FAIR for Machine Learning; Building on the Lessons from FAIR Software

Fotis E. Psomopoulos

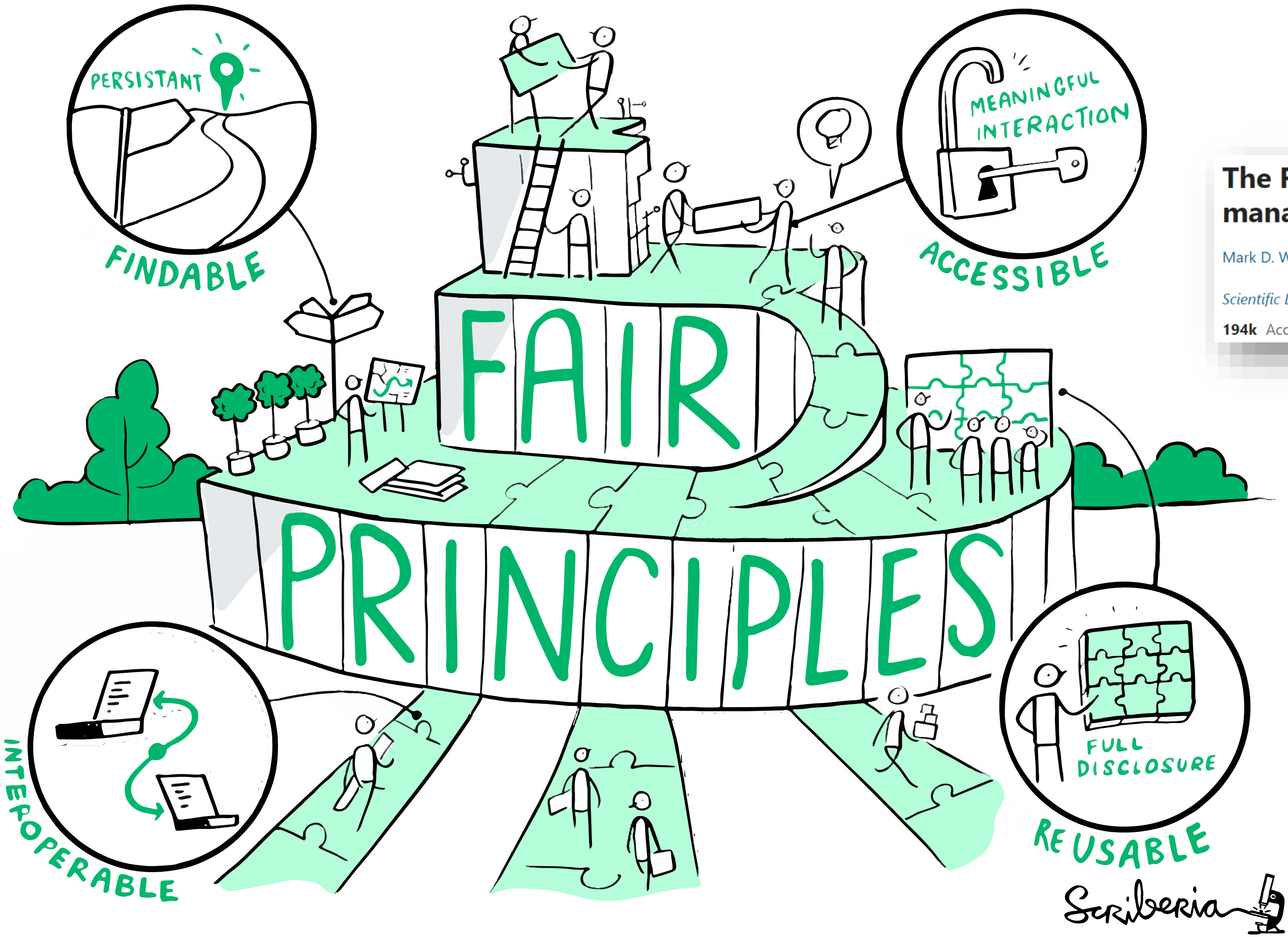
Institute of Applied Biosciences, CERTH, Greece

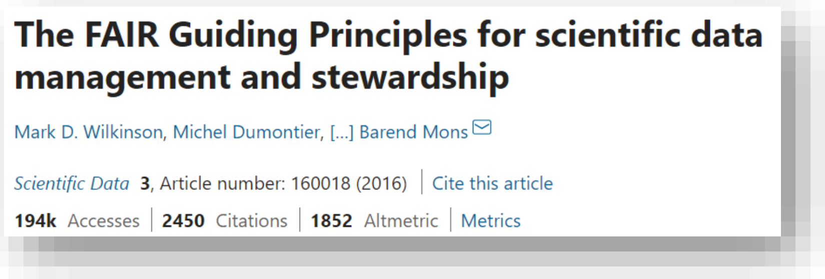
RDA
RESEARCH DATA ALLIANCE

elixir

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

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





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

The Turing Way Community, & Scriberia. (2024). Illustrations from The Turing Way
https://doi.org/10.5281/zenodo.10556824

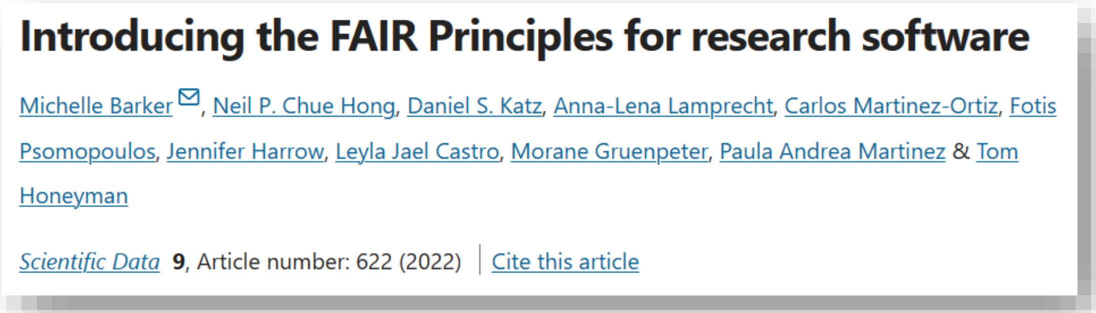
FAIR for non-data objects: some context

- 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

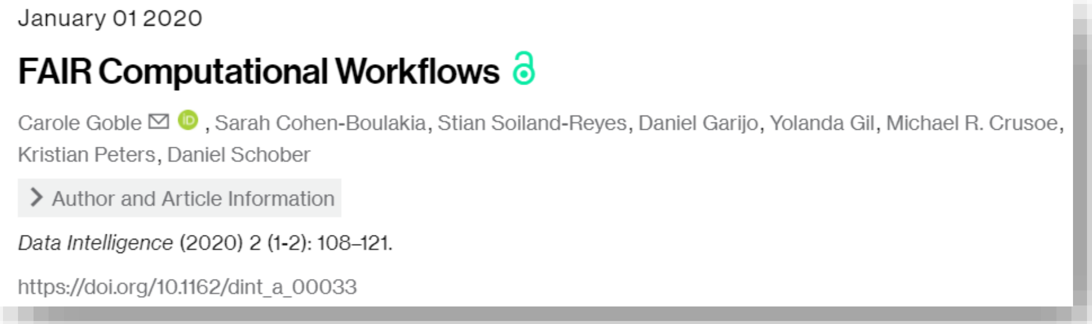
FAIR for non-data objects: an ongoing effort



