European XFEL

Satellite: Data Analysis at the European XFEL, January 25th

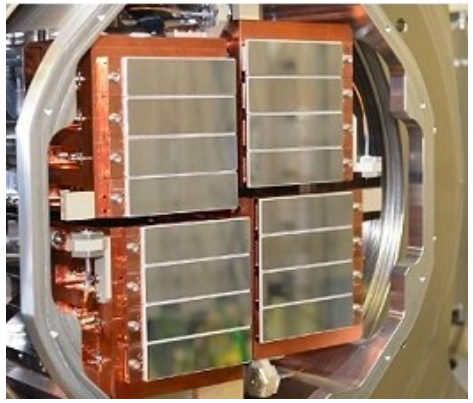
Automated SFX Analysis

Fabio Dall'Antonia, Oleksii Turkot
Data Analysis Group

European XFEL

Serial crystallography at EuXFEL

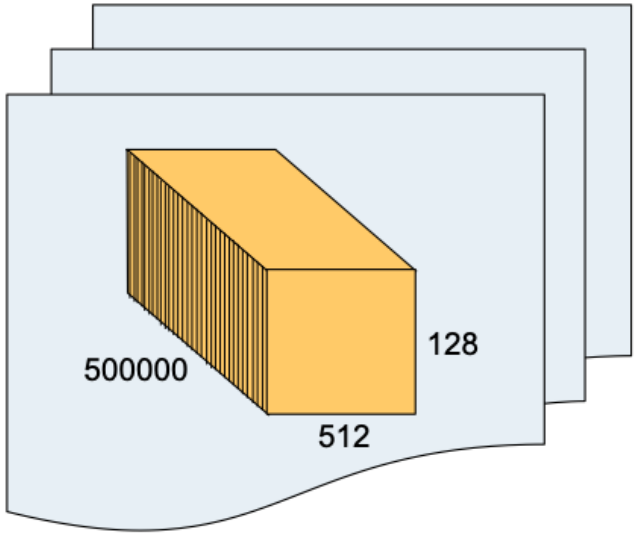
- Special case of crystallography:
  - still images
  - random orientation of crystals
  - typically low hit rate (1-10%)
- SW: CrystFEL, DIALS/cctbx
- Peculiarities at EuXFEL
  - Lots of data: for AGIPD-1M, 10^5 – 10^6 frames per run
  - MHz area detectors, but also JUNGFRAU-4M (SPB) have a modular layout with separate file writing to DAQ



Adapted from:
Wiedorn et al. (2018), Nat. Comm. 9

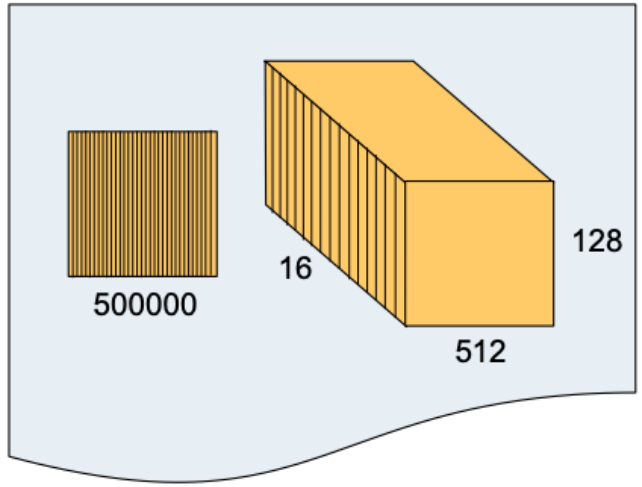
up to 3520 detector frames per second

Virtual dataset and geometry file



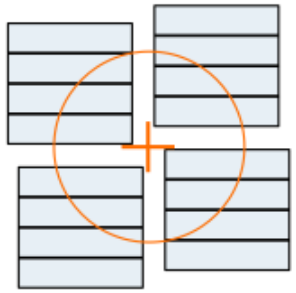
Create with EXtra-data*

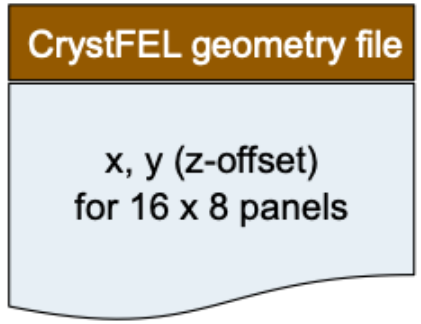
references



single HDF5 file: virtual dataset (~50 MB)

