

From FAIR data to FAIR Research Software, towards FAIR Machine Learning

Fotis E. Psomopoulos

Institute of Applied Biosciences, CERTH, Greece



THE FAIR PRINCIPLES

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, [...] Barend Mons 

Scientific Data 3, Article number: 160018 (2016) | [Cite this article](#)

194k Accesses | 2450 Citations | 1852 Altmetric | [Metrics](#)

A set of principles, to ensure that data are shared in a way that enables and enhances reuse by humans and machines

Findable

- F1.** (meta)data are assigned a globally unique and eternally persistent identifier.
- F2.** data are described with rich metadata.
- F3.** (meta)data are registered or indexed in a searchable resource.
- F4.** metadata specify the data identifier.

Accessible

- A1** (meta)data are retrievable by their identifier using a standardized communications protocol.
 - A1.1** the protocol is open, free, and universally implementable.
 - A1.2** the protocol allows for an authentication and authorization procedure, where necessary.
- A2** metadata are accessible, even when the data are no longer available.

Interoperable

- I1.** (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2.** (meta)data use vocabularies that follow FAIR principles.
- I3.** (meta)data include qualified references to other (meta)data.

Reusable

- R1.** (meta)data have a plurality of accurate and relevant attributes.
 - R1.1.** (meta)data are released with a clear and accessible data usage license.
 - R1.2.** (meta)data are associated with their provenance.
 - R1.3.** (meta)data meet domain-relevant community standards.

FAIR FOR NON-DATA OBJECTS: SOME CONTEXT

- FAIR Principles, at a high level, are intended to **apply to all research objects**; both those used in research and those that are research outputs
- Text in principles often includes "(Meta)data ..."
 - Shorthand for "metadata and data ..."
- Principles applied via dataset creators and repositories, collectively responsible for creating, annotating, indexing, preserving, sharing the datasets and their metadata
- What about non-data objects?
 - While they can often be stored as data, they are not **just** data
- While high level goals (F, A, I, R) are mostly the same, the details and how they are implemented depend on
 - How objects are created and used
 - How/where the objects are stored and shared
 - How/where metadata is stored and indexed
- Work needed to define, then implement, then adopt principles

Slide adapted from the [presentation](#) of the [RDA FAIR4RS steering group](#) at the International Funders Workshop (Nov 2022), <https://zenodo.org/doi/10.5281/zenodo.7350198>



FAIR FOR NON-DATA OBJECTS: SOME EFFORTS

Ten simple rules for making training materials FAIR

Leyla Garcia, Bérénice Batut, Melissa L. Burke, Mateusz Kuzak, Fotis Psomopoulos, Ricardo Arcila, Teresa K. Attwood, Niall Beard, Denise Carvalho-Silva, Alexandros C. Dimopoulos, Victoria Dominguez del Angel, Michel Dumontier, Kim T. Gurwitz, [...], Patricia M. Palagi [view all]

Published: May 21, 2020 • <https://doi.org/10.1371/journal.pcbi.1007854>

FAIR for AI: An interdisciplinary and international community building perspective

[E. A. Huerta](#), [Ben Blaiszik](#), [L. Catherine Brinson](#), [Kristofer E. Bouchard](#), [Daniel Diaz](#), [Caterina Doglioni](#), [Javier M. Duarte](#), [Murali Emani](#), [Ian Foster](#), [Geoffrey Fox](#), [Philip Harris](#), [Lukas Heinrich](#), [Shantenu Jha](#), [Daniel S. Katz](#), [Volodymyr Kindratenko](#), [Christine R. Kirkpatrick](#), [Kati Lassila-Perini](#), [Ravi K. Madduri](#), [Mark S. Neubauer](#), [Fotis E. Psomopoulos](#), [Avik Roy](#), [Oliver Rübél](#), [Zhizhen Zhao](#) & [Ruike Zhu](#)

Scientific Data **10**, Article number: 487 (2023) | [Cite this article](#)

January 01 2020

FAIR Computational Workflows

[Carole Goble](#), [Sarah Cohen-Boulakia](#), [Stian Soiland-Reyes](#), [Daniel Garijo](#), [Yolanda Gil](#), [Michael R. Crusoe](#), [Kristian Peters](#), [Daniel Schober](#)

