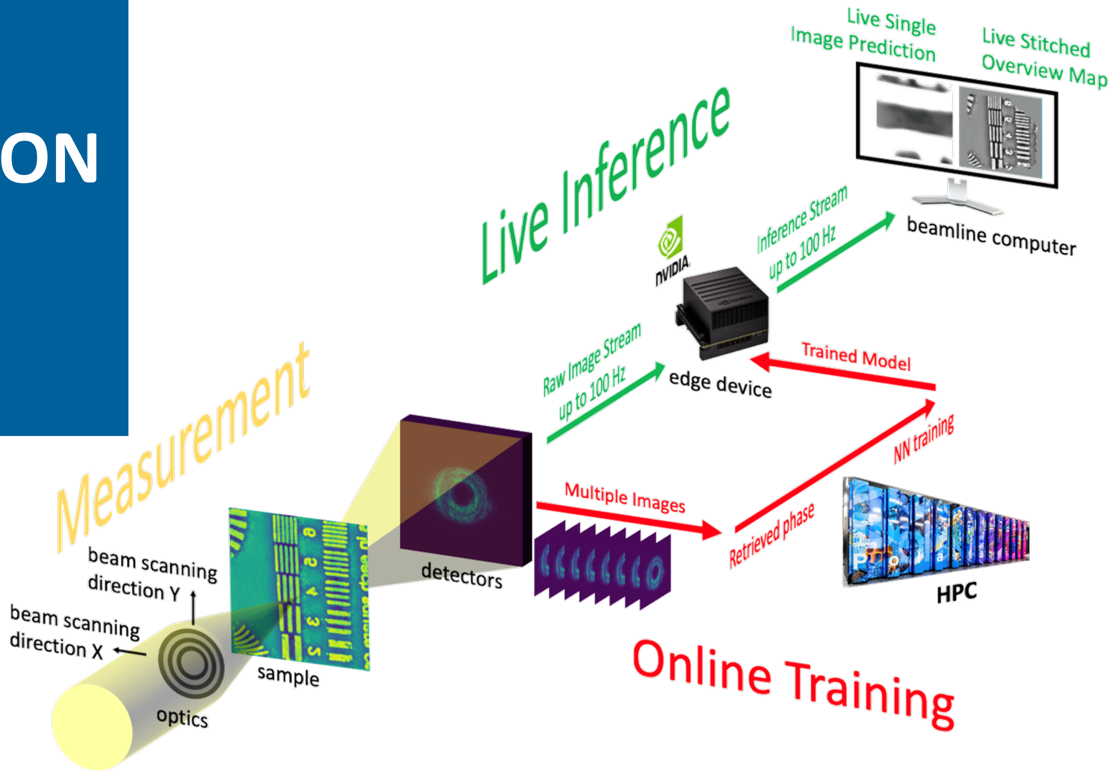
2024.03.19

IFDEPS 2024

MACHINE LEARNING BASED DATA REDUCTION AT THE EDGE AND ON CHIP AT THE APS

TAO ZHOU

Assistant Scientist
Center for Nanoscale Materials
Nanoprobe Beamline (26-ID-C @ APS)
Argonne National Laboratory



OUTLINE

Machine learning based data reduction methods are highly efficient

- Case 1 : High-throughput on-the-fly data reduction at the edge
- Case 2 : Real time feedback: rapid detection of rare events in situ
- Case 3 : Closing the loop : autonomous experiment steered at the edge

On-chip data compression to tackle challenges from the increasing data rate

- Case 4 : Co-designing ASIC with on-chip compression (“SparkPix-RT”)

You
What is the strength of machine learning based data reduction method, in one word

ChatGPT
$\boxed{\text{Efficiency}}$

OUTLINE

Machine learning based data reduction methods are highly efficient

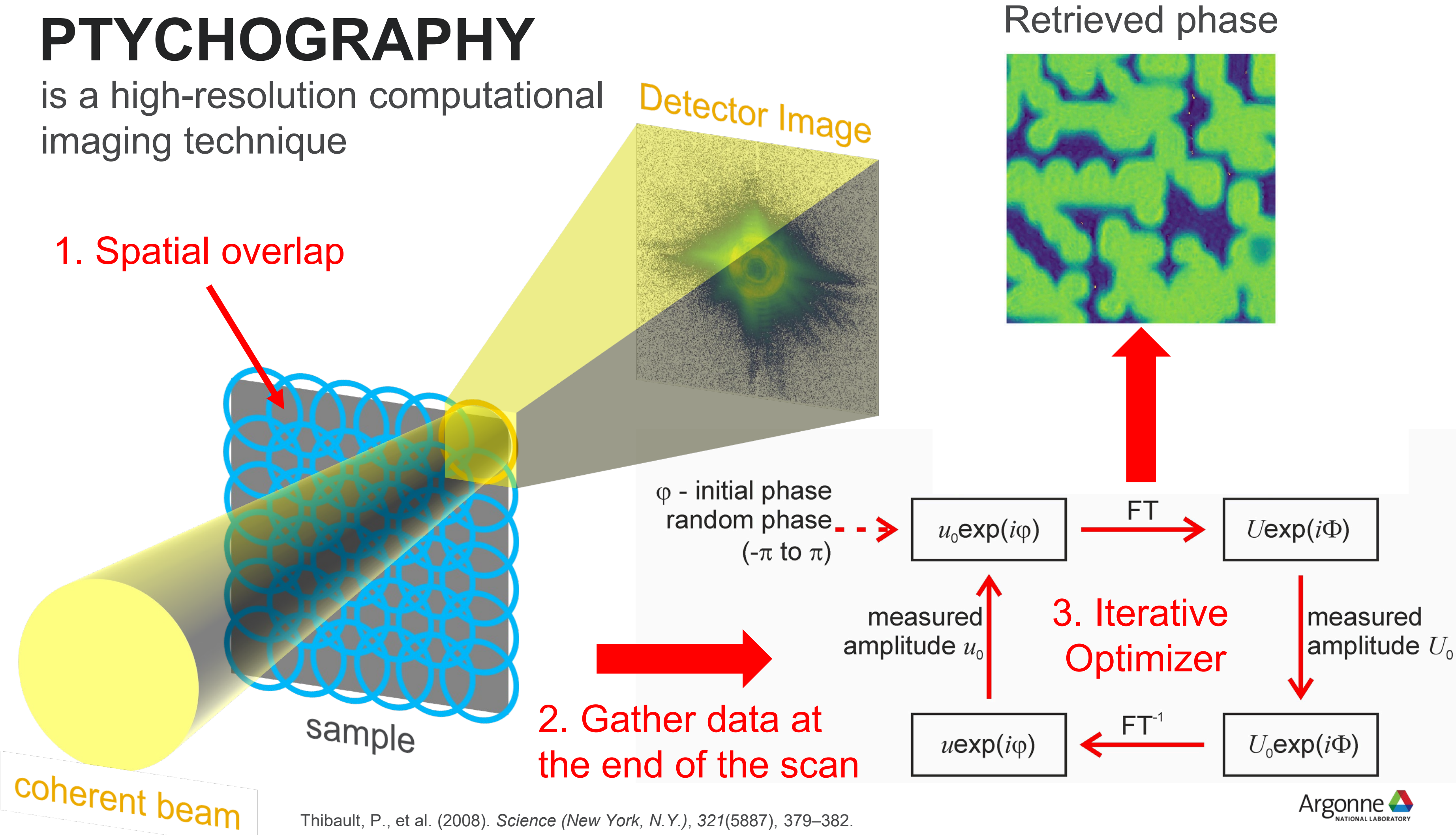
- Case 1 : High-throughput on-the-fly data reduction at the edge
- Case 2 : Real time feedback: rapid detection of rare events in situ
- Case 3 : Closing the loop : autonomous experiment steered at the edge

On-chip data compression to tackle challenges from the increasing data rate

- Case 4 : Co-designing ASIC with on-chip compression (“SparkPix-RT”)

