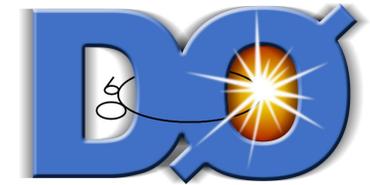


# Measurement and Discovery Techniques in Experimental Particle Physics



Thomas R. Junk  
*Fermilab*



Conference on Data Analysis 2014  
Santa Fe, New Mexico  
March 5-7, 2014



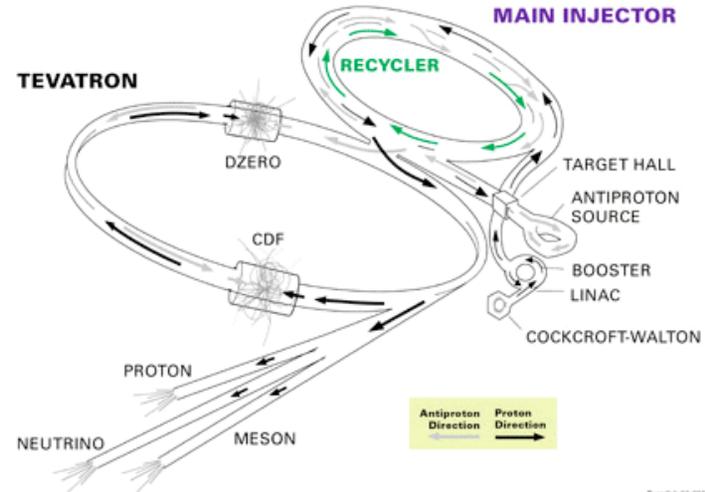
With some public material from the ATLAS and CMS experiments at CERN

mass →	$\approx 2.3 \text{ MeV}/c^2$	$\approx 1.275 \text{ GeV}/c^2$	$\approx 173.07 \text{ GeV}/c^2$	0	$\approx 126 \text{ GeV}/c^2$
charge →	$2/3$	$2/3$	$2/3$	0	0
spin →	$1/2$	$1/2$	$1/2$	1	0
	<b>u</b> up	<b>c</b> charm	<b>t</b> top	<b>g</b> gluon	<b>H</b> Higgs boson
<b>QUARKS</b>	$\approx 4.8 \text{ MeV}/c^2$	$\approx 95 \text{ MeV}/c^2$	$\approx 4.18 \text{ GeV}/c^2$	0	
	$-1/3$	$-1/3$	$-1/3$	0	
	$1/2$	$1/2$	$1/2$	1	
	<b>d</b> down	<b>s</b> strange	<b>b</b> bottom	<b><math>\gamma</math></b> photon	
	$0.511 \text{ MeV}/c^2$	$105.7 \text{ MeV}/c^2$	$1.777 \text{ GeV}/c^2$	$91.2 \text{ GeV}/c^2$	
	-1	-1	-1	0	
	$1/2$	$1/2$	$1/2$	1	
	<b>e</b> electron	<b><math>\mu</math></b> muon	<b><math>\tau</math></b> tau	<b>Z</b> Z boson	
<b>LEPTONS</b>	$< 2.2 \text{ eV}/c^2$	$< 0.17 \text{ MeV}/c^2$	$< 15.5 \text{ MeV}/c^2$	$80.4 \text{ GeV}/c^2$	
	0	0	0	$\pm 1$	
	$1/2$	$1/2$	$1/2$	1	
	<b><math>\nu_e</math></b> electron neutrino	<b><math>\nu_\mu</math></b> muon neutrino	<b><math>\nu_\tau</math></b> tau neutrino	<b>W</b> W boson	
					<b>GAUGE BOSONS</b>

# The Tevatron and Associated Accelerators



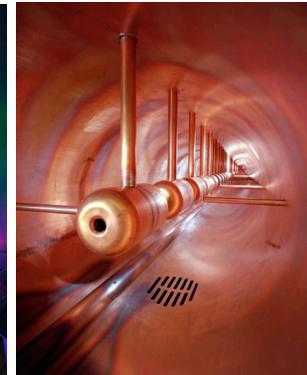
FERMILAB'S ACCELERATOR CHAIN



Fermilab 00-025



Cockcroft-Walton



Linac



Booster

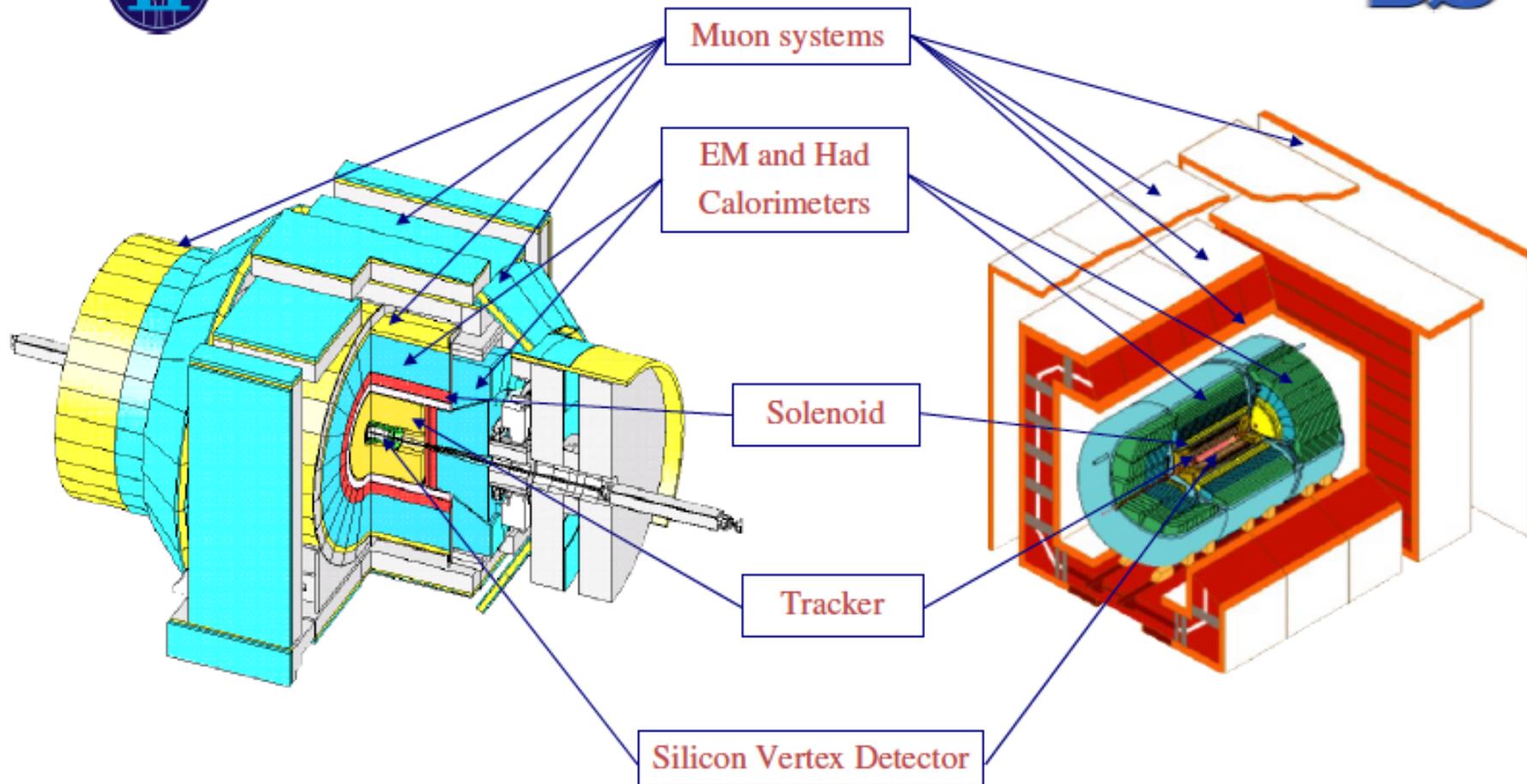


pbar debuncher and accumulator

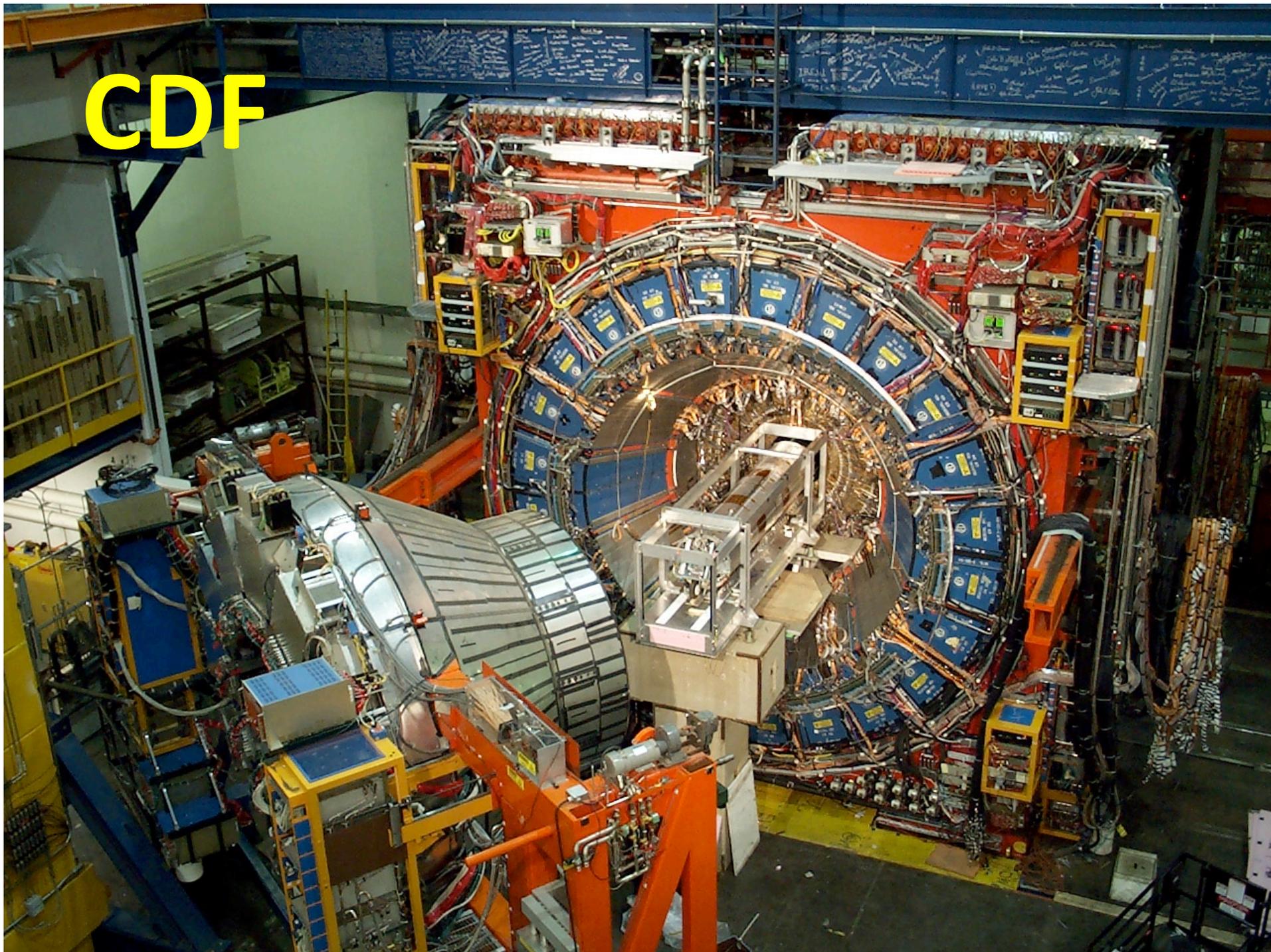
ppbar collisions at 1.96 TeV  
 Luminosity up to  $400 \times 10^{30} \text{ cm}^{-2} \text{ s}^{-1}$   
 A very good week:  $\sim 80 \text{ pb}^{-1}$

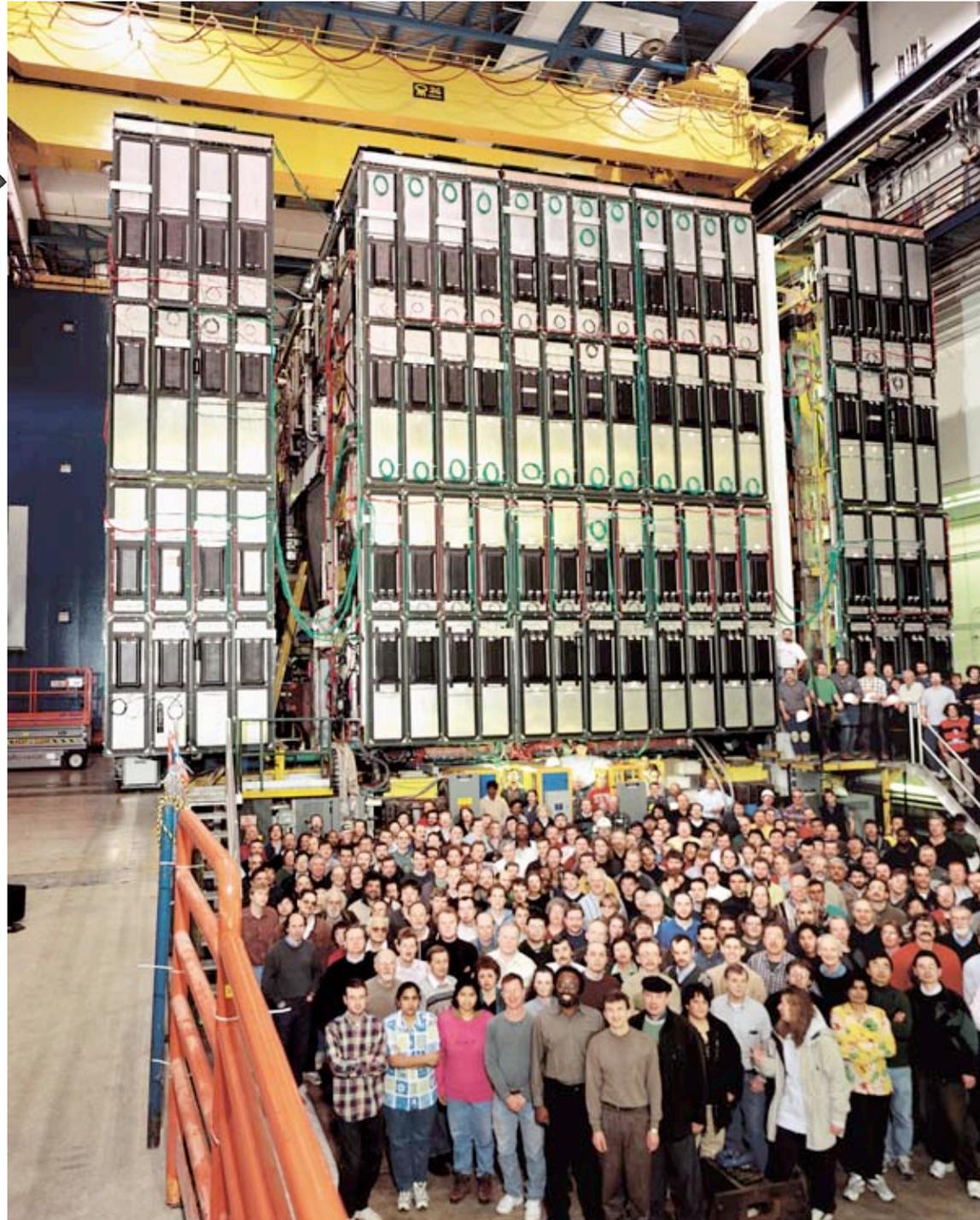
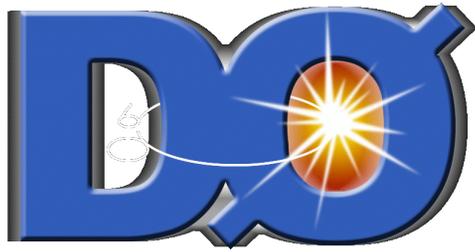
T. Junk CoDA HEP

# The CDF and D0 Detectors



CDF





3/5/14

# CDF Run II Trigger System

**Bunch Crossing Rate: ~1.7 MHz**

**Level 1 trigger ~15 KHz**

tracking

calorimeter: jets & electrons

muons

**Level 2 trigger ~800 Hz**

L1 information (tracks, e,  $\mu$ )

calorimeter shower max

silicon information

algorithms run in L2 processor

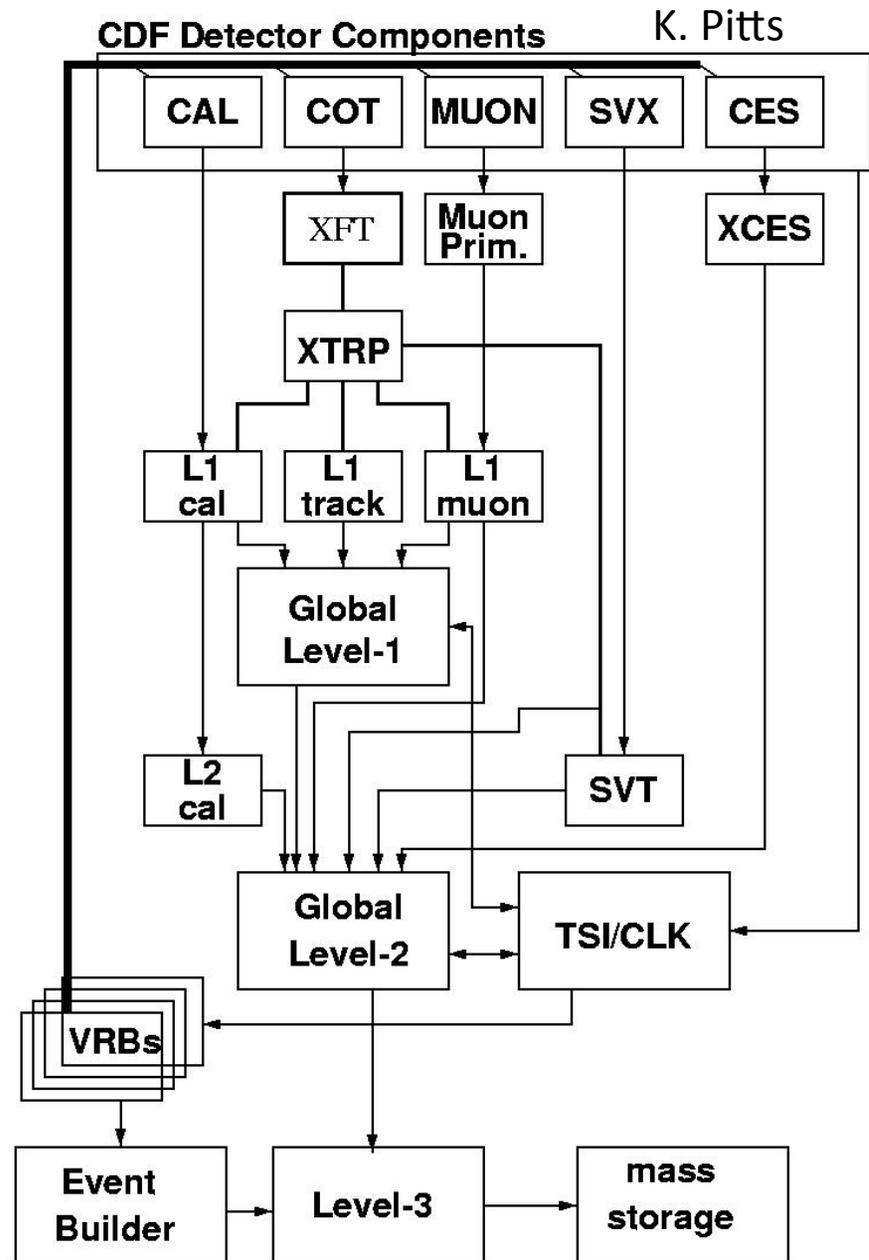
**Level 3 trigger ~200 Hz to tape**

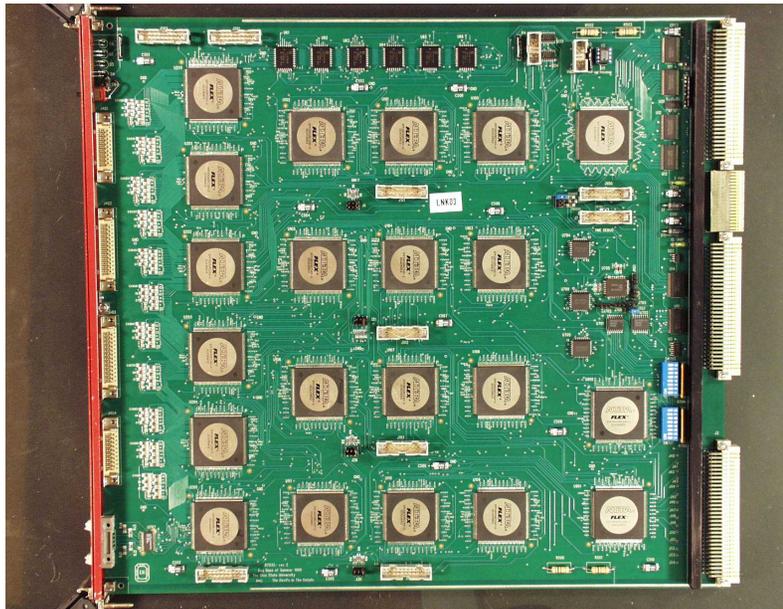
full detector readout

event building

“offline” processing

3/5/14

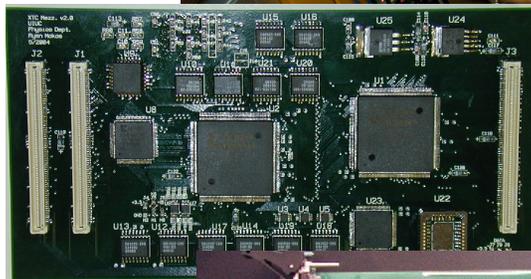




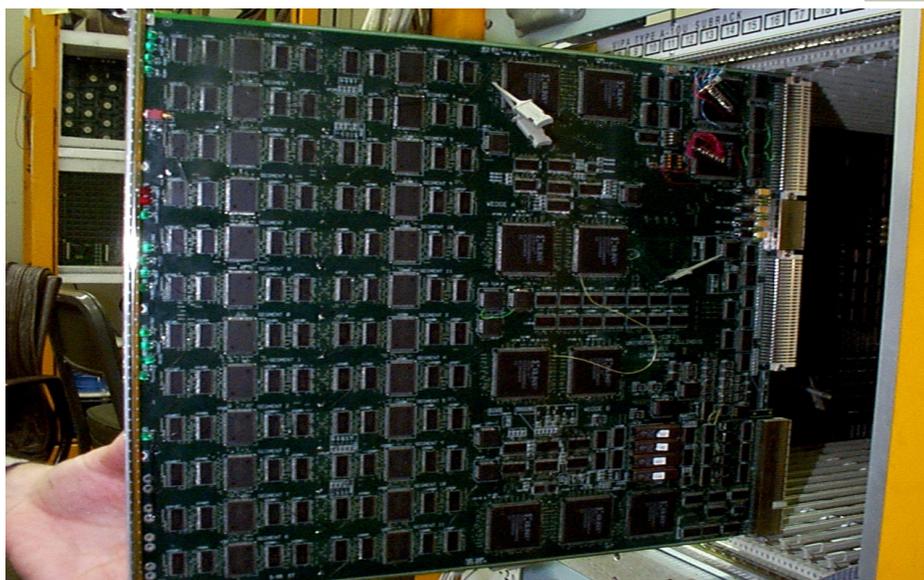
CDF  
Trigger  
Electronics



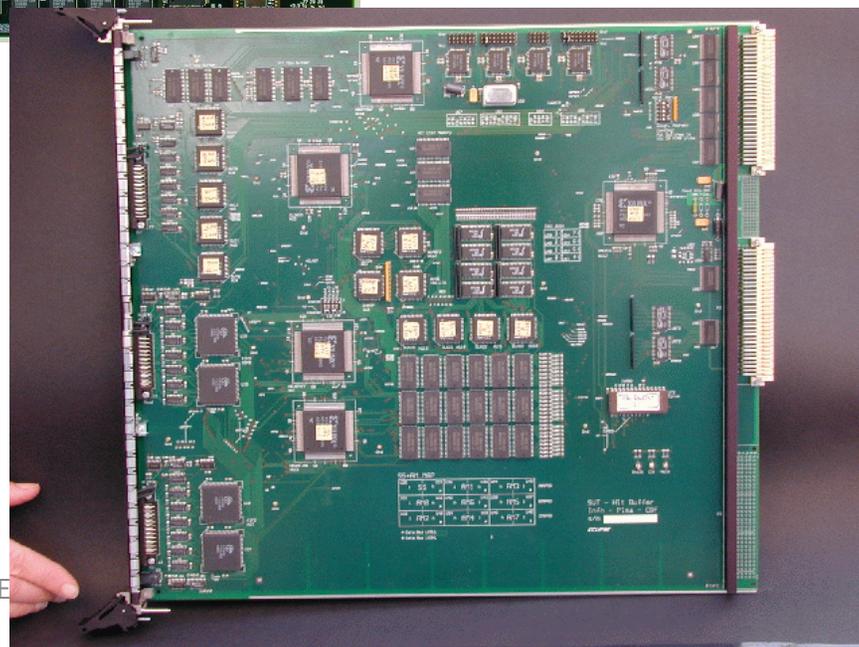
Hundreds of application- specific boards  
working together, passing and processing  
data on the nanosecond timescale...