Introducing the FAIR Principles for research software

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

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



January 01 2020

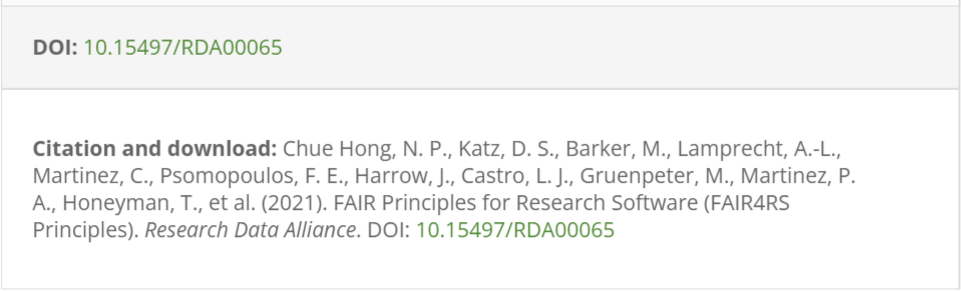
FAIR Computational Workflows

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

> Author and Article Information

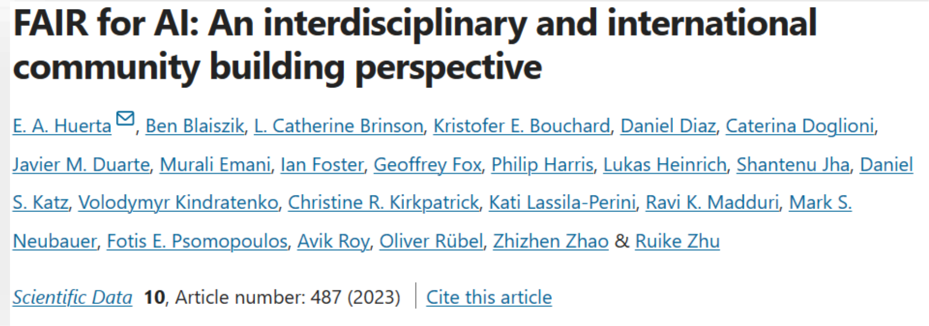
Data Intelligence (2020) 2 (1-2): 108–121.

https://doi.org/10.1162/dint_a_00033



DOI: 10.15497/RDA00065

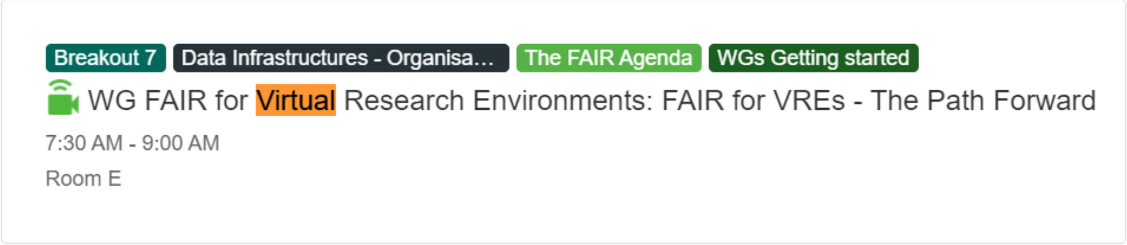
Citation and download: Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., Honeyman, T., et al. (2021). FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*. DOI: 10.15497/RDA00065



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 Rübel, Zhizhen Zhao & Ruike Zhu

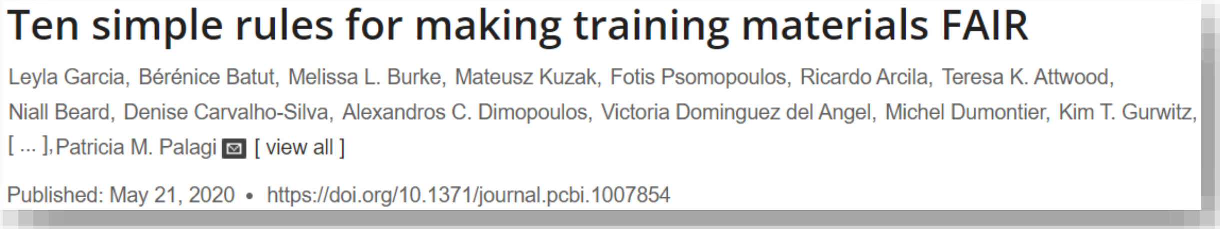
Scientific Data **10**, Article number: 487 (2023) | Cite this article



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

7:30 AM - 9:00 AM

Room E



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

On the road to Define FAIR for Research Software

RDA
RESEARCH DATA ALLIANCE

- Efforts to adapt and adopt the FAIR principles to research software (RDA FAIR4RS)

Recommendation n°2 :

*Make sure **<u>the specific nature of software</u>** is recognized and not considered as "just data" particularly in the context of discussion about the notion of FAIR data.*

***2019:** the **Opportunity Note** by the French national Committee for Open Science's Free Software and Open Source Project Group*

(Clément-Fontaine, 2019)

Recommendation n°5 :

*Recognise that FAIR guidelines will require **<u>translation for other digital objects</u>** and support such efforts.*

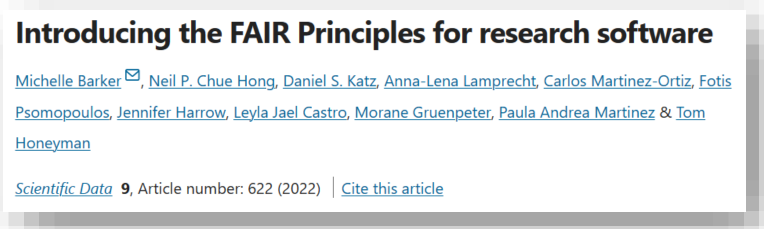
2020: 'Six Recommendations for Implementation of FAIR Practice'

(FAIR Practice Task Force EOSC, 2020)

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*)



Introducing the FAIR Principles for research software

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

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

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

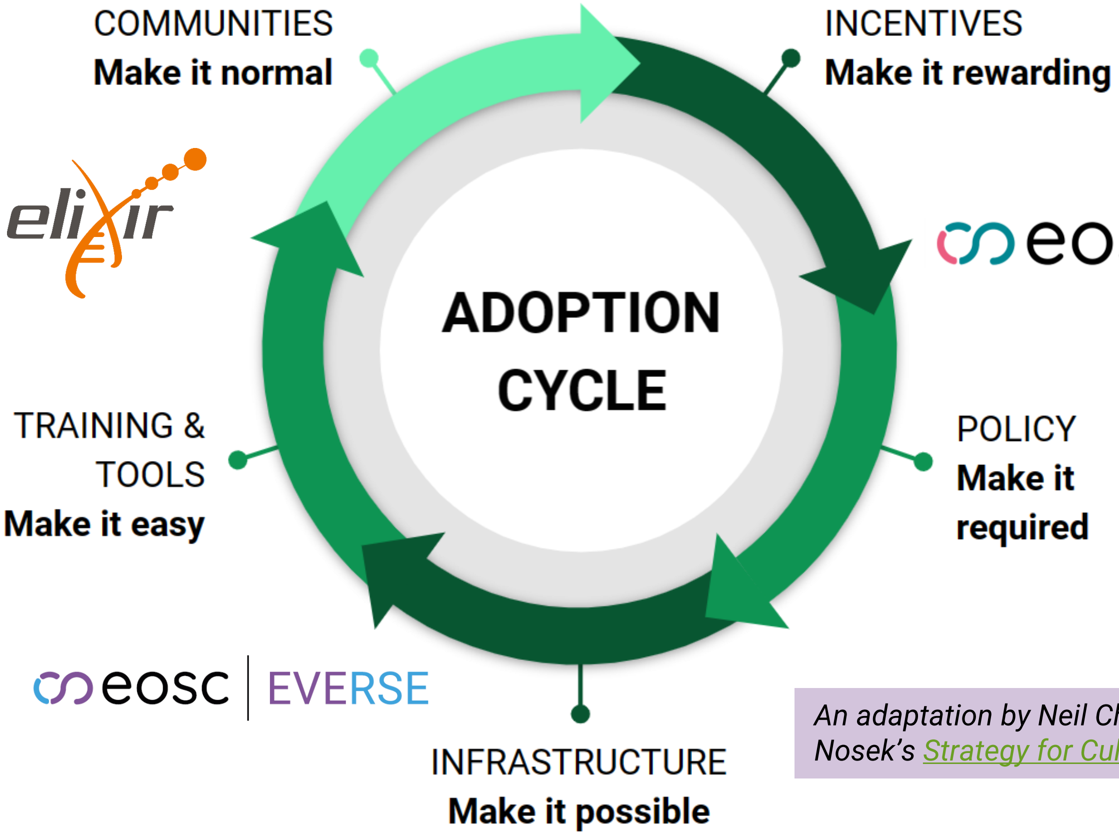
Slide adapted from the presentation of the RDA FAIR4RS steering group at the International Funders Workshop (Nov 2022), https://zenodo.org/doi/10.5281/zenodo.7350198

