DATA REDUCTION

R&D project

at European XFEL

Egor Sobolev
Data Analysis Group, European XFEL

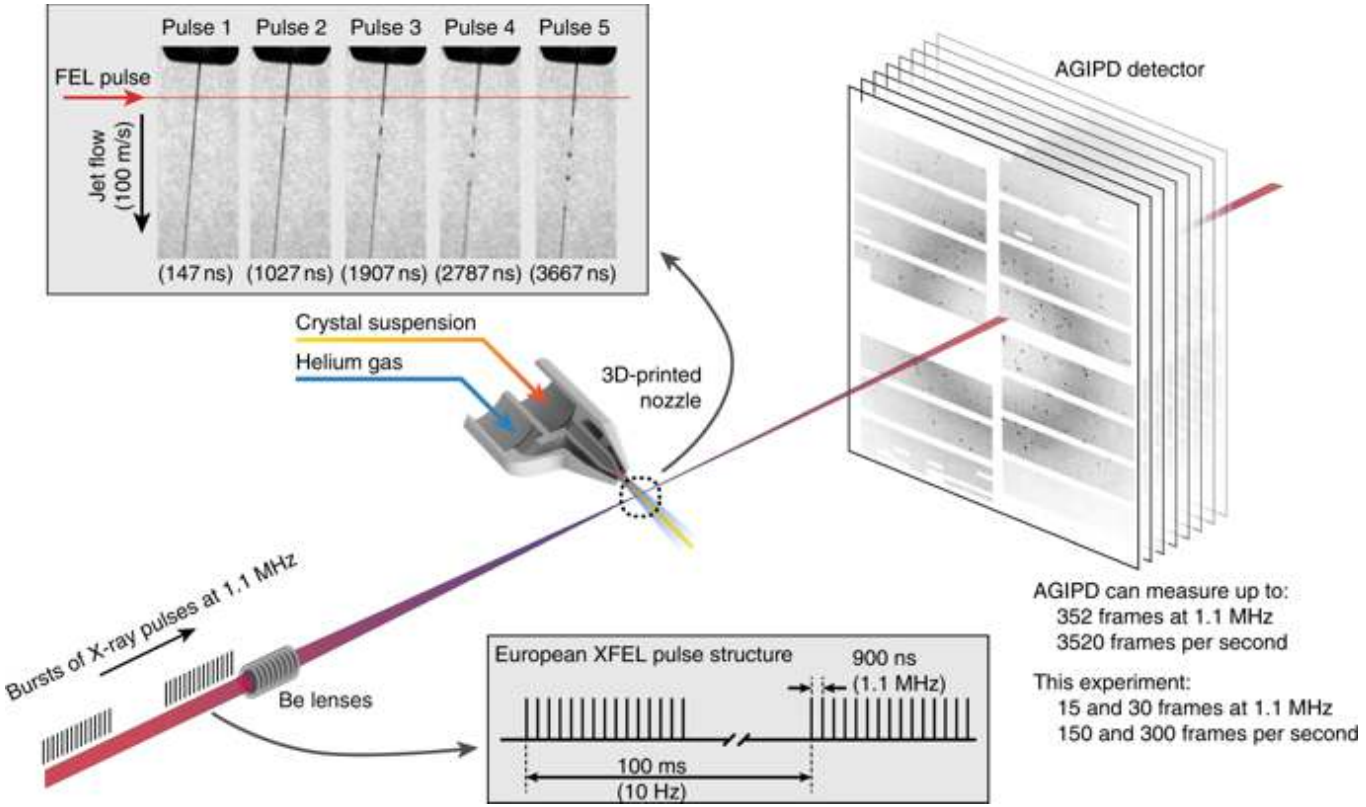
INNOV WP7 Workshop
11.10.2021

Structure of the presentation

- overview of the data production at European XFEL
- the issues and challenges related to the big data
- overview of the simplest and most effective reduction methods
- tools

