Machine Learning and Big Scientific Data Benchmarks: The Scientific ML Group and the Turing Hub at Harwell

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

The AI and Machine 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

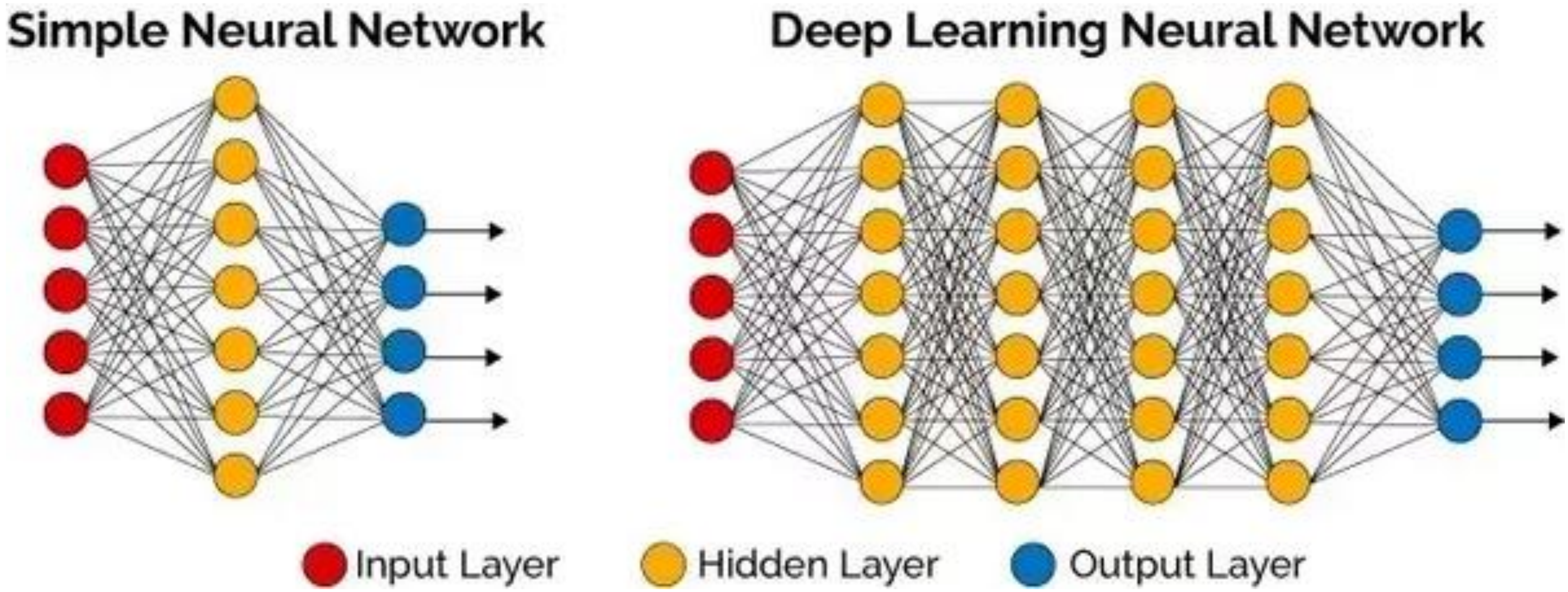
Boltzmann machines

Decision trees

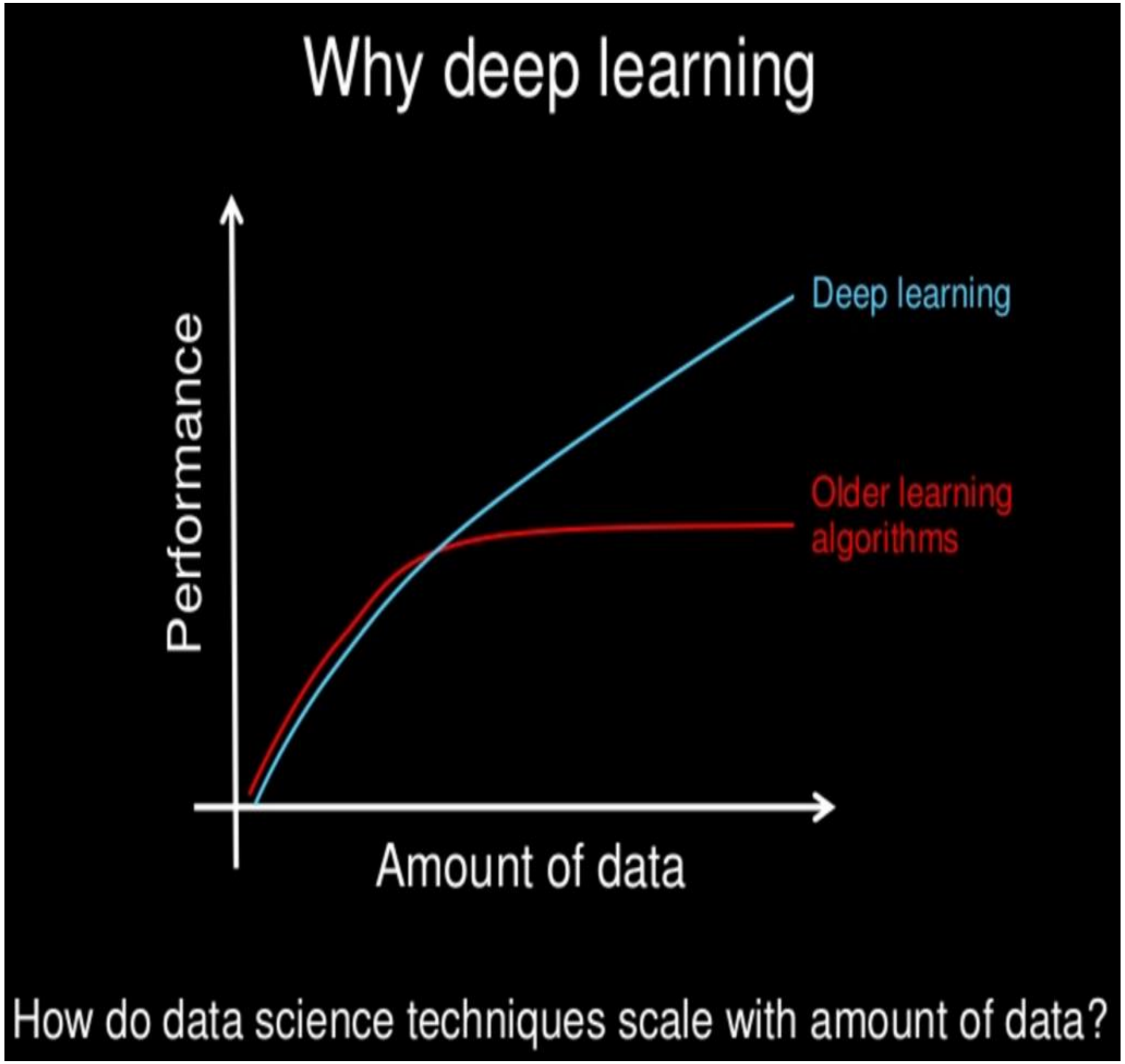
Radial basis functions

Hidden Markov Models

The Deep Learning Revolution



Why Deep Learning?



Slide by Andrew Ng

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.

