Convergence of HPC, Big Data and Machine Learning: A Science and Engineering Perspective

Professor Tony Hey
Chief Data Scientist
Rutherford Appleton Laboratory,
Science and Technology Facilities Council (STFC)
tony.hey@stfc.ac.uk

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…

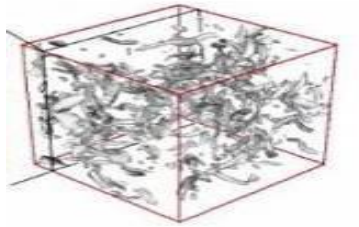
Last few decades – **Computational Science**

- 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

Particle Physics and Astronomy

kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

Higgs Boson Machine Learning Challenge

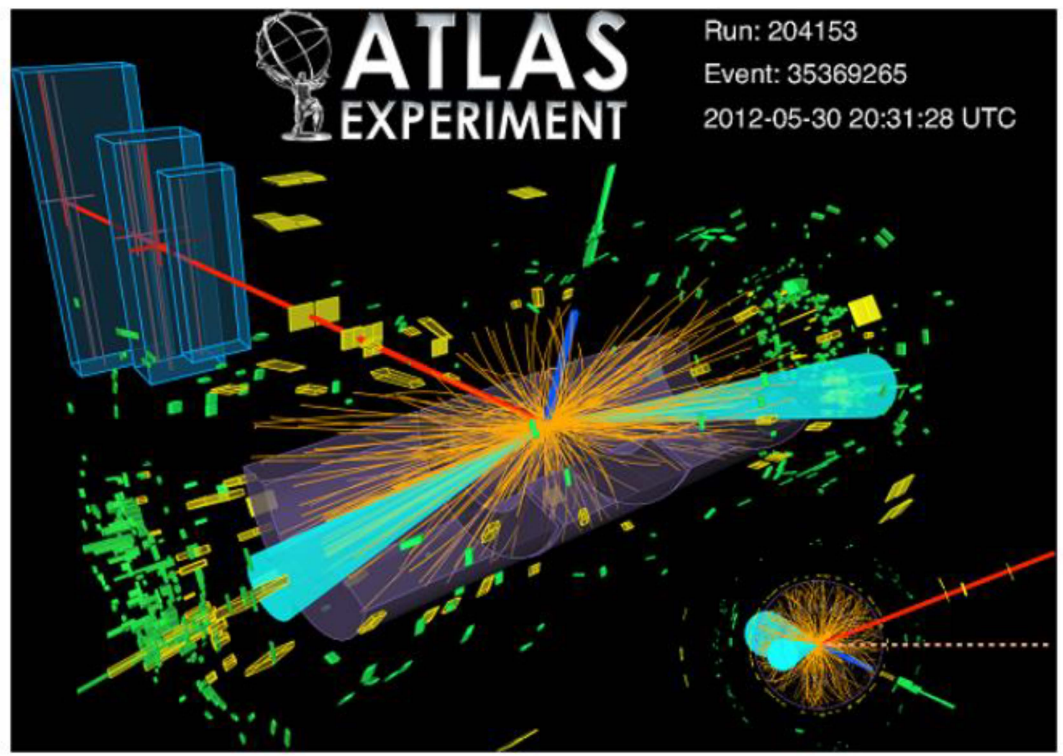
Use the ATLAS experiment to identify the Higgs boson

$13,000 · 1,785 teams · 4 years ago

Overview Data Discussion Leaderboard Rules

Overview

- Description
- Evaluation
- Prizes
- About The Sponsors
- Timeline
- Winners



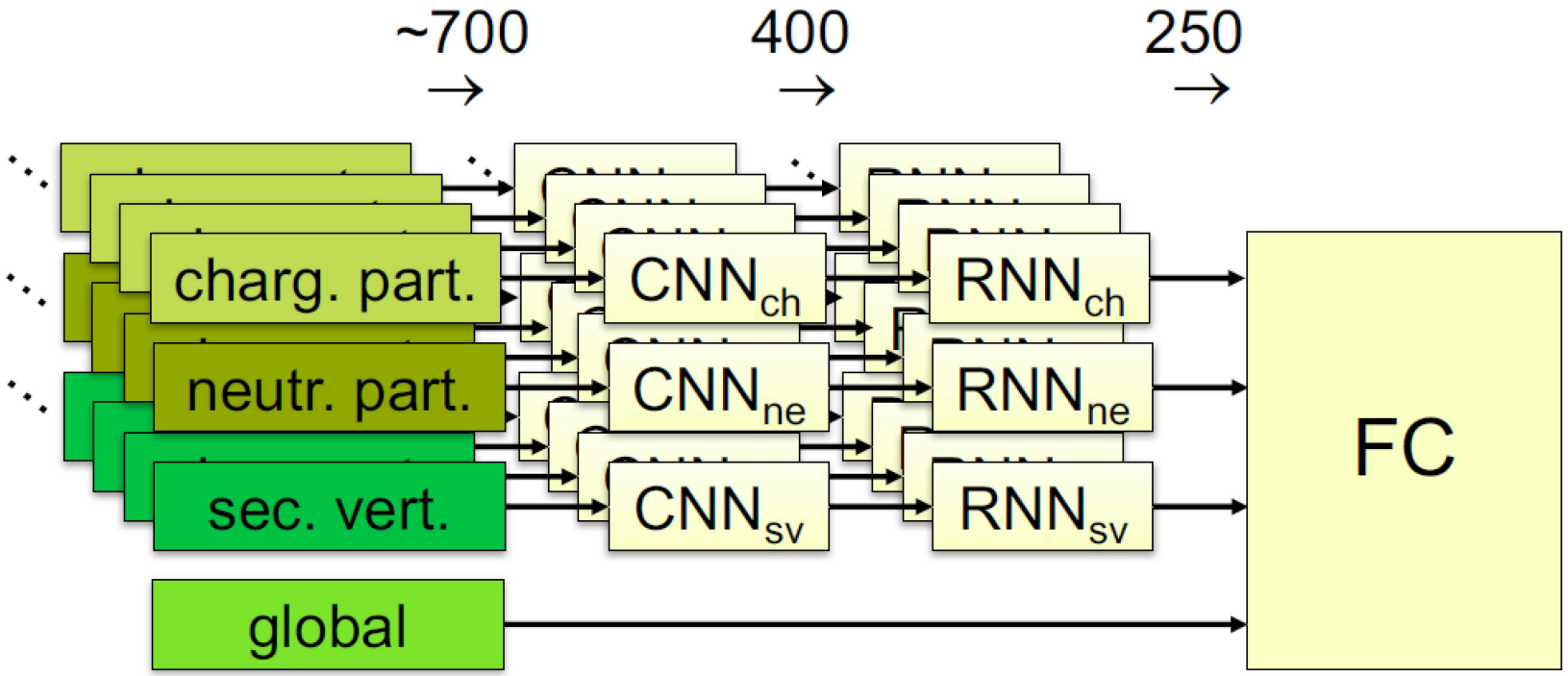
Imperial College London Data Science Institute

DeepJet: Jet classification with the CMS experiment

Markus Stoye
Imperial College London, DSI

"Big data science in astroparticle physics", HAP workshop, Aachen, Germany, 20th Feb. 2018

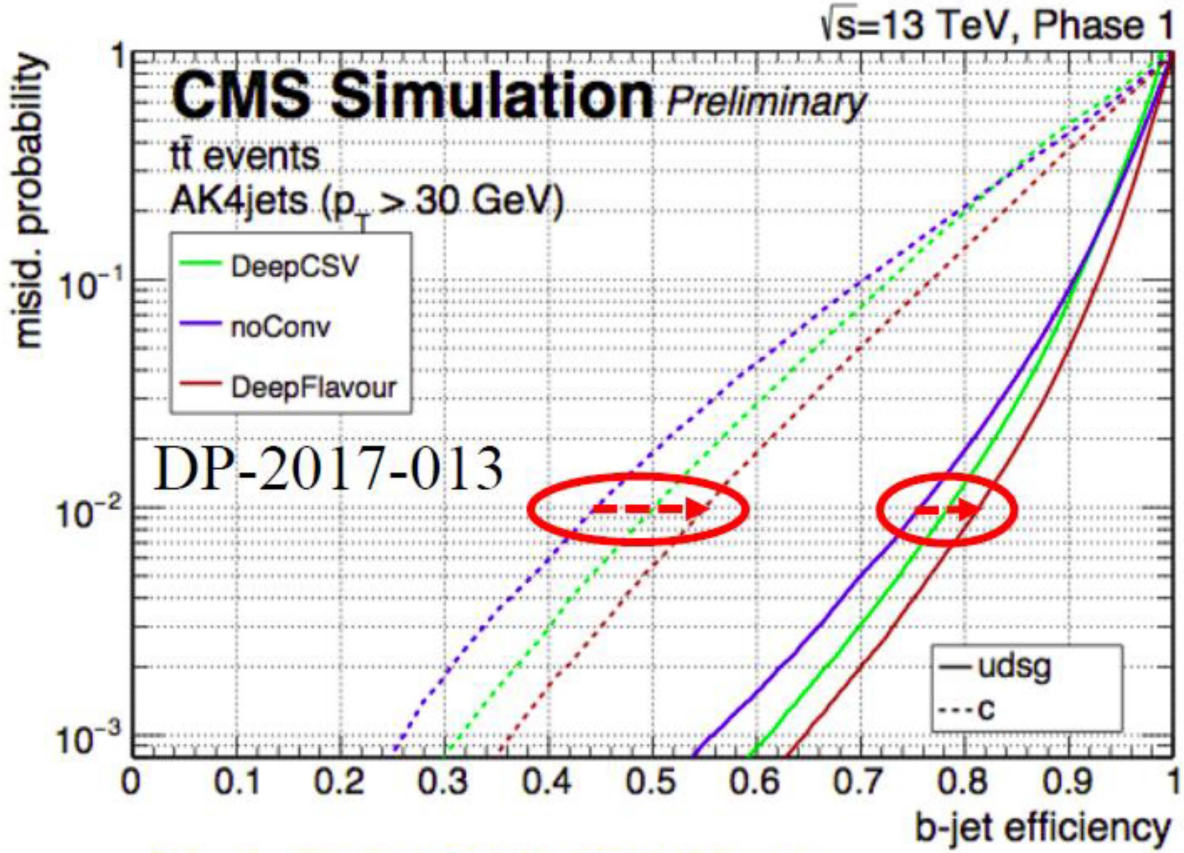
Particle and vertex based DNN: DeepJet



~ 700 inputs and 250.000 model parameters

- Particle and vertex based DNN has factor 10 less free parameters than a generic Dense DNN would have
- 100M jets used for training, overtraining is not an issue

