The Alan Turing Institute

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)	

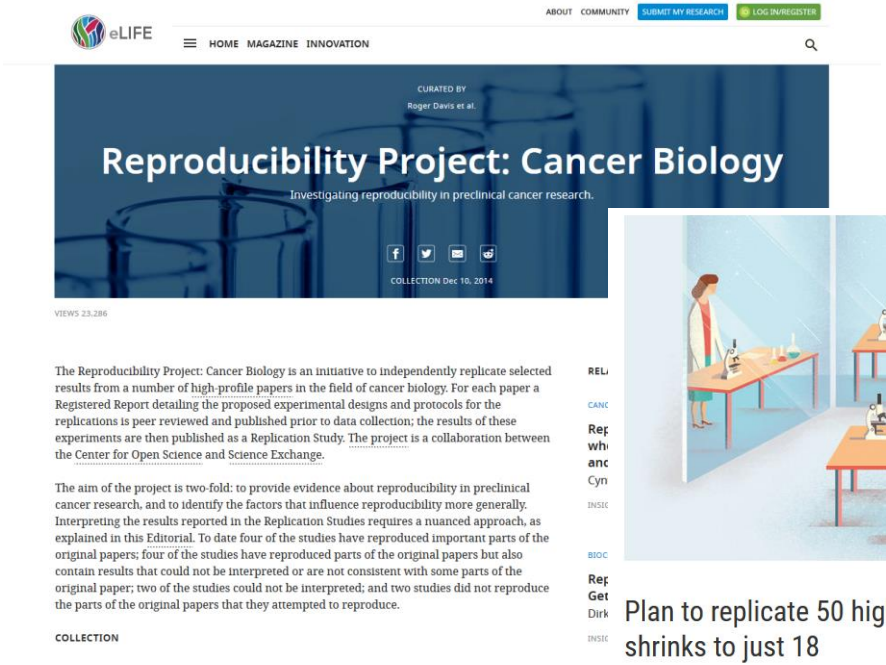
https://statmodeling.stat.columbia.edu/2013/04/16/memo-to-reinhart-and-rogoff-i-think-its-best-to-admit-your-errors-and-go-on-from-there
https://www.bbc.co.uk/news/magazine-22223190

Errors in the literature have real world effects

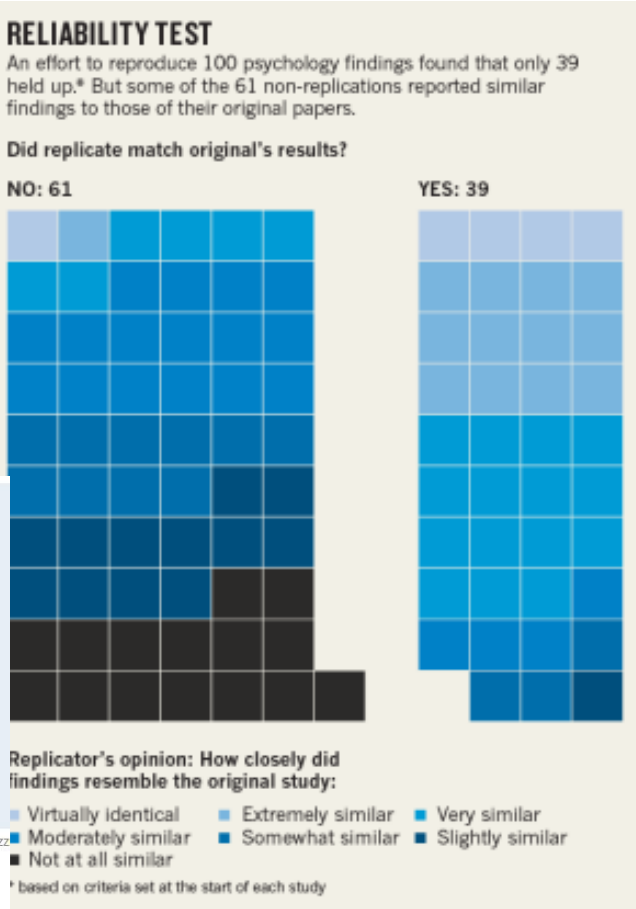
	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.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