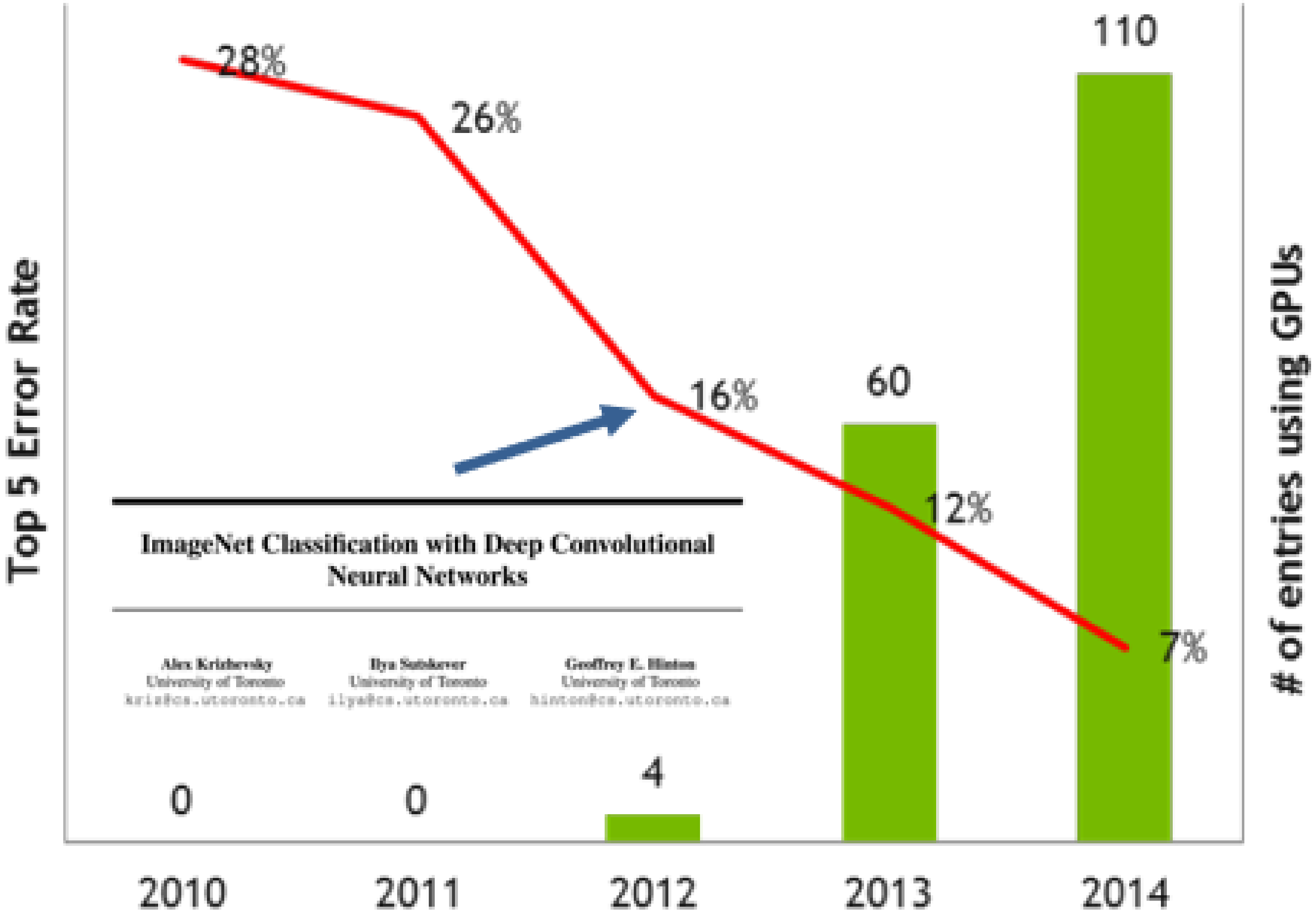
What do these images have in common? *Find out!*

Check out the ImageNet Challenge 2017

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

Krishevsky et al., http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

IMAGENET



https://devblogs.nvidia.com/parallelforall/nvidia-ibm-cloud-support-imagenet-large-scale-visual-recognition-challenge/

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

Research Prediction Competition

ImageNet Object Localization Challenge

Identify the objects in images

ImageNet · 25 teams · 11 years to go

Overview | Data | Discussion | Leaderboard | Rules

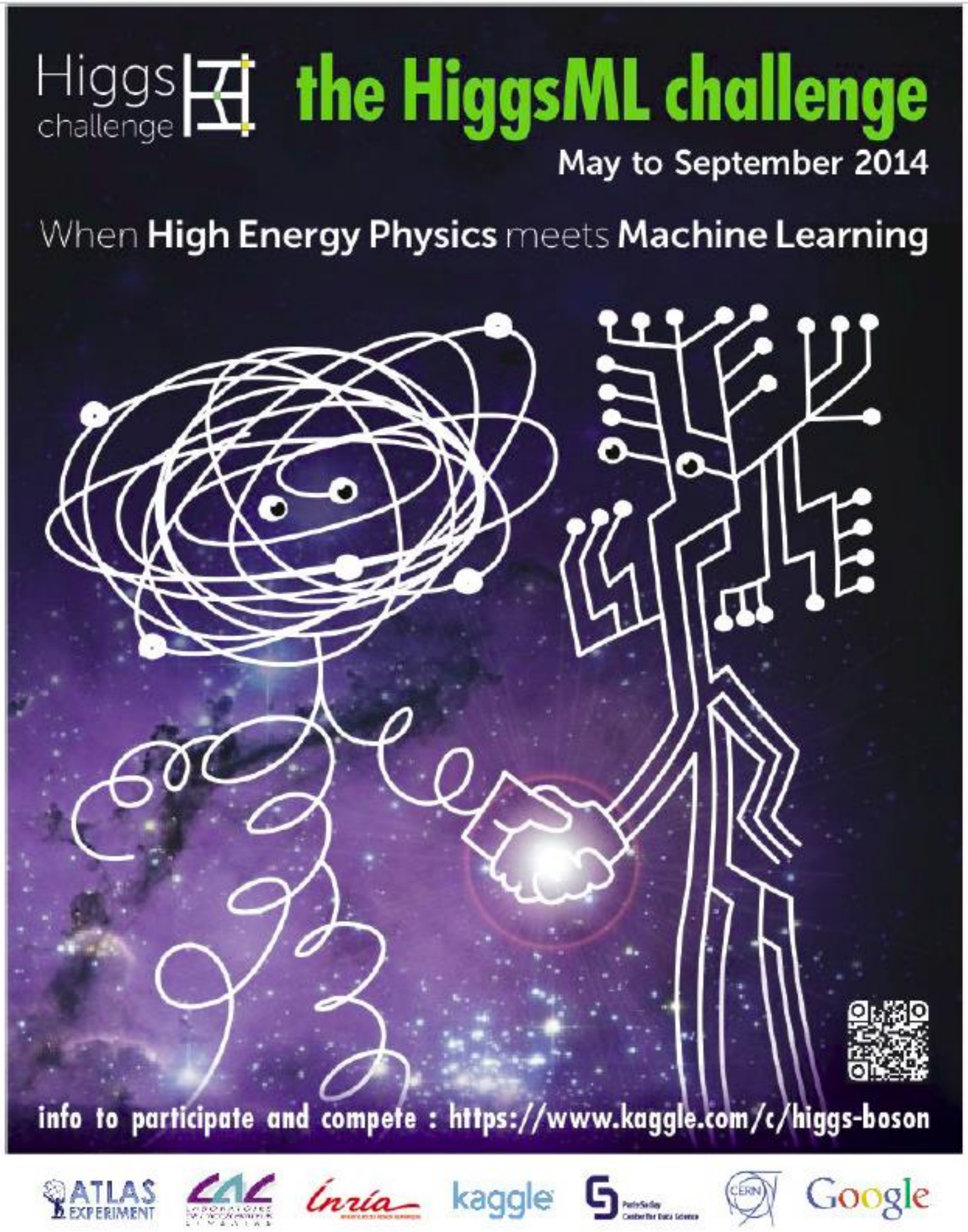
Overview

- Description
- Evaluation
- Timeline

Competition Description

While It's pretty easy for people to identify subtle differences in photos, computers still have a ways to go. Visually similar items are tough for computers to count, like this overlapping bunch of bananas

AI and Data-Intensive Science: Three Examples


Higgs challenge
the HiggsML challenge
May to September 2014
When High Energy Physics meets Machine Learning
info to participate and compete : https://www.kaggle.com/c/higgs-boson
ATLAS EXPERIMENT
Inria
kaggle
CERN
Google

Machine Learning winners of the Higgs Challenge

- Winner Gábor Melis, a graduate in software engineering and mathematics, developed an algorithm that is an ensemble of deep neural networks trained on random subsets of data provided with very little feature engineering and no physics knowledge

- Runner-up Tim Salimans, who has a PhD in Econometrics and works as a data science consultant, developed a solution he describes as a combination of a large number of boosted decision tree ensembles

- A Special High Energy Physics meets Machine Learning Award was presented to Tianqi Chen and Tong He of Team Crowwork. Their XG Boost algorithm was an excellent compromise between performance and simplicity, which could improve tools currently used in high-energy physics.