Impact of DNN architecture



Blue: generic DNN (650 inputs)

Green: CMS tagger (~65 human made inputs)

Red: Physics inspired DNN (650 inputs)

Physics object based DNN performs best

kaggle Search kaggle Competitions Datasets Kernels Discussion Learn ••• Sign In

Featured Prediction Competition

TrackML Particle Tracking Challenge

High Energy Physics particle tracking in CERN detectors

$25,000
Prize Money

CERN · 656 teams · 22 days ago

Overview Data Kernels Discussion Leaderboard Rules

Overview

Description

Evaluation

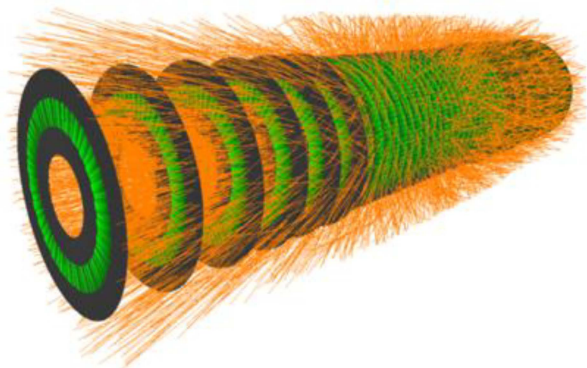
Prizes

About The Sponsors

Timeline

To explore what our universe is made of, scientists at CERN are colliding protons, essentially recreating mini big bangs, and meticulously observing these collisions with intricate silicon detectors.

While orchestrating the collisions and observations is already a massive scientific accomplishment, analyzing the enormous amounts of data produced from the experiments is becoming an overwhelming challenge.



Machine Learning in Astronomy

Machine learning and the Dark Energy Survey (DES)

- **Classification:**

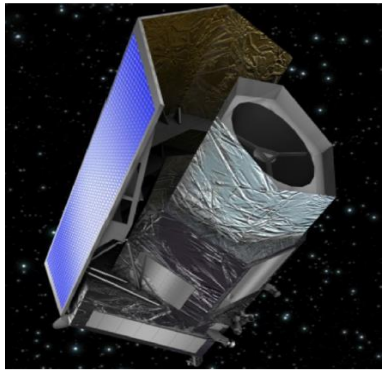
  galaxy type (0908.2033), star/galaxy (1306.5236),

  Supernovae Ia (1603.00882)

- **Photo-z** (1507.00490)

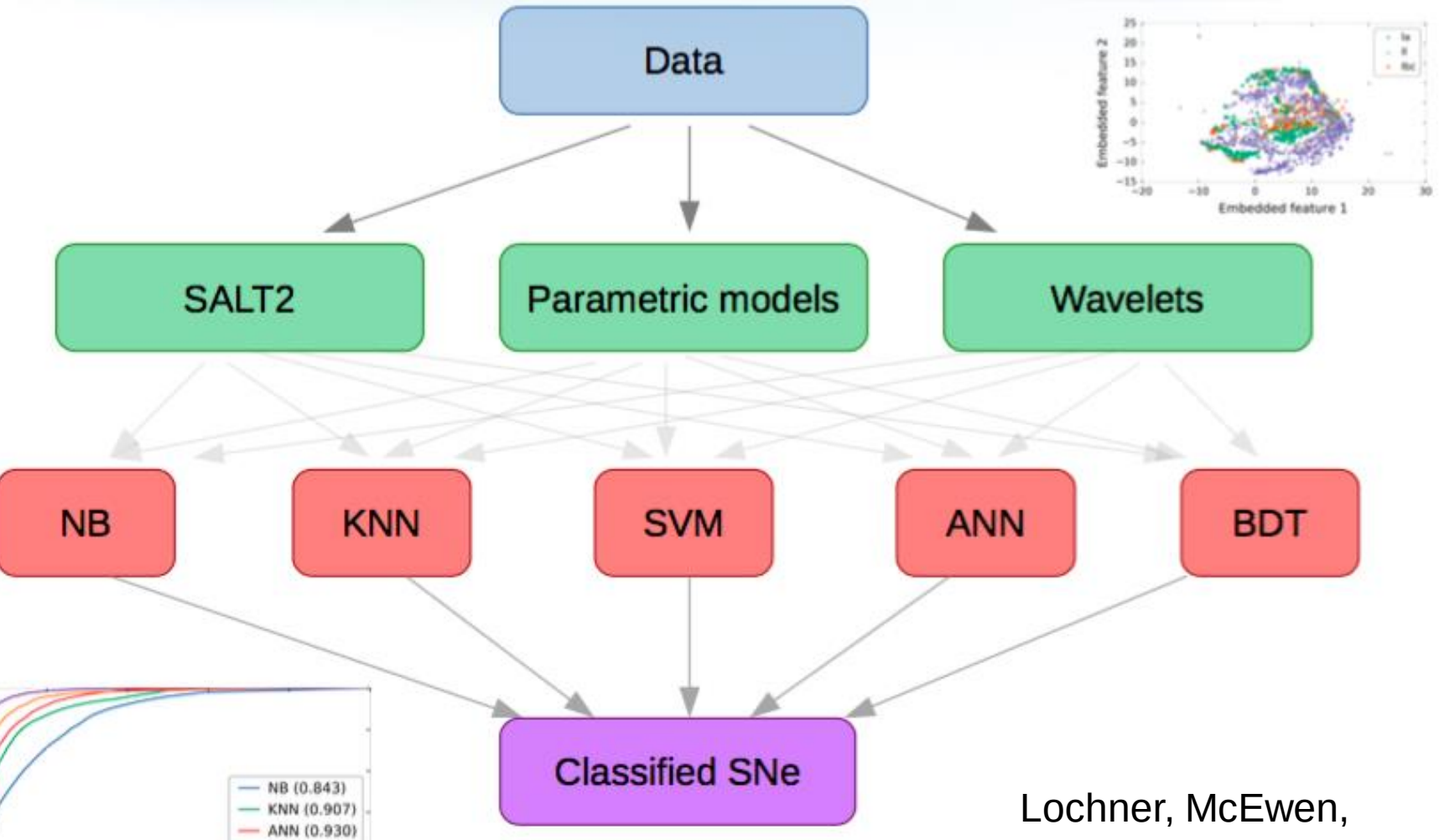
- **Mass of the Local Group** (1606.02694)

- **Search for Planet 9 in DES**



Slides thanks to Ofer Lahar

Photometric Classification of Supernovae



Lochner, McEwen,
Peiris, Lahav, Winter

arXiv: 1603.00882

Featured Prediction Competition

PLAsTiCC Astronomical Classification

Can you help make sense of the Universe?

$25,000
Prize Money

LSST Project · 34 teams · 3 months to go (2 months to go until merger deadline)

Overview

Help some of the world's leading astronomers grasp the deepest properties of the universe.

The human eye has been the arbiter for the classification of astronomical sources in the night sky for hundreds of years. But a new facility -- the Large Synoptic Survey Telescope (LSST) -- is about to revolutionize the field, discovering 10 to 100 times more astronomical sources that vary in the night sky than we've ever known. Some of these sources will be completely unprecedented!

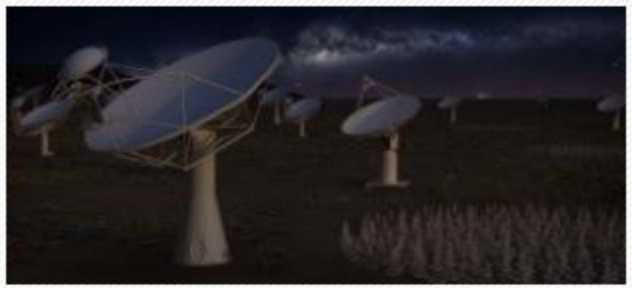
The Photometric LSST Astronomical Time-Series Classification Challenge (PLAsTiCC) asks Kagglers to help prepare to classify the data from this new survey. Competitors will classify astronomical sources that vary with time into different classes, scaling from a small training set to a very large test set of the type the LSST will discover.

Acknowledgements

PLAsTiCC is funded through LSST Corporation Grant Award # 2017-03 and administered by the University of Toronto. Financial support for LSST comes from the National Science Foundation (NSF) through Cooperative Agreement No. 1258333, the Department of Energy (DOE) Office of Science under Contract No. DE-AC02-76SF00515, and private funding raised by the LSST Corporation. The NSF-funded LSST Project Office for construction was established as an operating center under management of the Association of Universities for Research in Astronomy (AURA). The DOE-funded effort to build the LSST camera is managed by the SLAC National Accelerator Laboratory (SLAC).

The National Science Foundation (NSF) is an independent federal agency created by Congress in 1950 to promote the progress of science. NSF supports basic research and people to create knowledge that transforms the future.

