Data Challenges for accelerating scientific discovery at Neutron facilities

Jonathan Taylor

European Spallation Source

Outline

Introduction to scattering facilities

Data Challenges that user facilities face

Where Can ML technology help in accelerating discovery

What are the current ML challenges for scattering science

Neutron and Photon Scattering

Materials and Life Science Research

"Neutrons tell you where the atoms are and what the atoms are doing”

C. Shull & B Brockhouse '94 Nobel laureates

Photon scattering tells you where the electrons are and what they are doing.

The domains are characterised by many different experimental methods.

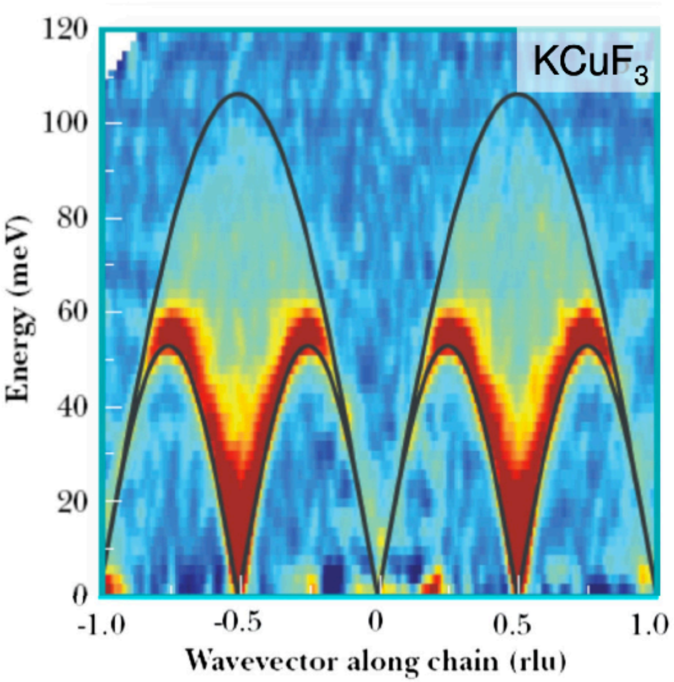
A direct probe of the quantum ground state (and excited states)

Quantum Magnetism

Haldane F D M 1983 Phys. Lett. 93A 464; 1983 Phys. Reu. Lett. 50 1153; 1985 J. Appl. Phys. 57 3359

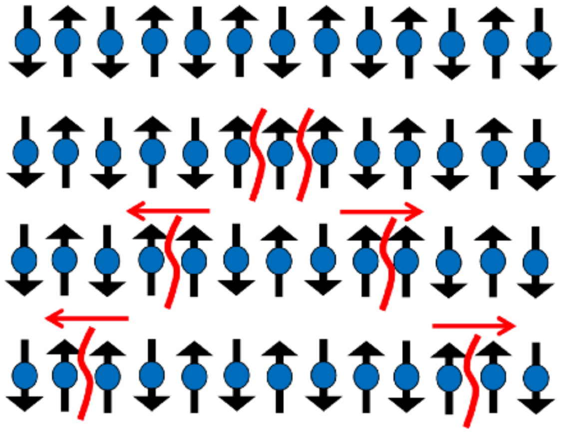
Spin=1/2 Heisenberg antiferromagnetic chain

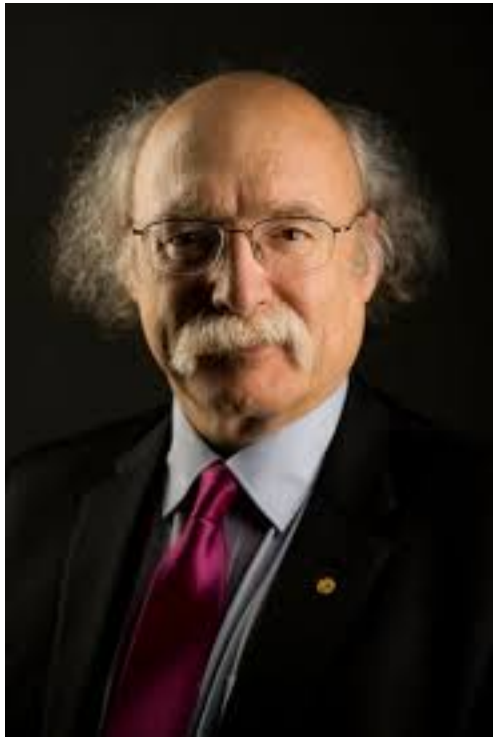
$$H = -J\sum_i \mathbf{S}_i.\mathbf{S}_{i+1}$$



