Addressing the Data Challenges at the Advanced Light Source

Alexander Hexemer
Senior Staff Scientist
Program Lead for Computing
Advanced Light Source
Center for Advanced Mathematics for Energy Research Applications (CAMERA)
Lawrence Berkeley National Lab

We can’t do it alone!!!

Acknowledgment

Lawrence Berkeley National Lab

- Ronald J Pandolfi
- Dinesh Kumar
- Dylan McReynolds
- Singanallur Venkatakrishnan (now ORNL)
- Luis Barrosso-Luque
- Holden Parks
- Austin Blair
- Dilworth Parkinson
- Shuai Liu
- Nathan Melton
- Andrew Wiedlea
- Debbie Bard
- Hari Krishnan
- Krishna Muriki
- Dani Ushizima
- James A Sethian

Argonne National Lab

- Zhang Jiang
- Doga Gursoy
- Francesco De Carlo
- Xianghui Xiao
- Ian Foster
- Nicholas Schwarz
- Ryan Chard

SLAC

Amanda Fournier
Fang Ren
Yury Kolotovsky
Apurva Mehta
Chris Tassone
Amedeo Perazzo

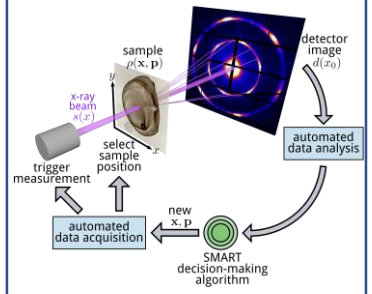
Brookhaven National Lab

Masafumi Fukuto
Kevin Yaeger
Thomas Caswell
Stuart Campbell

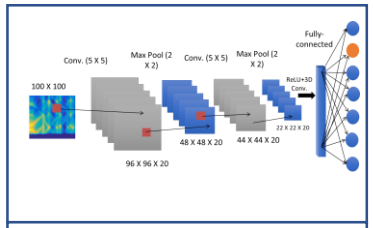
Funding Acknowledgments

- LBNL LDRD Program
- DOE Early Career Award
- CAMERA

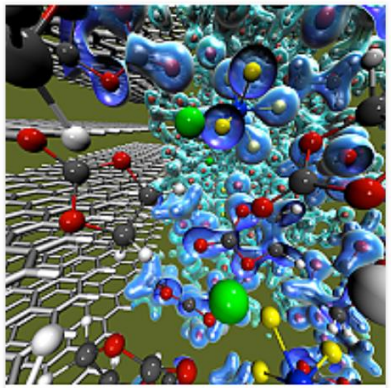
Center for Advanced Mathematics for Energy Research Applications



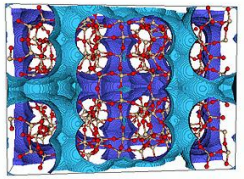
Autonomous Data Collection



Machine Learning

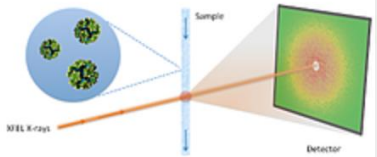


Electronic Structure

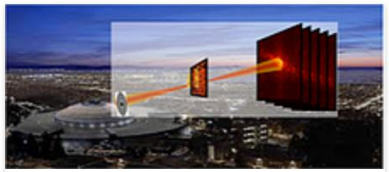


Material Informatics

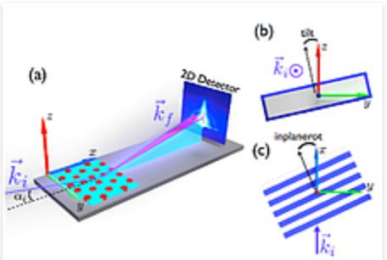
Nanocrystallography



Fluctuation Scattering



Ptychography



GISAXS

Image Analysis

Jamie Sethian
Head of CAMERA

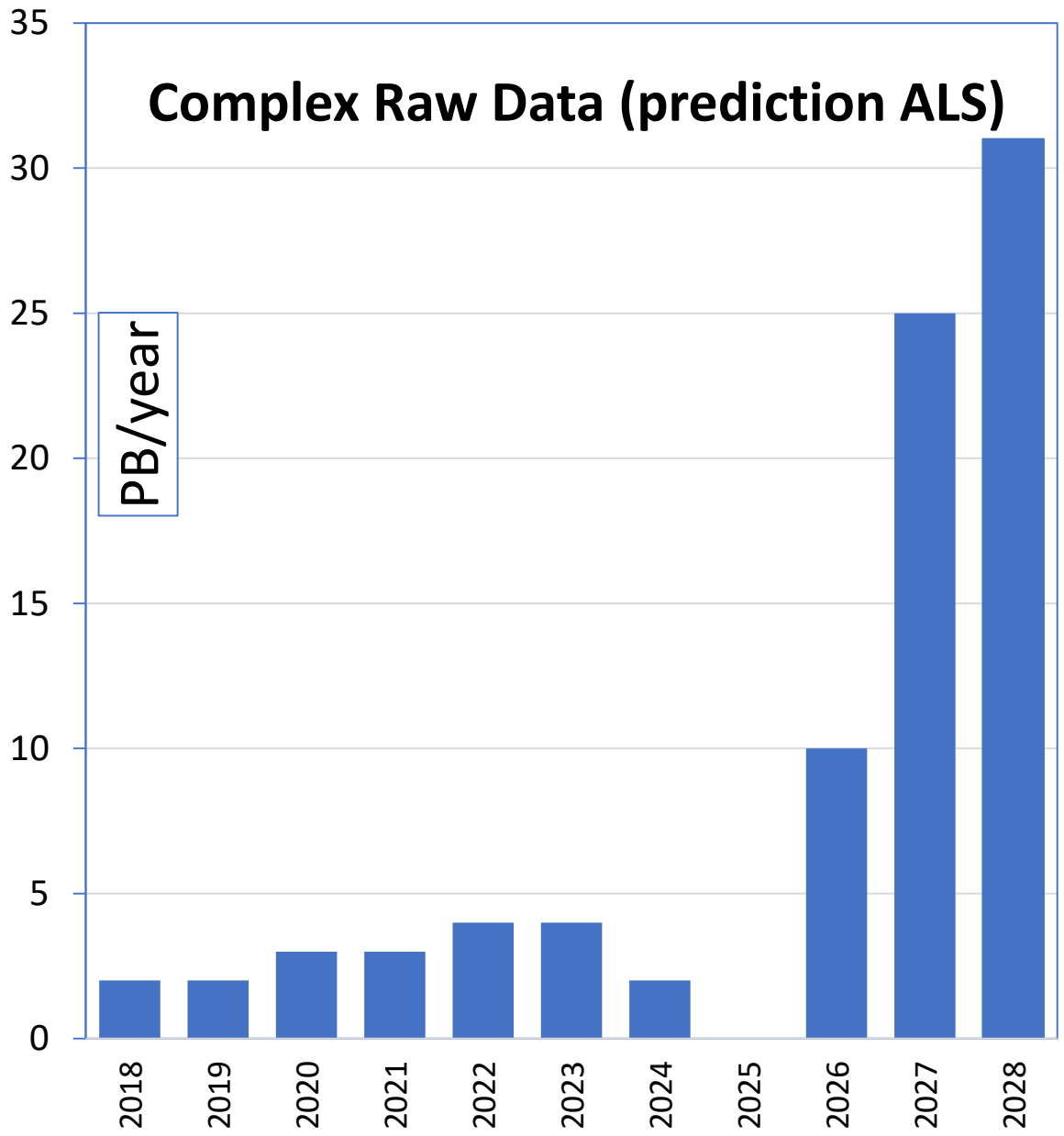
DOE: BES and ASCR
Renewed last year

Include latest theory and math, take advantage of latest architecture:
Multi CPU/GPU, open source, everything shared, many collaborations

