Metadata, workflows and machine learning

Thomas Proffen

Oak Ridge National Laboratory

tproffen@ornl.gov

Materials research crosses experimental and computing facilities

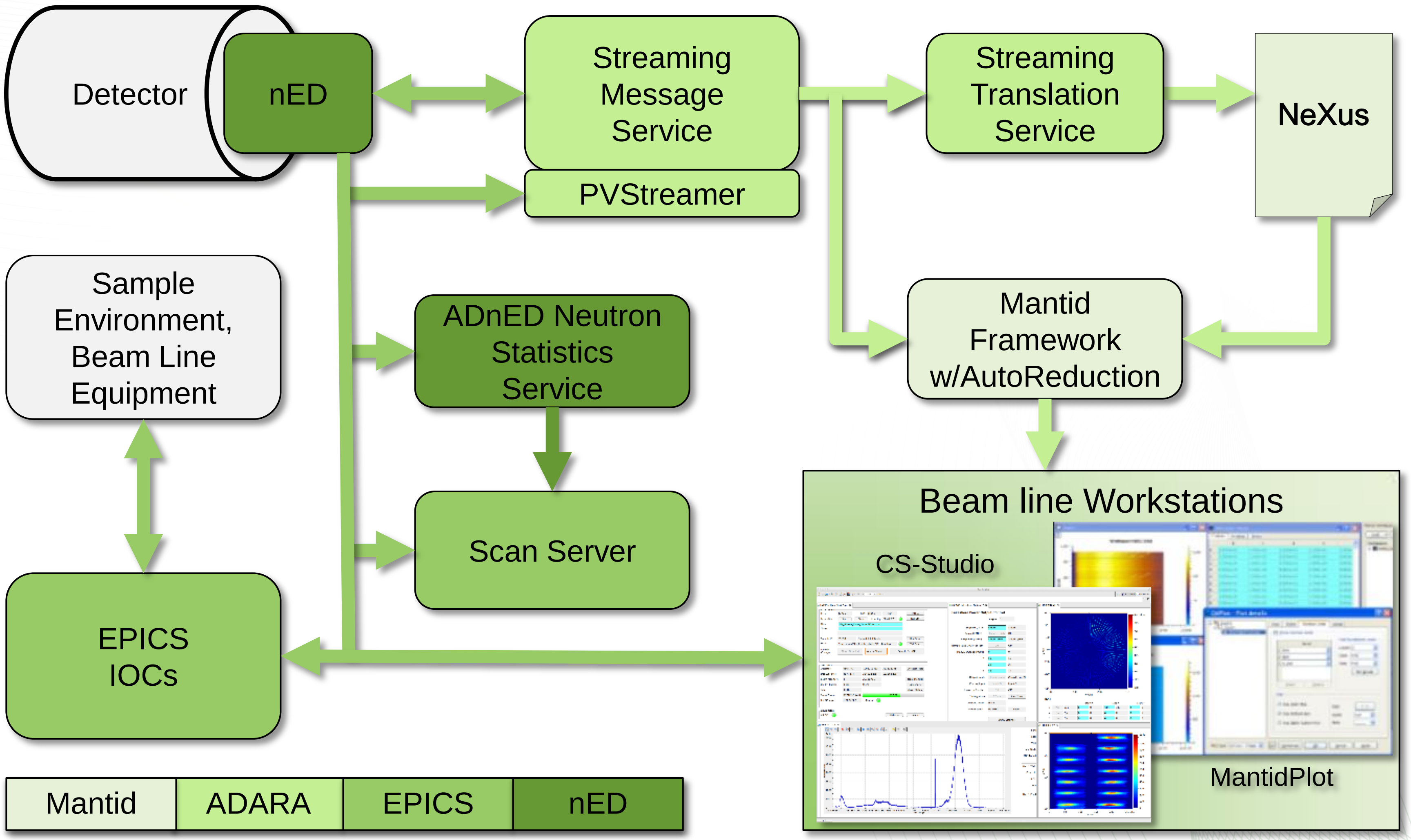
