Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477

# A summary of the workshop "AI at CERN and SKA“


Vladislav Stolyarov
Cavendish Astrophysics, University of Cambridge

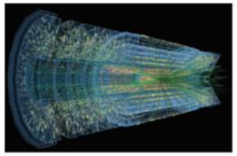

**3rd ASTERICS-OBELICS Workshop**
23-25 October 2018, Cambridge, UK.


H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).

# AI at CERN and SKA

## 17-18 September 2018, Alan Turing Institute, London

SKA Midrange telescope for the Karoo desert in South Africa



LHC timing detectors

**Overview**

This two-day workshop, to be held in central London at the Alan Turing Institute, will bring together the community of the Alan Turing Institute (ATI), AI researchers from industry, and scientists working with CERN or SKA surrounding the application of Artificial Intelligence (AI) for scientific discovery. The aim is to gain an understanding of what has been achieved in High Energy Physics (HEP) using data from CERN, and in astrophysics leveraging data obtained through radio astronomy. Areas of interest include processing of the raw instrument data, use of science data products for research and applicable technology. In discussions, possible new areas of collaboration and research will be explored.

The agenda will include a mix of invited talks, accepted submissions and generous time for discussions.

**Topics of interest**

- AI at very large scales
- Models used for scientific discovery
- Opening new doors in scientific discovery through AI
- Technological discoveries and implementations

**Organisers and program committee**

Prof Paul Alexander, (Cambridge, UK)
Prof Alan Barr, (Oxford, UK)
Dr Peter Braam, (Peter Braam)
Dr Maria Girone, (CERN)
Prof Terry Lyons, (Oxford, UK)
Dr Nicholas Rees, (SKA)
Prof Ian Shipsey, (Oxford, UK)

**Primary contact**

Dr Peter Braam pjb624@cam.ac.uk

**Registration & further information**

https://indico.cern.ch/event/745580/

Day 1

- General talks
- CERN/LHC talks
- Discussion

Day 2

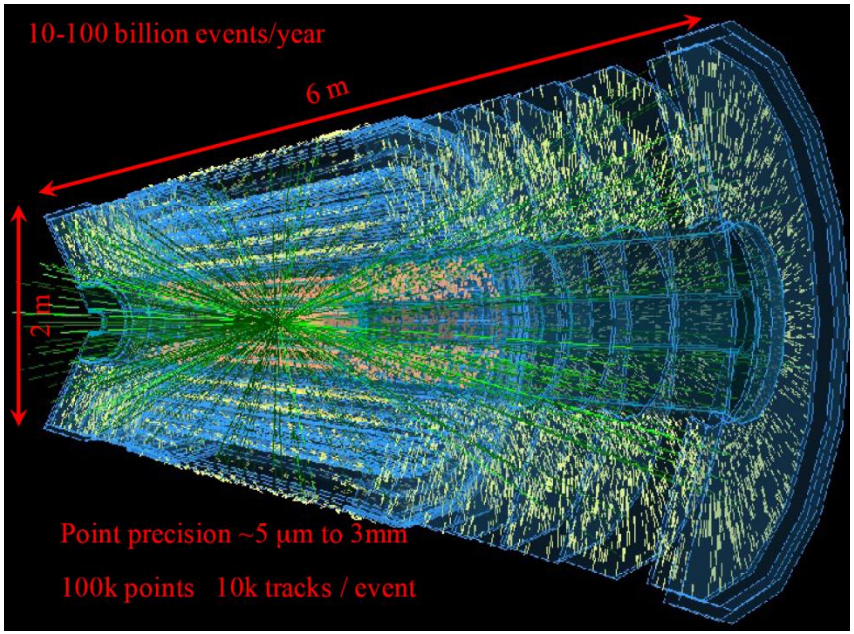
- Astro/SKA talks
- Discussion

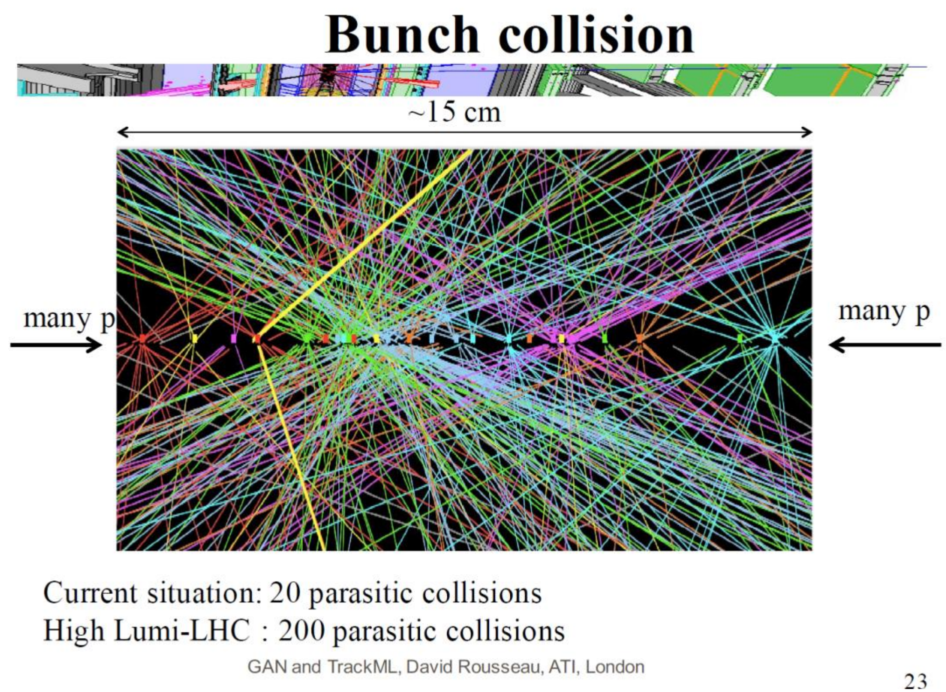
Presentations:
https://indico.cern.ch/event/745580/

# Machine Learning in HEP : GAN TrackML and more

**David Rousseau**
**LAL-Orsay**
rousseau@lal.in2p3.fr

**CERN/SKA workshop, Alan Turing Institute, London**
**17th Sep 2018**


10-100 billion events/year
6 m
2 m
Point precision ~5 µm to 3mm
100k points 10k tracks / event


Bunch collision
~15 cm
many p
many p
Current situation: 20 parasitic collisions
High Lumi-LHC : 200 parasitic collisions
GAN and TrackML, David Rousseau, ATI, London
23

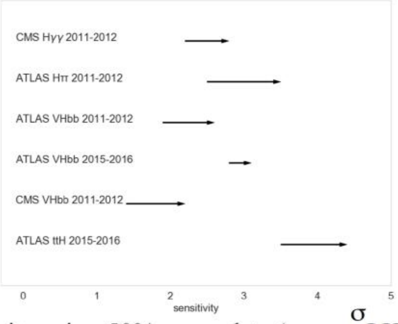
# Outline

- Boosted Decision Tree applied to Higgs physics
- Generative Adversarial Network for fast experiment simulation
- The Tracking Machine Learning challenge

GAN and TrackML, David Rousseau, ATI, London 2

# ML on Higgs Physics

- At LHC, Machine Learning used almost since first data taking (2010) for reconstruction and analysis
- In most cases, Boosted Decision Tree with Root-TMVA, on ~10 variables
- For example, impact on Higgs boson sensitivity at LHC:



