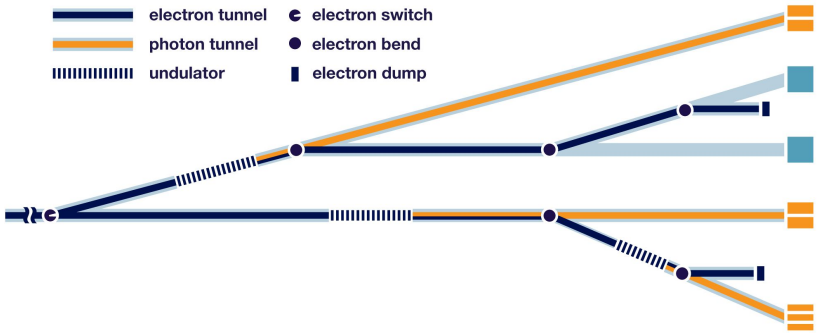
Data reduction activities at European XFEL: early results

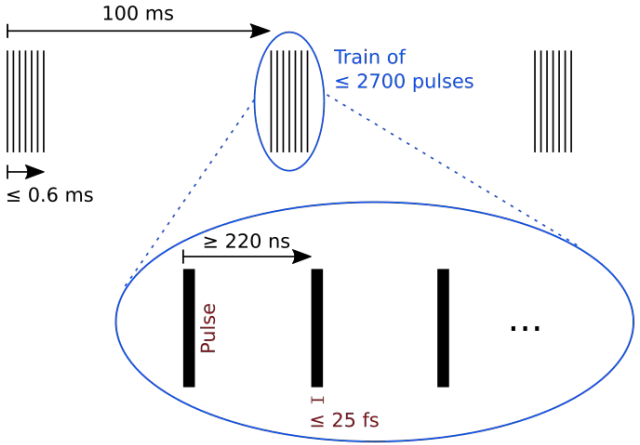
Philipp Schmidt
Data Analysis, European XFEL

On behalf of **many** others:
Egor Sobolev, 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, 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

19 March 2024

Facility overview



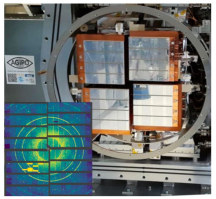
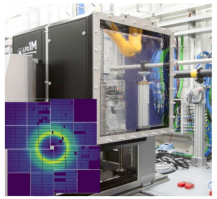
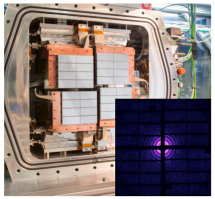


- 3 beamlines covering soft & hard X-rays
- 2-3 instruments per beamline, 7 in total
- Multiple endstations per instrument
- Pulses split across beamlines, all three operating at the same time

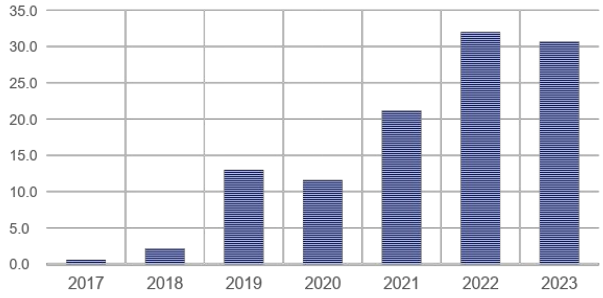
- Burst mode similar to FLASH
- 10 Hz trains with ≤ 2700 pulses each split across all beamlines
- Typically each instrument receives hundreds of pulses

