The Revolution in Experimental and Observational Science

The Convergence of Data-Intensive and Compute-Intensive Infrastructure

Professor Tony Hey

Chief Data Scientist

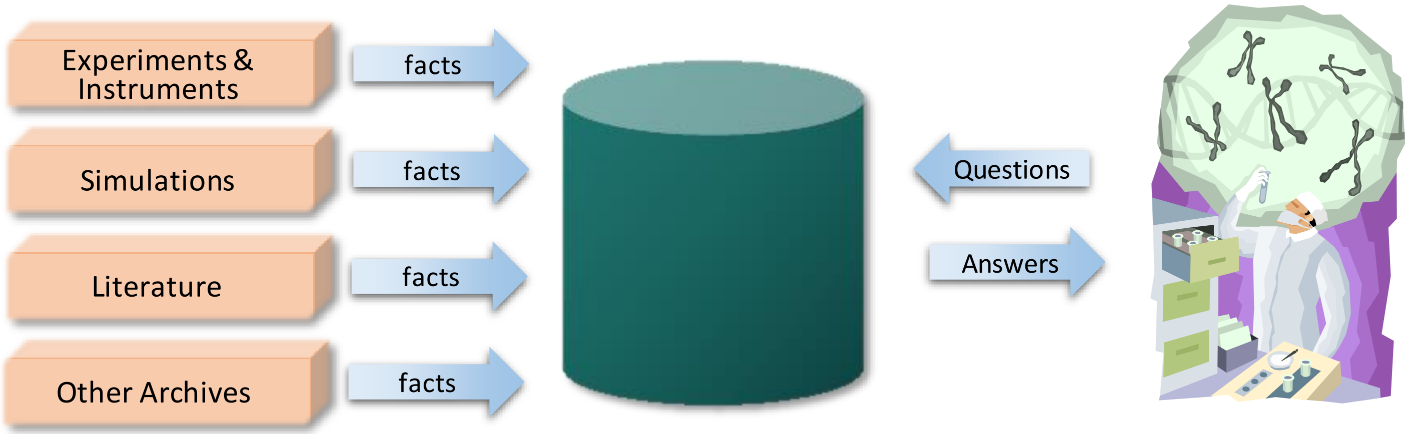
STFC

tony.hey@stfc.ac.uk

The Background

X-Info

- The evolution of X-Info and Comp-X for each discipline X
- How to codify and represent our knowledge



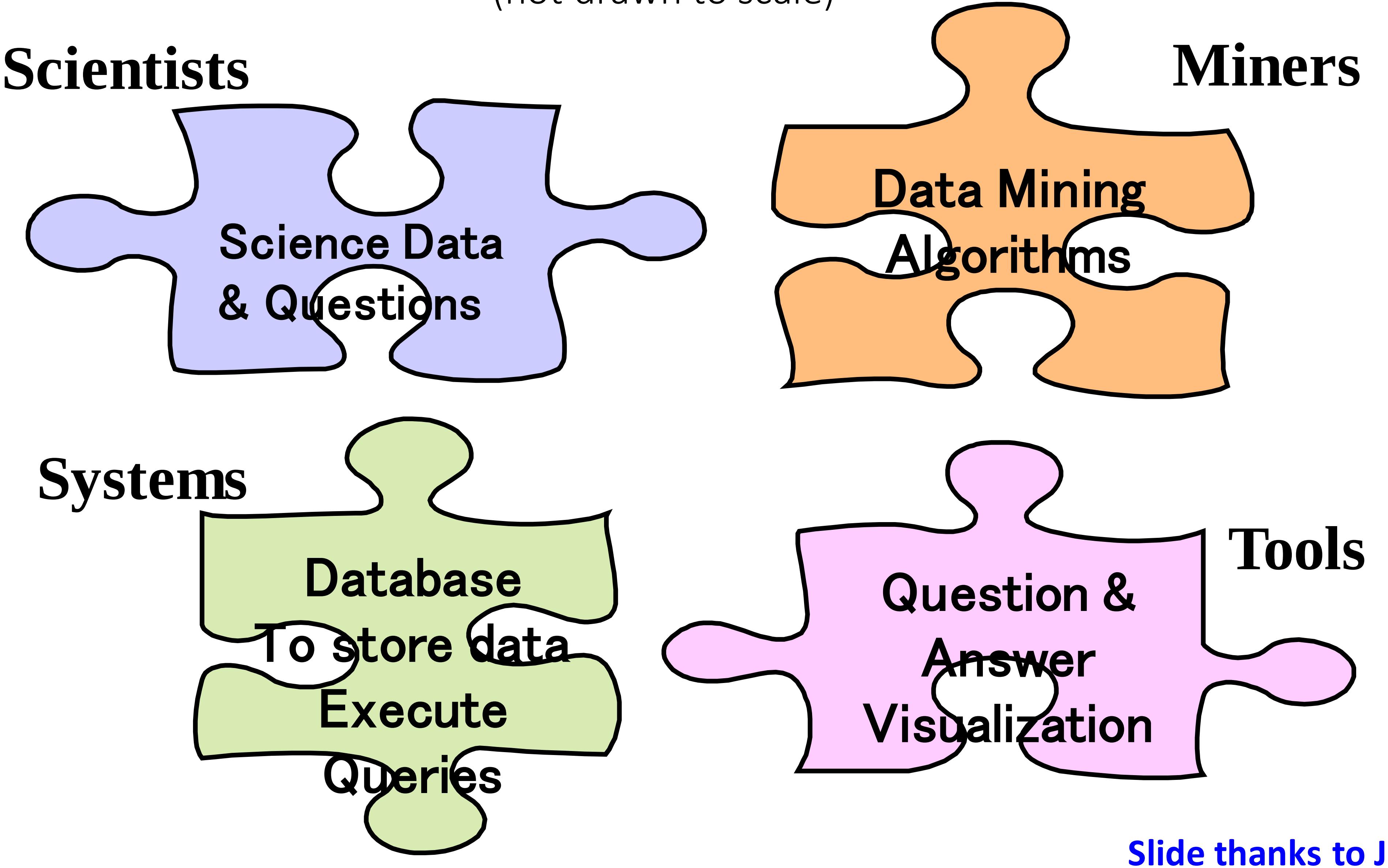
The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize it
- How to share with others
- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

Slide thanks to Jim Gray

What X-info Needs from Computer Science

(not drawn to scale)



Slide thanks to Jim Gray

e-Science and the Fourth Paradigm

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

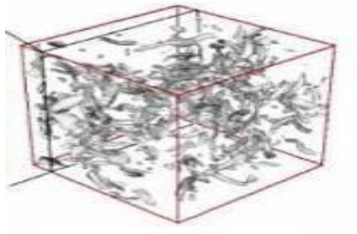
Last few decades – **Computational Science**

- Simulation of complex phenomena

Today – **Data-Intensive Science**

- Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

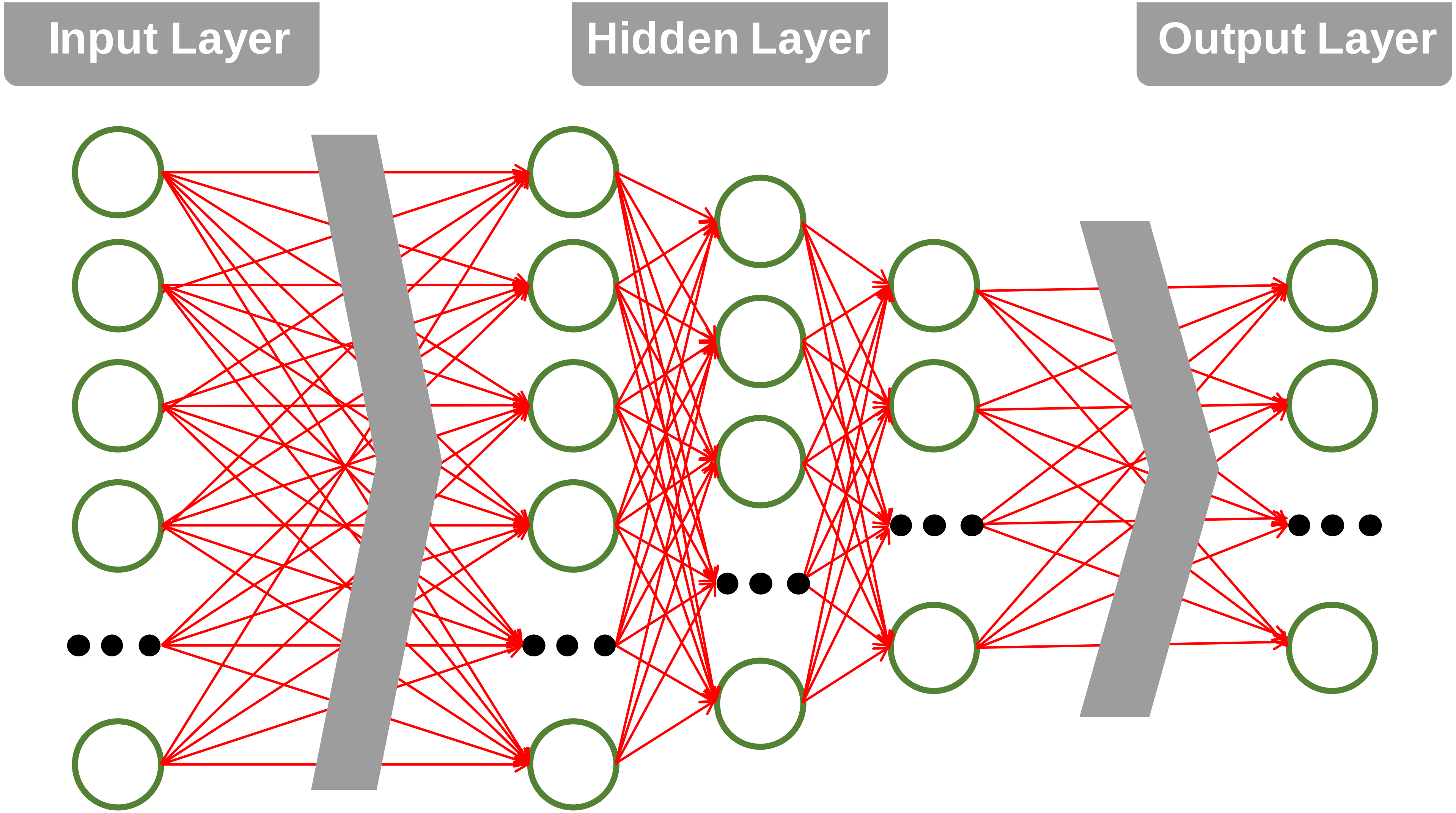


eScience is the set of tools and technologies to support data federation and collaboration

- For analysis and data mining
- For data visualization and exploration
- For scholarly communication and dissemination

With thanks to Jim Gray

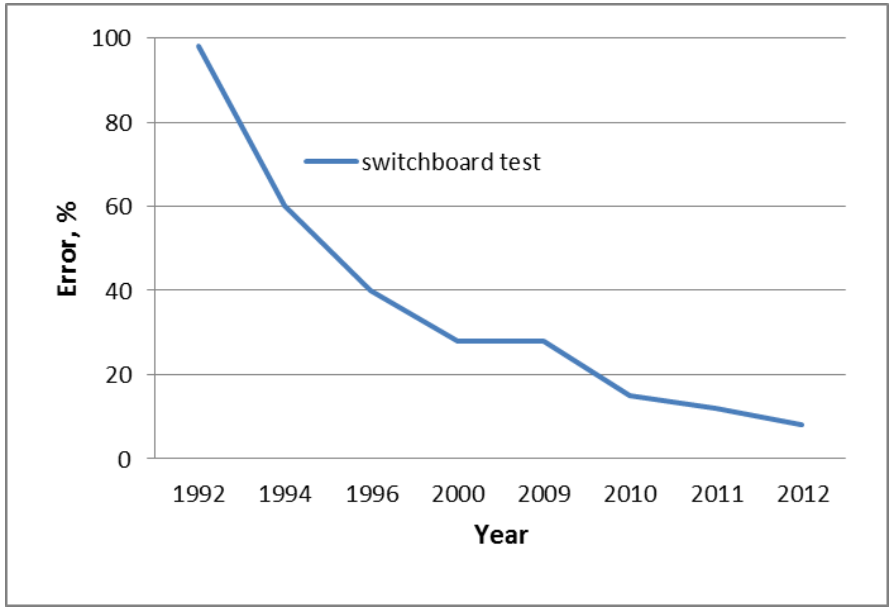
Artificial Neural Networks



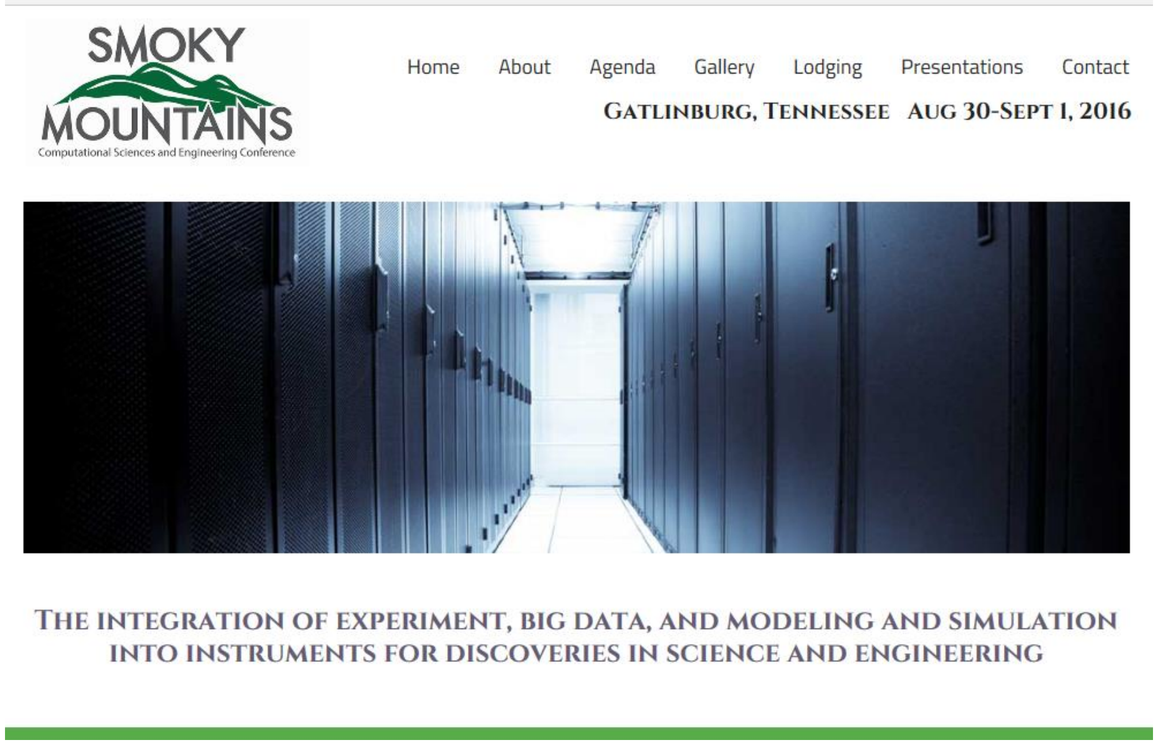
Machine Learning

- Neural networks are one example of a Machine Learning (ML) algorithm
- Deep Neural Networks are now exciting the whole of the IT industry since they enable us to:
    - Build computing systems that improve with experience
    - Solve extremely hard problems
    - Extract more value from Big Data
    - Approach human intelligence

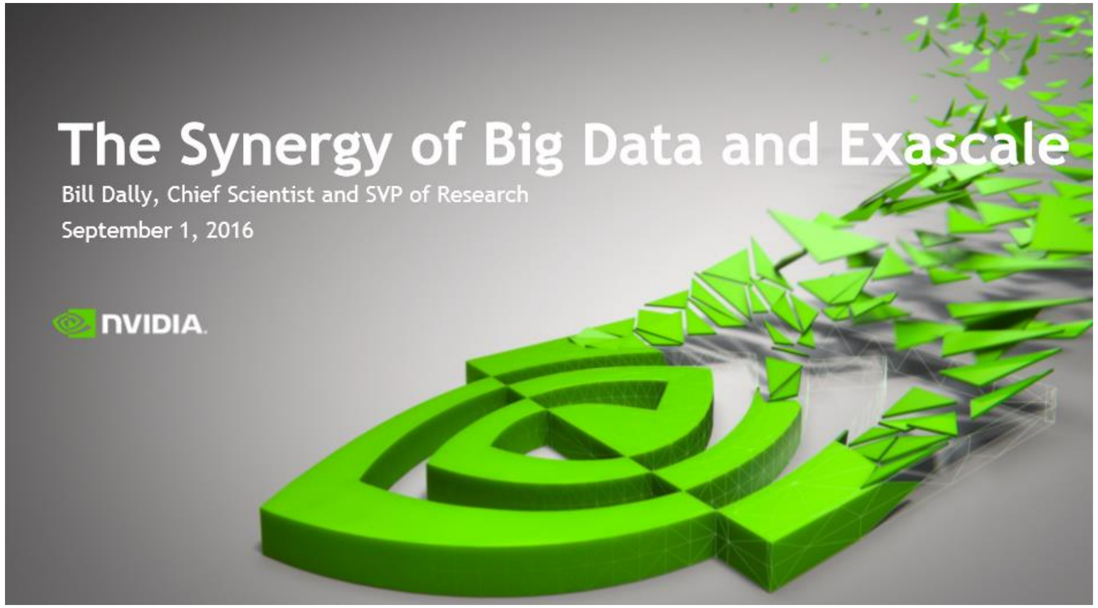
      e.g. natural language processing

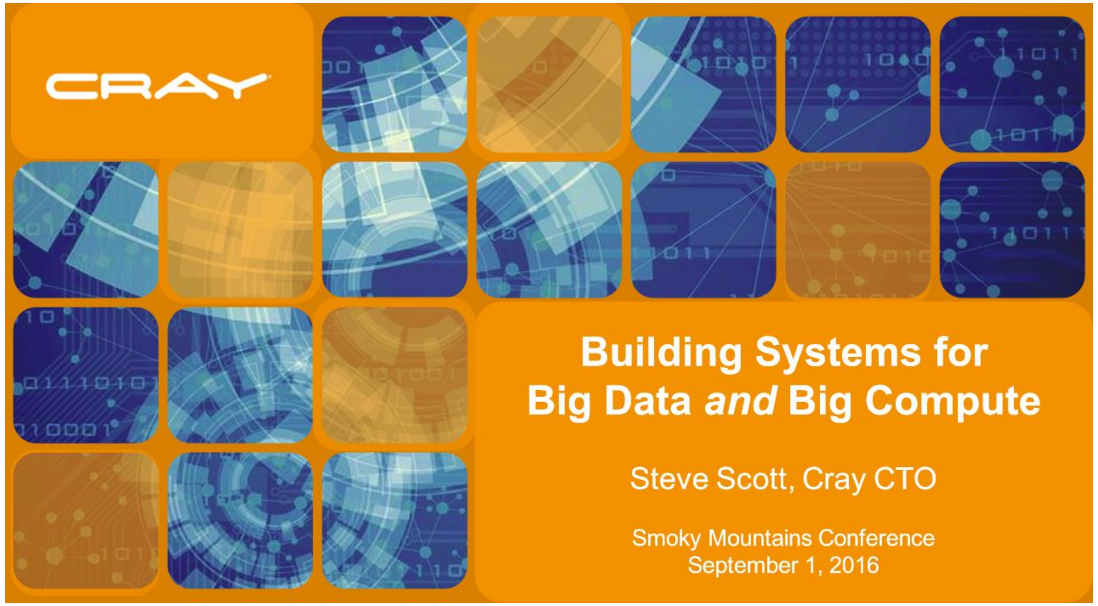


- The change in the Word Error Rate (WER) with time for the NIST “Switchboard” data.
- This shows the dramatic improvement made in the last few years using Deep Neural Networks

