Towards a BES Light Source Wide Event-triggered Tomography Data Analysis Pipeline Using a Sustainable Software Stack

Hari Krishnan, Lawrence Berkeley National Laboratory

CAMERA – Center for Advanced Mathematics for Energy Research Applications
ALS – Advanced Light Source
Data Pilot - DOE BES Light Source Pilot Project

Credits for slides goes to
BES Data Working Group and its members

ALS, Berkeley, US

Max IV, Lund, Sweden

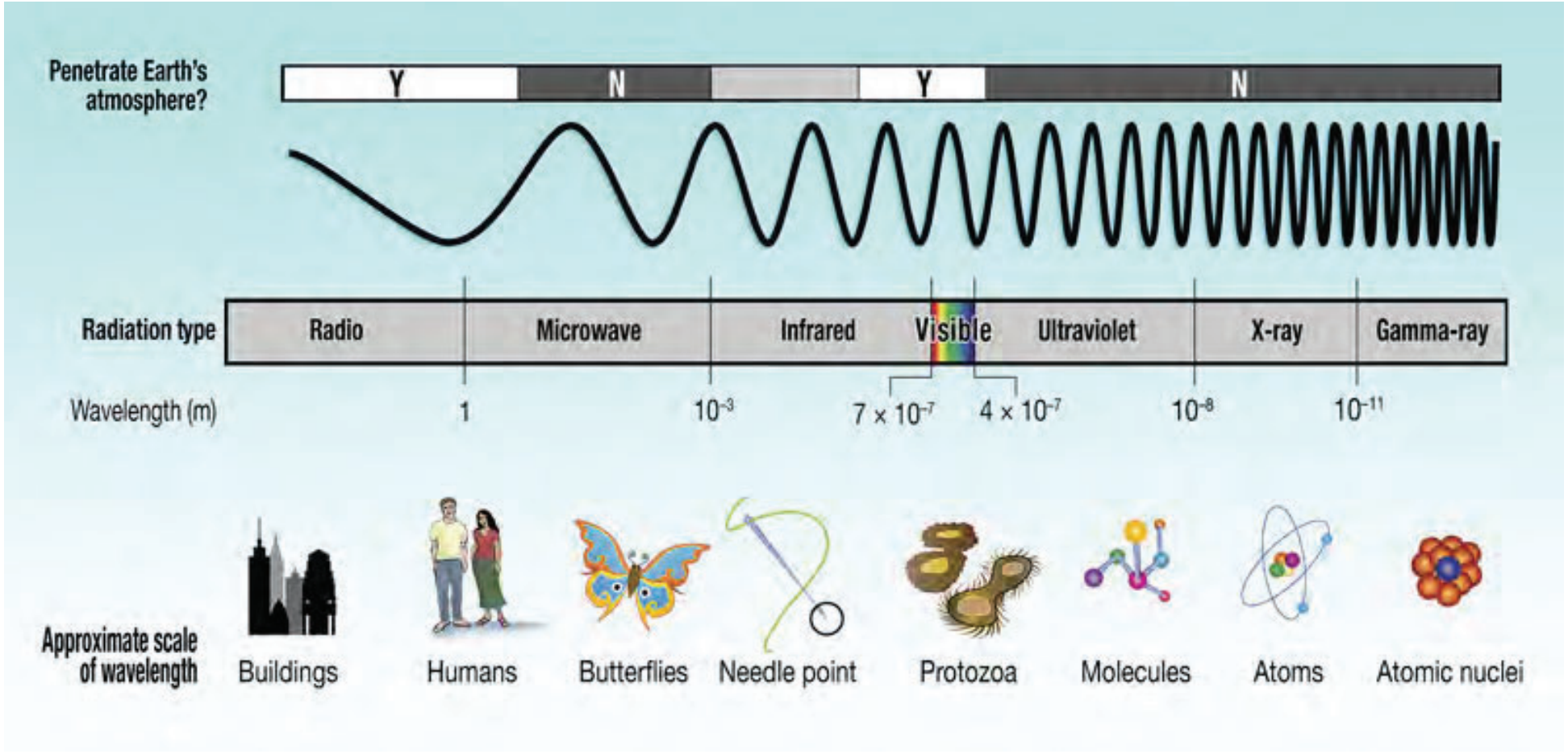
Sprint8/SACLA, Japan

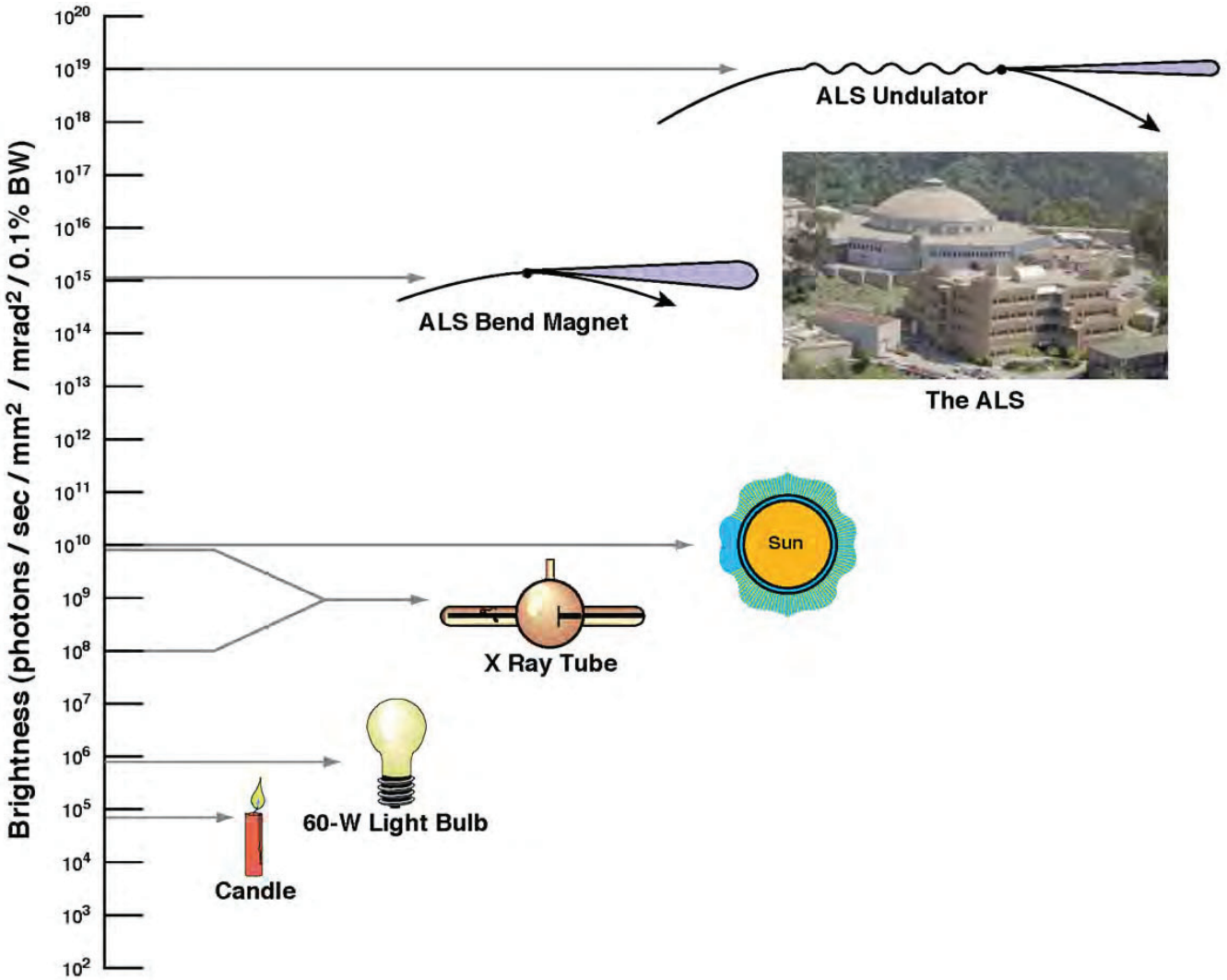
…and many more

SSRL/LCLS, Stanford, US

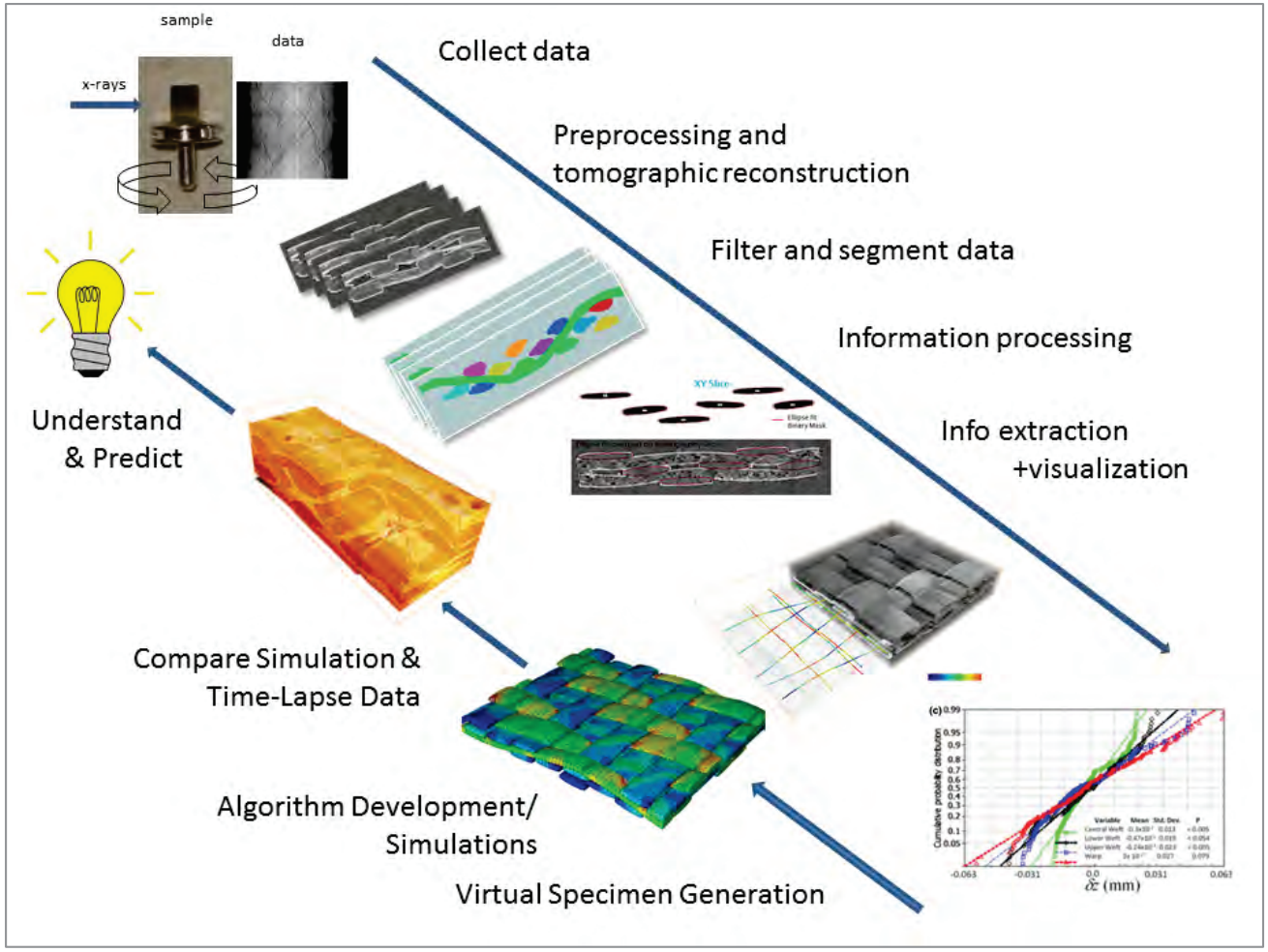
PSI, Switzerland

Elettra/FERMI, Trieste, Italy

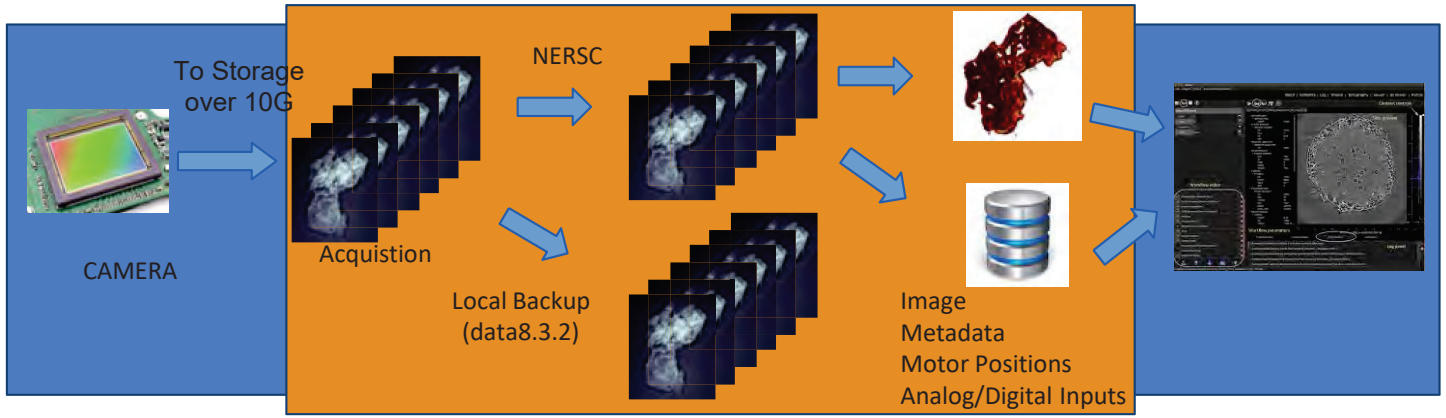

Penetrate Earth's atmosphere?
Y
N
Y
N
Radiation type
Radio
Microwave
Infrared
Visible
Ultraviolet
X-ray
Gamma-ray
Wavelength (m)
1
10^{-3}
7×10^{-7}
4×10^{-7}
10^{-8}
10^{-11}
Approximate scale of wavelength
Buildings
Humans