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).*

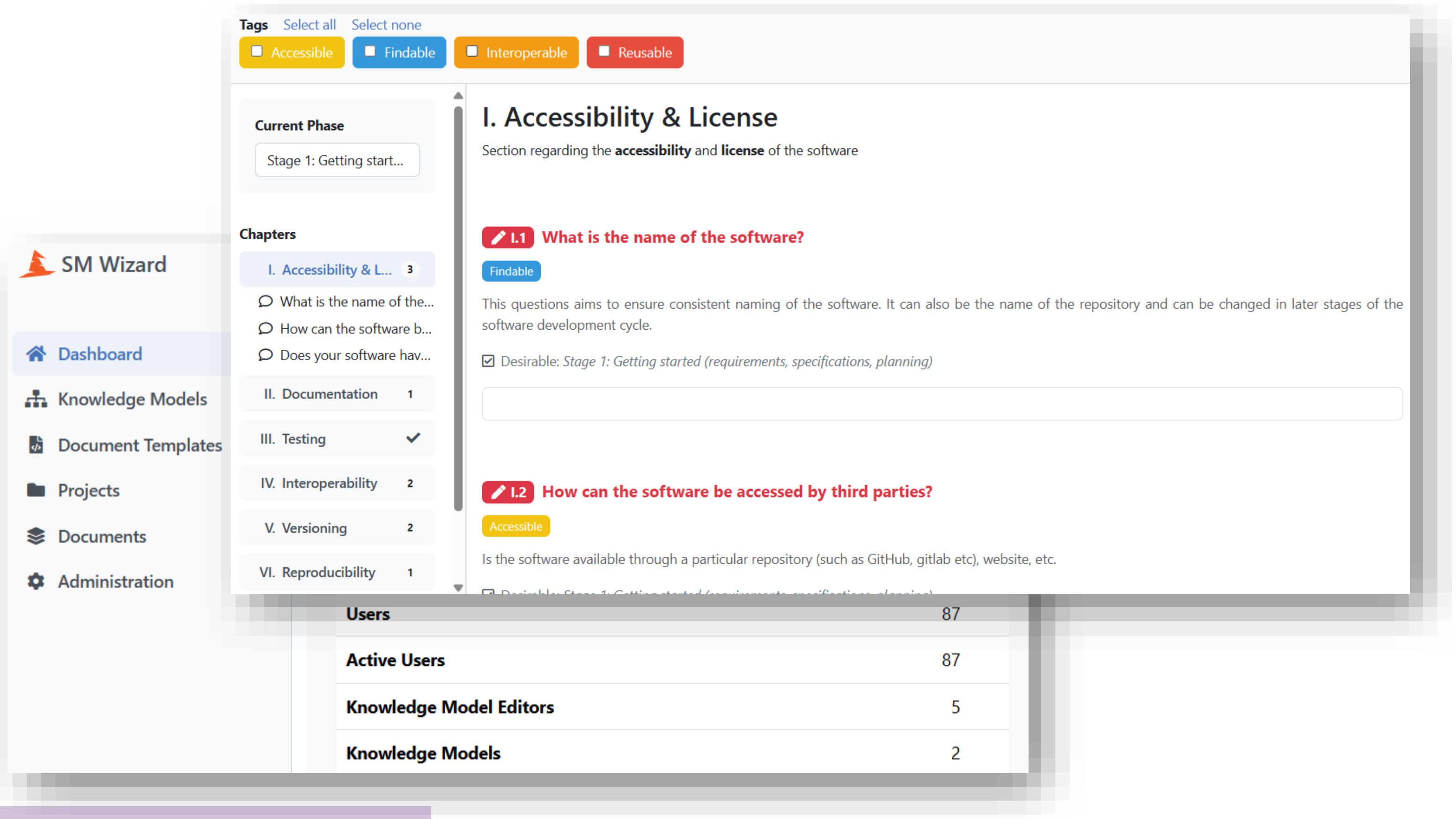
Slide adapted from the presentation of the RDA FAIR4RS steering group at the International Funders Workshop (Nov 2022), https://zenodo.org/doi/10.5281/zenodo.7350198



An adaptation by Neil Chue Hong of Nosek's Strategy for Culture change

Managing (FAIR) Software

- helps to **implement best practices** during software development
- ensures that software is **accessible** and **reusable** in the short and longer term
- contributes to the **reproducibility** of results
- stimulates **collaborative** work on open-source software for research.



Tags Select all Select none

□ Accessible □ Findable □ Interoperable □ Reusable

Current Phase

Stage 1: Getting start...

Chapters

I. Accessibility & L... 3

What is the name of the...

How can the software b...

Does your software hav...

II. Documentation 1

III. Testing ✓

IV. Interoperability 2

V. Versioning 2

VI. Reproducibility 1

I. Accessibility & License

Section regarding the **accessibility** and **license** of the software

I.1 What is the name of the software?

Findable

This questions aims to ensure consistent naming of the software. It can also be the name of the repository and can be changed in later stages of the software development cycle.

☑ Desirable: *Stage 1: Getting started (requirements, specifications, planning)*

I.2 How can the software be accessed by third parties?

Accessible

Is the software available through a particular repository (such as GitHub, gitlab etc), website, etc.

Users	87
Active Users	87
Knowledge Model Editors	5
Knowledge Models	2

Martinez-Ortiz, C. et al. (2022). Practical guide to Software Management Plans (1.0). https://doi.org/10.5281/zenodo.7248877

