Data reduction activities at European XFEL: early results

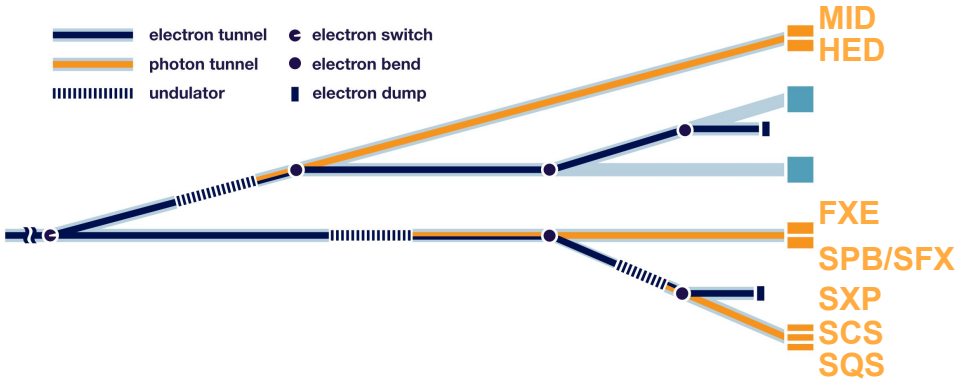
Ivette Bermudez
Data Analysis, European XFEL

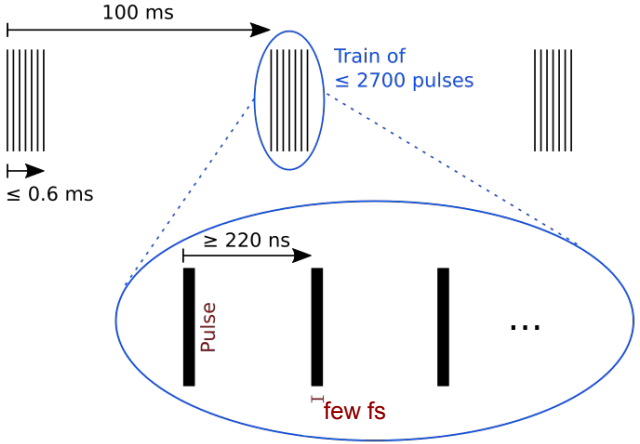
On behalf of **many** others:
Egor Sobolev, Janusz Malka, David Hammer, Djelloul Boukhelef, Johannes Möller, Karim Ahmed, Richard Bean, **Ivette Bermúdez Macias**, Johan Bielecki, Ulrike Bösenberg, Cammille Carinan, Fabio Dall'Antonia, Sergey Esenov, Hans Fangohr, Danilo Enoque Ferreira de Lima, Luís 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, James Wrigley, Steve Aplin, Steffen Hauf, Krzysztof Wrona, Luca Gelisio

08 April 2024

Sobolev E, et al. Data reduction activities at European XFEL: early results. Front. Phys. 12:1331329.
doi: 10.3389/fphy.2024.1331329

Facility overview



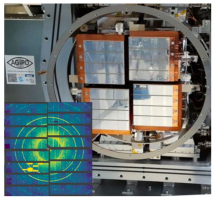
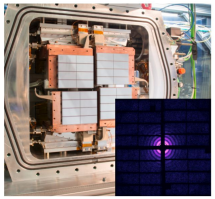


- 3 beamlines covering soft & hard X-rays
- 2-3 instruments per beamline, 7 in total
- Multiple endstations per instrument
- Pulses split across beamlines, all three operating at the same time