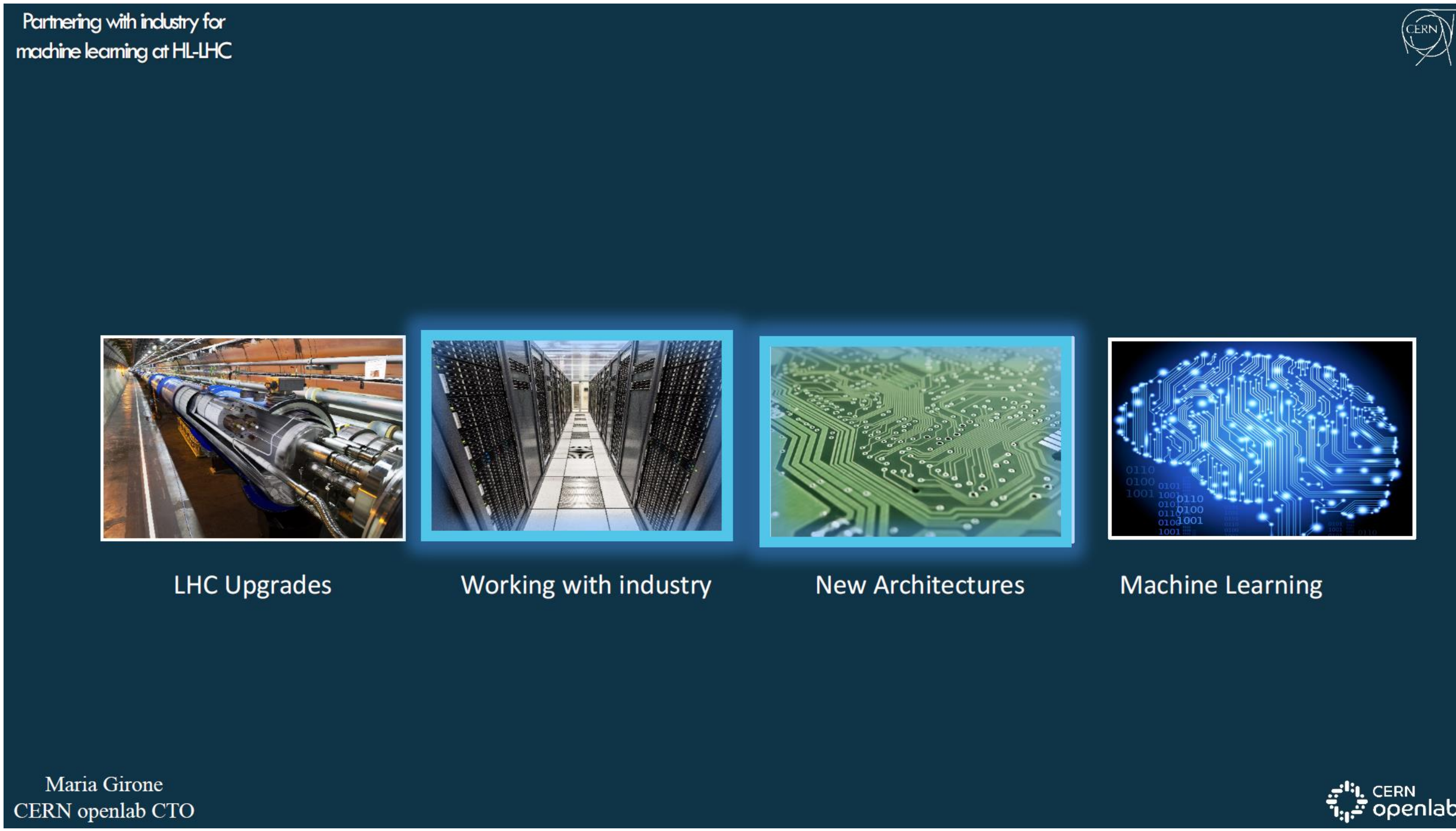

σ

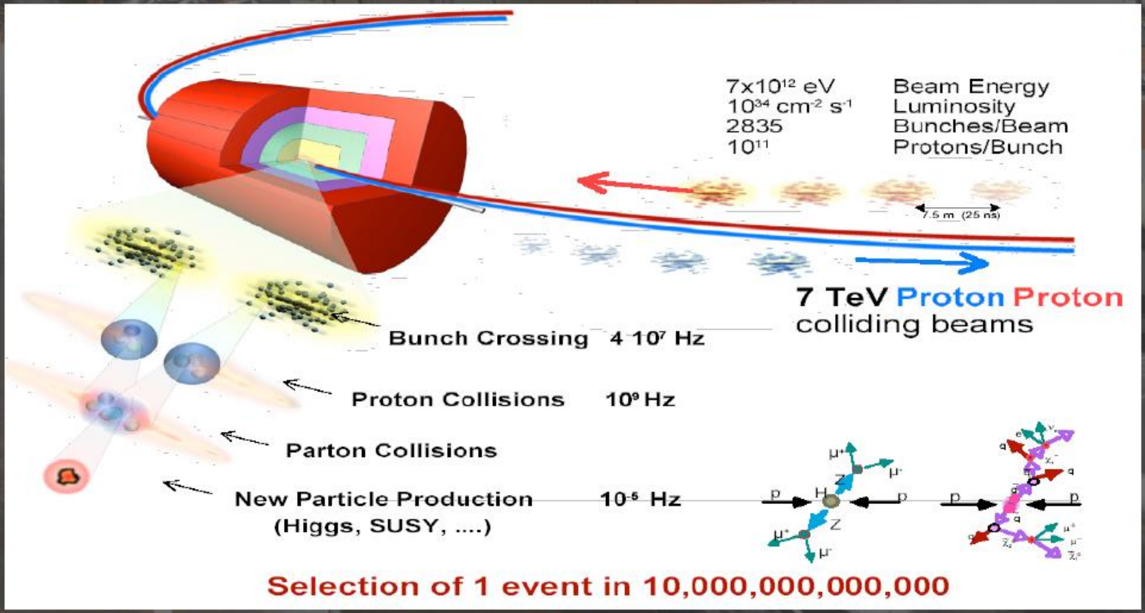
➔sensitivity gain ~50% more data (~ more LHC running time) GAN and TrackML, David Rousseau, ATI, London 4

# Conclusion

- We (in HEP) are analysing data from multi-billion € projects➔should make the most out of it!
- BDT work well, still the recommended tool for dozen variables classification/regression
- Recent explosion of novel (for HEP) ML techniques
- Focused here on two potential large speedup:
  - GAN for simulation
  - Possible novel tracking algorithms from TrackML challenge @trackmllhc : https://competitions.codalab.org/competitions/20112
- Never underestimate the time for :
  - (1) Great ML idea➔
  - (2) …demonstrated on toy dataset➔
  - (3) …demonstrated on semi-realistic simulation ➔
  - (4) …demonstrated on real experiment analysis/dataset ➔
  - (5) …experiment publication using the great idea

# Partnering with industry for machine learning at HL-LHC
## Maria Girone, CERN openlab CTO






Rate of new physics is 1 event in $10^{12}$

Selecting a new physics event is like choosing 1 grain of sand in 20 volley ball courts

More collisions help physicists to observe rare processes and study with greater precision.

Maria Girone
CERN openlab CTO

- The LHC experiments are working closely with industry via CERN openlab on machine learning
  - Focus on adoption of accelerators (GPUs, FPGAs)
  - Engineering resources dedicated to support the application porting and increase knowhow on deep learning techniques

**Data acquisition**
- Real time event categorization
- Data monitoring & certification
- Fast inference for trigger systems

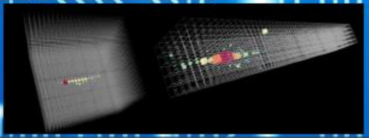
**Data Reconstruction**
- Calorimeter reconstruction
- Boosted object jet tagging

**Data Processing**
- Computing resource optimization
- Predicting data popularity
- Intelligent networking

**Data Simulation**
- Adversarial networks
- Fast simulation

**Data Analysis**
- Knowledge base
- Data reduction
- Searches for new physics



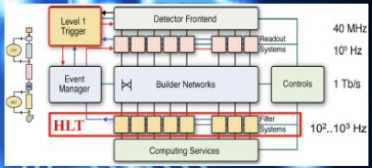




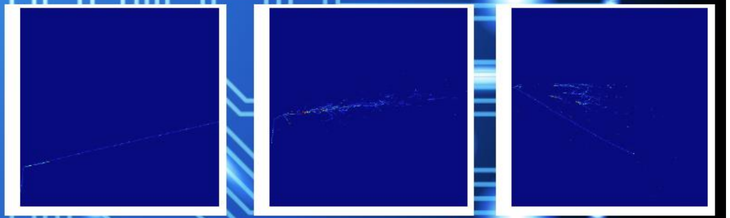
## Opportunities for collaboration.

CERN openlab is launching a a project with Micron to investigate applications of Machine Learning in data intensive environments

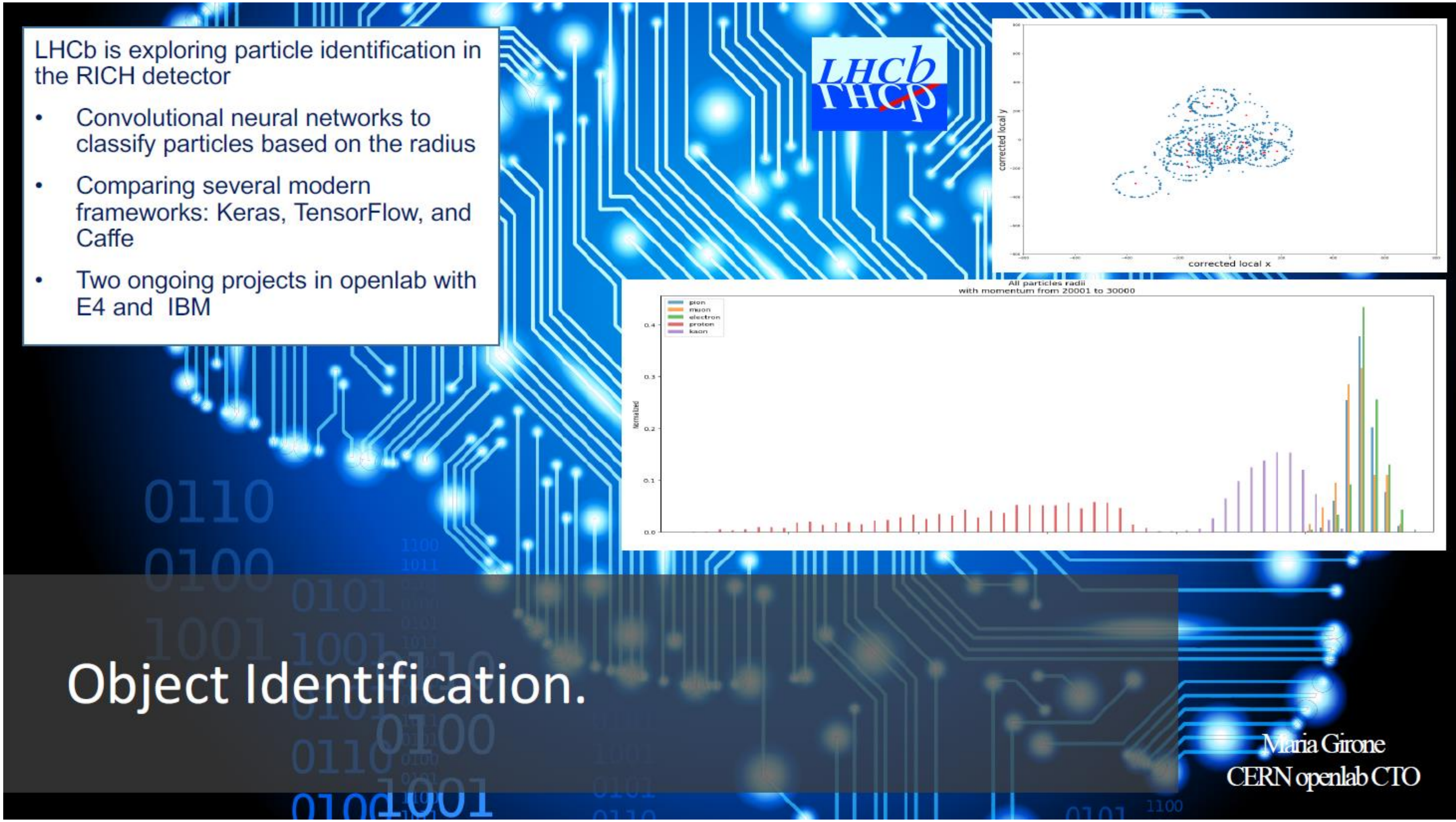
- Using an FPGA based co-processor from Micron
  - In CMS show that ML techniques can be applied in the Level-1 trigger. Complex decisions very close to real time (micro-seconds)
  - In DUNE the challenge is application in large data volumes
    - events are 5GB and the data rate is 5TB/s