Argonne NATIONAL LABORATORY

OAK RIDGE National Laboratory

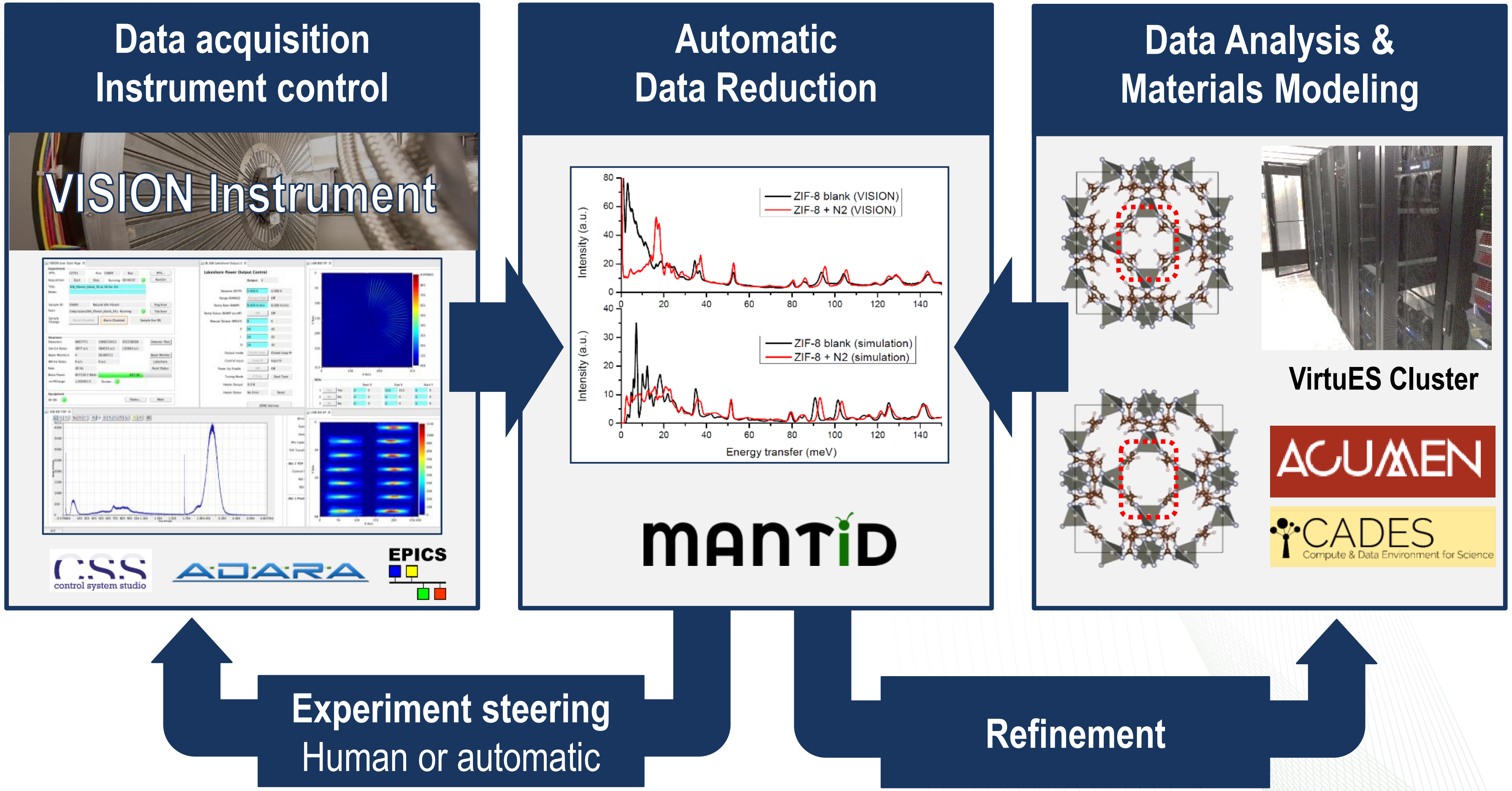
User Facilities

Variety of experiments, topics, methods and 'computer literacy' of users present significant challenge.

