

# **HEP Data Analysis Tutorial**

Roman Kogler (roman.kogler@desy.de) Christian Sander (christian.sander@desy.de) Alexander Schmidt (alexander.schmidt@desy.de)

Introduction to the Terascale – February 2013

# **Getting Started**

wget <u>http://www.desy.de/~csander/Files/HEPTutorial.tgz</u>

- tar -xvzf HEPTutorial.tgz
- cd HEPTutorial
- make
- ./example.x

UH

iİi

## **Motivation**

- Goal: This HEP Tutorial gives an introduction to ROOT based data analysis
- **Chosen example analysis:** *Top-Antitop cross section measurement* 
  - Comparison of background / signal MC
  - $\rightarrow$  Motivation of cuts
  - $\rightarrow$  Concepts of *purity* and *signal efficiency*
  - Application to data: simple cross section measurement
  - $\rightarrow$  Acceptance and trigger efficiencies

Based on publicly available HEP tutorial from CMS:

http://ippog.web.cern.ch/resources/2012/cms-hep-tutorial

# Outline



- Plot the invariant di-Muon Mass and determine the Z-mass
- Plot basic distributions of physic objects and compare data and simulation
- Find a selection which enriches the Top-Antitop events over the background from Standard Model processes. Do this on simulated events only! Never tune your signal selection on data!
- Obtain acceptance and trigger efficiencies from simulated events
- Measure the Top-Antitop cross section from data
- Evaluate a few systematic uncertainties
- Measure the Top-Mass

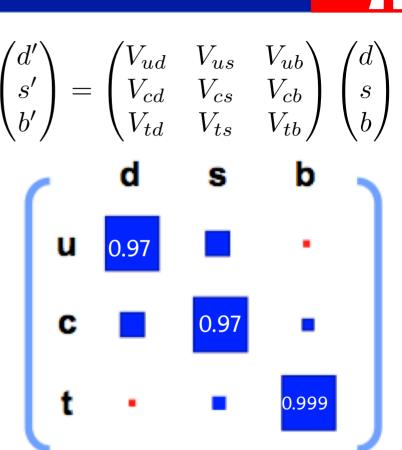
# The Top Quark

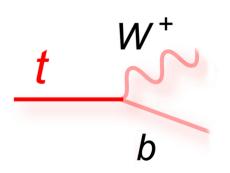
#### Why is the Top quark special?

- It is the heaviest SM particle!
- Coupling to Higgs field (Yukawa couplings) are ~
- Its live time is shorter than the characteristic hadronization time scale
  - Tops decay before fragmentation
  - Top quark decays carry information about spin correlations
- It decays exclusively in W+b

# Many searches for physics beyond the SM are connected to top physics:

- Searches for fourth generation quarks
- SUSY searches (important background, but also final states with tops)
- $Z' \rightarrow$  ttbar ...





# **Top Production at the LHC**

gg

qqbar

xf

0.8

0.6

0.4

0.2

 $10^{-4}$ 

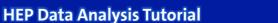
0000 0000  $\sigma_{tt}[pb]$ LHC (7 TeV) LHC (8 TeV) Tevatron 6.68+0.36+0.23 158.1+19.5+6.8 7.8+9.2 NLO 226 2+2**NNLO** 7.00 160.9 229  $0^+$ **NNLL** 162. 231

• LHC is a Top factory

**Tevatron and LHC** 

00000

gg and qqbar are complementary at



LHC (7 TeV)

~80%

~20%

xS (× 0.05)

**"Sea**"

 $10^{-3}$ 

 $\mathbf{x}^{1}$ 

UH

Tevatron (1.96 TeV)

~15%

~85%

exp. uncert. model uncert.

H1 and ZEUS HERA I+II PDF Fit

xg (× 0.05)

10-2

iİi

March 201

**HERAPDF Structure Function Working Group** 

 $Q^2 = 10000 \text{ GeV}^2$ 

xu,

HERAPDF1.5 NNLO (prel.)

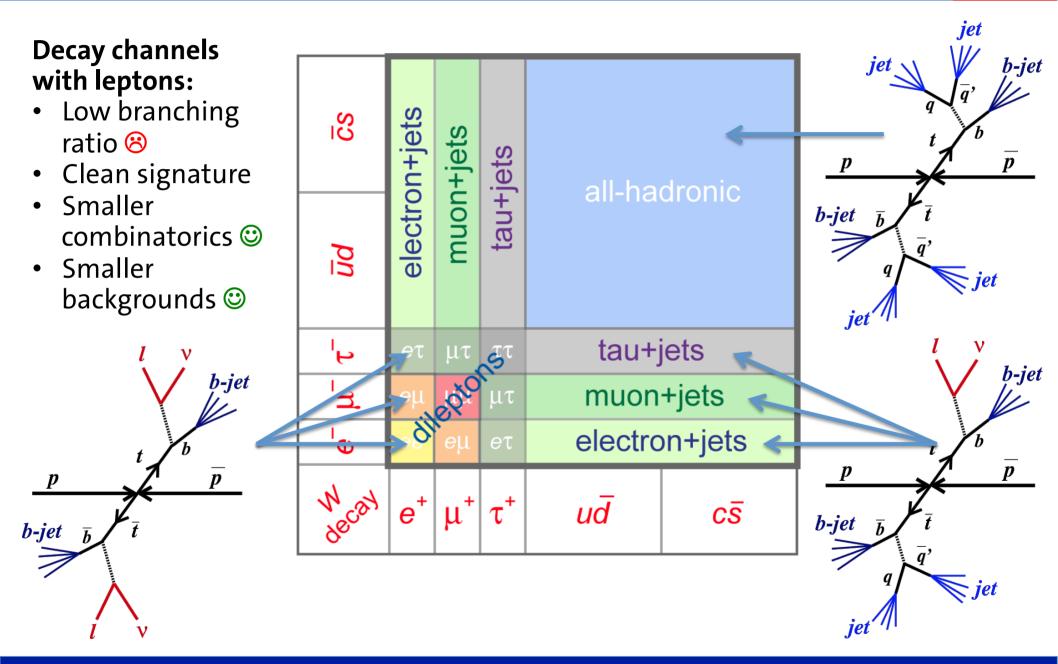
parametrization uncert.