$\varepsilon = \varepsilon(q_1) + \varepsilon(q_2)$

$q = q_1 + q_2$

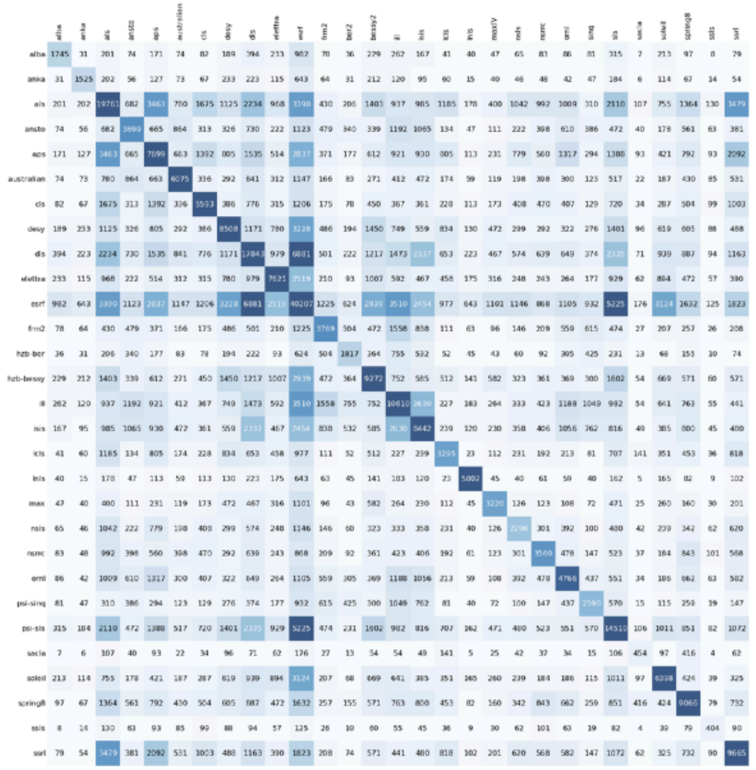




B. Lake, D.A. Tennant, C.D. Frost & S.E. Nagler

Nature Mater. **4,** 329–334 (2005)