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



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



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

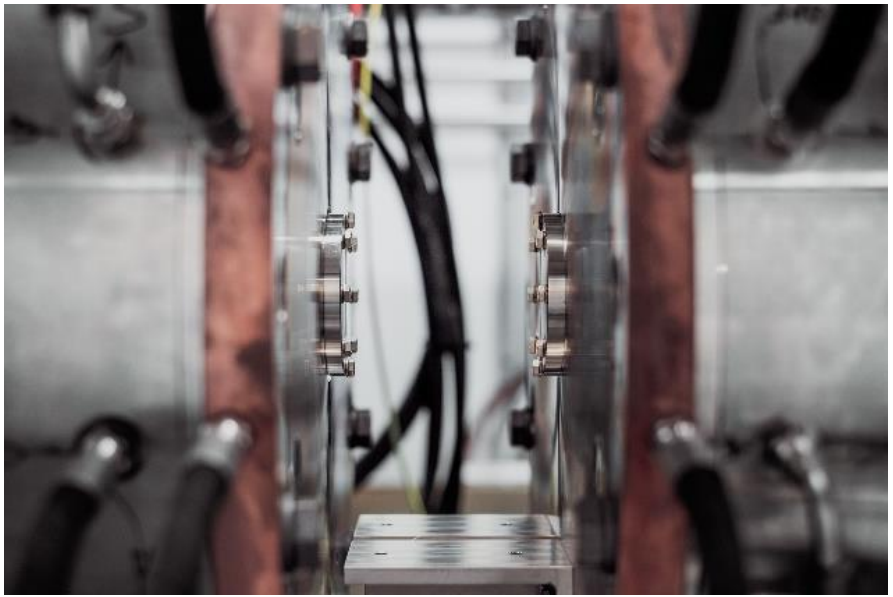
Fraud is not our biggest problem

https://doi.org/10.6084/m9.figshare.5558653

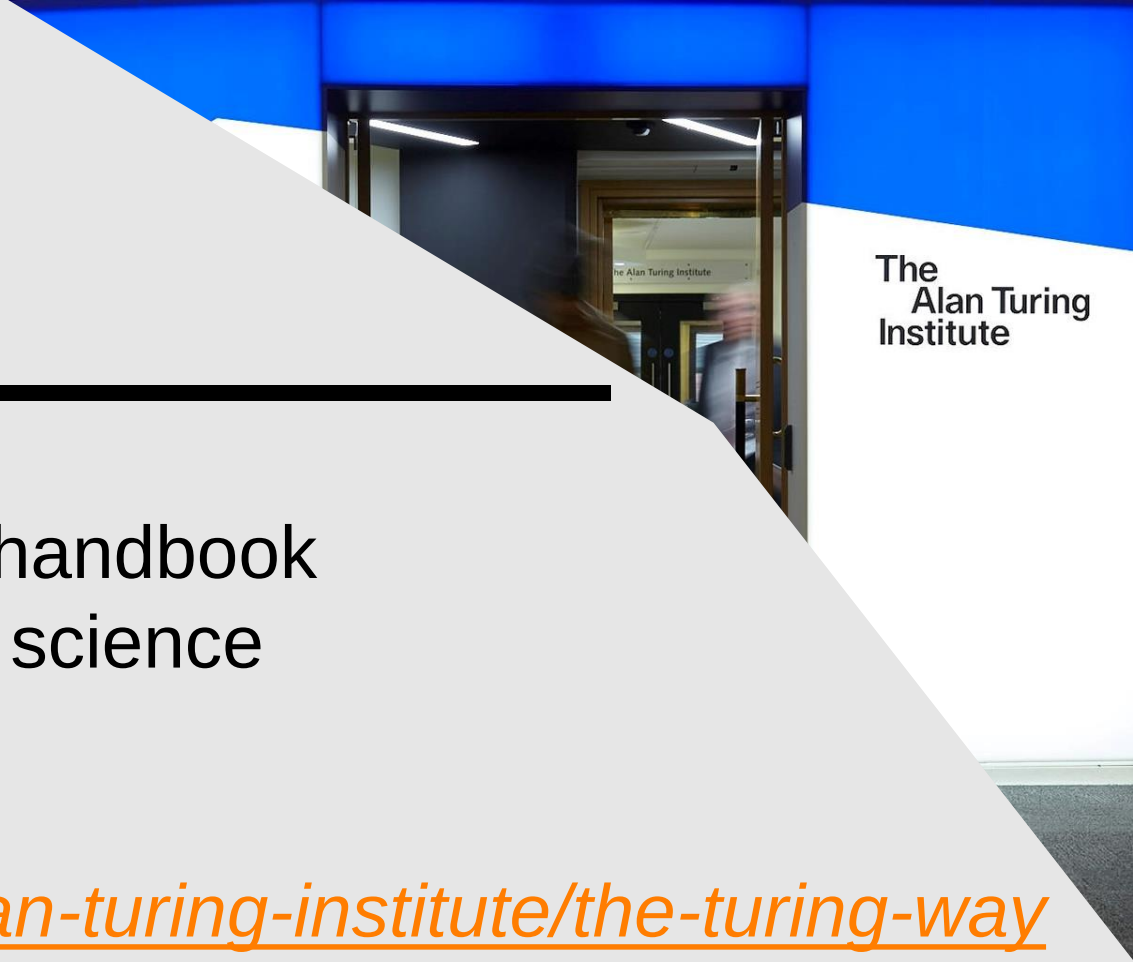
https://www.turing.ac.uk/news/alan-turing-institute-spearhead-new-cutting-edge-data-science-and-artificial-intelligence

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.



https://www.turing.ac.uk/research/interest-groups/tools-practices-and-systems



The Turing Way

A lightly opinionated handbook
for reproducible data science

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

What does reproducible mean?

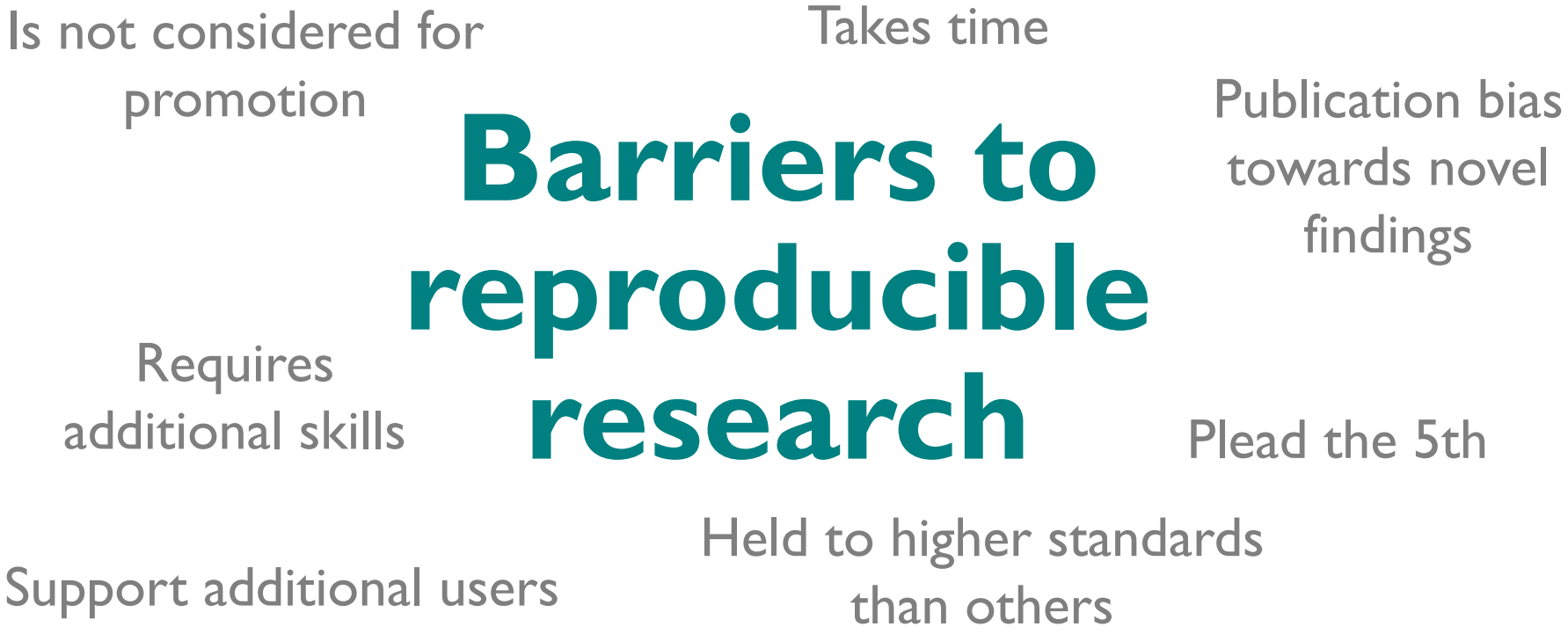
		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://doi.org/10.6084/m9.figshare.7140050
github.com/alan-turing-institute/the-turing-way/blob/master/chapters/open_research.md

