What does a (High Energy) Physicist mean when they say "Monte Carlo"

Jim Linnemann
Michigan State University

Banff Workshop:
Statistical Inference Problems in HEP & Astronomy
July 16, 2006

Monte Carlo =

Any calculation using random numbers

at least we always get *some* answer

Primary: physics simulation

estimate background (known physics) contributions

which could look like new physics

estimate signal (new physics) efficiency

fraction visible in detector

Secondary: statistics simulations

"Toy MC"

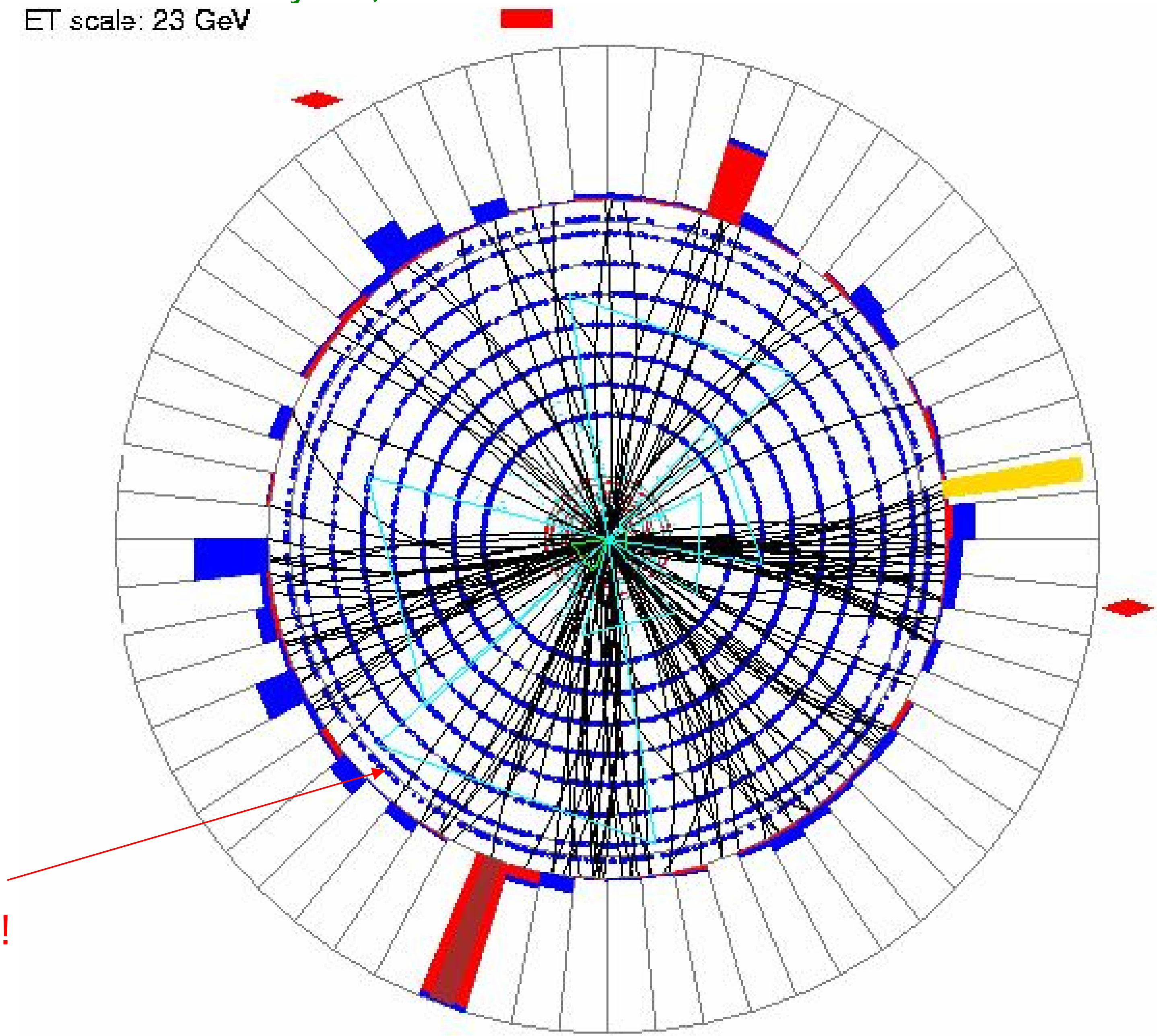
- Calculate a significance
- Calculate a property,
  - e.g. some integral with no closed form
  - Uncertainty of a derived quantity
  - Upper limit
- Evaluate performance of a technique
  - coverage; expected limit

