Análise Dados em HEP

Definindo a estratégia de análise

Felipe Torres Da Silva De Araujo

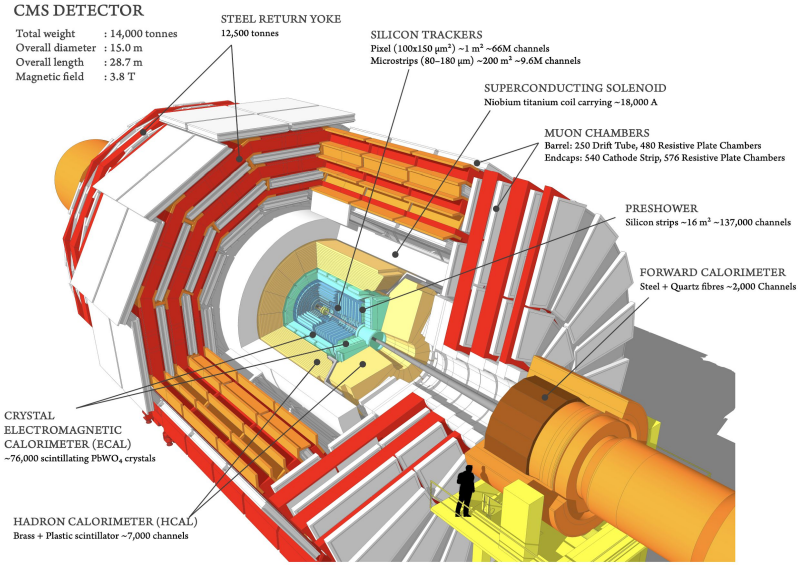
08/12/2020

Outline

- Signal and background process and samples
- Signal selection
- Correction Factors
- Statistical Modeling
- Control Regions
- ~~Systematics~~

Experimental Setup

Compact Muon Solenoid - CMS

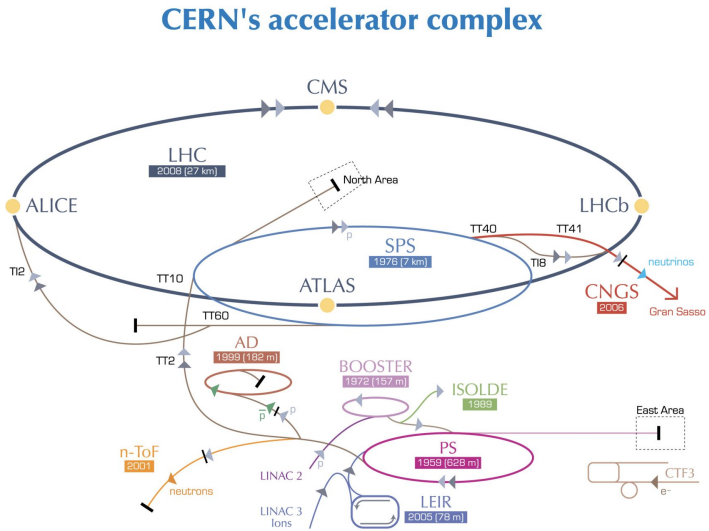


The Compact Muon Solenoid (CMS) is a multiple purpose experiment used to investigated pp as well as lead-lead collisions at the LHC.

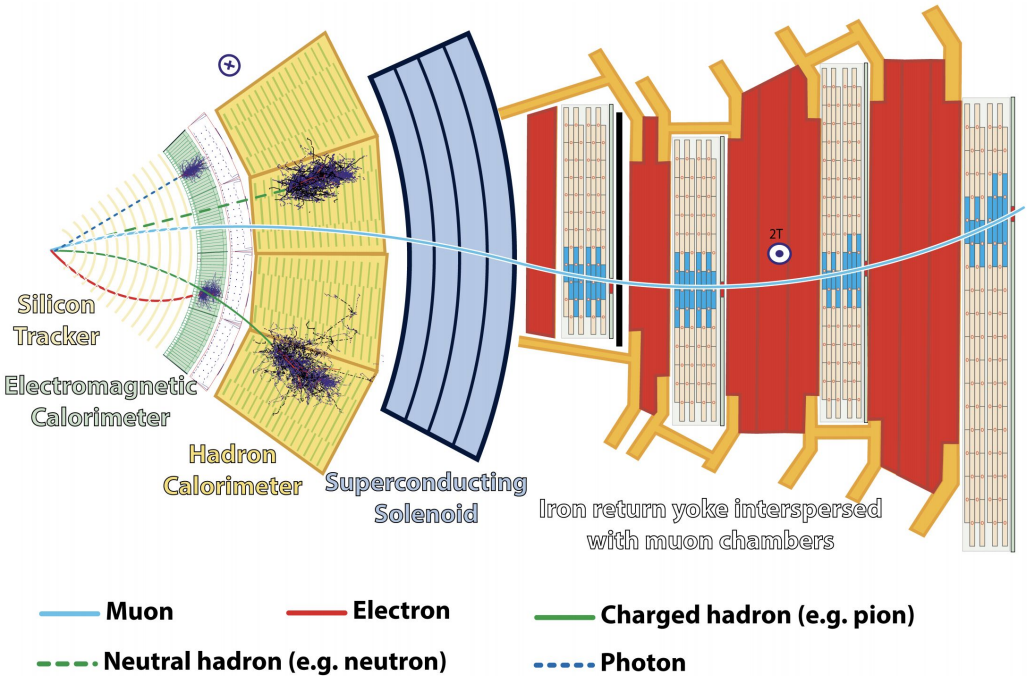
It is operated by the CMS Collaborations, composed by around 5000 researches and 20 institutes.

The CMS is located in the city of Cessy, France, 100 m below the surface.

- pp accelerator (27 km) designed, built and operated by CERN.
- Located in the border of Switzerland and France, close to Geneva.
- Capable of collide bunches of protons at center-of-mass energies up to 14 TeV (13 TeV for this study), with bunch crossing of 25 ns. Protons are accelerated inside vacuum pipes by RF cavities and a driven by a superconducting magnetic system.
- 4 main detectors installed at interaction points.



Particle Flow Algorithm



The global event reconstruction aims to reconstruct and identify each individual particle in an event, with an optimized combination of all subdetector information.

In this process, the identification of the particle type (photon, electron, muon, charged hadron, neutral hadron) plays an important role in the determination of the particle direction and energy.

Signal and background process and samples

Signal and Background

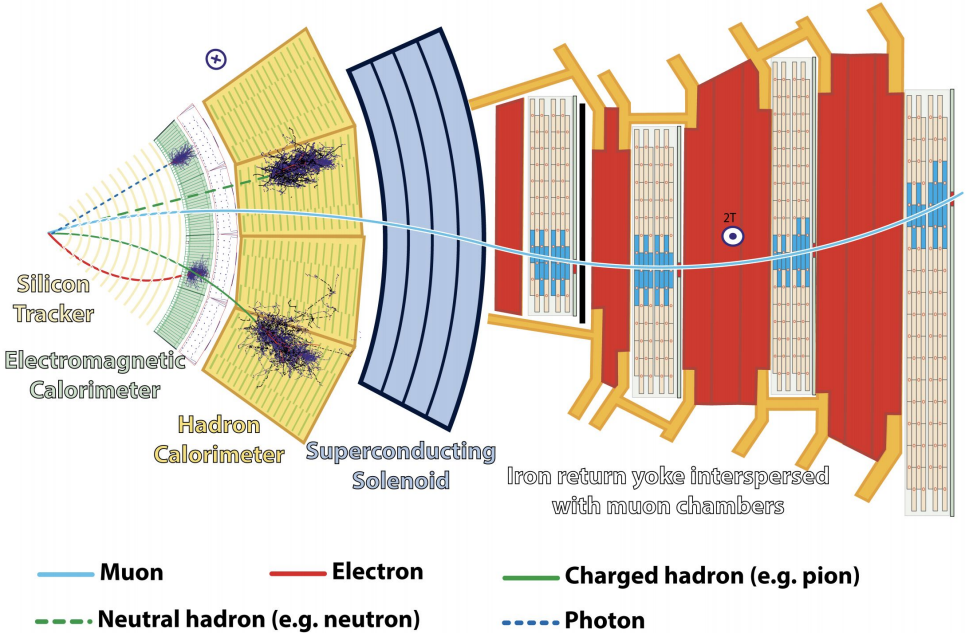
Given a certain process of interest:

- Signal is the process itself.
- Background is the set of process, accessible by the detector, that through the analysis would overlap (or mimic) with the signal distribution.
- Sources of background:
  - Detector effects: mis-identified objects.
  - Statistical effects: combinatorial background.
  - Physics processes: Process with similar/same final state.
  - Collider effects: Pile-up, beam halo.

