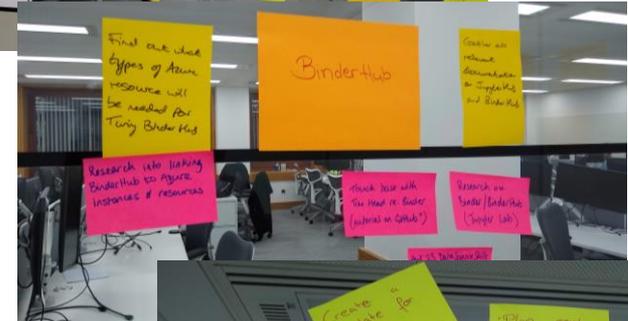
Support additional users

Held to higher standards  
than others

# 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



# Reproducible research champions



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

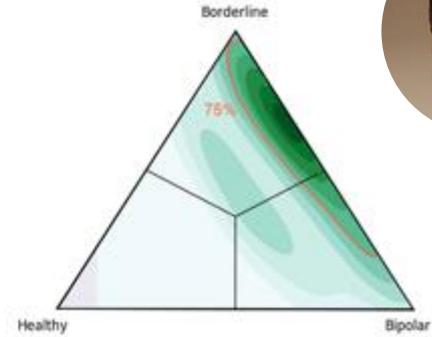
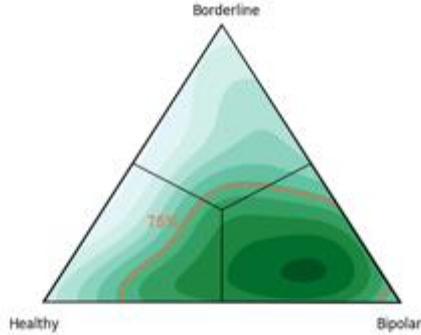
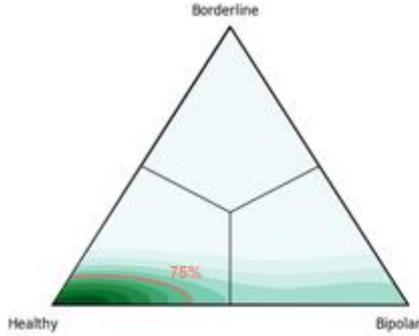
Imanol Perez Arribas, Guy Goodwin, John Geddes, Terry Lyons & Kate Saunders



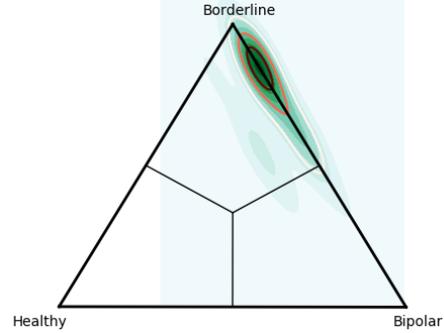
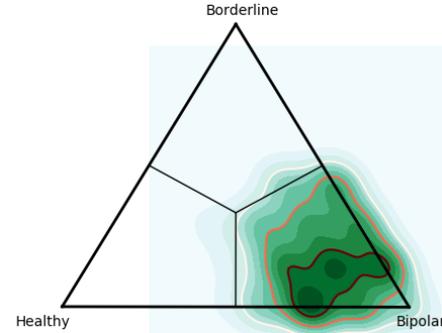
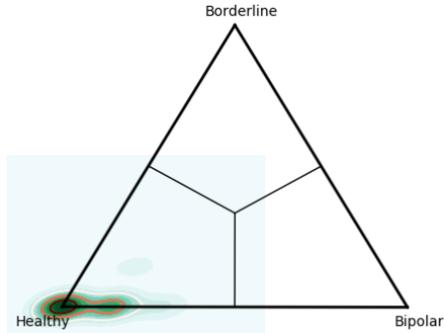
# Reproducible research champions



