

Finding a Common Data Language at Accelerator Facilities & PEARLS Update from DESY Hamburg

A visualization of a particle beam, showing a series of blue spheres (representing particles) moving along a diagonal path towards a bright, multi-colored starburst light source in the upper right corner.

Pardis Niknejadi

1st PEARLS Days Workshop
Freiburg 23.09.2026

HELMHOLTZ RESEARCH FOR
GRAND CHALLENGES



Bundesministerium
für Forschung, Technologie
und Raumfahrt

universität freiburg



Outline

- PEARLS DESY crew
- Case for Encoding Data at Large Accelerator Facilities like FLASH at DESY
- Available Diagnostics and Data Management Pipelines
- What's Missing?
- PEARLS Encoding objective
- Update from PEARLS DESY - Hamburg

PEARLS people at DESY + Current Collaborators



Dr. Pardis Niknejadi

DESY Hamburg
Project lead@DESY
(MPY)



Dr. Simran Gurdasani

DESY Berlin-Zeuthen



Gesa Goetzke

DESY Hamburg (MPY)



Dr. Kai Tiedtke

DESY Hamburg
(Photon Science)



Ruth Rudolph

DESY Hamburg
M - Finaces



Larissa Prochnow

DESY Hamburg (MPY)
@ Berlin-Zeuthen



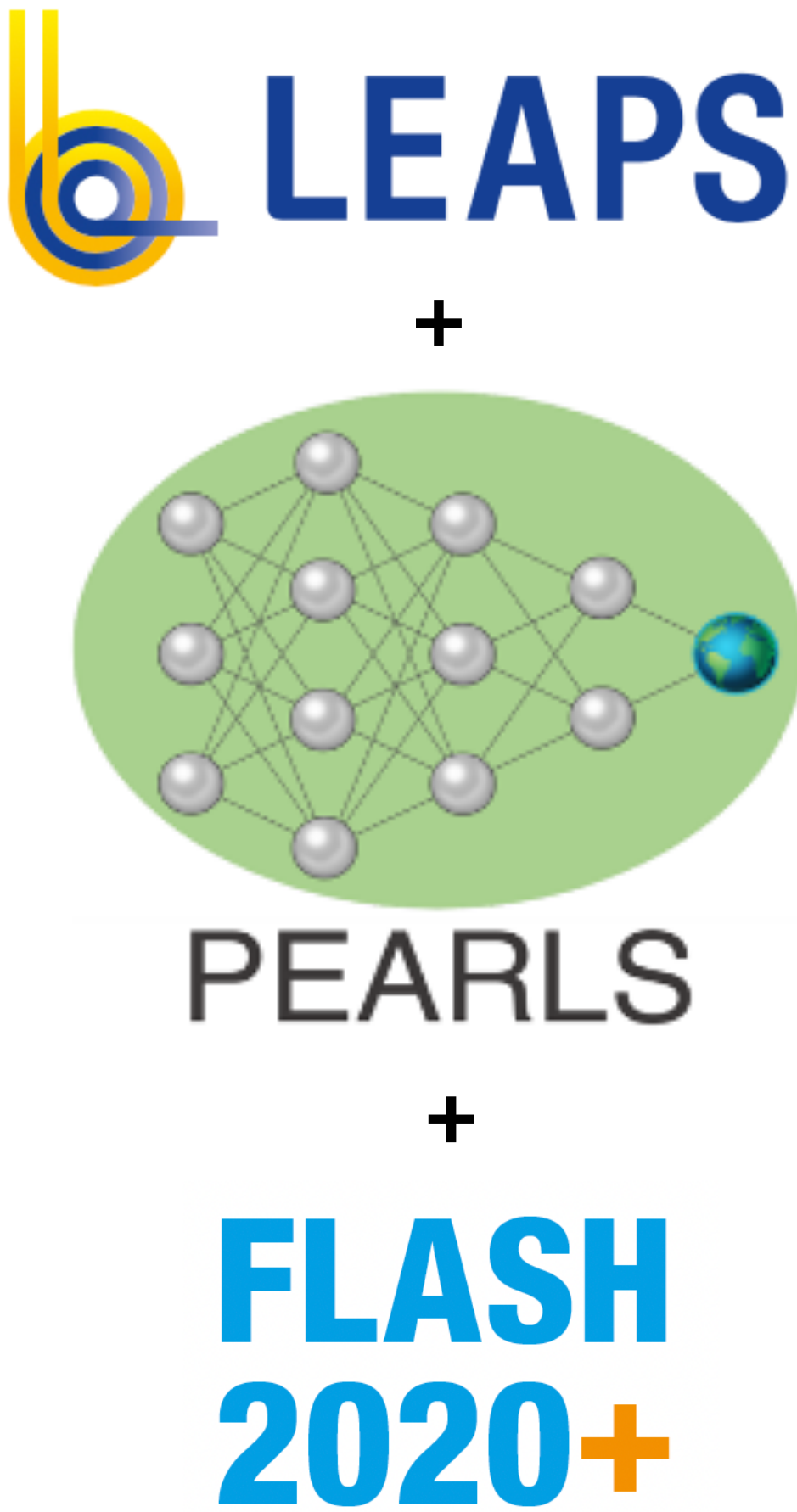
New Postdoc

DESY Hamburg (MPY)



Dr. Lucas Schaper

DESY Hamburg (MPY)



PEARLS people at DESY + Current Collaborators



Dr. Pardis Niknejadi

DESY Hamburg
Project lead@DESY
(MPY)



Ruth Rudolph

DESY Hamburg
M - Finances



**Dr. Simran
Gurdasani**

DESY Berlin-Zeuthen



Gesa Goetzke

DESY Hamburg (MPY)



New Postdoc

DESY Hamburg (MPY)



Larissa Prochnow

DESY Hamburg (MPY)
@ Berlin-Zeuthen



Dr. Kai Tiedtke

DESY Hamburg
(Photon Science)

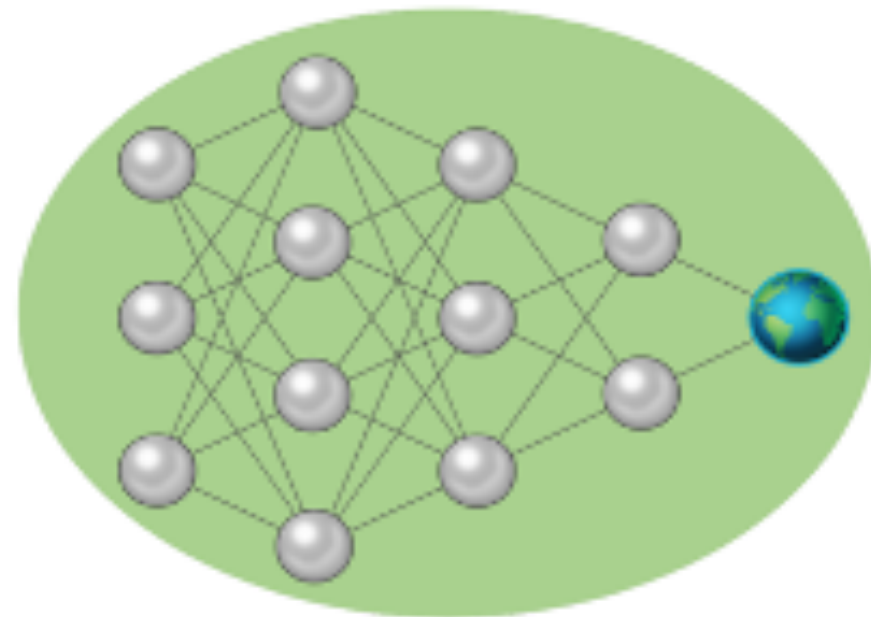


Dr. Lucas Schaper

DESY Hamburg (MPY)



+



PEARLS

+

**FLASH
2020+**

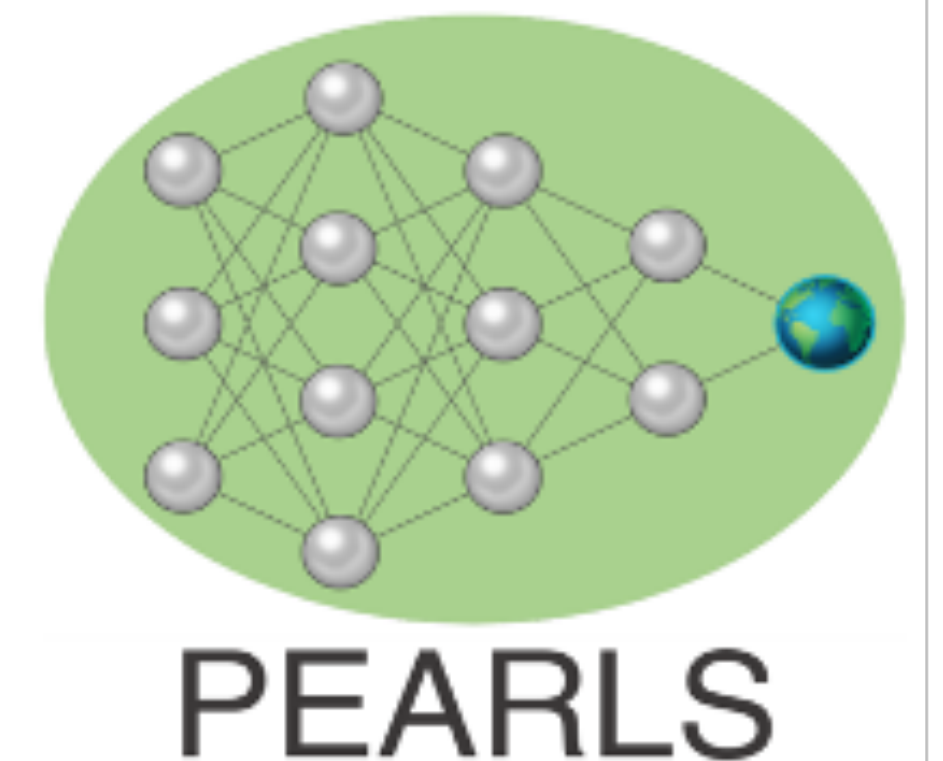
PEARLS: Precision in Energy-aware AI Research for Low-carbon Solutions

Goal: "Reduce the carbon footprint of AI & ML in high-energy physics research — without losing accuracy."

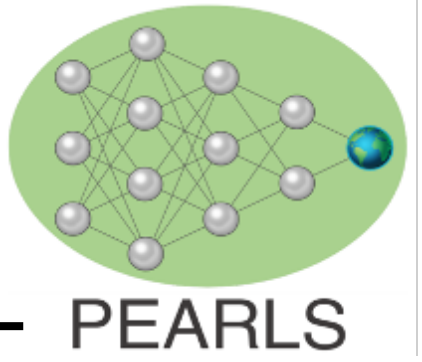
There is no shortage of motivation:

"Under the continuation of current trends, AI electricity consumption could reach up to 30% of global demand by 2050, corresponding to more than 8 gigatons of annual CO₂-equivalent emissions."

Scenario-based forecasting of the global energy demand and carbon footprint of artificial intelligence. *PLOS ONE*, 21 (3): e0343056. DOI: 10.1371/journal.pone.0343056



DESY Working Groups and Their Goals (Reduce the Model Size)



PRUNING:

The challenge: NNs have millions of parameters — many contribute little to predictions.

The approach: Use information flow mapping to identify and remove redundant nodes.

The payoff: Smaller model = faster inference = less energy = less CO₂. First time quantified in HEP.

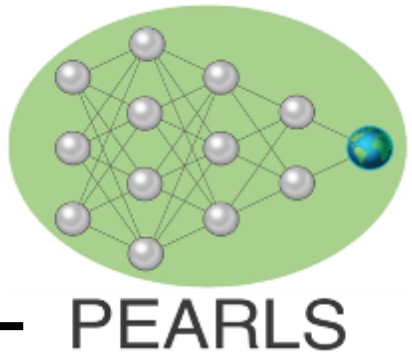
DATA ENCODING:

The challenge: HEP data formats (HDF5, ROOT) are not well-suited to most ML workflows.

The approach: Investigate model-based compression techniques that enable efficient retrieval and reuse of prior computations for sustainable ML workflows.

The payoff: Less preprocessing overhead, better GPU/CPU utilization, faster and greener training.

DESY Working Groups and Their Goals



PRUNING:

Covered by Simran yesterday

The challenge: NNs have millions of parameters — many contribute little to predictions.

The approach: Use information flow mapping to identify and remove redundant nodes.

The payoff: Smaller model = faster inference = less energy = less CO₂. First time quantified in HEP.



Topic for today: Why current data formats are not well suited for most ML workflows

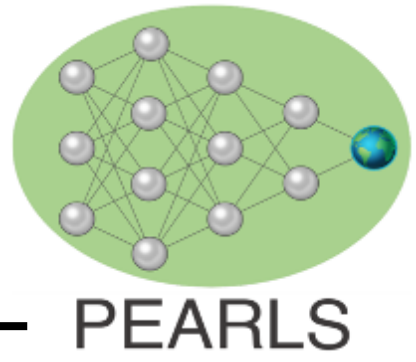
DATA ENCODING:

The challenge: HEP data formats (HDF5, ROOT) are not well-suited to most ML workflows.

The approach: Investigate model-based compression techniques that enable efficient retrieval and reuse of prior computations for sustainable ML workflows.