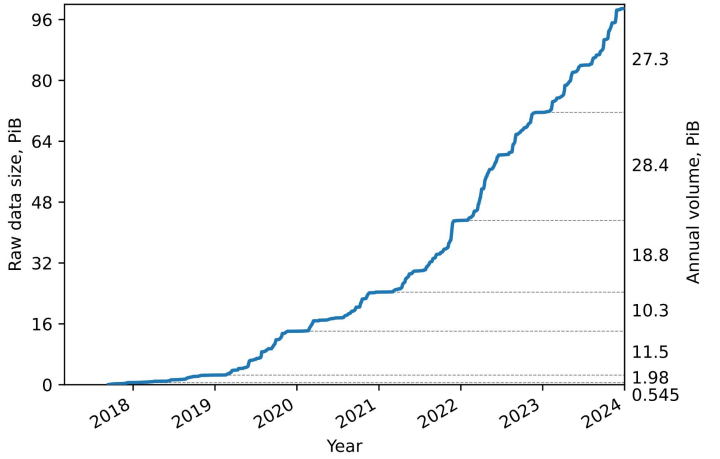
Growing Big Data challenges

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

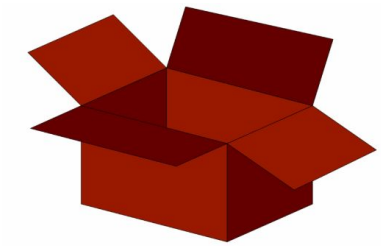
AGIPD-1M	LPD-1M	DSSC-1M	AGIPD-4M
			?
≥ 21 h/PiB	≥ 29 h/PiB	≥ 19 h/PiB	≥ 6 h/PiB

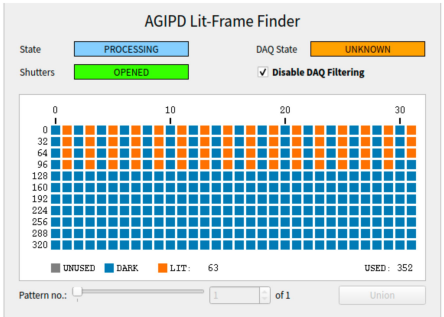
- Common bias towards storing raw data
- Growth of raw data production is **unsustainable**
- Upcoming upgrades:
  - AGIPD-4M detector
  - Duty cycle increasing up to 50%

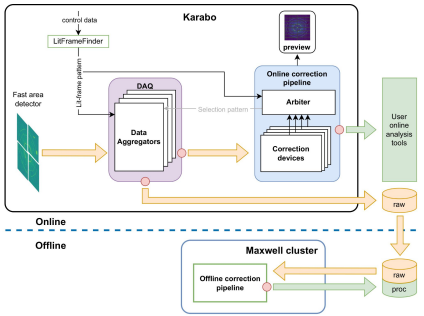




- Scientific Data Policy and the RED box
- Data reduction methods and current pilot projects
- Data reduction integration points online & offline

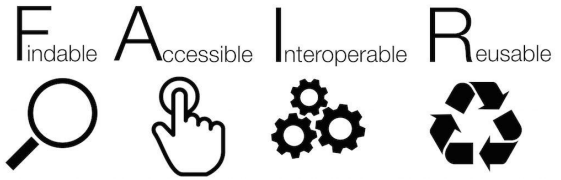




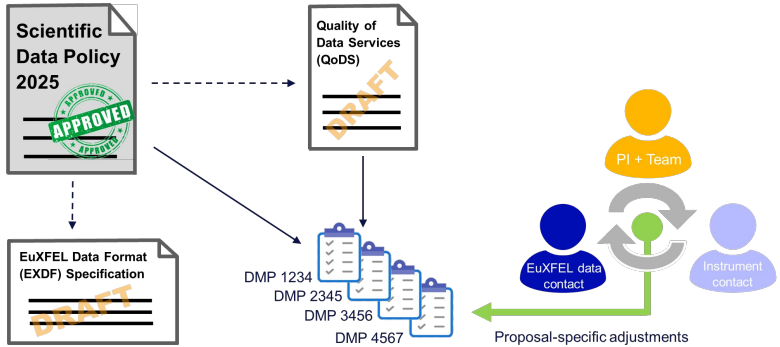


Scientific Data Policy 2025+

- 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).



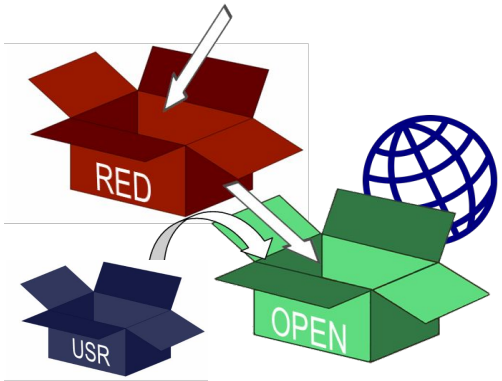
RED box and OPEN data

Please do not quote these exact numbers yet!