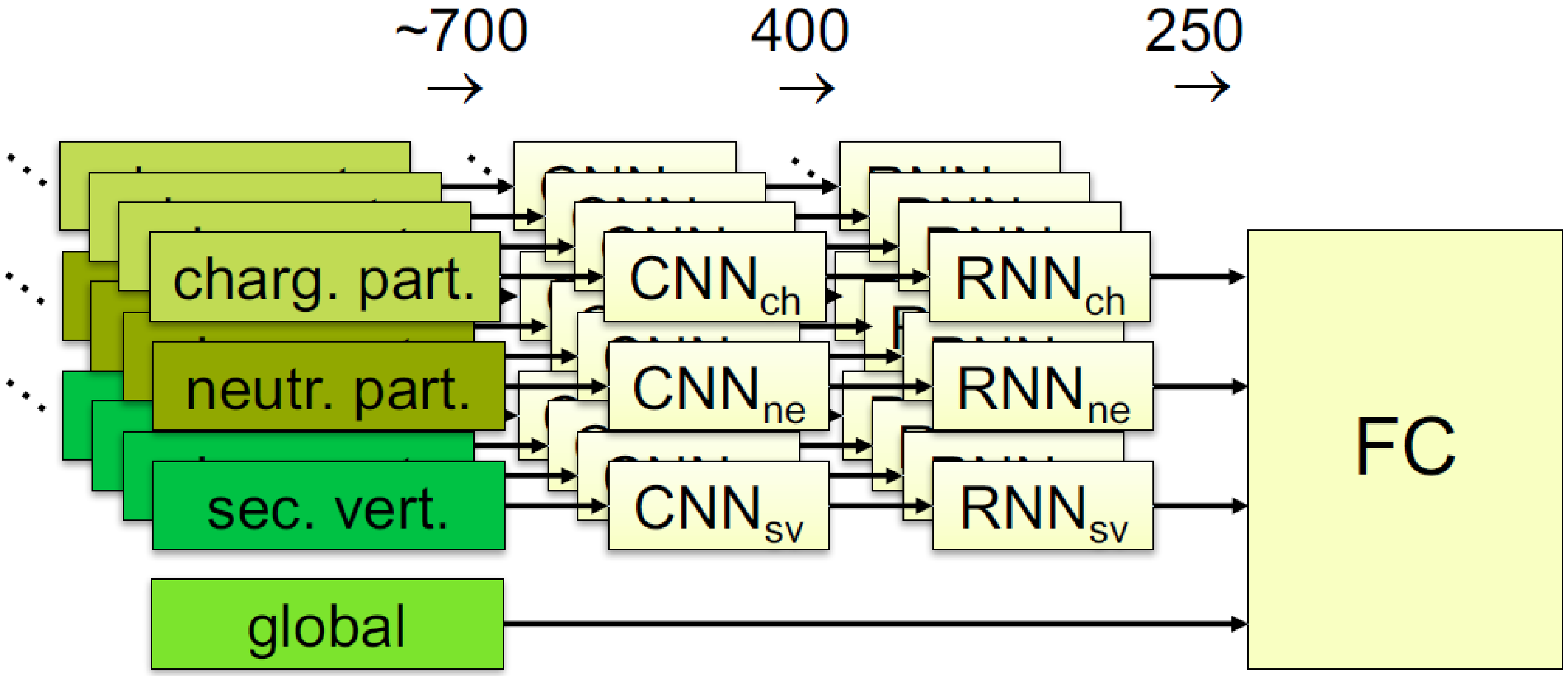
Winners of the Higgs Machine Learning Challenge: Gábor Melis and Tim Salimans (top row), Tianqi Chen and Tong He (bottom row).

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



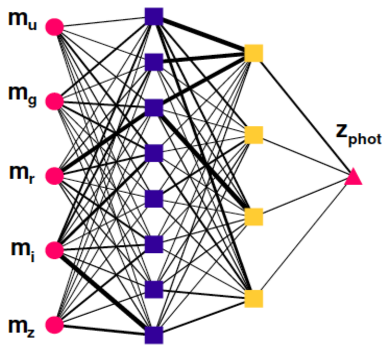
~ 700 inputs and 250.000 model parameters

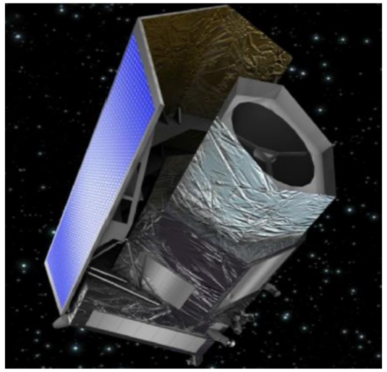
- Particle and vertex based DNN has factor 10 less free parameters than a generic Dense DNN would have
- 100M jets used for training, overtraining is not an issue

Machine Learning in Astronomy

Machine learning examples from Astronomy:

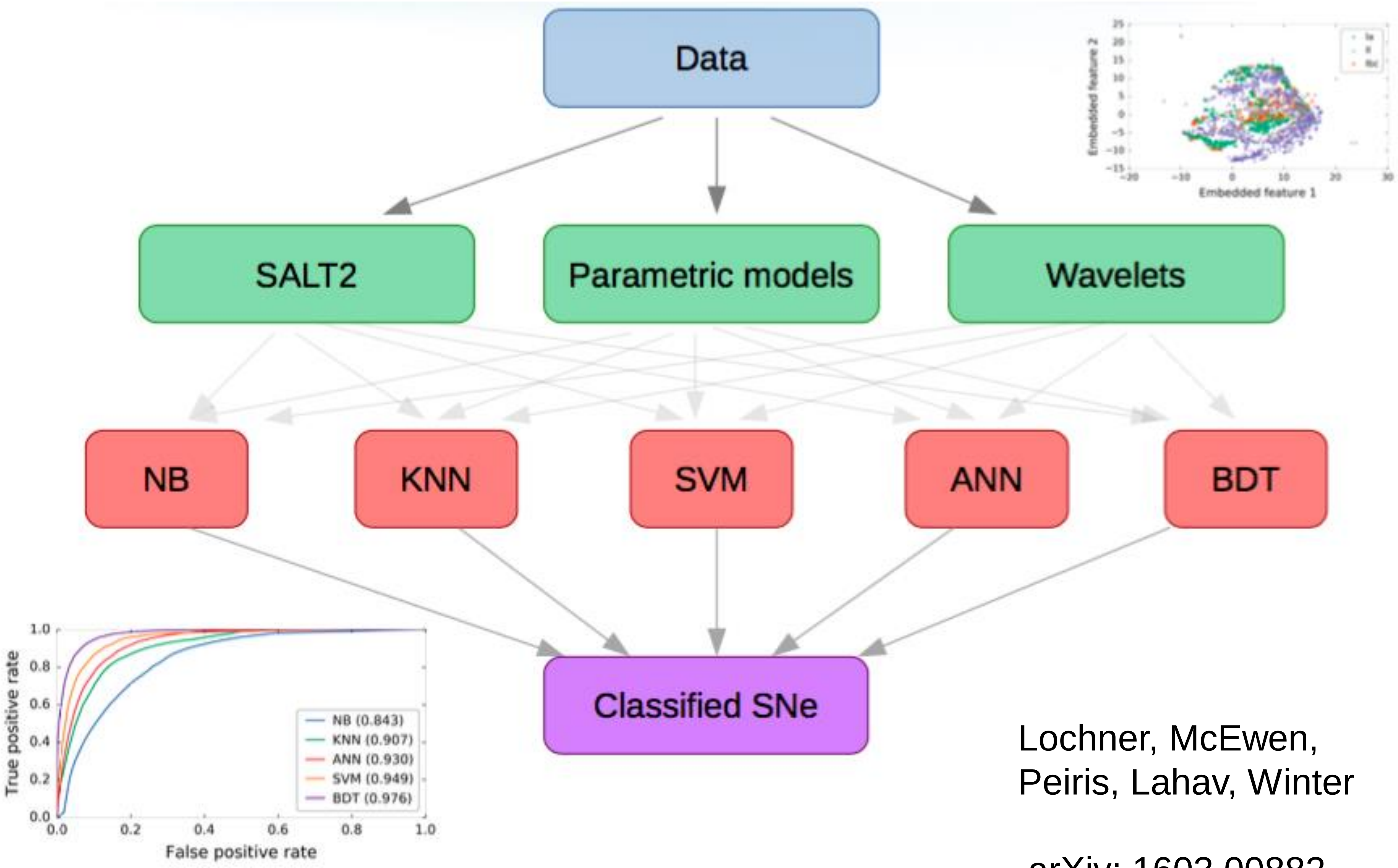
- **Classification:**
  galaxy type (0908.2033), star/galaxy (1306.5236),
  Supernovae Ia (1603.00882)
- **Photo-z** (1507.00490)
- **Mass of the Local Group** (1606.02694)
- **Search for Planet 9 in DES**





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 | Data | Kernels | Discussion | Leaderboard | Rules

Overview

- Description
- Evaluation
- Prizes
- Timeline

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.

Featured Prediction Competition

TrackML Particle Tracking Challenge

High Energy Physics particle tracking in CERN detectors

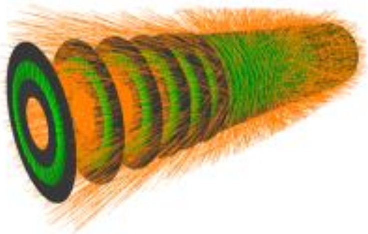
$25,000 Prize Money

CERN · 656 teams · 2 months ago

Overview | Data | Kernels | Discussion | Leaderboard | Rules

Overview

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



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.

Event rates have already reached hundreds of millions of collisions per second, meaning physicists must sift through tens of petabytes of data per year. And, as the resolution of detectors improve, ever better software is needed for real-time pre-processing and filtering of the most promising events, producing even more data.

To help address this problem, a team of Machine Learning experts and physics scientists working at CERN (the world largest high energy physics laboratory), has partnered with Kaggle and prestigious sponsors to answer the question: can machine learning assist high energy physics in discovering and characterizing new particles?

