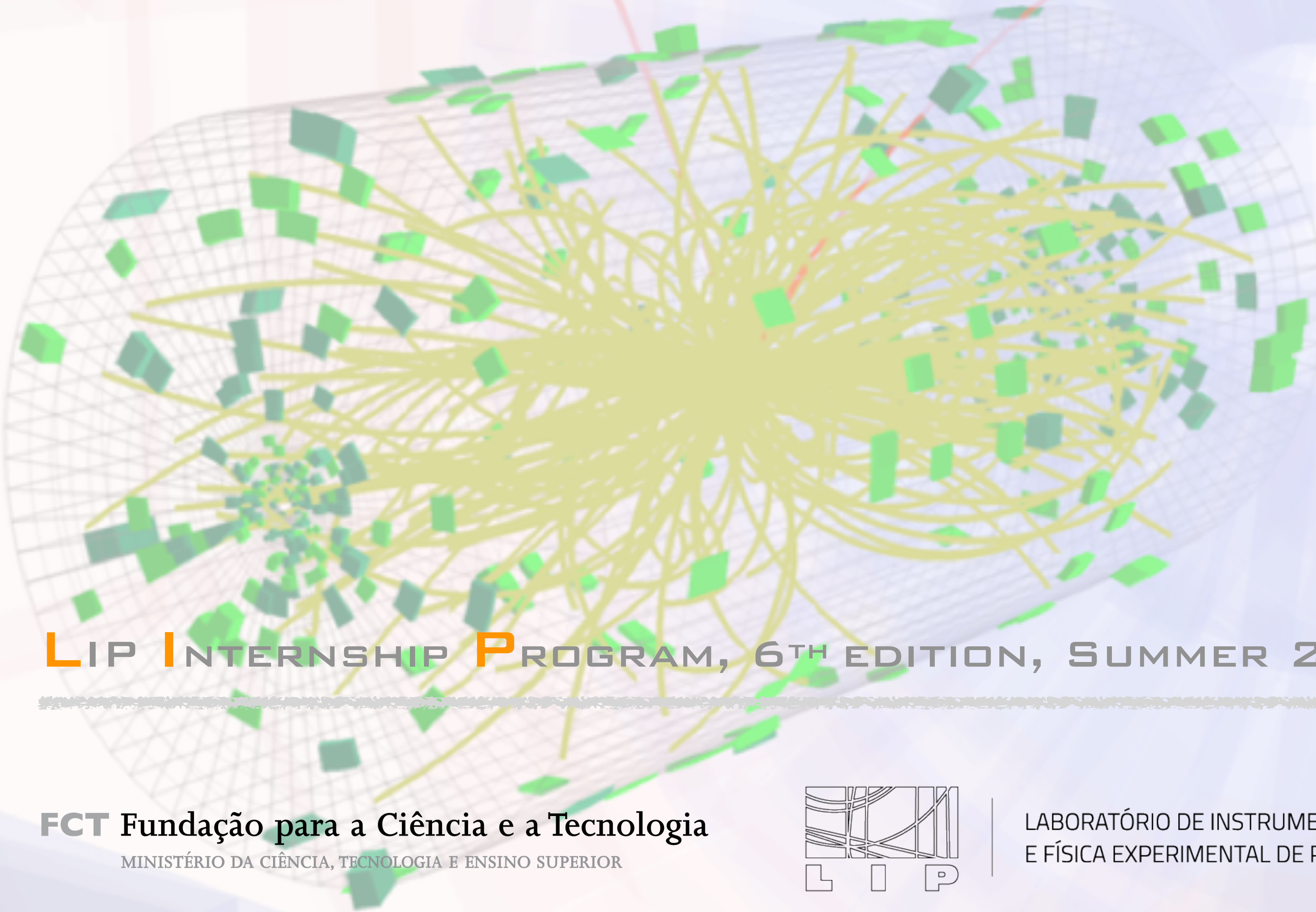
Tutorial on

Data Analysis



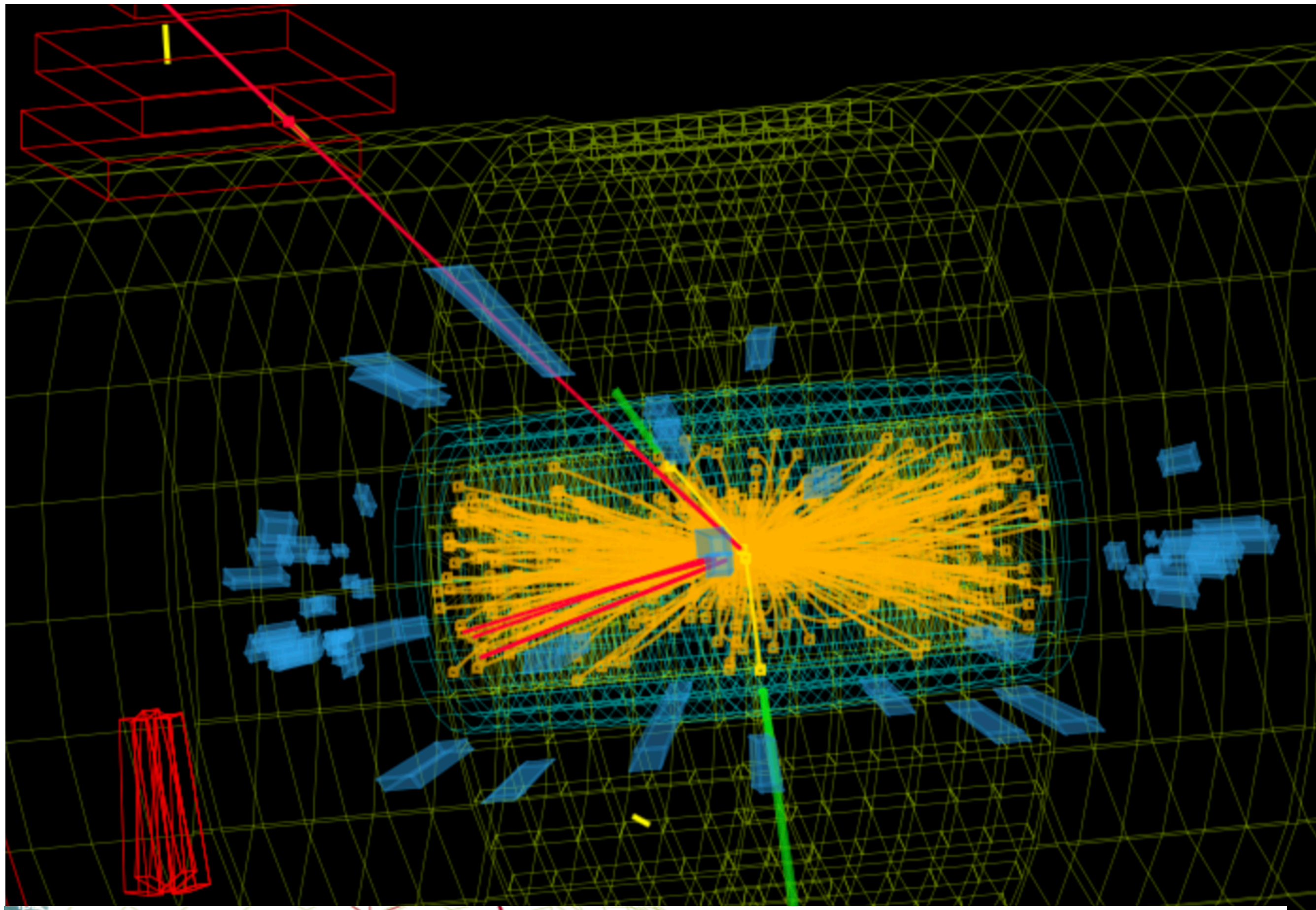
LIP Internship Program, 6th edition, Summer 2022

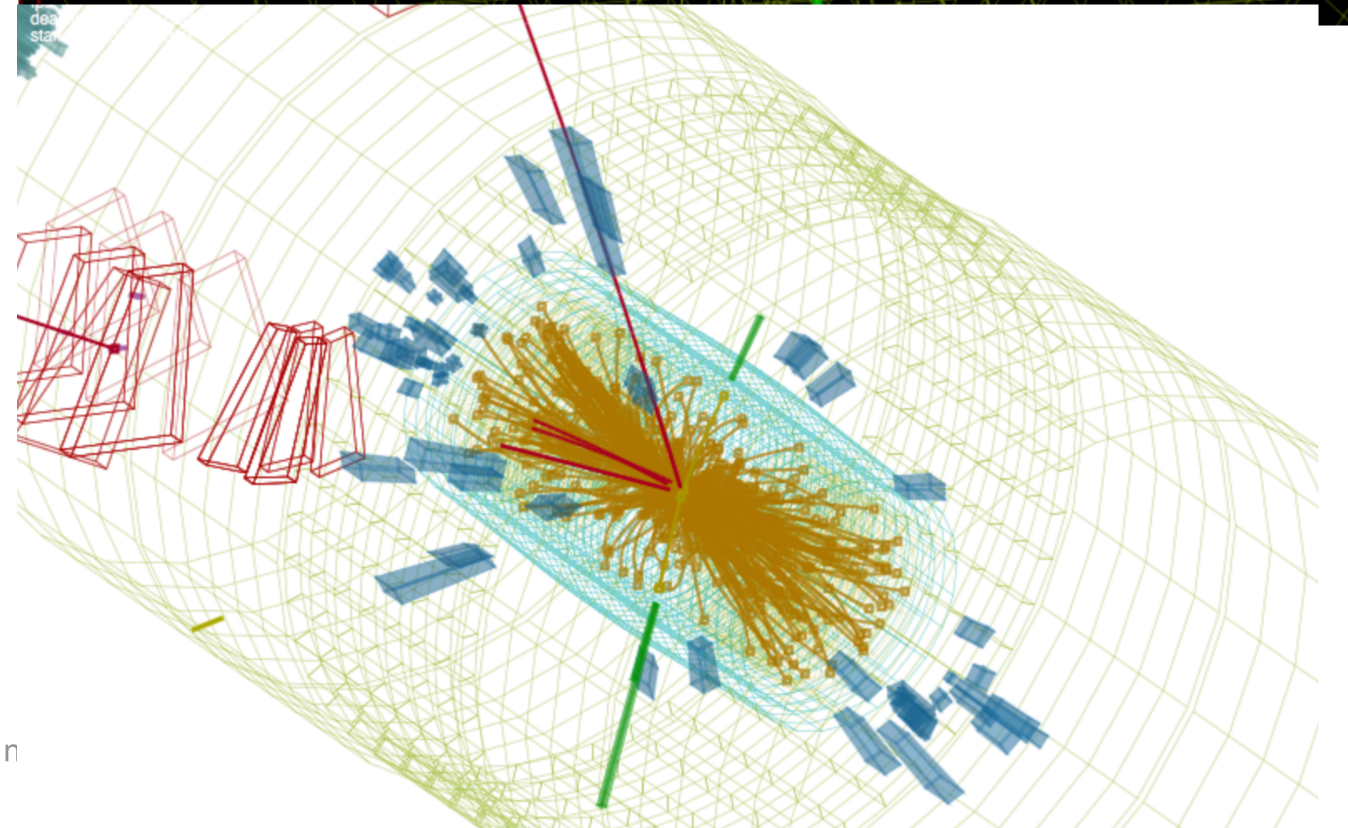
FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

LIP
LABORATÓRIO DE INSTRUMENTAÇÃO
E FÍSICA EXPERIMENTAL DE PARTÍCULAS

LHC Open Data

- the LHC collaborations make good chunks of their data publicly available
  - http://**opendata.cern**.ch/
- along with tools & software & examples
- for data visualisation and analysis
- from event reconstruction algorithms to machine learning challenges
- via virtual machines (with no need to install different software packages)
- few pointers
  - http://opendata.cern.ch/visualise/events/cms
  - http://www.i2u2.org/elab/cms/event-display/
- you're invited to **explore the LHC data** also on your own leisure



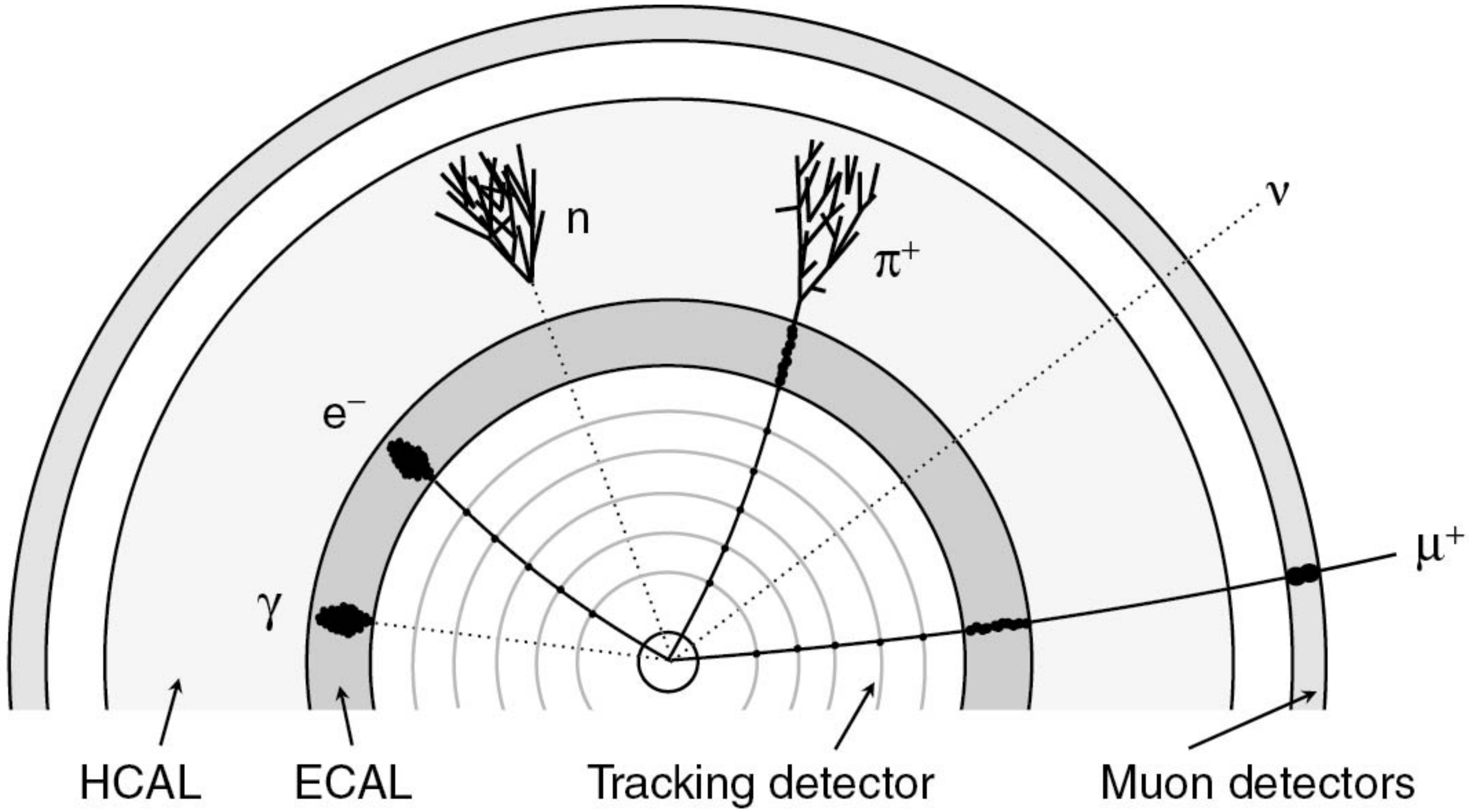


goals

perform a simple data analysis

- visualise the data
- manipulate data ntuples
- produce, process, and display data histograms
  - select different physics signals
  - plot kinematic distributions, inspect detector/trigger effects
- extract physics parameters from data
  - measure signal yields by performing a likelihood fit
  - inspect statistical and systematic errors
  - (extra) perform a differential measurement

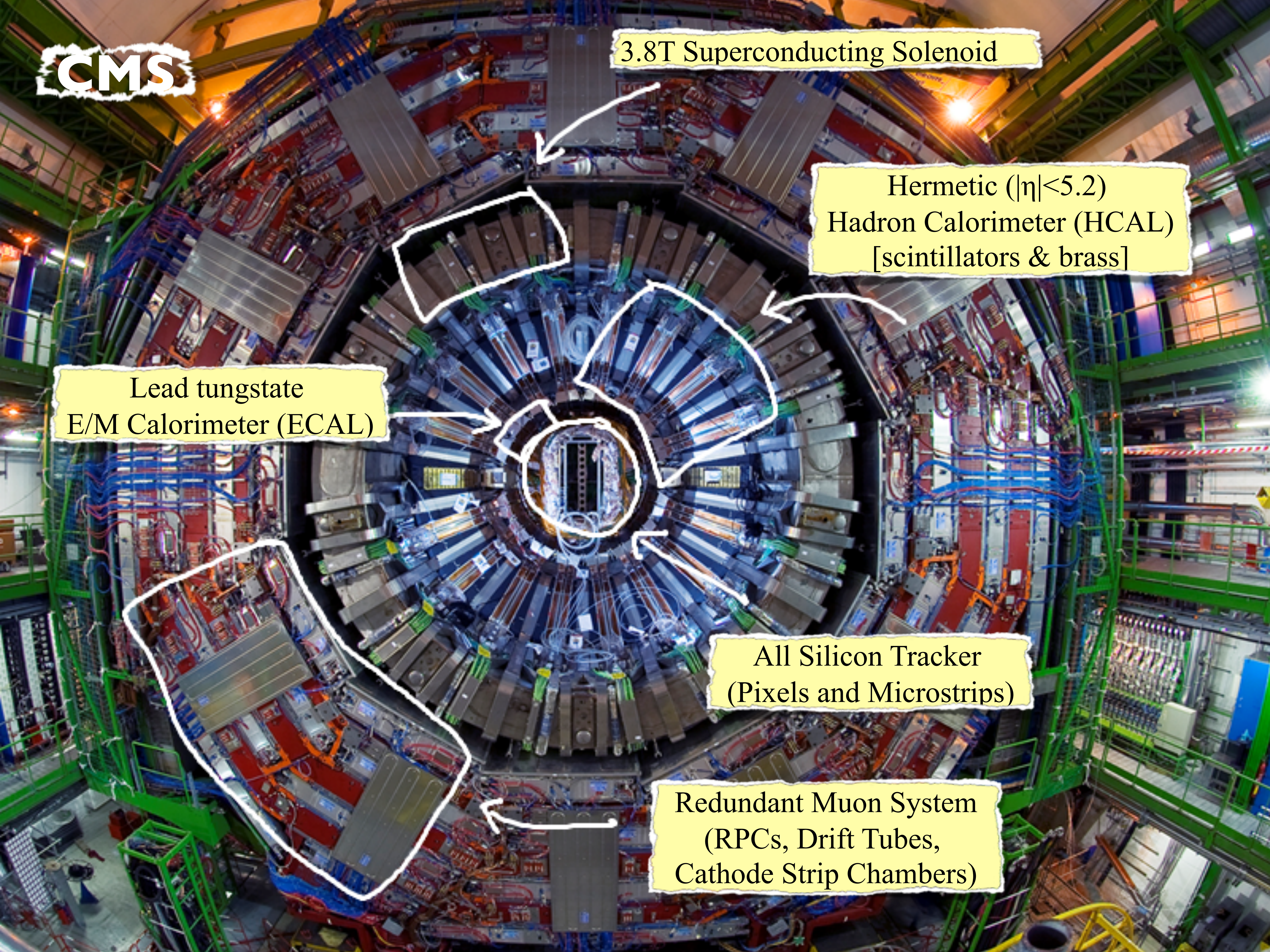
Detector & Event Reconstruction & Visualisation

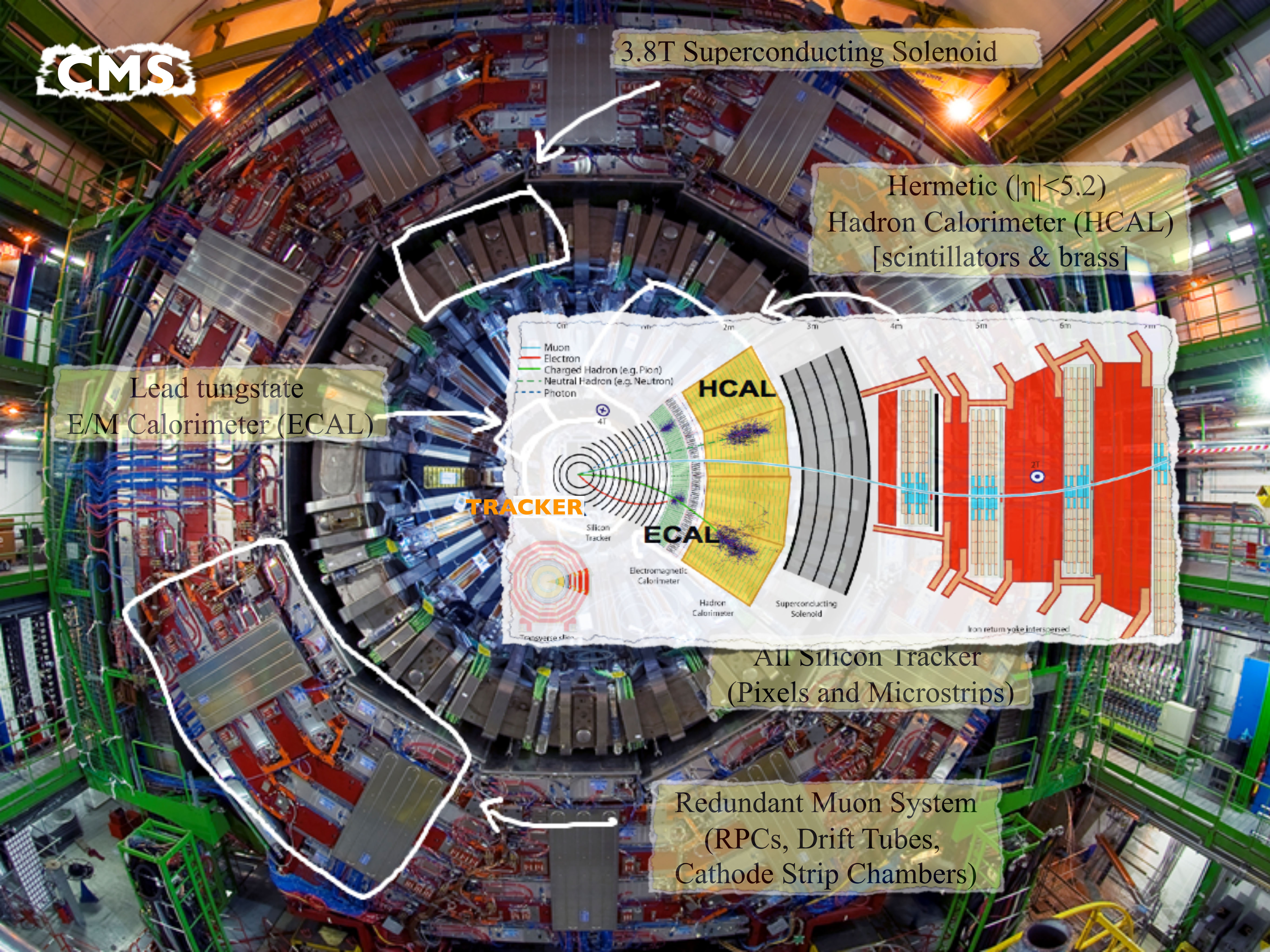


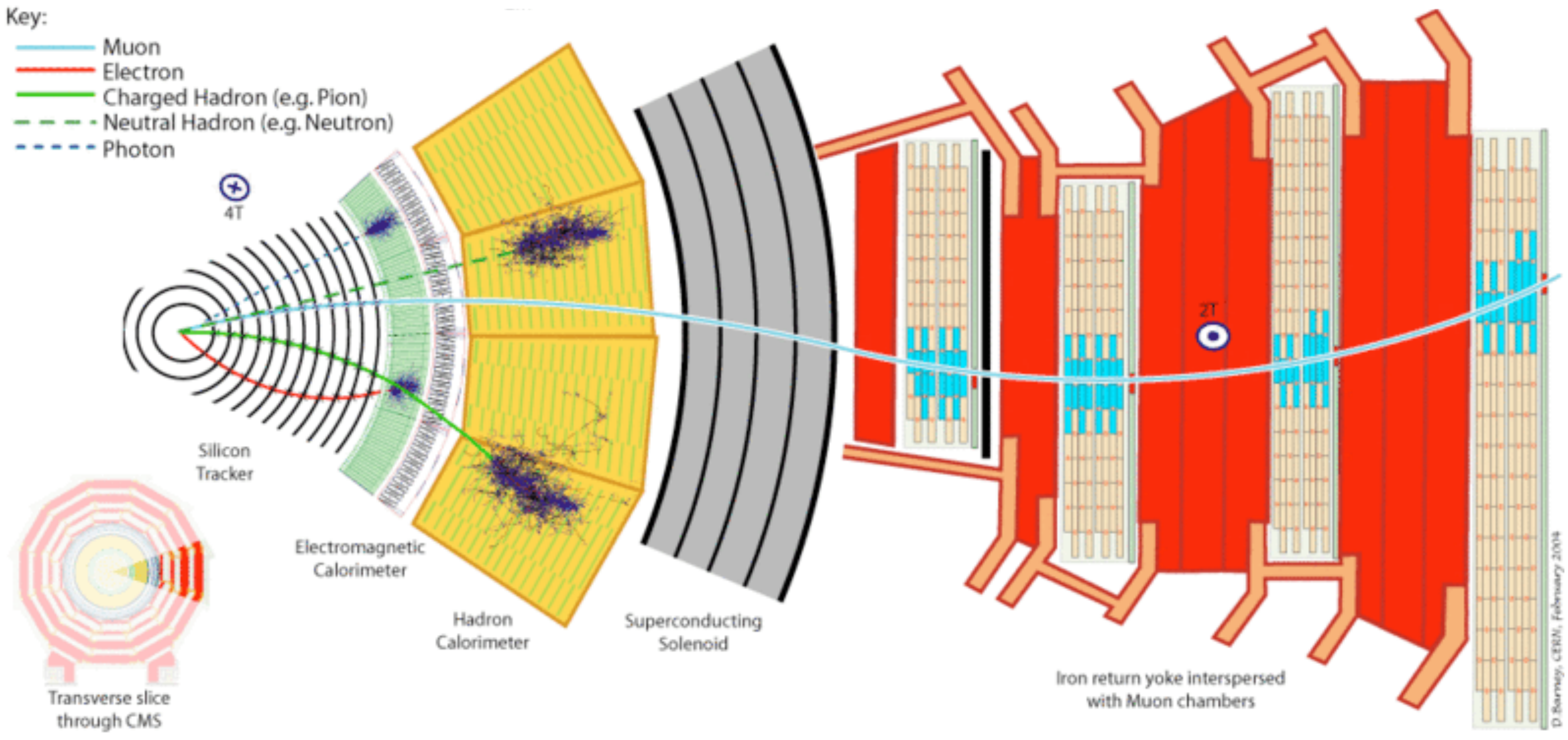
calorimeters:
measure particle's
energy by absorbing it

trackers:
detect trajectory
of charged particles

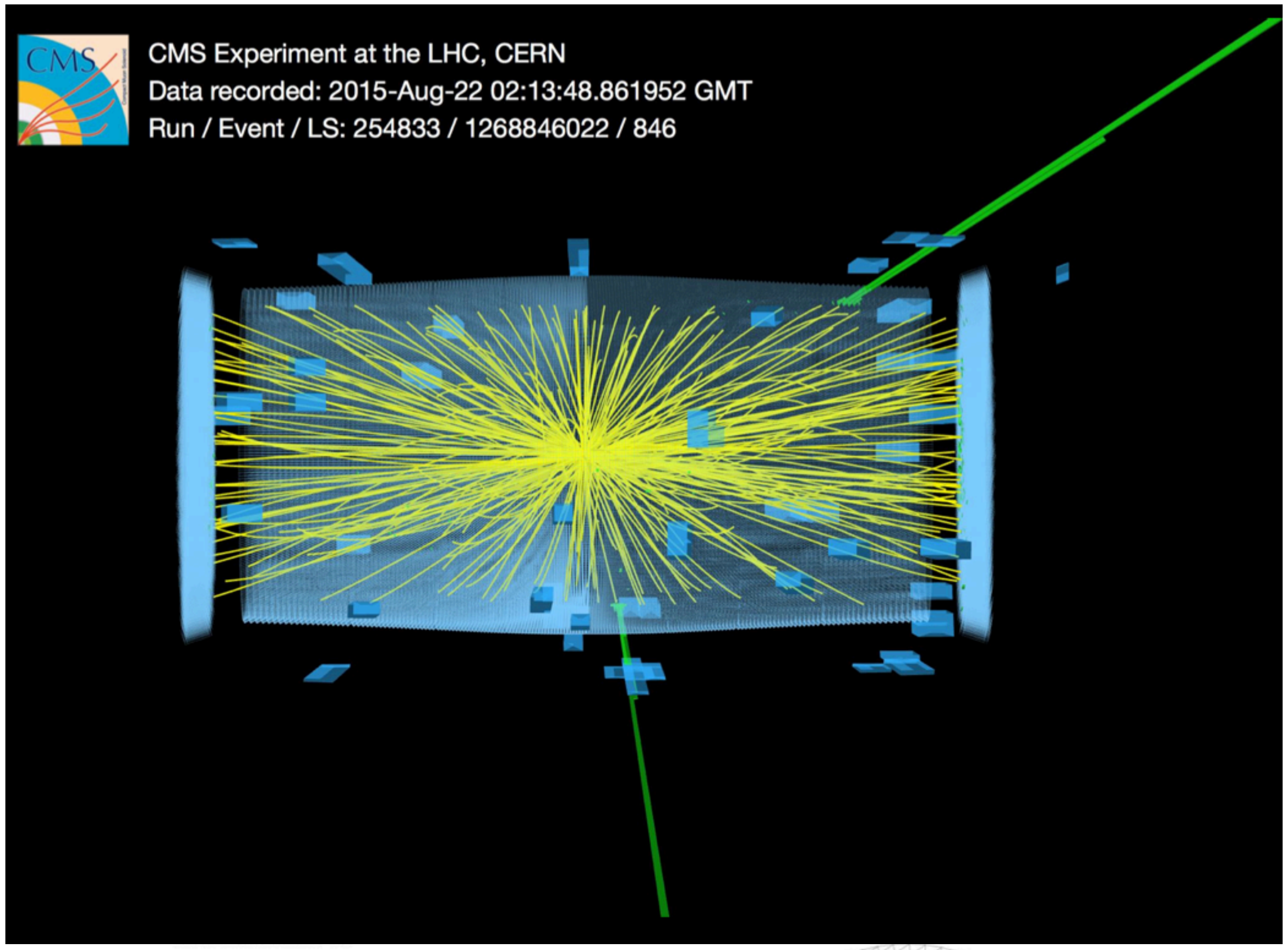
muons:
detected in outer
detector layers


CMS
3.8T Superconducting Solenoid
Hermetic (|η|<5.2)
Hadron Calorimeter (HCAL)
[scintillators & brass]
Lead tungstate
E/M Calorimeter (ECAL)
All Silicon Tracker
(Pixels and Microstrips)
Redundant Muon System
(RPCs, Drift Tubes,
Cathode Strip Chambers)

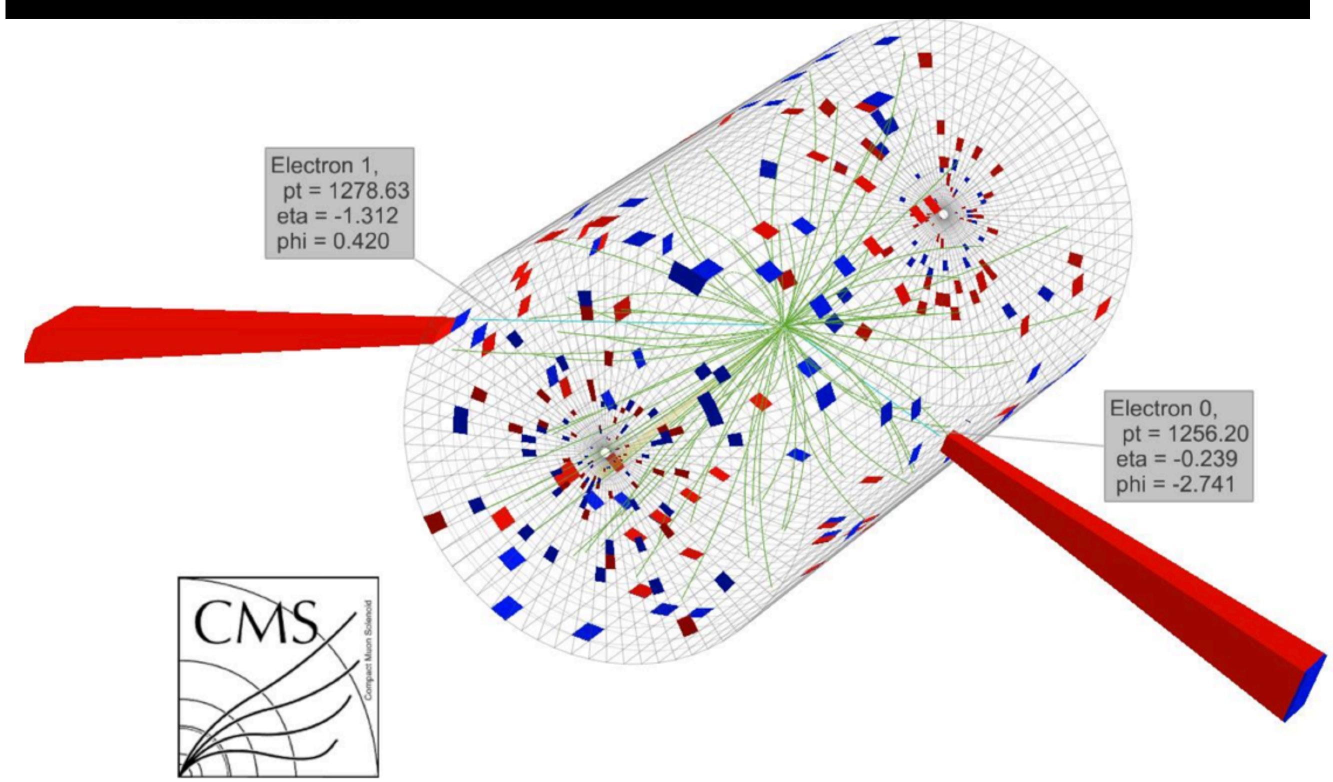

CMS
3.8T Superconducting Solenoid
Hermetic (|η|<5.2)
Hadron Calorimeter (HCAL)
[scintillators & brass]
Lead tungstate
E/M Calorimeter (ECAL)
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
HCAL
TRACKER
ECAL
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed
All Silicon Tracker
(Pixels and Microstrips)
Redundant Muon System
(RPCs, Drift Tubes,
Cathode Strip Chambers)

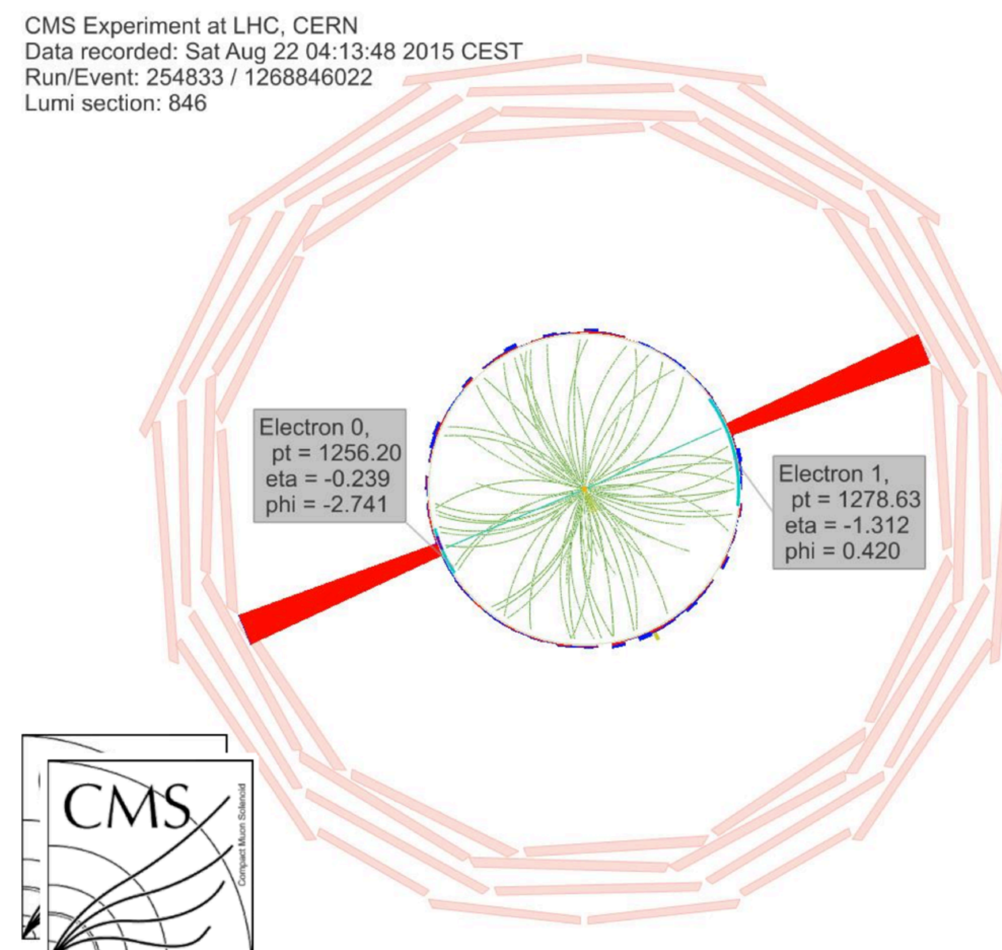

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
2T
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

a di-electron event

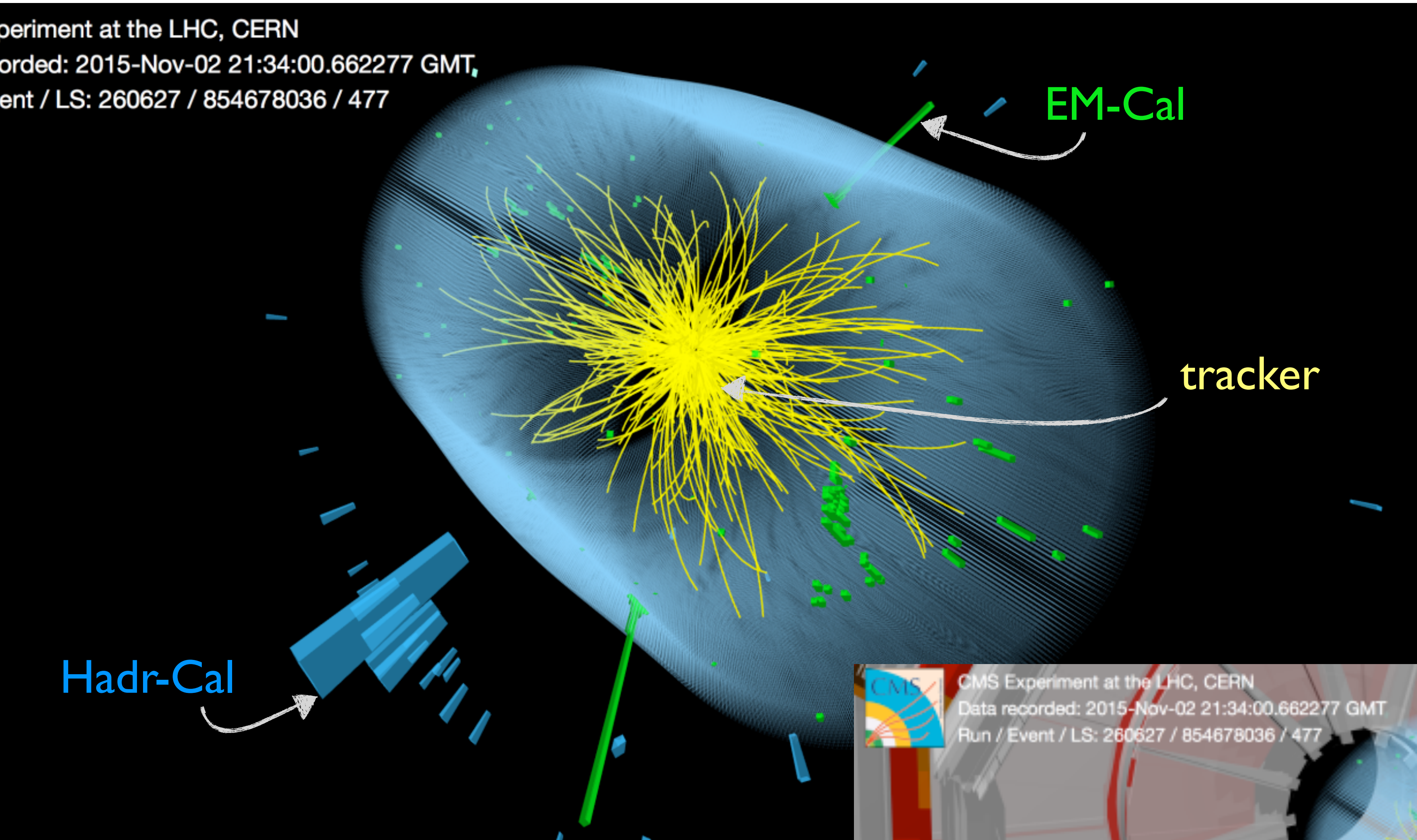


Event Display of a Candidate Electron-Positron Pair with an Invariant Mass of 2.9 TeV





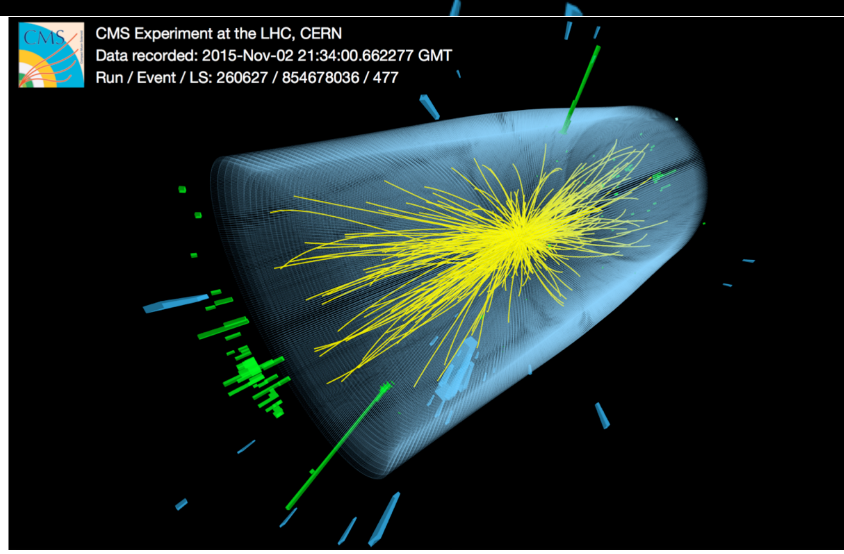
di-photons

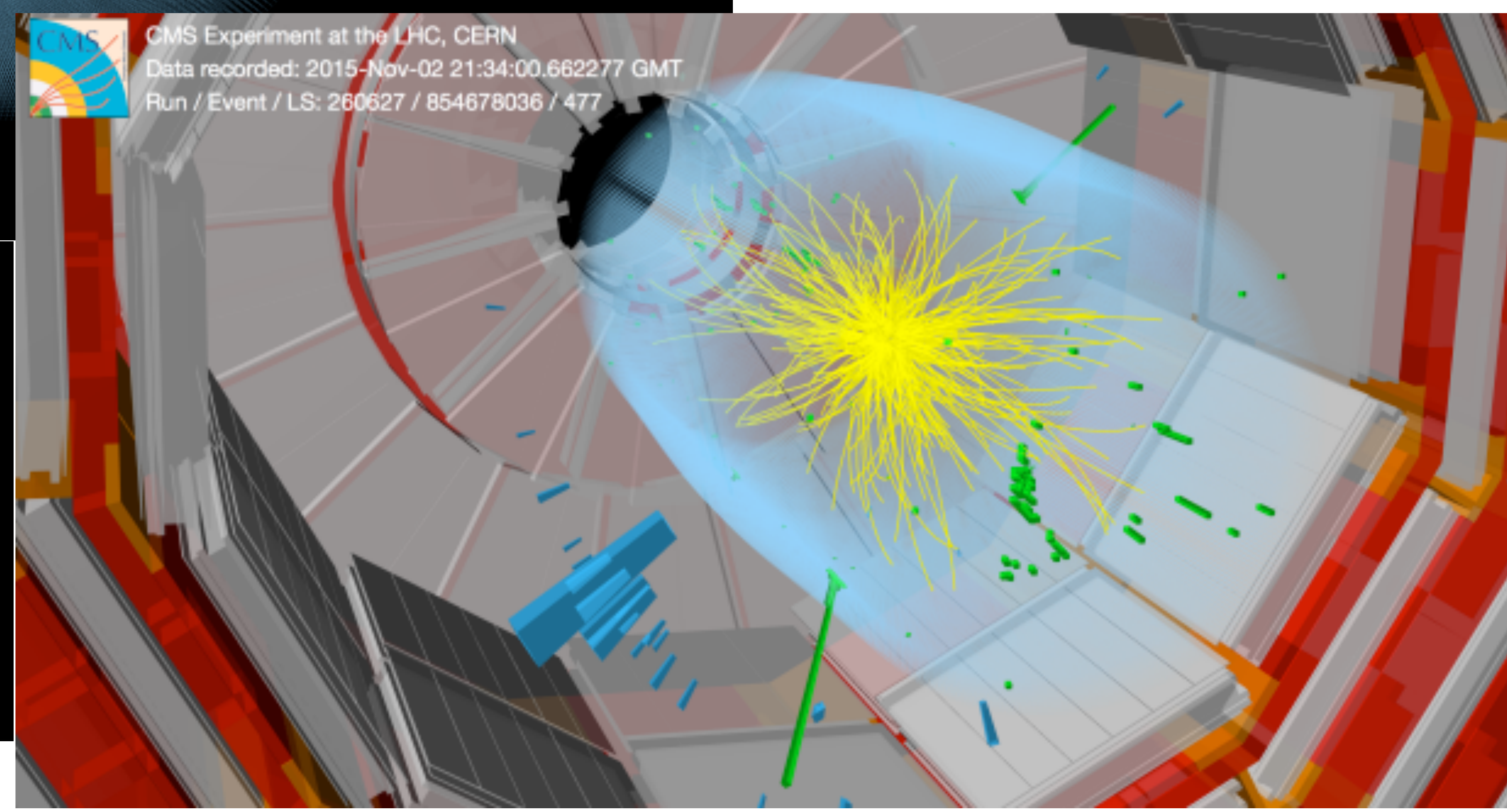


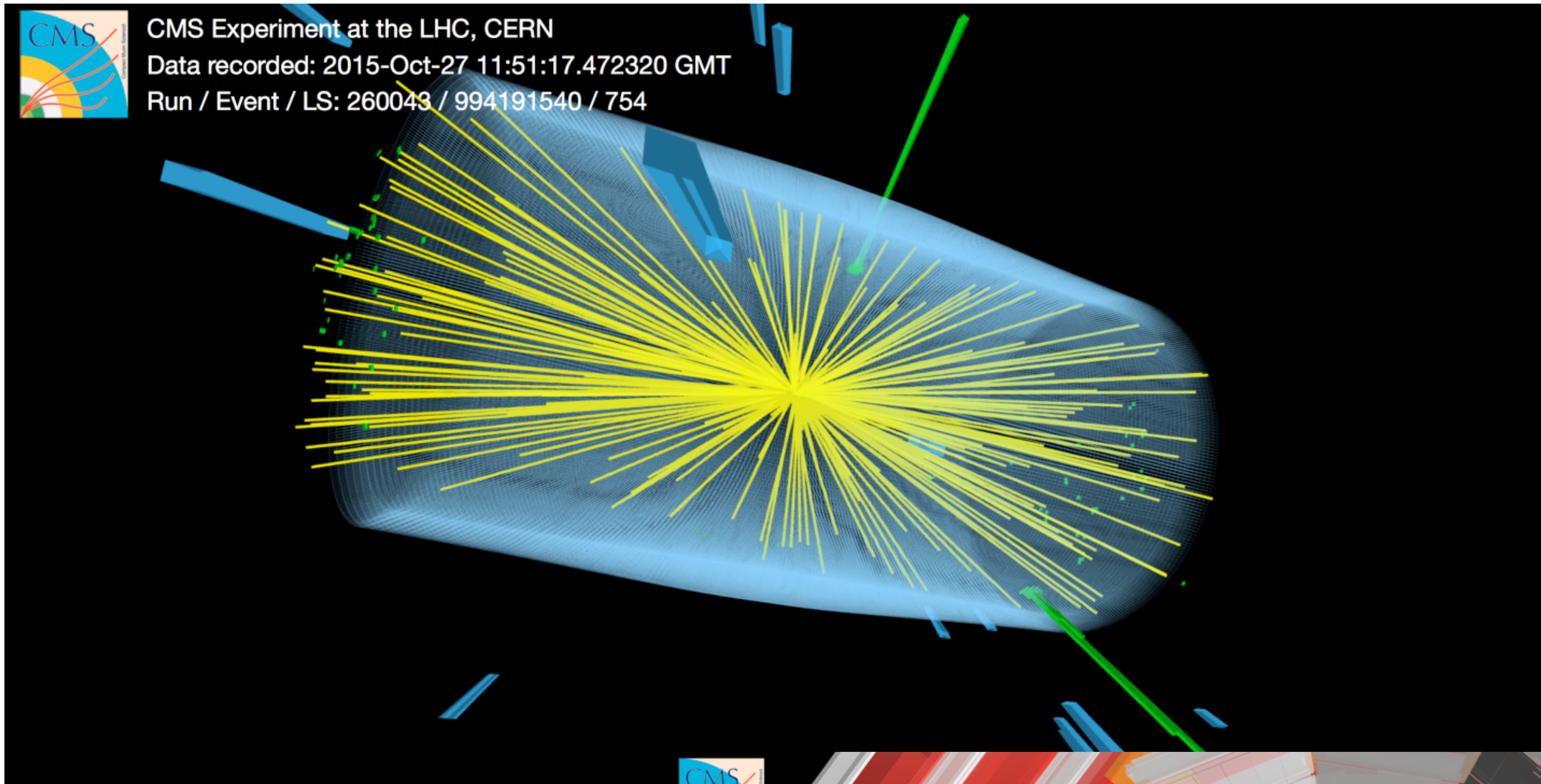
$X \to \gamma\gamma$

$m_{\gamma\gamma} \sim 750$ GeV

CMS-PHO-EVENTS-2015-007

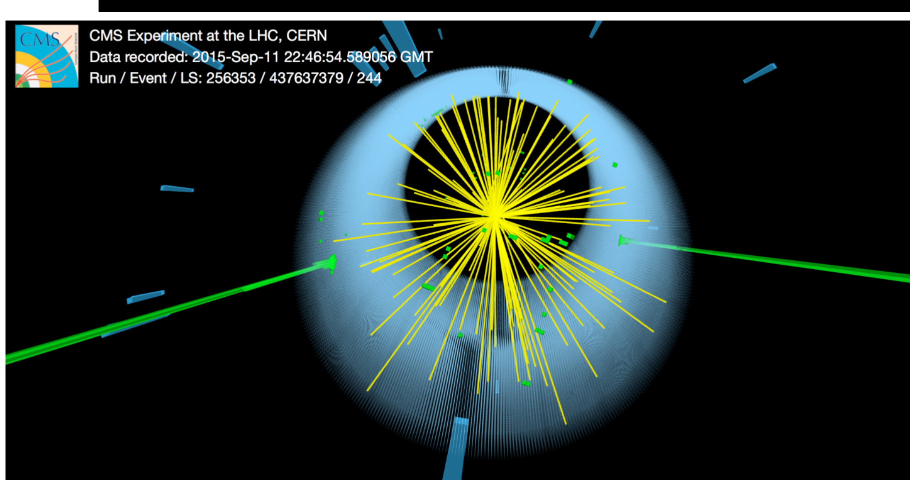


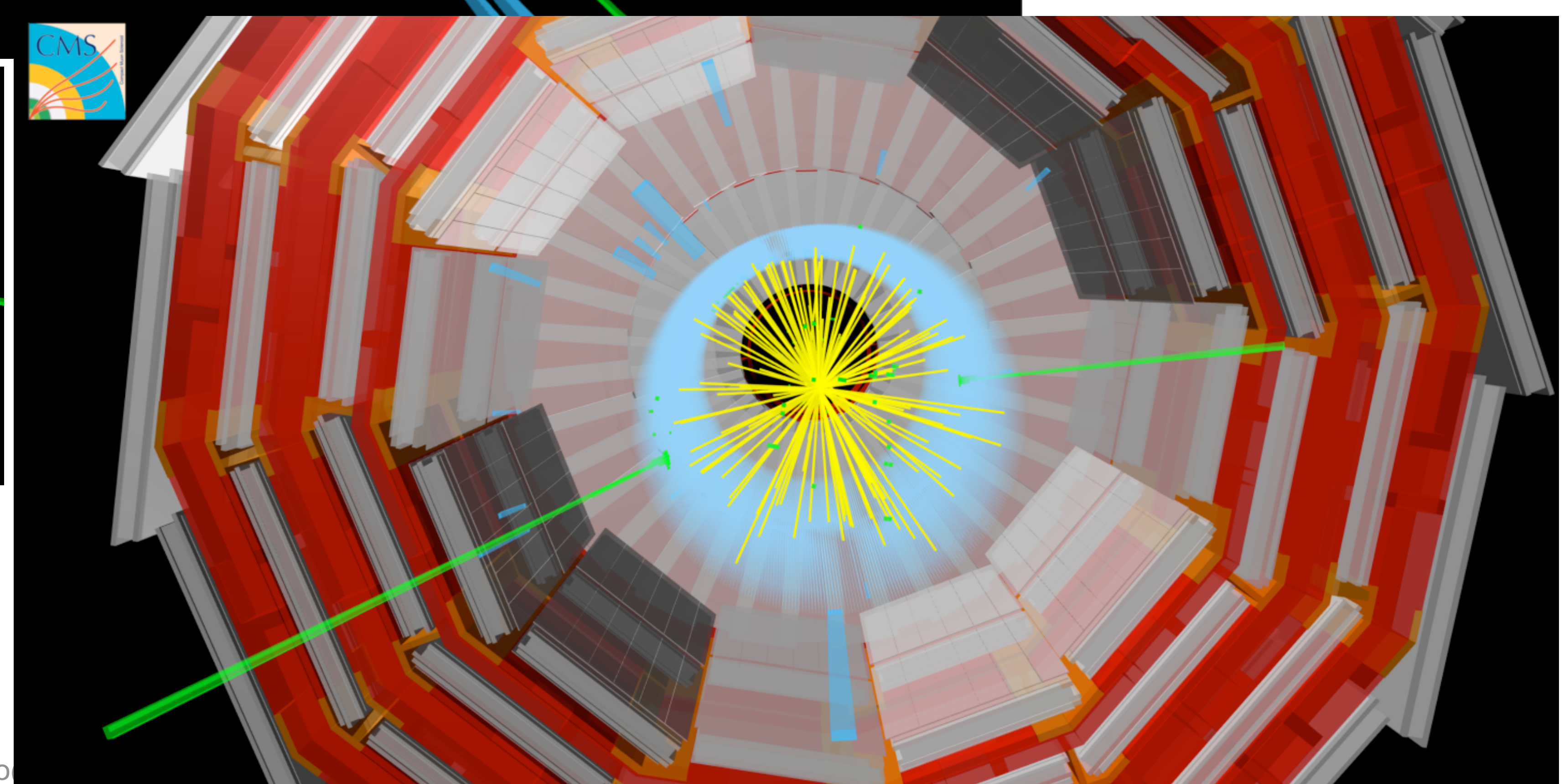


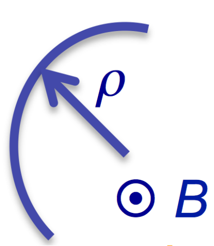


?

$m_{??} \sim 800$ GeV



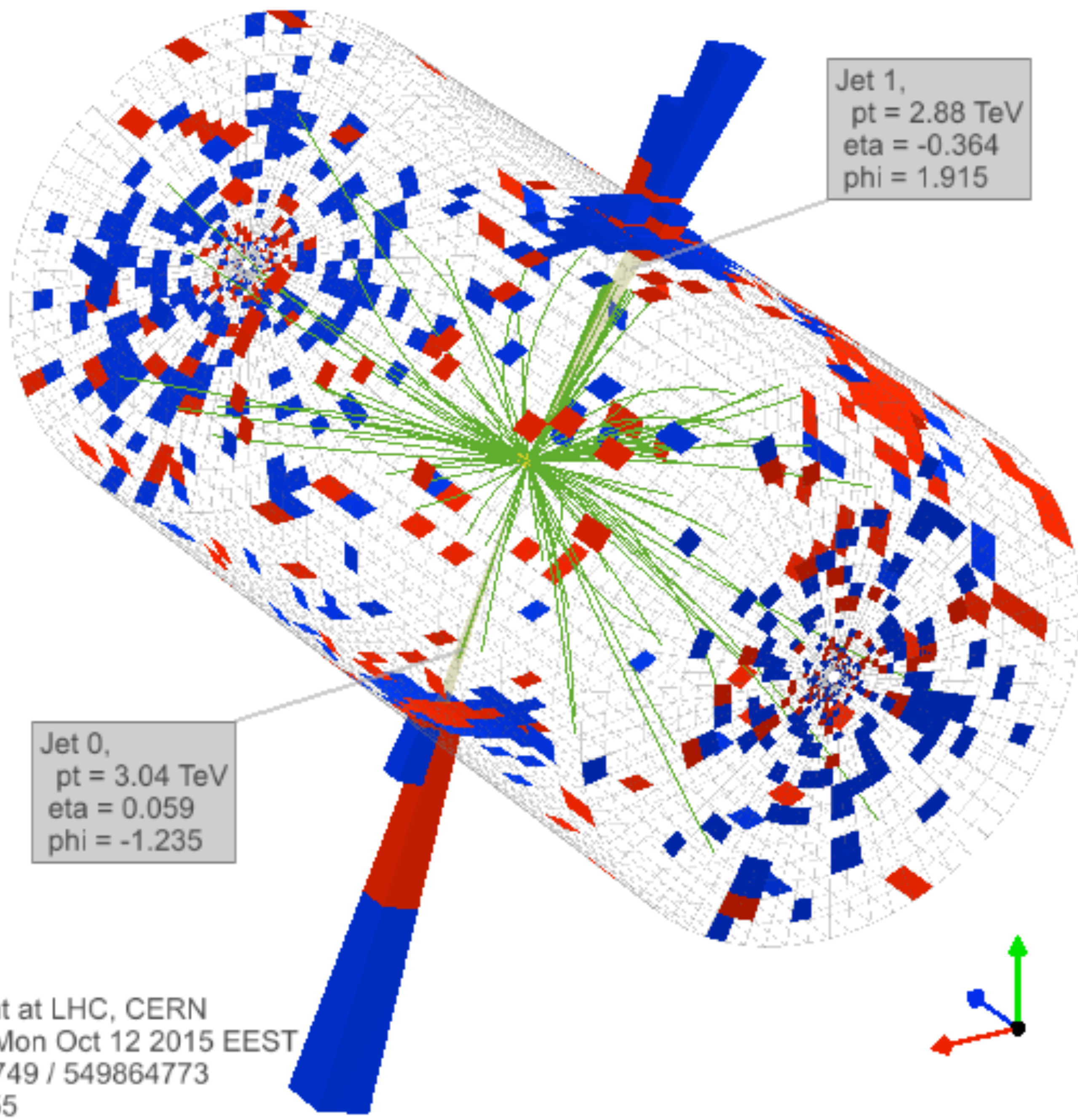




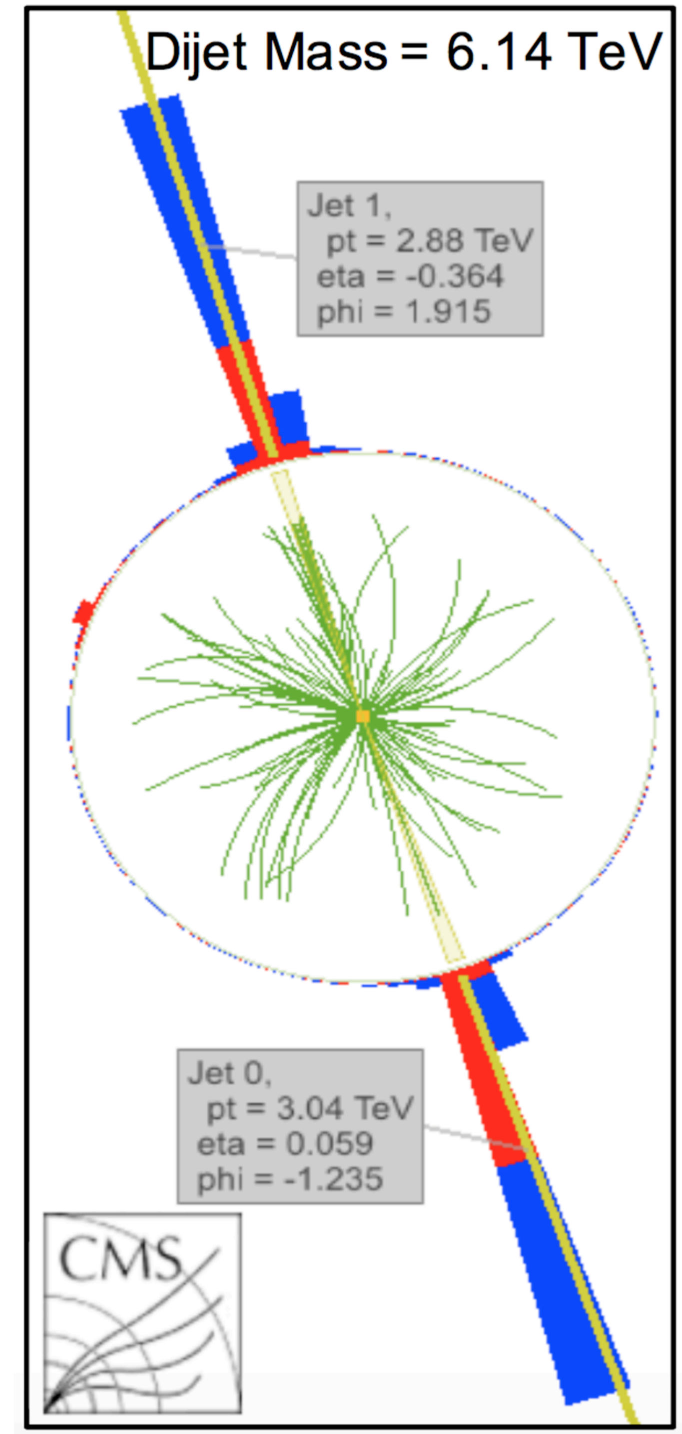
$$\rho = \frac{p}{ZeB}$$

di-jets

X → j j

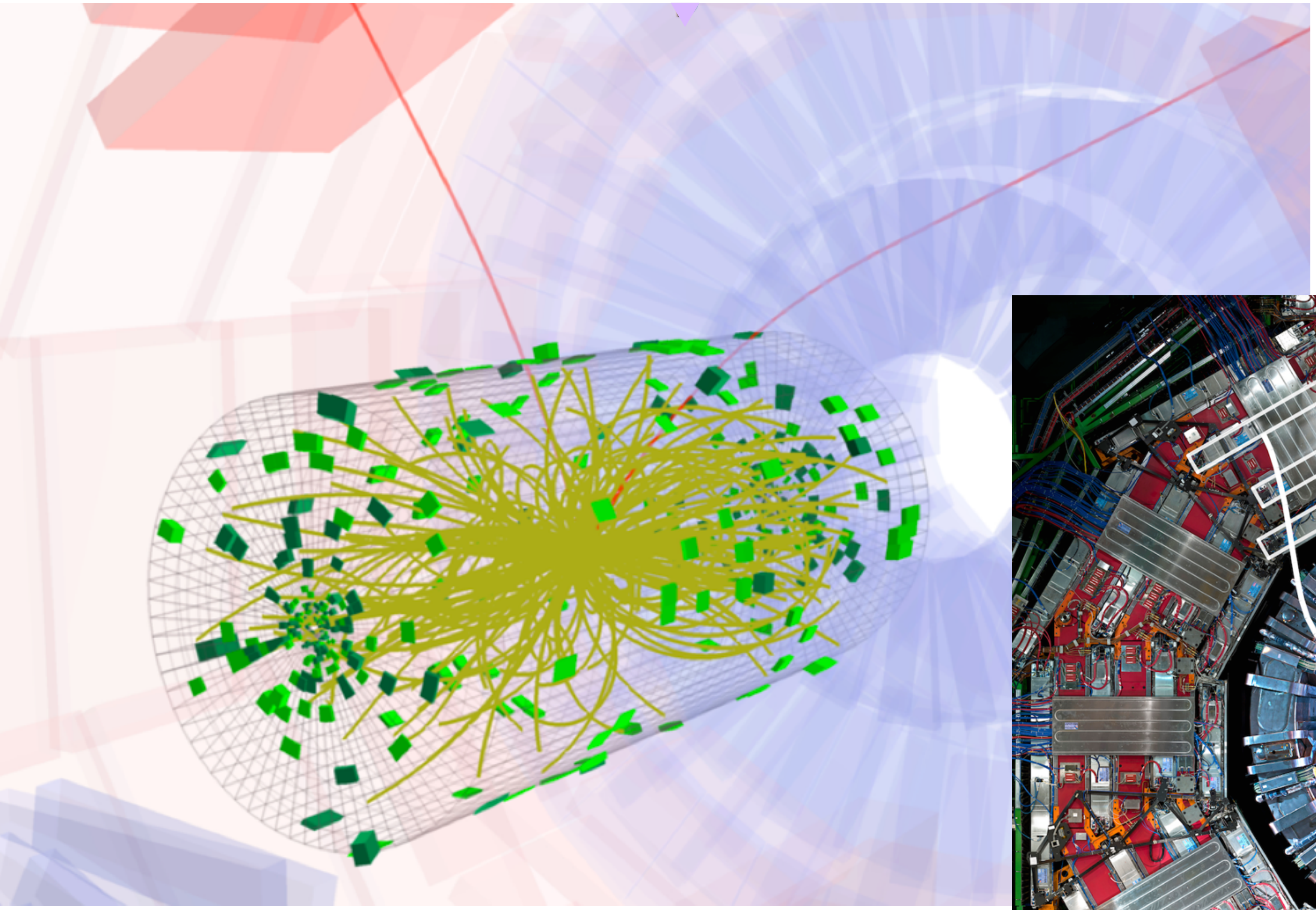


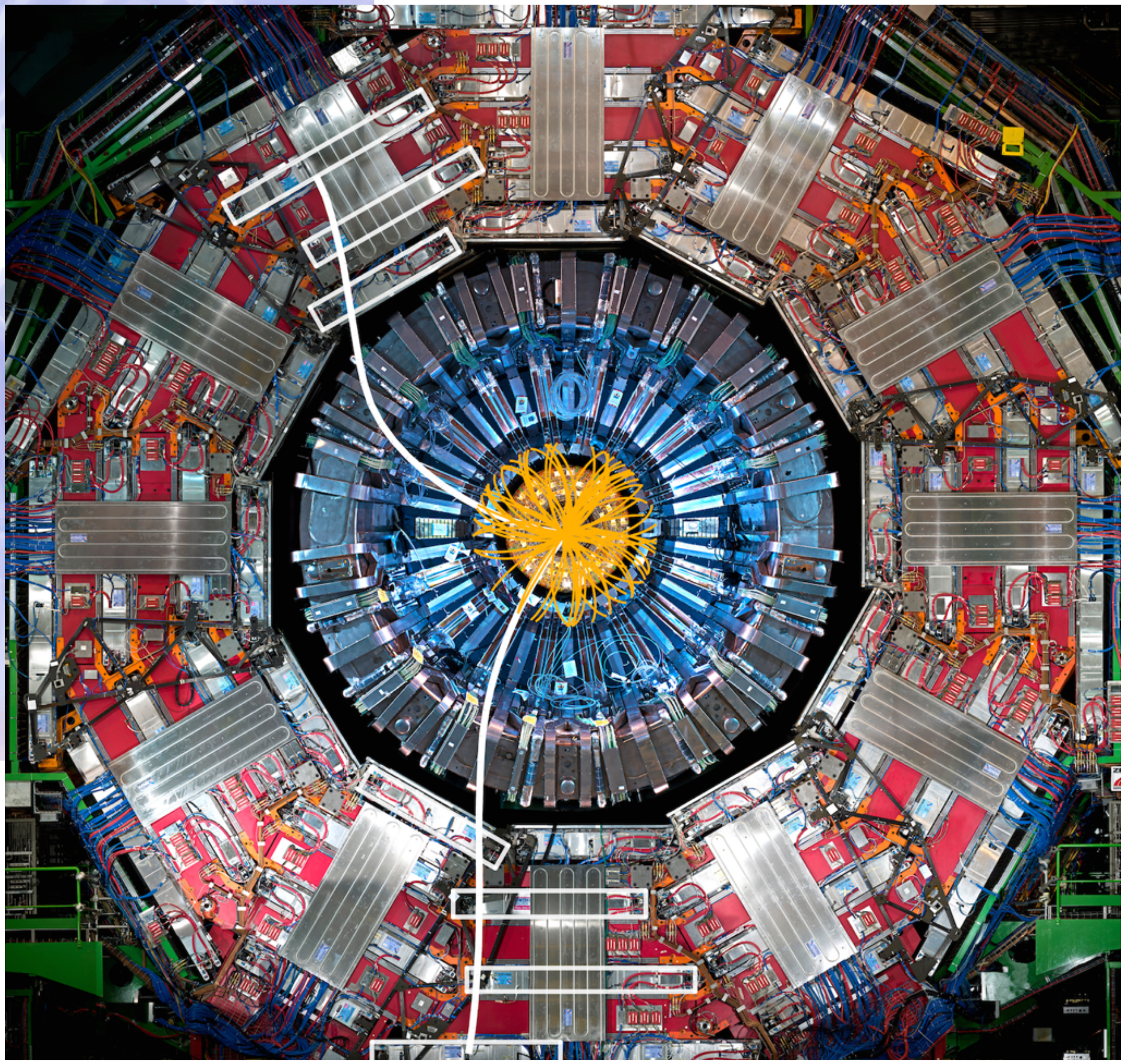
CMS Experiment at LHC, CERN
Data recorded: Mon Oct 12 2015 EEST
Run/Event: 258749 / 549864773
Lumi section: 355
Dijet Mass: 6.14 TeV



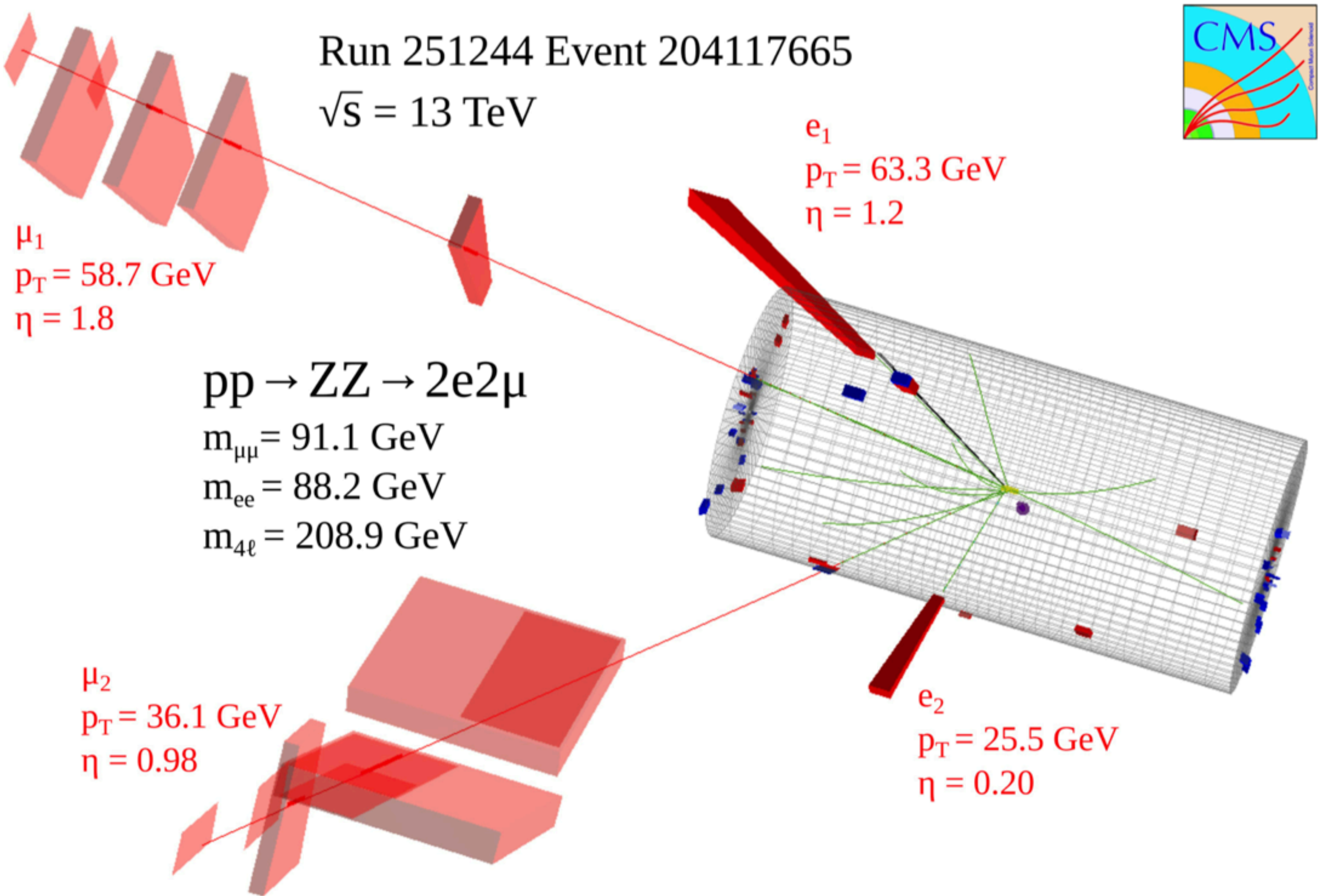
a di-muon event

X → μμ





a $\mu^+\mu^-e^+e^-$ event

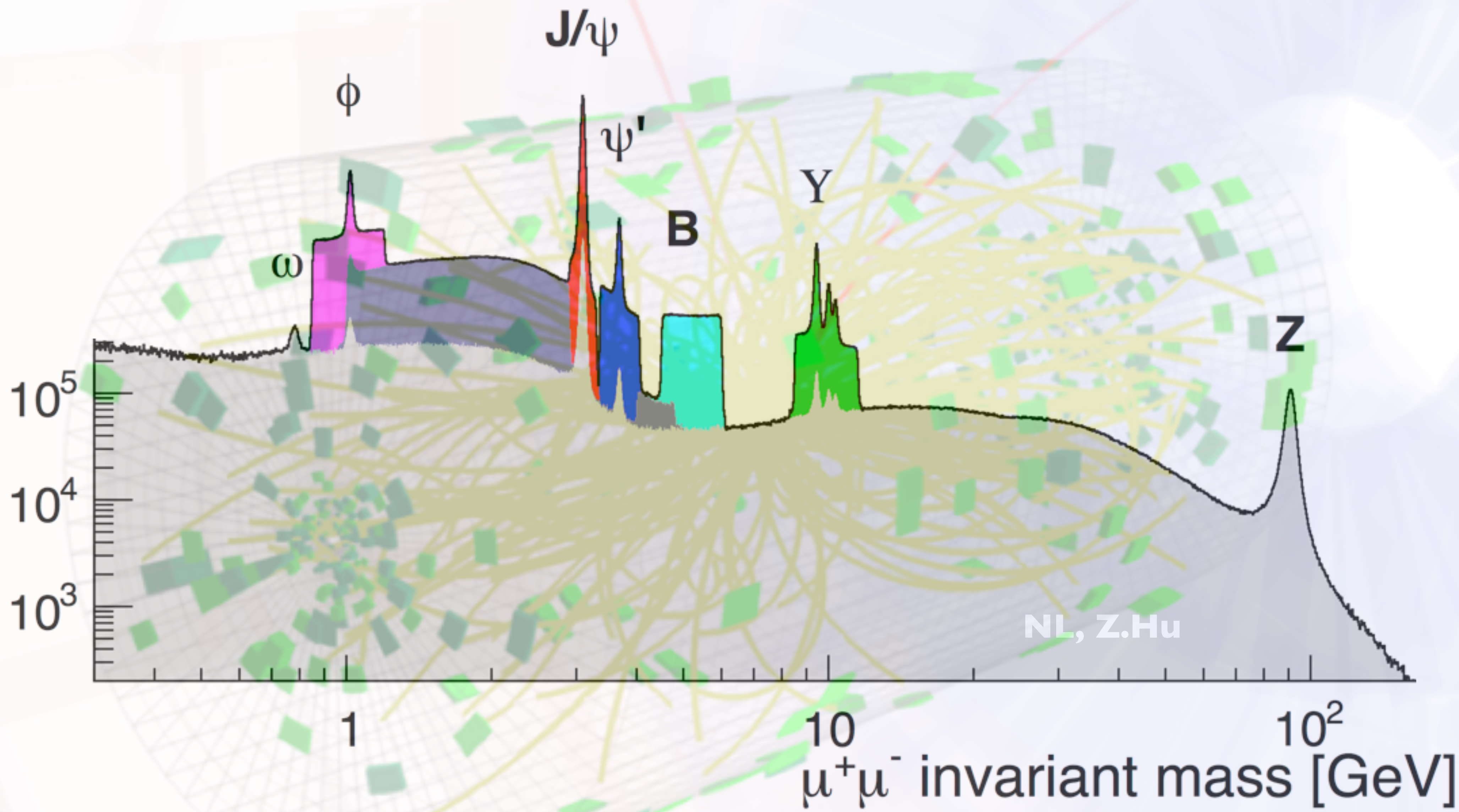


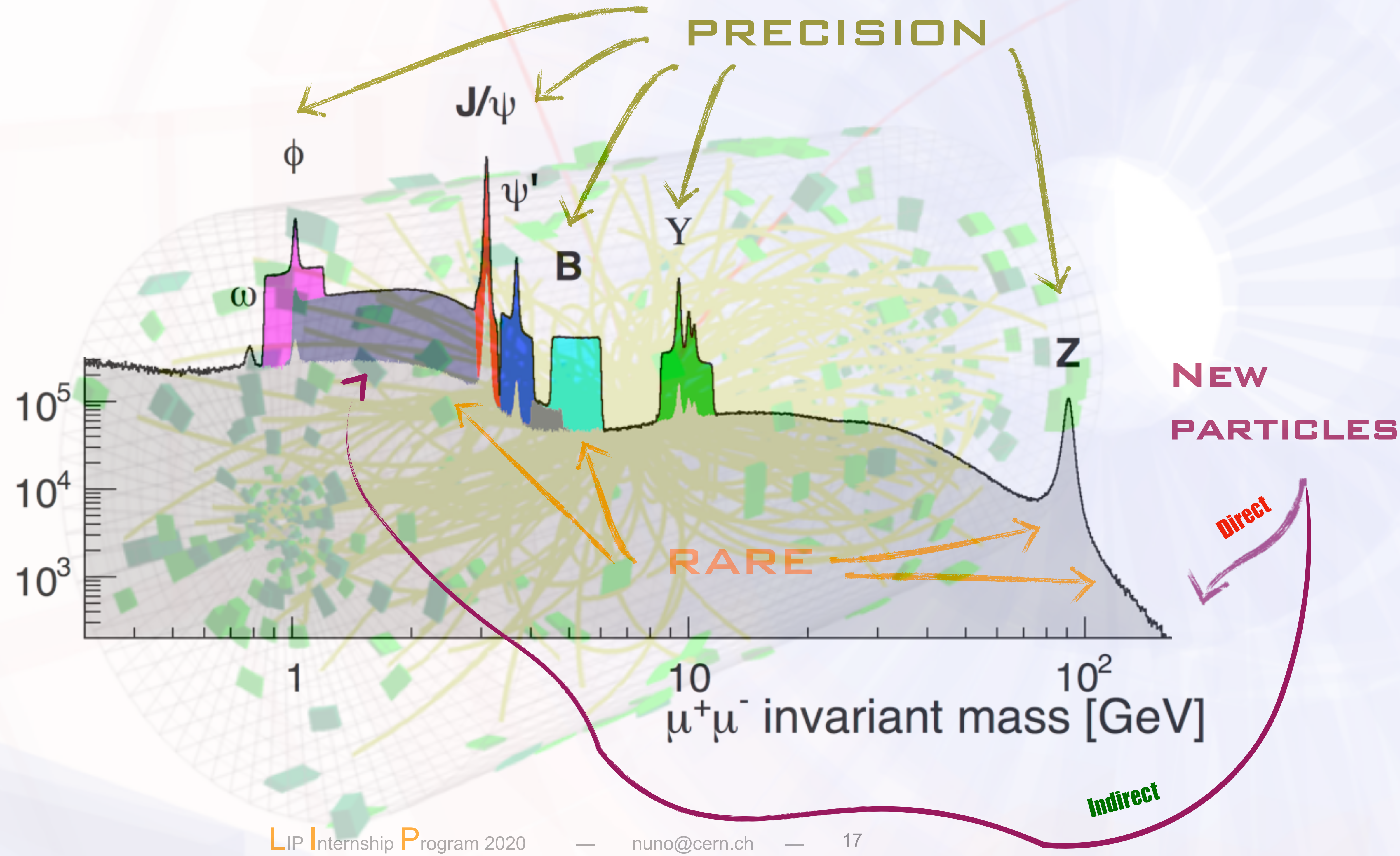
➡️ processes are explored with many (more) final state particles

the di-muon analysis

the di-muon spectrum ($X \to \mu\mu$)

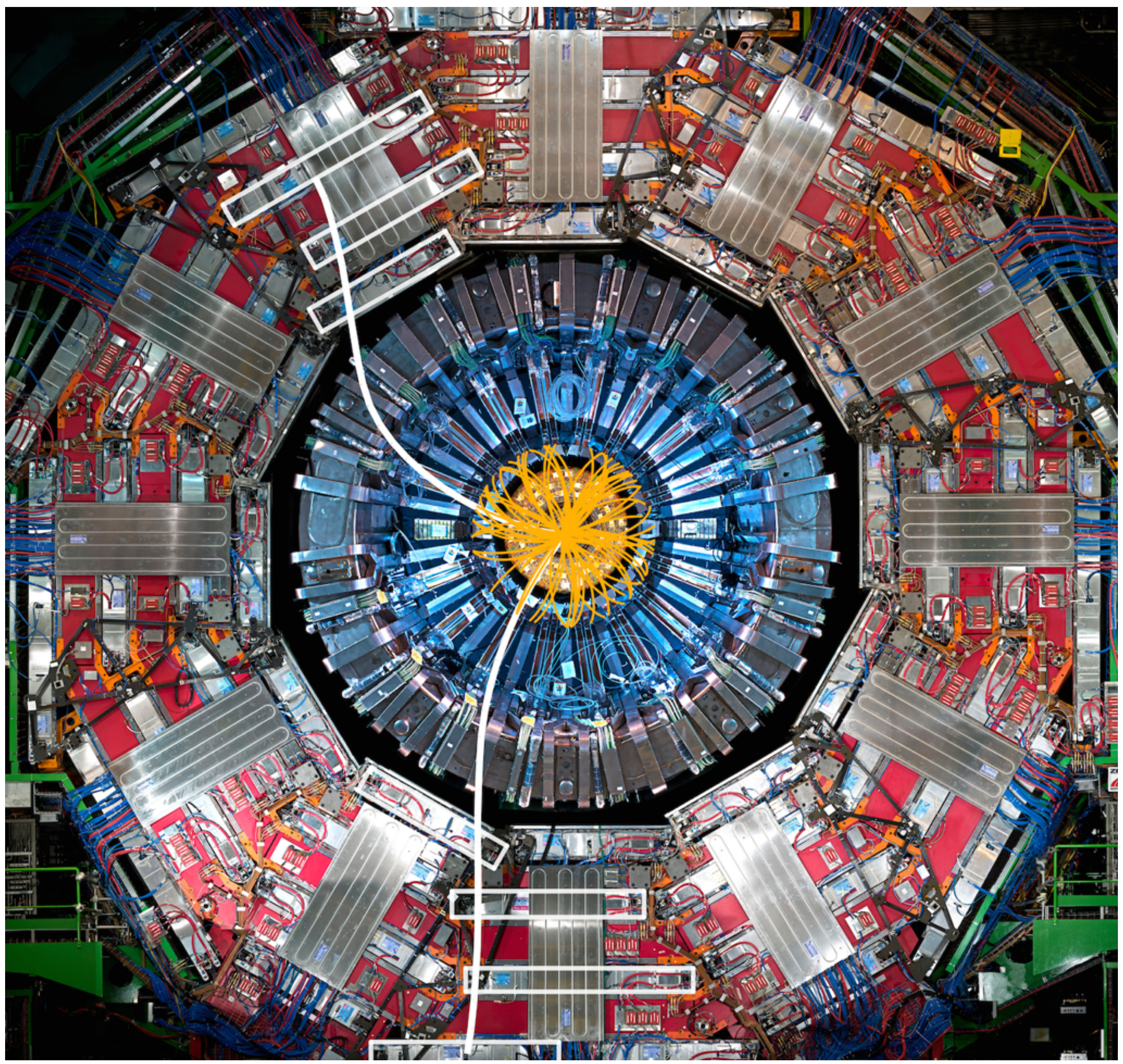
50 years of particle physics in one plot!

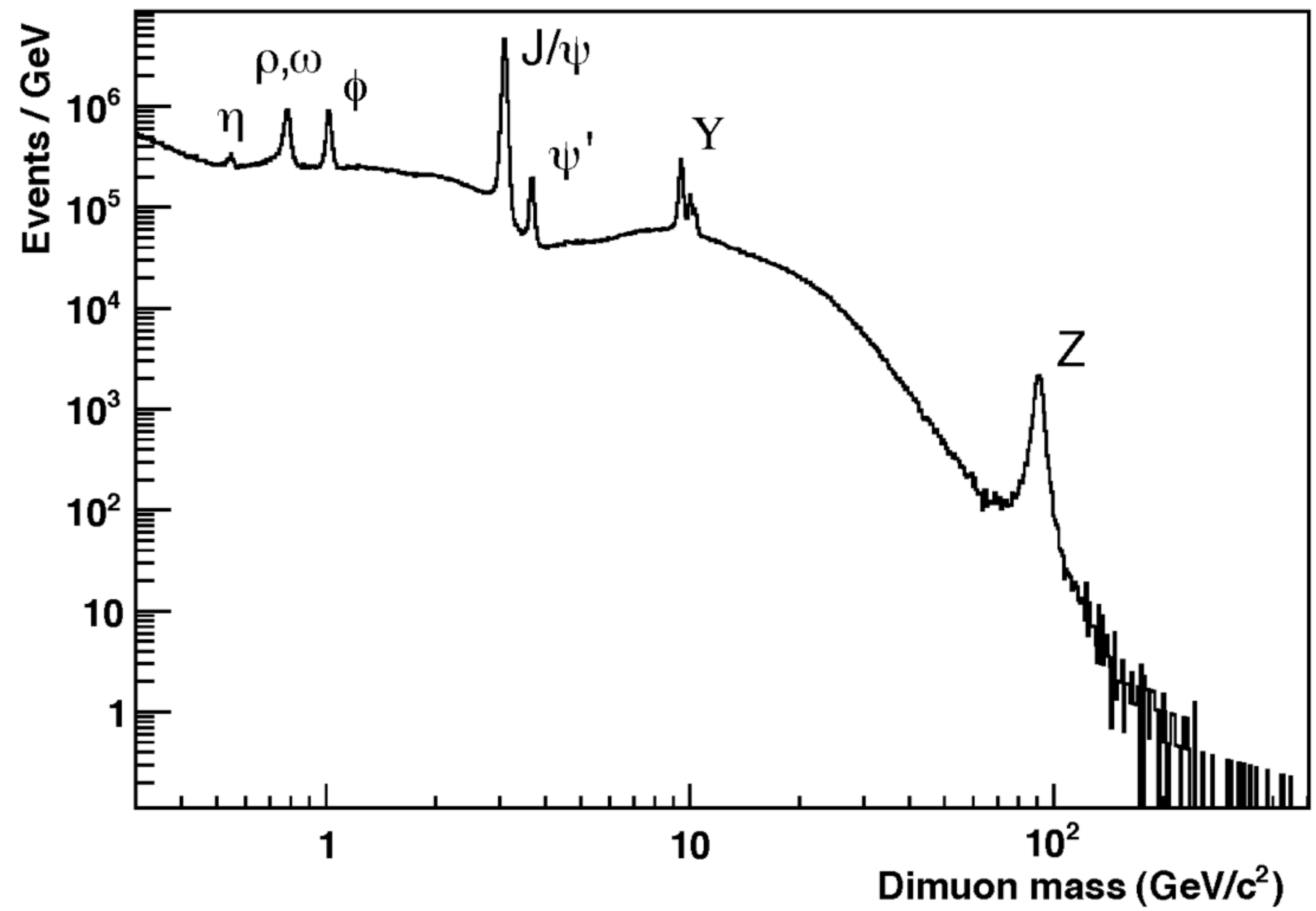



PRECISION
J/ψ
ϕ
ψ'
B
Y
ω
Z
New
particles
Direct
10^5
10^4
10^3
RARE
1
10
10^2
μ+μ- invariant mass [GeV]
Indirect

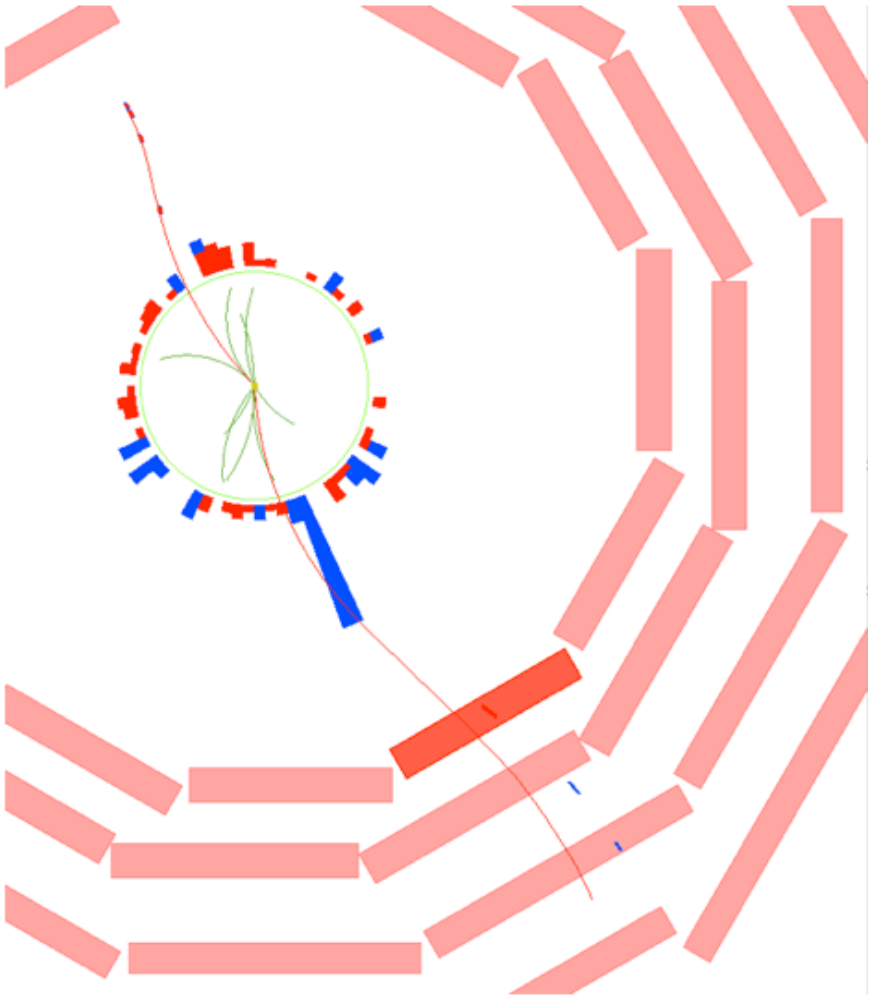
from detector to physics ...

?





di-muon ‘invariant mass’ ?

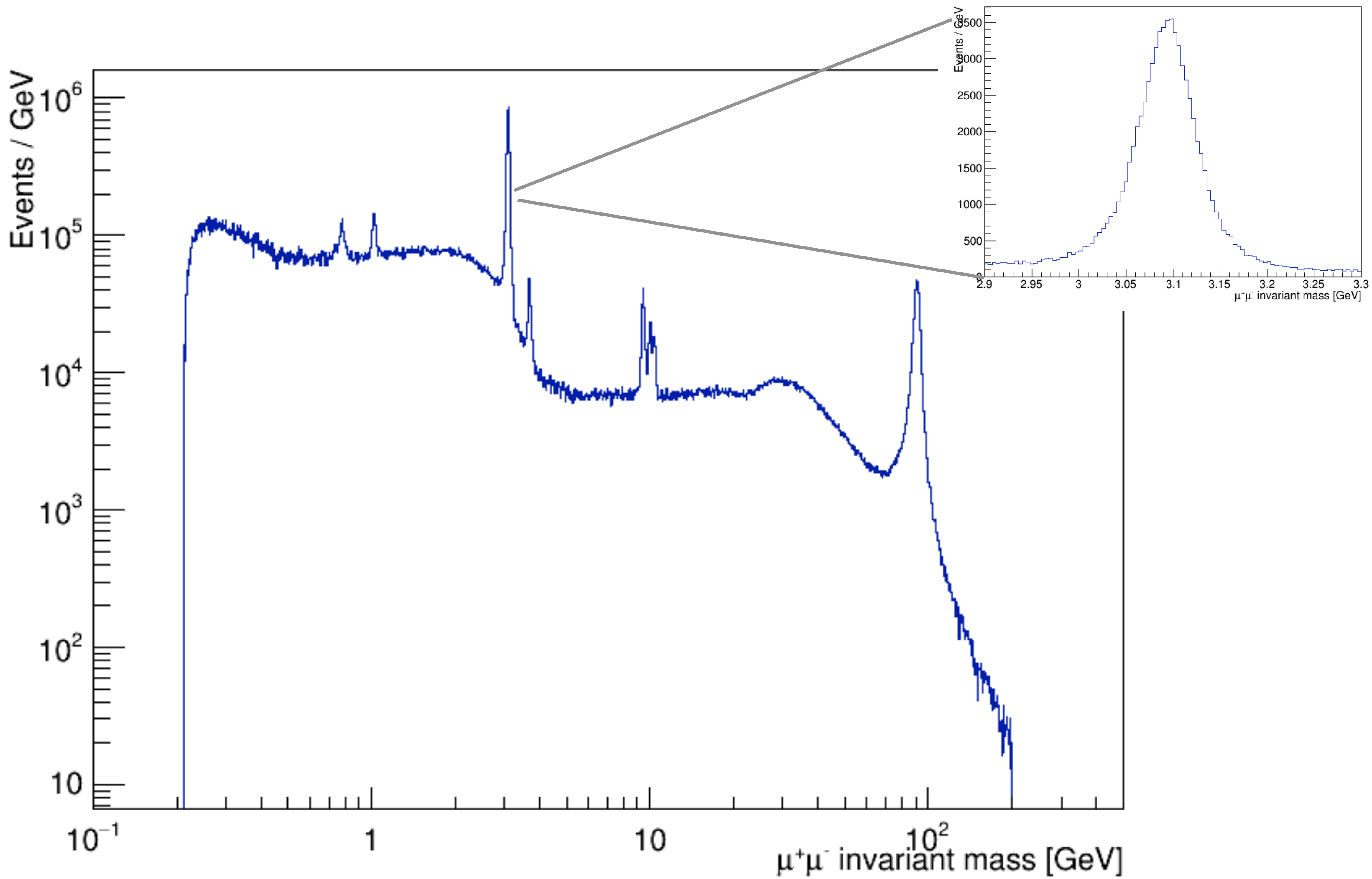


particle identification
- signal in muon chambers
→ it’s a muon!
➠ $m = m(\mu) \sim 106 MeV/c^2$

particle trajectory
- muon chambers but especially the silicon tracker
➠ linear momentum, $\underline{p} \equiv (p_x, p_y, p_z)$

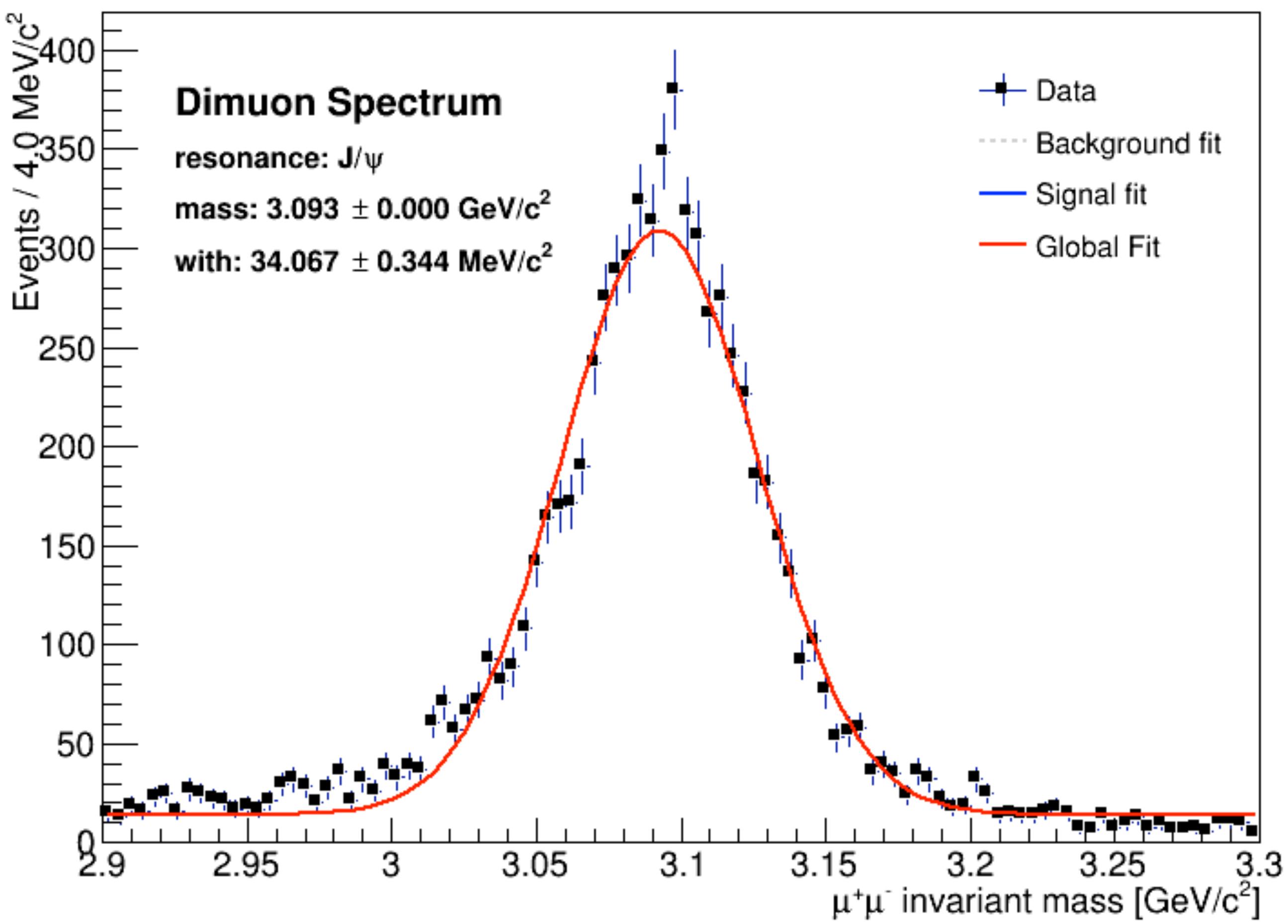
➠ form 4-momentum of each muon: $\mathbf{P}_\mu \equiv (E, p_x, p_y, p_z)$
➠ that of the di-muon pair $\mathbf{P}_{\mu\mu} = \mathbf{P}_{\mu 1} + \mathbf{P}_{\mu 2} = \mathbf{P}_{\mathbf{x} \to \mu\mu}$
➠ invariant mass $\mathbf{P}_{\mu\mu} \cdot \mathbf{P}_{\mu\mu} = \mathbf{M}_{\mu\mu}^2 = (\mathbf{M_x})^2$

the reconstructed di-muon spectrum



feature: variable bin widths, resolution-dependent, properly normalized, doubly-log scales

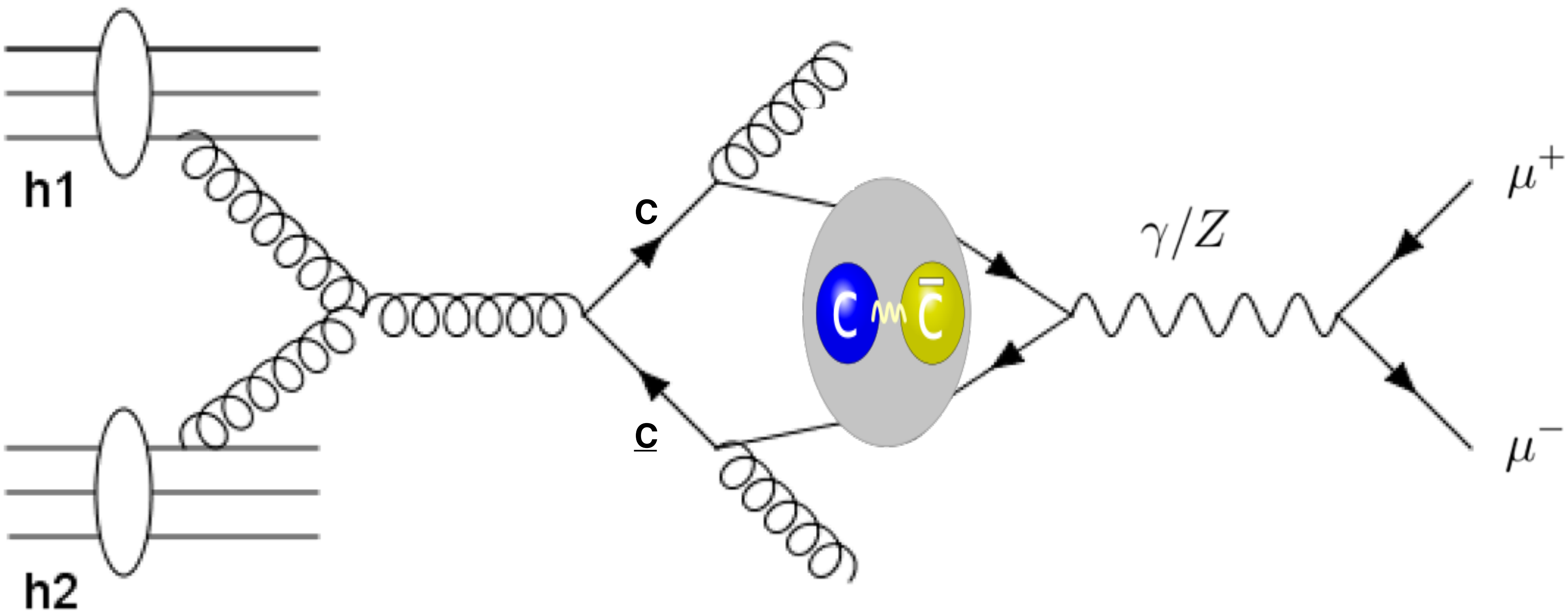
fit the data



- inspect quality of fit
  - can model be improved?
  - hint: final state radiation ($\mu \to \mu\gamma$) may distort shape

- establish a **fit model**
  - signal; Gaussian
  - background: polynominal
- extract **signal parameters**
  - yield ($N \pm \sigma_N$), mass ($m \pm \sigma_m$)
- estimate **systematic errors**
  - does the choice of fit model affect the measured results ?
  - quantify the systematic variations by employing different models
- quote **final measurements**
  - $N \pm \sigma_{stat} \pm \sigma_{syst}$

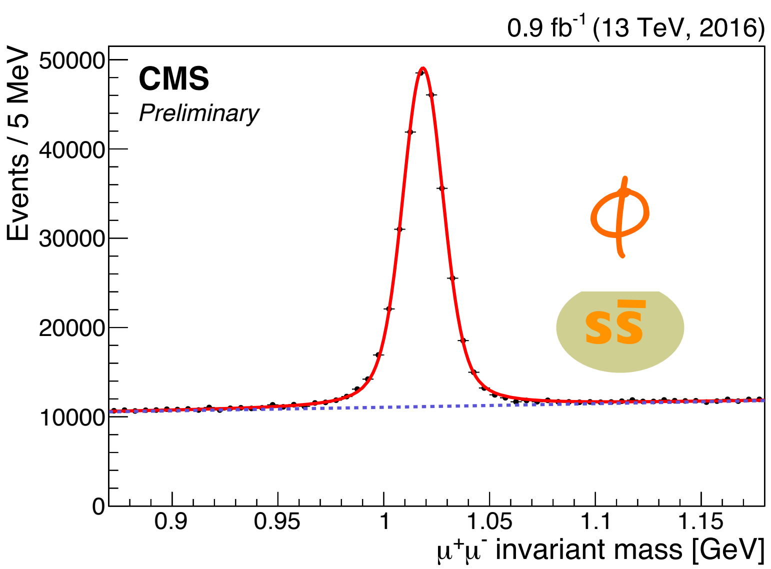
what's the physics process ?

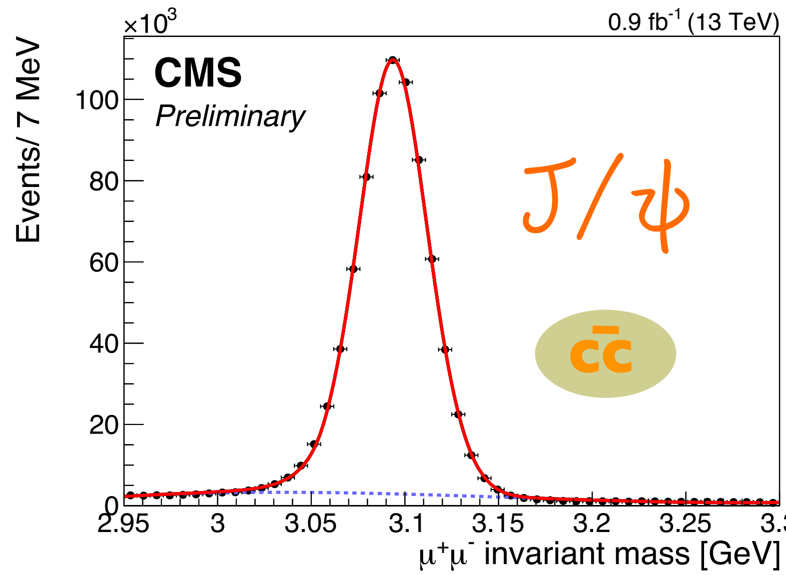


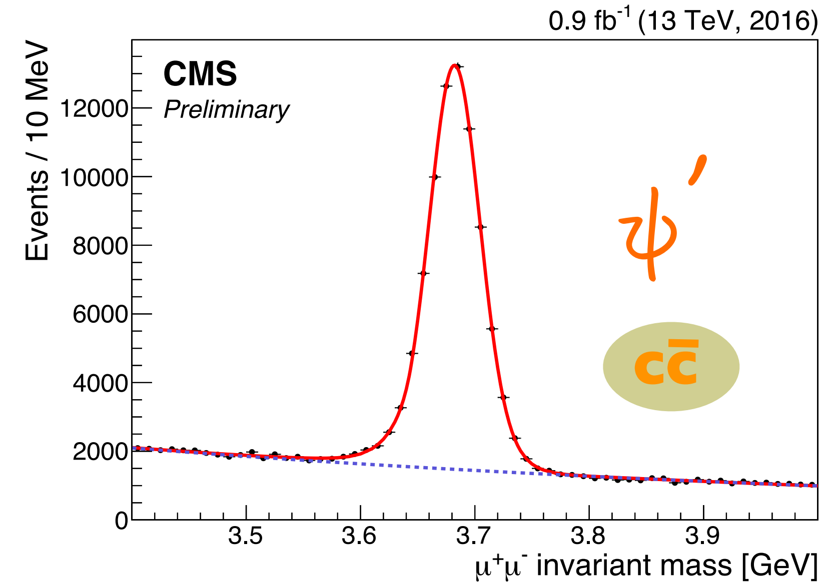
production: strong force

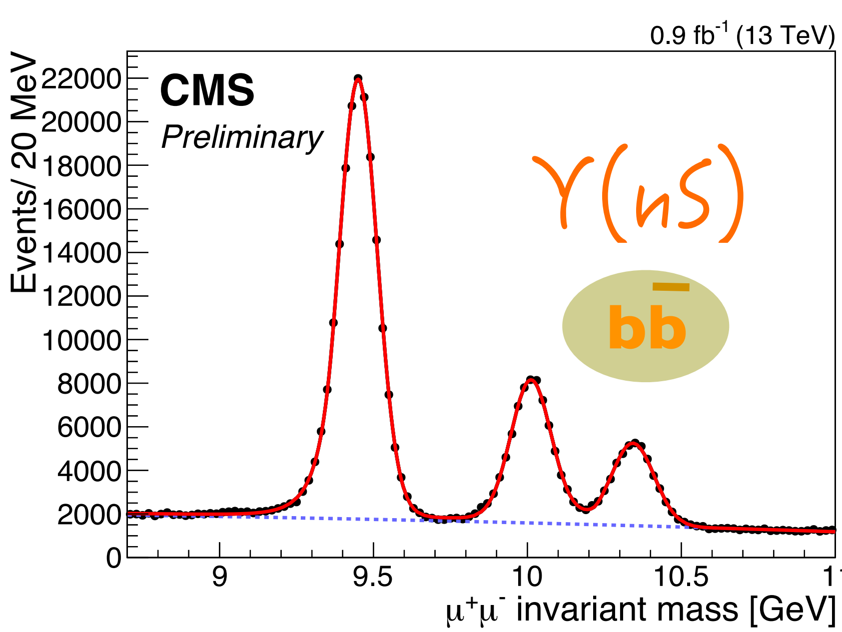
decay: electroweak force

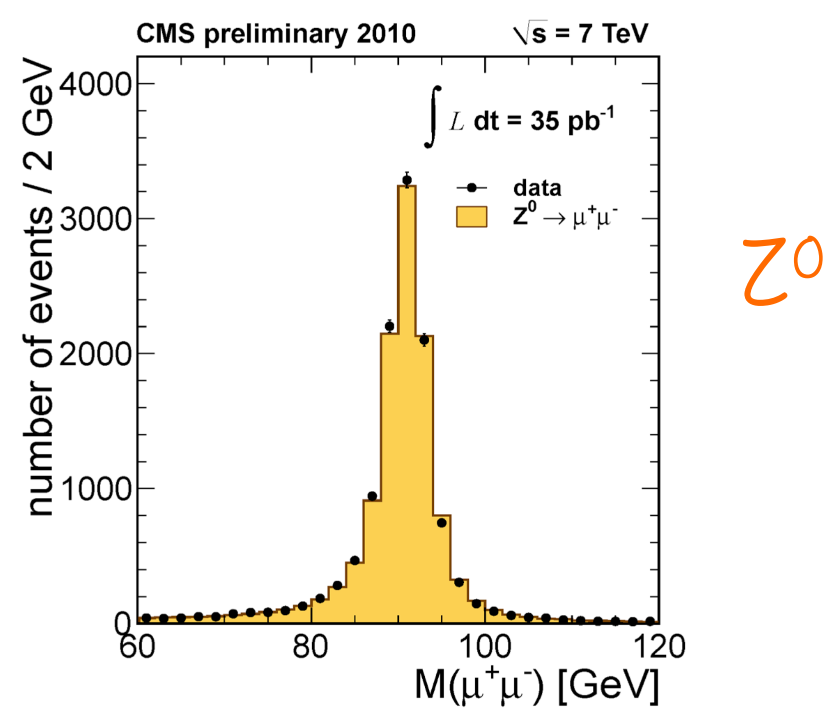
what are the peaks?











Check their measured properties from: http://pdglive.lbl.gov

GAUGE AND HIGGS BOSONS

INSPIRE search

Z $J = 1$

See related reviews:

Z Boson	PDF
Anomalous $ZZ\gamma$, $Z\gamma\gamma$, and ZZV Couplings	PDF
Anomalous W/Z Quartic Couplings (QGCs)	PDF

‣ Expand all sections

Z MASS	91.1876 ± 0.0021 GeV

55. Z Boson

Revised August 2018 by M. Grünewald (University Coll. Dublin) and A. Gurtu (CERN; TIFR Mumbai).

Precision measurements at the Z-boson resonance using electron–positron colliding beams began in 1989 at the SLC and at LEP. During 1989–95, the four LEP experiments (ALEPH, DELPHI, L3, OPAL) made high-statistics studies of the production and decay properties of the Z. Although the SLD experiment at the SLC collected much lower statistics, it was able to match the precision of LEP experiments in determining the effective electroweak mixing angle $\sin^2\overline{\theta}_W$ and the rates of Z decay to b- and c-quarks, owing to availability of polarized electron beams, small beam size, and stable beam spot.

The Z-boson properties reported in this section may broadly be categorized as:

- The standard 'lineshape' parameters of the Z consisting of its mass, M_Z, its total width, Γ_Z, and its partial decay widths, Γ(hadrons), and $\Gamma(\ell\overline{\ell})$ where $\ell = e, \mu, \tau, \nu$;
- Z asymmetries in leptonic decays and extraction of Z couplings to charged and neutral leptons;
- The b- and c-quark-related partial widths and charge asymmetries which require special techniques;
- Determination of Z decay modes and the search for modes that violate known conservation laws;
- Average particle multiplicities in hadronic Z decay;
- Z anomalous couplings.

The effective vector and axial-vector coupling constants describing the Z-to-fermion coupling are also measured in $p\overline{p}$ and ep collisions at the Tevatron and at HERA. The corresponding cross-section formulae are given in Section 39 (Cross-section formulae for specific processes) and Section 16 (Structure Functions) in this *Review*. In this minireview, we concentrate on the measurements in e^+e^- collisions at LEP and SLC.

The standard 'lineshape' parameters of the Z are determined from an analysis of the production cross sections of these final states in e^+e^- collisions. The $Z \to \nu\overline{\nu}(\gamma)$ state is identified directly by detecting single photon production and indirectly by subtracting the visible partial widths from the total width. Inclusion in this analysis of the forward-backward asymmetry of charged leptons, $A_{FB}^{(0,\ell)}$, of the τ polarization, $P(\tau)$, and its forward-backward asymmetry, $P(\tau)^{fb}$, enables the separate determination of the effective vector ($\overline{g}_V$) and axial vector ($\overline{g}_A$) couplings of the Z to these leptons and the ratio ($\overline{g}_V/\overline{g}_A$), which is related to the effective electroweak mixing angle

QUARKS

c $I(J^P) = 0(1/2^+)$

Charge = $\frac{2}{3}$ e Charm = $+1$

c-QUARK MASS	1.27 ± 0.02 GeV
m_c/m_s MASS RATIO	$11.76^{+0.05}_{-0.10}$
m_b/m_c MASS RATIO	4.58 ± 0.01
$m_b - m_c$ QUARK MASS DIFFERENCE	3.45 ± 0.05 GeV

$c\,\overline{c}$ MESONS
(including possibly non- $q\,\overline{q}$ states)

INSPIRE search

$J/\psi(1S)$ $I^G(J^{PC}) = 0^-(1^{--})$

$J/\psi(1S)$ MASS	3096.900 ± 0.006 MeV
$J/\psi(1S)$ WIDTH	92.6 ± 1.7 keV (S = 1.1)

$J/\psi(1S)$ **Decay Modes** ‣ Expand all decays

	Mode		Fraction (Γ_i/Γ)	Scale Factor/ Conf. Level	P(MeV/c)
Γ_1	hadrons		$(87.7 \pm 0.5)\%$		
Γ_2	virtual $\gamma \to$ hadrons		$(13.50 \pm 0.30)\%$		
Γ_3	ggg		$(64.1 \pm 1.0)\%$		
Γ_4	γgg		$(8.8 \pm 1.1)\%$		
Γ_5	e^+e^-		$(5.971 \pm 0.032)\%$		1548
Γ_6	$e^+e^-\gamma$	[1]	$(8.8 \pm 1.4) \times 10^{-3}$		1548
Γ_7	$\mu^+\mu^-$		$(5.961 \pm 0.033)\%$		1545

‣ Decays involving hadronic resonances

‣ Decays into stable hadrons

‣ Radiative decays

‣ Dalitz decays

‣ Weak decays

‣ Charge conjugation (C), Parity (P), Lepton Family number (LF) violating modes

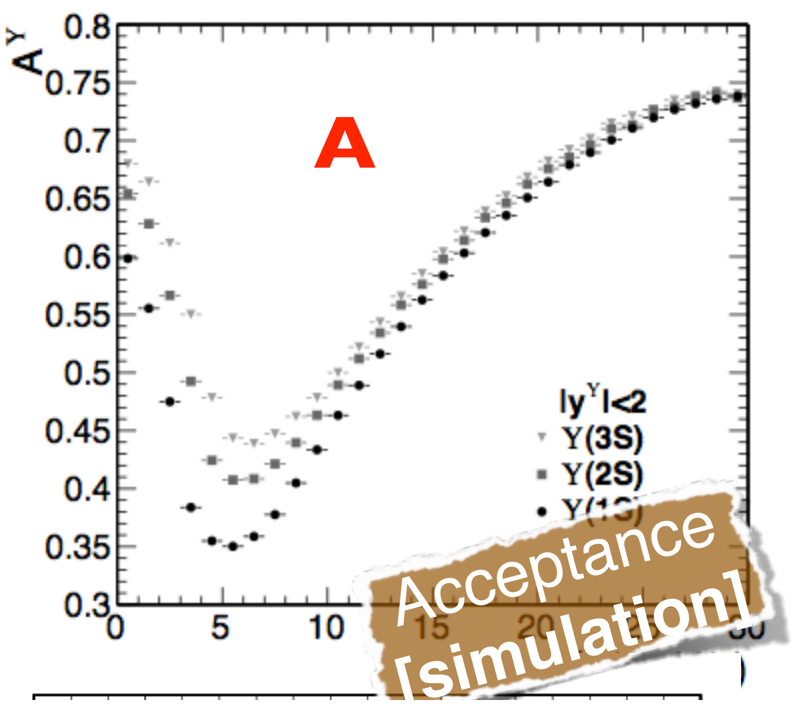
‣ Other decays

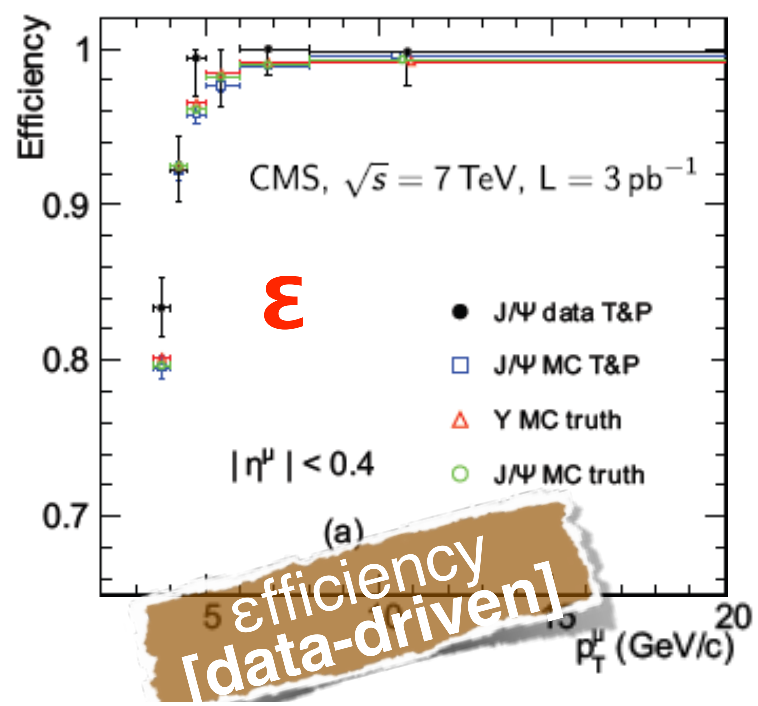
Cross section

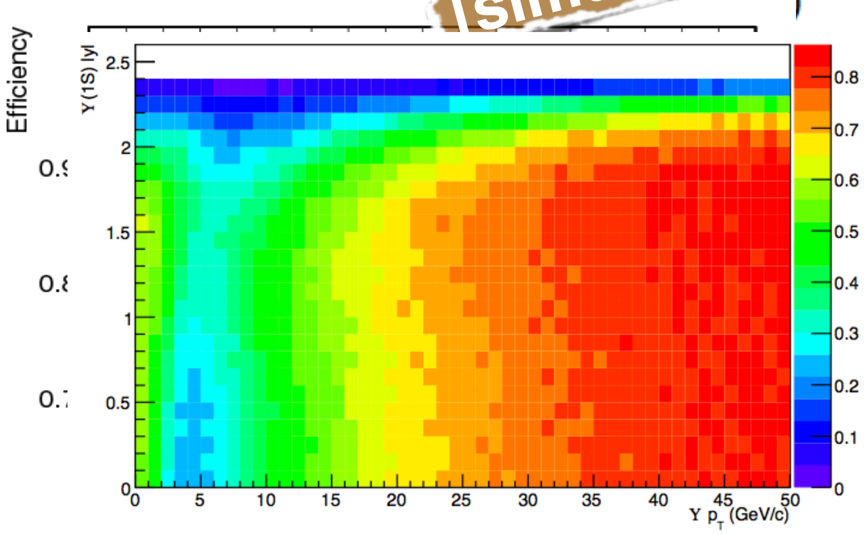
“N=L.σ”

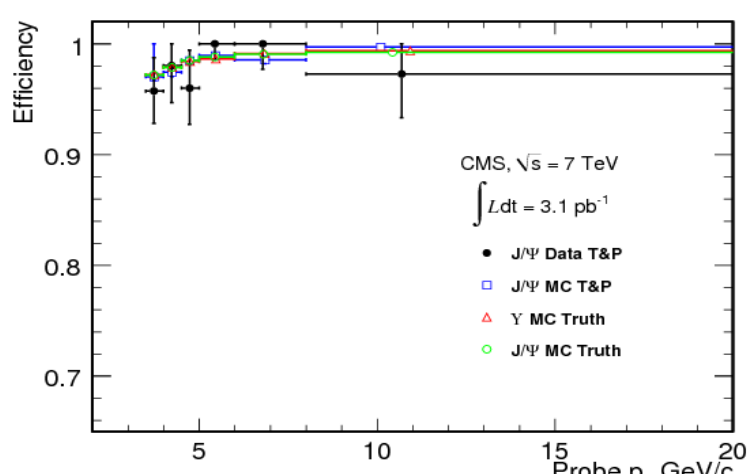
an effective area of interaction
unit: barn, 1b = 10^{-28} m^2 = 100fm^2

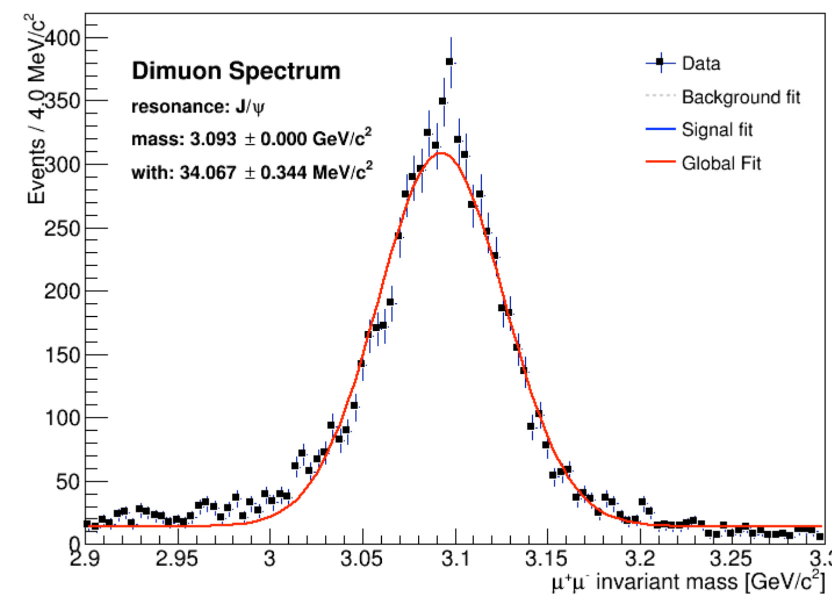
$$\frac{d^2\sigma(Q\overline{Q})}{dp_T\,dy}\mathcal{B}\left(Q\overline{Q}\to\mu^+\mu^-\right)=\frac{N_{fit}(Q\overline{Q})}{\mathcal{L}\cdot\mathcal{A}\cdot\epsilon\cdot\Delta p_T\cdot\Delta y}$$











- N: fitted signal yield
- A: detector acceptance from simulation
- ε: detector reconstruction and trigger efficiencies (simulation or data-driven)
- L: integrated sample luminosity

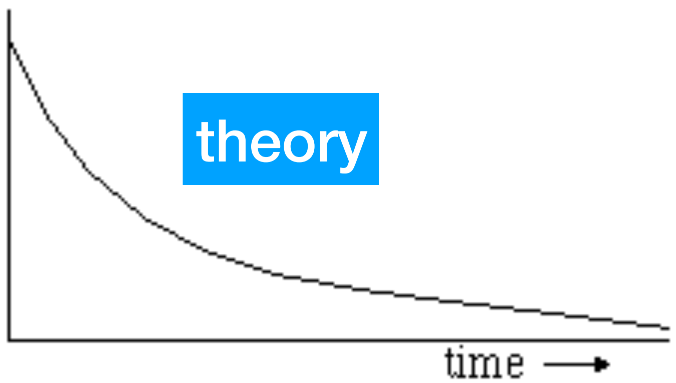
(extra) statistics

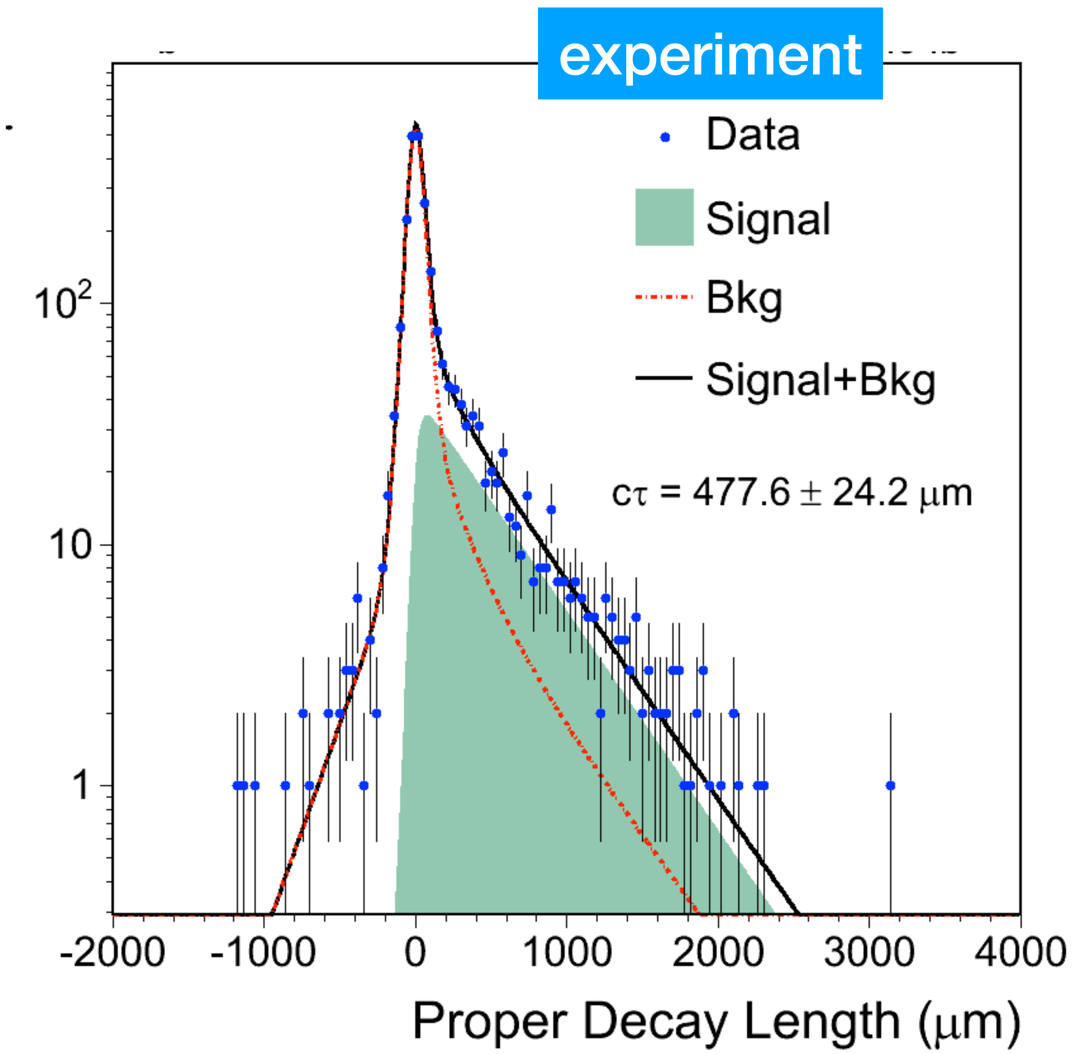
measurement: a lifetime example

signal model →

$$L(t|\sigma_t,\tau) = \frac{1}{\mathcal{N}} \cdot \left[\frac{1}{\tau} e^{-\frac{t}{\tau}} \theta(t) \otimes G(t;\sigma_t) \right] \cdot \mathcal{E}(t) \quad + \; L(\textbf{Background})$$

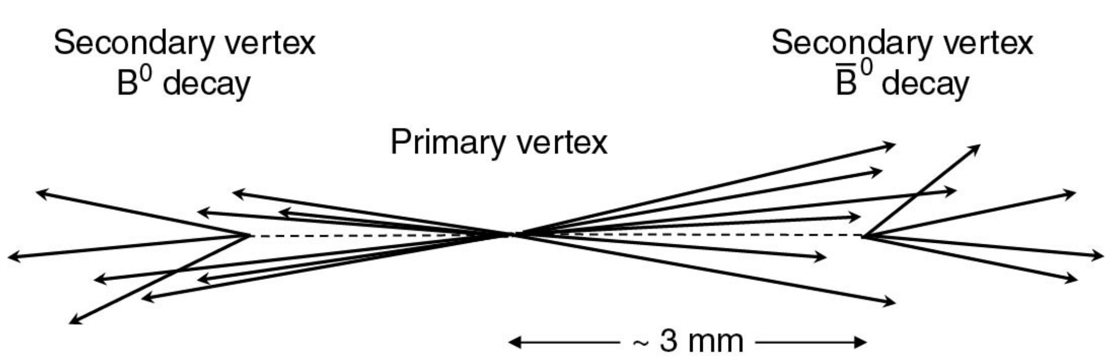
- $\frac{1}{\mathcal{N}}$ ↑ PDF normalization
- $\frac{1}{\tau} e^{-\frac{t}{\tau}} \theta(t)$ ↑ theory model
- $G(t;\sigma_t)$ ↑ t-resolution function
- $\mathcal{E}(t)$ ↑ t-acceptance function





$$t = \frac{L}{\beta\gamma} = L\frac{M}{p} = L_{xy}\frac{M}{p_T}$$

$\beta\gamma$ → Lorentz boost factor



statistics

- consider lifetime (or decay rate) distribution $$f(t;\tau) = \frac{1}{\tau} e^{-t/\tau}$$
- suppose we have n data points (measurements) $t_1, \ldots, t_n$
- likelihood function $$L(\tau) = \prod_{i=1}^{n} \frac{1}{\tau} e^{-t_i/\tau}$$

  maximum likelihood method

- value of τ for which L(τ) is maximum = maximizes log-likelihood

$$\ln L(\tau) = \sum_{i=1}^{n} \ln f(t_i;\tau) = \sum_{i=1}^{n} \left(\ln \frac{1}{\tau} - \frac{t_i}{\tau} \right)$$

- can find its maximum as $\dfrac{\partial \ln L(\tau)}{\partial \tau} = 0 \;\rightarrow\; \hat{\tau} = \dfrac{1}{n} \sum_{i=1}^{n} t_i$
- mean: $E[t] = \int_0^\infty t \, \frac{1}{\tau} e^{-t/\tau} \, dt = \tau$ variance: $V[t] = \int_0^\infty (t-\tau)^2 \frac{1}{\tau} e^{-t/\tau} \, dt = \tau^2$
- for ML estimator $$E[\hat{\tau}] = E\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[t_i] = \tau \;\longrightarrow\; b = E[\hat{\tau}] - \tau = 0$$

$$V[\hat{\tau}] = V\left[\frac{1}{n}\sum_{i=1}^{n} t_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[t_i] = \frac{\tau^2}{n} \;\longrightarrow\; \sigma_{\hat{\tau}} = \frac{\tau}{\sqrt{n}}$$

maximum likelihood

- expand ln L (θ) about its maximum

$$\ln L(\theta) = \ln L(\hat{\theta}) + \left[\frac{\partial \ln L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + ..$$

- first term is ln L_{max} , second term is zero, third term approximate

$$\ln L(\theta) \approx \ln L_{max} - \frac{(\theta - \hat{\theta})^2}{2\widehat{\sigma^2}_{\hat{\theta}}}$$

$$\ln L(\hat{\theta} \pm \hat{\sigma}_{\hat{\theta}}) \approx \ln L_{max} - \frac{1}{2}$$

from information inequality:

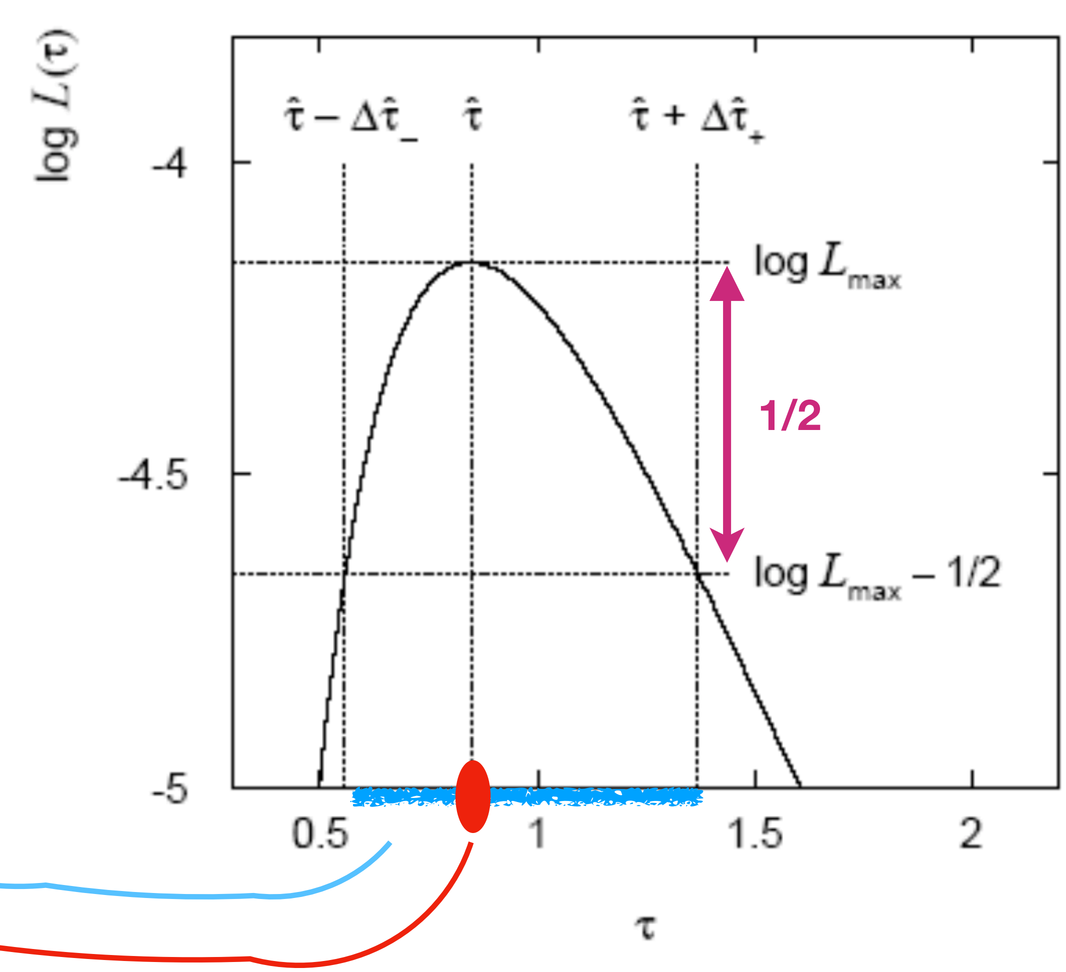
$$\hat{V}[\hat{\theta}] = -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$$

- error estimate: change θ away from $\hat{\theta}$ until ln L decreases by 1/2

- 1σ (68.3% CL) confidence interval

$$\ln L(\theta) = \ln L(\hat{\theta}) - \frac{1}{2} \quad \Rightarrow \quad [\hat{\theta} - \sigma_{\hat{\theta}}, \hat{\theta} + \sigma_{\hat{\theta}}]$$

$$\hat{\tau} = 0.85^{+0.52}_{-0.30}$$

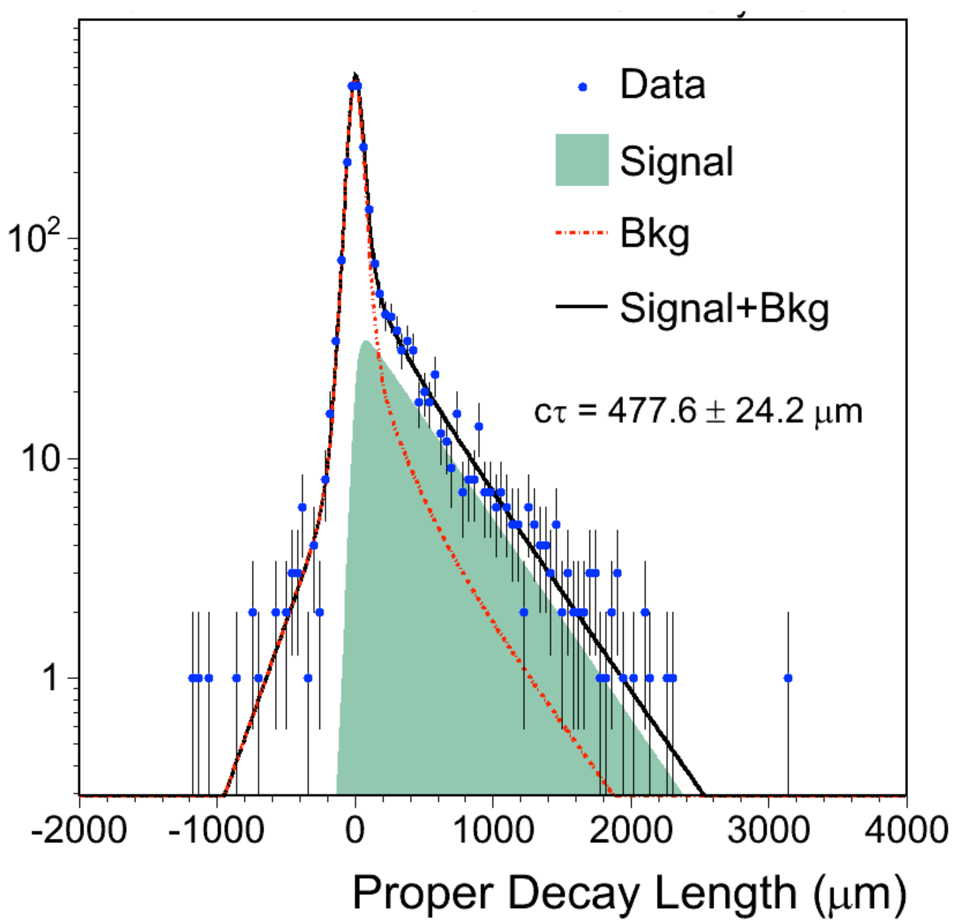


systematic uncertainties

- statistical uncertainty on τ is obtained from the maximum likelihood fit to the data

$$L(t|\sigma_t,\tau) = \frac{1}{\mathcal{N}} \cdot [\ \frac{1}{\tau} e^{-\frac{t}{\tau}} \theta(t) \otimes G(t;\sigma_t)\ ] \cdot \mathcal{E}(t) + L(\text{Background})$$

- systematic uncertainty quantifies any uncertainty in the procedure going from the raw data to a published result



Sources of systematic error:

- resolution calibration σ_t
- resolution function, G(t)
- t-efficiency $\epsilon(t)$
- background model
- …

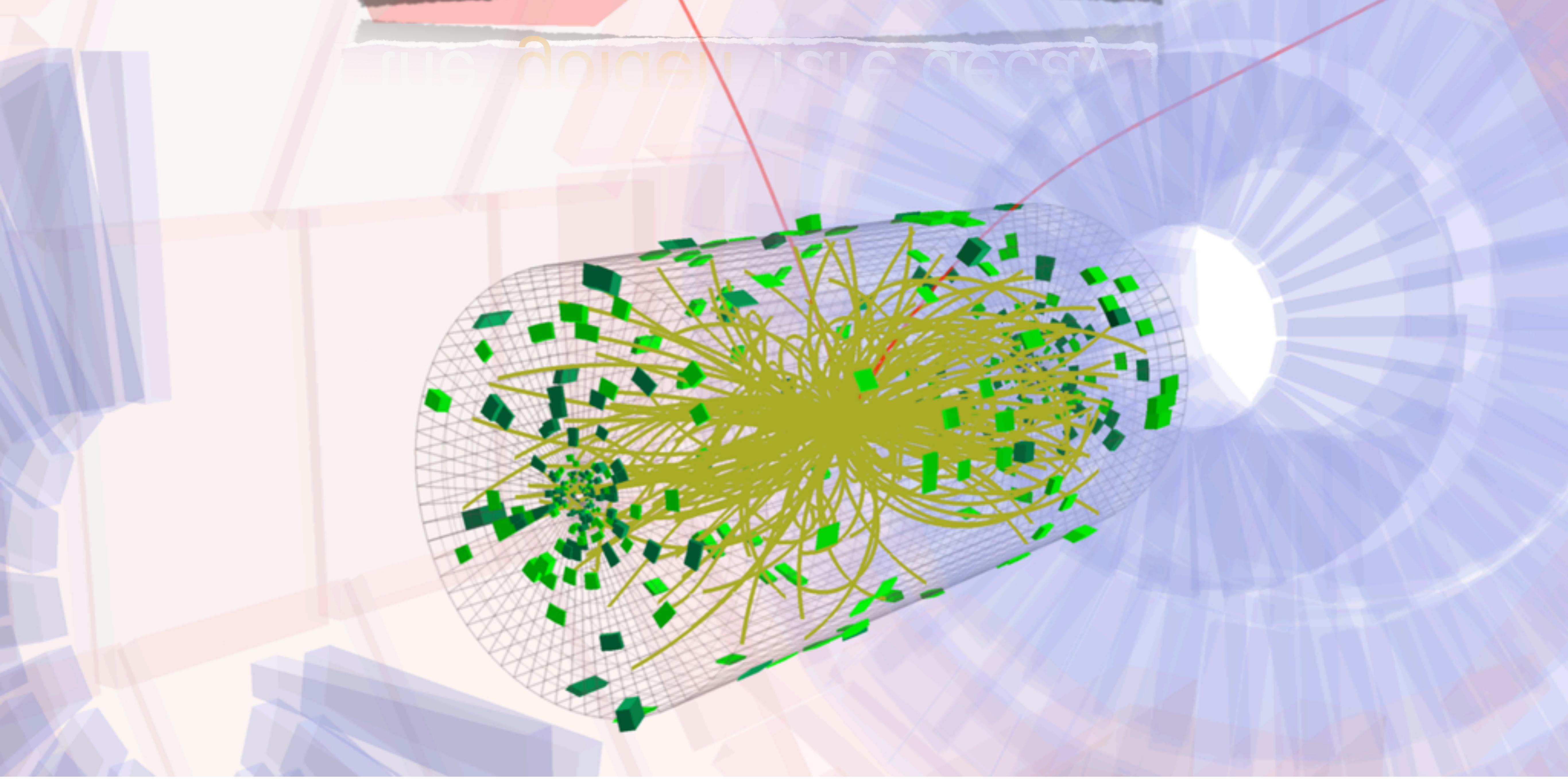
Inputs from Simulation and Data-driven

$$c\tau = 477.6 \pm 24.2\,(stat.) \pm 17.6\,(syst.)\ \mu m$$

(extra) ingredients of a physics measurement

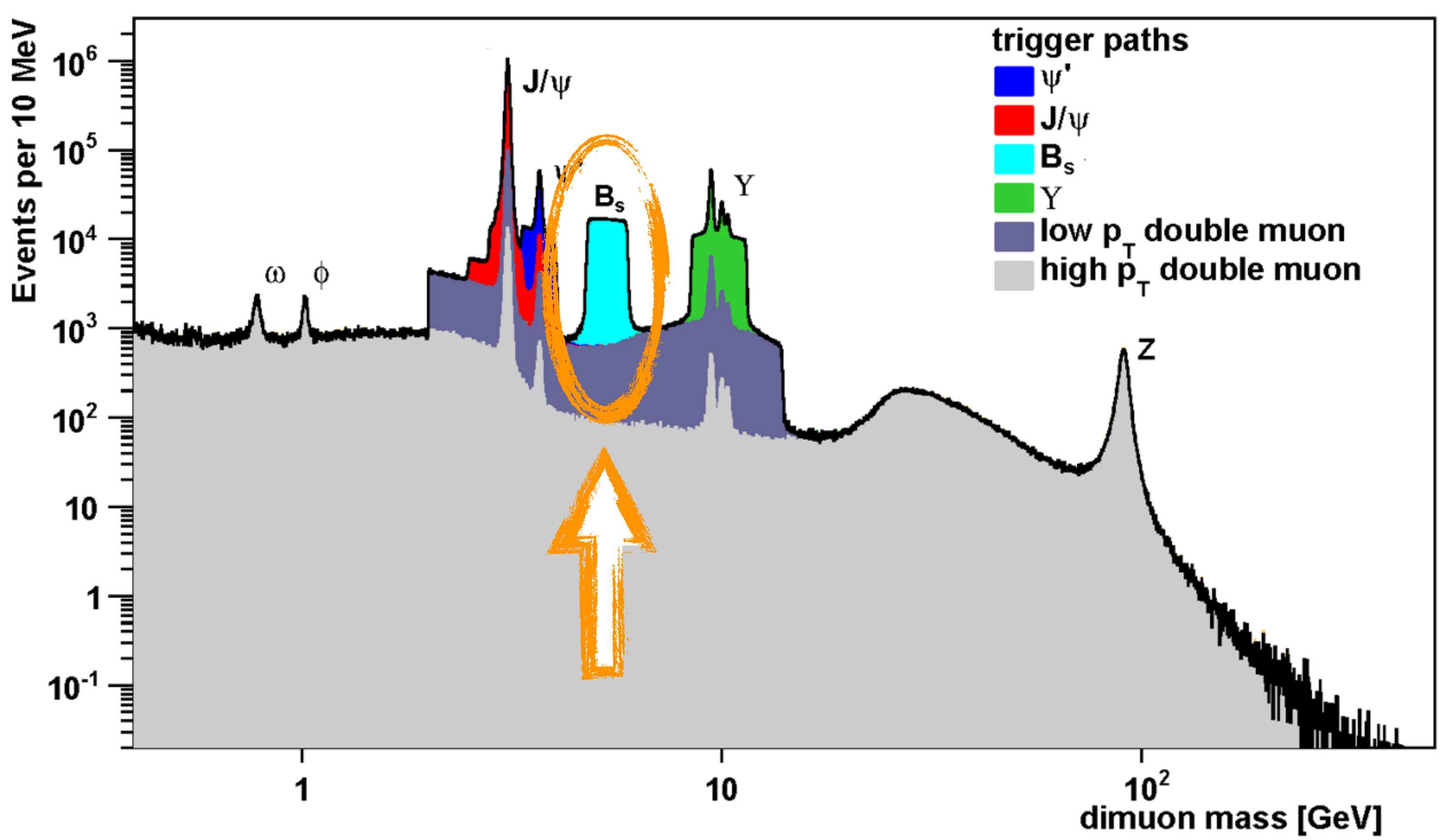
$B_s \to \mu\mu$

the ‘golden’ rare decay



searching for an *ultra-rare* decay: B→μμ

1. ONLINE SELECTION (TRIGGER)



Dimuon Trigger

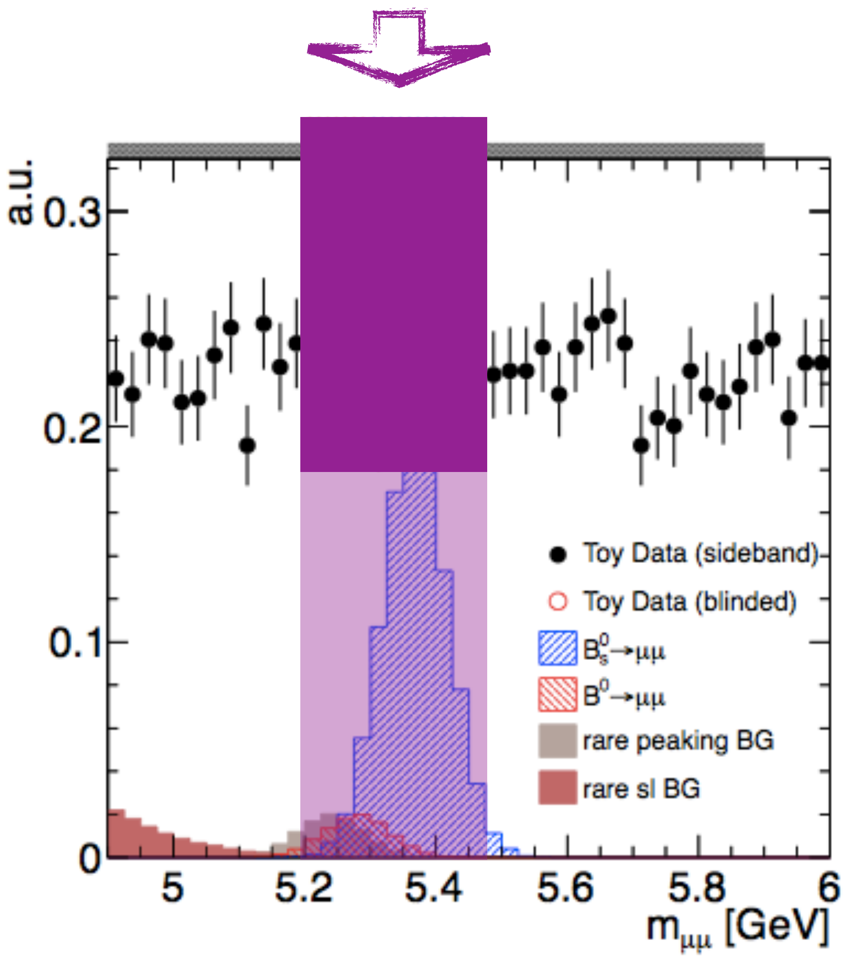
- L1 Hardware Trigger
  - p_T>3 GeV (few kHz)
- HLT Full tracking and vertexing
- HLT B_s→μμ
  - Leading and sub-leading μ p_T>3,4 (4,4) GeV $|\eta_{\mu\mu}|$<1.8 (1.8<$|\eta_{\mu\mu}|$<2.2)
  - p_T (μμ)>5 (4.8-6) GeV
  - 4.8 <m(μμ)< 6.0 GeV
  - $P(\chi^2/dof)$ >0.5%

searching for an ultra-rare decay: B→μμ

1. ONLINE SELECTION (TRIGGER)

2. BLIND THE DATA (AVOID BIAS)

SIGNAL
REGION
BLINDED

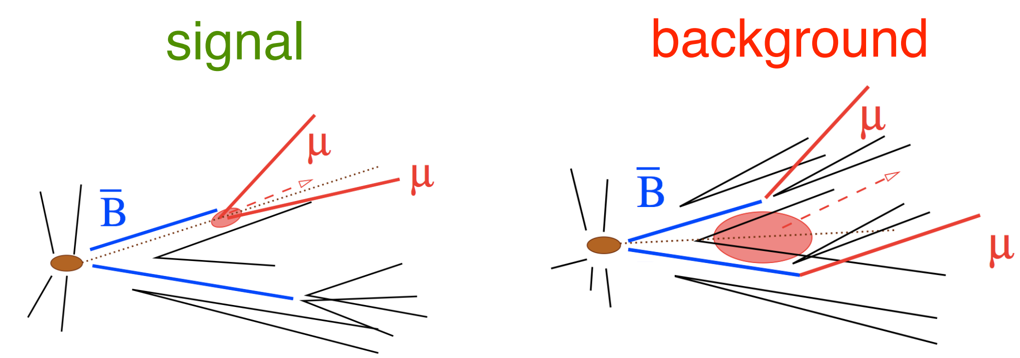


analysis procedure and event selection developed without inspecting the data in region where signal is expected

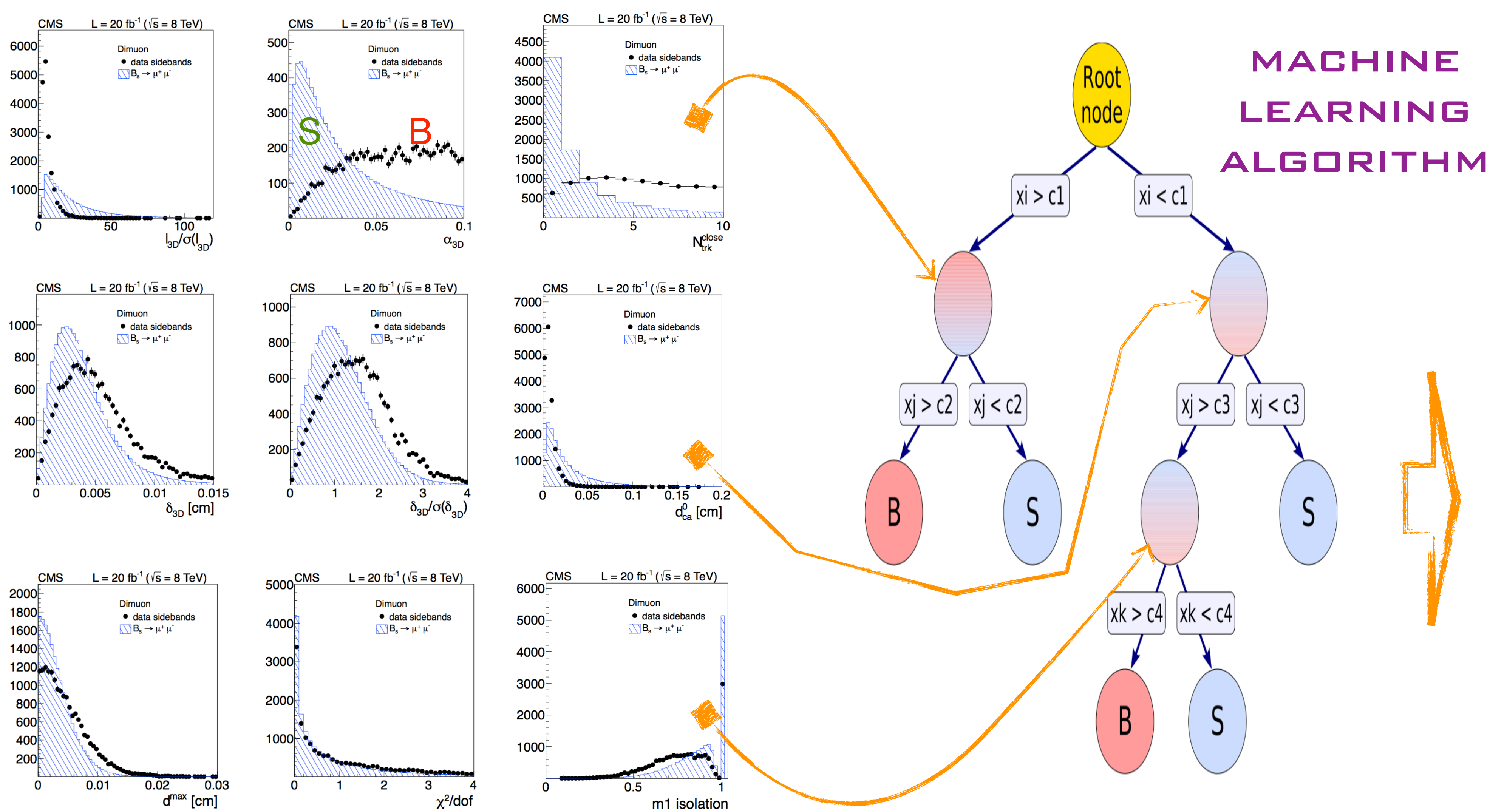
"box opening" only later, at final analysis stages

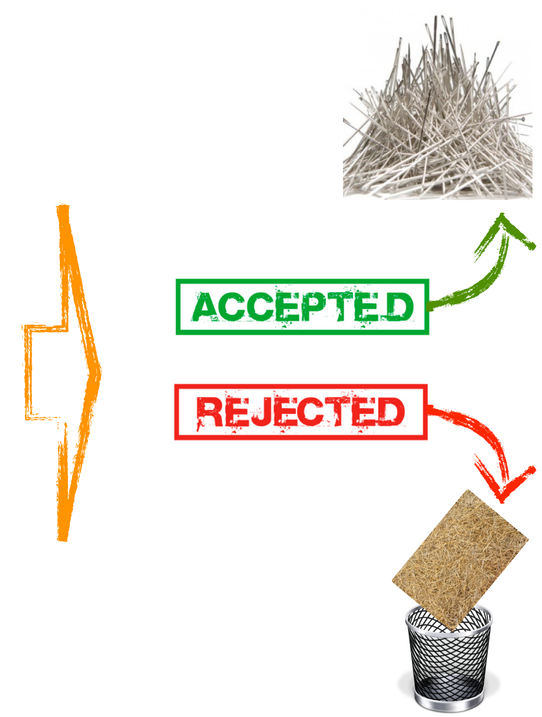
searching for an *ultra-rare* decay: B→μμ