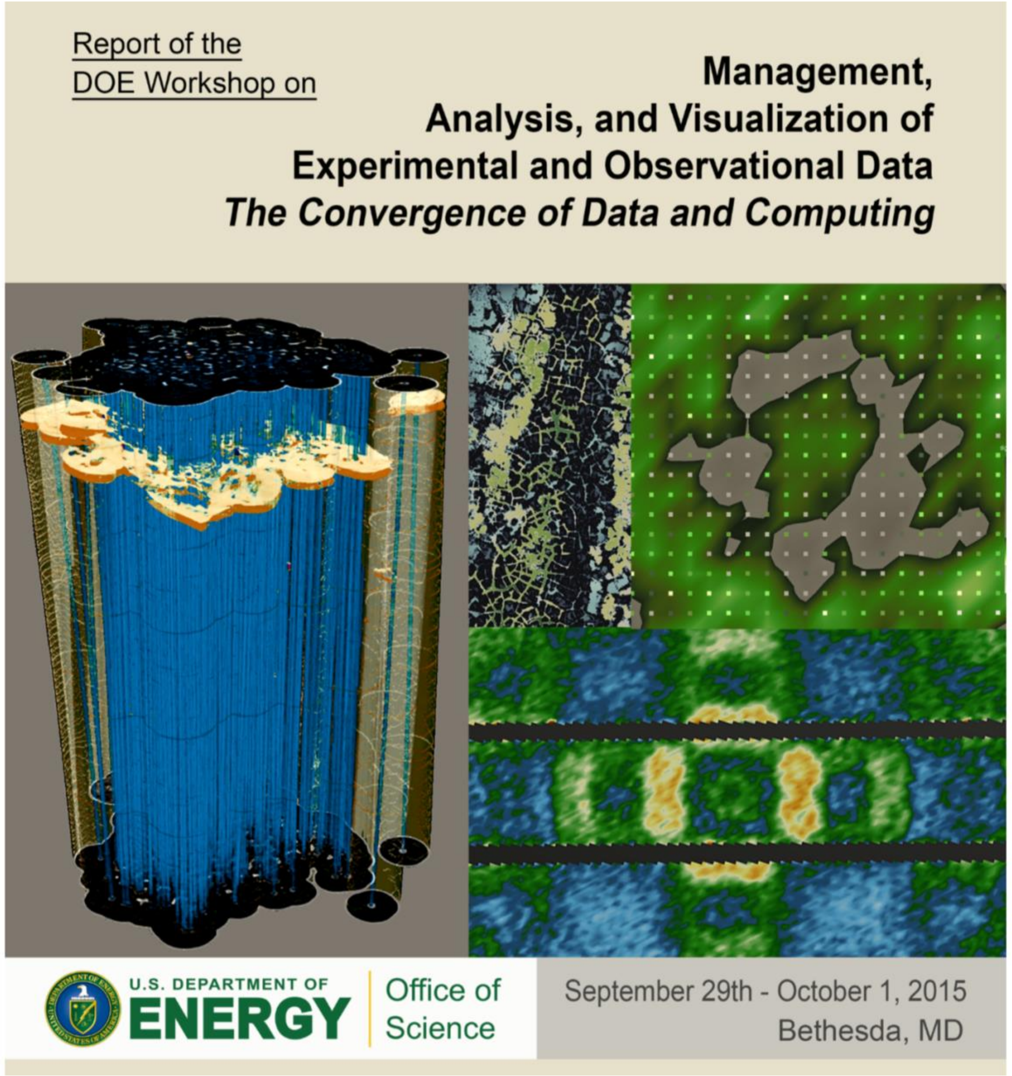

SMOKY MOUNTAINS
Computational Sciences and Engineering Conference
Home About Agenda Gallery Lodging Presentations Contact
GATLINBURG, TENNESSEE AUG 30-SEPT 1, 2016
THE INTEGRATION OF EXPERIMENT, BIG DATA, AND MODELING AND SIMULATION INTO INSTRUMENTS FOR DISCOVERIES IN SCIENCE AND ENGINEERING

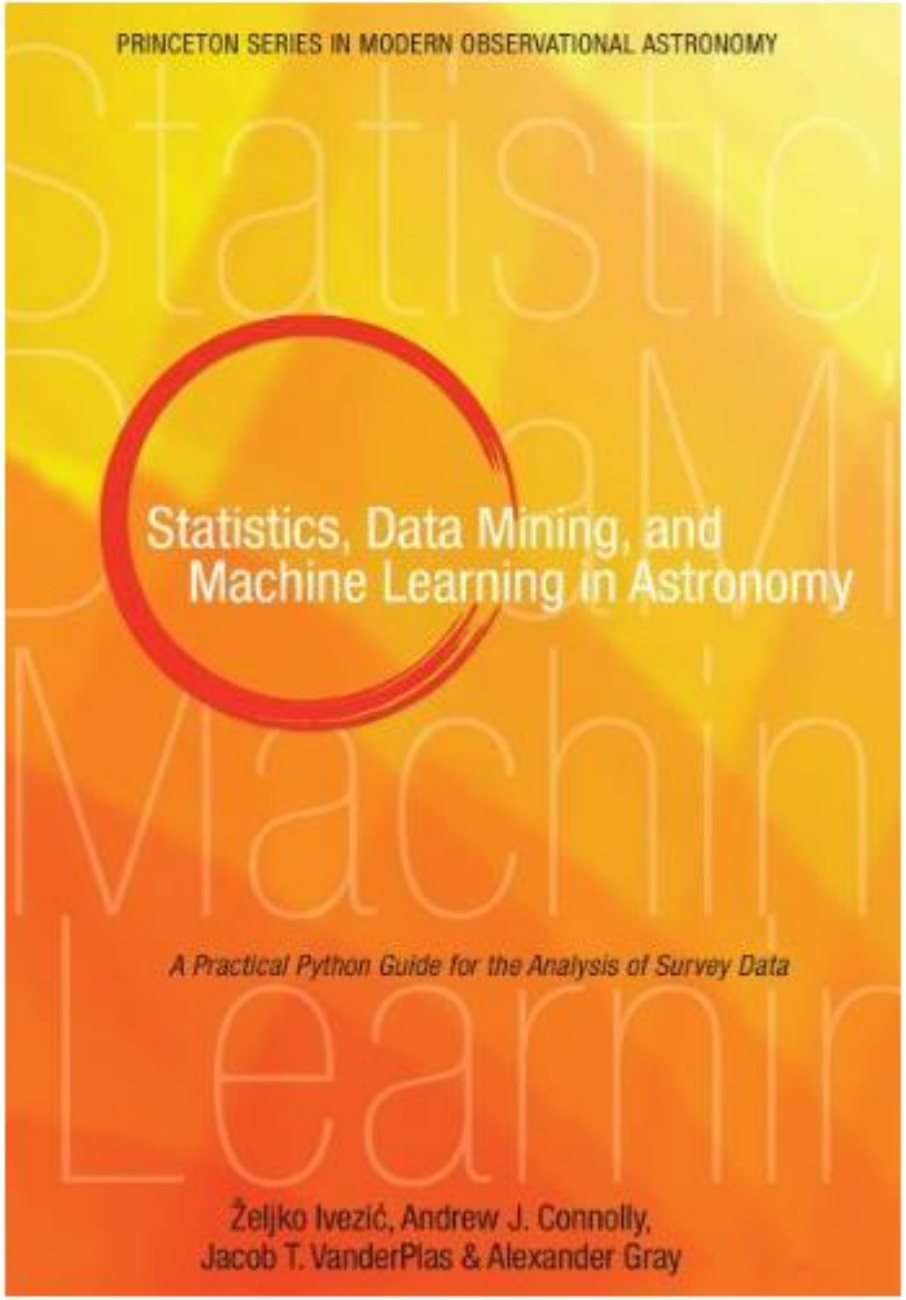

HPC and Big Data:
Better Together!
Kirill Malkin
Director of Storage Engineering
kmalkin@sgi.com
The SGI logos and SGI product names used or referenced herein are either registered trademarks or trademarks of Silicon Graphics International Corp. or one of its subsidiaries. All other trademarks, trade names, service marks and logos referenced herein belong to their respective holders. Any and all copyright or other proprietary notices that appear herein, together with this Legal Notice, must be retained on this presentation. The information contained herein is subject to change without notice.
sgi

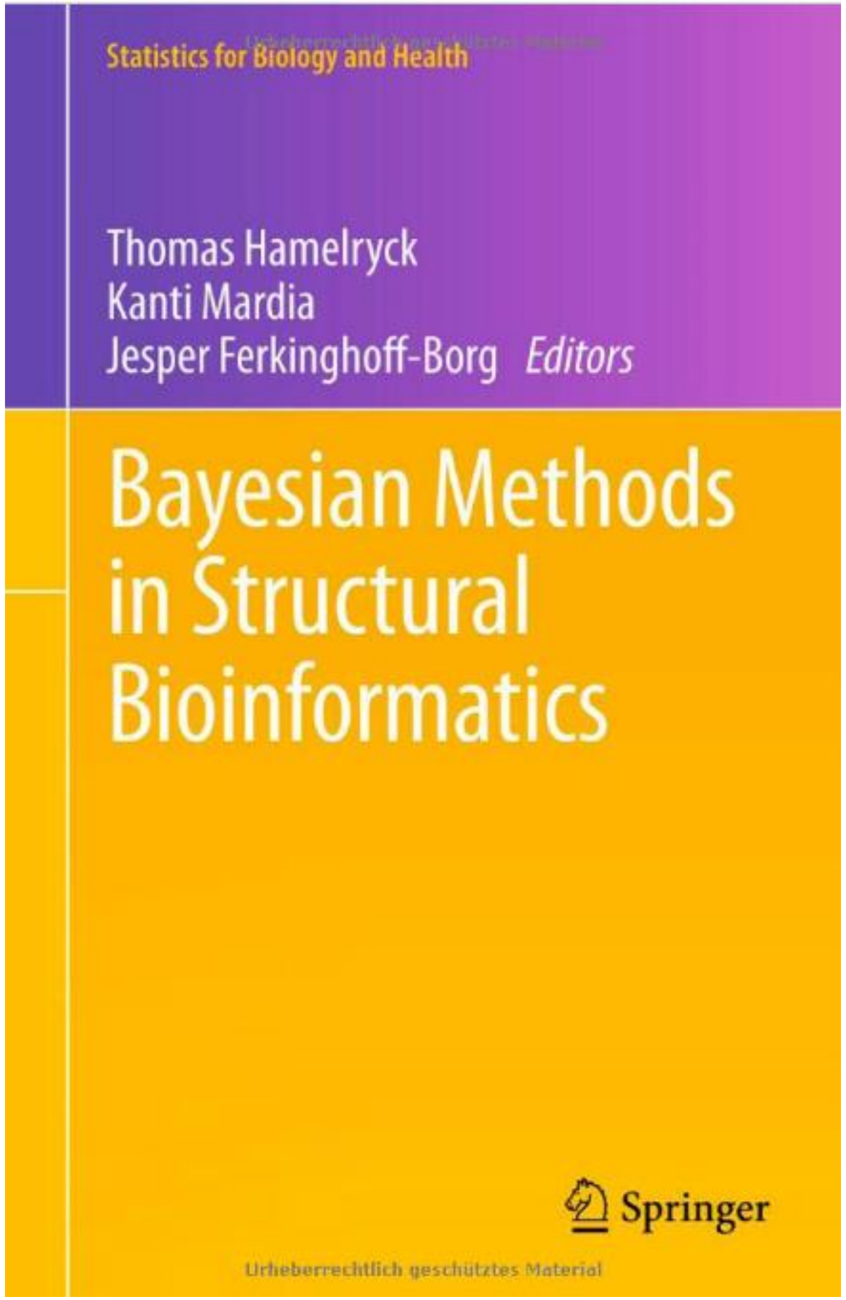

The Synergy of Big Data and Exascale
Bill Dally, Chief Scientist and SVP of Research
September 1, 2016
NVIDIA


CRAY
Building Systems for
Big Data and Big Compute
Steve Scott, Cray CTO
Smoky Mountains Conference
September 1, 2016


Future Directions for
NSF ADVANCED COMPUTING INFRASTRUCTURE
to Support U.S. Science and Engineering in 2017–2020
The National Academies of
SCIENCES · ENGINEERING · MEDICINE


Report of the DOE Workshop on
Management, Analysis, and Visualization of Experimental and Observational Data
The Convergence of Data and Computing
U.S. DEPARTMENT OF ENERGY
Office of Science
September 29th - October 1, 2015
Bethesda, MD


PRINCETON SERIES IN MODERN OBSERVATIONAL ASTRONOMY
Statistics, Data Mining, and Machine Learning in Astronomy
A Practical Python Guide for the Analysis of Survey Data
Željko Ivezić, Andrew J. Connolly,
Jacob T. VanderPlas & Alexander Gray


Statistics for Biology and Health
Thomas Hamelryck
Kanti Mardia
Jesper Ferkinghoff-Borg Editors
Bayesian Methods in Structural Bioinformatics
Springer
Urheberrechtlich geschütztes Material

Data Science and the UK Science and Technology Facilities Council

UK Science and Technology Facilities Council (STFC)

Big Data and Cognitive Computing:
Hartree Centre collaboration with IBM Research

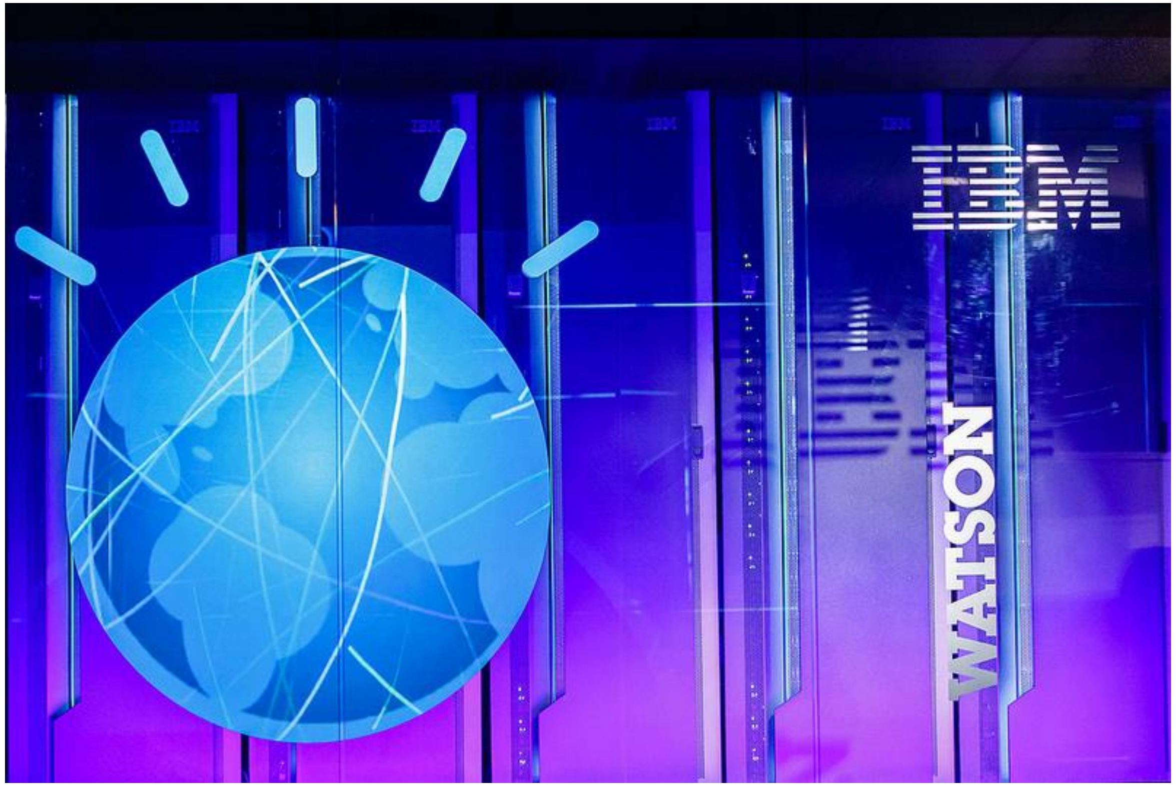
UK Hartree Center Partners with IBM on Big Data

June 4, 2015 by staff

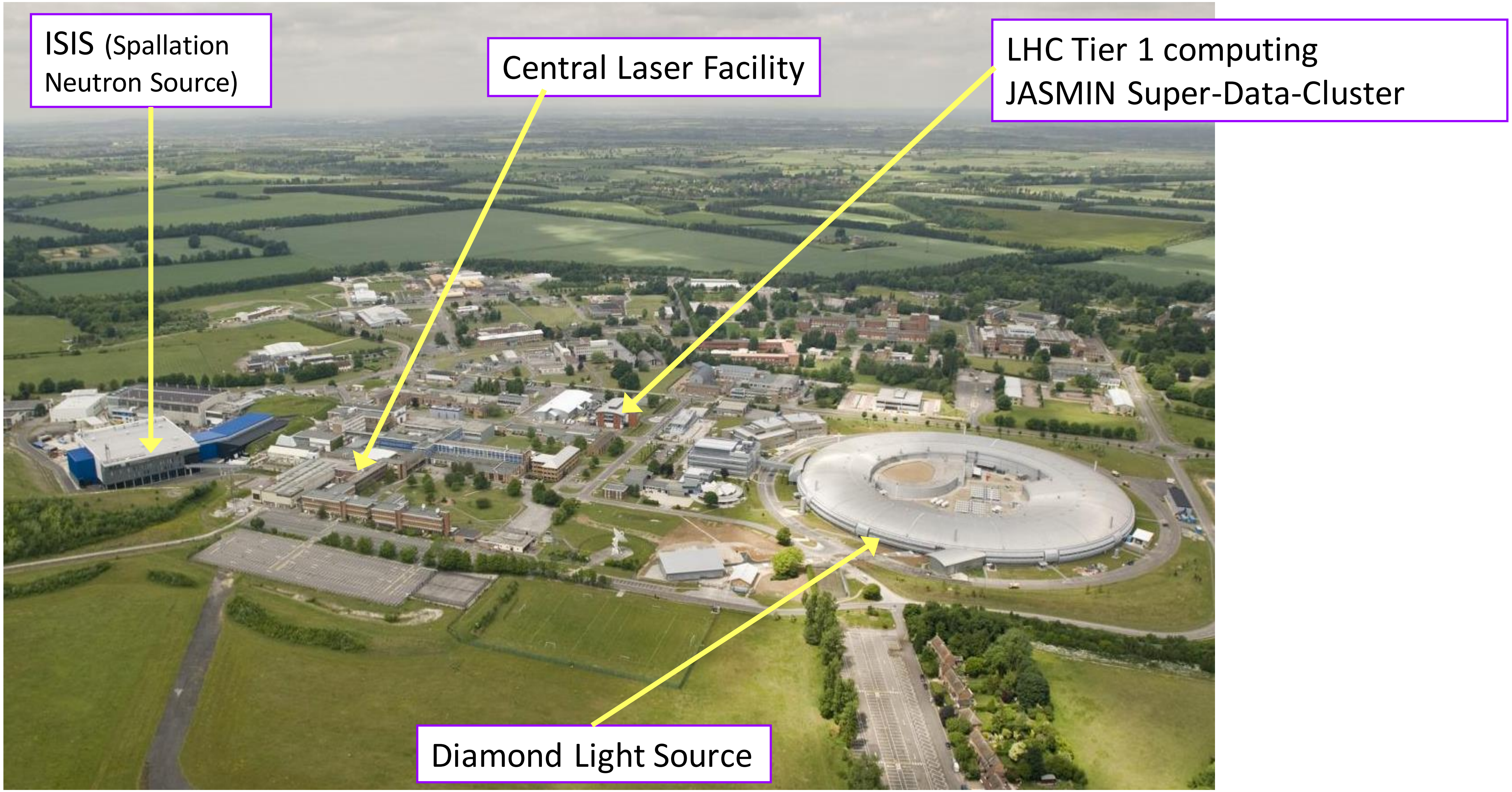
Today the UK government announced a £313 million partnership with information technology leader IBM to boost Big Data research in the UK.

