# Analysis Methods for Hadron Colliders I

Beate Heinemann

*UC Berkeley and Lawrence Berkeley National Laboratory*

TRIUMF Summer Institute, July 2009

# Introduction and Disclaimer

- Data Analysis in 3 hours !
  - Impossible to cover all…
    - There are gazillions of analyses
    - Also really needs learning by doing
      - That's why your PhD takes years!
  - Will try to give a flavor using illustrative examples:
    - What are the main issues
    - And what can go wrong
  - Will try to highlight most important issues
- Please ask during / after lecture and in discussion section!
  - I will post references for your further information also
    - Generally it is a good idea to read theses

# Outline

- Lecture I:
  - Measuring a cross section
    - focus on acceptance
- Lecture II:
  - Measuring a property of a known particle
- Lecture III:
  - Searching for a new particle
    - focus on backgrounds

# Cross Section: Experimentally

Number of observed events: counted

Background: Measured from data / calculated from theory

$$\sigma = \frac{N_{obs} - N_{BG}}{\int L dt \cdot \varepsilon}$$

Cross section σ

Luminosity: Determined by accelerator, trigger prescale, …

Efficiency: optimized by experimentalist

# Uncertainty on Cross Section

- You will want to minimize the uncertainty:

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

- 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

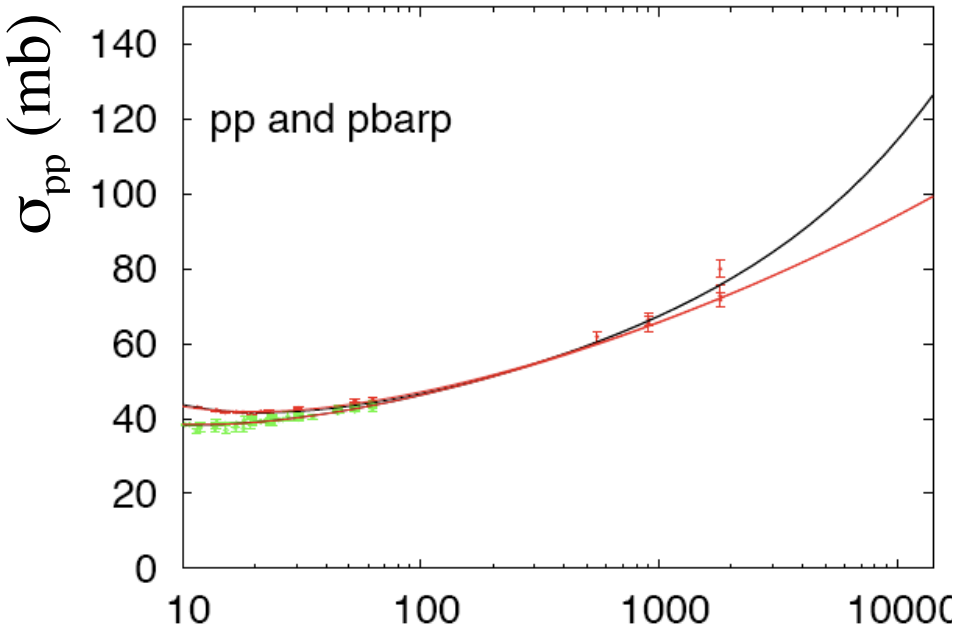
# Luminosity Measurement

- Many different ways to measure it:
  - Beam optics
    - LHC startup: precision ~10-30%
    - Ultimately: precision ~5%
  - Relate number of interactions to total cross section
    - absolute precision ~4-6%, relative precision much better
  - Elastic scattering:
    - LHC: absolute precision ~3%
  - Physics processes:
    - W/Z: precision ~2-3% ?
- Need to measure it as function of time:
  - $L = L_0 e^{-t/\tau}$ with $\tau \approx 14$h at LHC and $L_0$ = initial luminosity

# Luminosity Measurement

**Rate of pp collisions: $R_{pp} = \sigma_{inel}\ \varepsilon\ L_{inst}$**

- Measure fraction of beam crossings with no interactions
  - Related to $R_{pp}$
- Relative normalization possible
  - if Probability for no interaction>0 ($L<10^{32}$ $cm^{-2}s^{-1}$)
- Absolute normalization
  - Normalize to measured inelastic pp cross section
  - Measured by CDF and E710/E811
    - Differ by 2.6 sigma
    - For luminosity normalization use the error weighted average

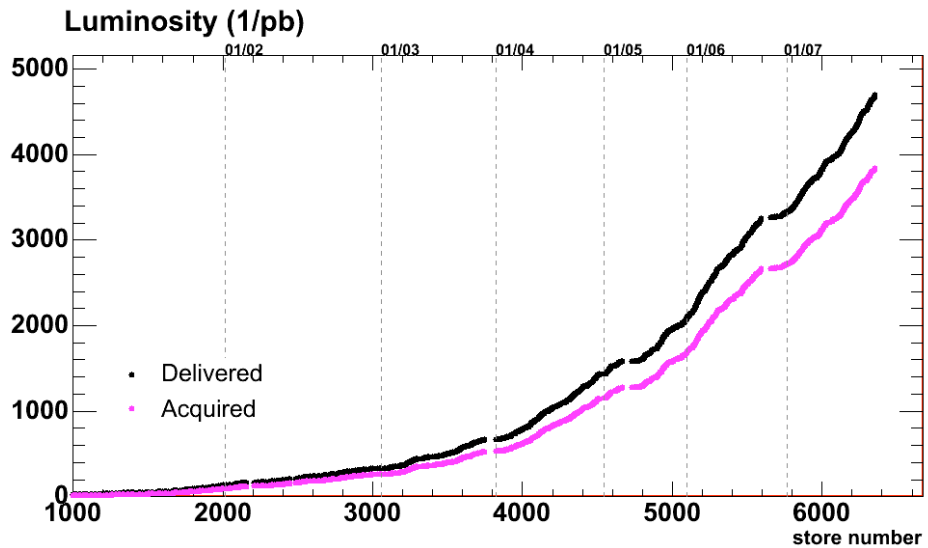


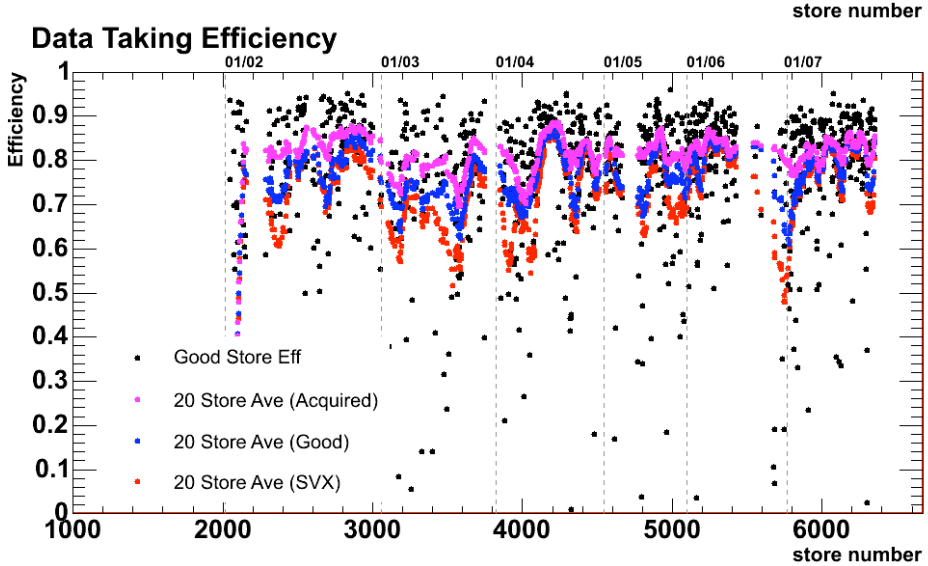

| | 1.96 TeV | 14 TeV |
|---|---|---|
| $\sigma_{inelastic}$ | 60.7±2.4 mb (measured) | 125±25 mb (P. Landshoff) |

# Your luminosity

- Your data analysis luminosity is not equals to LHC/Tevatron luminosity!
- Because:
  - Detector dead-time => live fraction $l_i$
  - The detector is not 100% efficiency at taking data: $\varepsilon_i$
  - Your trigger may have been off / prescaled at times: $p_i$
  - Some of your jobs crashed and you could not run over all events
- All needs to be taken into account
  - Severe bookkeeping headache

$$\int L dt = \sum_{LB_i} L_i \, l_i \, p_i \ldots \varepsilon_i \, \Delta t_i$$



# Acceptance / Efficiency

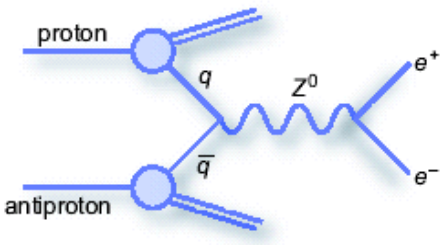
- Actually rather complex:
  - Many ingredients enter here
  - You need to know:

$$\varepsilon_{total} = \frac{\text{Number of Events used in Analysis}}{\text{Number of Events Produced}}$$

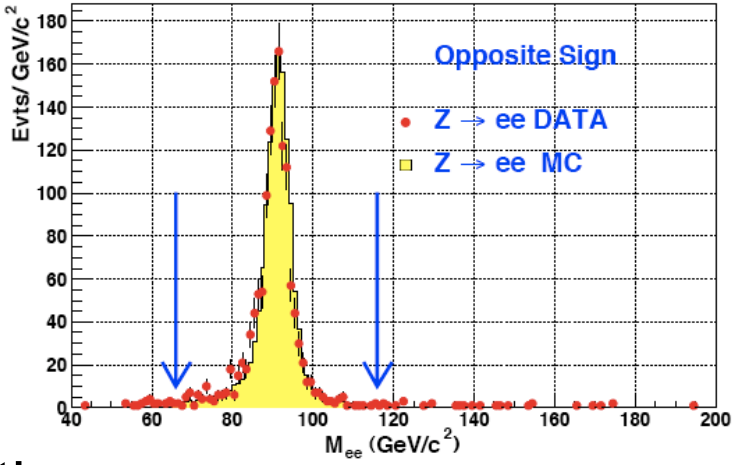
- Ingredients:
  - Trigger efficiency
  - Identification efficiency
  - Kinematic acceptance
  - Cut efficiencies
- Using three example measurements for illustration:
  - Z boson, top quak and jet cross sections

# Example Analyses

# Z Boson Cross Section




- Trigger requires one electron with $E_T$>20 GeV
  - Criteria at L1, L2 and L3/EventFilter
- You select two electrons in the analysis
  - With certain quality criteria
  - With an isolation requirement
  - With $E_T$>25 GeV and |eta|<2.5
  - With oppositely charged tracks with $p_T$>10 GeV
- You require the di-electron mass to be near the Z:
  - 66<M(ll)<116 GeV

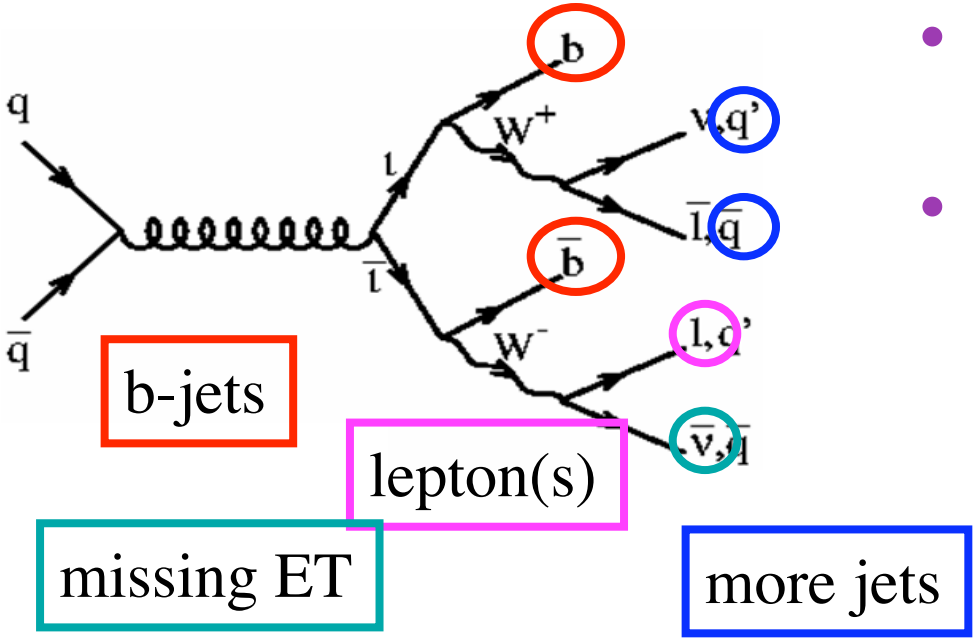



$$\Rightarrow \varepsilon_{total} = \varepsilon_{trig}\varepsilon_{rec}\varepsilon_{ID}\varepsilon_{kin}\varepsilon_{track}$$

# Top Quark Cross Section

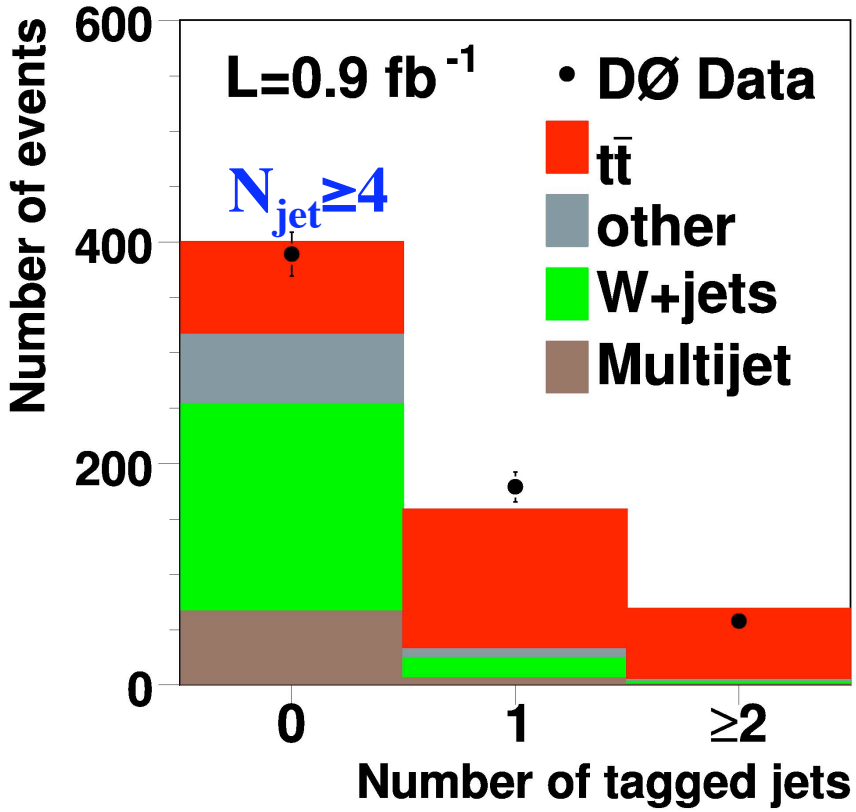
**SM: $t\bar{t}$ pair production, Br($t \rightarrow bW$)=100% , Br(W->lν)=1/9=11%**