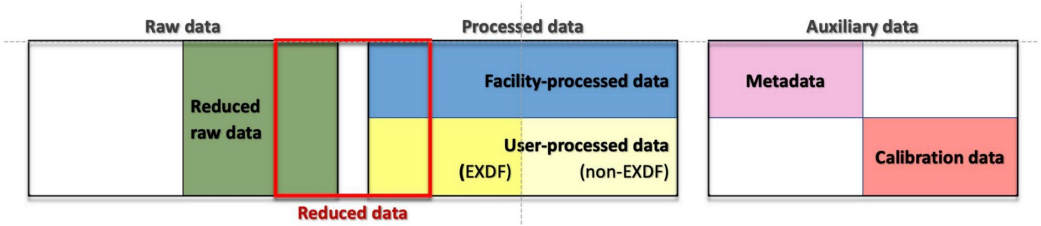
- The **size of raw data** determines the data volume retained **long-term** and **opened up** later:

  `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

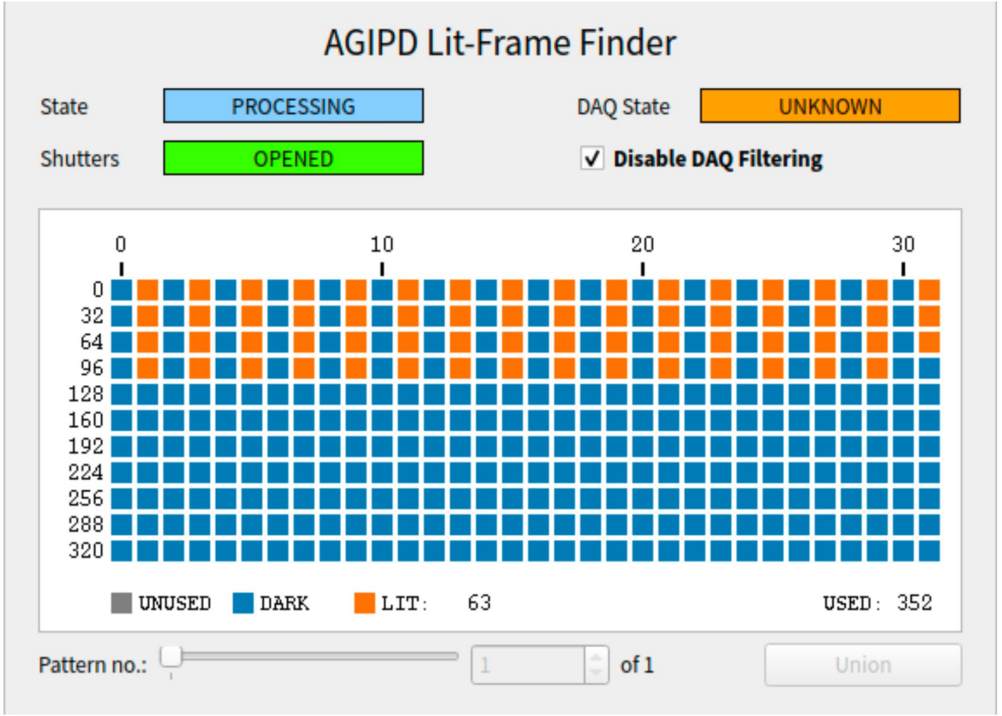
- Technique-specific methods
  Require analysis of experimental data, and typically involve tuning of certain parameters