Butterflies
Needle point
Protozoa
Molecules
Atoms
Atomic nuclei


Brightness (photons / sec / mm² / mrad² / 0.1% BW)
10²⁰
10¹⁹
10¹⁸
10¹⁷
10¹⁶
10¹⁵
10¹⁴
10¹³
10¹²
10¹¹
10¹⁰
10⁹
10⁸
10⁷
10⁶
10⁵
10⁴
10³
10²
ALS Undulator
ALS Bend Magnet
The ALS
Sun
X Ray Tube
60-W Light Bulb
Candle

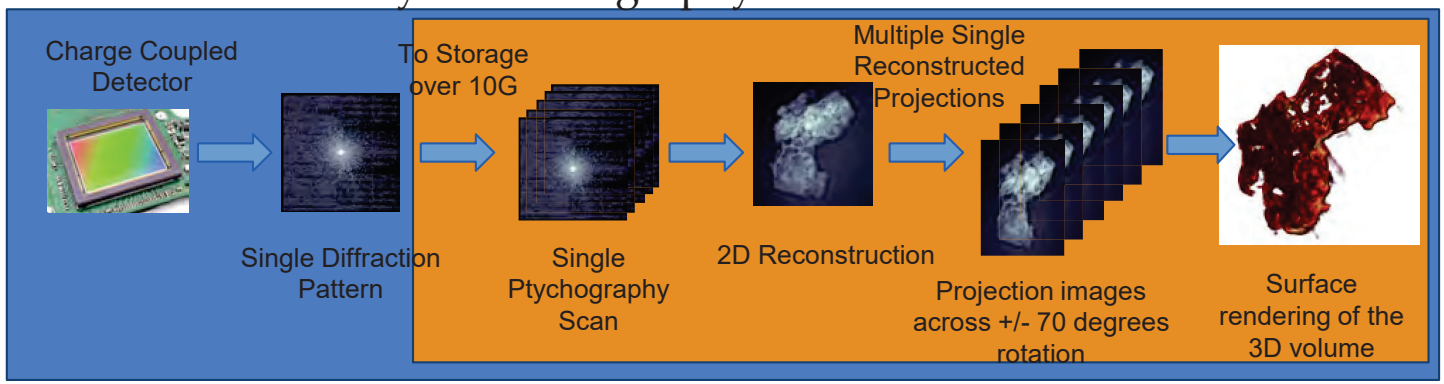
Introduction to Tomography


sample
data
x-rays
Collect data
Preprocessing and
tomographic reconstruction
Filter and segment data
Information processing
Info extraction
+visualization
Understand
& Predict
Compare Simulation &
Time-Lapse Data
Algorithm Development/
Simulations
Virtual Specimen Generation

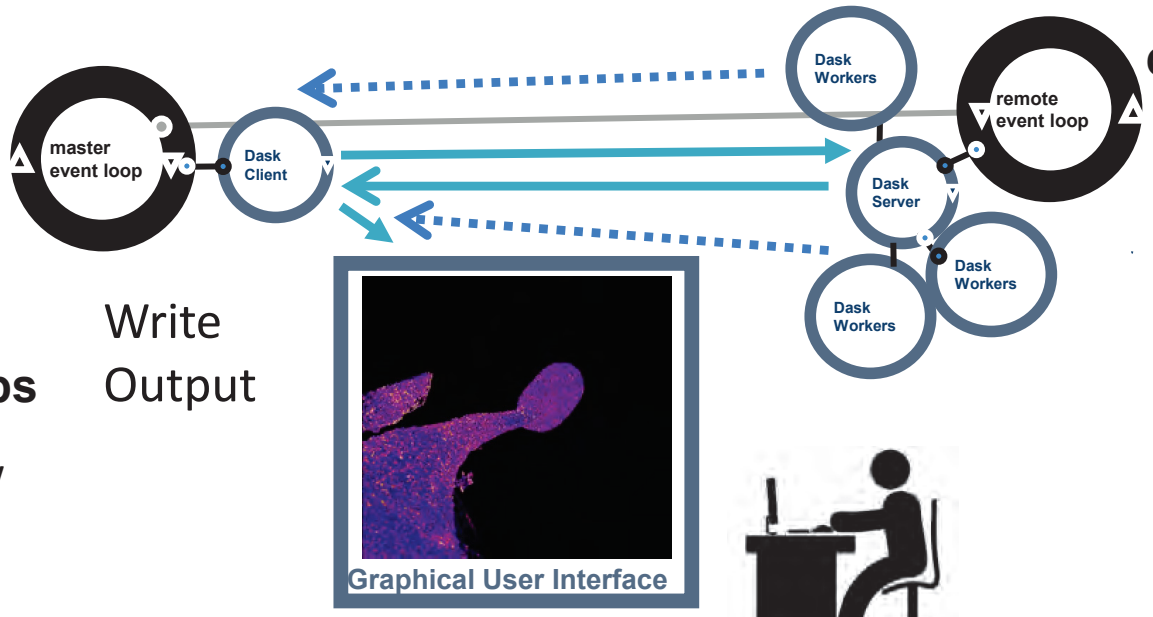
Tomography @ ALS - 8.3.2



Ptycho-Tomography @ COSMIC



Tomography Analysis



Compute cluster

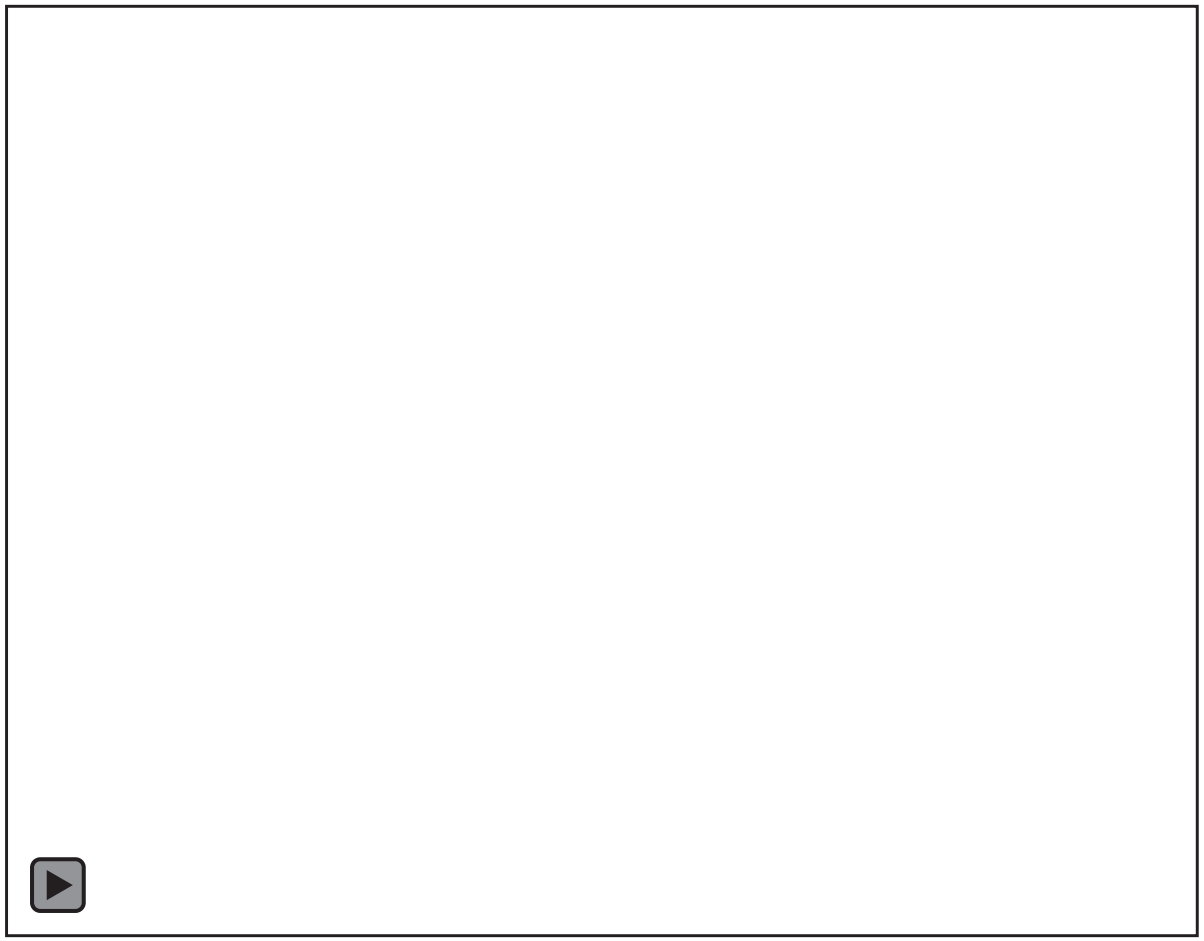
Tomography

1.Read
2.Normalize
3.Remote Outlier
4.Remove Stripe
5.Padding
6.Reconstruction
7.Crop
8.Circular Mask
9.Output