Signal and Background

Given a certain process of interest:

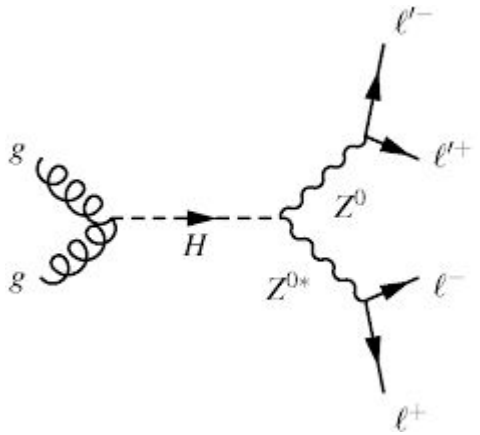
- Signal is the process itself.
- Background is the set of process, accessible by the detector, that through the analysis would overlap (or mimic) with the signal distribution.
- Sources of background:
  - **Detector effects: mis-identified objects.**
  - Statistical effects: combinatorial background.
  - Physics processes: Process with similar/same final state.
  - Collider effects: Pile-up, beam halo.



Signal and Background

Given a certain process of interest:

- Signal is the process itself.
- Background is the set of process, accessible by the detector, that through the analysis would overlap (or mimic) with the signal distribution.
- Sources of background:
  - Detector effects: mis-identified objects.
  - **Statistical effects: combinatorial background.**
  - Physics processes: Process with similar/same final state.
  - Collider effects: Pile-up, beam halo.

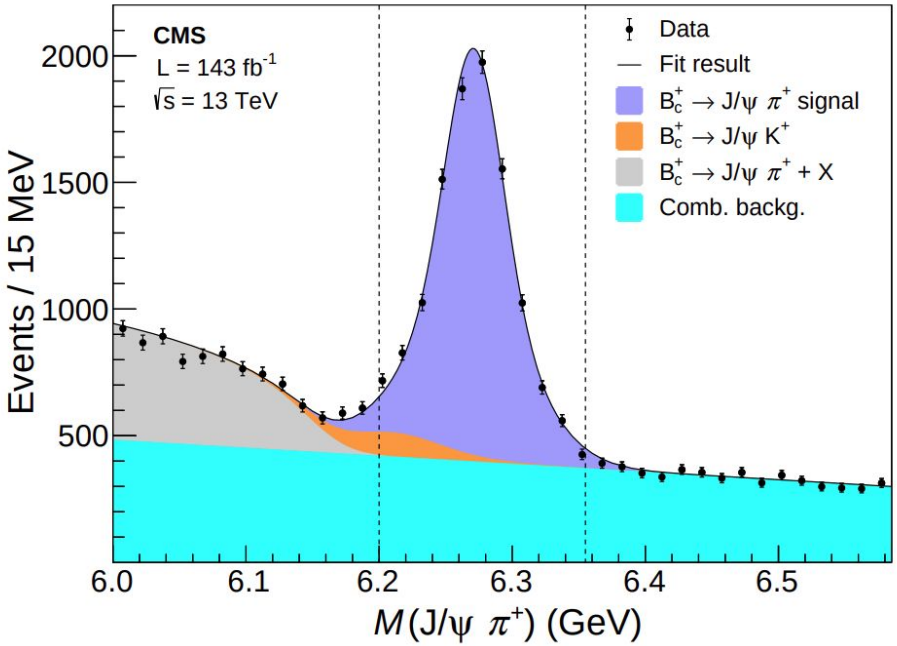


- H → ZZ → 4 leptons

Signal and Background

Given a certain process of interest:

- Signal is the process itself.
- Background is the set of process, accessible by the detector, that through the analysis would overlap (or mimic) with the signal distribution.
- Sources of background:
  - Detector effects: mis-identified objects.
  - Statistical effects: combinatorial background.
  - **Physics process: Process with similar/same final state.**
  - Collider effects: Pile-up, beam halo.



Signal and Background

Given a certain process of interest:

- Signal is the process itself.
- Background is the set of process, accessible by the detector, that through the analysis would overlap (or mimic) with the signal distribution.
- Sources of background:
  - Detector effects: mis-identified objects.
  - Statistical effect: combinatorial background.
  - Physics process: Process with similar/same final state.
  - **Collider effects: Pile-up, beam halo, MPI.**

- Beam Halo monitor at CMS

MC Samples

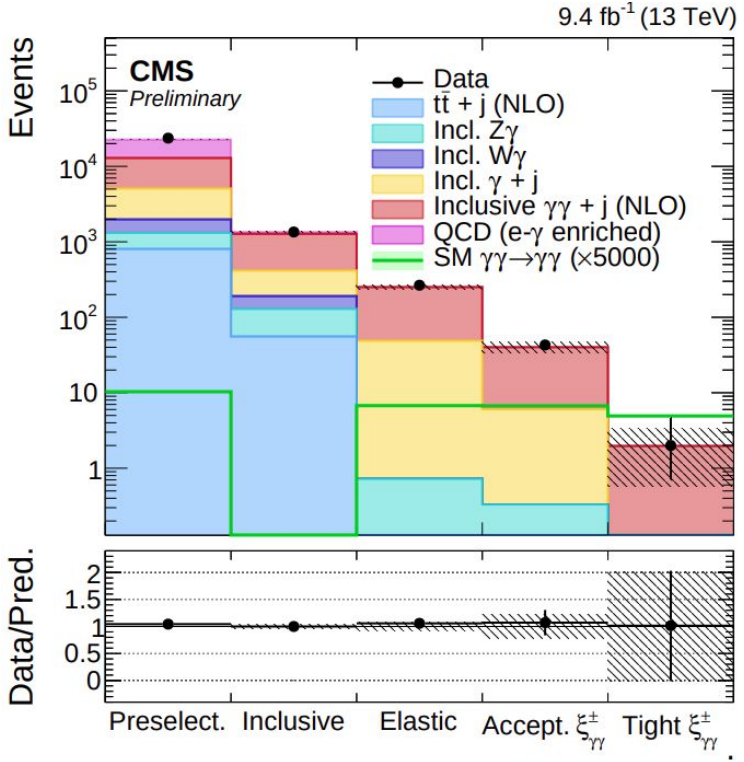
- In general you want a (set) of MC that is a very good representation of your signal.
- For the background, some MC is always welcome, but the most important is:
  - Not all background are simulatable.
  - Not all backgrounds are relevant (effective cross-section plays a role on this).
  - Some backgrounds are really easy to remove.
  - **As much as possible, background contributions should be extracted from Data.**
- **MC samples are just a representation of the contributing processes in your analysis.**
- **It is your responsibility, as the analyst, to prove that this representation is good enough.**

