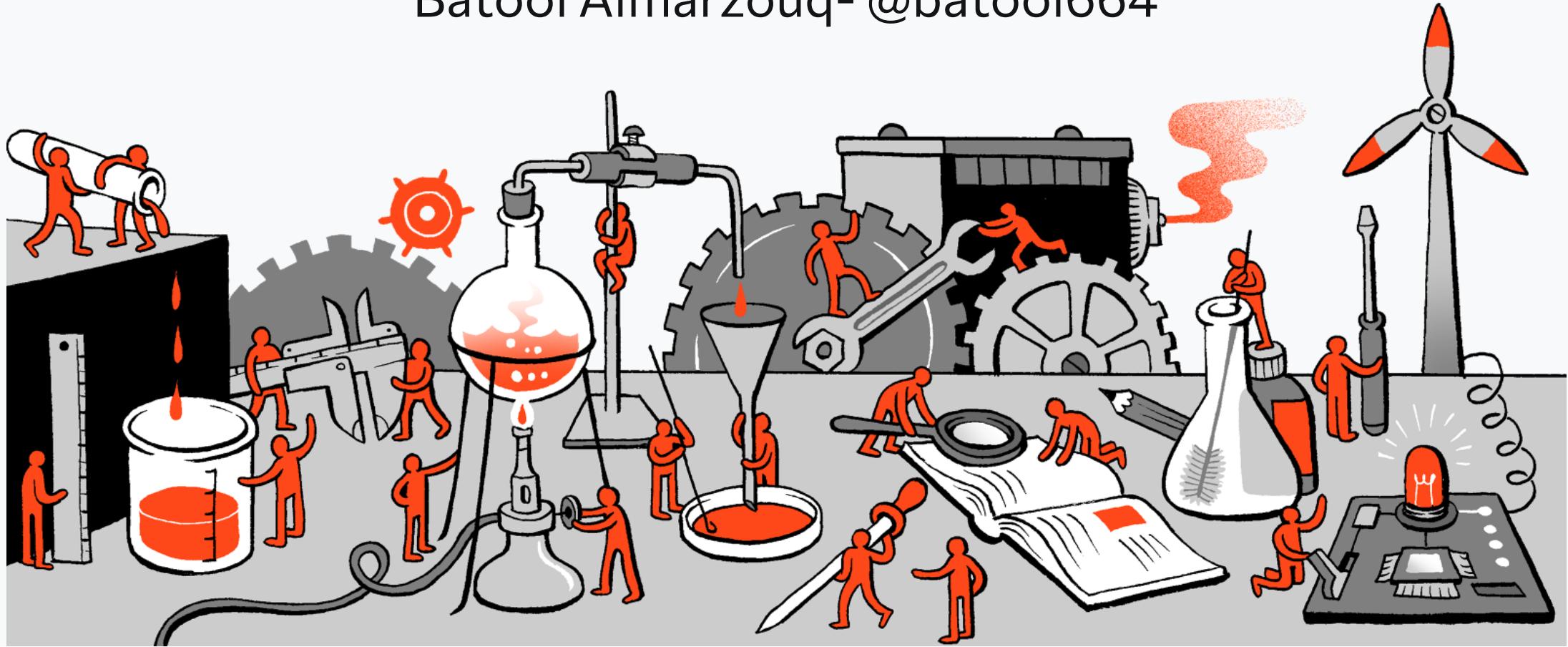


An Open Science Approach to Machine Learning in Biomedical Research

Batool Almarzouq- @batool664



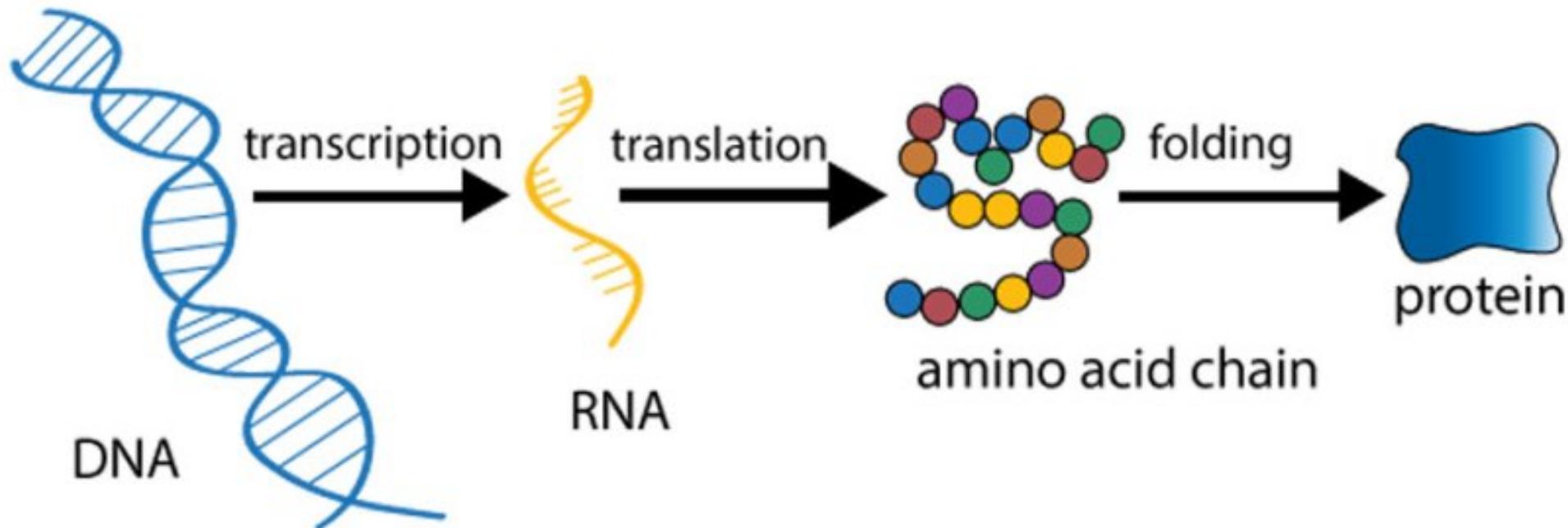
A little bit about me!