Specifically, in this competition, you're challenged to build an algorithm that quickly reconstructs particle tracks from 3D points left in the silicon detectors. This challenge consists of two phases:

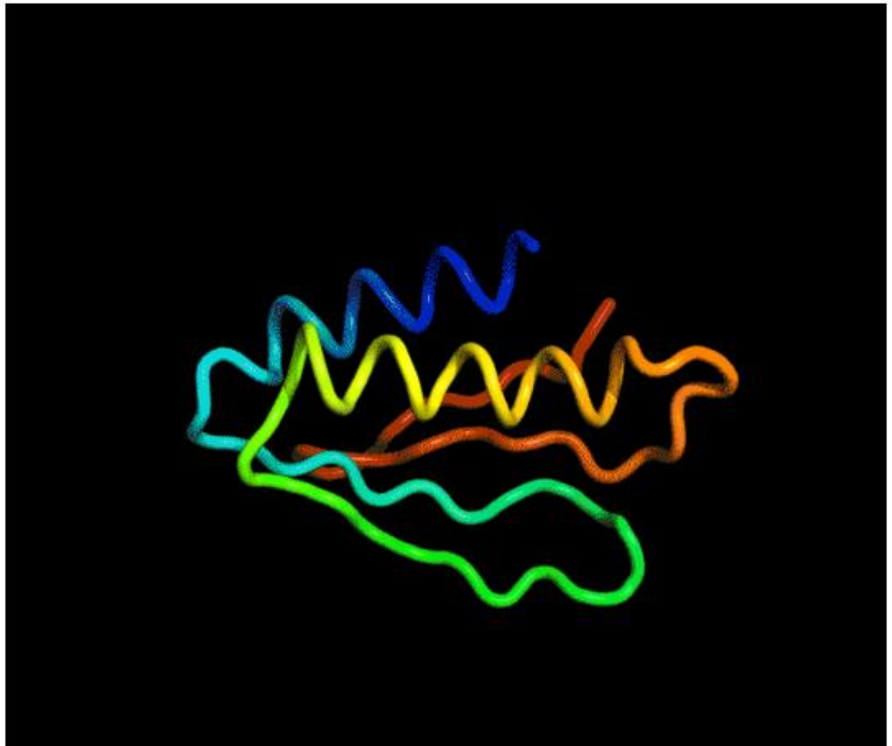
- The Accuracy phase has run on Kaggle from May to 13th August 2018 (Winners to be announced by end September). Here we'll be focusing on the highest score, irrespective of the evaluation time. This phase is an official IEEE WCCI competition (Rio de Janeiro, Jul 2018).
- The Throughput phase will run on Codalab starting in September 2018. Participants will submit their software which is evaluated by the platform. Incentive is on the throughput (or speed) of the evaluation while reaching a good score. This phase is an official NIPS competition (Montreal, Dec 2018).

All the necessary information for the Accuracy phase is available here on Kaggle site. The overall TrackML challenge web site is there.

Google DeepMind's AlphaGo Zero

AlphaFold: Using AI for scientific discovery

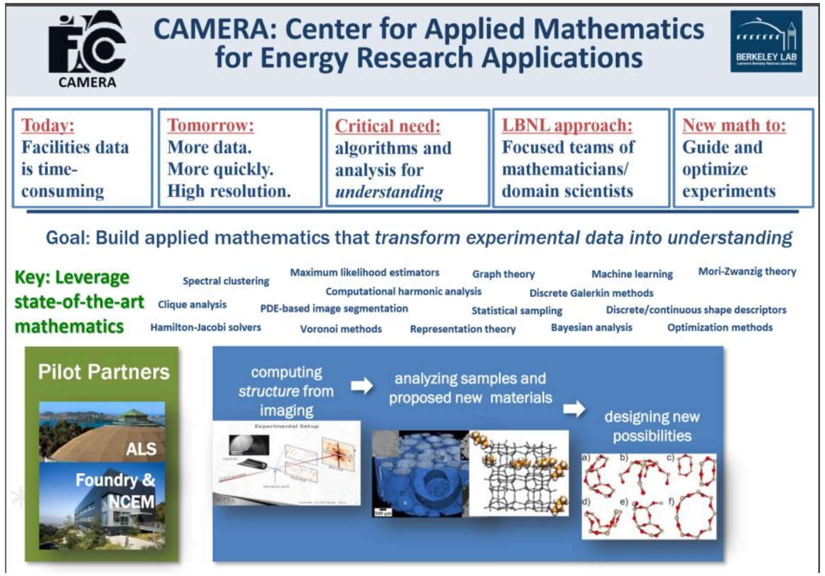
Our system, **AlphaFold**, which we have been working on for the past two years, builds on years of prior research in using vast genomic data to predict protein structure. The 3D models of proteins that AlphaFold generates are far more accurate than any that have come before—making significant progress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



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

Scientific Machine Learning (SciML) at the US DOE Labs

Data Science at the major US DOE Facilities Labs



BROOKHAVEN NATIONAL LABORATORY Newsroom Media & Communications Office

Newsroom | Photos | Videos | Fact Sheets | Lab History | News Categories | Contacts

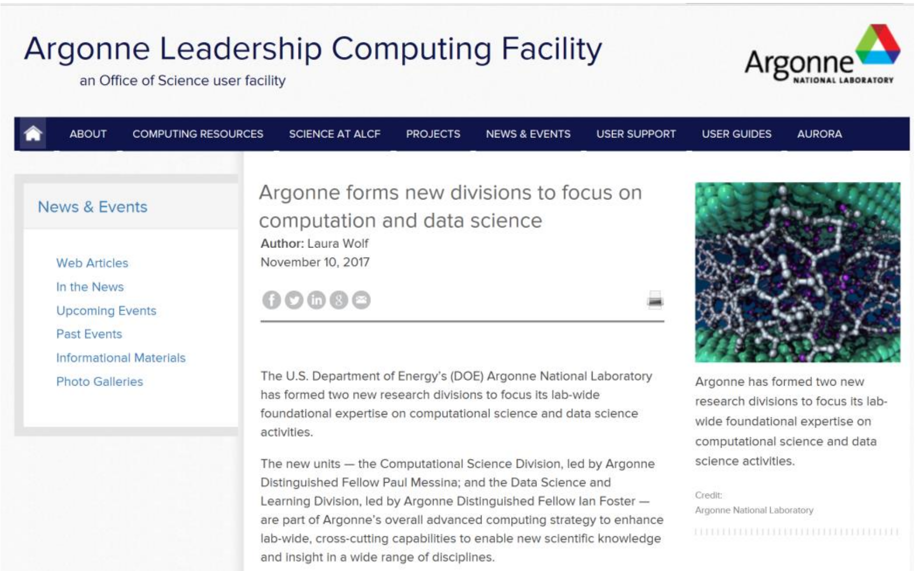
Contact: Ariana Tantillo, (631) 344-2347, or Peter Genzer, (631) 344-3174 share:

Shantenu Jha Named Chair of Brookhaven Lab's Center for Data-Driven Discovery

With expertise in high-performance and distributed computing, Jha will lead a center that provides the focal point for data science research and development

November 27, 2017

UPTON, NY—Computational scientist Shantenu Jha has been named the inaugural chair of the Center for Data-Driven Discovery (C3D) at the U.S. Department of Energy's (DOE) Brookhaven National Laboratory, effective October 1. Part of the Computational Science Initiative (CSI), C3D is driving the integration of domain, computational, and data science expertise across Brookhaven Lab's science programs and facilities, with the goal of accelerating and expanding scientific discovery. Outside the Lab, C3D is serving as a focal point for the recruitment of future data scientists and collaboration with other institutions.



Argonne Leadership Computing Facility
an Office of Science user facility
Argonne NATIONAL LABORATORY

ABOUT | COMPUTING RESOURCES | SCIENCE AT ALCF | PROJECTS | NEWS & EVENTS | USER SUPPORT | USER GUIDES | AURORA

News & Events
- Web Articles
- In the News
- Upcoming Events
- Past Events
- Informational Materials
- Photo Galleries

Argonne forms new divisions to focus on computation and data science

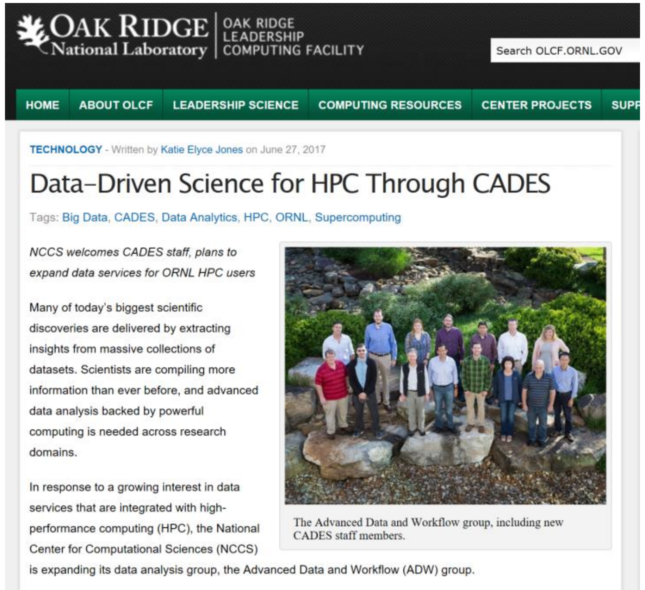
Author: Laura Wolf
November 10, 2017

The U.S. Department of Energy's (DOE) Argonne National Laboratory has formed two new research divisions to focus its lab-wide foundational expertise on computational science and data science activities.