Sample Type	Process
ZZ	$q\bar{q} \to ZZ \to 4\ell$ $q\bar{q} \to ZZ \to 4\ell$ $gg \to ZZ \to 4e$ $gg \to ZZ \to 4\mu$ $gg \to ZZ \to 4\tau$ $gg \to ZZ \to 2e2\mu$ $gg \to ZZ \to 2e2\tau$ $gg \to ZZ \to 2\mu2\tau$
WZ	$WZ \to 3\ell\nu$
Drell-Yan	$Z/\gamma^* \to \ell^+\ell^-$ jets $Z/\gamma^* \to \ell^+\ell^-$ jets
$t\bar{t}$	$t\bar{t}$ $t\bar{t} \to \ell\ell$ $t\bar{t} \to \ell\ell$
VVV	WWZ WZZ ZZZ ZZZ
Higgs	$gg \to H \to ZZ \to 4\ell$
ZZ + jets	ZZ + 2jets (EWK)

- pp → ZZ → 4 leptons

MC Samples

- In general you want a (set) of MC that is a very good representation of your signal.
- For the background, some MC is always welcome, but the most important is:
  - Not all background are simulatable.
  - Not all backgrounds are relevant (effective cross-section plays a role on this).
  - Some backgrounds are really easy to remove.
  - **As much as possible, background contributions should be extracted from Data.**
- **MC samples are just a representation of the contributing processes in your analysis.**
- **It is your responsibility, as the analyst, to prove that this representation is good enough.**

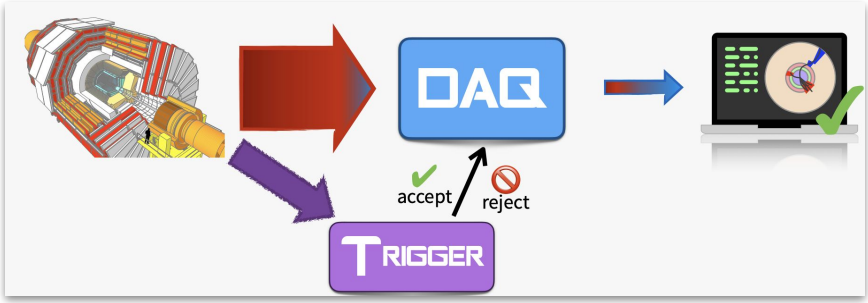


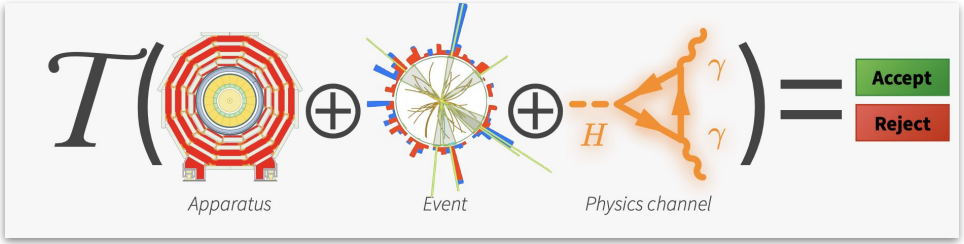
First search for exclusive diphoton production at high mass with intact protons in proton-proton collisions at √s = 13 TeV at the LHC.

Signal selection

Trigger

- Not all interactions can be recorded. The trigger menu of the experiment will define a set of characteristics that makes the event interesting, somehow.
- At CMS:
  - LHC @ 40 MHz → L1 Trigger (online) @ 1 kHz
  - L1 @ 1 kHz → HLT (offline) @ 100 Hz
- (Multi leveled) Trigger menus takes into account:
  - Cost
  - Event size * Expected Rate
  - Trigger efficiency (given a certain set in process of interest).
- In general:
  - Rate should be small.
  - Efficiency should be high.
- ISOTDAQ





The average event size determines the allowed trigger rate:

$$B_{DAQ} = R_T^{max} \times S_E$$

DAQ Bandwidth — Maximum Trigger rate — Event size

How many particles per event?
How many FE channels?

FE channels in GPD (ATLAS/CMS)
- O(10⁷) in inner detector
- O(10⁵) in calorimeters
- O(10⁶) in muon detectors

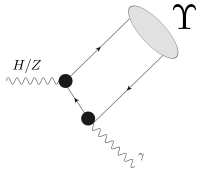
$$\epsilon_{trg} = \frac{N_{good}^{accepted}}{N_{good}^{prod/exp}}$$

Trigger

- Possible strategies for a analysis:
- Direct triggering.
- Indirect triggering.

Trigger

- Possible strategies for a analysis:
- **Direct triggering.**
  - One should try to be loose.
  - Don't trigger right just on your signal.
- Indirect triggering.



Single Muon (pT > 17 GeV) and Single Photon (ET > 33 GeV)

$\Upsilon(1S)\mu^+\mu^-$

Single Muon (pT > 0 GeV) and a Dimuon system with mass compatible with a Y(nS).

$B_c^{\pm} \rightarrow J/\psi\pi^{\pm}$ and $J/\psi \rightarrow \mu^+\mu^-$

Two muons with mass compatible with a J/Psi and a displaced track.

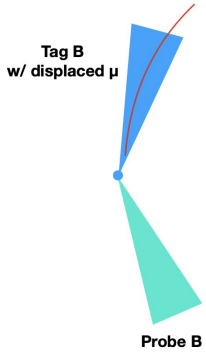
Trigger

- Possible strategies for a analysis:

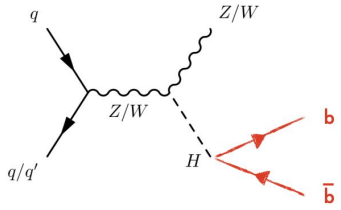
- Direct triggering.
  - One should try to be loose.
  - Don't trigger right just on your signal.

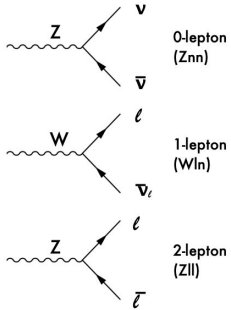
- **Indirect triggering.**
  - Instead of triggering on the signal process, trigger on associated production.

  - Not always possible, but if doable, would give you a unbiased signal sample.



CMS B-Parking strategy: 20% of the B hadrons decays goes to a final including a muon. This muon can be used as a tag for B hadrons enhanced events.



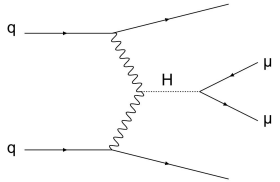


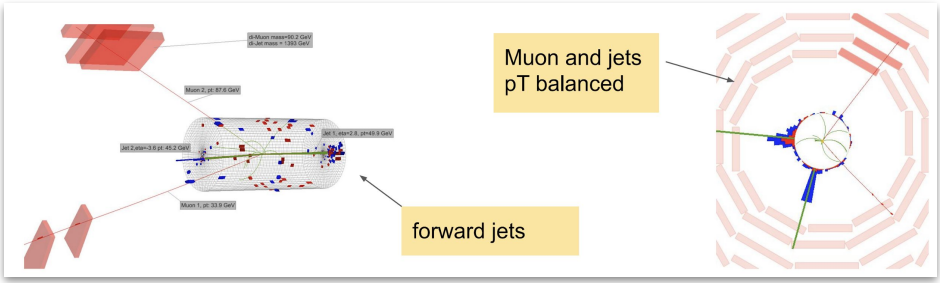
H → bb (inclusive) would be overwhelmed by QCD background. Triggering on the a high-pT lepton helps to explore the VH production mode and enhance sensitivity.

