V+jets as a Standard Candle

Gustaaf Brooijmans

Columbia University
In the City of New York

YETI 09

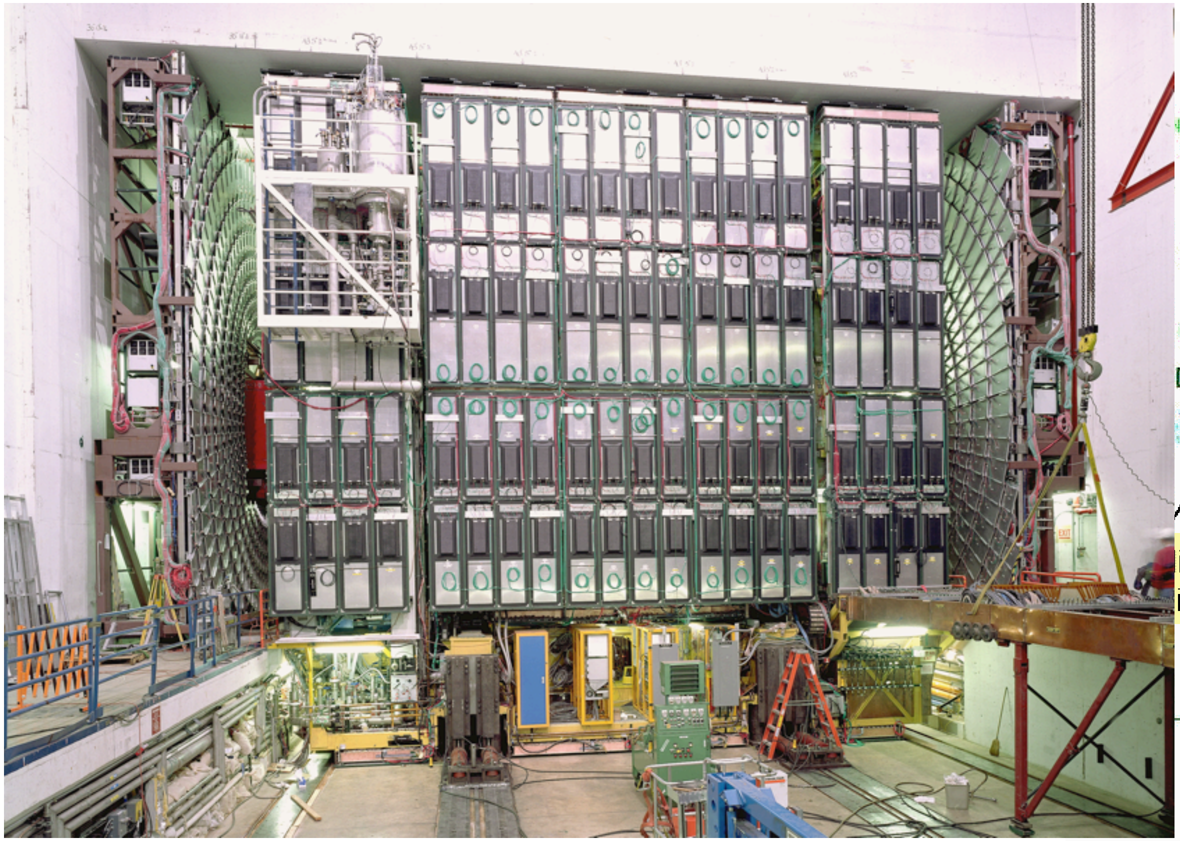
Outline

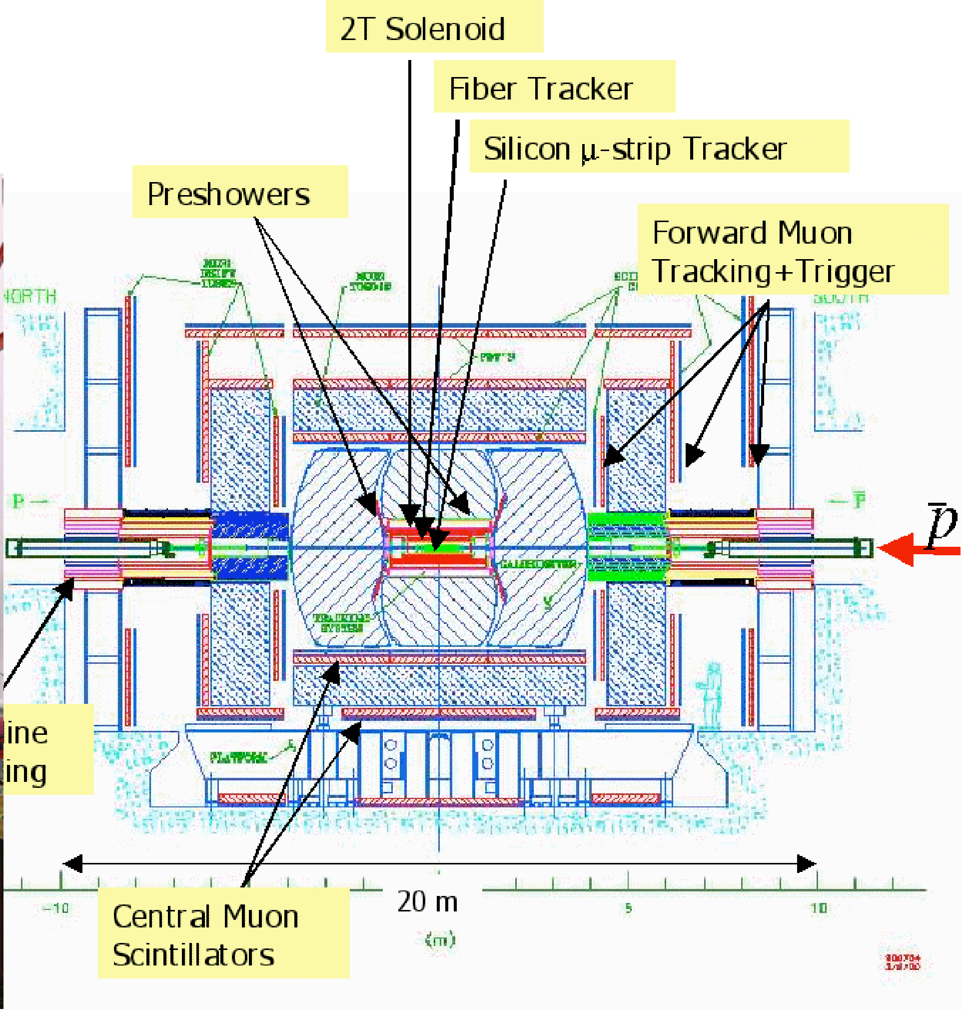
- Experiments
- Simulation
  - Why, how, how good?
- Generators & Data
  - V+jets at the Tevatron
- Top, LHC & new physics
- Summary

Experiments

DØ

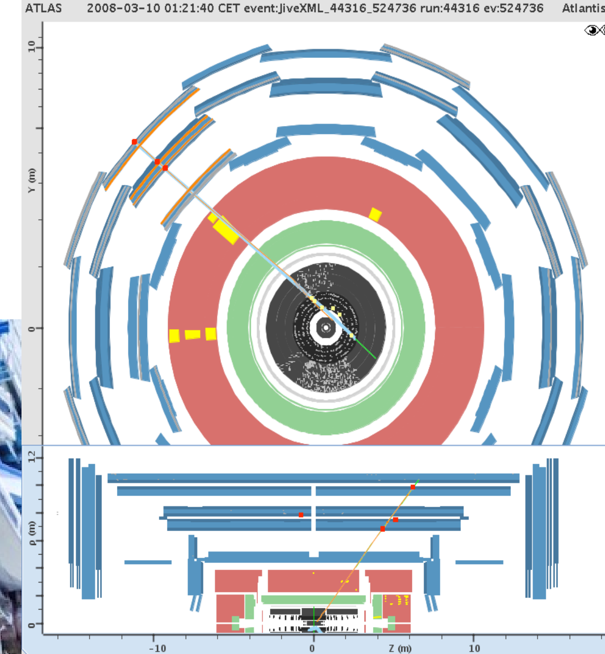
- Typical collider detector: aim to identify & measure all particles coming from the collision
  - ~6000 tons of material....

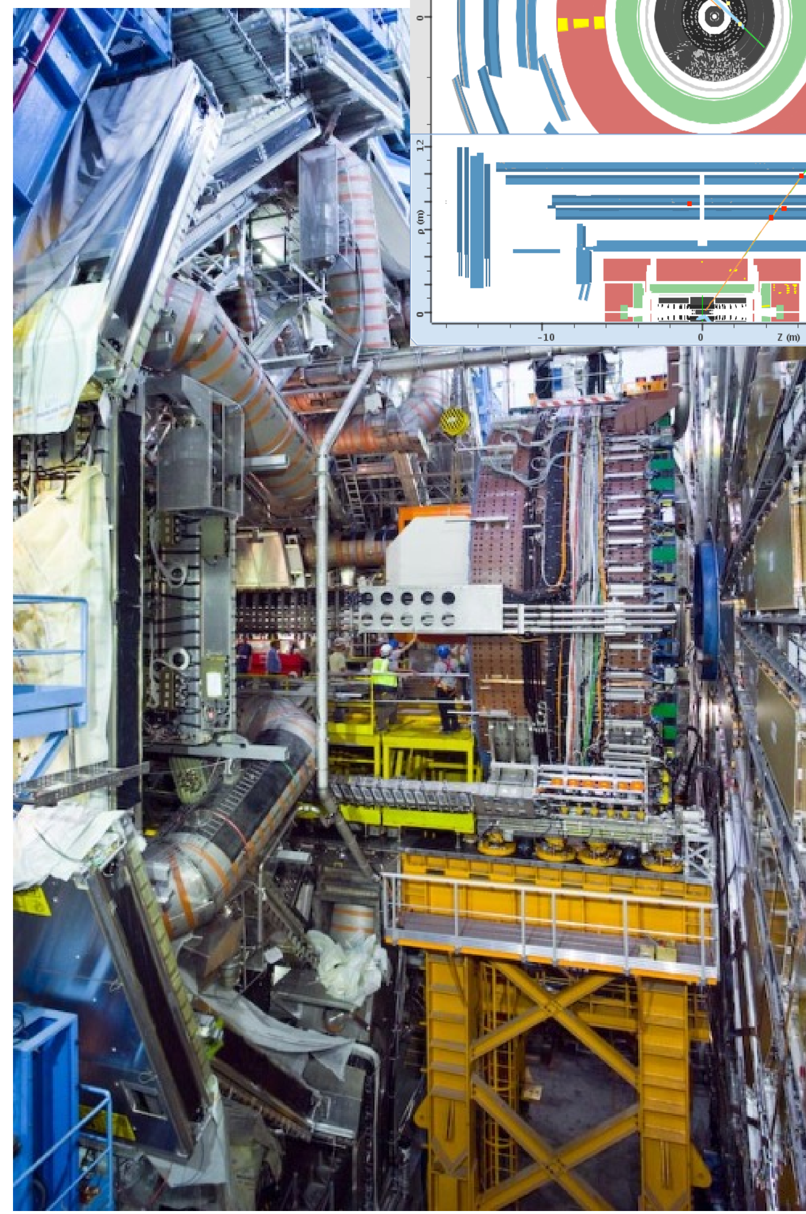


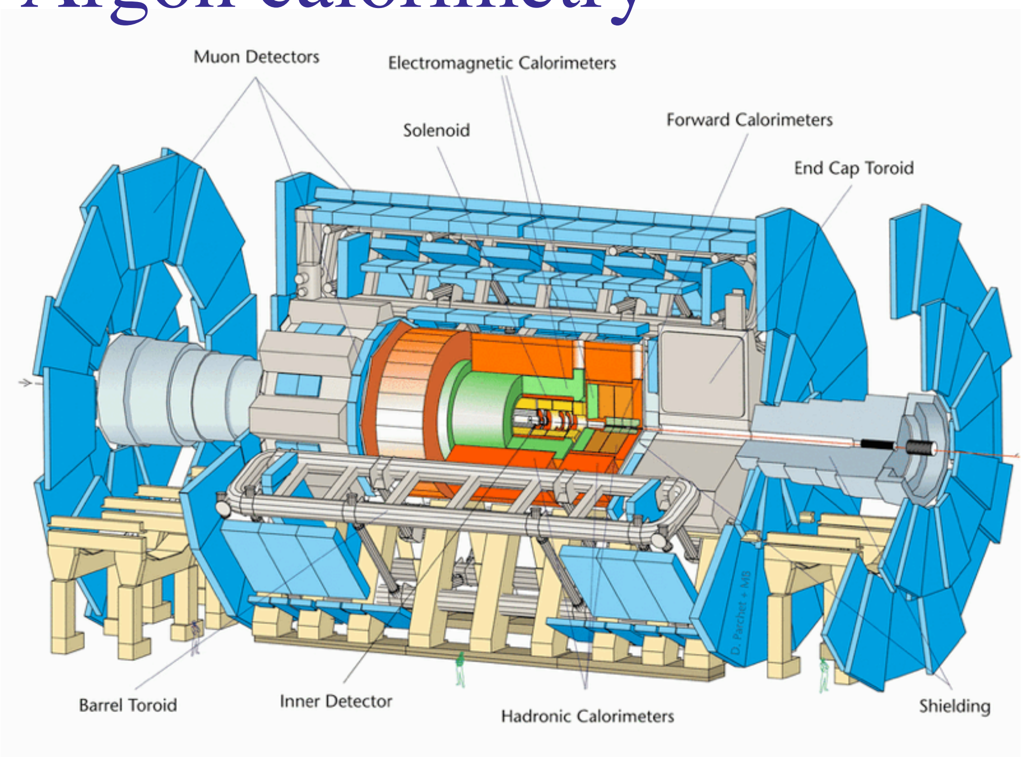


ATLAS

- Installed in-situ (installation essentially complete)
- Precise tracking
- Two magnet systems, liquid Argon calorimetry

