bTagging @ CMS (HEP)

Aim of the hackathon

- **Use realistic CMS@LHC@CERN data samples (from Monte Carlo simulation) to prepare a DL system able to “categorize” the flavour of the quark generating a jet of particles**

But first, a few wider range explanations

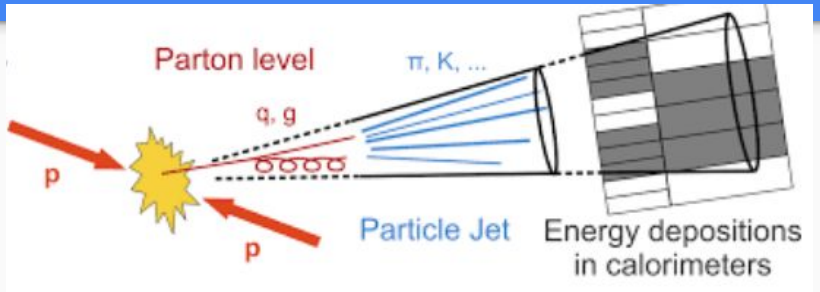
Largely based on the work by Leonardo Giannini (formerly INFN, now UCSD)

First seen here

CMS- Compact Muon Solenoid

- One of the 4 major LHC(CERN) experiments
- Collects data from proton proton collisions @ 7-13 TeV
- In particular, when you collide 2 bunch of protons, **quarks tend to be expelled**; due to QCD confinement, they cannot stay “naked”, and vest themselves with additional quarks and particles.
- Some of the quarks are unstable (t) or generate unstable particles (b, c, s and sometimes u and d), which generate even more particles close to the direction of the initial quark

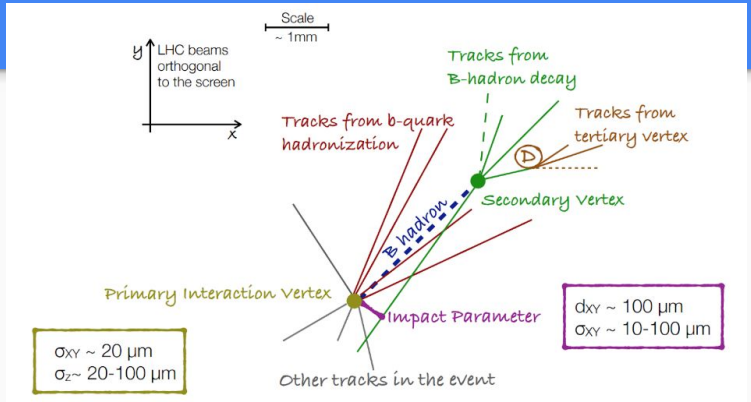
Here we try to recognize jets from b quarks from jets originating from other flavours



- Knowing (statistically) the favour if the initial high energy quark is very important in HEP
  - Many models of New Physics
  - Higgs physics (H decays preferentially to b quarks)
  - ...

Which are our handles?

- B hadrons have a “sizeable lifetime” (**ctau = 500 um**) so with the typical Lorentz boost @ LHC they **“decay”~ mm from the pp interaction point**
  - Light quarks (u, d) and gluons either much closer or much farther
- B hadrons have a **decay multiplicity** (number of daughter particles) ~5, higher than lighter quarks
- B hadrons have a higher fraction of decays containing **leptons**
- B hadron **mass is ~ 5 GeV**, higher than other quarks

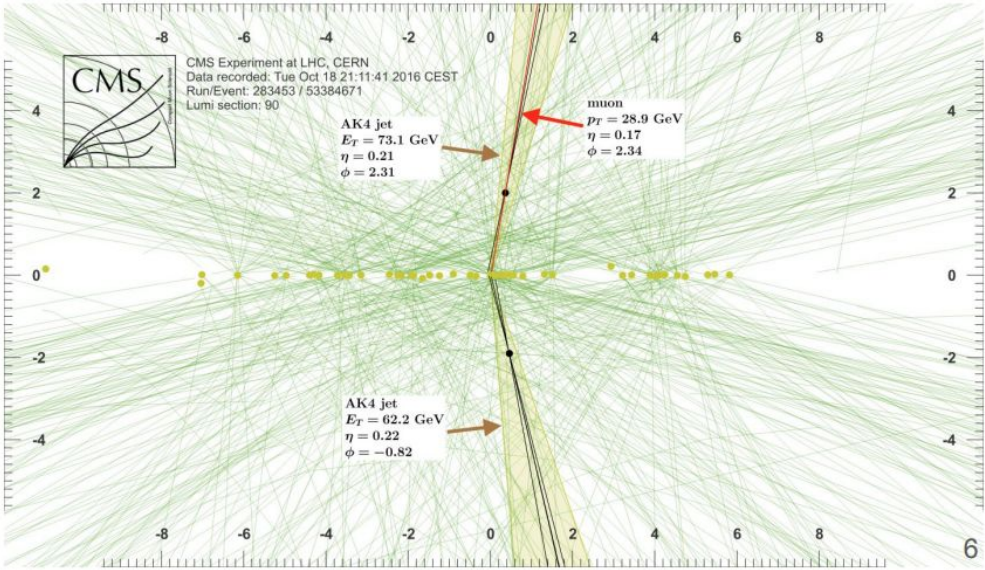


So we can search for

- Tracks **incompatible** from coming from the pp interaction point
- Presence of “**more tracks**”
- Presence of **leptons**
- Set of tracks with **large invariant masses**
- ...