> Author and Article Information

Data Intelligence (2020) 2 (1-2): 108–121.

https://doi.org/10.1162/dint_a_00033

Introducing the FAIR Principles for research software

[Michelle Barker](#), [Neil P. Chue Hong](#), [Daniel S. Katz](#), [Anna-Lena Lamprecht](#), [Carlos Martínez-Ortiz](#), [Fotis Psomopoulos](#), [Jennifer Harrow](#), [Leyla Jael Castro](#), [Morane Gruenpeter](#), [Paula Andrea Martínez](#) & [Tom Honeyman](#)

Scientific Data **9**, Article number: 622 (2022) | [Cite this article](#)

Breakout 7 Data Infrastructures - Organisa... The FAIR Agenda WGs Getting started

WG FAIR for Virtual Research Environments: FAIR for VREs - The Path Forward

7:30 AM - 9:00 AM

Room E

DOI: [10.15497/RDA00065](https://doi.org/10.15497/RDA00065)

Citation and download: Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martínez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martínez, P. A., Honeyman, T., et al. (2021). FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*. DOI: [10.15497/RDA00065](https://doi.org/10.15497/RDA00065)



SOFTWARE IS NOT JUST ANOTHER TYPE OF DATA

- FAIR Principles, are intended to be applied to all digital objects (Wilkinson et al. 2016)
- Efforts to adapt and adopt the FAIR principles to research software (RDA FAIR4RS)

Recommendation n° 5 :

Recognise that FAIR guidelines will require translation for other digital objects and support such efforts.

2020: 'Six Recommendations for Implementation of FAIR Practice'

([FAIR Practice Task Force EOOSC, 2020](#))

WHAT IS RESEARCH SOFTWARE?

Research software - Source code files, algorithms, scripts, computational workflows and executables that were created in either of two categories:

- A. Within a research project as a by-product to do the research, or*
- B. Through intentional development of a software product for general use in research by one or more projects.*

[Gruepenter et al. 2021](https://doi.org/10.5281/zenodo.5504016) Defining Research Software: a controversial discussion (Version 1). Zenodo. <https://doi.org/10.5281/zenodo.5504016>

Not all software used in research is research software

FAIR4RS PRINCIPLES



- **Findable:** Software, and its associated metadata, is easy for both humans and machines to find.
- **Accessible:** Software, and its metadata, is retrievable via standardized protocols.
- **Interoperable:** *Software interoperates with other software by exchanging data and/or metadata, and/or through interaction via application programming interfaces (APIs), described through standards.*
- **Reusable:** *Software is both usable (can be executed) and reusable (can be understood, modified, built upon, or incorporated into other software).*

(key differences from FAIR data principles in *italics*)

Output of [the FAIR principles for research software](#) (FAIR4S) - joint Research Software Alliance (**ReSA**), Research Data Alliance (**RDA**), **FORCE11** Working Group/Task force