Signal Selection

- A lot of works goes here.
- Your expertise as analyst plays a major role.
  - Similar analysis ➡ Similar Selection
- Control regions (sidebands) should also be foreseen in this phase.

- H ➡ dimuon (Run2)





A common preselection is applied on all events

- pT (μ) > 30, 20 GeV
- pT (jet) > 35, 25 GeV
- m(jj) > 400 GeV
- Δη(jj) > 2.5
- no bTag medium, < 2 bTag loose (ttH analysis)
- no additional leptons (VH analysis)

- **Signal Region** | m(μμ) - 125 | < 10
- **Sideband** 110 < m(μμ) < 115 or 135 < m(μμ) < 150
- **Z Region** | m(μμ) - 91 | < 15 *(only used to test bkg model from MC with higher statistics)*

Correction Factors

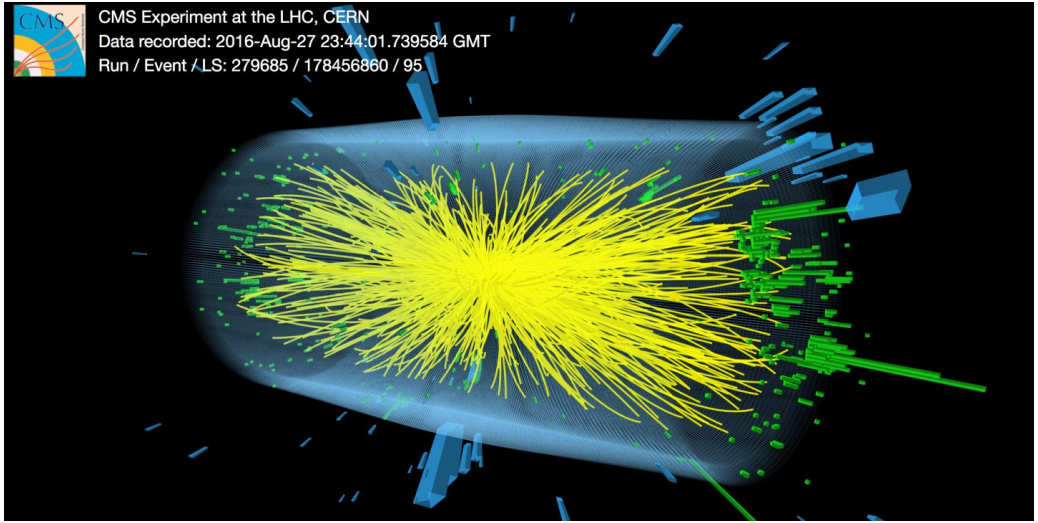
Corrections Factors

- There are intrinsic differences between MC and Data, that should be taken into account.

- **Some of them, should be corrected.**
- **Some should be quoted as systematics.**
- **Some are negligible.**

- Possible correction factors:
  - Pileup
  - Trigger simulation
  - Physics object reconstruction
  - ...

- Many are centrally produced by the collaborations.



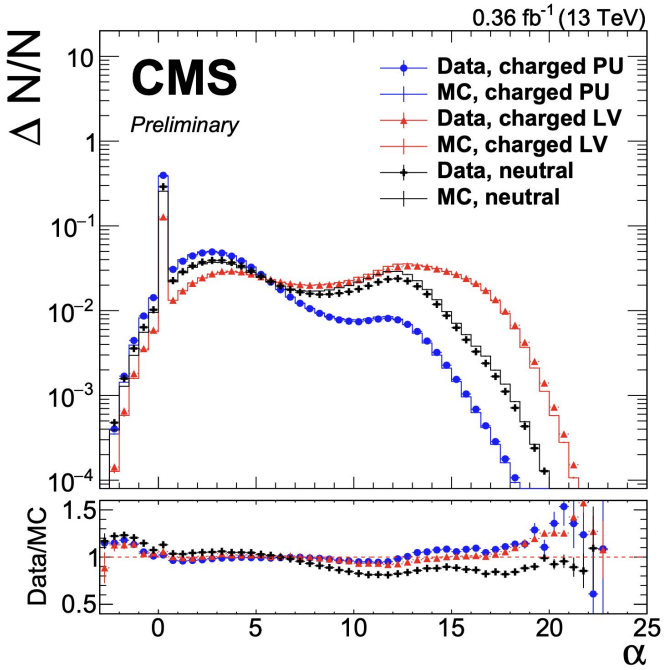
$$W_{\text{Event}} = W_e \times W_\mu \times W_\tau \times W_{\text{jets}} \times W_{\text{Pileup}} \times W_{\text{other}}$$

$$\text{where} \quad W_e = W_{\text{Trigger}} \times W_{\text{Reco}} \times W_{\text{ID}}$$

Pileup (PU)

- At CMS, a "event" is not a isolated pp collision.
- For each bunch crossing, many interactions happen. majority should not be interesting.
- There are intrinsic differences between MC and Data, that should be taken into account.
- MC (@ CMS) samples are already simulated including some pileup distribution.
- After comparison with measured PU at Data, the effective weight/event is updated.

$$w_{PU}(n) = \frac{P_{PU}^{Data}(n)}{P_{PU}^{Sim}(n)}$$

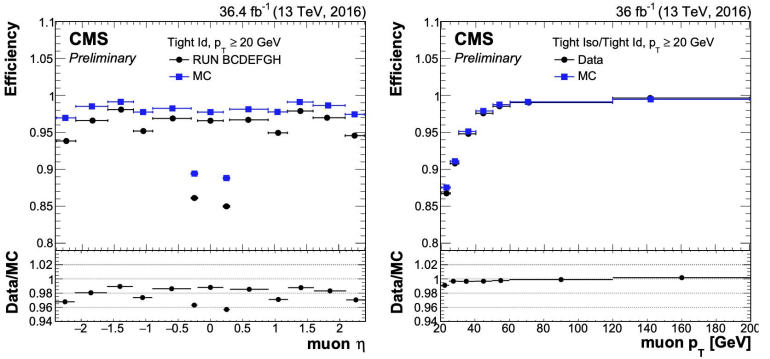


Scale Factors

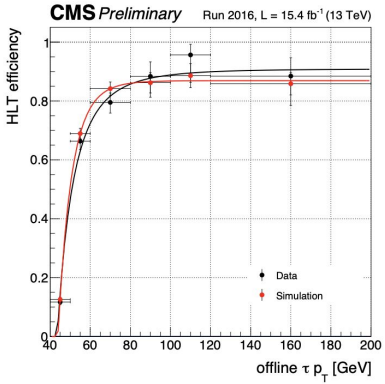
- Data and MC simulation might have different performance, in terms of efficiency.
- One can define a set of scale factors (per event weight) that would correct the the MC samples to match the Data samples.
- **What scale factors should be applied?**
  - **The ones that the analyst + collaboration define as important for the analysis.**
  - **In general (@ CMS): Trigger, object identification, object isolation, any other "decision" taken along selection procedure.**

