AI for Big Science: What can AI do for National Large-scale Experimental Facilities?

Professor Tony Hey
Chief Data Scientist
Rutherford Appleton Laboratory, STFC
tony.hey@stfc.ac.uk

The Deep Learning Revolution

Many Machine Learning Methods

K-means clustering

Markov random fields

Bayesian networks

Linear regression

Kalman filters

Random forests

Principal Component Analysis

Neural networks

Support Vector Machines

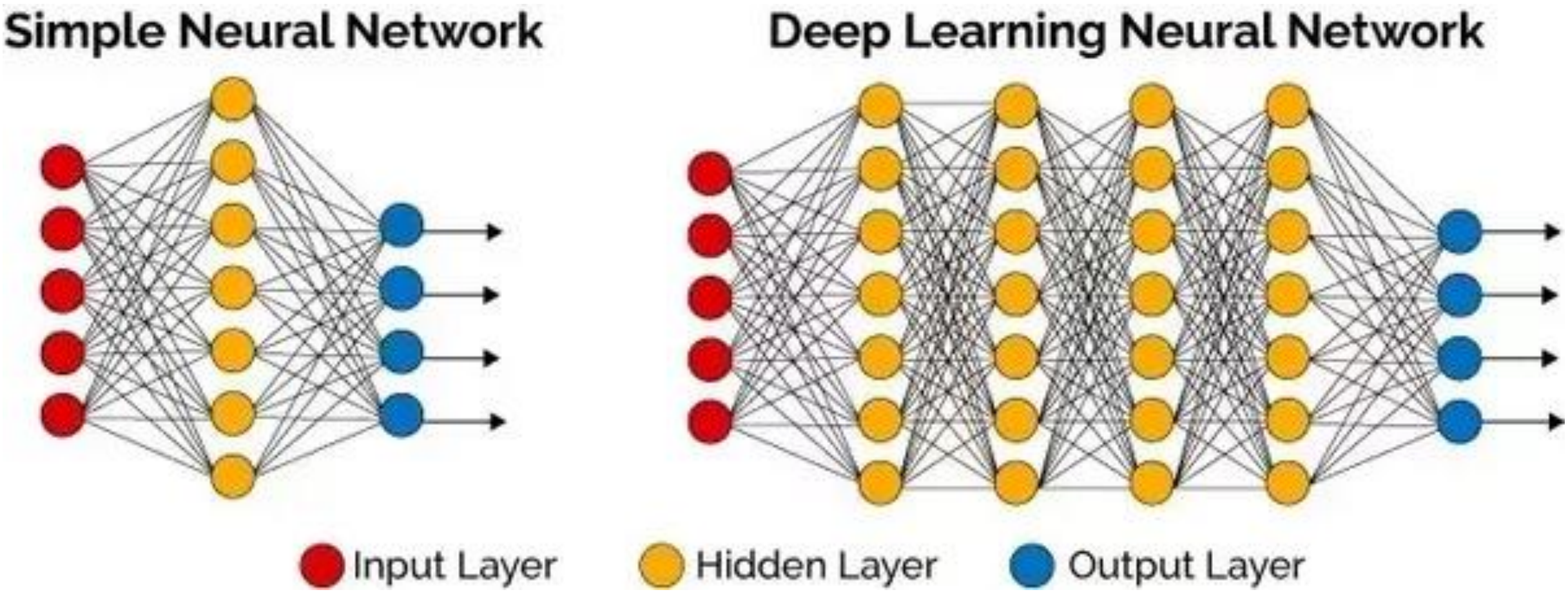
Boltzmann machines

Decision trees

Radial basis functions

Hidden Markov Models

The Deep Learning Revolution



IMAGENET

- ImageNet is an image dataset organized according to WordNet hierarchy. There are more than 100,000 WordNet concepts.
- ImageNet provides 1000 images of each concept that are quality-controlled and human-annotated.
- In competitions, ImageNet offers tens of millions of sorted images for concepts in the WordNet hierarchy.

What do these images have in common? *Find out!*

Check out the ImageNet Challenge 2017

- The ImageNet dataset has proved very useful for advancing research in computer vision

ImageNet Image Recognition Challenge

Image recognition challenge

ImageNet: 1000 categories, 1.2 million images

Classification error rate

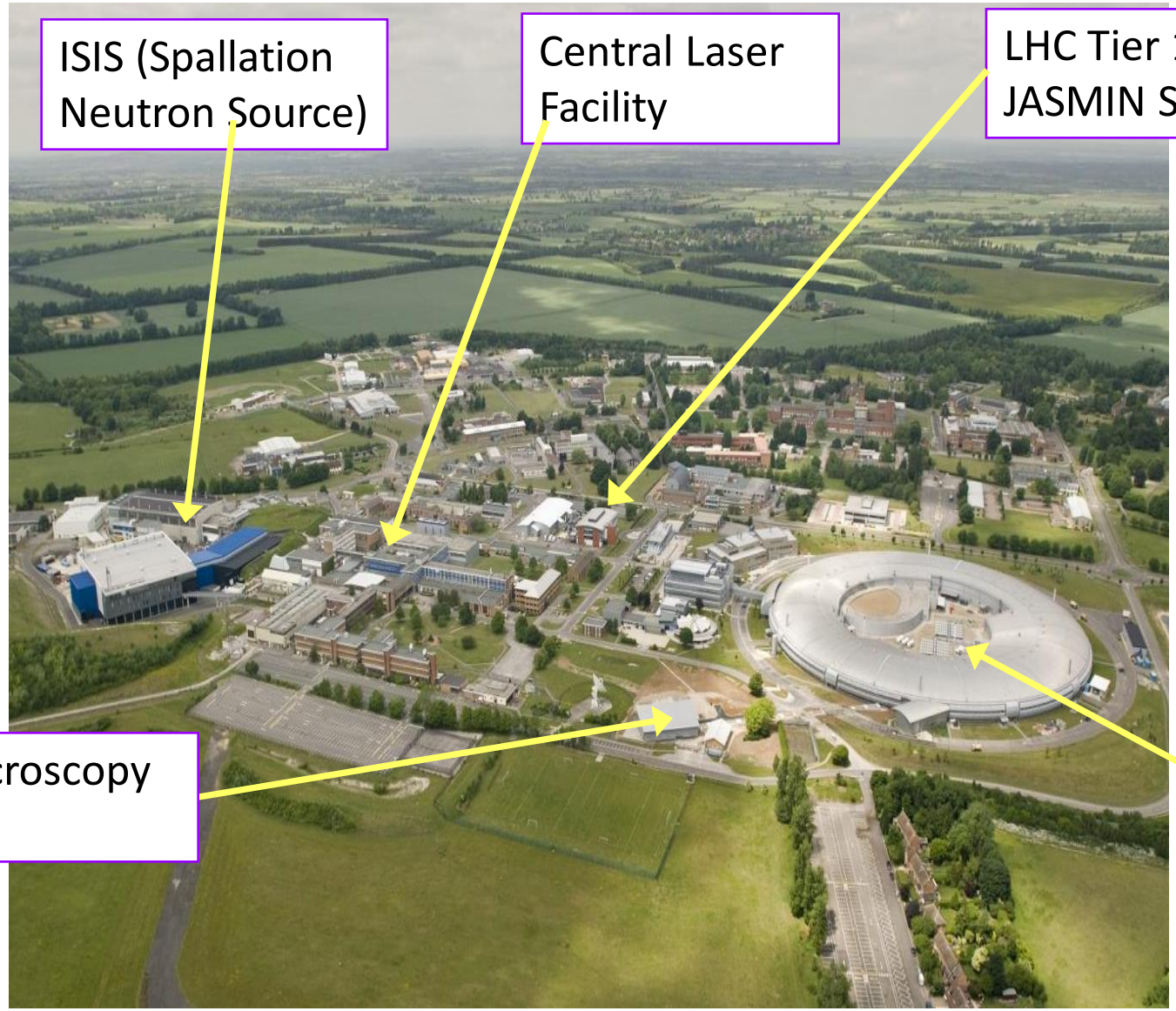
Deep learning errors < humans

O. Russakovsky et al, arXiv:1409.0575; K. He, X. Zhang, S. Ren, J. Sunar, arXiv:1512.03385

WMW Jie Hu, Li Shen (Oxford), Gang Sun, 2017

‘AI for Science’
at Rutherford Appleton Laboratory

Rutherford Appleton Laboratory (RAL)

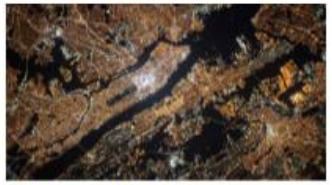


Scientific Machine Learning at RAL

- The mission of the Scientific Machine Learning (SciML) Group is to explore the use of AI and Machine Learning technologies on the 'Big Scientific Data' generated by the national large-scale experimental facilities hosted on the RAL site:
  - The Diamond Light Source
  - Electron Microscopy Facility
  - The ISIS Neutron and Muon Source
  - The Central Laser Facility
  - The Centre for Environmental Data Analysis (CEDA/JASMIN)
- SciML recently received funding from the Alan Turing Institute to act as a 'Turing Hub' for their 'AI for Science' program

Research Themes at the Alan Turing Institute

Artificial intelligence (AI) →
Advancing world-class research into artificial intelligence, its applications and its implications for society, building on our academic network's wealth of expertise.



Data science at scale →
Building upon advances in high-performance computer architectures, through algorithm-architecture co-design, with applications including health and life science.

Data-centric engineering →
Bringing together world-leading academic institutions and major industrial partners from across the engineering sector, to address new challenges in data-centric engineering.

Defence and security →
Collaborating with the defence and security community to deliver an ambitious programme of data science research, to deliver impact in real world scenarios.

Finance and economics →
Applying data science and AI techniques to how the financial sector and the economy work, and using these insights to address challenges of national and international importance.

Health →
Accelerating the scientific understanding of human disease and improving human health through data-driven innovation in AI and statistical science.