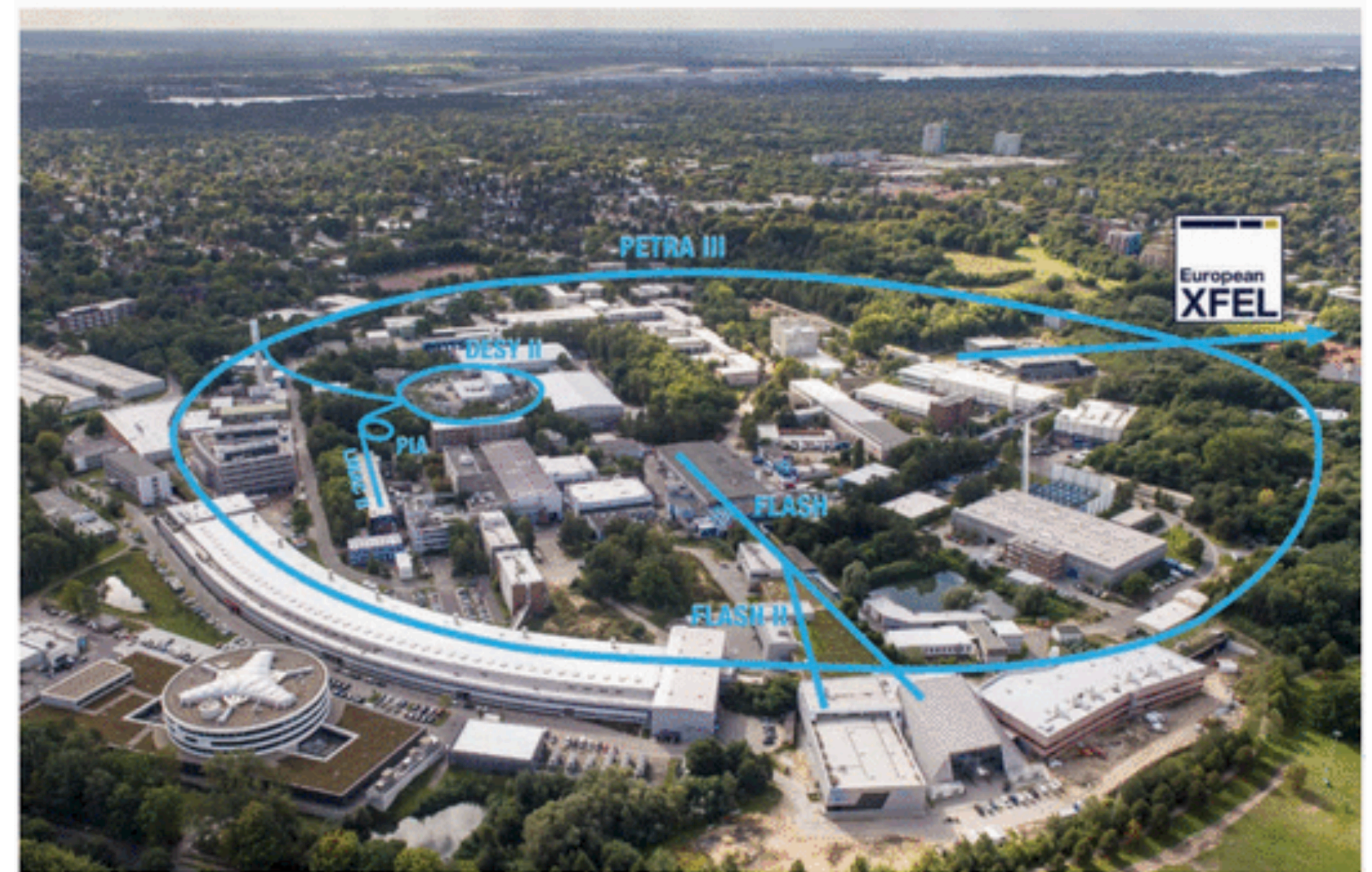
The payoff: Less preprocessing overhead, better GPU/CPU utilization, faster and greener training.

Why Encode Data at an Accelerator Facility/Light Source Facility



For example, DESY (which is Both an Accelerator and a Light Source Facility)

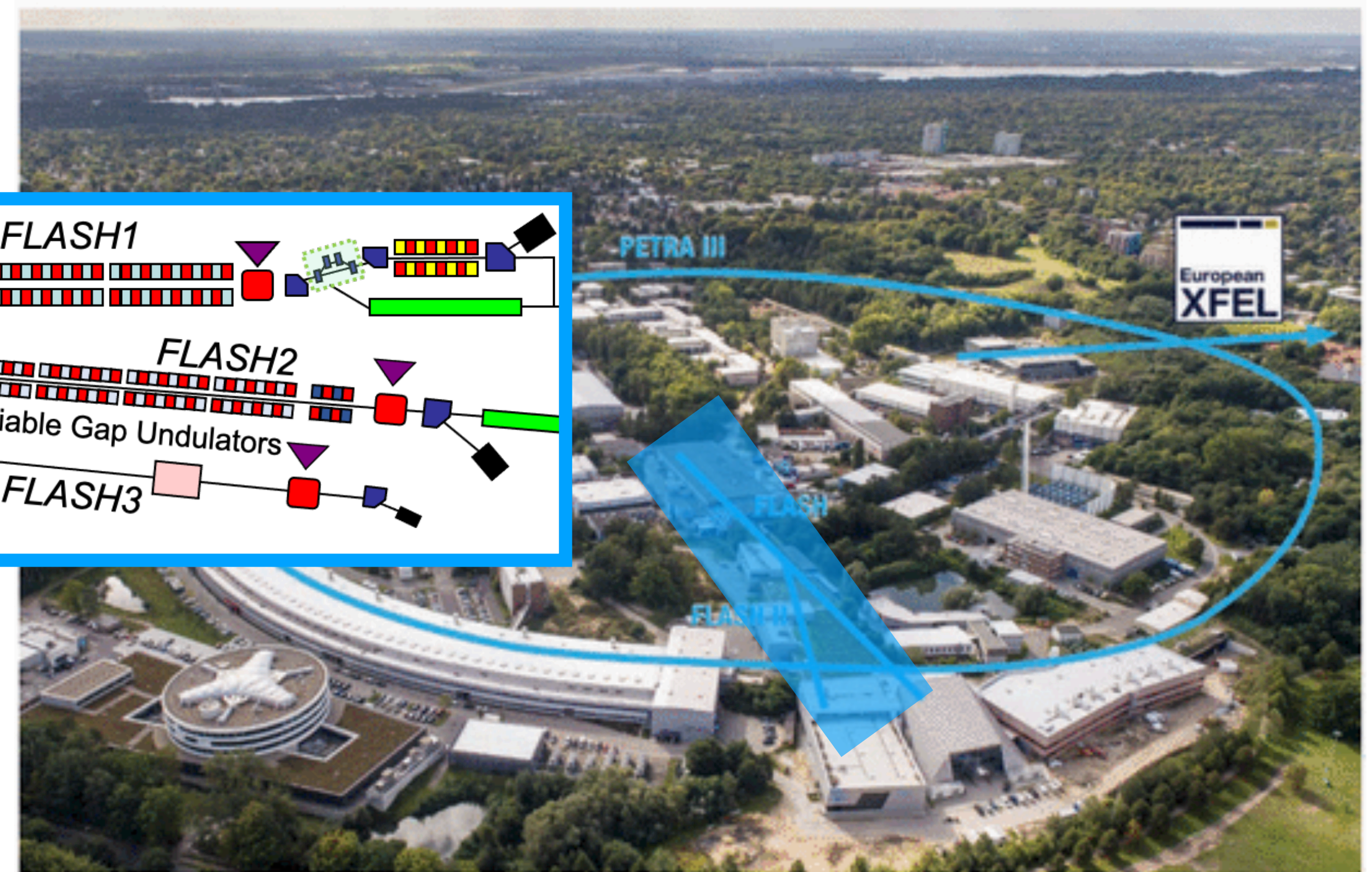
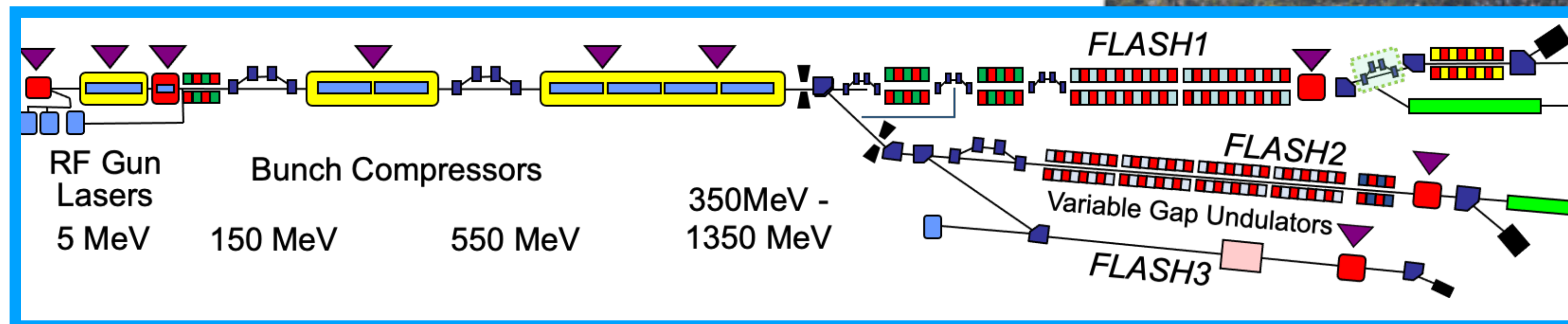
- Large infrastructure with lots of instrumentation
- Many researchers
- Many Users
- International Experiments
- Large Datasets
- Dedicated computing structures for control, postprocessing, and analysis



Why Encode Data at an Accelerator Facility/Light Source Facility

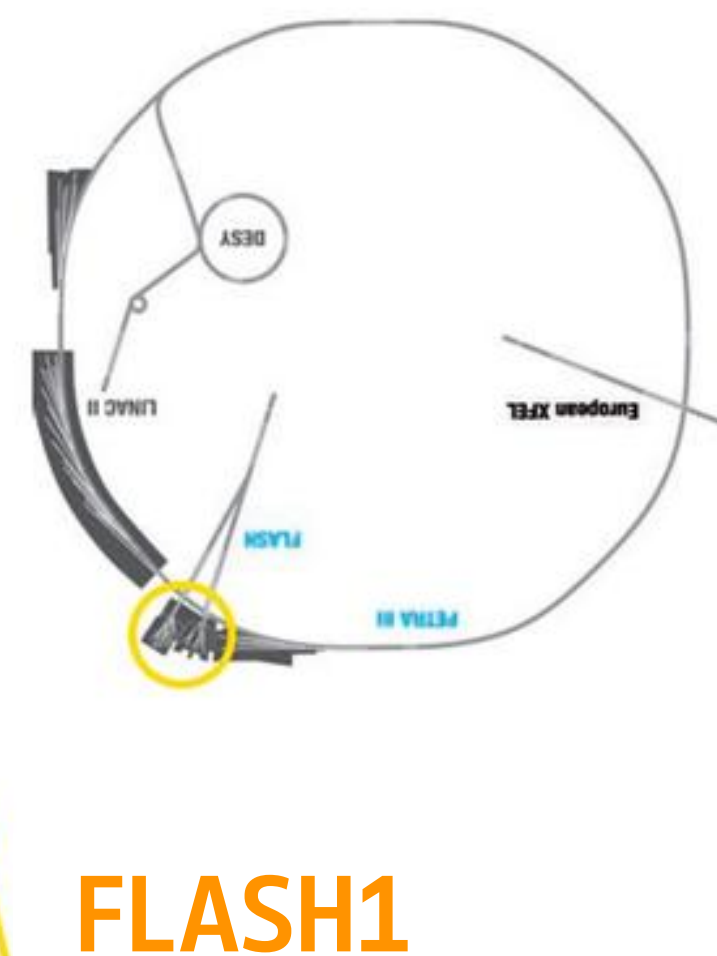
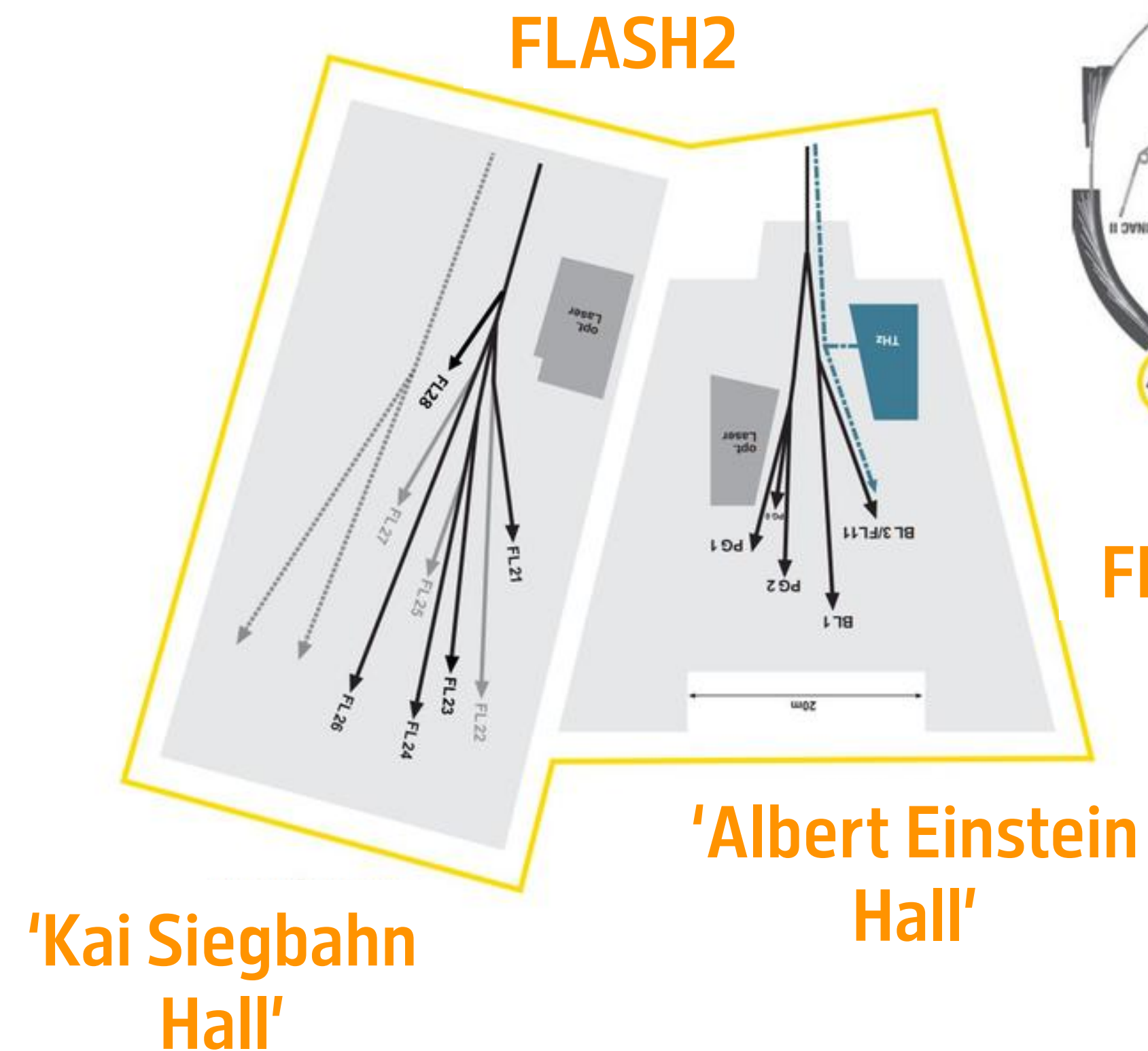
The Accelerator and Free-electron Laser Beamlines