"Valence"

xd,

**10**<sup>-1</sup>

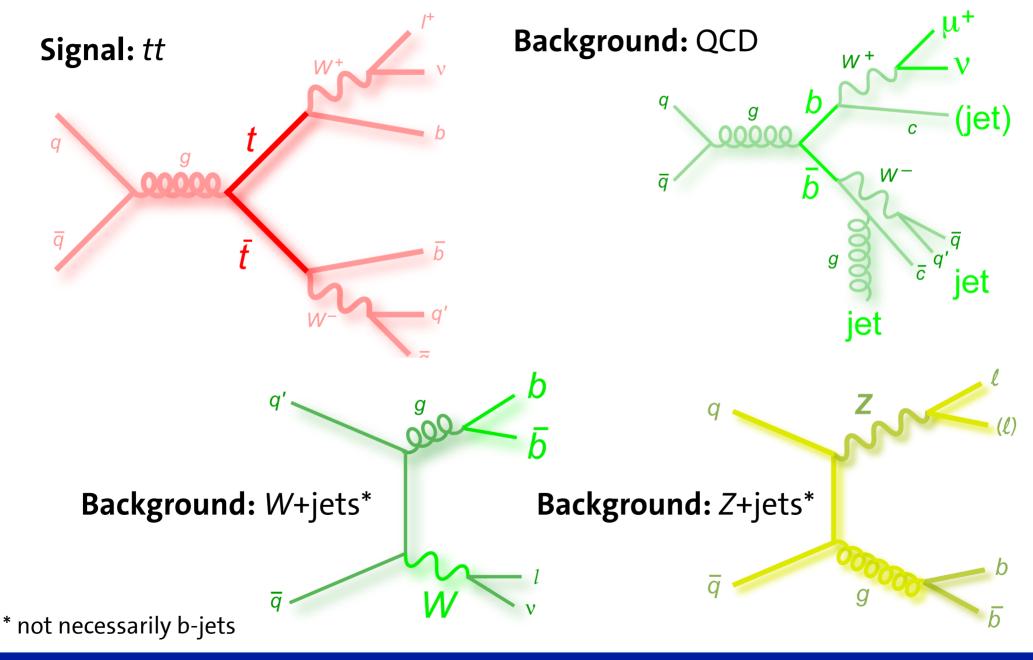
# **Decay Modes**



UH

iii

# **Signal and Background**



February 2013

UH

iii

# **b**-Jets

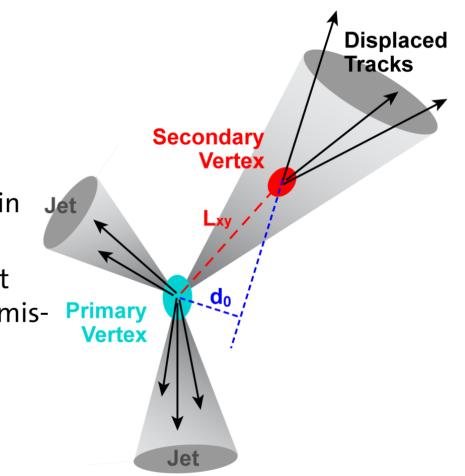


At hadron colliders (LHC) B-hadrons are produced inside of jets:

- Their lifetime (1.5ps) and
- The Lorentz boost
- → Displaced decay vertices

#### For example:

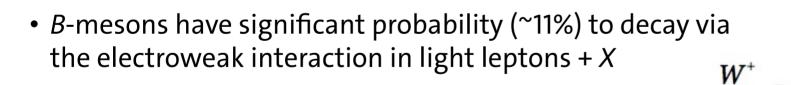
- Look for displaced tracks and vertices within jets (b-jet tagging)
- This tutorial: Track counting of high impact parameter tracks is used (efficiency ~50%, mis-Tag rate ~1%)



# Muons

#### Muons are very useful objects:

- Detected in the muon chambers which are "shielded" by thick absorbers (e.g. steel return yoke)
- $\rightarrow$  Other particles have negligible probability to reach this detector
- → Very clean object ID
- Processes with isolated muons are rare compared to QCD jet events
- Muon chambers can be used to trigger events (low  $p_T$  threshold compared to other objects) Iso<sub>µ</sub><sup>rel</sup> =  $\frac{p_T^{\mu}}{\sum_{\Delta B(i,\mu) < 0.3} p_T^i} (< 0.2)$



 $\rightarrow$  muons "inside" jets

ightarrow suppress by isolation

 $\rightarrow$  limited amount of additional activity around muon

B

1+

Ĉ

q

Jet

## Which, What, and How Much?

- Data: 50 pb<sup>-1</sup> (~1%) of CMS datasets
- Monte Carlo: Set of background processes are generated with full detector simulations

#### How is the information stored?

- Flat ROOT trees of only most fundamental object/event properties:
- → No CMSSW dependence, no reconstruction details accessible
- 4-vectors (px, py, pz, E) of leading objects
  - Jets: *p*<sub>T</sub> > 30 GeV; else: *p*<sub>T</sub> > 10 GeV
- +Isolation ( $\Delta R < 0.3$ ), +charge, +b-Tag, +Jet quality depending on object
- MC: +event weight, +IsoMuPt24 trigger bit (for tt only), +MC truth (parton level of semi-leptonic tt events)

UH

# **Available Samples**

| filename   | type                        | #events | x-section         | int. lumi.              | trig. only |
|------------|-----------------------------|---------|-------------------|-------------------------|------------|
| data.root  | data                        | 469384  |                   | $50 \ \mathrm{pb^{-1}}$ | yes        |
| ttbar.root | sim. $t\overline{t}$ signal | 36941   | 165  pb           | $50 \ { m pb}^{-1}$     | no         |
| wjets.root | sim. W plus jets background | 109737  | 31300  pb         | $50 \ \mathrm{pb^{-1}}$ | yes        |
| dy.root    | sim. Drell-Yan background   | 77729   | $15800~\rm{pb}$   | $50 \ { m pb}^{-1}$     | yes        |
| ww.root    | sim. WW background          | 4580    | 43  pb            | $50~{ m pb}^{-1}$       | yes        |
| wz.root    | sim. WZ background          | 3367    | 18  pb            | $50~{ m pb}^{-1}$       | yes        |
| zz.root    | sim. ZZ background          | 2421    | 6  pb             | $50~{ m pb}^{-1}$       | yes        |
| qcd.root   | sim. QCD multijet backgr.   | 142     | $10^8 \text{ pb}$ | $50 \text{ pb}^{-1}$    | yes        |

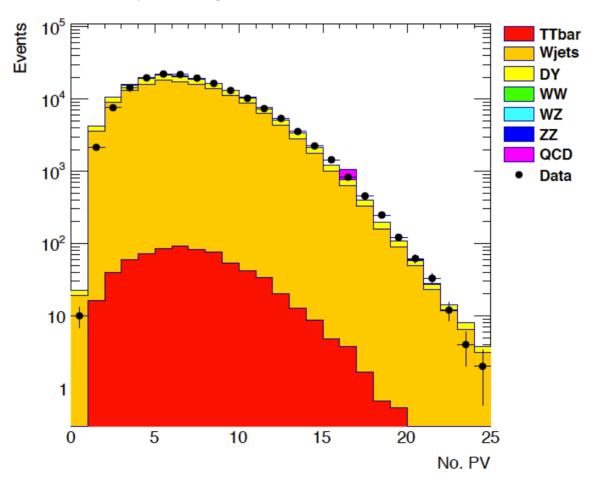
Table 1: Data and simulated Monte Carlo samples.

UH

iii

# **Pile-Up**

- Very high luminosity at LHC  $\rightarrow$  overlapping interactions (>50 for some events)
- Some physics objects depend on the "pile up", e.g. isolation of leptons
- MC generated with some PU distributions → reweight events to match distribution of number of primary vertices in data and simulation



UH

# **Cross Section Measurement – Strategy**

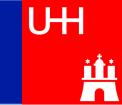
- 1) Selection of signal events (here: *tt*) to ensure
  - High trigger efficiency: require one isolated muon! Details later.
  - High signal acceptance = #(selected signal events)/#(all signal events) → low statistical uncertainties
  - High purity = #(selected signal events)/#(all selected events) → small uncertainty from unknown backgrounds
- 2) From simulation: acceptance, purity and, trigger efficiency
- 3) Count selected data events N<sub>data</sub>
- 4) Subtract expected background N<sub>background</sub> = N<sub>data</sub> (1 purity)
- 5) Correct for acceptance and trigger efficiency:

 $N_{\text{signal,corr}} = (N_{\text{data}} - N_{\text{background}}) / (\text{acceptance} \cdot \text{trigger efficiency})$ 