Integrating data acquisition, instrument control and data reduction



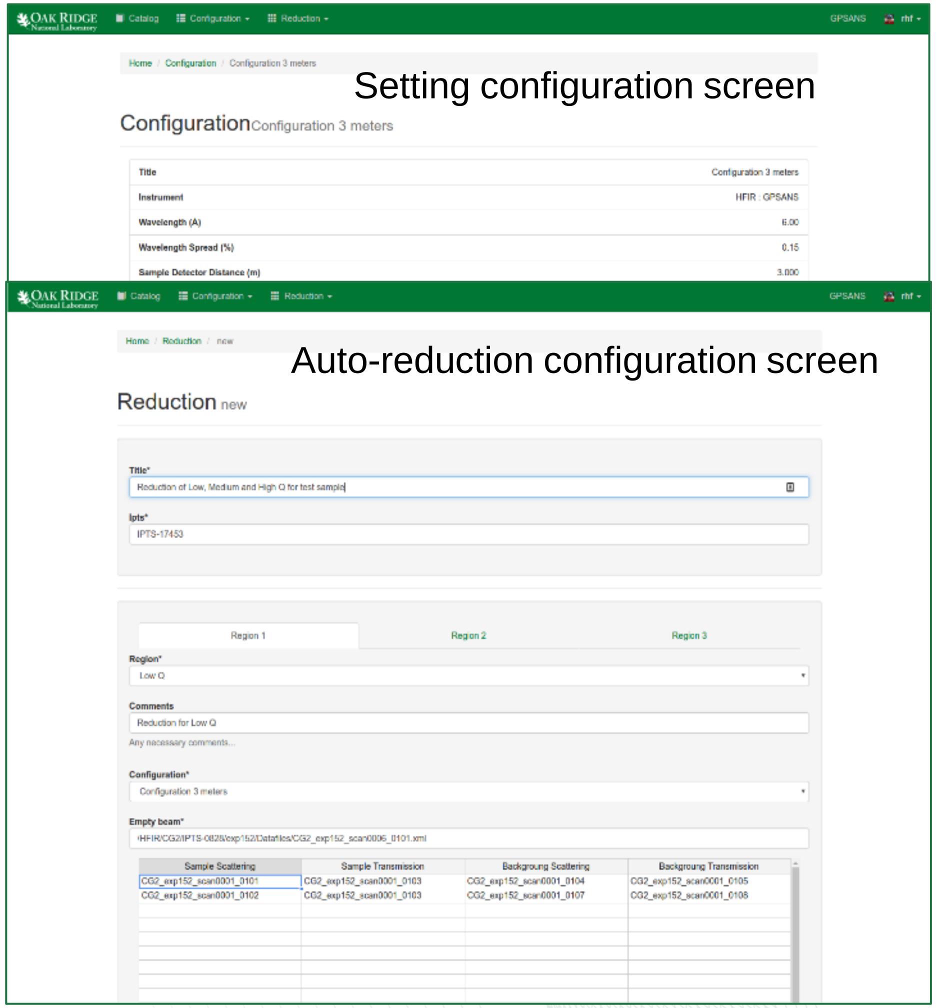
Supporting day-to-day needs from data collection, reduction and analysis to modeling.



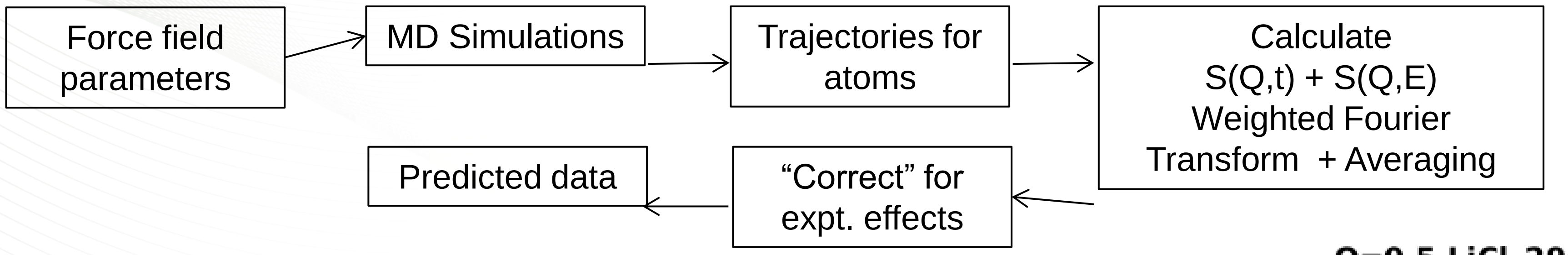
Data published in : M.E. Casco, Y.Q. Cheng, L.L. Daemen, D. Fairén-Jiménez, E.V. Ramos-Fernández, A.J. Ramirez-Cuesta, and J. Silvestre-Albero, Chem. Comm. (2016) 52, 3639

Automatic data reduction for HFIR SANS instruments implemented

- *Mantid* based automatic data reduction was implemented on GP-SANS and Bio-SANS.
- Configuration based setup
  1. Instrument staff sets up a configuration parameters for experiment.
  2. The users completed auto-reduction parameter table.
  3. Mantid script executed data reduction using parameters in table.
- Web based interface.
- Additional work under way to simplify operation by propagating meta data from Data Acquisition System to pre-fill setup parameters.

