Data analysis and Statistics at the LHC

Dr. Nicholas Wardle

VSOP *Quy Nhon, Vietnam*

15-26 July 2024

Overview (Exp)

	Morning	Afternoon
Tuesday 16th	**Lecture 1** **LHC Data Analysis**	**Exercise 1**
Wednesday 17th	**Lecture 2** **LHC Statistics**	**Exercise 2**
Thursday 18th		**Lecture 3** **LHC Statistics**
Friday 19th		**Exercise 3 (&4)**

Slides for today inspired by
Sourabh Dube (VSOP-28 2022)

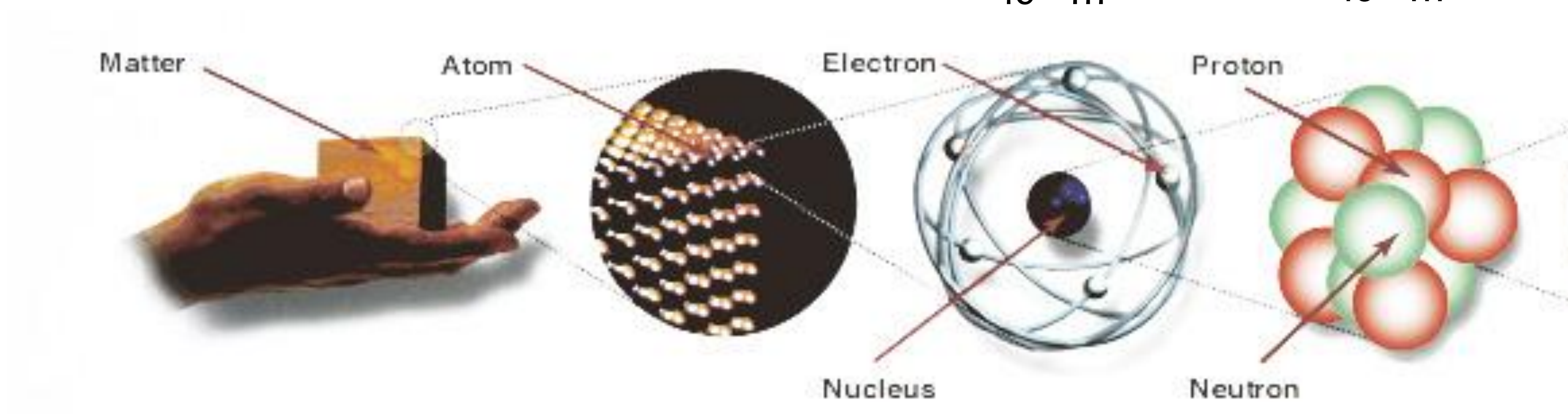
Our understanding of matter

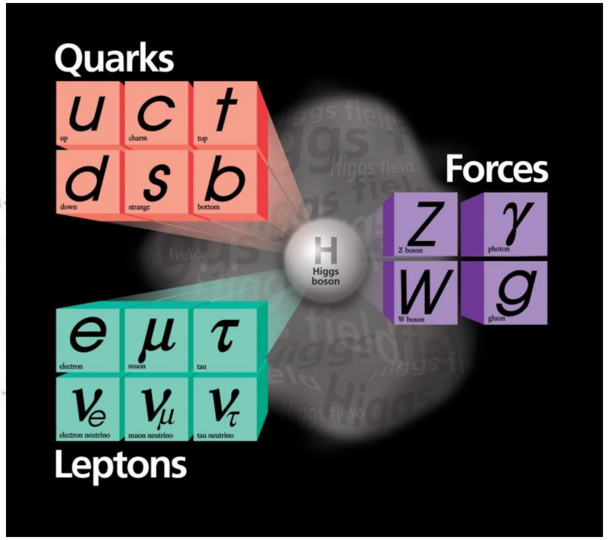
Physical Size

$< \sim 10^{-15}$m

$\sim 10^{-10}$m

$\sim 10^{-14}$m

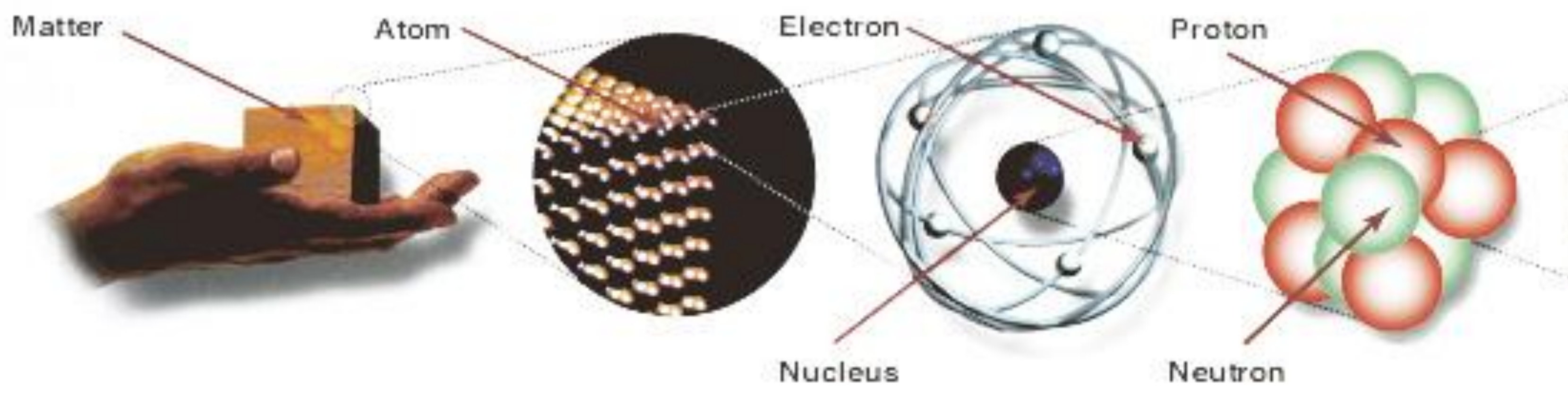




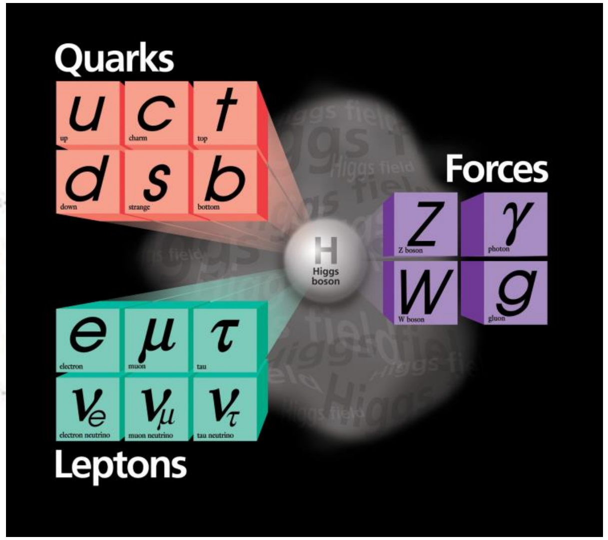
Our understanding of matter

Physical Size

$\sim 10^{-10}$m $\sim 10^{-14}$m



$< \sim 10^{-15}$m

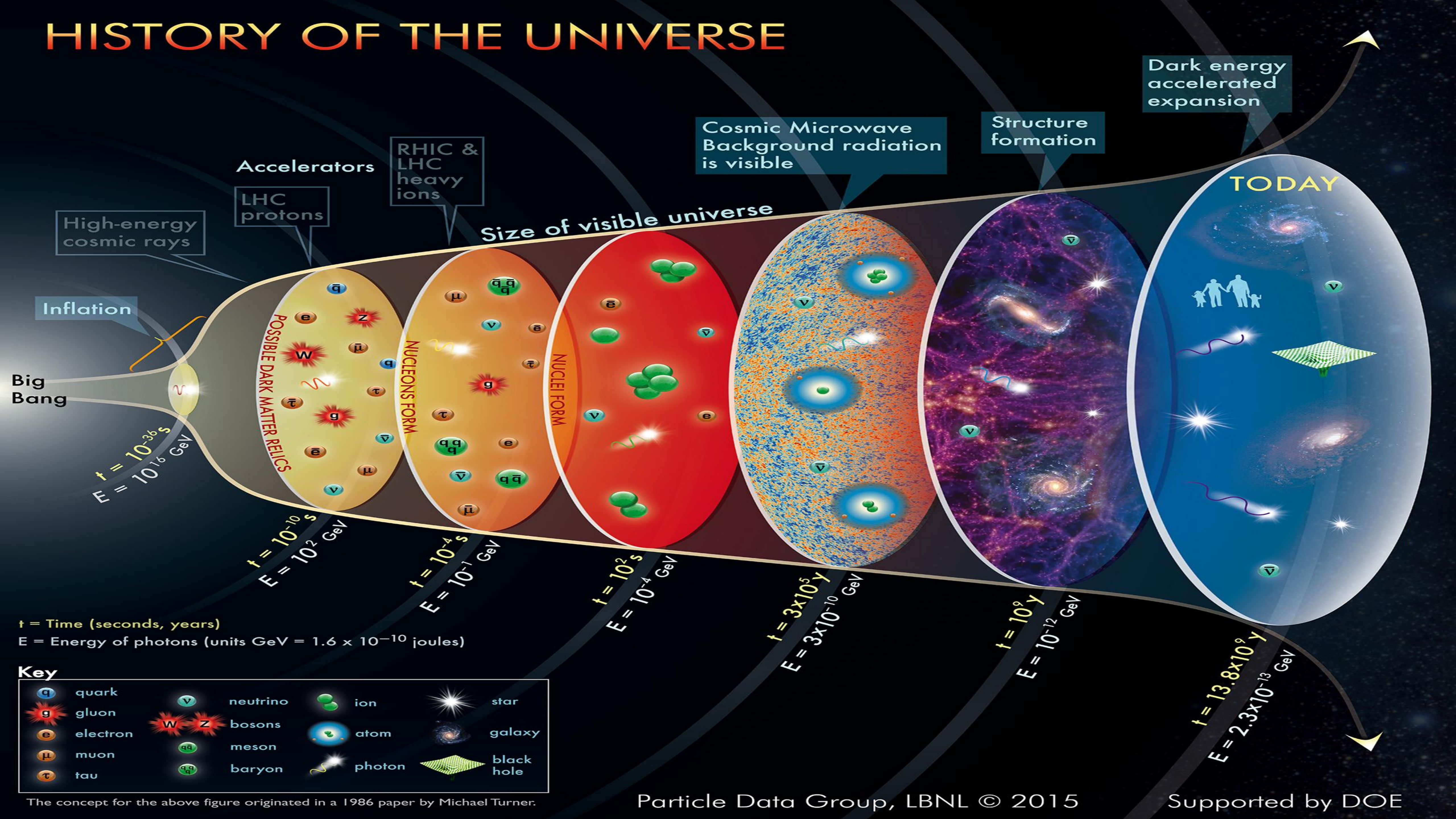


de Broglie relationship

$$\lambda = \frac{hc}{E}$$

Energy (density)

Probing finer structure requires higher energy densities → Particle Collisions probe fine structure of Nature


HISTORY OF THE UNIVERSE
Dark energy accelerated expansion
Cosmic Microwave Background radiation is visible
Structure formation
RHIC & LHC heavy ions
Accelerators
LHC protons
High-energy cosmic rays
Inflation
Size of visible universe
TODAY
Big Bang
POSSIBLE DARK MATTER RELICS
NUCLEONS FORM
NUCLEI FORM
$t = 10^{-36}$ s
$E = 10^{16}$ GeV
$t = 10^{-10}$ s
$E = 10^{2}$ GeV
$t = 10^{-4}$ s
$E = 10^{-1}$ GeV
$t = 10^{2}$ s
$E = 10^{-4}$ GeV
$t = 3\times10^{5}$ y
$E = 3\times10^{-10}$ GeV
$t = 10^{9}$ y
$E = 10^{-12}$ GeV
$t = 13.8\times10^{9}$ y
$E = 2.3\times10^{-13}$ GeV
t = Time (seconds, years)
E = Energy of photons (units GeV = 1.6×10^{-10} joules)
Key
quark
gluon
electron
muon
tau
neutrino
bosons
meson
baryon
ion
atom
photon
star
galaxy
black hole
The concept for the above figure originated in a 1986 paper by Michael Turner.
Particle Data Group, LBNL © 2015
Supported by DOE

The LHC

The Large Hadron Collider at CERN is a fundamental physics experiment!

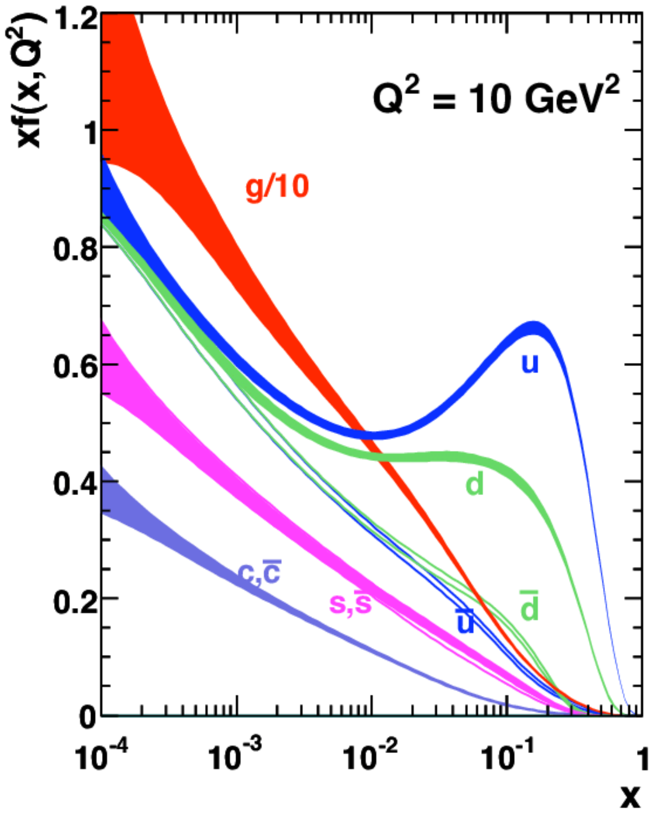
- 27 km in circumference
- 100m underground
- Accelerates protons to 99.9999991% x speed of light
- Proton circles 11,245 times per second!

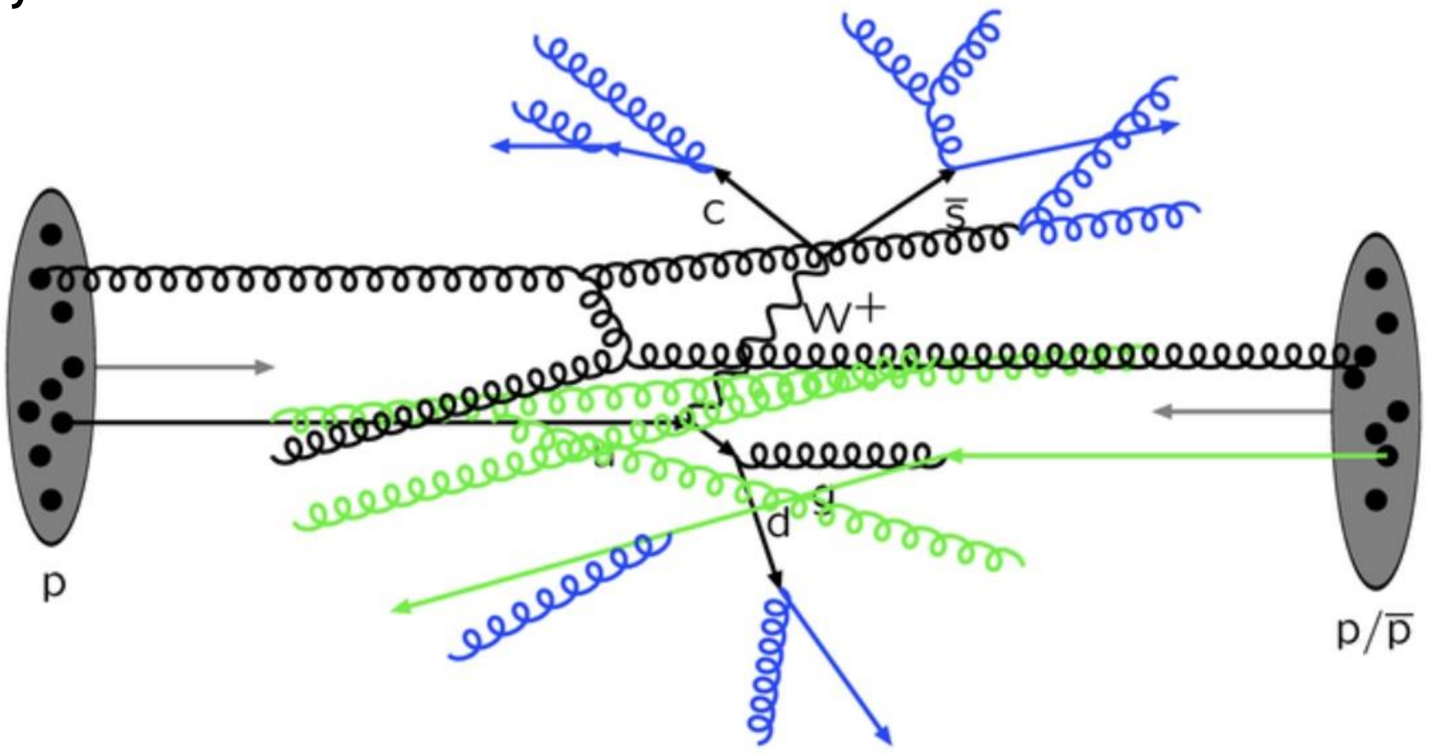
At center-of-mass energies of **13.6 TeV**, proton collisions probe **physics around the time of the big-bang!**

Proton Collisions

Unlike electron-positron colliders, proton collisions are messy but can probe **a huge range of energies simultaneously!**

F. Betchel

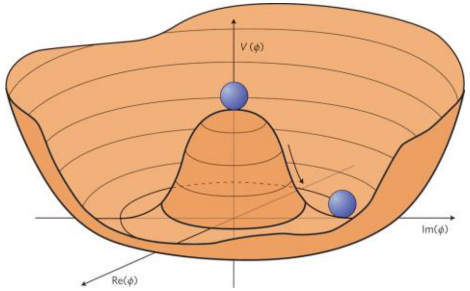


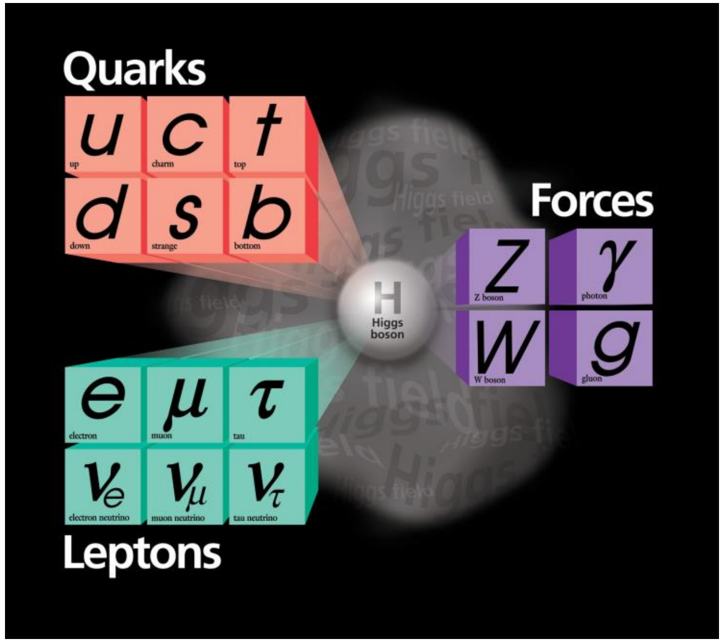


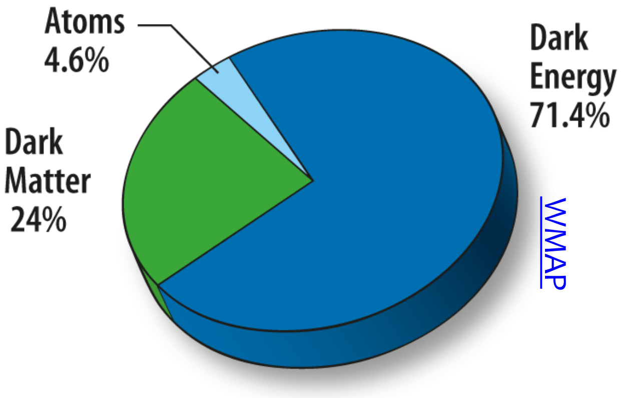
Fraction of proton momentum carried

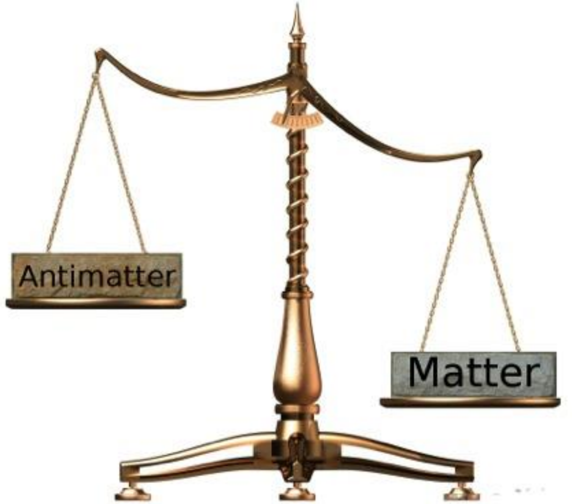
Open questions in Particle Physics

- Is the Higgs sector SM-like ?
- What is Dark Matter (DM)?
- Why is there more matter than anti-matter?
- What is the fundamental nature of neutrinos?
- What is (or is there) a quantum description of gravity?
- …



