Domain Adaptation to train model-independent classifiers in High Energy Physics

Third ML-INFN Hackathon: Advanced Level

21-24 November 2022
Bari (Italy)

Lorenzo Viliani - INFN Firenze
Benedetta Camaiani - UniFi & INFN Firenze
Piergiulio Lenzi - UniFi & INFN Firenze

INFN

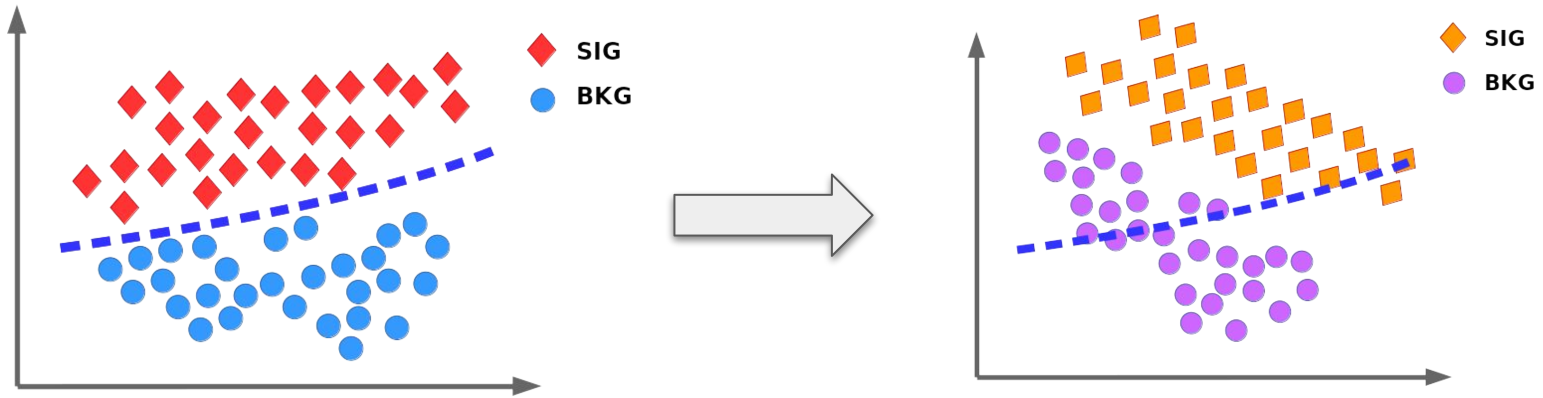
Introduction

- Domain Adaptation (DA) is a particular case of transfer learning (TL) that leverages labeled data in one or more related source domains, to learn a classifier for unseen or unlabeled data in a target domain.
- Source and Target domains are assumed to be related → distributions of source and target data are not completely different.
- Goal: train a NN in one dataset (source), securing good performance and accuracy in a different dataset (target).
- Different ways to achieve DA, unsupervised or (semi-)supervised:
  - we will focus on an Adversarial Deep Learning approach;
  - main idea: find a representation space that is common to source and target domains.

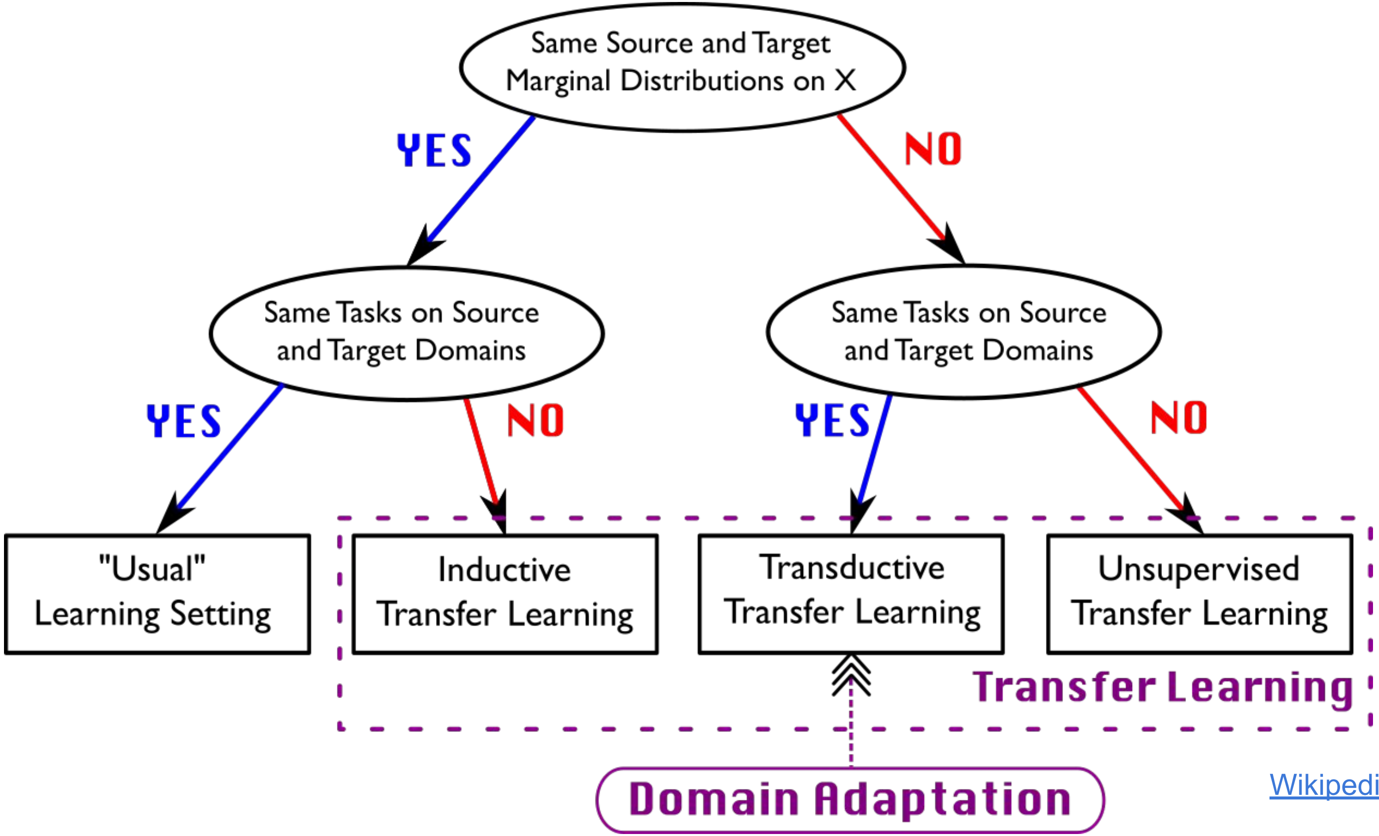
Some definitions

- **Domain**: a feature space $\mathbf{X}$ + marginal probability distribution P(x), with $x = \{x_1,...,x_n\}$ in $\mathbf{X}$
- **Task**: a label space Y and a function f: X->Y used to predicted the label y given the input x
- **Domain shift**: change in the data distribution between training and deployment.
  - Common AI algorithms do not perform well when domain shifts are present.

Source Domain **Target Domain**



Domain adaptation classes



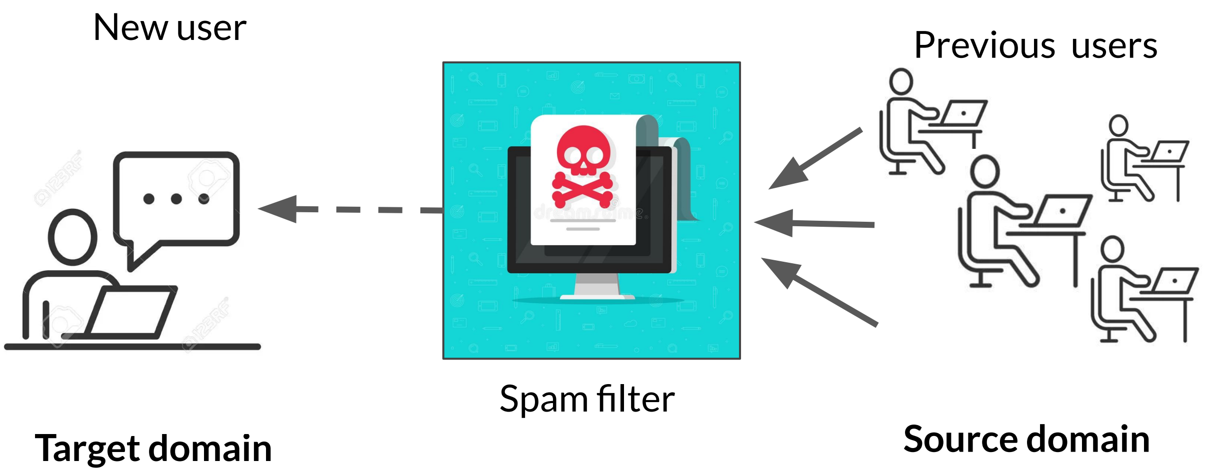
[Wikipedia (CC-BY-SA 4.0)]

Domain Adaptation

- Supervised Learning:
  - X input space
  - Y output space (or label space)
  - $(x_i,y_i) \in S$ i.i.d. from a distribution D_S (unknown and fixed) of support $X \times Y$
  - goal: learn $h: X \rightarrow Y$ from S such that it commits the least error possible for labelling new examples coming from D_S

- **Domain Adaptation**:
  - two different (but related) distributions D_S (source domain) and D_T (target domain) on $X \times Y$
  - objective: learn h from the two domains such that it commits as little error as possible on the target domain D_T

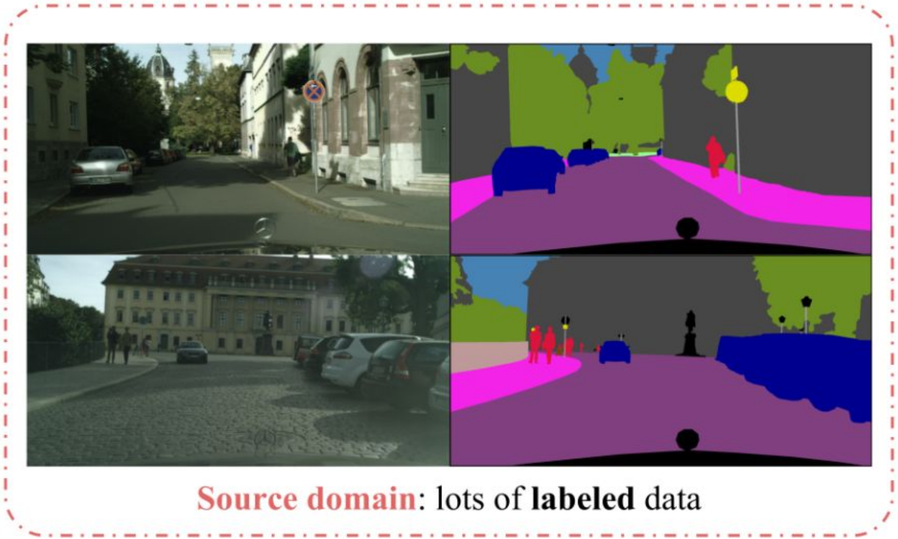
Examples of DA: spam filter

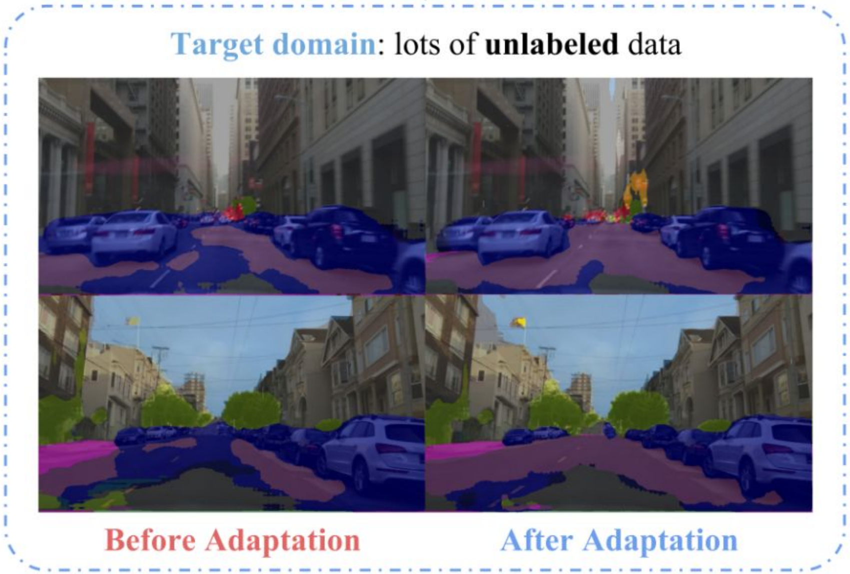


- Assumption: users generally agree on what is spam and what is not

The challenge is that the distributions of emails for the first set of users and for the new one are different

Examples of DA in computer vision



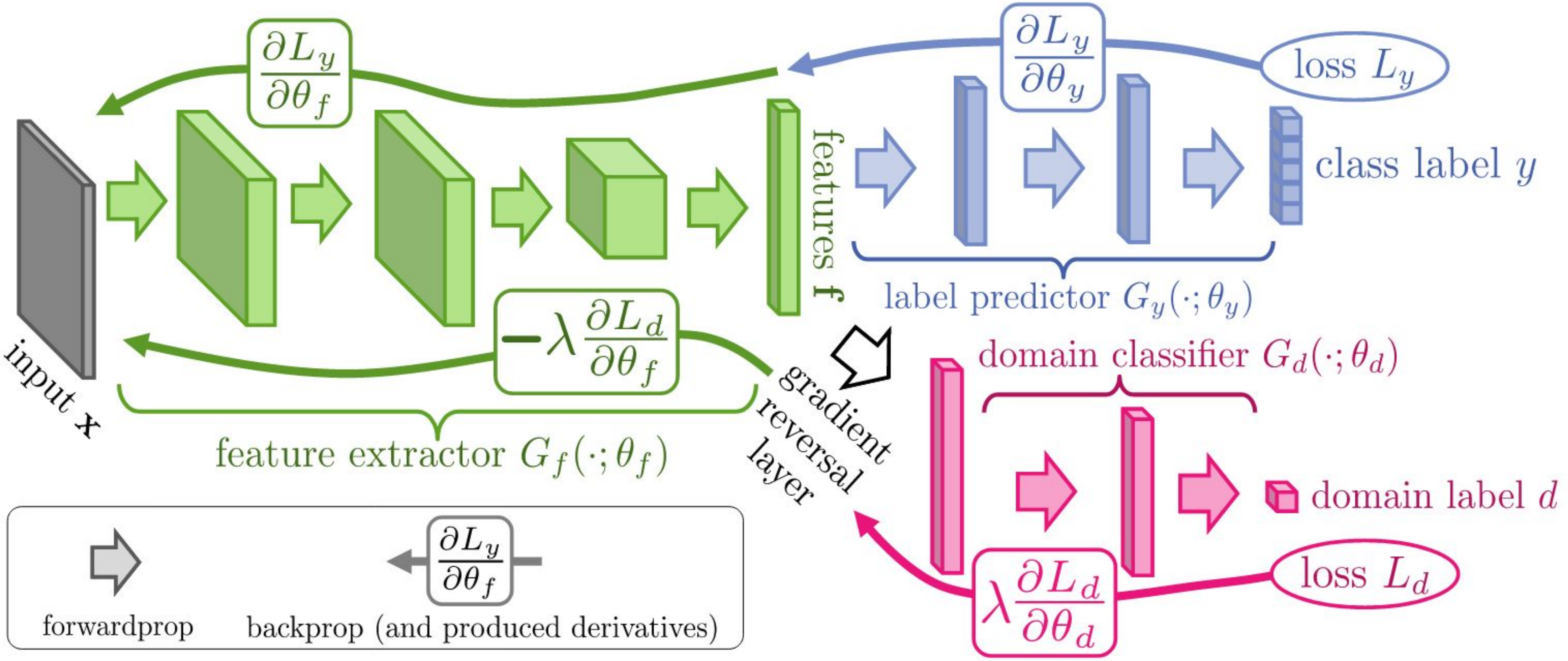


[arXiv:1612.02649]

DA with an adversarial approach

- We will focus on DA obtained with an adversarial approach…

[arXiv:1505.07818](https://arxiv.org/abs/1505.07818)



Application of DA in HEP experiments

The High Energy Physics case

- Usage of ML in HEP increased immensely in the last decade.
    - particle reconstruction;
    - event classification;
    - anomaly detection.
- For measurements and searches, the ML algorithms are commonly trained using synthetic data corresponding to the best current knowledge of the Standard Model (SM) physics processes.
    - Although the SM is pretty well established, this approach could be problematic because we don’t know which is the real model chosen by Nature.
- Example: which signal do we use in the training, if we are searching for/measuring a unknown/not well known physics process?

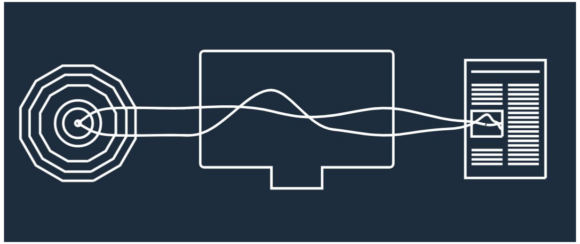
DA and re-interpretability

- HEPData is an open-access repository for High Energy Physics (HEP) data, that comprises results related to several thousand publications, including those from the Large Hadron Collider (LHC).

- Let's imagine someone in the future who wants to re-interpret a result from HEPData and compare it with a new physics model.

- If the analysis was performed with some physics model assumptions - which is often the case - the result will be biased towards that model, spoiling the re-interpretability of the measurement and leaving our friend confused.

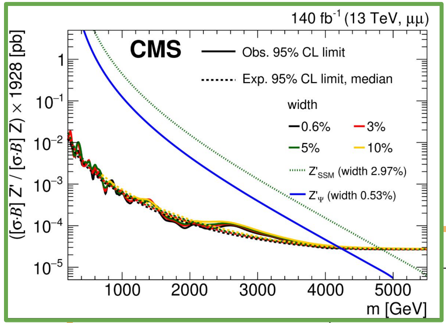
HEPData



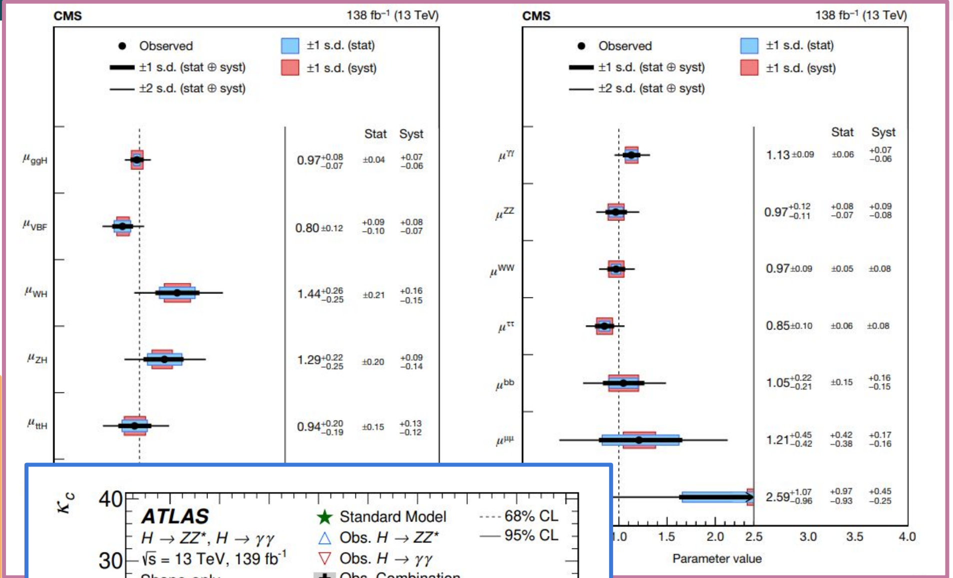
Our goal is to design a measurement that can be easily re-interpreted, and thus limiting the model dependence as much as possible.

Some recent results

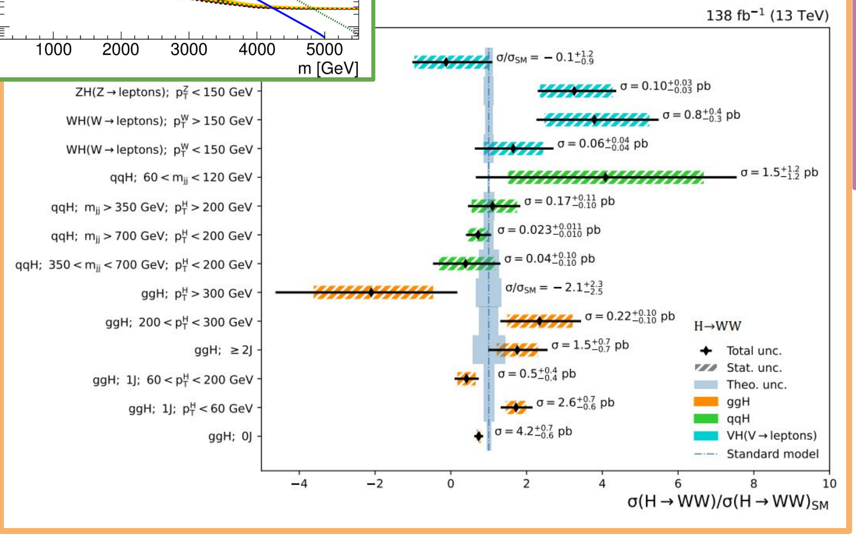
CMS,
Z' exclusion plot

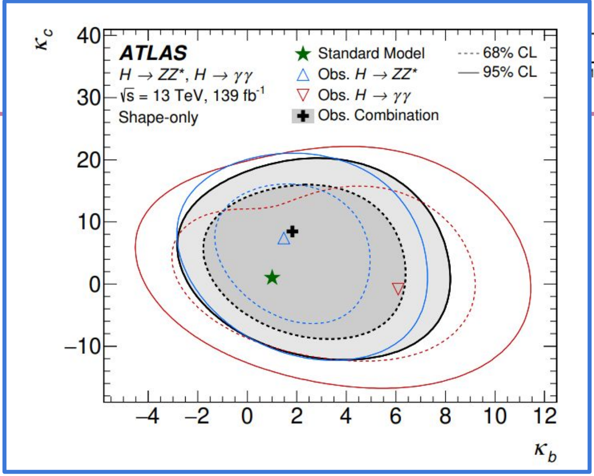


CMS,
measurement of the Higgs boson production modes



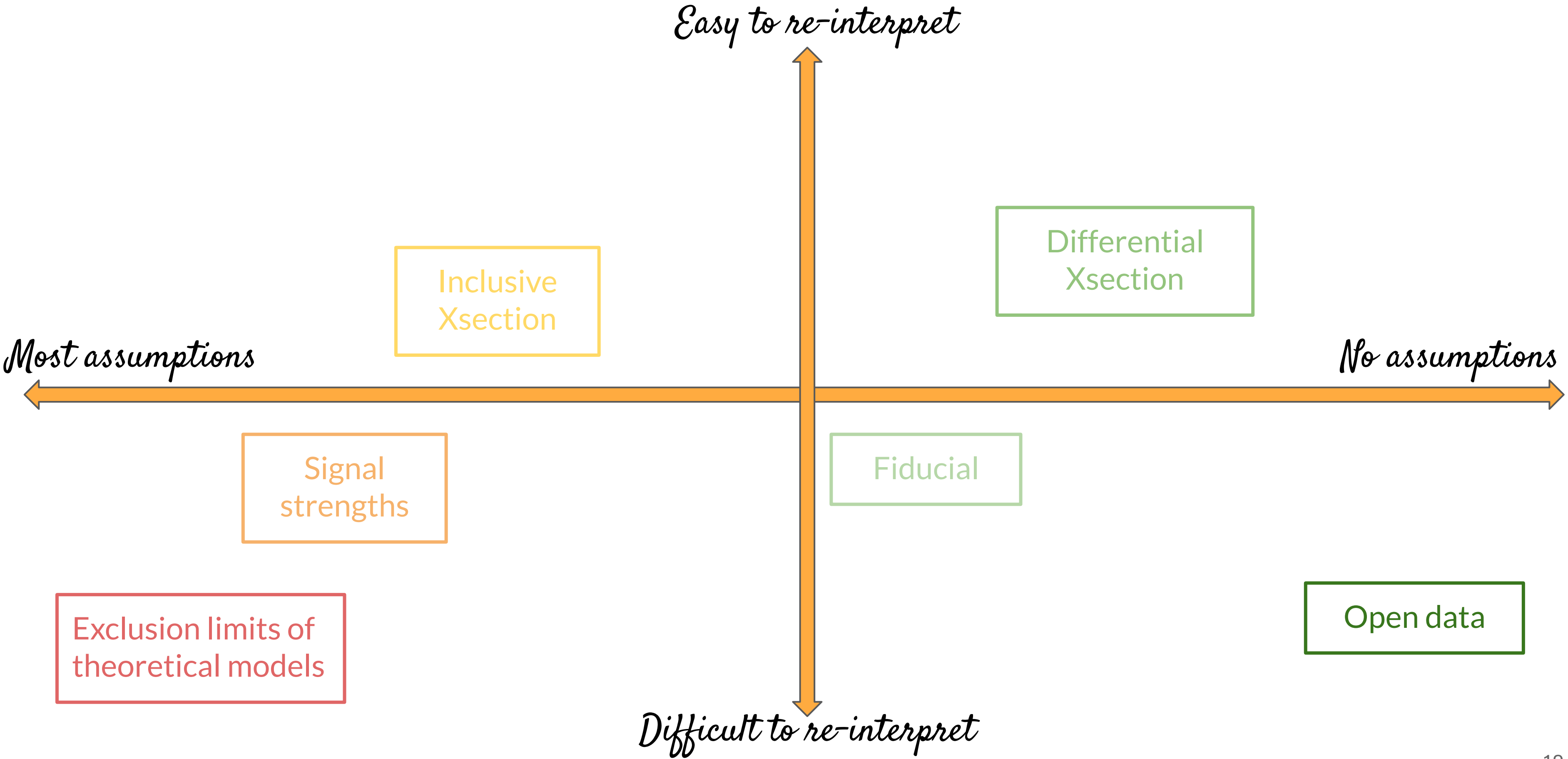
CMS,
STXS measurement





ATLAS,
constraints on Higgs boson couplings from diff XS.

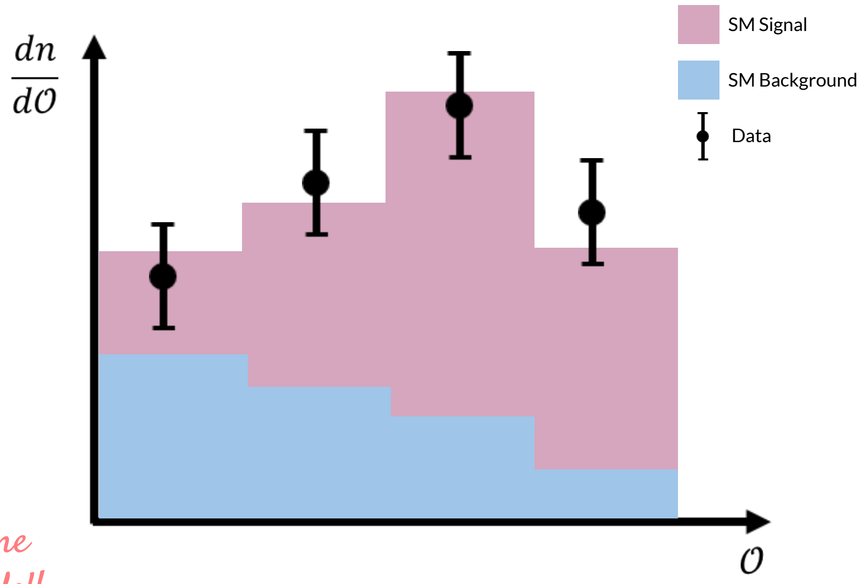
LHC “data” spectrum



Template fit

- Let $\mathcal{O}$ be an observable with good signal to background discriminating power
- The probability density function (PDF) is usually not known a priori and it is replaced by templates
- Monte Carlo simulations are produced for both signal and background process
- MC simulation of the signal is fitted to experimental data

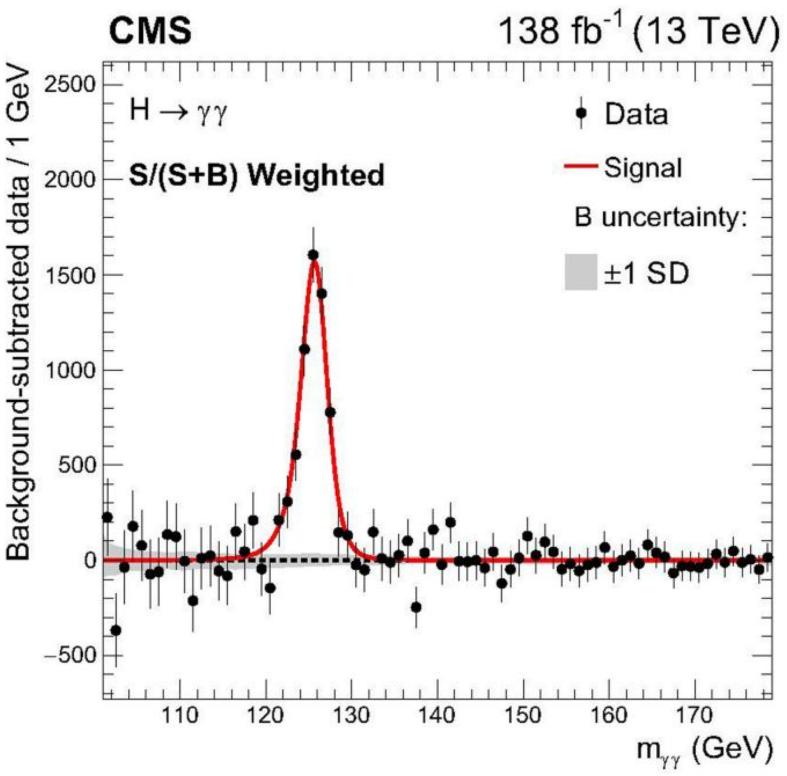
Need to assume a physics model!



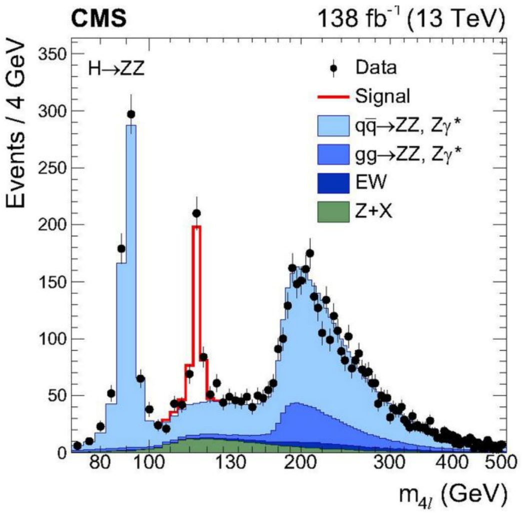
The shape of the observable distribution may in general depend on the properties of the physics theory governing the signal process under consideration

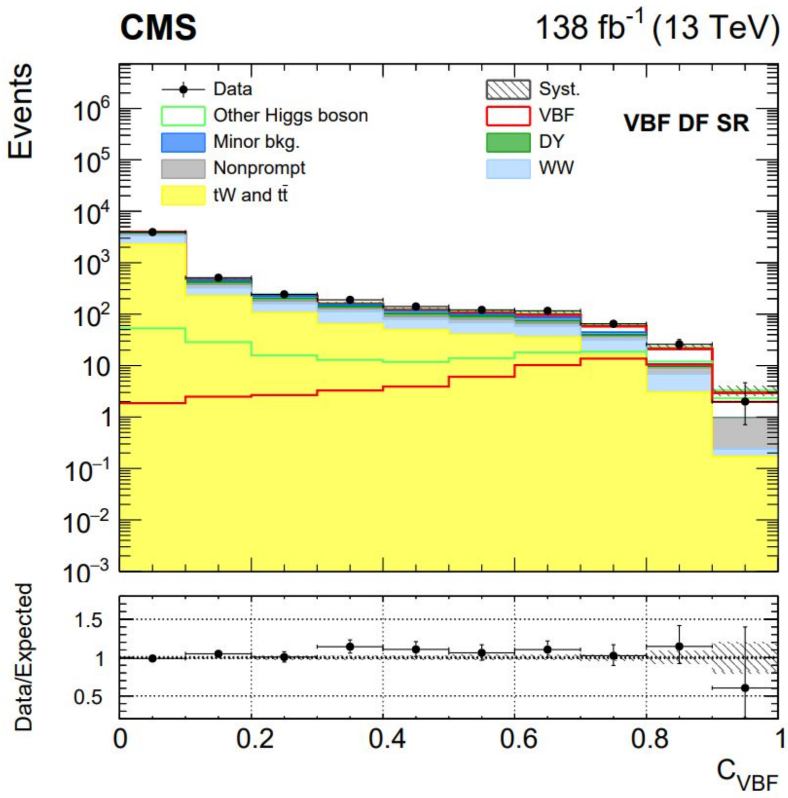
The fit result has a bias toward the prediction of the physics model used to generate the templates

H→γγ and H→ZZ



- The Higgs boson invariant mass can be reconstructed in H→γγ
- **The discriminating variable is model independent**
- No bias in the fit procedure
- Same for H→ZZ→4l



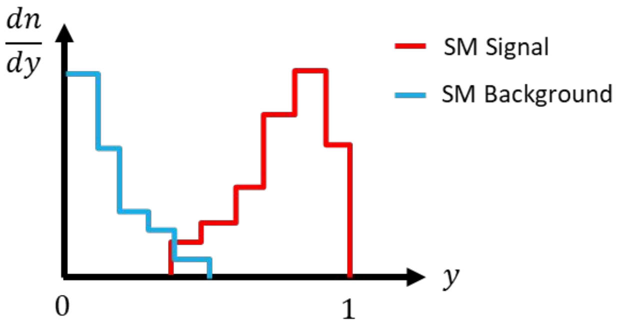


- Due to neutrinos, the Higgs invariant mass can not be reconstructed in H→WW→2l2v
- Deep Neural Network discriminant as fit variable
- Training set from Monte Carlo simulation
- **The shape of the discriminant depends on the physics hypothesis used to generate the training set**

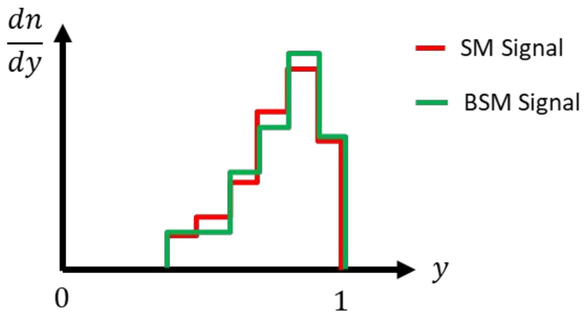
Main Objective

The main goal is to implement a new fit variable y that is agnostic with respect to the signal hypothesis:

- y must be able to discriminate signal from background events



- y must not be able to distinguish the physics model of signal events



y does not introduce a bias in the fit result since the shape of its distribution is roughly the same regardless of the theoretical model describing the data

How do we "adapt" our NNs? The use case of $H \rightarrow WW$

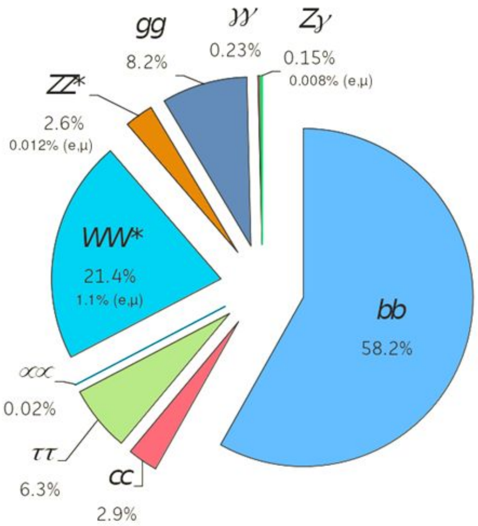
The H→WW channel

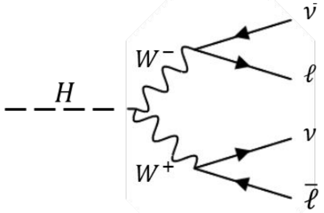
- The experimental sensitivity of a decay channel is determined by two factor: Branching ratio (BR) and composition of the final state

- The $H \to WW$ decay channel is suitable for measuring rare Higgs boson production modes and differential cross sections:
  - Second largest BR among the Higgs boson decays
  - Direct access to couplings with W bosons
  - Sensitive to the possible Higgs boson production mechanisms

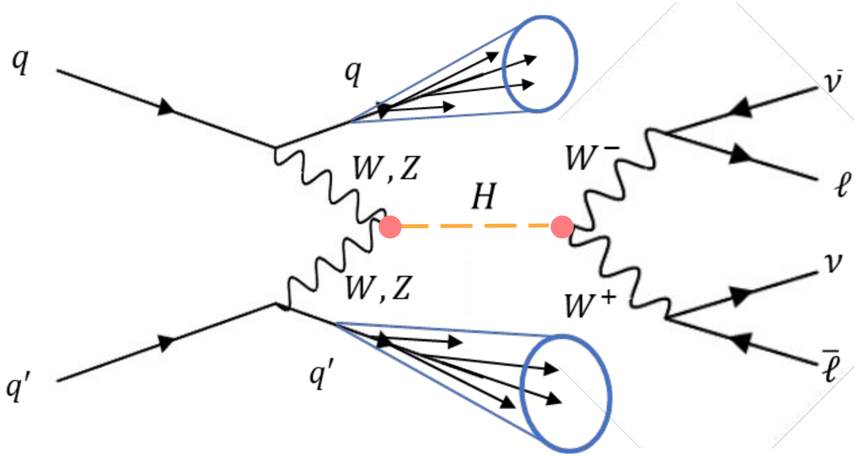
- Besides, considering the purely leptonic decay of the two Ws:
  - Low background contamination wrt hadronic channels
  - Charged leptons well reconstructed by the CMS detectors
  - No access to the full kinematics of the Higgs boson due to neutrinos

BRs for the Higgs boson decay modes





Signal process



Vector boson fusion (VBF)

- Second most probable Higgs boson production mode at the LHC
- Allows to test the SM predictions

Experimental signature in H→WW→2l2v

- ~ collinear charged (eμ) leptons due to a spin correlation effect
- missing transverse energy due to neutrinos
- $m_H < 2m_W \Rightarrow 1\ W$ boson is virtual
- 2 hadronic jets with large pseudorapidity gap

Some BSM theories predict Anomalous Couplings (AC) in the HVV vertex

Anomalous couplings in HVV

Scattering amplitude of one spin-0 Higgs boson (H) and two spin-1 gauge bosons ($V_1\ V_2$)

$$A(HVV) \sim \Bigg[\boxed{\boldsymbol{a}_{\boldsymbol{1}}^{\boldsymbol{VV}}} + \underbrace{\boxed{\frac{k_1^{VV} q_{V1}^2 + k_2^{VV} q_{V2}^2}{(\Lambda_1^{VV})^2}}}_{\mathbf{L_1}} \Bigg] m_{V1}^2 \epsilon_{V1}^* \epsilon_{V2}^* + \boxed{\boldsymbol{a}_{\boldsymbol{2}}^{\boldsymbol{VV}} f_{\mu\nu}^{*(1)} f^{*(2)\mu\nu}} + \boxed{\boldsymbol{a}_{\boldsymbol{3}}^{\boldsymbol{VV}} f_{\mu\nu}^{*(1)} \tilde{f}^{*(2)\mu\nu}}$$

This general structure has 4 couplings:

- $a_1^{VV} \neq 0$ SM couplings $J^{CP}=0^{++}$

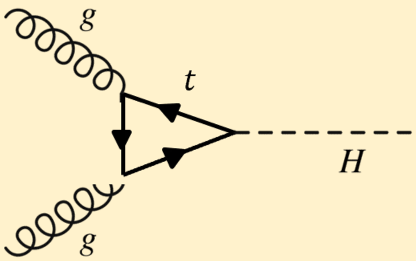
AC:

- $L_1 \neq 0$ $H{-}Vff$ or $H{-}ffff$ couplings predicted by **H0L1** model
- $a_2^{VV} \neq 0$ loop-induced ($HZ\gamma$, $H\gamma\gamma$, Hgg) CP-even coupling predicted by **H0PH** model
- $a_3^{VV} \neq 0$ three loop induced CP-odd coupling predicted by **H0M** model

+ **H0L1f05**, **H0PHf05** and **H0Mf05** theories which are mixtures between the SM and one of the previous model

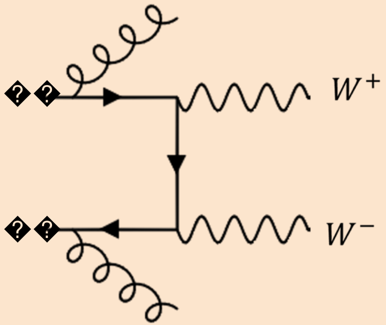
Main backgrounds

Gluon fusion (ggH)



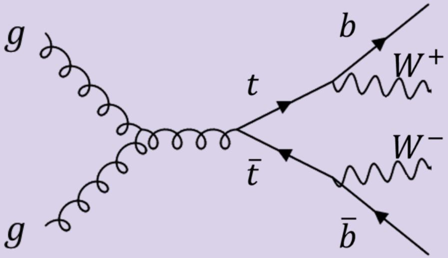
+ **2 jets from ISR**
- Main H production mode
- Large cross section

Non-resonant $W^+ W^-$



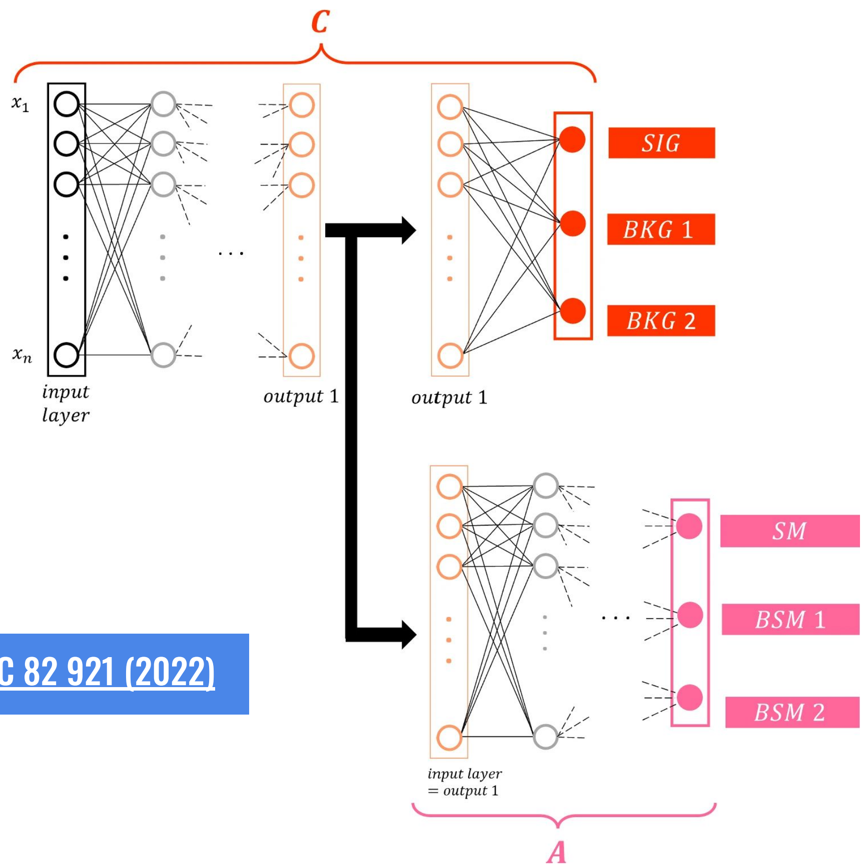
+ **2 jets from ISR**
- No spin correlation
- Different kinematics
- W bosons on-shell

Top pair production



- b-tagging algorithms to recognize b-jets
- Large cross section
- Dominant background

Adversarial deep neural network (ADNN)



[EPJC 82 921 (2022)]

Classifier

- Takes as input the measurable kinematic variables of an event
- Aims to determine if the event is signal- or background-like
- Each output represents the probability that an event belongs to the corresponding class
- Is trained on data sample including events coming from different "domains", i.e. different signal models

Adversary

- Is trained only on signal events
- Tries to guess the physics model of signal events, regressing the domain from the second-to-last layer of C

Competitive learning

- The classifier is penalized if its output contains too much information on the domain of origin of signal events
- This training approach fosters the emergence of features among the classifier input variables that provide discriminating power for the main learning task (signal-to-background separation) while not relying on the domain shift
- If C manages to prevent A from identifying the signal model, then the classification is independent of the domains of origin of the events

Two-step training procedure

1. $\mathcal{L}oss = \mathcal{L}oss(C) - \alpha \cdot \mathcal{L}oss(A)$
2. $\mathcal{L}oss(A)$

Competitive learning

Compute first the gradient of $\mathcal{L}$ with respect to the C weights. A weights frozen in this step.

The parameter α regulates the interplay between A and C

Two-step training procedure

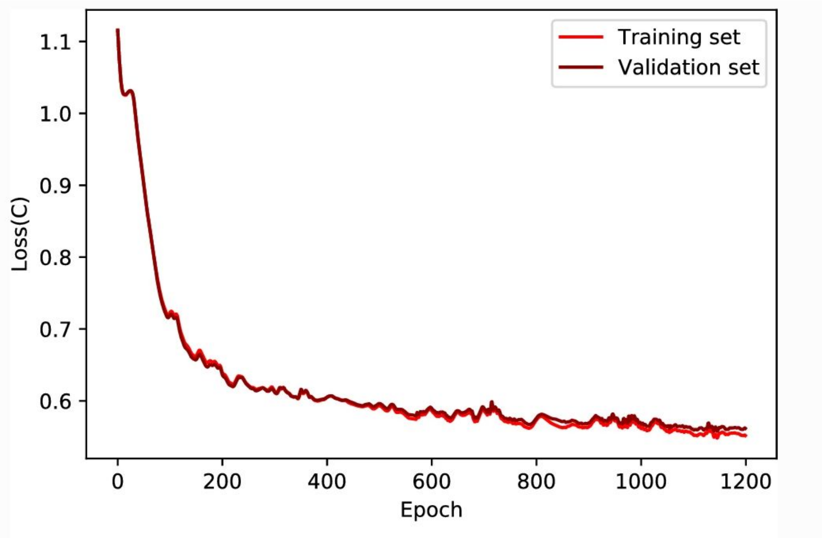
1. $\mathcal{L}oss = \mathcal{L}oss(C) - \alpha \cdot \mathcal{L}oss(A)$

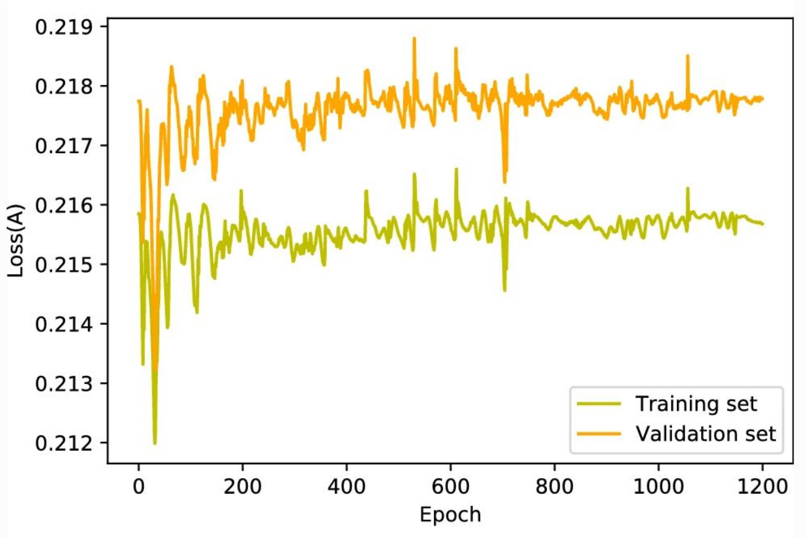
2. $\mathcal{L}oss(A)$

Compute the gradient of $\mathcal{L}(A)$ with respect to the A weights.

Behavior of the loss functions

- Example of the typical behavior of the A and C loss functions (categorical cross entropy in this case) when the two terms are balanced:
  - L(C) decreases as usual.
  - L(A) saturates to a constant value, meaning that the performance of A is equivalent to random guessing.





Summary

Summary

- Domain adaptation is a very active field of Machine Learning, especially in particular areas (such as computer vision).
- It is an emerging approach in the High Energy Physics field!
- We have seen just a specific application, but DA is also promising for a vast range of other applications currently under study.
- Enjoy (or have a look at) the dedicated hackathon exercise for a deeper immersion into DA!

Thanks for your attention!