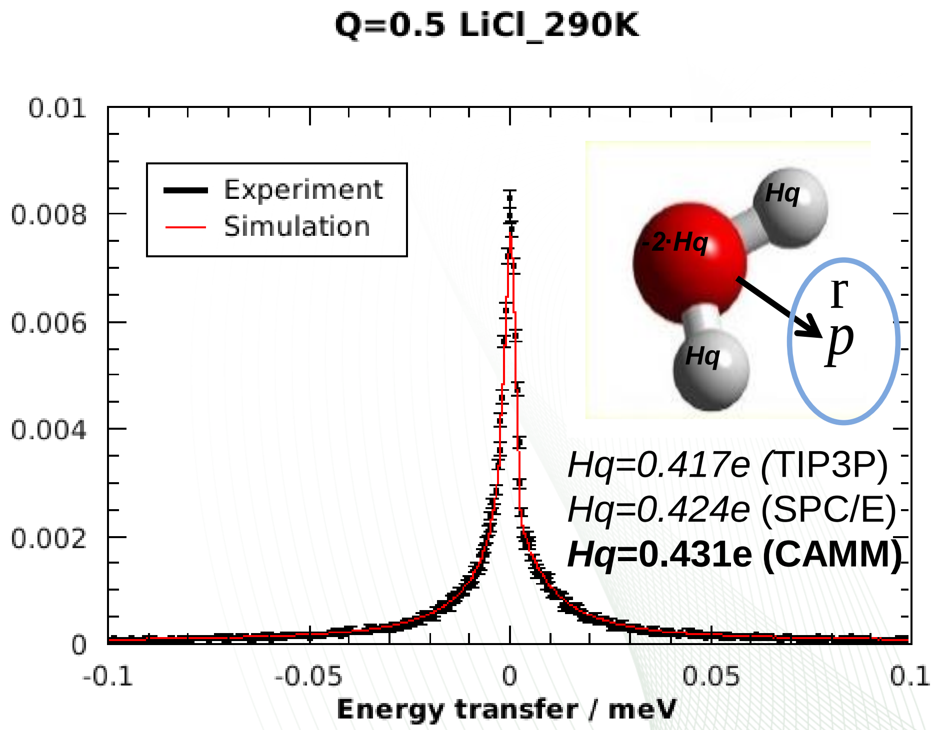


Refining force field parameters from neutron quasi-elastic data



- **First refinement framework test case.**
- High concentrations of LiCl allow studies of bulk water dynamics under 200K. LiCl induces polarization of the water.
- NAMD simulations

http://camm.ornl.gov



Data Catalog

Assessment of the Effectiveness of Data Collection, Reduction, and Analysis

Shelly Ren & Peter Parker
Scientific Information Systems

OAK RIDGE National Laboratory | HIGH FLUX ISOTOPE REACTOR | SPALLATION NEUTRON SOURCE