Publication



Synthetic Data



---

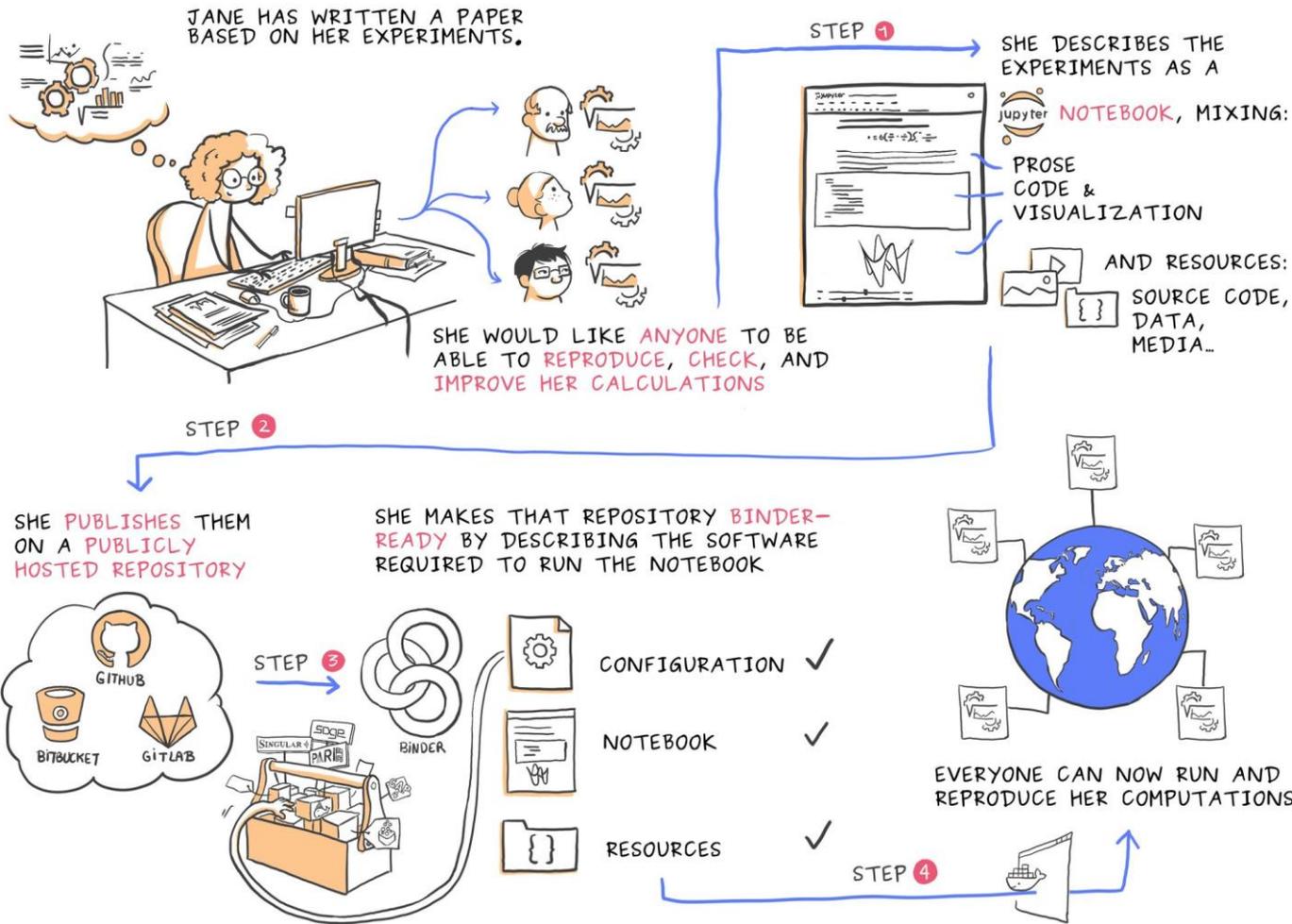
# Research engineering at the Turing