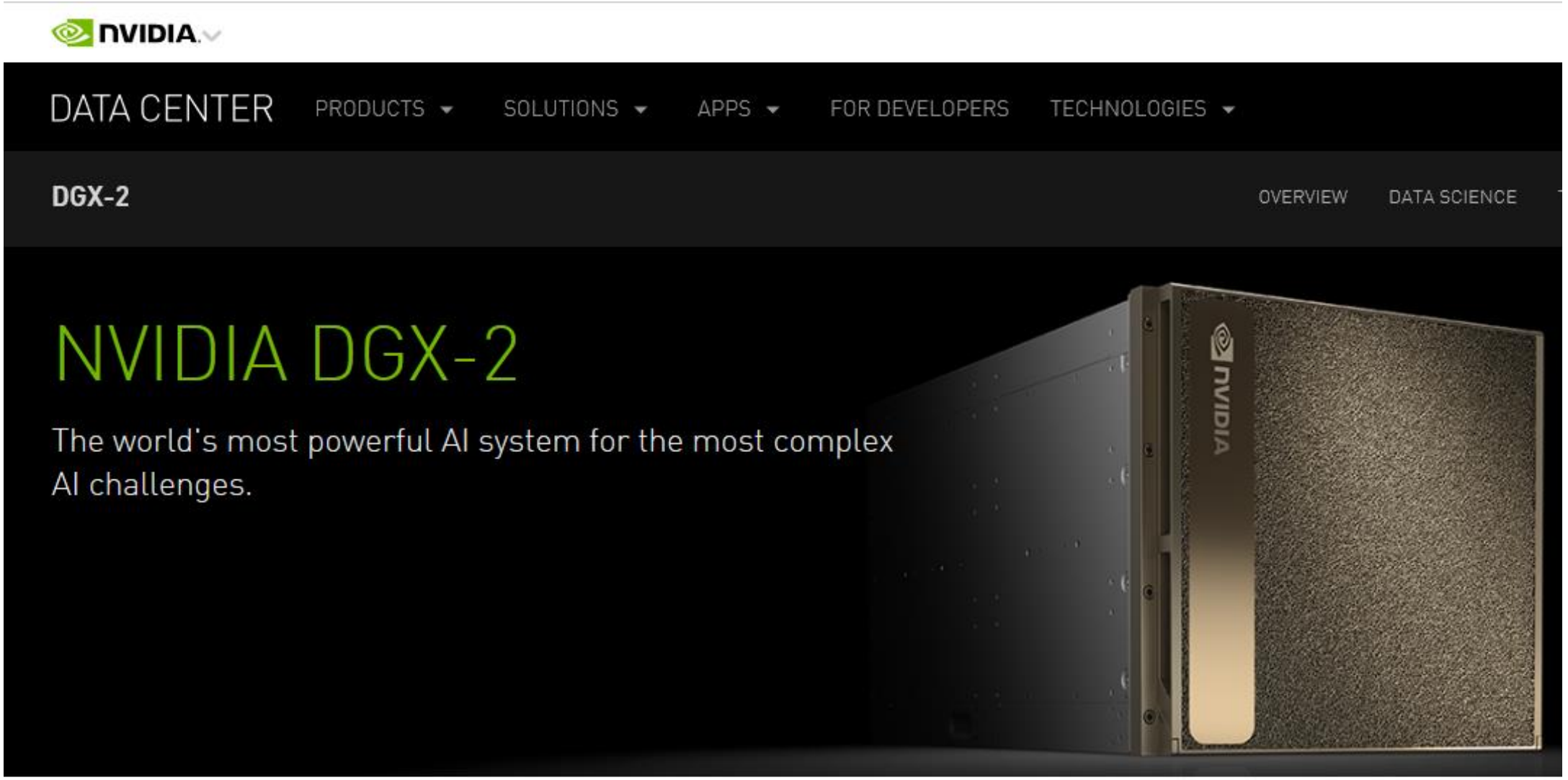
Public policy →
Working with policy makers on data-driven public services and innovation to solve policy problems, and developing ethical foundations for data science and AI policy-making.

Research Engineering →
Connecting research to applications, helping create usable and sustainable tools, practices and systems.

Urban analytics →
Developing data science and AI focused on the process, structure, interactions and evolution of agents, technology and infrastructure within and between cities.

Data science for science →
Ensuring that research across science and the humanities can make effective use of state of the art methods in artificial intelligence and data science.

'PEARL' GPU Computing Service for Turing Projects



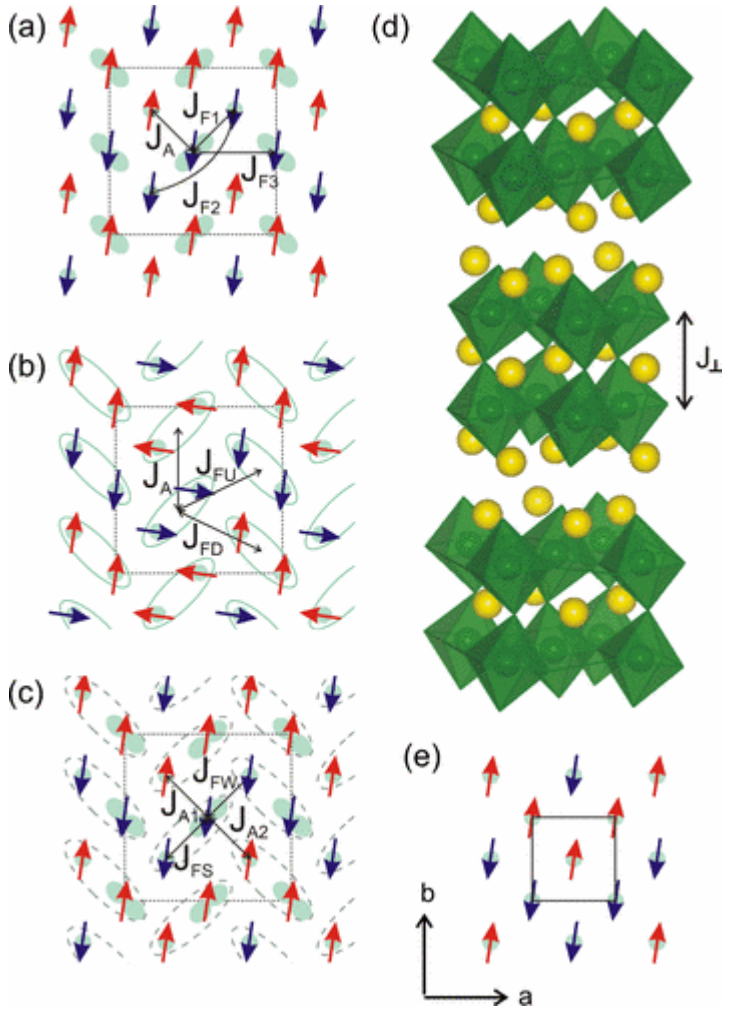
Machine learning for Inelastic Neutron Scattering

Keith Butler and Toby Perring

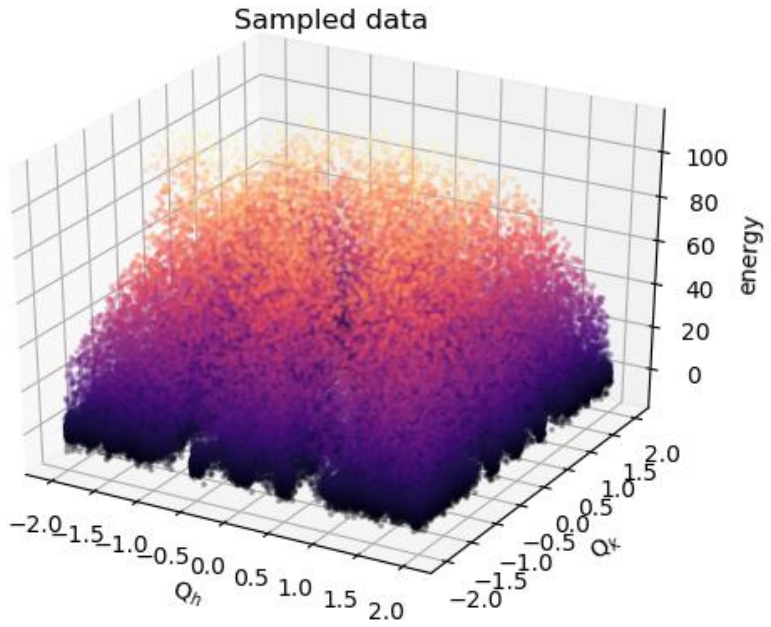
SciML Group and ISIS

Rutherford Appleton Laboratory

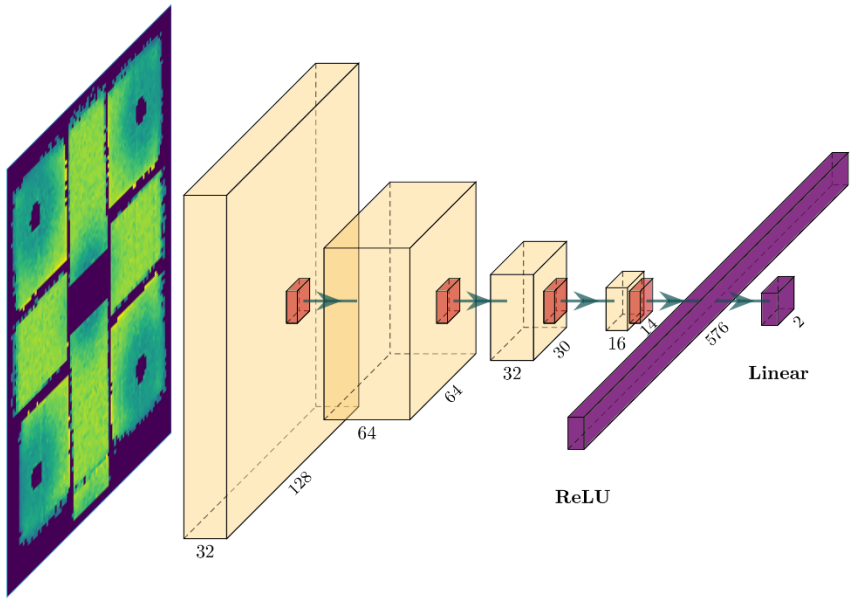
Magnons – finding a needle in a haystack



- Inelastic neutron scattering can, in principle distinguish between spin configurations – crucial for understanding quantum materials
- But the data is high-dimensional and finding the right areas for investigation is very difficult



Machine Learning for Hamiltonian parameters



- Given a physical model learn to infer parameters from data
- Train on a set of simulated spectra
- Range of possible J's set by prior knowledge
- Convolutional network, with linear activation function
- Convolutional filters learn to represent the data
- Dense ReLU + Linear layer learn a function from representation to measured property

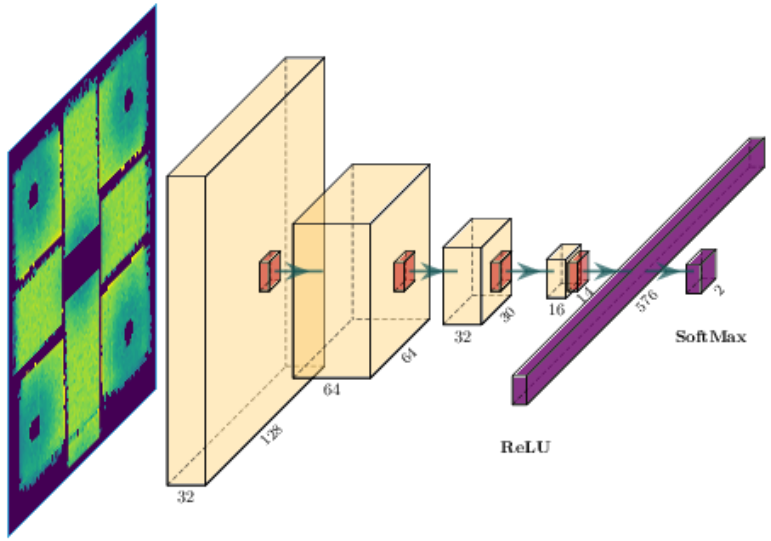
$$H = J_1 \sum_{<nn>} S_i \cdot S_j + J_2 \sum_{<nnn>} S_i \cdot S_j$$

- Given the experimental data the model predicts:
  J1=0.676; J2=0.014

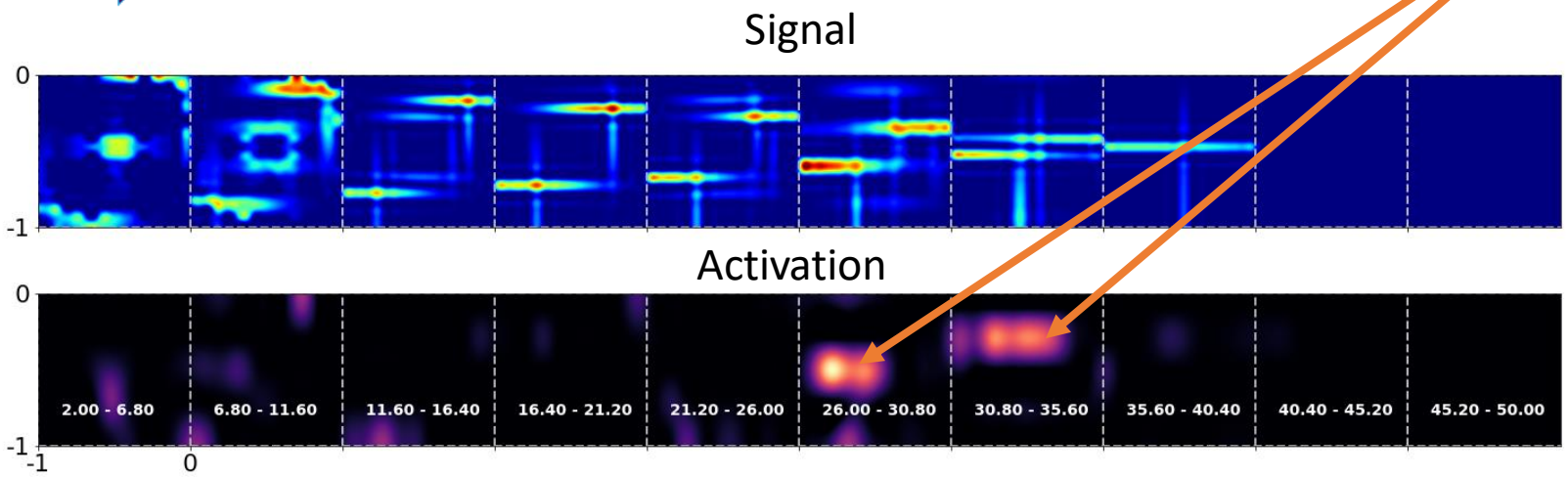
- Previous experimental analysis gives values of:
  J1=0.657±0.002; J2=0.003±0.009
  and
  J1=0.673±0.028; J2=0.012±0.002

ML can highlight important regions

Step 1:
CNN discriminates phases



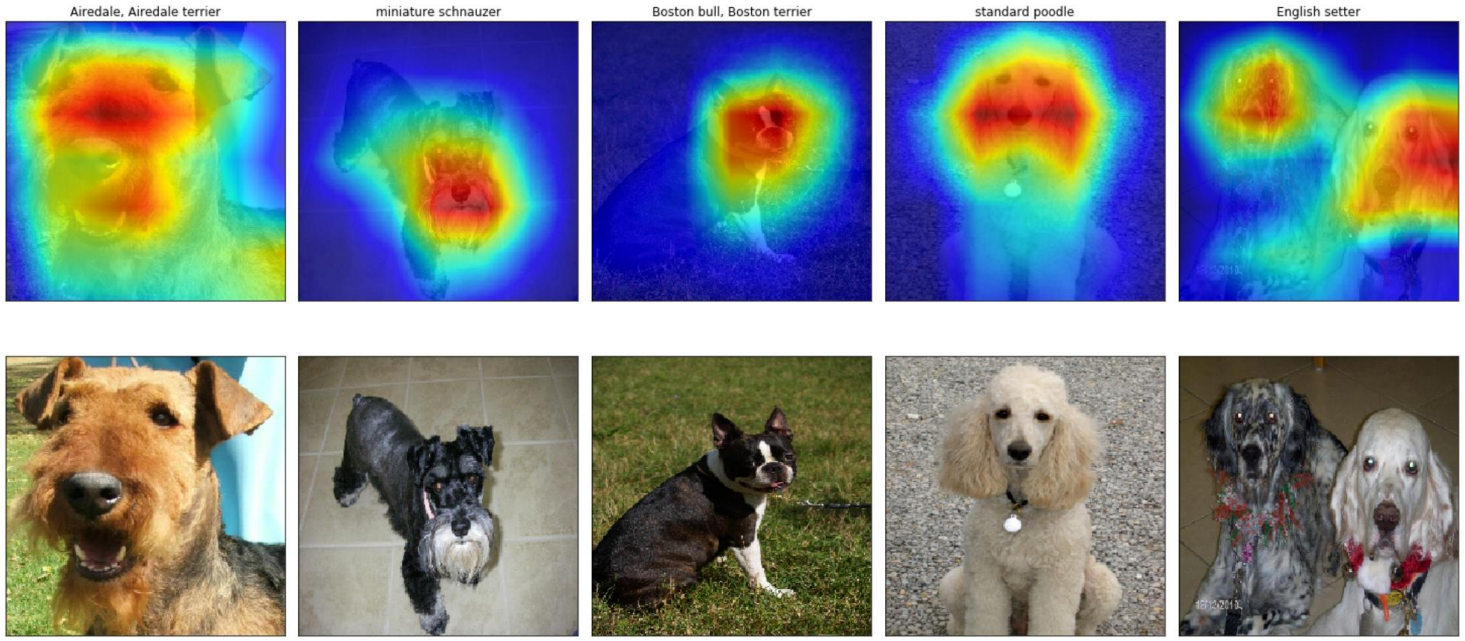
Step 2:
Interrogate CNN to produce activation maps



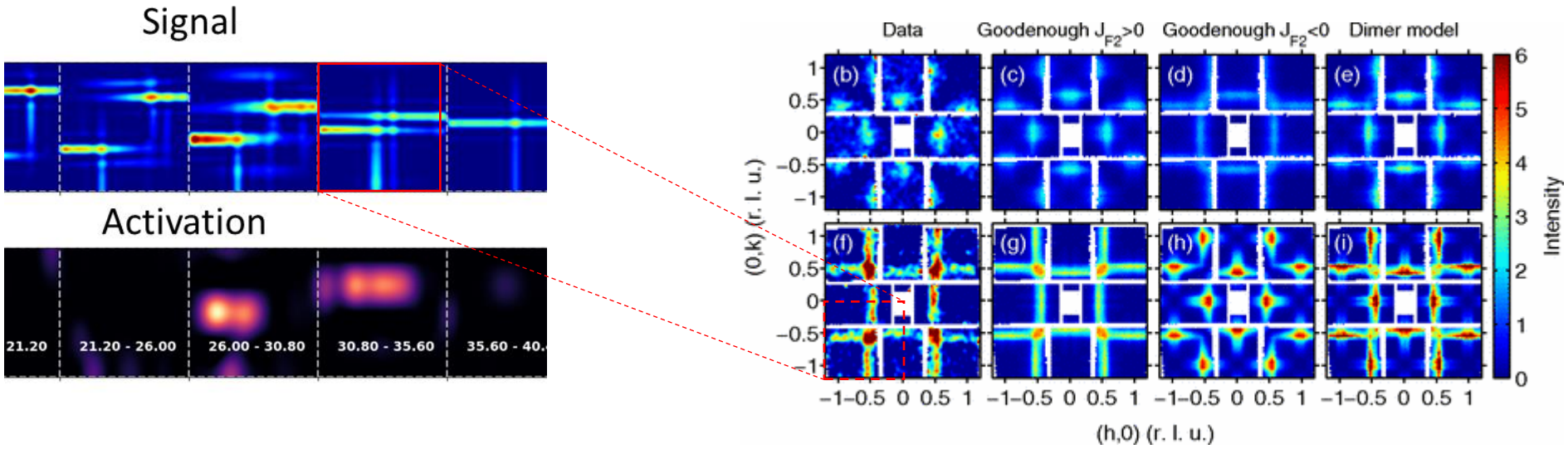
Network activation maps identify the regions in signal space that maximise discrimination

Learning from the result: Class Activation Maps

Class activation maps highlight which region of the image is most responsible for a classification



Compare our ML-generated heatmap to original publication



- The energy slice used in the original publication to discriminate the models is 34 – 36 meV
- The activation map finds the same region as a skilled physicist for discriminating between models.

Phys. Rev. Lett. 109, 237202

Machine learning for CryoEM

Jola Mirecka and Tom Burnley
SciML Group and Computational Biology Group
Scientific Computing Department,
Rutherford Appleton Laboratory

CCP-EM

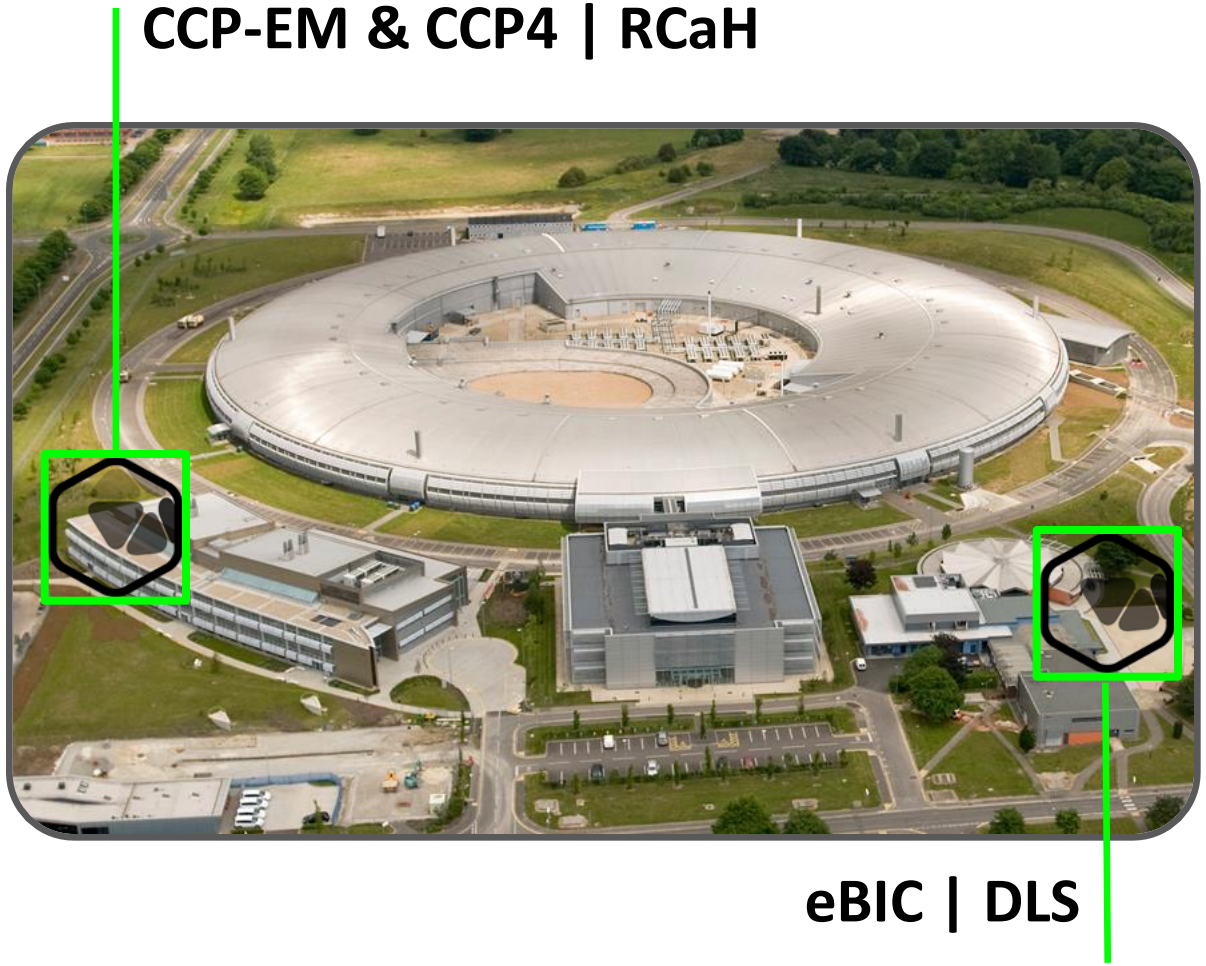
COLLABORATIVE COMPUTATIONAL PROJECT ON ELECTRON MICROSCOPY

Hosted by SCD / STFC @ RAL

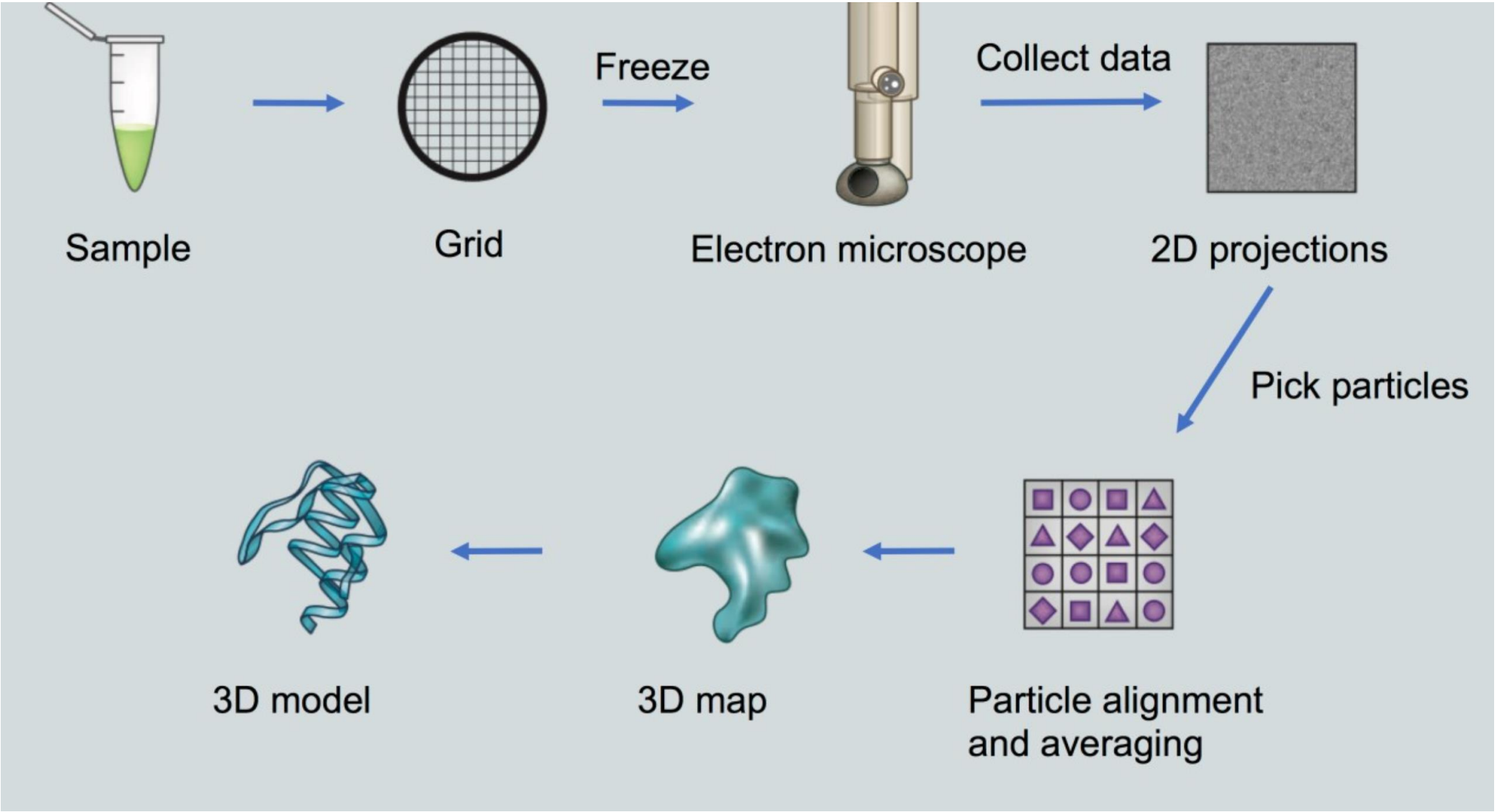
Support **users**, **developers** and **facilities** in computational aspects of biological cryo-EM

CCP-EM software suite: tools for EM data processing from image processing to atomic model building

MRC funded since 2012



CryoEM Data Processing



CCP-EM

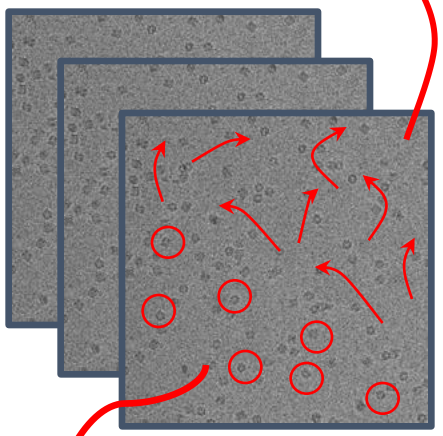
COLLABORATIVE COMPUTATIONAL PROJECT ON ELECTRON MICROSCOPY

SciML

MOTION CORRECTION:

Jason Yeung and Jeyan Thiyagalingam

- Kalman filter for motion correction
- generative convolutional networks for image denoising



diamond

PARTICLE PICKING:

Donovan Webb and Yuriy Chaban

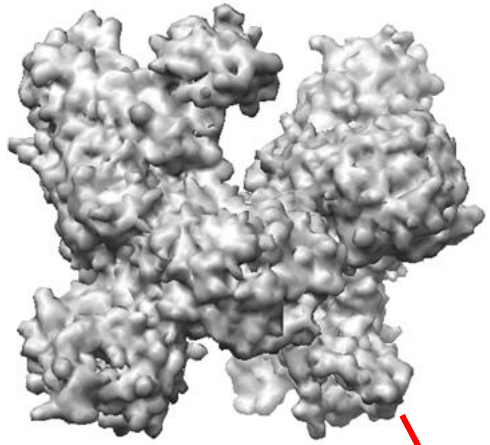
- benchmarking
- discriminative convolutional networks for particle picking

Julius-Maximilians-UNIVERSITÄT WÜRZBURG

SECONDARY STRUCTURE:

Andrea Thorn and Phillipp Mostosi

- encoder-decoder generative convolutional networks for secondary structure prediction

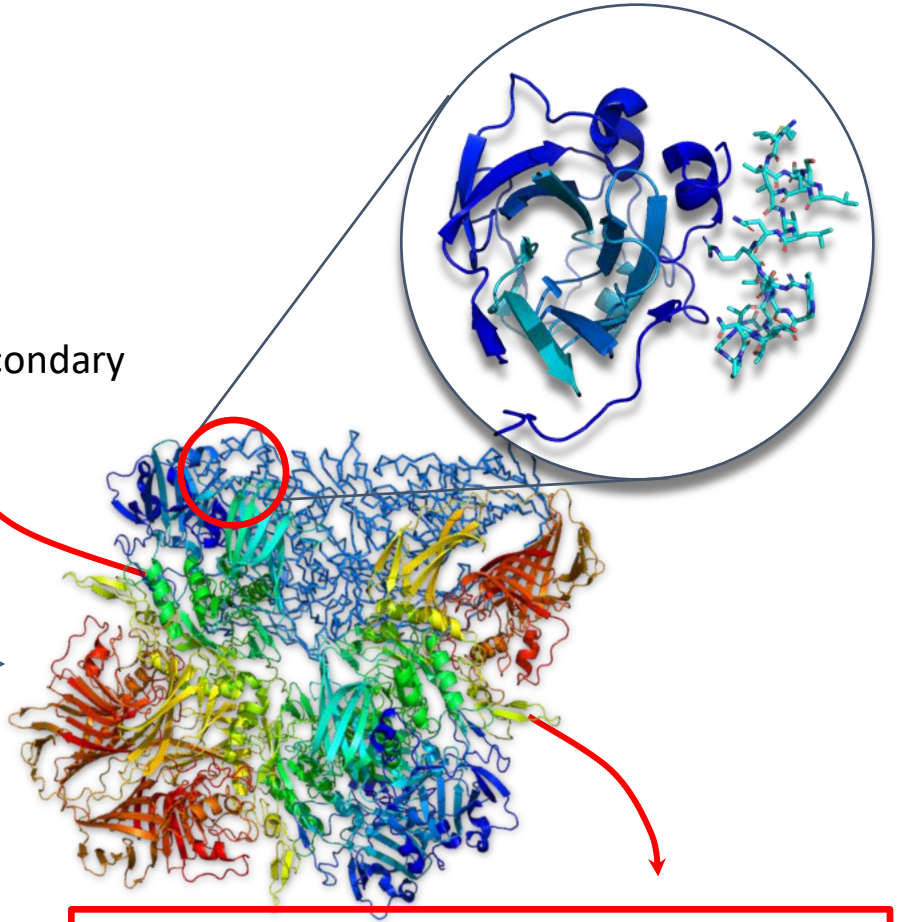


RELION MRC Medical Research Council

2D/3D RECONSTRUCTION:

Sjors Scheres, Dari Kaimanius and Liyi Dong

- neural network regression for improved 2D orientations selection



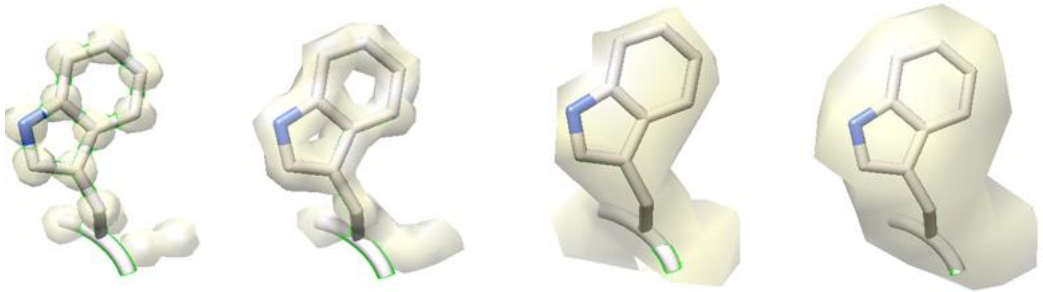
SciML

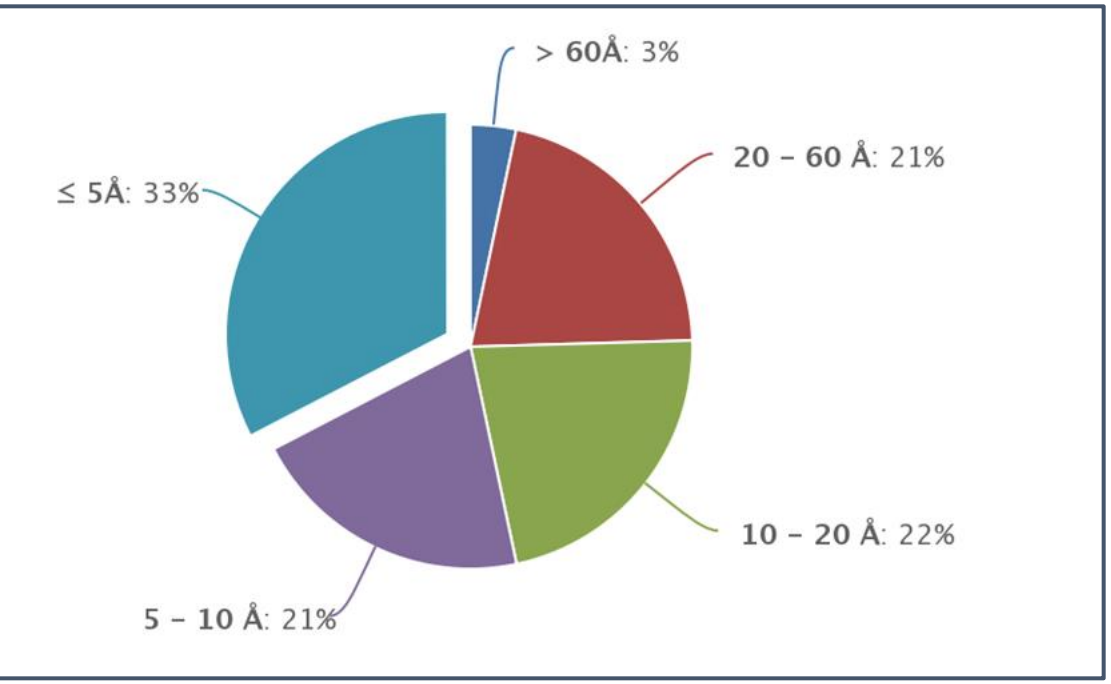
MODEL BUILDING:

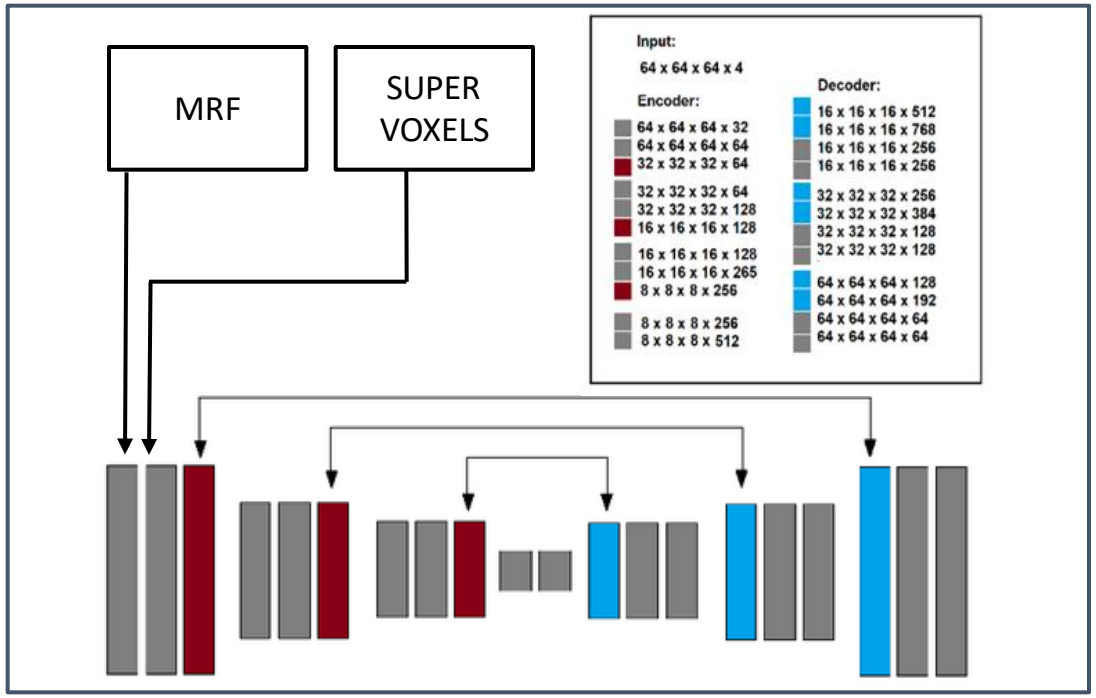
Jola Mirecka and Tom Burnley

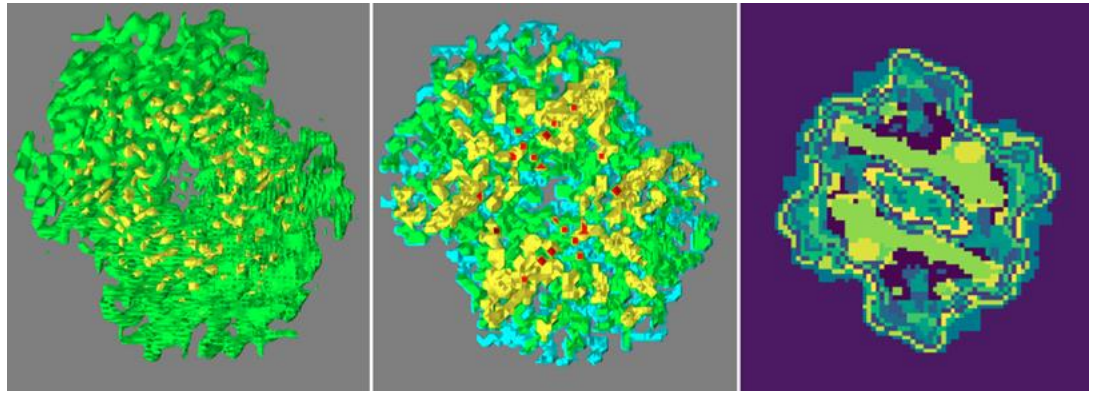
- hybrid generative neural networks for amino-acid identification at multi-resolutions

CCP-EM: MODEL BUILDING OVERVIEW







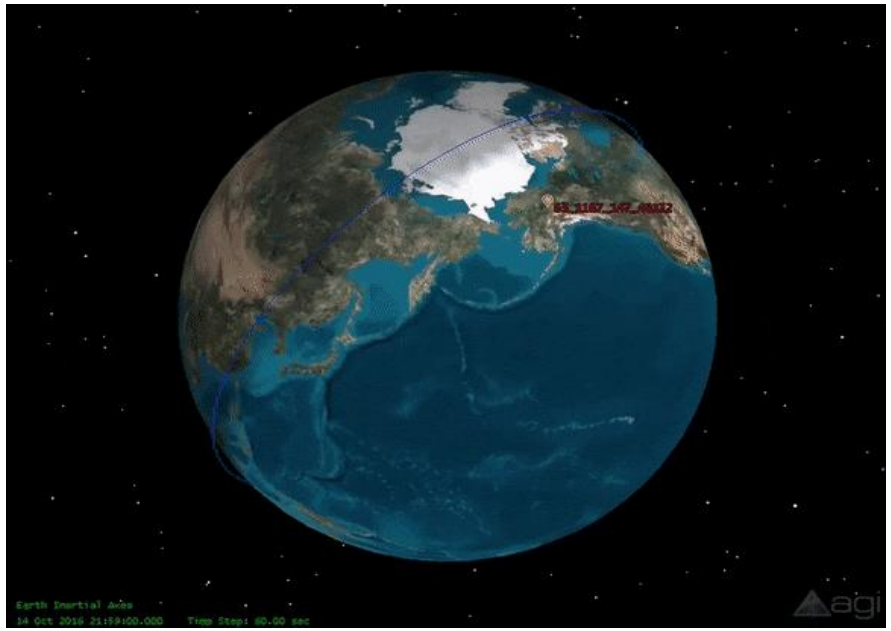


Machine learning for Cloud Identification from Satellite Data

Sam Jackson and Jeyan Thiyagalingam

SciML Group

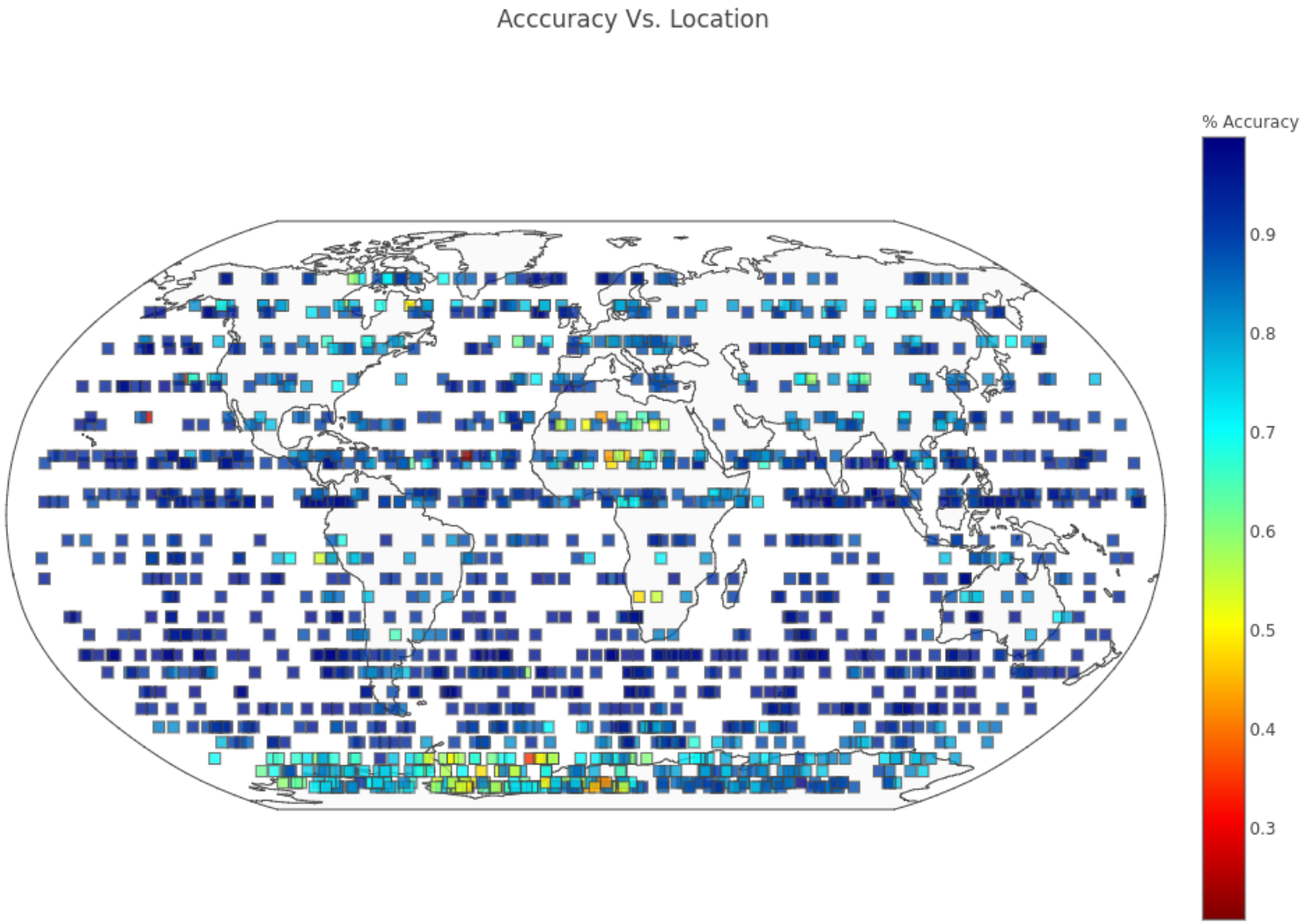
Rutherford Appleton Laboratory



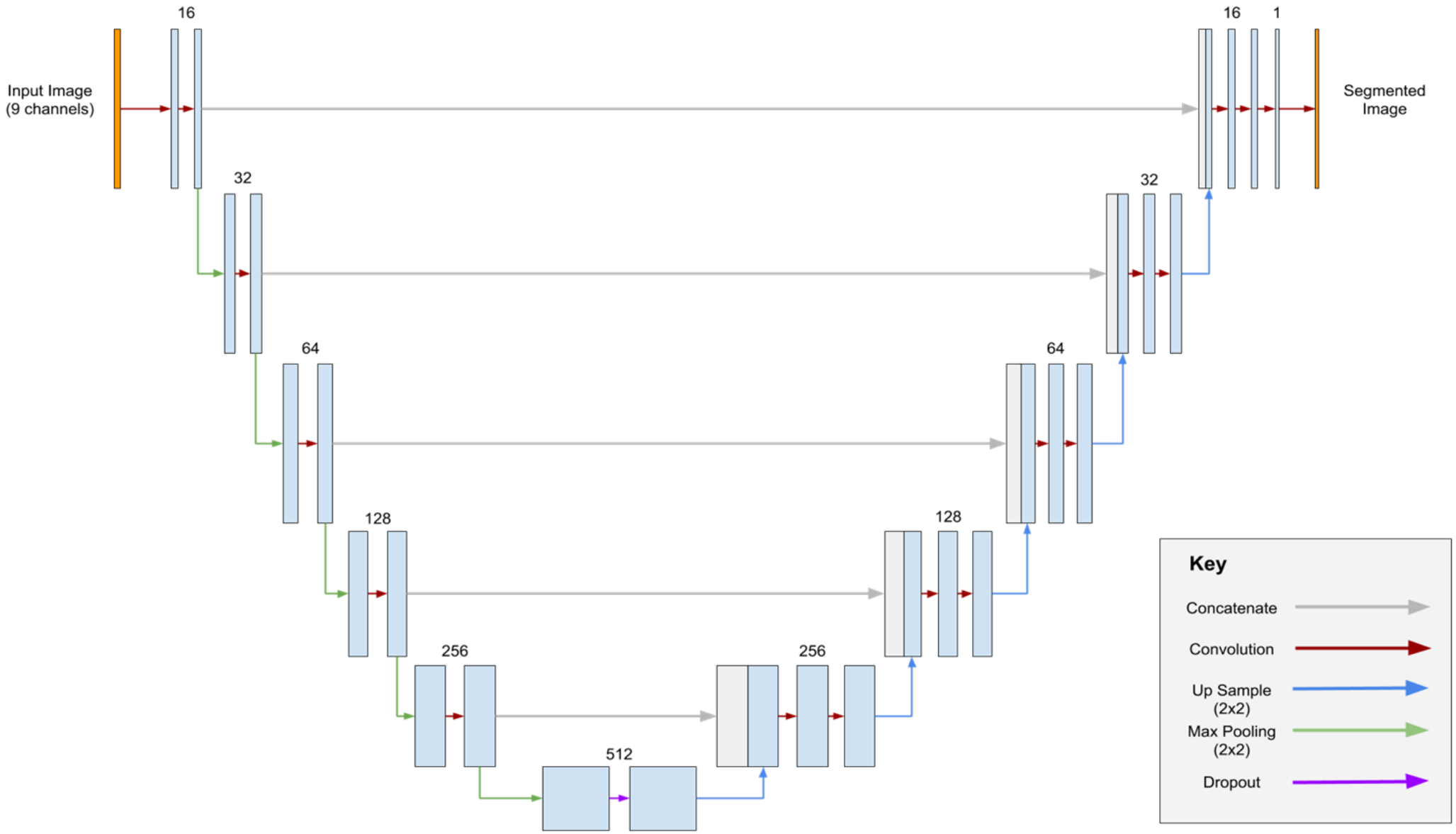
Sentinel 3A & 3B

- Two polar, Sun-synchronous satellites working in tandem
- Altitude: 815 km
- Carries
  - Sea and Land Surface Temperature Radiometer (SLSTR)
  - Ocean and Land Colour Instrument (OLCI)
- Revisit Times :
  - SLSTR: 1 day
  - OLCI : 2 days
  - Global coverage every 2 days (3A & 3B)

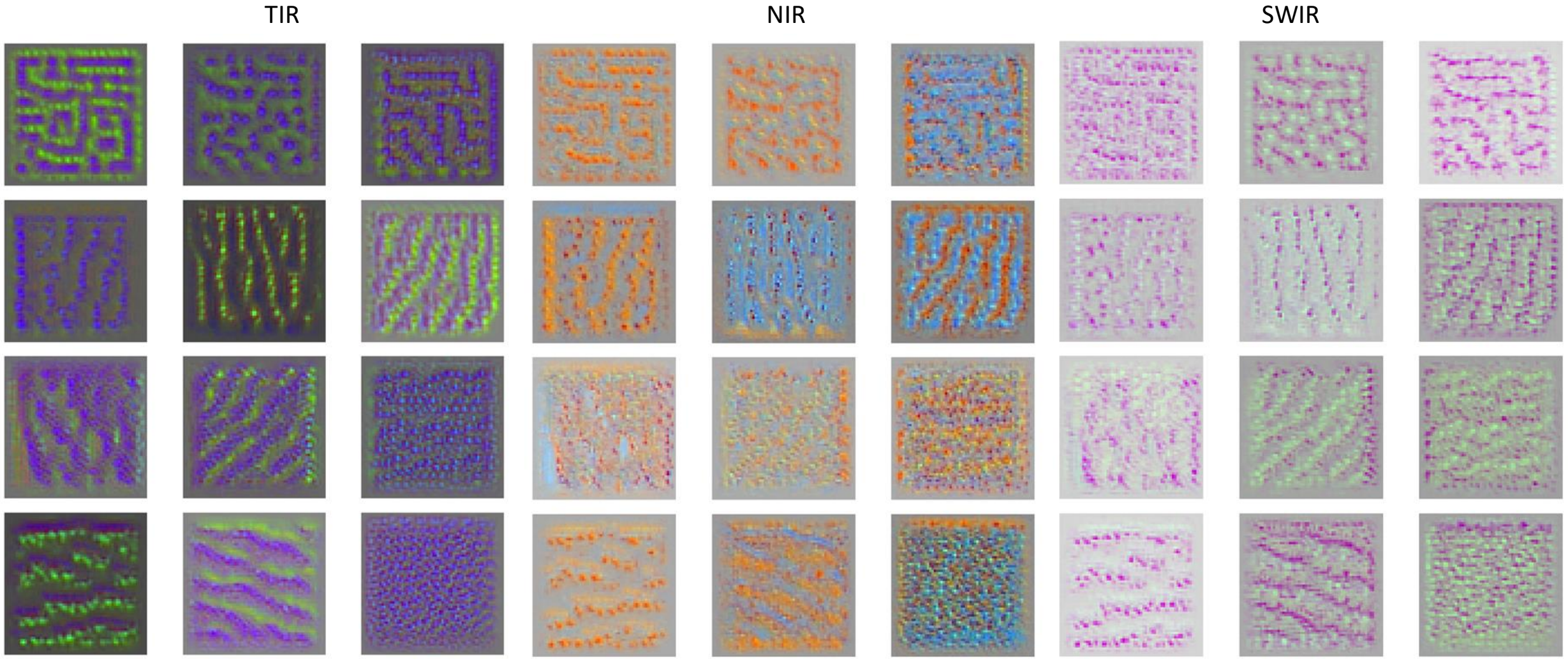
Global distribution of dataset against accuracy



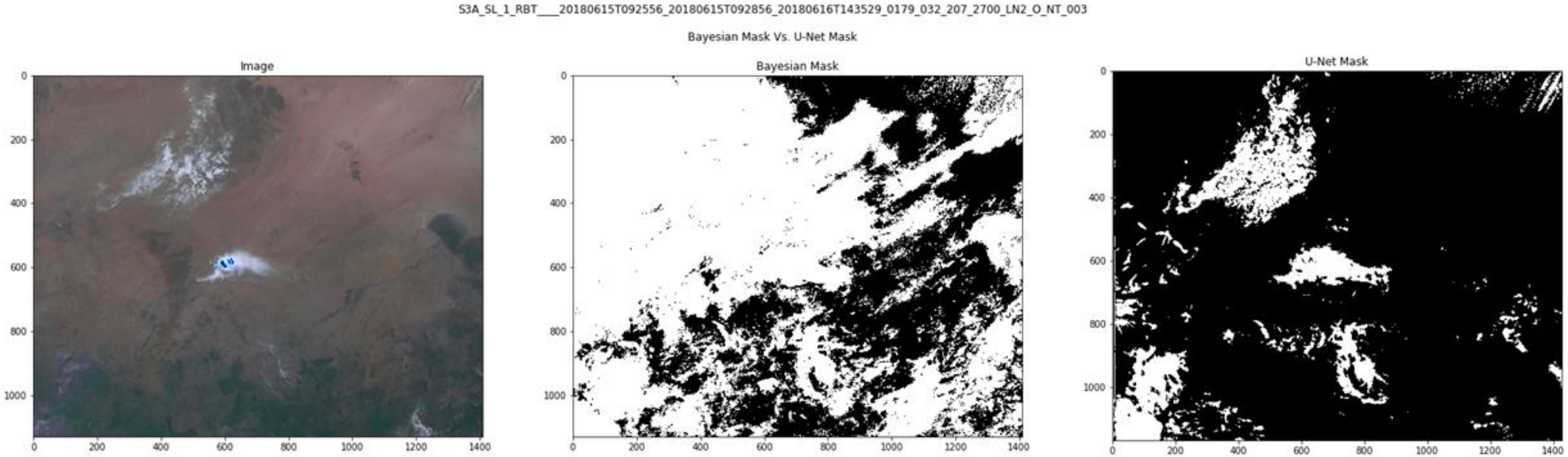
U-Net model for image segmentation



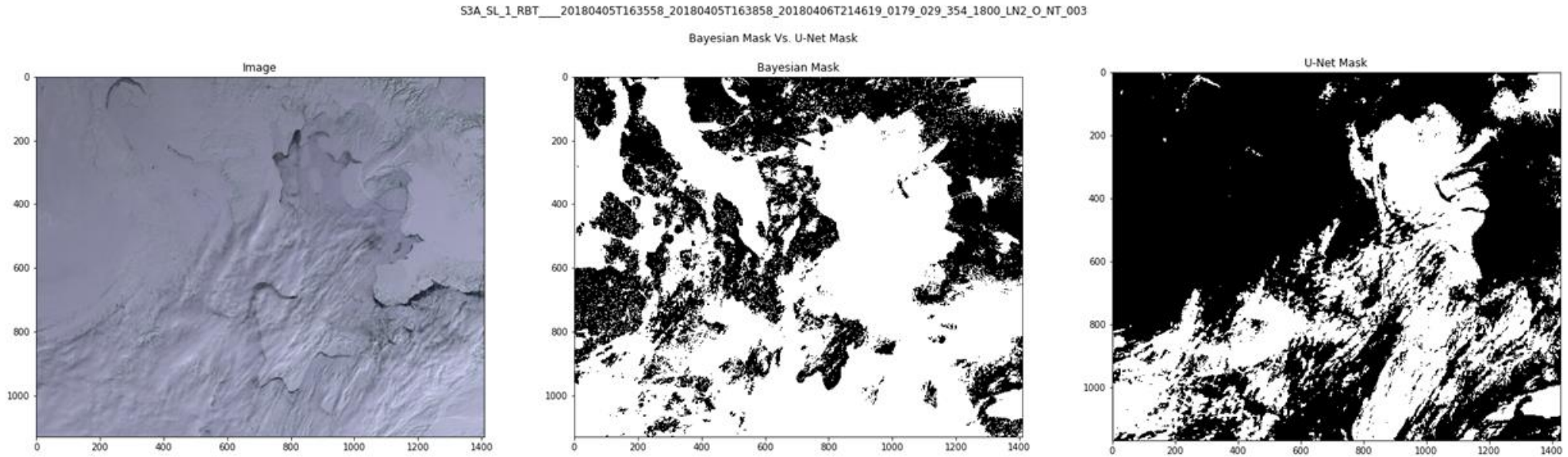
Texture Features From Final Layer



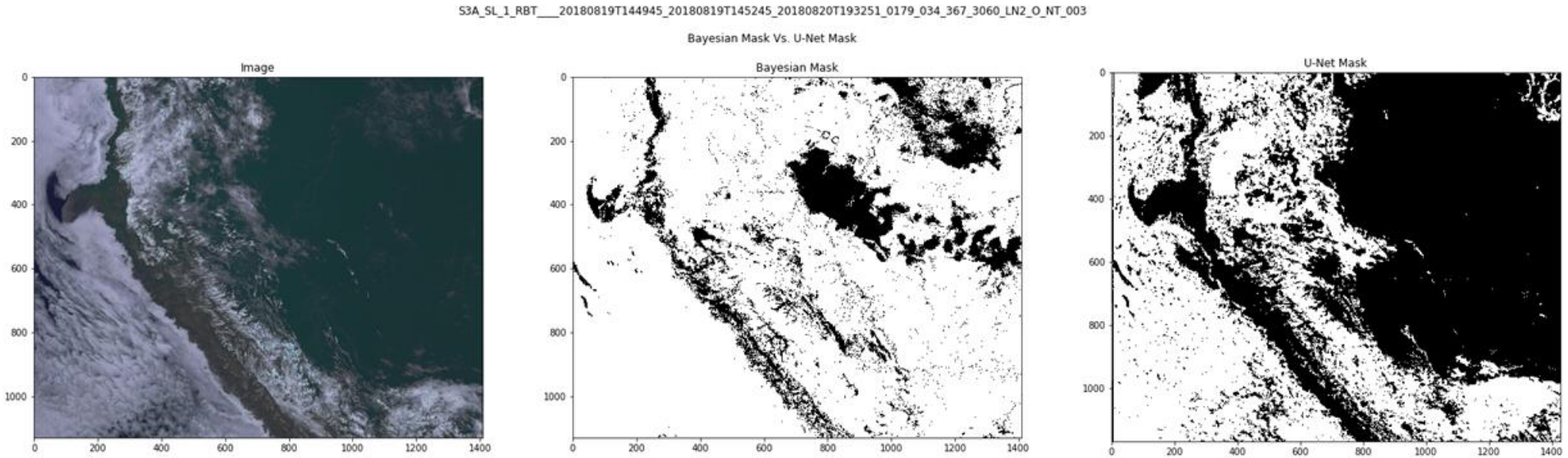
Clouds and Dust over the Sahara



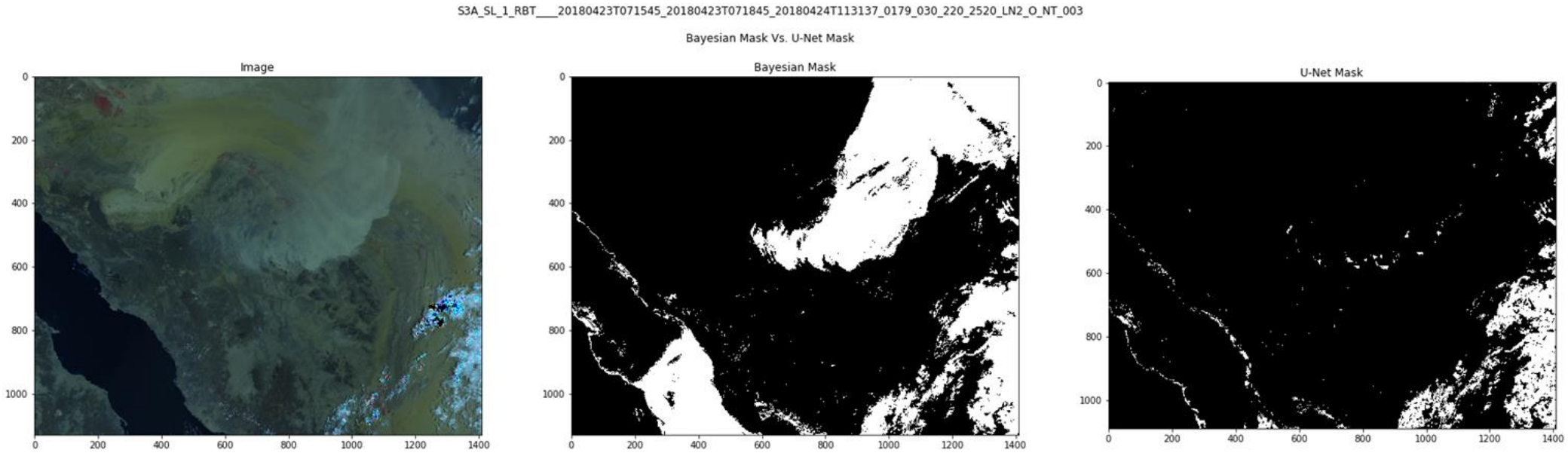
Ice Sheets over Northern Canada



Cloud front over the Amazon



Dust Storm over the Arabian Peninsula



Scientific Data Sets and Machine Learning Benchmarks

Scientific Datasets and ML Benchmarks at RAL

- Idea is to create scientific datasets that are sufficiently large and complex to provide a realistic testing ground for ML algorithms, software environments and hardware
- Currently have preliminary set of potential examples from several research communities:
  - ❑ Astronomy datasets from SDSS, DES, LSST, SKA, …
  - ❑ Particle Physics LHC datasets from ATLAS, CMS, DUNE, …
  - ❑ Large Scale Facilities datasets – DLS, ISIS and CLF
  - ❑ Environmental datasets from JASMIN
  - ❑ Datasets from Culham Centre for Fusion Energy

CANDLE Benchmarks

https://github.com/ECP-CANDLE

Benchmark Owners:
- P1: Fangfang Xia (ANL)
- P2: Brian Van Essen (LLNL)
- P3: Arvind Ramanathan (ORNL)

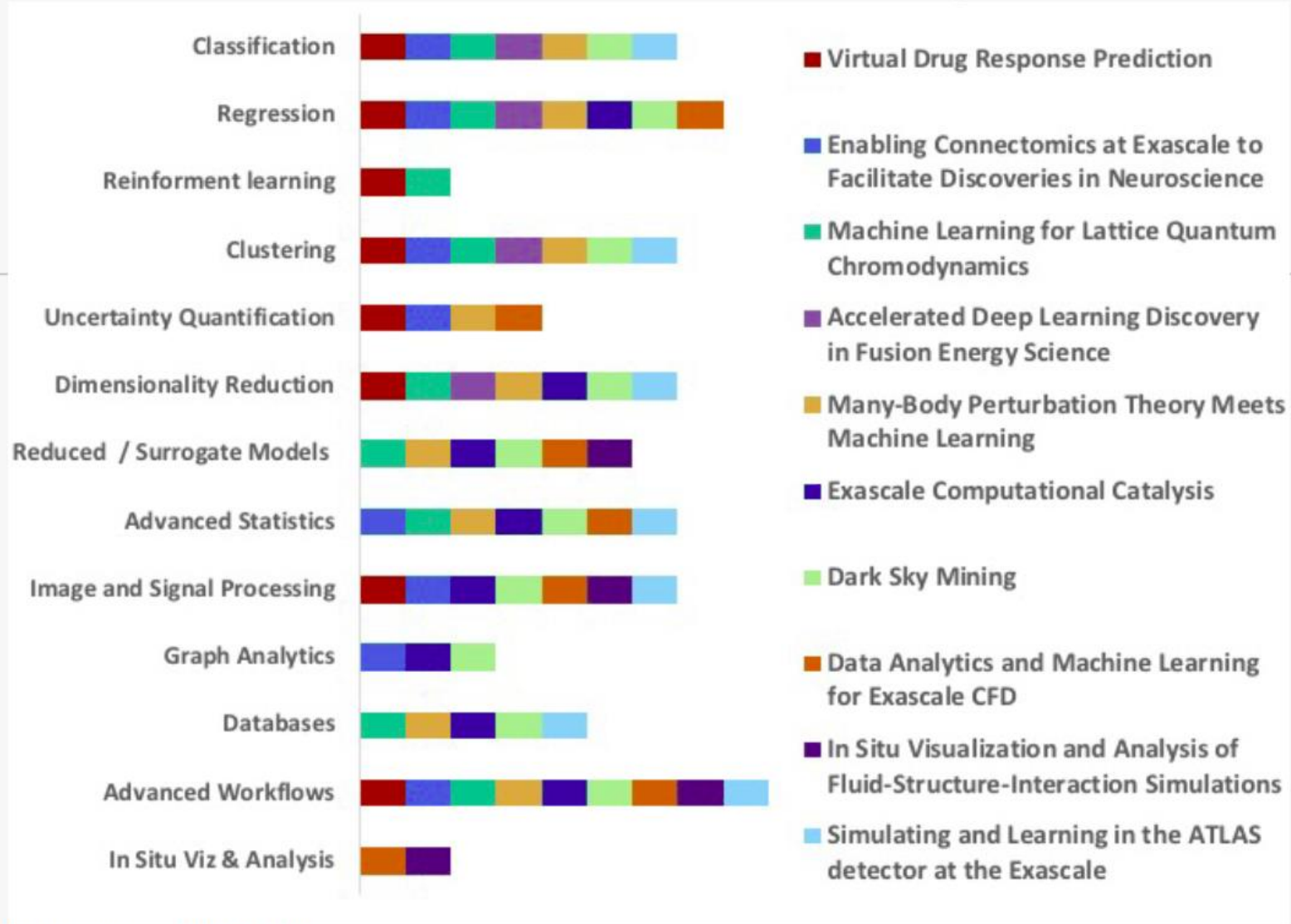
Benchmark	Type	Data	ID	OD	Sample Size	Size of Network	Additional (activation, layer types, etc.)
1. P1: B1 Autoencoder	MLP	RNA-Seq	10^5	10^5	15K	5 layers	Log2 (x+1) → [0,1] KPRM-UQ
2. P1: B2 Classifier	MLP	SN → Type	10^3	0	15K	5 layers	Training Set Balance issues
3. P1: B3 Regression	MLP+LCN	expression; drug descs	10^5	1	3M	8 layers	Drug Response [-100, 100]
4. P2: B1 Autoencoder	MLP	MD K-RAS	10^5	10^2	10^6-10^8	5-8 layers	State Compression
5. P2: B2 RNN-LSTM	RNN-LSTM	MD K-RAS	10^5	5	10^6	4 layers	State to Action
6. P3: B1 RNN-LSTM	RNN-LSTM	Path reports	10^3	5	5K	1-2 layers	Dictionary 12K +30K
7. P3: B2 Classification	CNN	Path reports	10^4	10^2	10^5	5 layers	Biomarkers

Drug Response

RAS Pathways

Patient Trajectories

AURORA ESP Data and Learning Methods



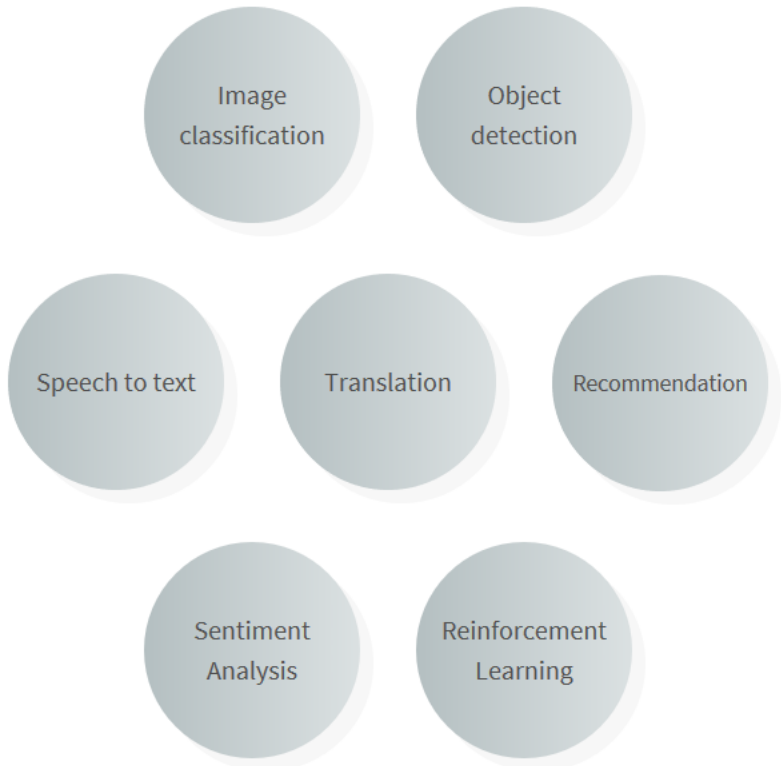
Learning

Data

Argonne NATIONAL LABORATORY

MLPerf

A broad ML benchmark suite for measuring performance of ML software frameworks, ML hardware accelerators, and ML cloud platforms.



Broad and Representative Problems

Like the Fathom Benchmark, the MLPerf suite aims to reflect different areas of ML that are important to the commercial and research communities and where open datasets and models exist. Here is our current list of problems.

MLPerf HPC Benchmarks

- Science domains (as many as possible)
  - HEP, NP, climate, cosmology, Precision Medicine (cancer)
- Problems
  - classification, regression, generative modeling, object detection
- Models
  - CNN, RNN, GAN, GraphNN
- Datasets
  - Various (open) scientific datasets, ImageNet
- Frameworks
  - Tensorflow, PyTorch, Keras, Horovod

Potential MLPerf HPC Benchmarks

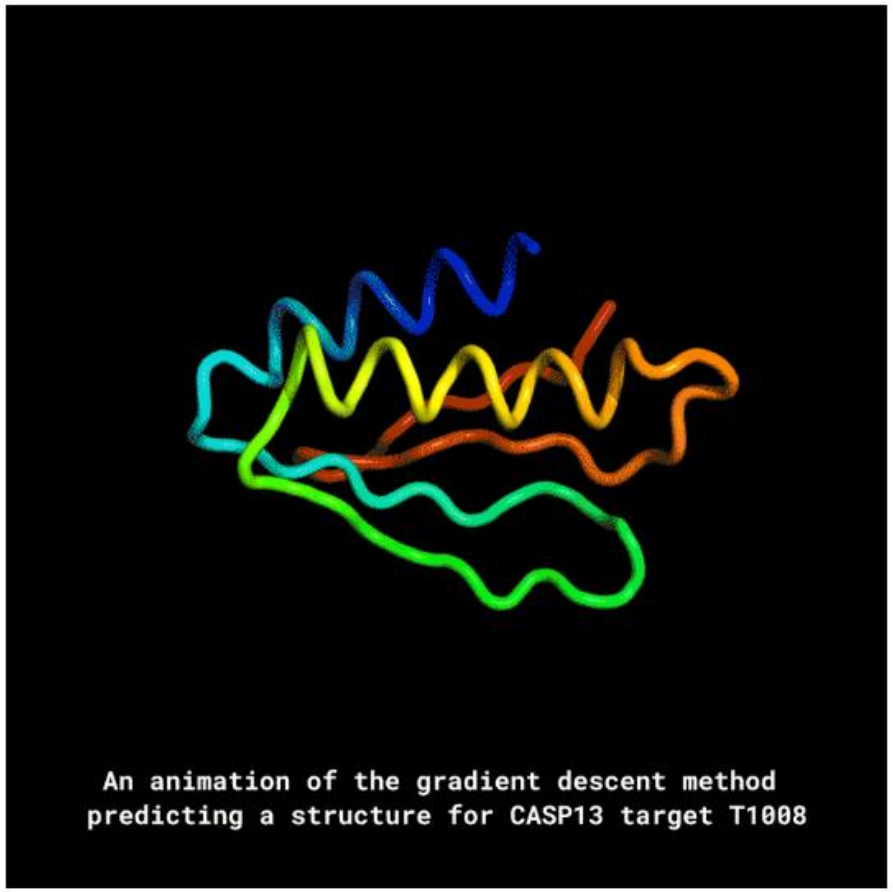
- CosmoFlow
  - Regression on cosmology volumes to predict cosmological parameters
  - https://arxiv.org/abs/1808.04728
  - Captures Scaling and I/O bottlenecks
  - 3D Convolutions
- DL Climate Analytics
  - Segmentation of climate images to identify extreme weather patterns
  - https://arxiv.org/abs/1810.01993
  - DeepLabv3 / Tiramisu 2D segmentation
  - Scaled to 16bit exaflop on summit, GB prize winner
- HEP-CNN
  - HEP image binary classification with generic CNN
  - Papers:
    - ACAT2017: https://arxiv.org/abs/1711.03573
    - SC17: https://arxiv.org/abs/1708.05256
    - CUG2018: https://cug.org/proceedings/cug2018_proceedings/includes/files/pap146s2-file1.pdf
- CosmoGAN
  - Generative model for cosmology weak lensing maps
  - https://arxiv.org/abs/1706.02390
  - 2D standard DCGAN

Scientific datasets forML Benchmarks will also help with …

- Understanding performance of new hardware for AI and Deep Learning – GPUs, TPUs, …
- Experimentation and training in Deep Learning technologies using significant size datasets executed on different hardware architectures
- Availability of challenging datasets for hands-on AI training for academia and industry
- Research on the uncertainties, robustness and explainability of DNNs
- Exploration of ML algorithms with scientific constraints for physics, chemistry and biology – example of AlphaFold!

AlphaFold: Using AI for scientific discovery

Our system, **AlphaFold**, which we have been working on for the past two years, builds on years of prior research in using vast genomic data to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—making significant progress on one of the core challenges in biology.



An animation of the gradient descent method predicting a structure for CASP13 target T1008