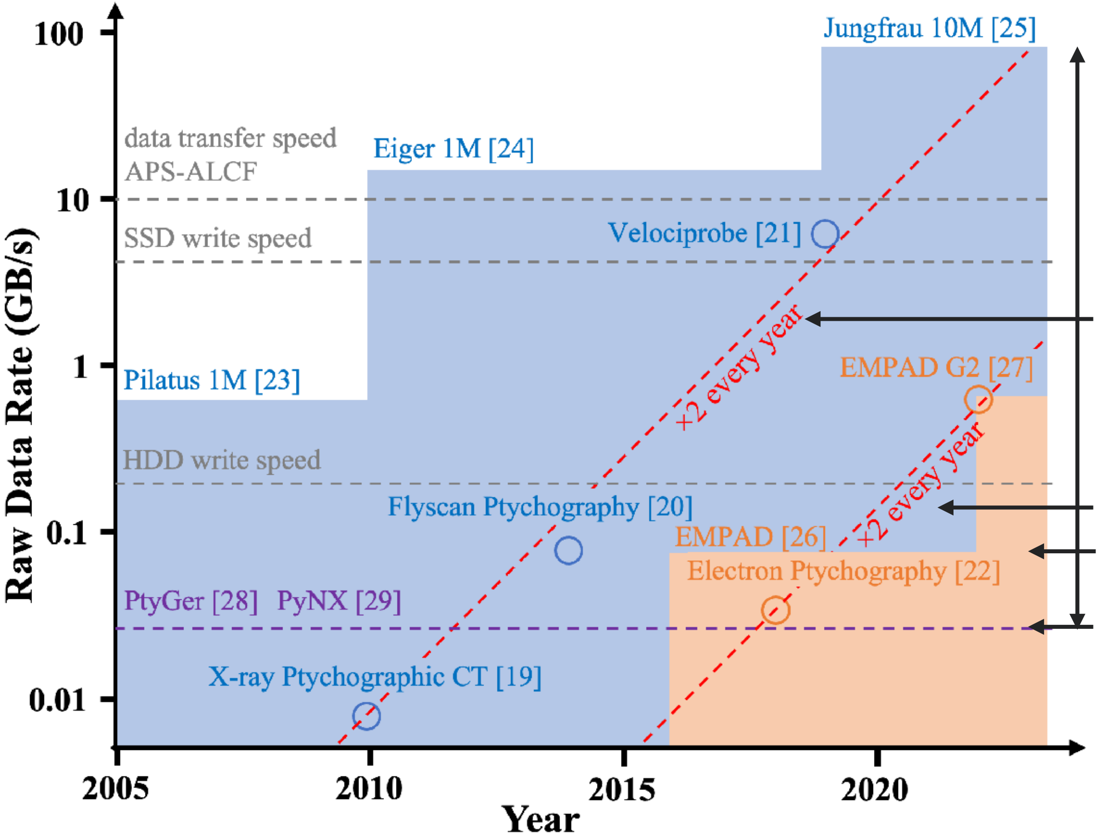
PTYCHOGRAPHY

is a high-resolution computational imaging technique



Thibault, P., et al. (2008). *Science (New York, N.Y.)*, *321*(5887), 379–382.

MOTIVATION



1 second of data acquired with the latest generation of detector, takes 1 hour to analyze

how to provide real time feedback during the experiment ?

raw data rate of x-ray ptychography doubles every year

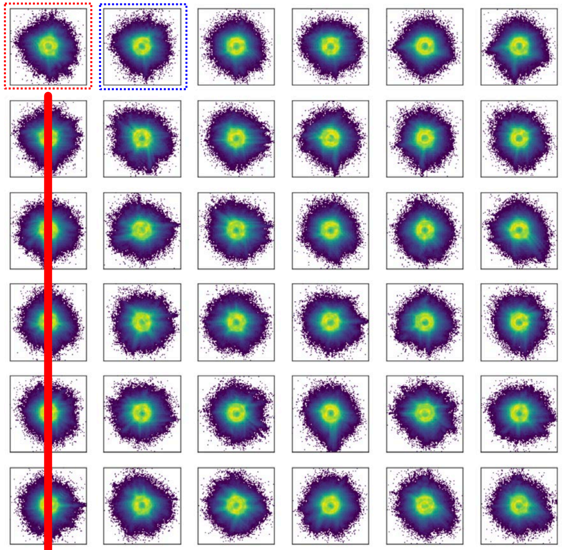
electron ptychography follows the same trend

for x-ray ptychography the raw data speed reached 100 MB/s in 2014

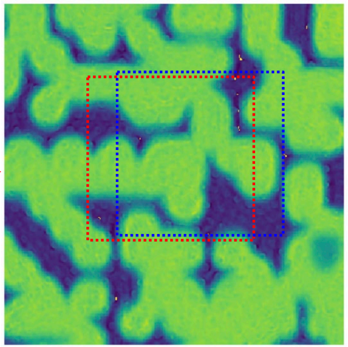
this exceeds the throughput of state-of-the-art iterative engines published in 2021