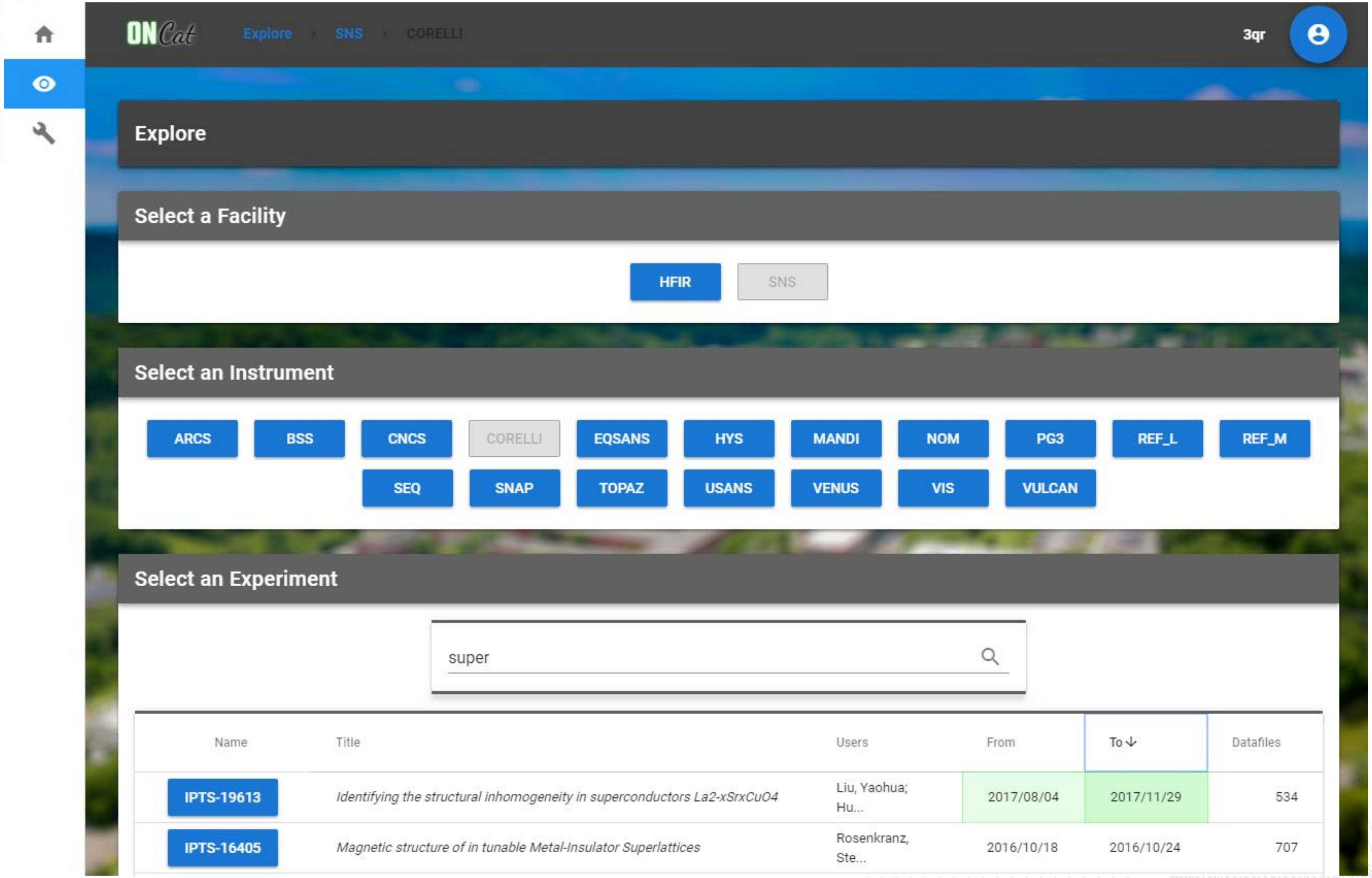
ICAT history and Limitations

- ICAT has been in production to catalog SNS metadata since 2006
- ICAT web services provide metadata to Mantid client, SNS monitor, user portal, and software tools
- Metadata Type has to be predefined and can only have a single type
- Retrieving metadata from datafiles involves a large number of entities in the relational database
- Need to manage fine-grained rules to enable authorization
- Deployment or upgrade is not a trivial task

ONCat Strategy and Development

- MongoDB, a “NoSQL” document store that preserves hierarchical metadata and scales well for our purposes
- Redis for caching and a relational DB for authorization
- Python/Flask to build API
- Vue.js and Vuetify to build ONCat web application
- Docker containers to ensure consistent environments across development, testing, and production.

