Validation of 2011 open data using measurements of top quark pair production

Oleksandr Zenaiev
(DESY)

Overview:

- Measurements of $t\bar{t}$ production in dilepton channel at 7 TeV [TOP-11-013, TOP-13-004] *code at Github*

Not covered in this talk:

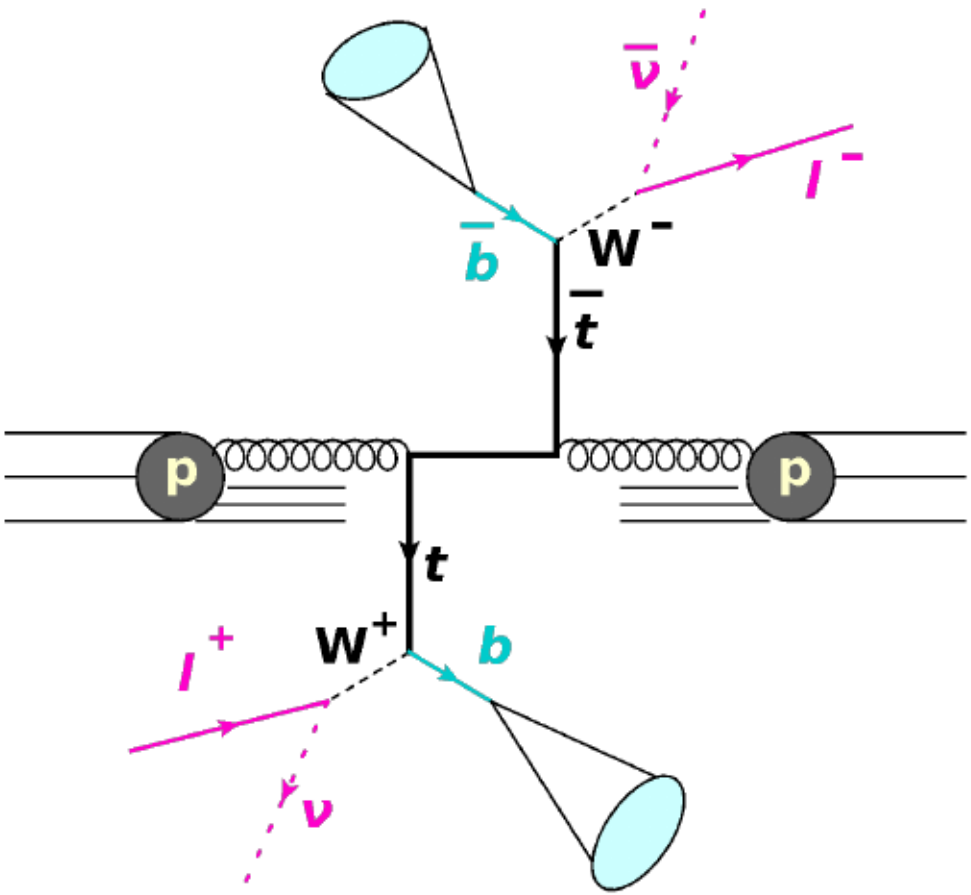
- Measurement of associated $W + c$ production [SMP-12-002]
  CERN Open Data meeting + DESY summer student O. Kot
- Jet tuple production applying jet energy corrections
  code at Github

CMS DESY Top Meeting
13.12.2016

Introduction

- Using 2011 open data, 2.5fb^{-1} + corresponding MC samples `http://opendata.cern.ch/about/cms`
- **Goal: reproduce published CMS results ('research' mode)**
- Running VM on private or office laptop, no usage of CMS, CERN and other resources
- Using information mainly from CMS papers, theses, public twiki pages, *sometimes needed to consult analysis notes*
- All analysis code has been written completely from scratch: no usage of CMS code, except for CMSSW provided with VM

Measurement of differential top-quark pair production cross sections in pp colisions at $\sqrt{s} = 7$ TeV



Dilepton channel: 3 possible final states:

- $e^{\pm}e^{\mp}$ + 2 b-jets + 2 neutrinos
- $\mu^{\pm}\mu^{\mp}$ + 2 b-jets + 2 neutrinos
- $e^{\pm}\mu^{\mp}$ + 2 b-jets + 2 neutrinos