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

Data Flow through the SKA

SKA1-MID

8.8 Tb/s

~50 PFLOPS

~5 Tb/s

100 PFLOPS

130 - 300 PB/yr

Users

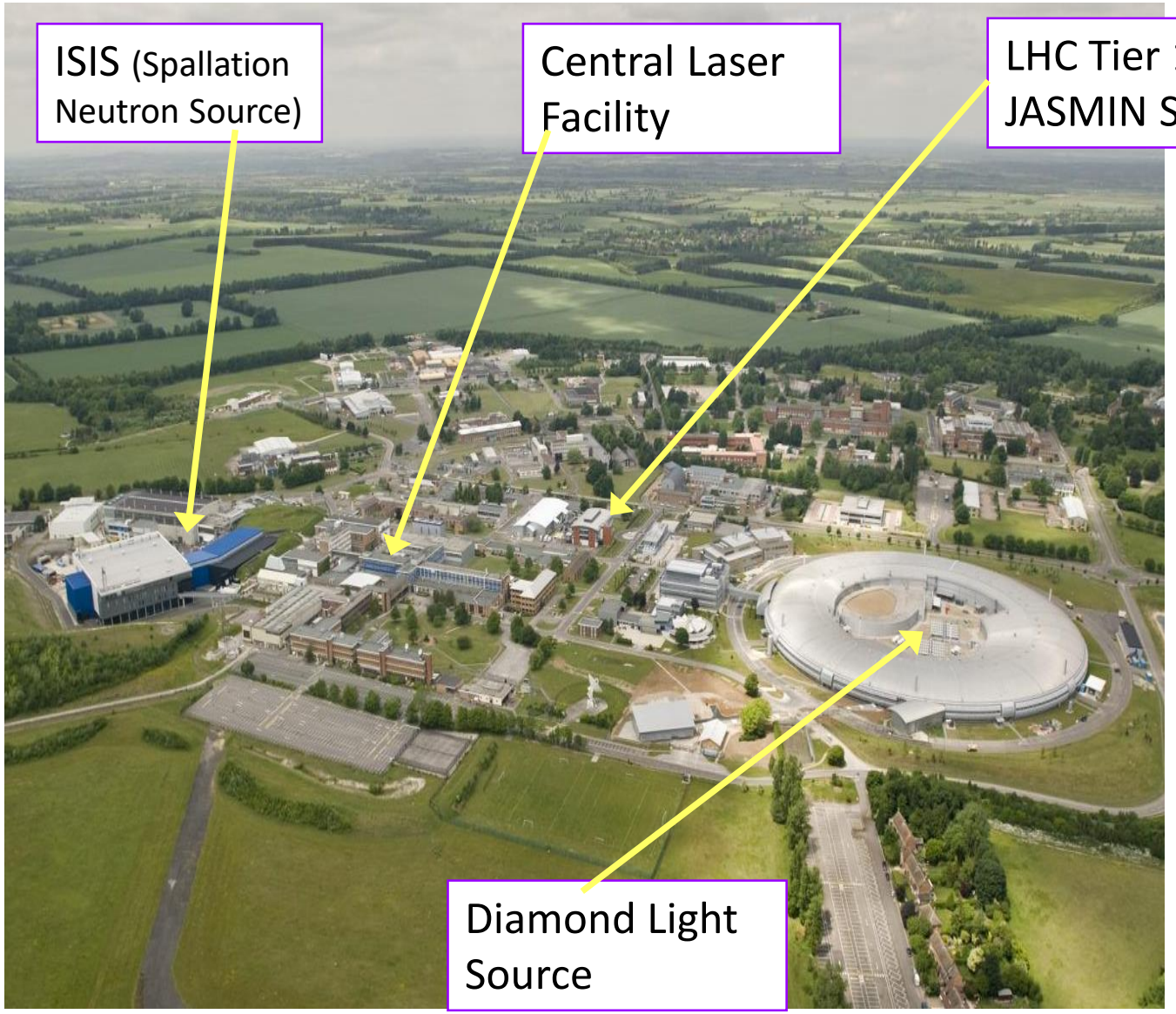
7.2 Tb/s

SKA1-LOW

~2 Pb/s

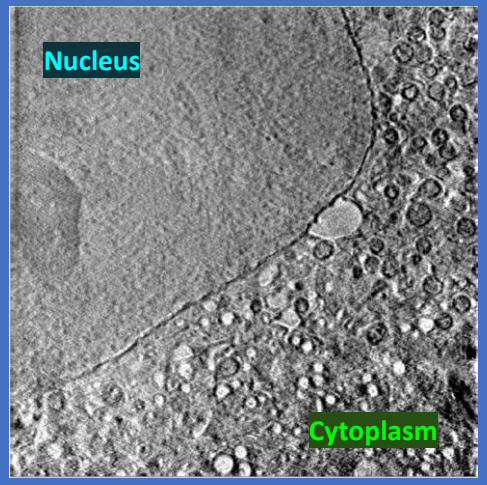
Big Scientific Data from Large Experimental Facilities in the UK

Rutherford Appleton Laboratory



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

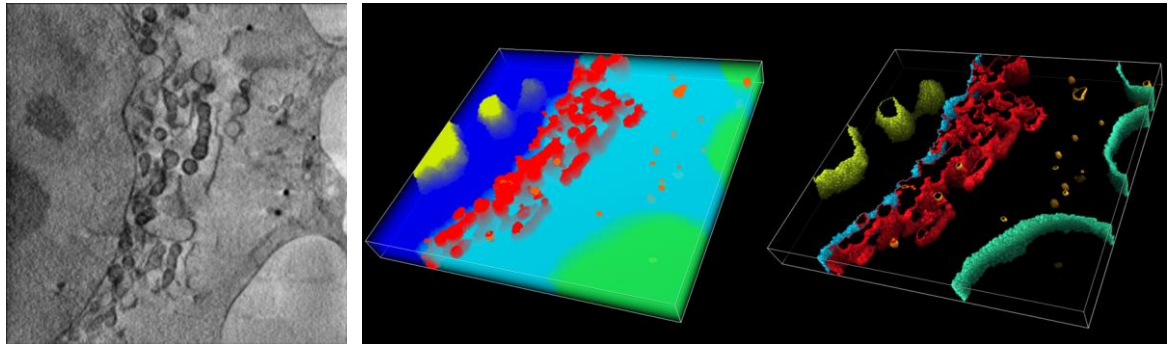
scientificsoftware@diamond.ac.uk

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

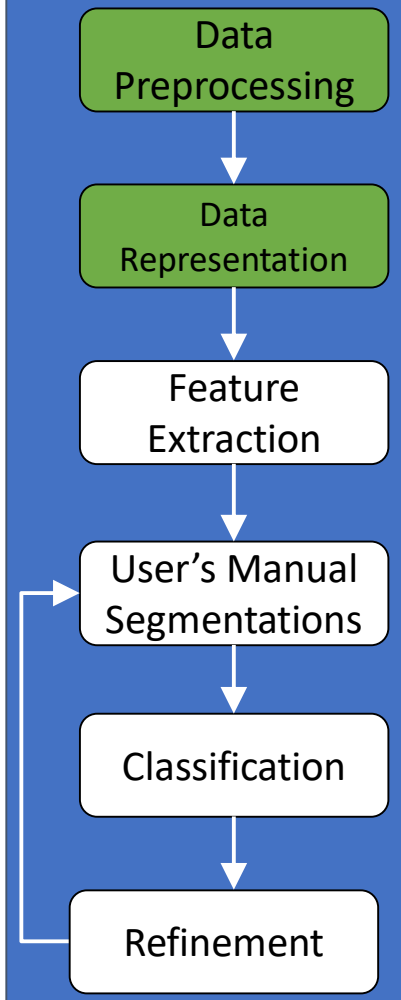


Tomographic Cell Analysis: Feature Extraction

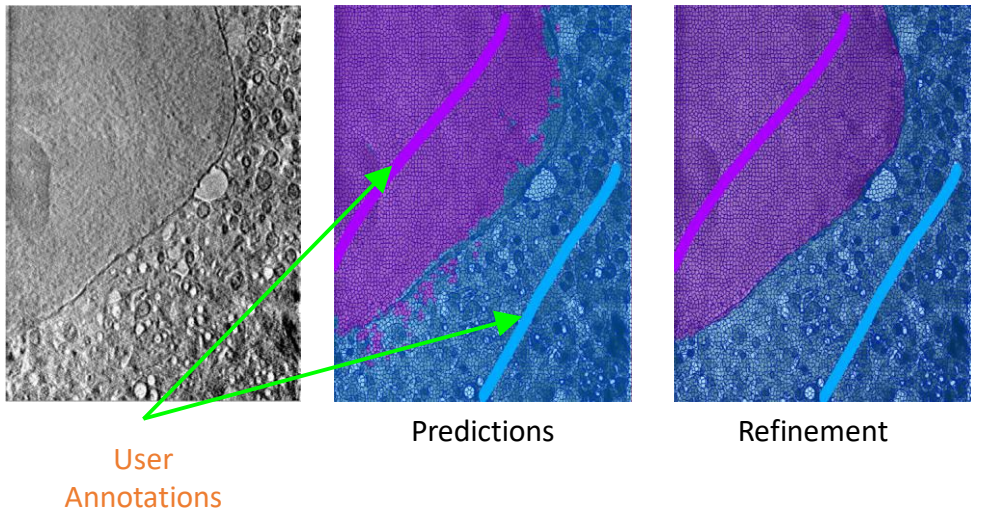
Features are extracted from voxels to represent their appearance:

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

Workflow



User Annotation + Machine Learning



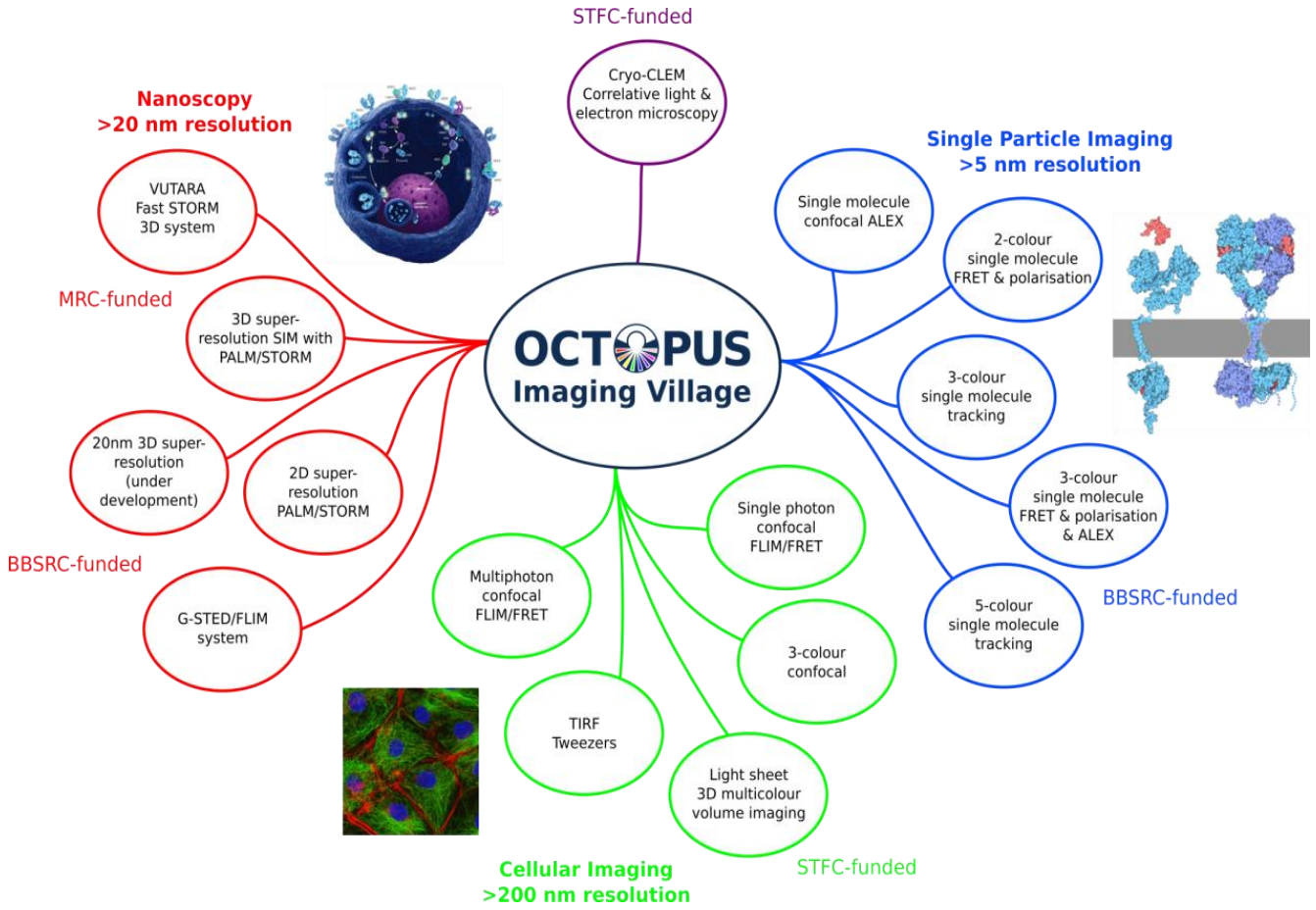
Using few user annotations as an input:

- Machine learning classifier (Random Forest) trained to discriminate between Nucleus and Cytoplasm and predict the class of each SuperVoxel
- Markov Random Field then used to refine the predictions

scientificsoftware@diamond.ac.uk

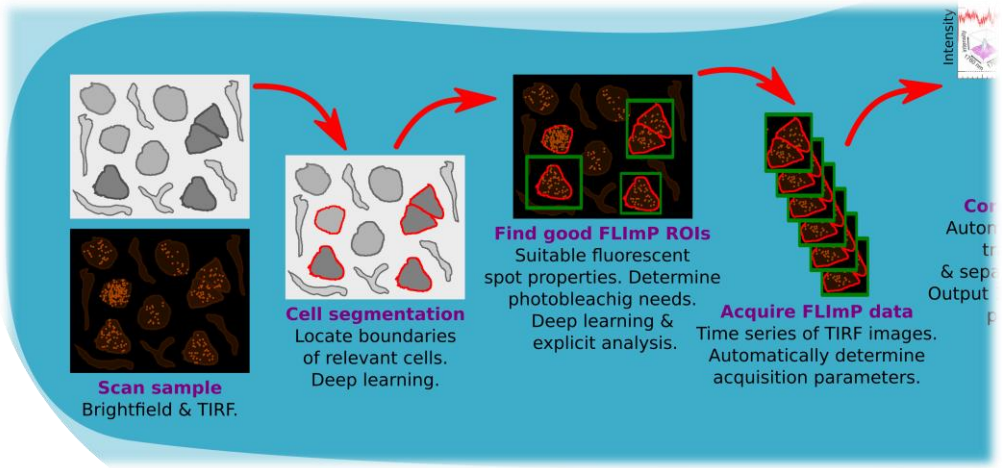
The OCTOPUS Imaging Village

- Part of UKRI STFC Central Laser Facility at Harwell Campus
- Cluster of microscopes and lasers and expert end-to-end multidisciplinary support
- National imaging facility with peer-reviewed, funded access



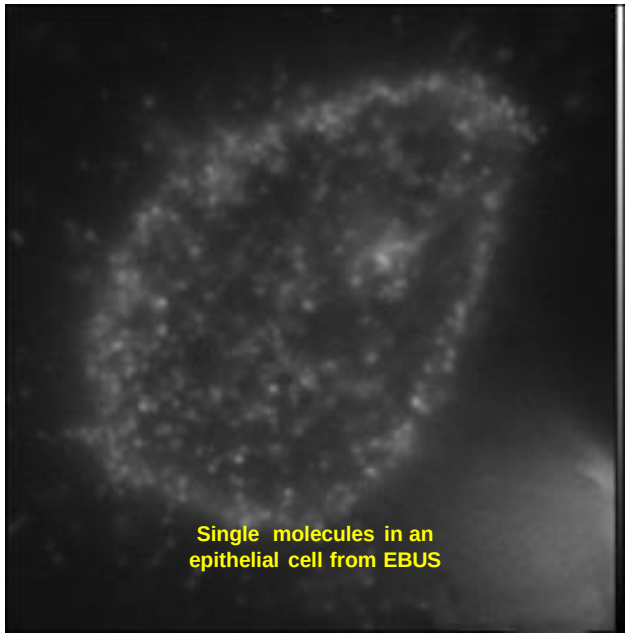
- ➢ Look at example of Single Molecule Super-resolution Microscopy
- ➢ Use technique of Fluorescence Localisation Imaging with Photobleaching (FLImP)
- ➢ 400 nm → 10 nm

Automated acquisition with Machine Learning



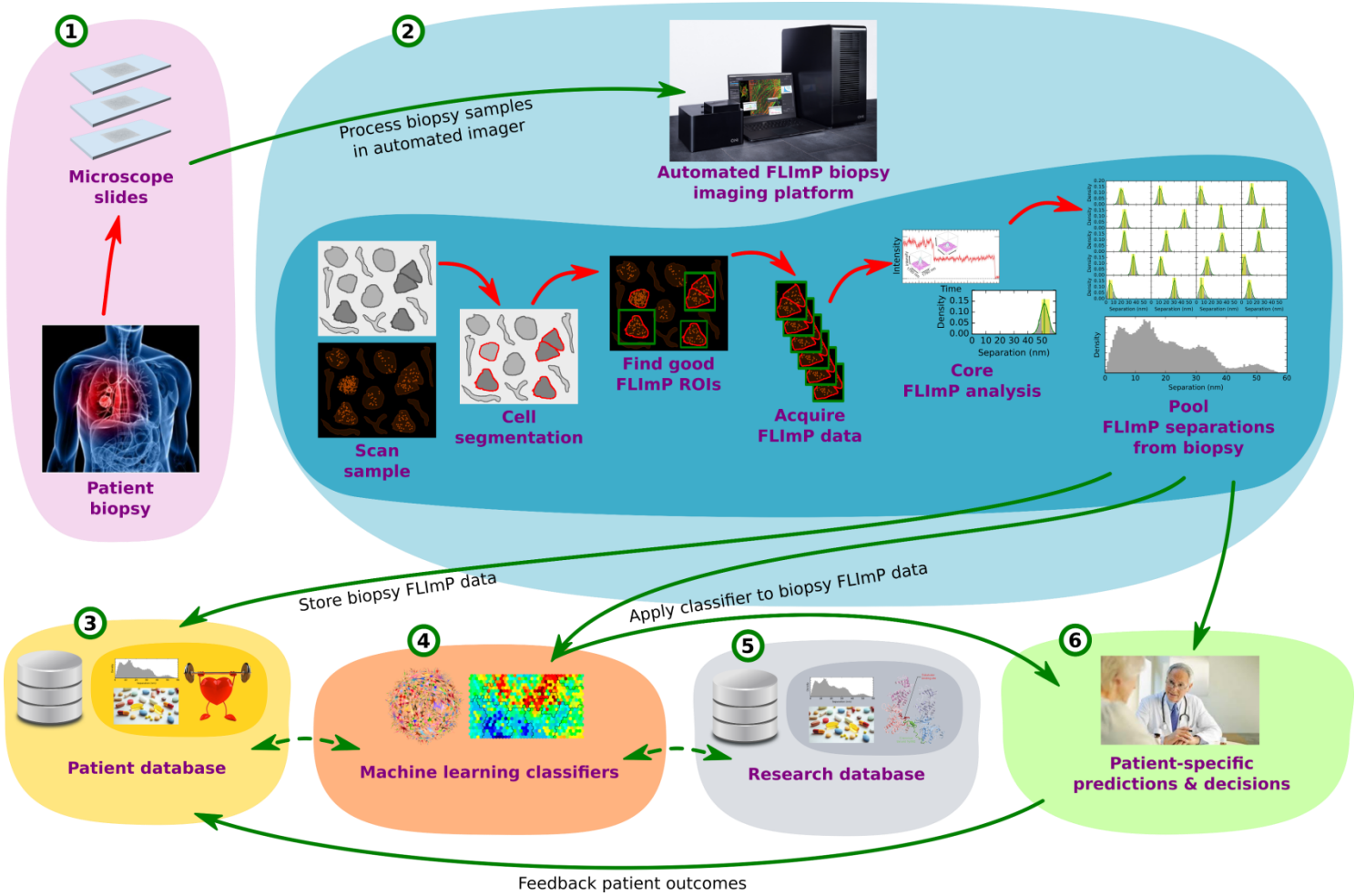
Key task - identify ROIs to acquire FLImP data

- Segment cell type of interest from in NSCLC biopsy
  - Fibroblasts, epithelial, endothelial, red blood
- Segment ROI good for single molecule FLImP
  - Well focussed, separated spots with smooth background
- Deep learning (CNNs) proven in cell segmentation
  - We wish to apply this to our system to identify the ROIs