16 (sequences of) HDF5 files in a run folder, several TB in total

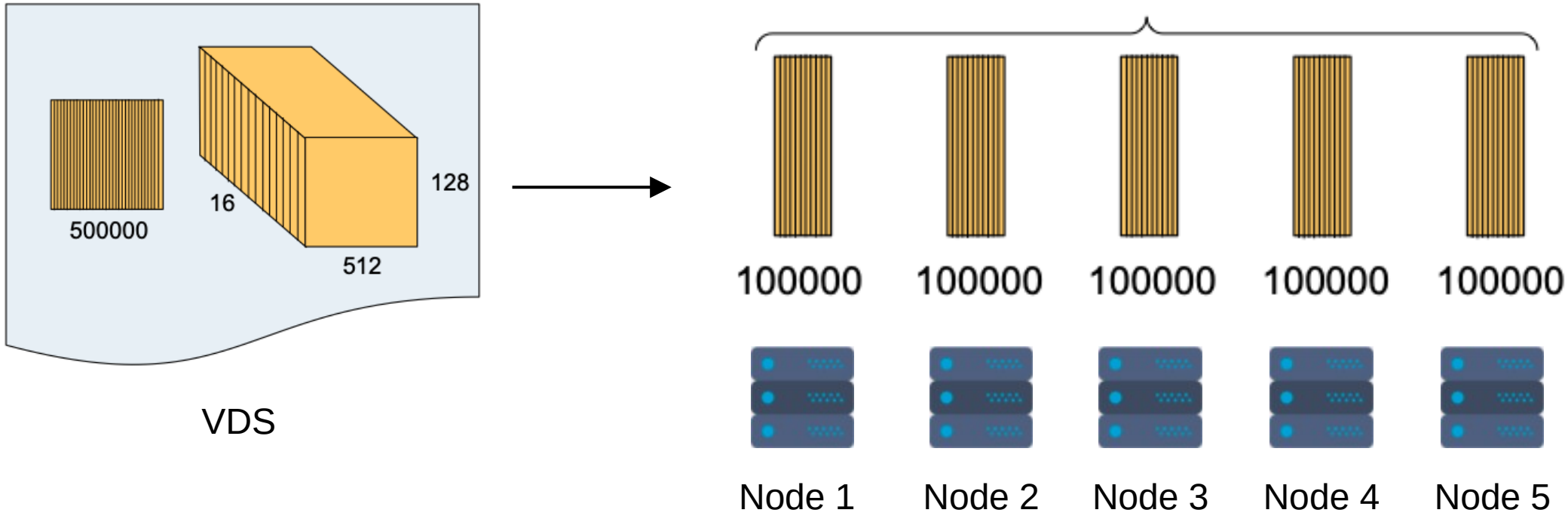




** cf. offline data analysis tutorial by Thomas Kluyver*

Distributed HPC computing

- CrystFEL features multi-processing (multiple workers) on a single computer
- Independent nature of frames → put into batches with subsets for parallel processing on multiple cluster nodes



What is Xwiz?

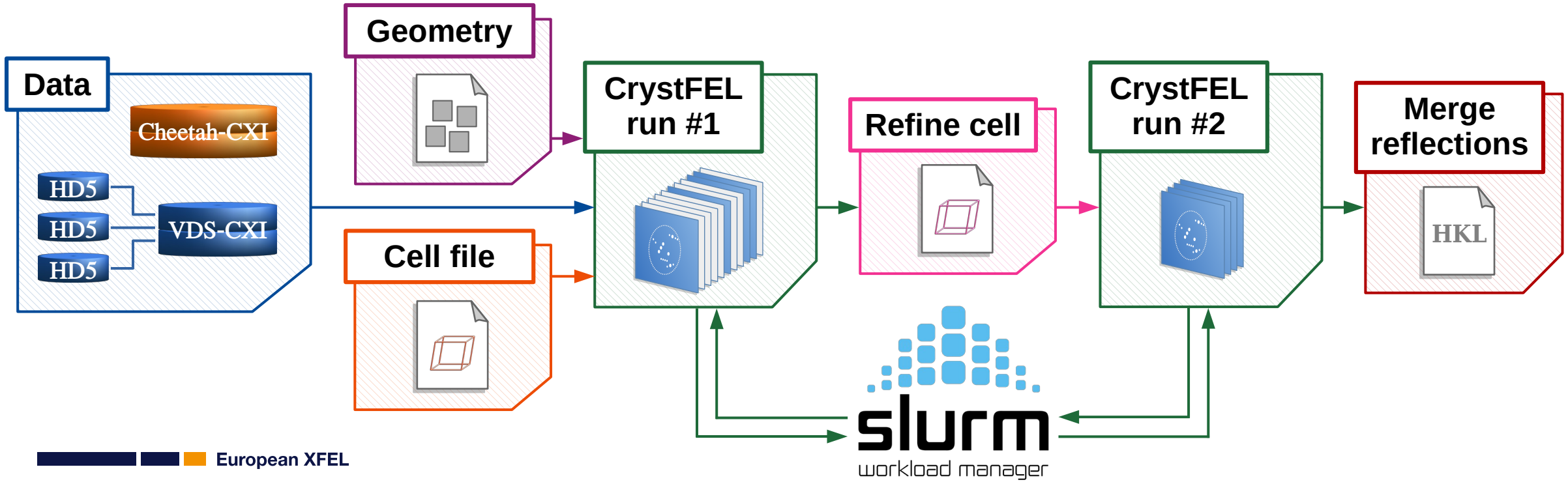
Key concepts:

Automatic

Robust

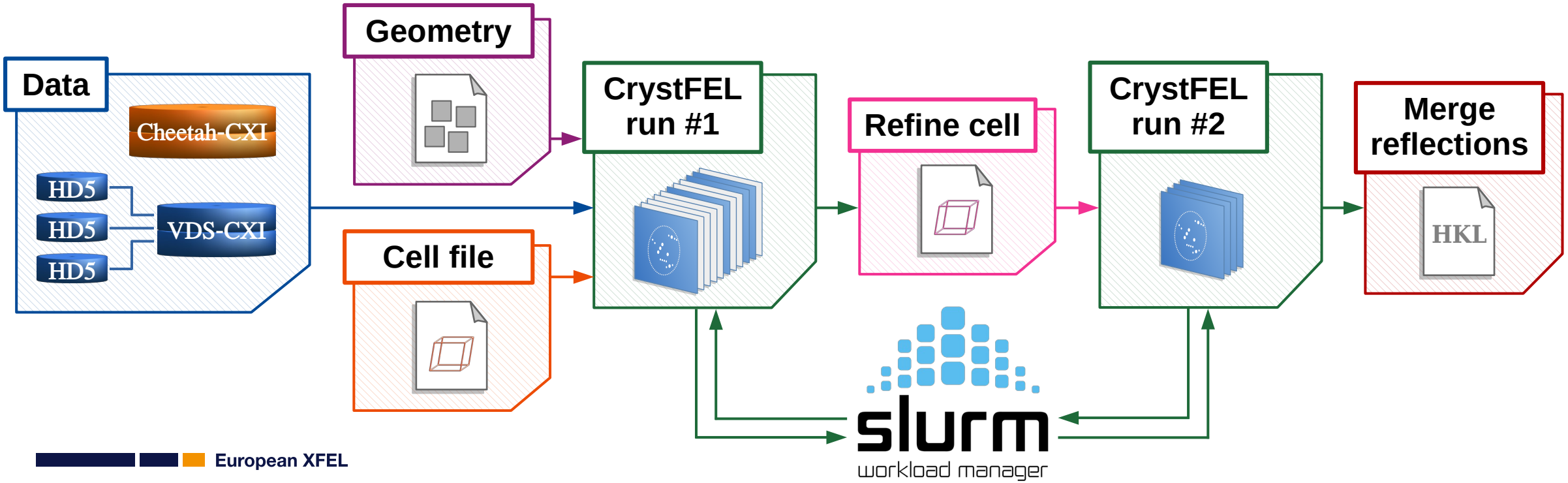
Reproducible

Extra-Xwiz is a framework for automated "offline" processing of the serial crystalography data.



What is Xwiz?

- Semi-automatic pipeline, interactive review of the workflow configuration is possible
- Uses CrystFEL – wrappers around indexamajig and partialator
- Distributed computing on Maxwell using SLURM for speed-up
- Developed in joint effort with SPB/SFX and CFEL

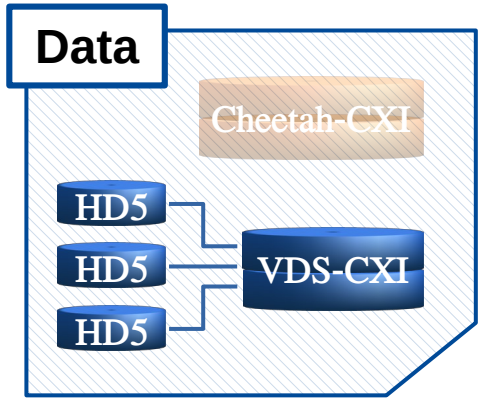


How to start using Xwiz?

```
[max-exfl001]~% module load exfel EXtra-xwiz
 - EXFEL modulepath enabled
(extra-xwiz_env) [max-exfl001]~% mkdir test_xwiz ; cd test_xwiz
(extra-xwiz_env) [max-exfl001]~/test_xwiz% xwiz-workflow -h
usage: xwiz-workflow [-h] [-a] [-d] [-r] [-p] [-c] [-adv]

optional arguments:
  -h, --help            show this help message and exit
  -a, --automatic       enable auto-pipeline workflow (skip configuration
                        review)
  -d, --diagnostic      keep SLURM stdout captures for diagnoses in case of
                        problems
  -r, --reprocess       enter workflow at the re-processing stage (refined
                        unit cell and frame selection exist)
  -p, --peak-input      use a CXI data set file from Cheetah including pre-
                        localized peaks
  -c, --cheetah-input   skip VDS generation and assemble input from Cheetah
                        folder contents
  -adv, --advance-config
                        Generate an advanced config instead of the base one.
(extra-xwiz_env) [max-exfl001]~/test_xwiz% xwiz-workflow
Configuration file is not present, will be created.
Configuration written to /home/turkot/test_xwiz/xwiz_conf.toml
Please rerun now.
(extra-xwiz_env) [max-exfl001]~/test_xwiz% ls
xwiz_conf.toml
(extra-xwiz_env) [max-exfl001]~/test_xwiz%
```