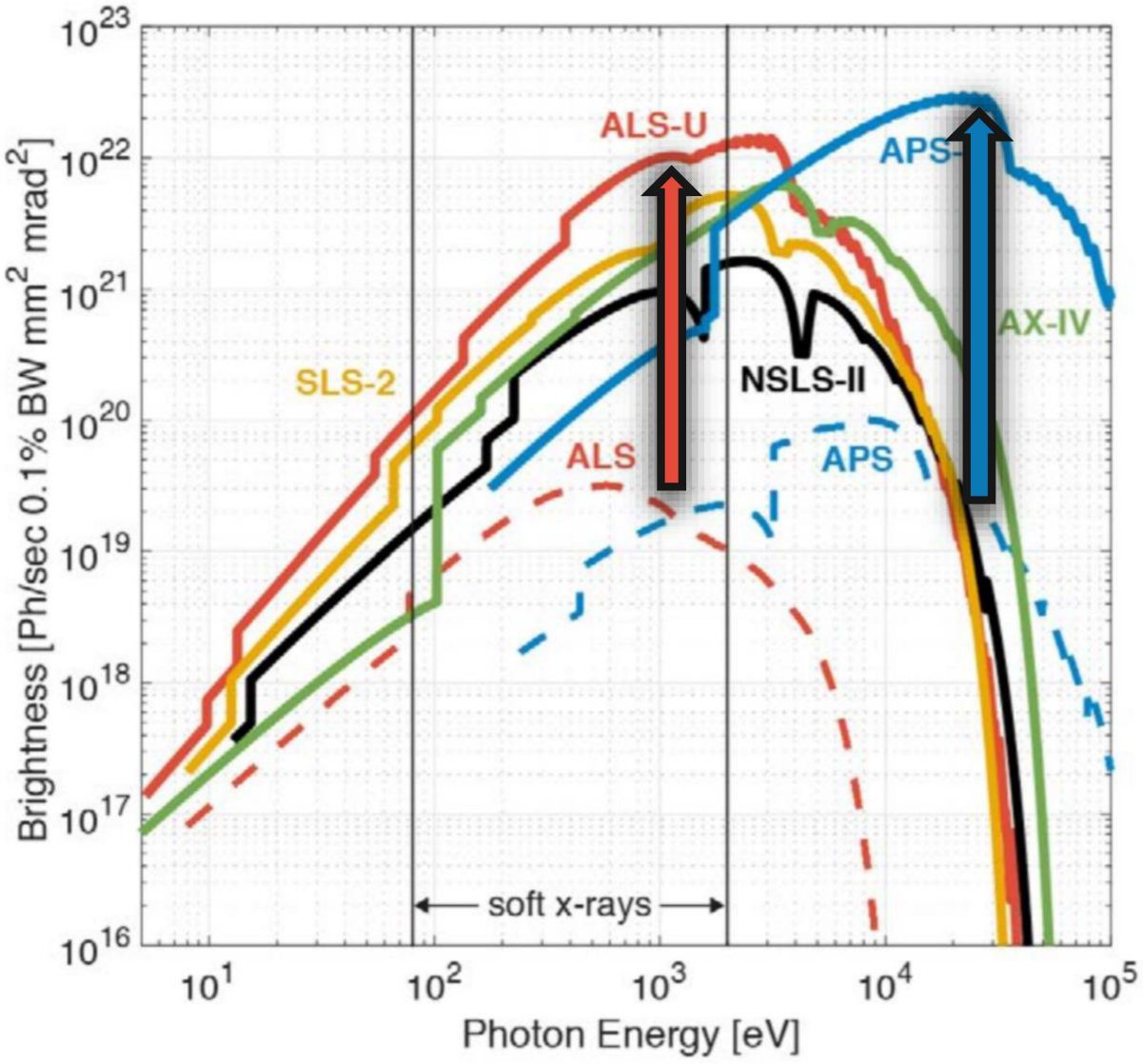
www.camera.lbl.gov

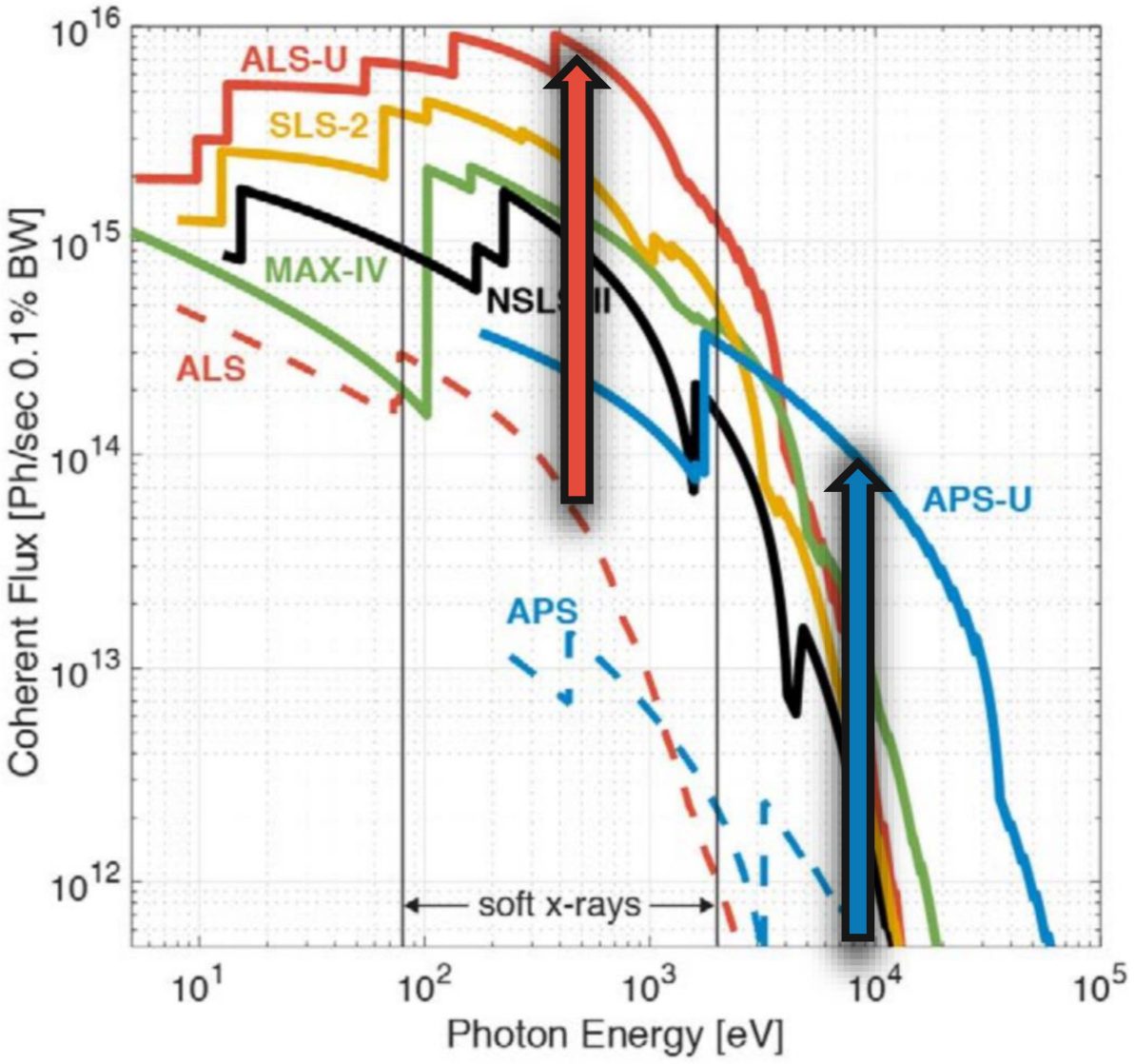
Challenges for User facilities

- >20% new users
- Provide very fast feedback and/or experiment combining data from different modalities
- Applying custom workflow to many data sets
- Need for new math and algorithms
- Make things easy and faster for users
- Data access across facilities
- Large amounts of data