1. ONLINE SELECTION (TRIGGER)
2. BLIND THE DATA (AVOID BIAS)
3. MULTIVARIATE SELECTION



signal vs background discriminating variables

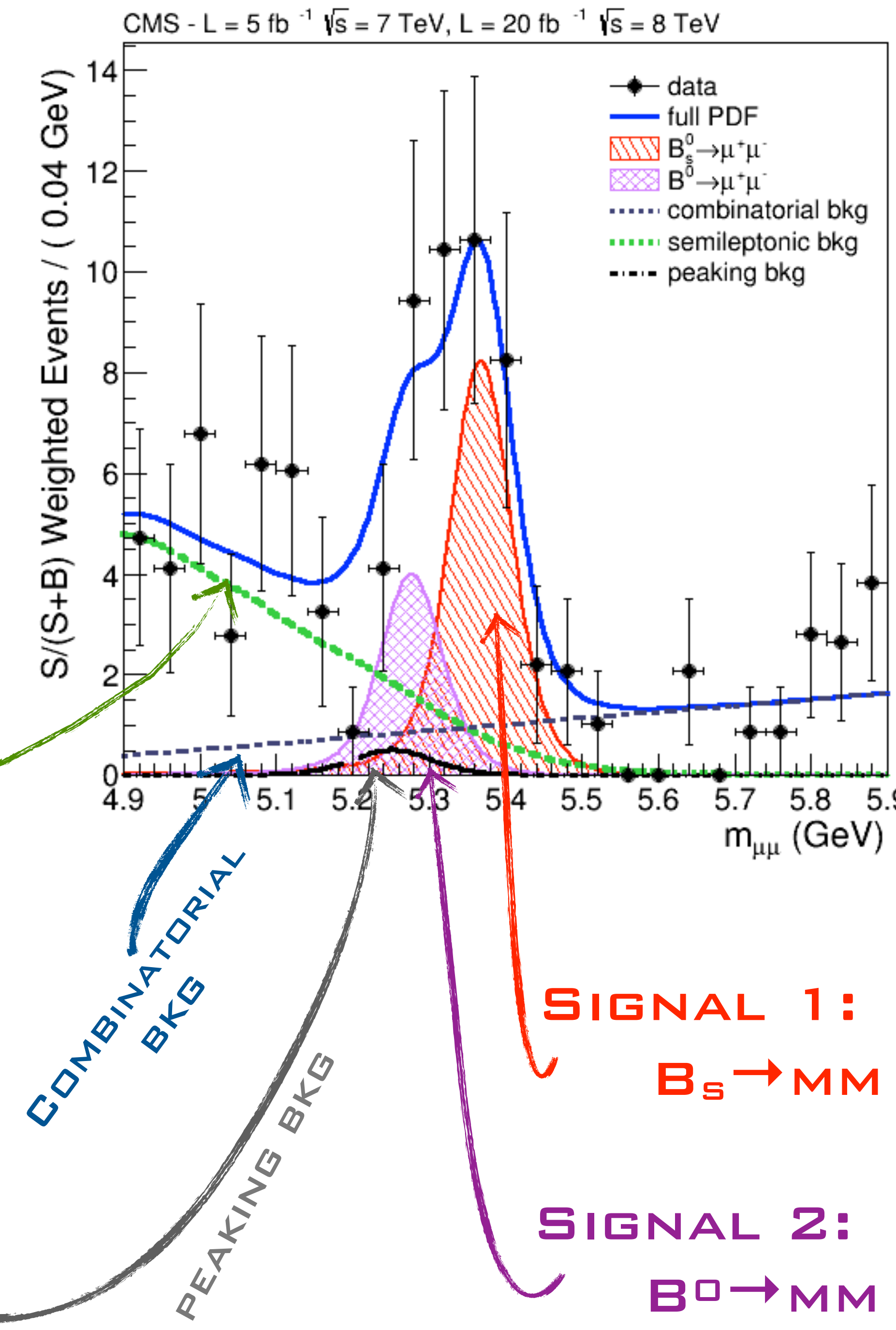


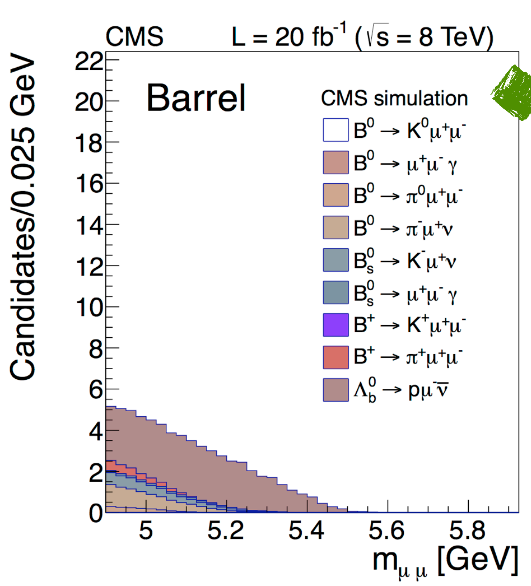


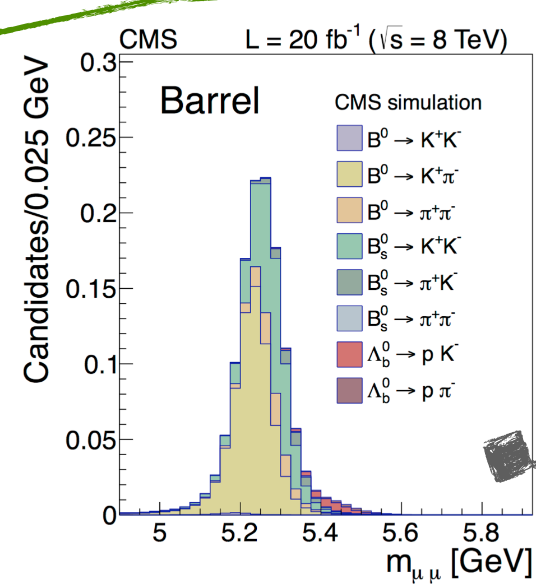
searching for an *ultra-rare* decay: $B \to \mu\mu$

1. ONLINE SELECTION (TRIGGER)
2. BLIND THE DATA (AVOID BIAS)
3. MULTIVARIATE SELECTION
4. **FIT THE DATA (LIKELIHOOD)**

Fit the data accounting for the various signal and background components

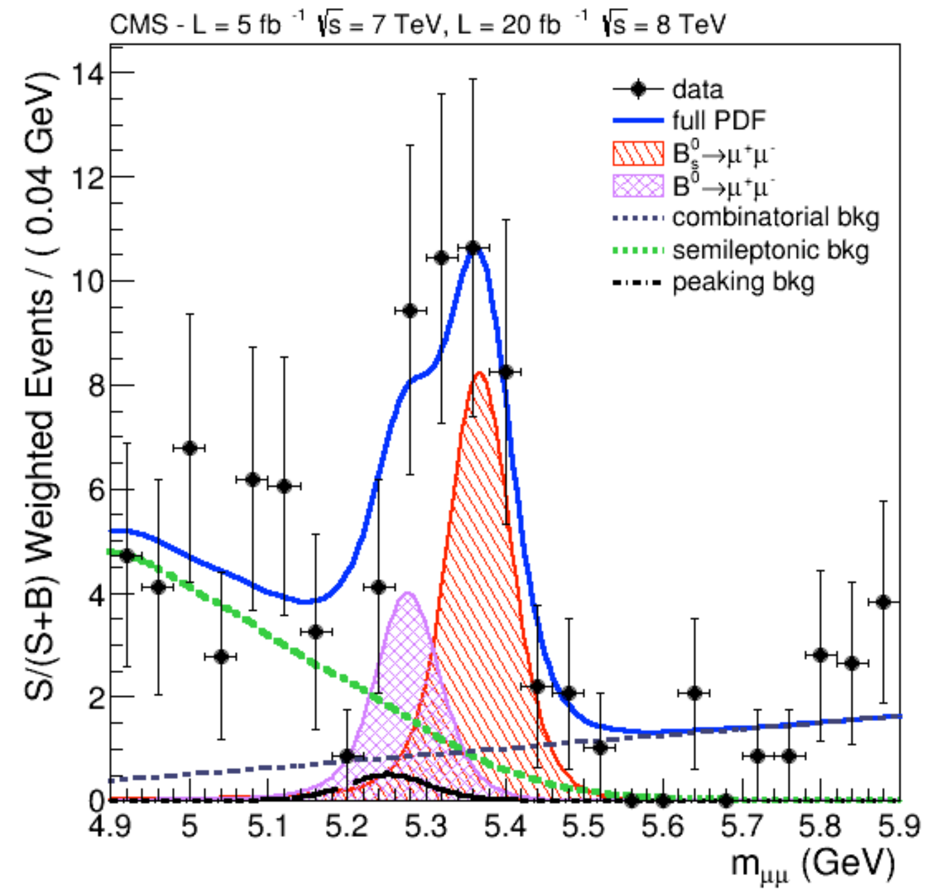




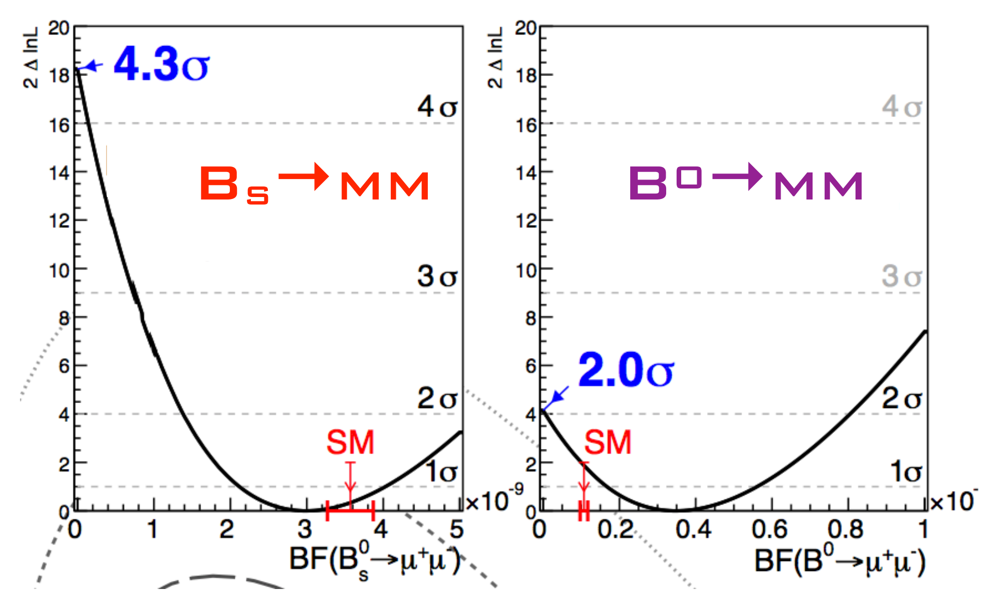


searching for an *ultra-rare* decay: B→μμ

1. ONLINE SELECTION (TRIGGER)
2. BLIND THE DATA (AVOID BIAS)
3. MULTIVARIATE SELECTION
4. FIT THE DATA (LIKELIHOOD)
5. **STATISTICAL SIGNIFICANCE**

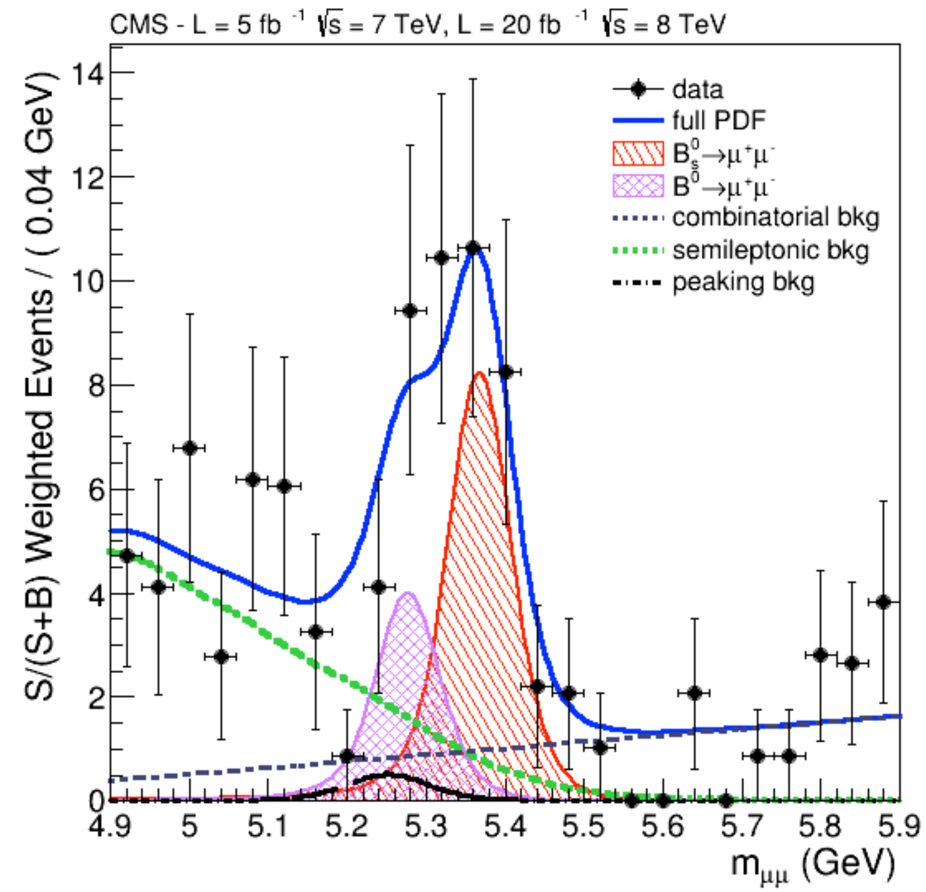


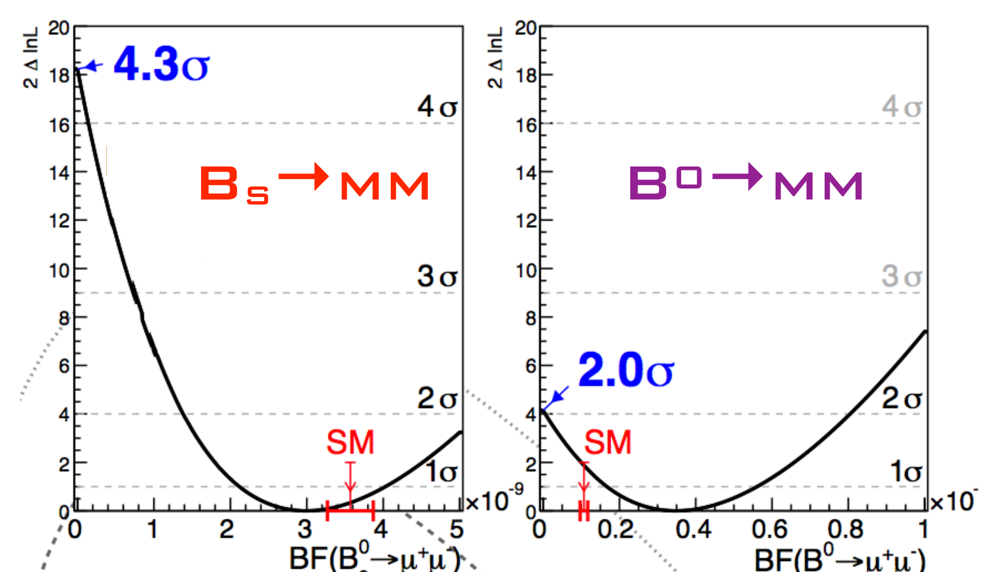
is the observed excess a genuine signal,
or just a fluctuation of the background?



searching for an ultra-rare decay: B→μμ

1. ONLINE SELECTION (TRIGGER)
2. BLIND THE DATA (AVOID BIAS)
3. MULTIVARIATE SELECTION
4. FIT THE DATA (LIKELIHOOD)
5. STATISTICAL SIGNIFICANCE
6. EXTRACT MEASUREMENT

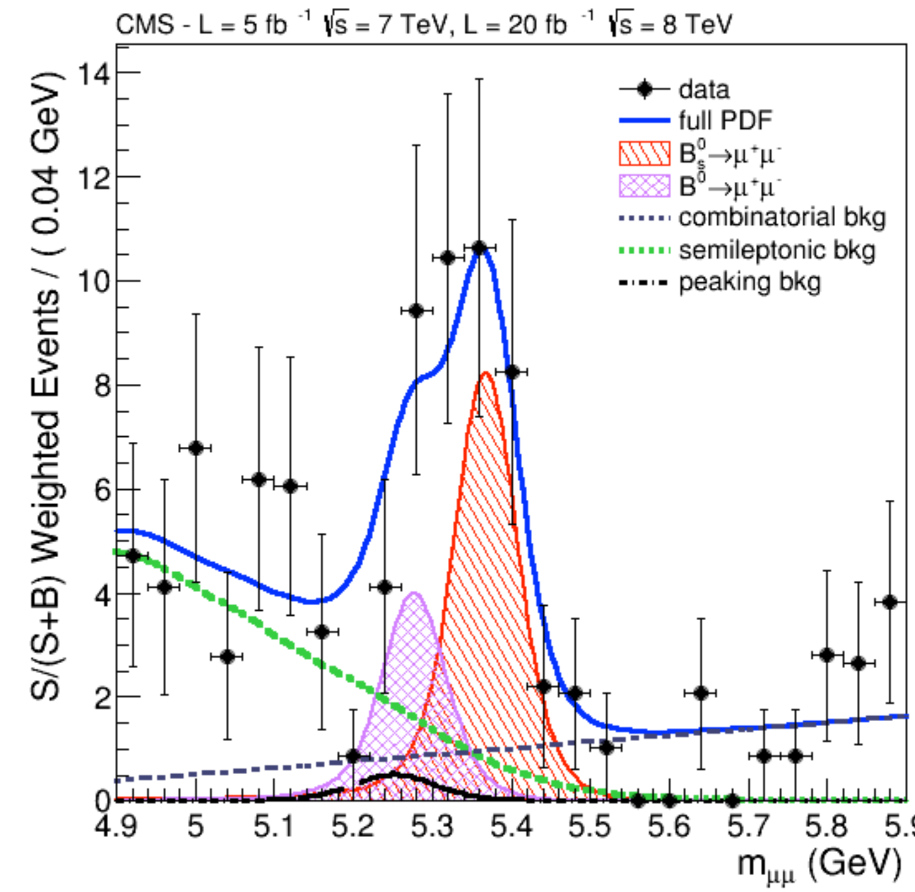




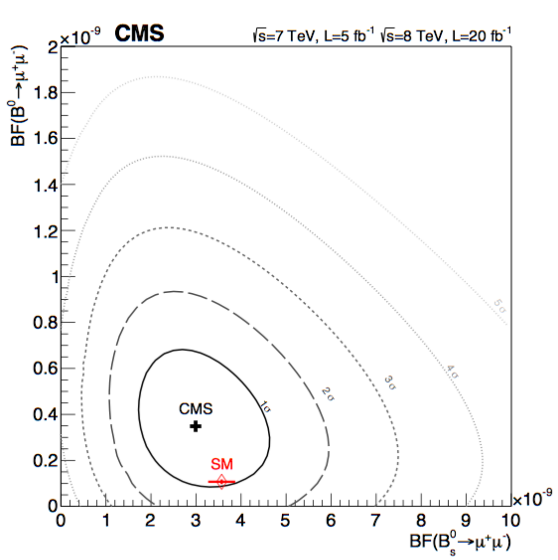
$$BR(B_s \to \mu\mu) = \left(3.0^{+0.9}_{-0.8}\ (\text{stat})^{+0.6}_{-0.4}\ (\text{syst})\right) \times 10^{-9}$$

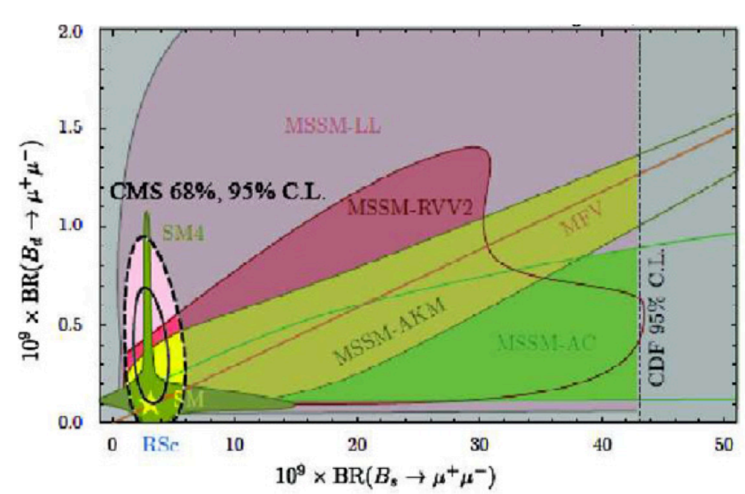
searching for an *ultra-rare* decay: B→μμ

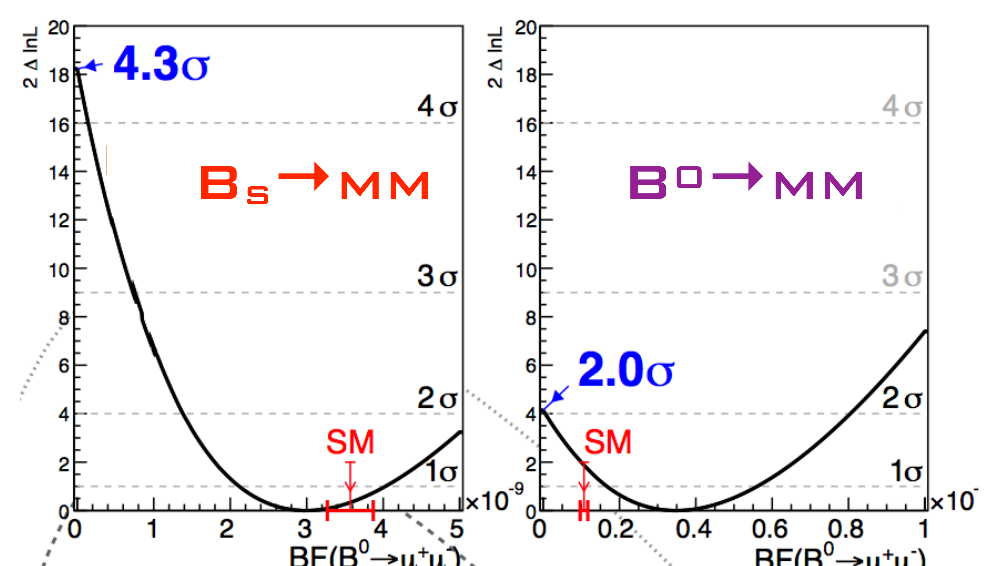
1. Online selection (trigger)
2. Blind the data (avoid bias)
3. Multivariate selection
4. Fit the data (likelihood)
5. Statistical significance
6. Extract measurement
7. **Compare to theory**



the SM... and Beyond







$$\mathrm{BR}(B_s \to \mu\mu) = \left(3.0^{+0.9}_{-0.8}\ (\mathrm{stat})^{+0.6}_{-0.4}\ (\mathrm{syst})\right) \times 10^{-9}$$