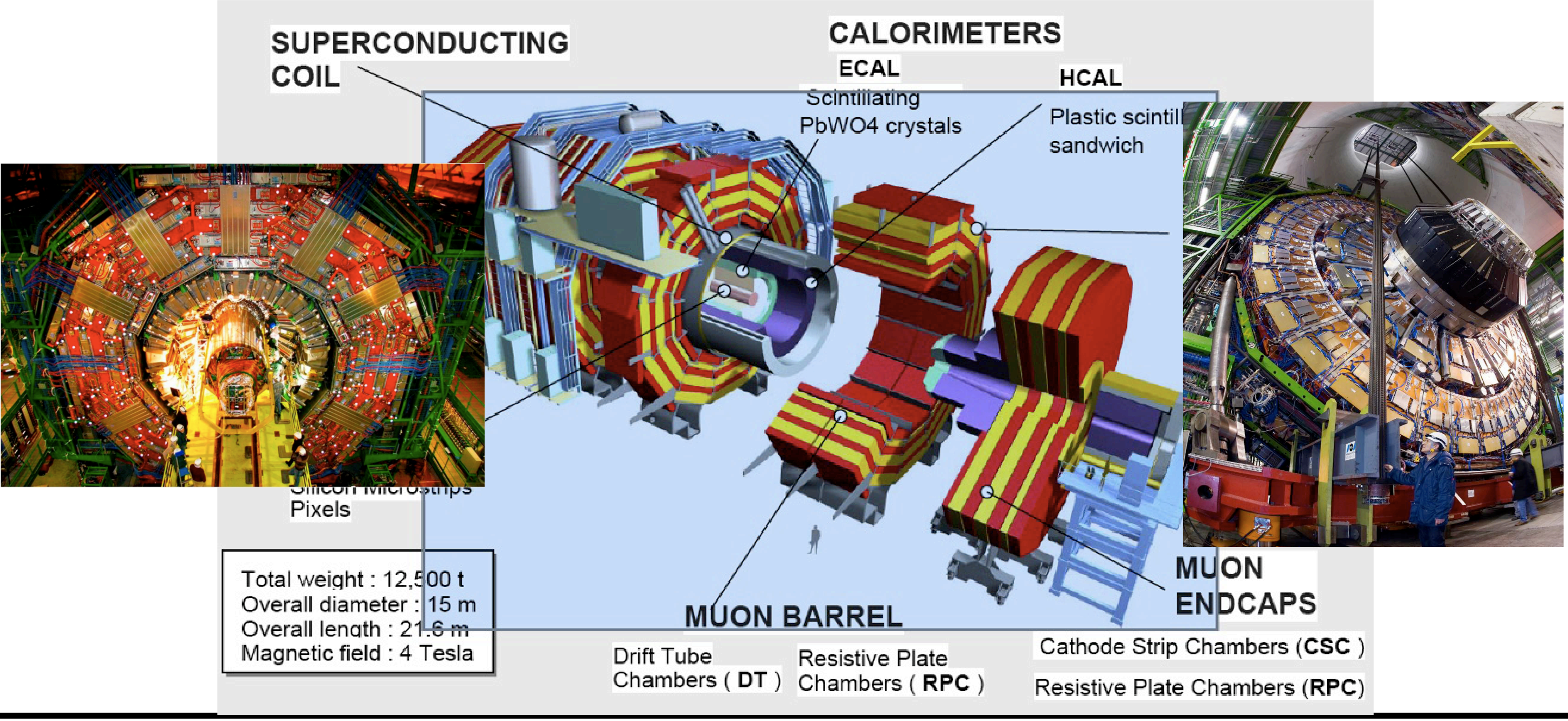




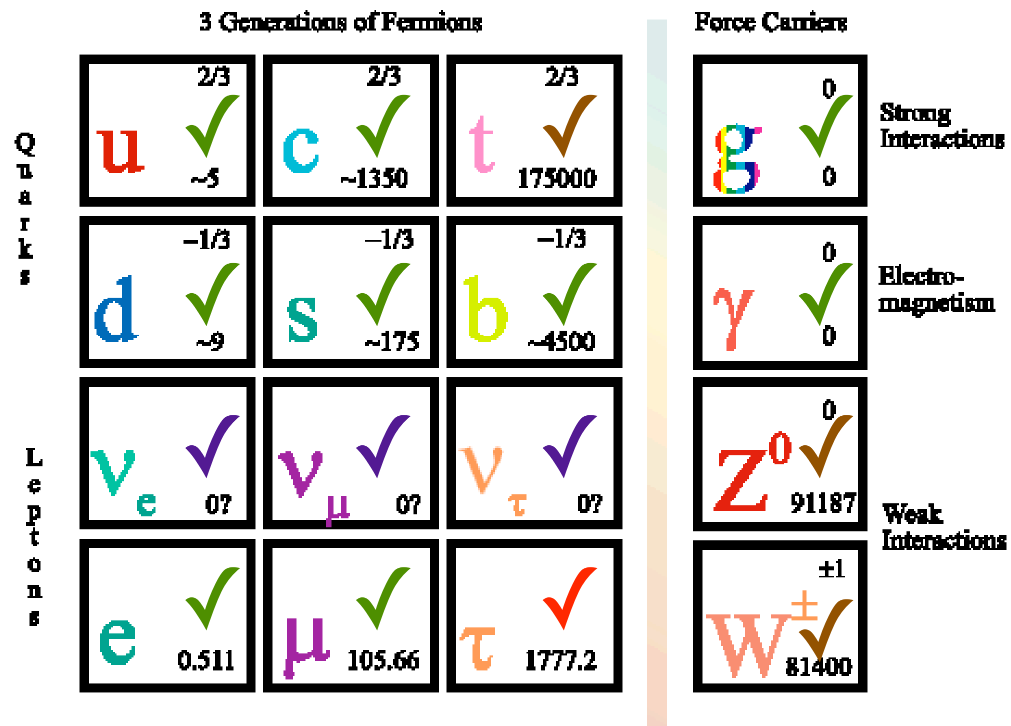


CMS

- Most of the assembly done on the surface, then lower in "slices"
- Key features: crystal calorimeter, all-silicon tracker



Detecting Particles



Masses are in MeV

√ : Detect with high efficiency

√ : Detect by missing transverse energy

√ : Detect through decays: $t \rightarrow Wb$, $W/Z \rightarrow$ leptons

Simulation: Technical

Monte-Carlo Simulation

- Our detectors produce lots of little electrical & optical signals which we try to translate back to Feynman diagrams
- It's much harder to unfold detector effects than to simulate Feynman diagrams and propagate the physics:
  - Generate physics process (according to differential x-section)
  - Parton shower: quarks & gluons radiate more quarks & gluons
  - Hadronization & decay: left with "stable" particles
  - Step each particle through the detector, simulate interactions with material according to measured cross-sections
    - Simulate little electrical & optical signals

Experimental Duality

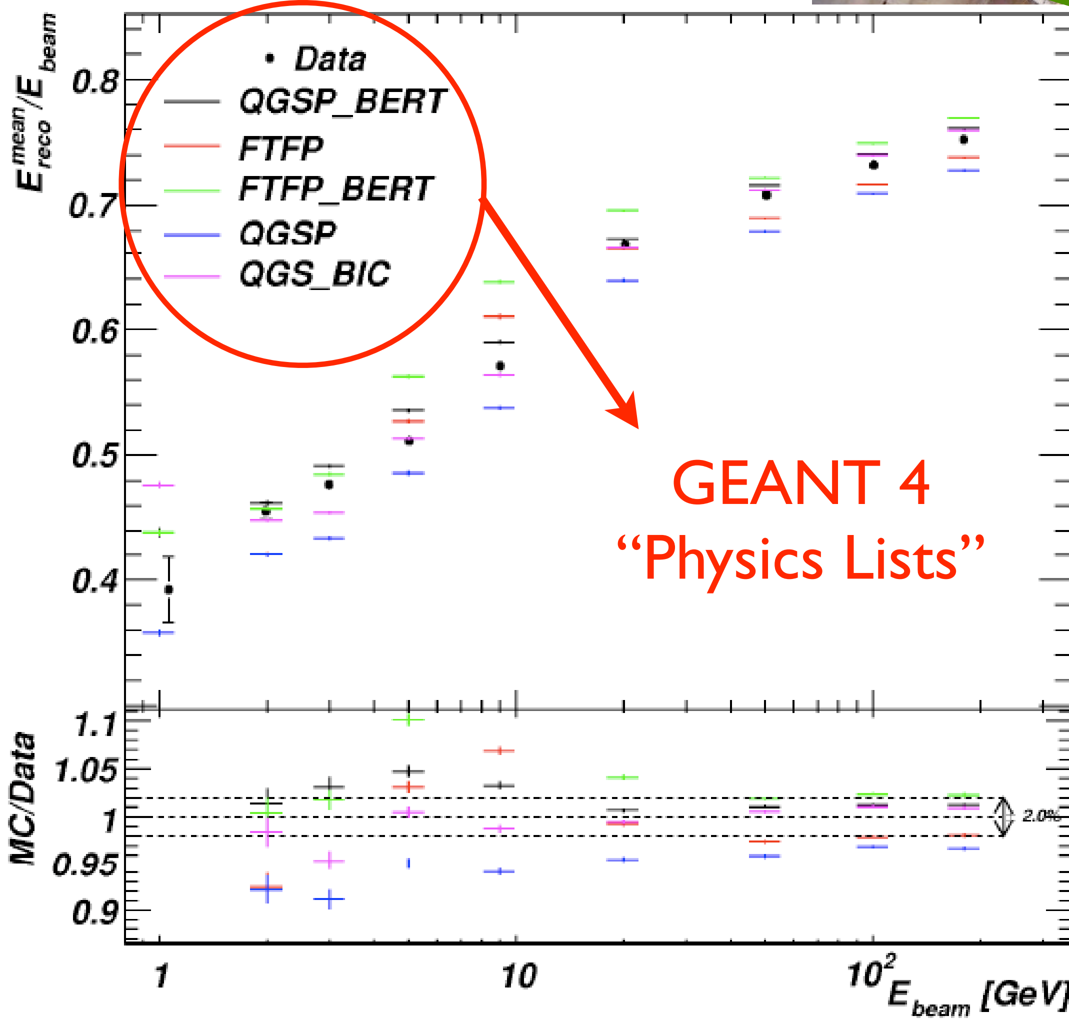
- Real Life
  - Physics event ("hard scatter")
  - ➡ (Parton shower)
  - ➡ Interactions of particles in detector lead to more particles and leave (tiny) electrical or optical signals (with bias)
  - ➡ Record some (biased) fraction of events
  - ➡ Pattern recognition to reconstruct showers, tracks (with unavoidable bias)
  - ➡ Infer physics

Simulation

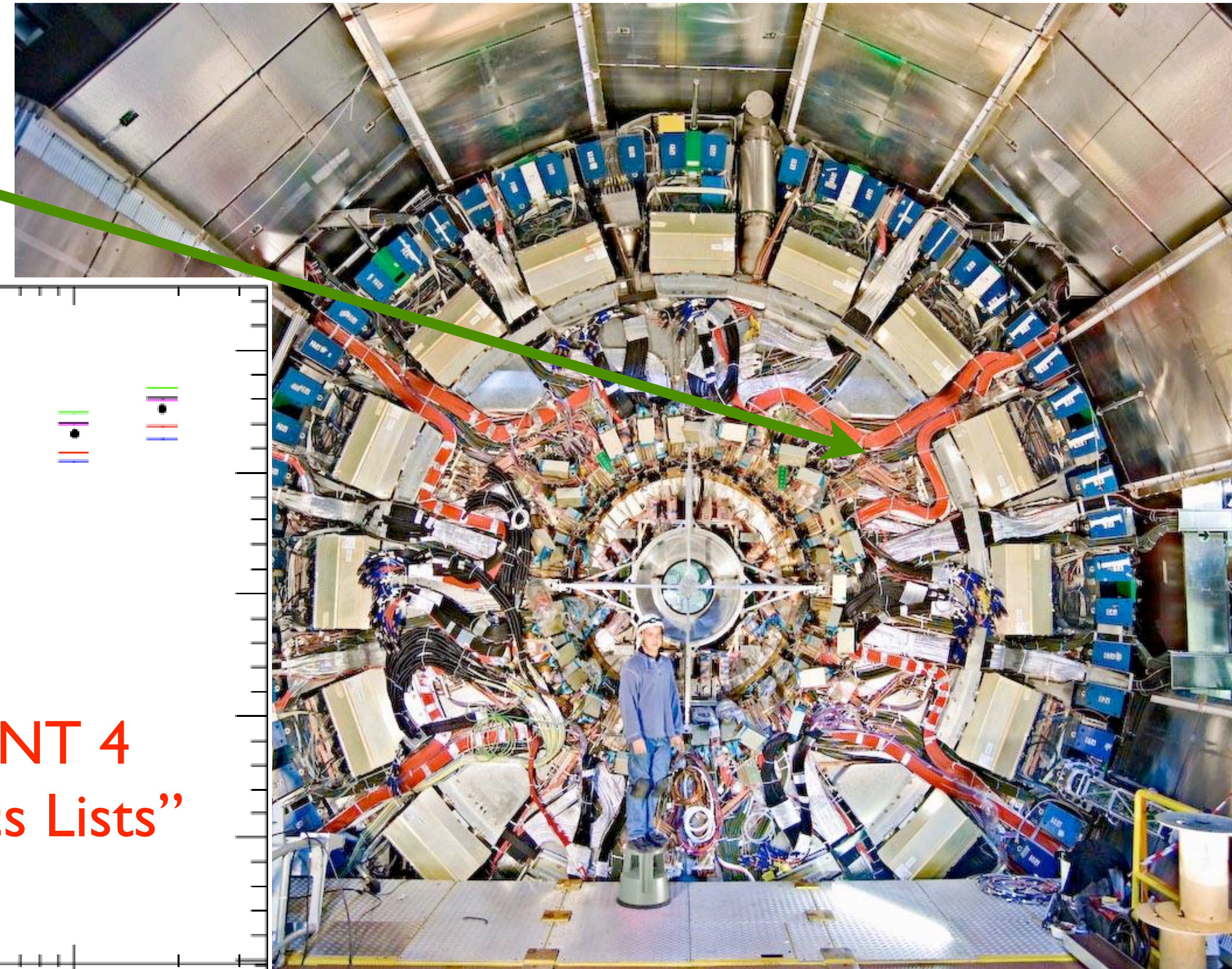
- Generator
- Pythia/Herwig (Matching)
- Geant/ Parametrized
- Trigger Sim.
- Reconstruction
- Analysis

Detectors & Interactions

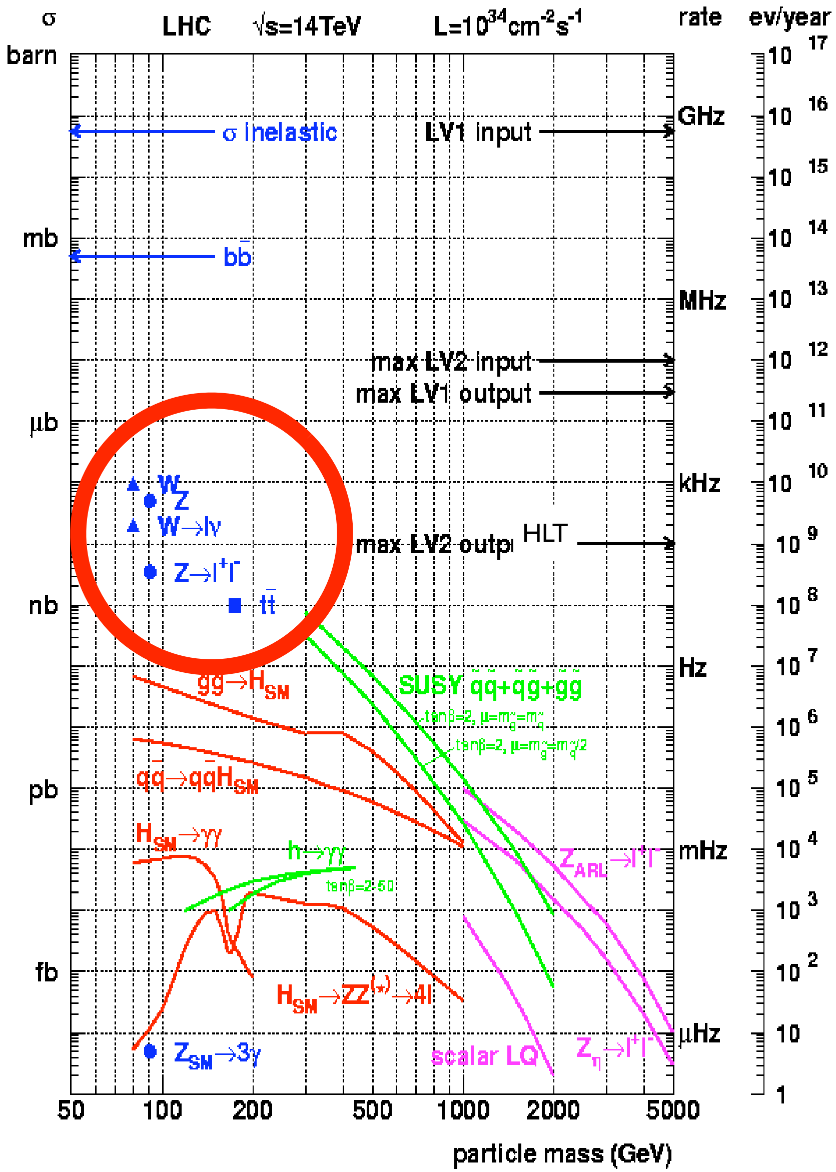
Not so easy to simulate exactly...



GEANT 4
“Physics Lists”



Physics Validation

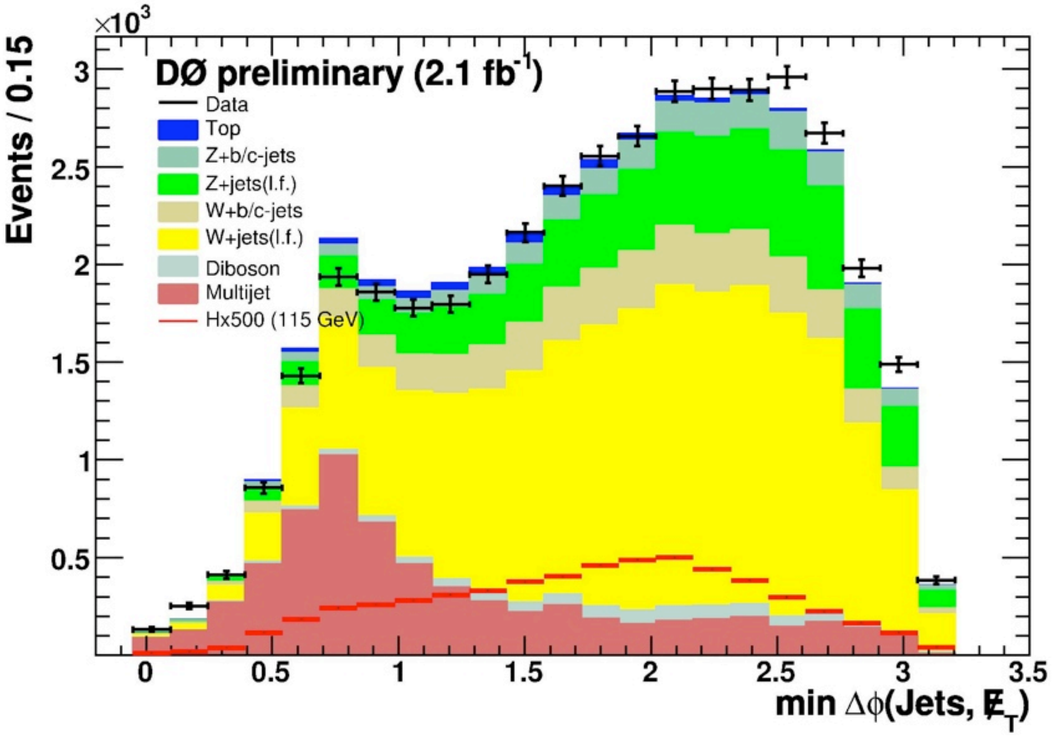


- Need to verify the accuracy of the simulation
- Good processes?
  - Close to things we're interested in measuring (produce all objects)
  - Large enough statistics so problems aren't hidden in error bars
  - Calculable & triggerable!
- W/Z production (+jets)
  - And, at the LHC, top pairs

Correcting for Biases

- To infer the physics, we need to correct for all the biases that have been introduced by the "event selection" (incl. detector, trigger, reconstruction)
  - Simulate all contributing processes and put them through the full simulation chain (which has its own biases)
  - Add all contributions, compare to data

Not always easy to determine why data doesn't agree with "expectation" (or: unfolding is hard!)



Biases, II

- In practice:
  - Determine MC efficiencies (easy)
  - Determine data efficiencies (not so easy)
  - Apply data/MC scale factors to MC
    - Generally depend on p^T, η, ϕ, ...
- For trigger and reconstruction separately, and efficiencies can be very dependent on topology
- Many different corrections that need to be applied
  - No too hard to make a mistake....

Reproducibility

- Corrections we apply are not always small: 10-20% effects rather common
  - Uncertainties on these corrections are a big issue - often major contributions to systematic uncertainties

➡ In addition to reproducibility by another experiment, require reproducibility within a single experiment

- Implies strict requirements on datasets used (some corrections applied centrally, others analysis-dependent), software used

➡ All datasets, including MC starting point, i.e. generator, produced by strictly controlled software

- In practice:
  - Getting new generator code into a software release is hard

$\mathcal{O}$(1 year)!

    - Code needs to be put into a release "in development"
    - Code needs to be validated (run events through, check things don't crash, make plots showing things are ok)
    - Release needs to go in production (requires release validated for many other things as well)

  ➡ Not necessarily easy to "try other models/versions"

  - Matrix element approach through LHEF should allow to reduce this
    - But (reproducible) inputs need to be archived somewhere
    - So getting .lhe files from a favorite theorist is not quite good enough....

Simulation: Physics

A Word About Event Generators

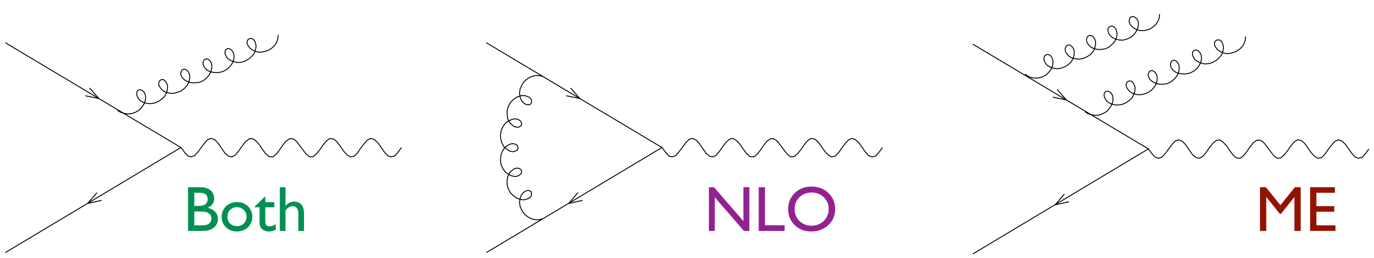
- We distinguish four kinds of Monte Carlo generators
  - "Calculators" (often NNLO) do not actually generate events, they just calculate some (limited) distributions
  - Traditional 2 → 2 generators: LO, e.g. $q\bar{q} \to e^+e^-$
    - Include parton shower, i.e. QCD radiation, and hadronization to jets
  - "Matrix Element" 2 → n (n < 9): LO, e.g. $q\bar{q} \to e^+e^-jjjj$
    - Require matching to parton shower to avoid double counting
  - NLOwPS 2 → 2 generators: include NLO corrections
    - I.e. in a sense they are 2 → 2 & 3 with virtual corrections

Matching

- The problem for "matrix element" (i.e. LO 2→n, n<9) generators:
  - If generate e.g. W+0j, W+1j, W+2j, W+3j, W+4j separately, then run parton shower, can get double counting of jets from parton shower and matrix element
  - So need to remove/suppress the extra events, two procedures
    - MLM (kind of ad-hoc)
    - CKKW (state of the art, but new & ~hard to use)
- Matching is, at this point, an art rather than a science
  - Will hopefully be ~solved by 2009

ME vs NLOwPS

- NLOwPS = NLO generator with parton shower
  - MC@NLO, PowHEG: no matching systematics!
  - Calculates *all* NLO diagrams (including virtual contributions), but *only* those



  - Use NLOwPS at low multiplicity, ME at high multiplicity
    - So all is good? For Higgs study we just need to use ME

Monte Carlo @ Tevatron

- A short word of history
  - madevent has been in use in top mass analysis since mid-late 90's (more later)
  - Start of Tevatron Run II (2001):
    - Pythia ("old shower") and herwig were the workhorses
      - Given Run I statistics, these were ok
  - ~2002, alpgen becomes available for users
    - For experimenters, need interface with parton shower
    - Double counting (i.e. "matching") comes up, and solution
- Developments happened during Run II

- ~2007 sherpa with all “required” features (radiation etc.)
- ~2007 madevent-pythia matching as well, MLM
  - (CKKW advertised but not really available)
- Late 2004 pythia with “new” p^T-ordered shower
  - Not used at Tevatron AFAIK, used in ATLAS

- ~2004 Run II statistics establish value of ME codes
  - ~million leptonic W’s, ME needed to cover phase space
- ~2007 increased stats → increased sensitivity
  - Million<u>s</u> of leptonic W’s, start seeing “issues” with MEs
  - Concurrent with theoretical studies of matching

Basic Physics Analysis

- Devise a set of selection cuts geared towards improving S/B
  - Often two sets: “loose” (control sample) and “tight”
- Determine the resulting sample’s composition
  - For high-p^T physics at a hadron collider:

Increasing difficulty ↓

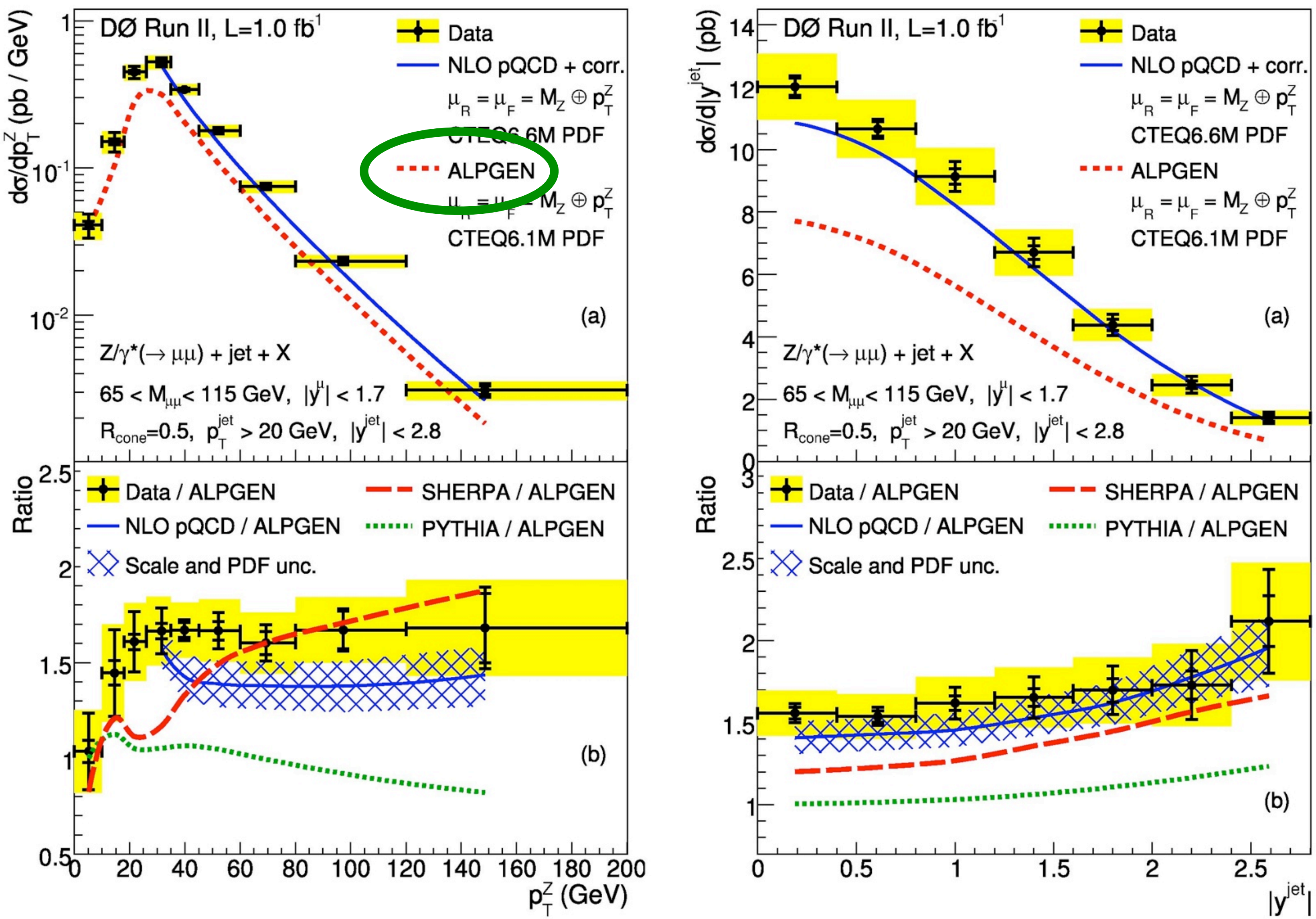
- Diboson from MC (usually small, + “trust” MC)
- At the Tevatron, top from MC (“large” statistical uncertainties)
- Z+jets from data & MC (“easy” to get a clean sample)
- QCD multijet from data
- W + jets from MC, but ....

Z (→ℓℓ) + jets

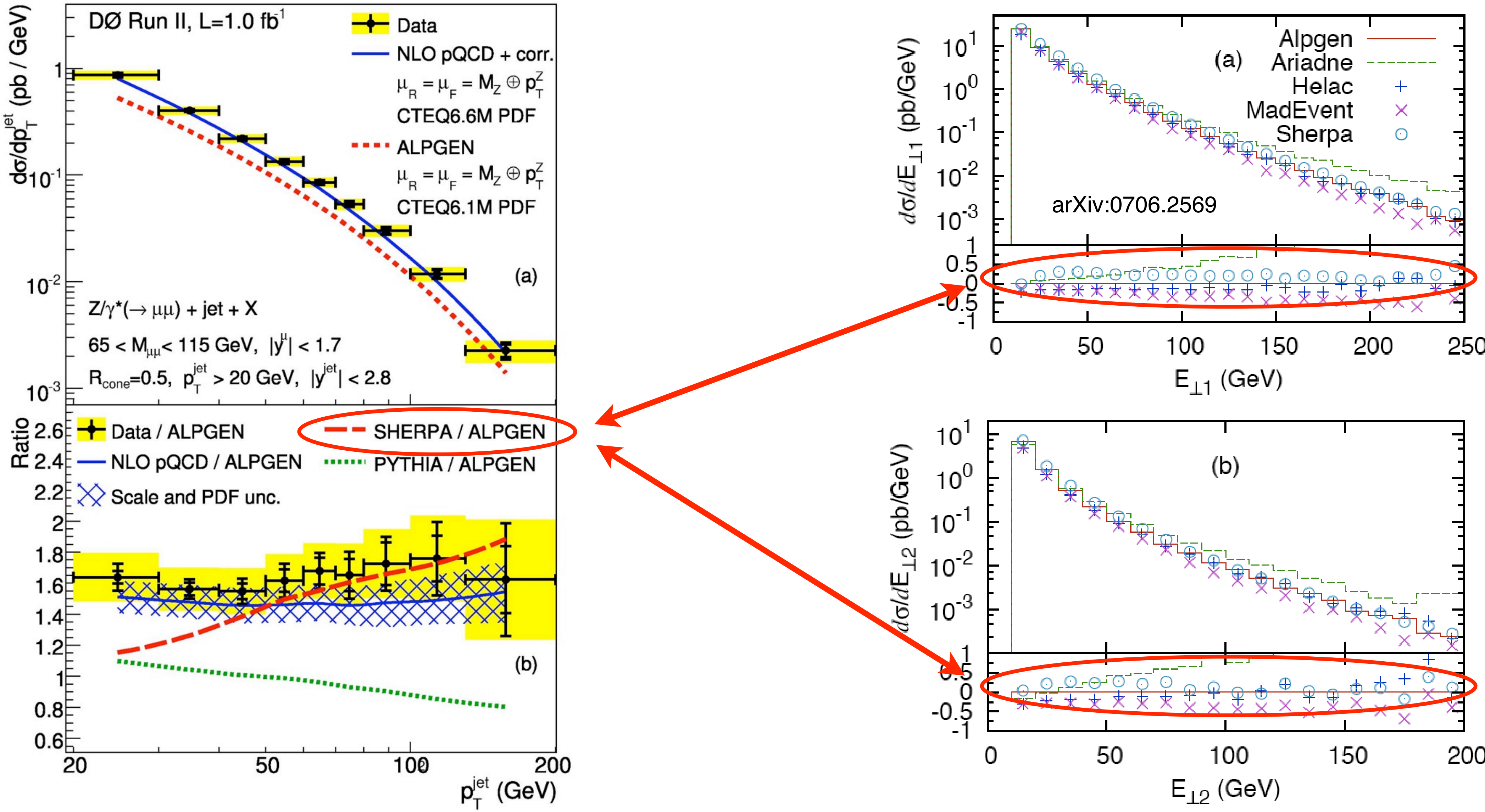
- Can get a clean sample, check if our simulation reproduces the data

⇒ yes!

(but not good enough)



Using MC Generators



- Clearly, ratio alpgen/sherpa depends on who runs the generator..... (there are many parameters!)

Correction Factors

- Of course, the ME's are LO, so "K-factors" needed
  - Different ones for heavy flavor etc..... convention to avoid confusion....

• **K-factor is purely theoretical, and denotes a (N)NLO/LO ratio of cross sections;**

• **K'-factor is also theoretical, and denotes a (N)NLO/LL ratio of cross sections.**
According to Steve, ALPGEN cross sections are Leading Log;

• **S-factor is empirical, and comes on top of K or K' to bring MC in agreement with data. MC should be initially normalized to luminosity, and all correction (a.k.a. scale) factors should be applied (trigger, ID...);**

• **HF-factor is, in principle, theoretical, but in practice only theory inspired. It tells you by how much heavy flavor production should be increased, on top of K or K', and possibly S;**

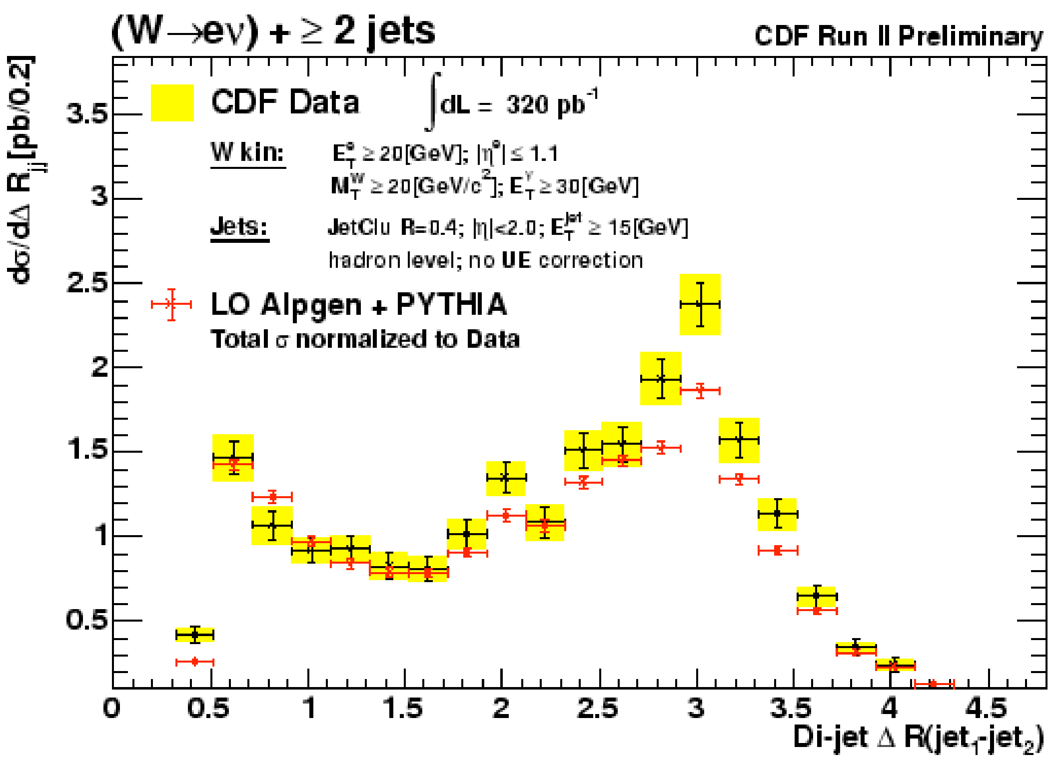
• **S_HF-factor is empirical, and comes on top of K or K', S, and HF, to bring MC in agreement with data, after b-tagging.**

In addition to W/Z p_T reweighting

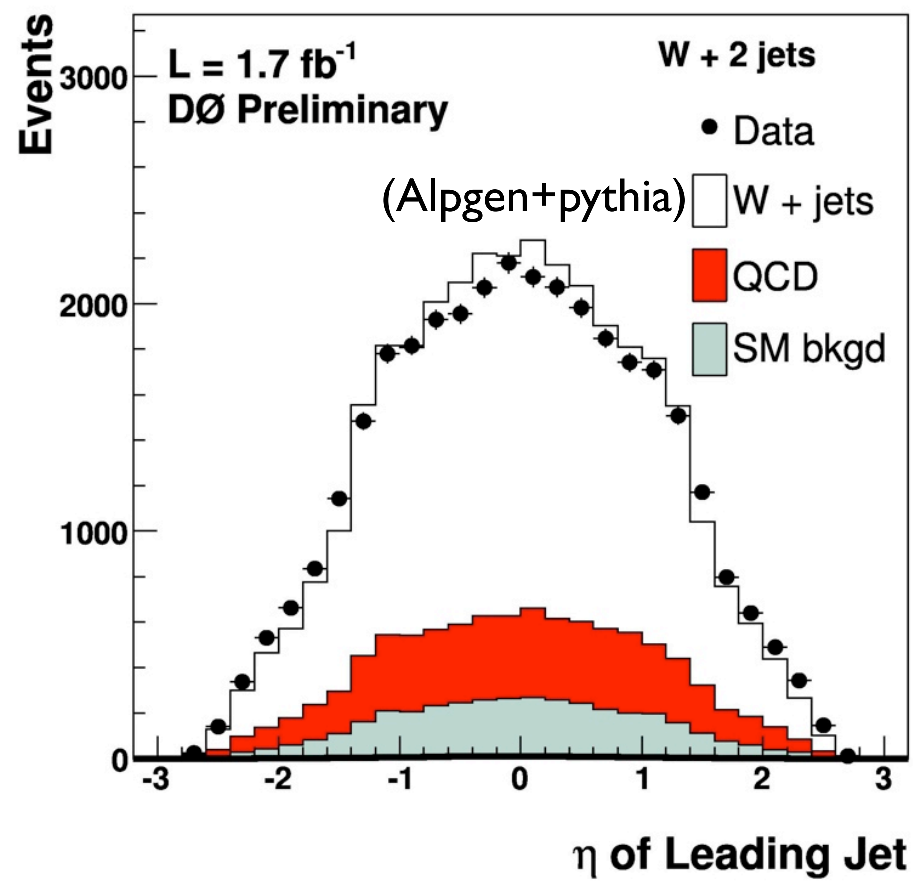
Data and ME

- Remember, alpgen currently the main generator used
  - Experiments have large “inertia” (rather have “known” problems…)

Hint of Trouble….

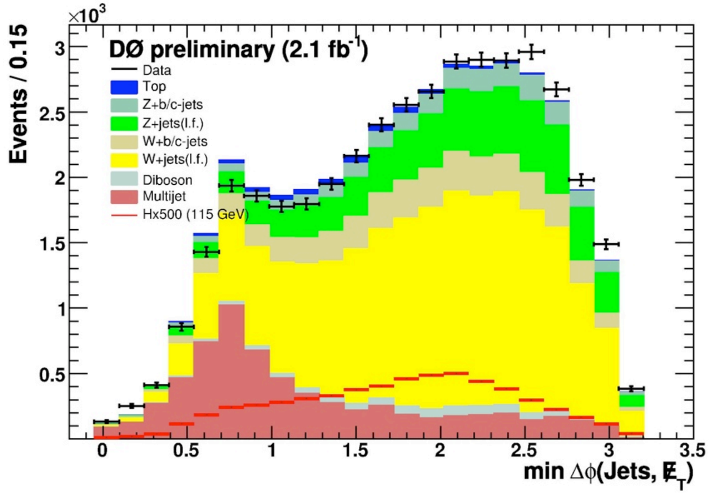


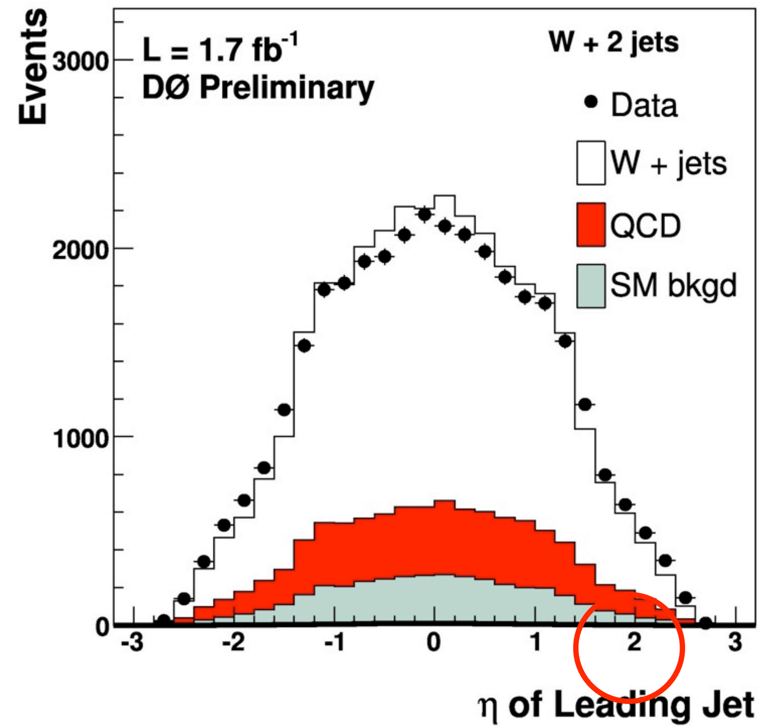
But Δφ sensitive to UE, MPI?



So...

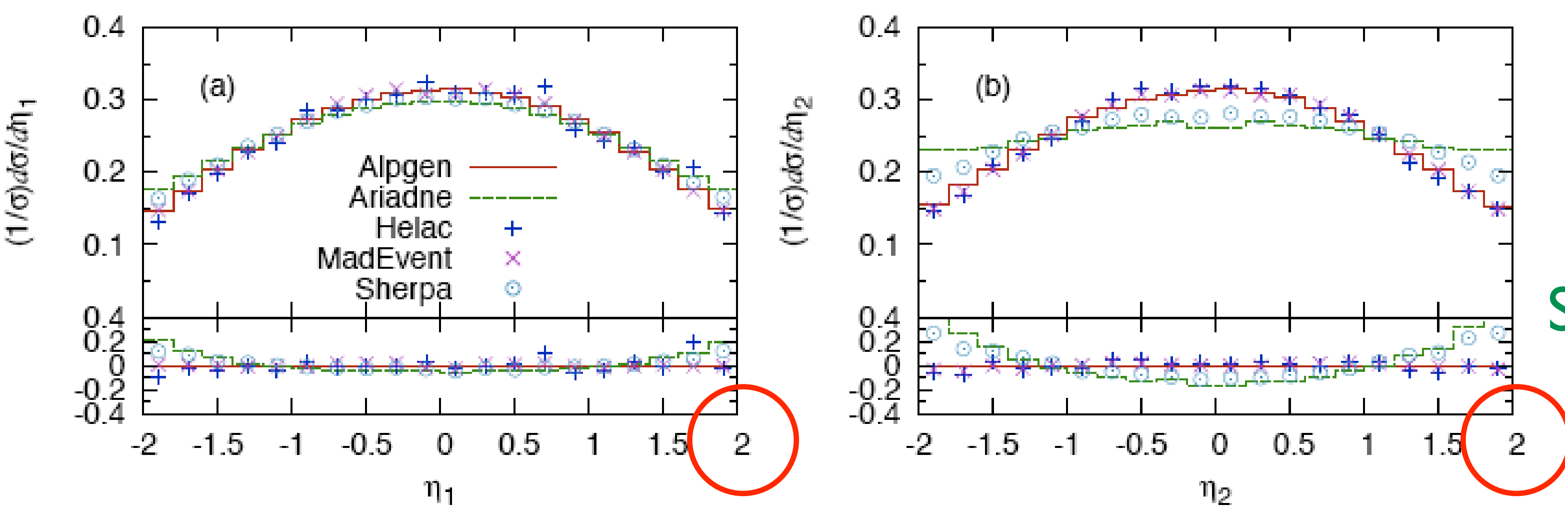
- After all these corrections....





- Maybe it's matching?

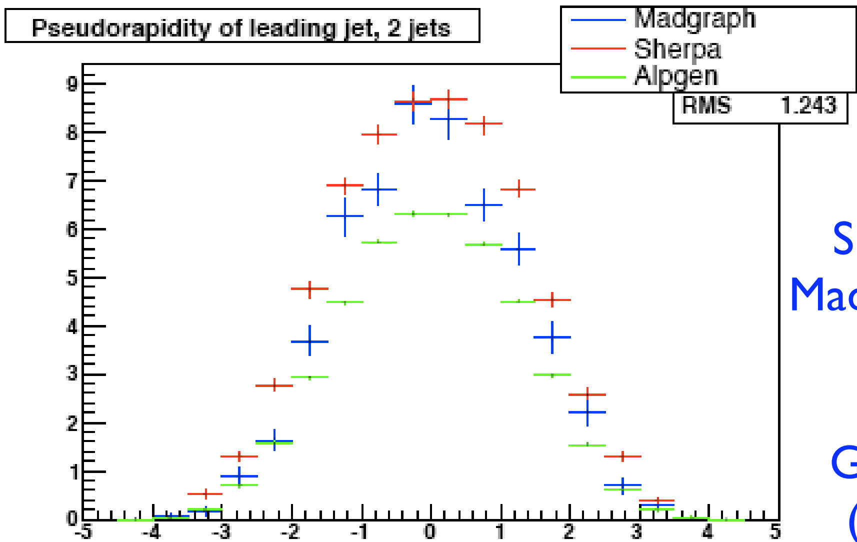
arXiv:0706.2569

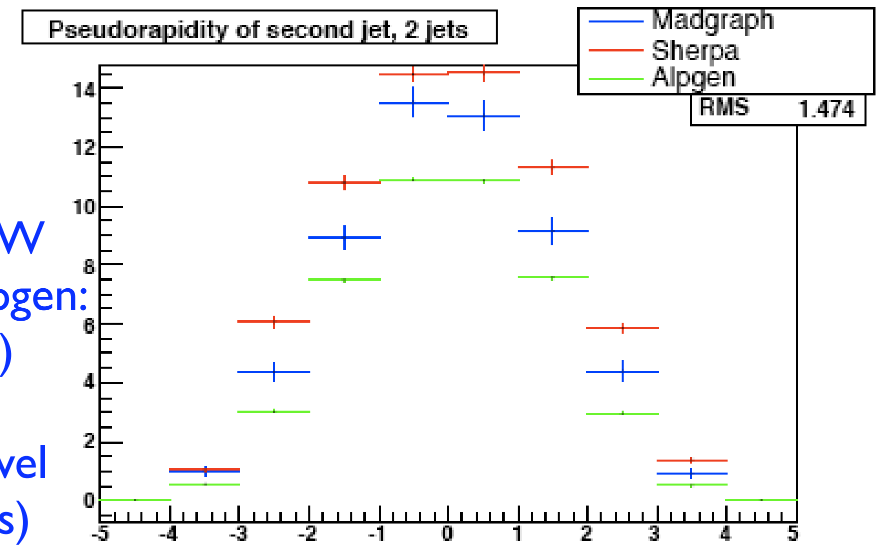


Alpgen, MadEvent, Helac with MLM, Sherpa and Ariadne with CKKW

You Can Do This at Home

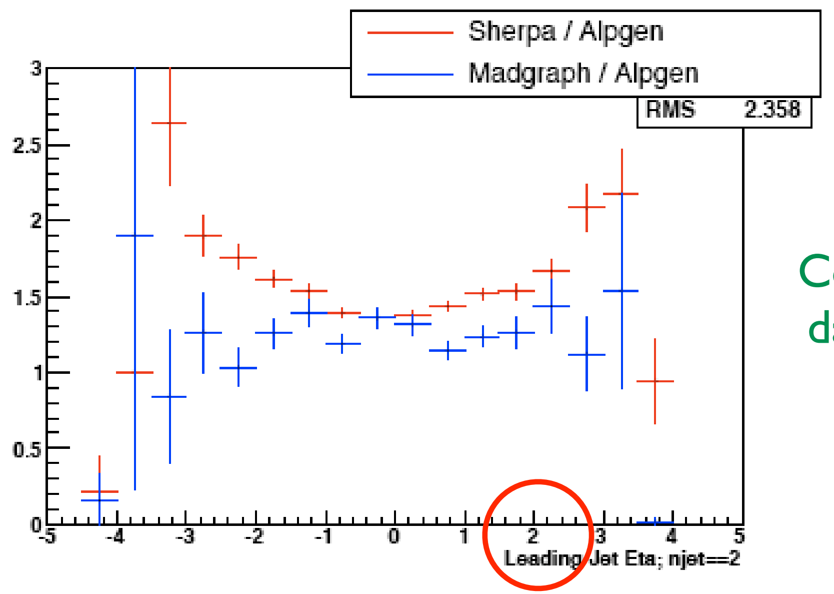
(In principle)



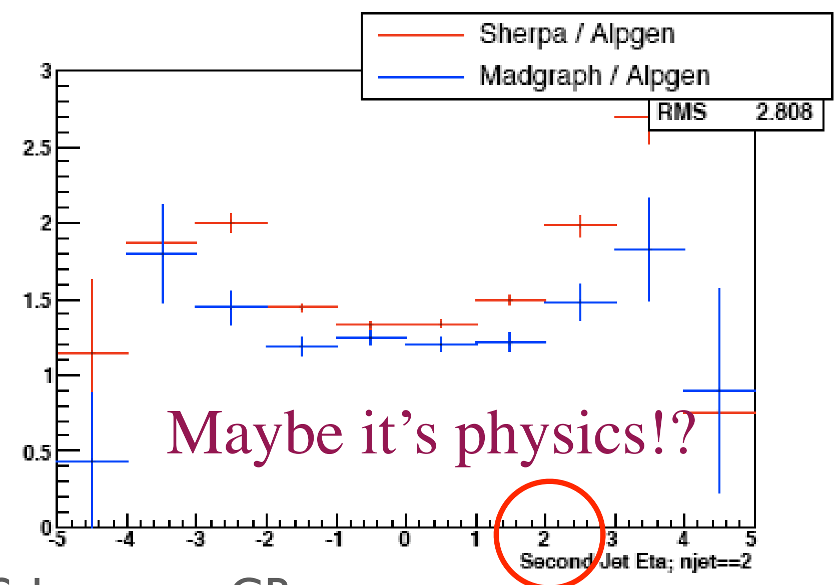


Sherpa: CKKW
MadEvent & Alpgen: MLM (cone)

Generator-level (SisCone jets)



Comparison to data by Fall (?)



W+jets, S. Lammers, GB

What Are We Learning?

- Tevatron samples large enough to do precision* V+jets physics

*Precision means: "can't hide in statistical uncertainty"

- We see differences between data & "ME MC"
  - After applying all "k"-factors we expect (+ p^T-dependent reweighting, heavy flavor)...
  - ... + some overall normalization factors we observe to be necessary
- (Eerily) similar differences can be observed between MC generators (at least in η distributions)
  - ➡ In principle it should be possible to understand their origin

Why Is This Bad?

- Experimentally, we determine contribution to "W+jets" from QCD multijet, Z+jets, top, ...
- But if we lack the necessary precision in understanding the shape of the actual W+jets contribution, we can't*
  - Measure WW → $\ell\nu jj$
  - Search for H → WW → $\ell\nu jj$
  - Search for qq → Wγqq → Wqq (the <u>only</u> VBF process accessible at the Tevatron...)
  - ...

