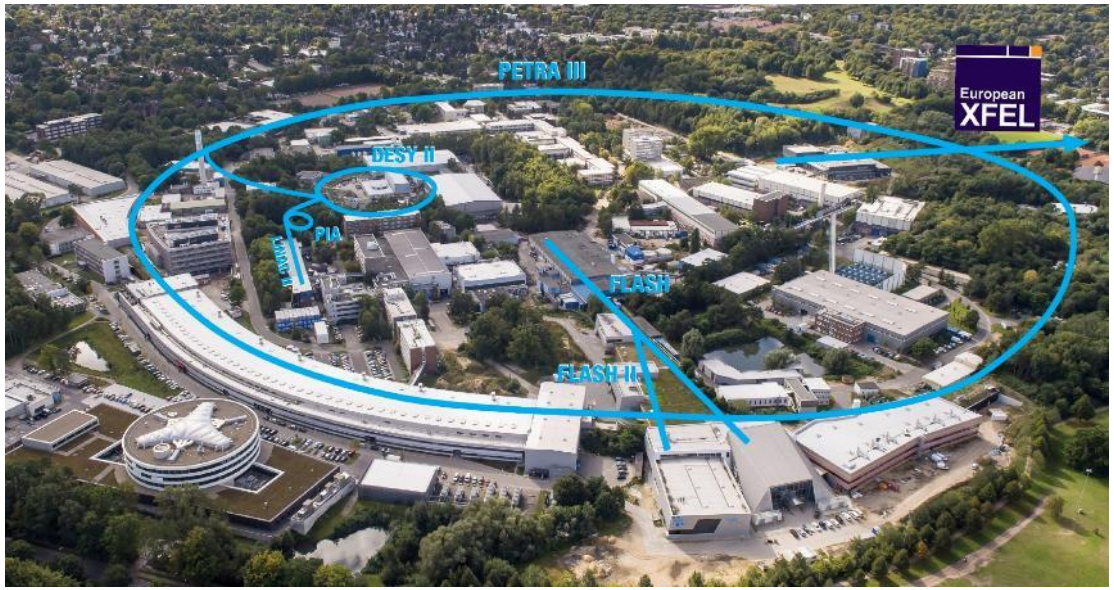
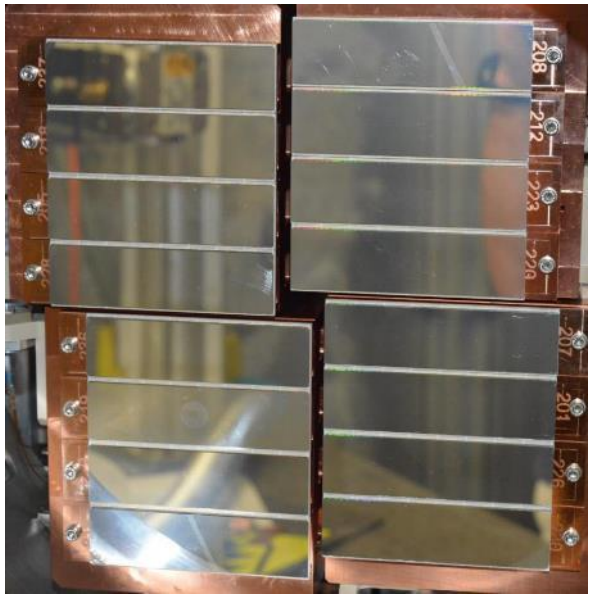
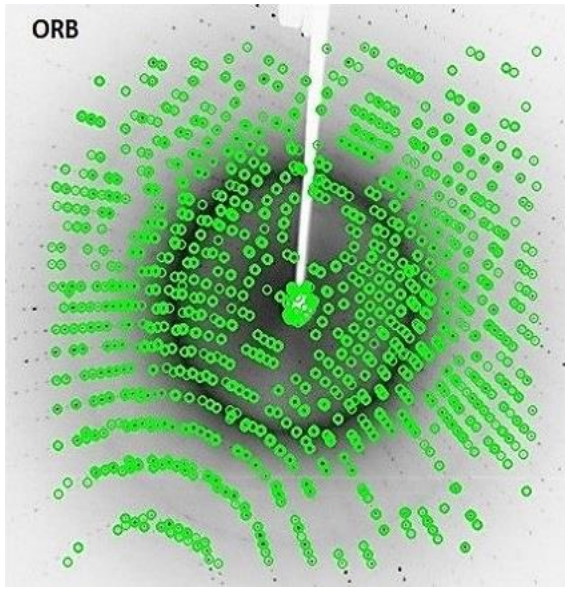
Data reduction in photon science

David Pennicard, Vahid Rahmani, Shah Nawaz, Shabarish Pala Setty, Heinz Graafsma - DESY







