Welcome to the training session on data preparation for ML

David Brunner, Isabell Melzer-Pellmann, Mykyta Shchedrolosiev

Download of the large datasets

Prelude

Before starting with the introduction: did you all download the large dataset?

If not, do the following:

- If you don't have a google account, get one.
- Get yourself a copy of the following notebook: https://colab.research.google.com/drive/1sCcDXuuffmuVO_rzApBwB_2uRyX3a31M#scrollTo=8k20EIosJpF0
- File → Save a copy in Drive
- Go through this notebook, make sure to follow the instructions when creating the directory, you have to go to the website that is given there, copy the code that will be displayed and paste it before the download can begin

Introduction

- The amounts of data taken and to be analyzed has grown significantly in the past, in science like high energy physics as well as in many other fields
- Data analysis makes more and more use of machine learning techniques
- High energy physics analyses were using machine learning since many (>20) years, but the real breakthrough of the technology came with the power of computers to allow for real deep networks
- At that point industry took over the lead
- Also the design and usage of complex models became easier in the past few years, driven by global players like google who invested much more manpower and hardware than science can dream of
- We can take advantage of the easy use, but before using machine learning, we have to prepare our data such that it can be used
- This is what this tutorial is about

Issues with big data

Typical pitfalls:

- When training your data, you have to use different datasets (e.g. signal and background) and different variables to isolate your signal
- Your background file might be much bigger than your signal file – how to treat this? We want to use as much information as possible, so do not simply reduce the number training events, but scale them. How is this done in practice?
- Your variables might have totally different scales – this might lead to a bad training result, since the gradient decent might be affected. Again rescaling does the trick, but how to do it in practice?

At the end of this tutorial, you will hopefully have working examples that you can keep as examples for your own work in the future.

Outline of the training session

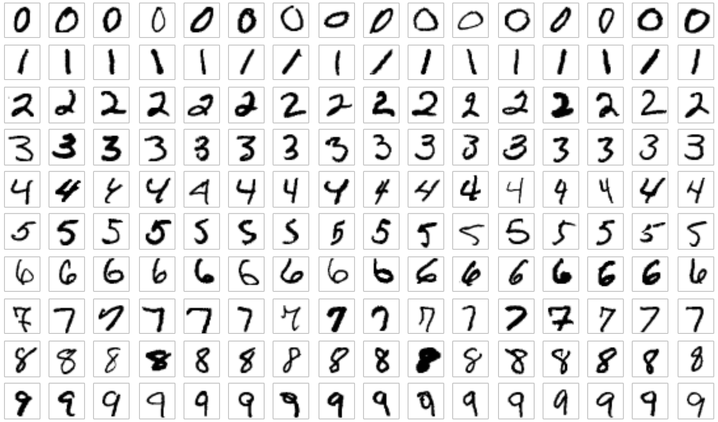
The training consists of two parts:

- Introduction to the techniques and example code using the MNIST dataset
- Exercises using Monte Carlo simulated events from the CMS experiment (real big data)