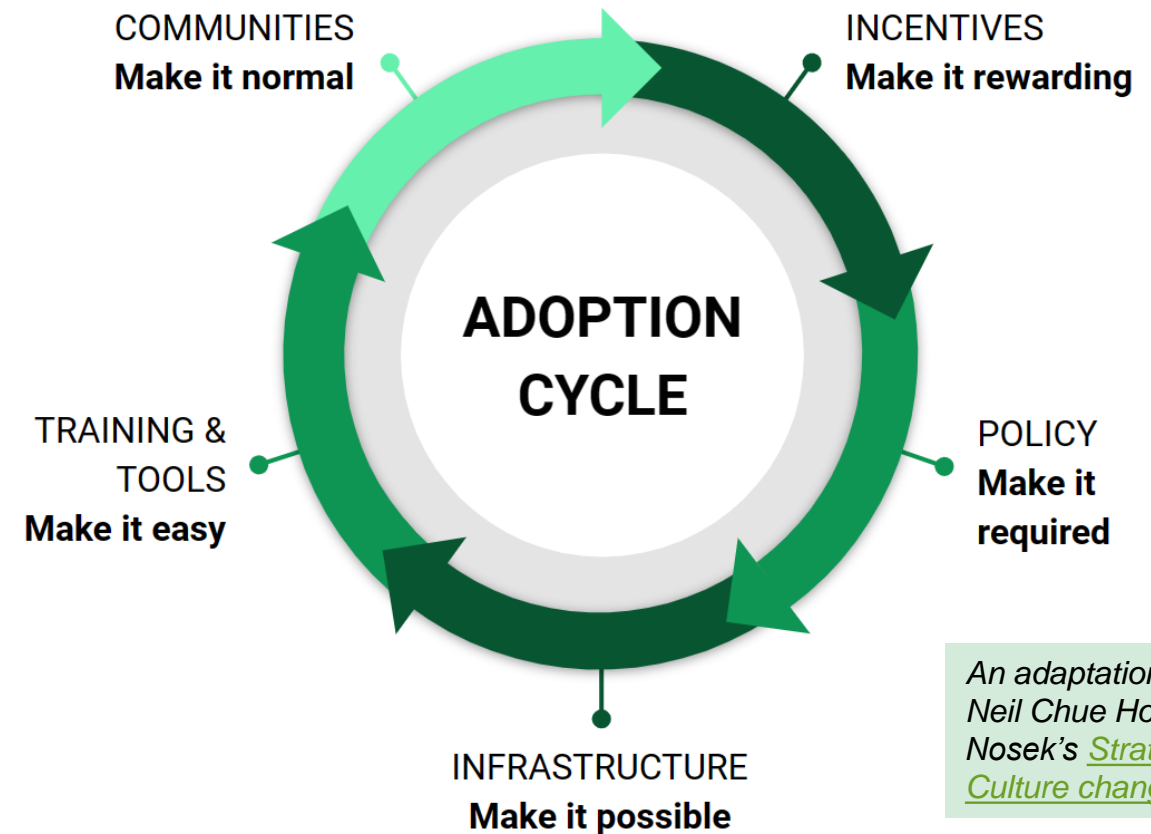
Slide adapted from the [presentation](#) of the [RDA FAIR4RS steering group](#) at the International Funders Workshop (Nov 2022), <https://zenodo.org/doi/10.5281/zenodo.7350198>

WHO IS RESPONSIBLE FOR FAIR SOFTWARE?

Who is expected to apply FAIR?

- *The application of the FAIR4RS Principles is the responsibility of the owners (who are often the creators) of the software, not the users.*
- *The FAIR4RS Principles are also relevant to, and require support from, the larger ecosystem and various stakeholders that support research software (e.g., repositories and registries).*

Slide adapted from the [presentation](https://zenodo.org/doi/10.5281/zenodo.7350198) of the [RDA FAIR4RS steering group](#) at the International Funders Workshop (Nov 2022), <https://zenodo.org/doi/10.5281/zenodo.7350198>

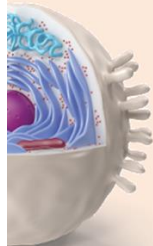


An adaptation by
Neil Chue Hong of
Nosek's [Strategy for
Culture change](#)

MOVING BEYOND SOFTWARE?

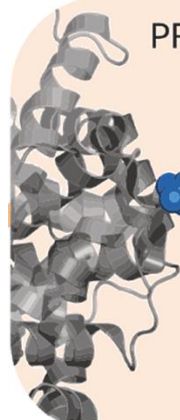
Machine learning is being used across many biological areas

CELL



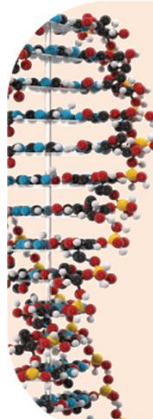
- Biotherapeutic cell culture optimisation
- Cellular Image Analysis
- Cell type Annotation

PROTEOME



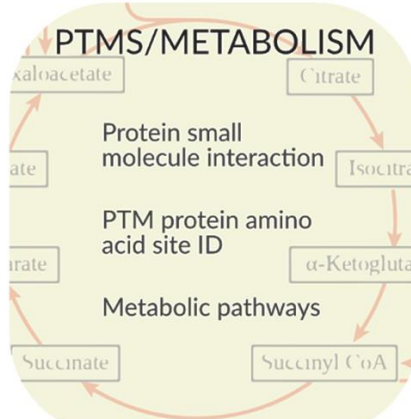
- Protein Structure
- Protein-protein interactions
- Binding site ID