CERTH CENTRE FOR RESEARCH & TECHNOLOGY HELLAS

elixir

European Commission

IEEE Advancing Technology for Humanity

EUROPEAN OPEN SCIENCE CLOUD

eosc

SCIENCE FOR AFRICA FOUNDATION

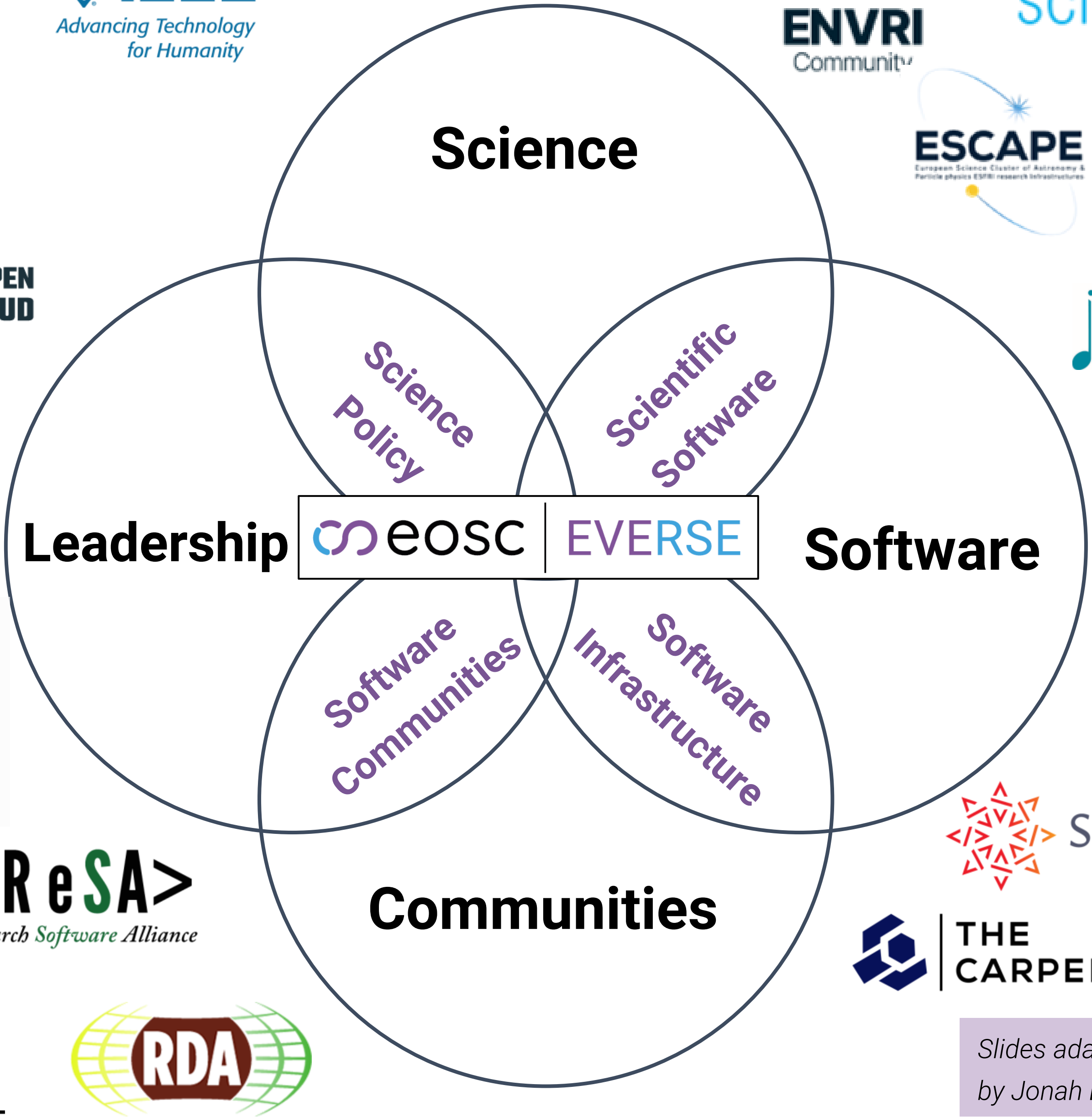
RSE

Software Sustainability Institute

<ReSA> Research Software Alliance

ORGANIZATIONAL MYCOLOGY

RDA RESEARCH DATA ALLIANCE



ENVRI Community

LIFE SCIENCE RI

panosc photon and neutron open science cloud

ESCAPE

SSHOpenCluster Social Sciences & Humanities Resources

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

julia

PyTorch

NUMFOCUS OPEN CODE = BETTER SCIENCE

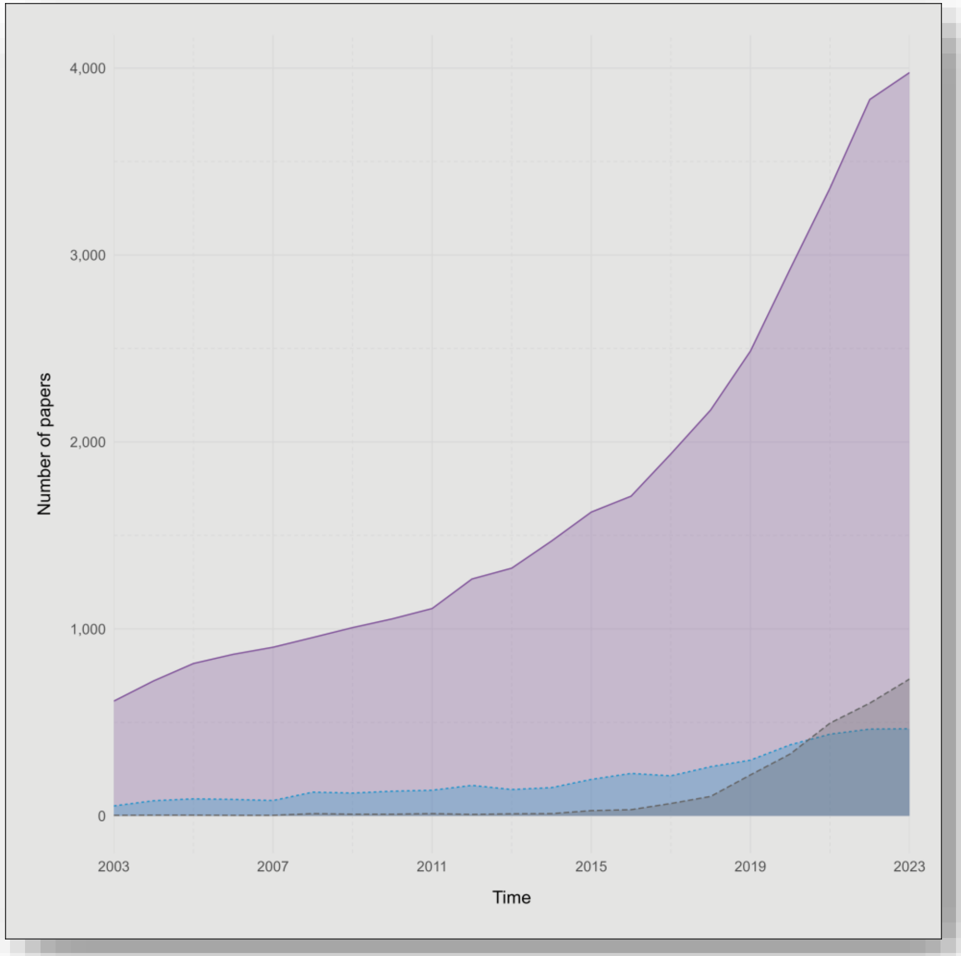
Software Heritage

THE CARPENTRIES

Open Life Science

Slides adapted from the "OrgMycology - eResearch NZ 2024" by Jonah Duckles (orgmycology)

How does FAIR fare (for Data / Software / ML)



Significant effort and push towards **FAIR data** (AI-ready datasets being a key demand)

Software: is only just beginning to get the support it needs as a first-class citizen in science

