Open and reproducible science

WHY AND HOW

Sara Villa

Postdoctoral researcher @Denk's group The Turing Way Contributor OLS Open Seeds programme mentor

ALL IMAGES TAKEN FROM The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. http://doi.org/10.5281/zenodo.3332807 <u>sara.villa@kcl.ac.uk</u>

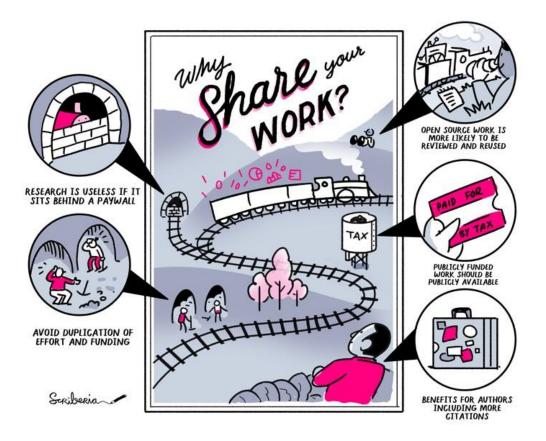


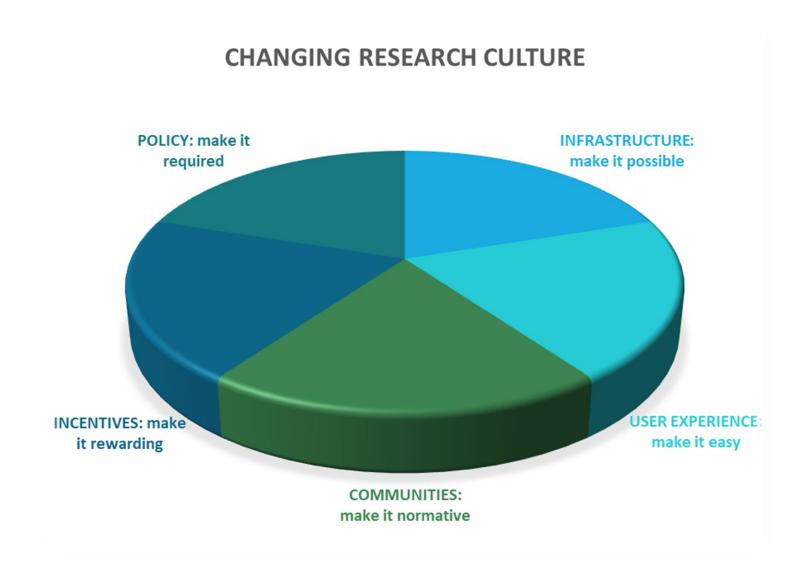


Scriberia

WHY Open science

School	Belief	Aim
Infrastructure	Efficient research depends on the available tools and applications.	Creating openly available platforms, tools, and services for researchers.
Pragmatic	Knowledge-creation could be more efficient if researchers worked together.	Opening up the process of knowledge creation.
Measurement	Academic contributions today need alternative impact measurements.	Developing an alternative metric system for research impact.
Democratic	The access to knowledge is unequally distributed.	Making knowledge freely available for everyone.
Public	Research needs to be made accessible to the public.	Making research accessible for citizens.