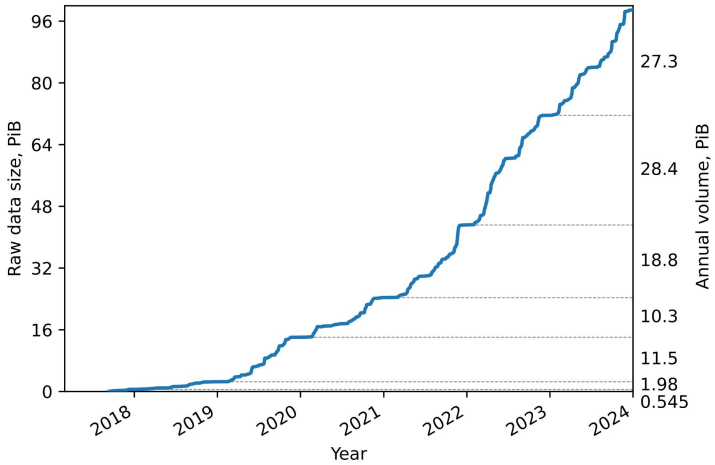
- Burst mode similar to FLASH
- 10 Hz trains with ≤ 2700 pulses each split across all beamlines
- Typically each instrument receives hundreds of pulses

Growing Big Data challenges

- Multiple fast area detectors at data rates ≫ **100 Gb/s**

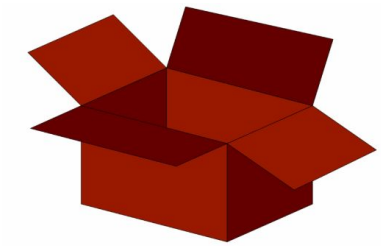
AGIPD-1M	LPD-1M	DSSC-1M	AGIPD-4M
			?
≥ 21 h/PiB	≥ 29 h/PiB	≥ 19 h/PiB	≥ 6 h/PiB

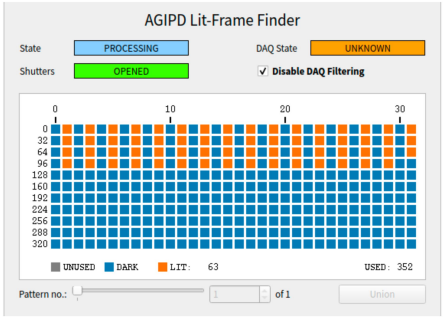
- Growth of raw data production is **unsustainable**
- Upcoming upgrades:
  - AGIPD-4M detector
  - Duty cycle increasing up to 50%

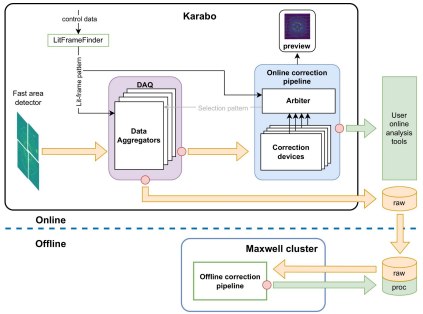


Addressing Big Data challenges

- Scientific Data Policy and the RED box
- Data reduction methods and current pilot projects
- Data reduction integration points in the infrastructure

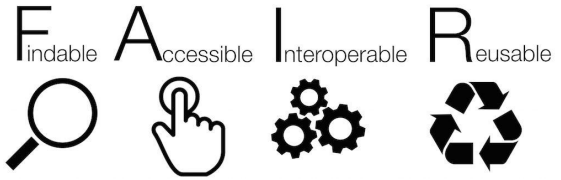




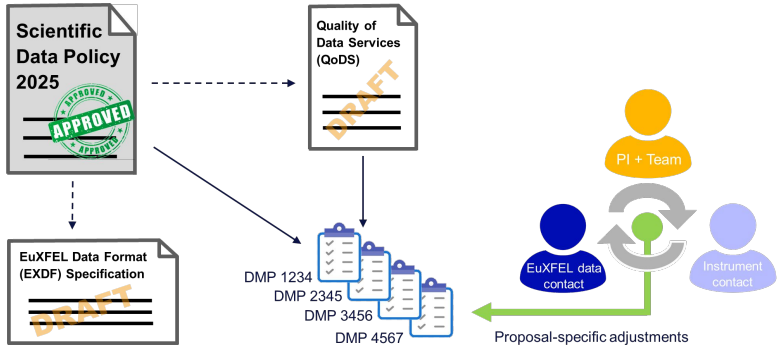


Scientific Data Policy 2025+

- New Scientific Data Policy (SDP) taking effect in 2025
- Data reduction becomes an (within limits) obligatory **early step** in the life cycle of experiments
- Implement FAIR principles, help users with ubiquitous **requirements** to make published data **available openly**
- Customizing to the needs of each experiment
  by **Data Management Plan** (DMP)



Wilkinson, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).