$\Rightarrow$ selecting events with two leptons, two jets, also expect missing transverse energy (MET)

t and $\bar{t}$ are then reconstructed from measured final state particles and assumptions on m_W, m_t

Data samples:

- $e^{\pm}e^{\mp}$: '`DoubleElectron`'
- $\mu^{\pm}\mu^{\mp}$: '`DoubleMu`'
- $e^{\pm}\mu^{\mp}$: '`MuEG`'

Primary vertex:

- dof > 4
- impact parameter < 2 cm in transverse plane and < 24 cm in z

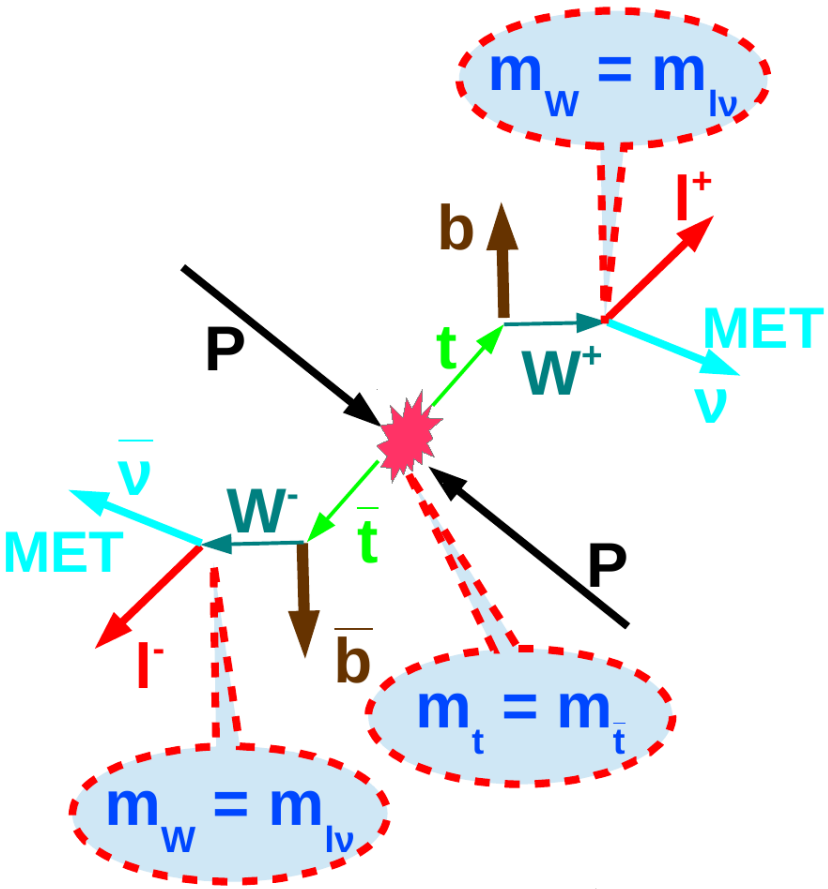
Leptons: two leading p_T opposite signed $p_T > 20$ GeV, $|\eta| < 2.4$

- Electrons ('`gsfElectrons`')
  - isolated with $I_{\text{rel}}^{\Delta R<0.3} < 0.17$ ($\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$)
  - no missing hits in the silicon tracker
- Muons ('`muons`')
  - isolated with $I_{\text{rel}}^{\Delta R<0.3} < 0.20$
  - required to be global muons
  - at least 10 valid tracker hits and 2 pixel hits
  - global track fit χ^2/dof < 10
  - impact parameter to PV $<$ 0.02 cm in transverse plane and $<$ 0.5 cm in z
- $M(ll) > 12$ GeV
- in ee and $\mu\mu$ channels exclude $16 < M(ll) < 106$ GeV
- in ee and $\mu\mu$ channels require MET > 30 GeV

Jets ('`ak5PFJets`'): at least two with $p_T > 30$ GeV, $|\eta| < 2.4$

- anti-k_T with clustering parameter 0.5
- Jet energy correction: '`ak5PFL1FastL2L3Residual`'
- at least one b-tagged using CSVL (discriminant $>$ 0.244)
- CSV information stored only for '`ak5CaloJets`': perform matching choosing closest jet in ΔR