6) Cross section  $\sigma_{signal} = N_{signal,corr} / Luminosity_{integrated}$ 

UH

# **Trigger Efficiency**



- Event rate of a process with cross section  $\sigma \rightarrow f = \sigma \cdot L_{instantaneous}$
- With  $L_{\rm instantaneous} = 10^{34} \, {\rm s}^{-1}$  and  $\sigma_{\rm total} = 10^8 \, {\rm pb} \rightarrow f = 10^6 \, {\rm s}^{-1}$
- Typical recorded rate ~100 s<sup>-1</sup>  $\rightarrow$  online preselection (trigger)
- For the presented example: trigger on isolated muon with  $p_{T}$  > 24 GeV
- Online-offline differences→ Determine trigger efficiency (e.g. from MC)

#### Trigger efficiency = #(triggered and selected)/#(selected)

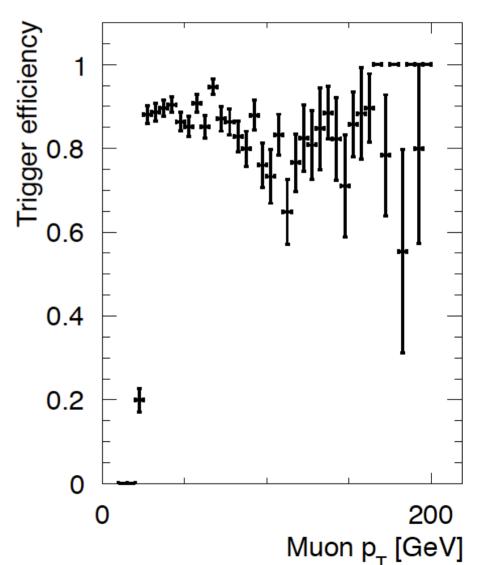
- Selection not necessarily signal selection (as long as independent)  $\rightarrow$  e.g. isolated muon in  $p_{T}$  interval  $\rightarrow$  "turn on" curve
- To evaluate statistical uncertainties correctly:
  - Don't use error propagation for ratio of two quantities; use binomial errors. **Better:** confidence intervals (otherwise error for efficiency of 1 is zero by definition)

```
TGraphAsymmErrors* g_eff = new TGraphAsymmErrors (h_SelTrig, h_Sel);
g_eff->GetXaxis()->SetTitle("Muon p_{T} [GeV]");
g_eff->GetYaxis()->SetTitle("Trigger efficiency");
g_eff->UseCurrentStyle();
g_eff->Draw("AP");
c2->Print("Trigger_eff.pdf");
```

# Trigger "Turn On"

- Trigger efficiency in bins of muon  $p_{\rm T}$
- Events with two muons lead to too large trigger efficiency
- → Correct: trigger matching (online to offline object)
- → Approximation: require events with exactly one muon
- Turn on curve reaches plateau short after passing  $p_{\rm T}$  threshold
- Efficiency not 1.0 at high  $p_{\rm T}$
- → Reconstruction (ID) or isolation inefficiencies on trigger level

**Example:** Trigger turn on for isolated muons (iso/ $p_T$ <0.05) as a function of  $p_T$ 



UH



#### 1 Warmup

The trigger for this tutorial selects events which contain one or more muons as discussed in the documentation and explanation.

- Find out how often there is more than one isolated, reconstructed muon in data (histogram of the muon multiplicity)! Where could these additional muons come from?
- Calculate the invariant mass of two muons of opposite charge (manually and/or using the ROOT functionality of adding two fourvectors)! Only use isolated muons.
- Display the invariant mass distribution of two muons in a histogram (hint: try different axis ranges)!
- Compare your results to MC simulation (display simulation and data in the same histogram). Make sure you select triggered events only for the simulated samples!

# **Files to Edit**

The only files to be modified by you:

- MyAnalysis.h
  - Define the histograms, e.g. **TH1F** \*h\_myVariable;
  - Define any auxiliary variable, e.g. event counter like int TotalEvents
- MyAnalysis.C
  - Book the historgrams:

```
h_myVariable = new TH1F(" myVariable", "myVariable", 100, 0, 10);
```

h myVariable->SetXTitle("myVariable");

h\_myVariable->Sumw2();

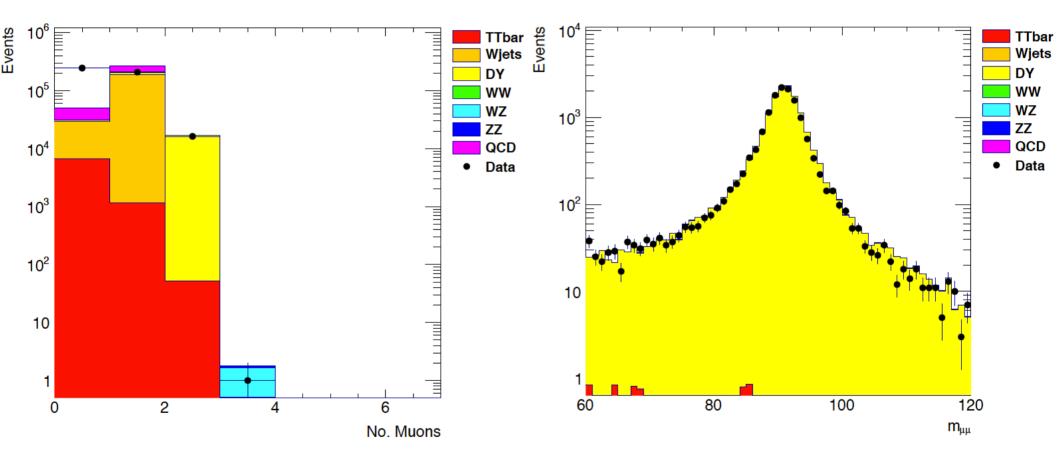
histograms.push\_back(h\_myVariable);

histograms\_MC.push\_back(h\_myVariable);