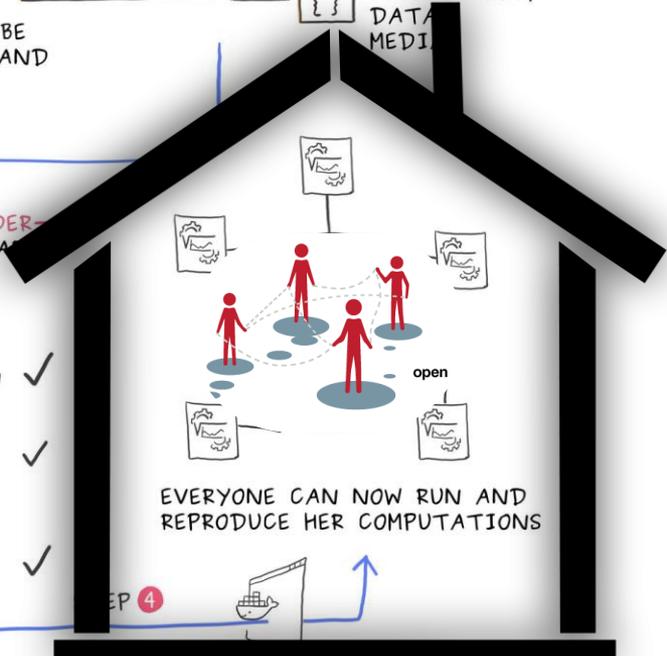
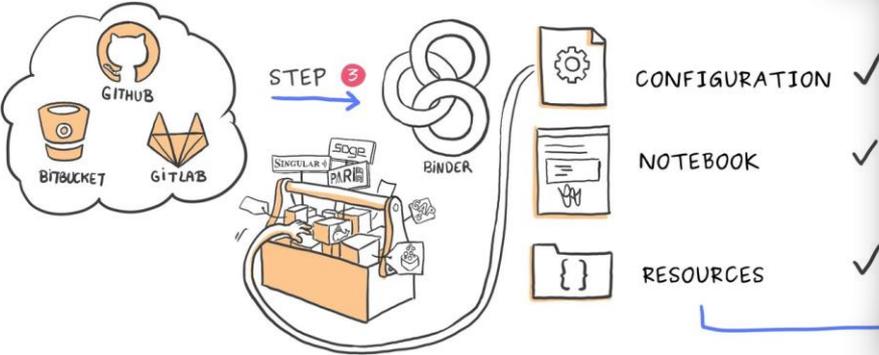
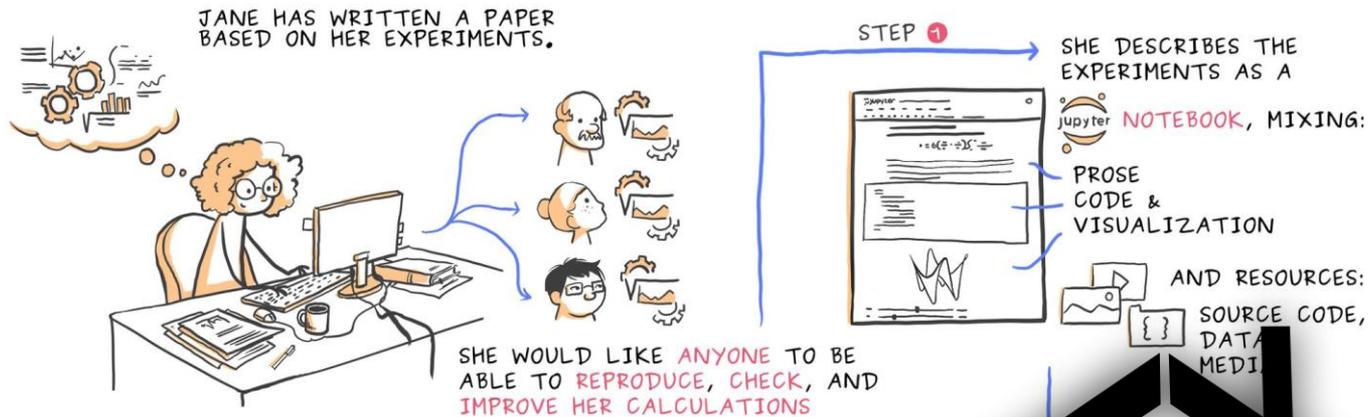


