ЛАБОРАТОРИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ имени М.Г. Мещерякова

Joint Institute for Nuclear Research

SCIENCE BRINGS NATIONS TOGETHER

Computing in High Energy Physics

Sergei Shmatov (MLIT JINR, Dubna)
shmatov@jinr.ru shmatov@cern.ch

MLIT, JINR, Dubna
2 December, 2024

What Do Particle Physicists Do?

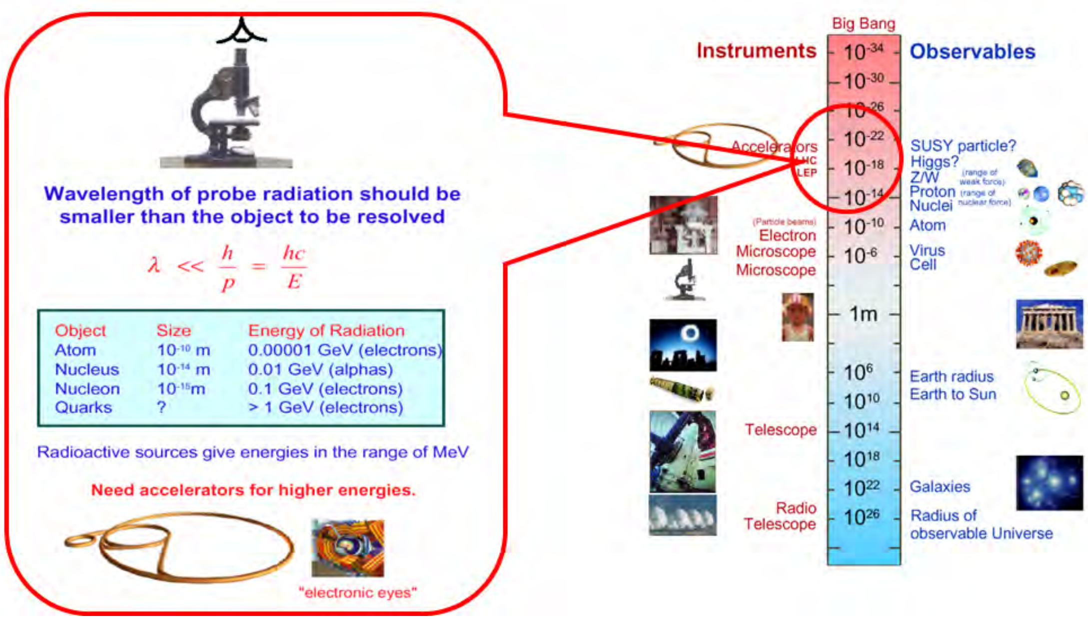
People have long asked,

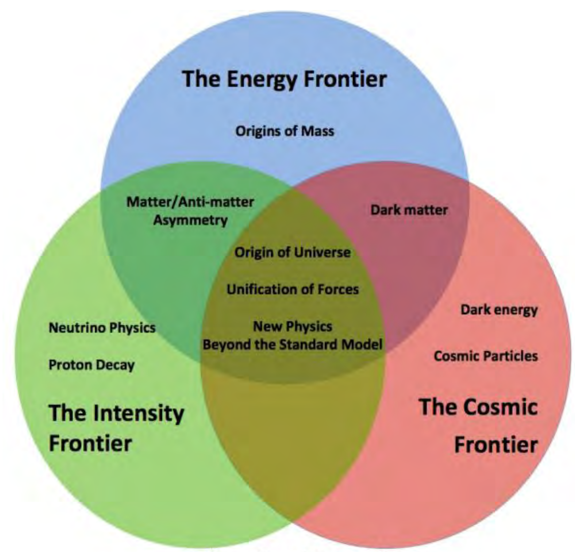
- "What is the world made of?"
- "What holds it together?"

Some eternal questions

Physicists hope to fill in their answers to these questions through the analysis of data from High Energy Physics experiments

Particle physics have focused on the inner space frontier, pursuing the questions of the construction of matter and the fundamental forces at the smallest scale accessible.





Physics Frontiers

Do Physicists Really Need Computers?

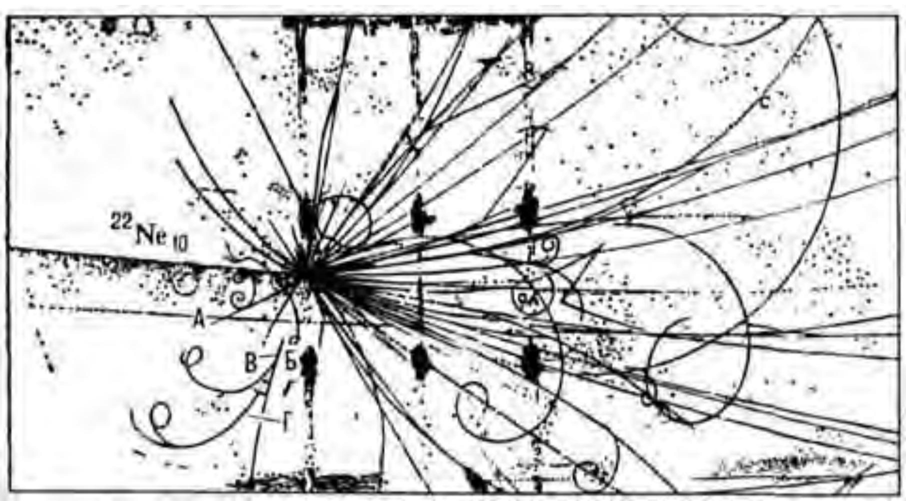
Computing is any goal-oriented activity requiring, benefiting from, or creating computing machinery (c) Wiki

- general-purposed devices (computers/laptops/mobile..) and software
- dedicated tools

Bubble chamber, Synchrophasotron (JINR),

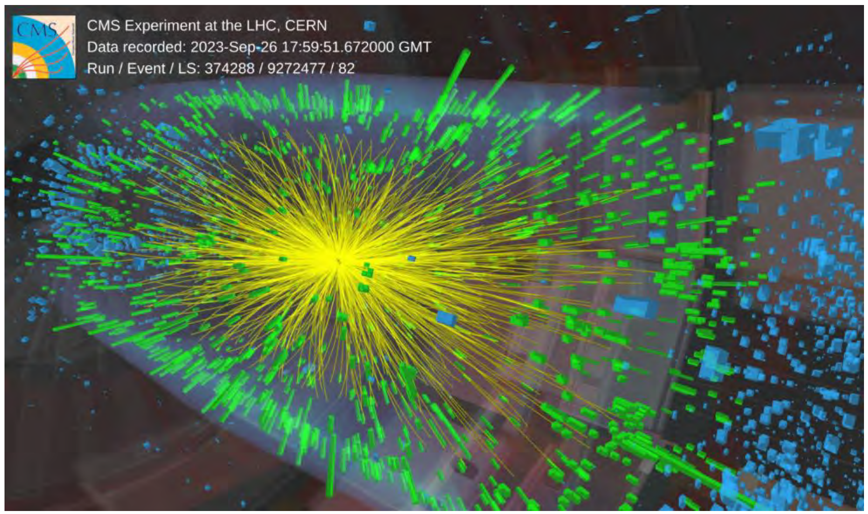
^{22}Ne (p=92,4 ГэВ/с) + Ta

- 50 particles
- only one photo per second



CMS @ LHC, PbPb (5,5 TeV), RUN3

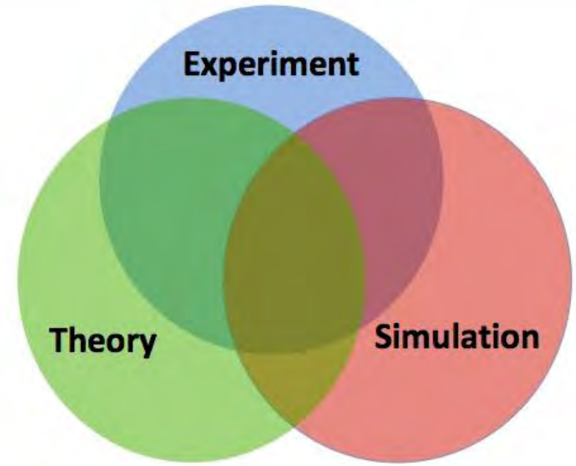
- thousands particles
- 140 million electronic channels
- 3-dimensional “camera” able to shot 40 million “pictures” per second



Particle Physics Tools

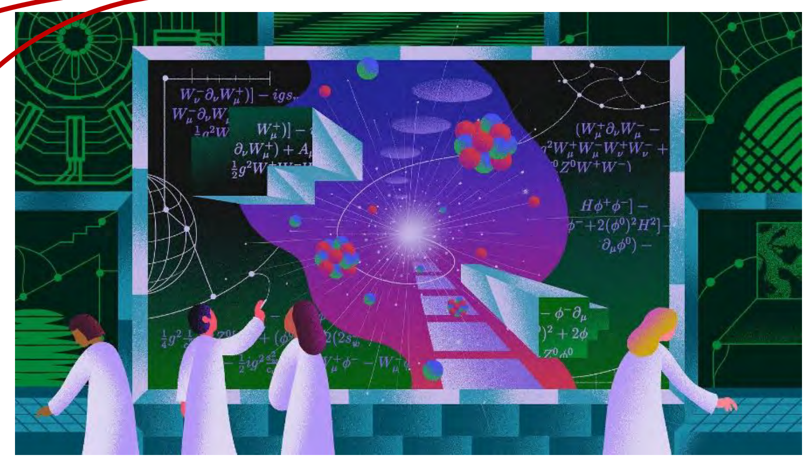
Particle physics or high energy physics is the study of fundamental particles and forces that constitute matter (c) Wiki

- Where can I get elementary particles?
  - ✓ in Nature (cosmic sources, earth sources, i.e. natural radioactivity)
  - ✓ man-made sources (reactors, accelerators)
- How can you catch particles ⇒ detector facilities
- What is needed for data processing?
  - ✓ algorithms and software for reconstruction of physics objects and processes
- What is needed for data analysis?
  - ✓ Theory
  - ✓ Monte Carlo Tools
  - ✓ Statistics Tools

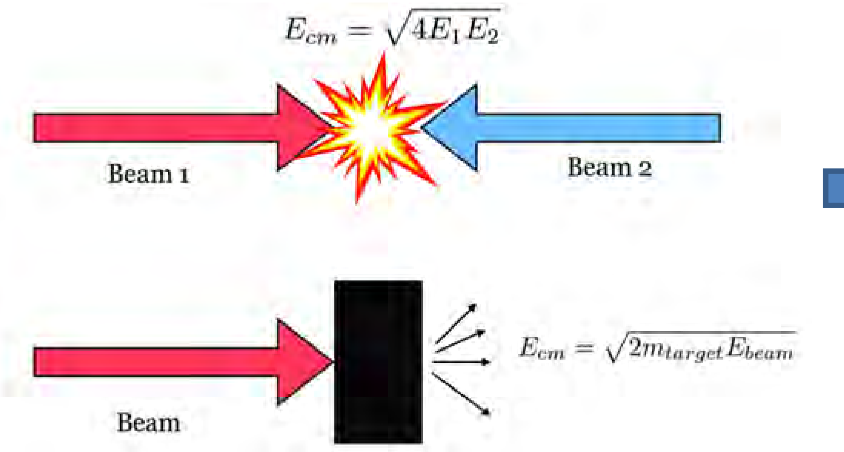


Along Three Paths

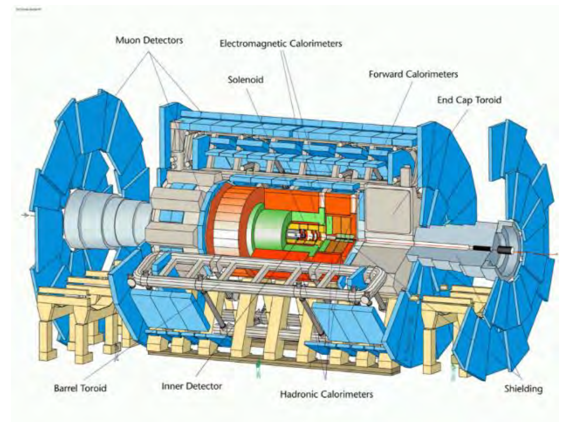
Information Technologies



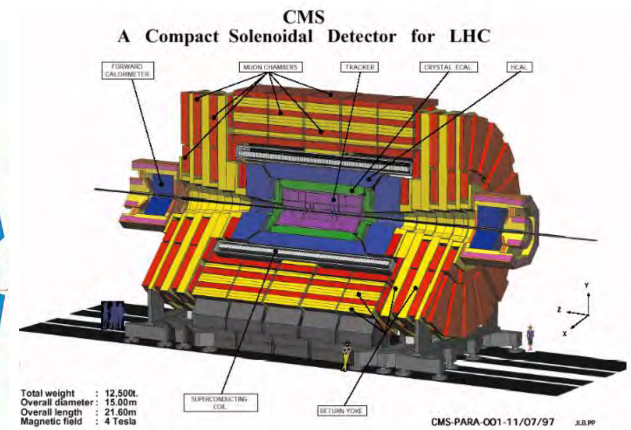
Examples of Experimental Facilities



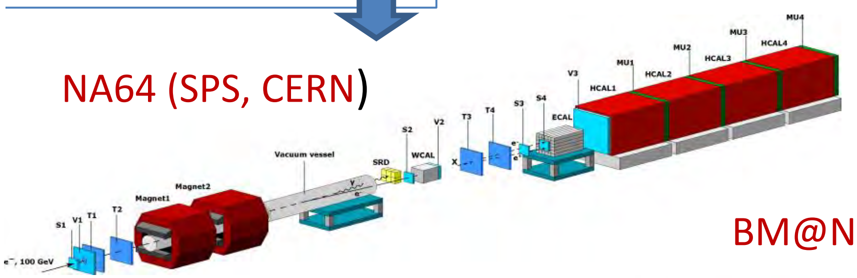
ATLAS (LHC, CERN)



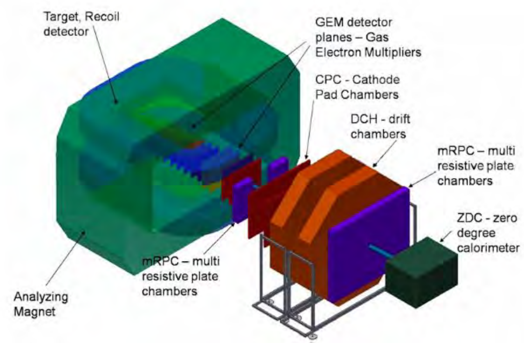
CMS (LHC, CERN)



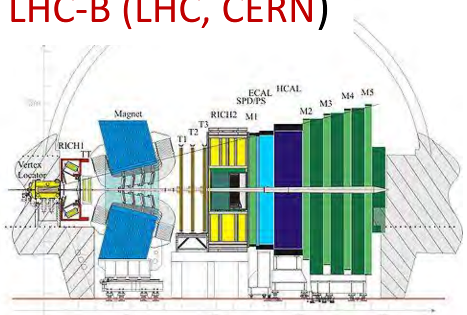
NA64 (SPS, CERN)



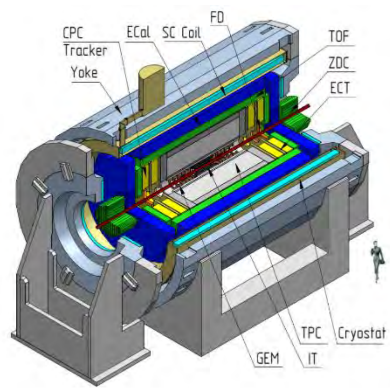
BM@N (NICA, JINR)



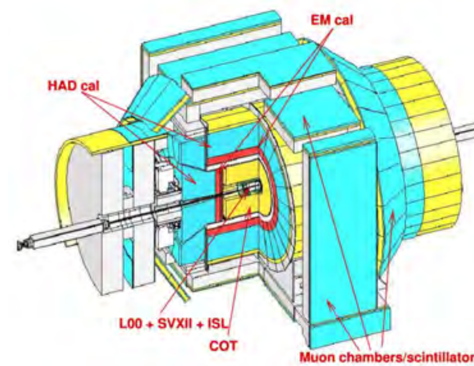
LHC-B (LHC, CERN)



MPD (NICA, JINR)



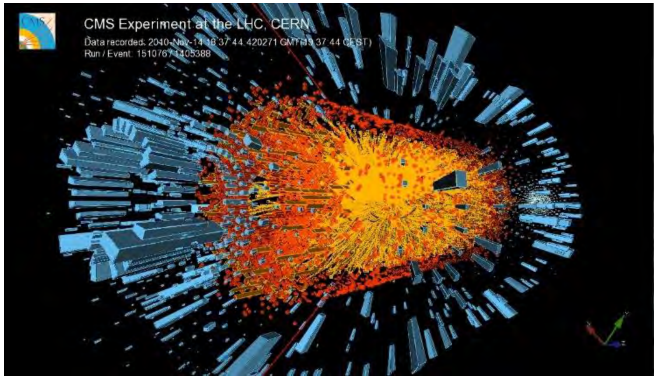
CDF (Tevatron, FLab)



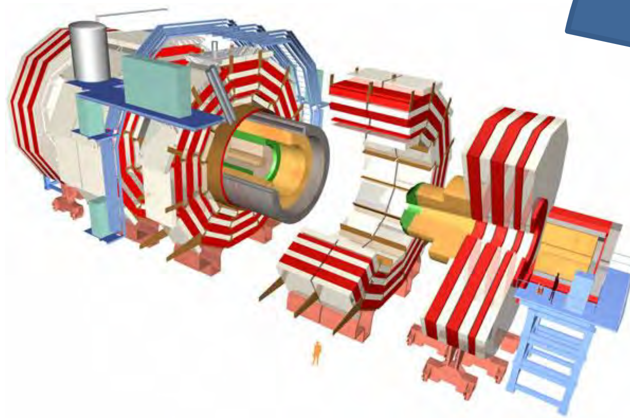
Scale without preservation of proportions

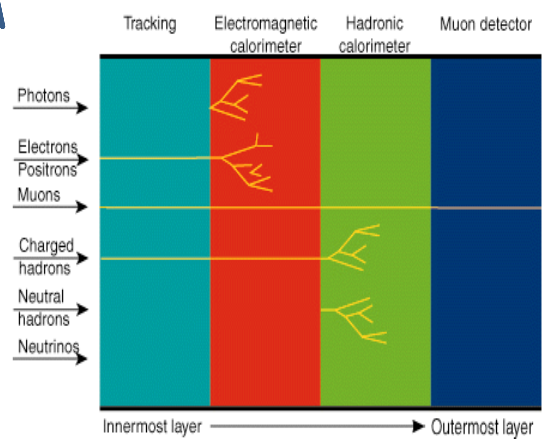
Workflow in Detectors

LHC Collisions

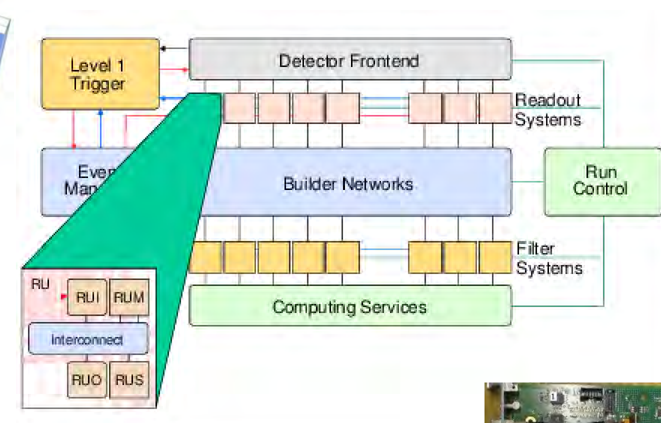


Interactions with detector elements





Readout of detector electronics

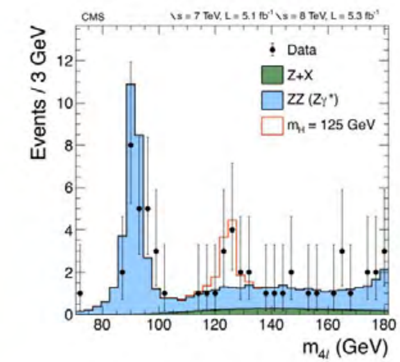


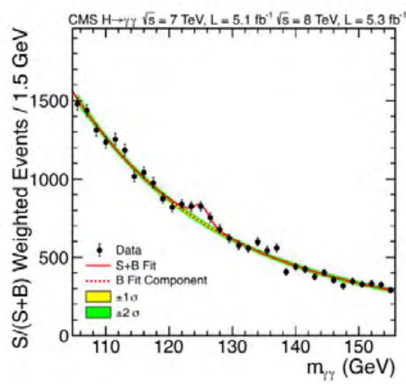
What do physicist want to see?

