ada lovelace centre

UKT0 workshop

13th March 2018

David Corney. Director, Scientific Computing Dept. STFC

ALC steering group

- Philip King (Chair) (ISIS)
- Sean Langridge (ISIS)
- Dave Clarke (CLF)
- Cristina Hernandez Gomez (CLF)
- Rob Akers (CCFE)
- Martin O'Brien (CCFE)
- Dave Stuart (DLS)
- Laurent Chapon (DLS)
- Gordon Brown (SCD)
- Barbara Montanari (SCD)
- Robin Pinning (Hartree)
- Roger Downing (Hartree)
- Tony Hey (STFC Chief Data Scientist)

Terms of Reference for ALC Steering committee. Draft 14-Dec-2017

To agree the purpose and aims of the ALC.
To agree and define the work program for the ALC, and any associated priorities, including a work program based on increased budget plans.
To oversee and where necessary guide the outcomes of the ALC activities.
Concentrate resources on delivering what is uniquely needed and deliverable on the campus for facilities and their users, supporting/facilitating the use of external resources/services where they meet requirements.
Define a model for the ALC that will best support the needs of all facilities, for infrastructure, data curation and data analysis.
Define key areas of expertise to be developed and how they will help complement the skillset already existing at facilities for software projects and setup a collaborative model for future projects.
To raise the profile and understanding of the aims and purpose of the ALC within STFC, CCFE, BEIS and UKRI.
To continue to make the case for ALC within STFC, government, and UKRI
To seek out, share and coordinate all relevant funding requests for ALC activities

To provide oversight of ALC projects and usage of funds. More specifically:

a) To ensure alignment of ALC activities to the needs of the facilities (ISIS, Diamond, CLF, CCFE)
b) In this phase of ALC, to provide a coherent and substantial set of projects which fall under the ALC remit ('badged' as ALC) as demonstrators of the potential of ALC to significantly impact on facility data analysis and modelling.
c) To ensure the work of ALC covers the spectrum of facility needs – from basic infrastructure supporting facility data analysis and modelling such as DAaaS, and development of modelling codes and their integration with facility data, through to capability building in areas such as big data and machine learning
d) To ensure these projects demonstrate the added value of working across facilities in a joined-up way
e) To prioritise ALC funds as they become available (for new/and/or existing projects)
f) To ensure that ALC projects are run as projects, with definite aims, timescales and outputs.
g) To receive periodic updates on ALC projects
h) To advise on projects priorities and governance of these projects

Would like to better understand what this means.

Ada Lovelace Centre

- Significantly enhance the capability to support the Facilities’ science programme

  - ***Theme 1: Capacity in advanced software development for data analysis and interpretation***
    - Middleware and visualisation
    - Code parallelisation and adaptation

  - ***Theme 2: A new generation of data experts and software developers, and science domain experts***
    - Provide research and development of computational codes and algorithms, with science domain experts
    - Training in HPC skills and code development

  - ***Theme 3: Compute infrastructure, for managing, analysing and simulating the data generated by the facilities and for designing next generation Big-Science experiments.***
    - Access to HPC resource including Hartree
    - Access to significantly enhanced cluster at RAL for “routine” data analysis
    - Management, integration and access to data

- ***Focused on the science drivers and computational needs of facilities.***

ALC core concept

- “Industrialising” simulation, modelling and analysis tool set for Facilities users
- Easy access to data, analysis tools and computation on site or remotely
- Improving productivity of investment in STFC facilities
- Releasing the bottleneck on science delivery from the national facilities
- Delivering necessary user training

Projects

- Data Analysis as a Service
  - Bringing data, software & compute together through the cloud
- ULTRA
  - High throughput HPC platform for tomographic image analysis
- Octopus
  - ICAT, Job Portal and associated infrastructure for CLF

Projects

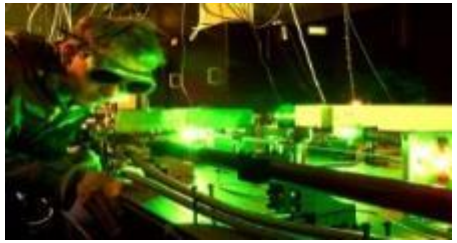
- Complex Molecules
- Least Squares Fitting for Mantid
- Modelling and simulation for the interpretation of muon experiments
- Materials Workbench
- Magnetic excitations in transition metals
- Proper Analysis of Coherent Excitations (PACE)

Services

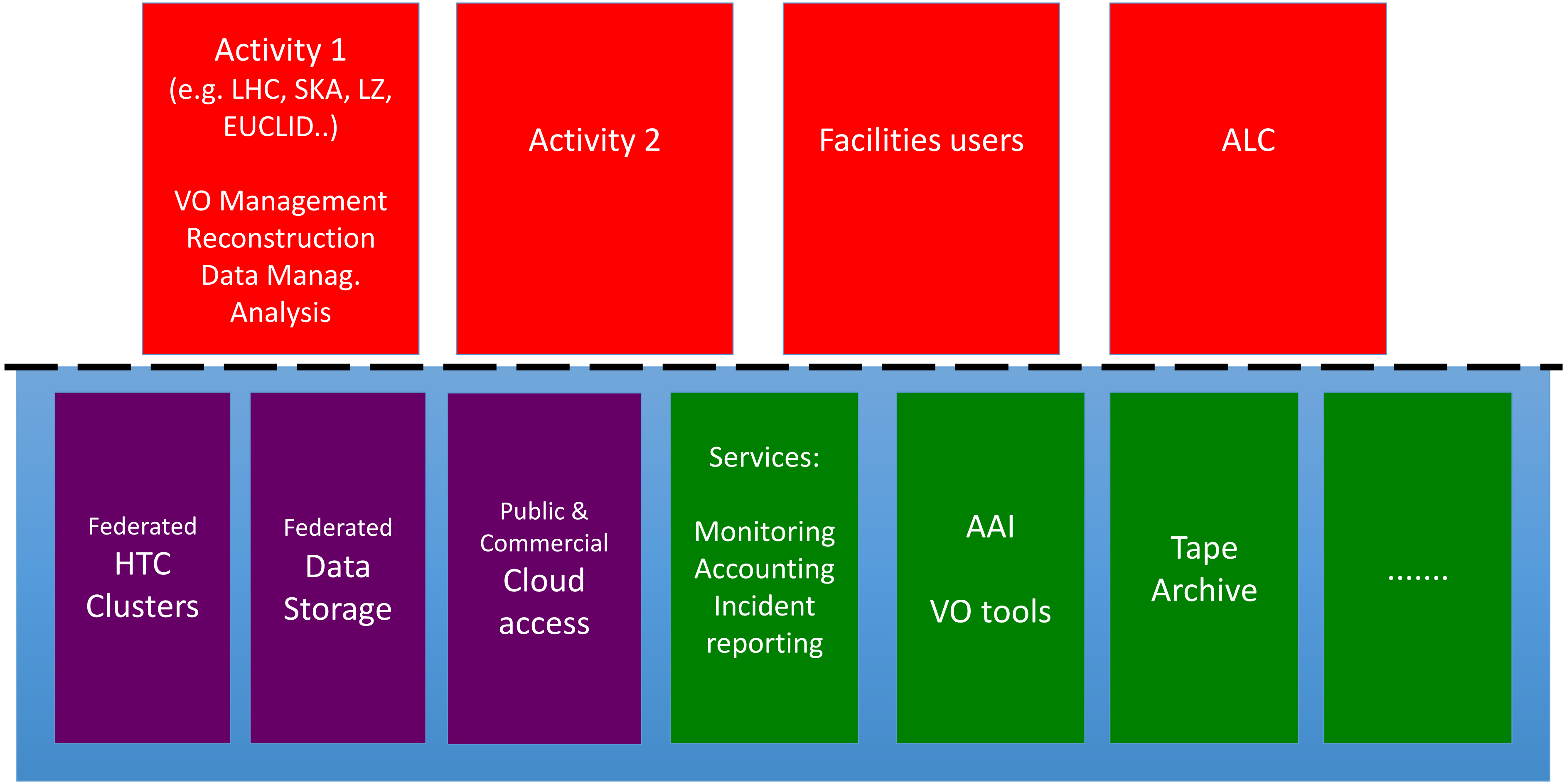
- SCARF
  - HPC cluster
- ICAT/TopCat (ECAT for CLF)
  - Data catalogue & interface
- eData
  - Repository for scientific data behind publications
- Cloud
  - OpenStack provision
- Database Services
  - Underpinning numerous production systems
- CEPH / CASTOR / Tape
  - Storage & Archive

Supporting Facilities Data Management

- STFC Scientific Computing Department
  - Support three STFC Funded facilities on the RAL campus
  - Provide data archiving and management tools
- ISIS Neutron and Muon Source
  - Provide tools to support ISIS's data workflows
  - Support through the science lifecycle
  - Rich metadata
  - Provide Data archiving
- DLS Synchrotron Light Source
  - Data Archiving
  - Limited metadata in archive
  - Managing the scale of the archive
- Central Laser Facility
  - Real-time data management and feedback to users
  - Rich metadata on laser configuration
  - Access to data



Science Domains remain "sovereign" where appropriate



Share in common where it makes sense to do so

Note: this does NOT necessarily mean centralised

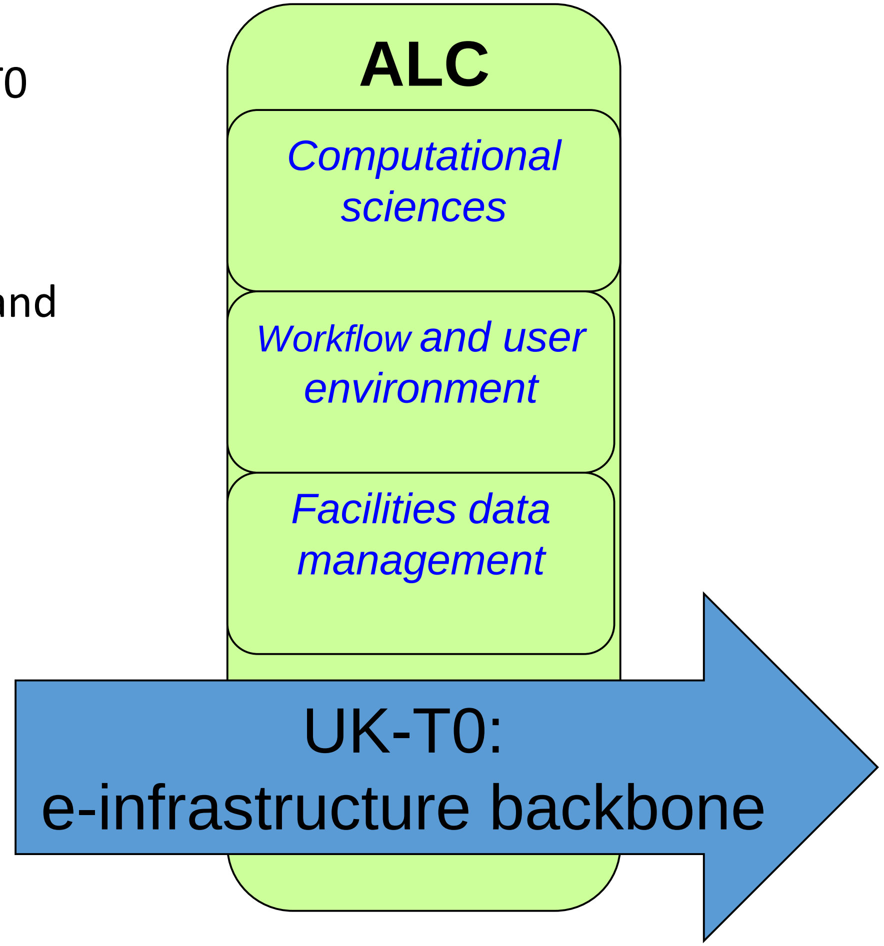
ALC and UK-T0 complementary

- ALC would pool infrastructure platform with UK-T0
  - HTC, HPC, networks, data storage and archive
  - Cloud provisioning, resource brokering, AAI

- ALC would provide the facilities specific services and training

- ALC Data Science Services
  - Metadata services and data access
  - Workflow and user environments
  - Provenance and data publication
  - Visualisation

- ALC Computational Science services
  - Computational facilities science research
  - Modelling and simulation algorithms
  - Data analysis algorithms
  - Code optimisation and parallelisation



STFC e-Infrastructure Strategy

Table 3 – The Ada Lovelace Centre

ALC- indicative costs £m	2018/19	2019/20	2020/21	2021/22	2022/23	
Unfunded CPU	2.70	2.70	2.70	2.70	2.70	
Unfunded Disk	0.50	0.70	0.70	0.90	0.30	
Unfunded Tape	0.14	0.10	0.10	0.20	0.10	
Other infrastructure	0.40	0.60	0.70	0.80	0.90	
Operations and middleware staff	0.80	0.90	1.20	1.20	1.25	
Software engineering	0.85	2.40	2.75	3.15	3.40	
Licencing, maintenance	0.50	0.50	0.50	0.50	0.50	
Power and cooling	0.30	1.00	1.20	1.50	1.50	
Total	6.19	8.90	9.85	10.95	10.65	46.54

Ada Lovelace Computer Programmer, Mathematician (1815–1852)

- The daughter of famed poet Lord Byron, Augusta Ada Byron, Countess of Lovelace—better known as "Ada Lovelace"—was born in London on December 10, 1815. Ada showed her gift for mathematics at an early age. She translated an article on an invention by Charles Babbage, and added her own comments. Because she introduced many computer concepts, Ada is considered the first computer programmer. Ada died on November 27, 1852.

ALC: Why

- Data rates are increasing, facilities science more data intensive
  - Handling and processing data has become a bottleneck to produce science
  - Need to compare with complex models and simulations to interpret the data

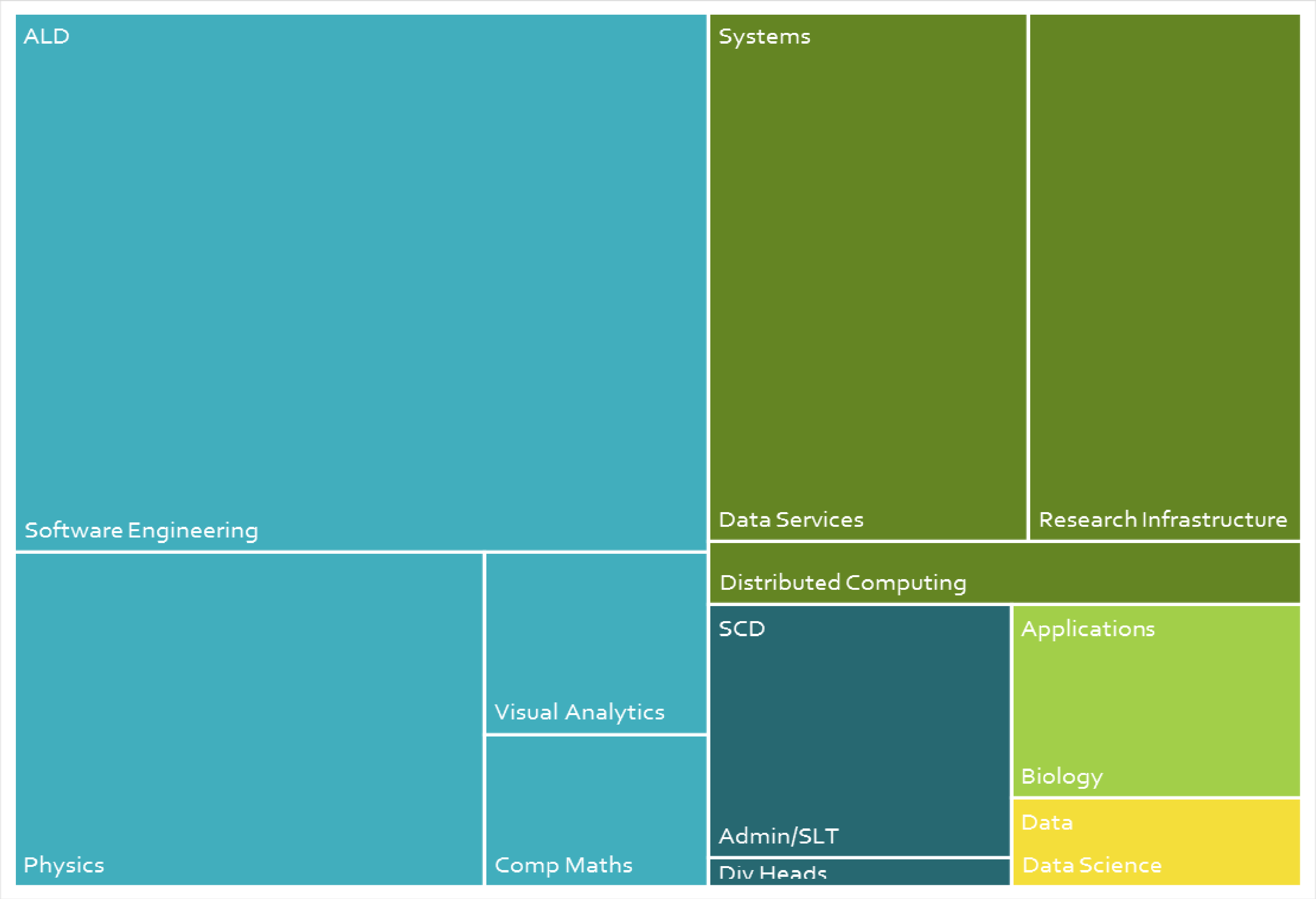
- Computing provision at home-institution highly variable
  - Consistent access to HTC/HPC to process and interpret experimental data
  - Computational algorithms more specialised
  - More users without the facilities science background

- Access to data, compute and software services
  - Allow more timely processing of data
  - Use of HPC routine not "tour de force"
  - Generate more and better science

- Provide within the facilities infrastructure
  - Remote access to common provision
  - Higher level of support within the centre
  - Core expertise in the computational science
  - More efficient than distributing computing resources to around individual facilities and research groups.

- Others in the world are also developing this super-facility capability
  - Lawrence Berkeley Laboratories
  - Oakridge National Laboratories
  - PanDaas Proposal expressed the desire to do this in Europe
    - Now in recent CALYPSO+ (X-Rays) and NMI3 (Neutrons) INFRAIA-01-2016 proposals

Actual effort on Facilities

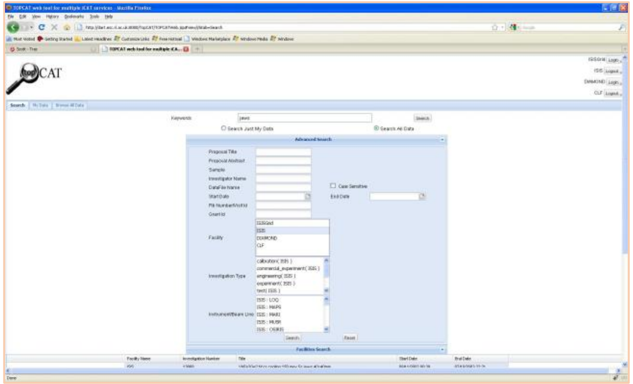


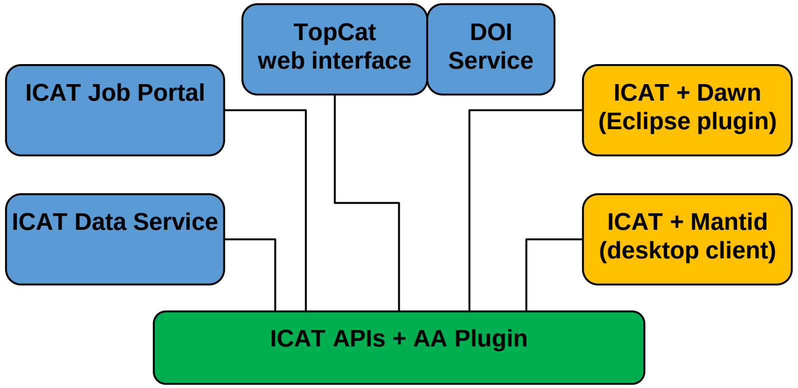
Allocation

Activity	FTE
Facilities Data Services (inc Datastore, CASTOR, Databases)	5.0
Machine Room Support	1.9
Software Engineering (ICAT, TopCat, Data Management, Repositories)	4.0
RIG Infrastructure, Service & Fabric	3.0
SCARF HPC Cluster	2.0
Cloud Development & Support	1.3
ULTRA	0.7
Applications Support	0.6
Graduates/Apprentices	1.0
Services Coordinator	0.7
Total	~21 FTE

Data Management tools and processes

- **Integrated data management pipelines for data handling**
  - From data acquisition to storage
- **A Catalogue of Experimental Data**
  - ICAT Tool Suite: Metadata as Middleware
  - Automated metadata capture
  - Integrated with the User Office system
- **Providing access to the user**
  - TopCat web front end
  - Integrated into Analysis frameworks
    - Mantid for Neutrons, DAWN for X-Rays
- **In daily production use**
  - CLF, ISIS, DLS
- **Also internationally**
  - ESRF, ILL,.SNS, …
  - PaNData Consortium





ICAT