First User Facility FEL and Blueprint for the Machines that came after



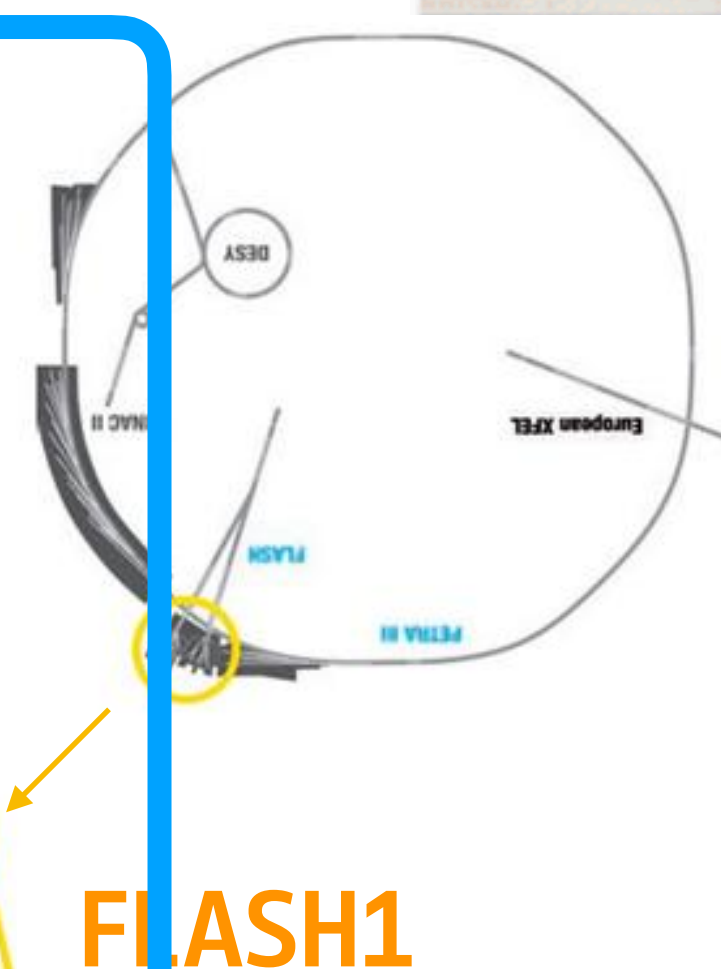
Why Encode Data at an Accelerator Facility/Light Source Facility

The Photon Science Beamlines

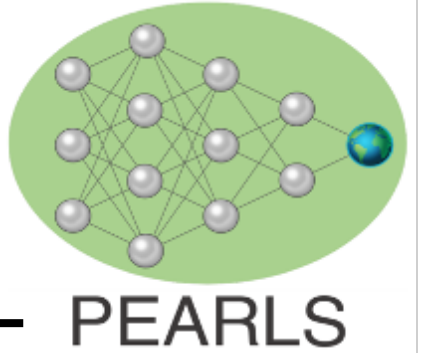


Why Encode Data at an Accelerator Facility/Light Source Facility

The Photon Science Beamlines

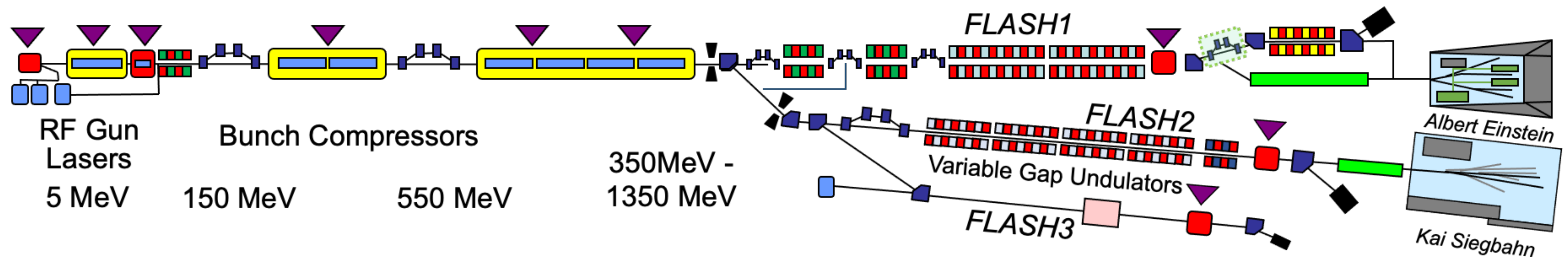


Think of FLASH as a camera with very high resolution & Shutter Speed



The Accelerator and Free-electron Laser Beamlines

FLASH Has A Swiss Army FEL Design philosophy



FLASH1:

Narrow Spectral Bandwidth
Shot-to-shot Stability
Longitudinal Coherence

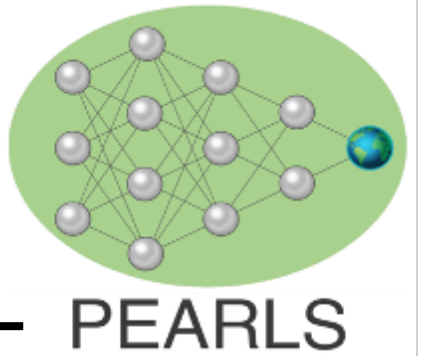
FLASH2:

High Pulse Energy
Short pulse schemes
Highest Photon Energy (with afterburner)

FLASH3:

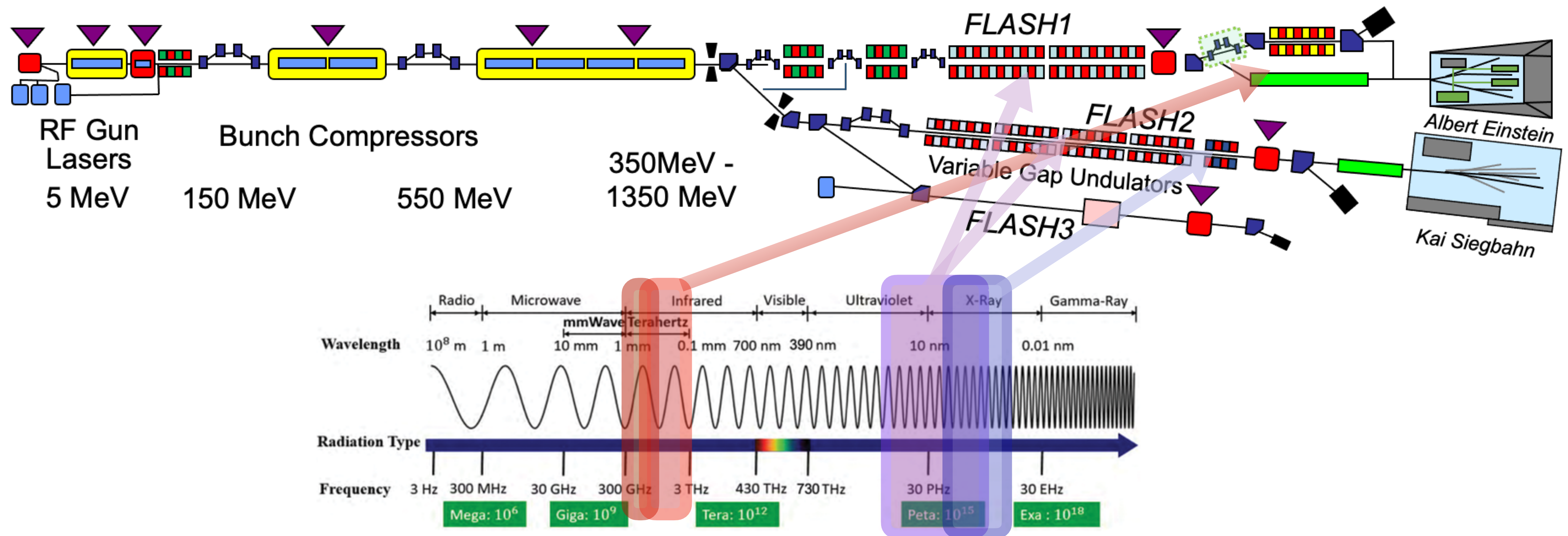
Plasma Acceleration
(investing in future generation of accelerators)

Why Encode Data at an Accelerator Facility/Light Source Facility

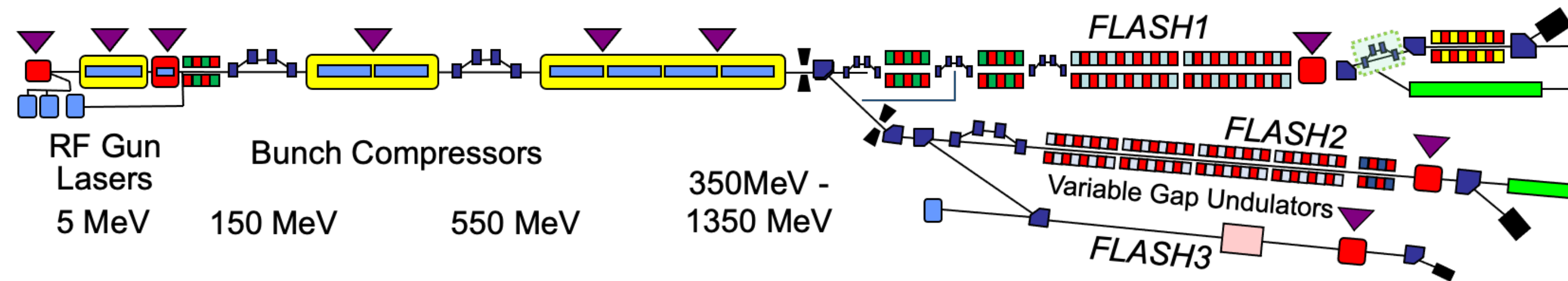


The Accelerator and Free-electron Laser Beamlines

FLASH Has A Swiss Army FEL Design philosophy

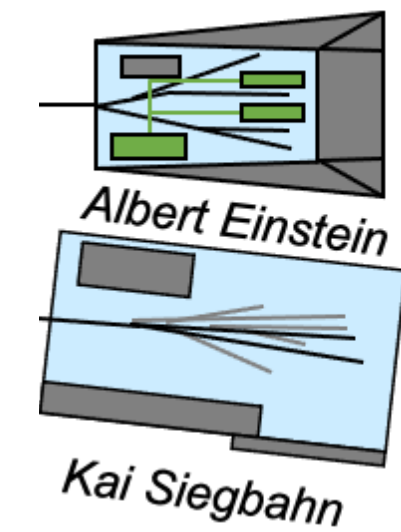


One Facility, Two Communities



Accelerator

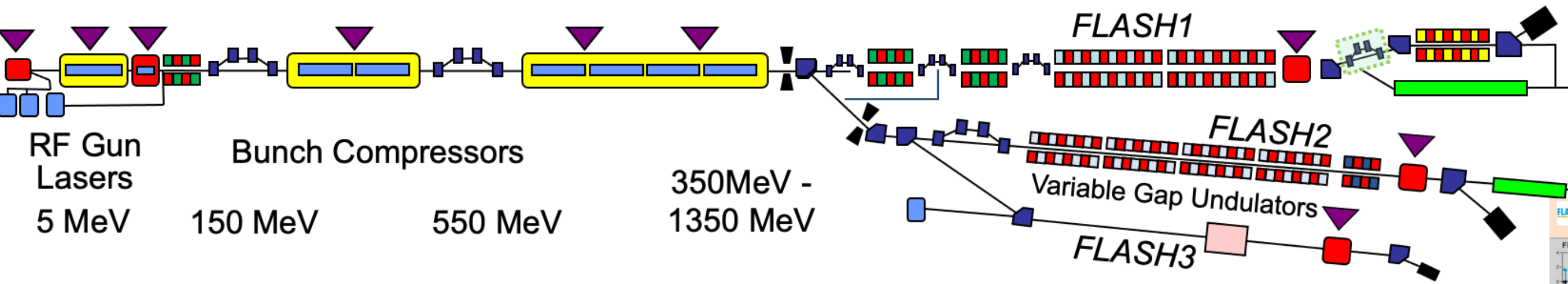
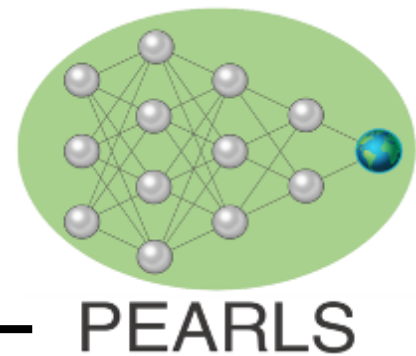
- ▶ Machine operation
- ▶ Beam dynamics
- ▶ Controls
- ▶ Setpoints and readbacks
- ▶ Time series and waveforms
- ▶ Bunch/train parameters
- ▶ Feedback and optimization



