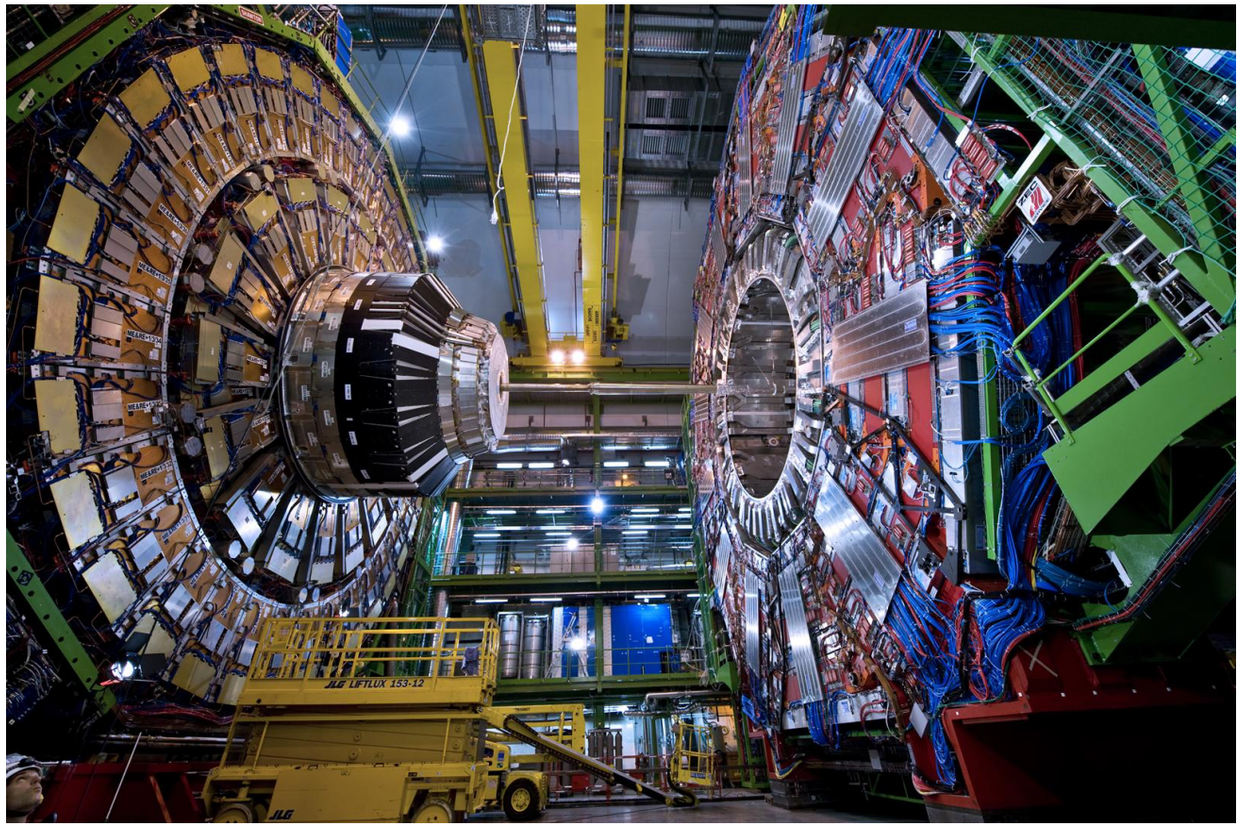
Machine Learning in CMS at DESY

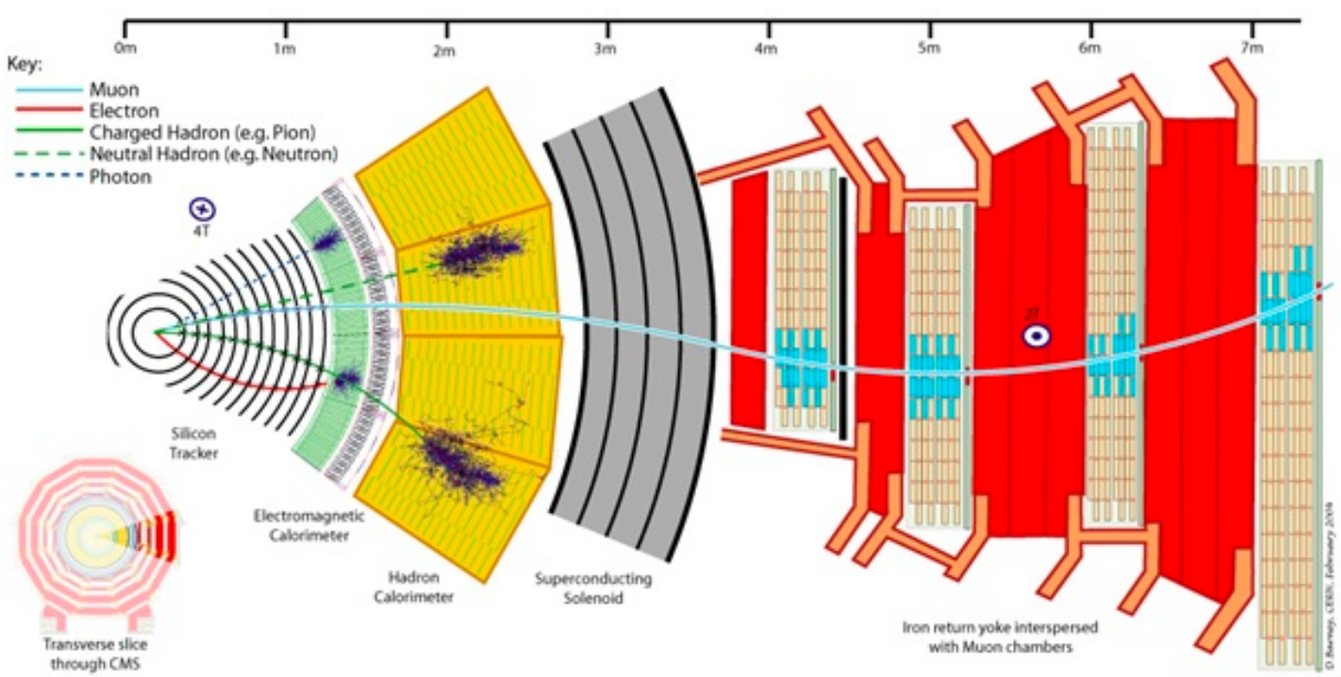
Adam Elwood

GPU computing & Machine Learning @ DESY
Nov 2018

adam.elwood@desy.de

The CMS experiment at the LHC





- Along with ATLAS, one of the 'Energy frontier' flagship experiments
- ~3000 scientists from 35 countries
- Extremely high data output rate
  - The majority of 'uninteresting' events thrown away by triggers
  - Very large datasets O(10PB) left for physics analysis
- Good simulation available for supervised training