- A computational biologist affiliated with the University of Liverpool.
- Founder of RLadies Chapter in Saudi Arabia (Dammam).
- A curator in the R Weekly team.
- Member of MiR accessibility committee.
- Member in the turing way community.
- Working on establishing an Open Science community in Saudi Arabia.

Acknowledgment

- Anelda Van der
- Malvika Sharan, Kirstie Whitaker and Martina G. Vilas
- The Turing Way Community
- Alison Presmanes Hill (slides)

Why do we use ML in Biomedical
Research?



DNA

- DNA sequence alignment
- DNA sequence classification
- DNA sequence clustering
- DNA pattern mining

Algorithms includes fuzzy sets, neural networks, genetic algorithms.

Image Credit: [ABC Science](#)

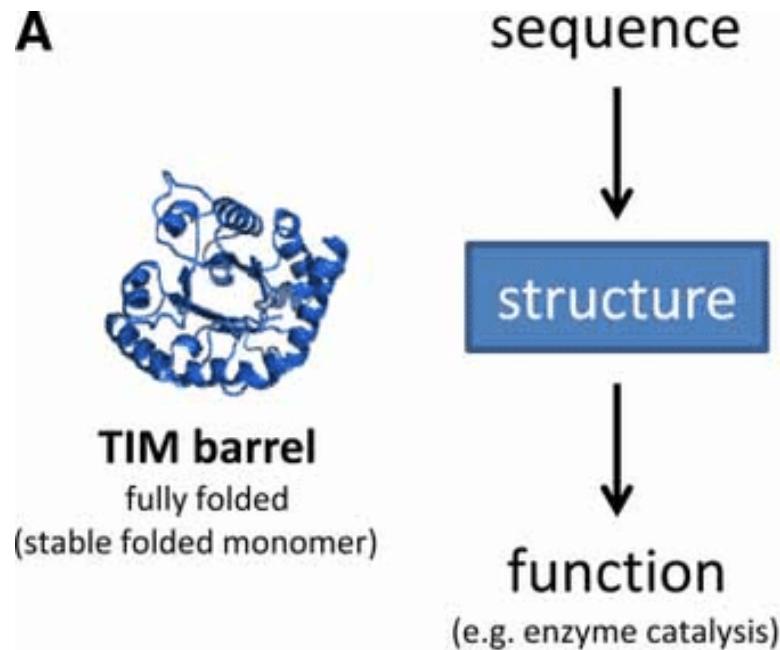
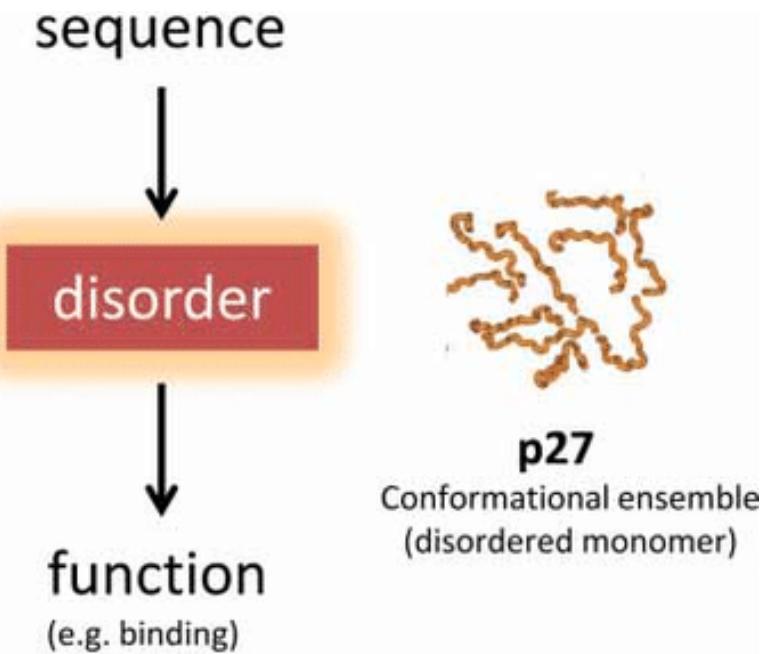
RNA

- Mainly RNA-sequencing (RNA-seq)
- Differentially expressed genes (DEGs)
- Alternative splicing
- Small RNA expression

Algorithms include Logistic Regression, Random Forest, LMT, Random Subspace.

Solving the sequence is not enough!

We need to know the structure and function of the protein!