$$Eff = \frac{\text{selected events/objects}}{\text{good events/objects}}$$

Muon Identification and Isolation @ CMS - 2016

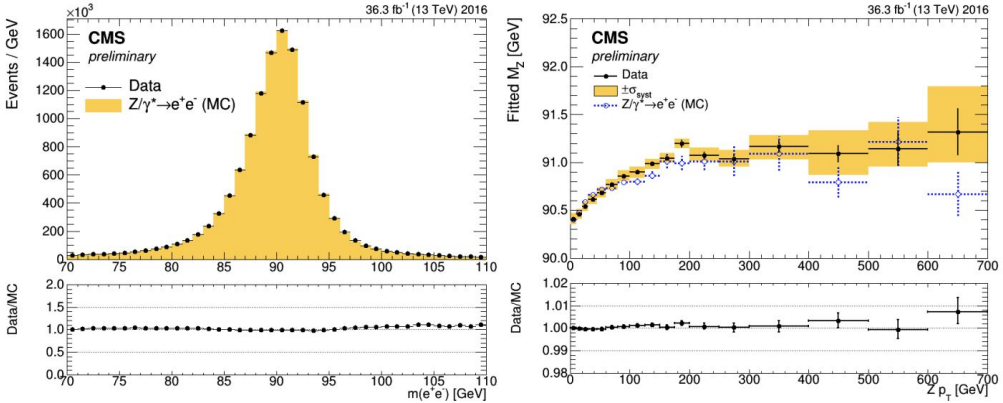


HLT Trigger tau efficiency x pT

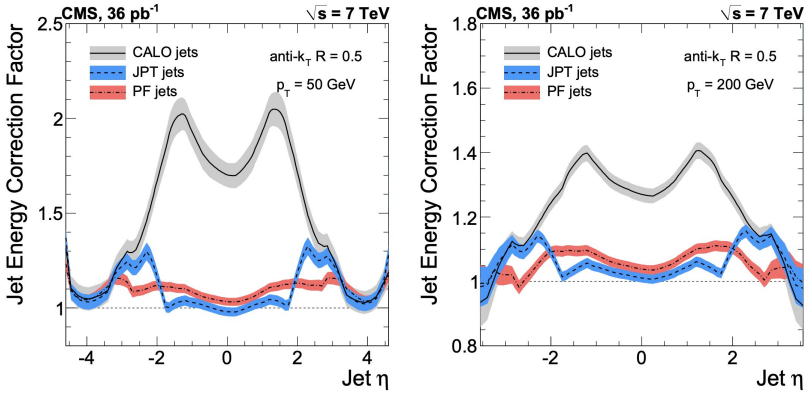


Scale and resolution

- Experimental constraints might affect the scaling and resolution of reconstructed objects.
- In the Figure (a), the MC mass distribution of dielectrons, around Z peak, is compared with data.
- MC have been corrected by scaling and resolution.
  - Breit-Wigner convoluted with a Gaussian
  - Corrections are propagated to electron pT.
- Figure (b) the fitted Z mass is plotted as a function of the Z pT.
  - Higher pT, implies in higher differences MC x Data.
- Jets usually are corrected in their energy.



$Z \to e^+e^-$ mass distribution (a) and fitted Z mass versus Z p_T (b)



Statistical Modeling

Modeling

- For a analysis is interested in measure a cross-section (or derived):
- One should be able to measure the number of expected/observed events.
- Different options:
  - Cut and count
  - Binned shape
  - Parametric shape

$$\sigma = \frac{N_{obs} - N_{bkg}}{\mathcal{L} \,.\, \epsilon \,.\, BR}$$

- Where:
  - N_{obs} is the number of observed events
  - N_{bkg} is the number of expected background events
  - $\mathcal{L}$ is the total integrated luminosity
  - ϵ is the acceptance efficiency
  - BR is the branching ratio

$$A \cdot \epsilon \equiv \frac{N_{\text{acc}}}{N_{\text{gen}}} \cdot \frac{N_{\epsilon}}{N_{\text{acc}}} = \frac{N_{\epsilon}}{N_{\text{gen}}},$$

Modeling

- For a analysis is interested in measure a cross-section (or derived):
- One should be able to measure the number of expected/observed events.
- Different options:
  - **Cut and count**
    - Count the number of events that pass full selection.
    - Event counts (rates) are modeled as a Poisson distribution.
    - Example: DY cross-section at 7 TeV http://cdsweb.cern.ch/record/1372196/files/arXiv:1108.0566.pdf
  - Binned shape
  - Parametric shape

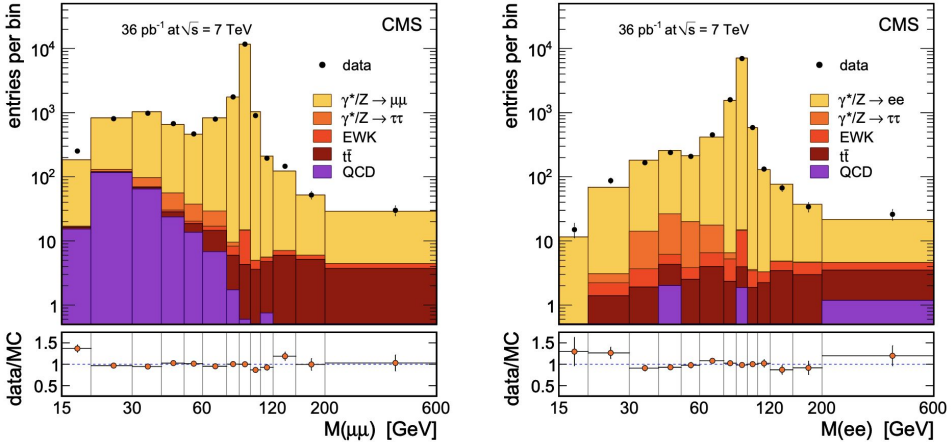
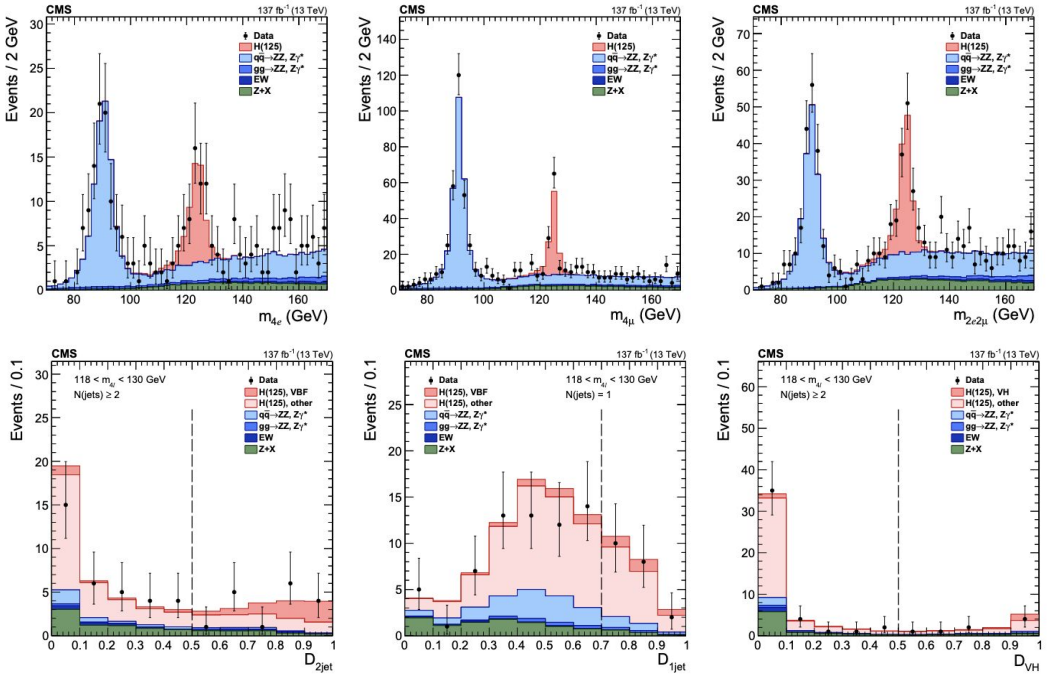


Figure 1: The observed dimuon (left) and dielectron (right) invariant mass spectra. No corrections are applied to the distributions. The points with error bars represent the data, while the various contributions from simulated events are shown as stacked histograms. By "EWK" we denote $Z/\gamma^* \to \tau\tau$, $W \to \ell\nu$, and diboson production. The "QCD" contribution results from processes associated with QCD and could be genuine or misidentified leptons. The lower panels show the ratios between the measured and the simulated distributions including the statistical uncertainties from both.

