Understanding the Monte Carlo methods applied in HEP analysis

Sandro Fonseca de Souza

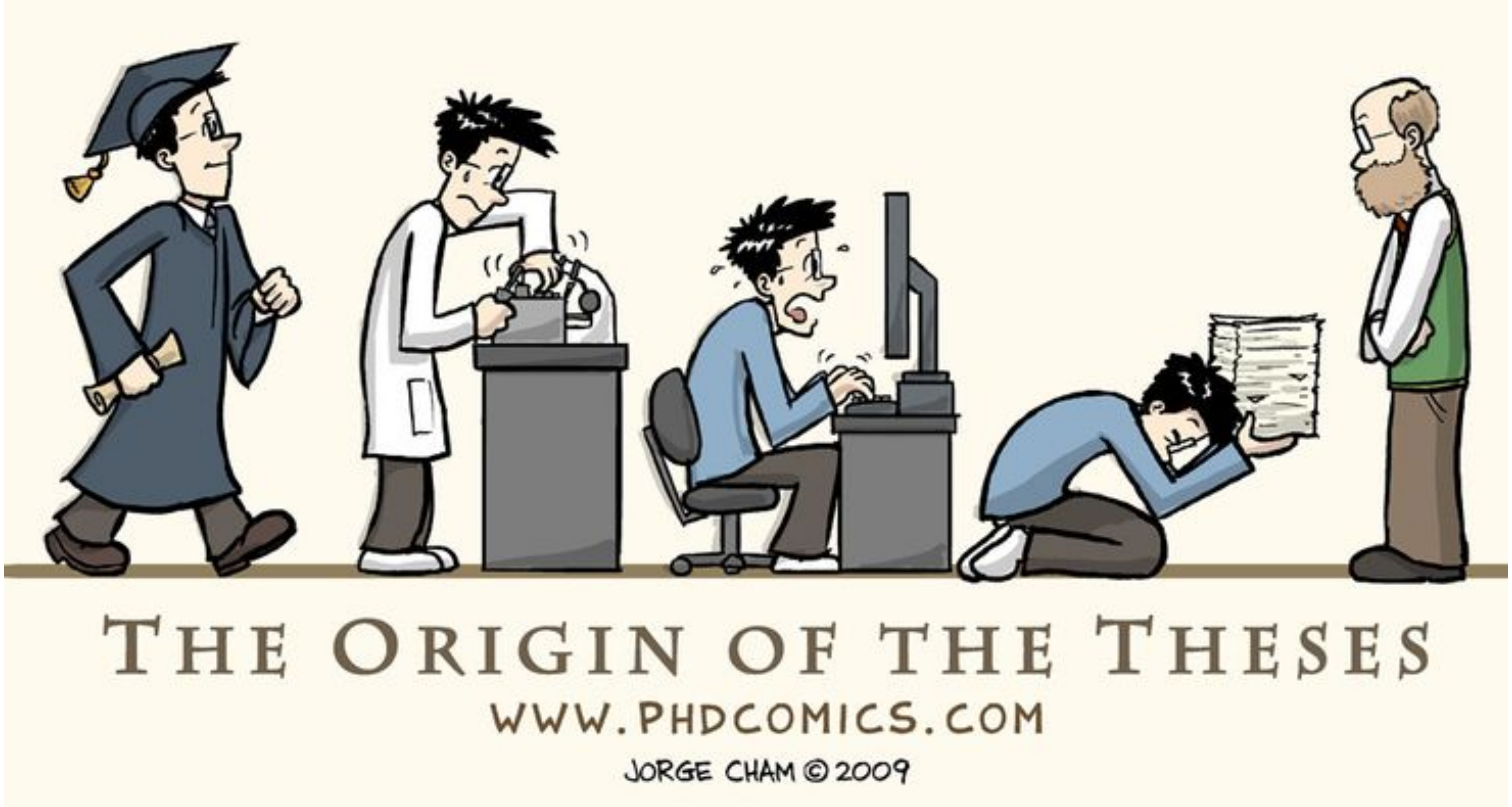
16 June 2020

Outline

Understanding the Monte Carlo methods applied in HEP analysis

- Motivations
- The Higgs Boson Discovery Explained - Youtube
- Monte Carlo Simulation in HEP
- Detector Simulation - CMS Experiment
- Measurement of Cross Section and evaluate these uncertainties
- Weighting Procedures
- MC Scale Factors
- Pile-Up reweighing contributions
- Acceptance and Purity (preliminary definitions)
- Summary and Conclusions
- References

Motivations



Motivations

PhD Defense Thesis April 25 2010

Motivations

Eram Rizvi

Atlas Collaboration

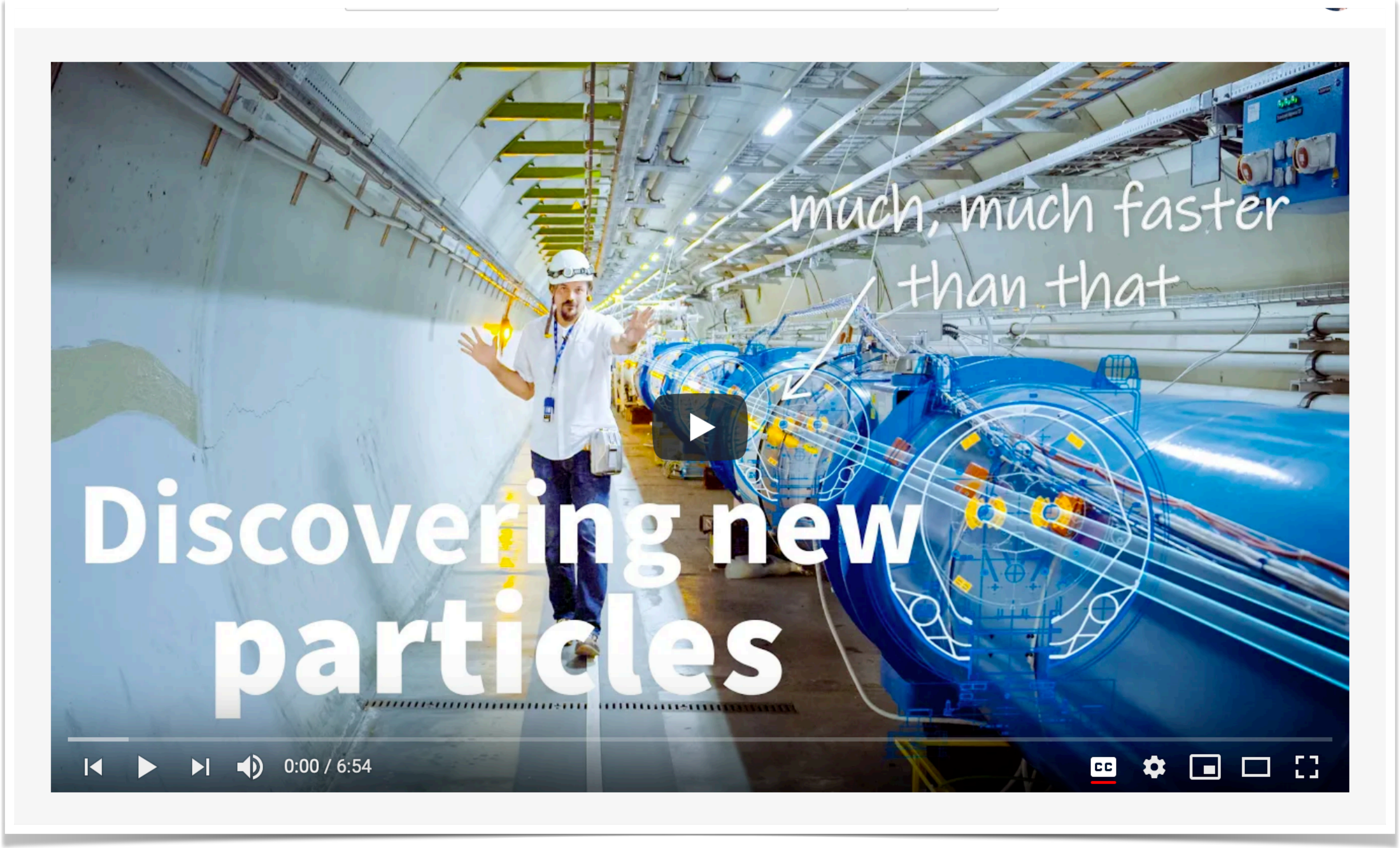
Aim to introduce methods useful for your analysis

Not found in textbooks - HEP folklore...

Details depend crucially on analysis

Only generalities given

The Higgs Boson Discovery Explained - Youtube



reading suggestion

Although Everett didn't see Sauerland's presentation, he suggests that the remarks could equally well be translated as two separate, direct sentences, such as: "The nut is under the banana leaf. Or so Oope says", or "Oope doesn't know. Where is the nut?"

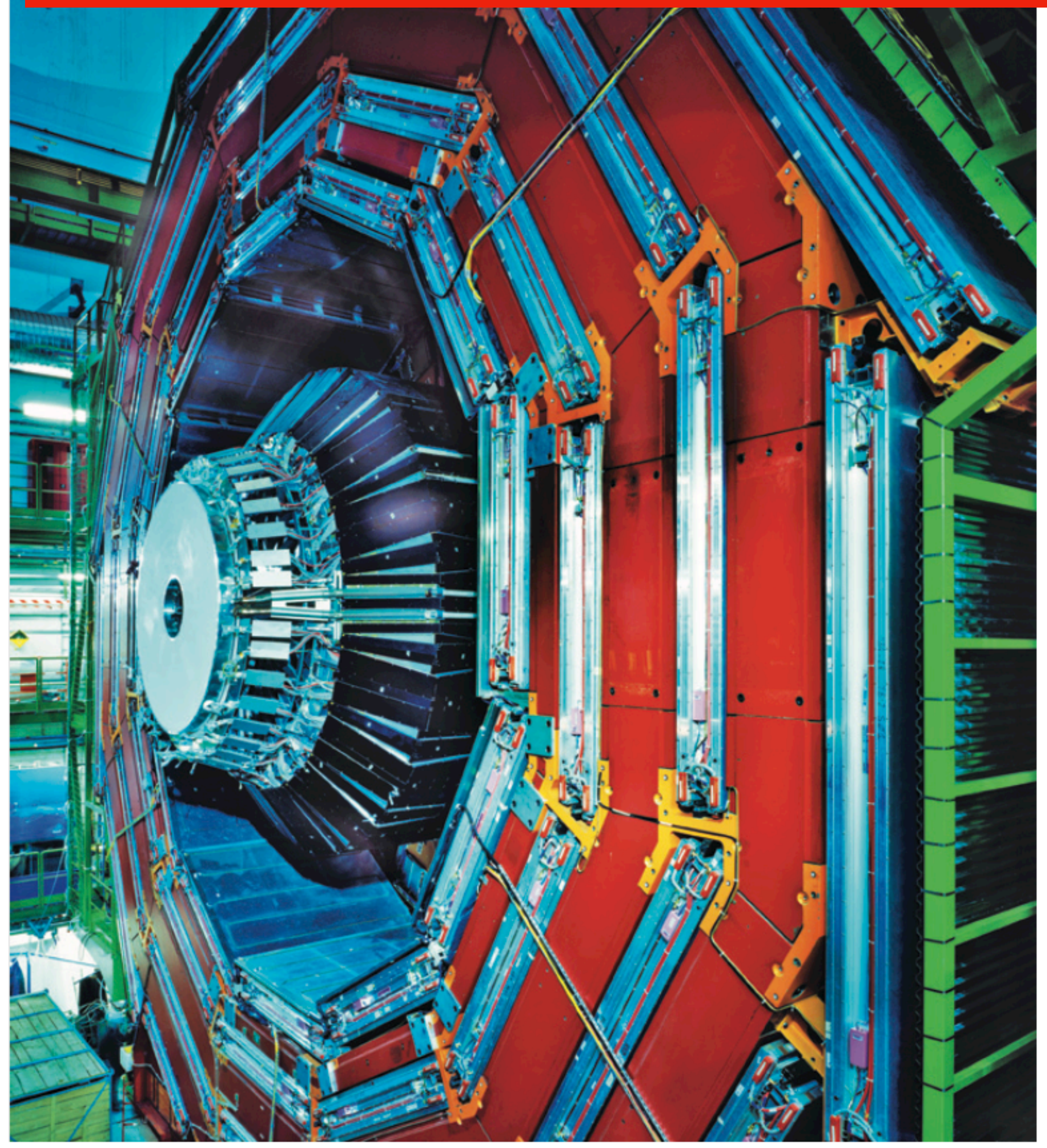
The difference is subtle, but it cuts to the heart of Everett's case against Chomsky's theory. Embedded clauses can be instances of recursion, an iterative process that Chomsky says is essential to all language because it enables ever more complex sentences to be built up out of individual words or sounds. Everett also says that Pirahã lacks colour and number terms and has no perfect tense, which is used in English for events that have been completed. Although many linguists say that Chomsky's theory of a universal grammar would hold even if Everett is right about those features, Everett believes that such a profound interplay between culture and language conflicts with Chomsky's theory of language as innate.

The situation underscores the potential difficulty in settling important claims about minority languages. The United Nations Educational, Scientific and Cultural Organization lists 2,473 languages as endangered, meaning either that they are spoken by only small communities of people or that the elderly people who speak them have not passed them on to subsequent generations. Many such languages have been studied by just a single linguist, so that other researchers must rely on that person's translations.

"For a lot of languages we have extremely poor documentation," says Lyle Campbell, a linguist at the University of Hawaii at Manoa who is leading ELCat, an online project supported by the US National Science Foundation that aims to catalogue endangered languages. Expected to launch later this month, ELCat will serve as a centralized repository for original data such as recordings, video, text, transcripts and translations. Campbell says that such documentation makes it possible for linguists to test each others' statements.

Thomas Roeper, a linguist at the University of Massachusetts in Amherst, says that linguists will inevitably have to work with data from a limited number of sources. "There are many languages that only one, two or three people have studied, with Western prejudices. It would be a great mistake if we didn't include their experiences in our knowledge," he says.

Everett and his colleagues are now testing his arguments using data on Pirahã collected by his missionary predecessor, Steve Sheldon. Everett is also working on making his own records available. "I have data recorded, and am translating more and more," he says. ■



ENRICO SACCHETTI

The Compact Muon Solenoid experiment detects hundreds of millions of particle collisions every second.

PARTICLE PHYSICS

LHC prepares for data pile-up

Physicists scramble to see through fog of collisions.

BY GEOFF BRUMFIEL

The world's largest particle accelerator is roaring along at an unprecedented pace, delivering torrents of data to its physicist handlers. But the hundreds of millions of collisions happening inside the machine every second are now growing into a thick fog that, paradoxically, threatens to obscure a fabled quarry: the Higgs boson.

The problem is known as pile-up, and it promises to be one of the greatest challenges this year for scientists working on the Large Hadron Collider (LHC) at CERN, Europe's main high-energy physics laboratory near Geneva, Switzerland.

Huge amounts of computing power, cunning software and technical tricks are helping scientists to stay ahead of the problem. But researchers may still need to scale back the collisions to find the long-sought Higgs, the manifestation of a field that is believed to confer mass on other particles.

If it exists, the Higgs will appear fleetingly inside the machine before decaying into lighter particles. Last year, the two biggest detectors at the LHC saw hints of a Higgs with a mass of about 125 gigaelectronvolts (energy and mass are interchangeable in particle physics). This year, researchers want to collect more data to see whether that signal grows into a certainty, or withers back to nothing.

Since it began its latest science run last month, the LHC has been squeezing trillions of protons into ever-smaller bunches, and smashing those bunches together tens of millions of times per second. The resultant data are measured in inverse femtobarns (fb^{-1}), a unit roughly equivalent to 100 trillion collisions. In the past month alone, the LHC recorded 1 fb^{-1} worth of collisions. By the end of the year it aims have captured at least 15 fb^{-1} (see 'Smashing!').

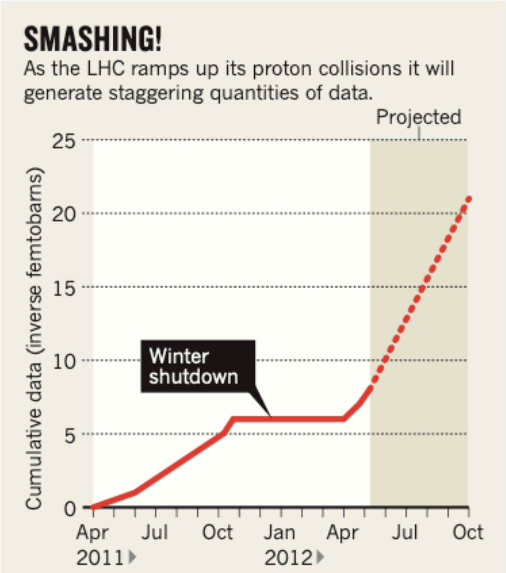
➲ NATURE.COM
To read more articles on the LHC, see:
go.nature.com/uzvtfd

To gather these data, researchers are pushing the collider in two ways: by accelerating the particles to ever-greater energies and by increasing the number of collisions. Higher energies allow heavier particles to pop into being, but it is the number of collisions that will determine whether physicists have enough data to declare a discovery. In the weeks ahead, scientists will pack more protons inside the machine and focus the particles as tightly as possible onto the collision points at the centre of the LHC's two biggest detectors. Already, "we've done humongously better than we thought we could", says Mike Lamont, the head of accelerator operations at CERN.

Every time two tightly packed bunches of protons cross, they generate not one collision, but on average 27, Lamont says. But within a few weeks, that number is expected to rise into the mid-30s, peaking at around 40 collisions per crossing. The two main detectors at the LHC were designed to handle only around two dozen collisions at once. But they have managed to cope so far.

Each detector is made up of layers of smaller detectors that record the tracks of debris coming from their centre. When a collision occurs, computers above the machine decide whether the data are interesting and, if so, reconstruct the collision from the tracks. But when dozens of collisions occur at once, the computers must disentangle them.

SOURCE: CERN



Last year, researchers working with the ATLAS detector formed a task force to tackle the pile-up problem, rewriting computer code so that the detector could cope with the extra collisions. Team member Andreas Salzburger says that the group has been working hard to weed out the 'ghost' particles that appear when the paths of several particles align, creating the illusion of a particle that is not actually there. Eliminating these ghosts as early as possible reduces the amount of computing power needed to crunch useful data, he says.