Nancy seems to be have calculated the odds well

Physics Simulation Parts

1. Event Generator
   - Produce events from specific physics processes
     - preferably un-weighted sampling from theoretical distributions
   - Event: set of elementary particles produced
     - could be hundreds in one event
2. Detector Simulation
   - Ideally, write simulated digitized data in real detector format
   - Physical interactions of each particle with detector
   - Detector model:
     - geometry, materials of sensitive regions or “inert” parts
     - production of secondary particles (up to millions)
   - Model detection behavior of sensitive regions of detector
     - efficiency: usually measured with data
   - Model response of electronics to sensitive region signals
   - Model detector calibration
     - position, energy measurement uncertainties, temporal variation
   - Model confounding effects not due to the physics event
     - other physics interactions happening simultaneously
     - cosmic ray interactions
   - That is: efficiency and measurement uncertainties (stat, systematic)

A real event July 15, 2006



Lots of hits to simulate!

Poisson: why we insist

Each event is independent of the next

Quantum Mechanics: satisfy the assumptions for Poisson process (rare, independent)

However, event selection sometimes can misinterpret two independent events with one single more-complex event

Physics Simulation: (Event) Generators

1. Production of fundamental particles:

   leptons, quarks, gluons, …

   approximately distribute according to

   approximate theoretical distributions from "Feynman diagrams"

   complex angular and energy correlations

2. Hadronization

   Coupled decay of the unobservable fundamental particles

   into ordinary elementary particles ("jets" of "stable" hadrons)

   Adjustable parameters have been "tuned" empirically
   to match previous data

There Are Competing Models

Different ways to approximate what's known

Imperfect: MC generators miss some Quantum Correlations

Different input Parton Distribution Functions (PDF !)

fits to data about longitudinal momentum of quarks, gluons in protons

Still, they are reasonably accurate (~ 10%)

distributions span many orders of magnitude

Differences among generators are part of our systematic errors

Feynman Diagrams

Here: production of a gluon + a W Boson from a quark-quark collision.

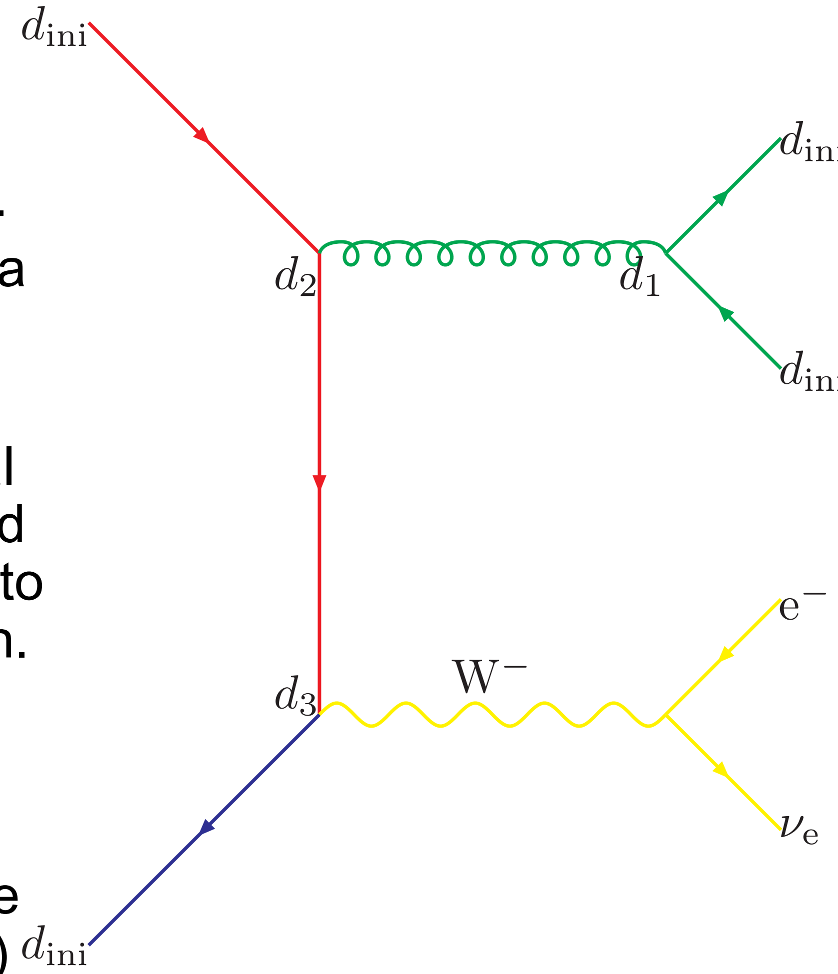
The gluon decays into 2 quarks, which will turn into jets of elementary particles, largely pions $\pi^+ \pi^- \pi^0$ (hadronization).

The W Boson decays into an electron and a (probably invisible) neutrino

The simulation picks momenta for the initial quarks from a PDF, and momenta and angles of the gluon and W according to a "matrix element" physics distribution. Similarly, physics governs the final decays.