Higgs Boson

From design

to discovery





4 July 2012

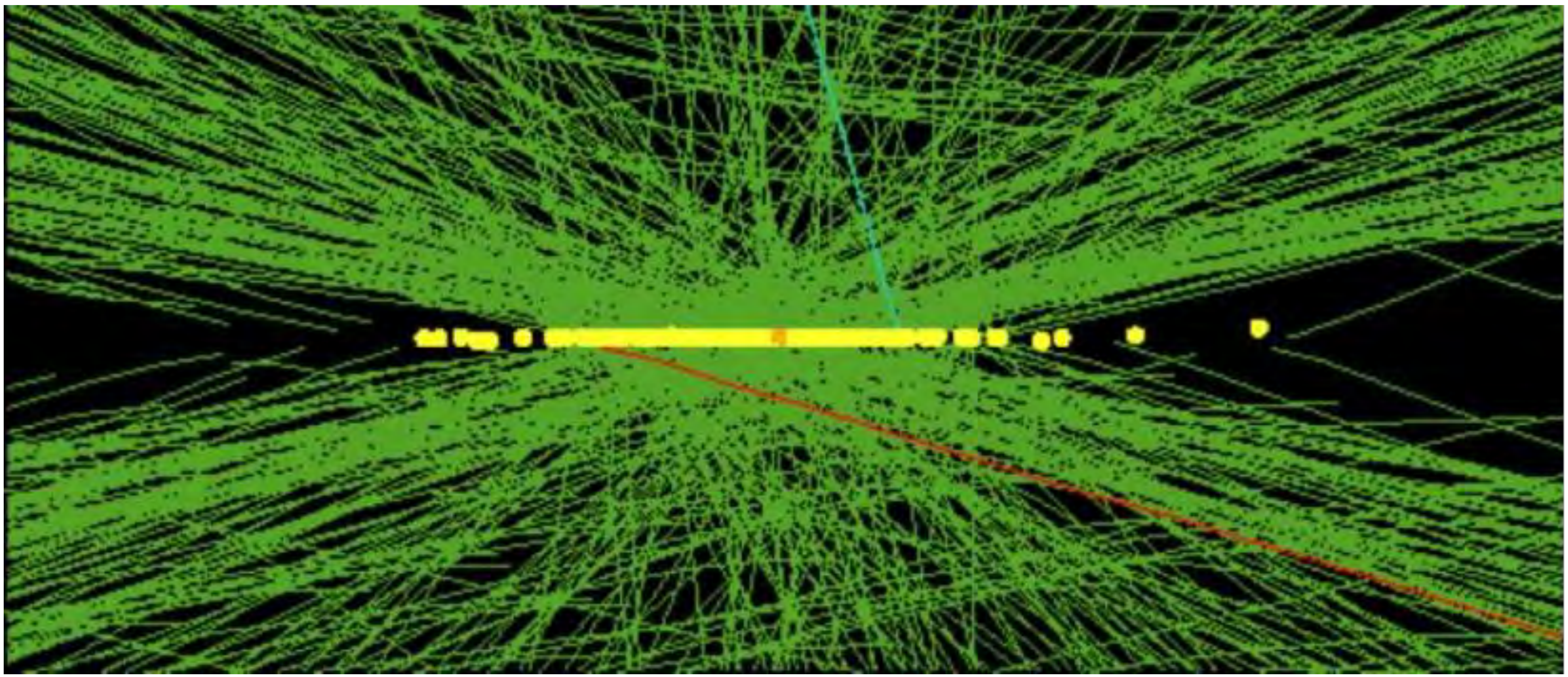
Higgs announcement at CERN

	Int. Luminosity at 7, 8 TeV	mH [GeV]	Expected [st. dev.]	Observed [st. dev.]
ATLAS	10.7 fb^{-1}	126.0 ± 0.6	4.6	5.0
CMS	10.4 fb^{-1}	125.3 ± 0.6	5.9	4.9

What do they actually see?

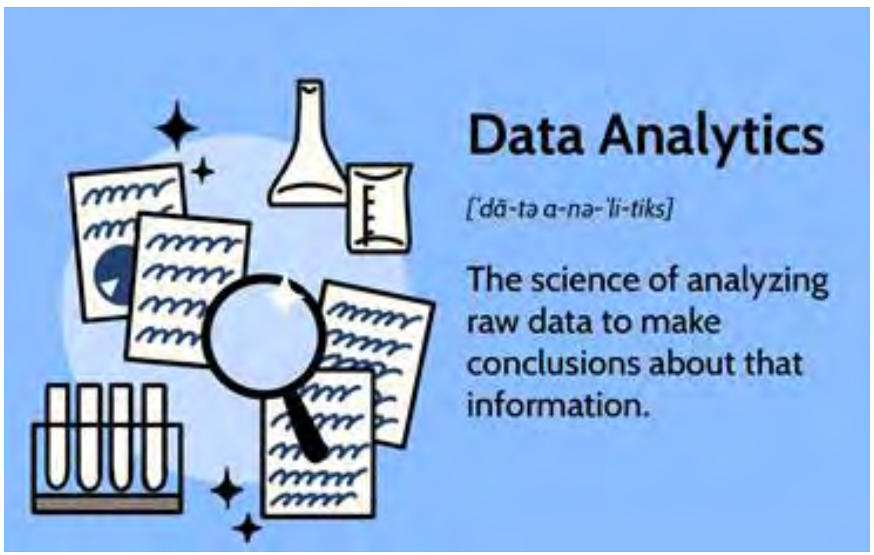
Real CMS Event with High Pile-up

High pileup event with 78 reconstructed vertices taken in 2012 by CMS

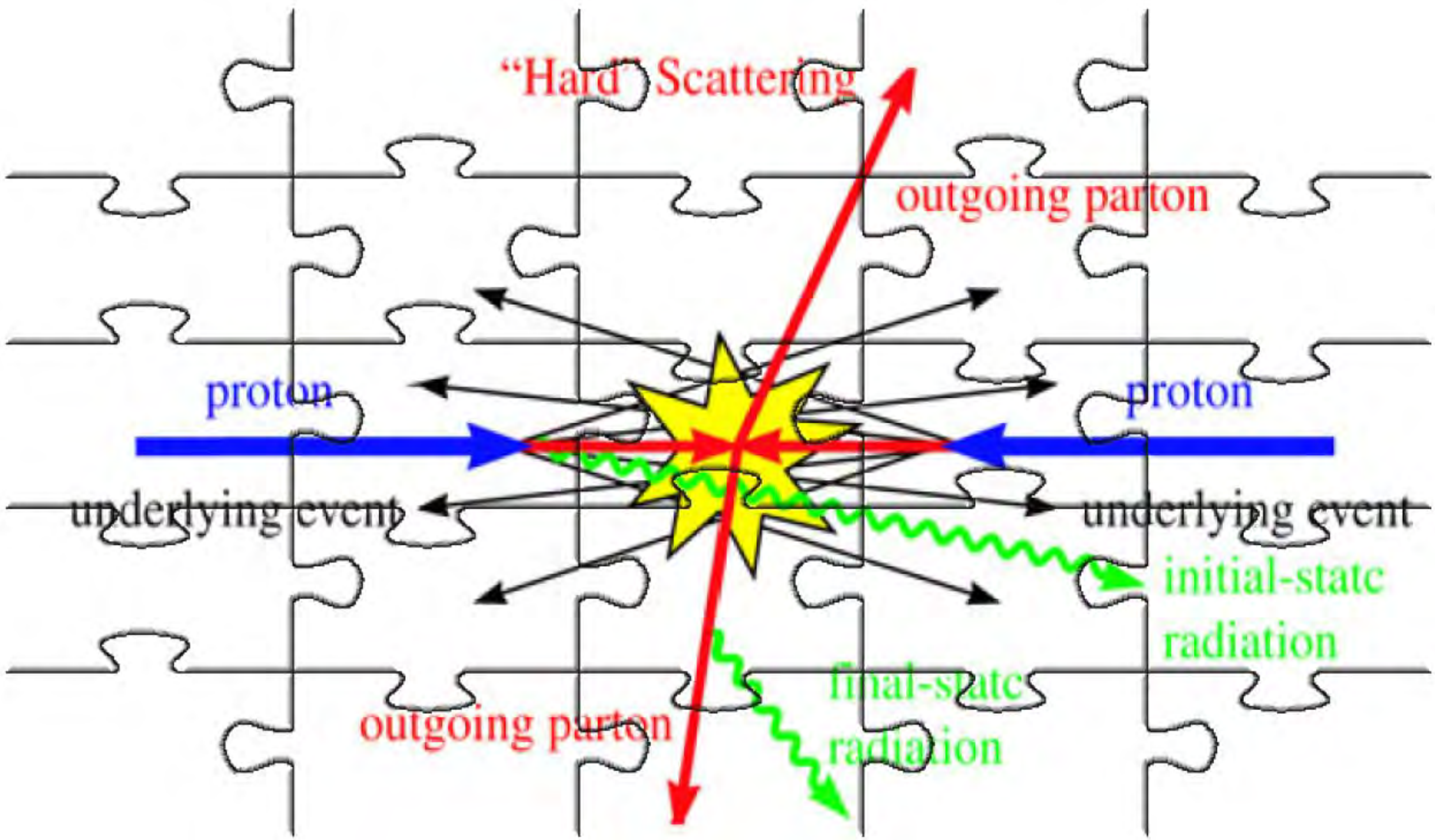


What is happening and and what we can do about it…

- Physics objects
- Event Selection
- Reconstruction and Processing
- Data Analysis

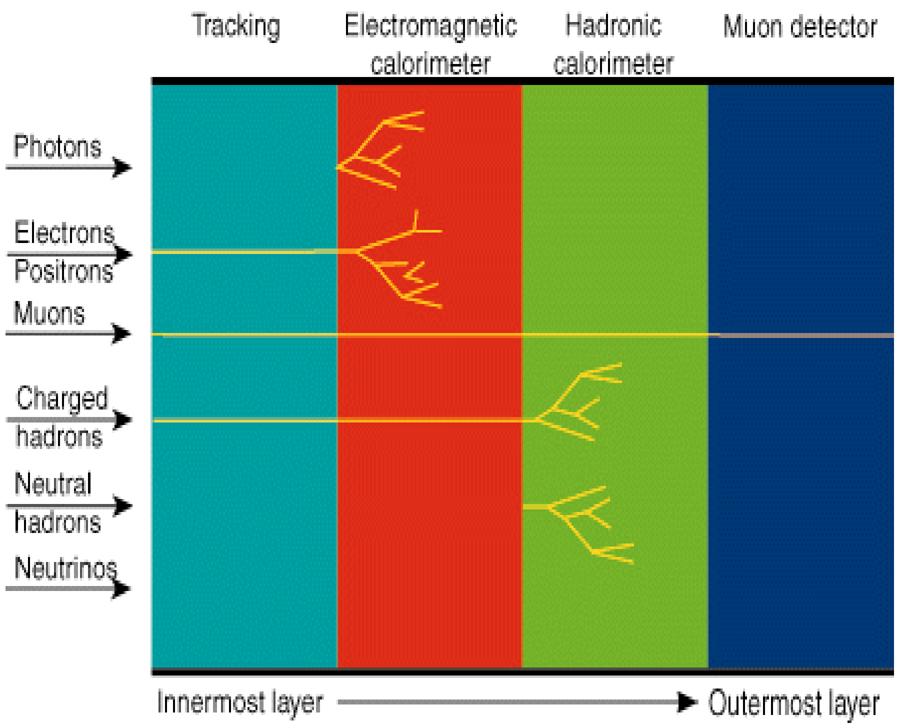


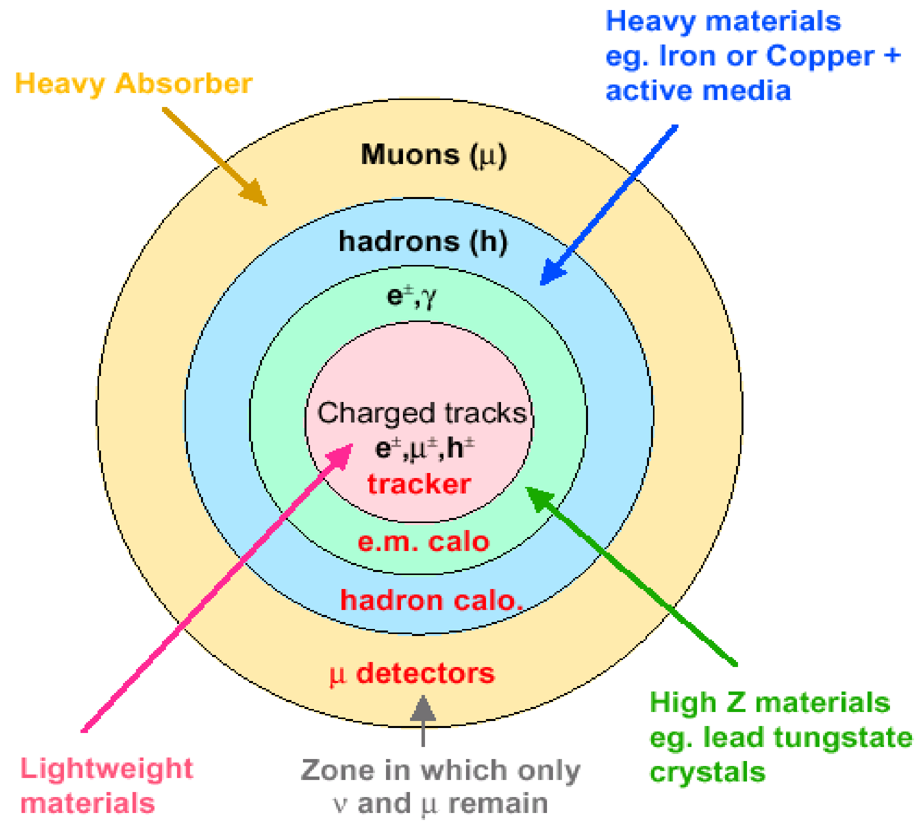
Mosaic of Collisions



Modus Operandi for Experiments

Onion structure of detector layers placed in B-field



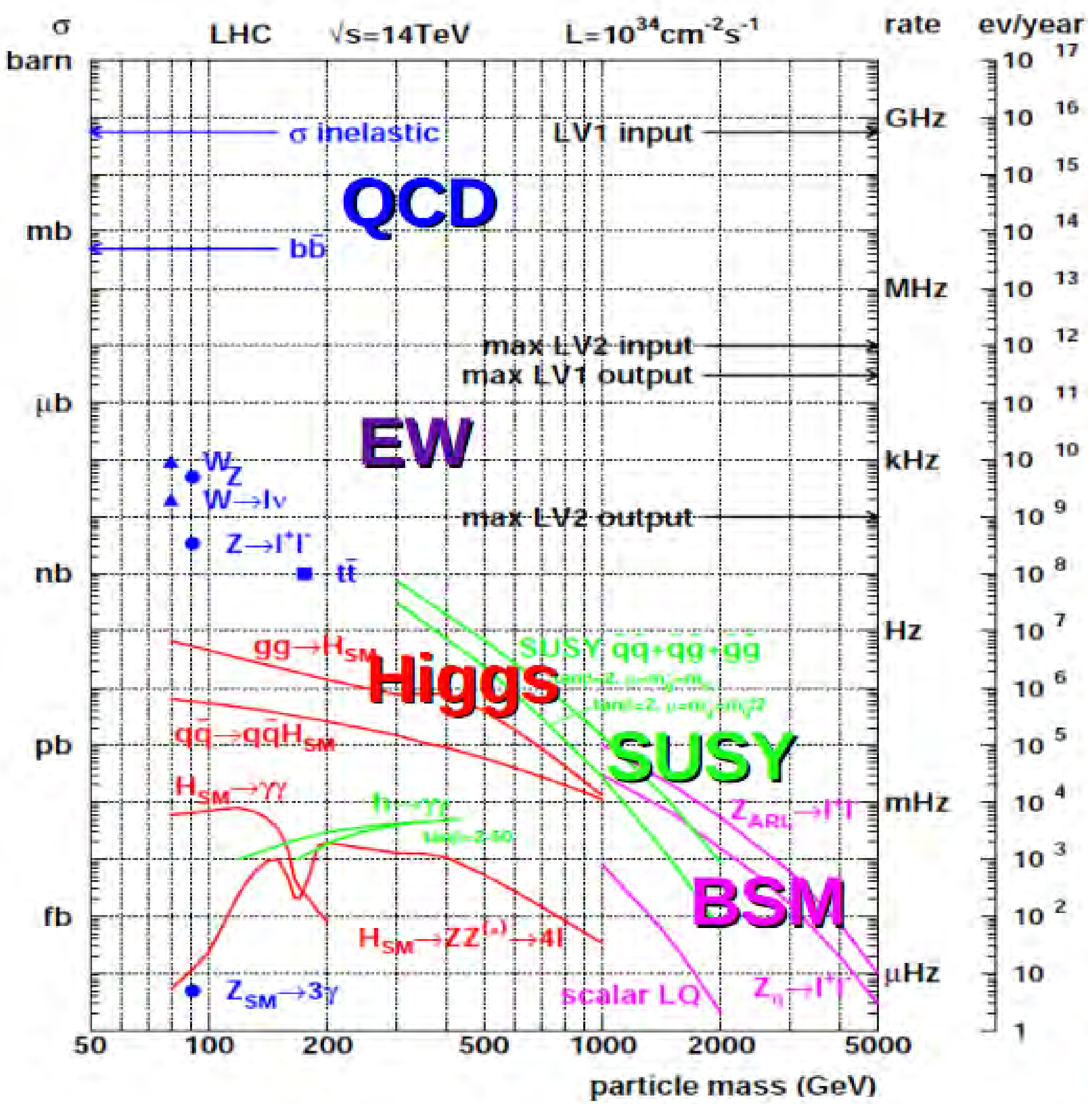


Each layer identifies and measures (or remeasures) the energy of particles unmeasured by the previous layer

No single detector can determine identity and measure energies/momenta of all particles

Event Selection and Data Flow

Physics Processes at LHC



■ SM processes:

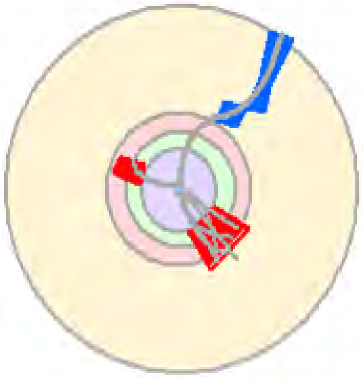
$\sigma \sim 1/(100\ \text{MeV})^2$

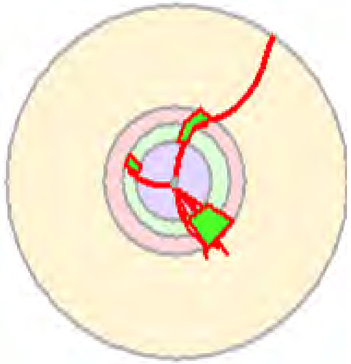
10^{-8} !