European Photon and Neutron Landscape

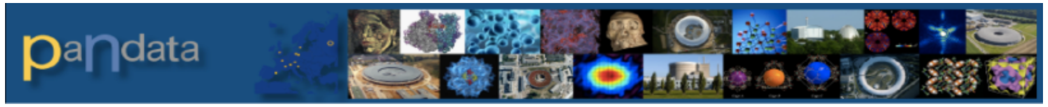


40k users per year

Users visit for short periods 1-3d - one time deal to get a result

10’s PB of data

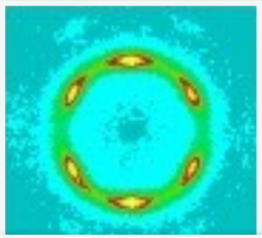
Data rate increase exceeds moore’s law



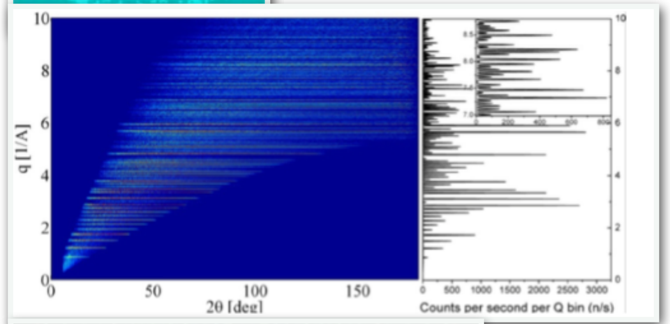
Data Challenges

Variety

SANS

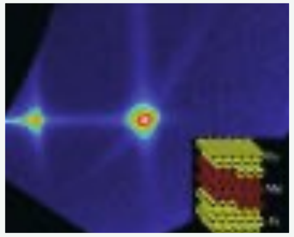


Powder diffraction

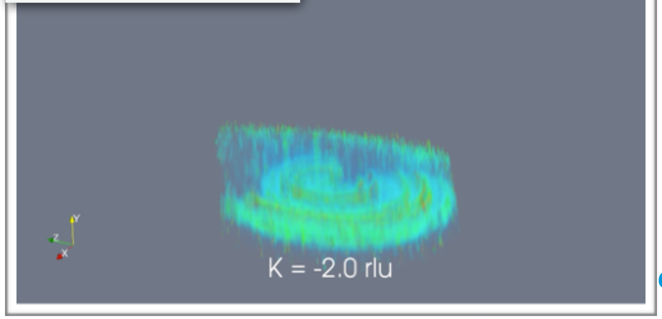


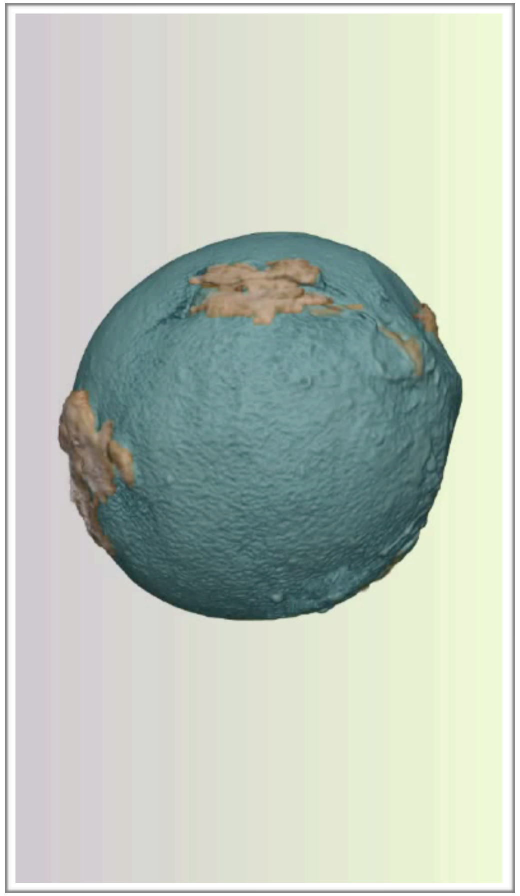
Single crystal diffraction

Reflectometry



Spectroscopy





Imaging

PSI Imaging Group

Data Challenges

Volume

Each experiment visit creates large data volumes

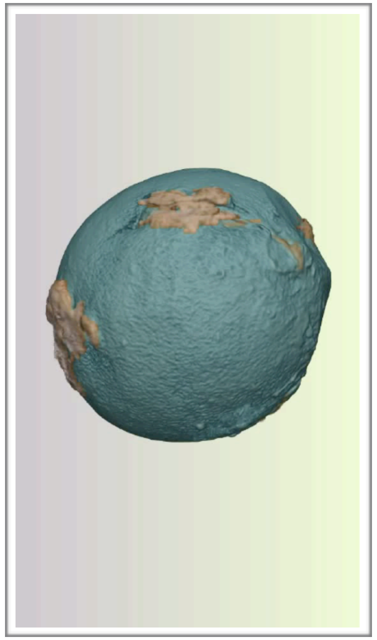
Neutron >5T per visit

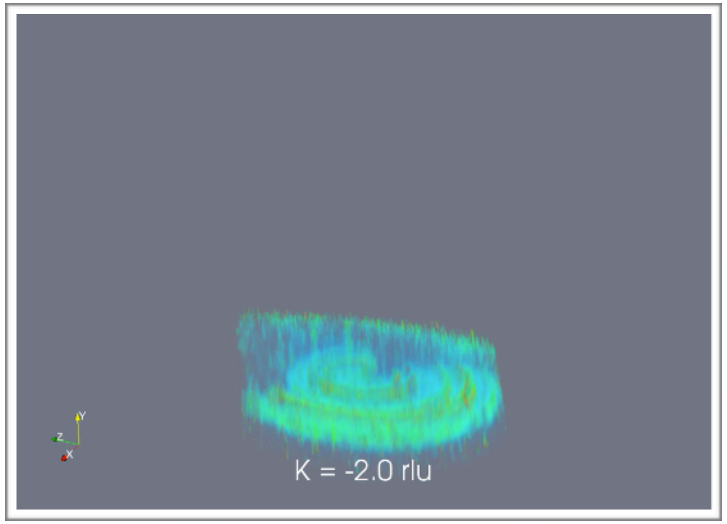
X-Fel >500TB per visit

Data processing becomes a limiting factor.

For data processing many corrections are ‘black box’ algorithms

Artefacts or ‘bad’ data may be included or influence the output





Neutron Data Challenges

Velocity of input data

Neutron detectors convert incident neutrons to charge
(or photons)

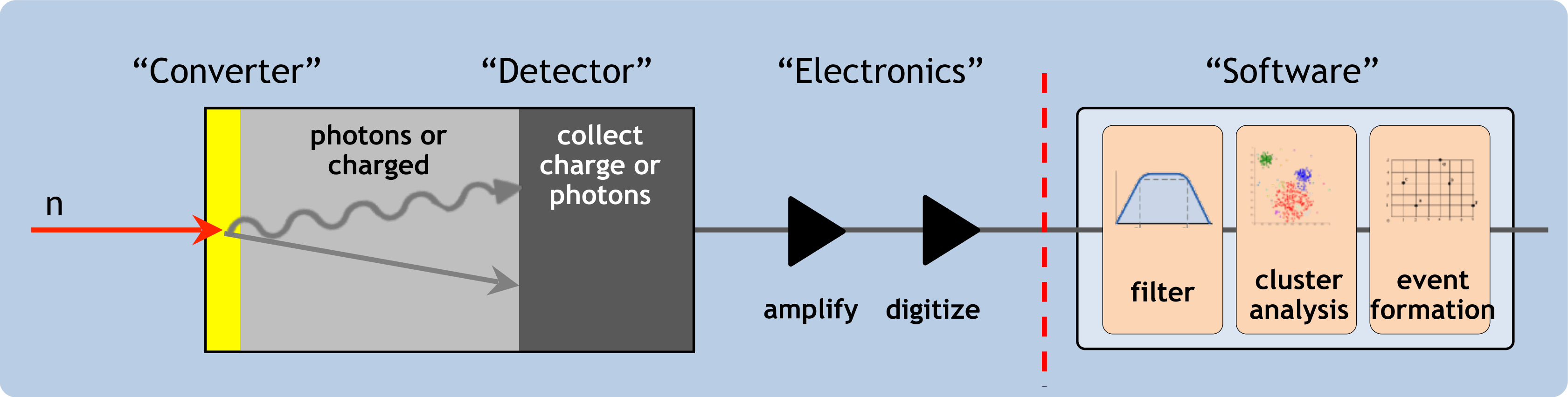
Processing algorithms then determine the spatial location
and Time of neutron arrival.

The input data rate can be as high as (Flux * # readout channels) * ADC bit rate
i.e. > (1e5 * 2) * 12 ~1MHz

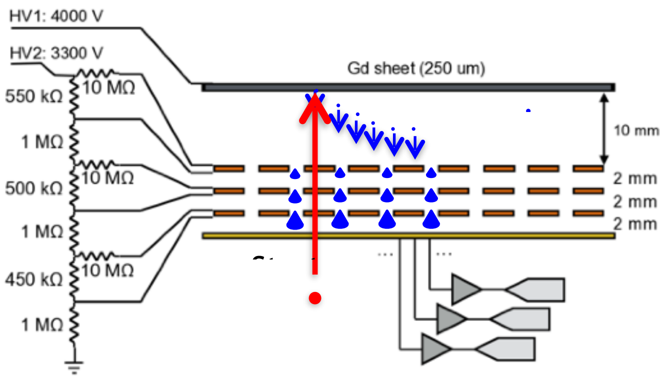
Processing triggers are currently 'simple algorithms'
Processing pipeline must maintain low latency
(for ESS this latency budget is 73msec)