HELMHOLTZ ASSOCIATION

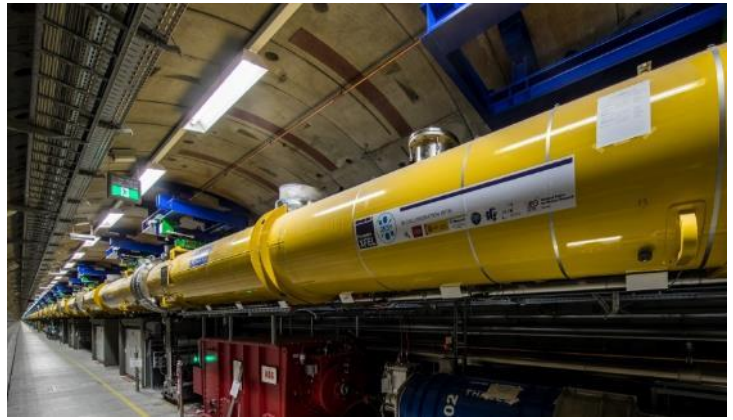
DESY

Introduction

- Why data is a challenge in photon science
- Connection to detector development
- Current and future data reduction methods

Drivers for increasing data rates in photon science

New facilities



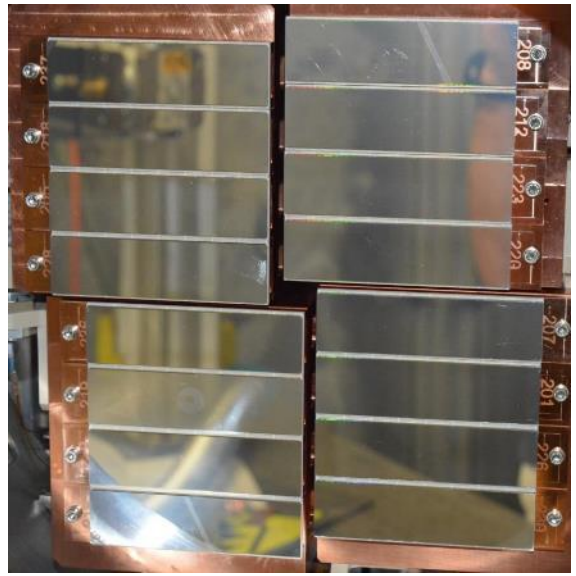
Free electron lasers (Eu.XFEL)

4th generation synchrotrons (ESRF-EBS, PETRA-IV)