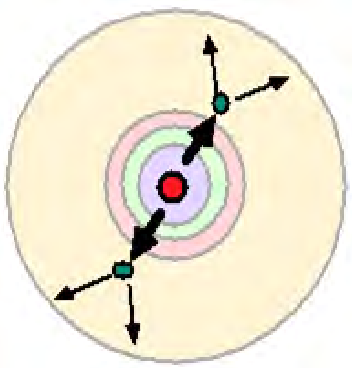
■ New Physics:

$\sigma \sim 1/(1\ \text{TeV})^2$

Multi-Layer Selection and Triggering

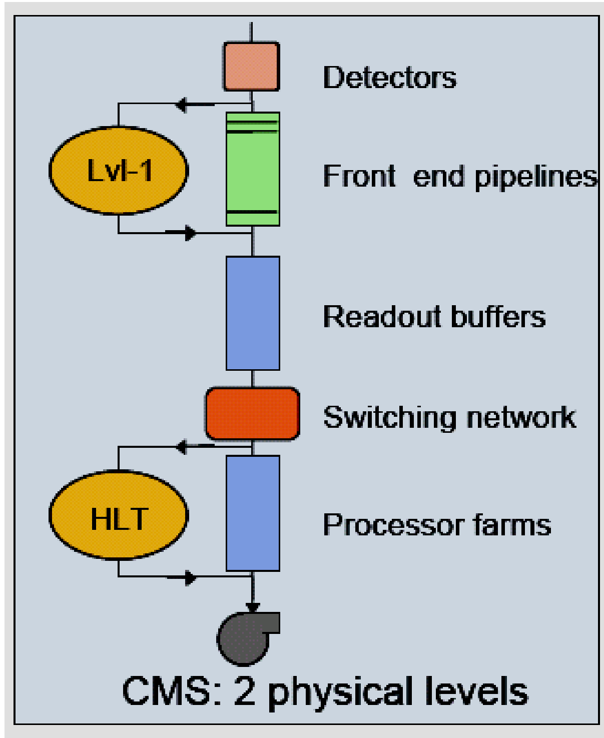




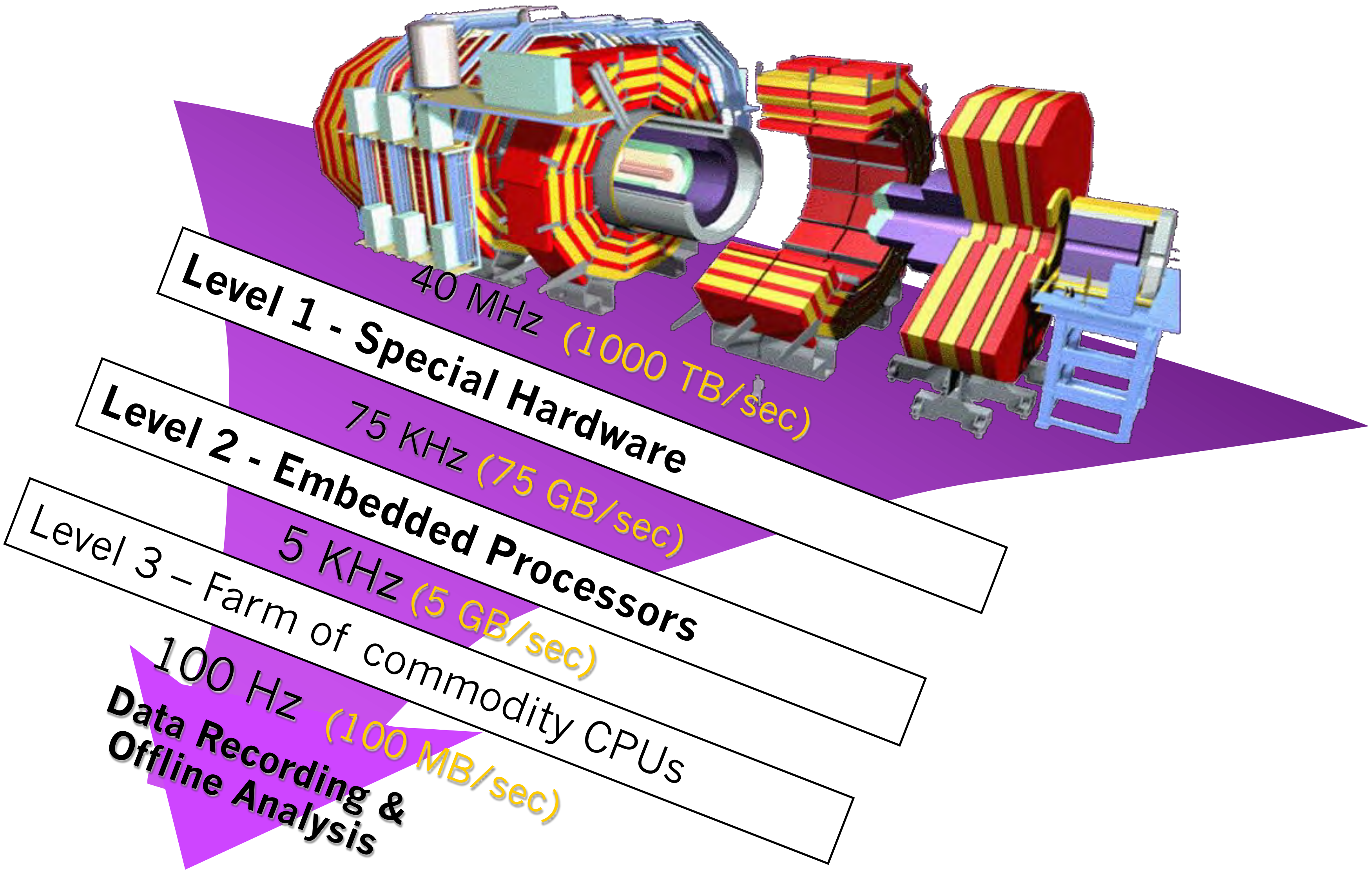


❑ Level-1:
Hardware selection is comprised of custom electronics that process data from detectors, rough cutoffs

❑ High Level Trigger:
Software selection based on reconstruction of physics objects, event topology

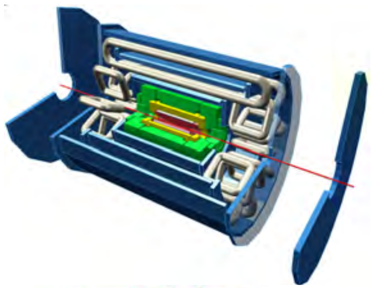


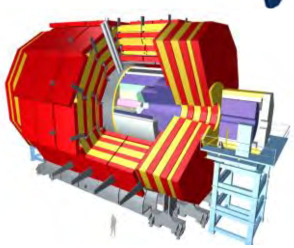
Trigger Rates

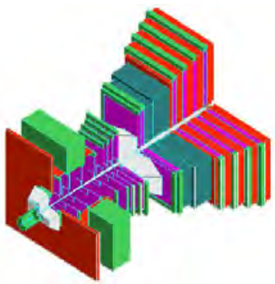


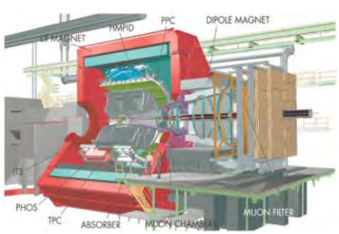
Data Flows

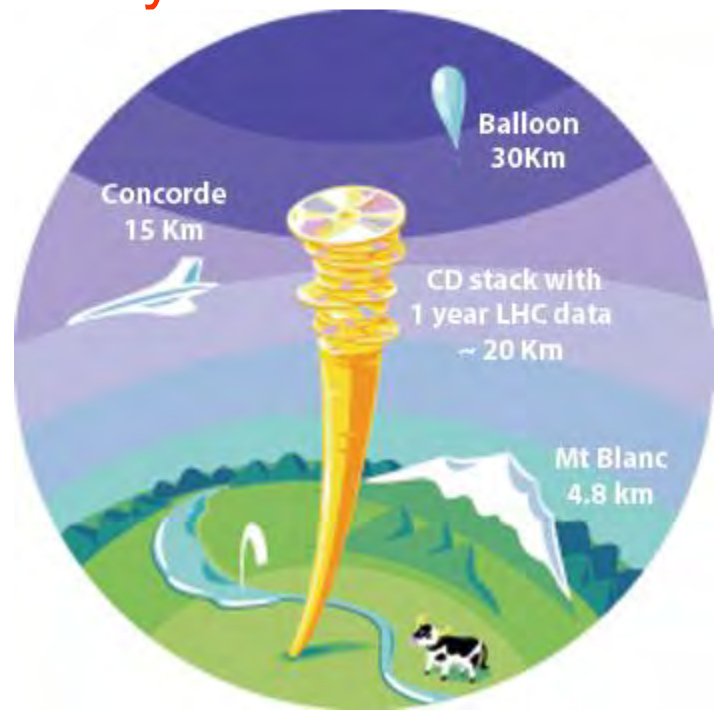
	Level-1 kHz	Event MByte	Storage MByte/s	
ATLAS	100	1	100	~ 3PB/year
CMS	100	1	100	
LHCb	400	0.1	20	
ALICE	1	25	1500	







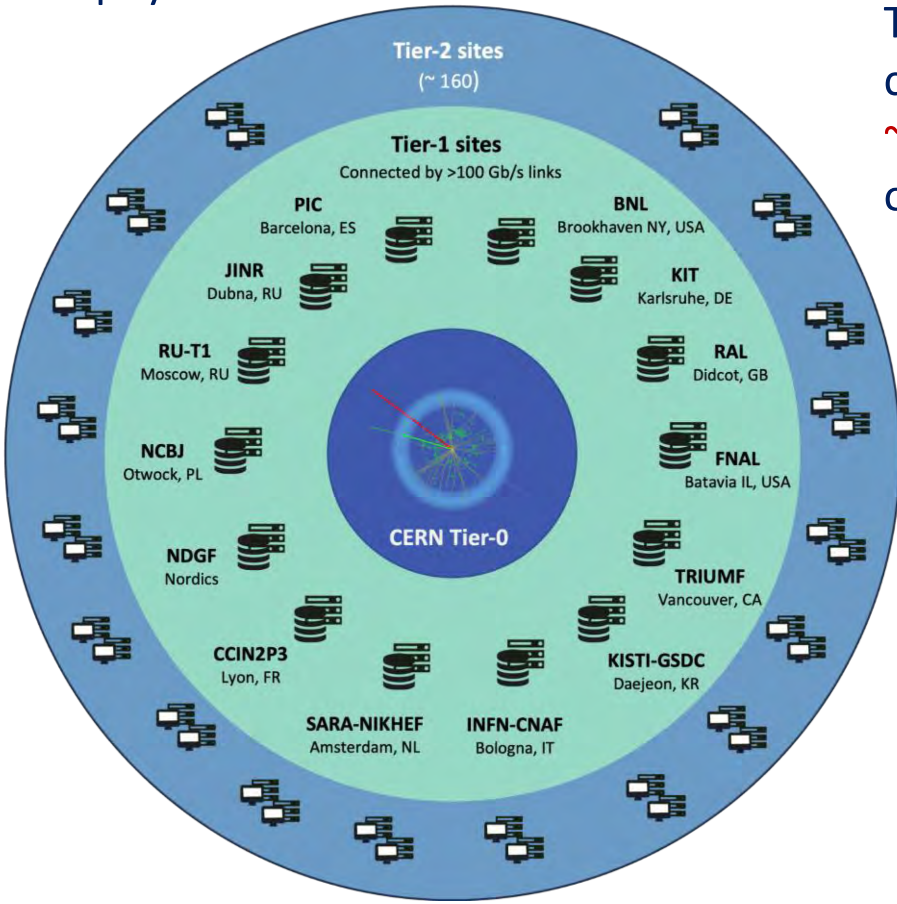




Big Data/GRID Computing

The Worldwide LHC Computing Grid

WLCG is an International collaboration to distribute and analyze LHC data. Integrates computer centers worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists



The mission of the WLCG project is to provide global computing resources to store, distribute and analyze the ~50-70 Petabytes of data expected every year of operations from the Large Hadron Collider.

WLCG computing enabled physicists to announce the discovery of the Higgs Boson.

- 170 sites
- 42 countries
- > 12k physicists
- ~1.4 M CPU cores
- 1.5 EB of storage
- > 2 million jobs/day
- 100-250 Gb/s links

Worldwide LHC Computing Grid - 2019

Tier0 (CERN): data recording, reconstruction and distribution

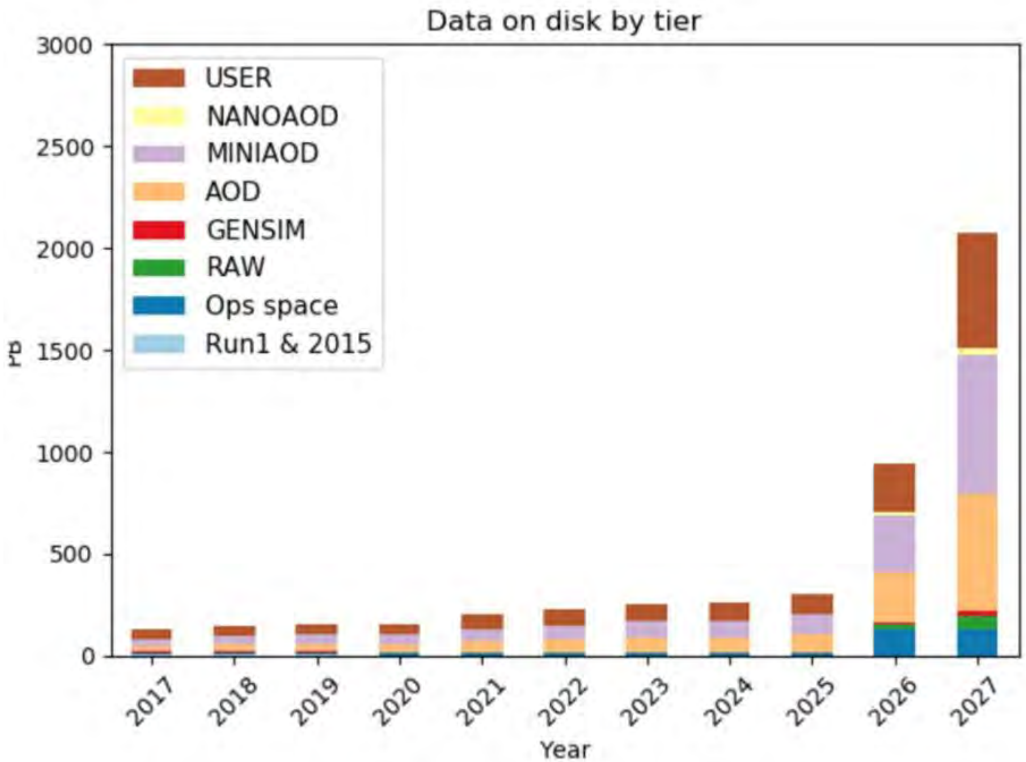
Tier1: permanent storage, re-processing,

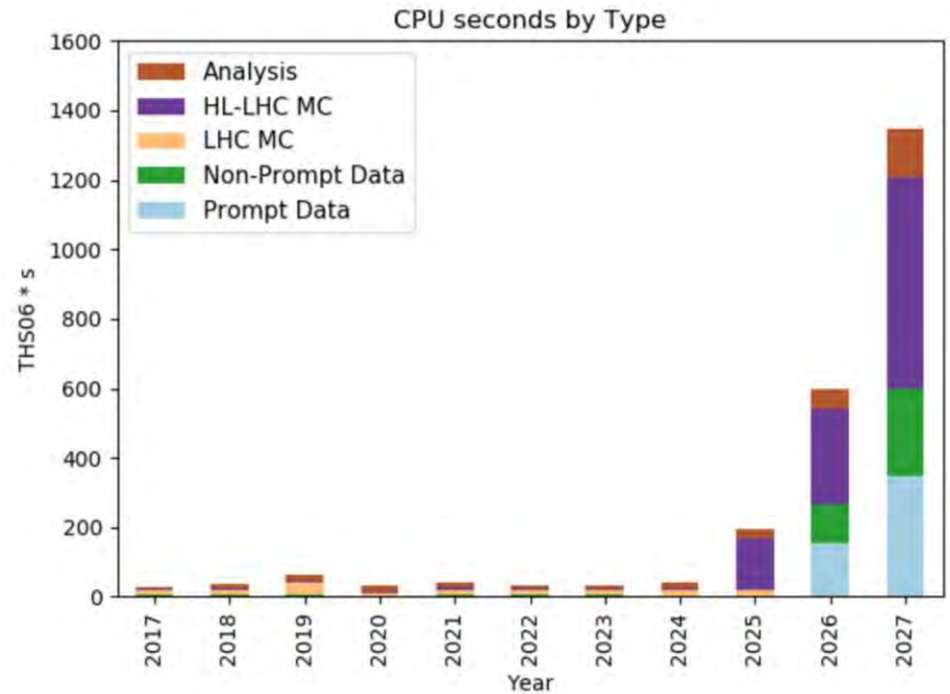
Tier2: Simulation, end-user analysis

Multifunctional Information and Computing Complex in LIT JINR

What Do We Expect?

CMS @ LHC Example



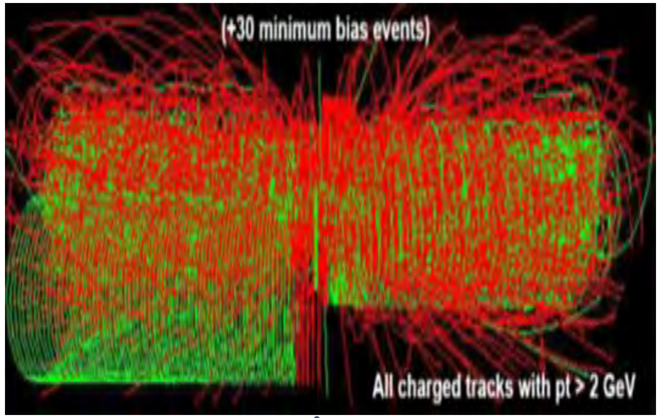


- Facing up to the exabyte (10^{18} bytes) era $\Rightarrow$ required computing capacity is roughly 10 times higher than today
- an improvement of around a factor 10 in processing capabilities