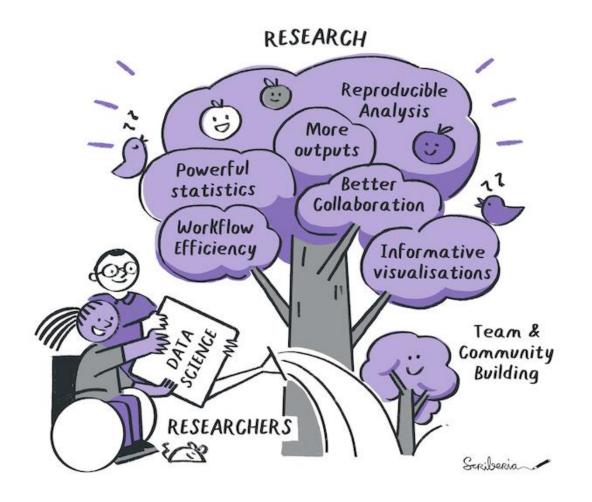




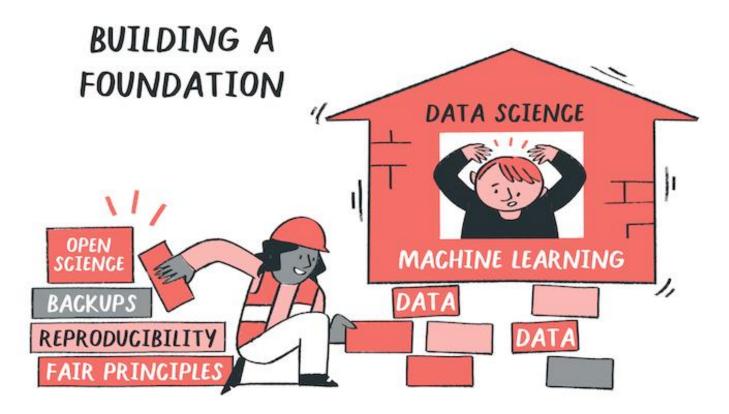
Reinterpreted from https://www.cos.io/blog/cos-celebrates-10-years

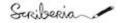
Every research project is a data science project





OS principles can be applied to ANY research field





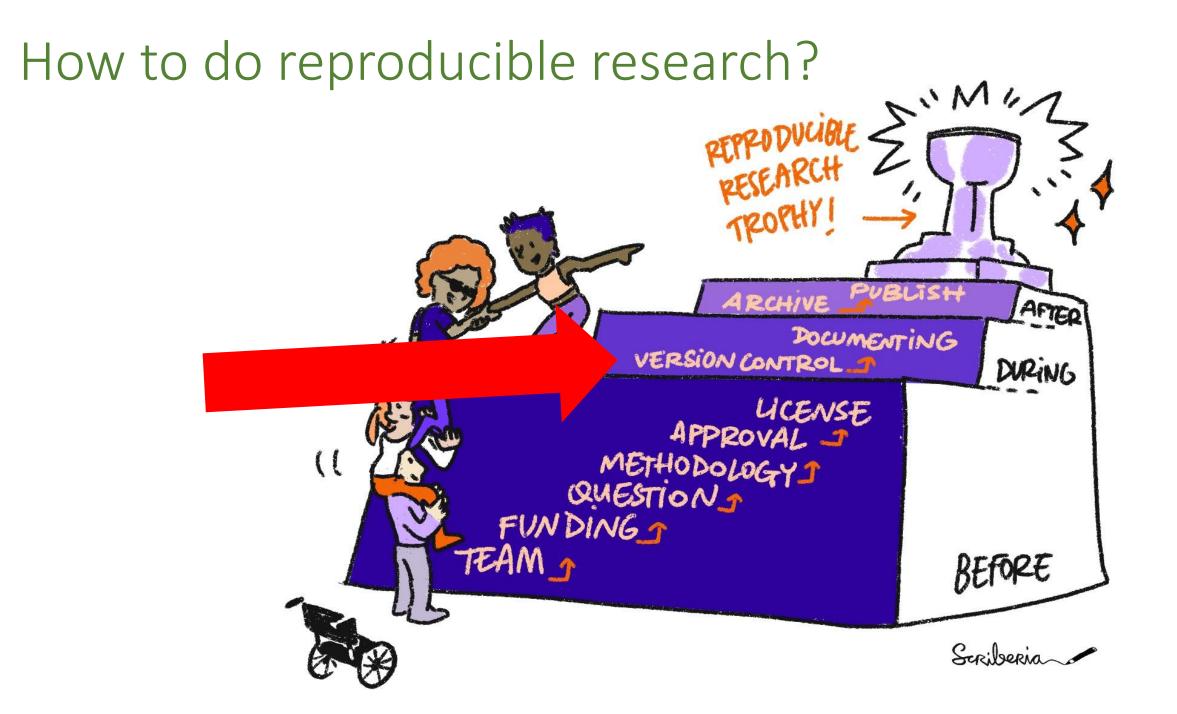
https://www.gofair.org/fairprinciples/i1metadata-use-formalaccessible-sharedbroadly-applicablelanguage-knowledgerepresentation/