New Sources are on the Horizon





The Light Source Data Working Group

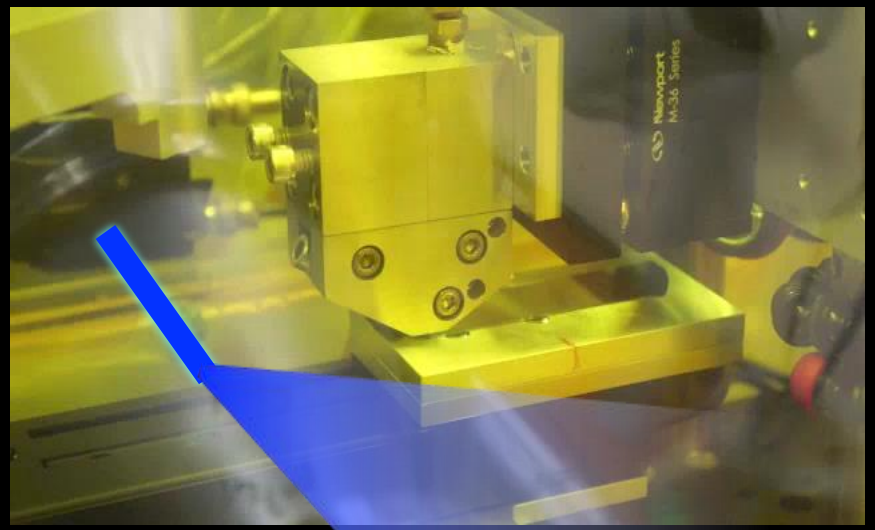
The purpose of the DWG is to serve as a resource for the 5 BES light source Facility Directors to call upon to provide information and recommendations on working together in the areas of data and computing.

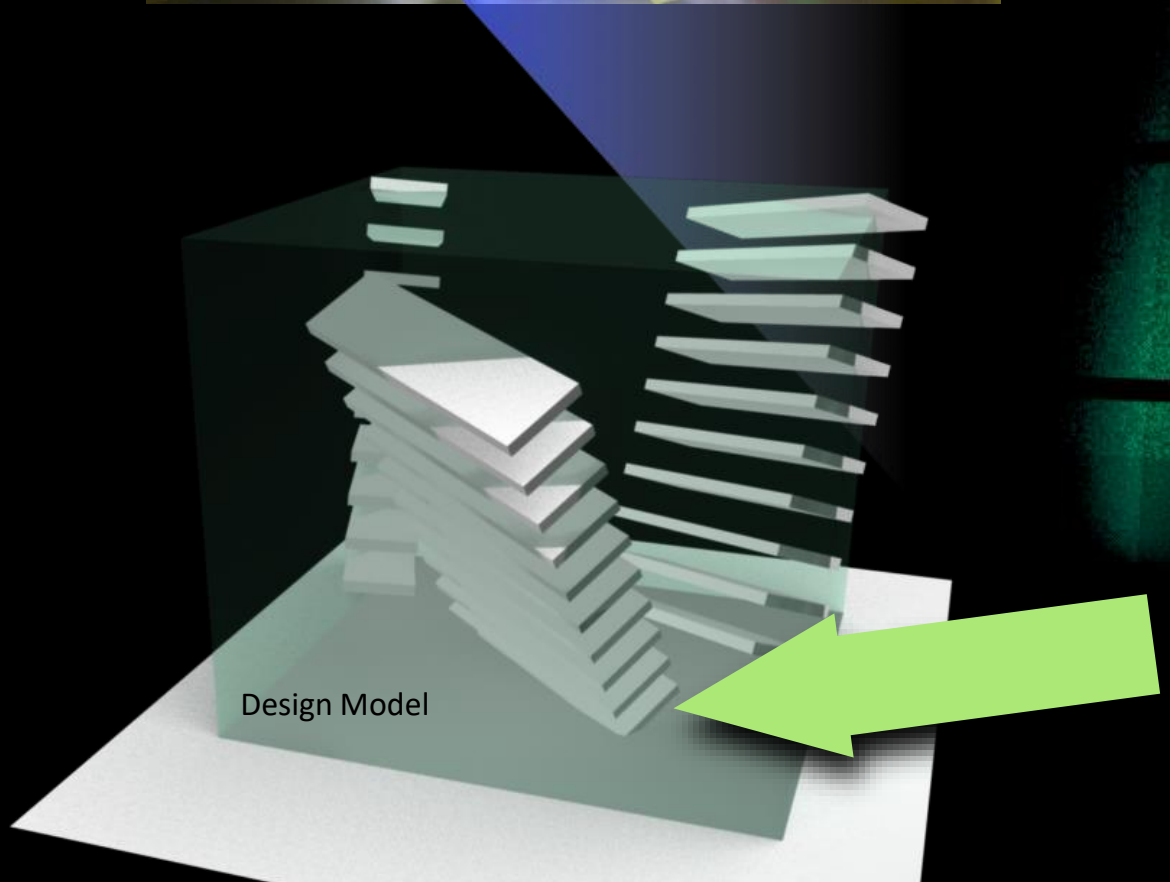
- Development of the standards for data storage and file formats
- Create reports and outline ways for the light sources to work together.
- Explore the topic of real time computing at the 5 light sources and into the future.
- Data access across facilities
- Regular meetings online and in person.

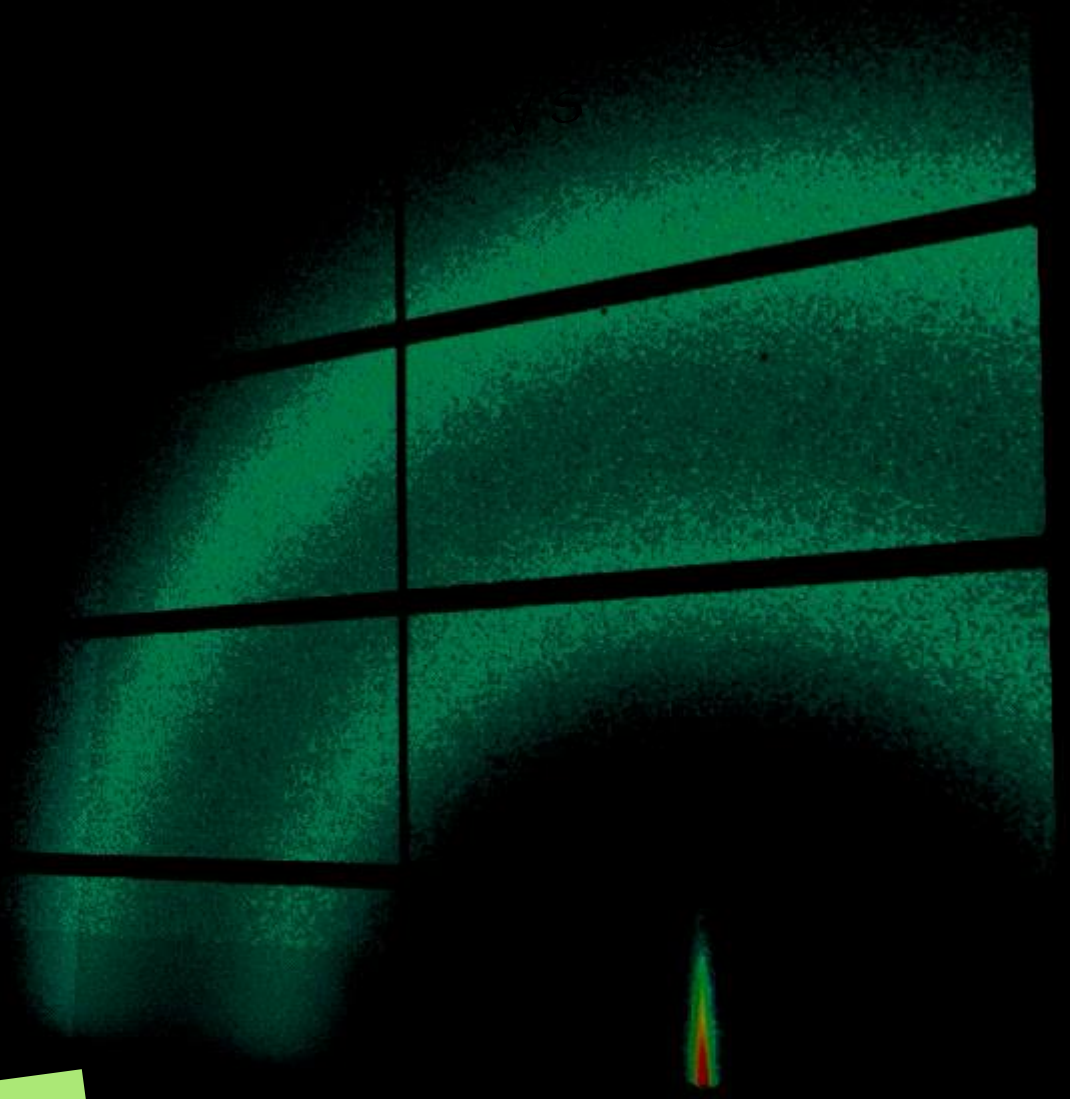
ALS (Alexander Hexemer), **APS** (Nicholas Schwarz) **LCLS II** (Amedeo Perazzo), **NSLS II** (Stuart Campbell)

