European XFEL Users' Meeting 2025, Hamburg, 20-24 January 2025

Data reduction activities at European XFEL

Egor Sobolev, Philipp Schmidt, Janusz Malka, David Hammer, Djelloul Boukhelef, Johannes Möller, Karim Ahmed, Richard Bean, Ivette Jazmín Bermúdez Macías, Johan Bielecki, Ulrike Bösenberg, Cammille Carinan, Fabio Dall'Antonia, Sergey Esenov, Hans Fangohr, Romain Letrun, Danilo Enoque Ferreira de Lima, Luís Gonçalo Ferreira Maia, Hadi Firoozi, Gero Flucke, Patrick Gessler, Gabriele Giovanetti, Jayanath Koliyadu, Anders Madsen, Thomas Michelat, Michael Schuh, Marcin Sikorski, Alessandro Silenzi, Jolanta Sztuk-Dambietz, Monica Turcato, Oleksii Turkot, Jose Luis Vazquez-Garcia, James Wrigley, Steve Aplin, Steffen Hauf, Krzysztof Wrona and Luca Gelisio

European XFEL, Schenefeld, Germany

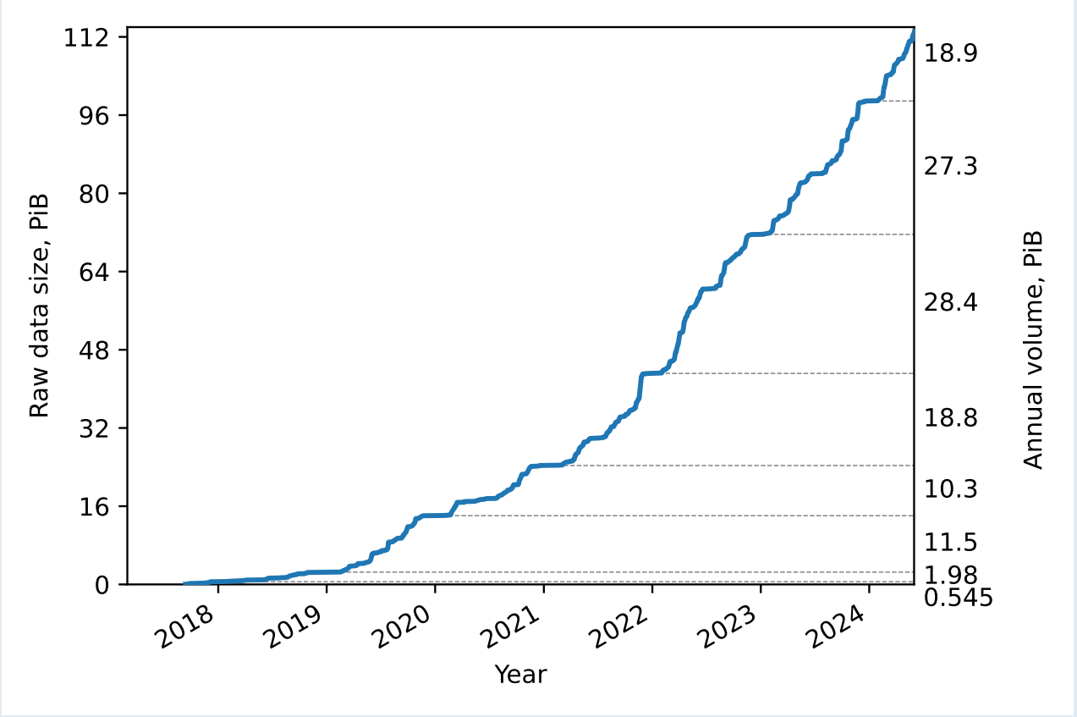
Introduction

The European XFEL is a megahertz repetition-rate facility producing extremely bright and coherent pulses of duration of the order of few femtoseconds or less. Owing to its X-ray imagers, specifically built to operate at these repetition rates (AGIPD, DSSC and LPD), the amount of data generated in the context of user experiments can exceed hundreds of gigabits per second, resulting in tens of petabytes stored every year. These rates and volumes pose significant challenges both for the facility and its users. In fact, if unaddressed, extraction and interpretation of scientific content is hindered, and investments and operational costs quickly becomes unsustainable.

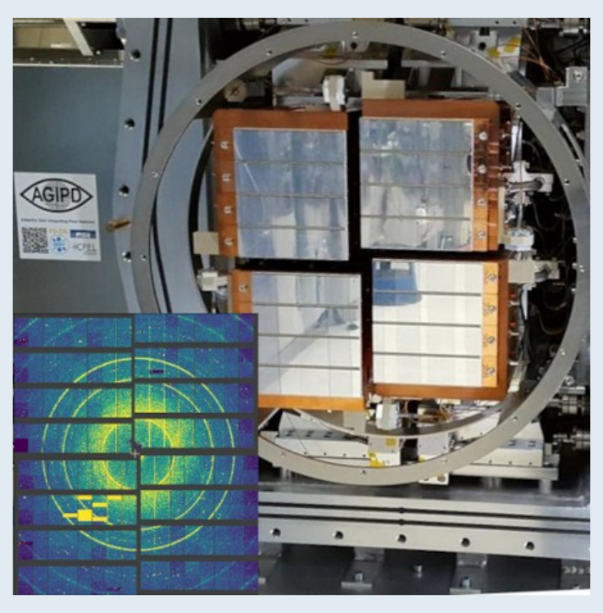
Growing big data challenges

Multiple fast area detectors at rates >> 100 Gb/s

- Common bias towards storing raw data
- Growth of raw data production is unsustainable
- Upcoming upgrades
  - AGIPD 4M detector
  - Duty cycle increasing up to 50 %

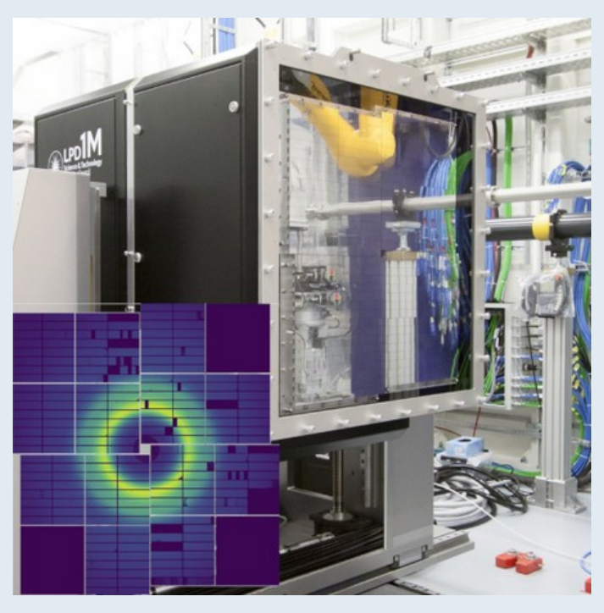


AGIPD-1M



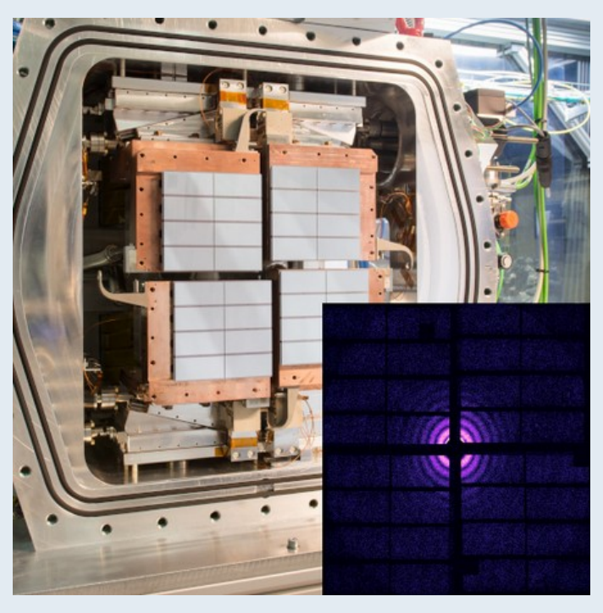
≥ 21 h/PiB

LPD-1M



≥ 29 h/PiB

DSSC-1M



≥ 19 h/PiB

AGIPD-4M

coming

≥ 6 h/PiB

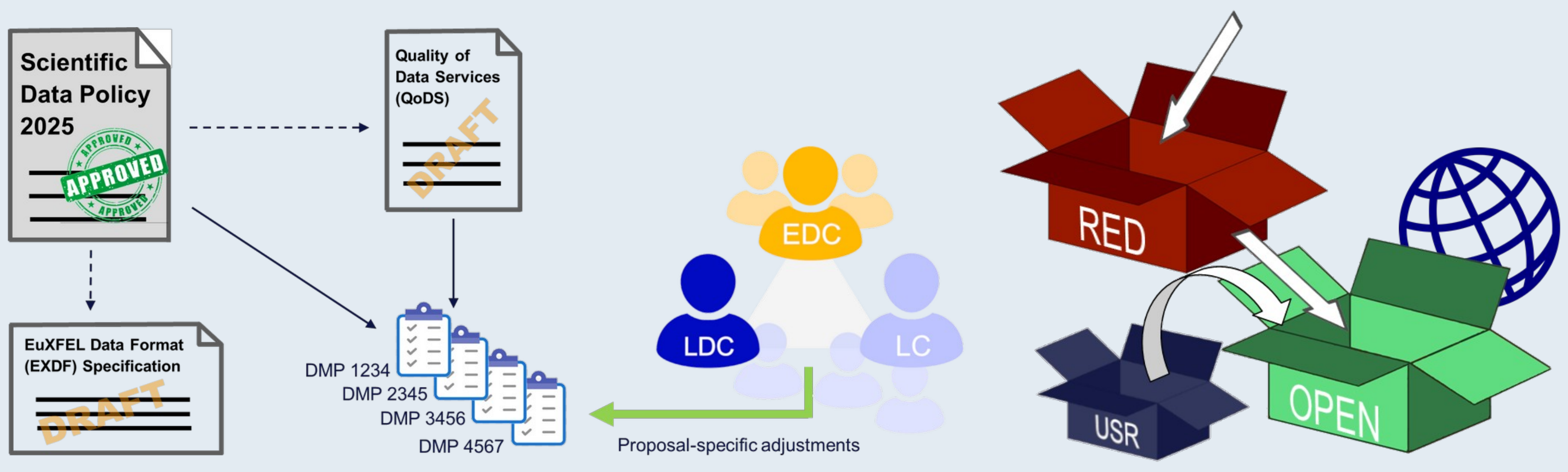
Strategy

- Scientific Data Policy and the RED box concept
- Co-development data reduction methods with and for users
- Extensive evaluation with users through pilot projects
- Integration of data reduction tools in the data handling infrastructure
- Provision of quality metrics to sustain automation

Scientific data policy 2025+

New Scientific Data Policy taking effect in 2025

- Data reduction becomes an (within limits, see bellow) obligatory early step in the lifecycle of experiment
- Implement FAIR principles, help users with ubiquitous requirements to make published data available openly
- Customized to the needs of each experiment by **Data Management Plan**


Scientific Data Policy 2025
APPROVED
Quality of Data Services (QoDS)
DRAFT
EuXFEL Data Format (EXDF) Specification
DRAFT
DMP 1234
DMP 2345
DMP 3456
DMP 4567
EDC
LDC
LC
Proposal-specific adjustments
RED
USR
OPEN

RED box and OPEN data

- The size of collected raw data determines the data volume which will be retained long-term and opened up later

  ***RED* = max 10 % *RAW*, (min 50 TiB, *RAW*)**

- RED box may consist of any raw or processed data in approved formats

Data reduction methods

- **Operation-specific methods**

  Related to instrument operation itself, little or no analysis of experimental data is usually required. This methods are robust, low risk, and the feedback latency is compatible with online requirements.
  - **Region of interest, e.g. module: 1-16 times**
  - **Lit-frame selection: 1-100 times**
  - **Compression: up to 40 times**
  - **Gain suppression: 2 times**
- **Technique-specific methods**

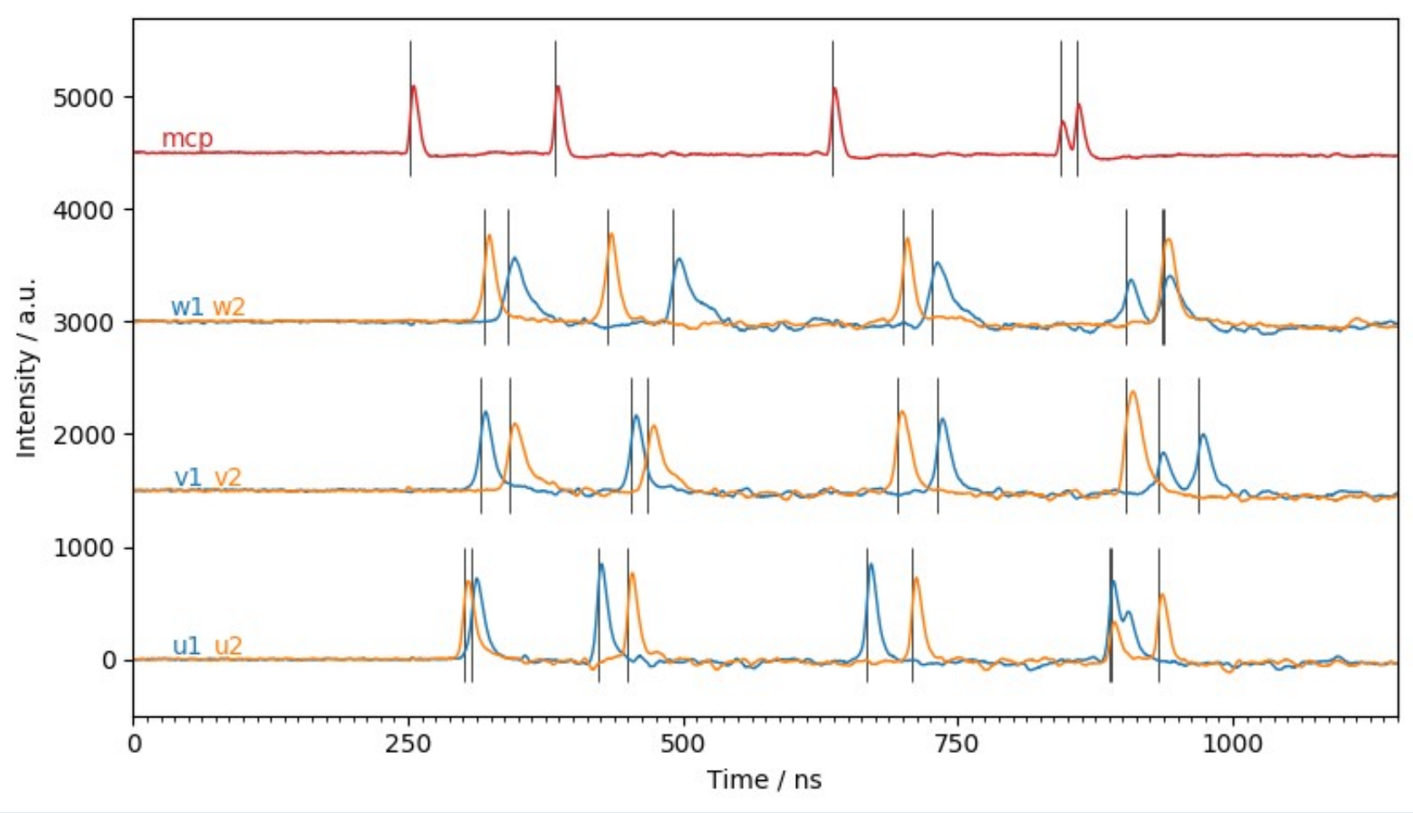
  Require analysis of experimental data, and typically involve tuning of certain parameters. The associated risks are generally higher, computational complexity is higher as well, and there are challenges for automation
  - **Hit finding: > 10 times**

    SFX, SPI
  - **Event reconstruction: ~2000 times**

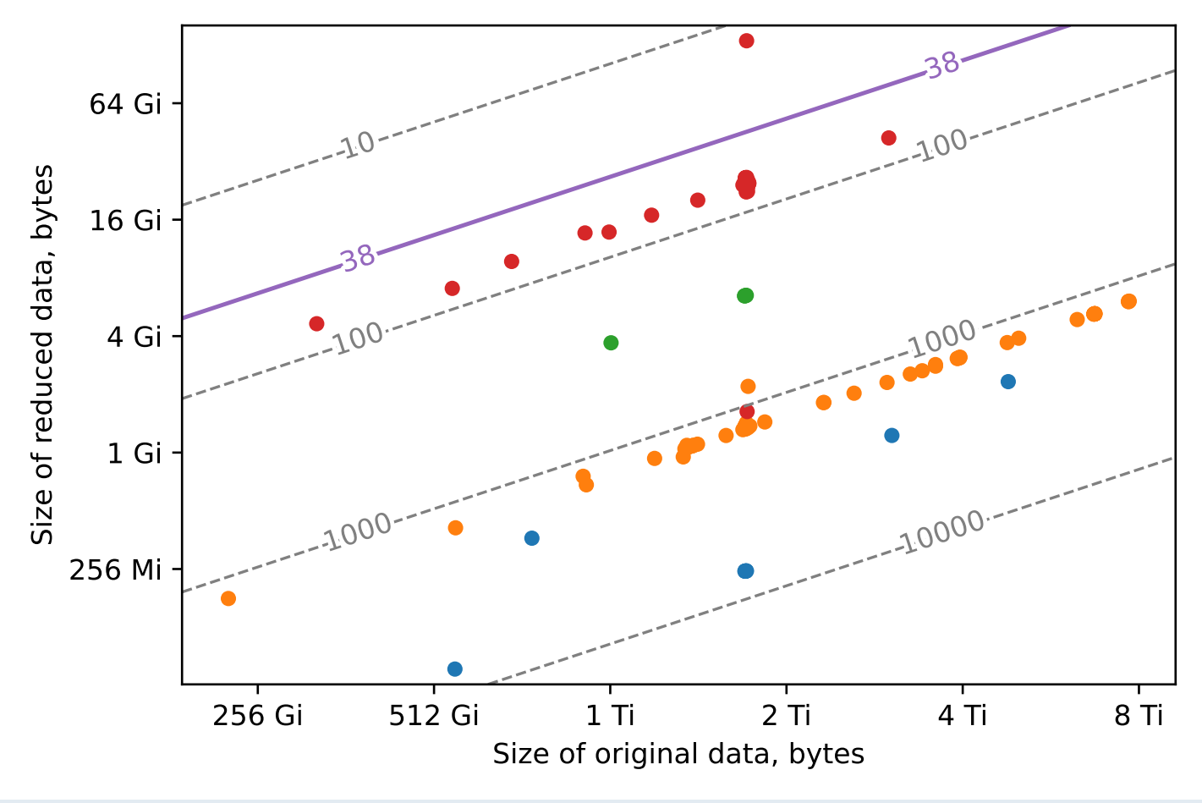
    REMI/COLTRINS, (tr-)RIXS
  - **Azimutal integration: ~1000 times**

    SAXS, WAXS, Powder diffraction, XPCS
  - **Correlation functions: ~1000 times**

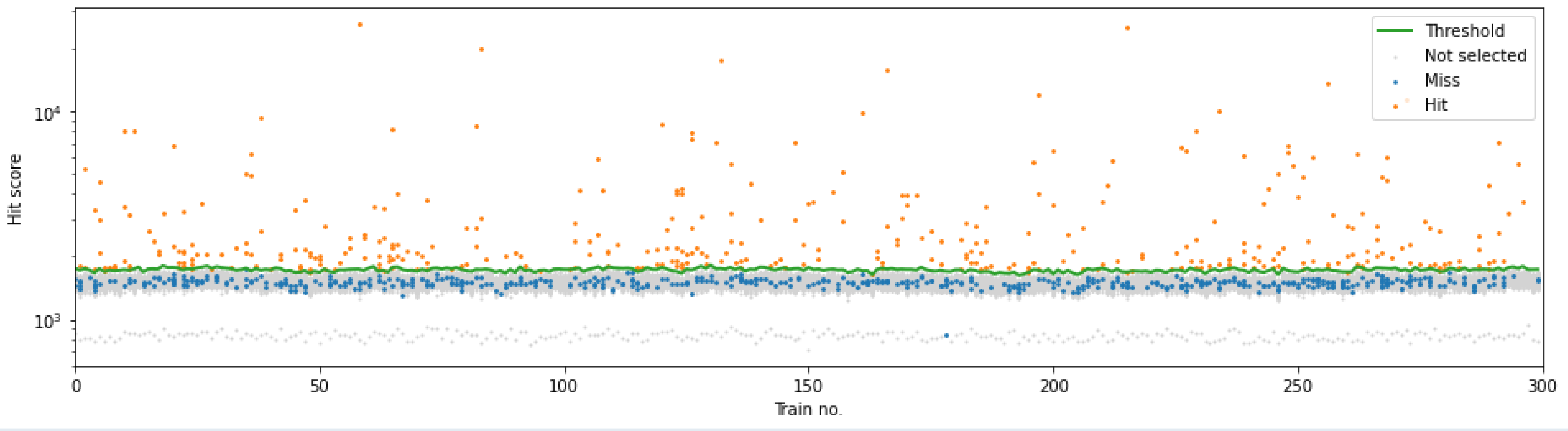
    XPCS, XCCA



REMI/COLTRINS: traces to edges transformation



Lit-frame selection and compression



Single particle hit-finding

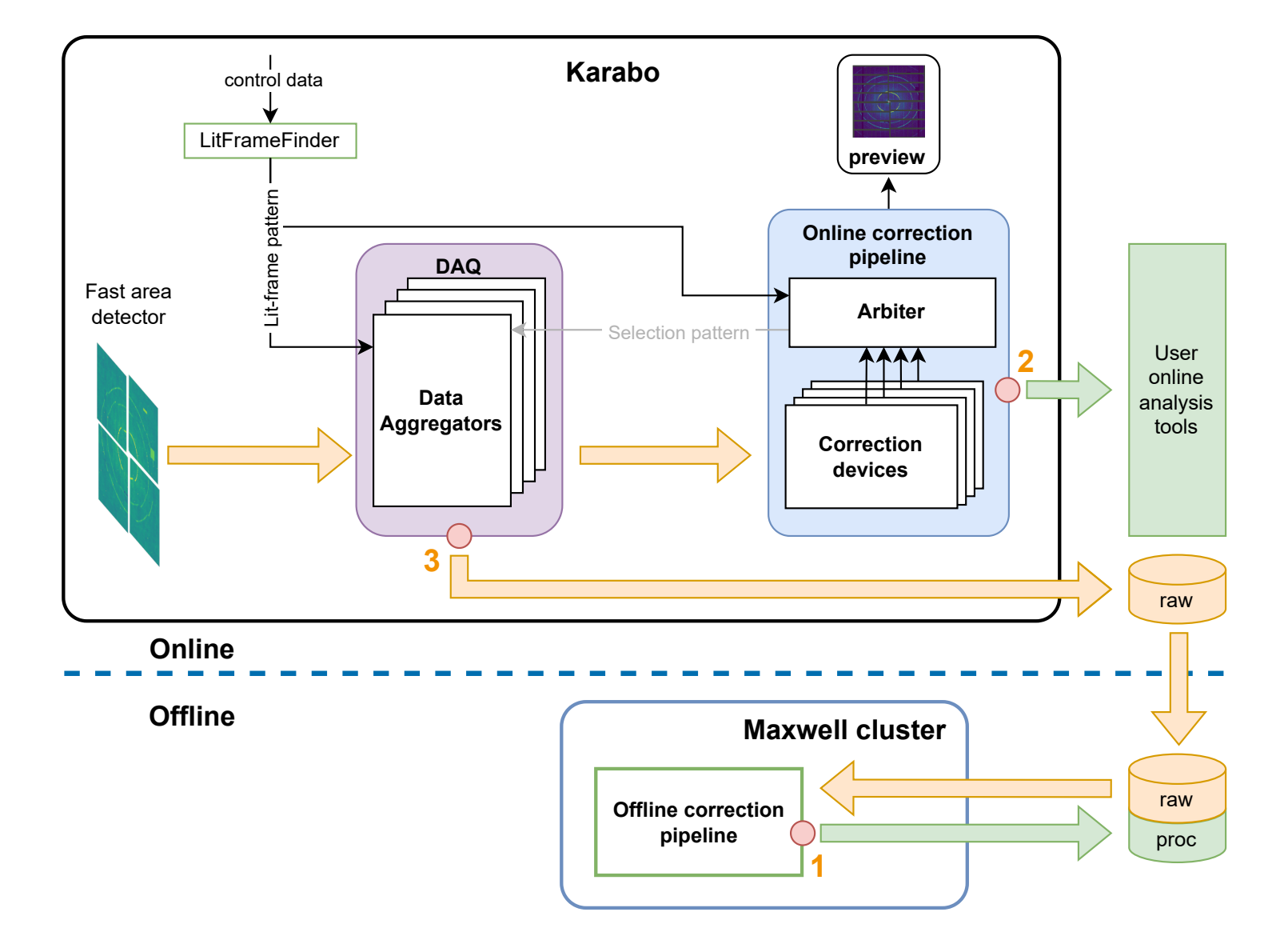
Data reduction integration points

1. **Offline processing**

   Most reproducible and safe
2. **Online processing**

   Mitigate bandwidth and computing power limits
3. **Acquisition**

   Maximal impact downstream, no turning back



Publications

Sobolev, Schmidt et al: Data reduction activities at European XFEL: early results, Front. Phys. 12, 1331329 (2024)

Acknowledgements

Special thanks to European XFEL users joining in pilot projects: Jonas Seilberg, Duane Loh, Filipe Maia, Xavier Paulraj, Kartik Ayyer and many others

Data reduction activities at the European XFEL have been partially funded through the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101004728.

European XFEL GmbH, Egor Sobolev, Holzkoppel 4, 22869 Schenefeld, Germany, Phone +49 40 8998-6556, egor.sobolev@xfel.eu
www.xfel.eu