A**B**

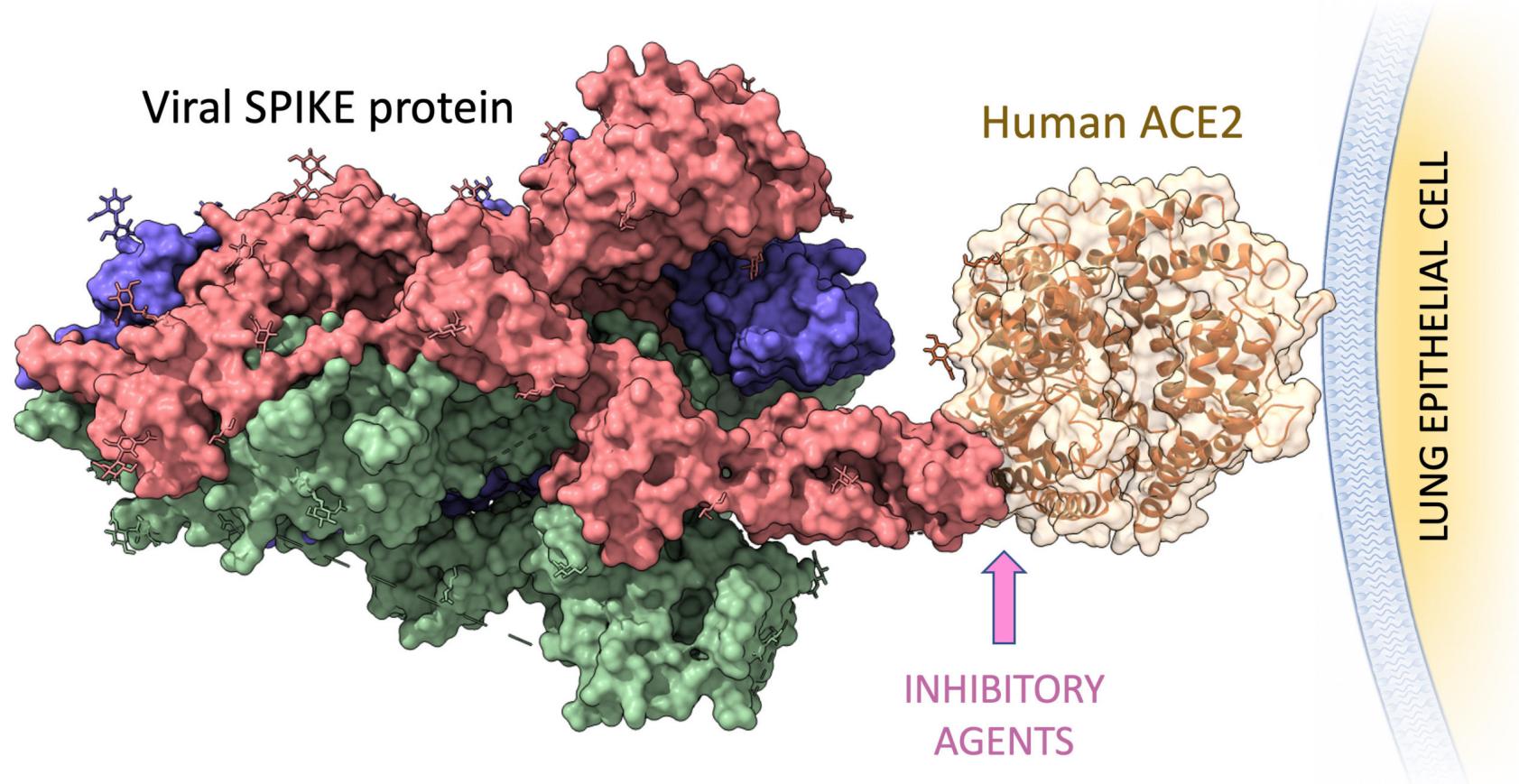
structure-function paradigm
(established)

disorder-function paradigm
(emerging)

HOW CAN WE PREDICT FUNCTION FROM STRUCTURE?

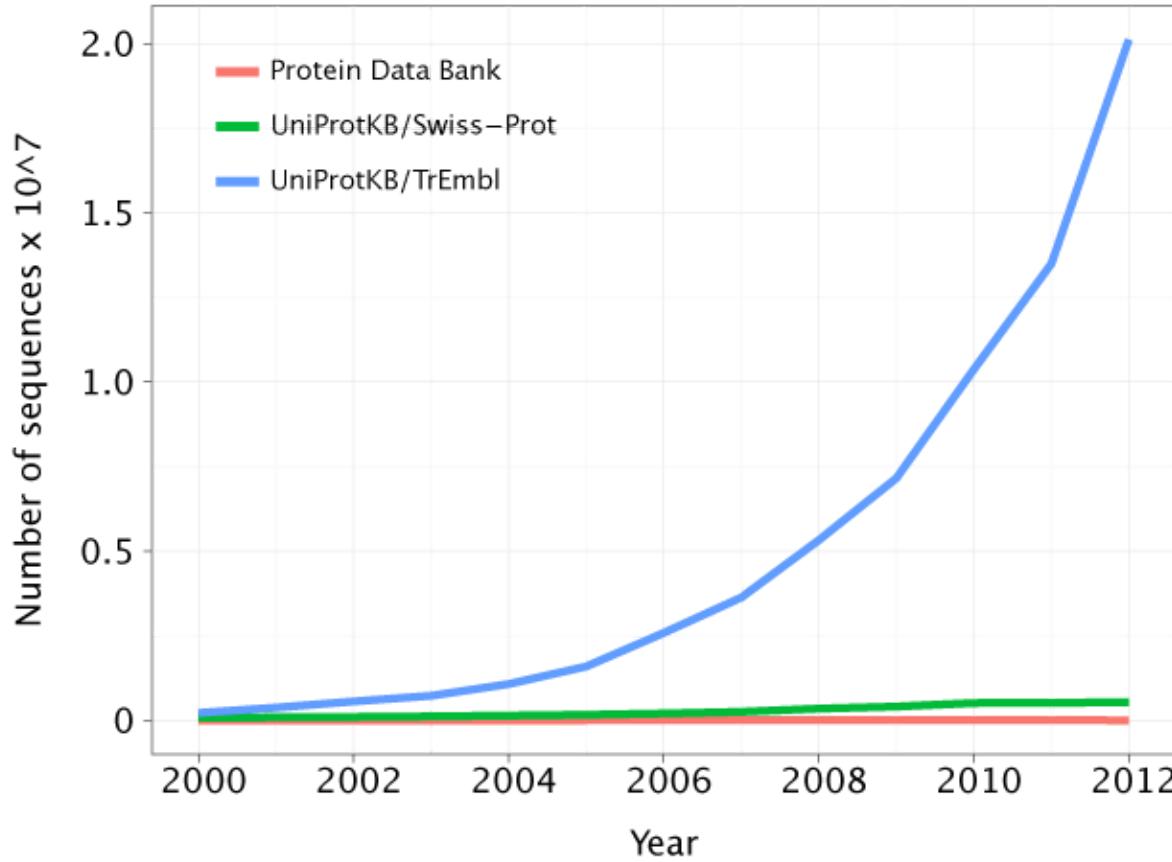
To predict the function from the structure, scientists use different approaches including machine learning (ML) and deep learning algorithms .