DEEP UNDERGROUND NEUTRINO EXPERIMENT



# Investigating Accelerated Architectures for real-time selection and image recognition.


LHCb is exploring particle identification in the RICH detector
• Convolutional neural networks to classify particles based on the radius
• Comparing several modern frameworks: Keras, TensorFlow, and Caffe
• Two ongoing projects in openlab with E4 and IBM
LHCb
corrected local y
corrected local x
All particles radii
with momentum from 20001 to 30000
pion
muon
electron
proton
kaon
Object Identification.
Maria Girone
CERN openlab CTO

# Deep Learning for the future of the Large Hadron Collider

Maurizio Pierini
CERN

European Research Council

# A Topology Classifier

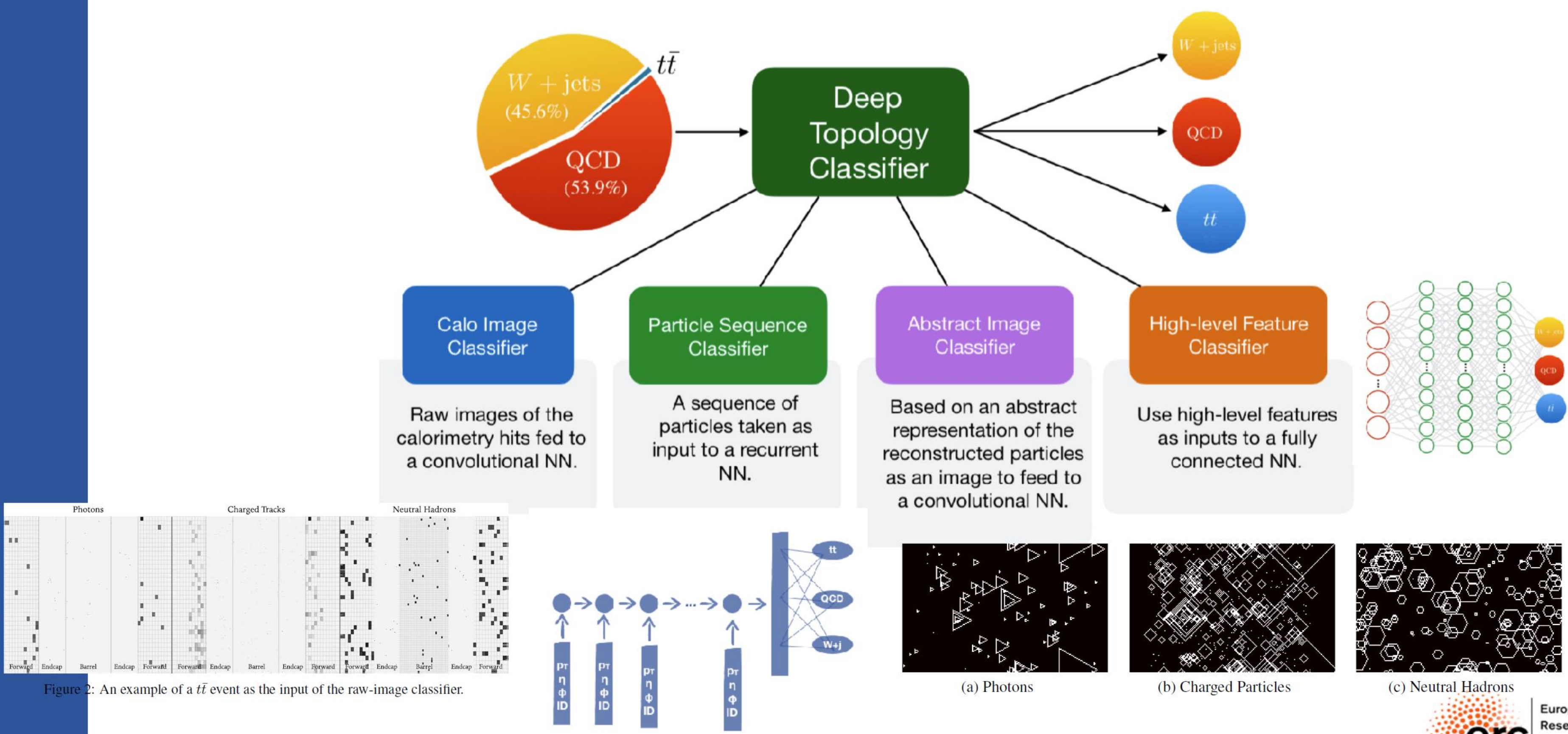


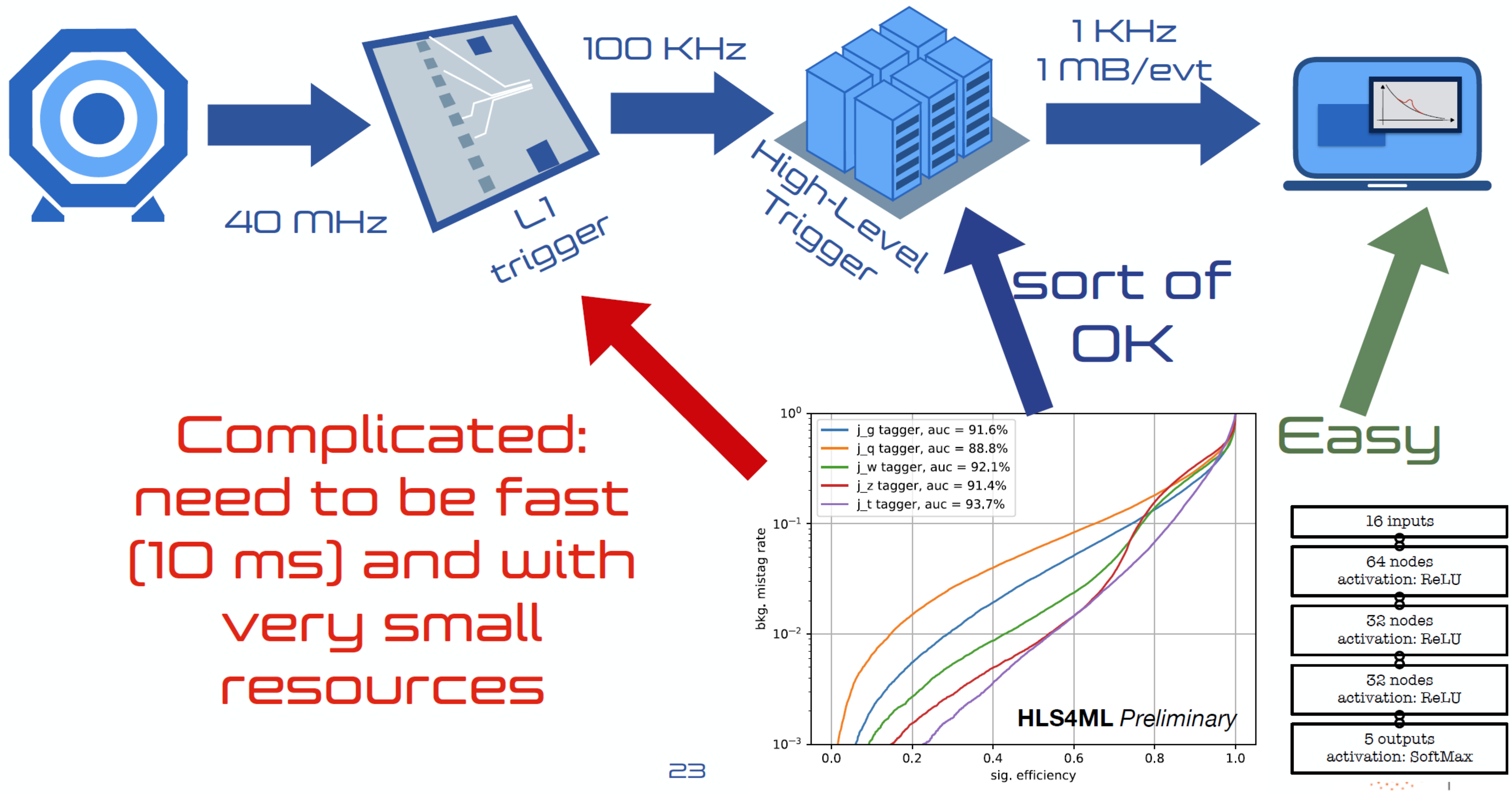

Figure 2: An example of a $t\bar{t}$ event as the input of the raw-image classifier.