Development of the Data Movement in Collaboration with other Divisions of LBNL

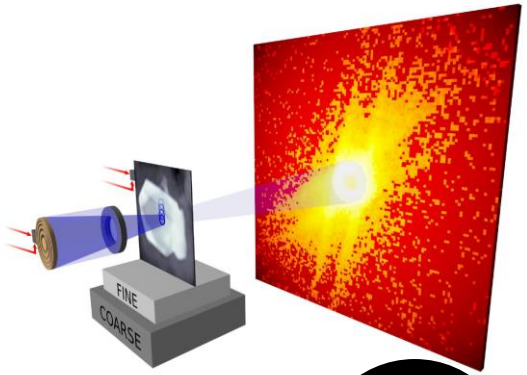
- **NERSC** (Debbie Bard et al.): HPC API specs are currently developed, GPGPUs
- **Lab science IT** (Andrew Wiedlea et al.): GPGPUs, data storage, infrastructure, Mongo and many other ideas
- **ESNet** (Kate Mace et al.): Network challenges
- **Data Acquisition Group** (Kevan Anderson et al.): Data and Metadata access (NEXUS), development of a python/scripting interface to allow for analysis driven control of the beamline
- **CAMERA** (Jamie Sethian et al.): Analysis, Xi-CAM, cam link, HPC API, sharp and many more


Newport
M-36 Series


Design Model



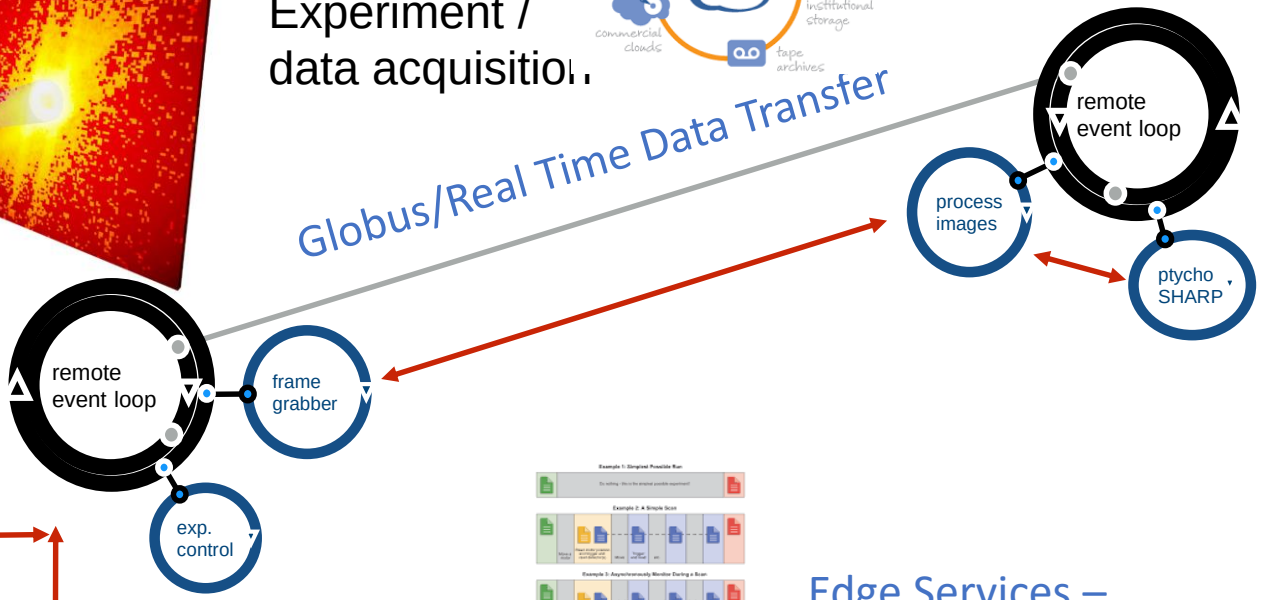
Full data lifecycle: Xi-CAM-Link [Data]

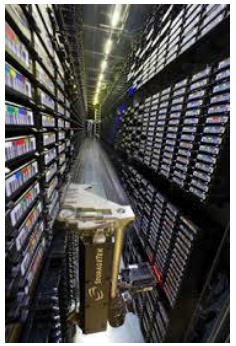


Experiment / data acquisition

Compute cluster

Globus/Real Time Data Transfer





Remote Storage/Tape Archive

- Compute
- Publishing
- Archiving

Local Storage

QC, Triage, Local Analysis (FFB)

Edge Services –

DataBroker: Event Based

- Metadata
- Provenance
- Data Tracking

MongoDB

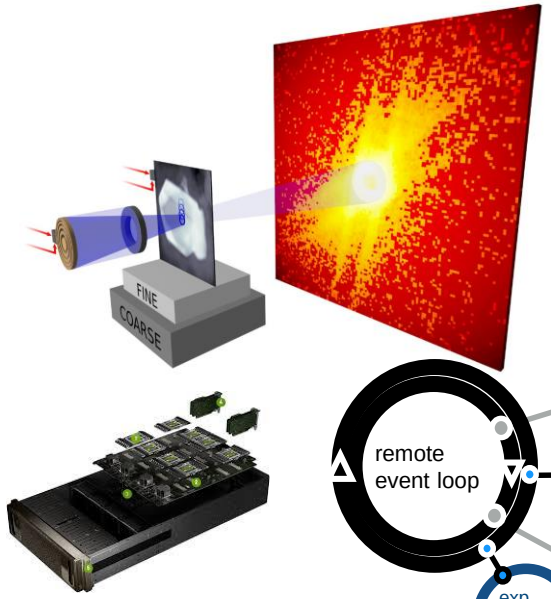
- Workflow State
- Data Cataloging
- Data Tagging
- Logging

CAM-Mart - Algorithm Marketplace

Execution steps

1. Identify and setup resources
2. Launch services
3. Connect network
4. Execute graph

Full data lifecycle: Xi-CAM-Link [Analysis]



Experiment / data acquisition

Compute cluster

remote event loop

frame grabber

exp. control

CAM-Link

remote event loop

process images

ptycho SHARP

elog

remote control

master event loop

User / local computer

Graphical User Interface (GUI)

- Compression
- Preprocessing
- Analysis close to detector

Python-based
- Remote Workers
- Support Docker/Singularity (WF Provenance)

Python-based
- Perform Queries, Searches
- Programmable Pipelines
- Visualization

Execution steps

1. Identify and setup resources
2. Launch services
3. Connect network
4. Execute graph

Deployment of Mathematical Algorithms: Xi-cam

Result of DOE Early Career Award

Scientific Achievement

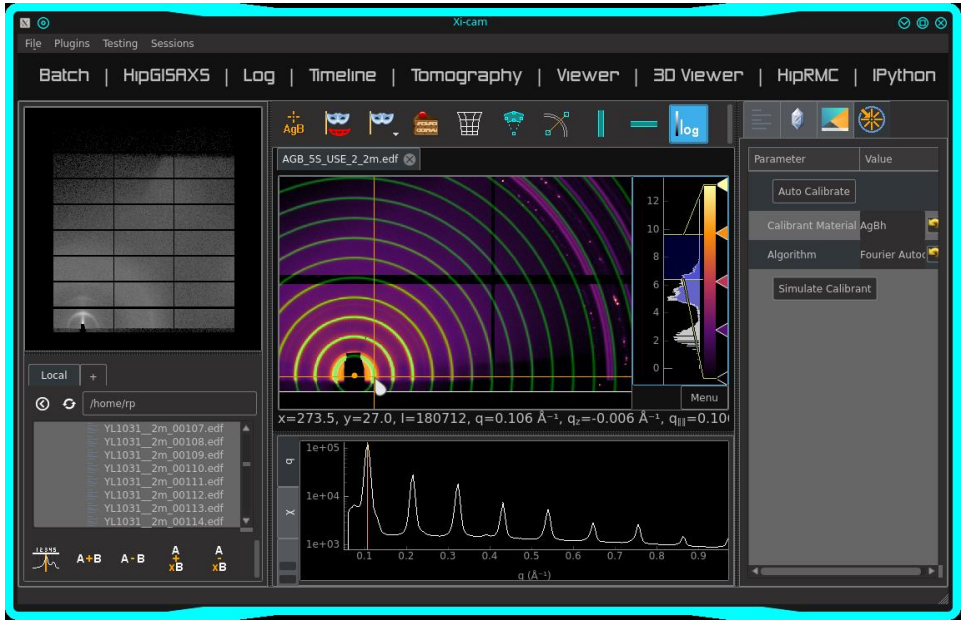
Development of a **community-maintainable** platform for new analysis and visualization techniques for synchrotrons.

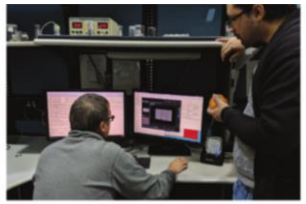
Research Details

- –Remote processing with HPC for real-time high data rate analysis
- –remote data access for high-volume data retrieval
- –Highly interactive design

Xi-CAM plug-ins

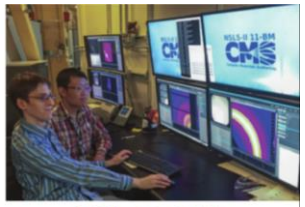
Tomography
Time resolved SAXS
Electron Microscopy
NEXAFS
iPython
Globus Online
Databroker
Remote Computing
GISAXS
HIPGISAXS
GIWAXS simulator
Reverse Monte Carlo
CD-GISAXS
Plug-in Store





APS 2-BM

APS 8-ID-E



BNL 11-BM

BNL SMI Beamline

Xi-cam: A versatile interface for data visualization and analysis

Ronald J. Pandolfi,[a]* Daniel B. Allan,[e] Elke Arenholz,[a] Luis Barroso-Luque,[a] Stuart I. Campbell,[e] Thomas A. Caswell,[e] Austin Blair,[a] Francesco De Carlo,[c] Sean Fackler,[a] Amanda P. Fournier,[b] Guillaume Freychet,[a] Masafumi Fukuto,[e] Doğa Gürsoy,[c,h] Zhang Jiang,[c] Harinarayan Krishnan,[a] Dinesh Kumar,[a] R. Joseph Kline,[g] Ruipeng Li,[e] Christopher Liman,[g] Stefano Marchesini,[a] Apurva Mehta,[b] Alpha T. N'Diaye,[a] Dilworth (Dula) Y. Parkinson,[a] Holden Parks,[a] Lenson A. Pellouchoud,[a] Talita Perciano,[a] Fang Ren,[b] Shreya Sahoo,[a] Joseph Strzalka,[c] Daniel Sunday,[g] Christopher J. Tassone,[a] Daniela Ushizima,[a] Singanallur Venkatakrishnan,[d] Kevin G. Yager,[f] James A. Sethian[a] and Alexander Hexemer[a]

On-the-fly Data Assessment for High-Throughput XRD

Collaboration with C. Tassone (SSRL) and M. Apurva (SSRL)

Scientific Aims

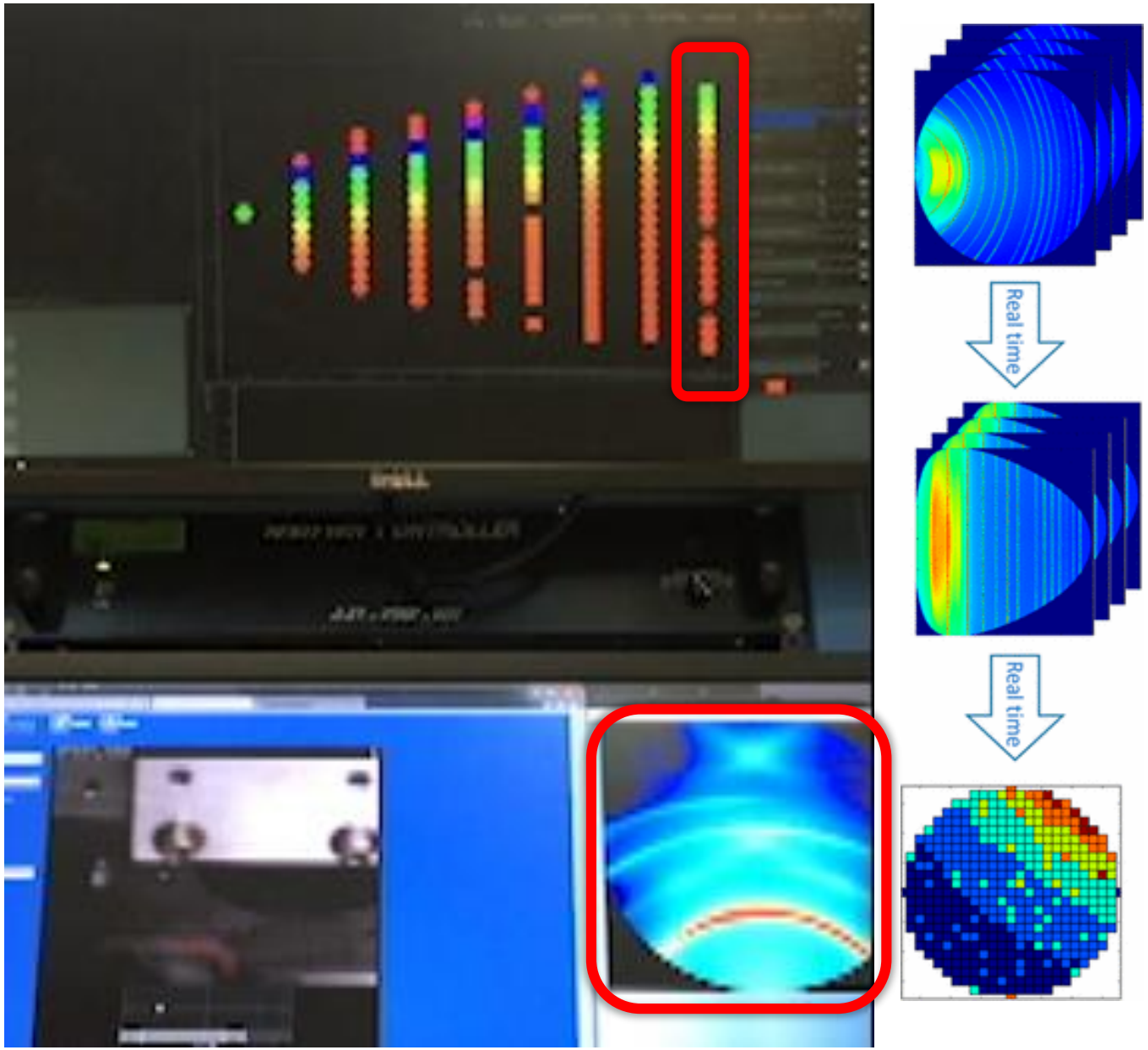
Efficient platform for rapid parallel screening and materials discovery.

Data-driven feedback directly drives material synthesis in-situ.

Significance and Impact

Prototype system at GIWAXS/XRD beamline at SSRL for development.

Material properties and characterizations computed on-the-fly; ~1 second per sample.

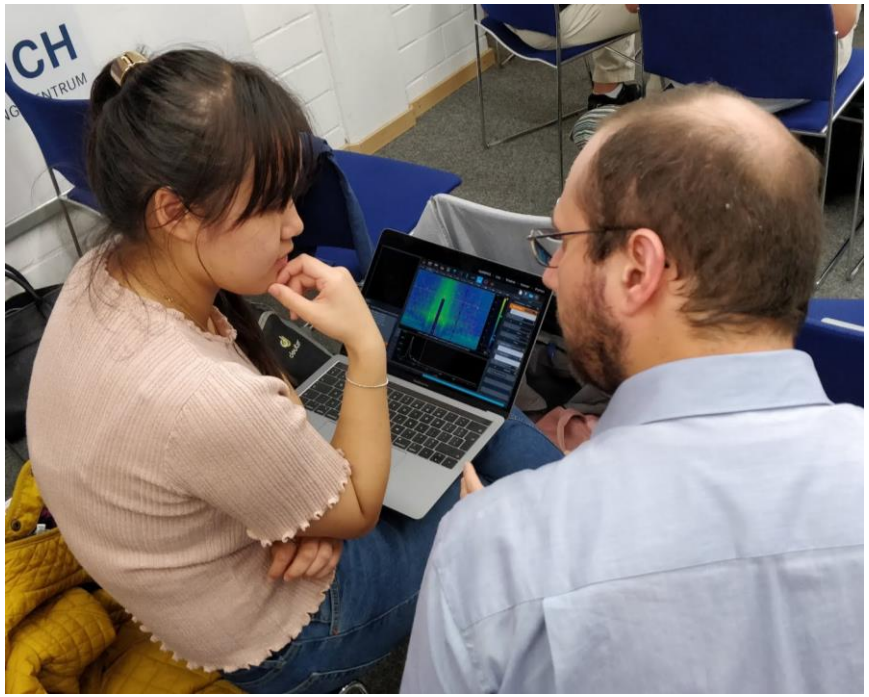


http://pubs.acs.org/doi/pdf/10.1021/acscombsci.7b00015

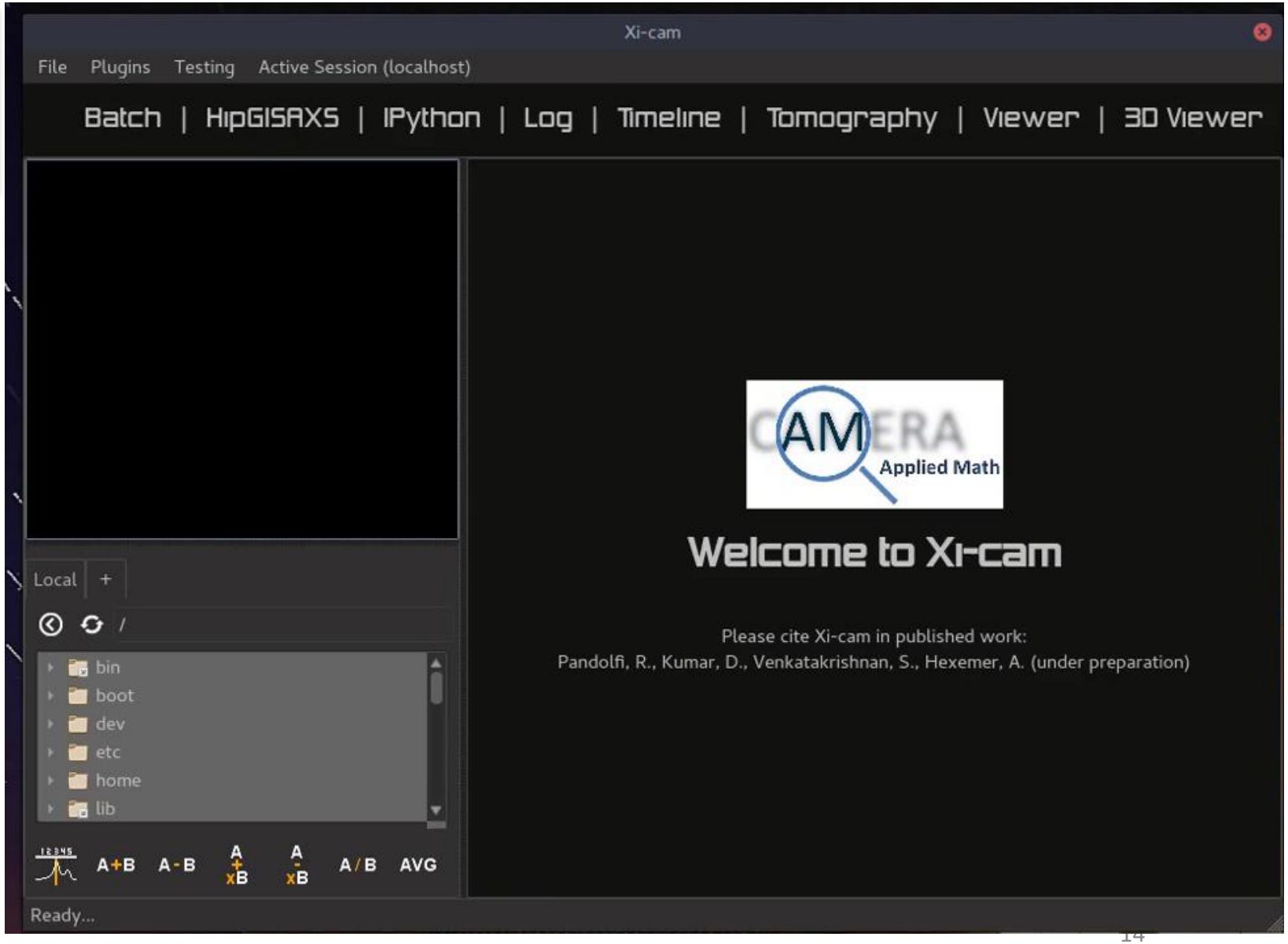
Remote Execution

UNIVERSITÄT BAYREUTH

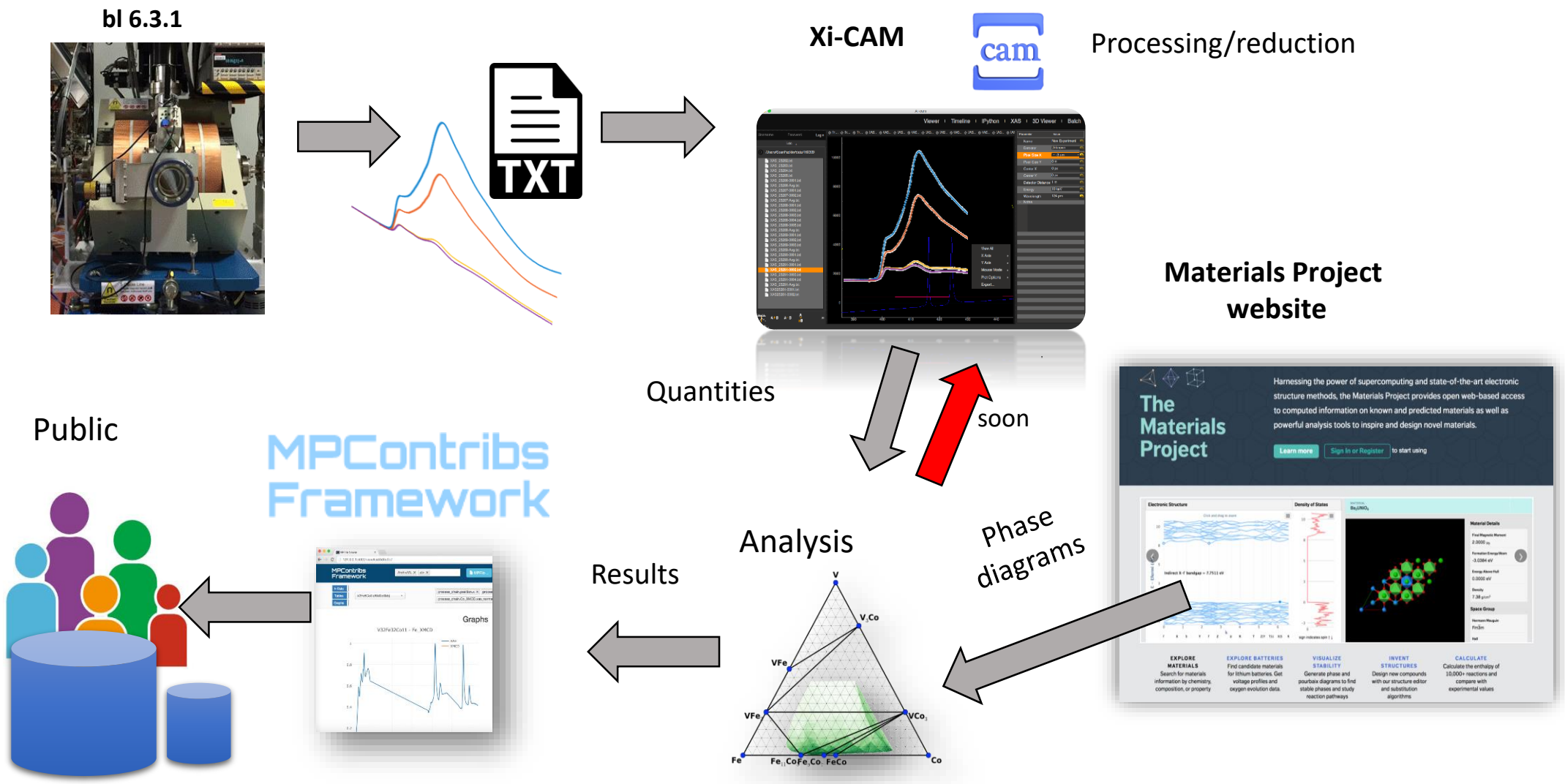
GISAS Summer School 2018



GISAXS Workshop in Bayreuth in collaboration with CAMERA

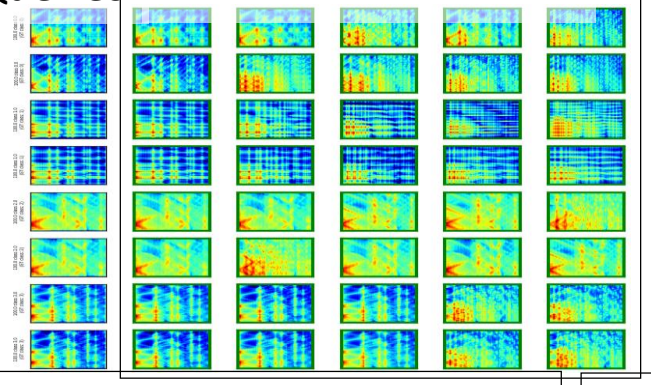


The High-throughput NEXAFS workflow collaboration with Materials Project

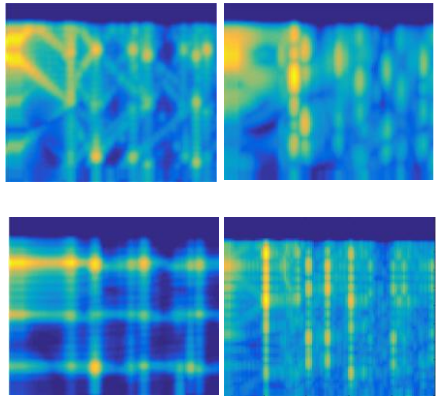


Deep learning for X-ray Scattering

Queries Best matches

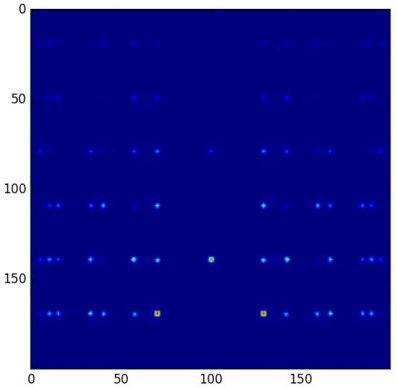


Method	Original Data Set	Noisy Data
Convolutional Neural Networks	92.3%	91.6%
Histogram of Oriented Gradients	92%	80%



Example 1: Classify nanostructures by packing

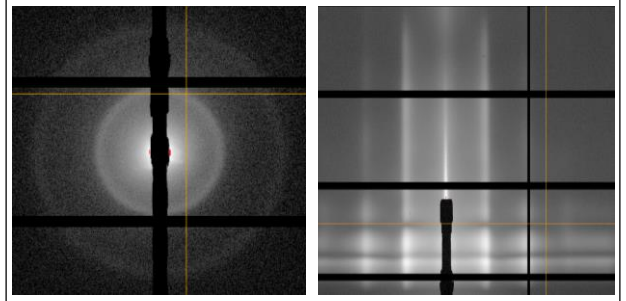
Input : 2D image of cristal diffraction (cubic, hexagonal … packing)



Generate a model which predict with 80% accuracy the different space groups

Example 2: Classify between SAXS/GISAXS data

Automatize the data classification on beamlines (7.3.3 beamline)

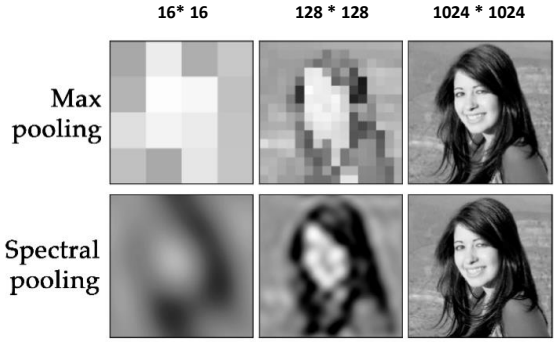


Generate a model which predict with 98% accuracy the different space groups

Collaboration K.Yager and M. Fukuto (NSLS2)

Example 3: Using Fourier transform to improve learning

Truncate frequencies in the Fourrier space to retain more information

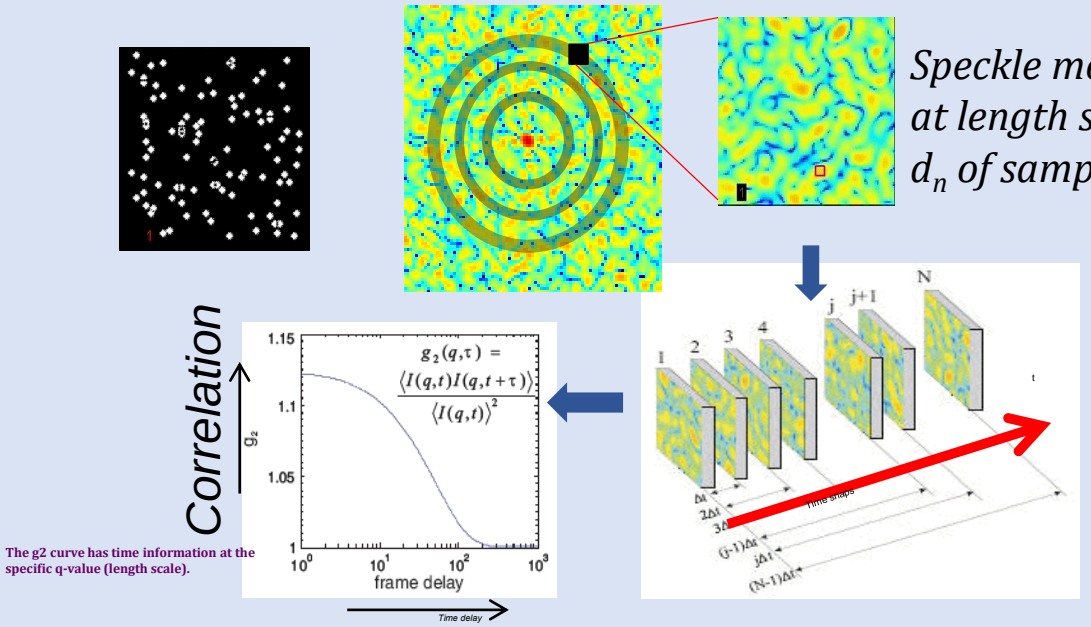


Reduce image size for classification in Fourier space to limit the lost of information

Rippel et al. (2015)

Ushizima, Kumar, Venkatakrishnan, Hexemer. "At the convergence of GISAXS and deep learning", IX CANSAS Conference, Berkeley and Menlo Park, CA, Jun 2017.
Ushizima, "High Throughput Reverse Image Search: Quantification, Search, Retrieval and Ranking for Multi-modal Imaging", BNL NSLS User Meeting, NY, May 2017.

Use Case in Development:
X-ray Photon Correlation Spectroscopy (XPCS)



*M. Sutton, C. R. Physique, **9**, 657 (2008).*
*G. Grübel and F. Zontone, J. Alloys Compd. **362**, 3 (2004).*

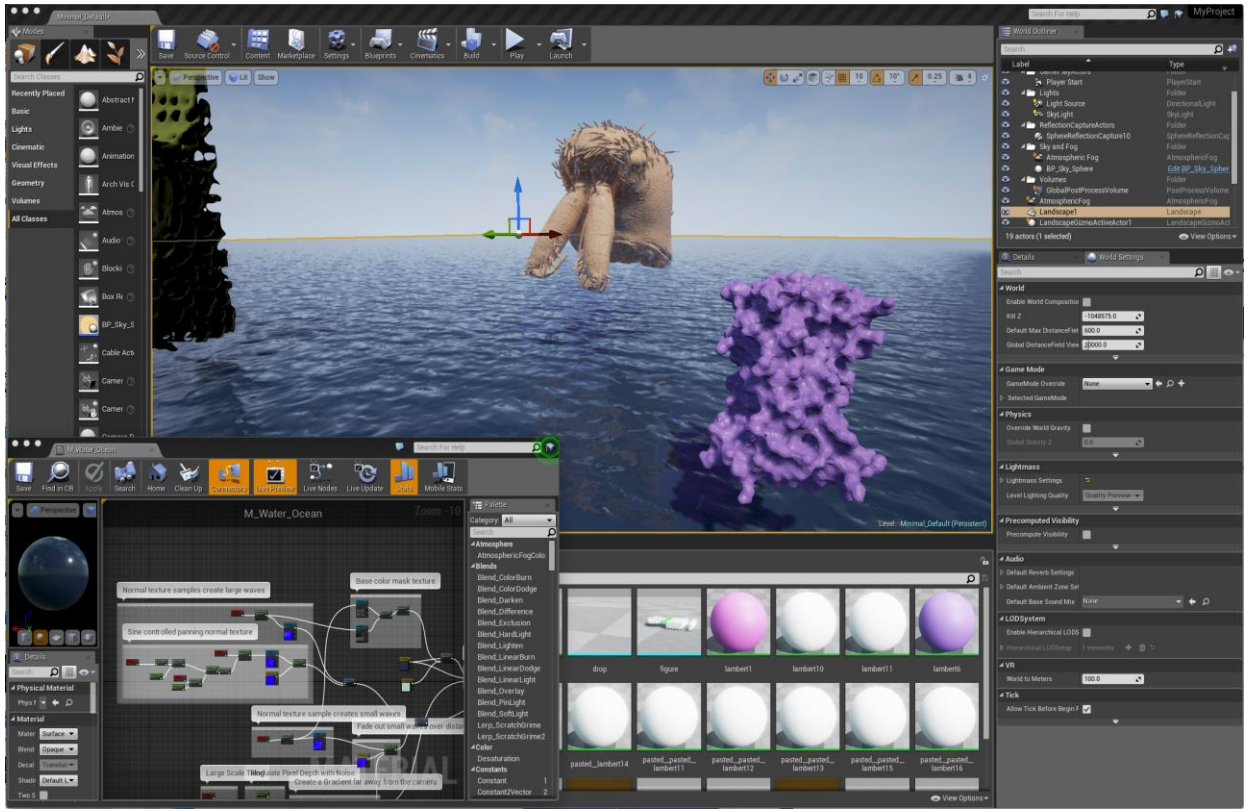
New Beamline development

Interface: Xi-CAM

Control:
Bluesky and DataBroker

HPC Code for XPCS

Future of Creative Discovery



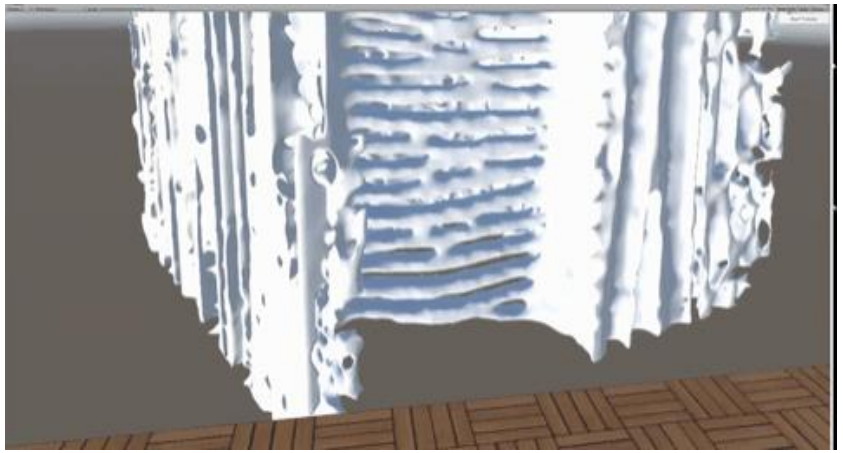
Rick Perry Explores Tomography Data
Spring 2018

KAREN ROSS
Secretary of the California Department of Food and Agriculture

Next Generation of Creative Tools for Analysis

- Visual and Interactive
- Multi modal

- Physics and Math engine by design
- Full integration of ML



Thanks

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov