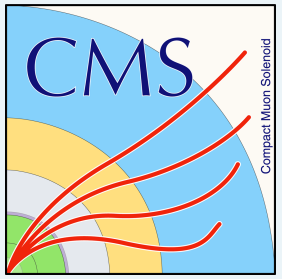
Higgs searches @LHC exercise: Signal/background discrimination for the VBF Higgs four lepton decay channel with the CMS experiment

B. D'Anzi[1,2] , N. De Filippis[3,2] , D. Diacono[1,2] , G. Miniello[1,2] , W. Elmetenawee[1,2], A. Sznajder[4]

[1]University of Bari; [2]INFN BARI; [3]Politecnico of Bari; [4]University of Rio de Janeiro

Second ML-INFN Hackathon

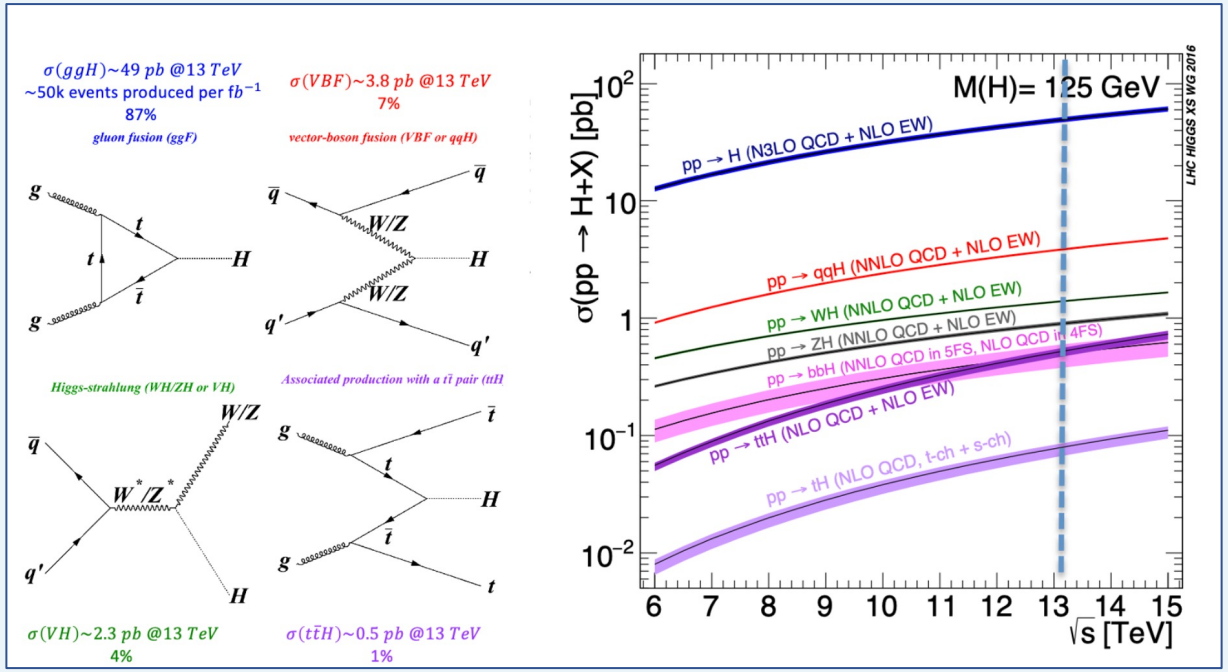
15th November 2021



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 **events** with mass hypothesis of 125 GeV in the four muons / four electrons (optional exercise) decay channel (4mu/4e) @CMS experiment (Monte Carlo simulated samples at generator level) ;
- **Irriducible backgrounds** : $gg \rightarrow H \rightarrow ZZ \rightarrow 4mu$, QCD $ZZ \rightarrow 4mu$, $ttbarH \rightarrow ZZ \rightarrow 4mu$ (optional exercise) .



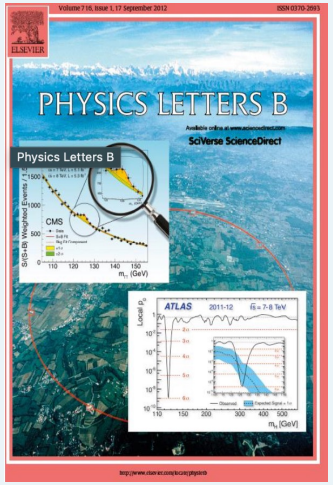
You will learn more about the Standard Model of particle physics!

