Karlsruher Institut für Technologie

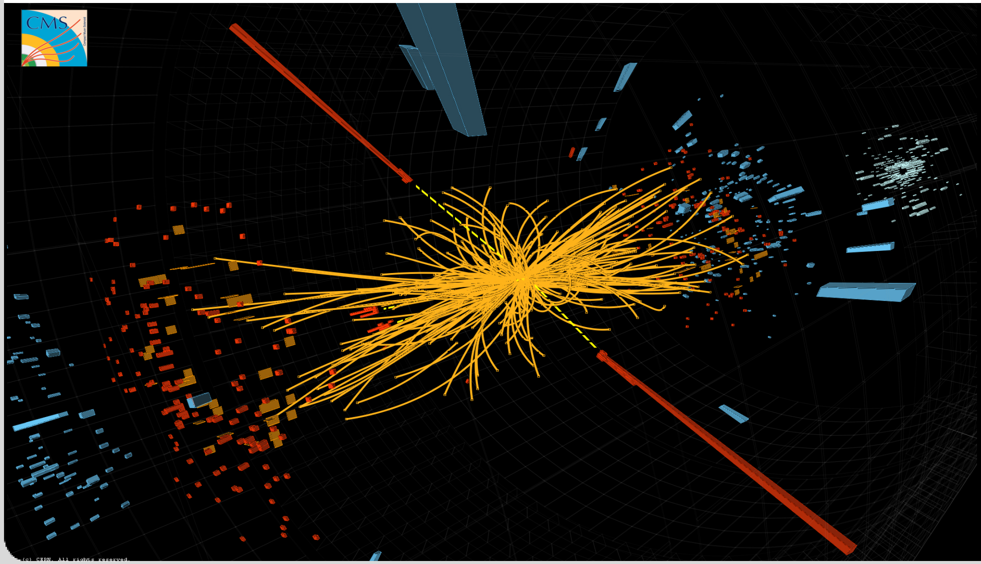
Higgs Boson Physics

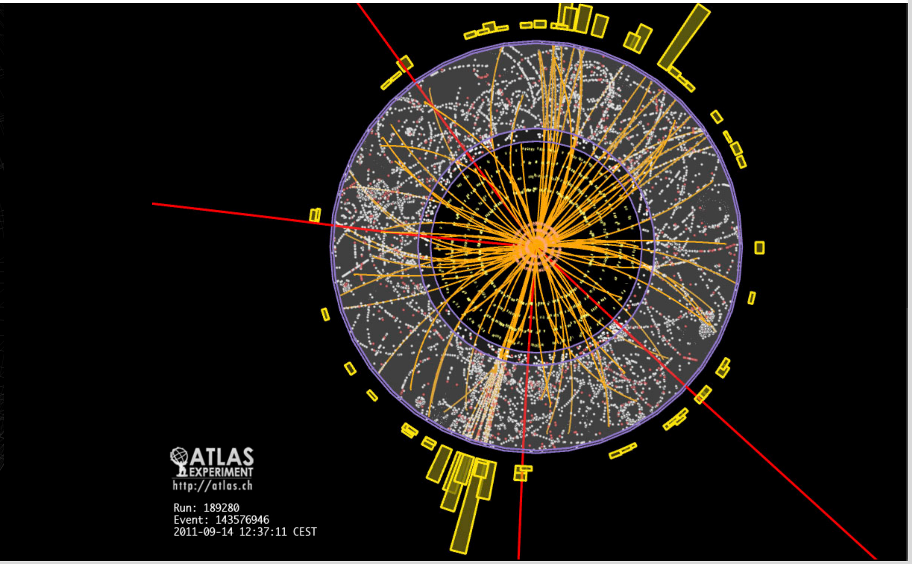
Analysis Techniques

Günter Quast, Roger Wolf, Andrew Gilbert

Master-Kurs
SS 2015

Institut für Experimentelle Kernphysik

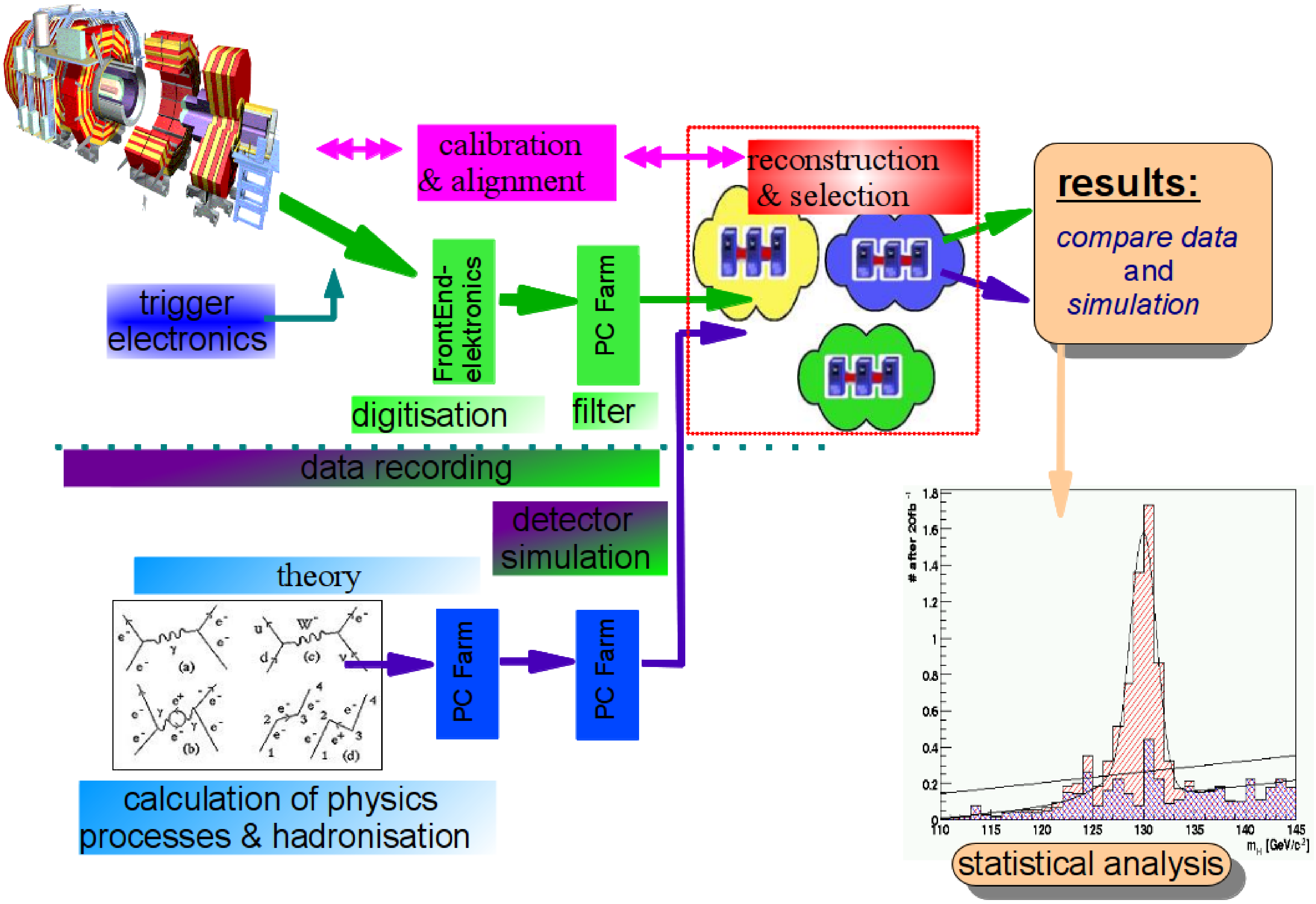




KIT – Universität des Landes Baden-Württemberg und
nationales Forschungszentrum in der Helmholtz-Gemeinschaft

www.kit.edu

Recap: Simulation and Analysis Chain



calibration & alignment

reconstruction & selection

results:

compare data and simulation

trigger electronics

FrontEnd-elektronics

PC Farm

digitisation

filter

data recording

detector simulation

theory

PC Farm

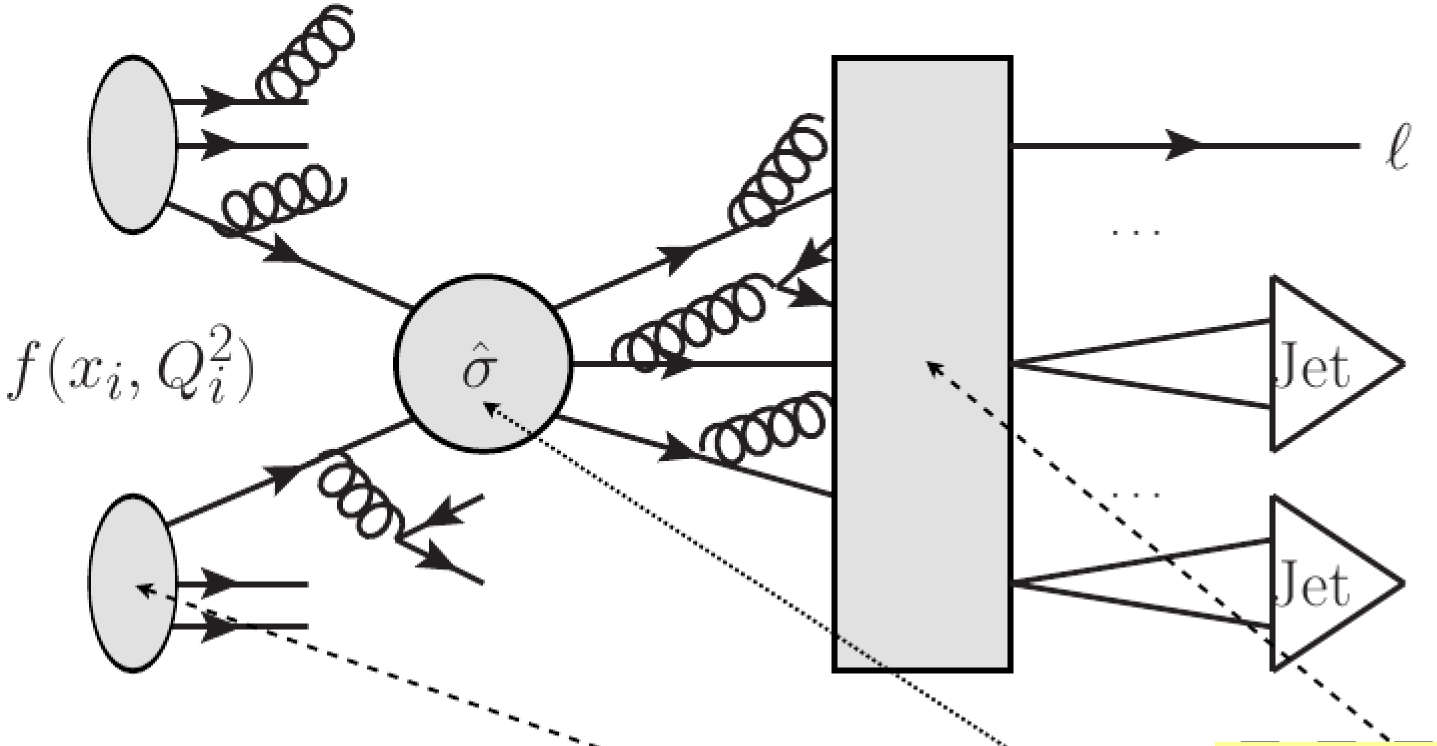
PC Farm

calculation of physics processes & hadronisation

statistical analysis

Recap: Event Simulation

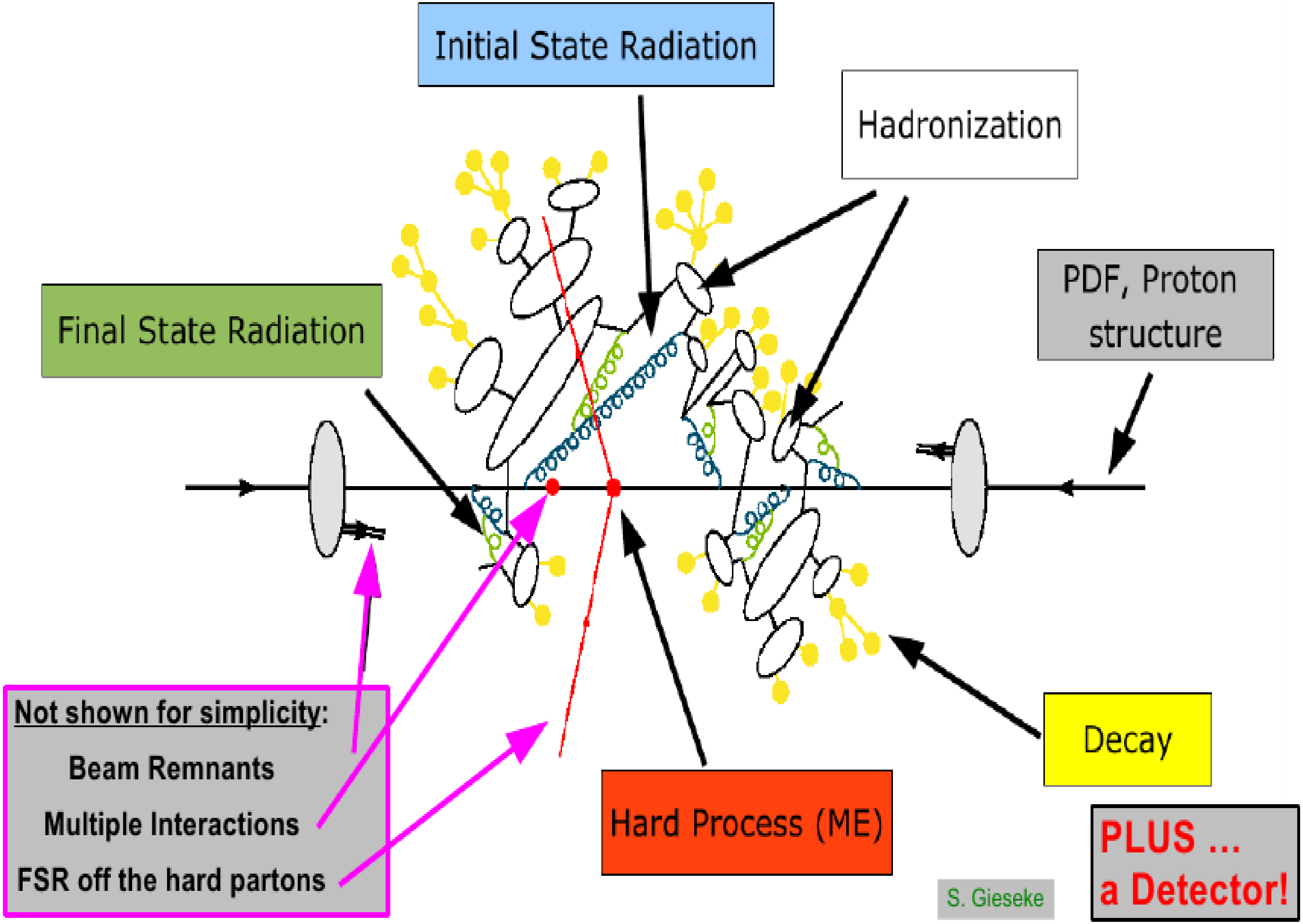
σ = PDFs ⊗ 2→n process ⊗ hadroniszation



$$\sigma_{\mathrm{QCD}} = \sum_{jk} \int \mathrm{d}x_j\, \mathrm{d}x_k\, f_j(x_j, \mu_F^2)\, f_k(x_k, \mu_F^2) \cdot \hat{\sigma}(x_j x_k s, \mu_F^2, \mu_R^2) \otimes \text{hadronization}$$

Complicated process – use MC techniques to calculate cross sections, phenomenological modes to describe hadronization process (quarks → jets)

Summary: pp collision



Initial State Radiation

Hadronization

Final State Radiation

PDF, Proton structure

Decay

Hard Process (ME)

Not shown for simplicity:

Beam Remnants

Multiple Interactions

FSR off the hard partons

PLUS … a Detector!

S. Gieseke

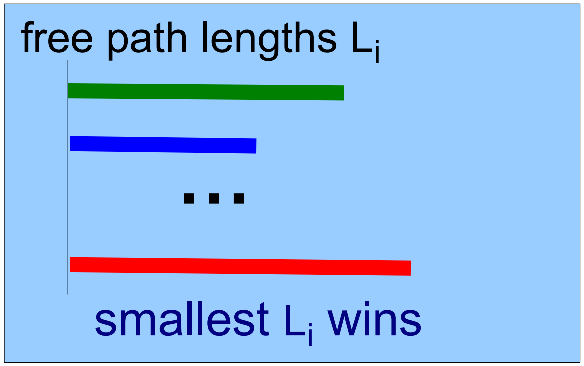
Recap: Detector Simulation

- Generate interaction points along a particle path according to distribution of path length in matter until next interaction (free path length):

$$w(L) = \rho_n \sigma \exp\left(-\rho_n \sigma L\right) = \frac{1}{\lambda} \exp\left(-L/\lambda\right)$$

$\lambda = (\rho_n\, \sigma)^{-1}$: **interaction length**

- in case of many competing processes, the one with the smallest free path length is selected to occur


free path lengths L_i

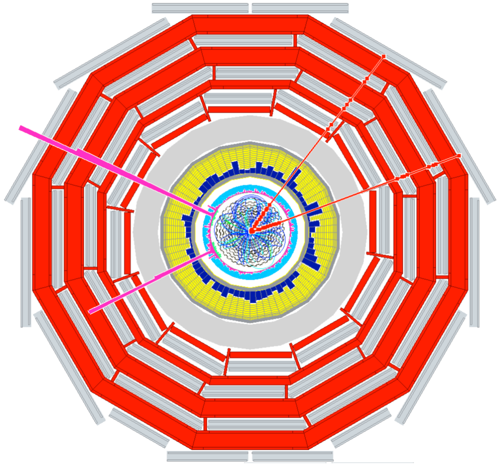
...

smallest L_i wins

- follow each particle, including newly produced daughter particles, until energy is below a cut-off threshold

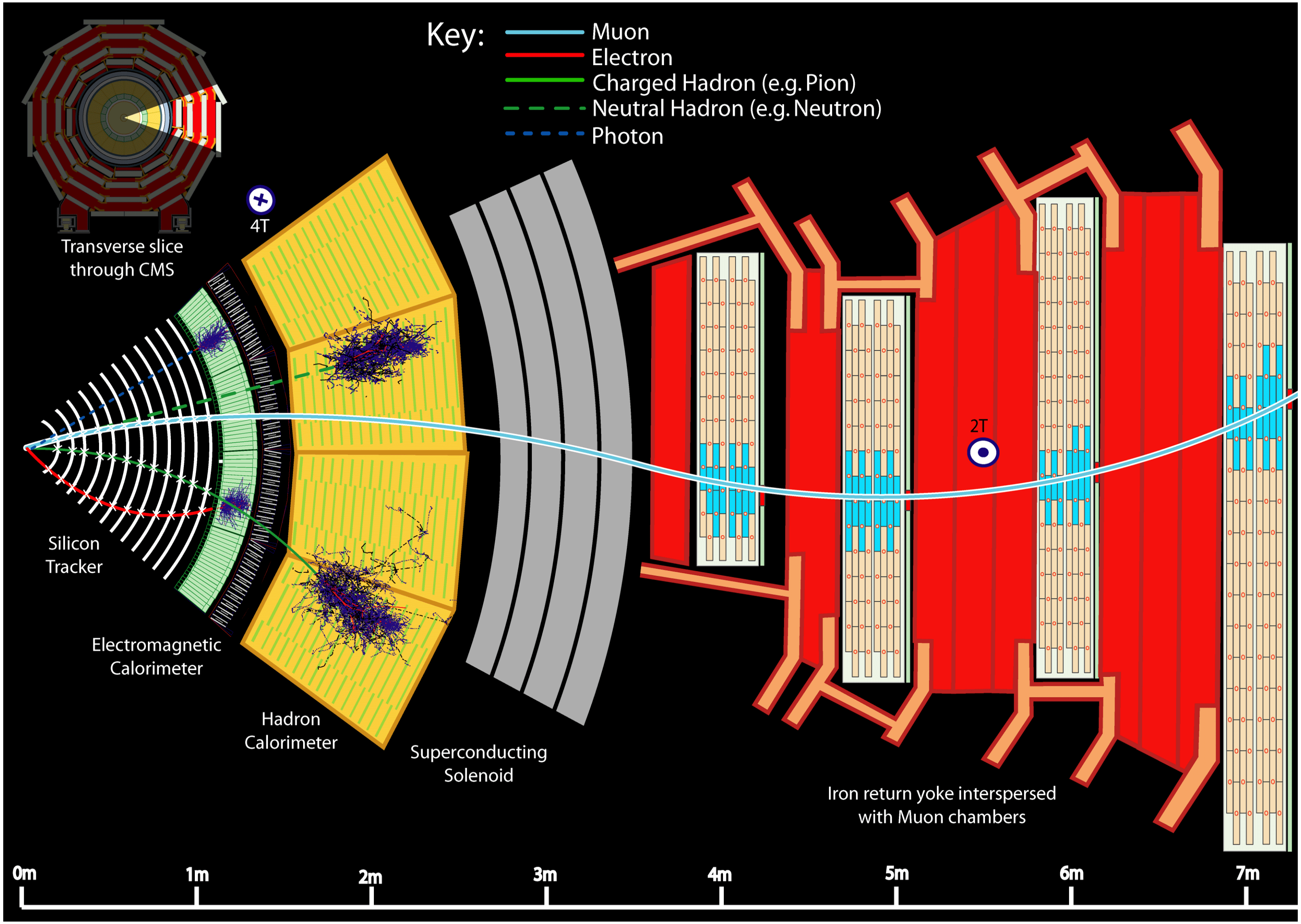
- calculate deposited energy in detector cells

- simulate observable signal (free charges or light)



The real experiment and data analysis

Particle reconstruction



Detector registers only „stable particles“,
i.e. with life times long enough to traverse the detector

7 stable particles:
γ, e, μ , p, n, $\pi^{\pm}$, $K^{\pm}$

Steps of Event selection

- **hardware Trigger** and **on-line selection** identify „interesting“ events with particles in the sensitive area of the detector (events not selected are lost)

  ⟶ detector acceptance and online-selection efficiency

- physics objects are **reconstructed** off-line

  ⟶ reconstruction efficiency

- **Analysis** procedure identifies physics processes and rejects backgrounds

  ⟶ selection efficiency and purity

- **statistical inference** to determine confidence intervals of interesting parameters (production cross sections, particle properties, model parameters, ...)

All steps are affected by systematic errors !

Cross section measurement

Master formula:

signal candidates

background events

$$\sigma = \frac{N_{cand} - N_{bkg}}{\epsilon \cdot \int L}$$

acceptance and efficiency

integrated Luminosity

product of many factors:
- – detector acceptance
- – trigger efficiency
- – reconstruction efficiency
- – selection efficiency
- – background rejection efficiency

usually determined using well-known reference reaction:

$$\int L = N_{ref} \,/\, \sigma_{ref}$$

Cross Section measurement: errors

by error propagation →

$$\frac{\delta\sigma}{\sigma} = \sqrt{\frac{\delta N_{cand}^2 + \delta N_{bkg}^2}{(N_{cand} - N_{bkg})^2} + \left(\frac{\delta\epsilon}{\epsilon}\right)^2 + \left(\frac{\delta\int L}{\int L}\right)^2}$$

This is the error you want to <u>minimize</u>

- with signal as large as possible
- background as small as possible
- nonetheless, want large efficiency
- luminosity error small (typically beyond your control, also has a “theoretical” component)

(Integrated) Luminosity

Luminosity, $\mathcal{L}$, connects event rate, r, and cross section, $\boldsymbol{\sigma}$:

$r = \mathcal{L} \cdot \sigma$, unit of [$\mathcal{L}$] = cm^{-2}/s oder 1/nb /s

Integrated luminosity, $\int \mathcal{L}\, dt$, is a measure of the total number of events at given cross section, $N = \int \mathcal{L}\, dt \cdot \sigma$

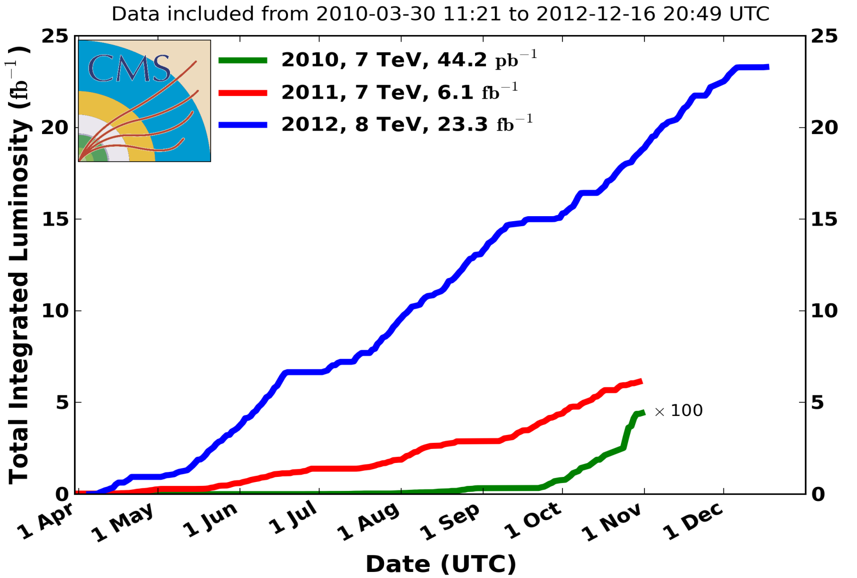
$\mathcal{L}$ is a property of the accelerator:

$$\mathcal{L} = \frac{f_{\mathrm{rev}} n_b N_p{}^2}{4\pi A_{\mathrm{bunch}}} = \frac{f_{\mathrm{rev}} n_b N_p{}^2}{4\pi \epsilon \beta^*}$$

f_{rev}: revolution frequency of beams
n_b: number of bunches
N_p: number of particles in a bunch
A_{bunch}: area of bunches
ε: emittance of beam
β*: beta-function at collision point

LHC design Luminosity: 10^{34} /cm²/s

$\int \mathcal{L}$ recorded by the CMS experiment



The total integrated Luminosity of 29.4 fb^{-1} corresponds to $1.8 \cdot 10^{15}$ pp collisions (assuming 60 mb inelastic pp cross section)

Determination of Luminosity

Luminosity is, however, not determined from machine parameters

(precision only ~10%)

but by simultaneous measurements of a **reference reaction** with well-known cross section:

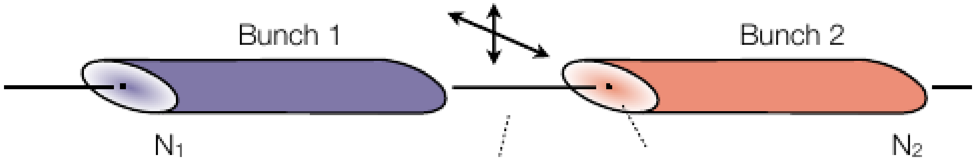
$$\int L = N_{ref} \,/\, \sigma_{ref}$$

absolute value from

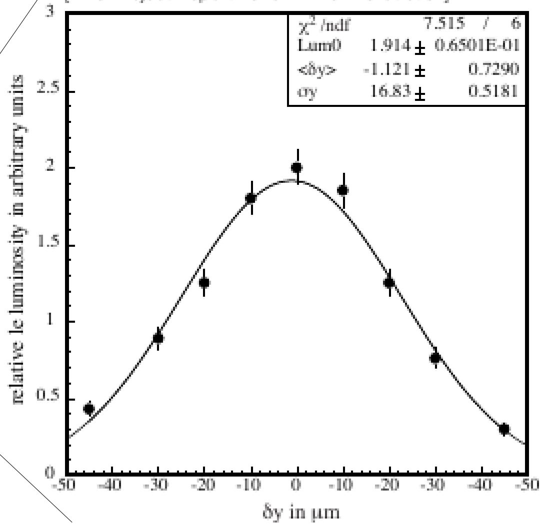
- elastic proton-proton scattering at small angles
- production of W or Z bosons
- production of photon or muon pairs in $\gamma\gamma$-reactions
- ...

measurement of luminous beam profile:

- van-der-Meer scans by transverse displacement of beams, record $\mathcal{L}$ vs. δx, δy



[LHC Project Report 1019: H. Burkhardt et al.]



relative methods:

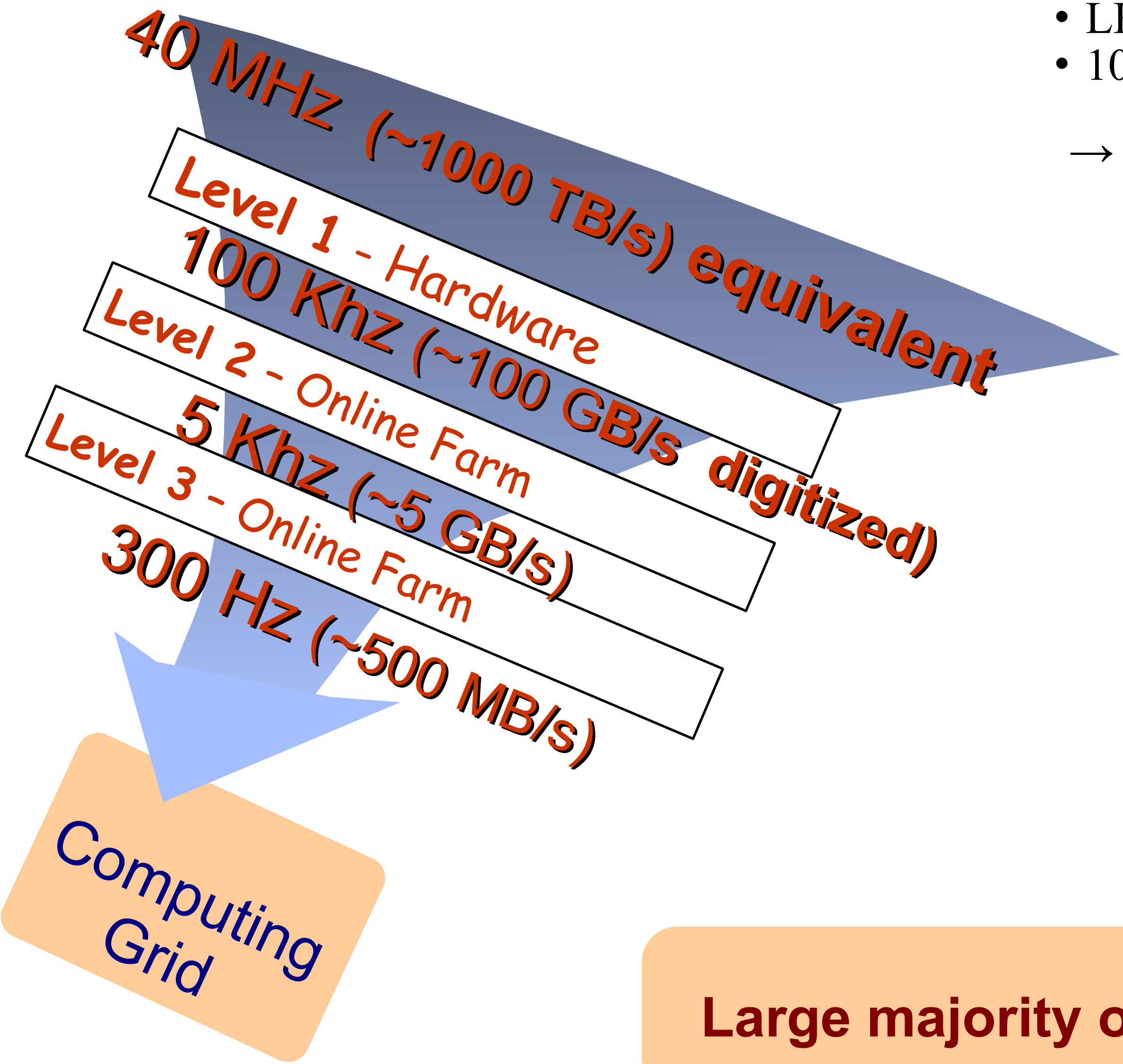
- particle counting or current measurements in detector components with high rates

(need calibration against one of the absolute methods)

accuracy on $\int\mathcal{L}$ (CMS experiment): 2.2% (7 TeV, 2011) and 2.6% (8TeV, 2012)

Trigger

Online Data Reduction



40 MHz (~1000 TB/s) equivalent

Level 1 - Hardware

100 Khz (~100 GB/s digitized)

Level 2 - Online Farm

5 Khz (~5 GB/s)

Level 3 - Online Farm

300 Hz (~500 MB/s)

Computing Grid

- ~ 100 million detector cells
- LHC collision rate: 40 MHz
- 10-12 bit/cell

→ **~1000 Tbyte/s raw data**

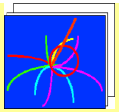
Zero-Suppression & **Trigger** reduce this to „only“ some 100 Mbyte/s

i.e. 1 /sec

Large majority of events is not stored!

CMS Trigger & Data Acquisition

every 25 ns



40 MHz
COLLISION RATE

100 kHz
LEVEL-1 TRIGGER

DAQ accepts
Level-1 Rate of 100kHz

1 Terabit/s
(50000 DATA CHANNELS)

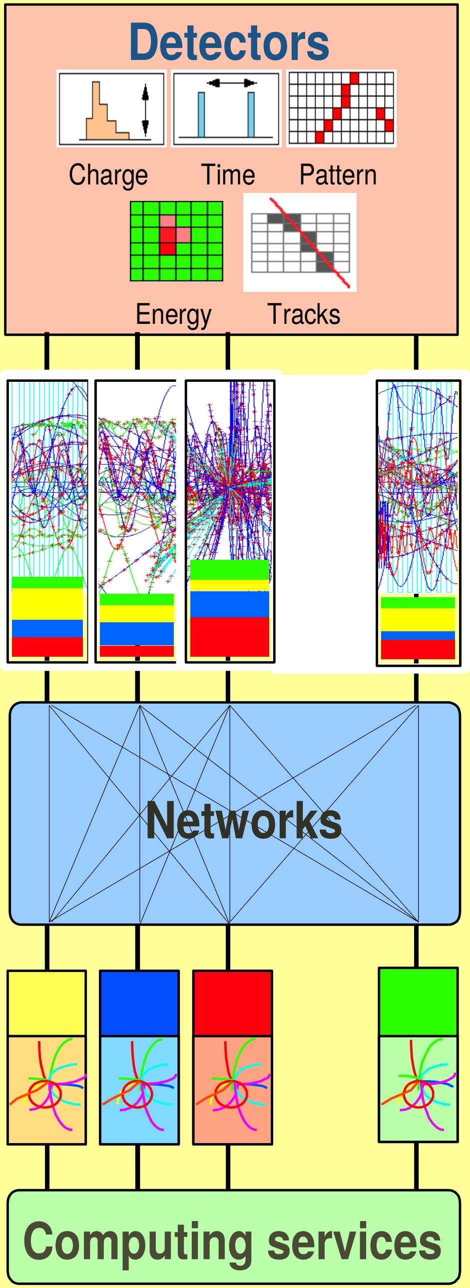
500 Gigabit/s

HLT (High Level Trigger) designed for O(100Hz)

- suppression factor ~1000

~2000 CPUs

Gigabit/s SERVICE LAN



16 Million channels
3 Gigacell buffers

1 Megabyte EVENT DATA

200 Gigabyte BUFFERS
500 Readout memories

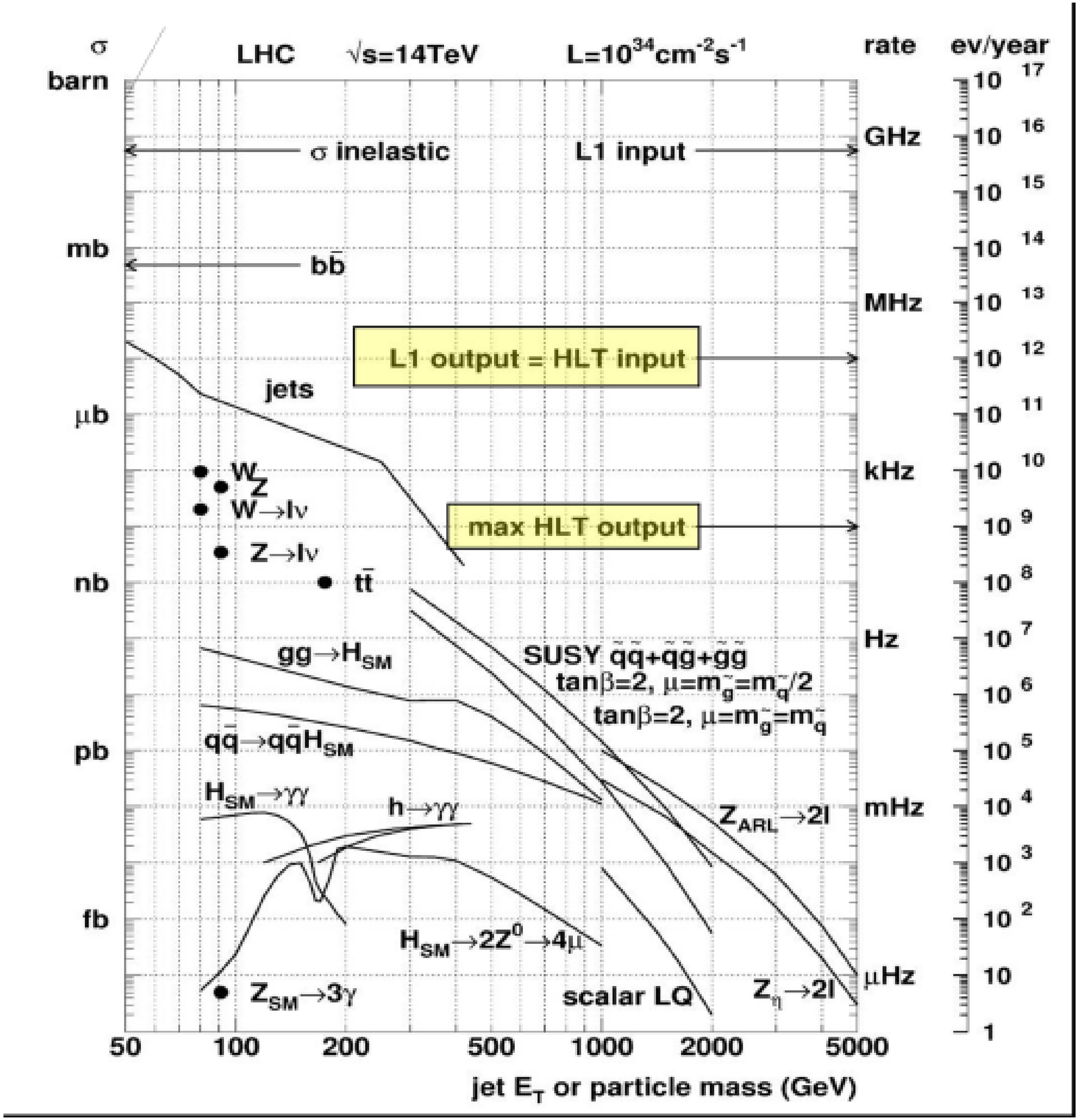
EVENT BUILDER. A large switching network (512+512 ports) with a total throughput of approximately 500 Gbit/s forms the interconnection between the sources (Readout Dual Port Memory) and the destinations (switch to Farm Interface). The Event Manager collects the status and request of event filters and distributes event building commands (read/clear) to RDPMs

5 TeraIPS

EVENT FILTER. It consists of a set of high performance commercial processors organized into many farms convenient for on-line and off-line applications. The farm architecture is such that a single CPU processes one event

Petabyte ARCHIVE

Trigger Rate vs. Cross section



Much of the
“interesting physics”
limited by maximum
possible trigger rate !

What is easy to trigger ?

Trigger thresholds rise as luminosity goes up,
and are a topic of permanent debate !

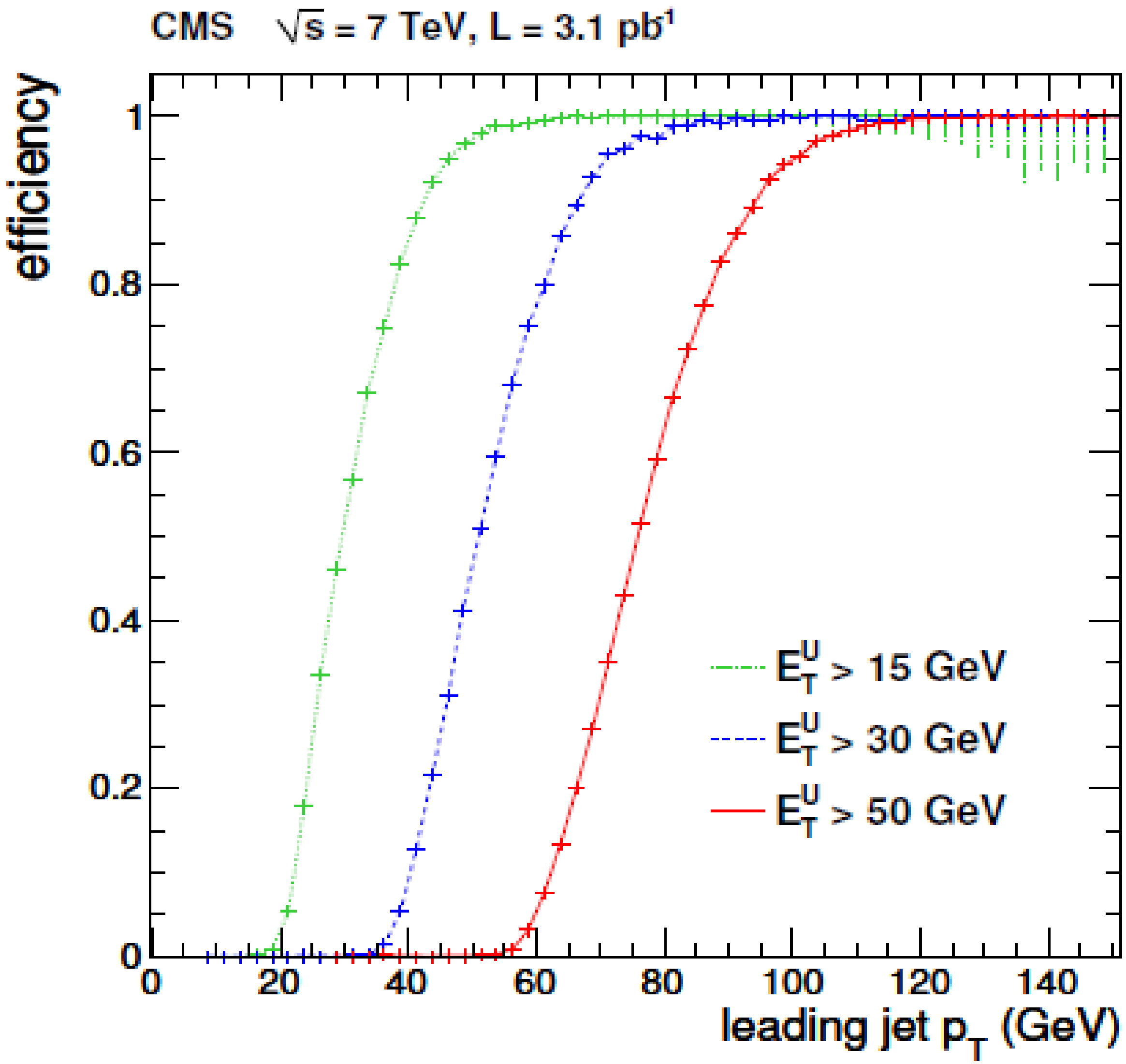
- isolated leptons with large transverse momentum > ~20 GeV (from W, Z, top)
- di-lepton events with transverse momentum > ~10 GeV
- jets with very high transverse momentum (several 100 GeV)
- events with large missing energy (~100 GeV)
- isolated photons with transverse energy >~50 GeV

lower-threshold triggers typically pre-scaled

Rest is difficult and probably not in recorded data !

for analysis, must know trigger efficiencies

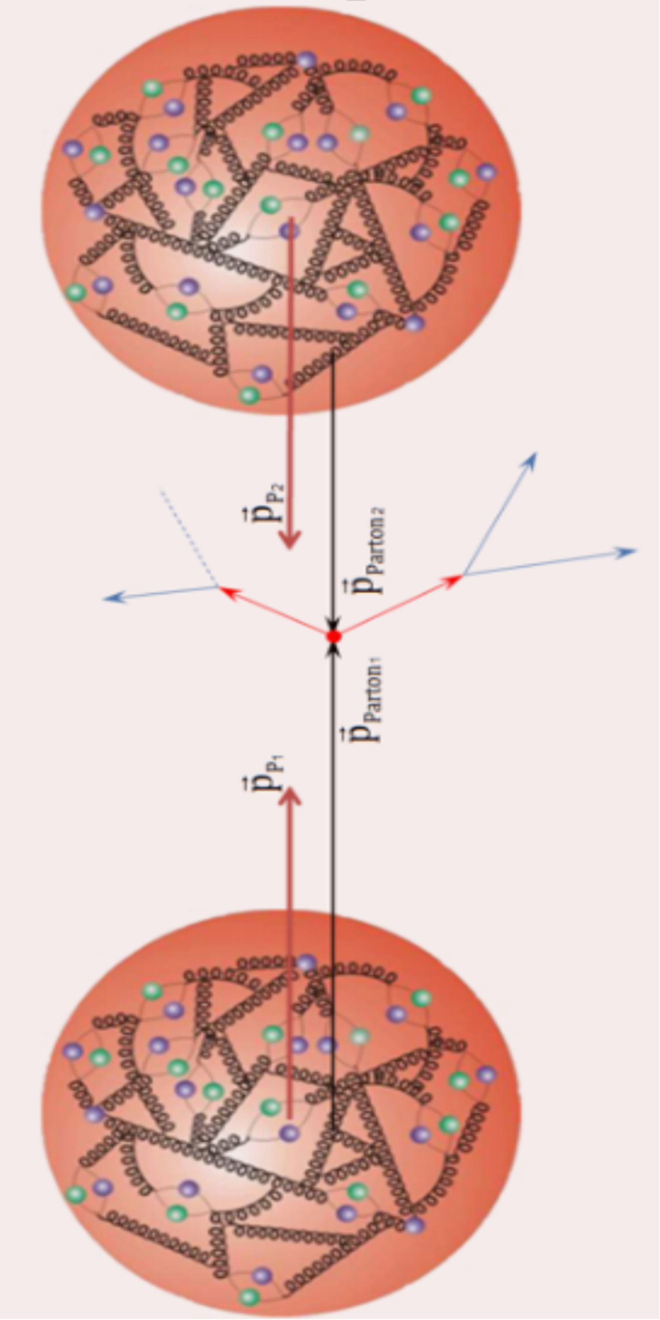
Example: trigger “turn-on” for jets

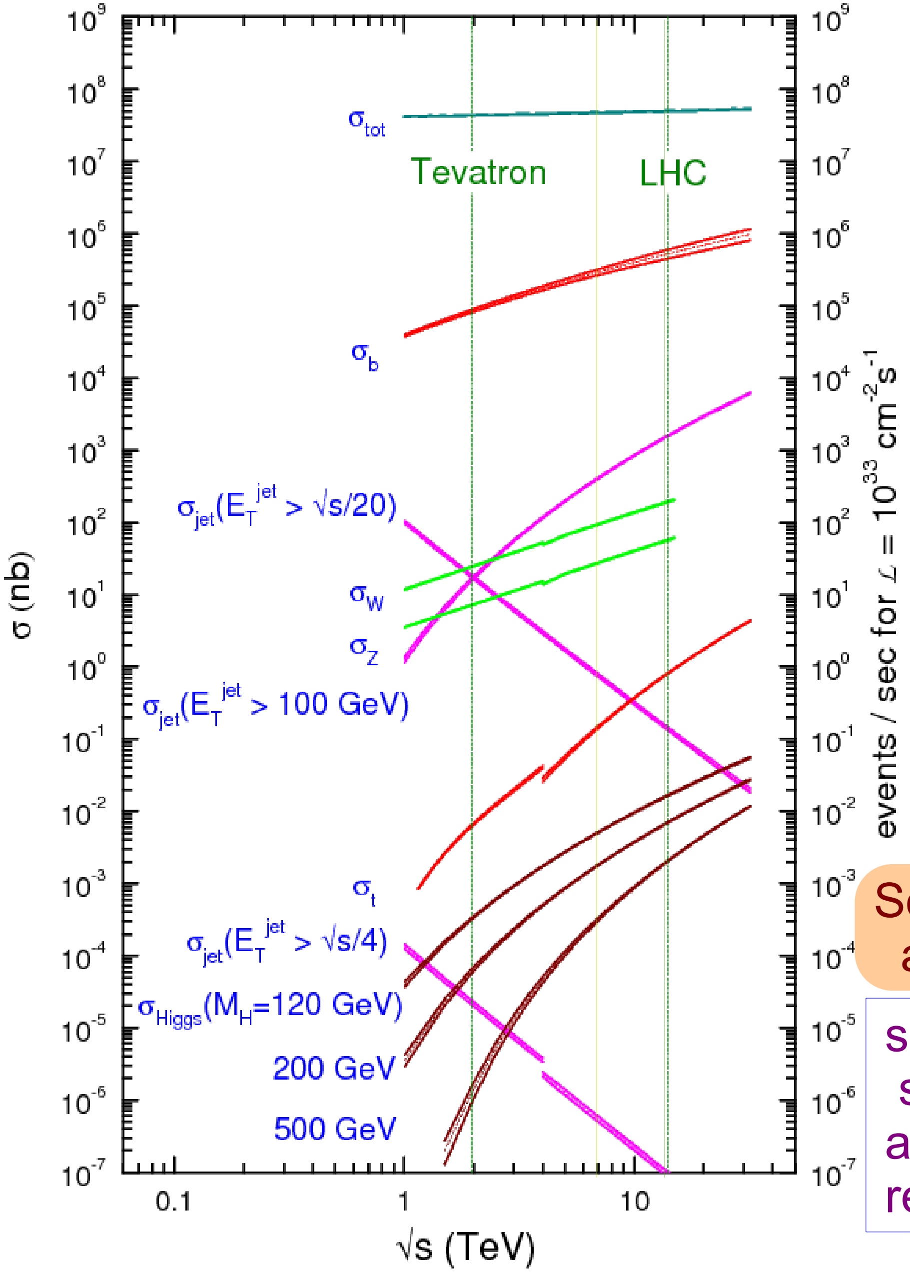


typical knee-shaped trigger efficiency curves (CMS, 2010), rising from 0 to 1

Data Analysis

Event Selection in the Analysis





Some processes are very rare !

sophisticated signal selection and background rejection needed.

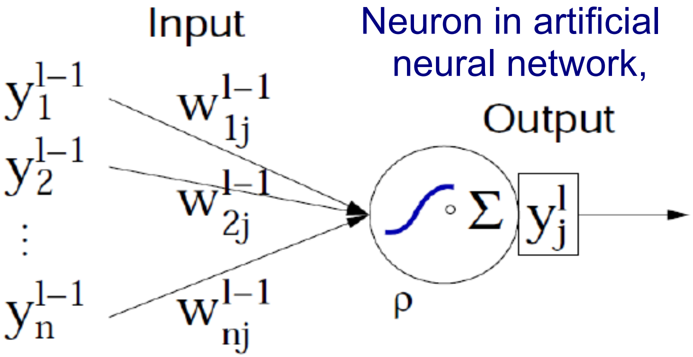
Analysis Steps

- recorded events are **reconstructed**: “detector hits” → physical objects like
  electrons, muons, photons, hadrons, jets, missing energy …
  need to know reconstruction efficiency and resolution

- **selection** of “interesting events” and objects for a particular analysis
  affected by selection efficiencies for signal and background processes

- last step of analysis involves advanced algorithms for the optimal **separation of signal from background** and **extraction of parameters** of interest from the background-corrected signal distribution
  (multivariate analysis, MVA, like discriminant methods, decorrelated likelihood, artificial neural networks, boosted decision trees)

  understanding the systematics involved is required !



Neuron in artificial neural network,

- Finally, arrive at a result with statistical and systematic errors
  evaluation of systematics requires much hard work
  Much use of **simulated data** is made in this process to evaluate known or suspected sources of uncertainties and propagate them to the final results.

see e.g. lecture

“Datenanalyse”

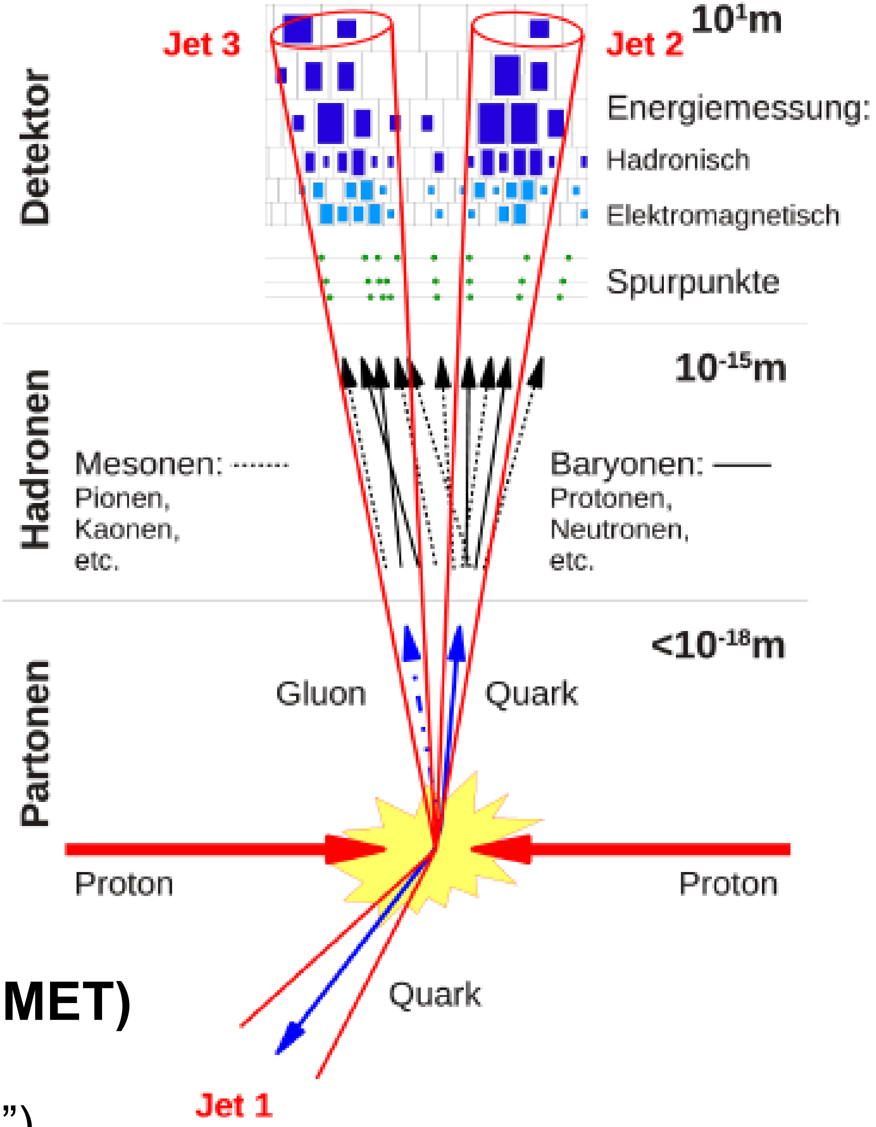
Reconstruction of Objects

1. **combine sub-detectors** to classify all stable objects, i.e. find electrons, muons, photons, hadrons.

2. **cluster objects** into “jets”
   relation between
   measured final state objects
   & hard partons
   two types of algorithms:
   1. **“cone”**: geometrically assign objects to the leading object
   2. **sequentially combine** closest pairs of objects – different measures of “distance” exist (kT, anti-kT) with some variation of resolution parameter, which determines “jet size”

   CMS does this across detector components (“particle flow” analysis)

3. determine **missing transverse energy (MET)** carried away by undetectable particles (neutrinos, or particles signalling “new physics”)



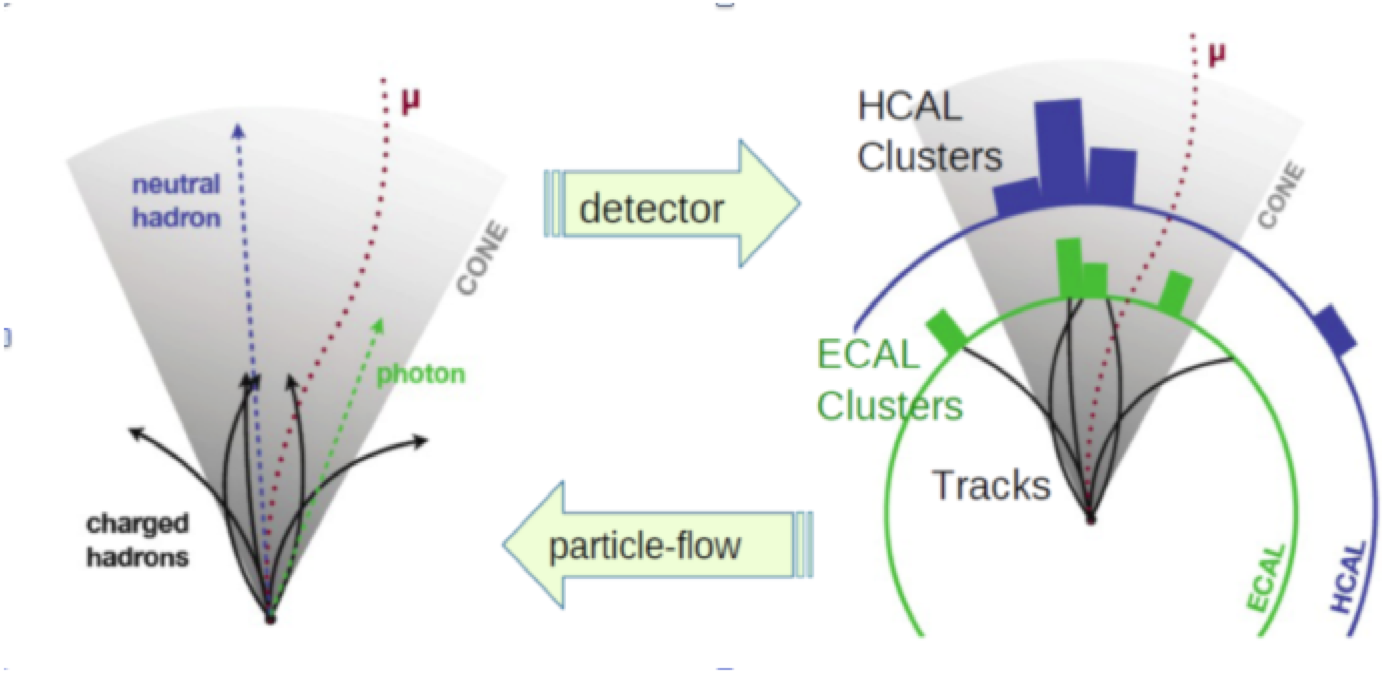
High-level Reconstruction

Particle Flow

- Attempts to reconstruct and identify all particles in the event
- Optimally combines information from all sub-detectors to give best four-momentum measurement of each particle type:

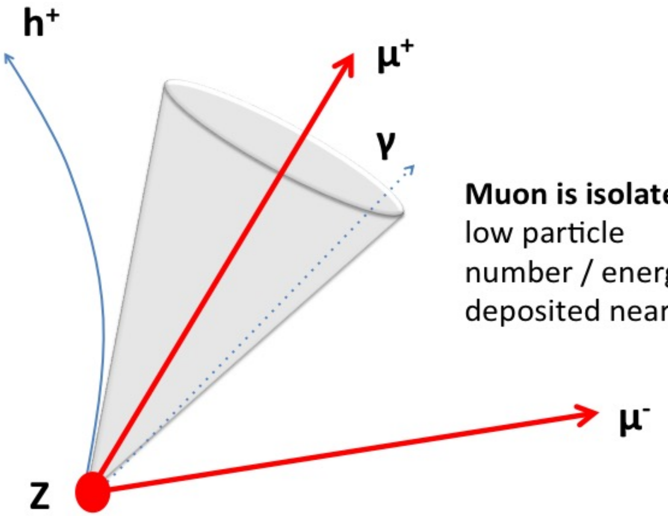
Charged hadrons, **neutral hadrons**, **electrons**, **photons** and **muons**

- Also improves performance for higher-level composite objects e.g. jets, MET

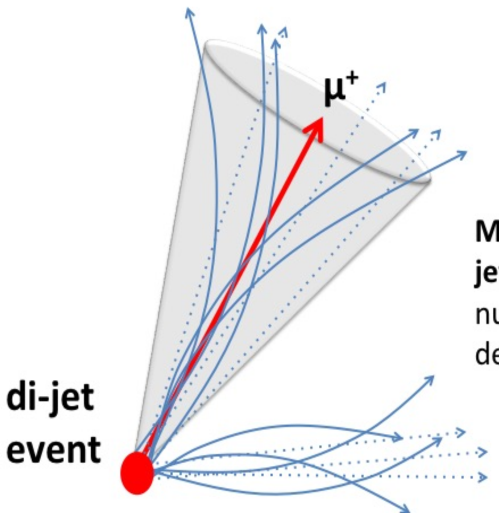


Event Selection

- Key concepts are: object identification and object isolation
- **Identification:** The true particle type can be ambiguous
  - “Is it an electron or a pion?” → can apply object criteria to increase purity of a particle type, e.g. small hadronic energy / EM energy → more likely to be an electron
- **Isolation:** powerful handle to reduce background from jets
  - We are often interested in leptons produced from decays of top quarks, W bosons, Z bosons, Higgs etc
  - These electroweak processes are 'clean' compared to QCD → less activity in the region around lepton direction

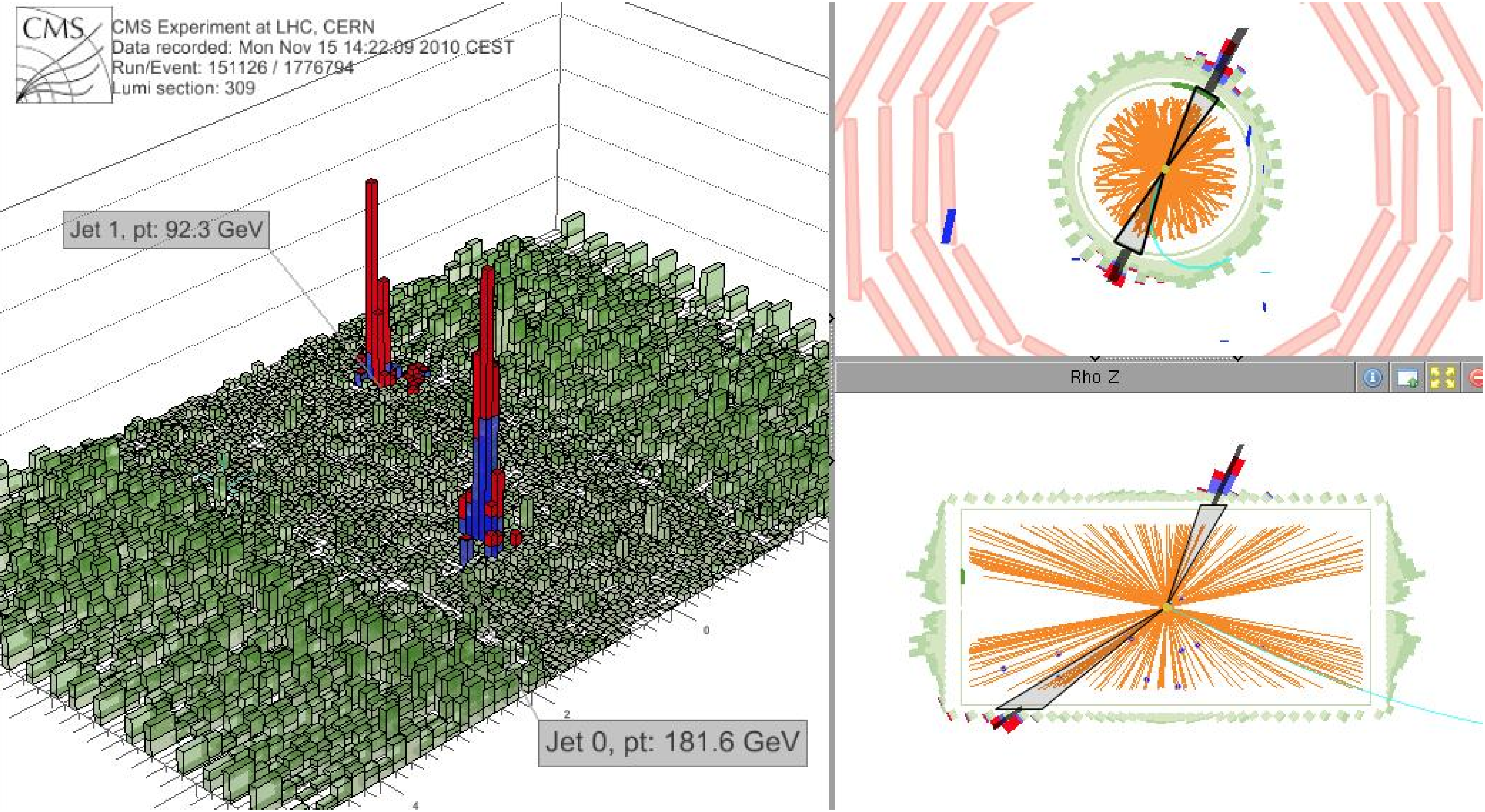


Muon is isolated: low particle number / energy deposited nearby

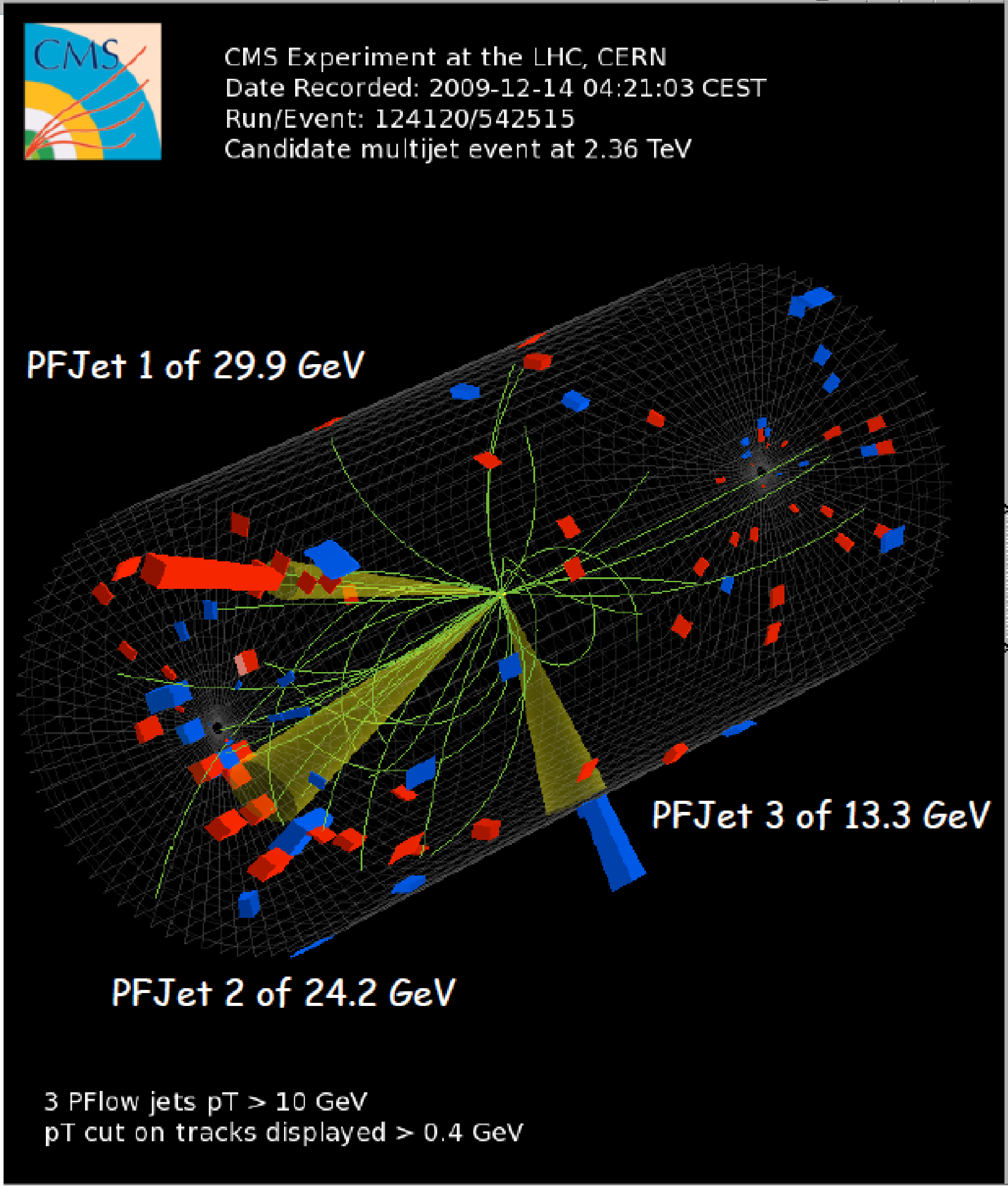


Muon produced in jet: high particle number / energy deposited nearby

Two-Jet Event in the CMS detector



Three-jet event



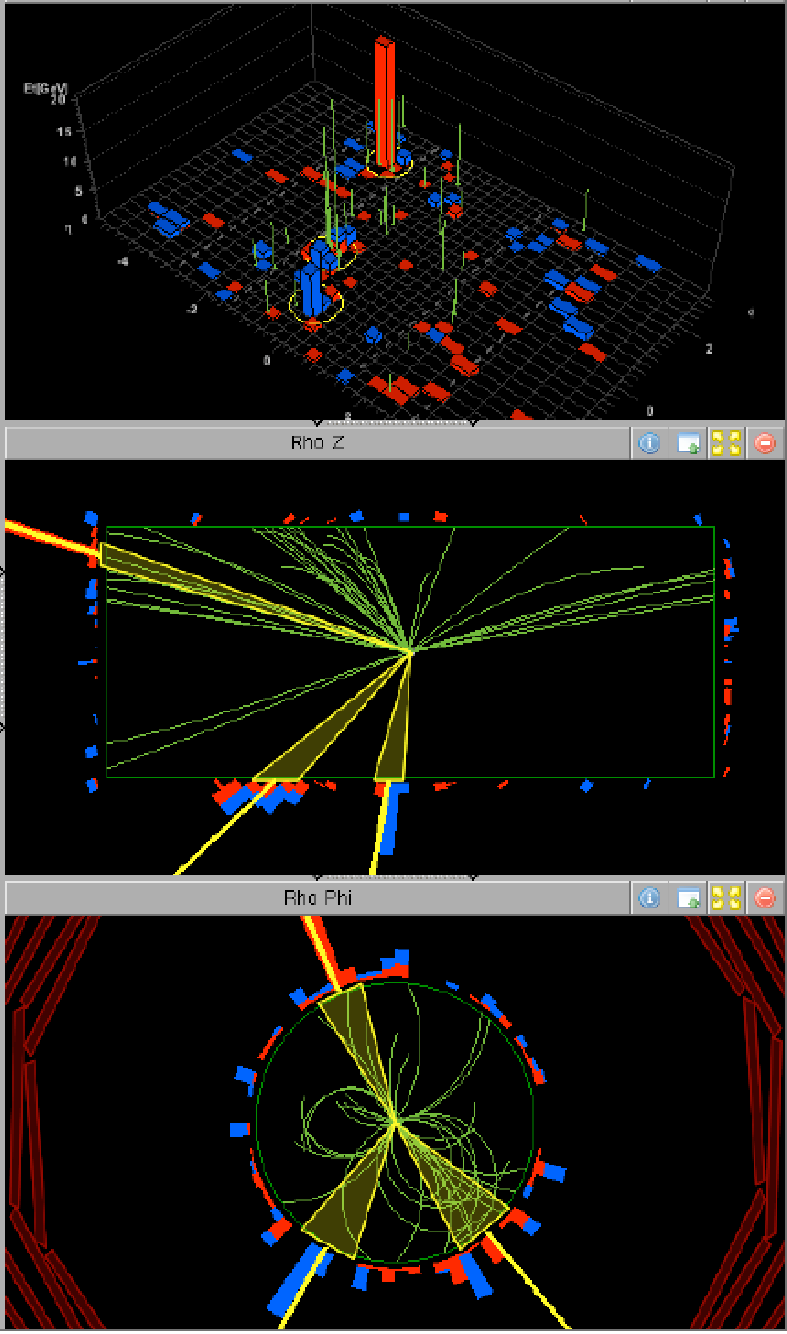
CMS Experiment at the LHC, CERN
Date Recorded: 2009-12-14 04:21:03 CEST
Run/Event: 124120/542515
Candidate multijet event at 2.36 TeV