Built by a team….and you!

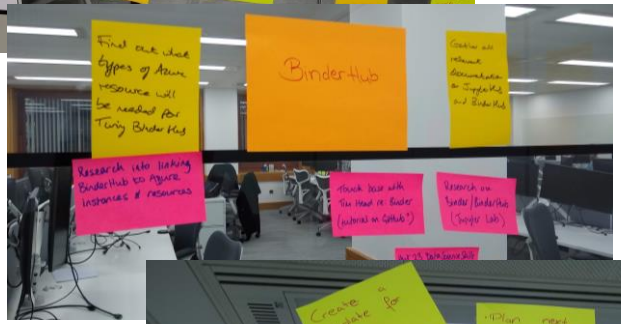
- Rachael Ainsworth
- Becky Arnold
- Louise Bowler
- Sarah Gibson
- Patricia Herterich
- Rosie Higman
- Anna Krystalli
- Alex Morley
- Martin O’Reilly
- . . .

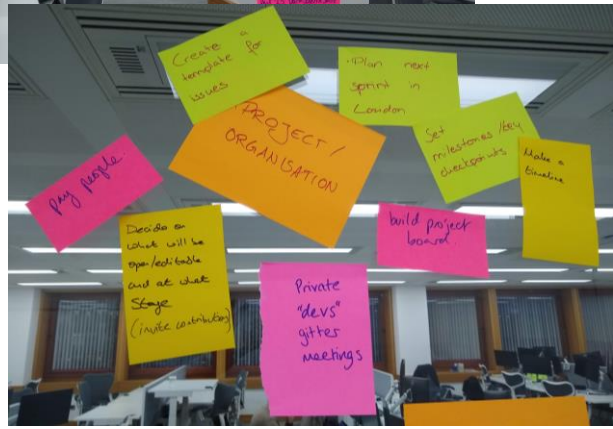
Why don’t people do this already?



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

Requires additional skills





Chapters will include:

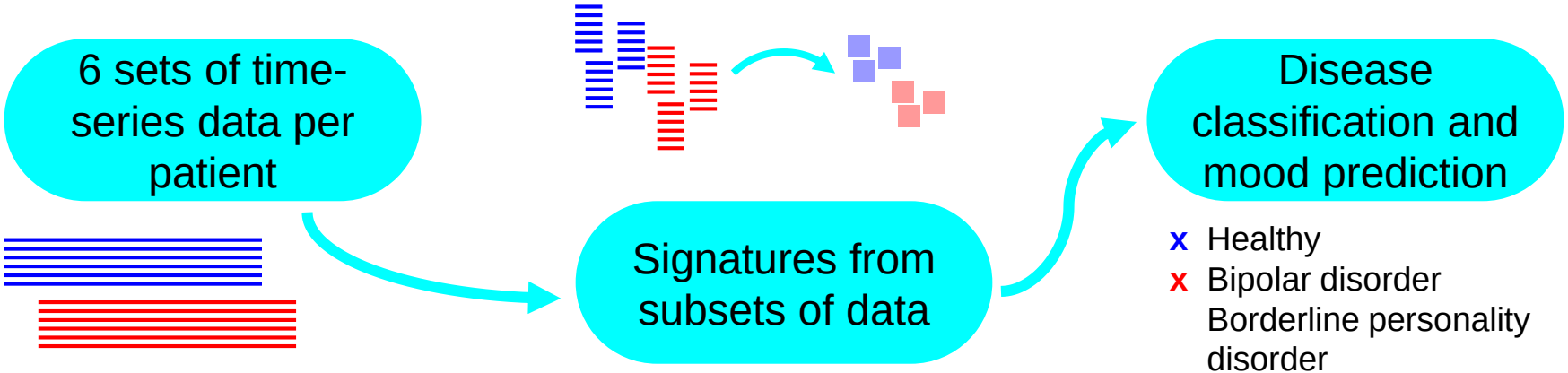
- Research data management
- Open science
- Reproducibility
- Version control with git
- Your working environment (IDE, notebooks etc)
- Capturing your compute environment
- Testing for research
- Continuous integration
- Collaborating through GitHub/GitLab