System continually  
upgraded to deal with  
increasing luminosity/  
pileup



CoDA HE

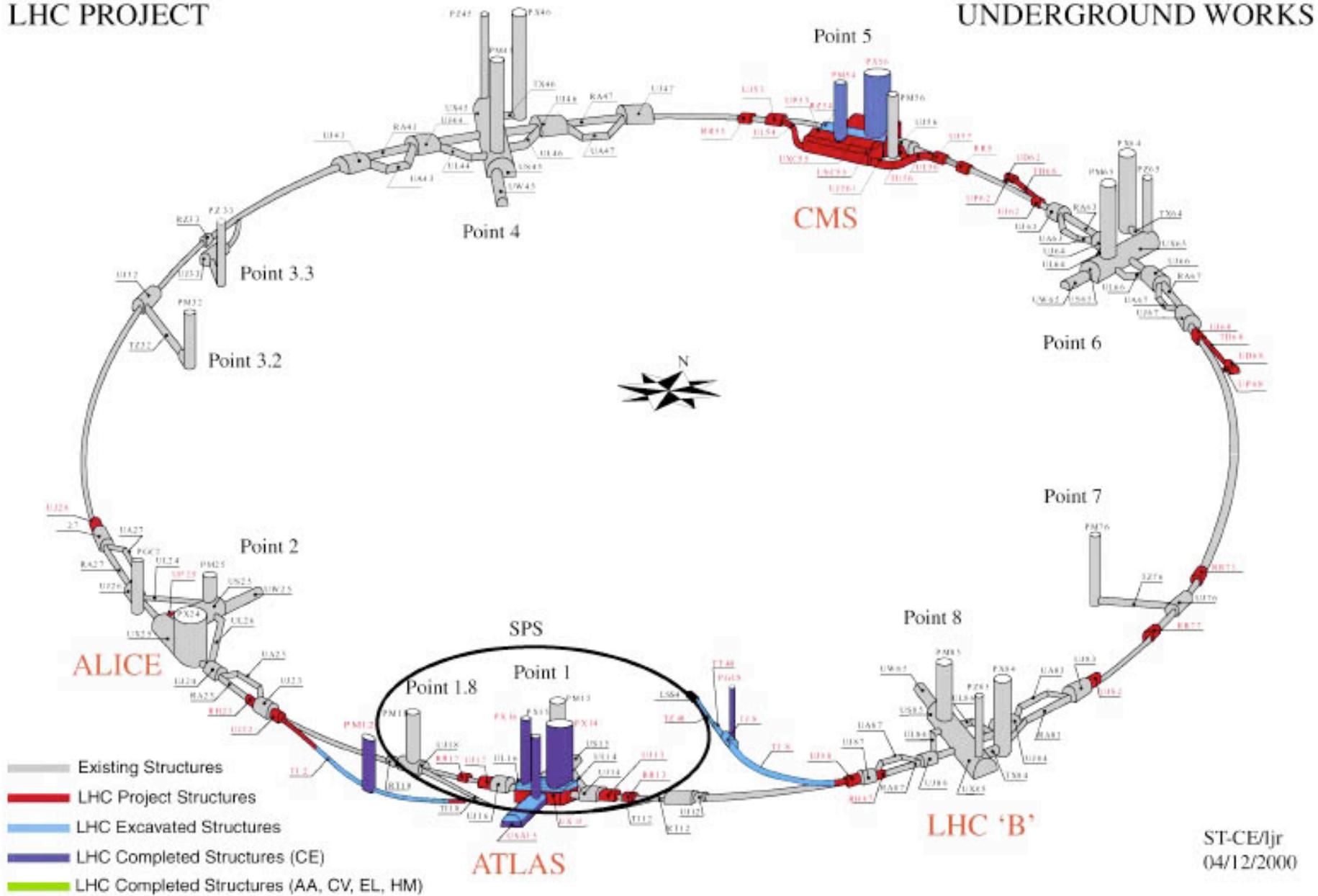


# The Large Hadron Collider at CERN from the air



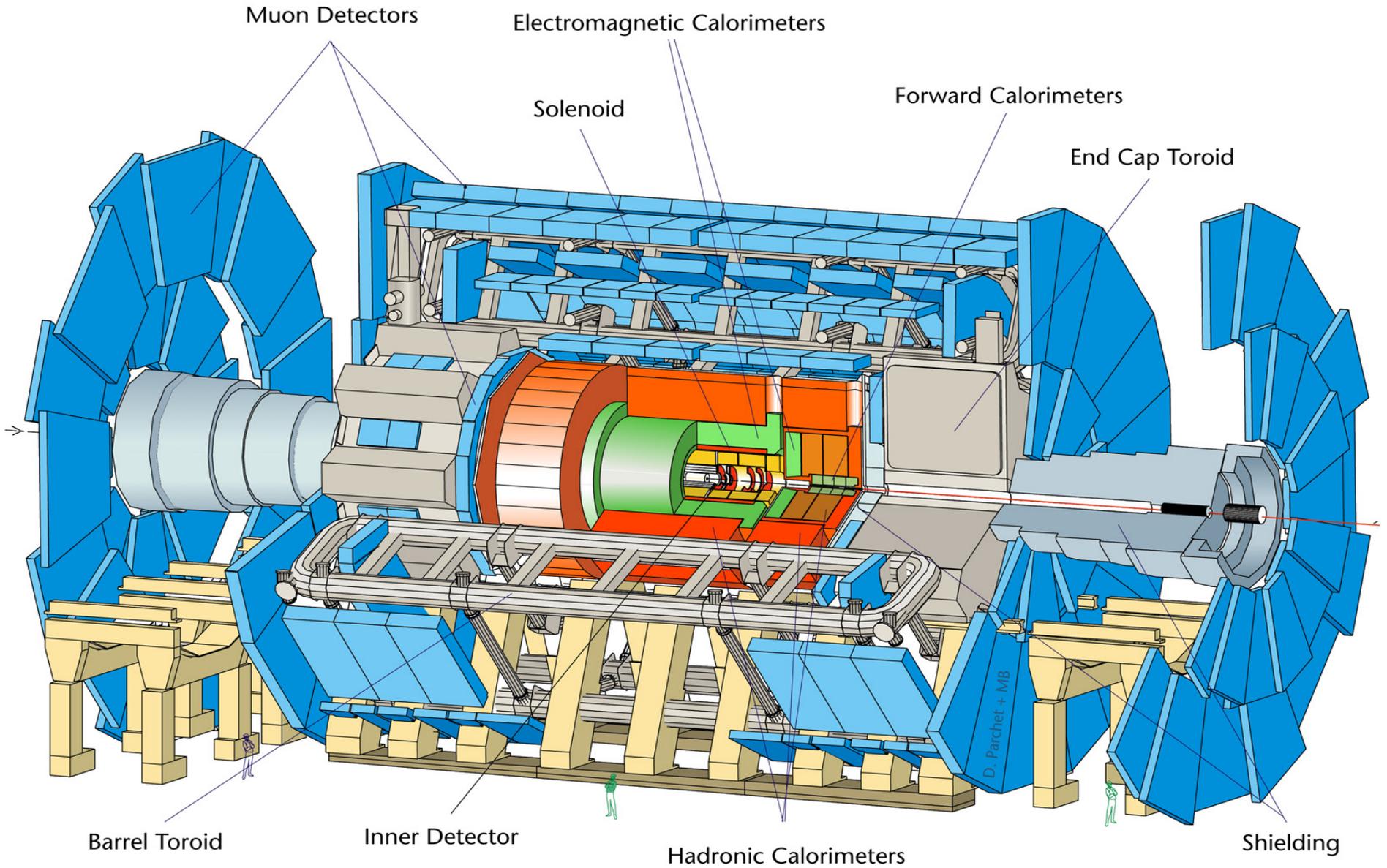
Re-uses the 27 km LEP tunnel

Strong superconducting magnets for  
pp collisions at 7, 8, and 13-14 TeV



ST-CE/ljr  
04/12/2000

# The ATLAS Detector at the LHC



# CMS DETECTOR

Total weight : 14,000 tonnes  
Overall diameter : 15.0 m  
Overall length : 28.7 m  
Magnetic field : 3.8 T

STEEL RETURN YOKE  
12,500 tonnes

SILICON TRACKERS  
Pixel (100x150  $\mu\text{m}$ )  $\sim 16\text{m}^2$   $\sim 66\text{M}$  channels  
Microstrips (80x180  $\mu\text{m}$ )  $\sim 200\text{m}^2$   $\sim 9.6\text{M}$  channels

SUPERCONDUCTING SOLENOID  
Niobium titanium coil carrying  $\sim 18,000\text{A}$

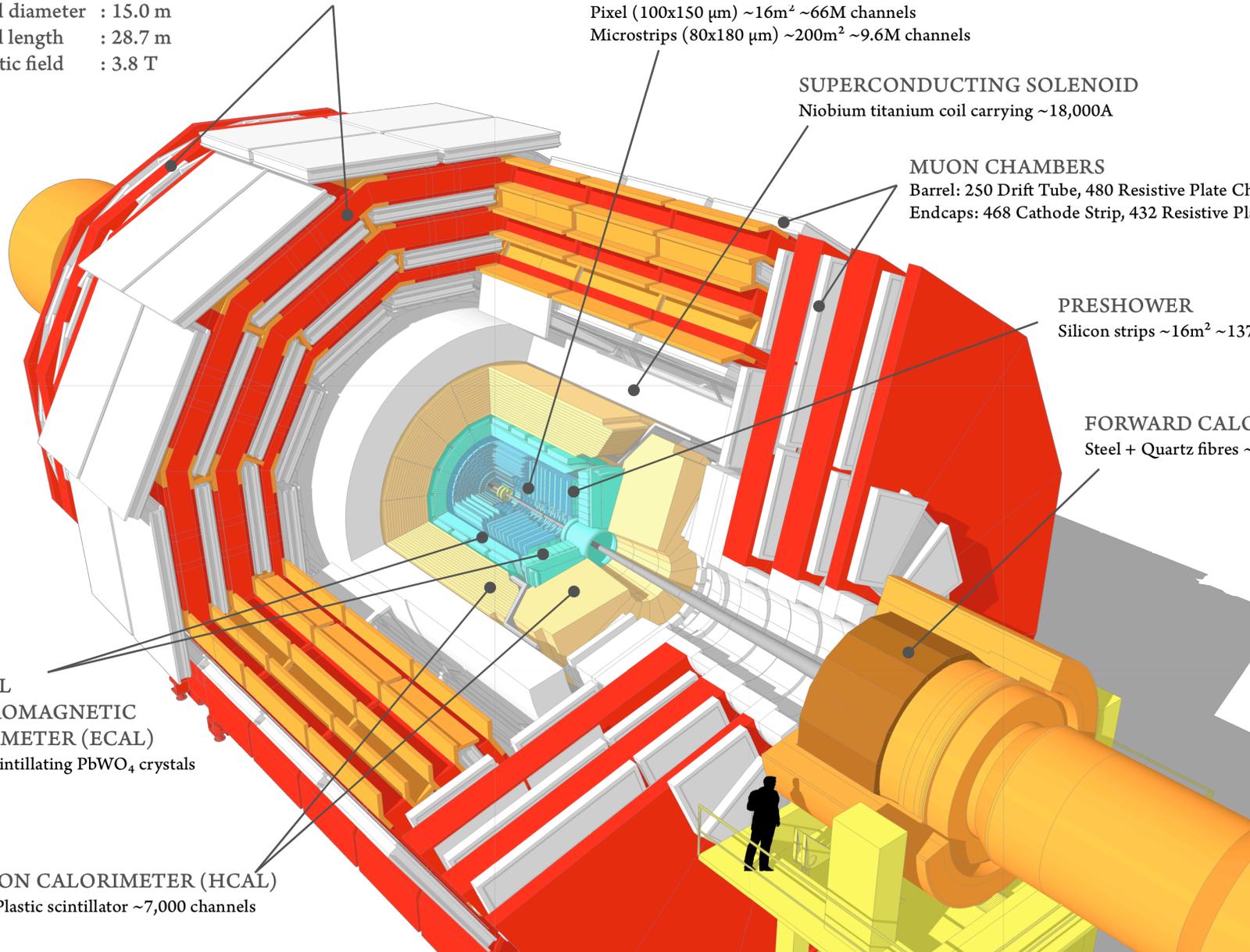
MUON CHAMBERS  
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers  
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers

PRESHOWER  
Silicon strips  $\sim 16\text{m}^2$   $\sim 137,000$  channels

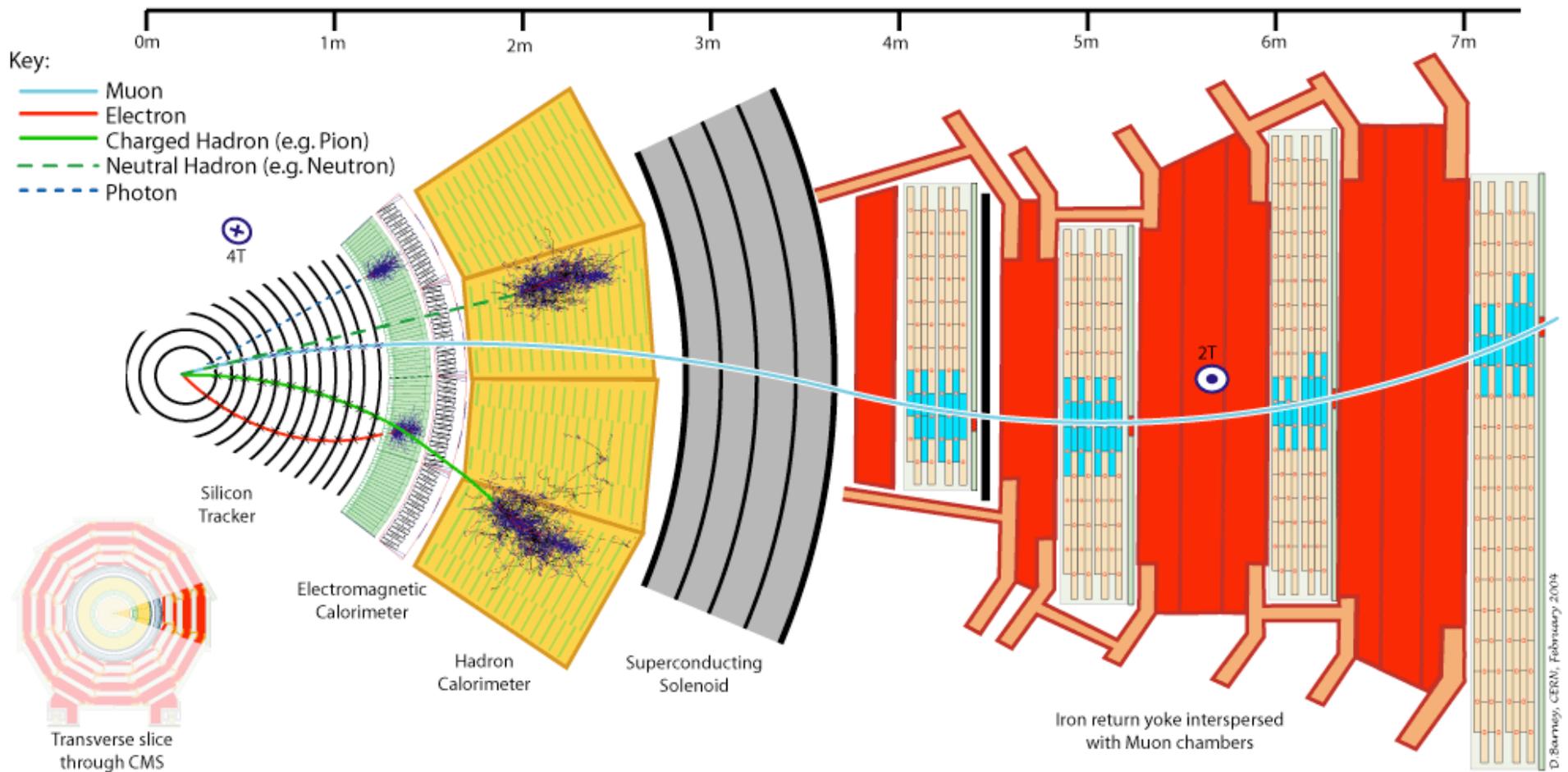
FORWARD CALORIMETER  
Steel + Quartz fibres  $\sim 2,000$  Channels

CRYSTAL  
ELECTROMAGNETIC  
CALORIMETER (ECAL)  
 $\sim 76,000$  scintillating  $\text{PbWO}_4$  crystals

HADRON CALORIMETER (HCAL)  
Brass + Plastic scintillator  $\sim 7,000$  channels

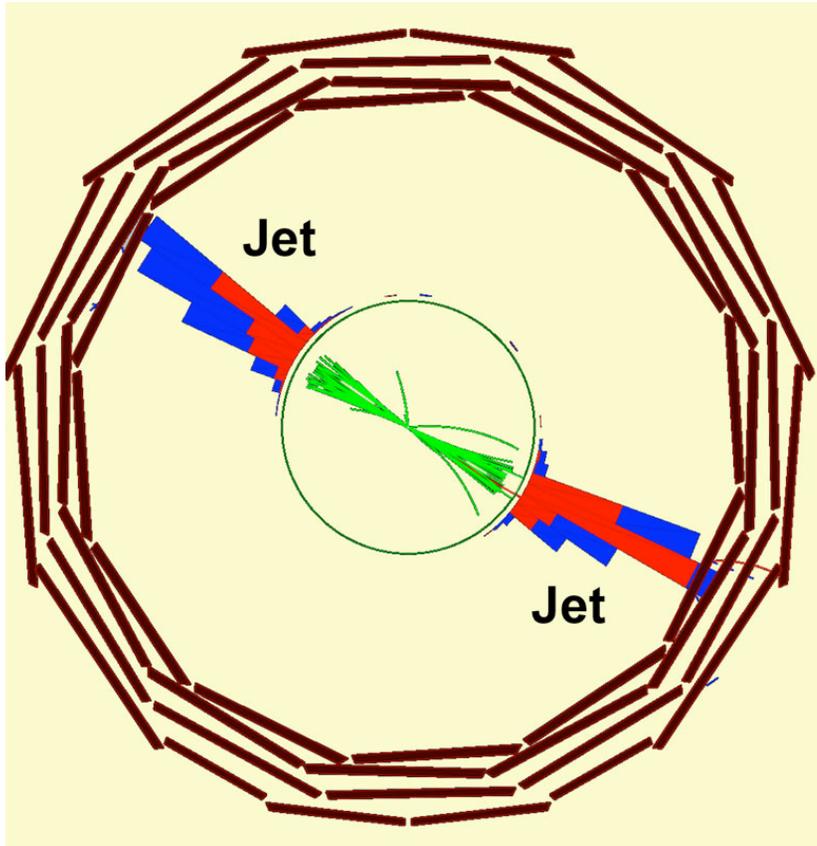


# A Slice Through CMS

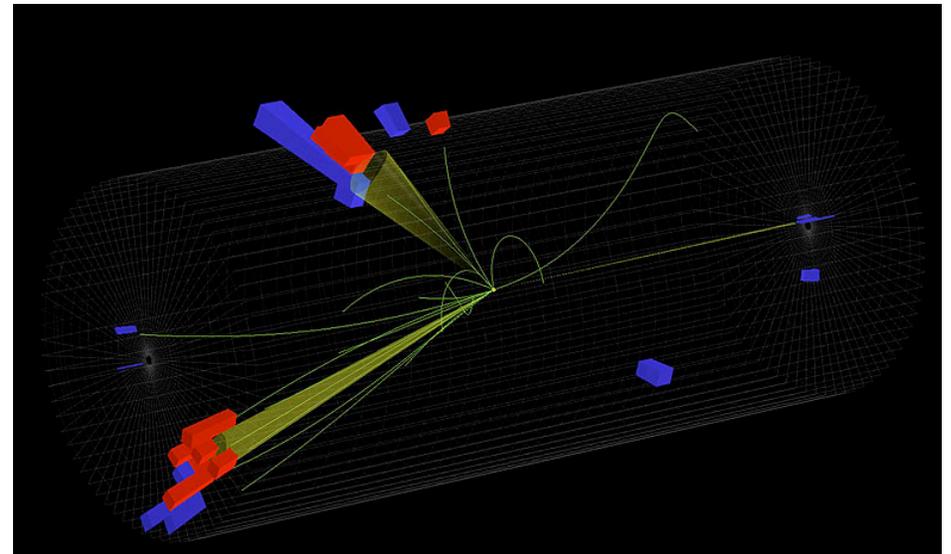
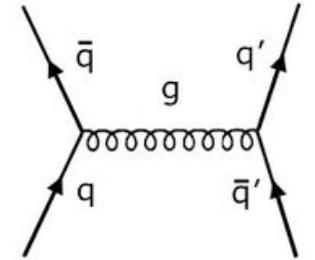
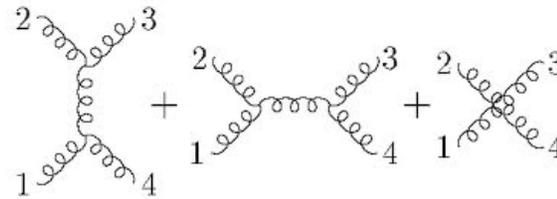


# A Couple of Di-Jet Events at a Hadron Collider

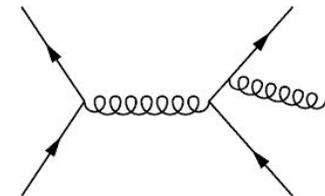
Most events are like this.



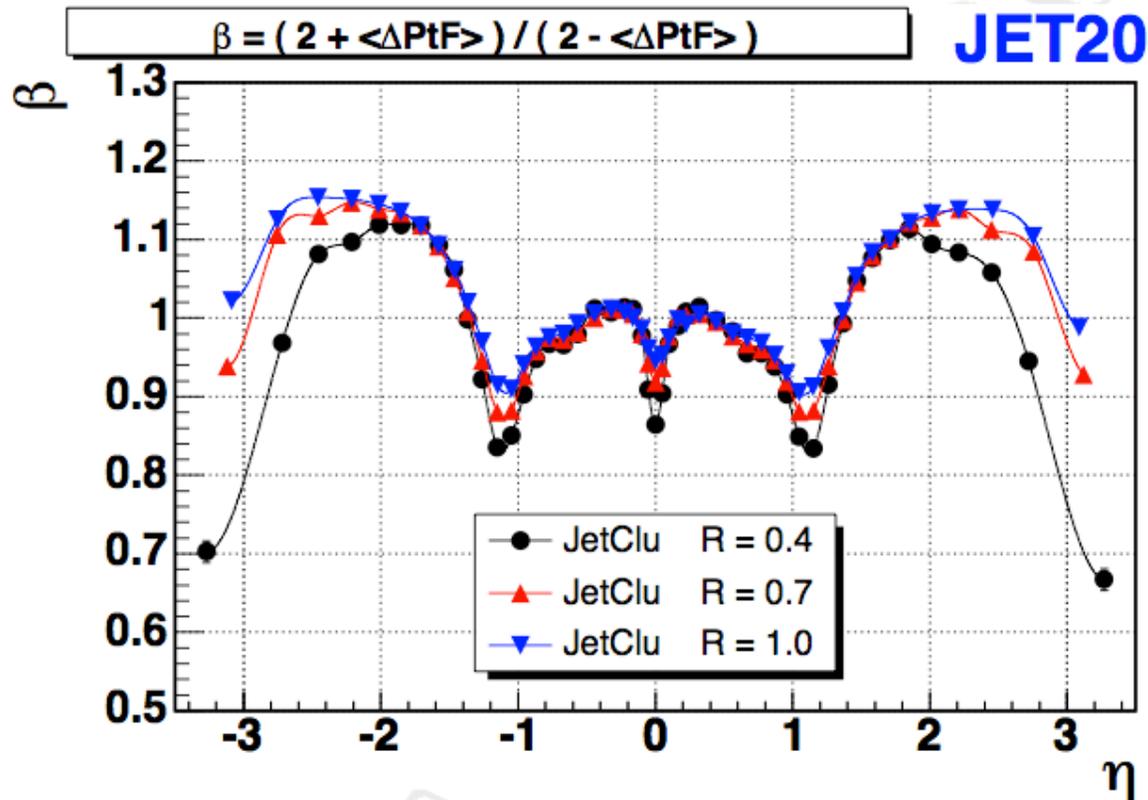
Some Contributing Processes:



Also: three or more jets possible.



# Non-Uniform Detector Response: Jet Energy Response for CDF



Data are more useful if some kinds of effects are calibrated out, like this one.

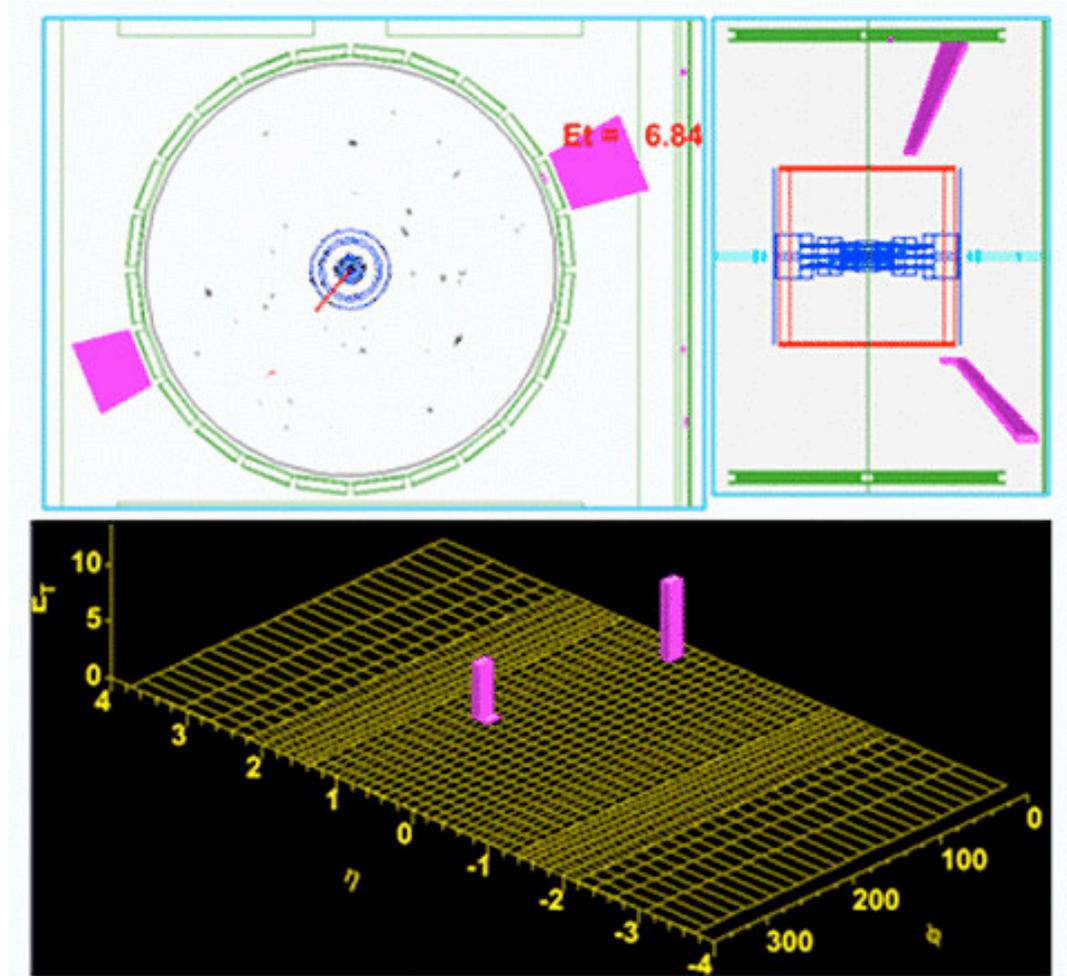
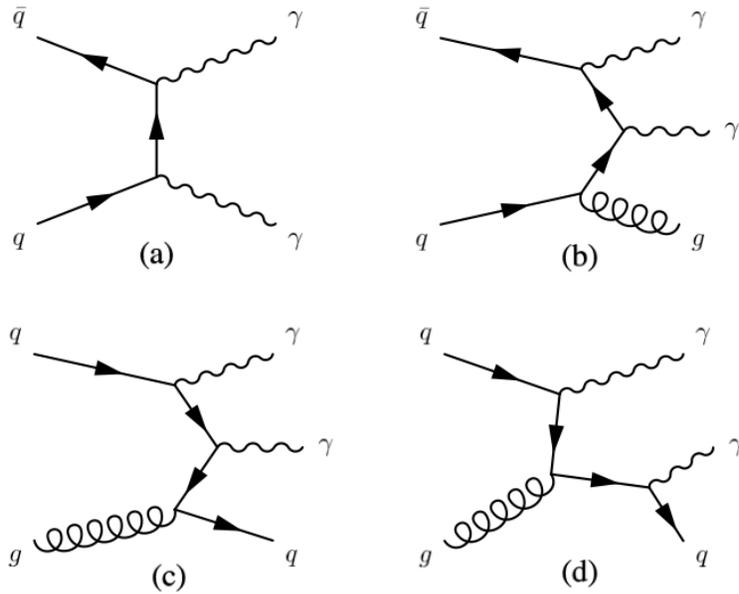
Energy resolution is better if all jets have the same energy scale.

A conundrum:

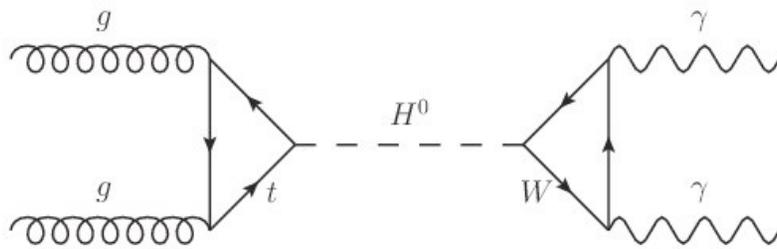
- Calibrations and corrections may fail to fully do their job
- Underlying detector performance may be mismodeled in the MC
- Easiest to understand modeling issues when the data are not corrected
- “Unfolding” can introduce model dependence that’s hard to undo

# A Di-Photon Event Collected by CDF

Some common nonresonant processes



A rare but exciting process:  
Higgs boson production and decay:

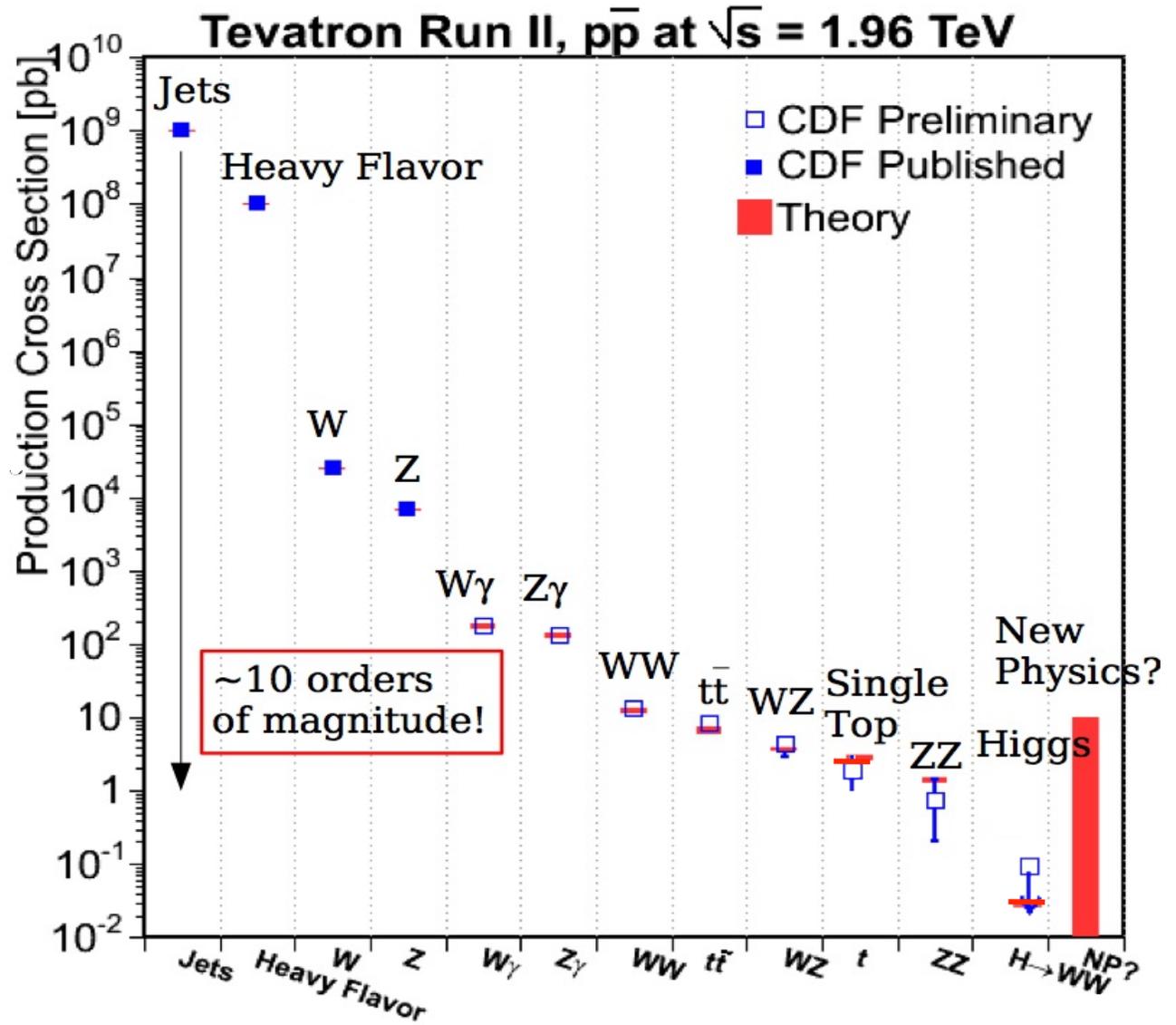


# Some Processes are Much More Rare than Others

The strong force is very strong  $\rightarrow$  high interaction rates.

Weak interactions suppressed by coupling strength, and heavy particle masses

Multiple heavy particles make events even more rare



# Data Processing Flow

- Detector – Raw Signals are Digitized
- Trigger – Some Collisions Retained, Most Discarded
- Online Monitoring – Data Quality
- Storage – Disk and Tape
  - Splitting data into separate streams depending on trigger types targeting specific event signatures
- Reconstruction
  - Digitized hits are clustered, energies summed, helical tracks fit to observed signals
  - Particle identification algorithms: jets, leptons, missing energy
  - Displaced vertex identification
- Calibration using reconstructed objects
- Re-Reconstruction using calibration
- Analysis

Each collision event is independent of all others. “Embarrassingly Parallel!”

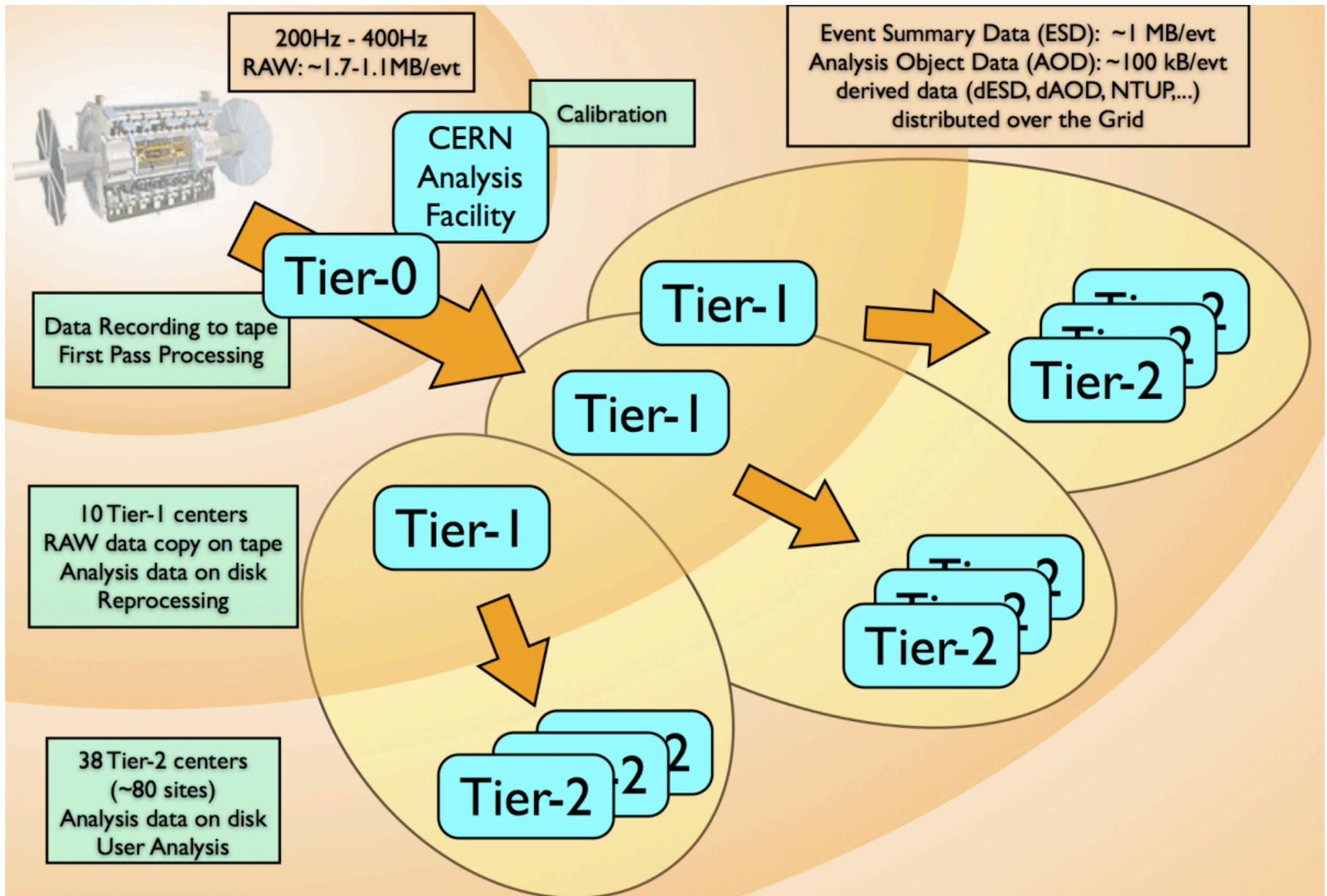
# Monte Carlo (“MC”) Simulation

Science output needs a comparison of the data with predictions.

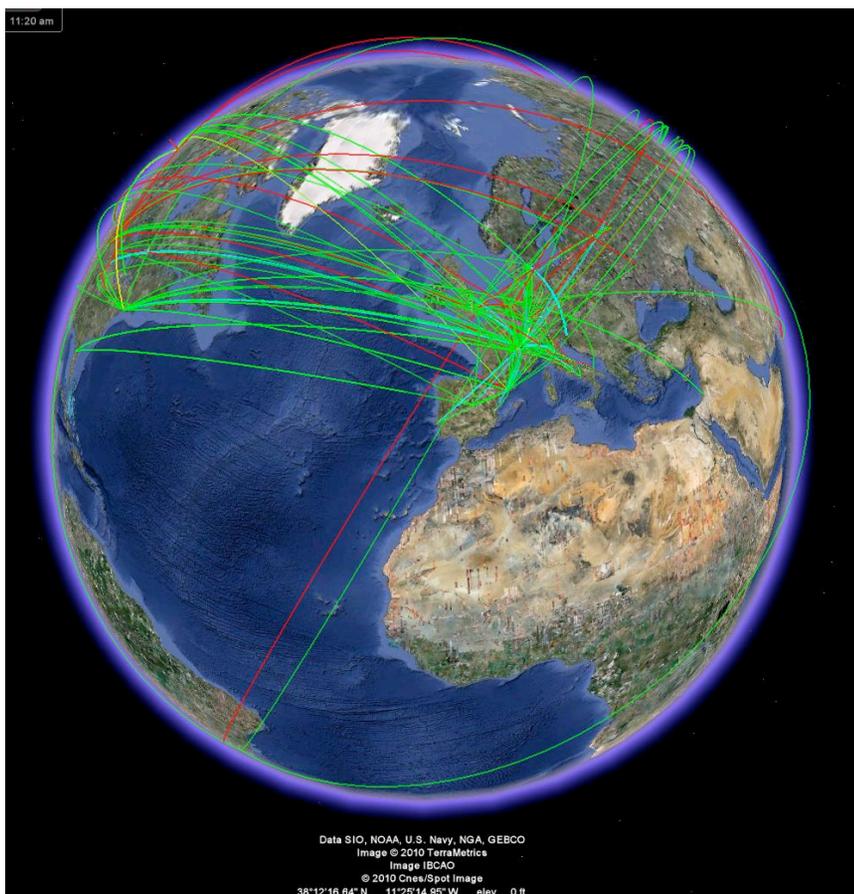
Many uses:

- Design the Experiment: Which sorts of signals is the equipment most sensitive to? How do we build our experiment to meet our goals?
- Develop Reconstruction Algorithms
- Predicting signal and background data yields
  - Need input from theoretical calculations and other experiments
  - Need detailed simulation of experimental apparatus and trigger
  - Once data are available, use data control samples to “tune” Monte Carlo predictions.
  - Ideally some predictions of signals and backgrounds are entirely “data-driven”, but there’s usually an extrapolation step from one sample to another, and that needs MC.
- Most (all?) systematic uncertainties in particle physics analyses enter via the signal and background predictions.
- Frequently need many alternate MC samples to estimate systematic uncertainties

# LHC Experiments' Multi-Tiered Computing Infrastructure



# Global WLCG Sites in Europe, the Americas, Asia, Australia



Current WLCG sites



>150 computing centers in nearly 40 countries

<http://wlcg.web.cern.ch/>

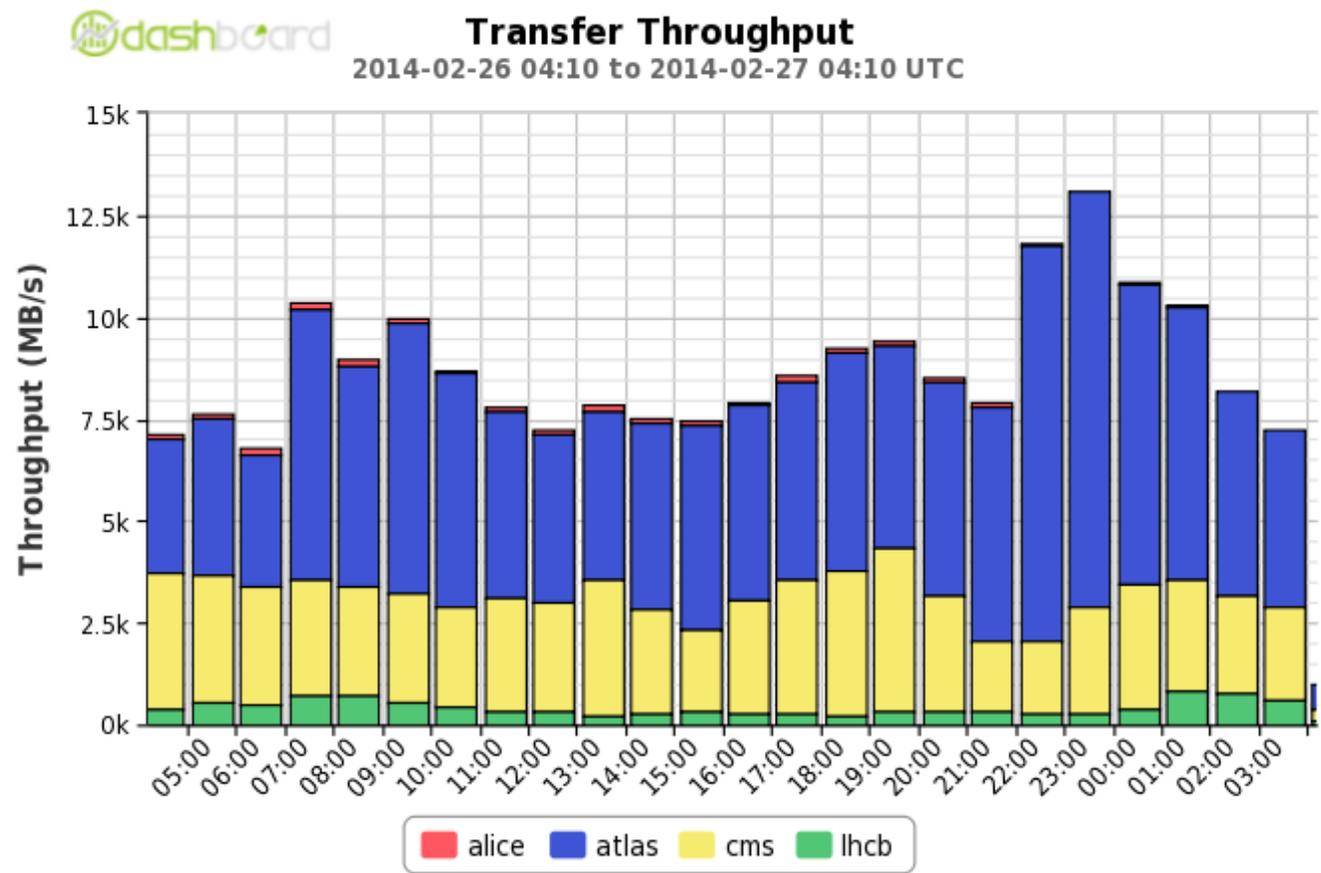
# WLCG Data Handling Rate

<http://wlcg.web.cern.ch/>

10 GB/second data analysis throughput,  
and the LHC is not even taking data!

25 PB per year  
produced by  
the LHC  
Experiments

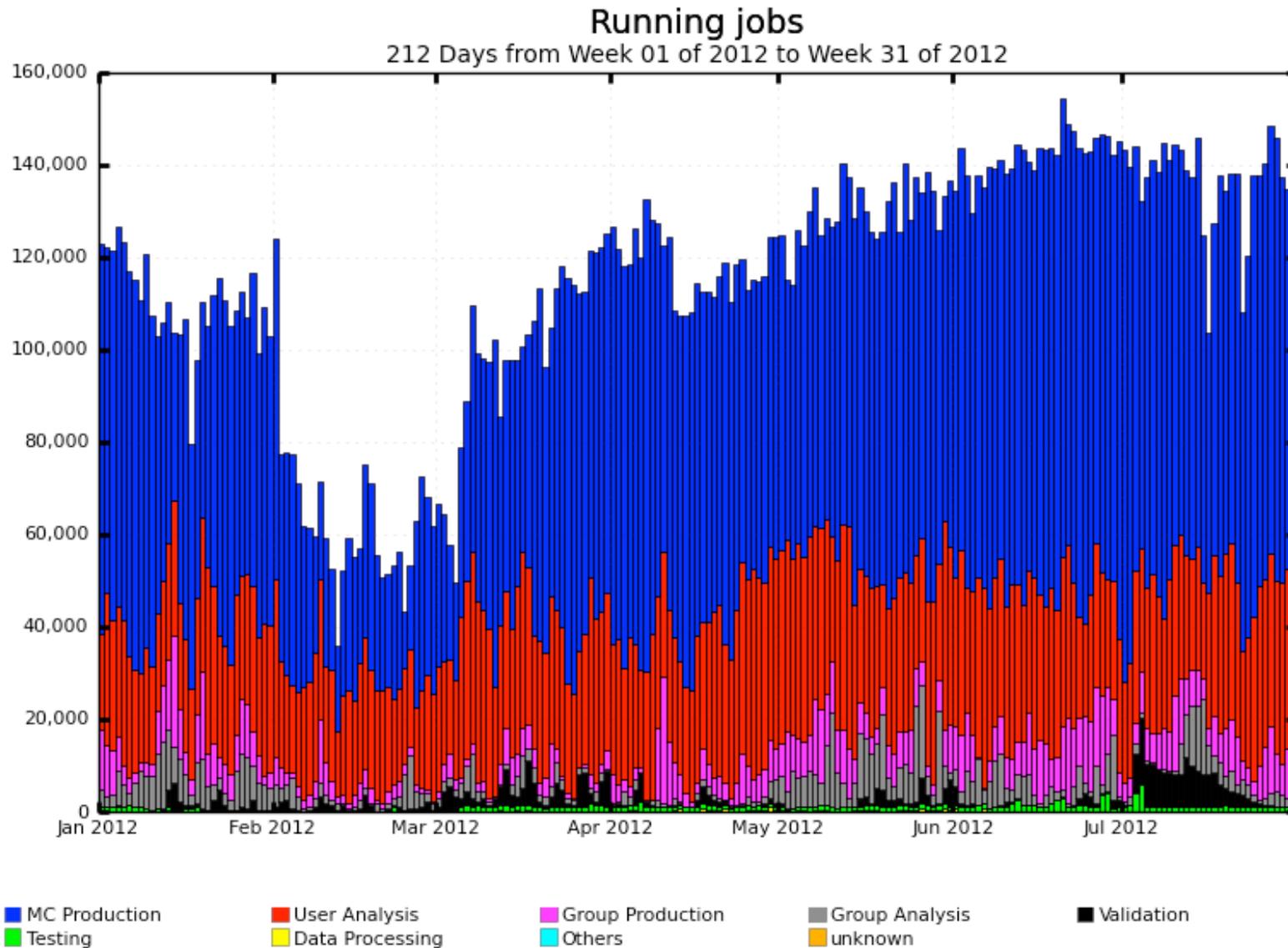
~3000 Collaborators  
each on ATLAS, CMS



# ATLAS: Before the 2012 Higgs Observation, up to 150,000 jobs running at a time

CMS: Similar

CDF maxed out at ~10K jobs running simultaneously



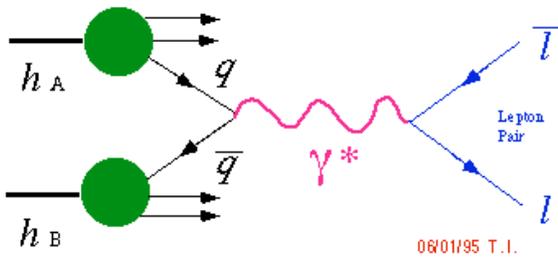
Maximum: 154,378 , Minimum: 35,776 , Average: 114,517 , Current: 137,942

# A Very Nice Dimuon Event from CMS

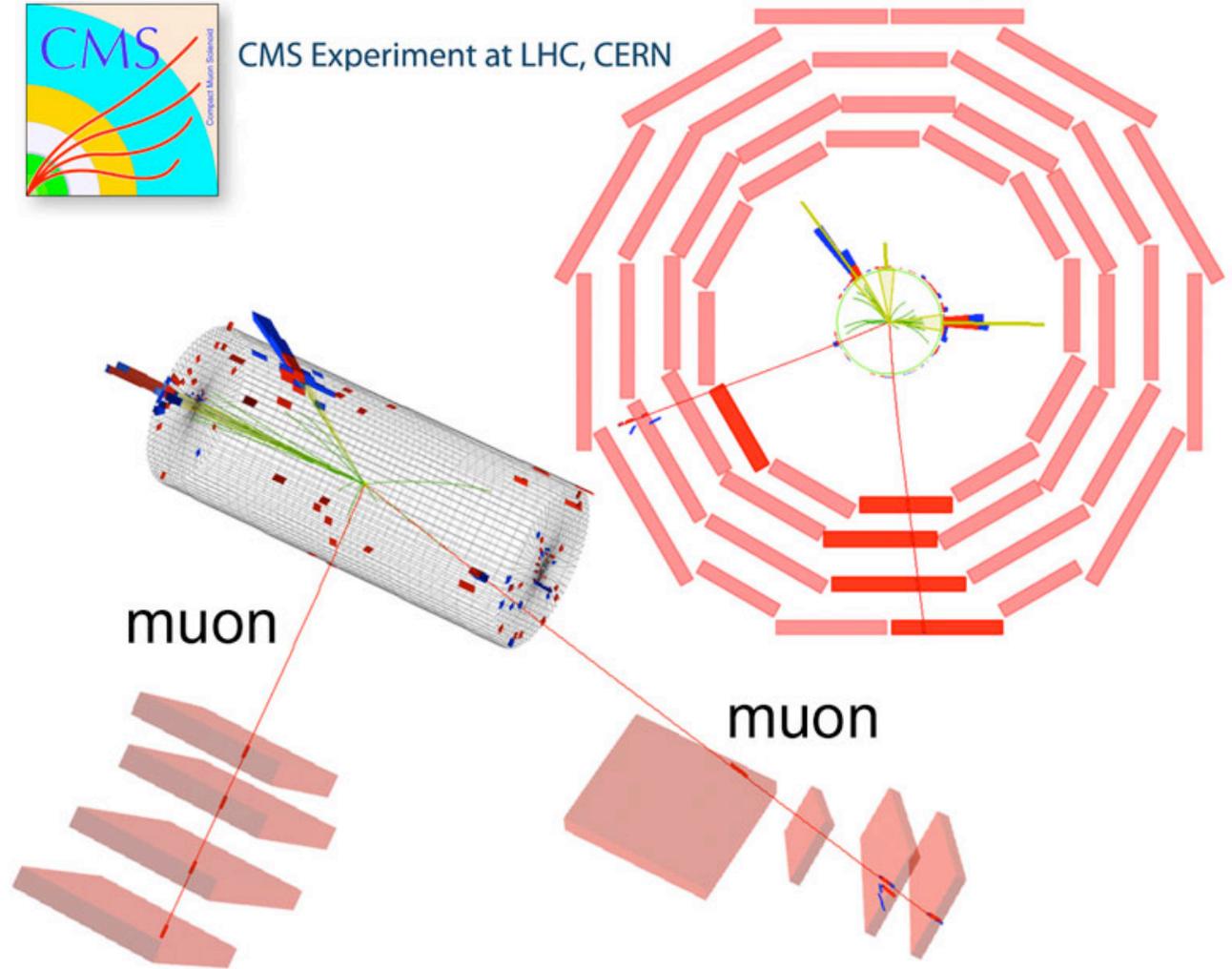
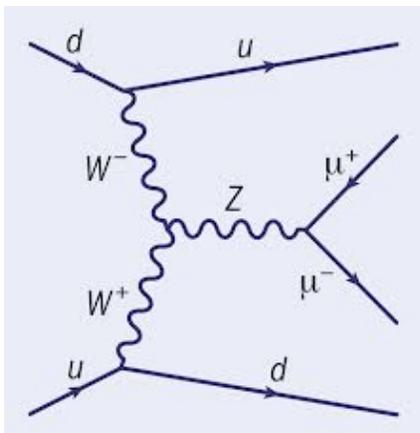


CMS Experiment at LHC, CERN

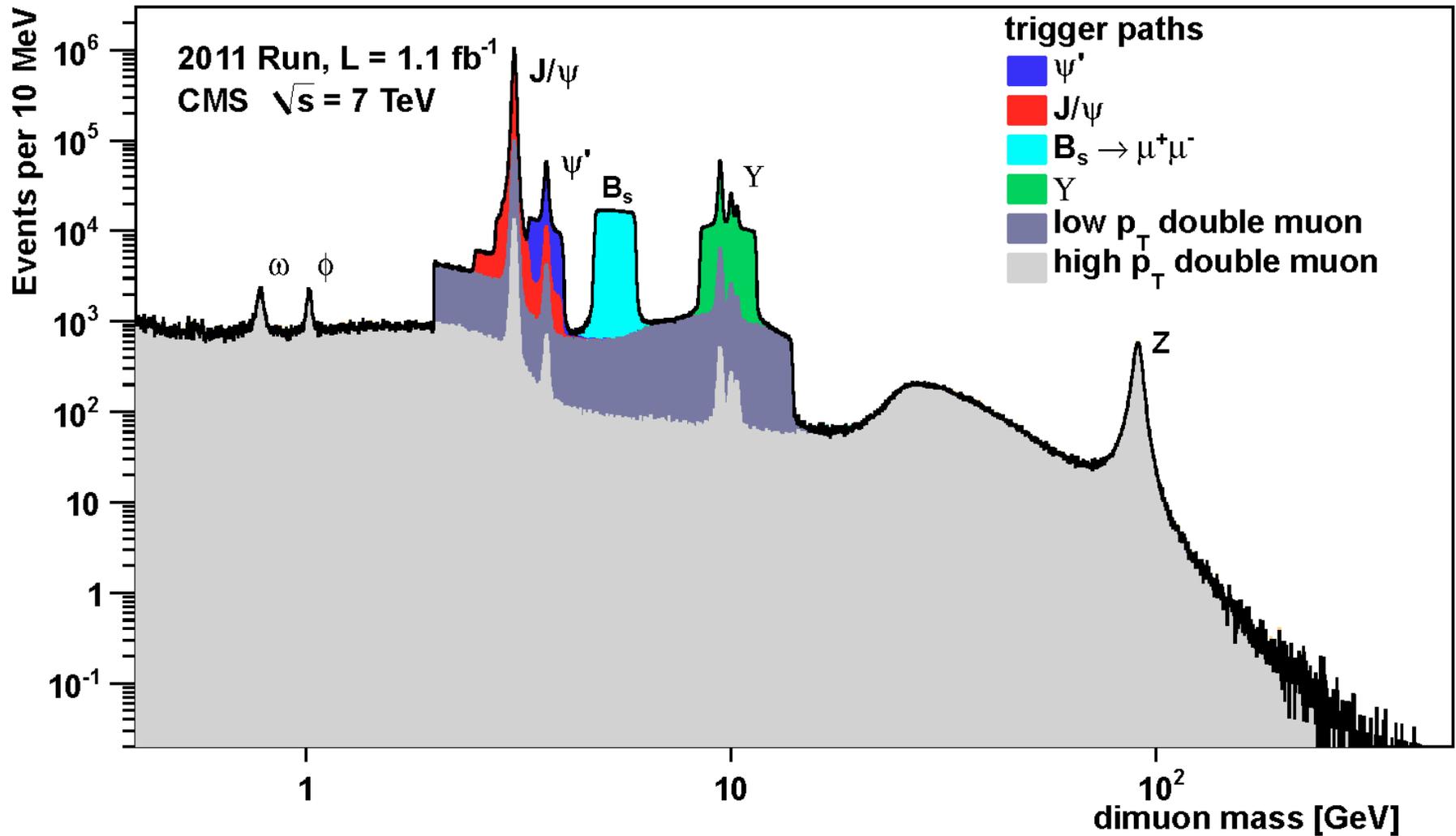
## The Drell-Yan Process



## And Vector-Boson Fusion



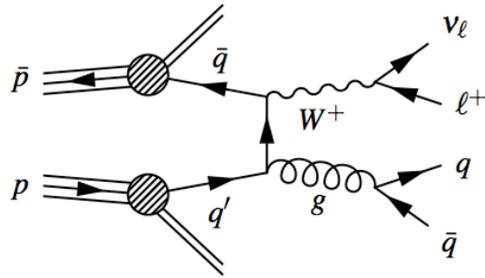
# Dimuon Mass Spectrum in a Small Initial Data Sample



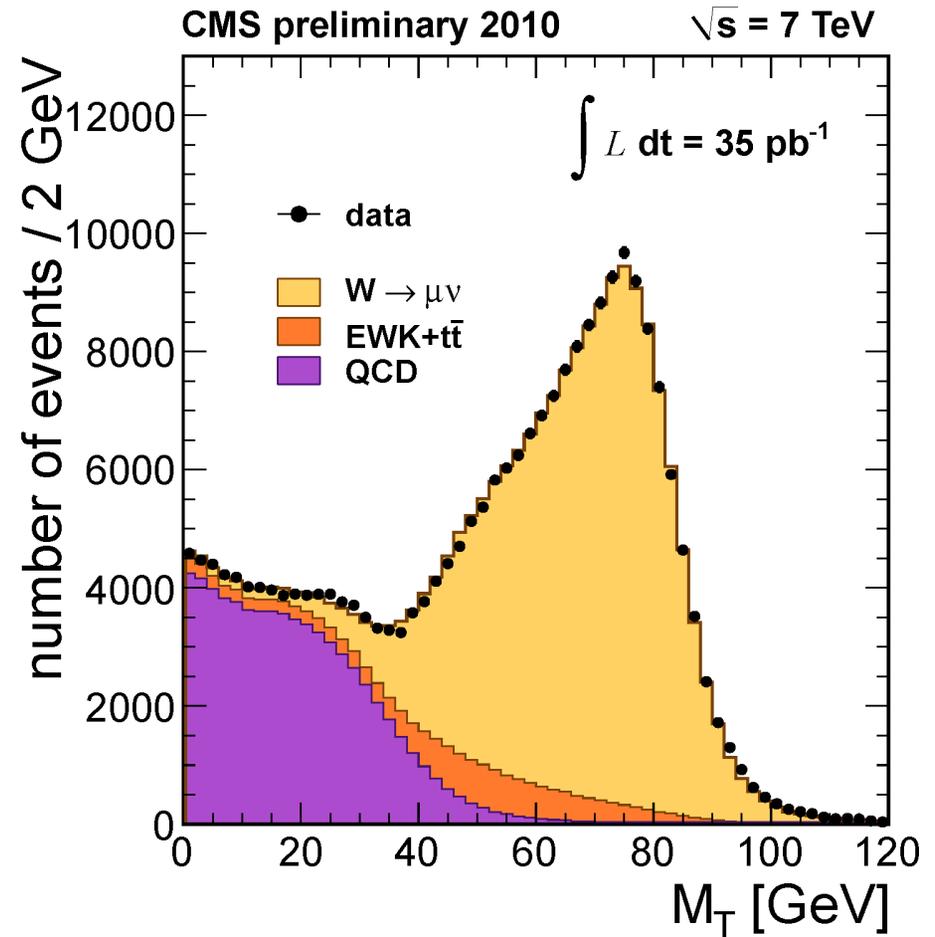
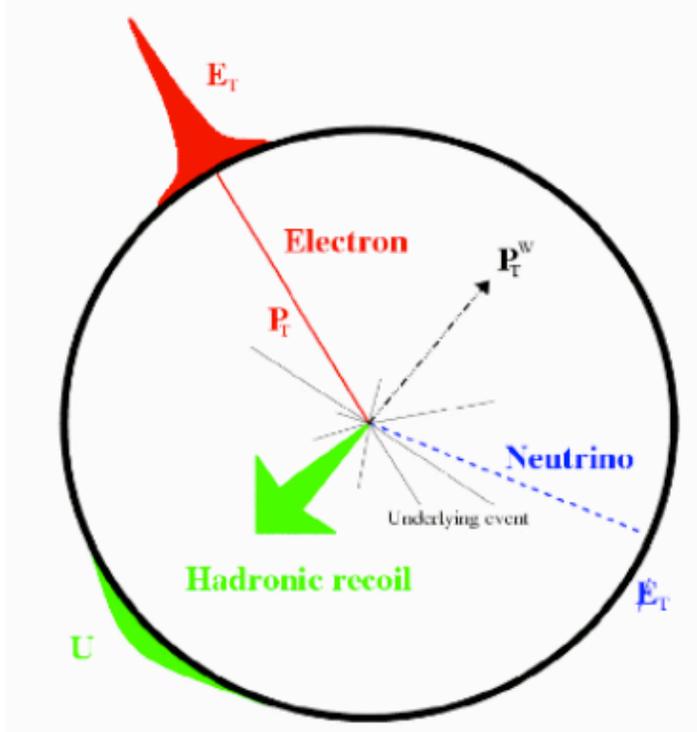
Triggers have different acceptances for different mass ranges

[https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsMUO#Full\\_invariant\\_mass\\_spectrum\\_of](https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsMUO#Full_invariant_mass_spectrum_of)

# Partially Reconstructed $W^\pm \rightarrow l^\pm \nu$ Bosons (missing the neutrino!)



Event signature: Identified lepton (e,  $\mu$ , sometimes  $\tau$ ), plus missing transverse momentum, plus hadrons

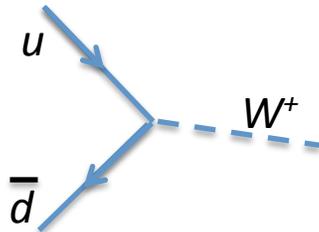


$$M_T = \sqrt{2E_T^l E_T (1 - \cos \Delta\phi)}$$

# Measurement of the Forward-Backward Asymmetry of W Bosons with D0

MSTW 2008 NLO

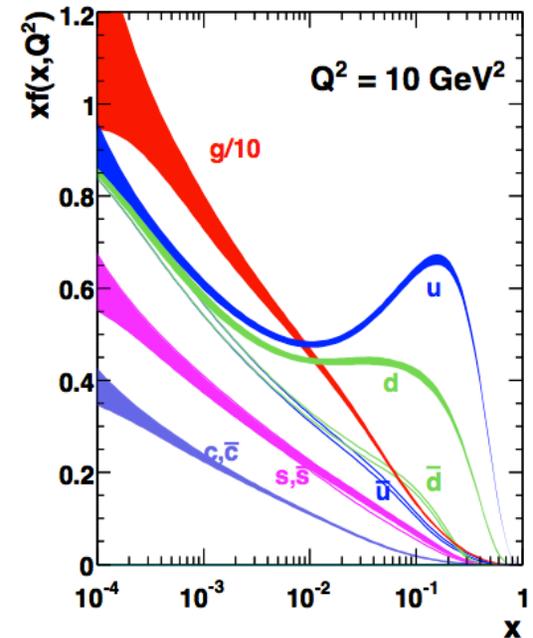
Tevatron: p-pbar collider



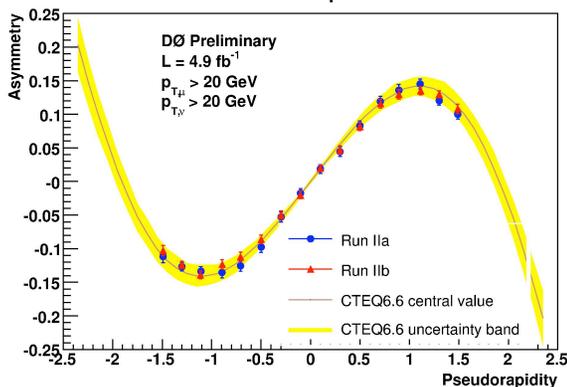
Forward-Backward Asymmetry comes largely from the difference between u and d PDF's. The Compton diagram  $gq \rightarrow Wq$  also participates.

$$A(\eta) = \frac{1}{(1 - 2g)} \left[ \frac{N^+(\eta) - N^-(\eta)}{N^+(\eta) + N^-(\eta)} \right]$$

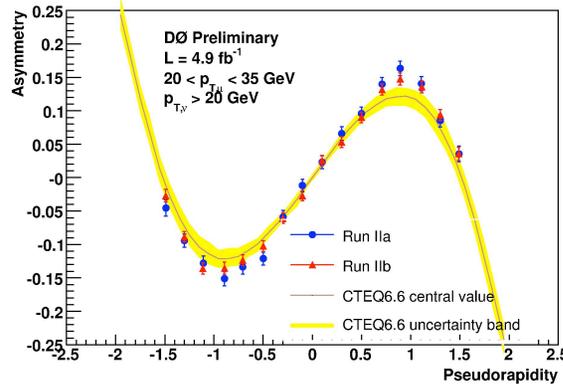
Is the LEPTON asymmetry (directly observable).  $p_{z,W}$  ambiguous due to neutrino solution.



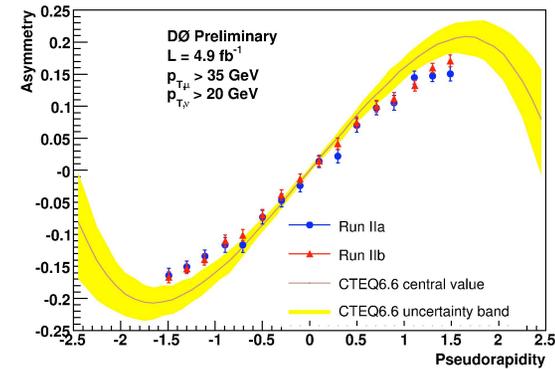
Inclusive  $p_{T\mu} > 20$  GeV



$20 < p_{T\mu} < 35$  GeV



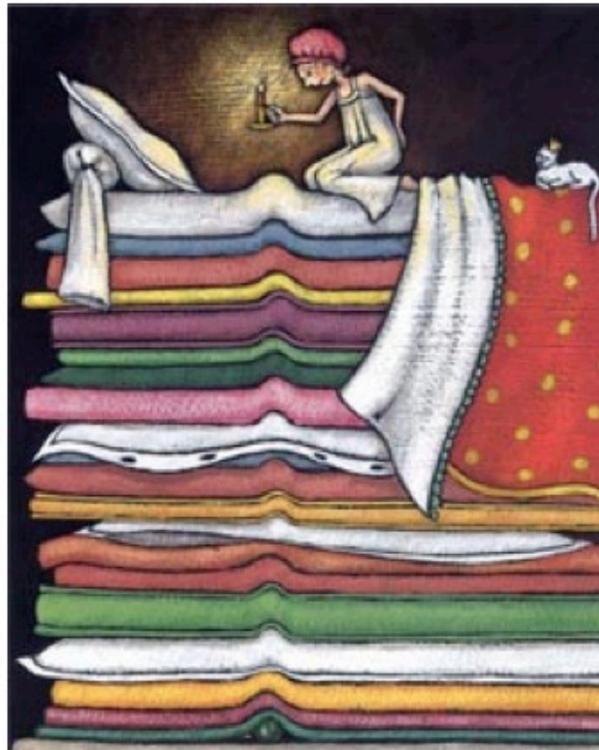
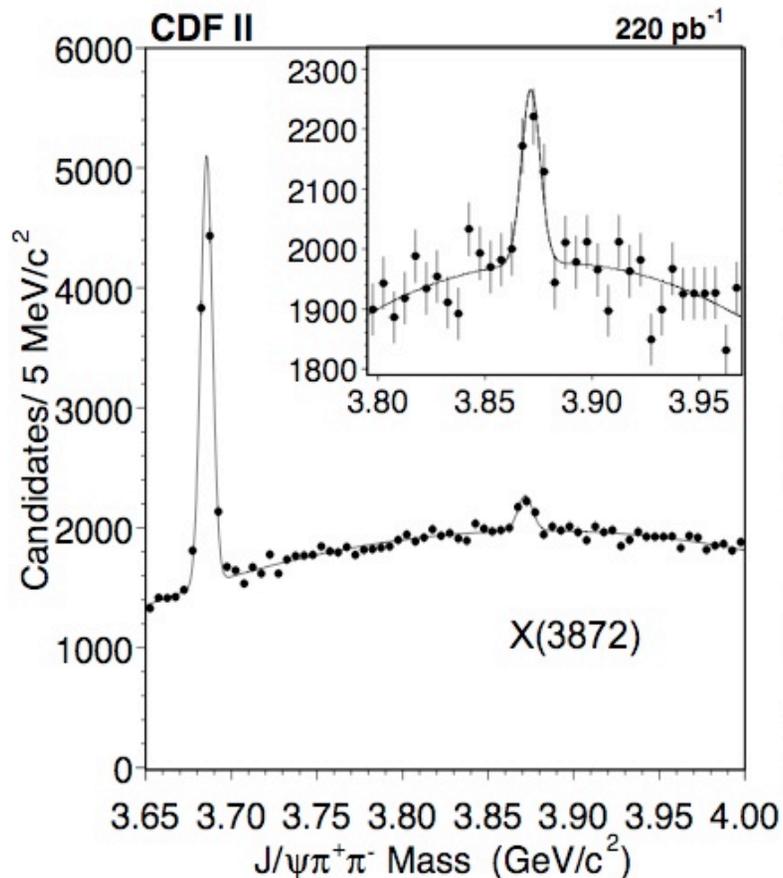
$p_{T\mu} > 35$  GeV



$g$  accounts for the charge misidentification rate, determined with like-sign  $Z \rightarrow \mu\mu$  candidate events. Solenoid reversal results are consistent.

# “On-Off” Example

Select events with  $J/\psi(\rightarrow \Pi) \pi^+\pi^-$  candidates. Lots of nonresonant background which is poorly understood *a priori*, but there's a lot of it.

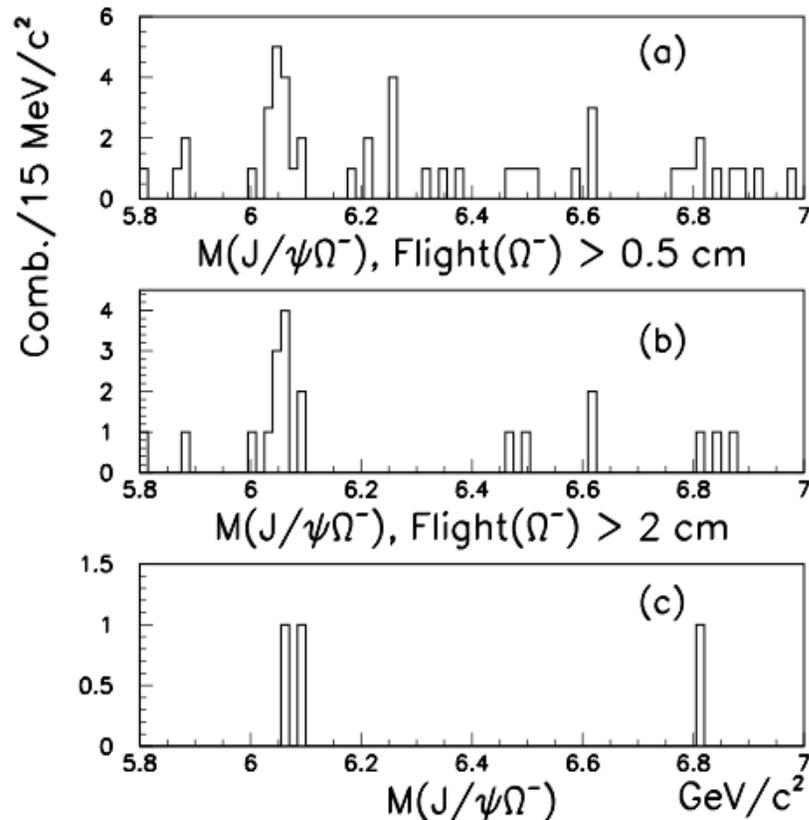


Typical strategy:  
Fit the background outside of the signal peak,  
and interpolate the background under the signal to subtract it off.

The ratio of events in the sidebands to the background prediction under the signal is called  $\tau$

Guess a shape that fits the backgrounds, and fit it with a signal.

# “Weak” Sideband Constraints



CDF's  $\Omega_b$  observation  
paper:

**Phys.Rev. D80 (2009) 072003**

FIG. 8: (a,b) The invariant mass distribution of  $J/\psi\Omega^-$  combinations for candidates where the transverse flight requirement of the  $\Omega^-$  is greater than 0.5 cm and 2.0 cm. (c) The invariant mass distribution of  $J/\psi\Omega^-$  combinations for candidates with at least one SVXII measurement on the  $\Omega^-$  track. All other selection requirements are as in Fig. 5(c).

# No Sideband Constraints?

Example: Counting experiment, only have a priori predictions of expected signal and background

All test statistics are equivalent to the event count – they serve to order outcomes as more signal-like and less signal-like. More events == more signal-like.

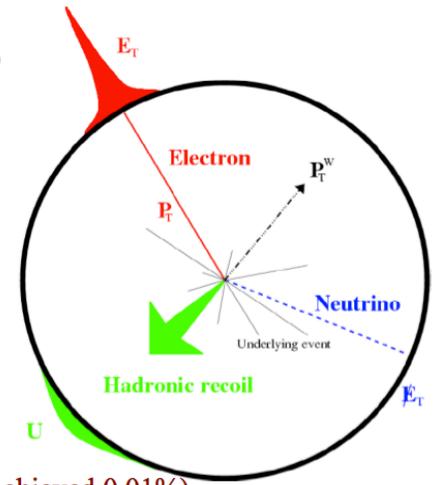
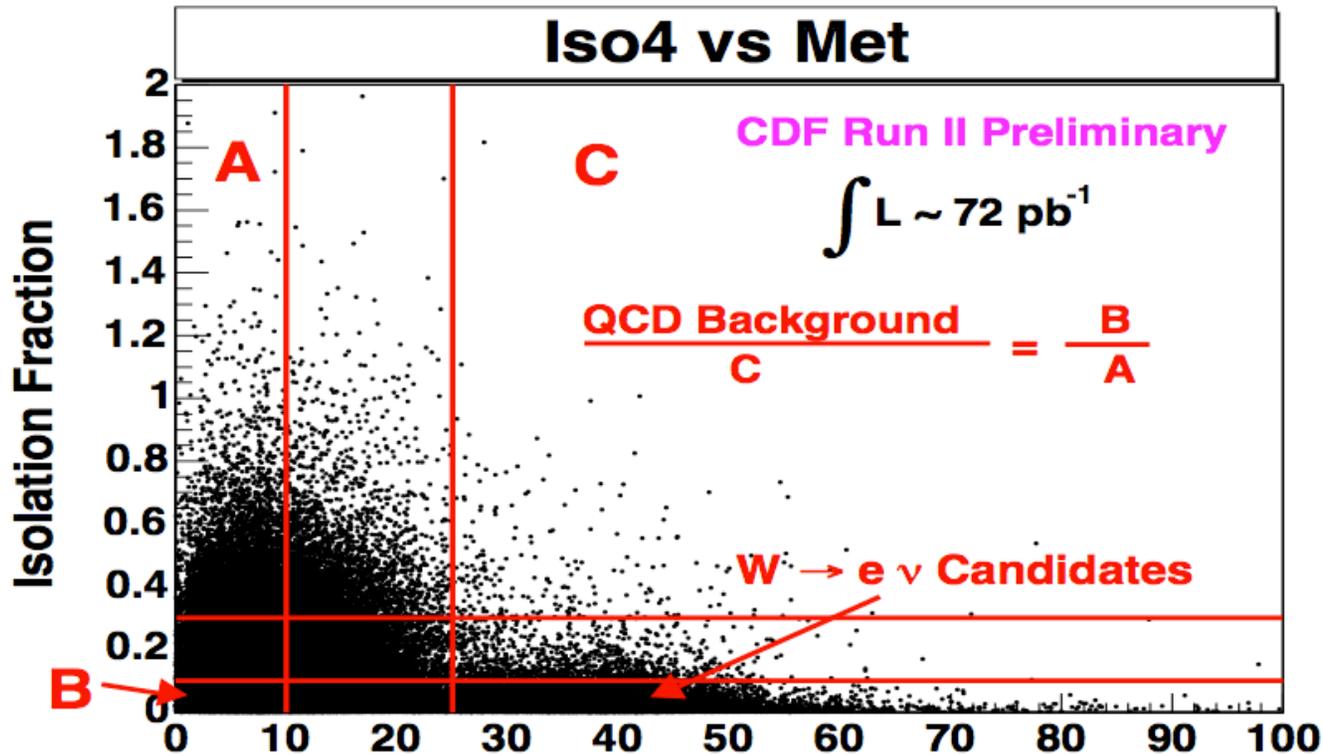
Classical example: Ray Davis's Solar Neutrino Deficit observation. Comparing data (neutrino interactions on a Chlorine detector at the Homestake mine) with a model (John Bahcall's Standard Solar Model). Calibrations of detection system were exquisite. But it lacked a standard candle.

How to incorporate systematic uncertainties? Fewer options left.

Another example: Before you run the experiment, you have to estimate the sensitivity. No sideband constraints yet (except from other experiments).

# “ABCD” Methods

CDF’s W Cross Section Measurement



Isolation fraction =

Energy in a cone of radius 0.4 around lepton candidate not including the lepton candidate / Energy of lepton candidate

Want QCD contribution to the “D” region where signal is selected.

Assumes: MET and ISO are uncorrelated sample by sample

Signal contribution to A, B, and C are small and subtractable

ABCD methods are really just on-off methods where  $\tau$  is measured using data samples

# “ABCD” Methods

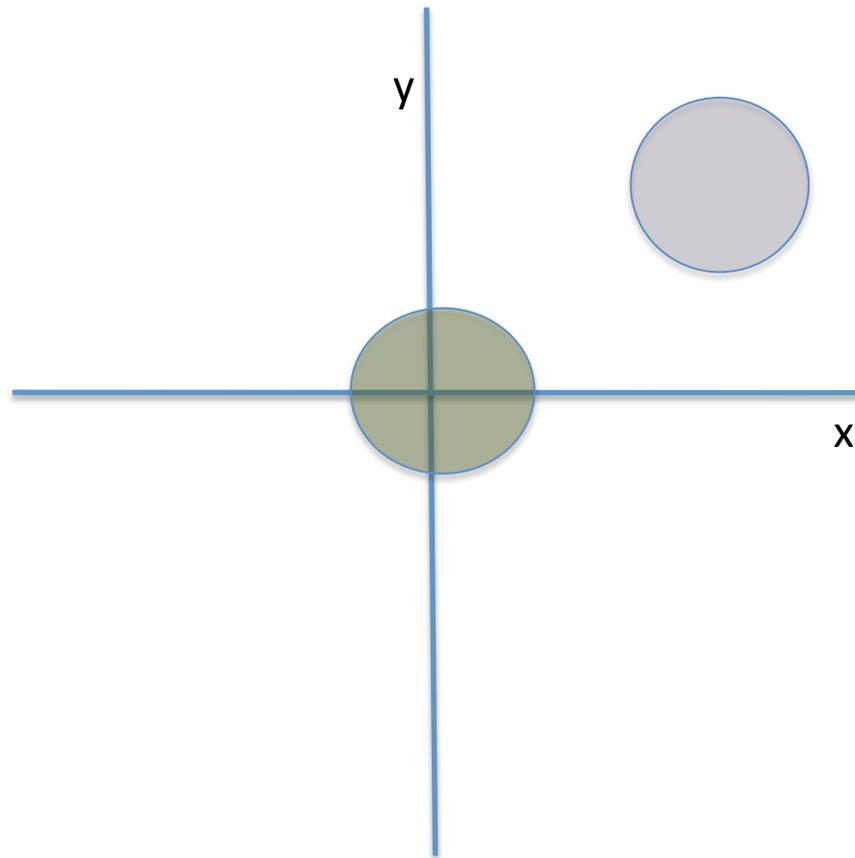
## Advantages

- Purely data based, good if you don't trust the simulation
- Model assumptions are injected by hand and not in a complicated Monte Carlo program (mostly)
- Model assumptions are intuitive

## Disadvantages

- The lack of correlation between MET and ISO assumption may be false. e.g., semileptonic B decays produce unisolated leptons and MET from the neutrinos.
- Even a two-component background can be correlated when the contributions aren't by themselves.
- Another way of saying that extrapolations are to be checked/assigned sufficient uncertainty
- Works best when there are many events in regions A, B, and C. Otherwise all the problems of low stats in the “Off” sample in the On/Off problem reappear here. Large numbers of events → Gaussian approximation to uncertainty in background in D
- Requires subtraction of signal from data in regions A, B, and C → introduces model dependence
- Worse, the signal subtraction from the sidebands depends on the signal rate being measured/tested.
  - A small effect if  $s/b$  in the sidebands is small
  - You can iterate the measurement and it will converge quickly

# The Sum of Uncorrelated 2D Distributions may be Correlated



Knowledge of one variable helps identify which sample the event came from and thus helps predict the other variable's value even if the individual samples have no covariance.

# Multivariate Analyses

These are an important tool for optimizing sensitivity

- Reduce expected uncertainties on measurements
- Raise chances of discovering particles that are truly there
- Improve the ability to exclude particles that are truly absent

## **BUT:**

- There are many ways to make a mistake with them: More work!
  - Optimizing them
    - Best input variables
    - Best choice of MVA
  - Validating them
    - Validate modeling of inputs *and* outputs
    - Check for overtraining
  - Propagate systematic uncertainties through them
    - Rates
    - Shapes
    - Bin-by-bin

# Example MVA Methods

Coded up in TMVA – comes with recent versions of ROOT

- Feed-Forward Neural Networks (multi-layer perceptrons)  
Abbreviations: NN, ANN, MLP

All are just functions of the reconstructed event observables.

- Boosted Decision Trees

We could devise our own functions if it suited our needs and we were smart enough.

- Matrix Elements

These are machine derived, so we call it *machine learning*.

See, for example, P. Bhat, **Ann.Rev.Nucl.Part.Sci. 61 (2011) 281-309**

# A Neural Network

Inputs to node  $i$  have weights  $w_j$ . Outputs are sigmoid functions of the weighted inputs.

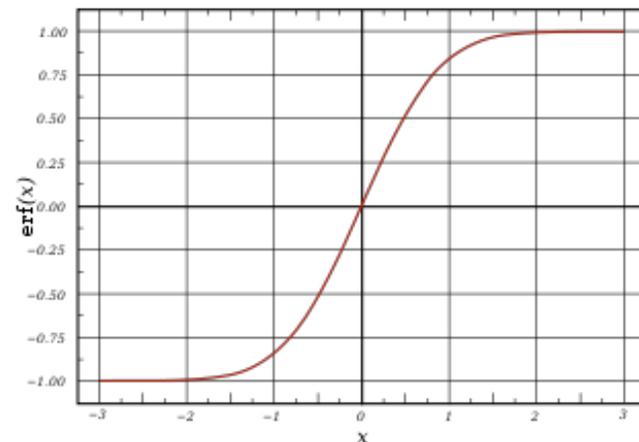
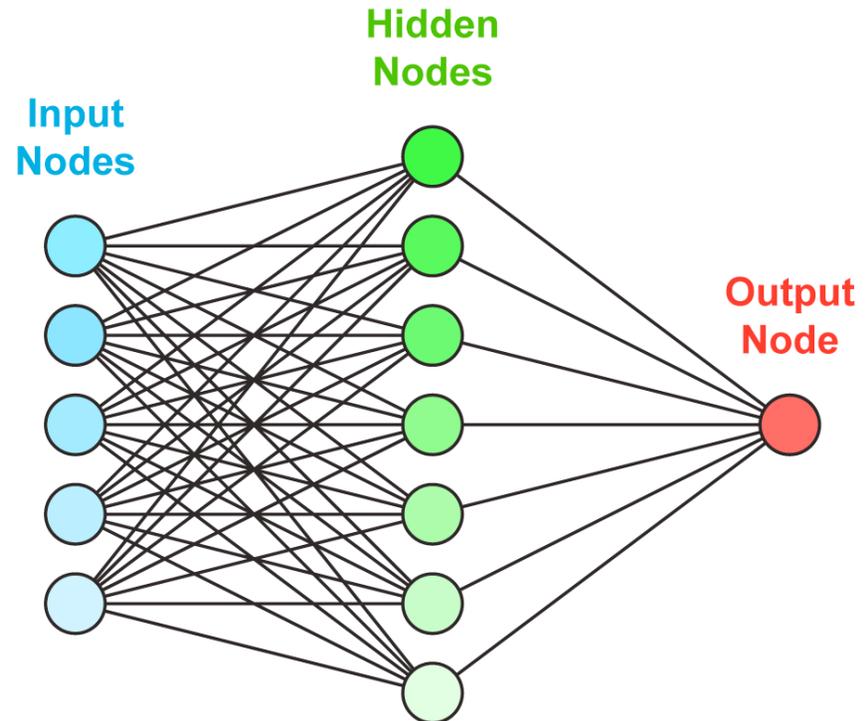
$$o_i = S\left(\sum_j w_{ij} v_j\right)$$

S is any of these:

$S(x) = \text{Tan}^{-1}(x)$ ,  
 $x/\text{sqrt}(1+x^2)$   
 $1/(1+\exp(-x))$   
 $\tanh(x)$

Or any other s-shaped function

Main features: Nonlinearity,  
3/5/14  
monotonicity



# Training a Neural Network

The weights  $w_{ij}$  are arbitrary. We may choose them, as well as the structure of the network, to optimize our analysis.

We would like to classify events as signal (output = 1) or background (output = 0).

Ad-hoc figure of merit: Minimize the sum of squares of errors made by the network:

$$E = \sum_{\text{events}} (O_{\text{desired}} - O_{\text{obtained}})^2$$

Why this function?

Well, it's easy to differentiate with respect to the weights for each event.

Back-propagation training: Loop over training events (some signal, some background) and adjust the weights each time according to how the adjustment will improve  $E$ .

Weighted events are okay with most MVA training programs. But it's worth checking to see how they respond to negative-weight events!

Adjustable parameters: "learning rate" – how big the steps in  $w_{ij}$  are scaled by the derivative. How many events to use to train, how many spins through the training

# Training a Neural Network

Critique of standard Neural Networks:

- No one really cares about  $E = \sum_{events} (O_{desired} - O_{obtained})^2$

We care about the best expected uncertainty

on cross section or property measurements

Best expected limits if a particle is not there

Best expected chances of discovery if a particle is there

- Addition of non-useful variables (random noise) can hurt overall performance
- Inputs can have very broad ranges of behavior  
discrete, large ranges, small ranges, mixtures ..

(can be mitigated by clever preprocessing)

- Advantages – can make use of correlations between input variables by forming nearly arbitrary functions of them.
- Experience with them shows that it is usually better to
  - Give it the best variables already as inputs
  - Pre-select the data into samples so the NN has less work to do  
(fewer sources of background that are important)

# Overtraining

If a training sample is small, and the NN has many nodes and weights, it is possible for the NN to “learn” the properties of individual events in the training sample and get them classified correctly all the time.

This may not be representative of any other sample (like the data).

The network may not perform as well as it thinks it is performing if only the training sample is used to judge.

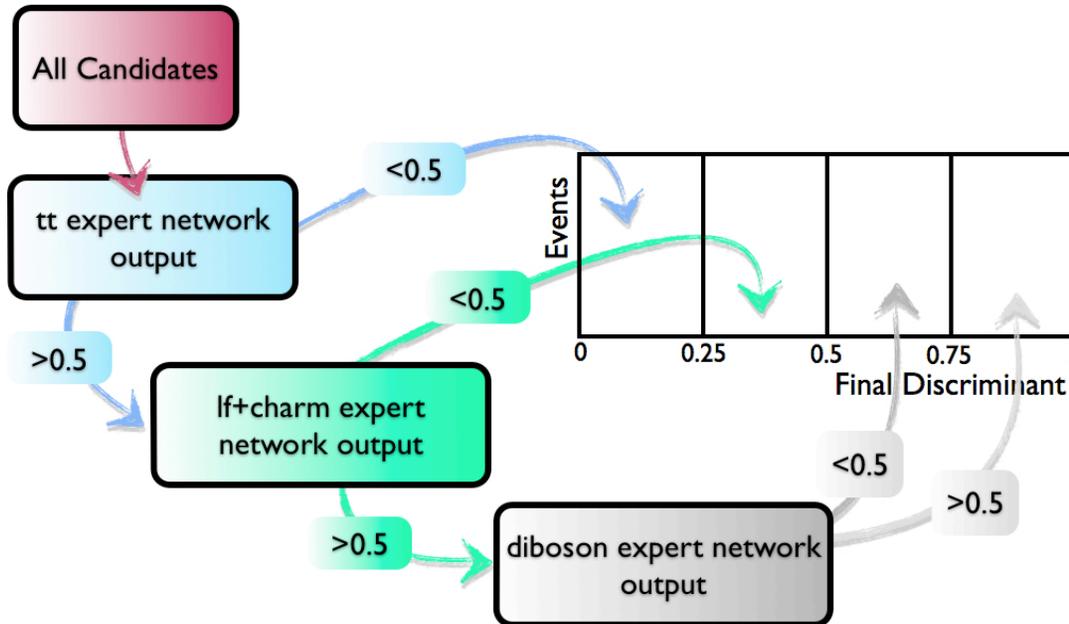
Ensure that overtraining does not affect correctness:

*Use different events to train a NN and to test it.*

Even if it's overtrained, then the independent evaluation of its performance is not systematically biased by this effect.

The NN may not be fully optimal, however.

# Example of Giving NN's Some Help – Cascading NN Stages



CDF's

ZH → llbb search

Further help:

Event selection is lljj, with  $m_{ll}$  near  $M_Z$ .  
One or two b-tags, with loose or tight b-tagging requirements.

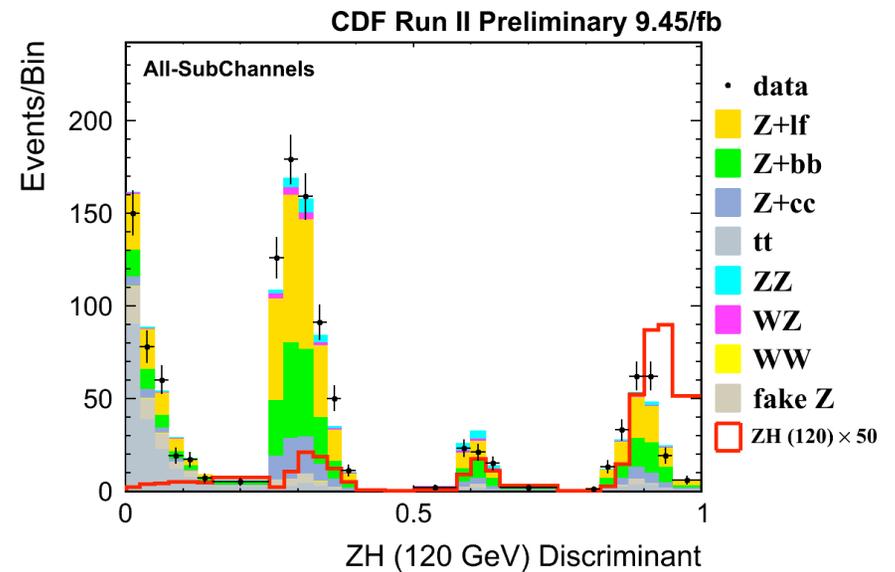
Split sample up into b-tag categories:

Tight-Tight

Tight-Loose

Single Tight

Loose-Loose



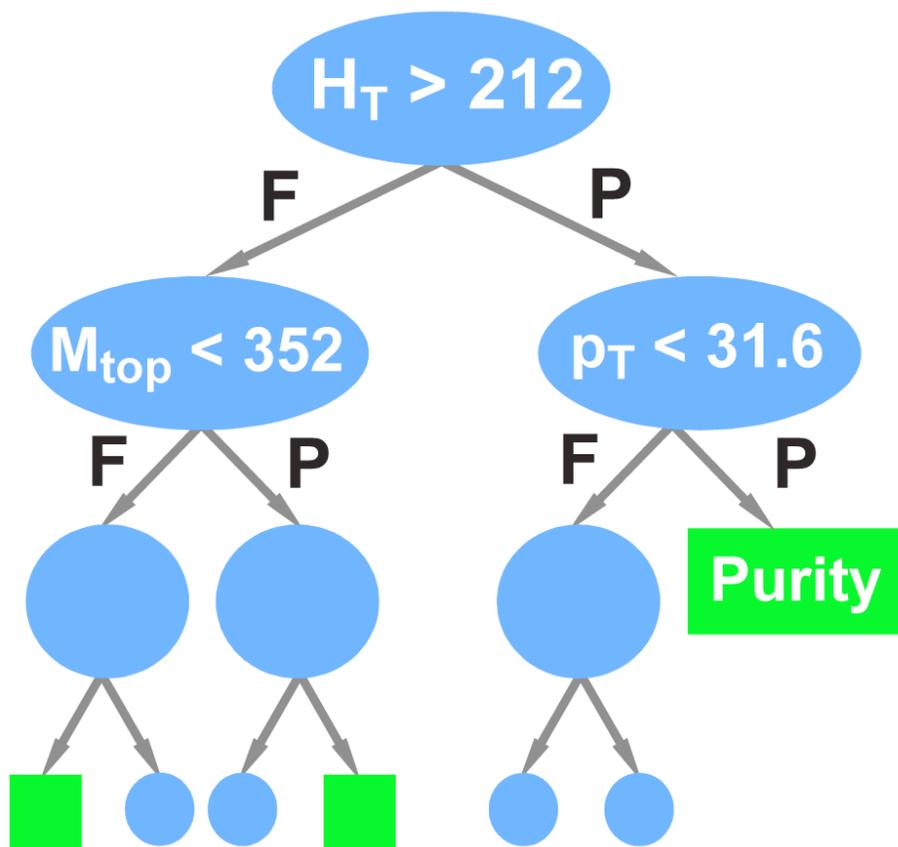
# (Boosted) Decision Trees

Original work by J. Friedman in the 1980's

Look through the list of input variables; Try sliding a cut along each one and find the cut on a variable that maximizes the purity difference on both sides of the cut.

“Gini index” –  $p(1-p)$ , where  $p$ =purity zero for perfect separation.

Iterate the search for the best cut on the best variable for each subset of events thus divided. Stop when you run out of enough MC to predict the contents of a sample.



- Advantages over NN's: not as sensitive to the addition of “noise” variables – they just never get cut on
- The Gini index is also just a proxy for what we really care about.

# Matrix-Element Discriminants

- Calculate probability density of an event resulting from a given process

$$P(p_l^\mu, p_{j1}^\mu, p_{j2}^\mu) = \frac{1}{\sigma} \int d\rho_{j1} d\rho_{j2} dp_v^z \sum_{comb} \phi_4 |M(p_i^\mu)|^2 \frac{f(q_1)f(q_2)}{|q_1||q_2|} W_{jet}(E_{jet}, E_{part})$$

Phase space factor:  
Integrate over unknown  
or poorly measured  
quantities

Parton distribution functions

Inputs:  
lepton and jet 4-vectors -  
no other information  
needed!

Matrix element:  
Different for each process.  
Leading order, obtained from  
MadGraph

Transfer functions:  
Account for  
detector effects in  
measurement of jet  
energy

- The input variables are the same for all matrix elements – adding a new matrix element requires more calculation but does not use any different information from the data

# Matrix-Element Discriminants

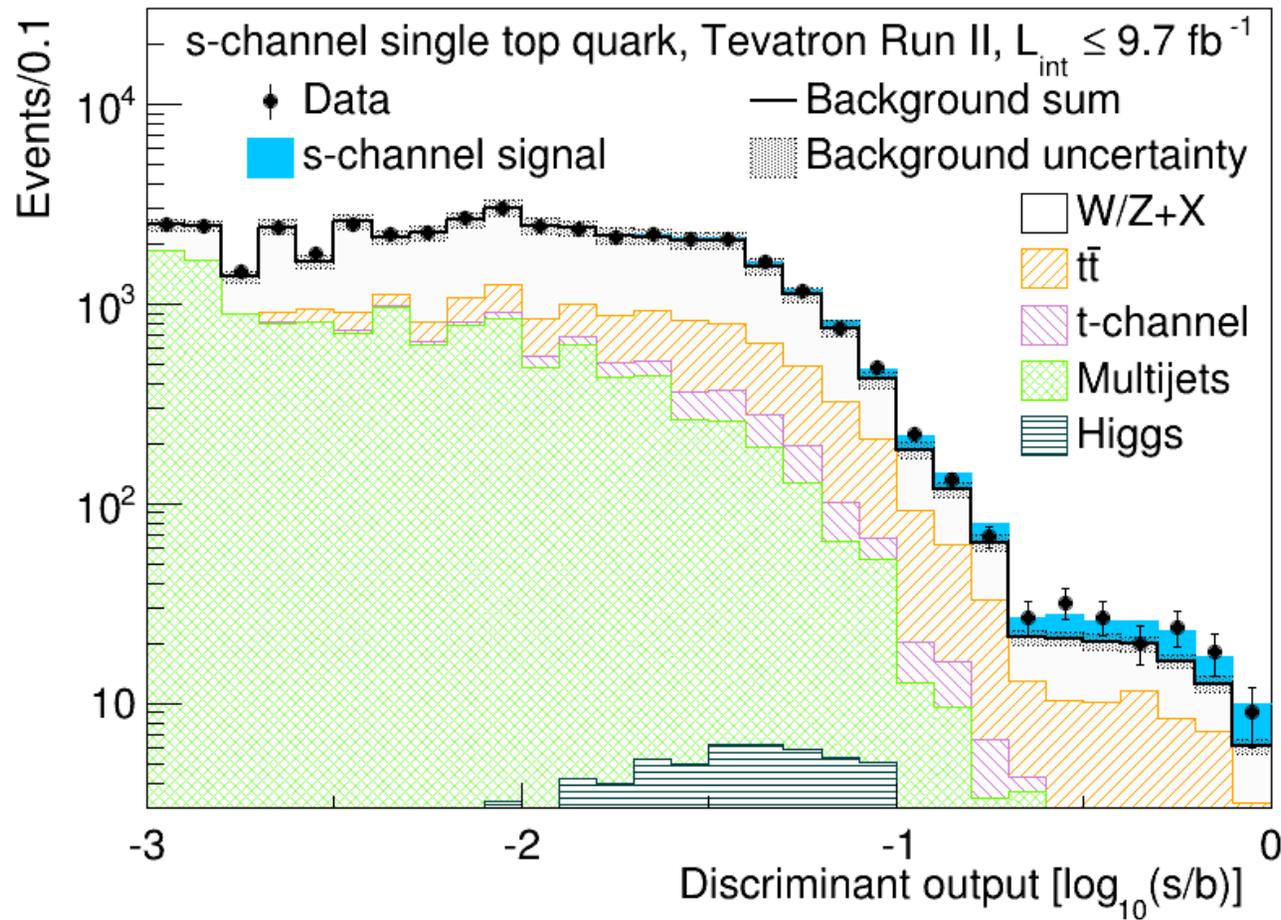
In principle, nothing performs better than these.

If processes cannot be separated because they contribute to the final state in the same way, this is all there is.

## **BUT:**

- Four-vectors are imperfectly measured. Transfer functions are also imperfect.
- Only the modeling needs systematics; construction of the discriminant does not incur additional systematics, so even if the discriminant is imperfect or naive, it's okay – just an optimization question.
- Matrix elements are usually leading-order only.
- Particles are sometimes not reconstructed at all, even when they should be
- Some processes do not have well defined matrix elements – like data-derived fakes.
- Non-kinematic information is important, too, such as b-tags (help reduce combinatorics)
- Not clear whether integrating over all possibilities or just picking the best one is the most optimal for the purposes we set out for.

# An Example Multi-Component Background with a Small Signal using Multiple MVA's



Fitting background shapes in situ reduces uncertainty.

Uncertainty on shapes of all templates is included.

Plot is a sum over several selected samples, and CDF+D0

Actual combined result uses separated analyses

# The Classical Two-Hypothesis Likelihood Ratio

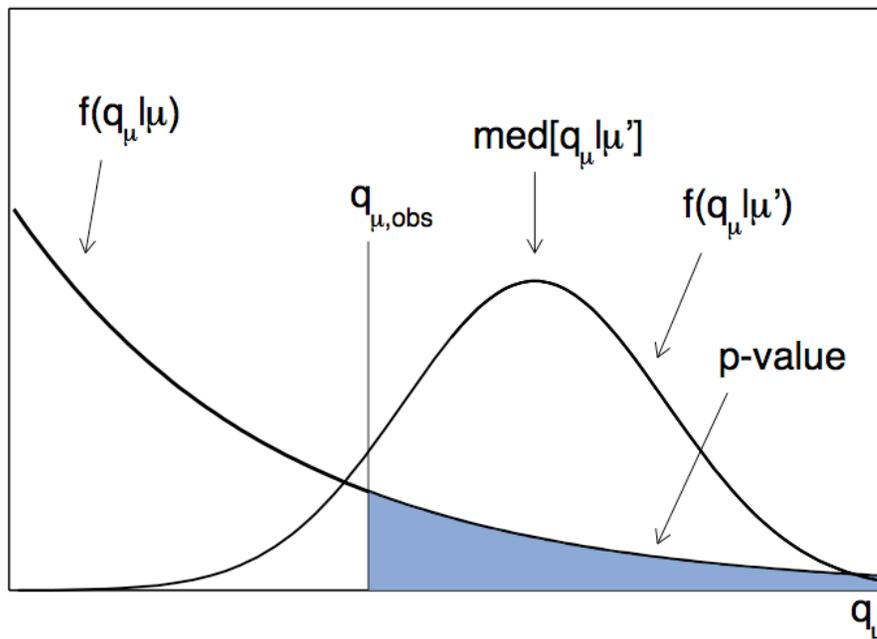
Distinguishing between  $\mu=0$  (zero signal, SM, Null Hypothesis) and  $\mu>0$  (the test hypothesis)

Assumption Warning!  
Signal rates scale with a single parameter  $\mu$

$$q_\mu \equiv 2 \ln \left( \frac{L(\text{data} | \hat{\mu}, \hat{\nu})}{L(\text{data} | \mu, \hat{\nu})} \right)$$

$\hat{\mu}$  is the best-fit value of the signal rate. Can be zero. Your choice to allow it to go negative.

$\mu$  is quadratically dependent on coupling parameters (or worse.).



Larger  $q_0$  is more signal-like

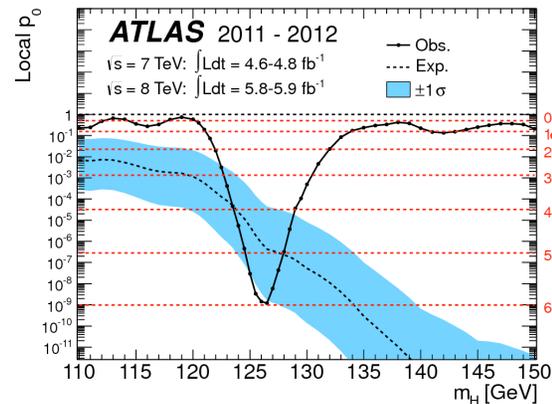
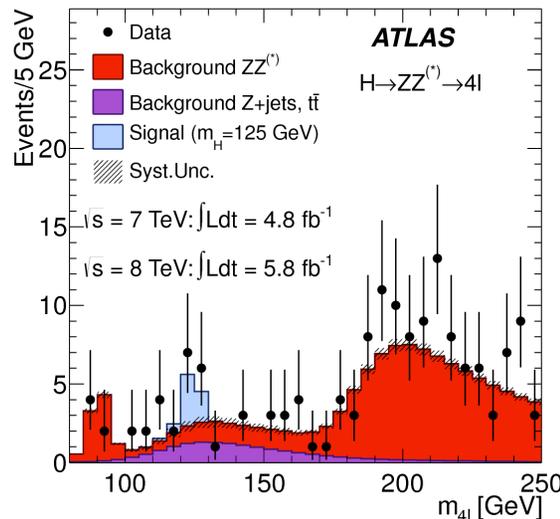
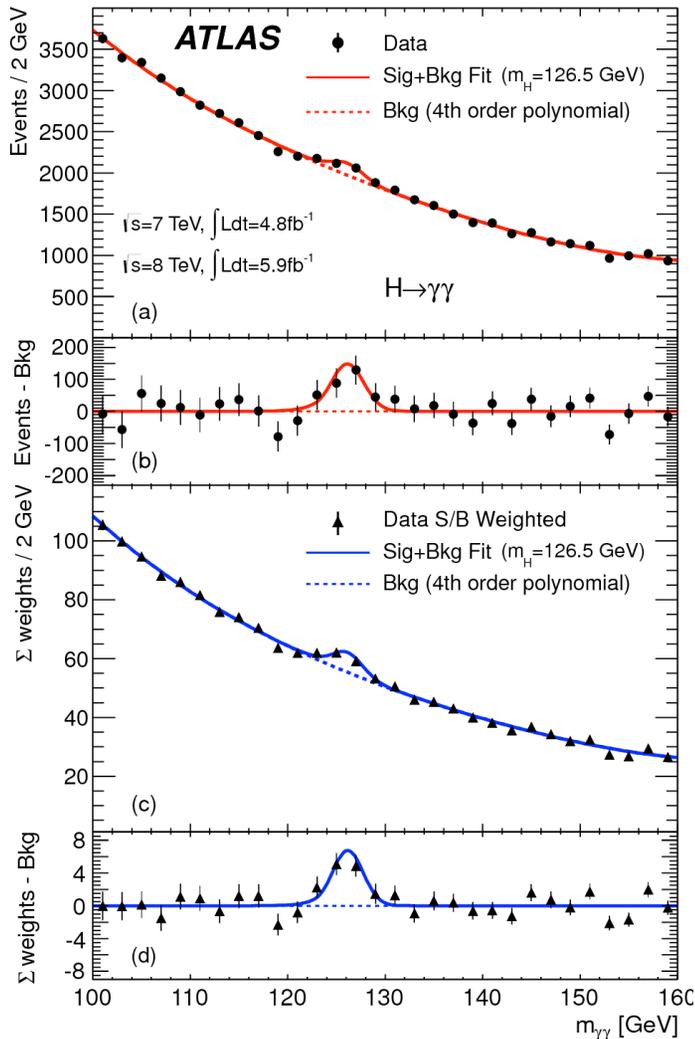
$q_\mu > 0$  always because  $H_1$  is a superset of  $H_0$  and therefore always fits at least as well.

$\hat{\nu}$  is the best-fit set of nuisance parameter values allowing  $\mu$  to also float.

$\hat{\hat{\nu}}$  the same, but with  $\mu$  fixed (zero for  $q_0$ )

ATLAS performance projections, CERN-OPEN-2008-020

# Observation of a New Particle: P-Value for the Null Hypothesis



Combining channels: Joint likelihood  $L$  with shared nuisance parameters.

Popper: you can only falsify models, never prove one right.

p-value:

$$p_0 = p(q_0 \geq q_0^{\text{obs}} \mid \text{no Higgs})$$

Criterion for discovery:

$p_0 < 2.77 \times 10^{-7}$  corresponds to 5 standard deviations

Bayesian Discovery Techniques (like the Bayes Factor) aren't popular in HEP (we like to know our error rates)

# Measuring an Interaction Rate

Frequently using the maximum posterior method, and taking a 68% credibility interval.

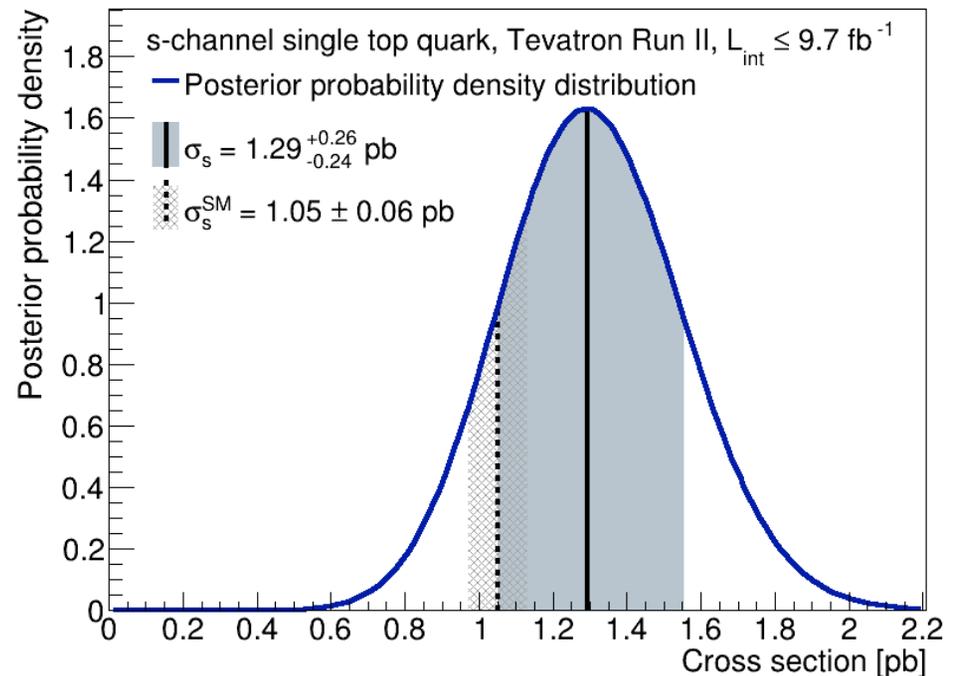
$$L'(\theta) = \int L(\theta, \nu) d\nu$$

Where  $\theta$  are the parameter(s) of interest and  $\nu$  are nuisance parameters

$$\text{Posterior Probability Density} = \frac{L'(\theta)\pi(\theta)}{\int L'(\theta)\pi(\theta)d\theta}$$

Example – a recent measurement combining data from CDF and D0. Some nuisance parameters are shared.

Alternatively, some people just run a maximum likelihood fit and quote uncertainties  $\Delta \ln L = \frac{1}{2}$ . This is not guaranteed to cover, esp.  $L$  is multimodal or just not very Gaussian shaped (even as a function of a nuisance parameter).



# Setting Limits on a Production Rate

Most hypothetical particles proposed by speculative theorists do not exist!

But we really should look for them because some of them might exist.

How do we express our results? Upper limits!

Different approaches, usually with similar results:

Bayesian

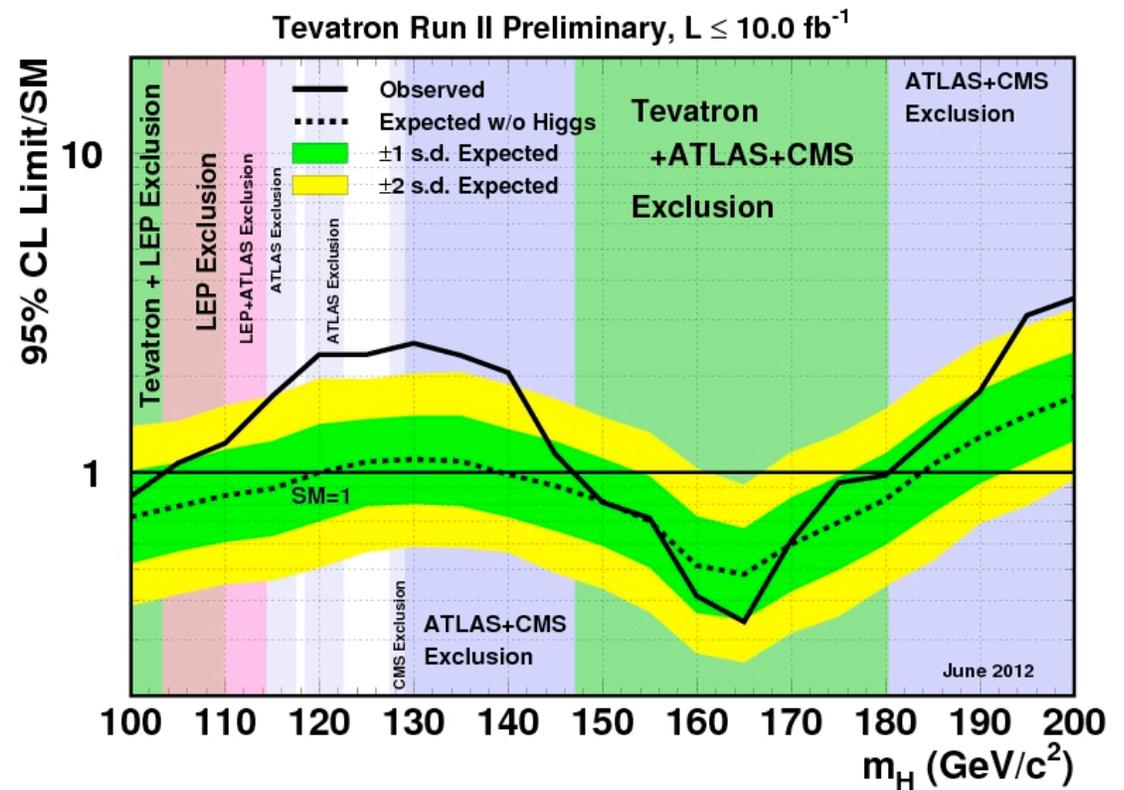
$CL_s$

Frequentist

Coverage: Exclude a true signal no more than 5% of the time at 95% CL.

Credibility

Power

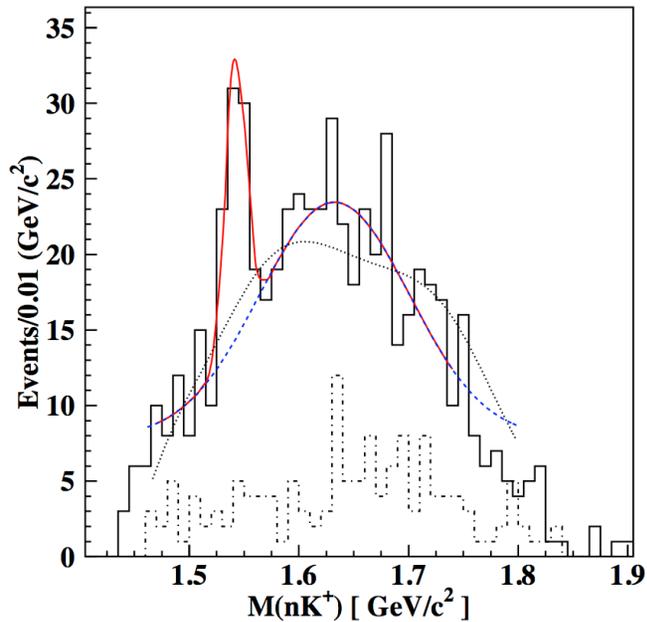


# Blind Analyses

(Sub)conscious bias on the part of analyzers can distort results.  
Even the decision to publish or not can constitute cherry-picking of results.

- Changing data selection requirements after looking at the data is very difficult to justify
- If a surprising result is obtained and it depends on a small number of events, then we like to check it in a statistically independent sample
  - Run the experiment some more, retain the original data selection requirements and analysis tools and see if the result can be reproduced
  - Check with the other collaboration(s) on the other side of the storage ring.

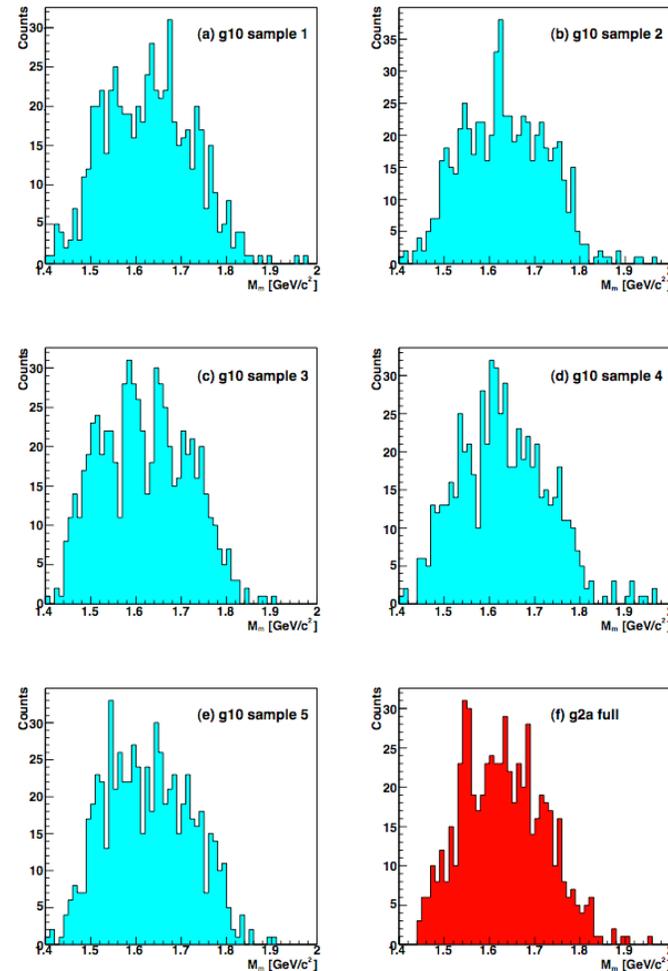
# A Cautionary Tale – The Pentaquark “Discoveries”



CLAS Collab., **Phys.Rev.Lett. 91 (2003) 252001**

Significance =  $5.2 \pm 0.6 \sigma$

Watch out for the  
background function  
parameterization!



Five times the data sample  
CLAS Collab., **Phys.Rev.Lett. 100 (2008) 052001**

n.b. the Bayesian analysis in this paper is flawed –  
see the criticism by R. Cousins, **Phys.Rev.Lett. 101 (2008) 029101**

# Blind Analyses

But sometimes we would like to eliminate possible bias at the outset!

- Hide the data in the signal region from the analyzers
- Allow analyzers to look at control samples (“sidebands”) (“calibration samples”)
- Introduce hidden offsets to measured quantities so analyzers do not know what the measured answer is so they cannot make it more (or less) like a prediction

# Blind Analysis Procedures

Validate analysis as much as possible with simulation and control sample data

Collaboration sign-off on the analysis without looking at signal-region data

Data are “unblinded” (or hidden offsets revealed)

A hard-line approach: Collaboration must approve the unblinded result and submit for publication, even if it contains mistakes that are obvious only when the signal region data are investigated.

“Blind”, not “deaf and dumb”: Allow review of possible mistakes. But then we’re not really blind, are we?

A practical concern: One analysis group’s calibration sample is another’s signal sample!

They can accidentally unblind each other!

Do we need to keep people out of each others’ meetings?

Collaboration by-laws usually prohibit denying access to data or to analysis meetings.

Usually a “good-faith effort”

# Look-Elsewhere

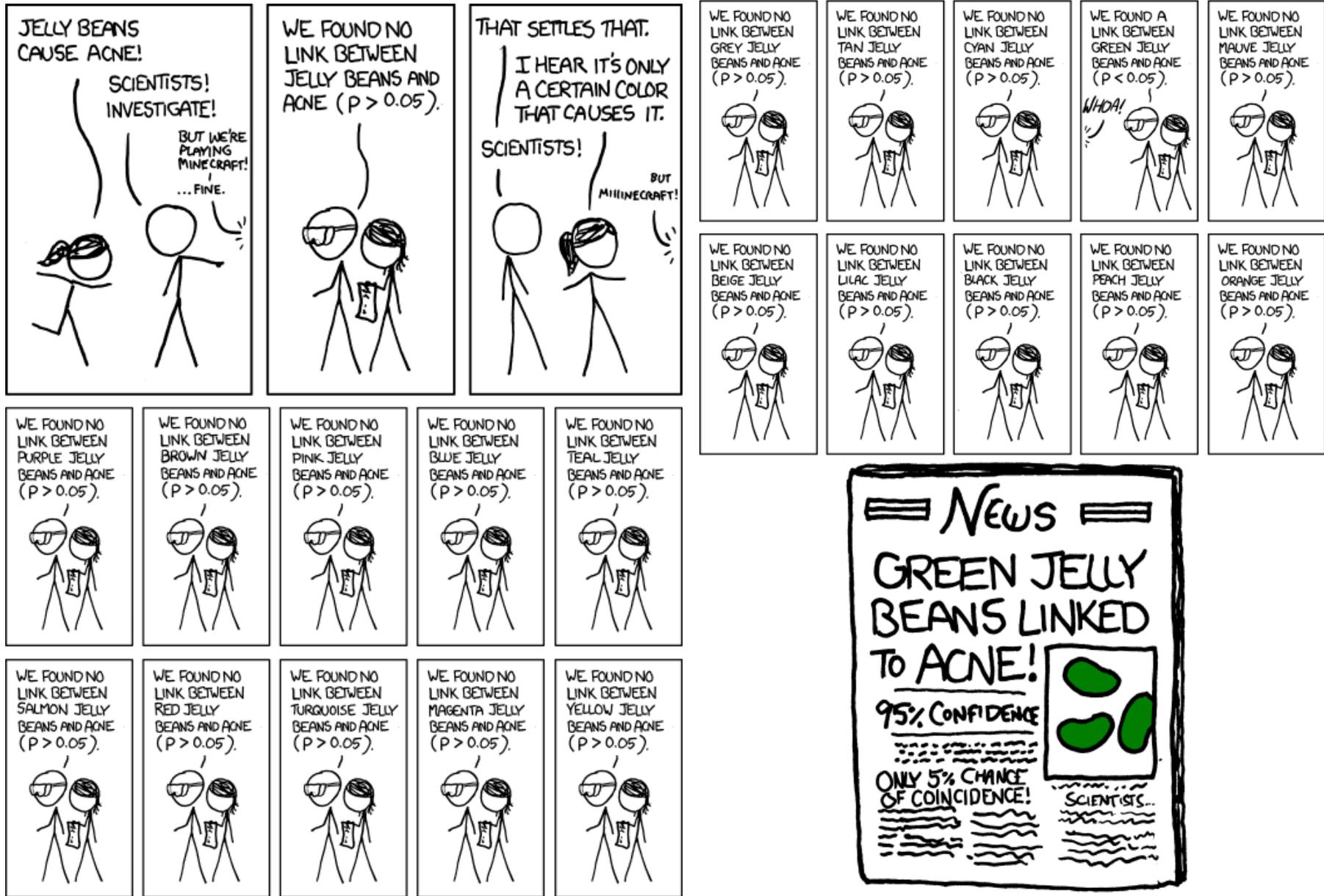
Multiple Independent Tests, each with a specified Type-1 Error Rate (false signal), ought to produce Errors at the specified rate (or they aren't powerful enough, or the error rate can be claimed to be lower).

A single analysis can involve many multiple tests. Classic example: A bump-hunt on a histogram.

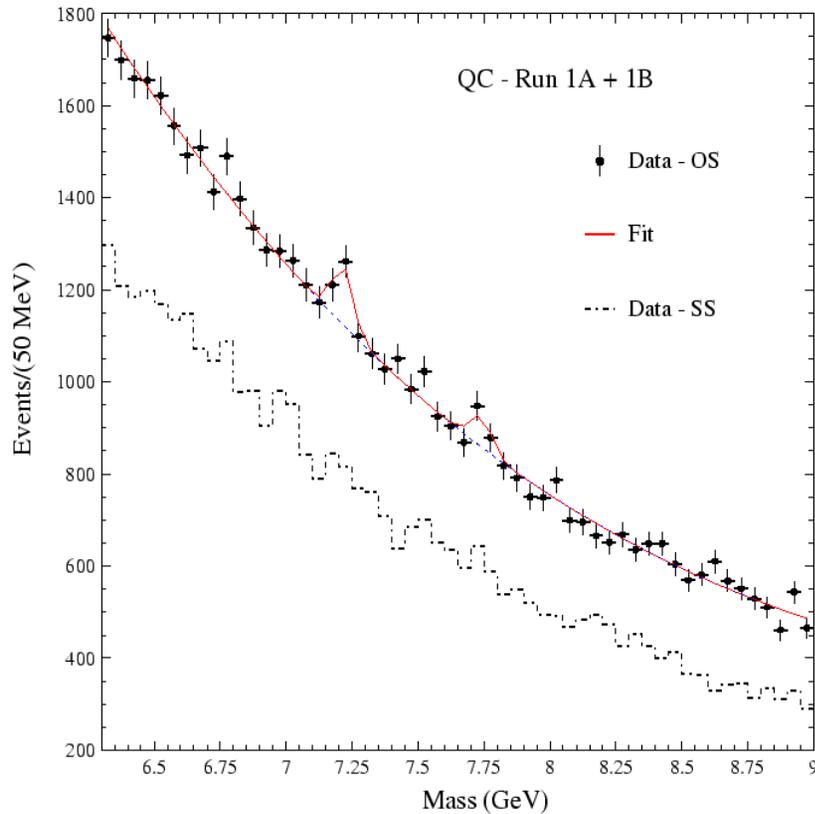
Old-fashioned way to handle it (Bonferroni) – multiply p-value by the number of independent searches:

Histogram width / resolution

This is an approximation to the right way to do it which is to compute the p-value of p-values. What's the chance of observing an excess as significant as the one I saw anywhere in the histogram?

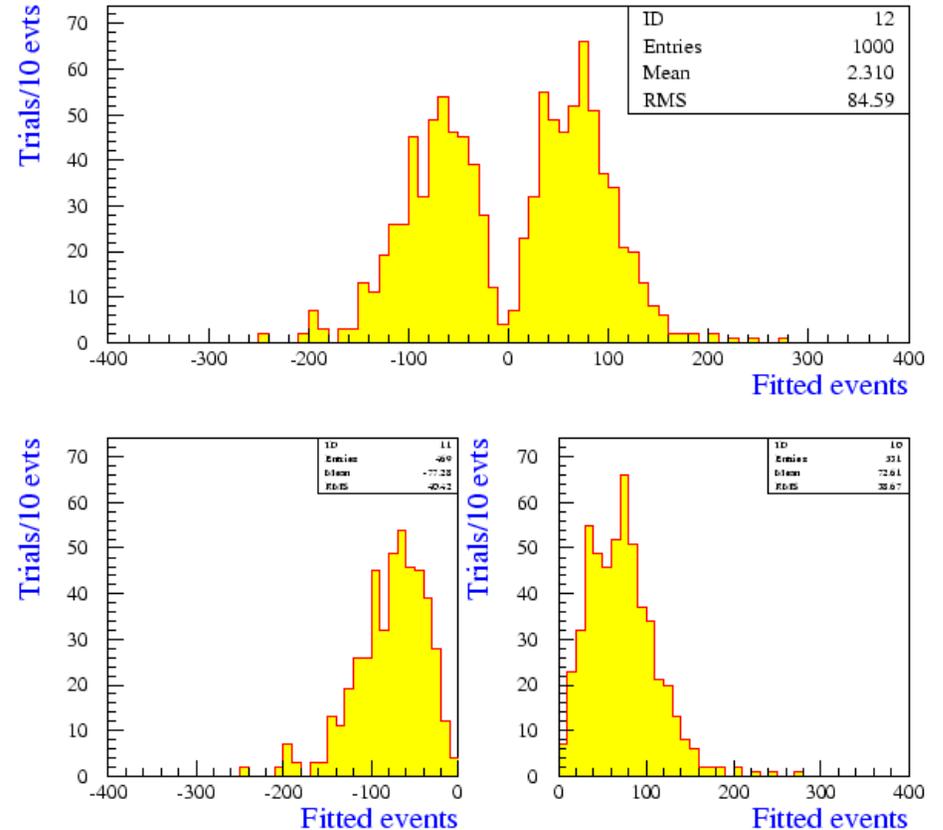


# An internal CDF study that didn't make it to prime time – dimuon mass spectrum with signal fit (not enough PE's)



249.7 ± 60.9 events fit in bigger  
 signal peak (4σ? No!)

Significance Tests on the Dimuon Mass Bump



Null hypothesis pseudoexperiments  
 with largest peak fit values

# Where is “Elsewhere?”

A collider collaboration is typically very large; >1000 Ph.D. students. ATLAS+CMS is another factor of two. (Four LEP collaborations, Two Tevatron collaborations).

Many ongoing analyses for new physics. The chance of seeing a fake bump somewhere is large. What is the LEE?

Do we have to correct our previously published p-values for a larger LEE when we add new analyses to our portfolio?

How about the physicist who goes to the library and hand-picks all the largest excesses? What is LEE then?

“Consensus” at the Banff 2010 Statistics Workshop: LEE should correct only for those models that are tested within a single published analysis. Usually one paper covers one analysis, but review papers summarizing many analyses do not have to put in additional correction factors.

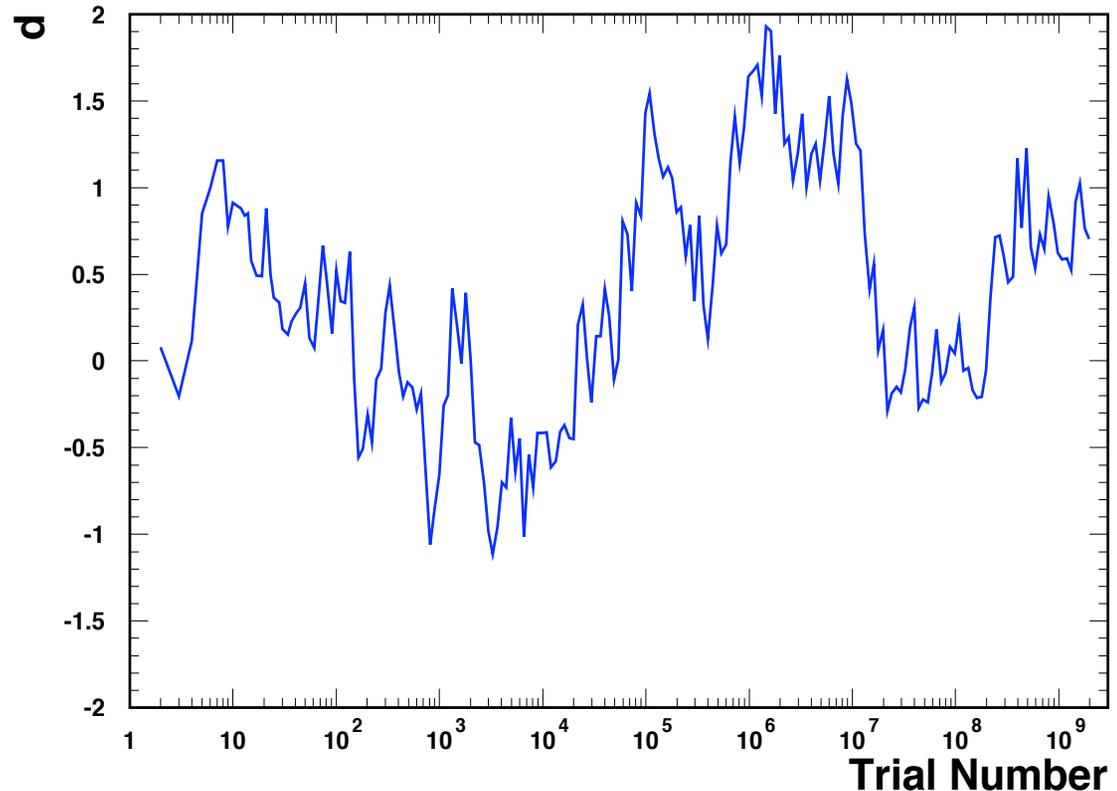
*Caveat lector.*

Running averages converge on correct answer, but the deviations in units of the expected uncertainty have a random walk in the logarithm of the number of trials

$$d_n = \frac{\sum_{k=1}^n r_k / n}{1/\sqrt{n}}$$

The  $r_k$  are IID numbers drawn from a unit Gaussian.

# Look ElseWHEN



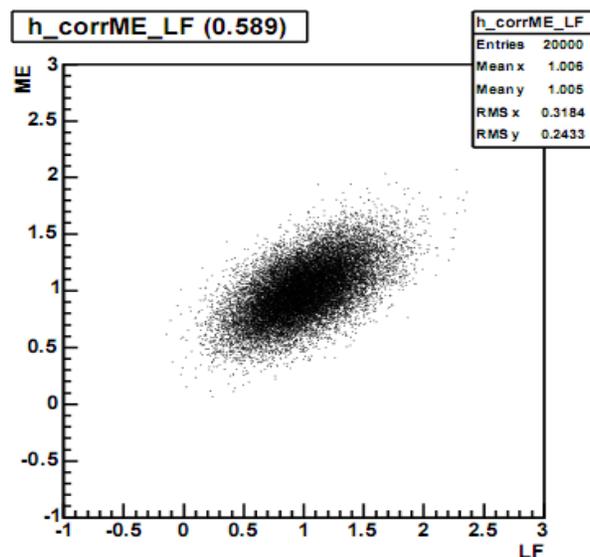
It's possible to cherry-pick a dataset with a maximum deviation. "Sampling to a foregone conclusion"

Stopping Rule: In HEP, we (almost always!) take data until our money is gone. We produce results for the major conferences along the way. Some will coincidentally stop when the fluctuations are biggest. We take the most recent/largest data sample result and ignore (or should!) results performed on smaller data sets. p-values still distributed uniformly from 0 to 1. A recipe for generating "effects that go away"

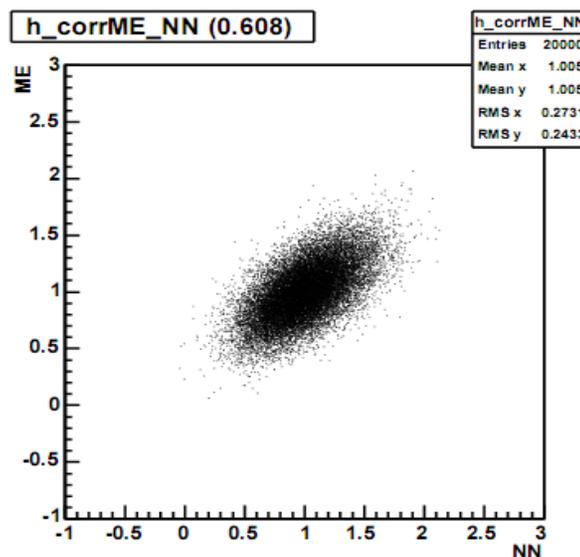
# Extras

# An Example of Running Three Analyses on the Same Events in Monte Carlo Repetitions

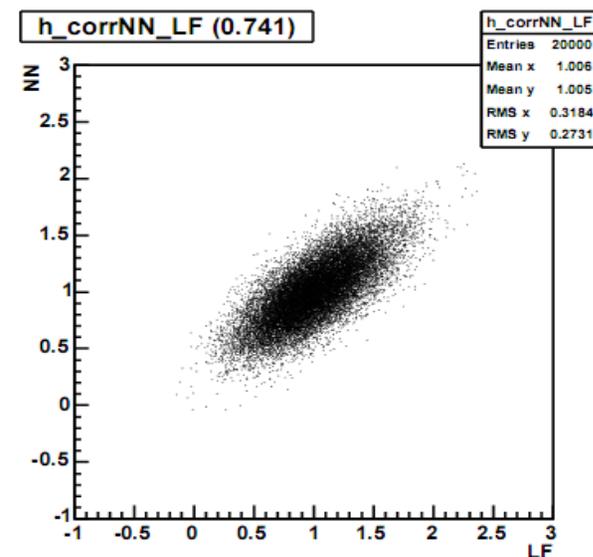
LF-ME 58.9%



ME-NN 60.8%



LF-NN 74.1%

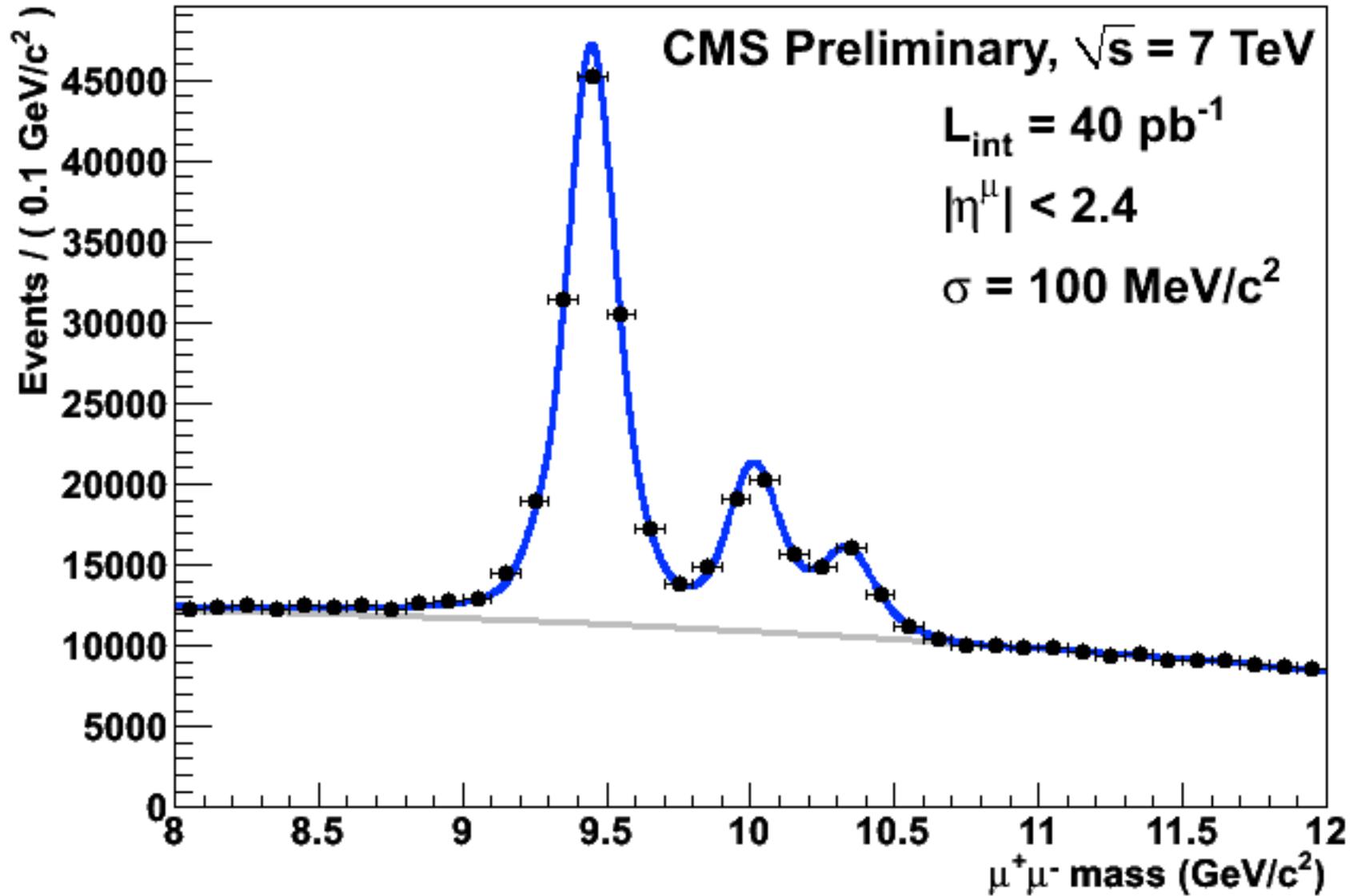


Different questions can be asked: What's the distribution of the maximum difference between the measurements any two teams? What's the quadrature sum of the pairwise differences? Condition on the sum? (Probably not..)

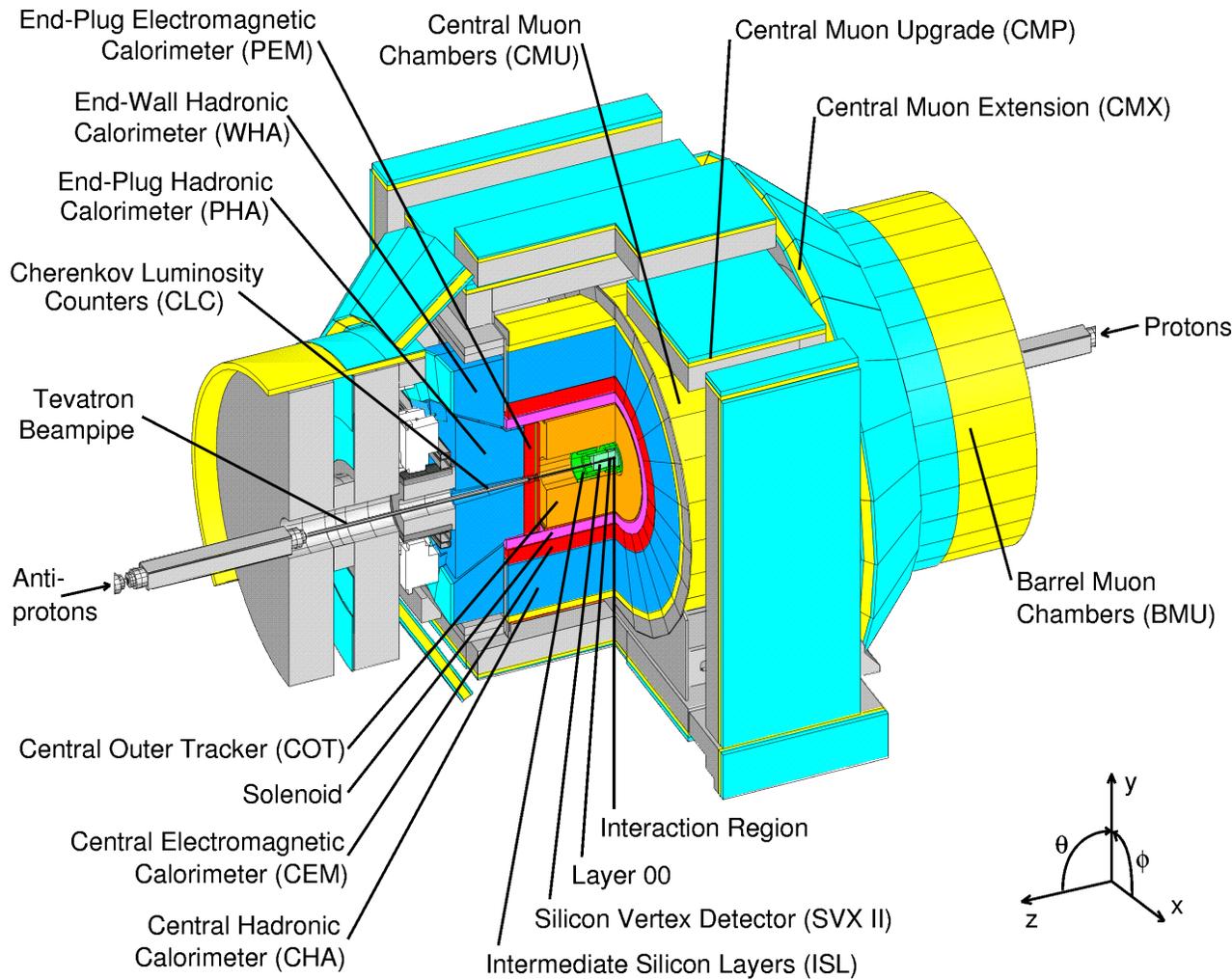
# Several Analyses on the Same Data

- Different groups are interested in the same search/measurement using the same data.
- May have slightly different selection requirements (Jet energies, lepton types, missing  $E_t$ , etc).
- Usually have different choices of MVA or even training strategies for the same MVA
- Always will give different results!
- What to do?
  - Pick one and publish it – criterion: best sensitivity. Median expected limit, median expected discovery sensitivity, median expected measurement uncertainty. How to pick it if the result is 2D? Need a 1D figure of merit.
  - Can check consistency with pseudoexperiments. A p-value using  $\Delta(\text{measurement})$  as a test statistic. What's the chance of running two analyses on the same data and getting a result as discrepant as what we got?
  - Combine MVA's into a super-MVA
    - Keeps everyone happy and involved
    - Usually helps sensitivity
    - Requires coordination and alignment of each event in data and MC
    - Easiest when overlap in data samples is 100%. Otherwise have to break sample up into shared and non-shared subsets and analyze them separately
- What not to do: Pick the one with the “best” observed result. (LEE!)

# States With A Bottom and an Anti-Bottom Quark: the Upsilon System



# The Detector



**Lepton coverage:**

$|\eta| < 1.5$  (muons)

$|\eta| < 2.0$  (electrons)

**b-tagging with**

$|\eta| < \sim 1.4$

**Jets to**

$|\eta| < 2.8$

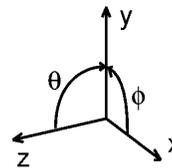
**Higgs analyses**

**restrict to**

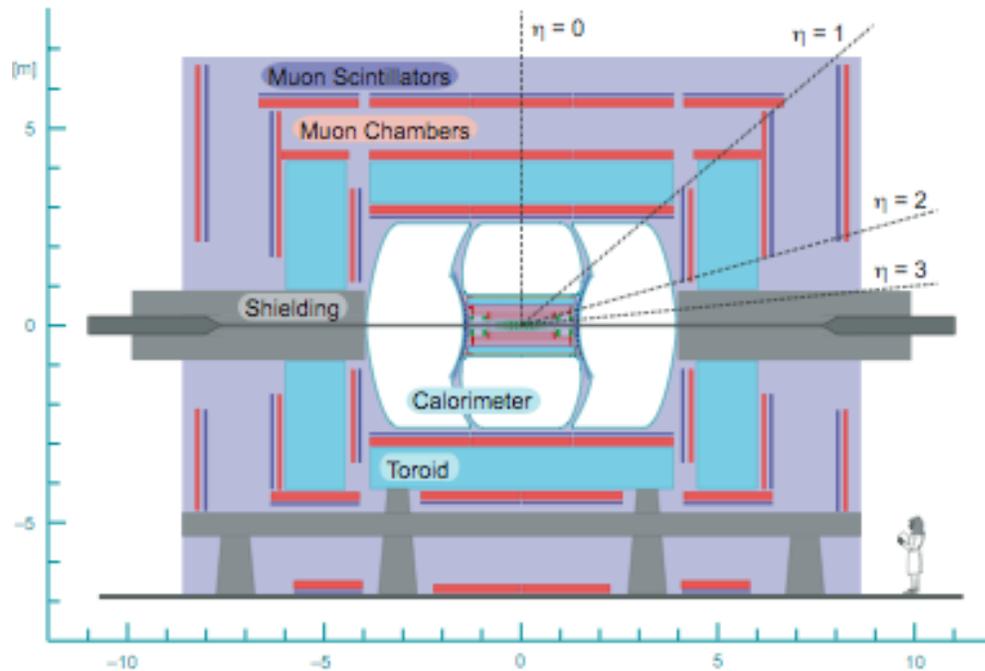
$|\eta| < 2.0$

**Dijet mass**

**resolution:  $\sim 16\%$**



# The **DO** Detector



**Lepton coverage:**

$|\eta| < 2$  (muons)

$|\eta| < 2.6$  (electrons)

Scintillating  
fiber tracker

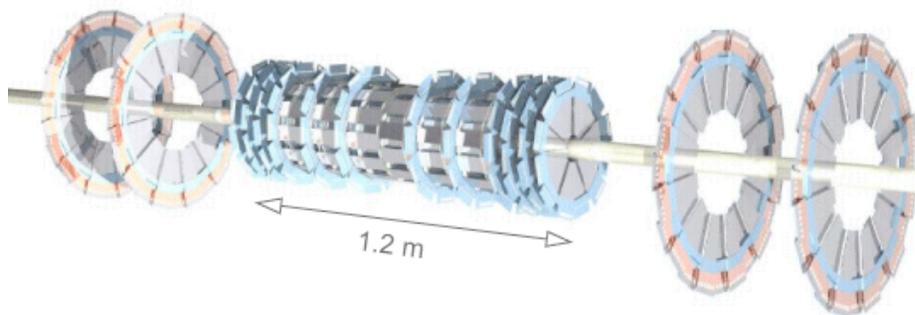
Trigger similar  
to CDF's

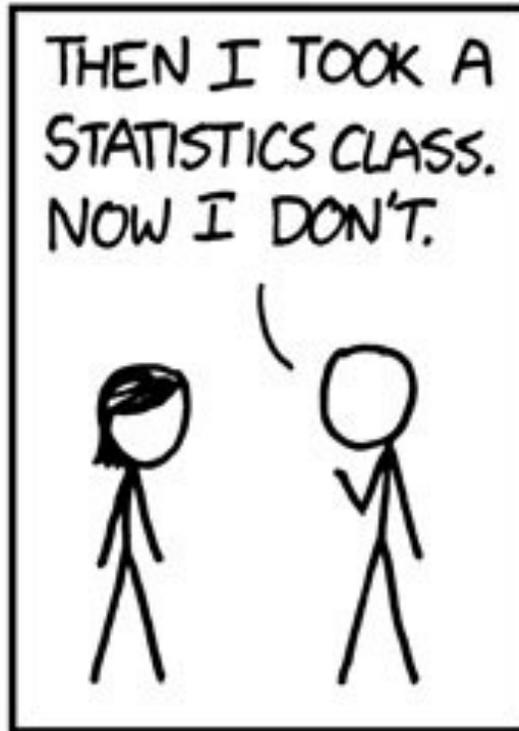
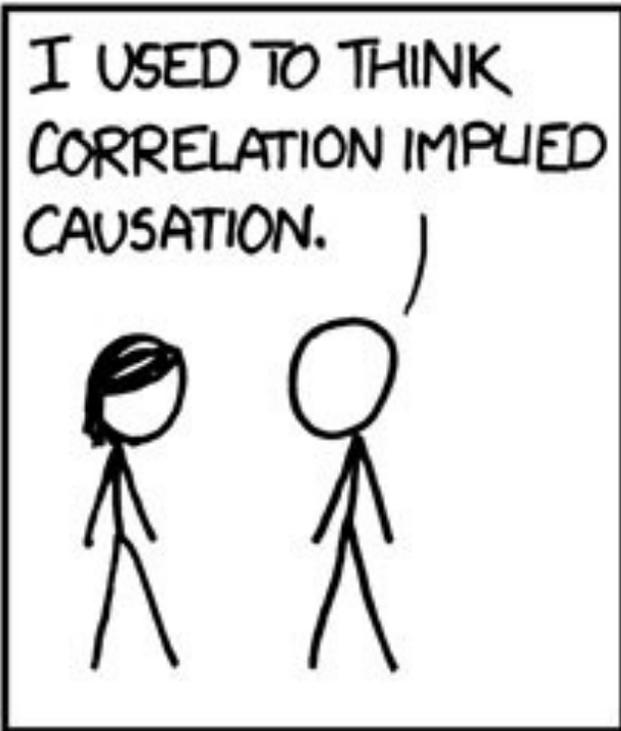
**b-tagging with**

$|\eta| < \sim 2$

**Jets to**

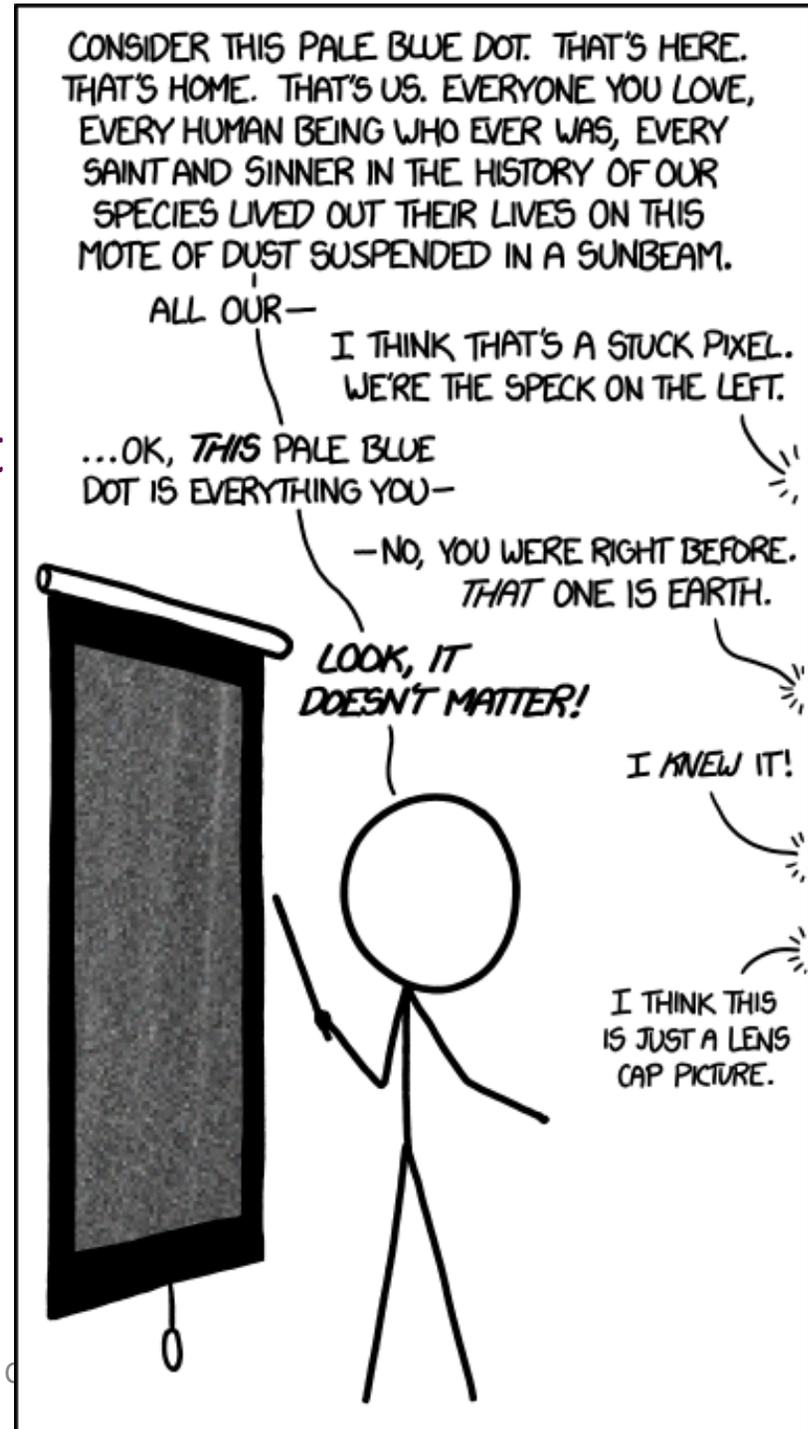
$|\eta| < 3$





XKCD

A good sample of purified signal is important if you don't have many events.



