LEAPS-Innov WP7.2 (M7-M18)

"Assessment of **future needs** and **development of metrics** for data compression and reduction"

Task coordinators:

- **Nicolas Soler** (ALBA)
- **Vincent Favre-Nicolin** (ESRF)

"Information on site-specific scientific needs of experiments at LEAPS facilities will be gathered through **interviews with domain experts and scientists** taking into account the past and future needs: **bandwidth, latency, connectivity, available storage volume and compute infrastructure**.

Downstream processing and analysis needs will be identified to serve as **upper boundary limits for information loss during compression**.

Detector manufacturers will be involved in the discussion to ensure the compatibility, and future integration in production, of the resulting components."

“Research will also be conducted to **identify the techniques which generate the most data** and to develop metrics which assess the effect of data reduction on the quality of the final result. The most likely candidates are **serial crystallography and single-particle imaging**, multiple types of **tomography**, and most **high-resolution imaging** techniques.

Metrics will be developed to evaluate the data explosion problem and to assess the **impact of the data reduction and compression solutions** developed in task 7.3. The performance of the various reduction strategies will be investigated, with the goal of **achieving close to real-time performance** in data reduction.

This task will be carried out in consultation with **scientists and external industrial experts** in data reduction. Contacts have been already established with industrial experts in data compression and specifically the Blosc community, IBM and StreamHPC”

LEAPS-Innov WP7.2 (M7-M18) D7.2

D7.2 Report on metrics for data reduction and compression (M18)

M18 = end of September 2022 !

Time is short to define metrics…

Need volunteers for the different parts !

NB: this presentation is intended for discussion, so please interrupt !

Data *reduction* metrics

This can either be:

- 2D->1D azimuthal integration (powder diffraction, SAXS..)
- 2D -> single crystal integrated intensities

The biggest issue for the community is the choice of keeping the raw data or not.

After discussion: there are still cases where metrics should be introduced to evaluate the information loss, e.g. comparing 2D->1D->2D

Data compression metrics

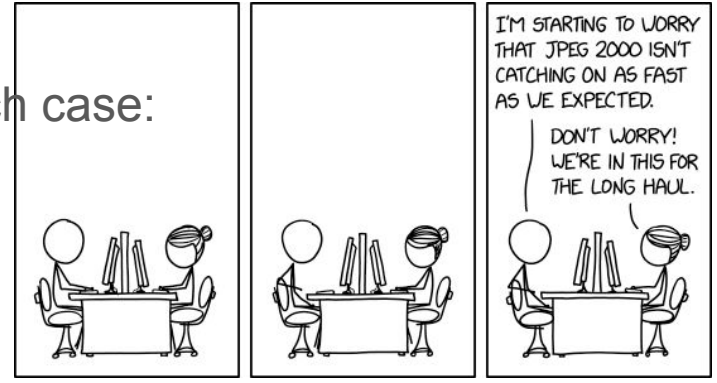
‘Obvious’ performance metrics:

- (de)compression speed:
  - this is simple to define/measure and should be compared with other relevant figures (I/O and processing times)
  - we should give a few tables with example compression times for
    - established compression methods (gzip, lz4, blosc, jpeg2000…)
    - chosen techniques: 2D (raw projections) and 3D tomography, sparse data (XPCS),...
- compression ratio
  - again rather trivial to define, but it’s useful to have a few figures
- Volunteers for the above ? Create a benchmark table, ideally a python script which auto-generates the results and can be reproduced.
- It would be interesting to include some hardware-accelerated results (power9, FPGA, etc..)

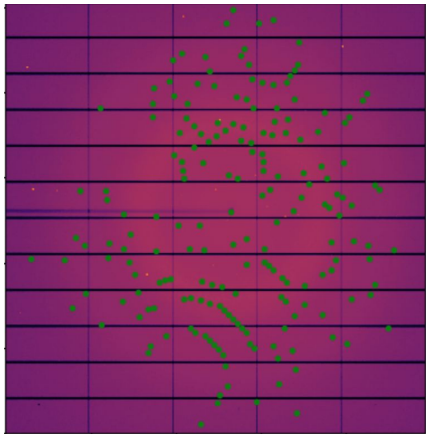
lossy compression metrics: *imaging*

This is more interesting and more elaborate. As per the proposal text, we may focus on relevant techniques (imaging, serial crystallography :

- Imaging base metrics:
  - resolution from knife-edge scan on known edge (using simulated data ?)
  - resolution from FSC/FRC
  - for both: comparison with simulated data or caomparison between un/compressed data ?
- tomography:
  - compression of raw images (unphased)
  - compression of final 3D volume
- need to setup scripts/analysis pipelines for each case:
  - basic knife-edge/FSC analysis functions
  - simple 2D or 3D compression & comparison
  - complete phasing+tomo pipeline
- Volunteers ?



lossy compression metrics: *serial Xtallography*



Compression in (S)SX can be done by on-the-fly:

- data reduction and complete integration
- on-the-fly lossy data compression (sigma-clipping)

In the latter case, the data is integrated from lossy-compressed images

Metrics: usual ones for MX are available

- CC
- Rmerge, Rfree
- etc..

this is a less well-defined case as the relevant metrics really depend on the analysis pipeline, if the lossy-compressed data actually are used for analysis or not. Compression may be only relevant for the archival of raw images.

Conclusion: timeline & volunteers

- report to be submitted end of September
- need (by early September):
  - scripts
  - example datasets
  - final figures / tables
- Volunteers to work on the tests & deliverable data
  - data reduction (2D->1D)
  - generic compression performance
  - lossy:
    - Imaging
    - (S)SX