STFC
The Hartree Centre

> *"We live in an information economy – from the smart devices we use every day to the super-computers that helped find the Higgs Boson, the power of advanced computing means we now have access to vast amounts of data," said Minister for Universities and Science Jo Johnson. "This partnership with IBM, which builds on our £113 million investment to expand the Hartree Centre, will help businesses make the best use of Big Data to develop better products and services that will boost productivity, drive growth and create jobs."*



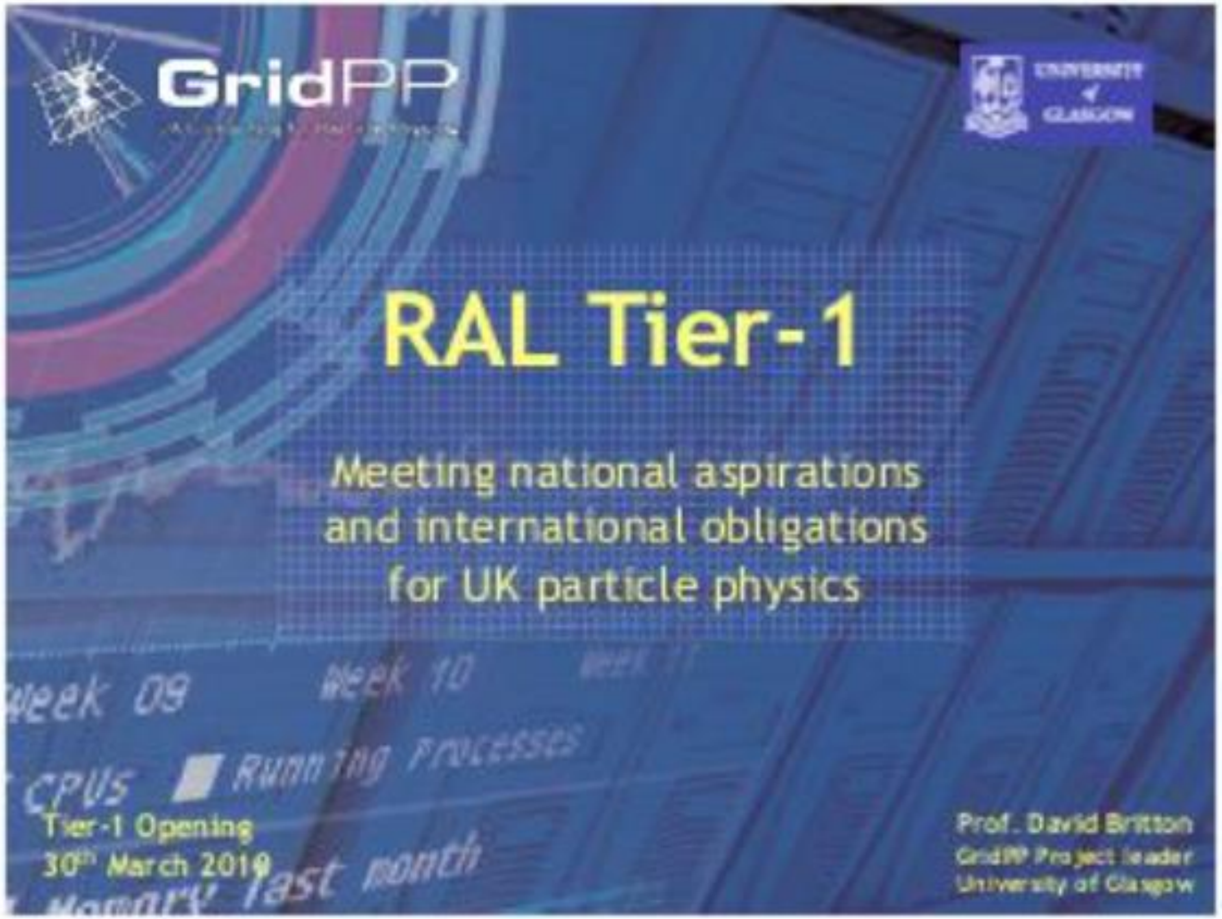
Rutherford Appleton Lab and the Harwell Campus



Applications Division

The Applications Division brings together four groups which develop and apply computational science software packages to solve problems in the physical and biological sciences. The groups are:

- **Computational Biology**, including structural biology, molecular simulation and bioinformatics
- **Theoretical and Computational Physics**, including electronic structure of the solid state and surfaces, atomic and molecular physics
- **Computational Engineering**, focusing on HPC solutions in fluid flow modelling, with particular strength in turbulence and microfluidics
- **Computational Chemistry**, including molecular dynamics, quantum chemistry and QM/MM techniques, and mesoscale methods



Collaborative Computational Projects: The CCP's

- Assist universities in developing, maintaining and distributing computer programs
- Promoting the best computational methods
- Each focuses on a specific area of research
- Funded by the UK's EPSRC, PPARC and BBSRC Research Councils

CCP	Chair	Title
CCP4	Prof David Brown	Macromolecular Crystallography
CCP5	Prof Neil Allan	The Computer Simulation of Condensed Phases
CCP9	Prof Mike Payne	Computational Electronic Structure of Condensed Matter
CCP12	Prof Mark Savill	High Performance Computing in Engineering
CCP-BioSim	Prof Adrian Mulholland	Biomolecular Simulation at the Life Sciences Interface
CCP-EM	Dr Martyn Winn	Electron Cryo-Microscopy
CCPi	Prof Phillip Withers	Tomographic Imaging
CCPN	Prof Geerten Vuister	NMR
CCP-NC	Dr Jonathan Yates	NMR Crystallography
CCP-Plasma	Dr Tony Arber	Computational Plasma Physics
CCPQ *	Prof Graham Worth	Quantum Dynamics in Atomic, Molecular and Optical Physics
CCP-SAS	Prof Steve Perkins	Analysis of Structural Data in Chemical Biology and Soft Condensed Matter
CCPForge	Catherine Jones	Collaborative Software Development Environment Tool
CCPPET/MR	Dr Kris Thielemans	Positron Emission Tomography (PET) and Magnetic Resonance (MR) Imaging
CCP CoDiMa	Prof Steve Linton	Computational Discrete Mathematics
CCP-WSI	Prof Deborah Greaves	A Collaborative Computational Project in Wave/Structure Interaction
CCPmag	Prof Julie Staunton	Computational Magnetism

CCP4

Collaborative Computational Project No. 4
Software for Macromolecular X-Ray Crystallography

CCP4 exists to produce and support a world-leading, integrated suite of programs that allows researchers to determine macromolecular structures by X-ray crystallography, and other biophysical techniques. CCP4 aims to develop and support the development of cutting edge approaches to experimental determination and analysis of protein structure, and integrate these approaches into the suite. CCP4 is a community based resource that supports the widest possible researcher community, embracing academic, not for profit, and for profit research. CCP4 aims to play a key role in the education and training of scientists in experimental structural biology. It encourages the wide dissemination of new ideas, techniques and practice.

Resources

Download CCP4 Software Suite
[Current Release] [Updates to Current Release]

Usage
[Documentation] [Referencing CCP4] [Licensing]
[Tutorials inc. ccp4i2]
[BAG training]
[CCP4 Wiki] [Community Wiki]

Help and Support
[Report a Problem]

Web Services
BALBES, MrBUMP, AMPLE, CRANK2, SHELX, Zanuda, PISA

CCP4 Bulletin Boards
[CCP4bb Mailing List] [Developers List] [Archives]

CCP4 Newsletter
[Current issue] [Back issues] [Contribute!]

CCP4 Working Groups 1 & 2, and Exec
[How to join and meeting minutes]

Study Weekend and Courses
[Study Weekend] [Workshops and conferences]

Commercial Use of CCP4
[Information on Commercial Use]

CCP4 is supported by

Science & Technology Facilities Council — Scientific Computing Department

bbsrc — Biotechnology and Biological Sciences Research Council

MRC Medical Research Council — Medical Research Council

CCP4 Twitter Feed

New! CCP4 twitter feed for all the latest updates and community announcements Follow

CCP4 Software Suite

The current version is CCP4 7.0 (07 January 2016). The new packages include:

- SHELX suite: co distribution for academic users
- CCP4I2: new ccp4 interface
- DIALS: data processing and integration
- ARCIMBOLDO-LITE: molecular replacement pipeline

The updated packages include PHASER, XIA2, REFMAC, MONOMER LIBRARY, AMPLE, COOT, AIMLESS POINTLESS, CTRUNCATE, DIMPLE, BLEND and many more

CCP4 Study Weekend 2017

Earlier bird registration ends 20th November.

- **From Crystal to Structure** with scientific organisers Keith Wilson (York Uni., UK) and Mike Hough (Essex Uni., UK). Nottingham UK, 9-11 Jan. 2017
  [Information] [Programme] [Registration] [Bursaries] [Satellite]

Proceedings of the 2015 Study Weekend on **Advances in Experimental Phasing** are now available. For more information about past events see details of past Study Weekends and proceedings.

Upcoming Courses and Events

- CCP4/DLS Data Collection and Structure Solution Workshop, Diamond Light Source, UK *13/12/2016 to 20/12/2016*
- CCP4/Spring-8 Structure Solution Workshop, Spring-8, Japan *23/01/2017 to 27/01/2017*

Browse all forthcoming and past courses and events

CCP4 Online Automated Webservices

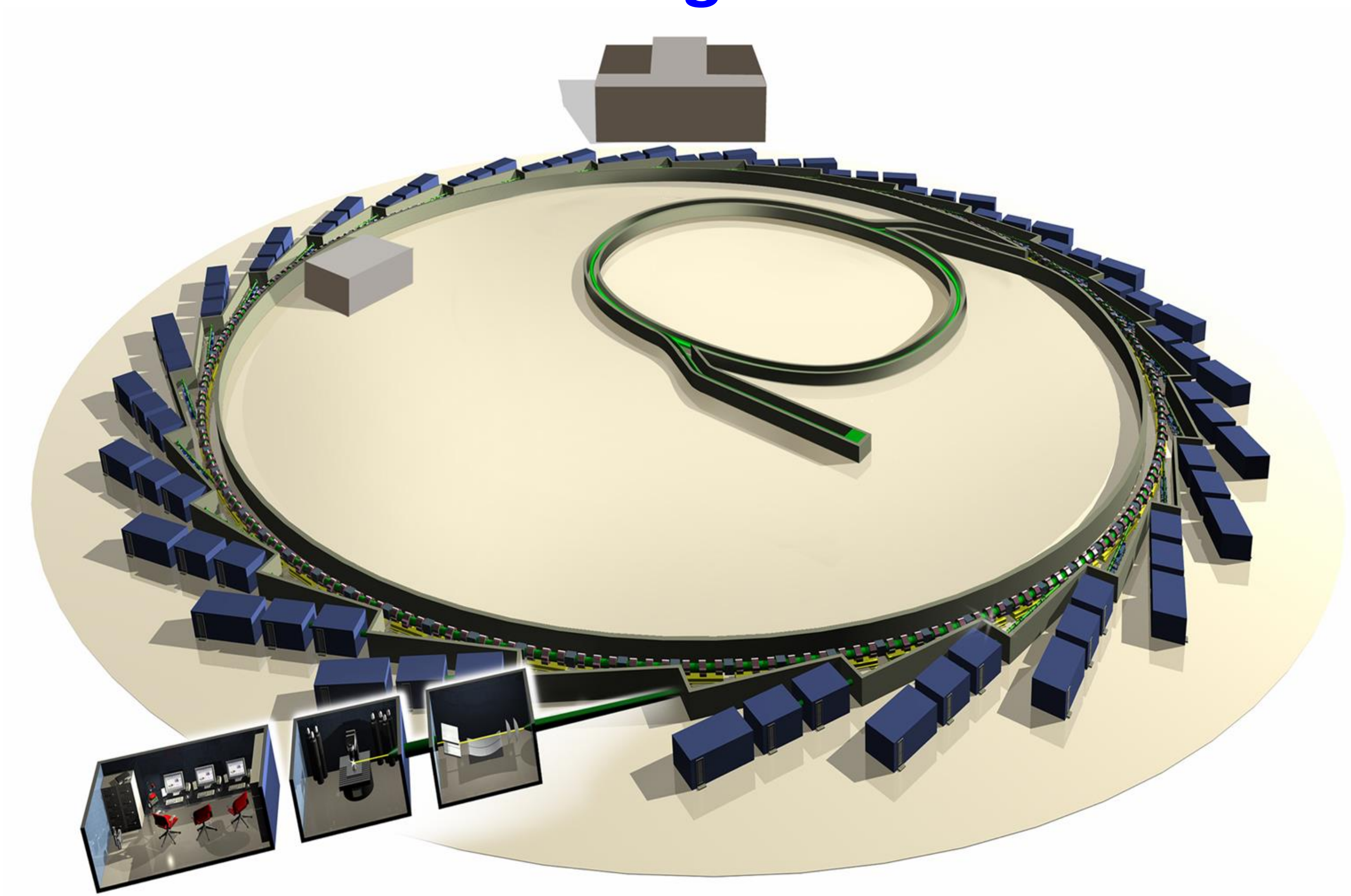
CCP4 on-line NEW! The CCP4-online webserver is now available. Users can make use of BALBES and MrBUMP, the automated molecular replacement services, Zanuda, the refinement result checking software and PISA for the calculation and analysis of macromolecular surfaces and interfaces are also available. To access the services please click here.

CCP4 Documentation Wiki

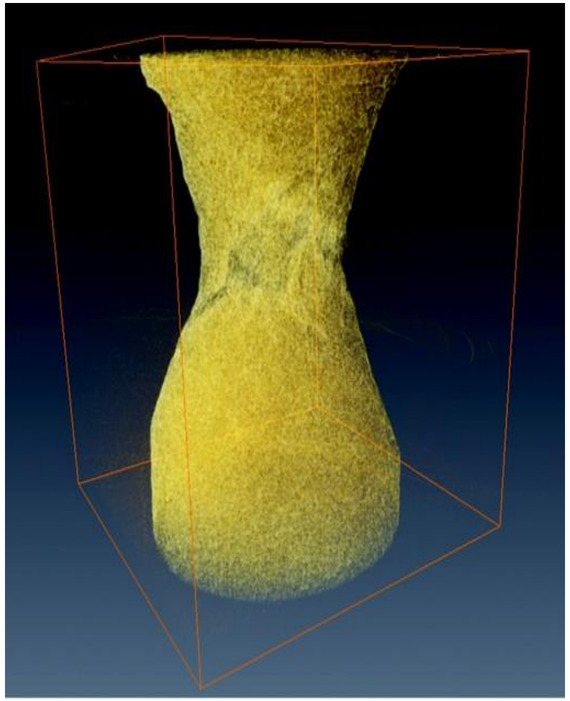
For up-to-date documentation on all of CCP4's software and lots of other useful information for X-ray crystallographers see the CCP4 Wiki.

The Diamond Synchrotron

Diamond Light Source

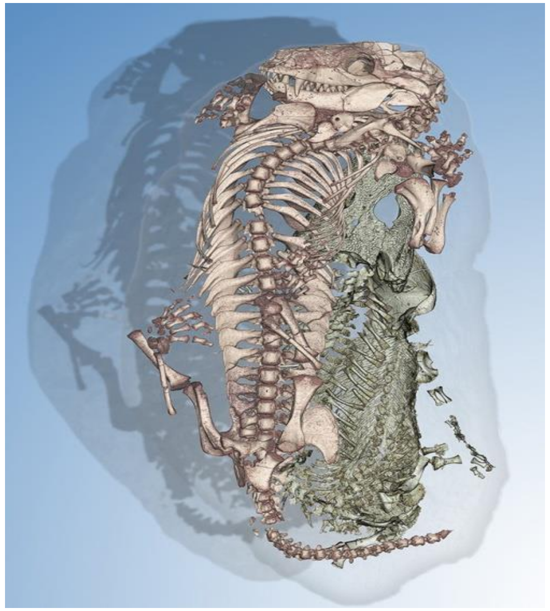


Science Examples

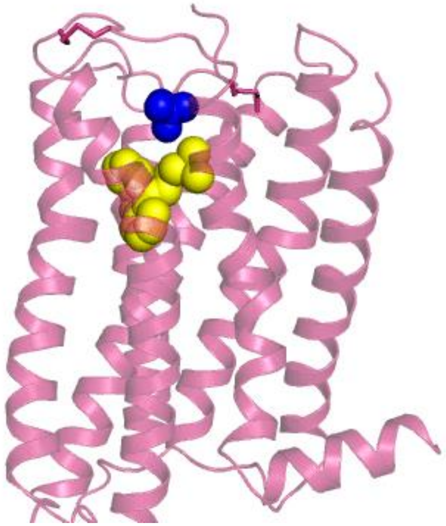


Casting aluminium

Pharmaceutical manufacture & processing

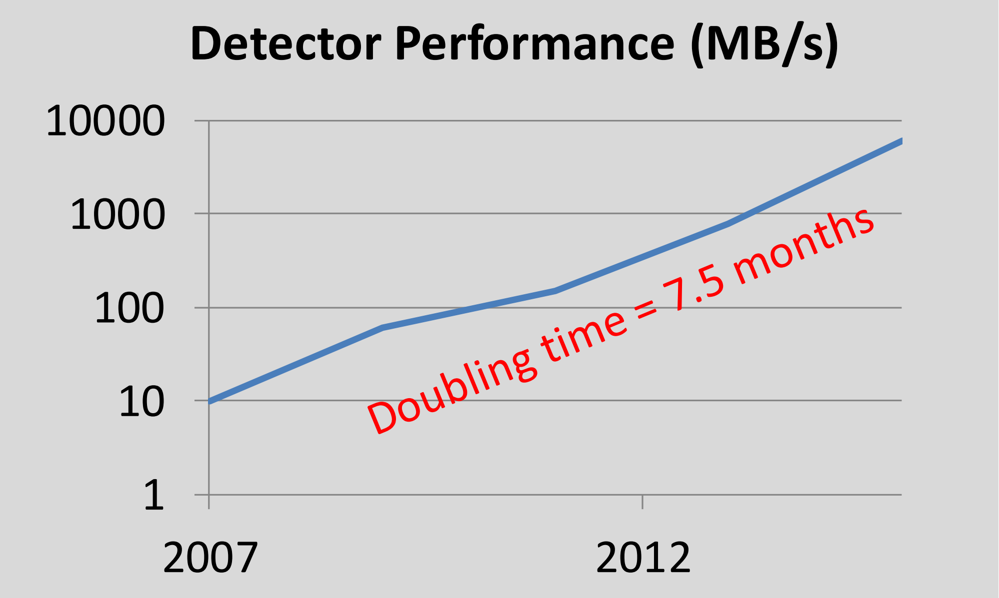


Non-destructive imaging of fossils



Structure of the Histamine H1 receptor

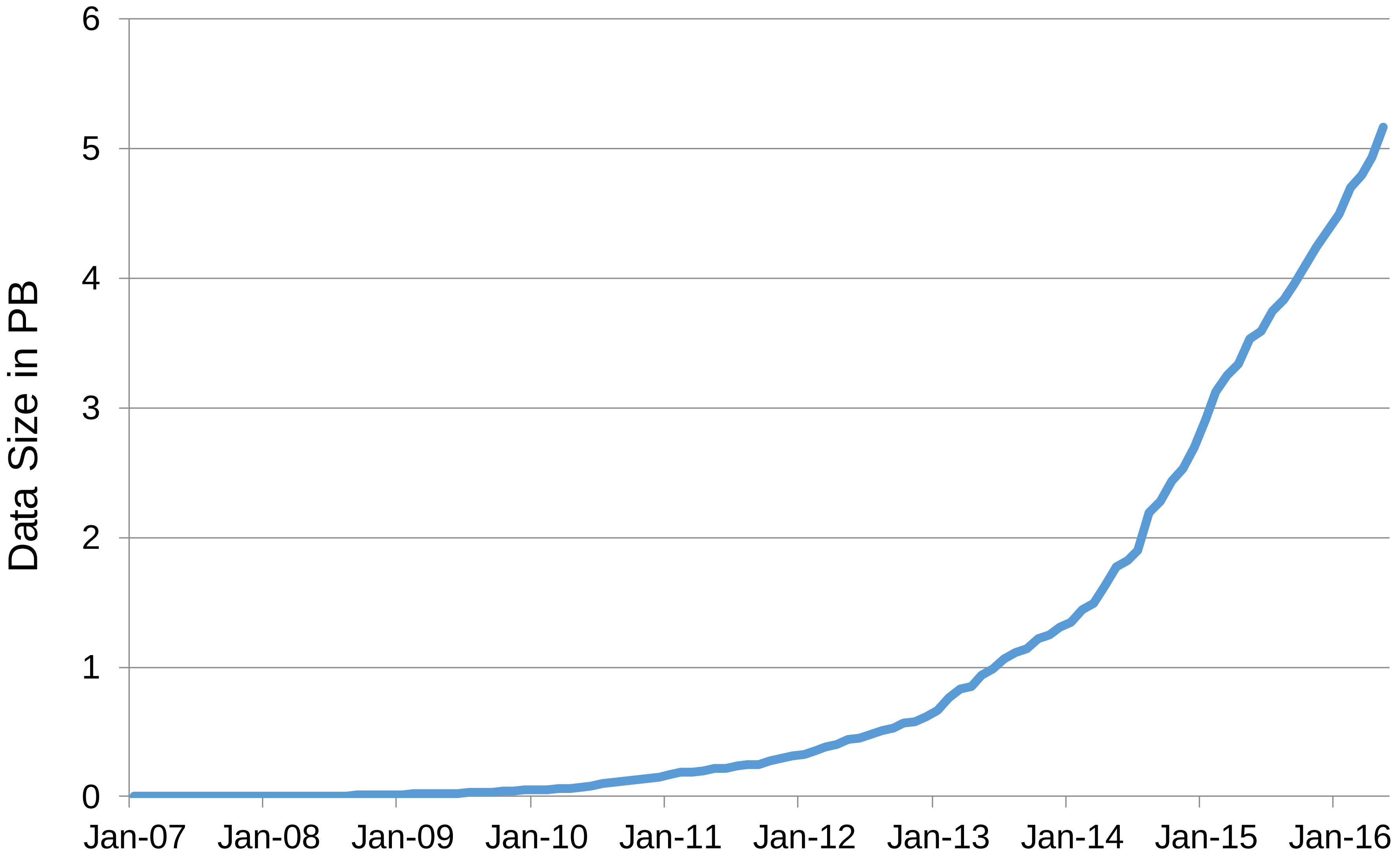
Data Rates



- 2007 No detector faster than ~10 MB/sec
- 2009 Pilatus 6M system 60 MB/s
- 2011 25Hz Pilatus 6M 150 MB/s
- 2013 100Hz Pilatus 6M 600 MB/sec
- 2013 ~10 beamlines with 10 GbE detectors (mainly Pilatus and PCO Edge)
- 2016 Percival detector 6GB/sec

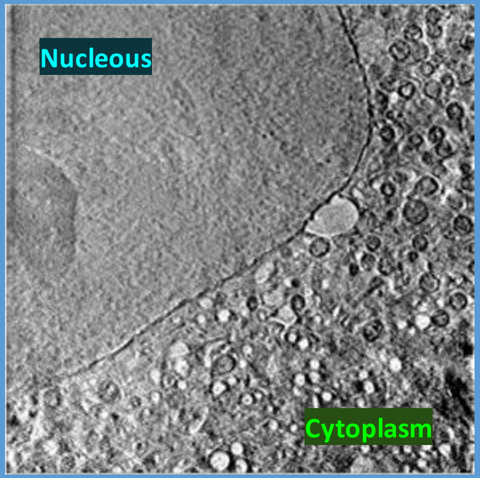
Thanks to Mark Heron

Cumulative Amount of Data Generated By Diamond



Thanks to Mark Heron

Cryo-SXT Data



Neuronal-like mammalian cell line; single slice

Challenges:

- Noisy data, missing wedge artifacts, missing boundaries
- Tens to hundreds of organelles per dataset
- Tedious to manually annotate
- Cell types can look different
- Few previous annotations available
- Automated techniques usually fail

scientificsoftware@diamond.ac.uk

Segmentation of Cryo-soft X-ray Tomography (Cryo-SXT) data

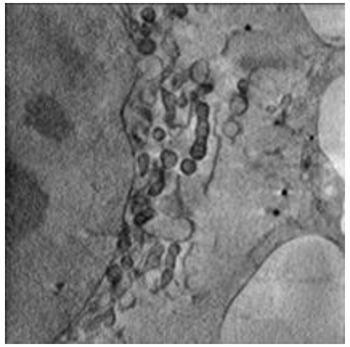
Data

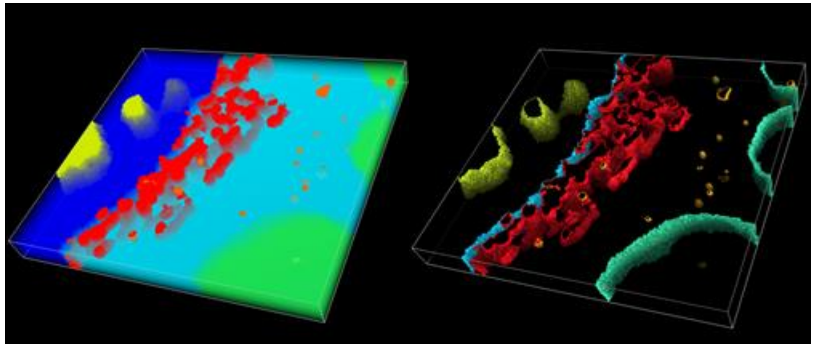
- **B24:** Cryo Transmission X-ray Microscopy beamline at DLS
- Data Collection: Tilt series from ±65° with 0.5° step size
- Reconstructed volumes up to 1000x1000x600 voxels
- Voxel resolution: ~40nm currently
- Total depth: up to 10μm
- **GOAL: Study structure and morphological changes of whole cells**

diamond
B24 beamline
Data Analysis Software Group

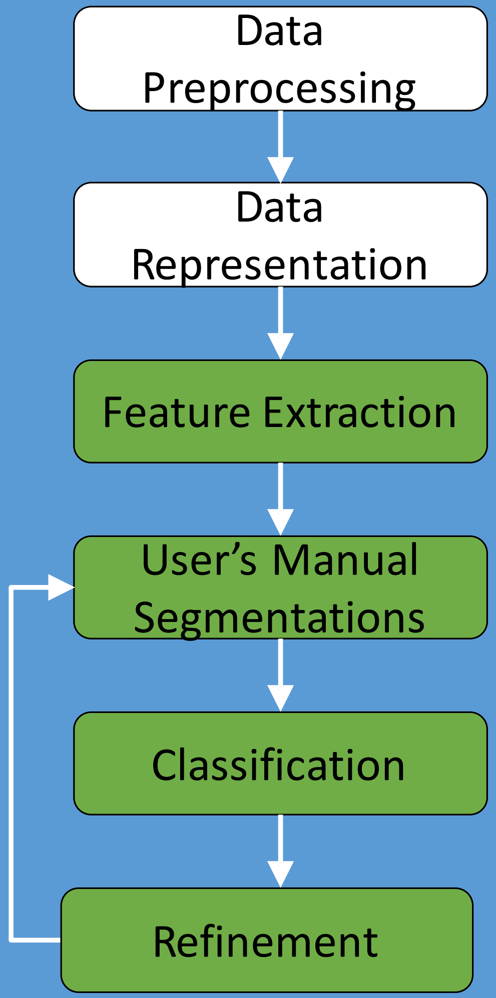
The University of Nottingham
Computer Vision Laboratory

3D Volume Data → Segmentation



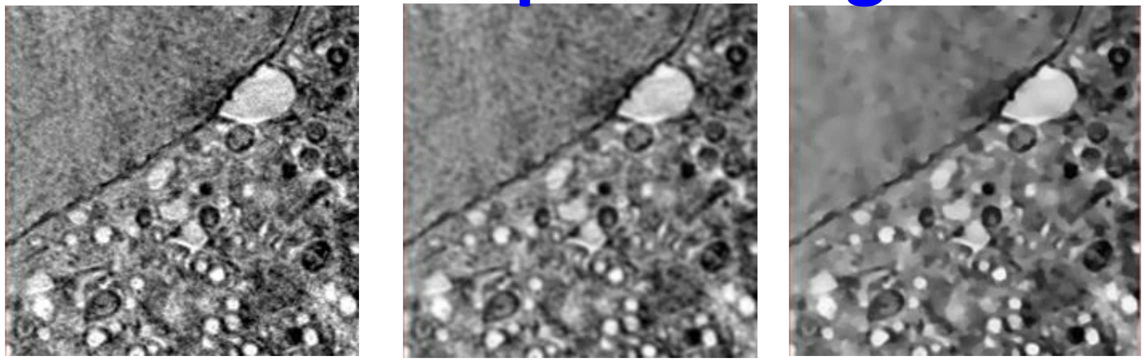


Workflow



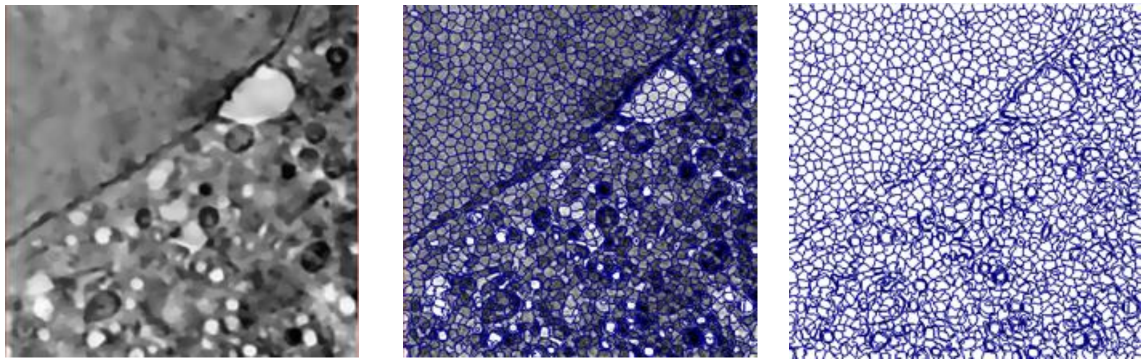
scientificsoftware@diamond.ac.uk

Data Preprocessing



Raw Slice | Gaussian Filter | Total Variation

Data Representation



SuperVoxels (SV) | SV Boundaries

SuperVoxels:

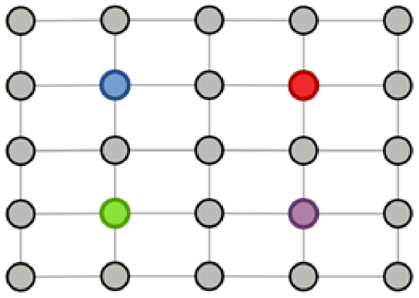
- Groups of similar and adjacent voxels in 3D
- Preserve volume boundaries
- Reduce noise when representing data
- Reduce problem complexity several orders of magnitude
- Use Local clustering in *{xyz + λ * intensity}* space

Workflow

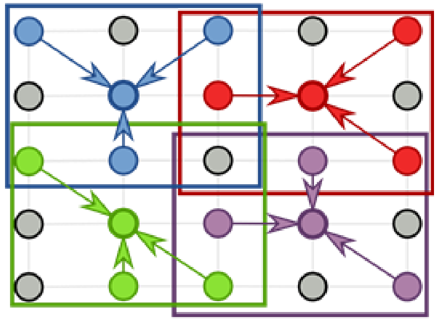
- Data Preprocessing
- Data Representation
- Feature Extraction
- User's Manual Segmentations
- Classification
- Refinement

scientificsoftware@diamond.ac.uk

Data Representation

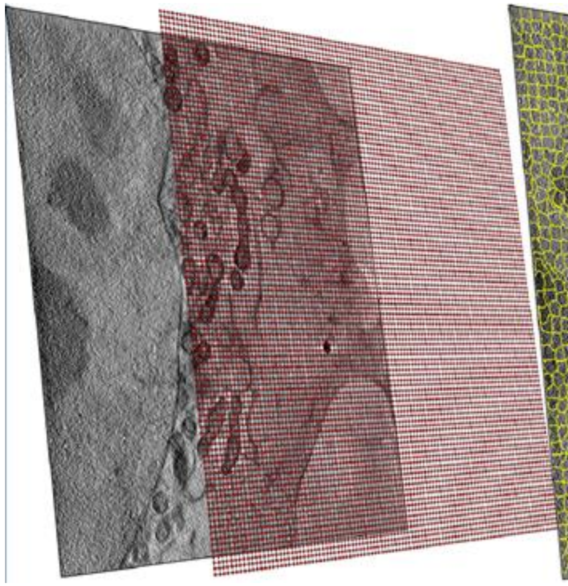


Initial Grid with uniformly sampled seeds



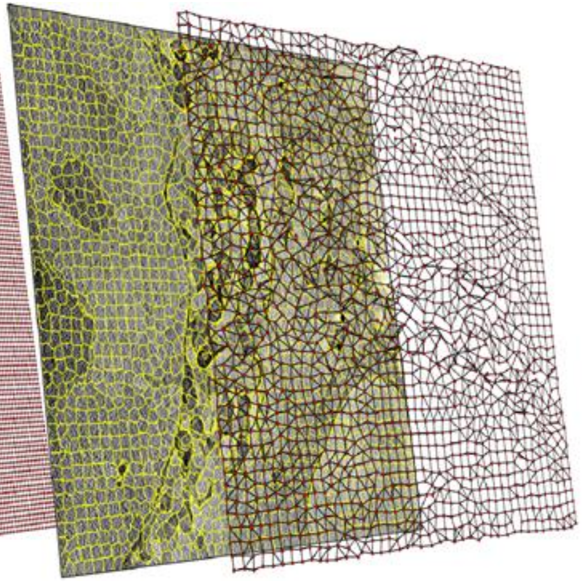
Local *k*-means in a small window around seeds

Voxel Grid



946 x 946 x 200 = 180M voxels

Supervoxel



180M / (10x10x10) = 180K supervoxels

Workflow

Data Preprocessing → Data Representation → Feature Extraction → User's Manual Segmentations → Classification → Refinement → User's Manual Segmentations

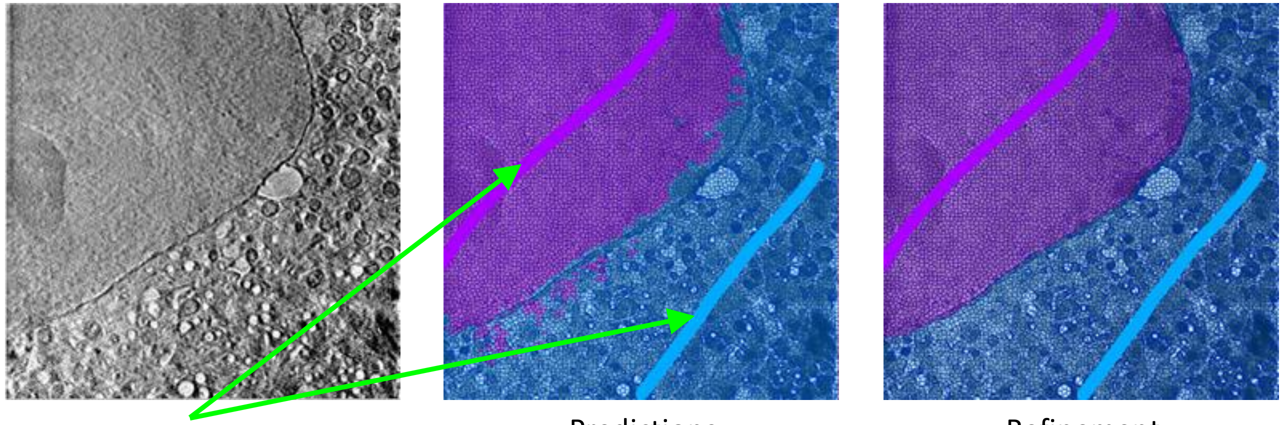
scientificsoftware@diamond.ac.uk

Feature Extraction

Features are extracted from voxels to represent their appearance:

- Intensity-based filters (Gaussian Convolutions)
- Textural filters (eigenvalues of Hessian and Structure Tensor)

User Annotation + Machine Learning



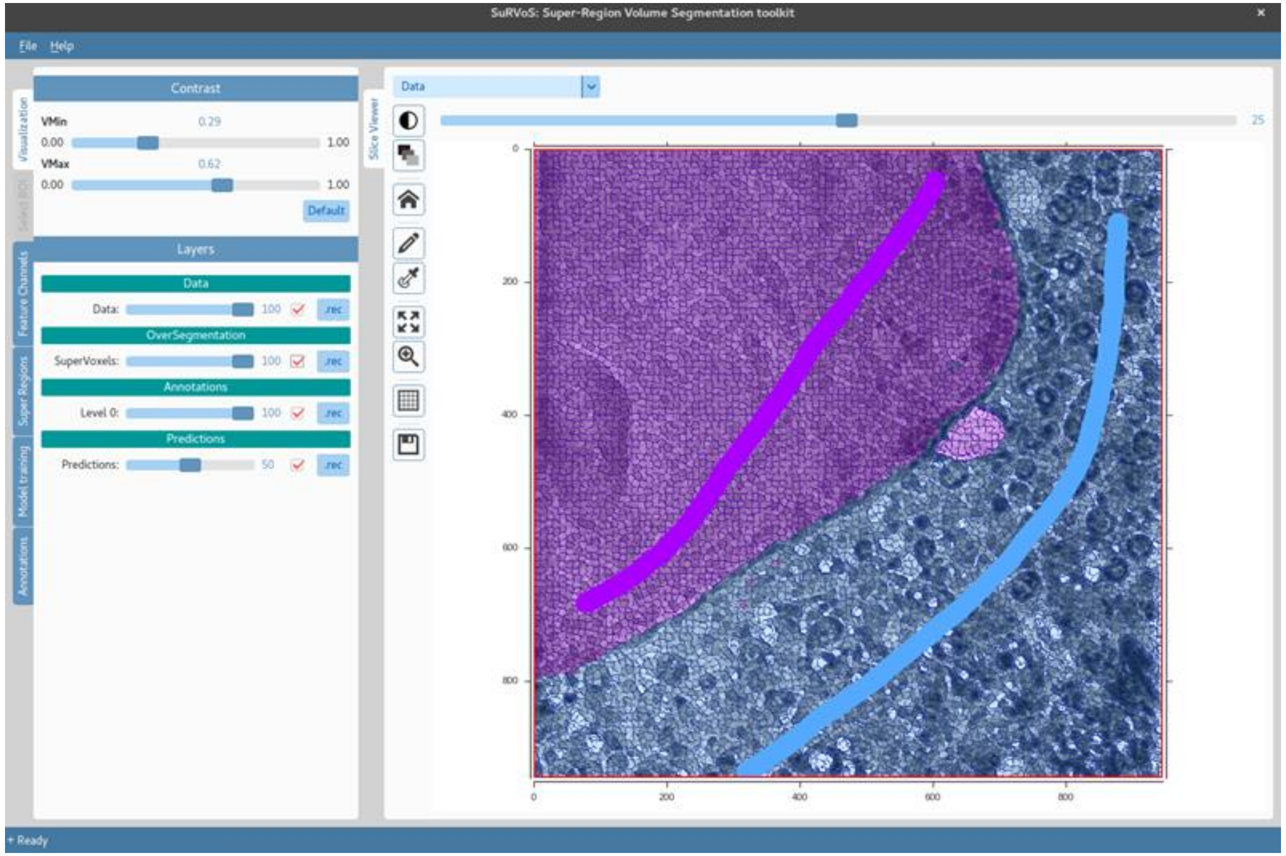
User Annotations Predictions Refinement

Using a few user annotations along the volume as an input:

- A machine learning classifier (i.e. Random Forest) is trained to discriminate between different classes (i.e. Nucleus and Cytoplasm) and predict the class of each SuperVoxel in the volume.
- A Markov Random Field (MRF) is then used to refine the predictions.