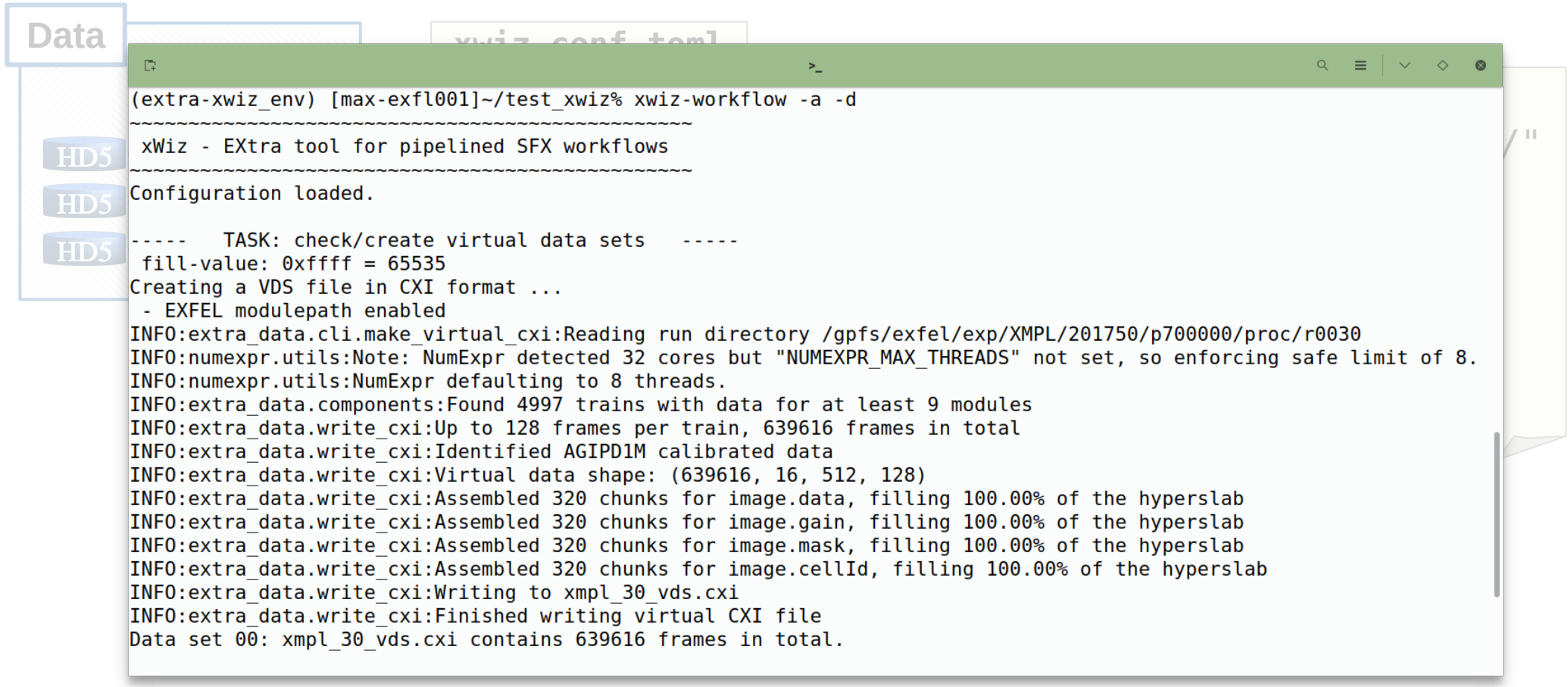
Prepare the data



xwiz_conf.toml

```
[data]
path = "/gpfs/exfel/exp/XMPL/201750/p700000/proc/"
runs = "30"
n_frames_percent = 100
n_frames_total = 300000
vds_names = "xmpl_30_vds.cxi"
vds_mask_bad = "0xffff"
list_prefix = "xmpl_30"
```

Prepare the data



Detector geometry and unit cell parameters

Geometry

xwiz_conf.toml

```
[geom]
file_path =
"/gpfs/exfel/data/user/turkot/store/geom/agipd_2120_v1.geom"
template_path = ""
```

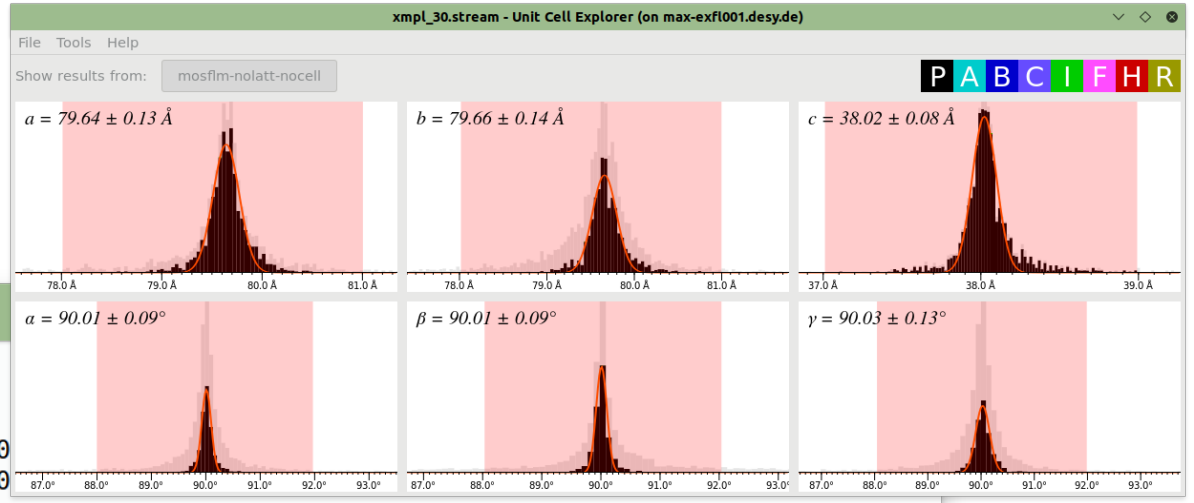
Cell file

xwiz_conf.toml

```
[unit_cell]
file = "hewl.cell"
run_refine = true
```

No cell file? No problem.

Extra-Xwiz will make an extra CrystFEL run without cell parameters and start cell explorer.




```
-----   TASK: prepare distributed computing   -----
Total number of frames to process: 300000
Split into: 10000 10000 10000 10000 10000 10000 10000 100
000 10000 10000 10000 10000 10000 10000 10000 10000 10000

-----   TASK: run CrystFEL (I)   -----
Geometry file is format-compatible to corresponding data
/gpfs/exfel/sw/software/extra-xwiz_env/lib/python3.7/site-packages/extra_xwiz/workflow.py:202: UserWarning: Processi
ng without unit cell due to invalid cell file.
  warnings.warn('Processing without unit cell due to invalid cell file.')
 Waiting for job-array 9873245
 |##################################################| 100.0%, ◆ 8764, Indexing rate: 2.9%

-----   TASK: determine initial unit cell and re-run CrystFEL
xmpl_30.stream: Loaded 8764 cells from 300000 chunks
Loaded 8764 cells from 300000 total chunks
Selected 8764 of 8764 cells
 Name of the cell file created with cell explorer > hewl_01.cell
```

CrystFEL run #1

xwiz_conf.toml

```
[slurm]
# Available partitions: 'upex', 'exfel'
partition = "exfel"
duration_all = "1:00:00"
n_nodes_all = 30
duration_hits = "0:30:00"
n_nodes_hits = 4
```

xwiz_conf.toml

```
[crystfel]
# Available versions: '0.8.0', '0.9.1', '0.10.0', 'cfel_dev'
version = '0.10.0'
```

CrystFEL run #1

xwiz_conf.toml

```
[proc_coarse]
resolution = 4.0
peak_method = "peakfinder8"
peak_threshold = 800
peak_snr = 5
peak_min_px = 1
peak_max_px = 2
peaks_hdf5_path = "entry_1/result_1"
index_method = "mosflm"
n_cores = 40
local_bg_radius = 3
max_res = 1200
min_peaks = 0
extra_options = "--no-non-hits-in-stream"
```