A. Babu, T. Zhou et al., Nature Communications 14 (1), 7059, 2023

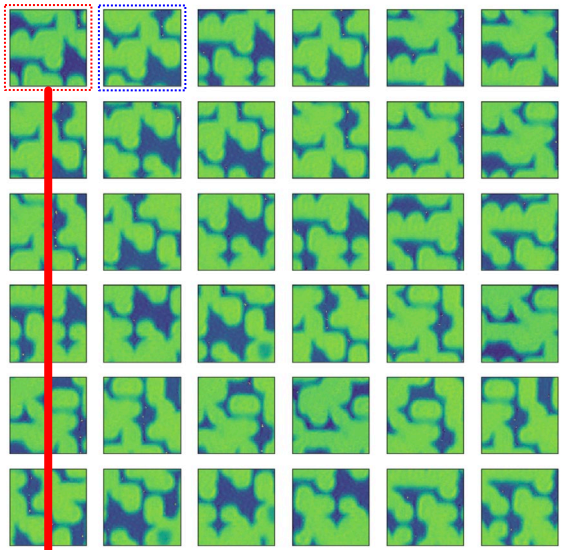
TEACH AI ABOUT PHASE RETRIEVAL



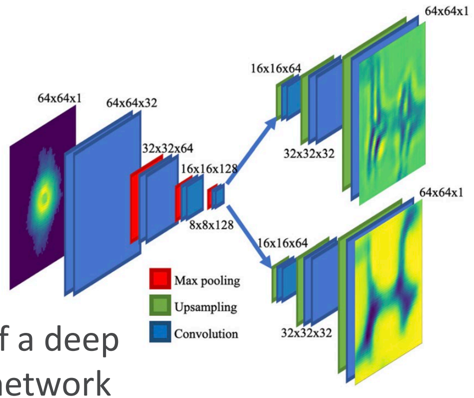
diffraction images
of the scan



result of iterative
phase retrieval



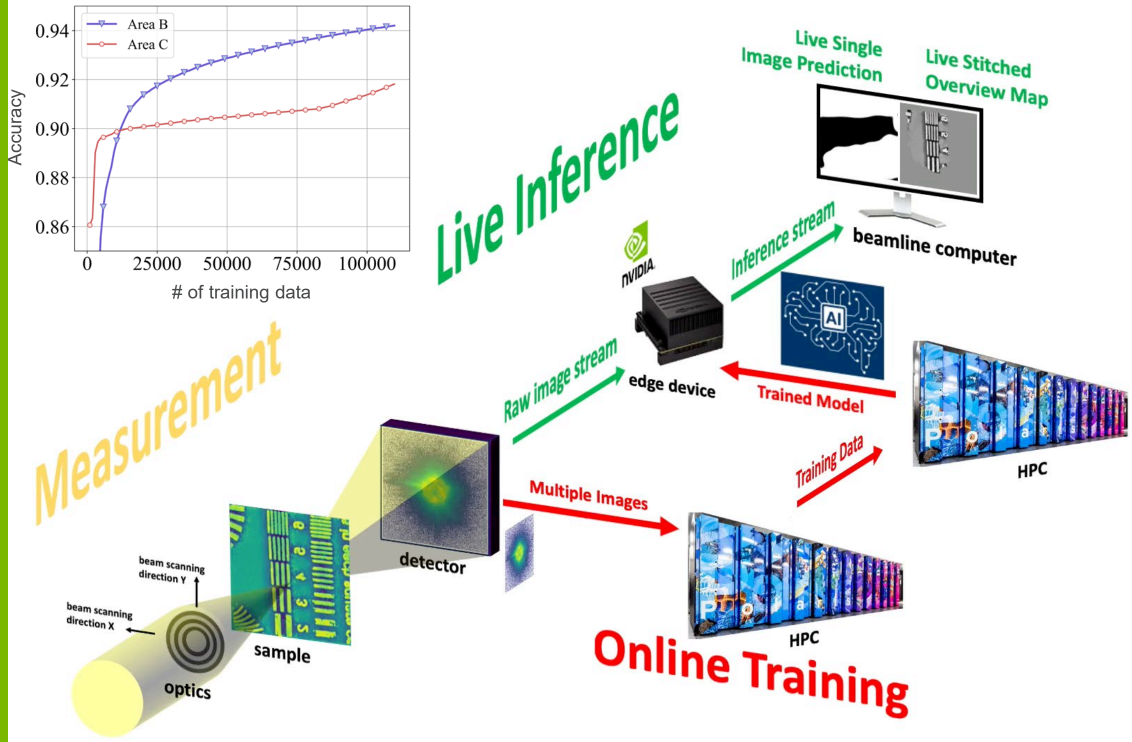
corresponding real space
(phase) images



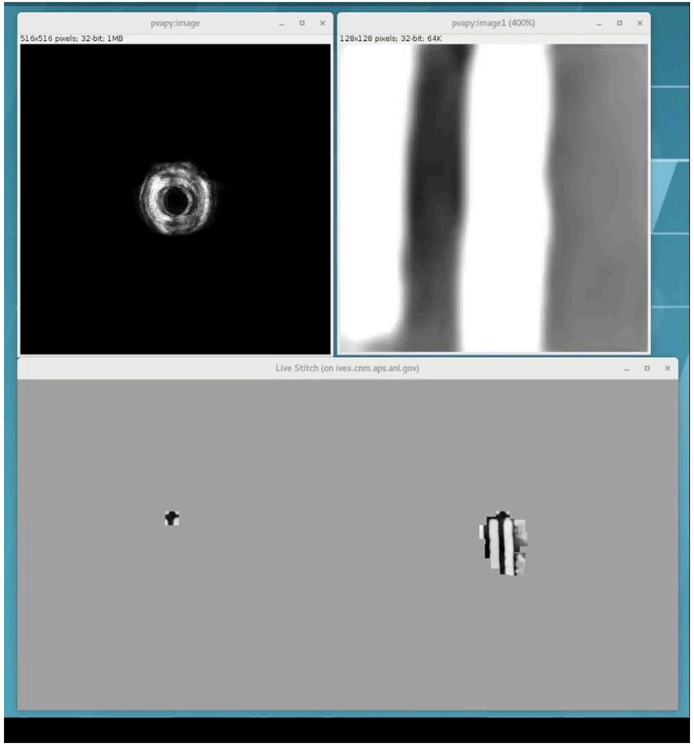
training of a deep
learning network

M. J. Cherukara, T. Zhou, et al. Appl. Phys. Lett. 117, 044103 (2020)

REAL TIME STREAMING PTYCHOGRAPHY



schematic of the workflow



video recording CNM-APS HXN beamline

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

A. Babu, T. Zhou et al., Nature Communications 14 (1), 7059, 2023

SCALABLE TO HANDLE HIGH DATA RATE

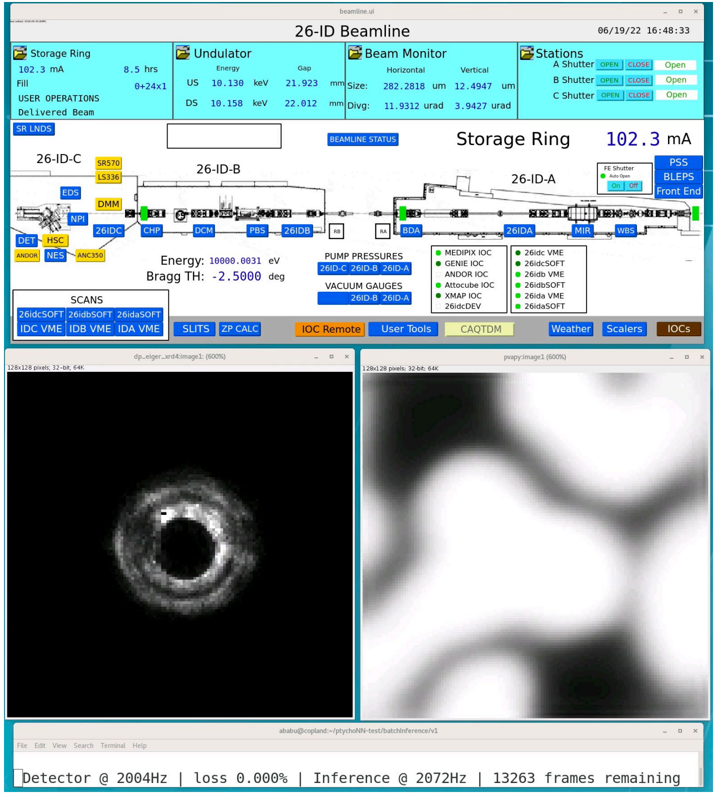
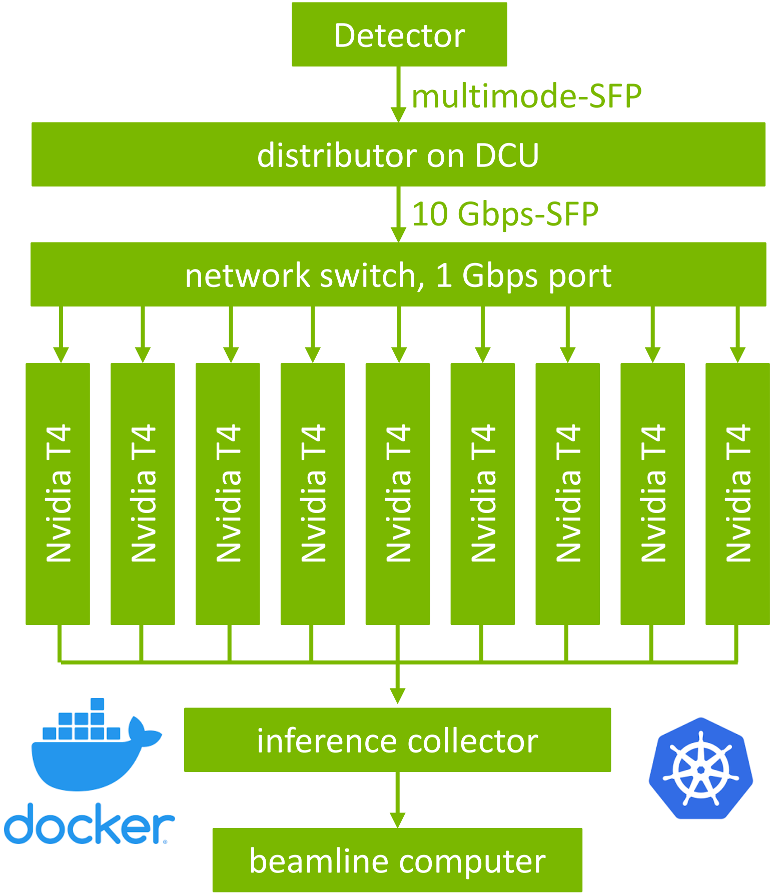


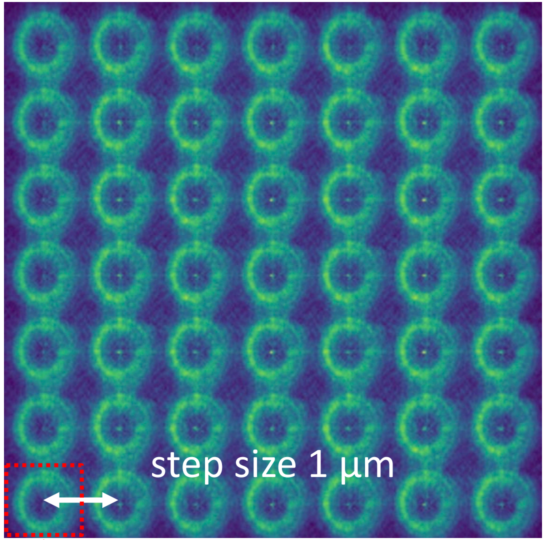
image size 512 x 512, 16 bit-depth, 2 kHz
raw data rate = 8 Gbps