The new units — the Computational Science Division, led by Argonne Distinguished Fellow Paul Messina; and the Data Science and Learning Division, led by Argonne Distinguished Fellow Ian Foster — are part of Argonne's overall advanced computing strategy to enhance lab-wide, cross-cutting capabilities to enable new scientific knowledge and insight in a wide range of disciplines.

Argonne has formed two new research divisions to focus its lab-wide foundational expertise on computational science and data science activities.

Credit: Argonne National Laboratory



OAK RIDGE National Laboratory | OAK RIDGE LEADERSHIP COMPUTING FACILITY
Search OLCF.ORNL.GOV

HOME | ABOUT OLCF | LEADERSHIP SCIENCE | COMPUTING RESOURCES | CENTER PROJECTS | SUPP

TECHNOLOGY - Written by Katie Elyce Jones on June 27, 2017

Data-Driven Science for HPC Through CADES

Tags: Big Data, CADES, Data Analytics, HPC, ORNL, Supercomputing

NCCS welcomes CADES staff, plans to expand data services for ORNL HPC users

Many of today's biggest scientific discoveries are delivered by extracting insights from massive collections of datasets. Scientists are compiling more information than ever before, and advanced data analysis backed by powerful computing is needed across research domains.

In response to a growing interest in data services that are integrated with high-performance computing (HPC), the National Center for Computational Sciences (NCCS) is expanding its data analysis group, the Advanced Data and Workflow (ADW) group.

The Advanced Data and Workflow group, including new CADES staff members.

ORNL Summit System Overview

System Performance

- Peak of 200 Petaflops (FP_{64}) for modeling & simulation
- Peak of 3.3 ExaOps (FP_{16}) for data analytics and artificial intelligence

The system includes

- 4,608 nodes
- Dual-rail Mellanox EDR InfiniBand network
- 250 PB IBM file system transferring data at 2.5 TB/s

Each node has

- 2 IBM POWER9 processors
- 6 NVIDIA Tesla V100 GPUs
- 608 GB of fast memory
  (96 GB HBM2 + 512 GB DDR4)
- 1.6 TB of non-volatile memory

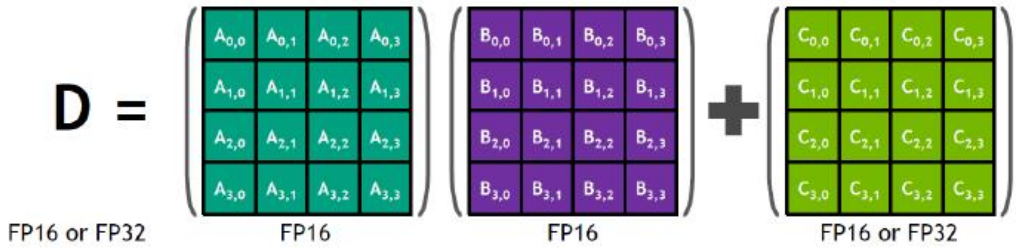
OAK RIDGE National Laboratory

Summit Contains 27,648 NVIDIA Tesla V100s

Each Tesla v100 GPU has:

- 300 GB/s total BW (**NVLink** v2.0)
- 5,120 CUDA cores (64 on each of 80 SMs)
- 640 **Tensor cores** (8 on each of 80 SMs)
- 20MB Registers | 16MB Cache | 16GB HBM2 @ 900 GB/s
- 7.5 DP TFLOPS | 15 SP TFLOPS | 120 FP_{16} TFLOPS

- Tensor cores do mixed precision multiply-add of 4x4 matrices



D = AB + C

Type	Size	Range	$u = 2^{-t}$
half	16 bits	$10^{\pm 5}$	$2^{-11} \approx 4.9 \times 10^{-4}$
single	32 bits	$10^{\pm 38}$	$2^{-24} \approx 6.0 \times 10^{-8}$
double	64 bits	$10^{\pm 308}$	$2^{-53} \approx 1.1 \times 10^{-16}$
quadruple	128 bits	$10^{\pm 4932}$	$2^{-113} \approx 9.6 \times 10^{-35}$

- The Modeling & Simulation community can benefit from better utilizing mixed / reduced precision
- Eg: Possible to achieve 4x FP64 peak for 64bit LU on V100 with iterative mixed precision (Dongarra et al.)

Summit Excels Across Simulation, Analytics, AI

- Data analytics – CoMet bioinformatics application for comparative genomics. Used to find sets of genes that are related to a trait or disease in a population. Exploits cuBLAS and Volta tensor cores to solve this problem 5 orders of magnitude faster than previous state-of-art code.
  - **Has achieved 2.36 ExaOps** mixed precision (FP_{16}-FP_{32}) on Summit
- Deep Learning – global climate simulations use a half-precision version of the DeepLabv3+ neural network to learn to detect extreme weather patterns in the output
  - **Has achieved a sustained throughput of 1.0 ExaOps (FP_{16})** on Summit
- Nonlinear dynamic low-order unstructured finite-element solver accelerated using mixed precision (FP_{16} thru FP_{64}) and AI generated preconditioner. Answer in FP_{64}
  - **Has achieved 25.3 fold speedup** on Japan earthquake – city structures simulation
- **Half-dozen Early Science codes are reporting >25x speedup on Summit vs. Titan**

‘AI for Science’ at Harwell and the Scientific Machine Learning Group

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, …

Scientific Machine Learning at Harwell

- SciML activity at Harwell is complementary to industry-focused Data Analytics activity at the Hartree Supercomputer Centre on the Daresbury Campus
- Collaborative ML projects with joint funding from Turing-Diamond-SciML initiative:
  - Cryo-EM particle picking
  - 4D Tomography
  - Mutimodal
- Working with ISIS Neutron Facility on SANS, Reflectometry and magnetic scattering experiments
  - Overlap with needs of SAXS and Reflectometry at Diamond

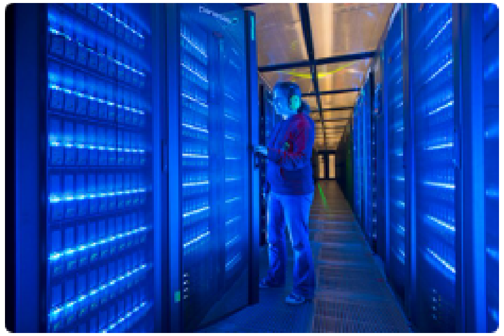
➢ Now awarded funding in Turing's 'AI for Science' theme in the UKRI Strategic Priorities Fund Wave 1.

STFC speeding up analysis of experimental data by using AI technologies

7 January 2019

STFC is about to harness the power of artificial intelligence (AI) and machine learning to more efficiently sort through the swathes of experimental data produced at its national multidisciplinary science facilities with the aim of making quicker scientific breakthroughs.

This work will be carried out in collaboration with the Alan Turing Institute, which has recently been awarded £40million from UK Research and Innovation to fund research into developing AI technology to benefit the engineering, health, science and criminal justice sectors.



JASMIN.
(Credit: STFC/RAL Space)

STFC's Chief Data Scientist Tony Hey will co-direct the project. He said: "There are many areas of science that now generate such large volumes of data that processing it is laborious and inefficient. There is an opportunity here for us to use the tools of data science and AI to assist scientists create new scientific knowledge more quickly and efficiently.

"It is vital that the UK develops suitable systems for mining and exploiting data at our national experimental facilities in order to maintain its position at the forefront of the global research community."

STFC will be specifically focusing on applying AI and advanced machine learning technologies to the experimental data generated by the facilities at the Harwell Campus – Diamond, ISIS neutron and muon source, the UK's Central Laser Facility and the NERC Centre for Environmental Data Analytics with its JASMIN super data cluster. This AI capability will be known as the 'Turing Hub', and will be hosted at STFC's Scientific Computing Department.

The Alan Turing Institute has allocated funding to set up the Turing Hub at Harwell, which includes funding for four data scientists and an AI computer system. The Hub will give users of the facilities the new ability to utilise AI technologies to collect and analyse their data, which will significantly increase the efficiency and productivity of users from both academia and industry.

The STFC Hartree Centre is also collaborating with the Alan Turing Institute to work on AI technologies for industry, and the new Turing Hub will be working closely with the Hartree Centre.