---

# Research engineering at the Turing







# AS OPEN AS POSSIBLE, AS CLOSED AS NECESSARY

Grantees have the right to opt-out, but need to say **why**



Top three reasons for opt-out:

privacy

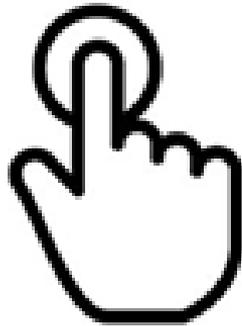
intellectual  
property rights

might jeopardise  
project's main  
objective

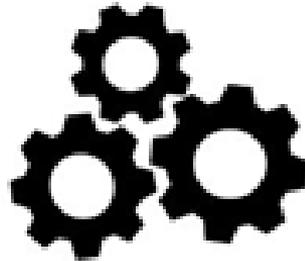
F  
Findable



A  
Accessible



I  
Interoperable



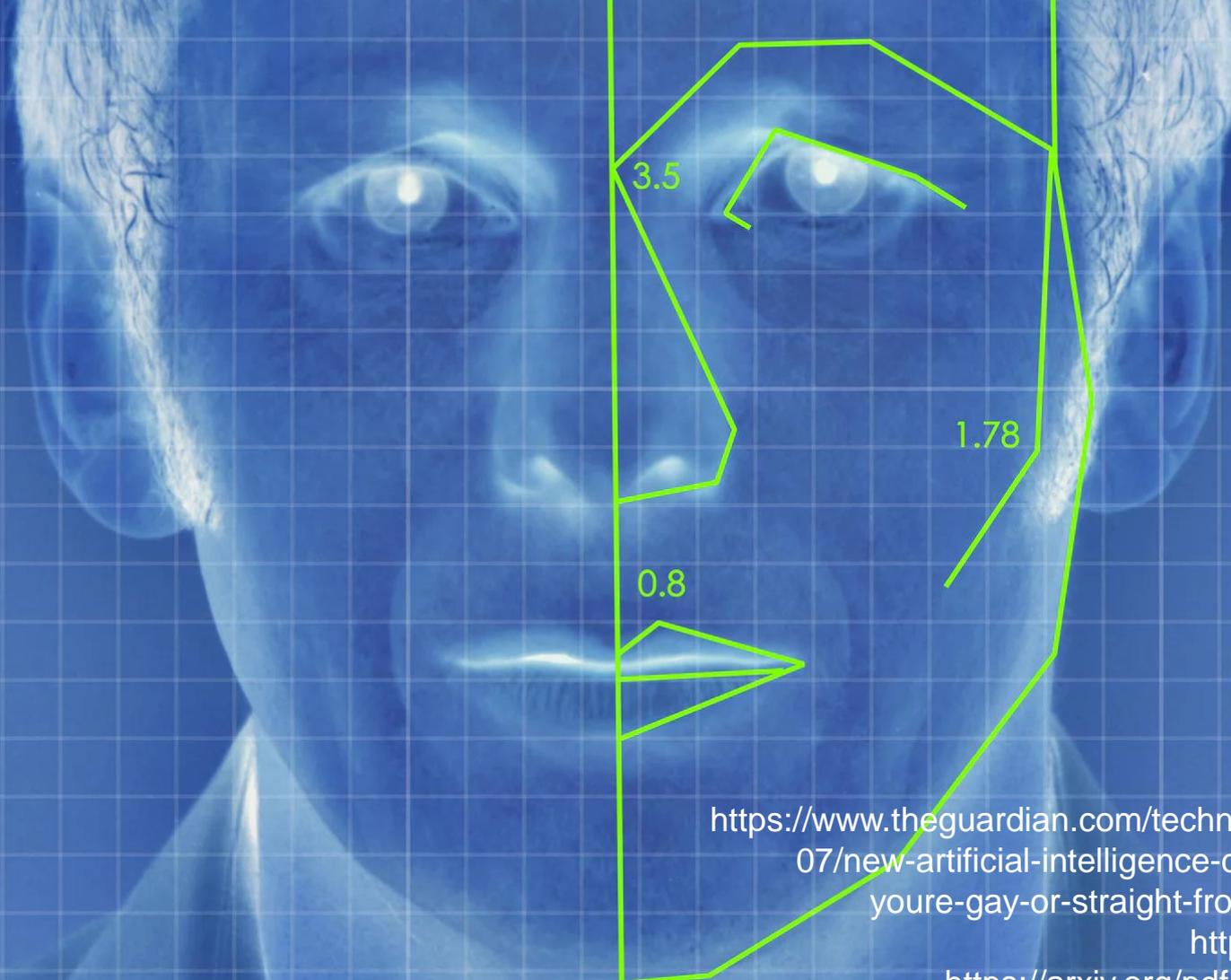
R  
Reusable



Wilkinson, M. D. *et al. Sci. Data* doi: 10.1038/sdata.2016.18 (2016).