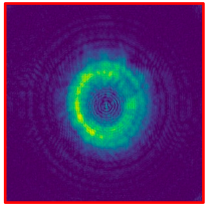


Yongho Kim, Seongha Park et al., publication in progress

REMOVING THE OVERLAP CONSTRAINT

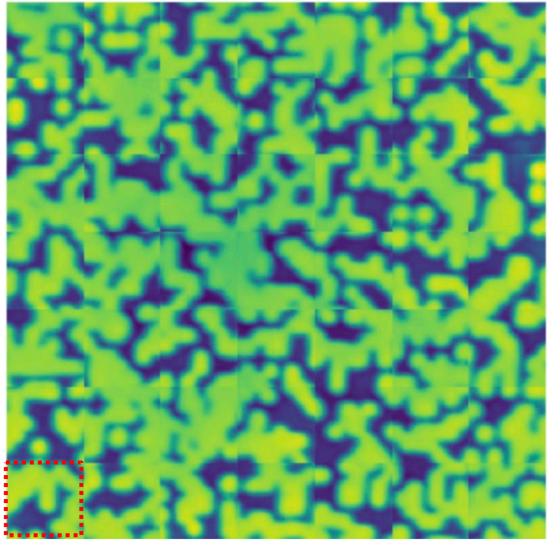
a scan with 7x7 points





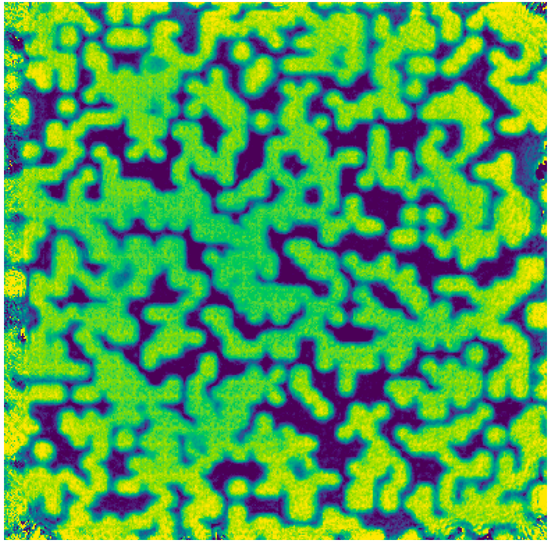
probe FWHM ~ 700 nm

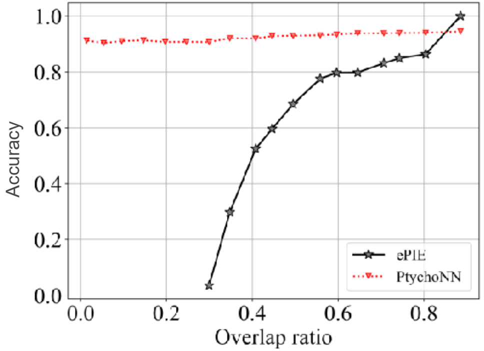
NN prediction (0% overlap)



step size 5×
reduce beam damage by 25×
increased FoV by 25×

ePIE (71% overlap)





A. Babu, T. Zhou et al., Nature Communications 14 (1), 7059, 2023
Neural network methods for radiation detectors and imaging, Front. Phys., 22 February 2024

OUTLINE

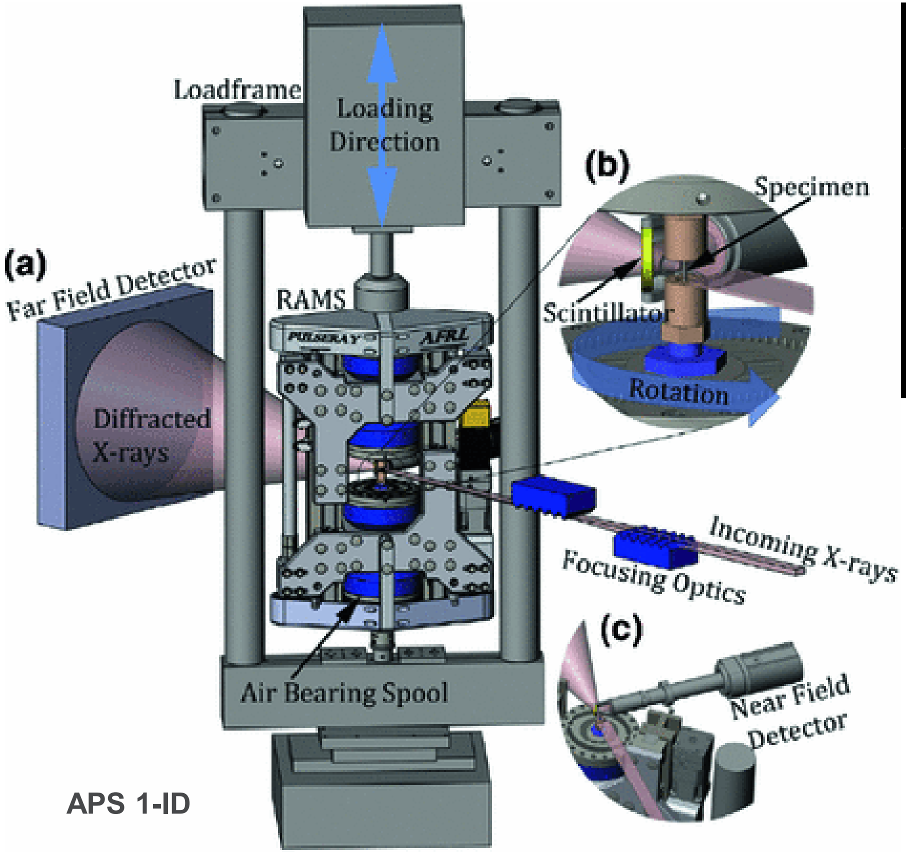
Machine learning based data reduction methods are highly efficient

- Case 1 : High-throughput on-the-fly data reduction at the edge
- Case 2 : Real time feedback: rapid detection of rare events in situ
- Case 3 : Closing the loop : autonomous experiment steered at the edge

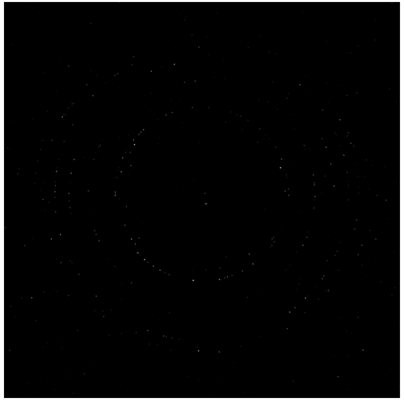
On-chip data compression to tackle challenges from the increasing data rate

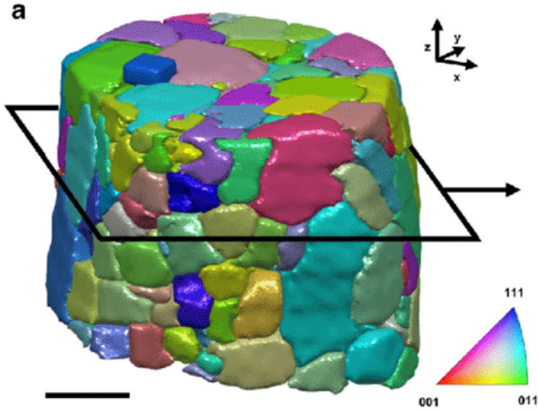
- Case 4 : Co-designing ASIC with on-chip compression ("SparkPix-RT")

HIGH ENERGY DIFFRACTION MICROSCOPY/3D-XRD



APS 1-ID



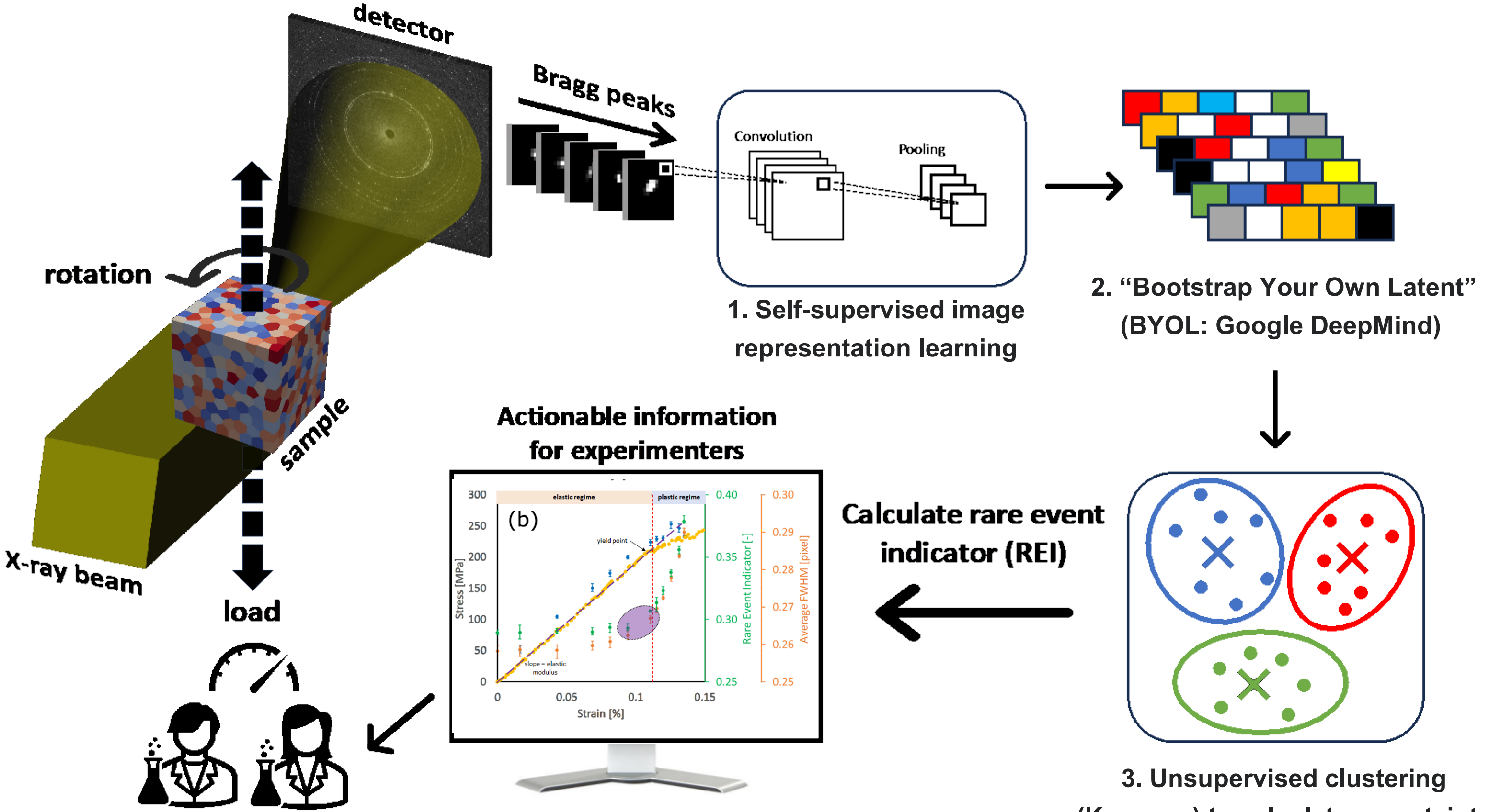


12 GBs dataset per loading acquired in 5 min

Efficient data reduction is required in *in situ* experiments to pause the loading for detailed measurements

Rapid detection of structural deformation with AI/ML method on the fly

RARE EVENT INDICATOR (REI)



50 times faster than conventional methods

Weijian Zheng, Jun-Sang Park, Peter Kenesei, Ahsan Ali, Zhengchun Liu, Ian T Foster, Nicholas Schwarz, Rajkumar Kettimuthu, Antonino Miceli, Hemant Sharma, Preprint: https://arxiv.org/abs/2312.03989

OUTLINE

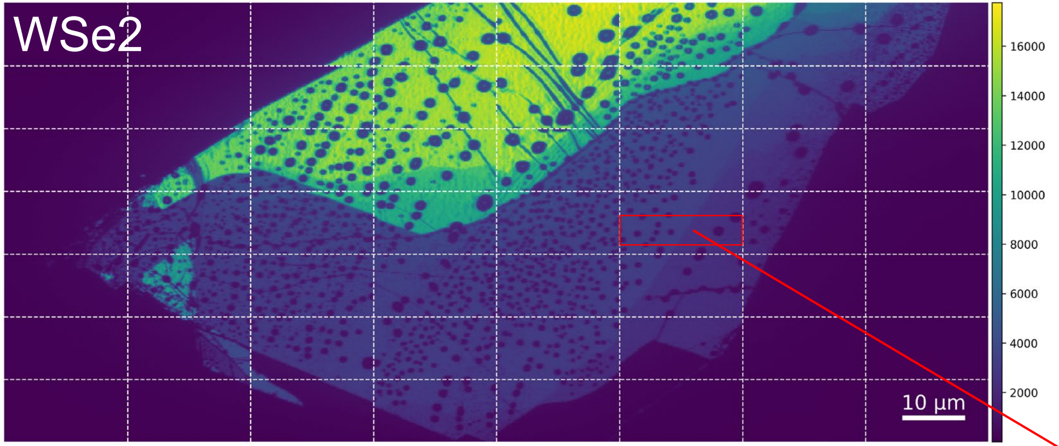
Machine learning based data reduction methods are highly efficient

- Case 1 : High-throughput on-the-fly data reduction at the edge
- Case 2 : Real time feedback: rapid detection of rare events in situ
- Case 3 : Closing the loop : autonomous experiment steered at the edge

On-chip data compression to tackle challenges from the increasing data rate

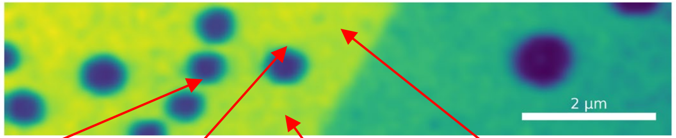
- Case 4 : Co-designing ASIC with on-chip compression (“SparkPix-RT”)

SCANNING X-RAY DIFFRACTION MICROSCOPY



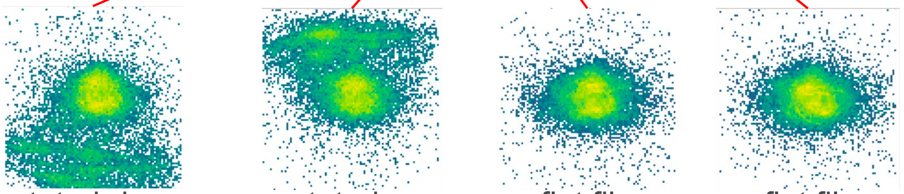
entire area is > 100 times larger
will take > 10 days to measure

40 x 200 points at 100 nm step
at 1 sec exposure time
takes > 2 hrs



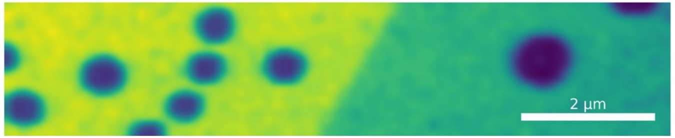
each points of measurement is associated
with a 2D detector image
each image carries information about
strain and lattice rotation

What if we only measure the most
important points (non-flat area) ?



rotated-down rotated-up flat film flat film

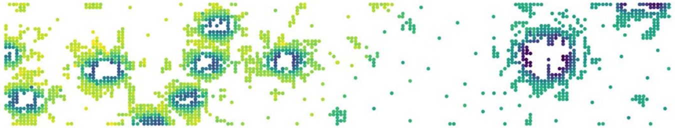
THE MOST IMPORTANT 25% POINTS



ground truth

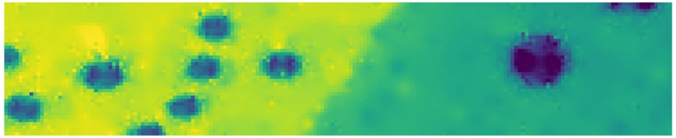
first 25%

random 25%



“smart” 25%

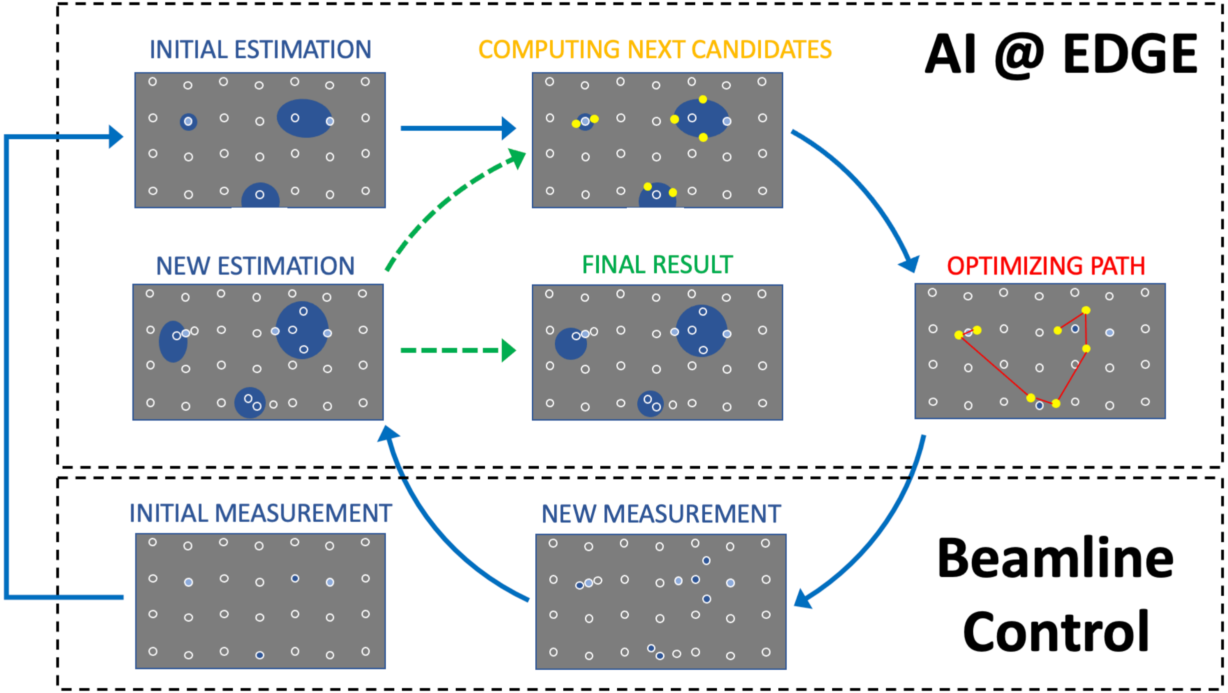
interpolation



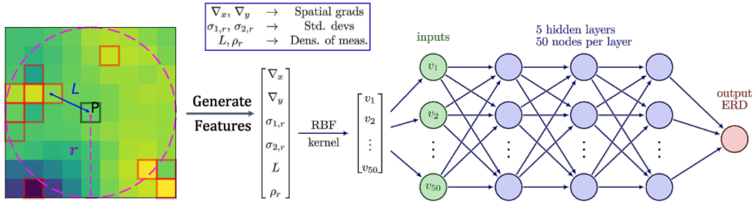
estimated results

S Kandel, T Zhou, et al., Nature Communications 14 (1), 5501

FAST AUTONOMOUS SCANNING TOOLKIT (FAST)

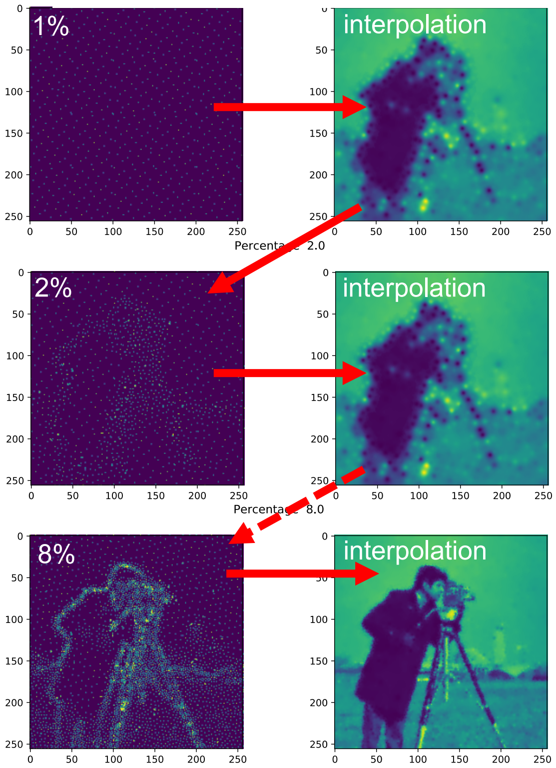


S Kandel, T Zhou, et al.,
Nature Communications 14 (1), 5501



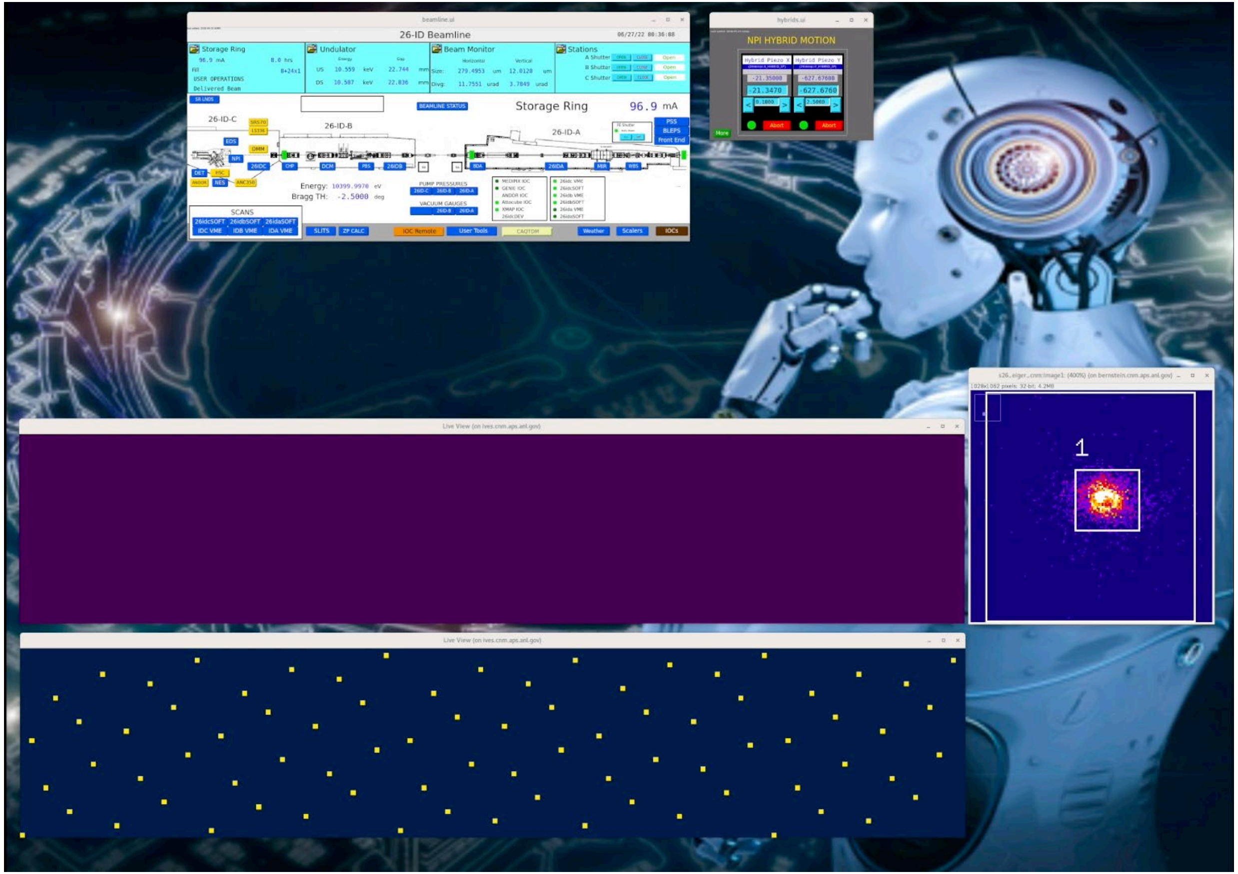
supervised learning approach for dynamic sampling (SLADS)
IEEE Transactions on Computational Imaging (Volume: 4, Issue: 1, March 2018)

Training using non-experimental data



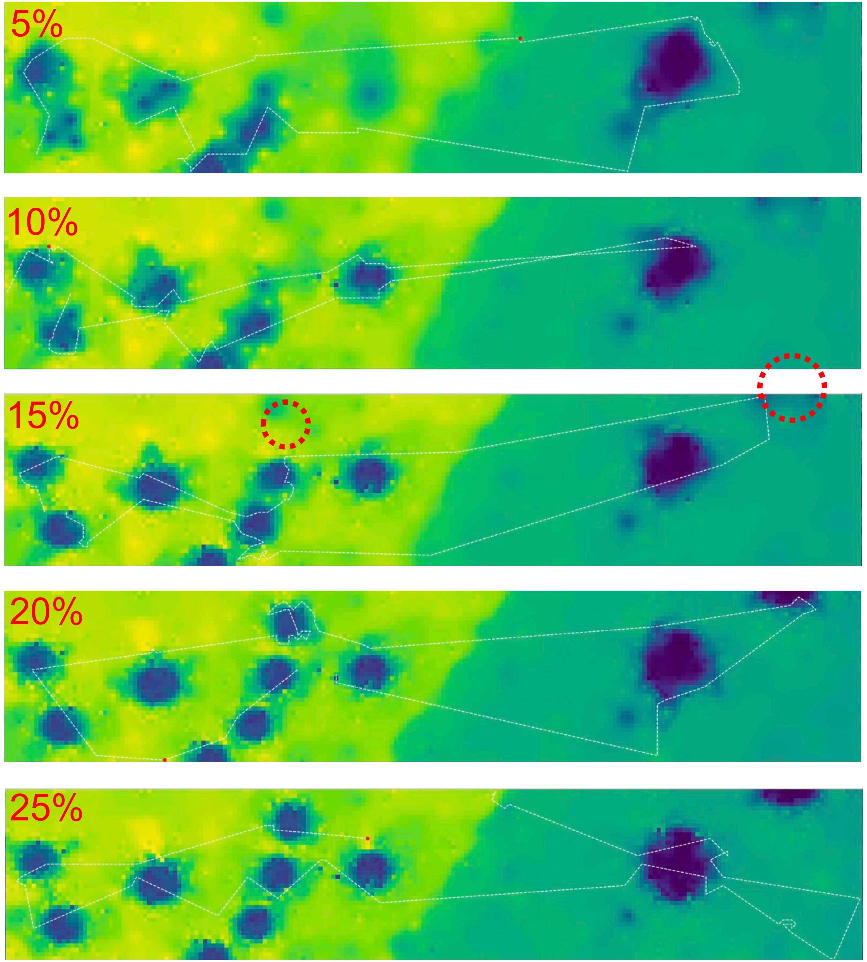
AI learns the relationship between available measured points and the choice of the next X% points that reduces the distortion the most

EXPERIMENTAL STEERING



S Kandel, T Zhou, et al., Nature Communications 14 (1), 5501

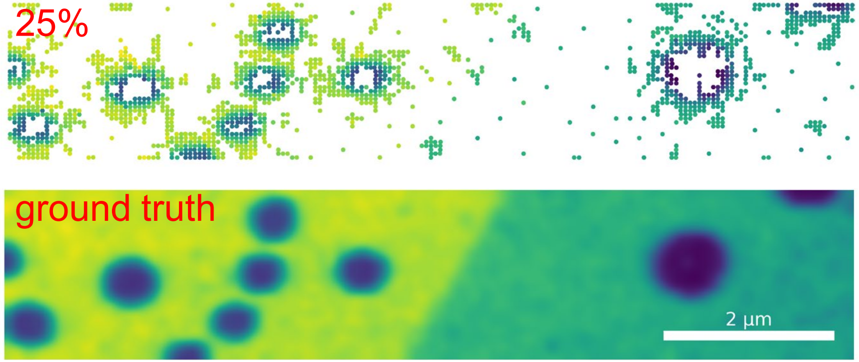
SUMMARY



ideal for sparse data
saves 80%* of the measurement time
* depends on the sparsity of the sample features

fast decision making (< 1s per 50 points)
with AI at the edge

exploratory experimental steering
on a synchrotron beamline
transferrable to other instruments
trained on camera man image and not on specific exp data



OUTLINE

Machine learning based data reduction methods are highly efficient

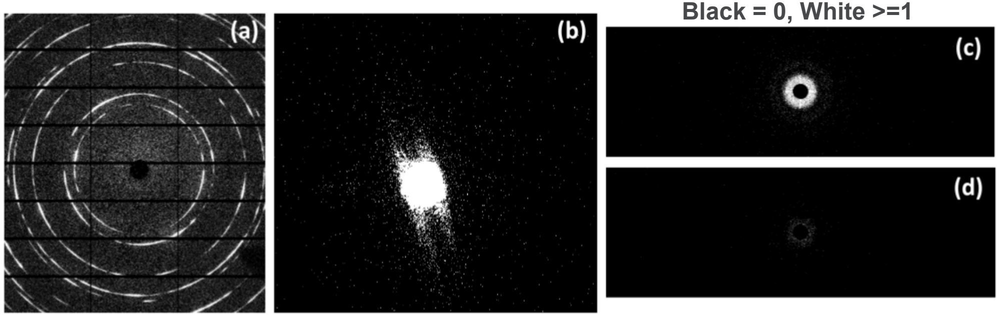
- Case 1 : High-throughput on-the-fly data reduction at the edge
- Case 2 : Real time feedback: rapid detection of rare events in situ
- Case 3 : Closing the loop : autonomous experiment steered at the edge

On-chip data compression to tackle challenges from the increasing data rate

- Case 4 : Co-designing ASIC with on-chip compression (“SparkPix-RT”)

CO-DESIGNING ASIC w/ ON-CHIP COMPRESSION

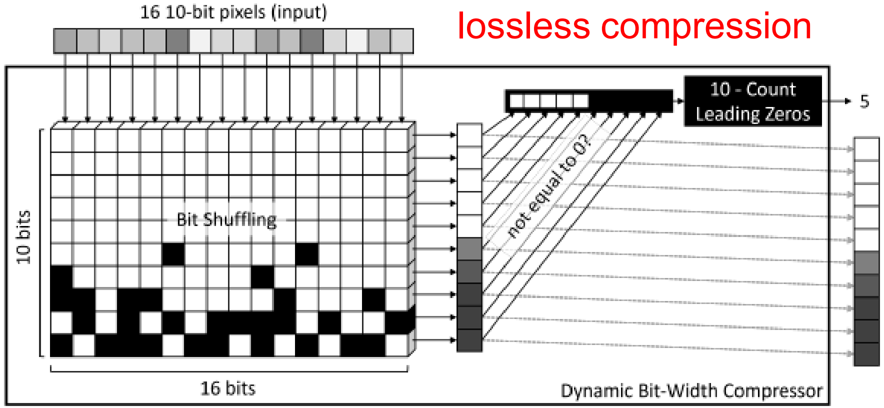
Black = 0, White >=1



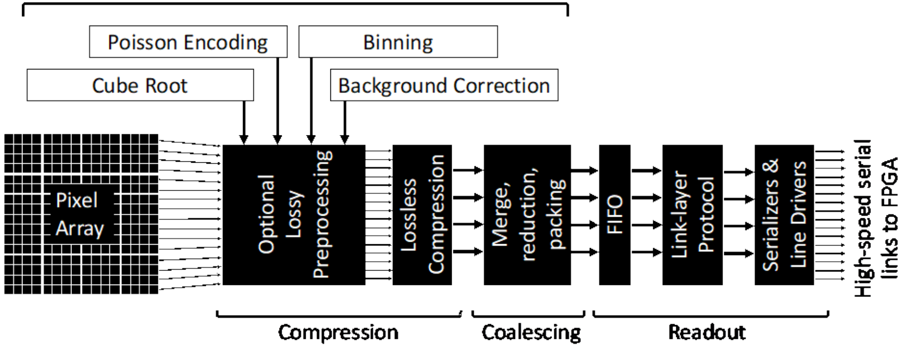
Statistics of zero-valued pixels per frame

Dataset	Mean %	Std Dev %	Minimum %	Maximum %
High-energy XRD	83.42	2.49	68.96	85.03
Ptychography	97.42	0.23	96.2	97.77
XPCS concentrated	98.62	0.03	98.02	98.68
XPCS dilute	99.9	0.01	99.8	99.91