PFJet 1 of 29.9 GeV

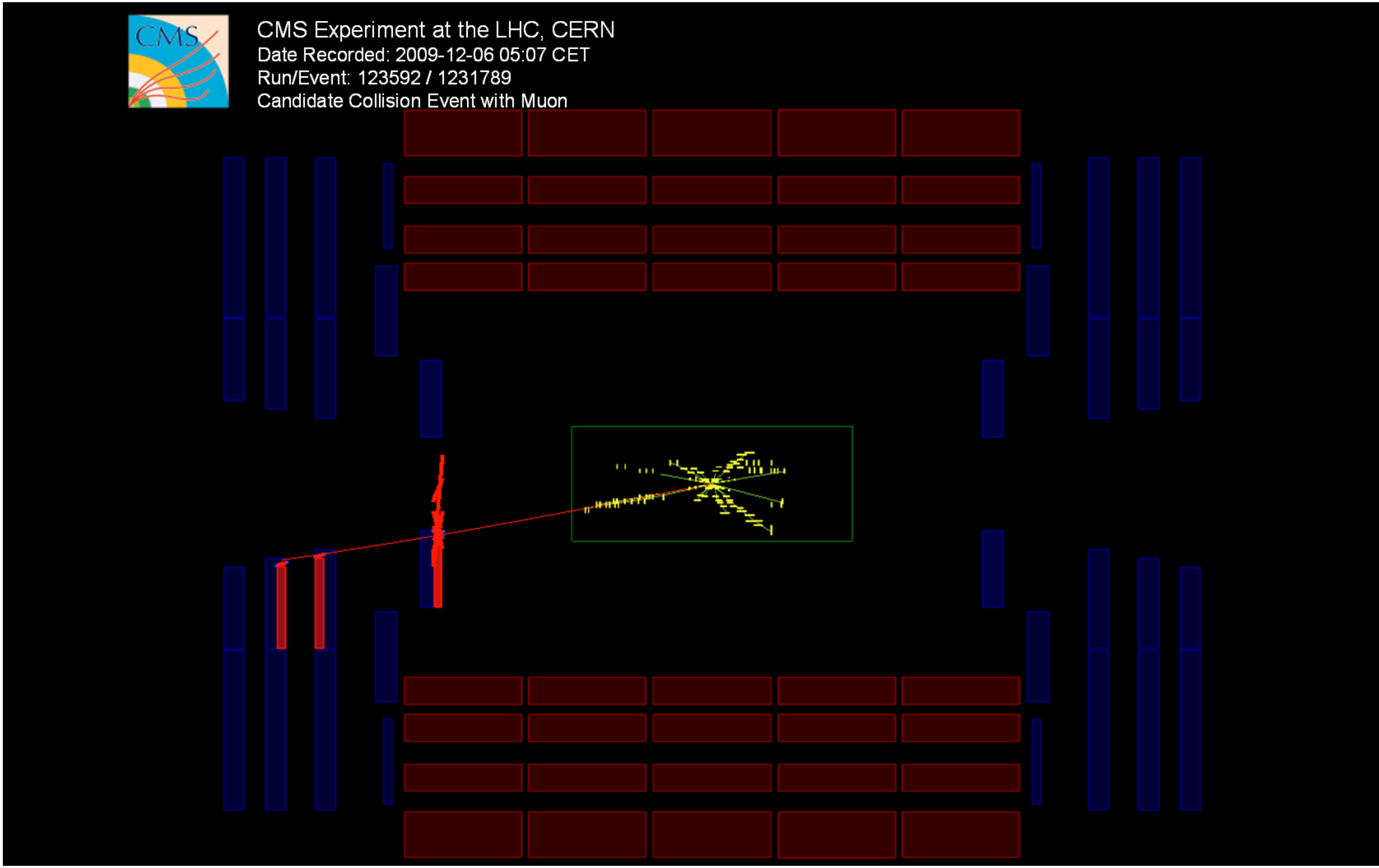
PFJet 3 of 13.3 GeV

PFJet 2 of 24.2 GeV

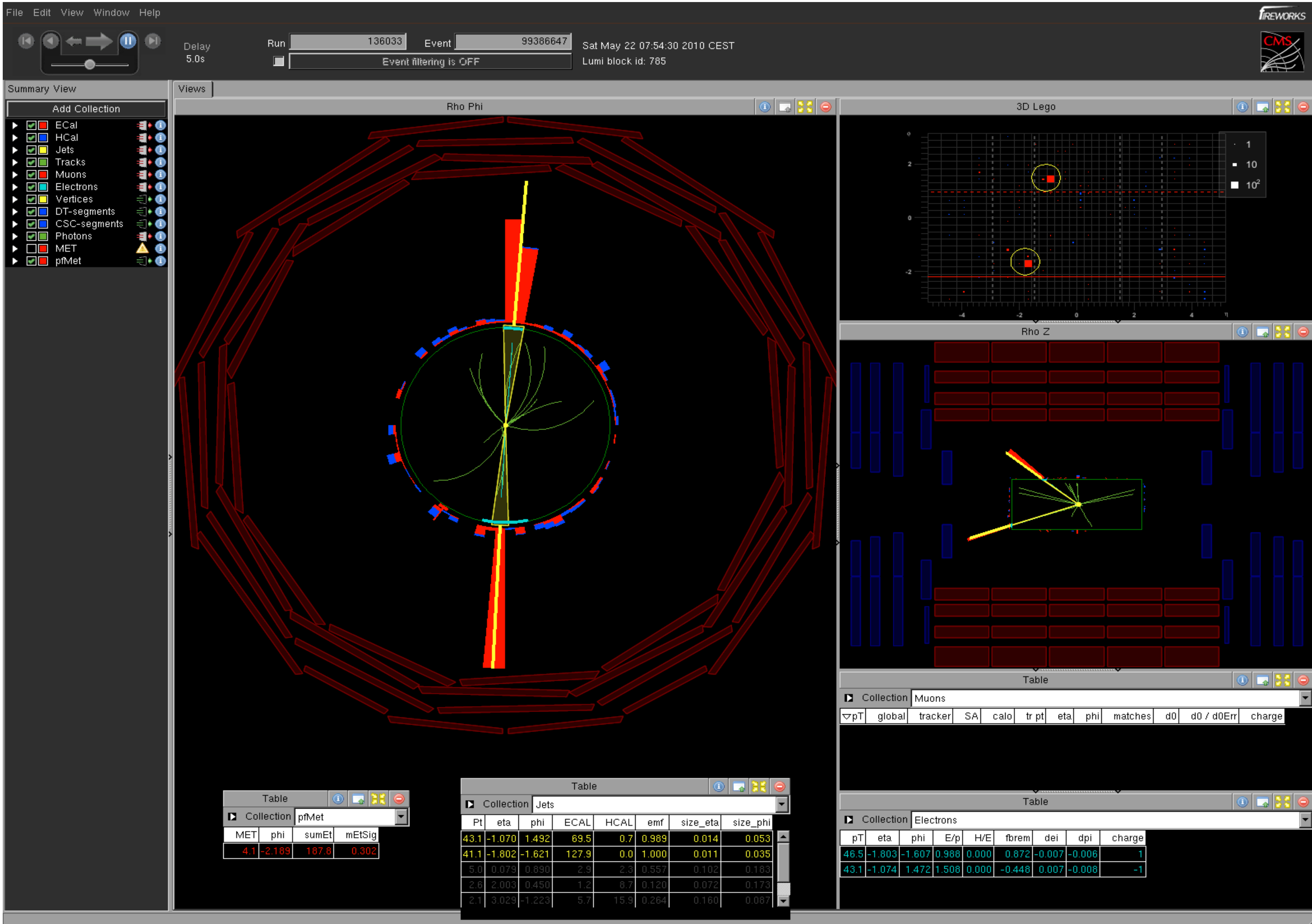
3 PFlow jets pT > 10 GeV
pT cut on tracks displayed > 0.4 GeV



event with end-cap muon



2 electrons in CMS



Delay 5.0s

Run 136033 Event 99386647

Event filtering is OFF

Sat May 22 07:54:30 2010 CEST

Lumi block id: 785

Collection: Muons

▽pT	global	tracker	SA	calo	tr pt	eta	phi	matches	d0	d0 / d0Err	charge

Collection: Electrons

pT	eta	phi	E/p	H/E	fbrem	dei	dpi	charge
46.5	-1.803	-1.607	0.988	0.000	0.872	-0.007	-0.006	1
43.1	-1.074	1.472	1.508	0.000	-0.448	0.007	-0.008	-1

Collection: pfMet

MET	phi	sumEt	mEtSig
4.1	-2.189	187.8	0.302

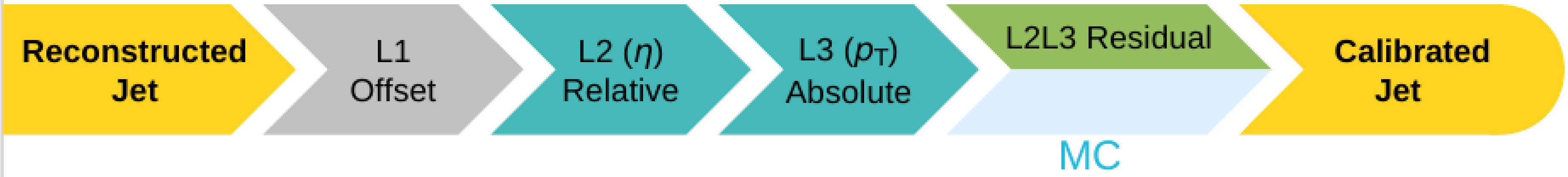
Collection: Jets

Pt	eta	phi	ECAL	HCAL	emf	size_eta	size_phi
43.1	-1.070	1.492	69.5	0.7	0.989	0.014	0.053
41.1	-1.802	-1.621	127.9	0.0	1.000	0.011	0.035
5.0	0.079	0.890	2.9	2.3	0.557	0.102	0.183
2.6	2.003	0.450	1.2	8.7	0.120	0.072	0.173
2.1	3.029	-1.223	5.7	15.9	0.264	0.160	0.087

Calibration

Energy/momentum of objects must be calibrated

Calibration of the jet energy in CMS ...



data

Reconstructed Jet → L1 Offset → L2 (η) Relative → L3 (p_T) Absolute → L2L3 Residual → **Calibrated Jet**

MC

... is a multi-step procedure, driven by data

Level 1: offset correction for pile-up and electronic noise
Level 2: relative (η) corrections
Level 3: absolute p_T correction
MC and special balanced events

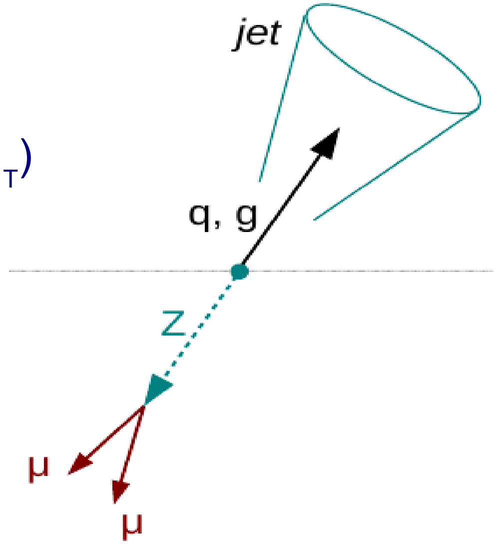
residual corrections from events with selected topology:

Level 2 residual η

from measured di-jet events, assuming the two jets have the same E_T)

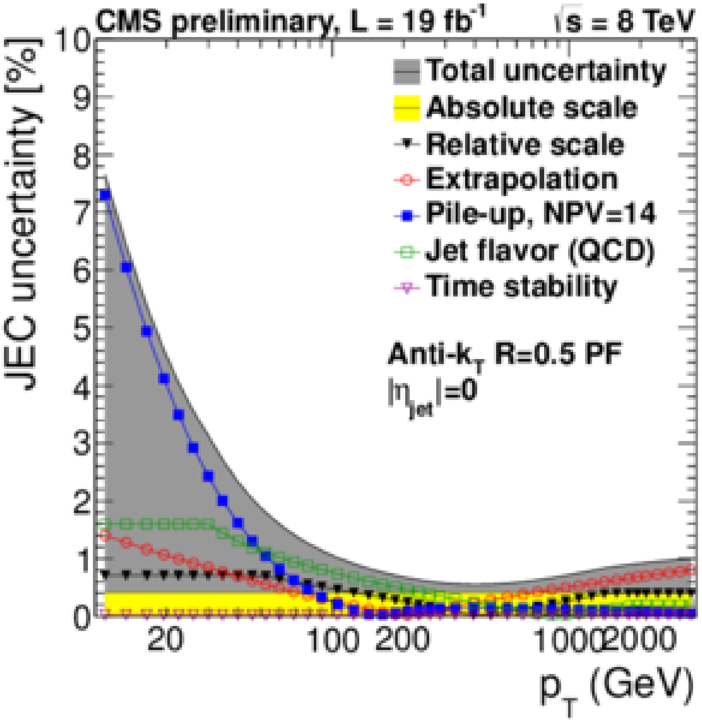
Level 2 residual p_T

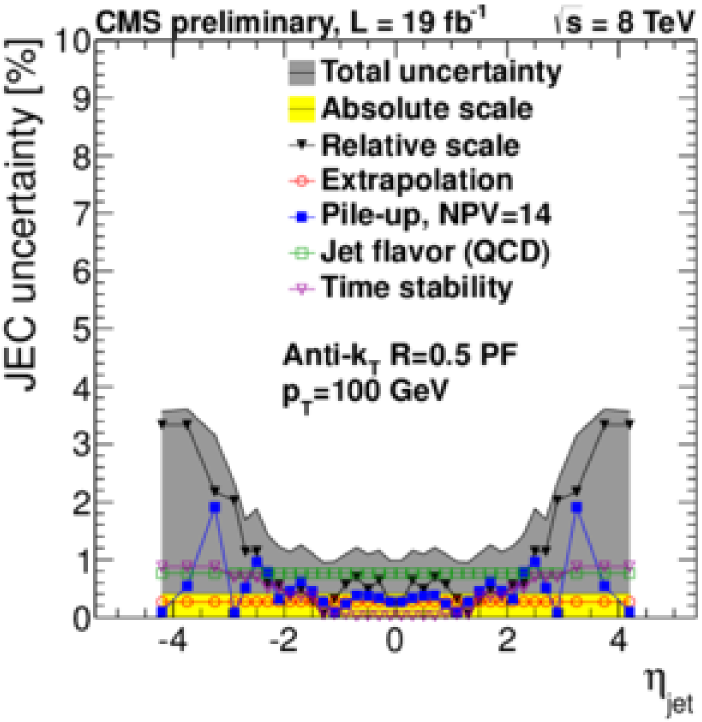
from measured Z+jet & photon+jet, jet blanced by Z/γ



Precision of Jet energy calibration better than 1 % !

Jet-energy calibration





Precision of Jet energy calibration reaches 1 % !

More complicated observables

Calculate **derived quantities** from objects,

examples:

– invariant masses of groups of objects to reconstruct decaying particles

– transverse momentum or energy, $\vec{p_T} = \sum_i \vec{p_T}_i^2$, $E_T = \sum_i {m_i}^2 + \vec{p_T}_i^2$

at hadron colliders where rest system of an interaction is boosted along z direction

– missing transverse energy, from all particles in an event, *assuming total transverse momentum of zero in each event, measures effects of invisible particles*

(neutrinos in the SM, but there are others in extended theories)

$$E_{T\,\mathrm{miss}} = - \sum_{\mathrm{all\ \ partilces}} {m_i}^2 + \vec{p_T}_i^2$$

– “transverse mass” (${M_T}^2 = \sum_i {E_T}_i^2 - \sum_i {\vec{p}_T}_i^2$) of groups of objects

– scalar sum of jet energies or sum of transverse jet energies to quantify the energy scale of the hard process in an interaction

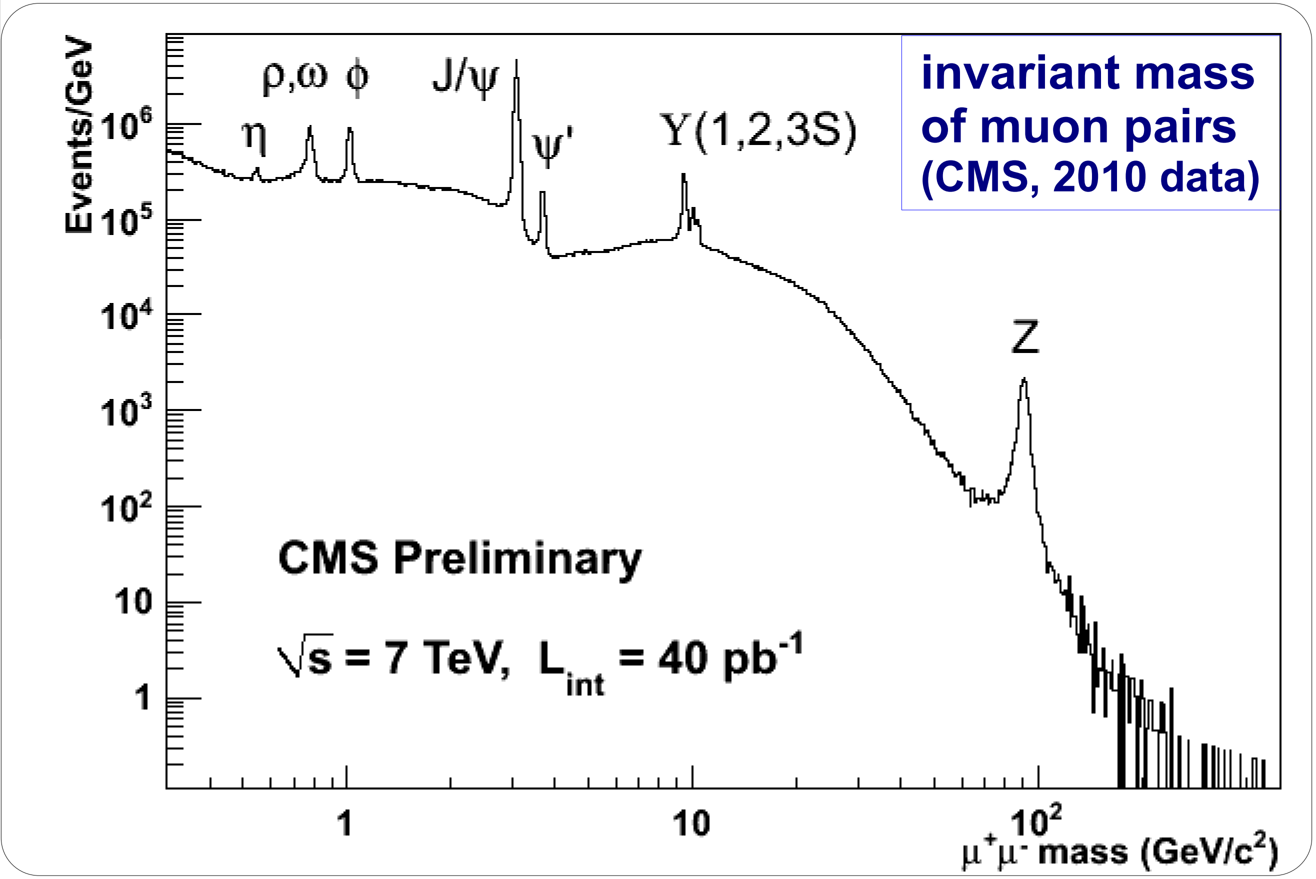
– event shape variables (for QCD analyses) to classify jet topologies

