Data reduction activities at European XFEL

Egor Sobolev

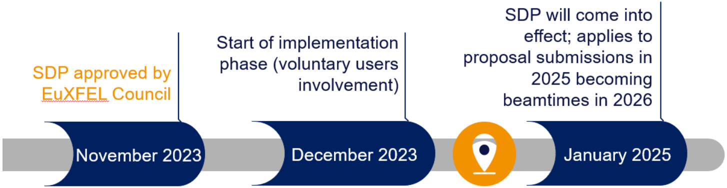
European XFEL

Data management, analysis, and reduction at European XFEL
European XFEL Users’ Meeting

26 January 2024

...what?!

As you lernt, data reduction is enabled as part of the forthcoming scientific data policy.



SDP approved by EuXFEL Council

November 2023

Start of implementation phase (voluntary users involvement)

December 2023

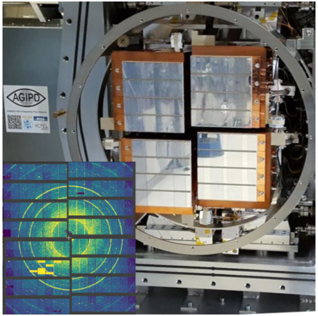
SDP will come into effect; applies to proposal submissions in 2025 becoming beamtimes in 2026

January 2025

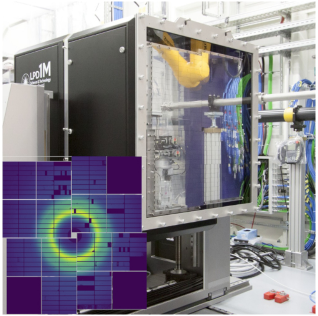
...no worries, we will do this together!

But why do we need to start reducing experimental data?

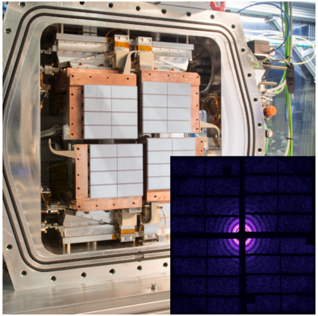
Big data producers – fast area detectors



AGIPD-1M – 21 hours/PiB

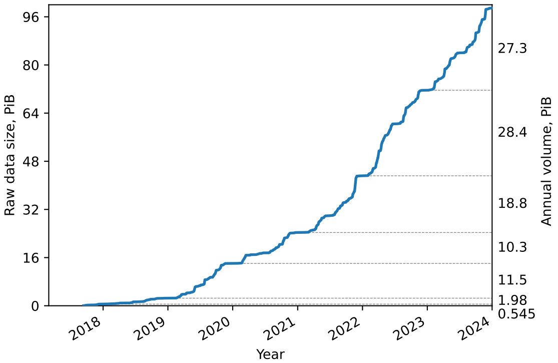


LPD-1M – 29 hours/PiB



DSSC-1M – 19 hours/PiB

Raw data production at European XFEL



Increasing at this rate is not sustainable.

Is there any benefit for me?

- The quality of your data is higher, as the fraction not contributing to answer your scientific question are not stored.
- Many data reduction steps are in any case also part of your data analysis pipeline: we are integrating these as part of our infrastructure, and aim at making them as robust and performant as possible, and at providing extensive validation of results.
- Any future analysis should be simpler, and scientific results should be obtained faster.
- We can keep data on disks for longer time, and open data can be exploited better.

OK, who does what?

We will try our best to

- work with you to provide a portfolio of tools and services;
- design extensive validation metrics;
- support you as much as we can.

You will

- decide.

Yes!

- you know best your scientific question;
- you know best your data and details of the experiment.

Red box – the draft concept

Within **six months** users have to decide on how to respect the storage quota defined as

$$\max\left[10\%\,\mathrm{raw\ data\ volume}, \min\left(\mathrm{raw\ data\ volume}, 50\mathrm{TiB}\right)\right]$$

We have been working over the last years with several of you, and designed and tested tools together

- for example, slicing data from SPI/SFX using the output of your analysis,
- or compressing your XPCS data,
- – many thanks for your engagement so far!