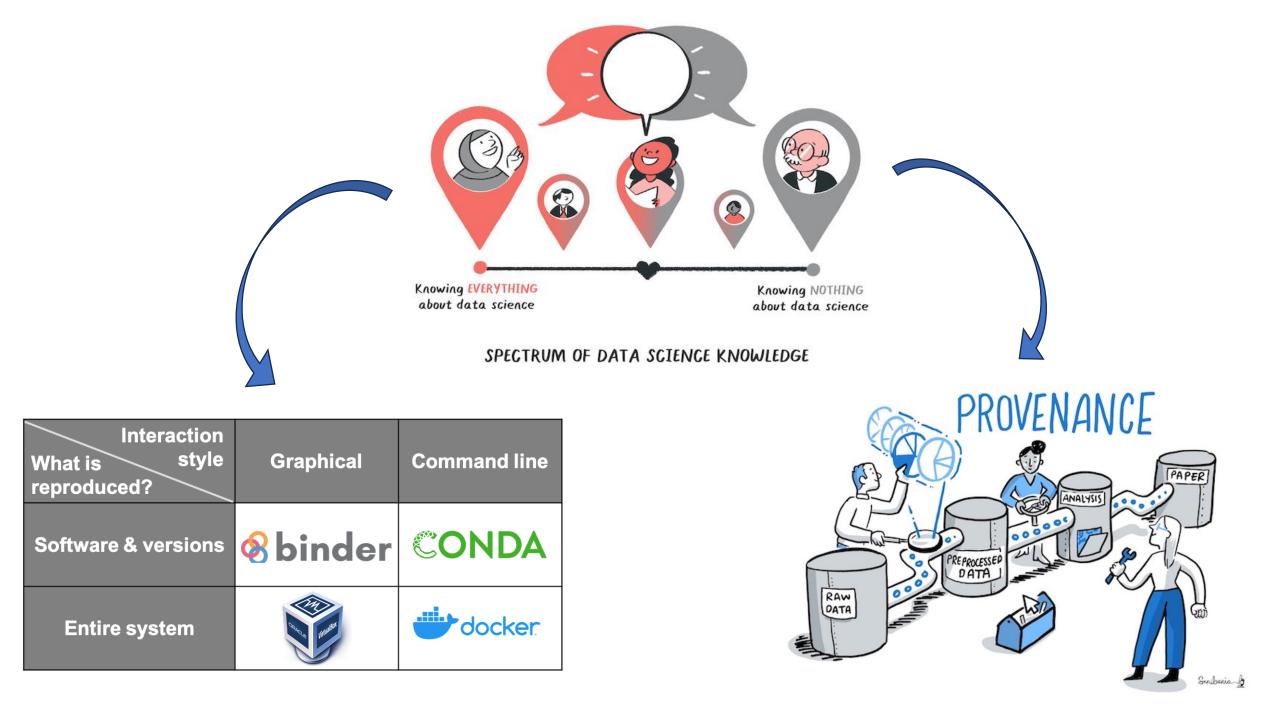


https://osf.io/

Is reproducibility becoming a buzz word?