– all kinds of “classifiers” using MVA techniques *for object or event classification*

More complicated observables

60 years of particle physics in only one year:

Example of a very simple selection:
just the invariant mass of muon pairs in events with one muon trigger



invariant mass of muon pairs
(CMS, 2010 data)

Determination of efficiencies

two options:

1. take efficiencies from simulation not always believable !

   check classification in simulated data vs. truth, i.e. determine

   ε_{MC} = fraction of correctly selected objects

   (probability to select background determined in the same way)

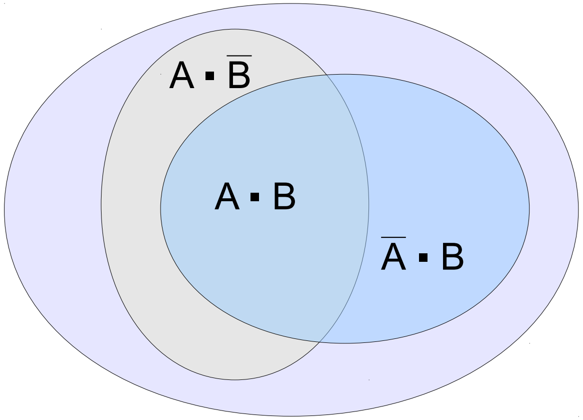
2. **design data-driven methods** using redundancy of at least two variables discriminating signal and background

   – tag & probe method:

   select very hard on one criterion, even with low efficiency, check result obtained by second criterion

Illustration: two **independent** criteria A, B

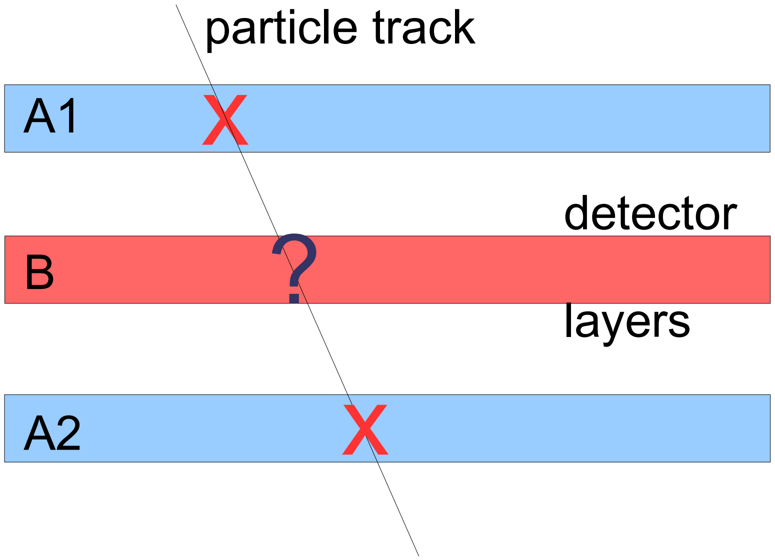
$$\boxed{\epsilon_B = \frac{n(A\cdot B)}{n(A\cdot B)+n(A\cdot\bar{B})}}$$



Important: selecting on A must not affect B, i.e. A and B must be uncorrelated !

Tag and Probe: Examples

Example 1:



Hits in layers A1 and A2 define valid particle track (tag)

probe hit in layer B

Coincidence of Layers A1 and A2 guarantees high purity of the tag (protects against random noise)

allows determination of efficiency of layer B

$$\Rightarrow \quad \epsilon_B = \frac{n_B}{n_{A1\cdot A2}}$$

Trigger efficiencies

Determination of trigger efficiencies depends on existence of independent selection methods

Important to ensure redundancy when building trigger systems !

Trigger information must be stored for later use in efficiency determination !

typical methods:

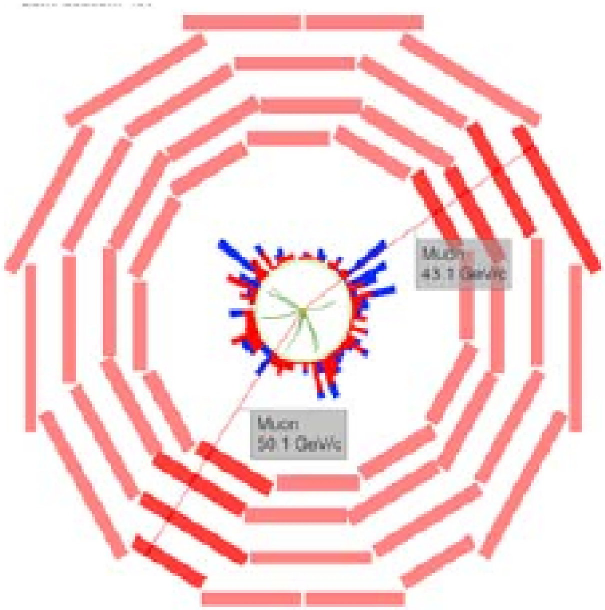
- use trigger from independent sub-systems
- trigger at lower threshold (typically pre-scaled to run at acceptable rates)
  → probe higher-threshold triggers
- trigger on pairs of objects at low threshold,
  → probe higher threshold on each member of the pair
  !!! potential bias, because higher-threshold trigger depends on same input signals as the tag !!!
- trigger only one object of a pair and use an off-line criterion to identify 2nd member of the pair and probe trigger decision on it

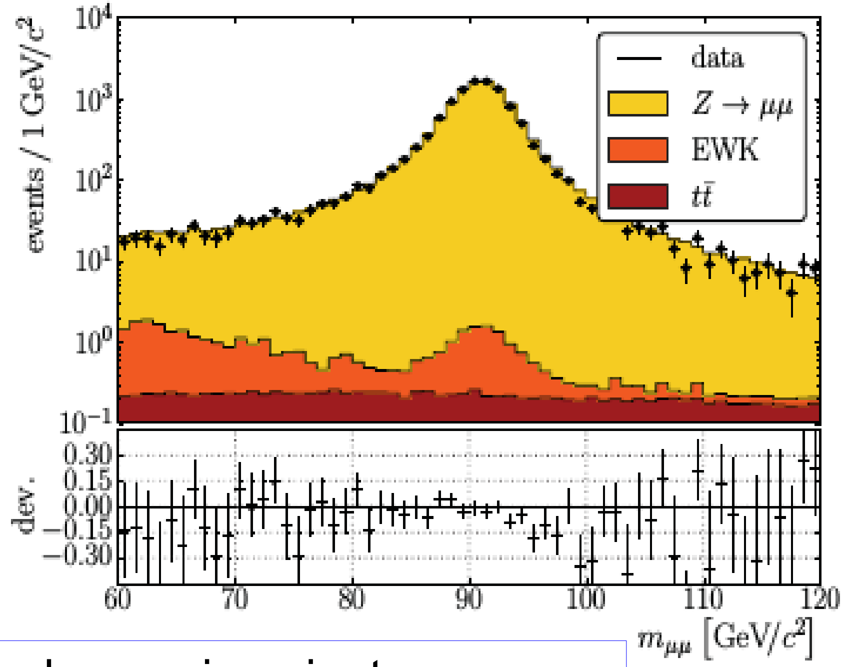
Examples

Example 2:

criterion A: a tight muon/electron and
one other track with tight selection on Z mass **(“tag”)**
thus selecting Z → μμ or Z → ee events
(which is possible with very high purity)
→ 2nd track also is a muon/electron with very high probability

criterion B: 2nd track selected by trigger (or analysis) **(“probe”)**
allows measurement of trigger efficiency
(or selection efficiency) of second muon





Z → μμ event in the CMS detector and invariant μμ mass

Statistical error on efficiency

determination of efficiencies is a clear application of **binomial statistics**:

number of successes ***k*** *in* ***n*** *trials at probability* ***p*** *per trial*

Binomial Distribution

$$\boxed{P(k;p,n) = \binom{n}{k} p^k (1-p)^{n-k}, k = 1, \ldots, n} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Expectation value

$$\boxed{\mathrm{E}[k] = np}$$

Variance

$$\boxed{\mathrm{V}[k] = np(1-p)}$$

Error on efficiency: insert measured efficiency $\epsilon = k/n$ in formula for variance

(instead of true (but unknown) selection efficiency p !)

$$\rightarrow \boxed{\sigma_\epsilon = \frac{\sqrt{\epsilon(1-\epsilon)}}{\sqrt{n}}}$$

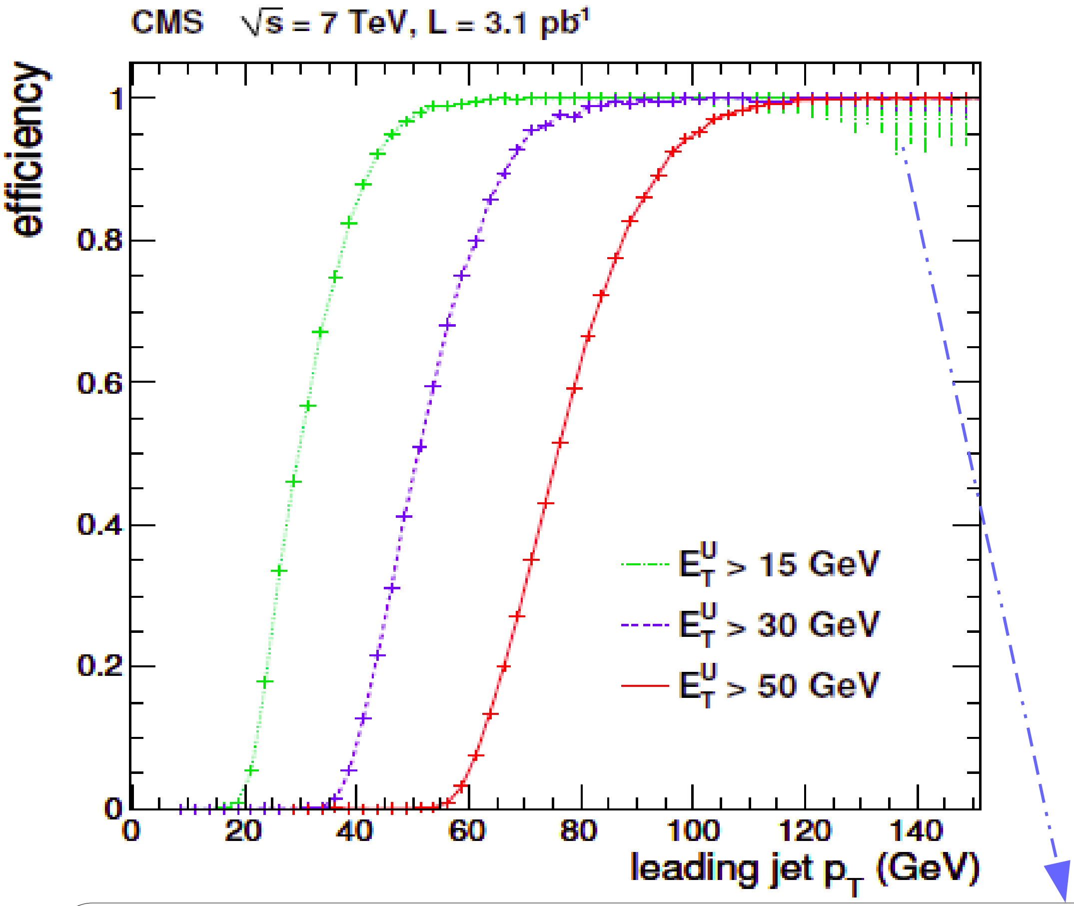
if this is not justified due to very small statistics, a more sophisticated method of “interval estimation” is needed to specify a confidence range on the measured efficiency:

→ Clopper-Pearson method

Example 3: Trigger efficiencies

Typical “turn-on” curves of trigger efficiencies

(calorimeter jet trigger on transverse energy of jets, CMS experiment)



Remarks:

- efficiency at 100% only far beyond “nominal” threshold
- trigger efficiencies vary with time (depend on “on-line” calibration constants)
- to be safe and independent of trigger efficiencies, analyses should use cuts on reconstructed objects that are tighter than trigger requirements

2nd remark: errors determined as 68% confidence interval by application of Clopper-Person method per bin; this explains the (counter-intuitive) large uncertainties on the >15 GeV trigger at high pT:

there were just no events observed where trigger was inefficient.

LESSON: sophisticated methods are not always plausible !

Determination of background

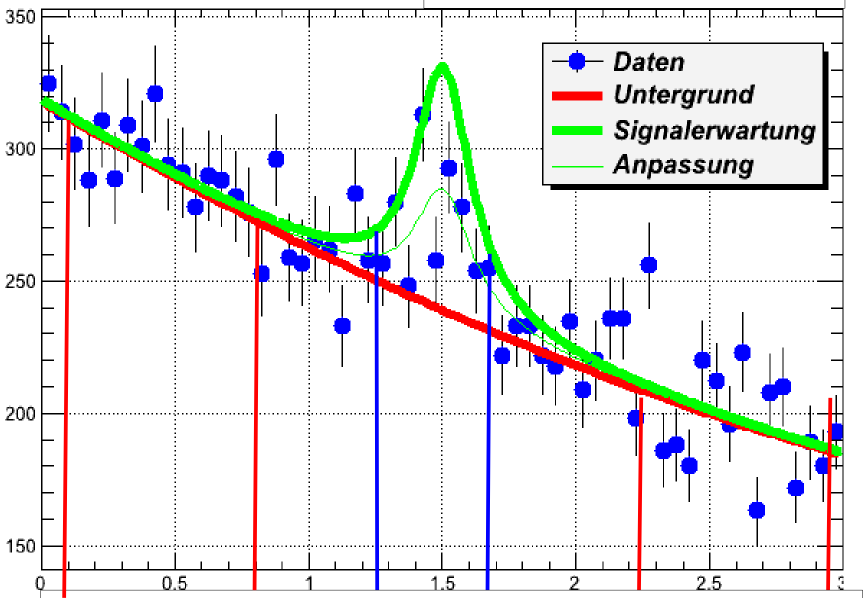
– **take from MC** (same comments as above)

– **extrapolation from "side band"**
assuming "simple" background shape or by taking background shape from simulation

- **event counting in background regions, extrapolation under signal assuming (simple) model**

- **fit of signal + background model to the observed data**

signal on background



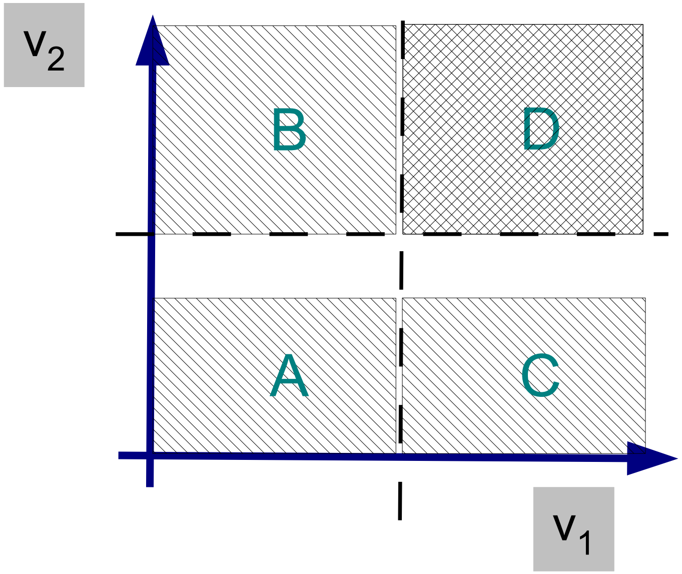
background signal background
regions

– **if a second, independent variable for separation of signal from background can be found, background determination purely from data becomes possible**

→ **ABCD method**

Determination of background

– ABCD – Method ...



