The Fourth Paradigm and Big Scientific Data

Professor Tony Hey

Chief Data Scientist

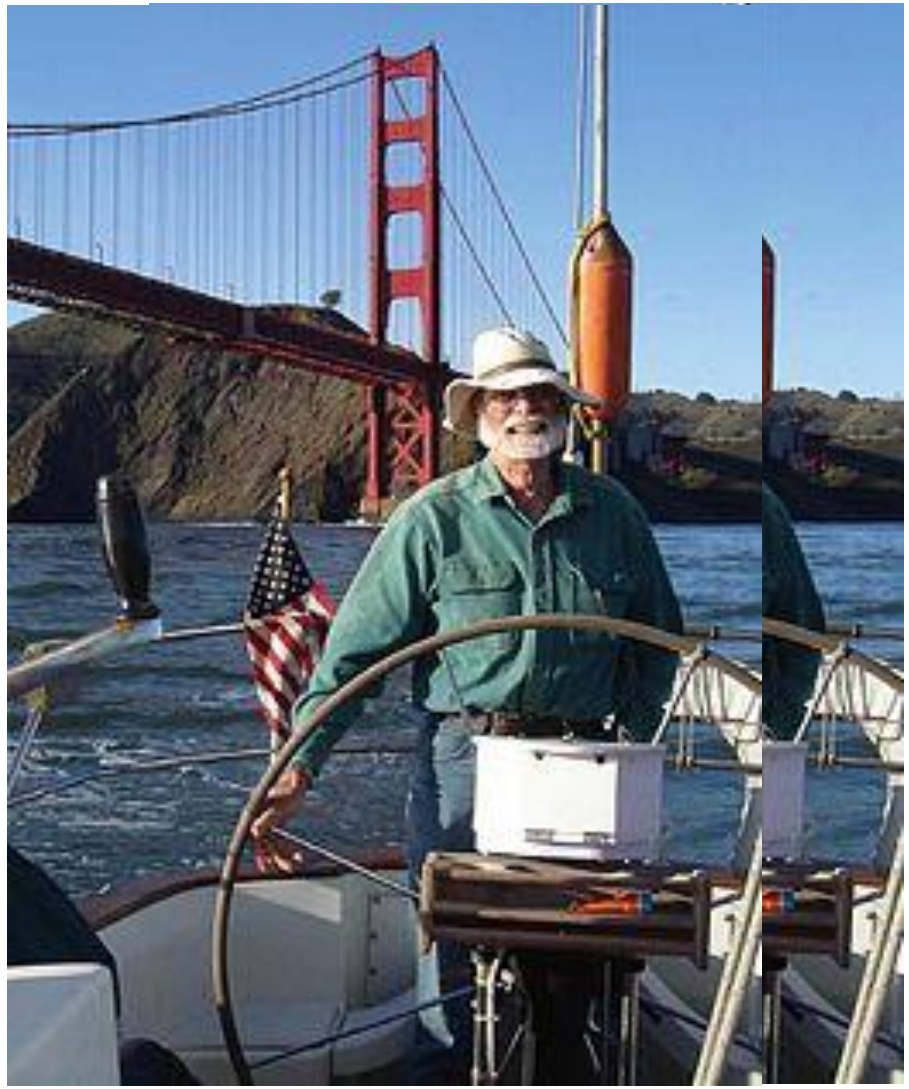
Rutherford Appleton Laboratory

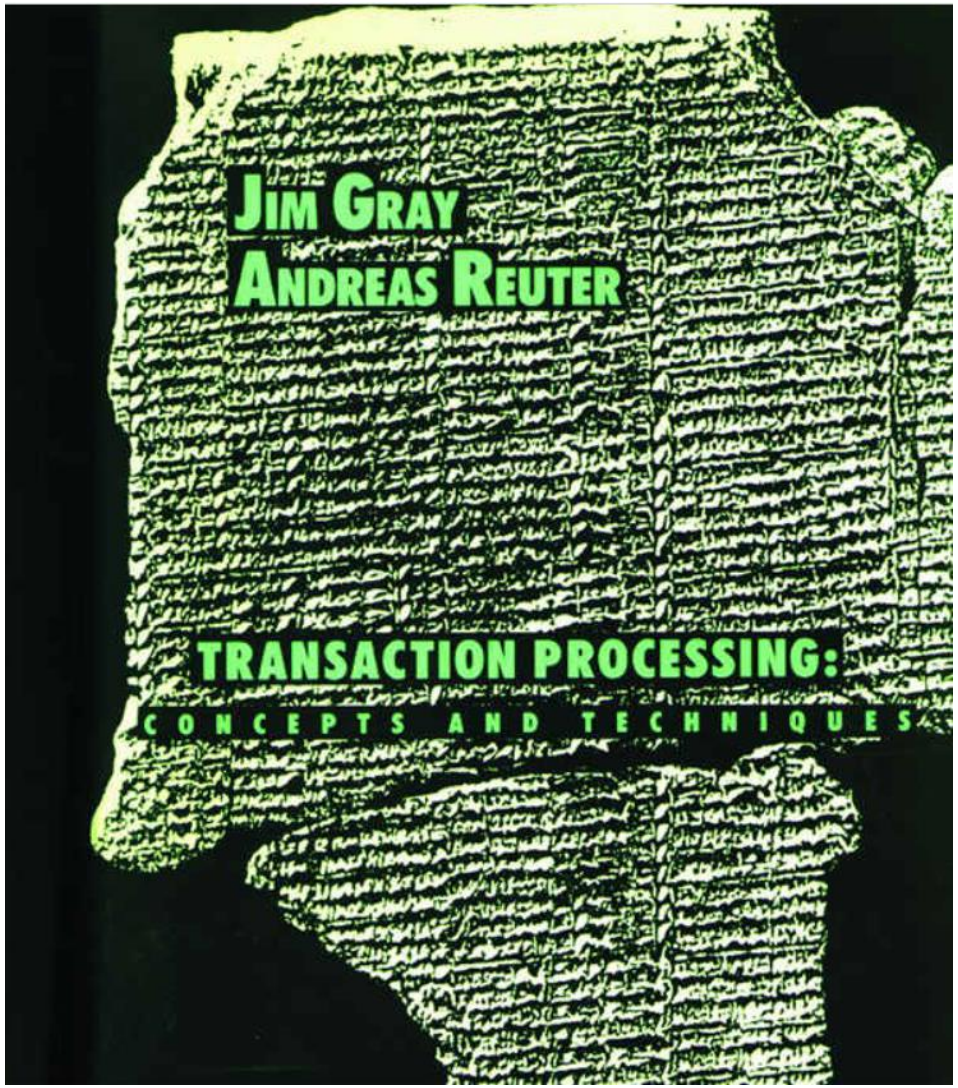
Science and Technology Facilities Council

tony.hey@stfc.ac.uk

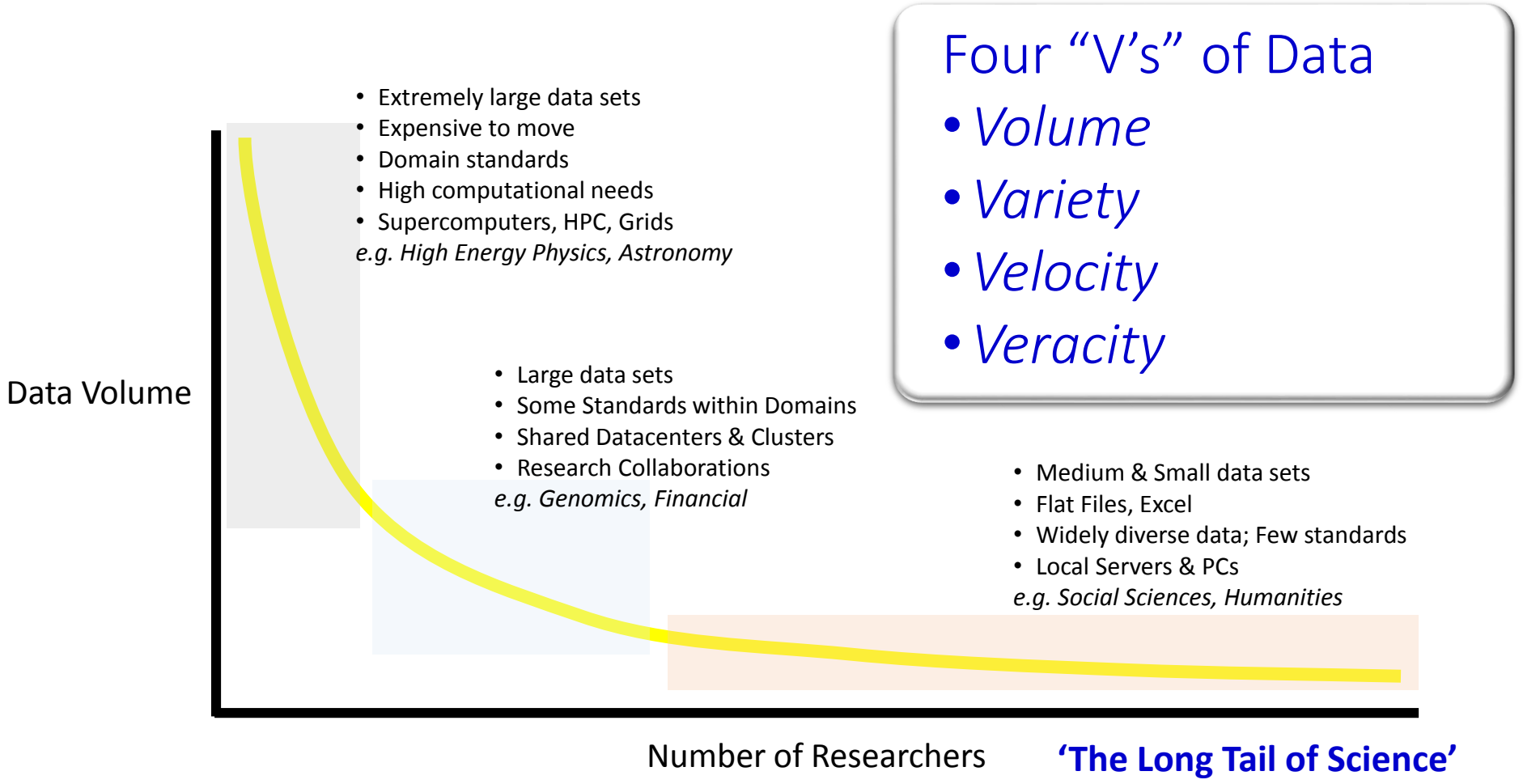
Jim Gray and the Fourth Paradigm

Jim Gray, Turing Award Winner





Much of Science is now Data-Intensive



The ‘Cosmic Genome Project’: The Sloan Digital Sky Survey

- Survey of more than ¼ of the night sky
- Survey produces 200 GB of data per night
- Two surveys in one – images and spectra
- Nearly 2M astronomical objects, including 800,000 galaxies, 100,000 quasars
- 100’s of TB of data, and data is public
- Started in 1992, ‘finished’ in 2008

➢ The SkyServer Web Service was built at JHU by team led by Alex Szalay and Jim Gray

The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA

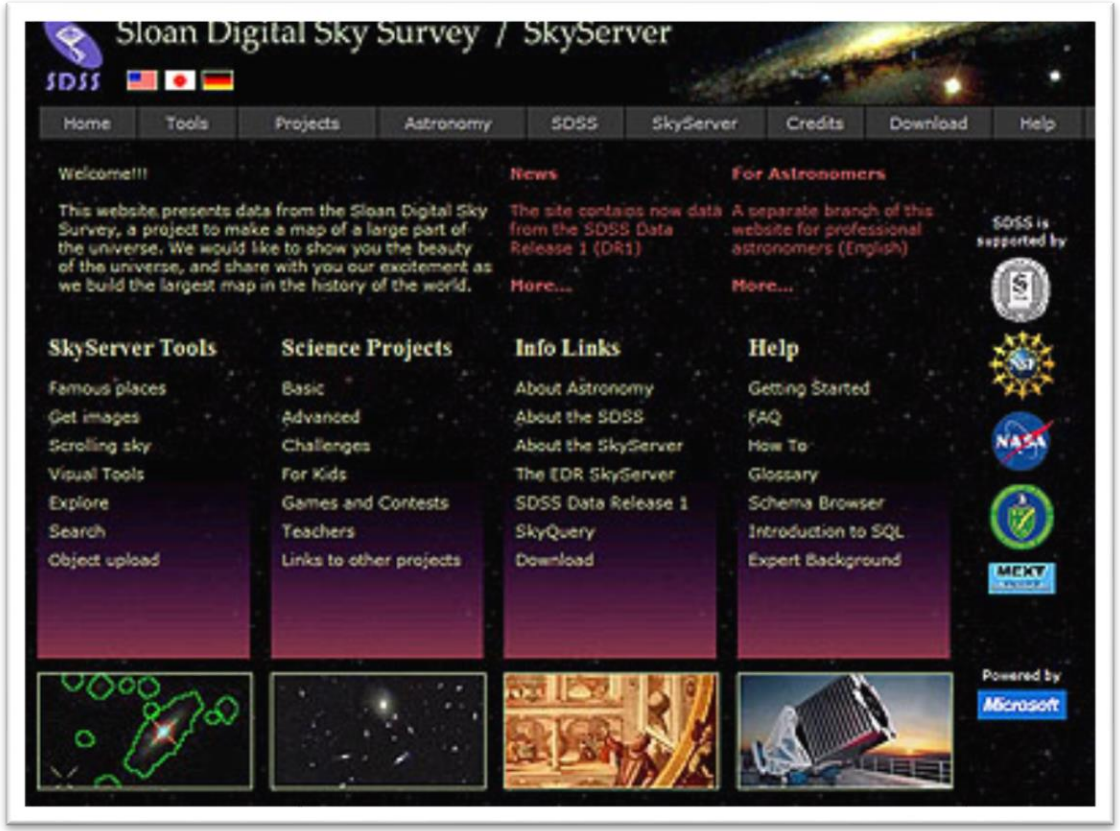
Open Data: Public Use of the Sloan Data

Posterchild for 21st century data publishing

- SkyServer web service has had over 400 million web
- About 1M distinct users vs 10,000 astronomers
- >1600 refereed papers!
- Delivered 50,000 hours of lectures to high schools

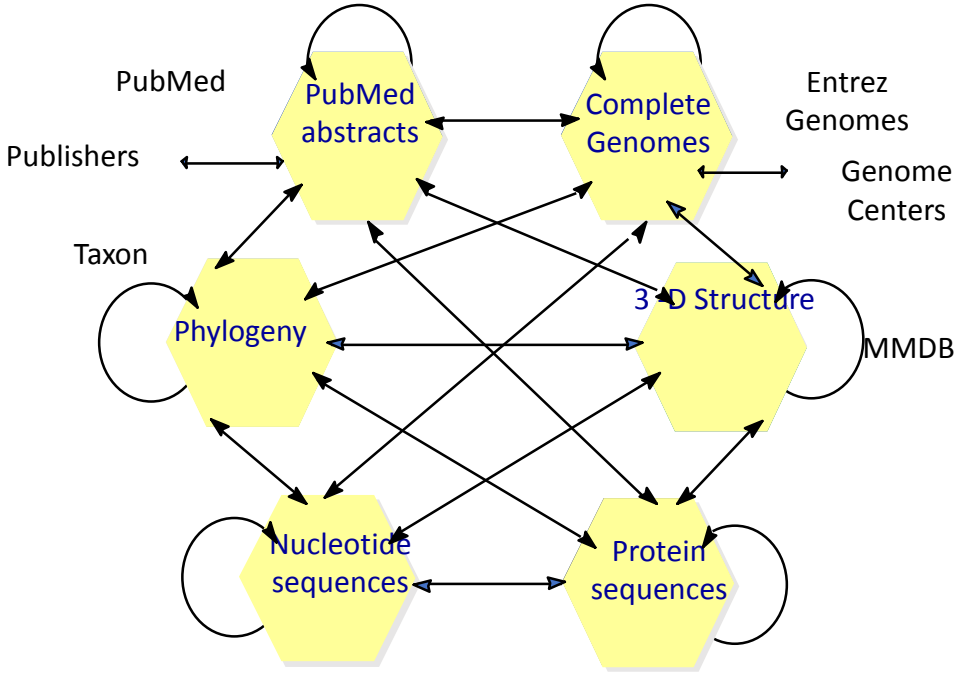
➢ New publishing paradigm: data is published before analysis by astronomers

➢ Platform for 'citizen science' with GalaxyZoo project



The US National Library of Medicine

- The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research.
- Requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central *upon acceptance for publication*.
- Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.