NIH National Lil	brary of Medicine Biotechnology Information
Pub	reproducibiliy X Search Advanced Create alert Create RSS User Guide
RESULTS BY YEAR	Save Email Send to Sorted by: Best match Display options 122,317 results
	2018 2023
My NCBI FILTERS 🖪	Filters applied: in the last 5 years. Clear all
TEXT AVAILABILITY	Showing results for <i>reproducibility</i> Your search for <i>reproducibility</i> retrieved no results
 Free full text Full text 	 Reproducibility Starts from You Today. Turkyilmaz-van der Velden Y, Dintzner N, Teperek M. Patterns (N Y). 2020 Sep 11:1(6):100099. doi: 10.1016/j.patter.2020.100099. eCollection 2020 Sep 11.
ARTICLE ATTRIBUTE	PMID: 33205134 Free PMC article. Review. Share Who hasn't yet heard about the debates on research reproducibility, or, perhaps even more, about the
Associated data	research reproducibility crisis?Reproducibility starts from you. Today! Reproducibility in Neuroimaging Analysis: Challenges and Solutions.







Advanced

Search

Q

Understand the experiment and

Nucleic Acids Research

JOURNAL ARTICLE

stem cell data

Issues	Soction	browse v
issues	Section	browse v

More Content V Submit v Purchase About **v**

```
Nucleic Acids Research 🛛 🔻
```

Nucleic Acids Research Volume 47, Issue D1 08 January 2019

Article Contents

Abstract

INTRODUCTION

REPORTS

DATA MINING

INTEGRATIVE DATA ANALYSIS TOOLS

COMPUTATIONAL RESOURCES

SUPPLEMENTARY DATA

ACKNOWLEDGEMENTS

A PDF

Abstract

INTERACTIVE ANALYSIS

CONCLUSION

DATA AVAILABILITY

Stemformatics: visualize and download curated gather available raw data - expression: fastq, Affy CEL files, ... - samples: GEO, publications, ... Jarny Choi, Chris M Pacheco, Rowland Mosbergen, Othmar Korn, Tyrone Chen, Isha Nagpal, Steve Englart, Paul W Angel, Christine A Wells 🖾 Nucleic Acids Research, Volume 47, Issue D1, 08 January 2019, Pages D841–D846, Process raw data -> summarised at https://doi.org/10.1093/nar/gky1064 the gene/probe level Published: 08 November 2018 Article history v - align reads, background correct, ... - counts matrix, expression matrix, ... Split View 66 Cite Permissions < Share v Quality control (QC) evaluation Stemformatics is an established gene expression data portal containing over - read depth, library quality, ... 420 public gene expression datasets derived from microarray, RNA sequencing - PCA, HC, expression density, ... and single cell profiling technologies. Developed for the stem cell community, it has a major focus on pluripotency, tissue stem cells, and staged differentiation. Stemformatics includes curated 'collections' of data relevant to cell Upload dataset into Stemformatics reprogramming, as well as hematopoiesis and leukaemia. Rather than simply - add sample data, (ontology terms) rehosting datasets as they appear in public repositories, Stemformatics uses a - add reports, test stringent set of quality control metrics and its own pipelines to process handpicked datasets from raw files. This means that about 30% of datasets BloodSpot: a database of healthy and processed by Stemformatics fail the quality control metrics and never make it malignant haematopoiesis updated with purified and single cell mRNA sequencing to the portal, ensuring that Stemformatics data are of high quality and have profiles been processed in a consistent manner. Stemformatics provides easy-to-use and intuitive tools for biologists to visually explore the data, including SC2disease: a manually curated database of

Failure reason (%) no raw data (8.19%)

> poor experiment design (10.88%) bad data quality (8.07%)

annotation problems (2.56%)

https://lakens.github.io /statistical inferences/

Insufficient sample size, power calculations... are basics often driving to a poor quality study

REGISTERED REPORTS

PEER REVIEW BEFORE RESULTS ARE KNOWN TO ALIGN SCIENTIFIC VALUES AND PRACTICES

ELIMINATE BIAS & INCREASE RIGOR



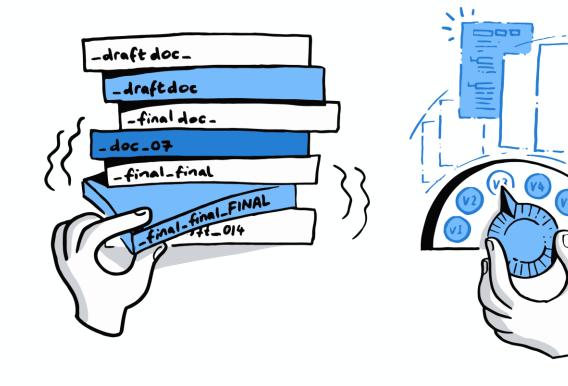
Feedback from project design and publish negative results tool

https://www.bnacredibility.org.uk/academia https://www.cos.io/initiatives/registered-reports