New detectors

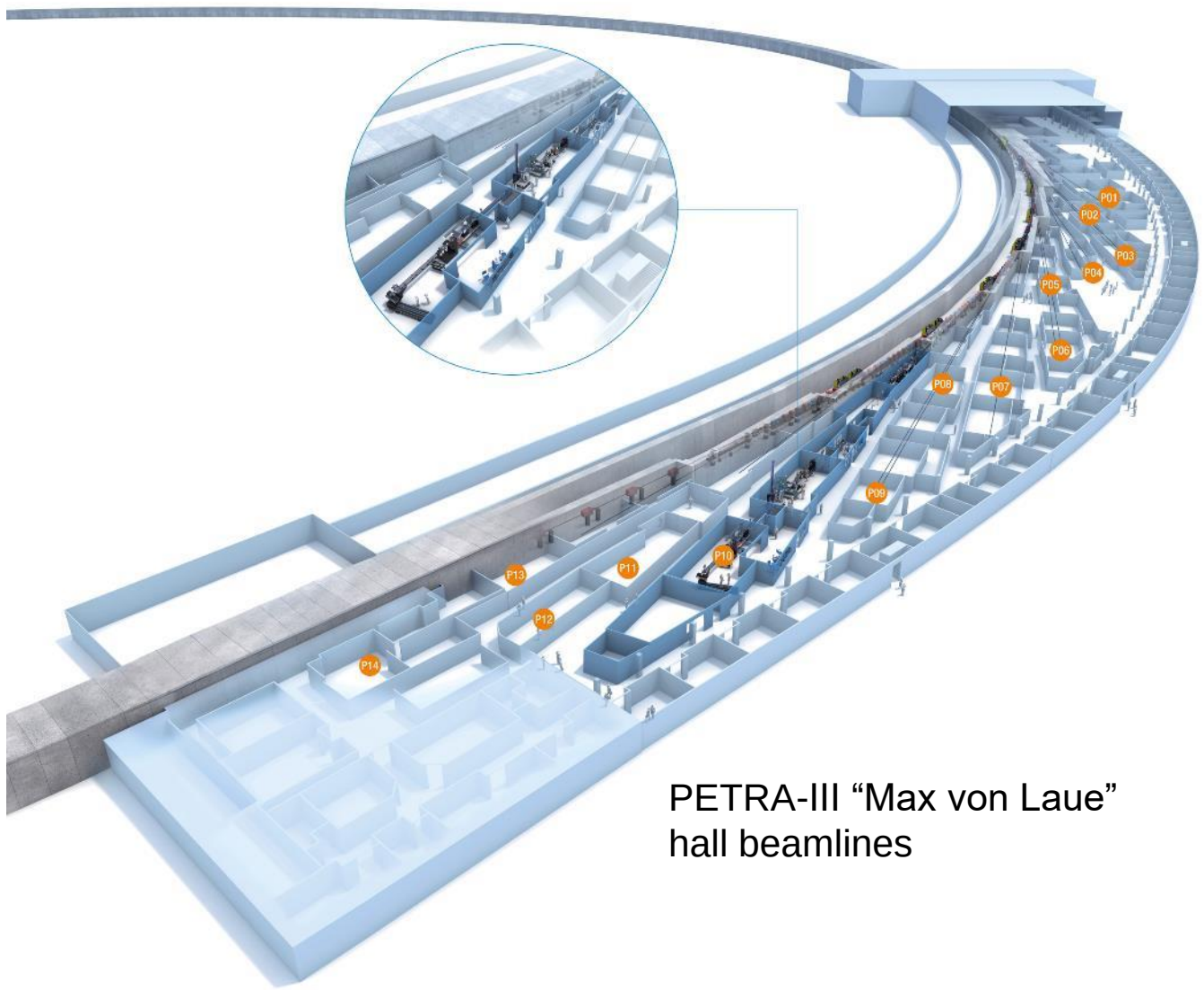
Multi-megapixel cameras

- Currently 1000s of frames/second
- Developing CoRDIA – over 100,000 fps



AGIPD detector

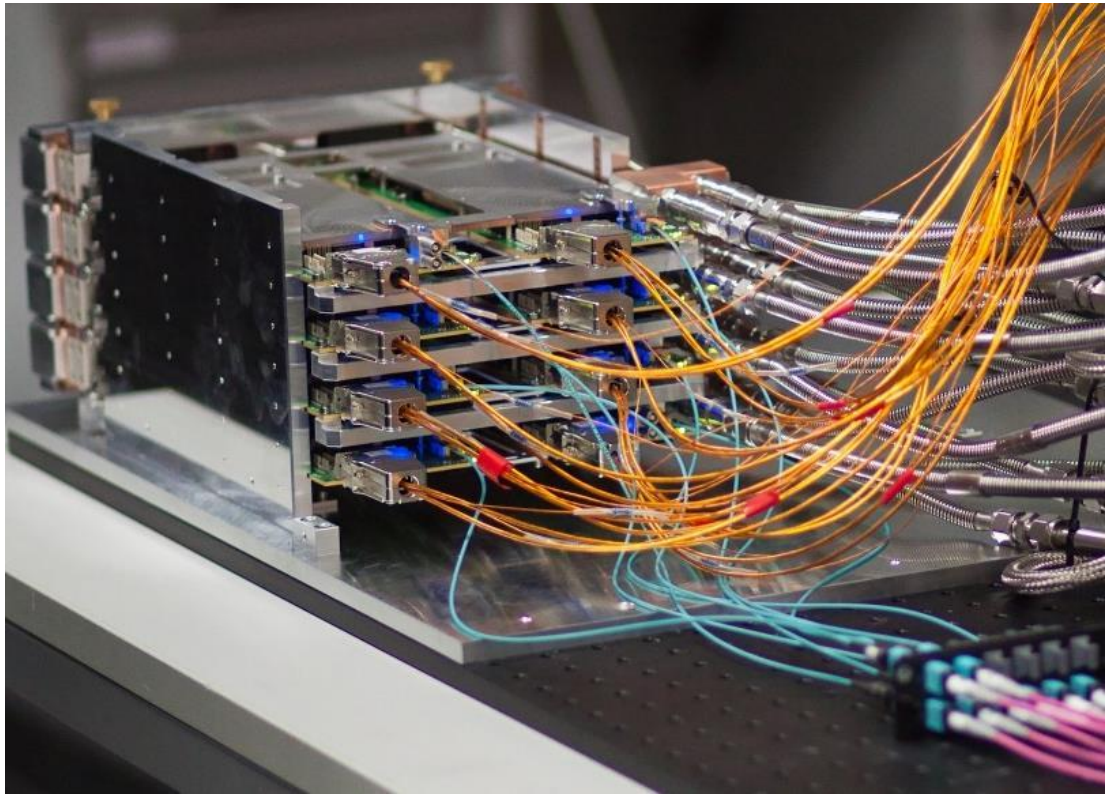
Experiments at light sources



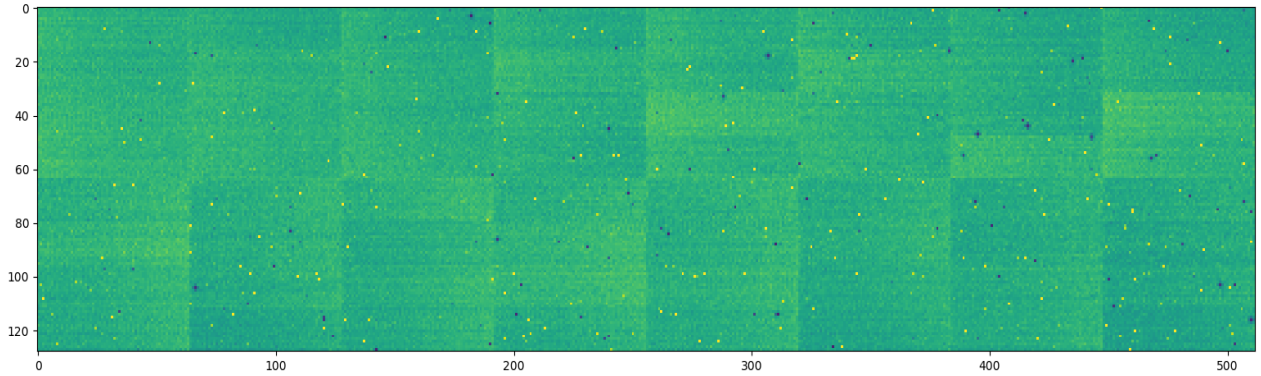
PETRA-III “Max von Laue” hall beamlines

- Wide variety of X-ray techniques and application areas
