Karlsruher Institut für Technologie

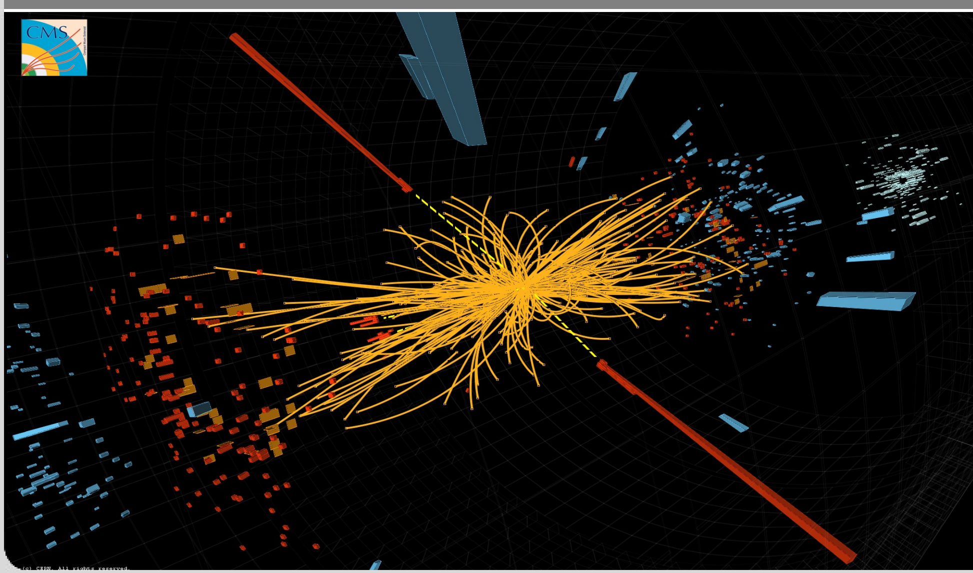
Higgs Boson Physics

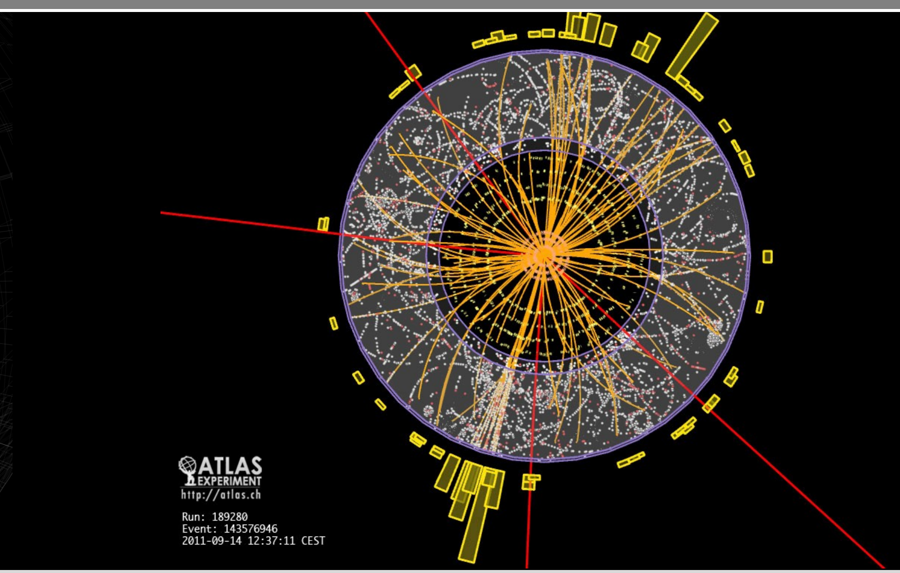
Analysis Techniques

Günter Quast, Alexei Raspereza

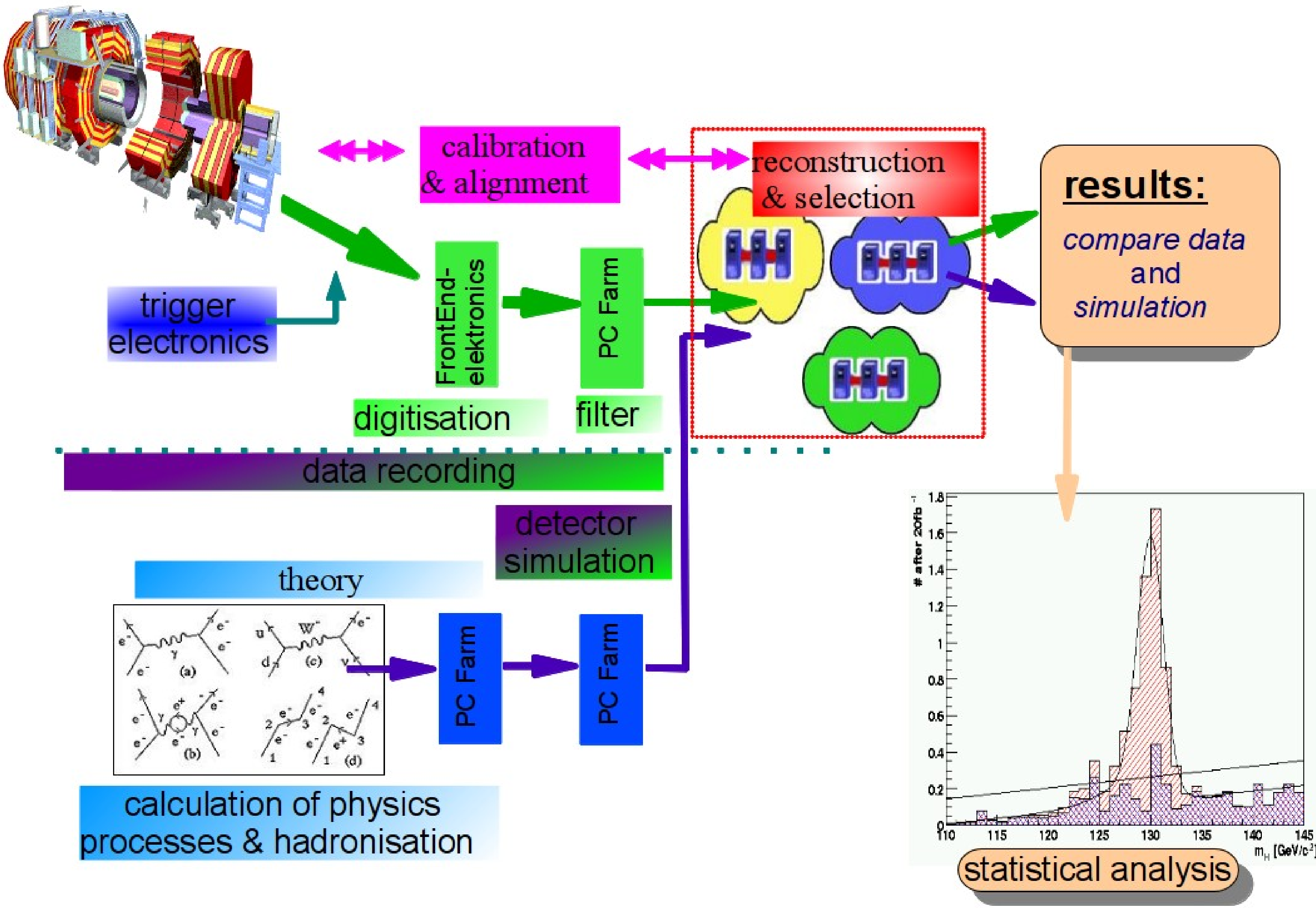
Master-Kurs
SS 2012

Institut für Experimentelle Kernphysik





Overview: Components of Analysis Chain



Components of Analysis Chain

- **Digitizers** record data from detector cells
  - remove empty cells („zero-suppression“ and „noise reduction“
- **Trigger** and **Filter** select „interesting“ events „on-line“
  to be stored for „off-line“ analysis
  (events not stored at this point are lost forever !)
- **Reconstruction** process constructs physical objects
  (electrons, muons, jets, …)
  (this and subsequent steps can be repeated many times)
- **Pre-selection** identifies interesting events and objects in
  events for further processing and analysis
- **Analysis** compares measured distributions with
  theoretical expectations

Experiment

Theory

- **theoretical calculation** of production cross sections
- **hadronisation** of quarks and gluons into jets

- **Detector simulation**

**same reconstruction, selection and analysis steps
for simulated events as for real events**

Monte-Carlo Generators

Calculation of differential cross sections

want to understand

$\mathcal{L}_{int} \quad \rightarrow$ final states

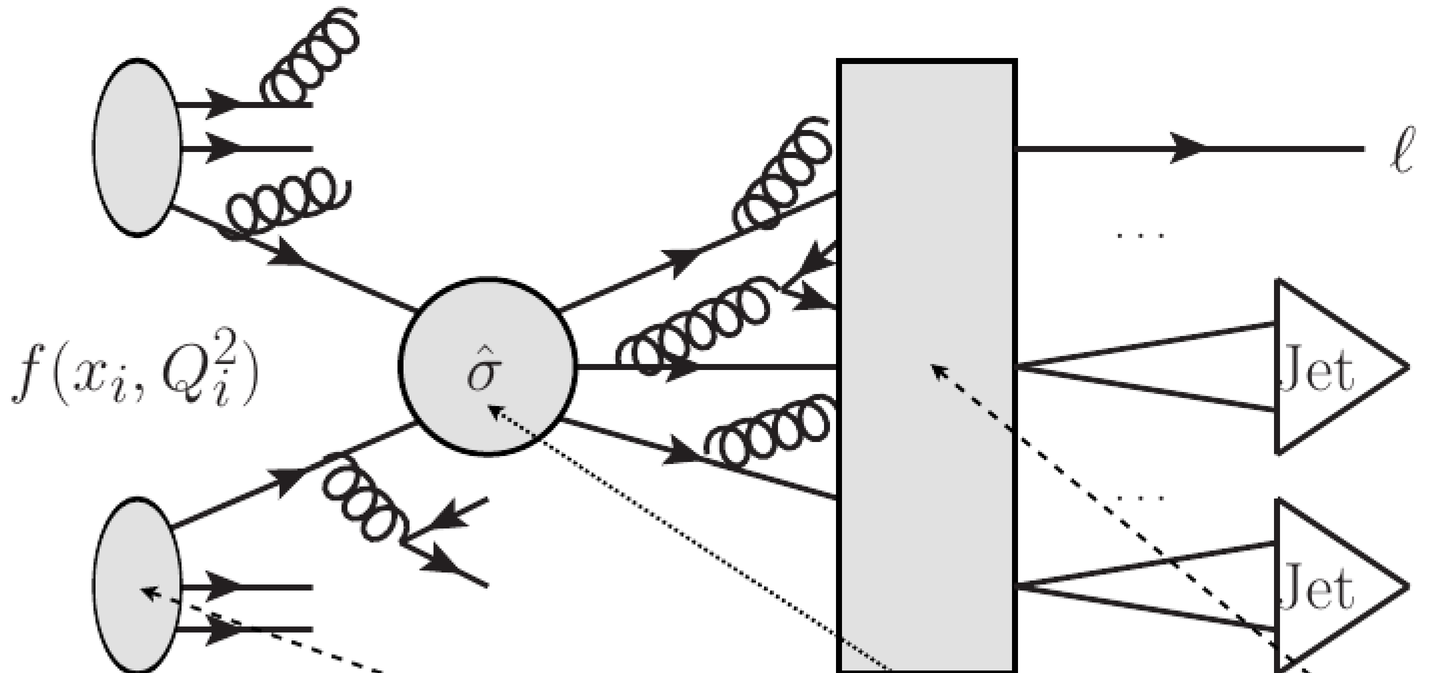
and predict measurable quantities

$$\frac{\partial \sigma}{\partial O_i} \quad = \text{differential cross section}$$

O_i: production angles of final state particles,
momenta of final state particles,
invariant mass of (groups of) final state particles;
…

Calculation of Cross sections

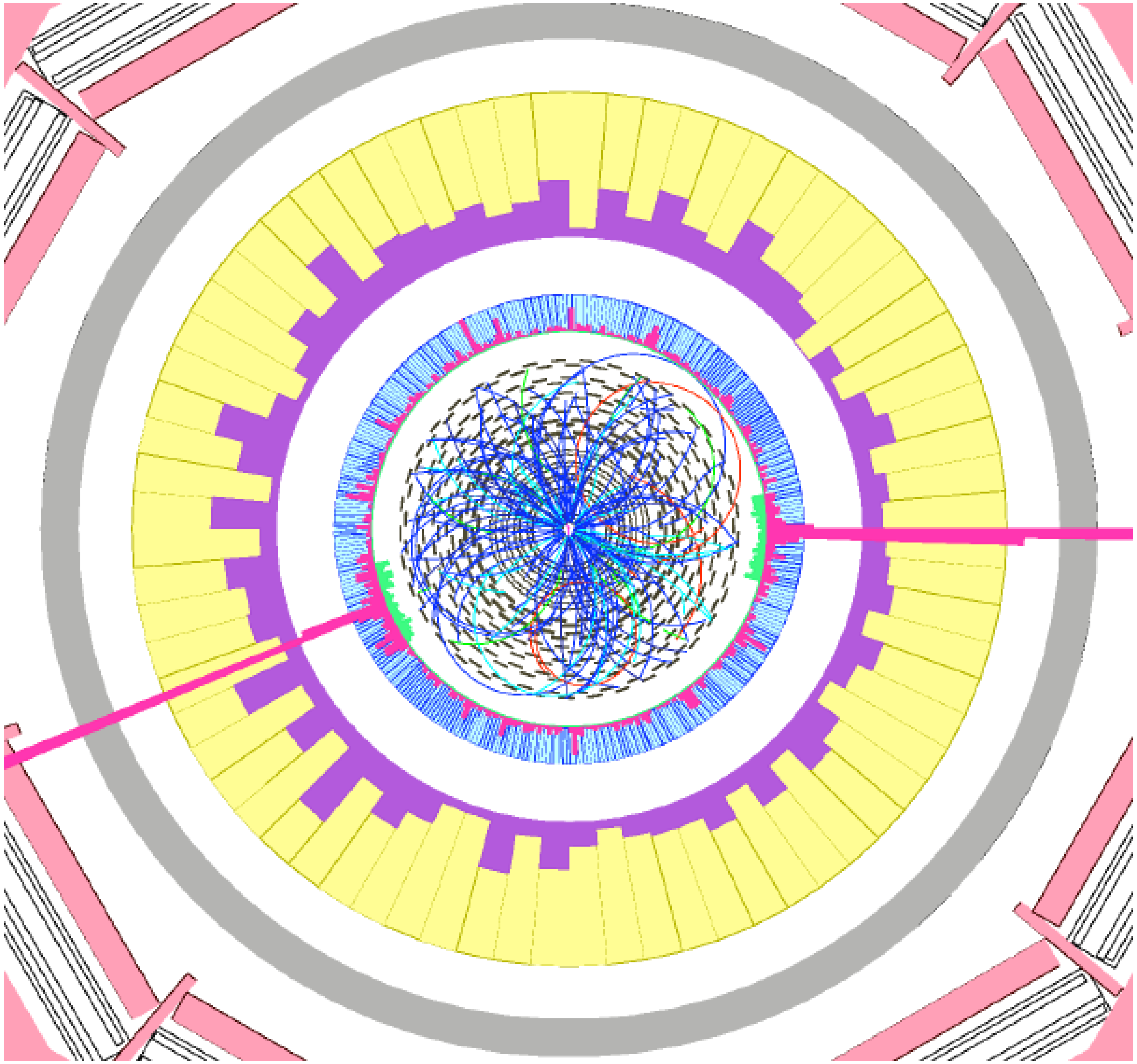
σ = PDFs ⊗ 2→n process ⊗ hadroniszation



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

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

Example: simulated Higgs Decay in CMS



Can you see the Higgs?