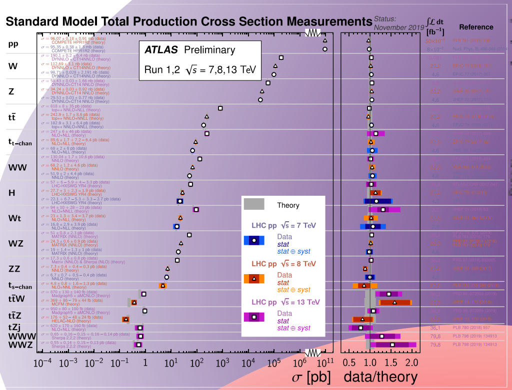


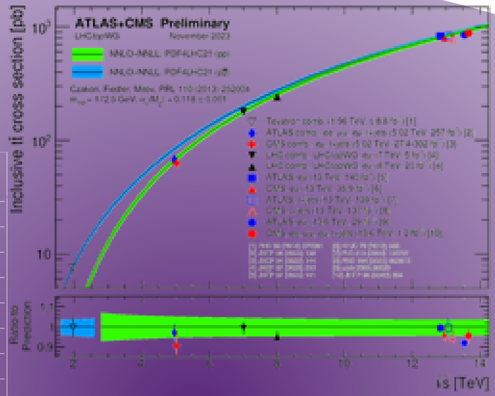


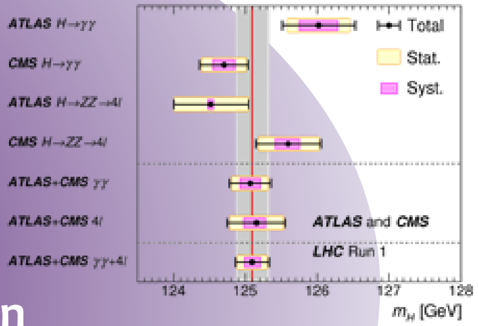


Data analyses at the LHC

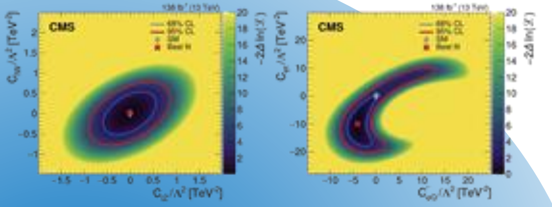
Precision SM/Top/Higgs

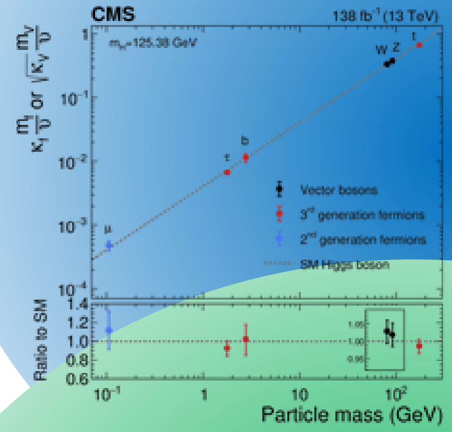


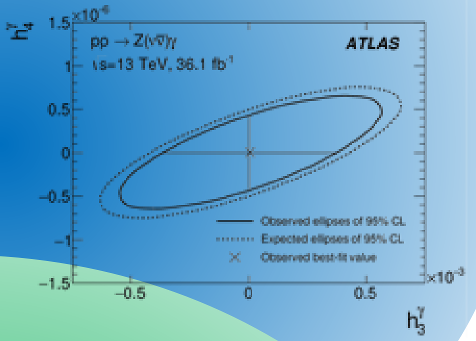




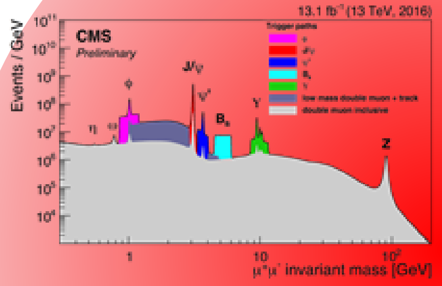
Indirect searches for BSM

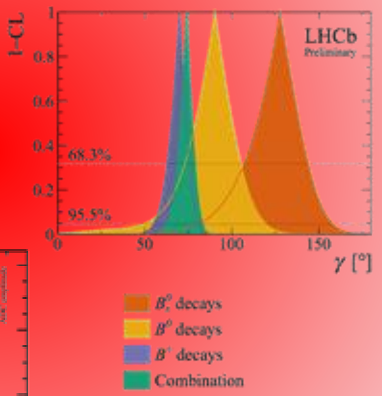


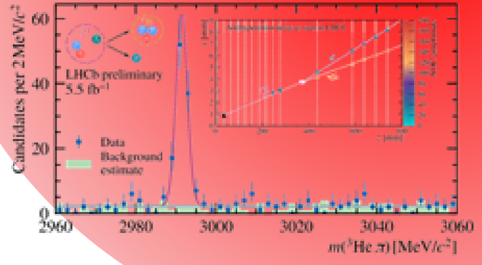




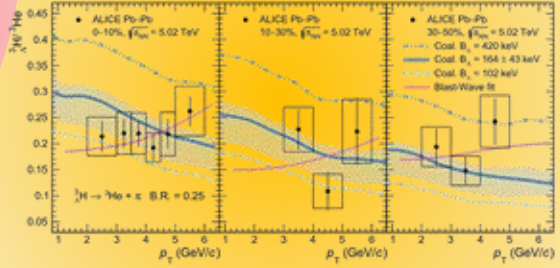
Spectroscopy & Flavour Physics



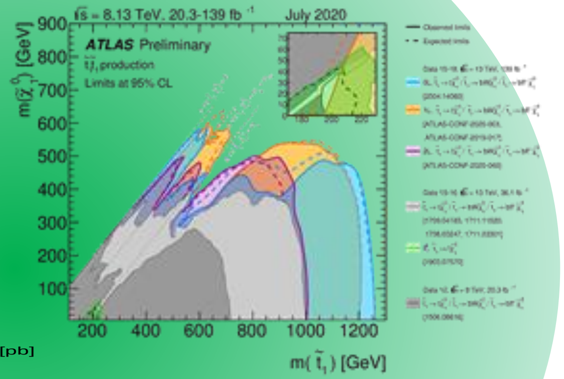


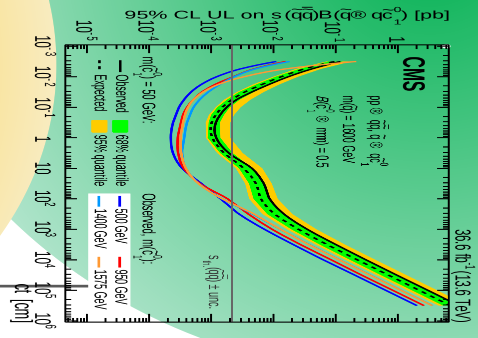


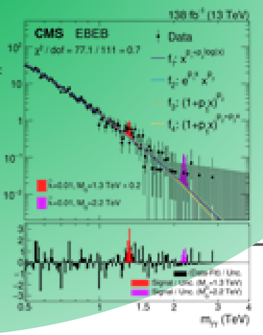
Heavy Ions



Direct searches for BSM/Exotic particles





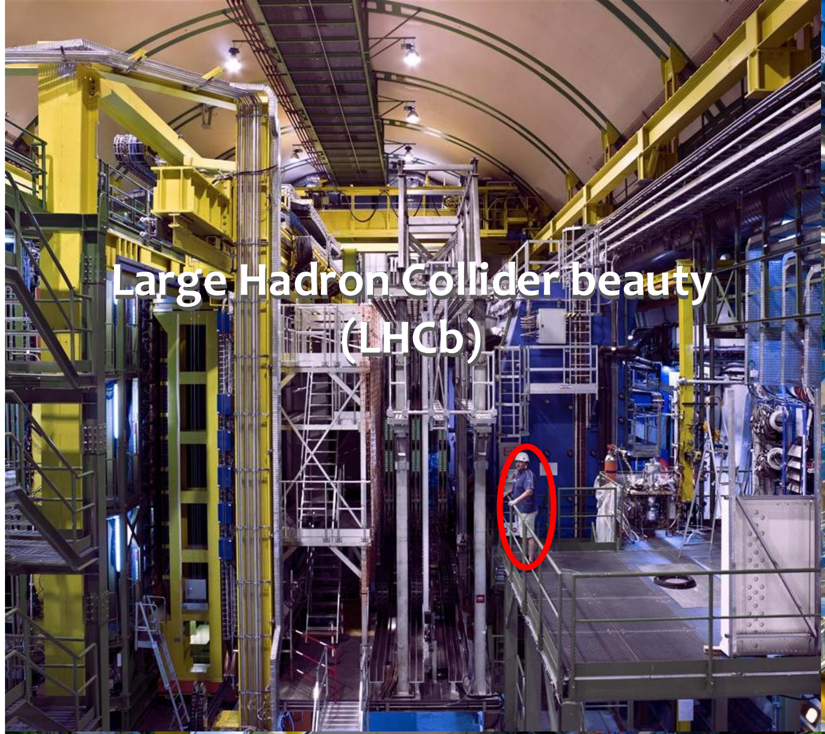


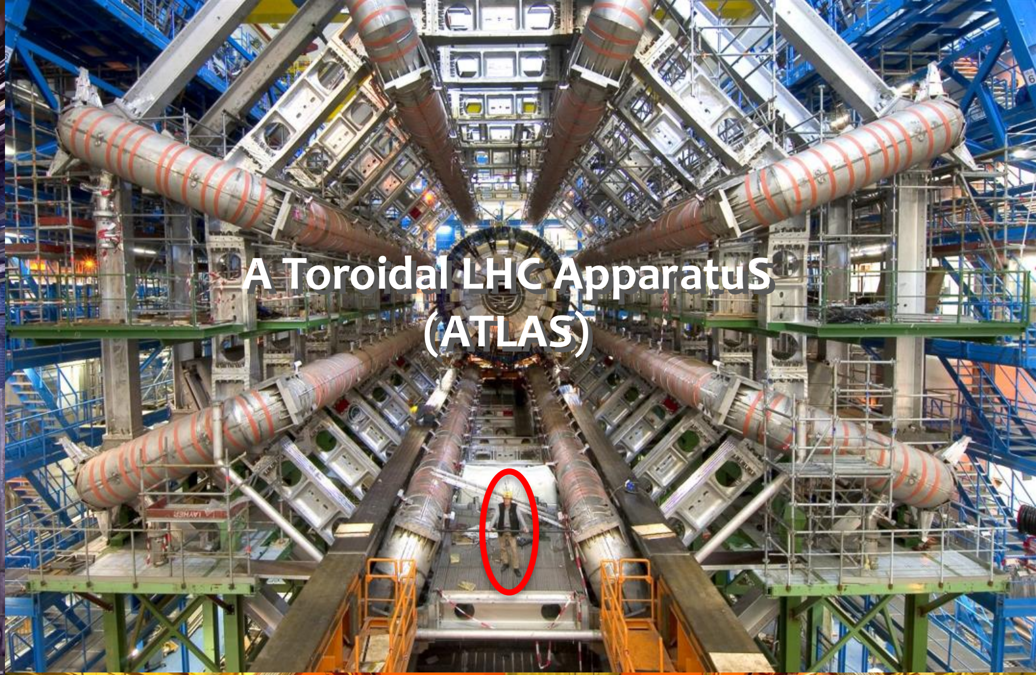
ATLAS and CMS are the two **General Purpose Detectors** at the LHC

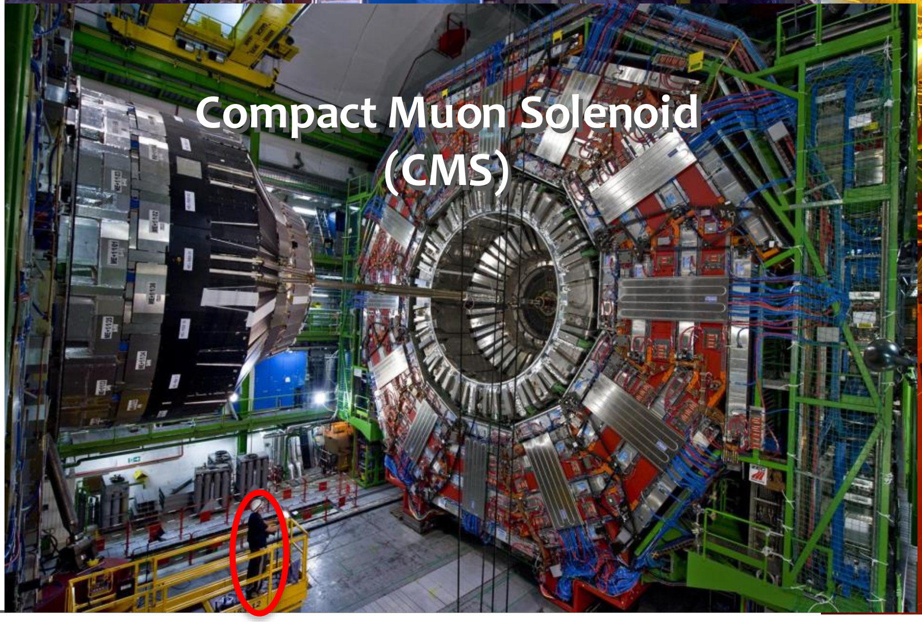
LHCb optimized for **flavour** physics and ALICE optimized for Heavy Ion collisions

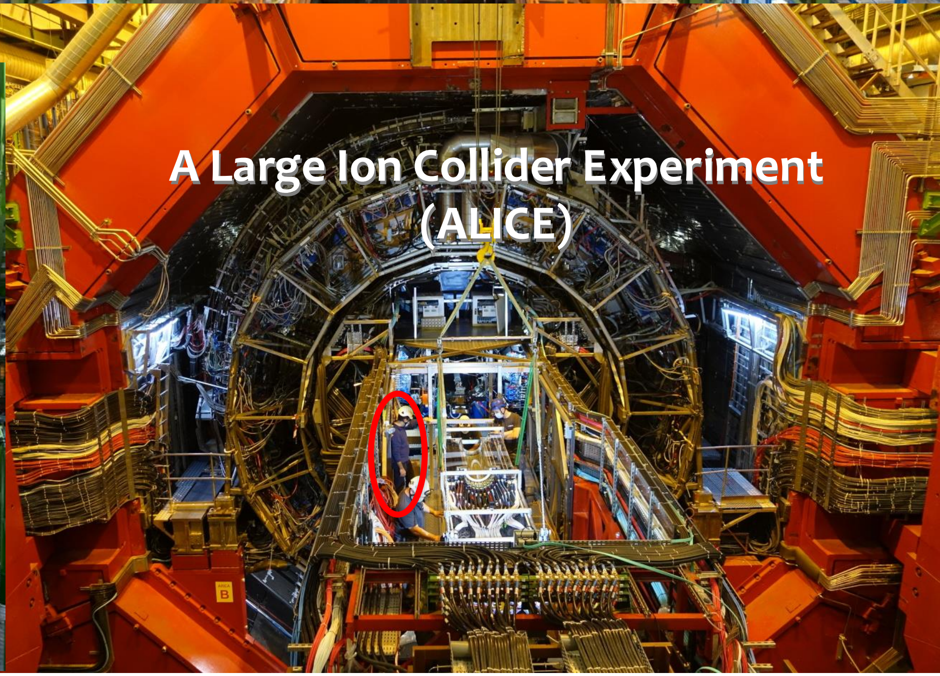
Each is designed to detect the products that are produced in the proton-proton collisions

Extremely large-scale machines are required to reconstruct the microscopic events









Data

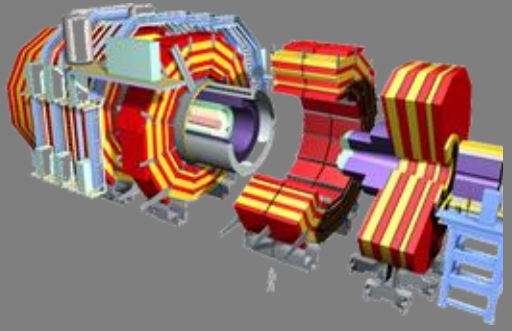
Reconstruction
&
Particle ID

Calibrations

Event Selections
&
Distributions

RESULTS

Data



Reconstruction

&

Particle ID

Calibrations

Event Selections

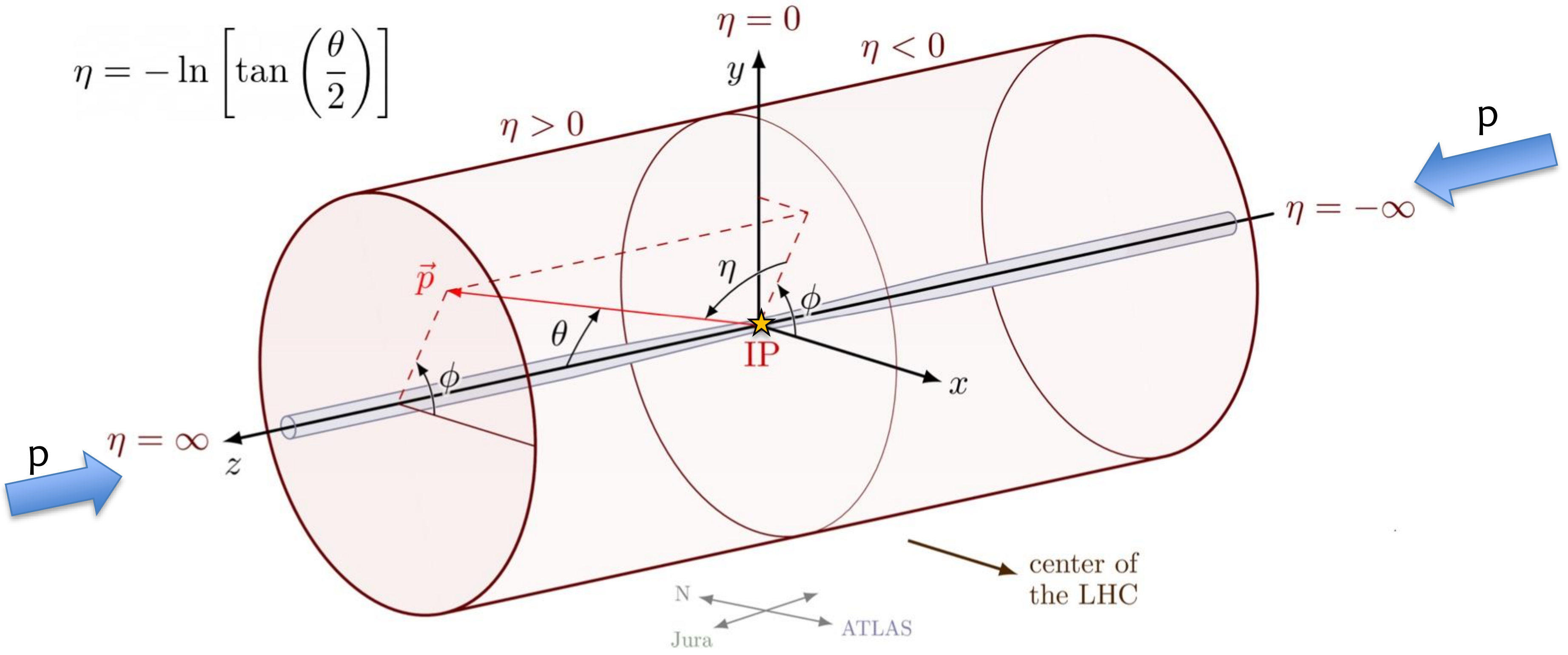
&

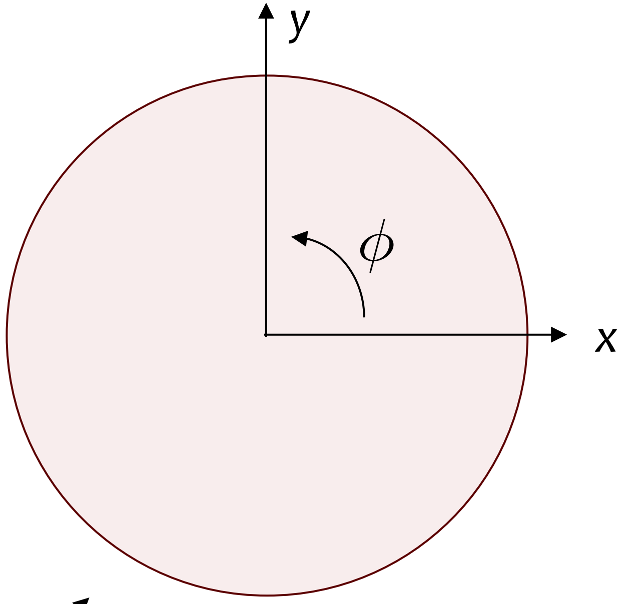
Distributions

RESULTS

Co-ordinate system

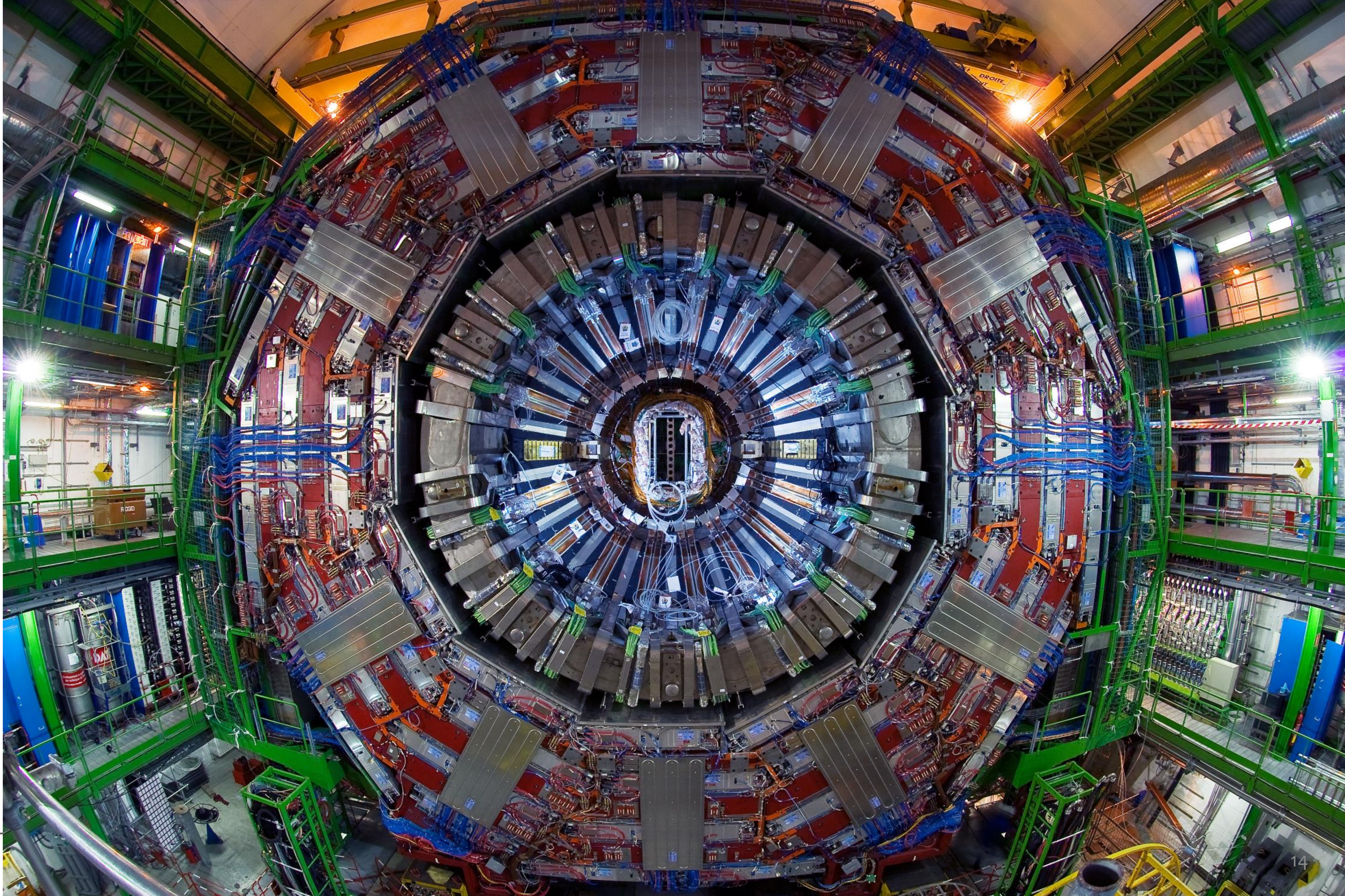
Co-ordinate system chosen around design of detector & collision system



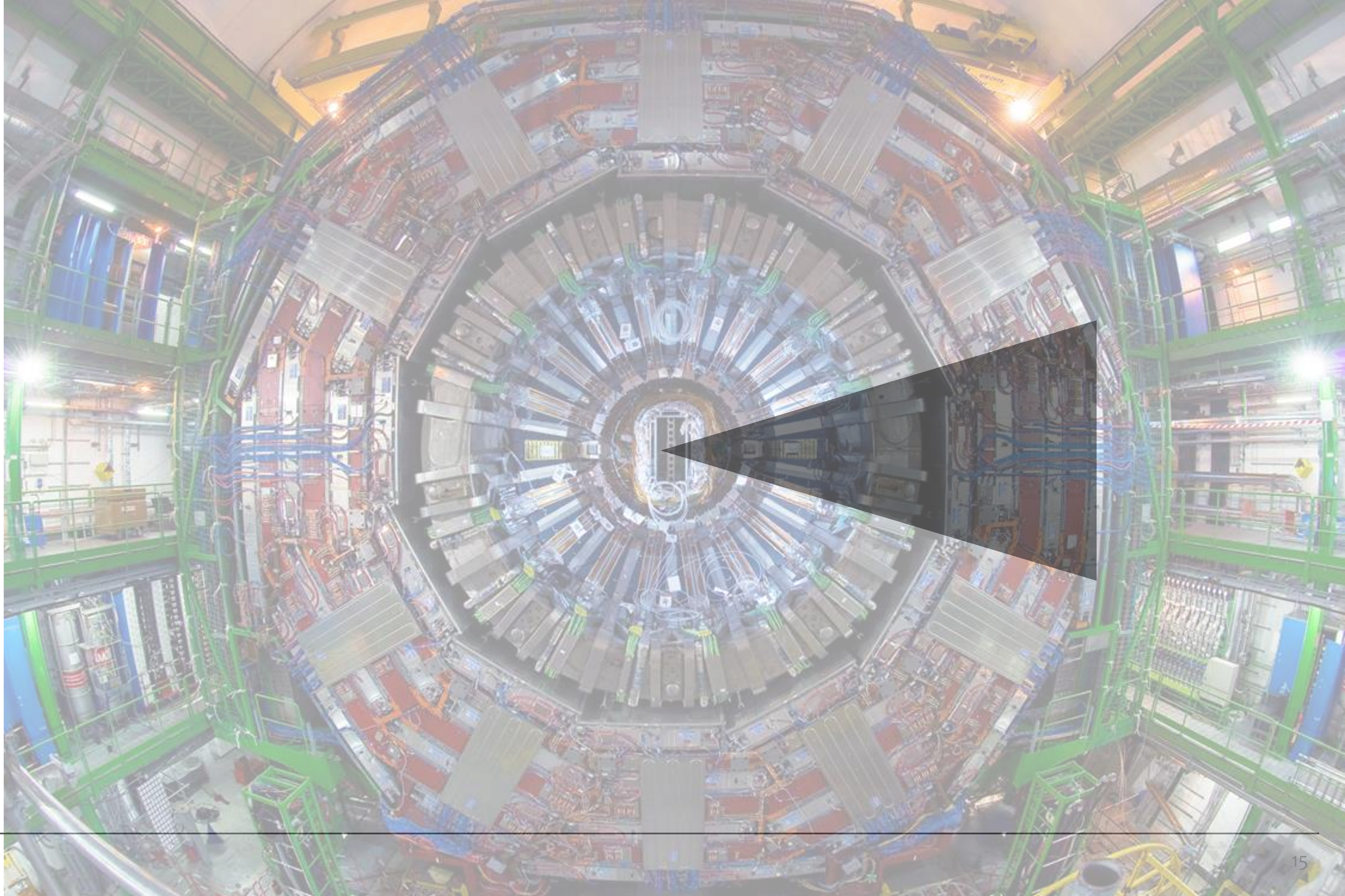


Typically deal with ***transverse*** projections as in this plane the incoming momentum is zero!

The CMS Detector

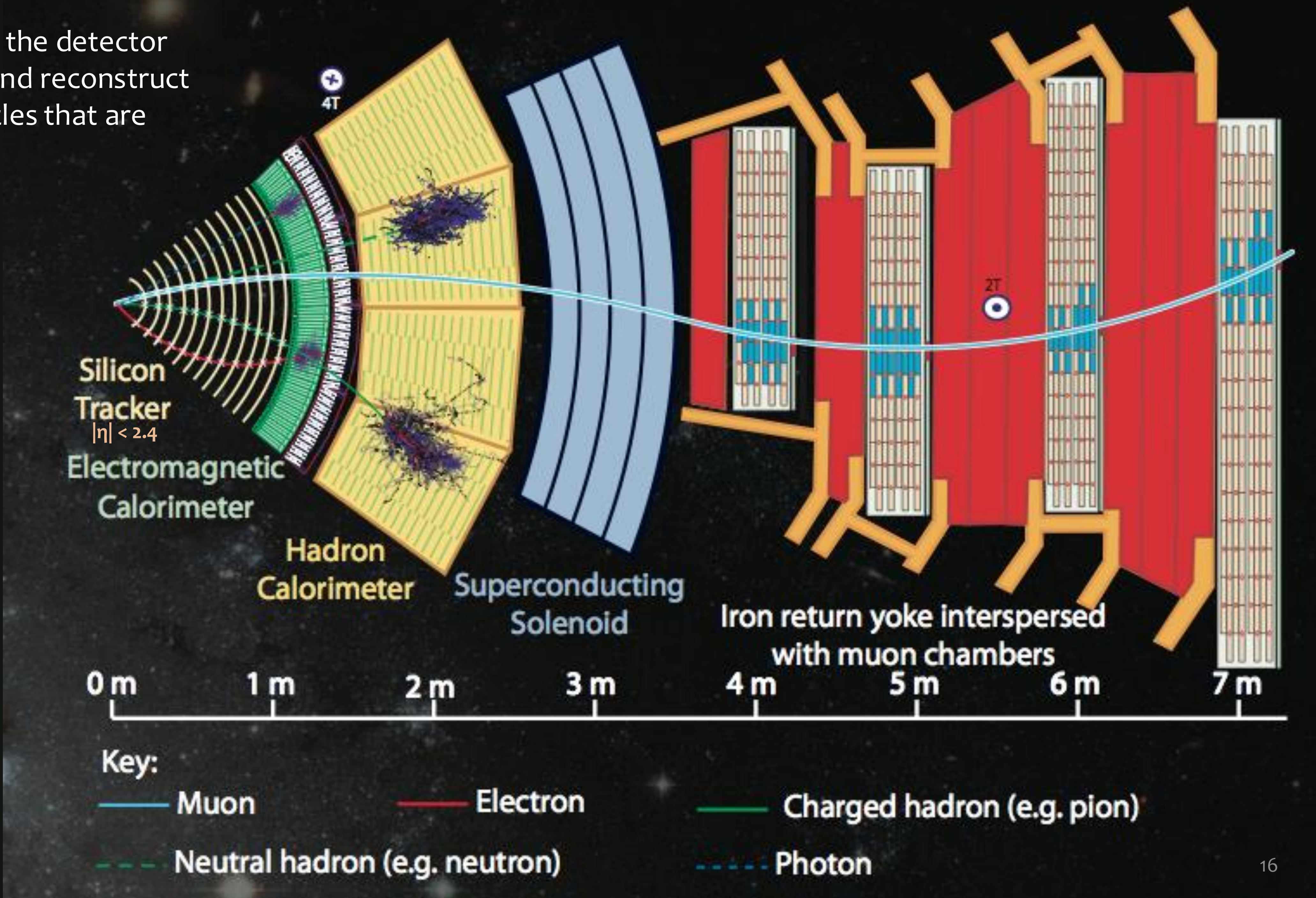


The CMS Detector

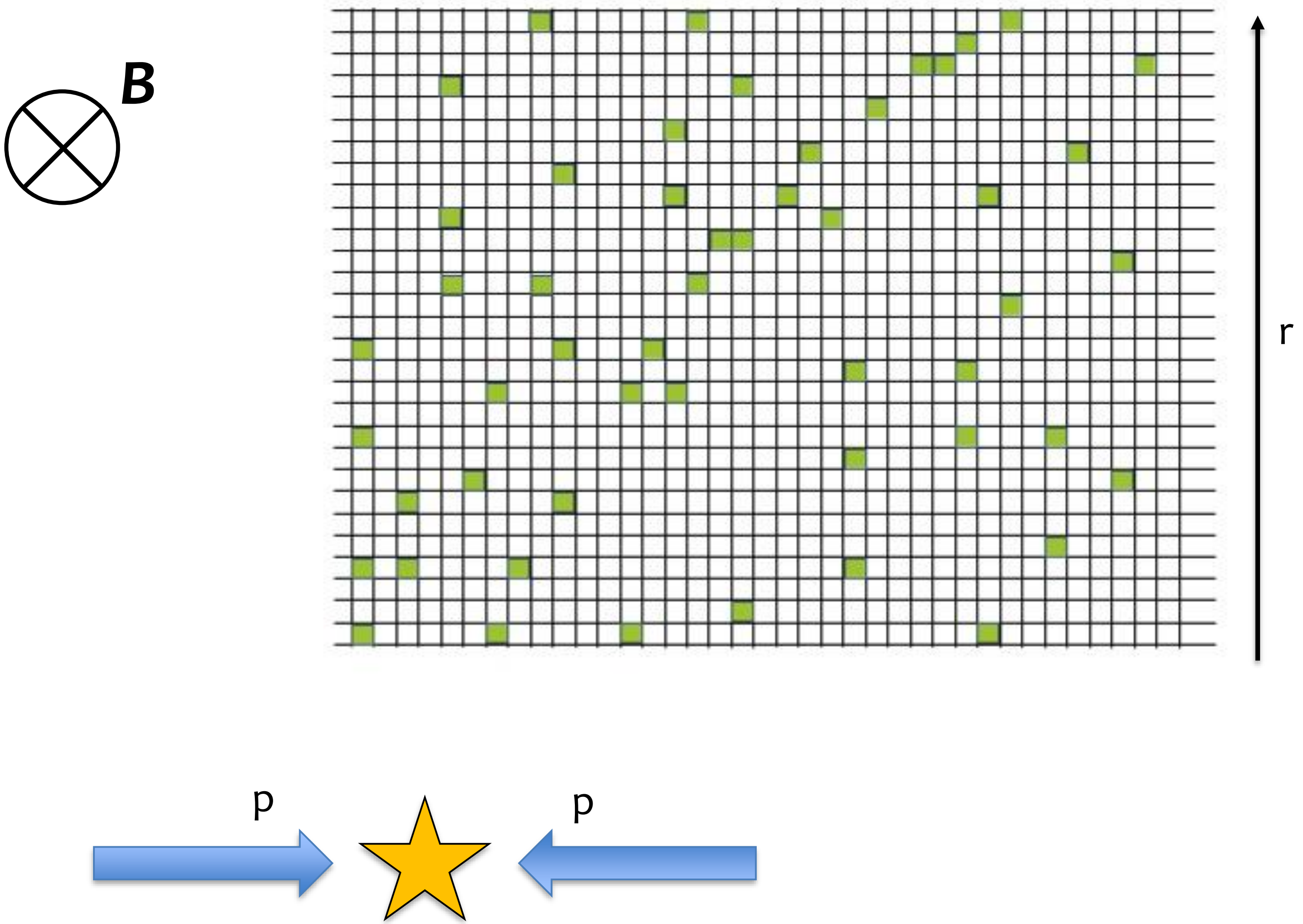


Different elements of the detector designed to identify and reconstruct different stable particles that are produced

The CMS Detector



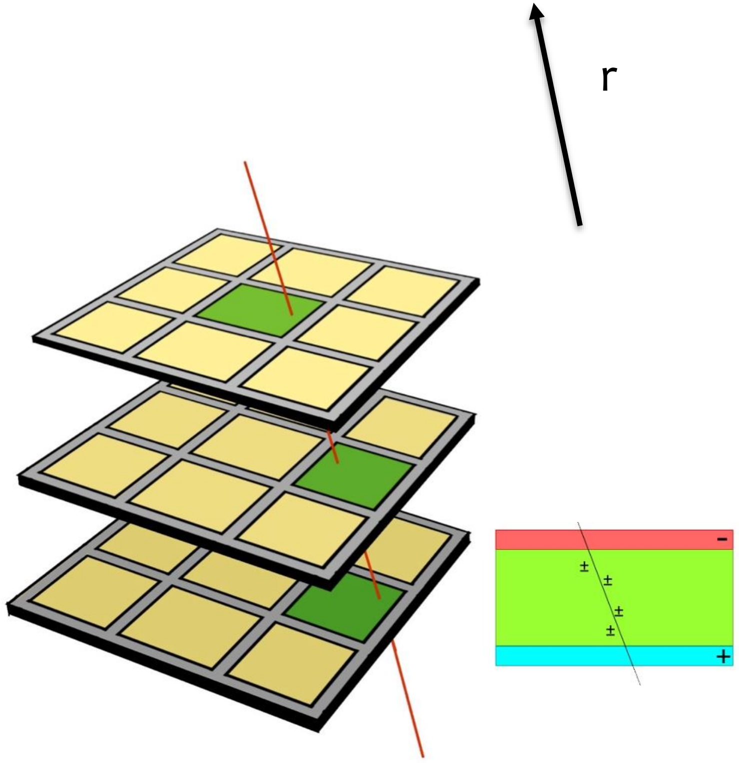
Forming Tracks



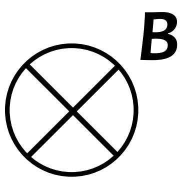
Charged particles travelling through silicon track layers (pixels/strips) will create electrons / hole pairs

→ Electrons drift where charge can be read-out

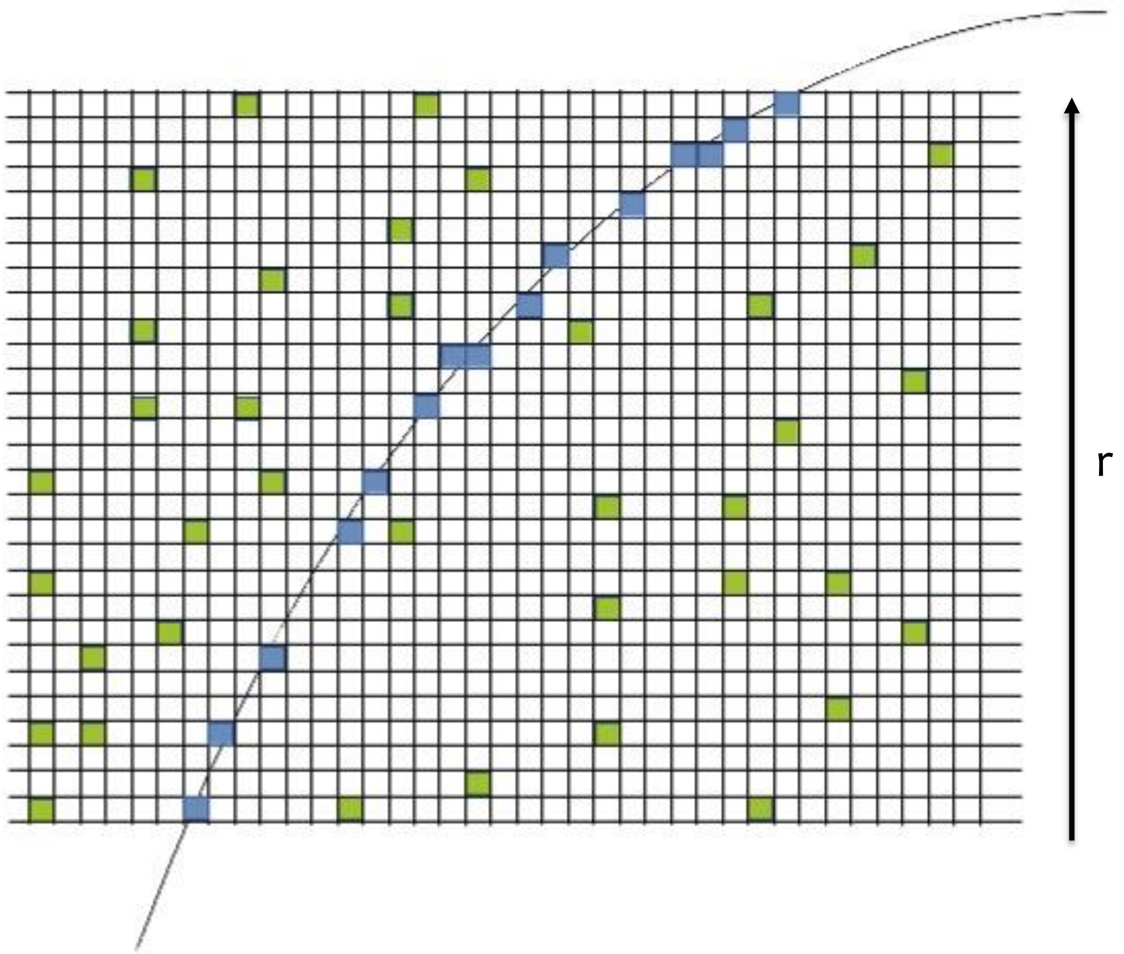
→ Localized "**hits**" in the tracker layer



Forming Tracks

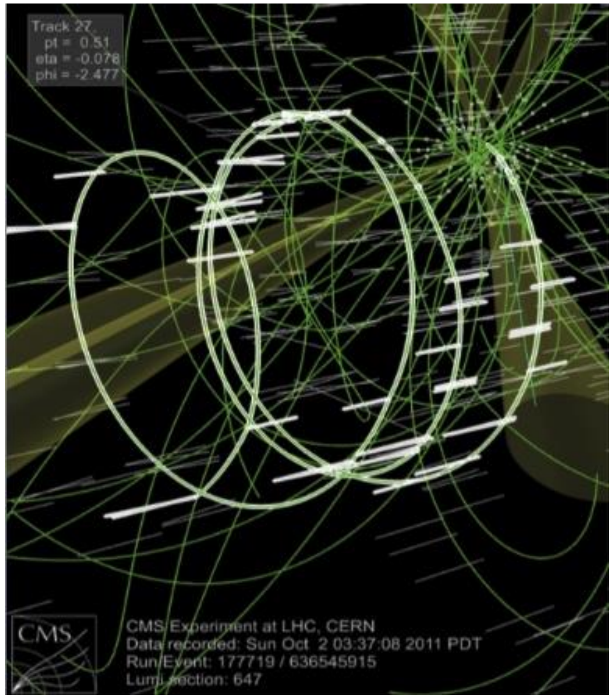


$\mathbf{F} = q\mathbf{v} \times \mathbf{B}$

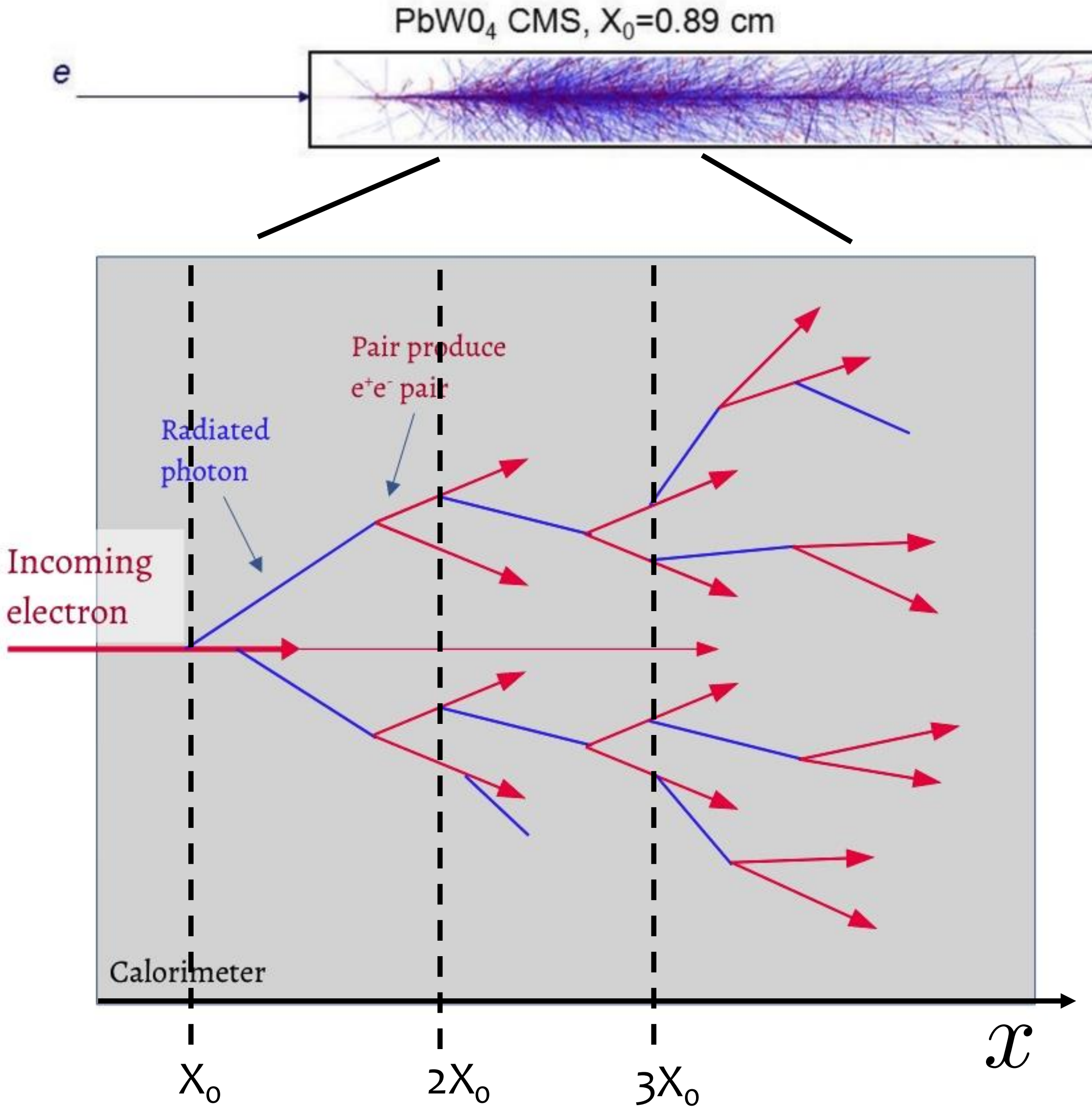


Tracking algorithm combines hits along path → track is formed!

- Radius of curvature → momentum
- Charge ID from direction of bending
- Angles of trajectory wrt beamline
- Impact parameters (offset wrt interaction point)



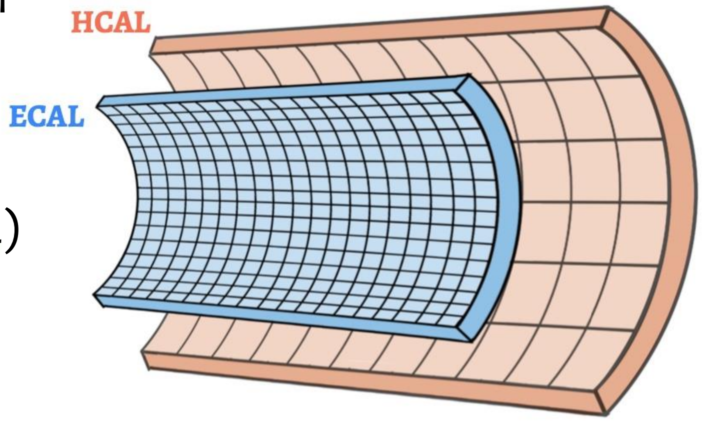
Calorimetery



Calorimeter layers are designed to absorb particle energy: E.g electron bremsstrahlung in ECAL / pair production produces showers which evolve through calorimeter material

Electromagnetic Calorimeter (ECAL) designed to stop electrons/photons

Hadronic Calorimeter (HCAL) designed to stop hadrons



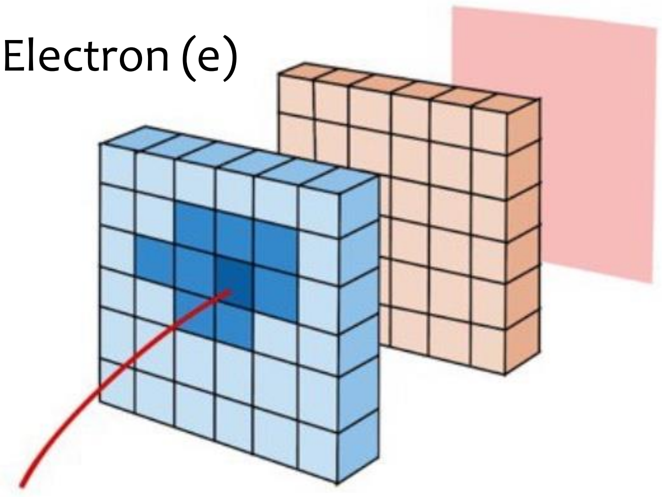
Different materials have different radiation lengths (X_0)

$$-\frac{dE}{dx} = \frac{E}{X_0}$$

Calorimetery

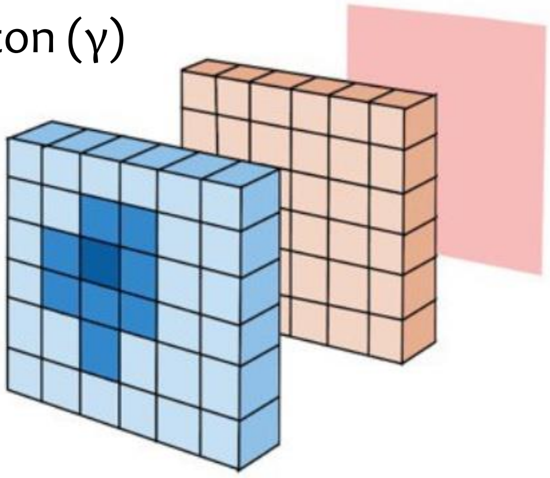
Remember that different components of our detector will respond differently to different particles

Electron (e)



Electrons and photons can be identified by deposits of energy in the ECAL without **NO** deposits in the HCAL

Photon (γ)



Track

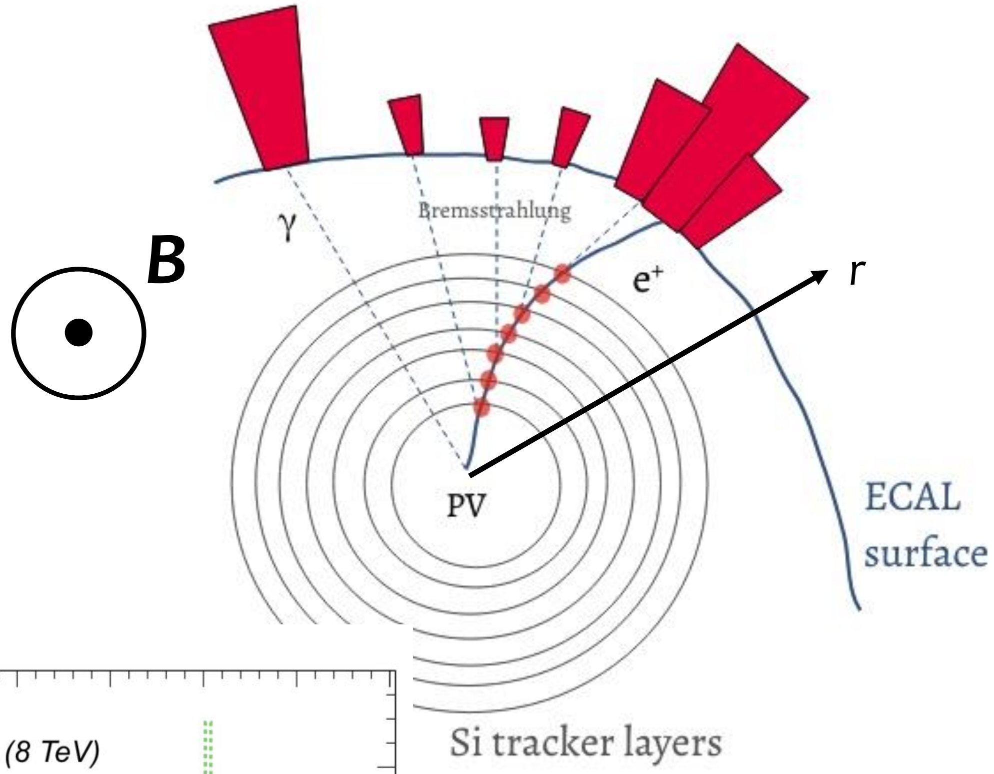
ECAL

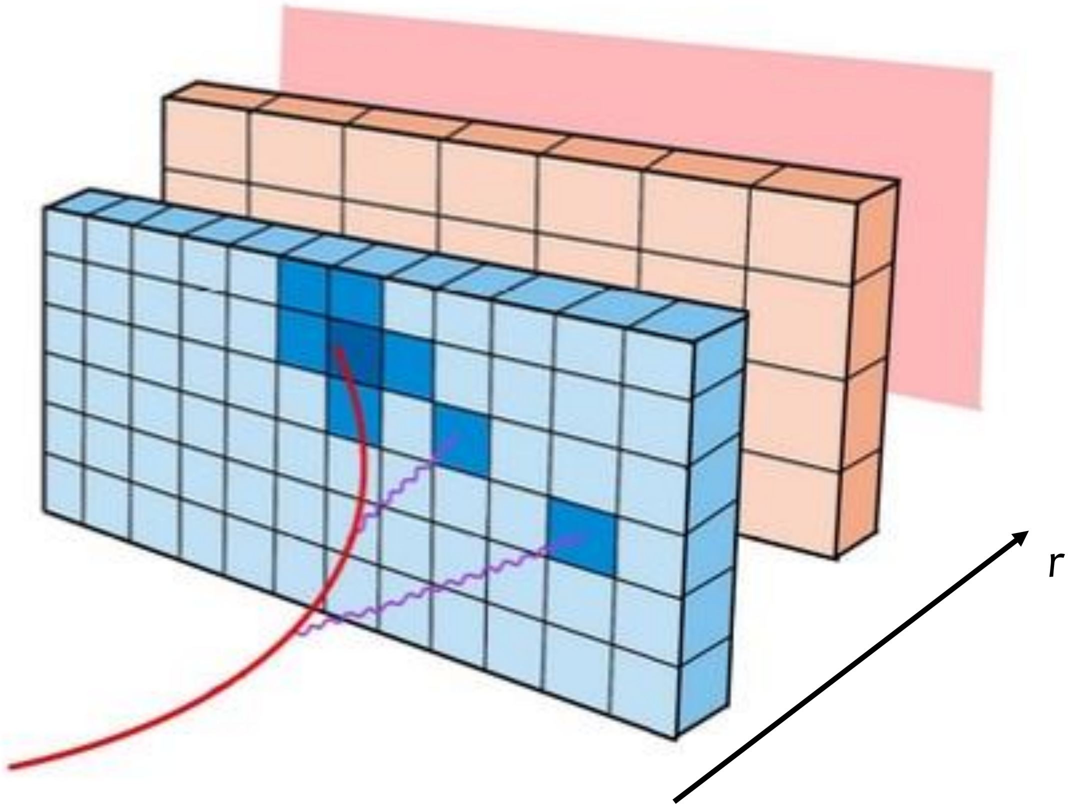
HCAL

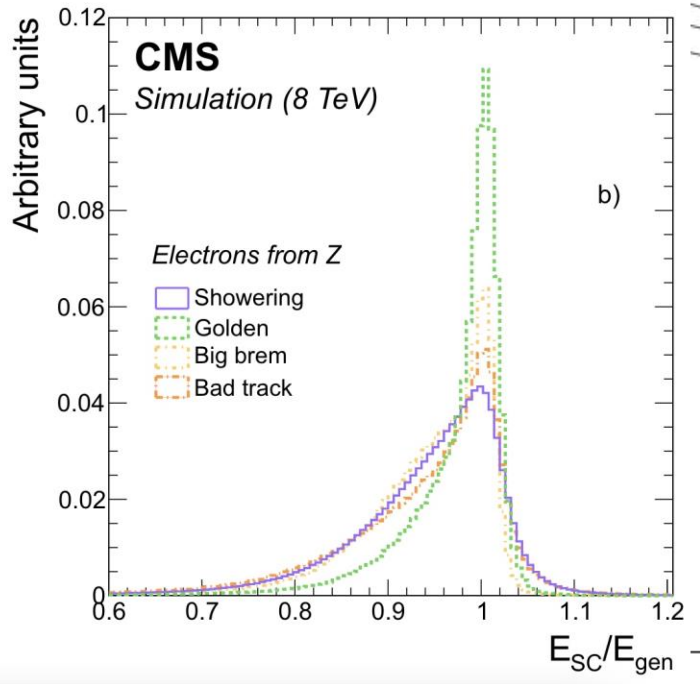
Muon Chamber

“Super clustering”

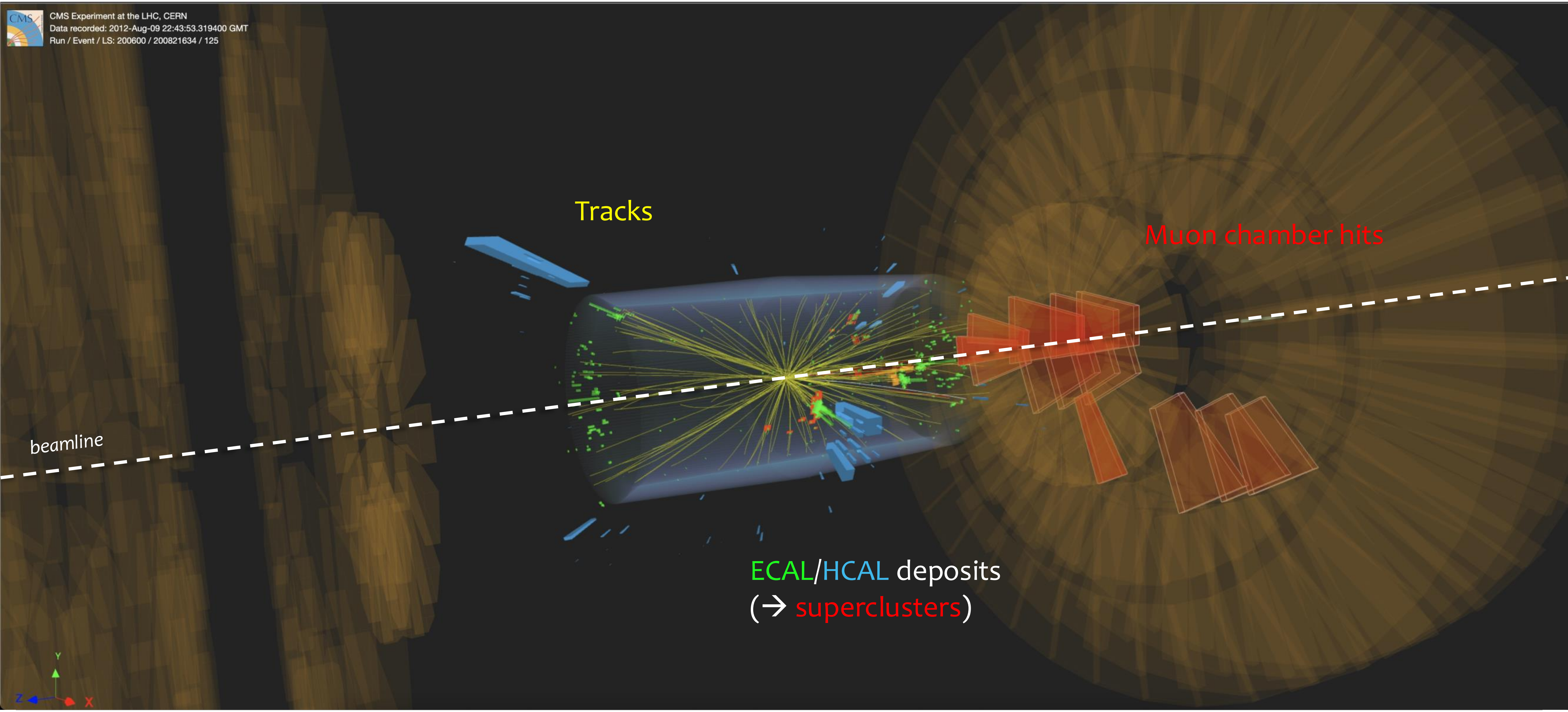
Electrons bend in the presence of a magnetic field B
→ Radiation from acceleration of charged particle
→ Photons must be included in reconstruction of electrons to maintain a good energy measurement





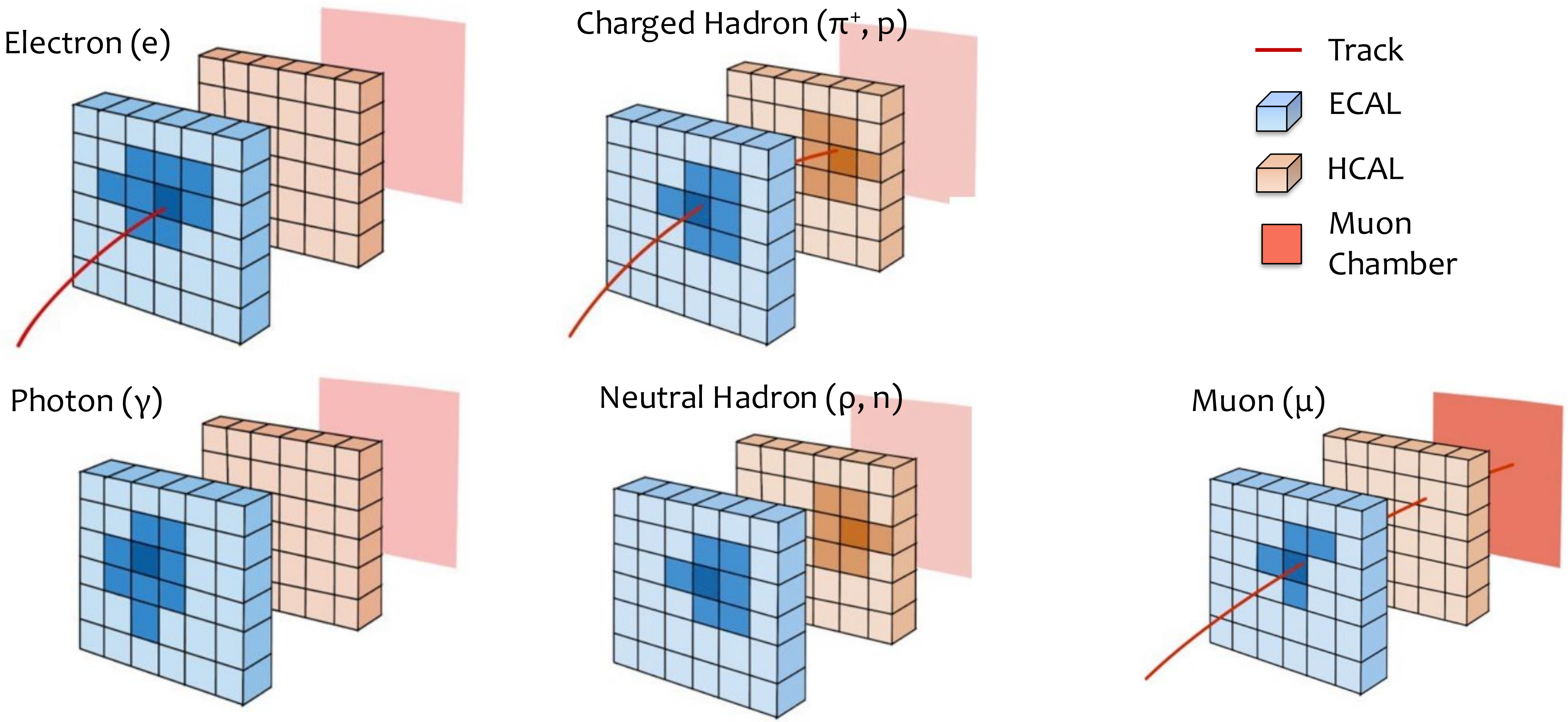


In each collision, the detector components measure energy deposits forming hits / tracks



Calorimeters

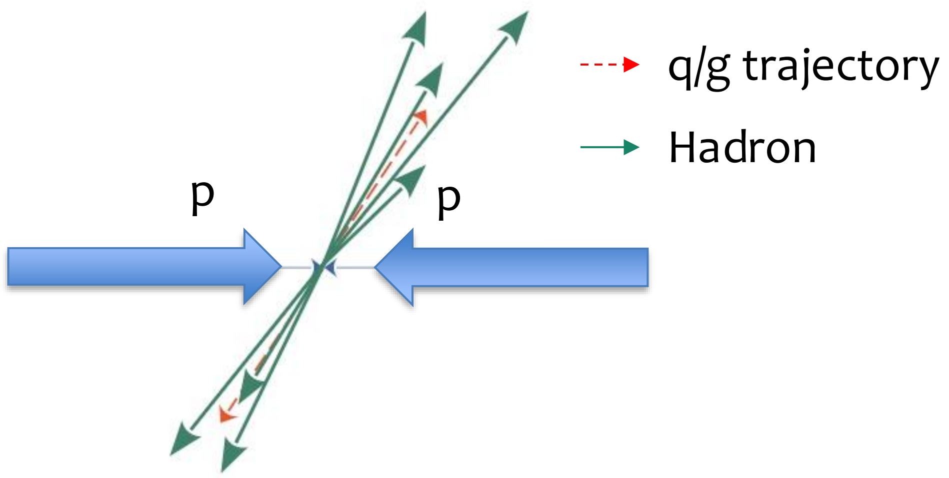
Remember that different components of our detector will respond differently to different particles

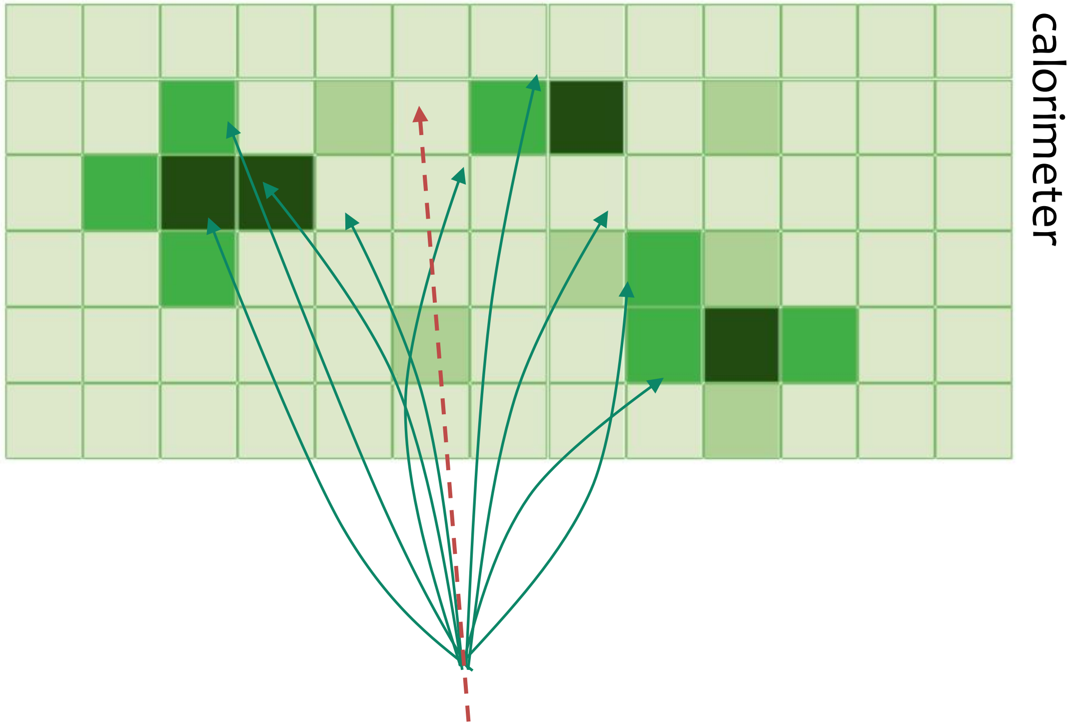


Jet Clustering

Coloured particles (quarks & gluons) produced in proton collisions do not reach the detector components

→ Part of the production energy/momentum is used to produce additional quark/antiquark pairs – which then form hadrons. It is the hadrons that exist/escape from the collision and can be detected



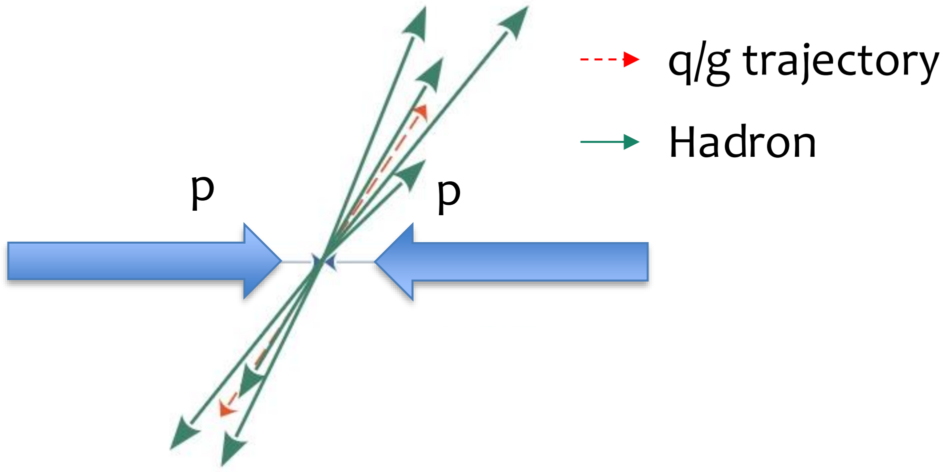


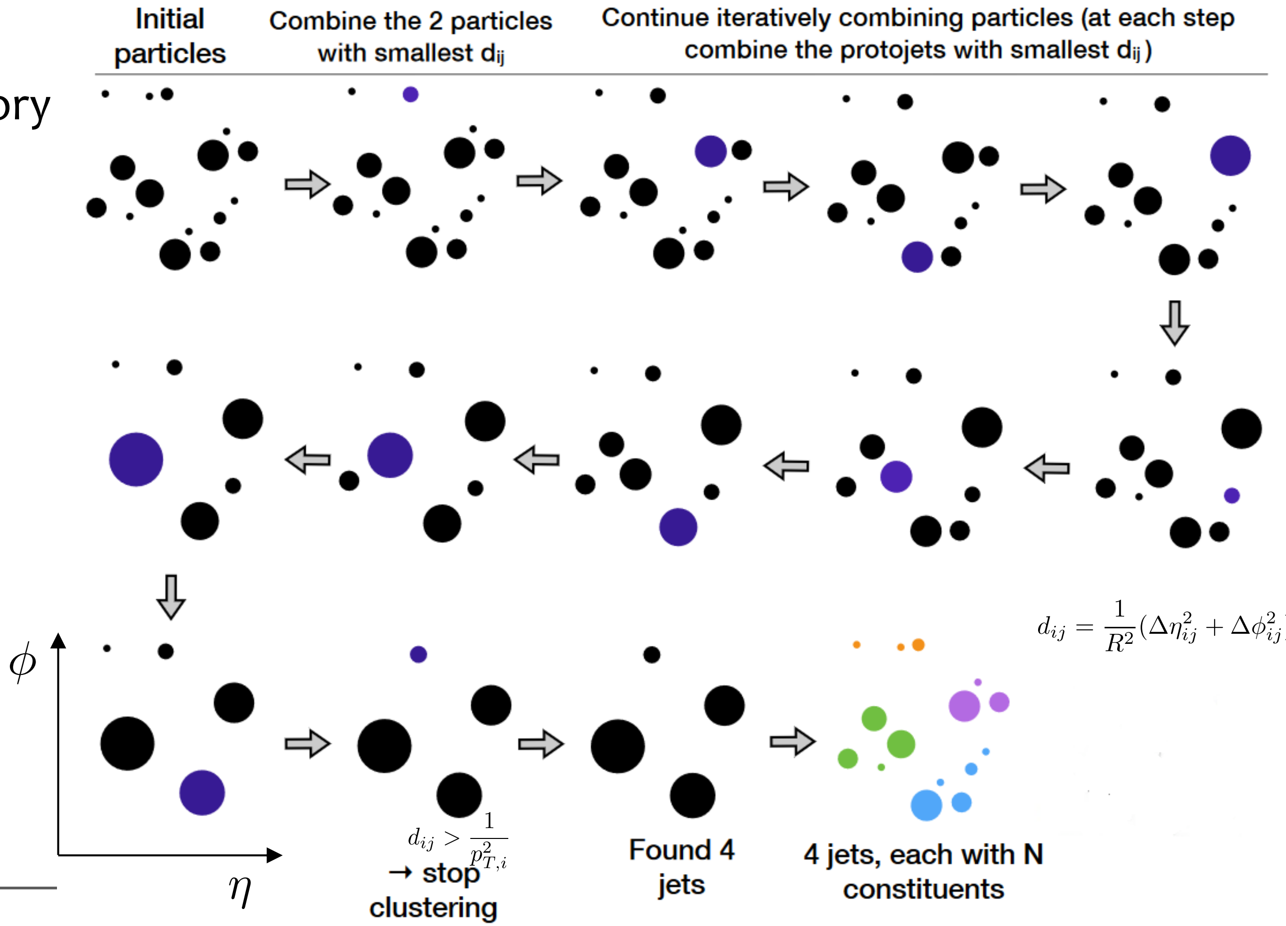
How can we determine energy & momentum of the original coloured particle?

Jet Clustering

Coloured particles (quarks & gluons) produced in proton collisions do not reach the detector components

→ Part of the production energy/momentum is used to produce additional quark/antiquark pairs – which then form hadrons. It is the hadrons that exist/escape from the collision and can be detected





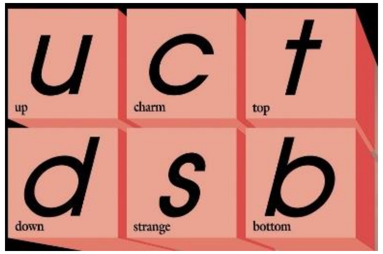
$$d_{ij} = \frac{1}{R^2}(\Delta\eta_{ij}^2 + \Delta\phi_{ij}^2) \cdot \min\left(\frac{1}{p_{T,i}^2}, \frac{1}{p_{T,j}^2}\right)$$

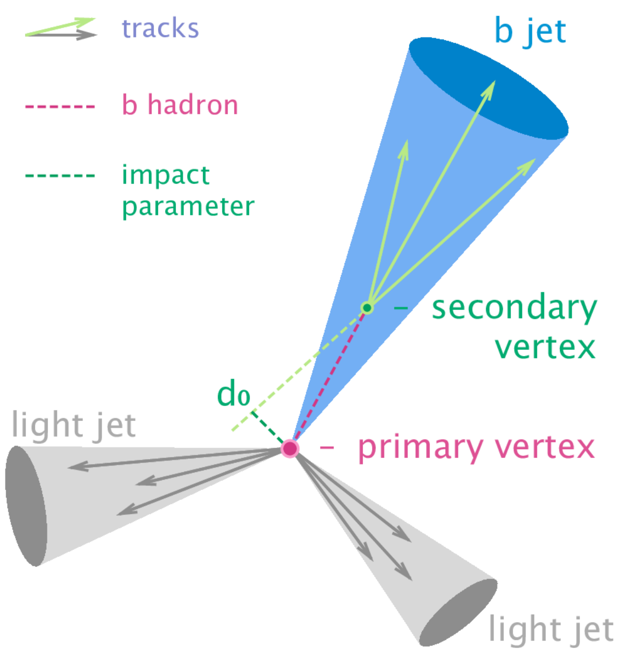
Clustering collects particles* with original quark/gluon into single four-vector using energy-momentum conservation!

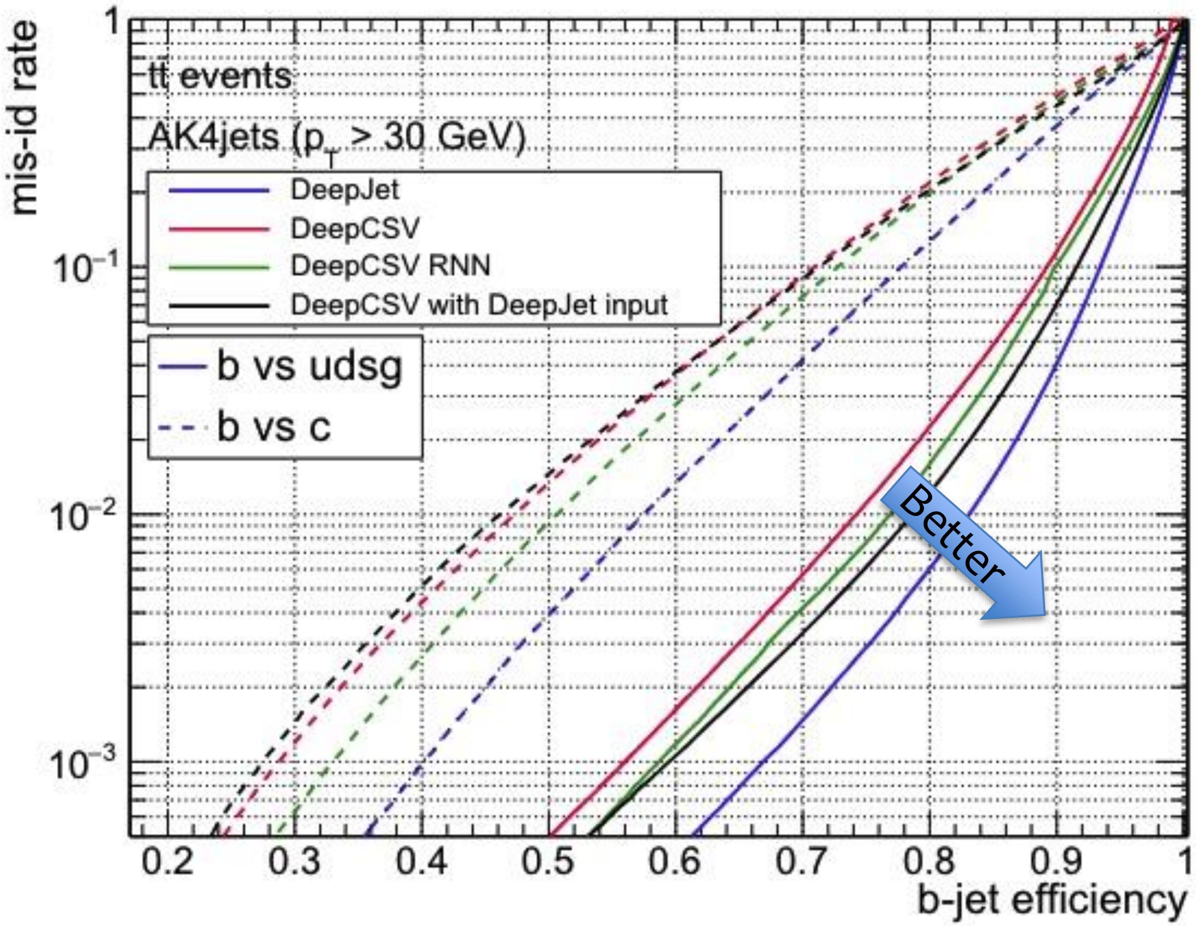
*or tracks/energy deposits …

b/c-jet

Identifying which particle initiated each jet requires lots of combined information about the constituents of the jet and the vertices it contains



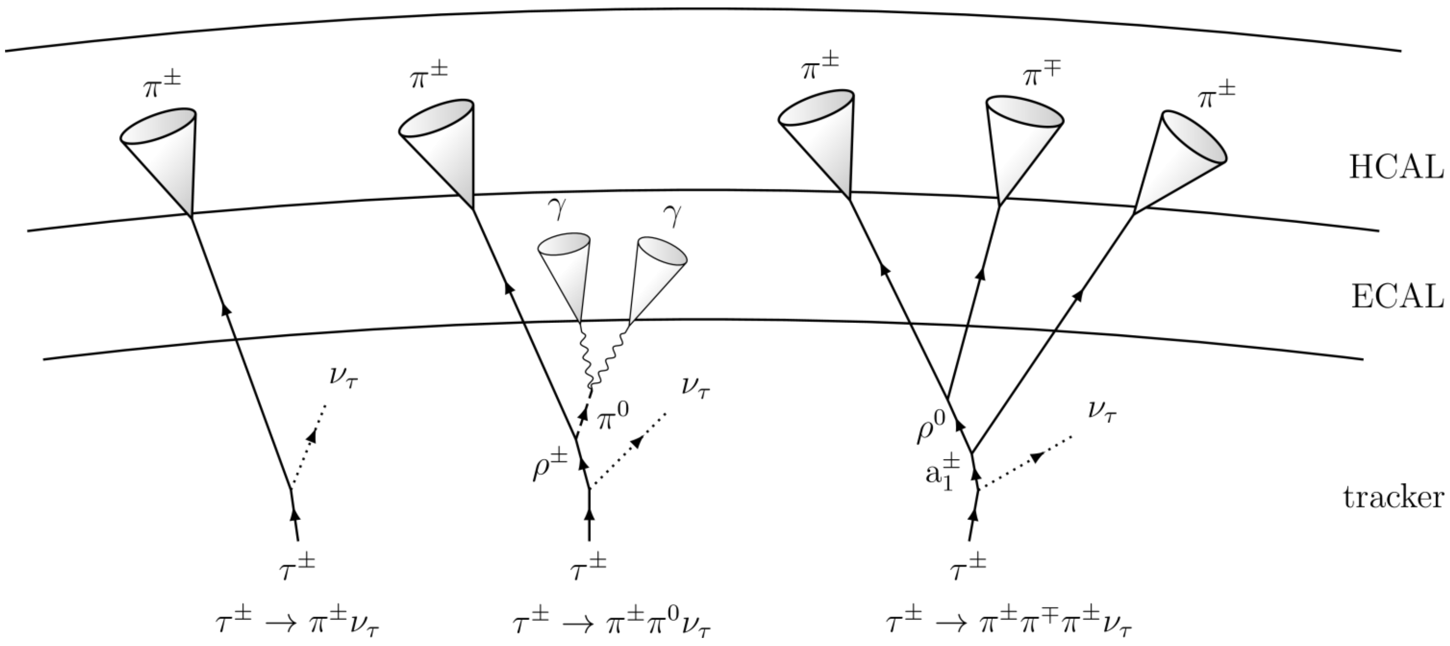




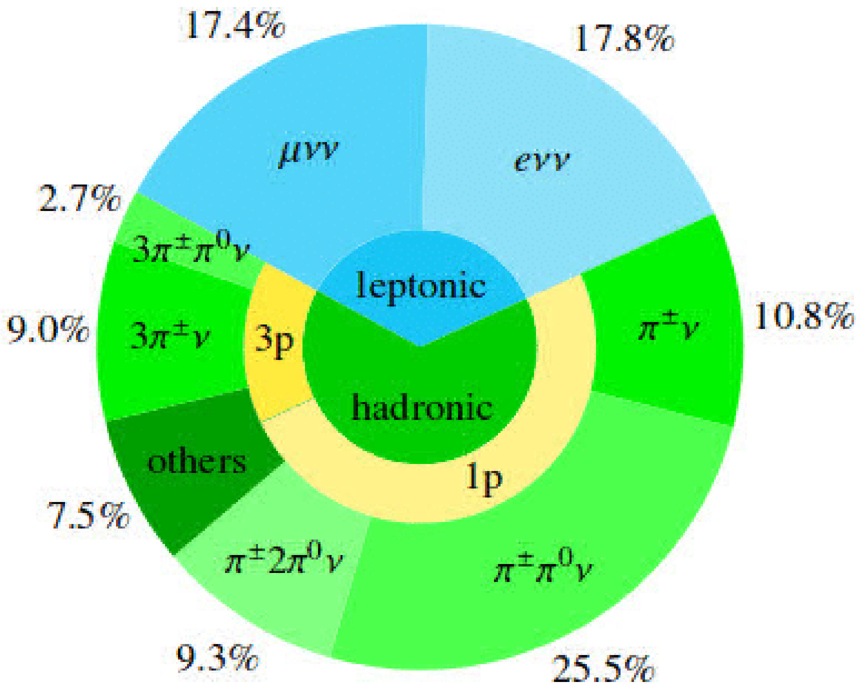
We use sophisticated **machine learning methods** to perform this task

Tau-leptons

τ leptons have very short lifetime → they decay into **leptons** or **hadronically**

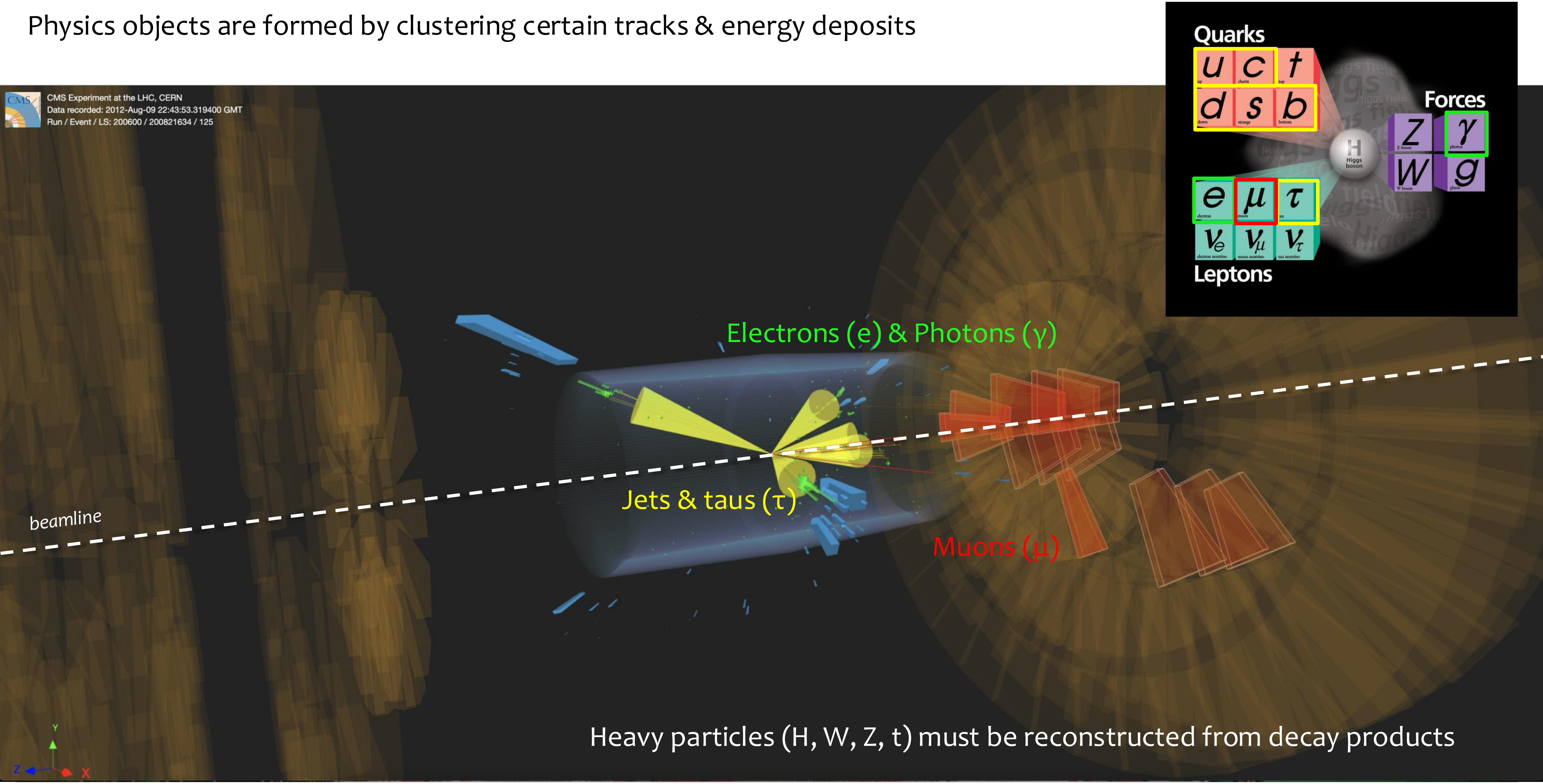


τ decay mode probabilities

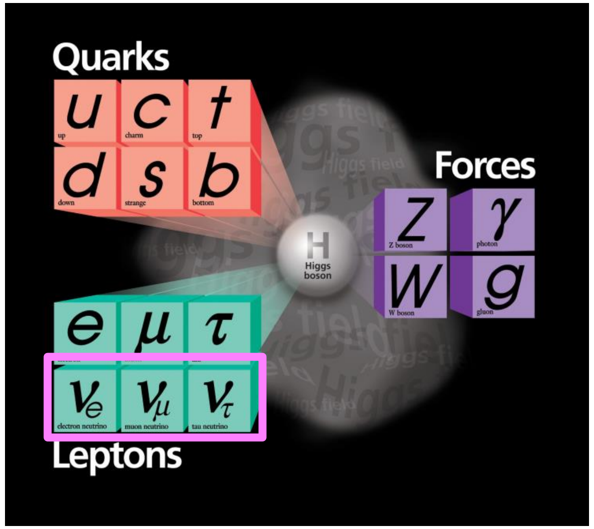


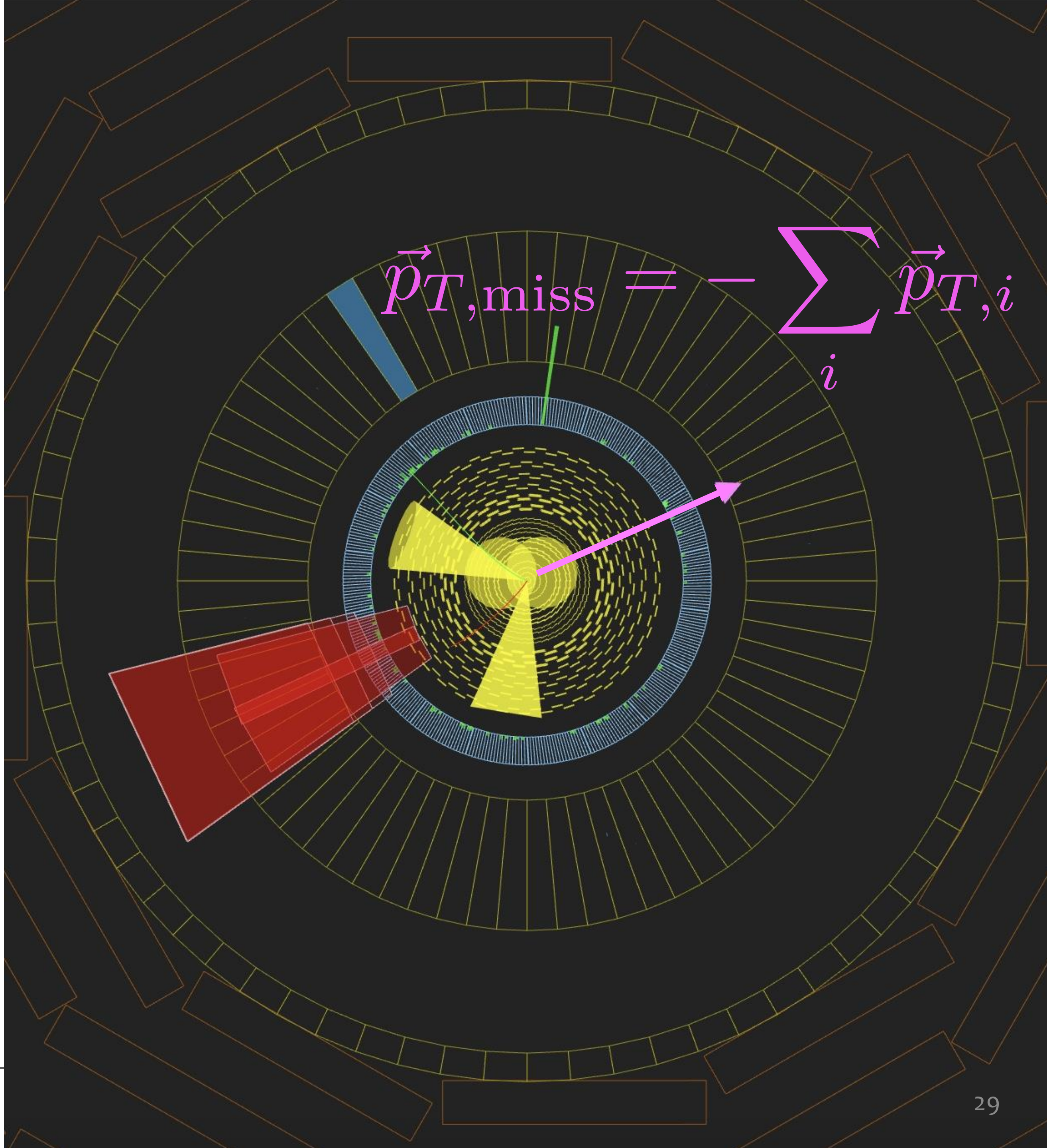
Most modern τ-ID strategies use **machine learning** to identify the decay mode and reconstruct the τ four-momentum

Physics objects are formed by clustering certain tracks & energy deposits



Missing momentum

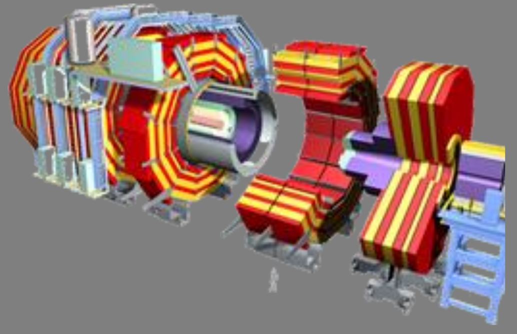




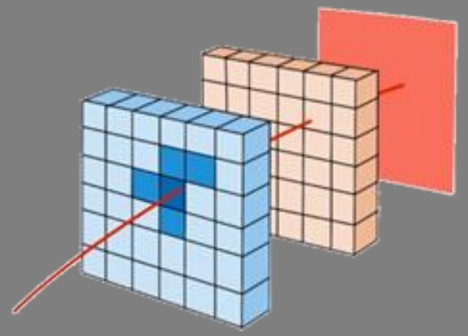
Neutrinos do not interact with any component of the detector

We infer the presence of neutrinos through an imbalance of momentum in the transverse plane
→ **missing transverse momentum**

Data



Reconstruction
&
Particle ID

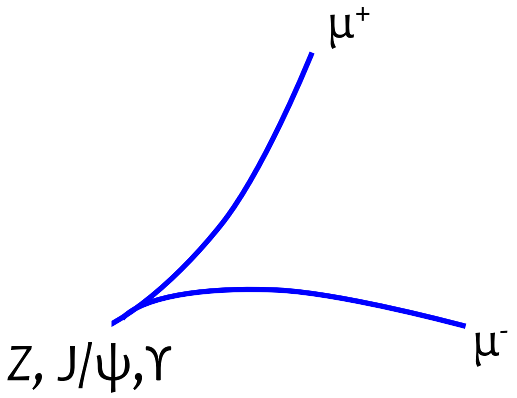


Calibrations

Event Selections
&
Distributions

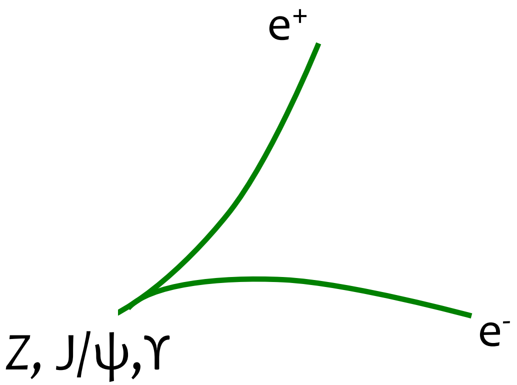
RESULTS

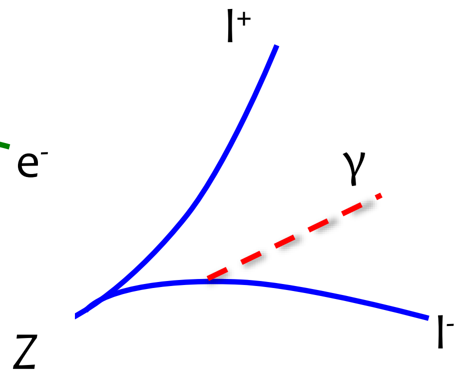
Standard Candles

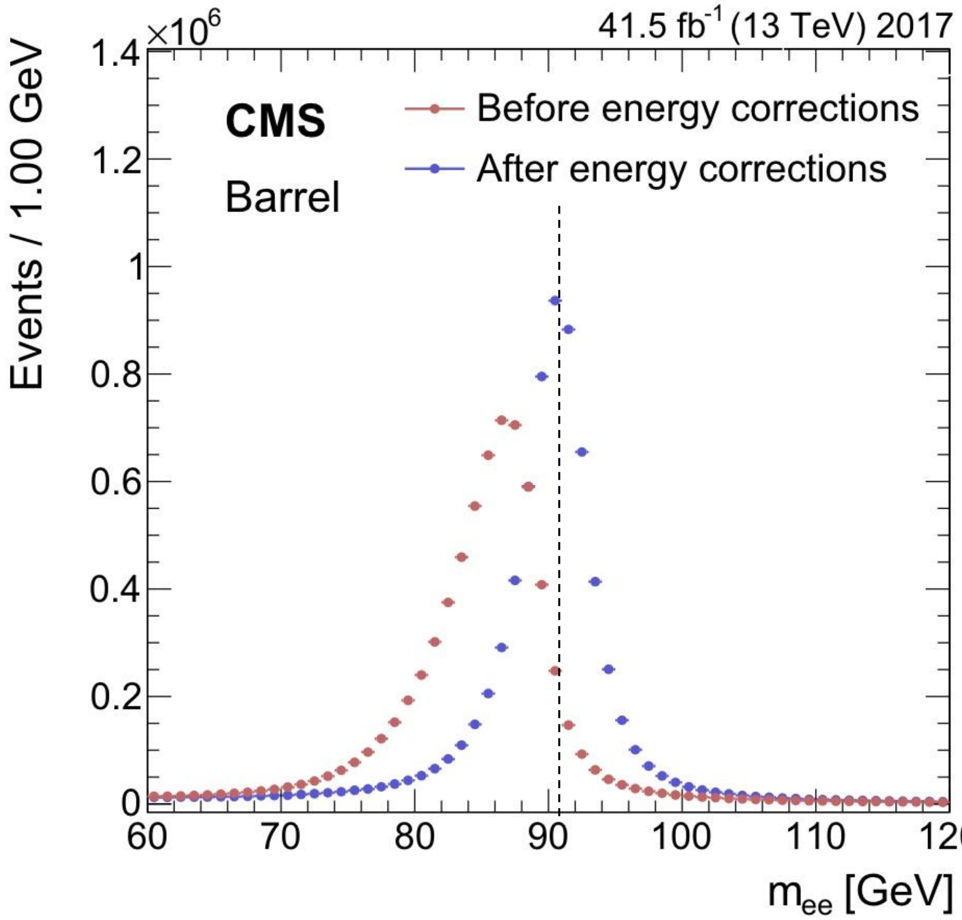


Z, J/ψ and ϒ decays in data provide standard candles to calibrate energy/momentum measurements

Large source of clean events with a well described mass peak

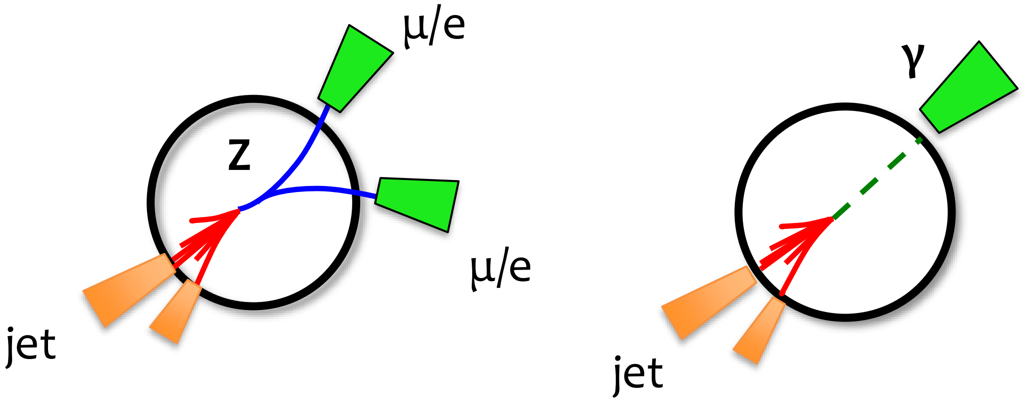




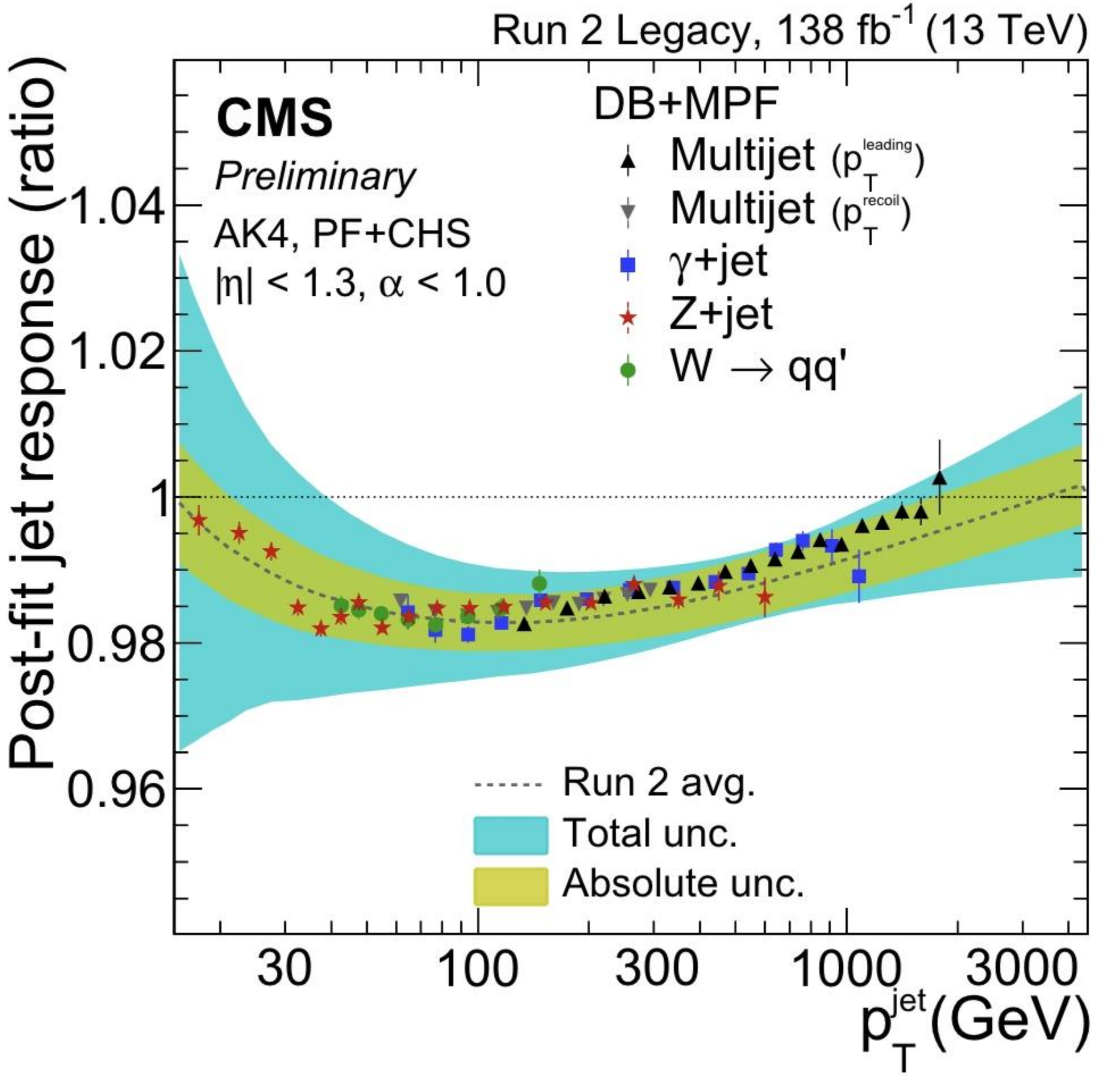


Relative corrections

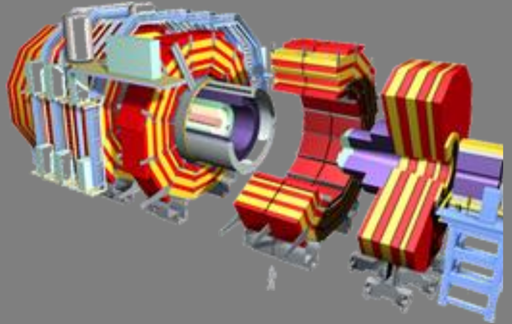
More complicated objects (eg jets) require several stages of correction → Use previously calibrated objects to calibrate jet momentum!



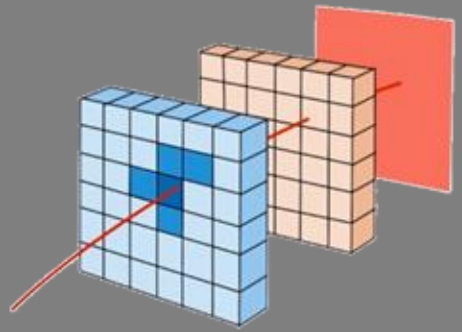
$$R_{\text{jet},p_T} = \frac{p_{T,jet}}{p_{T,Z/\gamma}}$$



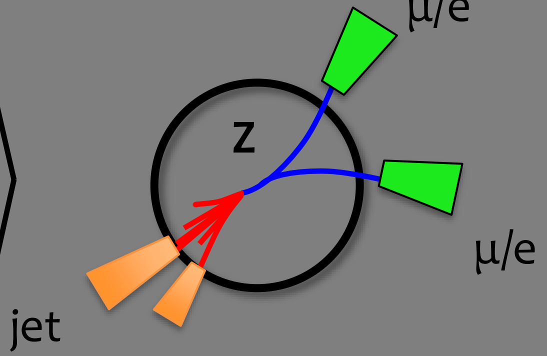
Data



Reconstruction & Particle ID

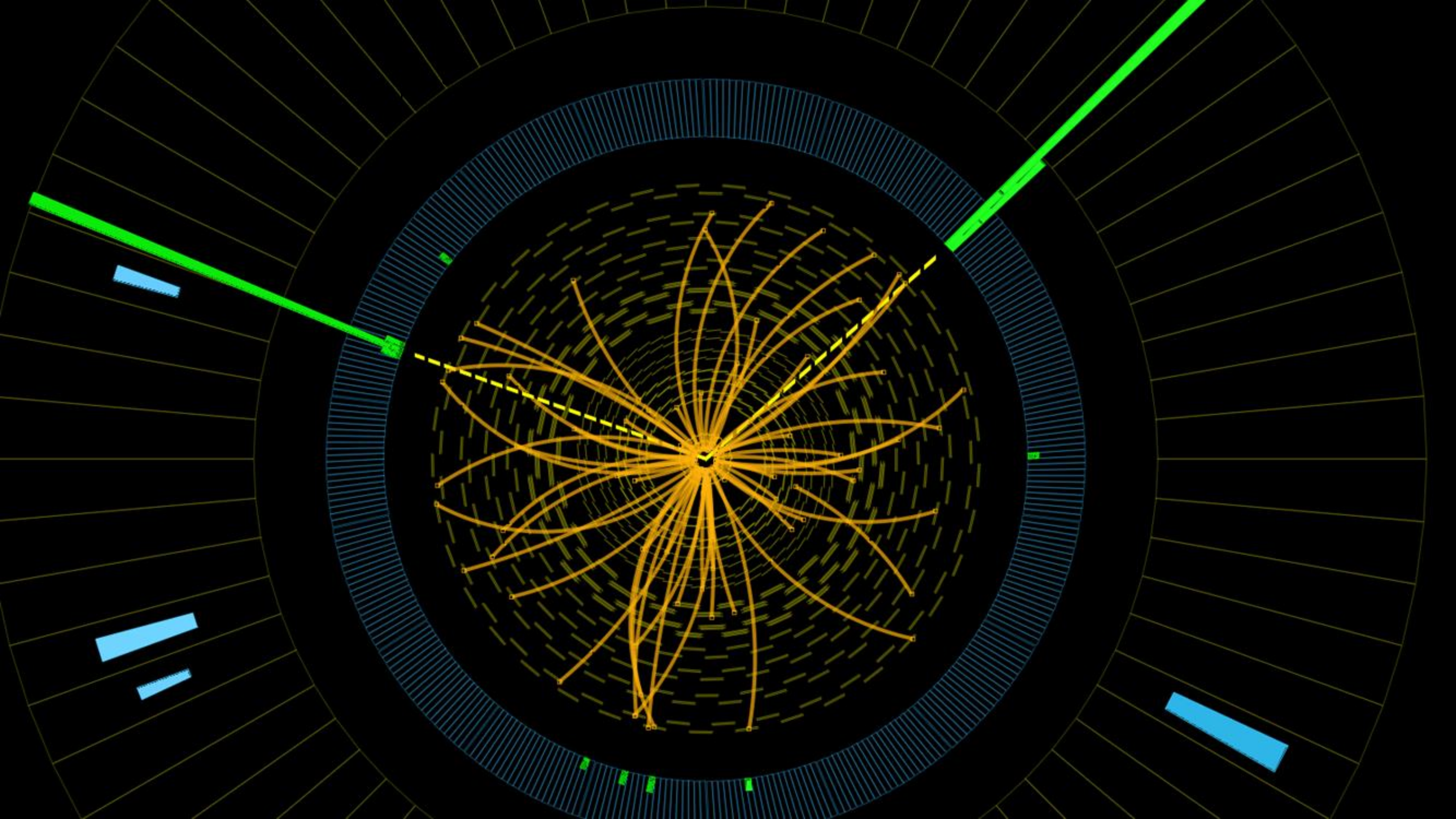


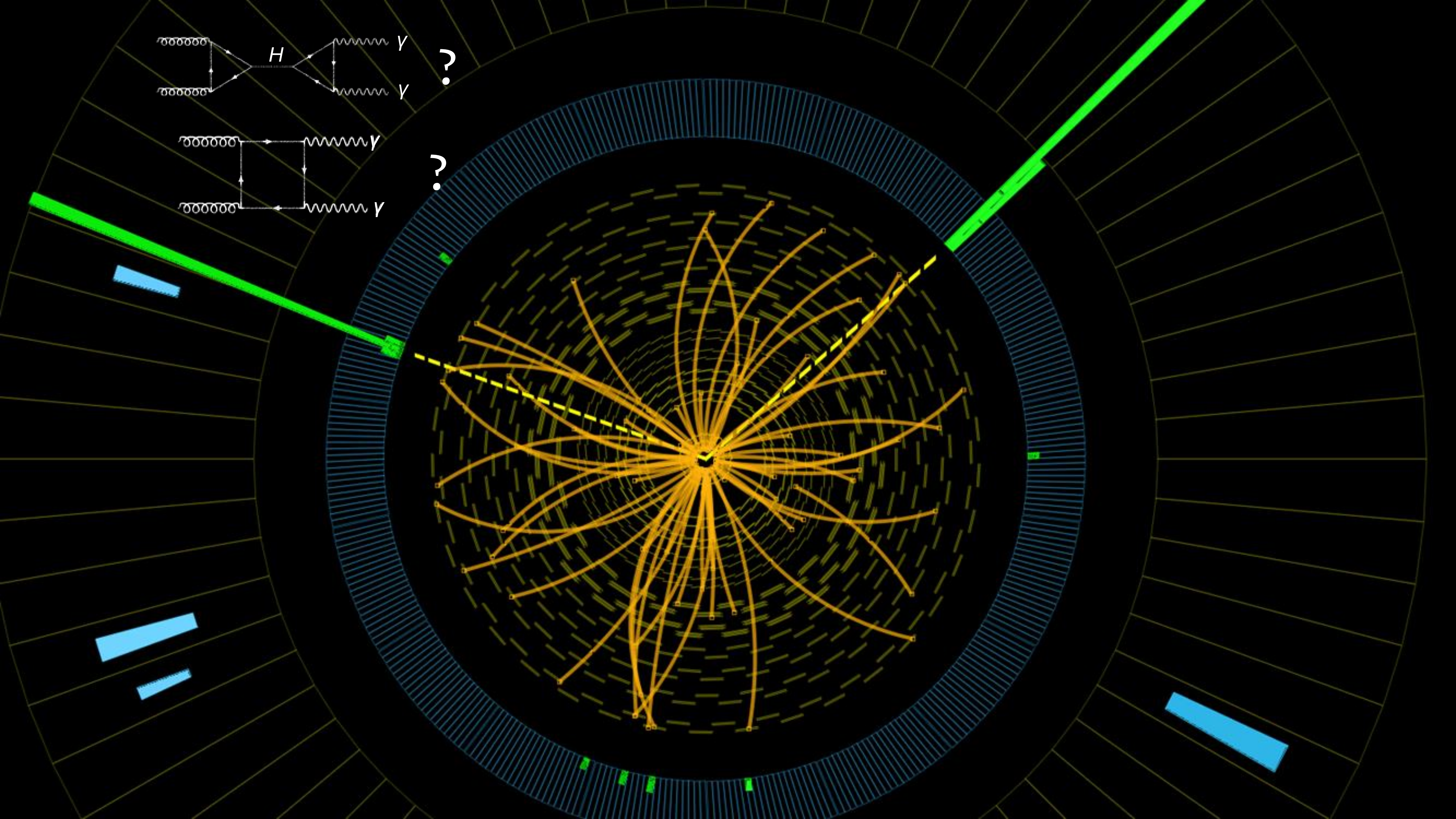
Calibrations



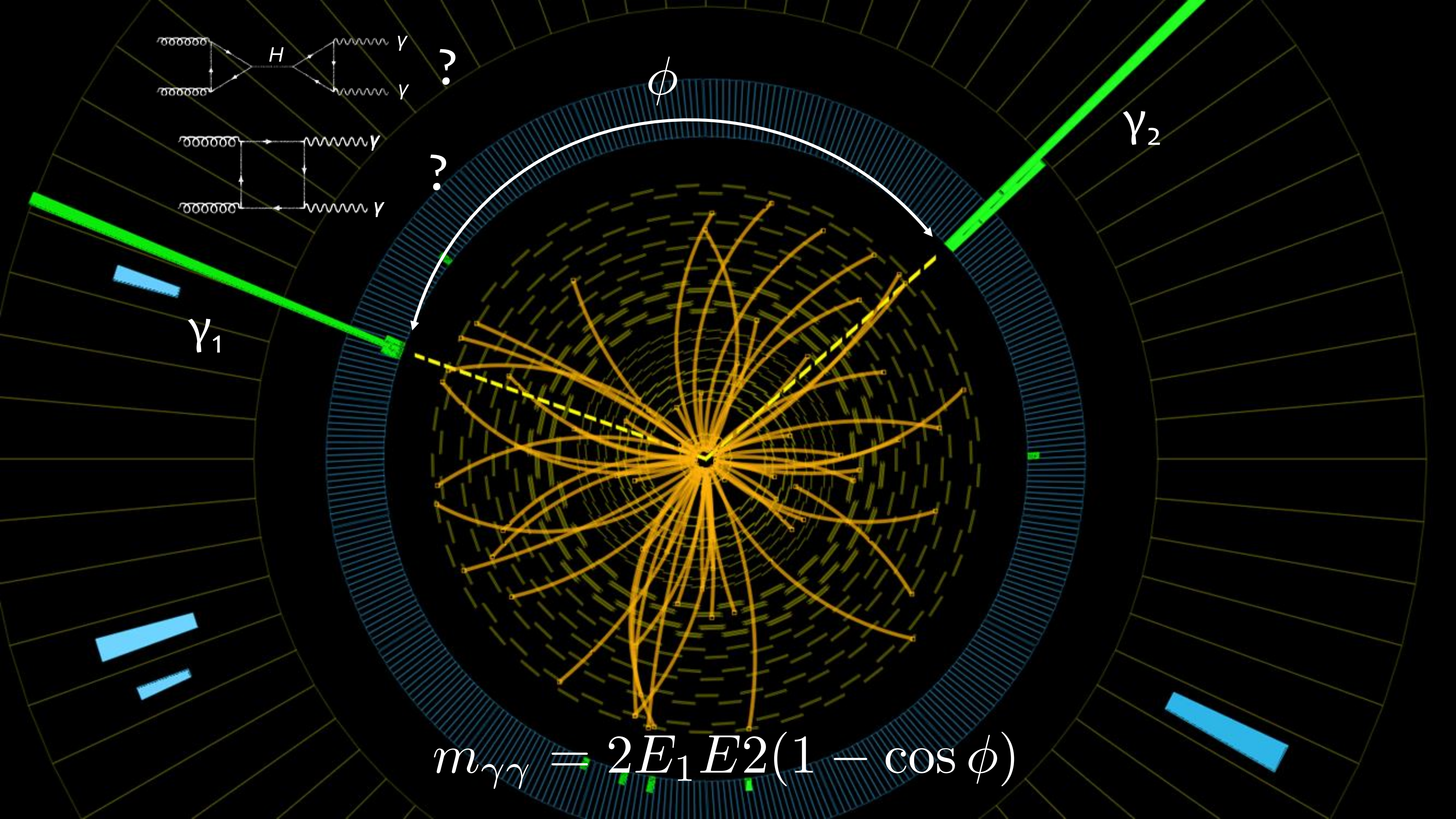
Event Selections & Distributions

RESULTS

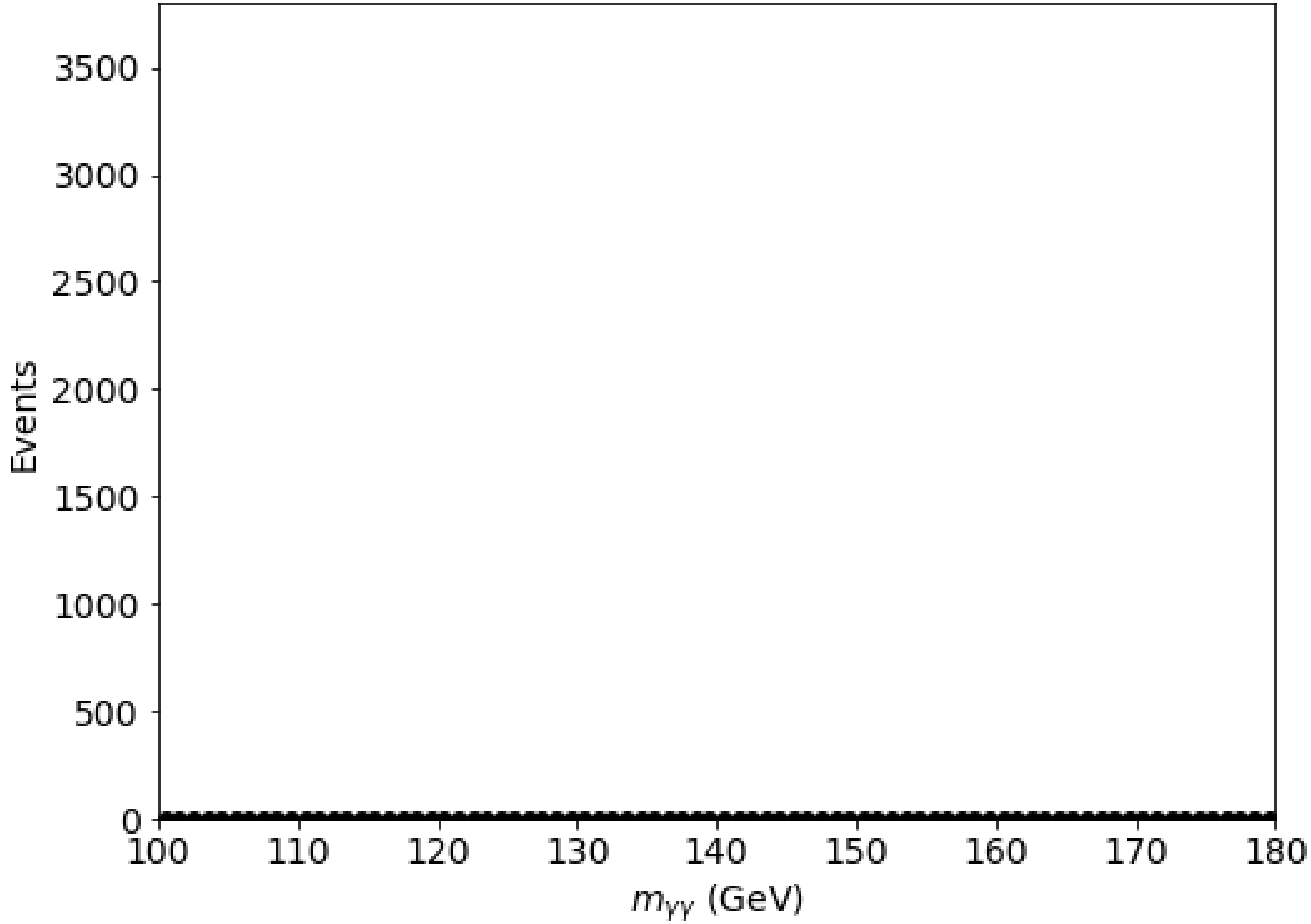




H
γ
γ
?
γ
γ
?


H
γ
γ
?
γ
γ
?
ϕ
γ_2
γ_1
$m_{\gamma\gamma} = 2E_1E2(1 - \cos\phi)$

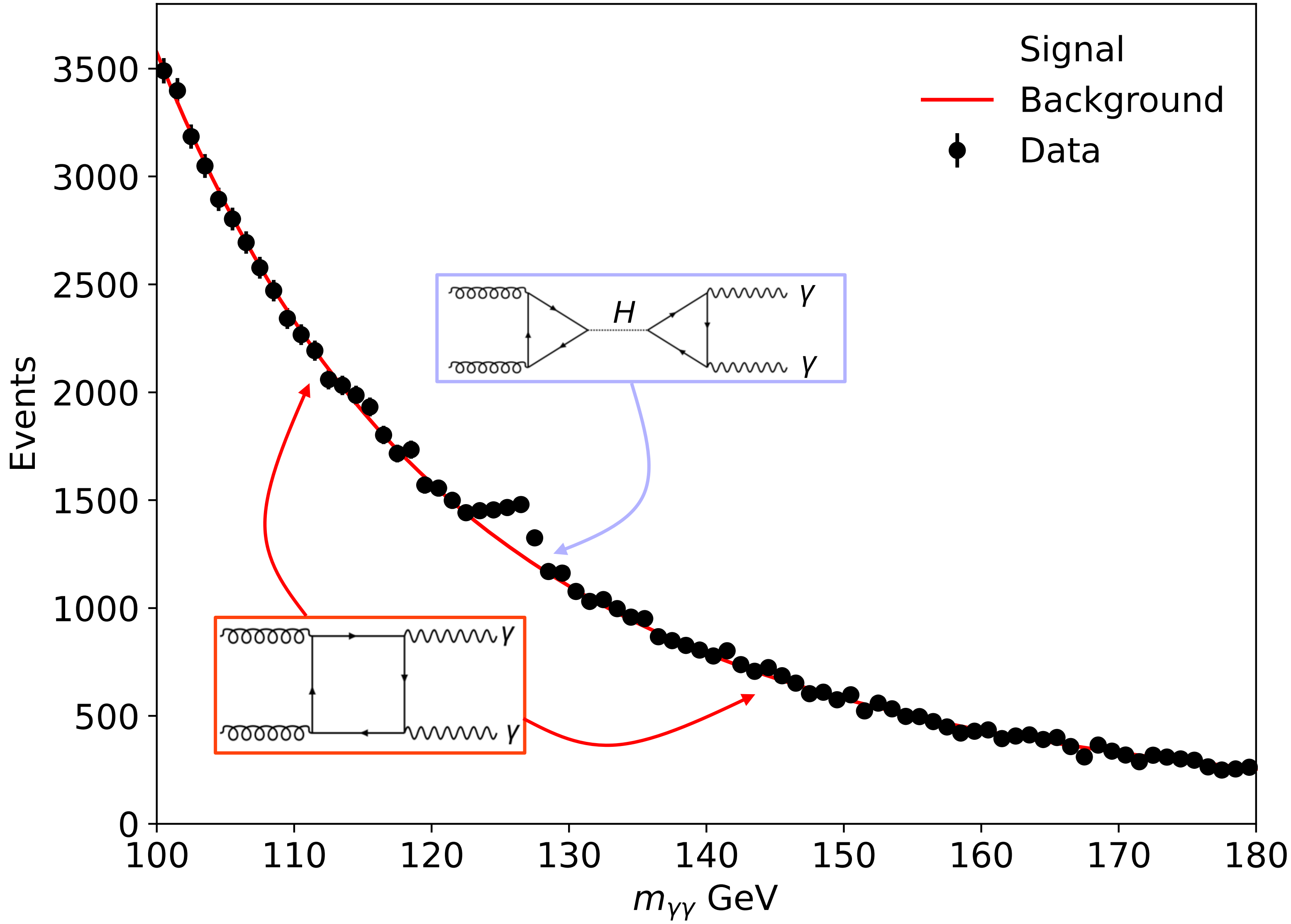
Collecting events



Each event that we select this way builds a picture of the underlying physics → if we're lucky, we might find something new

To extract the Physics, we use **distributions of observables** across many events

Collecting events



Each event that we select this way builds a picture of the underlying physics → if we're lucky, we might find something new

To extract the Physics, we use **distributions of observables** across many events

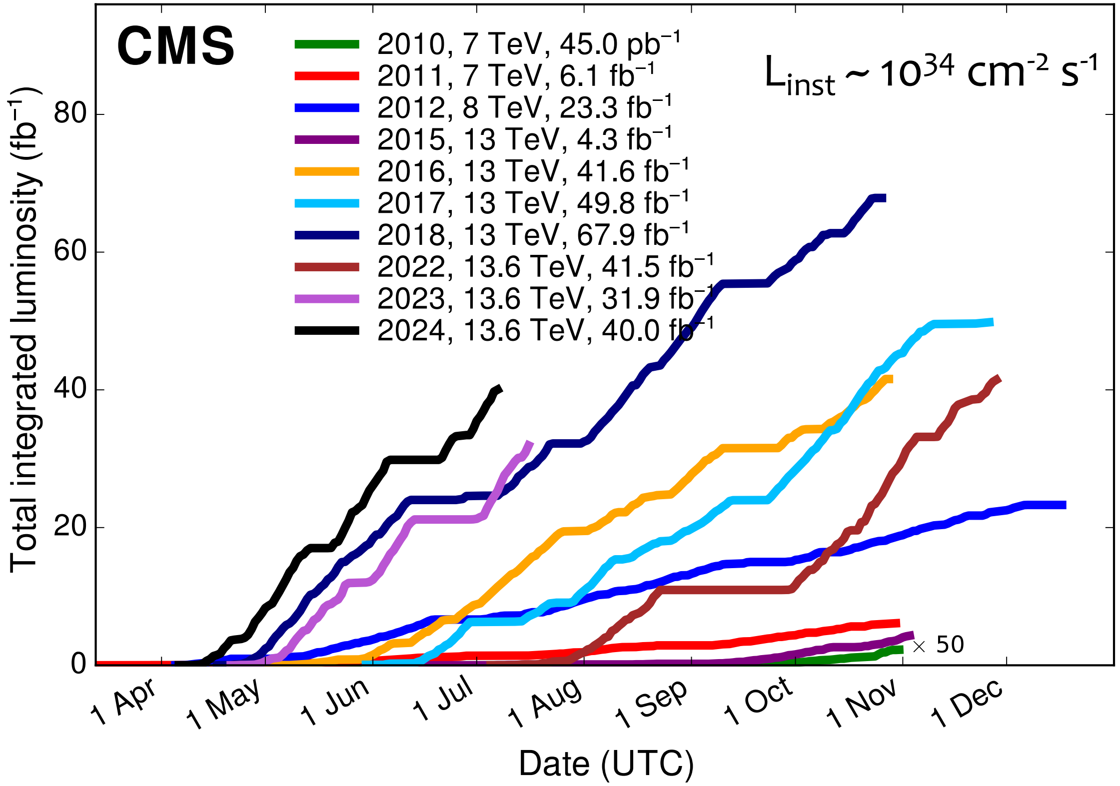
Collecting events

Collisions (i.e. bunch crossings) happen at **40 MHz**

Integrated luminosity (units of 1/area) $= \int c \cdot \frac{N_1 N_2}{4\pi\delta_{x,y}} \cdot f n_b dt$

$$N = L\sigma$$

Number of events



Collecting events

Collisions (i.e. bunch crossings) happen at **40 MHz**

Integrated luminosity (units of 1/area) $= \int c \cdot \frac{N_1 N_2}{4\pi\delta_{x,y}} \cdot f n_b dt$

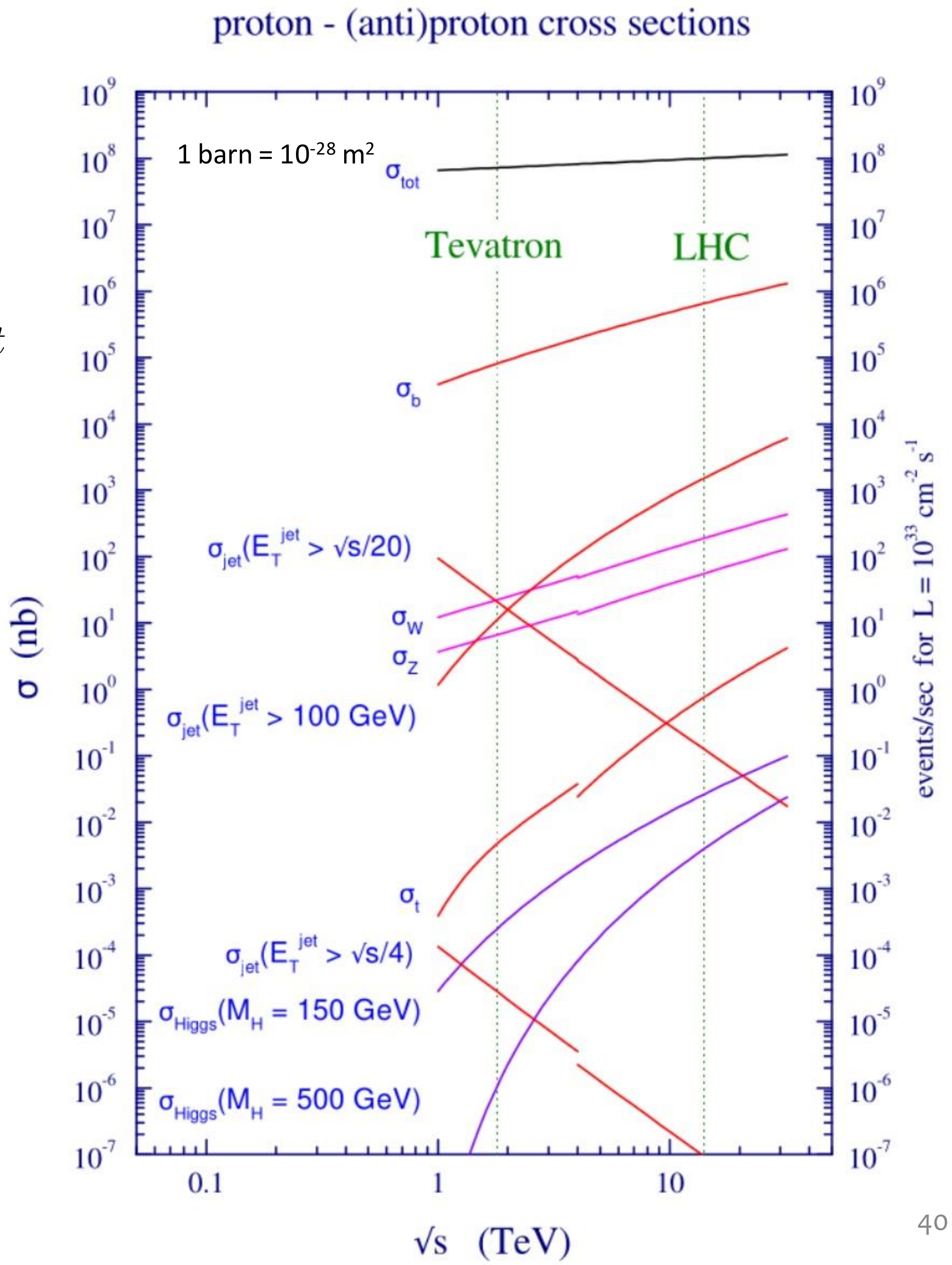
$$N = L\sigma$$

Number of events

Cross-section (units of area)

Example: For every **1,000,000,000** inelastic proton-proton collisions, **only expect one of them to produce a Higgs boson**!

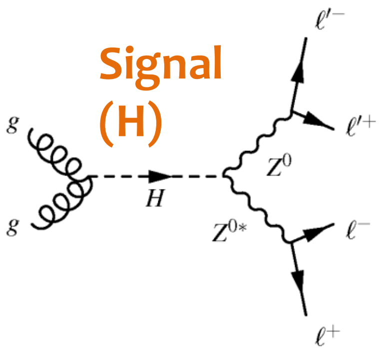
We typically have to select events based on these observables to dig out the signal from the background (noise!)

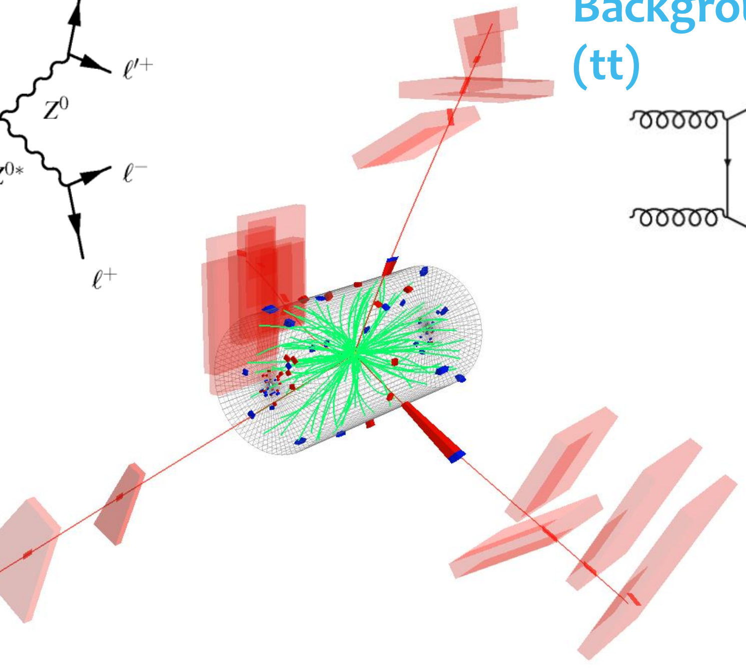


Event Selection

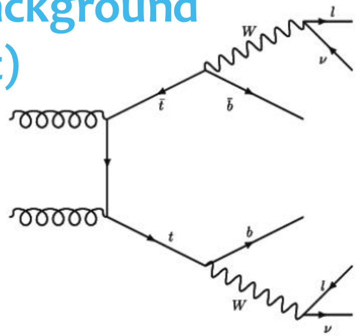
By knowing ahead of time the kind of events we are interested in, we impose selections on the events to reduce the background as much as possible while maintaining the signal

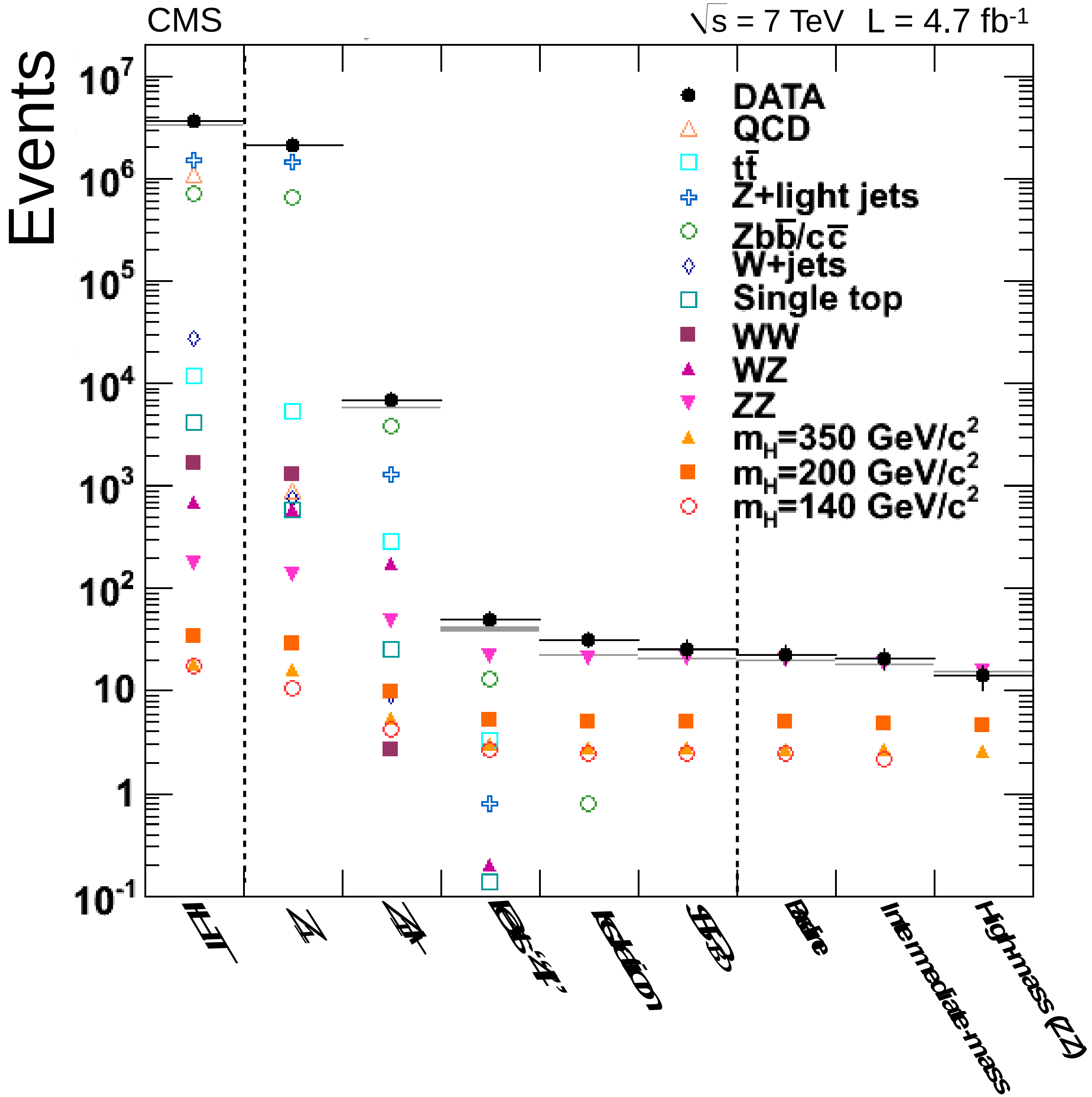
Signal (H)





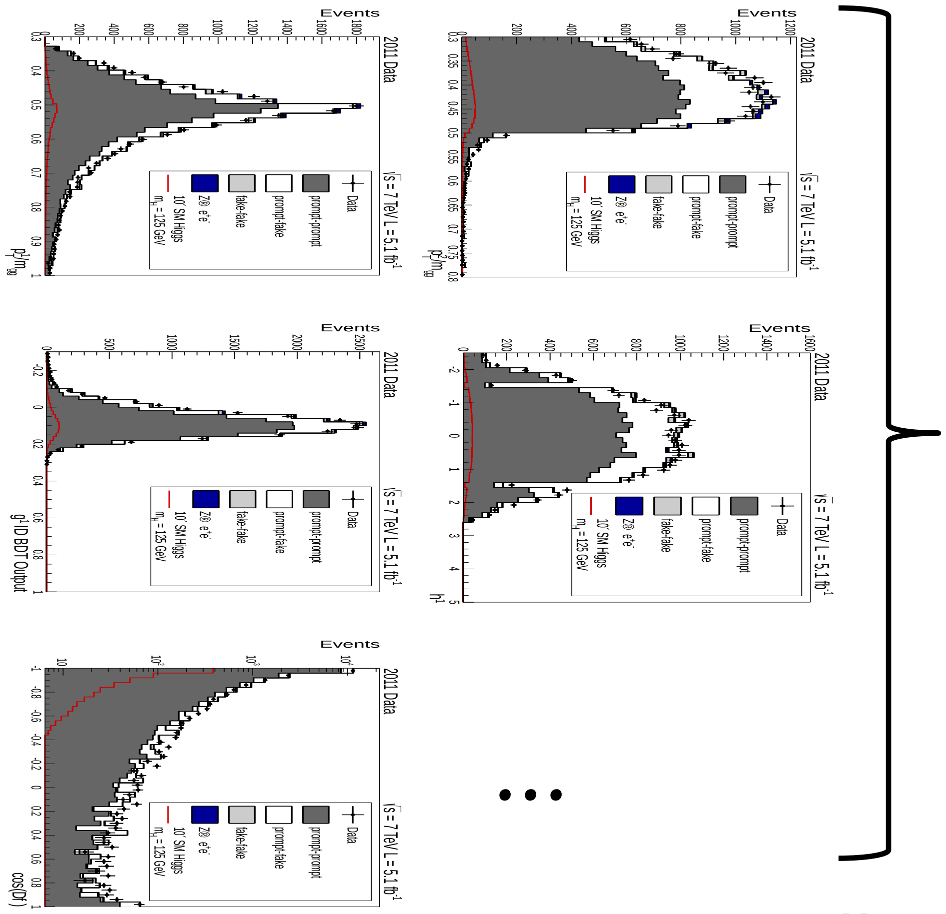
Background (tt)

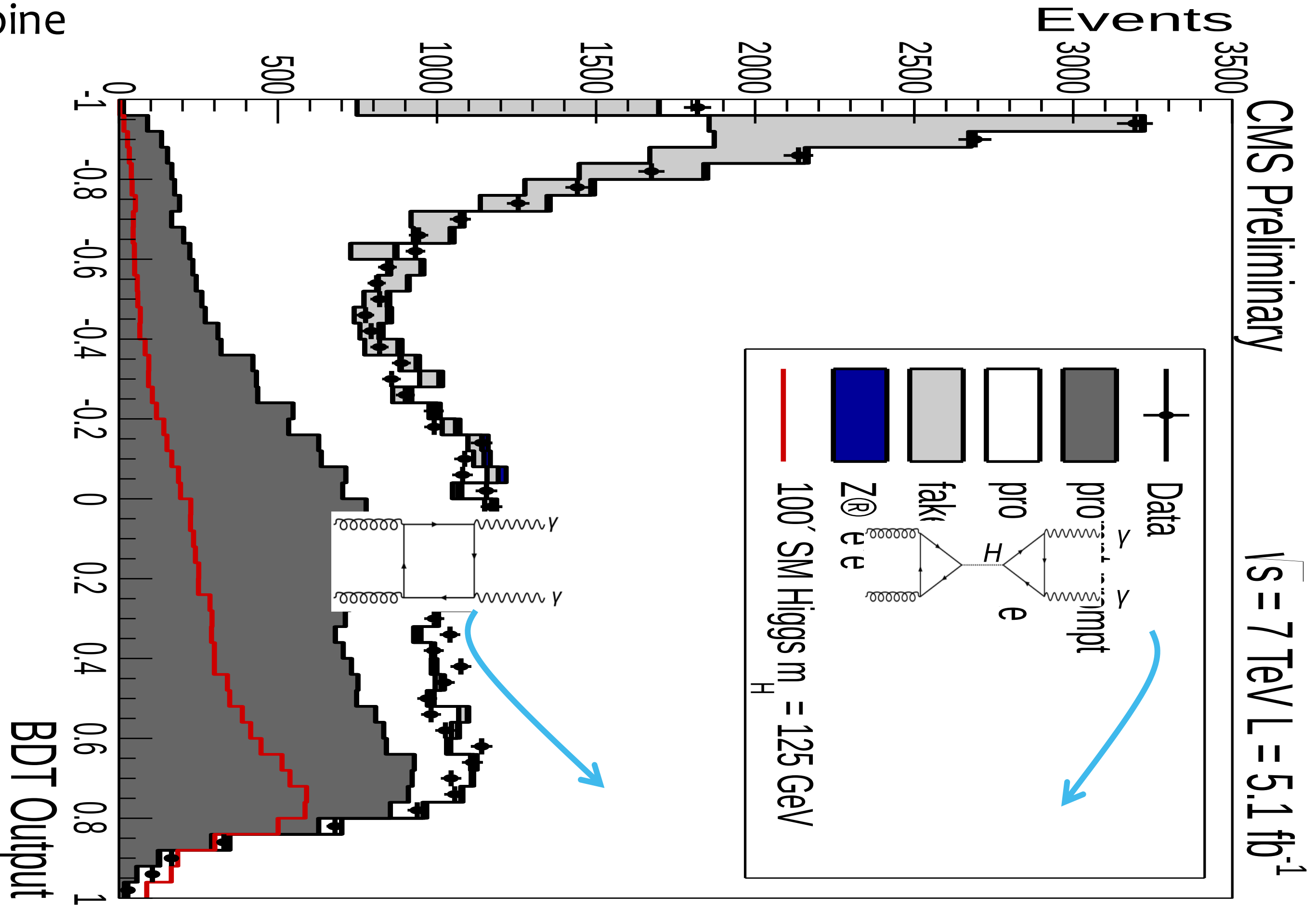




Event Selection

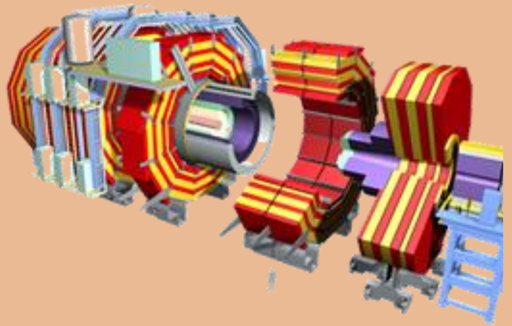
We often use **Machine Learning techniques** to combine as much information as possible for this task





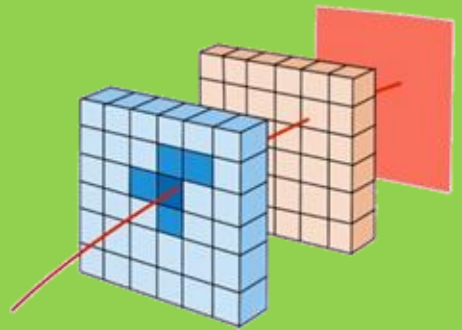
How can we choose these selections/train our Machine Learning models?

Data

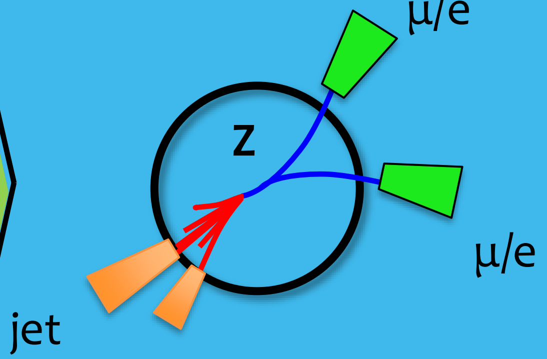


Simulation

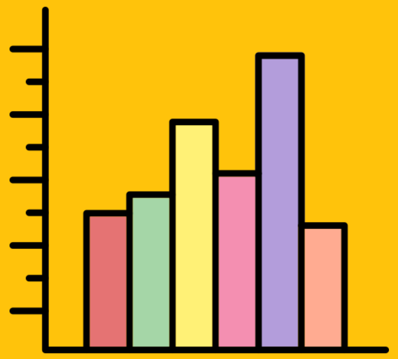
Reconstruction & Particle ID



Calibrations



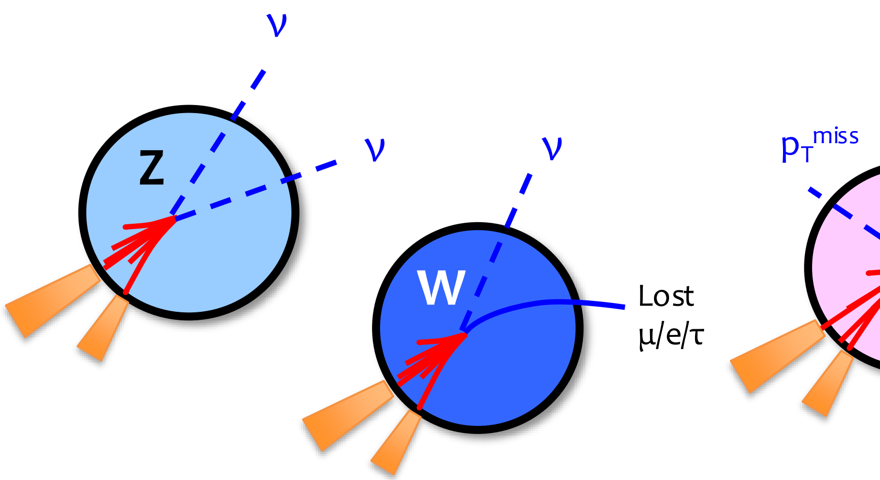
Event Selections & Distributions



RESULTS

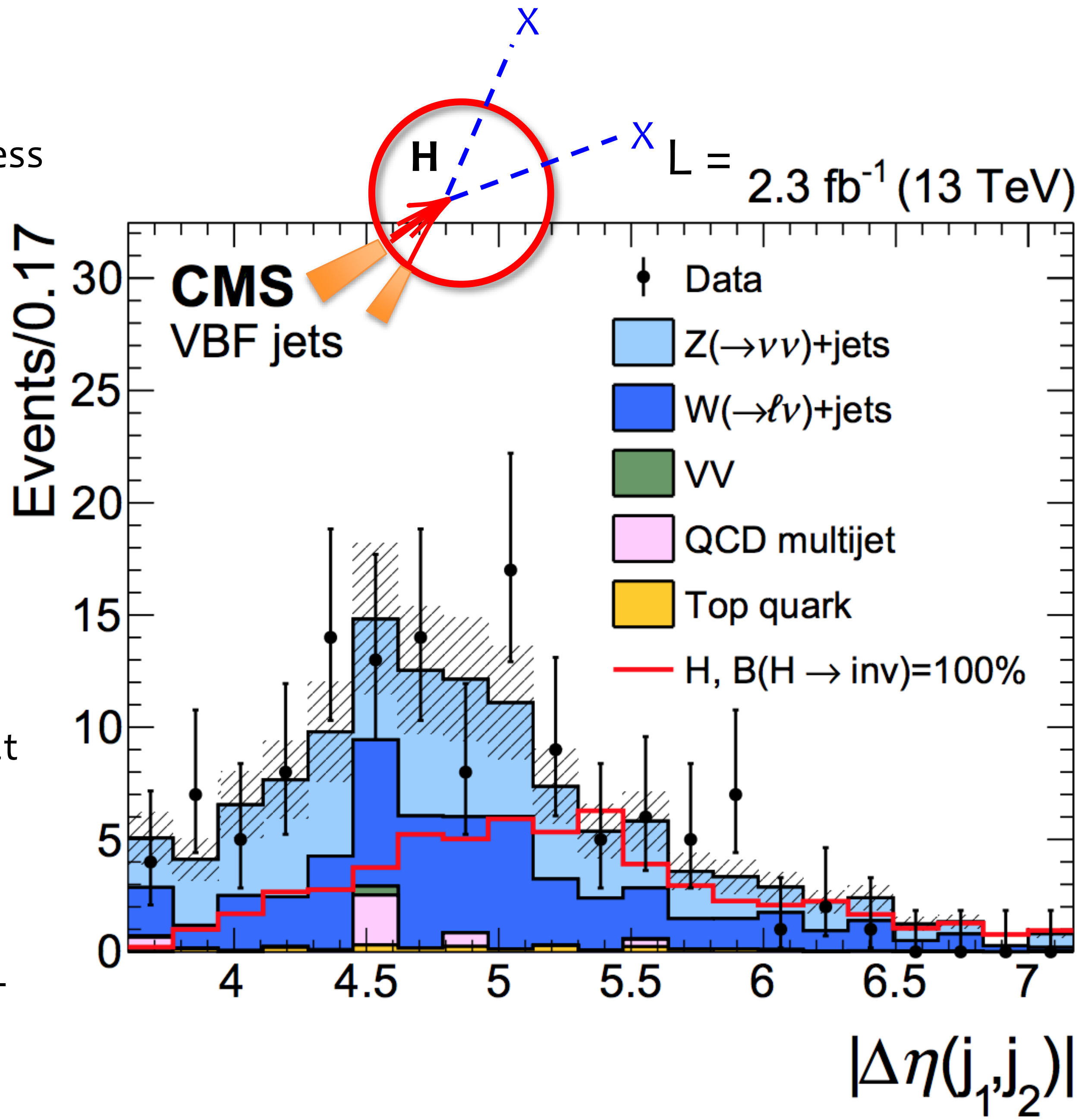
Simulation

Generate large number of simulated events for each process contributing to our analysis (**signals** and backgrounds)