GENOME



- ID gene coding regions
- Pharmacogenomics
- CRISPR Target sequence ID
- Methylation side ID
- Gene expression (microarrays)
- RNA structure

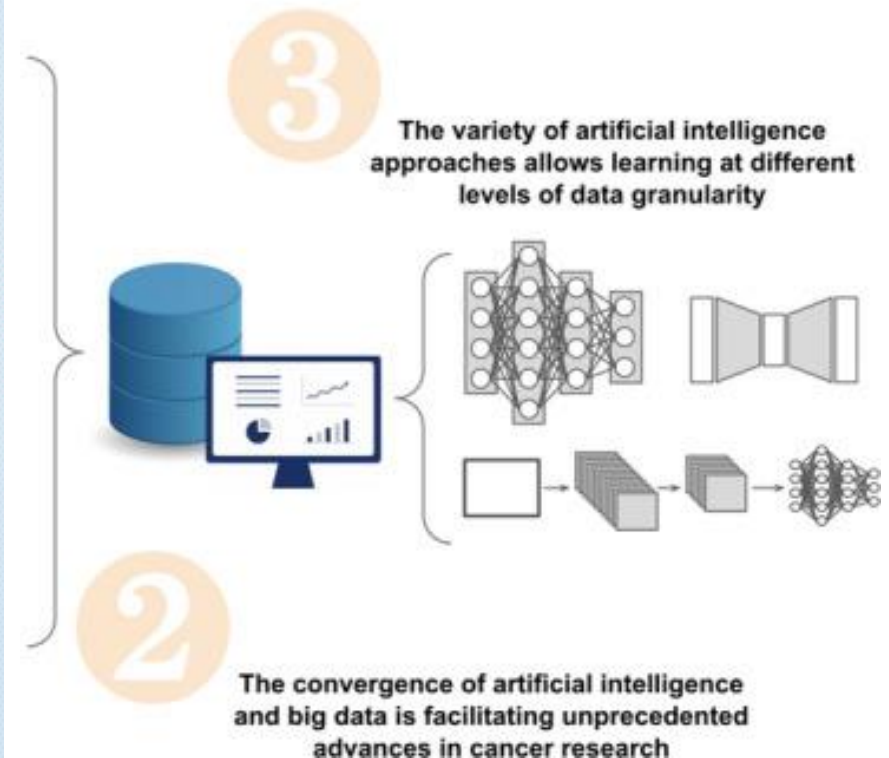
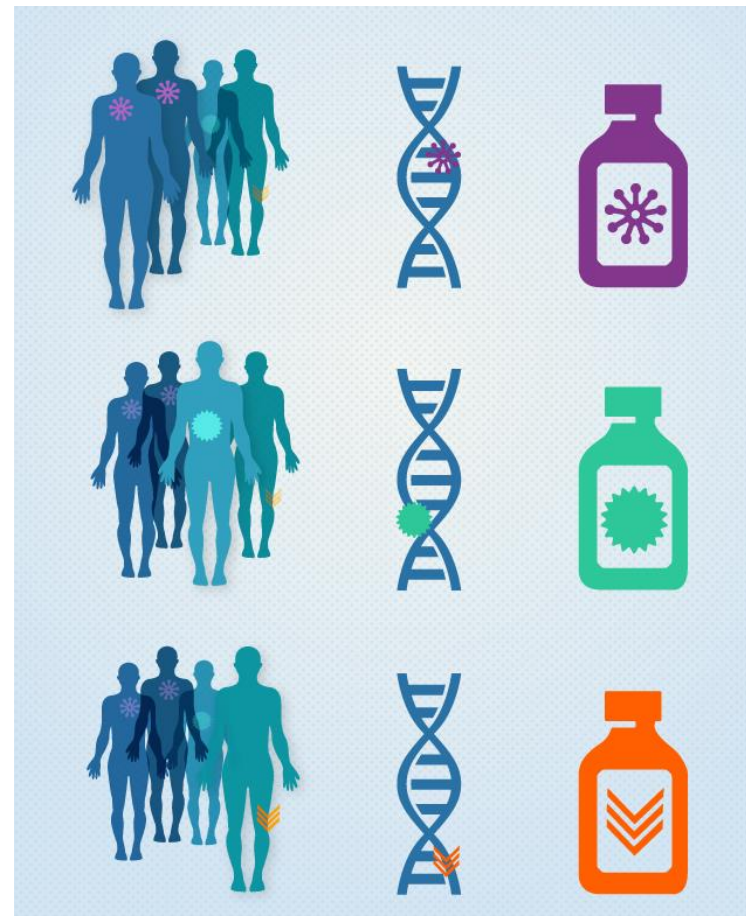
PTMS/METABOLISM



