An example of PAT analysis

Joaquim Speck

Journées CMS France, Strasbourg

27, 28 mai 2009

Outline

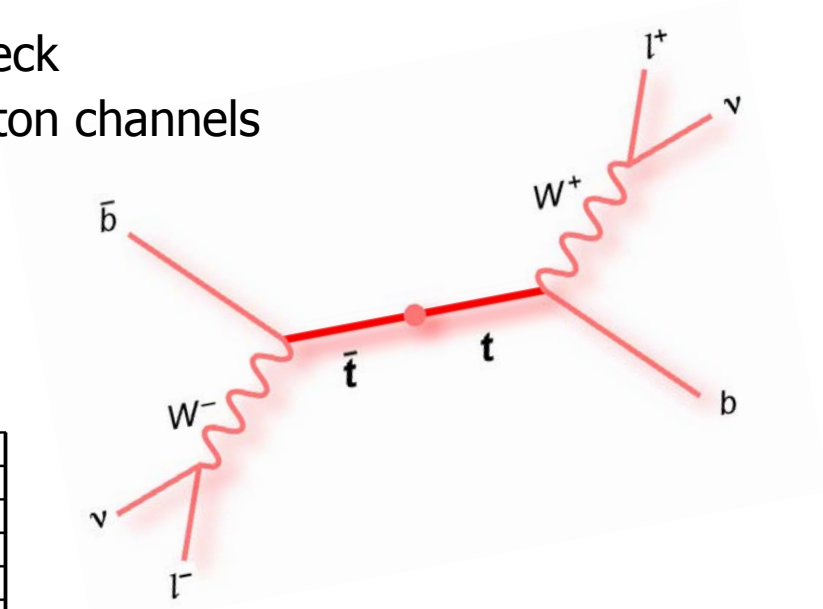
- A presentation of how one can organize an analysis around PAT
  - starting from scratch up to the final plots
- Very brief description of the top dilepton cross section measurement

- I will show how we proceeded during our analysis
  - concentrating on the computational point of view
- Different stages and crosschecks with other contributors
- Conclusion

Top dilepton analysis

- People involved: D. Gelé, J. Andrea, J. Speck
- Extract the ttbar cross section in the dilepton channels
  - 2 isolated, high p_T leptons
  - MET from the 2 neutrinos
  - 2 jets from b quarks (b-tagging)
- Backgrounds with a similar signature

Process	Number of events	Cross-sections (pb)	$\mathcal{L}_e$(pb-1)
MADGRAPH $t\bar{t}$	1028K	414	2483
PYTHIA $t\bar{t}$	147K	414	355
MADGRAPH Z+jets	1163K	3700×1.14	276
MADGRAPH W+jets	9854K	40000×1.14	216
MADGRAPH $(Z/W \to l^+l^-)bb$	1005K	289×1.14	3050
PYTHIA WZ	249K	17.3×1.7	8466
PYTHIA $WW \to l^+l^-$	106K	4.8×1.5	14722
PYTHIA $ZZ \to 4l^\pm$	267K	0.1039×1.3	1976
PYTHIA $ZZ \to 2l^\pm 2\nu$	113K	0.318×1.3	273
MADGRAPH Single-top tW	169K	29	5827
MADGRAPH Single-top t-channel	282K	130	2169
MADGRAPH Single-top s-channel	12K	5	2400
QCD EMenriched p_T = 20-30 GeV	5166K	$4 \cdot 10^8$	0.0129
QCD EMenriched p_T = 30-80 GeV	13090K	$1 \cdot 10^8$	0.131
QCD EMenriched p_T = 80-170 GeV	3412K	$1.9 \cdot 10^9$	1.796
Inclusive$\mu Pt15$	5232K	$1.21 \cdot 10^6$	4.324
QCD $b, c \to e$ p_T = 20-30 GeV	2218K	192000	11.5
QCD $b, c \to e$ p_T = 30-80 GeV	1933K	240000	8.1
QCD $b, c \to e$ p_T = 80-170 GeV	798K	22800	35