CrystFEL run #1

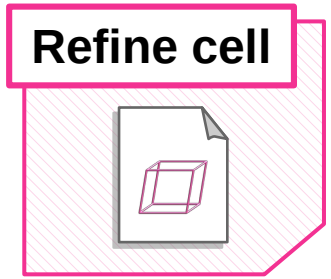
xwiz_conf.toml

```
[proc_coarse]
resolution = 4.0
peak_method = "peakfinder8"
peak_threshold = 800
n_cores = 40
local_bg_radius = 3
max_res = 1200
min_peaks = 0
extra_options = "--no-non-hits-in-stream"
```

```
>_

-----   TASK: prepare distributed computing   -----
Total number of frames to process: 300000
Split into: 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10
000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000 10000
 [cell-file check o.k.]
 Waiting for job-array 9873275
|##################################----------------| 67.1%, ◆ 2846, Indexing rate: 1.4%
```

[optionally] Refine cell parameters



xwiz_conf.toml

```
[unit_cell]
file = "hewl.cell"
run_refine = true
```

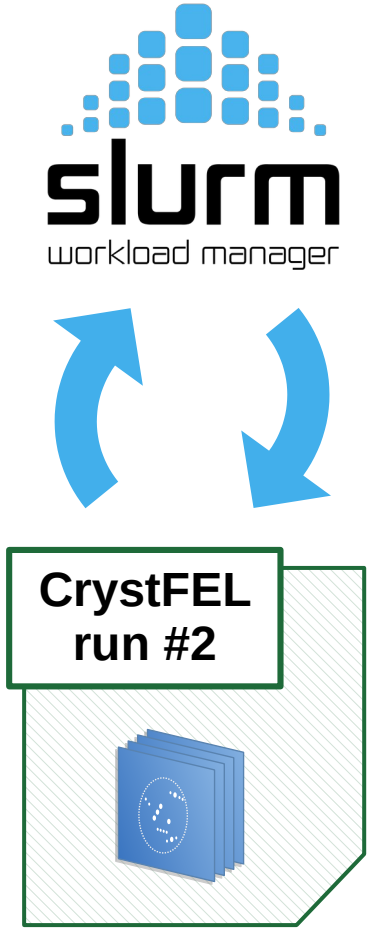
[optionally] Refine cell parameters

```
-----   TASK: filter crystal frames according to the unit cell parameters   -----
 check against CrystFEL unit-cell format:
4228 frames with (reasonable) crystals found
Overall indexing rate is 0.014093333333333333

-----   TASK: refine unit cell parameters   -----
Distribution for a
(2114, 79.6542160832545, 0.3914363564078666)
 77.54
 78.08
 78.61 ■
 79.15 ■■■■■■■■■
 79.69 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
 80.23 ■■■■■■
 80.76 ■
 81.30
 81.84
 82.38
[3.02305214e+03 7.96616646e+01 2.82309882e-01]

Distribution for b
(2114, 79.69934351939452, 0.4005336516808145)
 77.67
 78.32
 78.97 ■■■■■
 79.62 ■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■■
 80.27 ■■■■■■■■■
```

CrystFEL run #2



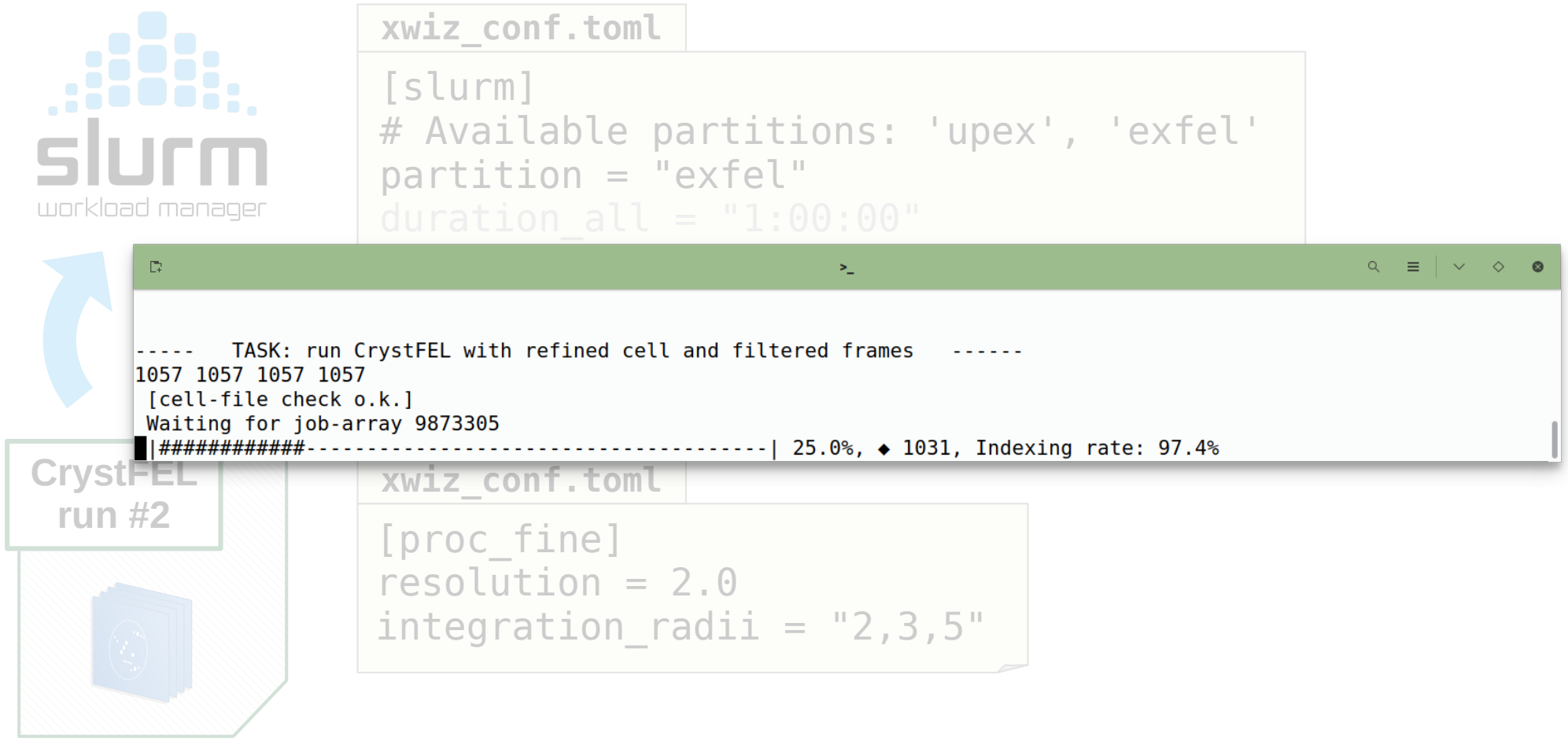
xwiz_conf.toml

```
[slurm]
# Available partitions: 'upex', 'exfel'
partition = "exfel"
duration_all = "1:00:00"
n_nodes_all = 30
duration_hits = "0:30:00"
n_nodes_hits = 4
```