European XFEL is the fastest X-ray laser in the world

27000 pulses per second

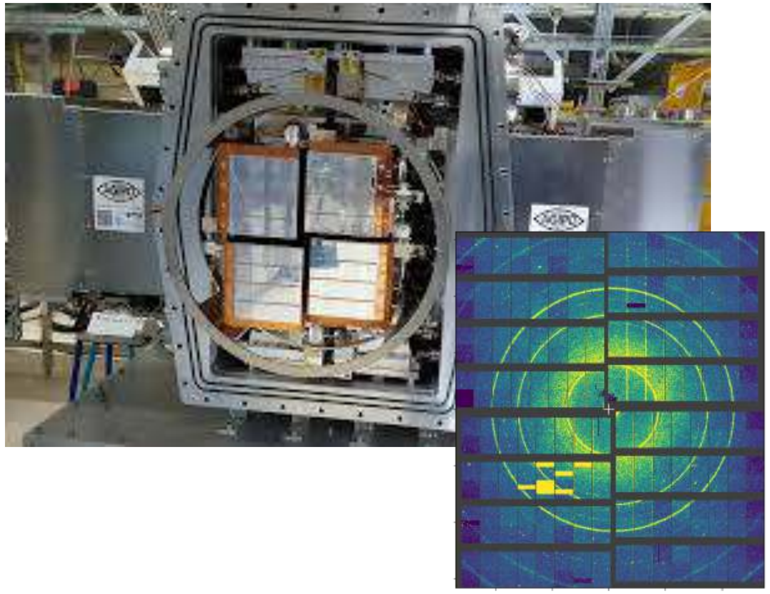


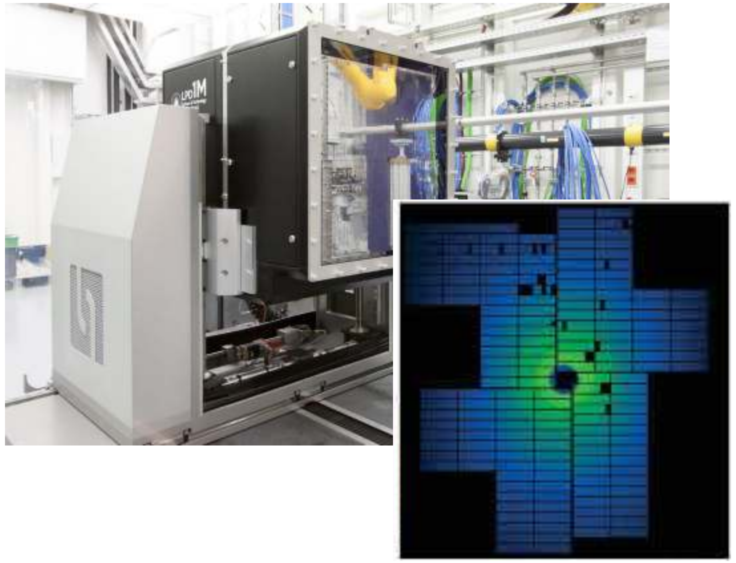
AGIPD can measure up to:
352 frames at 1.1 MHz
3520 frames per second

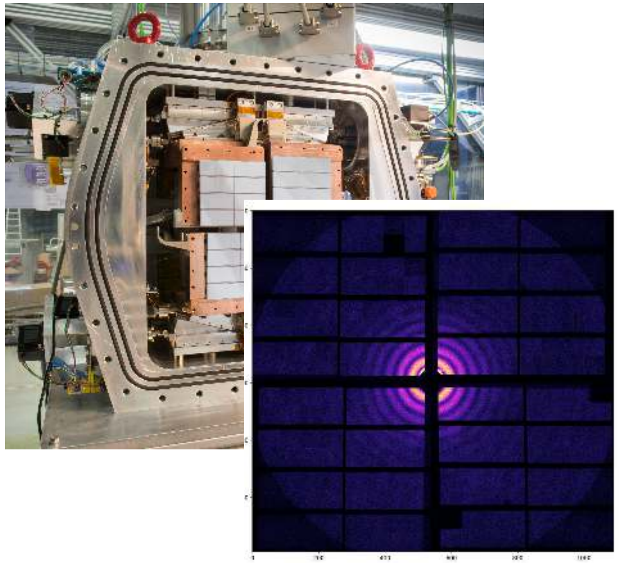
This experiment:
15 and 30 frames at 1.1 MHz
150 and 300 frames per second

Where from do the big data come?

- Fast area detectors:
  - up to 8000 1Mpx frames per second with 14-30 GiB/s (up to 100 TiB/hour)
  - typical amount is about 120TiB per experiment, the biggest > 1PiB (1 week)