Monte Carlo Generators: School

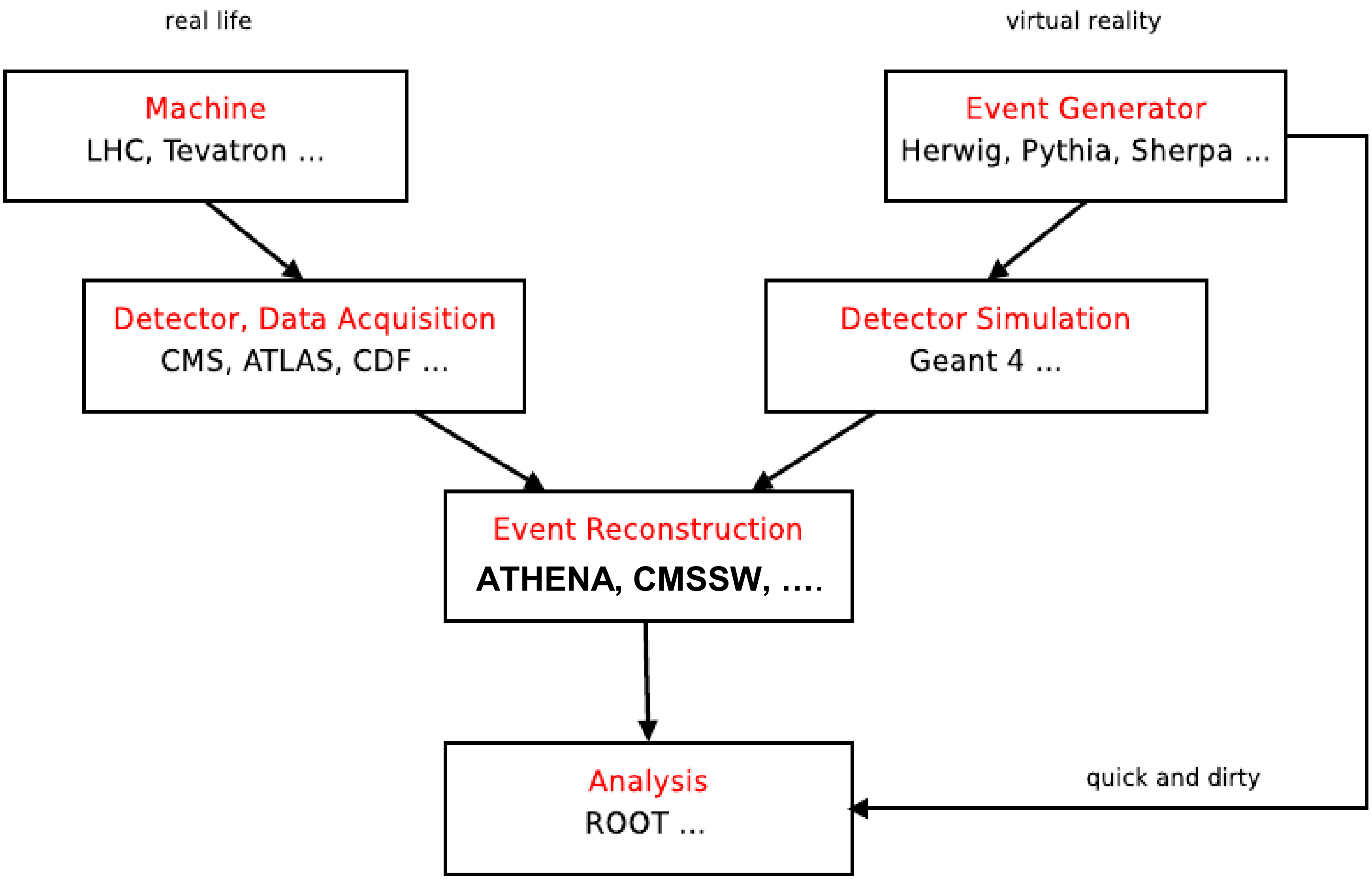
nice lecture, much more detailed than what can be shown here:

Monte Carlo School 2012, Helmholtz Alliance „Physcis at the Terascale“
lecture by Stefan Giesecke, KIT

Technique in particle physics:

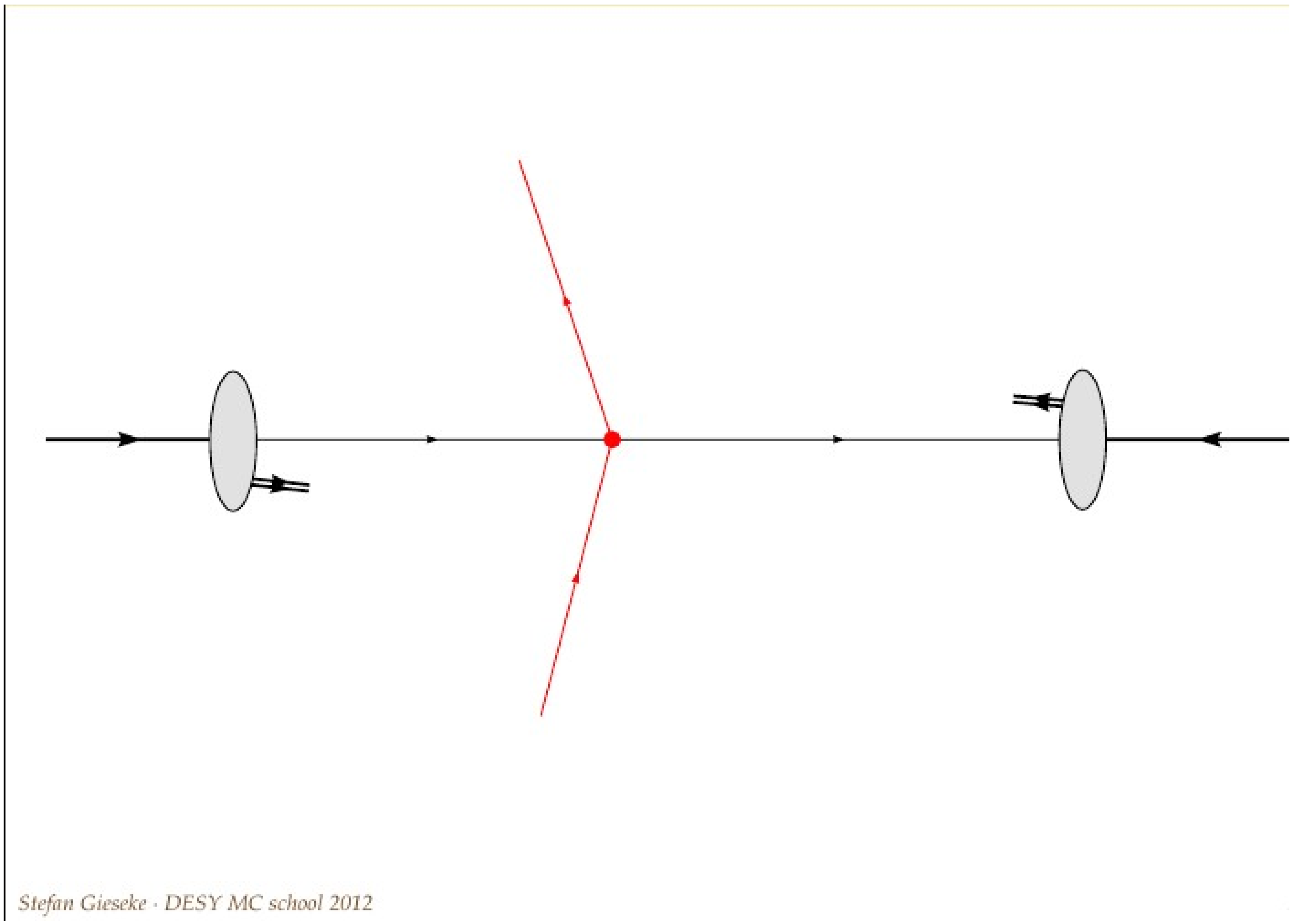
- Generate artificial events
  reflecting all processes in the Lagrangian
  using the Monte Carlo Technique
- obtain arbitrary distributions from simulated final state particles
- and compare with measurements