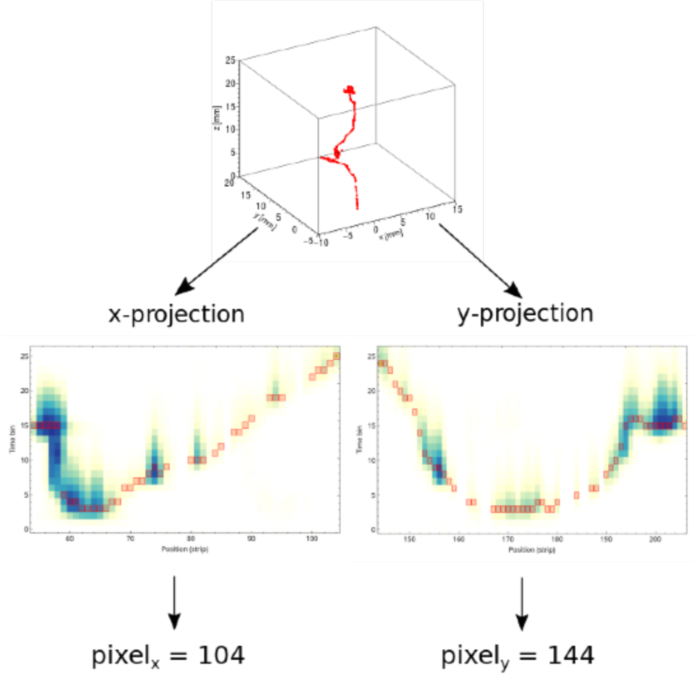
Input Data– Event formation

Particle tracks in x,y,t



There are examples Of ML on the detector backplane in the Photon Community.





Neutron Data Challenges

Velocity of output data

Data are converted into scientifically meaningful units.

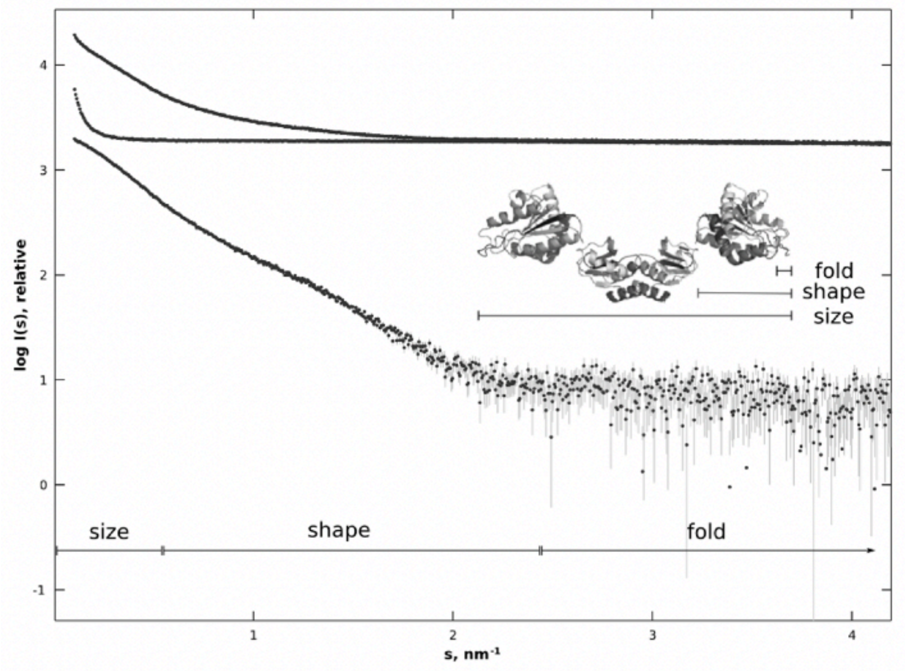
These must then be analysed.