Example: Data selection by user analysis

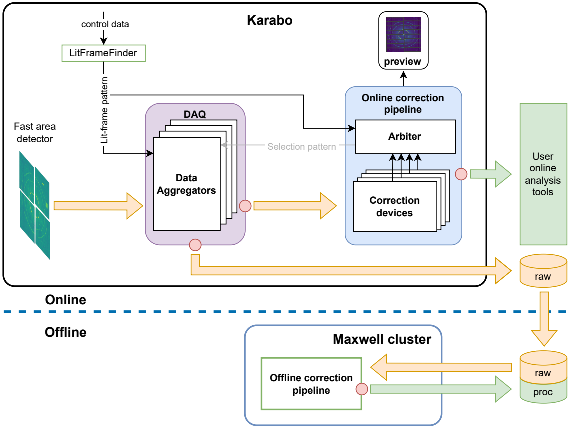
Use your data analysis pipeline.

- We provide a tool to slice stored data (EXDF-tool).
- In the future there will be web-service to select data.
- We are validating the approach with our instruments and volunteered users (Thanks: Jonas Sellberg, Duane Loh, Kartik Ayyer)

Data infrastructure

Offline correction pipeline

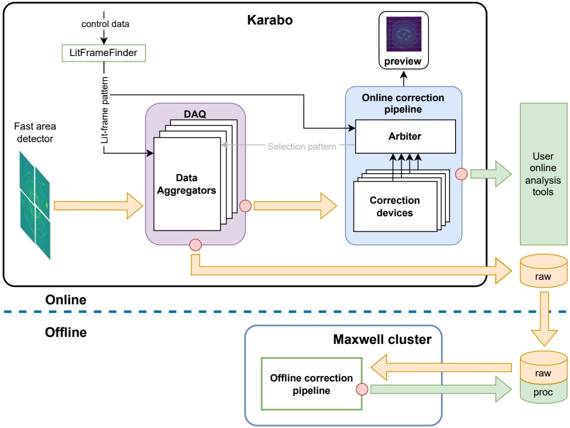
- Data annotation, transformation
- Processed data filtering
- Parameters are set for all runs in proposal



Data infrastructure

Online correction pipeline

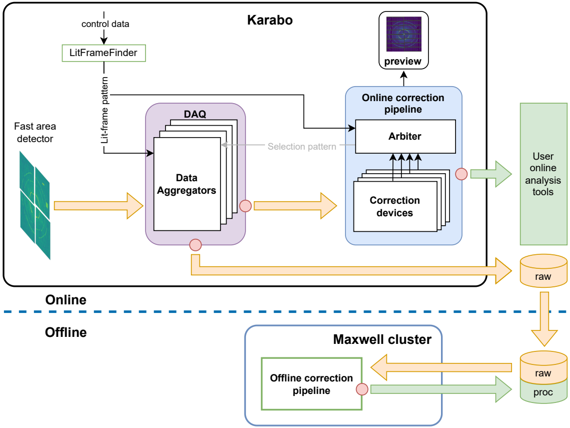
- Data annotation, transformation
- Filtering online data stream
- Small time budget!



Data infrastructure

Data acquisition system (DAQ)

- Raw data filtering
- Successfully tested at users' experiments
- Aiming at implementing the feedback from online correction pipeline



Data reduction methods

- Operation-specific methods

  are related to instrument operation itself, no analysis of detector images is usually required.