NIH National Institutes of Health *Office of Data Science Strategy*

NIH STRATEGIC PLAN FOR DATA SCIENCE 2023-2028

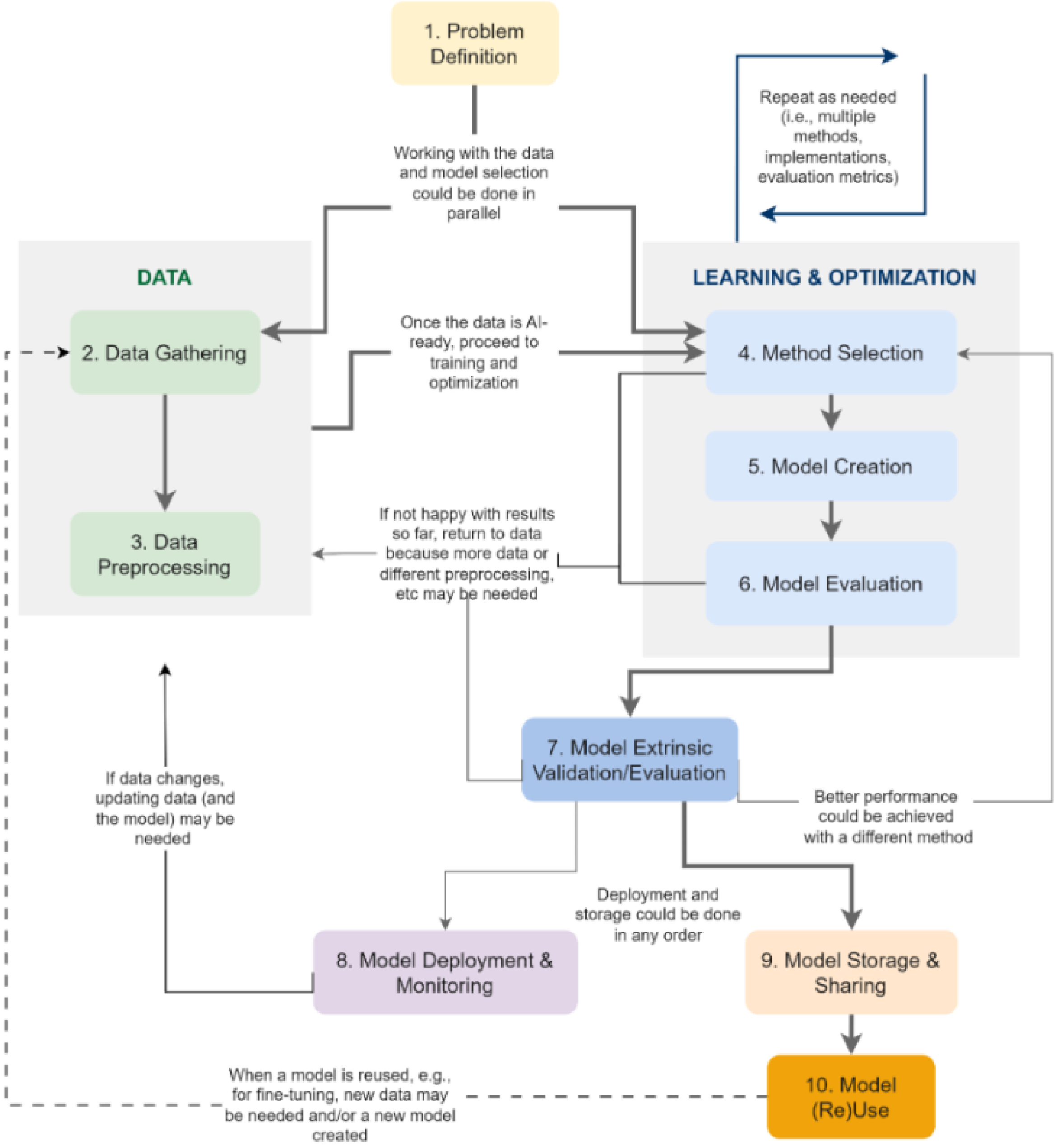
elixir STEER eosc | EVERSE

ML/AI: community just started realizing the challenges

EU Artificial Intelligence Act

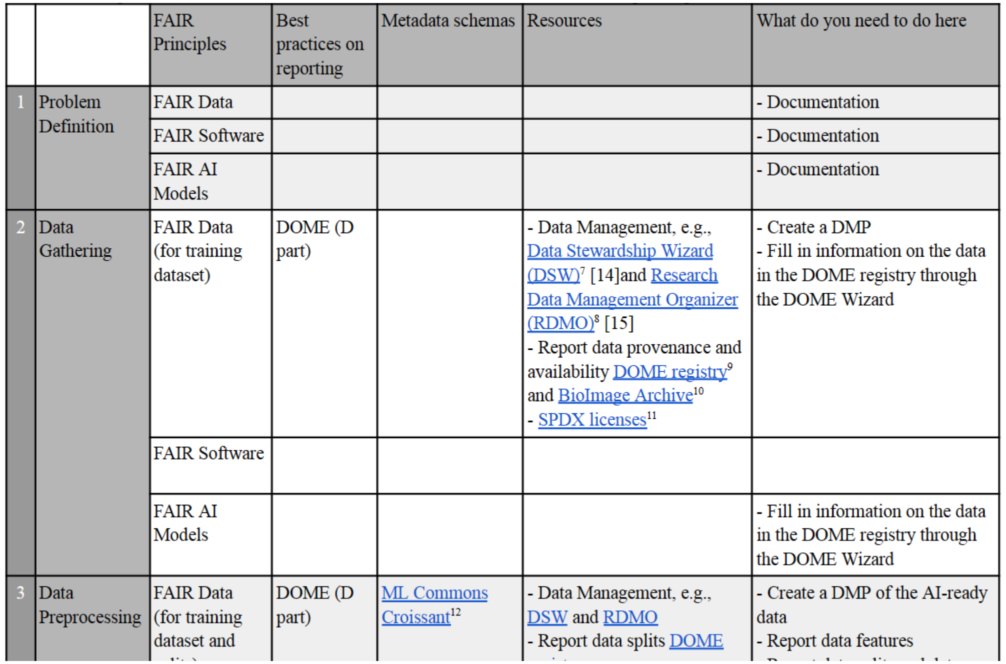
SAPEA Science Advice for Policy by European Academies

elixir


1. Problem Definition
Working with the data and model selection could be done in parallel
Repeat as needed (i.e., multiple methods, implementations, evaluation metrics)
DATA
2. Data Gathering
3. Data Preprocessing
LEARNING & OPTIMIZATION
4. Method Selection
5. Model Creation
6. Model Evaluation
Once the data is AI-ready, proceed to training and optimization
If not happy with results so far, return to data because more data or different preprocessing, etc may be needed
7. Model Extrinsic Validation/Evaluation
Better performance could be achieved with a different method
If data changes, updating data (and the model) may be needed
Deployment and storage could be done in any order
8. Model Deployment & Monitoring
9. Model Storage & Sharing
10. Model (Re)Use
When a model is reused, e.g., for fine-tuning, new data may be needed and/or a new model created

FAIR in the ML Lifecycle

Different FAIR principles apply to different aspects of the Life Cycle



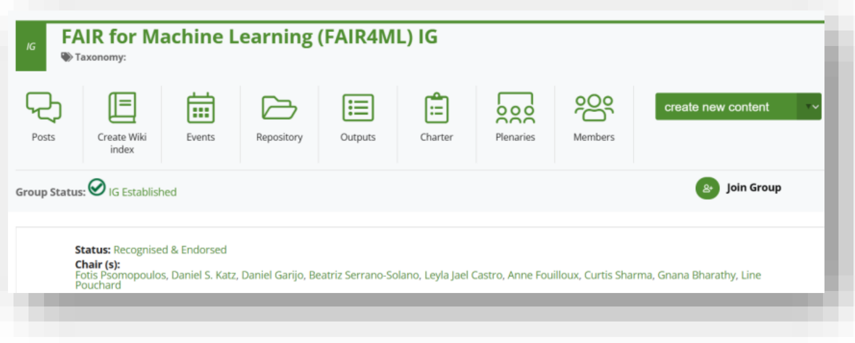
		FAIR Principles	Best practices on reporting	Metadata schemas	Resources	What do you need to do here
1	Problem Definition	FAIR Data				- Documentation
		FAIR Software				- Documentation
		FAIR AI Models				- Documentation
2	Data Gathering	FAIR Data (for training dataset)	DOME (D part)		- Data Management, e.g., Data Stewardship Wizard (DSW)[7] [14]and Research Data Management Organizer (RDMO)[8] [15] - Report data provenance and availability DOME registry[9] and BioImage Archive[10] - SPDX licenses[11]	- Create a DMP - Fill in information on the data in the DOME registry through the DOME Wizard
		FAIR Software				
		FAIR AI Models				- Fill in information on the data in the DOME registry through the DOME Wizard
3	Data Preprocessing	FAIR Data (for training dataset and	DOME (D part)	ML Commons Croissant[12]	- Data Management, e.g., DSW and RDMO - Report data splits DOME	- Create a DMP of the AI-ready data - Report data features

Castro, L. J, et.al., & Zhang, Y. (2023). Lifecycle for FAIR Machine Learning.. https://doi.org/10.5281/zenodo.10407265

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?

RESEARCH DATA ALLIANCE



FAIR for Machine Learning (FAIR4ML) IG

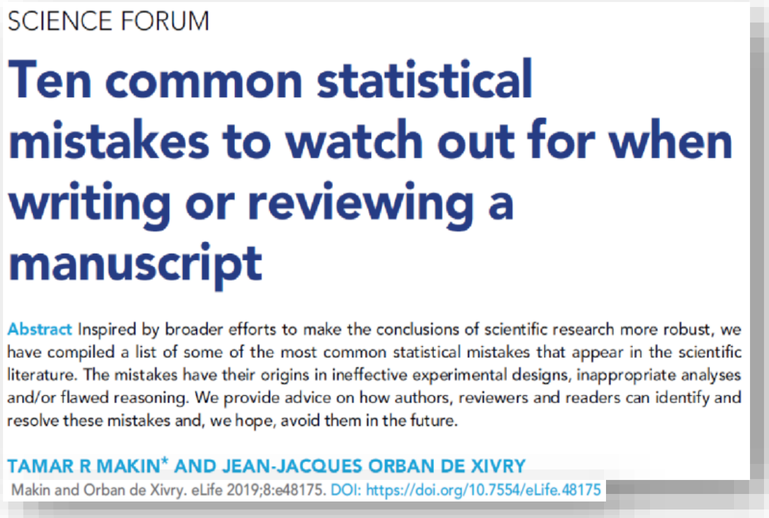
Taxonomy:

Group Status: IG Established

Status: Recognised & Endorsed
Chair (s):
Fotis Psomopoulos, Daniel S. Katz, Daniel Garijo, Beatriz Serrano-Solano, Leyla Jael Castro, Anne Fouilloux, Curtis Sharma, Gnana Bharathy, Line Pouchard