Modeling

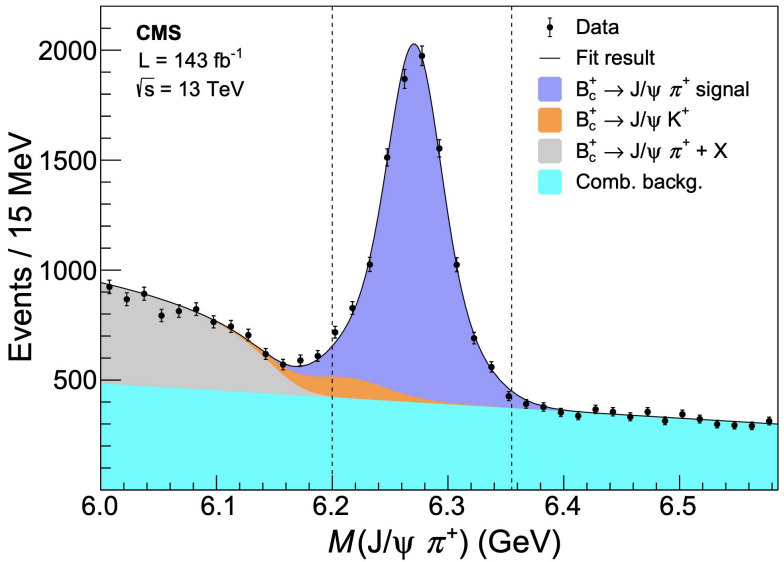
- For a analysis is interested in measure a cross-section (or derived):
- One should be able to measure the number of expected/observed events.
- Different options:
  - Cut and count
  - **Binned shape**
    - Similar to the previous approach.
    - Data/MC is binned at their distribution (histograms) are used as model the infere the result.
    - Number of events is extracted from histograms relative normalizations.
    - Example: H → ZZ→ 4l http://cds.cern.ch/record/2668684
  - Parametric shape



$$\mathcal{L}_{2D}(m_{4\ell}, \mathcal{D}^{\text{kin}}_{\text{bkg}}) = \mathcal{L}(m_{4\ell})\mathcal{L}(\mathcal{D}^{\text{kin}}_{\text{bkg}}|m_{4\ell}).$$

Modeling

- For a analysis is interested in measure a cross-section (or derived):
- One should be able to measure the number of expected/observed events.
- Different options:
  - Cut and count
  - Binned shape
  - **Parametric shape**
    - Functional forms of a pdfs are used.
    - Binning doesn't play a role in the result, but there are the pdf parameters.
    - Example: Study of excited B_c meson decays to B_c π π https://arxiv.org/pdf/1902.00571.pdf



- **Signal**: two Gaussian functions with a common mean.
- **Combinatorial Background:** a first-order Chebyshev polynomial.
- **Partially reconstructed B_c decay:** ARGUS function convolved with a Gaussian resolution function + MC derived.

The unbinned maximum-likelihood fit gives a B_c^+ signal yield of 7629 ± 225 events, a B_c^+ mass of $M(B_c^+) = 6271.1 \pm 0.5\,\text{MeV}$, and a mass resolution of $33.5 \pm 2.5\,\text{MeV}$, where the uncertainties are statistical only.

Background Modeling from Data

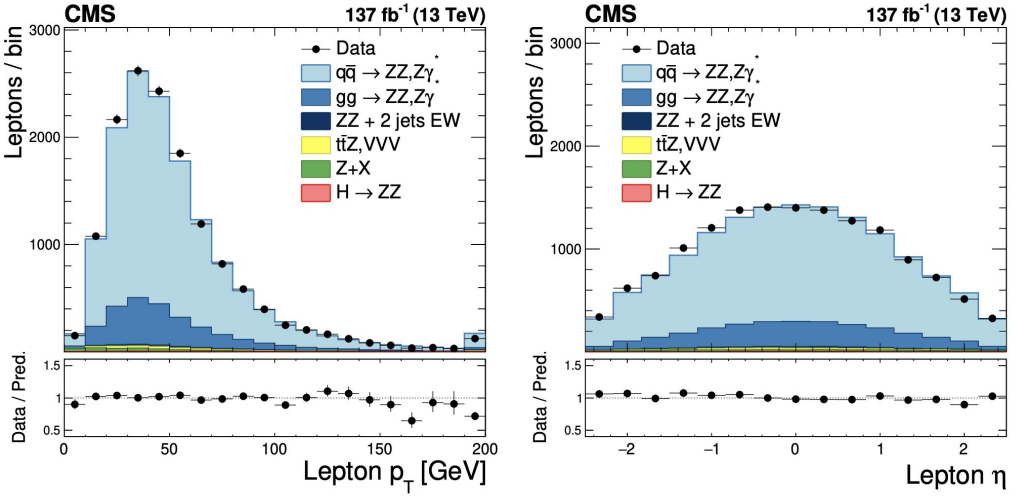
- **Unless with a very good justification, background contributions should be modeled from Data.**
- MC inputs are, of course, welcome.
- Normalization ➜ Data

- Two strategies (there should be more):
  - Normalization: From Data; Shape: From MC
  - Normalization: From Data; Shape: From Data

Background Modeling from Data

- **Unless with a very good justification, background contributions should be modeled from Data.**
- MC inputs are, of course, welcome.
- Normalization ➡ Data

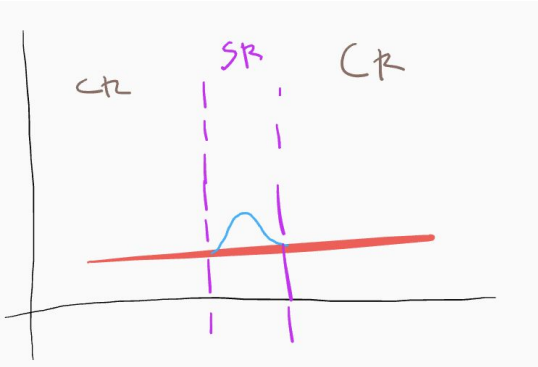
- Two strategies (there should be more):
  - **Normalization: From Data; Shape: From MC**
    - First step: Use MC to get the a matching on distributions shape.
    - Second step: Use a non-signal selection to estimate signal region normalization.
    - For background, comparing a signal and non-signal regions, the normalizations should be, at least proportional.
    - Another possibility: ABCD Method [REF]
  - Normalization: From Data; Shape: From Data



ZZ cross section with full Run 2 data: Distributions of (left) transverse momentum and (right) pseudorapidity for individual leptons. [REF]

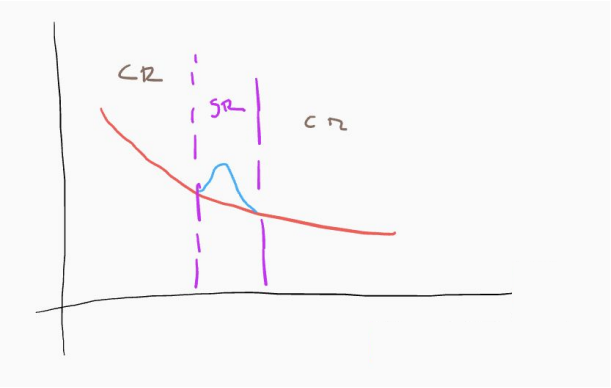
To account for all sources of background events, two control samples are used to estimate the number of background events in the signal regions. Both are defined as samples that contain events with a dilepton candidate satisfying all requirements (Z_1) and two additional lepton candidates $\ell^+\ell^-$. In one control sample, enriched in WZ events, one ℓ candidate is required to satisfy the full identification and isolation criteria and the other must fail the full criteria and instead satisfy only the relaxed ones; in the other, enriched in Z+jets events, both ℓ candidates must satisfy the relaxed criteria, but fail the full criteria. The additional leptons must have

Background Modeling from Data



Consider a region in CR with the same size as SR.
The normalization in this region should be the SR.