Version control



TRACK PROJECT HISTORY



Scriberia_

Learning new tools is worthy for your research most of the times...

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE? (ACROSS FIVE YEARS)

	HOW OFTEN YOU DO THE TASK						
		50/ _{DAY}	5/day	DAILY	WEEKLY	MONTHLY	YEARLY
ſ	1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
	5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
	30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
HOW MUCH		8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
TIME. YOU		9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
SHAVE OFF			6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
	1 HOUR		10 months	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
	6 HOURS				2 MONTHS	2 WEEKS	1 DAY
	1 DAY					8 WEEKS	5 DAYS

https://xkcd.com/1205/

Read literature Write papers Bring in funding Collect / access data Design What does a experiments Set strategy quantitative researcher do? Run statistical analyses Mentor colleagues Promote work to decision makers @TuringWay @kirstie j Manage a team Interpret results https://doi.org/10.5281/zenodo.7749650

Peer review

Read literature Write papers Bring in funding Collect / access data Design What does a experiments Set strategy quantitative Testing & researcher do? quality control Mentor Manage budgets Run statistical analyses Promote work to colleagues decision makers Report on @TuringWay @kirstie j Manage a team Interpret results progress https://doi.org/10.5281/zenodo.7749650

	Μ	anage (cloud) compute			Peer review
Read	Deliver	resources			
literature	trainings	Write papers		Bri	ng in funding
Design		What	does	a Collect	z / access data
experime	nts				$\sigma \vee $
Testing	, &	quant	titativ	e Set Strate	5У
quality co	ontrol	researc	cher d	0?	
Mento	r	Manage k	oudgets	Run statist	cical analyses
colleagu	es	Promote woi	rk to		Scope user
General	ise code to	decision ma	kers V	isualisation	requirements
sof	tware	Report on	Manage		• @TuringWay @kirstie_j
Interpret	results	progress	0		/10.5281/zenodo.7749650

	Μ	anage (cloud)				Peer review
Read	Deliver	105001005	Archive d			
literature	trainings	Write paper	s code	3	Brir	ng in funding
Design experime		What	t doe	sa	Collect	z / access data
Testing	<u></u> 8	•	ntitati	IVE	et strateg	37
quality co	ontrol	resear	cher	do?		
Mento	r	Manage	budget	s Ri	un statist	cical analyses
colleagu		Promote wo	ork to			Scope user
Generali	ise code to		akers	Visuali	sation	requirements
	tware	Report on	Mana	ge a tea	m	<pre>@TuringWay @kirstie_j</pre>
Interpret	results	progress		h	ttps://doi.org/	/10.5281/zenodo.7749650

		anage (cloud) compute	pr	ercialise a oduct	Peer review
Read	Deliver	resources A	Archive da	ita &	
literature	trainings	Write paper	s code	B	Bring in funding
Design		Wha [·]	t does	Colle	ect / access data
experime	nts				tegv
Testing	<u></u> 8	quar	ntitativ	/e	CC01
quality co	ontrol	resear	cher o	do?	
Mento	rTrack imp	bact Manage	budgets	Run stat	istical analyses
colleagu	es	Promote wo	ork to		Scope user
General	ise code to	decision m	akers	Visualisation	requirements
sof	tware	Report on		e a team	ہ @TuringWay @kirstie_j
Interpret	results	progress	0		org/10.5281/zenodo.7749650

Attend / chair Manage	e (cloud) Comr	nercialise a	Peer review
meetings com	pute P	product Pror	note EDI and
Read Deliver resou			k-life balance
	ite papers code	e Bri Bri	ng in funding
Design experiments Share work with the	What doe	Colloc	t / access data
experiments public		Set strate	gy Maintain ethical &
Testing & public	quantitati	ve serstrate	ethical &
quality control	researcher	do?	egal principles
MentorTrack impact	Manage budgets	s Run statist	tical analyses
colleagues Pro		Host events	Scope user
Generalise code to de		Visualisation	requirements
JUILVUIL	ort on Manag	ge a team	@TuringWay @kirstie_j
Interpret results prog	gress	https://doi.org	/10.5281/zenodo.7749650





But they don't exist

We can't expect individuals to be able to do all these tasks to the highest standard

@TuringWay @kirstie_j

https://doi.org/10.5281/zenodo.7749650

@TuringWay@kirstie

https://doi.org/10.5281/zenodo.7749650

There is one way... Learning new tools might be worthy... or it might not sometimes

HOW LONG CAN YOU WORK ON MAKING A ROUTINE TASK MORE EFFICIENT BEFORE YOU'RE SPENDING MORE TIME THAN YOU SAVE? (ACROSS FIVE YEARS)

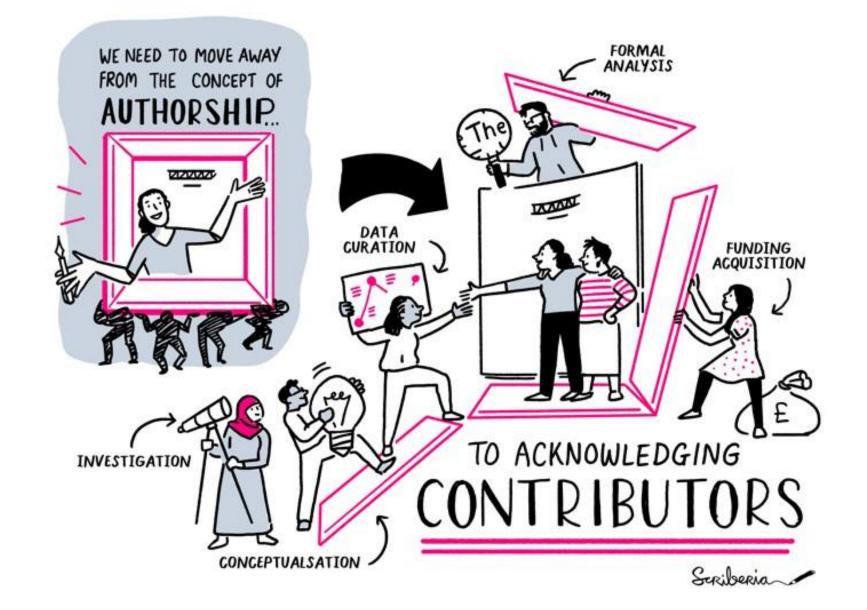
	HOW OFTEN YOU DO THE TASK						
		50/ _{DAY}	5/DAY	DAILY	WEEKLY	MONTHLY	YEARLY
	1 SECOND	1 DAY	2 HOURS	30 MINUTES	4 MINUTES	1 MINUTE	5 SECONDS
	5 SECONDS	5 DAYS	12 HOURS	2 HOURS	21 MINUTES	5 MINUTES	25 SECONDS
	30 SECONDS	4 WEEKS	3 DAYS	12 HOURS	2 HOURS	30 MINUTES	2 MINUTES
400 МОСМ		8 WEEKS	6 DAYS	1 DAY	4 HOURS	1 HOUR	5 MINUTES
TIME YOU		9 MONTHS	4 WEEKS	6 DAYS	21 HOURS	5 HOURS	25 MINUTES
SHAVE OFF	30 MINUTES		6 MONTHS	5 WEEKS	5 DAYS	1 DAY	2 HOURS
	1 HOUR		10 months	2 MONTHS	10 DAYS	2 DAYS	5 HOURS
	6 HOURS				2 MONTHS	2 WEEKS	1 DAY
	1 DAY					8 WEEKS	5 DAYS

https://xkcd.com/1205/

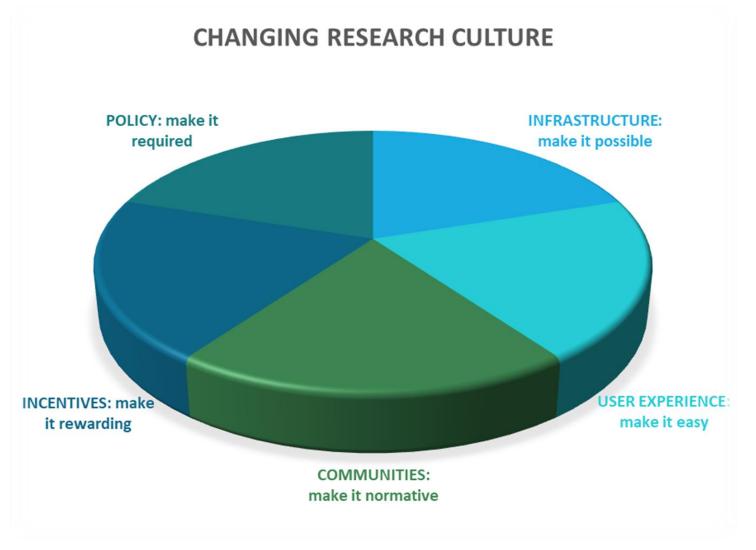
There is another way... The Turing way!



Key aspect for collaboration in neuroscience



It's never THAT easy



Reinterpreted from https://www.cos.io/blog/cos-celebrates-10-years



FOR YOU TO START... plenty of resources/tools out there... HAPPY TO HELP!

https://the-turing-way.netlify.app/index.html https://software.ac.uk/



The Turing Way

Search this book...

:ome

e for Reproducible Research e for Project Design e for Communication e for Collaboration e for Ethical Research munity Handbook word

Welcome

Welcome to The Turing Way handbook to reproducible, ethical and collaborative data science.

The Turing Way project is open source, open collaboration, and community-driven. We involve and support a diverse community of contributors to make data science accessible, comprehensible and effective for everyone. Our goal is to provide all the information that researchers and data scientists in academia, industry and the public sector need to ensure that the projects they work on are easy to reproduce and reuse.

🍷 Top Tip

The Turing Way is not meant to be read from start to finish. Start with a concept, tool or method that you need now, in your current work. Browse the different guides that make up the book, or use the search box to search for whatever you would like to learn about first.

All stakeholders, including researchers, software engineers, project leaders and funding teams, are encouraged to use *The Turing Way* to understand their roles and responsibility of reproducibility in data science. You can inspect our resources on GitHub, contribute to the project as described in our contribution guidelines and re-use all materials (see the License).

Please join our Slack Workspace to connect and discuss your ideas or suggestions with *The Turing Way* members.



https://openlifesci.org/



The OLS program is for people interested in applying open principles in their work and becoming Open Science ambassadors in their communities.

About

C O 🗶

Our

Histo Citino

This is a 16-week long personal mentorship and cohort-based training, where participants (organisers, hosts, mentors and project leads/mentees) of this program will:

- share their expertise and gain knowledge essential to create, lead, and sustain an Open Science project
- connect with members across different projects, communities, backgrounds, and identities
- empower each other to become effective Open Science ambassadors in their communities

Participants join this program with a project that they either are already working on or want to develop during this program individually or in teams.



Don't be overwhelmed... just choose from the Open science buffet!



Don't try to stuff yourself on everything, select what works for this study and let's steadily improve our fields..