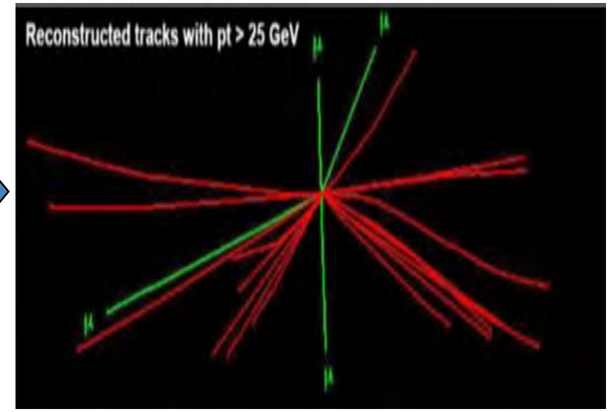
Data Processing and Analysis

RAW Data

Reconstruction

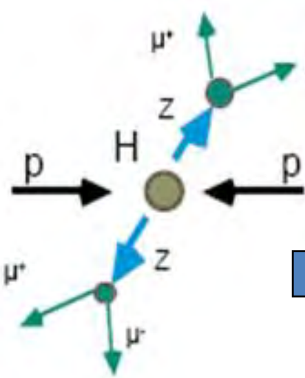


Event Selection

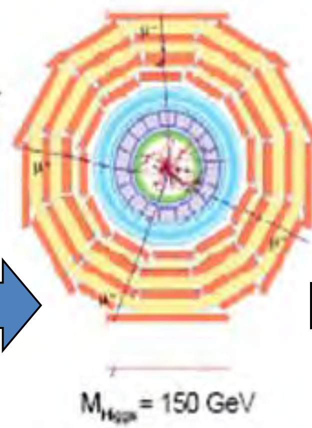


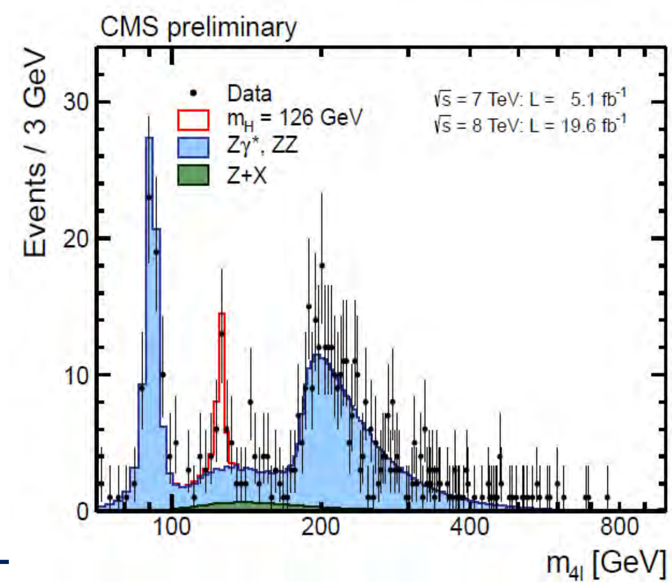
Calibration/Condition/etc
Data Bases

Theory/
Monte Carlo
Simulation

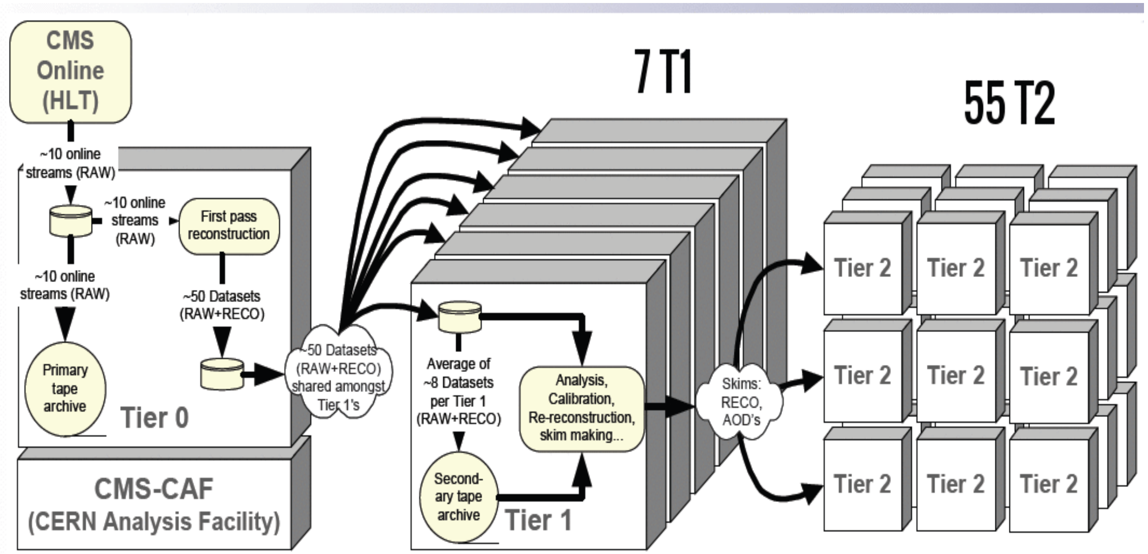


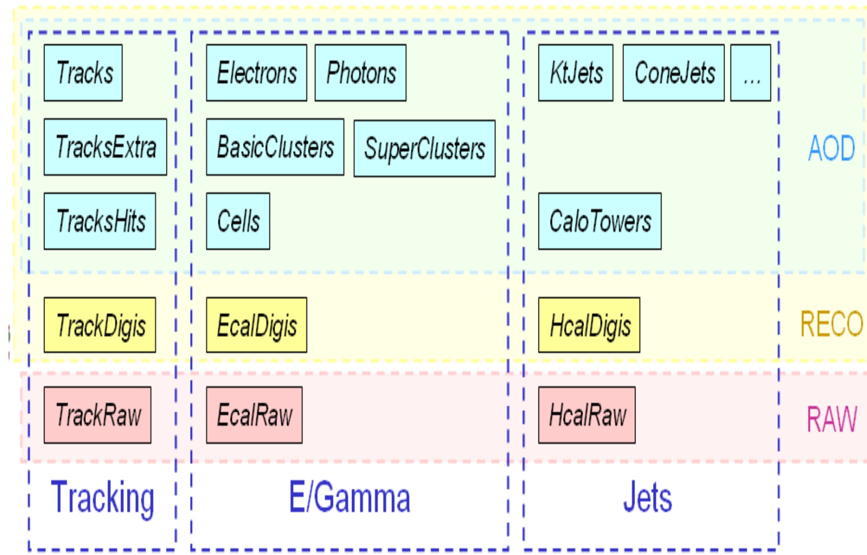
Detector Response
Simulation



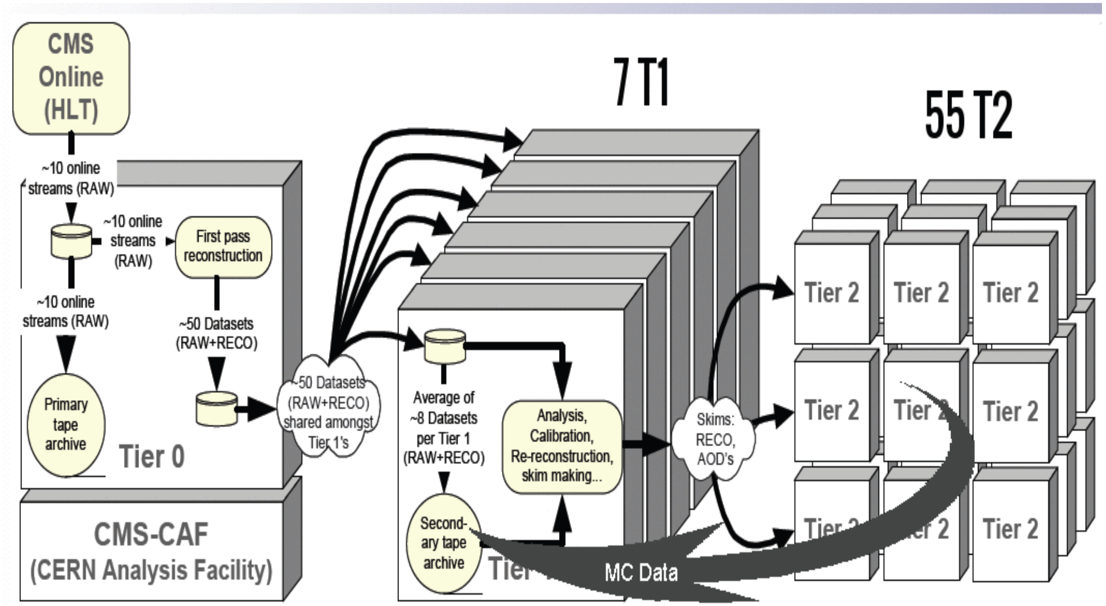


Data Model and Data Flow through Tiers



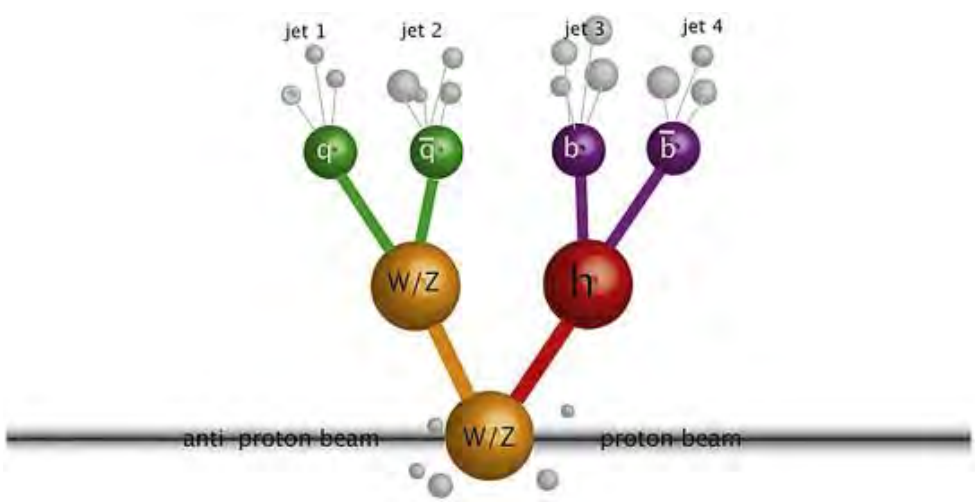


- T0 ⇒ T1
  - ✓ scheduled, time-critical, will be continuous during data-taking periods
  - ✓ reliable transfer needed for fast access to new data, and to ensure that data is stored safely
- T1 ⇒ T1:
  - ✓ redistributing data, generally after reprocessing (e.g. processing with improved algorithms)
- T1 ⇒ T2:
  - ✓ Data for analysis at Tier-2s

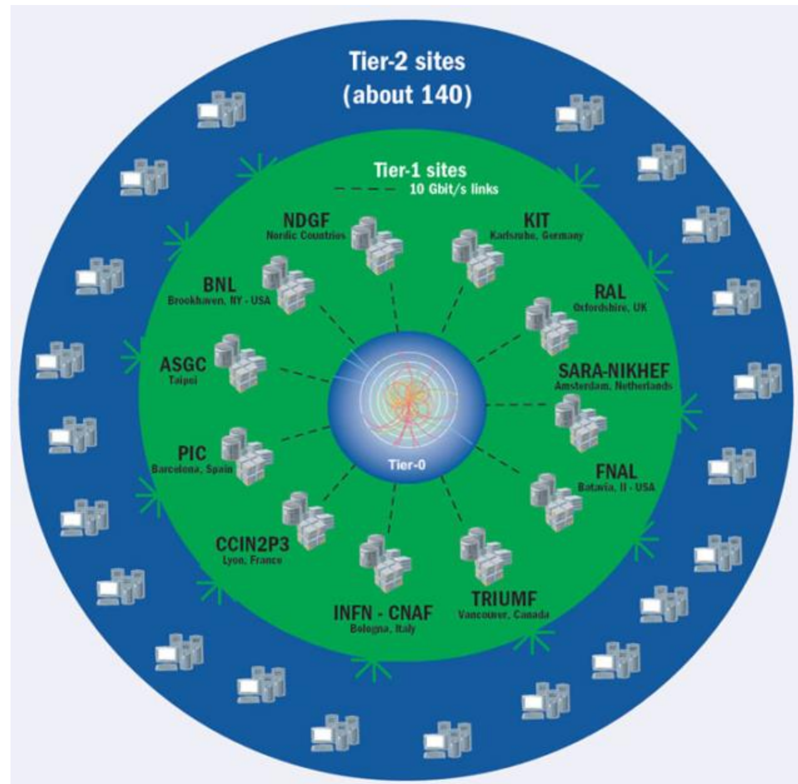


Event Reconstruction

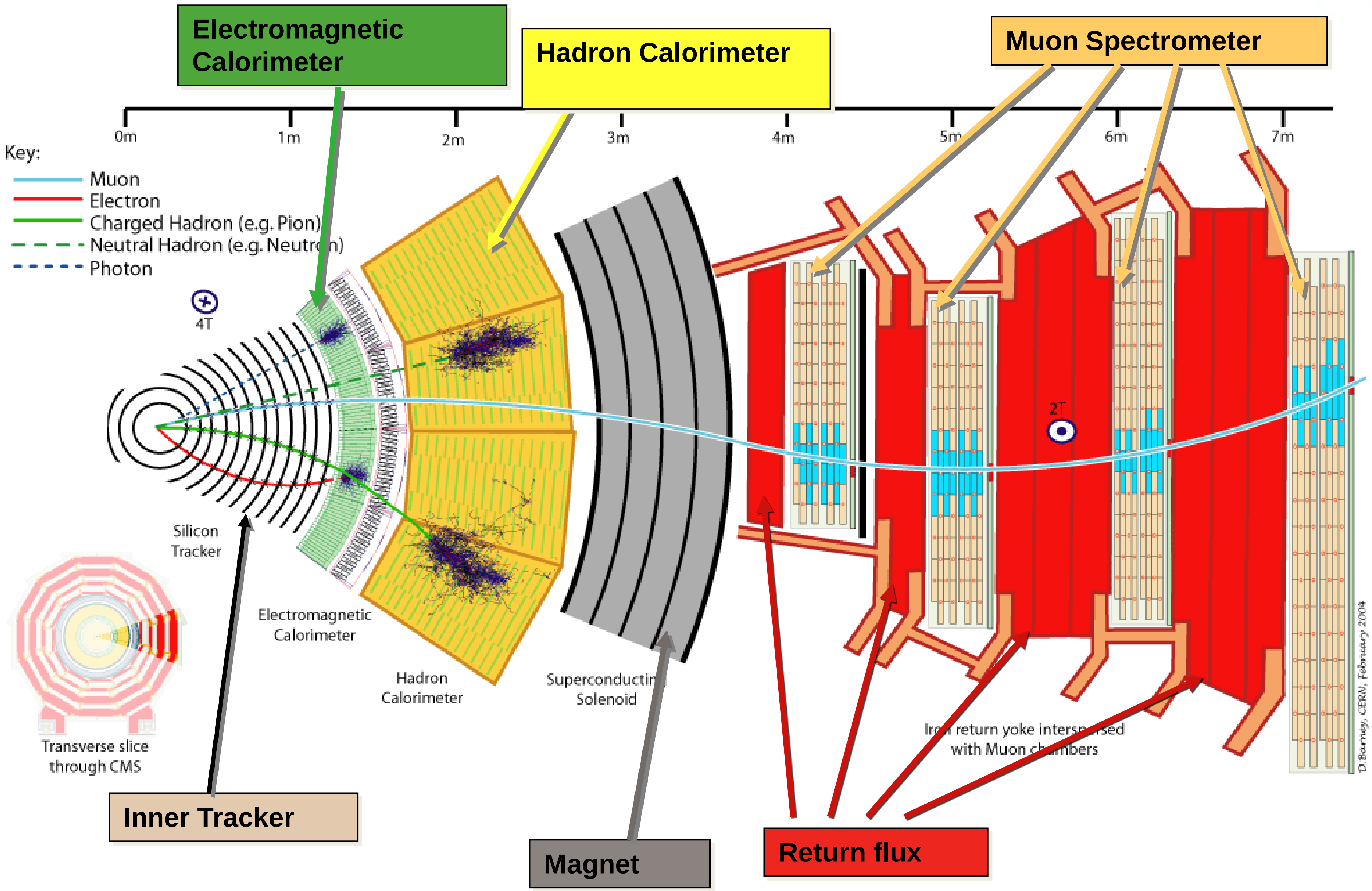
- Reconstruction (mathematical methods/algorithms/SW)
  - ✓ physics objects - stable particles (e, μ, γ), clusters of particles (energy), vertexes, etc
  - ✓ unstable particles/ physics processes



- Data Processing

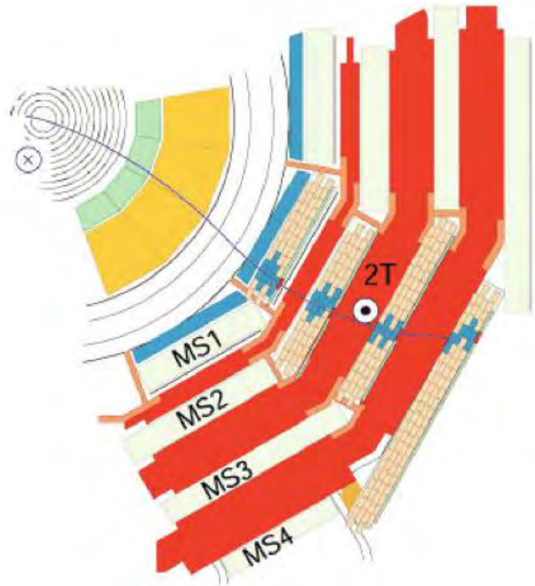


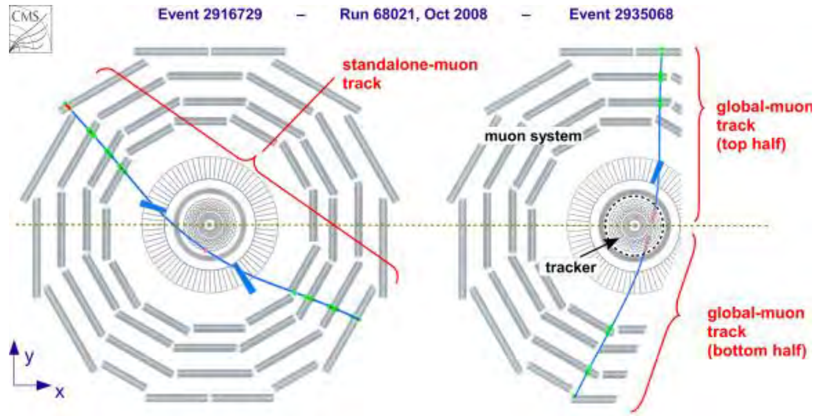
Particles in Detectors

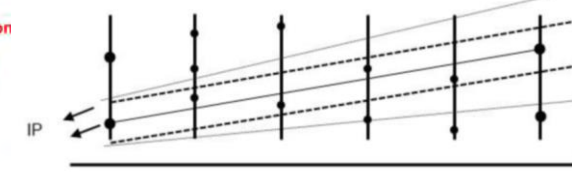


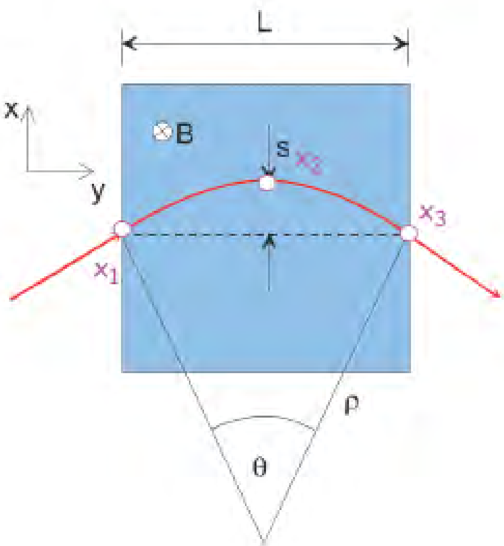
Muon Track and Dumuons Reconstruction

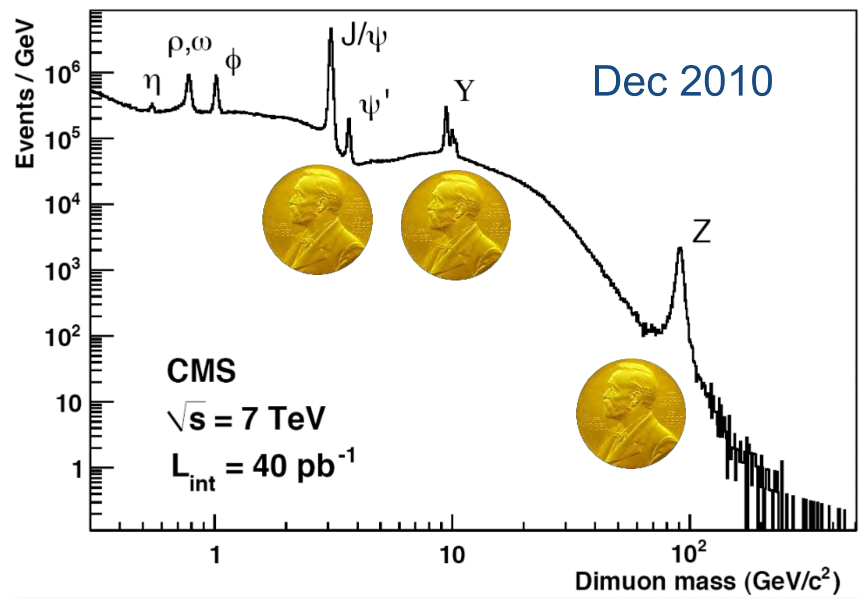
CMS Muon System shows a excellent performance to detect different resonances







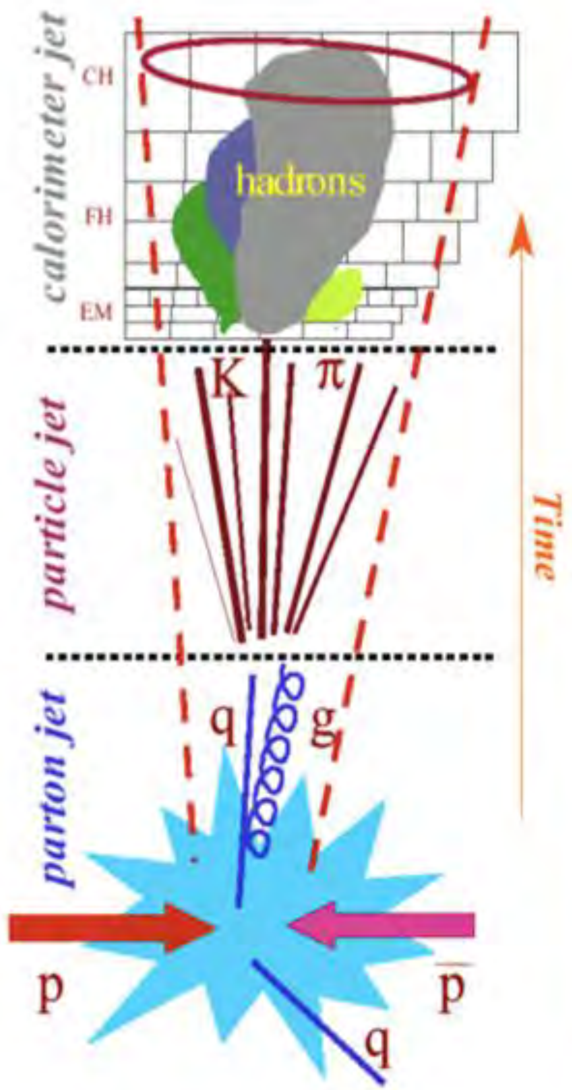




$$\frac{\sigma(p_T)}{p_T} = \frac{\sigma_x p_T}{0.3BL^2}\sqrt{\frac{720}{N+4}}$$

https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsMUO

Jet Finding



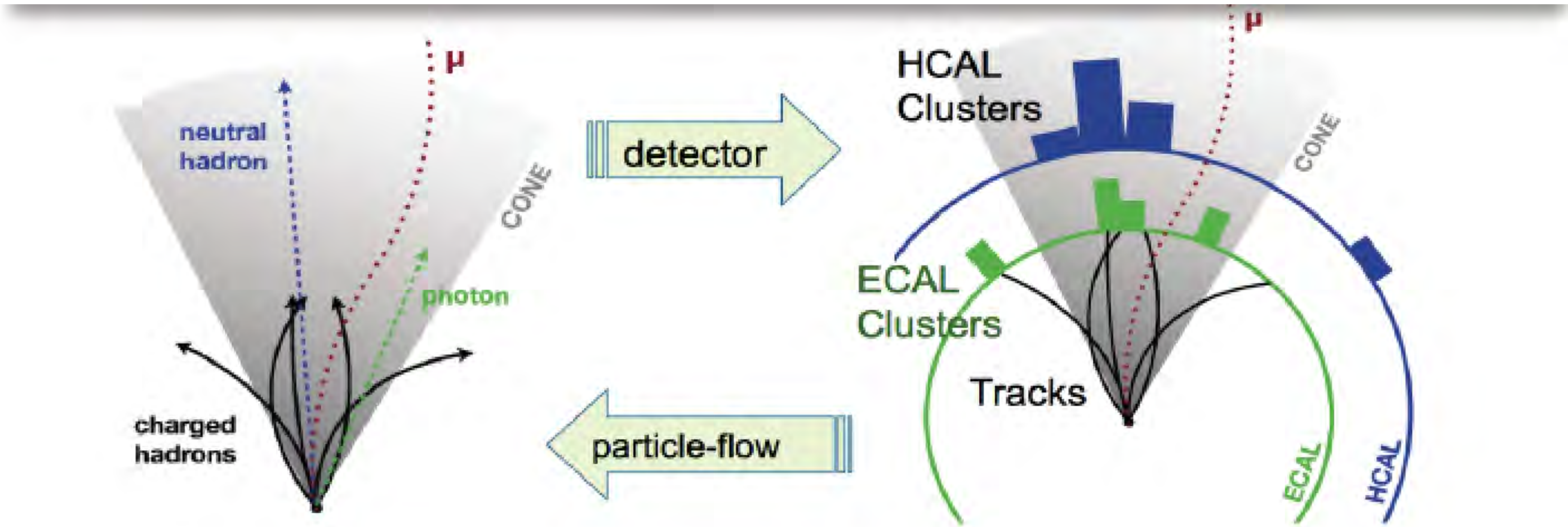
- **Calorimeter jet (cone)**
  - jet is a collection of energy deposits with a given cone *R:* $R = \sqrt{\Delta\varphi^2 + \Delta\eta^2}$
  - cone direction maximizes the total E_T of the jet
  - various clustering algorithms