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

Reproducible research champions

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

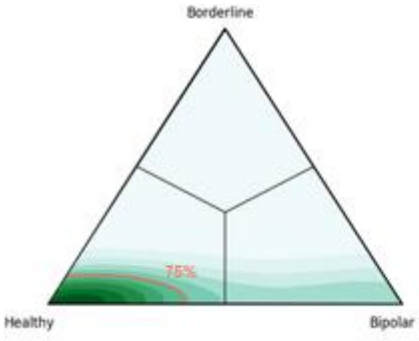
Imanol Perez Arribas, Guy Goodwin, John Geddes, Terry Lyons & Kate Saunders

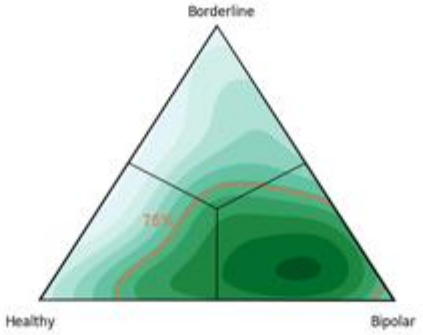


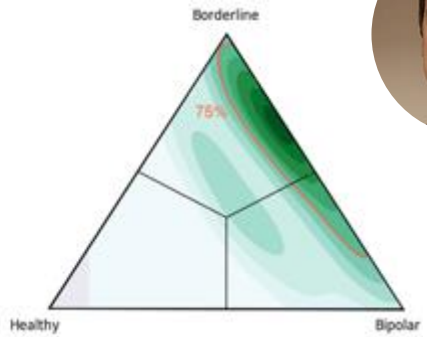
Perez Arribas et al, Translational Psychiatry, 2018

Reproducible research champions

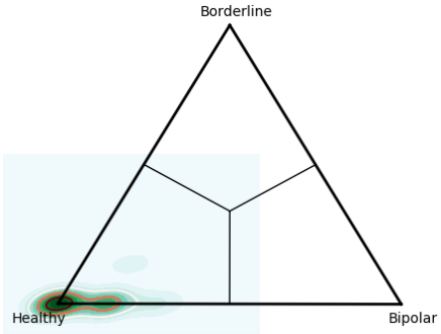
Publication

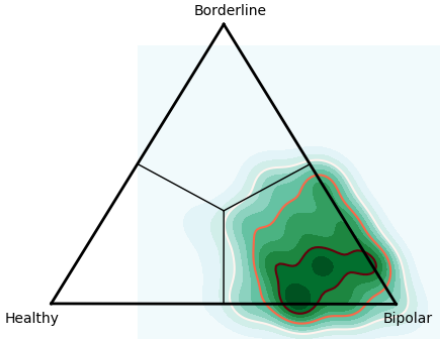


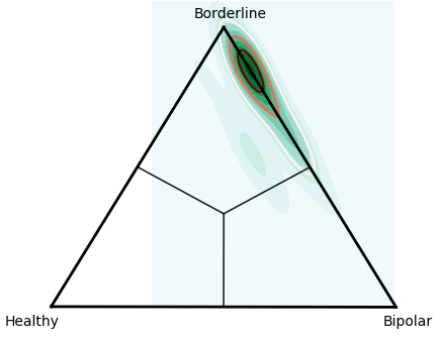




Synthetic Data







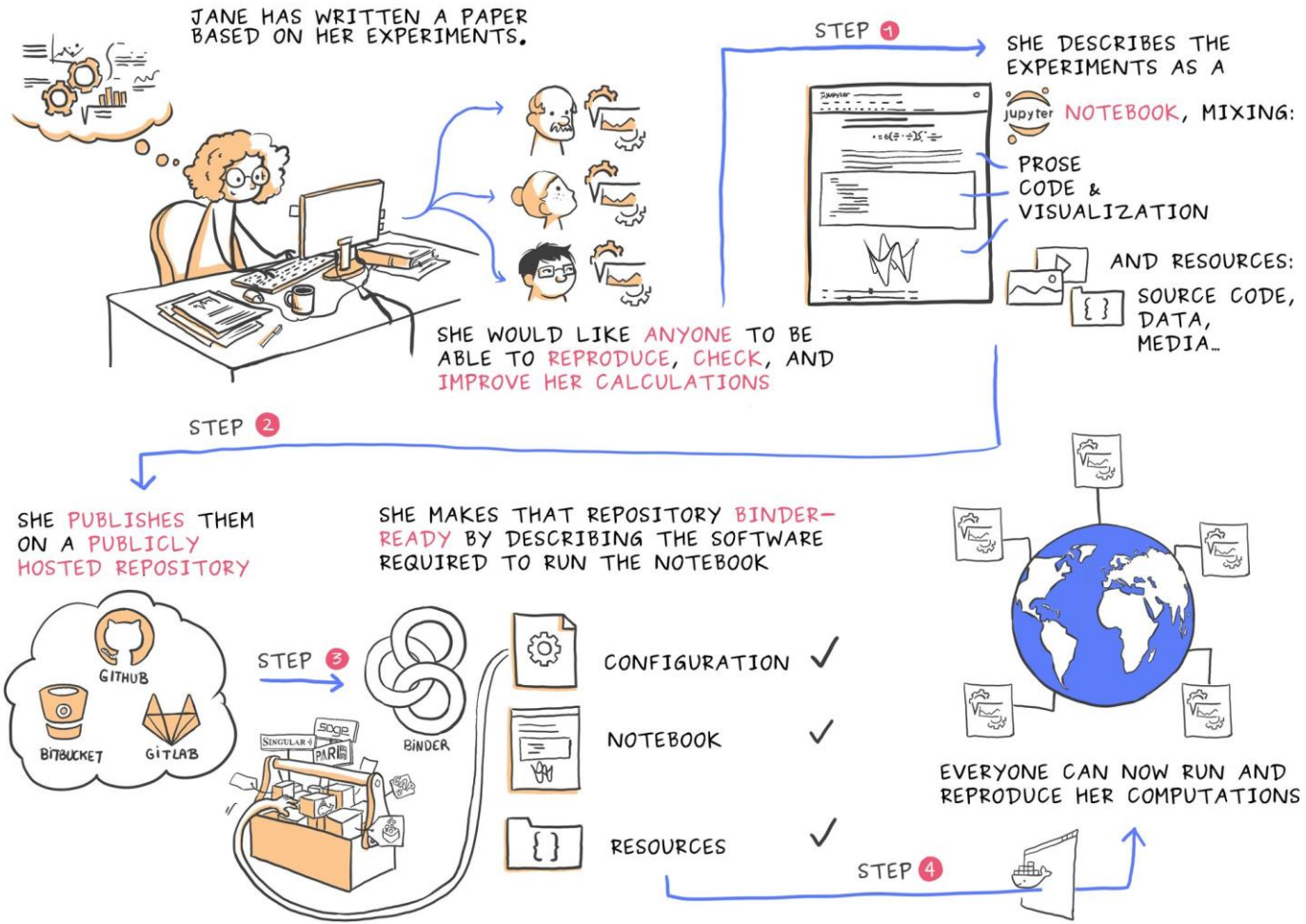
https://github.com/alan-turing-institute/signatures-psychiatry