At each step, various conservation laws are enforced (e.g. no change in Σ charge)

part of W+2j



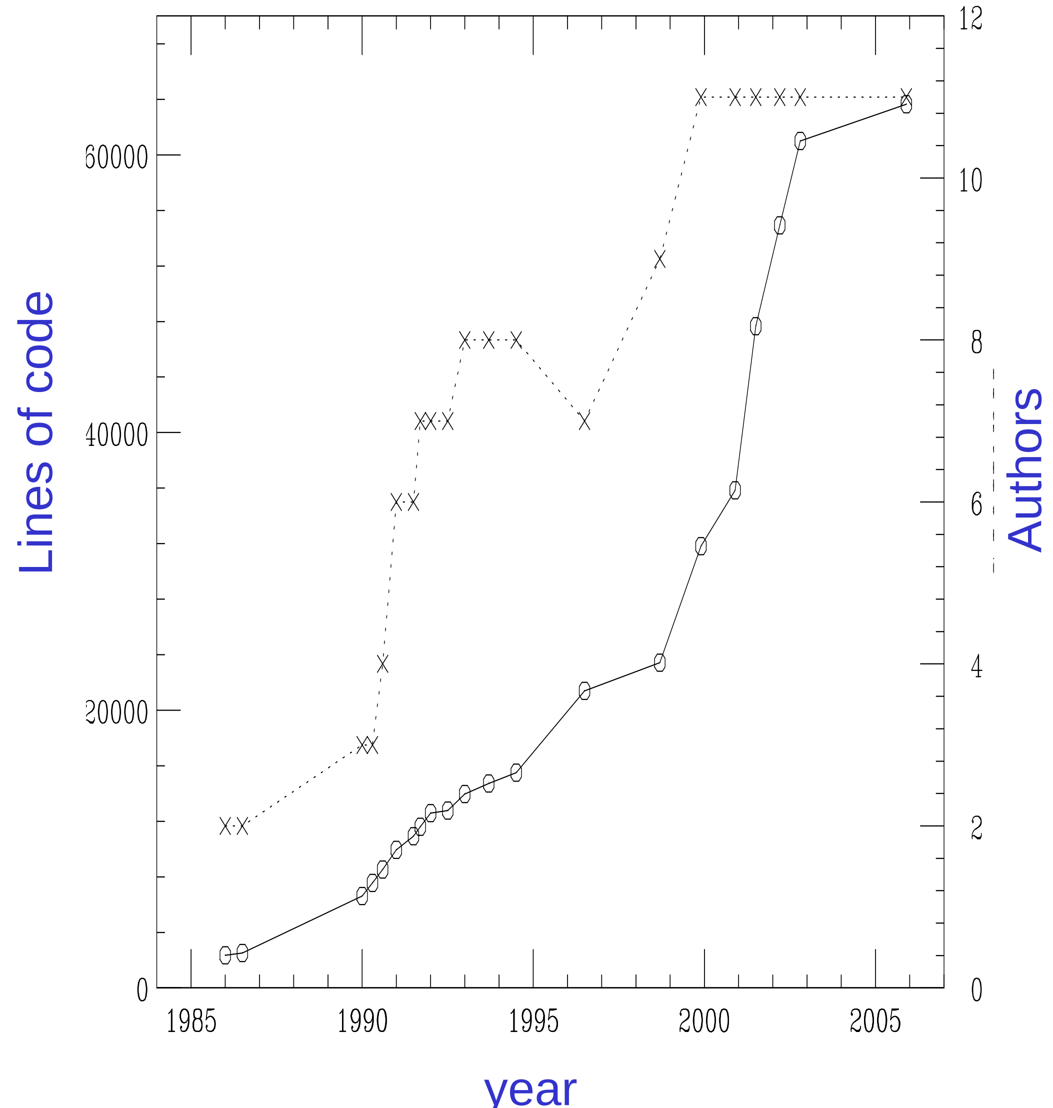
An Event Generator Takes some Effort

~ 100K lines code

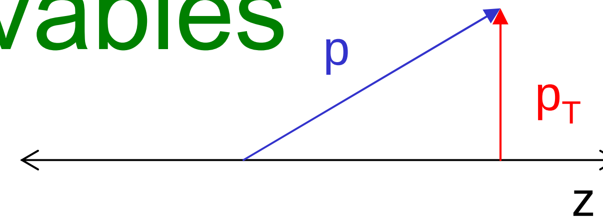
Fortran 77 or C++

~ 10 authors (theorists)

years to develop



Kinematics: particle observables



$\mathbf{p} = \{p_x, p_y, p_z\}$ or $\{p_T, \theta, \phi\}$ or $p = |\mathbf{p}|$

p_T momentum $\perp$ to beam $= \sqrt{(p_x^2 + p_y^2)}$

(high p_T is a rare violent collision)