Background Modeling from Data

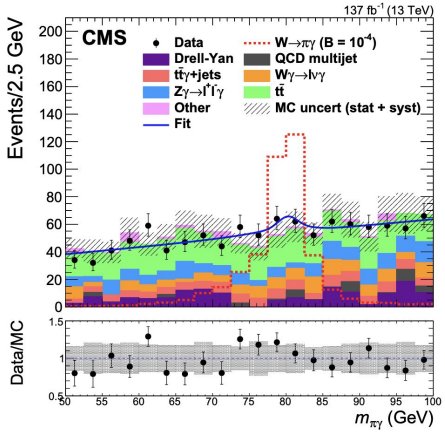


What is the proportionality????

Background Modeling from Data

- **Unless with a very good justification, background contributions should be modeled from Data.**
- MC inputs are, of course, welcome.
- Normalization ➔ Data

- Two strategies (there should be more):
  - Normalization: From Data; Shape: From MC
  - **Normalization: From Data; Shape: From Data**
    - **Side-bands:** Fit Data, with a functional form, excluding the signal region.
    - **Same sign:** Build a Background model, by combining same sign objects (good for neutral decays).
    - **Opposite flavour:** Build a Background model, by combining opposite flavour objects (eg. muon + electron).
    - **Event mixing:** Build a Background model, by combining objects from different events.



Search for the rare exclusive hadronic decay W->pi+gamma at 13 TeV: Event distribution as a function of $m_{\pi\gamma}$ for the combination of the lepton channels. [REF]

Divide and Conquer

- Different sensitivities will be found when:
  - objects are reconstructed in different regions of the detector;
  - tagging in production modes or associated processes;
  - selecting different kinematical regions.

- One could explore those higher sensitivities and divide the analysis in different categories (bins) and then combine the results.

- More or less equivalent to divide your data in bins of a histograms, analyse each bin separately and combine the results.

Too much categorization could jeopardize your measurement (less statistics → higher statistica error).
There should be a balance between categorization and sensitivity.

- Consider N measurements of the same quantity. The combination of this measurements is the one that minimizes [REF]:

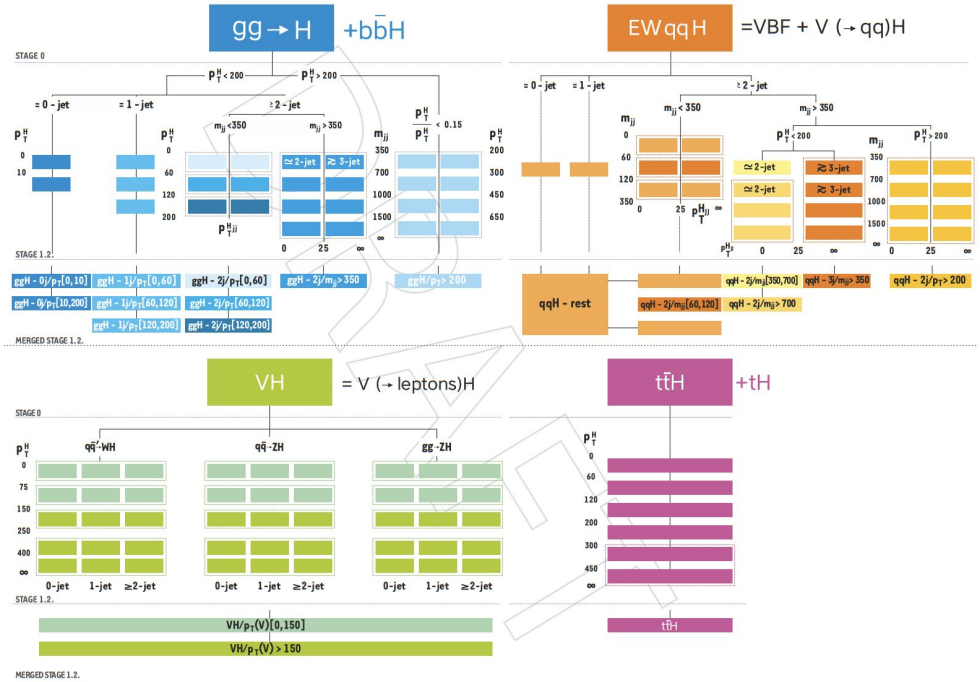
$$S(x) = \Sigma[(x - x_i)/\sigma_i]^2,$$

- This results in:

$$x_{comb} = \Sigma w_i x_i / \Sigma w_i, \quad w_i = 1/\sigma_i^2$$

$$1/\sigma^2_{comb} = \Sigma(1/\sigma_i^2)$$

Categories for H → ZZ → 4l



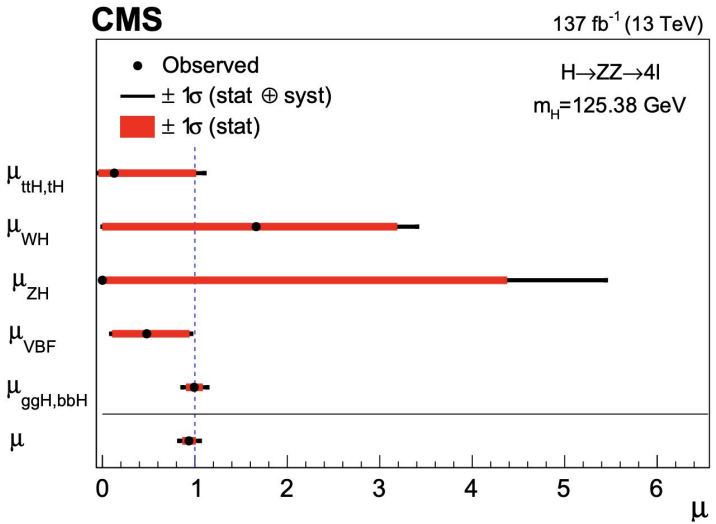


Figure 1: Binning of the gluon fusion production process, the electroweak production process (combines VBF and VH with hadronic V decay), the VH production process with leptonic V decay (combining WH, ZH, and gluon fusion ZH production), and the $t\bar{t}H$ production process in the merged stage 1.2 of the STXS framework [94] used in the analysis.

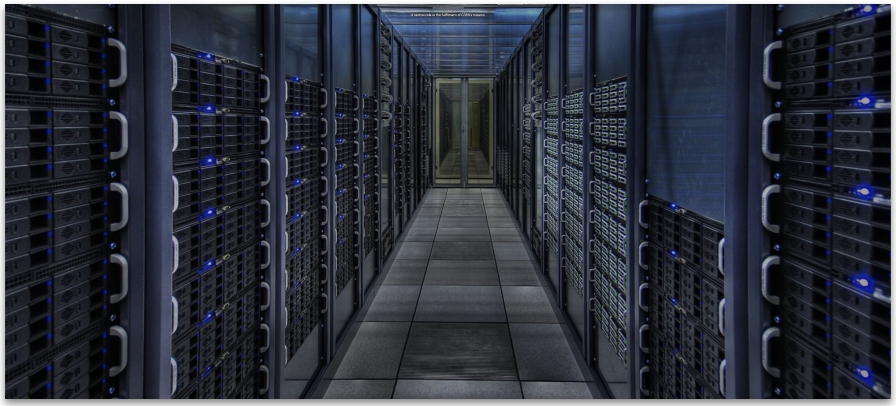
Practicalities and summary

Computing resources

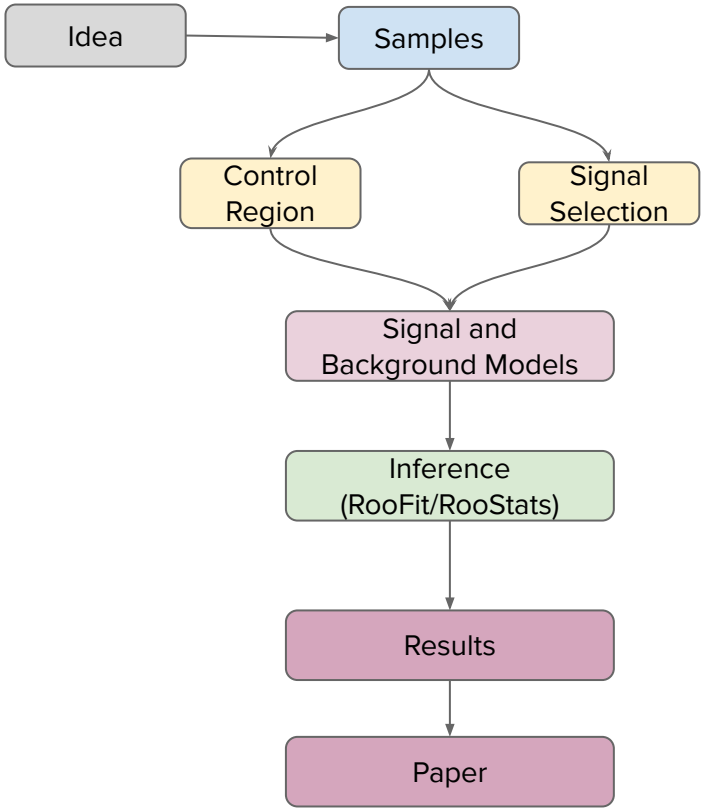
- Computing is "expensive".
- Real analysis demands a lot of data/computation. Your notebook is not enough.

- To think ahead:
  - Where is Data stored?
  - Where can I store my own samples (MC/ntuples).
  - What are the MC samples needed? Centrally produced?
- Throughout:
  - **Cloud (Google Colab/CERN's SWAN):** Good for prototyping. Won't scale.
  - **Multiprocessing:** Given enough storage, the easiest solution. Expensive, also.
  - **Cluster:** Shared resource, but with enough planning is productive.
  - **Grid:** Super-shared resource. Your analysis can be jammed in traffic. Has access to virtually any data.

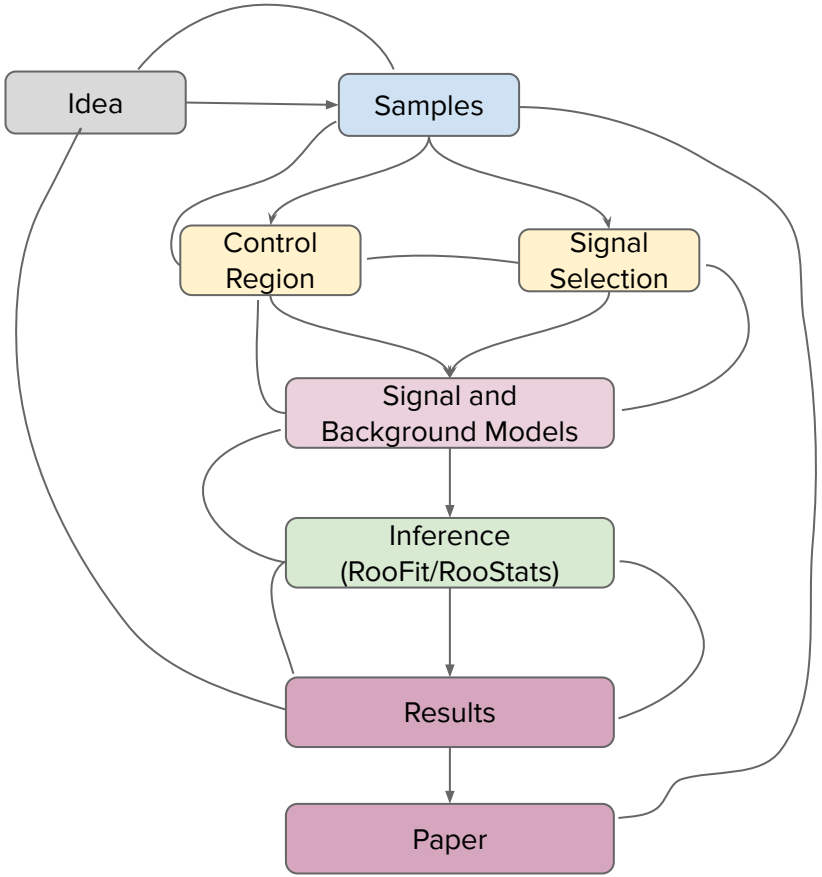
- Your analysis code should already foreseen your throughout from the beginning.
  - To rethink you analysis is easier than rewrite your code.



Ideally

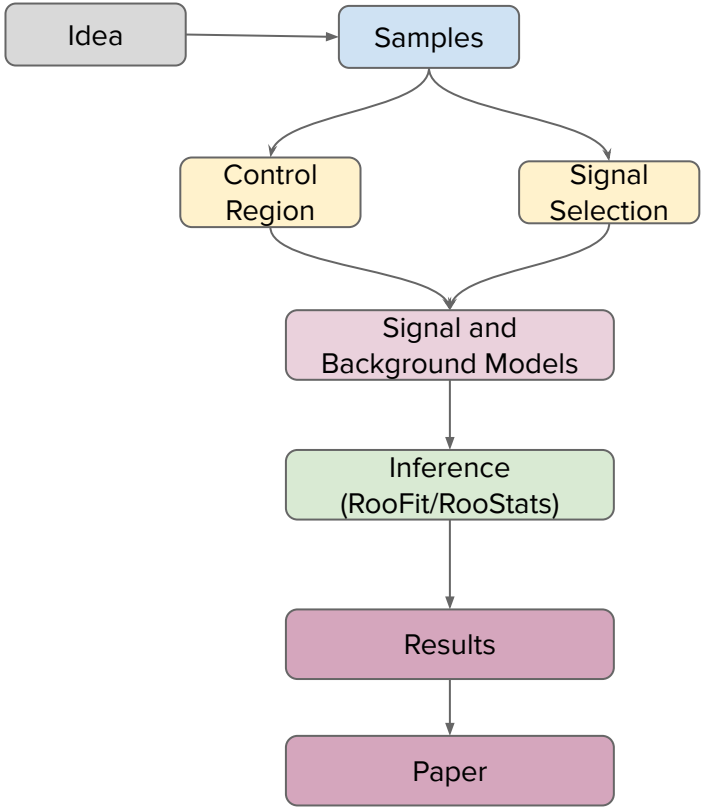

Idea
Samples
Control Region
Signal Selection
Signal and Background Models
Inference (RooFit/RooStats)
Results
Paper

Reality


Idea
Samples
Control Region
Signal Selection
Signal and Background Models
Inference (RooFit/RooStats)
Results
Paper

Summary

- The strategy for an analysis is a mix of common sense and personal opinions.
- **In the majority needs to agree.**
- **There is no magic formula.**
- Tip 1: Standing on the shoulders of giants.
  - Bigger analysis, previous/similar results should be the primary reference.
- Tip 2: Be ready to change your plans over, and over and over...



BACKUP SLIDES