Image credit: [Sangya Pundir](#)

45  
50  
55  
60  
65  
70  
75  
80  
85  
90  
95



3.5

1.78

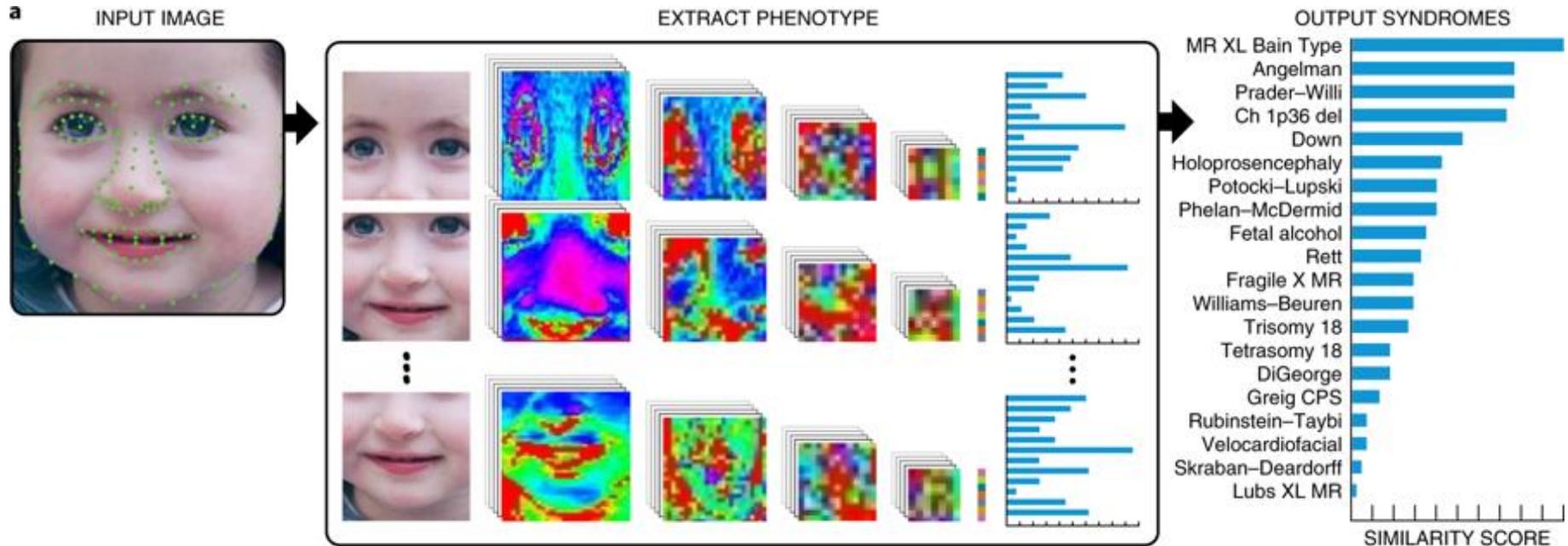
0.8

<https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph>

<https://osf.io/zn79k>