- User facilities with a broadening user base over time
- 2010: User takes home copy of data on hard drive
- 2020: User has remote access to full dataset stored at facility (lossless compression)
- >2025: Non-lossless reduction unavoidable

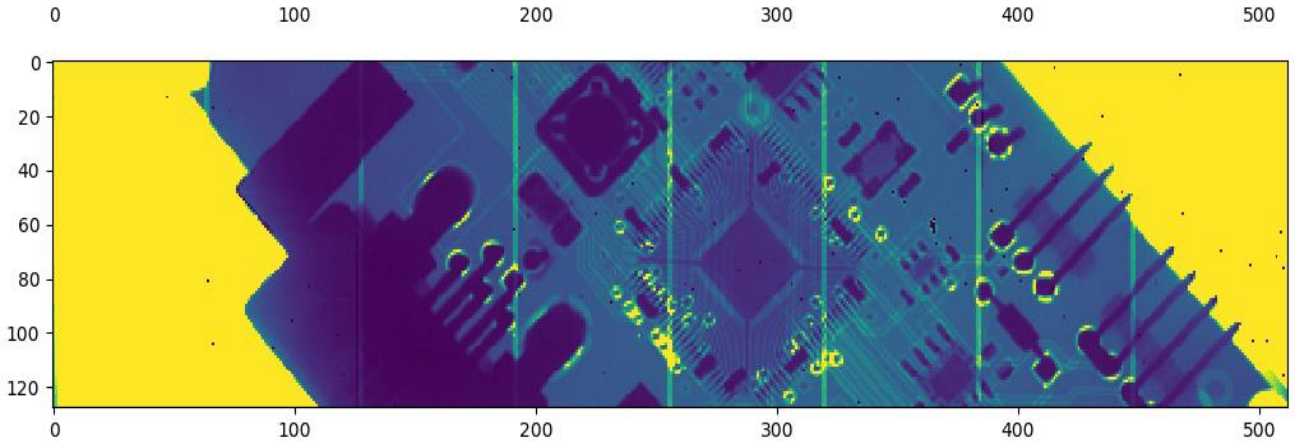
Detector-specific data acquisition and processing