Image Credit: [doi:10.1021/cr400525m](https://doi.org/10.1021/cr400525m)



Prediction of protein structure is important to develop small molecules and targeted therapy for diseases.

Why not only rely on Experemtal Methods?



Because of the growing gap between the newly-sequenced and characterized sequences in the genome databases, computational methods in gene functional annotation are indispensable. Moreover, given the drop in the genome sequencing techniques' cost, this gap is only destined to grow.

Credit: Data for UniProtKB obtained from Claire O'Donovan via EBI database support

Biology has become a highly data-intensive science, dependent on complex, computational, and statistical methods!

So, how can we make these methods available and accessible for researchers, while ensuring that scientific results remain reproducible?

What is the percentage of reproducible research?

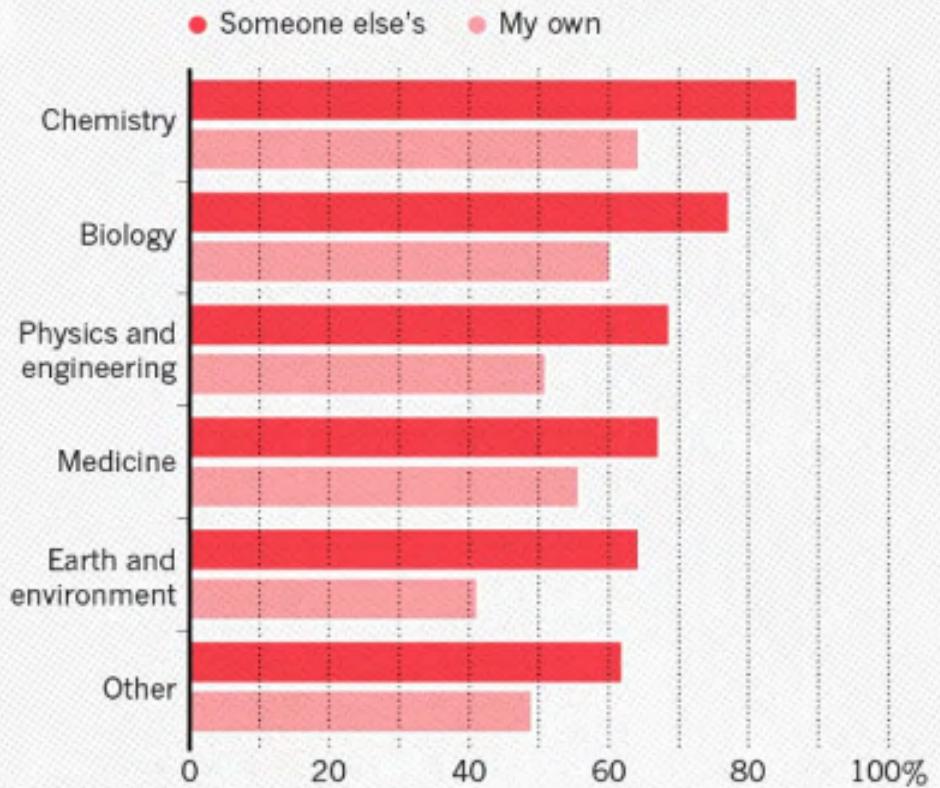
IS THERE A REPRODUCIBILITY CRISIS?



©nature

HAVE YOU FAILED TO REPRODUCE AN EXPERIMENT?

Most scientists have experienced failure to reproduce results.



Credit: Key results of the survey on reproducibility conducted by Nature in 2016

How can we overcome the reproducibility
crisis?

How can you improve the reproducibility of your data science project?