Protein small molecule interaction

PTM protein amino acid site ID

Metabolic pathways



The integration of artificial intelligence (AI) and machine learning approaches within life sciences is making drug discovery and development more innovative, time-effective, and cost-effective.





NEW SET OF CHALLENGES

SCIENCE FORUM

Ten common statistical mistakes to watch out for when writing or reviewing a manuscript

Abstract Inspired by broader efforts to make the conclusions of scientific research more robust, we have compiled a list of some of the most common statistical mistakes that appear in the scientific literature. The mistakes have their origins in ineffective experimental designs, inappropriate analyses and/or flawed reasoning. We provide advice on how authors, reviewers and readers can identify and resolve these mistakes and, we hope, avoid them in the future.

TAMAR R MAKIN* AND JEAN-JACQUES ORBAN DE XIVRY
Makin and Orban de Xivry. *eLife* 2019;8:e48175. DOI: <https://doi.org/10.7554/eLife.48175>

Briefings in Bioinformatics, 17(5), 2016, 831–840

doi: 10.1093/bib/bbv082
Advance Access Publication Date: 26 September 2015
Paper

Correct machine learning on protein sequences: a peer-reviewing perspective

Ian Walsh, Gianluca Pollastri and Silvio C. E. Tosatto

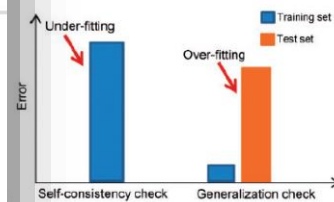
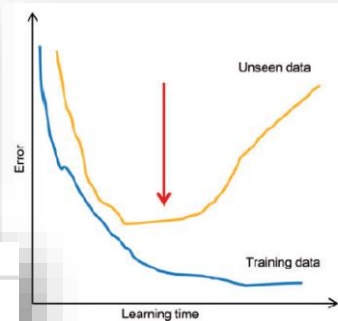
Corresponding author: Silvio C. E. Tosatto, Dept. of Biomedical Sciences, University of Padua, viale G. Colombo 3, 35131 Padova, Italy. Tel.: +39 049 827 6269; Fax: +39 049 827 6260; E-mail: silvio.tosatto@unipd.it

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

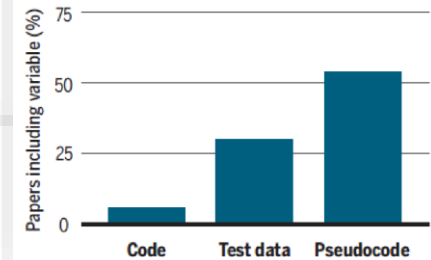
Unpublished code and sensitivity to make many claims hard to verify

By **Matthew Hutson** SCIENCE [sciencemag.org](https://www.sciencemag.org) 16 FEBRUARY 2019



Code break

In a survey of 400 artificial intelligence papers presented at major conferences, just 6% included code for the papers' algorithms. Some 30% included test data, whereas 54% included pseudocode, a limited summary of an algorithm.



nature
machine intelligence

PERSPECTIVE

<https://doi.org/10.1038/s42256-019-0139-8>

Validity of machine learning in biology and medicine increased through collaborations across fields of expertise