Raw AGIPD data stream values

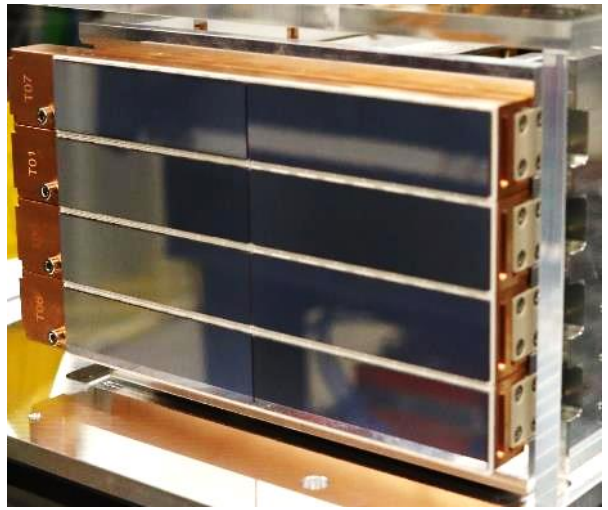


Corrected image



Goal – first phase of data reduction close to the detector

Detector module
e.g. CoRDIA



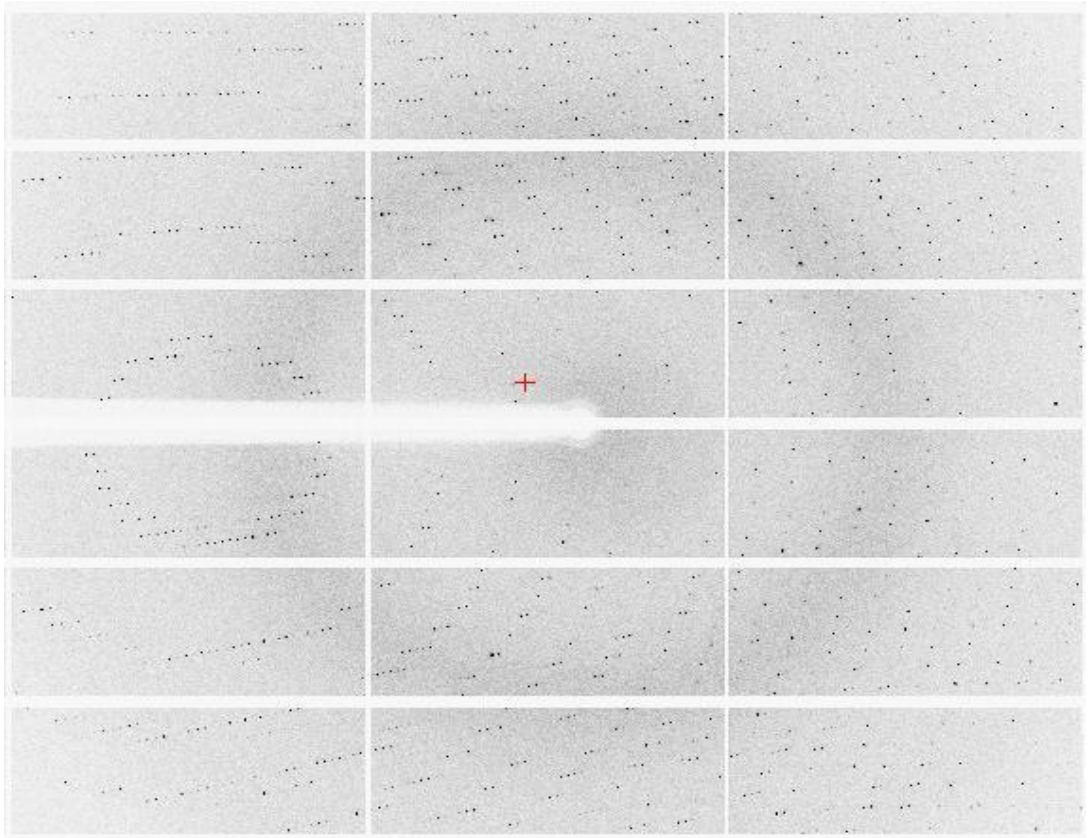
Detector PC layer with accelerator cards

Facility central computing

Data reception, image correction, data reduction

Current lossless compression methods

- Lossless compression currently works well on X-ray diffraction images (~10 compression at high frame rates):
  - Discrete numbers of X-ray photons per pixel
  - Most pixels close to 0
  - Long runs of 0 in low-intensity images