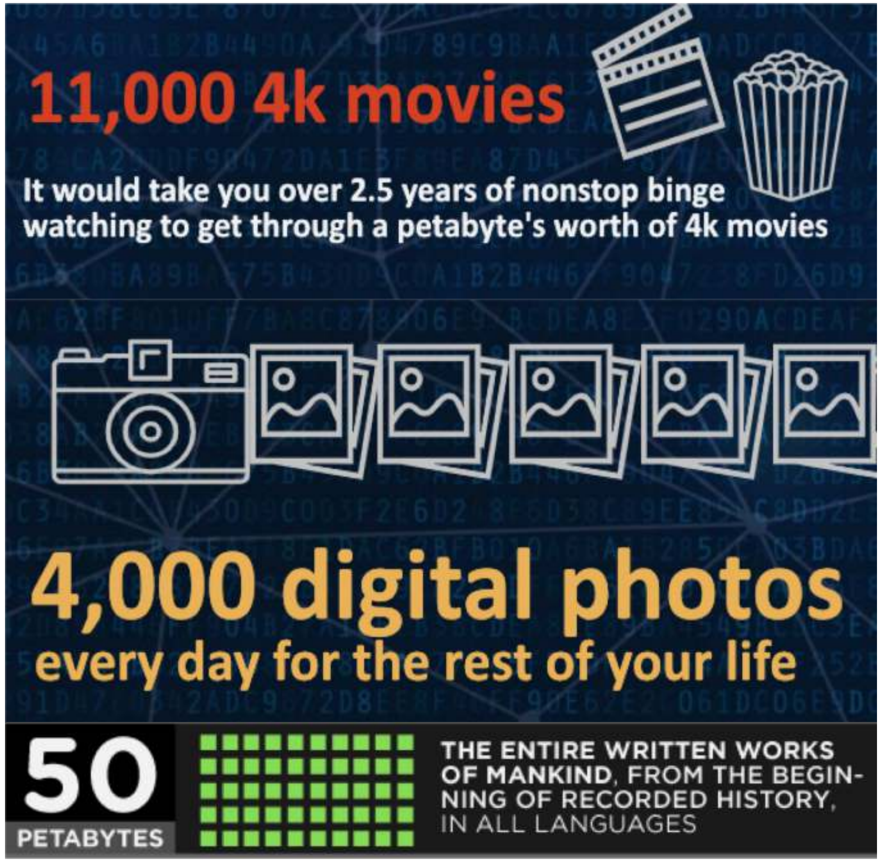






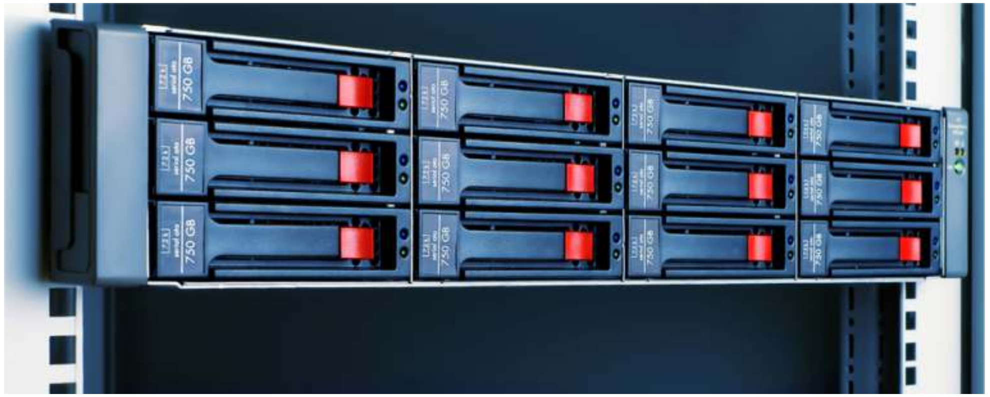
- AGIPD 4M is coming

How big is a petabyte?

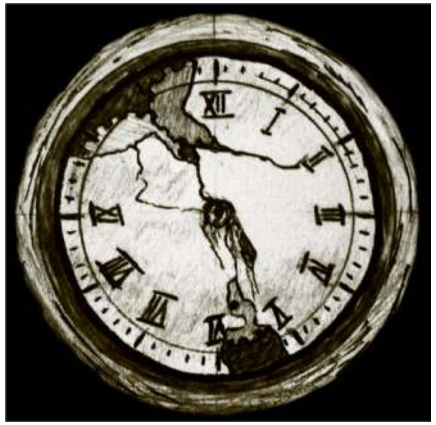

11,000 4k movies
It would take you over 2.5 years of nonstop binge watching to get through a petabyte's worth of 4k movies
4,000 digital photos
every day for the rest of your life
50
PETABYTES
THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

What is needed to analyse the data of the singe experiment?