Slide adapted from various presentations of the RDA FAIR4ML interest group during Plenary events

New set of challenges



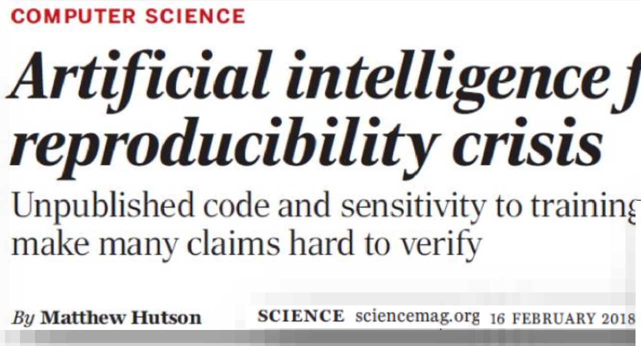
SCIENCE FORUM

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



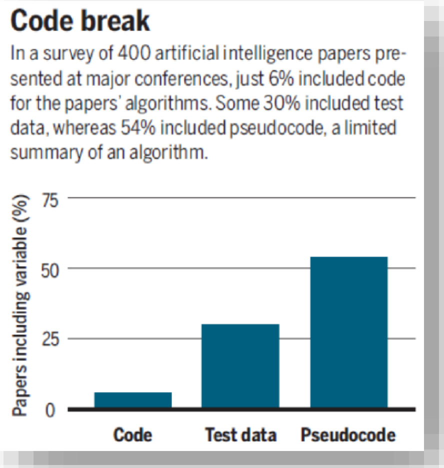
COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

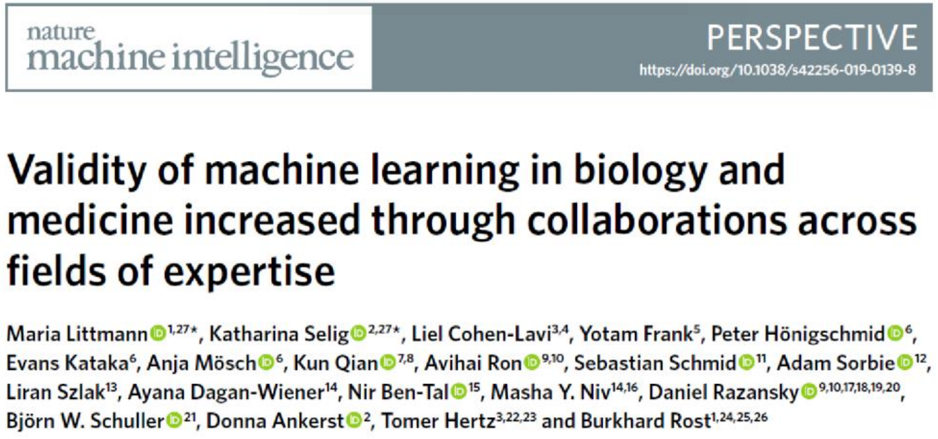
By Matthew Hutson

SCIENCE sciencemag.org 16 FEBRUARY 2018



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.



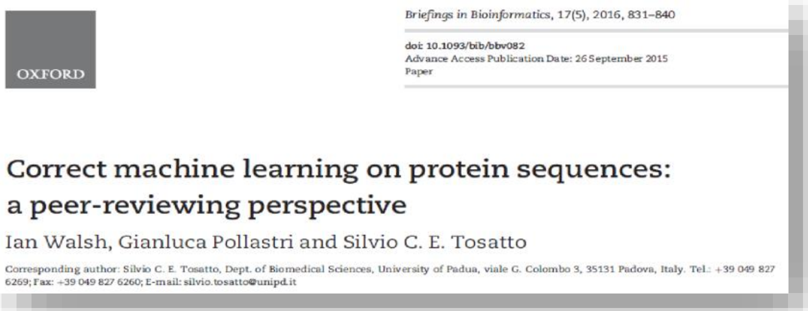
nature machine intelligence

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 Frank[5], Peter Hönigschmid[6], Evans Kataka[6], Anja Mösch[6], Kun Qian[7,8], Avihai Ron[9,10], Sebastian Schmid[11], Adam Sorbie[12], Liran Szlak[13], Ayana Dagan-Wiener[14], Nir Ben-Tal[15], Masha Y. Niv[14,16], Daniel Razansky[9,10,17,18,19,20], Björn W. Schuller[21], Donna Ankerst[2], Tomer Hertz[3,22,23] and Burkhard Rost[1,24,25,26]



OXFORD

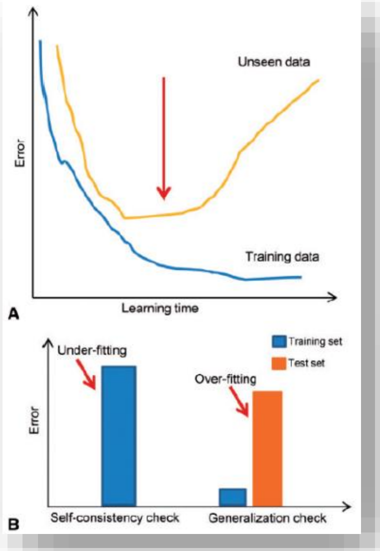
Briefings in Bioinformatics, 17(5), 2016, 831–840

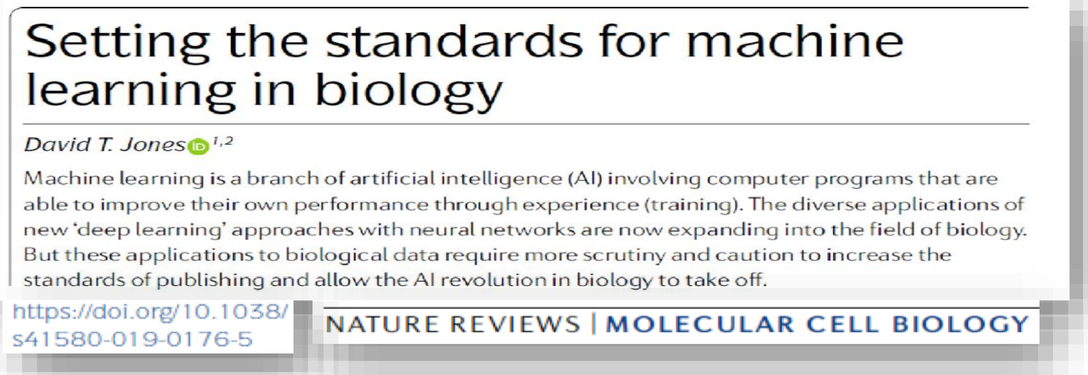
doi: 10.1093/bib/bbv082
Advance Access Publication Date: 26 September 2015
Paper

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





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.

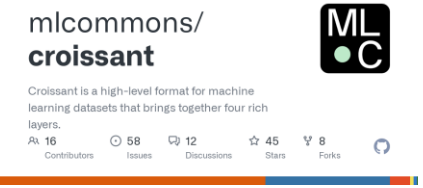
https://doi.org/10.1038/s41580-019-0176-5

NATURE REVIEWS | MOLECULAR CELL BIOLOGY

15 April 2024

Need for Community-led Standards and Best Practices (1/2)

ML Commons



mlcommons/
croissant

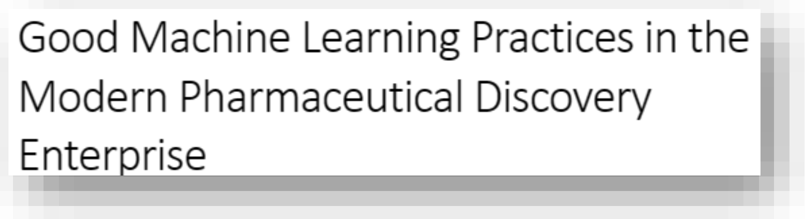
Croissant is a high-level format for machine learning datasets that brings together four rich layers.