Steps of MC simulation



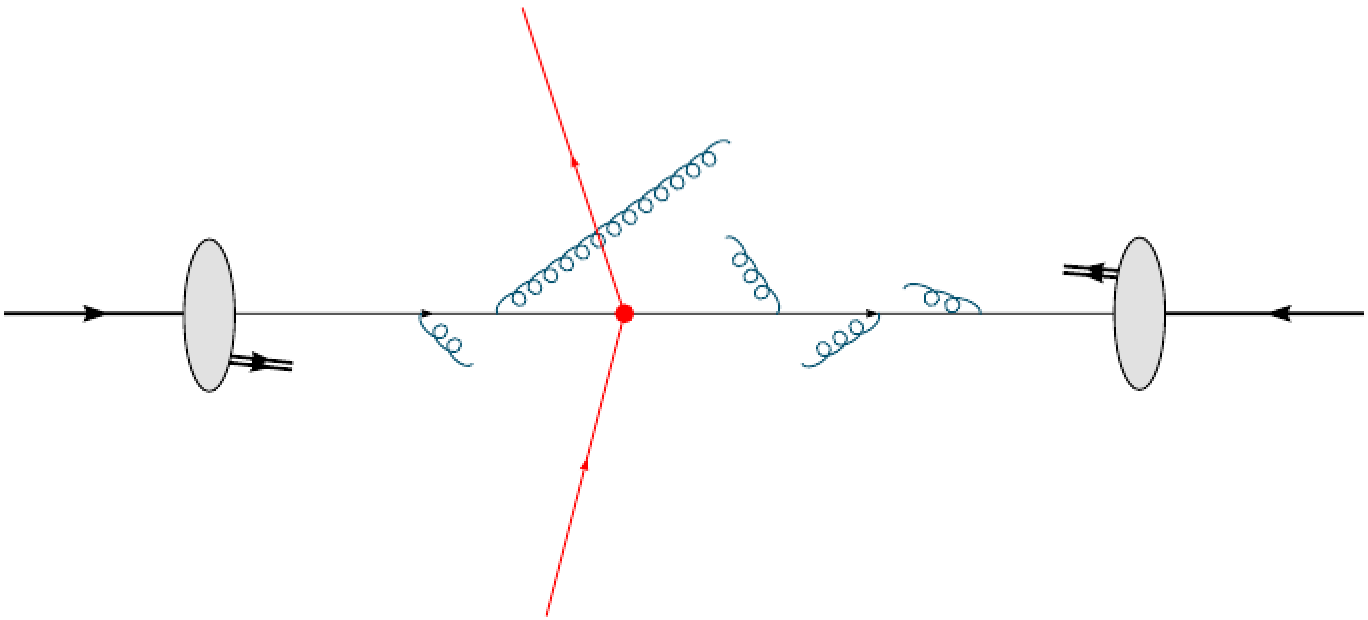
Example: pp collision

Example: pp collision



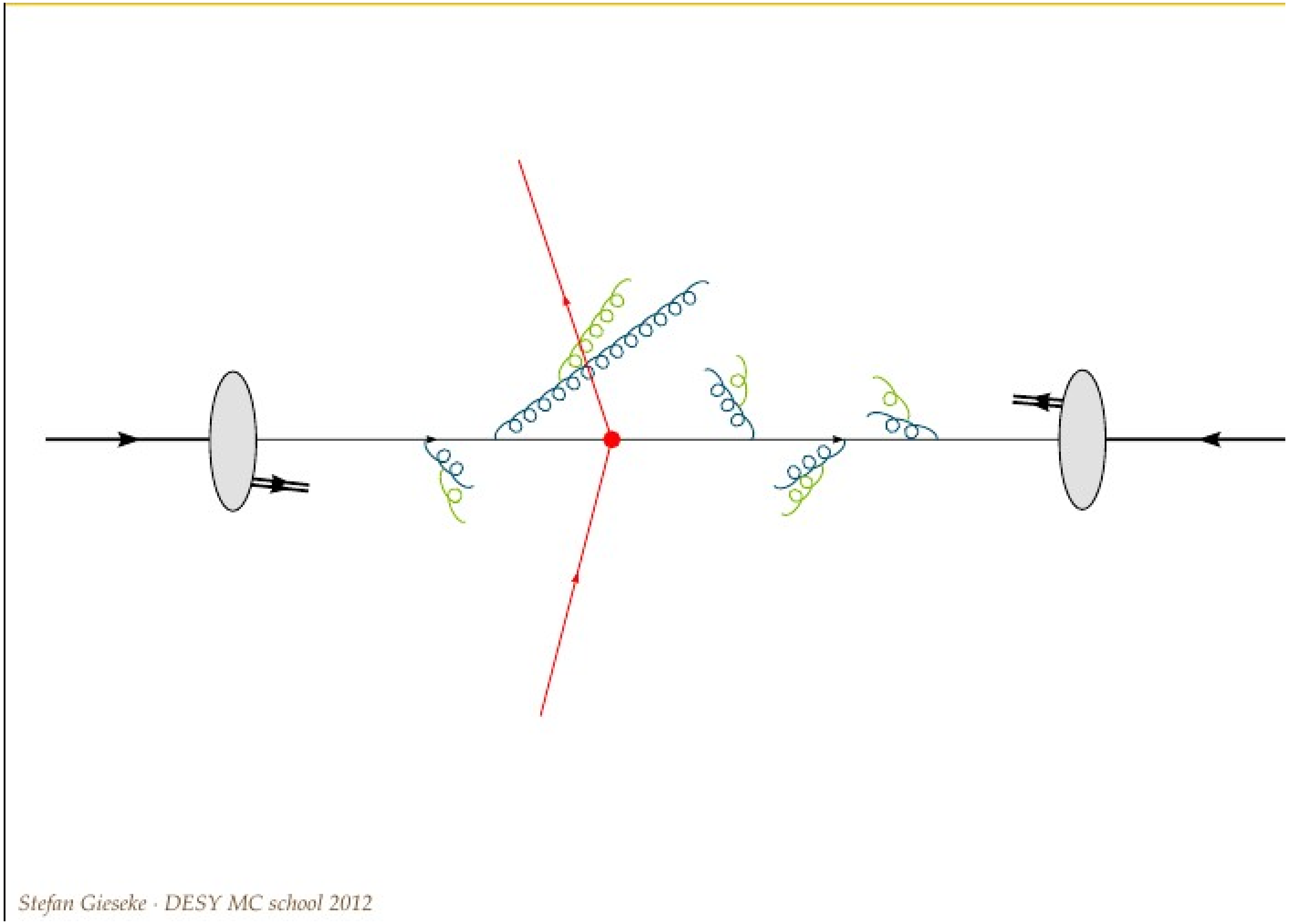
matrix element of hard process

Example: pp collision



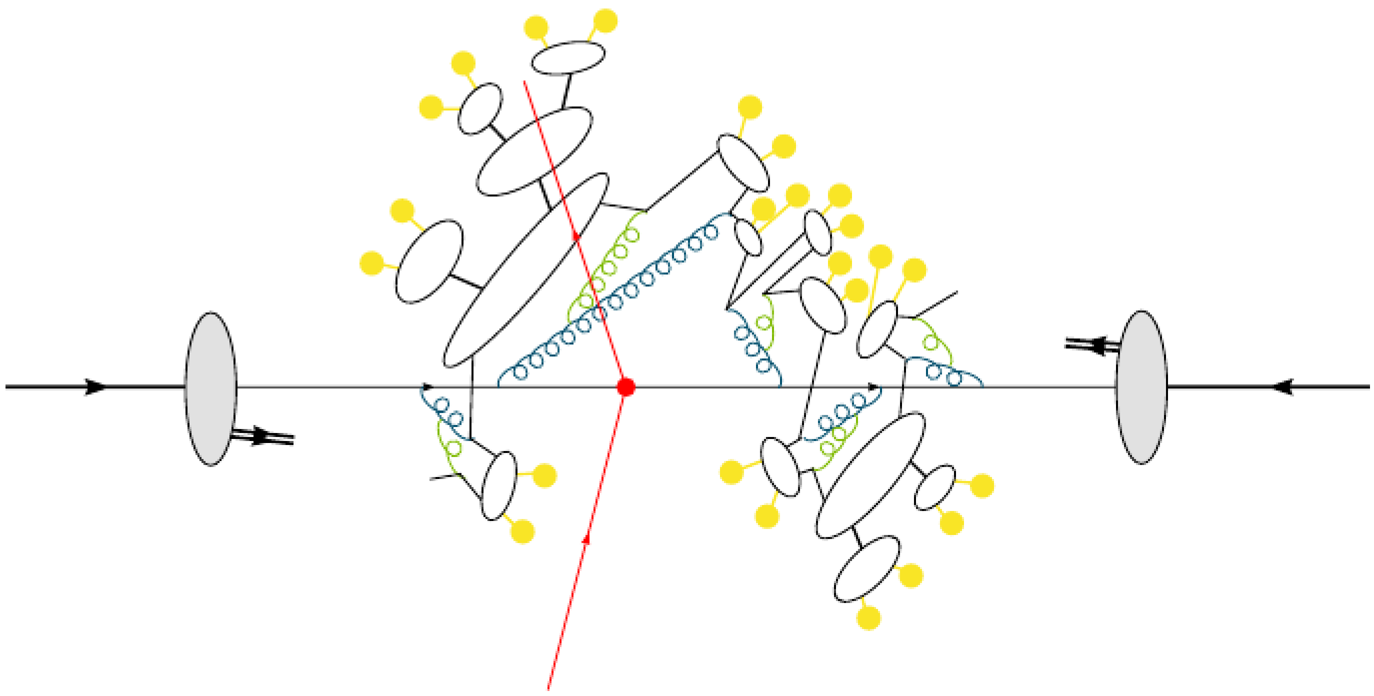
parton shower

Example: pp collision



parton shower

Example: pp collision

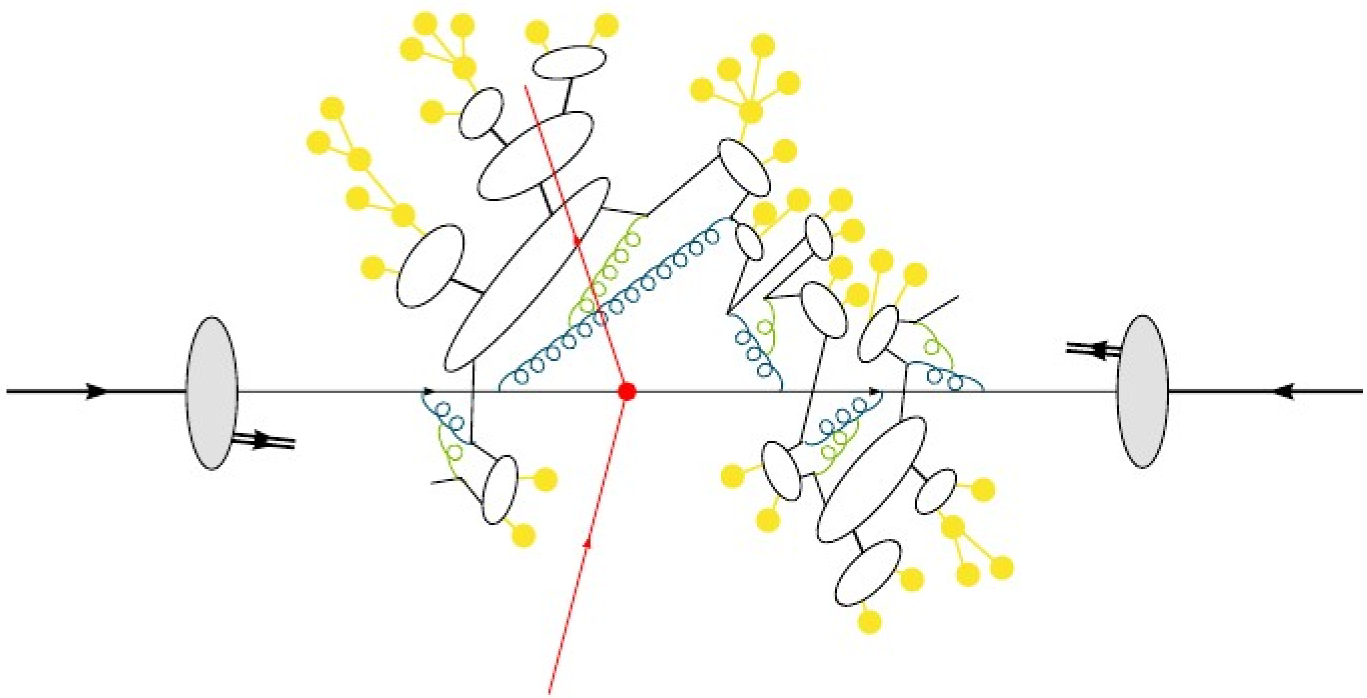


Stefan Gieseke · DESY MC school 2012

hadronization

phenomenological:
Lund string model
(Pythia)
or
cluster hadronisation
(Herwid(++))

Example: pp collision

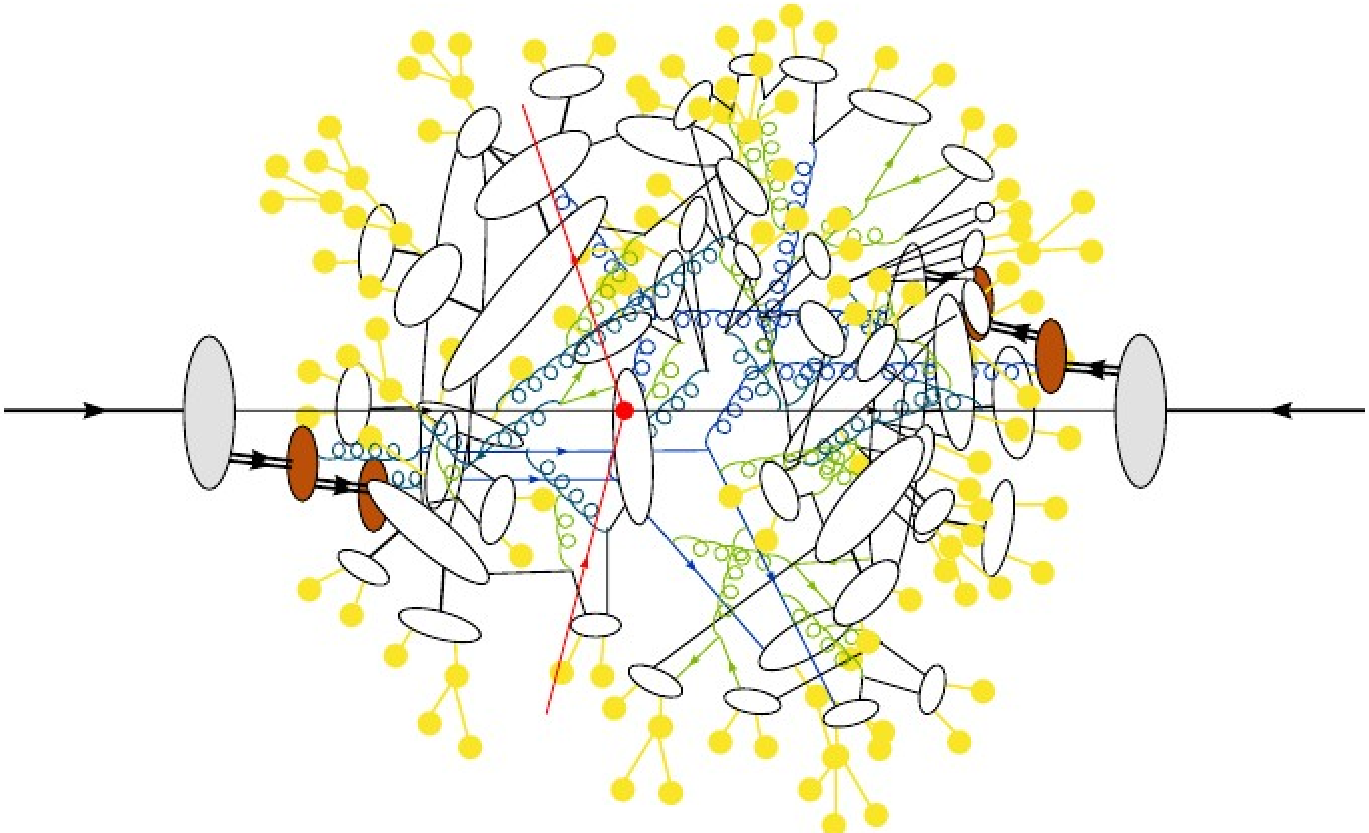


Stefan Gieseke · DESY MC school 2012

hadron decays

tedious -
relies on
measurements

Example: pp collision

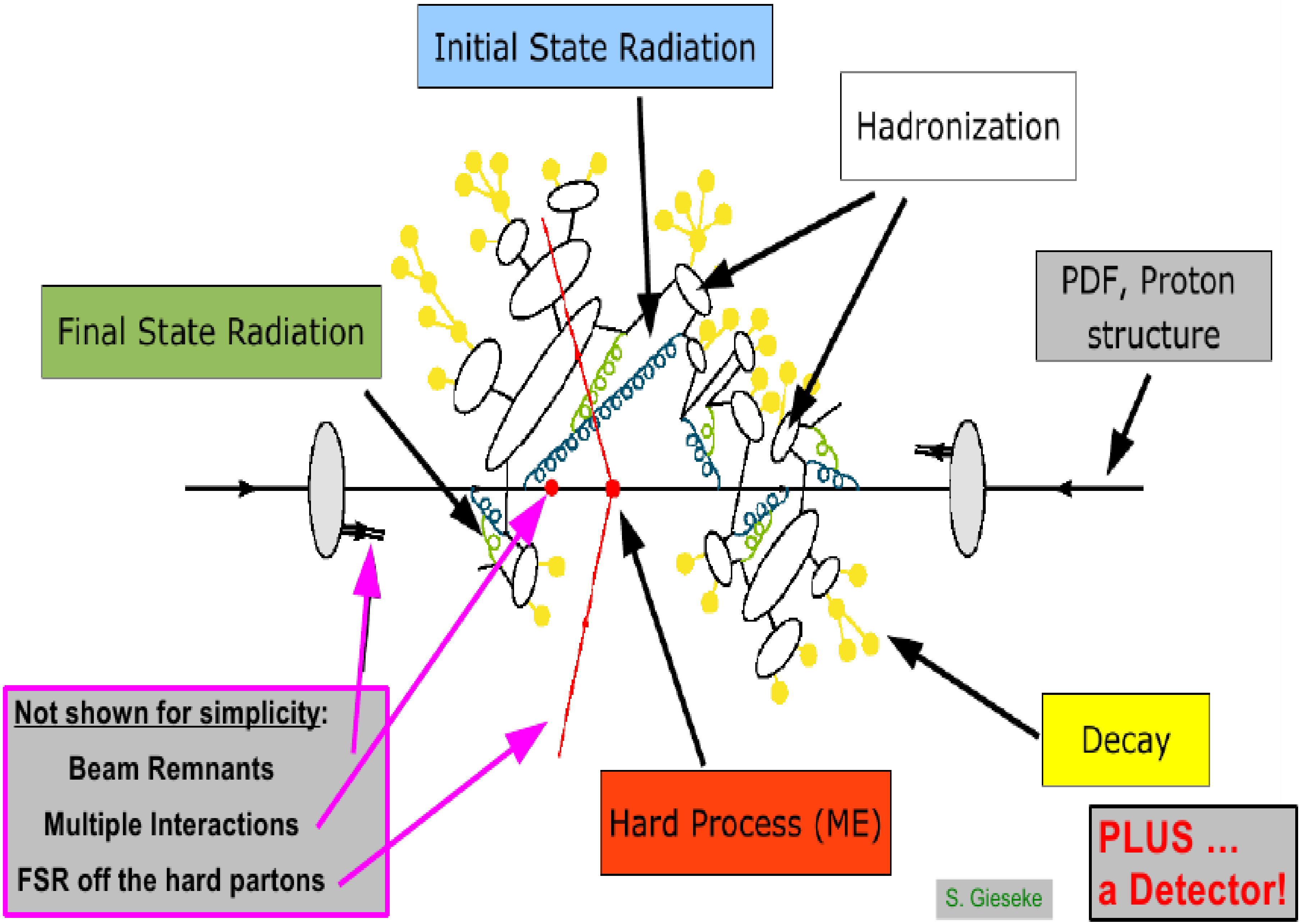


Stefan Gieseke · DESY MC school 2012

relies on models
& measurements
→ needs „tunig“

Multi-parton interactions and underlying event

Summary: pp collision



Example: pp collision

last step:

- process stable particles through detector simulation
  to obtain „hits“ in detector cells;
- run reconstruction software
  to obtain „reconstructed objects“
- run selection procedures („Analysis“)
  to obtain „identified reconstructed objects“

in total:

true properties of objects from hard process at parton level are folded with

- parton distribution functions,
- hadronization effects,
- detector acceptance and efficiency,
- reconstruction efficiency and resolution,
- identification efficiency and purity

to obtain **reconstructed properties**

all steps involve multi-dimensional integrations;
Monte Carlo is the only choice !

Result of Simulation

Distribution(s) of

– signal events

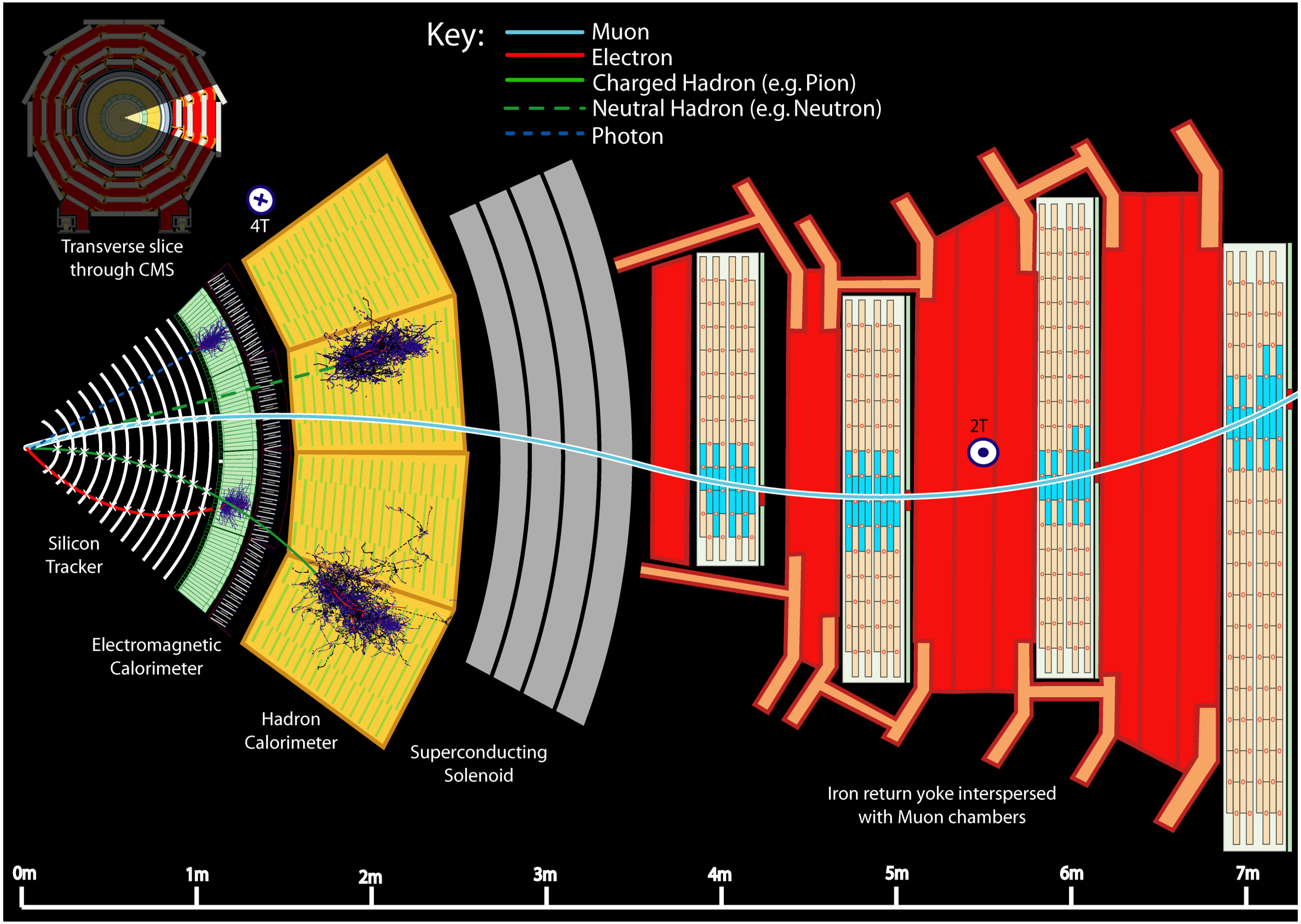
– background events

Used to formulate
„signal+backbground“ (S+B)
and
„background-only“ (B)
hypotheses for comparison
with data and
statistical inference

Hint: in the real experiment, only very small numbers
are expected (see y-axis) – the question will be:
are they best descreibed by the S+B or the B-only shape?