At the Compact Muon Solenoid (CMS), ATLAS's rival detector, physicists have trained their algorithms to triage data on the fly, analysing particle tracks in order of complexity. "Did you ever play the game 'pick-up sticks'?," asks Joe Incandela, the spokesman for the CMS. "You pick up the easiest ones first, and it makes it simpler to deal with the other ones." The team is also working on ways to get rid of signals from 'loopers', low-energy particles that spiral along the detector's magnetic field lines, generating data that are irrelevant to the Higgs hunt.

Such tricks are likely to be less effective as the number of collisions rise. At the outer edges of the machine, the detector segments are larger and have coarser resolution, so it might not be possible to disentangle some of the tracks. That could reduce a detector's ability to pick up one signature of the Higgs: a decay to a pair of W bosons, which causes a cascade of particles that need to be caught by these outer segments.

For now, the mountains of extra data should offset what is lost to pile-up. Researchers expect to miss no more than 15% of events from the most likely Higgs decay pathway, which produces two γ-rays. And if ATLAS and the CMS can't handle the extra particles surging through the machine, Lamont says, the accelerator physicists are ready to dial it back. But "if they can take it, we will give it to them", he says. ■

PHYSICS

A boost for quantum reality

Theorists claim they can prove that wavefunctions are real states.

BY EUGENIE SAMUEL REICH

The philosophical status of the wavefunction — the entity that determines the probability of different outcomes of measurements on quantum-mechanical particles — would seem to be an unlikely subject for emotional debate. Yet online discussion of a paper claiming to show mathematically that the wavefunction is real has ranged from ardently star-struck to downright vitriolic since the article was first released as a preprint in November 2011.

The paper, thought by some to be one of the most important in quantum foundations in decades, was finally published last week in *Nature Physics* (M. F. Pusey, J. Barrett & T. Rudolph *Nature Phys.* http://dx.doi.org/10.1038/nphys2309; 2012), enabling the authors, who had been concerned about violating the journal's embargo, to speak about it publicly for the first time. They say that the mathematics leaves no doubt that the wavefunction is not just a statistical tool, but rather, a real, objective state of a quantum system. "People have become emotionally attached to positions that they defend with vague arguments," says Jonathan Barrett, one of the authors and a physicist at Royal Holloway, University of London. "It's better to have a theorem."

The authors have some heavyweights in their corner: their view was once shared by Austrian physicist and quantum-mechanics pioneer Erwin Schrödinger, who proposed in his famous thought experiment that a quantum-mechanical cat could be dead and alive at the same time. But other physicists have favoured an opposing view, one held by Albert Einstein: that the wavefunction reflects the partial knowledge an experimenter has about a system. In this interpretation, the cat is either dead or alive, but the experimenter does not know which. This 'epistemic' interpretation, many physicists and philosophers argue, better explains the phenomenon of wavefunction collapse, in which a quantum state is fundamentally changed by measuring it.

Barrett and his colleagues are following the approach of physicist John Bell, who in 1964 proved that quantum mechanics has another counterintuitive implication: that measurements on one particle can influence the state of another, distant particle, faster than the speed of light should allow. Bell's was a 'no-go' theorem: its strategy was to show that theories that do not allow faster-than-light ▸

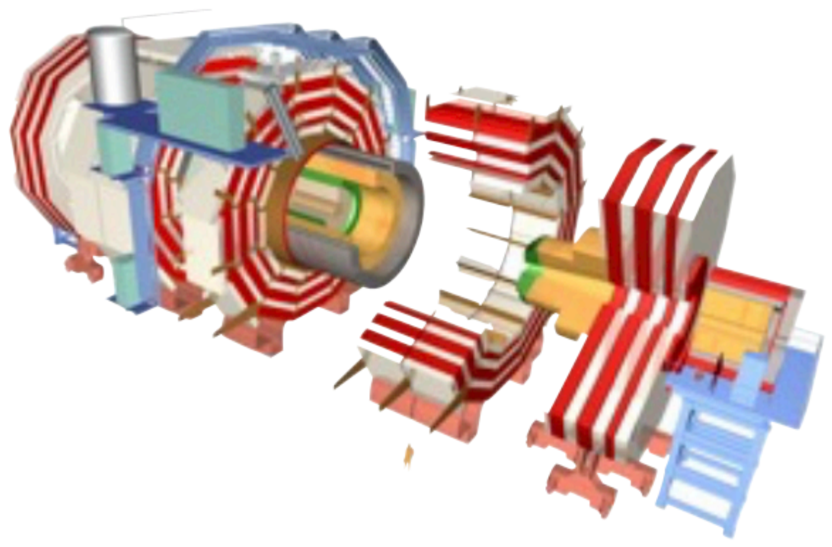
Monte Carlo Simulation in HEP

Choose model, constraints, parameters, decay chain of interest

Proposed Theory

Kinematics, information from a known (detectable) particles

Generator



Experiment Triggering

Detector Simulation
- Hardware
- Software

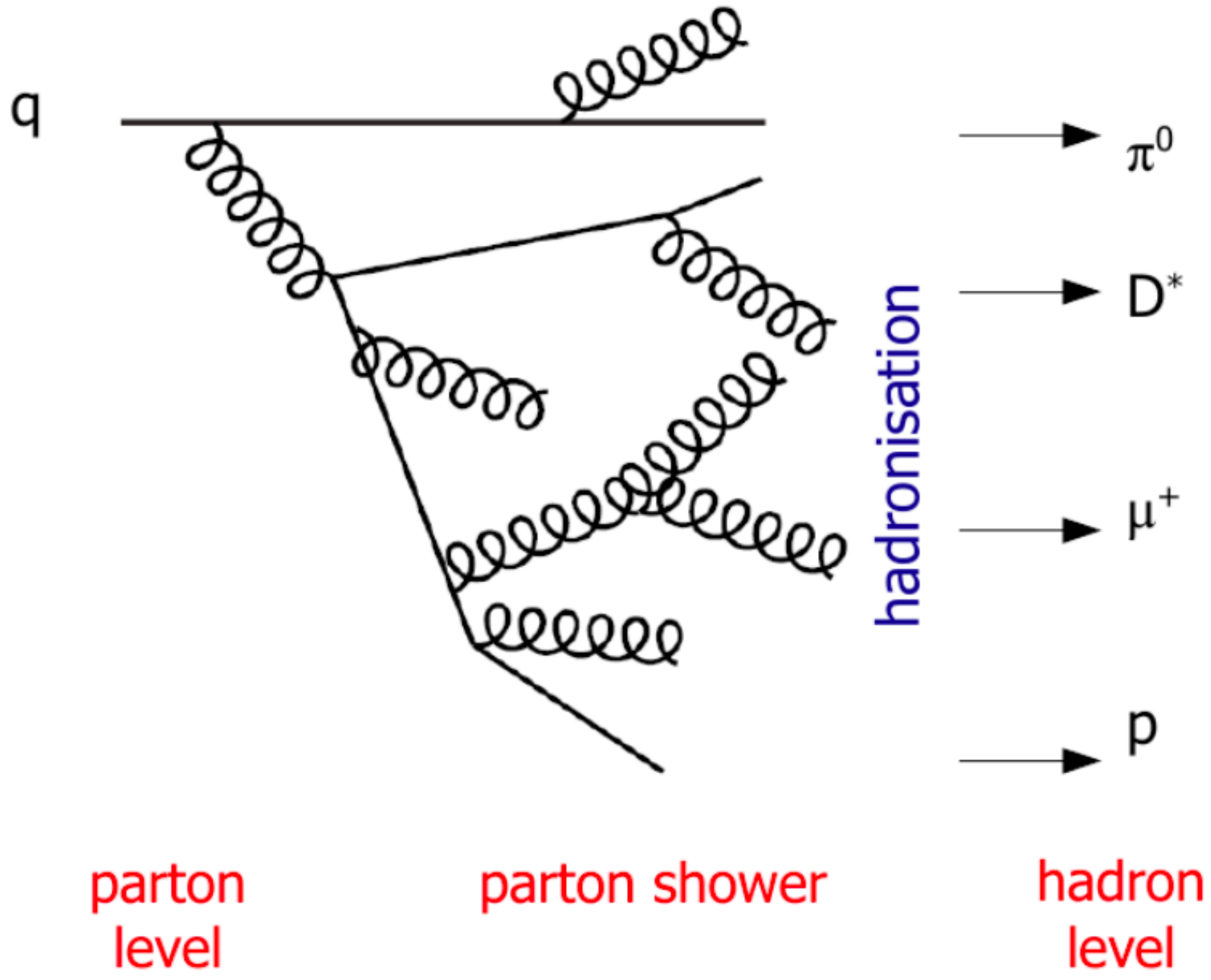
Simulation Digitization Triggering

Offline software
- Event selection

Reconstruction Analysis

Results
Improvement

MC Generation

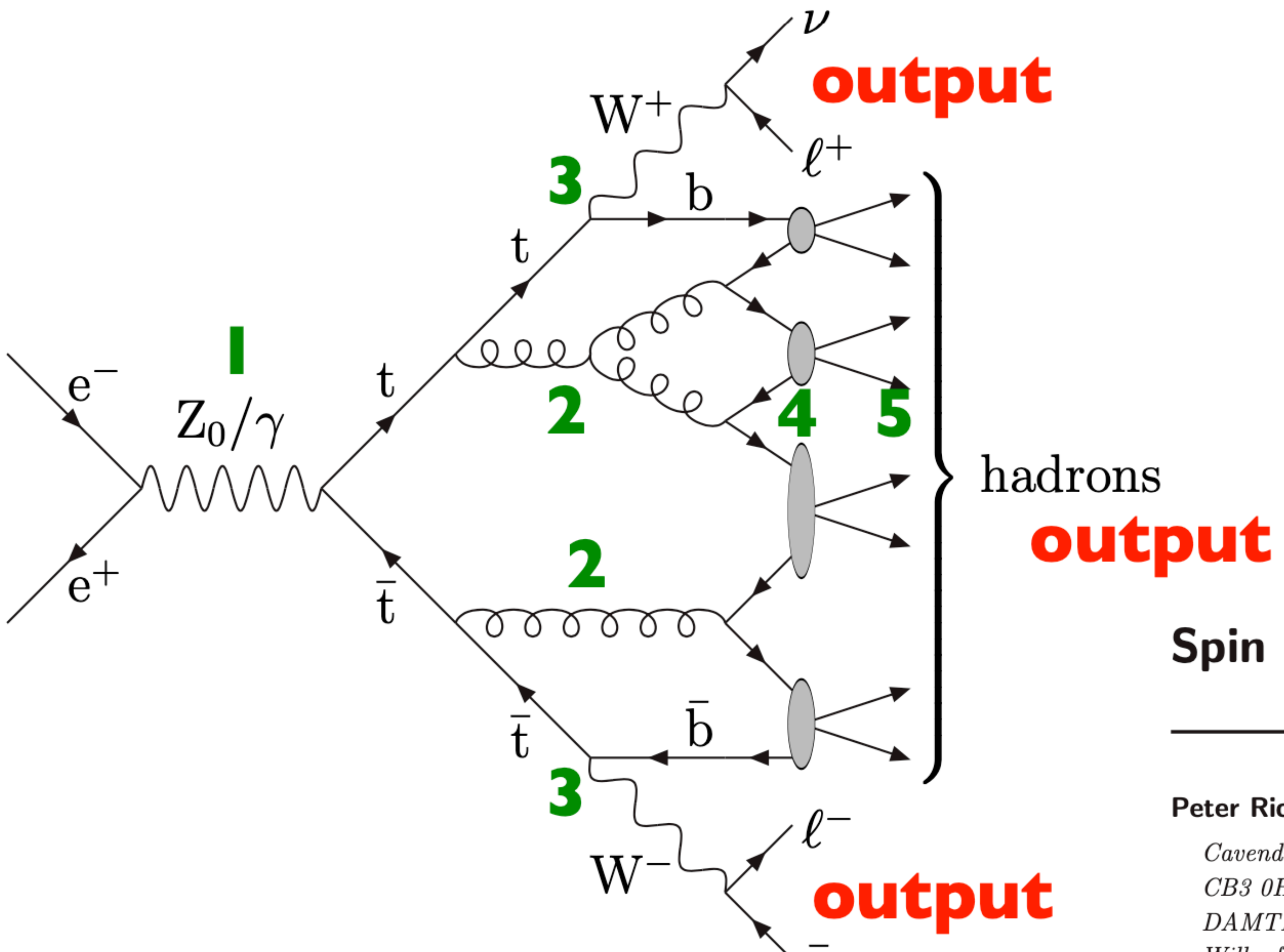


- MC Generator stops with set of "stable" final state particles
- Complete 4-vector info is known about every particle
- All parent-daughter relations are known and stored
- High energy parton state known as parton level
- Stable particle state known as hadron level

- This level of information is often called the truth record
- This is the pure event before it interacts with any apparatus

Monte Carlo event generator process

(1) Hard process: What do you want to study?
(2) Parton-shower phase:
(3) Hard particles decay before hadronizing: e.g. top, SUSY
(4) Hadronization: form observed hadron
(5) Unstable hadrons decay: Experimentally measured BR, phase-space distribution of the decay product



Spin Correlations in Monte Carlo Simulations

Peter Richardson

Cavendish Laboratory, University of Cambridge, Madingley Road, Cambridge, CB3 0HE, UK, and
DAMTP, University of Cambridge, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WA, UK.

Monte Carlo generators (theoretician)

Remember: You need PDF to build MC program
- Learn to describe particle production/decay by matrix element (amplitude) of that process (Explain in QFT)
- Calculate PDF by squaring matrix element, integrate, approximate
- Learn QCD (to deal with jet fragmentation, parton)
- Computaional + Mathematical skills are needed

<u>Real life</u>
There are few theoretical groups who provide us the MC event generator:
- Lund University, Sweden (PYTHIA)
- INP, Krakow, Poland (TAUOLA, PHOTOS)
- Other groups
    - HERWIG
    - ISAJET
    - ...

Monte Carlo generators

Generator (study @ parton level)

↓ ↓

Theoretical Model | **Shower generator**

Input: Model parameters.

Output: Four-vector of momenta of stable/quasi-stable particles produced in interactions

Example MC generator Inteface with CMSSW

Pythia6 ★	Phantom
Herwig6 ★	Hydjet
Pythia8 ★	Pyquen
ThePEG (Herwig++, Ariadne 5) ★	Cosmic Muon Generator
ALPGEN	Beam Halo Muon Generator
MadGraph	ExHuME
MC@NLO	Pomwig
POWHEG	BcGenerator
SHERPA	HARDCOL

A list of MC generator can be found in [http://www.hepforge.org/]

9

Monte Carlo generators (experimentalist)

HEP experimentalists do not write MC generator ourselves (^_^)!!

We use/modify MC generators proposed by theoretical groups.

Your best MC generator is a generator which serve you results that you believe/study.

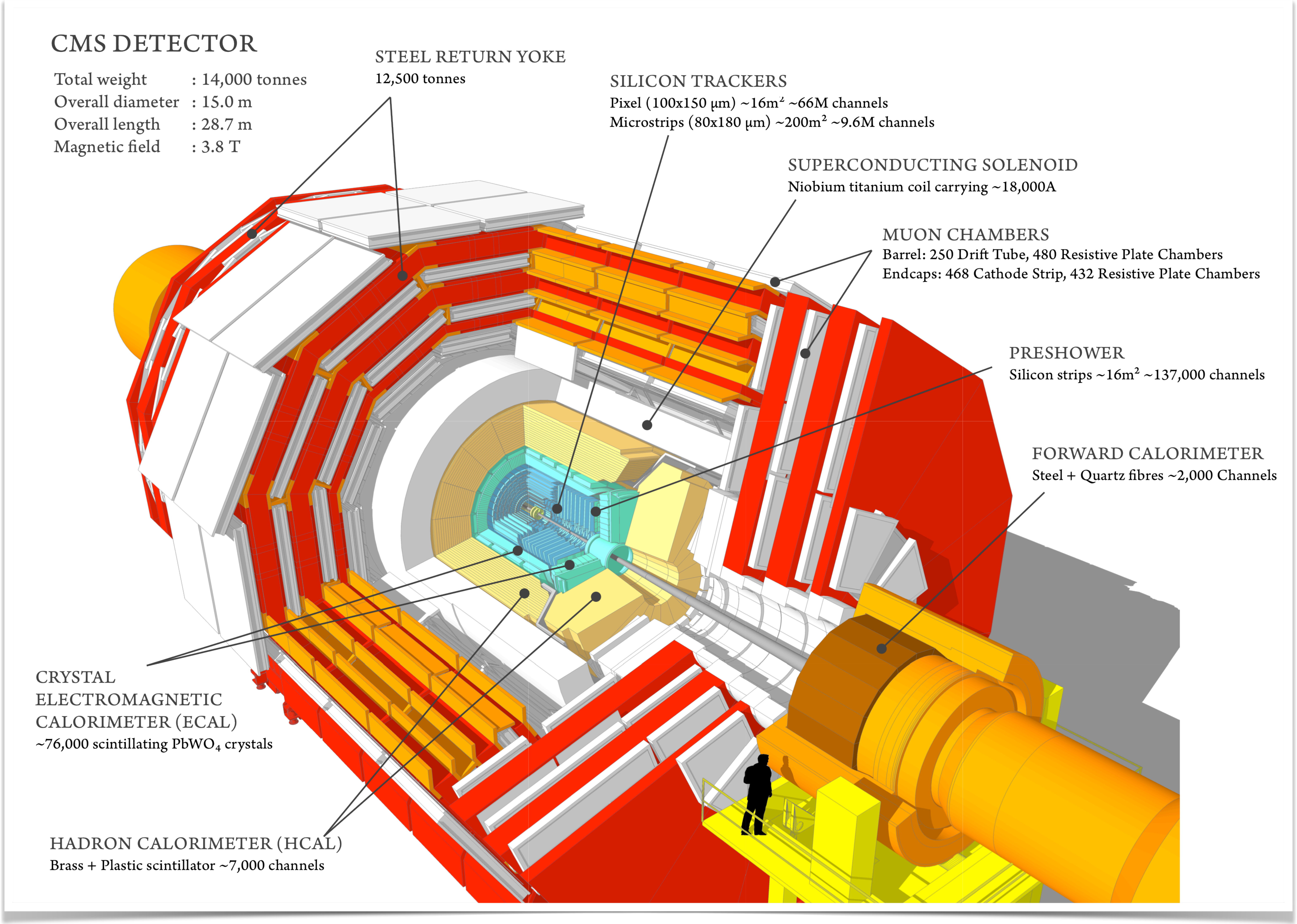
Read manual (VERY IMPORTANT)

You may need more than a MC generator to generate particles for the new physics.