Application of ML techniques in CMS

- Object reconstruction

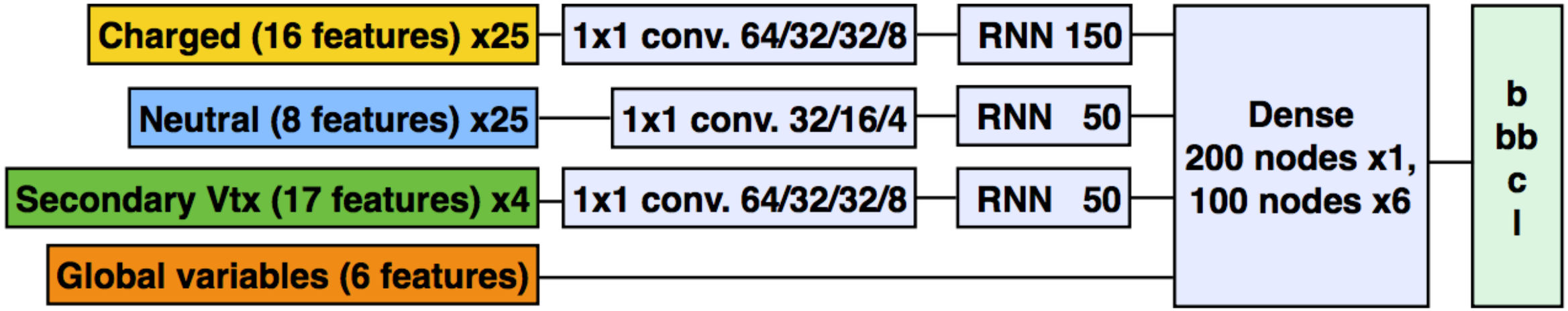
Current

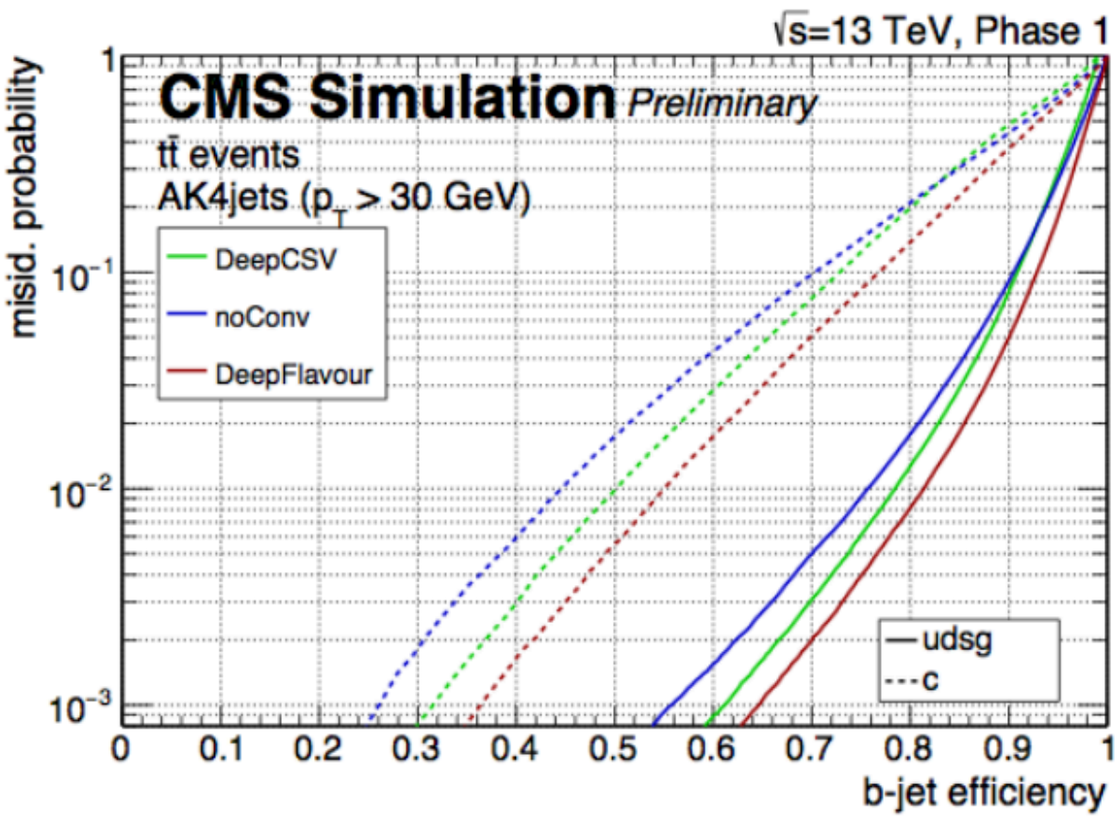
  - ML systems with input features taken from detector output used to reconstruct fundamental particles produced in detector
  - Significant success in using DNNs for jet tagging in CMS (DeepJet)
- Physics analysis
  - In searches for rare physics processes use for Signal and Background classification
  - Lots of success with BDTs but now considering DNNs as they are well suited to non-binary classification and accounting for complex correlations
- Generating simulation quickly

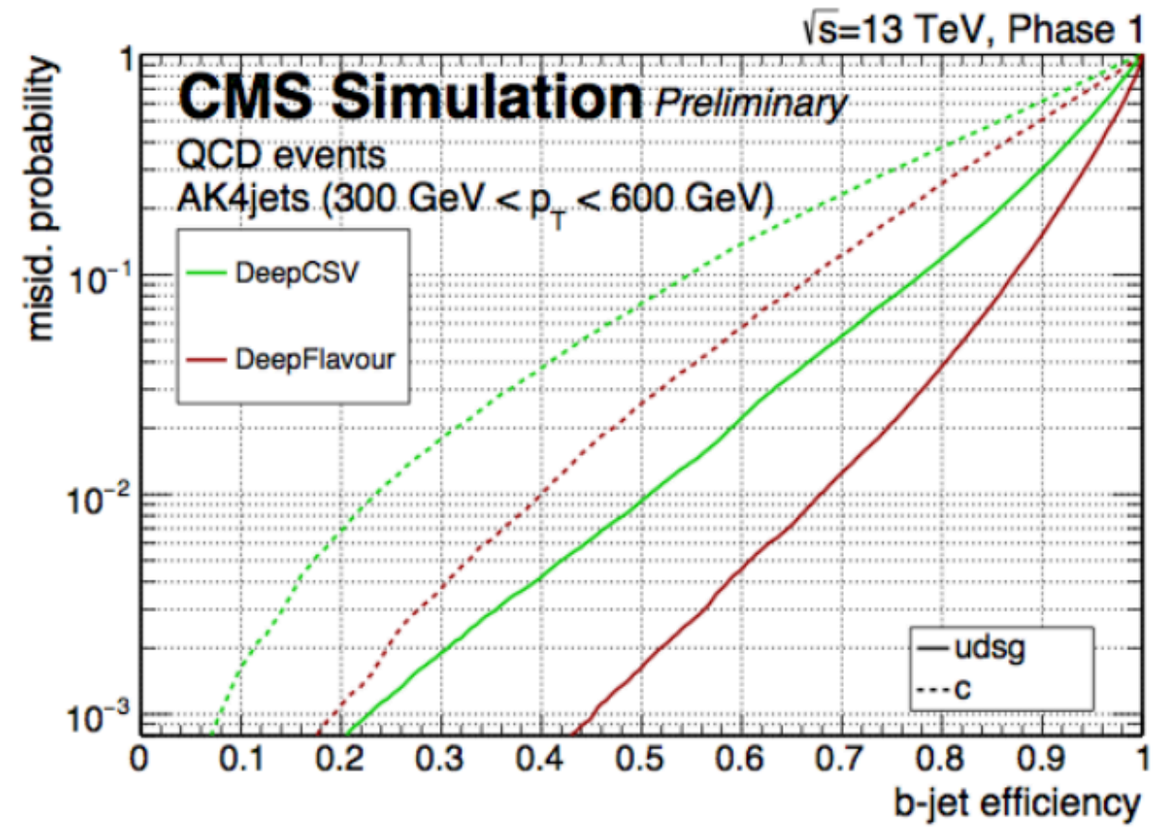
Future

  - Can minimise full detector simulation and significantly speed up generation
- More intelligent triggering
  - Making intelligent decisions without time for an involved reconstruction