ONCat Web Application



ONCat Web Application

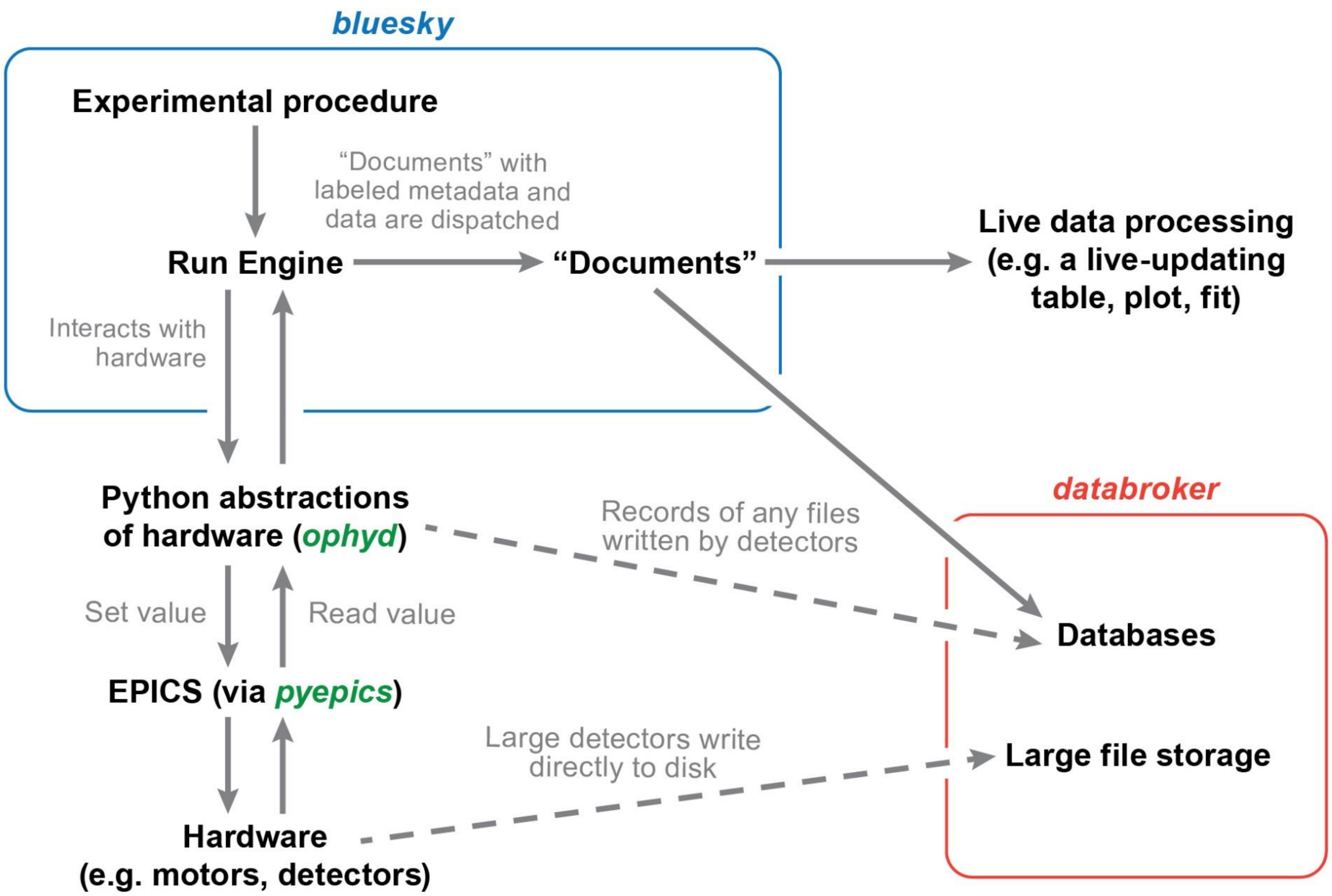

```
{
  "name": "IPTS-19613",
  "id": "IPTS-19613",
  "indexed": [
    "run_number"
  ],
  "exts": [
    ".nxs.h5"
  ],
  "earliest": {
    "modified": "2017-08-09T09:12:57.269000-04:00",
    "ingested": "2017-08-09T09:12:57.269000-04:00",
    "created": "2017-08-04T16:37:59.352000-04:00"
  },
  "tags": [
    "type/raw"
  ],
  "type": "experiment",
  "latest": {
    "modified": "2017-11-29T14:10:57.775000-05:00",
    "ingested": "2017-11-29T14:10:57.775000-05:00",
    "created": "2017-11-29T14:10:46.865000-05:00"
  },
  "size": 534,
  "title": "Identifying the structural inhomogeneity in superconductors La2-xSrxCuO4",
  "users": [
    {
      "name": "Liu, Yaohua",
      "id": "ynl"
    },
```


JSON data from Experiment

NSLS-II
Data Broker concept

Curtesy of Stuart Campbell.