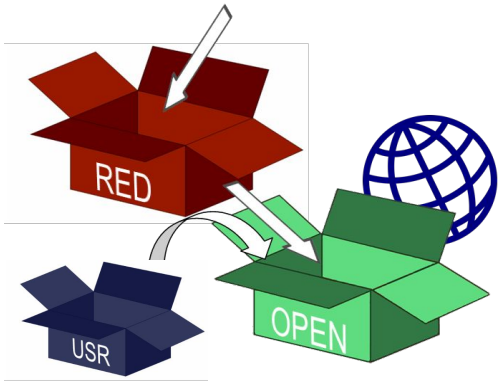
RED box and OPEN data

Please do not quote these exact numbers yet!

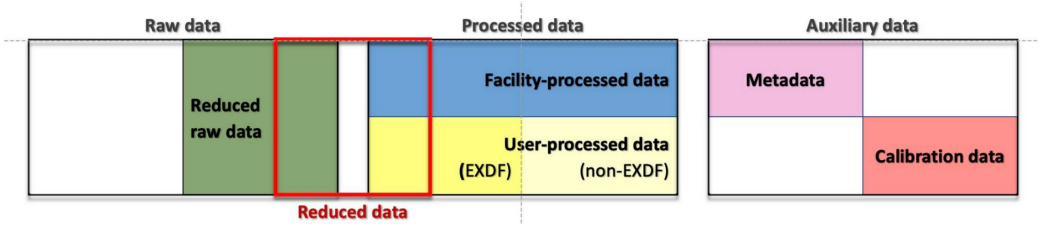
- The **size of raw data** determines the data volume
  retained **long-term** and **opened up** later:

  $RED = \max 10\% \; RAW, \; (\min 50 \; TiB, \; RAW)$

  If the raw data recorded for your proposal is
  - **below 50 TiB**, you can retain up to the **size of raw data**
  - **between 50 TiB and 500 TiB**, you can retain **50 TiB**
  - **above 500 TiB**, you can retain **10% of raw data**



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



Data reduction methods

Operation-specific methods

Related to instrument operation itself, little or no analysis of experimental data is usually required.

These methods are robust, low risk, and the feedback latency is compatible with online requirements.

- **ROIs**, e.g. module: 1-16
- **Lit-frame selection**: 1 - 100
- **Compression***: up to 40
  *often requires technique-specific preparation
- Gain suppression: 2

Technique-specific methods

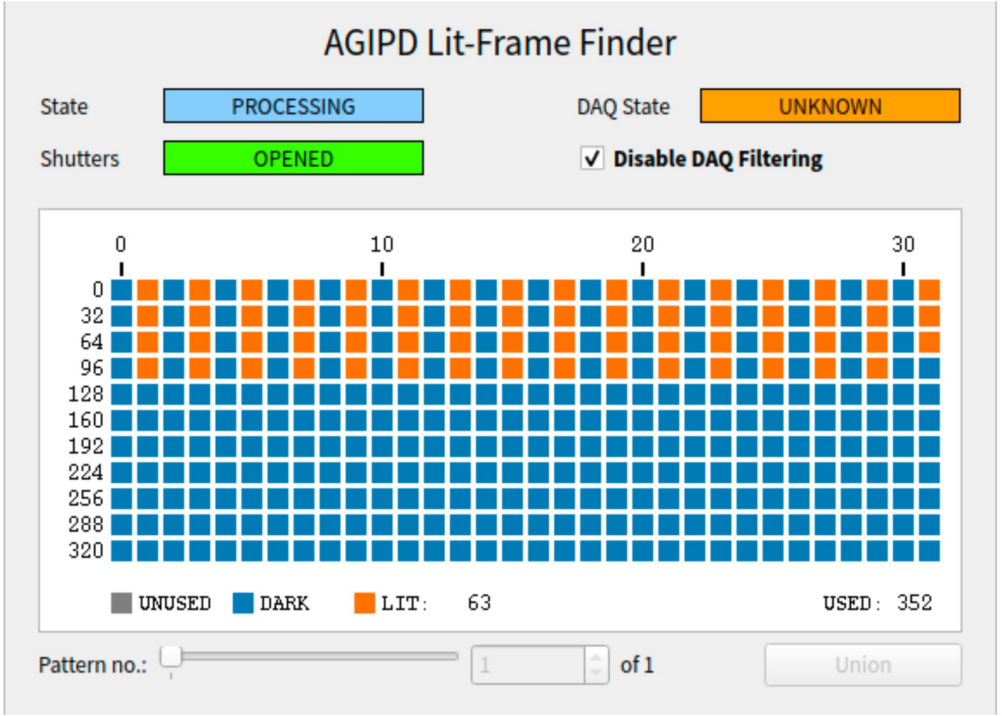
Require analysis of experimental data, and typically involve tuning of certain parameters

The associated risks are generally higher, computational complexity is higher as well, and there are challenges for automation.

- **Hit finding**: 10 - ~1000
  SFX, SPI
- **Event reconstruction**: ~2000
  REMI/COLTRIMS, (tr-)RIXS
- Azimuthal integration: ~1000
  SAXS, WAXS, Powder diffraction, XPCS
- Correlation functions: ~1000
  XPCS, XCCA

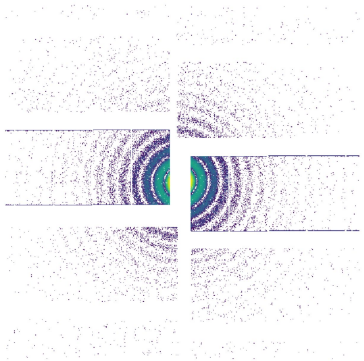
Lit frame selection

- Realtime annotation of detector frames based on
  - Pulse pattern
  - Detector configuration
  - Trigger timing
  - Shutter states
- Deployed in production to only consider lit frames for processing
- Used in pilot experiments to filter on DAQ level, or as initial input to further online data reduction

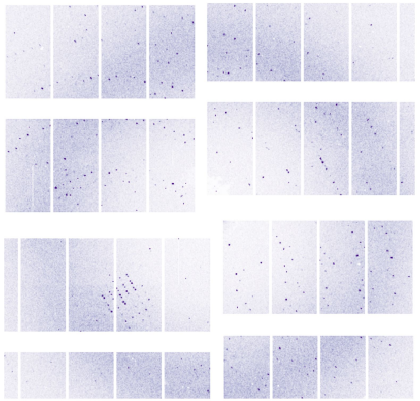


Compression

- Detector data, especially once calibrated to absolute energy, does not compress well. Depending on the illumination pattern, some technique allow reducing the entropy:
  - Low intensity scattering
    Conversion and rounding to integer photon counts
    XPCS, Bragg CDI, SPI

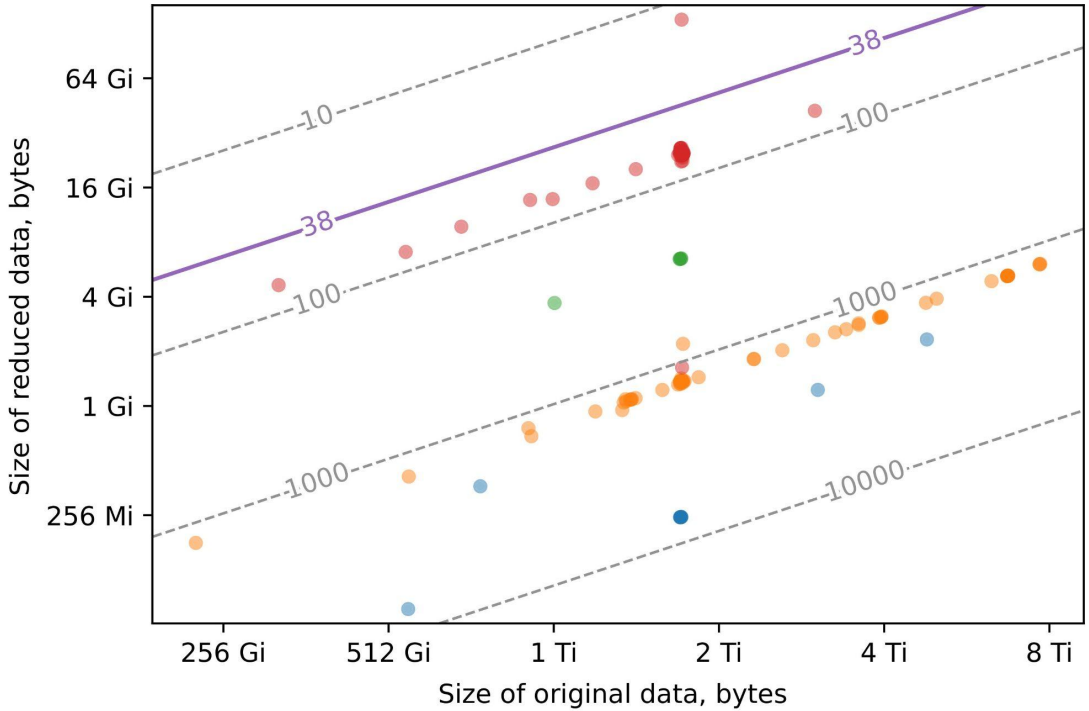


  - High intensity scattering
    Rounding to few highest significant bits
    SFX



Typical results for XPCS experiments

- AGIPD lit-frame selection to automatically account for varying illumination patterns
- Rounding to nearest photon count after gain correction and compression
- Routinely applied now to XPCS experiments at MID



Retrospective reduction

- Select data by the data keys or event Ids
- Data sliced according to the results of any external processing pipeline
- Tools to semantically reduce recorded data *EXDF-tools*
  - Maintain EXDF data structure
  - Reduction on series of operations on data or its structure
  - Easy extendable

```
r0585_proc_events.h5
└entry_1
  ├cellId          [uint16: 1089088 × 16]
  ├do_integrate    [int8: 1089088]
  ├hit_indices     [int64: 12214]
  ├litpixels       [uint64: 4 × 1089088]
  ├modules         [int64: 4]
  ├pulseId         [uint64: 1089088]
  ├pulse_energy_uJ      [float64: 1089088]
  └trainId         [uint64: 1089088]
```

Results: applied reductions

- Avoided storage of 7.4 PiB (as of 11.2023)

Reduction method	Type	Instrument	Experiments	Original data size, PiB	Reduction factor
Lit-frame selection	raw	SPB/SFX	2	0.88	3.8
	proc	SPB/SFX	12	3.8	1.2
		MID	10	5.8	2.5
Train selection	proc	HED	4	0.52	19
Conversion to ph. and compression	proc	MID	10	5.8	17

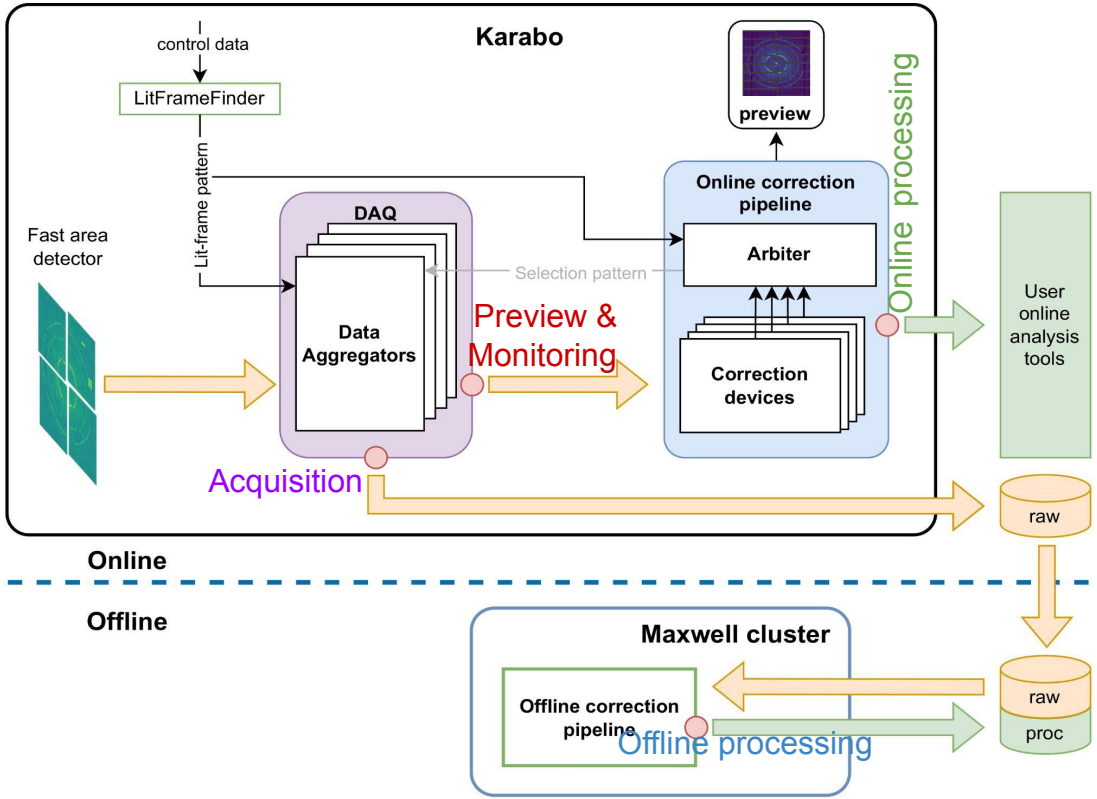
Results: Candidates to retroactive reductions

- 17 PiB expected to be freed (as of 11.2023)

Reduction method	Type	Instrument	Experiments	Original data size, PiB	Reduction factor
Lit-frame selection	raw	SPB/SFX	27	9	1.11
		MID	23	14	1.9
Gain information suppression	raw	SPB/SFX	5	1.2	2
		MID	12	7.4	2
Train selection	raw	HED	4	0.52	19
Module selection	raw	MID	5	2.3	5
SPI hit finding	raw	SPB/SFX	4	5.5	19

Data reduction integration points

- **Acquisition**
  Maximal impact downstream, no turning back
- **Preview & Monitoring**
  Simplify real-time analysis
- **Online processing**
  Mitigate bandwidth and computing power limitations
- **Offline processing**
  Most reproducible and safe, still large impact on user analysis

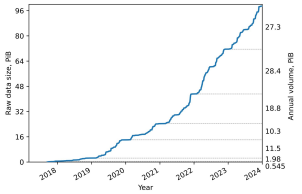


Future developments

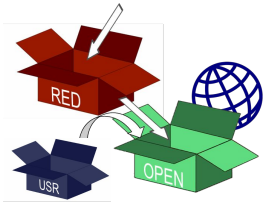
We are already reducing data :).. But we still record too fast.

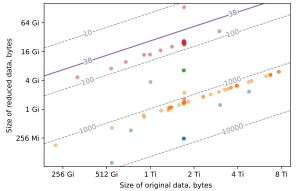
- Development of new methods to cover most experiments
- Infrastructure developments
  - Faster data transferring from DAQ to online correction pipeline (RDMA)
  - Feedback to DAQ with reduction decision or modified data
  - Online validation metrics, eg. conditional accumulators
  - Track online parameters to use them in offline reduction
  - Extend the file format document data reduction
- Development of DMP services and interface to fill RED box
- Extensive validation

Conclusions



- **Retaining full raw data** is **not sustainable** and future upgrades make recording impossible
- **Scientific Data Policy 2025+** makes data reduction a first-class citizen of scientific data curation and management
- **Facilities** must develop and **provide** operation- and technique-specific **reduction methods**
- Integration points for **data reduction** and **validation** both online & offline, **open to** and **extendable by users.**
- Reducing data attractive to users: faster road between experiment and publication.





Thank you for your attention

Data analysis group: da@xfel.eu, www.xfel.eu/data_analysis

Get in touch about SDP, DMP and data reduction: data-policy@xfel.eu