DeepJet







CMS-DP-2017-013

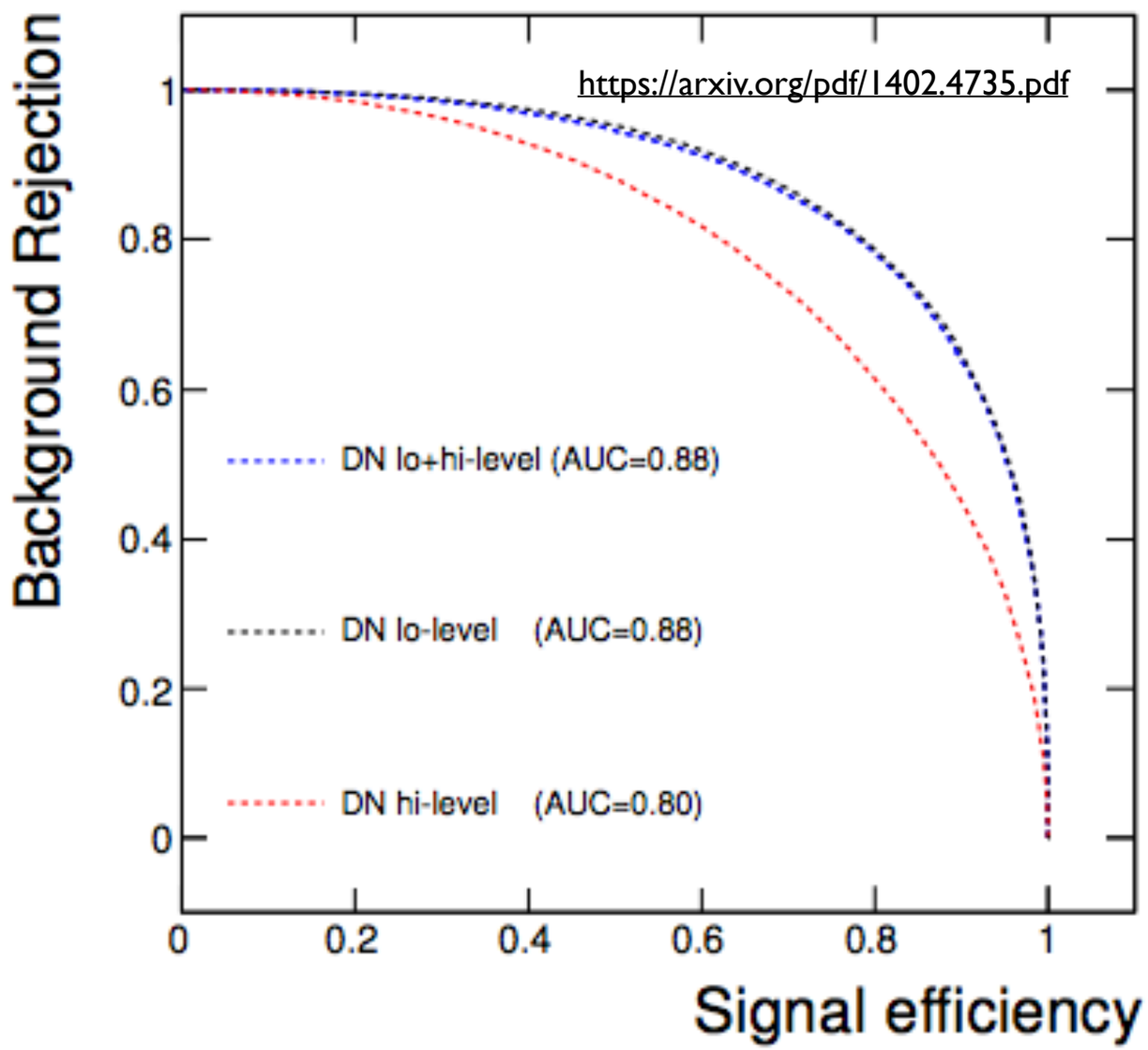
10

M. Verzetti (CERN and FWO)

https://indico.cern.ch/event/668017/contributions/2947023/attachments/1628927/2595479/IML_2018.pdf

Use in searches

- Networks can learn the sophisticated physics knowledge that goes into the design of high-level variables for HEP analyses



- CMS ttH result has made good use of DNNs (Phys. Rev. Lett. 120, 231801)

Current CMS work in DESY

Learning for Discovery

Cross-entropy and Asimov significance

- When optimizing a classifier a typical approach is to optimize the accuracy: $\mathrm{Acc}=strue+btrue/all$
- For neural networks standard approach for training a binary classifier is the **cross-entropy**
- Accuracy maximizing is equivalent to minimizing the cross-entropy

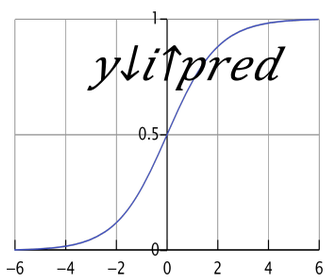
- **Can we optimize directly for the Asimov significance, i.e. can we use it as a <u>loss function</u> ?**
- **Caveat**: To define the number of signal and background events we need to cut on the discriminator output
  - Makes it non-differentiable ??
  - Differentiability is needed for gradient descent learning

$$Z_A = \left[2\left((s+b)\ln\left[\frac{(s+b)(b+\sigma_b^2)}{b^2+(s+b)\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2 s}{b(b+\sigma_b^2)}\right]\right)\right]^{1/2}$$

Learning for Discovery - Main Idea

Asimov significance as loss function

- A single sigmoid output neuron
- Replace the discrete number of signal and background events by a smooth function of the predicted label $y_i^{pred} \in [0,1]$



$$s = W_s \sum_i^{N_{batch}} y_i^{pred} \times y_i^{true}$$

$$b = W_b \sum_i^{N_{batch}} y_i^{pred} \times (1 - y_i^{true})$$

$1/Z_A(s,b)$ becomes a smooth function of y_i^{pred}

- Network and **Asimov loss** function defined in Keras with Tensorflow
- Dense Networks
- Here results for 1 hidden layer (21 nodes, ReLU)
- 1 Mill events each for stop and ttbar
- 1/3 for testing
- Less then 200 epochs (early stopping)

W_s, W_b some weights to get the physical number of events depending on luminosity and efficiency to get the

Learning for Discovery

Asimov Loss function and Classifier Output

arXiv:1806.00322 [hep-ex]