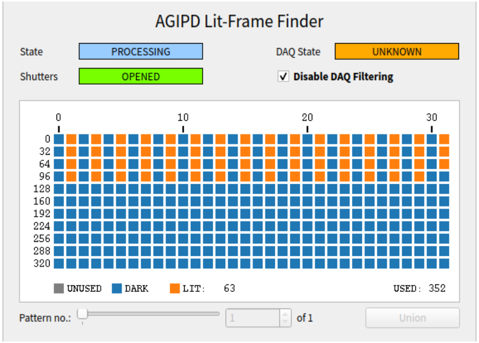
These methods are robust, low risk, and the feedback latency is compatible with online requirements.

- Technique-specific methods

  require analysis of detector data, and typically involves tuning of certain parameters.

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

Lit-frame selection



AGIPD Lit-Frame Finder

State: PROCESSING
DAQ State: UNKNOWN
Shutters: OPENED
☑ Disable DAQ Filtering

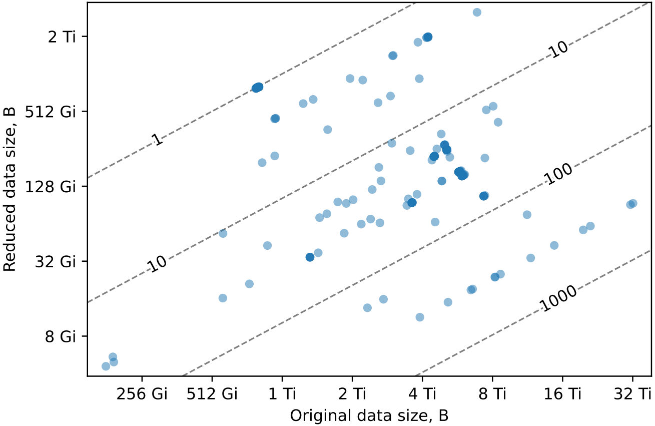
	0										10										20										30	
0	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧
32	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧
64	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧
96	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟧	🟦	🟦
128	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
160	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
192	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
224	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
256	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
288	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
320	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦

◼ UNUSED 🟦 DARK 🟧 LIT: 63 USED: 352

Pattern no.: 1 of 1

- Karabo device *LitFrameFinder* annotates detector frames
- The annotation is used to filter data at:
  - DAQ
  - Online correction
  - Offline correction
- The annotation can be used for retroactive reduction of stored data

Lit-frame selection results



Other operation-specific methods

- Module selection
  - Ratio: 1-16
  - EXDF-tools

- Gain suppression
  - Ratio: 2
  - EXDF-tools

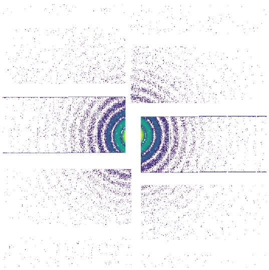
- Shutter-based train selection
  - Ratio: 10-1000
  - EXDF-tools, Offline correction