Facilities can generate ~ 500MB /s of processed data

For 1D data like Small Angle Scattering IvQ

1 data set per pulse -14 /s

24 Hours collection ~1M 1D datasets

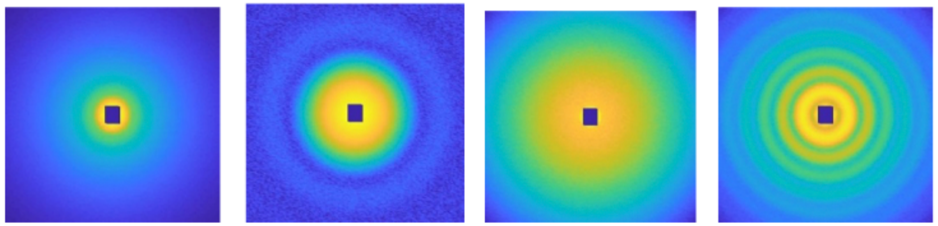


This is too much for a human to process in real time during an experiment

How do we know the experiment is working or collecting useful data

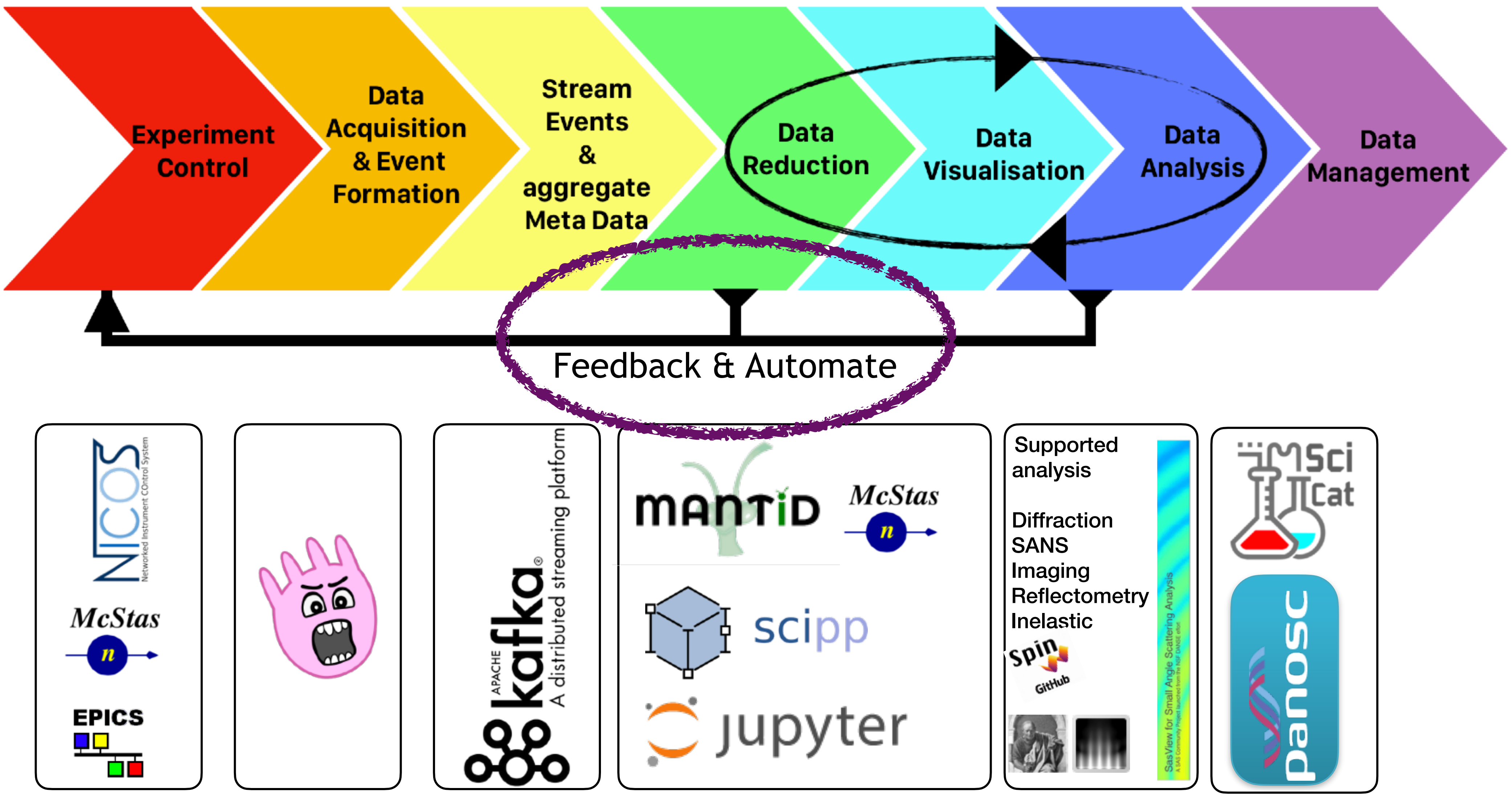
ESS has developed a realtime data pipeline*