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

ONNX

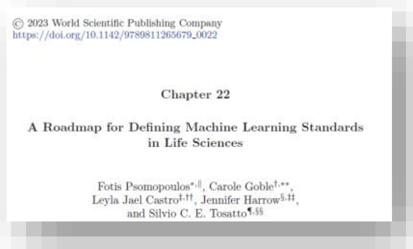
Hugging Face

Pistoia Alliance



Good Machine Learning Practices in the Modern Pharmaceutical Discovery Enterprise

https://doi.org/10.31219/osf.io/kuz8p

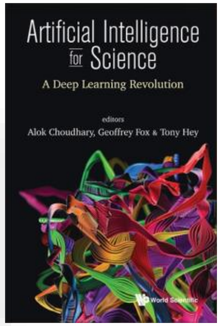


© 2023 World Scientific Publishing Company
https://doi.org/10.1142/9789811265679_0022

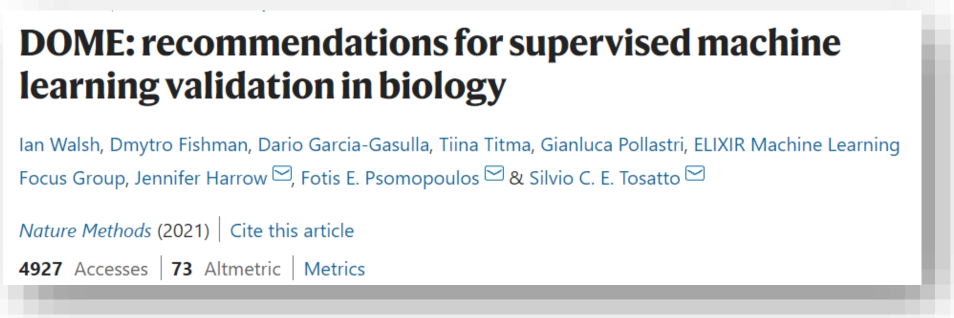
Chapter 22

A Roadmap for Defining Machine Learning Standards in Life Sciences

Fotis Psomopoulos*,‖, Carole Goble†,**, Leyla Jael Castro‡,††, Jennifer Harrow§,‡‡, and Silvio C. E. Tosatto¶,§§



Artificial Intelligence for Science
A Deep Learning Revolution
editors
Alok Choudhary, Geoffrey Fox & Tony Hey



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

elixir

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

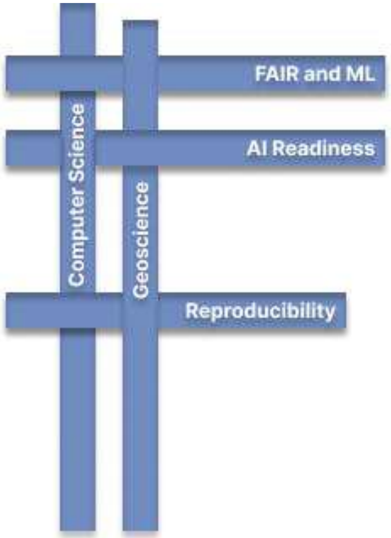
(GIGA)^n SCIENCE

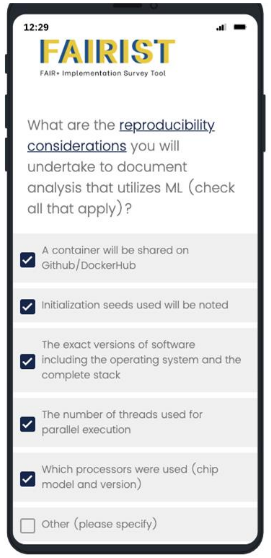
DOME adopted as part of the **submission system** for **GigaScience**
(*see example here: http://gigadb.org/dataset/102404*)
Online registry of annotated papers: https://registry.dome-ml.org

Need for Community-led Standards and Best Practices (2/2)

FARR
FAIR, AI Readiness & Reproducibility

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





12:29

FAIRIST

FAIR+ Implementation Survey Tool

What are the reproducibility considerations you will undertake to document analysis that utilizes ML (check all that apply)?

☑ A container will be shared on Github/DockerHub

☑ Initialization seeds used will be noted

☑ The exact versions of software including the operating system and the complete stack

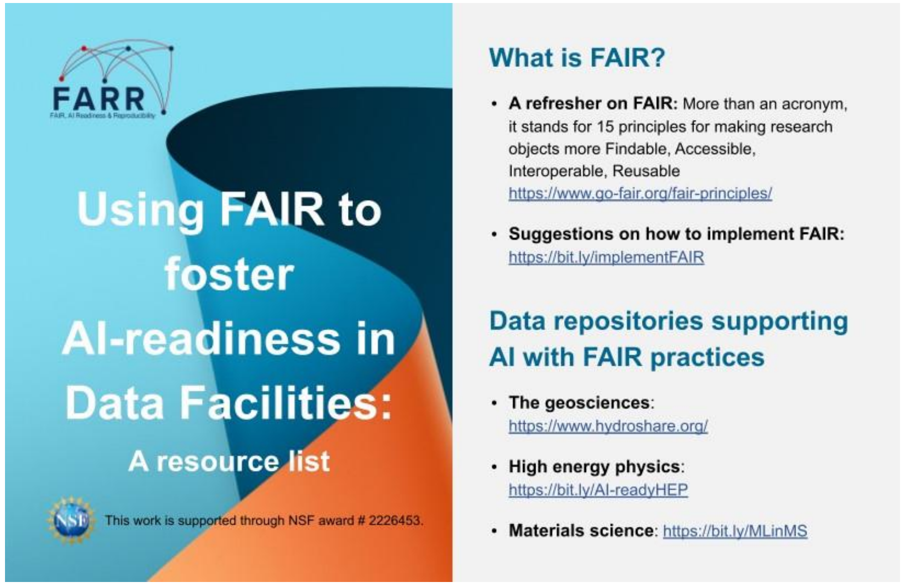
☑ The number of threads used for parallel execution

☑ Which processors were used (chip model and version)

☐ Other (please specify)

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?

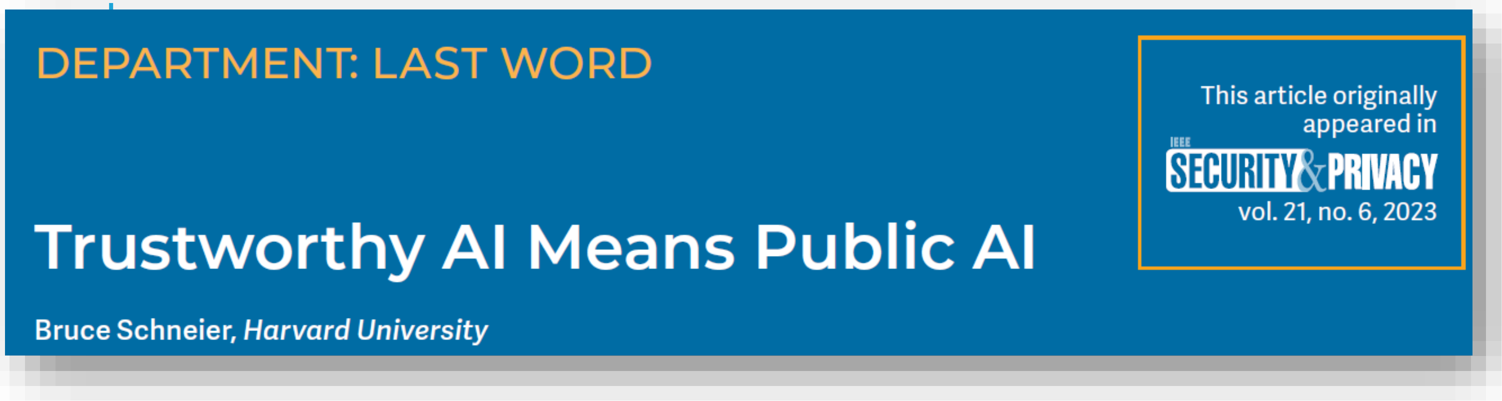
- **A refresher on FAIR:** More than an acronym, it stands for 15 principles for making research objects more Findable, Accessible, Interoperable, Reusable https://www.go-fair.org/fair-principles/
- **Suggestions on how to implement FAIR:** https://bit.ly/implementFAIR

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

This work is supported through the NSF award #2226453.

Contact:
https://www.farr-rcn.org/
community@farr-rcn.org

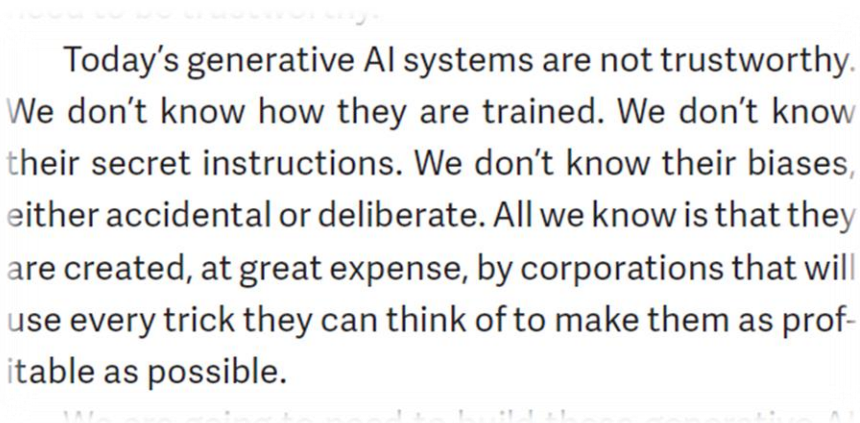


DEPARTMENT: LAST WORD

Trustworthy AI Means Public AI

Bruce Schneier, *Harvard University*

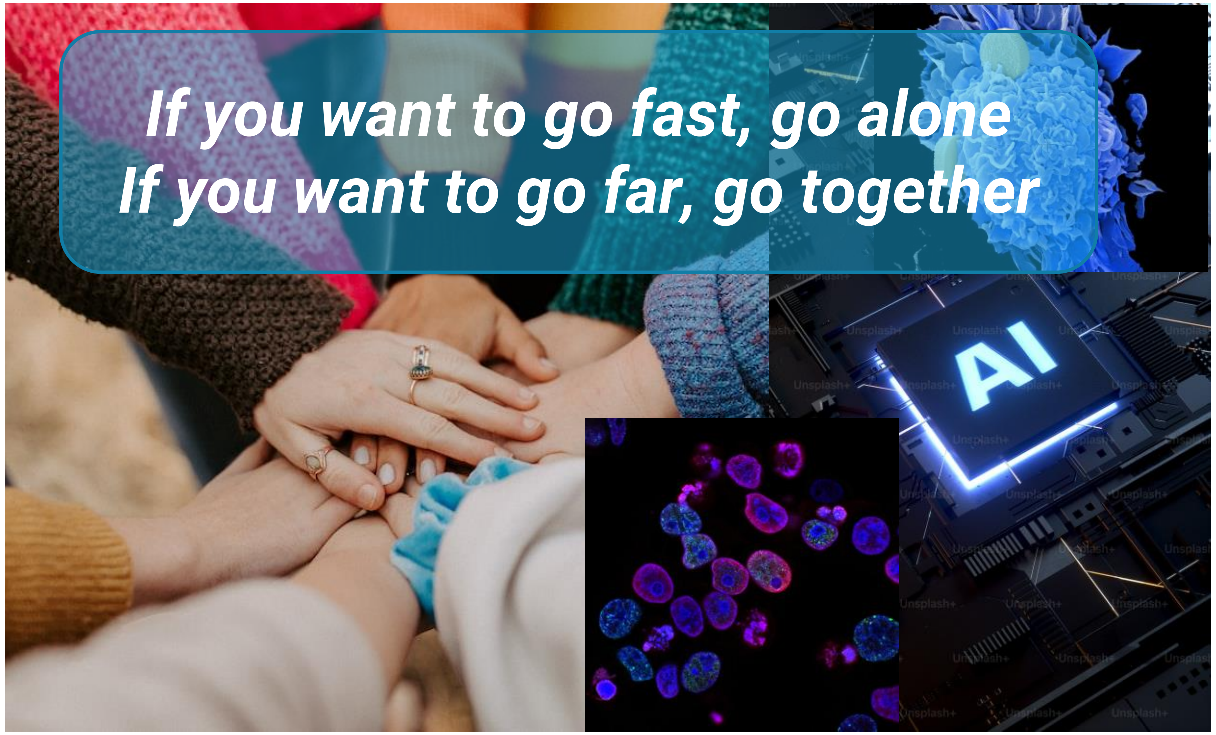
This article originally appeared in IEEE SECURITY & PRIVACY vol. 21, no. 6, 2023



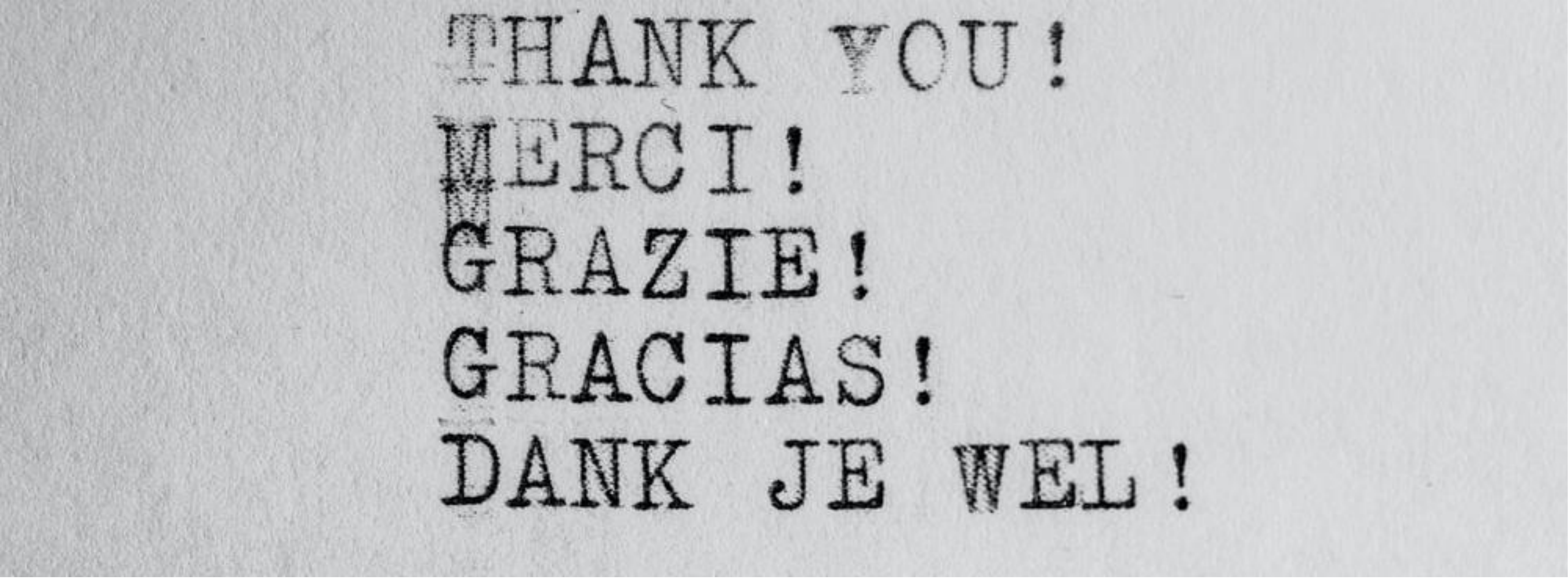
need to be trustworthy.

Today's generative AI systems are not trustworthy. We don't know how they are trained. We don't know their secret instructions. We don't know their biases, either accidental or deliberate. All we know is that they are created, at great expense, by corporations that will use every trick they can think of to make them as profitable as possible.

We are going to need to build these generative AI



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

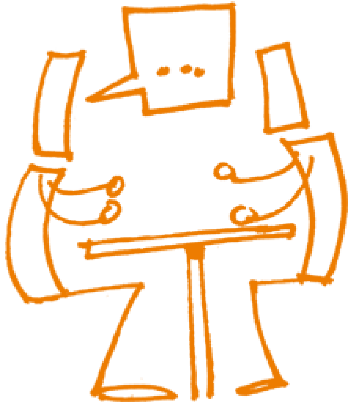


THANK YOU!
MERCI!
GRAZIE!
GRACIAS!
DANK JE WEL!

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

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

elixir



Slides available at: https://zenodo.org/records/10953108

@fopsom@genomic.social

@fopsom