Research engineering at the Turing

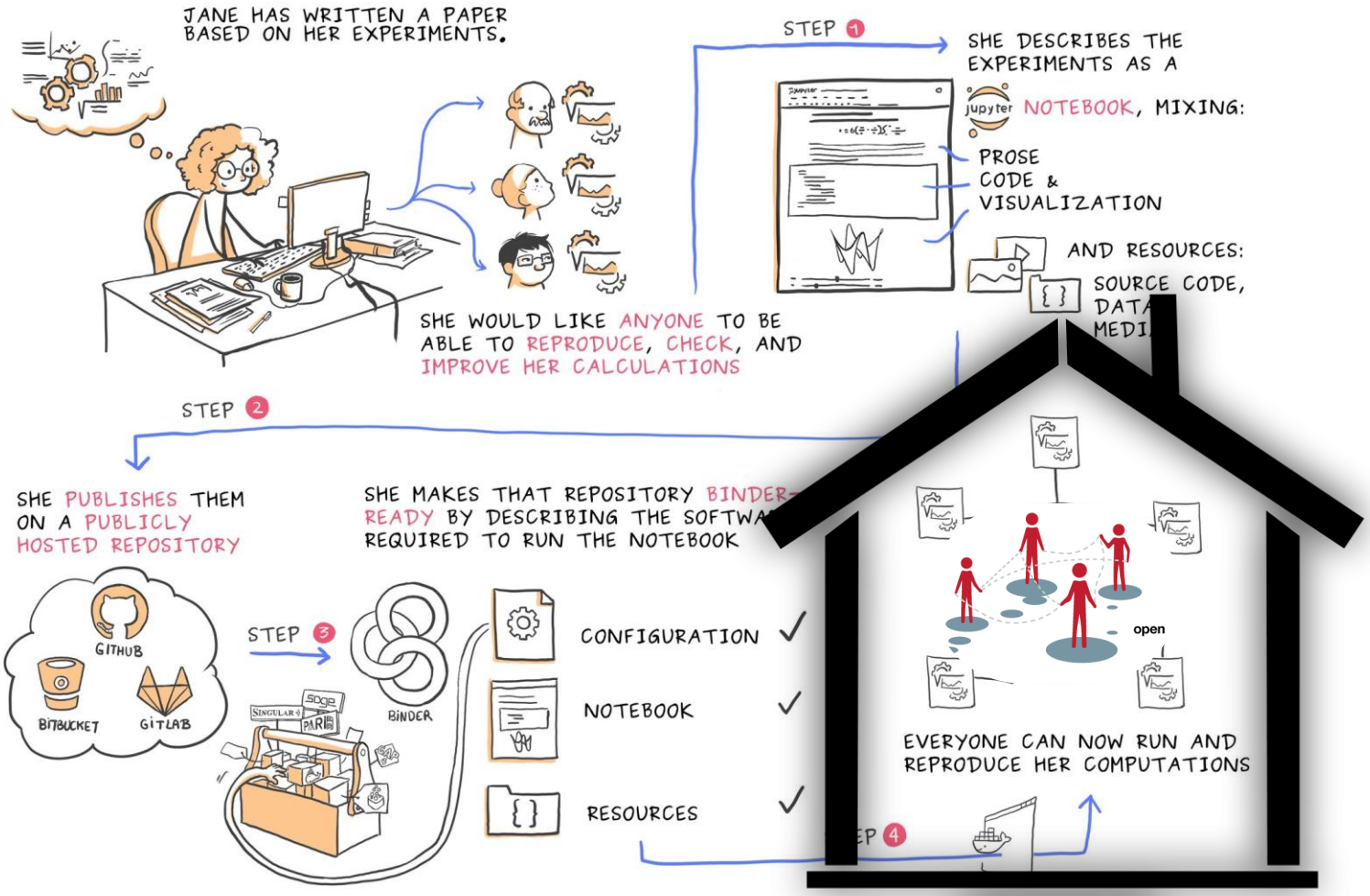
https://www.turing.ac.uk/research/research-programmes/research-engineering