* Our system has no intelligent way of automating feedback

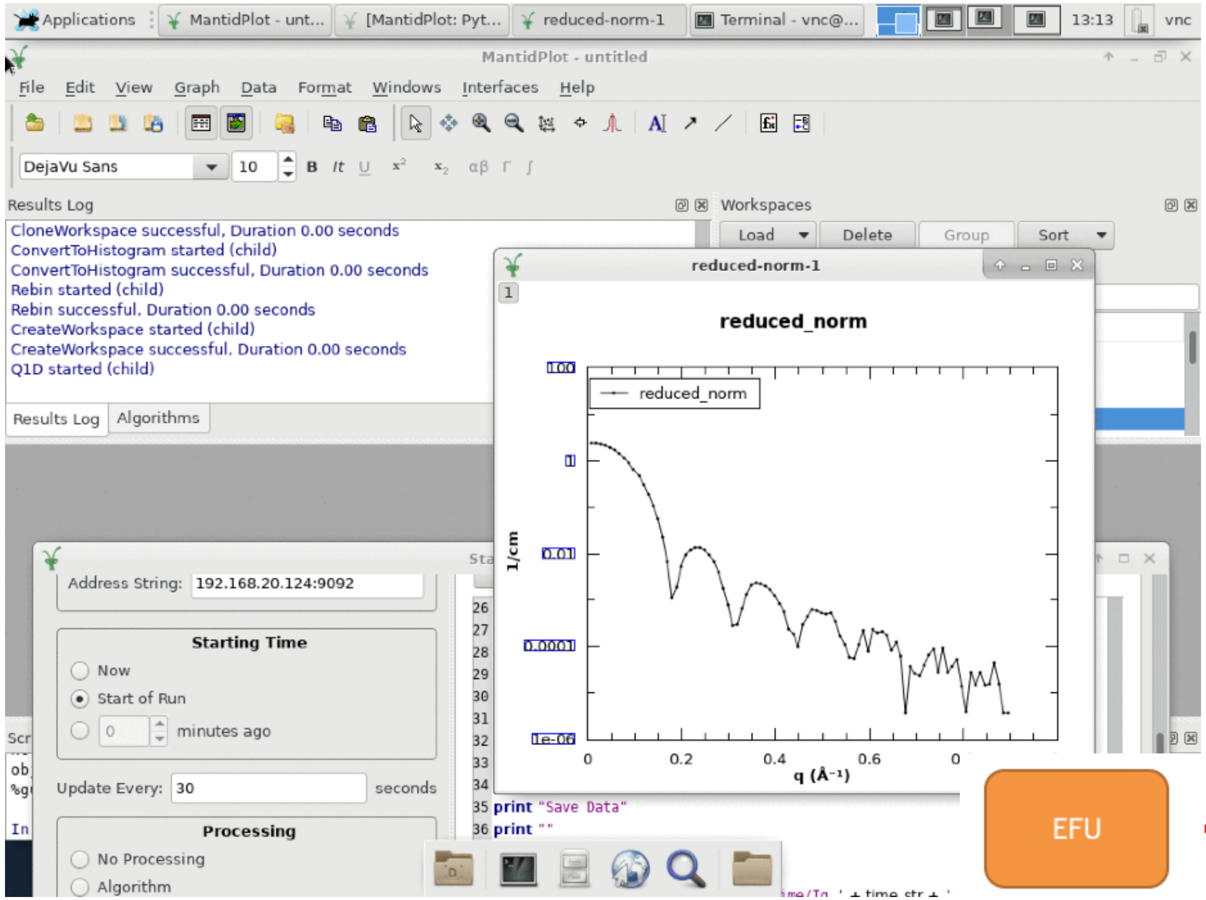


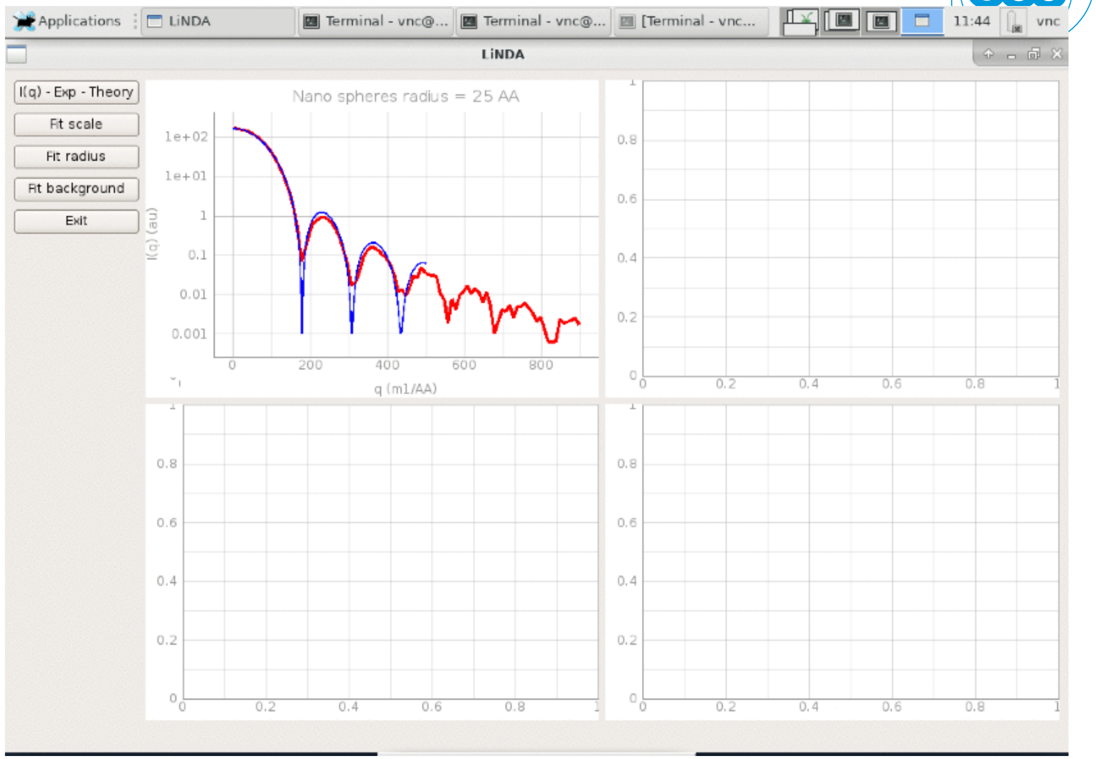
Scientific Computing Pipeline

Pipeline build to provide near realtime processing



Real Time processing of Neutron Event Data for SANS





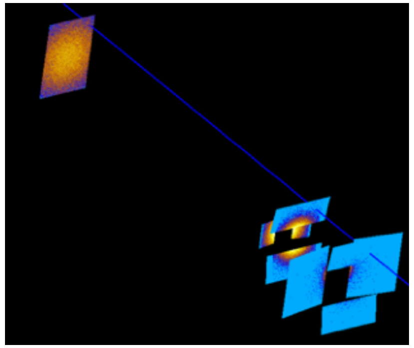
EFU → Kafka → Mantid → SasView → Results