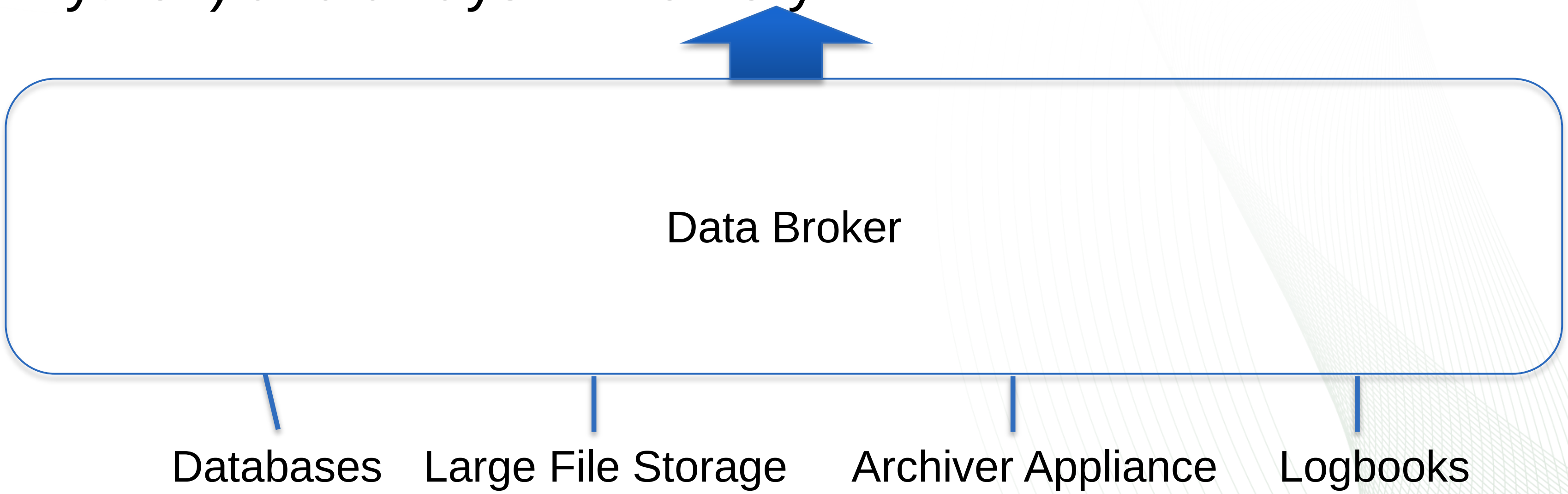
…and how the components work together



http://nsls-ii.github.io/

Data Broker: A Unified Interface to Data

- The databroker keeps I/O concerns separate from scientific code.
- The system is un-opinionated about data formats.
- It provides metadata/data as key-value pairs ("dictionaries" in Python) and arrays in memory.



Opportunities using Machine Learning

AI is about how we use and process data. It will be, and is, transformative in knowledge-based disciplines. AI will not replace scientists, but scientists who use AI will replace those who don’t.*

*Modified from a quote in the Microsoft report, “The Future Computed: Artificial Intelligence And Its Role In Society”

MACHINE LEARNING

A machine learning method takes a bunch of data and "learns" from it!

Data

Label: Not a pig

Label: Pig

Label: Pig

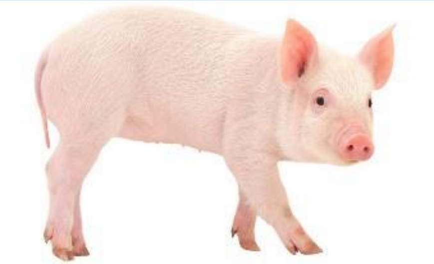
Label: Not a pig

Machine Learning Method

Machine's rules for determining whether or not an image is a pig

DID IT “LEARN” SOMETHING?

Label: Not a pig



Label: Pig

Label: Pig

Label: Not a pig

Training Data

The data we give to the machine learning method to learn from

Label: Not a pig

Label: Pig

Testing Data

The data we hold out and use to check to see if the method actually learned something!

DEEP LEARNING

Simulated scattering 'images'

- Small Angle Scattering
- Diffraction
- Diffuse Scattering
- Quasi Elastic Scattering

Labels

- Relate to model / parameters
- Related to topology
- Good/Bad

Training Data

The data we give to the machine learning method to learn from

Testing Data

The data we hold out and use to check to see if the method actually learned something!