$\tan\theta = p_z / p_T$ $\quad \eta = \mathrm{Ln}(\tan\theta/2)$ ("rapidity")

Jets: combinations of clustered particles

attempt to reconstructed quark or gluon $\quad H_T = \Sigma\, |p_T|$

2 or more particles (or jets):

p_T important since $\Sigma\, \mathbf{p}_T = 0$ $\quad$ imbalance is Missing E_T ($\not{E}_T$ $\not{P}_T$ MET)

some particles may not register in detector

longitudinal momentum usually unconstrained

η important since $\Delta\eta \sim$ invariant (under Lorentz transformations)

Many combinations of $\mathbf{p}_1$, $\mathbf{p}_2$ could be useful $\quad$ more with n particles

$\Delta\, \phi_{12}$ $\quad$ invariant

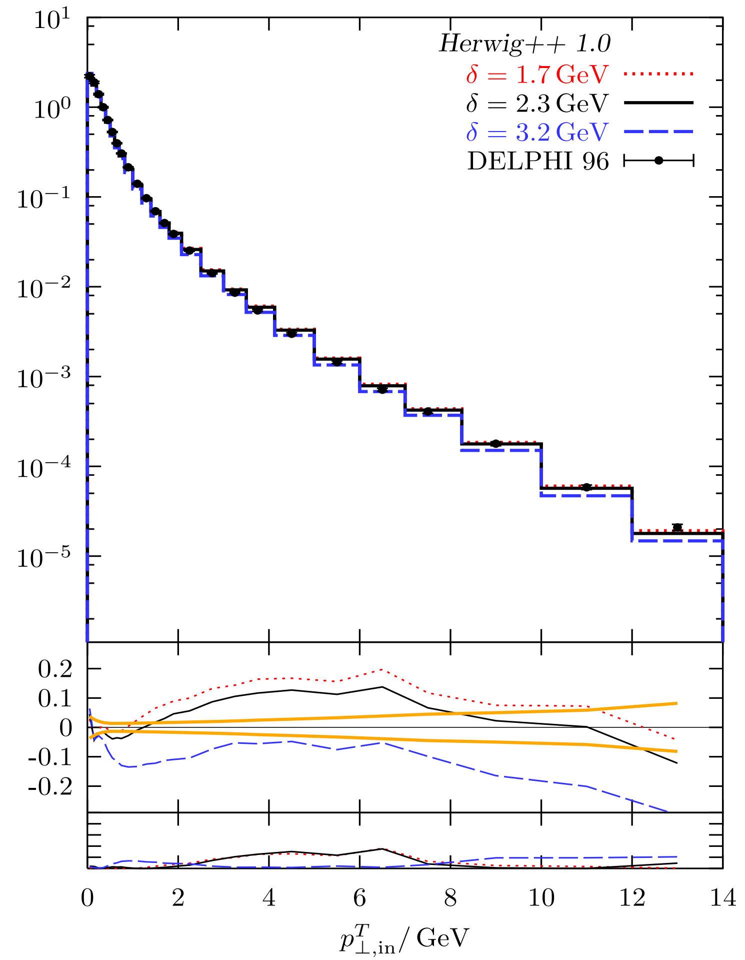
$m_{12} = \sqrt{[\, (p_1 + p_2)^2 - (\mathbf{p}_1 + \mathbf{p}_2)^2 \,]}$ $\quad$ invariant

sometimes expect a sharp peak, sometimes broader…

Masses are subject to combinatoric backgrounds

many candidates to try as p_1, p_2

Event Generators are checked against data



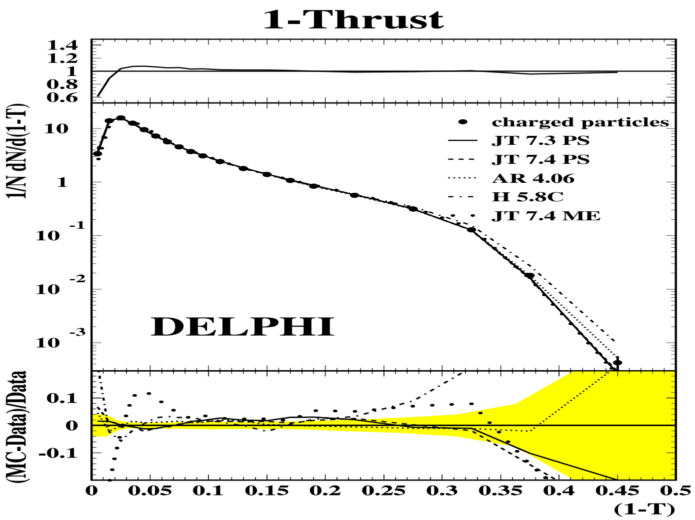
Momentum transverse to the thrust axis in the event plane.

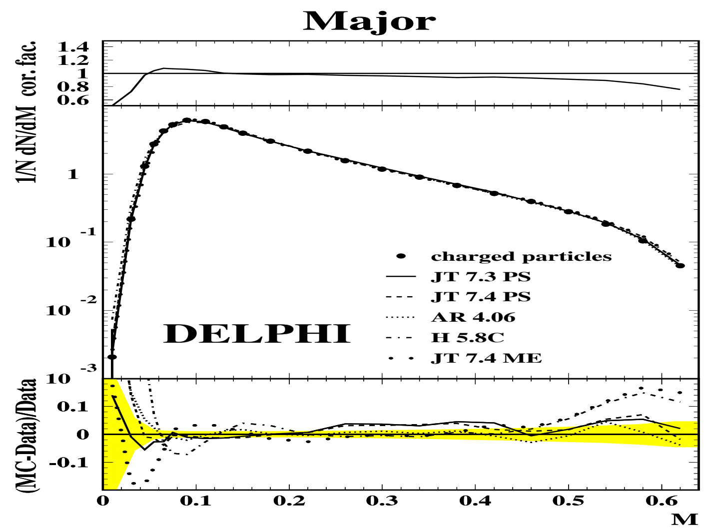
Often distributions fall steeply

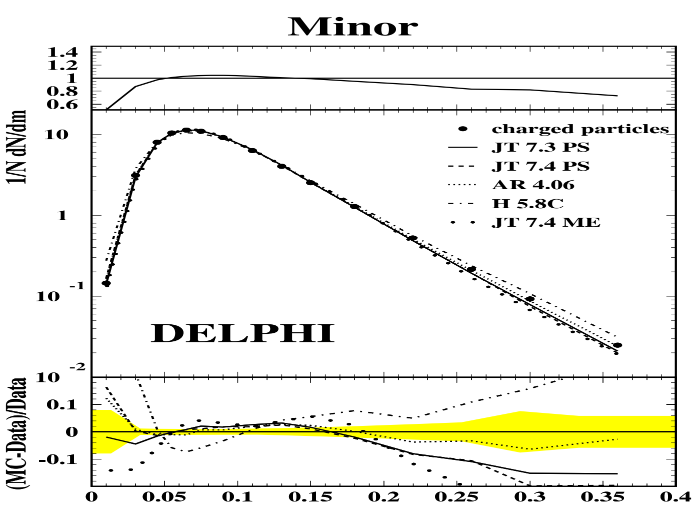
Here quasi-exponential;

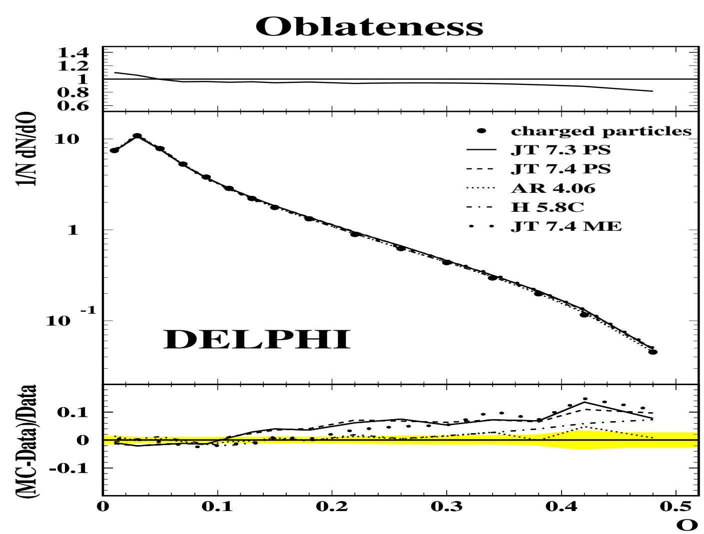
5 orders of magnitude range

Many many distributions are checked…

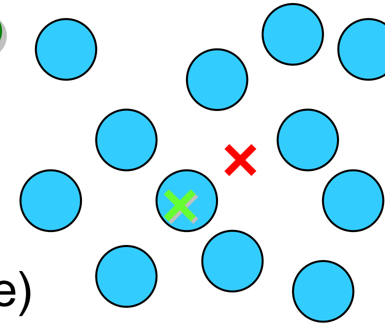








What are "Cross Sections"?