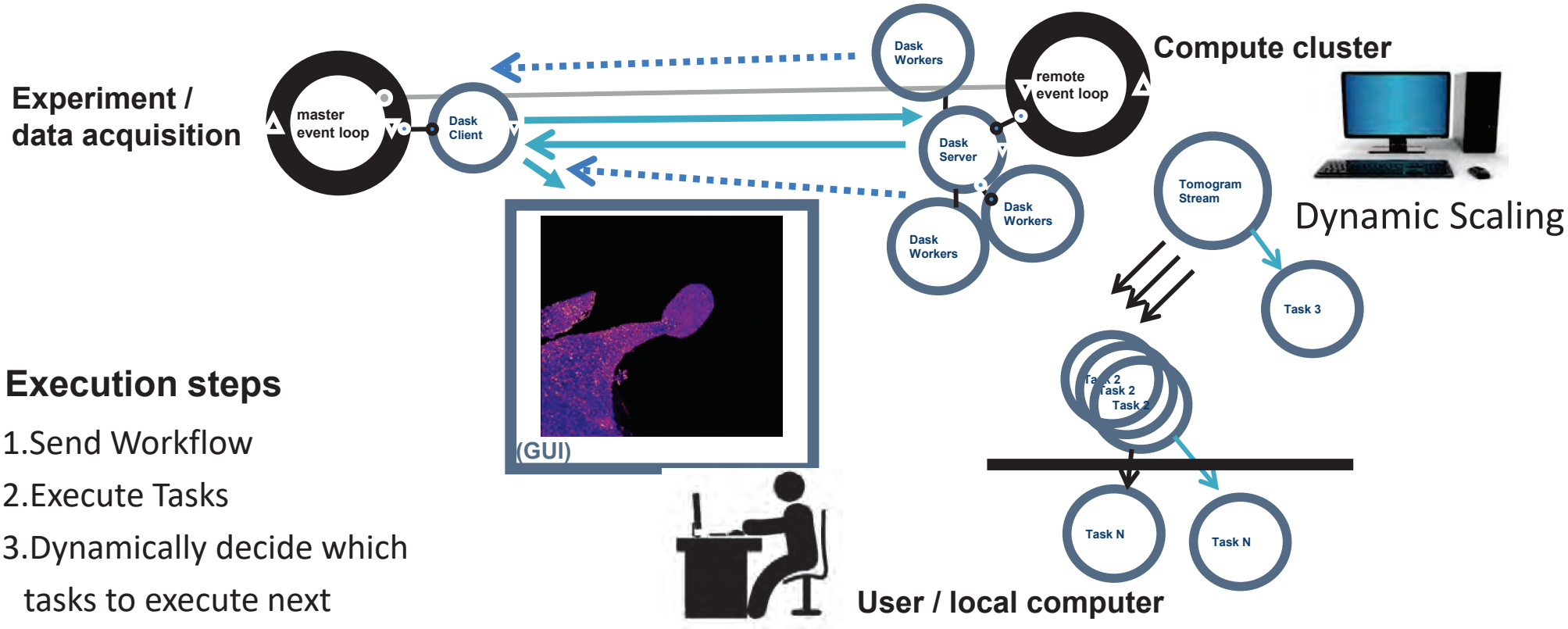
Execution steps

1.Send Workflow
2.Execute Graph
3.Return result
4.Visualization updates

Post Processing Tomography

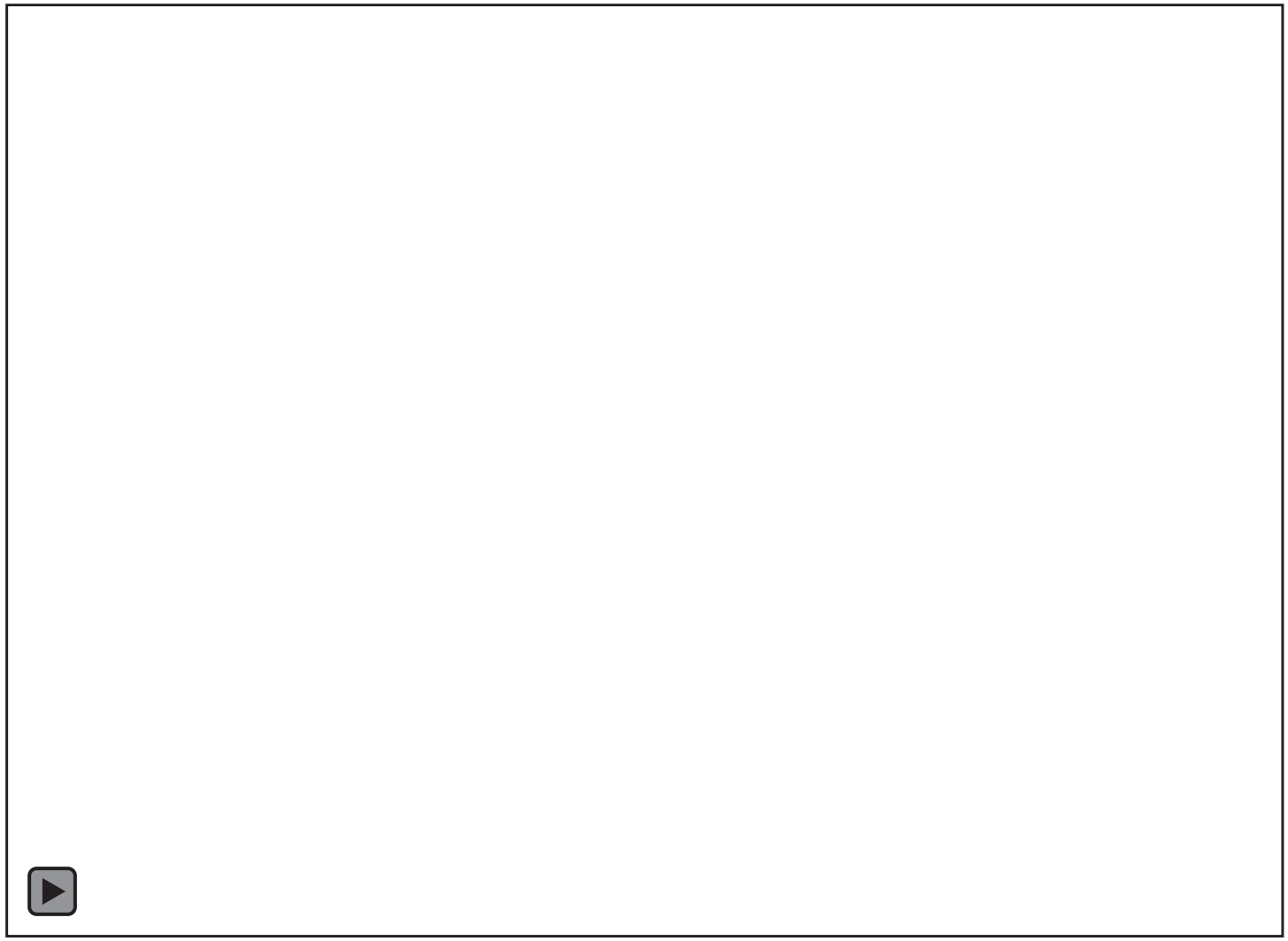


Streaming Analysis

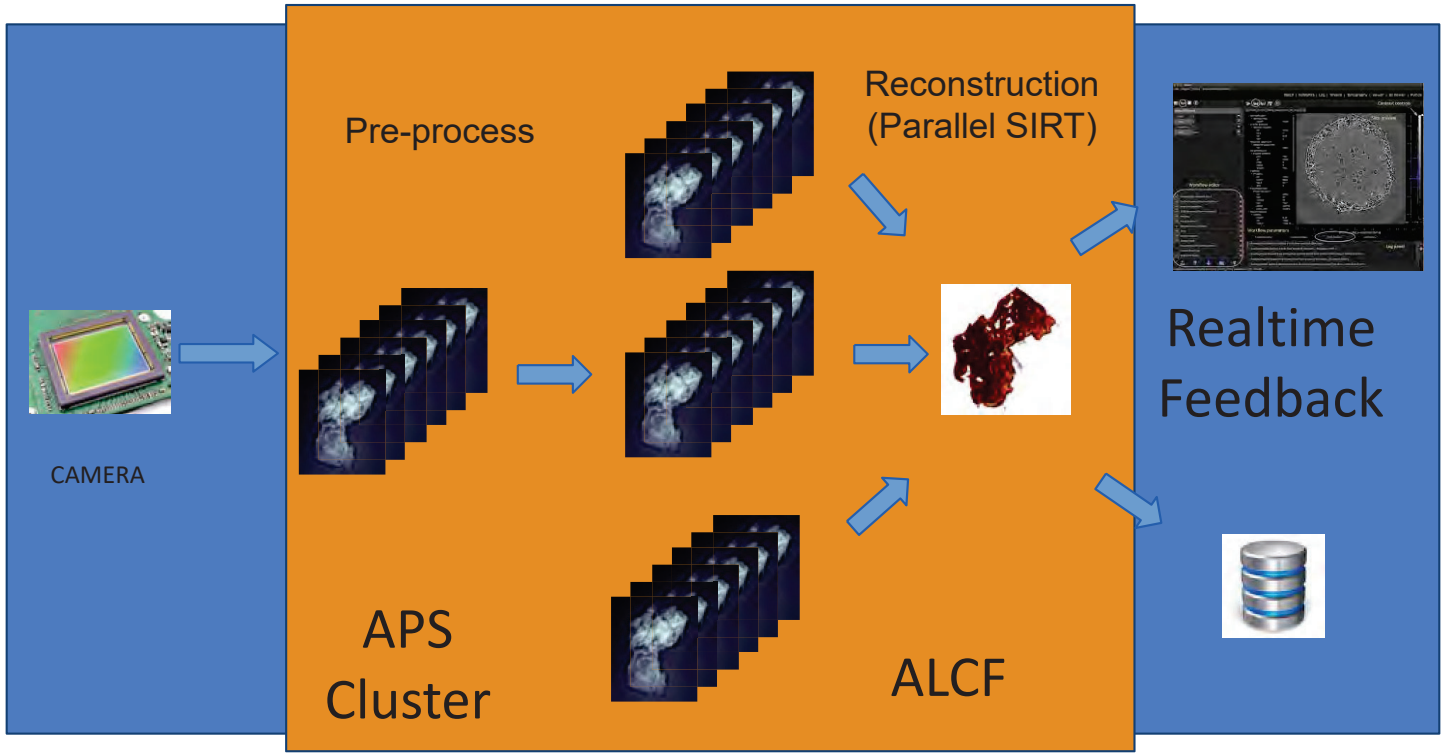


Execution steps

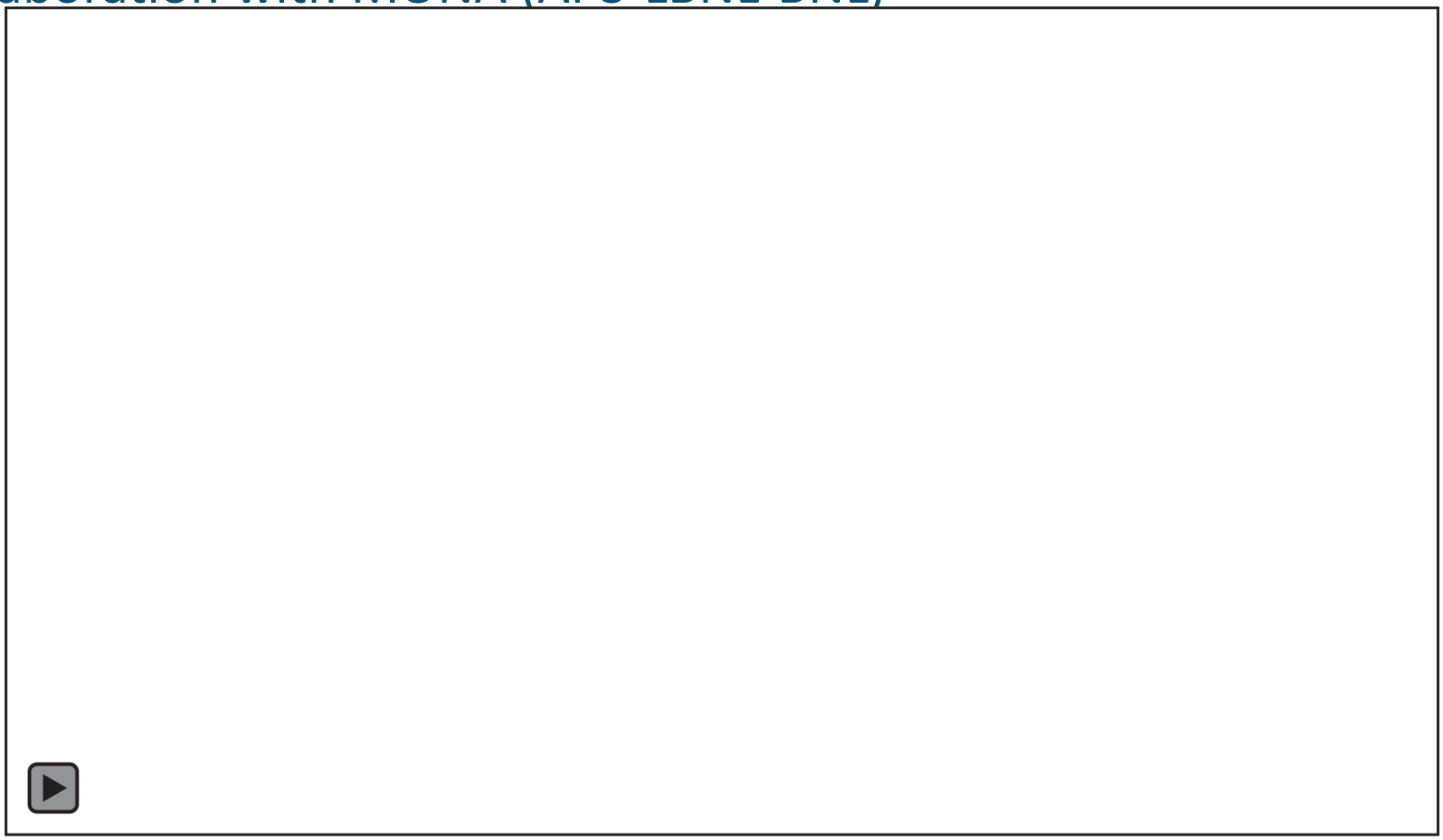
1. Send Workflow
2. Execute Tasks
3. Dynamically decide which tasks to execute next



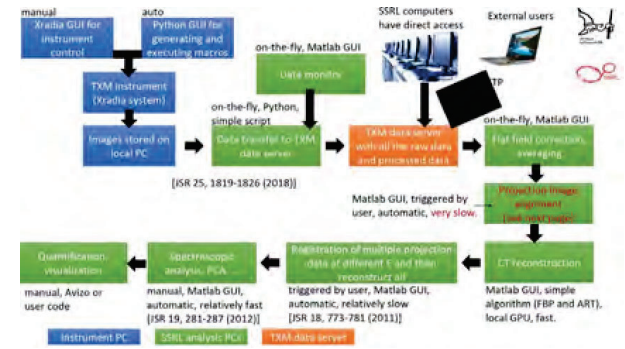
Tomography @ APS - 2BM

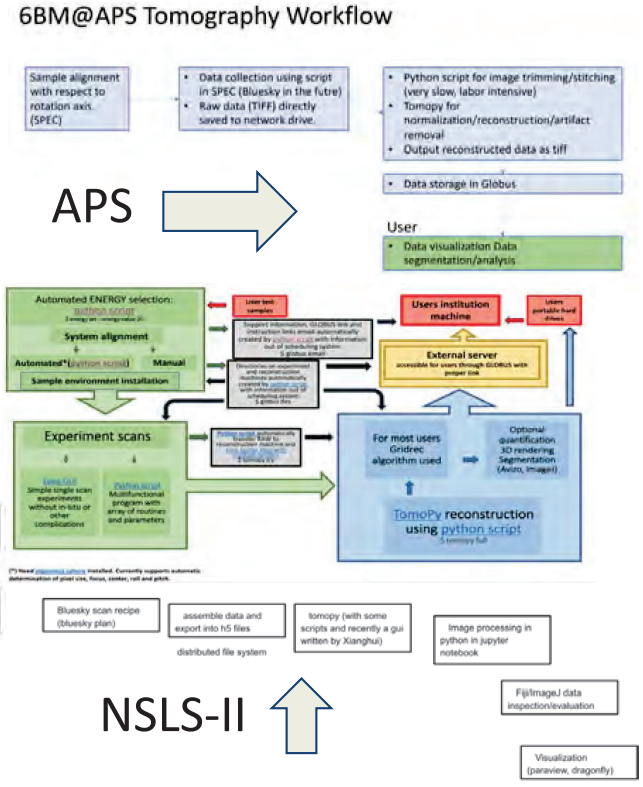


Streaming Tomography at APS
- Collaboration with MONA (APS-LBNL-BNL)

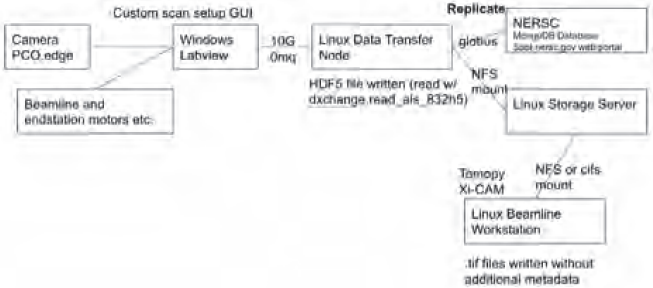


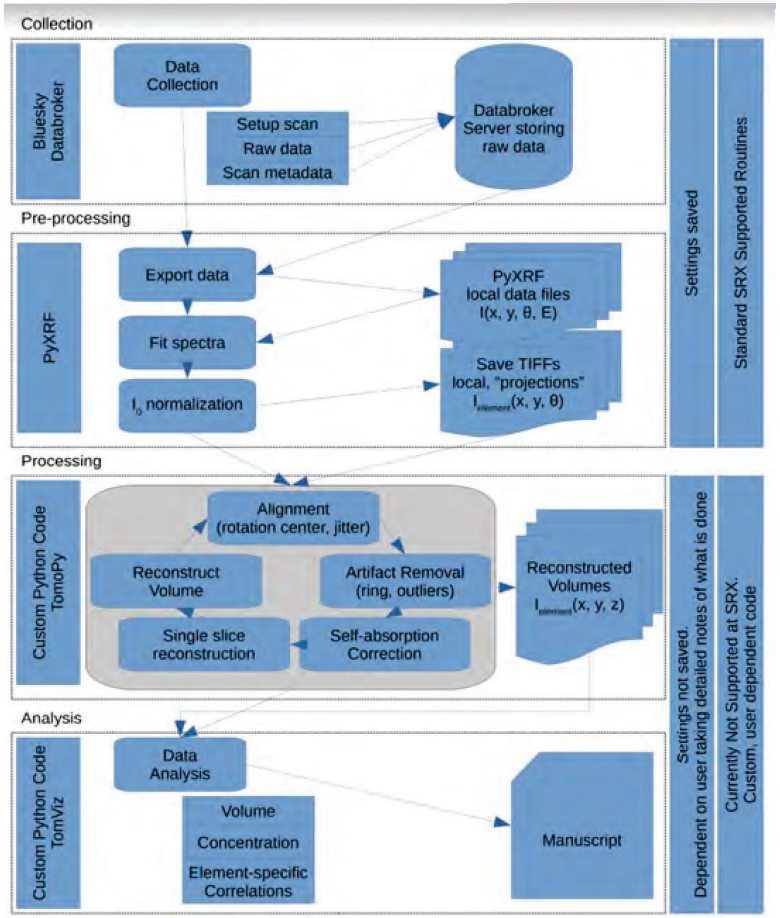
Workflows – Software View





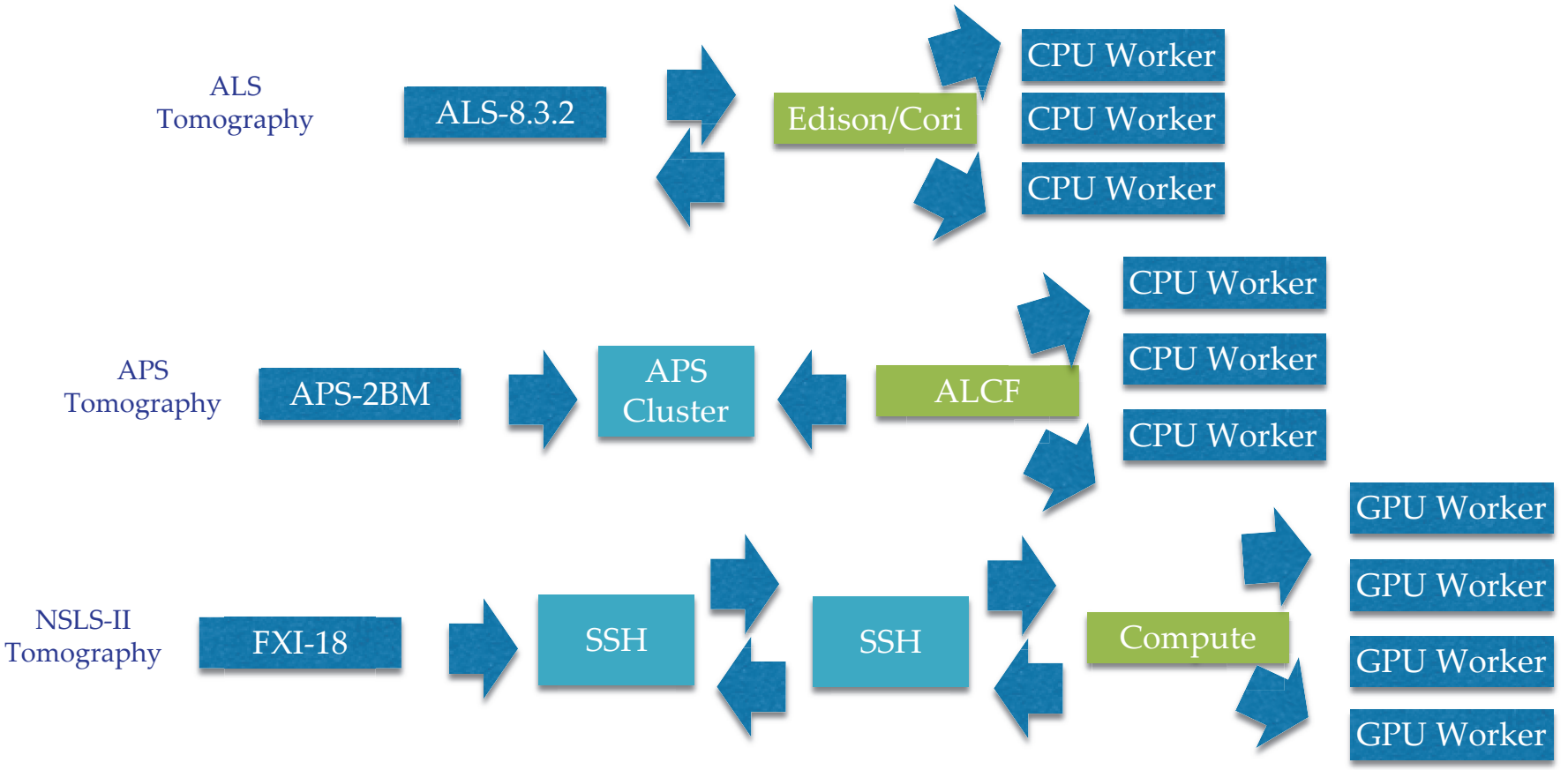
SSRL ALS





Credit: Data Pilot Tomography Breakout Report (All credit goes to respective authors)

Workflows – Hardware View



Challenges – Current

- Acquisition
  - Custom Instrumentation, Detectors, Drivers
- Networks
  - Custom Network Infrastructure & Authentication
- Hardware
  - Custom Hardware (FPGAs, GPUs, CPUs, …)
- Workflows
  - Custom analytics & Software dependencies

BES Light Source Data Generation and Computing Estimates

Year	Facility				
	ALS	APS	LCLS/LCLS-II	NSLS-II	SSRL
2021	3 PB	7 PB	30 PB	42 PB	15 PB
2028	31 PB	243 PB	300 PB	85 PB	15 PB

Estimated data generation rates per year at the BES Light Sources. At the ALS and APS, data generation will stop during 2025 and 2023, respectively, due to installations of new storage rings. Aggregate data generation across the BES Light Sources will approach the exabyte (EB) range.