Maria Littmann^{1,27*}, Katharina Selig^{2,27*}, Liel Cohen-Lavi^{3,4}, Yotam Frank⁵, Peter Höningschmid⁶, Evans Kataka⁶, Anja Mösch⁶, Kun Qian^{7,8}, Avihai Ron^{9,10}, Sebastian Schmid¹¹, Adam Sorbie¹², Liran Szlak¹³, Ayana Dagan-Wiener¹⁴, Nir Ben-Tal¹⁵, Masha Y. Niv^{14,16}, Daniel Razansky^{9,10,17,18,19,20}, Björn W. Schuller²¹, Donna Ankerst², Tomer Hertz^{3,22,23} and Burkhard Rost^{1,24,25,26}

Setting the standards for machine learning in biology

David T. Jones^{1,2}

Machine learning is a branch of artificial intelligence (AI) involving computer programs that are able to improve their own performance through experience (training). The diverse applications of new 'deep learning' approaches with neural networks are now expanding into the field of biology. But these applications to biological data require more scrutiny and caution to increase the standards of publishing and allow the AI revolution in biology to take off.

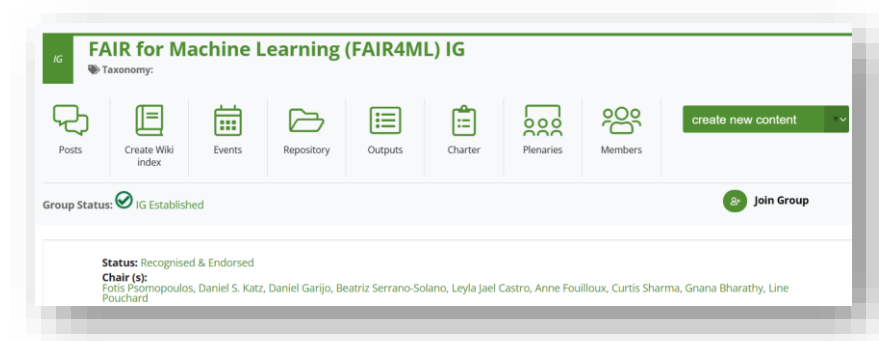
<https://doi.org/10.1038/s41580-019-0176-5>

NATURE REVIEWS | MOLECULAR CELL BIOLOGY



FAIR IN MACHINE LEARNING (MODELS)?

- What does FAIR apply to?
 - Are they data?
 - E.g., a set of parameters and options for a particular framework
 - Are they software?
 - E.g., an executable object that takes input and provides output
 - Are they something else?
- How does FAIR apply?
 - Searched and shared via repositories?
 - Searched and shared via executable platforms?
 - Searched and shared via something else? (e.g., DLHub, OpenML, HuggingFace...)
 - Models and training data are linked - should they be shared together?




Slide adapted from various presentations of the [RDA FAIR4ML](#) interest group during Plenary events



NEED FOR COMMUNITY-LED STANDARDS (1/2)

ML Commons

mlcommons/
croissant



Croissant is a high-level format for machine learning datasets that brings together four rich layers.

16 Contributors | 58 Issues | 12 Discussions | 45 Stars | 8 Forks



the DOME Recommendations



Hugging Face

ONNX

DOME: recommendations for supervised machine learning validation in biology

Ian Walsh, Dmytro Fishman, Dario Garcia-Gasulla, Tiina Titma, Gianluca Pollastri, ELIXIR Machine Learning Focus Group, Jennifer Harrow, Fotis E. Psomopoulos & Silvio C. E. Tosatto

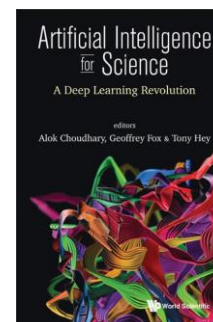
Nature Methods (2021) | Cite this article

4927 Accesses | 73 Altmetric | Metrics



DOME Registry

A database of annotations for published papers describing machine learning methods in biology.