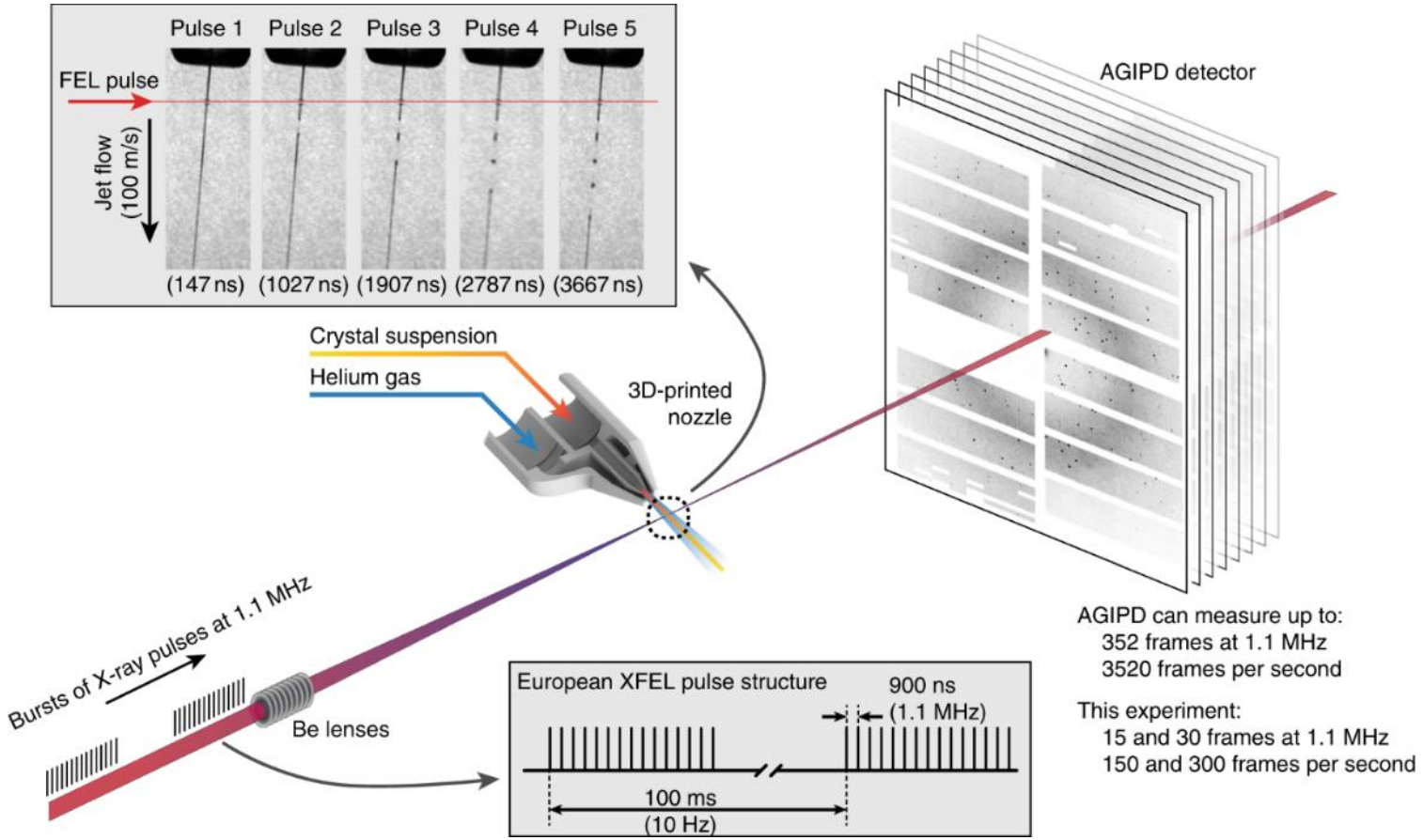
Lossy compression

- Lossless compression is not effective e.g. for imaging and CT
  - Also, reconstructions are roughly as large as raw data!
  - 3 beamlines at PETRA (out of 24) account for 50% of storage
- Synchrotron labs evaluating lossy compression – does it affect reconstruction?
  - E.g. standard methods like JPEG2000
    - PSI demonstrated compression factor of 4-8 without quality loss
  - E.g. new error-bounded methods like SZ

Impact of lossy compression of X-ray projections onto reconstructed tomographic slices, F. Marone, J. Vogel, M. Stampanoni
Journal of Synchrotron Radiation 27 5 (2020).

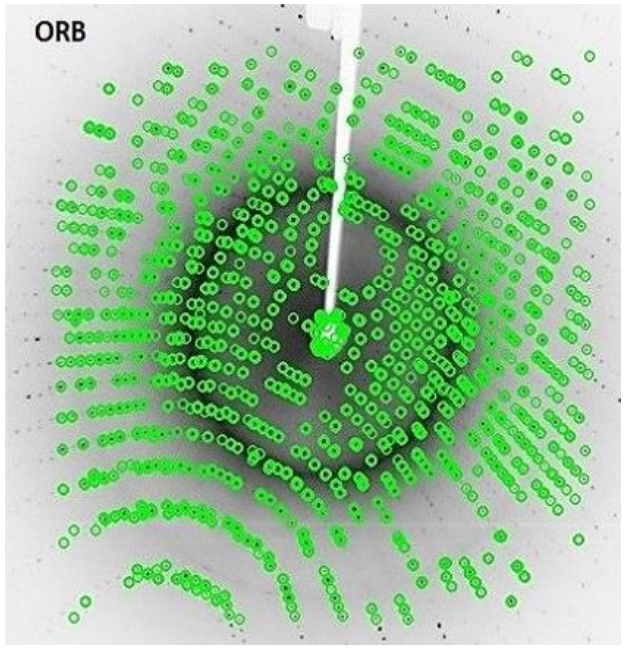
Rejection of bad data (“triggering”)

- In serial crystallography at FELs, many images are bad (sometimes >99%) due to beam missing protein crystals