OPEN SOURCE SOFTWARE

SHARE CODE/ANALYSIS

SHARE COMPUTATIONAL ENVIRONMENT

VERSION CONTROL

TESTING

DOCUMENTATION

OPEN DATA/FAIR DATA

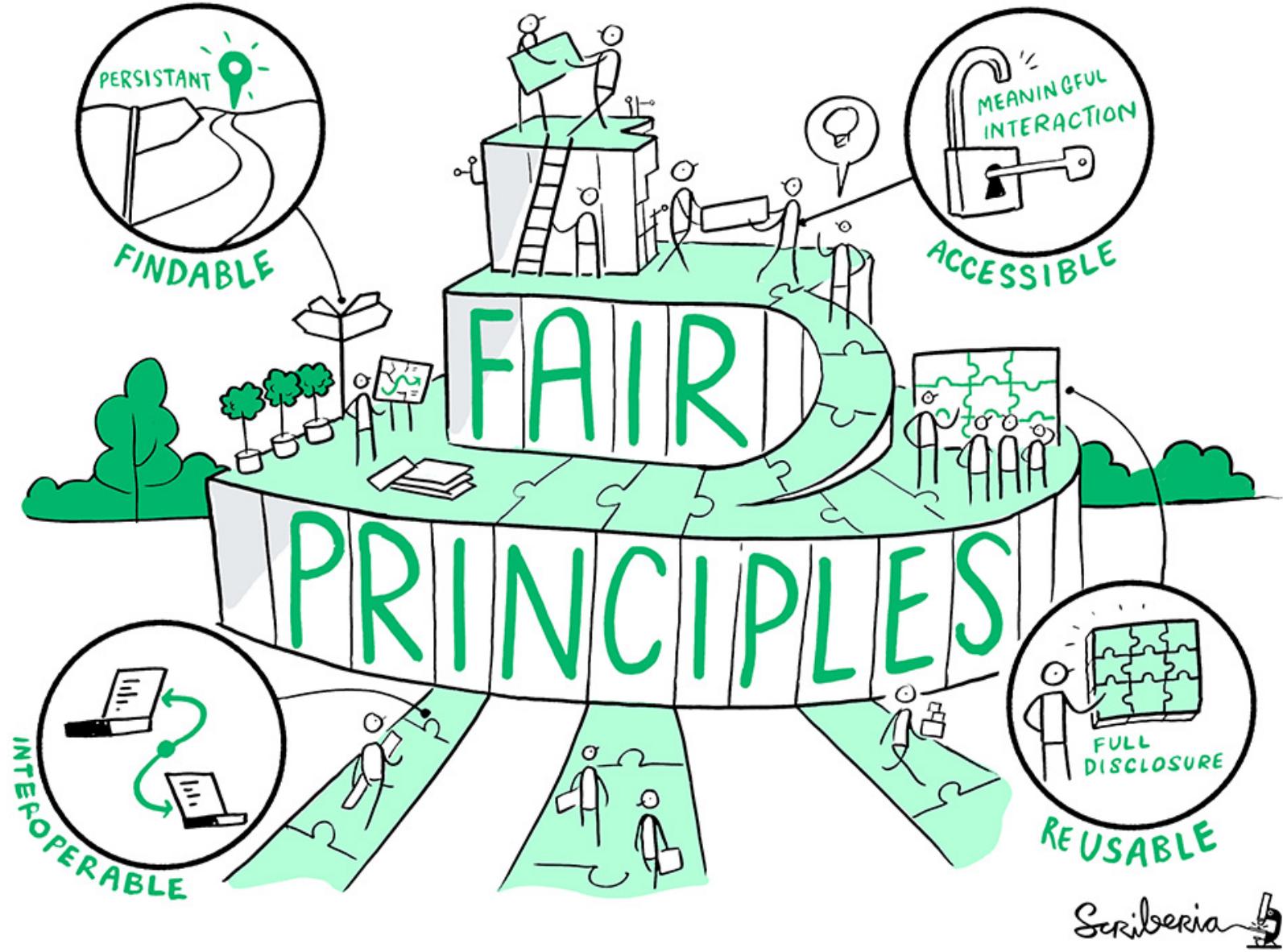
OPEN ACCESS

This is called Open Science.

Open Science is about extending the principles of openness to the whole research cycle, fostering sharing and collaboration as early as possible thus entailing a systemic change to the way science and research is done

-- FOSTER Plus

What are the FAIR principles?

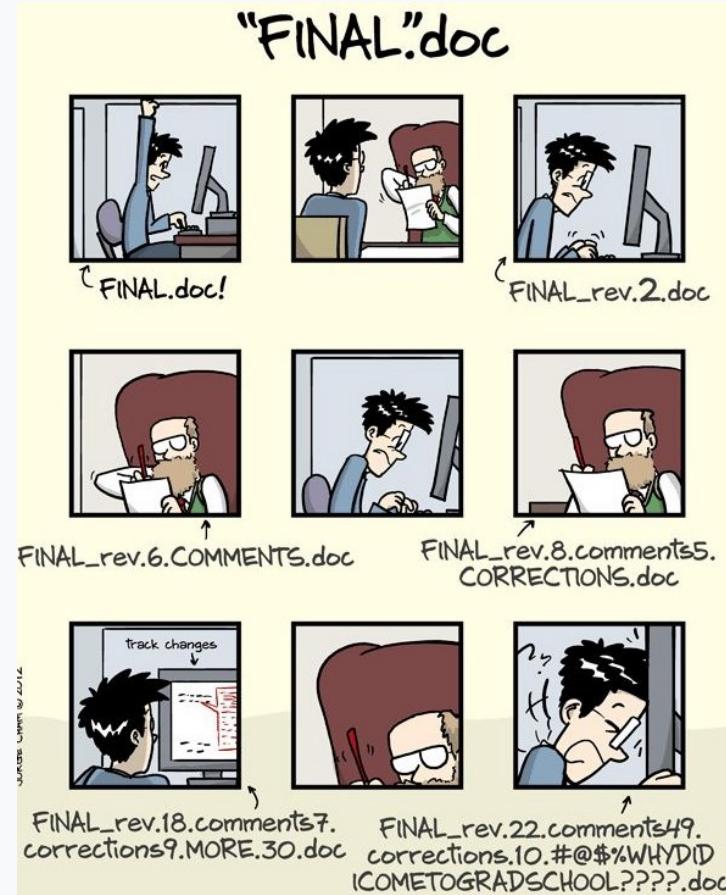


* The Turing Way project illustration by Scriberia. Zenodo. <http://doi.org/10.5281/zenodo.3332807>

Why do we use version control (git)?

Version Control in the Old Days ..

FINAL
FINAL.FINAL
final.FOR REAL
FINAL.version 2
absolutely.FINAL
FINAL.2
FINAL.3
FINAL.3.01
FINAL.3.02
FINAL.working



Real Version Control (including backup)

BatoolMM / An-Open-Science-Approach-to-Machine-Learning

Unwatch 1 Star 0 Fork 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

adding the first part of the talk

main

BatoolMM committed 10 hours ago

1 parent d4b57f7 commit b035487ef1f2deae91a8c58149a30b83ff622eac

Showing 2 changed files with 89 additions and 1,154 deletions.

Unified Split

569 bin/presentation.Rmd

... ... @@ -1,6 +1,6 @@

1 1 ---

2 - title: "An Open Science Approach to Machine Learning in Medical and Biological Research"

3 - subtitle: "Saudi Data Community"

2 + title: "An Open Science Approach to Machine Learning in Biomedical Research"

3 + subtitle: "Talk @ Saudi Data Community"

4 4 author: Dr. Batool Almarzouq

5 5 date: ``r Sys.Date()``

6 6 output:

@@ -22,9 +22,8 @@ output:

22 22 ---

The screenshot shows a GitHub repository page for 'BatoolMM / An-Open-Science-Approach-to-Machine-Learning'. The commit 'adding the first part of the talk' by BatoolMM is highlighted, showing 1 parent commit (d4b57f7) and a commit message 'commit b035487ef1f2deae91a8c58149a30b83ff622eac'. Below the commit, it says 'Showing 2 changed files with 89 additions and 1,154 deletions.' A detailed diff view for 'bin/presentation.Rmd' is shown, comparing two versions of the file. The changes include updates to the title from 'Medical and Biological Research' to 'Biomedical Research', and the subtitle from 'Saudi Data Community' to 'Talk @ Saudi Data Community'. The diff view uses color coding to highlight additions (green) and deletions (red).