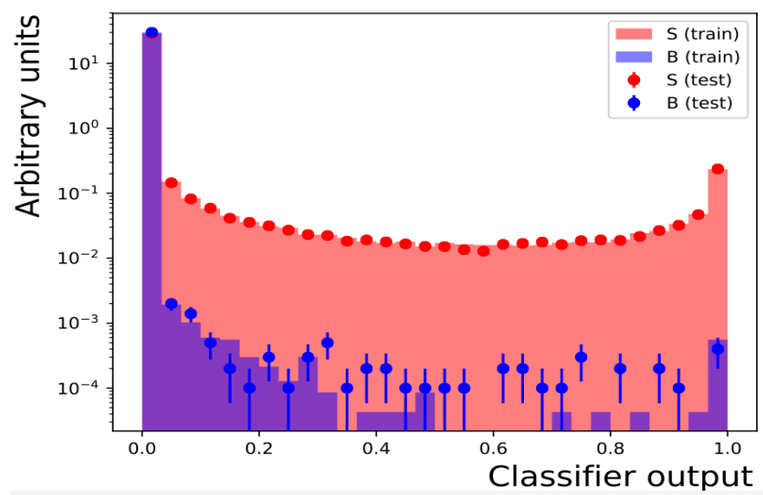
Asimov loss training

- best $\mathbf{Z_A = 6.2} \pm 0.6$
- Acc = 59%
- AUC=0.80
- Tries to find a background free region

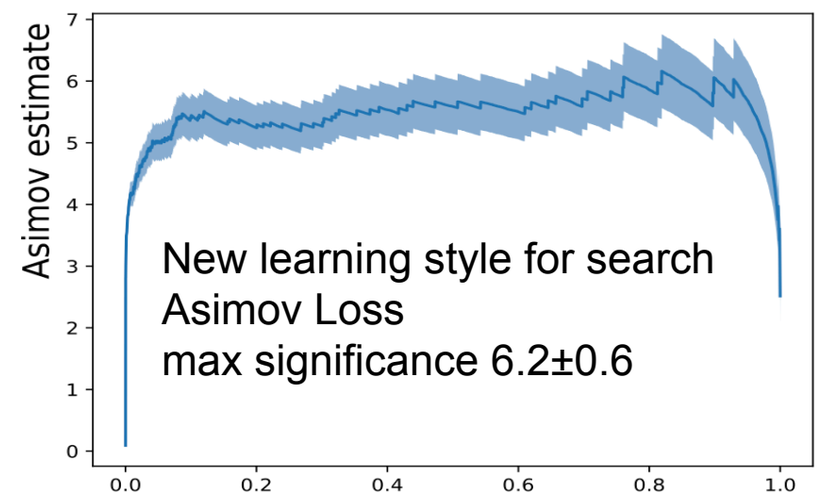
Cross-entropy training + purity cut

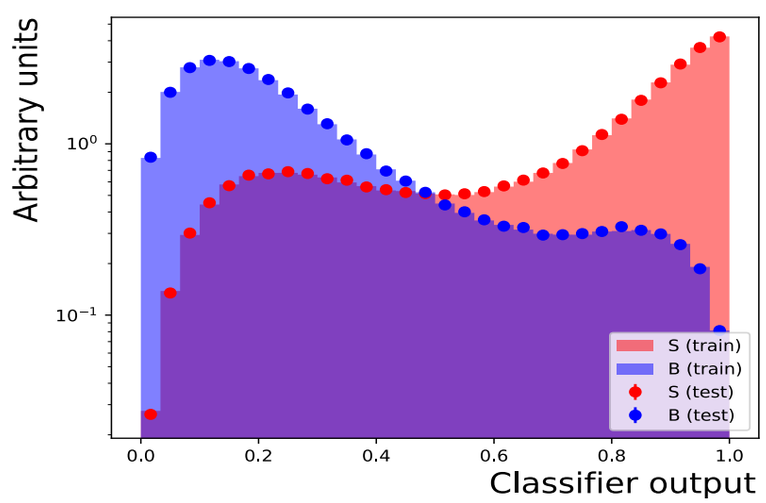
- best $\mathbf{Z_A = 4.8} \pm 0.3$
- Acc = 92%
- AUC=0.87

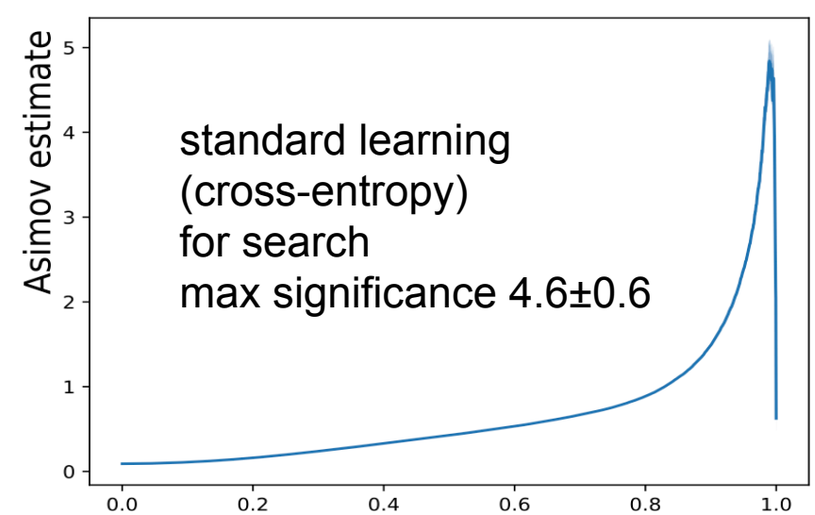
Systematic uncertainty 50%, Differences in Z_A shrink for small systematic uncert.



Asimov score vs. cut on classifier



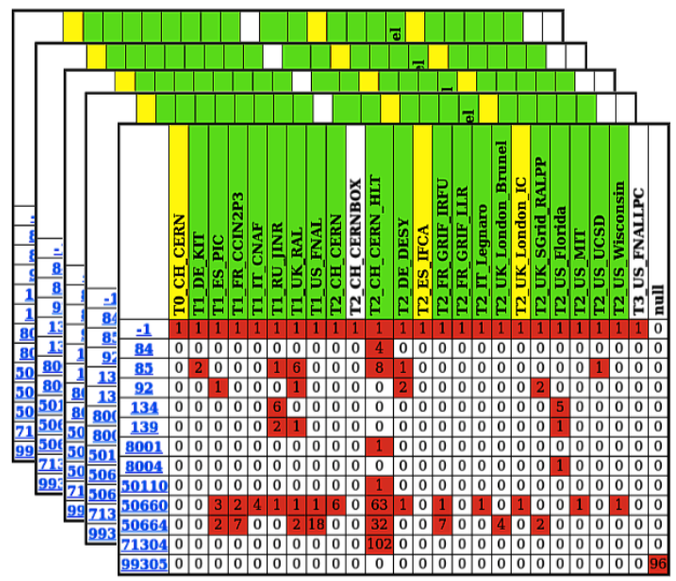


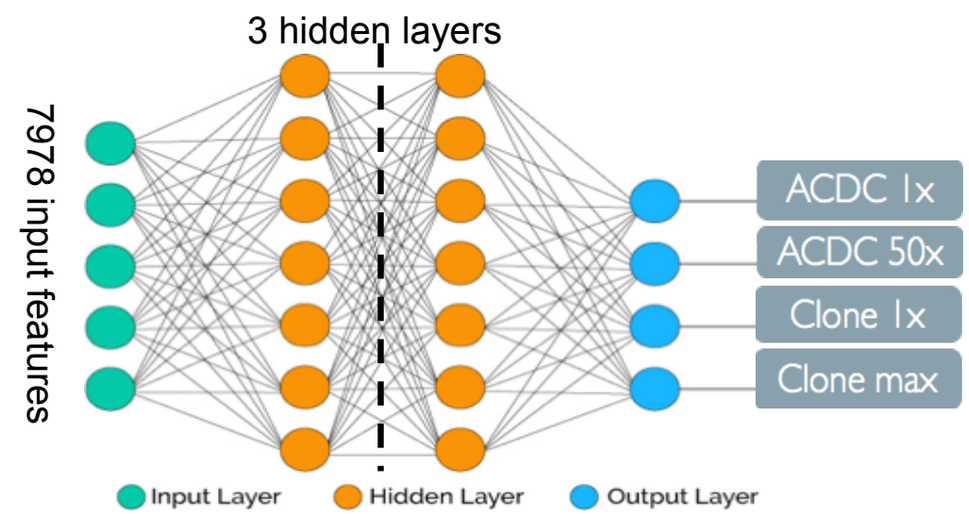


compressed SUSY model point

CMS Workflow Failures Recovery Panel Towards AI-assisted Operation

Christian Contreras et al. CHEP 2018





Motivation

- CMS central production sytem
  - ~200K grid cores, ~hundred sites
  - thousands of work flows with t housands of jobs
- Failures are unavoidable
  - Manual operator intervention
  - Can we have an automated system that give **advice how to recover a job**?
- Train a Deep Neural Network

Trainings data (15,000 workflow tasks)

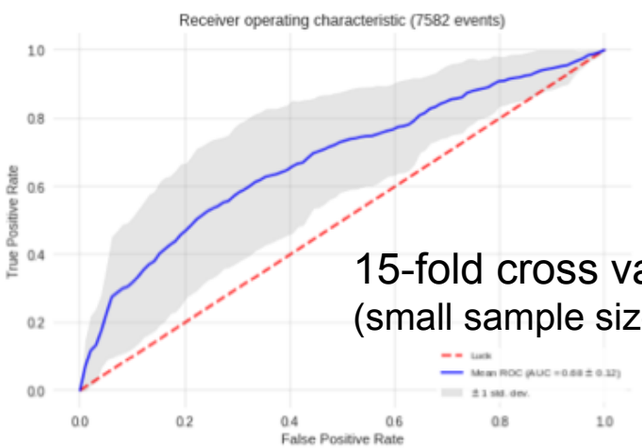
- Logs (json files) pulled from Workflow Team Web Tools which extracts information from Site-Readiness report
  - **Job failure codes and site information**
  - **Operator actions**

Multiclass Training
2 stage classifier
Binary + multi-class

- ACDC (rerun failed)
  - splitted
- Clone (total retry)
  - splitted

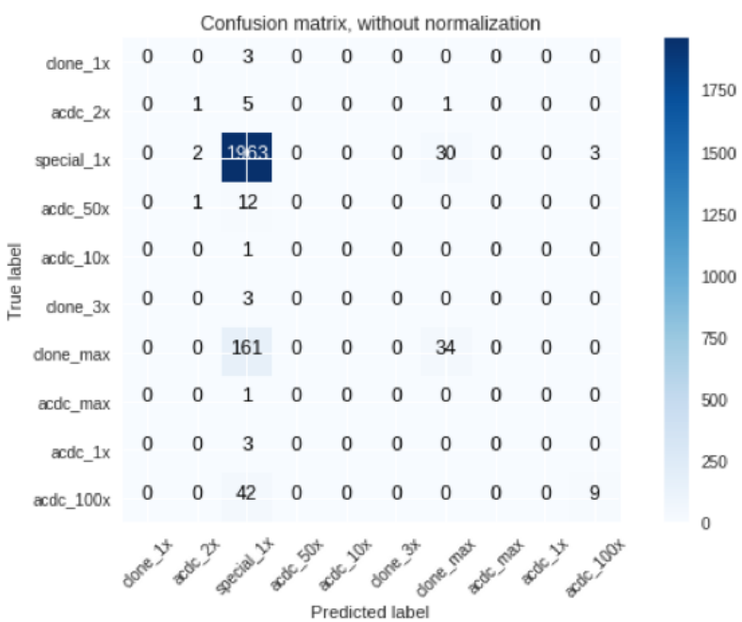
CMS Workflow Failures Recovery Panel Towards AI-assisted Operation

Christian Contreras et al. CHEP 2018



binary stage

15-fold cross validation on ROC
(small sample size)



Confusion matrix
for the multi-class stage
70/30% split for train/test

- A first pass for the supervised learning in error handling prediction. The operator's procedure will be automatized further by applying the decisions that are predicted with acceptable confidence.
- **Improve current WTC web interface**
  - To start using Machine Learning Model
  - Include the prediction for recommended action Start recovery from trivial cases
  - Monitor performance for model re-training
- Add GUI display for diagnostic summary reports

Model type	accuracy	Recall	Precision
Binary	87%	87%	83%
Multi-class	87%	86%	83%

1st Terascale Machine Learning School

Last week we run the first school on ML and Deep Learning successfully

95 participants! Educating people that will need resources.

With lectures and tutorials on

- **Deep Learning**
  - **Basics and advanced**
- **PyTorch, Keras and Tensorflow**

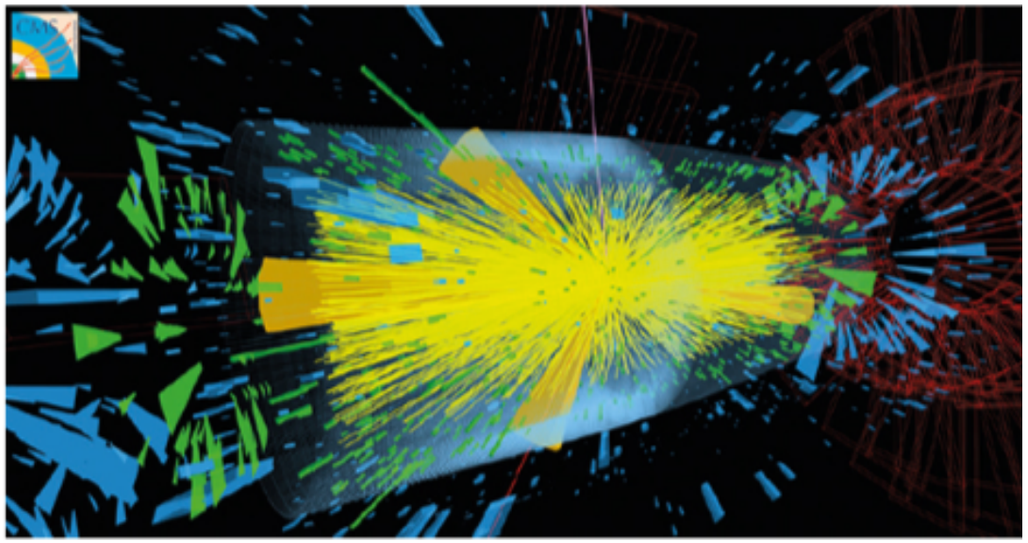
and many more subjects

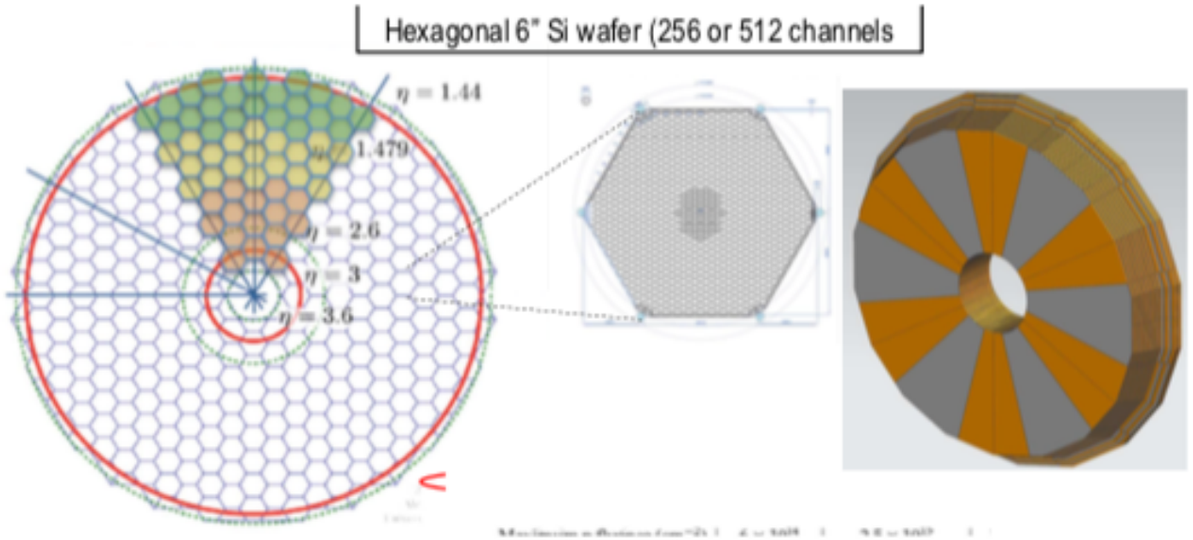
by Vincent Brenner, Christian Contreras Campana, Adam Elwood, Simone Frintrop, Gregor Kasieczka, Dirk Krücker, Gilles Louppe, Lorenzo Moneta, Benjamin Nachman, Fedor Ratnikov and Andrey Ustyuzhanin

Future CMS work in DESY

Extra motivation: the HL-LHC

- The HL-LHC is looming and presents a huge computing challenge
  - ~40 collisions/event to ~200 collisions/event
  - With current software ~10s/event to ~1 minute/event
  - On top of that… flat budget and more sophisticated detector as part of upgrade





- Current rule-based reconstruction algorithms and simulation techniques will not be sustainable
- **We have to leverage modern ML techniques to make up the gap**

M. Pierini https://indico.cern.ch/event/745580/contributions/3119763/

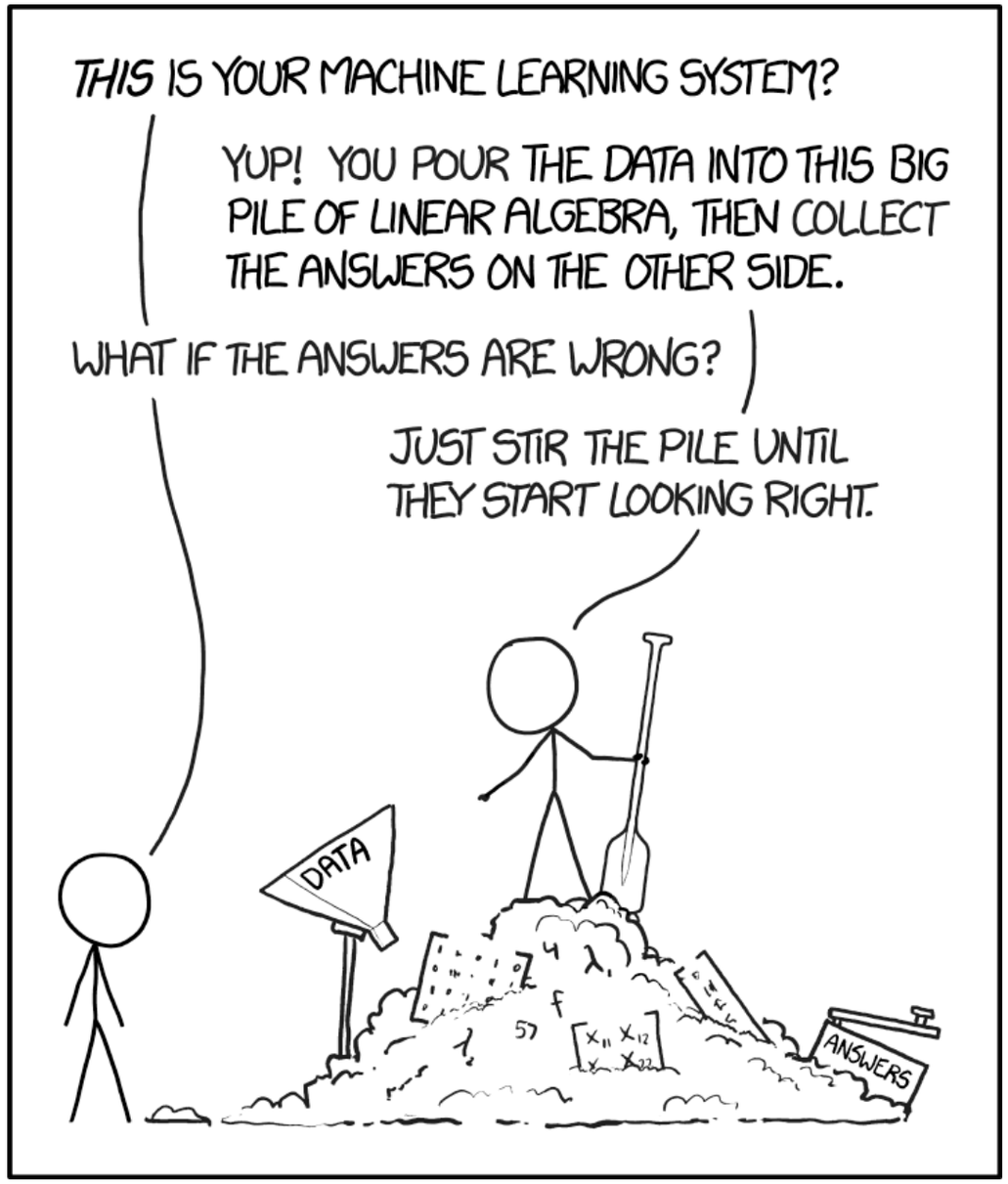
Future work at DESY

- Concrete projects:
  - Using more sophisticated architectures for physics analyses
    - E.g. adversarial networks for discrimination and data driven background estimation of a SUSY search
- Areas of interest to start to pursue:
  - Investigating the feasibility of GANs and VAEs to speed up all steps of MC simulation
  - Using event-level Autoencoders to see if the encoded latent space can help us to learn novel things about physics
  - Putting neural networks into FPGA to make intelligent trigger decisions with a microsecond latency

