Data Management Makes Machine Learning Easier

Till Korten, Helene Hoffmann, Peter Steinbach, Özlem Özkan

Data management days, HZDR
October 29, 2024

Helmholtz AI Consulting for Matter

- Based at Helmholtz-Zentrum Dresden-Rossendorf
- Working in research field matter
- **Exploration voucher: 80 working hours of our time**
- Data management makes our collaboration much more effective
- A few practical examples for each letter in FAIR Data

FAIR

Findable data gets used

- AI benchmarks
- Training material for foundational models
- As teaching material in machine learning courses
- For AI research projects
- If your data is findable, you don’t need to find AI experts to collaborate with - they will find you!

😀

Collaborators find you

FAIR

Findable data gets used

Example (>4000 citations) :

Deng, L. ***The MNIST Database of Handwritten Digit Images for Machine Learning Research*** *[Best of the Web]. IEEE Signal Processing Magazine 2012, 29 (6), 141–142. https://doi.org/10.1109/MSP.2012.2211477.*

😀

Used data gets cited

FAIR

Accessible data saves time and effort

Where is your data?		How can you share it?
On paper	➔	Digitise it ~ hours to weeks
On a computer	➔	Upload to cloud ~ 1h/GB
On a portable harddisk	➔	Borrow hard disk ~ min to days
On a Fileserver	➔	Upload to cloud ~ 20min/GB
On cloud storage	➔	Send link ~ 10 s
In a public repository	➔	Send DOI ~ 10 s

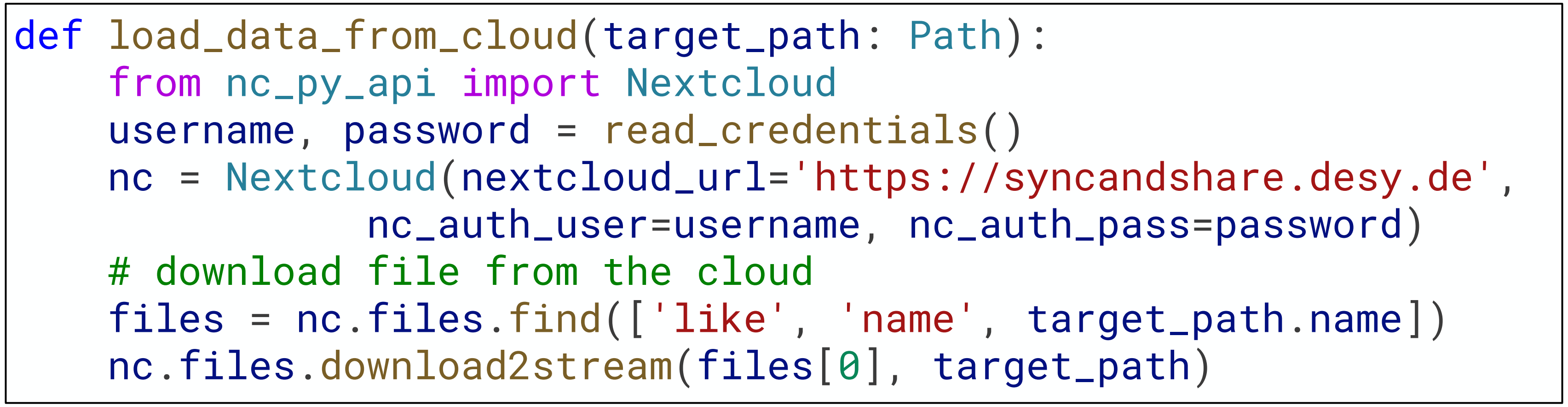
Common example

- 6 GB on a fileserver:
  - 2h: find a suitable cloud storage
  - 2h: file upload
  - 4h: write a data loader
  - 2h: file download
- 10h total