For SUSY, you need mass spectrum + BR + hadronization

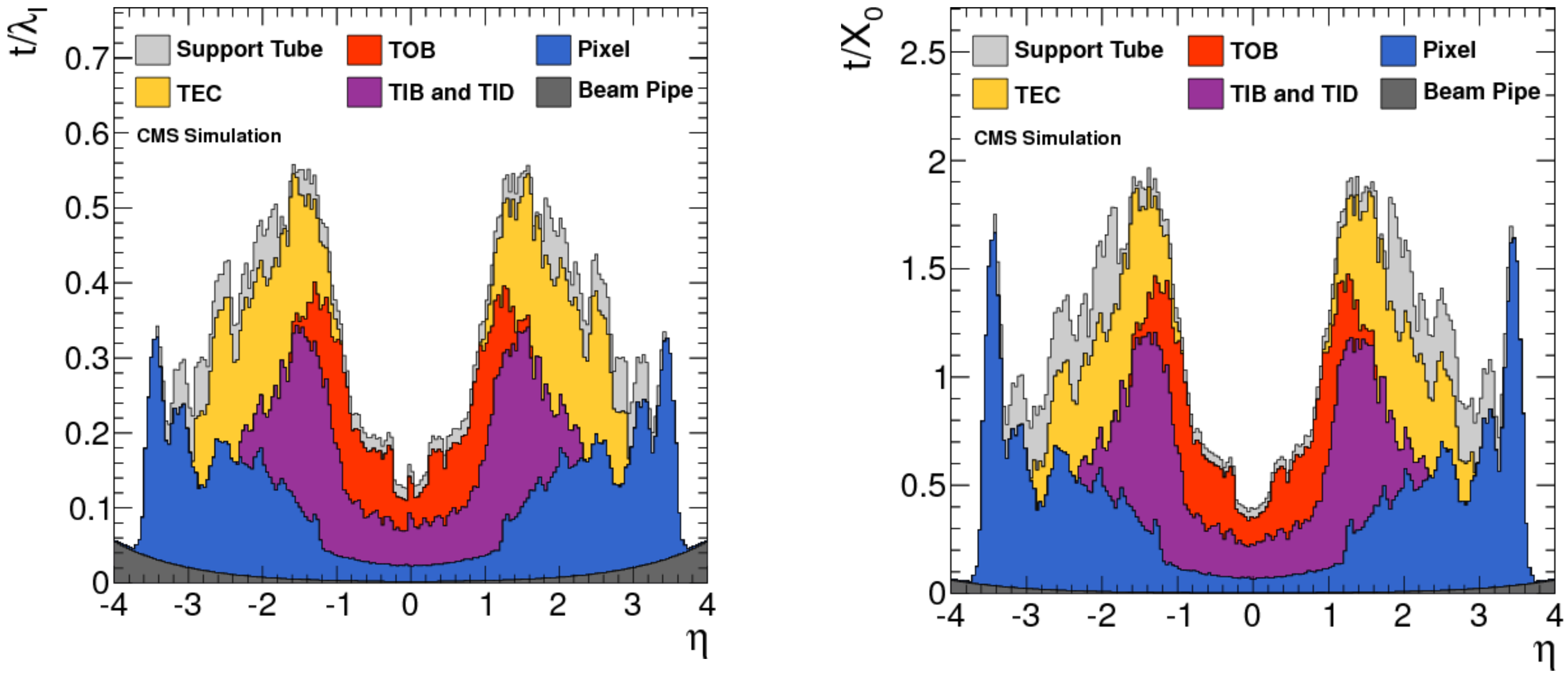
Link with standard txt files, i.e. LHE

Detector Simulation- CMS Experiment



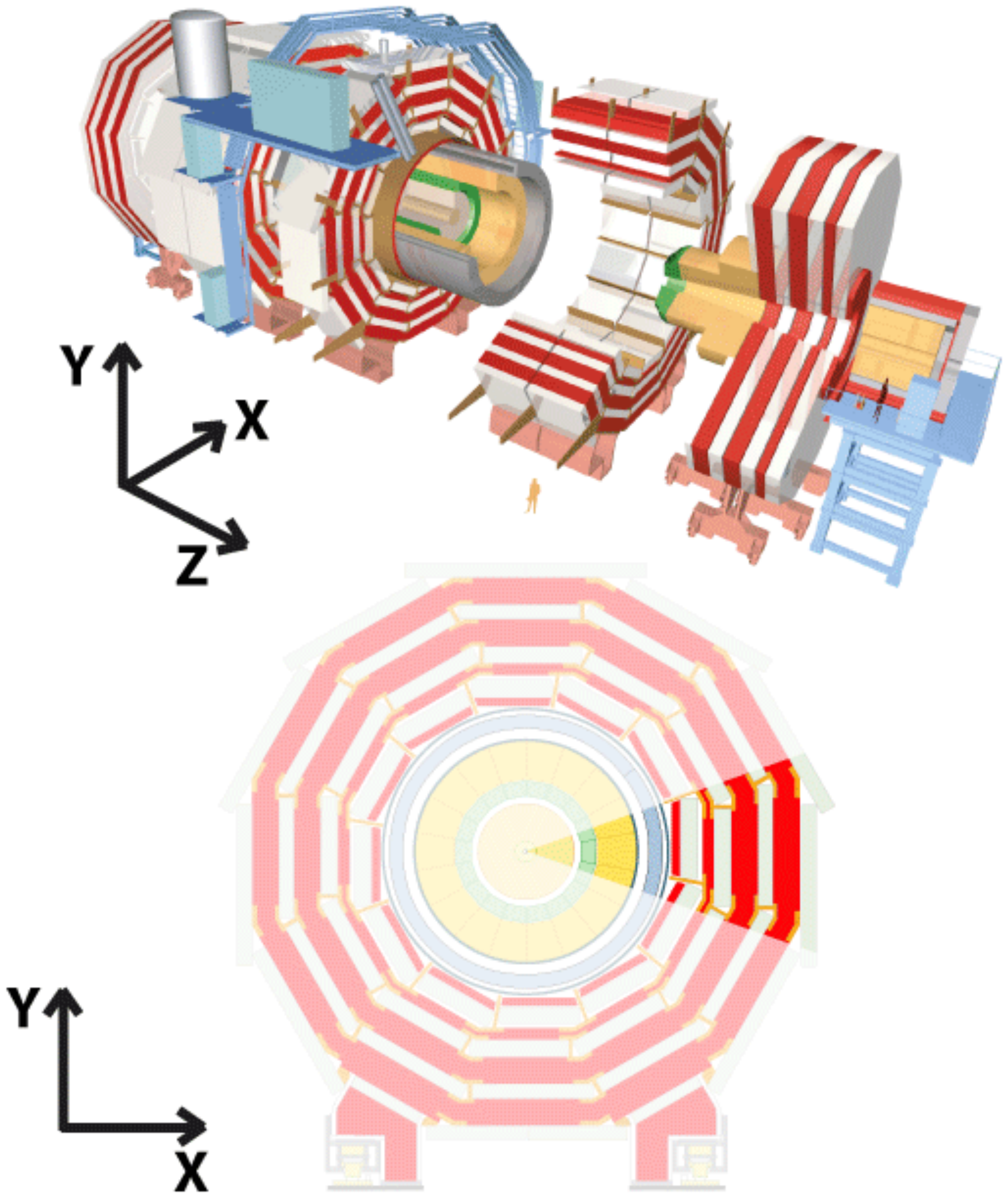
Now need to simulate the detector

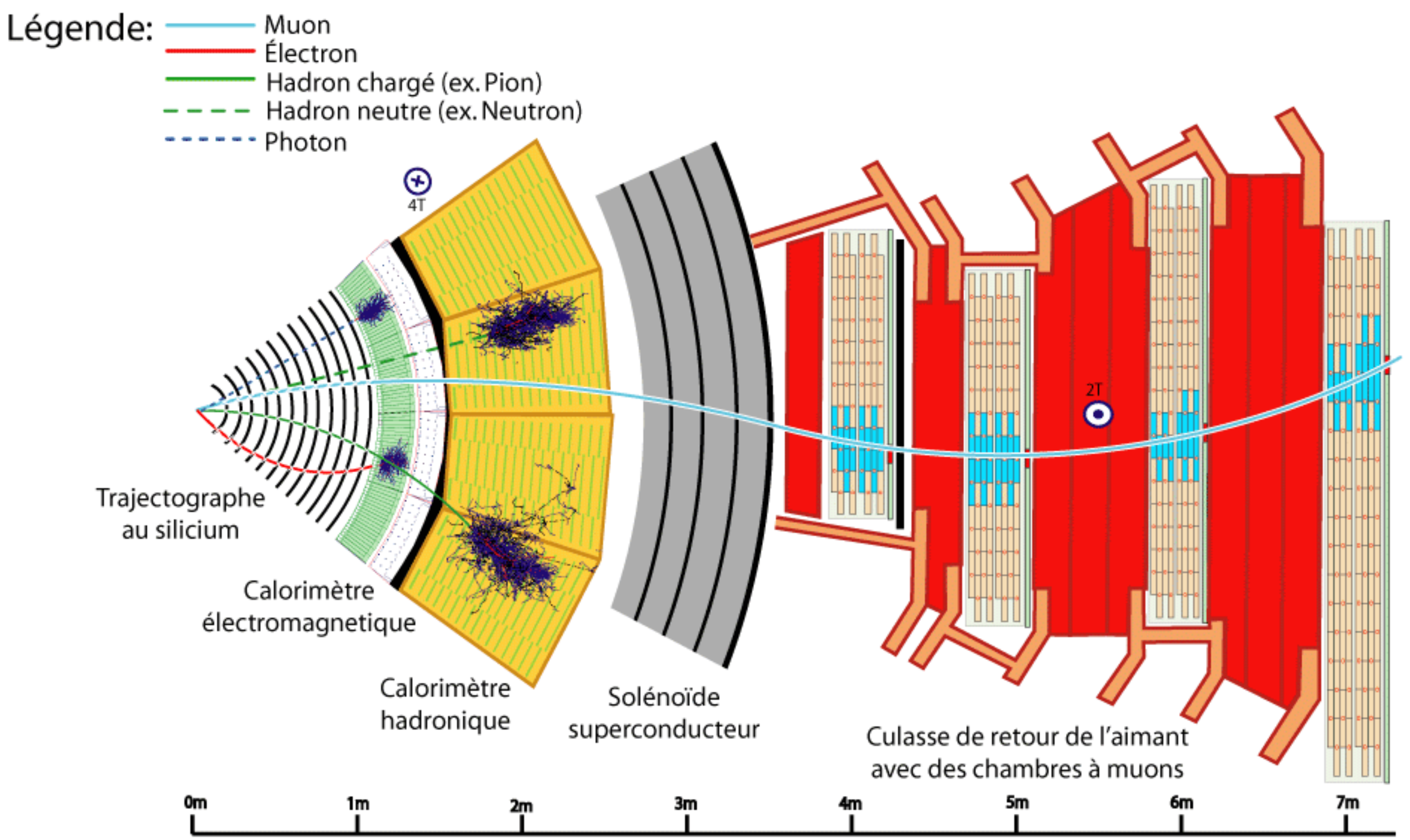
https://twiki.cern.ch/twiki/bin/view/CMSPublic/TrackerMaterialBudgetplots



Material budget in number hadronic interaction length (t/λ0, left) of and radiation length(t/X0, right) as a function of the pseudorapidity η. The contribution of the support tube (light gray), the beam pipe (dark gray), and sub-detectors: TOB (red), Pixel Phase 1 (blue), TEC (yellow) and TID+TIB (magenta) are stacked.

Detector Simulation- CMS Experiment





Monte Carlo for detector simulation

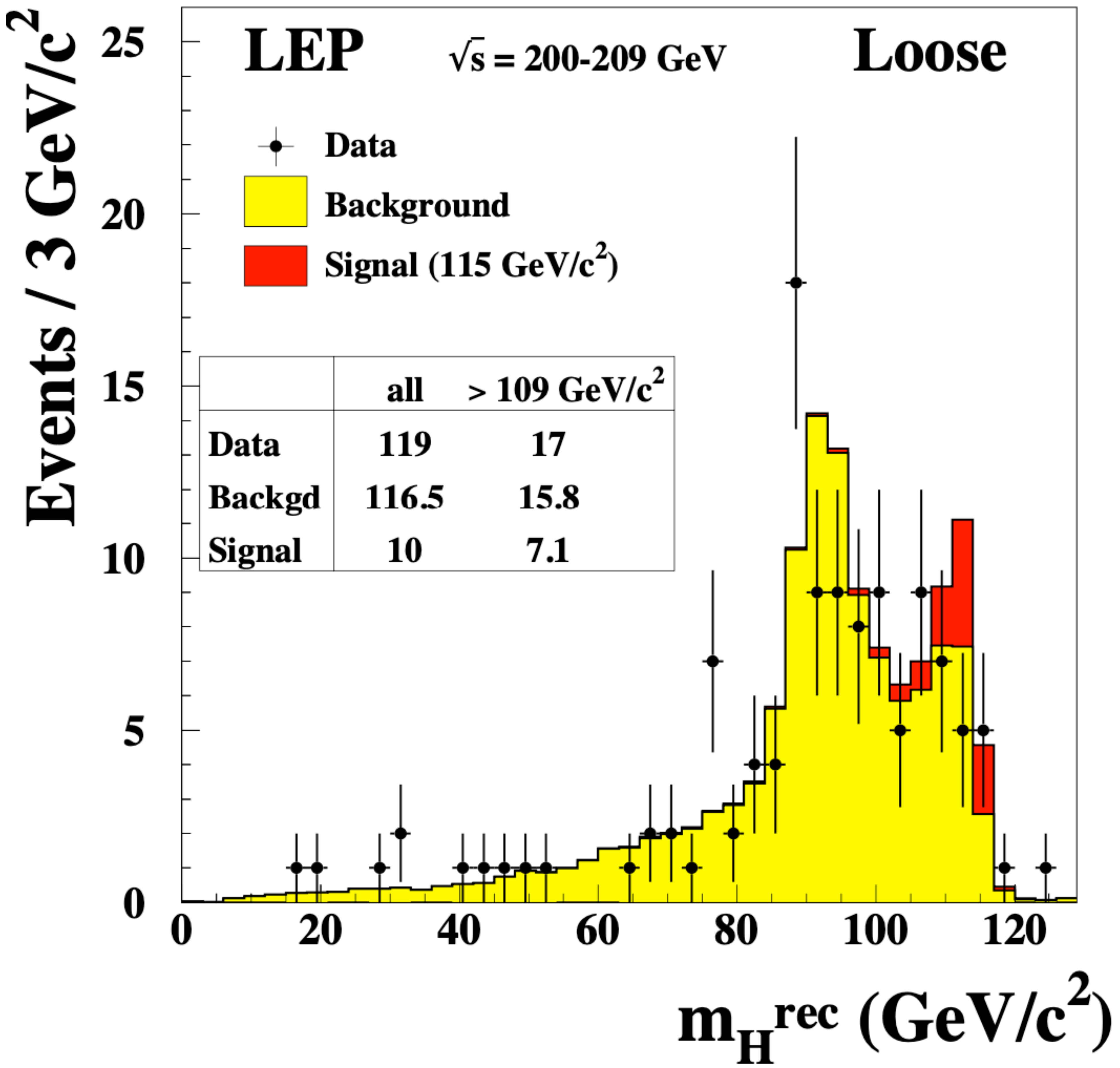
We need to know how our detector will see the productions from collisions.