Last updated: 15 January 2019

Initial Focus on Big Scientific Data ML Benchmarks

Idea is to create scientific datasets that are sufficiently large and complex to provide a realistic testing ground for ML algorithms

- 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 CEDA data
- Datasets from Culham Centre for Fusion Energy

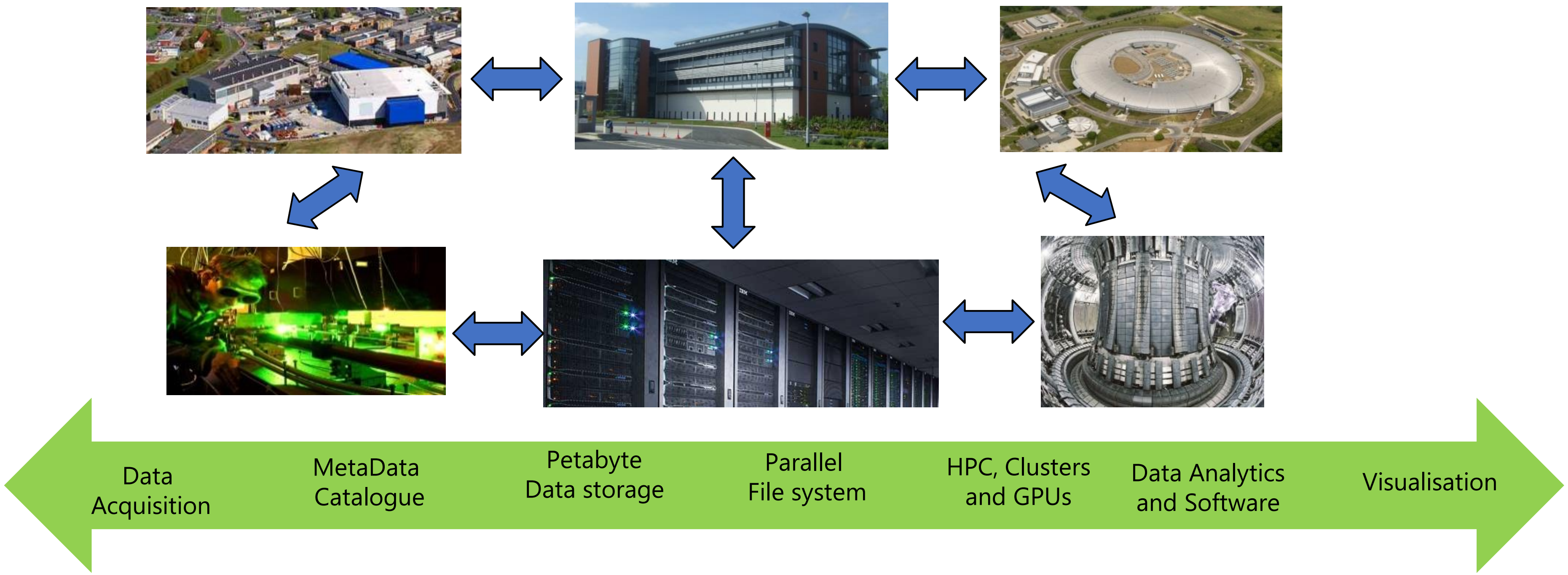
➢ Experimentation and training in Machine Learning technologies executed on different hardware architectures

➢ Use as basis for training courses for academia and industry

➢ R&D on optimization, robustness and transparency

Vision: A Harwell Campus 'AI for Science Centre'

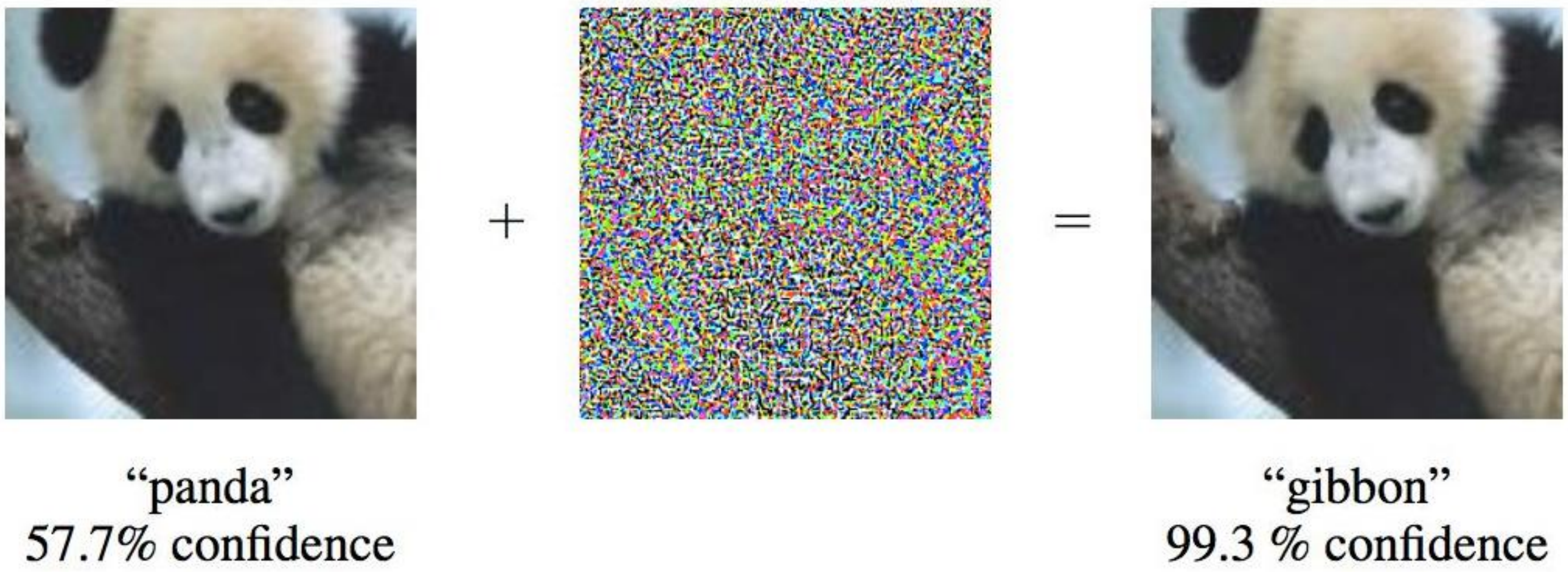
Compute and Data Infrastructure + Software + Expertise



- **Support university users of the Facilities for new science**
- **Expertise and Training for Industry**
- **Focus on R&D in Applied AI and Machine Learning**

Some concerns …

Adversarial Noise and Deep Learning?