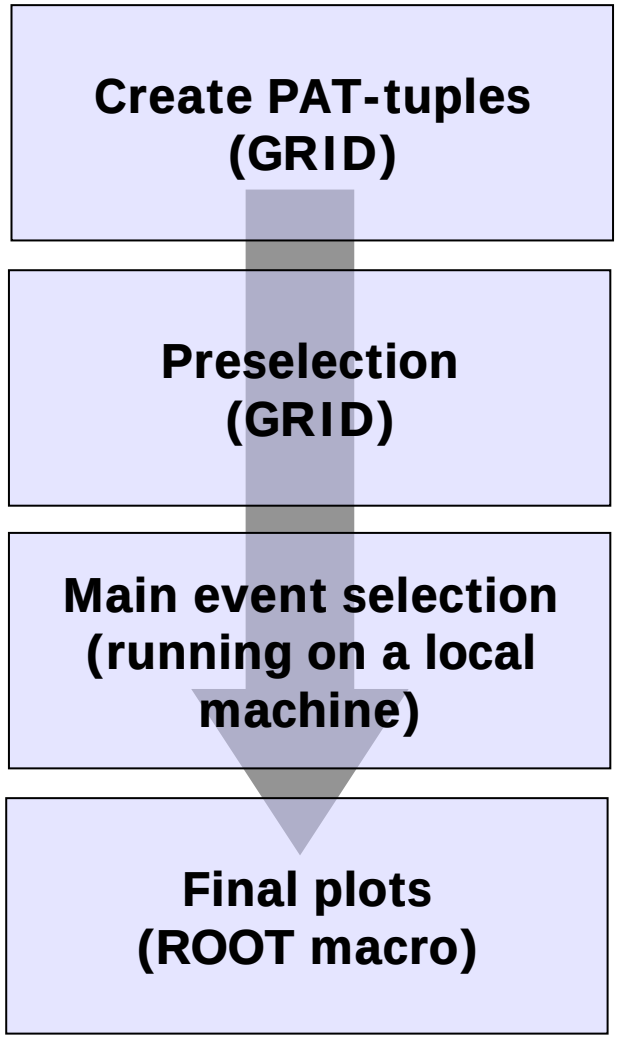
More than **40 million** events!

We need a unique and coherent code that can run on all these samples.

It needs to be **fast** as well.

Main code architecture

- The aim is to reduce the available information to its necessary minimum
  - by keeping only the quantities we are interested in
    - to reduce disc space
    - to reduce processing times
    - to avoid having to re-run on everything from scratch
      - whenever we want to implement a new feature
      - whenever we need to correct a bug!



Physics Analysis Tools

- The **Physics Analysis Toolkit** (PAT) is a high-level analysis layer providing the PAGs with easy access to the algorithms developed by the POG in the CMSSW framework
- It aims at fulfilling the needs of most CMS **physics** analyses
- Easier file content than RECO, but is hence not suited for hardware/software performance and efficiency studies

- We only use PAT in the **early stage** of our analysis
- PAT-tuples are created using standard configuration files
  - official PAT-tuples should be provided by the different production Tiers (at some point)
  - we have produced our own **official** ones
  - PATLayer0 & 1 ran using the Grid
- No personal code is used yet
- PAT is the way to keep people official as long as possible

Preselection

- Dedicated event preselection: We fill up a ROOT Tree with rewritten C++ classes containing
  - MC and event information
  - collection of muons, electrons
    - loose cuts:
      - p_T>10GeV, |η|<2.4, quality cuts
  - tracks, jets and MET
  - still keeping some relevant inheritances from CMSSW classes

/TTJets-madgraph/Fall08_IDEAL_V9_v2/GEN-SIM-RECO

Created 24 Nov 2008 08:38:26 CET, contains 1028322 events, 533 files, 15 block(s), 513.3G

Release info , Block info , Run info , Conf. files , Parents , Children , Description , PhEDEx , C

Location	Events	Files	size	LFNs
T2_US_UCSD : srm-3.t2.ucsd.edu	1028322	533	513.3GB	cff plain
T2_ES_IFCA : storm.ifca.es	1028322	533	513.3GB	cff plain
T2_BE_IIHE : maite.iihe.ac.be	1028322	533	513.3GB	cff plain
No SiteDB name : srm.ucr.edu	1028322	533	513.3GB	cff plain
T2_DE_RWTH : grid-srm.physik.rwth-aachen.de	1028322	533	513.3GB	cff plain
T2_FR_IPHC : sbgse1.in2p3.fr	1028322	533	513.3GB	cff plain
T3_US_FNALLPC : cmsdca2.fnal.gov	1028322	533	513.3GB	cff plain
T2_US_Wisconsin : cmssrm.hep.wisc.edu	1028322	533	513.3GB	cff plain
T2_US_Nebraska : srm.unl.edu	1028322	533	513.3GB	cff plain
T2_DE_DESY : dcache-se-cms.desy.de	1028322	533	513.3GB	cff plain
T3_US_TTU : sigmorgh.hpcc.ttu.edu	1028322	533	513.3GB	cff plain
T1_DE_FZK : gridka-dCache.fzk.de	1028322	533	513.3GB	cff plain

- PAT-tuples and Trees are created using the Grid and are copied back to our local SE in Strasbourg

GEN-SIM-RECO	ROOT Tree
513GB	10GB

Crosschecks

- Reaching this stage is the computationally most difficult task
  - because of the data quantity, we need to do this using the Grid
  - If the Grid is performing OK, processing all the signal and background samples can be done within **2 days**
- The main **advantage** is that we can process the data up to this stage once and for all
  - Each contributor (UCSB, Louvain-la-Neuve, Oviedo) has developed his own dedicated preselection code
  - At this stage, we froze a specific sample and **crosschecked** our results
    - to be sure, we got the PATLayer0 & 1 set up correctly
    - all using the same patches
    - and albeit we use different codes, to verify that we got the same end-results (since we are applying the same cuts)
- Once this has been checked, we can just continue to work with the reduced ROOT Trees ☺