FAIR

Why machine learning experts love **MNIST**

- Data is ready for machine learning in 2 min -> see notebook
- You can achieve this by uploading a simple python package with a few lines of code to https://pypi.org



```python
def load_data_from_cloud(target_path: Path):
    from nc_py_api import Nextcloud
    username, password = read_credentials()
    nc = Nextcloud(nextcloud_url='https://syncandshare.desy.de',
                   nc_auth_user=username, nc_auth_pass=password)
    # download file from the cloud
    files = nc.files.find(['like', 'name', target_path.name])
    nc.files.download2stream(files[0], target_path)
```

- Example: https://github.com/psteinb/b3get

Interoperable data is easy to open

- **Open** file format
  - **Open** source (ideally python) libraries for opening the data exist
  - Metadata is **well structured** and **machine readable**
  - **Good Practice**: have script on how to load small example of your dataset
    (this can also be automatically tested)
- For supervised machine learning:
  - Data is **labeled**

Writing a workflow to read and convert an obscure format can take days

FAIR

Interoperable data is easy to open

- **Open** file format
  - **Open** source (ideally python) libraries for opening the data exist
  - Metadata is **well structured** and **machine readable**
  - **Good Practice**: have script on how to load small example of your dataset *(this can also be automatically tested)*
- For supervised machine learning:
  - Data is **labeled**

Ex: OME Zarr

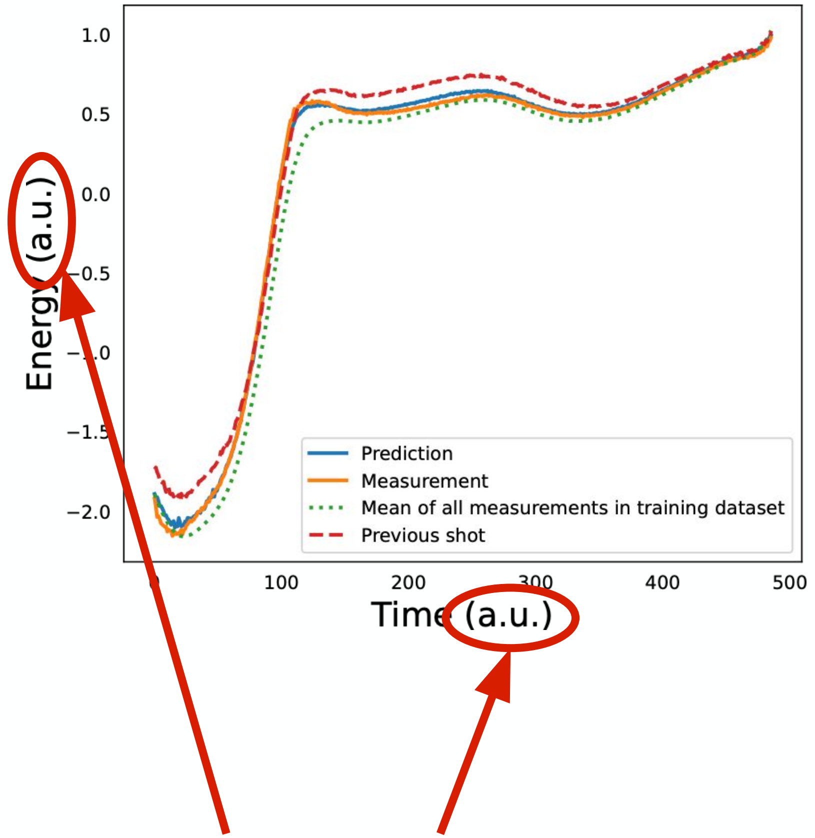
for Microscopy data

FAIR

Reusable data is easy to work with

Documentation

- The data is **well described** and **easy to understand**
- Descriptions are **detailed enough** that data can be used for more than the originally intended purpose