lossless compression

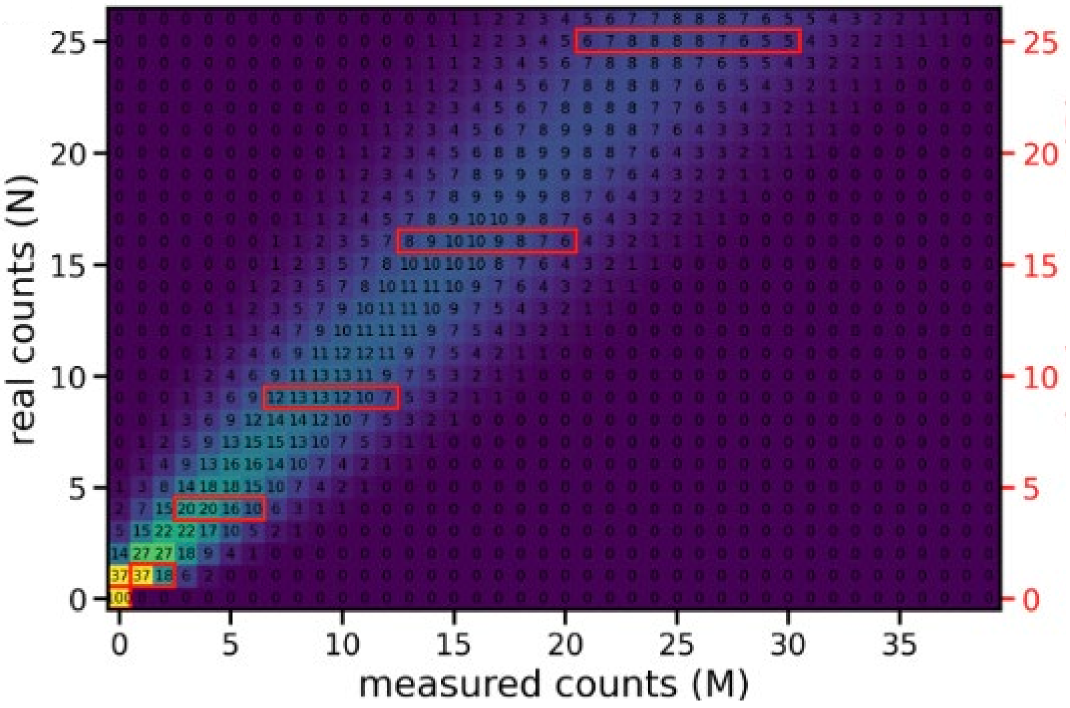


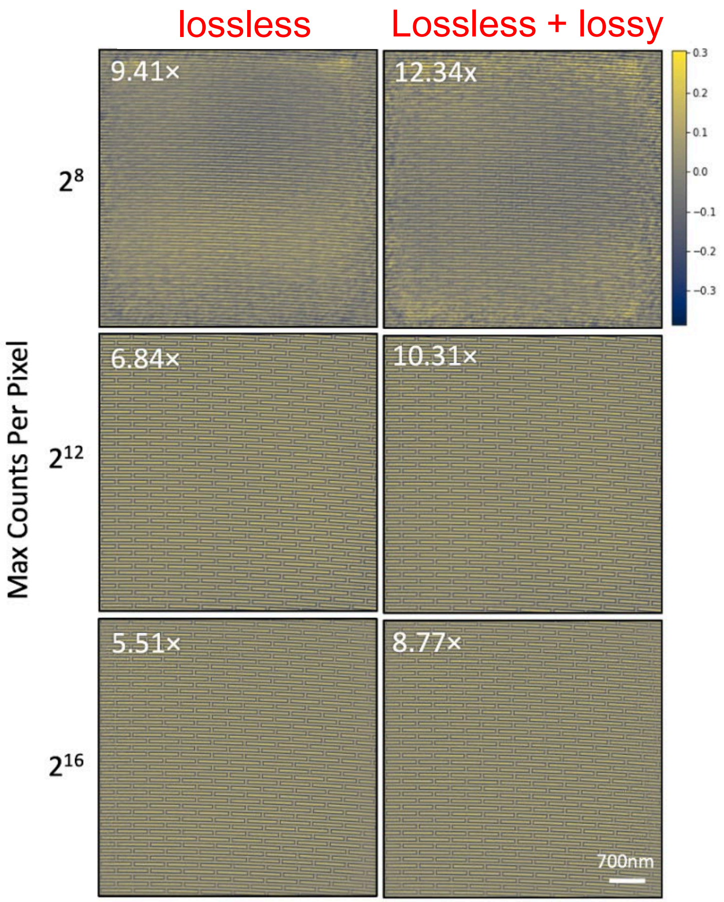
User configurable
lossy compressions



S Strempfer, T Zhou, et al., Journal of Instrumentation 17 (10), P10042

POISSON ENCODING



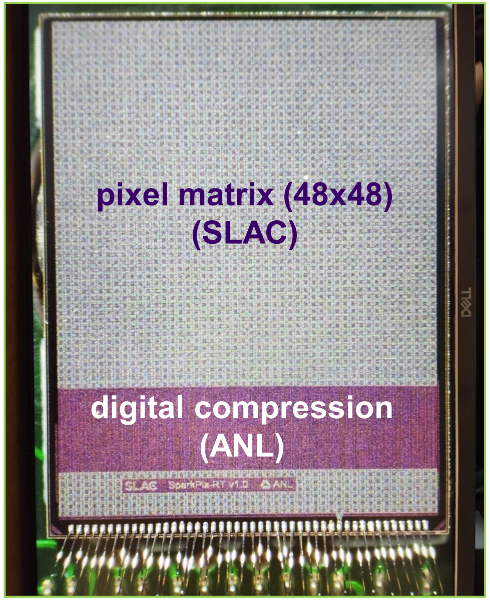


Taking into account the counting statistics of x-ray photon science

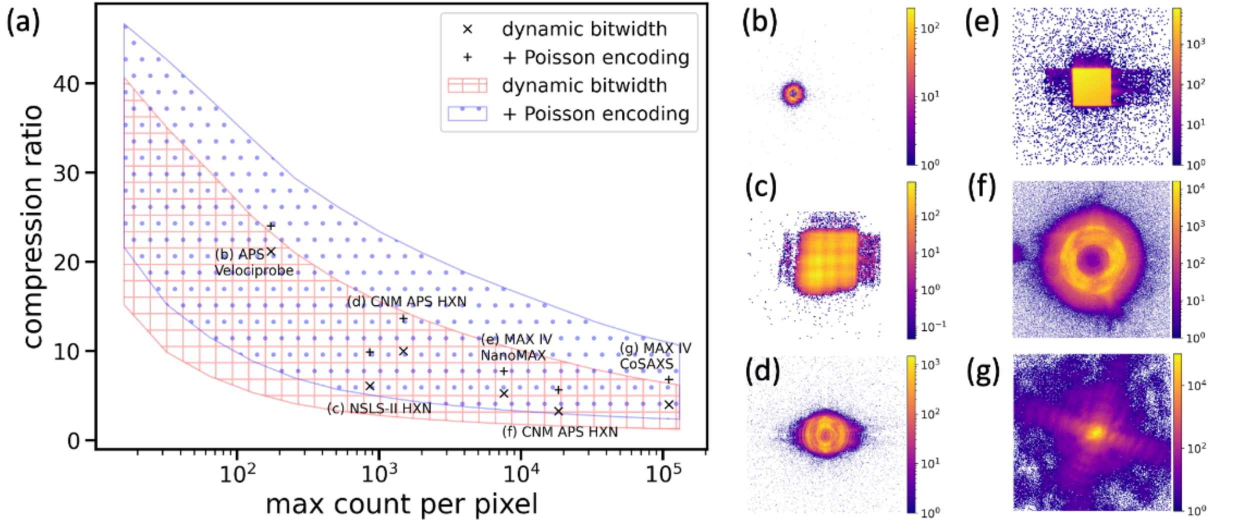
S Strempfer, T Zhou, et al., Journal of Instrumentation 17 (10), P10042

SPARKPIX-RT

Implementation ("RT1")



"RT2" (192x168)
tape-out in June 2024



Estimated compression performance based on experimental ptychography data from different beamlines and synchrotrons (APS, NSLS-II, MAX IV)

ACKNOWLEDGEMENT

Case 1: High throughput data reduction at the edge

- (APS) Anakha Babu, Saugat Kandel, Tekin Bicer, Zhengchun Liu, Willian Judge, Daniel J. Ching, Yi Jiang, Sinisa Veseli, Steven Henke, Yudong Yao, Antonino Miceli, Mathew J. Cherukara
- (DSL) Ryan Chard, Ian T. Foster
- (CNM) Martin V. Holt
- (NVidia) Ekaterina Sirazitdinova, Geetika Gupta

Case 2: Rapid detection of rare events

- (APS & DSL) Weijian Zheng, Jun-Sang Park, Peter Kenesei, Ahsan Ali, Zhengchun Liu, Ian T. Foster, Nicholas Schwarz, Rajkumar Kettimuthu, Antonino Miceli, Hemant Sharma

Case 3: Closing the loop : Experimental steering at the edge

- (APS) Saugat Kandel, Anakha Babu, Antonino Miceli, Mathew J. Cherukara
- (CNM) Xinxin Li, Xuedan Ma, Martin V. Holt
- (MCS) Zichao Di
- (MSD) Charudatta Phatak

SPARKPIX-RT DESIGN TEAM

SLAC ASIC/FPGA Design

- **Dionisio Doering –** system architecture, project leader
- **Aseem Gupta –** RT cluster logic design, RT/RT2 top-level layout, RT/RT2 readout and full-chip verification, mixed-signal verification
- **Hyunjoon Kim –** cocotb, verification
- **Pietro King –** RT/RT2 FPGA design, emulation, verification, test setup
- **Lorenzo Rota –** architecture, cluster logic design, simulations, analog, etc.

Argonne ASIC/FPGA Design

- **Antonino Miceli** – project leader ANL
- **Mike Hammer –** RT/RT2 digital design and verification
- **Henry Shi –** RT/RT2 physical implementation and PnR, RT verification, RT2 AXI control design and verification
- **John Weizeorick –** RT AXI control design and verification, RT carrier board design, RT FPGA design and testing
- **Kazutomo Yoshii –** compression algorithm development, compressor design and verification, formal verification
- **Tao Zhou –** beamlines scientist, compression algorithms development, compressor design, compression usability

Students

- **Senthil Gnanasekaran –** cocotb verification/emulation, serial control (SUGOI) driver, control & data path checking
- **Sebastian Strempfer –** RT/RT2 digital design, analysis of compression techniques and designs, cocotb verification

THANK YOU FOR YOUR ATTENTION !