→ correct for finite energy resolution
→ subtract underlying event
→ add out of cone energy

- **Particle jet**
  - a spread of particles running roughly in the same direction as the parton after hadronization

Global Event Reconstruction

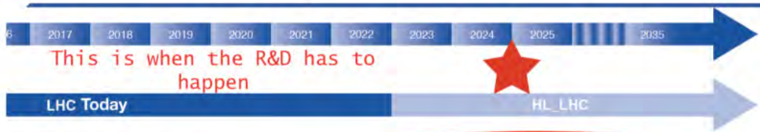
Using all information of the detector together for optimal measurement



- Optimal combination of information from all subdetectors
- Returns a list of reconstructed particles
  - e,μ,γ, charged and neutral hadrons
    - Used in the analysis as if it came from a list of generated particles
    - Used as building blocks for jets, taus , missing transverse energy , isolation and PU particle identification

Machine Learning

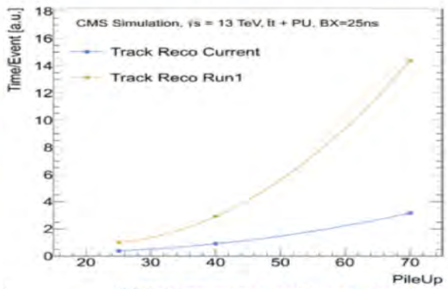
HL-LHC: elephant in the room



- ~40 collisions/event
- ~10 sec/event processing time
- (at best)Same computing resources as today

- ~200 collisions/event
- ~minute/event processing time(*)
- (at best)Same computing resources as

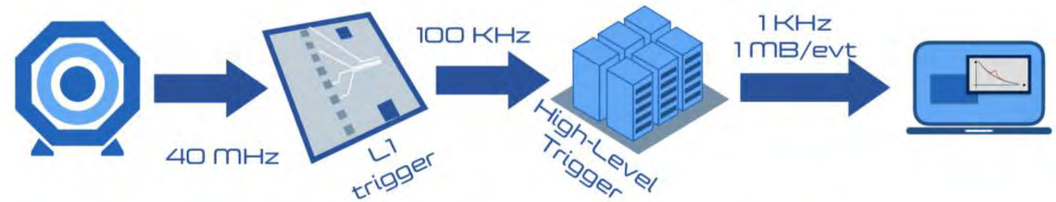
- *Flat budget vs. more needs =* ***current rule-based reconstruction algorithms will not be sustainable***
- ***Adopted solution:*** *more granular and complex detectors → more computing resources needed → more problems*
- ***Modern Machine Learning might be the way out***



(*)With nowadays software development

The LHC Big Data Problem

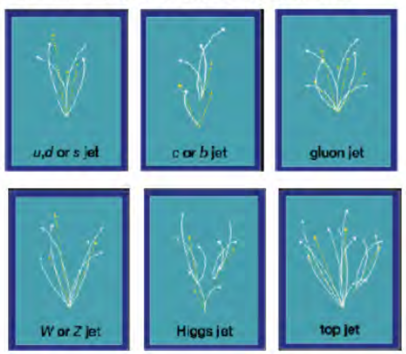
- *Too many data, too large data -> need to filter online*

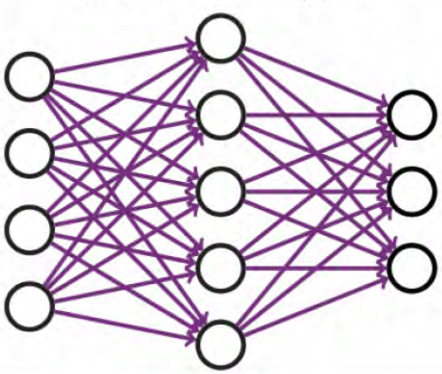


- *The solution to the HL-LHC problem: modern Machine Learning as a fast shortcut between the data and the right answer (the outcome of our traditional & slow algorithms)*

DEEP LEARNING TECHNIQUES

Deep neural networks
based on many low-level features with large training data sets to classify jets

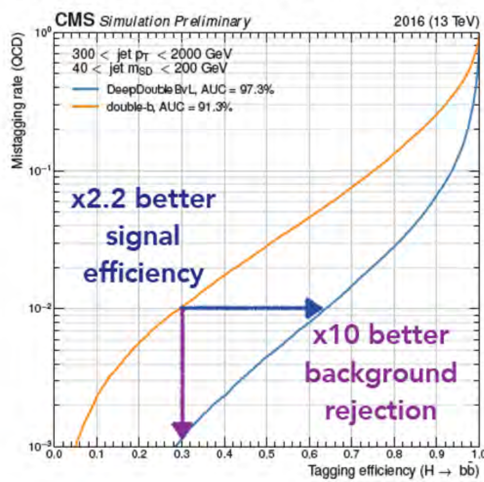




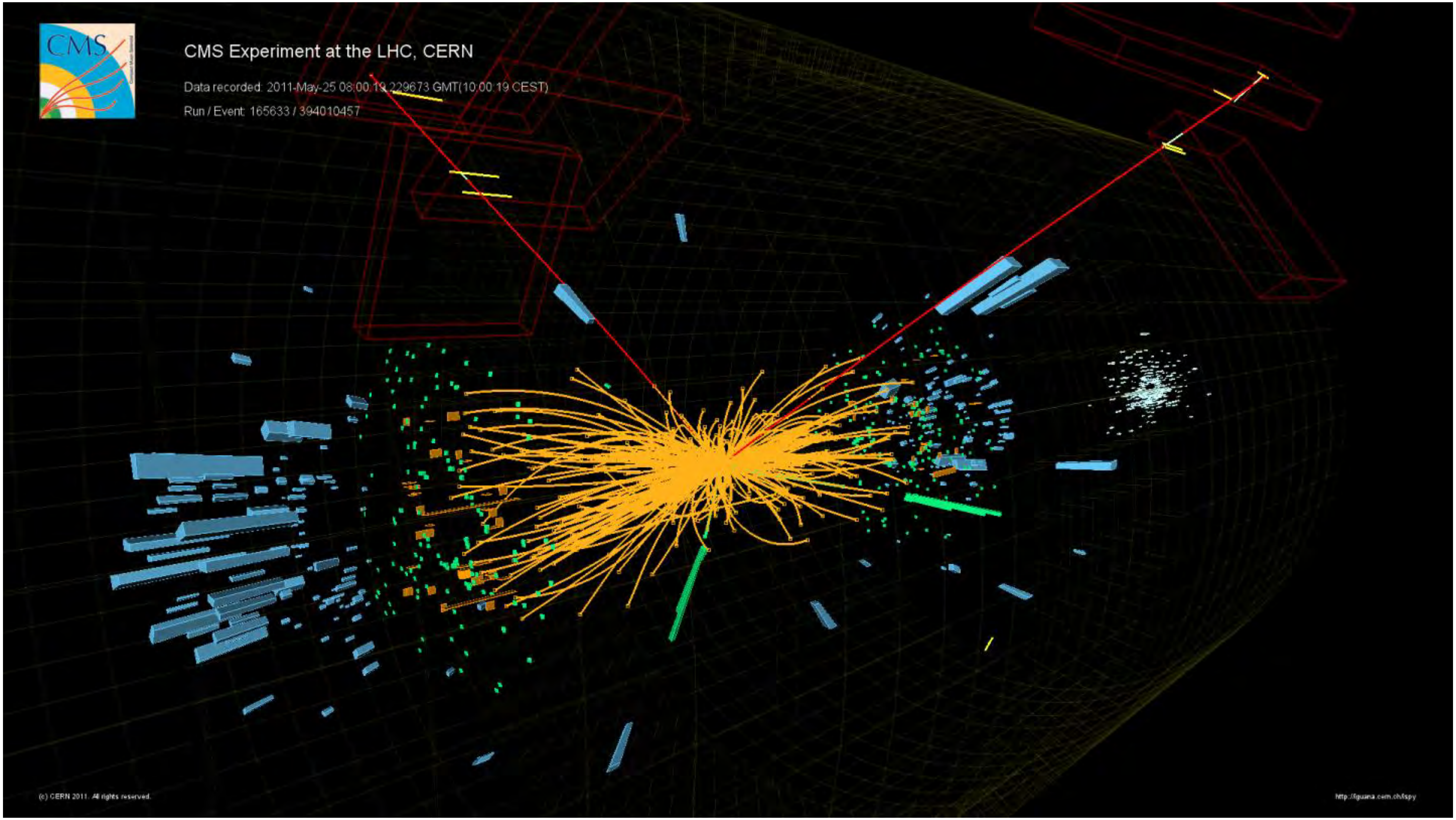
DP-2018/033

DEEP DOUBLE-B TAGGER

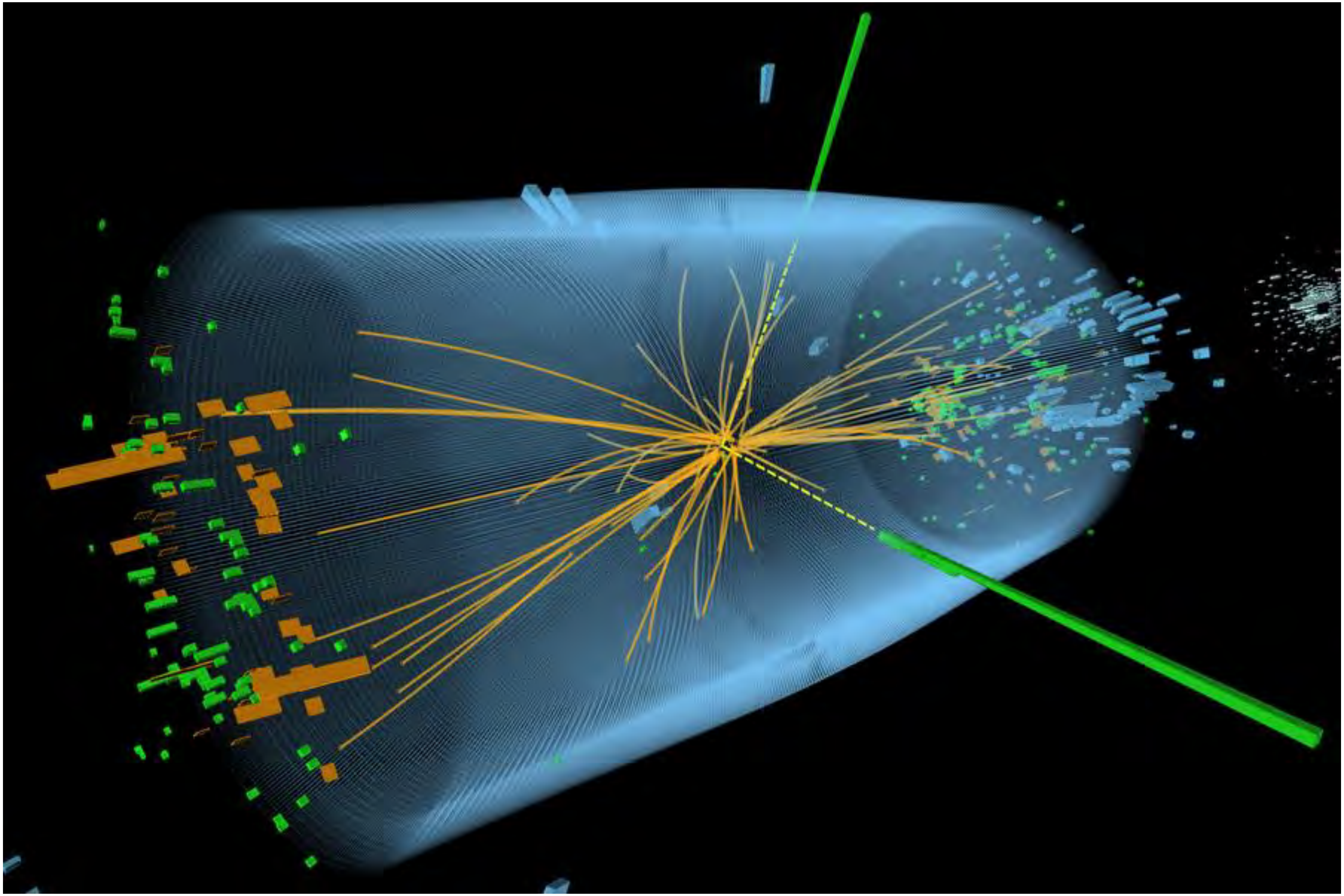
- Large performance gain over previous algorithm



Example of h → ZZ → 2e 2μ

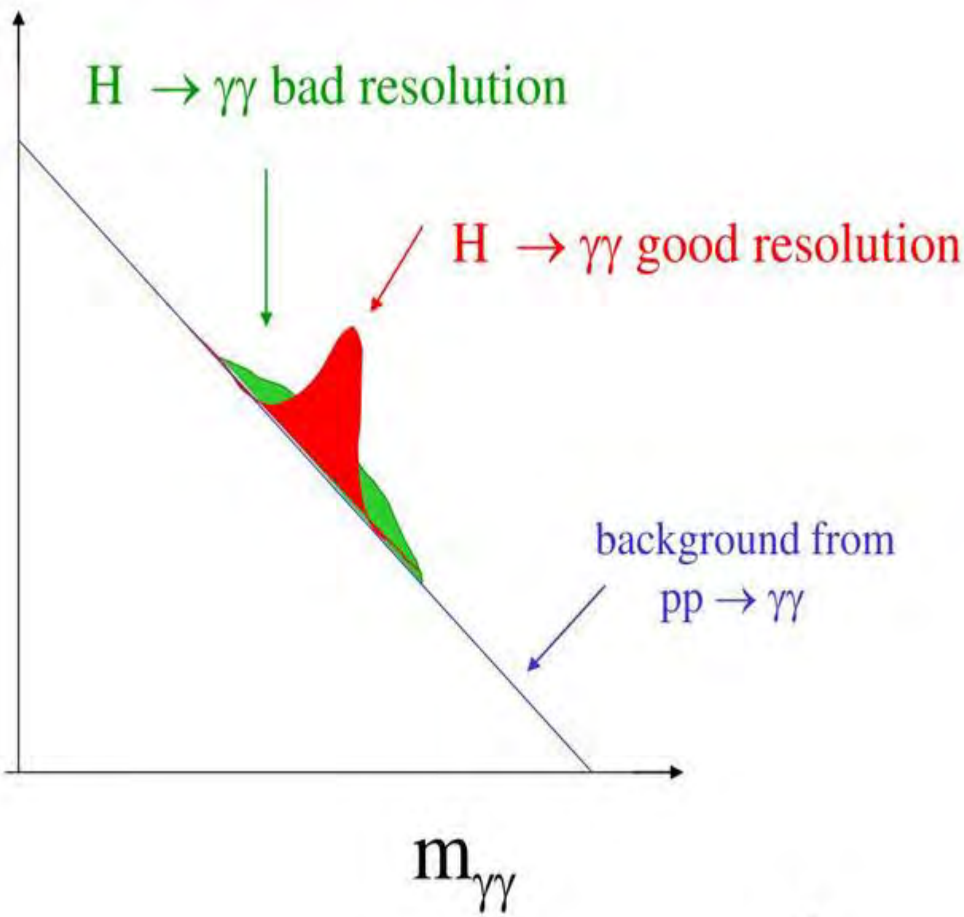


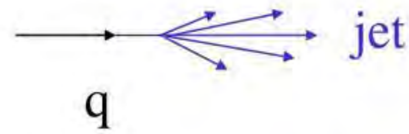
Example of h $\to$ 2γ



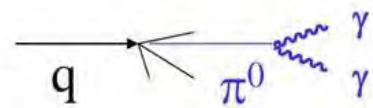
Challenge to the Detector/SW (Example)

Example : $H \rightarrow \gamma\gamma$

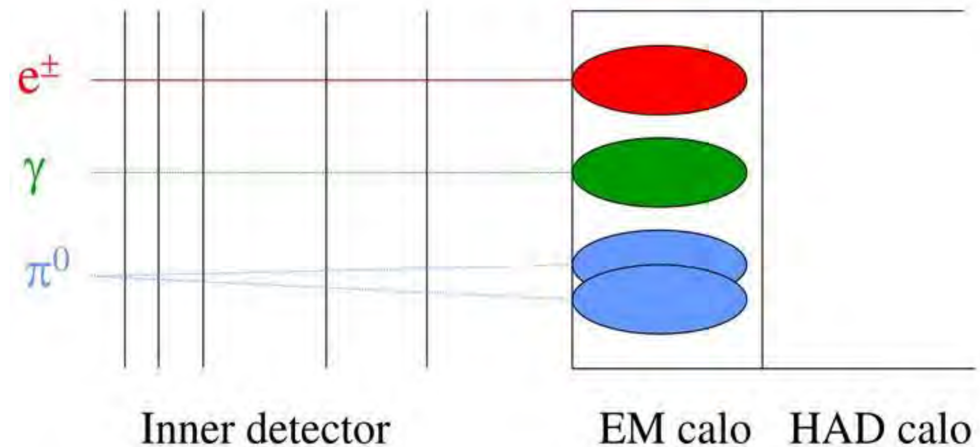




number and p_T of hadrons in a jet have large fluctuations



in some cases: one high-p_T π^0; all other particles too soft to be detected

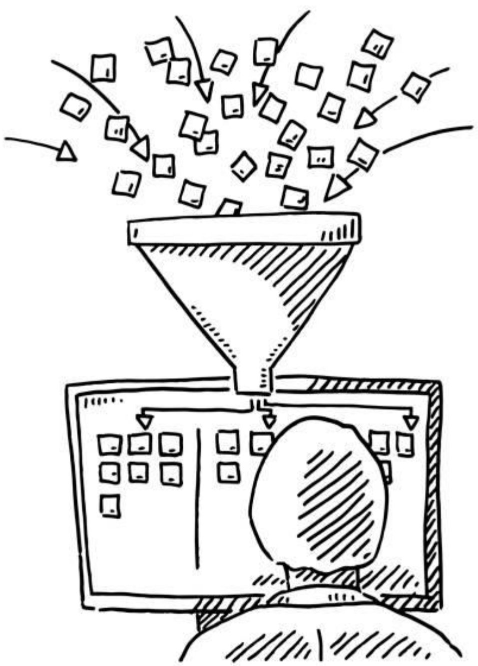


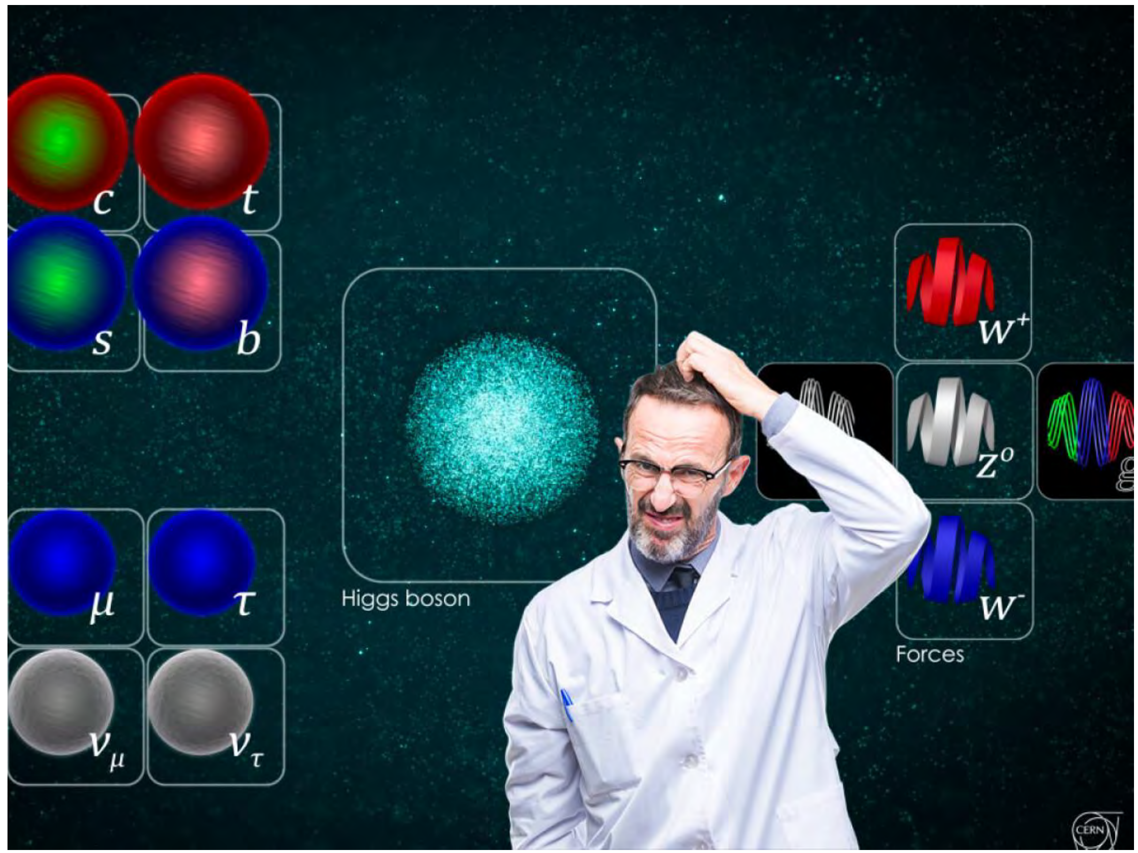
$d(\gamma\gamma) < 10$ mm in calorimeter → QCD jets can mimic photons. Rare cases, however:

$$\frac{\sigma_{jj}}{\sigma(H \rightarrow \gamma\gamma)} \sim 10^8 \qquad m_{\gamma\gamma} \sim 100 \text{ GeV}$$

Data Analysis

- Data vs Theory ⇒ which theories you believe vs. reject
- Significance of final results ⇒ do you trust your analysis or not?