The Real Experiment

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 facotrs:
– 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 minimize

– with signal as large as possible

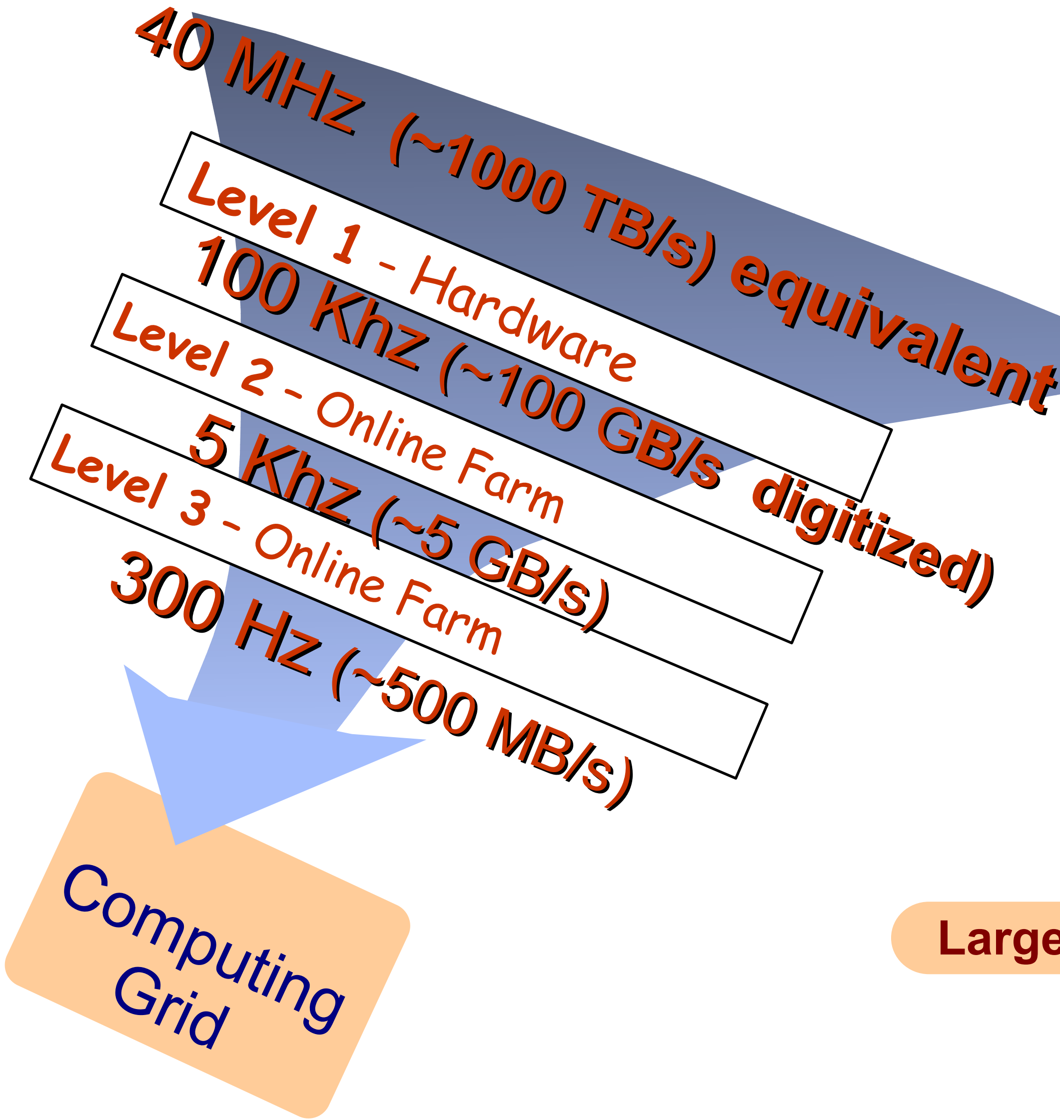
– background as small as possible

– nonetheless, want large efficiency

– luminosity error small (typically beyond your control, also has a “theoretical” component)

Trigger

Online Data Reduction



- ~ 100 millionen 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 Acquisistion

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)

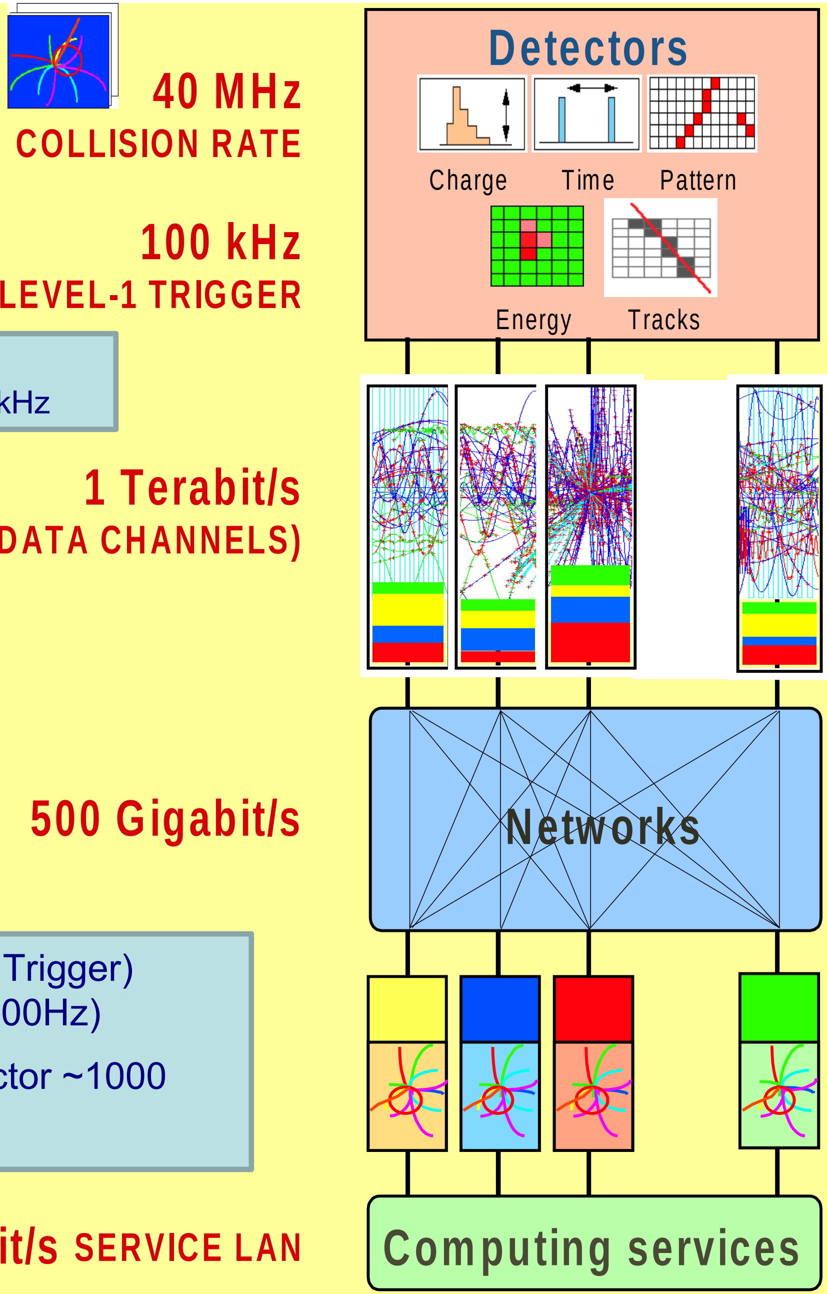
500 Gigabit/s

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

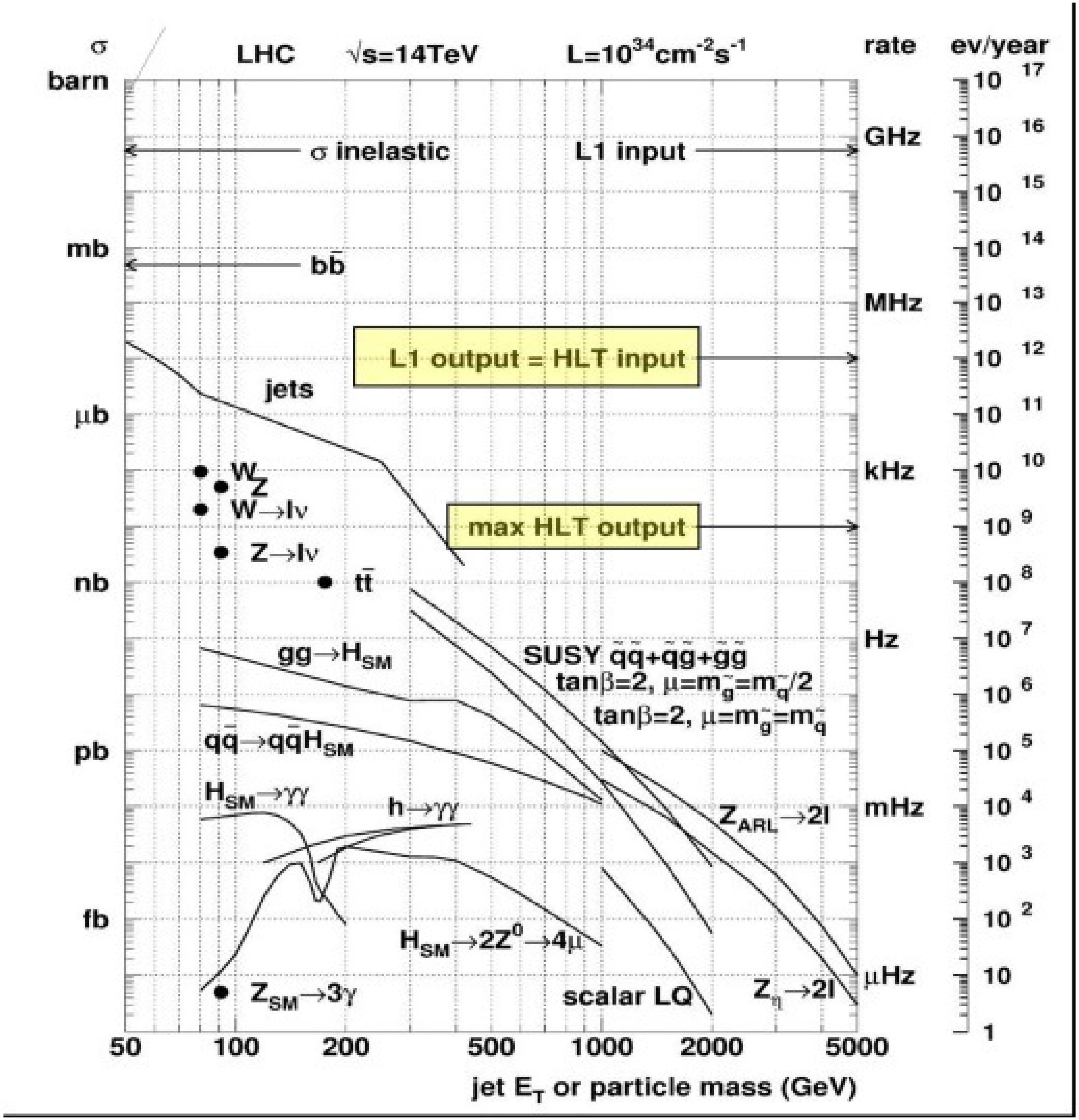
EVENT BUILDER. A large switching network (512+512 ports) with a total throughput of approximately 500 Gbit/s forms the interconnection between the sources (Readout Dual Port Memory) and the destinations (switch to Farm Interface). The Event Manager collects the status and request of event filters and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

What is easy to trigger ?

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

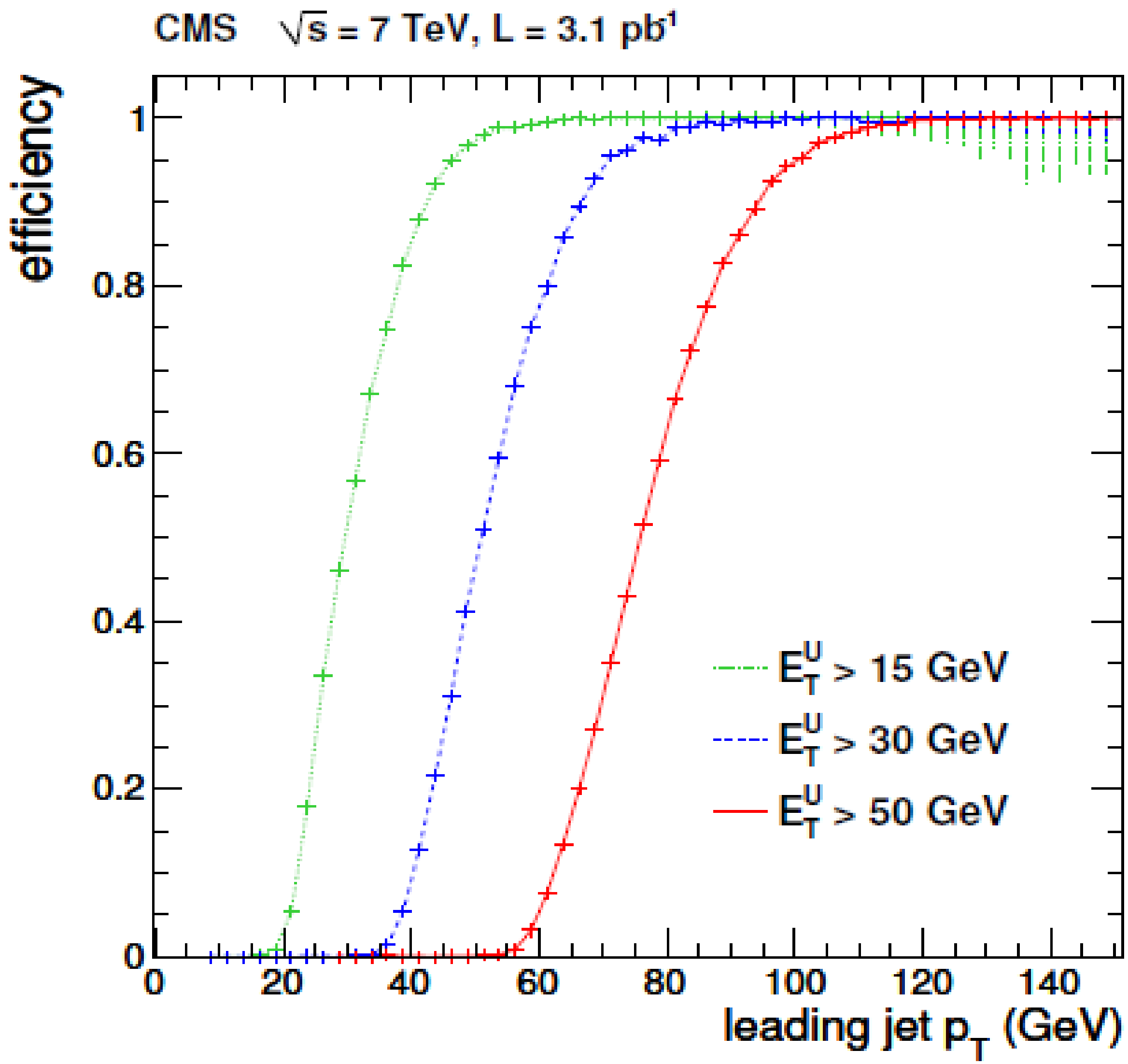
- isolated leptons with large transverse momentum > ~20 GeV
  (from W, Z, top)
- di-lepton events with transverse momentum > ~10 GeV
- jets with very high transverse momentum (several 100 GeV)
- events with large missing energy (~100 GeV)
- isolated photons with transverse energy >~50 GeV

lower-threshold triggers typically pre-scaled

Rest is difficult and probably not in recorded data !