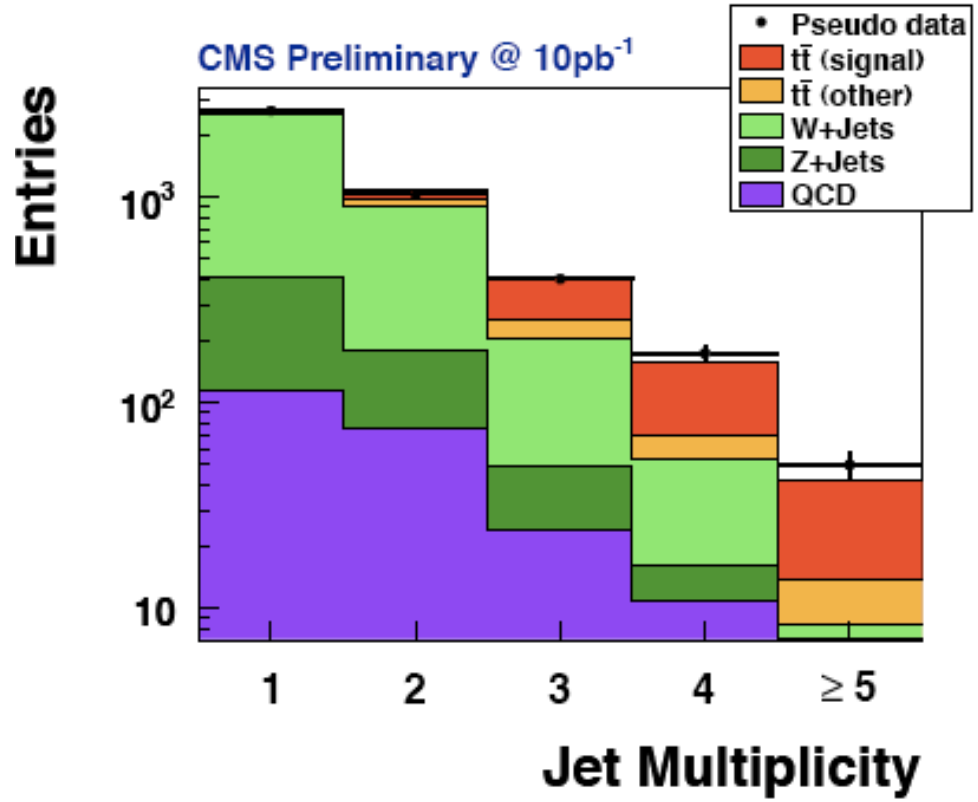
| | | |
|---|---|---|
| **dilepton** | ***(4/81)*** | **2 leptons + 2 jets + missing $E_T$** |
| **lepton+jets** | ***(24/81)*** | **1 lepton + 4 jets + missing $E_T$** |
| **fully hadronic** | ***(36/81)*** | **6 jets** |




- Trigger on electron/muon
  - Like for Z's
- Analysis cuts:
  - Electron/muon $p_T$>25 GeV
  - Missing $E_T$>25 GeV
  - 3 or 4 jets with $E_T$>20-40 GeV

# Finding the Top Quark

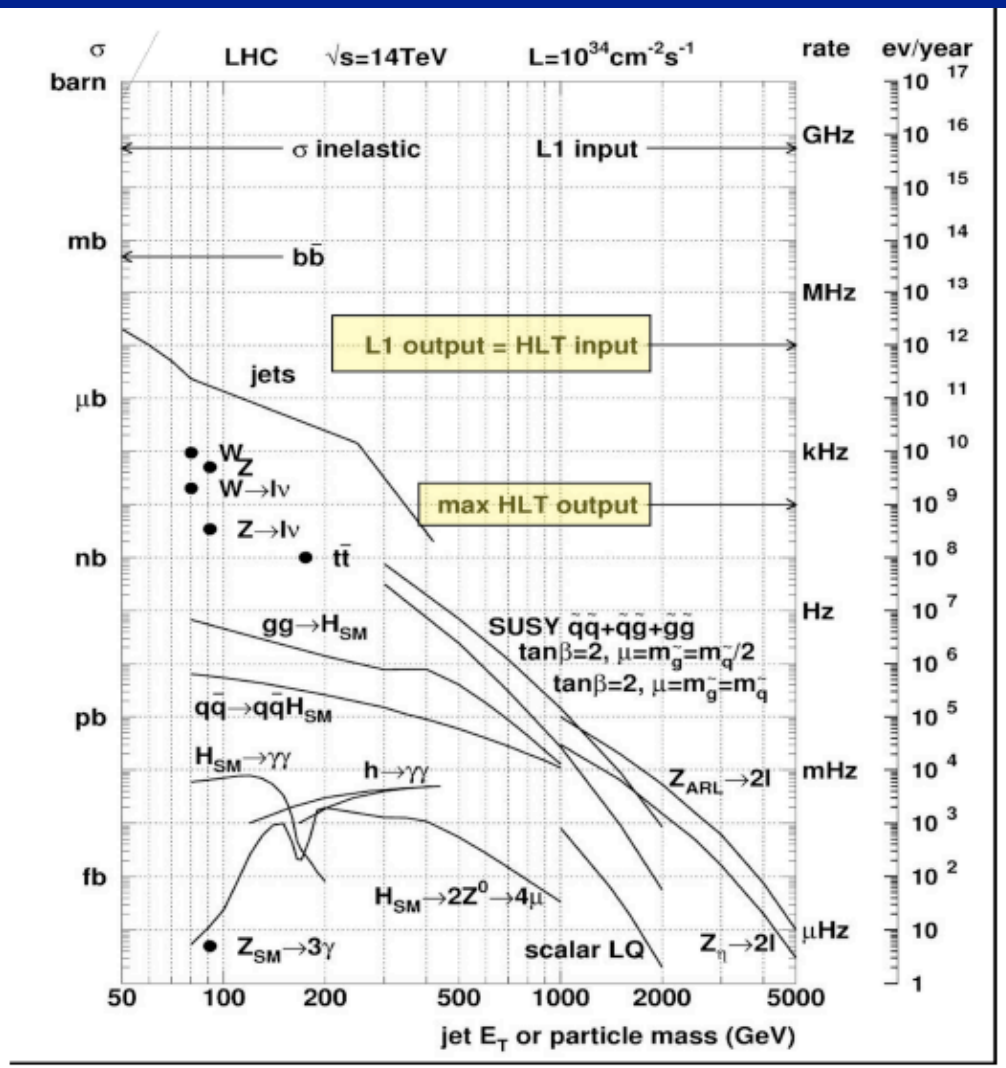



- **Tevatron**
  - Top is overwhelmed by backgrounds:
  - Top fraction is only 10% (≥3 jets) or 40% (≥4 jets)
  - Use b-jets to purify sample => purity 50% (≥3 jets) or 80% (≥4 jets)
- **LHC**
  - Purity ~70% w/o b-tagging (90% w b-tagging)

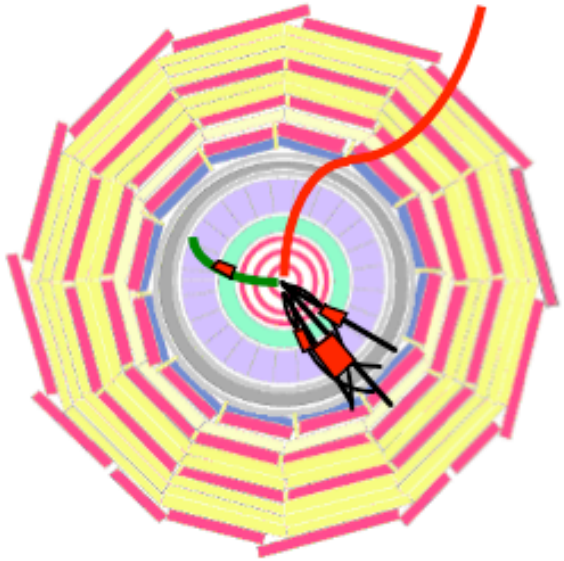
# Trigger

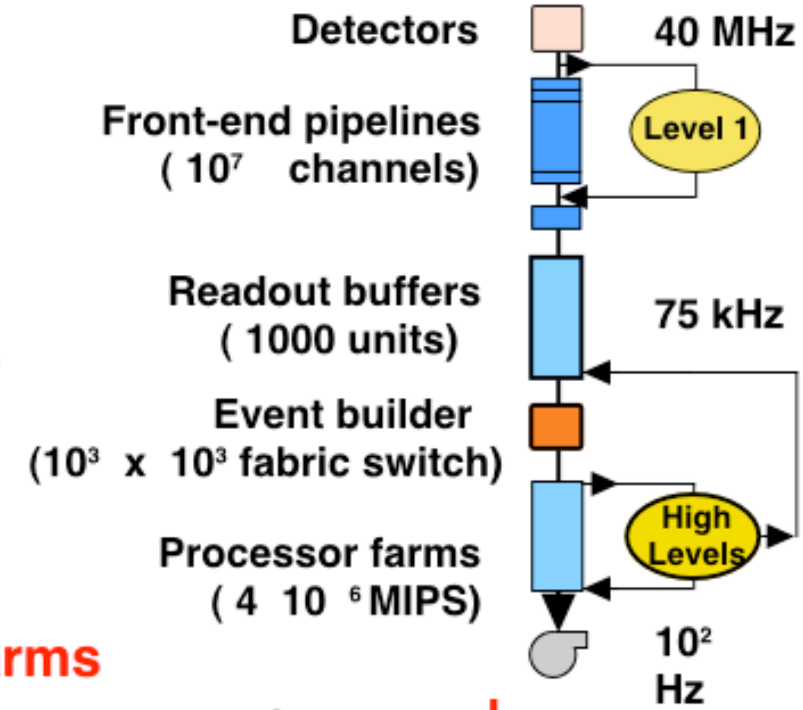
# Trigger Rate vs Physics Cross Section




- Acceptable Trigger Rate << many physics cross sections

# Example: CMS trigger






**High level triggers. CPU farms**

- Finer granularity precise measurement
- Clean particle signature ($\pi^0$-$\gamma$, isolation, ...)
- Kinematics. Effective mass cuts and topology
- Track reco and matching, b,$\tau$-jet tagging
- Full event reconstruction and analysis

**Successive improvements : background event filtering, physics selection**

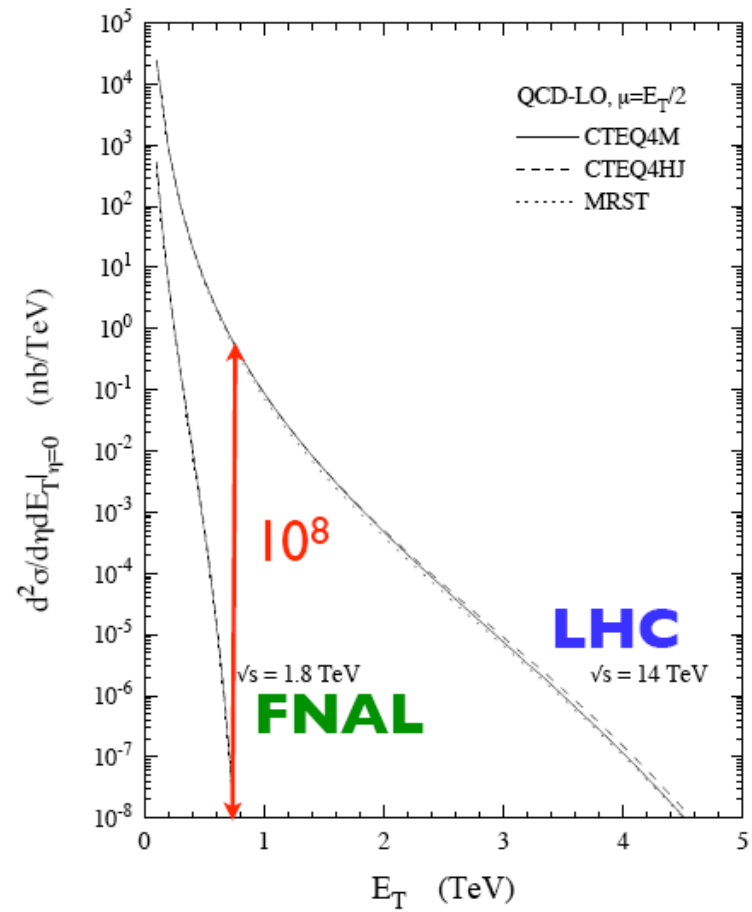
**NB: Similar output rate at the Tevatron**

# Tevatron versus LHC Cross Sections

**Cross Sections of Physics Processes (pb)**

| | Tevatron | LHC | Ratio |
|---|---|---|---|
| $W^{\pm}$ (80 GeV) | 2600 | 20000 | 10 |
| $t\bar{t}$ (2x172 GeV) | 7 | 800 | 100 |
| $gg \rightarrow H$ (120 GeV) | 1 | 40 | 40 |
| $\tilde{\chi}^{+}_{1}\tilde{\chi}^{2}_{0}$ (2x150 GeV) | 0.1 | 1 | 10 |
| $\tilde{q}\tilde{q}$ (2x400 GeV) | 0.05 | 60 | 1000 |
| $\tilde{g}\tilde{g}$ (2x400 GeV) | 0.005 | 100 | 20000 |
| Z' (1 TeV) | 0.1 | 30 | 300 |

**Jet Cross Section**




- **Amazing increase for strongly interacting heavy particles!**
- **LHC has to trigger >10 times more selectively than Tevatron**

# Are your events being triggered?

- Typically yes, if
  - events contain high $p_T$ isolated leptons
    - e.g. top, Z, W
  - events contain very high $p_T$ jets or very high missing $E_T$
    - e.g. SUSY
  - …
- Possibly no, if
  - events contain only low-momentum objects
    - E.g. two 20 GeV b-jets
      - Still triggered maybe at Tevatron but not at LHC
  - ….
- This is the first thing you need to find out when planning an analysis
  - If not then you want to design a trigger if possible