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

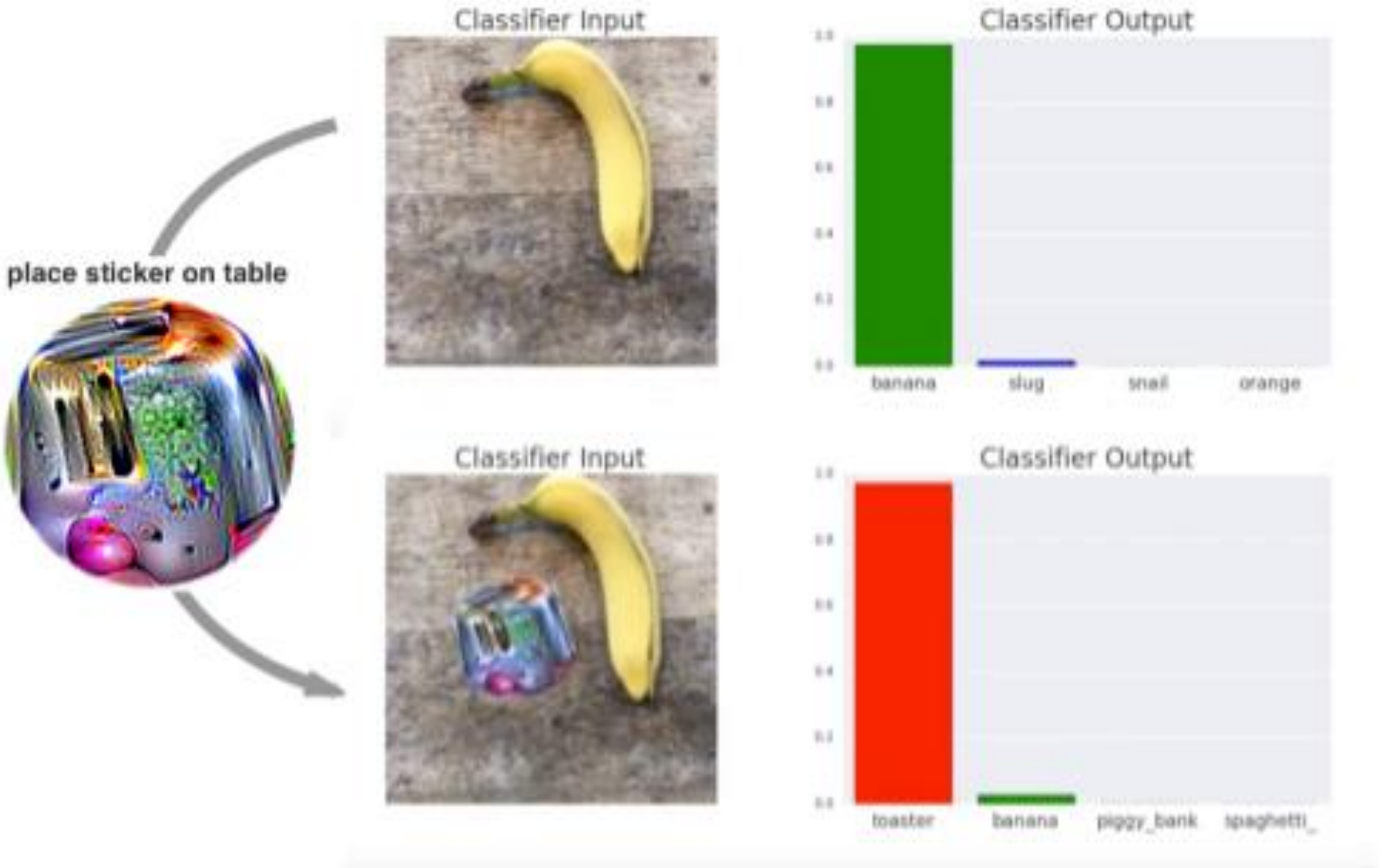
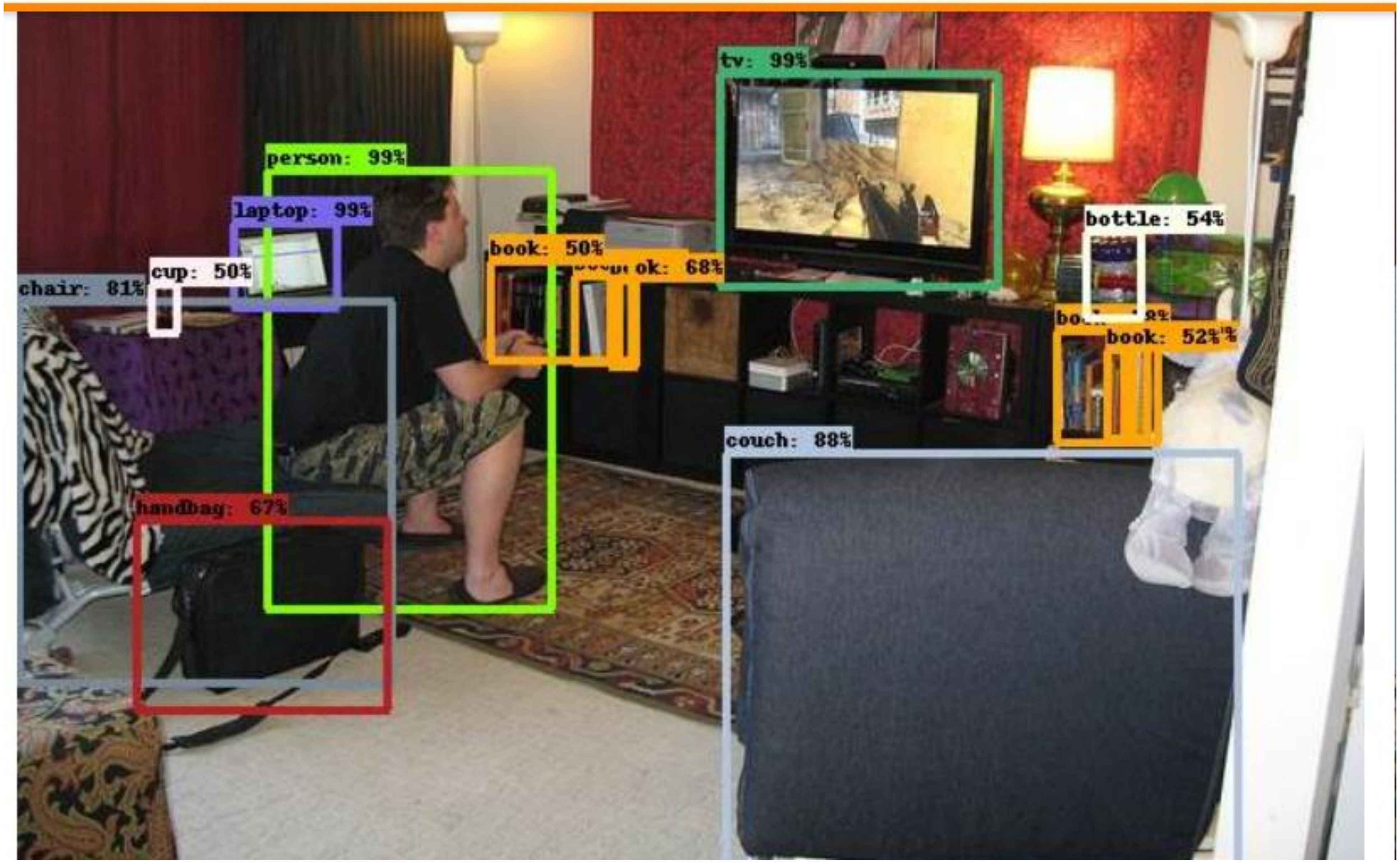


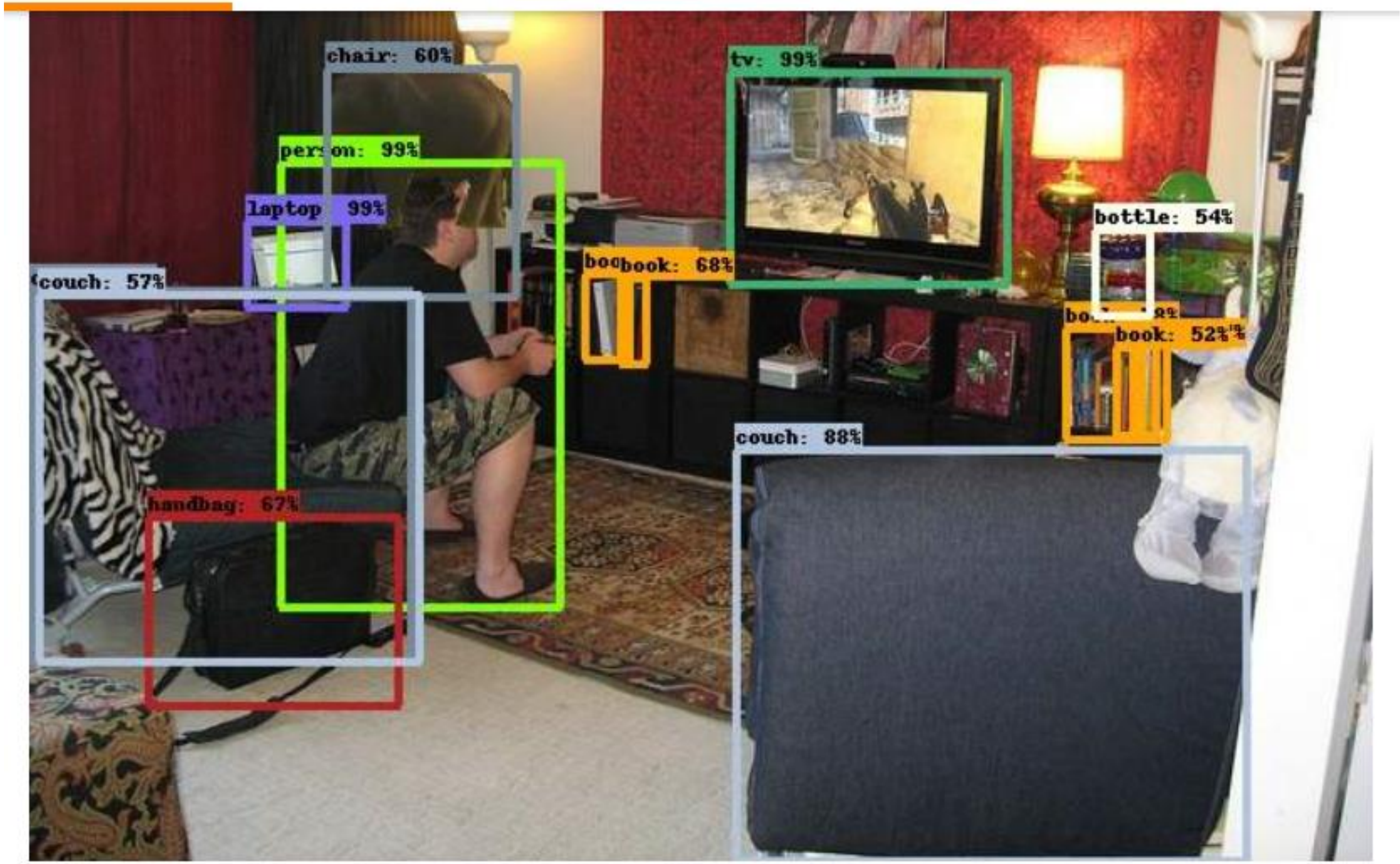
Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence (bottom plot). See the following video for a full demonstration: `https://youtu.be/i1sp4X57TL4`

https://www.theverge.com/2018/1/3/16844842/ai-computer-vision-trick-adversarial-patches-google