MNIST dataset

MNIST database: Modified national Institute of Standards and Technology database

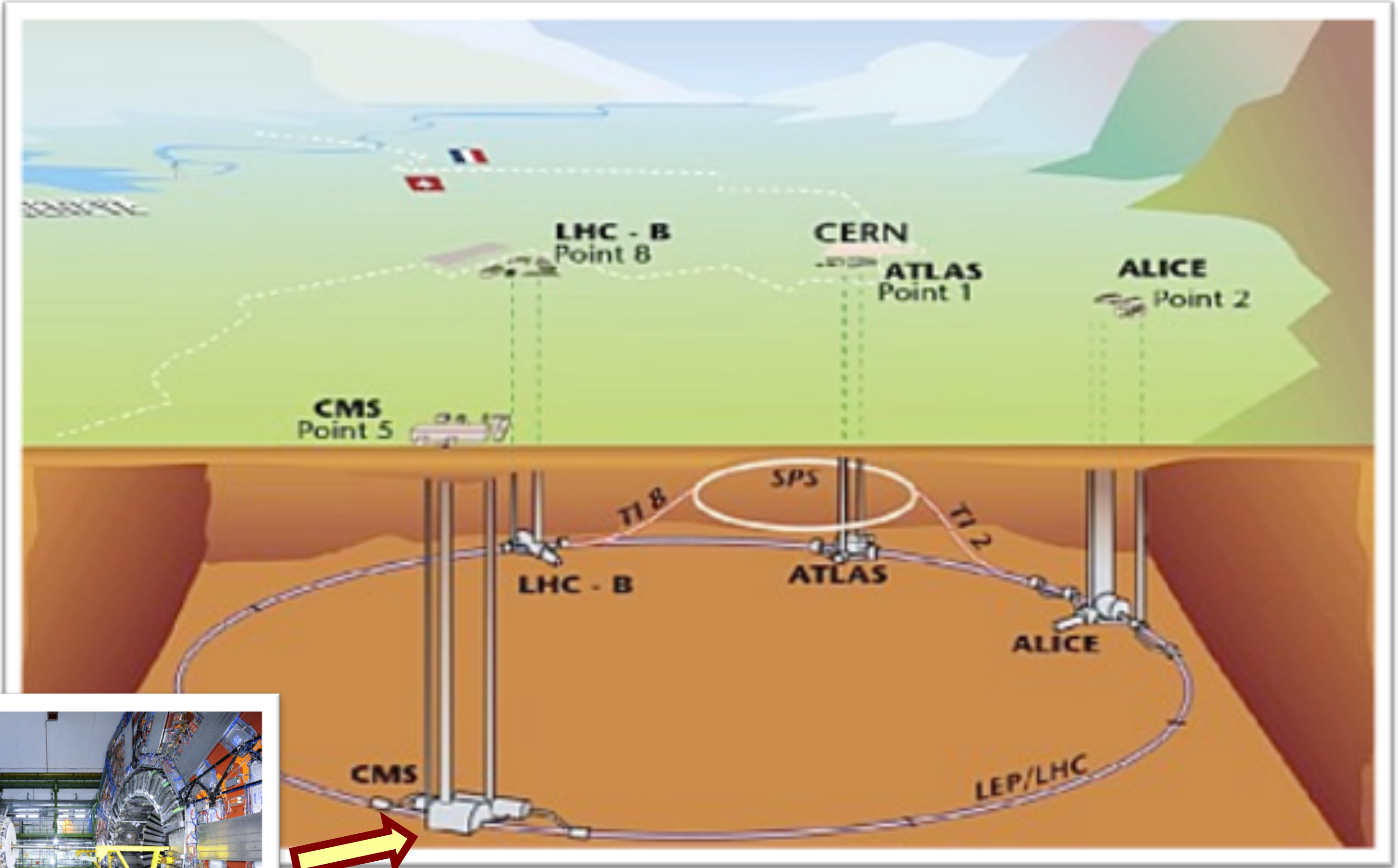
- Large database of handwritten digits that is commonly used for training various image processing systems
- This database is also widely used for training and testing in the field of machine playing
- Created by "re-mixing" the samples from NIST's original datasets, because the original NIST training dataset was taken from American Census Bureau employees, while the testing dataset was taken from American high school students → not well-suited for machine learning
- The black and white images from NIST were normalized to fit into a 28x28 pixel bounding box and anti-aliased, which introduced grayscale levels
- 60,000 training images and 10,000 testing images

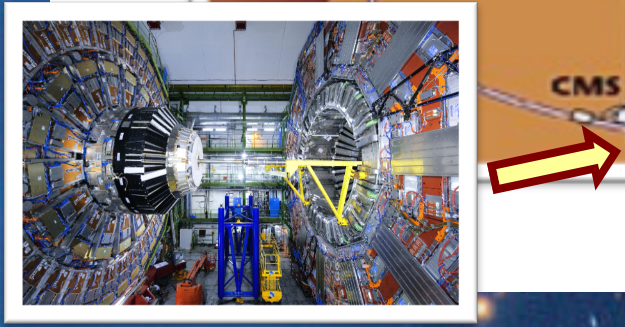


From Wikipedia; read more here:
https://en.wikipedia.org/wiki/MNIST_database

Introduction to CMS

CMS: Compact Muon Solenoid – experiment at the **Large Hadron Collider**

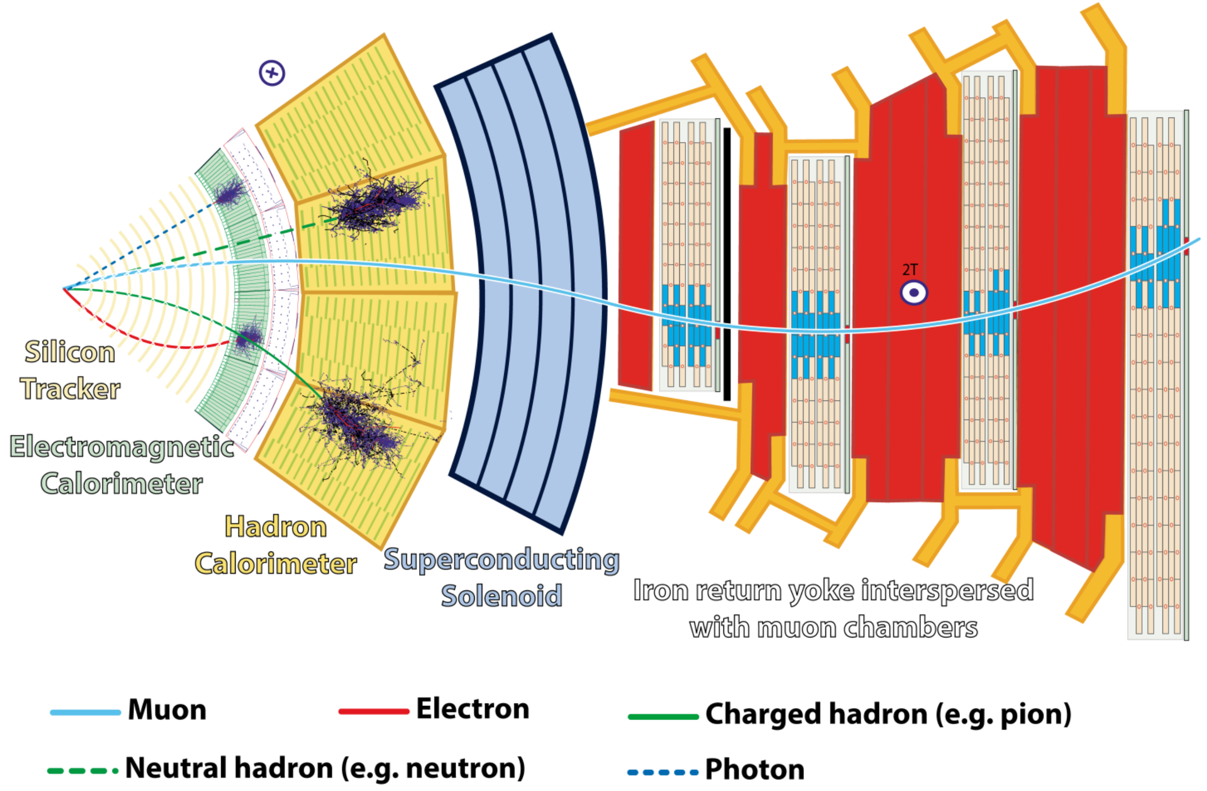




The CMS Experiment

Multi-purpose detector:

- 15m high, 21m long
- Weighs 14 000 tons
- Biggest solenoid magnet with 6m diameter and a magnetic field of 4T
- Different particles leave different traces in the detector

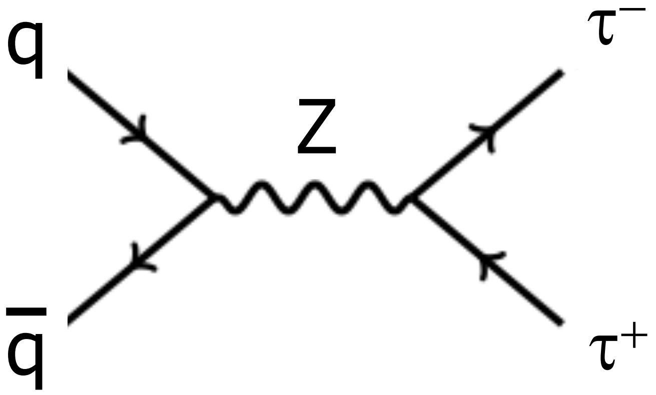


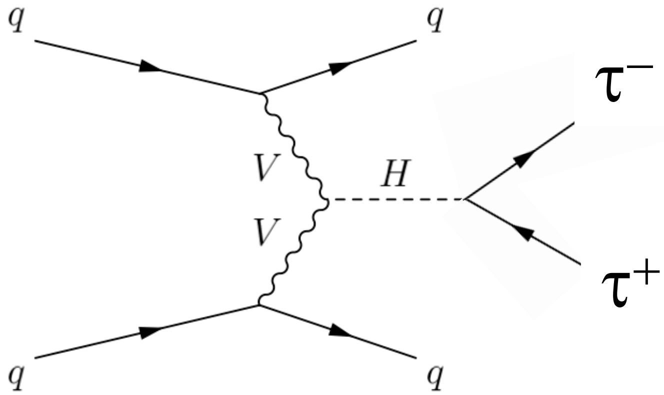
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with muon chambers
2T
Muon
Electron
Charged hadron (e.g. pion)
Neutral hadron (e.g. neutron)
Photon

Particle production and decay

Your task today: train a network that can distinguish two different physics processes – find a Higgs boson in a background of Z boson events!

- Good to know: Z boson production is several orders of magnitude more likely than Higgs boson production
- The CMS experiment made part of their data and simulated events publicly available – we will use these data





- Quarks (q) hadronize to jets of particles
- Tau lepton: heavier sibling of the electron (and muon): decays, and visible decay product is either a jet (slightly different from quark jet) or a muon or electron (here: use jets)

CERN open data

The full list of variables that is available for each event in this dataset can e.g. be found here:

http://opendata.cern.ch/record/12353

Not all variables make sense for the training, let's focus on a few:

- Variables connected with the tau leptons:
  - Transverse momentum (p_T)
  - Pseudorapidity (eta)
- Variables connected to jets:
  - Transverse momentum (p_T)
  - Pseudorapidity (eta)
  - Phi
- Missing transverse momentum (since in tau decays appear neutrinos which leave the detector without a trace)
  - Transverse momentum (p_T)
  - Phi

Simple selection: require two tau leptons and two jets

Tutorial and exercises

All exercises will be done in colab:

- Go to the following address:
  - https://colab.research.google.com/drive/1iCBmGDpuI6TPrvt6nNsFEN_leNxXN1nG#scrollTo=ZcfvZHSMGMed
  - File → Save a copy in Drive

Let’s get started!

Tutorial and exercises

Backup if you failed so far to download the large dataset:

- Do you have a google account? – If not, get one now
- Try to use the files in David’s googledrive following these instructions before running the CoLab notebook:
  - Go to this link: https://drive.google.com/drive/folders/1Iw4Bh3g8cUij20VT8uR_7QAWNcLRqTSx?usp=sharing
  - On the top of the page you see:
    Shared with me > LiverpoolSchool
  - Click on the little arrow to the right of “LiverpoolSchool” and select “Add a shortcut to drive”
  - Select “My Drive” and click on “ADD SHORTCUT”
- **Caveat:** We encountered problems with this, when more than 2 people access David’s file (seems to be no bug, but an unwanted feature)