Photon Science

- ▶ Scientific experiments
- ▶ Photon-sample interaction
- ▶ Detector acquisition
- ▶ Images and spectra
- ▶ Event streams
- ▶ Scans and coordinates
- ▶ Reconstruction and analysis

One Facility, Two Communities



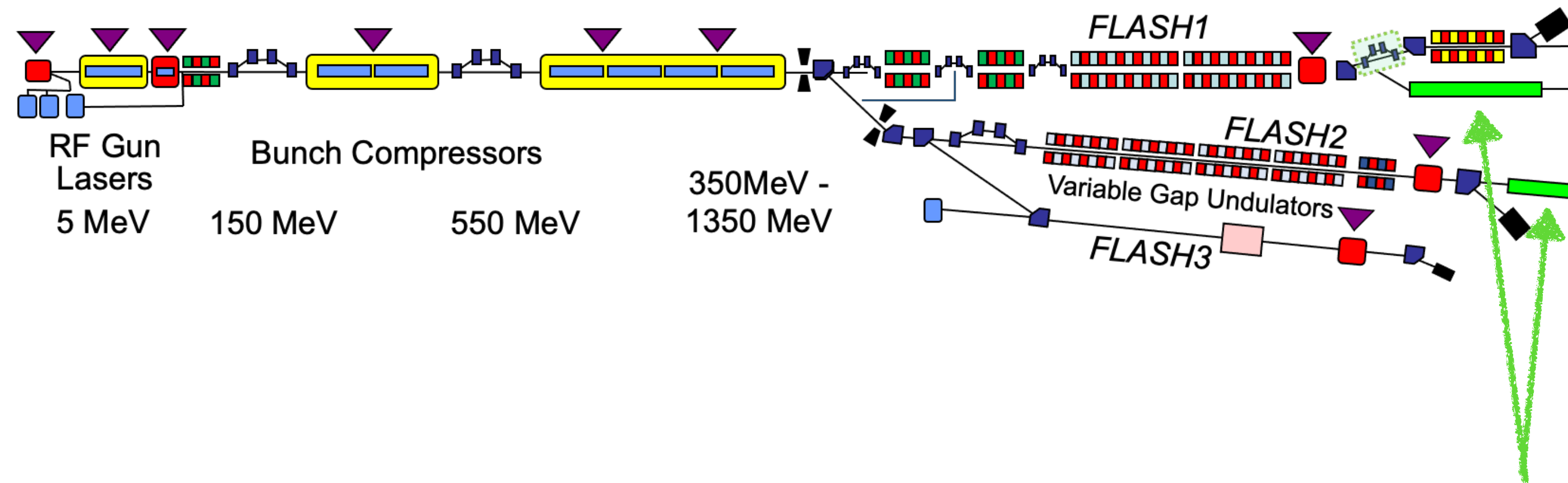
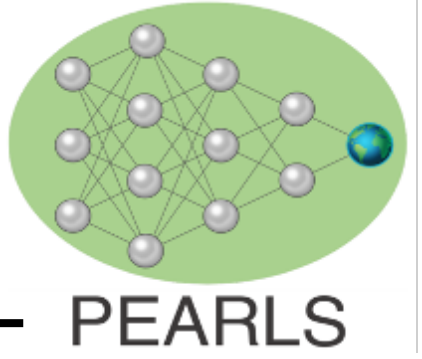
Accelerator

- ▶ Machine operation
- ▶ Beam dynamics
- ▶ Controls
- ▶ Setpoints and readbacks
- ▶ Time series and waveforms
- ▶ Bunch/train parameters
- ▶ Feedback and optimization

Magnet / RF	: $x(t)$	setpoint / readback
BPM	: (x_b, y_b)	beam position
Toroid / BCM	: Q_b	bunch charge
BAM	: t_b	arrival time
LLRF	: $V(t), \phi(t)$	waveforms
BLM	: $I(t)$	waveform
Screen	: $I(x, y)$	image



Photon diagnostics sit between the accelerator and the experiments



For one train containing up to 800 FEL pulses

GMD : $D[b, q]$ pulse-resolved parameters

VLS : $I[b, \lambda]$ pulse-resolved spectra

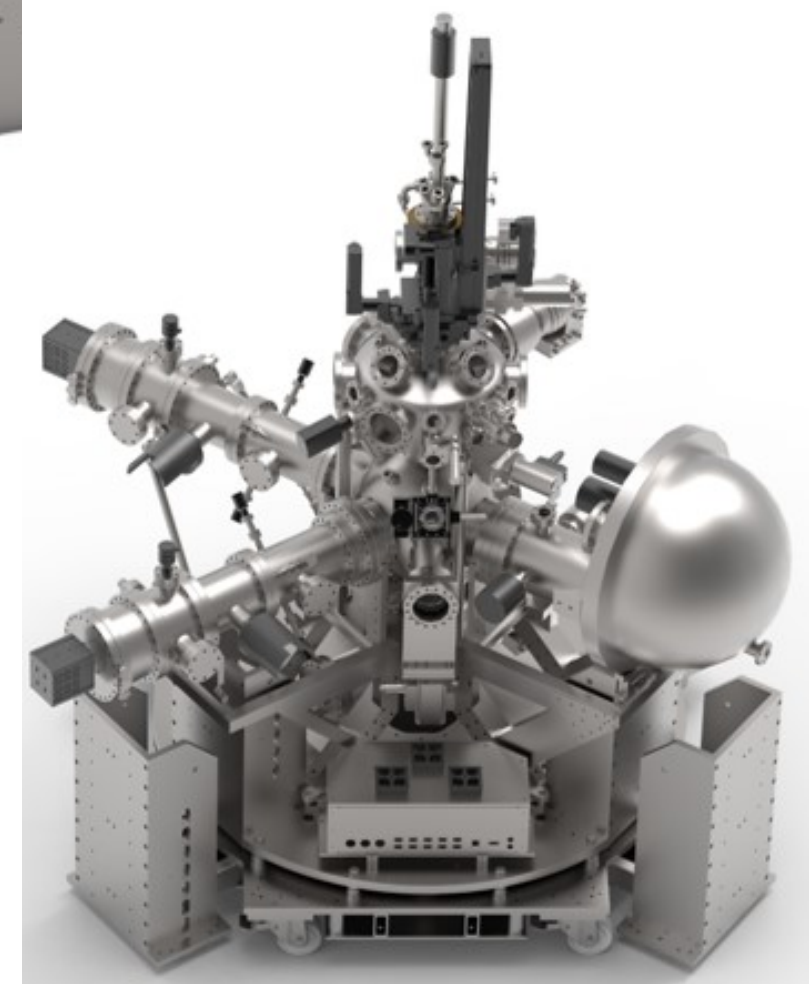
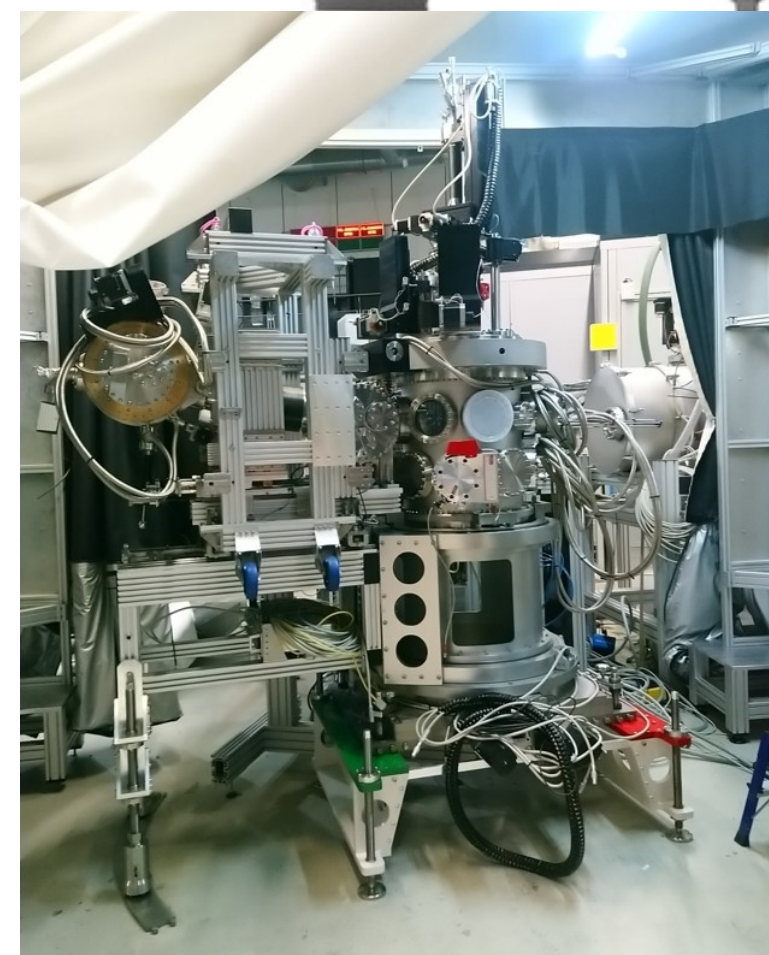
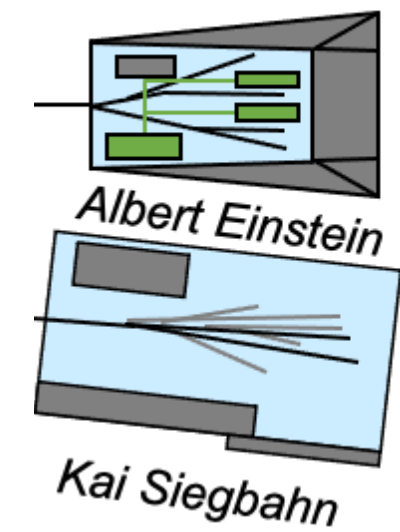
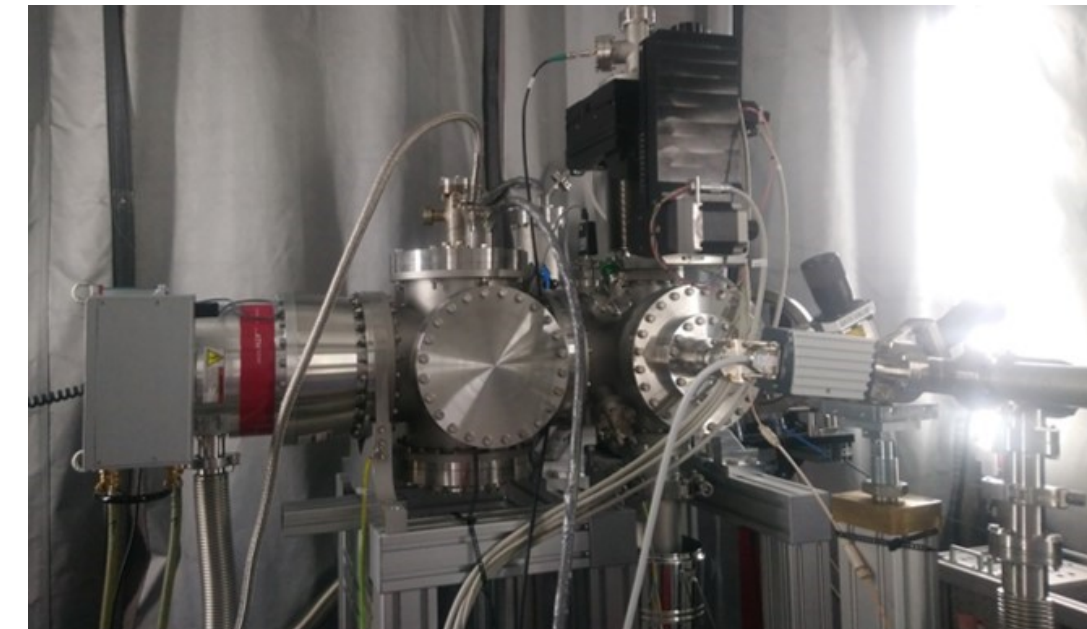
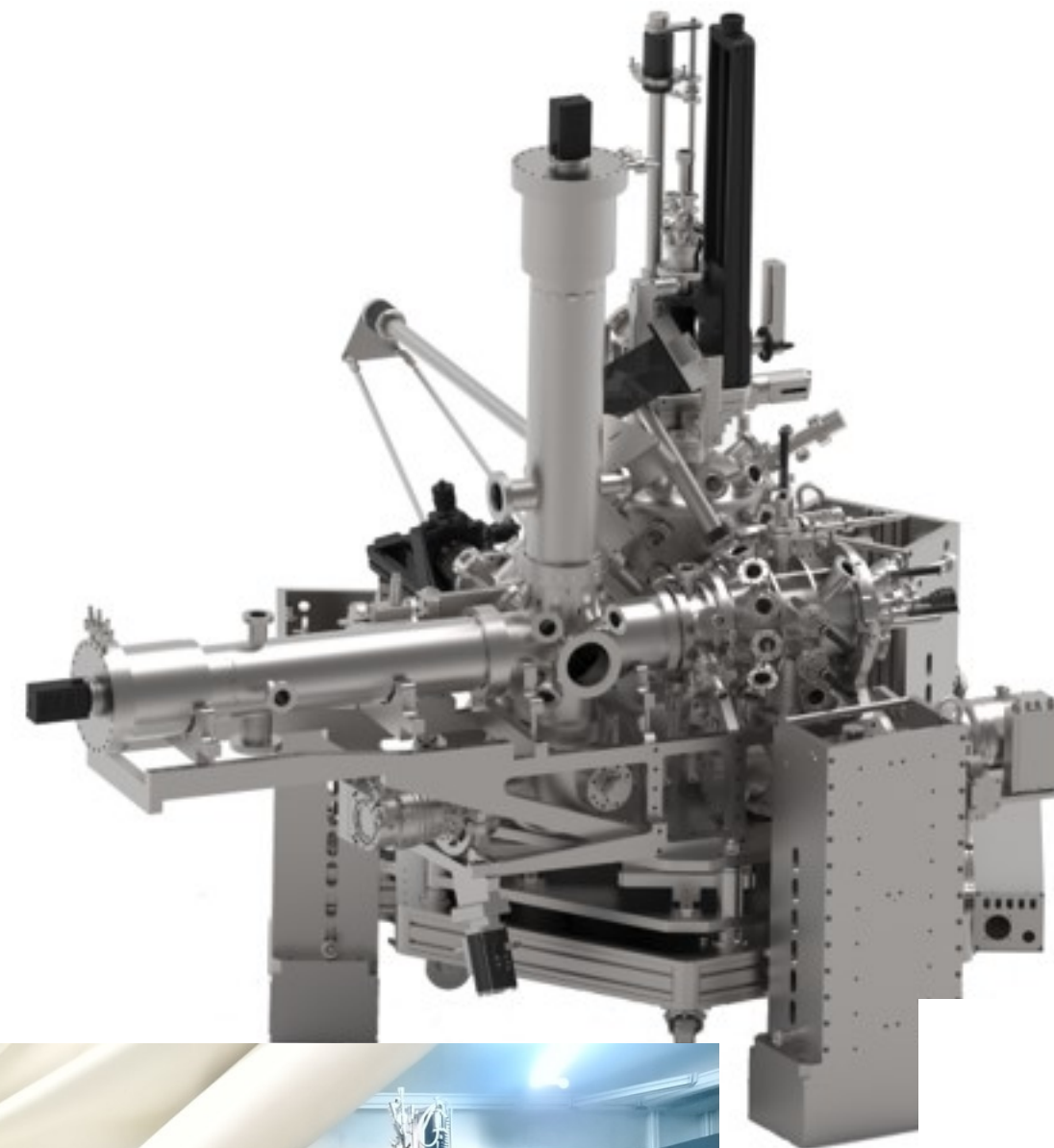
OPIS : $V[b, c, t]$ multi-channel time traces

Ce:YAG : $I(x, y)$ 2D image, possibly integrated

Mobile "semi-permanent" endstations

Photon Science

- ▶ Scientific experiments
- ▶ Photon-sample interaction
- ▶ Detector acquisition
- ▶ Images and spectra
- ▶ Event streams
- ▶ Scans and coordinates
- ▶ Reconstruction and analysis

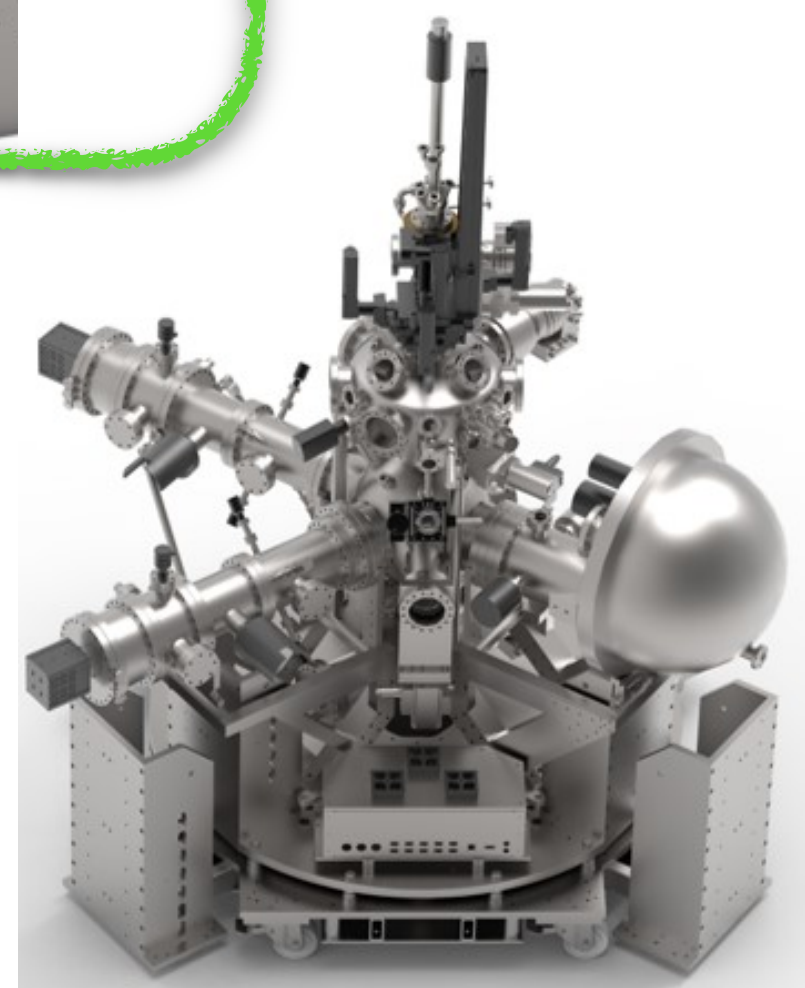
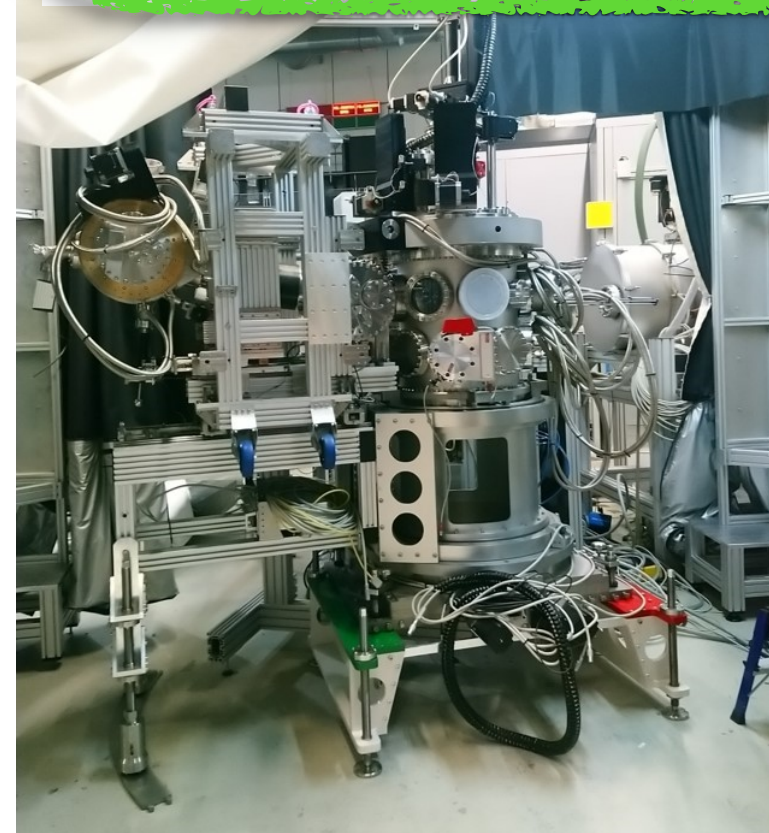
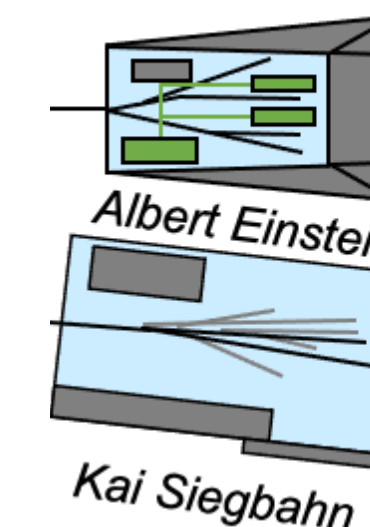
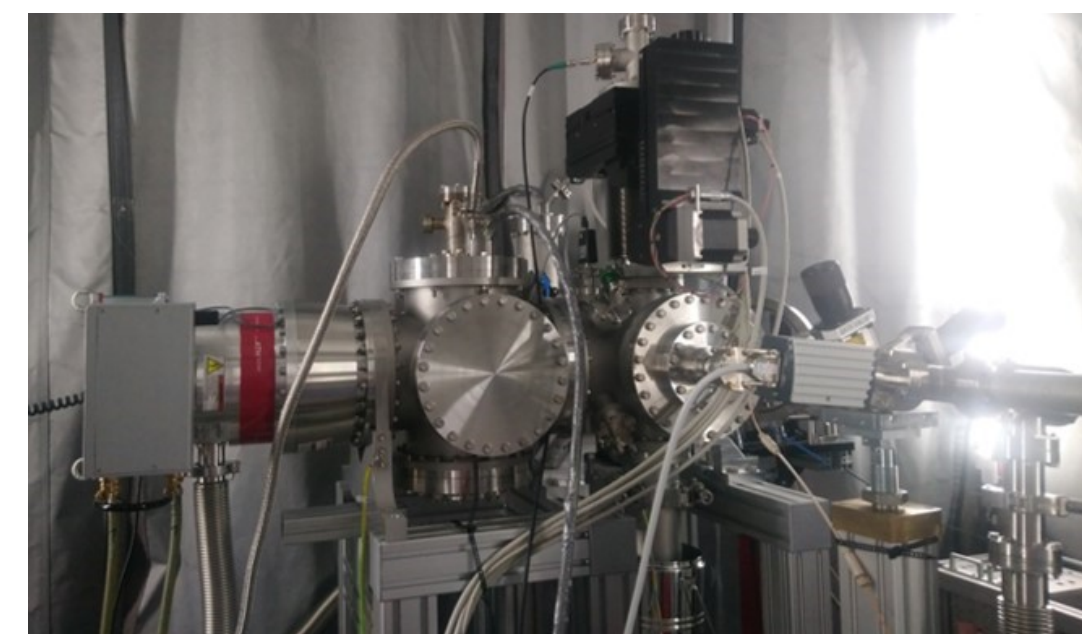


Mobile "semi-permanent" endstations

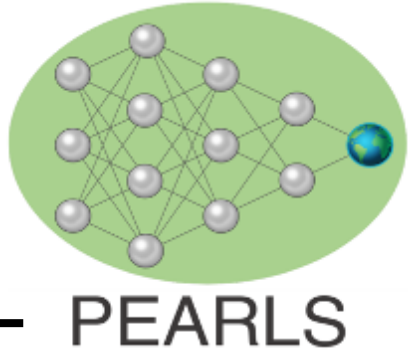
Photon Science

- ▶ Scientific experiments
- ▶ Photon-sample interaction
- ▶ Detector acquisition
- ▶ Images and spectra
- ▶ Event streams
- ▶ Scans and coordinates
- ▶ Reconstruction and analysis

HEXTOF



HEXTOF (High Energy X-ray Time-Of-Flight) end station



Space-, Time-, and Spin-resolved photoemission spectroscopy using a time-of-flight momentum microscope

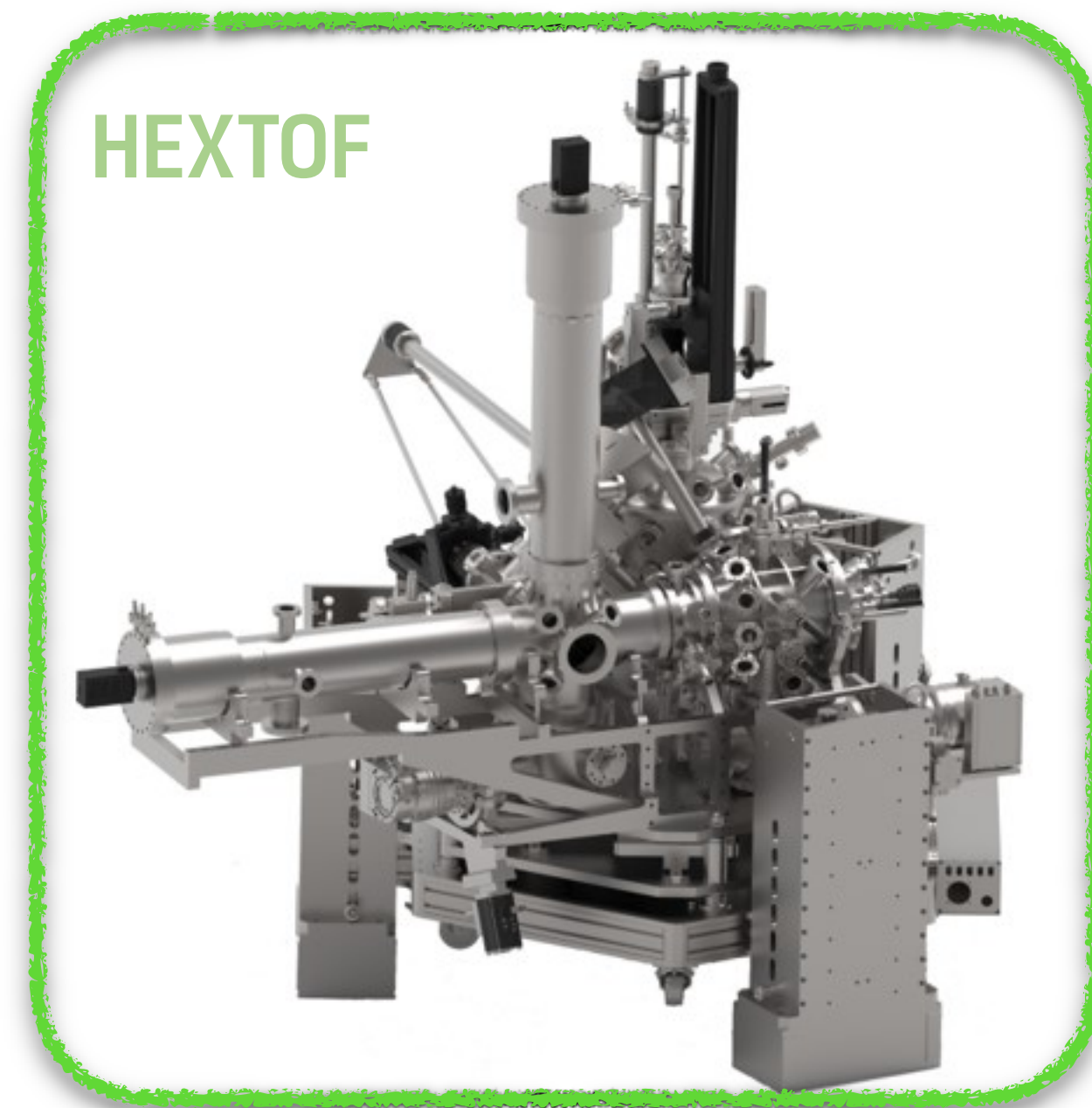
$$\text{event}_i = (x_i, y_i, t_{\text{TOF},i}, \dots)$$

After calibration:

$$(x, y, t_{\text{TOF}}) \longrightarrow (k_x, k_y, E)$$

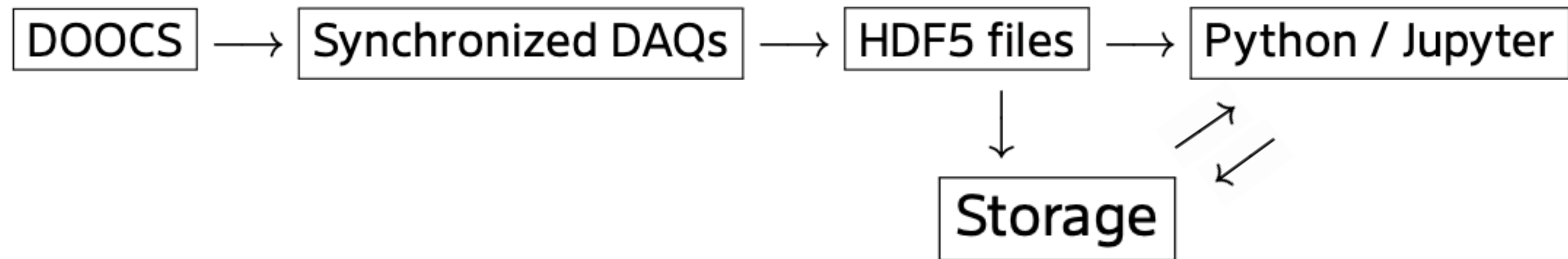
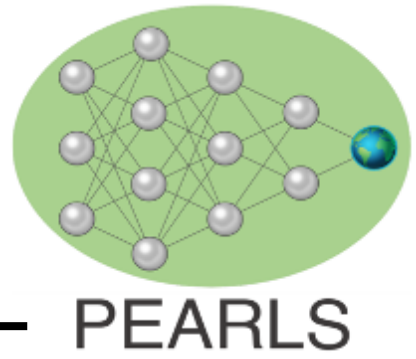
The same event stream can produce:

$I(E)$	spectrum
$I(k_x, k_y)$	momentum image
$I(k, E)$	band-dispersion image
$I(k_x, k_y, E)$	3D volume
$I(k_x, k_y, E, \Delta t)$	4D pump-probe data

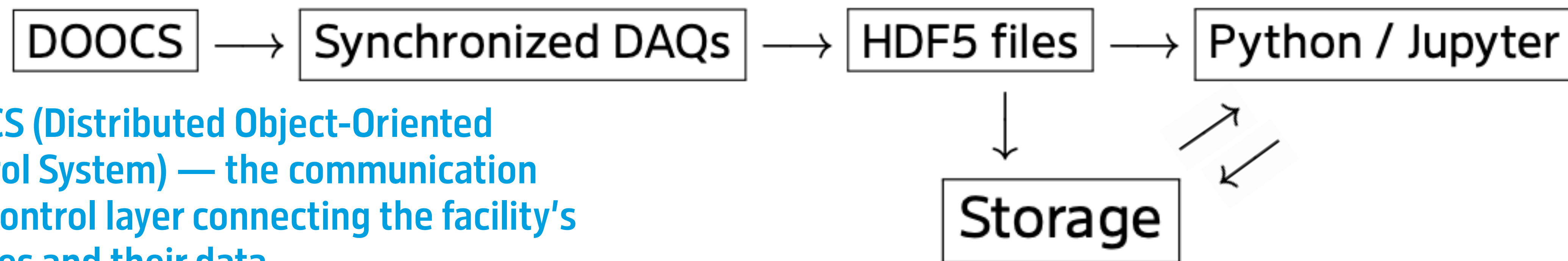
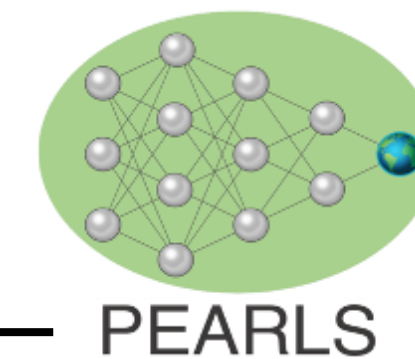


Sparse event stream $\rightarrow$ Multidimensional histogram (Similar to HEP detector Data structure?)

What does DESY Computing and Controls already provide?



What does DESY Computing and Controls already provide?

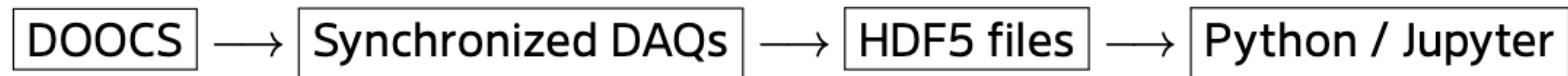
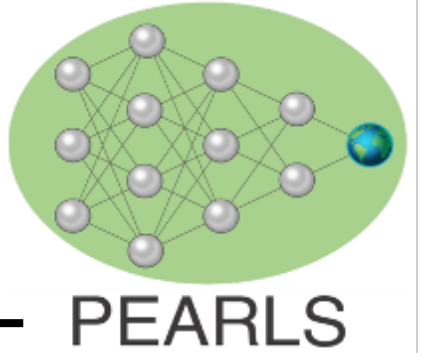


DOOCS (Distributed Object-Oriented Control System) — the communication and control layer connecting the facility's devices and their data.



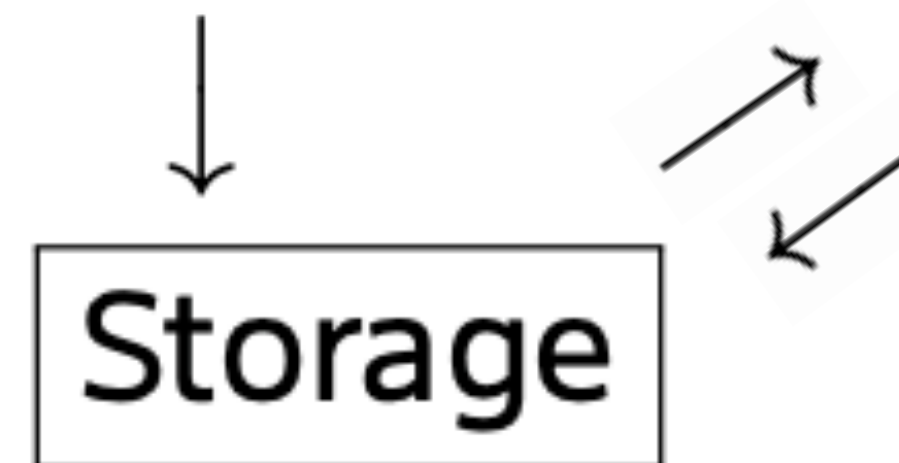
JDDD (Java DOOCS Data Display) — control-room dashboard for DOOCS operators use to tune the machine

What does DESY Computing and Controls already provide?



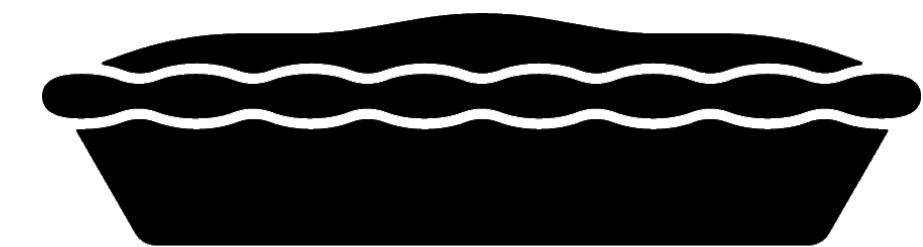
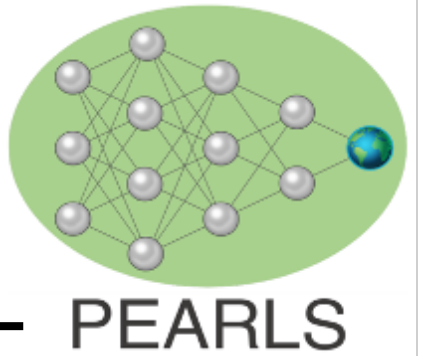
Different DAQs!!!!

Machine (linac) DAQ saves data continuously, and then this data is deleted after 2 weeks unless it is requested to be saved in dCache

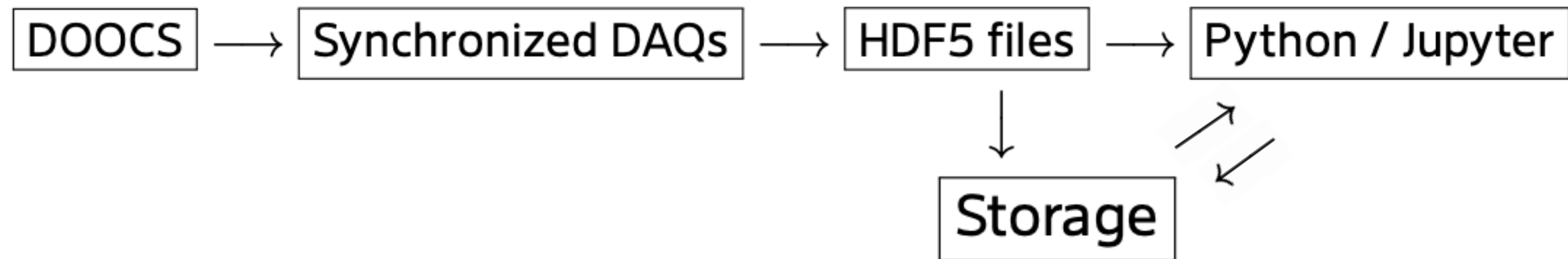


Experiment DAQ (varies for different photon science experiments)

What does DESY Computing and Controls already provide?



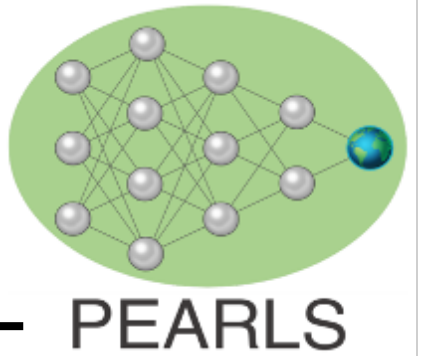
doocspie



Machine-readable $\neq$ Machine-learnable

Stored Data is not automatically AI-Ready Data

Before Machine Learning, we must solve a data-engineering problem

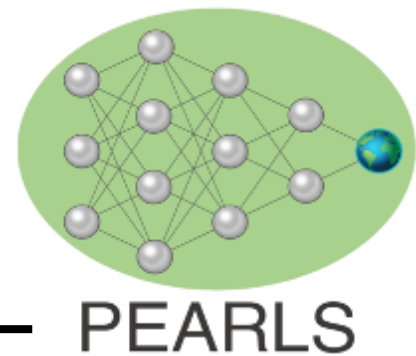


An HDF5 file can store

$X_{\text{accelerator}}[n, b]$	accelerator parameters
$E_{\text{GMD}}[n, b]$	FEL pulse energy
$T(t)$	slow conditions
$\{\text{HEXTOF events}\}_{n,b}$	sparse events

without sufficiently describing how these data relate to each other physically.

Before Machine Learning, we must solve a data-engineering problem



An HDF5 file can store

$X_{\text{accelerator}}[n, b]$ accelerator parameters

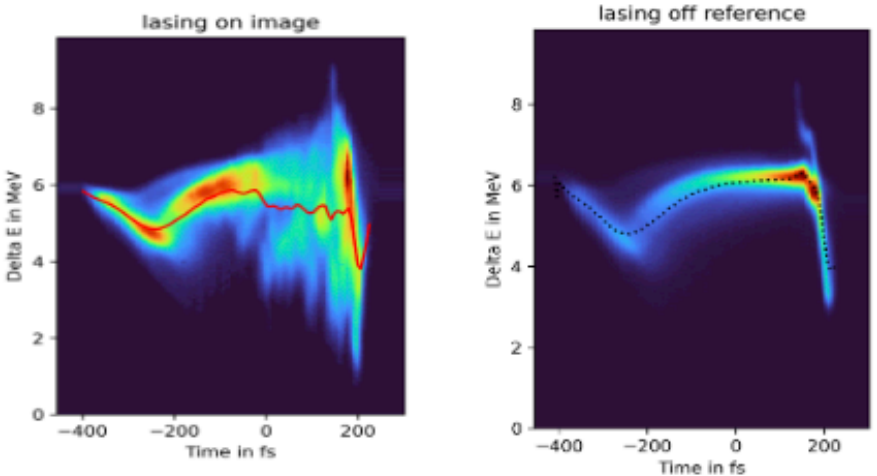
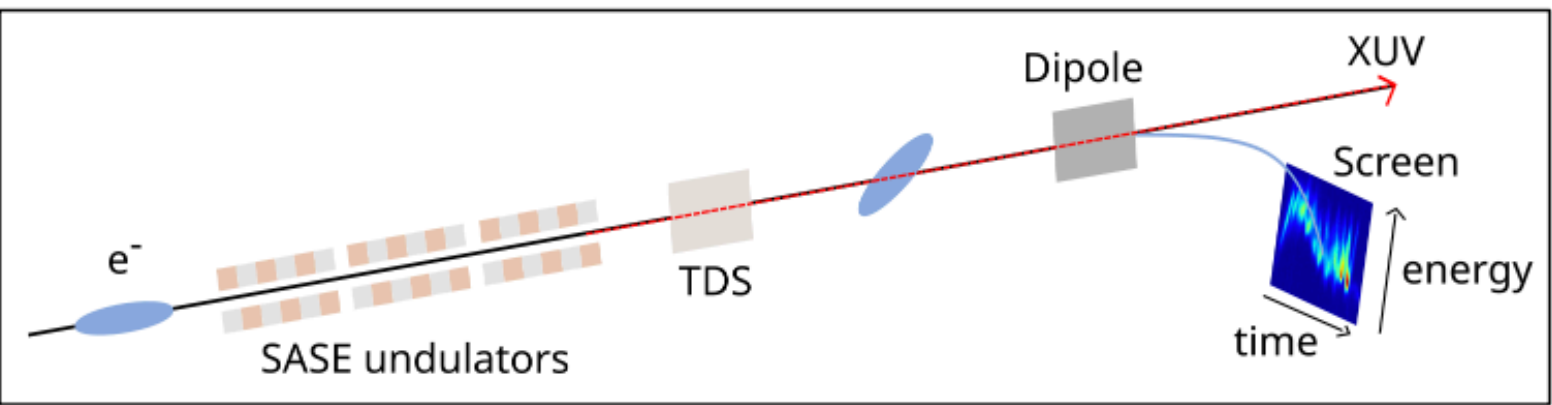
$E_{\text{GMD}}[n, b]$ FEL pulse energy

$T(t)$ slow conditions

$\{\text{HEXTOF events}\}_{n, b}$ sparse events

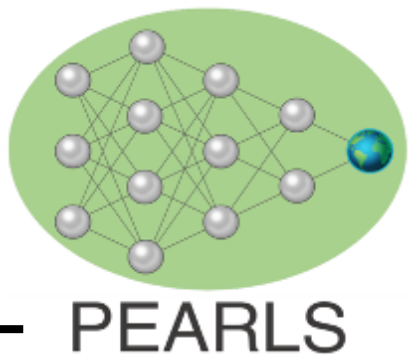
Use a subset of parameters from the machine that have a correlation with the FEL pulse diagnostic to add:

- Event association,
- validity,
- Quality,
- Uncertainty,
- Calibration



Gesa Goetzke's Talk after coffee break

Goal of Encoding: preserve physical meaning while creating a representation that ML models can benefit from

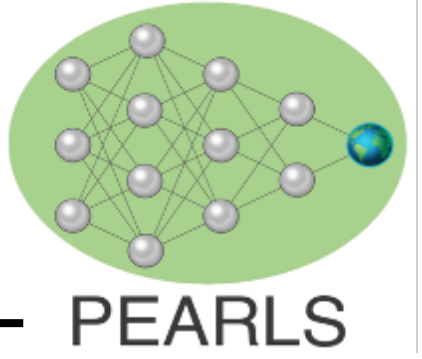


Encoding	Facility example	Representation
One-hot	Machine / operation mode	FLASH1 Stability $\rightarrow [1, 0, 0]$
Ordinal	Losses (Discrete levels)	low, medium, high $\rightarrow 1, 2, 3$
Cyclic	RF phase/ phase shifters	$\phi \rightarrow [\sin \phi, \cos \phi]$
Physics-based	HEXTOF detector events	$(x, y, t_{\text{TOF}}) \rightarrow (k_x, k_y, E)$
Summary features	Waveform / spectrum	$V(t) \rightarrow$ peak, integral, width, arrival time

Master's level methods

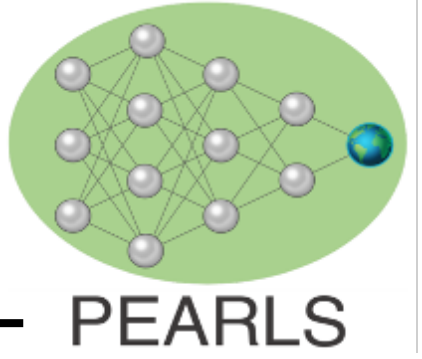
Postdoc level methods

How about before AI-assisted machine setup?



These train of 3-7 HEXTOF electrons came from macropulses 427-530 for which the GMD measured $41 \mu\text{J}$, with aperture set to 3 mm. Go to machine state during macropulses 427.

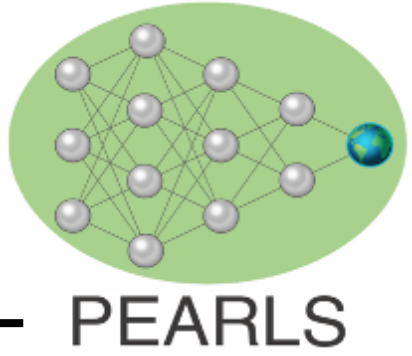
How about before AI-assisted machine setup?



These train of 3-7 HEXTOF electrons came from macropulses 427-530 for which the GMD measured $41 \mu\text{J}$, with aperture set to 3 mm. Go to machine state during macropulses 427.

I will discuss the challenges of this tomorrow.

UPDATE from PEARLS DESY - Hamburg



- Established Collaboration with the photon science group at DESY (LEAPS)
 - 1 postdoc for 3 years $\implies$; 2 postdocs for 2 years
 - Hiring process takes time, and its nice to have the flexibility
- Paper at the IPAC 2026 conference
 - Discusses motivation and goals



SUSTAINABLE DATA TRANSFORMATION STRATEGIES FOR PARTICLE AND FIELD MODELING FREE-ELECTRON LASERS

P. Niknejadi*, S. S. Gurdasani, J. L. Prochnow, L. Schaper
Deutsches Elektronen-Synchrotron DESY, Hamburg and Zeuthen, Germany
N. K. Bhalla, V. S. Lang
Albert-Ludwigs-Universität Freiburg, Freiburg, Germany

Abstract

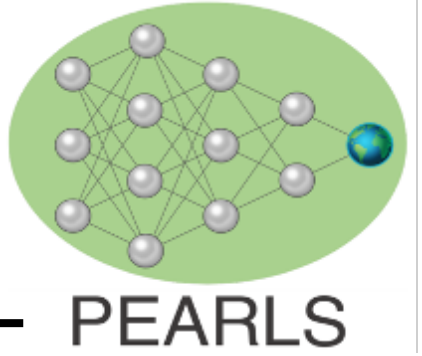
Scientific data is commonly stored in formats such as CSV for small to medium datasets, HDF5 for complex hierarchical structures, and NetCDF for gridded and temporal variables. These formats provide robust mechanisms for the structured storage of hierarchical or multidimensional data, but they do not, by themselves, produce the task-specific tensors, masks, and metadata needed for efficient machine learning (ML) or GPU-accelerated computation. In practice, workflows built around these formats often incur substantial transformation overhead. Recent studies have demonstrated the promise of tailored data-transformation and encoding approaches in scientific domains. In particular, convolutional autoencoder-based compression of 3D particle-tracking data can achieve significant reductions in post-processing as well as storage requirements while keeping the essential spatial features for downstream workflows. The key bottleneck is not only data volume, but the mismatch between code-specific simulation output and the physically meaningful tensor structures required for ML. Motivated by the need to reduce the carbon footprint of increasingly large-scale simulations, this work proposes a domain-specific encoding perspective for interfacing particle and field data across different simulation tools, aiming to enhance the scalability and integration of scientific data into more sustainable workflows and future machine learning pipelines.

ticularly significant role in the carbon footprint of digitally intensive research, as simulations, processing, and storage all contribute to the total. This issue is also connected to broader community efforts in resource-aware computation, such as minimizing resource consumption for experiments in Universe and Matter research and improving the availability and usability of scientific data [2, 3]. While a great deal of research-sustainability effort has been spent on facility-level initiatives such as infrastructure and lab efficiency, procurement, and travel policy, these efforts don't necessarily inform the impact of individual researchers' workflows. This perspective led to the Precision in Energy-aware AI Research for Low-carbon Solutions (PEARLS) project, which ties carbon accounting together with sustainable scientific machine learning and resource-aware computation. PEARLS considers research emissions to be part of a workflow of interconnected activities. Data encoding is one part of PEARLS' sustainability strategy, which is part of our focus on data transformation discussed here.

DEFINING THE ENCODING GAP

In practice, accelerator-simulation data often undergo repeated conversion from code-specific output to array- or tensor-format structures amenable to analysis, machine learning training, optimization, and GPU-based workflows. The plasma-based accelerator community, motivated in part by the transition from traditional linear accelerators toward

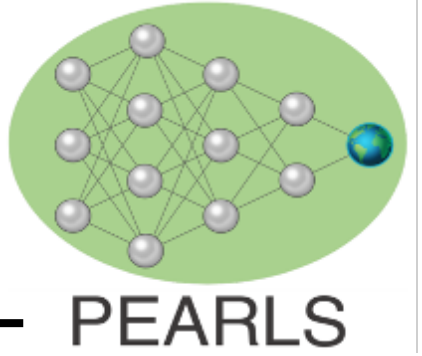
UPDATE from PEARLS DESY - Hamburg



- Computing Access:
 - JUWELS (juwels regular CPU nodes) + Data Storage
 - JUPITER (jupiter booster GPU nodes)
 - DESY - MAXWELL
- The hiring process took a long time.
 - 1 postdoc hired
 - 1 is currently being advertised: https://v22.desy.de/e67416/records313935/index_eng.html
 - Supported DESY-Zeuthen hiring as well

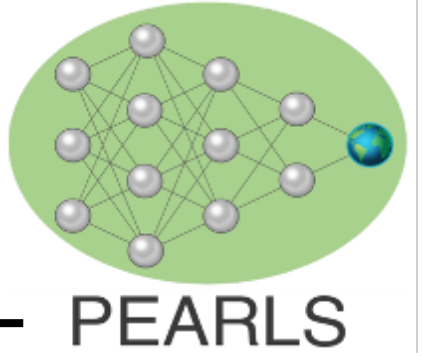


Summary and Outlook



- Goal of Encoding: preserve physical meaning while creating a representation that ML models can benefit from
 - We are not starting from zero. We are building on existing infrastructure.
- Through collaboration, it's possible to do fairly advanced physics-based encoding
 - Gesa has a lot of experience with ML and NN; PEARLS can benefit from it
- If PEARLS timeline allows, it will be very exciting to implement encoding+pruning+uncertainty calculation + CO2 footprint
- Actively looking for future funding and collaborations (Thursday discussions)

Acknowledgements



Many Thanks to

- DESY - PEARLS Zeuthen
- FLASH Team, especially L. Schaper, E. Ferrari, M. Vogt, J. Rönsch-Schulenburg
- Supporting M groups
- Supporting FS groups
- ...
- And in memory of S. Schreiber

Q&A and Discussion

The impact of Artificial Intelligence

4.32 g

CO₂ per ChatGPT message

While a typical Google search emits ~ 0.2 grams of CO₂

Source: [Smartly AI](#)