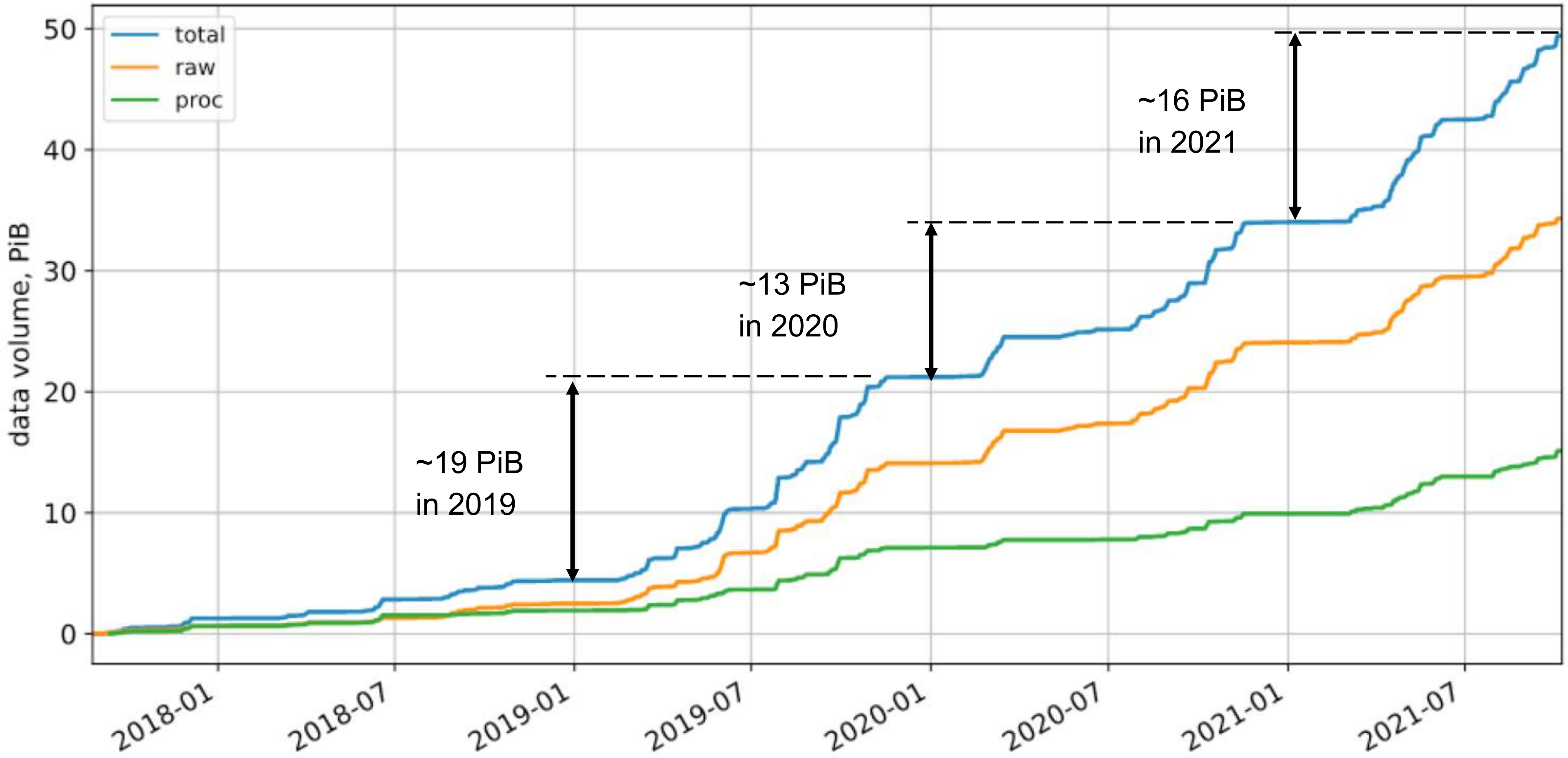
- To store
  - 10 TiB HDD x 12
  - Disk array system



- To read or write
  - 12-18 hours (in parallel)



Total data generated by European XFEL



~19 PiB in 2019

~13 PiB in 2020

~16 PiB in 2021

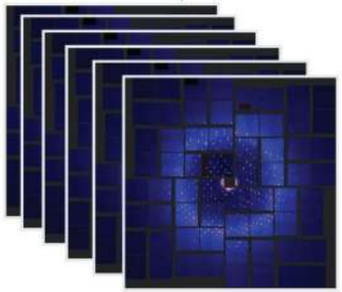
© D.Fletcher for CloudTweaks.com

What happens with data during analysis?

Serial (femto-second) X-ray crystallography workflow

Raw Data

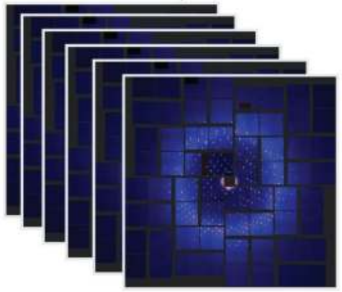
~100 TB

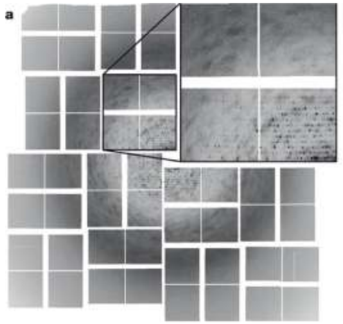


What happens with data during analysis?

Serial (femto-second) X-ray crystallography workflow

Raw Data	Hit finding	Extract peaks
~100 TB	~10 TB	~TB





What happens with data during analysis?

Serial (femto-second) X-ray crystallography workflow

Raw Data

Hit finding

Extract peaks

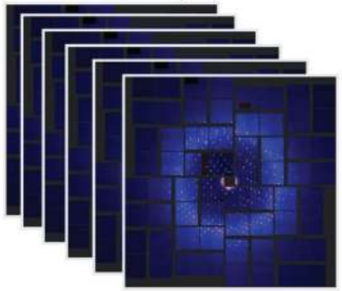
Indexable pattern

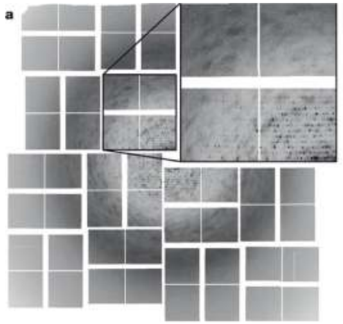
~100 TB

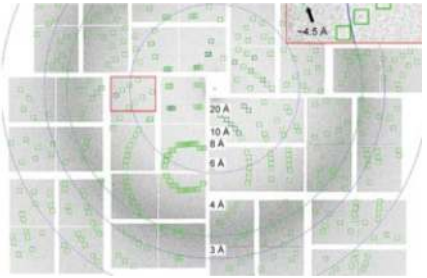
~10 TB

~TB

< 1TB



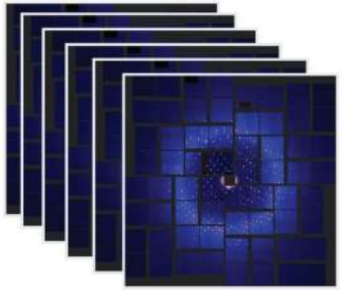


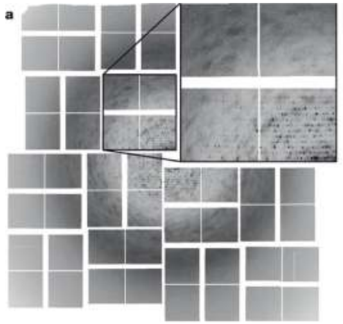


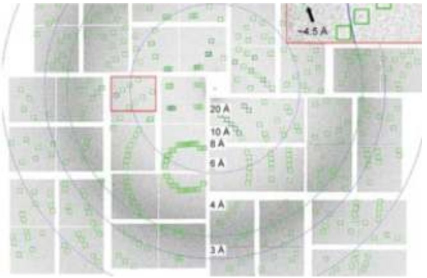
What happens with data during analysis?

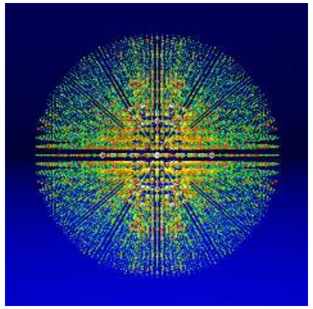
Serial (femto-second) X-ray crystallography workflow

Raw Data	Hit finding	Extract peaks	Indexable pattern	Index/integrate/merge
~100 TB	~10 TB	~TB	< 1TB	~GB





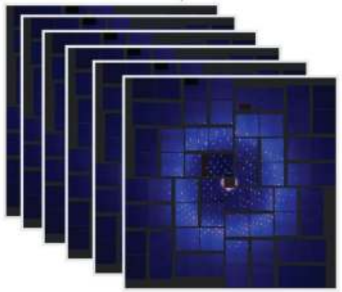


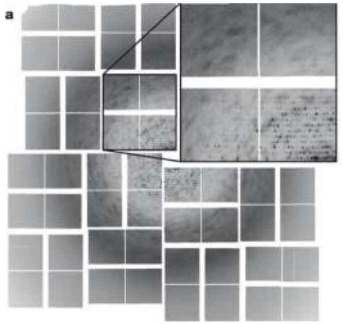


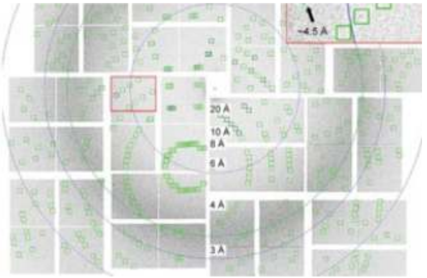
What happens with data during analysis?

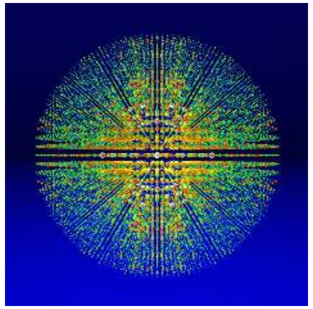
Serial (femto-second) X-ray crystallography workflow

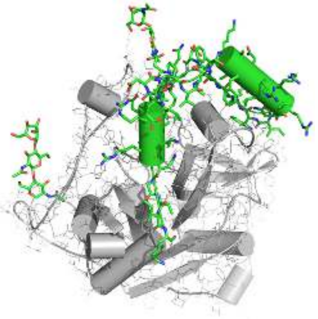
Raw Data	Hit finding	Extract peaks	Indexable pattern	Index/integrate/merge	Model
~100 TB	~10 TB	~TB	< 1TB	~GB	~MB