Compression

Low intensity scattering

Conversion and rounding to photon counts

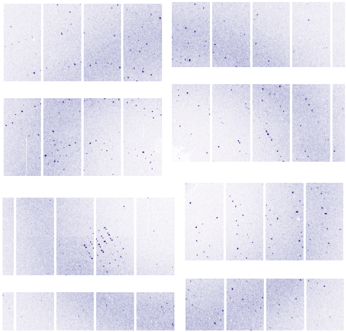
XPCS, Bragg CDI, SPI



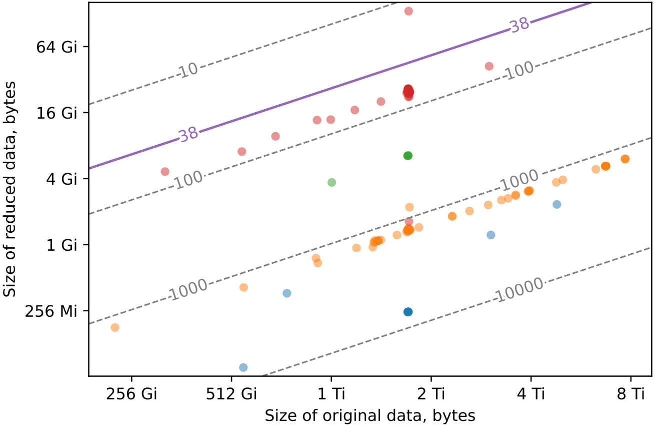
High intensity scattering

Rounding to a few highest significant bits

SFX



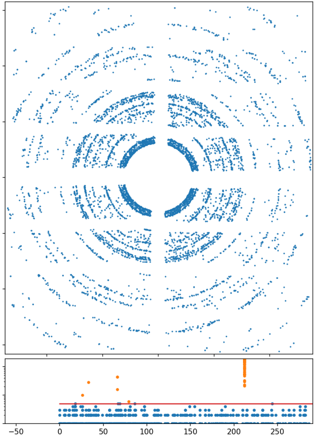
Rounding and compression

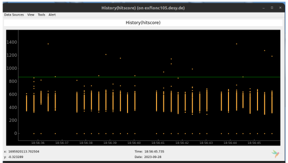


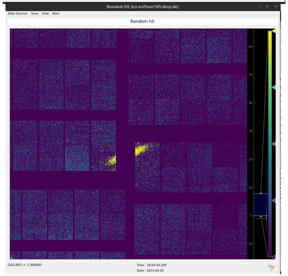
Hit-finding

- SFX hit-finding @ EXtra-xWiz
- SFX and SPI hit-finding @ online correction pipeline

  user teams: Jonas Sellberg, Filipe Maia, Xavier Paulraj, Kartik Ayyer







Other technique-specific methods

- Azimuthal integration
  - SAXS/WAXS, Powder diffraction, XPCS
  - Ratio: 1000
  - Standalone tools

- Correlation functions
  - XPCS, XCCA
  - Ratio: 1000
  - Standalone tools

- Event reconstruction: reaction microscope
  - COLTRIMS
  - Ratio: 1000
  - Offline correction

Results: applied reductions

Avoided storage of 7.4 PiB

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

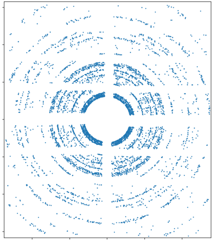
Results: candidates to retroactive reduction

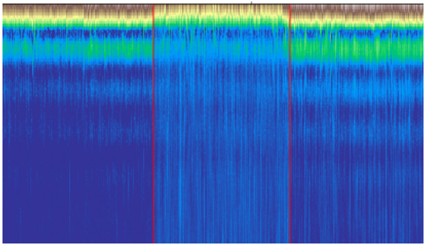
17 PiB expected to be freed

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

Validation

- Pass-thru small percent for data filtered out
- Design and automatically compute quality metrics





Documentation

- User documentation for data reduction at European XFEL
  `https://rtd.xfel.eu/docs/data-reduction-user-documentation/en/latest/`
- EuXFEL Data Analysis User Documentation
  `https://rtd.xfel.eu/docs/data-analysis-user-documentation/en/latest/`
- Data Analysis group web-page
  `https://www.xfel.eu/data_analysis/`
- Article “Data reduction activities at European XFEL: early results” submitted in Frontiers in Physics

Conclusions

- Data reduction is enabled in the new Scientific Data Policy.
- The development and integration of tools and services has been started a few years ago, together with you.
- Some data reduction methods have been extensively applied.
- You decide, we empower you by providing tools, validation, and support.
- Come to the round table, and work with us to shape services that are useful for you!

Acknowledgements

Many people inside and outside the facility have contributed shaping ideas and activities so far, in particular

- EuXFEL Users: Jonas Sellberg, Duane Loh, Filipe Maia, Xavier Paulraj, Kartik Ayyer, and many others
- EuXFEL Instrument scientists: Johannes Möller, Johan Bielecki, Ulrike Bösenberg, Jayanath Koliyadu, and many others
- EuXFEL Data department
- Participants of LEAPS-INNOV WP7

Thank you for you 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`