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
  - Assumes separate and sequential creator/publisher (repository) roles
- 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

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

Need for FAIR for non-data objects

- FAIR Principles, are intended to apply to all digital objects (Wilkinson et al. 2016)

FAIR Practice Task Force EOSC, "Six Recommendations for Implementation of FAIR Practice," 2020:

Recommendation 5:

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

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

FAIR and ML Models

- As previously stated, original FAIR principles
  - Claim to apply to "scholarly digital research objects"
  - But actually focus on metadata and data
- FAIR for Research Software work and FAIR Workflows focusing on how to translate/interpret the principles for research software & workflows
- What about machine learning (ML) models?
  - 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 a combination of data+ software + workflows?
  - Are they something else?

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

How does FAIR apply?

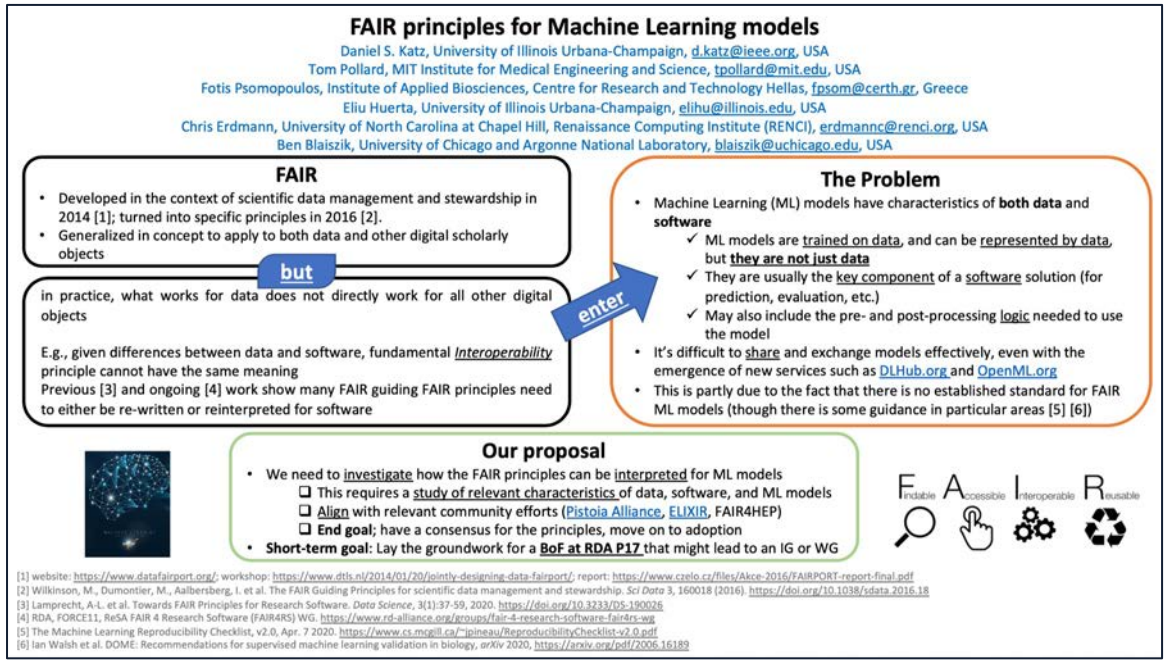
- Large elements of FAIR for data are dependent on archival repositories (e.g. Zenodo, re3data.org)
  - Hold data and/or metadata, provide search and access capabilities
- Software is different, since it typically isn't shared via archival repositories but instead via social coding platform (e.g., GitHub) and package management systems (e.g. PyPI, CRAN)
- What about ML models?
  - Searched and shared via repositories?
  - Searched and shared via executable platforms?
  - Searched and shared via something else? (e.g., DLHub, OpenML, …)
- Models and training data are linked - should they be shared together?

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

Work to-date and going forward

- Poster at RDA VP16 (Nov 2020):
- BoF at RDA VP17 (Apr 2021):
- FAIR for Machine Learning Models (Jun 2021), FAIR Festival
- 1st Community call (Jul 2021)
- DaMaLOS talk (24 Oct 2021)
- BoF at RDA VP18 (4 & 9 Nov 2021)
- BoF at SC21 (18 Nov 2021)
- BoF at RDA P19 (23 Jun 2022)

- Discussing a possible new interest group
- Discussing a potential white paper on FAIR 4 ML



FAIR principles for Machine Learning models

Daniel S. Katz, University of Illinois Urbana-Champaign, d.katz@ieee.org, USA
Tom Pollard, MIT Institute for Medical Engineering and Science, tpollard@mit.edu, USA
Fotis Psomopoulos, Institute of Applied Biosciences, Centre for Research and Technology Hellas, fpsom@certh.gr, Greece
Eliu Huerta, University of Illinois Urbana-Champaign, elihu@illinois.edu, USA
Chris Erdmann, University of North Carolina at Chapel Hill, Renaissance Computing Institute (RENCI), erdmannc@renci.org, USA
Ben Blaiszik, University of Chicago and Argonne National Laboratory, blaiszik@uchicago.edu, USA

FAIR

- Developed in the context of scientific data management and stewardship in 2014 [1]; turned into specific principles in 2016 [2].
- Generalized in concept to apply to both data and other digital scholarly objects

but

in practice, what works for data does not directly work for all other digital objects

E.g., given differences between data and software, fundamental *Interoperability* principle cannot have the same meaning
Previous [3] and ongoing [4] work show many FAIR guiding FAIR principles need to either be re-written or reinterpreted for software

enter

The Problem

- Machine Learning (ML) models have characteristics of **both data** and **software**
  - ✓ ML models are trained on data, and can be represented by data, but **they are not just data**
  - ✓ They are usually the key component of a software solution (for prediction, evaluation, etc.)
  - ✓ May also include the pre- and post-processing logic needed to use the model
- It's difficult to share and exchange models effectively, even with the emergence of new services such as DLHub.org and OpenML.org
- This is partly due to the fact that there is no established standard for FAIR ML models (though there is some guidance in particular areas [5] [6])

Our proposal

- We need to investigate how the FAIR principles can be interpreted for ML models
  - ❑ This requires a study of relevant characteristics of data, software, and ML models
  - ❑ Align with relevant community efforts (Pistoia Alliance, ELIXIR, FAIR4HEP)
  - ❑ **End goal**; have a consensus for the principles, move on to adoption
- **Short-term goal**: Lay the groundwork for a **BoF at RDA P17** that might lead to an IG or WG

Findable Accessible Interoperable Reusable

[1] website: https://www.datafairport.org/; workshop: https://www.dtls.nl/2014/01/20/jointly-designing-data-fairport/; report: https://www.czelo.cz/files/Akce-2016/FAIRPORT-report-final.pdf
[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[3] Lamprecht, A-L. et al. Towards FAIR Principles for Research Software. *Data Science*, 3(1):37-59, 2020. https://doi.org/10.3233/DS-190026
[4] RDA, FORCE11, ReSA FAIR 4 Research Software (FAIR4RS) WG. https://www.rd-alliance.org/groups/fair-4-research-software-fair4rs-wg
[5] The Machine Learning Reproducibility Checklist, v2.0, Apr. 7 2020. https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf
[6] Ian Walsh et al. DOME: Recommendations for supervised machine learning validation in biology, *arXiv* 2020, https://arxiv.org/pdf/2006.16189

Defining FAIR for Machine Learning (ML)

Home

25 JAN 2021

Defining FAIR for Machine Learning (ML)

Submitted by Daniel S. Katz

Meeting objectives:
Discuss:

- Current projects (both research and infrastructure) in machine learning (ML) that are considering FAIR,
- If there's value in and a need for defining FAIR for ML, and if so,
- How to move forward to do so, ideally under the RDA umbrella based on the current role of RDA in FAIR activities

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

Relevant projects & stakeholders

- Platforms
  - DLHub - Find, share, publish, and run machine learning models and discover training data for science
  - Kipoi - API & repository of ready-to-use trained models for genomics
  - OpenML - Build open source tools to discover (and share) open data, draw them into machine learning environments, build models, analyse results, get advice on better models
- Communities
  - Pistoia Alliance - a global, not-for-profit members' organization working to lower barriers to innovation in life science and healthcare R&D through pre-competitive collaboration
  - ELIXIR - An intergovernmental organisation that brings together life science resources (including databases, software tools, training materials, cloud storage and supercomputers) from across Europe
  - CLAIRE - Confederation of Laboratories for Artificial Intelligence Research in Europe
- Projects
  - FAIR4HEP - Using high-energy physics (HEP) as the science driver, developing a FAIR framework to advance understanding of AI, applying AI techniques, and exploring approaches to AI
  - HPC-FAIR - Providing a generic HPC data management framework to make both training data and AI models of scientific applications FAIR, focusing on the domain of program analyses/optimizations using AI/ML
- Others? Please contact Dan

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