First seen here.

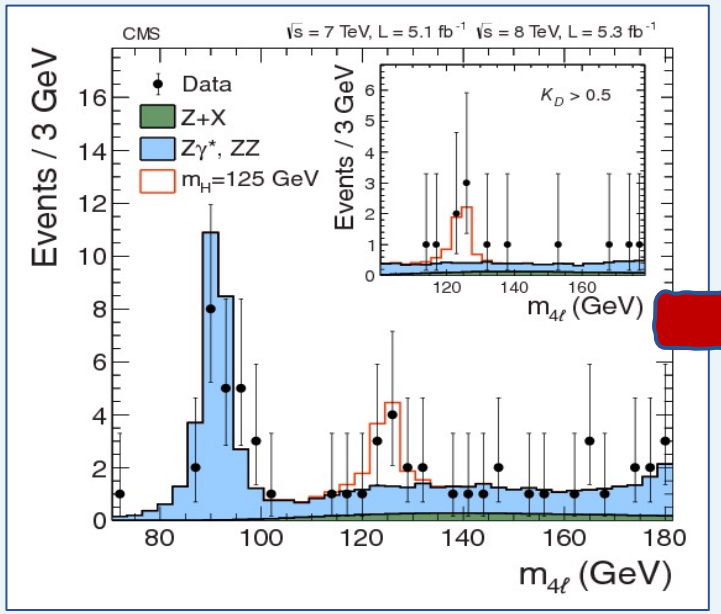
- The Higgs boson signal is a «rare» physical process @LHC, difficult to discriminate from the background with standard multivariate analysis techniques (i.e. by imposing cuts on single physical observables)
- We use **Machine Learning algorithm,** a **Deep Neural Network (DNN),** for our multivariate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

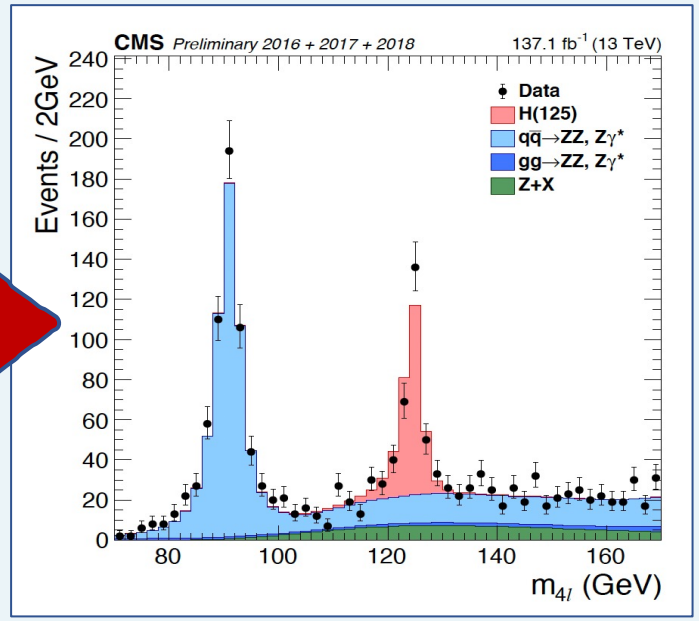
Higgs searches @ LHC: Aim of the hackathon exercise