# Examples for Unprescaled Triggers

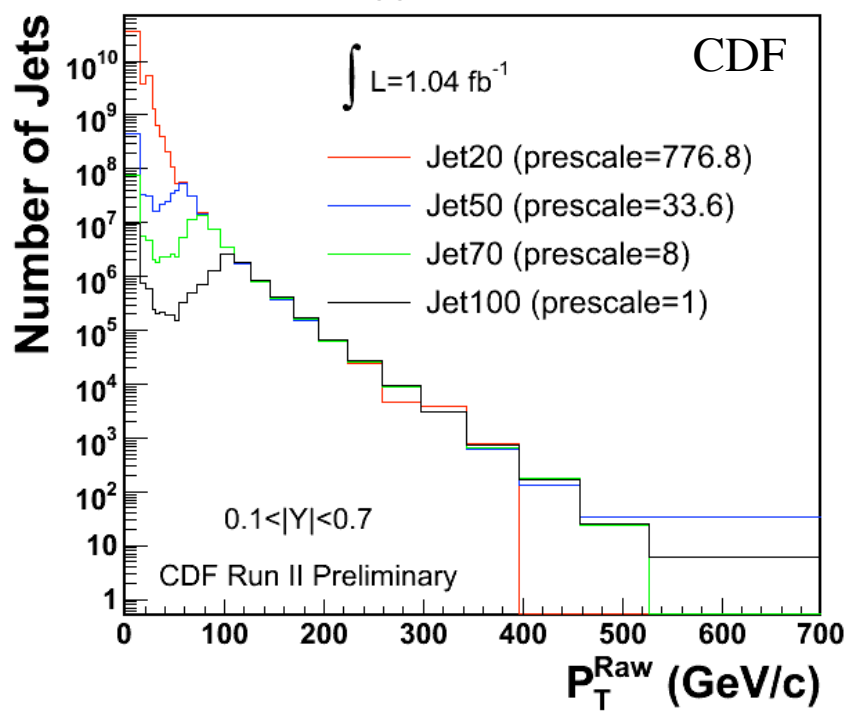
| | ATLAS(*) ($L=2\times10^{33}$ $cm^{-2}s^{-1}$) | CDF ($L=3\times10^{32}$ $cm^{-2}s^{-1}$) |
|---|---|---|
| MET | > 70 GeV | > 40 GeV |
| Jet | > 370 GeV | > 100 GeV |
| Photon (iso) | > 55 GeV | > 25 GeV |
| Muon | iso + $p_T$> 20 GeV | > 20 GeV |
| Electron | Iso + $E_T$> 22 GeV | > 20 GeV |
| incl. dimuon | > 10 GeV | > 4 GeV |

- Increasing luminosity leads to
  - Tighter cuts, smarter algorithms, prescales
  - Important to pay attention to this for your analysis!

# Typical Triggers and their Usage

- Unprescaled triggers for primary physics goals, e.g.
  - Inclusive electrons, muons $p_T$>20 GeV:
    - W, Z, top, WH, single top, SUSY, Z',W'
  - Lepton+tau, $p_T$>8-25 GeV:
    - MSSM Higgs, SUSY, Z
    - Also have tau+MET: W->taunu
  - Jets, $E_T$>100-400 GeV
    - Jet cross section, Monojet search
    - Lepton and b-jet fake rates
  - Photons, $E_T$>25 GeV:
    - Photon cross sections, Jet energy scale
    - Searches (GMSB SUSY), ED's
  - Missing $E_T$>45-100 GeV
    - SUSY

- Prescale triggers because:
  - Not possible to keep at highest luminosity
  - But needed for monitoring
  - Prescales depend often on Luminosity
- Examples:
  - Jets at $E_T$>20, 50, 70 GeV
  - Inclusive leptons >8 GeV
  - Backup triggers for any threshold, e.g. Met, jet ET, etc…
    - At all trigger levels

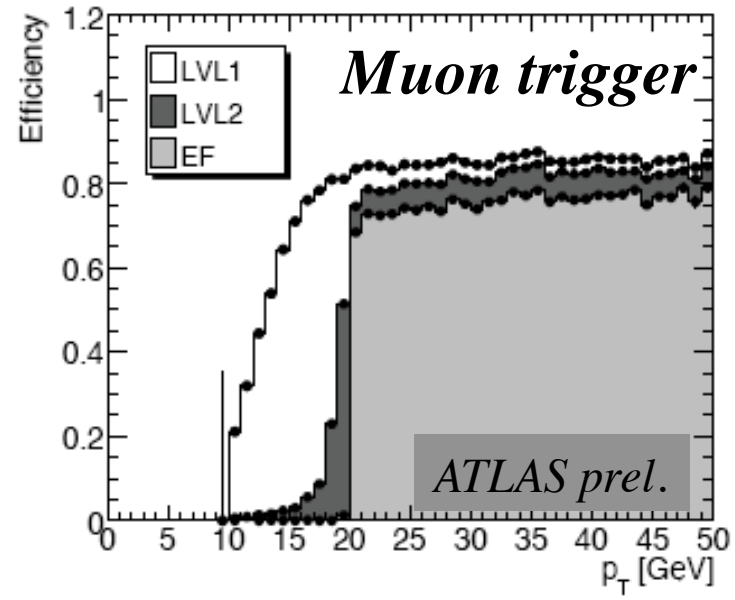


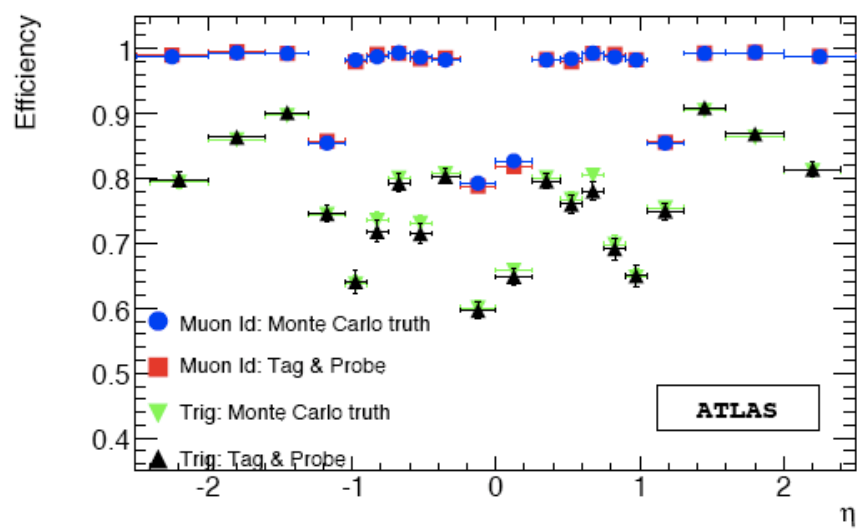
# Trigger Efficiency for e's and μ's

- Can be measured using Z's with tag & probe method
  - Statistically limited

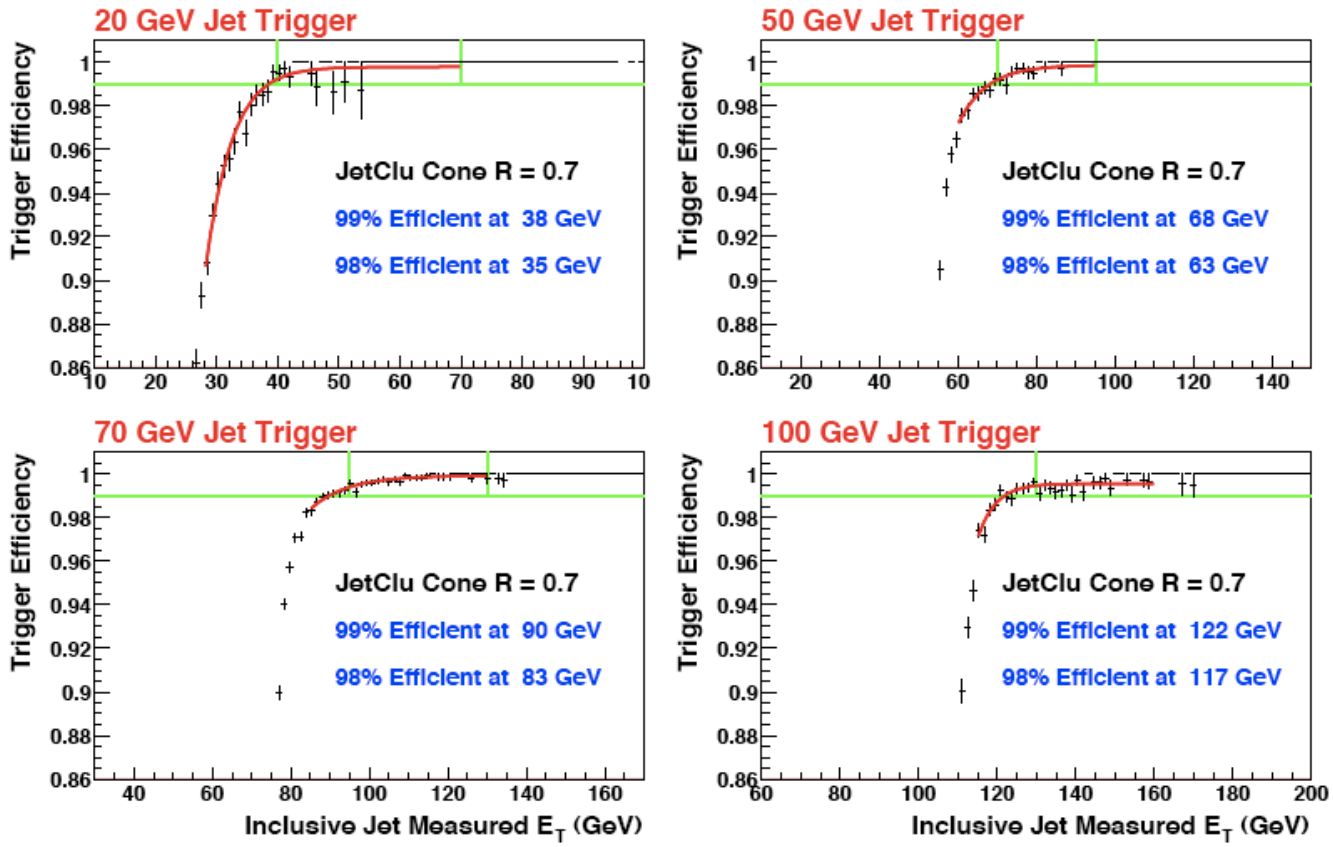
$$\varepsilon_{trig} = \frac{N_{trig}}{N_{ID}}$$

- Can also use trigger with more loose cuts to check trigger with tight cuts to map out
  - Energy dependence
    - turn-on curve decides on where you put the cut
  - Angular dependence
    - Map out uninstrumented / inefficient parts of the detectors, e.g. dead chambers
  - Run dependence
    - Temporarily masked channels (e.g. due to noise)



# Jet Trigger Efficiencies




- Bootstrapping method:
  - E.g. use MinBias to measure Jet-20, use Jet-20 to measure Jet-50 efficiency … etc.
- Rule of thumb: choose analysis cut where $\varepsilon$>90-95%
  - Difficult to understand the exact turnon

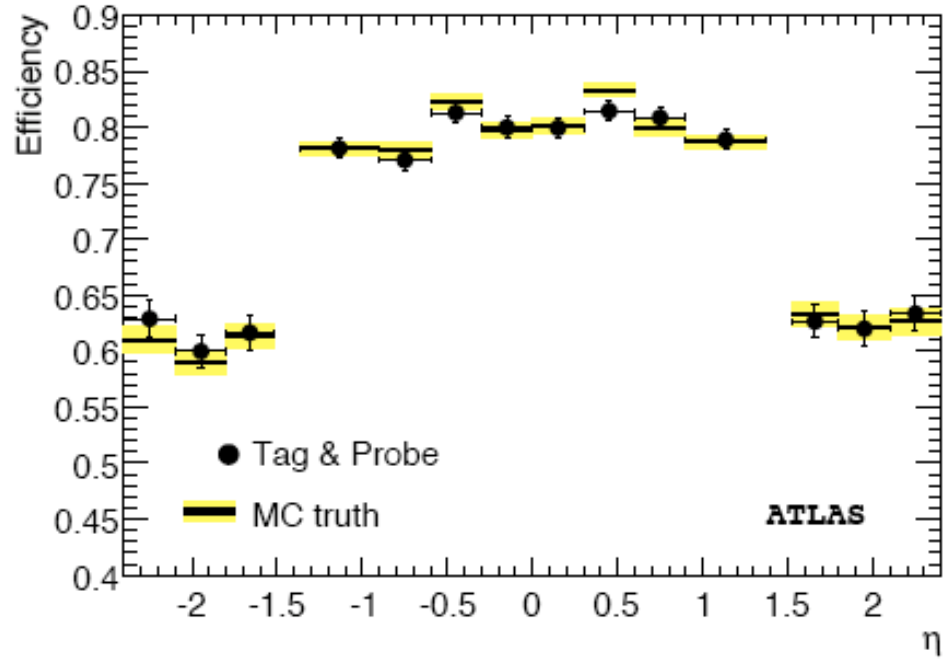
# Efficiencies

Two Examples

- Electrons
- B-jets

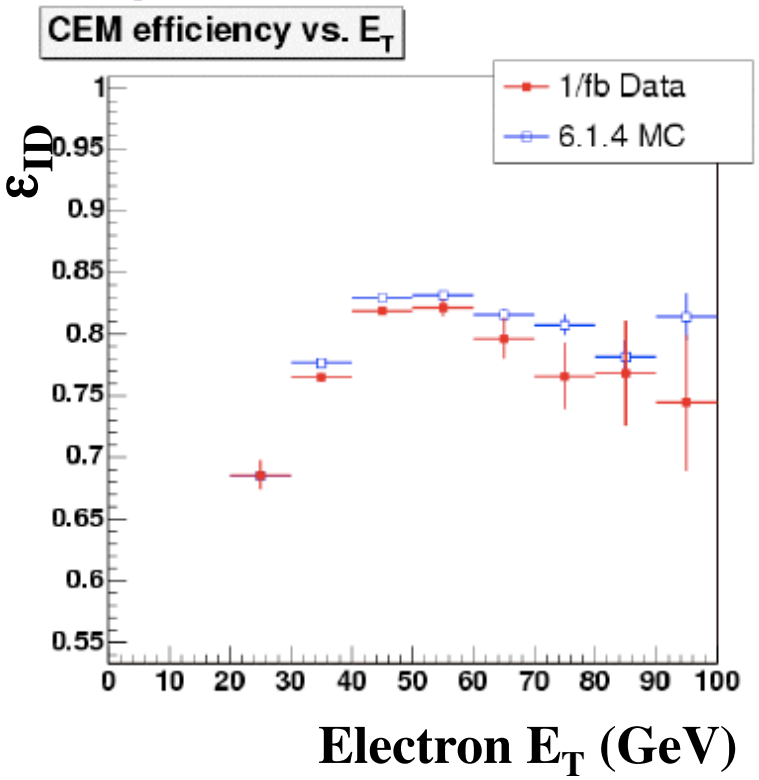
# Electron Identification

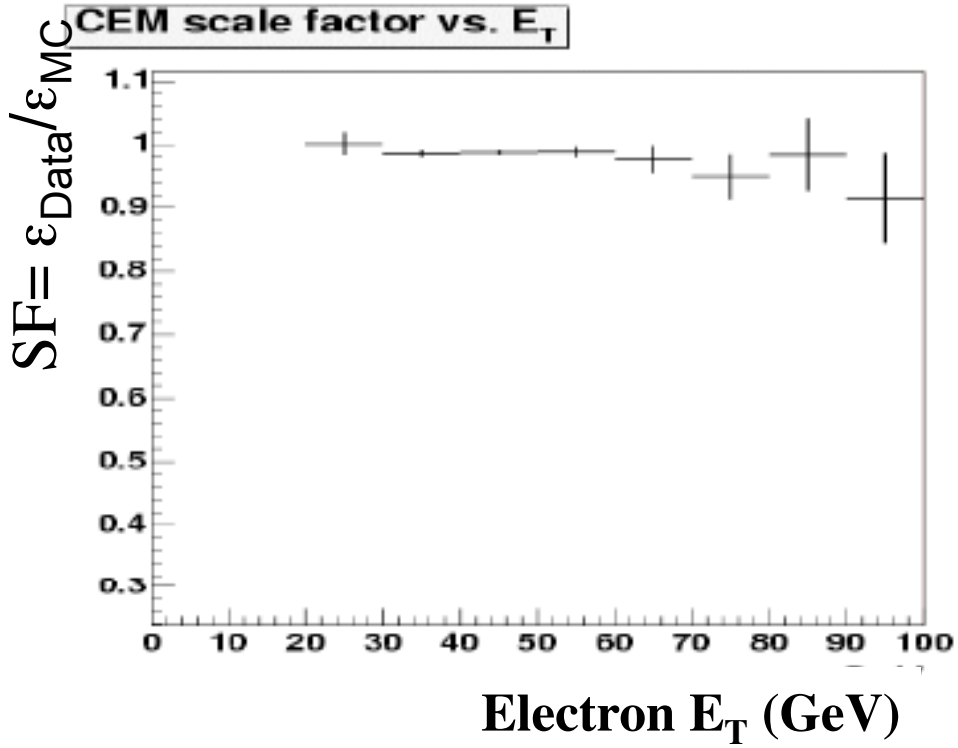
- Desire:
  - High efficiency for (isolated) electrons
  - Low misidentification of jets
- Cuts:
  - Shower shape
  - Low hadronic energy
  - Track requirement
  - Isolation
- Performance:
  - Efficiency measured from Z's using "tag and probe" method
  - Usually measure "scale factor":
    - $SF=\varepsilon_{Data}/\varepsilon_{MC}$ (=1 for perfect MC)
    - Easily applied to MC




| | CDF | ATLAS |
|---|---|---|
| Loose cuts | 85% | 88% |
| Tight cuts | 60-80% | ~65% |

# Electron ID "Scale Factor"

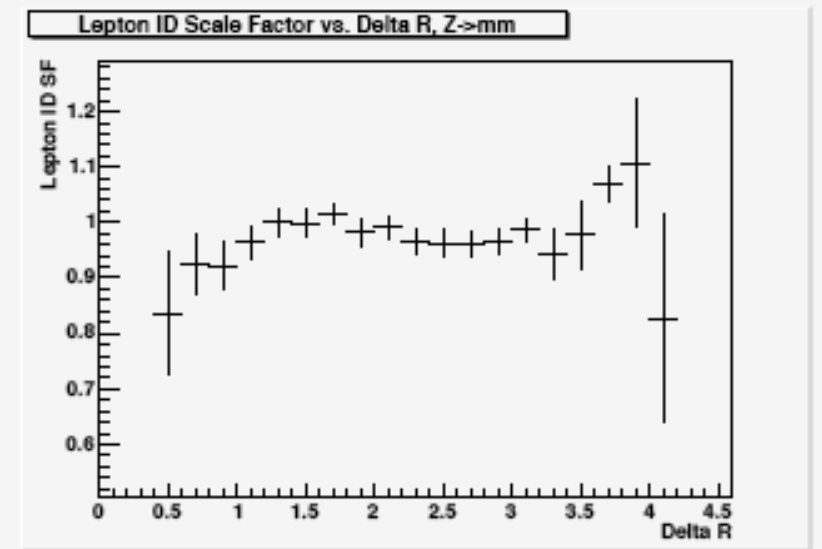


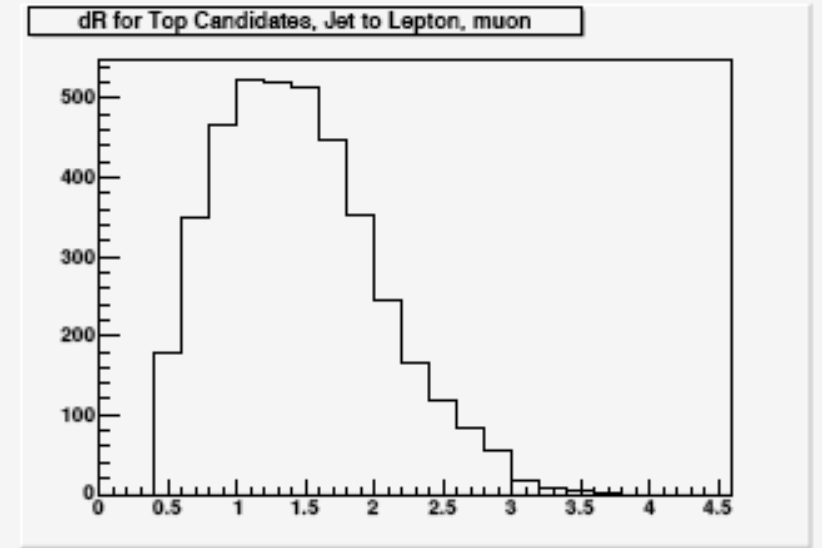




- Efficiency can generally depend on lots of variables
  - Mostly the Monte Carlo knows about dependence
- Determine "Scale Factor" = $\varepsilon_{Data}/\varepsilon_{MC}$
  - Apply this to MC
  - Residual dependence on quantities must be checked though

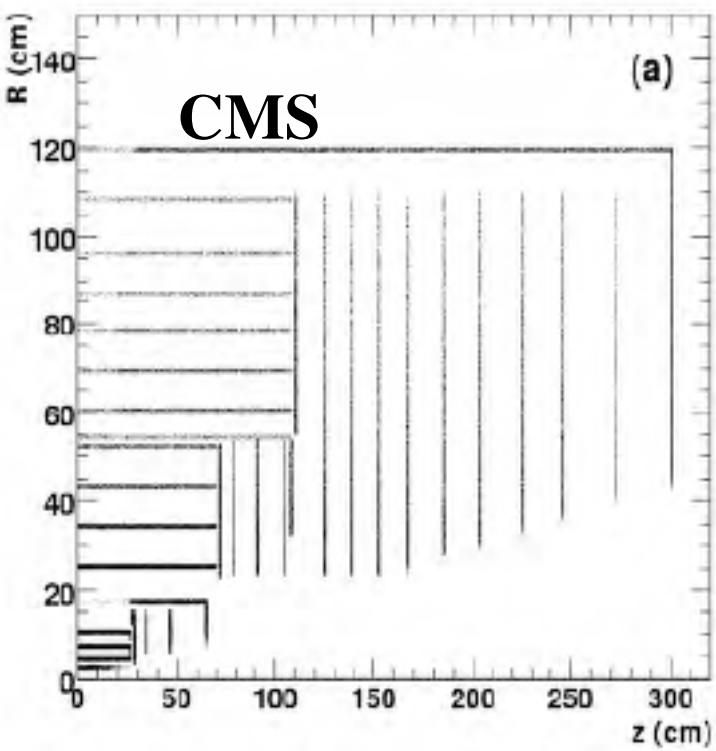
# Beware of Environment

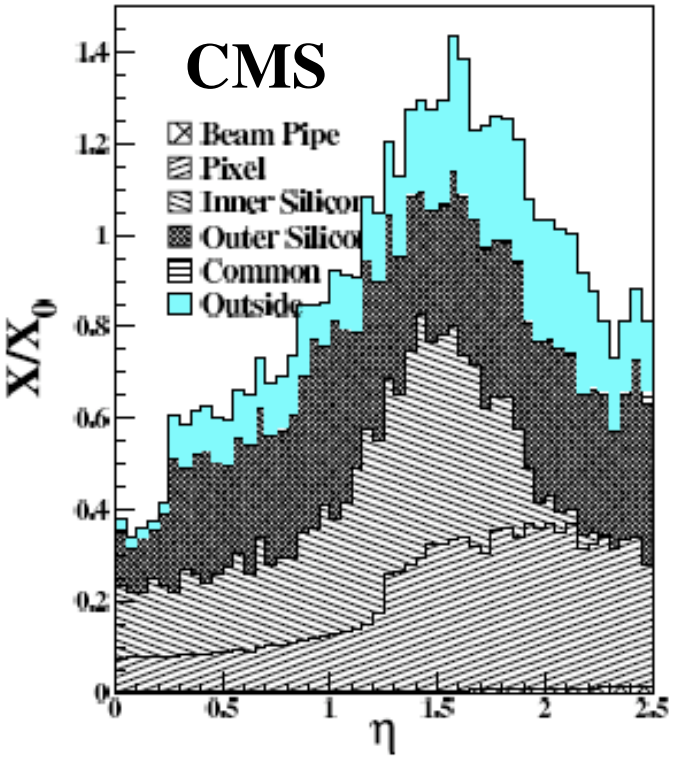
- Efficiency of e.g. isolation cut depends on environment
  - Number of jets in the event
- Check for dependence on distance to closest jet



# Material in Tracker

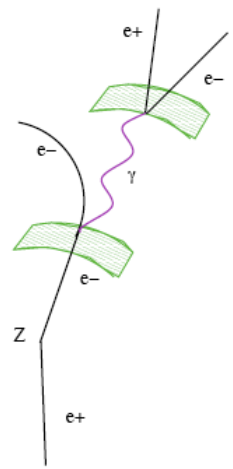


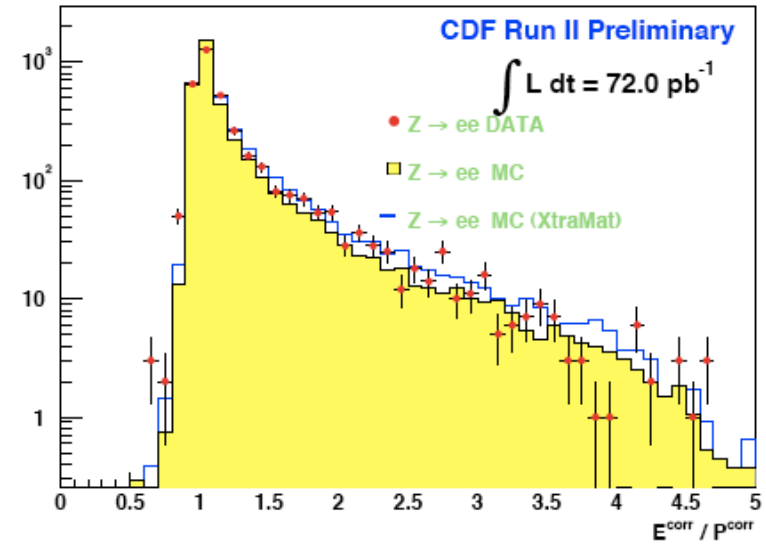



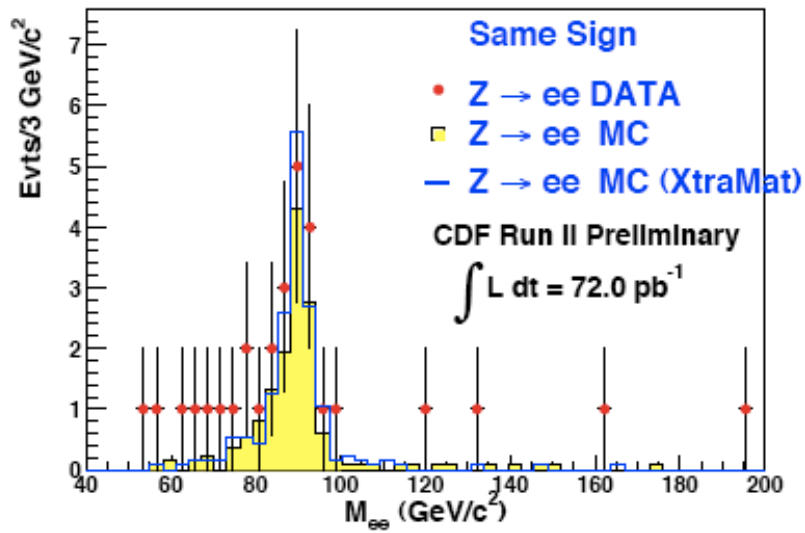

- Silicon detectors at hadron colliders constitute significant amounts of material, e.g. for R<0.4m
  - CDF: ~20% $X_0$
  - ATLAS: ~20-90% $X_0$
  - CMS: ~20-100% $X_0$

# Effects of Material on Analysis

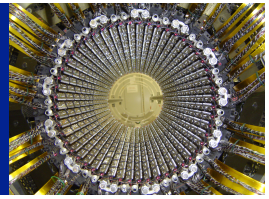
- Causes difficulties for electron/photon identification:
  - Bremsstrahlung
  - Photon conversions
- Constrained with data:
  - Photon conversions
  - E/p distribution
  - Number of $e^{\pm}e^{\pm}$ events