Goal: obtain $\vec{p_t}$ and $\vec{p_{\bar{t}}}$



Efficiency determined in signal MC:
$\approx 70\%$ vs 90% in the paper

- Measured input: 2 leptons, 2 jets, MET
- Unknowns: $\bar{p}_\nu$, $\bar{p}_{\bar{\nu}}$ (6)
- Constraints:
  - m_t, $m_{\bar{t}}$ (2)
  - m_{W+}, m_{W-} (2)
  - $(\bar{p}_\nu + \bar{p}_{\bar{\nu}})_T = \mathrm{MET}$ (2)
- For each pair of jets, solve this using the method from [Phys. Rev. D 73 (2006) 054015]
- If there are several solutions in event (either because of many jet combinations, or several solutions for one configuration), prefer:
  - with 2 b-tagged jets
  - with 1 b-tagged jets
  - with highest weight, weight is determined according to the MC neutrino energy spectrum
- Difference from the paper: no m_t scan $\Rightarrow$ worse efficiency due to detector effects

[DESY-THESIS-2012-037]

MC samples

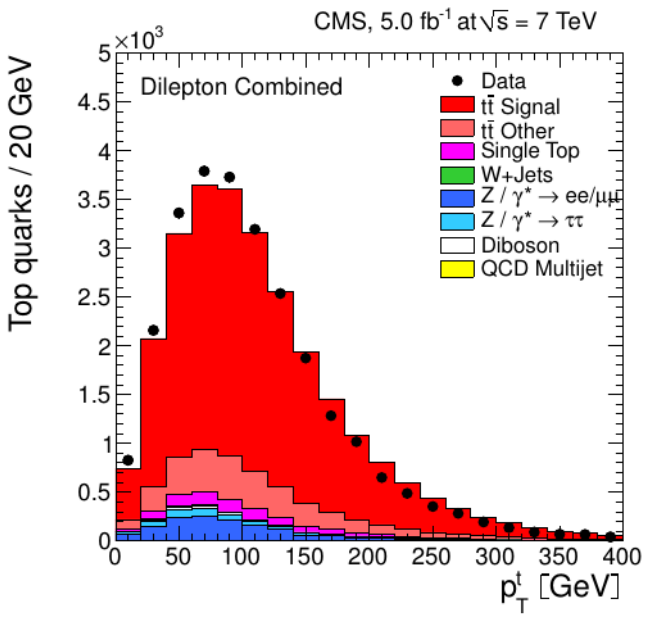
Signal:

- MadGraph + Pythia6, 55M

Background:

- $t\bar{t}$ 'other', mainly via τ decays (MadGraph + Pythia6)
- single top (POWHEG + Pythia6), 1.5M
- Drell-Yan (DY) (MadGraph + Pythia6), 44M
- W + jets (MadGraph + Pythia6), 55M
- Diboson and QCD multijet considered in the paper, but contribute negligibly: not used

TOP-11-013



In total processed:

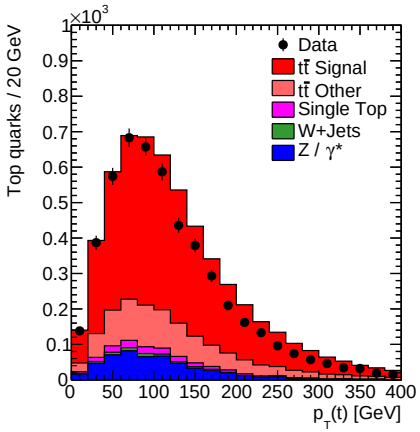
- Data: 33M ($e\mu$) + 50M (ee) + 40M ($\mu\mu$) $\approx$ 123M
- MC: $\approx$ 155M (processing MC took $\times 5$ more time than data: busy events, larger fraction selected)

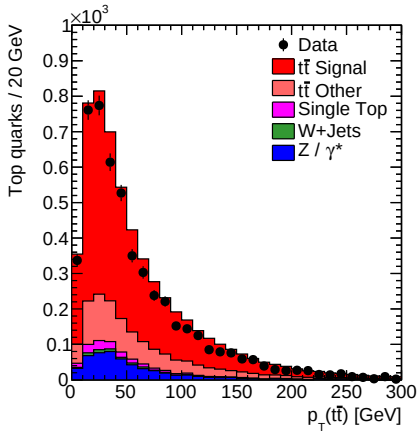
Overall $\sim$ 2 weeks (not CPU time!), running several jobs in parallel on one machine, but also gaps between running jobs. Some jobs needed to be resubmitted. Bottleneck: data network access (latent server responce?).