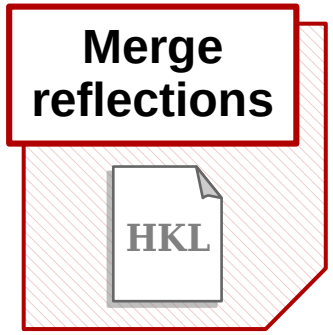
xwiz_conf.toml

```
[proc_fine]
resolution = 2.0
integration_radii = "2,3,5"
```

CrystFEL run #2



Scale and merge reflections



xwiz_conf.toml

```
[merging]
point_group = "422"
scaling_model = "unity"
scaling_iterations = 1
max_adu = 100000
```

```
-----   TASK: scale/merge data and create statistics -----
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
 Workflow complete.
 See: xmpl_30.summary
(extra-xwiz_env) [max-exfl001]~/test_xwiz% ls partialator
hewl_01.cell_refined  xmpl_30_ccstar.dat        xmpl_30_hits.stream  xmpl_30_merged.hkl1  xmpl_30_rsplit.dat
xmpl_30_cchalf.dat    xmpl_30_completeness.dat  xmpl_30_merged.hkl   xmpl_30_merged.hkl2
(extra-xwiz_env) [max-exfl001]~/test_xwiz%
```

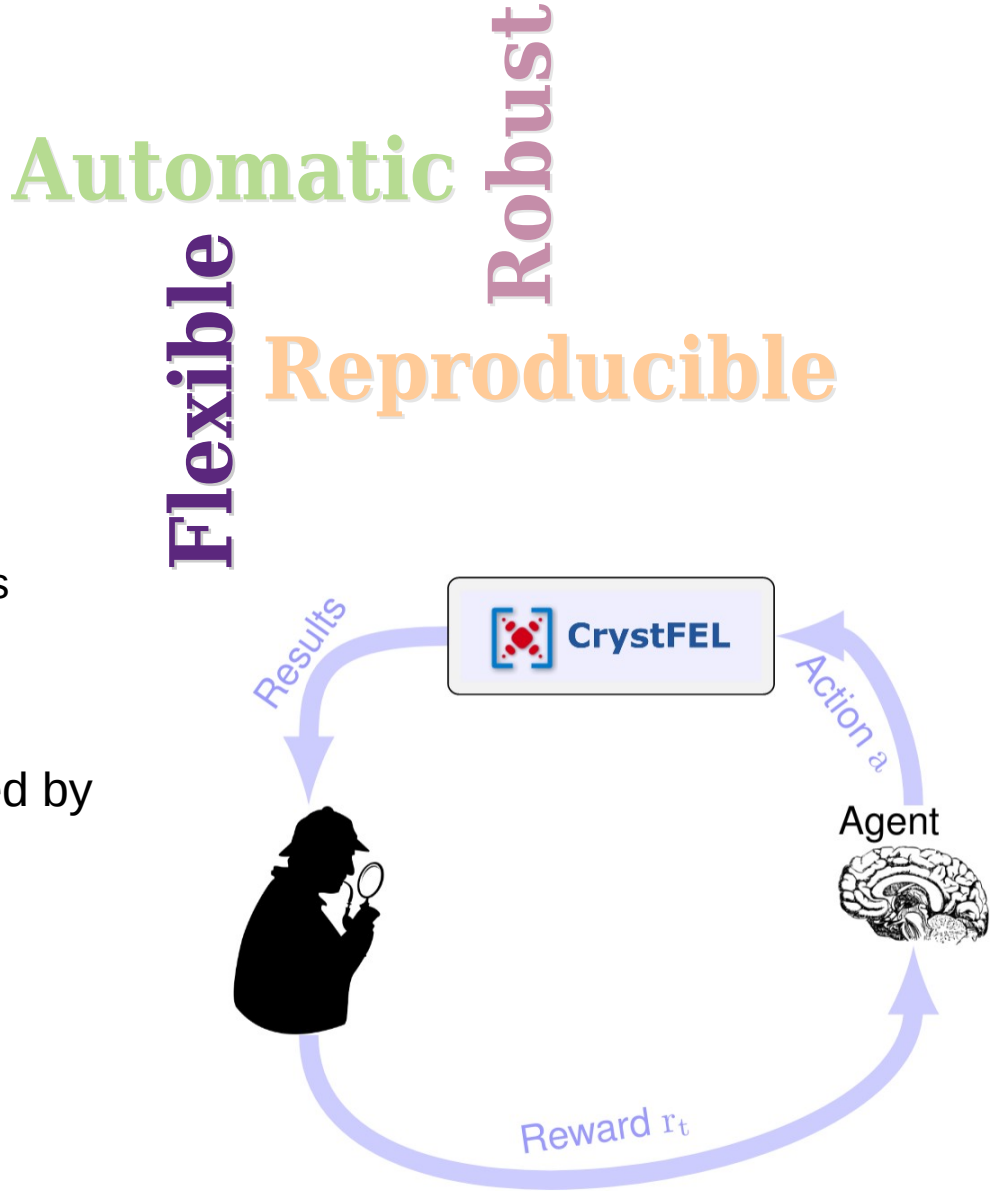
Scale and merge reflections

xmpl_30.summary

```
Crystallographic FOMs:
                                 overall    outer shell
Completeness                      100.01         100.00
Signal-over-noise                   3.13           2.03
CC_1/2                            0.7113         0.0677
CC*                               0.9118         0.3562
R_split                            40.58          89.08
```