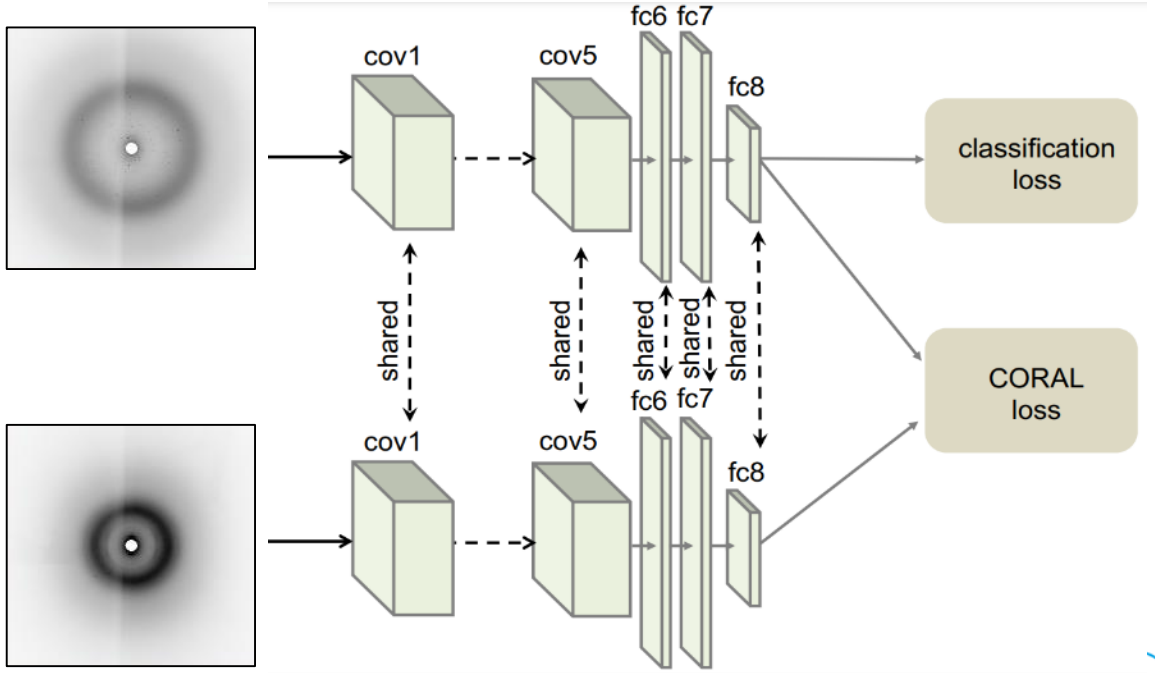


Rejection of bad data (“triggering”)

- In data processing, bad images can be rejected based on counting Bragg peaks
  - Some expert tuning of analysis parameters required
- Working on machine learning methods to reject “clearly bad” images even before saving to disk
  - See posters by Vahid Rahmani, Shah Nawaz