Year	Facility				
	ALS	APS	LCLS/LCLS-II	NSLS-II	SSRL
2021	0.1 PFLOPS	4 PFLOPS	1 - 100 PFLOPS	2.5 PFLOPS	< 1 PFLOPS
2028	30 PFLOPS	50 PFLOPS	1 - 1,000 PFLOPS	45 PFLOPS	< 1 PFLOPS

Estimated PFLOPS of on-demand computing resources required by each of the BES Light Sources by 2021 and 2028. Compute jobs requiring < 10 PFLOPS are common and best run on local resources; compute jobs requiring > 10-20 PFLOPS are best suited to run at a high-end computing facility.

High-Priority Shared Needs

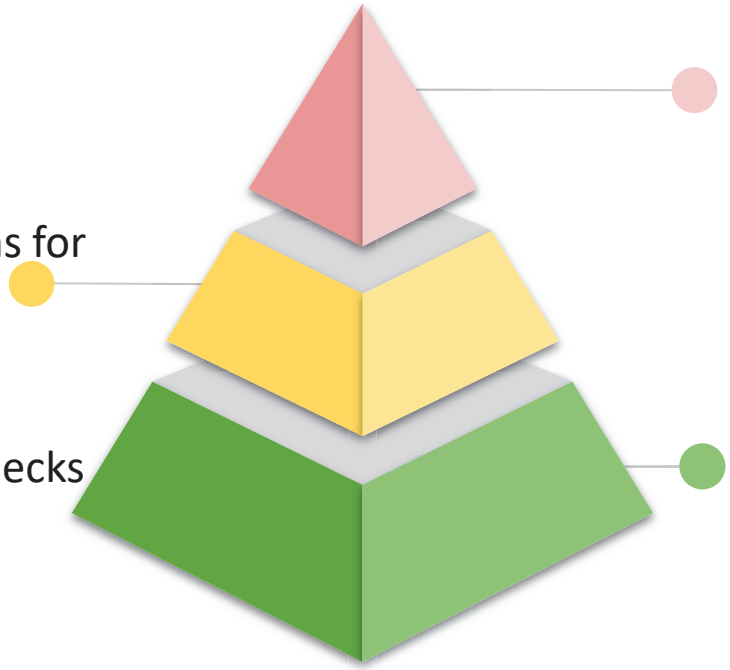
Mission requires computing advances in four main areas

- **Data management and workflow tools**
  - Integrate beamline instruments with compute and storage
- **Real-time data analysis capabilities**
  - Reduce data volumes
  - Provide feedback during experiments
  - Apply tools to steer data collection (algorithms, ML, simulation)
- **On-demand utilization of computing environments**
- **Data storage and archival**

Building on Common Software Tools (BES Data Pilot Project)

Advanced Analysis & Automation

- ❑ Advanced analysis & visualization (Post-reconstruction)
- ❑ Multi-modal analysis
- ❑ Streaming (real-time) analysis
- ❑ Automated Acquisition (ML support)



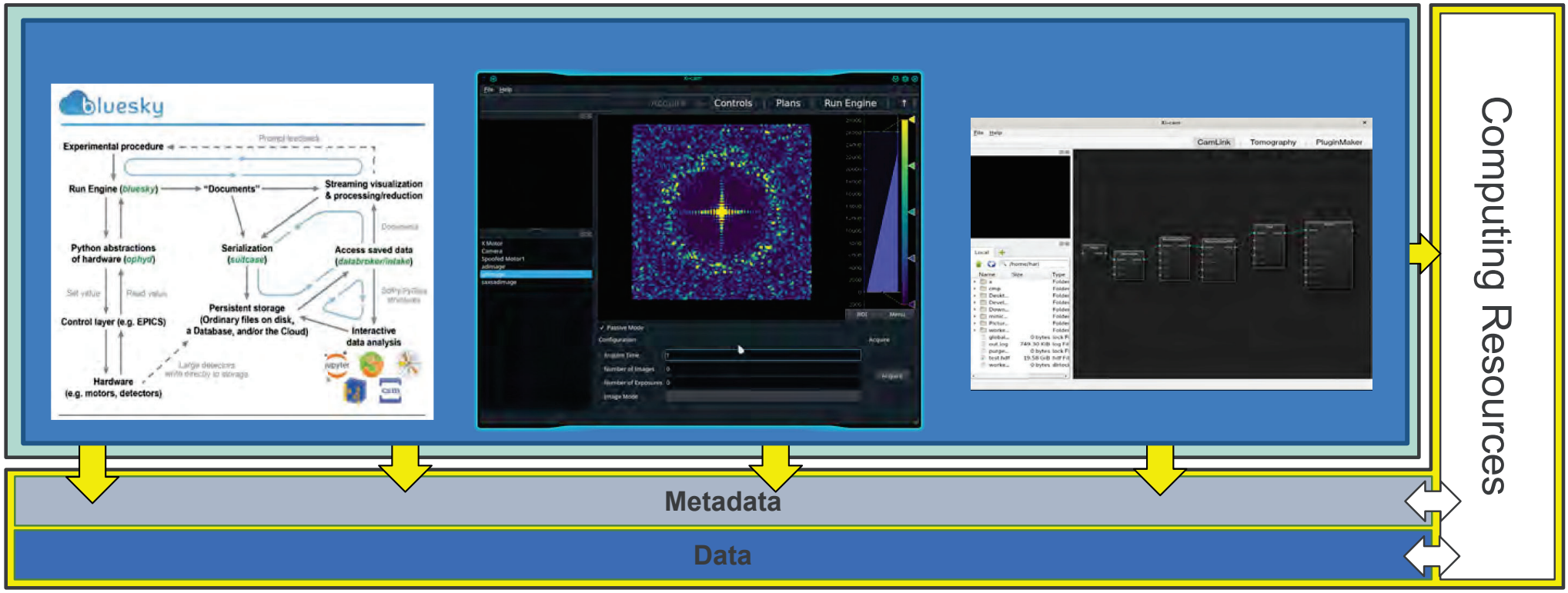
Algorithms & Data Quality

- ❑ Implementation of reconstruction algorithms for shared use
- ❑ Advanced visualization features
- ❑ Real-time data quality checks

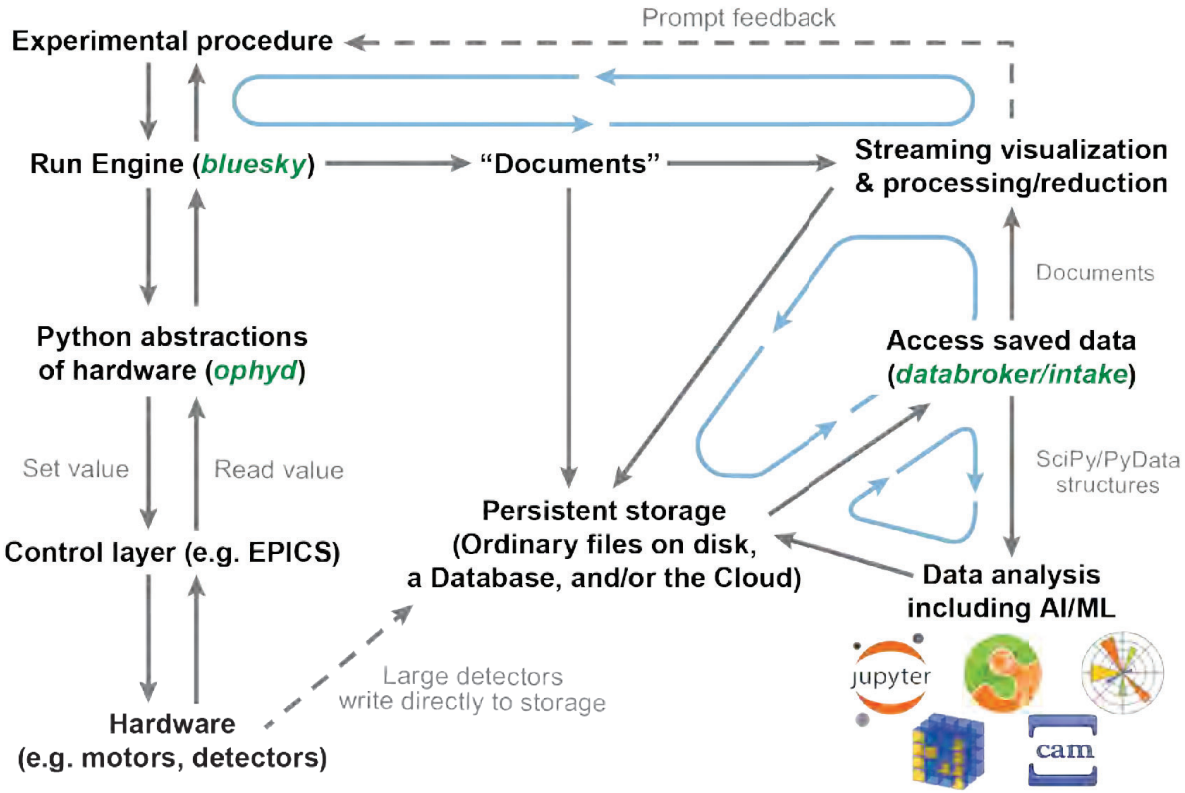
Exchange & Standardization

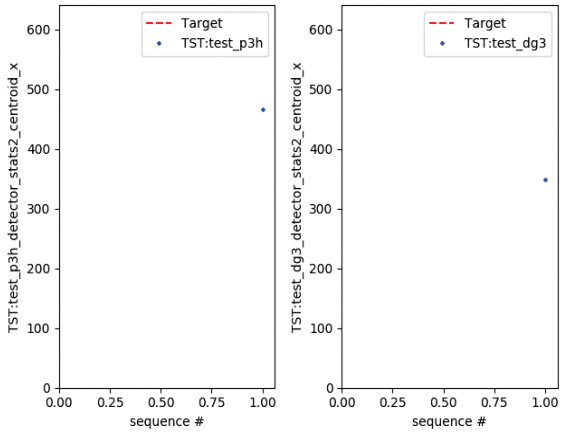
- ❑ Common database
- ❑ Knowledge base of software & algorithms
- ❑ Standardizes data structures
- ❑ Standardized acquisition

Towards a Sustainable Software Stack



Part 1: Acquisition & Controls



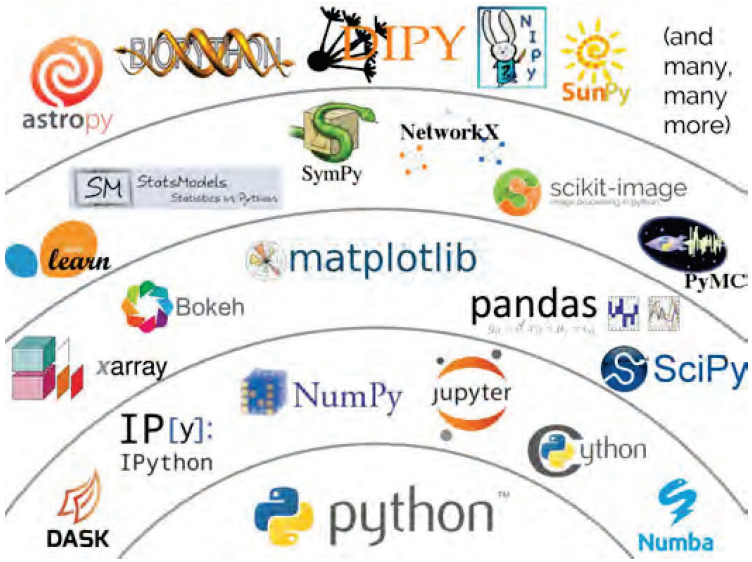


Goal of the Bluesky Project overall:

Make it easy for synchrotrons to leverage the ecosystem of freely available, open-source scientific Python community tools.

bluesky

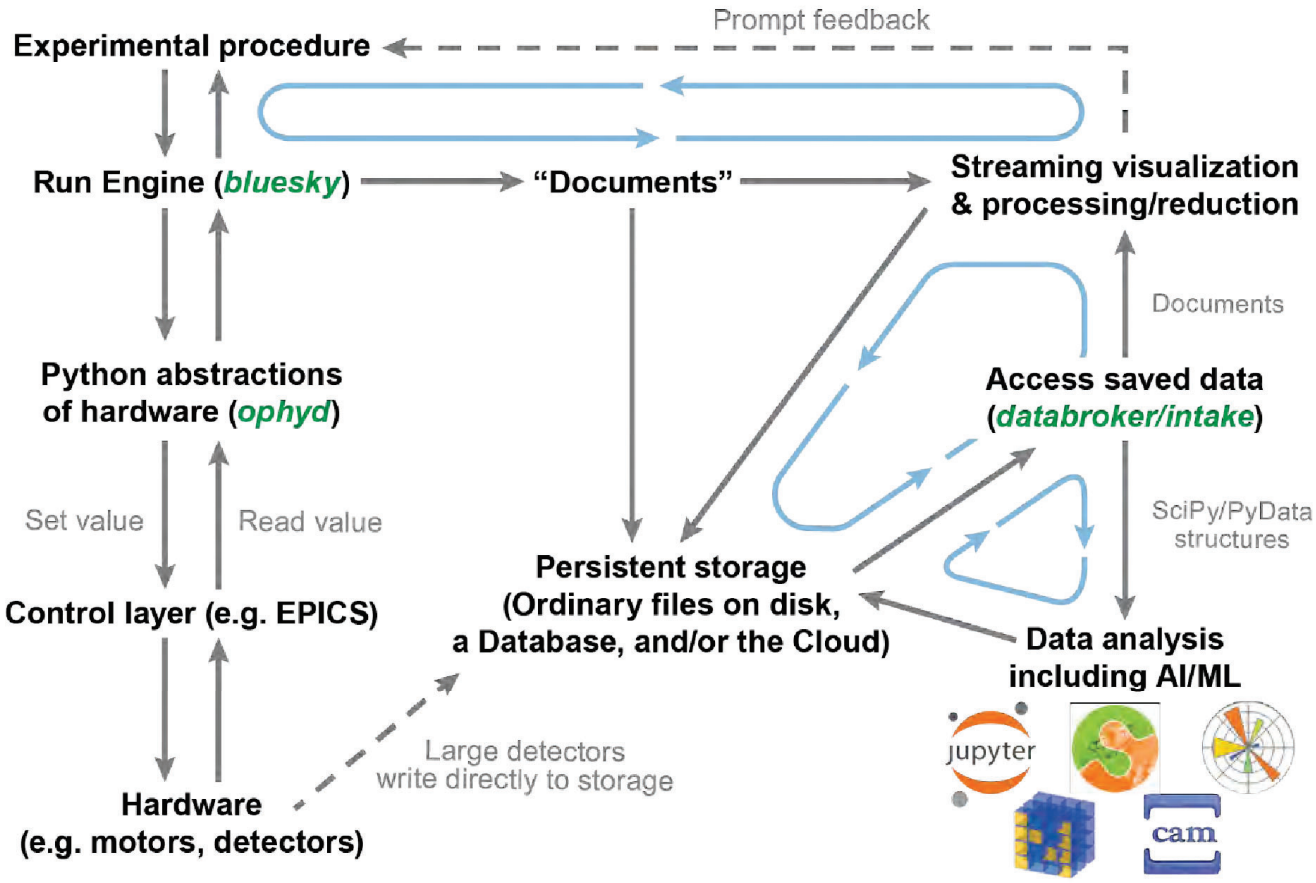
DOE Light Sources



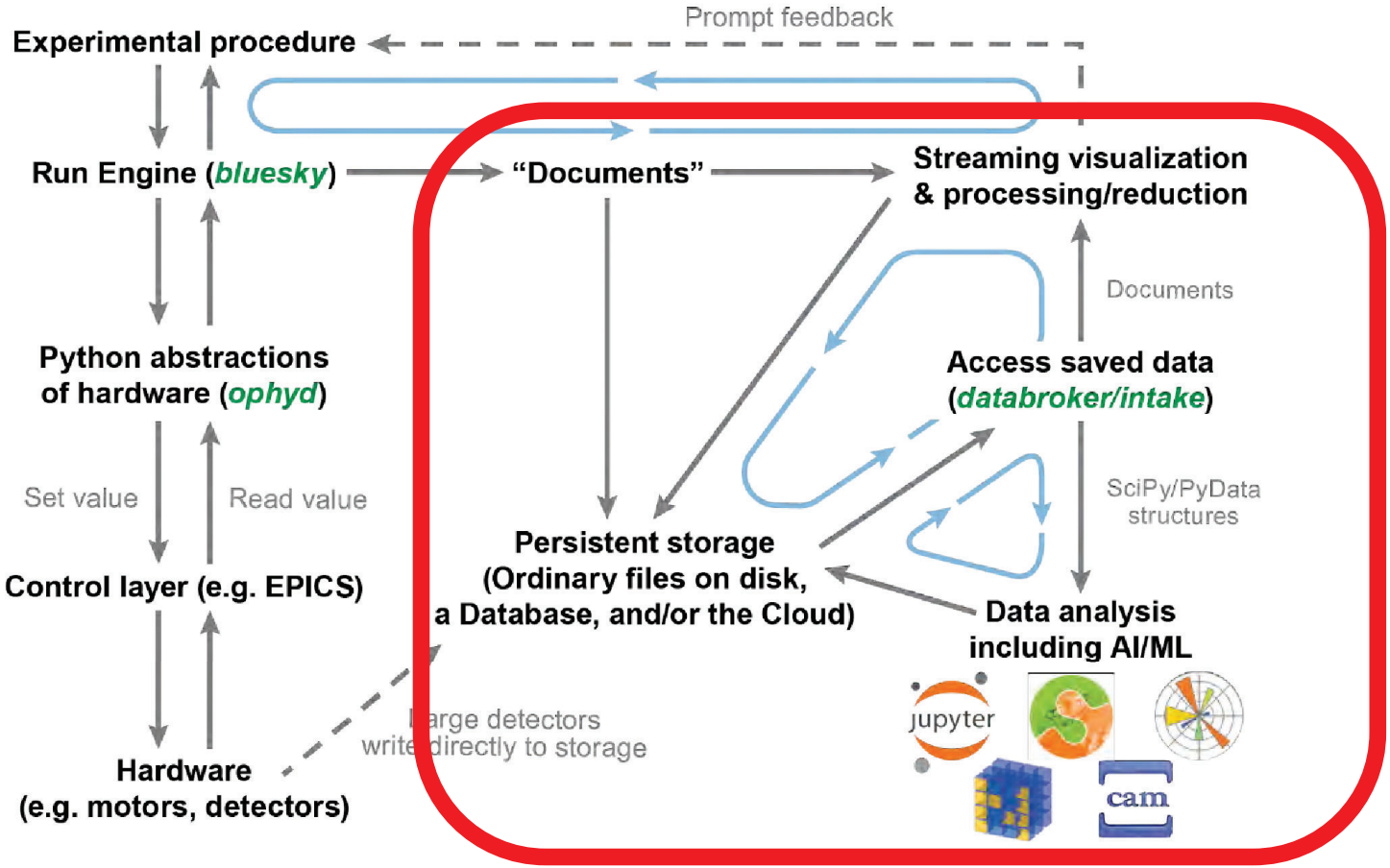
***Bluesky Data Broker**: Search and retrieve scientific data for interactive and automated data analysis.*

Figure Credit: Jake vanderPlas, "The Unexpected Effectiveness of Python in Science", PyCon 2017

Bluesky Ecosystem

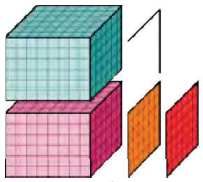


Data Broker


Prompt feedback
Experimental procedure
Run Engine (bluesky)
"Documents"
Streaming visualization & processing/reduction
Documents
Python abstractions of hardware (ophyd)
Access saved data (databroker/intake)
Set value
Read value
SciPy/PyData structures
Persistent storage (Ordinary files on disk, a Database, and/or the Cloud)
Data analysis including AI/ML
Large detectors write directly to storage
Control layer (e.g. EPICS)
Hardware (e.g. motors, detectors)
jupyter
cam

How Data Broker fits into this

- Released Data Broker Version 1.0 installed at the five DOE Light Sources
- **Unify data access across the facilities**
- Improved **usability**, incorporating 5 years of user feedback on “beta” versions
- New, hands-on **tutorial materials for scientists** at blueskyproject.io/tutorials
- Leverages **community scientific Python projects** under the hood for...



xarray

Labeled, physically-meaningful data structures

DASK

Scaling across thousands of nodes on HPC, cloud, or traditional servers

INTAKE

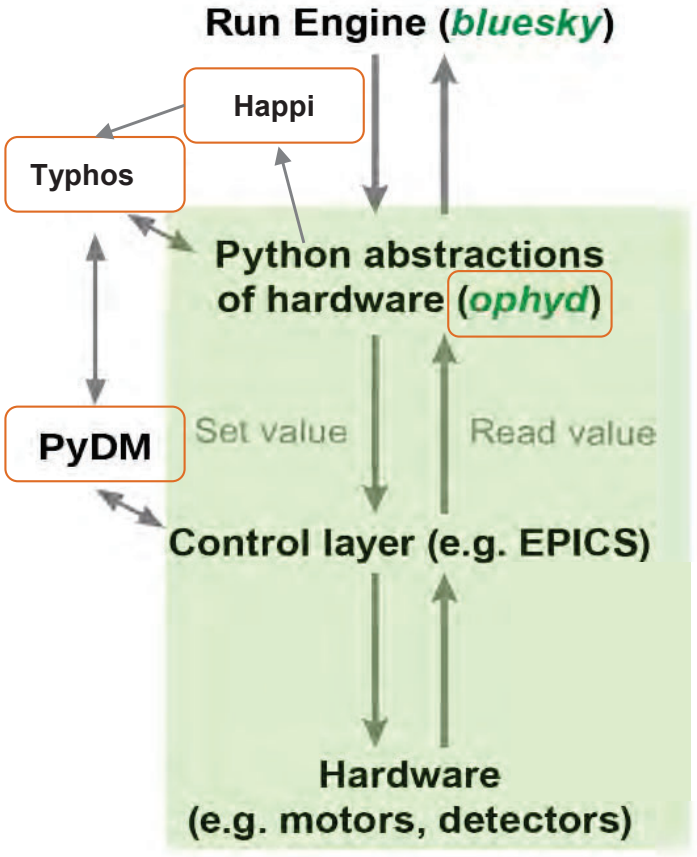
Unopinionated about data formats

(This work also funded partly by light source facility operations.)