• Fill the histograms in Bool\_t MyAnalysis::Process (Long64\_t entry)

#### • Example.C

- Here the plotting is done. You have to edit only ...
  - when plotting the trigger efficiency
  - when calculating the cross section



UΗ

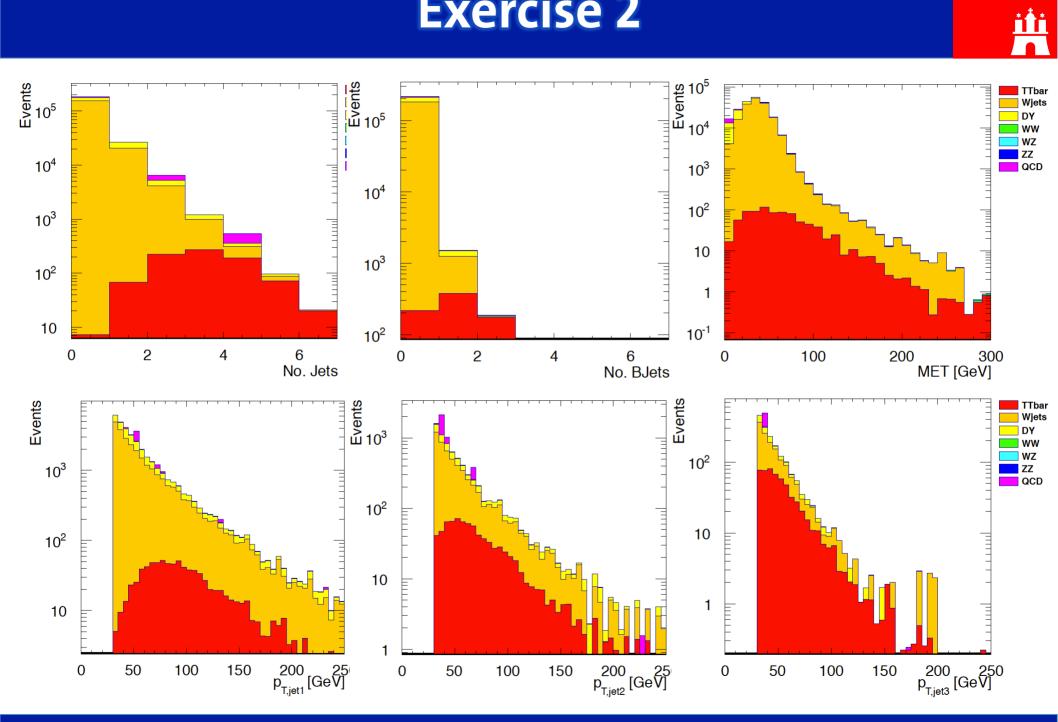
iii



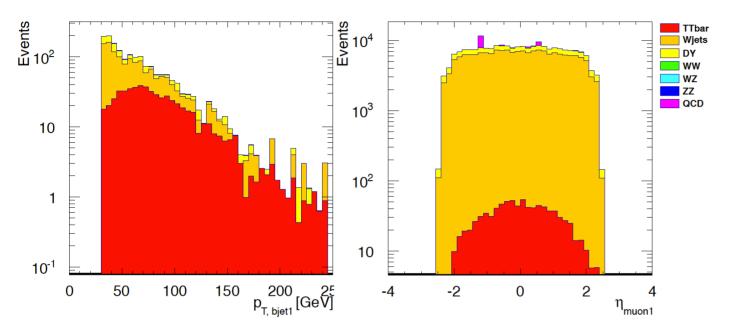
#### 2 Properties of top quark events

In this exercise we take the first steps towards a real measurement using top quark events. We need to understand how we can efficiently select top quark events and reject events without top quarks (background rejection) at the same time.

- Starting from the requirement of at least one isolated muon, compare several other distributions of event variables for simulated signal ( $t\bar{t}$  events) and background.
- Try to find variables which are especially sensitive to separate signal from background (jet multiplicity, transverse momenta of jets and leptons, lepton isolation, b-tagging, missing transverse energy, angular distributions). Fill all these distributions into histograms and compare between signal, background and data.
- Apply cuts on these variables to enrich the signal over background. Try to optimize the signal over background ratio and estimate the purity that can be achieved (based on simulation only).
- Apply your selection cuts also on data. Compare the selection efficiency between data and simulation.



υH





#### 3 Cross-section of top quark production

In this exercise we will calculate the cross-section of top quark pair production at the LHC. The necessary ingredients are developed step by step.

- The first ingredient is the trigger efficiency  $\epsilon^{\text{trig}}$ . We can trust the MC simulation to reproduce this efficiency correctly. Produce the trigger "turn-on" curve which shows the trigger efficiency depending on the muon transverse momentum  $p_{\text{T}}$ . Calculate the efficiency of triggering top quark events with a reconstructed and isolated muon of  $p_{\text{T}} > 25$  GeV?
- The second ingredient is the acceptance  $\epsilon^{acc}$  (not including the trigger). This includes the fact that we only select semi-leptonic top quark decays with muons. The branching fraction is well known, so we can take it from simulation. In addition, the acceptance includes all the selection cuts that have been found in Exercise 2. You can calculate the acceptance by comparing the number of generated top quark events with the number of selected events, after all your cuts.
- background subtraction: we also trust the simulation to correctly predict the number of background events after selection. Subtract the expected background from the observed (selected) data events.
- You can calculate the cross section now using the purity corrected observed events in data  $N_{\text{data}}^{\text{obs}} * \text{purity}$ . You have to apply corrections for trigger efficiency  $\epsilon^{\text{trig}}$  and acceptance  $\epsilon^{\text{acc}}$ .
- Compare your result with official publications of the ATLAS and CMS Collaborations.

UH

Monte Carlo:

Signal Number of generated events: 7928.61

Signal Number of selected events: 253.915

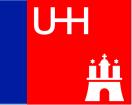
Signal Number of selected and triggered events: 214.014

Signal Acceptance: 0.0320252 Trigger efficiency:0.842856

Background Number of selected and triggered events:10.3363 Purity:0.953928

Data:

Number of selected and triggered events: 231 Number of selected and triggered events (bg subtracted): 220.357 Acceptance and trigg efficiency corrected yield: 8163.62 Cross section [pb]: 163.272

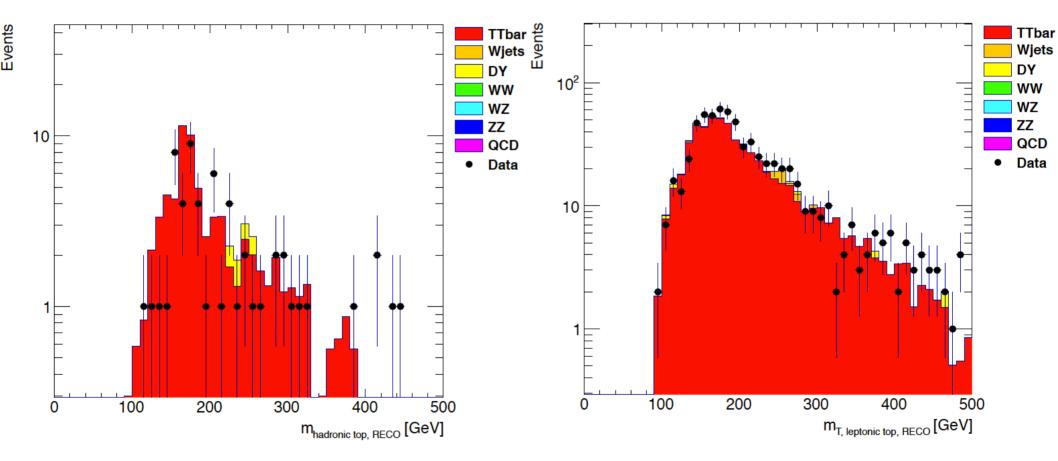


#### 4 Top quark mass reconstruction

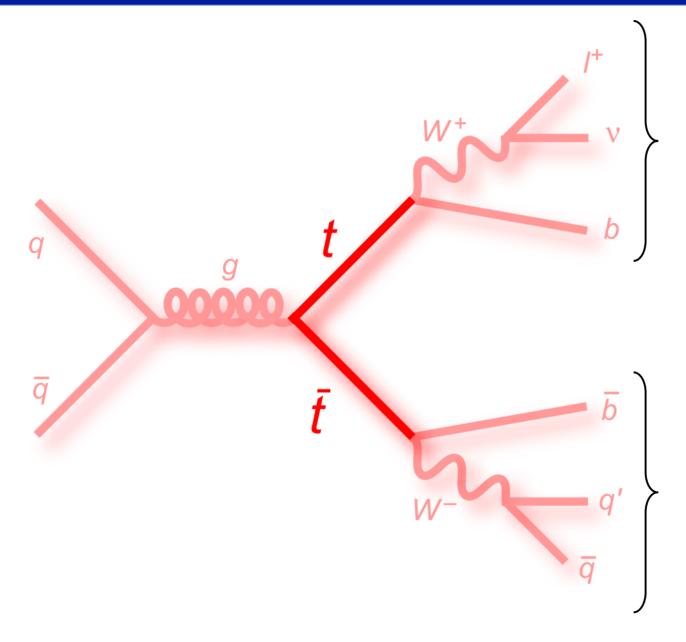
In this exercise we will reconstruct the fourvectors of the top quarks by assigning the detector objects (jets, leptons, missing energy) to the hypothetical  $t\bar{t}$  decay tree. As we only consider semi-leptonic decays with muons in the final state, we expect four jets, one muon plus missing energy in the final state. Two of the four jets are b-jets (b-tagged).

