## Machine learning workshop

- "Exploiting open data for machine learning training: can the Photon and Neutron community do it?"

- Hybrid event: SOLEIL + online

- Duration: 2 half days + network dinner

    2023-10-17 / 2023-10-18, just before the GA.

- Registration:

    - Cost: On-site registration €50; online registration is free

    - Capacity: 40 for on-site registrants (current 3), no limits for online participation.

    - Deadline: 2023-10-01 for on-site registration; no deadline for online participation.

Leveraging open data from PaN facilities for machine learning

17–18 Oct 2023
Europe/Paris timezone

Overview
Timetable
Programme Committee
Registration
Call for Abstracts
Accommodation

Contact

✉ majid.ounsy@synchrotr...
✉ sophie.servan@desy.de

Exploiting open data for machine learning training: can the Photon and Neutron community do it?

Introduction

During the last decade, most European Photon and Neutron (PaN) facilities have adopted **open data policies**, making data available for the benefit of the entire scientific community. At the same time, **machine learning** (ML) is seen as an essential tool to address the exponential growth of data volumes from PaN facilities.

**Exploitation of experimental training datasets** is a key component of machine learning. The combination of ML algorithms and open data can therefore be seen as an ideal marriage that would ultimately help the entire community to tackle 'big data' challenges with more automation.

https://indico.synchrotron-soleil.fr/event/67/

# We also have a flyer (available from Indico home page)



## Leveraging open data from PaN facilities for machine learning

Oct 17 - 18, 2023

Synchrotron SOLEIL - CNRS - CEA, Paris-Saclay, France

### Exploiting open data for machine learning training: can we do it?

During the last decade, most European Photon and Neutron (PaN) facilities have adopted open data policies, making data available for the benefit of the entire scientific community. Exploitation of experimental training datasets is a key component of machine learning. However, finding the right data to train algorithms is a challenge and one of the motivations for making data FAIR is exactly that: to provide scientists working on AI applications with quality training datasets.

But what does 'quality' mean to PaN science communities? What metadata fields are needed to find the data, to understand if it is suitable for our research, and ultimately to be able to ingest it in our training models? How can we provide sufficiently rich metadata?

### Objectives

With this workshop, we aim to discuss these questions, among staff and users of the LEAPS and LENS facilities, across disciplines and across Europe.

We will present projects and teams that have successfully used open datasets from PaN facilities to train their specific ML application (data consumers), as well as domain scientists (data producers) who have published curated data specifically for ML applications. We will also look at cases where it hasn't worked so well, and discuss why and how it could be better in the future.

### Call for abstracts

During the workshop, we have slots available for 20 minute presentations. We are particularly interested in contributions that can address the following points:

- Have you used open data for machine learning? What has been your experience?
- Do you feel your research could benefit from more and better curated open data? Tell us how.
- Have you already made open datasets available to your community for ML training purposes?
- Have you managed to enrich the metadata of open datasets for better training?

Whatever your discipline, as long as the data originally came from experiments at a photon or neutron source, please submit an abstract in Indico A

**Programme**

- Current structure:
  - Two keynote talks: one on day 1, one on day 2.
  - First day has a 02:15 session on "community talks".
  - Second day has three parallel sessions, over a two hour period.
- Call for abstracts is out.
  - Requesting presentations for the "community talks" section.
  - Deadline for submitting abstracts is 2023-09-08.

**Dissemination**

- Programme committee has a list of organisations and people we are contacting.

    - Probably best to try to coordinate this.

    - Please let "us" (the programme committee) know if you can think of someone we should contact. We can add them to the list.

- Please help to spread the word!

    - Print out flyers and put them around in your facility

    - Let people know who you think might be interested.