Detector simulation tracks the particles through detector material (simulating their interactions with material)

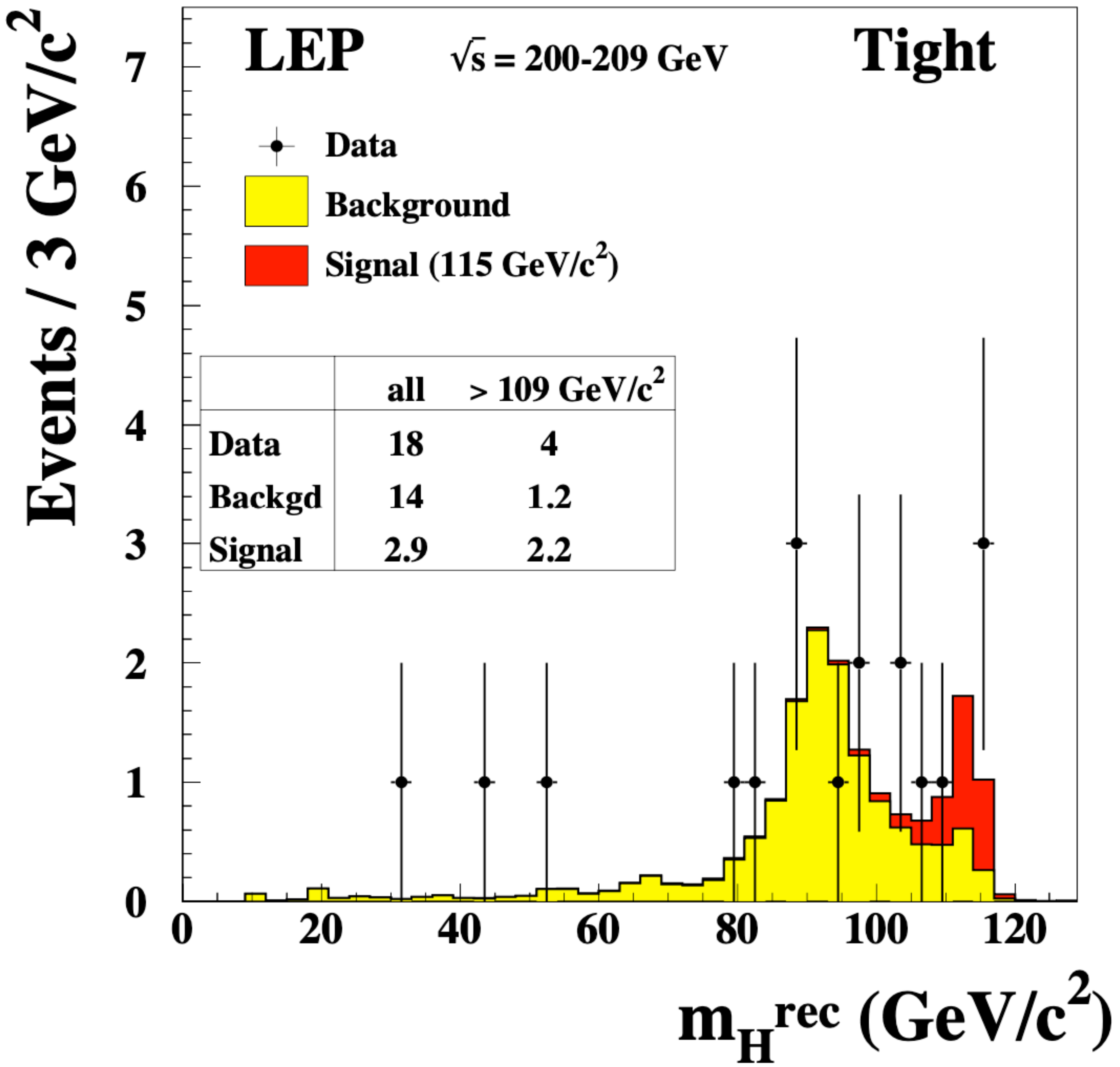
Input: group of (quasi)-stable of particles from MC generator
group of particles from particle-gun
Output: Depend on your analysis

Popular detector simulation
- GEANT4 (http://geant4.web.cern.ch)
- FLUKA (http://www.fluka.org/fluka.php)

Future Lecture will present by Prof. Gabriela Hoff



	all	> 109 GeV/c²
Data	119	17
Backgd	116.5	15.8
Signal	10	7.1



	all	> 109 GeV/c²
Data	18	4
Backgd	14	1.2
Signal	2.9	2.2

Measuring a Cross Section

Number of observed events: counted

Background: Measured from data / calculated from theory

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

$$"L" = \int L\,dt$$

Measured Cross section σ

Efficiency: optimized by experimentalist

Integrated Luminosity: Determined by accelerator, trigger prescale, …

Uncertainty on the Measured Cross section

- You will want to minimize the uncertainty:

$$\frac{\delta\sigma}{\sigma} = \sqrt{\frac{\delta N_{obs}^2 + \delta N_{BG}^2}{(N_{obs} - N_{BG})^2} + \left(\frac{\delta\mathcal{L}}{\mathcal{L}}\right)^2 + \left(\frac{\delta\epsilon}{\epsilon}\right)^2}$$

"Fractional Uncertainties Add in Quadrature"

- Thus you need:
  - N_{obs}-N_{BG} small (I.e. N_{signal} large)
    - Optimize selection for large acceptance and small background
  - Uncertainties on efficiency and background small
    - Hard work you have to do
  - Uncertainty on luminosity small
    - Usually not directly in your power

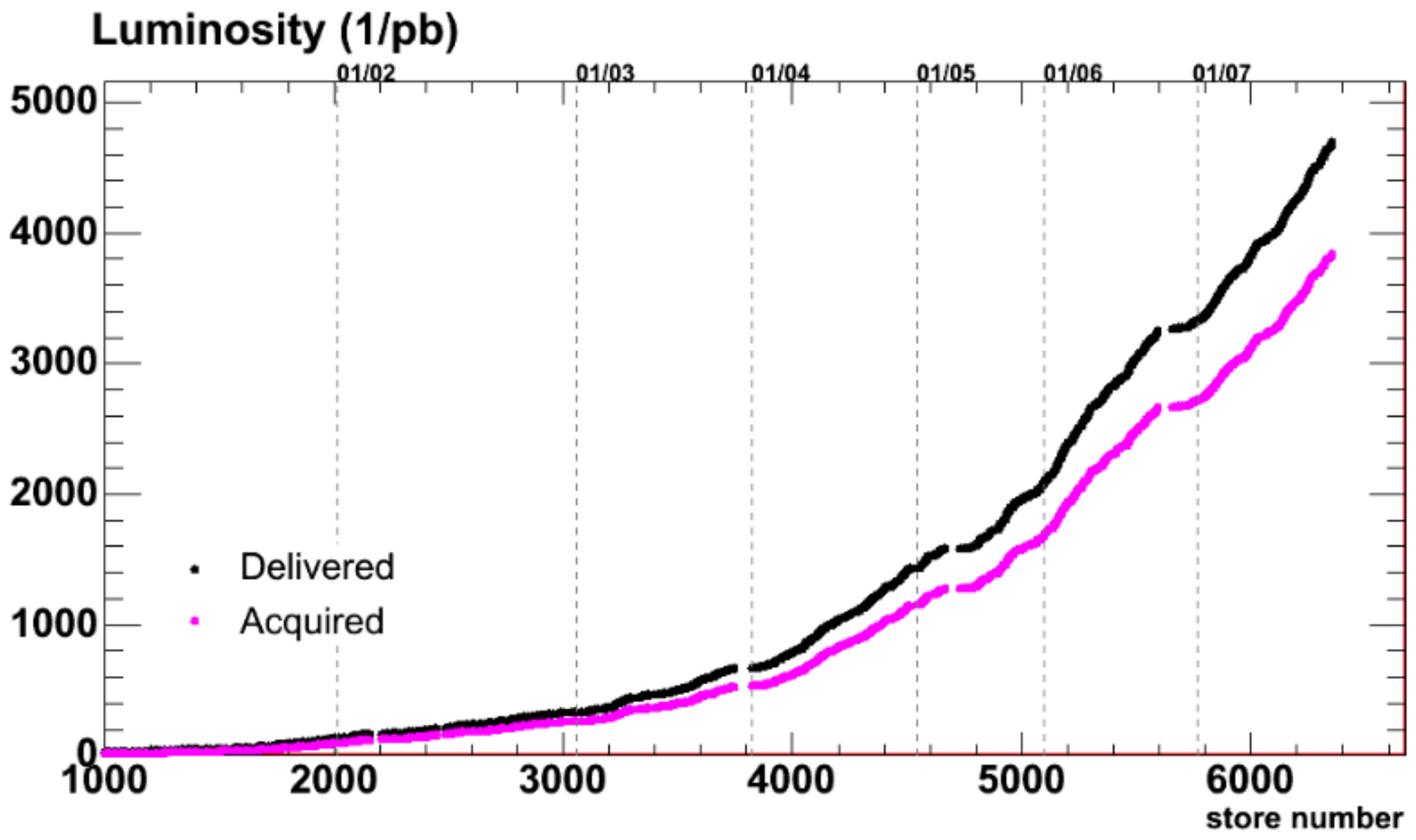
Luminosity Measurements and Uncertainties

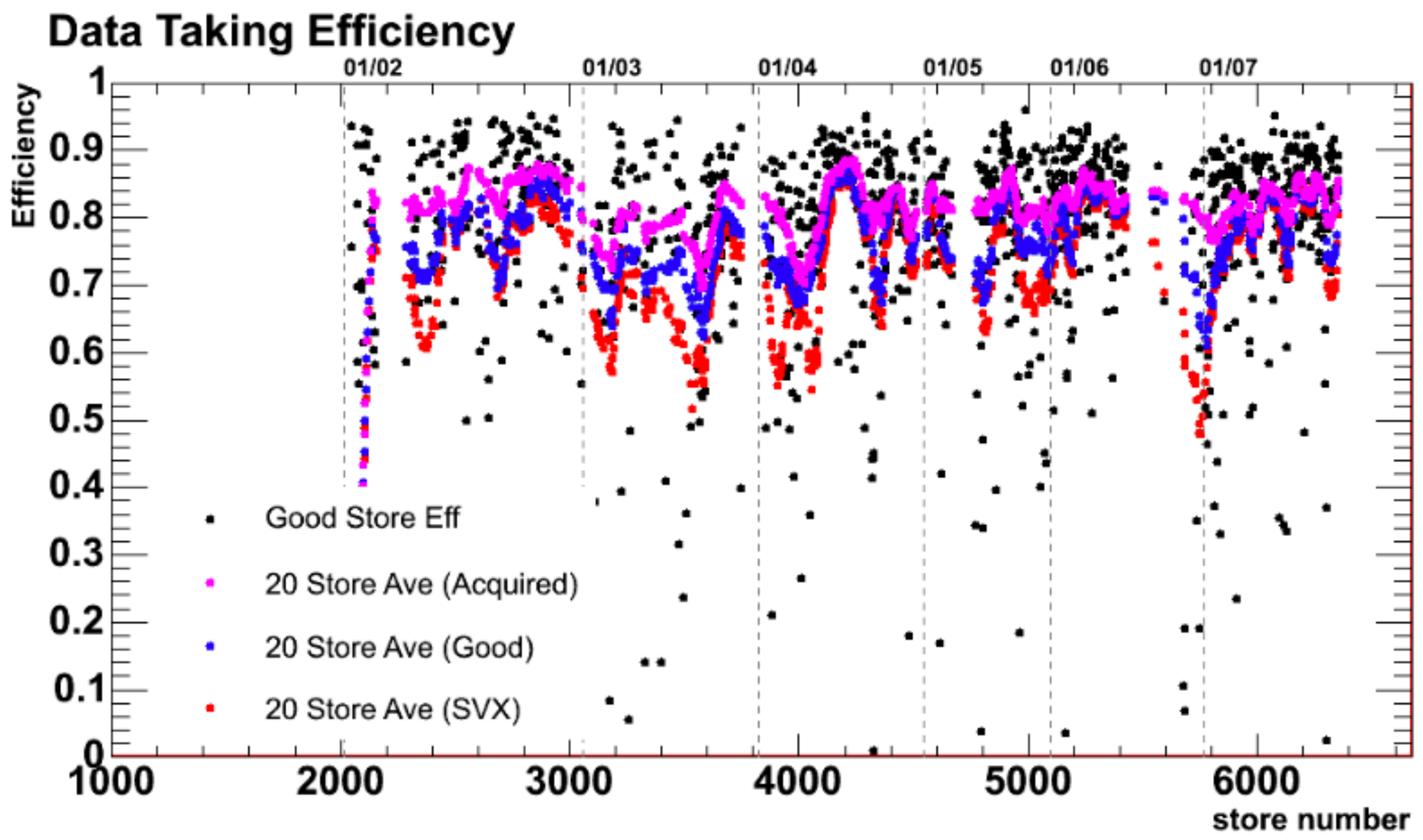
- Many different ways to measure it:
  - Beam optics
    - LHC startup: precision ~20-30%
    - Ultimately: precision ~5%
  - Relate number of interactions to total cross section
    - absolute precision ~4-6%, relative precision much better
  - Elastic scattering:
    - LHC: abslute precision ~3%
  - Physics processes:
    - W/Z: precision ~2-3% ?
- Need to measure it as function of time:
  - $L(t) = L_0 \, e^{-t/\tau}$ with $\tau \approx 14h$ at LHC and L_0 = initial luminosity

Luminosity Estimates are a "Shared Resource" – One example of a calibration shared by many groups

Your Luminosity

- Your data analysis luminosity is not equals to LHC/Tevatron luminosity!
- Because:
  - The detector is not 100% efficiency at taking data
  - Not all parts of the detector are always operational/on
  - Your trigger may have been off / prescaled at times
  - Some of your jobs crashed and you could not run over all events
- All needs to be taken into account
  - Severe bookkeeping headache





A Problem with that Uncertainty Formula

$$\frac{\delta\sigma}{\sigma} = \sqrt{\frac{\delta N_{obs}^2 + \delta N_{BG}^2}{(N_{obs} - N_{BG})^2} + \left(\frac{\delta\mathcal{L}}{\mathcal{L}}\right)^2 + \left(\frac{\delta\epsilon}{\epsilon}\right)^2}$$

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

Both the integrated luminosity in the denominator and the N_{BG} in the numerator depend on the luminosity estimate, because some backgrounds are estimated using

Theory cross section x Integrated Luminosity x branching ratios x cut acceptance.

Other backgrounds may be estimated using data-based techniques (more on this later)

→ Missing a correlation!

Handling Correlations the Easy Way

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

1) Identify ***independent*** sources of systematic uncertainty. Usually they have names and are listed in tables of systematic uncertainties. These are called ***nuisance parameters***

Luminosity estimate depends on:

- Inelastic pp (or ppbar) cross section
- Luminosity monitor acceptance

or, if using a data-based luminosity extraction

- Inclusive W or Z cross section theory prediction, and
- Lepton identification systematic uncertainty

Note – you cannot measure the inclusive Z cross section using the second method.

continued:

Nuisance parameter

From Wikipedia, the free encyclopedia

In statistics, a **nuisance parameter** is any parameter which is not of immediate interest but which must be accounted for in the analysis of those parameters which are of interest. The classic example of a nuisance parameter is the variance, σ^2, of a normal distribution, when the mean, μ, is of primary interest.[citation needed]

https://en.wikipedia.org/wiki/Nuisance_parameter

Handling Correlations the Easy Way

$$\sigma^{meas} = \frac{N_{obs} - N_{BG}}{L \cdot \varepsilon}$$

2) Evaluate the impact of each nuisance parameter on your answer, holding the others fixed:

$$\frac{d\sigma^{meas}}{d\nu_i}$$

were ν_i is the i^{th} nuisance parameter.

Tip – you can often collect nuisance parameters together if they all affect the result in the same way. "Integrated Luminosity" is a perfectly good nuisance parameter most of the time, as predictions depend on it.

But sometimes you can't. Suppose pp→Z is one of the background sources, and you are using the measured Z rate to constrain the luminosity in the data.

These can even be non-overlapping data. Z→ee constrains the lumi, while Z→hadrons is a background, for example.

Then the inclusive Z cross section assumed becomes the nuisance parameter (and its impact partially cancels!)

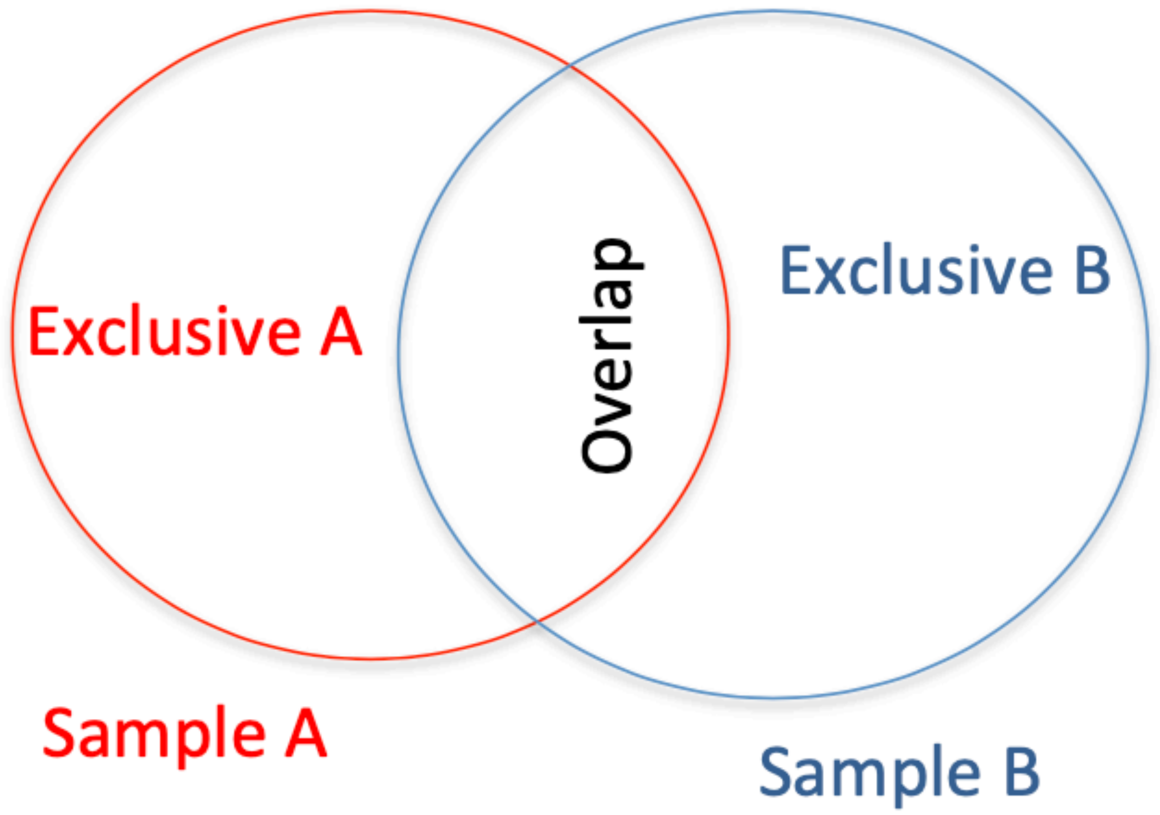
Handling Correlations the Easy Way

Tip: Sometimes uncertainties are correlated in a nontrivial way:

Example: Two sources of background are estimated using data control samples, but these control samples share some but not all of their data events.

Suggestion: Always seek an uncorrelated parameterization. We know that those control samples are partially correlated due to the overlaps. We estimate the correlation by knowing the fractions in the exclusive and inclusive samples.

You can always break down partially correlated uncertainties into pieces – a fully correlated piece and uncorrelated pieces.



In this case, Exclusive A, Exclusive B, and Overlap may be the nuisance parameters for evaluating the stat. uncertainty from these control samples

Handling Correlations the Easy Way

Putting it all together:
Once you have an uncorrelated basis, just add the uncertainties in quadrature.

$$\delta\sigma^{meas} = \sqrt{\sum_i \left(\frac{d\sigma^{meas}}{d\nu_i} \delta\nu_i \right)^2}$$

A nuisance parameter is any value you assumed in order to do your analysis which you do not know the exact value of (usually all of them).

Much of the work is devoted to identifying a proper set of nuisance parameters, and constraining their possible values, preferably with data.

We must frequently ask theorists for help!

Weighting MC Factor procedures

- Whenever we want to compare more than one MC sample together, they need to be properly weighted.
  - All samples should be normalized to correspond to the same amount of integrated luminosity.

$$\text{weight}\left(\int \mathcal{L}dt\right) = \frac{\int \mathcal{L}dt \cdot \text{cross section}}{\text{number of events considered}}$$

- For example, I want to compare distributions of W + jets and top. Assuming an integrated luminosity of 1 fb^{-1}, I get:

Sample	Cross section	Number of MC Events	Weight
W + jets	24 nb	1,204,434	19.926
Top	165 pb	221,331	0.745

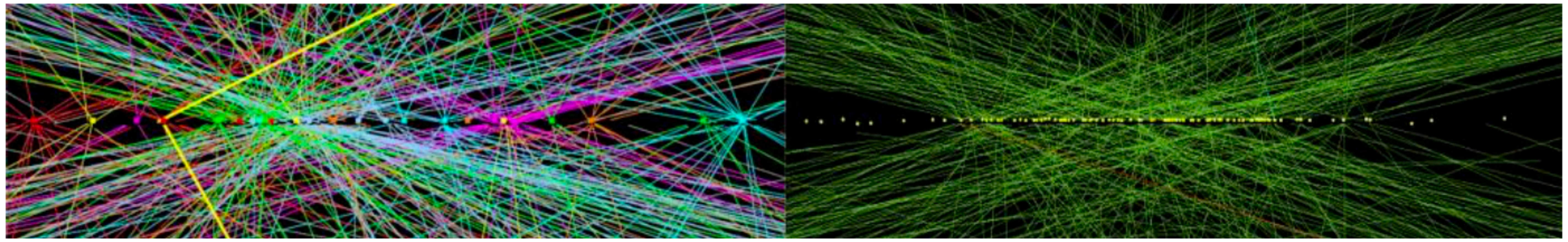
(Remember: 1 nb = $1 \cdot 10^6$ fb and 1 pb = $1 \cdot 10^3$ fb)

- After running analysis jobs, make sure we scale/multiply all histograms/numbers by the proper weight.
- This information is currently **not** stored in CMS MC.

Pile-up!

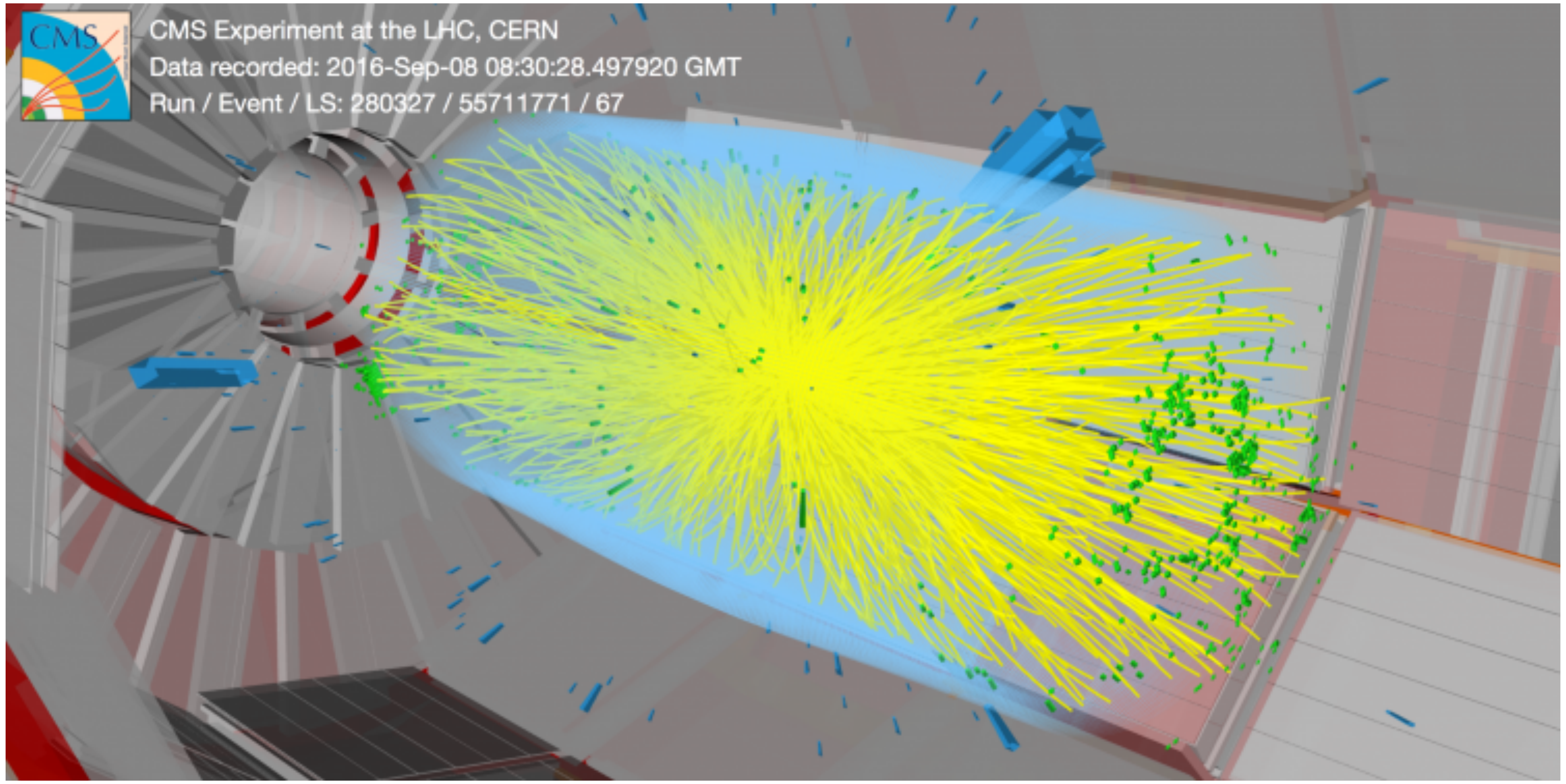
- **Multiple events in a photo**
  - **Pileup in Firework**
    - ✓ **How many explosions can you identify?**

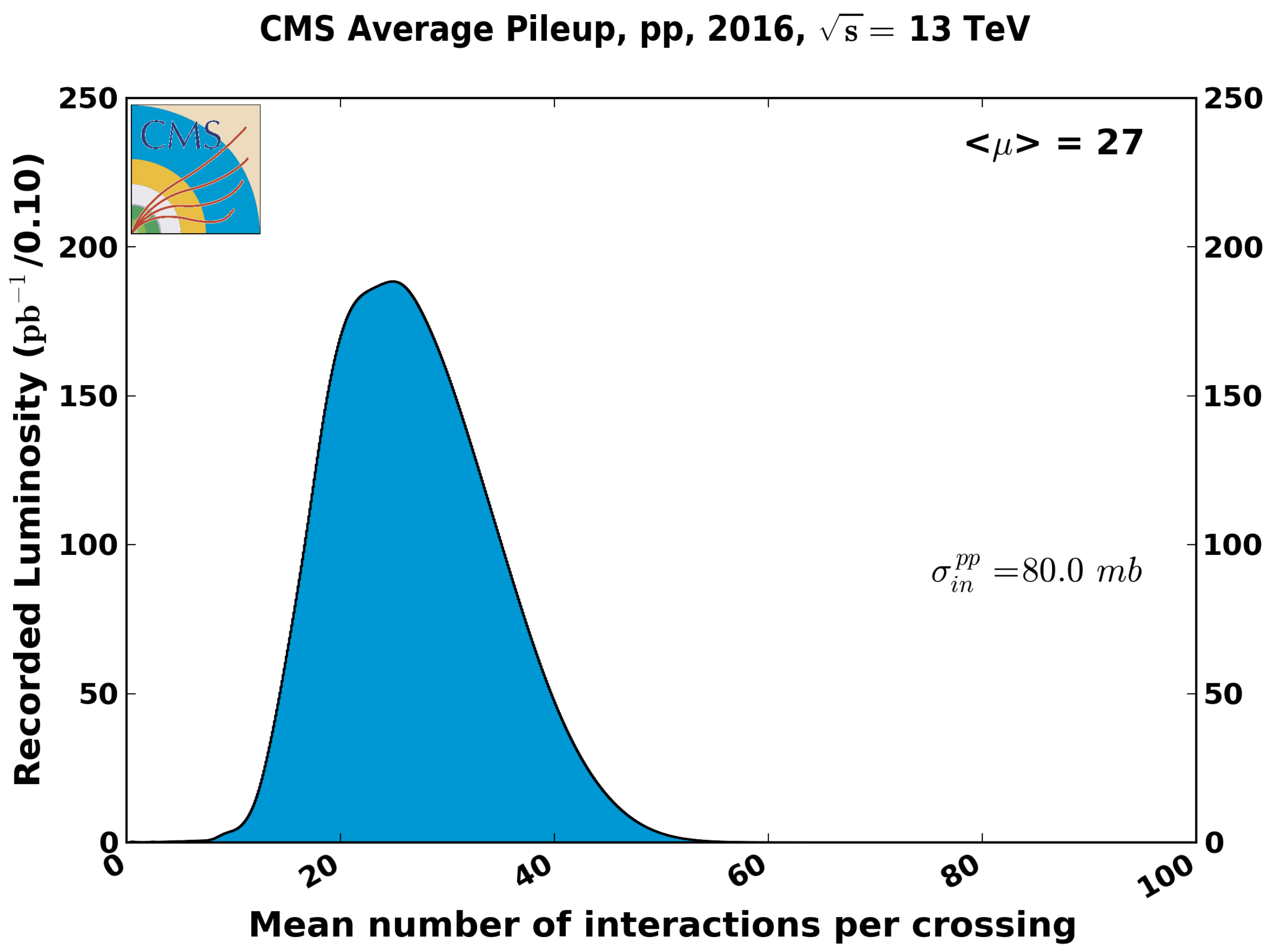
  - **Pileup in LHC**
    - ✓ **How many vertices can you reconstruct?**

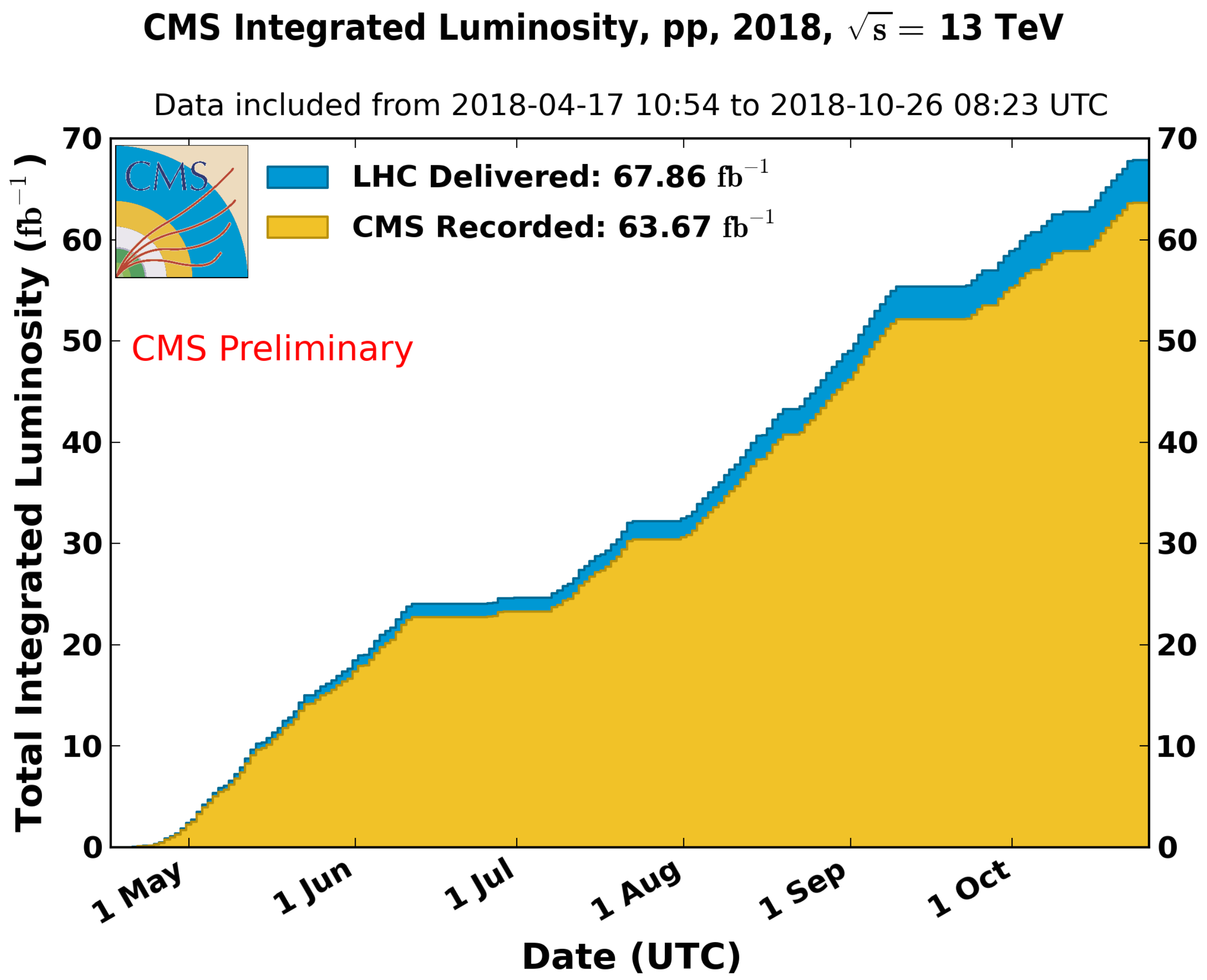


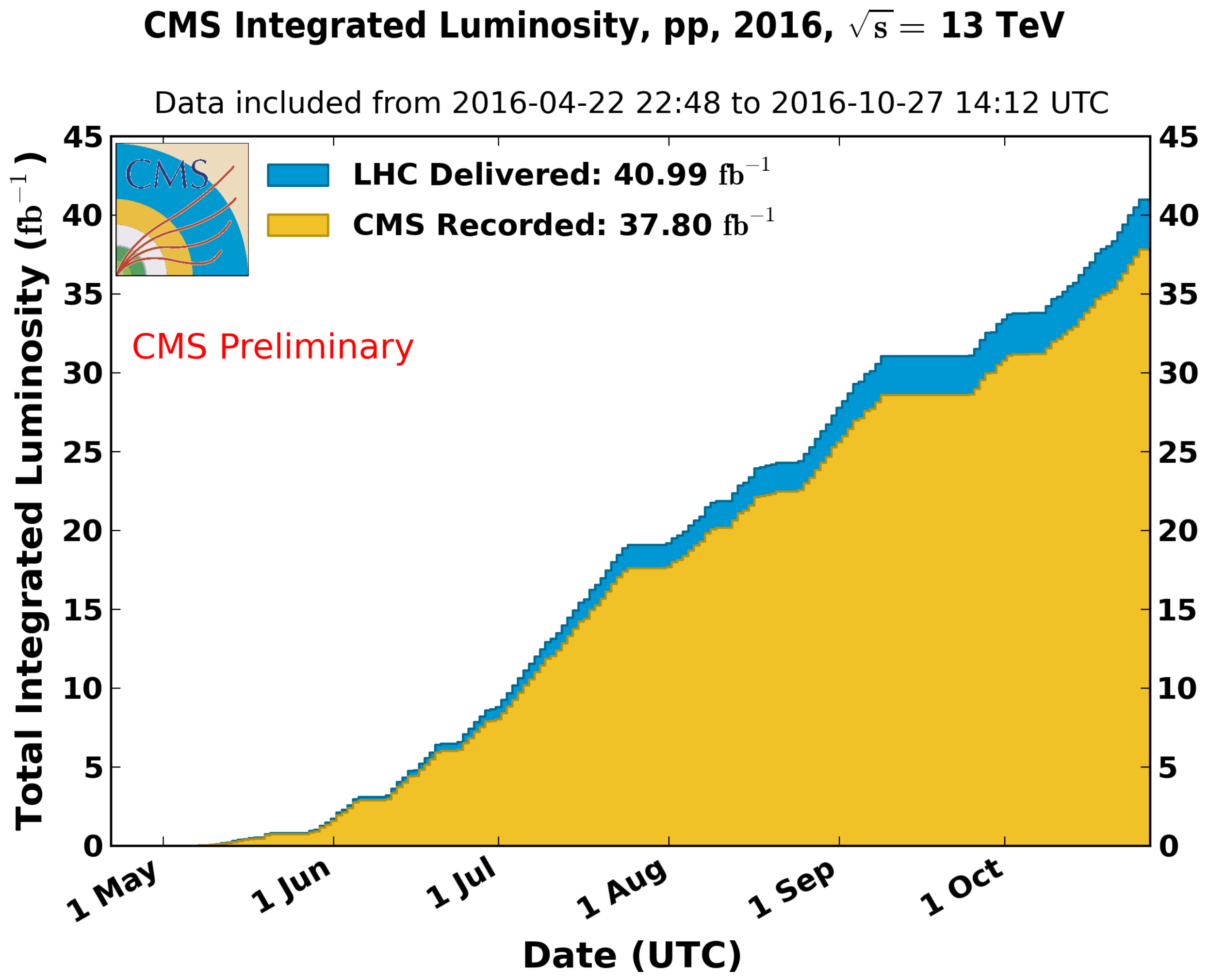
Pile-up estimation

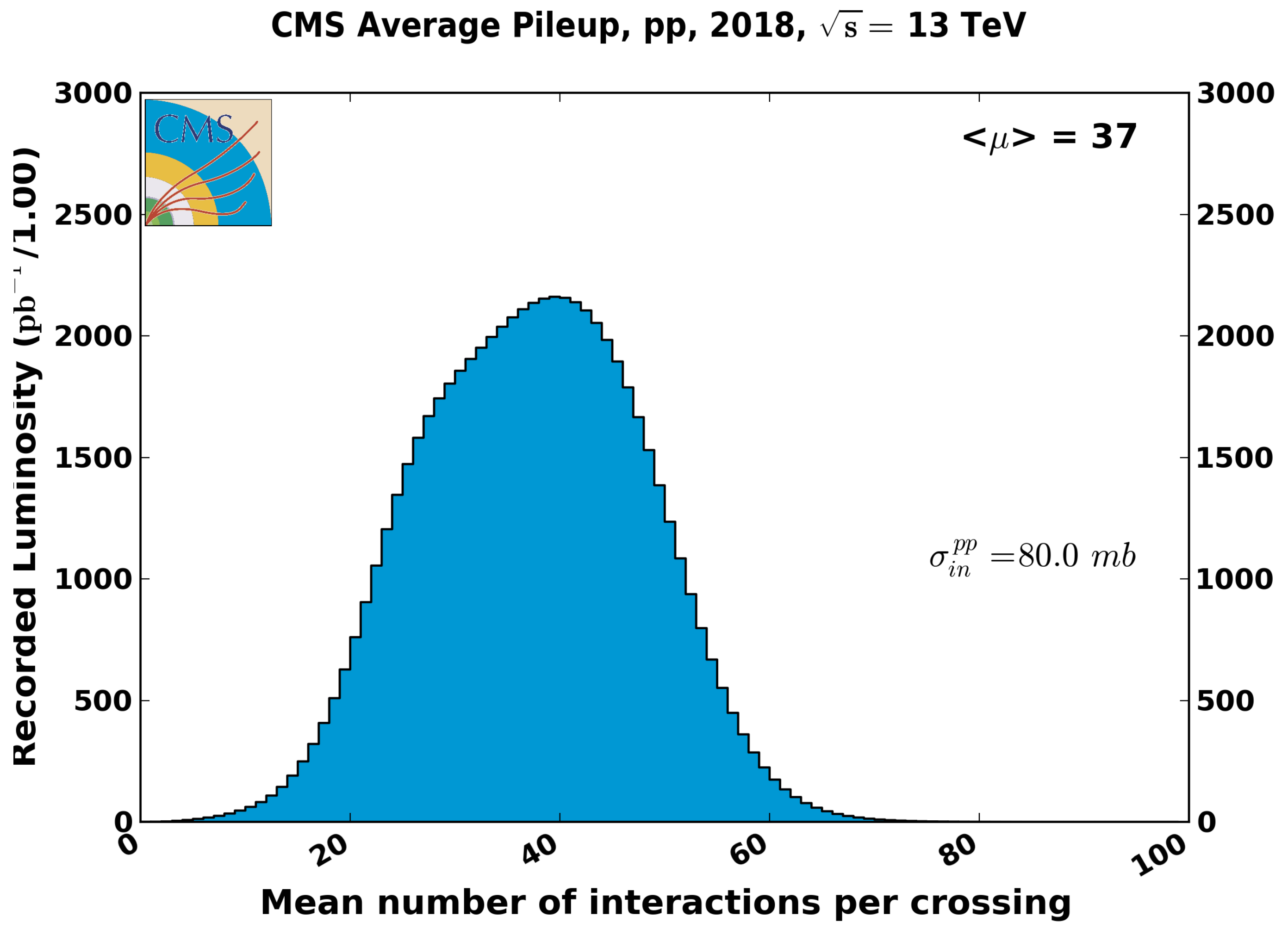
- **Pileup: average number of collisions produced per bunch crossing**
  - **Bunch crossing rate, R_{BC}**
    - ✓ $R_{BC} = \frac{c}{27km} \times n_b = 11253Hz \times n_b$
  - **For the nominal LHC**
    - ✓ $R = \mathcal{L} \times \sigma_{inel} = 10^{34}cm^{-2}s^{-1} \times 70mb \sim 7 \times 10^{8}Hz$
      - – $Pileup = R/R_{BC} = 7 \times 10^{8}/(11253 \times 2808) \sim 22$
  - **For example, LHC 2012 run**
    - ✓ $R = \mathcal{L} \times \sigma_{inel} = 7.7 \times 10^{33}cm^{-2}s^{-1} \times 70mb \sim 5.4 \times 10^{7}$
      - – $Pileup = 5.4 \times 10^{7}/(11253 \times 1380) \sim 35$



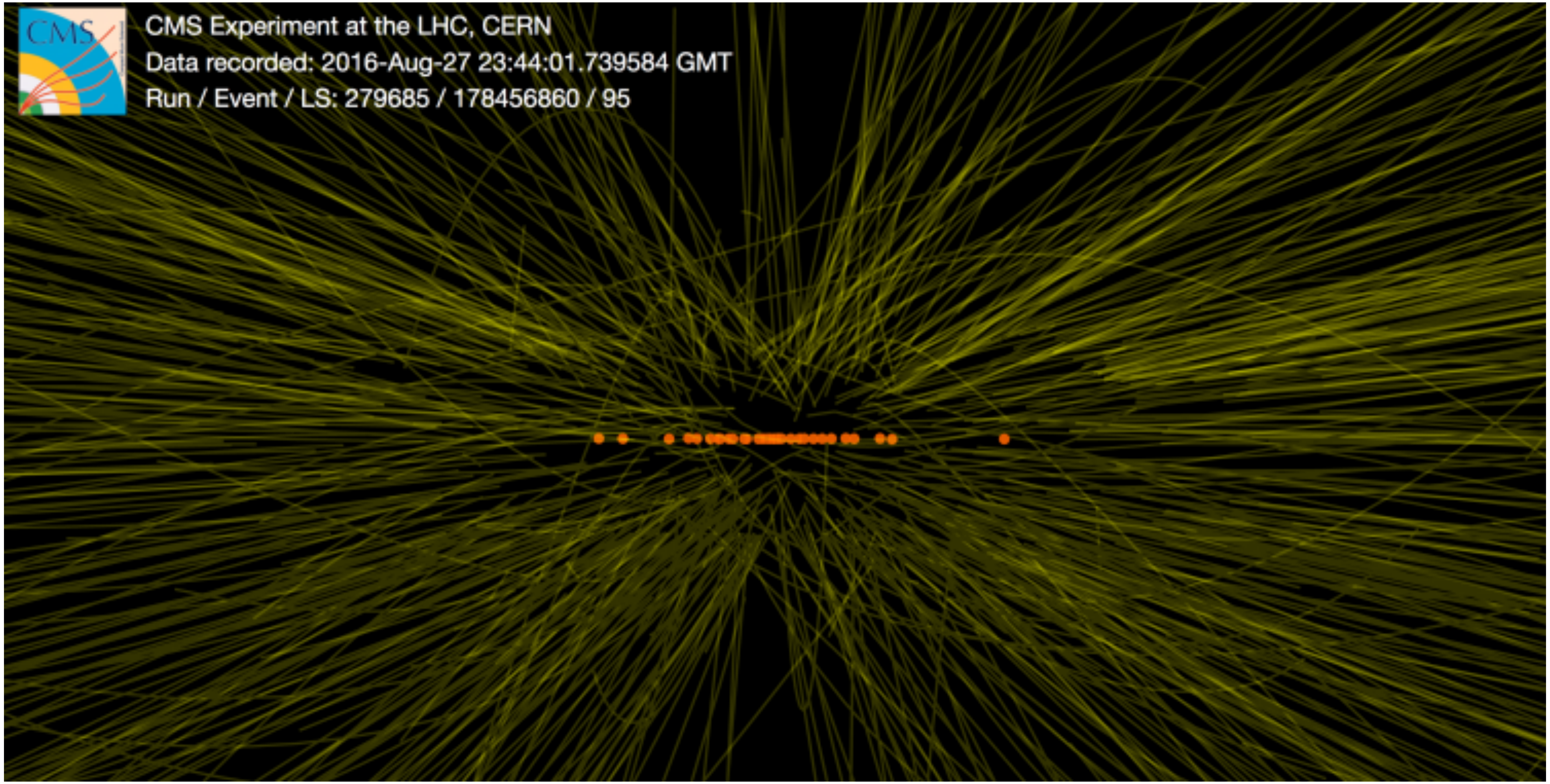








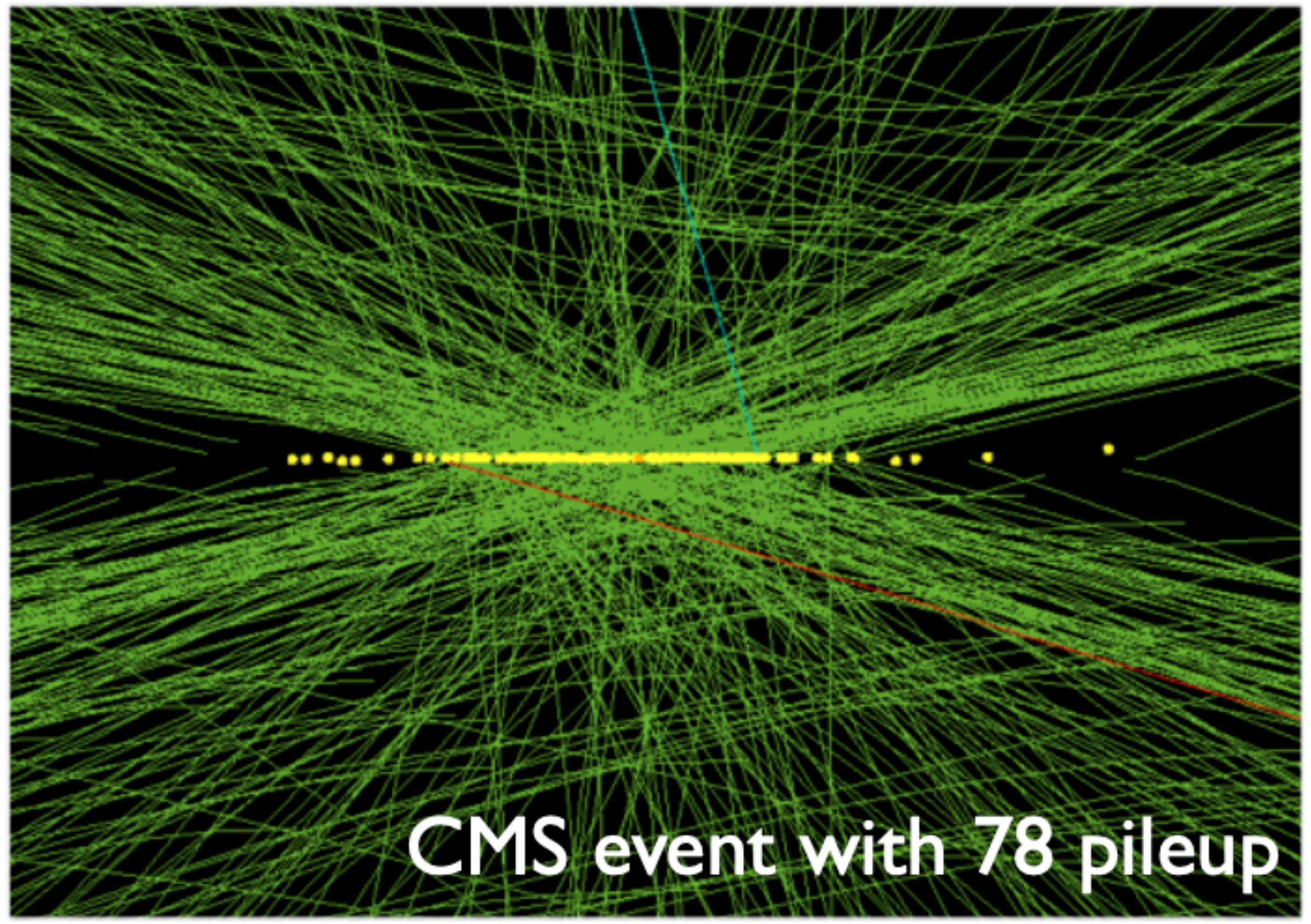
https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults#Overview



https://cms.cern/news/how-cms-weeds-out-particles-pile

Pileup basics and key questions

- Luminosity of $5\times10^{34}cm^{-2}s^{-1}$ corresponds to an *average* pileup of 140 events
  - Upper estimate of average number of pileup events for this lumi – partly accounts for bunch-to-bunch variation
  - Average of a Poisson distribution with a sigma of about 12 events
- Key questions:
  - Can the detectors work with even higher (average) pileup to allow 3000 /fb to be delivered more quickly?
  - Can a longer beam spot help pileup mitigation?
- Need to take into account in-time pileup (same bunch crossing) and out-of-time pileup (previous crossings) – particularly for ATLAS colorimeter and for muon spectrometers

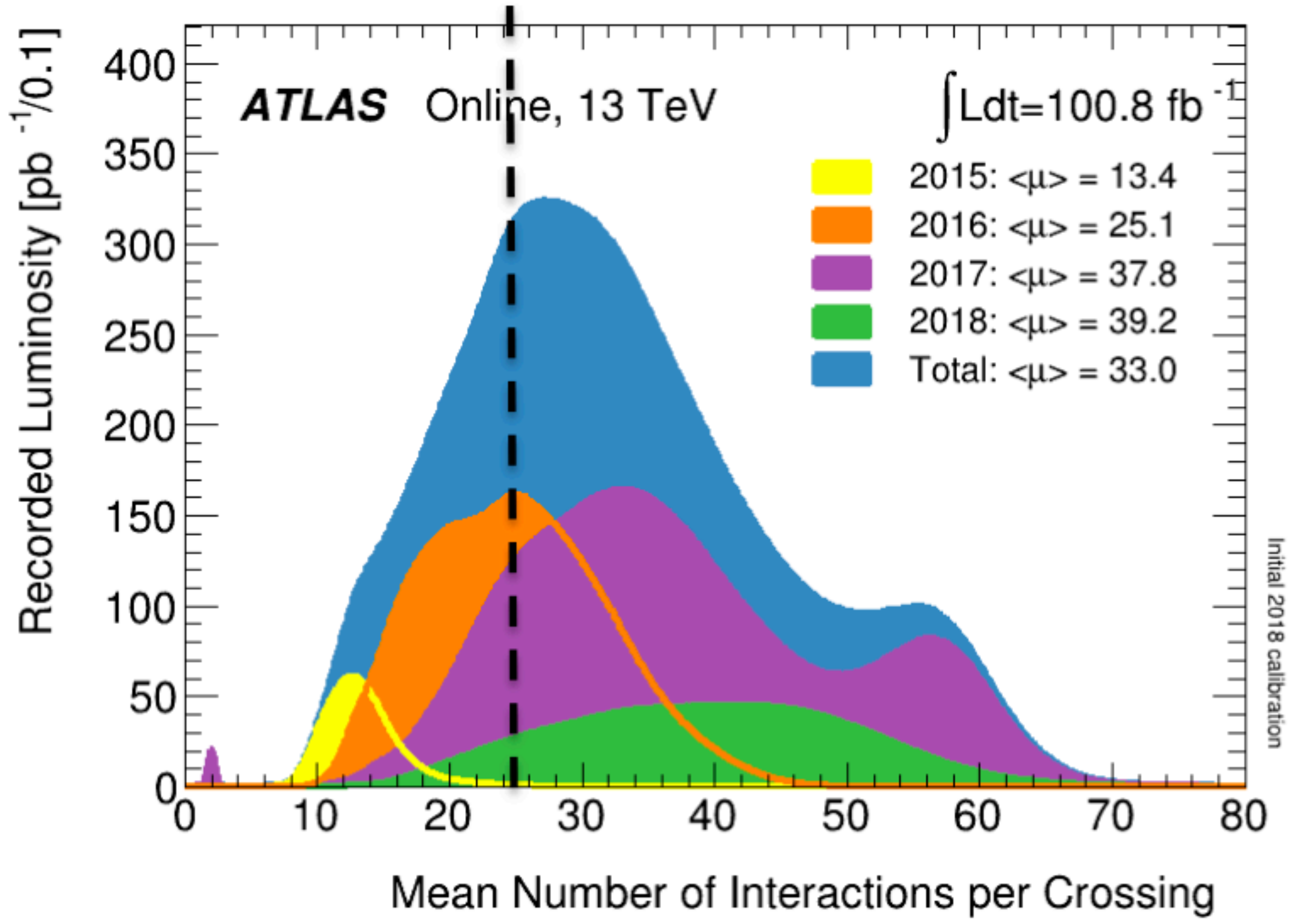


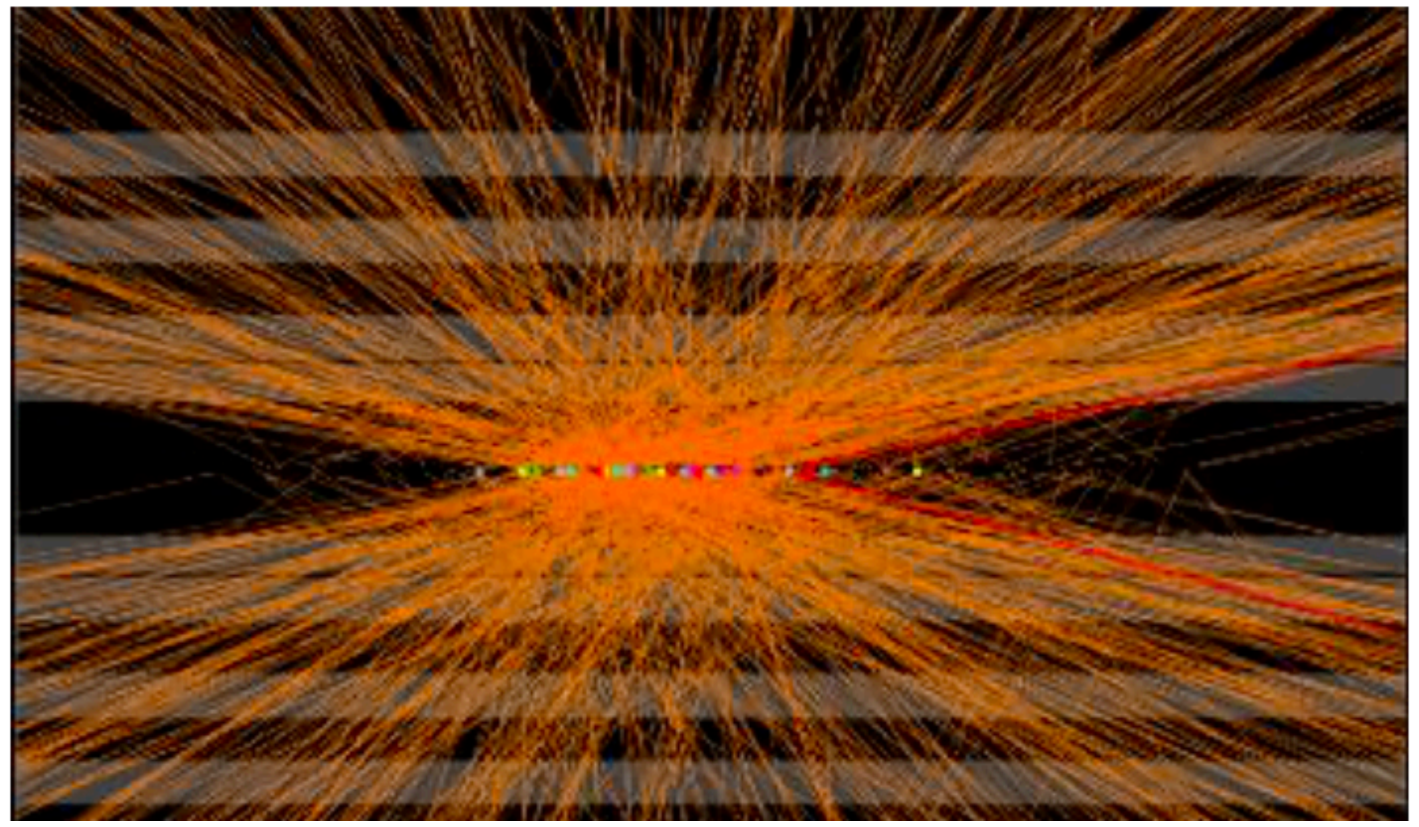
CMS event with 78 pileup

Run-2 pileup in ATLAS

The average pileup in the Run-2 dataset is ~33. But with a peak above 60, mostly coming from 2017 running with fewer, high intensity bunches. In 2018 running with more bunches allows higher luminosity, but with less pileup (much better for physics!)