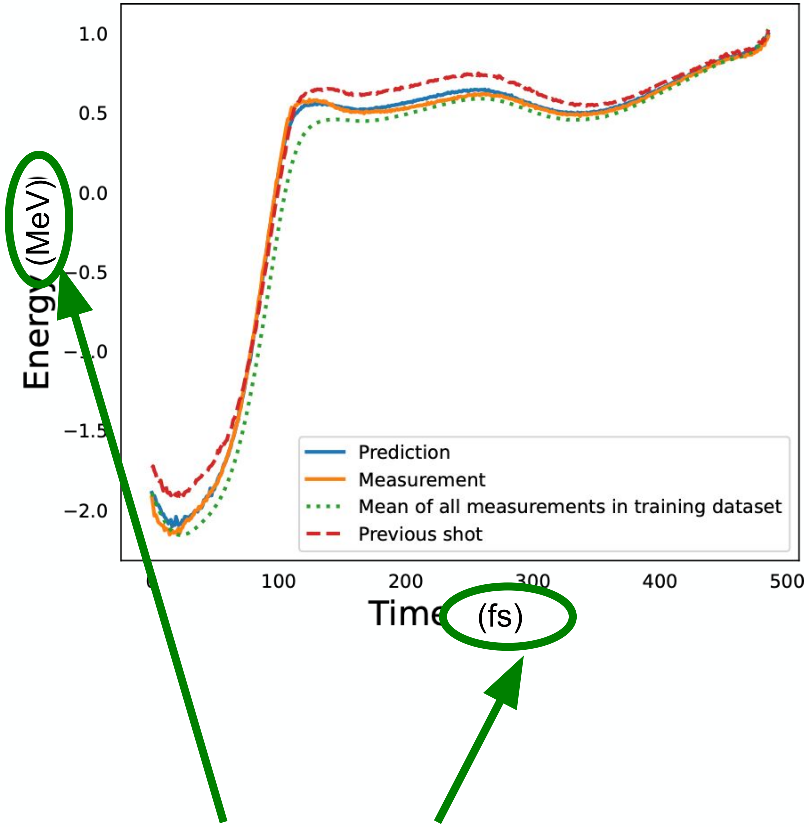


FAIR

Reusable data is easy to work with

Documentation

- The data is **well described** and **easy to understand**
- Descriptions are **detailed enough** that data can be used for more than the originally intended purpose

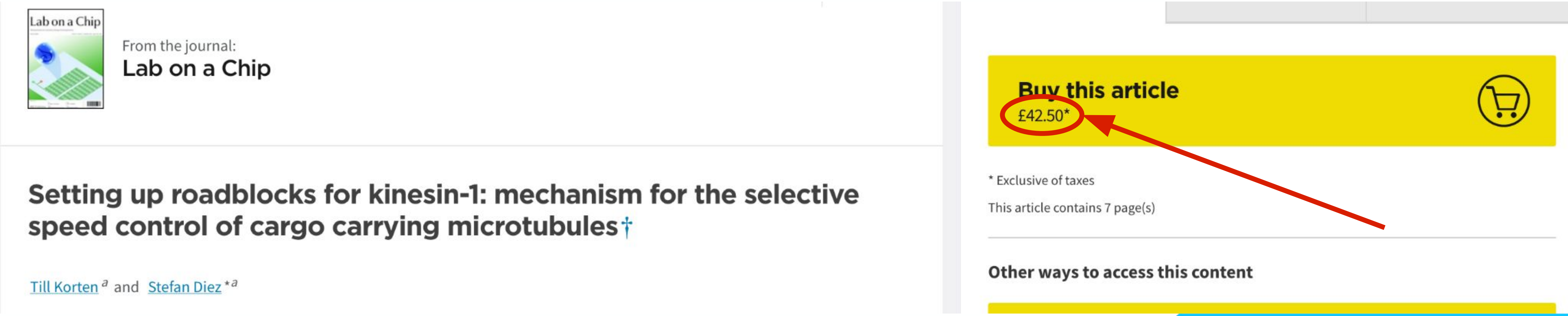


FAIR

Reusable data is easy to work with

Licensing

- Data is **licensed** *(ideally with a permissive creative commons license)*
  - I once paid 35 dollars for my own paper because of a **restrictive license**