- What is the mass of the top quark in MC simulation (in  $t\overline{t}$  events)? Use the generator-level truth information to calculate the top quark fourvector in the hadronic and leptonic branch.
- As a next step try to use detector objects only. Find out which (not b-tagged) jets come from the hadronic W boson decay using the W boson mass.
- Combine this W boson with a b-jet. As there are two b-jets, simply use both solutions, and fill the reconstructed top quark mass in histograms, comparing data to simulation.
- Reconstruct the top quark from the leptonic branch as well. The z-component of the neutrino is not measured, as we only have transverse missing energy. You can calculate the z-component using a W mass constraint (two solutions).





#### **Top Mass Measurement**



#### • Leptonic top decay:

 Invariant mass of (*b*-jet, muon, and neutrino)

UΗ

ΪÌ

- Unknown p<sub>z</sub> from neutrino
- Combinatorics: two bjets

#### Hadronic decay:

- Invariant mass (*b*-jet and two non-*b*-jets)
- All momentum components known
- But:
  - Worse resolution
  - Larger combinatorics

## **Mass Determination – Options**

- Hadronic Mass:
  - All combinations (*b* + two non-*b*)
  - Invariant mass of two non-b jet in W-mass window (e.g. 60 ... 100 GeV)
- Leptonic Mass:
  - Assume  $p_z$  of neutrino to be 0
    - All combinations (both *b*-jets)
    - Optional: use angular variables to enrich right combination
  - Calculate:  $p_z$  of neutrino from invariant W-mass constraint
    - ( (  $E, p_x, p_y, p_z$ )<sub>muon</sub> + (  $E, p_x, p_y, p_z$ )<sub>neutrino</sub>)<sup>2</sup> = ( $M_W$ )<sup>2</sup>
    - With  $m_{\text{neutrino}} = 0 \rightarrow E^2 = p_x^2 + p_y^2 + p_z^2 \rightarrow \text{quadratic equation in } p_z$  (with 0, 1, or 2 solutions)
- Use combinations which yield smallest difference of leptonic and hadronic mass

UHH

# Backup

# **Stored Information – Jets**

- NJet (integer): number of jets in the event.
- Jet\_Px[NJet] (float): x-component of jet momentum. This is an array of size NJet, where a maximum of twenty jets are stored (NJet < 21). If there are more than twenty jets in the event, only the twenty most energetic are stored. Only jets with  $p_{\rm T} > 30$  GeV are stored.
- Jet\_Py[NJet] (float): y-component of jet momentum, otherwise same as Jet\_Px[NJet].
- Jet\_Pz[NJet] (float): z-component of jet momentum, otherwise same as Jet\_Px[NJet].
- Jet\_E[NJet] (float): energy of the jet, otherwise same as Jet\_Px[NJet]. Note that the four components Jet\_Px, Jet\_Py, Jet\_Pz and Jet\_E constitute a fourvector which fully describes the kinematics of a jet.
- Jet\_btag[NJet] (float): b-tagging discriminator. This quantity is obtained from an algorithm that identifies B-hadron decays within a jet. It is correlated with the lifetime of the B-hadron. Higher values indicate a higher probability that the jet originates from a b-quark. The discriminator has small performance differences in data and simulation. To account for this, simulated events have to be reweighted by a factor of ~ 0.9 per required b-tagged quark.
- Jet\_ID[NJet] (bool): Jet quality identifier to distiguish between real jets (induced by hadronic interactions) and detector noise. A good jet has true as value.

UH

# **Stored Information – Muons**

- NMuon (integer): number of muons in the event.
- Muon\_Px[NMuon] (float): x-component of muon momentum. This is an array of size NMuon, where a maximum of five muons are stored (NMuon < 5). If there are more than five muons in the event, only the five most energetic are stored.
- Muon\_Py[NMuon] (float): y-component of muon momentum, otherwise same as Muon\_Px[NMuon].
- Muon\_Pz[NMuon] (float): z-component of muon momentum, otherwise same as Muon\_Px[NMuon].
- Muon\_E[NMuon] (float): energy of the muon, otherwise same as Muon\_Px[NMuon]. Note that the four components Muon\_Px, Muon\_Py, Muon\_Pz and Muon\_E constitute a fourvector which fully describes the kinematics of a muon.
- Muon\_Charge[NMuon] (integer): charge of the muon. It is determined from the curvature in the magnetic field and has values +1 or -1.
- Muon\_Iso[NMuon] (float): muon isolation. This variable is a measure for the amount of detector activity around that muon. Muons within jets are accompanied by close-by tracks and deposits in the calorimeters, leading to a large values of Muon\_Iso. On the other hand, muons from W bosons are isolated and have small values of Muon\_Iso.

UH

## **Stored Information – Electrons / Photons**

- NElectron (integer): same as for muons above, but for electrons.
- Electron\_Px[NElectron] (float): same as for muons above, but for electrons.
- Electron\_Py[NElectron] (float): same as for muons above, but for electrons.
- Electron\_Pz[NElectron] (float): same as for muons above, but for electrons.
- Electron\_E[NElectron] (float): same as for muons above, but for electrons.
- Electron\_Charge[NElectron] (integer): same as for muons above, but for electrons.
- Electron\_Iso[NElectron] (float): same as for muons above, but for electrons.
- NPhoton (integer): same as for muons above, but for photons.
- Phtoton\_Px[NPhoton] (float): same as for muons above, but for photons.
- Photon\_Py[NPhoton] (float): same as for muons above, but for photons.
- Photon\_Pz[NPhoton] (float): same as for muons above, but for photons.
- Photon\_E[NPhoton] (float): same as for muons above, but for photons.
- Photon\_Iso[NPhoton] (float): same as for muons above, but for photons.