Event Formation

Event Stream to Mantid

Data Reduction Workflows

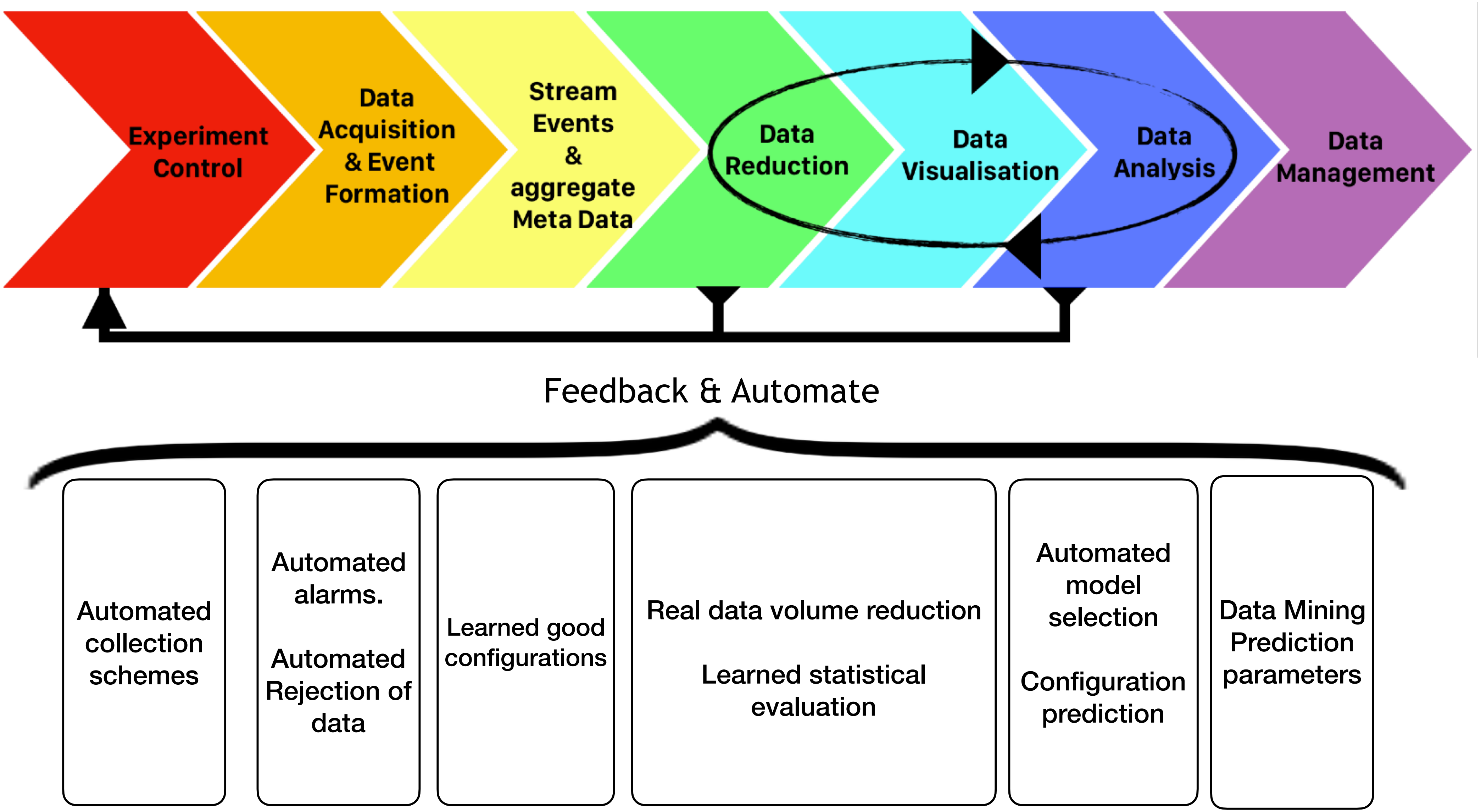
Data Analysis Workflows



Nano Spheres Model
Radius 25 Å

Scientific Computing Pipeline

ML use cases and impact areas



ML challenges for Scattering data.

Classification and Segmentation methods have been successfully applied many types of scattering data.

https://workshops.ill.fr/event/209/overview

There is a lack of Labelled training data.

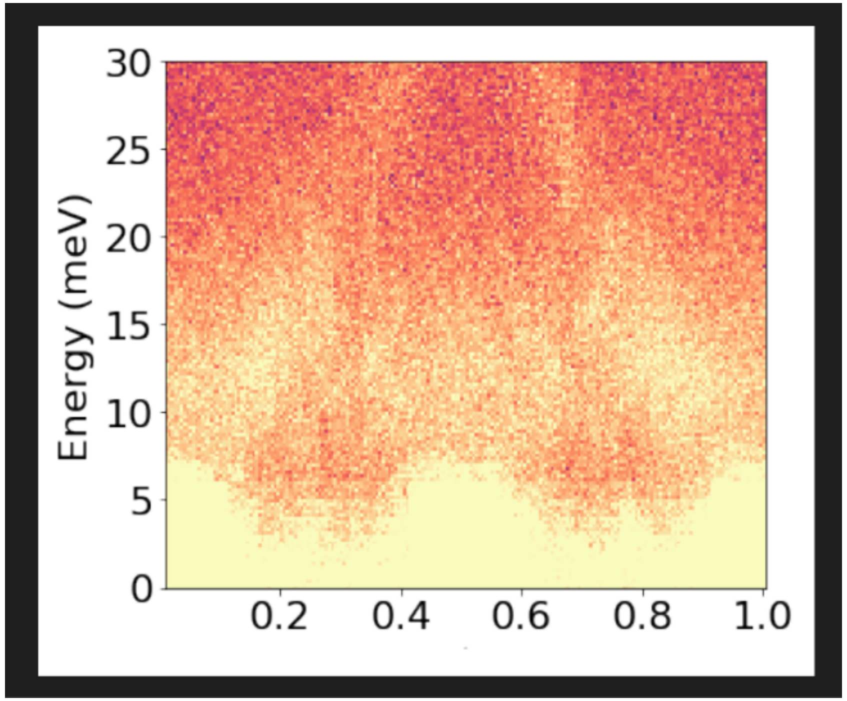
Experimental Noise is an issue.

Experimental backgrounds are problematic.

Simulated data has been used with limited success.

Analytical understanding of the results.

- We need an analytical understanding of the process.
- What is the confidence level on any output.