(a) Photons (b) Charged Particles (c) Neutral Hadrons

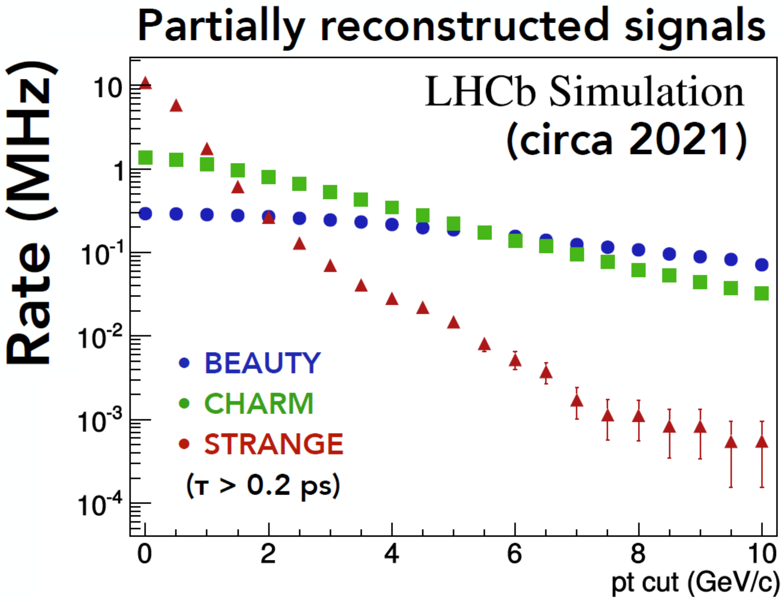
19

European Research Council

# The LHC Big Data Problem



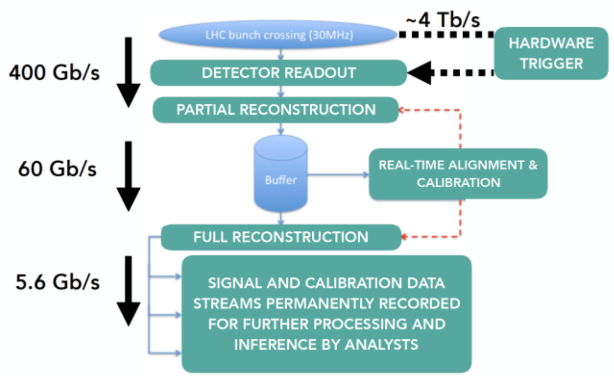
# Too much of a good thing
# (how to drink New Physics from a 40 Tb/s firehose)


Vladimir V. Gligorov, CNRS/LPNHE
On behalf of the LHCb collaboration

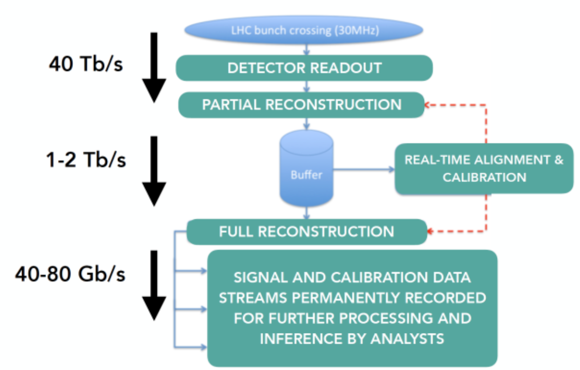
AI @ CERN and SKA, Alan Turing Institute, London 17.09.2018

## A lot of signal ⇒ a lot of data to process!



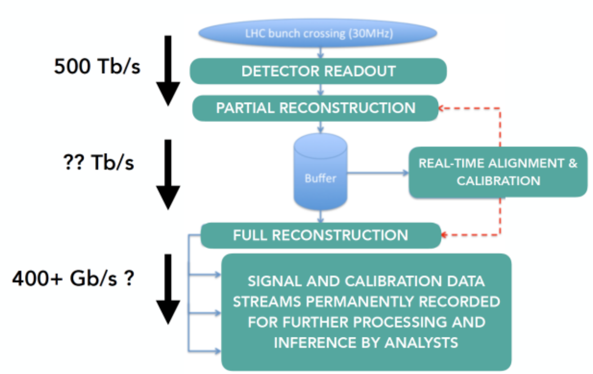

**LHCb real-time data processing today**

## And this data volume will only rise…




**LHCb real-time data processing circa 2021**

## …and rise again




**LHCb real-time data processing circa 2032**

# Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders

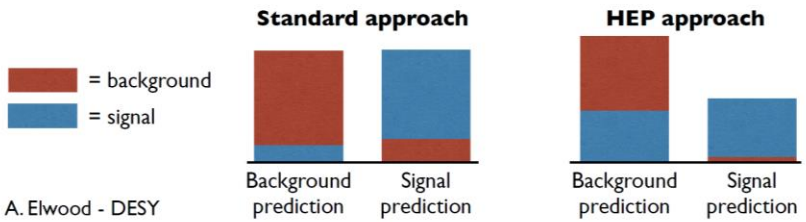
Adam Elwood and Dirk Krücker

adam.elwood@desy.de

1

# A HEP search approach

- Searches for new physics can be framed as signal/background classification problems
- Typical machine learning approach to classification is to optimise accuracy or area under ROC curve
  - Correct classification of signal or background has the same weight
- When searching for new physics actually care about statistical significance of signal counts over background counts
  - Don't care about the purity of the background classification (particularly when there are large systematic uncertainties)






# Loss functions to directly optimise significance

- To train neural networks that take account of this, can design a loss function based around the direct optimisation of the significance
  - Try to obtain the optimal bin to count signal and background events
  - For a batch of N training events:
    - ***s*** = (N correctly classified signal events)
    - ***b*** = (N incorrectly classified background events)
  - Maximise standard estimate of exclusion significance based on gaussian statistical uncertainties:

$$s/\sqrt{s+b}$$

  - Maximise Asimov estimate of discovery significance (including systematic uncertainty):

$$Z_A = \left[2\left((s+b)\ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{1/2}. \quad *$$

H2020-Astronomy ESFRI and Research Infrastructure Cluster
(Grant Agreement number: 653477).

# Accelerating Science by Repurposing Machine Learning Software

Bojan Nikolic

ATI London – 18 September 2018

## PyTorch

PyTorch

*Tensors and Dynamic neural networks in Python with strong GPU acceleration*

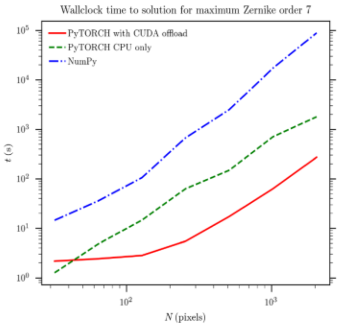
Installation:

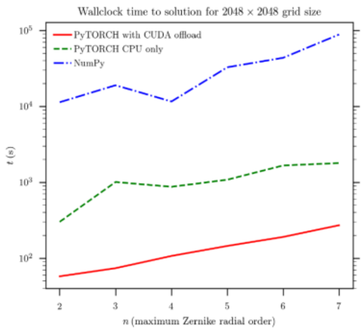
**conda install pytorch cuda91 –c pytorch**

Automatic differentiation
Trivially easy to offload to GPUs:

UNIVERSITY OF CAMBRIDGE REPURPOSING MACHINE LEARNING SOFTWARE London 18 September 2018 17

## Performance comparison




UNIVERSITY OF CAMBRIDGE REPURPOSING MACHINE LEARNING SOFTWARE London 18 September 2018 25

## Summary

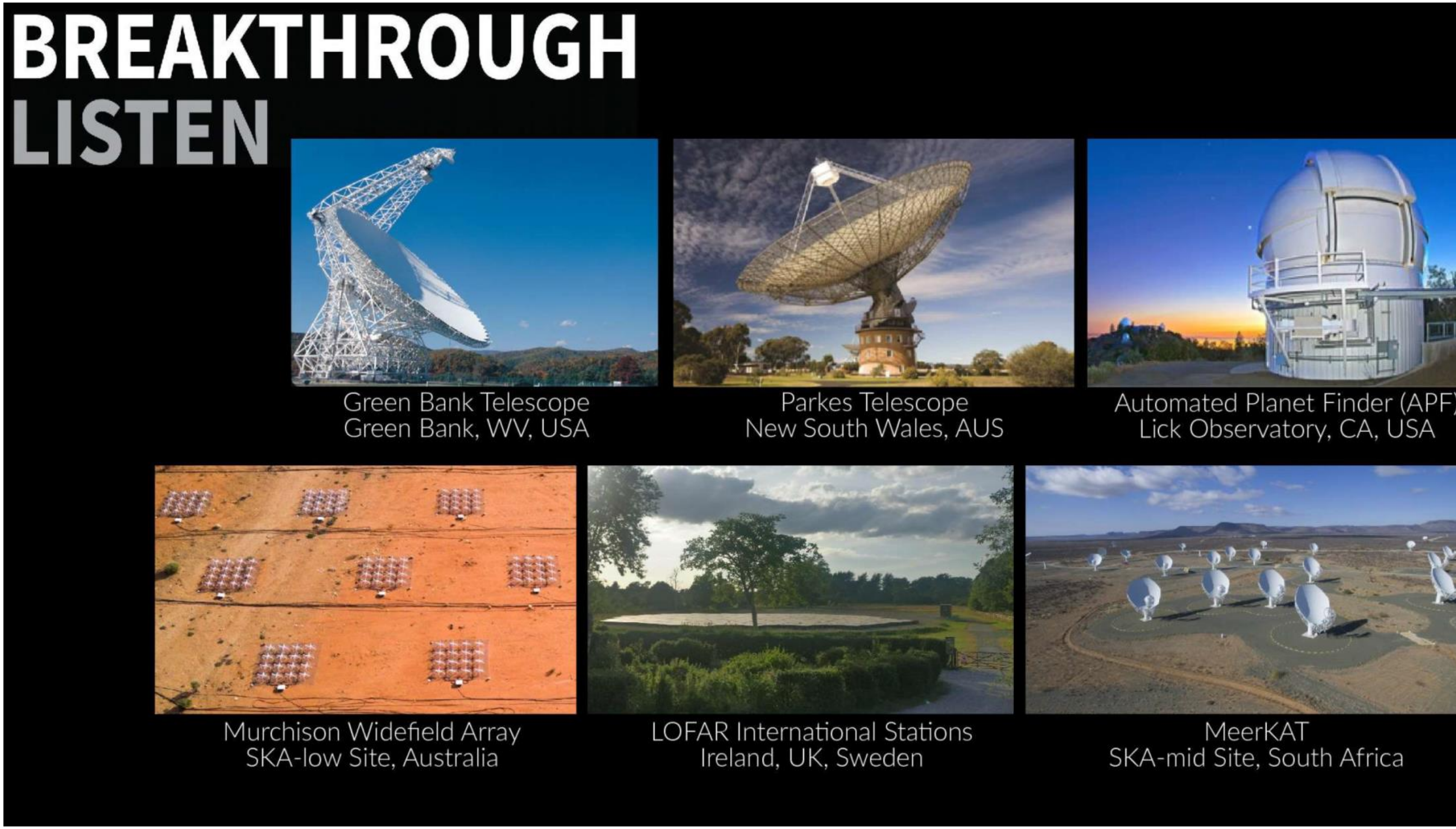
- >100x performance improvement in minimising functions
- Small, contained, software effort needed
  - Perfect integration with standard Python environment
- Out-of-box support for GPUs and multi-threaded CPUs
- Easy to use (& install!)
- More details: arXiv:1805.07439

UNIVERSITY OF CAMBRIDGE REPURPOSING MACHINE LEARNING SOFTWARE London 18 September 2018 27

# ML in Breakthrough Listen

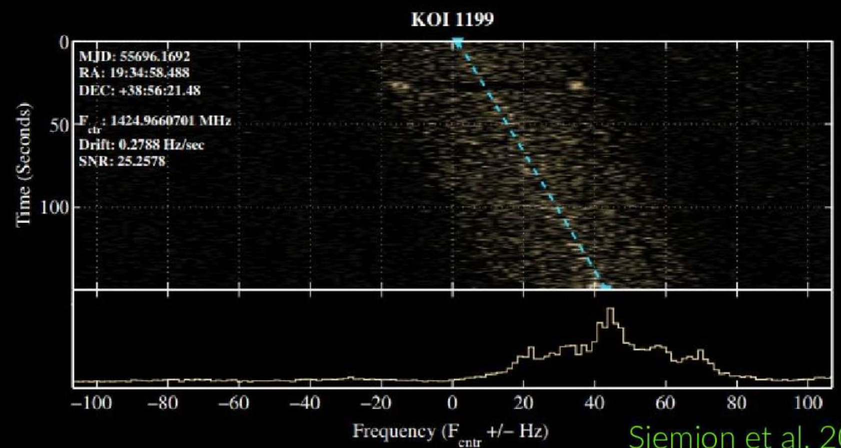
## Building Search Pipelines and Labelled Datasets for Transient Discovery

Griffin Foster
University of Oxford, Department of Physics
University of California at Berkeley, Department of Astronomy


BREAKTHROUGH
LISTEN
Green Bank Telescope
Green Bank, WV, USA
Parkes Telescope
New South Wales, AUS
Automated Planet Finder (APF)
Lick Observatory, CA, USA
Murchison Widefield Array
SKA-low Site, Australia
LOFAR International Stations
Ireland, UK, Sweden
MeerKAT
SKA-mid Site, South Africa

Doppler Drift Search
(typically ~1 Hz resolution)

$$\dot{f} = \frac{d\vec{V}}{dt}\frac{f_{\mathrm{rest}}}{c}$$




Siemion et al. 2013

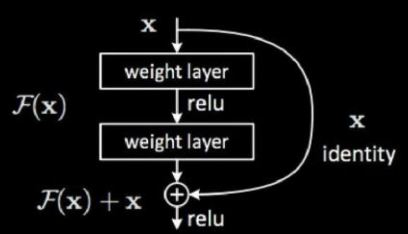
**The signal is unknown (anomaly detection problem)**

**Radio receiver noise is not Gaussian**

**RFI -dominated false detections**

**We want to find transforms to maximized S/N**

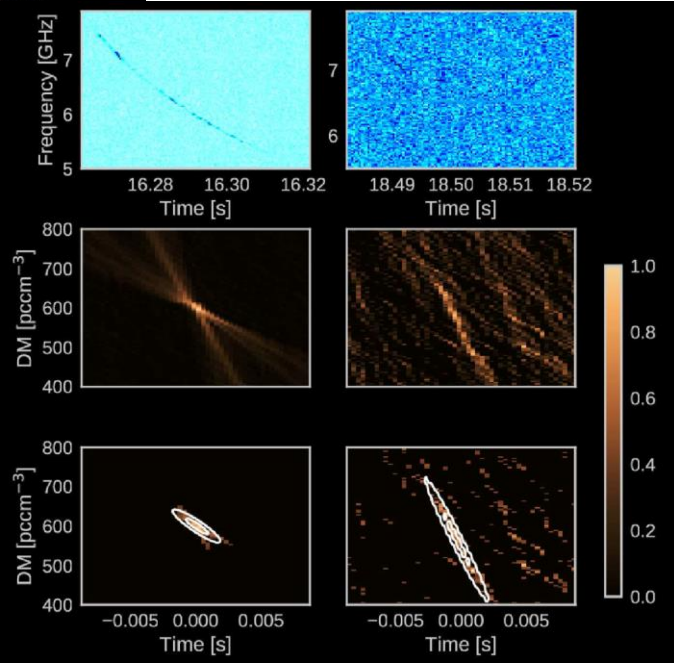
# Deep Feature Model: Residual Network




ResNets (He et al. 2015)

FRB CNN ResNet Model

| Group Name | Output Size | Stack Type |
|---|---|---|
| conv0 | $342 \times 256 \times 1$ | $[32 \times 1] \times 1$ |
| conv1 | $171 \times 128 \times 32$ | $[7 \times 7] \times 1$ |
| conv2 | $42 \times 32 \times 32$ | $[3 \times 3] \times 2$ |
| conv3 | $10 \times 8 \times 64$ | $[3 \times 3] \times 3$ |
| conv4 | $5 \times 4 \times 128$ | $[3 \times 3] \times 2$ |
| avg-pool | | |
| fc | | |



# AstroStatistics & AstroInformatics

## in the context of the SKA and LSST

Jason McEwen
www.jasonmcewen.org
@jasonmcewen

*Mullard Space Science Laboratory (MSSL)*
*University College London (UCL)*

AI for CERN and SKA, Alan Turing Institute
September 2018

## Deep learning methods for radio interferometric imaging

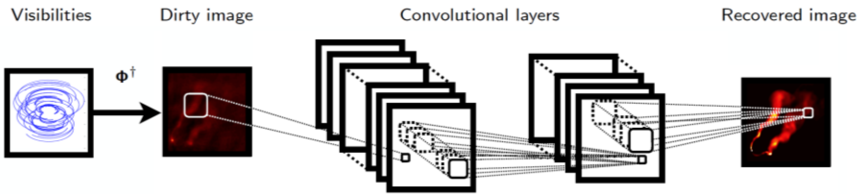



Figure: Deep learning architecture for interferometric imaging (Allam & McEwen, in prep.)

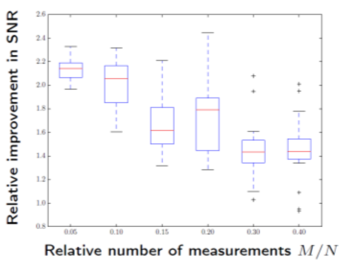



Figure: Improvement in signal-to-noise-ratio (SNR)



## Outline

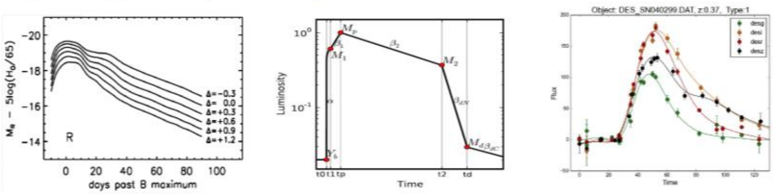
1. Distributed and parallelised algorithms
2. Online algorithms
3. Uncertainty quantification
4. Machine learning



## Supernova classification

### Photometric classification

- Photometric Supernova classification by machine learning (Lochner, McEwen, Peiris, Lahav & Winter 2016)
- Limited training data.
- Go beyond single techniques to study classes.




(a) Templates (b) Generic parameterisations (c) Wavelets (non-parametric)

Figure: Feature selection classes (in order of increasing model independence)

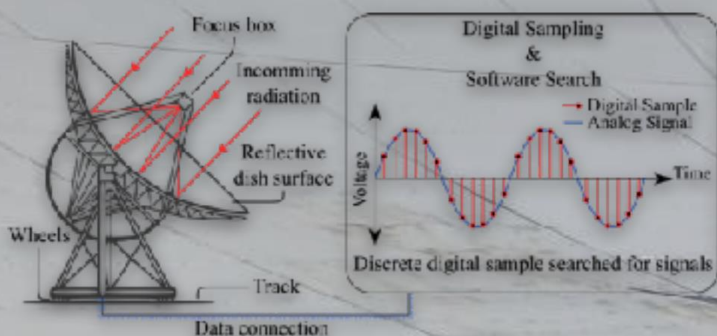
- Integrate physics into machine learning (scale and dilation invariance).
- Understand physical requirements: representative training, redshift.


MANCHESTER
1824
The University of Manchester
Time-domain Machine Learning - Opportunities and Challenges for the SKA
Rob Lyon
University of Manchester (SKA Group)
@scienceguyrob
robert.lyon@manchester.ac.uk

MANCHESTER
1824
The University of Manchester

# Finding Pulsars

- Analog to digital samples.
- Complex search pipeline applied (RFI removal, sifting…).
- Identifies significant signal detections.
- Detections recorded as “candidates”.
- Candidates are summary detections that must be filtered.




1 stream per beam

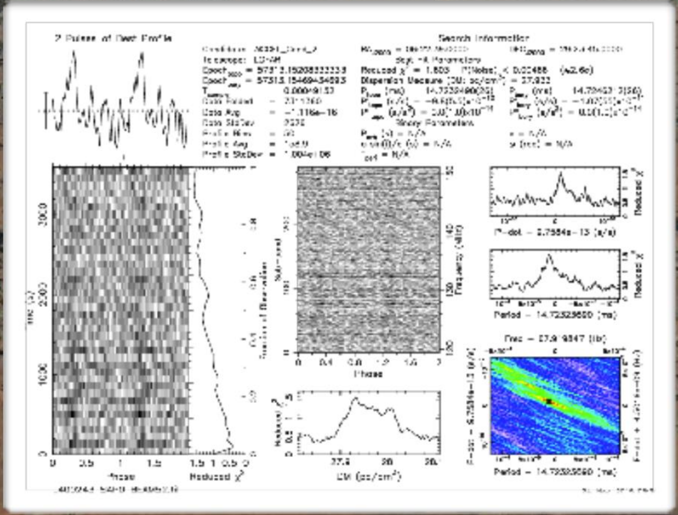
AI at CERN and SKA Sept. 17th-18th 2018 8/35

Credit: SKAO

MANCHESTER 1824
The University of Manchester

# Candidate Volumes - Increasing!

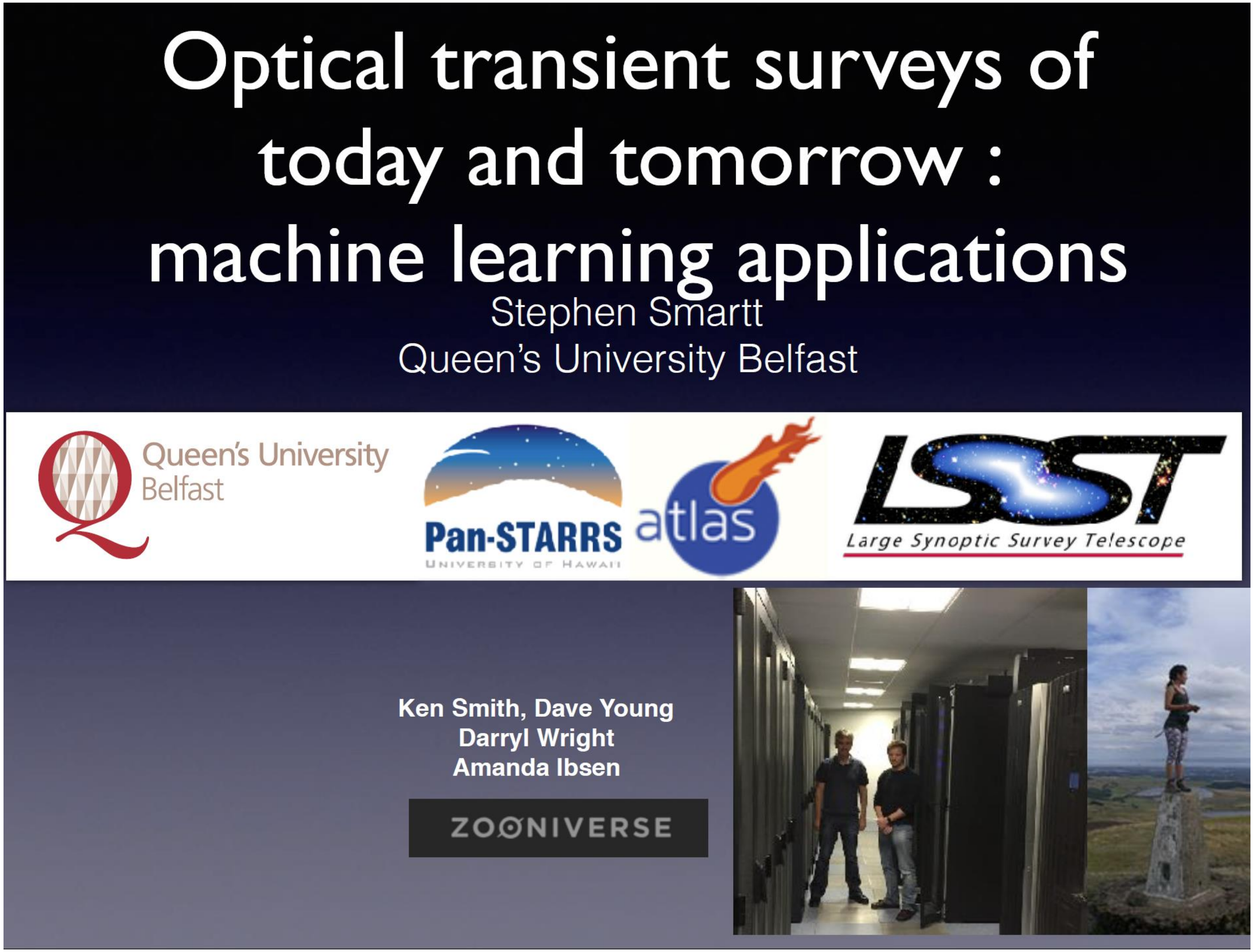
- 1,500 search beams for SKA Mid. Each producing 1,000 candidates per scan.
- 1.5 million candidates per scan.
- A scan can be as short as 180 seconds = 48 per day.
- That's 72 million per day.
- There are >10,000 non-target examples for each interesting candidate.
- At best 150 good candidates per beam, per scan. That's 7,200 per day.

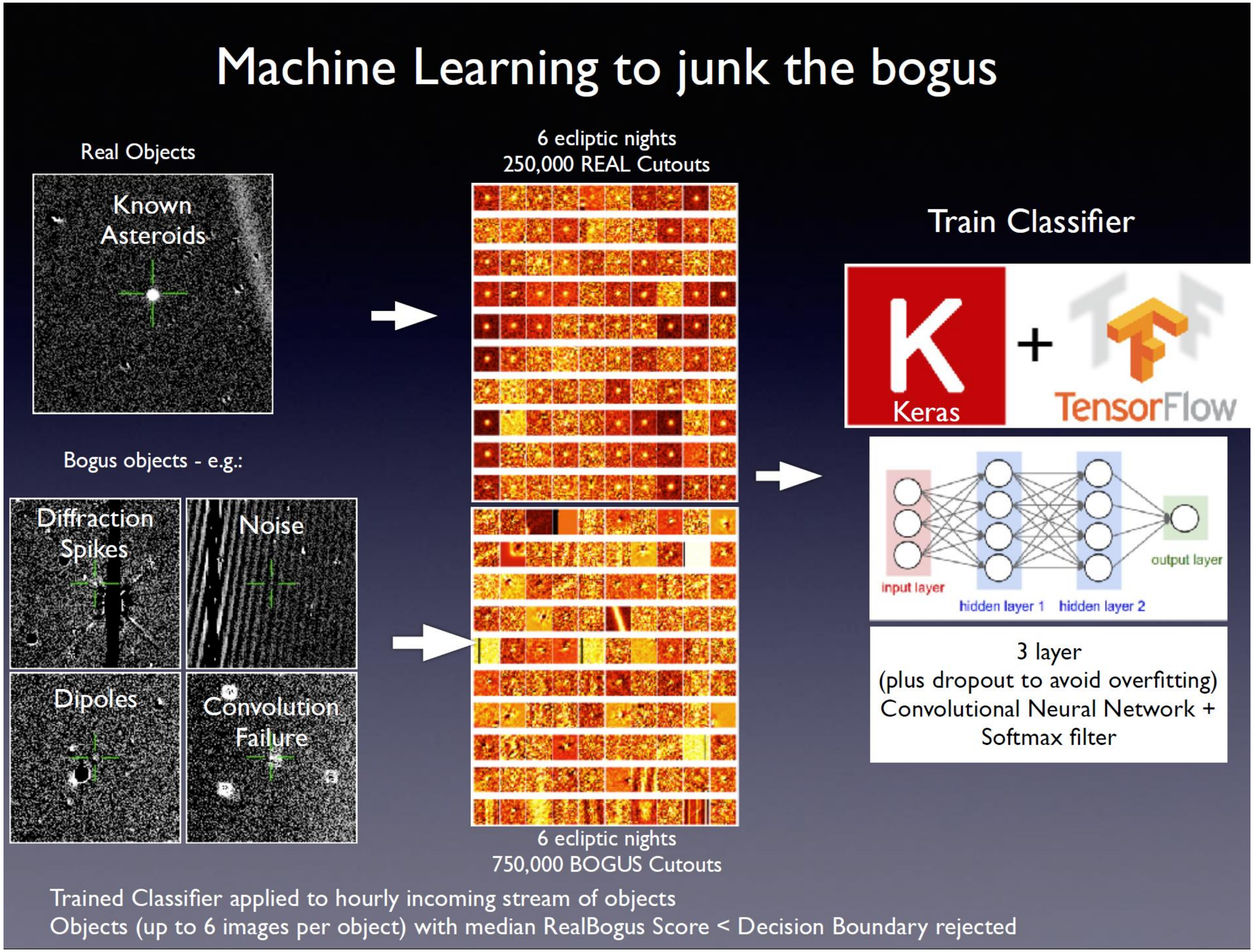


Credit: LOTAAS Collaboration (Chia Min Tan et. al.).

AI at CERN and SKA    Sept. 17th-18th 2018    24/35

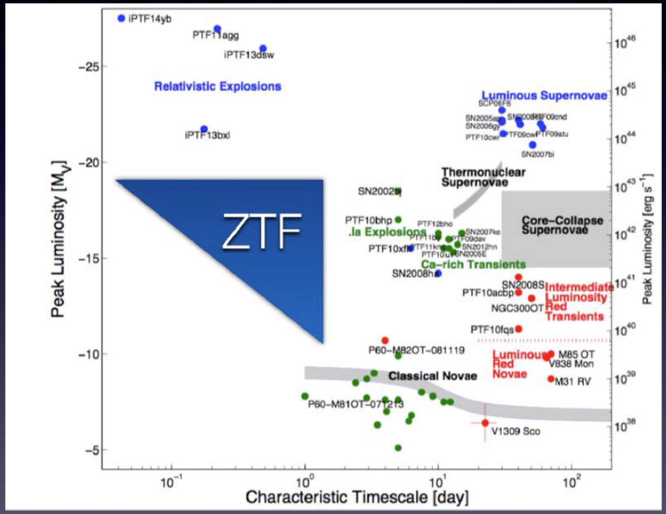

MANCHESTER
1824
The University of Manchester
Recent Efforts (
• On-line classifiers designed to address real-time issues.
• Classifiers designed to overcome class imbalance.
• Deep Neural Networks (CNNs, GANs) applied to the classification problem with some success.
• Gold-standard data sets being obtained.
• Gold-standard test-vector generators now exist.
• Today performance is very good, but the false-positive rate remains too high for SKA processing scenarios…
AI at CERN and SKA
Sept. 17th-18th 2018
28/35