Machine Learning for classification

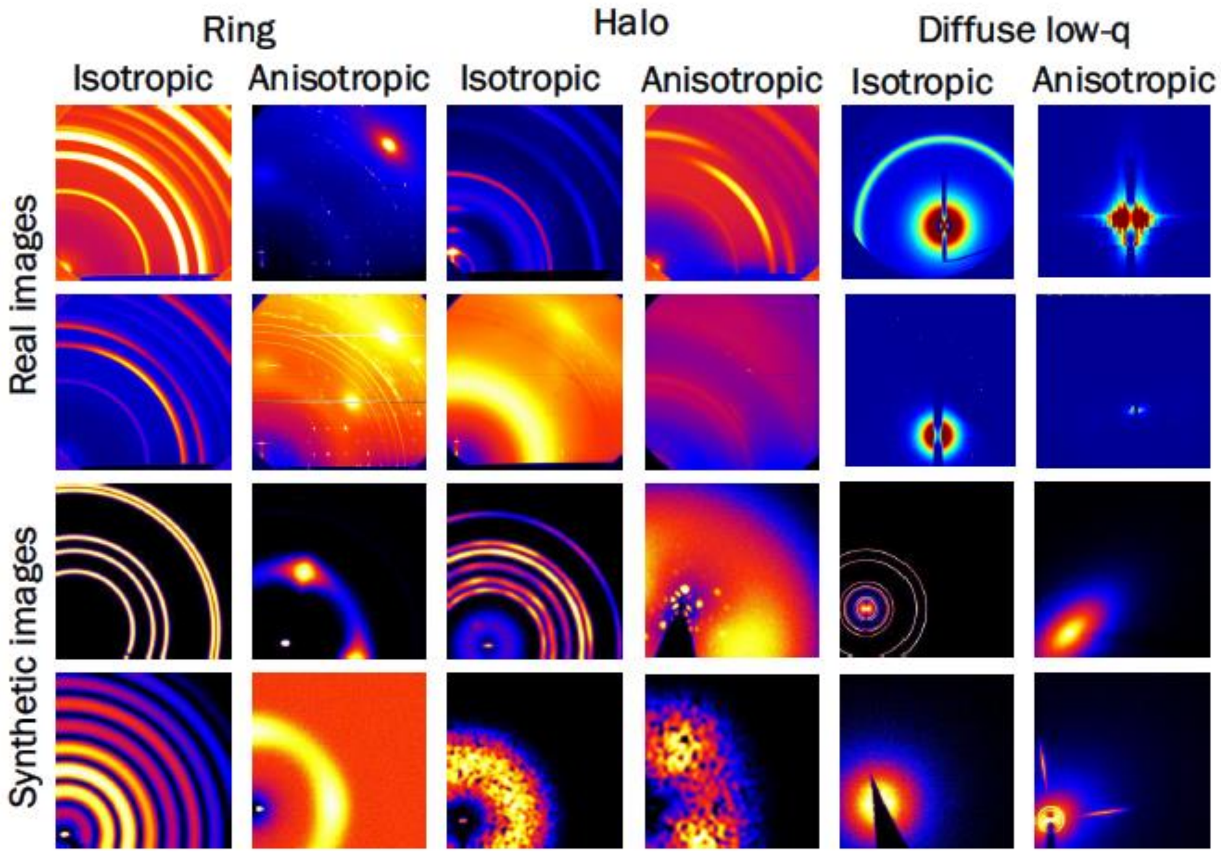


Figure 2: Comparison between synthetic images and real experimental images. The first and second rows are real experimental images, while the third and forth rows are synthetic images. Images in the same column have the same attribute. From left to right, the attributes are: Ring: Isotropic, Ring: Anisotropic, Halo: Isotropic, Halo: Anisotropic, Diffuse low q: Isotropic, and Diffuse low q: Anisotropic. Visually, synthetic and real images are indiscernible.

2017 IEEE Winter Conference on Applications of Computer Vision

X-ray Scattering Image Classification Using Deep Learning

Boyu Wang[1], Kevin Yager[2], Dantong Yu[2], and Minh Hoai[1]
[1]Stony Brook University, Stony Brook, NY, USA
{boywang, minhhoai}@cs.stonybrook.edu
[2]Brookhaven National Laboratory, Upton, NY, USA
{kyager,dtyu}@bnl.gov

Thank you

Thomas Proffen
tproffen@ornl.gov

http://neutrons.ornl.gov