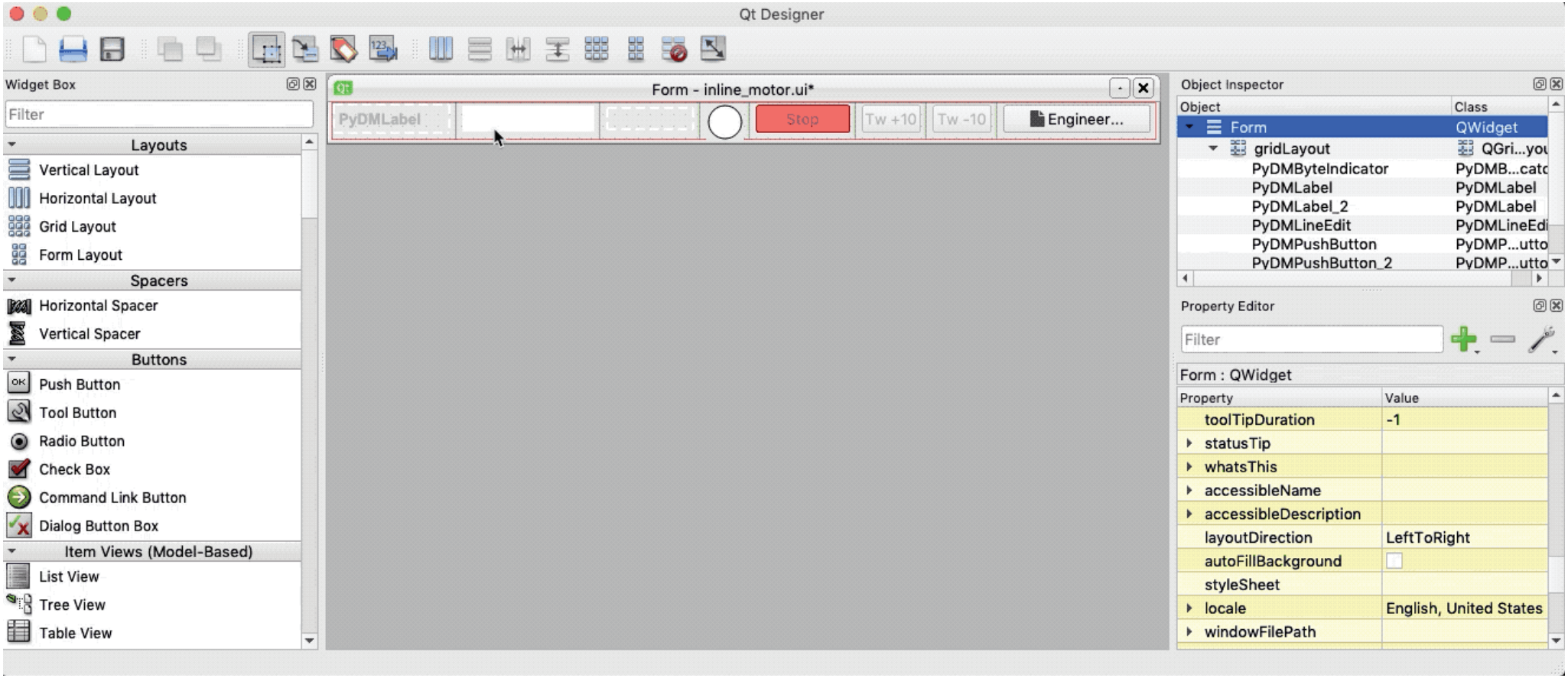
Controls Integration within the Bluesky Ecosystem

Development of user interface frameworks that facilitate data acquisition and intelligent beamline control applications across the DOE light-sources, including:

Happi	Device location/attribute DB
ophyd	Device abstraction layer
PyDM	Python Display Manager
Typhos	User Interface generator

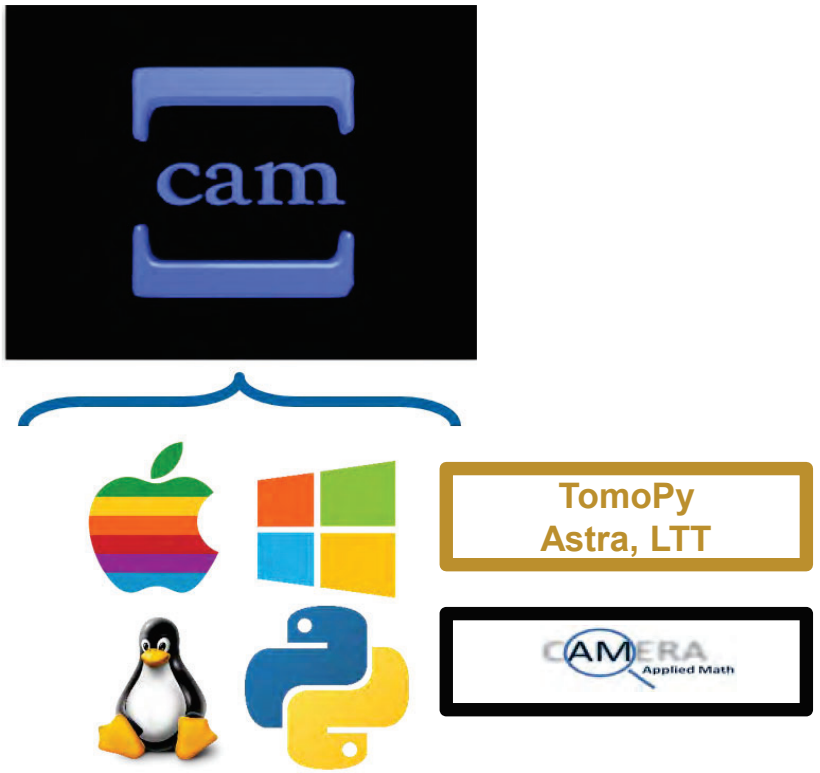


Overview - Ease of use (PyDM & Qt Designer)



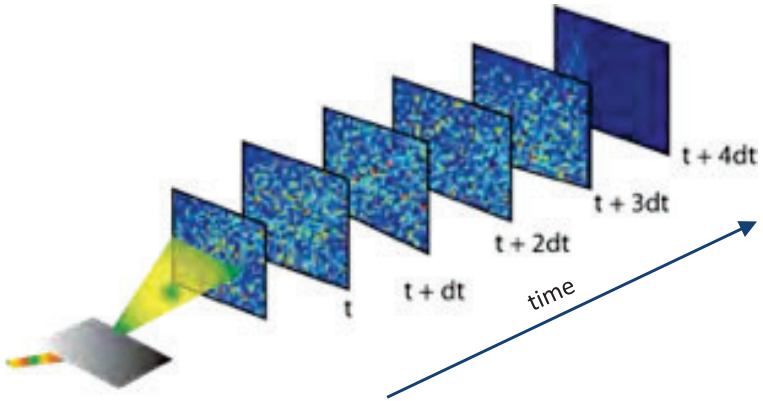
Part 2: Analysis & Algorithms – Xi-CAM & Workflows

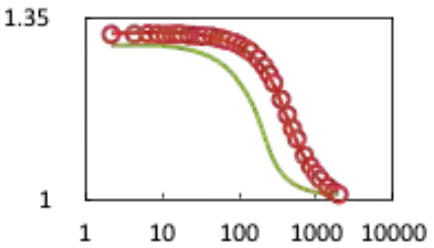
- GUI frontend and extensible framework for synchrotron data…
  - acquisition
  - analysis
  - visualization
  - management
- Utilizes software components developed by many external groups, including NSLS-II, APS, ALS, and SLAC
- Deployment platform for analysis algorithms,
  such as those from CAMERA



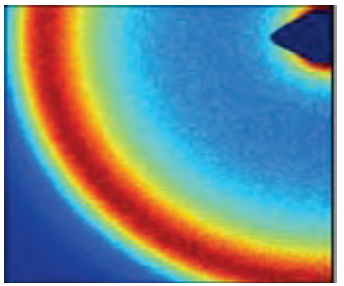
XPCS

- **X**-ray **P**hoton **C**orrelation **S**pectroscopy
- Probes dynamics/fluctuations in materials
  length scale: <1μm - nms
  time scale: minutes - milliseconds
- X-ray data are 2D image series that exhibit speckle fluctuations (sample dynamics)
- 1st XPCS in 1995 and emerging technique
  - Increasing coherent flux
  - Faster time scales (nanoseconds)
  - Tunable beamline energies for atomic species
  - *in-situ* or *in-operando* experiments





g_2 calculation at two different length scales



2D small-angle scattering pattern from a suspension of silica spheres.

Ptychography

- Scanning, coherent diffractive imaging technique (CDI)