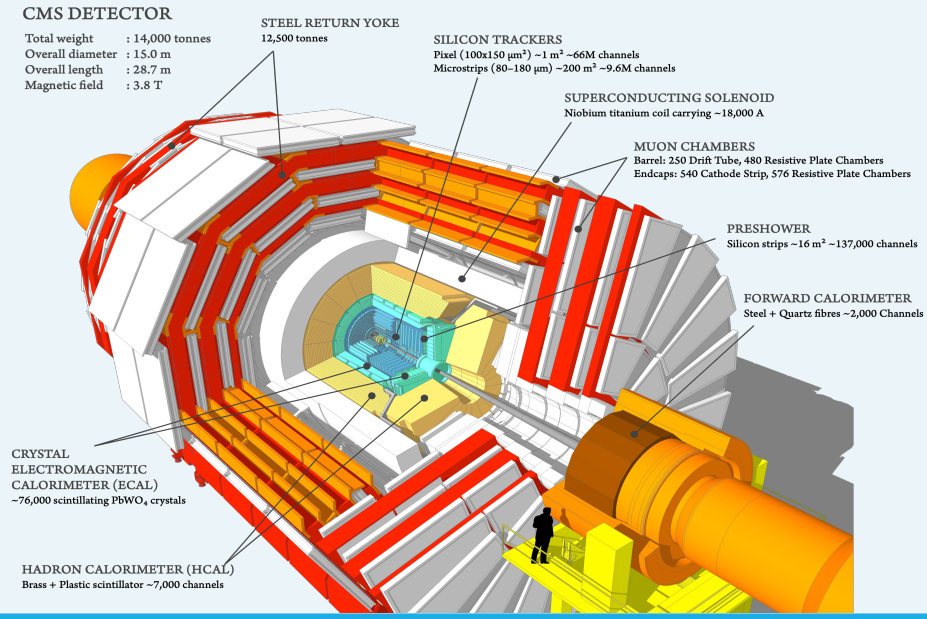
The Compact Muon Solenoid (CMS)



- 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!





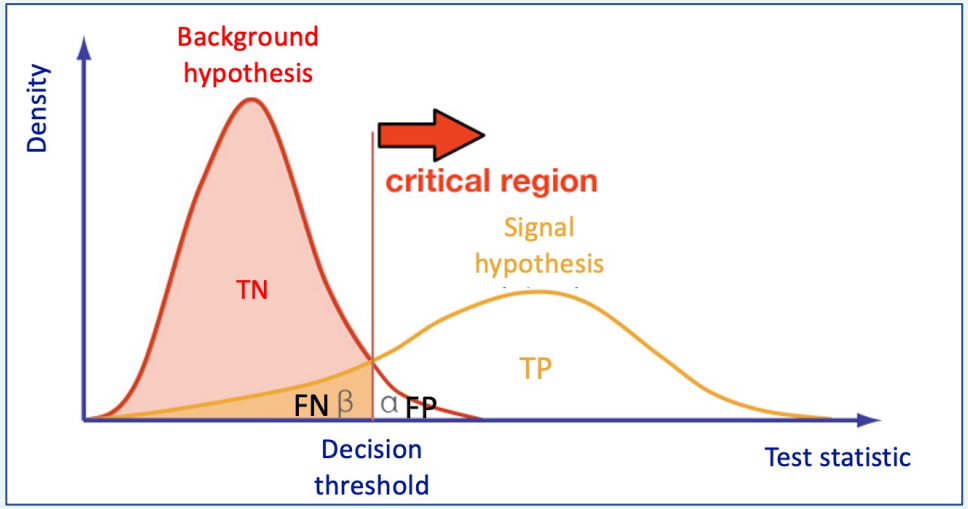


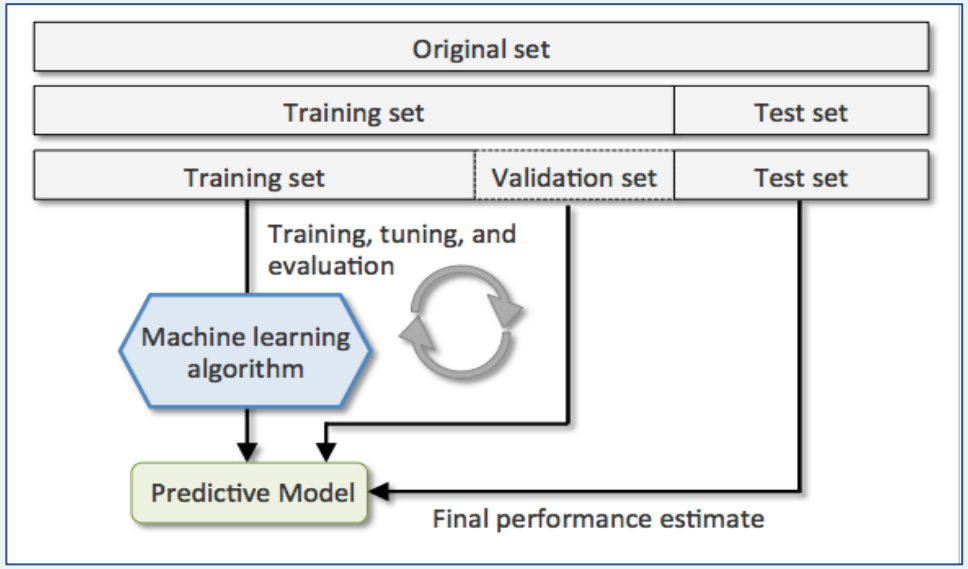
Hypothesis test and ML data-sets

- Multivariate Analysis algorithms receive as input a set of discriminating variables. Each single variable does not allow to reach an optimal discrimination power between two categories (**signal** and **background**). Therefore the algorithms compute an output that combines the input variables. This is what every **Multivariate Analysis (MVA) discriminator** does.

- The discriminant output, also called ***discriminator, score*** , *or* ***classifier***, is used as a test statistic and is then adopted to perform the signal selection. It could be used as a variable on which a cut can be applied under a particular hypothesis test.

- In particular, 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.



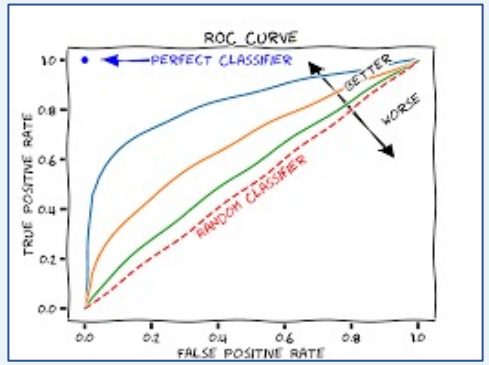


Evaluation Metrics

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

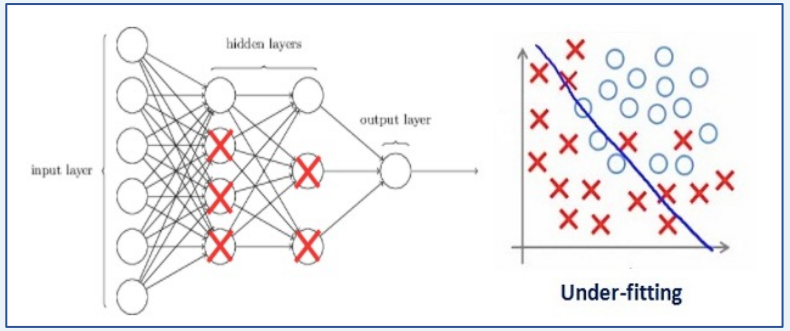
- It is extremely important to use **multiple evaluation metrics** to evaluate your model. This is because a model may perform well using one measurement from one evaluation metric, but may perform poorly using another measurement from another evaluation metric.

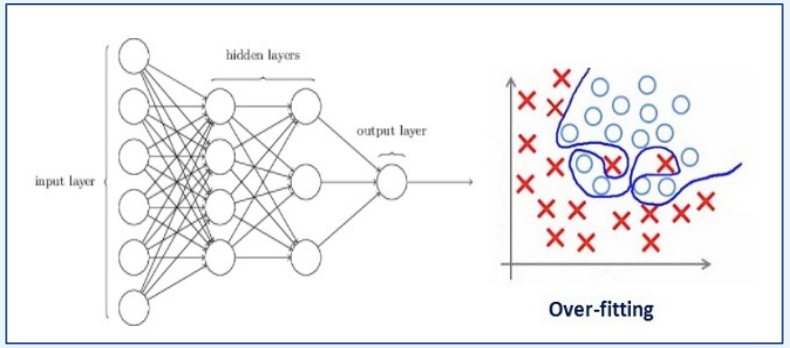


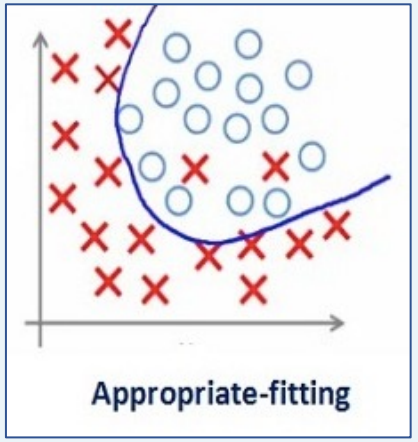
NOTE: Precision == purity; recall==sensitivity == TPR == signal efficiency; FPR = FP/(FP+TN)

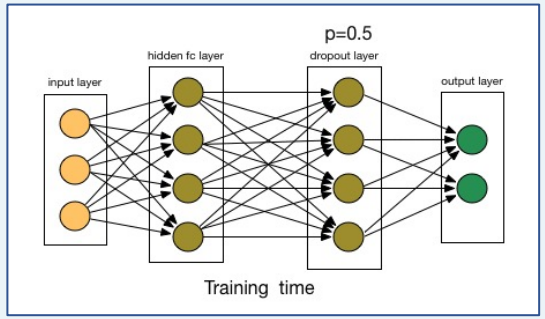
Optimization and regularization algorithms

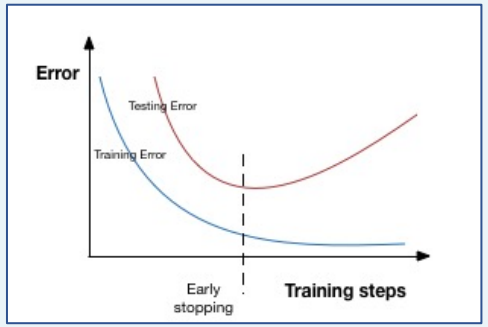
- One of the most common problems ML techniques users face is to avoid **overfitting**. You have to come across this problem when your model performs exceptionally well on train data but it is not able to predict test data.
- We will overcome overfitting problem by using algorithms such as the Dropout or EarlyStopping regularization techniques.

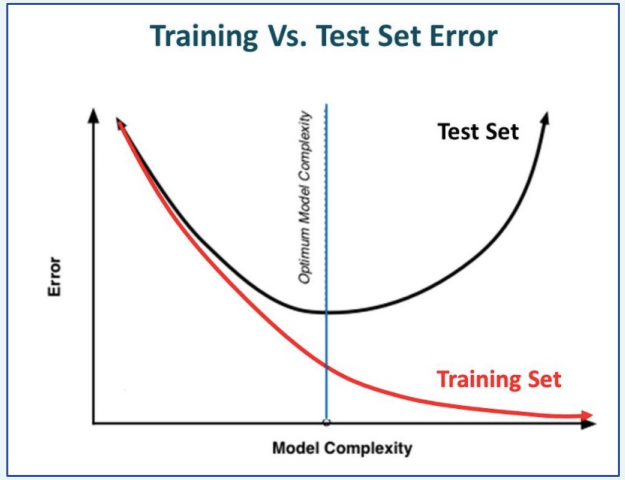






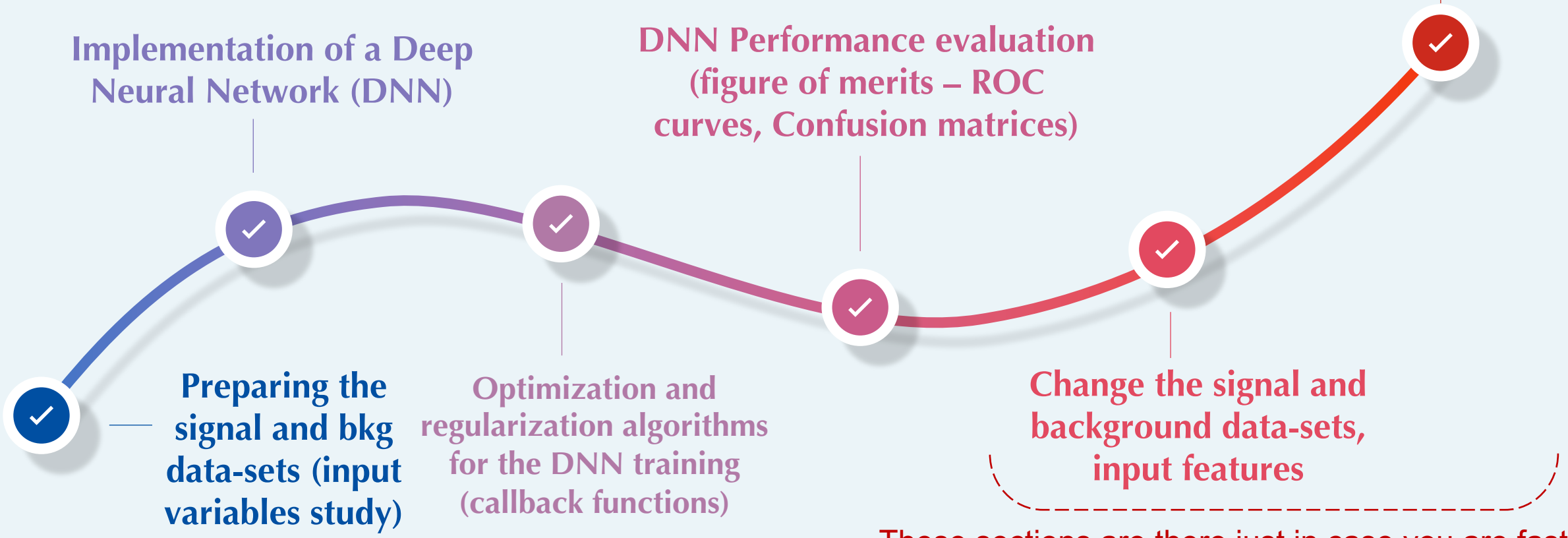






Guideline to the exercise (1)

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

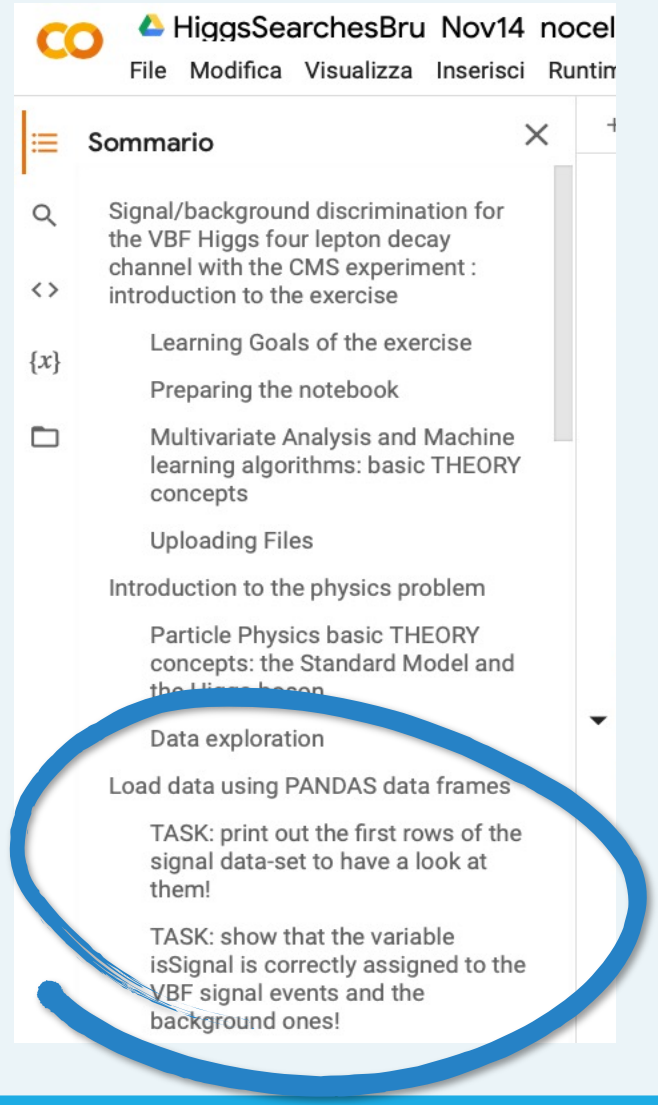


Guideline to the exercise (2)

- You will find in the colab shared area the notebook **HiggsSearchesBru_Nov14_nocelloutput_students.ipynb**
- It contains already **a lot of code** (and particle physics, statistical and machine learning theory support), especially to help you in preparing the data samples and making plots by using scientific libraries (**scikit-learn, Keras, matplotlib**)
- You will need to complete parts (from 1 to 10 lines of codes usually) where a **TASK for you** is outlined. They will test your acquired knowledge in **Deep Learning/Machine Learning** and use of **Jupyter Notebook** by using the python programming language!
- You will have **some questions** and a **final Machine Learning challenge** where you have simply to upload your improved ML algorithm on the link which we indicate to you at the end of the exercise!

Upload your results here:

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



Workgroup and reporting

- You are supposed to interact with each other and collaborate to arrive to a shared solution.
- For the afternoon session, we would like to see some results and/or if it was impossible to solve the problem. Please use the template here (one for each group) for guidance.
- Please do not hesitate to ask for help to the tutors: Brunella, Giorgia Domenico/Nicola and Walaa.

Ph.D. student in CMS collaboration @ UniBA