Intuition:

balloons on a dart board at a fair

probability of dart hitting balloon (dart: incident particle)

area (cross section) of balloons

× balloons / area (balloon: target particle)

Rate = cross section x Intensity of beam (luminosity)

so cross section **proportional to probability**

calculable intrinsic property of type of particle interaction

Use: $N = L\,\sigma\,\varepsilon$

N = number of events

L = integrated luminosity (exposure)

correct for deadtime: $L = L_{tot} \times$ (1-dead fraction)

σ = cross section another notational nuisance

sometimes: $\sigma \times$ BR

cross section x branching ratio

ε = efficiency of detection

Differential cross section: $\frac{d\sigma}{dx}$

How Detector Simulations are done

Geant4 is standard framework

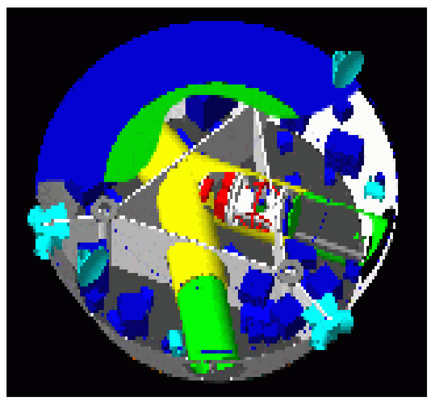
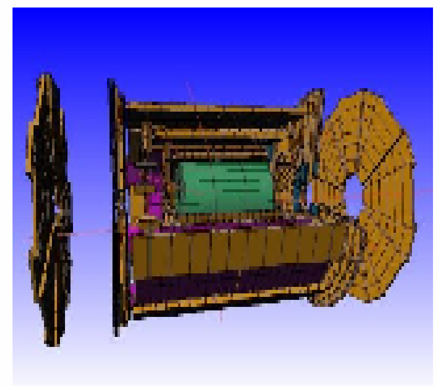
detector description package

physics of particle interactions with materials

10 X source code of an event generator

1M lines

International effort ~100 authors, 10 HEP labs


Used in medical (detectors, cancer treatment) and space physics

Substantial physics model verification effort

comparisons with experiment, other codes; lots of 1-D plots

Detector description: ~ 100K lines code

for each experiment

How Long does it take?

About 1 minute per event

2GHz Opteron

1-3% in event generator

90-95% in detector simulation

5% in analysis/reconstruction (~5 sec/event)

note: real data 50/sec

250 node farm to keep up with reconstruction

Need ~ 10^5-10^6 MC events

Few standard model events fake new physics

months on substantial processor farm

Need very long cycle random number generators

$2^{32} \sim 10^9$ is not nearly enough

Sometimes we can speed up detector simulation

Model, not directly simulate, 1-particle resolution distribution:

$$\text{prob}(x_{\text{measured}} \mid y_{\text{true}})$$

- parameterize detector and reconstruction in one step
  - as function of energy, location in detector
    - typically need non-Gaussian tails

Speedups of x 10-100

But: approximate (miss correlations, details)

- substantial effort to tune and certify

Worrisome if:

- backgrounds due to resolution tails of detector effects

Background Estimation

Run Monte Carlo for specific physics processes

can't always run as much as we should

ones which fake new physics can be rare, 10^{-5}

few pass the selection criteria ("cuts")

Or, scale from auxiliary data samples

"tails" of very common processes

occasionally, bootstrap-like calculations

"mixing" parts of separate events

Partial list of backgrounds generated one analysis…

Process	p_T	events	σ
$Z/\gamma^* \to ee + 2j$	[15-60]	86,250	26.2
$Z/\gamma^* \to ee + 2j$	[60-130]	203,450	28.3
$Z/\gamma^* \to ee + 2j$	[130-250]	93,500	0.271
single-top $e\nu bb$	-	15,500	0.115
single-top $e\nu bqb$	-	15,500	0.259
single-top $\mu\nu bb$	-	29,000	0.115
single-top $\mu\nu bqb$	-	15,500	0.259
$Z \to \nu\bar{\nu} + 1j$	-	245,250	529.
$W \to \tau\nu + 1j$	-	145,500	840.
$W \to \mu\nu + 1j$	-	98,750	840.
$W \to e\nu + 1j$	-	97,750	840.
$Z/\gamma^* \to \tau\tau + 1j$	[10-15]	97,249	67.5
$Z/\gamma^* \to \tau\tau + 1j$	[15-60]	90,250	80.8
$Z/\gamma^* \to \tau\tau + 1j$	[60-130]	96,500	81.1
$Z/\gamma^* \to \tau\tau + 1j$	[130-250]	0	0.760
$Z/\gamma^* \to \mu\mu + 1j$	[10-15]	146,500	67.5
$Z/\gamma^* \to \mu\mu + 1j$	[15-60]	255,000	80.8
$Z/\gamma^* \to \mu\mu + 1j$	[60-130]	325,995	81.1
$Z/\gamma^* \to \mu\mu + 1j$	[130-250]	24,000	0.760
$Z/\gamma^* \to ee + 1j$	[10-15]	147,750	67.5
$Z/\gamma^* \to ee + 1j$	[15-60]	171,000	80.8
$Z/\gamma^* \to ee + 1j$	[60-130]	186,750	81.1
$Z/\gamma^* \to ee + 1j$	[130-250]	39,000	0.760
QCD	[5-10]	103,000	7,359,000,000.
QCD	[10-20]	104,747	536,600,000.
QCD	[20-40]	104,239	30,360,000.
QCD	[40-80]	103,984	1,289,000.
QCD	[80-160]	114,988	38,599.

We try to share MC across analyses

What are "Cuts"?

Conditions which **select a subset of the data**

jet $P_T > 100$ GeV (GeV is an energy unit)

Reduce data volume

- so we can afford to store or process
- so it fits on my disk

Concentrate on where signal/background favorable

- or regions previously explored (with less data)

Remove regions hard to simulate accurately

- reduce systematic errors
  - want background not dominated by instrumental effects

Optimize statistical significance, say $\langle s \rangle / \sqrt{\langle b \rangle}$

oops…we say $\langle x \rangle$ for $E[x]$

Blame Dirac Quantum Mechanics notation for matrix elements

$\langle a \mid \text{Operator} \mid b \rangle$

Examples of cuts in an analysis

full sample ~ 10^9 events

cut applied	events left
CSskim NP without duplicate events	37,178,272
S1 : Remove bad runs and bad lbns JetMet	32,009,538
S2 : cal_event_quality	29,544,233
S3 : Trigger (*)	14,706,155
S4 : "sqgl" sub-skim	934,440
S5: Trigger (*)	877,067
S6: Acoplanarity < 165 degrees	504,390
P1 : Vertex $\lvert z\rvert < 60cm$	441,817
P2 : 1st leading jet $P_t > 60$ GeV/c	290,452
P3 : 2nd leading jet $P_t > 40$ GeV/c	99,851
P4 : 1st leading jet $\lvert\eta_{\rm det}\rvert < 0.8$	52,469
P5 : 2nd leading jet $\lvert\eta_{\rm det}\rvert < 0.8$	24,703
: 1st leading jet EMF < 0.95	24,480
: 2nd leading jet EMF < 0.95	23,913
: 1st leading jet CPF > 0.05	19,505
: 2nd leading jet CPF > 0.05	19,161
: Bad jet veto ($P_t > 15$ GeV/c)	14,811

Coarse selection

Jet quality

cut applied	events left	efficiency (%)
Common pre-selection	14,811	27.8
DI1: 2nd leading jet $P_t > 50$ GeV/c	10,423	27.0
DI2: $\not{E}_T > 60$ GeV	1,375	25.9
DI3: EM veto	1,242	21.7
DI4: Muon veto	1,162	19.3
DI5: $\Delta\Phi(\not{E}_T, jet1) > 90$ degrees	1,119	19.0
DI6: $\Delta\Phi(\not{E}_T, jet2) > 50$ degrees	543	18.1
DI7: $\Delta\Phi\min(\not{E}_T, any\,jet) > 40$ degrees	275	14.7
DI8: $H_T > 275$ GeV	33	10.1
DI9: $\not{E}_T > 175$ GeV	6	6.23

Optimized analysis cuts

What is  Triggering?

Collections of cuts applied before data is recorded

events occur at 10^6/s; we can record 10^2/s, so which are interesting?

often implemented in specialized hardware—electronic "exposure button"

analogy to "triggering" an oscilloscope to record a trace

only one chance for this event—so cuts usually conservative

May use hundreds of trigger conditions (encoded as tag bits on event)

a single analysis uses groups of them

one for each related channel

Z→ pairs of electrons, or mu, or tau particles (3 channels)

related triggers for cross checks of analyses

different analyses look for distinct final states

Examples:

3 jets > 25 GeV

electron > 20, another e > 5, missing E_T > 10

Estimating Signal Efficiency

Use physics generator for predicted new processes

predicted but undiscovered (possibly nonexistent!)

less verified than standard model

but OK for discovery typically

efficiency = probability it passes cuts (few %)

Often model has unknown parameters (e.g. masses)

Need ~ 10^4 events per parameter setting

Calculate efficiency of few % with reasonably accurately

Samples also used to determine/optimize cuts

Thou shalt determine thy cuts on simulation, not data!