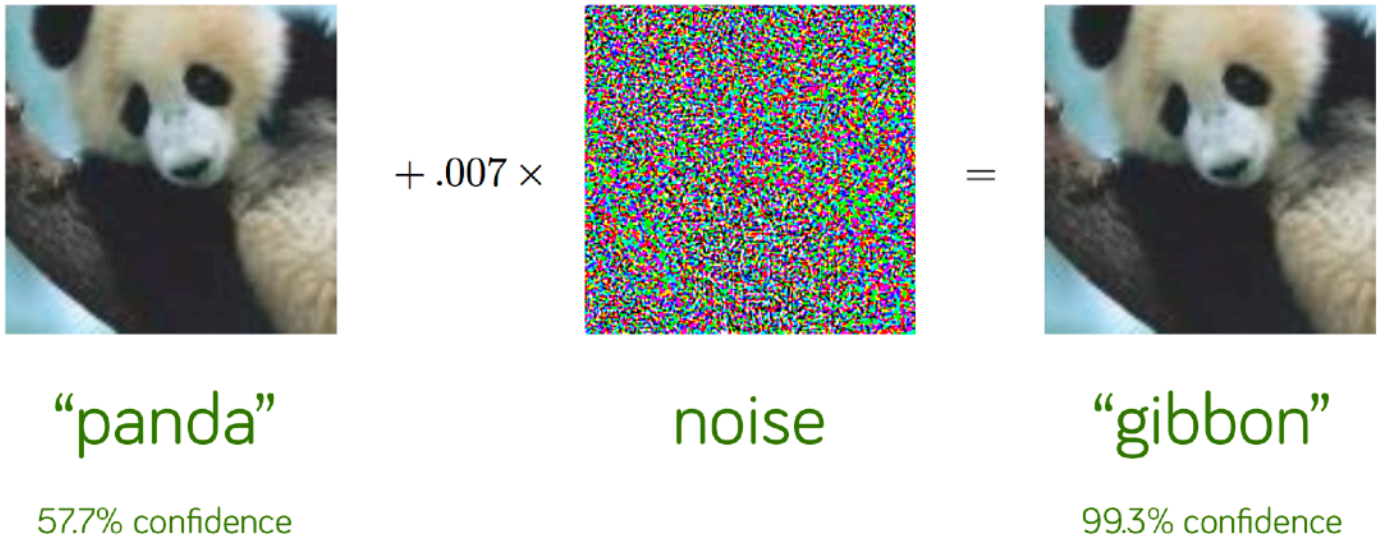
Useful data + noise + background(s)

Noise

Statistical noise is inherent in scattering experiments

The impact of noise on trained systems is well documented

This impacts the use of simulated data as training data



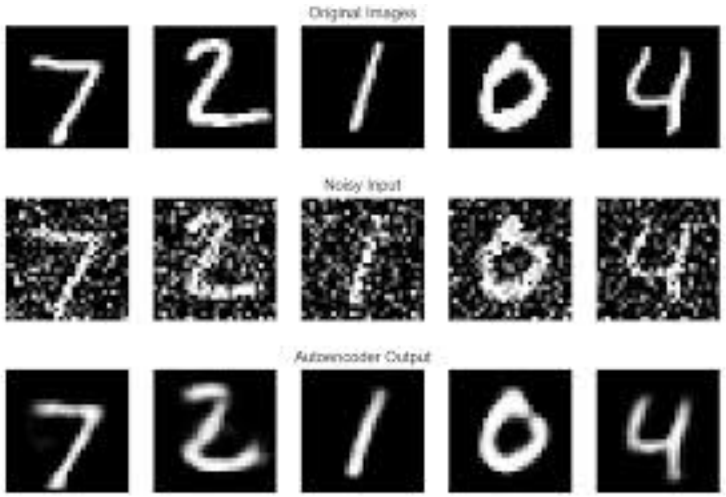
Solutions

De-Noising is a ML use case.

i.e. Noise removal using auto encoders

Trained systems require a ground truth

For the variety of scattering techniques this is a challenge.



Trust & Reliability

In many cases a black box classification is not useful.

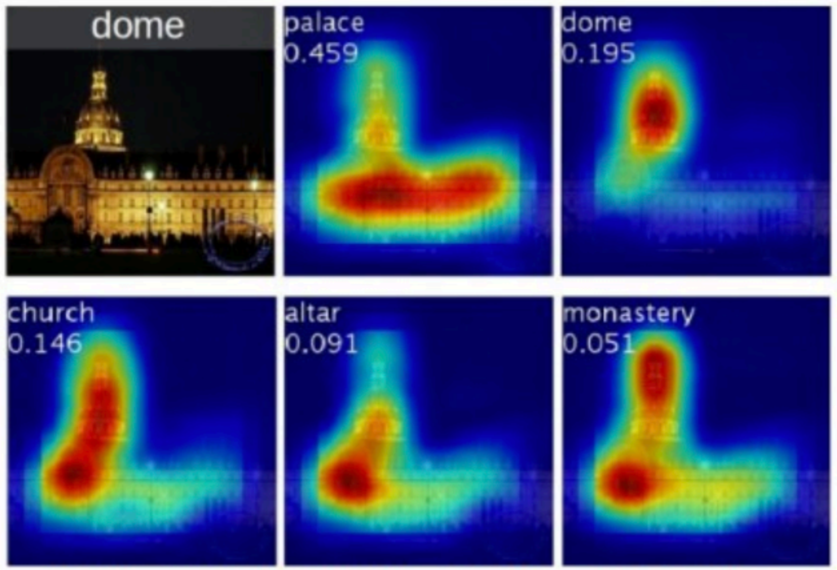
Scientifically we need to know why a classification has been made.

What part of the data influenced the process.

http://cnnlocalization.csail.mit.edu/

How reliable is a classification

- How do we decide good enough



Class activation maps of top 5 predictions

ML tech is code and code needs to be tested to some level of QA

Materials discovery

Data Mining

Mining Existing measurements from Open Data or the literature

- Requires excellent data management
- Examine trends & correlations in parameter space
- Predict future experiments

Mining databases from atomistic calculations.

- Atomisic codes can calculate specific properties of materials
- ML can then be used predict new materials with specific properties
- https://materialsproject.org/
- Organic Materials Data base https://omdb.mathub.io/

Future work

Standardised training data for the photon and neutron domain.

- Essential for future progress.

Benchmarks for performance and reliability.

New methods to add to the existing tools available.

- Focusing on methods that require less or no training data.

Summary

Can ML accelerate scientific discovery at Scattering facilities

- Is the only technology that can provide a good level of autonomous control and feed back.

- We can create too much data for a single human to meaningfully interpret on the time scale of an experiment.

- The outputs and choices need to have traceability and variance.

-It is as important to know why as to know what.