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

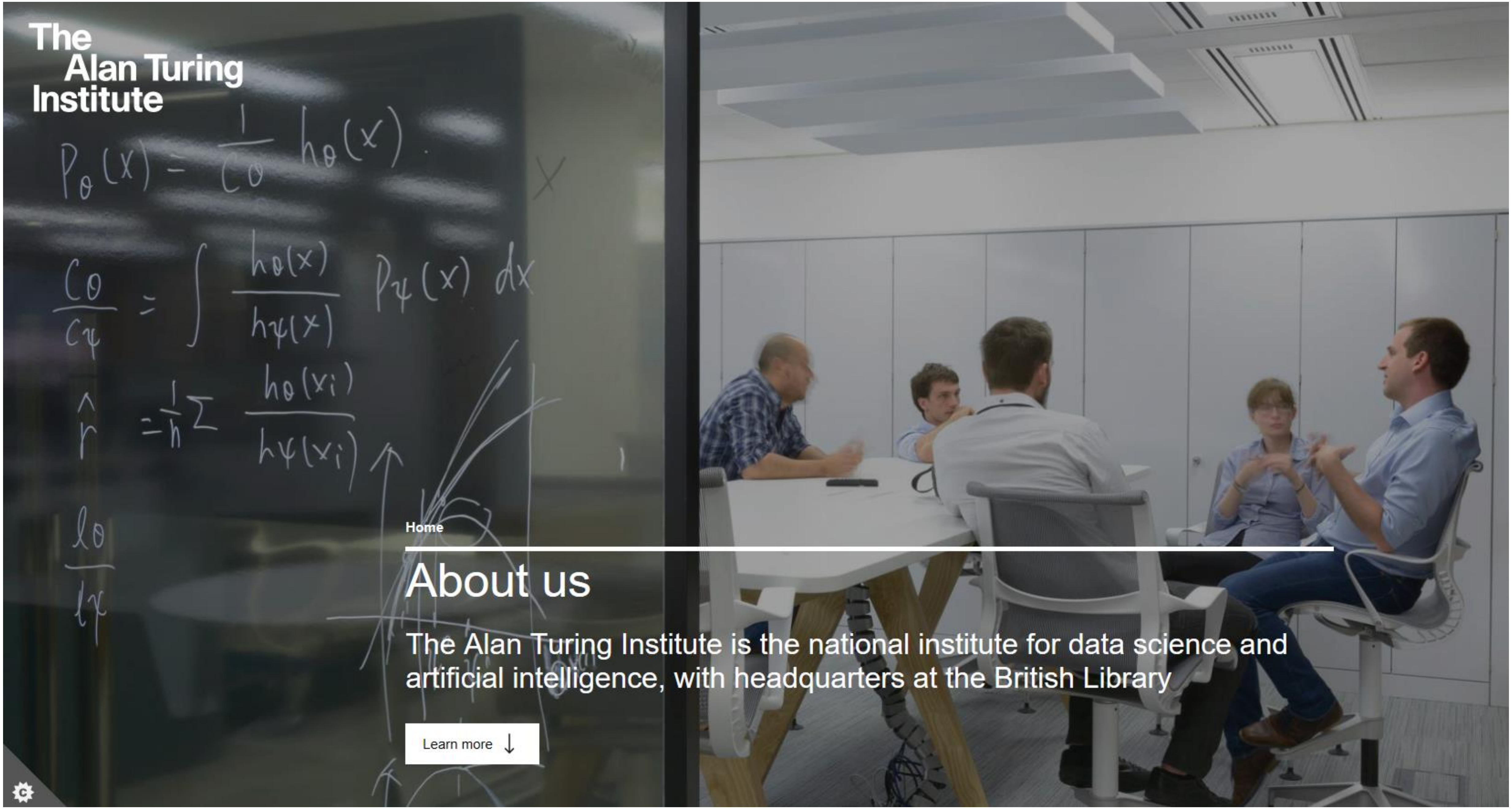
arXiv:1808.03305v1 [cs.CV] 9 Aug 2018



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

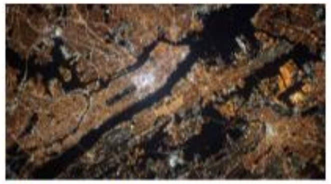
[arXiv:1808.03305v1 [cs.CV] 9 Aug 2018](arXiv:1808.03305v1 [cs.CV] 9 Aug 2018)

‘AI for Science’ at Turing


The Alan Turing Institute
Home
About us
The Alan Turing Institute is the national institute for data science and artificial intelligence, with headquarters at the British Library
Learn more

Research Themes at Turing

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.

NVIDIA

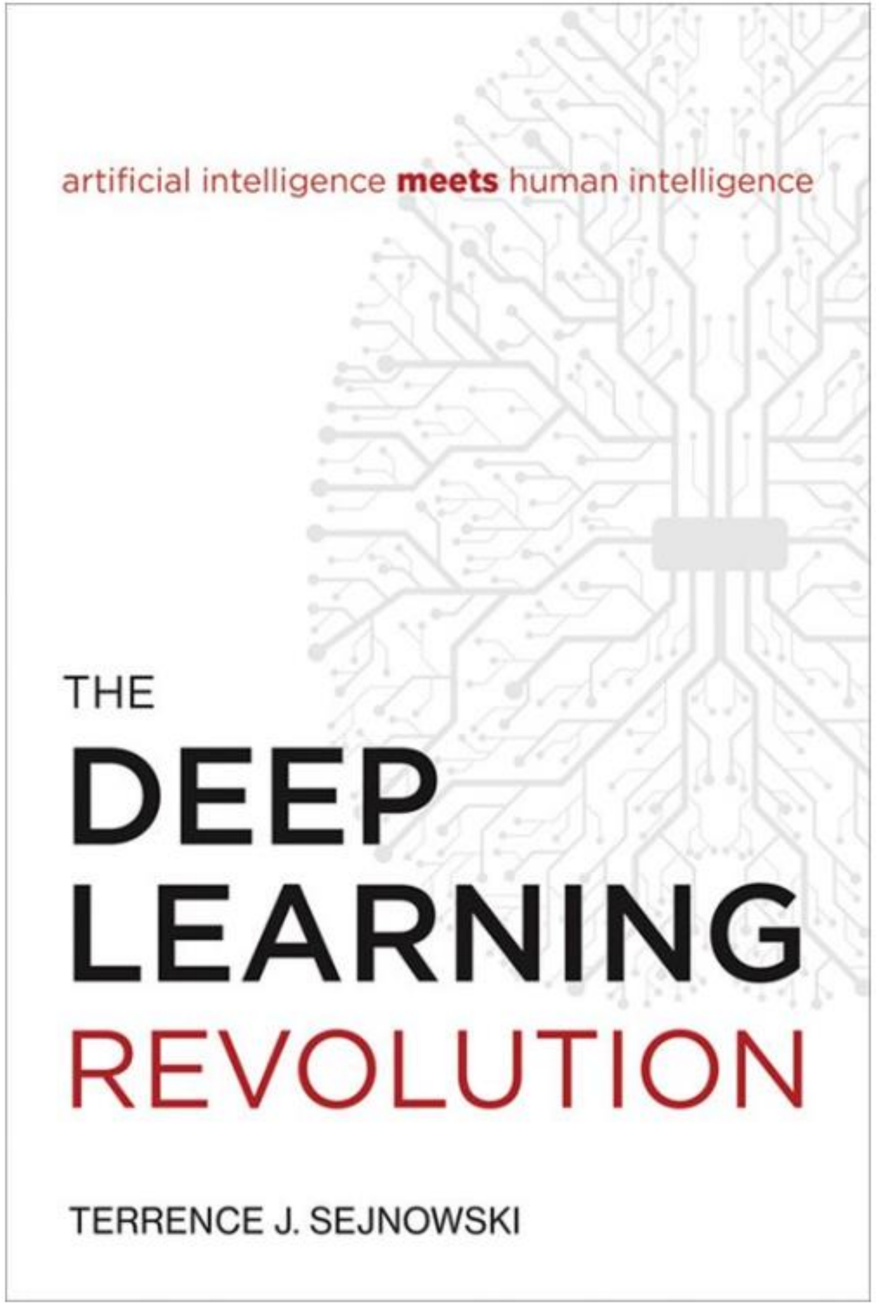
DATA CENTER PRODUCTS SOLUTIONS APPS FOR DEVELOPERS TECHNOLOGIES

DGX-2 OVERVIEW DATA SCIENCE TOUR CONTACT US

NVIDIA DGX-2

The world's most powerful AI system for the most complex AI challenges.

Experience new levels of AI speed and scale with NVIDIA® DGX-2™, the first 2 petaFLOPS system that combines 16 fully interconnected GPUs for 10X the deep learning performance. It's powered by NVIDIA® DGX™ software and a scalable architecture built on NVIDIA NVSwitch, so you can take on the world's most complex AI challenges.



"What made deep learning take off was big data. … The explosion of data is having an influence not just on science and engineering but also on every area of society."

Terry Sejnowski

Terry Sejnowski and Geoffrey Hinton in 1990