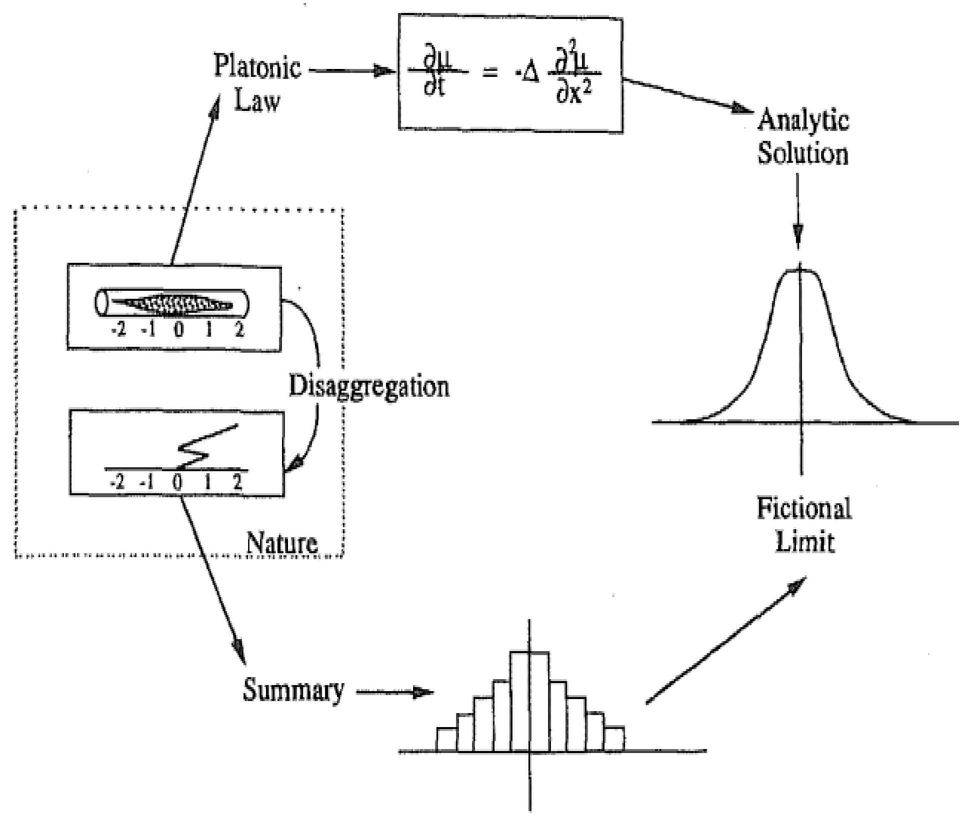
Data Analysis: Theory and Modeling (Monte Carlo Simulation)

Three main goals

- experiment planning
- algorithm's training
- data/MC comparison

Digital Twin of Experiments

- physics in a collision point
- models of detector systems
- response from detectors including digitization
- processing of MC data (simulation of data flow)



Chain of Simulation

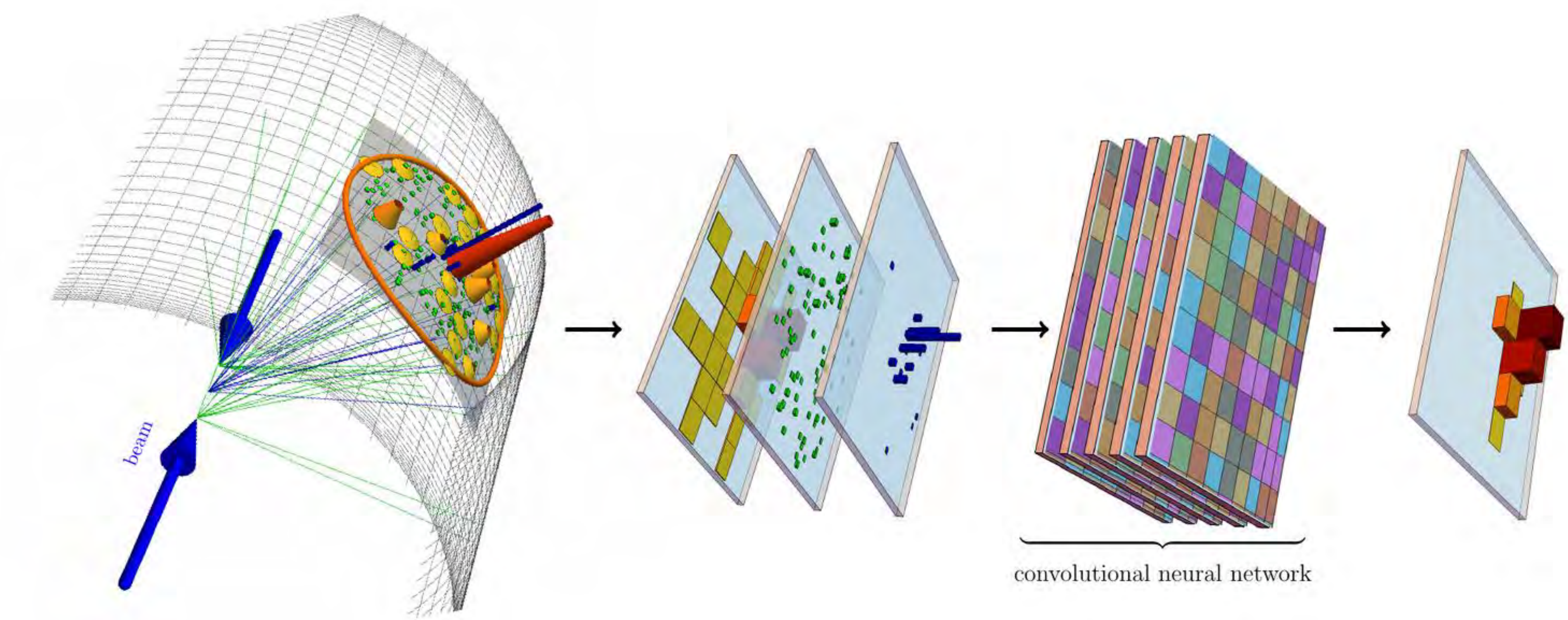
LHE Data

Physics Processes
(Event Generators)

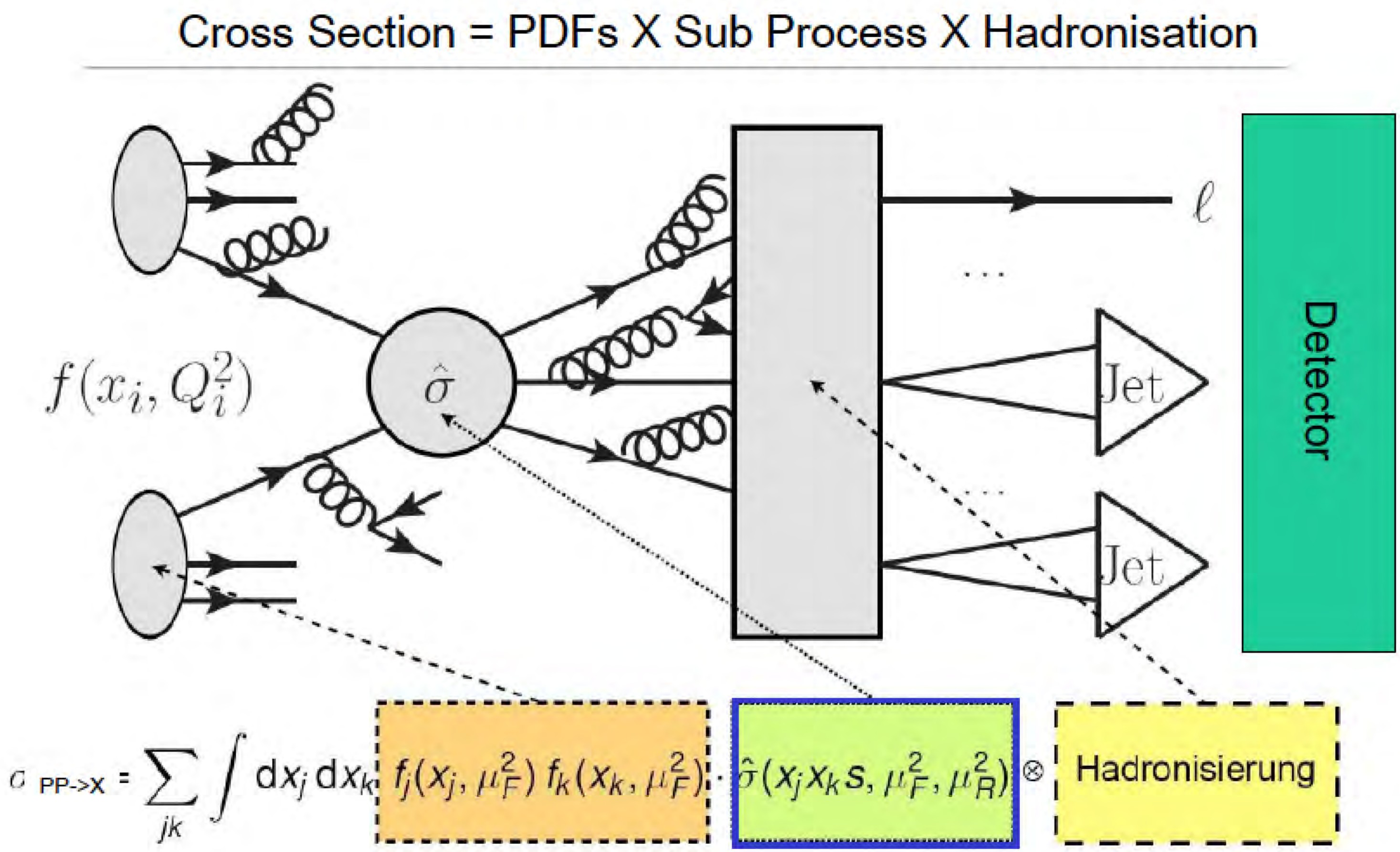
Detector Response
(tracing a particle through matter)

RAW Data

Reconstruction



Theory of Collisions



Event Generators

Three general-purpose generators:

- HERWIG
- PYTHIA
- SHERPA

Many others good/better at some specific tasks.

Generators to be combined with detector simulation (GEANT)

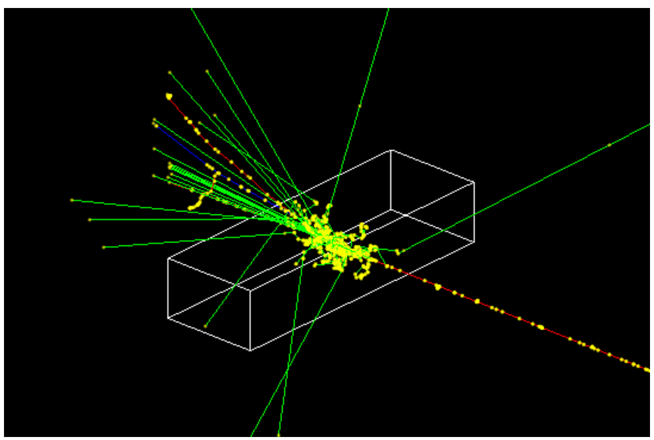
accelerator/collisions ⇔ event generator
detector/electronics ⇔ detector simulation

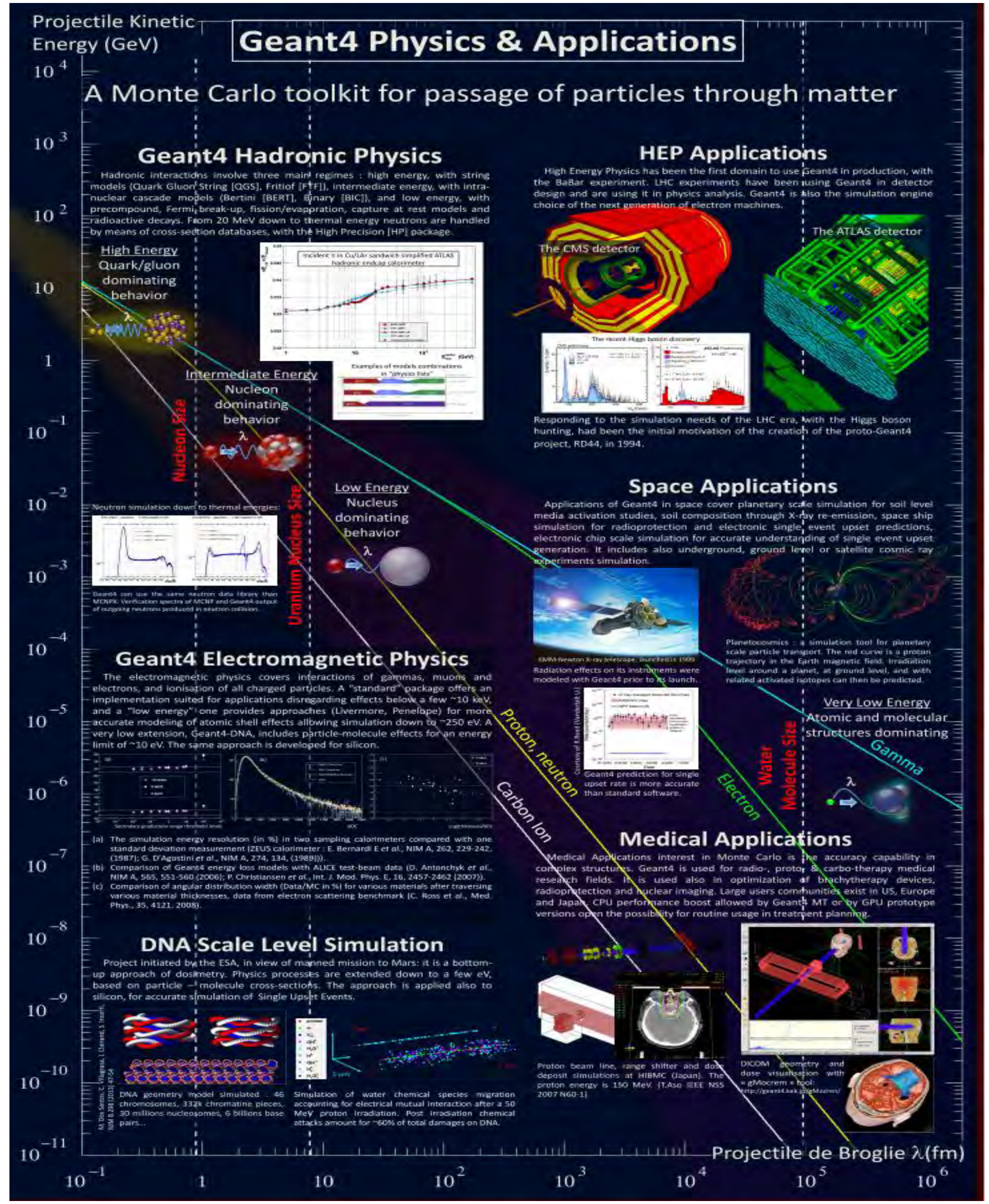
to be used to
- predict event rates and topologies
- simulate possible backgrounds
- study detector requirements
- study detector imperfections

Detector Modeling

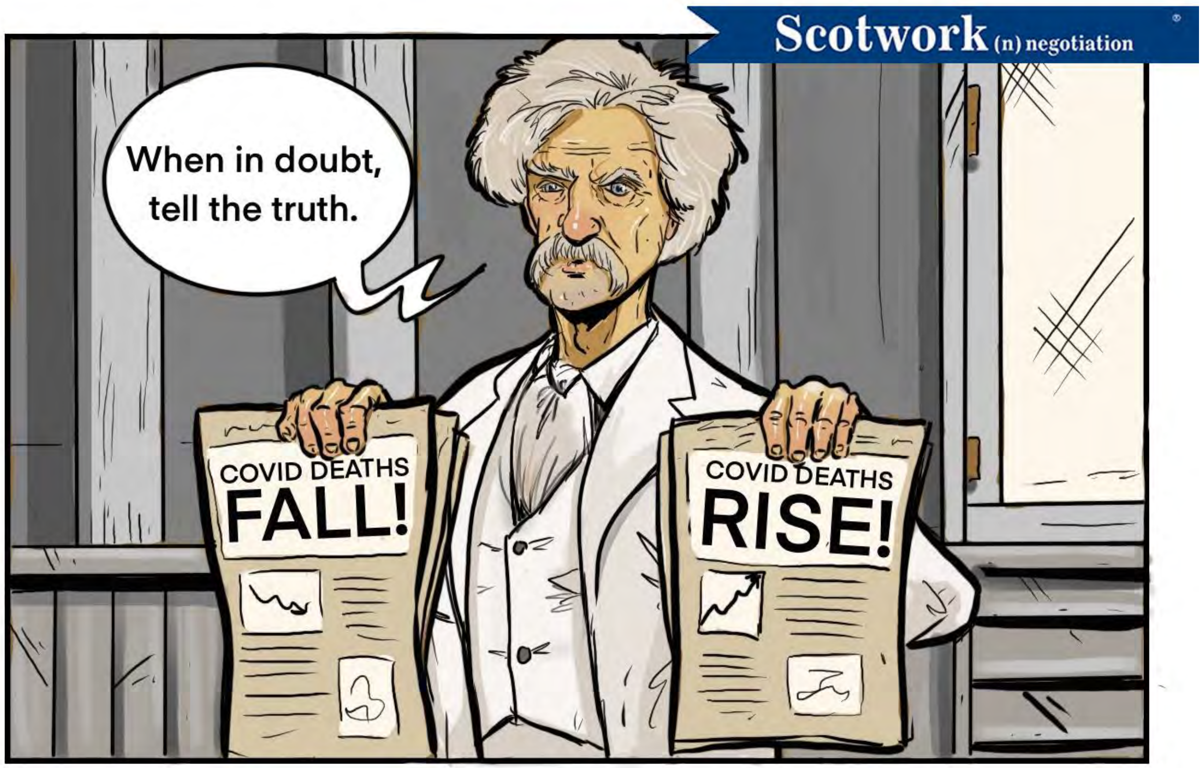
GEANT4

- Toolkit created by CERN to simulate the passage of particles through matter.
- Designed to make the physics used transparent within the toolkit, handle a wide range of geometries, and enable an easy adaptation of different physics to fit the application.