From the journal:
Lab on a Chip

Setting up roadblocks for kinesin-1: mechanism for the selective speed control of cargo carrying microtubules†

Till Korten [a] and Stefan Diez *[a]

Buy this article
£42.50*

* Exclusive of taxes

This article contains 7 page(s)

Other ways to access this content

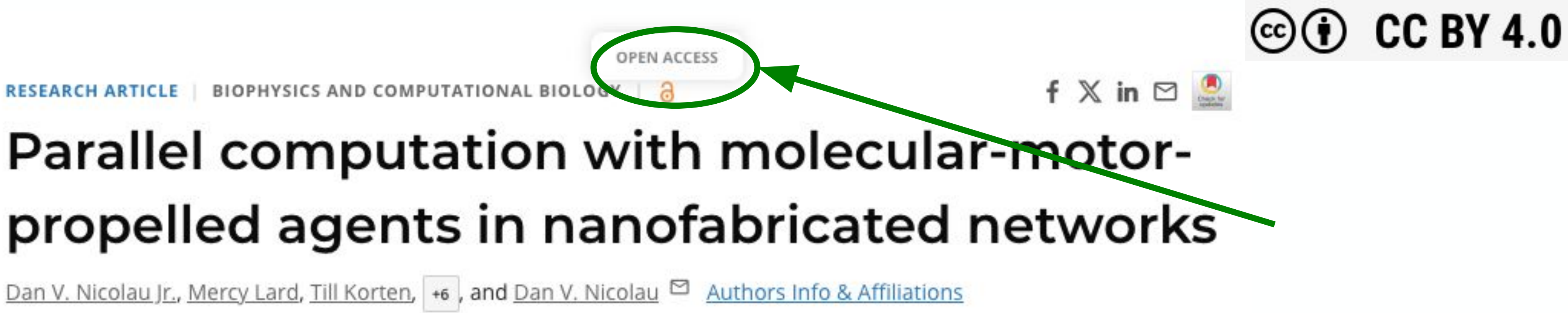
Reusable data is easy to work with

FAIR

Licensing

- Data is **licensed** *(ideally with a permissive creative commons license)*
  - My most cited paper is open access
    - ideally under a permissive license like CC BY 4.0

→ the lower the hurdles, the more citations



CC BY 4.0

OPEN ACCESS

RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Parallel computation with molecular-motor-propelled agents in nanofabricated networks

Dan V. Nicolau Jr., Mercy Lard, Till Korten, +6 , and Dan V. Nicolau Authors Info & Affiliations

FAIR

Reusable data is easy to work with

Citable

- Data has a **DOI** (DFG counts data citations)