  → latest generation light sources with bright beams

- Extremely high spatial resolution (low nanometer)
- Versatile application

  → flexible field of view, geometry, energy range

  → many kinds of samples

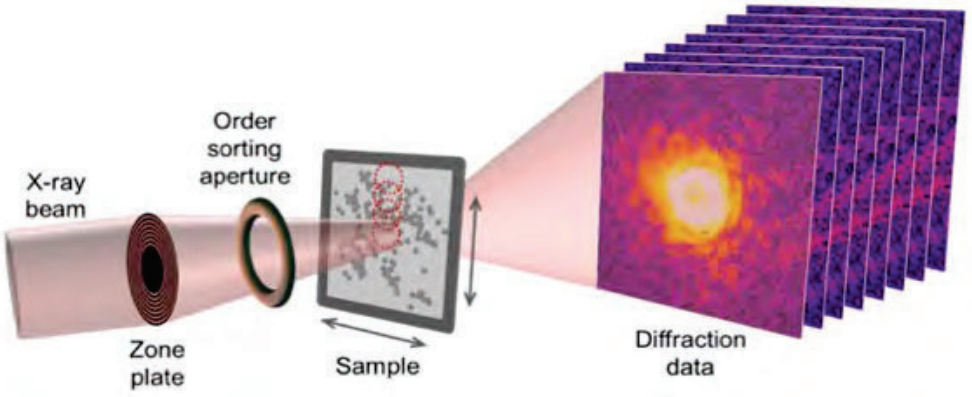
- Complementary techniques

versatility

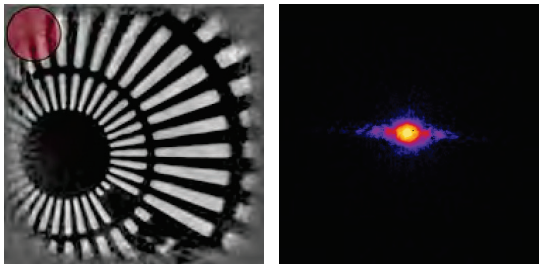
very popular

many algorithms

high data rates

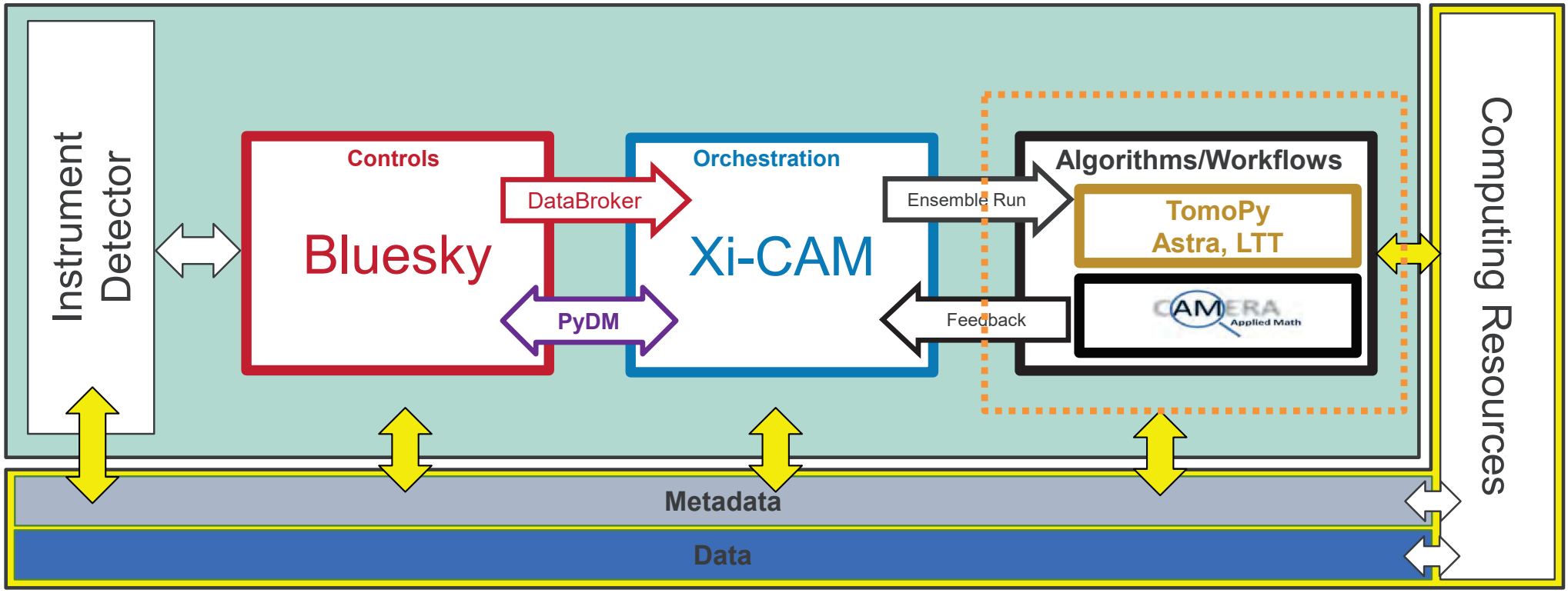


Exemplary ptychography setup, source: Weker Group, SSRL

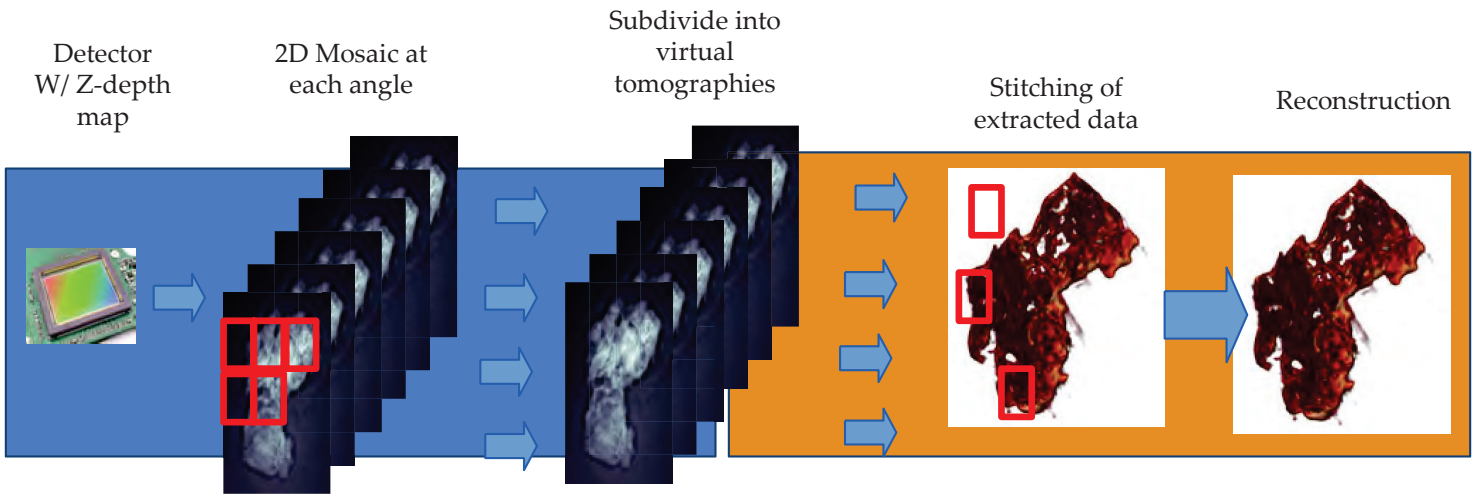


Scanning the sample and corresponding diffraction patterns

Putting it all together



From Design to Execution: Tomography @ FXI-18 (NSLS-II)



Interleaved:
1. Take flats
2. Re-calibrate X-ray
3. Align system

For each image:
a. Apply flat field correction
b. Image quality check, re-acquire if failed
c. Remove outliers
d. Perform ring removal
e. Apply distortion correction
f. Potentially apply point spread function deconvolution

For each tomography scan:
a. Perform inter-angle alignment (rigid x/y shifts to align images)
b. Quick reconstruction to estimate angle of IC in theta and phi
c. Warp sinograms to align IC layers in reconstruction for extraction
d. Reconstruct
e. Layer extraction, segmentation, etc…

Deploying the Standardized Stack

Highlight: Successfully ran real time analytics Tomography pipeline.

Live Processing mode - as data is acquired

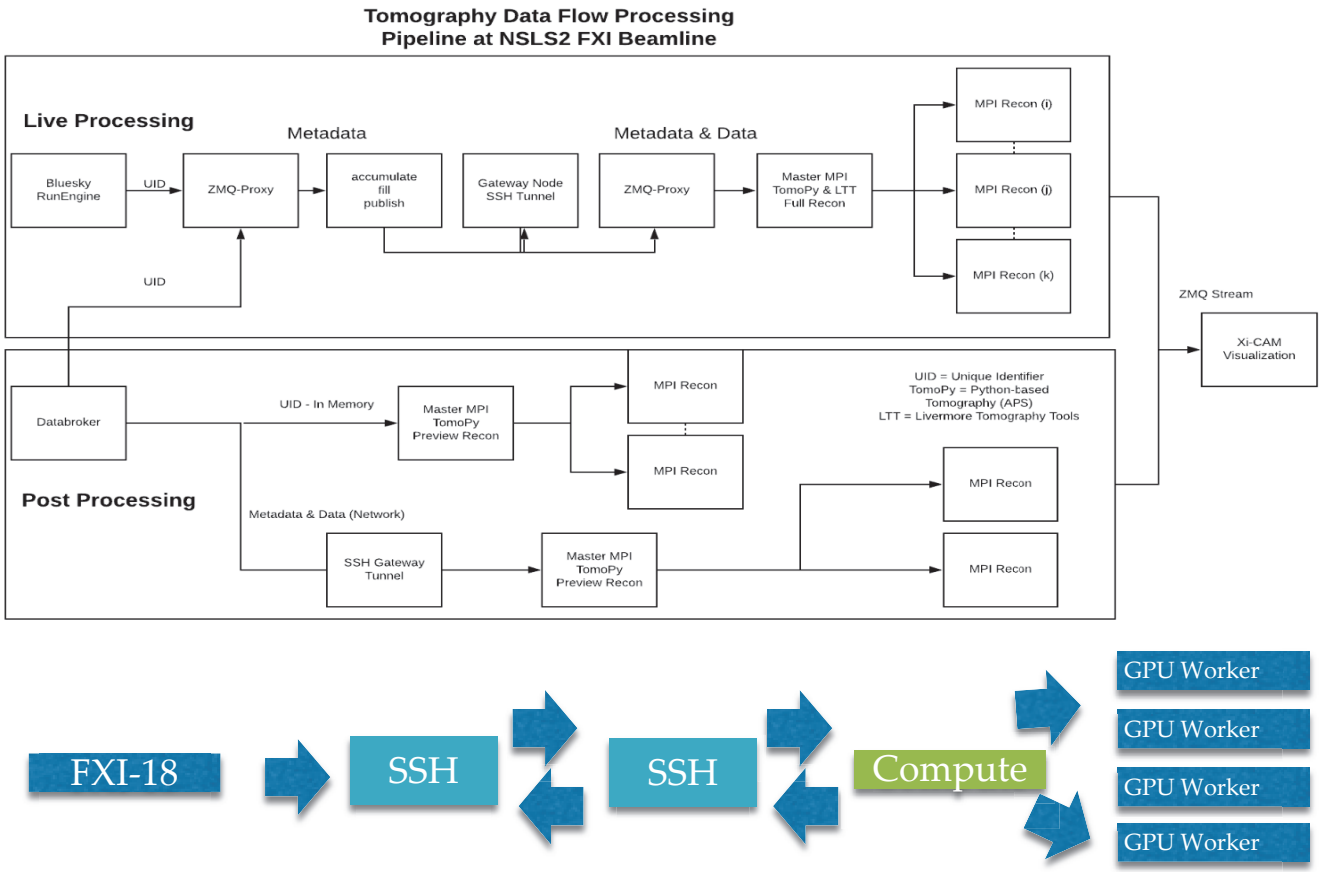
Post Processing mode - using data broker to fill the live analysis pipeline and trigger reconstructions on remote computational hardware.

Lessons Learned: Demonstration to Deployment

Disk: Resolving SWMR issues at the detector level would enable true streaming.

Network: Resolving Network issues, would enable overlap

Algorithm - Binning 1 data requires 211 GB of memory (Requires HPC to run in real time)



CREDITS & ACKNOWLEDGEMENTS

Thank You!

APS
Pete Jemian
Faisal Khan
Suresh Narayanan
Alec Sandy
Nicholas Schwartz
Qingteng Zhang

LBNL
Alexander Hexemer
John Joseph
Roland Koch
Sujoy Roy
Dula Parkinson
Dylan McReynolds
Charles Melton

CAMERA
Dinesh Kumar
Ian Humphrey
Harinarayan Krishnan
Ronald Pandolfi
Pablo Enfaduque
Marcus Noack
James Sethian
Dani Ushizima

BNL
Daniel Allan
Stuart Campbell
Thomas Caswell
Maksim Rakitin
Andi Barbour
Andrei Fluerasu

SLAC
Kenneth Lauer
Teddy Rendahl
Hugo Slepicka
Jana Thayer
Apurva Mehta

Acknowledgements
- CAMERA
- ALS, APS, LCLS, NSLS-II, SSRL, …
- LBNL, ANL, BNL, SLAC

Community effort:
Many others…

Daniel Allan
Thomas Caswell

Robert Tang-Kong

Ronald Pandolfi

Bluesky Data Broker Pilot Breakout Summary

Lead — Daniel Allan, NSLS-II

Co-lead — Dylan McReynolds, ALS

Dylan McReynolds

Pete Jemian

Jana Thayer

Acquisitions & Controls GUI Breakout

Co-Leads: Daniel Flath[a], Robert Tang-Kong[b]

Hugo Slepicka[a], Juliane Reinhardt[c], Thomas Caswell[d], Apurva Mehta[b], Ronald Pandolfi[e], Jana Thayer[a], Zachary Lentz[a], Ken Lauer[a], Pete Jemian[f]

a LCLS

b SSRL

c ALS ADVANCED LIGHT SOURCE

d NSLS-II NATIONAL SYNCHROTRON LIGHT SOURCE II

e CAMERA Applied Math

f Argonne NATIONAL LABORATORY Advanced Photon Source

XPCS Pilot Breakout

Lead: Andi Barbour

Co-Leads: Faisal Khan
Qingteng **Z**hang

Shared success by the hard work of many:

Eric Dufresne
Pete Jemian
Faisal Khan
Suresh Narayanan
Qingteng Zhang

Argonne NATIONAL LABORATORY

Ian Humphrey
Roland Koch
Dinesh Kumar
Dylan McReynolds
Sophie Morley
Ronald Pandolfi
Juliane Reinhardt
Sujoy Roy

BERKELEY LAB

Dan Allan
Andi Barbour
Garrett Bischof
Tom Caswell
Andrei Fluerasu
Josh Lynch
Maksim Rakitin
Yugang Zhang

BROOKHAVEN NATIONAL LABORATORY

Data / Computer Scientist in Breakout Discussions
X-ray Beamline Scientist in Breakout Discussions
Contributor to XPCS effort of Data Solutions Pilot