In the pandemic, some publishers have “opened” their journals to make certain articles freely available.

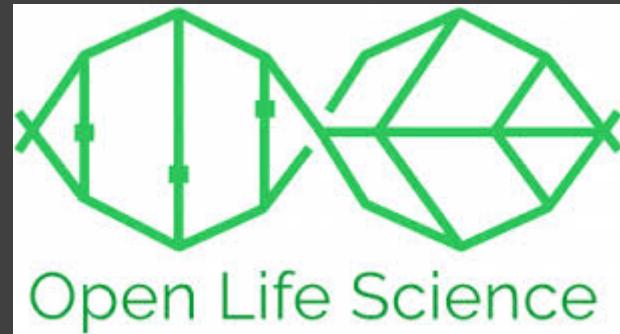
Databases have been created that are completely open access, such as the Open COVID Pledge.

UNESCO is launching international consultations
aimed at developing a Recommendation on Open
Science for adoption by member states in 2021

There is a network of Open Science Communities
in Netherlands, Sweden, Germany, UK and others

In line with vision 2030, we are starting an Open Science Community in Saudi Arabia.

It's created and developed with the help of the
"Open Life Sciences"



Open Life Sciences (OLS3) program helps individuals and stakeholders in research to become Open Science ambassadors.

We want to provide a place where newcomers and experienced peers interact, inspire each other to embed open science (research) practices and values in their workflows and provide feedback on policies, infrastructures and support services. Together working to make Open Science the norm. So we are calling out to researchers and colleagues in Saudi Arabia.



Batool Almarzouq
The University of
Liverpool



Founder and director
of Talarify, Mentor
OLS3



Paula Moraga,
Assistant Professor in
Statistics for Public
Health (KAUST)

Join me on the 24th
of Feb for a
workshop titled
"Collaborating on
Open Data Science
Projects" as part of
the Datathon for
WiDS2021.

How can you start learning about Open Science?

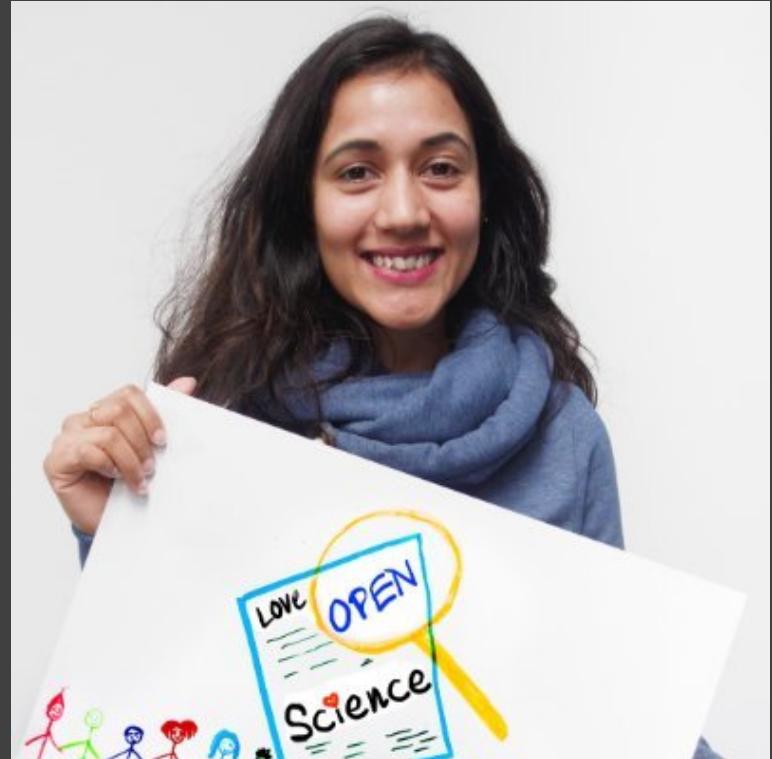


* The Turing Way project illustration by Scriberia. Zenodo. <http://doi.org/10.5281/zenodo.3332807>

Kirstie Whitaker,
Project Lead



Malvika Sharan,
Community Manager



So, What is the turing way?



Book:
the-turing-way.netlify.app/

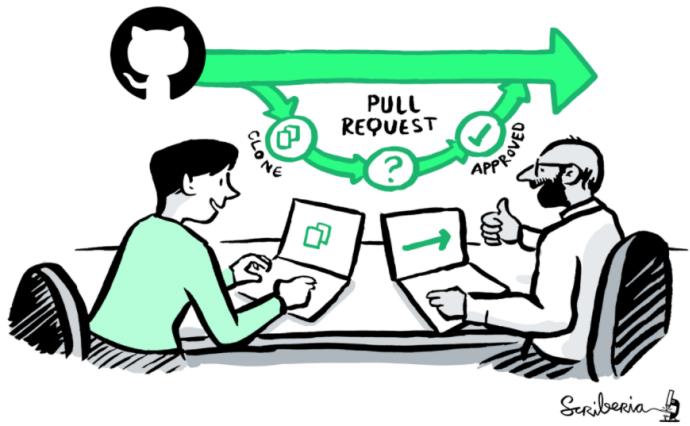
GitHub:
github.com/alan-turing-institute/the-turing-way

Twitter:
twitter.com/turingway

Email:
theturingway@gmail.com

CC-BY 4.0, *The Turing Way*

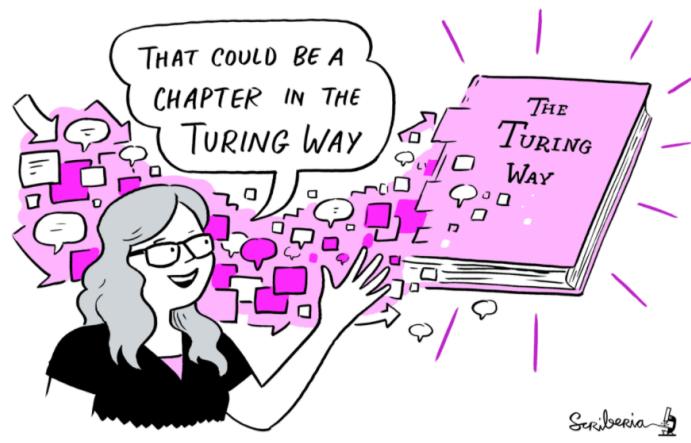
An Open Source Project



A Community



A Book



A Culture of Collaboration



Moonshot Goal: Reproducibility “too easy not to do”



The Turing Way

Search this book...