Btagging @ CMS

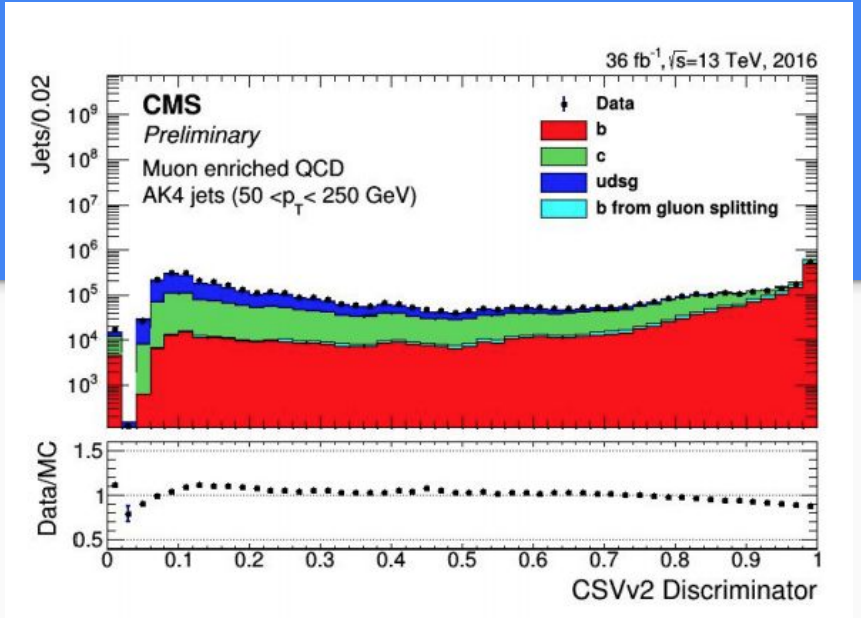
- The CMS detector has been designed with capabilities targeted (also) at optimally finding b quark decay jets:
  - Efficient particle tracking to reconstruct charged tracks
  - Capability to reconstruct decay vertices with < 100 um resolution
  - Electron and muon systems to spot leptons
- … even in the complex LHC scenario



Btagging algorithms

- Simple algorithms look for **leptons, tracks incompatible from coming from the pp vertex, secondary vertices**
  - They have a limited sensitivity and were used in the initial data taking periods
- Today's algorithm try to **combine all** the available info in order to have a more performant and stable result
  - Probabilistic, Likelihoods, MVA, BDT, … all used in the past. Now it is ML/DL turn!

In the CMS framework, the output of an algorithm is called "**discriminator**": it is just a variable which, when tested on Monte Carlo Simulated samples, tends to be different when the algorithm is applied to jets coming from b quarks … or not.

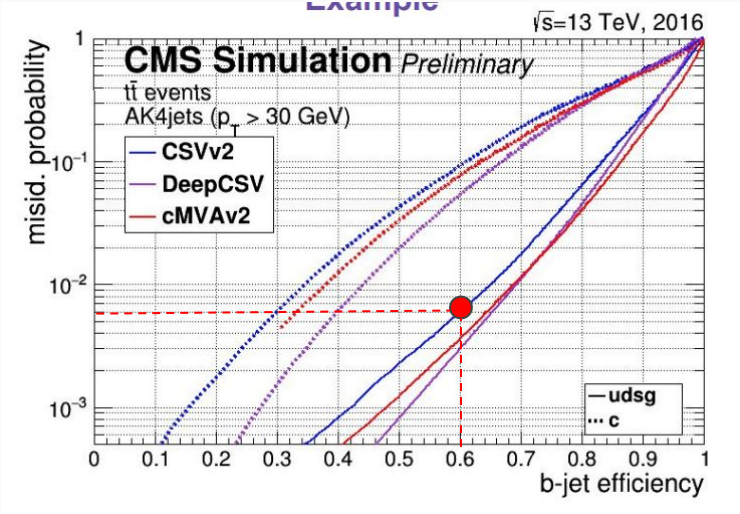


You can see here that most b quarks populate the "high value region" (it is on log scale!!!), while light quarks (uds) and gluons mostly the low value

- The higher the separation, the better you can be at selectin / categorizing / reducing the udsg background in the analysis

How to evaluate how good is an algorithm?

- You need to tune the selection on the “discriminator” in order to select more/less efficiently b quark Jets, with more/less selected events from lighter jets
- These can be inferred from the ROC curve:
  - You scan the discriminator cuts, and doing so you plot a line in a (X,Y) plot
  - Given a point, you can get the efficiency on signal on the X axis, and on background on the Y axis



“This algorithm, with a given discriminator cut, is able to select 60% of the jets from b which selecting just 0.6% of jets from lighter quarks”

The exercise

- **3 parts, 2nd and 3rd are there just if you are very fast**
- First part:
  - Consider some Jet variables and the output of some already existing complex features
  - Build a ROC curve using some of them
  - Assuming that “the more the info, the better the result”, try to prepare a DL system which has in input many of the info, to see if you can do better than standard algorithms
    - Better: plot a new ROC curve
- Second and third part (probably not today):
  - Instead of using only jet related quantities, try and add jet constituent quantities (like the tracks inside the jet), with 2 different DL technologies
  - Is that even better? → ROC curve

The notebook ..

- You will find in the collabshared area the notebook **btaggingTom_May26_nocelloutput_students.ipynb**
- It contains already a lot of code, especially those to load the data samples and do some lower level actions (plots, ROC curve, …)
- You will need to complete parts, mostly DL related
- You will have some questions, hints for improvements, ideas …

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 (make one copy per group!) for guidance

The tutors

Group 4: Silvio Donato
INFN Pisa researcher , CMS Trigger Coordinator

Group 5: Leonardo Giannini
PhD with SNS in CMS, now @ UCSD

Group 4: Valerio Bertacchi
INFN Pisa, PhD-epsilon with SNS