© 2023 World Scientific Publishing Company
https://doi.org/10.1142/9789811265679_0022

Chapter 22

A Roadmap for Defining Machine Learning Standards in Life Sciences

Fotis Psomopoulos^{*||}, Carole Goble^{1,***},
Leyla Jael Castro^{1,††}, Jennifer Harrow^{3,‡‡},
and Silvio C. E. Tosatto^{4,§§}



(GIGA)ⁿ SCIENCE

DOME adopted as part of the **submission system** for **GigaScience**

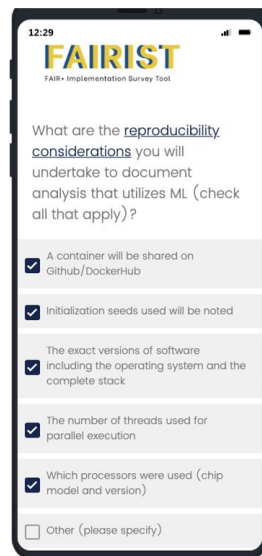
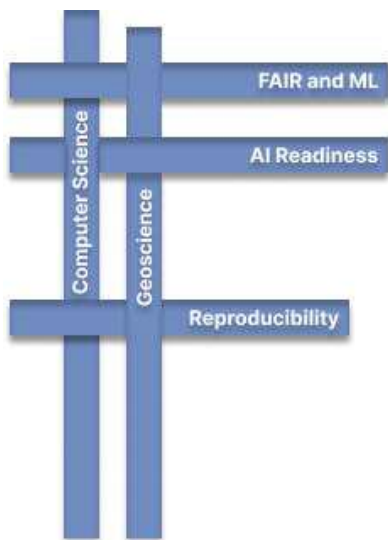
(see example here: <http://gigadb.org/dataset/102404>)

Online registry of annotated papers: <https://registry.dome-ml.org>



NEED FOR COMMUNITY-LED STANDARDS (2/2)

FARR: FAIR in ML, AI Readiness, & Reproducibility Research Coordination Network



Ways to Get Involved

- **Input** on community needs, gaps & roadmap
- **Suggest use cases** and let us promote your project's use of AI and FARR-related practices
- Let us feature you in a **science story**

FARR
FAIR, AI Readiness & Reproducibility

Using FAIR to foster AI-readiness in Data Facilities:
A resource list

This work is supported through NSF award # 2226453.

What is FAIR?

- **A refresher on FAIR:** More than an acronym, it stands for 15 principles for making research objects more Findable, Accessible, Interoperable, Reusable
<https://www.go-fair.org/fair-principles/>
- **Suggestions on how to implement FAIR:**
<https://bit.ly/implementFAIR>

Data repositories supporting AI with FAIR practices

- **The geosciences:**
<https://www.hydroshare.org/>
- **High energy physics:**
<https://bit.ly/AI-readyHEP>
- **Materials science:** <https://bit.ly/MLinMS>