ROI = Region Of Interest

FLImP Vision for Personalized Cancer Care

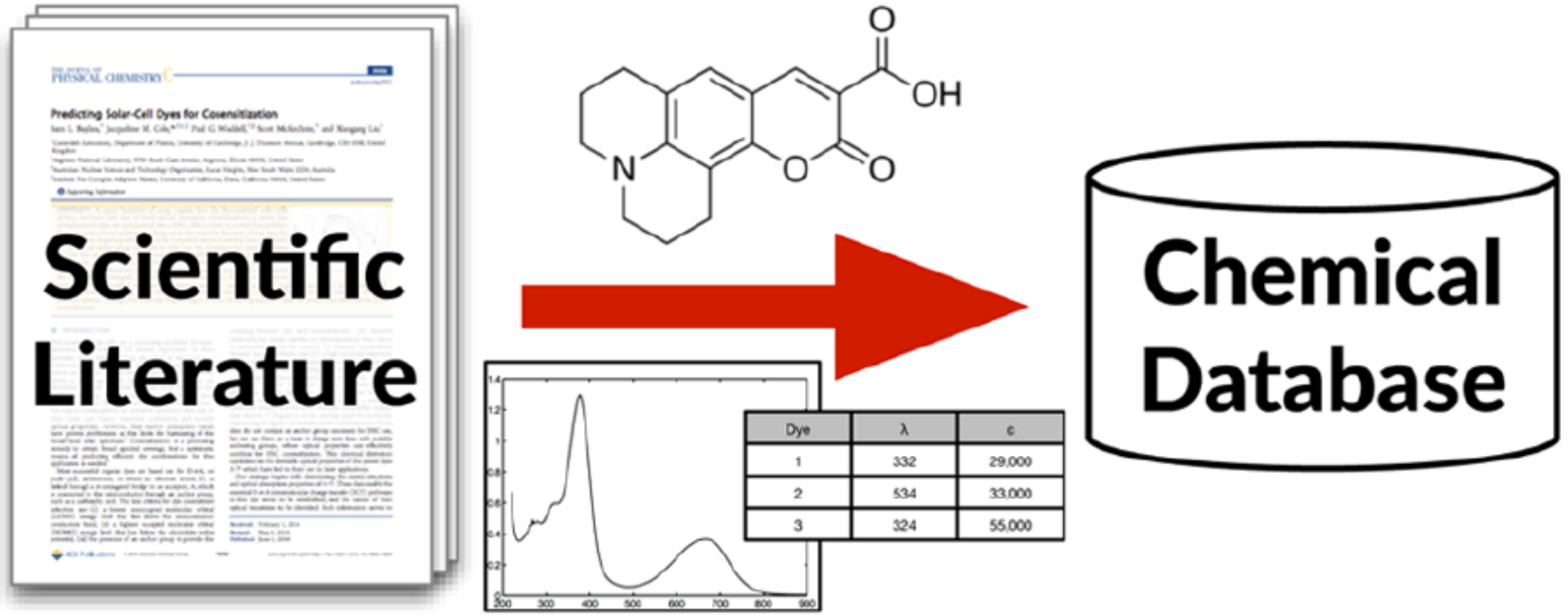


Fully automate from acquisition to output

- Increase efficiency and impact in lab research
- Automated structure fingerprint for personalised cancer care in clinic

Another type of Big Scientific Data and Machine Learning

ChemDataExtractor



M. C. Swain, J. M. Cole *J. Chem. Inf. Model*. 56 (2016) 1894-1904

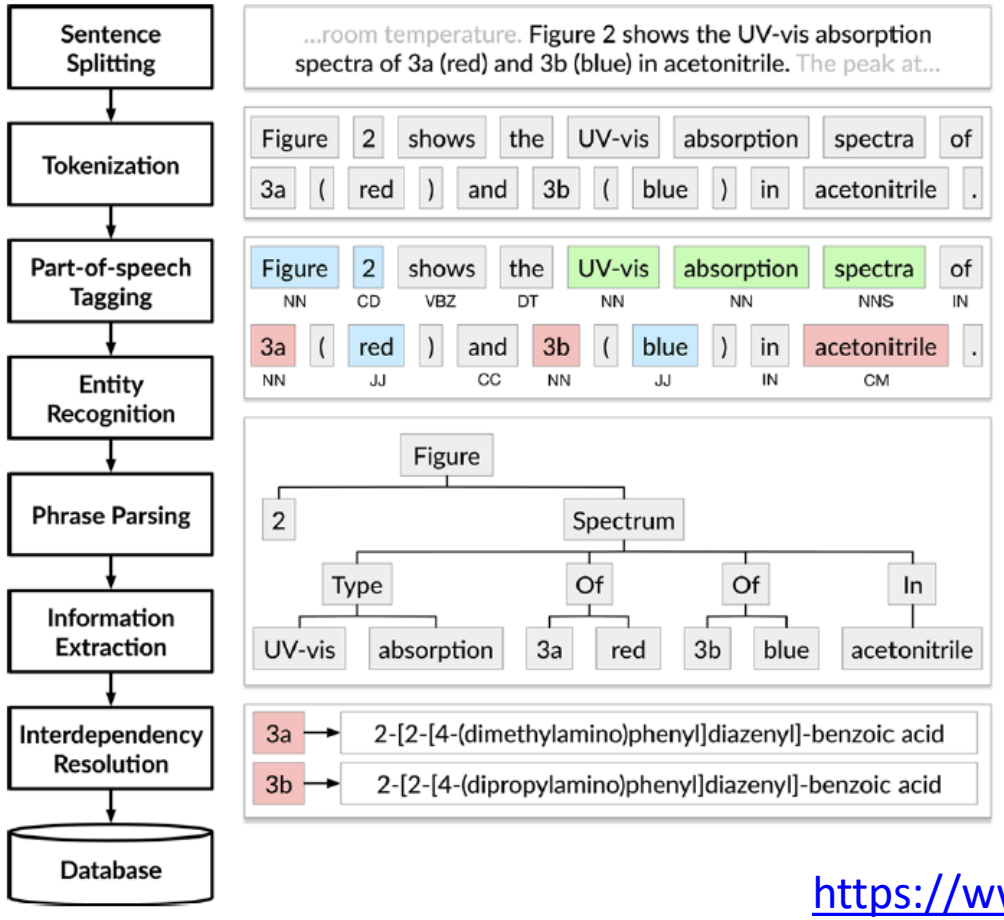
www.chemdataextractor.org

Slide thanks to Jacqui Cole

‘Chemistry-Aware’ aspect of ChemDataExtractor

Input

ChemDataExtractor Process



Output

Chemical	Property # 1	Property # 2
Name	(e.g. wavelength)	(e.g. intensity)
<Chemical name 1>	453 nm	10,030
<Chemical name 2>	357 nm	12,404
<Chemical name 3>	846 nm	10,204
<Chemical name 4>	743 nm	30,049
….	…	….

Slide thanks to Jacqui Cole

https://www.nature.com/articles/sdata2018111

- **Natural language processing w/ interdependency resolution to resolve chemical names & match them to material properties**
- **Machine learning is incorporated to improve text-mining extraction performance as the tool is progressively used**

Data-Driven Materials Discovery Workflow

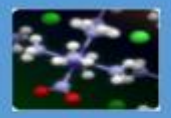
Academic Literature

Data Source

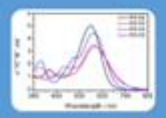
ChemDataExtractor

Stage 1

Data Extraction



Material structure



Physical properties

Databases

Stage 2

Data Enrichment

- Machine learning
- Quantum-chemical calculations (φ)

Enriched Data

Feedback

- Materials Prediction
- Experimental Validation

Photovoltaics

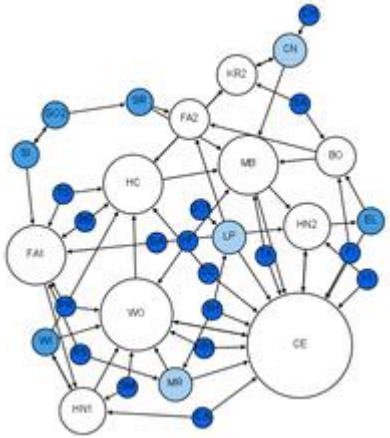
Magnetism

Catalysis

Applications

Stage 4

Prediction & Validation



Stage 3

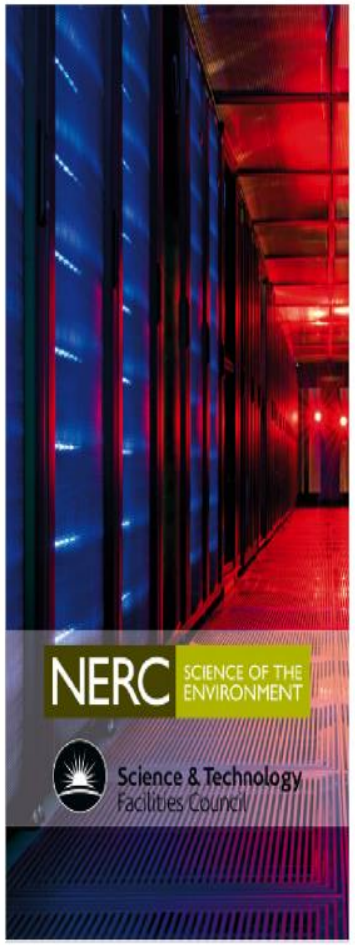
Create algorithms for materials prediction

Slide thanks to Jacqui Cole

Environmental Science and the JASMIN ‘Super Data Cluster’

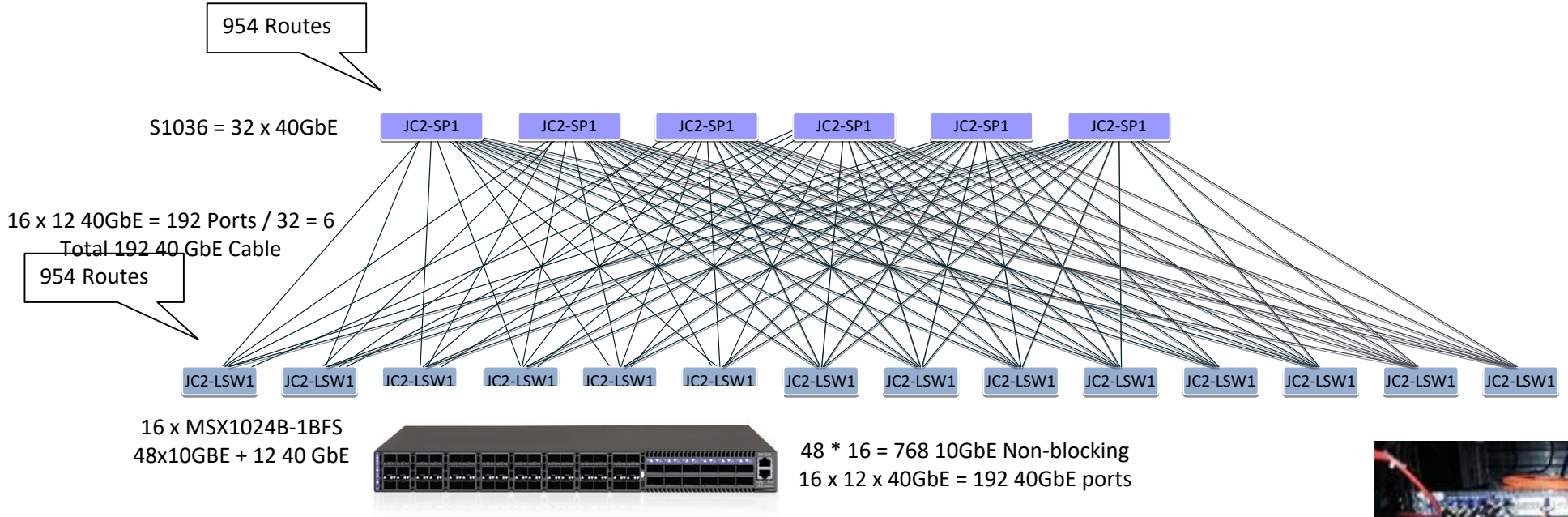
Centre for Environmental Data Analytics

JASMIN Super-Data Cluster infrastructure



- 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.