Assumptions:

- two independent variables v1 and v2 for background
- signal only in region D

$$\rightarrow \quad n_D^{bkg} = n_C \frac{n_B}{n_A}$$

... a **data driven estimate** of *background under a signal*

Example: invariant mass of two unlike-sign particles, combinatorial background from sample with like-sign particles.

– **more advanced methods** exist to **exploit two uncorrelated variables** to predict the background shape under a signal, see e.g. “sPlot method” in ROOT.

Example of improved background modelling

Hybrid events: data + Monte Carlo

example: Z → ττ background in the H → ττ search

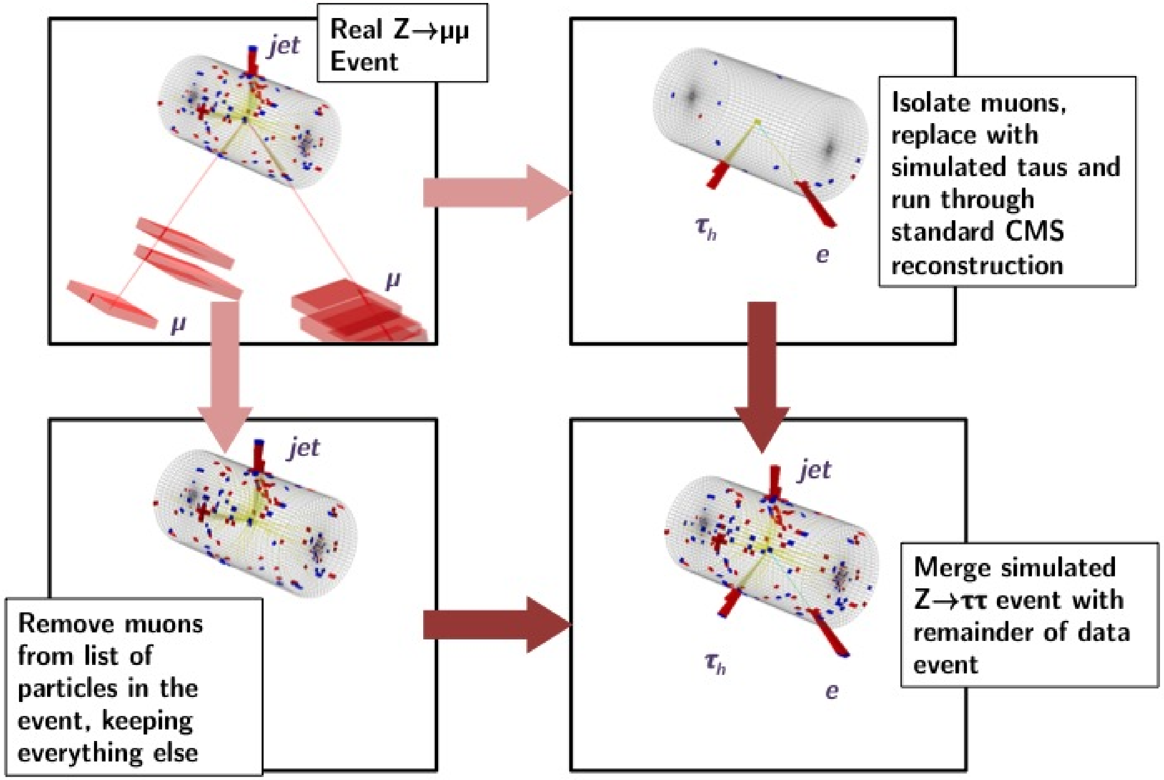
- H → μμ has very low cross section, hence there is no H → μμ under H → μμ
- Z → μμ and Z → ττ are very similar (lepton universality of weak decay)

idea:

replace real μ in Z→μμ events with simulated τ to model Z background under H signal

advantages:

- non-leptonic part of event is from real data, esp. important in presence of pile-up
- leptonic part can be well and easily modelled
- important cross check of full simulation via MC



Real Z→μμ Event

Isolate muons, replace with simulated taus and run through standard CMS reconstruction

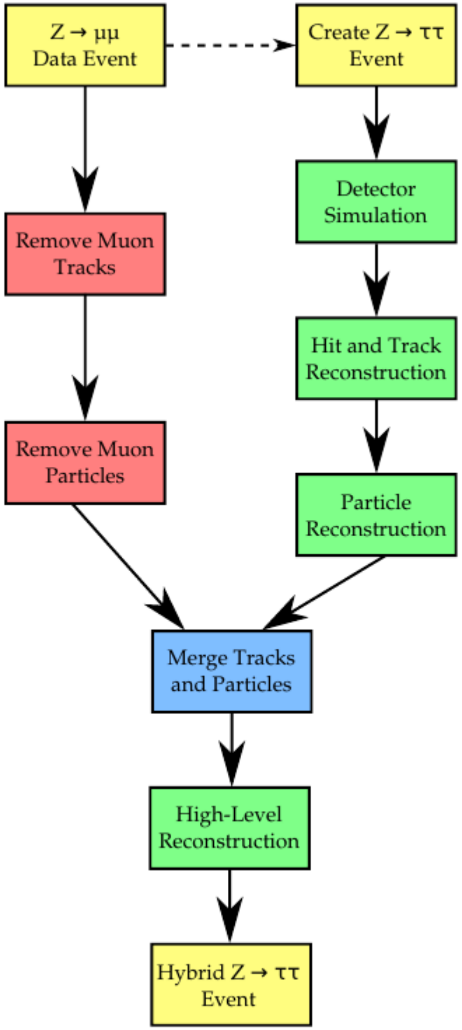
Remove muons from list of particles in the event, keeping everything else

Merge simulated Z→ττ event with remainder of data event

Embedding: two options

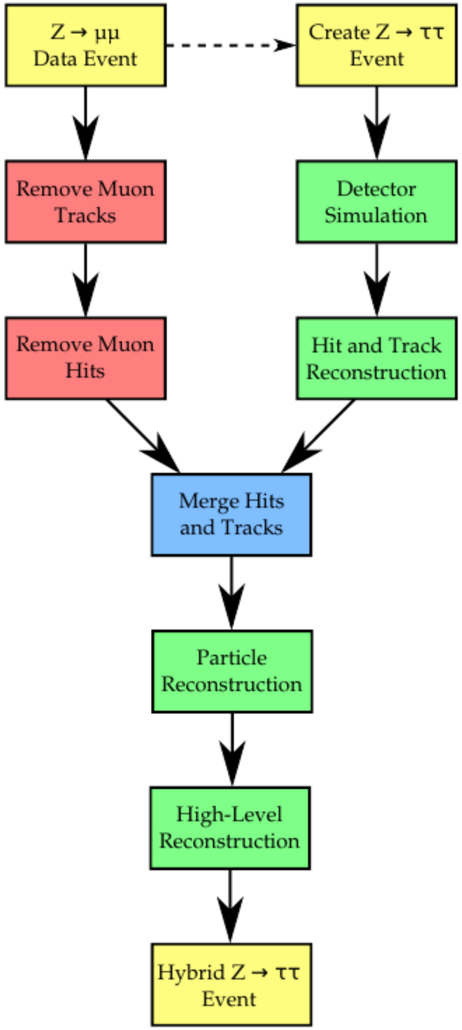
Embedding based on

reconstructed objects



- $Z \to \mu\mu$ Data Event → Remove Muon Tracks → Remove Muon Particles → Merge Tracks and Particles
- Create $Z \to \tau\tau$ Event → Detector Simulation → Hit and Track Reconstruction → Particle Reconstruction → Merge Tracks and Particles
- Merge Tracks and Particles → High-Level Reconstruction → Hybrid $Z \to \tau\tau$ Event

detector hits



- $Z \to \mu\mu$ Data Event → Remove Muon Tracks → Remove Muon Hits → Merge Hits and Tracks
- Create $Z \to \tau\tau$ Event → Detector Simulation → Hit and Track Reconstruction → Merge Hits and Tracks
- Merge Hits and Tracks → Particle Reconstruction → High-Level Reconstruction → Hybrid $Z \to \tau\tau$ Event

- more difficult

+ also simulates reconstruction efficiency

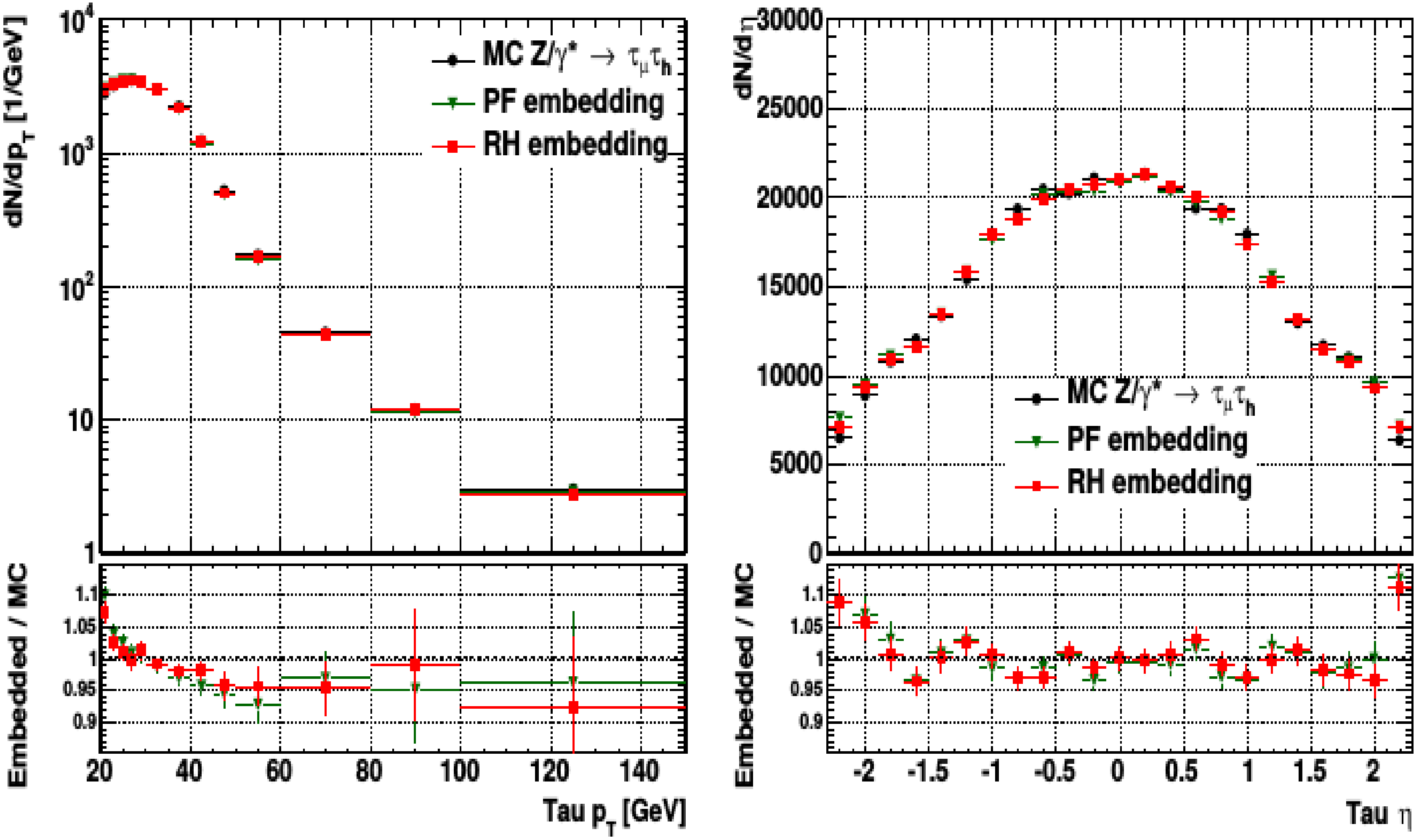
+ can take into account extra clusters due to “pile-up” (i.e. multiple pp collisions in an event)

from PhD thesis Armin Burgmeier, Karlsruhe - DESY, June 2014

Validation of Method with MC

"Closure Test"

demonstrate that method works on simulated events



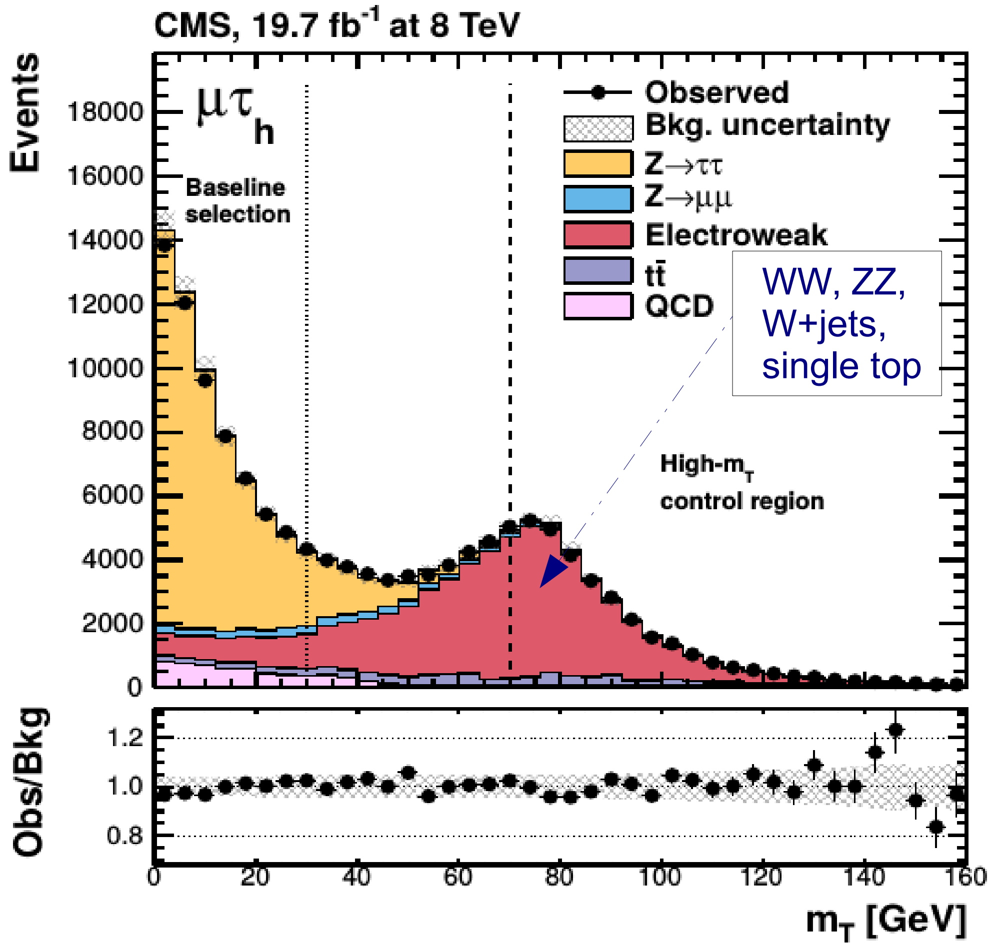
from PhD thesis Armin Burgmeier, Karlsruhe - DESY, June 2014

Embedding method: compare with data

Distribution of transverse mass in H → $\tau\tau$ candidate events

- $\tau\tau$ events are expected at low values of m_T
- Z → $\tau\tau$ events are well described by embedding method

(almost no H events are expected in this distribution)



WW, ZZ, W+jets, single top

Example illustrates usage of a background control region in a sensitive variable.

Coming Next:

statistical analysis of rare signals