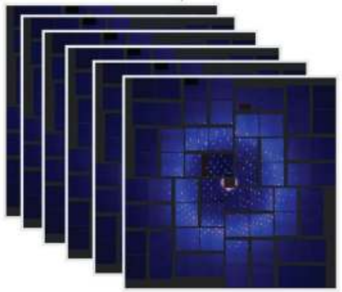


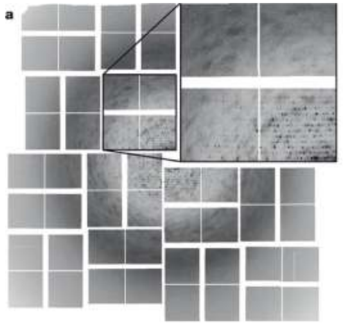


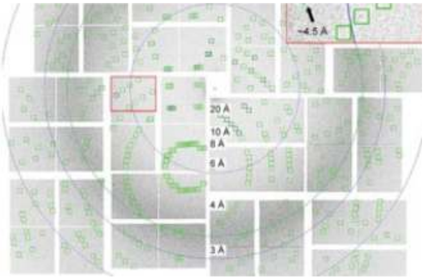
What happens with data during analysis?

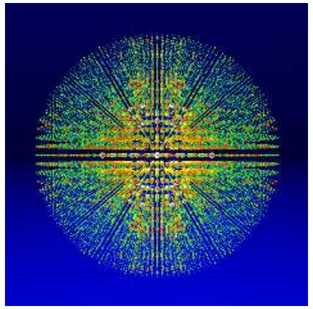
Serial (femto-second) X-ray crystallography workflow

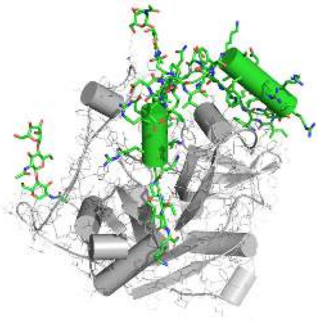
Raw Data	Hit finding	Extract peaks	Indexable pattern	Index/integrate/merge	Model
~100 TB	~10 TB	~TB	< 1TB	~GB	~MB











- May be done straightforward per frame basis
- Required iterative refinement of many parameters by analysis of many frames simultaneously

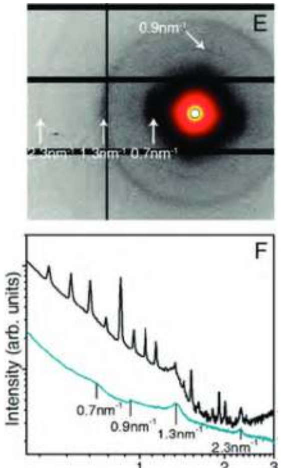
What happens with data during analysis?

Integration many 2D images to a line plot gives up to million times reduction

Small angle scattering (SAXS)

~GB

~KB



Radial direction

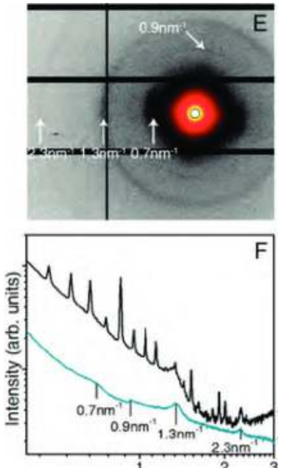
What happens with data during analysis?

Integration many 2D images to a line plot gives up to million times reduction

Small angle scattering (SAXS)

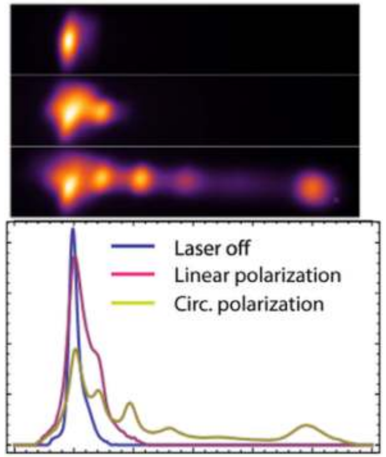
~GB

~KB



Radial direction

X-ray absorption spectroscopy (XAS)



Horizontal direction

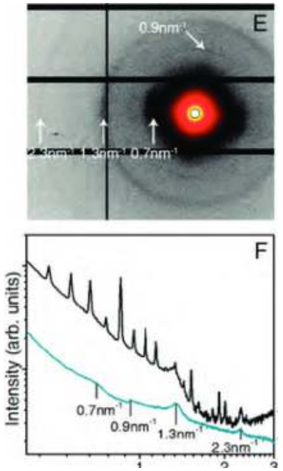
What happens with data during analysis?