Larger samples to train complex signal/background discriminants

A page from a typical talk

Search for generic Squarks and Gluinos in the multi-jet -- MET topology (D0)

Search for high MET and $H_T = \Sigma_{jet} E_T$ events in 3 regions of mSUGRA

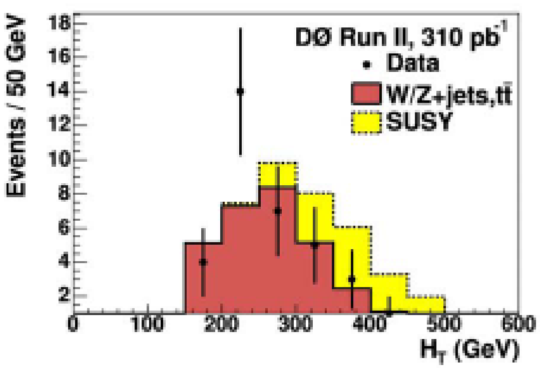
Small m_0: $M_{sq} < M_{gl}$	2 acoplanar jets	(>60,50 GeV)	H_T>275, MET>175 GeV
$M_{sq} \sim M_{gl}$:	3 jets	(>60,40,30 GeV)	H_T>350, MET>100 GeV
Large m_0: $M_{gl} < M_{sq}$	4 jets	(>60,40,30,20 GeV)	H_T>225, MET> 75 GeV

Data and SM bg are in agreement

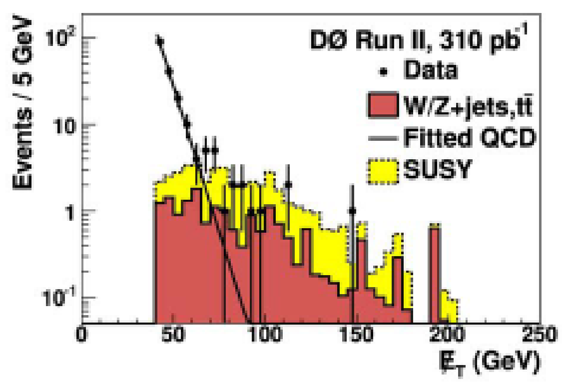
2jet:	6	$4.8^{+4.5}_{-2.1}$
3jet:	4	$3.9^{+1.5}_{-1.3}$
4jet:	10	$10.3^{+2.4}_{-2.9}$

Calculate limits:

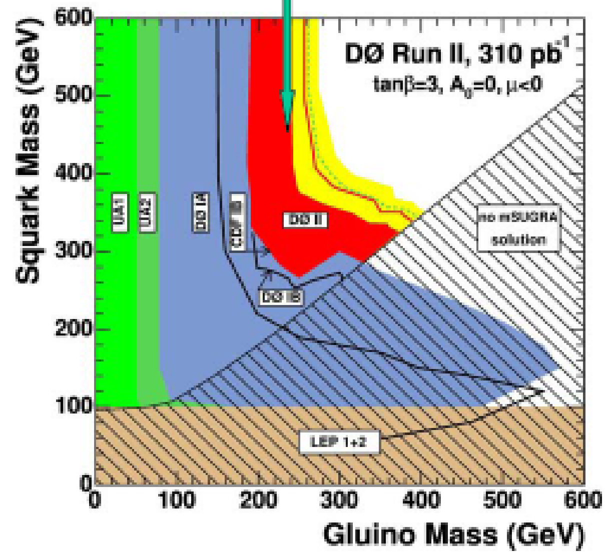
Theoretical cross section reduced by its uncertainties



"3-jet"



"4-jet"



M_{gl} > 241 GeV/c² ; M_{sq} > 325 GeV/c²

Physics at LHC, Cracow, July 2006 | E.Nagy - Tevatron SUSY Results | 21

Physicists: share code at phystat.org

site map accessibility

Phystat

Phystat Physics Statistics Code Repository

log i

An open, loosely moderated repository for code, tools, and documents relevant to statistics in physics applications. Search and download access is universal; package submission is loosely moderated for suitability.

Using the Site

- Lists of packages
- Search for a package
- Submit a Package
- Comment on a package (not yet available)

About the Repository

- Repository Policies and Procdures
- The Phystat Repository Steering Committee
- Comment on the repository site or policies

PHYSTAT Conference Links

- PHYSTAT 05 (Oxford, 2005)
- PHYSTAT 03 (SLAC 2003)
- More Conferences and Workshops ...

Lists and Statistics Resources

- The R Project For Statistical Computing
- StatCodes (Center for Astrostatistics)
- More resources ...

"Toy" MC for statistical methods

"Toy" because simpler (and faster) than full MC simulation

Examples:

Pr($\Delta\chi^2$ > 7.5) for adding fit to a normal distribution to a smooth background:

smoothly in falling e^{-x} spectrum with 300 events

plus, in data, a bump

Pr(k > 10) observed events if the background mean is 5.3, but known to 30%

same, but with NN discriminant > 0.65

retrained with "similar" training samples?

Have to choose relevant ensemble correctly!

what to hold constant? $e^{-1.0\,x}$? 1000 events? <1000>?