SuRVoS Workbench

(Su)per-(R)egion (Vo)lume (S)egmentation



Coming soon: https://github.com/DiamondLightSource/SuRVoS

scientificsoftware@diamond.ac.uk

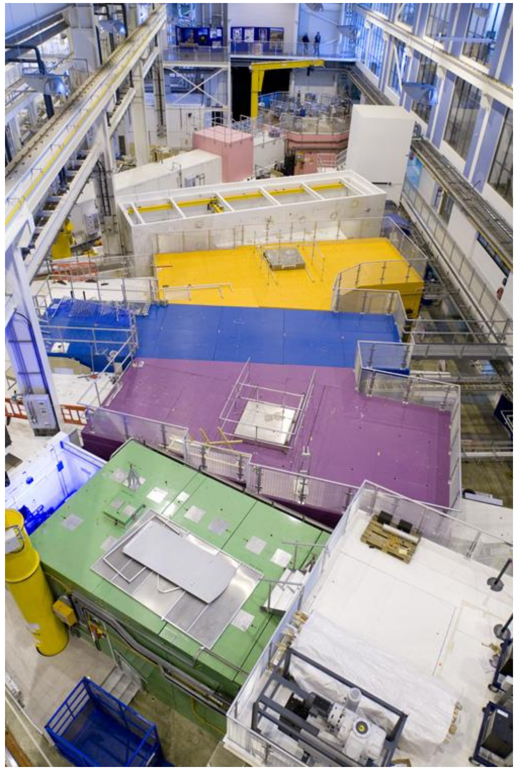
Imanol Luengo *<imanol.luengo@nottingham.ac.uk>*, Michele C. Darrow, Matthew C. Spink, Ying Sun, Wei Dai, Cynthia Y. He, Wah Chiu, Elizabeth Duke, Mark Basham, Andrew P. French, Alun W. Ashton

The ISIS Neutron and Muon Facility

ISIS

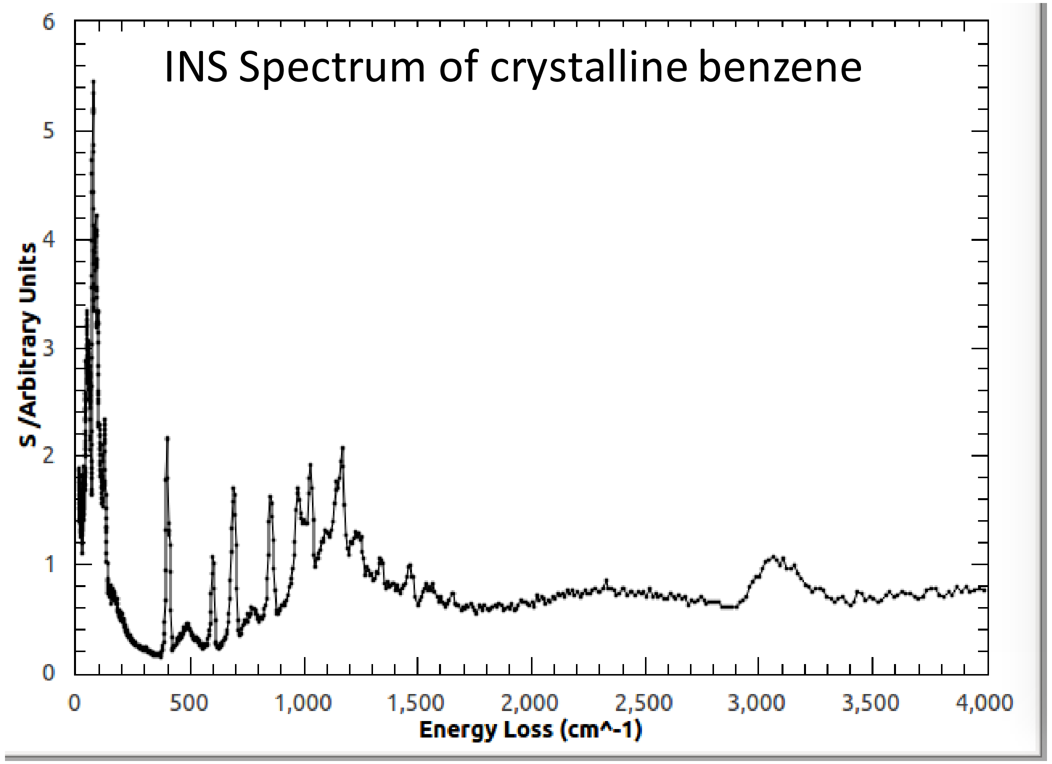
ISIS

- ≈30 neutron instruments
- 3 muon instruments
- 1400 individual users per year making 3000 visits
- 800 experiments per year resulting in 450 publications
- Diverse science
    - Fundamental condensed matter physics
    - Functional materials e.g. multiferroics, spintronics
    - Chemical spectroscopy e.g. catalysis and hydrogen storage
    - Engineering e.g. stress and fatigue in power plants and transportation
    - Solvents in industry
    - Structure of pharmaceutical compounds, biological membranes



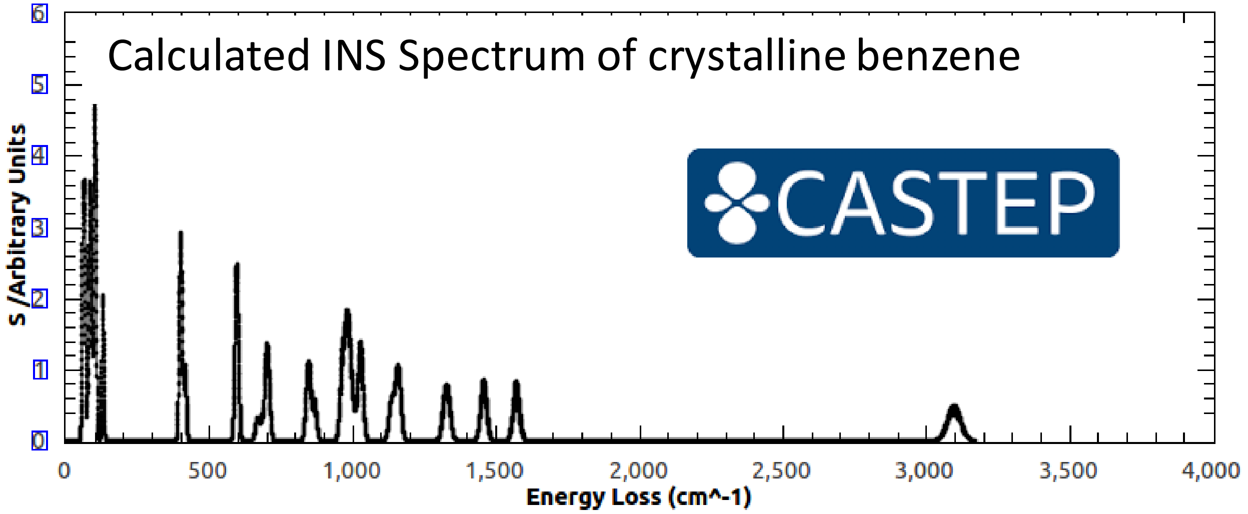
Peak Assignment in Inelastic Neutron Scattering

- Vibrational motion of atoms crucial for many properties of a material -e.g., how well it conducts electricity or heat
- Peaks in INS spectrum correspond to specific atomic vibrations
- Peak assignment: what specific vibrational motions of atoms give rise to specific peaks ?

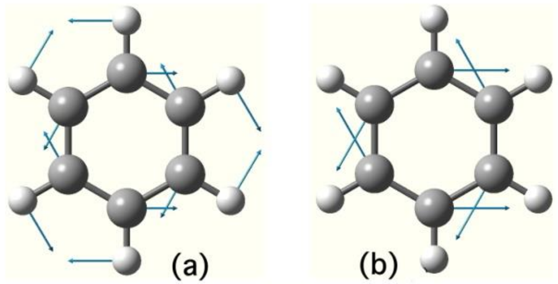


S. Parker and S. Mukhopadhyay (ISIS)

Modelling & Simulation for INS Peak Assignment

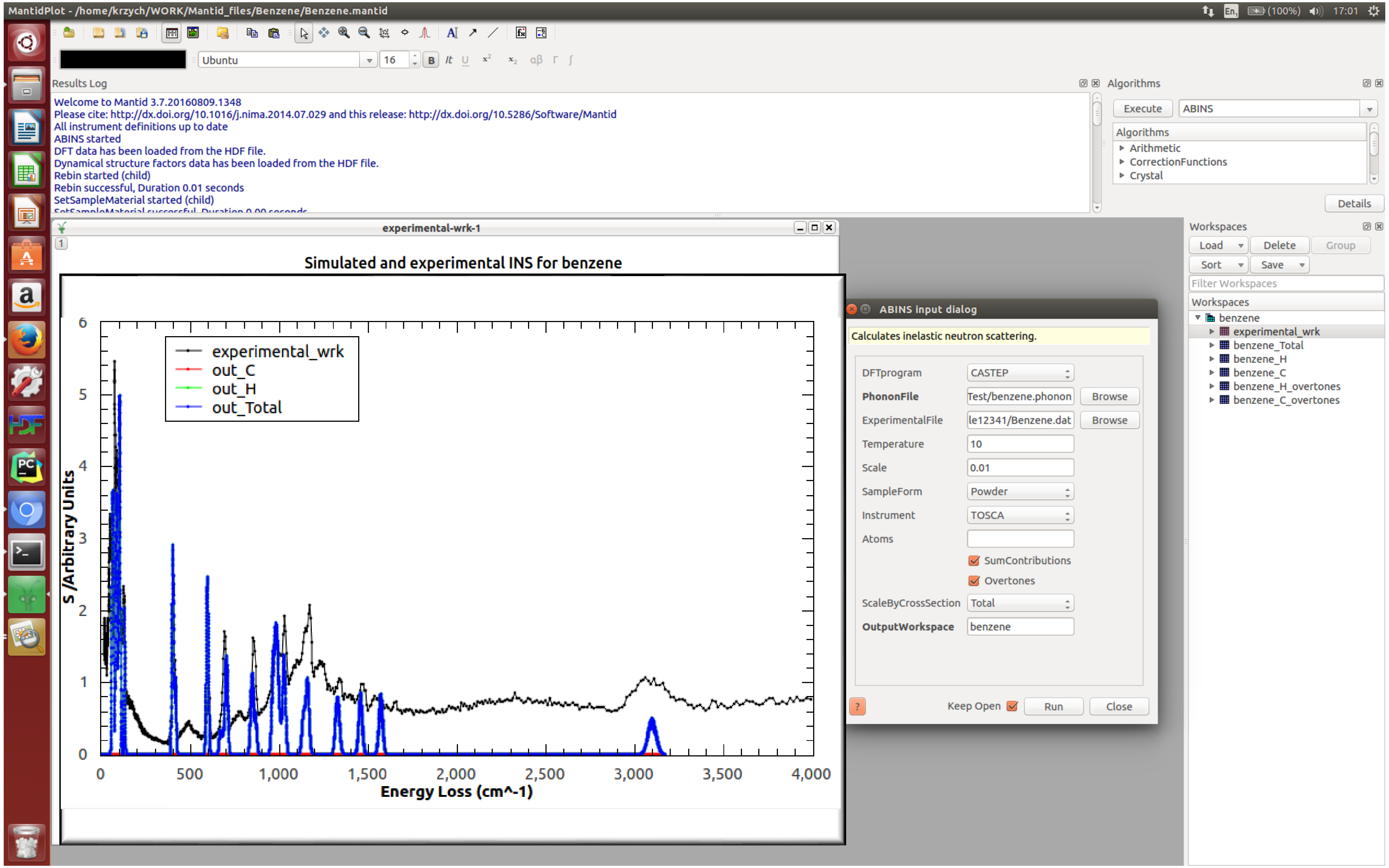


- INS spectra can be computed for a given atomic structure
- Calculations allow us to see what specific vibrational motion of atoms occur, and at what frequency



L. Liborio

Materials Workbench

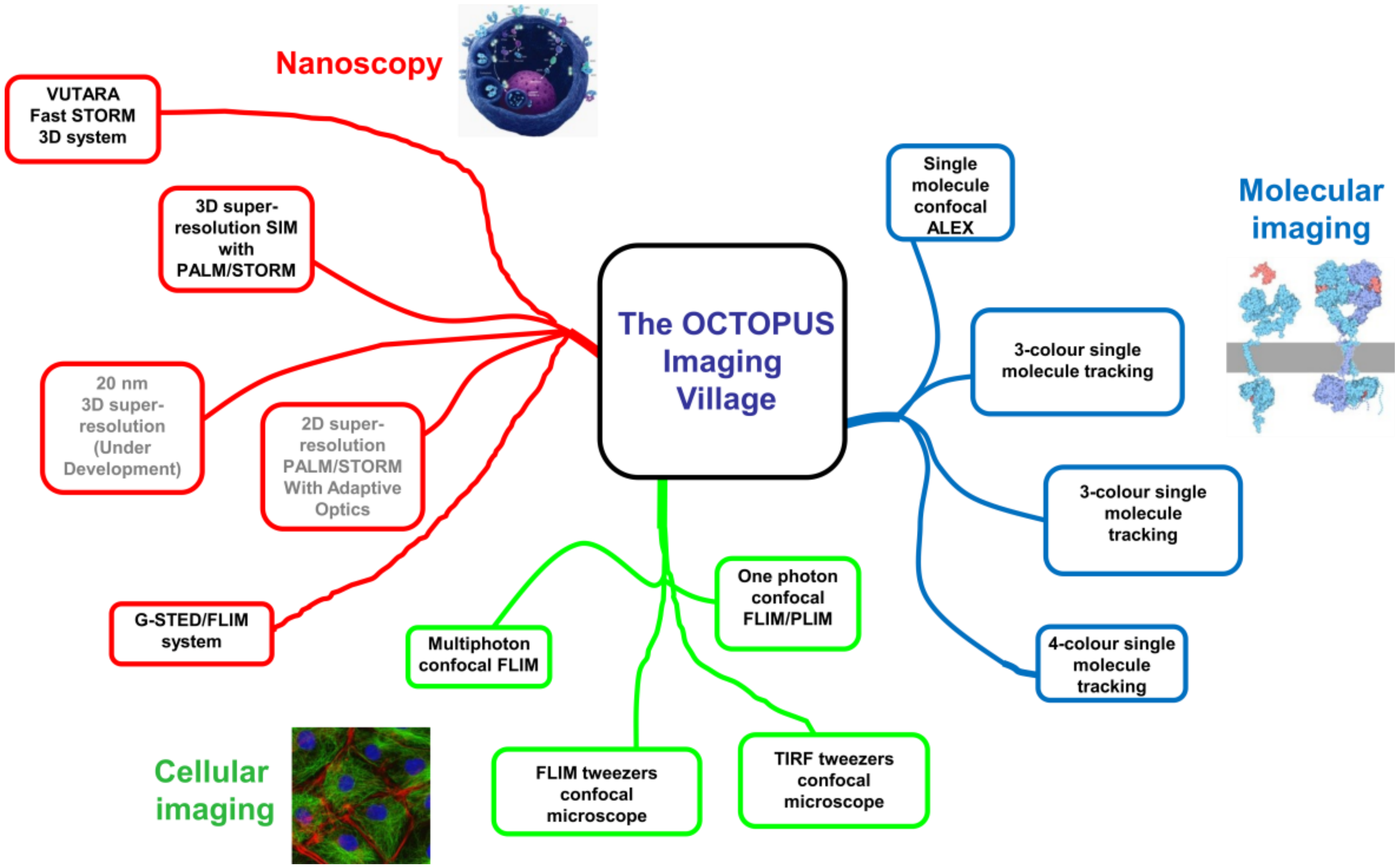


K. Dymkowski

The Central Laser Facility

OCTOPUS Facility in the CLF

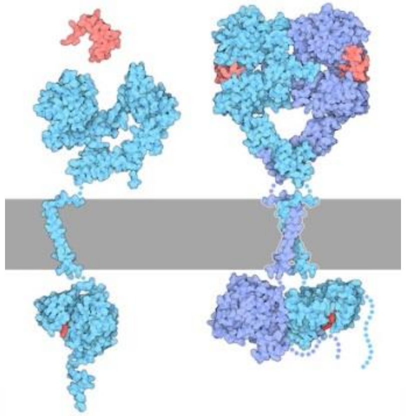
- National imaging facility with peer-reviewed, funded access
- Located in Research Complex at Harwell
- Cluster of microscopes and lasers and expert end-to-end multidisciplinary support
- Operations and some development funded by STFC
- Key developments funded through external grant – BBSRC, MRC

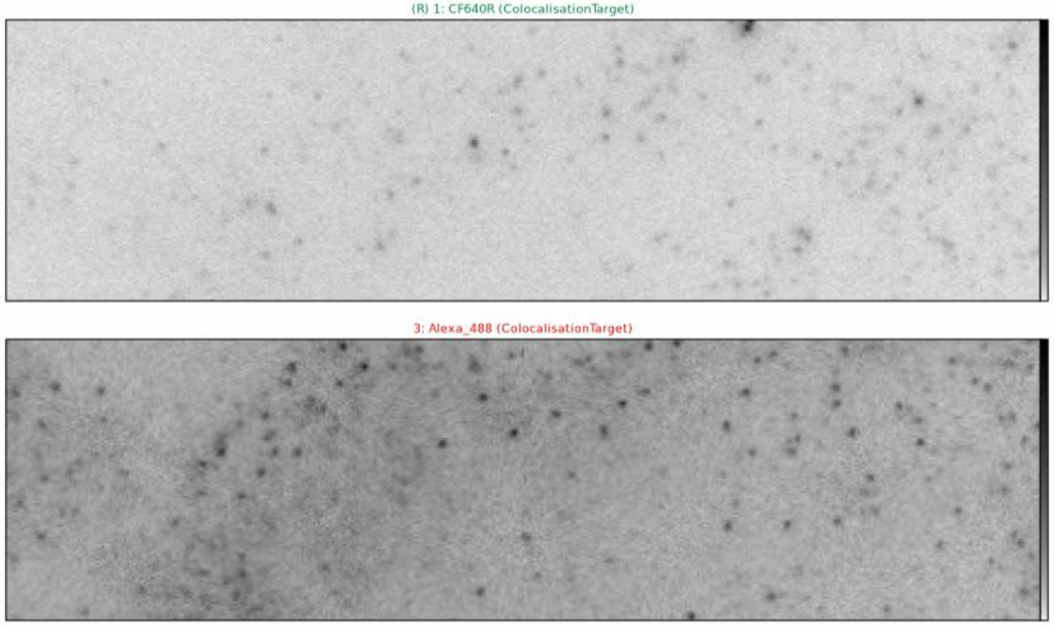


With thanks to Dan Rolfe

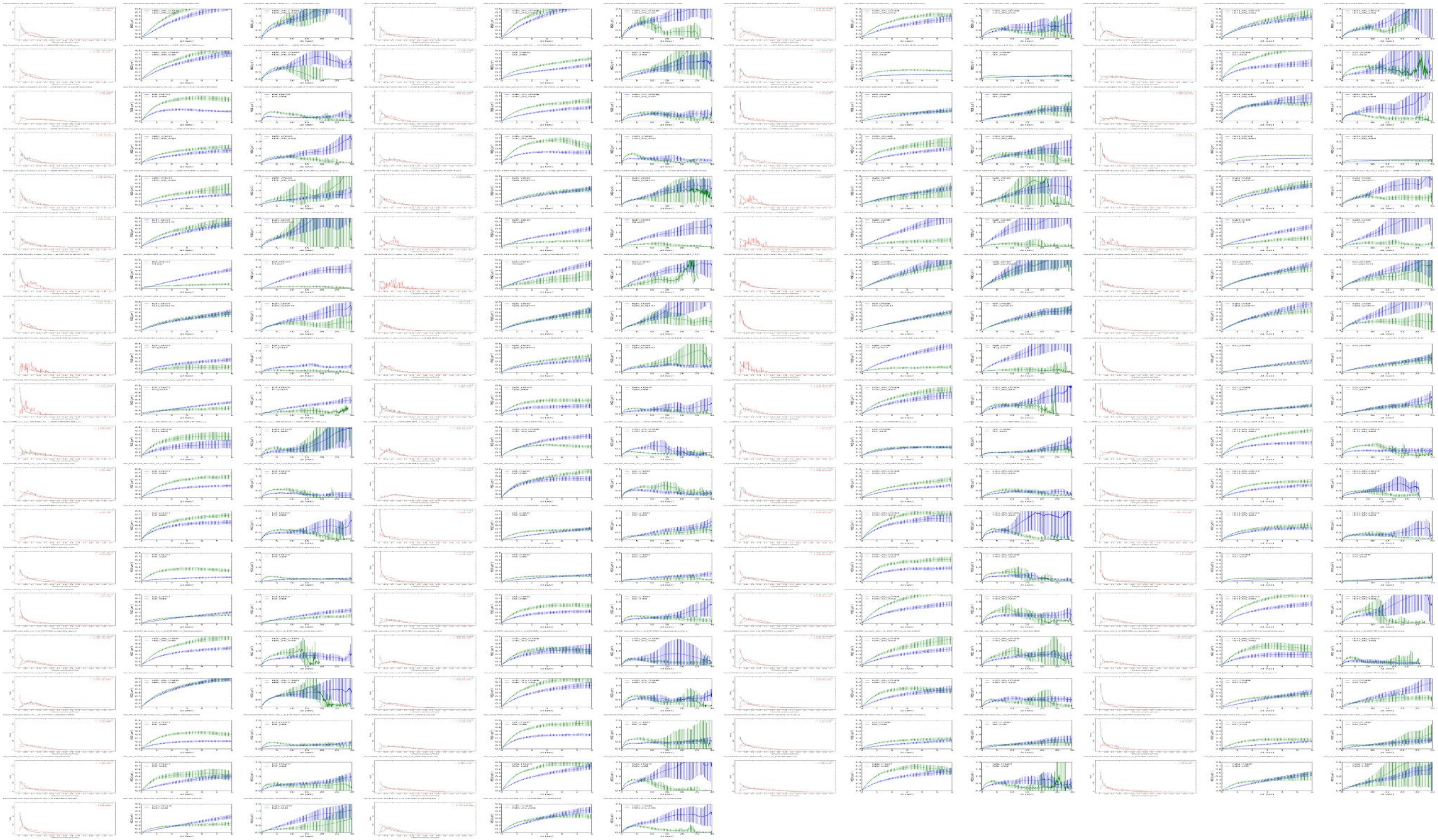
Example: EGFR cell signalling in cancer

- Driven OCTOPUS single molecule developments
- User in plant cell imaging now catching up in scale of challenge
- Part of a PhD project:
  - 1 experimental technique
  - 50 experimental conditions
  - 30 datasets for each condition
  - 1000 single molecule tracks for each condition
  - Multiple properties & events of interest in each track
  - Comparison of just one property…

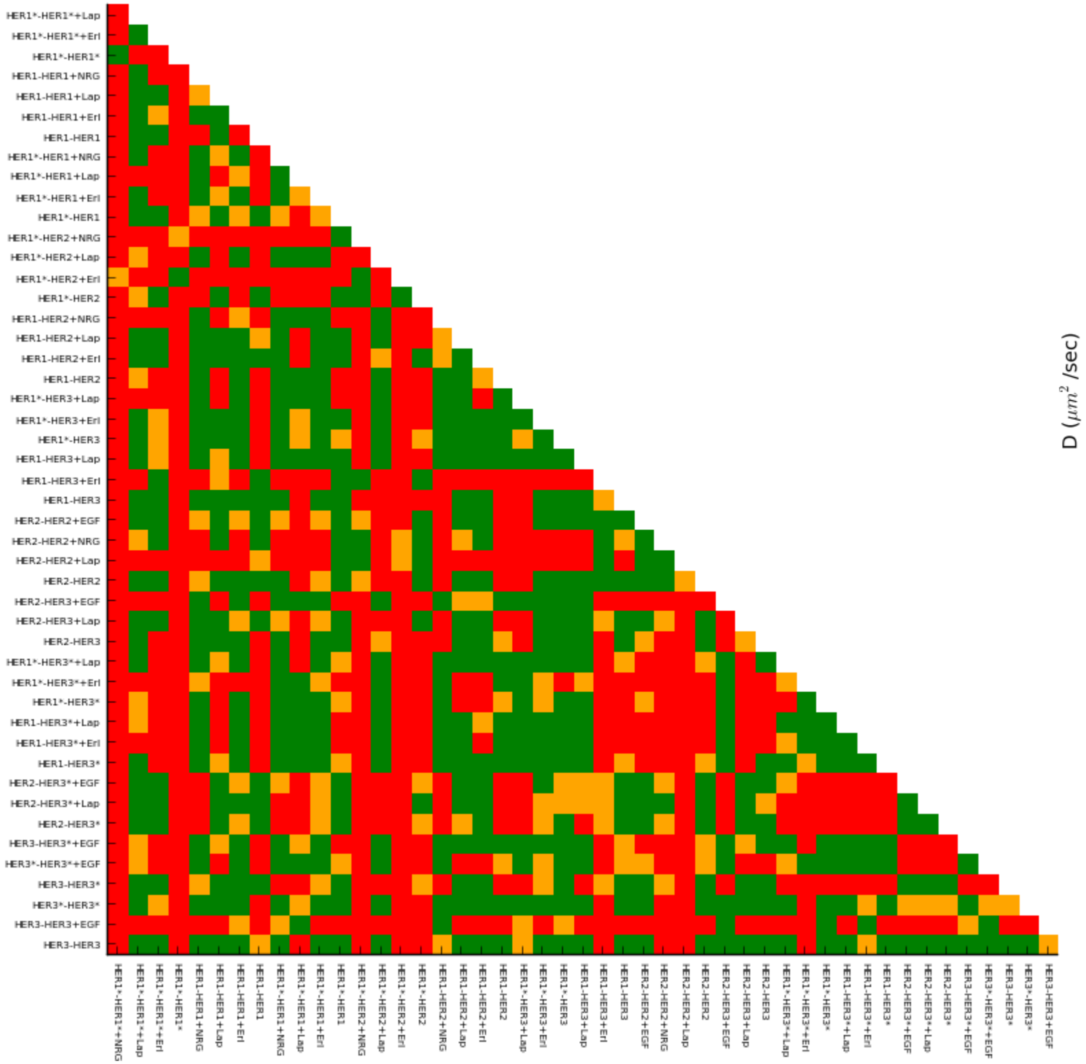


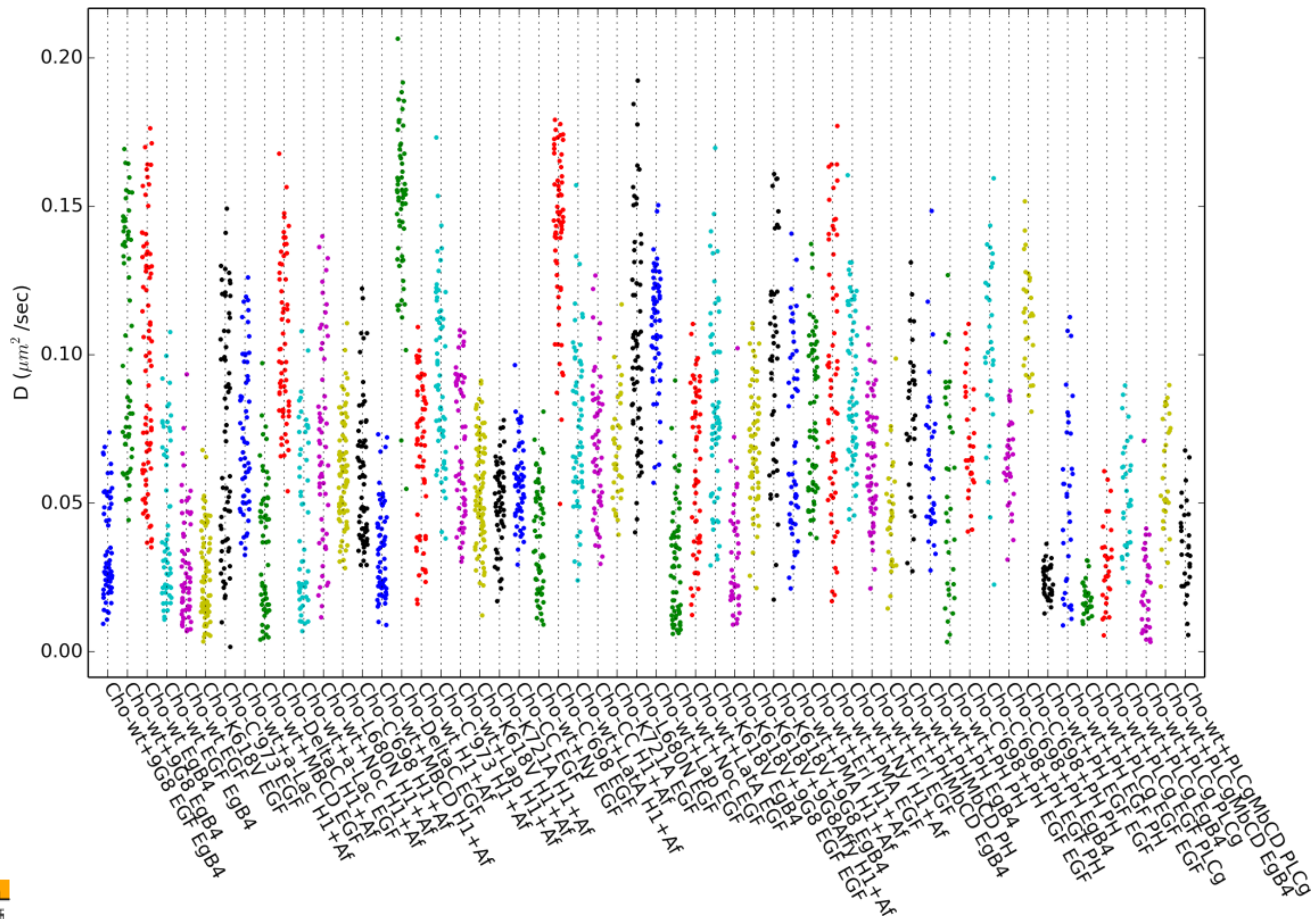


With thanks to Dan Rolfe



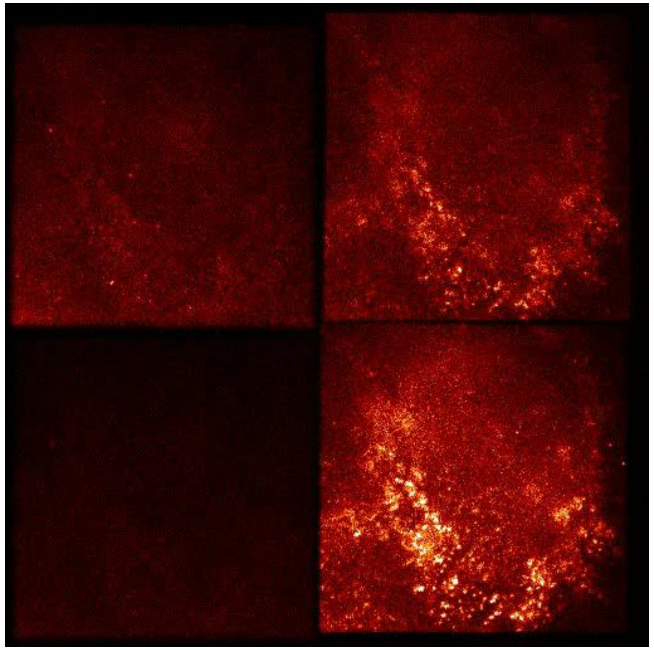
Large scale comparisons

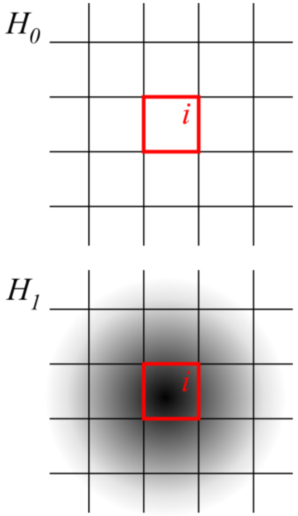


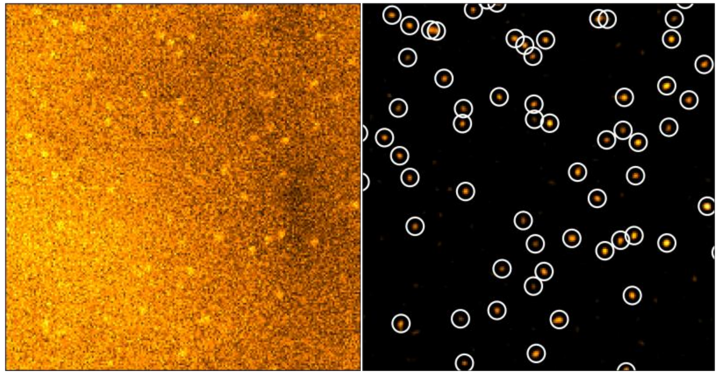


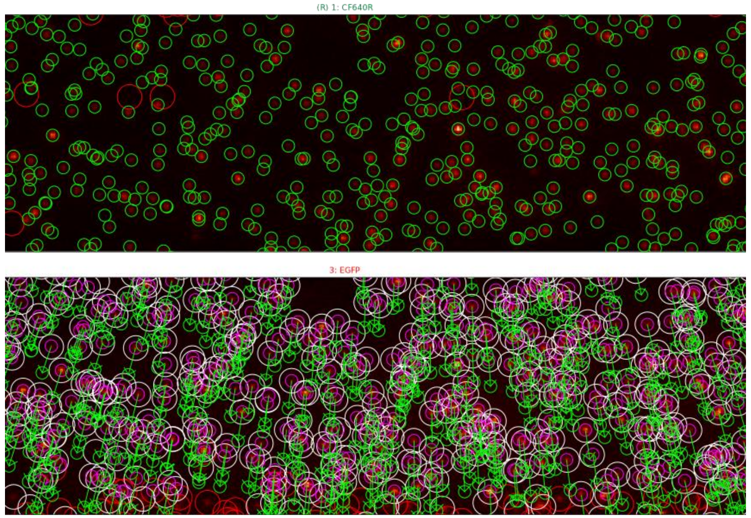
With thanks to Dan Rolfe

Multidimensional single molecule tracking









- Automated registration & tracking in multiple channels
  - Computer vision
  - Bayesian feature detection from astronomical galaxy detection
- Instrumental metadata from acquisition
  - Flexible specification of many instrument configurations

Rolfe et al 2011, *Euro Biophys J, 2011*

With thanks to Dan Rolfe

The JASMIN Environmental Science Super Data Cluster

Centre for Environmental Data Archival

SCIENCE AND TECHNOLOGY FACILITIES COUNCIL
NATURAL ENVIRONMENT RESEARCH COUNCIL

Data Centres

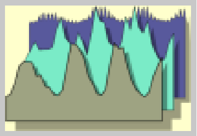
The Centre for Environmental Data Archival is responsible for the running of the following data centres:

The British Atmospheric Data Centre

The British Atmospheric Data Centre (BADC), NERC's designated data centre for the UK atmospheric science community, covering climate, composition, observations and NWP data.

NERC Earth Observation Data Centre

The NEODC is NERC's designated data centre for Earth Observation data and is part of NERC's National Centre for Earth Observation.



The UK Solar System Data Centre

The UK Solar System Data Centre, co-funded by STFC and NERC, curates and provides access to archives of data from the upper atmosphere, ionosphere and Earth's solar environment.

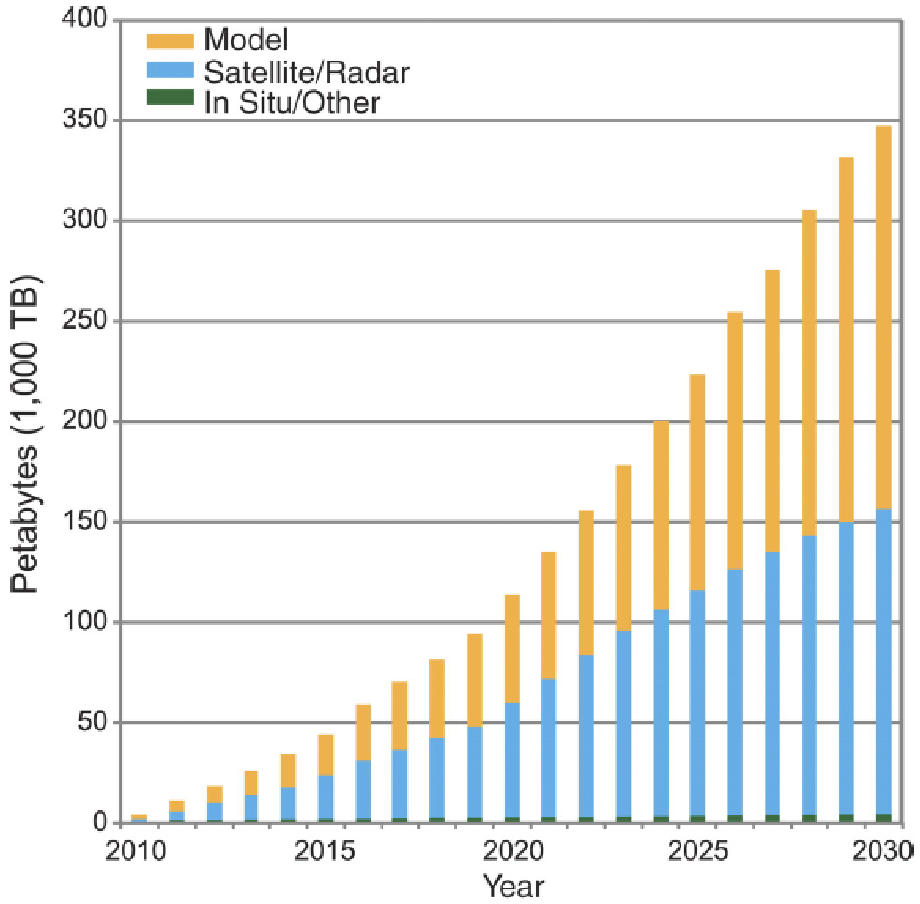
IPCC Data Distribution Centre

The Intergovernmental Panel on Climate Change (IPCC) DDC provides climate, socio-economic and environmental data, both from the past and also in scenarios projected into the future. Technical guidelines on the selection and use of different types of data and scenarios in research and assessment are also provided.

More Data

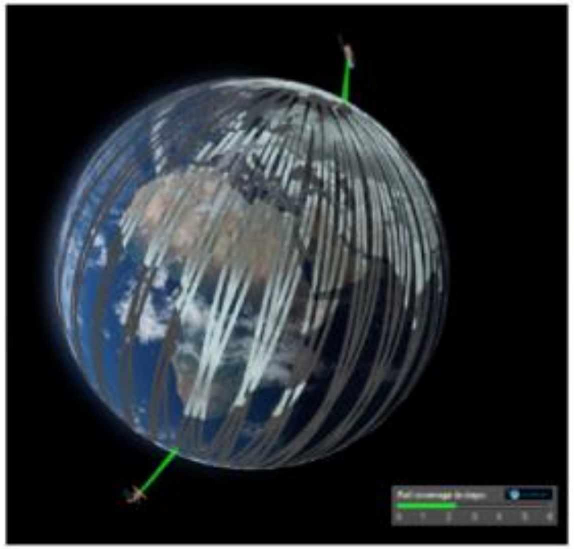
Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

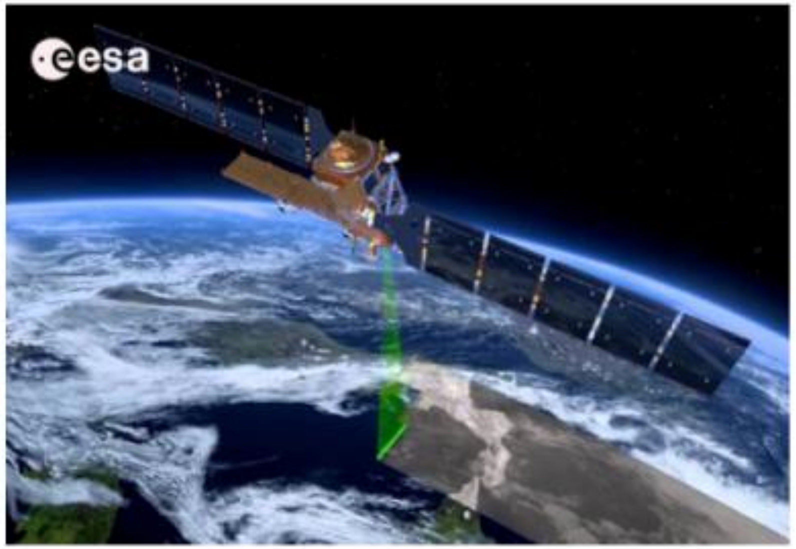
(BNL: Even if you are?)



J T Overpeck et al. Science 2011;331:700-702

Large data sets: satellite observations



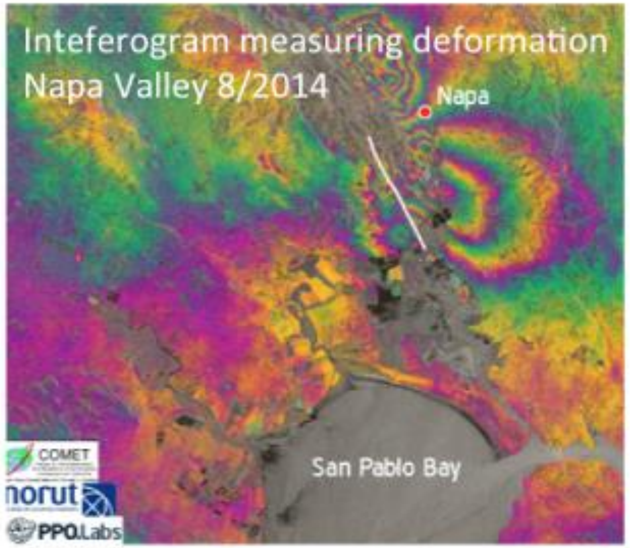


Sentinel 1A: Launched 2014
(1B due 2016)

- Key instrument: Synthetic Aperture Radar
- Data rate (two satellites: raw 1.8 TB/day, archive products ~ 2 PB/year)

NERC SCIENCE OF THE ENVIRONMENT

COMET: Centre for Observation and Modelling of Earthquakes, Volcanoes, and Tectonics



(Picture credits: ESA, Arianespace.com, PPO.labs-Norut–COMET-SEOM Insarap study, ewf.nerc.ac.uk/2014/09/02/new-satellite-maps-out-napa-valley-earthquake/)

Core Science Requirements

Today:	**Observations**	**Models**
Volume	20 million = **2×10^7**	5 million grid points 100 levels 10 prognostic variables = **5×10^9**
Type	98% from 60 different satellite instruments	physical parameters of atmosphere, waves, ocean

Soon:	**Observations**	**Models**
Volume	200 million = **2×10^8**	500 million grid points 200 levels 100 prognostic variables = **1×10^{13}**
Type	98% from 80 different satellite instruments	physical and chemical parameters of atmosphere, waves, ocean, ice, vegetation

→ Factor 10 per day

→ Factor 2000 per time step

→ but many more time steps needed

Big International Drivers:

WWRP World Weather Research Programme

GAW

Copernicus Europe's eyes on Earth

WCRP World Climate Research Programme

- aerosol cci
- cloud cci
- fire cci
- ghg cci
- glaciers cci
- antarctic ice sheet cci
- ice sheets greenland cci
- land cover cci
- ocean colour cci
- ozone cci
- sea ice cci
- sea level cci
- sst cci
- soil moisture cci
- cmug cci

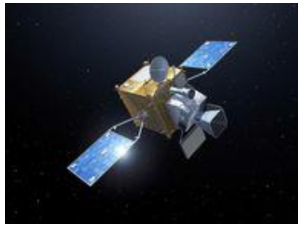
Why JASMIN?

- Urgency to provide better environmental predictions
- Need for higher-resolution models
- HPC to perform the computation
- Huge increase in observational capability/capacity

But…

- Massive storage requirement: observational data transfer, storage, processing
- Massive raw data output from prediction models
- Huge requirement to process raw model output into usable predictions (post-processing)

Hence JASMIN…

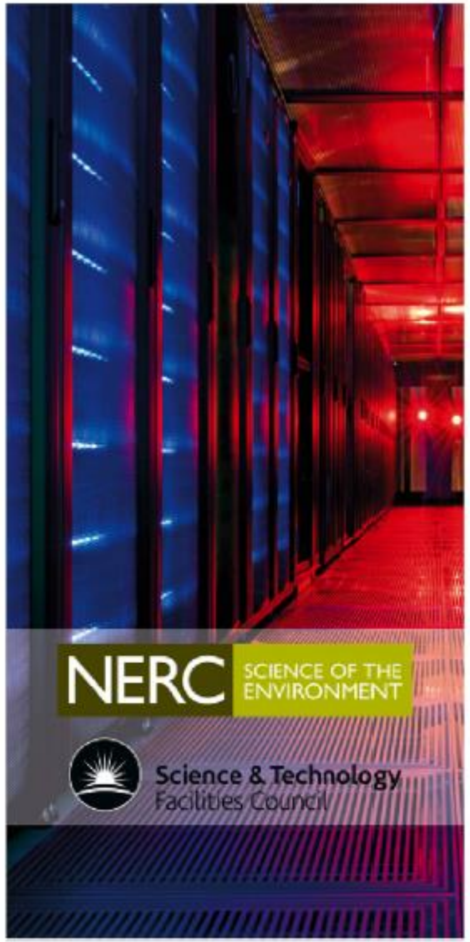


ARCHER supercomputer (EPSRC/NERC)

JAMSIN (STFC/Stephen Kill)

JASMIN infrastructure

Part data store, part HPC cluster, part private cloud…

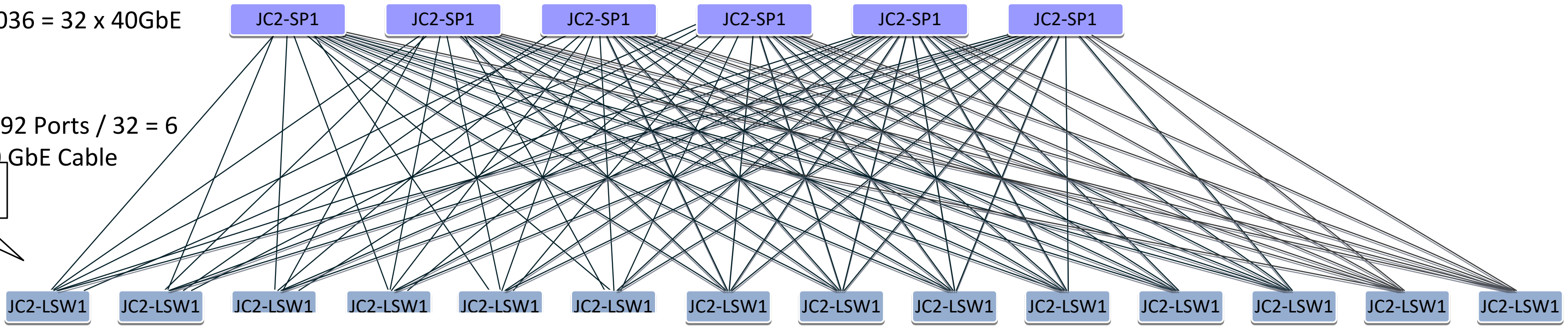


- 16 PB Fast Storage (Panasas, many Tbit/s bandwidth)
- 1 PB Bulk Storage
- Elastic Tape
- 4000 cores: half deployed as hypervisors, half as the "Lotus" batch cluster.
- Some high memory nodes, a range, bottom heavy.

Non-blocking, low latency, CLOS Tree Network

954 Routes

S1036 = 32 x 40GbE



16 x 12 40GbE = 192 Ports / 32 = 6
Total 192 40 GbE Cable

954 Routes

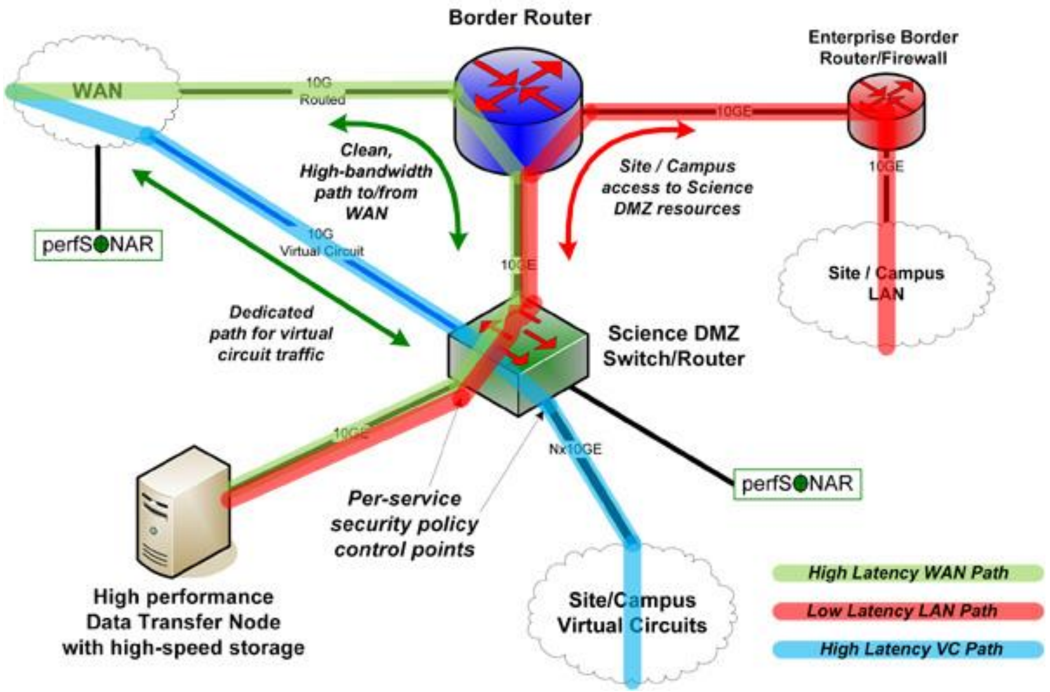
16 x MSX1024B-1BFS
48x10GBE + 12 40 GbE

48 * 16 = 768 10GbE Non-blocking
16 x 12 x 40GbE = 192 40GbE ports

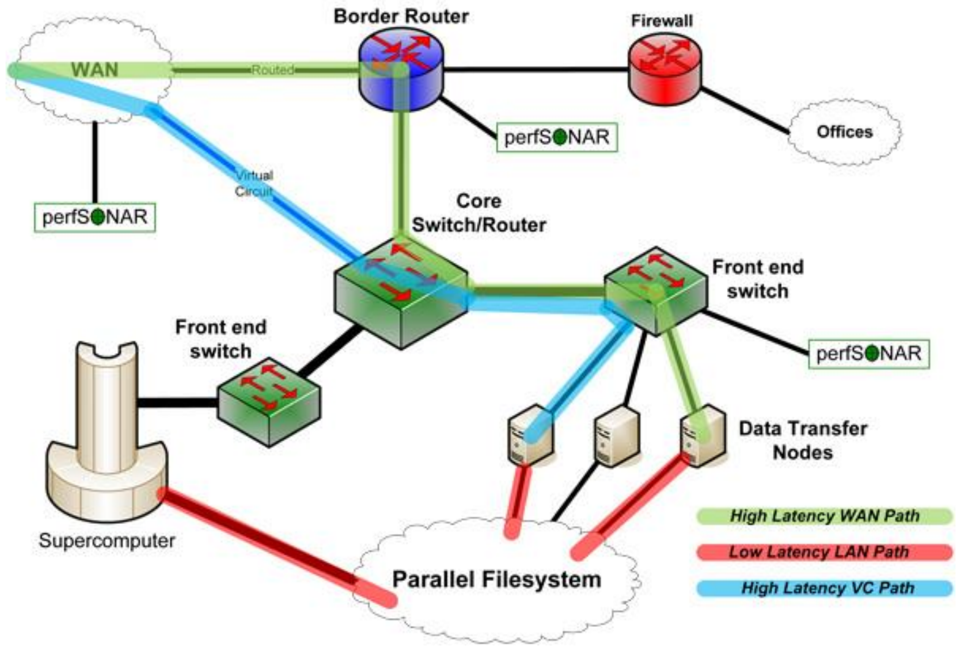
1,104 x 10GbE Ports CLOS L3 ECMP OSPF

- ~1,200 Ports expansion
- Max 36 leaf switches :1,728 Ports @ 10GbE
- Non-Blocking, Zero Contention (48x10Gb = 12x 40Gb uplinks)
- Low Latency (250nS L3 / per switch/router) 7-10uS MPI

JASMIN "Science DMZ" Architecture



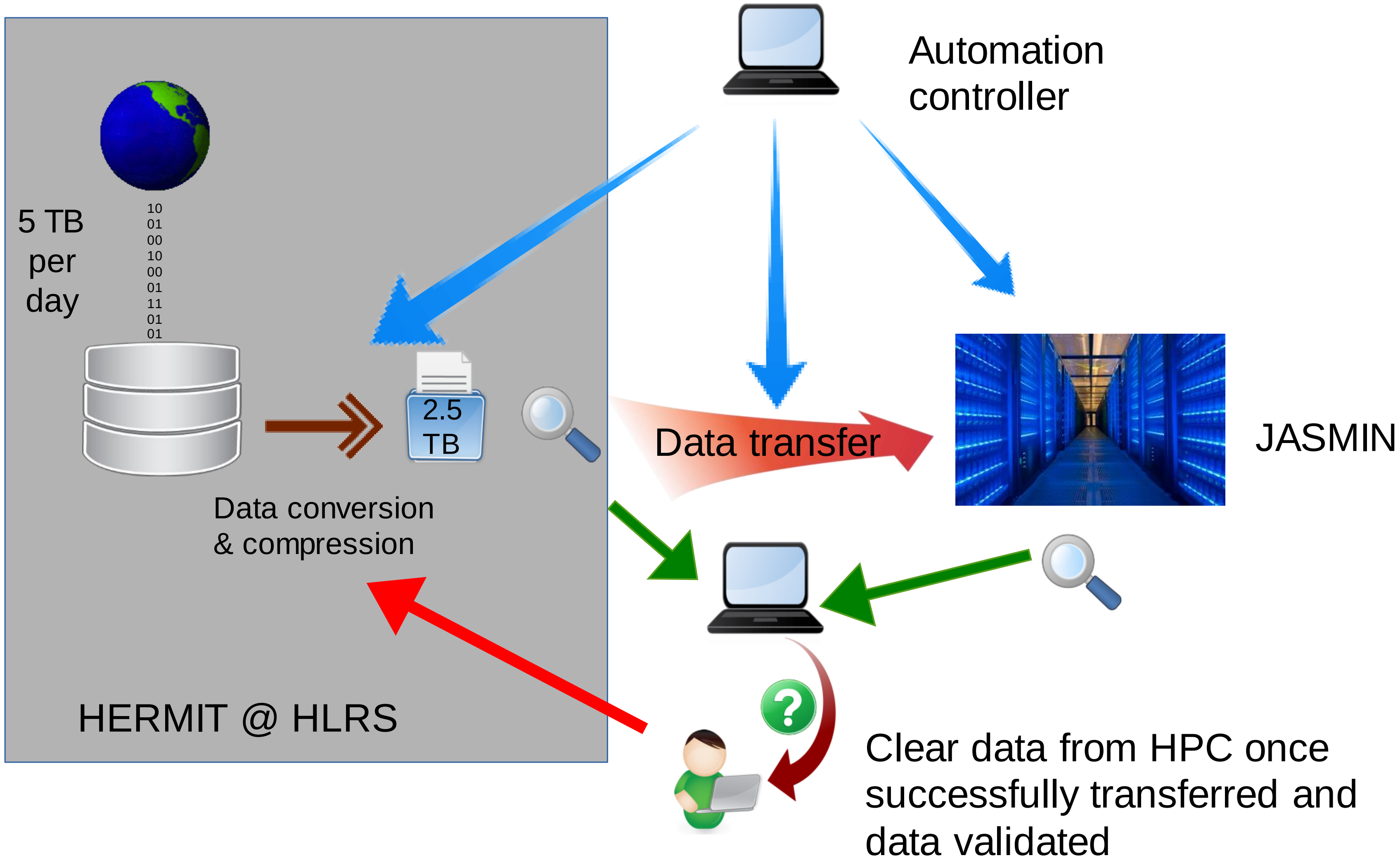
Simple Science DMZ



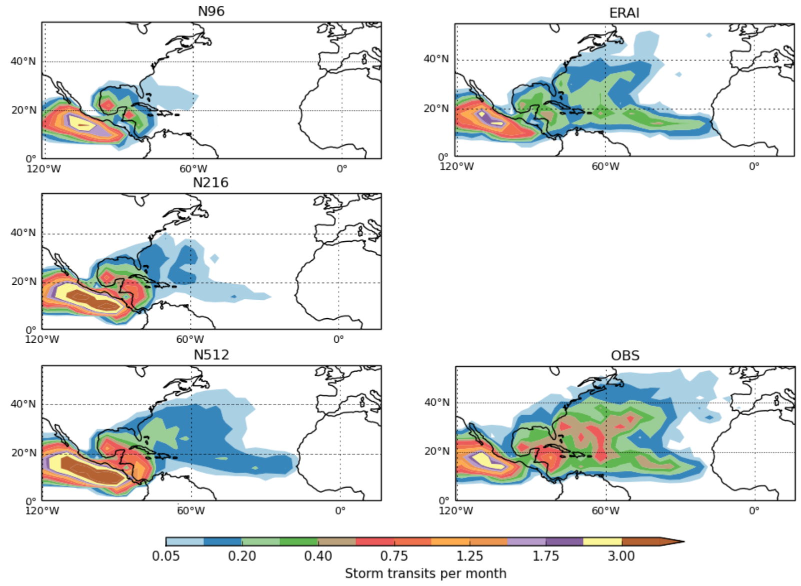
Supercomputer Center

http://fasterdata.es.net/science-dmz-architecture

The UK Met Office UPSCALE campaign



Example Data Analysis



- Tropical cyclone tracking has become routine; 50 years of N512 data can be processed in 50 jobs in one day
- Eddy vectors; analysis we would not attempt on a server/workstation (total of 3 months of processor time and ~40 GB memory needed) completed in 24 hours in 1,600 batch jobs
- JASMIN/LOTUS combination has clearly demonstrated the value of cluster computing to data processing and analysis.

M Roberts et al: Journal of Climate 28 (2), 574-596

The Ada Lovelace Center

ada lovelace centre

The Experimental Data Challenge

- 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

➢ Need 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

Ada Lovelace Centre

The ALC will significantly enhance our capability to support the Facilities' science programme:

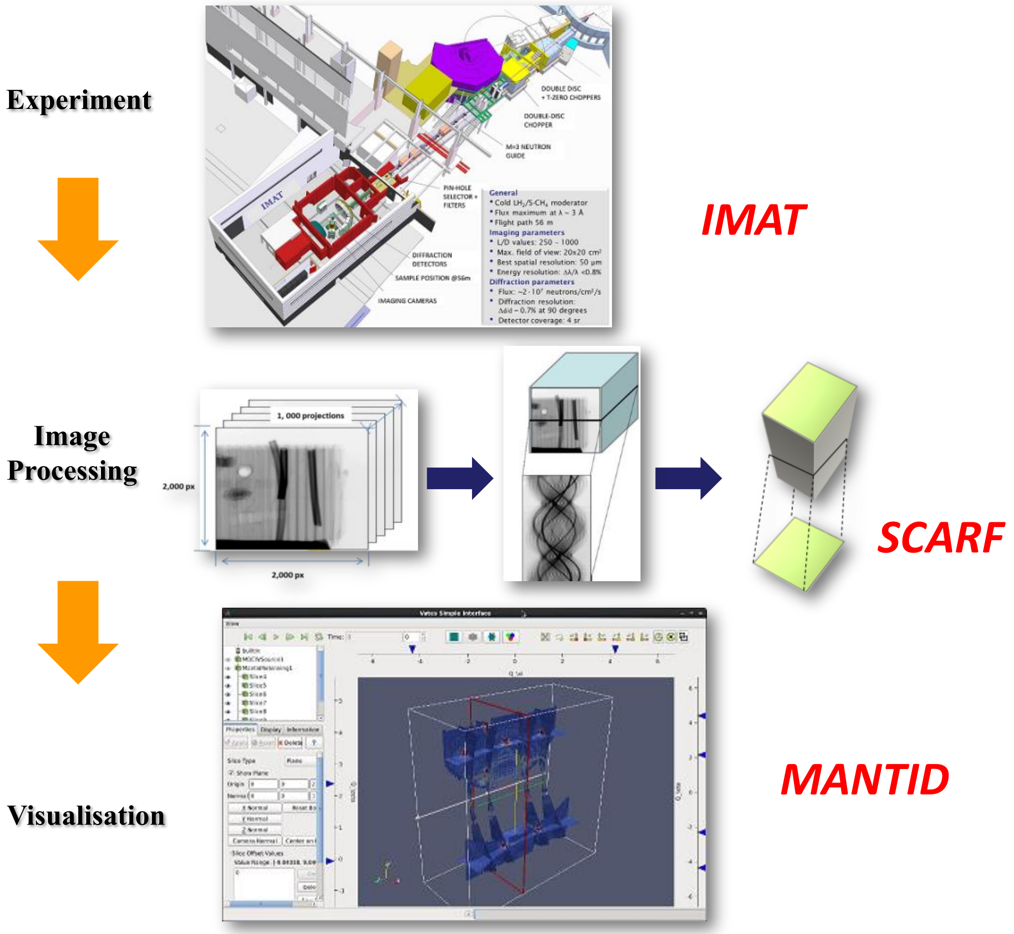
- *Theme 1: Increases capacity in advanced software development for data analysis and interpretation*

- *Theme 2: Develop new generation of scientific data experts and scientific software engineers who can interact with science domain experts*

- *Theme 3: Provide significant compute infrastructure for managing, analysing and simulating the data generated by the facilities and for designing next generation Big-Science experiments*

➢ Focus is the science drivers and computational needs of Facilities

ALC Pathfinder: Tomographic Reconstruction



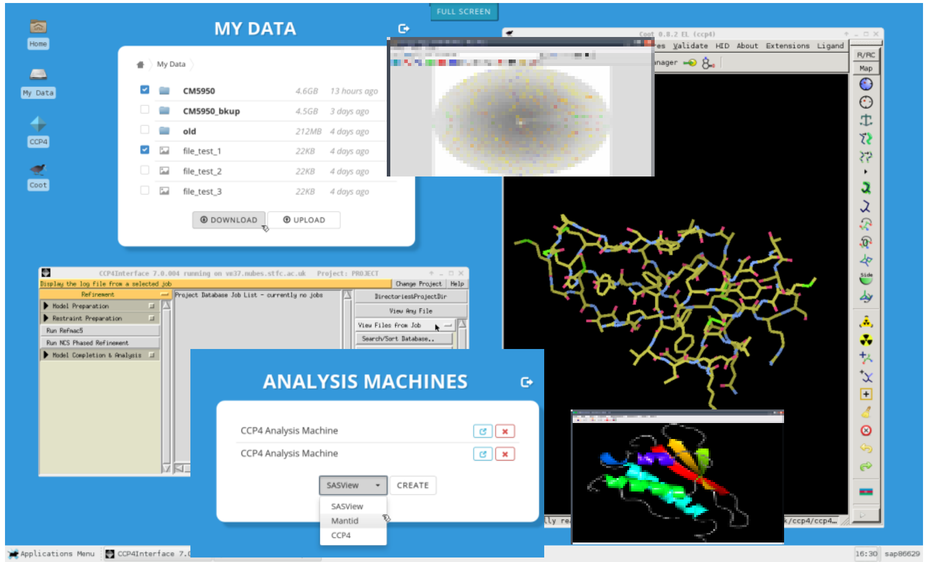
- Support in-experiment and post-experiment tomographic reconstruction
  - Round-trip the data to HPC CPU/GPU clusters in experiment time
  - Tomographic image reconstruction toolbox with different algorithms
  - High throughput image reconstruction framework – time scheduled
  - Visualisation on the beamline or remote
  - An integral component of IMAT’s in-experiment data analysis capability through the ISIS Mantid software suite
- Goal is to maximise the science from data collected on facility instruments

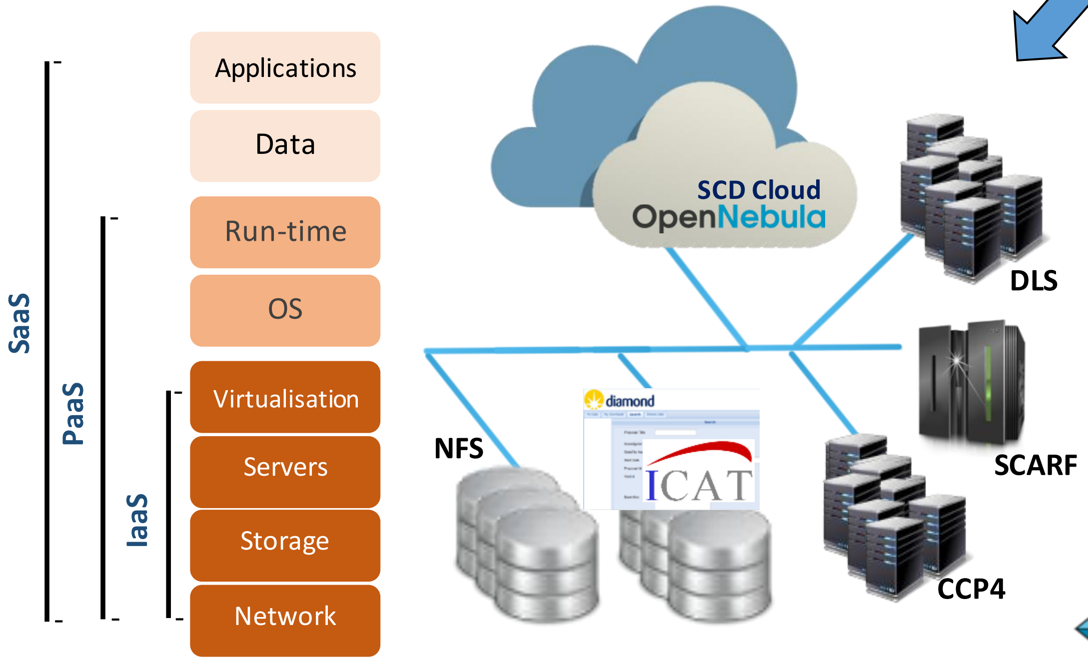
STFC Scientific Computing: Erica Yang, Srikanth Nagella, Martin Turner, Derek Ross
STFC ISIS: Winfried Kockelmann, Genoveva Burca, Federico Montesino Pouzols
DLS: Mark Basham

ALC Pathfinder: CCP4-DAaaS

CCP4 – Macro-Crystallography suite

- proteins, viruses and nucleic acids
- determine macromolecular structures by X-ray crystallography
- Used by DLS users
  - But need post-experimental access





Data Analysis as a Service

- Remote access to data and compute via SCD Cloud
- CCP4 s/w maintained on Cloud via VM packaging and distribution (CVMFS)
- User Portal provides access to right data and compute and workflows

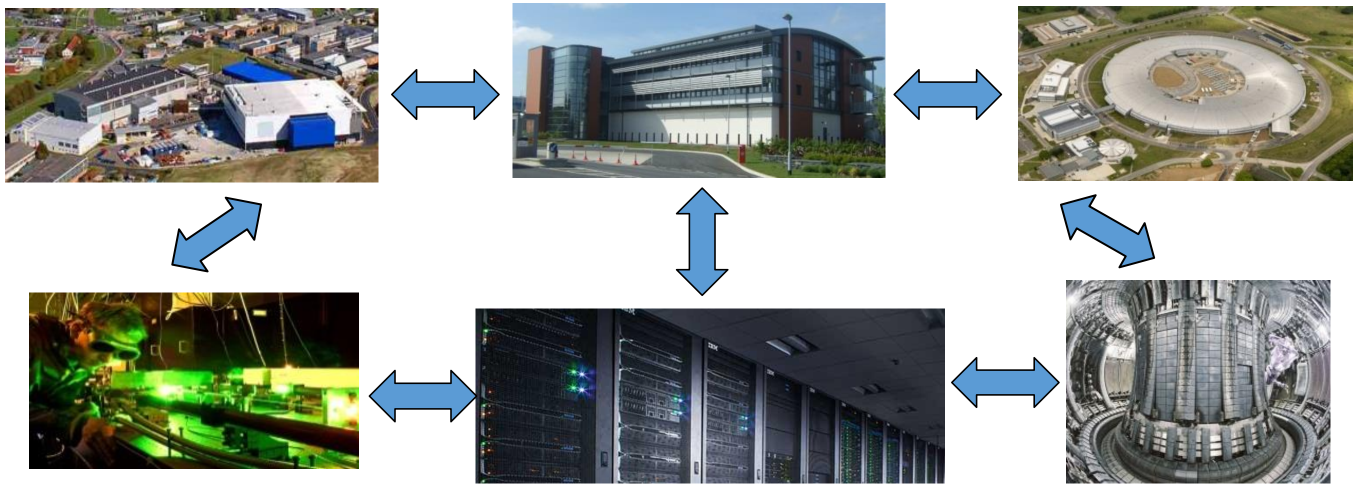
CCP4

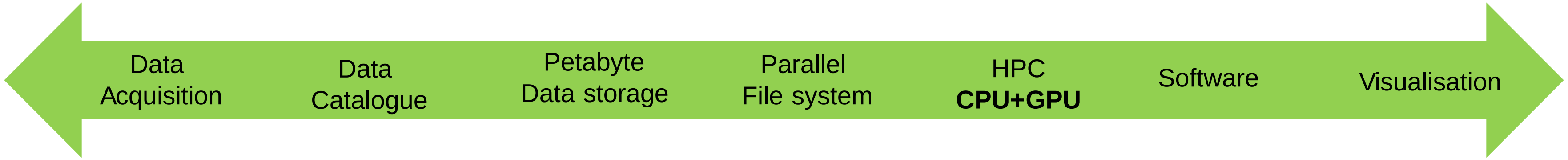
Frazer Barnsley, Shirley Crompton, CCP4, et al

The ALC - Towards a “Super-facility”?

Infrastructure + Software + Expertise

With Common Interfaces and Transparent Access



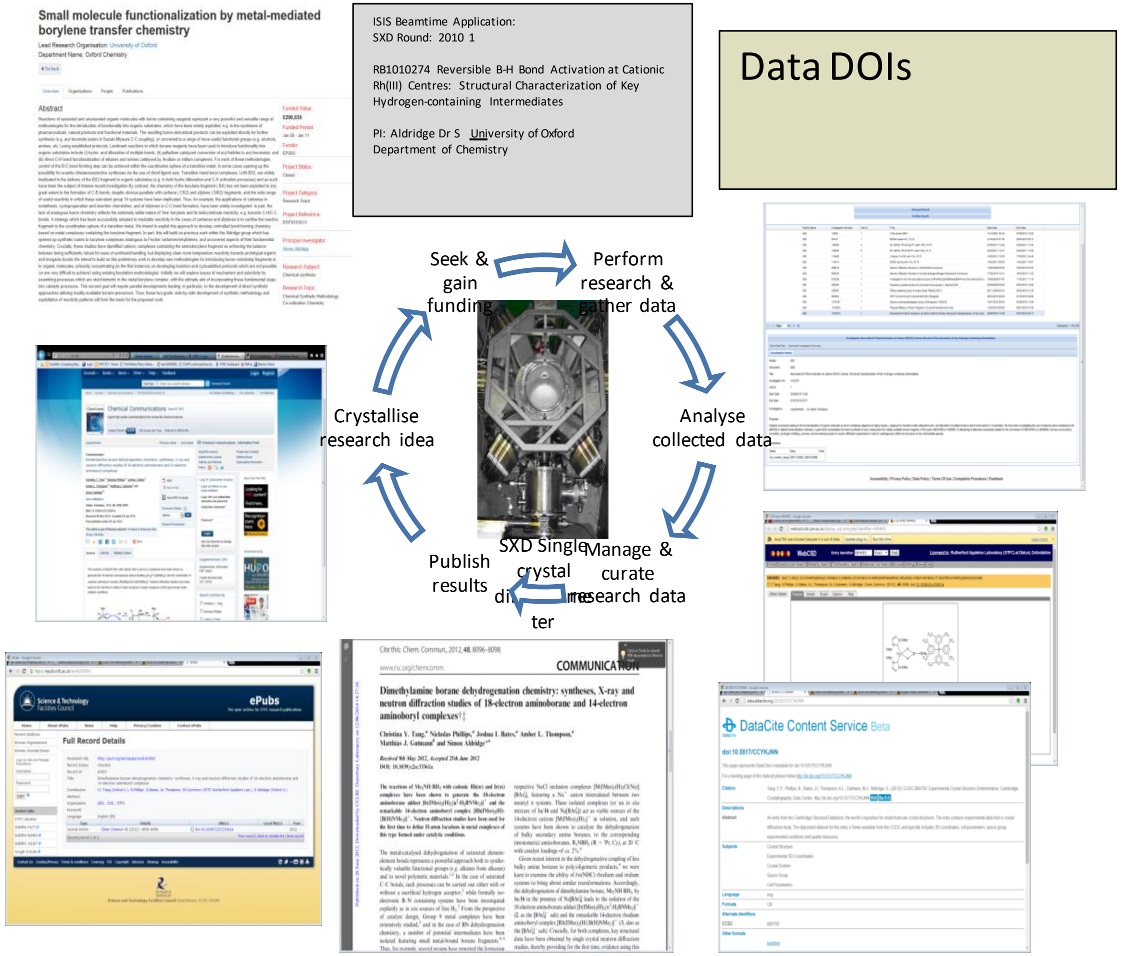


“A network of connected facilities, software and expertise to enable new modes of discovery”

Katie Antypas, Inder Monga, Lawrence Berkeley National Laboratory

New Opportunities: Reproducible Science

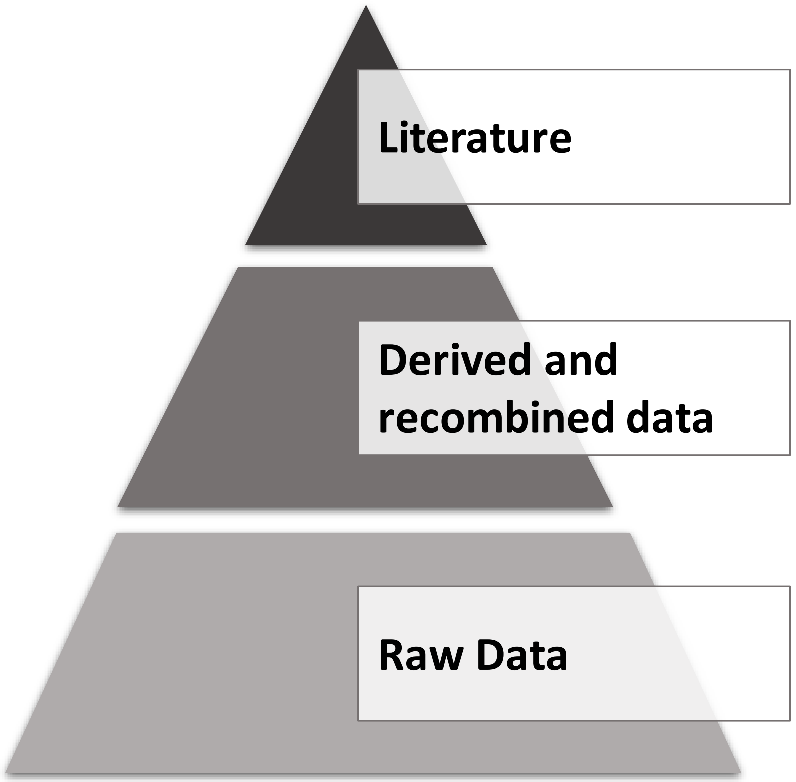
- Traceable science
  - Preservation
  - Provenance
  - Publishing
- A tool for the user
  - Tracking progress
- ‘RARE’ research
  - Robust
  - Accountable
  - Reproducible
  - Explainable



➢ ALC can build in support for reproducible science

Jim Gray's Vision: All Scientific Data Online

- Many disciplines overlap and use data from other sciences.
- Internet can unify all literature and data
- Go from literature *to* computation *to* data *back to* literature.
- Information at your fingertips – For everyone, everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity



(From Jim Gray's last talk)

Acknowledgements:

With thanks to Mark Basham, David Corney, Jonathan Churchill, Imanol Luengo, Barbara Montanari, Brian Matthews and Dan Rolfe