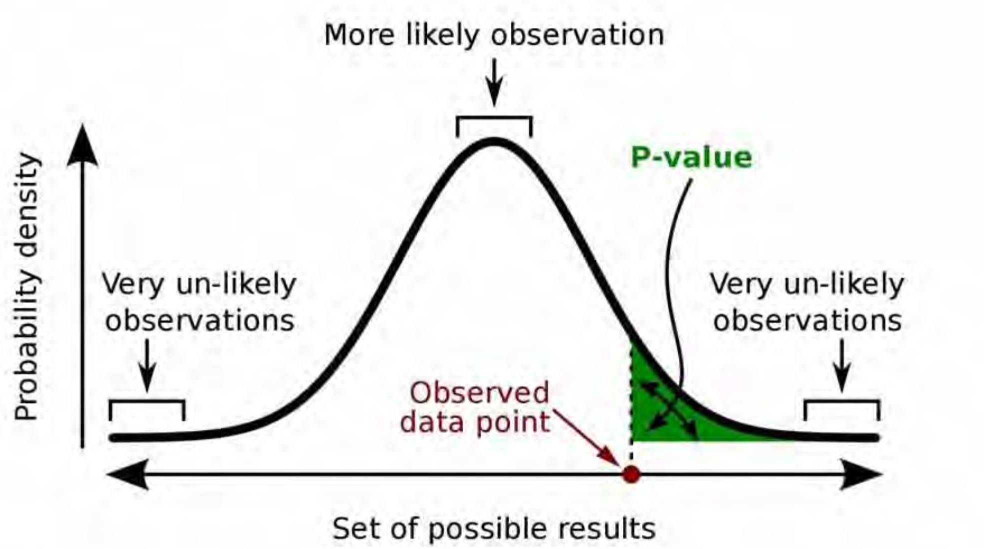
Data Analysis: Statistics



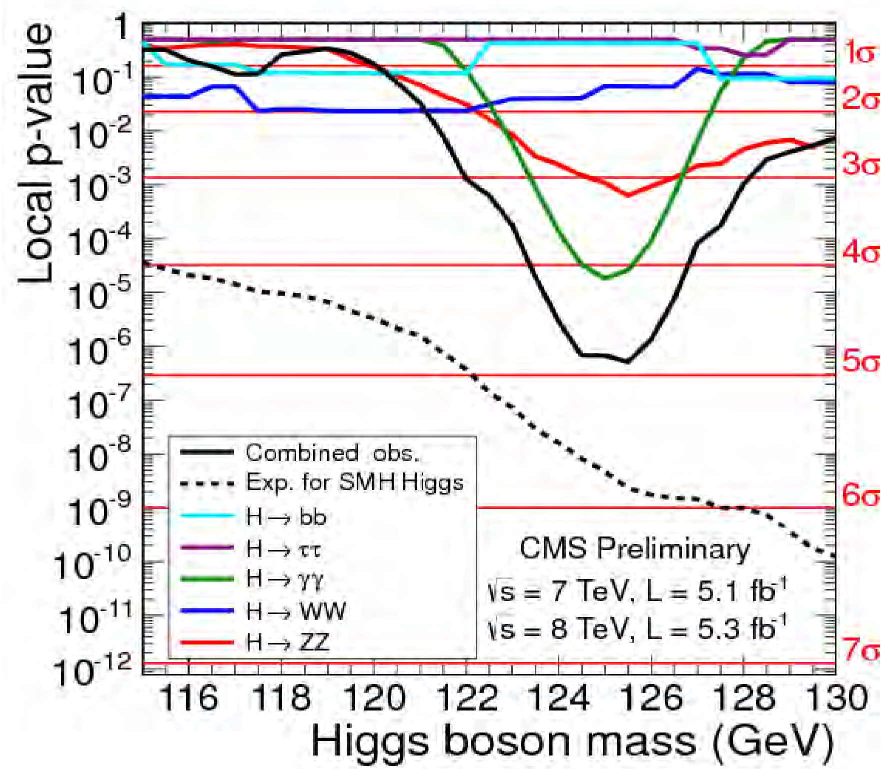
There are three kinds of lies: lies, damned lies, and statistics (c) Benjamin Disraeli

Significance of Discovery

The probability that an observed excess was a statistical fluctuation of the background (p-value)

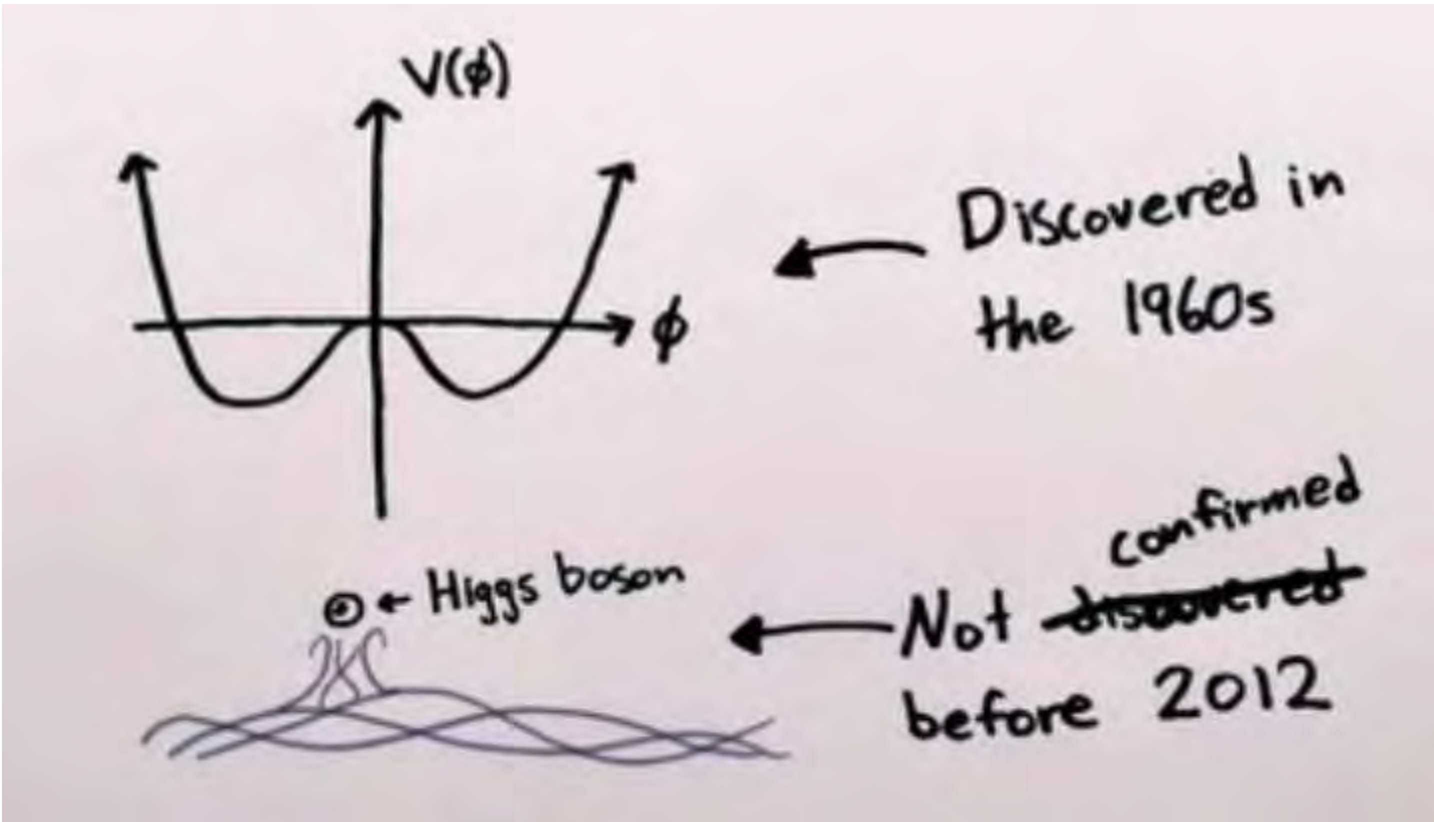


A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

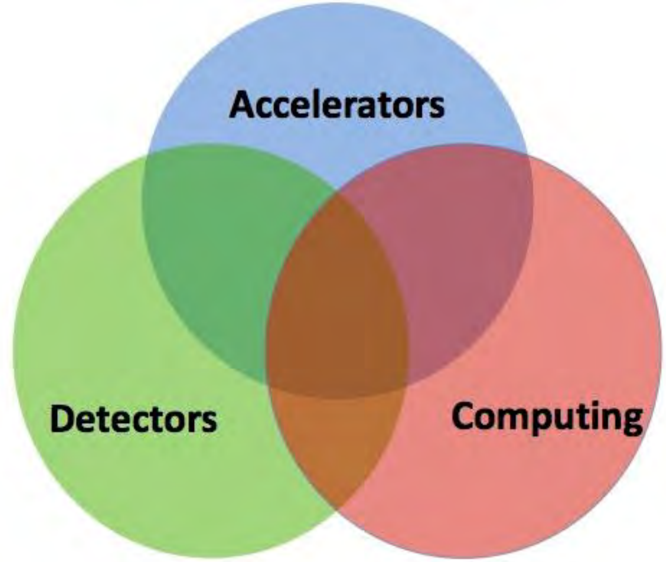


Notable values for an excess in particle physics are **3σ, or p-value = 0.0013**; and **5σ, or p-value = 2.87 x 10^{-7}. When we have an excess of 3σ we talk about an evidence, and when we have an excess of 5σ, we are facing a discovery**.

… and as a result…



Essential Parts of the Success



Accelerators : powerful machines that accelerate particles to extremely high energies and bring them into collision with other particles

Detectors : gigantic instruments that record the resulting particles as they "stream" out from the point of collision.

Computing : to collect, store, distribute and analyse the vast amount of data produced by these detectors

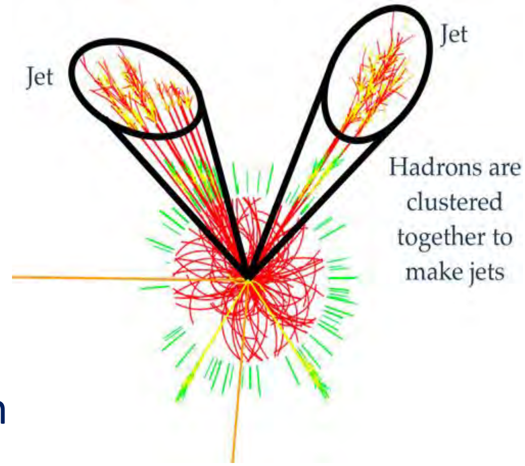
It's been a global effort, a global success. It has only been possible because of the extraordinary achievements of the experiments, infrastructure and the grid computing"
(c) Rolf Heuer, the Director General of CERN, when the discovery of the Higgs

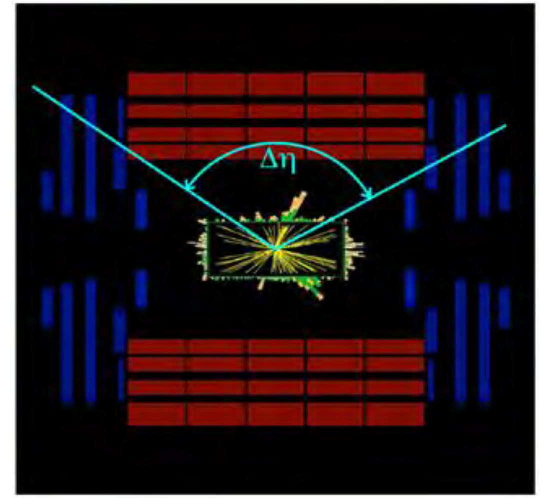
Collaborative Science on Worldwide scale : thousands of scientists, engineers, technicians and support sta**ff to design, build and operate these complex** "machines".

THANK YOU FOR YOUR ATTENTION!

Physics Objects

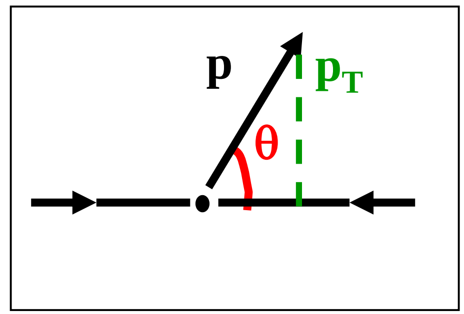
- Muons (transverse momentum p_T)
- Electrons (energy and tr. momentum p_T)
- Photons (energy)
- Jets (energy and coordinates)
- .....
- Missing energy and p_T
  - vectorial sum of all transverse momentum

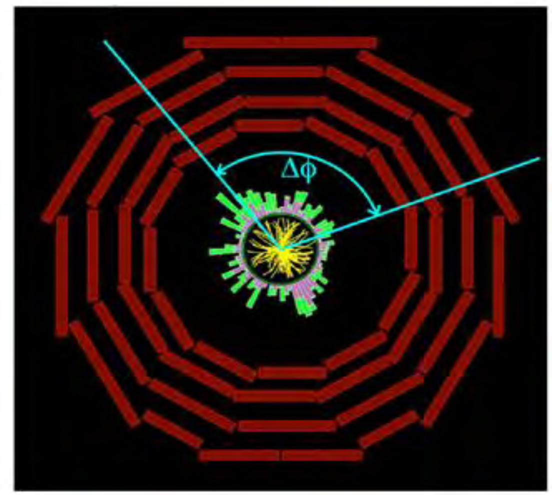




Kinematic Variables

- Transverse momentum p_T (energy)
  - particles that escape detection have p_T=0
  - total visible p_T = 0
- Longitudinal momentum p_Z and energy E_Z
  - particles that escape detection have p_T=0
  - visible p_Z is not conserved (not so usefull variable)
- Angles
  - azimuthal and polar angles
  - polar angle θ is not Lorenz invariant ⇒
  - rapidity y
  - or (or m=0) pseudorapidity η





$$y \equiv \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_z}\right)$$

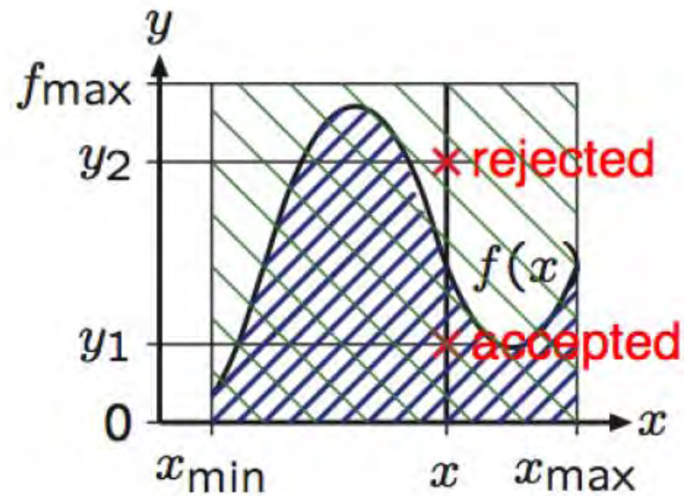
$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]$$

4π-experiments cover 360^0 over ϕ and large pseudorapidity range, |η| ≤ 5.0 (0.8^0)

Hit-and-miss Monte Carlo

If $f(x) \leq f_{\max}$ in $x_{\min} < x < x_{\max}$
use interpretation as an area

1. select
   $x = x_{\min} + R\,(x_{\max} - x_{\min})$
2. select $y = R\, f_{\max}$ (new R!)
3. while $y > f(x)$ cycle to 1



Integral as by-product:

$$I = \int_{x_{\min}}^{x_{\max}} f(x)\,\mathrm{d}x = f_{\max}\,(x_{\max} - x_{\min})\,\frac{N_{\mathrm{acc}}}{N_{\mathrm{try}}} = A_{\mathrm{tot}}\,\frac{N_{\mathrm{acc}}}{N_{\mathrm{try}}}$$

Binomial distribution with $p = N_{\mathrm{acc}}/N_{\mathrm{try}}$ and $q = N_{\mathrm{fail}}/N_{\mathrm{try}}$,
so error

$$\frac{\delta I}{I} = \frac{A_{\mathrm{tot}}\sqrt{p\,q/N_{\mathrm{try}}}}{A_{\mathrm{tot}}\,p} = \sqrt{\frac{q}{p\,N_{\mathrm{try}}}} = \sqrt{\frac{q}{N_{\mathrm{acc}}}} \longrightarrow \frac{1}{\sqrt{N_{\mathrm{acc}}}} \quad \text{for } p \ll 1$$

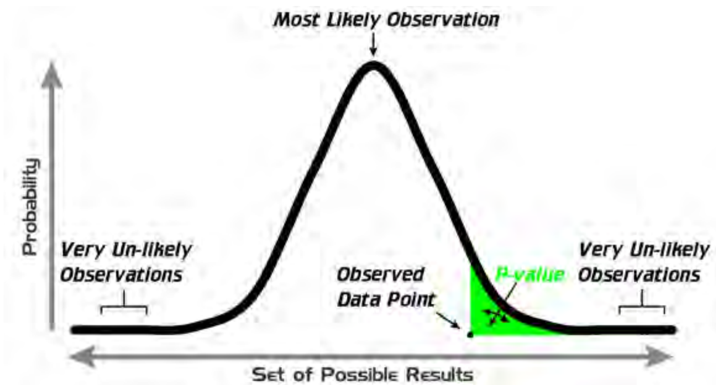
Статистический анализ

Событие (результат) называется “статистическими значимым”, если оно вряд ли произошло случайно

p-value - вероятность получить результат, такой как наблюдается (или выше) в предположении, что нуль-гипотеза верна

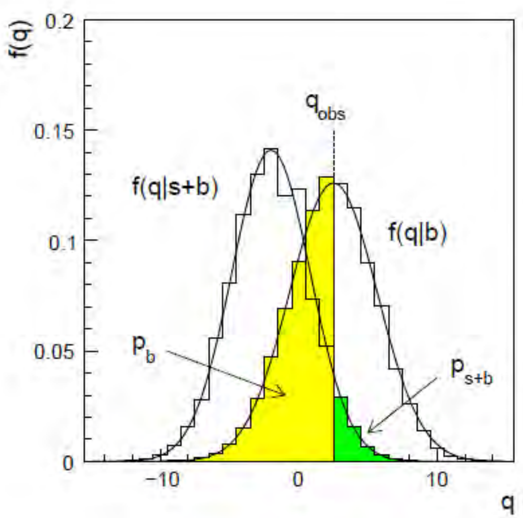
⇒ в нашем случае вероятность, того, что флуктуация фона достигли (или превысили) наблюденное значение

$$p = P(\, n \geq n_{obs} \mid b \,)$$



Нуль-гипотеза – основная проверяемая гипотеза (фон)

⇒ Нулевая гипотеза отвергается, когда значение p-value меньше уровня стат. значимости α (по соглашению <0.05)



Масштабный фактор (strength factor)

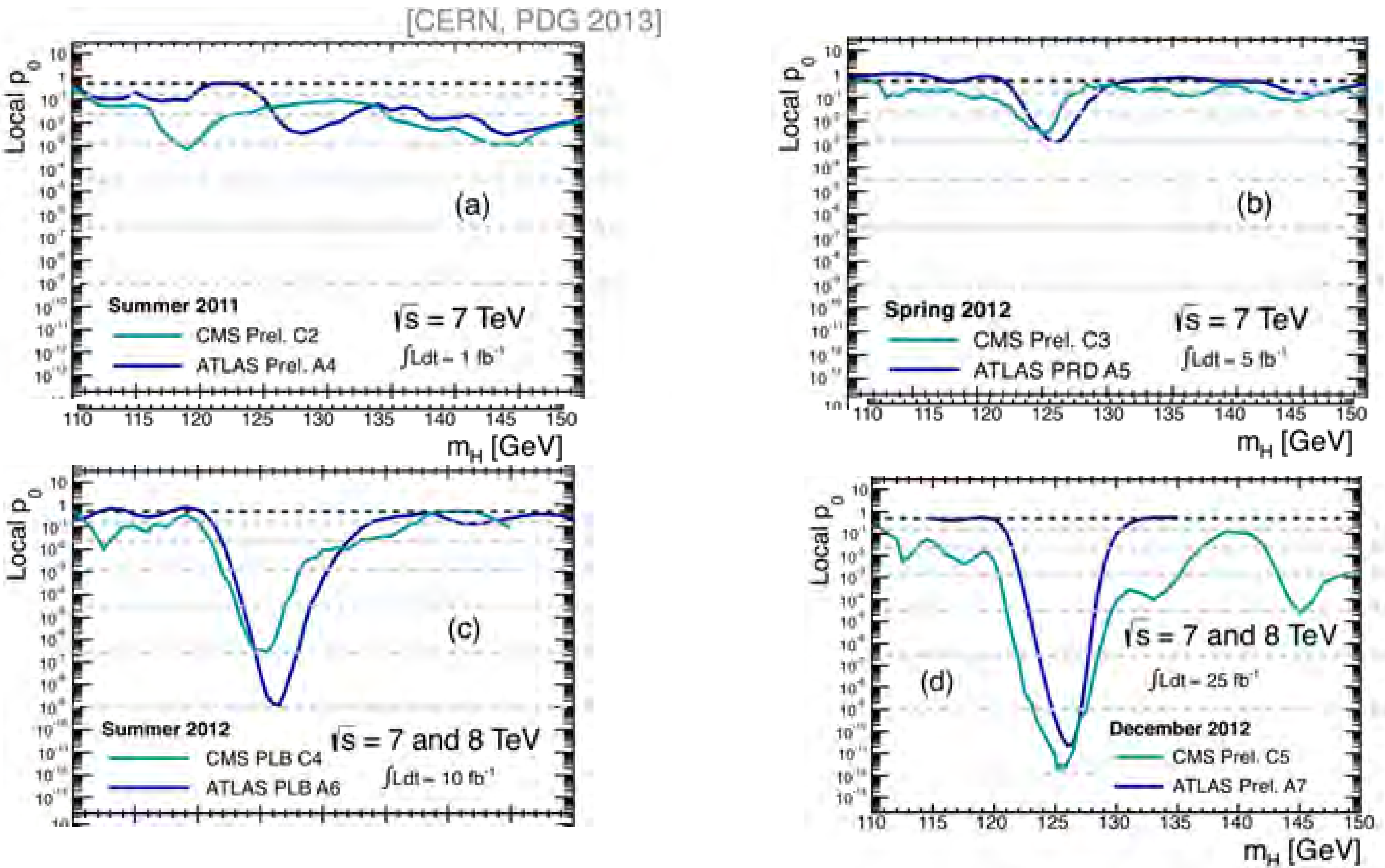
$$\mu = \frac{\sigma}{\sigma_{SM}} < \mu^{95\%} \text{ at 95\% C.L., e.g. } \mu^{95\%} = 1 \Rightarrow \text{exclusion}$$

σ_{SM} – сечение бозона Хиггса в СМ, σ - гипотетическое сечение бозона Хиггса

$$CL_S(\mu^{95\%}) = \frac{CL_{S+B}}{CL_B} = \frac{P(q_\mu > q_\mu{}^{obs} | B + \mu^{95\%} \times S)}{P(q_\mu > q_\mu{}^{obs} | B\ )} = 0.05$$

$$q_\mu = -2 \ln \frac{\mathcal{L}(\text{data} \mid \mu, \hat{\hat{\theta}}_\mu)}{\mathcal{L}(\text{data} \mid \hat{\mu}, \hat{\theta})},$$

Story at Higgs Discovery



What does Brazilian Flag mean?