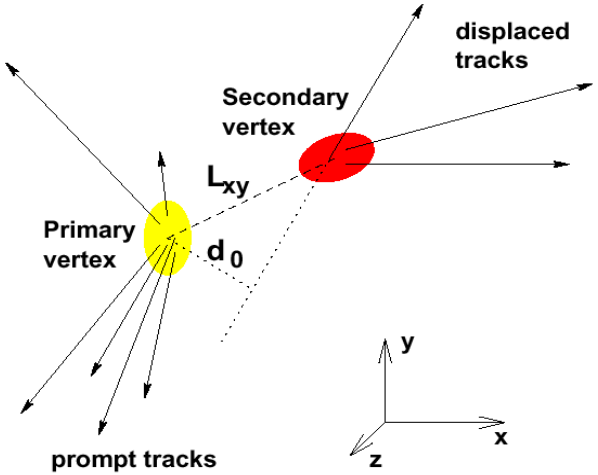



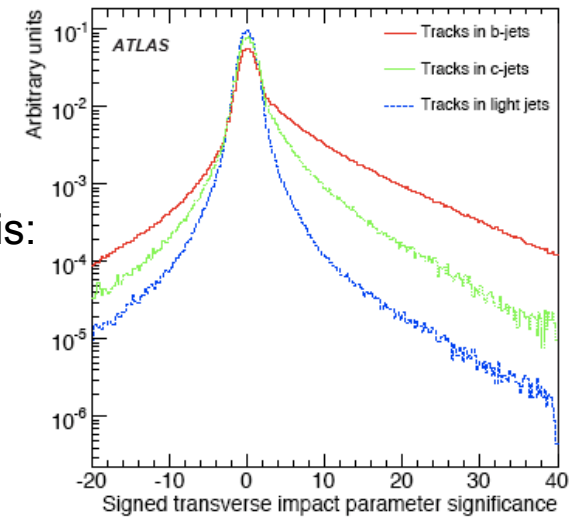



# Finding the b-jets



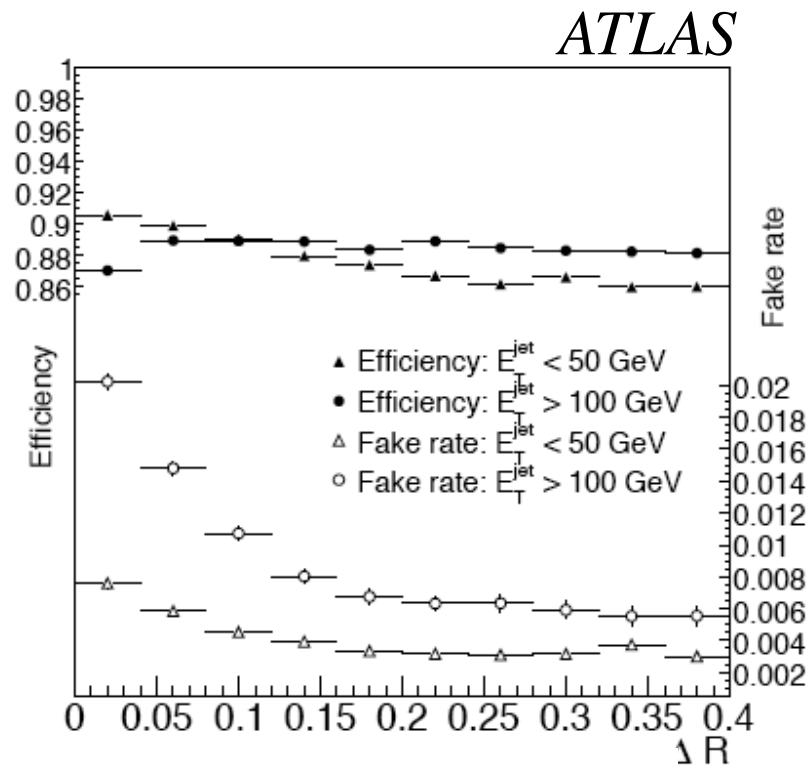
- Exploit large lifetime of the b-hadron
  - B-hadron flies before it decays: $d=c\tau$
    - Lifetime $\tau$ =1.5 $ps^{-1}$
    - $d=c\tau$ = 460 $\mu$m
    - Can be resolved with silicon detector resolution
- Procedure "Secondary Vertex":
  - reconstruct primary vertex:
    - resolution ~ 30 $\mu$m
  - Search tracks inconsistent with prim. vtx (large $d_0$):
    - Candidates for secondary vertex
    - See whether those intersect at one point
  - Require distance of secondary from primary vertex
    - Form $L_{xy}$: transverse decay distance projected onto jet axis:
      - $L_{xy}$>0: b-tag along the jet direction => real b-tag or mistag
      - $L_{xy}$<0: b-tag opposite to jet direction => mistag!
    - Significance: e.g. $\delta L_{xy} / L_{xy}$ >7.5
- More sophisticated techniques exist
  - Neural networks, likelihoods, etc.



# B-tagging relies on tracking in Jets

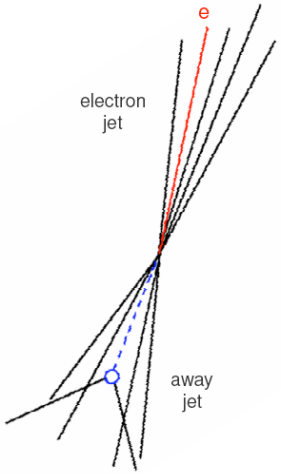
- Finding "soft" tracks inside jets is tough!
  - Difficult pattern recognition in dense environment
- Trade-off of efficiency and fake rate
- Difficult to measure in data
  - Only method I know is "track embedding"
  - Embed a MC track into data and check if one can find it
    - Requires well tuned simulation

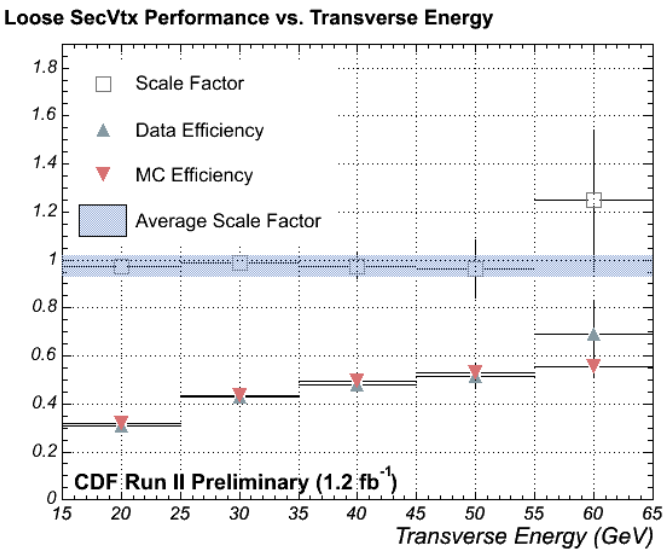



**Distance to closest jet**

# Characterize the B-tagger: Efficiency

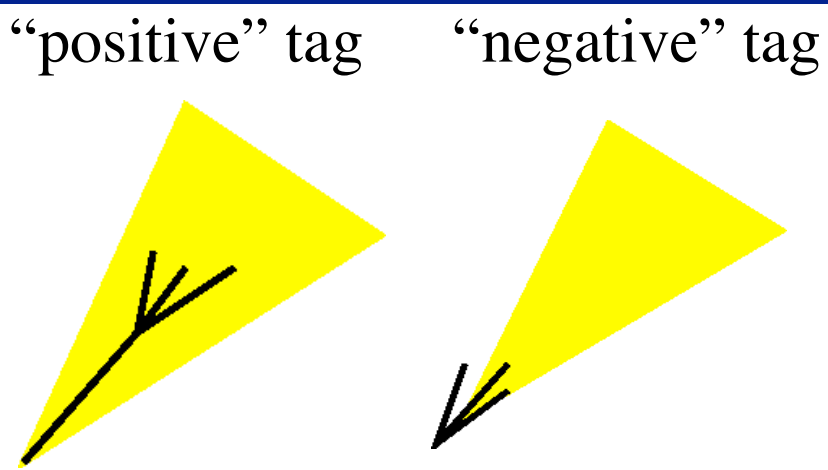
- Efficiency of tagging a true b-jet
  - Use Data sample enriched in b-jets
  - Select jets with electron or muons
    - From semi-leptonic b-decay
    - And b-jet on the opposite side
  - Measure efficiency in data and MC
    - Determine Scale Factor
  - Can also measure it in top events
    - Particularly at LHC ("top factory")

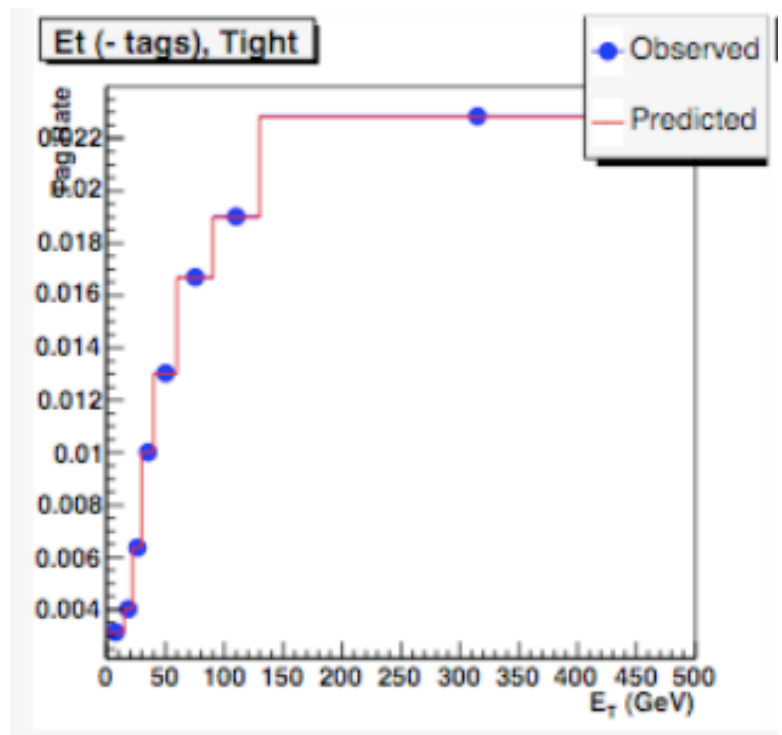


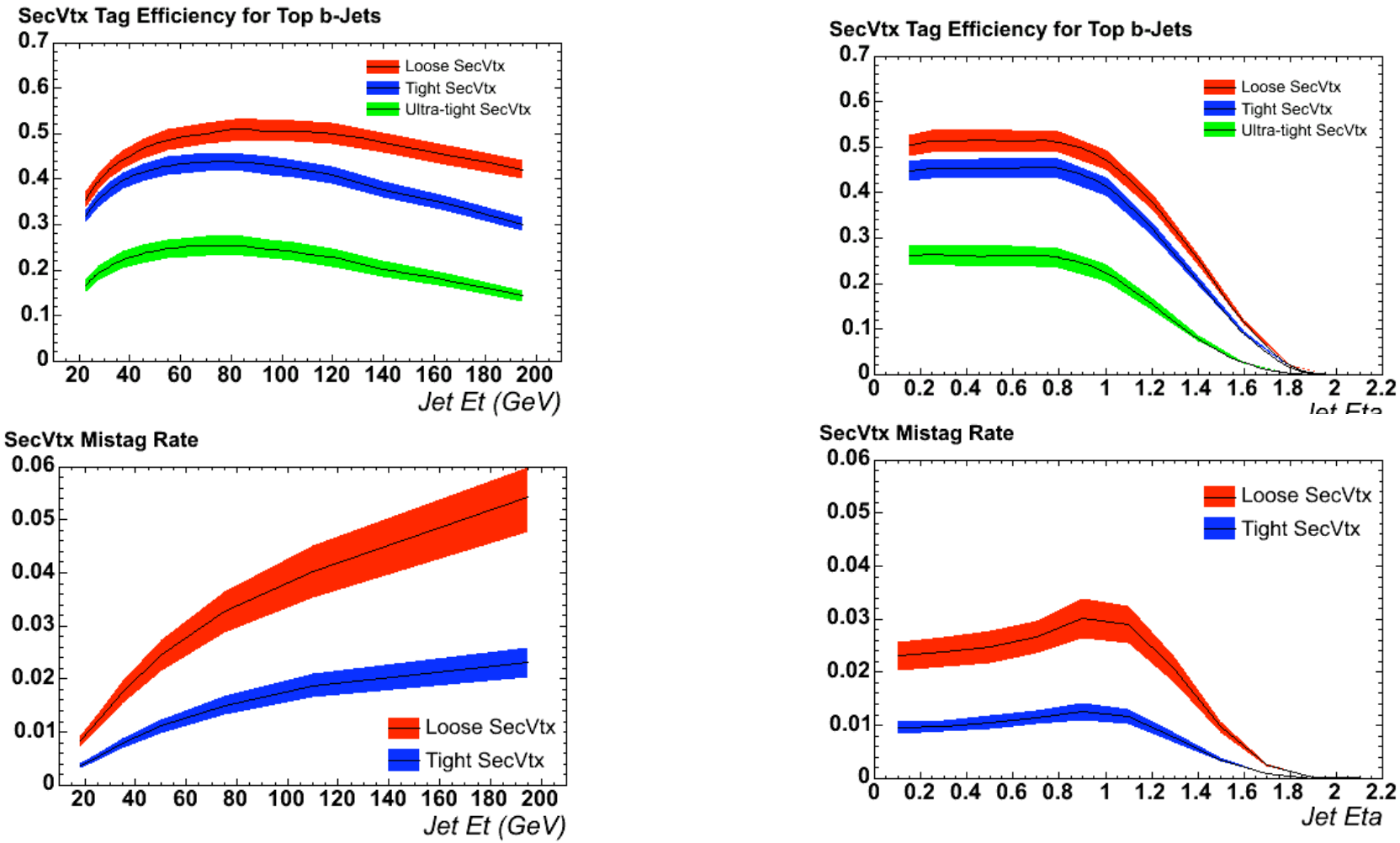
# Characterize the B-tagger: Mistag rate

- Mistag rate measurement:
  - Probability of light quarks to be misidentified
  - Use "negative" tags: $L_{xy}<0$
    - Can only arise due to misreconstruction
  - Need to correct to positive $L_{xy}$
    - Material interactions, conversions etc …
- Determine rate as function of all sorts of variables
  - Apply this to data jets to obtain background



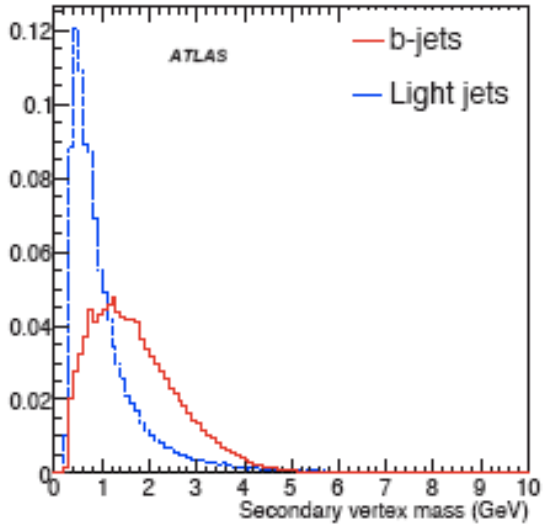



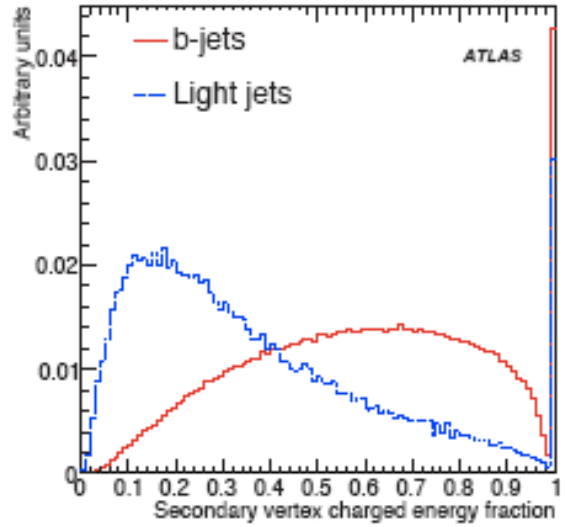
# Final Performance

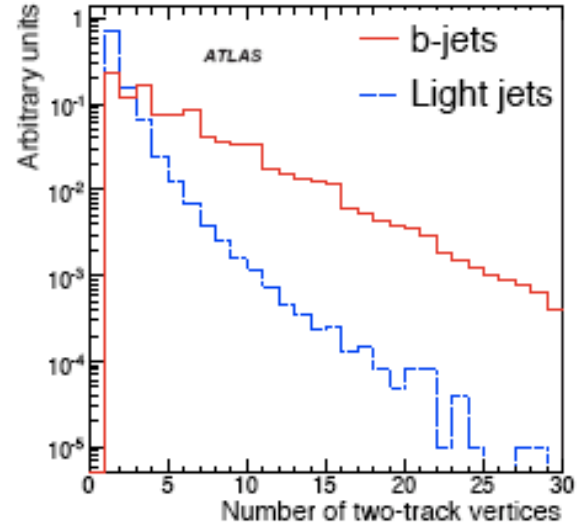



- Choose your operating point depending on analysis
  - Acceptance gain vs background rejection

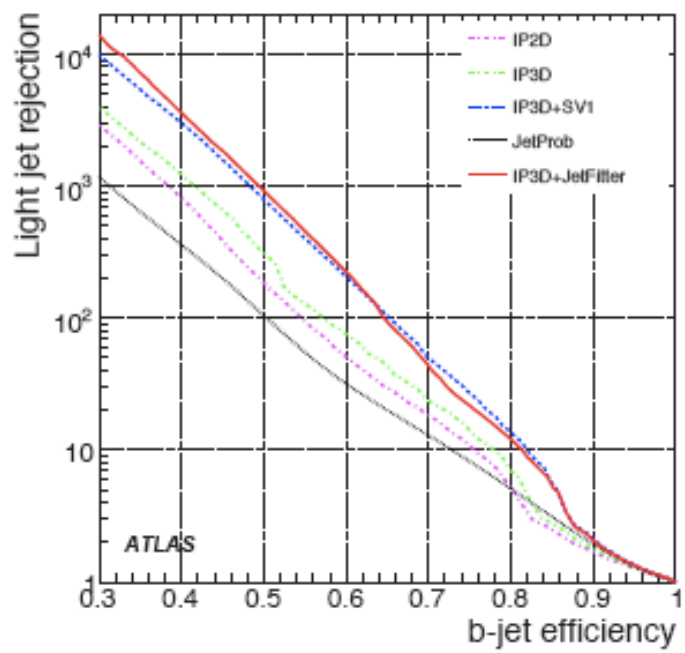
# Improving B-tagging



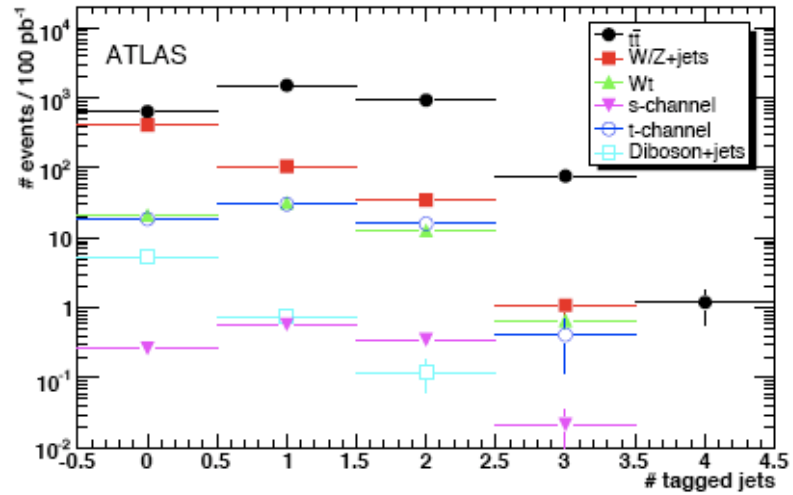



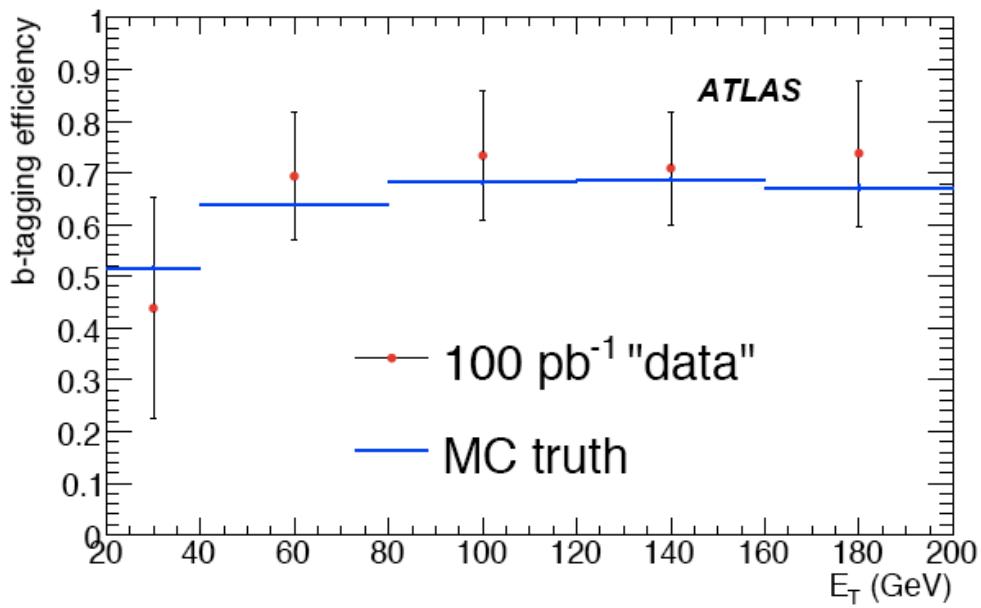

- Use more variables to achieve higher efficiency / higher purity
  - Build likelihood or Neural Network to combine the information
- E.g. for 50% efficiency
  - Mistag rate 0.1%



# Measure b-tag Efficiency in top

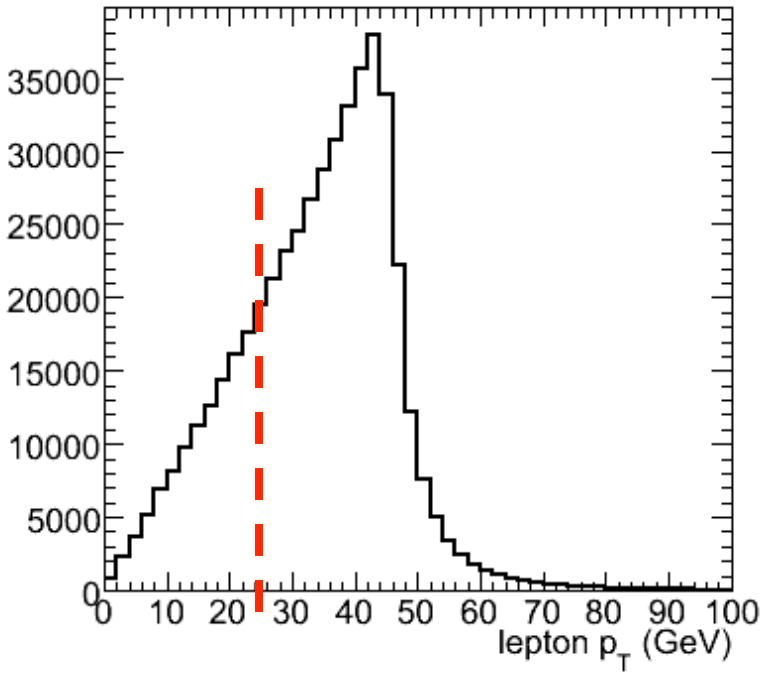
- At LHC high purity of top events
  - Ntop(0-tag)$\propto (1-\varepsilon_b)^2$
  - Ntop(1-tag)$\propto 2\varepsilon_b(1-\varepsilon_b)$
  - Ntop(2-tag)$\propto \varepsilon_b^2$
- => Solve for $\varepsilon_b$
- Backgrounds are complicating this simple picture
  - But it is doable!

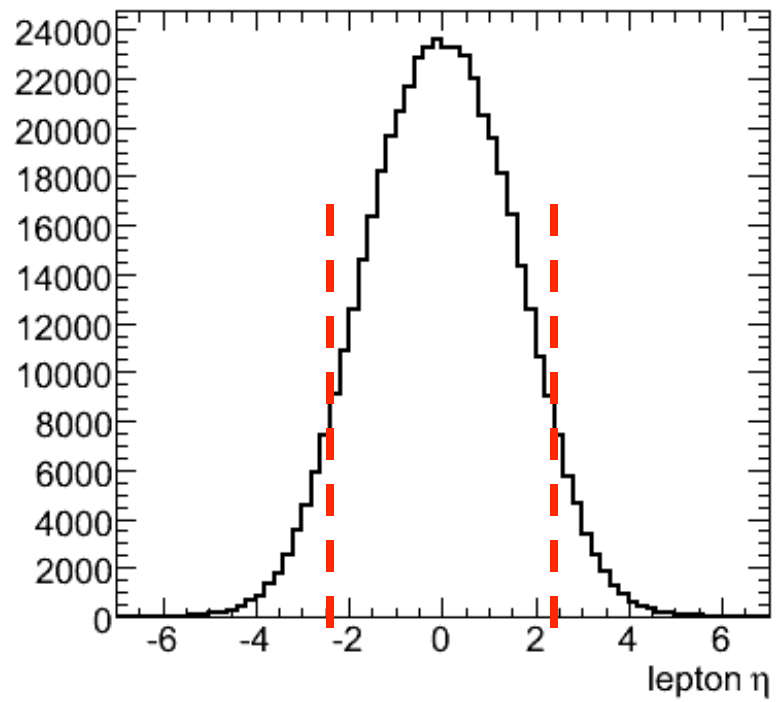


# Acceptance of kinematic cuts

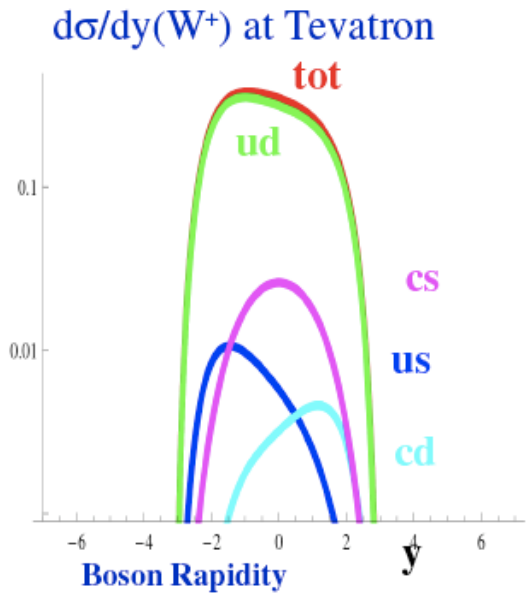
# Acceptance of Kinematic Cuts: Z's

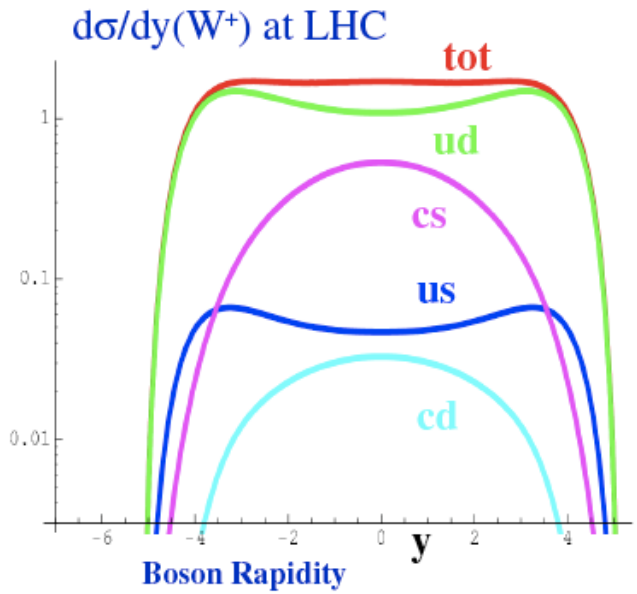


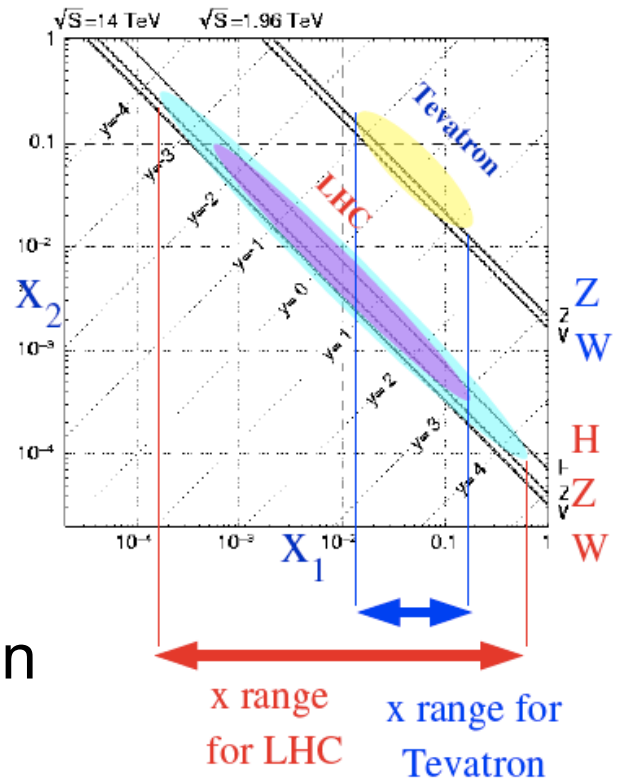




- Some events are kinematically outside your measurement range
- E.g. at Tevatron: 63% of the events fail either $p_T$ or $\eta$ cut
  - Need to understand how certain these 63% are
  - Best to make acceptance as large as possible
    - Results in smaller uncertainties on extrapolation

# Parton Distribution Functions



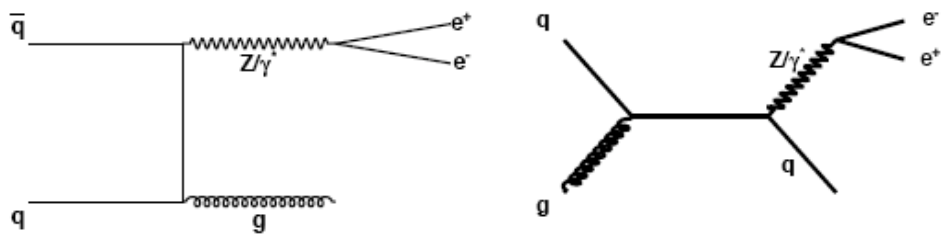



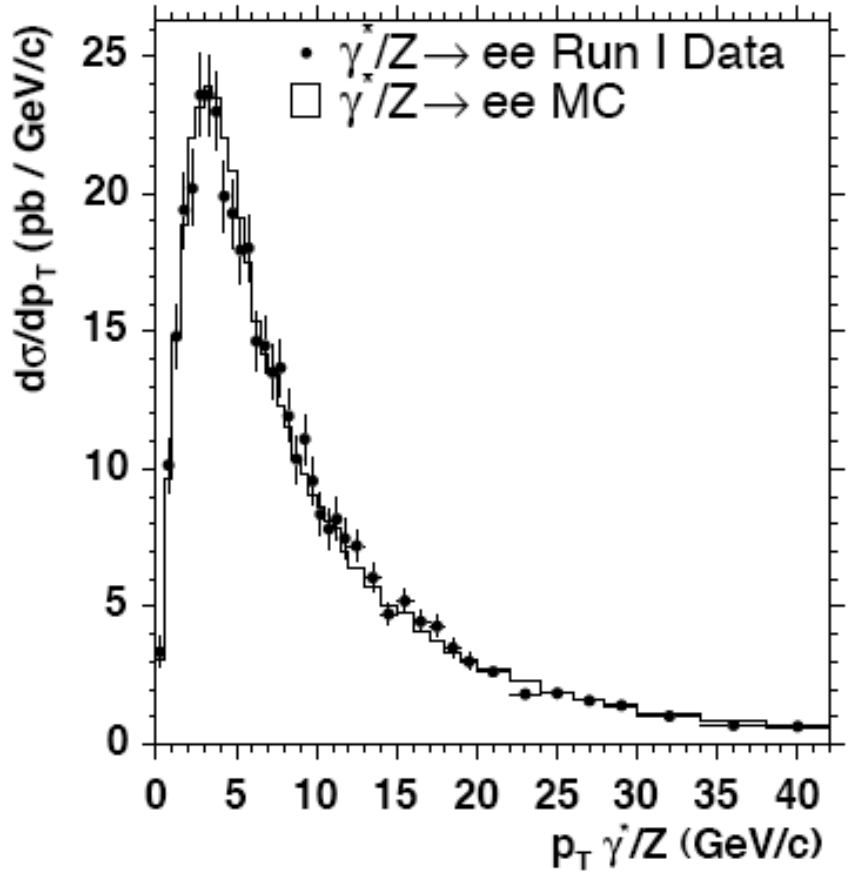




- Acceptance sensitive to parton distribution functions
  - At LHC charm and strange quark densities plays significant role but not well constrained
  - Typical uncertainties on c and s pdf: ~10%
- Can result in relatively large systematic uncertainties

# QCD Modeling of Process

- Kinematics affected by $p_T$ of Z boson
  - Determined by soft and hard QCD radiation
    - tune MC to describe data
- Limitations of Leading Order Monte Carlo
  - Compare to NNLO calculation



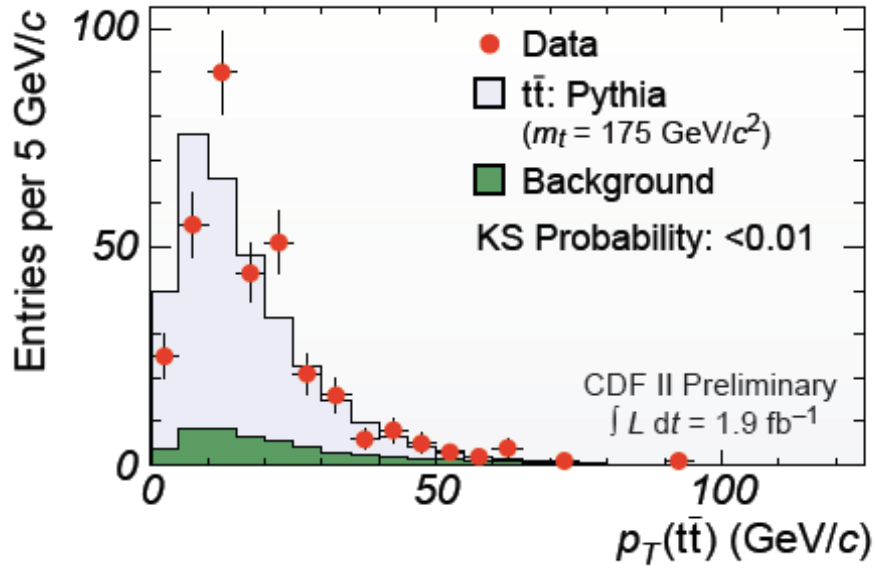

**CDF**

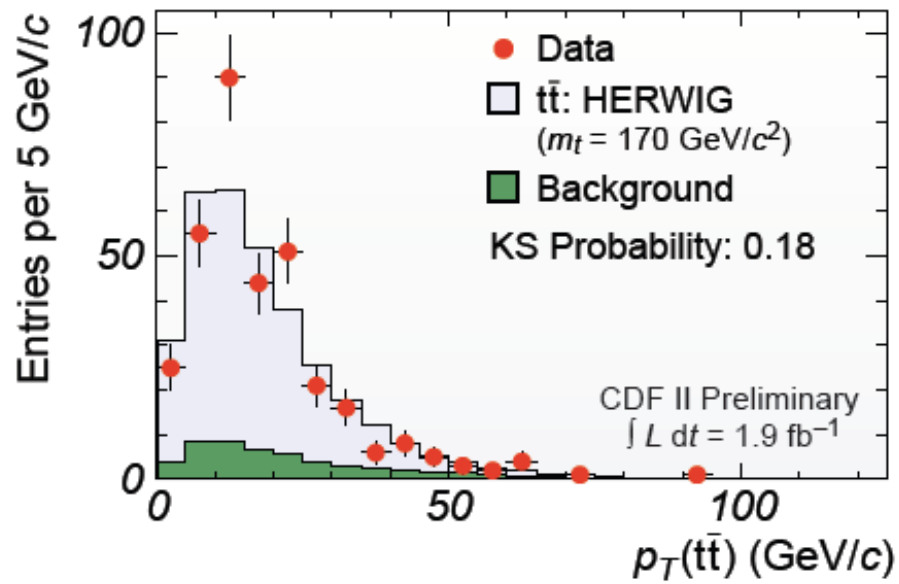
TABLE XII: Central acceptance values for our candidate samples based on $d\sigma/dy$ distributions obtained from both NNLO and PYTHIA simulation.

| Acceptance | NNLO Calc. | PYTHIA | Difference (%) |
|---|---|---|---|
| $A_{W\to\mu\nu}$ | 0.1970 | 0.1967 | +0.15 |
| $A_{W\to e\nu}$ | 0.2397 | 0.2395 | +0.08 |
| $A_{Z\to\mu\mu}$ | 0.1392 | 0.1387 | +0.36 |
| $A_{Z\to ee}$ | 0.3182 | 0.3185 | -0.09 |
| $A_{Z\to\mu\mu}/A_{W\to\mu\nu}$ | 0.7066 | 0.7054 | +0.17 |
| $A_{Z\to ee}/A_{W\to e\nu}$ | 1.3272 | 1.3299 | -0.20 |

# MC Modeling of top

- Use different MC generators
  - Pythia
  - Herwig
  - Alpgen
  - MC @ NLO
  - …
- Different tunes
  - Underlying event
  - Initial/final state QCD radiation
  - …
- Make many plots
  - Check if data are modelled well

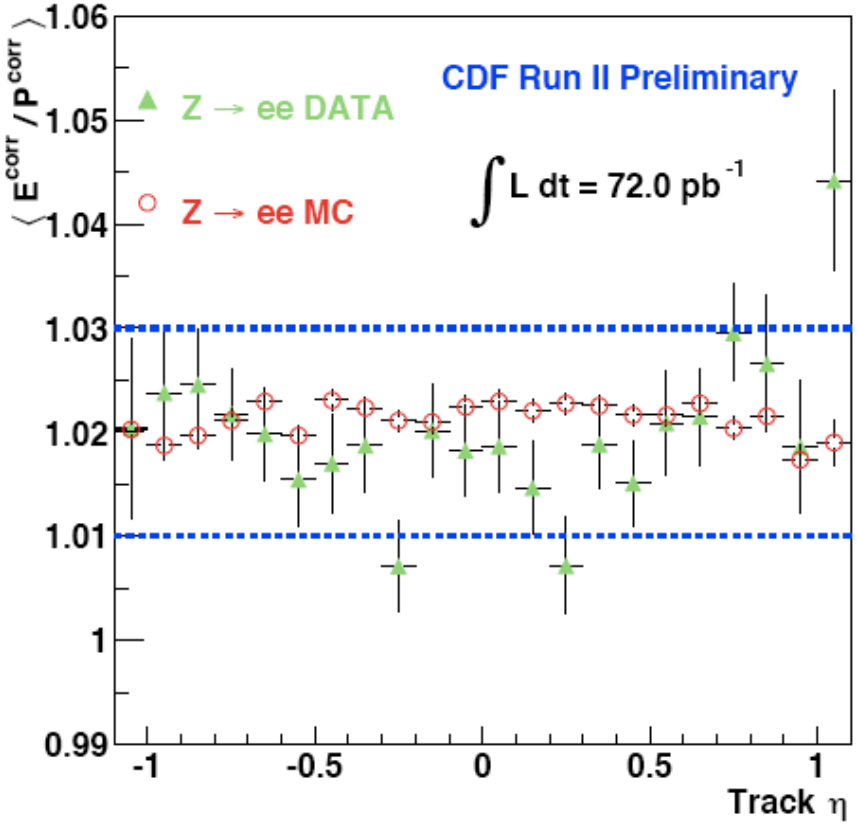


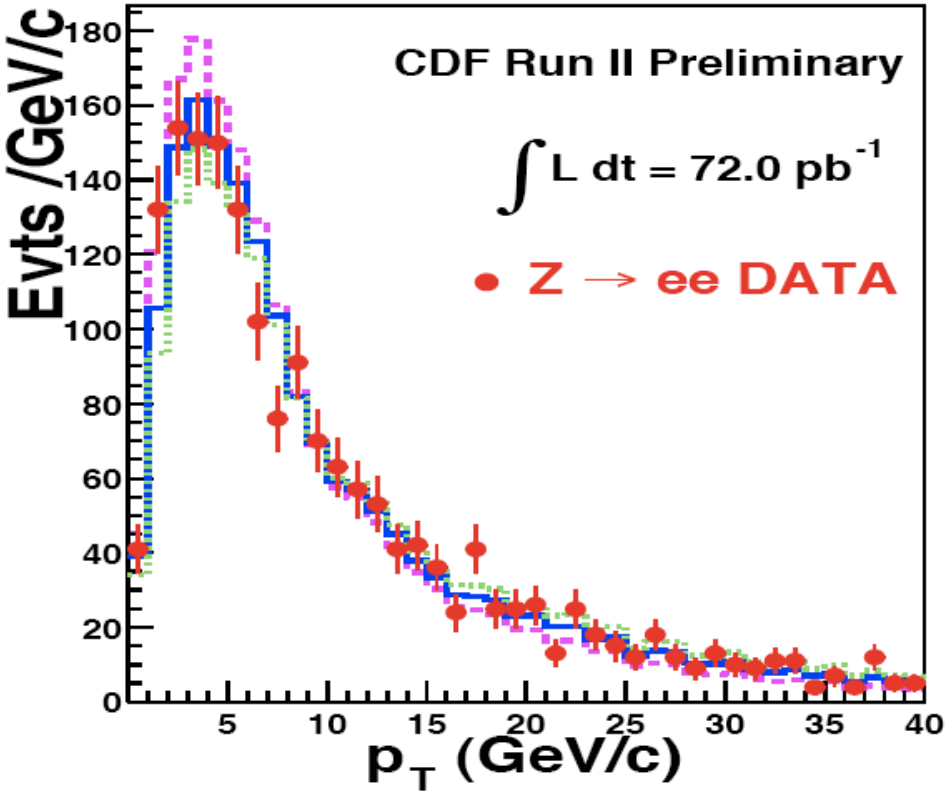



# Systematic uncertainties

- This will likely be >90% of the work you do
- **Systematic errors cover our lack of knowledge**
  - need to be determined on every aspect of measurement by varying assumptions *within sensible reasoning*
  - Thus there is no "correct way":
    - But there are good ways and bad ways
    - You will need to develop a feeling and discuss with colleagues / conveners / theorists
    - There is a lot of room for creativity here!
- What's better? Overestimate or underestimate
  - Find New Physics:
    - it's fine to be generous with the systematics
    - You want to be really sure you found new physics and not that "Pythia doesn't work"
  - Precision measurement
    - Need to make best effort to neither overestimate nor underestimate!

# Examples for Systematic Errors




- Mostly driven by comparison of data and MC
  - Systematic uncertainty determined by (dis)agreement and statistical uncertainties on data

# Systematic Uncertainties: Z and top

**Z cross section** (not all systematics)

| source | variation | $\Delta A_Z$ | $\Delta A_Z / A_Z$ |
|---|---|---|---|
| $E_T^e$ scale | 1% variation | 0.03% | 0.3% |
| $E_T^e$ resolution | 2% extra smearing | 0.02% | 0.2% |
| $p_T^e$ scale | 1% variation | 0.01% | 0.1% |
| $p_T$ modelling | | 0.01% | 0.1% |
| Material | 5.5 % $X_0$ | 0.54% | 4.7% |
| PDFs | reweighting of y | 0.34% | 2.9% |
| **overall** | | 0.64% | 5.5% |

**top cross section**

| Systematic | Inclusive (Tight) | Double (Loose) |
|---|---|---|
| Lepton ID | 1.8 | |
| ISR | 0.5 | 0.2 |
| FSR | 0.6 | 0.6 |
| PDFs | 0.9 | |
| Pythia vs. Herwig | 2.2 | 1.1 |
| Luminosity | 6.2 | |
| JES | 6.1 | 4.1 |
| $b$-Tagging | 5.8 | 12.1 |
| $c$-Tagging | 1.1 | 2.1 |
| $l$-Tagging | 0.3 | 0.7 |
| Non-$W$ | 1.7 | 1.3 |
| $W$+HF Fractions | 3.3 | 2.0 |
| Mistag Matrix | 1.0 | 0.3 |
| Total | 11.5 | 14.8 |

- Relative importance and evaluation methods of systematic uncertainties are very, very analysis dependent

# Final Result: Z cross section

- Now we have everything to calculate the final cross section

TABLE XXXVII: Summary of the input parameters to the $\gamma^*/Z \to \ell\ell$ cross section calculations for the electron and muon candidate samples.

| | $\gamma^*/Z \to ee$ | $\gamma^*/Z \to \mu\mu$ |
|---|---|---|
| $N_Z^{obs}$ | 4242 | 1785 |
| $N_Z^{bck}$ | $62 \pm 18$ | $13 \pm 13$ |
| $A_Z$ | $0.3182\ ^{+0.0039}_{-0.0041}$ | $0.1392\ ^{+0.0027}_{-0.0033}$ |
| $\epsilon_Z$ | $0.713 \pm 0.012$ | $0.713 \pm 0.015$ |
| $\int \mathcal{L}dt$ (pb$^{-1}$) | $72.0 \pm 4.3$ | $72.0 \pm 4.3$ |

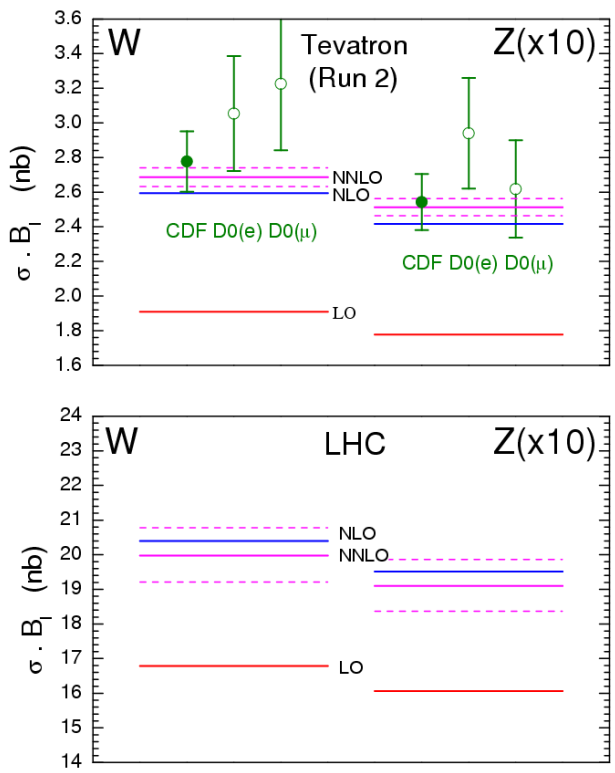
$$\sigma_{\gamma^*/Z} \cdot Br(\gamma^*/Z \to ee) = 255.8 \ \pm \ 3.9(stat.) \ \pm \ {}^{5.5}_{5.4}(syst.) \ \pm \ 15.3(lum.) \text{ pb}$$

**Measurement gets quickly systematically limited**

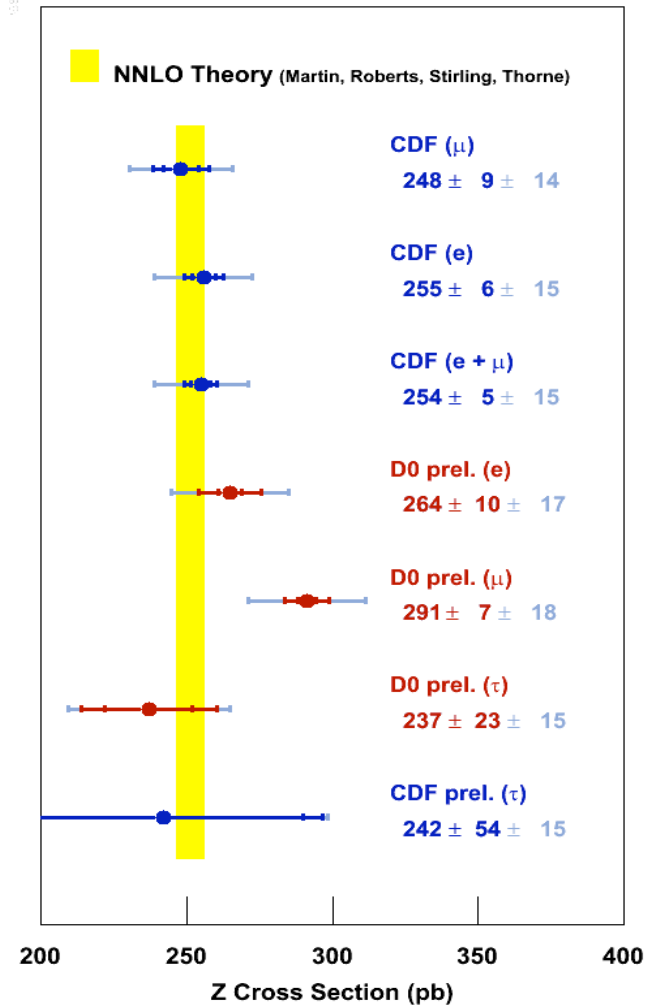
# Comparison to Theory

- Experimental uncertainty: ~2%
- Luminosity uncertainty: ~6%
- Theoretical uncertainty: ~2%

$\sigma_{Th,NNLO}$=251.3±5.0pb
(Martin, Roberts, Stirling, Thorne)

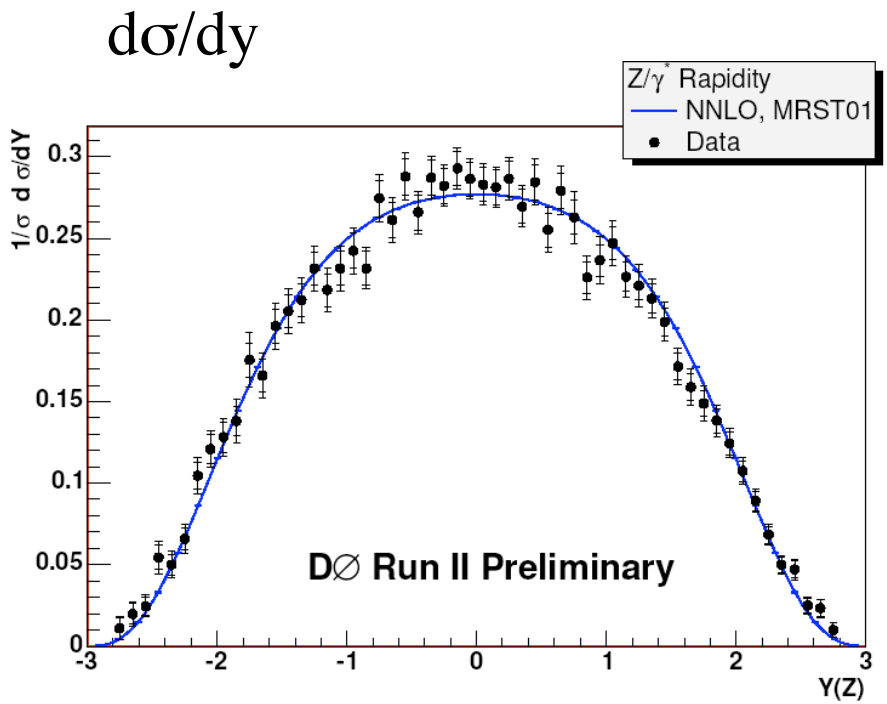


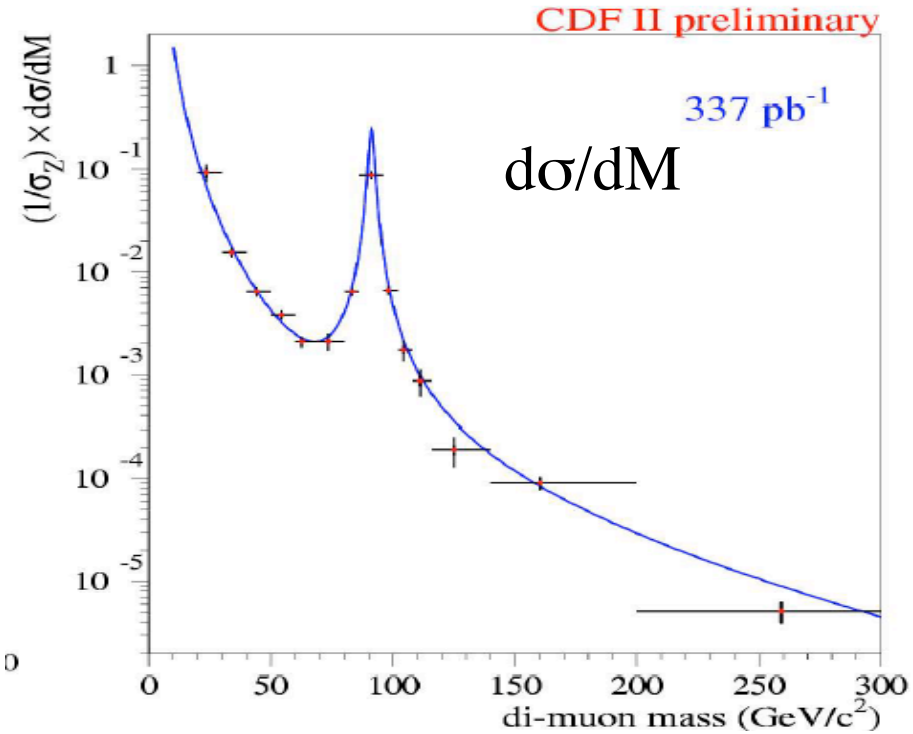

partons: MRST2002
NNLO evolution: Moch, Vermaseren, Vogt
NNLO W,Z corrections: van Neerven et al. with Harlander, Kilgore corrections




- Can use these processes to normalize luminosity absolutely
  - However, theory uncertainty larger at LHC and theorists don't agree (yet)

# More Differential σ(Z) Measurements

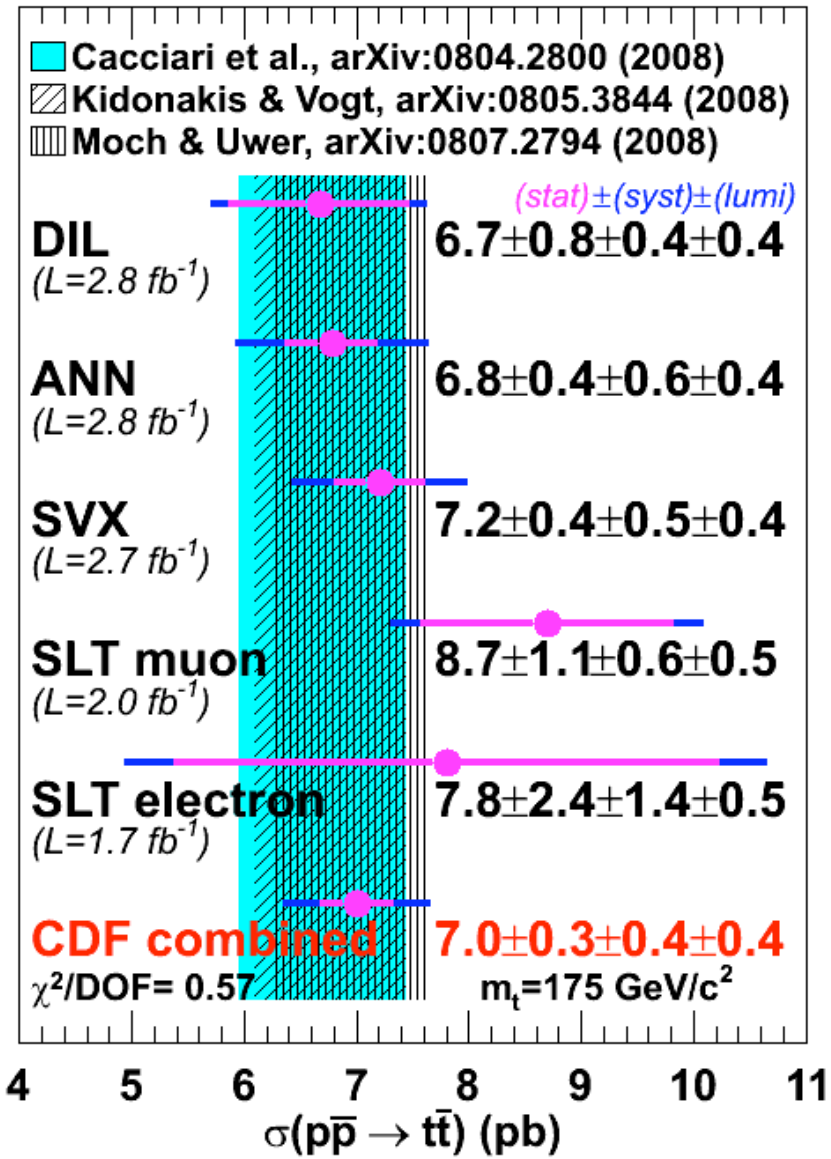



**Differential measurements in principle very similar**

**But now need to understand all efficiencies as function of y or mass**

# Final Results: Top Cross Section




- **Tevatron**
  - Measured using many different techniques
  - Good agreement
    - between all measurements
    - between data and theory
  - Precision: ~9%

- **LHC:**
  - Cross section ~100 times larger
  - Measurement will be one of the first milestones (already with 10 $pb^{-1}$)
    - Test prediction
    - demonstrate good understanding of detector
  - Expected precision
    - ~4% with 100 $pb^{-1}$

# Conclusions of 1st Lecture

- Cross section measurements require
  - Selection cuts
    - Optimized to have large acceptance, low backgrounds and small systematic uncertainties
  - Luminosity measurement
    - Several methods of varying precision
  - Trigger
    - Complex and critical: what we don't trigger you cannot analyze!
  - Acceptance/efficiency has many subcomponents
    - Estimate of systematic uncertainties associated with each
    - Dependence on theory assumptions and detector simulation particularly critical
    - Minimize extrapolations to unmeasured phase space
  - Background estimate
    - See final lecture
- Systematic uncertainties are really a lot of work