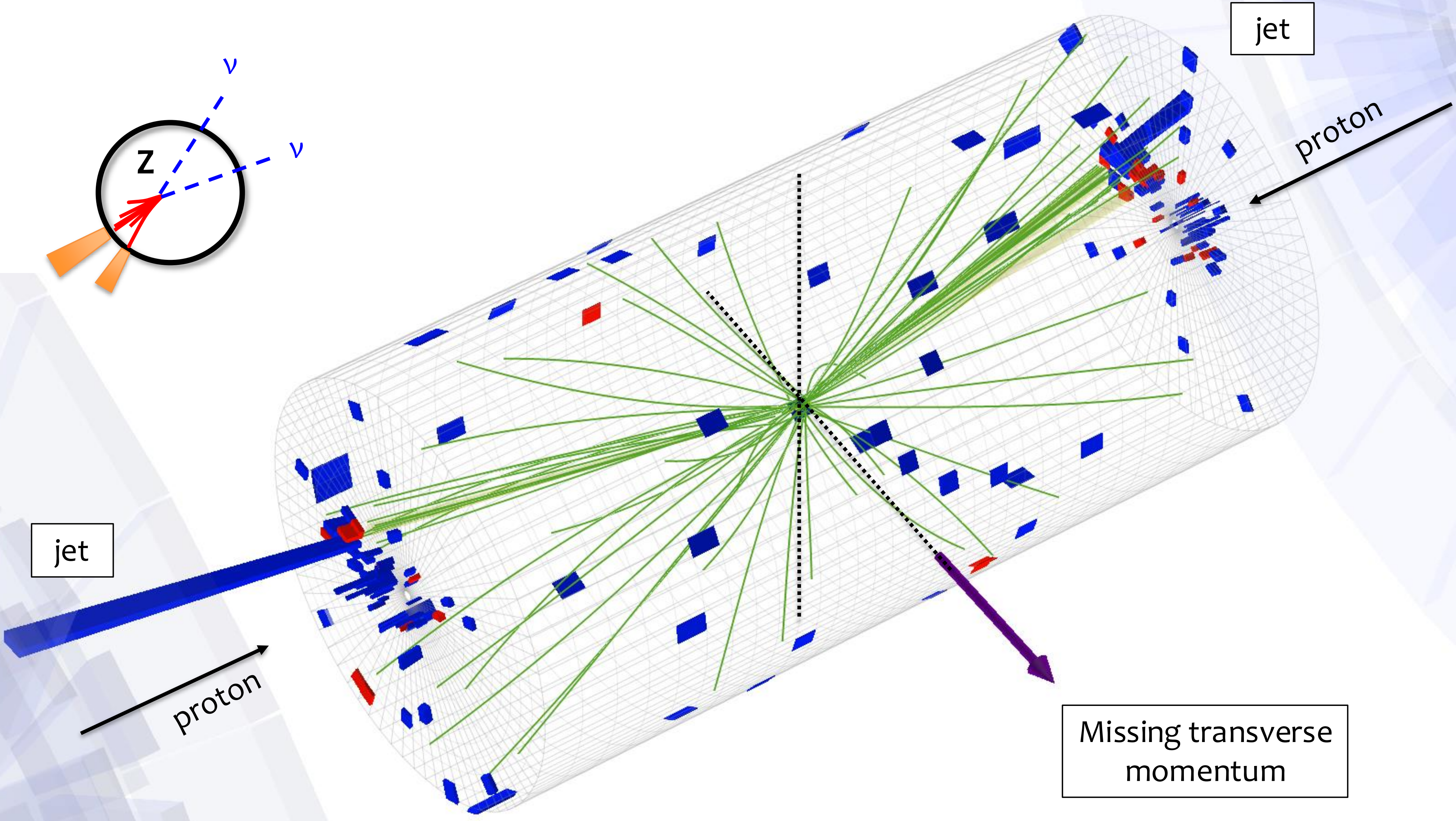


Simulated events must be ***weighted*** to the get the correct predicted yield for a given dataset

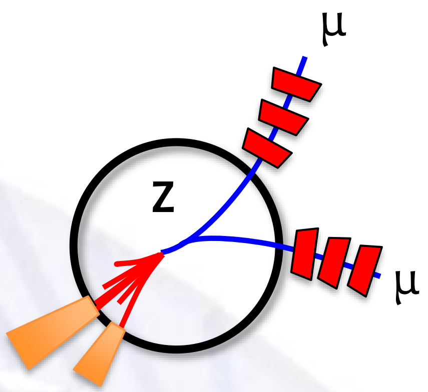
$$L_{\mathrm{eff}} = \frac{N_{\mathrm{gen}}}{\sigma} \Rightarrow \mathrm{weight} = \frac{L}{L_{\mathrm{eff}}}$$

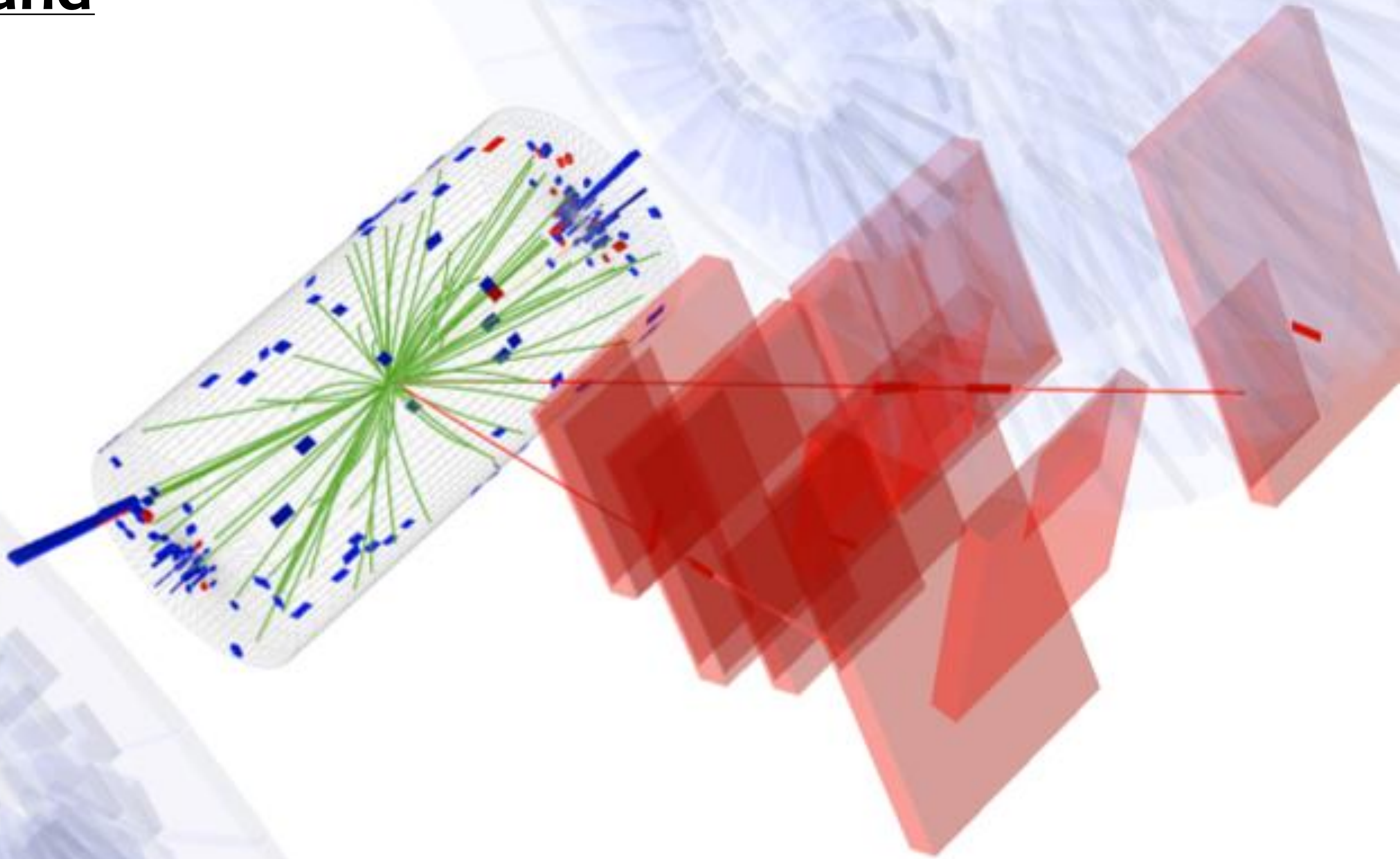


Data-driven background



Data-driven background

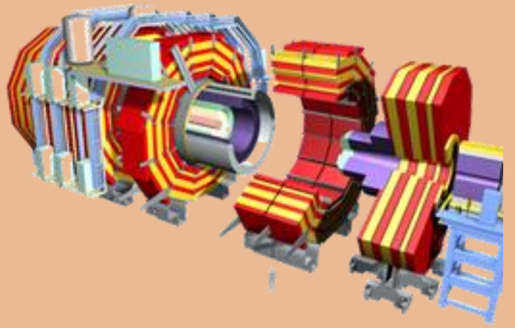




Estimate the normalization of the Z→neutrinos background using data!

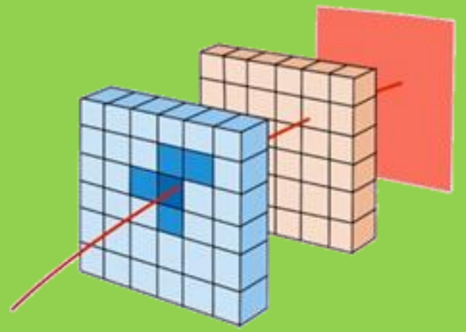
$$N_{Z(\to\nu\nu)} \approx N_{Z(\to\mu\mu)} \frac{B(Z\to\nu\nu)}{B(Z\to\mu\mu)} A(\mu)\epsilon(\mu)$$

Data

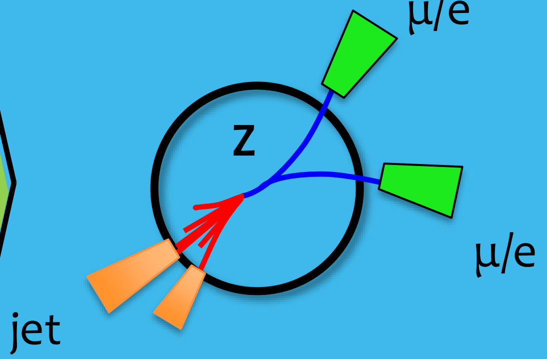


Simulation

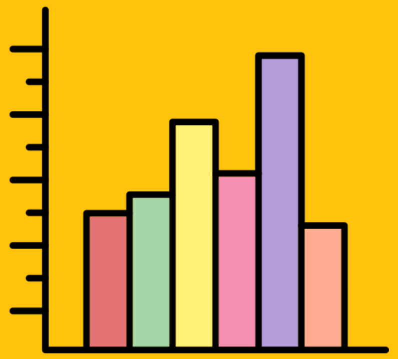
Reconstruction & Particle ID



Calibrations

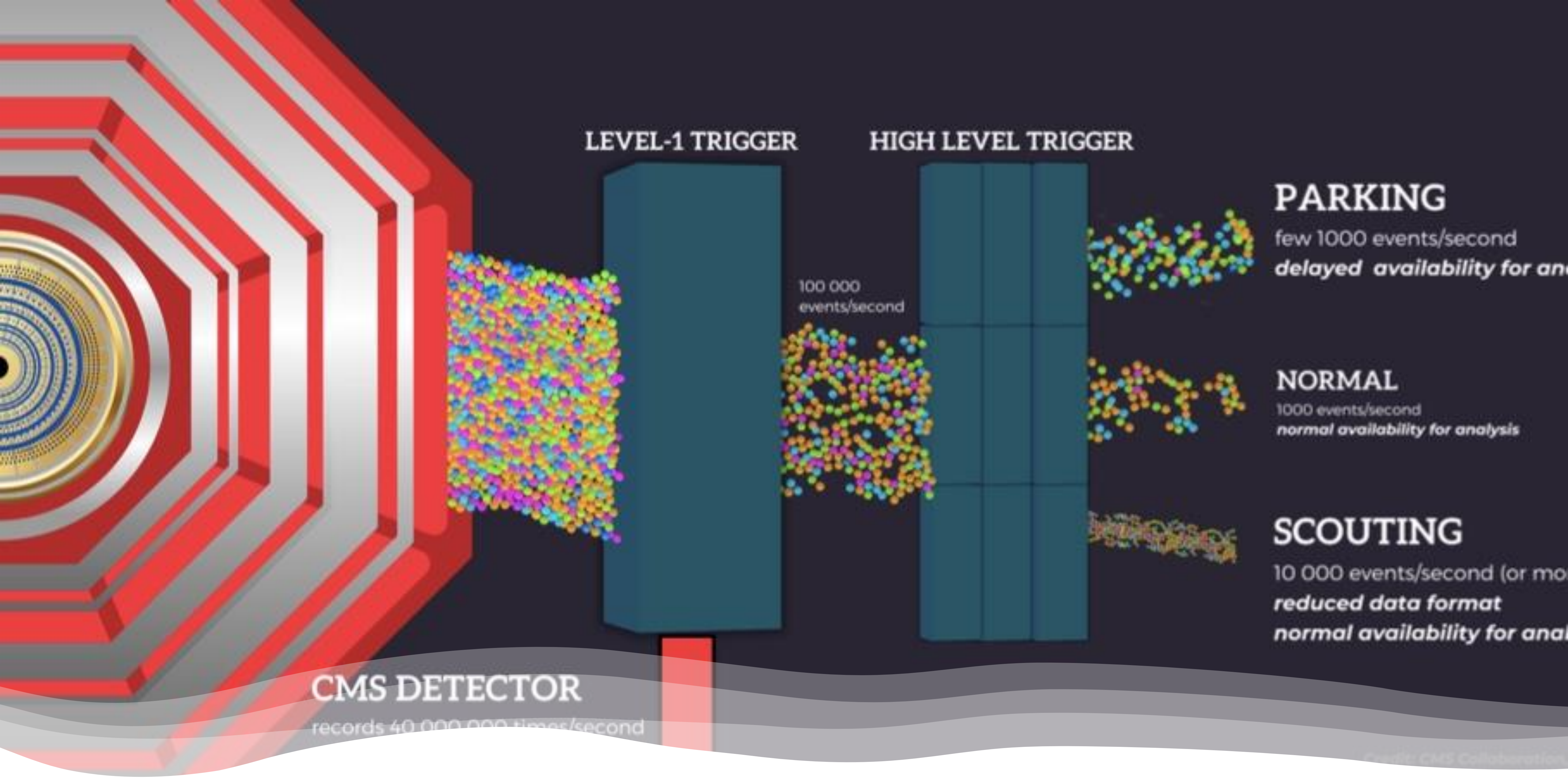


Event Selections & Distributions



RESULTS

Huge computing power required to acquire and analyse LHC data



Online Selection of events "Trigger" determines which events to keep in around 4 micro-seconds!

Huge collaborations of people required for Data Analysis at the LHC

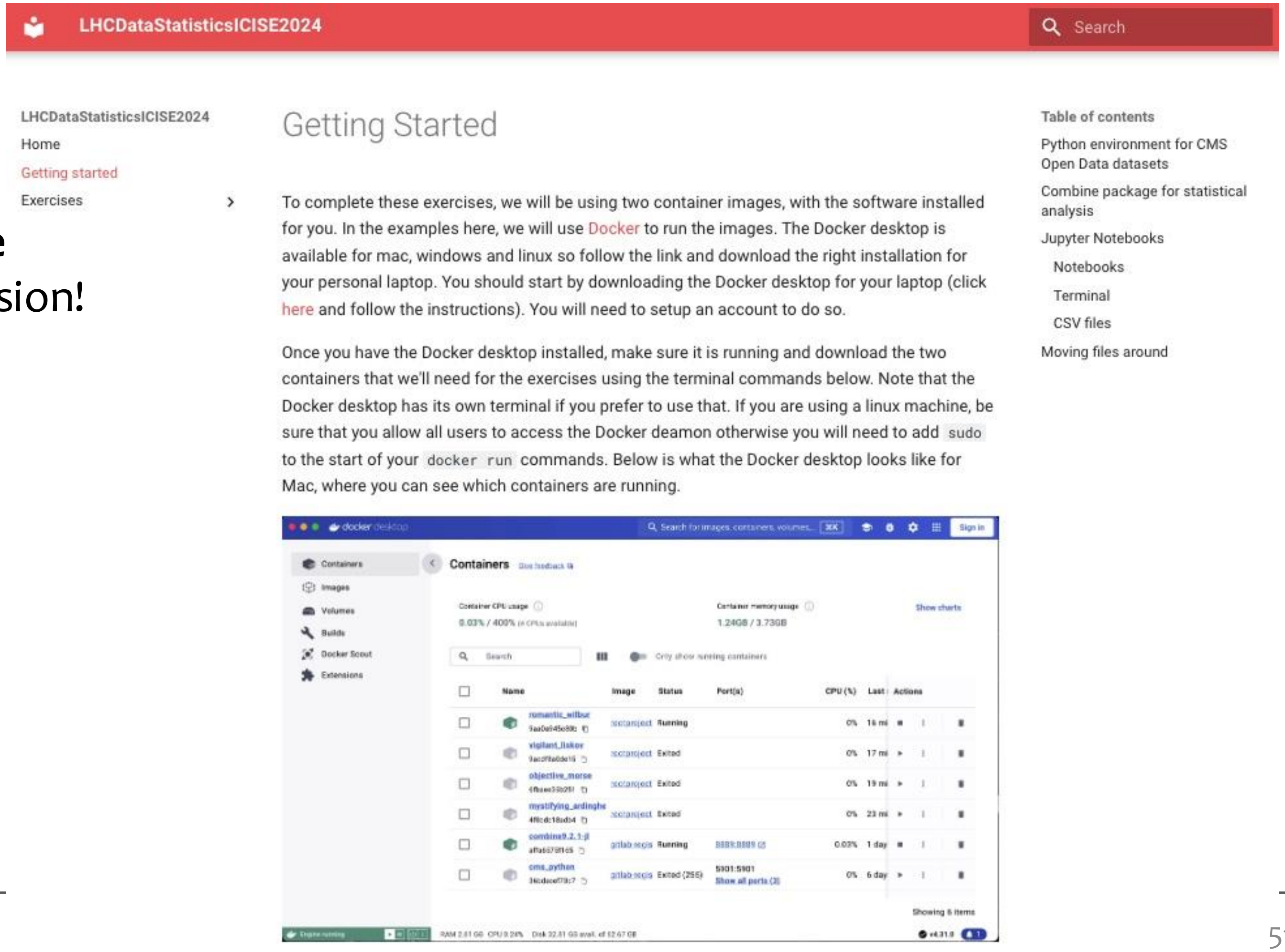
Now it’s your turn!

This afternoon, we are going to have a go at doing a data analysis with some real CMS proton-proton collision data!

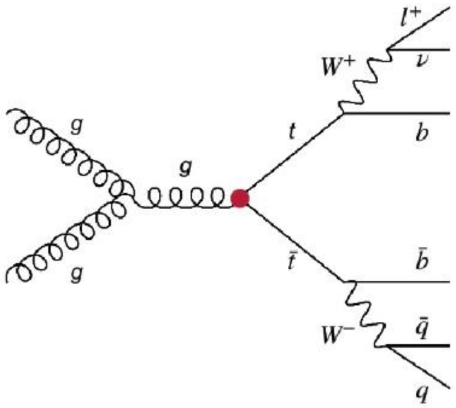
All of the instructions for getting setup and the exercises are available here:
https://nucleosynthesis.github.io/LHCDataStatisticsICISE2024/

You will also see links to the lecture slides (**password VSOPLHC2024**)

If you haven’t already done so, **please go through the “Getting started” section** before this afternoon’s session!



This afternoon, we will be working through **Exercise 1**



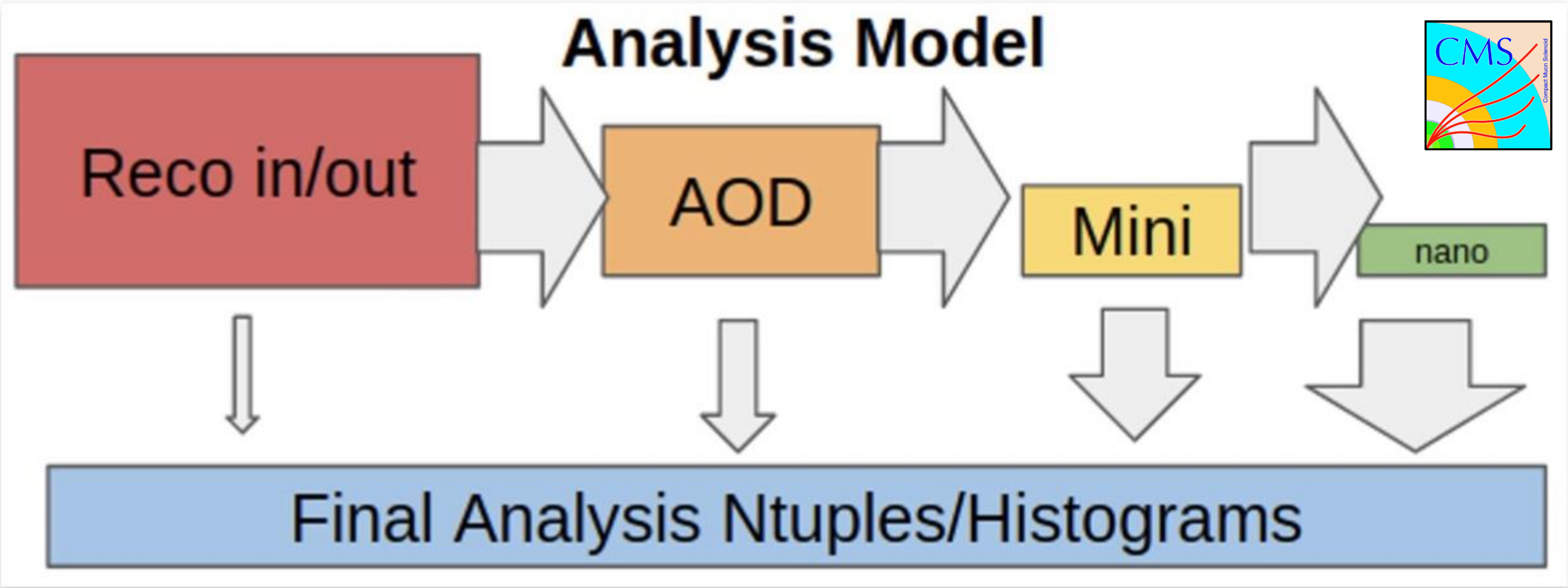
(Extra Slide) Data Tiers

RAW data output from experiments is too large for direct analysis

e.g 2018 data from CMS O(10) PB at RAW data level

→ Reduce content through processing at different data tiers to make analysis manageable

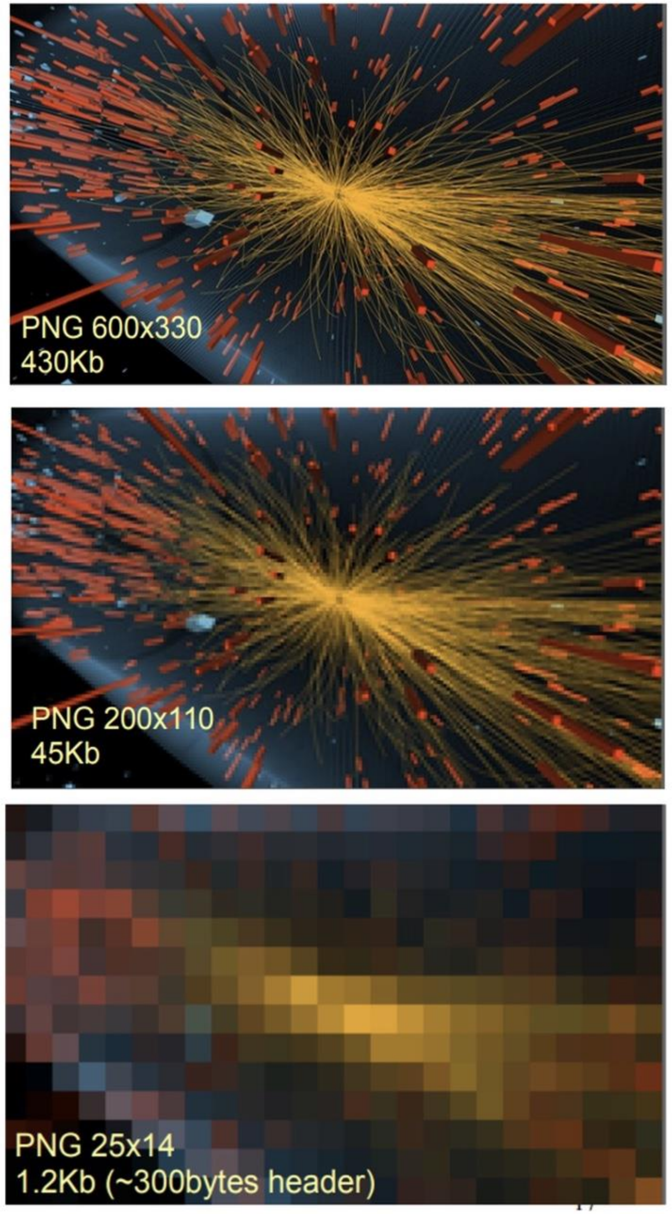
→ Less information but content is closer to final analysis objects (hits → particles)



AOD 450kb/ev

MiniAOD 50kb/ev

NanoAOD 1-2kb/ev



(Extra Slide) A Real CMS analysis selection flow