for analysis, must know trigger efficiencies

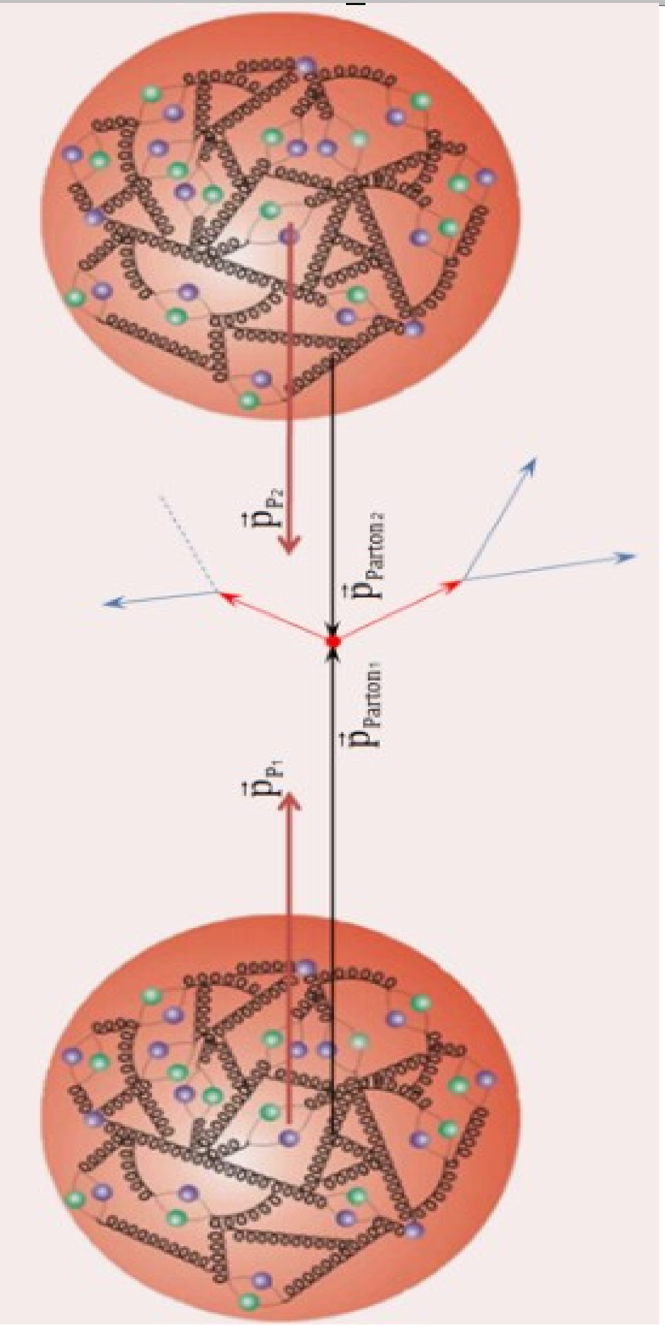
Example: trigger “turn-on” for jets

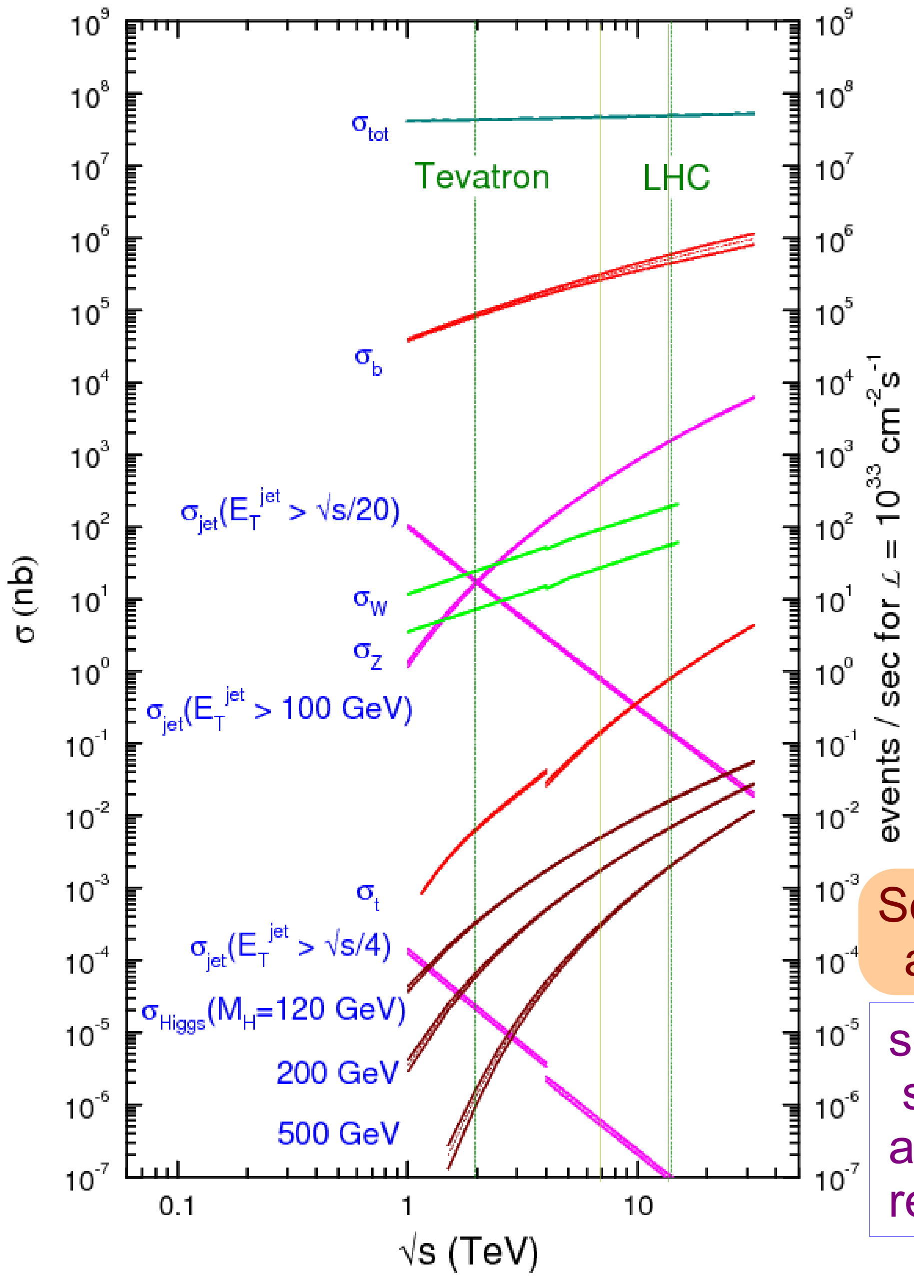


typical knee-shaped trigger efficiency curves (CMS, 2010), rising from 0 to 1

Data Analysis

Event Selection in the Analysis





Some processes are very rare !

sophisticated signal selection and background rejection needed.

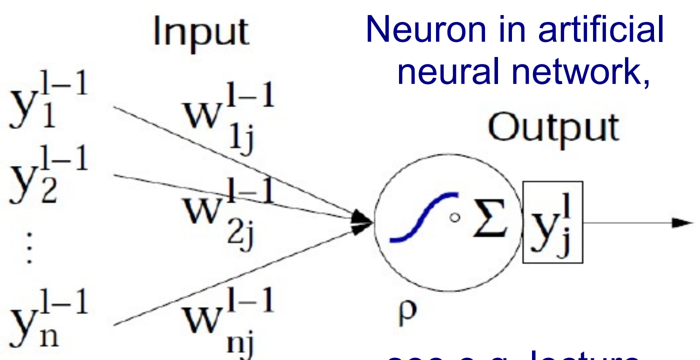
Analysis Steps

- recorded events are **reconstructed**: 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,



see e.g. lecture “Datenanalyse”

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.

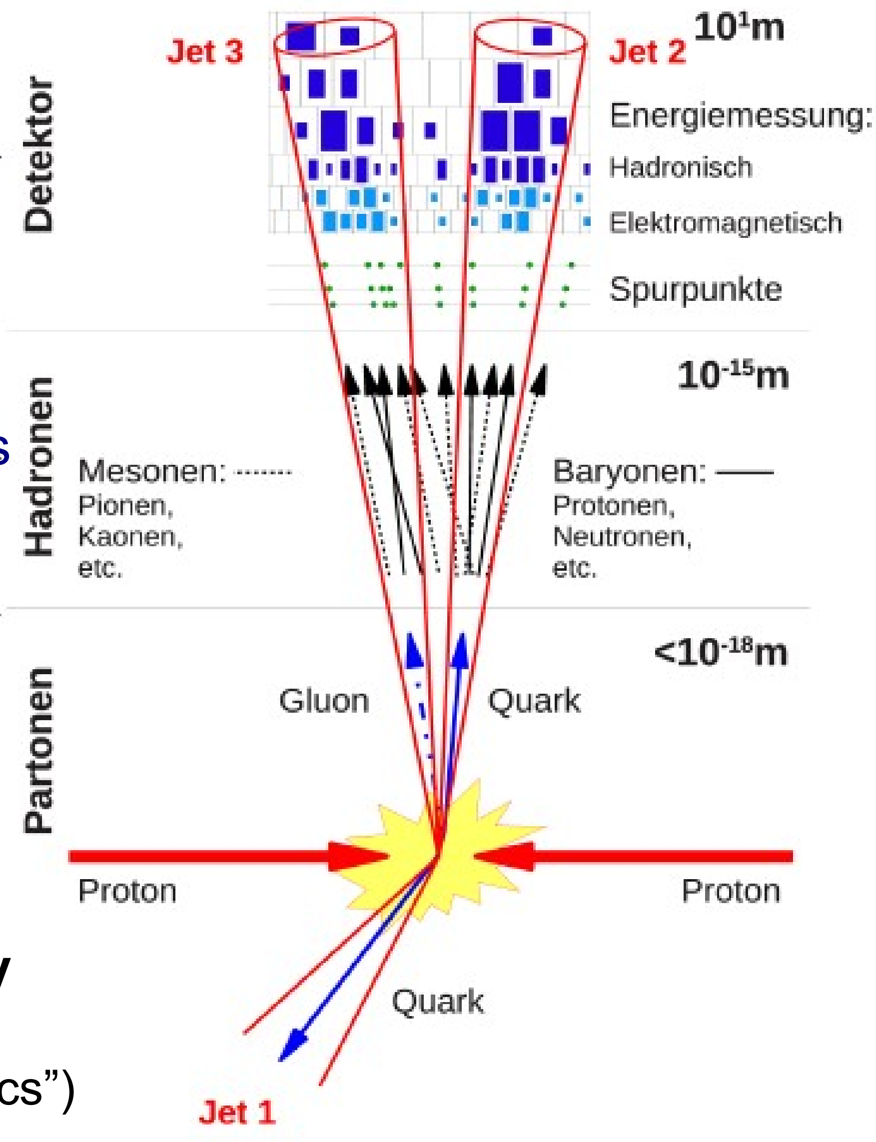
Reconstruction of Objects

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

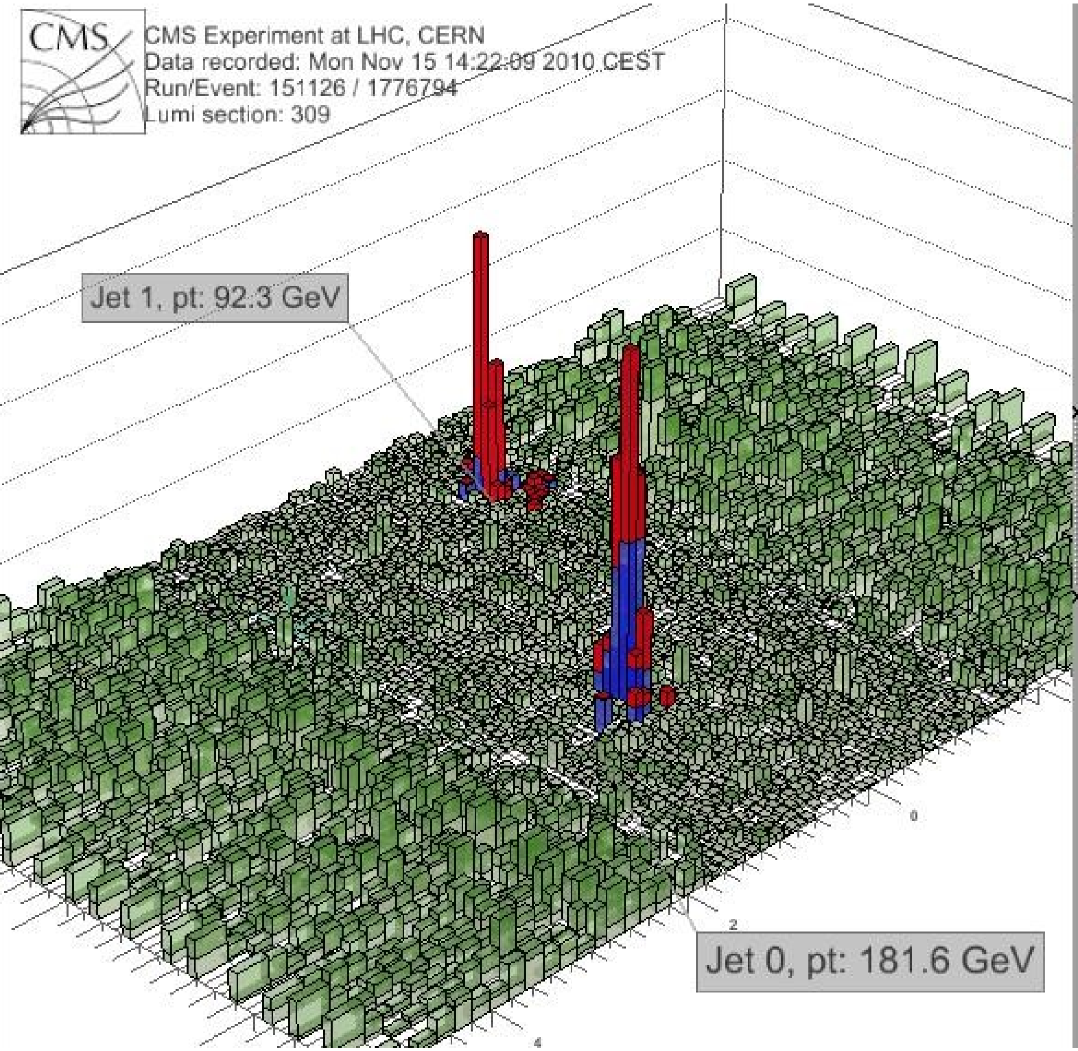
2. **cluster objects** into “jets”
   relation between
   measured final state objects
   & hard partons
   two types of algorithms:
   1. **“cone”**: geometrically assign objects to the leading object
   2. **sequentially combine** closest pairs of objects – different measures of “distance” exist (kT,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

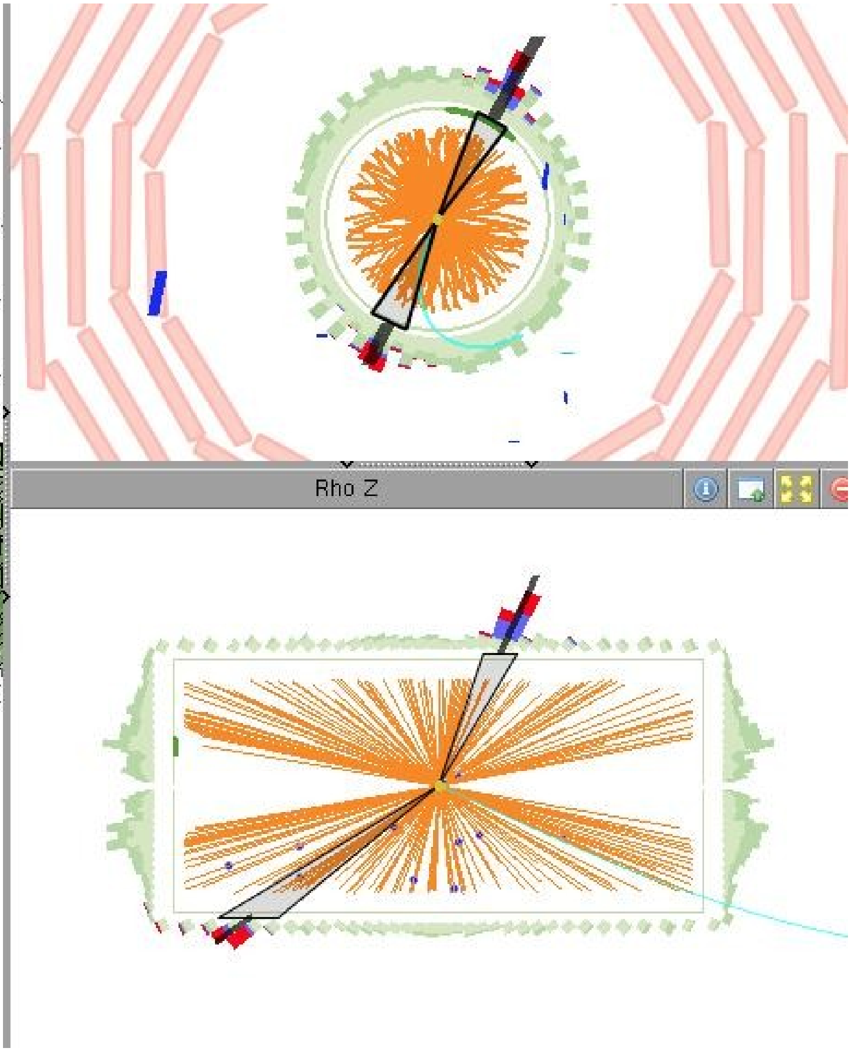
   CMS does this across detector components (“particle flow” analysis)