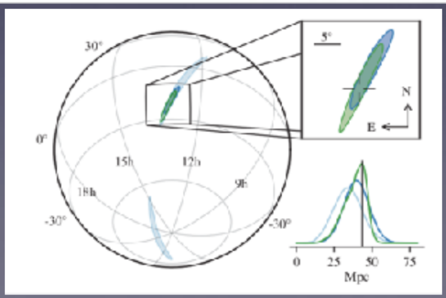

Optical transient surveys of today and tomorrow : machine learning applications
Stephen Smartt
Queen's University Belfast
Queen's University Belfast
Pan-STARRS
University of Hawaii
atlas
LSST
Large Synoptic Survey Telescope
Ken Smith, Dave Young
Darryl Wright
Amanda Ibsen
ZOONIVERSE


Machine Learning to junk the bogus
Real Objects
Known Asteroids
Bogus objects - e.g.:
Diffraction Spikes
Noise
Dipoles
Convolution Failure
6 ecliptic nights
250,000 REAL Cutouts
6 ecliptic nights
750,000 BOGUS Cutouts
Train Classifier
Keras
+
TensorFlow
input layer
hidden layer 1
hidden layer 2
output layer
3 layer
(plus dropout to avoid overfitting)
Convolutional Neural Network +
Softmax filter
Trained Classifier applied to hourly incoming stream of objects
Objects (up to 6 images per object) with median RealBogus Score < Decision Boundary rejected

# Summary : scratching the surface of potential for LSST

- Multiple machine learning to provide probability of physical classification
- Identify the outliers
- Real-time, every day - ability to trigger rapid follow-up
- Cross match with radio, x-ray, gamma-ray surveys, LIGO-Virgo sky localisation maps for GW
- Enormous discovery potential in data, but completeness and probabilistic approach essential
- **Understand the population - repeatability and completeness**




GW170817

# Classifying radio galaxies with deep learning

Vesna Lukic, Marcus Brüggen
University of Hamburg
Beatriz Mingo, Judith Croston
The Open University, UK

## Current work

- Sources from the LOFAR catalogue of sky survey in 120-240MHz, eventually all sky
  - A new window into the universe
  - > 100 million galaxies

## Conclusions

- Machine learning is essential in analysing data from future astronomical surveys
- Previous work focused on using convolutional neural networks
  - Our first work used PyBDSF to characterise sources into different # components
- Current work shows performance of convolutional network surpasses that of capsule network models
- Sources have varying levels of noise as well as potential intruding sources

## Current work

### Morphological classification for LOFAR surveys: Capsule Networks versus Convolutional Neural Network

V. Lukic,[1*] M. Brüggen,[1†] B. Mingo,[2] J. Croston[2]
[1] Hamburger Sternwarte, University of Hamburg, Gojenbergsweg 112, Hamburg 21029, Germany
[2] School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK

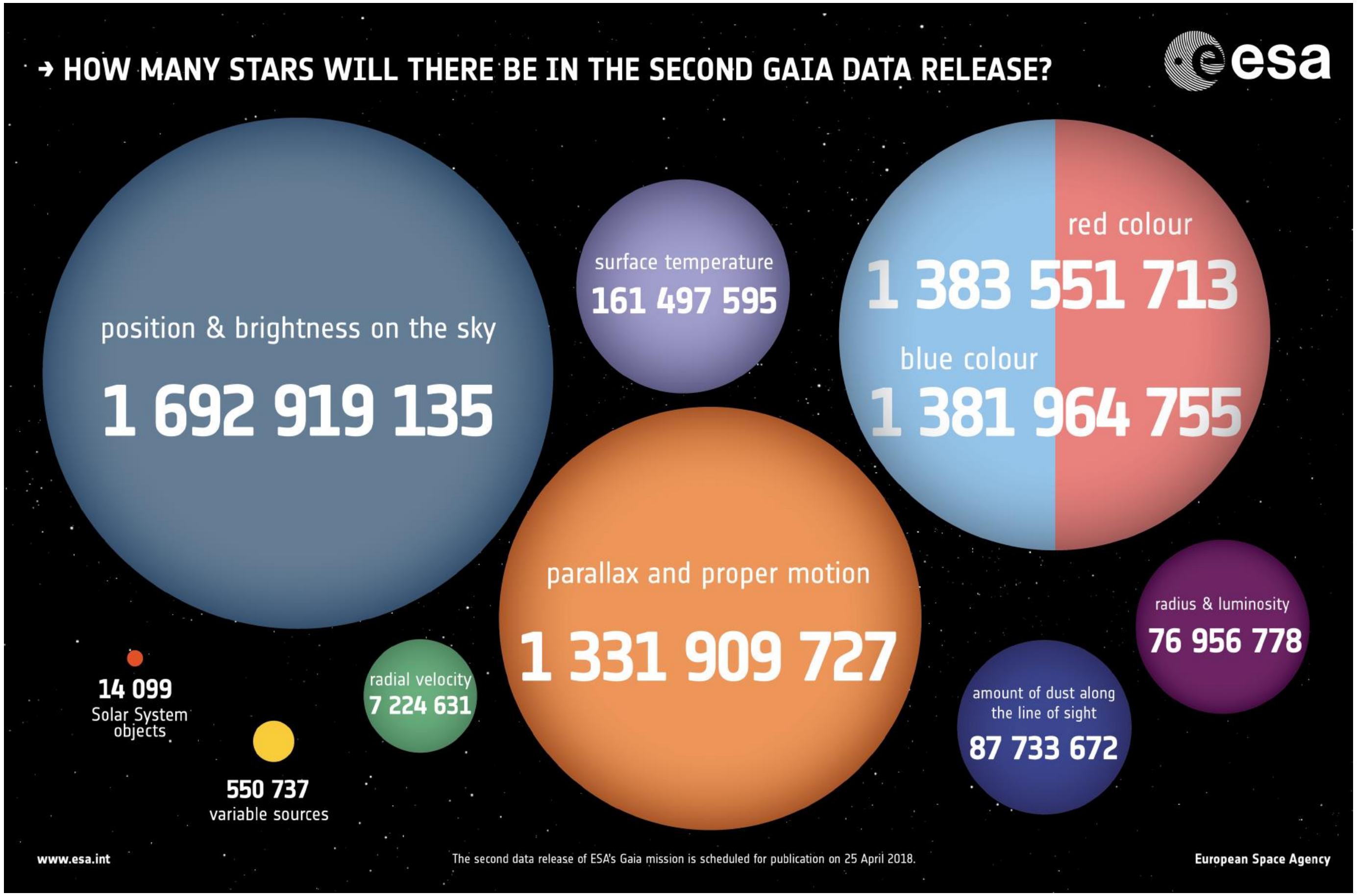
Accepted XXX. Received YYY; in original form ZZZ

**ABSTRACT**
Next generation radio surveys will yield an unprecedented amount of data, warranting analysis by use of machine learning techniques. Convolutional neural networks are the deep learning technique that has proven to be the most successful in classifying image data. However, they are limited in their capacity to model hierarchical information in an image and are not rotationally invariant. Capsule networks have been developed to address these issues by using capsules comprised of groups of neurons, that describe properties of an image including the relative spatial locations of features. We utilise images from the LOFAR galaxy zoo, which has attained higher resolution images revealing richer and more complex morphologies compared with previous surveys. The increased complexity results in a broader variety of morphology within each class, as well as added noise and increased potential contamination with intruding sources, presenting further challenges for machine learning algorithms. The current work explores the performance of different capsule network architectures against a standard convolutional neural network consisting of 8 convolutional layers, in reproducing the classifications into three classes of data. We obtain an overall precision of 92.8% and 85.1% using the convolutional network and default capsule network architecture respectively, when training on the original and augmented images. The convolutional network almost always outperforms any variation of the capsule network, as it proves to be more robust to noisy images and the pooling operation may not cause such a significant diminishment in performance as radio galaxy morphologies display some amount of intra-class variability.

**Key words:** Astronomical instrumentation, methods, and techniques; radio continuum: galaxies

# Some ML and AI challenges in current and future optical and near infra imaging datasets

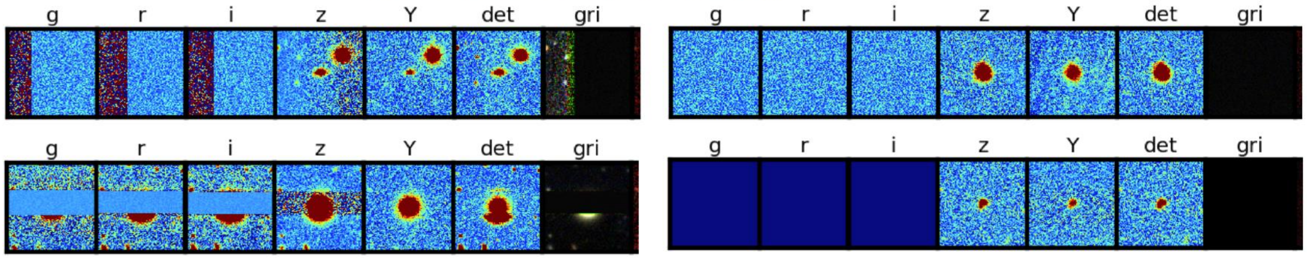
Richard McMahon (Institute of Astronomy,
University of Cambridge)
and
Cameron Lemon, Estelle Pons, Fernanda Ostrovski, Matt Auger, Manda Banerji, Vidhi Lalchand


→ HOW MANY STARS WILL THERE BE IN THE SECOND GAIA DATA RELEASE?
esa
position & brightness on the sky
1 692 919 135
surface temperature
161 497 595
red colour
1 383 551 713
blue colour
1 381 964 755
parallax and proper motion
1 331 909 727
radius & luminosity
76 956 778
amount of dust along the line of sight
87 733 672
radial velocity
7 224 631
14 099
Solar System objects
550 737
variable sources
www.esa.int
The second data release of ESA's Gaia mission is scheduled for publication on 25 April 2018.
European Space Agency