# Combining the Work of Two Teams Analyzing the Same Set of Events

- How do we get the most out of our work?
- Typical NN error function is not something we care about:

$$E = \sum_{events} (meas - desired)^2$$

- But it is easy to back-propagate for efficient training
- Instead we want
  - Discovery
  - failing that, exclusion
- This figure of merit works better:

$$F = \sum_{bins} s^2 / b$$

- But how do you train to optimize that?

# Neuro-Evolution to the Rescue!

Kenneth O. Stanley and Risto Miikkulainen (2002).

"Evolving Neural Networks Through Augmenting Topologies".

Evolutionary Computation 10 (2): 99–127;

[http://en.wikipedia.org/wiki/NeuroEvolution\\_of\\_Augmented\\_Topologies](http://en.wikipedia.org/wiki/NeuroEvolution_of_Augmented_Topologies)

- Figure of merit difficult to calculate
- Test one configuration, set of weights against others, pick features from the best performers
- Handles to optimize:
  - Network topology
  - Network weights
  - Output binning
- Inputs MEBDT, NN outputs for each event.
- Sensitivity improvement -- 9% in expected limit

# Matrix Element Basics

Predictions given by QM matrix element and phase space.

Many processes (signal and background) give the same observable quantities in the detector -- cannot assign an event to be signal or background (if we could, we would!)

Instead, ask what the ratio of chances of getting an event from signal or background processes. Need to incorporate experimental resolution effects.

# Imperfect Reconstruction

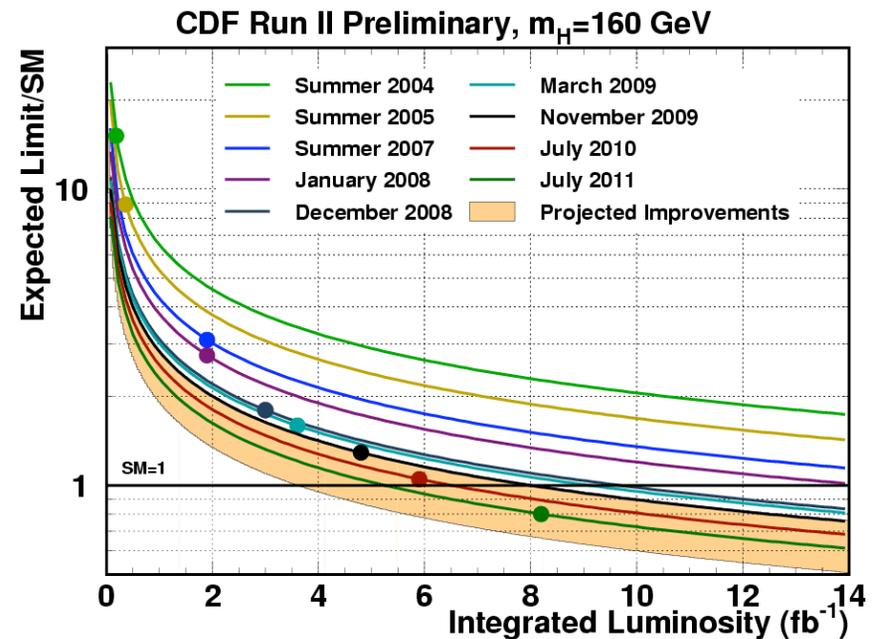
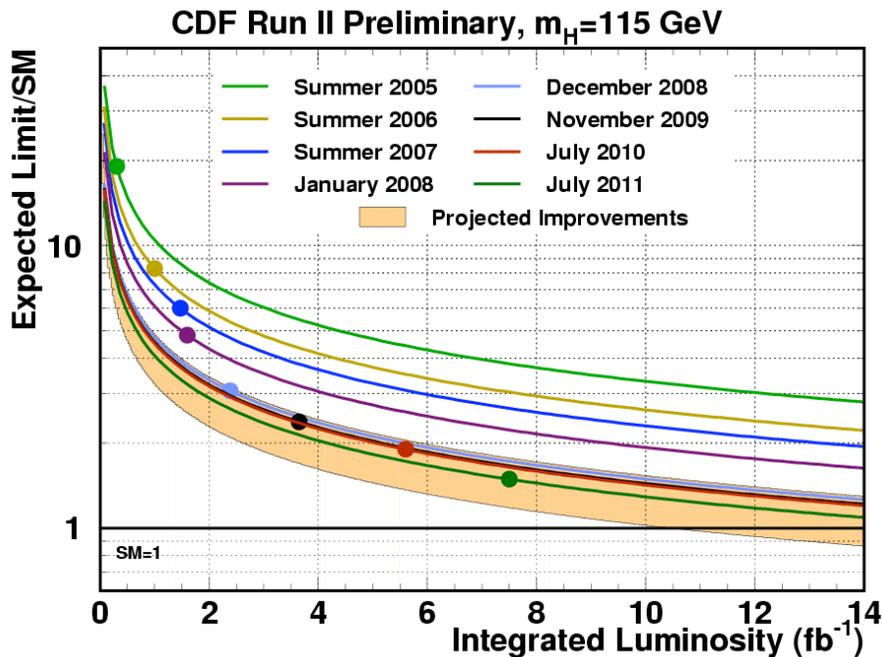
- Missing neutrinos! Missing  $E_T$  resolution not perfect.
- Jet energies not perfectly measured. Directions are pretty good, and leptons are measured well.
- What parton 4-vectors could have given us the measured events?

$$P(x) = \frac{1}{\sigma} \int 2\pi^4 \overset{\text{ME}}{|M|^2} \overset{\text{PDFs}}{\frac{f(y_1)}{|E_{q_1}|} \frac{f(y_2)}{|E_{q_2}|}} \overset{\text{transfer function}}{W(y, x)} d\Phi_4 dE_{q_1} dE_{q_2}$$

$y$ =parton (or neutrino momenta),  $x$ =measured jet quantities.

Do this for each physics process -- form a likelihood ratio from them

# Sensitivity Improves Over Time



Naive Expectation: Expected rate limit scales as  $1/\sqrt{L_{\text{int}}}$ . Assumes  $b \gg 1$  event.

Exceptions to the rule:

- $b < 1$ : Expected limit scales as  $1/L_{\text{int}}$ . As  $L_{\text{int}}$  grows,  $b$  grows, and the dependence departs from  $1/L_{\text{int}}$  anyway. Not an issue for our searches (trilepton ones have low  $b$  though)
- Systematics could hit a “brick wall”. Background and signal efficiency systematics are constrained by data, so we expect these to scale as  $1/\sqrt{L_{\text{int}}}$ .
- **Analyses improve!** New taggers, more acceptance, trigger improvements, smarter MVA's.
- Theorists give us new cross sections and b.r.'s (we scale these out so it's apples to apples).

# Why 5 Sigma for Discovery?

From what I hear: It was proposed in the 1970's when the technology of the day was bubble chambers.

Meant to account for the Look Elsewhere Effect. A physicist estimated how many histograms would be looked at, and wanted to keep the error rate low.

Also too many  $2\sigma$  and  $3\sigma$  effects “go away” when more data are collected.

Some historical recollections:

[http://www.huffingtonpost.com/victor-stenger/higgs-and-significance\\_b\\_1649808.html](http://www.huffingtonpost.com/victor-stenger/higgs-and-significance_b_1649808.html)

Not all estimations of systematic uncertainties are perfect, and extrapolations from typical  $1\sigma$  variations performed by analyzers out to  $5\sigma$  leave room for doubt.

Some effects go away when additional uncertainties are considered. Example – CDF Run I High- $E_T$  jets. Not quark compositeness, but the effect could be folded into the PDFs.

If a signal is truly present, and data keep coming in, the expected significance quickly grows ( $s/\sqrt{b}$  grows as  $\sqrt{\text{integrated luminosity}}$ ).

# A Useful Tip about Limits

It takes almost exactly 3 expected signal events to exclude a model.

If you have zero events observed, zero expected background, then the limit will be 3 signal events.

$$p_{Poiiss}(n = 0, r) = \frac{r^0 e^{-r}}{0!} = e^{-r}$$

If  $p=0.05$ , then  $r=-\ln(0.05)=2.99573$

You can discover with just one event and very low background, however!

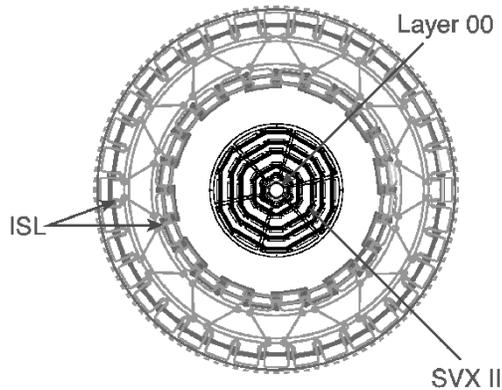
Example: The  $\Omega^-$  discovery with a single bubble-chamber picture.

Cut and count analysis optimization usually cannot be done simultaneously for limits and discovery.

But MVA's take advantage of all categories of s/b and remain optimal in both cases; but you have to use the entire MVA distribution

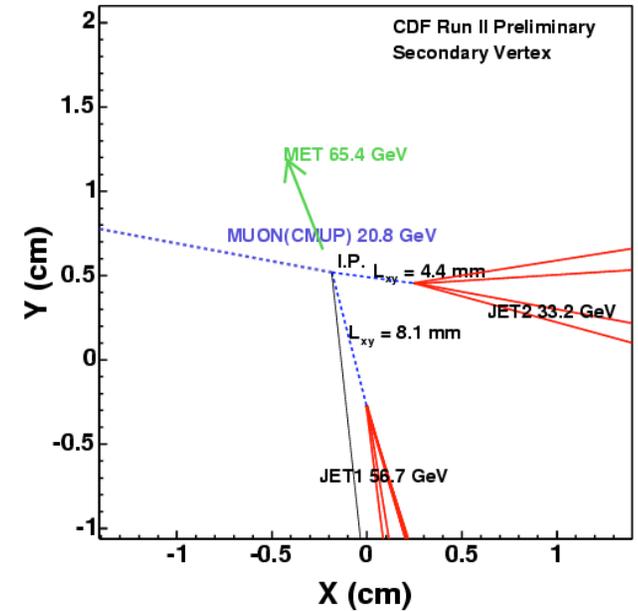
# B-Tagging

L00 single-sided silicon +  
5-layer double-sided silicon+  
2-layer ISL

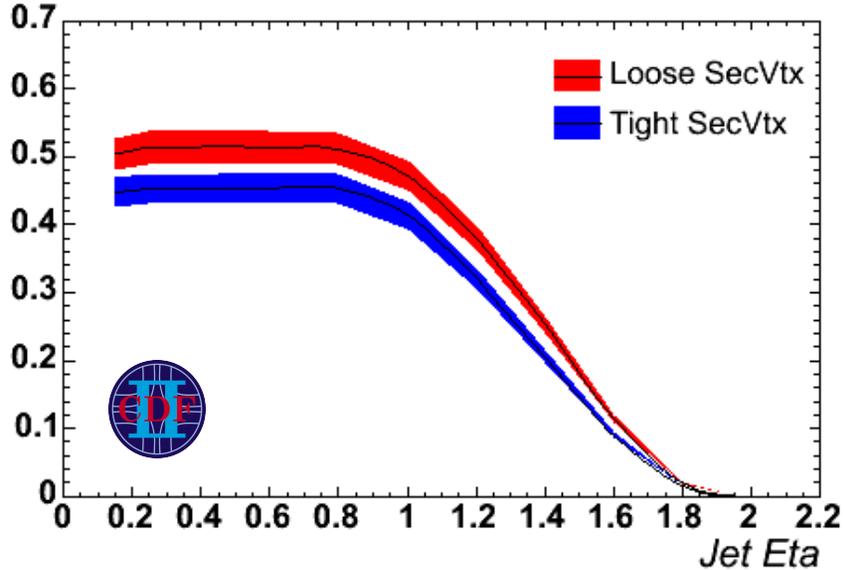


Impact  
parameter  
resolution  
for high- $p_T$   
tracks  $\sim 18\mu\text{m}$

B-tagging relies on  
displaced vertex  
reconstruction:  
high mass, long lifetime



SecVtx Tag Efficiency for Top b-Jets



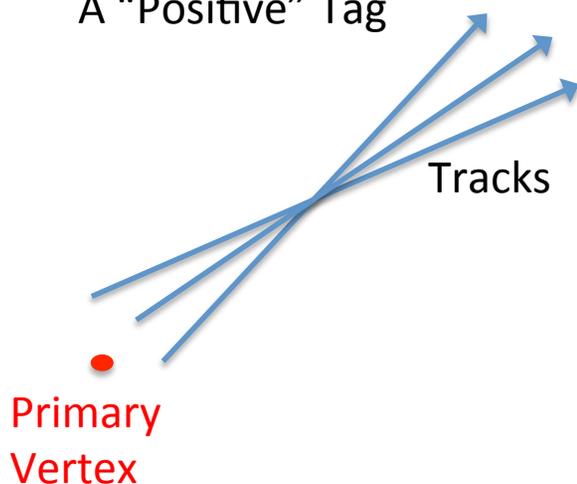
Mistag rates  
typically  
 $\sim 1\%$  for  
light-flavor jets

Example  
candidate  
event (lvbb)

$D0$  B-tagging per-jet  
efficiency = 50-70% (of taggable  
jets) for 1-5% Mistag rate

# B-Tagging Calibration

A "Positive" Tag

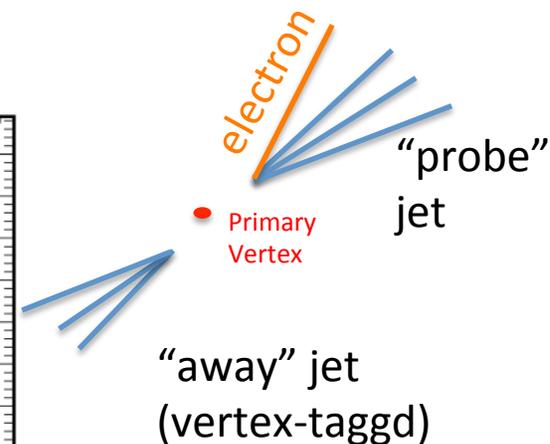
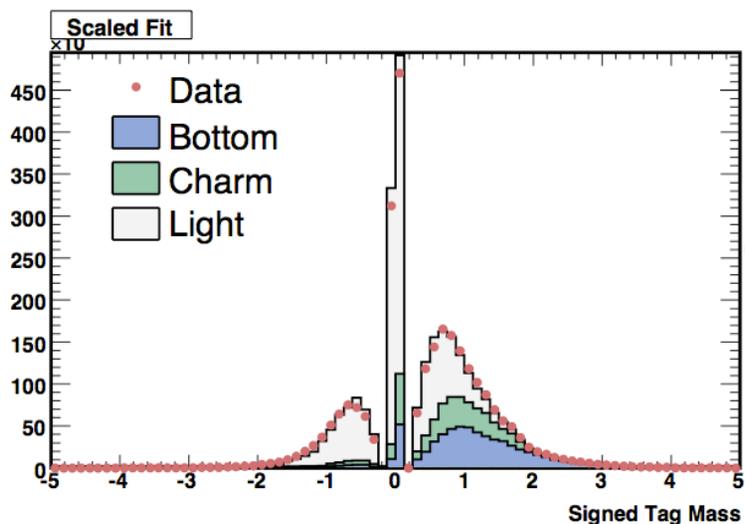
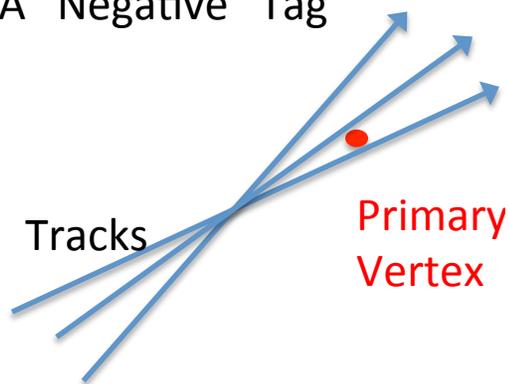


Fake tags calibrated with data – resolution dominated negative tags.

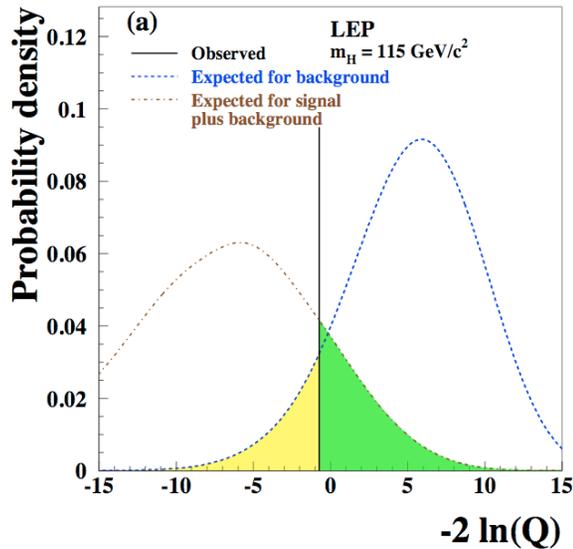
Watch out for asymmetry: scatters in detector material and long-lived strange particles cause more positive mistags than negative ones

Positive b-tags  
Calibrated in data with a sample of dijets –  
"away" jet is b-tagged, "probe" jet has a high- $P_{T,rel}$  electron.  
Tag rates compared in data and MC

A "Negative" Tag



# CL<sub>s</sub> Limits -- extension of the p-value argument



(apologies for the notation)

p-values:

$$CL_b = P(-2\ln Q \geq -2\ln Q_{obs} \mid b \text{ only})$$

Green area =  $CL_{s+b} = P(-2\ln Q \geq -2\ln Q_{obs} \mid s+b)$

Yellow area = "1-CL<sub>b</sub>" =  $P(-2\ln Q \leq -2\ln Q_{obs} \mid b \text{ only})$

$$CL_s \equiv CL_{s+b} / CL_b \geq CL_{s+b}$$

Exclude at 95% CL if  $CL_s < 0.05$

Scale  $r$  until  $CL_s = 0.05$  to get  $r_{lim}$  ←

This step can take significant CPU

- Advantages:

- Exclusion and Discovery p-values are consistent.  
Example -- a  $2\sigma$  upward fluctuation of the data with respect to the background prediction appears both in the limit and the p-value as such
- Does not exclude where there is no sensitivity (big enough search region with small enough resolution and you get a 5% dusting of random exclusions with  $CL_{s+b}$ )

# A Useful Tip about Limits

It takes almost exactly 3 expected signal events to exclude a model.

If you have zero events observed, zero expected background, then the limit will be 3 signal events.

$$p_{Poiiss}(n = 0, r) = \frac{r^0 e^{-r}}{0!} = e^{-r}$$

If  $p=0.05$ , then  $r=-\ln(0.05)=2.99573$

You can discover with just one event and very low background, however!

Example: The  $\Omega^-$  discovery with a single bubble-chamber picture.

Cut and count analysis optimization usually cannot be done simultaneously for limits and discovery.

But MVA's take advantage of all categories of s/b and remain optimal in both cases; but you have to use the entire MVA distribution

# Rule of Three

---

From Wikipedia, the free encyclopedia

**Rule of three** may refer to:

- [Rule of three \(aviation\)](#), a rule of descent in aviation
- [Rule of three \(C++ programming\)](#), a rule of thumb about class method definitions
- [Rule of three \(computer programming\)](#), a rule of thumb about code refactoring
- [Rule of three \(economics\)](#), a rule of thumb about major competitors in a free market
- [Rule of three \(mathematics\)](#), a computation method in mathematics
- [Rule of three \(medicine\)](#), for calculating a confidence limit when no events have been observed
- [Rule of Three \(Wicca\)](#), a tenet of Wicca
- [Rule of three \(writing\)](#), a principle of writing
- *Rule of Three*, a series of one-act plays by [Agatha Christie](#)



## See also

---

- [Rule of thirds](#), a compositional rule of thumb in photography
- [Rule of thirds \(diving\)](#), a rule of thumb for scuba divers

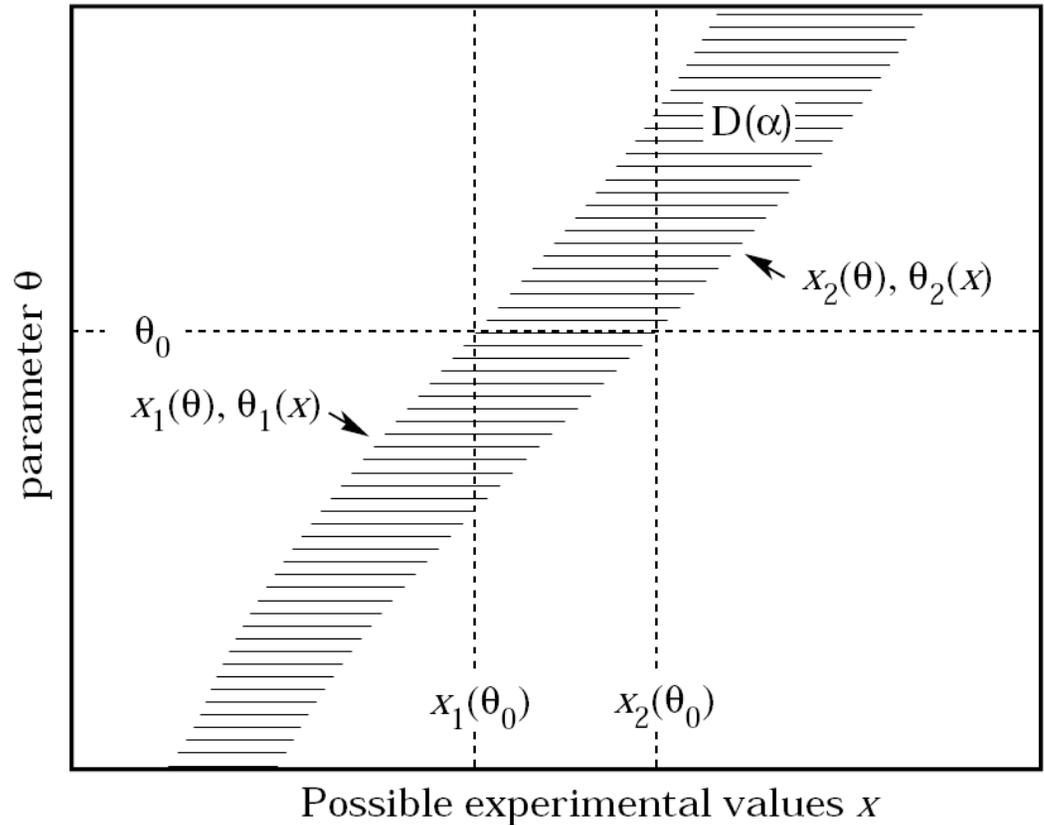
# The “Neyman Construction” of Frequentist Confidence Intervals

Essentially a  
“calibration curve”

- Pick an observable  $x$  somehow related to the parameter  $\theta$  you’d like to measure
- Figure out what distribution of observed  $x$  would be for each value of  $\theta$  possible.
- Draw bands containing 68% (or 95% or whatever) of the outcomes
- Invert the relationship using the prescription on this page.

**Proper Coverage is Guaranteed!**

3/5/14



A pathology: can get an empty interval. But the error rate has to be the specified one.

Imagine publishing that all branching ratios between 0 and 1 are excluded at 95% CL.

# Some Properties of Frequentist Confidence Intervals

- Really just one: *coverage*. If the experiment is repeated many times, the intervals obtained will include the true value at the specified rate (say, 68% or 95%).

Conversely, the rest of them ( $1-\alpha$ ) of them, must not contain the true value.

- But the interval obtained on a particular experiment may obviously be in the unlucky fraction. Intervals may lack credibility but still cover.

Example: 68% of the intervals are from  $-\infty$  to  $+\infty$ , and 32% of them are empty. Coverage is good, but power is terrible.

FC solves some of these problems, but not all.

Can get a 68% CL interval that spans the entire domain of  $\theta$ .

Imagine publishing that a branching ratio is between 0 and 1 at 68% CL.

Still possible to exclude models to which there is no sensitivity.

FC assumes model parameter space is complete -- one of the models in there is the truth. If you find it, you can rule out others even if we cannot test them directly.

## A Special Case of Frequentist Confidence Intervals: Feldman-Cousins

Each horizontal band contains 68% of the expected outcomes (for 68% CL intervals)

But Neyman doesn't prescribe which 68% of the outcomes you need to take!

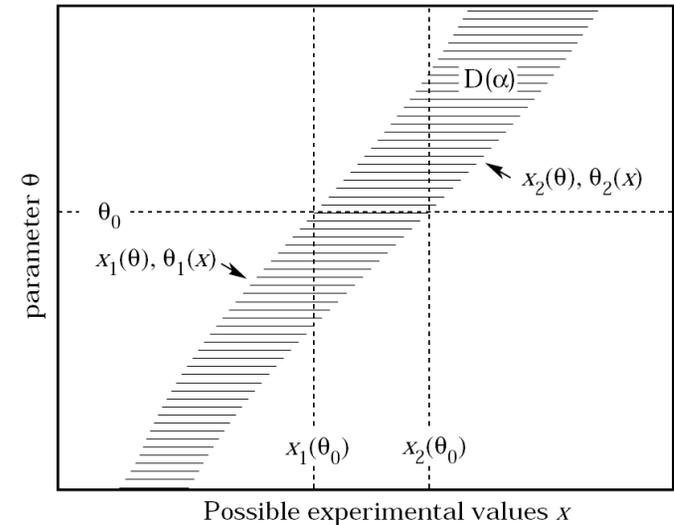
Take lowest  $x$  values: get lower limits.  
Take highest  $x$  values: get upper limits.

Cousins and Feldman: Sort outcomes by the likelihood ratio.

$$R = L(x|\theta)/L(x|\theta_{\text{best}})$$

$R=1$  for all  $x$  for some  $\theta$ .

Picks 1-sided or 2-sided intervals --  
no flip-flopping between limits and 2-sided intervals.



G. Feldman and R. Cousins,  
“A Unified approach to the  
classical statistical  
analysis of small signals”  
Phys.Rev.D57:3873-3889,1998.  
arXiv:physics/9711021

No empty intervals!