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

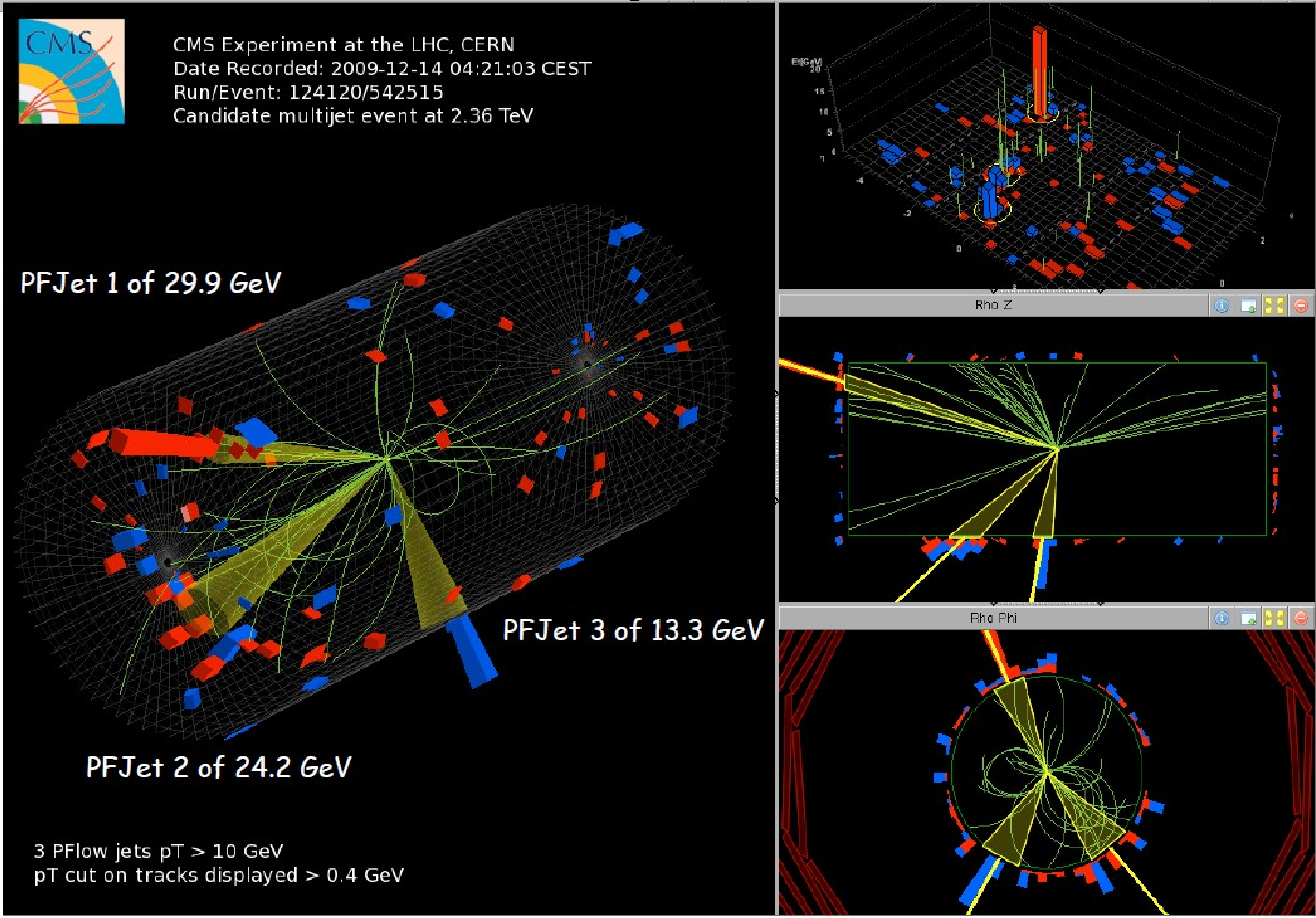


Two-Jet Event in the CMS detector

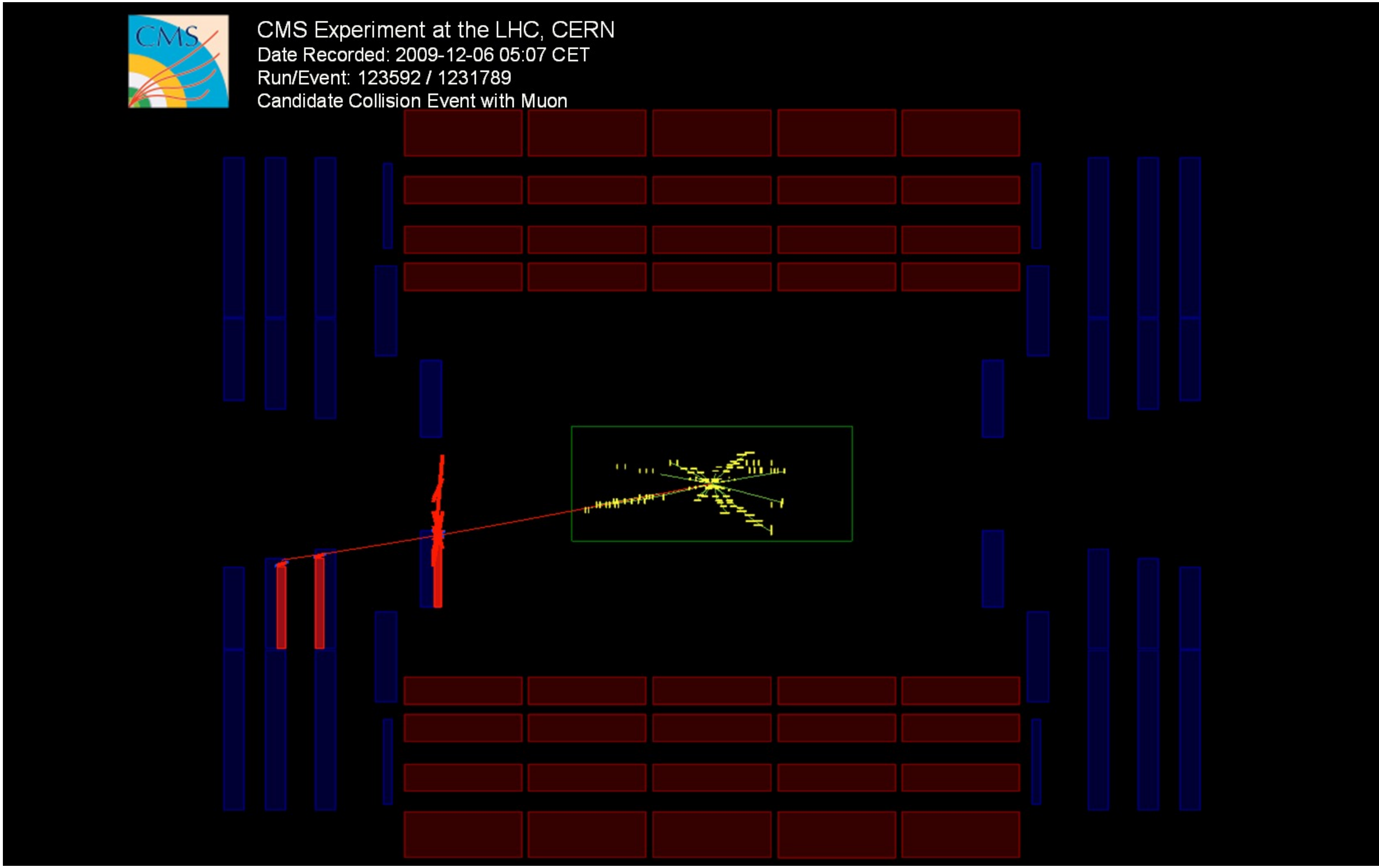




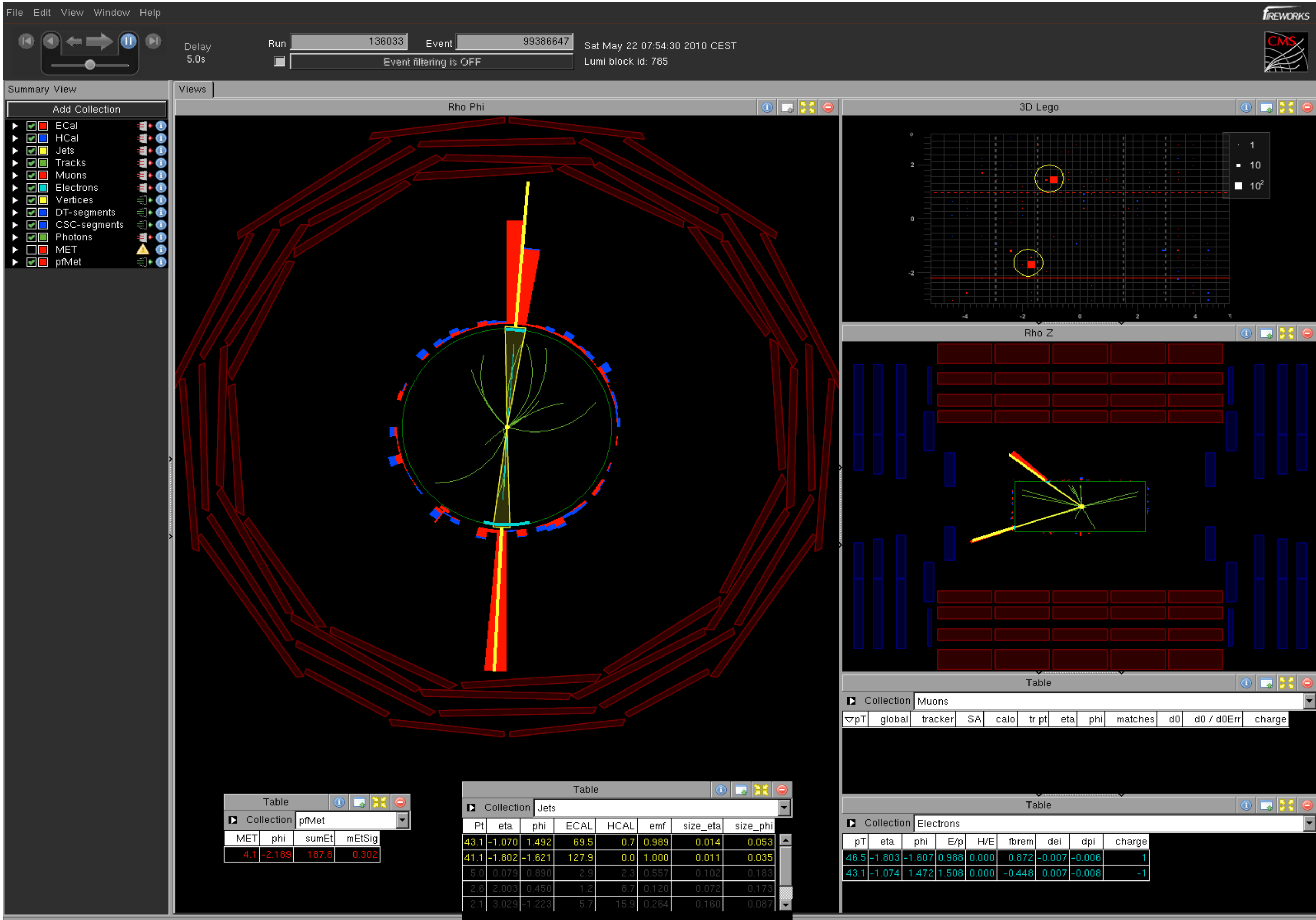
Three-jet event



event with end-cap muon



2 electrons in CMS



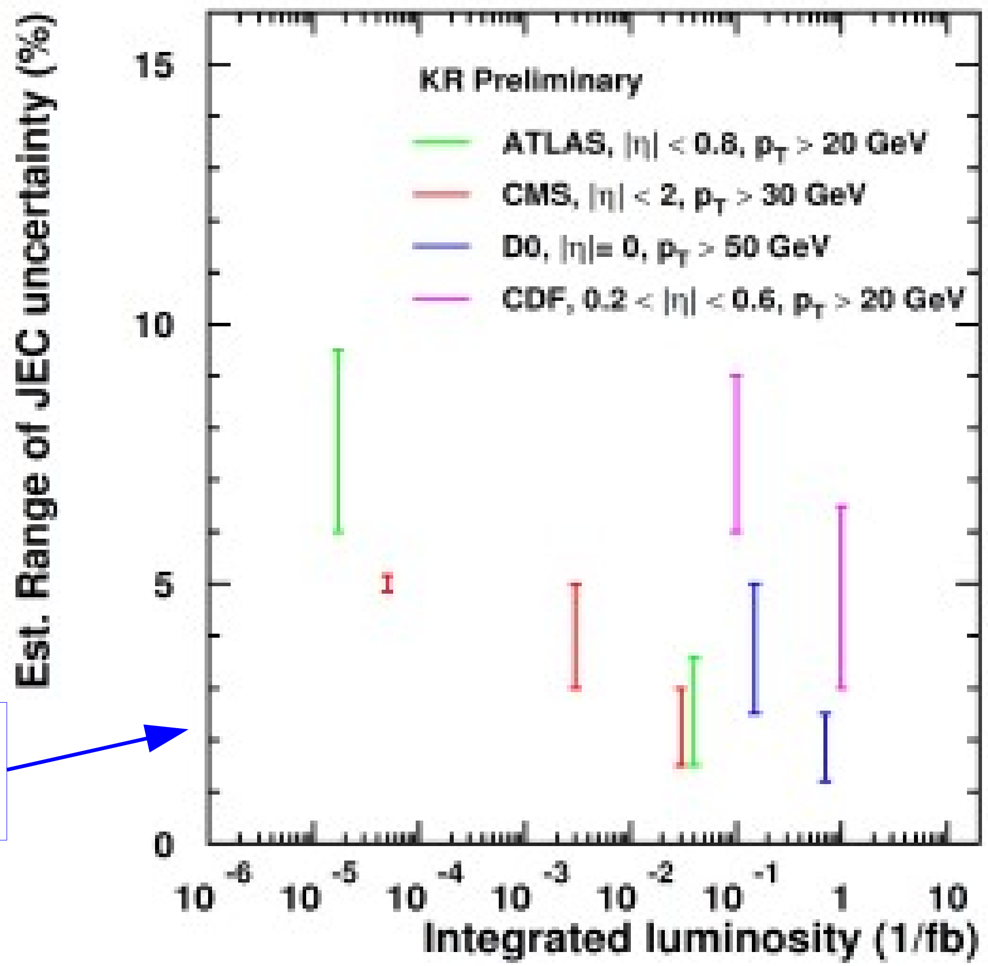
Calibration

Jets and missing transverse energy must be calibrated

relies on special topologies:

- **di-jet events** to
  equalize detector response
- **Z or γ balanced by a jet** to
  determine absolute scale
- events with
  genuine missing energy
  ($Z \to \nu\bar{\nu}$, W, Top)

Precision of Jet energy calibration reaches level of a few % !

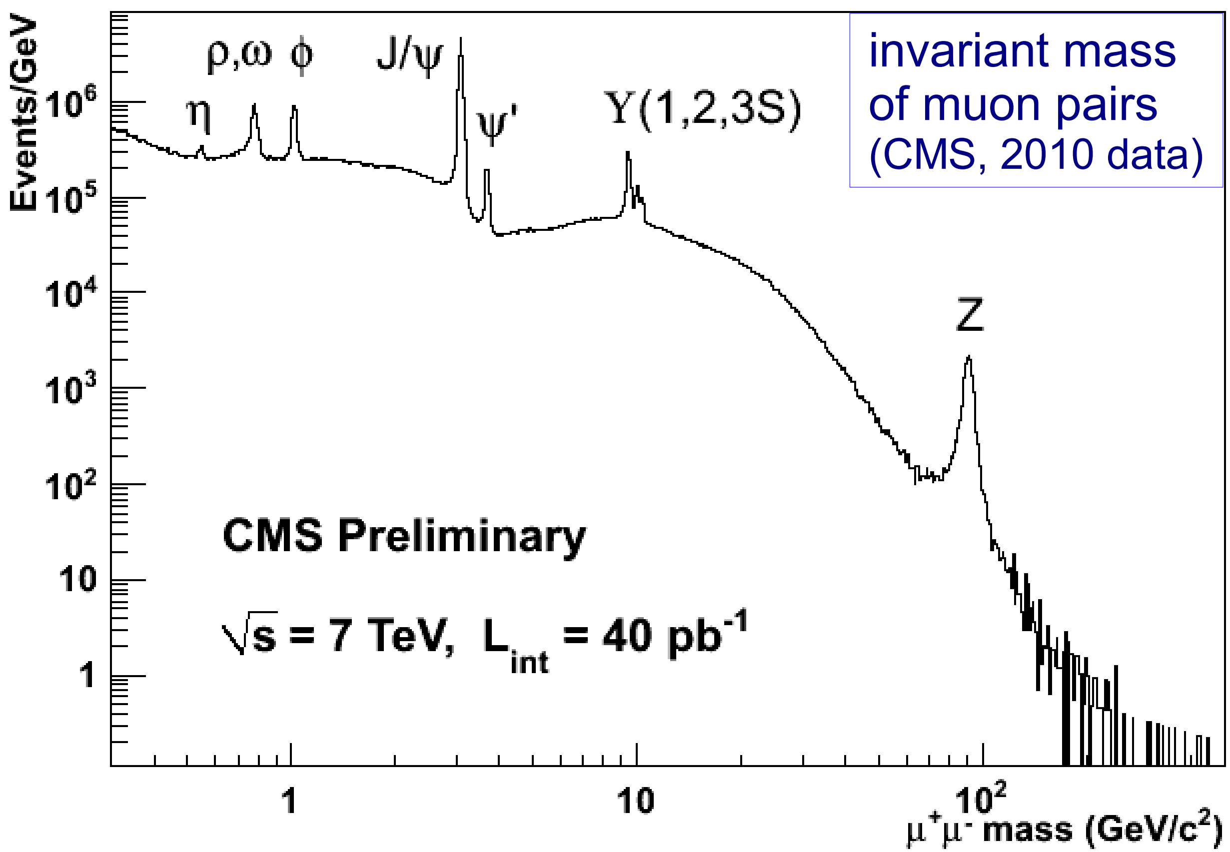


More complicated observables

Calculate **derived quantities** from objects:

– invariant mass of groups of objects
– missing energy or missing transverse momentum
– scalar sum of jet energies
– event shape variables (for QCD analyses)
– all kinds of “classifiers” for event classification

60 years of particle physics in only one year:



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

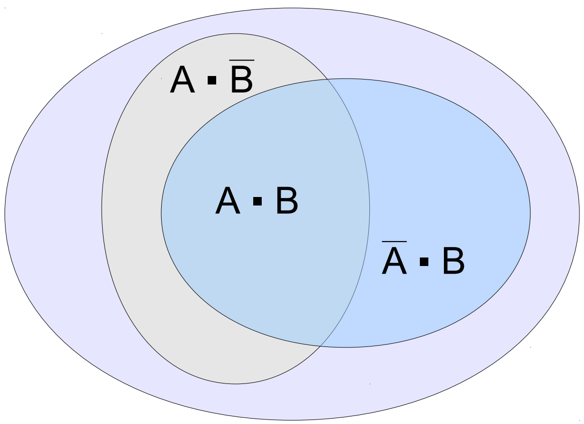
    – tag & probe method:

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

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

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

(statistical errors governed by Binomial distribution)

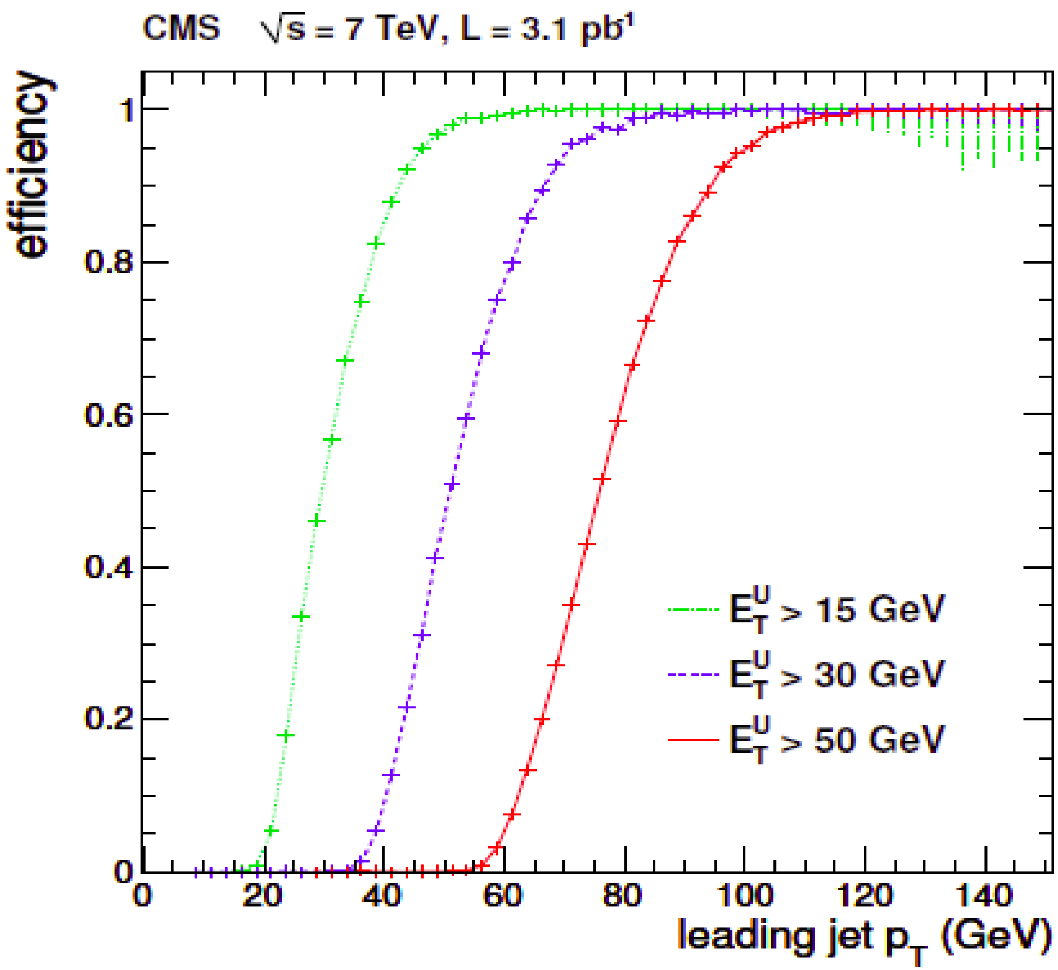


Example: 1 tight muon and one loose muon with tight selection on Z mass (“tag”) allows to measure the selection or trigger efficiency of second muon (“probe”)

Example: Trigger efficiencies

Typical “turn-on” curves of trigger efficiencies

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



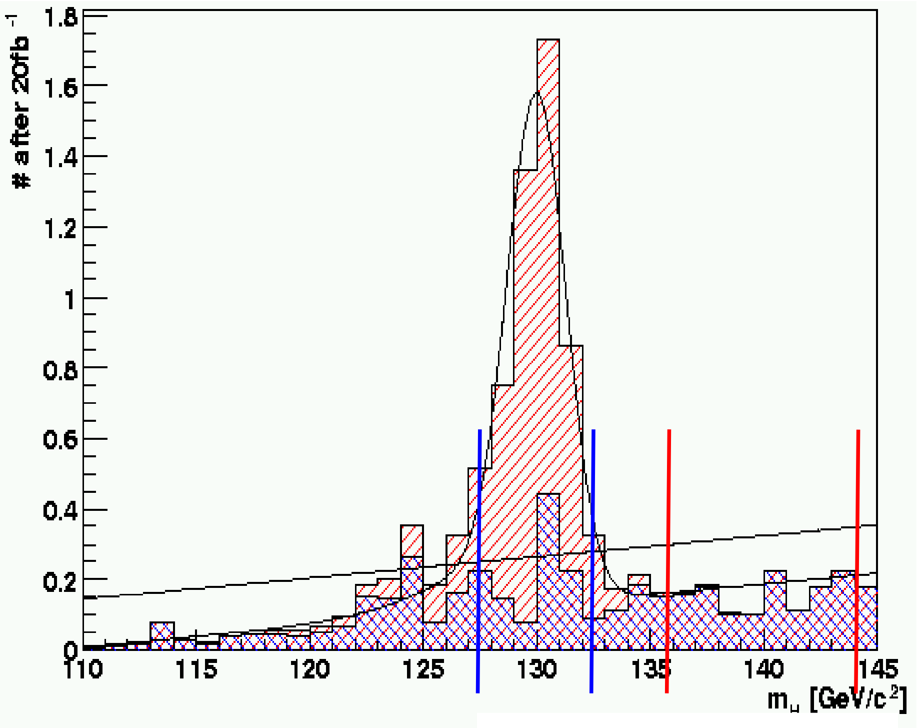
Remarks:

- 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

Determination of background

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

– **extrapolation from "side band"**
assuming "simple" signal shape
or taking signal shape from MC

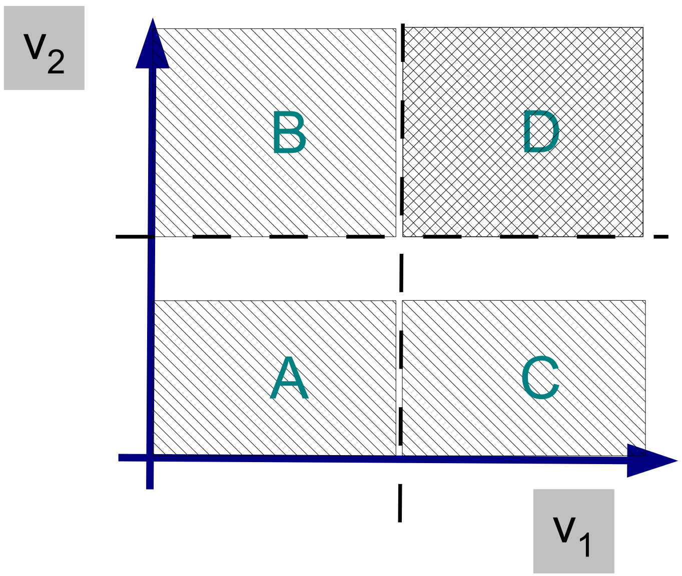


signal background
regions

– **if a second, independent variable**
can be found, background extrapolation 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*

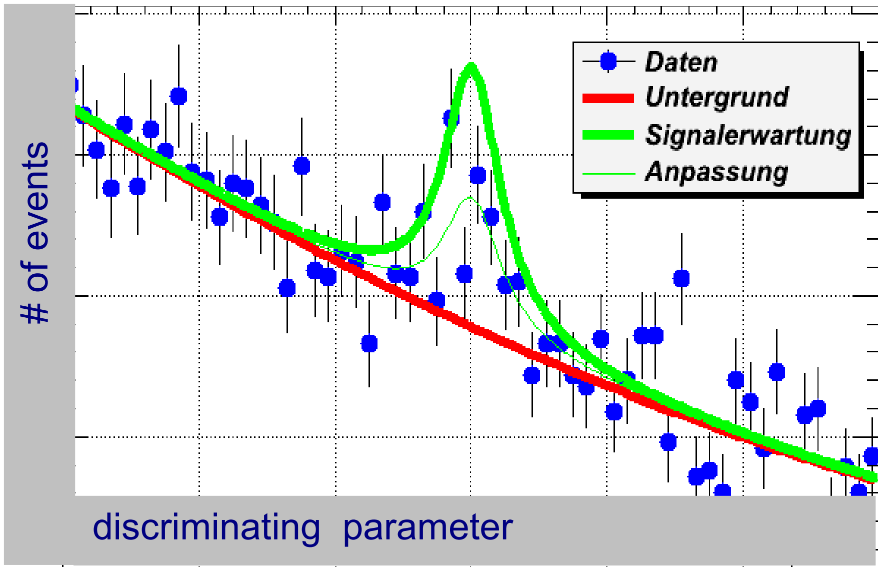
Example: invariant mass of two unlike-sign particles, combinatorial background from sample with like-sign particles.

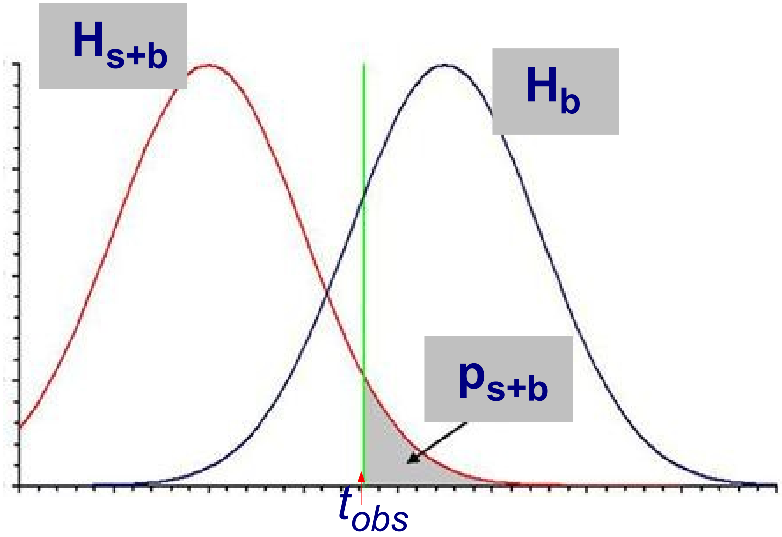
– **more advanced methods** exist to **exploit two uncorrelated variables** to predict the background shape under a signal, see e.g. "sPlot method" in ROOT.

Statistical analysis

The Problem: an excess of observed events can have two sources:

1. signal in addition to expectation
2. a statistical upward fluctuation or insufficient understanding of background distribution (systematic error)





from the statistical view point, a **Hypothesis Test** H_{s+b} vs. H_b

Definition of a suitable teststatistic t as a function of the data: t_{obs}

calculation of probabilities $p_{s+b} = \text{Prob}(t > t_{obs} \mid H_{s+b})$

resp. $p_b = \text{Prob}(t < t_{obs} \mid H_b)$

„p-values“ w.r.t. of S+B resp. B-only hypothesis

To postulate the observation of a **new signal**, background fluctuations must be excluded with very high probability !

Ein Beispiel: Würfel

Dodekaeder,

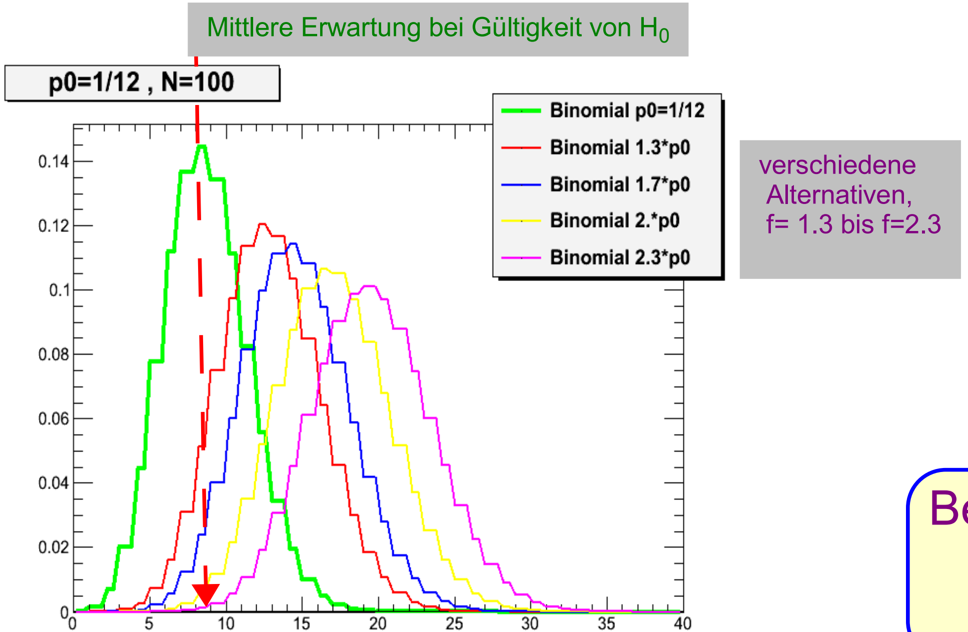
Frage: funktioniert so etwas als „Würfel“ im Spiel ?

Statistischer Test mit einer Zahl (z.B. 12), sollte mit Wahrscheinlichkeit p=1/12 auftreten

Nullhypothese, H_0: $p_{12} = {}^1/_{12}$

Alternative, H_a: $p_{12} = f \times {}^1/_{12}$

Teststatistik: Zahl der geworfenen 12en in N=100 Versuchen, n_{12}, folgt Binomialverteilung $B(n_{12}; N, {}^1/_{12})$



Was können wir über f sagen ?