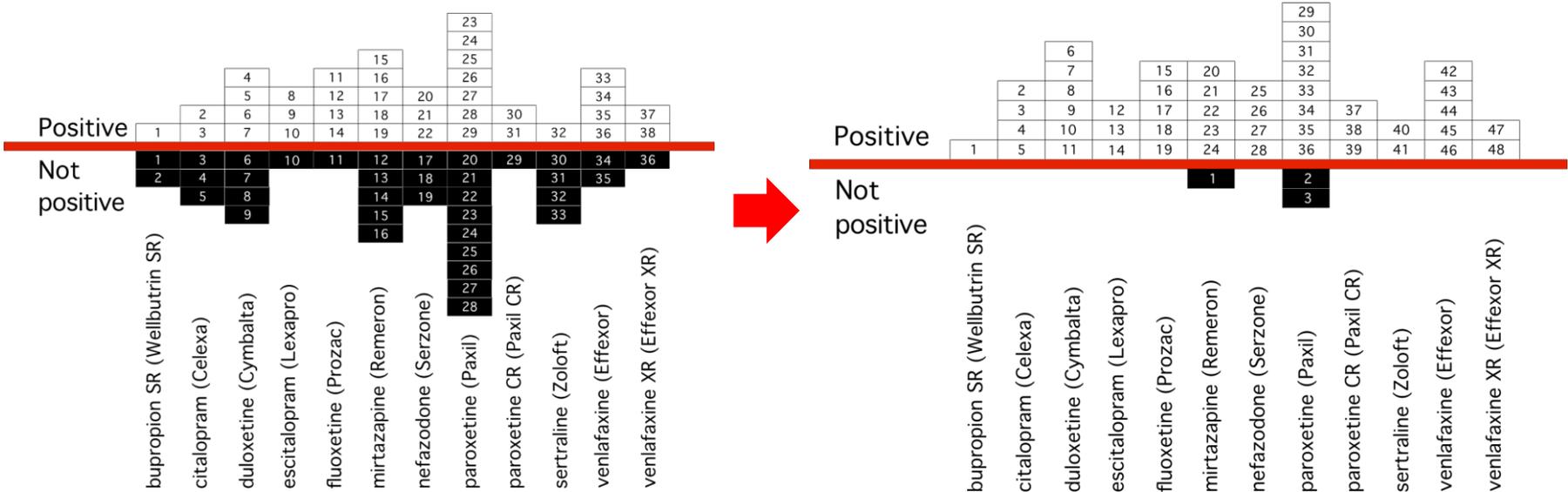
<https://arxiv.org/pdf/1902.10720.pdf>

# Facial recognition used to identify genetic disorders



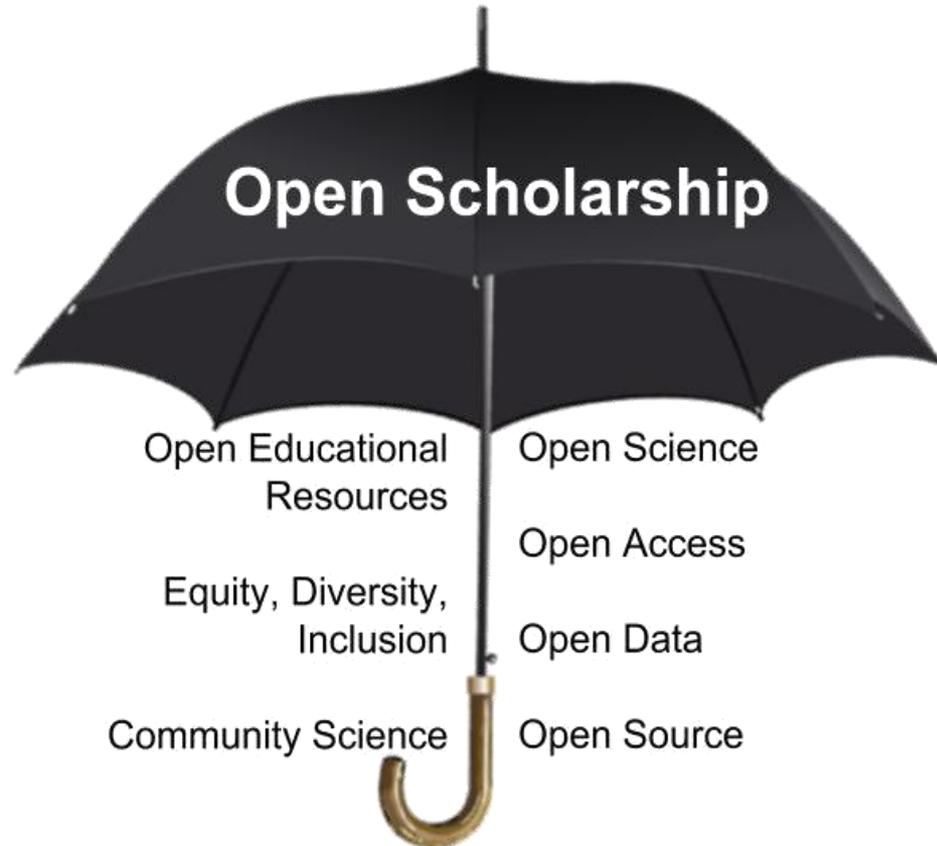
Gurovich et al, Nature Medicine, 2019  
<https://www.face2gene.com>