Welcome

- Guide for Reproducible Research
- Guide for Project Design
- Guide for Communication
- Guide for Collaboration
- Guide for Ethical Research
- Community Handbook
- Afterword

Visit our [GitHub Repository](#)
This book is powered by [Jupyter Book](#)

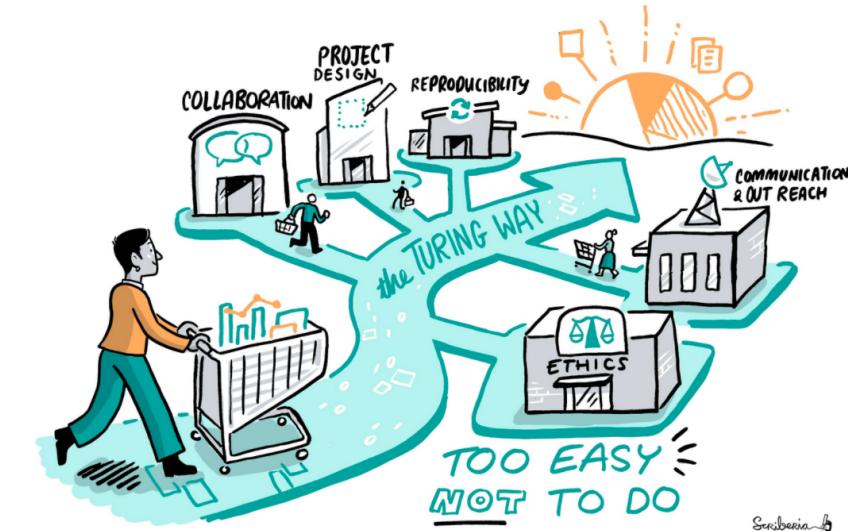
Welcome

The Turing Way is an open source community-driven guide to reproducible, ethical, inclusive and collaborative data science.

Our goal is to provide all the information that data scientists in academia, industry, government and the third sector need at the start of their projects to ensure that they are easy to reproduce and reuse at the end.

The book started as a guide for reproducibility, covering version control, testing, and continuous integration. However, technical skills are just one aspect of making data science research “open for all”.

In February 2020, *The Turing Way* expanded to a series of books covering reproducible research, project design, communication, collaboration, and ethical research.



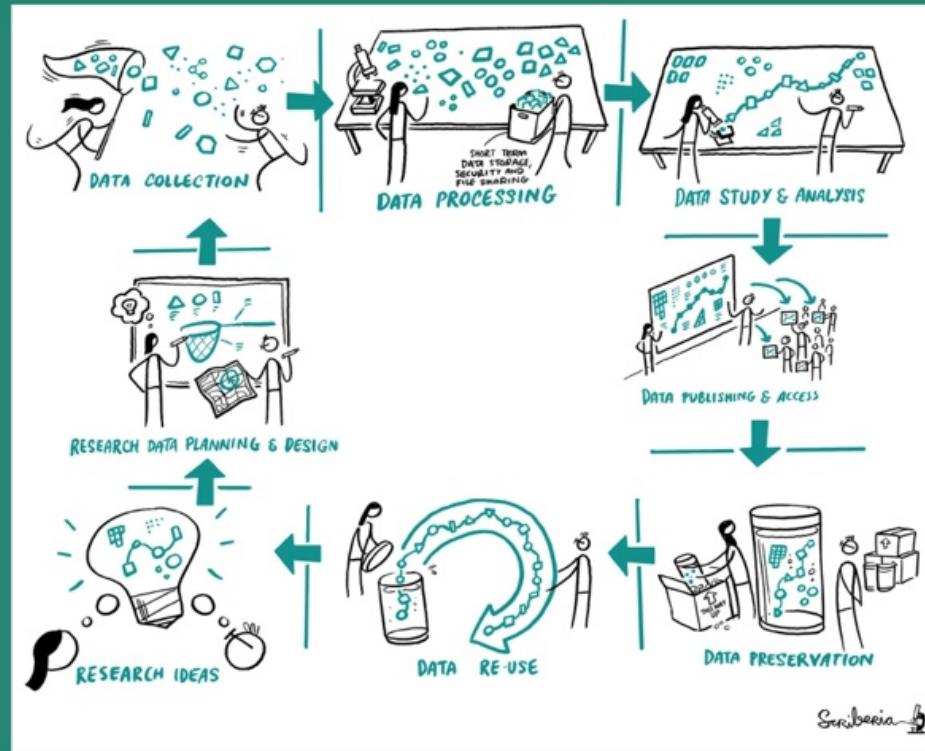
Scrubberia

The Turing Way book with five guides: <https://the-turing-way.netlify.app/welcome>, @turingway, DOI: 10.5281/zenodo.4650221

A book on reproducibility

Same analysis steps on the same dataset produces same answer

*The moonshot goal of the project
is to make reproducible research
"too easy not to do"*



Pathways for Collaboration



Connect with us



Start where you can



Edit, review, update



Help make it global



Join the community

Join the next book dash event!