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

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

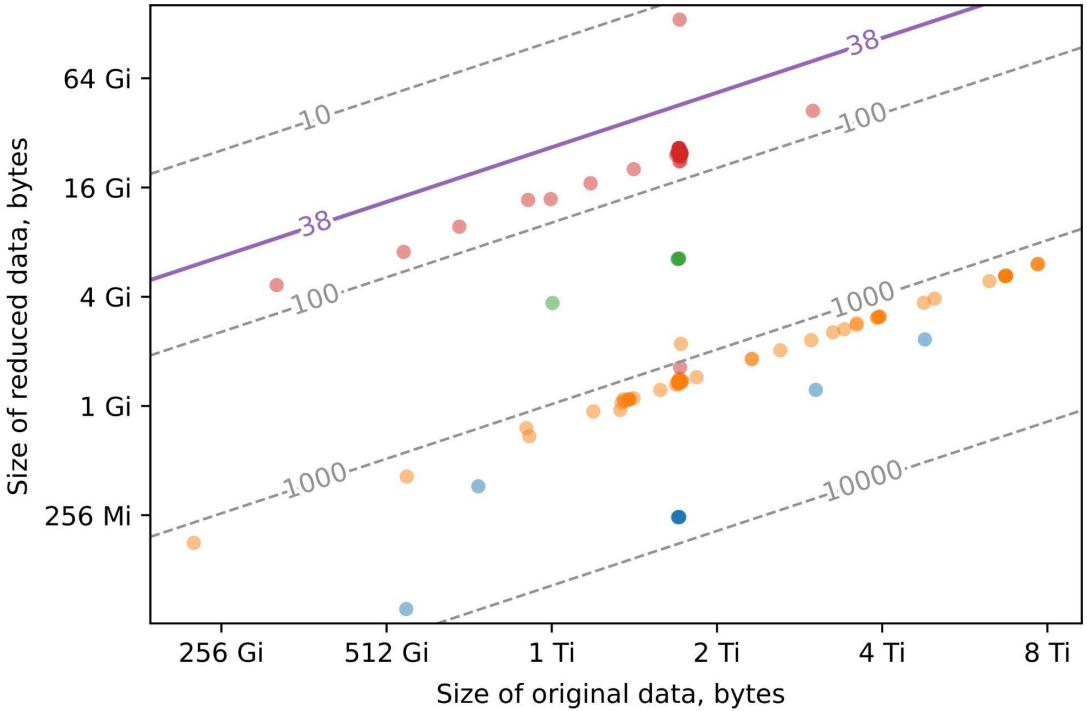
Lit frame selection

- 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



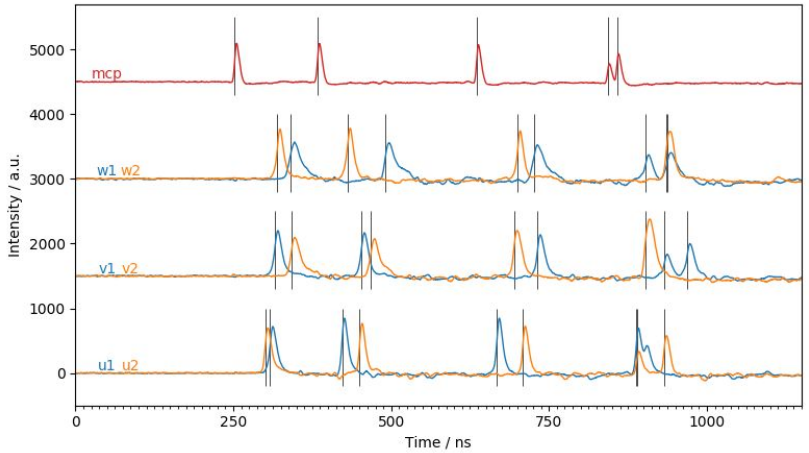
Typical results for XPCS experiments

- 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



Delay line detector event reconstruction

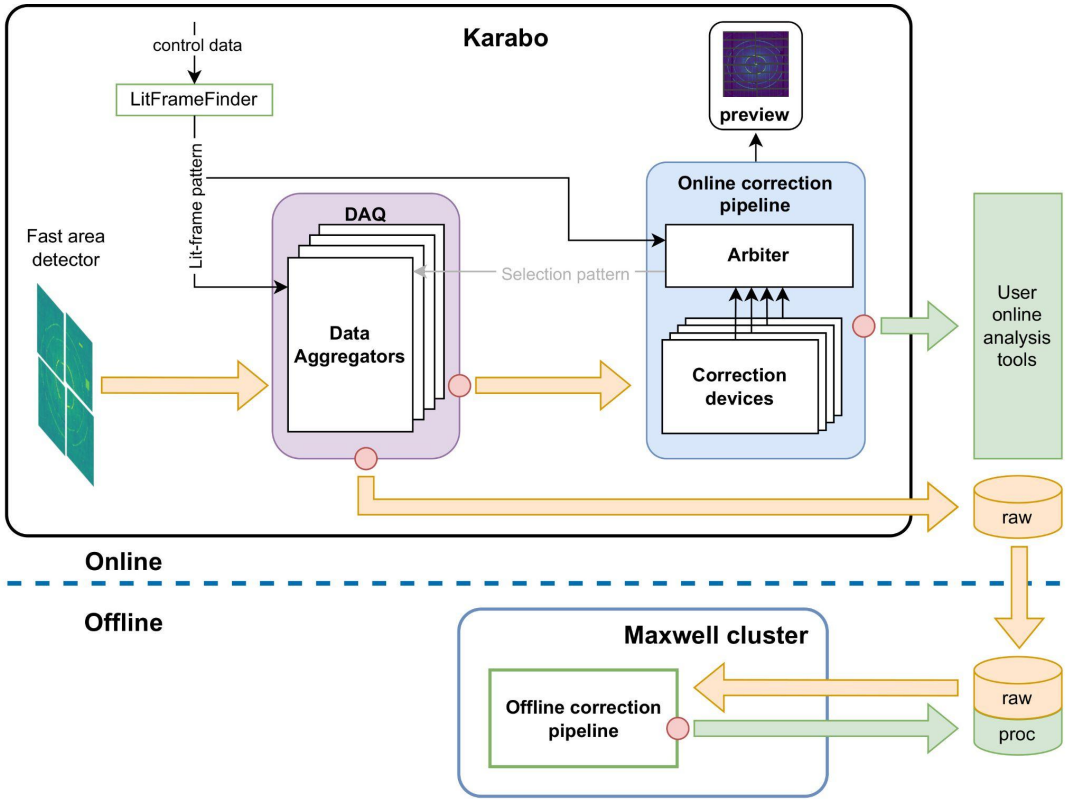
- Delay line detectors for REMI/COLTRIMS and RIXS sampled with GHz digitizers
  - Uncompressed raw data rate of ~3 Gb/s at current duty cycle of 0.6%
- On-FPGA zero suppression of analog signal allows reduction by ~50x
- Digitization and event reconstruction to `(x,y,t)` tuples reduces data by another >40x, compared to raw >2000x.




```