Dimuon example
$$R_\sigma = \frac{\sigma(pp \to Z' + X \to l^+l^- + X)}{\sigma(pp \to Z^0 + X \to l^+l^- + X)}$$

Extended gauge models

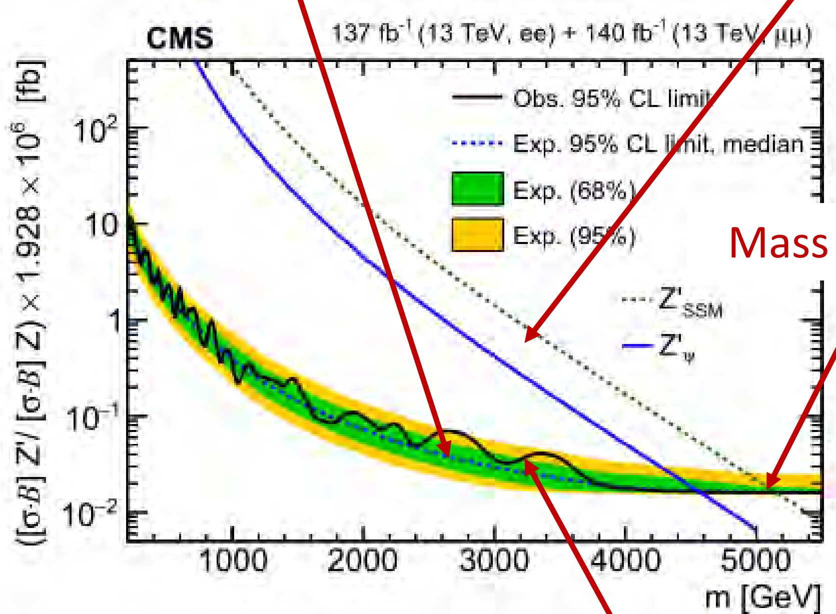
Models of low-energy gravity (RS1-type scenario of ED)

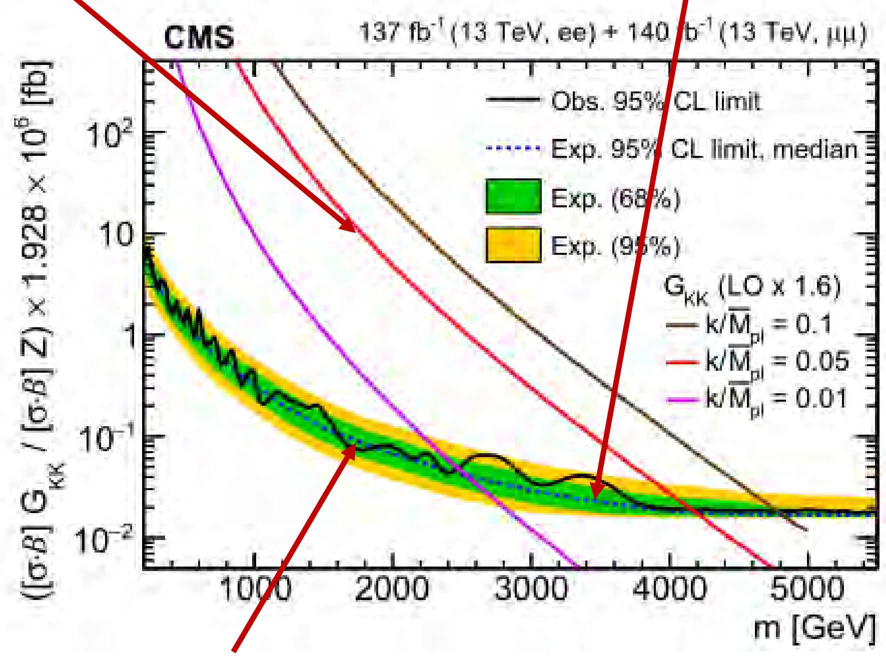
SM predictions

BSM predictions

SM predictions

Mass limit



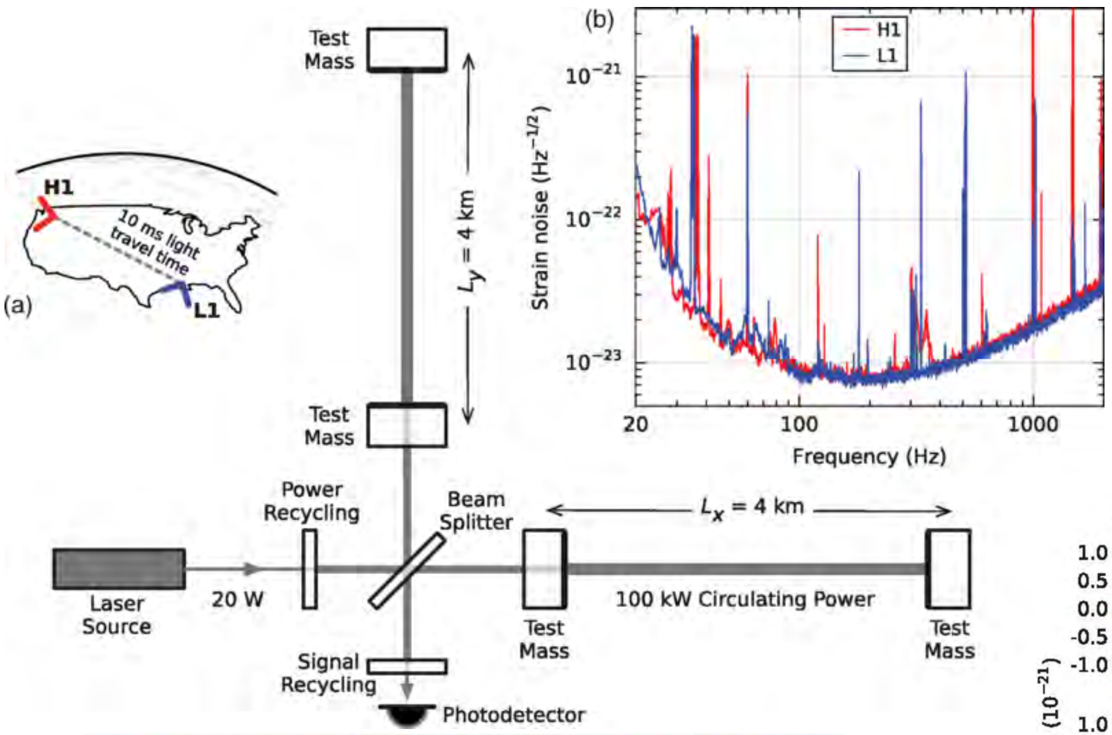


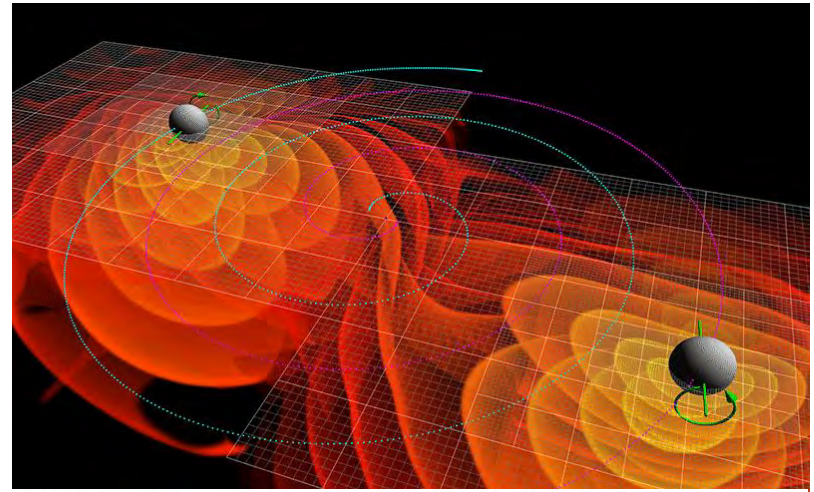
Model-independent limits on cross section (in narrow width approximation, NWA)

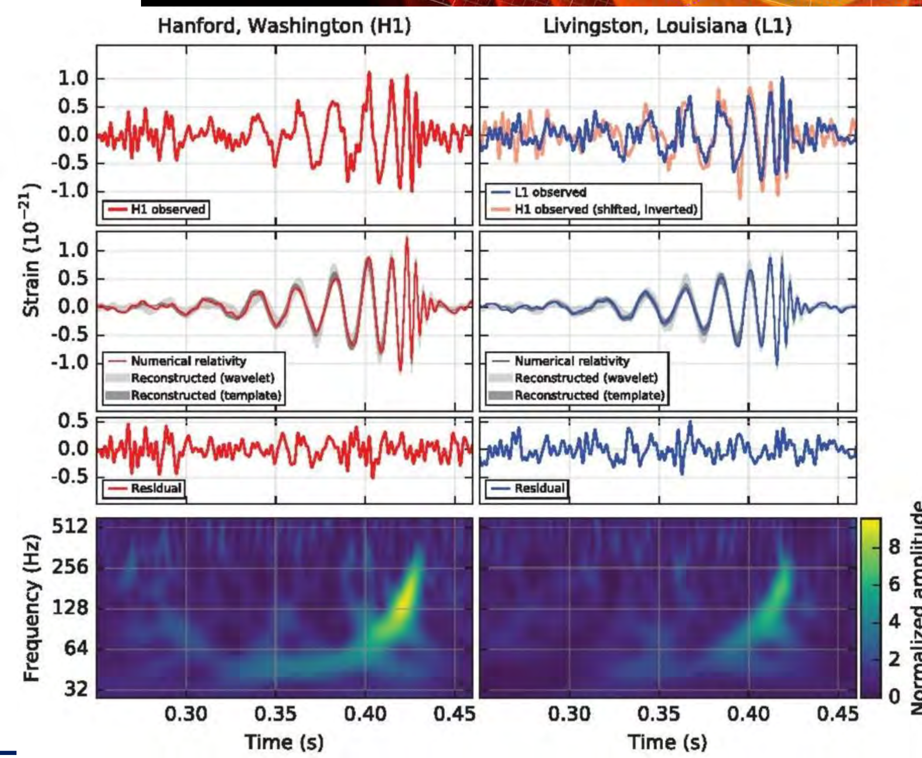
Channel	Z'_{SSM} Obs. [TeV]	Z'_{SSM} Exp. [TeV]	Z'_ψ Obs. [TeV]	Z'_ψ Exp. [TeV]
e e	4.72	4.72	4.11	4.13
$\mu^+\mu^-$	4.89	4.90	4.29	4.30
e e + $\mu^+\mu^-$	5.15	5.14	4.56	4.55

Channel	$k/\overline{M}_{Pl} = 0.01$ Obs. [TeV]	$k/\overline{M}_{Pl} = 0.01$ Exp. [TeV]	$k/\overline{M}_{Pl} = 0.05$ Obs. [TeV]	$k/\overline{M}_{Pl} = 0.05$ Exp. [TeV]	$k/\overline{M}_{Pl} = 0.1$ Obs. [TeV]	$k/\overline{M}_{Pl} = 0.1$ Exp. [TeV]
e e	2.16	2.29	3.70	3.83	4.42	4.43
$\mu^+\mu^-$	2.34	2.32	3.96	3.96	4.59	4.59
e e + $\mu^+\mu^-$	2.47	2.53	4.16	4.19	4.78	4.81

Observation of Gravitational Waves



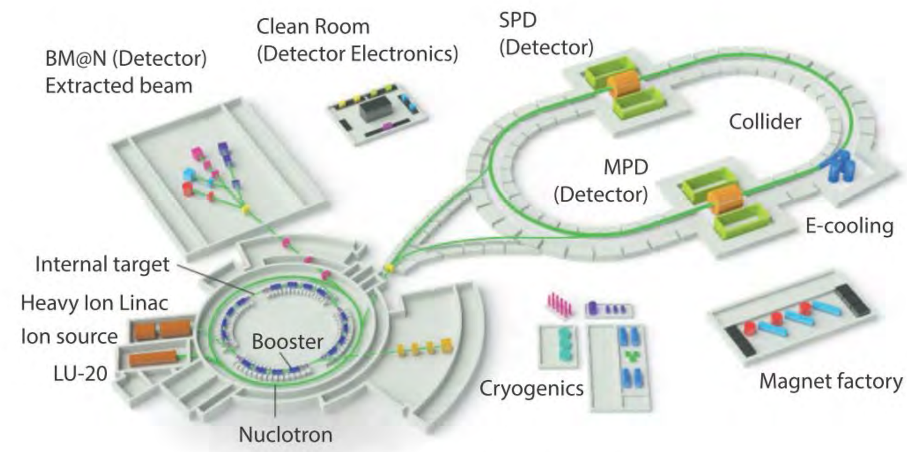




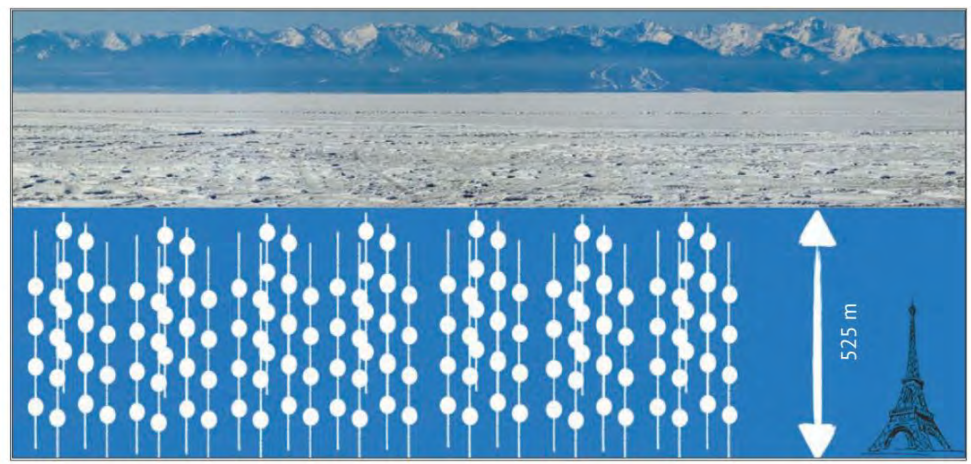
JINR in Particle Frontiers

JINR LONG-TERM DEVELOPMENT STRATEGIC PLAN UP TO 2030 AND BEYOND

- RELATIVISTIC HEAVY-ION PHYSICS AT NICA
- JINR PARTICIPATION IN FOREFRONT EXTERNAL EXPERIMENTS OFF-SITE
  - LHC, SPS, RHIC, and at facilities under construction, as for example the FAIR facility
- NICA SPIN PHYSICS
- PARTICLE PHYSICS AT THE LHC AND BEYOND
  - Accelerator-based research and frontier accelerator technologies (LHC, SPS, NICA, FAIR, etc)
  - Neutrino physics and astroparticle physics (Baikal-GVD, JUNO, NOvA, DUNE, etc)
  - Multi-messenger astronomy including gravitational wave detection (Baikal-GVD, TAIGA, VIRGO, etc)

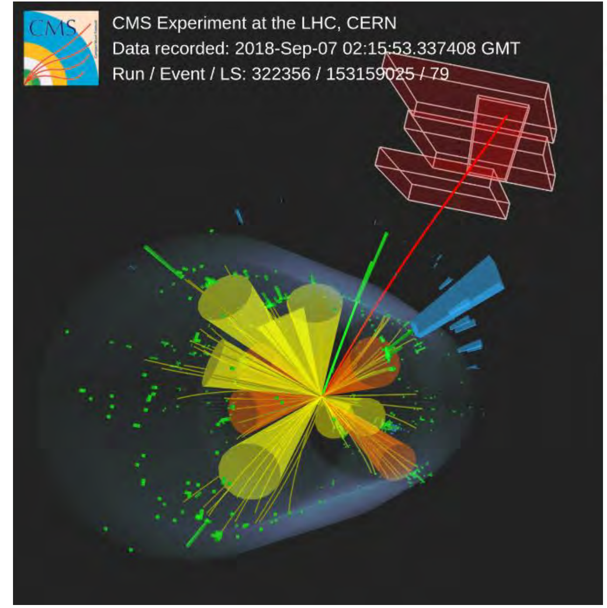


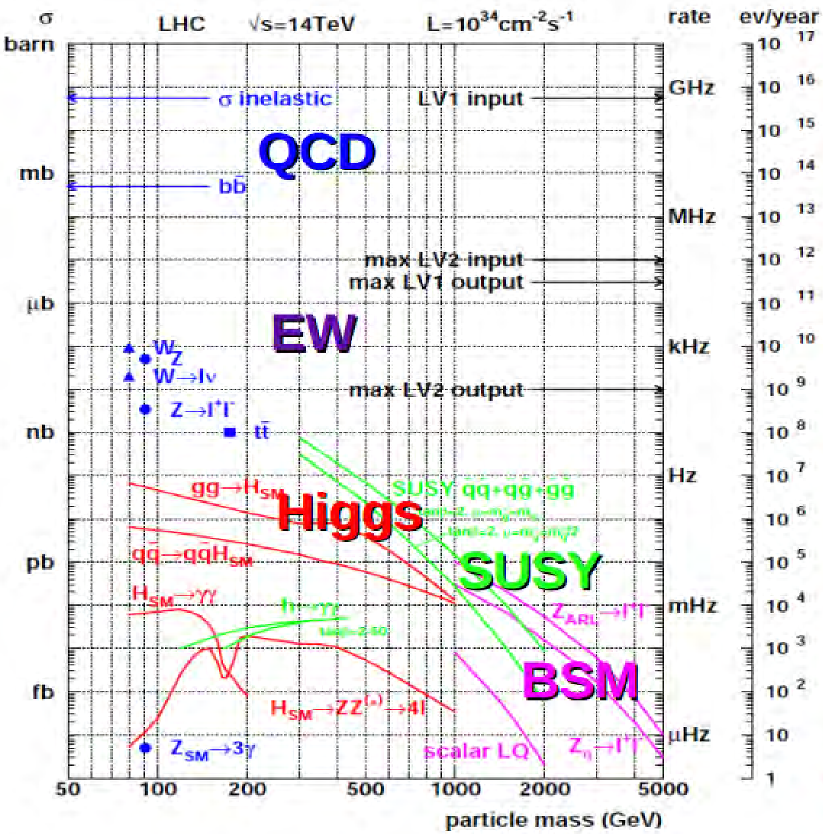
The NICA accelerator complex



Baikal-GVD (Gigaton Volume Detector)

What do we know today about the Standard Model from LHC?





- SM processes:

  $\sigma \sim 1/(100\ \text{MeV})^2$

10^{-8} !

- New Physics:

  $\sigma \sim 1/(1\ \text{TeV})^2$

During Run 2 the LHC produced 10^{16} collisions

Large samples of various particles produced:

- W bosons: 12 billion
- Z bosons: 2.8 billion
- Top quarks: 300 million
- B quarks: 40 trillion
- Higgs bosons: 7.7 million