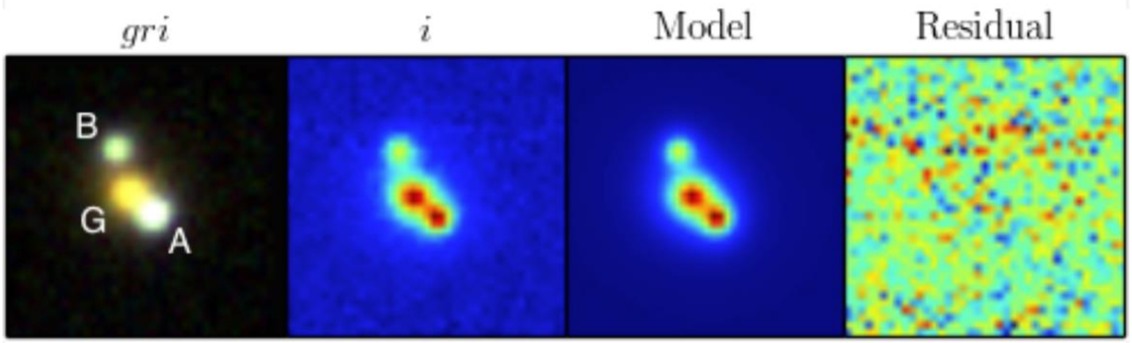
# Example of ‘junk’ issues with imaging Data

- XXL-SDSS-DES sample = 1497 sources
  - But 13% of them (197 sources) do not have a DES $i_{psf}$ magnitude

➡ Not due to a non-detection

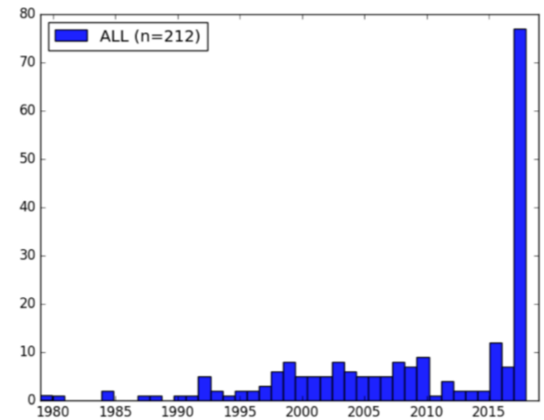


**VDES J2325−5229 a z = 2.7 gravitationally lensed quasar discovered using morphology-independent supervised machine learning: GMM**




J2325−5229 as a g, r, and i DES Y1 colour composite, an i-band image, an i-band image model and the residuals from subtracting the model from the image. All cutouts are 10.0 arcsec in size. North is up and East is left.

Rate of discovery of lensed quasars



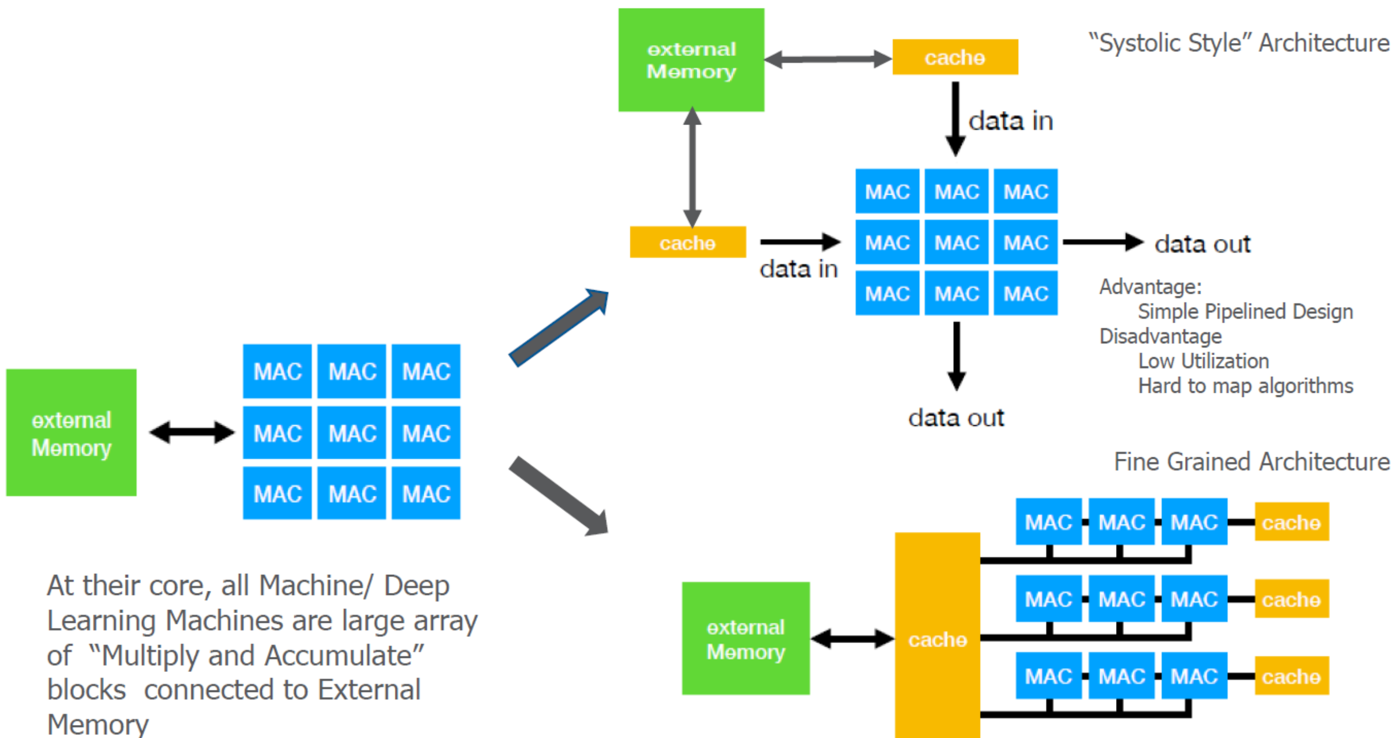
# Evolving Compute/Memory Analytics Architectures

Steve Pawlowski

Vice President, Advanced Computing Solutions

September 17, 2018

# Neural Network HW ➡ JBOM – " Just a Bunch of MAC's/ External Memory and on-chip cache"



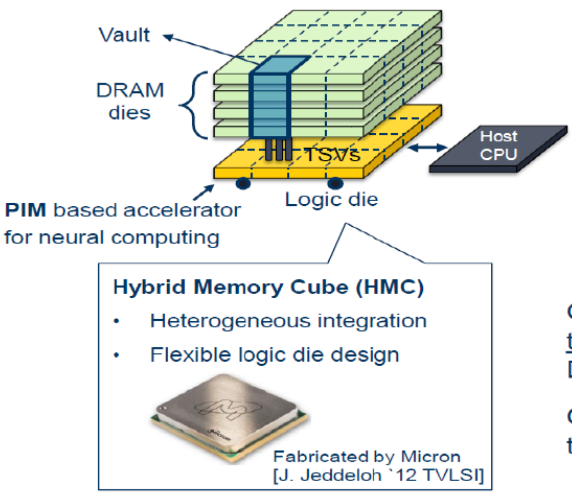

At their core, all Machine/ Deep Learning Machines are large array of "Multiply and Accumulate" blocks connected to External Memory

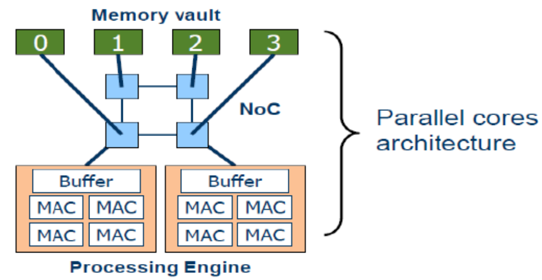
Advantage:
High Utilization/Efficiency
Disadvantage
Complex HW/Compiler

Source: FWDNXT Inc.

# Looking Forward –
# Stacking Memory on top of the ML Compute fabric, we can get high bandwidth, low energy and…yes…capacity.

**Programmable, scalable platform as processor in memory**




Q1. Neural computing layer should meet <u>thermal and area constraint</u> in 3D stacked DRAM

Q2. NeuroCube should be programmable to cover <u>different types of neural network</u>

[Kim et al., NeuroCube, ISCA 2016]

***"Combining memory and processing resources in a single device has huge potential to increase the performance and efficiency of DNNs as well as others forms of machine learning systems. It is possible to make a trade off between memory and compute resources to achieve a different balance of capability and performance in a system that can be generally useful across all problem sets."***

- https://www.graphcore.ai/blog/why-is-so-much-memory-needed-for-deep-neural-networks

## How can you implement Deep Learning now…

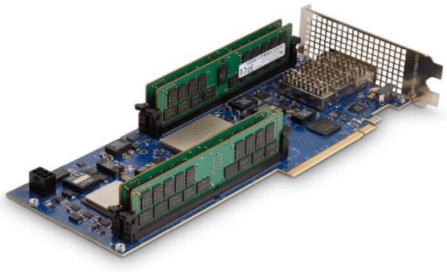
**Deep Learning processor**


- High Performance Memory
- Best performance per power
- Best utilization
- Efficient use of memory bandwidth
- Low latency
- Scalability: IoT to cloud

25 | 9/17/2018 Micron

## For larger networks/larger memory capacity

| FPGA board | Micron SB-852 / AC511 |
|---|---|
| Accelerator cores | 1024 MAC units @ 250 MHz |
| Peak Throughput | 512 G-ops/s |
| Architecture | Fined Grain Cached Architecture |
| Memory | DDR4 and High Perf Memory |
| Memory B/W | 120 GB/s |
| Power (Board) | 48 W |



26 | 9/17/2018 Micron

# Conclusion

- AI/ML/NN techniques are widely used in HEP, astrophysics and radio astronomy
- A collaboration with industrial partners is quite common
- Specialised hardware is under development
- There are many areas for collaboration in SKA and CERN data analysis related to ML in the frames of the OpenLAB.

# Acknowledgement

- H2020-Astronomy ESFRI and Research Infrastructure Cluster (Grant Agreement number: 653477).