BOOK DASH NOVEMBER 2020

Review README.md Arabic translation

Translate README.md to Arabic #1586

Changes from all commits ▾ File filter... ▾ Jump to... ▾ ⚙ ▾ 0 / 1 files viewed ⓘ Review changes ▾

171 README-translated/README-ARABIC.md

31 + **حول المشروع**
32 + البحث القابل للتكرار ضروري لضمان الوثوق بالعمل العلمي، بدأ الممоловون والناشرون في المطالبة بأن تتضمن المنشورات إمكانية الوصول إلى البيانات الأساسية وكود
33 + التحليل.
34 أي أن مشاركة مخرجات البحث مهمة في فهم إدارة البيانات، وعلوم المكتبات، وتطوير البرامج وتقنيات التكامل المستمر، وهي المهارات التي لم يتم تدريسيها على نطاق
واسع، أو لم يتم توقعها من الباحثين الأكاديميين وعلماء البيانات. نظرًا لأن هذه الأنشطة لا يتم تدريسيها بشكل شائع، فإننا ندرك أن عبء المتطلبات واكتساب المهارات
35 الجديدة يمكن أن يكون مخيّلاً للأفراد الجدد في هذا العالم.
36 أي أن مشاركة مخرجات البحث مهم في فهم إدارة البيانات وعلوم المكتبات وتطوير البرامج وتقنيات التكامل المستمر، وهي المهارات التي لم يتم تدريسيها على نطاق
واسع أو لم يتم توقعها من الباحثين الأكاديميين وعلماء البيانات، نظرًا لأن هذه الأنشطة لا يتم تدريسيها بشكل شائع، فإننا ندرك أن عبء المتطلبات واكتساب المهارات
37 الجديدة يمكن أن يكون مخيّلاً للأفراد الجدد في هذا العالم.
38 طريقة تورننج هي عبارة عن كتاب لدعم الطلاب ومشرفيهم ومموليهم ومحرري المجلات في شأن أن علم البيانات القابل للتكرار "من السهل جدًا عدم القيام به" حتى بالنسبة
لأشخاص الذين لم ي عملوا بهذه الطريقة من قبل، وسيشمل مواد تدريبية حول التحكم في الإصدار واختبار التحليل والتواصل المفتوح والشفافية مع المستخدمين المستقبليين،
والبناء على دراسات الحالة وورش العمل الخاصة بمعهد تورننج
39 تم تطوير هذا المشروع بشكل مفتوح ، نرحب بأي وجميع الأسئلة والتعليقات والتوصيات في بوابة حيث هب،
40 + **الفريق**
41 +
42 + (Turing Institute)
43 فريق العمل من معهد تورننج
44 +
45 لمزيد من المعلومات حول كيفية الاتصال بنا ، راجع القسم الأخير.
46 +
47 + **المعاونة**

Rachael Ainsworth Ainsworth 	Tarek Allam Allam 	Tania Allard Allard 	Diego Alonso Alvarez Alvarez 	Kristijan Armeni Armeni 	Becky Arnold Arnold 	Louise Bowler Bowler 	Paula Andrea Martinez Martinez 	Lachlan Mason Mason 	Rohit Mishra Mishra 	Javier Molina Molina 	Beth Montague-Hallinan Montague-Hallinan 	Alexander Morley Morley 	James Myatt Myatt 	
Alex Clarke Clarke 	Jez Cope Cope 	Eric Dahl Dahl 	Stephen Druskat Druskat 	Stephan DuPre DuPre 	Stephen Eglen Eglen 	Joe Fennell Fennell 	Ollie Jimenez Jimenez 	Martin O'Reilly O'Reilly 	Pickering Pickering 	Jude Jude 	Carmine Ranger Ranger 	Roxi Readoff Readoff 	James Robinson Robinson 	Susanna Asensio Asensio
<h1>Join our community of contributors & friends!</h1>														
Oliver Forrest Forrest 	Pooja Gadige Gadige 	Jason Gates Gates 	Sarah Gibson Gibson 	Oscar Giles Giles 	Richard Gilham Gilham 	Cassandra Gould van Praag Gould van Praag 	Ali Seyhun Saral Seyhun Saral 	Chamaki Illushka Seresinha Illushka Seresinha 	Nadia Soliman Soliman 	Andrew Stewart Stewart 	Sarah Stewart Stewart 	Oliver Strickson Strickson 	Natalie Thurlby Thurlby 	
Michael Grayling Grayling 	Liberty Hamilton Hamilton 	Tim Head Head 	Patricia Herterich Herterich 	Rosie Higman Higman 	Ian Hinder Hinder 	Hieu Hoang Hoang 	Gertjan van den Burg van den Burg 	Kirstie Whitaker Whitaker 	Tany Yang Yang 	Yo Yehudi Yehudi 	Mahika Sharan Sharan 	Esther Esther 	Anna Hadjithof Hadjithof 	
Dan Hebley Hebley 	Chris Holdgraf Holdgraf 	Will Hulme Hulme 	James Kent Kent 	Greg Kiar Kiar 	Danhee Kim Kim 	Anna Krystalli Krystalli 	Kevin Kunzmann Kunzmann 	Eric Leung Leung 	Clare Liggins Liggins 	Robin Long Long 	Christopher Lovell Lovell 	Eirini Mallaraki Mallaraki 	Chris Markiewicz Markiewicz 	

Wall of all Contributors
tagged by emojis

- eventOrganizing
- financial
- fundingFinding
- ideas
- review
- security
- tool
- translation
- test
- tutorial
- talk

Upcoming Workshop by the turing way

Register for the free workshop '[Boost your research reproducibility with Binder](#)' run by Sarah Gibson from the Turing Way as part of our Research Software Camp on research accessibility.



Resources:

- [The Turing Way](#)
- [The CMU ML Blog](#)
- [Redesign open science for Asia, Africa and Latin America](#)
- [Open Science Beyond Open Access: For and with communities, A step towards the decolonization of knowledge](#)
- [Embracing science as it is: beyond Nobel-like research](#)
- [Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA](#)

Thank you so much!

batool@liverpool.ac.uk

Twitter: @batool664

Join RLadiesDammam: @RLadiesDammam