Conclusions

- We are gaining experience with modern ML techniques
  - Currently making use of 4 DESY CMS specific P100 Nvidia GPUs in Maxwell, shared in common partition
  - A couple of successful concrete projects already finished
  - Lots of new ideas to pursue
- Future developments will be critical for the success of HEP as we move into the HL-LHC era
- **Expect a steady increase in resources used**
  - Currently not limited by resources but have started with low intensity applications
  - Lots of the future ideas would require more intense resource usage

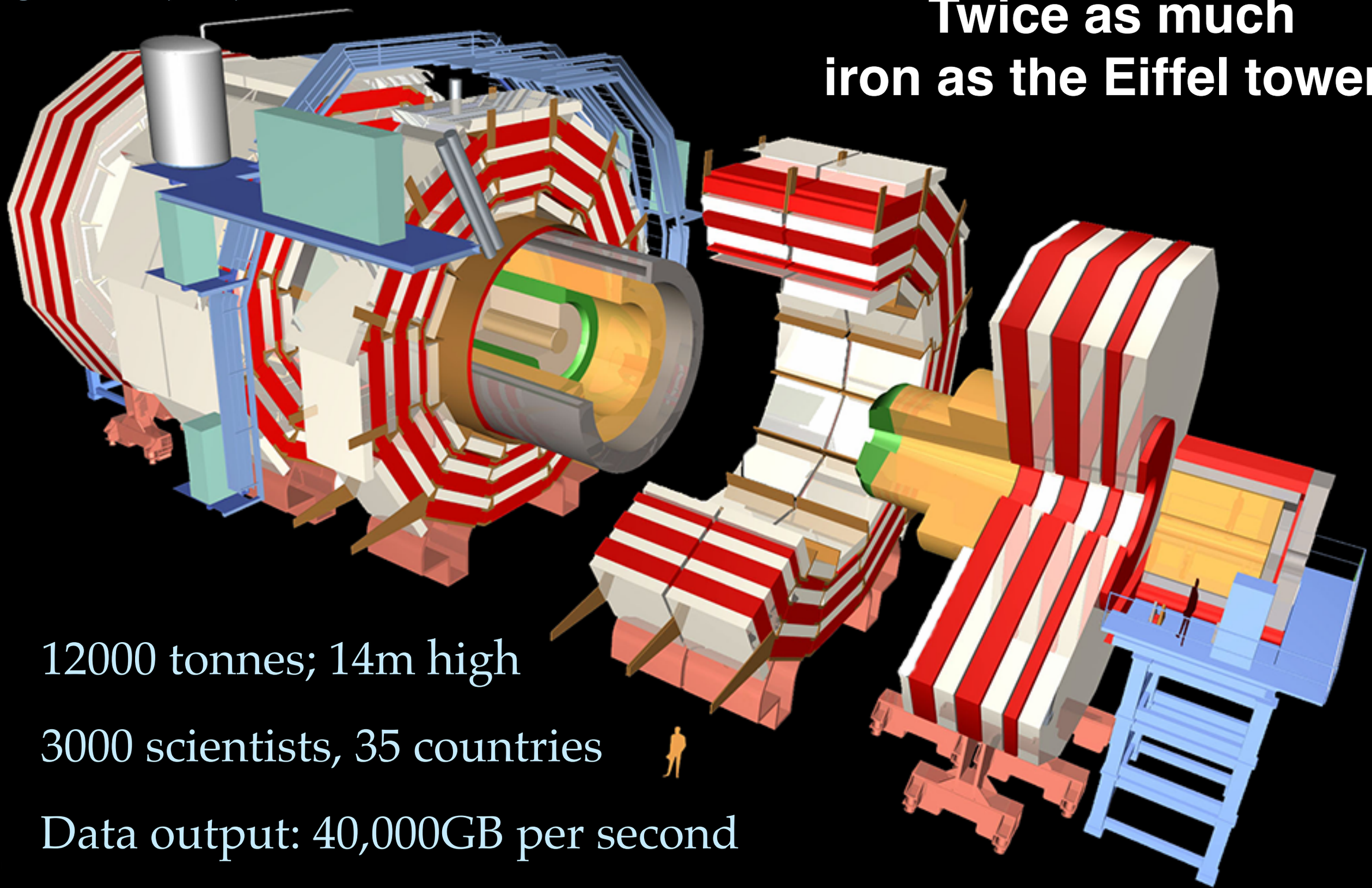
Backup


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS

At one of the collision points lies **CMS**

A general-purpose detector

Twice as much iron as the Eiffel tower!