(-46.99, -30.75, 1277.01, 0)
(45.44, 7.41, 1678.84, 3)
(38.30, 11.44, 3679.09, 1)
(-27.46, -23.33, 6249.36, 15)
(-33.62, -19.56, 6249.64, 19)
```

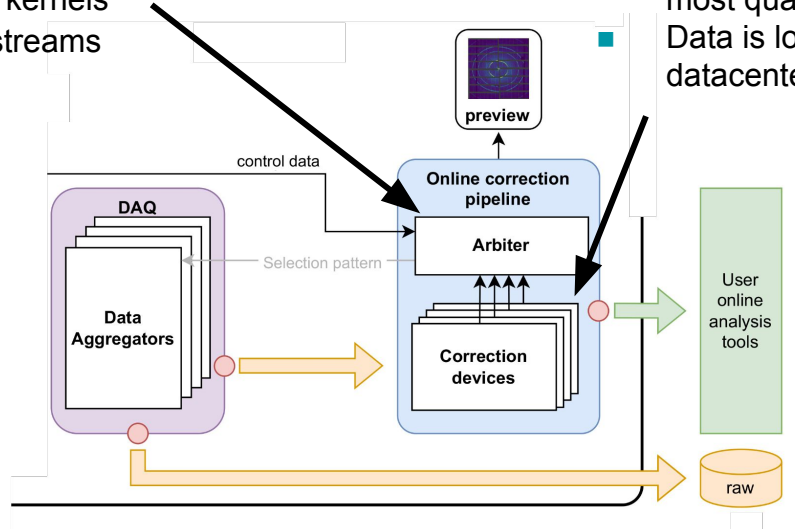
Data reduction integration points

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



Online correction addons and arbiter kernel

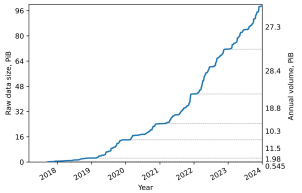
- Data reduction decision in central ***arbiter***
  - Combine metadata across all modules & components
  - Predefined reduction or custom kernels
  - Used to filter in DAQ or output streams
- Compute additional results in ***correction addon***
  - Must run per module or at most quadrant
  - Data is local on datacenter-grade GPUs
- Users can configure provided implementations or integrate their own fully custom code