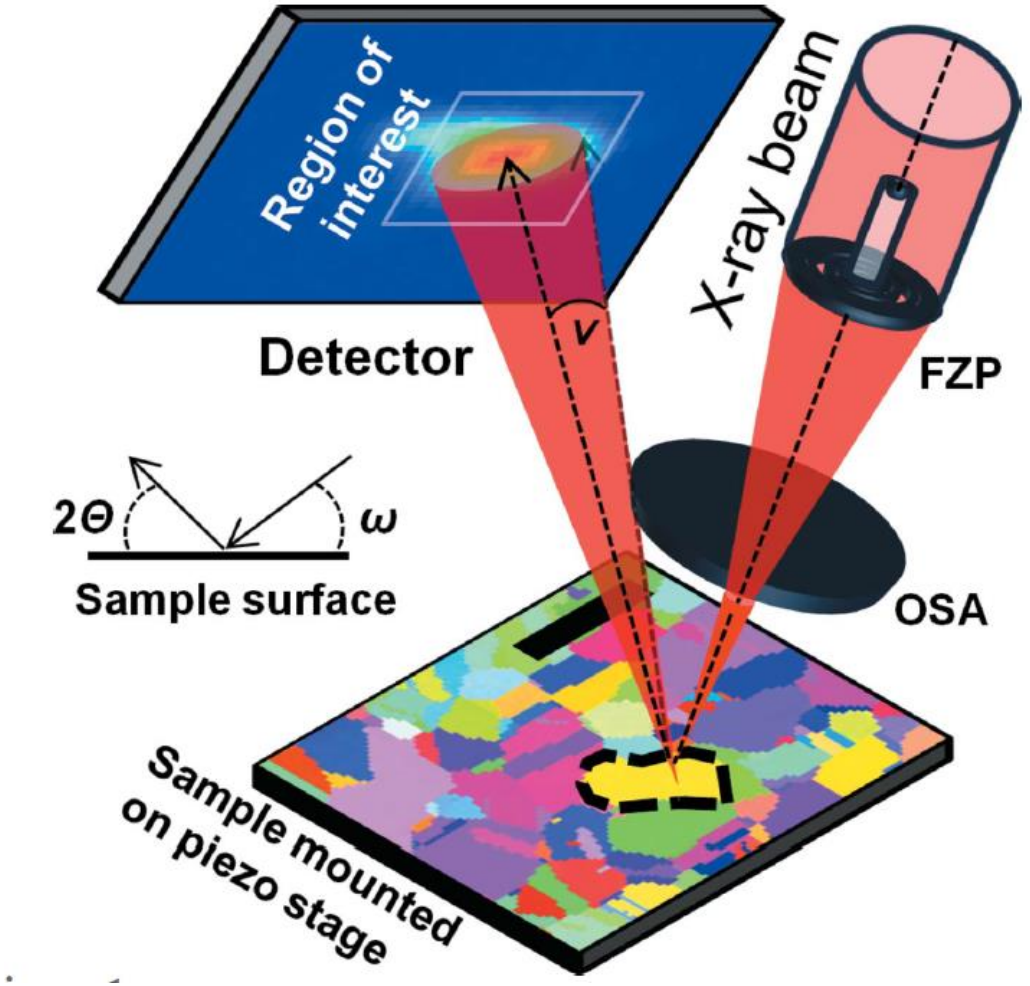


Peak detection with ORB algorithm from computer vision



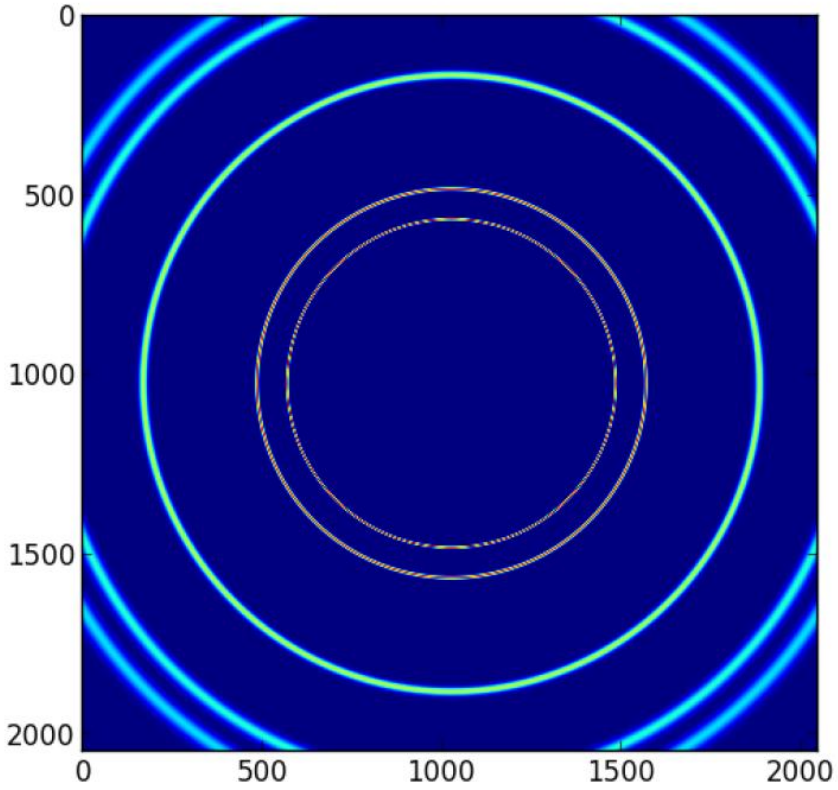
Autonomous data acquisition?

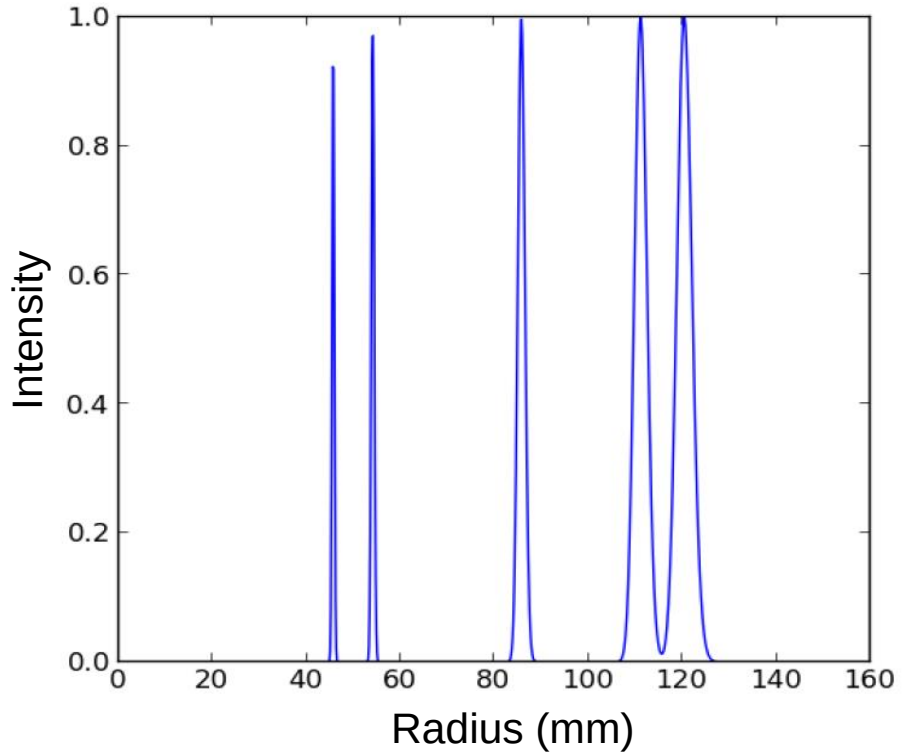
- Any X-ray technique can be used for microscopy when combined with raster scanning
- No “bad” data, but some regions more interesting than others
- Feedback could be used to automatically adjust scan parameters



“Compression by processing”

- Example – azimuthal integration
  - GPU implementation available – PyFai (pyfai.readthedocs.io)
  - MAX-IV working on FPGA implementation





Efficient representations of data?

- Usually, compressed data later needs to be decompressed for analysis
- Ideally, sparse representations of data could be used for compression, analysis, feedback etc.
- Simple example – XPCS
  - Many images are taken at high speed, but pixels are mostly 0
  - A list of hit pixels provides good compression, and can be analyzed directly
    - https://scripts.iucr.org/cgi-bin/paper?ay5566 (Zhang et al. 2021, Argonne National Lab)
  - Timestamping detectors could directly provide this list, with much higher time resolution!

Outlook

- Photon science data is varied, and in turn a variety of approaches are needed for data reduction
- “Close to the detector”, we focus on generic data compression, rejecting clearly bad images, and common preliminary analysis steps
- There are various collaborations working in this area:
  - LEAPS (League of European Accelerator-Based Photon Sources)
  - Innopool Data-X
  - Helmholtz Imaging Platform