LHC design peak pileup

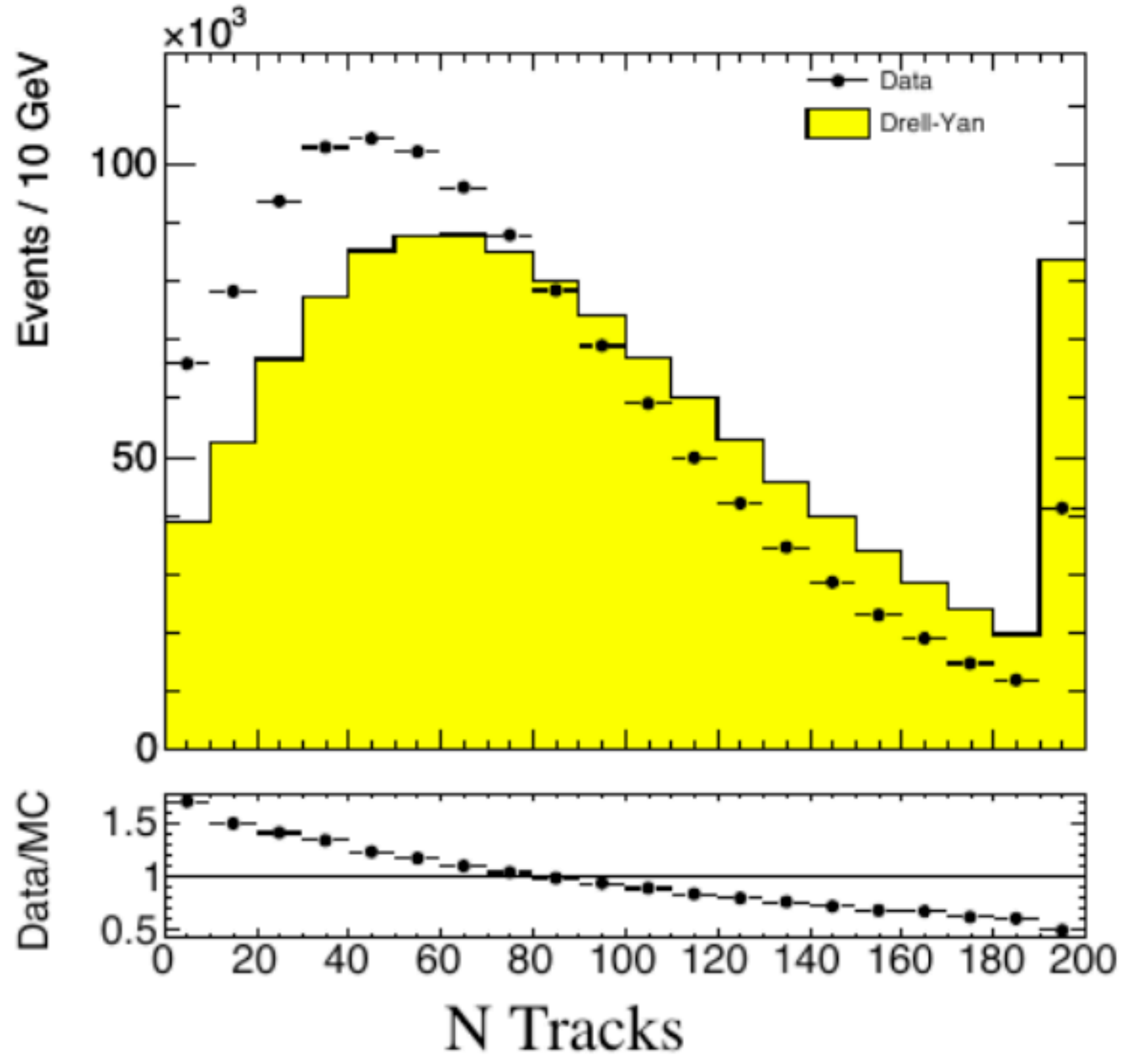




A zoom in of an event in ATLAS showing 30 reconstructed vertices. This event contains a Z->μμ process and many soft QCD interactions.

Re-Weighting MC

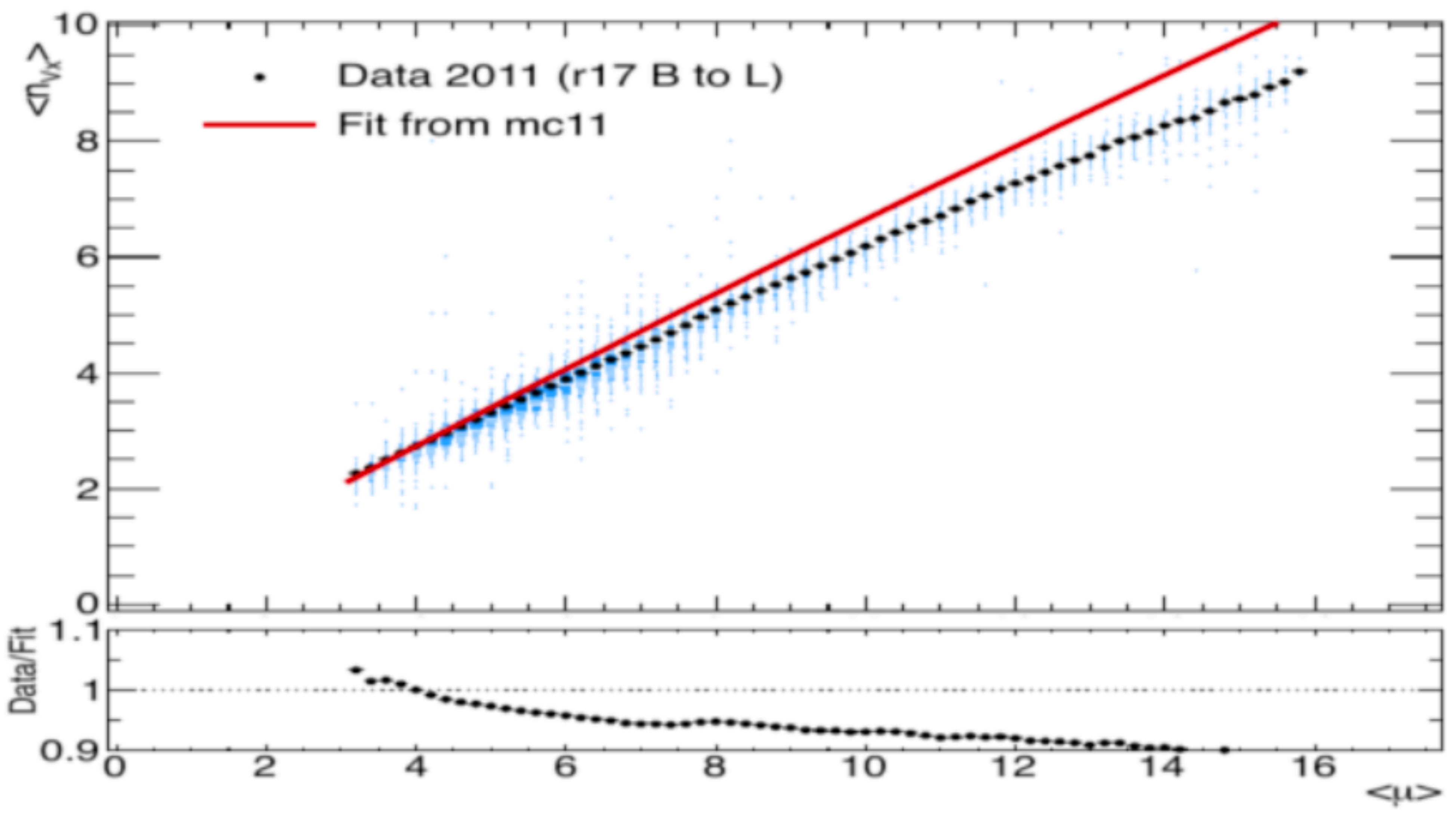
Number of tracks in *ATLAS* events



- The MC clearly does not describe the data
- What has gone so wrong?
- Is it parton showering, hadronisation or something else?
  - Need to understand the root cause of this disagreement

Pileup Re-Weighting

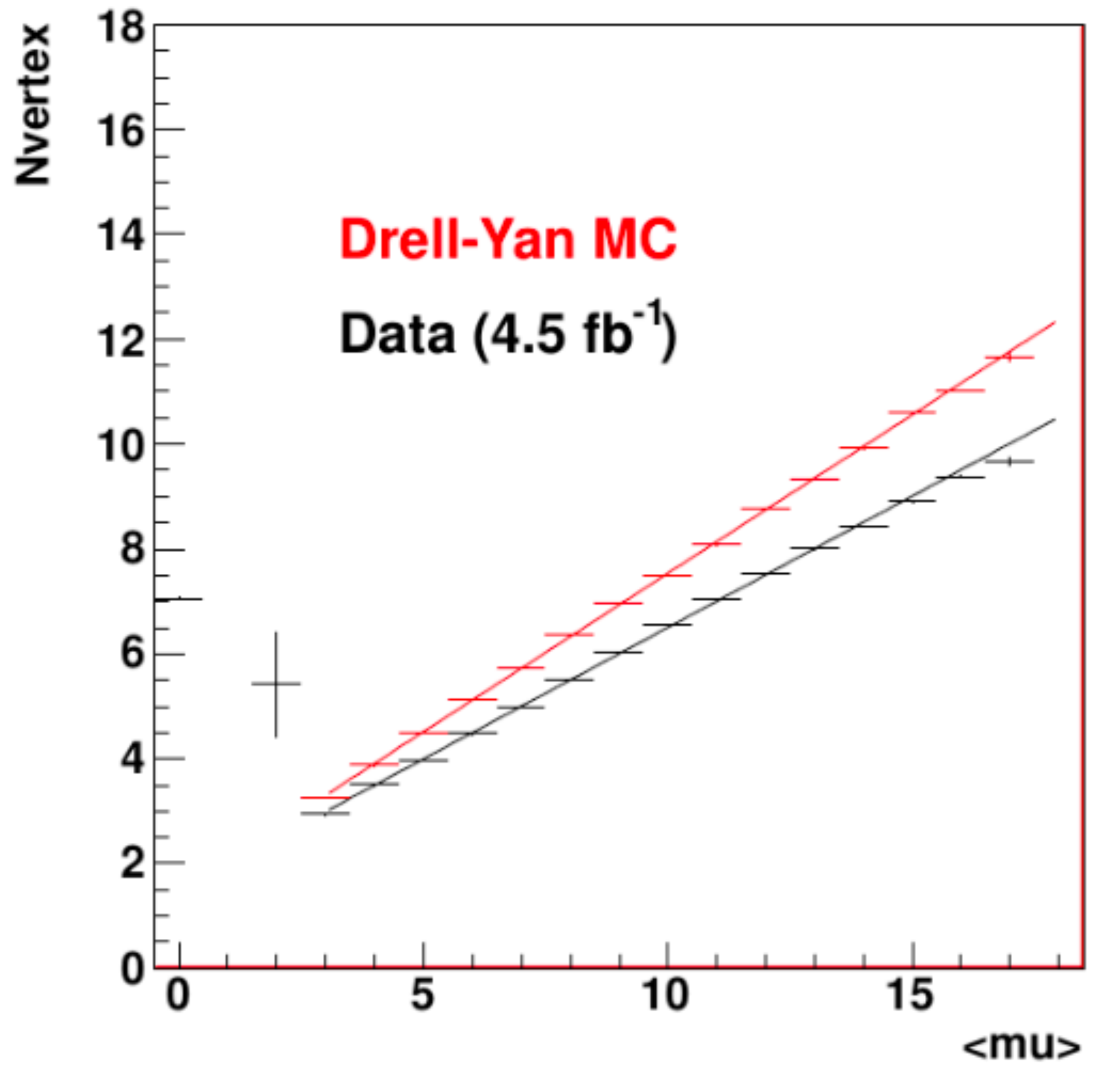
N vertices Vs average N interactions per bunch crossing



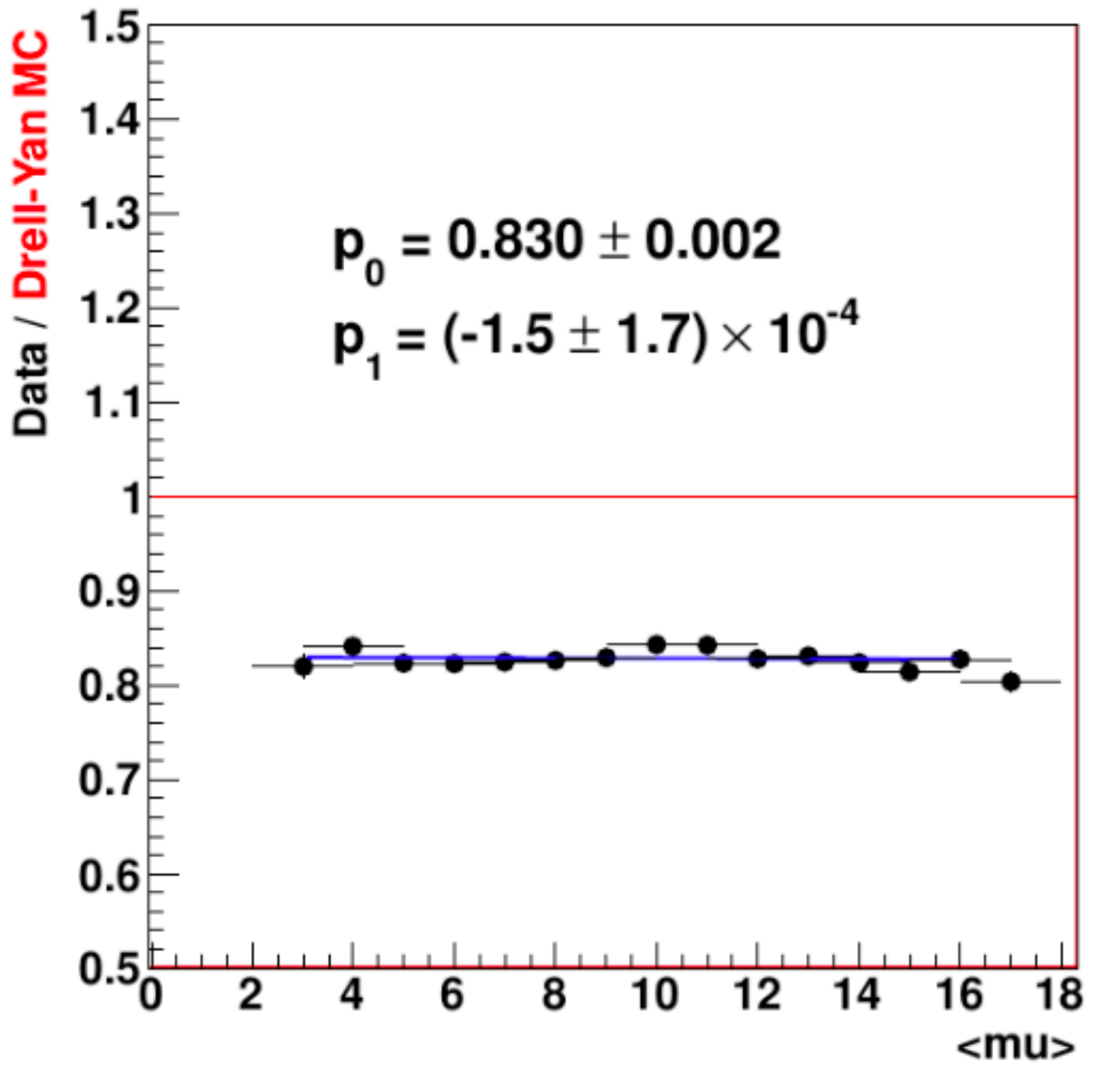
- Classic *ATLAS* example of MC not describing data accurately
- This shows that the MC gets the number of vertices wrong
  - Problem simulating proton bunches with 10^{11} protons
  - Understandably a very difficult task!
- Unfortunately this has big effects for many distributions