Integration many 2D images to a line plot gives up to million times reduction

Small angle scattering (SAXS)

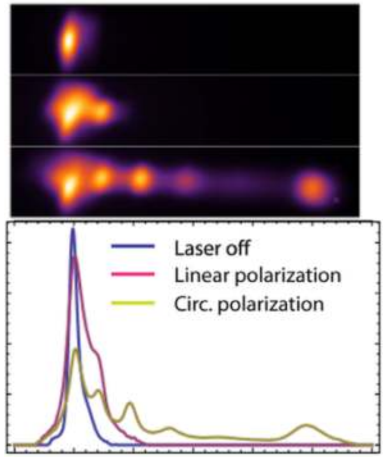
~GB

~KB



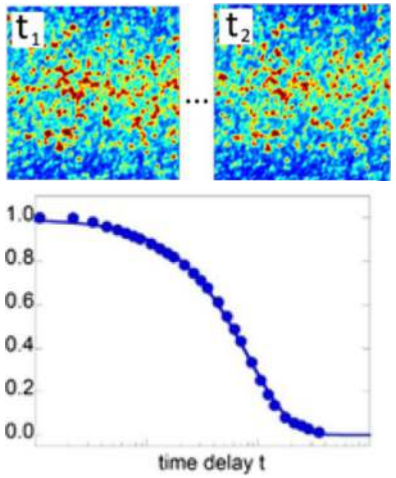
Radial direction

X-ray absorption spectroscopy (XAS)



Horizontal direction

X-ray photon correlation spectroscopy (XPCS)



time

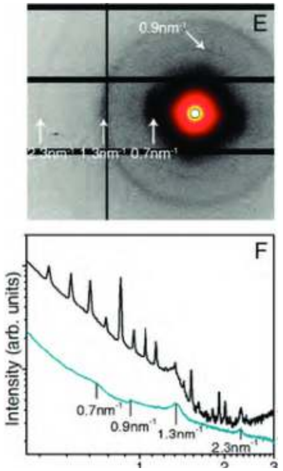
What happens with data during analysis?

Integration many 2D images to a line plot gives up to million times reduction

Small angle scattering (SAXS)

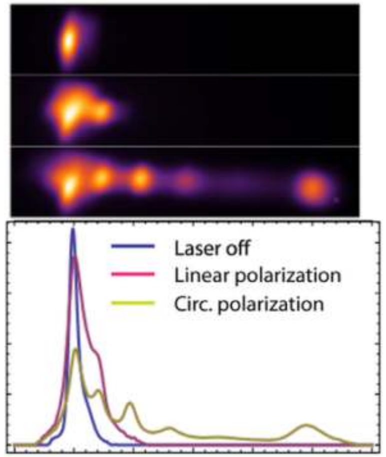
~GB

~KB



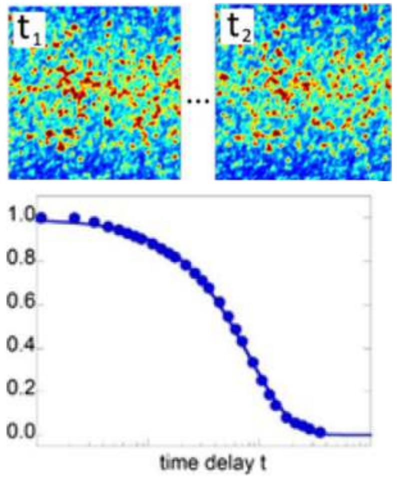
Radial direction

X-ray absorption spectroscopy (XAS)



Horizontal direction

X-ray photon correlation spectroscopy (XPCS)



time

🟧 Validation

Data reduction for different experiment types

Experimental techniques	Reduction method	Ratio	Aggregation method	Ratio
Spectroscopy XES, XAS, etc	ROI, Integration	$\sim 10^{-3}$	frames averaging	10^{-2}–10^{-3}
Powder diffraction, SAXS/WAXS	Azimuthal integration	10^{-2}–10^{-3}	frames averaging	10^{-2}–10^{-3}
Correlation analysis XPCS, XCCA	Correlation function integration	10^{-2}–10^{-3}	frames averaging	10^{-2}–10^{-3}
SFX	Hit finding	0.1–0.01		
SPI/CDI	Hit finding	10^{-2}–10^{-3}		

Data annotation

- Small data related quantities
  - XRay pulse flag
  - PPL pulse flag
  - Pulse energy

Data annotation

- Small data related quantities
  - XRay pulse flag
  - PPL pulse flag
  - Pulse energy
- Frame analysis related quantitates:
  - Number of photons
  - Number of lit pixels
  - Number of peaks
  - Hit flag

Data annotation

- Some quantities can be derived from small data
- Suits for triggering data analysis pipelines as well
- Zero-risk strategies
- Validation

- Small data related quantities
  - XRay pulse flag
  - PPL pulse flag
  - Pulse energy

- Frame analysis related quantitates:
  - Number of photons
  - Number of lit pixels
  - Number of peaks
  - Hit flag

Software tools for data analysis and reduction

Offline tools

- EXtra-data – access to the data
- EXtra-writer – writes data in EuXFEL format
- EXtra-geom, GeoAssembler – detector geometry tools
- Pasha – shared memory parallelisation
- Framework for offline analysis

Software tools for data analysis and reduction

Offline tools

- EXtra-data – access to the data
- EXtra-writer – writes data in EuXFEL format
- EXtra-geom, GeoAssembler – detector geometry tools
- Pasha – shared memory parallelisation
- Framework for offline analysis