Main event selection

- The main selection is then run on the Trees coming from the preselection
- Here, we apply the relevant cuts to extract a dileptonic ttbar signal
  - Lepton p_T > 20 GeV, isolation, Z veto, Njets ≥2, MET > 50GeV, b-tagging
- This can be done locally on **any machine running CMSSW** (2_2_3)
- Output:
  - a ROOT file containing all the quantities needed to produce plots
  - a T_EX file containing the selection numbers
- Processing ~1 million ttbar events corresponds to ~20 minutes/channel
  - If we want to implement an additional feature/cut or correct for some bug, we can **reprocess** the signal data **within the hour**
- This is a crucial point when it comes to **systematic studies**
  - the **JER** systematics study required to redo the analysis at least 500 times! This has been possible within a week and would have been unconceivable without the reformatted Trees

Conclusion

- Finally, simple ROOT macros produce the different plots and selection tables from the files produced during the last step
- Processing time is not a factor anymore (negligible)

- We have a dedicated workflow that splits up the different efforts needed in successive stages
- From its architecture, this workflow can be used directly on real life operations once the data arrives ☺
- Bonus: the code is flexible enough to plugin a different analysis
  - J. Andrea derived another study on the W charge asymmetry by turning the main selection code into a muon+track selection algorithm in only 2 days

Backup

Backup

Backup: Selection tables

Applied cuts	$t\bar{t}$ signal	$t\bar{t} \to \tau\tau$	$t\bar{t}$ bkg	Z+jets	W+jets	Vbb
Presel.+triggers	237.1 ± 2.5	1.0 ± 0.2	7.8 ± 0.6	43192.2 ± 118.6	34.4 ± 4.0	1731.1 ± 7.4
+inv. mass cut	171.9 ± 2.3	0.6 ± 0.2	6.1 ± 0.5	1261.7 ± 21.4	4.1 ± 1.4	59.1 ± 1.4
+number of jets	131.0 ± 2.1	0.4 ± 0.2	4.6 ± 0.5	103.4 ± 6.2	3.2 ± 1.2	4.6 ± 0.4
+$\not{E}_T$cut	71.1 ± 1.6	0.2 ± 0.1	2.3 ± 0.4	14.2 ± 2.3	1.4 ± 0.8	0.4 ± 0.2
+1 b-tag cut	66.1 ± 1.6	0.1 ± 0.1	2.0 ± 0.3	5.9 ± 1.5	0 ± 0	0.3 ± 0.1
+2 b-tag cut	39.0 ± 1.3	0.1 ± 0.1	1.2 ± 0.3	0.4 ± 0.4	0 ± 0	0.1 ± 0.1

Applied cuts	WZ	WW	$ZZ2l2\nu$	$ZZ4l$
Presel.+triggers	36.9 ± 0.7	30.2 ± 0.5	4.9 ± 0.1	2.4 ± 0.1
+inv. mass cut	11.8 ± 0.4	2 ± 0.2	0.4 ± 0.1	0.5 ± 0.1
+number of jets	0.9 ± 0.2	1.7 ± 0.2	0.1 ± 0.1	0.2 ± 0.1
+$\not{E}_T$cut	0.3 ± 0.1	0.8 ± 0.1	0.1 ± 0.1	0.1 ± 0.1
+1 b-tag cut	0.2 ± 0.1	0.3 ± 0.1	0.1 ± 0.1	0.1 ± 0.1
+2 b-tag cut	0 ± 0	0.1 ± 0.1	0.1 ± 0.1	0.1 ± 0.1

Applied cuts	tW	t-channel	s-channel	QCD	Total backgrounds	S/B
Presel.+triggers	24.8 ± 1	0.9 ± 0.2	0.1 ± 0.1	967.2 ± 895.9	46049.5 ± 991.7	0.005
+inv. mass cut	8.2 ± 0.6	0.3 ± 0.2	0 ± 0	0 ± 0	1354.2 ± 20.8	0.13
+number of jets	6.4 ± 0.5	0.3 ± 0.2	0 ± 0	0 ± 0	125.4 ± 6.4	1
+$\not{E}_T$cut	3.4 ± 0.4	0.3 ± 0.1	0 ± 0	0 ± 0	23.3 ± 2.5	3.1
+1 b-tag cut	2.9 ± 0.4	0.2 ± 0.1	0 ± 0	0 ± 0	12.0 ± 1.6	5.5
+2 b-tag cut	1.2 ± 0.2	0.1 ± 0.1	0 ± 0	0 ± 0	3.3 ± 0.6	11.8

Table 10: The tables give the expected number of signal and background events passing the different cumulated selection criteria for the ee-channel for 100 pb^{-1} of integrated luminosity, for which we expect around 700 ee events. The $t\bar{t}$ signal numbers include $\tau \to e$ decay. The contribution of $t\bar{t} \to \tau\tau \to ee$ is also given here. The important yield for QCD events is due to the high scale factor; there are only three events passing the preselection.