UH

iii

- MET\_px (float): x-component of the missing energy. Due to the hermetic coverage of the LHC detectors and the negligible transverse boost of the initial state, the transverse momentum sum of all detector objects (jets, muons, etc...) must be zero. This is required by energy and momentum conservation. Objects which escape the detector, such as neutrinos, are causing a "missing" transverse energy which can be measured and associated to the neutrino.
- MET\_py (float): y-component of the missing energy.
- NPrimaryVertices (integer): the number of proton-proton interaction vertices. Due to the high LHC luminosity several protons within one bunch crossing can collide. This is usually referred to as "pileup". The spread of these vertices is several centimeters in longitudinal direction and only micrometers in the transverse direction.
- triggerIsoMu24 (bool): the trigger bit. It is "true" if the event is triggerd and "false" if the event is not triggered (data can only contain triggered events).
- EventWeight (float): weight factor to be applied to simulated events due to different sample sizes.

Ш

# **Further Stored Information**

Ш

- MChadronicBottom\_px (float): x-component of the b-quark from the top decay belonging to the hadronic branch.
- MChadronicBottom\_py (float): y-component ...
- MChadronicBottom\_pz (float): z-compoment ...
- MChadronicWDecayQuark\_px (float): x-component of the quark from the hadronic W boson decay
- MChadronicWDecayQuark\_py (float): y-component ...
- MChadronicWDecayQuark\_pz (float): z-component ...
- MChadronicWDecayQuarkBar\_px (float): x-component of the anti-quark from the hadronic W boson decay
- MChadronicWDecayQuarkBar\_py (float): y-component ...
- MChadronicWDecayQuarkBar\_pz (float): z-component ...

#### + the analogous information for the leptonic decay branch

#### **Neutrino Momentum**



$$M_W^2 = 2p_\ell p_\nu = 2 \cdot (E_\ell E_\nu - \vec{p}_\ell \cdot \vec{p}_\nu) = 2 \cdot (E_\ell E_\nu - \vec{p}_{T,\ell} \cdot \vec{p}_{T,\nu}) = 2 \cdot (E_\ell E_\nu - p_{T,\ell} p_{T,\nu} \cos \Delta \phi - p_{z,\ell} p_{z,\nu}) .$$
(A.3)

By introducing the abbreviation  $\mu = \frac{M_W^2}{2} + p_{\mathrm{T},\ell} p_{\mathrm{T},\nu} \cos \Delta \phi$ , this can be further simplified to

$$E_{\ell}E_{\nu} = \mu + p_{z,\ell}p_{z,\nu}$$
 (A.4)

The energy  $E_{\nu}$  of the massless neutrino can be expressed in terms of its momentum components

$$E_{\ell} \sqrt{p_{\mathrm{T},\nu}^2 + p_{z,\nu}^2} = \mu + p_{z,\ell} p_{z,\nu} . \qquad (A.5)$$

After squaring the equation and rearranging its terms, a quadratic equation in  $p_{z,\nu}$  is obtained

$$p_{z,\nu}^2 - 2 \cdot \frac{\mu p_{z,\ell}}{E_\ell^2 - p_{z,\ell}^2} \cdot p_{z,\nu} + \frac{E_\ell^2 p_{\mathrm{T},\nu}^2 - \mu^2}{E_\ell^2 - p_{z,\ell}^2} = 0 .$$
 (A.6)

#### **Neutrino Momentum**

$$p_{z,\nu}^2 - 2 \cdot \frac{\mu p_{z,\ell}}{E_\ell^2 - p_{z,\ell}^2} \cdot p_{z,\nu} + \frac{E_\ell^2 p_{\mathrm{T},\nu}^2 - \mu^2}{E_\ell^2 - p_{z,\ell}^2} = 0 , \qquad (4.4)$$

with 
$$\mu = \frac{M_W^2}{2} + p_{\mathrm{T},\ell} p_{\mathrm{T},\nu} \cos \Delta \phi$$
 . (4.5)

 $E_{\ell}$  and  $p_{z,\ell}$  denote the energy and the z component of the momentum of the charged lepton, respectively. The azimuthal angle difference between the charged lepton and  $\vec{E}_{T}$  is given by  $\Delta \phi$ . This equation is solved by

$$p_{z,\nu}^{\pm} = \frac{\mu p_{z,\ell}}{p_{\mathrm{T},\ell}} \pm \sqrt{\frac{\mu^2 p_{z,\ell}^2}{p_{\mathrm{T},\ell}^4} - \frac{E_\ell^2 p_{\mathrm{T},\nu}^2 - \mu^2}{p_{\mathrm{T},\ell}^2}} , \qquad (4.6)$$

UH