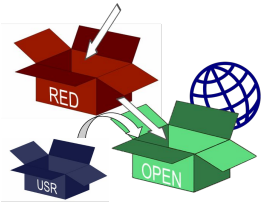
Extendable reduction of EXDF-structured data

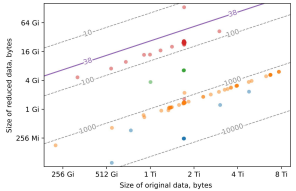
- Tools to **semantically** reduce and compare already recorded data offline
- Seamlessly maintain EXDF data structure and compatibility
- Reduction expressed by extendable series of operations on data or its structure
  - `select-trains, select-entries, subslice-keys, compress-keys, …`
- Reproducible and serializable representation
- Used to reduce multiple prior proposals, extendable by users

```
xfel.eu data-reduction-recipe v1.0

- AgipdGain <exdf.data_reduction.builtins.AgipdGain>
rechunk-keys      SPB*AGIPD1M*:xtdf image.data        (-1, 1, None, None)
subslice-keys     SPB*AGIPD1M*:xtdf image.data        [0, :, :]

- ManualPattern <exdf.data_reduction.builtins.ManualPattern>
select-entries    SPB*AGIPD1M*:xtdf image             by_id[:]    [10:60]

- PpuTrainSequences <exdf.data_reduction.builtins.PpuTrainSequences>
select-trains     SPB*AGIPD1M*:xtdf          by_id[111431123]
select-trains     SPB*AGIPD1M*:xtdf          by_id[111431433]
select-trains     SPB*AGIPD1M*:xtdf          by_id[111431713]

- SomeUserMethod <user.package.module>
...
```



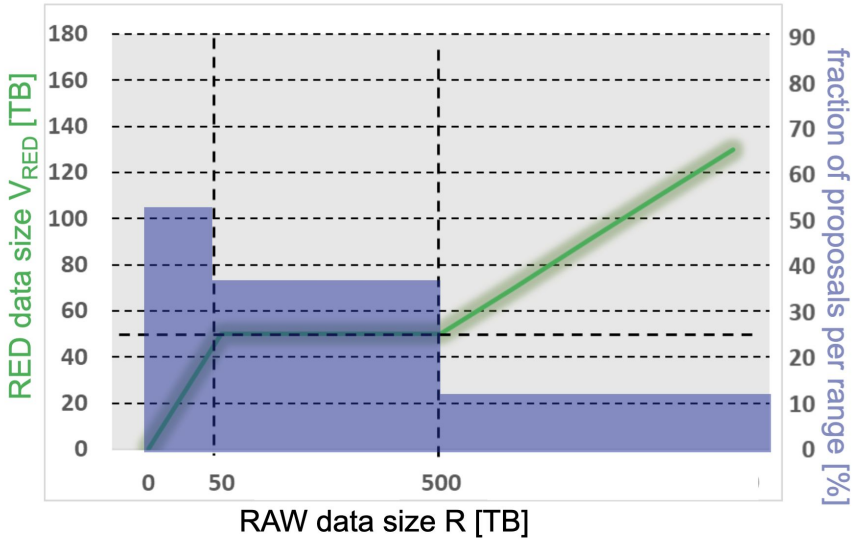
- **Retaining full raw data** is **not sustainable** and future upgrades make recording impossible
- **Scientific Data Policy 2025+** makes data reduction a first-class citizen of scientific data curation and management
- **Facilities** must develop and **provide** operation- and technique-specific **reduction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

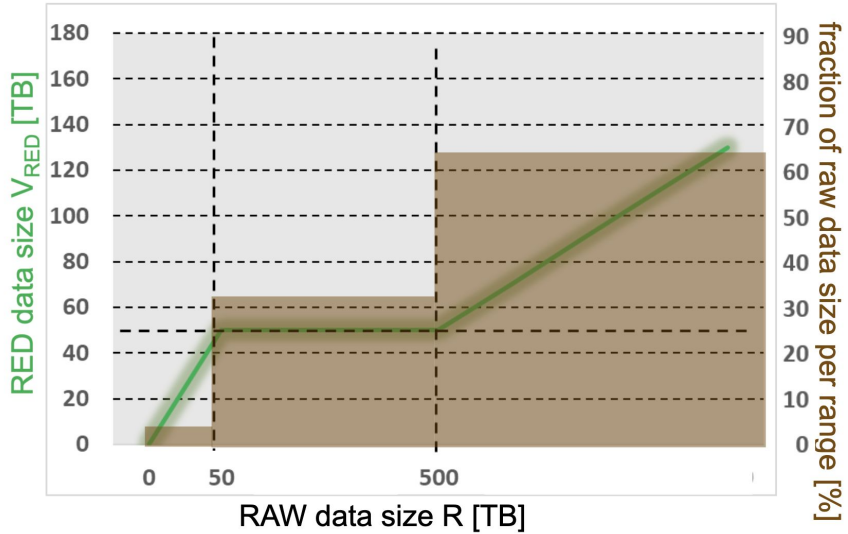




- Special thanks to European XFEL users joining in pilot projects: *Jonas Seilberg, Duane Loh, Filipe Maia, Xavier Paulraj, Kartik Ayyer and many others*
- Sobolev, Schmidt et al: *Data reduction activities at European XFEL: early results*, Front. Phys. **12**, 1331329 (2024)

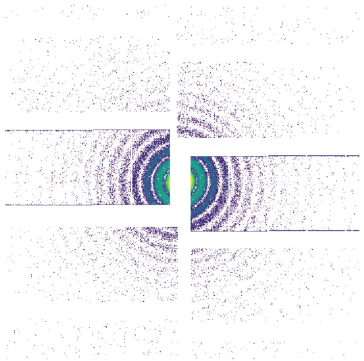
Proposal size distributions



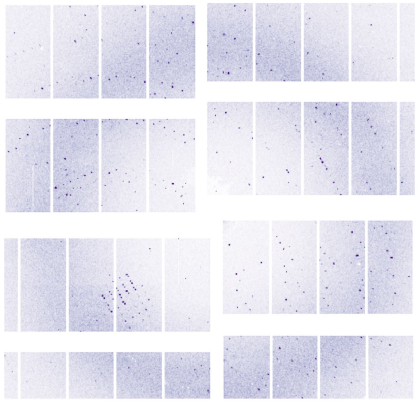


Compression

- 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

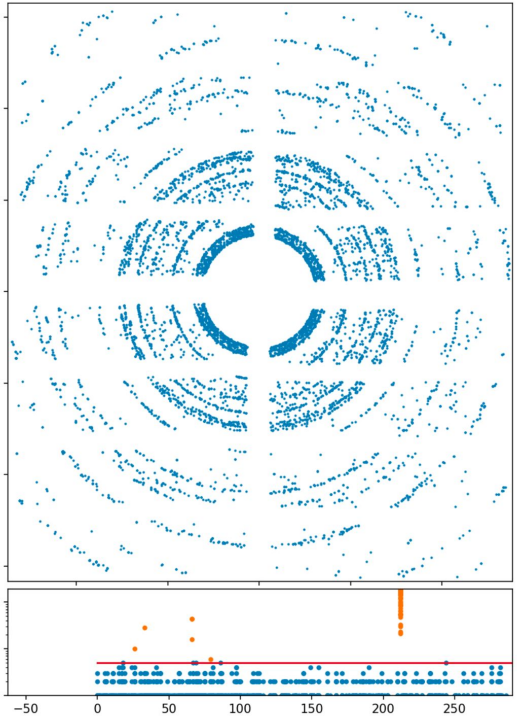


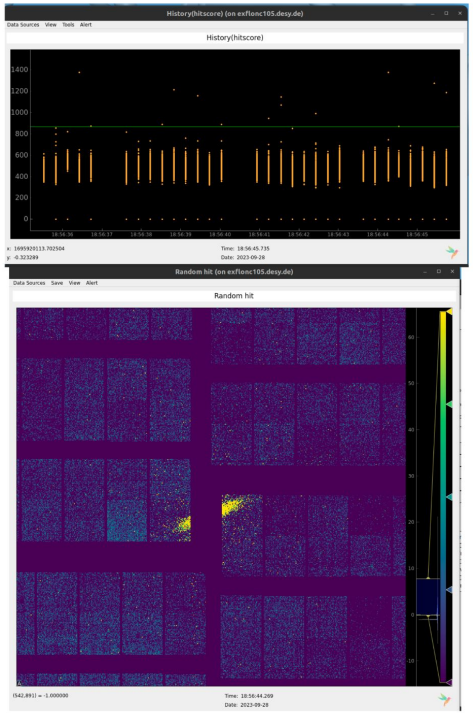
  - High intensity scattering
    Rounding to few highest significant bits
    SFX