Important!

*Can't is a strong word... we can reweigh & assign a systematic uncertainty of the same size as the effect

How Important Is This?

- The understanding of W+jets (i.e. the discrepancy between data and alpgen, and between various generators) is currently one of the major difficulties in many Tevatron analyses
  - Comparisons between the other generators and data will hopefully be available soon
- Based on the plots, I believe/hope the problem can be
  - Understood, and
  - Solved $\Rightarrow$ "Mega-W precision"
- IMHO it would be a mistake to postpone this to LHC
  - It will probably be harder, + no need to delay

Anyway...

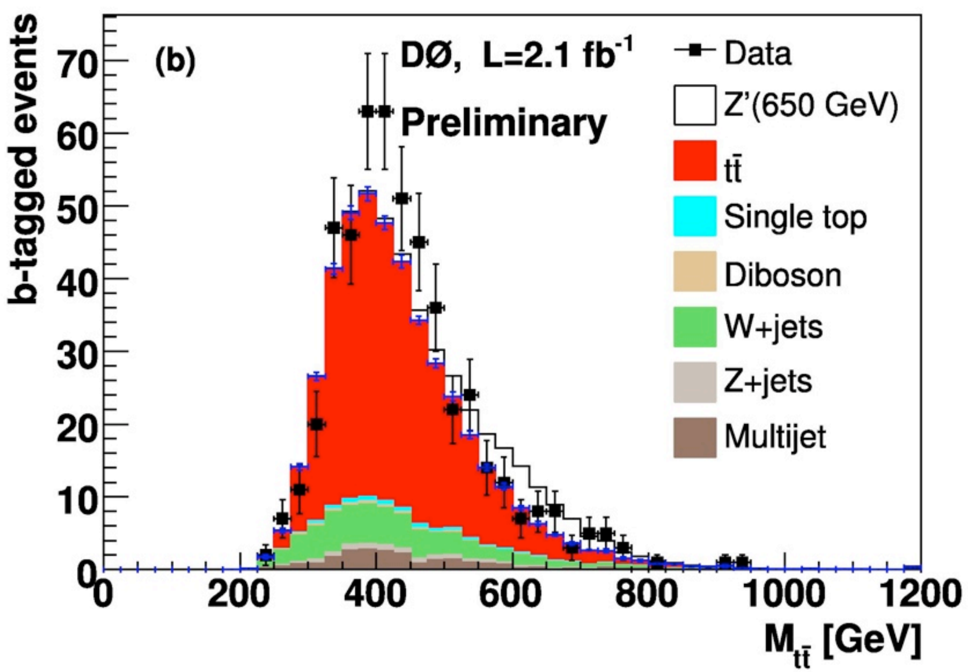
- Luckily, we can make signal-poor samples, and based on that adjust the MC to the data
  - Take the size of that adjustment as a systematic uncertainty
  - (This adjustment is not in places that are particularly sensitive to the signal BTW)
- Then proceed with the Higgs/single top/... search
  - Need to look at all channels (e.g. production, and decay of both H, W and Z)
  - Push sensitivity in each channel to the limit

Top

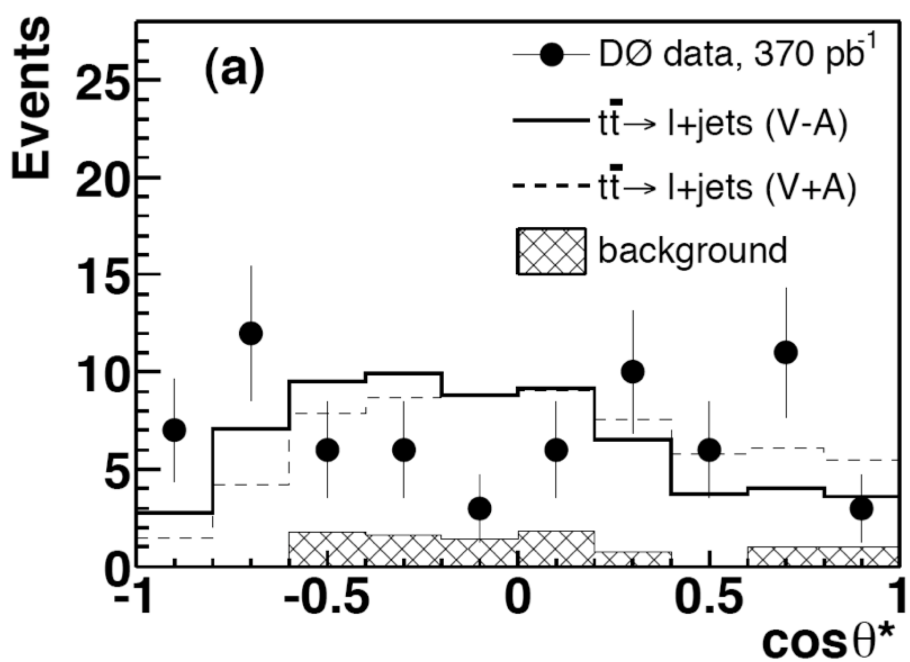
Top @ Tevatron: Production & Decay

- Top mass and cross-section measurements are very accurate
  - "Integral" measurements
- "Differential" measurements statistics limited

tt resonance search

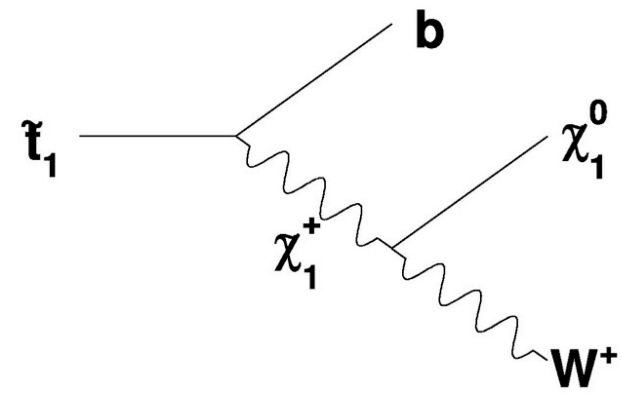


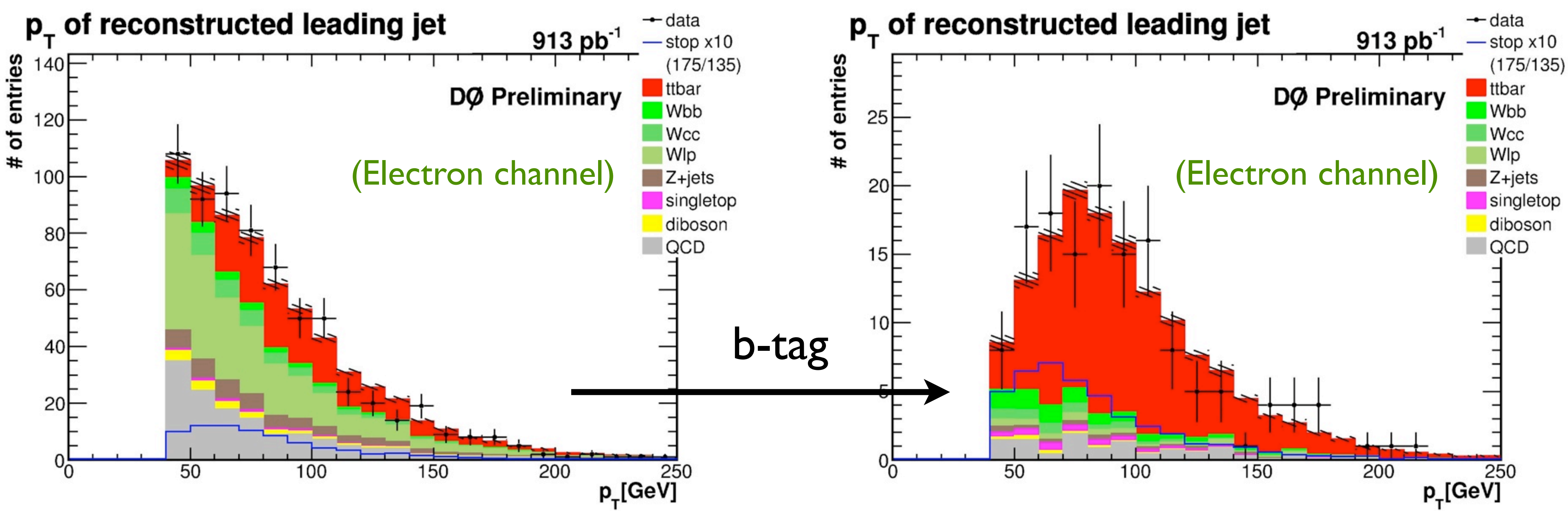
W helicity in top decay



Top @ Tevatron: Just top?

- Search for stop pair production
  - It looks like a top
  - Use multiple variables in *likelihood*





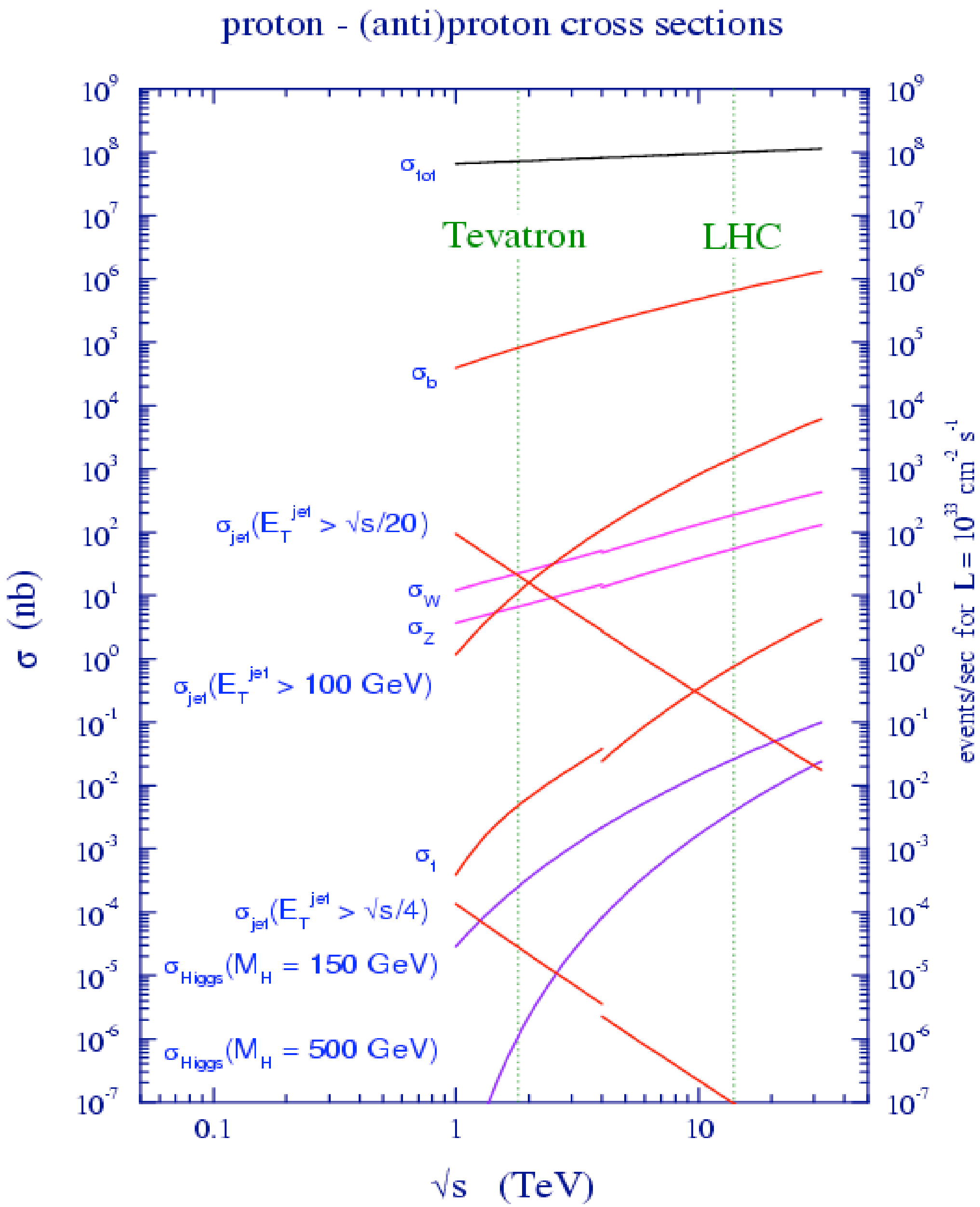
Top and Simulation

- At the Tevatron, small statistics → large statistical uncertainties
  - Accuracy of top simulation only needs to be that good
- But, *very* difficult to correct simulation based on data
  - For (non-top) W+jets some handle from Z+jets
    - Good for a counting experiment, i.e. how many lν + 4 jets from W +jets?
    - But reweighting in multiple variables tricky, and modern analyses all use some kind of multivariate technique
- What is the best way to validate top simulation?