Current developments and outlook

- Support for pump-probe experiments
- Rewriting the framework core to make it also as flexible as possible
- Upstream pipeline extensions
  - Automatic triggering once corrected detector data is available
  - Inclusion of detector geometry optimization
- Integration with reinforced learning tool developed by Arman and Danilo
- Downstream pipeline extension
  - Crystollographic phasing & model building
- Graphical representation of processing results



Automatically run the automated pipelline

We provide a tool which allows to run xwiz iteratively changing any of the config parameters over a list or range of values.


```
[max-exfl001]~% module load exfel EXtra-xwiz
 - EXFEL modulepath enabled
(extra-xwiz_env) [max-exfl001]~% xwiz-scan-parameters -h
usage: xwiz-scan-parameters [-h] [-xc XWIZ_CONFIG] [-o | -f]

Run xwiz scanning through the parameters specified inxwiz_scan_conf.toml.

optional arguments:
  -h, --help            show this help message and exit
  -xc XWIZ_CONFIG, --xwiz_config XWIZ_CONFIG
                        Use specified xwiz configuration file. If this option
                        is omitted configuration file specified in the scan
                        config will be used.
  -o, --output          Skip running xwiz jobs and just collect existing
                        output.
  -f, --force           Replace any existing job folders.
(extra-xwiz_env) [max-exfl001]~%
```

Set up xwiz-scan-parameters config file

xwiz_scan_conf.toml

```
[settings]
xwiz_config = '<some_path>/xwiz_conf.toml'
log_completion = 20

...

[output.store_xarray]
  output_file = "scan_data.nc"

[output.store_csv]
  output_file = "scan_data.csv"
```

Set up xwiz-scan-parameters config file

xwiz_scan_conf.toml

```
[scan.run]
  'data.runs' = ['35', '36', '37']
  'data.vds_names' = [
      'pXXXX_r35_vds.cxi',
      'pXXXX_r36_vds.cxi',
      'pXXXX_r37_vds.cxi'
  ]


[scan.SNR]
  'proc_coarse.peak_snr' = {
      start = 3, end = 9, step = 2
  }
```

xwiz_conf.toml

```
[data]
Path = "<proposal_path>/proc/"
runs = "35"
...
vds_names = "pXXXX_r35_vds.cxi"
```

xwiz_conf.toml

```
[proc_coarse]
...
peak_snr = 5
```

Run parameters scanner

xwiz_scan_conf.toml

```
[scan.run]
  'data.runs' = ['35', '36', '37']
  'data.vds_names' = [
  ]

[scan.SNR]
  'proc_coarse.peak_snr' = {
      start = 3, end = 9, step = 2
  }
```

```
(extra-xwiz_env) [max-exfl001]~/test_par_scan% xwiz-scan-parameters
INFO: Starting EXtra-xwiz parameters scanner.
INFO: Running with xwiz config from: /home/turkot/test_xwiz_03/xwiz_conf.toml
INFO: Finished job 3/12: /home/turkot/test_par_scan/aprep_00/run_00/SNR_02.
INFO: Finished job 6/12: /home/turkot/test_par_scan/aprep_00/run_01/SNR_01.
[############################\----------] 58.3% 7/12 Running in: aprep_00/run_01/SNR_03
```

xwiz_conf.toml

```
[data]
Path = "<proposal_path>/proc/"
runs = "35"
                                  s.cxi"
```