Pileup Re-Weighting

Need to determine re-weighting factors



(a) Data-MC comparison

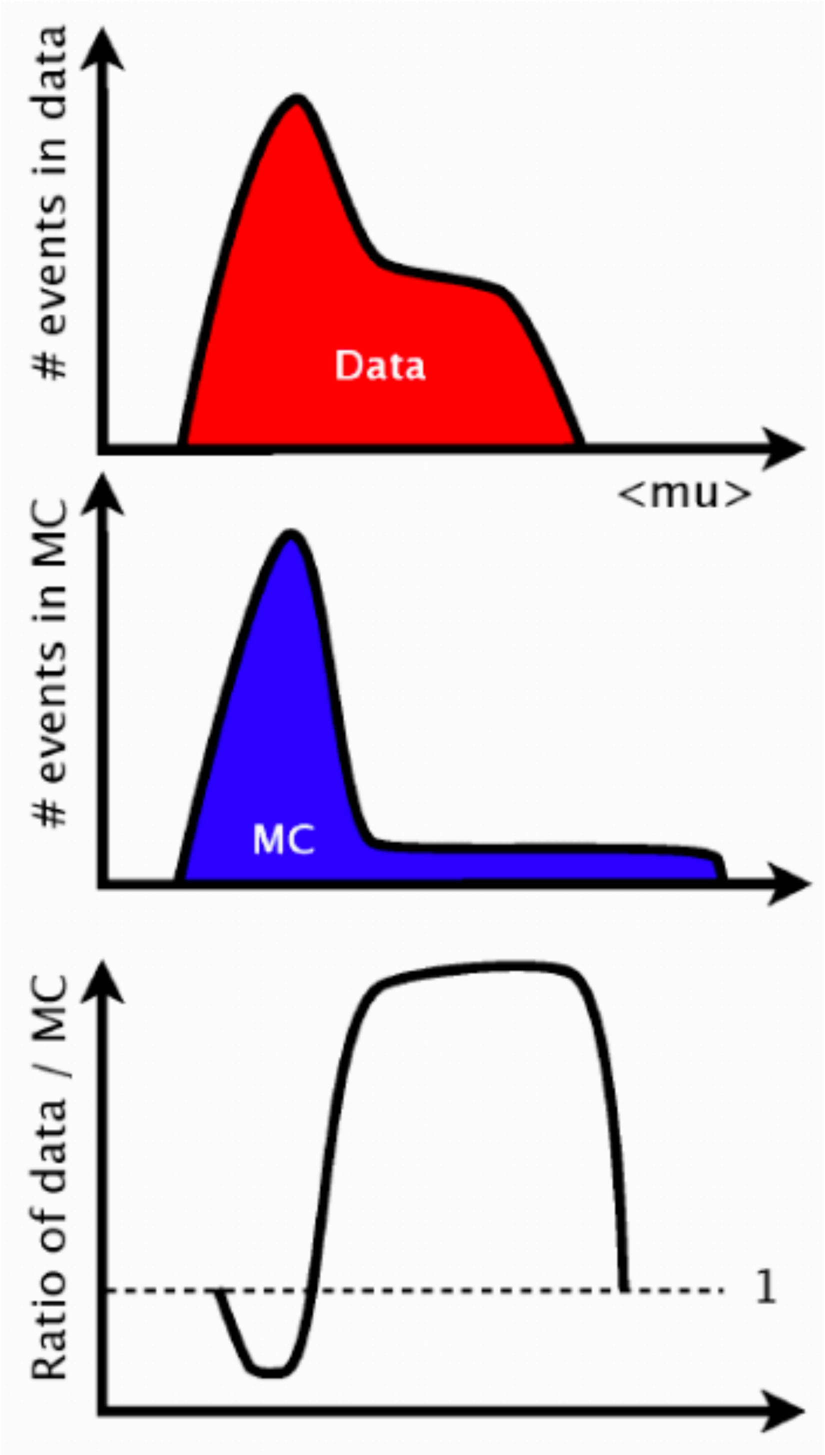


(b) Fit of the ratio of the distributions in (a)

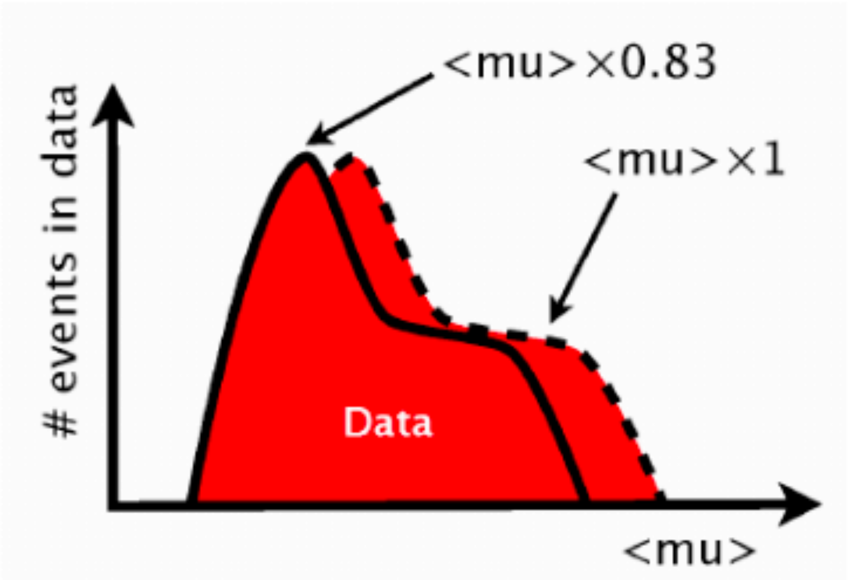
- Divide Data by MC to determine correction
- In this case, fit the ratio and determine a weight
- Use this weight for each MC event
  - histogram $\rightarrow$ Fill(x, weight);

Pileup Re-Weighting

Cartoon illustrating re-weighting procedure



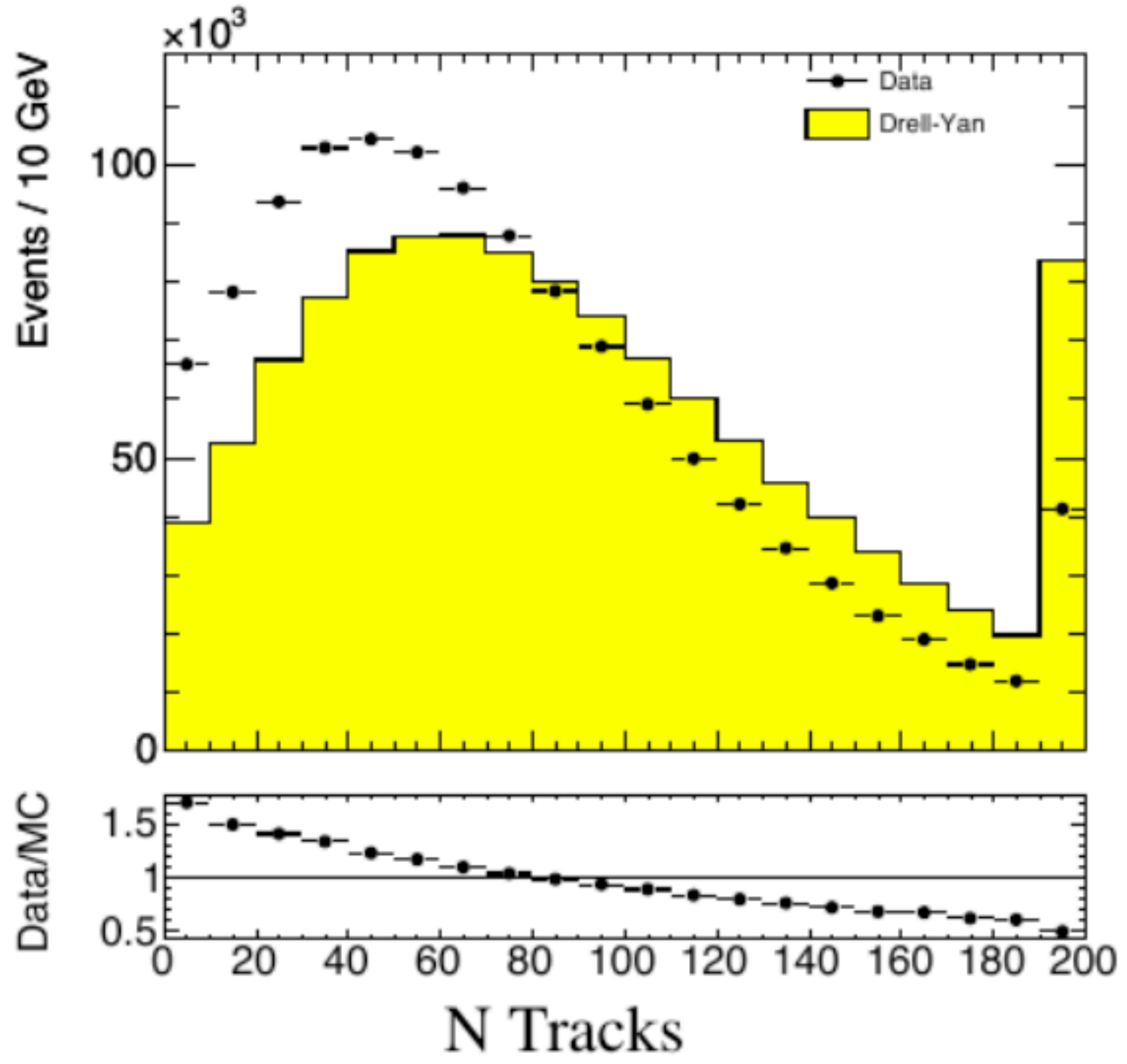
(a) Reweight procedure



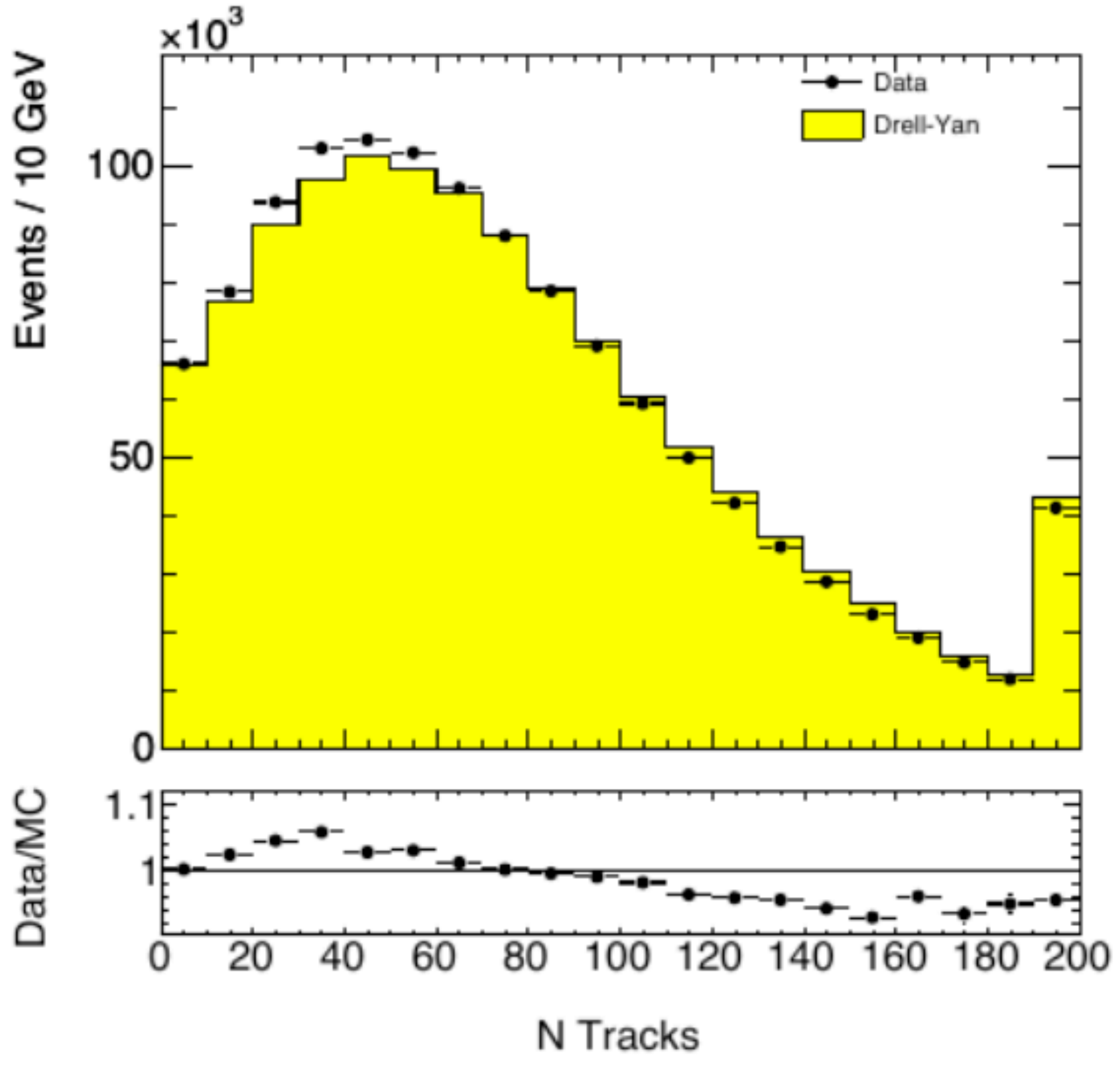
(b) Rescale procedure

Pileup Re-Weighting

Effect of re-weighting on N Tracks



(a) $\langle\mu\rangle$ rescale factor= 1 (no rescale)



(b) $\langle\mu\rangle$ rescale factor= 0.83

- (a) Before re-weighting MC describes the data poorly
- (b) After re-weighting the MC description is much better
- We have identified the underlying problem - N vertices
  - We have re-weighted the N vertices distribution
  - Can see effect in the number of tracks distribution

Re-Weighting

We require many different weights

- Pileup is just 1 example where we need weights
- Electrons, muons, taus, photons all require weights for:
  - Trigger, reconstruction, identification
- Jets require weights for:
  - Reconstruction, resolution, jet vertex fraction
- All of these weights provide a correction to specific aspects of data/MC disagreement
- Combined all (via multiplication) to provide an overall weight

$$W_{\text{Event}} = W_e \text{ x } W_\mu \text{ x } W_\tau \text{ x } W_{\text{jets}} \text{ x } W_{\text{Pileup}} \text{ x } W_{\text{other}}$$

$$\text{where } \; W_e = W_{\text{Trigger}} \text{ x } W_{\text{Reco}} \text{ x } W_{\text{ID}}$$

Acceptance

Acceptance (calculated from Monte Carlo) corrects for:

1. Finite resolution: events end up in the wrong bin.
2. Finite selection efficiency: some events end up in no bin at all.

$$\boxed{A_{CC}=N_{rec}/N_{gen}}$$

N_{rec} = # Events reconstructed (found) in a bin
N_{gen} = # Events generated in a bin

A_{CC} calculated from Monte Carlo

- A_{CC} depends on the shape of the cross section (smearing and migration).
- Monte Carlo must reproduce the data reasonably well.
- Systematic error should be calculated (vary input cross section: determine change on cross section).

Acceptance correction method reliable if bin widths greater than resolution. Careful as some of acceptance may be purely geometrical ! (Example: detector which only measures in one hemisphere has A_{CC}=0.5 even if everything else is perfect !)

Purity

Acceptance is not the whole story! Imagine a bin in which all events migrated out and were replaced by an equal number which had migrated in from elsewhere in the kinematic plane. $A_{CC} = 1$ but our binning is clearly stupid !

$$\text{Purity} = \frac{\#\,\text{events generated AND reconstructed in bin}}{\#\,\text{events reconstructed in bin}}$$

Flat cross section with perfect efficiency in 1D binning with 1σ width gives purity=68% (Gaussian errors).

Rule of thumb: ask for Purity > $(0.6)^n$ for an n dimensional binning

Purity is not used directly in evaluating the cross section. It is an example of a "control distribution" or "control plot".

Summary and conclusions

Eram Rizvi

Atlas Collaboration

Killer Questions for Physics Meetings

We can summarise some of what we have learned by listing questions which are always worth asking about your own or other people's analyses:

1. How have you defined the measured cross section ? If the answer relies on theoretical models, there may well be errors not properly treated.
2. What is the minimum purity in your measurement ? It's amazing how many people don't even know ! Be very suspicious if an evasive answer is given.
3. What is the largest source of systematic uncertainty ? Someone who has done a careful job will always know the answer, including:
   - what the next largest effect is,
   - what the largest effect is in different regions of phase space,
   - what they could do to reduce the error and why they didn't do it.
4. How did you check the size of your systematic errors ? For a good analysis, several different methods will have been used to check.
5. What is the physics message of your analysis ? Ask this question at the end for some fun discussion ! (You will know what I mean in a year...)

DEPARTAMENTO DE FÍSICA NUCLEAR E ALTAS ENERGIAS

Summary and conclusions

Eram Rizvi

Atlas Collaboration

Final Words

- Analysis in HEP is complicated ! Common sense is very important.
- HEP Collaborations are chaotic: much useful knowledge is "folklore". Getting to know people and their work is vital for you:
  - Attend working group and collaboration meetings
  - Try to understand other analyses !
- Make a useful contribution ! A Ph.D. which does not contribute to a journal publication is a waste of time scientifically !
  - Do not reinvent the wheel: use all that exists and build on it to invent something genuinely new !
  - Ask yourself where your analysis is heading: if it isn't contributing in some way to a publication, ask yourself why not. Ask your supervisor or RAs as well.


To DO list.
1. DRINK COFFEE.
2 TRY TO TAKE OVER THE WORLD!

References

[1] HEP Analysis Lectures by Dr. John Morris,Queen Mary University of London, (link)

[2] P*hysics at the High-Luminosity LHC*. *Pippa Wells*, CERN, on behalf of the. ATLAS and CMS Collaborations. UK HEP Forum (link)

[3] Introduction to Monte Carlo for Particle Physics Study, N. SRIMANOBHAS (Dept. of Physics, Faculty of Science, Chulalongkorn University) (link)

[4] Data Analysis and Statistical Methods in Experimental Particle Physics,Thomas R. Junk, Fermilab, Hadron Collider Physics Summer School 2012, (link)

[5] The LHC machine in Run-2, and physics highlights from the ATLAS experiment, Jamie Boyd (CERN), Erice School 2018 (link)

[6] Analysis Methods in High Energy Physics lectures by Eram Rizvi (link)