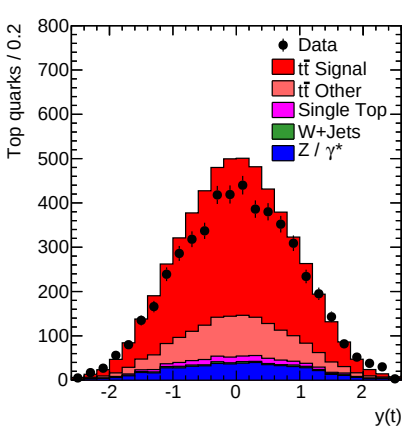
Any improvement here is very desirable: how will it be feasible for more complicated analyses, or 2012 data?..

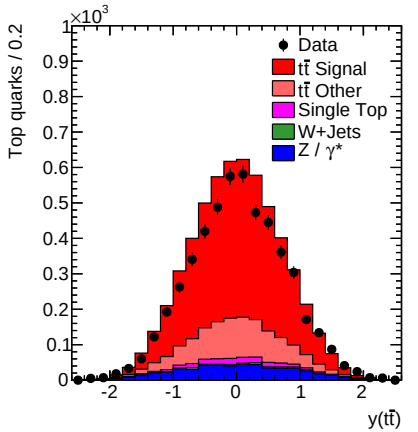
Control distributions

Open Data

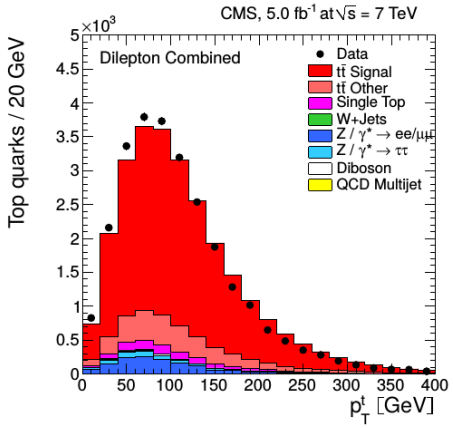


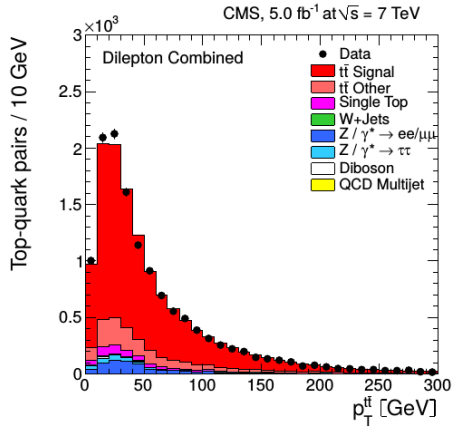


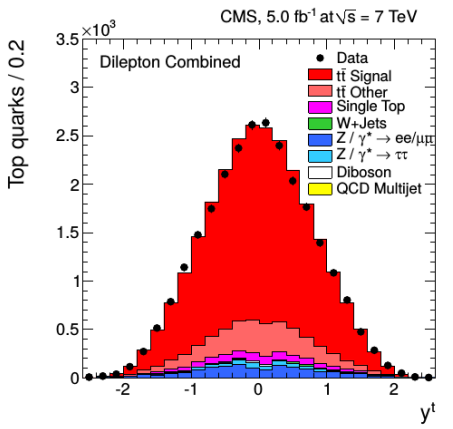


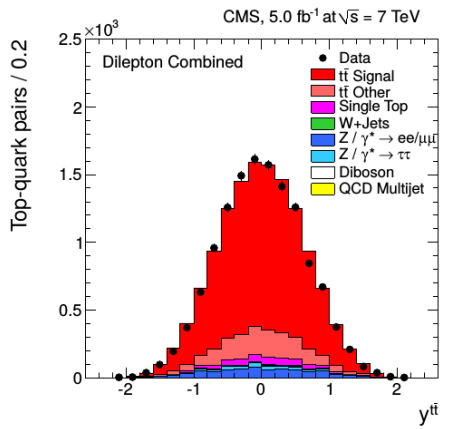


TOP-11-013









- $\approx$ 25% less events than expected from TOP-11-013: consistent with smaller kin. reco efficiency
- Larger MC / data: consistent with missing scale factors, corrections etc.
- Slightly larger background fraction
- Shapes very similar