# Ethical and responsible publication goes beyond reproducible analyses



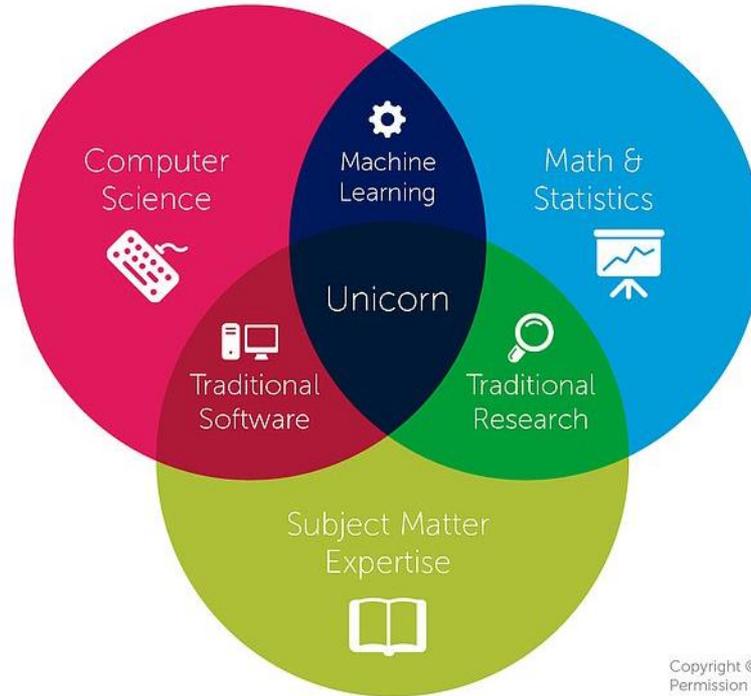
---

# The many dimensions of open scholarship



---

# 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?*

---

# Visit the Turing Health Programme Bazaar

- 15 tables
- 4 x 20 minute discussions



---

# Visit the Turing Health Programme Bazaar

- 15 tables
- 4 x 20 minute discussions



---

# Visit the Turing Health Programme Bazaar

- 15 tables
- 4 x 20 minute discussions
- Capture the conversation



---

# Visit the Turing Health Programme Bazaar

## Data sources

- Electronic patient records
- Internet of things
- Imaging

## Modelling

- Explainable algorithms
- Deep learning
- Causal inference
- Predictive modelling

## Working with data

- Secure data analysis
- Data linkage and integration
- Data quality, access and reuse

## Application areas

- Mental health
- Precision medicine
- Genomics/bioinformatics
- Epidemiology
- AI for diagnosis

---

# Visit the Turing Health Programme Bazaar

- 15 tables
- 4 x 20 minute discussions
- Capture the conversation
- Listen more than you speak
  - Step up, step down
- Come with an open mind
- Be constructive & creative



# Visit the Turing Health Programme Bazaar

- 15 tables
- 4 x 20 minute discussions
- Capture the conversation
- Listen more than you speak
  - Step up, step down
- Come with an open mind
- Be constructive & creative
- Pitch the challenge
  - Where can the Turing Health programme add value?



---

# Visit the Turing Health Programme Bazaar

## **Data sources**

- Electronic patient records
- Internet of things
- Imaging

## **Modelling**

- Explainable algorithms
- Deep learning
- Causal inference
- Predictive modelling

## **Working with data**

- Secure data analysis
- Data linkage and integration
- Data quality, access and reuse

## **Application areas**

- Mental health
  - Precision medicine
  - Genomics/bioinformatics
  - Epidemiology
  - AI for diagnosis
-

The  
Alan Turing  
Institute



---

# Thank you

[kwhitaker@turing.ac.uk](mailto:kwhitaker@turing.ac.uk)  
[@kirstie\\_j](https://twitter.com/kirstie_j)

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

[gitter.im/alan-turing-institute/the-turing-way](https://gitter.im/alan-turing-institute/the-turing-way)

doi: [10.6084/m9.figshare.7819442](https://doi.org/10.6084/m9.figshare.7819442)



**#IWD2019**