Research engineering at the Turing

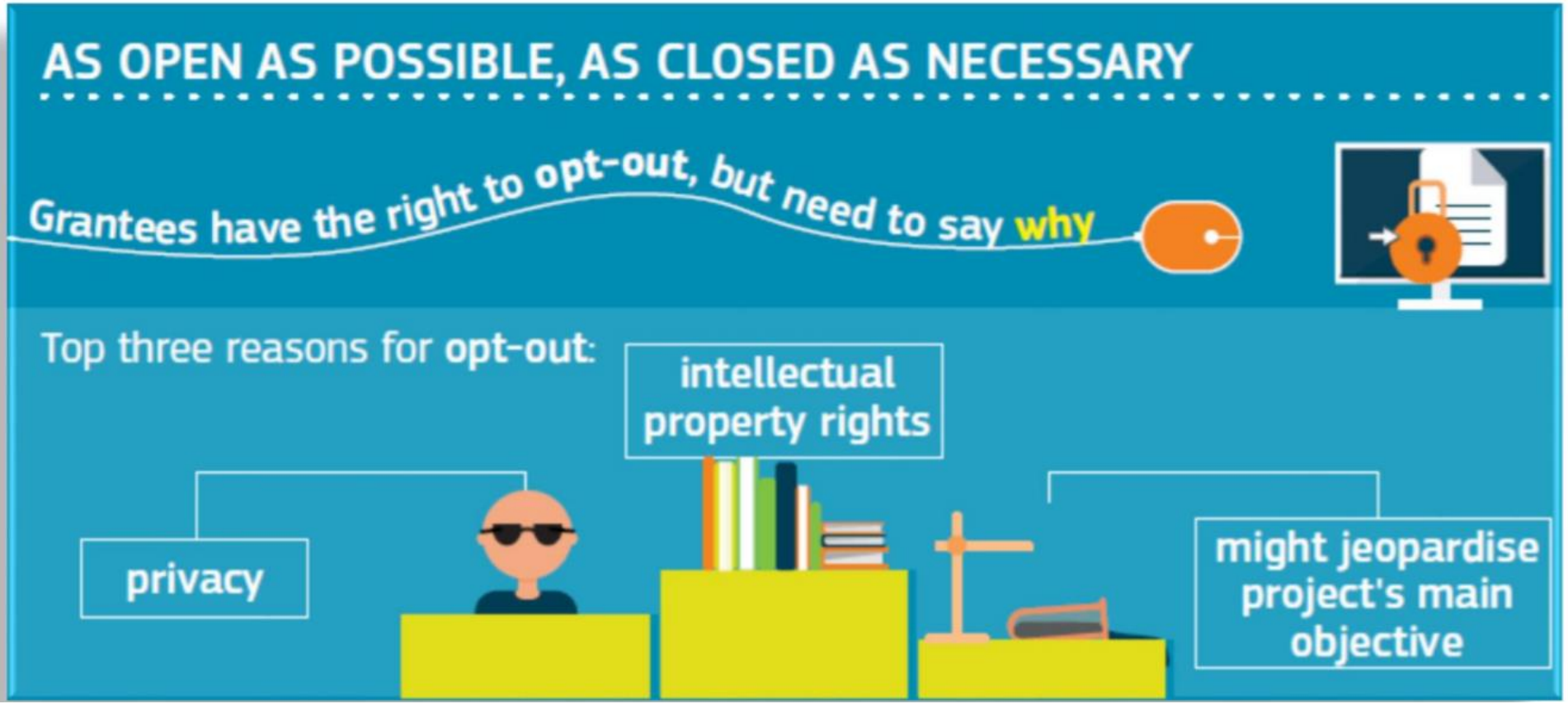
https://www.turing.ac.uk/research/research-programmes/research-engineering



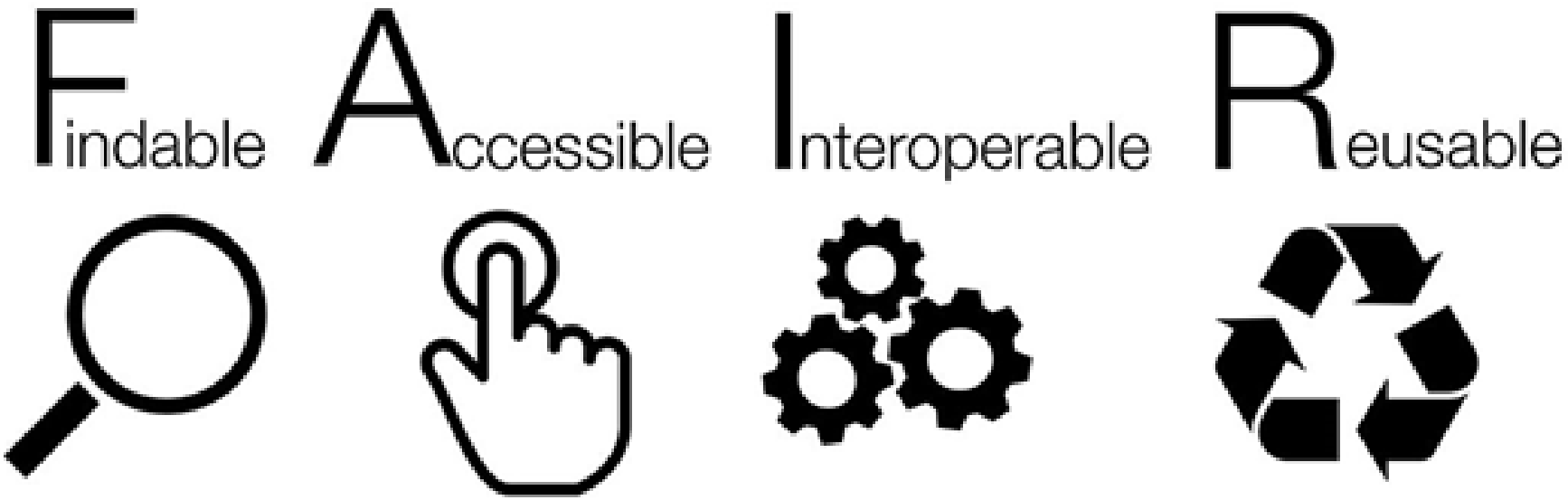
Courtesy of Juliette Belin: https://twitter.com/JulietteTaka/status/1082735653929000960



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

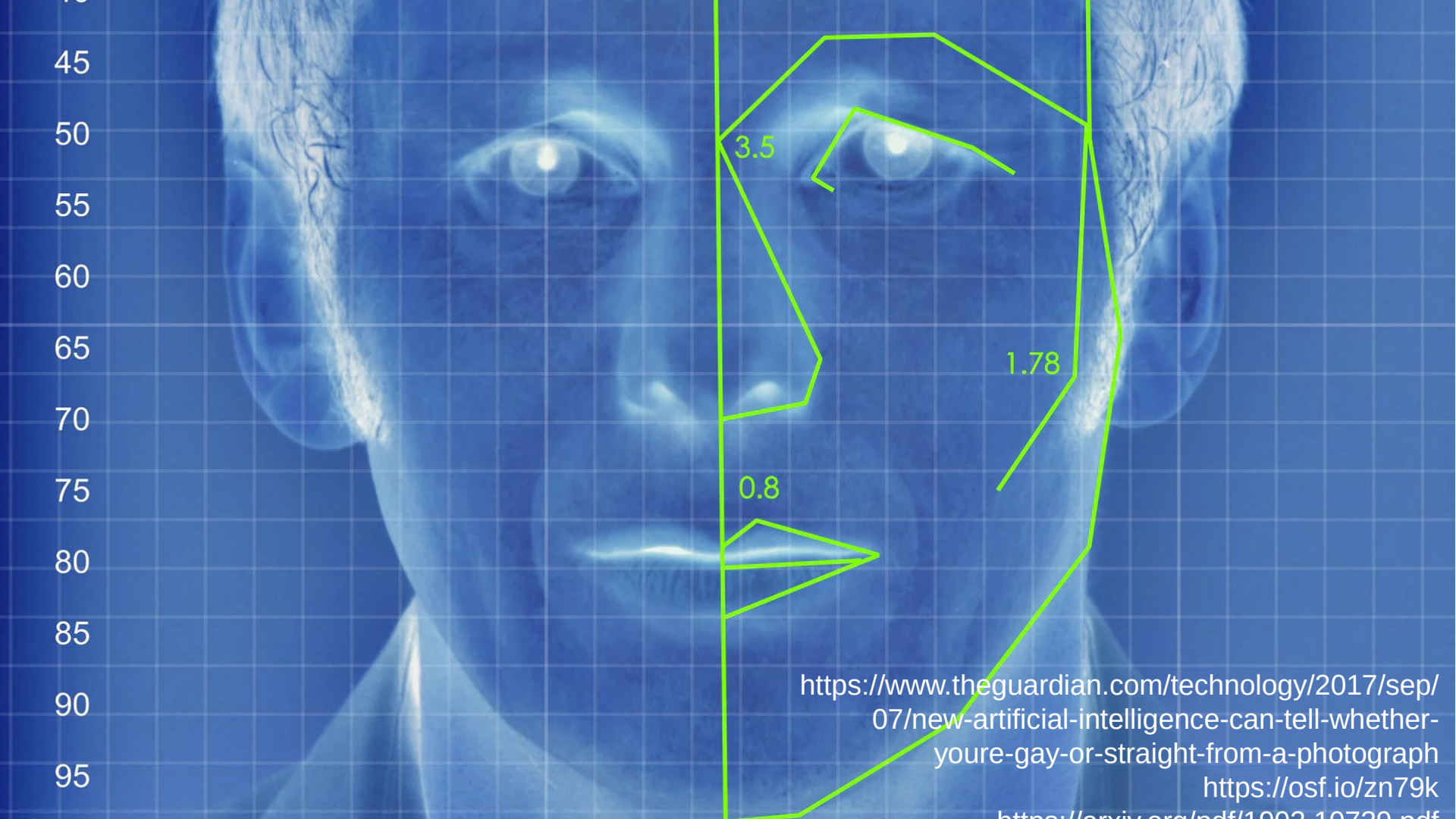


https://ec.europa.eu/research/press/2016/pdf/opendata-infographic_072016.pdf



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

Image credit: Sangya Pundir

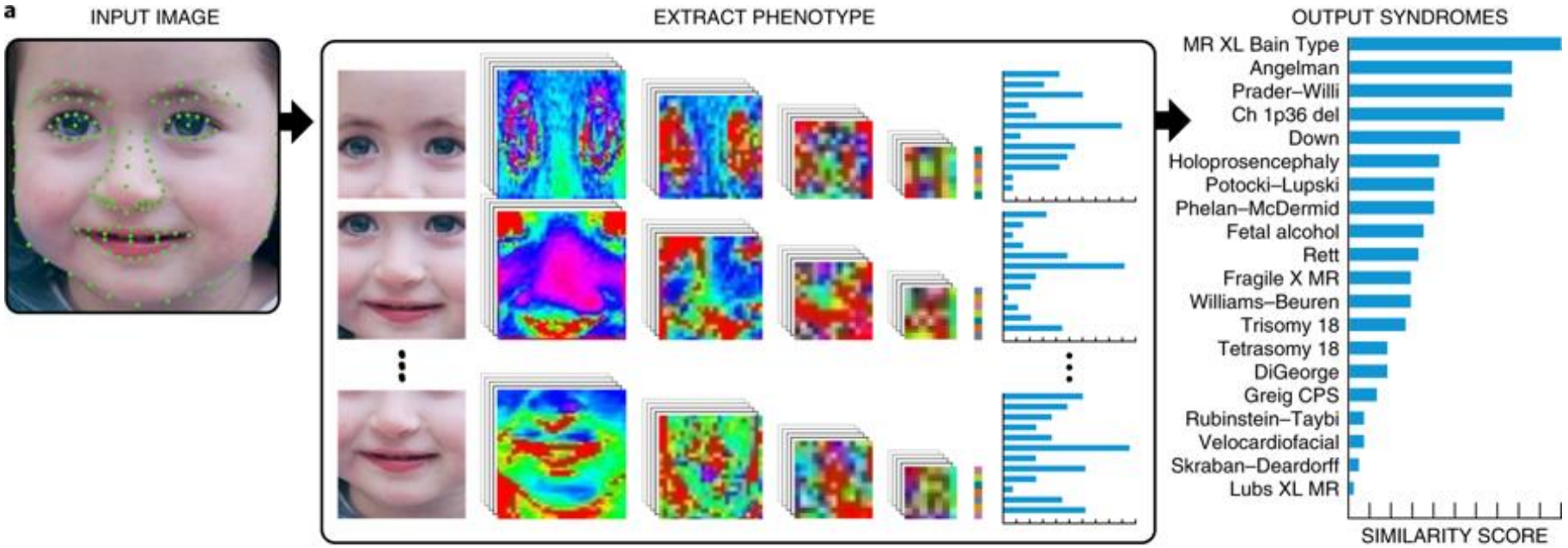


https://www.theguardian.com/technology/2017/sep/07/new-artificial-intelligence-can-tell-whether-youre-gay-or-straight-from-a-photograph

https://osf.io/zn79k

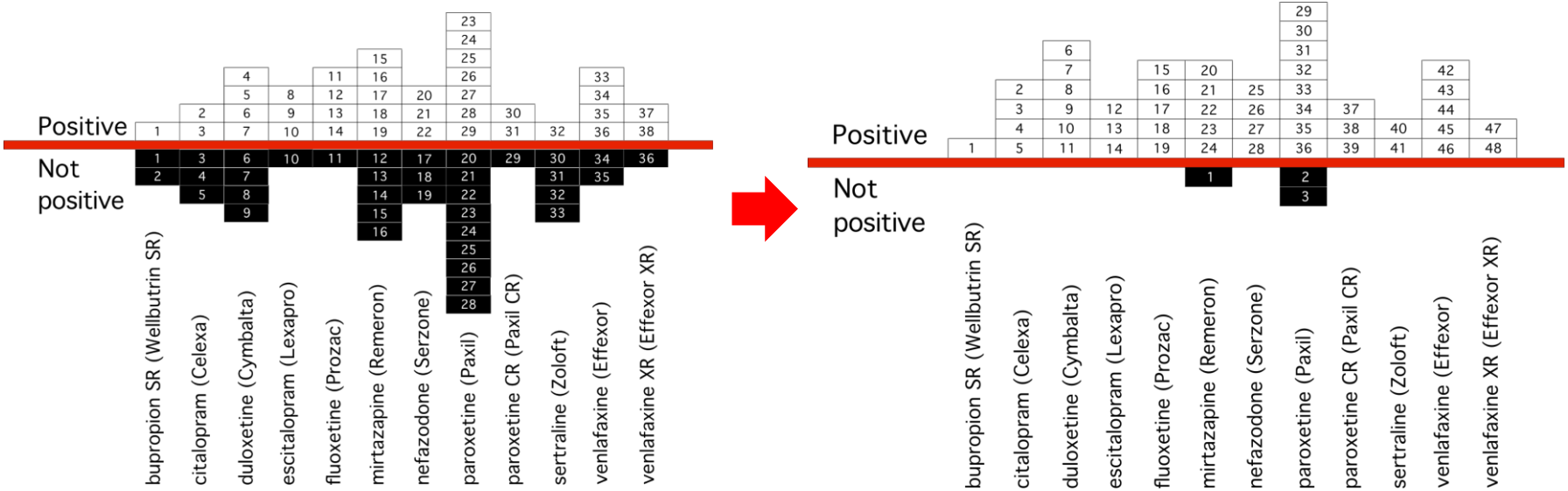
https://arxiv.org/pdf/1902.10739.pdf

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



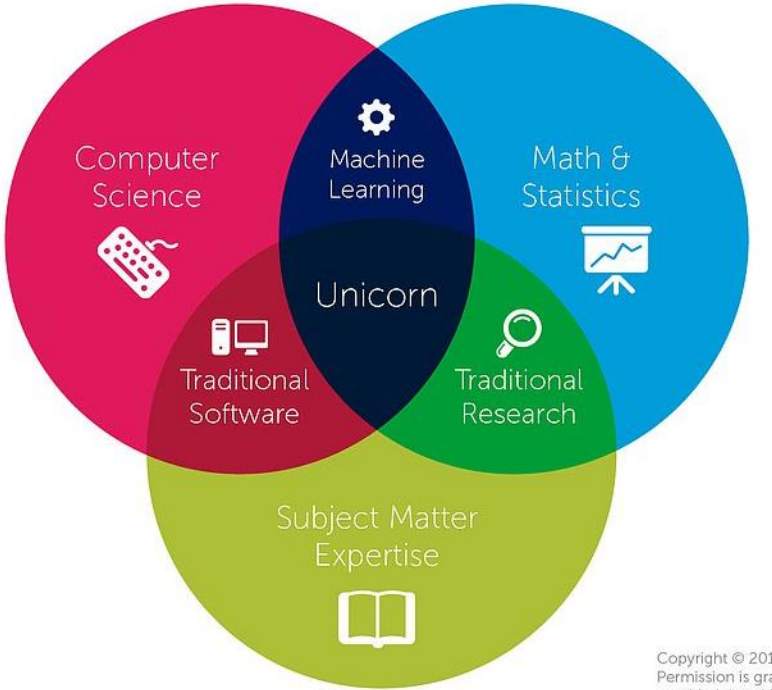
https://doi.org/10.6084/m9.figshare.3381379

The many dimensions of open scholarship

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

Robin Champieux and Danielle Robinson

The Data Science Unicorn



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

moz://a

UNIVERSITY OF CAMBRIDGE

Thank you

kwhitaker@turing.ac.uk
@kirstie_j
github.com/alan-turing-institute/the-turing-way
gitter.im/alan-turing-institute/the-turing-way
doi: 10.6084/m9.figshare.7819442