* $p_T(t)$, $y(t)$ vs $p_T(t), p_T(\bar{t})$ and $y(t), y(\bar{t})$ in paper

Cross section measured at parton level in full phase space:

$$\frac{\mathrm{d}\sigma}{\mathrm{d}Y} = \frac{N_{Sig}}{ALB\Delta Y},\ N_{Sig} = N_{DATA} - N_{MCbackgr},\ E = \frac{N_{MCreco}}{N_{MCgen}}$$

$$\sigma = \int \frac{\mathrm{d}\sigma}{\mathrm{d}Y}$$

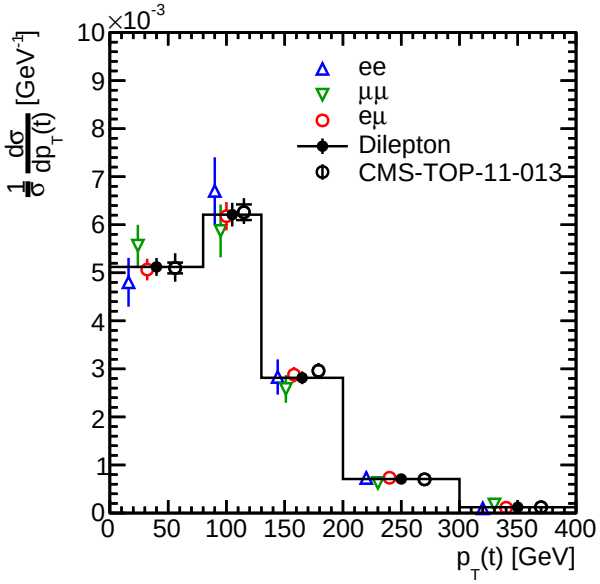
$L = 2.5\ \mathrm{fb}^{-1},\ B = 4.6\%$

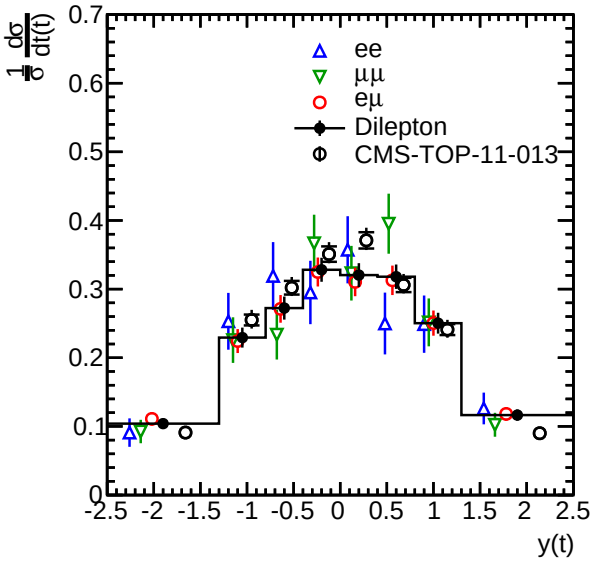
Efficiency E determined as bin-to-bin corrections: no 'unfolding'.
This should give underestimated stat. uncertainties and might bias central values in bins with small purity/stability and poor MC description.

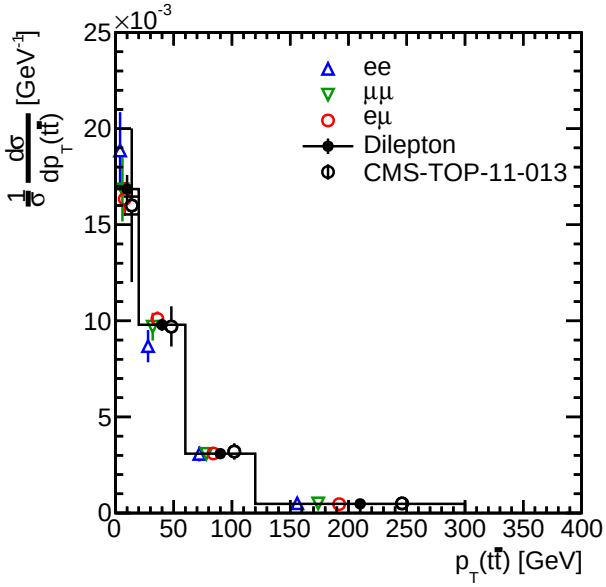
Measure:

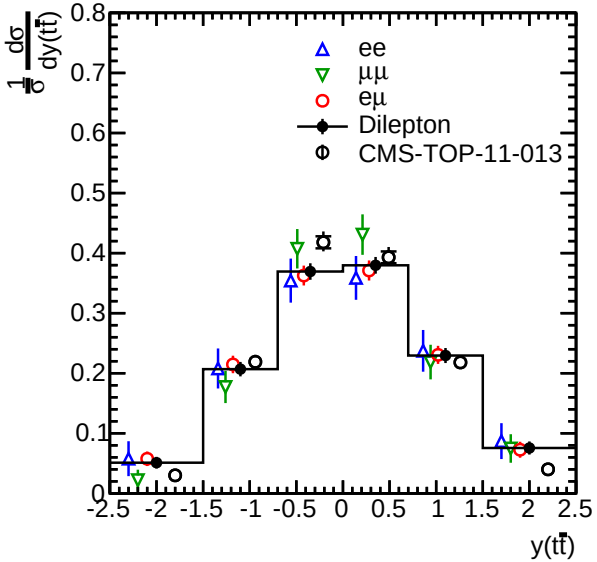
- normalised differential x-section $\frac{1}{\sigma}\frac{\mathrm{d}\sigma}{\mathrm{d}Y}$ (published in TOP-11-013)
- total x-section (published in TOP-13-004)

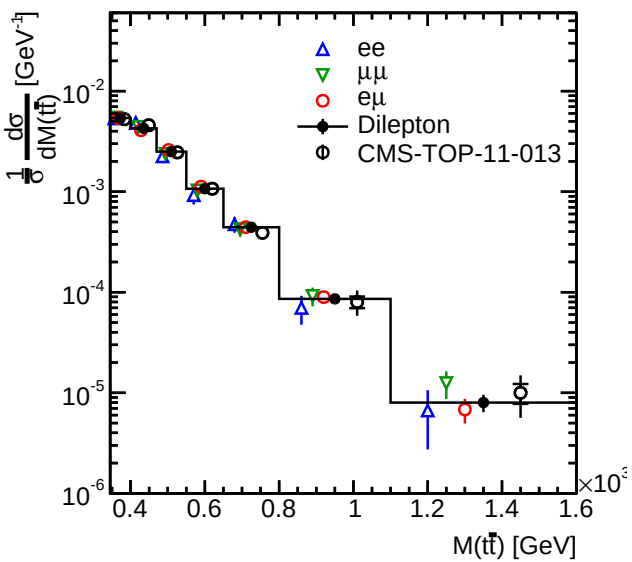
Normalised differential cross sections











- Consistent results in all three channels
- **Consistent with the paper!** (no difference larger than 3 σ)
- Trends at large $|y(t)|$, $|y(t\bar{t})|$: unfolding might be important here

Total cross section

How it is determined:

- Using $e\mu$ channel only (most precise)
- In this analysis obtained by integrating over differential x-section
- More sophisticated procedure in the paper

Open data

e.g. by integrating $p_T(t)$ diff. x-section:
$\sigma(t\bar{t}) = 163.3 \pm 4.3$ (stat) pb

additionally spread between integrations over different variables ≈ 10 pb:
$\sigma(t\bar{t}) = 163.3 \pm 4.3$ (stat) $\pm$ 5 (syst) pb

TOP-13-004

173.6 ± 2.1 (stat) $^{+4.5}_{-4.0}$ (sys) ± 3.8 (lum) pb

- Reasonable consistency. In agreement with larger MC / data rate, missing corrections etc.
- Total x-section sensitive to (in)efficiency, scale factor issues (cancel to large extend for normalised x-section).

Summary

- **Successfull validation of 2011 open data by re-doing published CMS measurement**
- Only similar event selection/reconstruction, no sophisticated corrections, sometimes even completely different procedures
- All key features reproduced, cross sections in reasonable agreement with published values
- **Analysis code available at** `Github`:
  `https://github.com/zenaiev/2011-ttbar`
  also to be available from the CMS open data webpage
  (good starting points for bachelor or master students, as well as interested physicists not from CMS)