- Festlegen eines Konfidenzniveaus, z.B. 5%
- Bestimme f_{95}, so dass Fläche links von n_{12} unter der Verteilung 5% ist
- → **$f < f_{95}$ mit 95% Konfidenzniveau „beobachtete Grenze“**

Legen wir statt der Beobachtung n_{12} die rote Linie (Erwartungswert $\langle n_{12} \rangle$ unter H_0) zu Grunde, erhalten wir die **„erwartete Grenze“**

Bedeutung für Higgs-Suche:

$\langle n_{12} \rangle$: erwartete Standard-Prozesse

f : „Signalstärke“ für Higgs-Beitrag

Nuisance Parameters

A complication:

expected signal and background processes and
event selection are affected by additional uncertainties
→ nuisance parameters

need to treat these properly in the interpretation of the
parameter of interest (e.g. the production rate of a new process)

Two main methods exist:

- *profile likelihood*
- *marginalization*
  of the Bayesian posterior probability density

Nuisance parameters with profile likelihood

Minimization of **- 2 ln $\mathcal{L}$(x| θ , λ_i) w.r.t. all λ_i**

for different values of the parameter of interest **θ** (e.g. # of signal events)

1. **profile likelihood**

provides all relevant quantities:

Best estimator: $\hat{\Theta}$

at minimum of likelihood

Sqare of signal significance:

$$-2(\ln \mathcal{L}(N_s = 0) - \ln \mathcal{L}(N_s = \hat{\Theta})$$

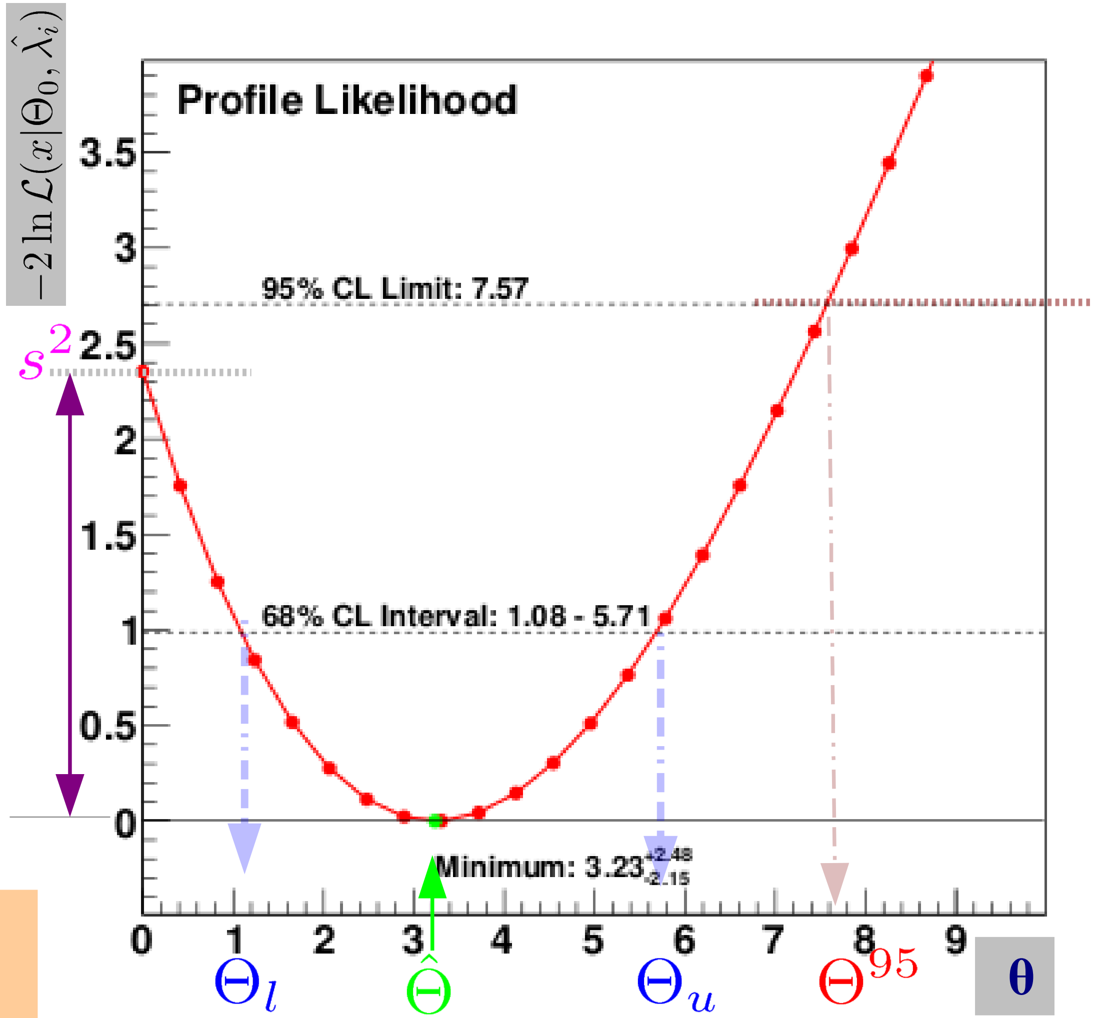
$$= 2 \ln \frac{\mathcal{L}(N_s = \hat{\Theta})}{\mathcal{L}(N_s = 0)} = 2 \ln Q = s^2$$

68% confidence interval:

$$-2\Delta \ln \mathcal{L}(\Theta_{l,u}) = 1$$

95% limit:

$$-2\Delta \ln \mathcal{L}(\Theta^{95}) = 1,645^2 = 2,71$$

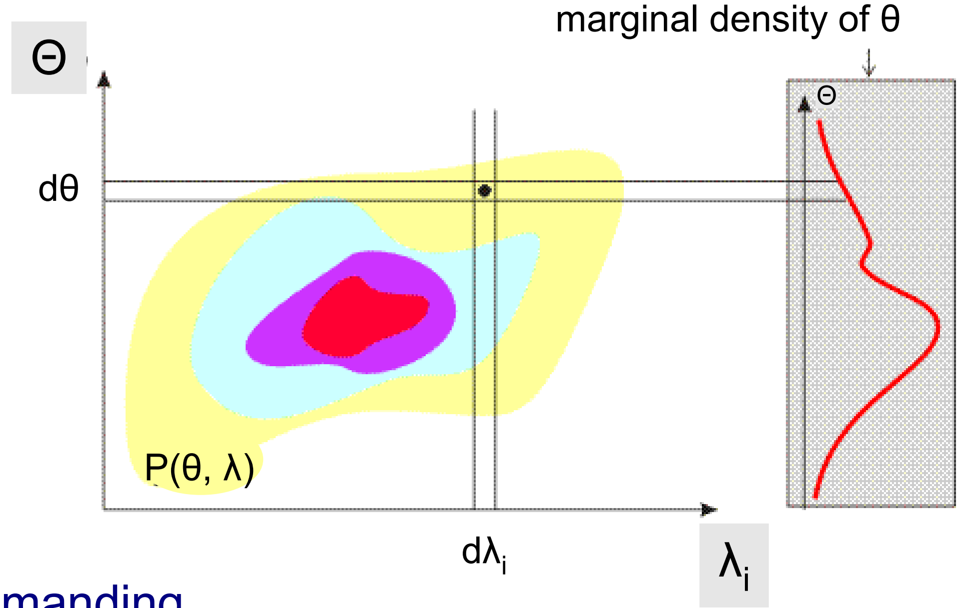


Problem, if minimum near or close to physical boundary
- unphysical for number of signal events

Nuisance parameters via marginalisation

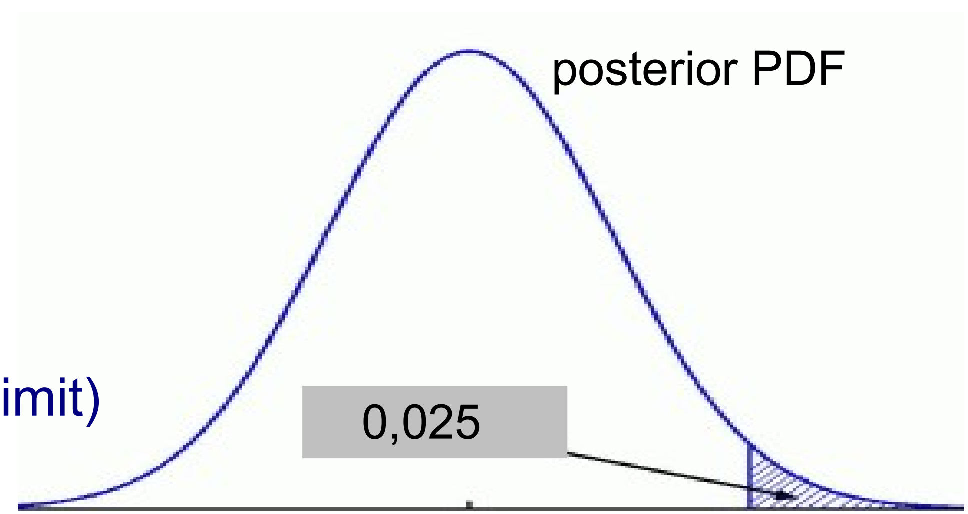
Integrate out the nuisance parameters in the Bayesian posterior probability density:

$$P(\theta) = \int \ldots \int P(\theta, \lambda_i)\, d\lambda_1 \ldots d\lambda_n$$



Multi-dimensional integral may be very demanding,
Markov-Chain Monte Carlo often used

confidence intervals, in Bayesian statistics often called **"credibility intervals"**,
by determination of the appropriate quantiles of the marginalized posterior PDF
(e.g. one-sided upper limit)



Teststatistic for LHC Higgs search

for Higgs search: use **likelihood ratio** as teststatistic

profile likelihood w.r.t. the signal strength μ (μ=0: no signal, μ=1: nominal signal) normalized to the global maximum of likelihood

teststatistic **for limits**

signal strength

Set of nuisance parameters (= systematic uncertainties)

condition to ensure μ ≥ 0 and a one-sided limit

$$q_\mu = -2\ln\frac{\mathcal{L}(data|\mu,\hat{\Theta}_\mu)}{\mathcal{L}(data|\hat{\mu},\hat{\Theta})}\ ,\quad 0 \le \hat{\mu} \le \mu$$

best-fit values of all parameters

data distribution

^ := value of parameter that maximizes likelihood

$$\mathcal{L}(\mathrm{data}|\mu,\Theta) = \prod \mathrm{Poisson}(\mathrm{N_i}|\mu\cdot s_i(\{\Theta\}) + b_i(\{\Theta\}))\cdot p(\{\tilde{\Theta}\}|\{\Theta\}))$$

*determination of the **distribution of q_μ , $f(q_\mu\,|\,\mu)$** , for background (μ=0) resp. signal hypothesis (μ≠0) , via pseudo-experiments or asymptotic formulae in the limit of large data sets*

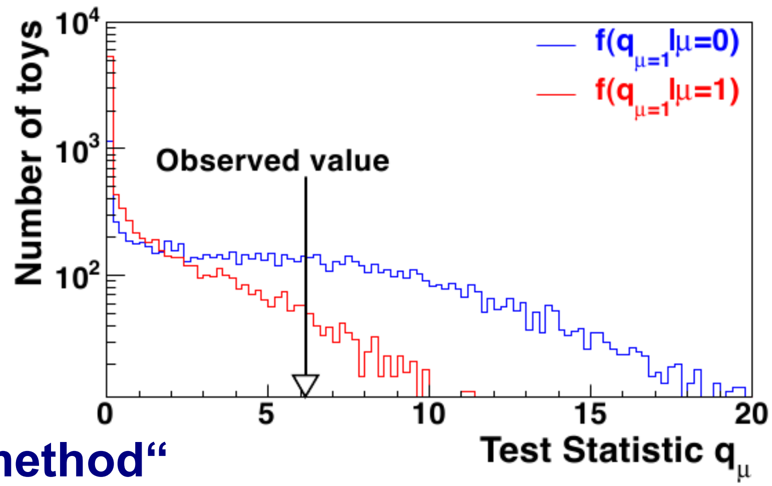
Statistical Analysis

next:

- determination of p-values:

$$p_\mu = \mathrm{Prob}(q_\mu > q_\mu^{\mathrm{obs}} | \mu \cdot s + b)$$

$$p_0 = \mathrm{Prob}(q_\mu > q_\mu^{\mathrm{obs}} | b)$$



- Calculation of confidence level with **„CL_s method“**

$$\mathrm{CL_s} = \frac{p_\mu}{p_0}$$

robust against down-ward fluctuations of background

CL_s quantifies agreement with signal hyphthesis

- for μ=1, CLs = α a Higgs Boson is excluded with confidence level (1- α)
  convention: α=0.05, exclusion 95% CL.

- **usually**: specify value of μ that is excluded at 95% CL

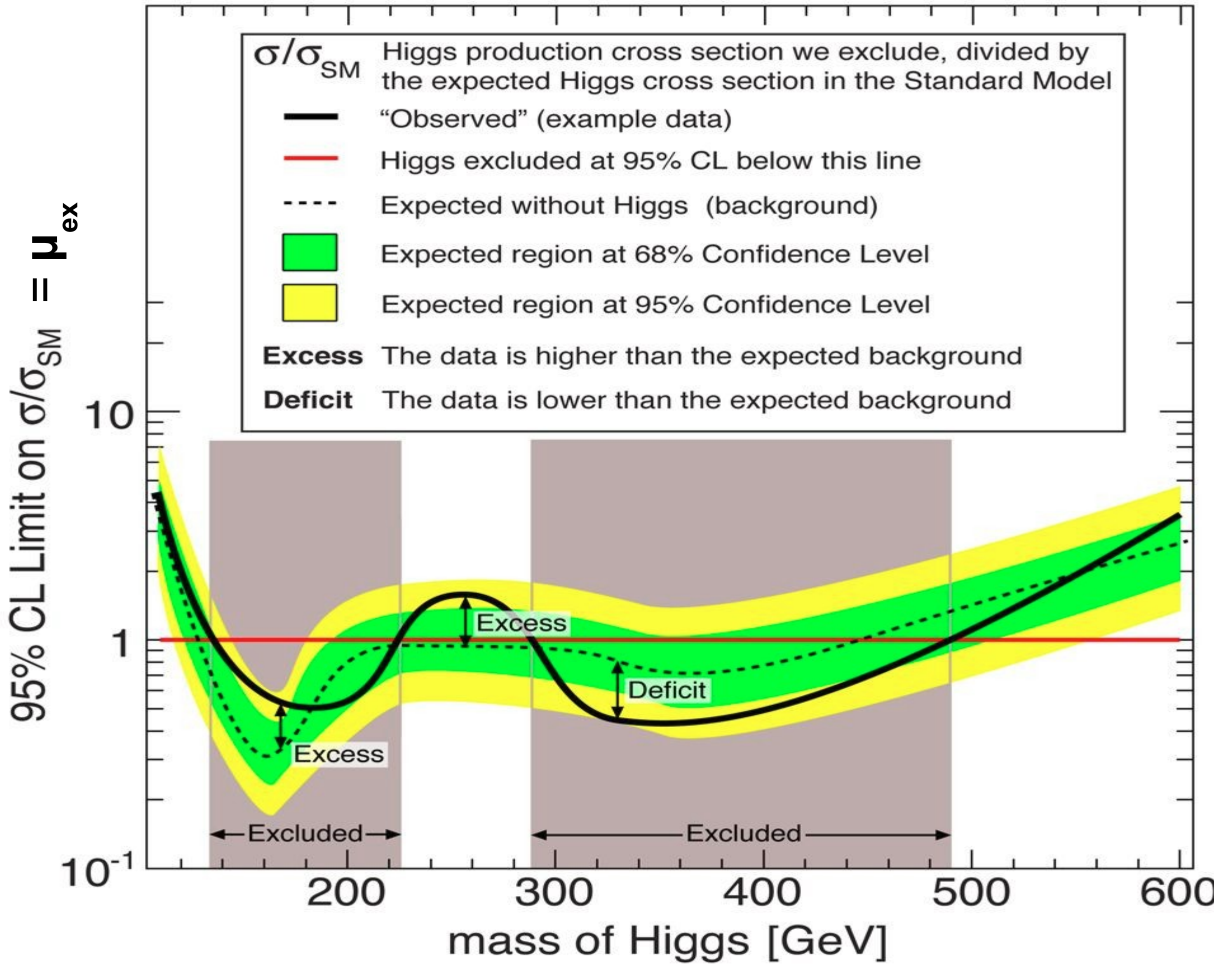
perform pseudo experiments to determine **expected limit,**
i. e. the median of the distribution of obtained limits (dashed line),
and the regions for 68% („1 σ“, green band) and 95% („2σ“, yellow band)

Statistical Analysis

- repeat all of the above for different values of Higgs mass

⇒

- Plot of the signal strength μ_{ex} excluded at **95% CL**
  – for data (black line)
  – for the median and 68%- und 95%-regions of the limit distribuion obtained from the pseudo experiments
  (dashed line and green resp. yellow bands)



Higgs boson excluded
for $\mu_{ex}<1$ (below red line, grey bands)

Statistical Analysis: Significance of a Discovery

If a signal cannot be excluded, what is the "significance" of a possible discovery?

Concept of "local significance":

assume that N events have been observed over an expected background N_b

→ number of signal events: $N_s = N - N_b$

compare N_s with statistical fluctuations of backbround (N_b):

in Gaussian limit (N_b > ~50)

$$\boxed{S = \frac{N_s}{\sqrt{N_b}}}$$

– measure of a signal excess in terms of "number of sigmas" ("z-value")

"The observed signal is S times larger than the standard deviation of the expected background fluctuations"

Statistical Analysis: Significance of a Discovery

typically, N_s and N_b are small numbers if a signal is just being discovered;

→ use Poisson statistics

$$P_0(N; N_b) = \frac{1}{N!} N_b^N \mathrm{e}^{-N_b}$$

background-only hypothesis

$$P_1(N; N_s + N_b) = \frac{1}{N!} (N_s + N_b)^N \mathrm{e}^{-(N_s+N_b)}$$

s+b hypothesis

and take logarithm of likelihood ratio:

$$2\ln\left(\mathcal{L}_1 - \mathcal{L}_0\right) = 2\ln Q = 2\left(N\ln\left(1+\frac{N_s}{N_b}\right) - N_s\right)$$

and assume $N_s = N - N_b$

"in the asymptotic limit"

2 lnQ can directly be interpreted as the z-value of the observation!

(as it is the twice the difference in log-likelihod for $N_s = N - N_b$ and $N_s = 0$)

works also if observation is made in many bins,

$N \rightarrow \sum N_i$: $2 \ln Q = \sum 2 \ln Q_i$

full treatment: define suitable test statistic ($q=q_0$ instead of $q=q_\mu$ for LHC),

determine distribution of q under signal and background hypotheses

and calculate p-values.

If a signal cannot be excluded, what is the "significance" of a possible discovery?

Concept of "local significance":

assume that N events have been observed over an expected background N_b

→ number of signal events: $N_s = N - N_b$

compare N_s with statistical fluctuations of backbround (N_b):

in Gaussian limit (N_b > ~50)

$$\boxed{S = \frac{N_s}{\sqrt{N_b}}}$$

– measure of a signal excess in terms of "number of sigmas" ("z-value")

"The observed signal is S times larger than the standard deviation of the expected background fluctuations"