UKCCB
Conference of Bioinformatics and Computational Biology

Leveraging

Open Science in Machine Learning and Bioinformatics

Batool Almarzouq

OSCSA
Open Science Community Saudi Arabia
مجتمع العلوم المفتوحة في المملكة العربية السعودية

Thank you!

- Yo Yehudi
- Piv Gopalasingam
- John Ogunsola
- Emily Angiolini
- Turing Way

Batool Almarzouq

Computational Biologist

King Abdullah International Medical Research Center

Honorary research Fellow

UNIVERSITY OF LIVERPOOL

Community Lead

OSCSA
Open Science Community Saudi Arabia

Education Committee

ISCB
INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Core Contributor

The Turing Way

Global Team

R Ladies

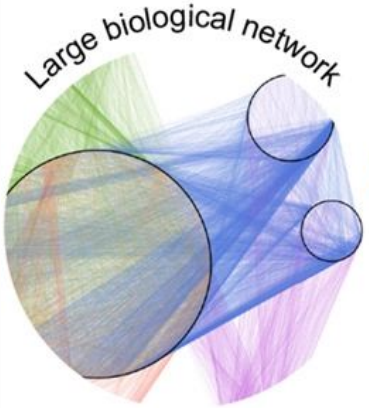
R LADIES DAMMAM
SAUDI ARABIA

Subject Matter Experts (SME)

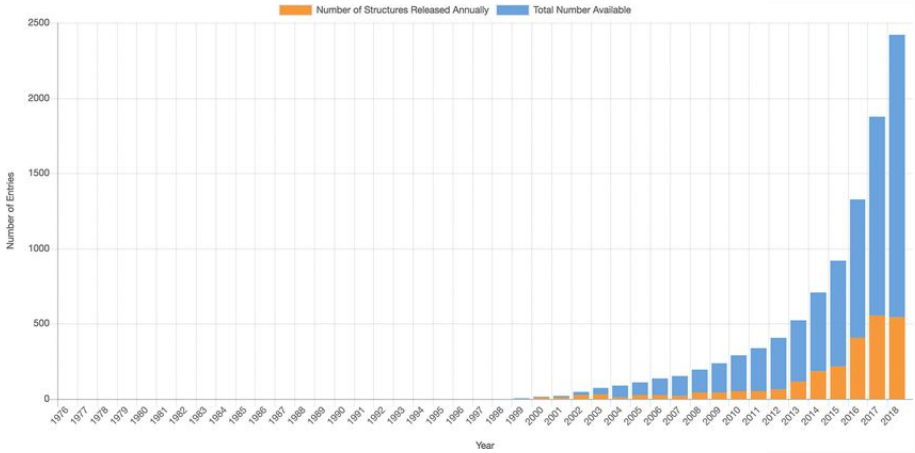
NASA

TOPS

Biology is **data-intensive** science



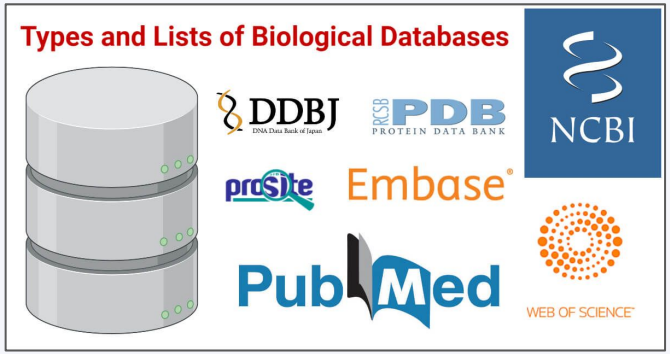
Descriptive biology
(data integration)



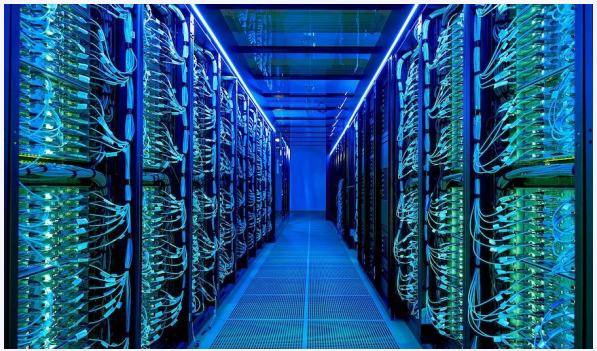
Source: Miquel Duran-Frigola *et al*, Formatting biological big data for modern machine learning in drug discovery

Can we **go beyond** that?

With the improvements in computing power, storage and exponential growth in biological data, can we **answer what we failed to solve** using conventional methods?

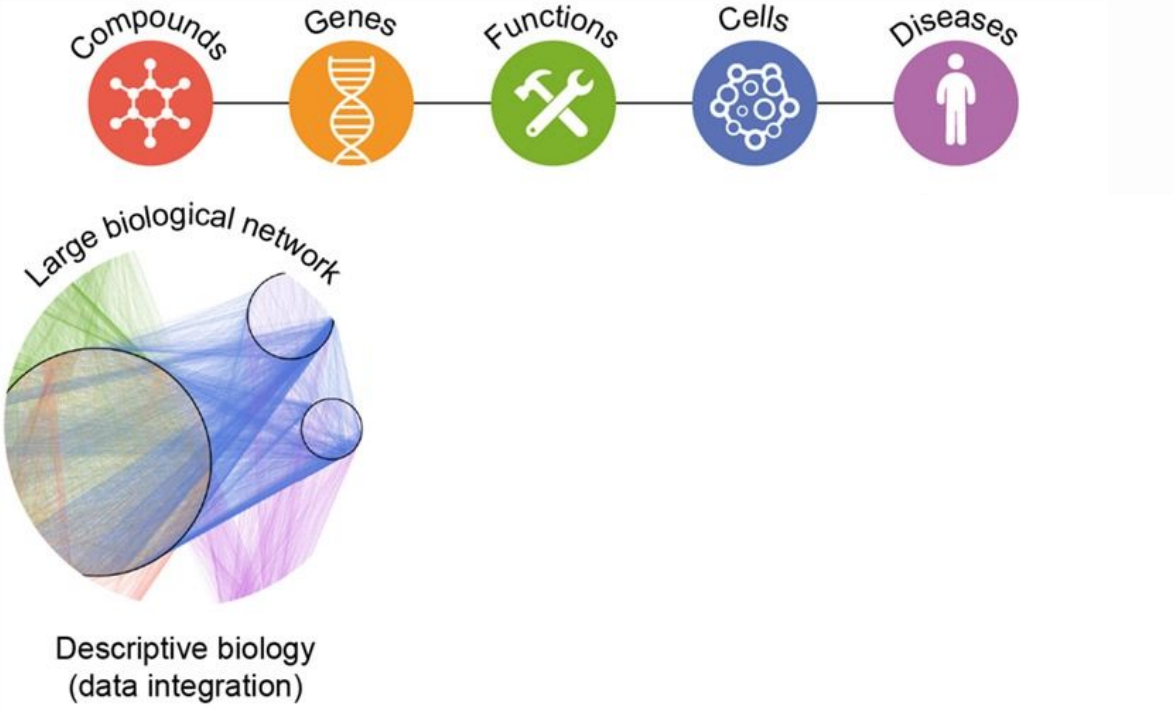


Biological Databases. Image Source: The Biology Notes



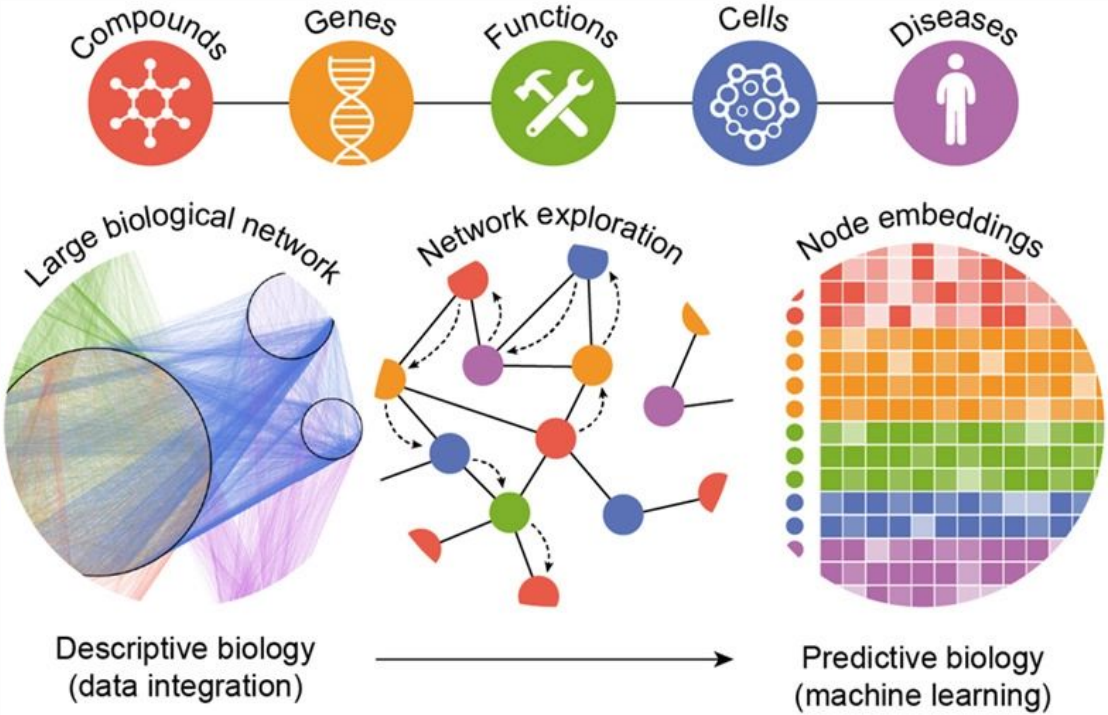
Credit: The supercomputer Hochleistungsrechner Karlsruhe“ (Author: Amadeus Bramsiepe, KIT)

Superpower!



Source: Miquel Duran-Frigola *et al*, Formatting biological big data for modern machine learning in drug discovery

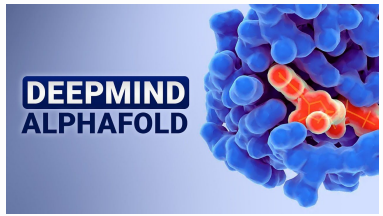
Superpower!



Source: Miquel Duran-Frigola *et al*, Formatting biological big data for modern machine learning in drug discovery

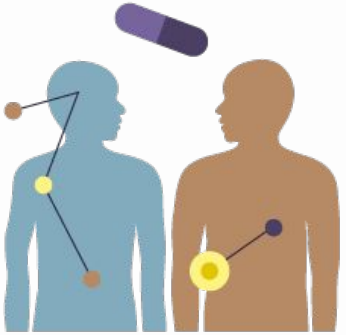
Success!

	Protein structure prediction	Protein function prediction	Genome engineering	Systems biology and data integration	Phylogenetic inference
Paradigm shifting	✓				
Major success		✓	✓		
Moderate success				✓	
Minor success					✓



DeepGOPlus: Improved protein function prediction from sequence

CRISPR
GENE EDITING TOOL



Source: Sapoval, N., Aghazadeh, A., Nute, M.G. *et al.* **Current progress and open challenges for applying deep learning across the biosciences.**

If that's the case!

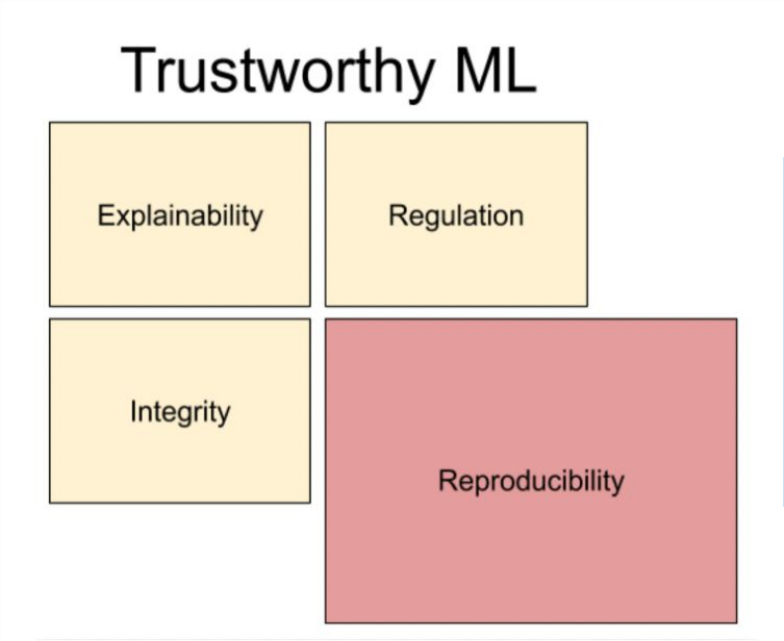
Source: Techlady

Challenges we have seen so far!

Challenge
Biased results
High infrastructure costs
Lack of explainability
Limited training data
Overfitting
Poor performance on novel data

Source: Sapoval, N., Aghazadeh, A., Nute, M.G. *et al.* **Current progress and open challenges for applying deep learning across the biosciences**
Cartoon Credit: Macloo.

What is trustworthy ML?



Resource: Reproducible Machine Learning by Preeti Hemant

If deep learning is to solve real scientific and technological problems, the community needs to build on each **others' tools and data** by implementing **Open Science practices.**

Reproducibility Crisis!

The $28 Billion a Year Research Reproducibility Crisis

October 19, 2015

The biggest crisis in the future of healthcare has
talk about, such as payment systems and access
conditions like heart disease, dementia, diabetes
treatments that rely upon genomics, nanotechn
that at least half of the research money we inves
implications? What are the solutions? We'll show

Comment | Published: 30 August 2021

Reproducibility standards for machine learning in t life sciences

Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene & Steph Hicks

Nature Methods **18**, 1132–1135 (2021) | Cite this article

15k Accesses | **16** Citations | **267** Altmetric | Metrics

Session 2: Bioimaging and Artificial Intelligence

Session Lead: Jean-Marie Burel, Senior Software Architect, University of Dundee

Talk 1: Yang Zhang, Project Scientist & Project Manager, Carnegie Mellon University (US

Nucleome Browser: An integrative and multimodal data navigation platform for 4D Nucleome

Talk 2: Matthew Hartley, BioImage Archive Team Leader, European Bioinformatics Institute (EMBL-EBI)

Open bioimaging data at scale: publication, analysis and reuse

Talk 3: Josh Moore, Senior Software Architect, University of Dundee

OME-NGFF (next-generation file format): Zarr as a cloud-native solution for FAIRer bioimaging data

Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush & Hugo J. W. L. Aerts

Nature **586**, E14–E16 (2020) | Cite this article

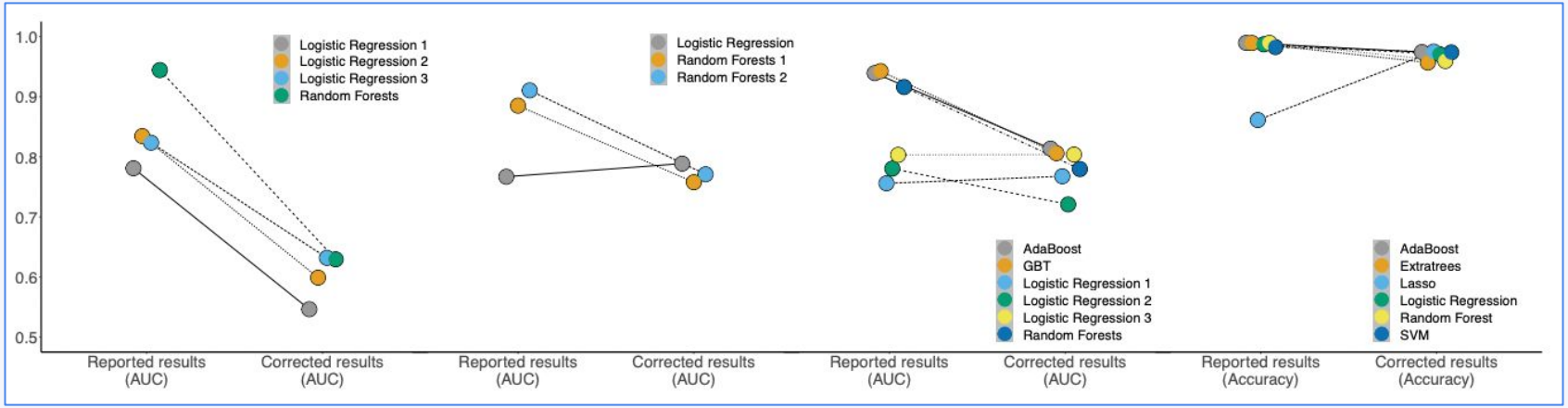
16k Accesses | **85** Citations | **530** Altmetric | Metrics

Reproducibility Crisis in ML?

Field	Paper	Number of papers reviewed	Number of papers with pitfalls	[L1.1] No test set	[L1.2] Pre-proc. on train-test	[L1.3] Feature sel. on train-test	[L1.4] Duplicates	[L2] Illegitimate features	[L3.1] Temporal leakage	[L3.2] Non-ind. b/w train-test	[L3.3] Sampling bias	Comput. reproducibility issues	Data quality issues	Metric choice issues	Standard dataset used?
Medicine	Bouwmeester et al. (2012)	71	27	○								○			
Neuroimaging	Whelan & Garavan (2014)	–	14	○		○									
Autism Diagnostics	Bone et al. (2015)	–	3				○				○		○	○	○
Bioinformatics	Blagus & Lusa (2015)	–	6		○										
Nutrition Research	Ivanescu et al. (2016)	–	4	○									○	○	
Software Eng.	Tu et al. (2018)	58	11					○				○	○		○
Toxicology	Alves et al. (2019)	–	1				○					○	○		
Satellite Imaging	Nalepa et al. (2019)	17	17						○				○		○
Tractography	Poulin et al. (2019)	4	2	○								○	○	○	○
Clinical Epidem.	Christodoulou et al. (2019)	71	48			○							○		
Brain-computer Int.	Nakanishi et al. (2020)	–	1	○											○
Histopathology	Oner et al. (2020)	–	1						○						
Neuropsychiatry	Poldrack et al. (2020)	100	53	○	○								○	○	
Medicine	Vandewiele et al. (2021)	24	21			○				○	○	○	○		○
Radiology	Roberts et al. (2021)	62	62	○			○				○	○			○
IT Operations	Lyu et al. (2021)	9	3					○							○
Medicine	Filho et al. (2021)	–	1						○						
Neuropsychiatry	Shim et al. (2021)	–	1			○						○			
Genomics	Barnett et al. (2022)	41	23			○							○		
Computer Security	Arp et al. (2022)	30	30	○	○	○		○		○	○		○	○	

Source: Leakage and the Reproducibility Crisis in ML-based Science by Sayash Kapoor, Arvind Narayanan

We should talk about data leak!