Online tools

- Varios Karabo devices
- AgipdLitFrameFinder – Agipd frame annotation
- EXtra-metro – runtime programmable processing pipeline
- EXtra-foam – online & offline data analysis and visualisation

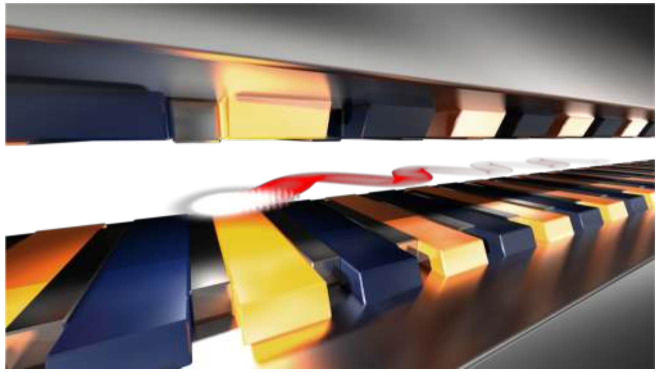
Processing abstraction in offline framework

```
class BatchAzimuthalInt(AlgorithmBase):
    ALG_ID = "azint"
    HELP = "Azimuthal Integration"

    @classmethod
    def add_arguments(cls, parser):
        pass

    def configure(self, args):
        nbuf = 2 * self._computer.nworker
        shm_map = dict()

        self._computer.configure(shm_map, nbuf)
```

```
def initialize(self):
    data_iterator = self.preprocessor.split_trains(
        self._computer.nworker)
    return data_iterator

def process_train_data(self, idx, train_no, train_id, first, img):
    return img.npulse

def finalize_chunk_processing(self, chunk_no, idx):
    pass

def finalize(self):
    pass
```

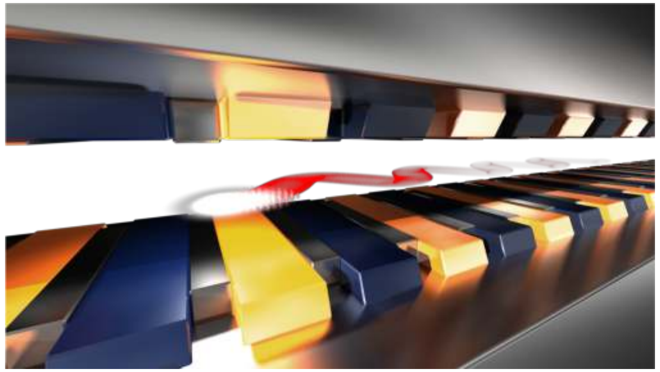
Data reduction

- Essential step to get scientific results from raw data



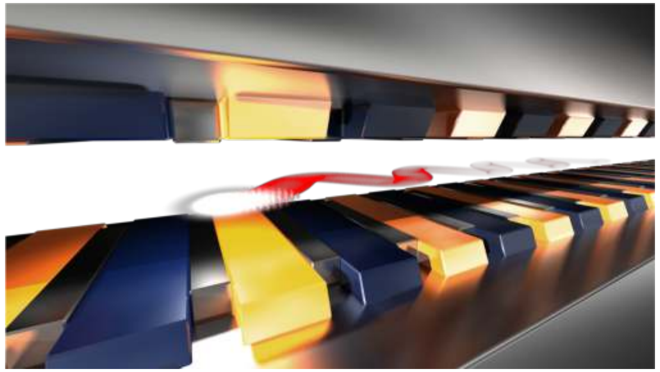
Data reduction

- Essential step to get scientific results from raw data
- Next level in automatization and speed of data analysis



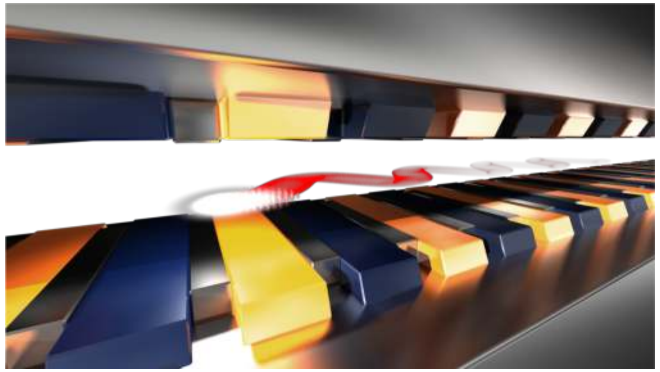
Data reduction

- Essential step to get scientific results from raw data
- Next level in automatization and speed of data analysis
- New responsibility of scientific facilities



Data reduction

- Essential step to get scientific results from raw data
- Next level in automatization and speed of data analysis
- New responsibility of scientific facilities
- Collaborations between facility and community



Thank you for you attention.
Questions?

Egor Sobolev, egor.sobolev@xfel.eu
Data Analysis Group, da-support@xfel.eu