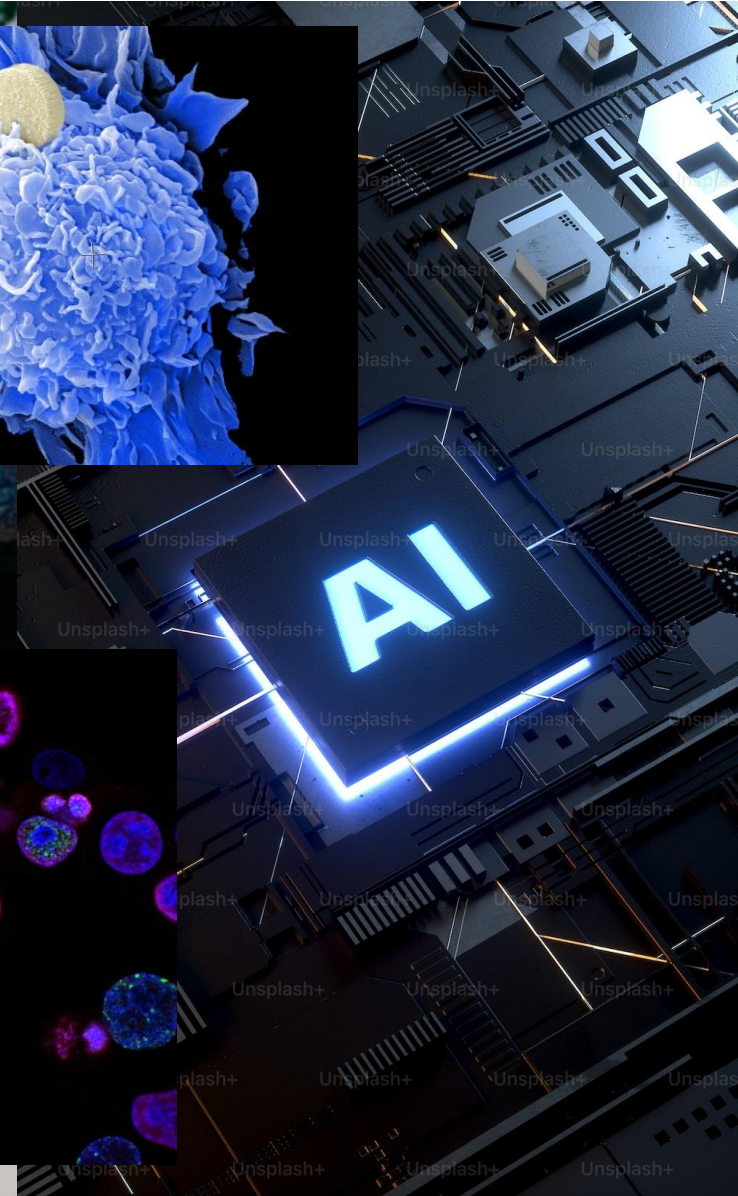
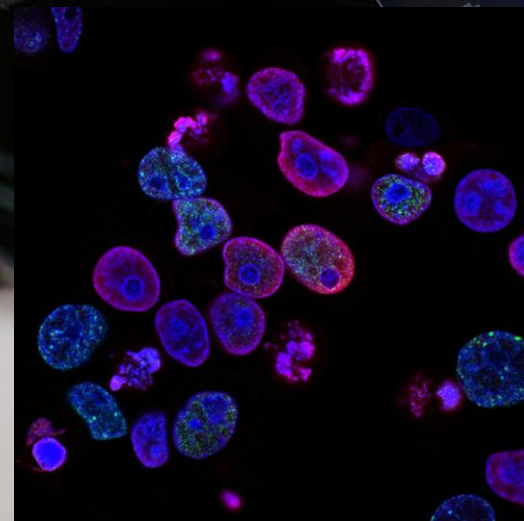
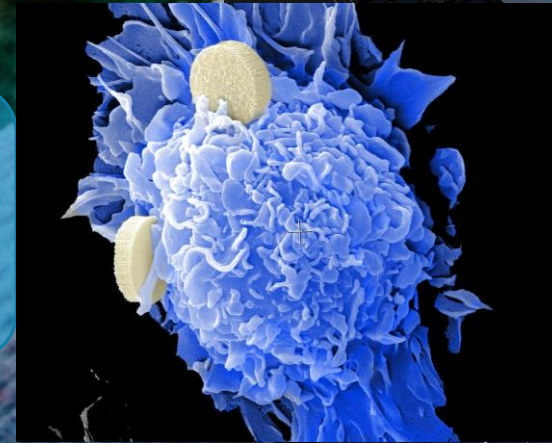
Contact:

<https://www.farr-rcn.org/>
community@farr-rcn.org

This work is supported through the NSF award #2226453.



*If you want to go fast, go alone
If you want to go far, go together*



THANK YOU!

MERCI!

GRAZIE!

GRACIAS!

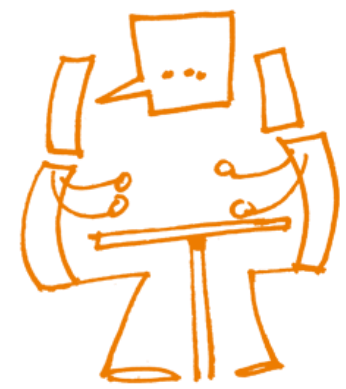
DANK JE WEL!



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS



INSTITUTE OF APPLIED BIOSCIENCES
ΙΝΣΤΙΤΟΥΤΟ ΕΦΑΡΜΟΣΜΕΝΩΝ ΒΙΟΕΠΙΣΤΗΜΩΝ
CENTRE for RESEARCH and TECHNOLOGY-HELLAS



@fopsom@genomic.social



@fopsom