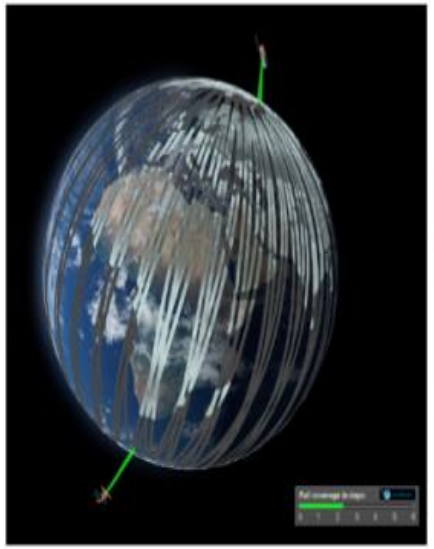
Non-blocking, low latency, CLOS Tree Network

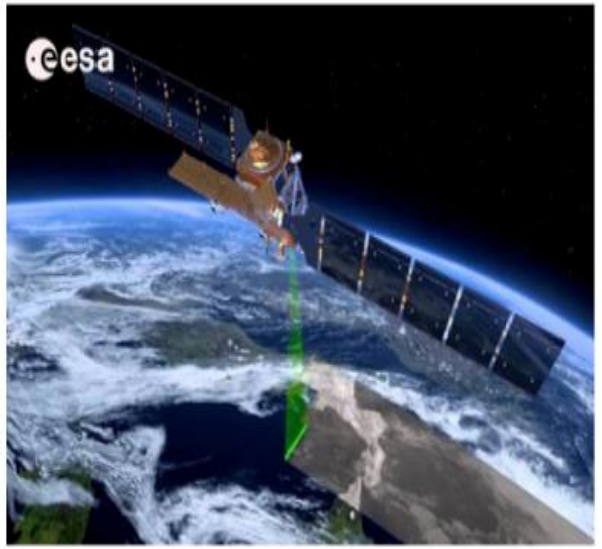


1,104 x 10GbE Ports CLOS L3 ECMP OSPF

- ~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

Large data sets: satellite observations



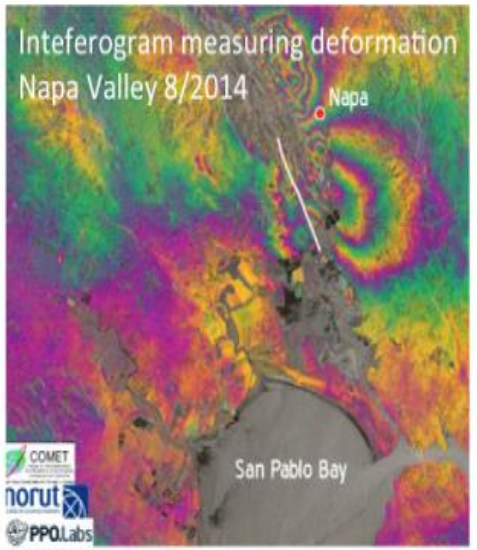


Sentinel 1A: Launched 2014
(1B due 2016)

- 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

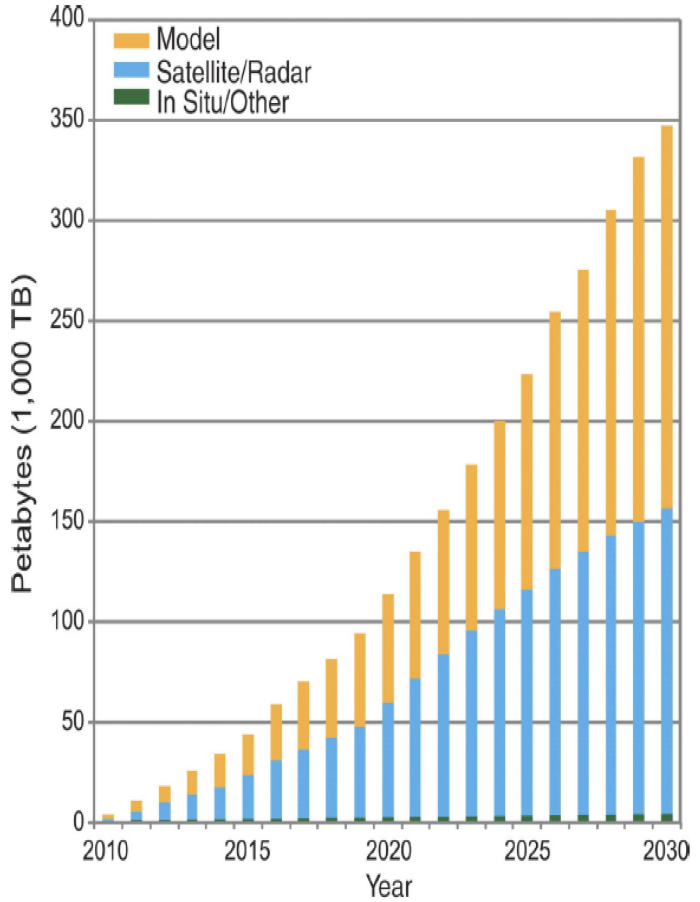


(Picture credits: ESA, Arianespace.com, PPO.labs-Norut–COMET-SEOM Insarap study, ewf.nerc.ac.uk/2014/09/02/new-satellite-maps-out-napa-valley-earthquake/)

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?)

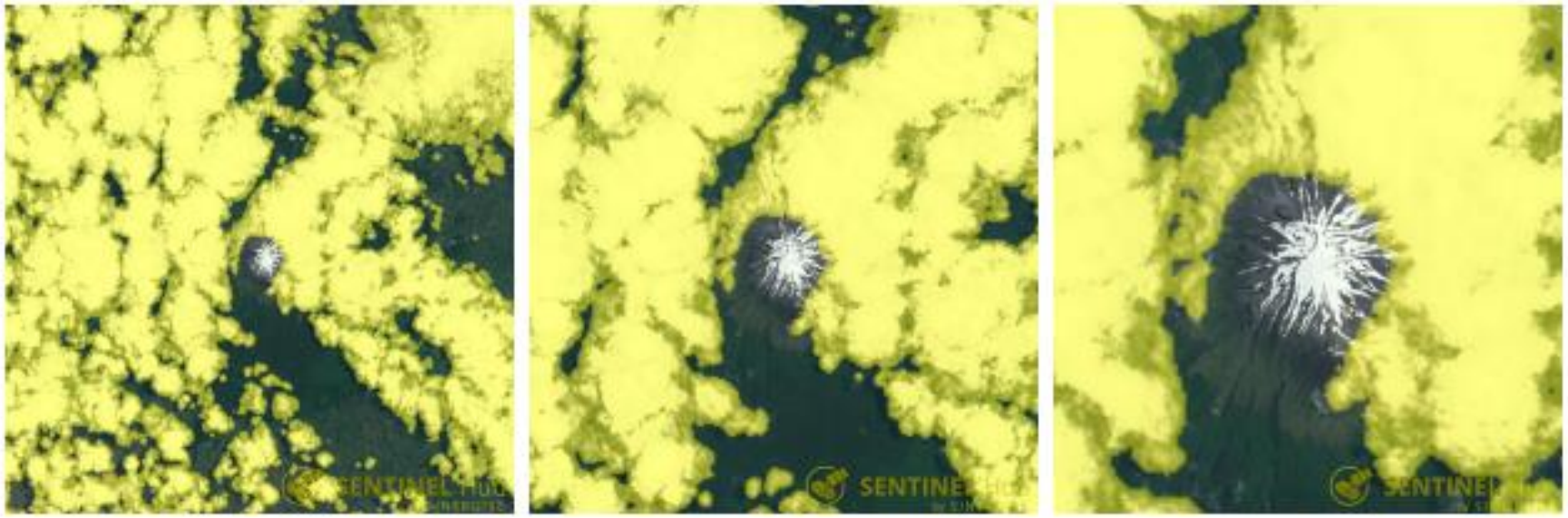


J T Overpeck et al. Science 2011;331:700-702

Example: The Sentinel Hub Cloud Detector

Improving Cloud Detection with Machine Learning

This is a story about clouds. Real clouds in the sky. Sometimes you love them and sometimes you wish they weren't there. If you work in remote sensing with satellite images you probably hate them. You want them gone. Always and completely. Unfortunately, the existing cloud masking algorithms that work with Sentinel-2 images are either too complicated, expensive to run, or simply don't produce satisfactory results. They misidentify clouds too often and/or like to identify clear sky over land as clouds.



Mount Taranaki cuts through the clouds. Sentinel-2 image from 2017–12–15 via Sentinel Hub with a cloud mask produced with the Sentinel Hub Cloud Detector at three different zoom levels.

https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13

Sentinel Hub Blog: Results and Concluding remarks

- Machine Learning approach can give state-of-the-art results if the training and validation sets are both of good enough quality and representative of the unseen data.
- Procurement of labeled samples in remote sensing that are suitable for development of models that perform well on a global scale is particularly challenging.
- With the new labeled data sets curated with the help of community, the performance of our cloud detector can only improve.

Fraction of classifications as clouds

True Label	Fmask	Sen2Cor	Sentinel Hub
Cloud	89.0%	97.5%	99.4%
Cirrus	88.3%	87.7%	83.8%
Land	7.2%	5.7%	2.2%
Water	2.0%	0.0%	0.1%
Snow	39.2%	30.7%	13.5%
Shadow	3.9%	3.9%	5.8%

Cloud and cirrus cloud detection rates and land, water, snow and shadow misclassification rates as clouds as determined using 108 Sentinel-2 scenes hand labeled by Hollstein et al.