LHC

At the LHC

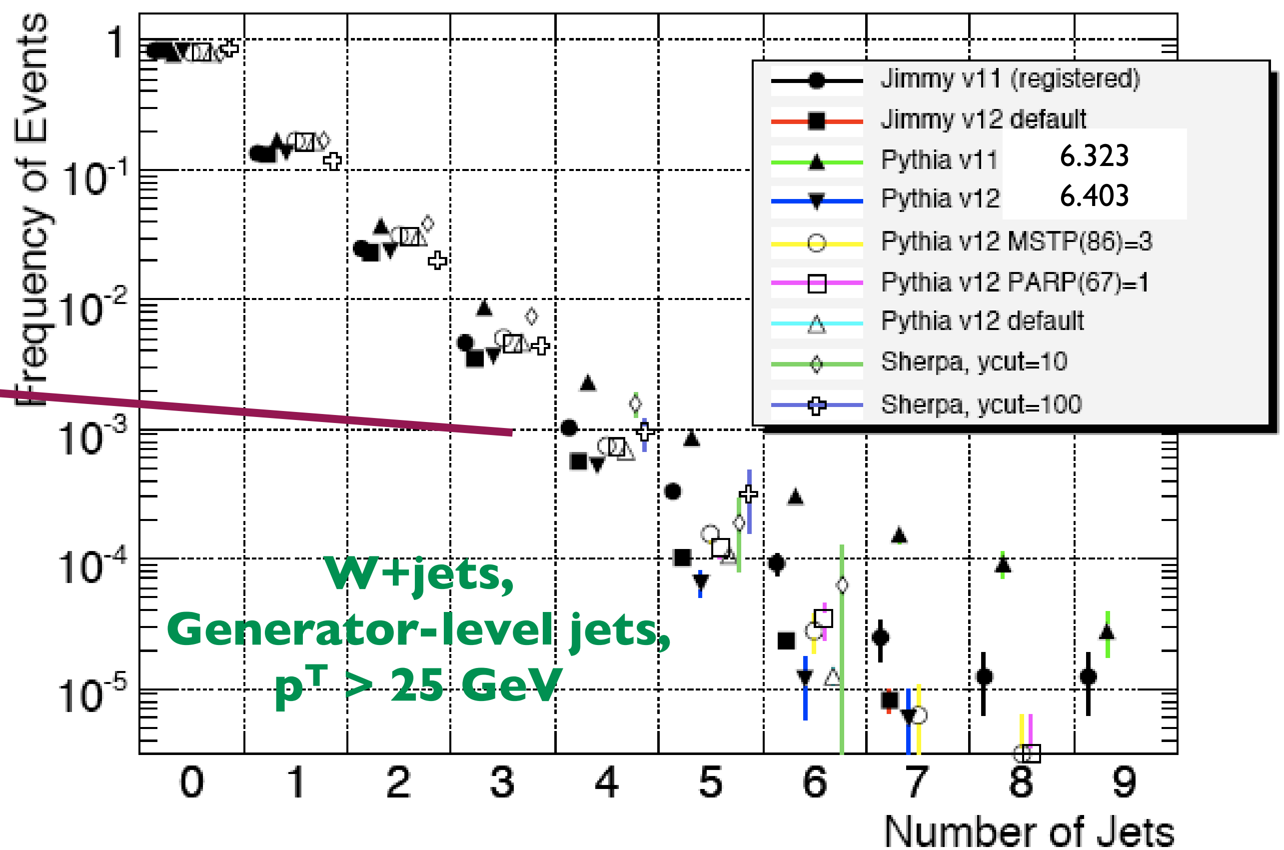
- Cross-sections:
  - W, Z x 10
  - top x 100+
  - ➡ 1 fb^{-1} yields
    - ~10^6 tt pairs (x 5-10% ε)
    - ~6 10^7 leptonic W's (x ε)
- Luminosity x 30
  - We expect 100's of fb^{-1}
  - ➡ "Giga-W, mega-top"



"First, the Standard Model"

- Common wisdom for LHC is to first re-establish the SM
  - Yes! But what is it?

ATLAS study, Berkeley group

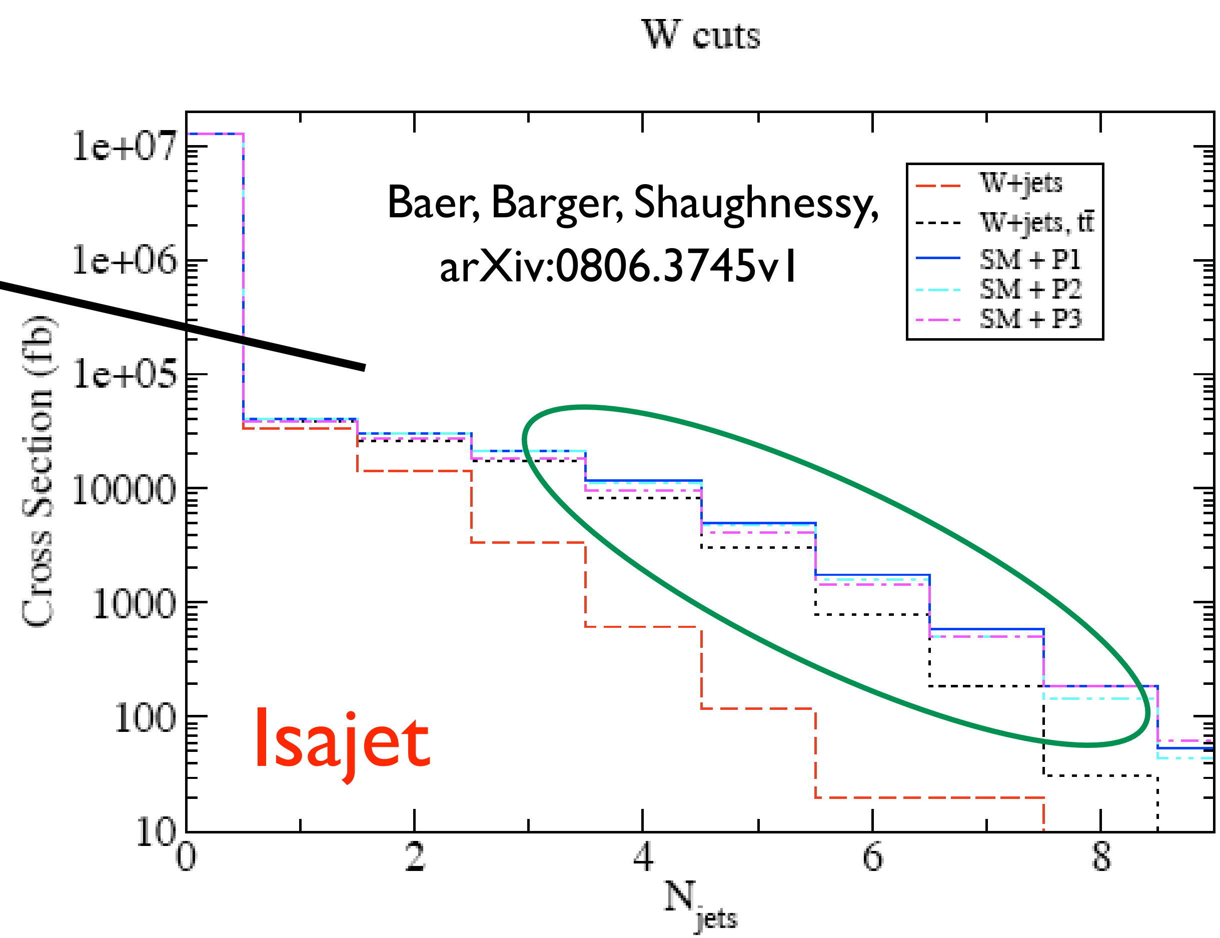


Large variation between generators, and within a generator significant sensitivity to parameters

- And there may be contamination!
  - Even for the relatively low mass SUSY points below, SUSY impact within generator differences

- $p_T(j_1) > 100$ GeV,
- $p_T(j_2, ..., j_n) > 50$ GeV.

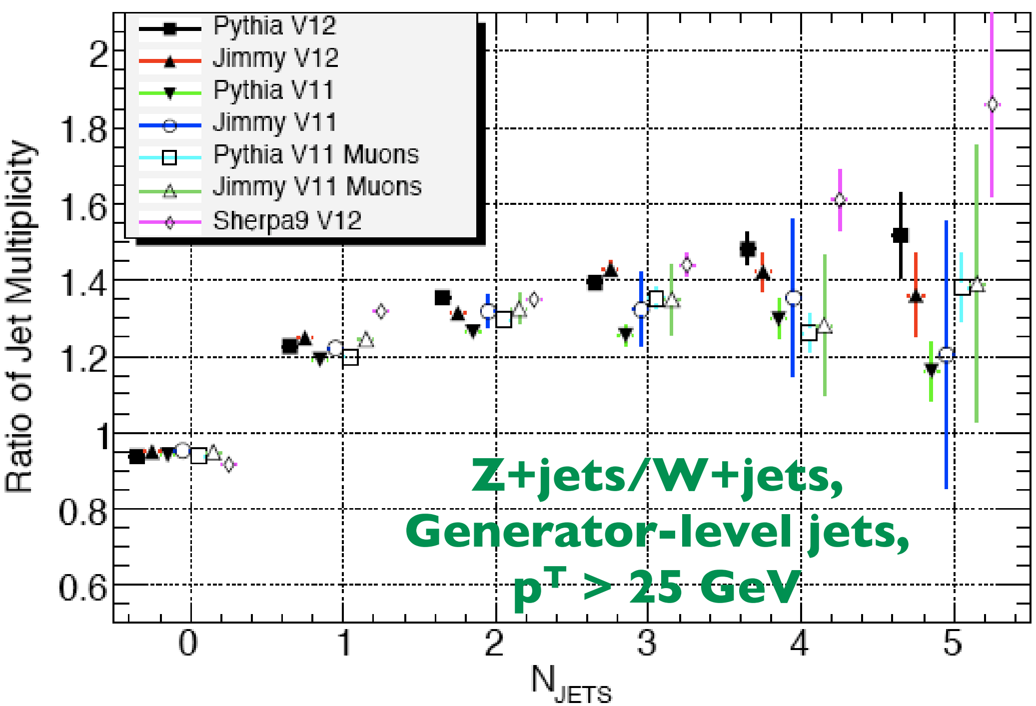
Not soft jets....



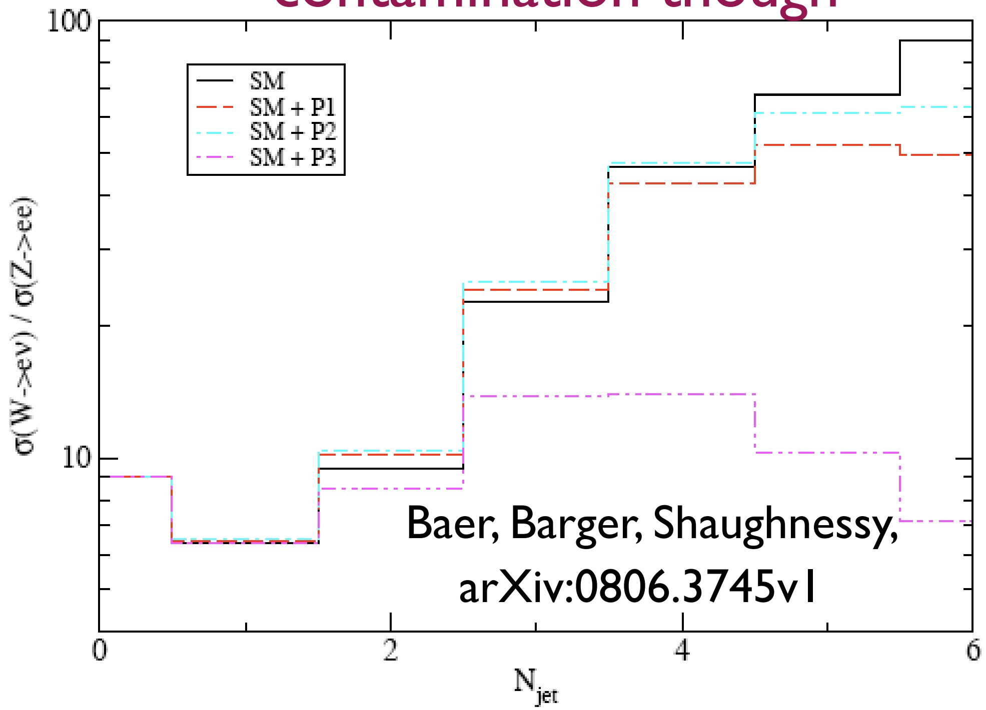
Ratios, Top

- Natural to try W/Z ratio in jet bins
  - Get much better agreement between generators
  - Driven by energy scale, usually set to boson mass

ATLAS study, Berkeley group



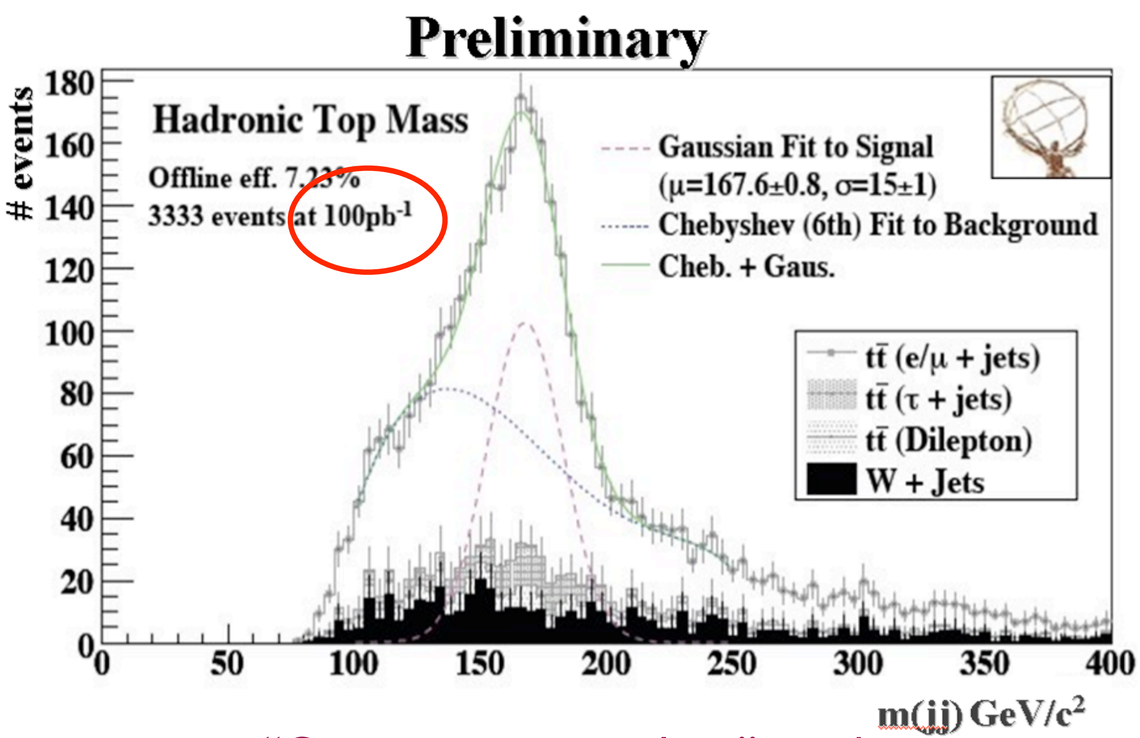
May not help isolate SUSY contamination though



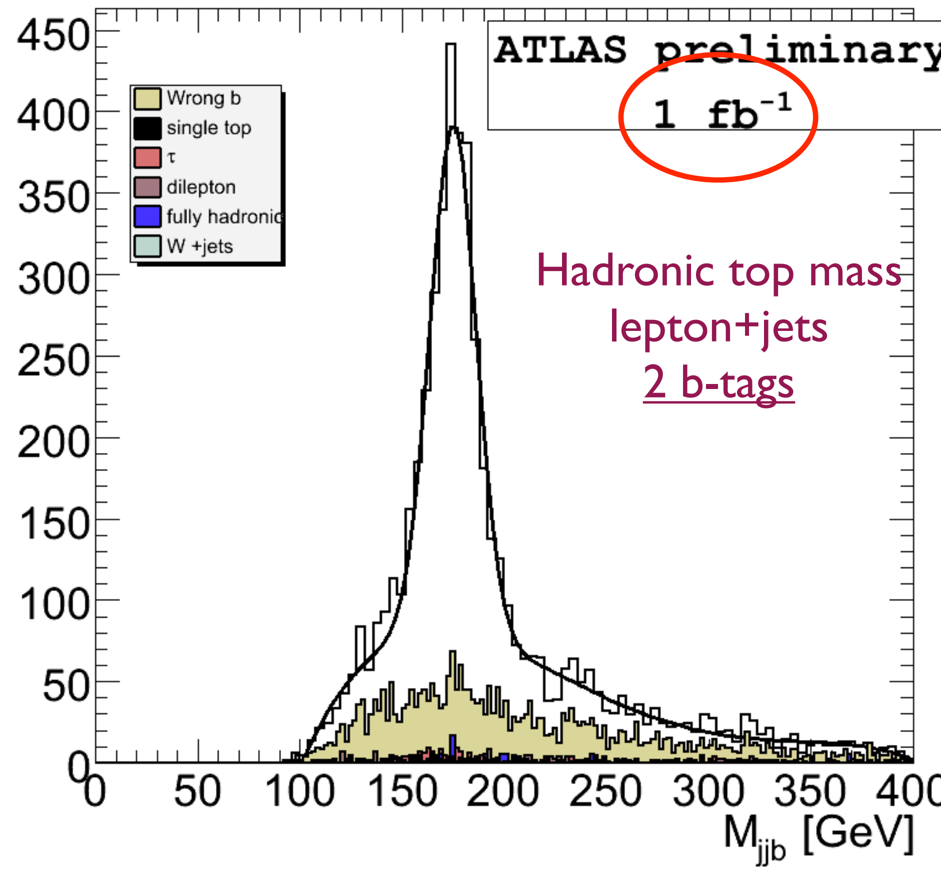
Top @ LHC

- Early data ⇒ divide Tevatron error bars by 10
  - Immediately get large samples

Early top x-section measurement

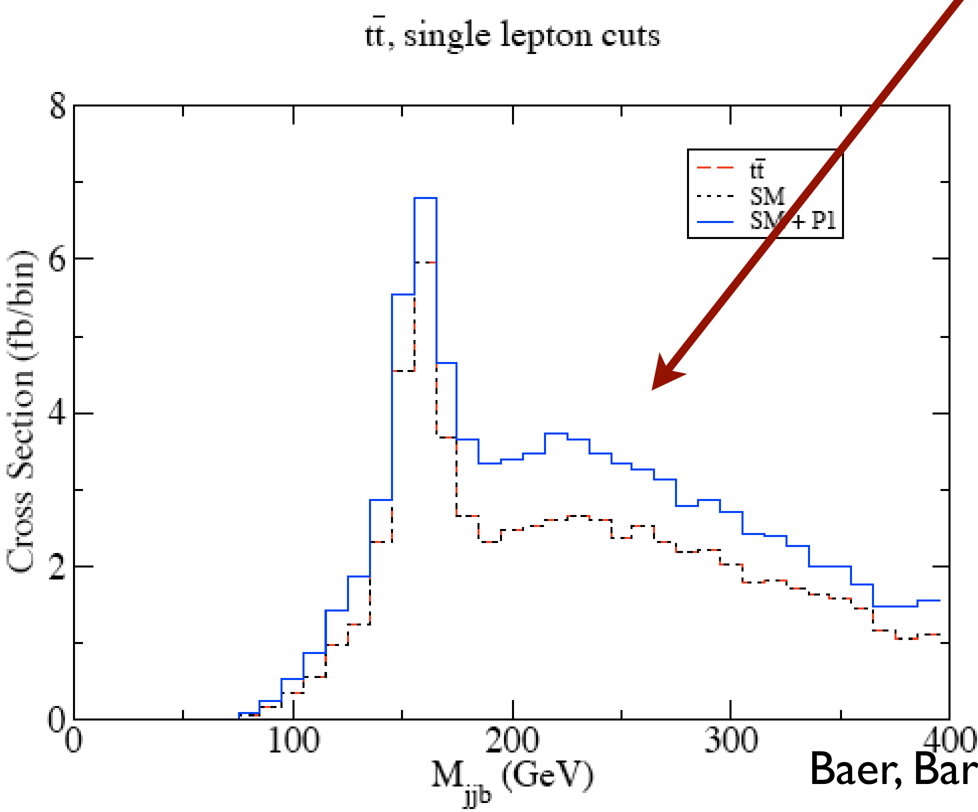


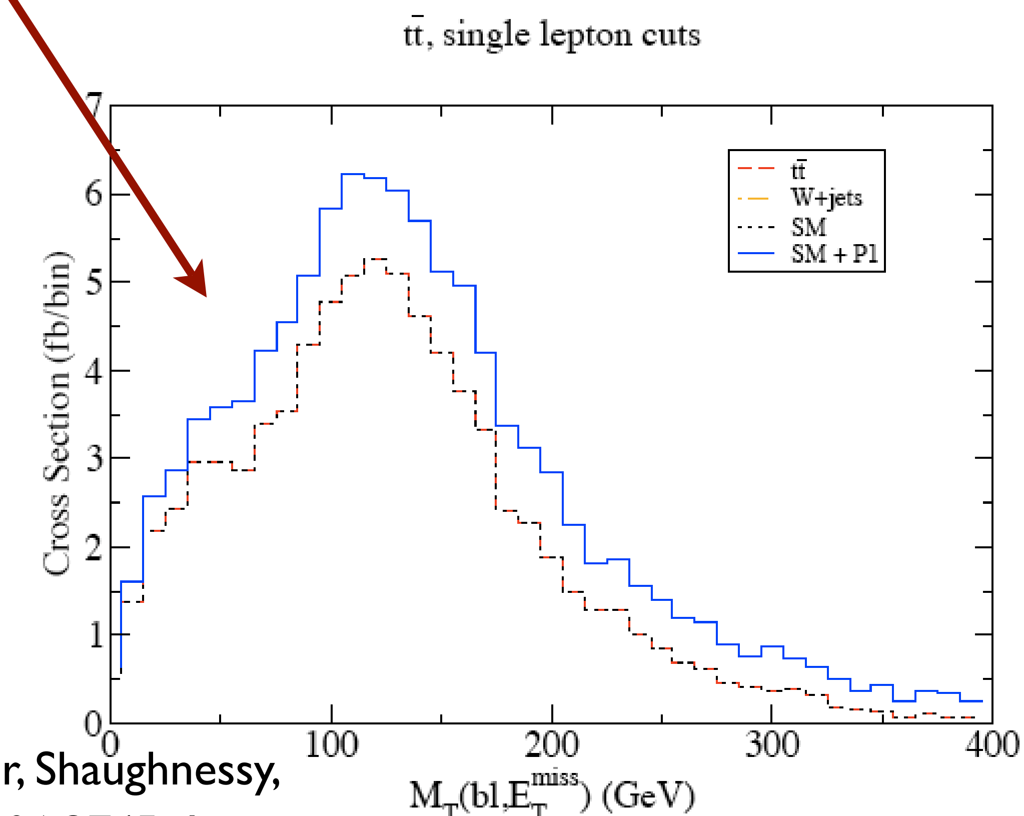
"Commissioning analysis", no b-tags



New Physics Pollution

- New Physics contribution may or may not be easy to isolate
  - How dependent are these on the MC generator?

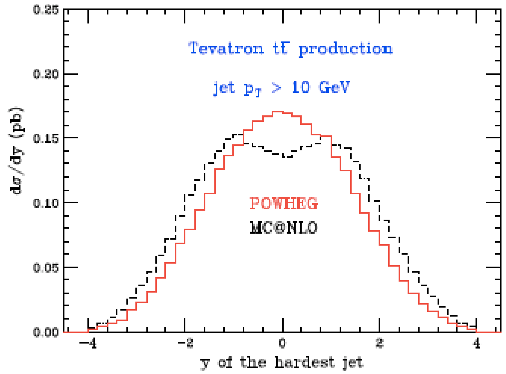




Baer, Barger, Shaughnessy, arXiv:0806.3745v1

Top Simulation

- Possible to get clean, large tt samples in data (if not polluted by new physics)
  - But unfolding is hard: Z+jets unfolding has been many years in the making at the Tevatron (see earlier)
  - Clean samples don't have statistics in the tails
  - Need to know which variables are particularly useful in identifying key uncertainties in modeling
    - Things we can measure well, like lepton p^T spectrum
    - 5th jet y is a scary variable
      - Scary mostly on "our" side
      - But what generates this? (Herwig...)



Event Generation @ LHC

- Rule of thumb: want 10x more MC events than real data
  - Not going to happen @ LHC (in first n years)!
  - $W \rightarrow l\nu$ exceeds rate-to-tape at design luminosity...
- Need to be very specific about samples that are most useful to "adjust" generators
  - Requires close interaction between experimenters & generator experts...
  - ... but of course we are limited in our ability to share "data"
    - We'll need to work our way through this

New Physics

- Generation of new physics in various models readily available
  - SUSY extensively covered
  - LRSM, some ED, ...
- Of course, exceptions
  - Is there publicly available code for $T_H T_H \rightarrow tt\ A_H A_H$?
- New models without generators (or not interfaced to PS) can't be tested by experimenters
  - LHEF are a good start, but ...
  - ... users should be able to change parameters

New Physics Precision

- "All current new physics models are wrong (at some level)", phenomenology is what's important
  - We are limited in the number of samples we can produce
- Many new search techniques use multivariate techniques, helicity variables
  - Need to get many distributions "right"
    - In signal & background: $g^1_{RS} \to t_R\, t_R \neq$ (wide) $Z' \to t\, t$
  - Requires e.g. decaying top in madgraph before feed to pythia $\Rightarrow$ reduces "slots" left for extra jets
    - Important to propagate spin information!

Summary

- Great datasets exist, fantastic ones will be collected soon
  - Mega-W, Kilo-top now, Giga-W & Mega-top soon
  - ➡ Precision physics in V+jets, top+jets
    - Critical to discovery and/or understanding of new physics
- Top quark is the next big challenge
  - Early LHC running will have lots of tt +1 jet, tt + 2 jets
  - How soon is tt + 3 jets important?

Summary (2)

- Multivariate techniques now very easy to use ("standard" software packages)
  - Requires fuller understanding of correlations between distributions
    - Maybe used a little too aggressively for the moment
    - But critical to improving sensitivity!
- Tremendous progress in MC description of data in past ~8 years
  - But need to keep going!
  - Dialog between experimenters and developers to identify variables most sensitive to modeling uncertainties

Generators @ LHC

- By now, many (~20) generators incorporated in ATLAS/CMS software
  - From pythia 6 & 8, herwig(++) to alpgen, madgraph, sherpa, whizard,....
- LHE files in principle allow reduced turn-around time
  - Changing pythia/herwig versions is hard!
- Thanks to Tevatron experience, LHC communities are getting used to ME generators & matching
  - Comparisons with new pythia shower also starting