Take your pick from the 'buffet' of open science practices from transparency, statistics, preregistration, multi-lab collaborations...

open science buffet" metaphor (h/t Christina Bergmann)

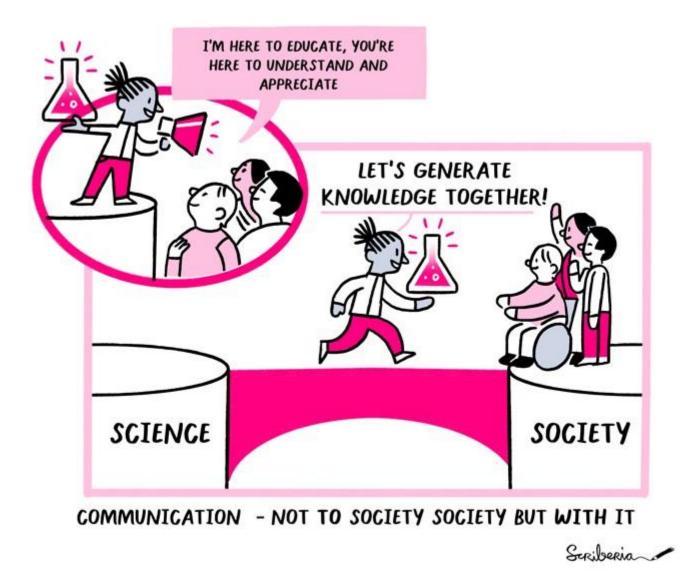


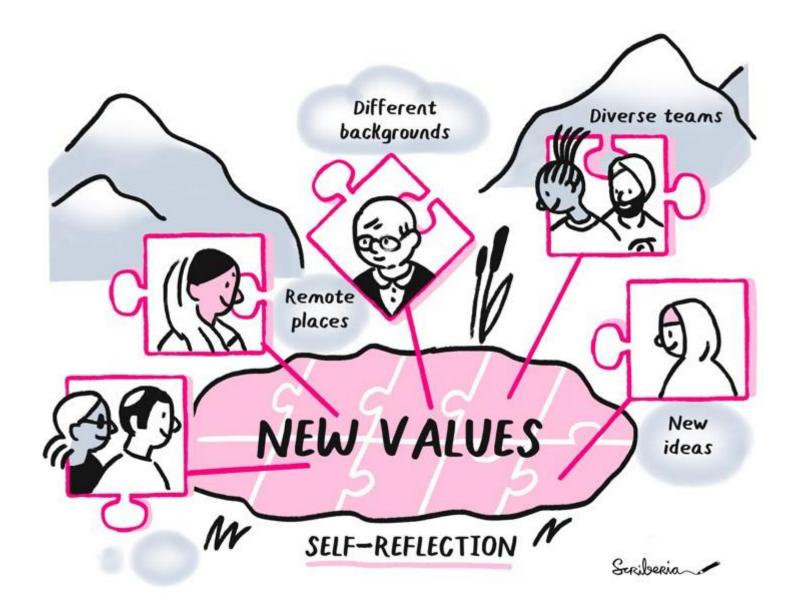
ALL IMAGES TAKEN FROM The Turing Way project illustration by Scriberia. Used under a CC-BY 4.0 licence. http://doi.org/10.5281/zenodo.3332807

There is another way... The Turing way!

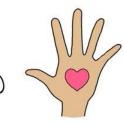


Self-reflection and change of mindset





Ladder of Participation (Arnstein, 1969)



		1
S	Citizen control	Stakeholders have the up the project
Citizen Power	Delegated power	Goal created by a faci resources and responsibilit citizens
Citiz	Partnership	Stakeholders have direct in decision making
	Placation	Stakeholders shape ideas decision sits with facili
Tokenism	Consultation	Stakeholder views are so decisions made by facil
. (Informing	Stakeholders are inform but no opportunity to con
Non participation	Educating	Assumption that the st are passive receipients
Non par	Manipulation	The illusion of participa actually power is denied
		n.

e idea and set

ilitator but ty given to

t involvement

s, but final litators

sought but ilitators med on decisions ntribute

stakeholders

ation when -

© creative . clinical . Psychologist

Drawn by Juliet youg