An Engineering Example: Electric Vehicle Batteries

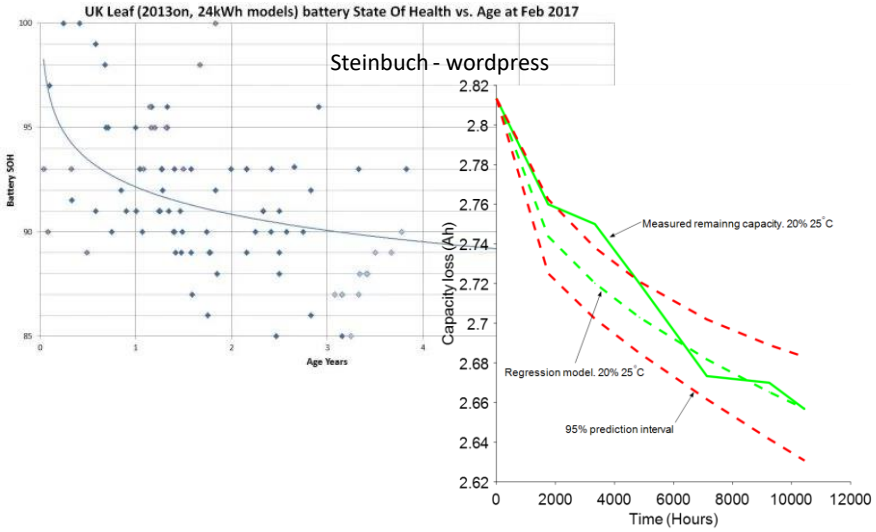
Machine Learning and EV Battery Lifetime Prediction at Warwick Manufacturing Group

Wide range of usage duty cycles

- Automotive (main sector)
- Off-highway
- Domestic renewables

Models used to predict lifetime validated at much shorter time-scales

- Lifetime prediction ~ 8-15 years
- Data collection ~ 6-14 months



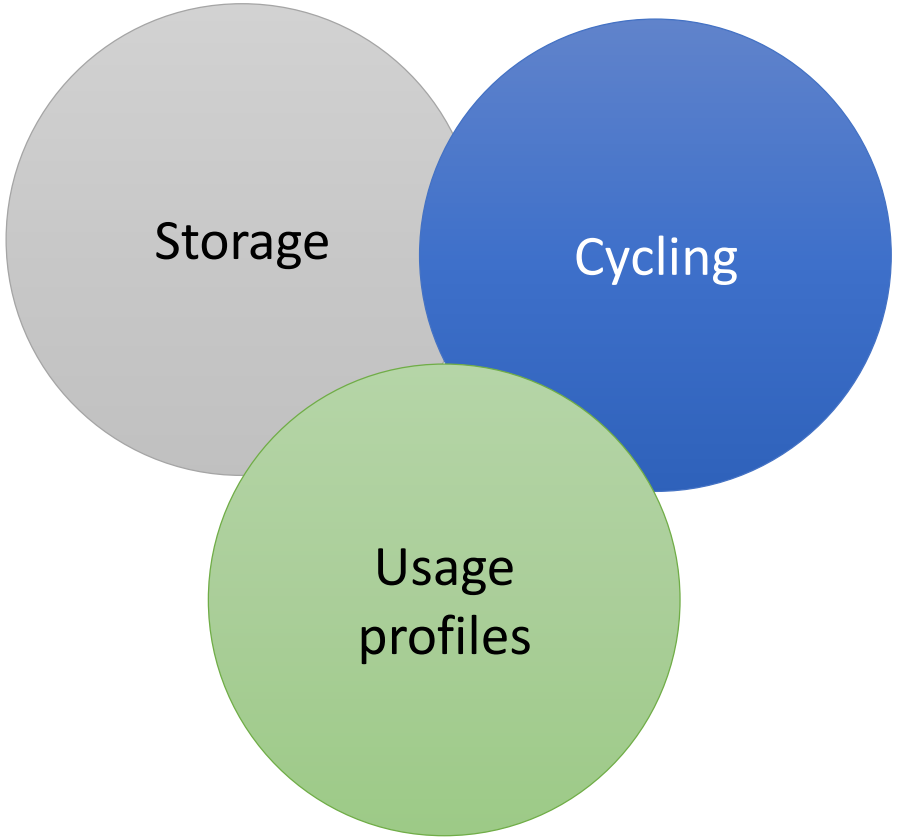
Proposed modelling approach

- Understand the interactions via Machine Learning methods
- Quantify prediction uncertainties
- Couple with underlying physical processes

Slide thanks to Dhammika Widanalage and Theo Damoulas

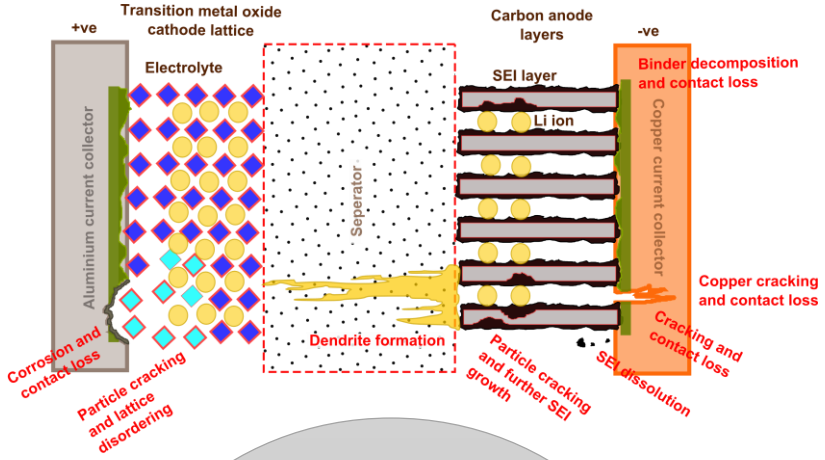
Battery Ageing Data

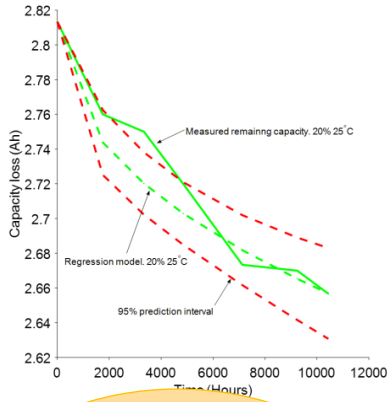
- Data sets to include a diverse set of usage profiles.
- Data sets to include
  - Cell ageing for future EV battery packs evaluation
  - Telemetry data of cell ageing from vehicle usage
- Combination of data, modelling and machine learning will link fundamental research to production and consumption decisions for EVs

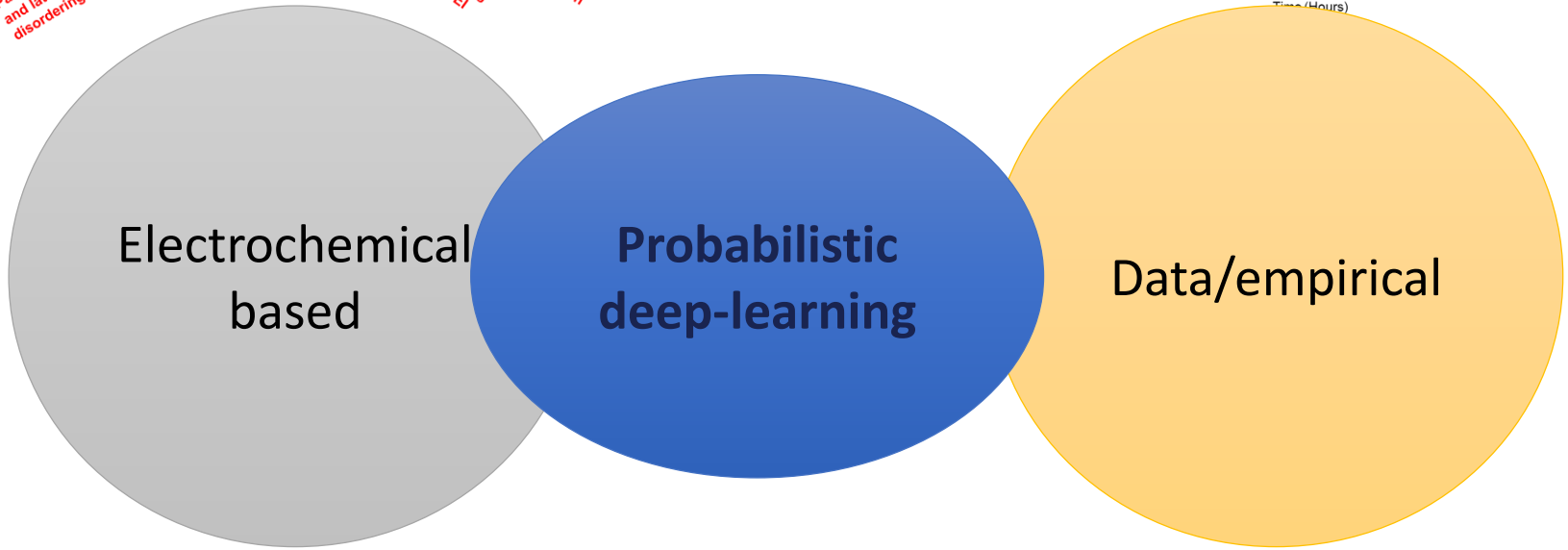


Slide thanks to Dhammika Widanalage and Theo Damoulas

Approach: Multi-scale Modelling, Data and Machine Learning







Slide thanks to Dhammika Widanalage and Theo Damoulas

Some concerns …

How can Academia compete with Industry on Machine Learning and AI?

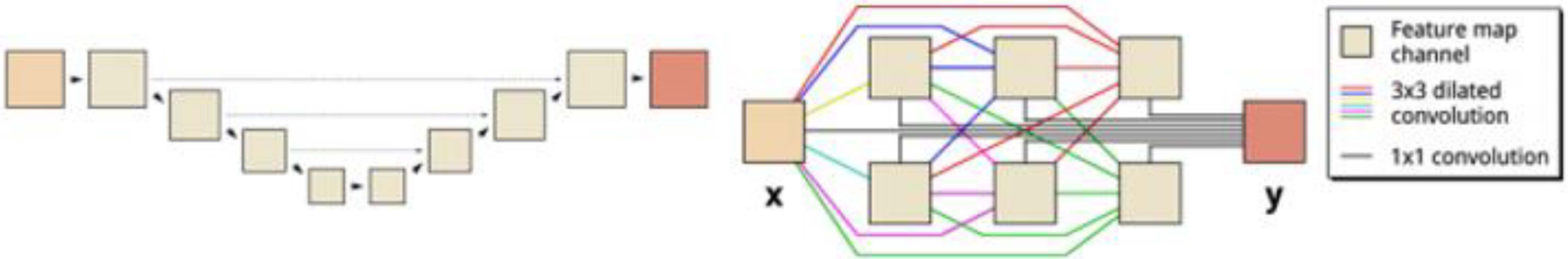
Companies like Facebook, Google, Amazon, Microsoft (and probably Baidu, Alibaba and Tencent) and have three key advantages over academia:

1. These companies all have many, very large, private datasets that they will never make publicly available
2. Each of these companies employs many hundreds of computer scientists with PhDs in Machine Learning and AI
3. Their researchers and developers have essentially unlimited computing power at their disposal

➢ NLP, Machine Translation, Image Recognition, …

Berkeley Lab 'Minimalist Machine Learning' Algorithms Analyze Images From Very Little Data

CAMERA researchers develop highly efficient convolution neural networks tailored for analyzing experimental scientific images from limited training data



Common DCNN Architecture

Mixed-Scale Dense Architecture

Left: A schematic representation of a common DCNN architecture with scaling operations; downward arrows represent downscaling operations, upward arrows represent upscaling operations, and dashed arrows represent skipped connections. Right: Schematic representation of an MS-D network with w=2 and d=3; colored lines represent 3x3 dilated convolutions, with each color corresponding to a different dilation: All feature maps are used for the final output computation.

Adversarial Noise and Deep Learning?

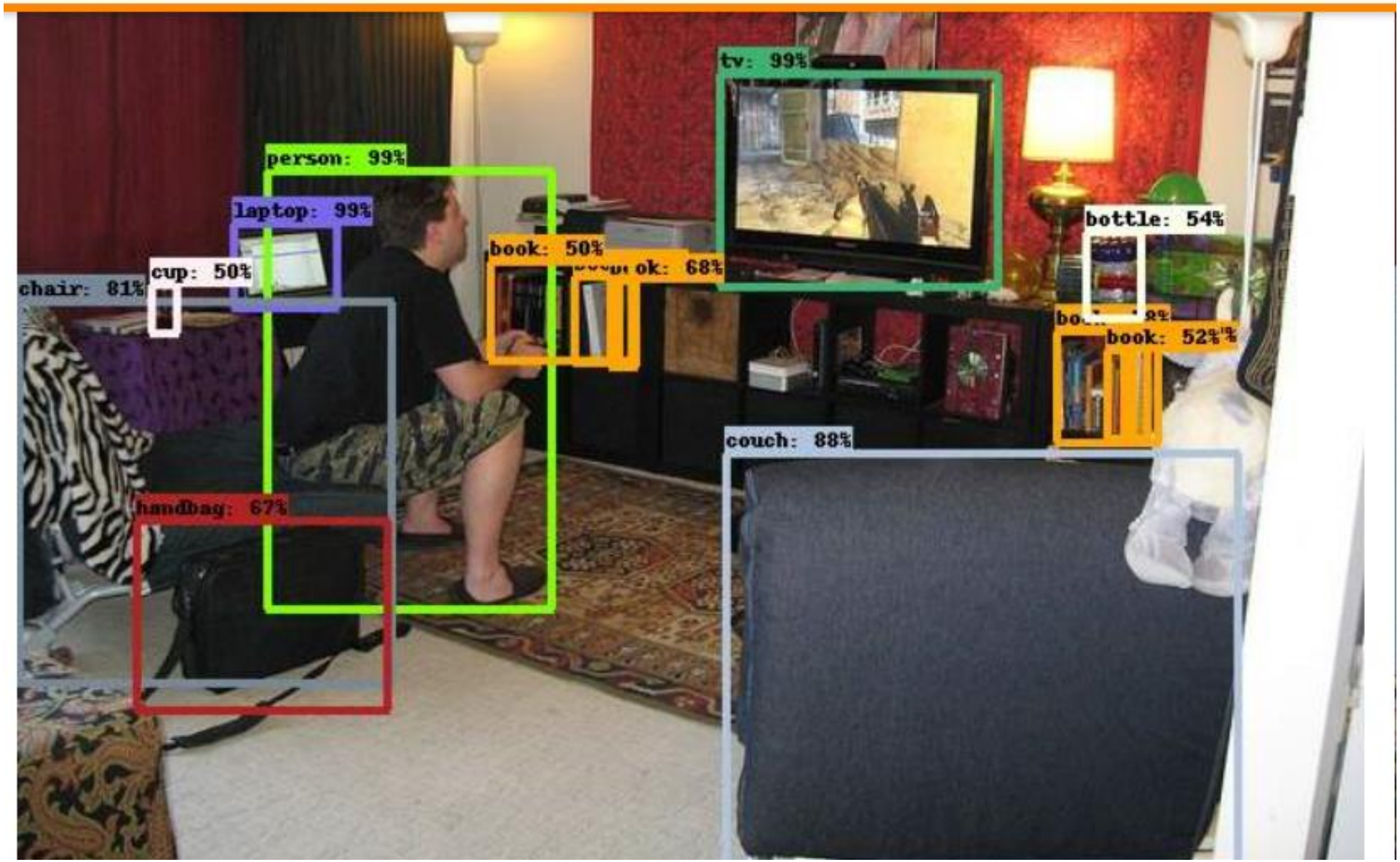
+

=

“panda”
57.7% confidence

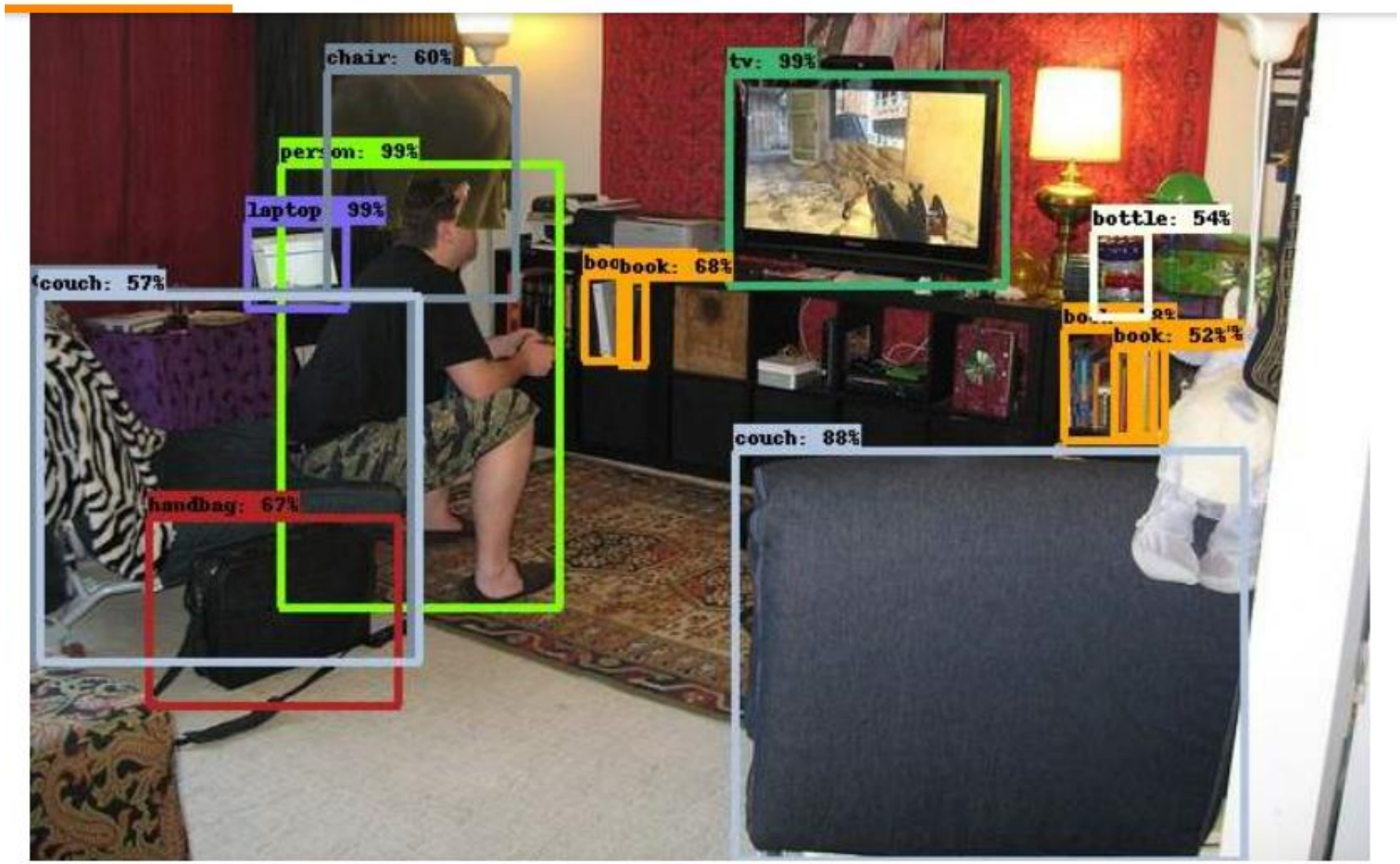
“gibbon”
99.3 % confidence

On the left is the original image; in the middle, the perturbation; and on the right, the final, perturbed image. | Image by Ian Goodfellow, Jonathon Shlens, and Christian Szegedy



The Elephant in the Room Amir Rosenfeld, Richard Zemel, and John K. Tsotsos

arXiv:1808.03305v1 [cs.CV] 9 Aug 2018

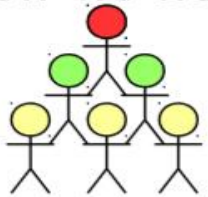


The Elephant in the Room Amir Rosenfeld, Richard Zemel, and John K. Tsotsos

arXiv:1808.03305v1 [cs.CV] 9 Aug 2018

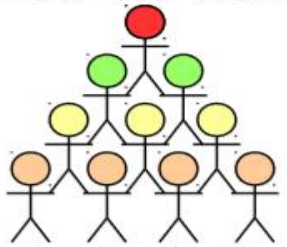
Career paths for Research Software Engineers and Data Scientists?

How we worked



PI stands on the shoulders of her postdocs and students (and as Newton would have said, the giants.)

How we work



PI stands on the shoulders of her postdocs, students, software engineers and data scientists. (Are the giants down with the turtles?).

- It's fair to say that our institutions have not really caught onto the necessity to have careers for everyone in that stack.
- From the people managing vocabularies and manually entering metadata, to the software engineers and data scientists, we have new careers appearing, and we're not really ready for it.
- Mercifully we're not alone, bioinformatics is blazing a similar trail, but we have much to do.

Slide thanks to Bryan Lawrence

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov