

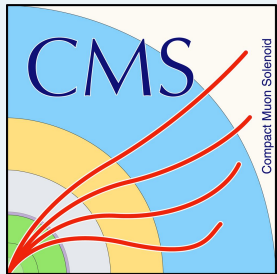
# Higgs searches @LHC exercise: Signal/background discrimination for the VBF Higgs four lepton decay channel with the CMS experiment

B. D'Anzi<sup>1,2</sup>, G. Miniello<sup>2</sup>, [W. Elmetenawee](#)<sup>2</sup>, N. De Filippis<sup>3,2</sup>, D. Diacono<sup>2</sup>, A. Sznajder<sup>4</sup>

<sup>1</sup>University of Bari; <sup>2</sup>INFN BARI; <sup>3</sup>Politecnico of Bari; <sup>4</sup>University of Rio de Janeiro

Fourth ML-INFN Hackathon: Starting Level

23<sup>rd</sup> June 2023



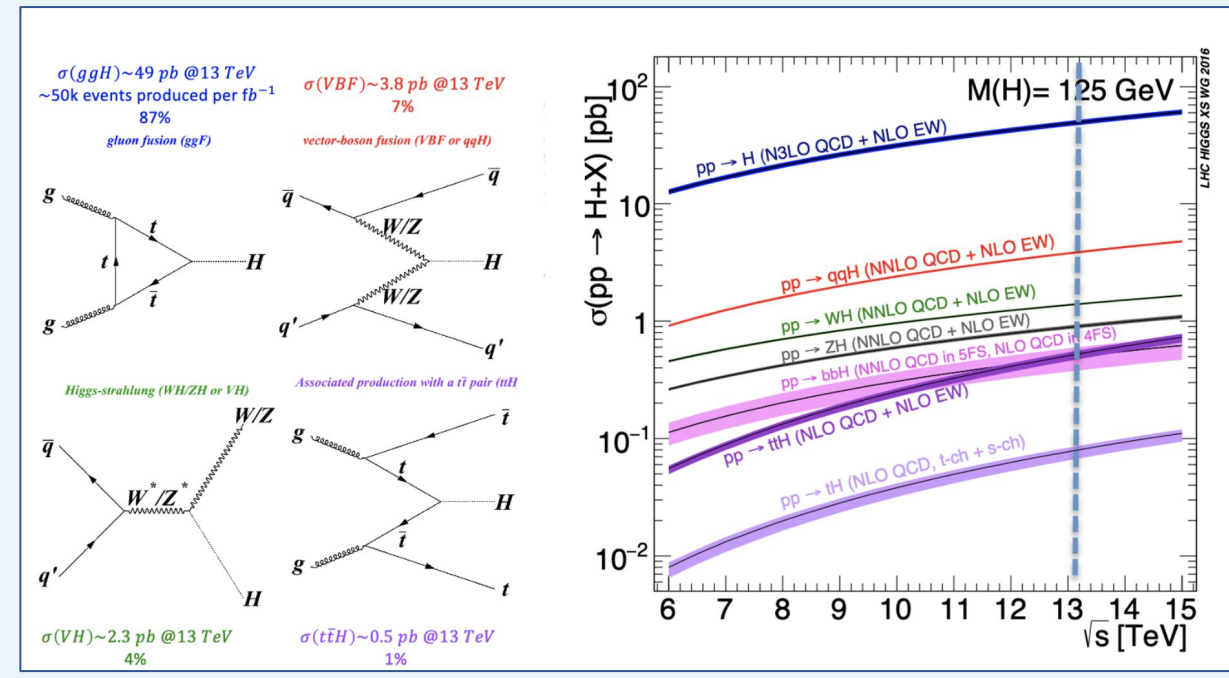
m l



Brief description of the exercise for the High Energy Physics (HEP) groups

# Binary classification task @CMS

- ❑ Selection of single **Higgs boson** via vector boson fusion (VBF) production mechanism **event** with mass hypothesis of 125 GeV in the 4mu/4e decay channel **@CMS Experiment @LHC** (Monte Carlo simulated sample at generator level)
- ❑ Irriducible backgrounds:
  - $gg \rightarrow ZZ \rightarrow 4\mu + \text{jets}$
  - $qq \rightarrow ZZ \rightarrow 4\mu + \text{jets}$
  - $WH \rightarrow qqZZ \rightarrow qq4\mu$  (optional exercise)
  - $ttH \rightarrow ttZZ \rightarrow qq4\mu$  (optional exercise)



You will learn more about the Standard Model of particle physics!

First seen [here](#).

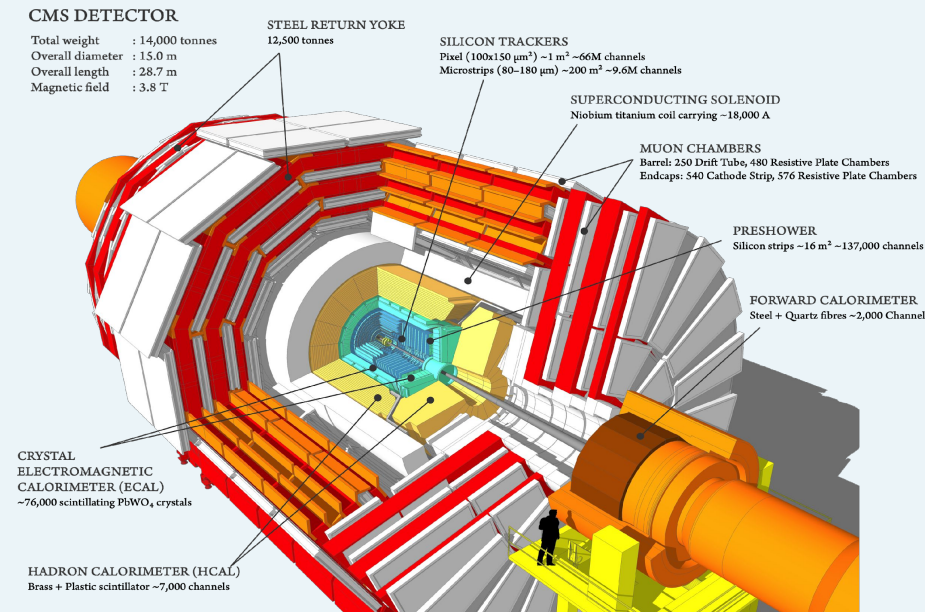
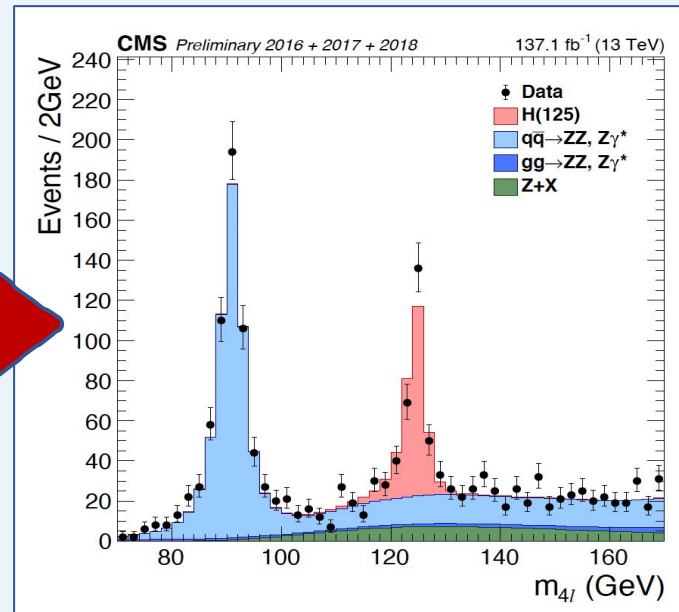
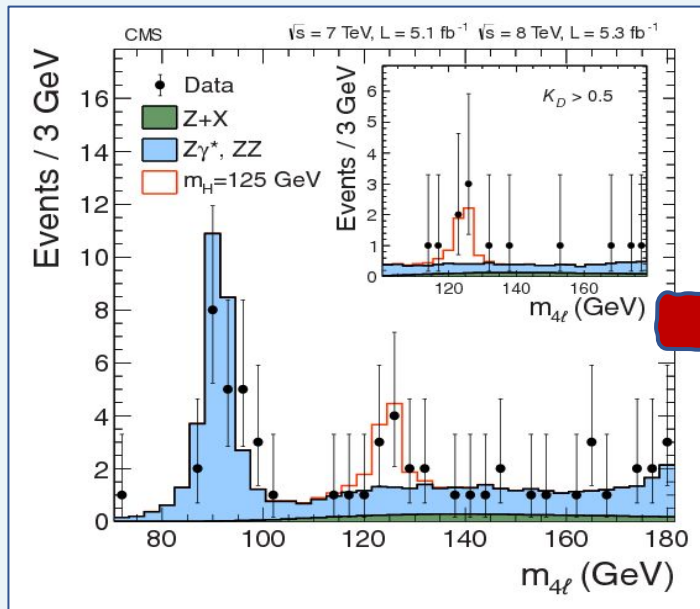
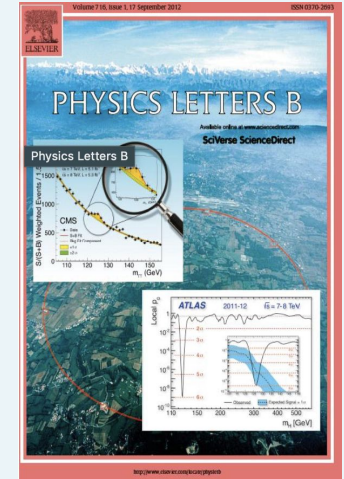


- ❑ The Higgs boson signal is a «rare» physical process @LHC, difficult to discriminate from the background with standard multivariate analysis techniques (i.e. by imposing cuts on single physical observables)
- ❑ We use **Machine Learning algorithm**, a **Deep Neural Network (DNN)**, for our multivariate analysis!

# The Compact Muon Solenoid (CMS) m l

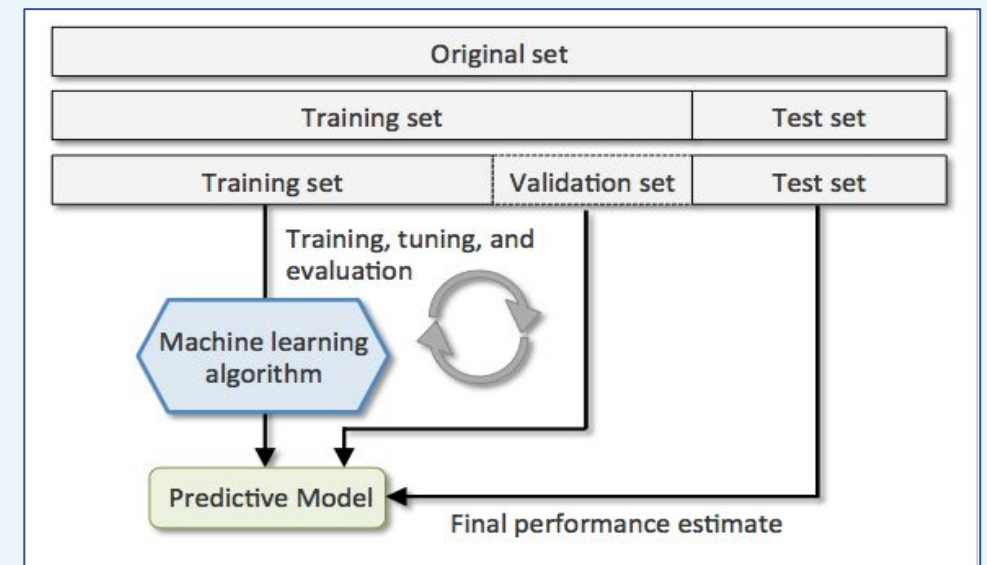
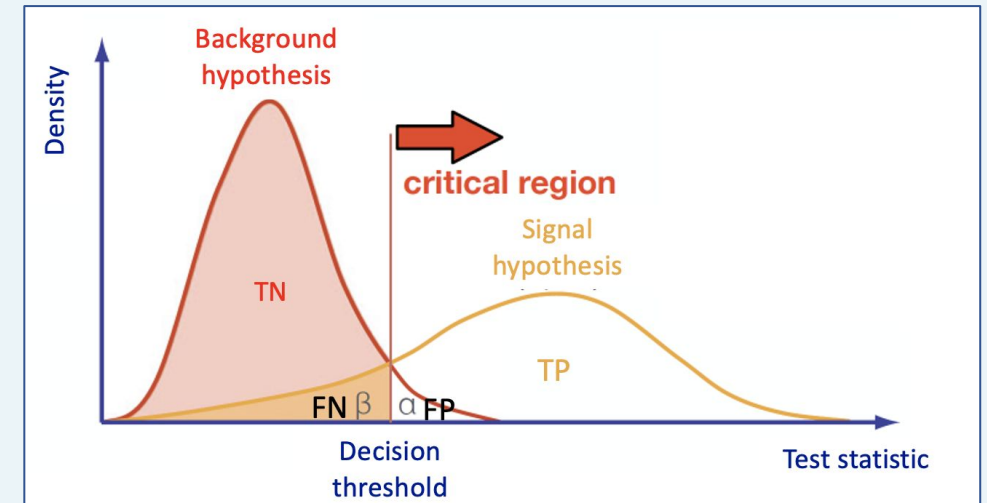


- One of the two general-purpose experiments which detected the Higgs boson in **2012 @ Large Hadron Collider (LHC), CERN.**
- We will use **2018 MC data-set** with which, after performing the binary classification task, you will be able to produce plots of physical observables (invariant mass, phi, eta distributions) for the Higgs signal and the main backgrounds as an actual particle physicist!



# Hypothesis test and ML data-sets

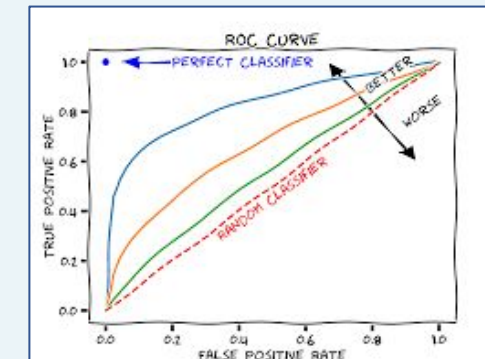
- In a **Multivariate Analysis (MVA)**, the MVA algorithm is fed by a set of discriminating variables (input) which are combined to reach an optimal discrimination power between two categories (**signal** and **background**).
- The discriminant output, also called **discriminator, score**, or **classifier**, is used as a test statistic and then adopted to perform the signal selection.
- The classifier can be used as a variable on which a cut can be applied under a particular hypothesis test.
- Machine Learning tools are models which have enough capability to define their own internal representation of data to accomplish two main tasks: **learning from data** and **make predictions** without being explicitly programmed to do so.



- Evaluation metrics are used to measure the quality of our machine learning model. There are many different types of evaluation metrics available to test a model. These include the **area under the Receiver Operating Characteristic – TPR vs FPR - curve (AUC)**, **confusion matrix, accuracy**, and others.

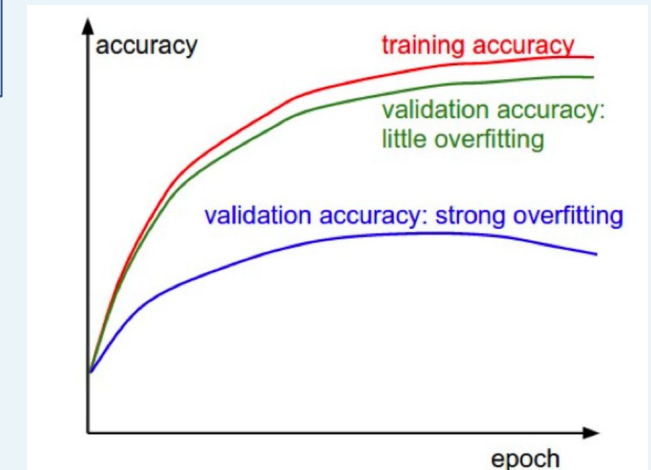
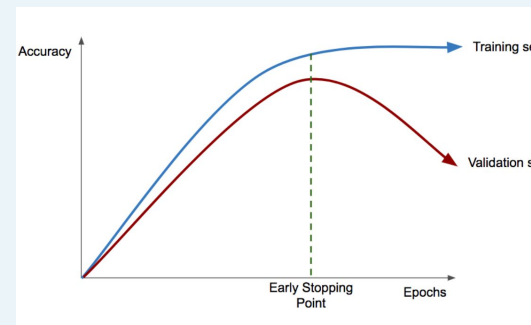
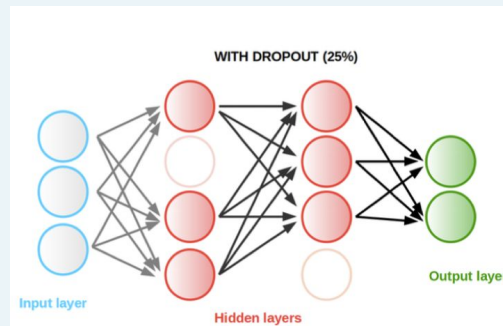
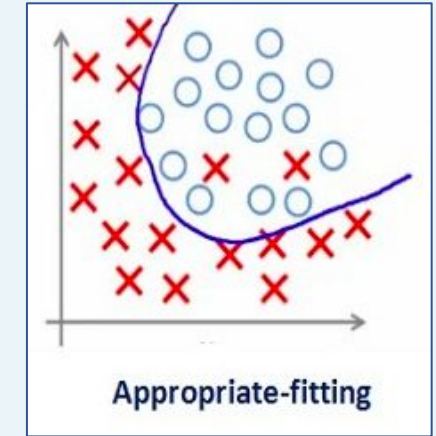
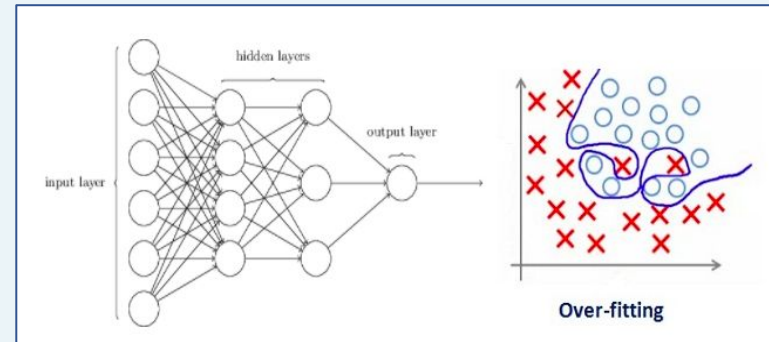
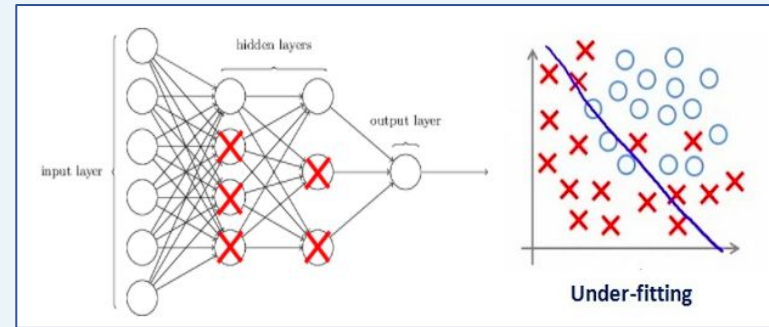
		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) <b>Type II Error</b>	<b>Sensitivity</b> $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) <b>Type I Error</b>	True Negative (TN)	<b>Specificity</b> $\frac{TN}{(TN + FP)}$
		<b>Precision</b> $\frac{TP}{(TP + FP)}$	<b>Negative Predictive Value</b> $\frac{TN}{(TN + FN)}$	<b>Accuracy</b> $\frac{TP + TN}{(TP + TN + FP + FN)}$

- It is extremely important to use **multiple evaluation metrics** to evaluate your model. This is because a model may perform well using one measurement from one evaluation metric, but may perform poorly using another measurement from another evaluation metric.



**NOTE:** Precision == purity; recall==sensitivity == TPR == signal efficiency; FPR = FP/(FP+TN)

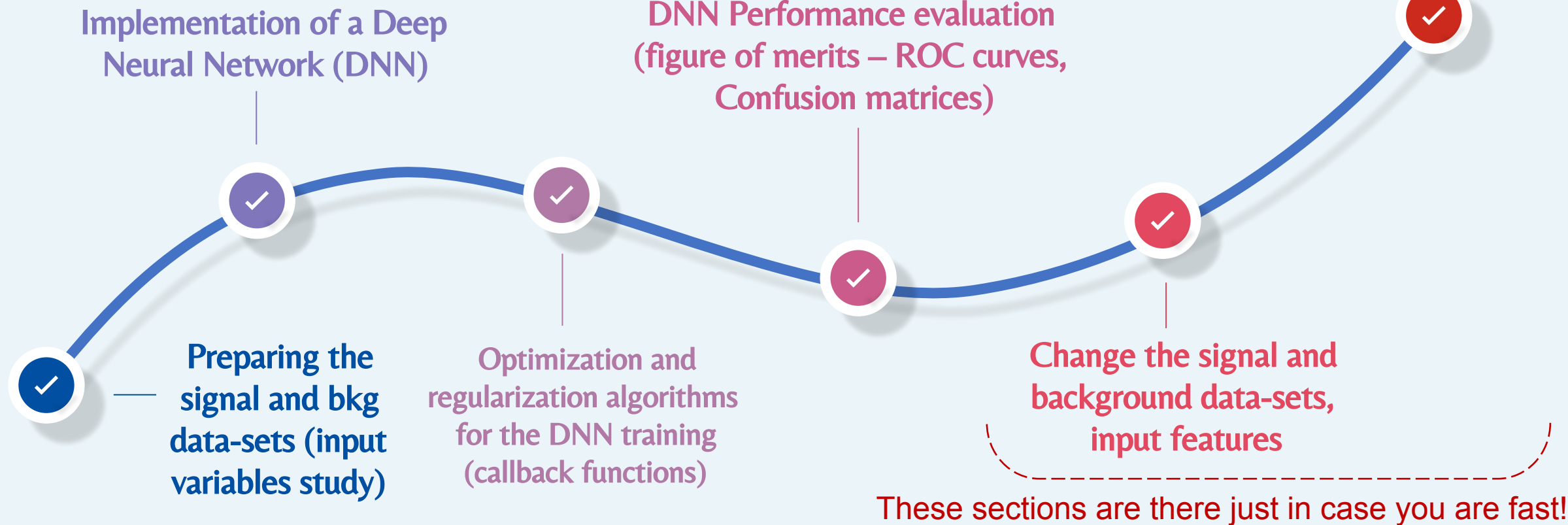
- One of the most common problems for ML technique users is the **overfitting**.
- The overfitting takes place when the model performs exceptionally well on train data but it is not able to predict unseen data .
- We overcome overfitting problem by using algorithms such as the Dropout or Early Stopping regularization techniques.



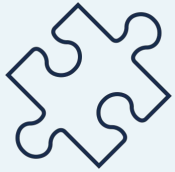
# Guideline to the exercise

- ❑ We propose you some requests throughout a **unique short exercise**
- ❑ Solving the requests is not mandatory to run the whole exercise and reach last cells of code where physical results are shown.
- ❑ We suggest to **complete the tasks by following their order** (which follows an increasing order of difficulty)

Improvement of a ML algorithm performance in terms of ROC curve (ML challenge)



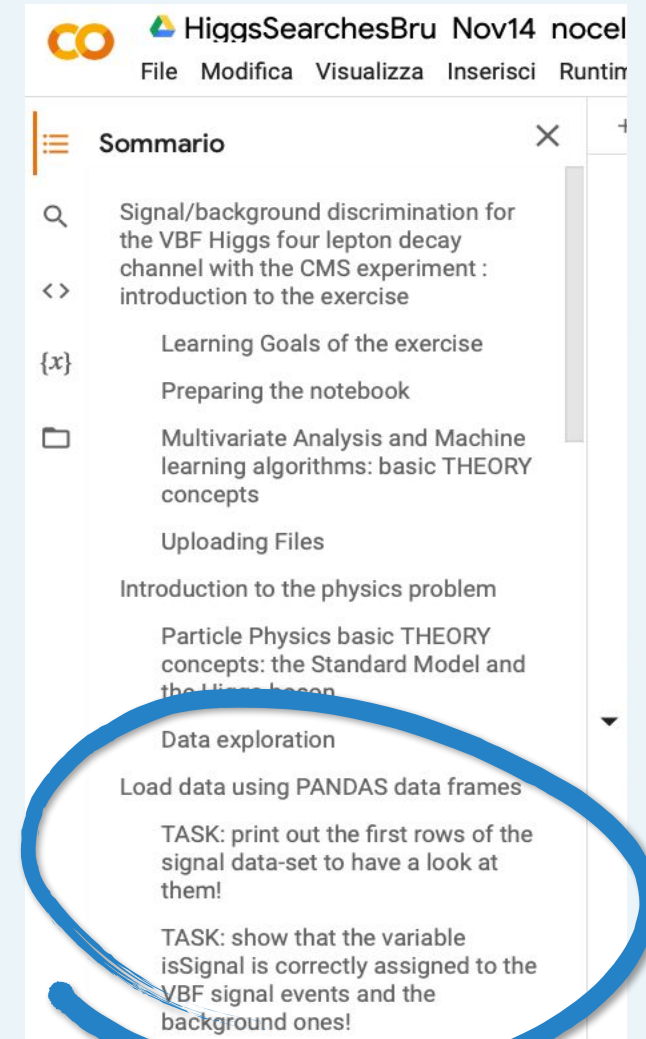
# Guideline to the exercise



- You will find in the colab shared area the notebook **4th\_MLHackathon2023\_nocelloutput\_students.ipynb**
- **The notebook contains a lot of code** already implemented (along with particle physics, statistical and machine learning theory support) in order to help you preparing the data samples and making plots by using scientific libraries (**scikit-learn, Keras, matplotlib**)
- When you find «**TASK for you**» in the exercise, you need to complete parts (from 1 to 10 lines of codes usually)
- You will have **some questions** and a **final Machine Learning challenge** where you have simply to upload your improved ML algorithm on the link you find at the end of the exercise!

Upload your results here:

<https://recascloud.ba.infn.it/index.php/s/CnoZuNrlr3x7uPI>



co HiggsSearchesBru Nov14 nocel  
File Modifica Visualizza Inserisci Runtin

Sommarrio

- Signal/background discrimination for the VBF Higgs four lepton decay channel with the CMS experiment : introduction to the exercise
- Learning Goals of the exercise
- Preparing the notebook
- Multivariate Analysis and Machine learning algorithms: basic THEORY concepts
- Uploading Files
- Introduction to the physics problem
- Particle Physics basic THEORY concepts: the Standard Model and the Higgs boson
- Data exploration
- Load data using PANDAS data frames

TASK: print out the first rows of the signal data-set to have a look at them!

TASK: show that the variable isSignal is correctly assigned to the VBF signal events and the background ones!



# Workgroup and reporting

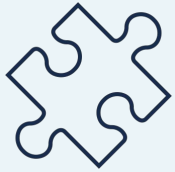
- You are supposed to strongly interact with one another and collaborate in order to come to a shared solution.
- Please do not hesitate to ask for help to the tutors: Brunella and Walaa.
- For the afternoon session, we will discuss together the issues that you found while running the exercise so don't be shy ;) and report to us any difficulties you faced.



Ph. D. in CMS  
collaboration @ UniBA



Dr. in CMS collaboration  
@ UniBA



- General ML library (Python):

- 1 <https://scikit-learn.org/stable>

- Deep learning libraries:

- 1 <https://www.tensorflow.org> ( Google )

- 2 <https://pytorch.org> ( Facebook )

- 3 <https://www.microsoft.com/en-us/cognitive-toolkit> ( Microsoft )

- 4 <https://mxnet.apache.org> ( Apache )

- 5 <https://github.com/Theano/Theano> ( Univ.Montreal )

- High level deep learning API:

- 1 <https://keras.io> ( Tensorflow ,CNTK,Theano)

- 2 <https://docs.fast.ai> ( Pytorch )

- Converting ROOT trees to Python numpy arrays or panda data frames

- 1 [https://github.com/scikit-hep/root\\_numpy](https://github.com/scikit-hep/root_numpy)

- 2 [https://github.com/scikit-hep/root\\_pandas](https://github.com/scikit-hep/root_pandas)