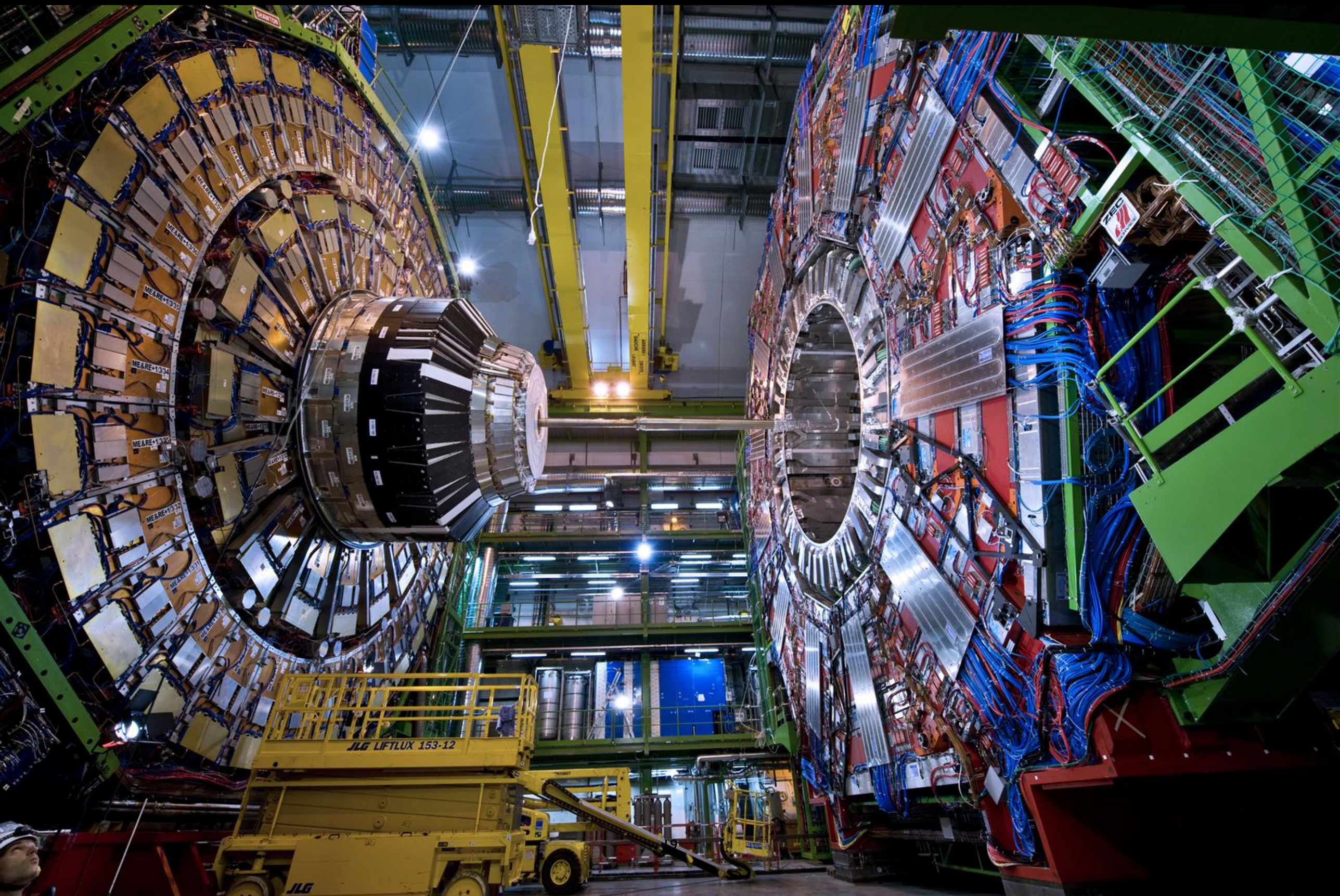
12000 tonnes; 14m high

3000 scientists, 35 countries

Data output: 40,000GB per second

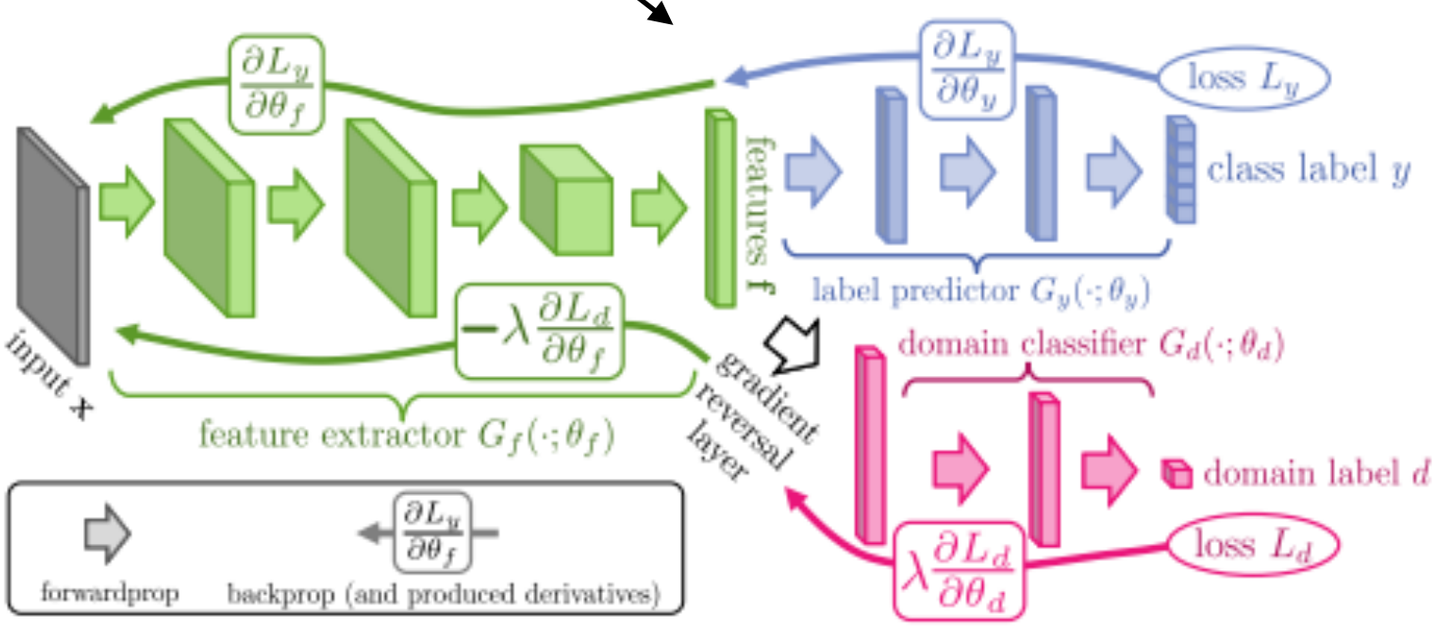
M.

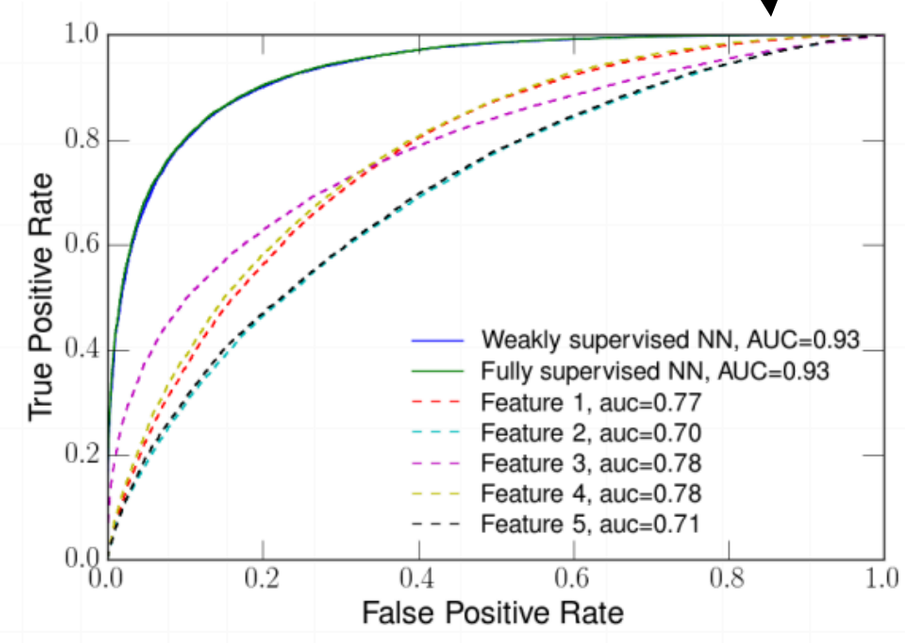
At one of the collision points lies **CMS**



Too much reliance on simulation?

- Ideas to remove systematics introduced by training on simulation:
  - Only use input variables with good MC/Data agreement * (but take care of correlations)?
  - Train purely on data, one sample signal enriched and one background enriched? **
  - Train an adversarial network that penalises differentiation between data and MC? ***





* https://indico.cern.ch/event/646801/contributions/2651014/attachments/1487701/2311252/6Jul2017_CMSML_systematics.pdf
** http://arxiv.org/abs/1702.00414 *** http://jmlr.org/papers/volume17/15-239/15-239.pdf , https://arxiv.org/abs/1702.05464