AI at CERN and SKA

Report of Contributions

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

Contribution ID: **2** Type: **not specified**

opening words

Monday 17 September 2018 10:00 (5 minutes)

Presenter: Dr PETER, Braam

Session Classification: Day 1

Contribution ID: **3** Type: **not specified**

Partnering with industry for machine learning at HL-LHC

Monday 17 September 2018 10:05 (30 minutes)

The High Luminosity LHC (HL-LHC) represents an unprecedented computing challenge. For the program to succeed the current estimates from the LHC experiments for the amount of processing and storage required are roughly 50 times more than are currently deployed. Although some of the increased capacity will be provided by technology improvements over time, the computing budget is expected to be flat and to close the gap huge gains in the efficiency for processing and analyzing events must be achieved. An area that has the potential for a significant breakthrough is Machine Learning. In recent years industry has invested heavily in both hardware and software to develop machine learning techniques to filter, process, analyze, and derive correlations from very large scale heterogeneous datasets. Through CERN openlab, with industry partners, and the R&D projects of the LHC experiments we are attempting to build on the industry investments to utilize these techniques for science. In this presentation we will discuss the activities of the CERN openlab industry partnerships in machine learning and how they can be extended to science applications.

Maria has a PhD in particle physics. She also has extensive knowledge in computing for high-energy physics experiments, having worked in scientific computing since 2002.Maria has worked for many years on the development and deployment of services and tools for the Worldwide LHC Computing Grid (WLCG), the global grid computing system used to store, distribute, and analyse the data produced by the experiments on the Large Hadron Collider (LHC).Maria was the founder of the WLCG operations coordination team, which she also previously led. This team is responsible for overseeing core operations and commissioning new services. Throughout 2014 and 2015, Maria was the software and computing coordinator for one of the four main LHC experiments, called CMS. She was responsible for about seventy computing centres on five continents, and managed a distributed team of several hundred people.Prior to joining CERN, Maria was a Marie Curie fellow and research associate at Imperial College London. She worked on hardware development and data analysis for another of the LHC experiments, called LHCb —as well as for an experiment called ALEPH, built on the accelerator that preceded the LHC.

Presenter: Dr GIRONE, Maria (CTO CERN OpenLab)

Session Classification: Day 1

Contribution ID: **4** Type: **not specified**

Emerging Use of DNNs

Monday 17 September 2018 10:35 (30 minutes)

Abstract: Data is being created at an alarming rate and we need faster and more efficient machines and algorithms to make sense of this data. Though we will still need the performance of traditional high performance computing, there are characteristics and relationships in the data that are needing more non-traditional computing approaches for greater efficiency. This need is also coupled with the fact that, with the slowing of Moore's Law, such new architectures not only have to be faster but need to compensate the changes in device reliability. This talk will focus on the emerging use of Deep Neural Networks (DNN's), their architecture, their memory requirements and the need to understand the application needs in order to improve the performance of such systems.

Bio: Steve Pawlowski is advanced computing solutions vice president at Micron Technology. He is responsible for defining and developing innovative memory solutions for the enterprise and high-performance computing markets. Prior to joining Micron in July 2014, Mr. Pawlowski was a senior fellow and the chief technology officer for Intel's Data Center and Connected Systems Group. Mr. Pawlowski's extensive industry experience includes 31 years at Intel, where he held several high-level positions and led teams in the design and development of next-generation system architectures and computing platforms. Mr. Pawlowski earned bachelor's degrees in electrical engineering and computer systems engineering technology from the Oregon Institute of Technology and a master's degree in computer science and engineering from the Oregon Graduate Institute. He also holds 58 patents.

Presenter: Mr PAWLOWSKI, Stephen (VP of Advanced Computing Solutions, Micron)

Session Classification: Day 1

Contribution ID: **5** Type: **not specified**

Learning Deep Generative Models of Graphs

Monday 17 September 2018 11:05 (30 minutes)

Abstract: Graphs are fundamental data structures which concisely capture the relational structure in many important real-world domains, such as knowledge graphs, physical and social interactions, language, and chemistry. Here we introduce a powerful new approach for learning generative models over graphs, which can capture both their structure and attributes. Our approach uses graph neural networks to express probabilistic dependencies among a graph's nodes and edges, and can, in principle, learn distributions over any arbitrary graph. In a series of experiments our results show that once trained, our models can generate good quality samples of both synthetic graphs as well as real molecular graphs, both unconditionally and conditioned on data. Compared to baselines that do not use graph-structured representations, our models often perform far better. We also explore key challenges of learning generative models of graphs, such as how to handle symmetries and ordering of elements during the graph generation process, and offer possible solutions. Our work is the first general approach for learning generative models over arbitrary graphs, and opens new directions for moving away from restrictions of vector- and sequence-like knowledge representations, toward more expressive and flexible relational data structures.

Bio: Yujia Li is a research scientist at DeepMind. He obtained his Ph.D. at University of Toronto, studying machine learning and deep neural networks. His current research focuses on developing structured neural models and solving problems on structured data. He also has experience in other aspects of machine learning, including structured prediction models, generative models, semi-supervised learning and application domains including computer vision and temporal signal processing.

Presenter: Dr LI, Yujia (DeepMind)

Session Classification: Day 1

Contribution ID: **6** Type: **not specified**

Infrastructure for ML panel

Monday 17 September 2018 11:50 (1h 10m)

Moderator: Dr Peter Braam
Panelists: Dr Maria Girone, Dr Yujia Li, Dr Bojan Nikolic, Stephen Pawlowski

Session Classification: Day 1

Contribution ID: **7** Type: **not specified**

Deep Learning for Science: Opening the Pandora Box

Monday 17 September 2018 14:00 (30 minutes)

Dr. Shirley Ho, Group Leader, Flatiron Institute, Cosmology X Data Science, CCA

Abstract: Scientists have always attempted to identify and document analytic laws that underlie physical phenomena in nature. The process of finding natural laws has always been a challenge that requires not only experimental data, but also theoretical intuition. Often times, these fundamental physical laws are derived from many years of hard work over many generation of scientists. Automated techniques for generating, collecting, and storing data have become increasingly precise and powerful, but automated discovery of natural laws in the form of analytical laws or mathematical symmetries have so far been elusive. Over the past few years, the application of deep learning to domain sciences –from biology to chemistry and physics is raising the exciting possibility of a data-driven approach to automated science, that makes laborious hand-coding of semantics and instructions that is still necessary in most disciplines seemingly irrelevant. The opaque nature of deep models, however, poses a major challenge. For instance, while several recent works have successfully designed deep models of physical phenomena, the models do not give any insight into the underlying physical laws. This requirement for interpretability across a variety of domains, has received diverse responses. In this talk, I will present our analysis which suggests a surprising alignment between the representation in the scientific model and the one learned by the deep model.

Bio: Dr. Shirley Ho is Group Leader for Cosmology X Data Science at Flatiron Institute, Visiting Research Associate at Princeton University and Associate (Adjunct) Professor at Carnegie Mellon University. https://users.flatironinstitute.org/sho/public_www/index.

Presenter: Dr HO, Shirley

Session Classification: Day 1

Contribution ID: **8** Type: **not specified**

Two computing challenges for particle physics : the tracking challenge and event simulation with Generative Adversarial Networks

Monday 17 September 2018 14:30 (30 minutes)

Dr. David Rousseau, HEP Physicist (ATLAS software and Higgs Physics) at LAL, Orsay, France

Abstract: I will expand on two specific lines of effort to solve the computational challenge at the LHC. (i) LHC experiments need to reconstruct the trajectory of particles from the few precise measurements in the detector. One major process is to « connect the dots », that is associate together the points left by each particle. The complexity of the process is growing exponentially with the LHC luminosity, so that new algorithms are needed. The TrackML challenge is a two phases competition to tackle the issue: 100.000 points to be associated into 10.000 tracks in less than 100 seconds. The first phase (with no speed incentive) has run on Kaggle over the summer, while the second one (with a strong speed incentive) is just starting on Codalab. I will summarize the preliminary findings and perspective. (ii) The growing LHC luminosity also increases the need of very high statistics and accurate event simulations. About 200.000 processor cores world wide are crunching numbers continuously to deliver event simulations, within the current baseline technique (Geant4 like) which is to simulate particles one by one. The recent Generative Adversarial Network technique allows to train an algorithm to generate images similar to an input set, whether celebrity faces, hotel rooms ⋯or particle showers in a calorimeter or even full LHC events. Once trained, the speed gain is potentially several order of magnitude. I will report on several strategies short-cutting the baseline approach that have now passed the proof of concept stage.

Bio: I am senior scientist at LAL-Orsay. After many years at the forefront of software developments for the ATLAS (CERN) experiment until the Higgs boson discovery in 2012, I was looking for something different. A chance encounter at the cafeteria with a Machine Learning (what was that?) scientist decided it. With Higgs physics always on my mind, I organized the Higgs ML challenge in 2014, and now the tracking ML challenge. I co-coordinate the ATLAS ML forum. I'm keen on both riding and promoting the ML wave in particle physics and science in general.

Presenter: Dr ROUSSEAU, David

Session Classification: Day 1

Contribution ID: **9** Type: **not specified**

Too much of a good thing : how to drink new physics from a 40 Tb/s firehose

Monday 17 September 2018 15:15 (30 minutes)

Dr. Vladimir "Vava"Gligorov, Research Scientist at CNRS/LPNHE

Abstract: LHCb is one of the four major experiments at the Large Hadron Collider (LHC) at CERN. It searches for particles and forces beyond our current physics theories, in particular by making highly precise measurements of the properties of the particles produced in the LHC collisions. Doing so requires analyzing an enormous quantity of data in real time : 8 Tb/s today and rising to 40 Tb/s in 2021. In this talk I will describe LHCb's real-time data analysis in the context of ongoing developments in both computer hardware and AI algorithms, and give some perspectives on the likely future evolution of this real-time analysis.

Bio: I spent my student years being bothered by quantum nonlocality, but eventually discovered that not being able to do mathematics would prove to be a problem if I became an experimental physicist. Now I divide my time between thinking about the myriad contradictions in our theories of the microscopic and macroscopic universe, and building real-time analysis systems to help LHCb probe these contradictions to ever higher precisions. I am also involved in the International Masterclass program and with the work of the Petnica Science Center, trying to make the next generation as much as possible. Life is too short for most social media, but I do tweet @particleist not necessarily about science.

Presenter: Dr GLIGOROV, Vladimir "Vava"

Session Classification: Day 1

Contribution ID: **10** Type: **not specified**

Deep Learning for the future of the Large Hadron Collider

Monday 17 September 2018 15:45 (30 minutes)

Dr. Maurizio Pierini, CERN Physicist working on the CMS experiment at the Large Hadron Collider

Abstract: The High-Luminosity LHC phase, scheduled for 2025, will challenge the commonly accepted solutions for tasks such as event reconstruction, real-time processing, large dataset simulation, etc. Deep Learning is considered as a strong candidate to solve this issue, by speeding up the task executions with potentially no performance loss. We will discuss a few preliminary studies, showing potential use cases for techniques like computing vision of natural-language processing to LHC-related physics problems.

Bio: I am a CERN physicist, working on the CMS experiment at the Large Hadron Collider. My main areas of interests are the search for physics beyond the standard model. I have a consolidated experience with optimizing real-time data processing and selection. In the recent past, I have been working on investigating and promoting the usage of Deep Learning in particle physics.

Presenter: Dr PIERINI, Maurizio

Session Classification: Day 1

Contribution ID: **11** Type: **not specified**

Panel: ML Specific to HEP

Monday 17 September 2018 16:45 (1h 15m)

Moderator: Dr Alan Barr
Panelists: Dr. David Rousseau, Dr. Vladimir "Vava" Gligorov, Dr. Maurizio Pierini, Dr. Jennifer Thompson

Session Classification: Day 1

Contribution ID: **12** Type: **not specified**

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

Tuesday 18 September 2018 10:00 (30 minutes)

Professor Richard McMahon, Professor of Astronomy, Director and Head of Department, Institute of Astronomy, University of Cambridge

Abstract: I will present an overview of non-radio imaging current and future challenges based on current ground and space based optical and near infrared imaging surveys; DES/VISTA/Gaia -> Euclid/LSST. Current surveys are producing PB scale imaging datasets at a range wavelengths, depths and spatial resolution. 1-3 billion row source catalogues per survey with many thousands of columns which are overwhelming incoming graduate students and Post Docs who have not dealt with data on these scales before. I will use some examples of rare object searches (the most luminous super massive black holes in the Epoch of Reionization; Gravitionally lensed quasars for measuring the rate of expansion of the Universe and characterising Dark Matter and Dark Energy) and new phenomenon which get swamped by gaussian and non-gaussian outliers and instrumental artefacts. We are starting to explore ML and AI techniques at the catalogue and image level. I speak as a newcomer to ML and AI, but with 30 years of domain experience working on the large scale datasets and will highlight some of the domain level challenges we face and lessons already learnt in terms of dealing with bad data, incomplete and ambiguous meta-data.

Bio: Richard McMahon is the Director of the Institute of Astronomy, University of Cambridge. He is an observational astronomer with over 30 years of experience in instrument design and data management starting with large scale of photographic surveys with giga-pixel images. He was a member of the team that discovered the accelerating expansion of the Universe through the discovery and observations of distant supernovae. Richard's research interests span the discovery and observation of the most distant objects in the Universe in the Epoch of Reionization; development of instrumentation and computational data analysis techniques centered around large scale data intensive techniques using optical and infra-red imaging and spectroscopic sensors on telescopes around the world (primarily in Chile) and in Space using Gigapixel cameras and Petsacale multi-wavelength datasets. Richard is PI of the VISTA Hemisphere Survey and has 'Builder'status membership of the Dark Energy Survey. He is also the 4MOST Extragalactic Project Scientist.

Presenter: Prof. MCMAHON, Richard

Session Classification: Day 2

Contribution ID: **13** Type: **not specified**

Panel: ML Specific to Astro Physics

Tuesday 18 September 2018 12:00 (1 hour)

Moderator: Prof. Richard McMahon
Panelists: Dr. Shirley Ho

Session Classification: Day 2

Contribution ID: **14** Type: **not specified**

Optical transient surveys of today and tomorrow : machine learning applications

Tuesday 18 September 2018 10:30 (20 minutes)

Prof. Stephen J. Smartt, Astrophysics Research Centre
School of Mathematics and Physics, Queen's University Belfast

Abstract: Wide-field optical telescopes routinely employ large format detectors (CCDs) with 0.1 to 1 gigapixels which provide images every 30 seconds. Such facilities are capable of surveying the whole sky every night and the scientific exploitation requires immediate processing of the data together with selection of real astrophysical transients and association with all catalogues of stars, galaxies, asteroids and all multi-wavelength surveys (gamma ray to radio) that exist. Machine learning is now playing a critical role in this field and I will describe some of the applications currently being employed and what the challenges are for the next major facility –the Large Synoptic Survey Telescope.

Bio: Stephen Smartt is an astrophysicist at Queen's University Belfast working on time domain surveys of the sky in the optical and infra-red wavelengths. His scientific interest is finding extreme types of exploding, merging and erupting stars in these sky surveys in real time to allow larger facilities on the ground and in space to be triggered rapidly. This involves application of machine learning techniques and algorithms to create billion row database catalogues and select objects from their catalogued properties.

Presenter: Prof. SMARTT, Stephen

Session Classification: Day 2

Contribution ID: **15** Type: **not specified**

Accelerating science by re-purposing machine learning software

Tuesday 18 September 2018 10:50 (20 minutes)

Dr. Bojan Nikolic, Principal Research Associate at the Cavendish Laboratory, Cambridge

Abstract: I show that a software framework intended primarily for training of neural networks, PyTorch, is easily applied to a general function minimisation problem in science. The qualities of PyTorch of ease-of-use and very high efficiency are found to be applicable in this domain and lead to two orders of magnitude improvement in time-to-solution with very small software engineering effort. This result demonstrates that re-purposing of machine learning software can lead faster scientific results even when using traditional scientific approaches (i.e., exact models motivated by hypotheses). Paper Link: https://arxiv.org/abs/1805.07439

Bio: Dr. Bojan Nikolic is a Principal Research Associate at the Cavendish Laboratory, Cambridge. Bojan's interests span astronomy, instrumentation, computing and software. He has played a major role in construction/commissioning of three prominent recent radio astronomy facilities: the GBT 100-m telescope, the ALMA 66-element array and the forthcoming Square Kilometre Array. His current astronomical research is focused on the formation of the earliest stars and black holes and the "Epoch of Reionisation" that is triggered by these objects. Bojan has also worked in industry as a senior software engineer and been a speaker at prominent computing conferences. His software is run daily in production at observatories and commercial organisations around the world.

Presenter: Dr NIKOLIC, Bojan

Session Classification: Day 2

Contribution ID: **16** Type: **not specified**

AstroStatistics and AstroInformatics in the content of the SKA and LSST

Tuesday 18 September 2018 11:10 (20 minutes)

Abstract: Over recent decades cosmology has transitioned from a data-poor to a data-rich field, which has lead to dramatic improvements in our understanding of the cosmic evolution of our Universe. Nevertheless, we remain ignorant of many aspects of the scenario that has been revealed. Little is known about the fundamental physics of structure formation in the early Universe or the formation of the first large-scale structure during the epoch of reionization. A complete understanding of dark energy and dark matter, which dominate the late evolution of our Universe, also remains elusive. In coming decades the field will transition from being data-rich to being overwhelmed by data as next-generation observational facilities come online. The emerging big-data era of cosmology – spearheaded by the SKA and LSST experiments – has the potential to lead to another dramatic improvement in our understanding, addressing unanswered fundamental questions about the content and evolution of our Universe – provided that we can make sense of the overwhelming data-sets that will be acquired. I will very briefly review a variety of cutting edge data science techniques developed and applied in astrophysics, including compressive sensing approaches, algorithms that scale to overwhelming large data-sets, uncertainty quantification techniques, and machine learning approaches.

Bio: Jason McEwen is a University Reader in the Mullard Space Science Laboratory (MSSL) at University College London (UCL). He is also Director of Research (Astrophysics) of the UCL Centre for Doctoral Training (CDT) in Data Intensive Science (DIS). He has broad multidisciplinary research interests in applied mathematics, statistics, machine learning, and astrophysics. He in heavily involved in numerous astrophysical experiments, including Planck, Euclid, LSST and the SKA. He is also the Founder of KageNova, a startup company specialised in core technology for virtual and augmented reality.

Presenter: Prof. MCEWEN, Jason (Astrophysics & Cosmology at Mullard Space Science Laboratory, University College London)

Session Classification: Day 2

Contribution ID: **17** Type: **not specified**

How to train taggers on data

Tuesday 18 September 2018 11:30 (15 minutes)

Dr. Jennifer Thompson, CERN

Abstract: The machine learning methods currently used in high energy particle physics often rely on Monte Carlo simulations of signal and background. A problem with this approach is that it is not always possible to distinguish whether the machine is learning physics or simply an artefact of the simulation. In this presentation I will explain how it is possible to perform a new physics search with a tagger that has been trained entirely on background data. I will show how jets of particles produced at the LHC, are prime targets for such an approach. To this end, we use an unsupervised learning method, trained entirely on background jets, to detect any anomalous result as a new physics signal. I will show a range of applications for this approach, and describe how to practically include it in an experimental analysis by mass-decorrelating the network output in an adversarial framework and performing a bump hunt.

Bio: I am a theorist in particle physics based at Heidelberg university for the second stage of my first postdoc. I graduated from the University of Durham with work on Monte Carlo simulations of events at high energy colliders, on NLO corrections in QCD, and on approximations to NLO EW corrections. From this, I moved into interfacing full NLO EW corrections before transitioning to using Machine Learning algorithms for phenomenological studies in a high energy particle physics setting.

Presenter: Dr THOMPSON, Jennifer

Session Classification: Day 2

Contribution ID: **18** Type: **not specified**

Coffee Break

Tuesday 18 September 2018 11:45 (15 minutes)

Session Classification: Day 2

Contribution ID: **19** Type: **not specified**

Panel: Areas for Possible Collaboration

Tuesday 18 September 2018 17:00 (1 hour)

Moderator: Prof. Paul Alexander
Panelists: Dr. Peter Braam, Prof. Terry Lyons, Prof. Richard McMahon, Prof. Ian Shipsey

Session Classification: Day 2

Contribution ID: **20** Type: **not specified**

Removing Clutter - Killing Dimension

Tuesday 18 September 2018 14:00 (30 minutes)

Professor Terry Lyons FLSW FRSE FRS, Wallis Professor of Mathematics, Mathematical Institute, University of Oxford

Abstract: It often happens that measurements have redundant information over and above the invariants of interest; we might measure a rigid object in cartesian co-ordinates although we are only interested in the shape. Representing in an invariant way that does not carry redundant information about its embedding can significantly enhance many processes. If there is a representation where the invariance are highly nonlinear this cleaning can be challenging and the ever changing clutter adds noise and degrades learning processes. One critical and ignored symmetry occurs when we sample a stream of data, often we care little about how the path was sampled or parametrised, and we just care about the trajectory (in space time). This is an infinite dimensional set of symmetries, and if the sampling changes with each observation, or the data comprises rare tranisents, or has missing data. The induced noise can be very considerable. Recent mathematics introduces a transform the signature that describes streamed data faithfully without introducing a parametrisation. It has considerable benefits in terms of dimension reduction etc.

Bio: Professor Terry Lyons is the Wallis Professor of Mathematics at the University of Oxford; he was a founding member (2007) of, and then Director (2011-2015) of, the Oxford Man Institute of Quantitative Finance; he was the Director of the Wales Institute of Mathematical and Computational Sciences (WIMCS; 2008-2011). He came to Oxford in 2000 having previously been Professor of Mathematics at Imperial College London (1993-2000), and before that he held the Colin Maclaurin Chair at Edinburgh (1985-93). Prof Lyons'long-term research interests are all focused on Rough Paths, Stochastic Analysis, and applications –particularly to Finance and more generally to the summarising of large complex data. More specifically, his interests are in developing mathematical tools that can be used to effectively model and describe high dimensional systems that exhibit randomness. Prof Lyons is involved in a wide range of problems from pure mathematical ones to questions of efficient numerical calculation.

Presenter: Prof. LYONS FLSW FRSE FRS, Terry

Session Classification: Day 2

Contribution ID: **21** Type: **not specified**

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

Tuesday 18 September 2018 14:30 (15 minutes)

Dr. Adam Elwood

Abstract:"We introduce two new loss functions designed to directly optimise the statistical significance of the expected number of signal events when training neural networks to classify events as signal or background in the scenario of a search for new physics at a particle collider. The loss functions are designed to directly maximise commonly used estimates of the statistical significance, s/√(s+b), and the Asimov estimate, Z_A. We consider their use in a toy SUSY search with 30˜fbˆ(−1) of 14˜TeV data collected at the LHC. In the case that the search for the SUSY model is dominated by systematic uncertainties, it is found that the loss function based on Z_A can outperform the binary cross entropy in defining an optimal search region."
arXiv: https://arxiv.org/abs/1806.00322

Bio: I'm a postdoctoral researcher at the DESY particle physics laboratory in Hamburg and a member of the CMS collaboration. My interests are in the applications of machine learning techniques to searches for Beyond the Standard Model (BSM) physics, particularly supersymmetry and dark matter. I completed my PhD at Imperial College London with a search for supersymmetry in the all hadronic final state at the CMS experiment.

Presenter: Dr ELWOOD, Adam

Session Classification: Day 2

Contribution ID: **22** Type: **not specified**

Coffee Break

Session Classification: Day 2

Contribution ID: **23** Type: **not specified**

Radio Galaxy Classification with Deep Learning

Tuesday 18 September 2018 14:45 (15 minutes)

Abstract: Machine learning techniques have proven to be increasingly useful in astronomical applications over the last few years, for example in object classification, estimating redshifts and data mining. A topic of current interest is to classify radio galaxy morphology, as it gives us insight into the nature of the AGN, surrounding environment and evolution of the host galaxy. The task of performing classifications manually is tedious, especially with the development of future surveys that probe more deeply and widely into search-space, such as the SKA and EMU. This necessitates the use of automated techniques. Convolutional neural networks are a machine learning technique that have been very successful in image classification, due to their ability to capture high-dimensional features in the data. A drawback of the technique is the use of the pooling operation, which results in information loss and does not preserve the relative position of features in the image. Capsule networks however are able to preserve this information with the use of dynamic routing via capsules. We explore a convolutional neural network architecture against variations of Capsule network setups and evaluate their performance in replicating the classifications of radio galaxies detected by LOFAR.

Biography: I completed a Bachelor of Engineering/Science and Masters of Physics at the University of Melbourne. Afterwards I worked as a research assistant in Bioinformatics for almost 4 years, implementing software algorithms and statistics to analyse genetic data, running NGS pipelines on sequencing data of individuals in pedigrees with rare diseases to identify rare potentially disease-causing variants, and researching in-silico gene prioritisation using Allen Human Brain Atlas data to identify potentially co-expressed genes. I am currently at the University of Hamburg completing a PhD on using deep learning techniques to classify radio galaxies, with the first publication exploring the classification of sources from the Radio Galaxy Zoo, where an accuracy of 94.8% was achieved on Data Release 1, when classifying into 3 classes of data.

Presenter: Ms LUKIC, Vesna (PhD candidate University of Hamburg)

Session Classification: Day 2

Contribution ID: **24** Type: **not specified**

Time-domain Machine Learning - Opportunities and Challenges for the SKA

Tuesday 18 September 2018 16:15 (30 minutes)

Abstract: To harness the discovery potential of data collected by the SKA, we require efficient and effective automated data processing methods. Machine learning tools have the potential to deliver this capability, as evidence via their successful application to similar problems in the astronomy domain. This talk introduces the machine learning required for successful time-domain data processing (pulsar / transient discovery), and the infrastructure required to support it. Here the overriding aim is to increase awareness of what is required to facilitate the execution of automated learning methods, which we'll need if we are to achieve the SKA's ambitious science goals.

Biography: Machine Learning Researcher in the SKA Group, School of Physics & Astronomy @ The University of Manchester. Member of the Central Signal Processor (CSP) Consortia, working on the design of the Pulsar Search Sub-system (PSS). Member of the SDP Consortia, Non-Imaging Processing Specialist (Pulsar/transient search, Pulsar Timing). Software engineer in a past life.

Presenter: Dr LYON, Robert

Session Classification: Day 2

Contribution ID: **25** Type: **not specified**

Networked data-science for research, academic communities and beyond

Tuesday 18 September 2018 15:30 (30 minutes)

Dr. Andrey Ustyuzhanin, Head of Yandex-CERN Joint Research Projects & Head of the Laboratory of Methods for Big Data Analysis at National Research University Higher School of Economics

Abstract: There is an exceptional way of doing data-driven research employing networked community. The following examples can illustrate the approach: Galaxy Zoo or Tim Gower's blog. However many cases of collaboration with the data-science community within competitions organised on Kaggle or Coda Lab platforms usually get limited by restrictions on those platforms. Common Machine Learning quality metrics do not necessarily correspond to the original research goal. Constraints imposed by the problem statement typically look artificial for ML-community. Preparing a perfect competition takes a considerable amount of efforts. On the contrary research process requires a lot of flexibility and ability to look at the problem from different angles. I'll describe the alternative research collaboration process can bridge the gap between domain-specific research and data science community. Notably, it can involve academic researchers, younger practitioners and all enthusiasts who are willing to contribute. Such research process can be supported by an open computational platform that will be described along with essential examples for the audience of the workshop.

Biography: Dr Andrey Ustyuzhanin - the head of Yandex-CERN joint research projects as well as the head of the Laboratory of Methods for Big Data Analysis at NRU HSE. His team is the member of frontier research international collaborations: LHCb - collaboration at Large Hadron Collider, SHiP (Search for Hidden Particles) - the experiment is designed for the New Physics discovery. His group is unique for both collaborations since the majority of the team members are coming from the Computer and Data Science worlds. The primary priority of his research is the design of new Machine Learning methods and using them to solve tough scientific enigmas thus improving the fundamental understanding of our world. Amongst the project he has been working on are efficiency improvement of online triggers at LHCb, speed up BDT-based online processing formula, the design of custom convolutional neural networks for processing tracks of muon-like particles on smartphone cameras. Development of the algorithm for tracking in scintillators optical fibre detectors and emulsion cloud chambers. Those project aid research at various experiments: LHCb, OPERA, SHiP and CRAYFIS. Discovering the deeper truth about the Universe by applying data analysis methods is the primary source of inspiration in Andrey's lifelong journey. Andrey is a co-author of the course on the Machine Learning aimed at solving Particle Physics challenges at Coursera and organiser of the annual international summer schools following the similar set of topics. Andrey has graduated from Moscow Institute of Physics and Technology in 2000 and received PhD in 2017 at Institute of System Programming Russian Academy of Sciences.

Presenter: Dr USTYUZHANIN, Andrey

Session Classification: Day 2

Contribution ID: **26** Type: **not specified**

Learning the solution to SDEs using rough paths theory and machine learning

Tuesday 18 September 2018 15:00 (15 minutes)

Abstract: In this talk, we consider the supervised learning problem where the explanatory variable is a data stream. We provide an approach based on identifying carefully chosen features of the stream which allows linear regression to be used to characterise the functional relationship between explanatory variables and the conditional distribution of the response; the methods used to develop and justify this approach, such as the signature of a stream and the shuffle product of tensors, are standard tools in the theory of rough paths and provide a unified and non-parametric approach with potential significant dimension reduction. Motivated by the numerical schemes of SDEs, we propose the new algorithm combining the recurrent neural network (RNN) with the signature feature to further improve the efficiency of the linear signature method. We apply it to the example of learning the unknown SDEs and demonstrate the superior effectiveness of this method benchmarked with RNN with raw data.

Bio: Dr Hao Ni is a senior lecturer in financial mathematics at UCL and the Turing Fellow at the Alan Turing institute since September 2016. Prior to this she was a visiting postdoctoral researcher at ICERM and Department of Applied Mathematics at Brown University from 2012/09 to 2013/05 and continued her postdoctoral research at the Oxford-Man Institute of Quantitative Finance until 2016. She finished her D.Phil. in mathematics at University of Oxford. Her research interests include stochastic analysis, financial mathematics and machine learning. More specifically she is interested in non-parametric modelling effects of data streams through rough paths theory and statistical models and its various applications, e.g. online Chinese handwritten character and financial data streams analysis.

Presenter: Dr NI, Hao (Senior Lecturer in Financial Mathematics at UCL and the Turing Fellow at The Alan Turing Institute)

Session Classification: Day 2

Contribution ID: **27** Type: **not specified**

Building Search Pipelines and Labelled Datasets for Transient Discovery

Tuesday 18 September 2018 16:00 (15 minutes)

Abstract: Breakthrough Listen is the largest effort to date to search for techno-signatures from extraterrestrial civilizations. We use extensive computing power to search at high frequency and time resolution for transients events in petabytes of observational data from the Green Bank Telescope, Parkes Telescope, LOFAR, and soon MeerKAT. Given the diverse manifestations of transient signals in an environment of increasing anthropogenic RFI, machine learning-based models are proving essential in filtering and detecting anomalous signals. We are actively using classic wide-feature models and deep neural network models to detect and classify astrophysical signals such as pulsars and FRBs. We are building large labelled datasets from a diverse sample of observations in order to facilitate ML research groups in developing new algorithms which we can incorporate into our search pipelines.

Bio: Griffin is a post-doctoral researcher in the physics department at the University of Oxford and a visiting scholar at the University of California at Berkeley. He is the project scientist for Breakthrough Listen on MeerKAT. He was previously an SKA Research Fellow at Rhodes University and SARAO in South Africa where he worked on interferometric calibration and imaging. His D.Phil thesis (Oxon 2013) was on FPGA-based correlator design and low-frequency aperture synthesis.

Presenter: Dr FOSTER, Griffin (University of Oxford, University of California at Berkeley)

Session Classification: Day 2