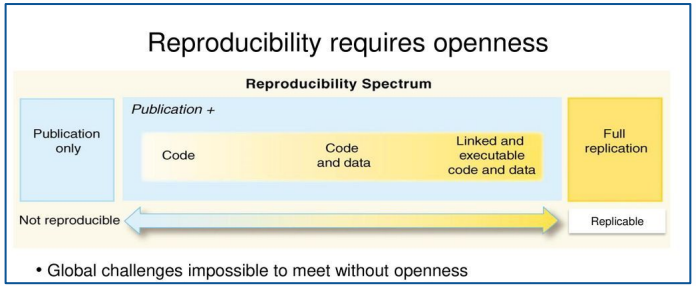
Source: Leakage and the Reproducibility Crisis in ML-based Science by Sayash Kapoor, Arvind Narayanan

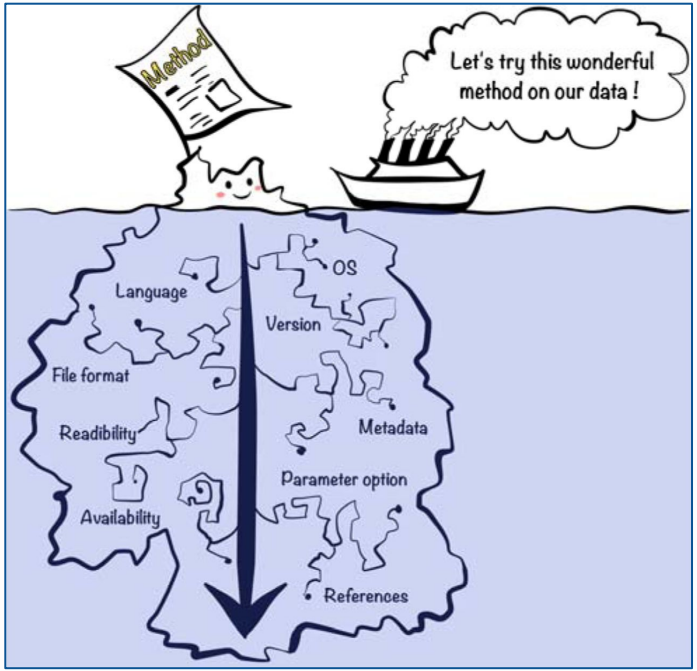
Peer Review is not enough!

> "To consult the [experts] **after an experiment is finished** is often merely to ask to **conduct a postmortem examination**. [...] can perhaps say what the experiment died of"
>
> - Ronald Fisher

This Slide is inspired from Talk by Malvika Sharan

Reproducibility require Openness!





Peng (2011)

Credit: Yang-Min Kim, Jean-Baptiste Poline, Guillaume Dumas, Experimenting with reproducibility: a case study of robustness in bioinformatics, *GigaScience*, Volume 7, Issue 7, July 2018, giy077, https://doi.org/10.1093/gigascience/giy077

What is Open Science?

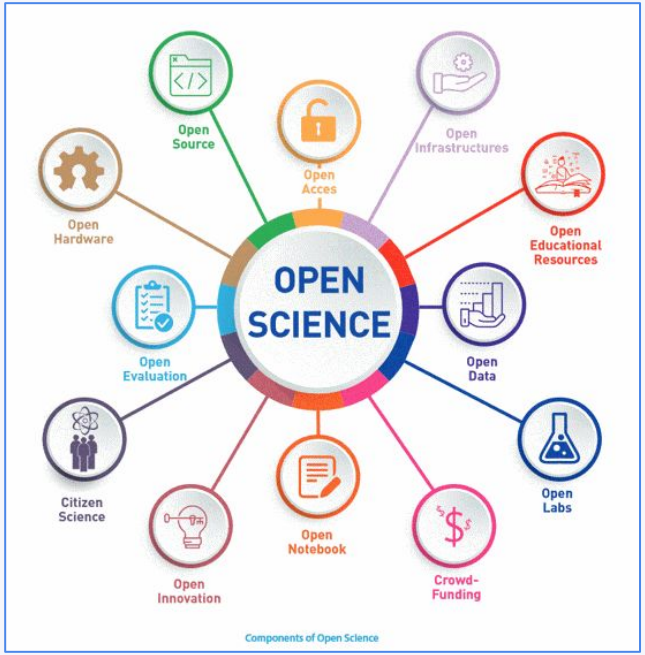
'Open Science' is an **umbrella** term that encopasses various partctises to increase the **visibility, the transparency, useability** of the scientific work.

NEWS | 03 February 2021

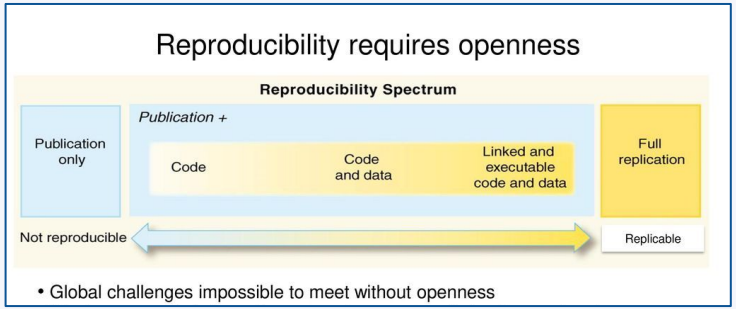
Scientists call for fully open sharing of coronavirus genome data

Other researchers say that restrictions at the largest SARS-CoV-2 genome platform encourage fast sharing while protecting data providers' rights.

Richard Van Noorden

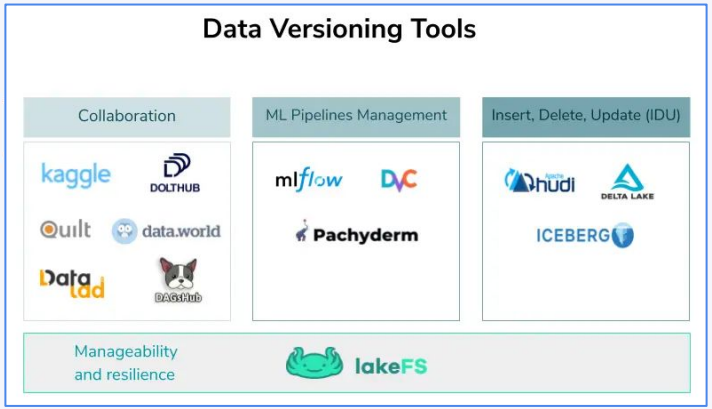


What can we do to implement Open Science?



Peng (2011)

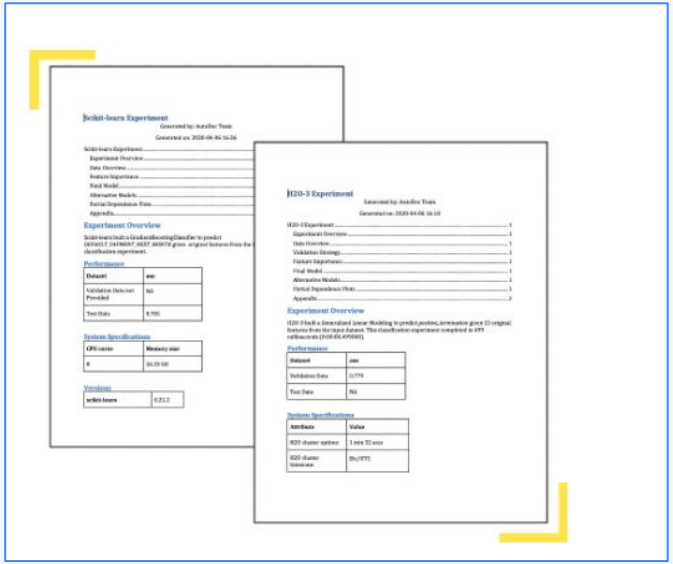
Versioning Data and Model



Source: KDnuggets

Source: MLOps Guide

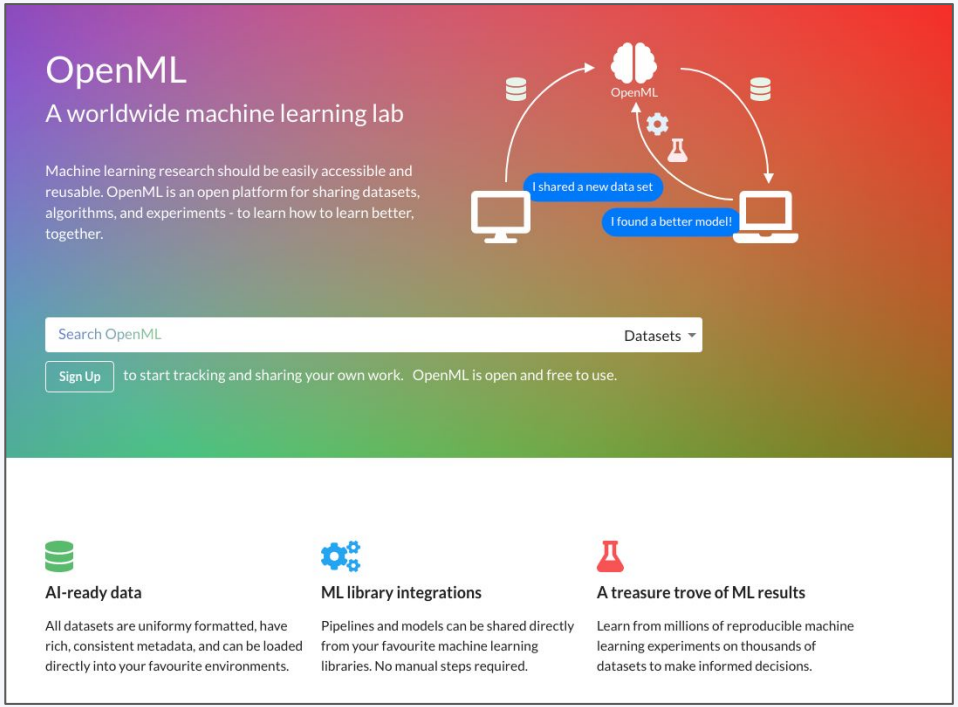
Documentation for Model Training



Documenting details of:

- How the model was trained will ensure repeatable results.
- Feature transformations, order of features,
- Hyperparameters and the method to select them

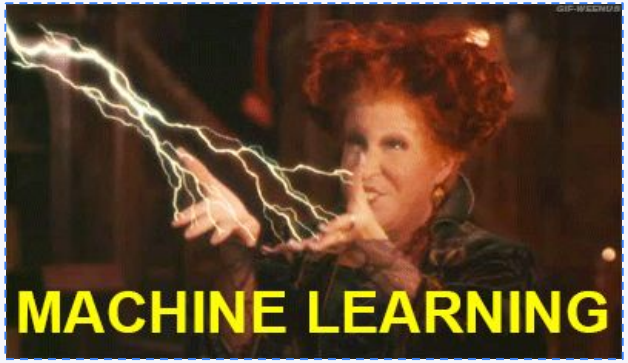
Community-based Projects!



Website: OpenML

Community-based Projects!

Reference: Ersilia Open Source Initiative – Strategic Plan 2021 – 2023 Main Points (Outreachy Contribution)



"So far, each research community has independently rediscovered these pitfalls. **Without fundamental changes** to research and reporting practices, **we risk losing public trust owing to the severity and prevalence of the reproducibility crisis across disciplines**"

– Sayash Kapoor and Arvind Narayanan

Discussion!

- What **concerns** do you have about sharing your bioinformatics research and are there any research objects you should not share?
- How to **support researchers and the community** creating a comprehensive open source machine learning environment?
- What is the **potential for innovation** when adopting open science in bioinformatics?

References

- Ersilia Open Source Initiative
- Reproducible Machine Learning by Preeti Hemant
- Leakage and the Reproducibility Crisis in ML-based Science by Sayash Kapoor, Arvind Narayanan
- AI slipping on tiles: data leakage in digital pathology by Nicole Bussola, Alessia Marcolini, Valerio Maggio, Giuseppe Jurman, Cesare Furlanello
- Experimenting with reproducibility: a case study of robustness in bioinformatics (GigaScience) by Yang-Min Kim, Jean-Baptiste Poline, Guillaume Dumas
- The Turing Way

Thank You