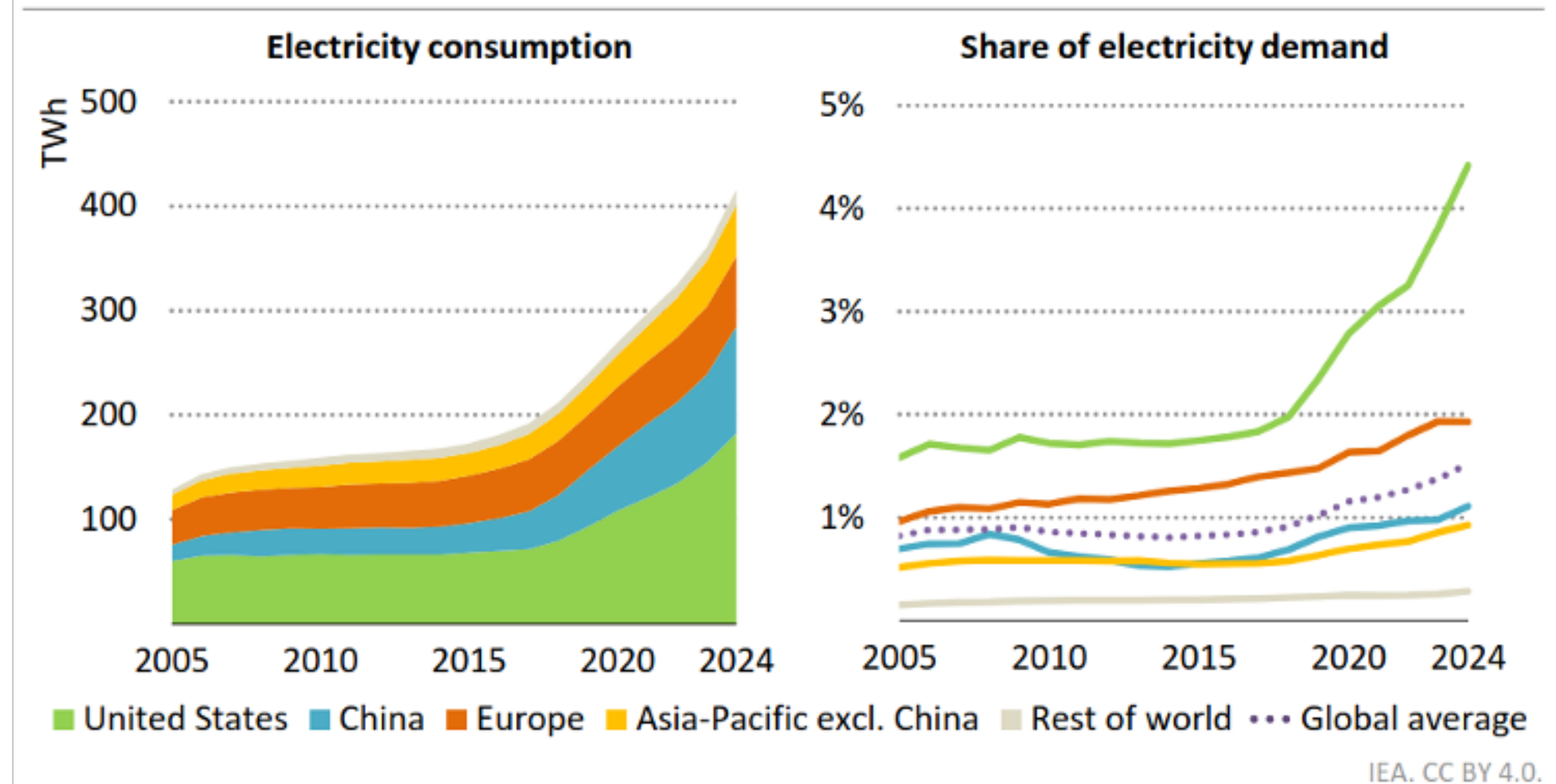
12%

Annual growth in data centre electricity

Which is 4x faster than total electricity consumption

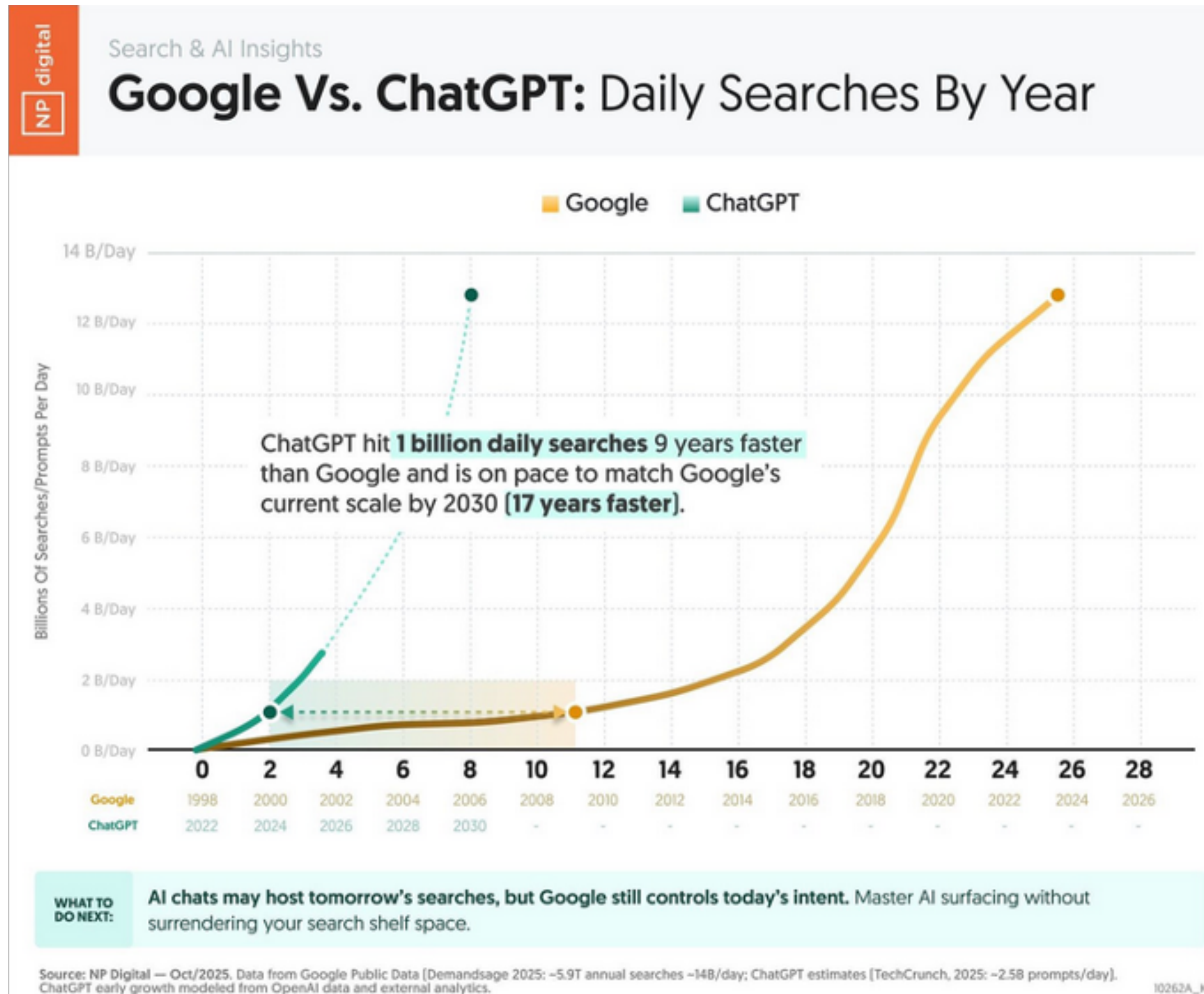
Source: [IEA's Energy & AI Report](#)

Figure 2.6 ▶ Electricity consumption of data centres by region, 2005-2024



The acceleration in data centre electricity consumption observed in 2017 was mainly driven by the United States and, to a lesser extent, by China

The impact of Artificial Intelligence



9 years

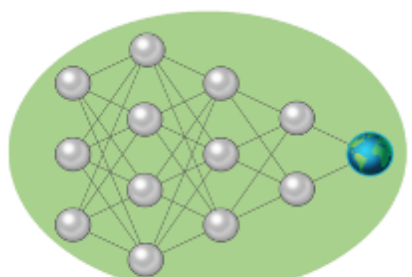
ChatGPT hit 1 billion daily searches 9 years faster than Google.

Adoption rate faster than any other technology so far.

Source: [NP Digital](#), [AI Adoption Rates](#)

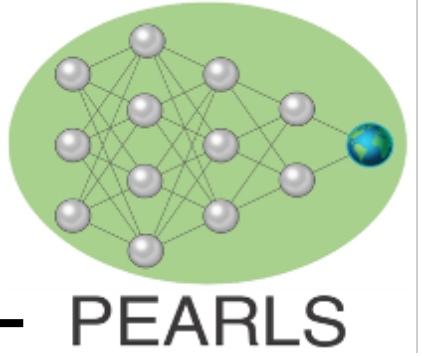
- Google still handles more searches: ~ 14 bn daily queries now.
- But AI tools, like ChatGPT, grew exponentially. From hundreds of millions to billions of daily queries.
- Growth even faster than Instagram, Tiktok

-> New information tools can replace old ones.



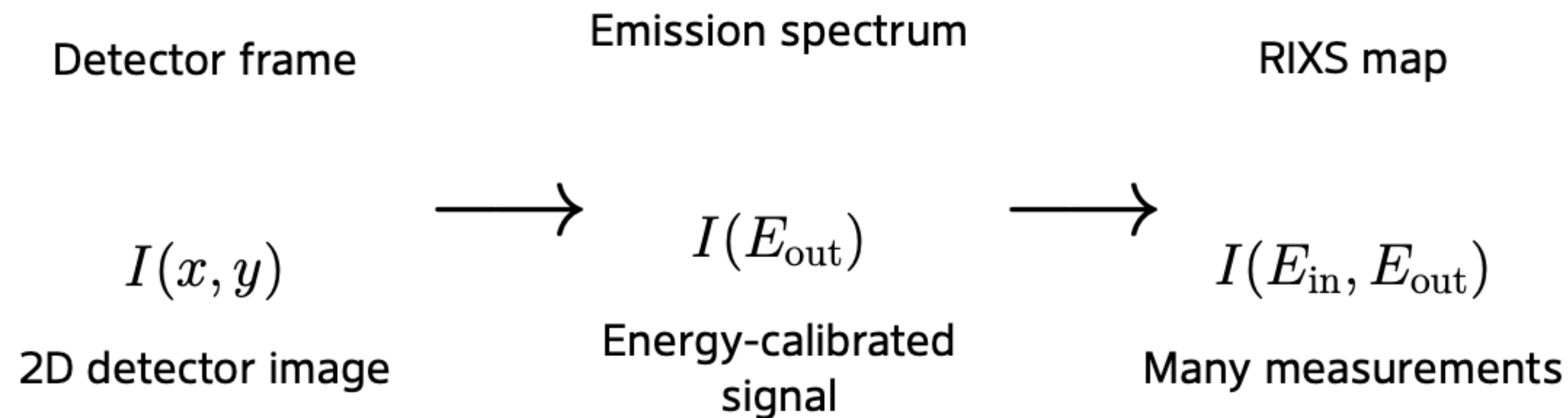
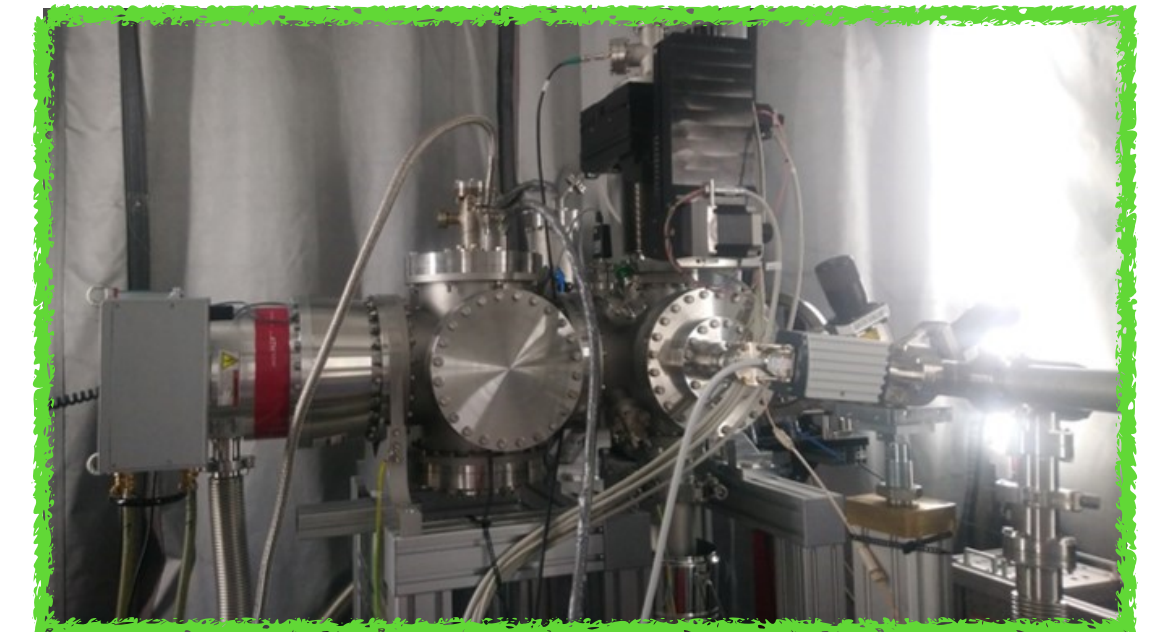
PEARLS

FLUXS (FLASH Liquid-phase Ultrafast X-ray Spectroscopy)



A flexible system for advanced solution-phase spectroscopy in the soft X-ray regime.
Used in different modes and machines

FLUXS



Different levels of representation

The same experiment produces data with different dimensionality and physical meaning as acquisition and analysis progress.

Detector data → physical coordinates → multidimensional scientific data