Universidade do Minho
Escola de Ciências

LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS
partículas e tecnologia

Big Data HEP

25 Abril
1974-2019

[Big data and machine learning at LIP]

Nuno Castro
nfcastro@lip.pt

(thanks to G. Milhano, A. Lindote and R. Conceição for the help in the preparation of the slides)

2nd Joint Workshop IGFAE/LIP, Santiago de Compostela, 26 April 2019

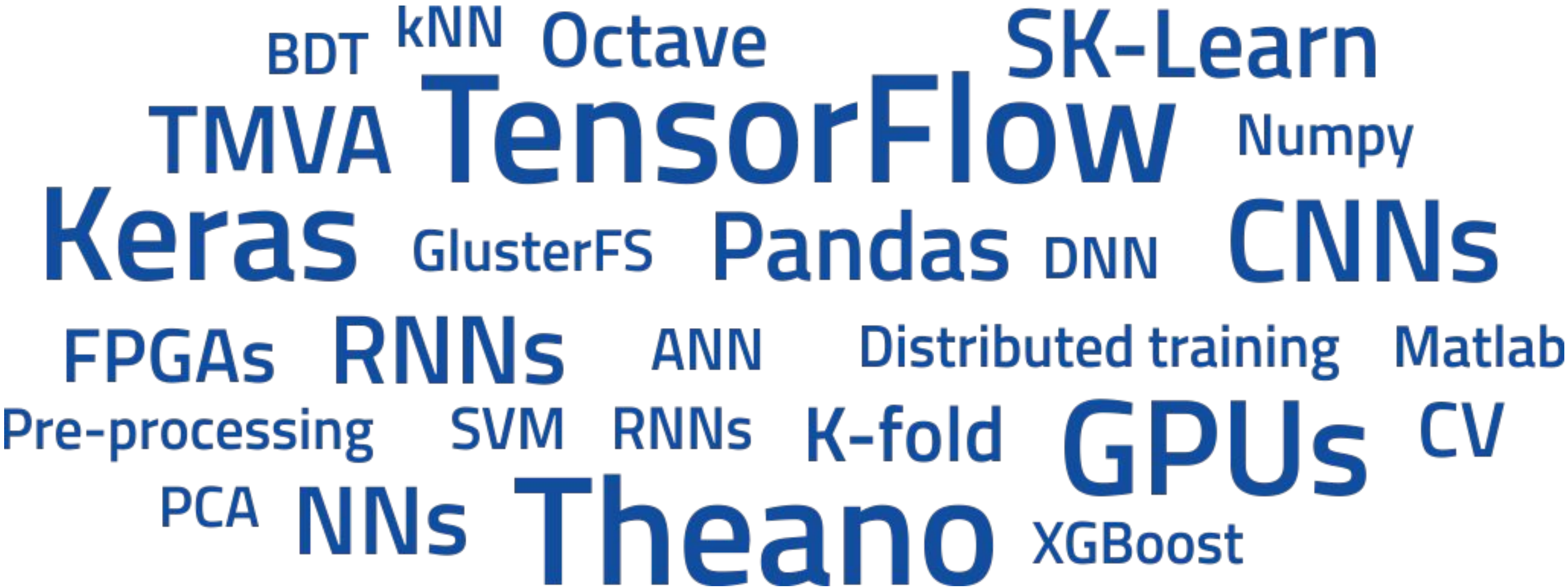
POCI/01-0145-FEDER-029147
PTDC/FIS-PAR/29147/2017

FCT Fundação para a Ciência e a Tecnologia
Lisb@20[20]
COMPETE 2020
PORTUGAL 2020
UNIÃO EUROPEIA
Fundo Europeu de Desenvolvimento Regional

Competence Center on Simulation and Big Data
data analysis and processing in particle physics

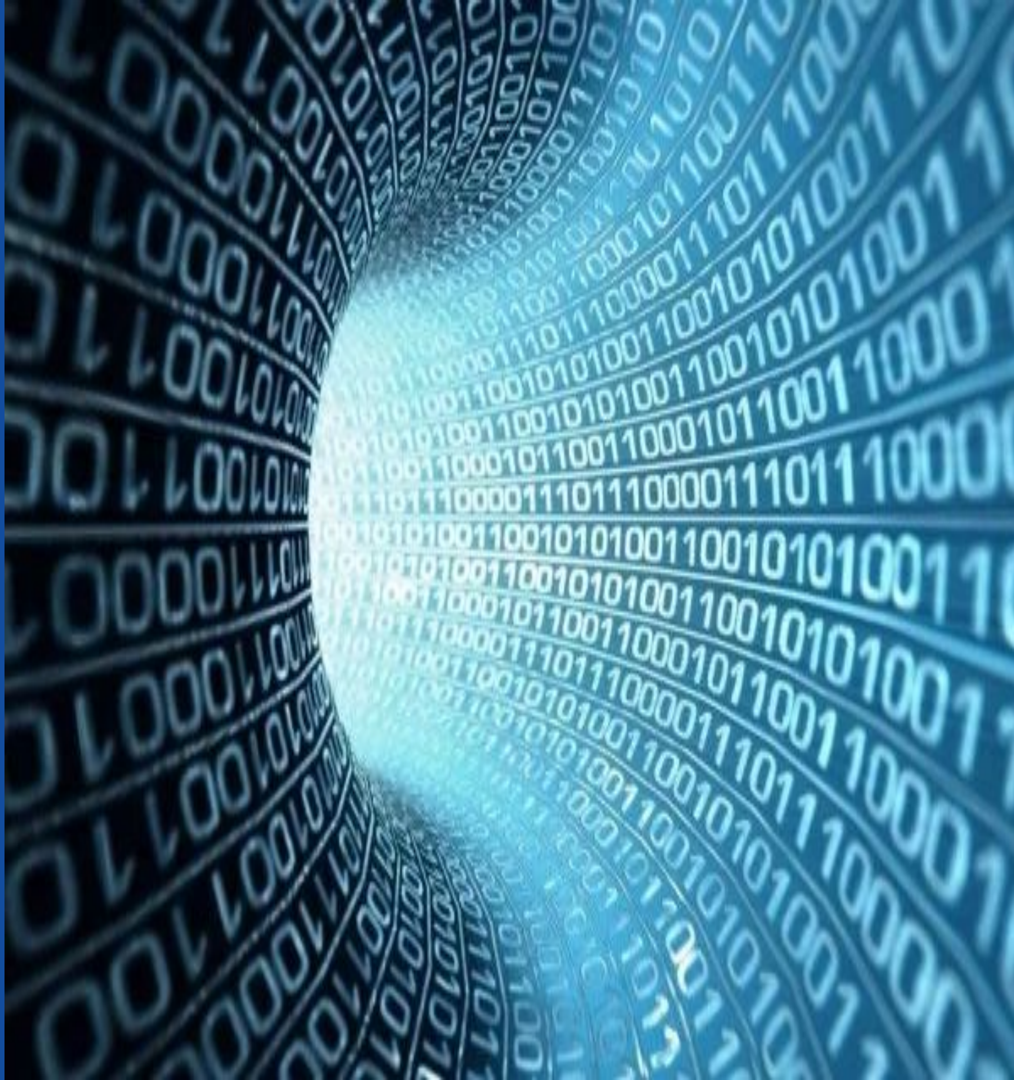
- LIP has been involved in the analysis of extremely large amounts of data produced by different experiments in High Energy Physics for a long time
- Expertise on the implementation and development of elaborate multivariate techniques aiming at a vast range of applications
- Competence in efficient data processing to better use the available computing resources

LIP competences
data analysis and processing in particle physics



Big Data

- LIP Computing group has a long experience in handling huge quantities of data
- Strong collaboration with CESGA - Centro de Supercomputación de Galicia



LIP
computing group

The LIP computing group provides IT services to LIP and its research groups:

- Integrated management of all scientific computing resources
- Typical IT services for users and administrative services
- Support LIP physics research projects
- R&D mostly in distributed computing
- e-Science and e-Infrastructures
- Grid Computing (driven by WLCG)
- Cloud Computing
- Technical coordination of INCD

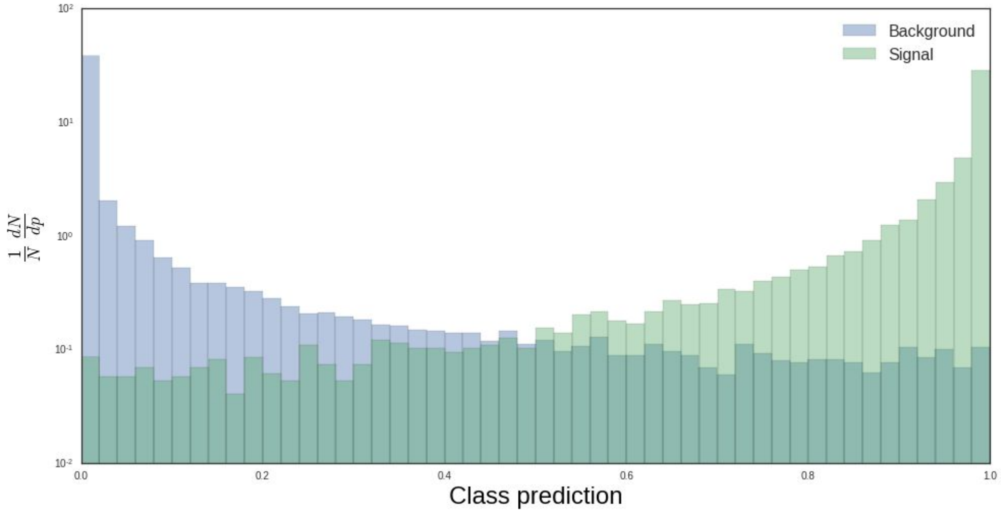
LIP
WLCG

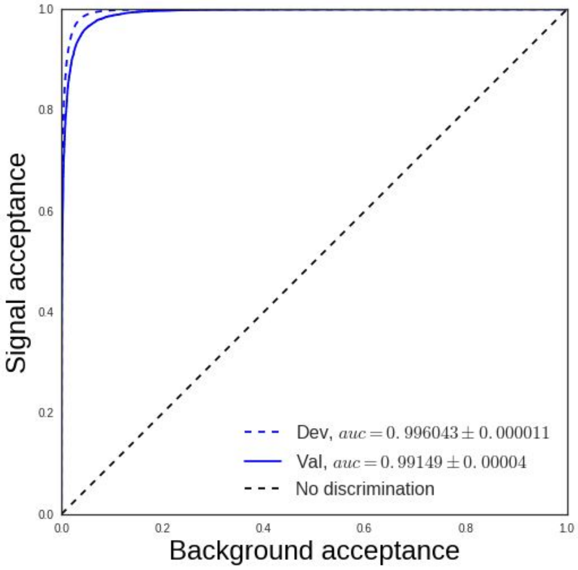
Worldwide LHC Computing Grid

Machine Learning at LIP

Machine learning at LIP
training on modern tools

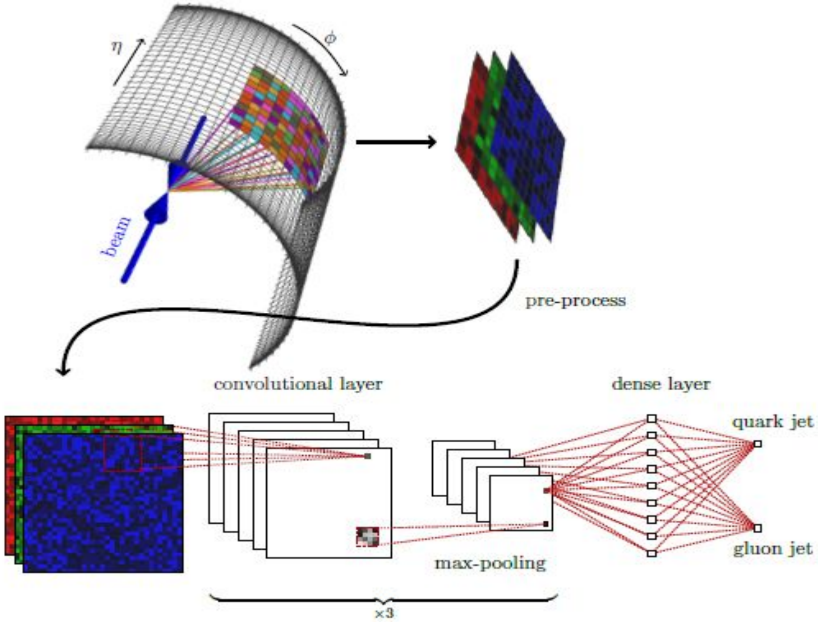
https://github.com/GilesStrong/ML_Tutorials
https://github.com/GilesStrong/LIP_DSS_Keras_Tutorial_2019





Studying jets at the LHC
using ML to understand very subtle effects

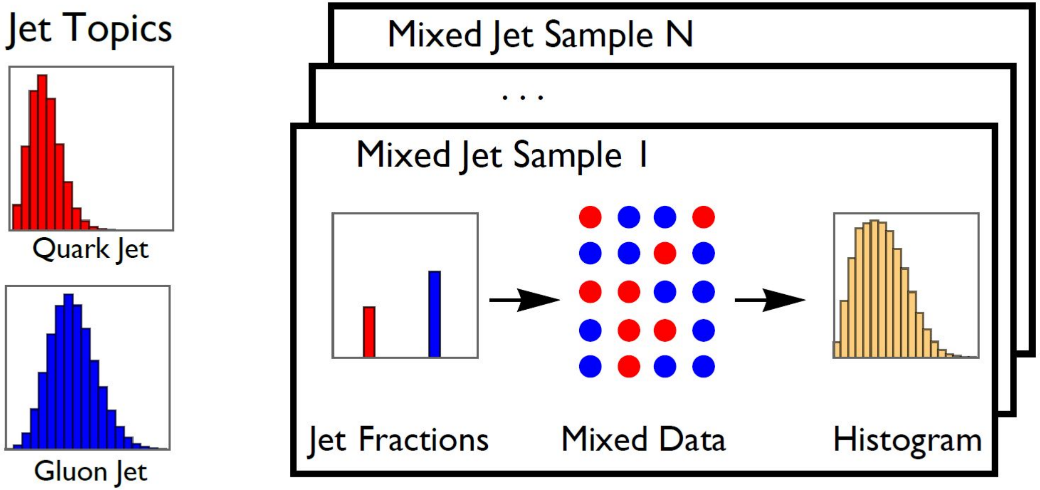
- Machine learning is used since a few years to study jets in colliders



JHEP 01 (2017) 110

Studying jets at the LHC
using ML to understand very subtle effects

- learning different topics from samples populated differently (Demix method)



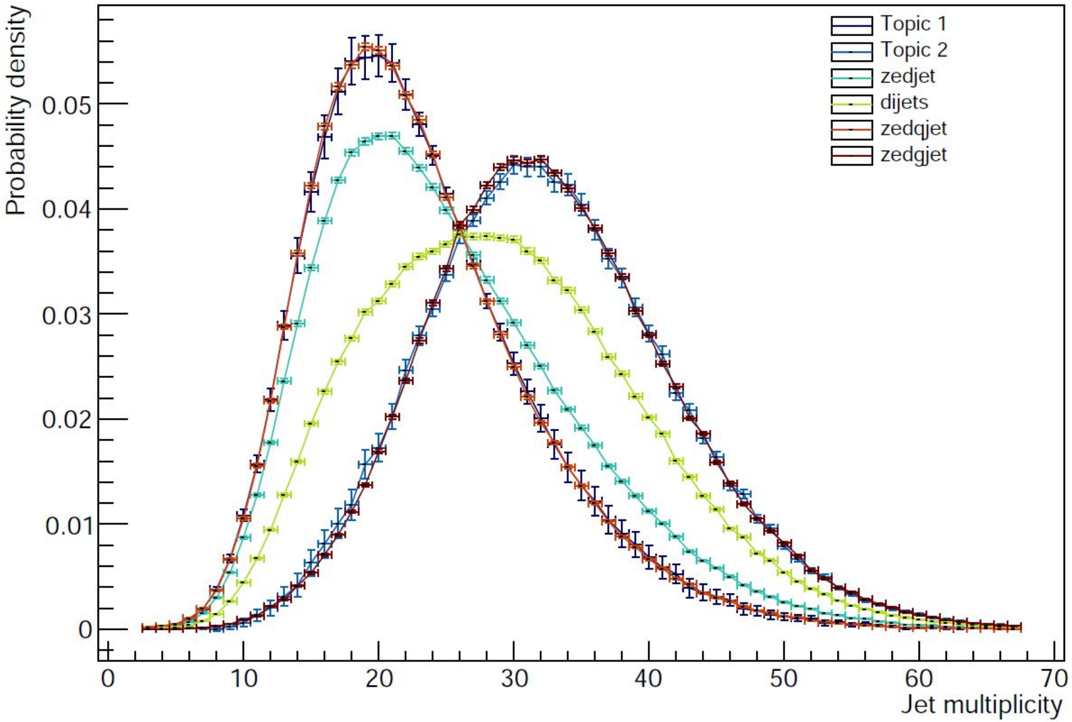
[Phys. Rev. Lett. 120, 241602 (2018)]

Studying jets at the LHC
using ML to discriminate between gluon and quark jets

- Use of the Demix method for extraction of quark/gluon jet distinction by demixing physical samples with different quark/jet fractions
  - New noise reduction strategy:
    - histograms trimmed to escape noisy areas by checking when two consecutive points at both tails are incompatible with their statistical error (2σ)

- The algorithm is able to extract two different topics from **jet multiplicity** in MC samples for **Z+jet** (quark jet dominated) and **dijet** (gluon jet dominated)
  - The accuracy of the separation is checked by comparison to pure Z-quark jet and Z-gluon jet samples

Studying jets at the LHC
using ML to discriminate between gluon and quark jets

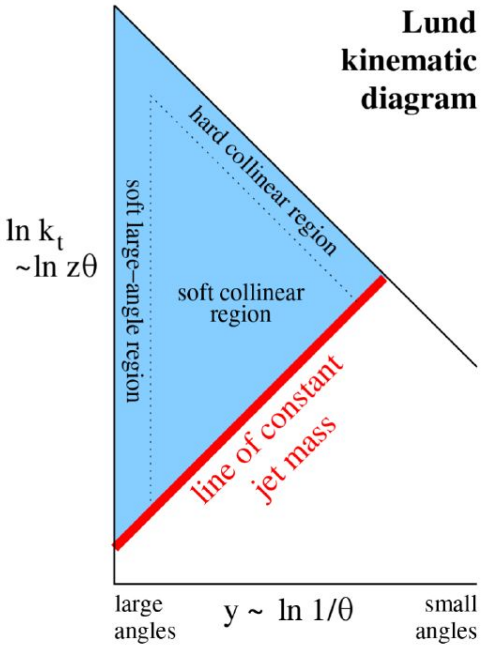


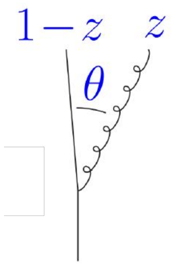
ongoing work by João Gonçalves, IST MSc student

Studying jets at the LHC
using ML to understand jet emissions

- Lund diagrams in the $(\ln z\theta, \ln\theta)$ plane are a very useful way of representing emissions.
- Different kinematic regimes are clearly separated, used to illustrate branching phase space in parton shower Monte Carlo simulations and in perturbative QCD resummations.
- Soft-collinear emissions are emitted uniformly in the Lund plane

$$dw^2 \propto \alpha_s \frac{dz}{z}\frac{d\theta}{\theta}$$



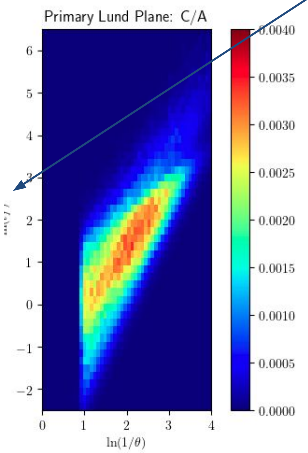


slide by Frédéric Dreyer, Zürich, May 2018

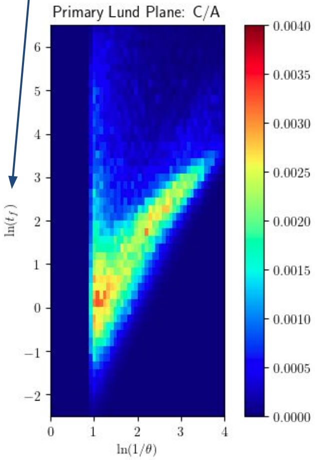
JHEP 1309 (2013) 029

Studying jets at the LHC
using ML to tag jets passing through a dense medium

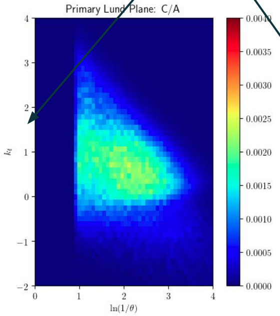
- distinction of quenched and unquenched jets using Lund planes
  - using $t_f = 1 / (p_T\, z\, \Theta^2)$ instead of the traditional k_T splitting



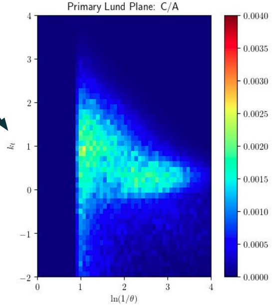
vacuum



QGP



vacuum

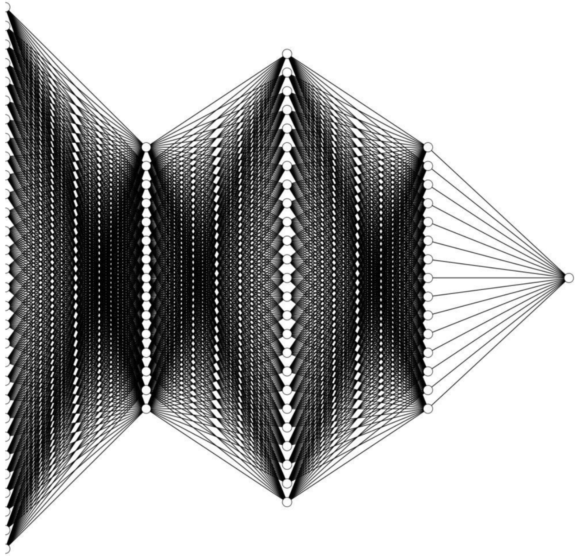


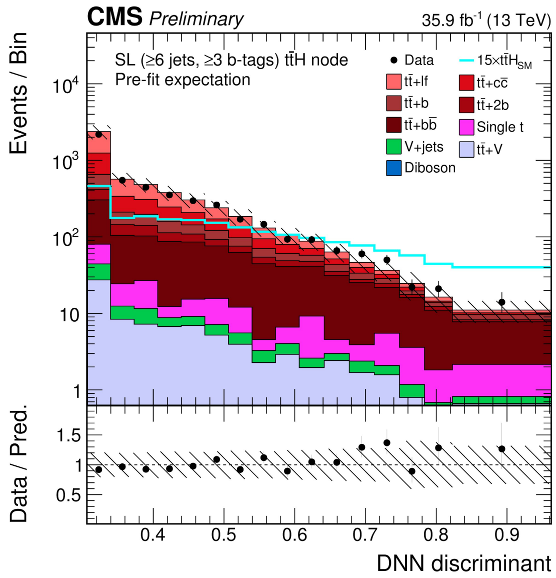
QGP

ongoing work by Filipa Peres, UMinho MSc student

Searching for rare events at the LHC
finding a needle in many haystacks

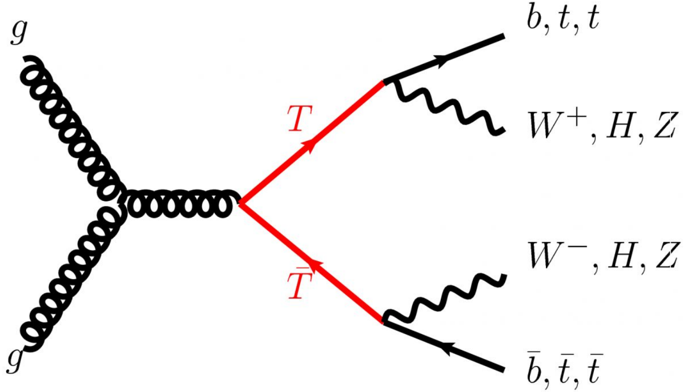
- the interesting collisions at the Large Hadron Collider are extremely rare so advanced multivariate techniques are required





Searching for the unknown

an example: vector-like quarks



charge 2/3

charge -1/3

Singlets $T_{L,R}$ $B_{L,R}$

charge 5/3

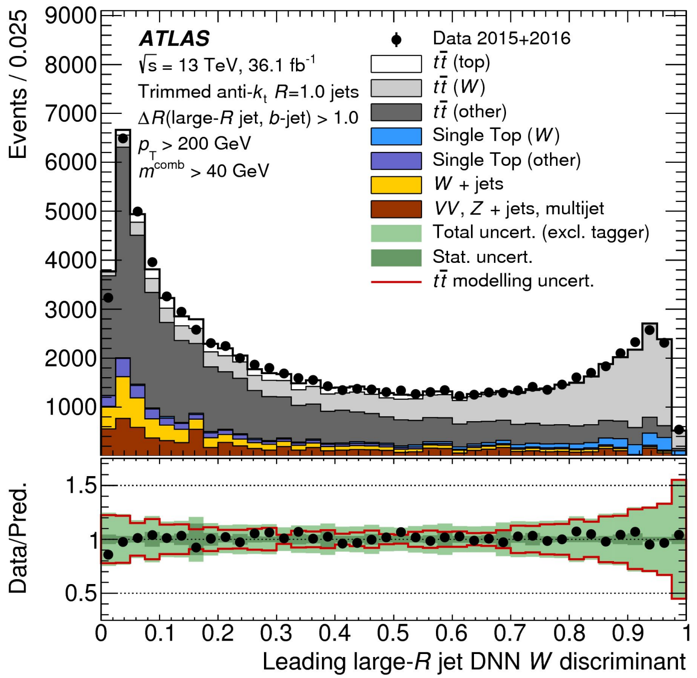
Doublets $\begin{pmatrix} X \\ T \end{pmatrix}_{L,R}$ $\begin{pmatrix} T \\ B \end{pmatrix}_{L,R}$ $\begin{pmatrix} B \\ Y \end{pmatrix}_{L,R}$

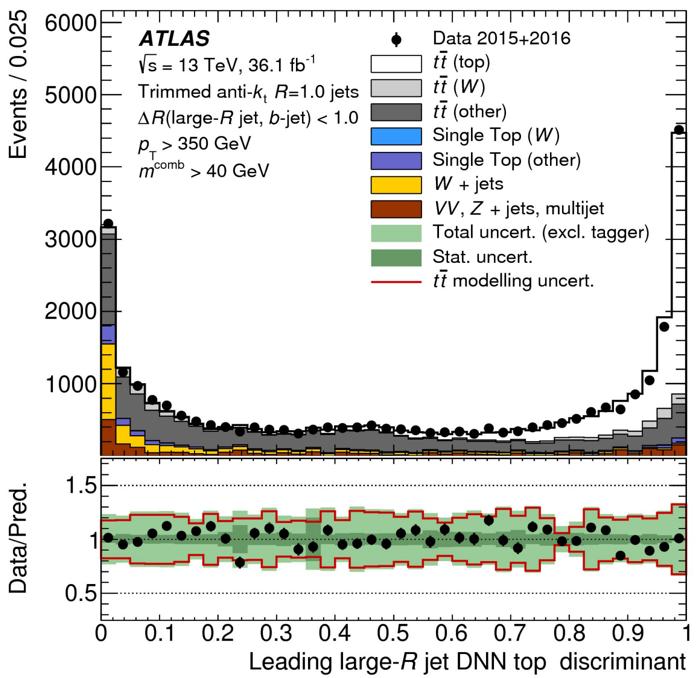
charge -4/3

Triplets $\begin{pmatrix} X \\ T \\ B \end{pmatrix}_{L,R}$ $\begin{pmatrix} T \\ B \\ Y \end{pmatrix}_{L,R}$

Searching for the unknown

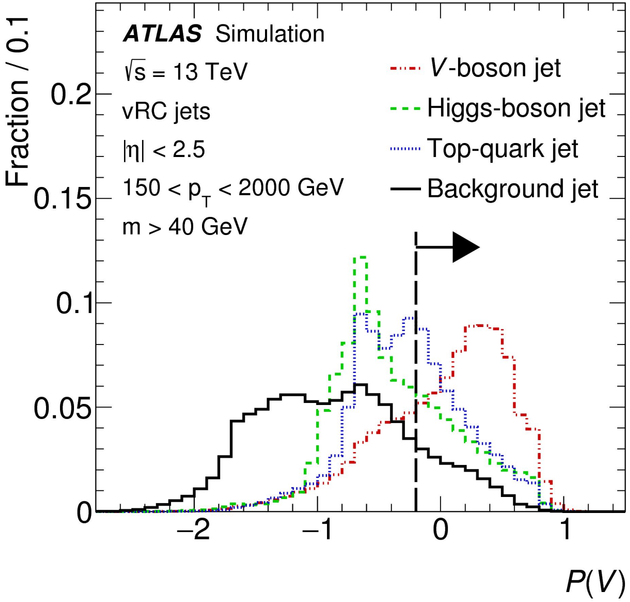
an example: use of neural networks in searches for object tagging





Searching for the unknown

an example: use of neural networks in searches for object tagging



$$P(V) = \log_{10}\left(\frac{D^{V}_{\mathrm{DNN}}}{0.9 \cdot D^{\mathrm{background}}_{\mathrm{DNN}} + 0.05 \cdot D^{t}_{\mathrm{DNN}} + 0.05 \cdot D^{H}_{\mathrm{DNN}}}\right)$$

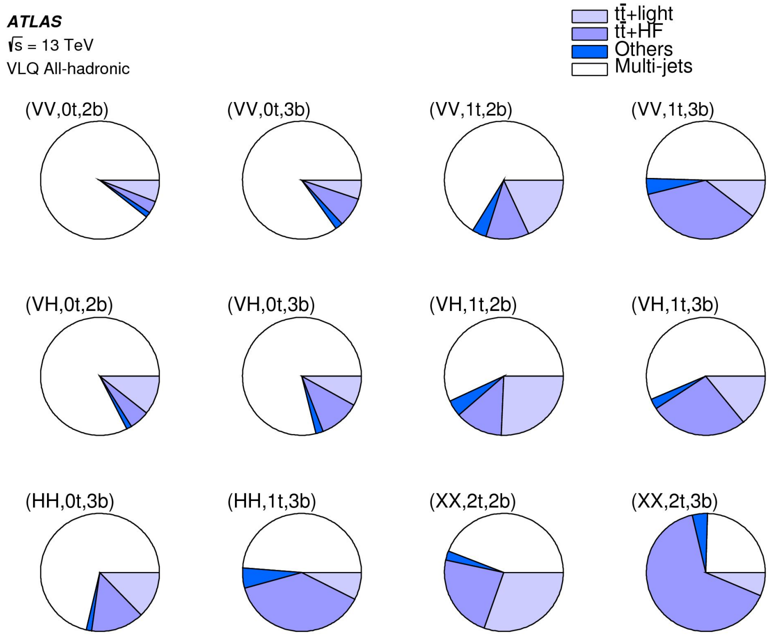
$$P(H) = \log_{10}\left(\frac{D^{H}_{\mathrm{DNN}}}{0.9 \cdot D^{\mathrm{background}}_{\mathrm{DNN}} + 0.05 \cdot D^{V}_{\mathrm{DNN}} + 0.05 \cdot D^{t}_{\mathrm{DNN}}}\right)$$

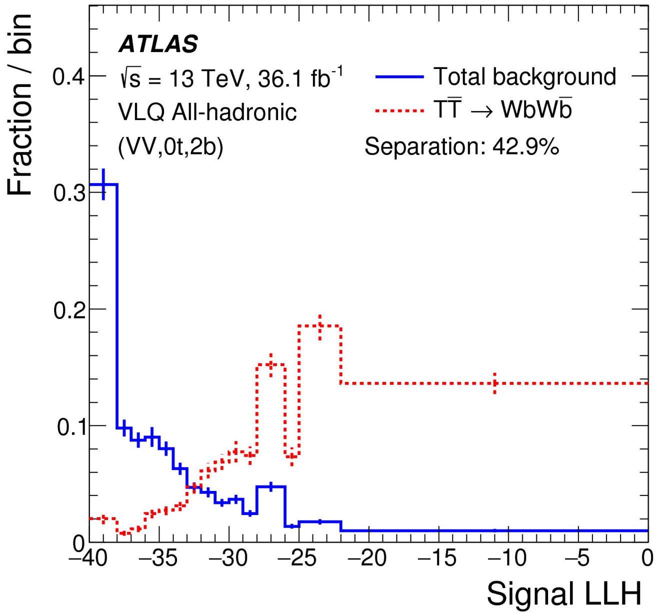
$$P(t) = \log_{10}\left(\frac{D^{t}_{\mathrm{DNN}}}{0.9 \cdot D^{\mathrm{background}}_{\mathrm{DNN}} + 0.05 \cdot D^{H}_{\mathrm{DNN}} + 0.05 \cdot D^{V}_{\mathrm{DNN}}}\right)$$

[Phys. Rev. D 98 (2018) 092005]

Searching for the unknown

an example: use of complex classification schemes in searches

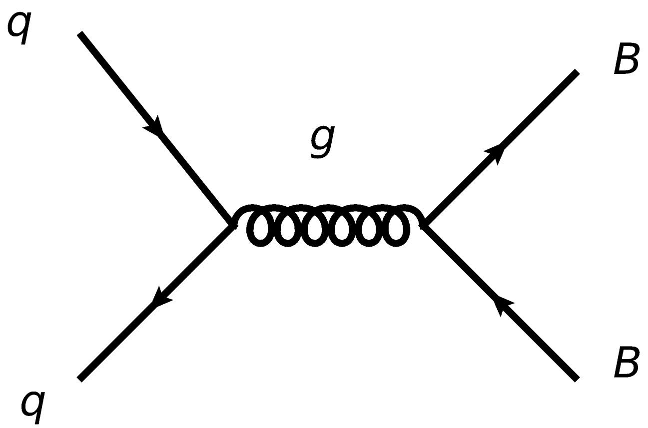


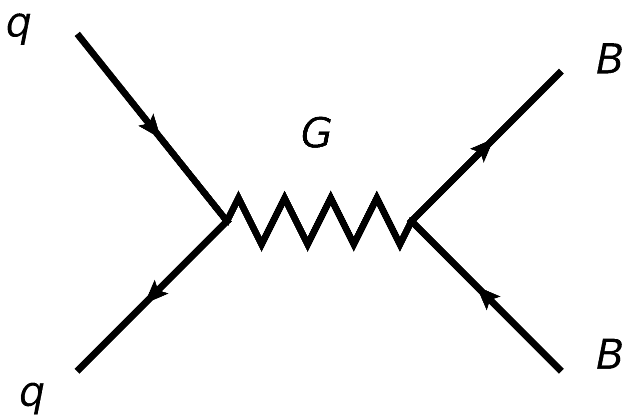


[Phys. Rev. D 98 (2018) 092005]

Searching for rare events
finding a needle in many haystacks

- ML can also help us to make sure we don’t miss subtle new physics signals





Searching for VLQ pair production
making sure we also target non-standard modes

- Use of low-level information instead of explicit final states reconstruction

- **Jets** (R = 0.4):
  - pT, mass, eta, phi, btag
  - 3 most energetic
- **Large-R** (1.0) jets:
  - pT, mass eta, phi, tau (1-5)
  - 3 most energetic
- **Leptons** (electrons and muons):
  - pT, eta, phi
  - 2 most energetic
- **MET**

ongoing work by Tiago Vale
MAP-Fis (UMinho) and
IDPASC PhD student

Searching for VLQ pair production
making sure we also target non-standard modes

- **Keras** with pandas and scikit-learn
  - Tensorflow as the backend

- Inputs are normalized, standardized and ran through PCA to decorrelate

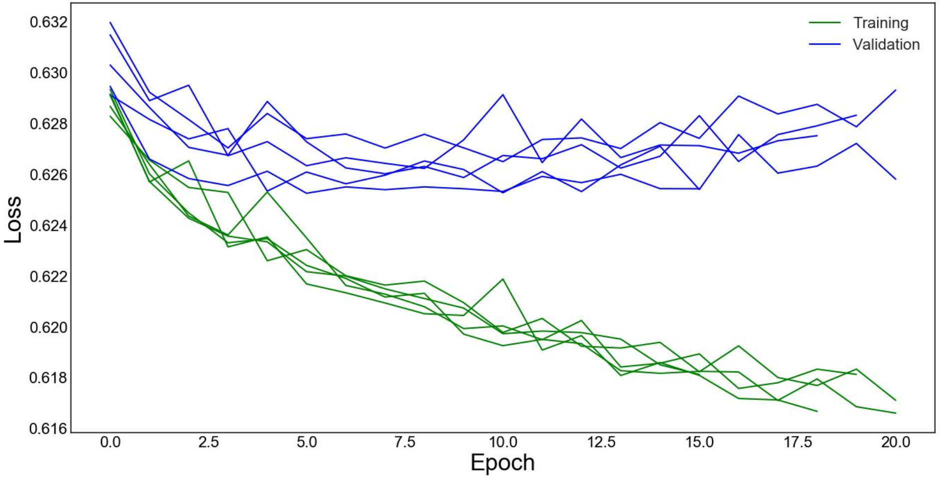
- Adamax with binary cross-entropy

- **First** architecture approach:
  - 3 layers of 100 nodes
  - selu as activation layer
  - Batch normalization in between each dense layer and its activation layer
  - Sigmoid in the output layer
  - Bayesian optimization machinery in place

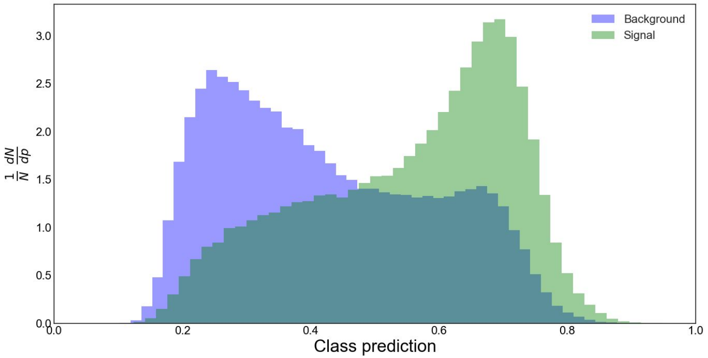
Searching for VLQ pair production
making sure we also target non-standard modes

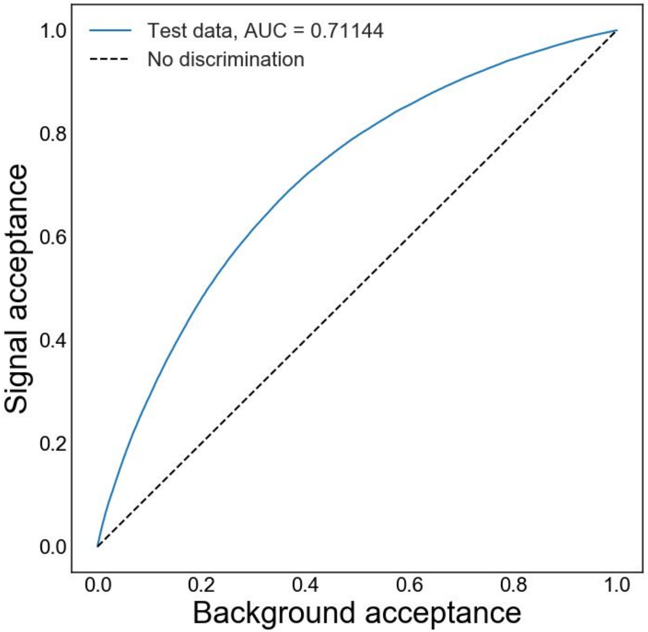
- First approach:

  - Test *pp-> g -> TT* against *pp->G->TT*
  - Stable training
  - ongoing work



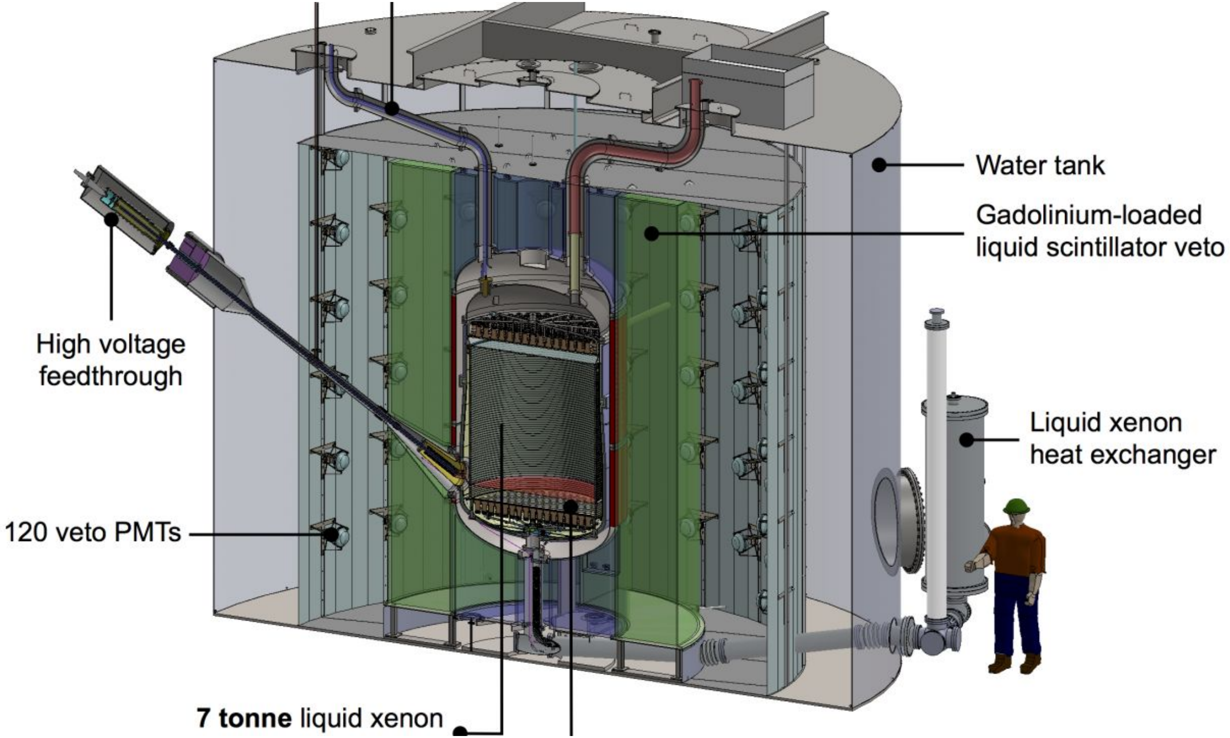
Searching for VLQ pair production
making sure we also target non-standard modes





- Decent discrimination
- Background and signal are HG and SM pair-production

Searching for Dark Matter
redefining the meaning of *rare events*

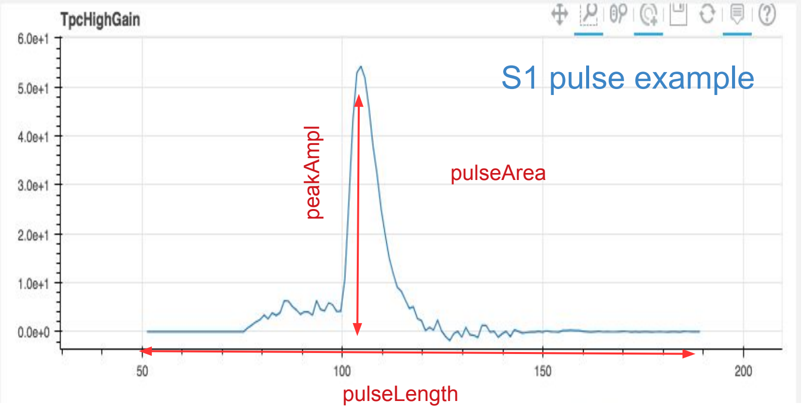


Searching for Dark Matter
using ML for pulse classification in LZ

Goal: Identify the nature of a given pulse based on its geometry, returning a prob. vector for different topologies [S1, S2, SPE, SE, MPE, Other]

Input: 17 geometric pulse parameters

Tools being used:
1. **Keras**
2. **Scikit-learn**



Data used for training/testing:
1. LZ simulated data - **7.3M** pulses
   (No pulse-level MCTruth available)
   Labels obtained by heuristic classifier with parameter selection criteria (decision tree)

ongoing work by
Paulo Brás,
UCoimbra PhD student

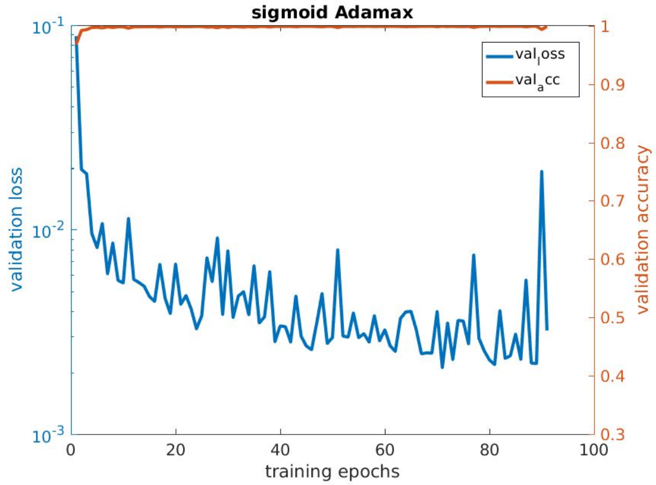
Searching for Dark Matter
using ML for pulse classification in LZ

Training:

- 7.6M pulses total
- 20% used for validation
- Learning rate = 0.001
- Batch size = 256

Optimization of the hidden section

- Layer size = 31
- depth = 3
- Activation = sigmoid
- Optimizer = Adamax
- Loss function = categorical_crossentropy



Confusion matrix

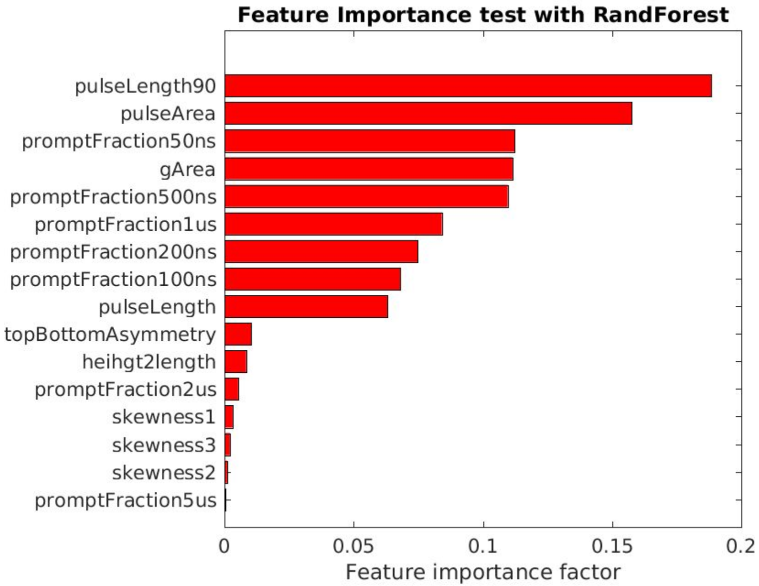
Predicted class	S1	S2	SE
Training label			
S1	2587379	262	86
S2	53	2502211	985
SE	118	4165	2540115

Efficiency loss dominated by S2/SE "misclassification", which doesn't impact the analysis

Average 99.93% accuracy

Searching for Dark Matter
other methods being evaluated for LZ

- **Random Forests**
  - Mainly used for finding the most relevant parameters (feature importance):



- **Isolation Forests**
  - Outlier detection: cleaning impure datasets
  - Used in tandem with other methods

- **SVM**
  - Optimization of selection regions in the parameter hyperspace.

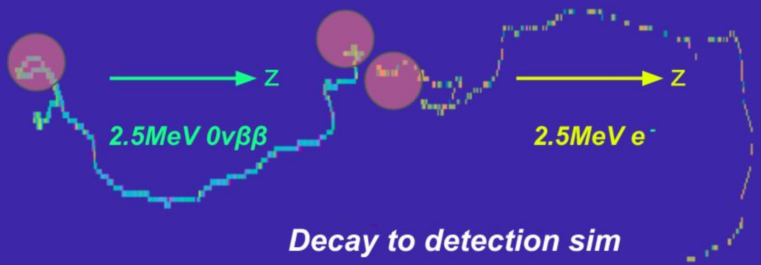
- **Convolutional Neural Nets**
  - Bypass pulse parametrization by reading pulse waveforms directly
  - Promising results with simplified synthetic pulses
- **Semi-supervised learning with Kernels (RKHS)** (work by Francisco Neves)
  - Classification generalization with only a small dataset of labeled data

Searching for Dark Matter

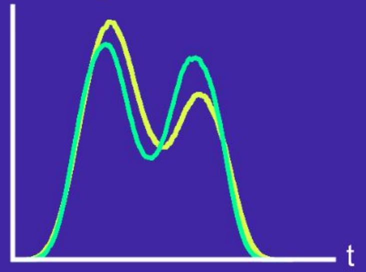
searching for Majorana Neutrinos with LZ

^{136}Xe decays via $2\nu\beta\beta$. If $\nu = \overline{\nu}$, $0\nu\beta\beta$ possible (beyond SM)

*In a LXe TPC, the **most significant bg src** for $0\nu\beta\beta$ is ~2.5MeV single electrons from scattering with high energy γ's*



Binary classification



Decay to detection sim

- *Energy deposition using GEANT4*
- *LXe secondary electron production, drift and diffusion*
- *Light propagation and PMT array signal using ANTS2 in distributed computing mode*

ML Plan

- *Use Keras for classification*
- *Parametrize signals so as to find best discrimination parameters, test using NN, Random Forests, etc.*
- *Feed waveform directly into CNN, sans parametrization*

ongoing work by Andrey Solovov, UCoimbra MSc student

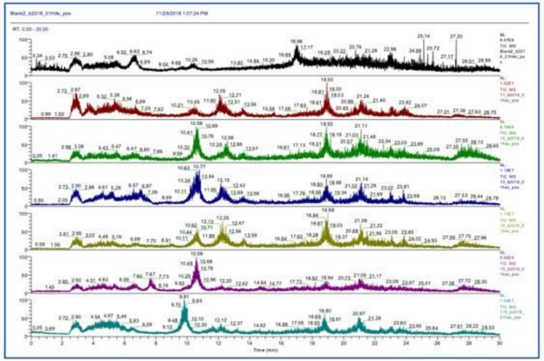
Collaborations beyond HEP

Machine Learning in Analytical Chemistry
collaborating with UMinho colleagues

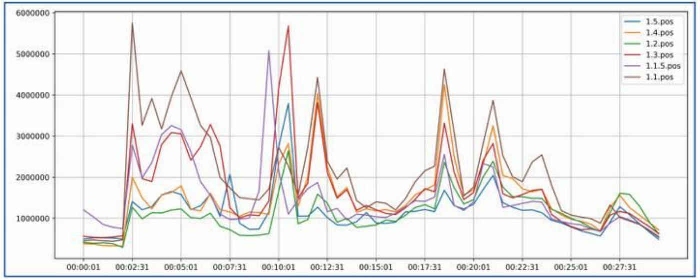
Study object:

- PCBs (Printed Circuit Boards)
- Train a model to classify PCBs as (not) contaminated fed with data obtained from chemical analysis.

Machine Learning in Analytical Chemistry
data analysis methods on chromatographic techniques



~12000 datapoints



60 datapoints

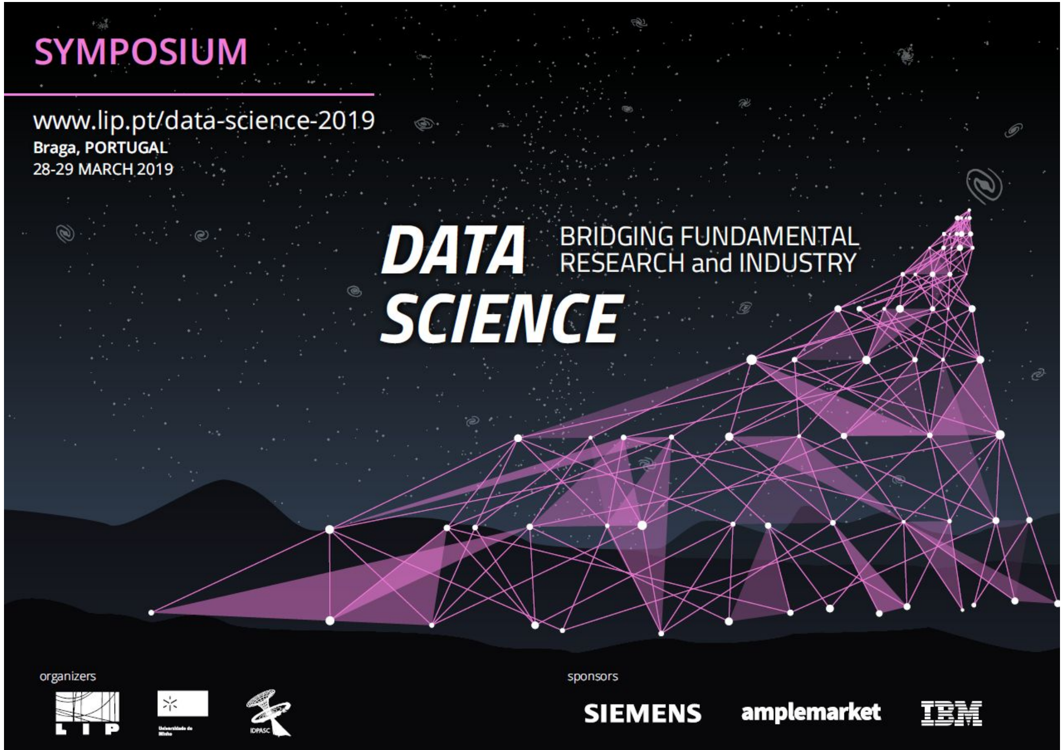
	00:00:01	00:00:31	00:01:01	00:01:31	00:02:01	00:02:31	00:03:01	00:03:31	00:04:01
1.5.pos	4.948382e+05	5.249579e+05	525739.883333	521650.775397	499838.630159	1.410234e+06	1.209399e+06	1.285888e+06	1.569353e+06
1.4.pos	3.820739e+05	3.748238e+05	338593.482540	332126.689683	324048.813492	2.001369e+06	1.498057e+06	1.230812e+06	1.562850e+06
1.2.pos	4.148494e+05	3.938719e+05	404375.784127	384809.158730	292515.057143	1.272567e+06	9.890456e+05	1.137694e+06	1.129999e+06
1.3.pos	5.633455e+05	5.368122e+05	533404.807143	546116.104762	574996.476984	3.303113e+06	1.968239e+06	1.891453e+06	2.795662e+06
1.1.5.pos	1.205078e+06	1.034886e+06	852045.073016	782434.719841	750664.167460	2.787587e+06	1.982766e+06	2.360404e+06	3.037262e+06

5 rows × 60 columns

ongoing work by
Diogo Barros,
UMinho
MSc student

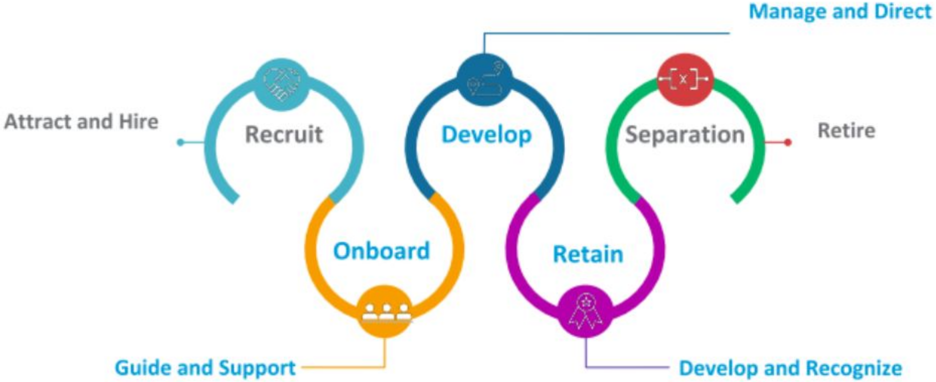
Exploring synergies between academia and industry

2nd edition of a workshop started last year



Machine Learning as a service
collaboration with Nielsen

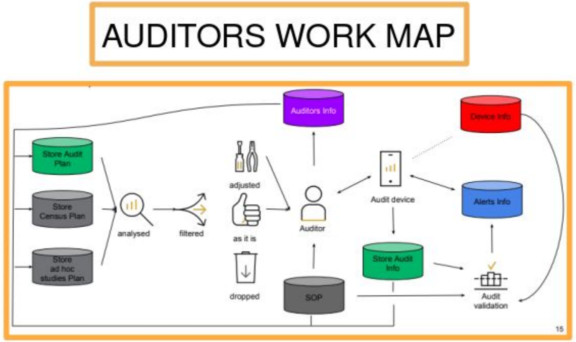
- How to predict Auditors attrition?



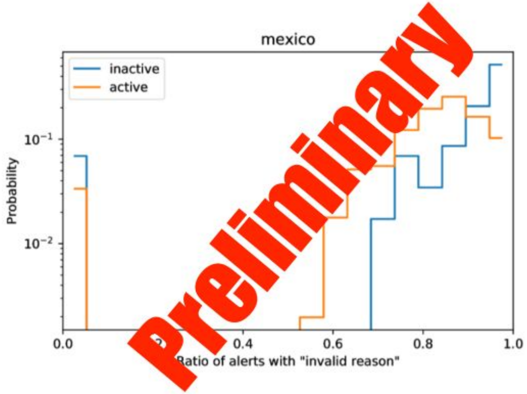
- Try to predict probability of an auditor to leave the company based on data related with his activities

Machine Learning as a service
collaboration with Nielsen

- Ongoing work



- Clean Data (make it trustworthy)
- Identify most sensitive quantities
  - 1st level: direct correlations
  - 2nd level: building up complex variables



summary

- in HEP we have a long time tradition (and expertise) in the analysis of large and complex data
- the most suitable technique has to be chosen for each problem
  - uncertainties and imperfect datasets
- synergies with other fields and activities possible/desirable