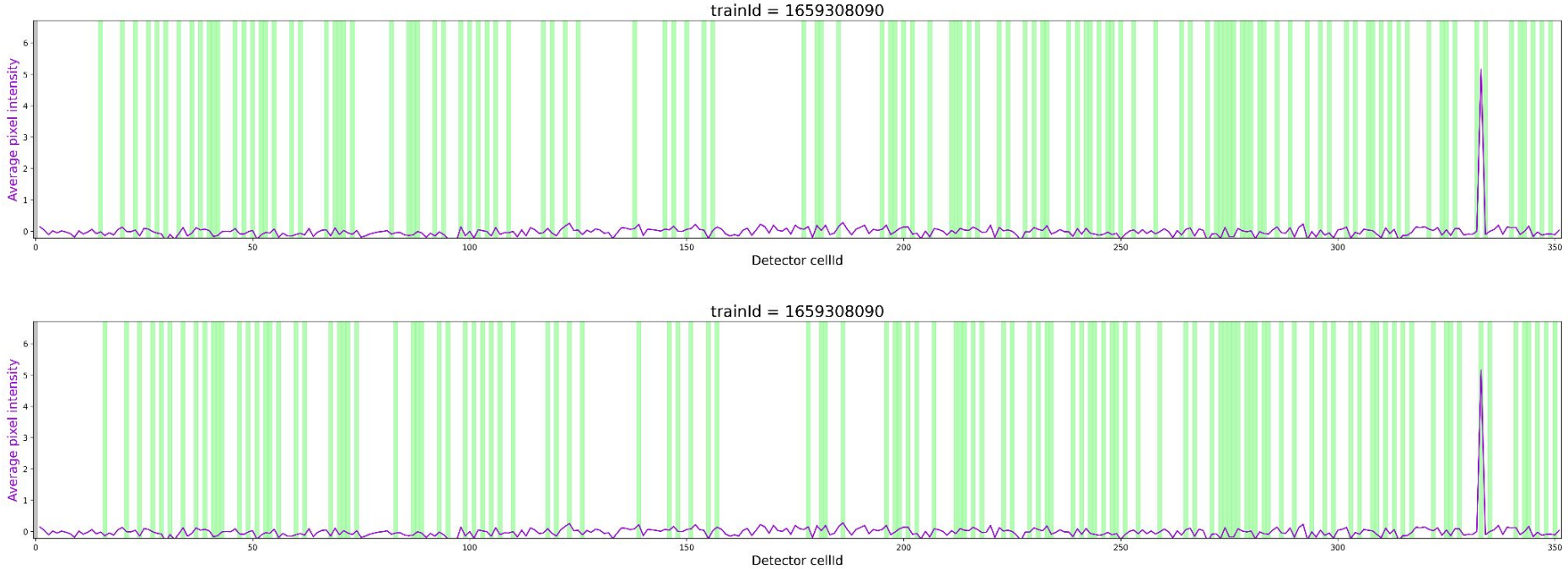
Hit finding

- SFX hit-finding offline by EXtra-Xwiz
  Turkot et al, Crystals 13, 1533 (2023)
- SFX and SPI hit-finding online in user tools and as part of online preview
- Pilot ongoing to apply user-supplied candidate frame lists to saved data





Validation of user-supplied candidate frame lists



Online correction addons and arbiter kernel

- *Correction addon* running within the online correction pipeline
  - Distributed across online cluster resources
  - Data is local on datacenter-grade GPUs
  - May morph correction result as needed
  - Compute statistics and attach metadata
- *Arbiter kernel* running in central arbiter
  - Combines metadata across all detectors
  - May take data reduction decision and feed back to DAQ or to filter output streams
- Configure and combine provided implementations or integrate your own fully custom code