Entrez - cross-database search tool

PMC Open Access Compliance?

- PMC Compliance Rate
  - Before legal mandate compliance was 19%
  - Signed into law by George W. Bush in 2007
  - After legal mandate compliance up to 75%
- NIH announced in 2013 that they

  '… will hold processing of non-competing continuation awards if publications arising from grant awards are not in compliance with the Public Access Policy.'
- Since NIH implemented their policy about continuation awards
  - Compliance rate increasing ½ % per month
  - By September 2016, compliance had reached close to 90%

The Fourth Paradigm: Data-Intensive Science

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

- Newton's Laws, Maxwell's Equations…

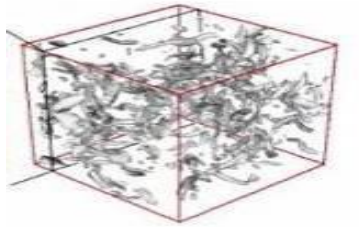
Last few decades – **Computational Science**

- 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

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

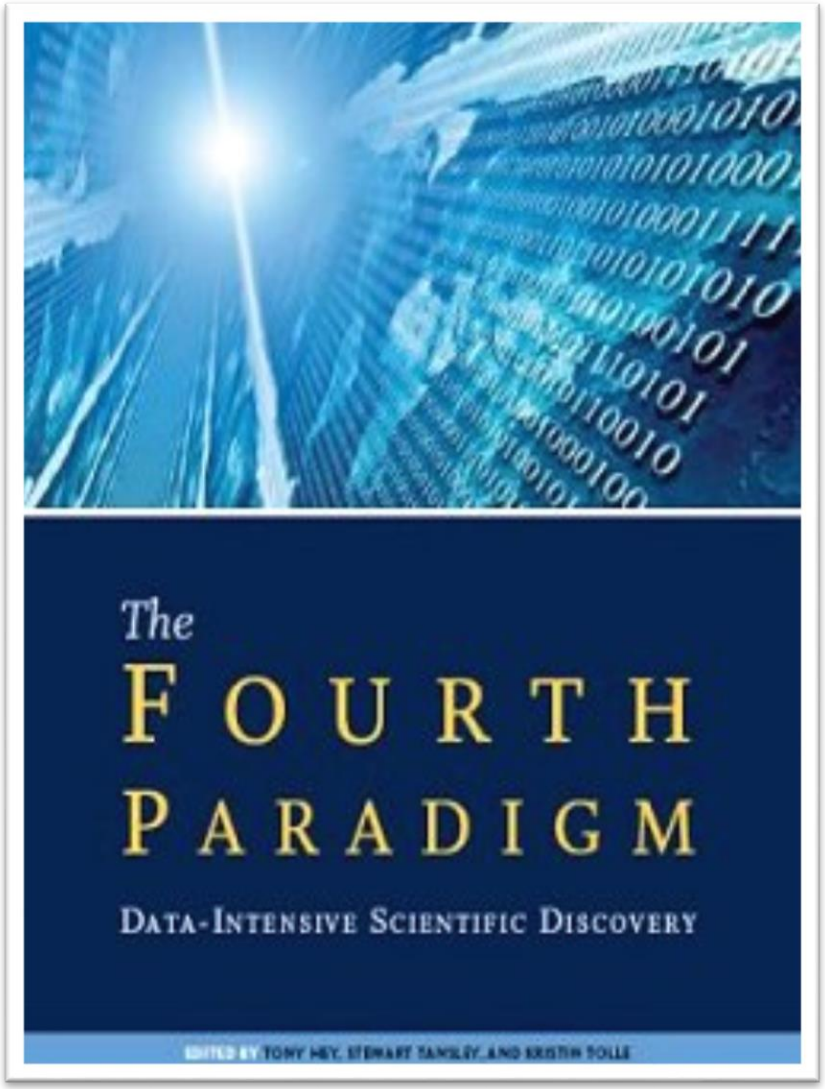


eScience is the set of tools and technologies to support data federation and collaboration

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

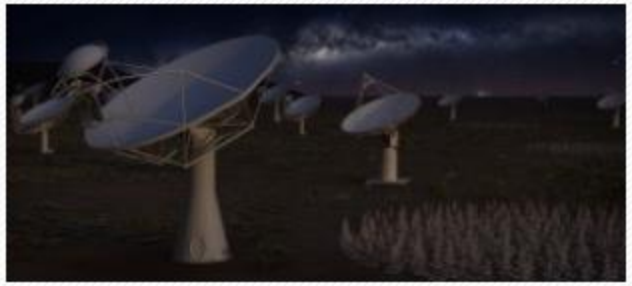
With thanks to Jim Gray

Data-Intensive Scientific Discovery



Examples of Data-Intensive Science

SKA Project



Artist impression of the Square Kilometre Array

The Square Kilometre Array (SKA) project is an international effort to build the world's largest radio telescope, with eventually over a square kilometre (one million square metres) of collecting area. The scale of the SKA represents a huge leap forward in both engineering and research & development towards building and delivering a unique instrument, with the detailed design and preparation now well under way. As one of the largest scientific endeavours in history, the SKA will bring together a wealth of the world's finest scientists, engineers and policy makers to bring the project to fruition.

Latest News

22nd December 2015
2015: a big year for ASKAP!

21st December 2015
Outcomes Of The 19th SKA Board Meeting

7th December 2015
Australia Announces AUS$293.7 Million for the SKA

SKA– Key Science Drivers: The history of the Universe

Extremely broad range of science!

Data Flow through the SKA

SKA1-MID

8.8 Tb/s

~50 PFLOPS

~5 Tb/s

100 PFLOPS

130 - 300 PB/yr

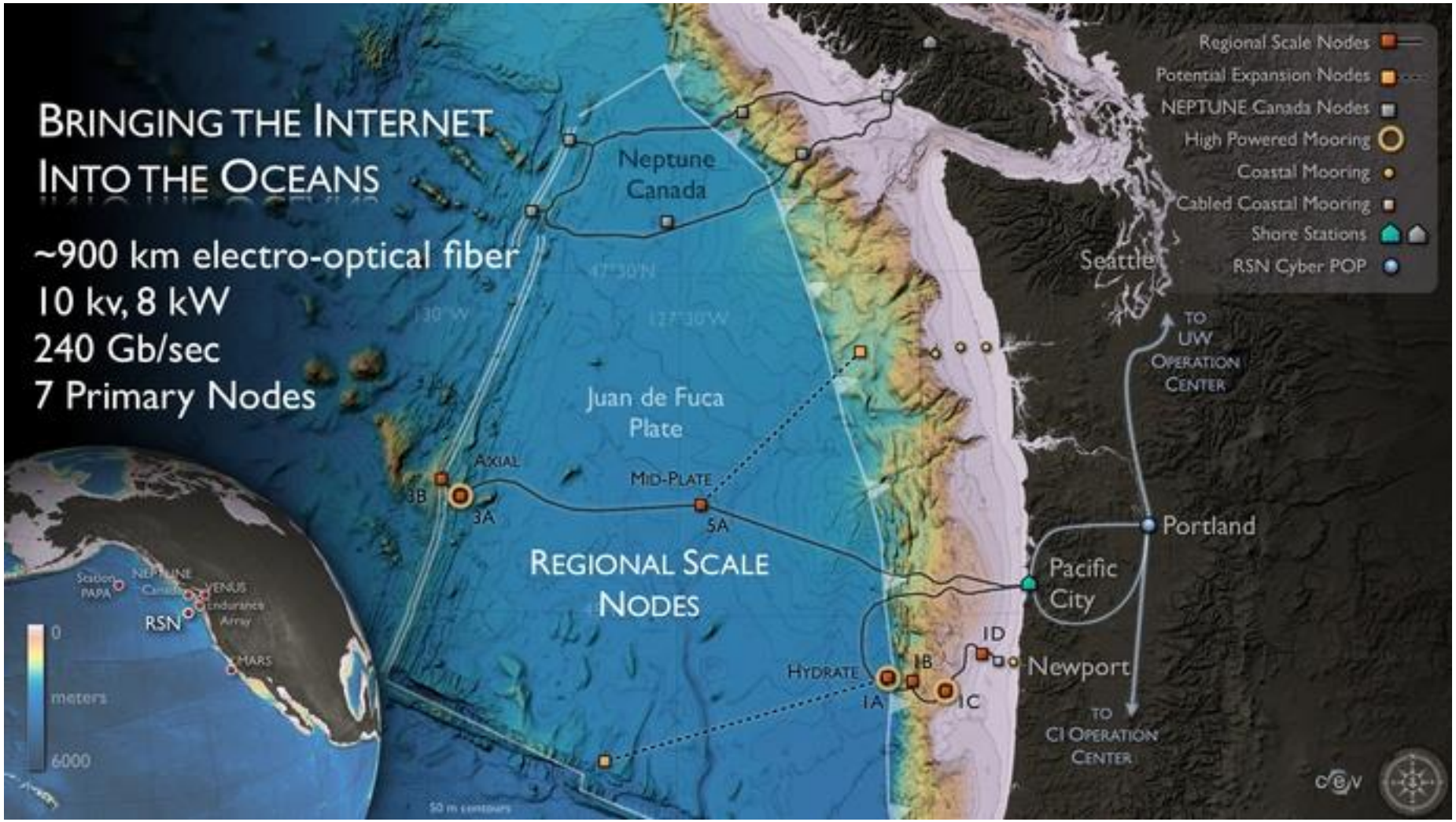
Users

7.2 Tb/s

SKA1-LOW

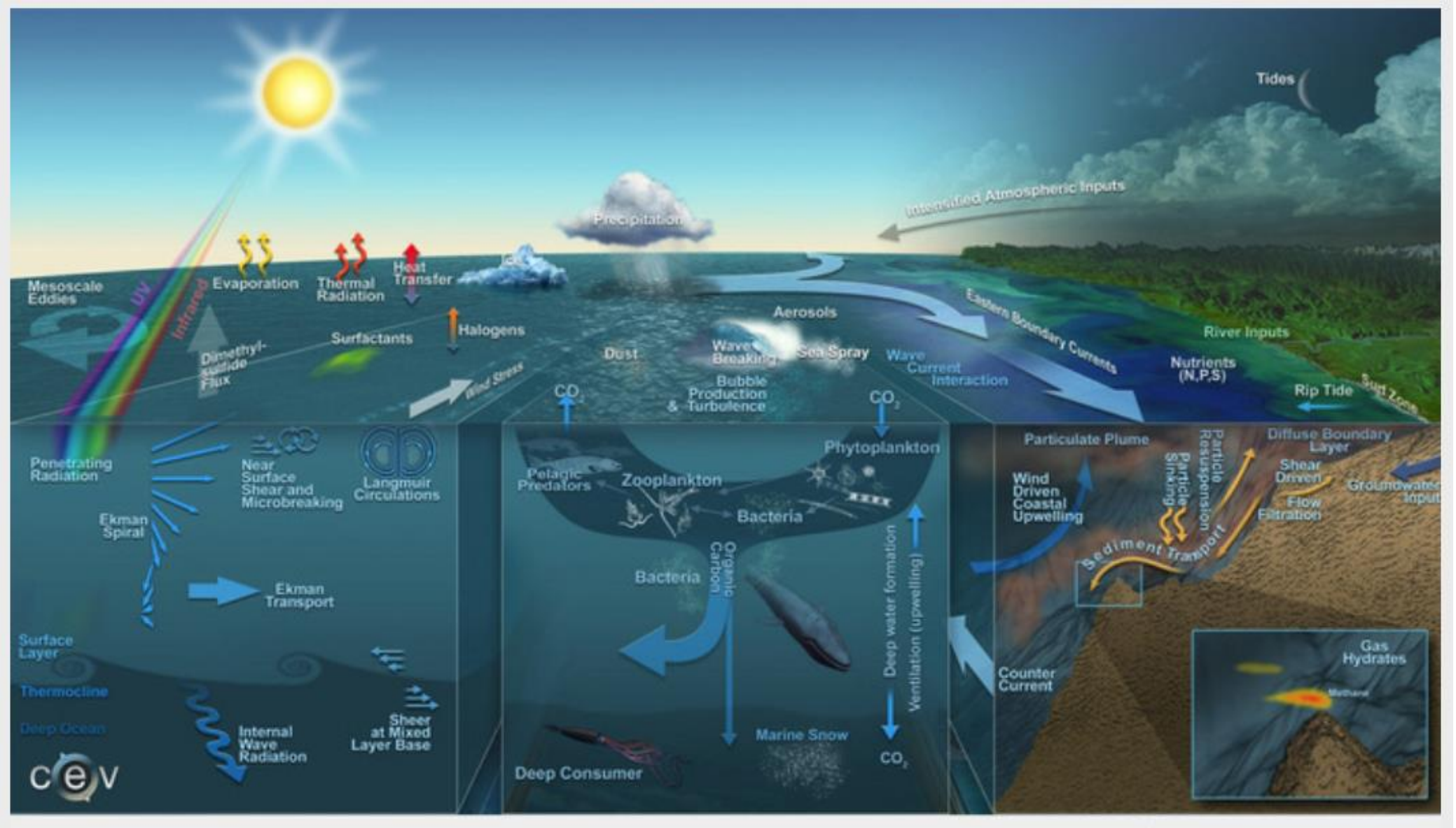
~2 Pb/s

NSF's Ocean Observatory Initiative



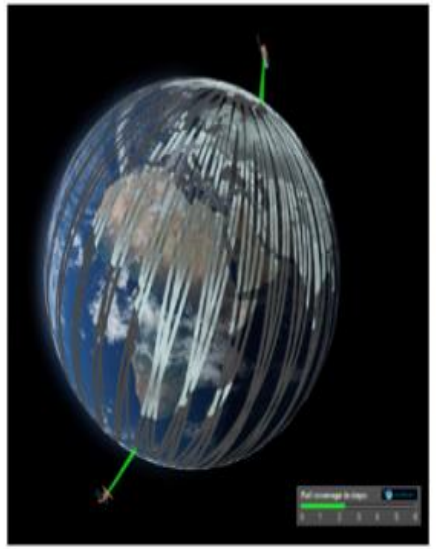
Slide courtesy of John Delaney

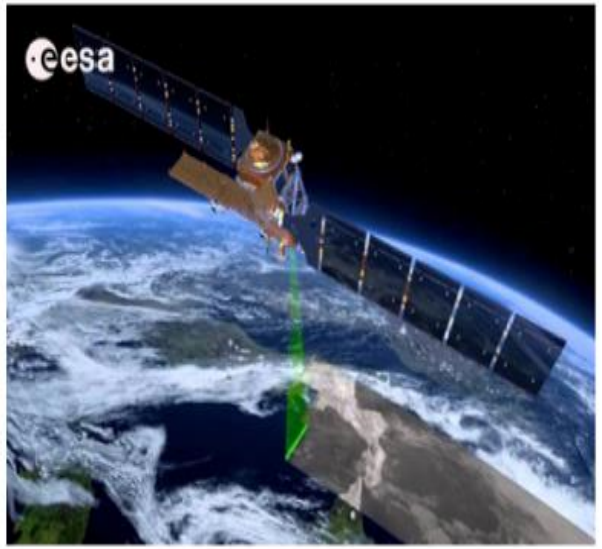
Science Drivers: Oceans and Life



Slide courtesy of John Delaney

Large data sets: satellite observations



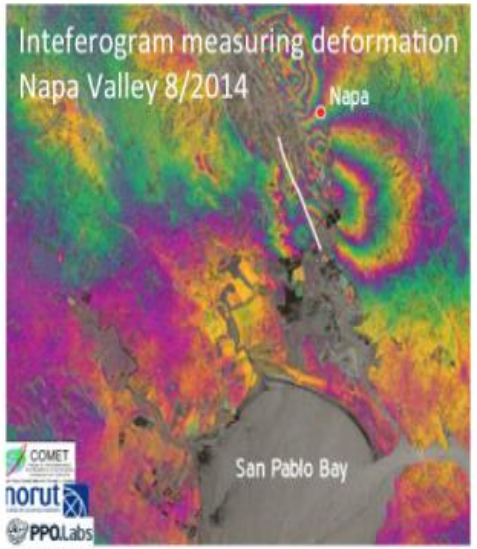


Sentinel 1A: Launched 2014
(1B due 2016)

- 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 x 10^7**	5 million grid points 100 levels 10 prognostic variables = **5 x 10^9**
Type	98% from 60 different satellite instruments	physical parameters of atmosphere, waves, ocean

Soon:	Observations	Models
Volume	200 million = **2 x 10^8**	500 million grid points 200 levels 100 prognostic variables = **1 x 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

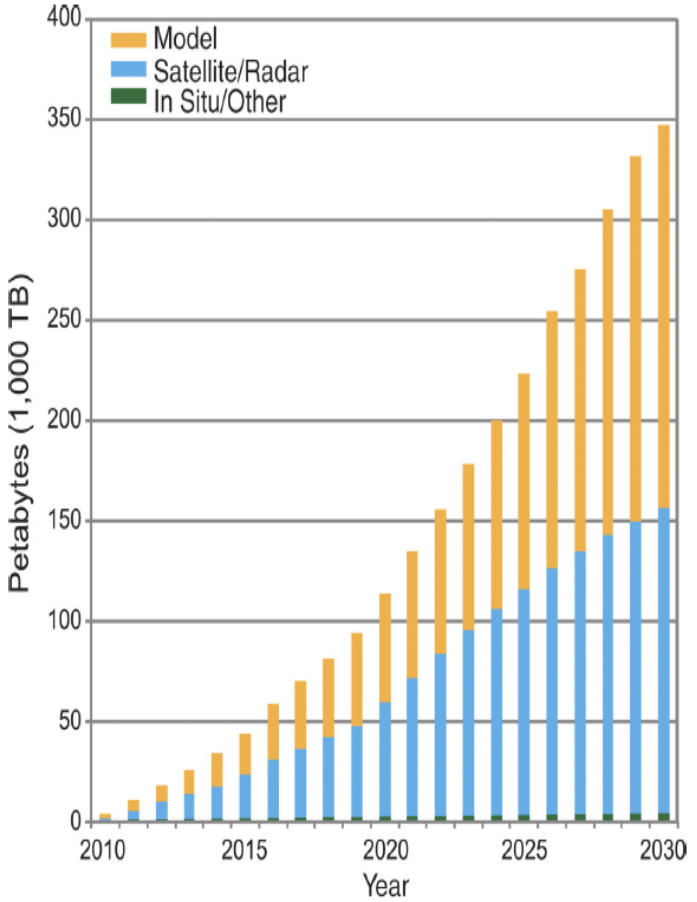
WCRP World Climate Research Programme

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

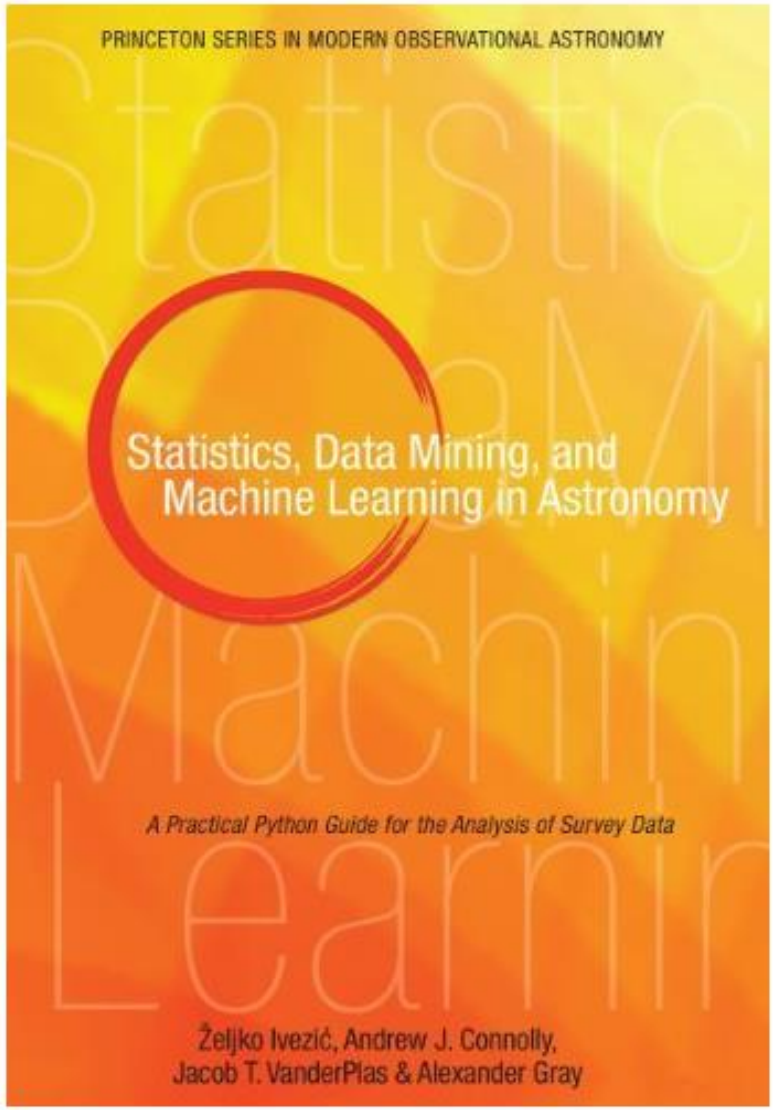
More Data

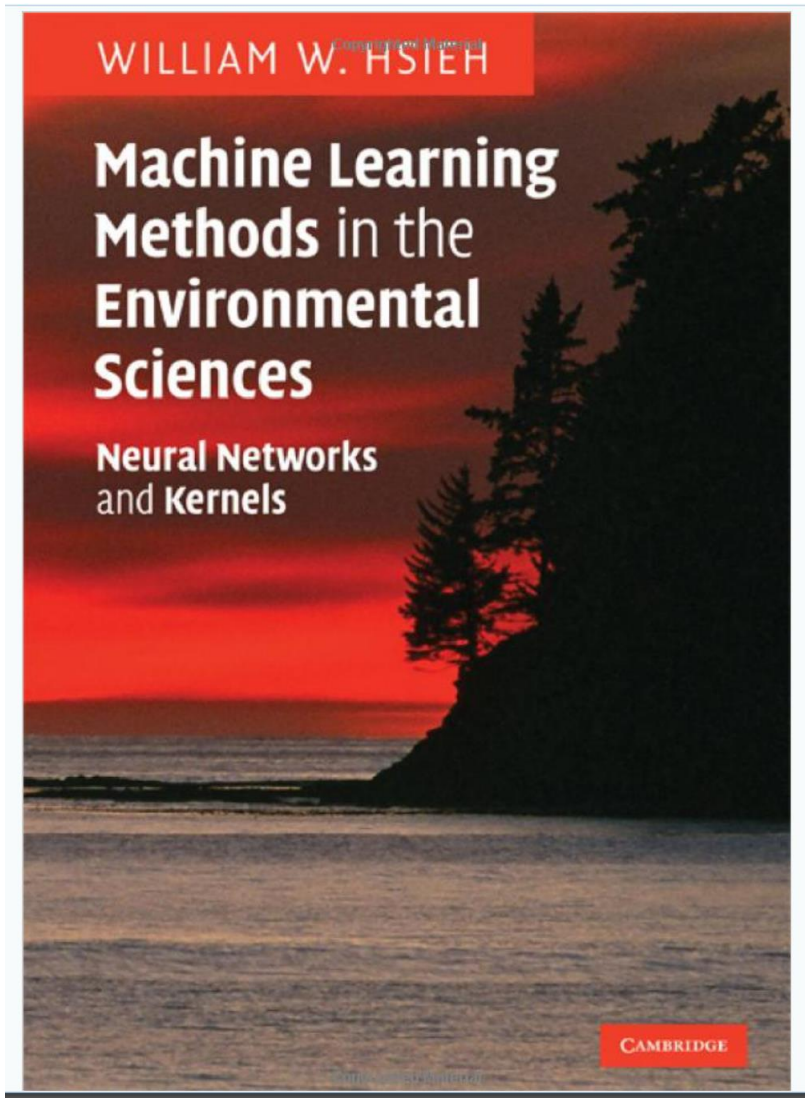
Fig. 2 The volume of worldwide climate data is expanding rapidly, creating challenges for both physical archiving and sharing, as well as for ease of access and finding what's needed, particularly if you're not a climate scientist.

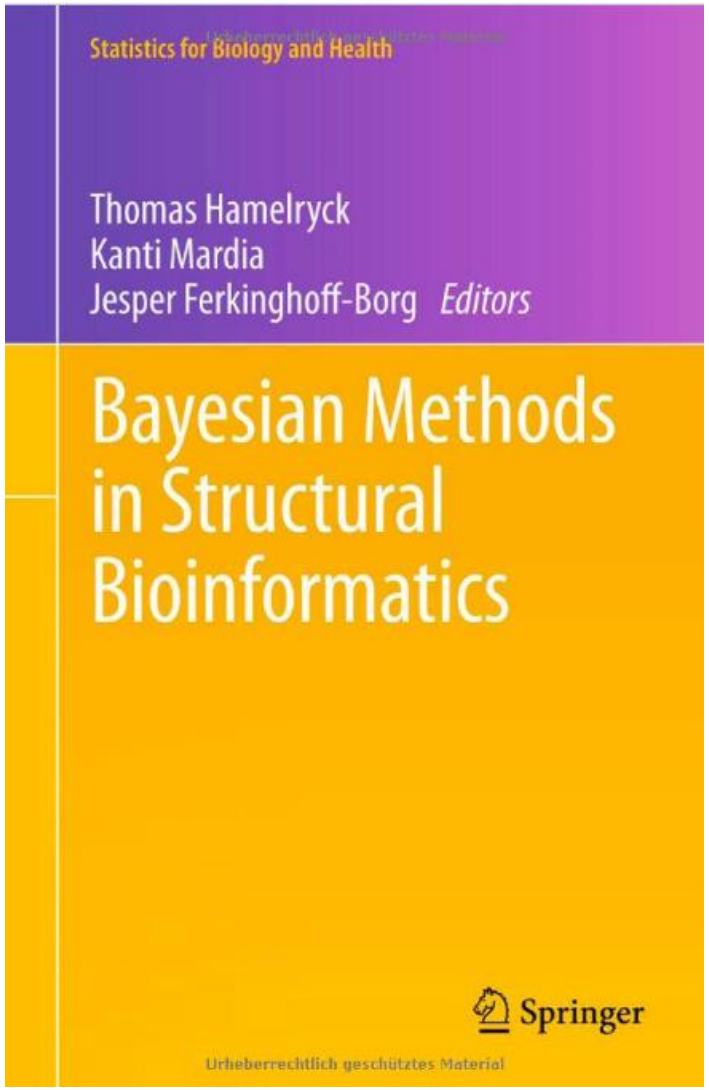
(BNL: Even if you are?)



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


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

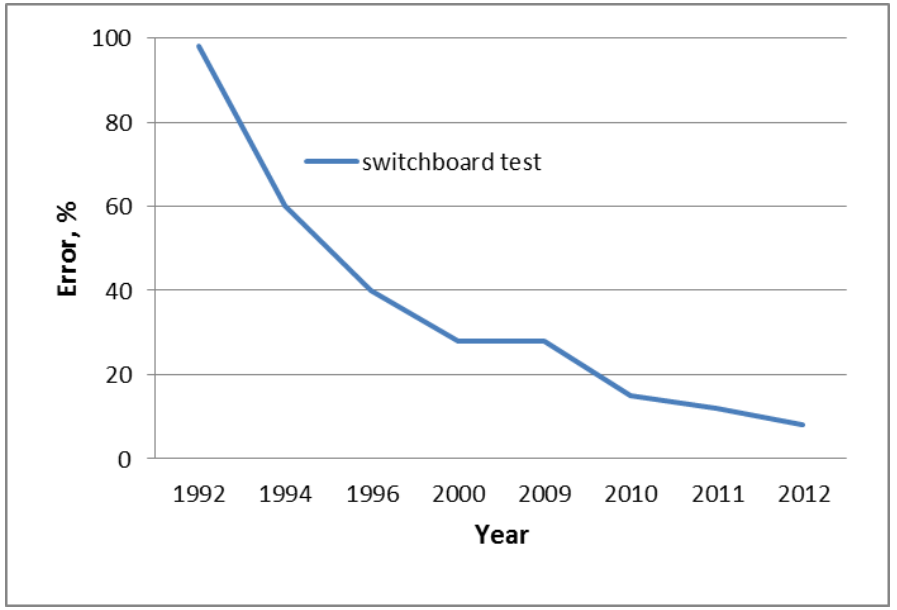

WILLIAM W. HSIEH
Machine Learning Methods in the Environmental Sciences
Neural Networks and Kernels
CAMBRIDGE


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

The Machine Learning Revolution

- Neural networks are just one example of a Machine Learning (ML) algorithm

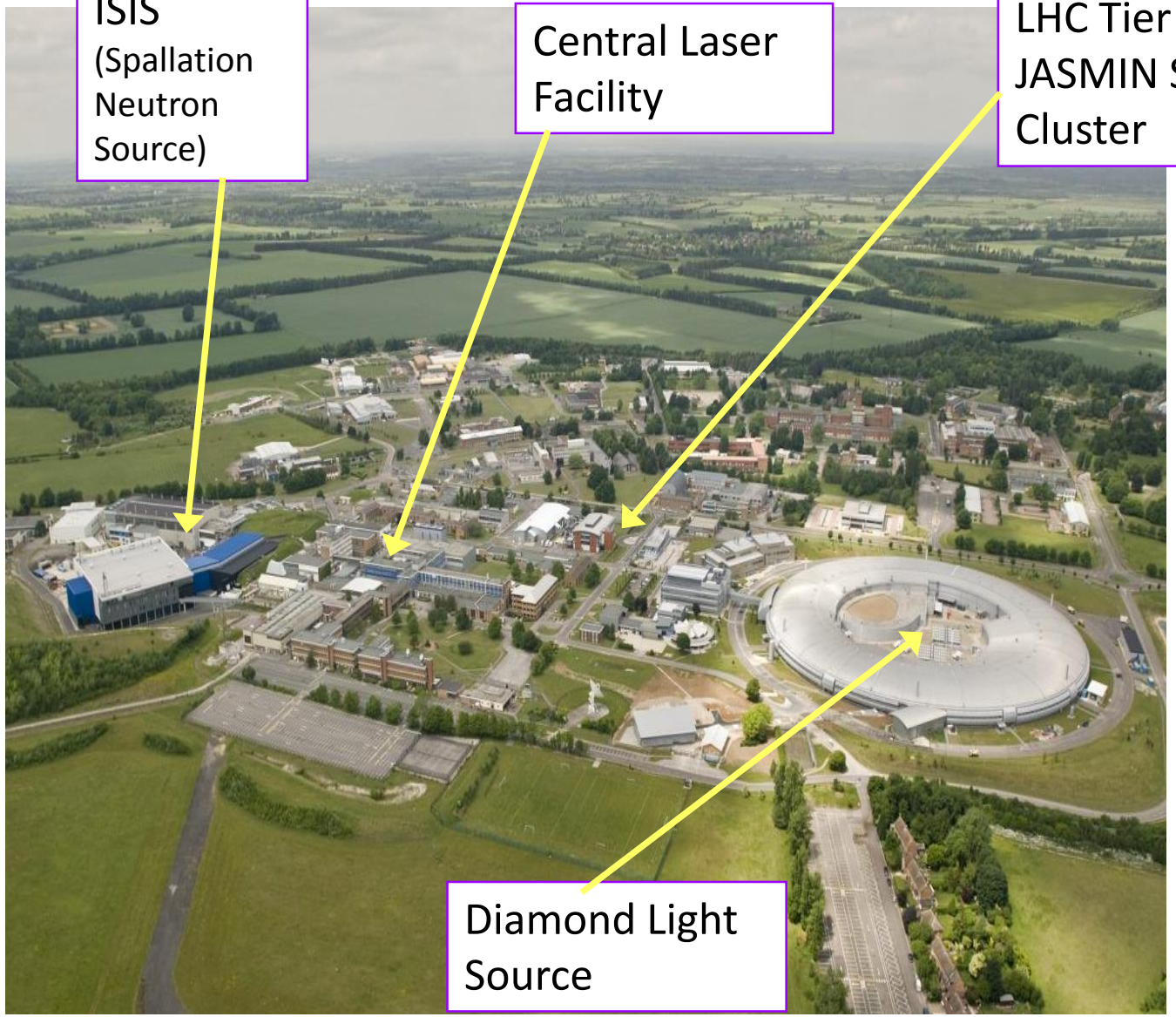
- Deep Neural Networks are now exciting the whole of the IT industry since they enable us to:
  - Build computing systems that improve with experience
  - Solve extremely hard problems
  - Extract more value from Big Data
  - Approach human intelligence
    e.g. natural language processing



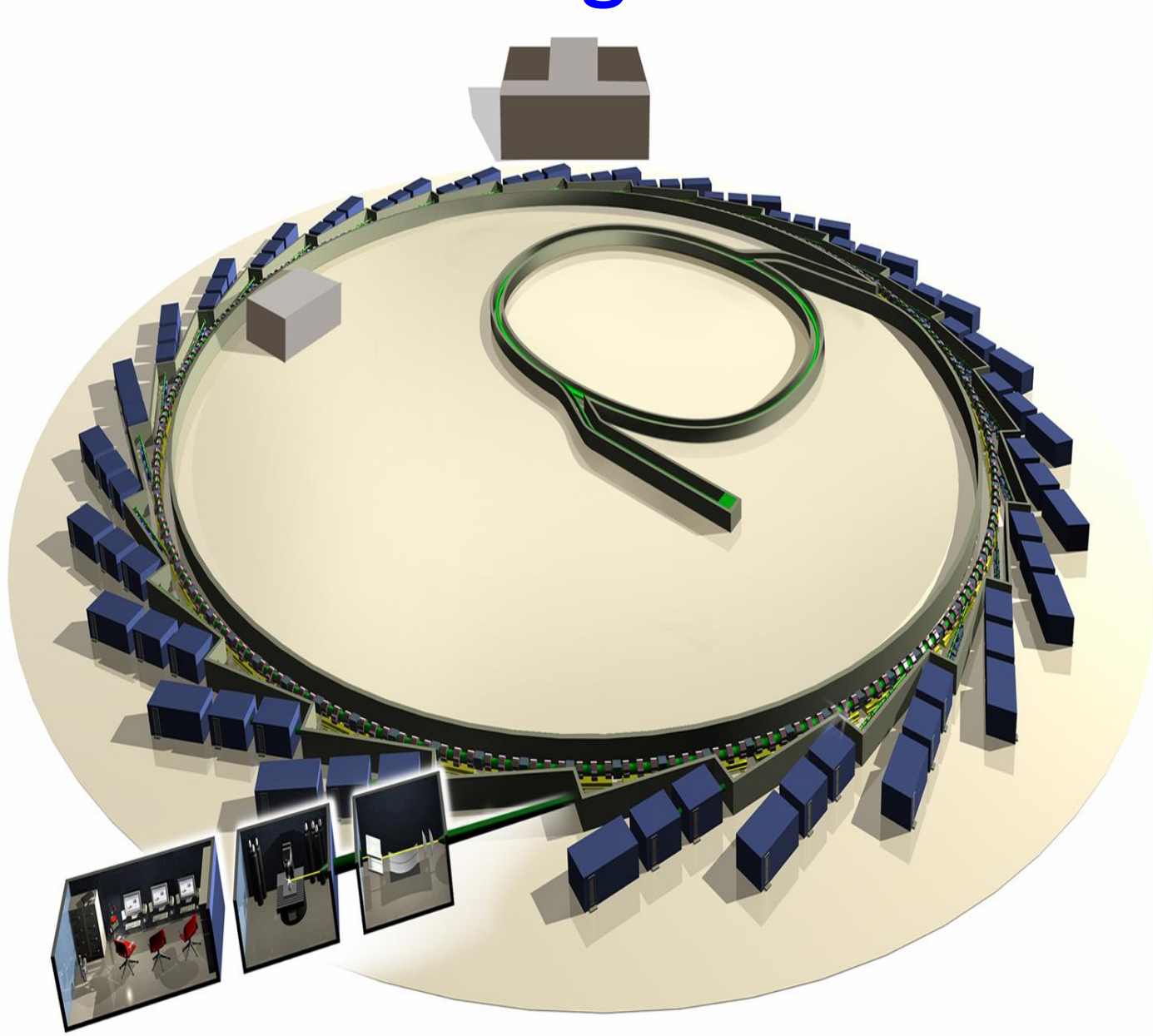
- The change in the Word Error Rate (WER) with time for the NIST "Switchboard" data.
- In 2016 Microsoft researchers achieved a word error rate (WER) of 6.3 percent, the lowest in the industry.

Big Scientific Data at Harwell

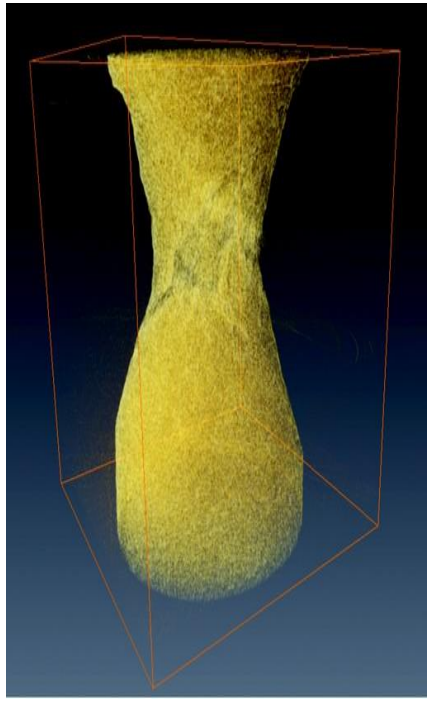
Rutherford Appleton Lab and the Harwell Campus



Diamond Light Source

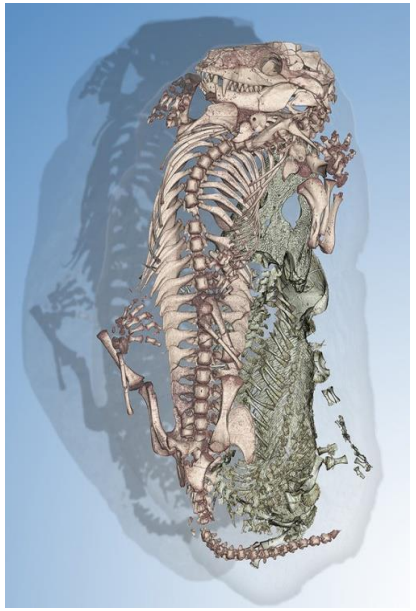


Science Examples

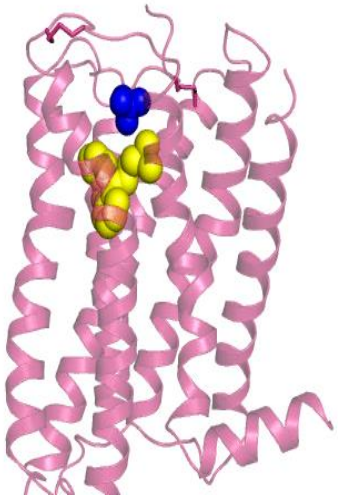


Casting aluminium

Pharmaceutical & manufacturing

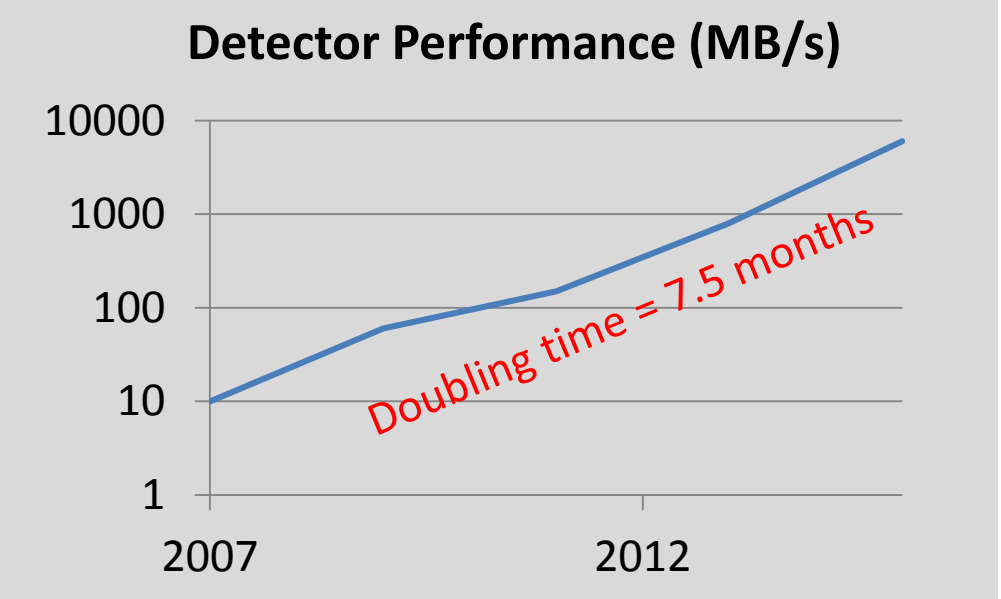


Non-destructive imaging of fossils



Structure of the Histamine H1 receptor

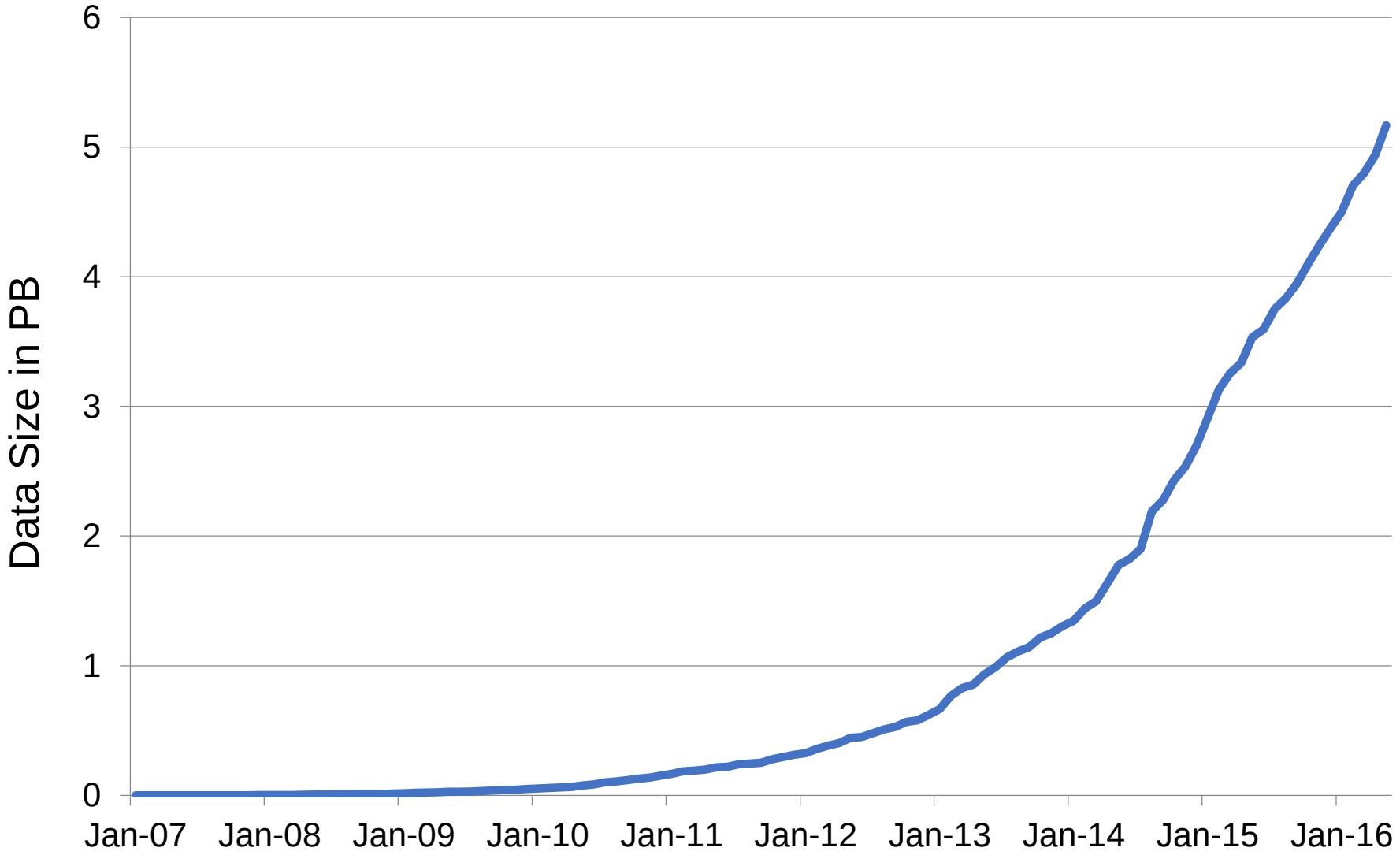
Data Rates



- 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
- 2016 Percival detector 6GB/sec

Thanks to Mark Heron

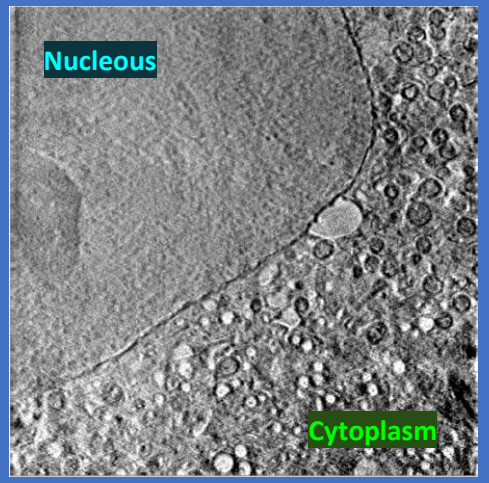
Cumulative Amount of Data Generated By Diamond



Thanks to Mark Heron

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

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

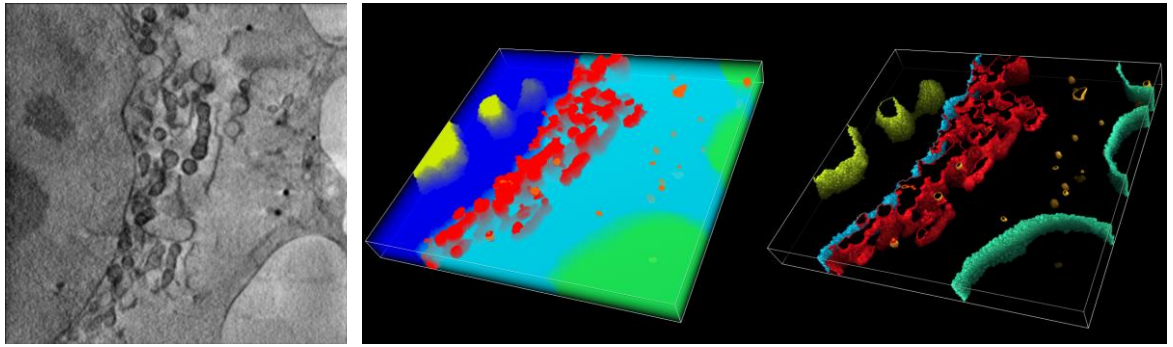
scientificsoftware@diamond.ac.uk

Data

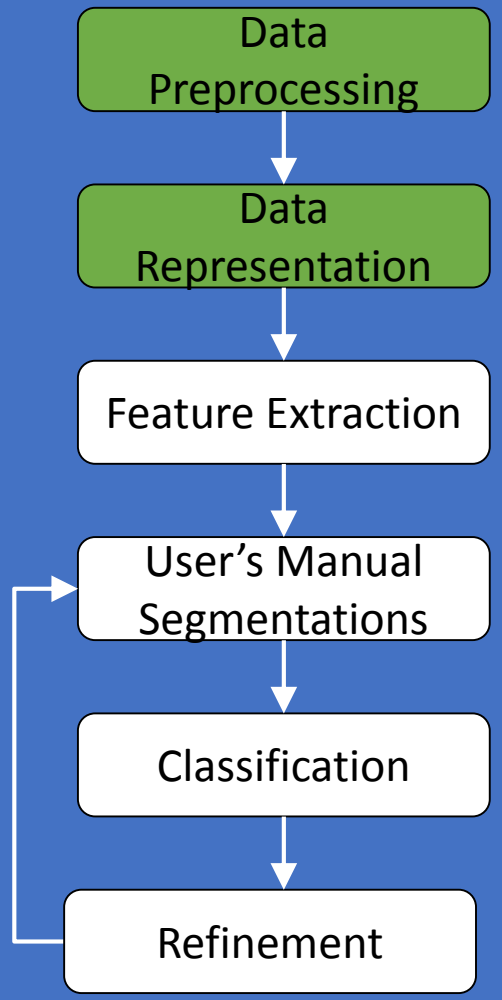
- **B24:** Cryo Transmission X-ray Microscopy beamline at DLS
- Data Collection: Tilt series from ±65° with 0.5° step size
- Reconstructed volumes up to 1000x1000x600 voxels
- Voxel resolution: ~40nm currently
- Total depth: up to 10µm
- **GOAL: Study structure and morphological changes of whole cells**

diamond
B24 beamline
Data Analysis Software Group
The University of Nottingham
Computer Vision Laboratory

3D Volume Data → Segmentation



Workflow



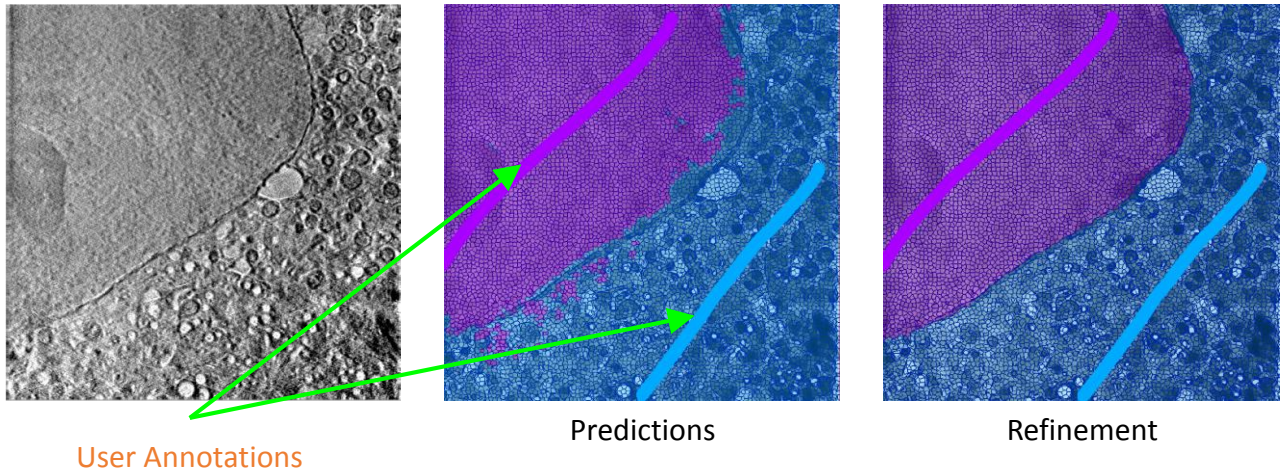
scientificsoftware@diamond.ac.uk

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



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

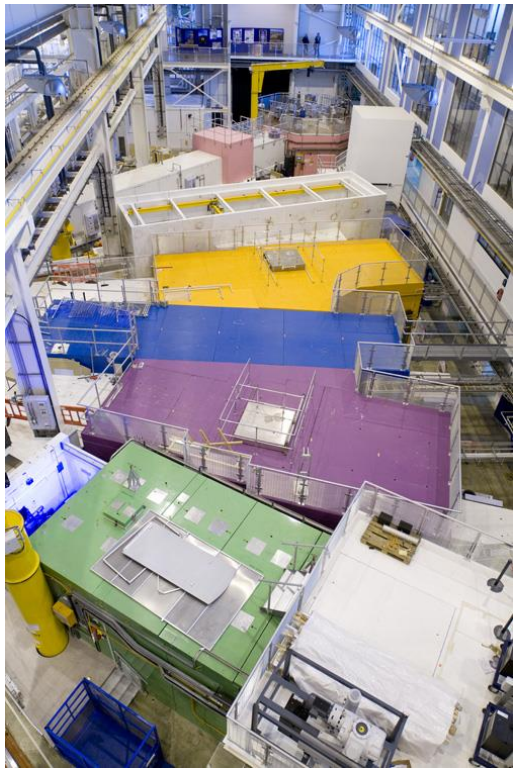
- A machine learning classifier (i.e. Random Forest) is trained to discriminate between different classes (i.e. Nucleus and Cytoplasm)
- A Markov Random Field (MRF) is then used to refine the predictions.

The ISIS Neutron and Muon Facility


ISIS

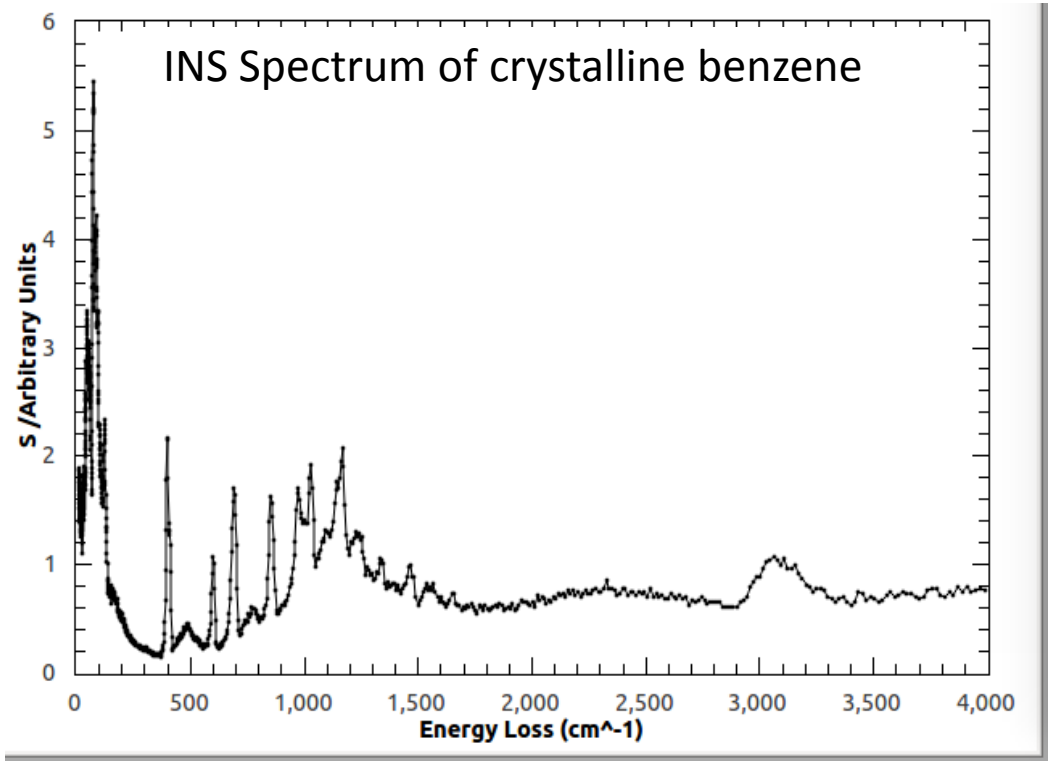
ISIS Neutron and Muon Source

- ≈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



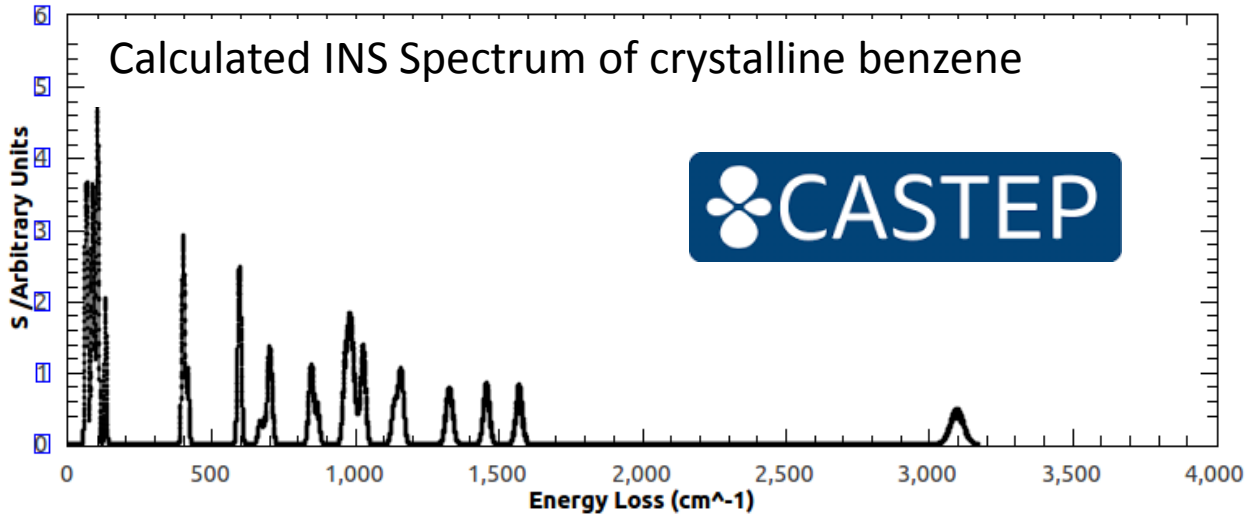
Peak Assignment in Inelastic Neutron Scattering

- 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 ?

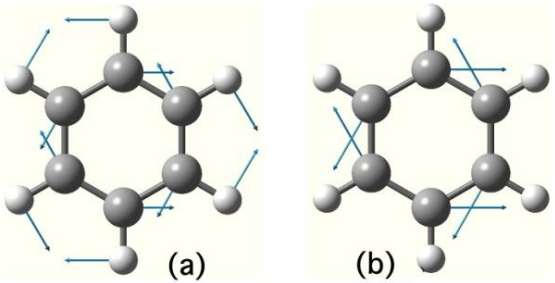


S. Parker and S. Mukhopadhyay (ISIS)

Modelling & Simulation for INS Peak Assignment

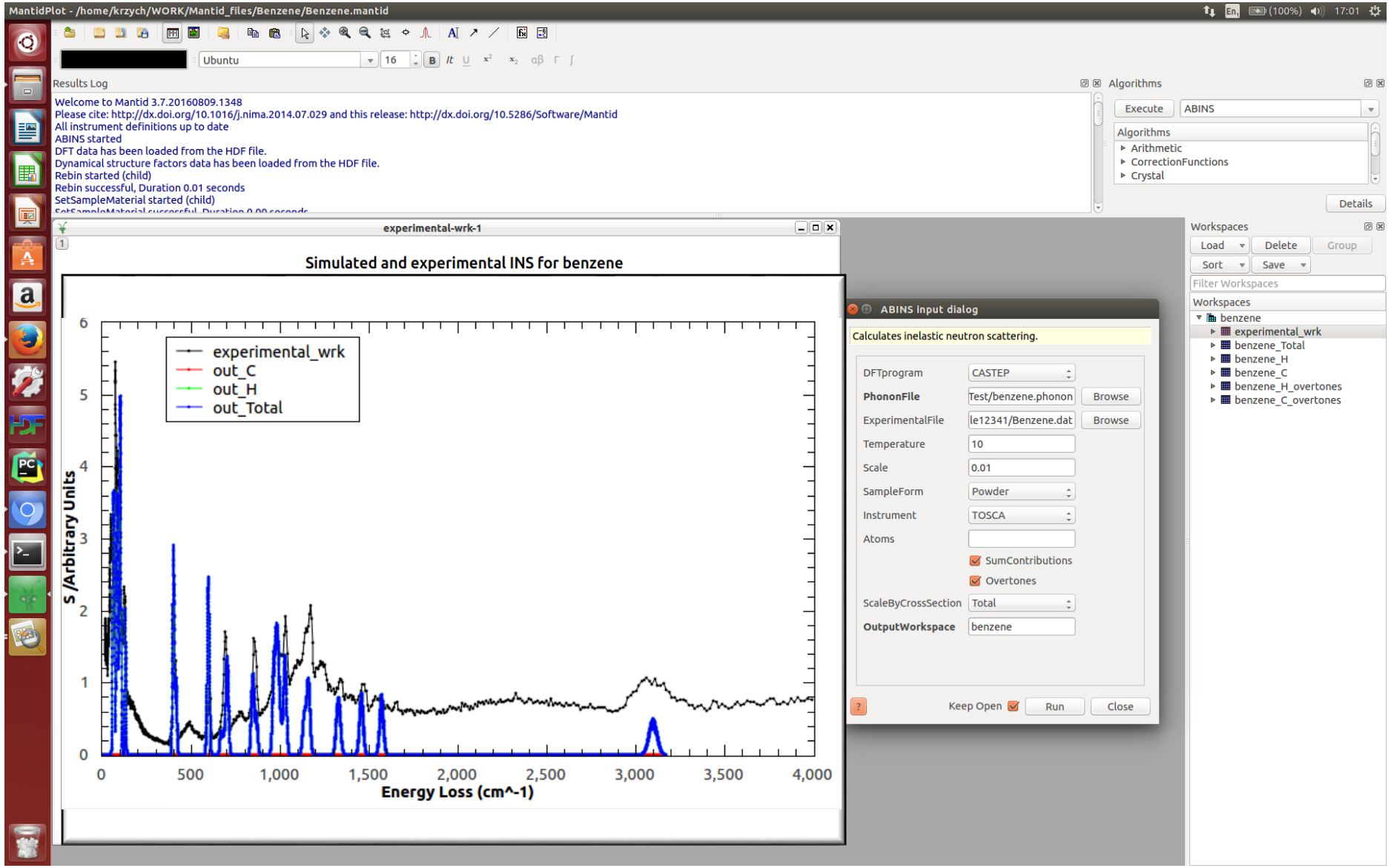


- INS spectra can be computed for a given atomic structure
- Calculations allow us to see what specific vibrational motion of atoms occur, and at what frequency



L. Liborio

Materials Workbench

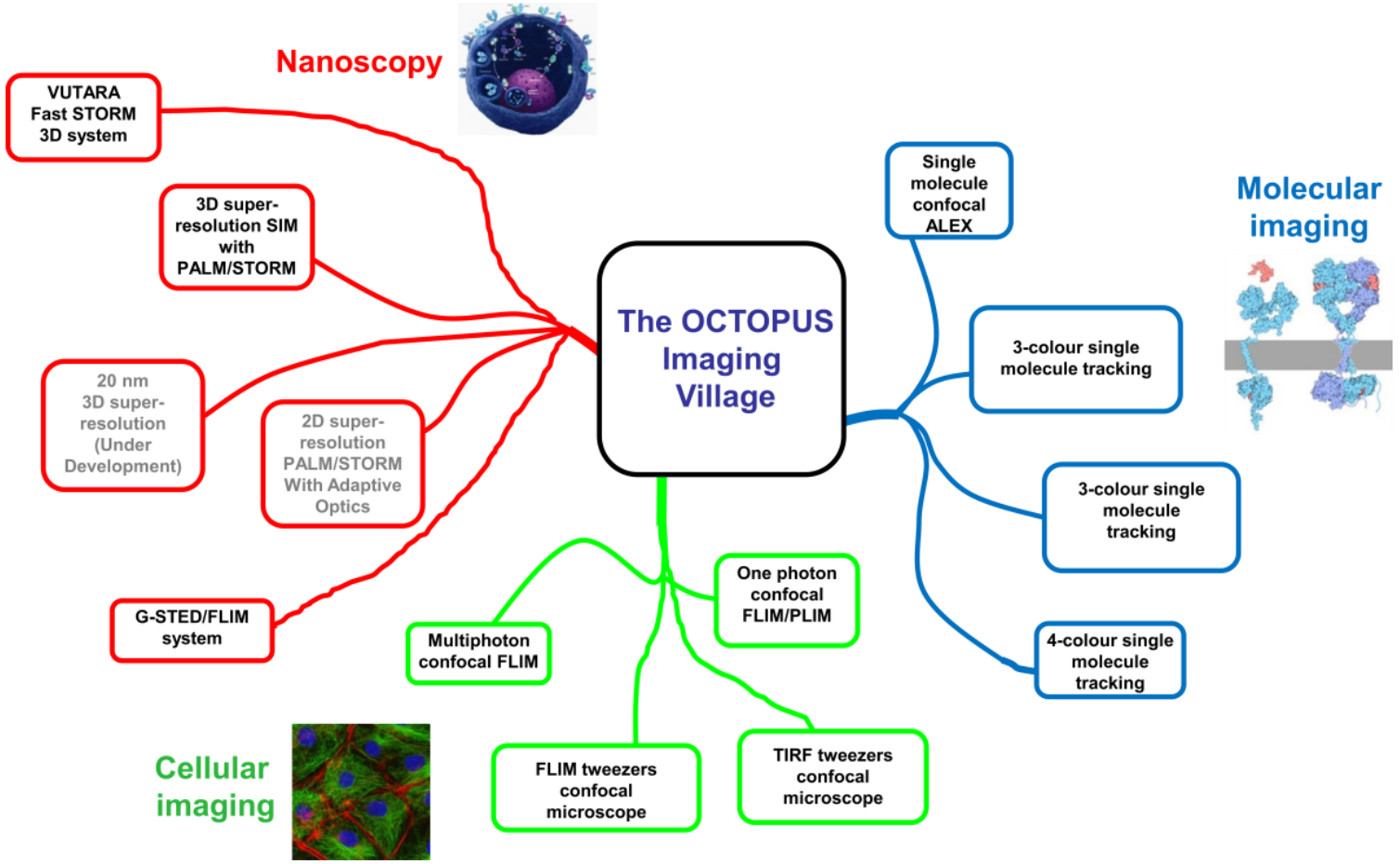


K. Dymkowski

The Central Laser Facility (CLF)

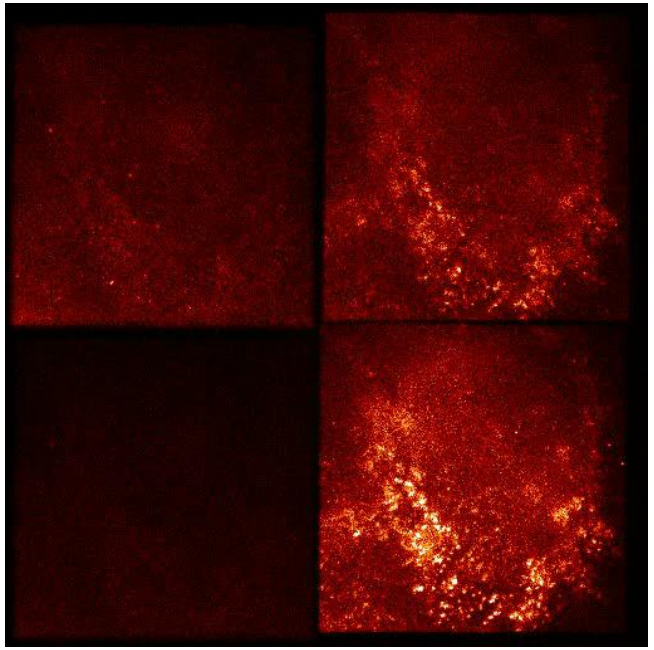
OCTOPUS Facility in the CLF

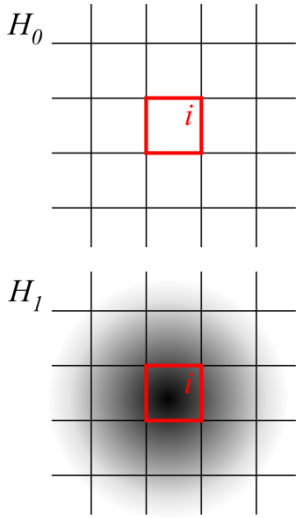
- National imaging facility with peer-reviewed, funded access
- Located in Research Complex at Harwell
- Cluster of microscopes and lasers and expert end-to-end multidisciplinary support
- Operations and some development funded by STFC
- Key developments funded through external grant – BBSRC, MRC

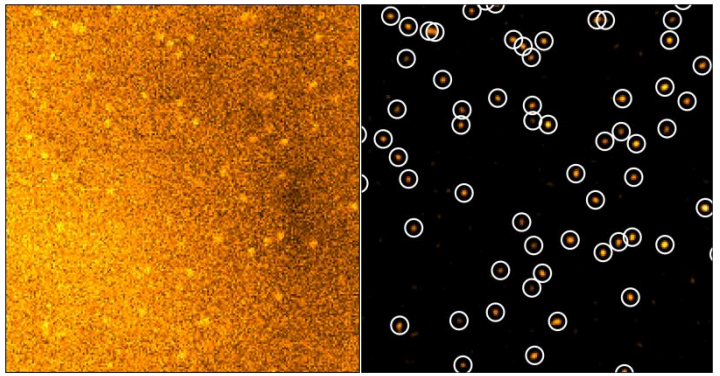


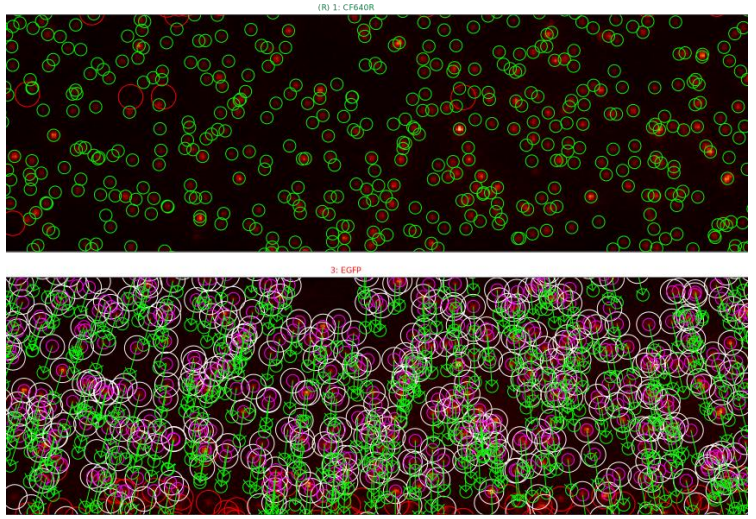
With thanks to Dan Rolfe

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

Cryo-Electron Microscopy and the CMOS Detector Revolution

With thanks to Nicola Guerrini

Transmission Electron Microscopy (TEM)

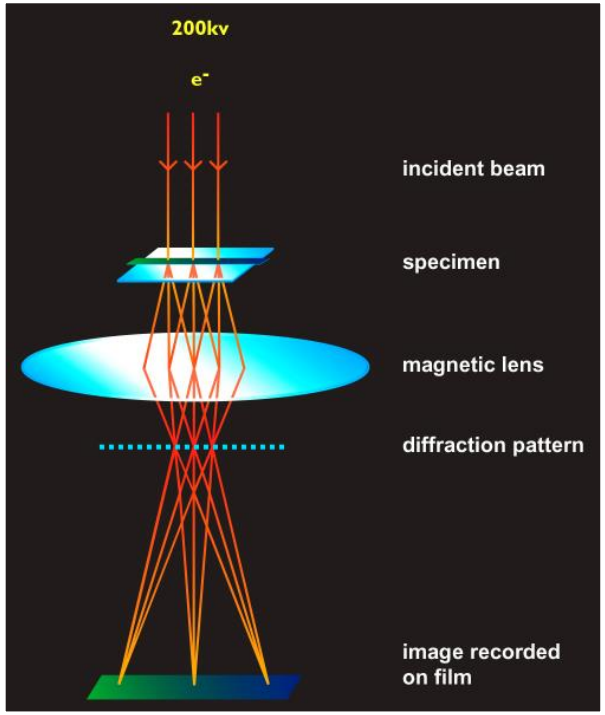


Image courtesy of LMB-Cambridge

- With visible light it is impossible to resolve points that are closer together than a few hundred nanometres
- Electron and ion microscopes use a beam of charged particles instead of light, and electromagnetic or electrostatic lenses to focus the particles

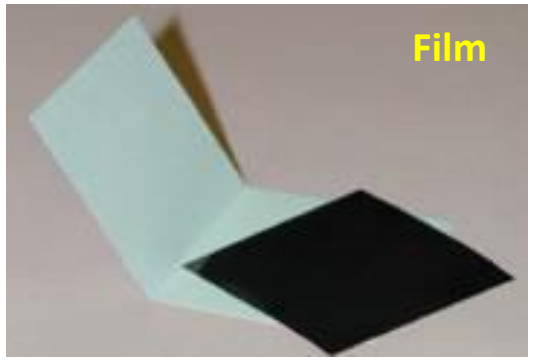
➢ The resulting image used to be recorded on film or with a CCD camera with phosphor and a fibre optics …

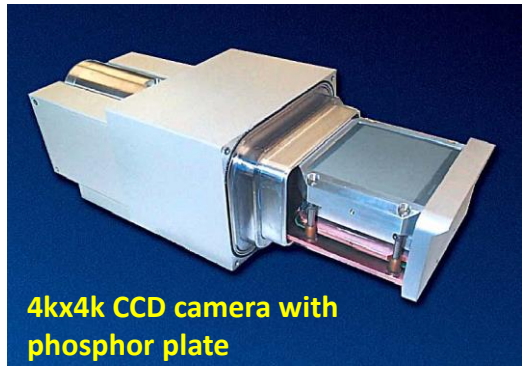
With thanks to Nicola Guerrini

The CMOS Detector Revolution

- Film: good resolution, non digital, needs time for development, poor Signal/Noise for weak exposure

- CCD with phosphor: not direct detection (radiation hardness), phosphor ruins spatial resolution, good for tomography



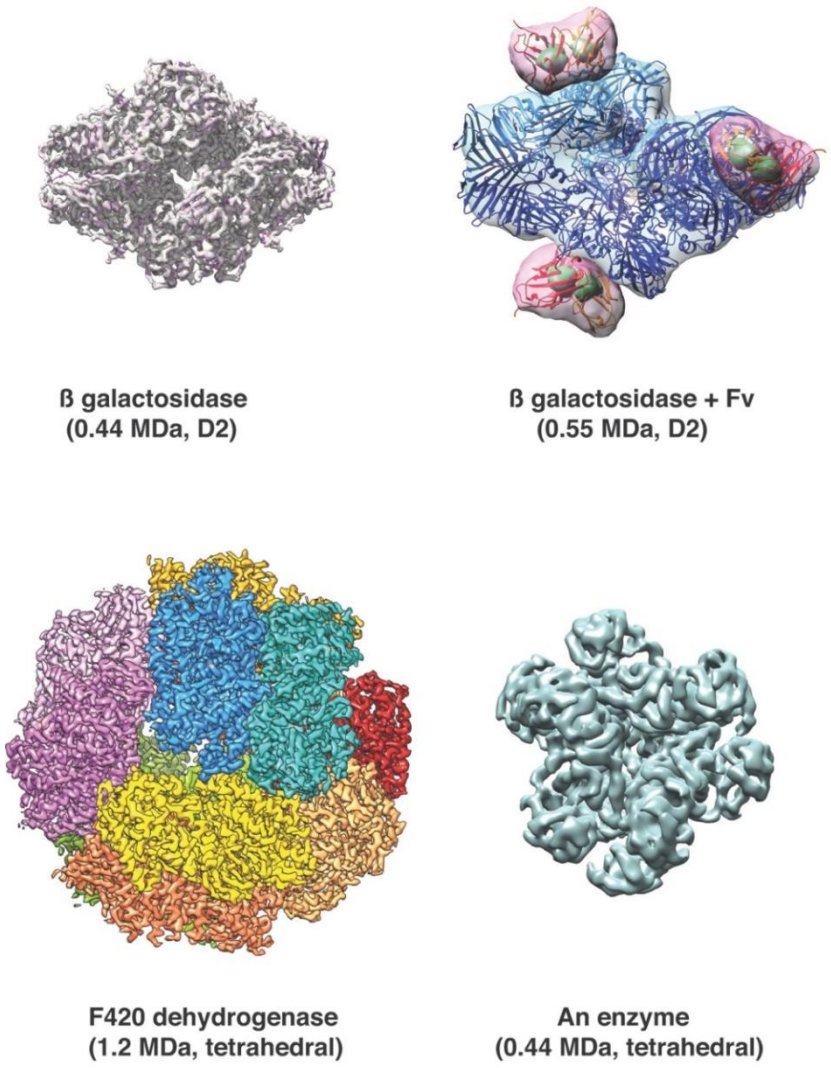


➢ CMOS sensors allow direct detection, digital, have good spatial resolution and good sensitivity (single electron)

With thanks to Nicola Guerrini

Cryo-Electron Microscopy (Cryo-EM)

- Electron Cryo-Microscopy is used to observe three-dimensional structure of macromolecules in conditions close-to-native
- Atomic structures of specimens can be obtained from frozen samples, without the crystals needed in X-ray crystallography
- The three-dimensional macromolecular structure is reconstructed by taking several images of many molecules. Such images are then processed to extract high resolution information



W. Kühlbrandt, Science, vol. 343, 28 March 2014

With thanks to Nicola Guerrini

CMOS Sensors for TEM

Direct detection enabled by CMOS sensors has led to a "resolution revolution"

Advances in detector technology and image processing are yielding high-resolution electron cryo-microscopy structures of biomolecules

BIOCHEMISTRY

The Resolution Revolution

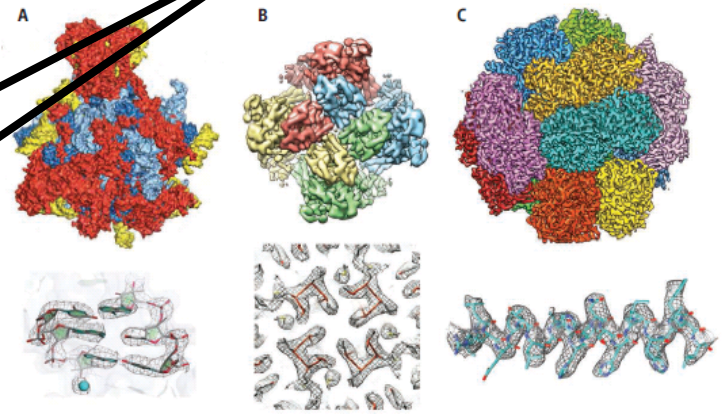
Advances in detector technology and image processing are yielding high-resolution electron cryo-microscopy structures of biomolecules.

Werner Kühlbrandt

Precise knowledge of the structure of macromolecules in the cell is essential for understanding how they function. Structures of large macromolecules can now be obtained at near-atomic resolution by averaging thousands of electron microscope images recorded before radiation damage accumulates. This is what Amunts *et al.* have done in their research article on page 1485 of this issue (*1*), reporting the structure of the large subunit of the mitochondrial ribosome at 3.2 Å resolution by electron cryo-microscopy (cryo-EM). Together with other recent high-resolution cryo-EM structures (*2*–*4*) (see the figure), this achievement heralds the beginning of a new era in molecular biology, where structures at near-atomic resolution are no longer the prerogative of x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy.

Ribosomes are ancient, massive protein-RNA complexes that translate the linear genetic code into three-dimensional proteins. Mitochondria—semi-autonomous organelles that supply the cell with energy—have their own ribosomes, which closely resemble those of their bacterial ancestors. Many antibiotics, such as erythromycin, inhibit growth of bacteria by blocking the translation machinery of bacterial ribosomes. When designing new antibiotics, it is essential that they do not also block the mitochondrial ribosomes. For this it is of great value to know the detailed structures. The same holds for heterogeneous samples or flexible complexes that do not crystallize readily, because cryo-EM images of different particles or conformations are easily separated at the image processing stage.

The new detectors offer another decisive advantage: Their fast readout makes it possible to compensate small movements that inevitably happen when the electron beam



Near-atomic resolution with cryo-EM. (**A**) The large subunit of the yeast mitochondrial ribosome at 3.2 Å reported by Amunts *et al.* In the detailed view below, the base pairs of an RNA double helix and a magnesium ion (blue) are clearly resolved. (**B**) TRPV1 ion channel at 3.4 Å (*2*), with a detailed view of residues lining the ion pore on the four-fold axis of the tetrameric channel. (**C**) F_{420}-reducing [NiFe] hydrogenase at 3.36 Å (*3*). The detail shows an α helix in the FrhA subunit with resolved side chains. The maps are not drawn to scale.

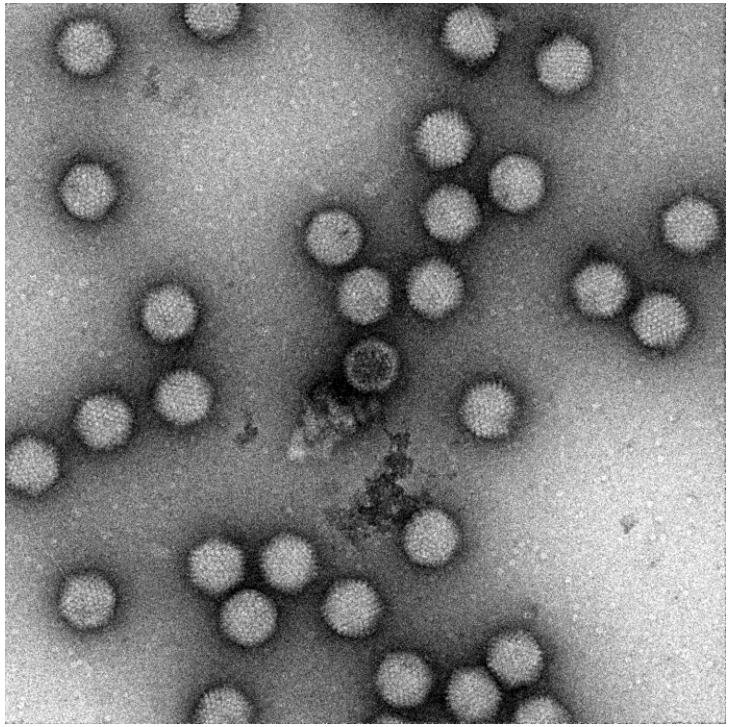
Photographic film works in principle much better for high-resolution imaging, but is incompatible with rapid electronic readout and high data throughput, which are increasingly essential.

Some 10 years ago, Henderson and Faruqi realized that it should be possible to design a sensor that detects electrons directly and that combines the advantages of CCD cameras and

W. Kühlbrandt, Science, vol. 343, 28 March 2014

With thanks to Nicola Guerrini

First CMOS image sensor for the TEM market in the FEI Falcon© Direct Electron Detector



Courtesy of G. McMullan (LMB, Cambridge, UK)

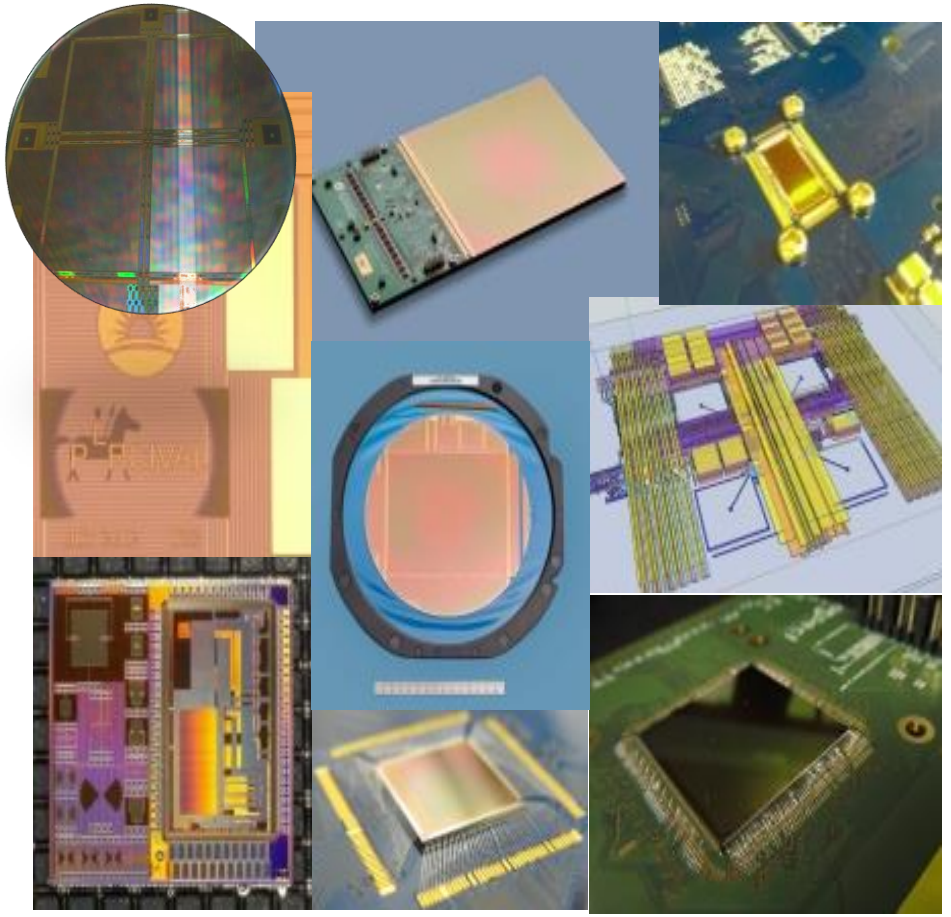
- Direct detection of electrons → high MTF and DQE
- 16 Mpixel, 14 µm pitch
- 40 fps or 640 Mpixel/sec
- Radiation hard → >20 Mrad
- CMOS image sensors are replacing film/CCD

With thanks to Nicola Guerrini

Conclusions

- CMOS sensors are a fast growing sector
- CMOS sensors have revolutionised the TEM field
- Exciting results already published and more to come

➢ Faster and less noisy sensors for better performance are the way forward



With thanks to Nicola Guerrini

The Ada Lovelace Center

ada lovelace centre

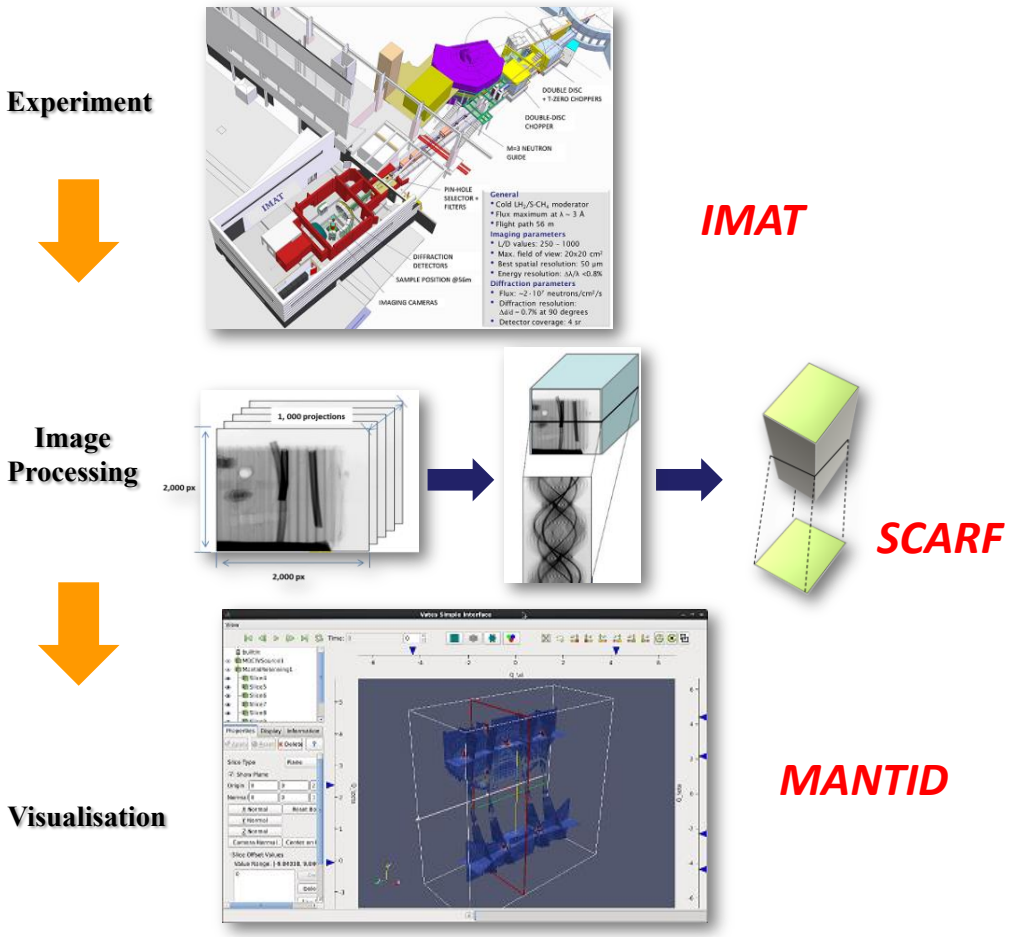
The Data Analysis Gap

- Complex Data
  - Too big to move in some cases
  - High CPU / memory requirements
  - May need to combine data from different sources
- Complex software environments
  - Variation in users' knowledge of HPC
  - Variation in home computing environments
  - Variation in the availability of Analysis and modelling Software
- Diverse science communities supported by the Facilities
  - Different analysis software requirements

➢ Users' access to computing resources and expertise is becoming a real barrier to extracting science from Big Scientific Data

With thanks to Brian Matthews, SCD

ALC Pathfinder: ‘Ultra’ for Tomographic Reconstruction



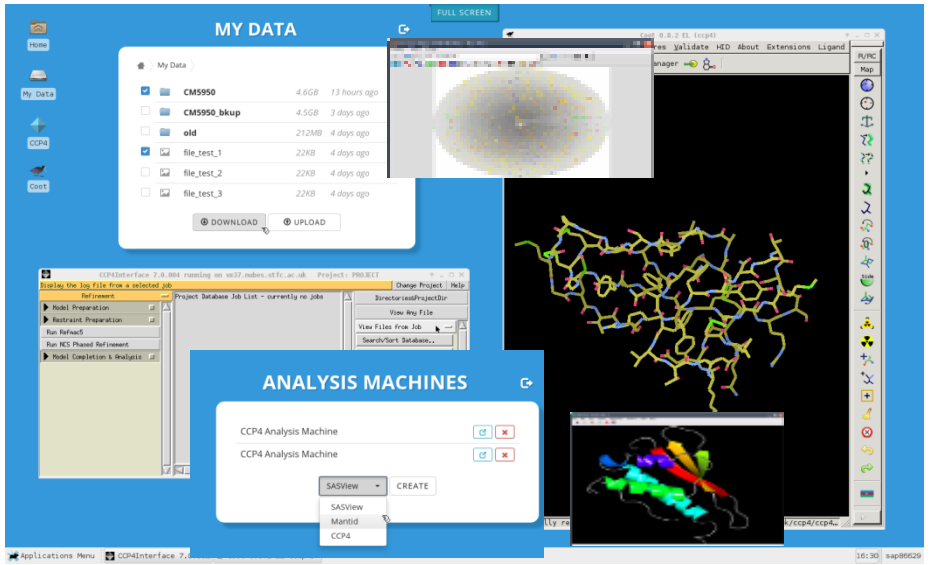
- Support in-experiment and post-experiment tomographic reconstruction
  - Round-trip the data to HPC CPU/GPU clusters in experiment time
  - Tomographic image reconstruction toolbox with different algorithms
  - High throughput image reconstruction framework – time scheduled
  - Visualisation on the 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

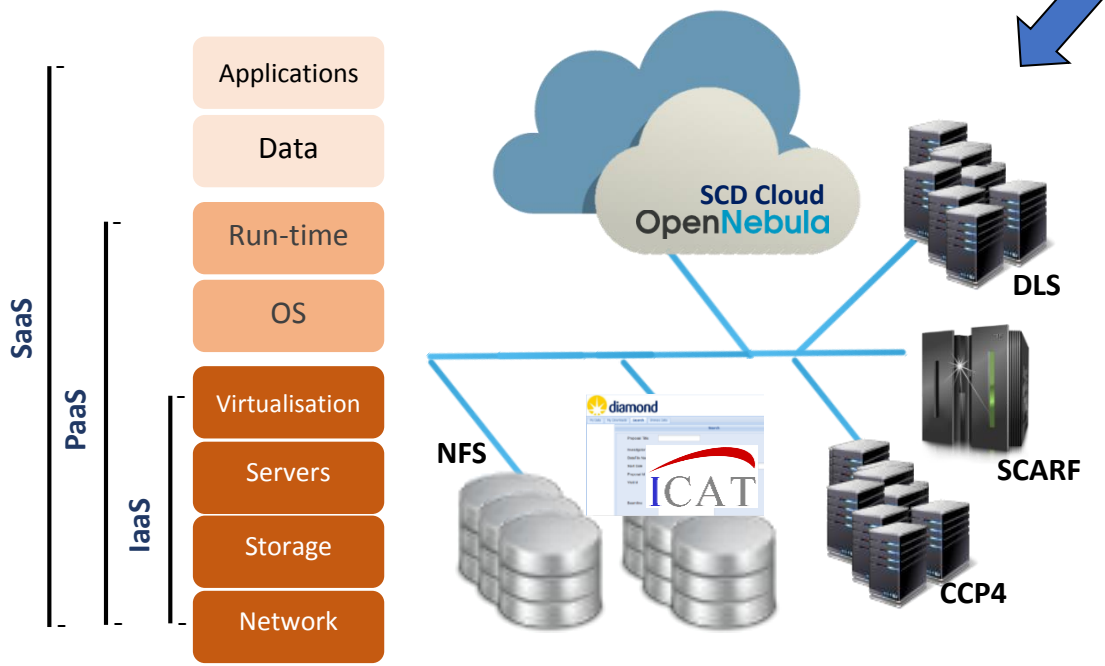
STFC Scientific Computing: Erica Yang, Srikanth Nagella, Martin Turner, Derek Ross
STFC ISIS: Winfried Kockelmann, Genoveva Burca, Federico Montesino Pouzols
DLS: Mark Basham

ALC Pathfinder: DAaaS Project

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





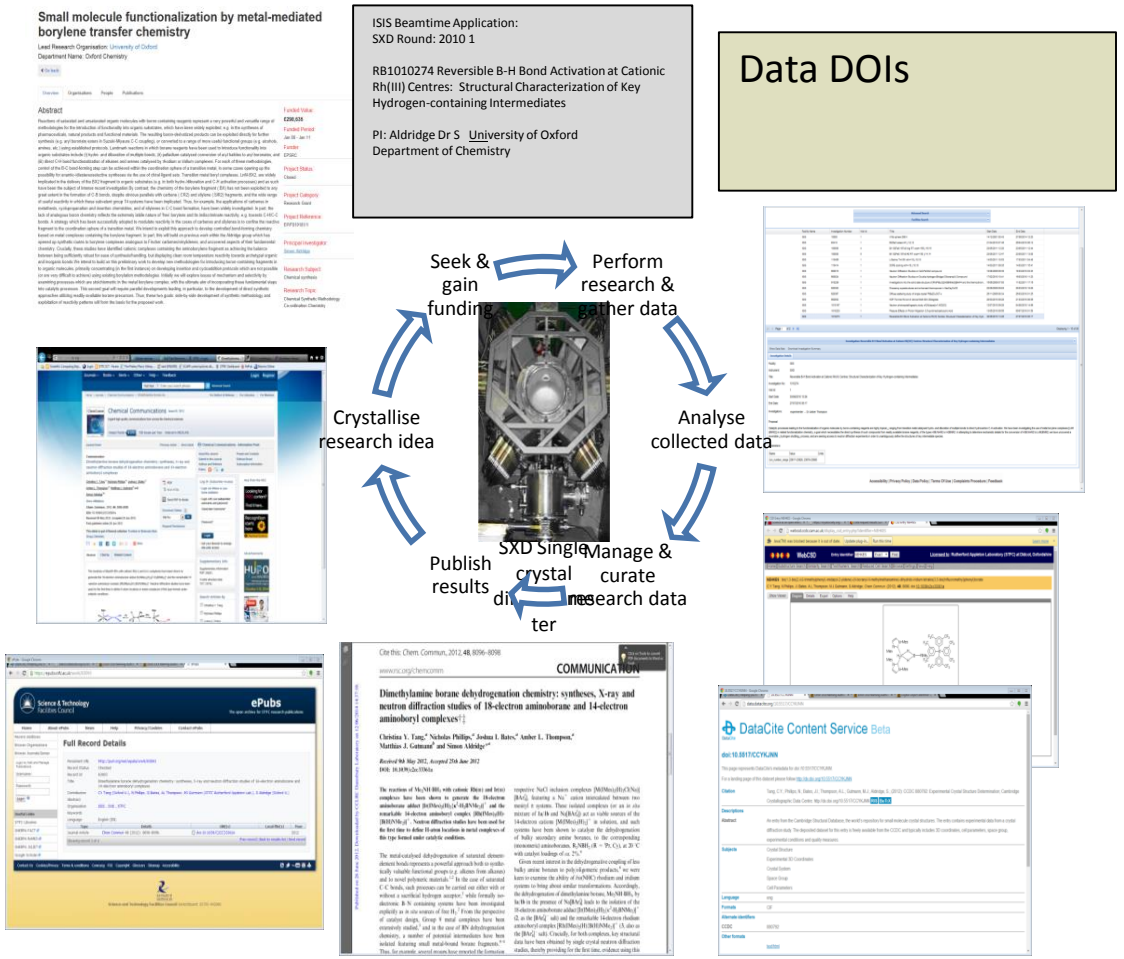
Data Analysis as a Service

- 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

Thanks to Frazer Barnsley, Shirley Crompton, CCP4, et al.

New Opportunities: Reproducible Science

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



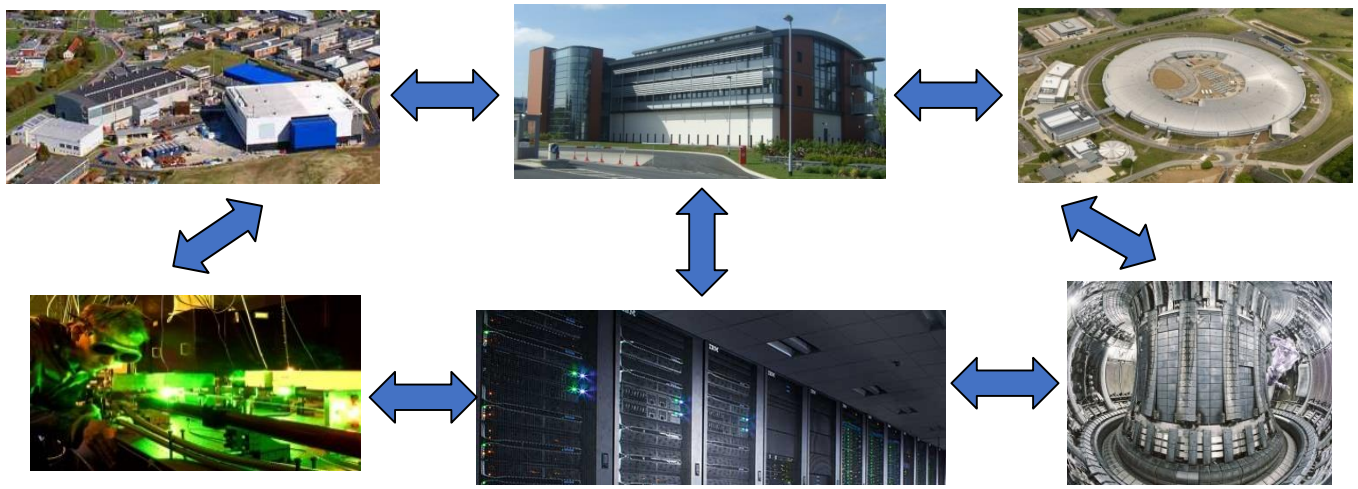
➢ ALC can build in support for open and reproducible science

Thanks to Catherine Jones

The ALC - Towards a "Super-facility"?

Infrastructure + Software + Expertise

With Common Interfaces and Transparent Access



Data Acquisition | Data Catalogue | Petabyte Data storage | Parallel File system | HPC **CPU+GPU** | Software | Visualisation

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

Katie Antypas, Inder Monga, Lawrence Berkeley National Laboratory

Acknowledgements

With thanks to Alun Ashton, Mark Basham, Gordon Brown, David Corney, Jonathan Churchill, Nicola Guerrini, Mark Heron, Imanol Luengo, Barbara Montanari, Brian Matthews and Dan Rolfe