Standardised

- Data and metadata **follow community standards**

Control.gif
sample-1.jpg
sample_two.jpg
anothersample.png

📂Control
wildtype.tif
📂Samples
<gene_code>-.tif
<gene_code>+.tif

Data management saves time - 80 h total

Accessing the data

Convert data to a usable, machine readable open format

Understanding the data

Check metadata and labels

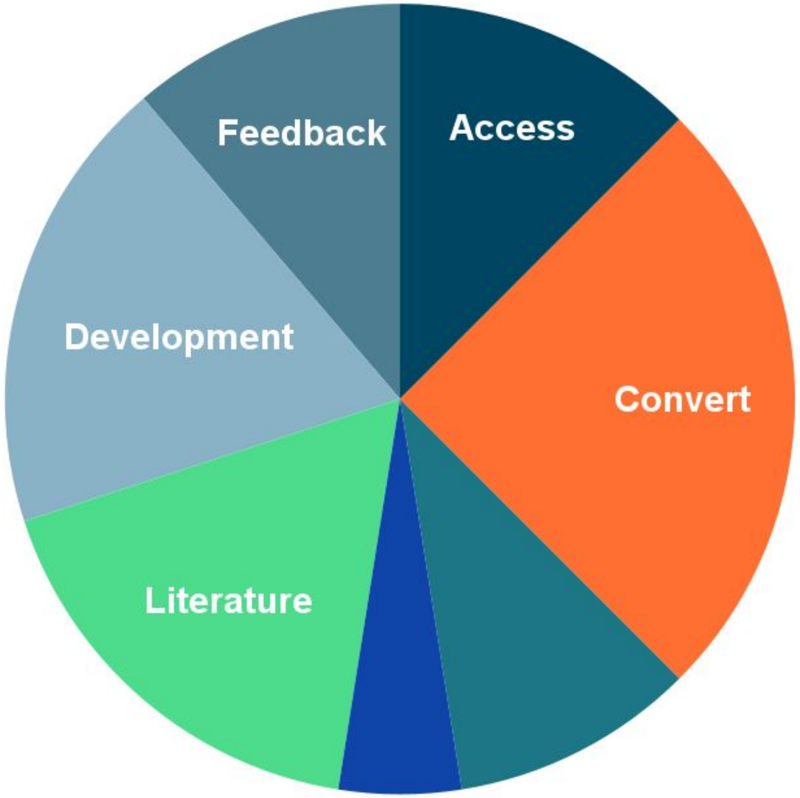
Literature search

Develop a machine learning workflow

Feedback and final improvements

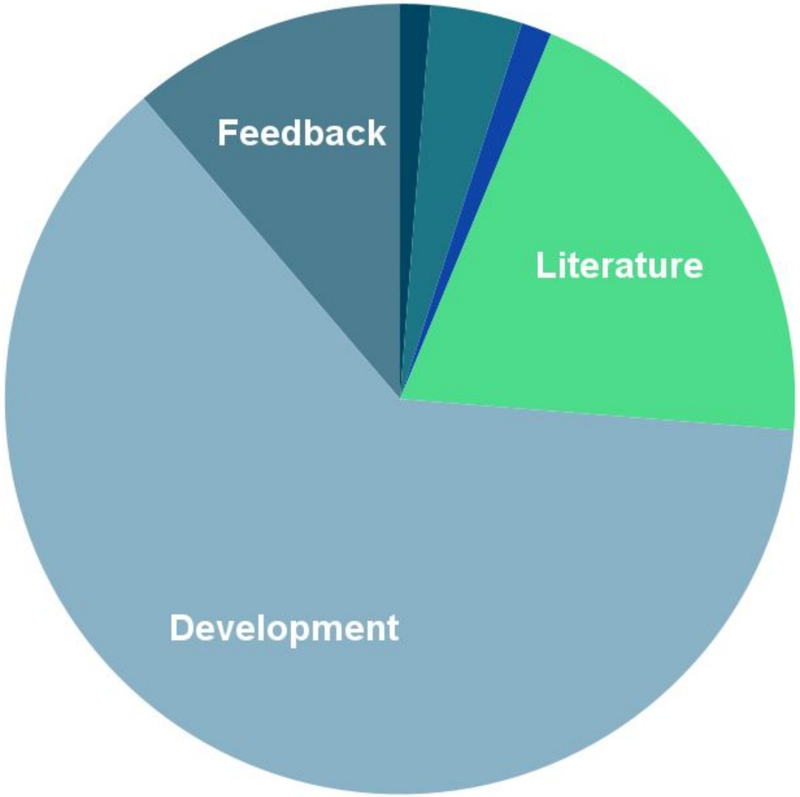
Less than half the time for the actual work

Room for improvement



Almost the entire time for the actual work

Good data management



Summary and Conclusion

- Good Data Management enables Good Machine Learning
  (reduce 80/20 split, FAIR data enables model building, can massively reduce time to solution, availability of data ensures transparency and progress)
- Curating a Data Set entails software and data science skills
  (collaborate where you can, the higher the load on the ML engineer - the less ML is done, well curated data ensures transparency and progress)

Slides on figshare
(CC-BY 4.0)

Thank you for attention!
We are happy to take questions, feedback or concerns.

Shout out to our collaborators!
Helmholtz AI Consulting Team HZDR
Helmholtz Metadata Collaboration (HMC)