xwiz_conf.toml

```
[proc_coarse]
...
peak_snr = 5
```

Parameters scanner results

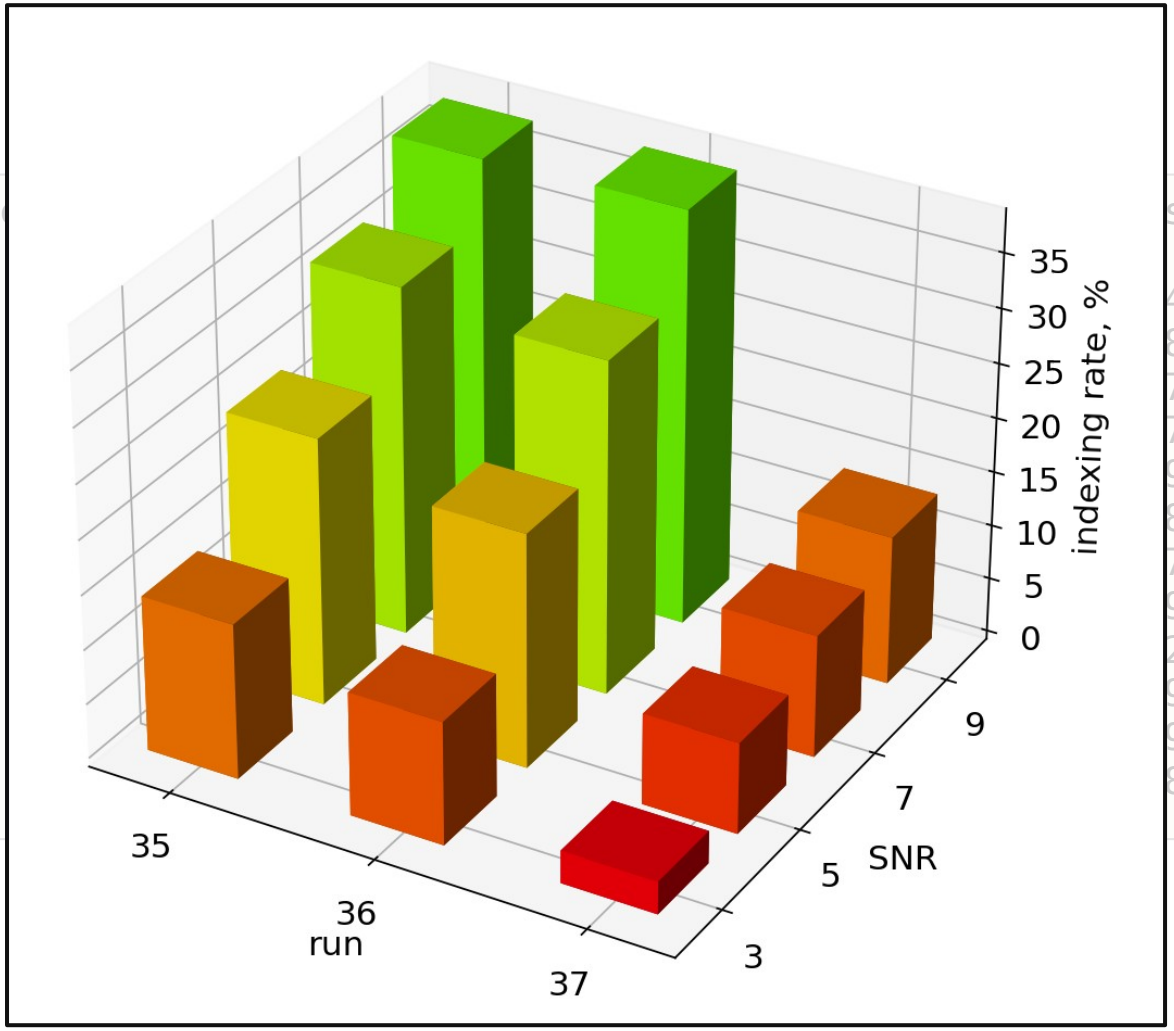
param_scan.log

SNR	run	index_rate	n_crystals	n_frames	completeness	snr	cc_half	cc_star	r_split
3	35	14.220000	711	5000	91.989998	1.04	0.1134	0.4513	116.739998
	36	11.280000	564	5000	90.260002	0.91	0.6296	0.8790	66.260002
	37	2.980000	149	5000	48.430000	2.35	0.3502	0.7202	85.110001
5	35	24.520000	1226	5000	97.309998	1.01	0.3484	0.7188	83.750000
	36	21.440001	1072	5000	96.510002	1.01	0.7205	0.9152	60.950001
	37	8.340000	417	5000	75.330002	1.75	0.4949	0.8137	72.879997
7	35	32.080002	1604	5000	98.110001	1.16	0.4679	0.7985	149.320007
	36	30.620001	1531	5000	98.010002	0.90	0.7499	0.9258	59.750000
	37	11.220000	561	5000	81.430000	1.84	0.0336	0.2550	88.540001
9	35	37.900002	1895	5000	98.300003	1.17	0.8641	0.9629	73.620003
	36	38.160000	1908	5000	98.410004	1.17	0.6890	0.9033	55.980000
	37	13.640000	682	5000	84.379997	1.51	0.5053	0.8193	72.930000

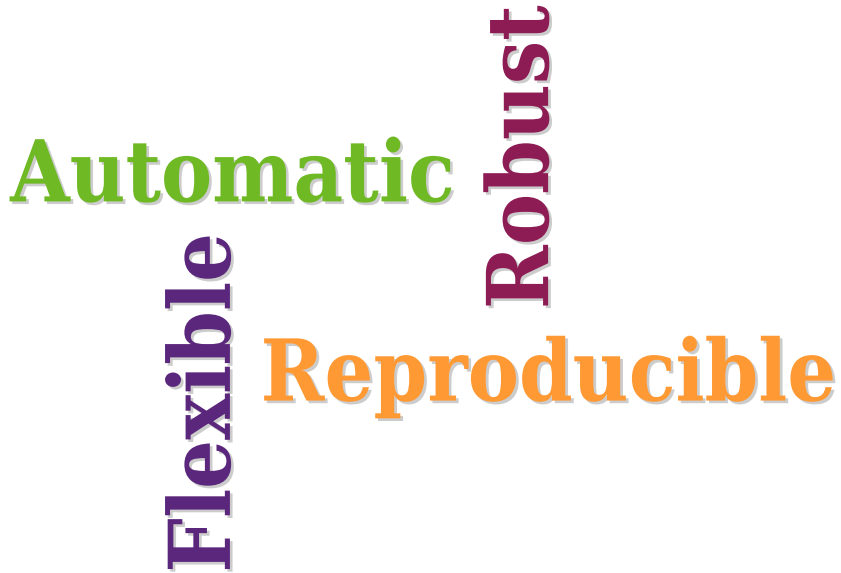
Parameters scanner results

param_scan.log

SNR	run	index_rate	n_	star	r_split
3	35	14.220000		4513	116.739998
	36	11.280000		8790	66.260002
	37	2.980000		7202	85.110001
5	35	24.520000		7188	83.750000
	36	21.440001		9152	60.950001
	37	8.340000		8137	72.879997
7	35	32.080002		7985	149.320007
	36	30.620001		9258	59.750000
	37	11.220000		2550	88.540001
9	35	37.900002		9629	73.620003
	36	38.160000		9033	55.980000
	37	13.640000		8193	72.930000



Summary



- EXtra-xwiz is an automatic, robust and reproducible pipeline for precessing crystallographic data
  - Undergoing developments to make it as flexible as possible
- Parameters scanner allows to run xwiz iteratively
  - User can define a scan over values for any xwiz config parameters
- It is a growing project and we will appreciate your suggestions - please contact us at da@xfel.eu.
- In the future we have plans to implement similar pipelines also for other experiments.