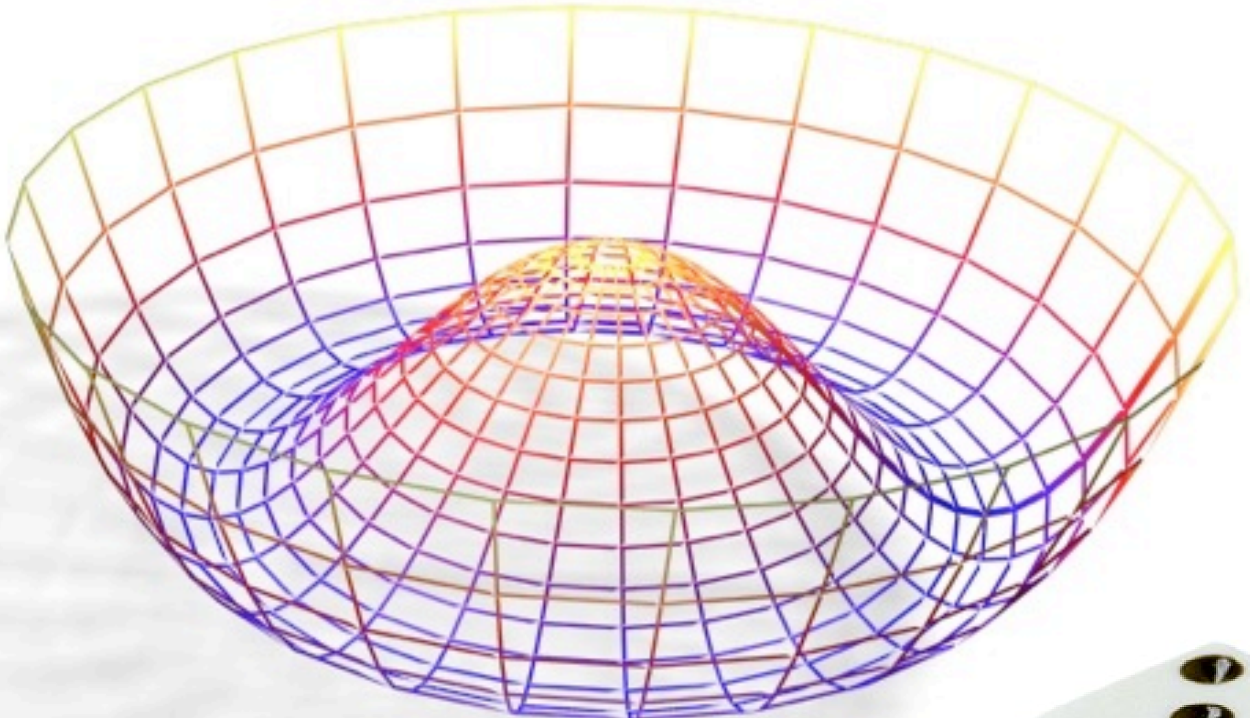
Center for
Cosmology and
Particle Physics

Practical Statistics for Particle Physics



Kyle Cranmer,
New York University

Introduction

Statistics plays a vital role in science, it is the way that we:

- quantify our knowledge and uncertainty
- communicate results of experiments

Big questions:

- how do we make discoveries, measure or exclude theory parameters, etc.
- how do we get the most out of our data
- how do we incorporate uncertainties
- how do we make decisions

Statistics is a very big field, and it is not possible to cover everything in 4 hours. In these talks I will try to:

- **explain** some fundamental ideas & prove a few things
- **enrich** what you already know
- **expose** you to some new ideas

I will try to go slowly, because if you are not following the logic, then it is not very interesting.

- Please feel free to ask questions and interrupt at any time

Further Reading

By physicists, for physicists

G. Cowan, Statistical Data Analysis, Clarendon Press, Oxford, 1998.

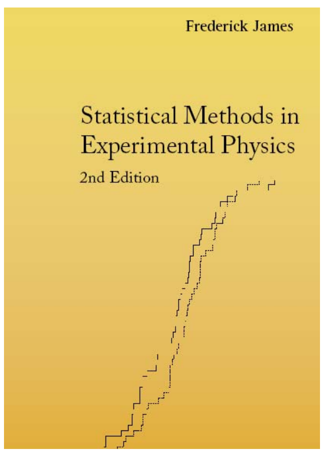
R.J.Barlow, A Guide to the Use of Statistical Methods in the Physical Sciences, John Wiley, 1989;

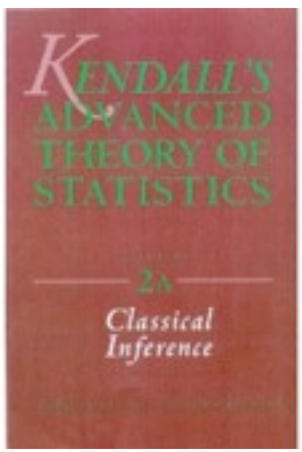
F. James, Statistical Methods in Experimental Physics, 2nd ed., World Scientific, 2006;

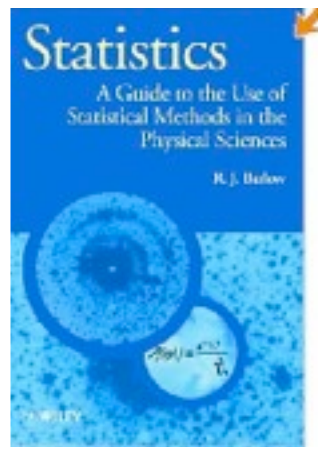
- W.T. Eadie et al., North-Holland, 1971 (1st ed., hard to find);

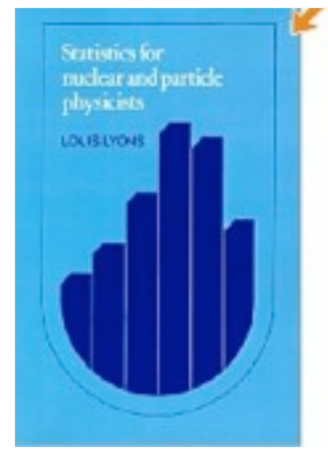
S.Brandt, Statistical and Computational Methods in Data Analysis, Springer, New York, 1998.

L.Lyons, Statistics for Nuclear and Particle Physics, CUP, 1986.









My favorite statistics book by a statistician:

Stuart, Ord, Arnold. "Kendall's Advanced Theory of Statistics" Vol. 2A *Classical Inference & the Linear Model.*

Other lectures

Fred James's lectures

http://preprints.cern.ch/cgi-bin/setlink?base=AT&categ=Academic_Training&id=AT00000799

http://www.desy.de/~acatrain/

Glen Cowan's lectures

http://www.pp.rhul.ac.uk/~cowan/stat_cern.html

Louis Lyons

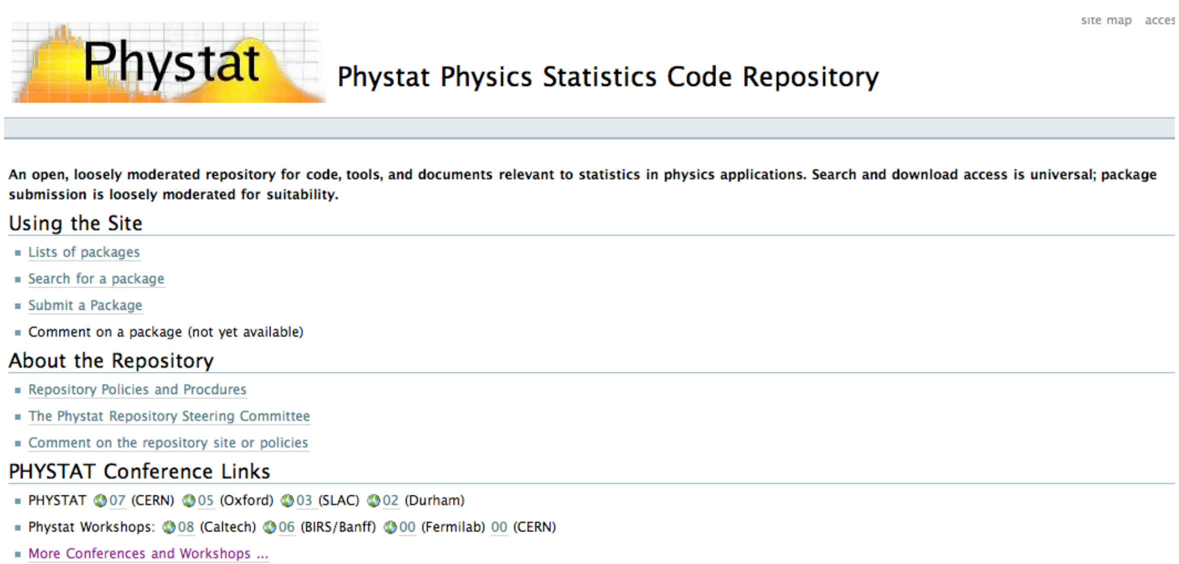
http://indico.cern.ch/conferenceDisplay.py?confId=a063350

Bob Cousins gave a CMS lecture, may give it more publicly

Gary Feldman "Journeys of an Accidental Statistician"

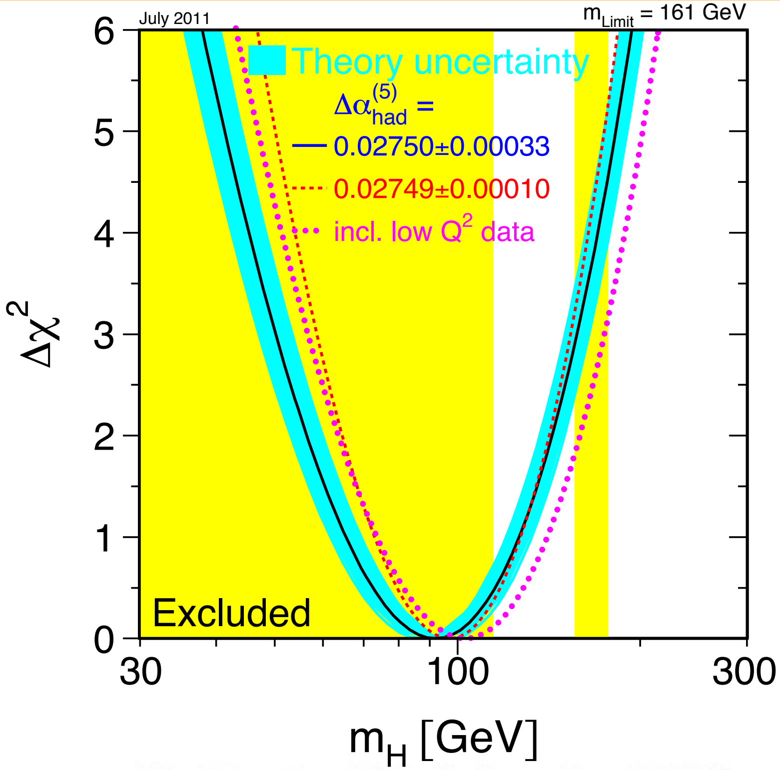
http://www.hepl.harvard.edu/~feldman/Journeys.pdf

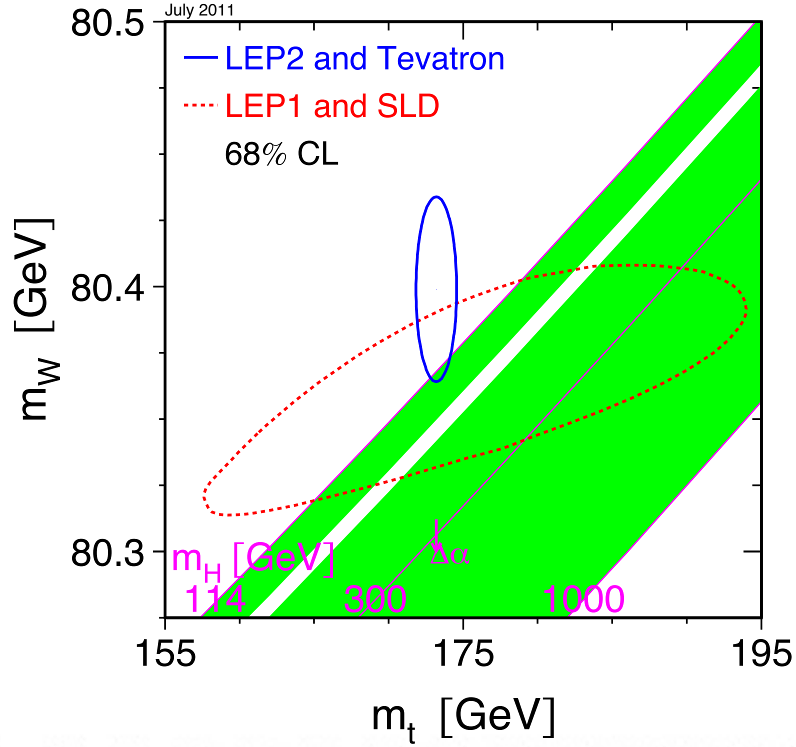
The PhyStat conference series at PhyStat.org:

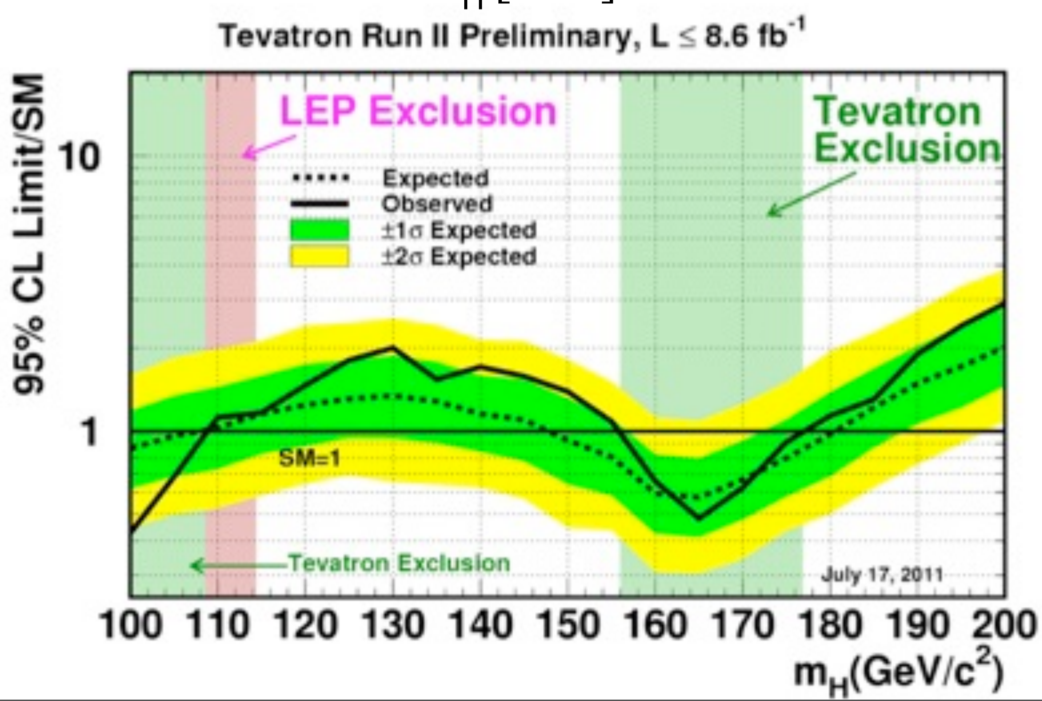


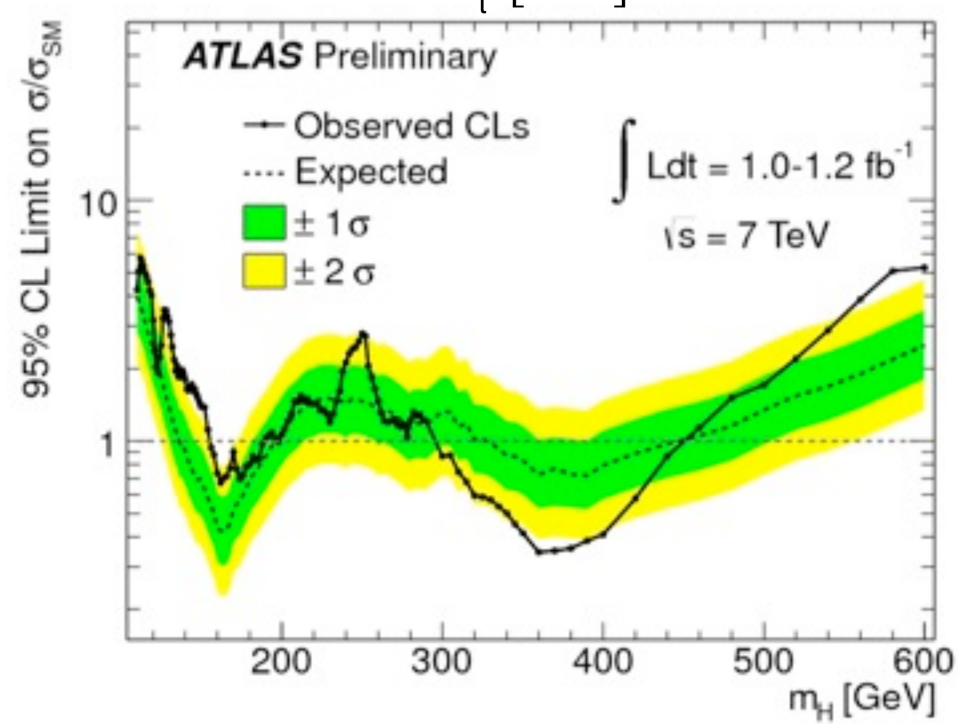
Lecture 1

What do these plots mean?









Preliminaries

Probability Density Functions

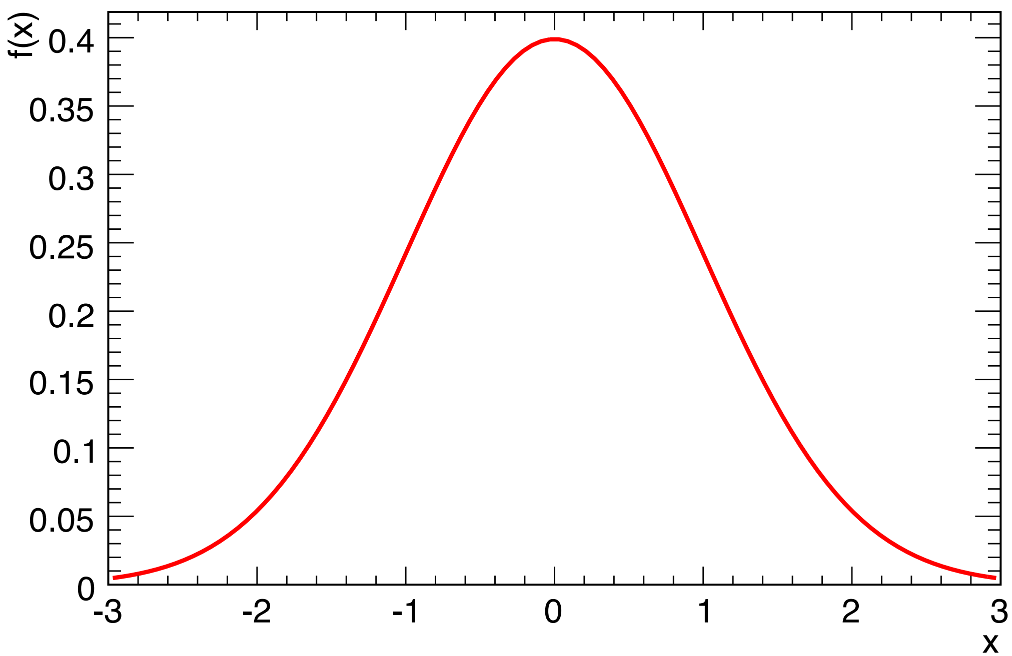
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function** (PDF... not parton distribution function)

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

PDFs are always normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



Probability Density Functions

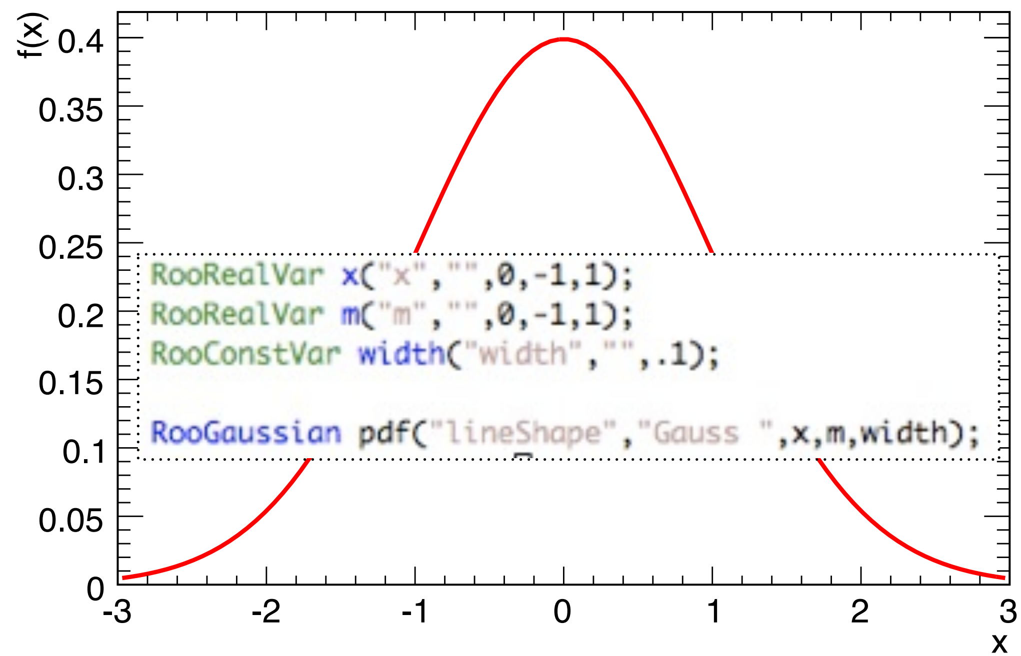
When dealing with continuous random variables, need to introduce the notion of a **Probability Density Function** (PDF... not parton distribution function)

$$P(x \in [x, x+dx]) = f(x)dx$$

Note, $f(x)$ is NOT a probability

PDFs are always normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$



Parametric PDFs

Center for Cosmology and Particle Physics

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu, \sigma)$ is parametrized by (μ, σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below

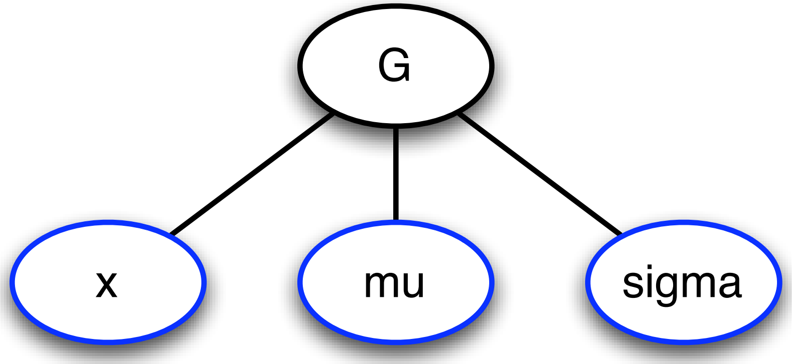
Parametric PDFs

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below



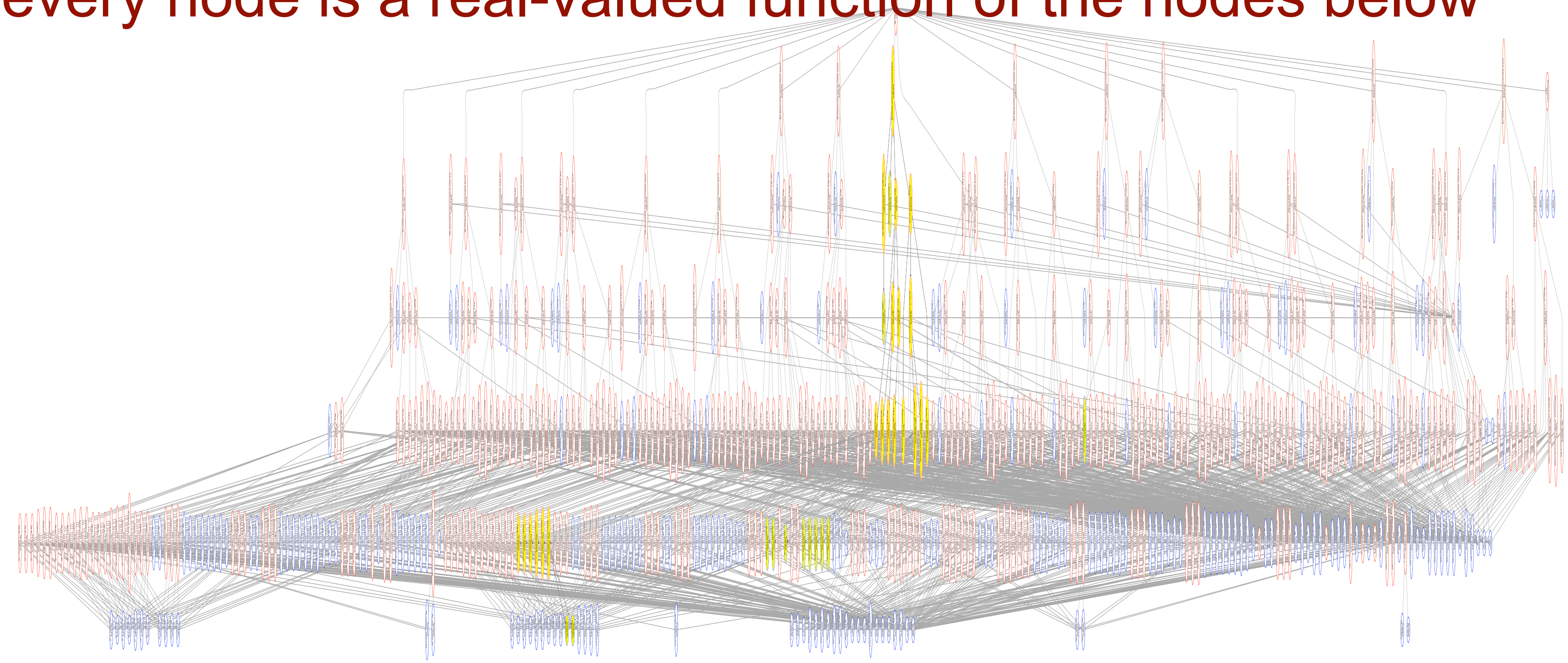
Parametric PDFs

Many familiar PDFs are considered **parametric**

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- defines a family of distributions
- allows one to make inference about parameters

I will represent PDFs graphically as below (directed acyclic graph)

- every node is a real-valued function of the nodes below



The Likelihood Function

A Poisson distribution describes a discrete event count n for a real-valued mean μ.

$$Pois(n|\mu) = \mu^n \frac{e^{-\mu}}{n!}$$

The likelihood of μ given n is the same equation evaluated as a function of μ

- Now it's a continuous function
- But it is not a pdf!

$$L(\mu) = Pois(n|\mu)$$

Common to plot the -2 ln *L*

- helps avoid thinking of it as a PDF
- connection to χ^2 distribution

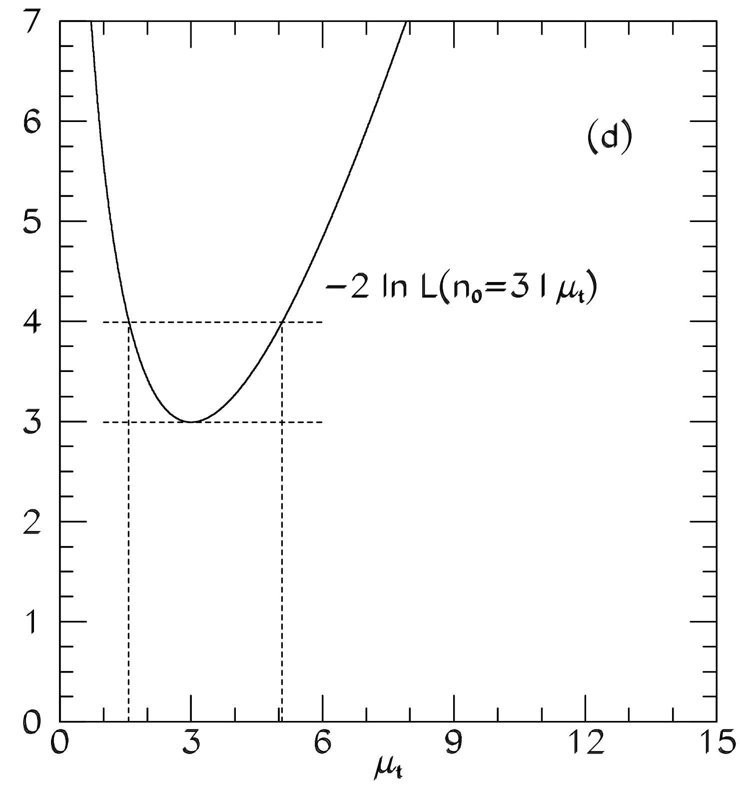


Figure from R. Cousins, Am. J. Phys. 63 398 (1995)

Change of variable x, change of parameter θ

- For pdf $p(x|\theta)$ and change of variable from x to y(x):

  $$p(y(x)|\theta) = p(x|\theta) \,/\, |dy/dx|.$$

  Jacobian modifies probability *density*, guaranties that

  $P(\, y(x_1) < y < y(x_2)\,) \;=\; P(x_1 < x < x_2\,)$, i.e., that

  Probabilities are invariant under change of variable x.

  - Mode of probability *density* is *not* invariant (so, e.g., criterion of maximum probability density is ill-defined).
  - Likelihood *ratio* is invariant under change of variable x. (Jacobian in denominator cancels that in numerator).

- For likelihood $\mathcal{L}(\theta)$ and reparametrization from θ to $u(\theta)$:

  $$\mathcal{L}(\theta) \;=\; \mathcal{L}(u(\theta)) \quad (!).$$

  - Likelihood $\mathcal{L}\,(\theta)$ is invariant under reparametrization of parameter θ (reinforcing fact that $\mathcal{L}$ is *not* a pdf in θ).

Bob Cousins, CMS, 2008

Probability Integral Transform

"...seems likely to be one of the most fruitful conceptions introduced into statistical theory in the last few years"
– Egon Pearson (1938)

Given continuous $x \in (a,b)$, and its pdf $p(x)$, let

$$y(x) = \int_a^x p(x')\, dx' .$$

Then $y \in (0,1)$ and $p(y) = 1$ (uniform) for all y. (!)

So there always exists a metric in which the pdf is uniform.

Many issues become more clear (or trivial) after this transformation*. (If x is discrete, some complications.)

The specification of a Bayesian prior pdf $p(\mu)$ for parameter μ is equivalent to the choice of the metric $f(\mu)$ in which the pdf is uniform. This is a *deep* issue, not always recognized as such by users of flat prior pdf's in HEP!

*And the inverse transformation provides for efficient M.C. generation of p(x) starting from RAN().

Bob Cousins, CMS, 2008

Different definitions of Probability

Frequentist

- defined as limit of long term frequency
- probability of rolling a 3 := limit of (# rolls with 3 / # trials)
  - you don't need an infinite sample for definition to be useful
  - sometimes ensemble doesn't exist
    - eg. P(Higgs mass = 120 GeV), P(it will snow tomorrow)
- Intuitive if you are familiar with Monte Carlo methods
- compatible with orthodox interpretation of probability in Quantum Mechanics. Probability to measure spin projected on x-axis if spin of beam is polarized along +z

$$|\langle \rightarrow | \uparrow \rangle|^2 = \frac{1}{2}$$

Subjective Bayesian

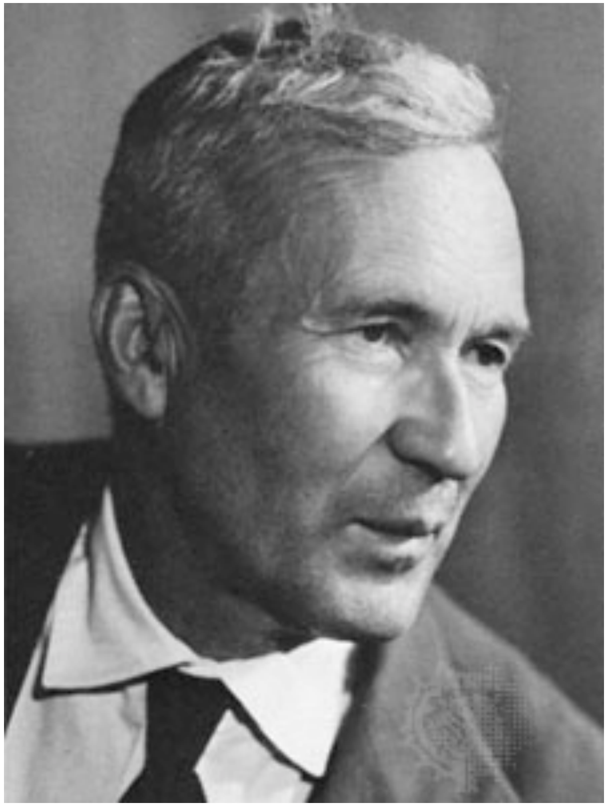
- Probability is a degree of belief (personal, subjective)
  - can be made quantitative based on betting odds
  - most people's subjective probabilities are not ***coherent*** and do not obey laws of probability

http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/#3.1

Axioms of Probability

These Axioms are a mathematical starting point for probability and statistics

1. probability for every element, E, is non-negative $P(E) \geq 0 \qquad \forall E \subseteq \mathcal{F} = 2^{\Omega}$
2. probability for the entire space of possibilities is 1 $P(\Omega) = 1.$
3. if elements E_i are disjoint, probability is additive $P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$



Kolmogorov axioms (1933)

Consequences:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(\Omega \setminus E) = 1 - P(E)$$

Bayes' theorem relates the conditional and marginal probabilities of events A & B

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

- P(A) is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- P(A|B) is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- P(B|A) is the conditional probability of B given A.
- P(B) is the prior or marginal probability of B, and acts as a normalizing constant

Derivation from conditional probabilities

To derive the theorem, we start from the definition of conditional probability. The probability of event A given event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Equivalently, the probability of event B given event A is

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

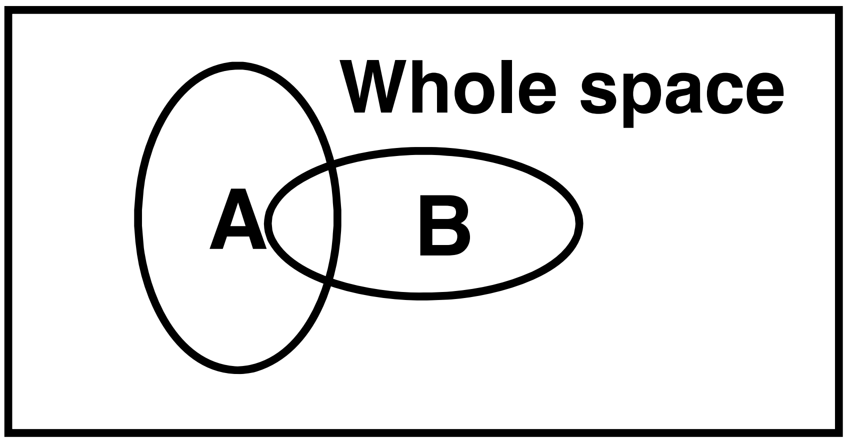
Rearranging and combining these two equations, we find

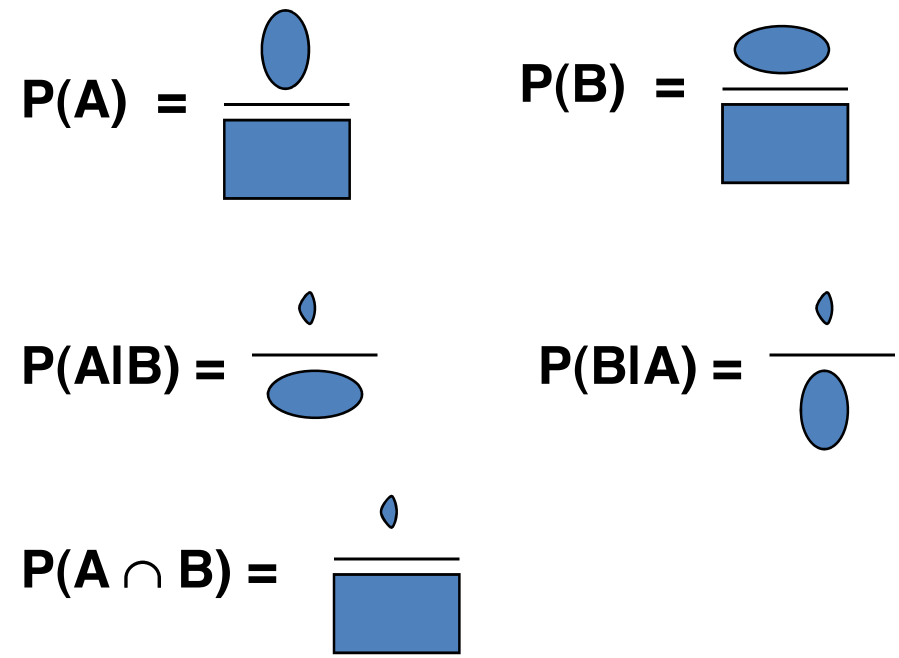
$$P(A|B)\,P(B) = P(A \cap B) = P(B|A)\,P(A).$$

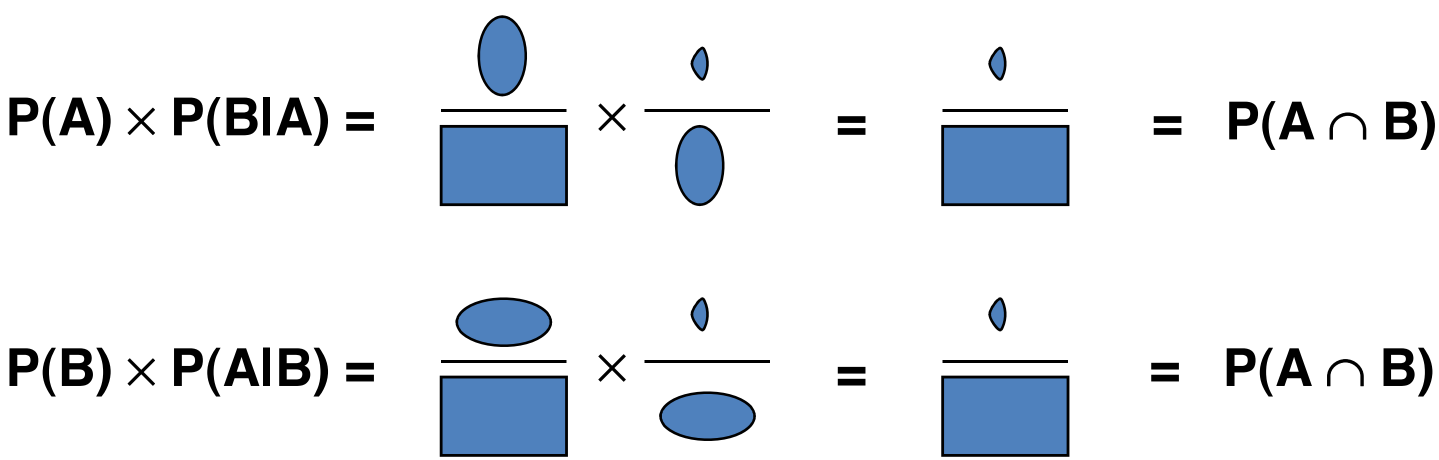
This lemma is sometimes called the product rule for probabilities. Dividing both sides by P(B), providing that it is non-zero, we obtain Bayes' theorem:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)\,P(A)}{P(B)}.$$

P, Conditional P, and Derivation of Bayes' Theorem in Pictures





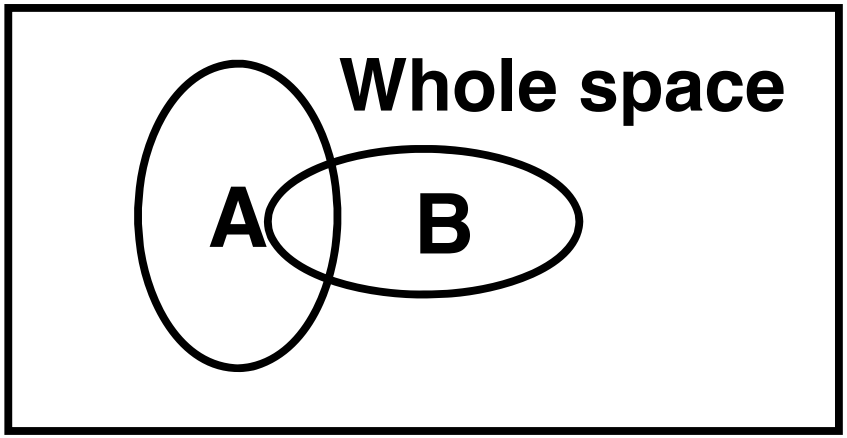


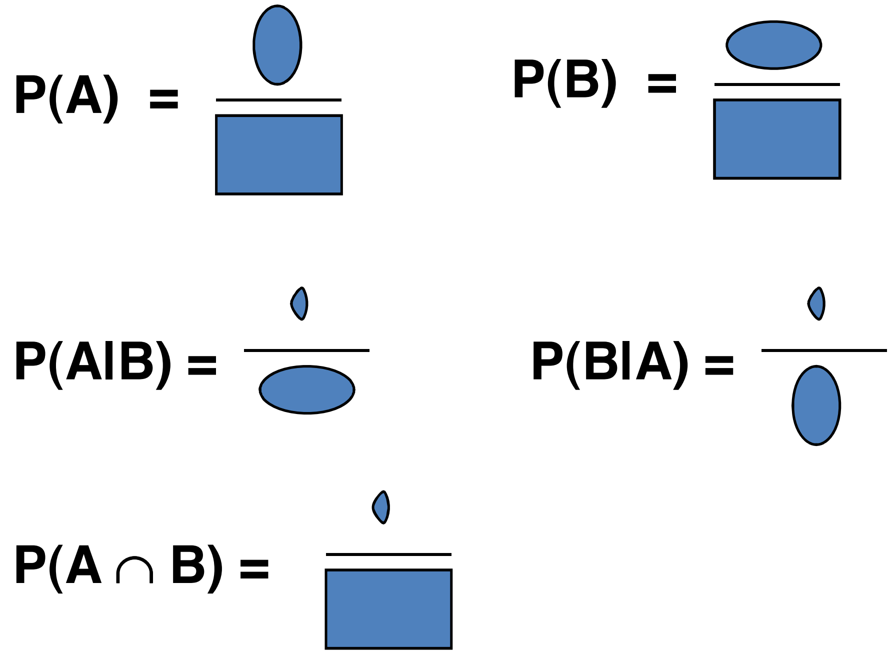
Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

... in pictures (from Bob Cousins)

P, Conditional P, and Derivation of Bayes' Theorem in Pictures





Don't forget about "Whole space" Ω. I will drop it from the notation typically, but occasionally it is important.

Bob Cousins, CMS, 2008

$\Rightarrow$ **P(B|A) = P(A|B) × P(B) / P(A)**

Louis's Example

P (Data;Theory) $\neq$ P (Theory;Data)

Theory = male or female

Data = pregnant or not pregnant

P (pregnant ; female) ~ 3%

but

P (female ; pregnant) >>>3%

Modeling:
The Scientific Narrative

Building a model of the data

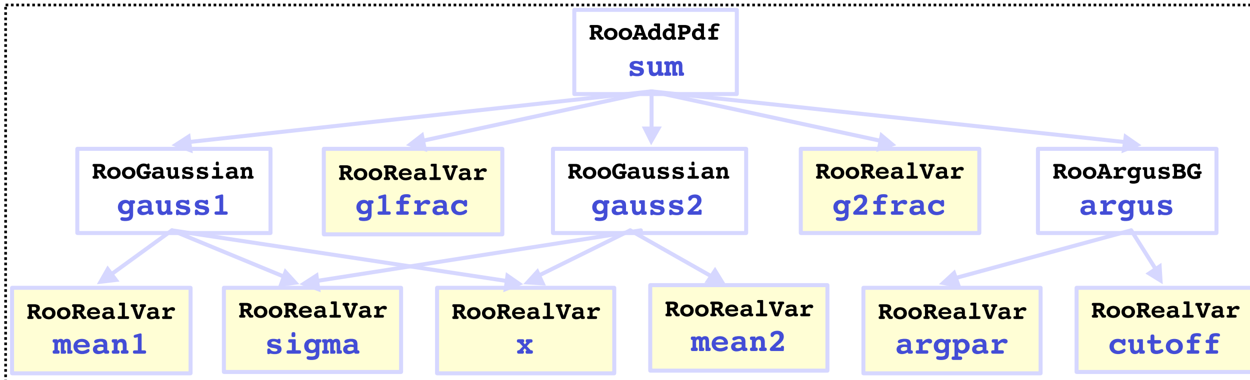
Before one can discuss statistical tests, one must have a "**model**" for the data.

- by "model", I mean the full structure of P(data | parameters)
  - holding parameters fixed gives a PDF for data
  - ability to evaluate generate pseudo-data (Toy Monte Carlo)
  - holding data fixed gives a **likelihood function** for parameters
    - note, likelihood function is not as general as the full model because it doesn't allow you to generate pseudo-data

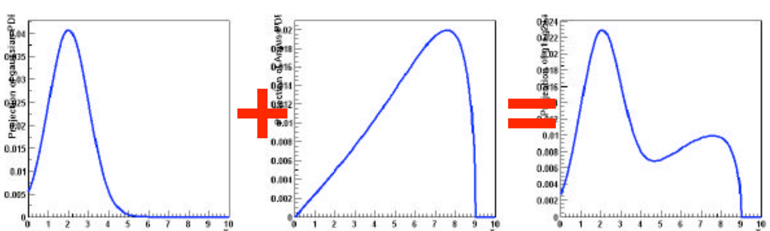
Both Bayesian and Frequentist methods start with the model

- it's the objective part that everyone can agree on
- it's the place where our physics knowledge, understanding, and intuiting comes in
- building a better model is the best way to improve your statistical procedure

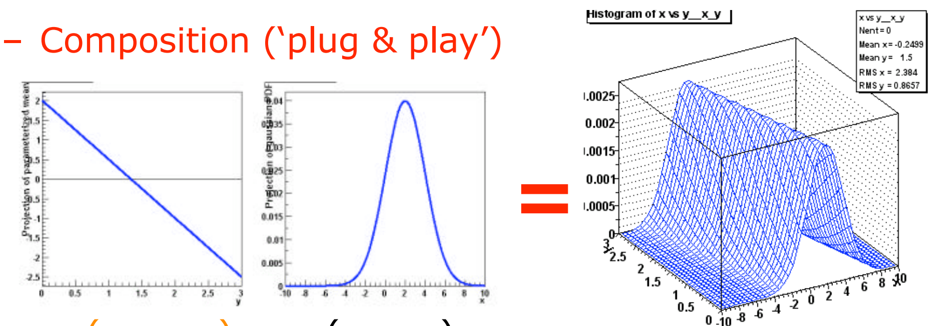
RooFit is a major tool developed at BaBar for data modeling.
RooStats provides higher-level statistical tools based on these PDFs.



– Addition



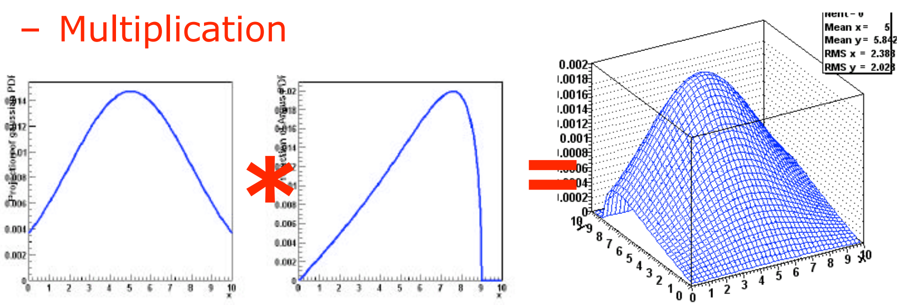
– Composition ('plug & play')



$m(y;a_0,a_1)$ $g(x;m,s)$ $g(x,y;a0,a1,s)$

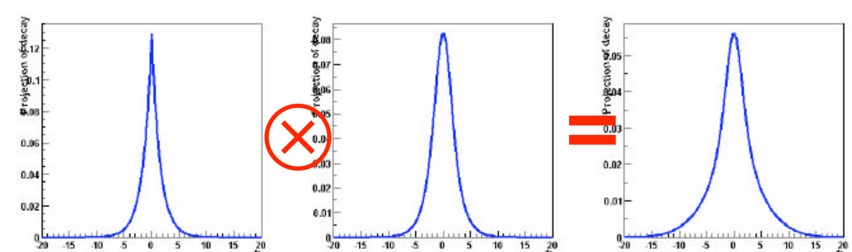
Possible in *any* PDF
No explicit support in PDF code needed

– Multiplication



Wouter Verkerke,

– Convolution



Wouter Verkerke, UCSB

The Scientific Narrative

The model can be seen as a quantitative summary of the analysis

- If you were asked to justify your modeling, you would tell a **story** about why you know what you know
  - based on previous results and studies performed along the way
- the quality of the result is largely tied to how convincing this story is and how tightly it is connected to model

I will describe a few “narrative styles”

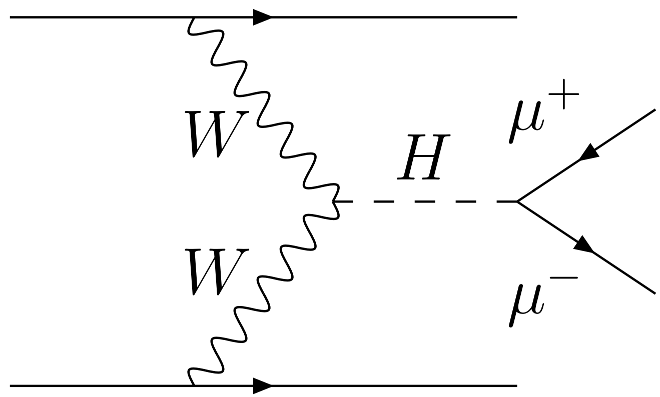
- The “Monte Carlo Simulation” narrative
- The “Data Driven” narrative
- The “Effective Modeling” narrative
- The “Parametrized Response” narrative

Real-life analyses often use a mixture of these

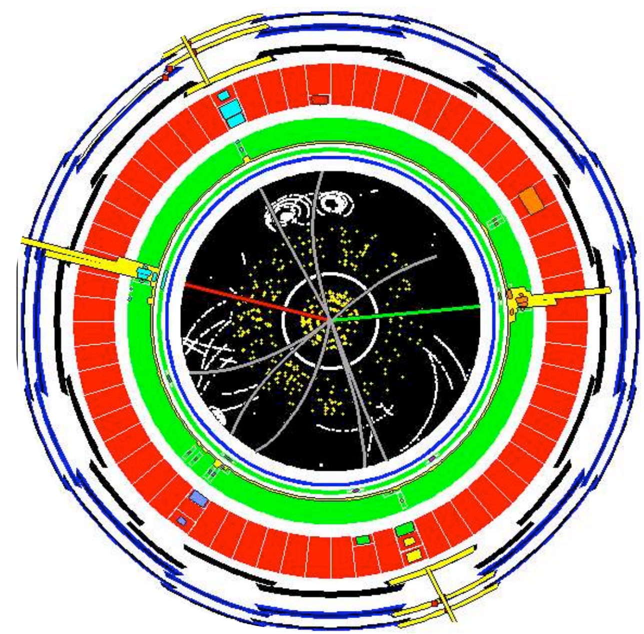
The Monte Carlo Simulation narrative

Let's start with "the Monte Carlo simulation narrative", which is probably the most familiar

$P =$



$\oplus$



The simulation narrative

1) The language of the Standard Model is Quantum Field Theory
Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$

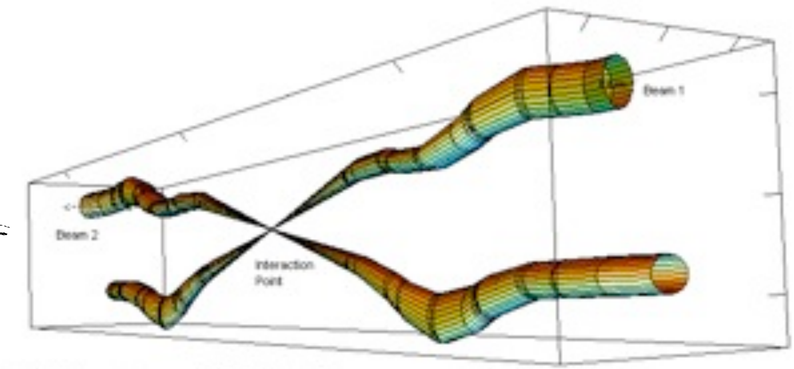
The simulation narrative

1) The language of the Standard Model is Quantum Field Theory
Phase space Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f | i \rangle|^2}{\langle f | f \rangle \langle i | i \rangle}$$

$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$



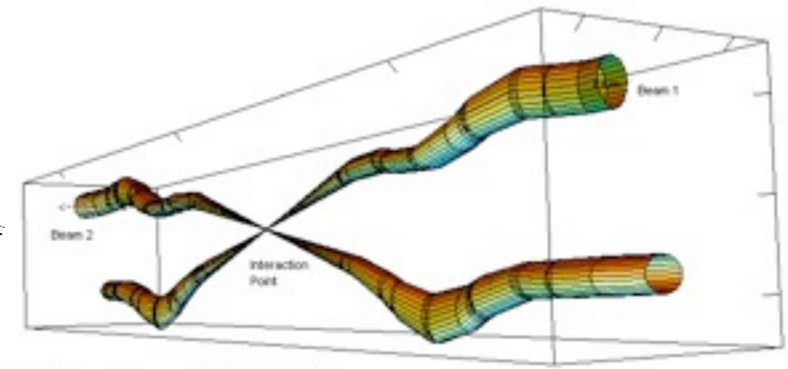
Relative beam sizes around IP1 (Atlas) in collision

1) The language of the Standard Model is Quantum Field Theory **Phase space** Ω defines initial measure, sampled via Monte Carlo

$$P = \frac{|\langle f|i\rangle|^2}{\langle f|f\rangle\langle i|i\rangle}$$

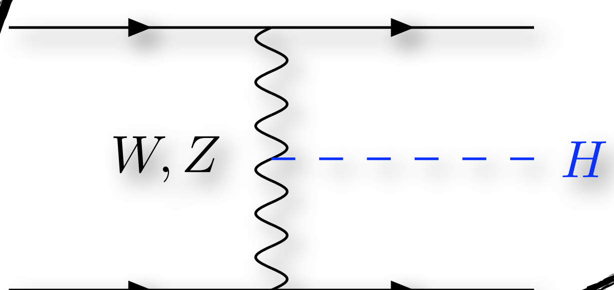
$$P \to L\sigma$$

$$d\sigma \to |\mathcal{M}|^2 d\Omega$$



Relative beam sizes around IP1 (Atlas) in collision

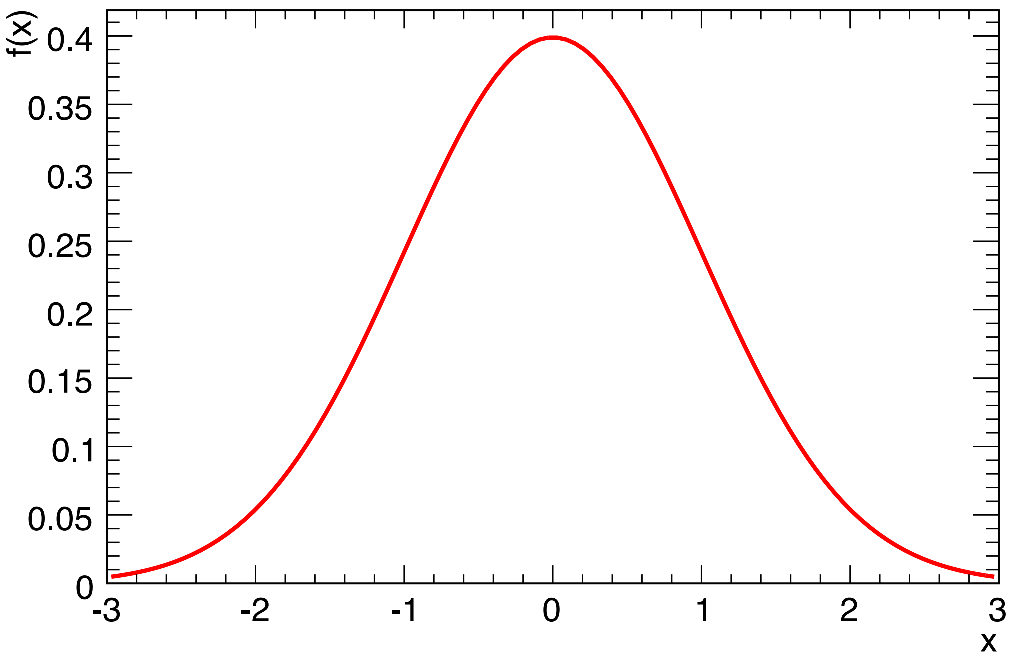
$$\begin{aligned}
\mathcal{L}_{SM} = \; & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
& + \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
& + \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}
\end{aligned}$$

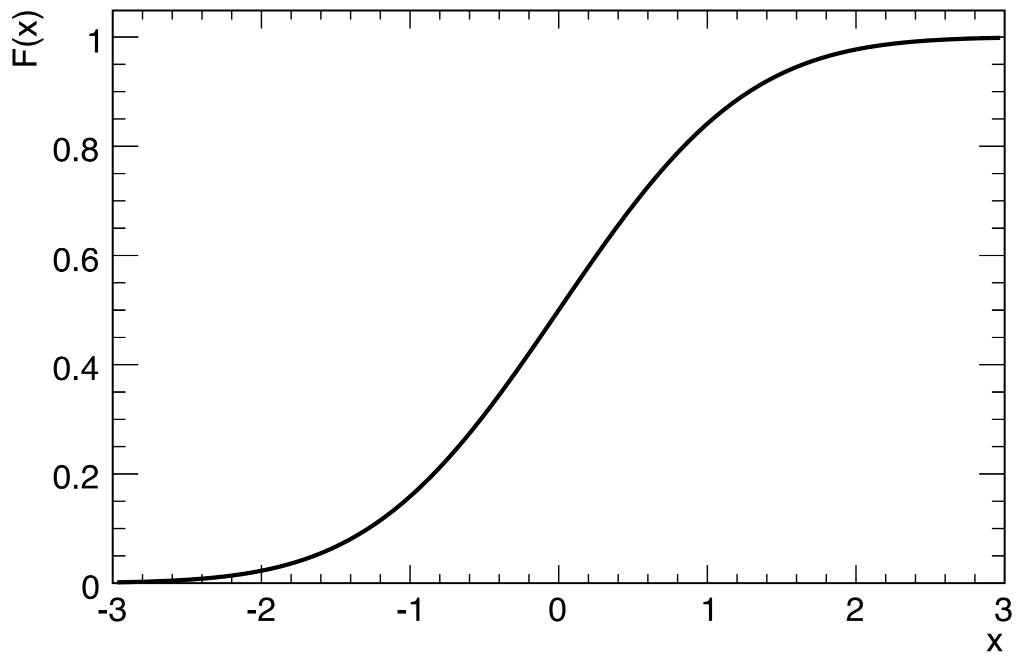


Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension: $$\int_{-\infty}^{x} f(x')dx' = F(x)$$



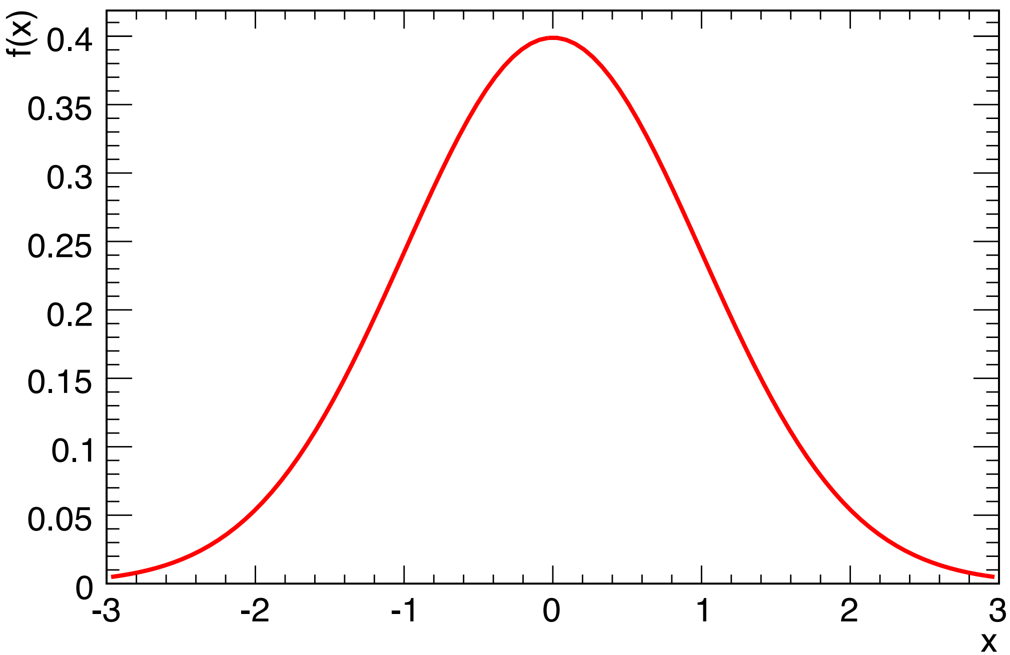


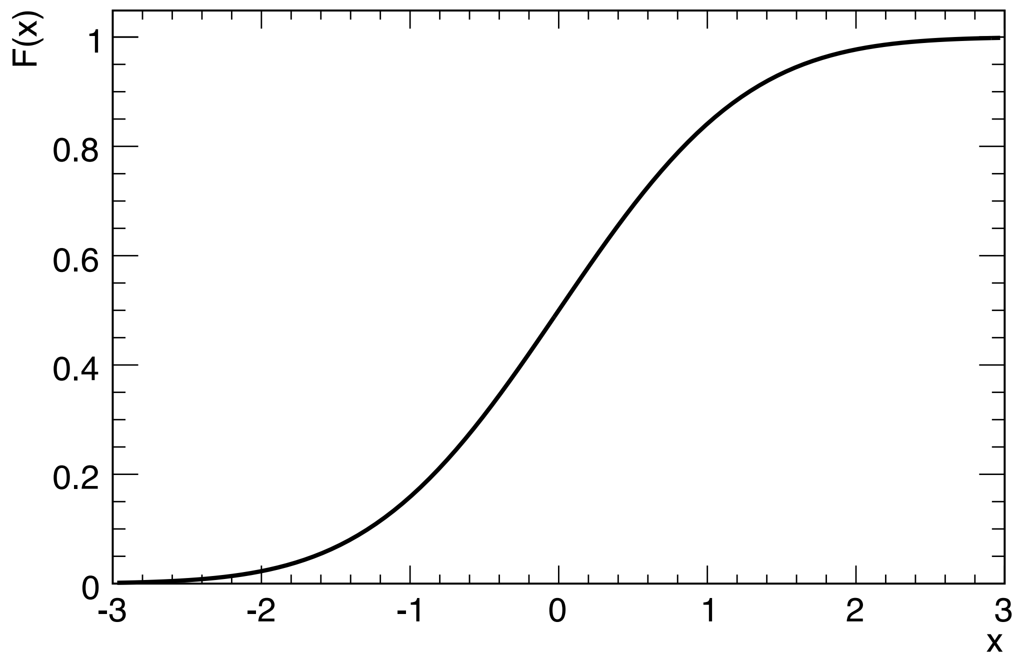
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

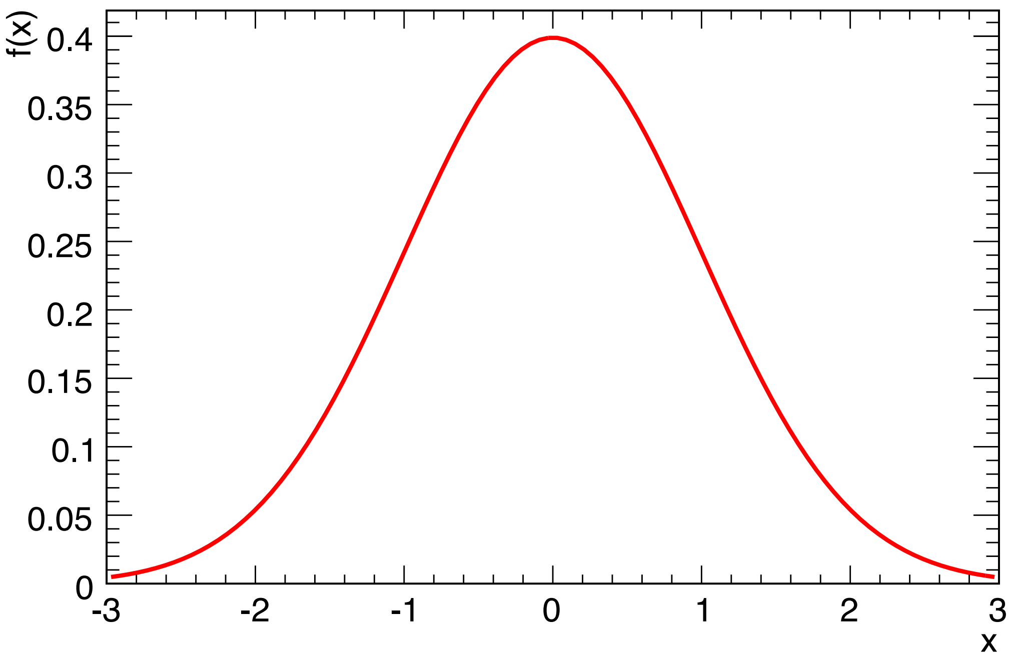
$$f(x) = \frac{\partial F(x)}{\partial x}$$

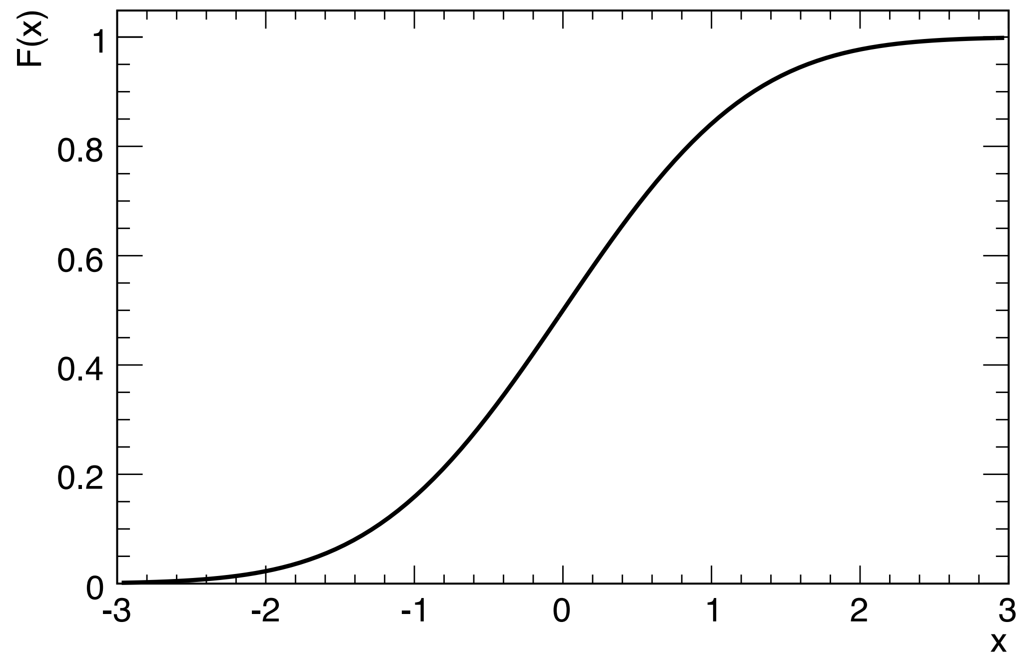
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

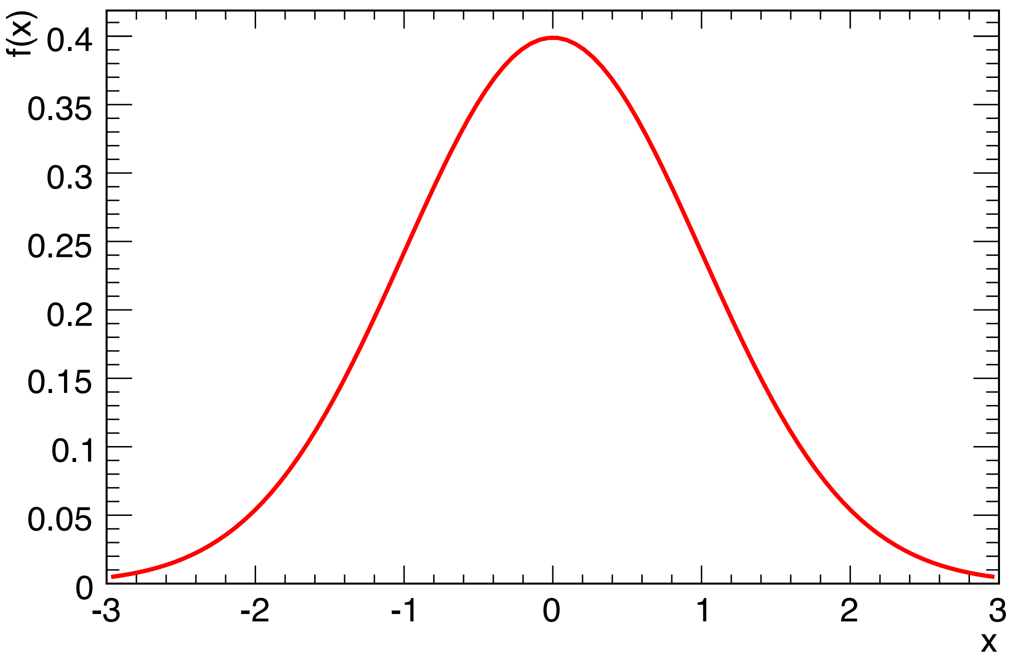
$$f(E) = \frac{1}{\sigma}\frac{\partial \sigma}{\partial E}$$

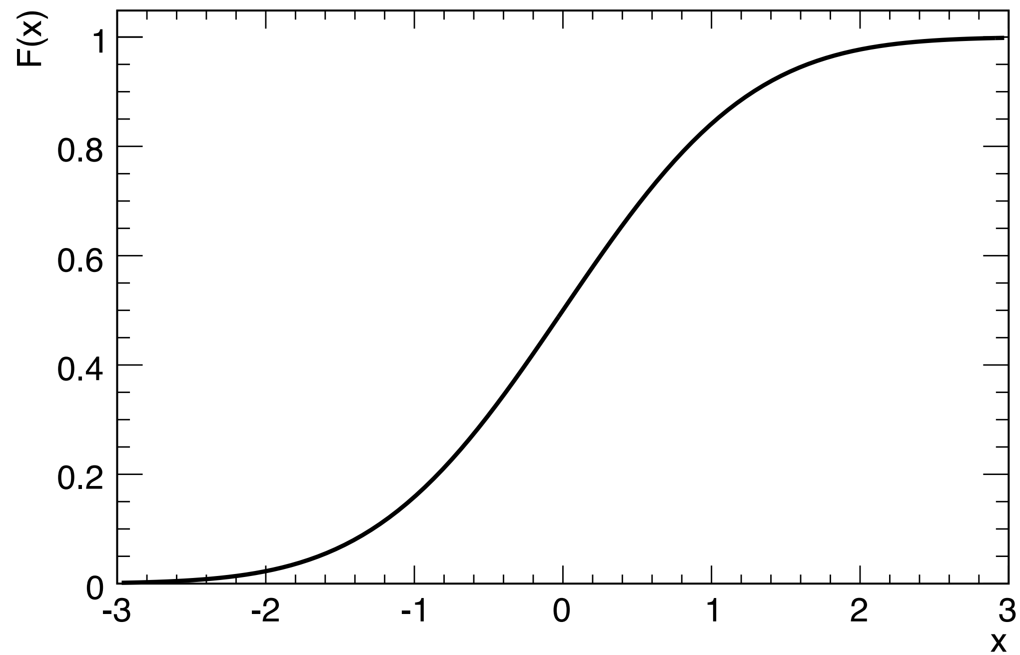
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

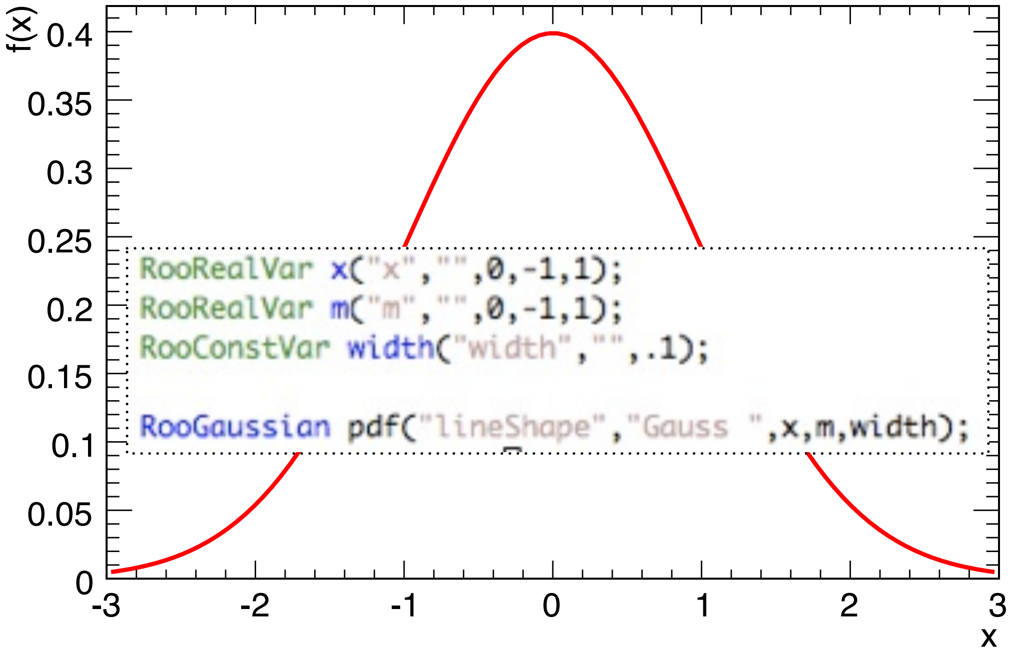
$$f(E,\eta) = \frac{1}{\sigma}\frac{\partial^2 \sigma}{\partial E \partial \eta}$$

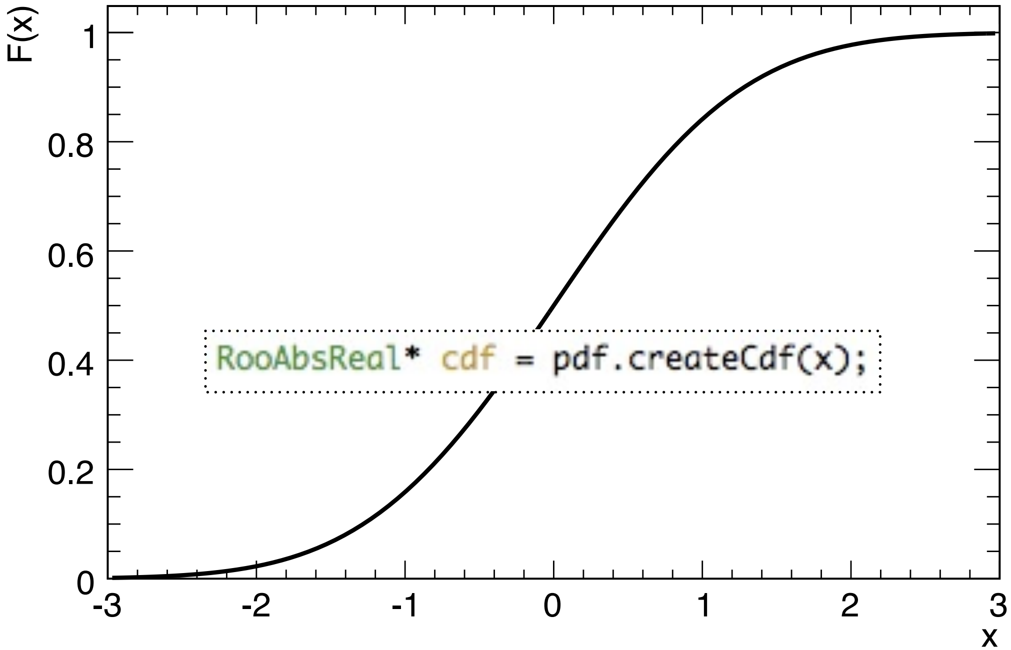
Cumulative Density Functions

Often useful to use a cumulative distribution:

- in 1-dimension:

$$\int_{-\infty}^{x} f(x')dx' = F(x)$$





- alternatively, define density as partial of cumulative:

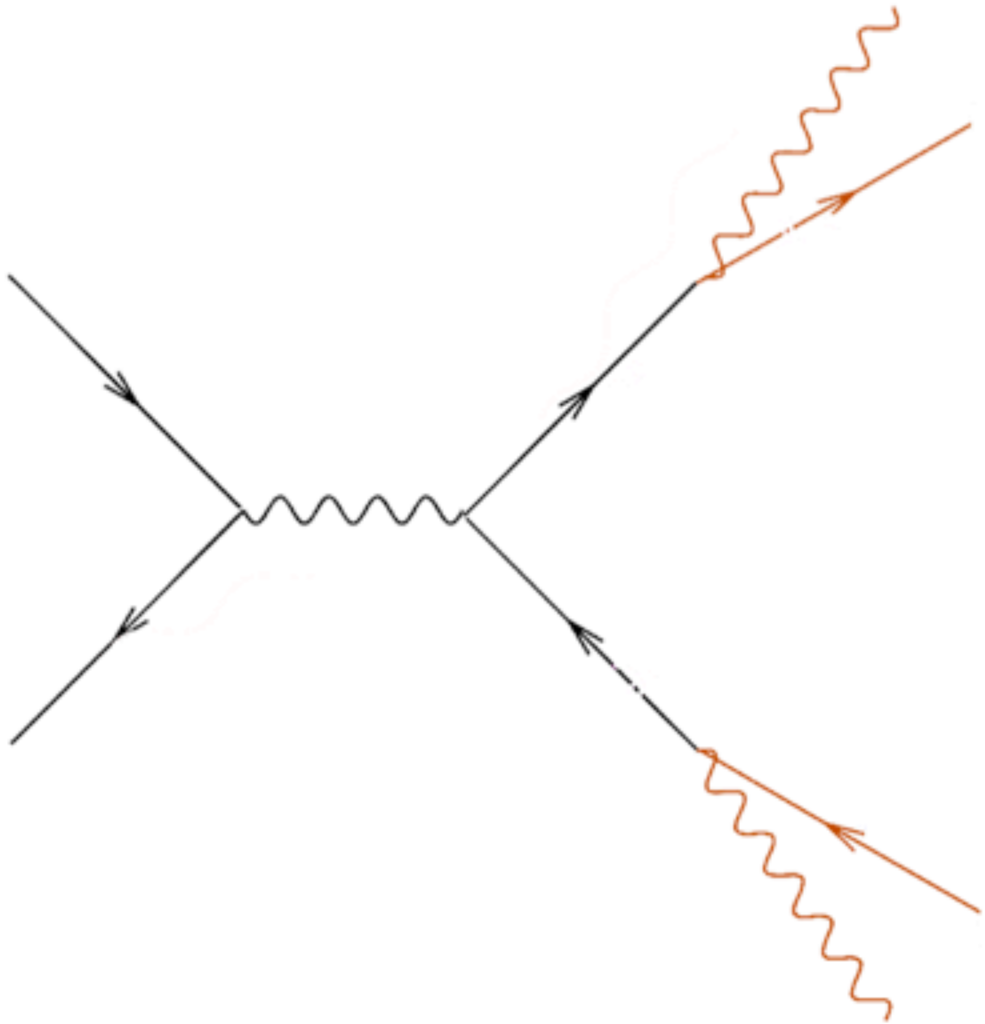
$$f(x) = \frac{\partial F(x)}{\partial x}$$

- same relationship as total and differential cross section:

$$f(E,\eta) = \frac{1}{\sigma}\frac{\partial^2 \sigma}{\partial E \partial \eta}$$

The simulation narrative

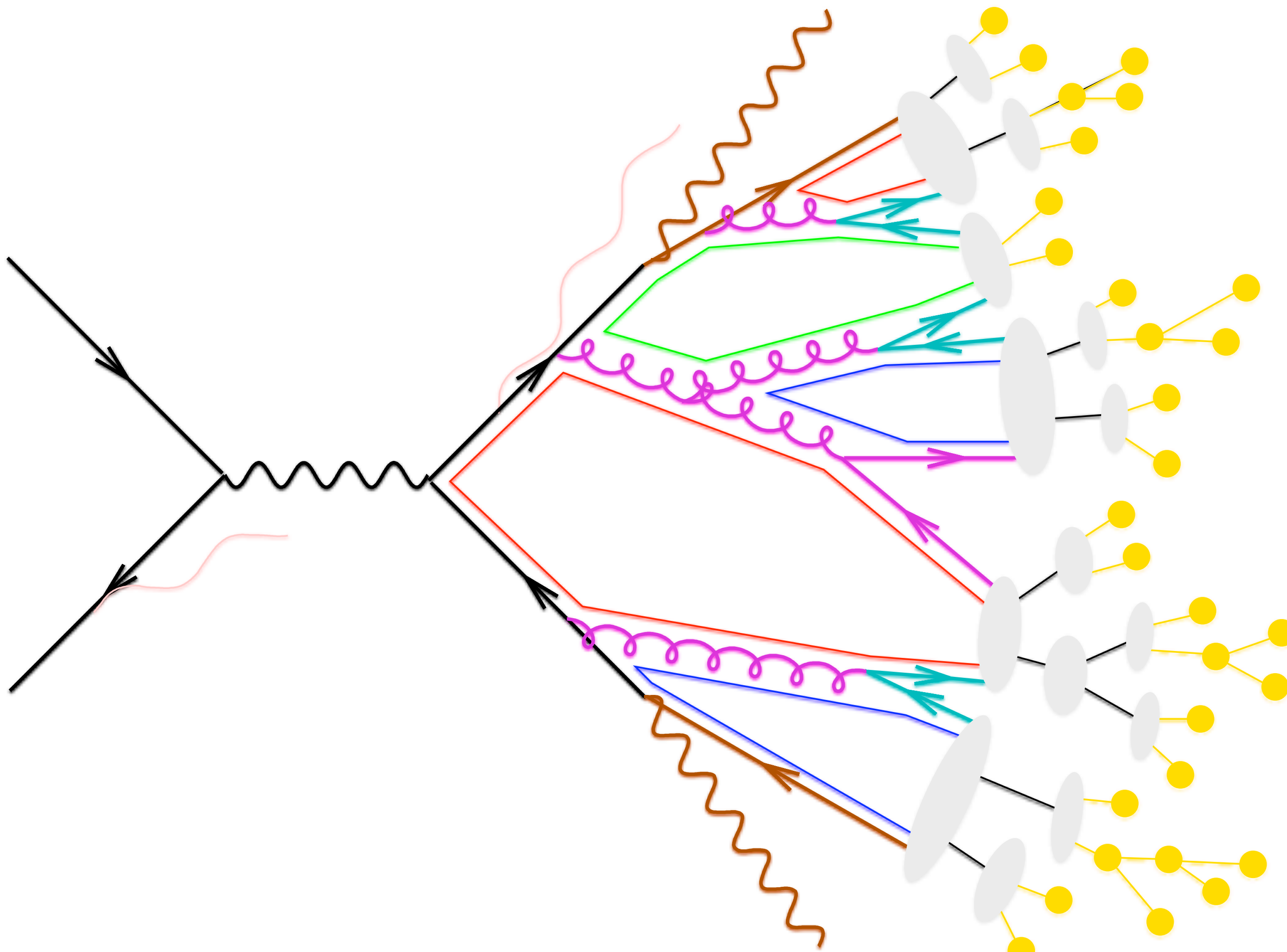
2) a) Perturbation theory used to systematically approximate the theory.
b) splitting functions, Sudokov form factors, and hadronization models
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**



- hard scattering
- partonic decays, e.g. $t \to bW$

The simulation narrative

2)
a) Perturbation theory used to systematically approximate the theory.
b) splitting functions, Sudokov form factors, and hadronization models
c) all sampled via accept/reject Monte Carlo **P(particles | partons)**

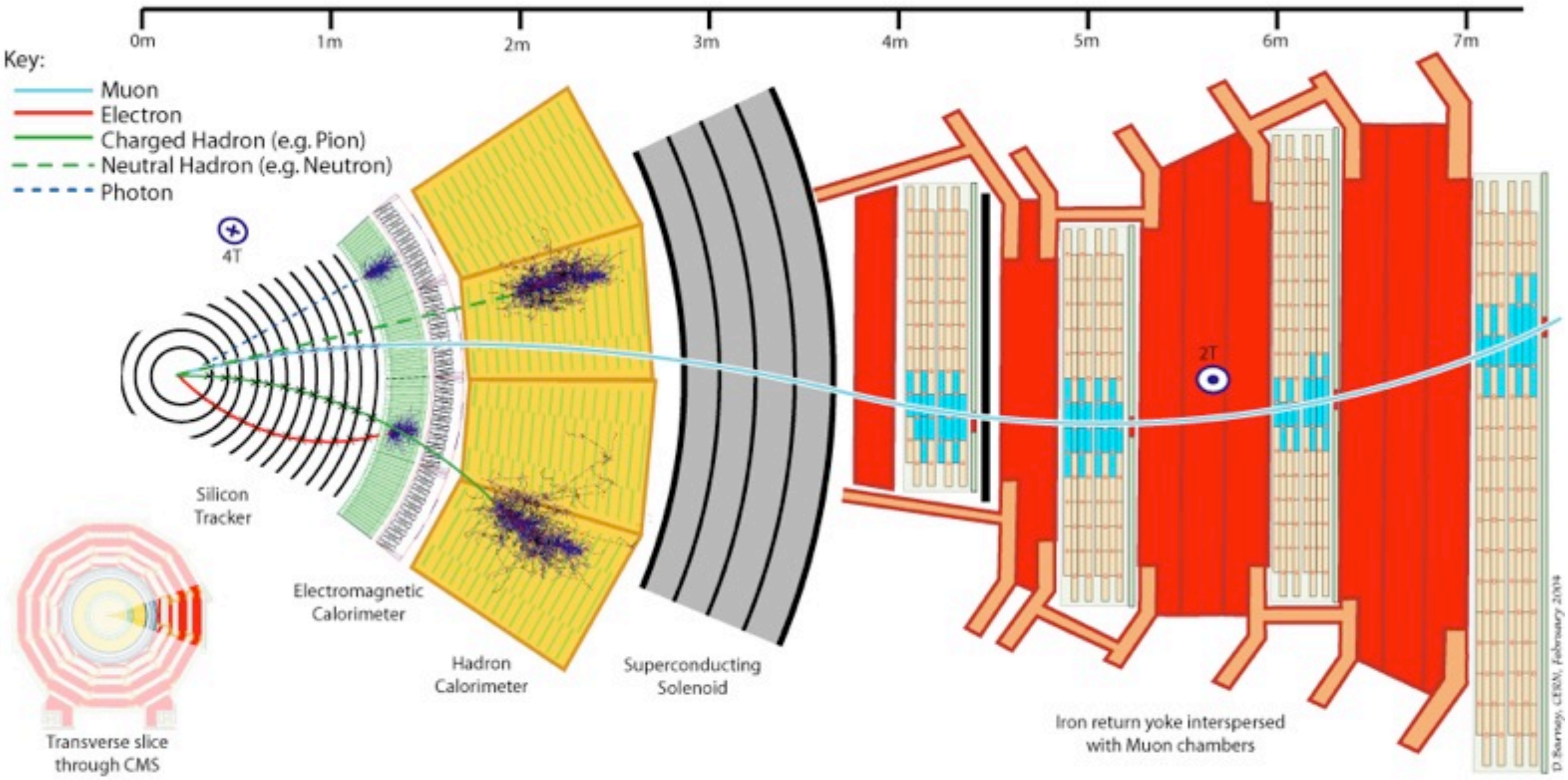


- hard scattering
- (QED) initial/final state radiation
- partonic decays, e.g. $t \to bW$
- parton shower evolution
- nonperturbative gluon splitting
- colour singlets
- colourless clusters
- cluster fission
- cluster $\to$ hadrons
- hadronic decays

The simulation narrative

Center for Cosmology and Particle Physics

3) Next, the interaction of outgoing particles with the detector is simulated.
Detailed simulations of particle interactions with matter.
Accept/reject style Monte Carlo integration of very complicated function
P(detector readout | initial particles)



A "number counting" model

From the many, many collision events, we impose some criteria to select n candidate signal events. We hypothesize that it is composed of some number of signal and background events.

$$\mathrm{Pois}(n|s+b)$$

The number of events that we expect from a given interaction process is given as a product of

- L : a time-integrated luminosity (units 1/cm^2) that serves as a measure of the amount of data that we have collected or the number of trials we have had to produce signal events
- σ : "cross-section" (units cm^2) a quantity that can be calculated from theory
- ε : fraction of signal events satisfying selection (efficiency and acceptance)

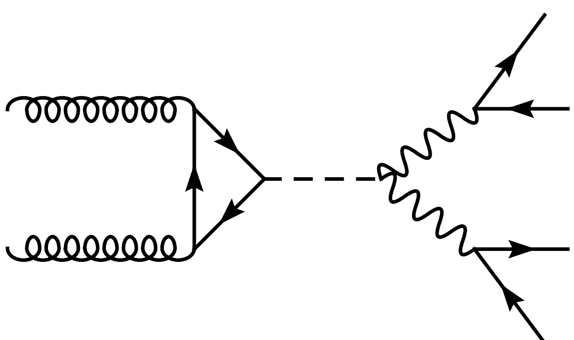
Including "shape" information

In addition to the rate of interactions, our theories predict the distributions of angles, energies, masses, etc. of particles produced

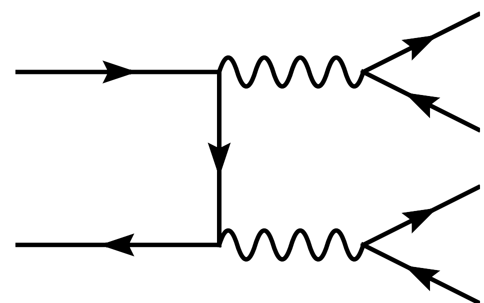
- we form functions of these called **discriminating variables** *m*,
- and use Monte Carlo techniques to estimate *f(m)*

In addition to the hypothesized signal process, there are known background processes.

- thus, the distribution of *f(m)* is a **mixture model**
- the full model is a **marked Poisson process**



signal process



background process

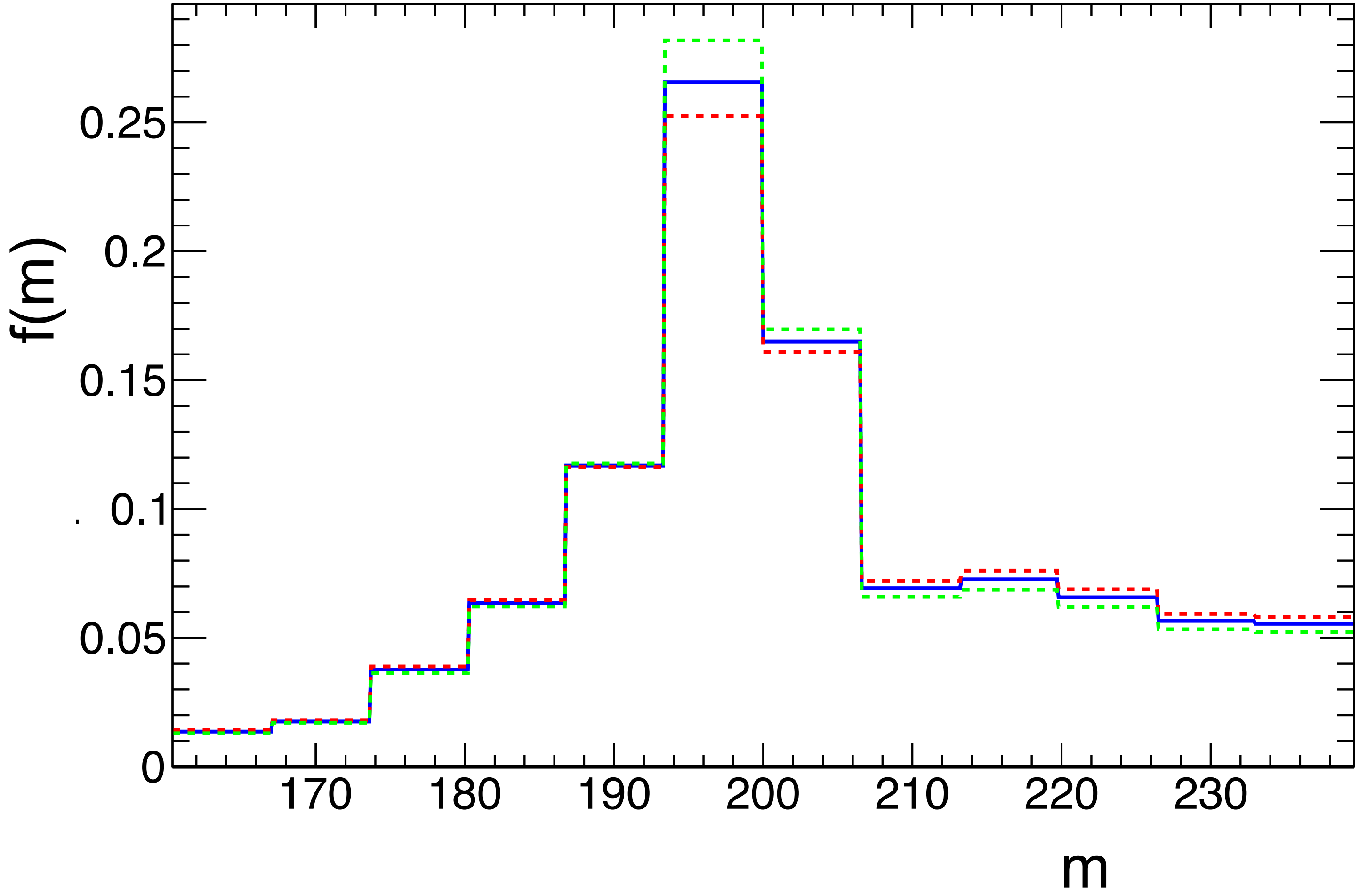
$$P(\mathbf{m}|s) = \mathrm{Pois}(n|s+b) \prod_{j}^{n} \frac{s f_s(m_j) + b f_b(m_j)}{s+b}$$

Incorporating Systematic Effects

Center for Cosmology and Particle Physics

Of course, the simulation has many adjustable parameters and imperfections that lead to systematic uncertainties.

- one can re-run simulation with different settings and produce **variational histograms** about the **nominal prediction**



Explicit parametrization

Important to distinguish between the **source** of the systematic uncertainty (eg. jet energy scale) and its **effect.**

- The same 5% jet energy scale uncertainty will have different effect on different signal and background processes
  - not necessarily with any obvious functional form
- Usually possible to decompose to independent “uncorrelated” sources

Imagine a table that **explicitly quantifies** the effect of each source of systematic.

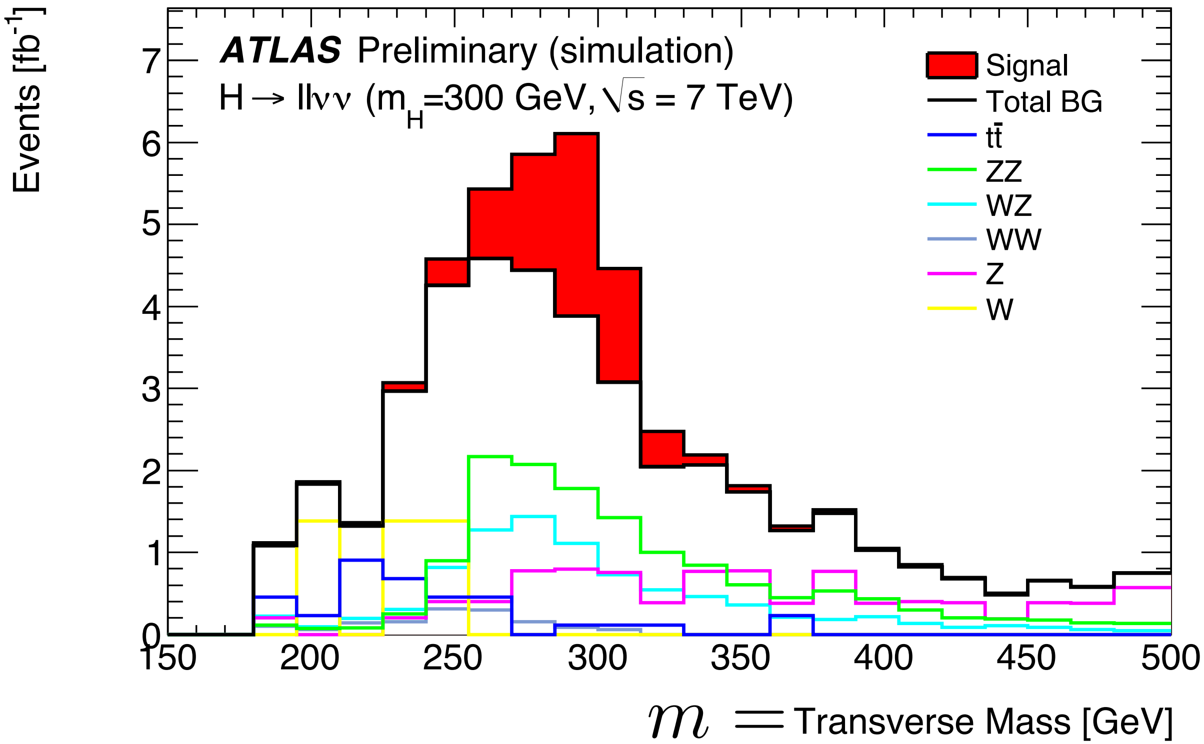
- Entries are either normalization factors or variational histograms

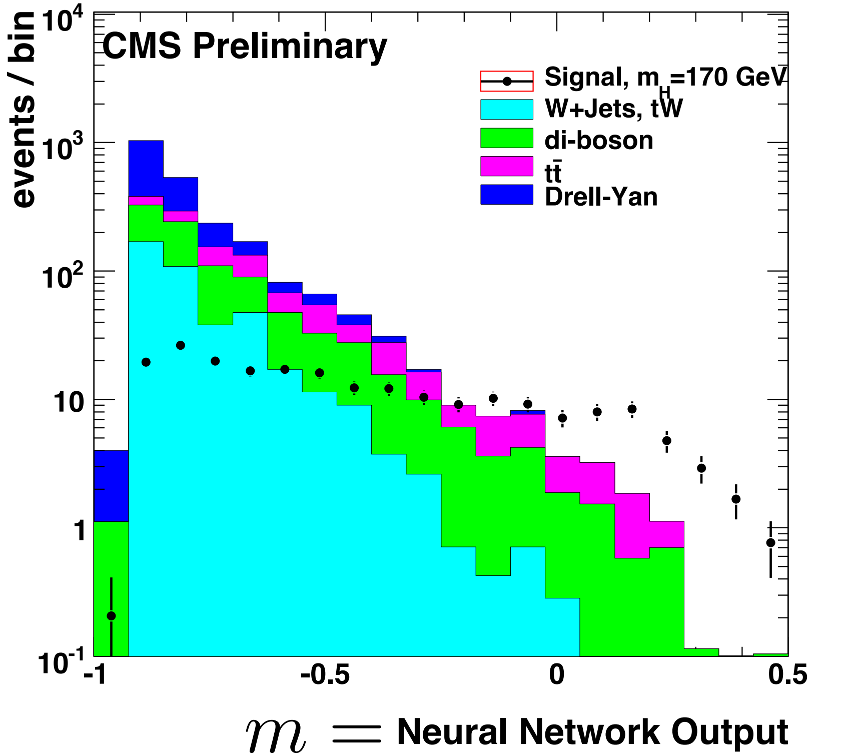
	sig	bkg 1	bkg 2	...
syst 1				
syst 2				
...				

CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

Here is an example prediction from search for H→ZZ and H→WW

- sometimes multivariate techniques are used





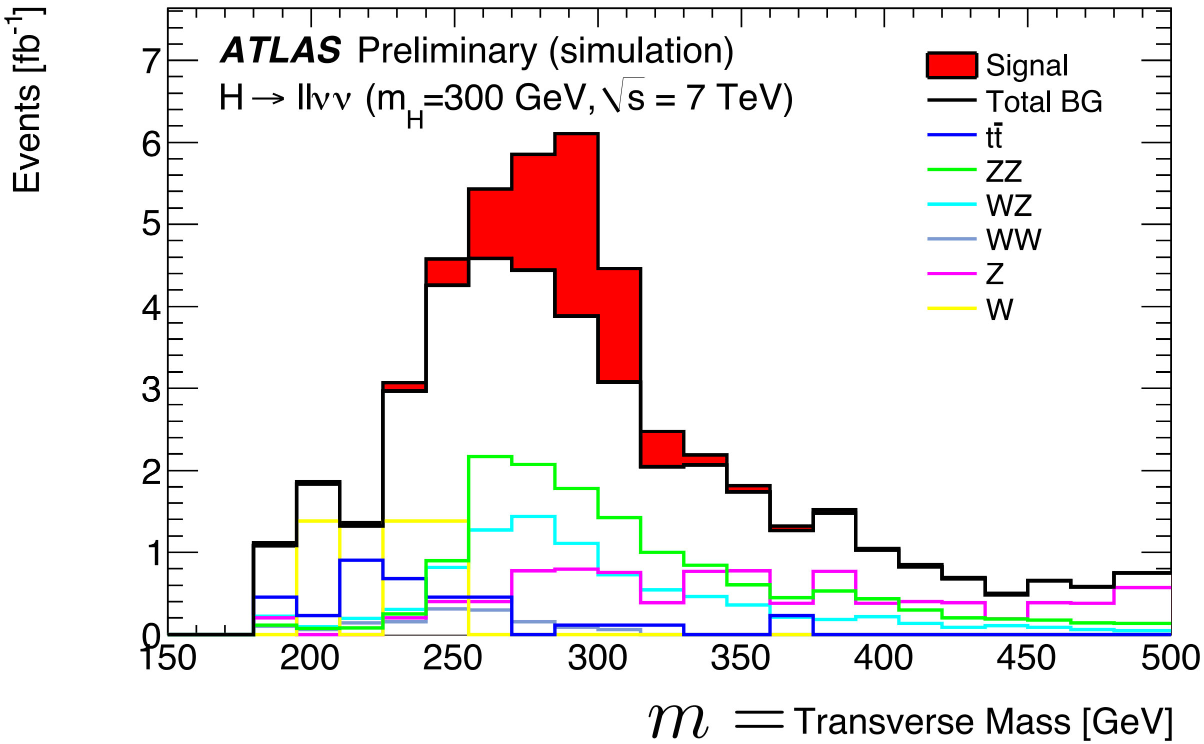
$$P(\mathbf{m}|s) = \mathrm{Pois}(n|s+b) \prod_{j}^{n} \frac{s f_s(m_j) + b f_b(m_j)}{s+b}$$

Simulation narrative overview

Center for
Cosmology and
Particle Physics

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of **nuisance parameter** α_i



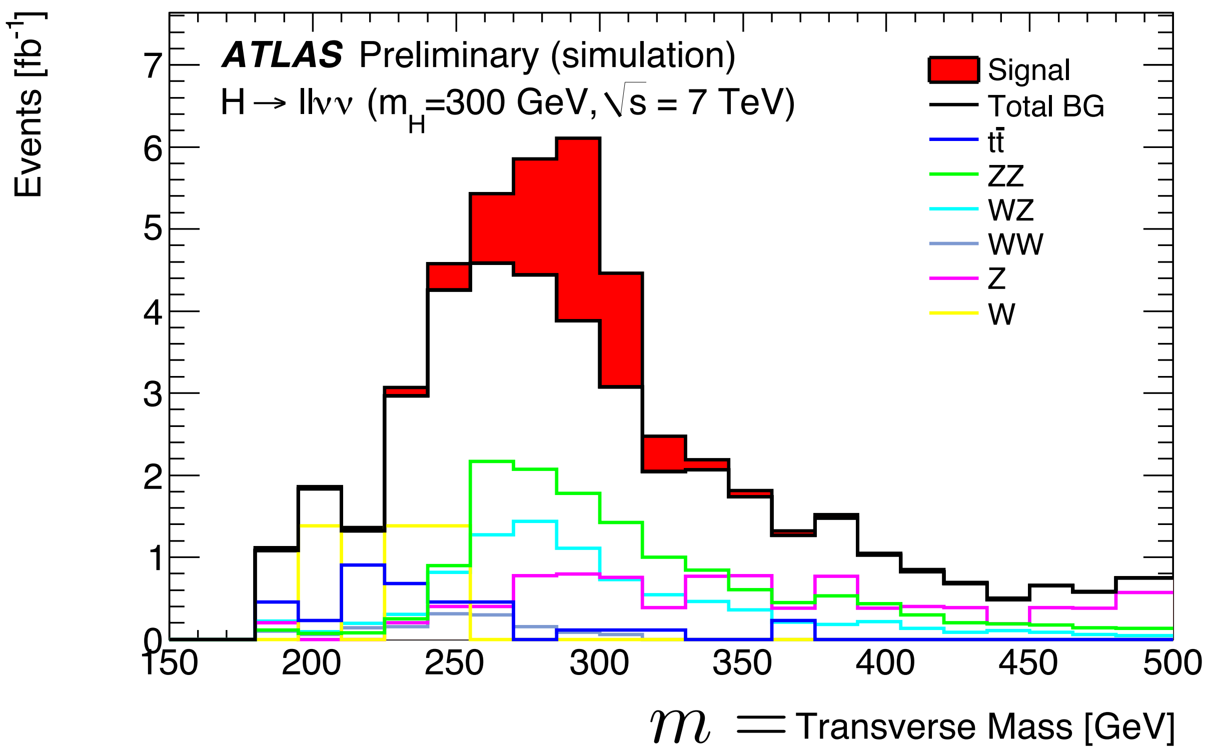
	sig	bkg 1	bkg 2	...
syst 1				
syst 2				
...				

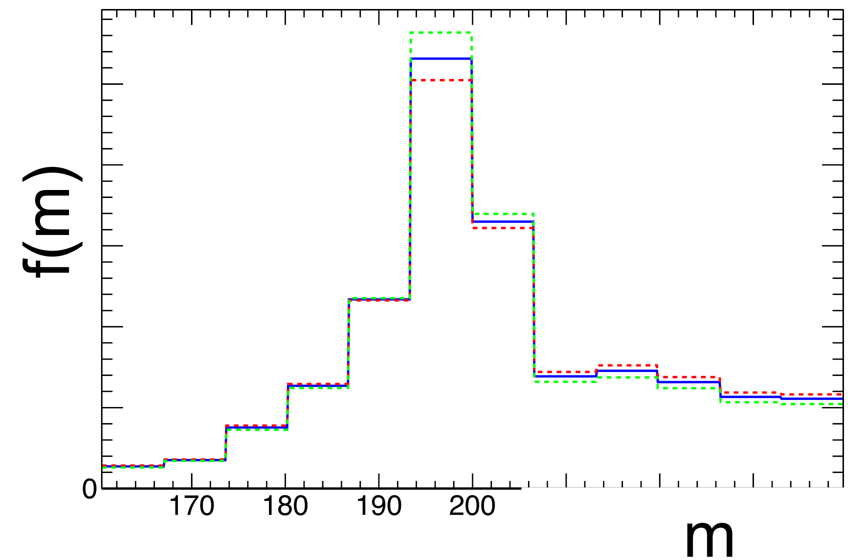
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \text{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

Simulation narrative overview

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of **nuisance parameter** α_i



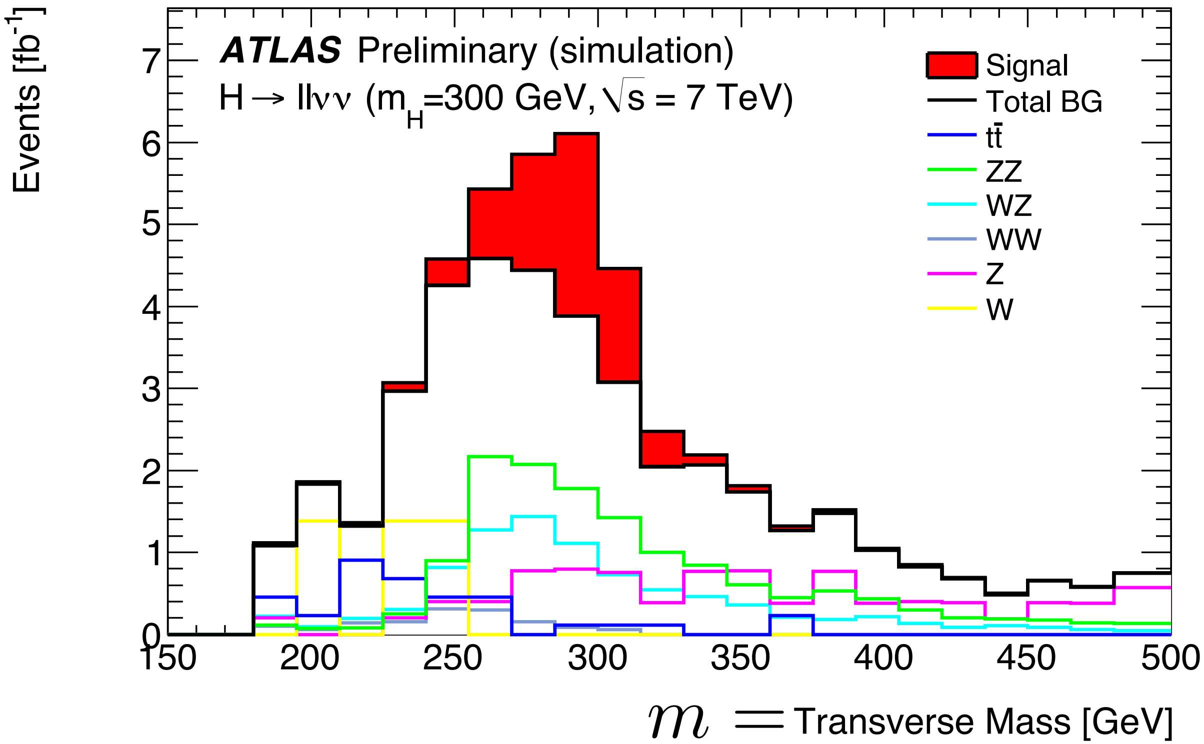


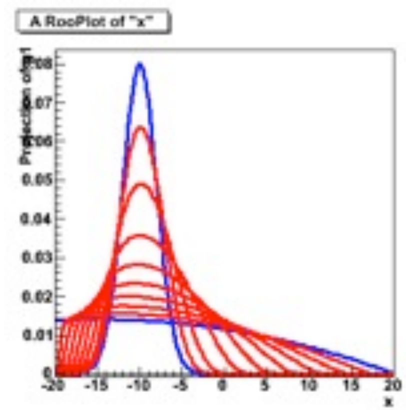
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha}) f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha}) f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})}$$

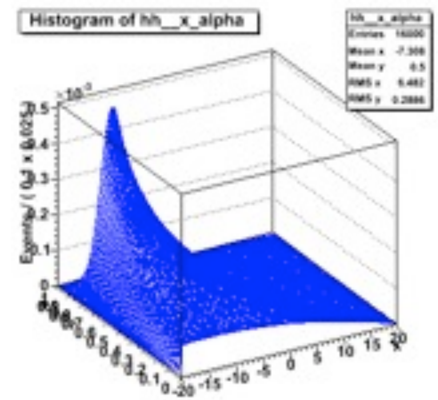
Simulation narrative overview

Tabulate effect of individual variations of sources of systematic uncertainty

- use some form of interpolation to parametrize i^{th} variation in terms of **nuisance parameter** α_i







$$P(\mathbf{m}|\boldsymbol{\alpha}) = \text{Pois}(n|s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha})f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})}$$

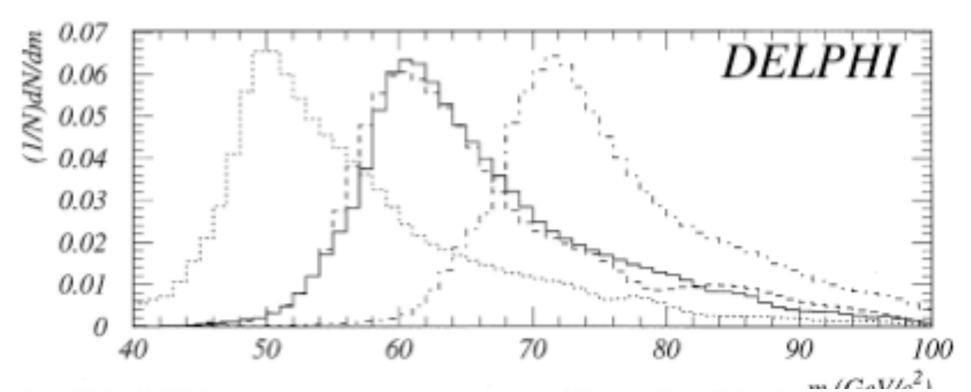
Histogram Interpolation

Center for
Cosmology and
Particle Physics

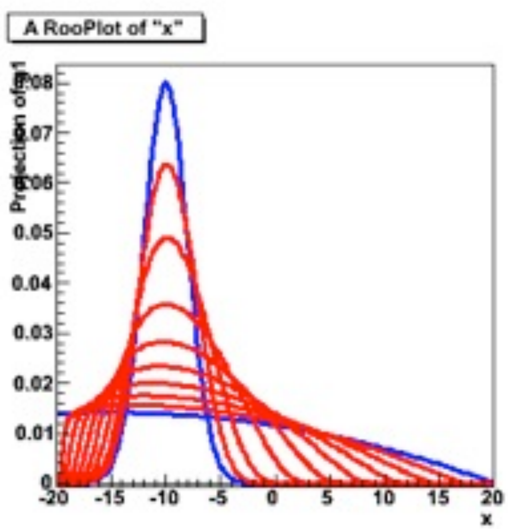
Several interpolation algorithms exist: eg. Alex Read's "horizontal" histogram interpolation algorithm (RooIntegralMorph in RooFit)

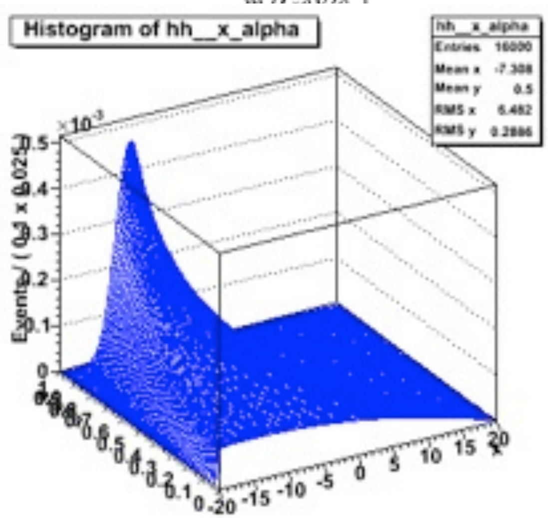
- take several PDFs, construct interpolated PDF with additional **nuisance parameter** α

A.L. Read / Nuclear Instruments and Methods in Physics Research A 425 (1999) 357–360



Simple "vertical" interpolation bin-by-bin.





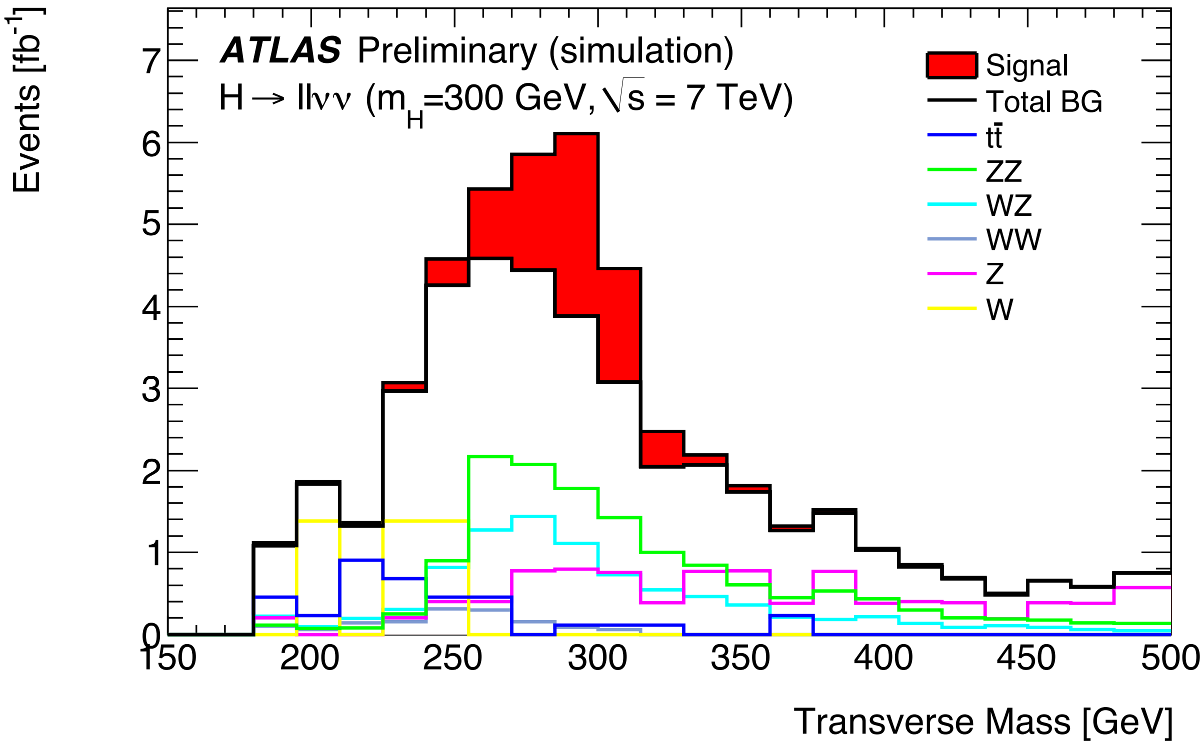
Alternative "horizontal" interpolation algorithm by Max Baak called "RooMomentMorph" in RooFit (faster and numerically more stable)

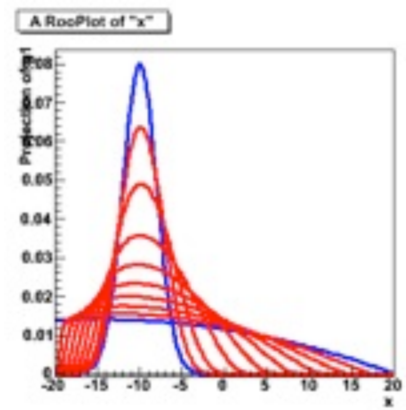
Simulation narrative overview

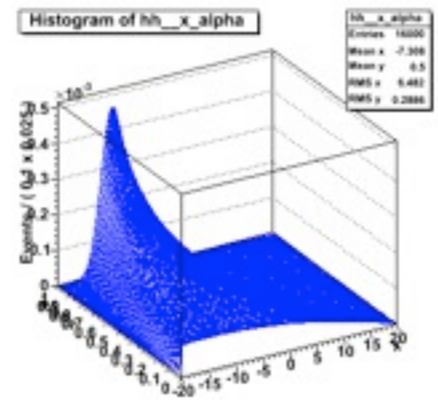
CENTER FOR COSMOLOGY AND PARTICLE PHYSICS

Something must 'constrain' the nuisance parameters α

- the data itself: sidebands; some control region
- "**constraint terms**" are added to the model... this part is subtle.







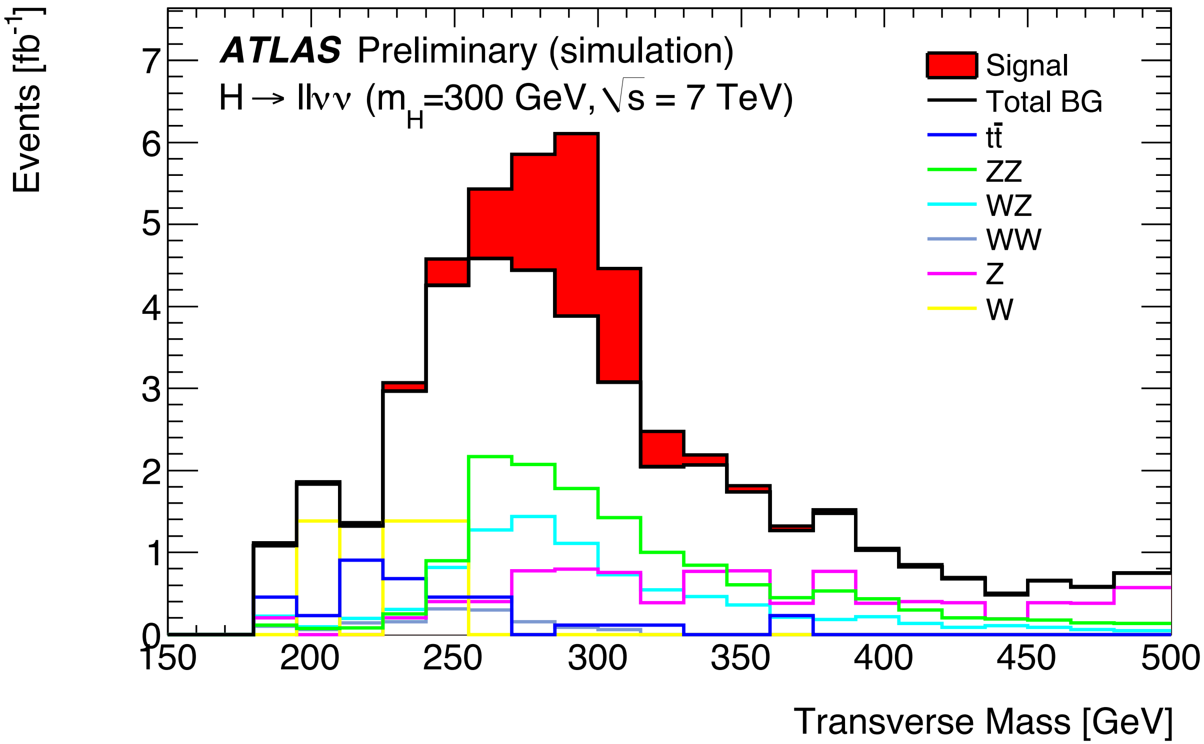
$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha}))\prod_{j}^{n}\frac{s(\boldsymbol{\alpha})f_s(m_j|\boldsymbol{\alpha})+b(\boldsymbol{\alpha})f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})}$$

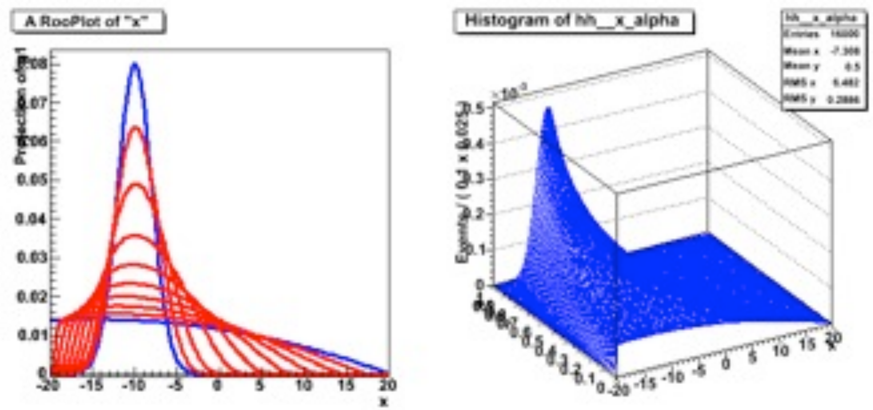
Simulation narrative overview

Center for Cosmology and Particle Physics

Something must 'constrain' the nuisance parameters α

- the data itself: sidebands; some control region
- "**constraint terms**" are added to the model... this part is subtle.





$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})) \prod_j^n \frac{s(\boldsymbol{\alpha})f_s(m_j|\boldsymbol{\alpha}) + b(\boldsymbol{\alpha})f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})}$$
$$\times G(a|\alpha,\sigma)$$

Constraint Terms
Auxiliary Measurements and Priors

What do we mean by uncertainty?

Let's consider a simplified problem that has been studied quite a bit to gain some insight into our more realistic and difficult problems

- number counting with background uncertainty
  - in our main measurement we observe n_{on} with $s+b$ expected

$$\mathrm{Pois}(n_{\mathrm{on}}|s+b)$$

- and the background has some uncertainty
  - but what is "background uncertainty"? Where did it come from?
  - maybe we would say background is known to 10% or that it has some pdf $\pi(b)$
    - then we often do a **smearing** of the background:

$$P(n_{\mathrm{on}}|s) = \int db\, \mathrm{Pois}(n_{\mathrm{on}}|s+b)\, \pi(b),$$

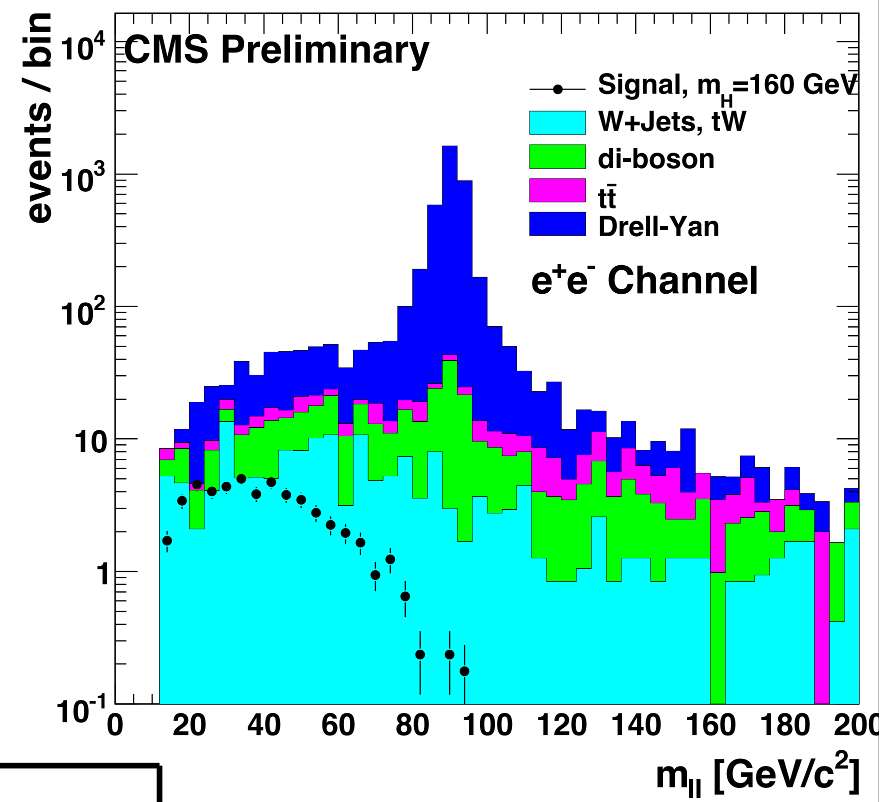
  - Where does $\pi(b)$ come from?
    - did you realize that this is a Bayesian procedure that depends on some prior assumption about what b is?

The Data-driven narrative

Center for Cosmology and Particle Physics

Regions in the data with negligible signal expected are used as control samples

- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties



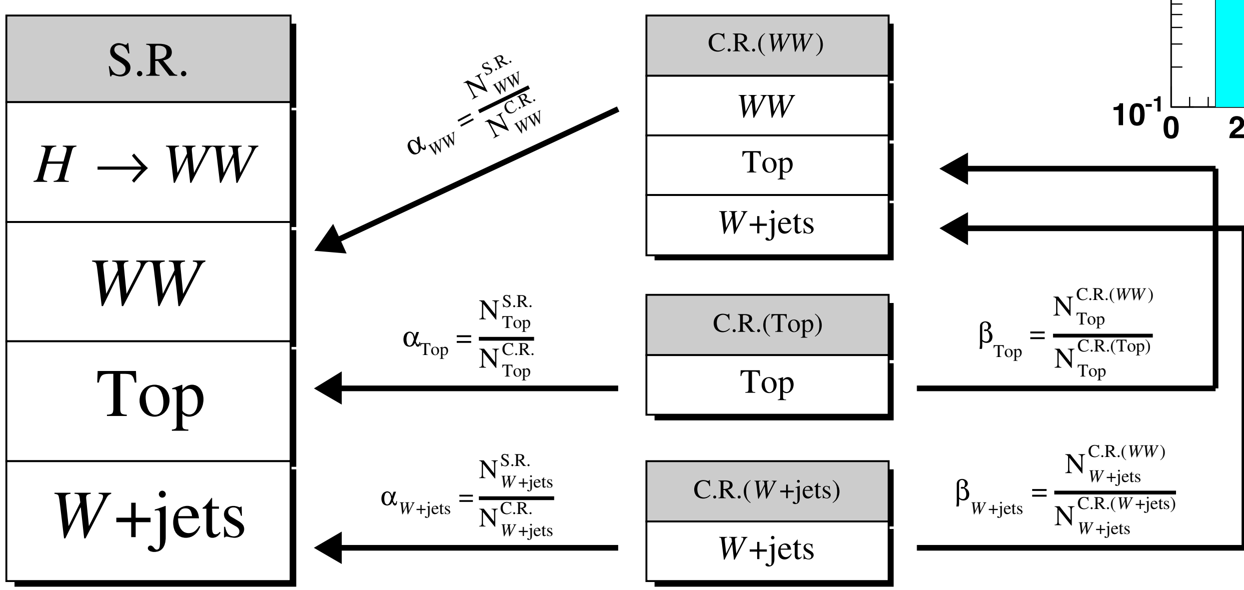
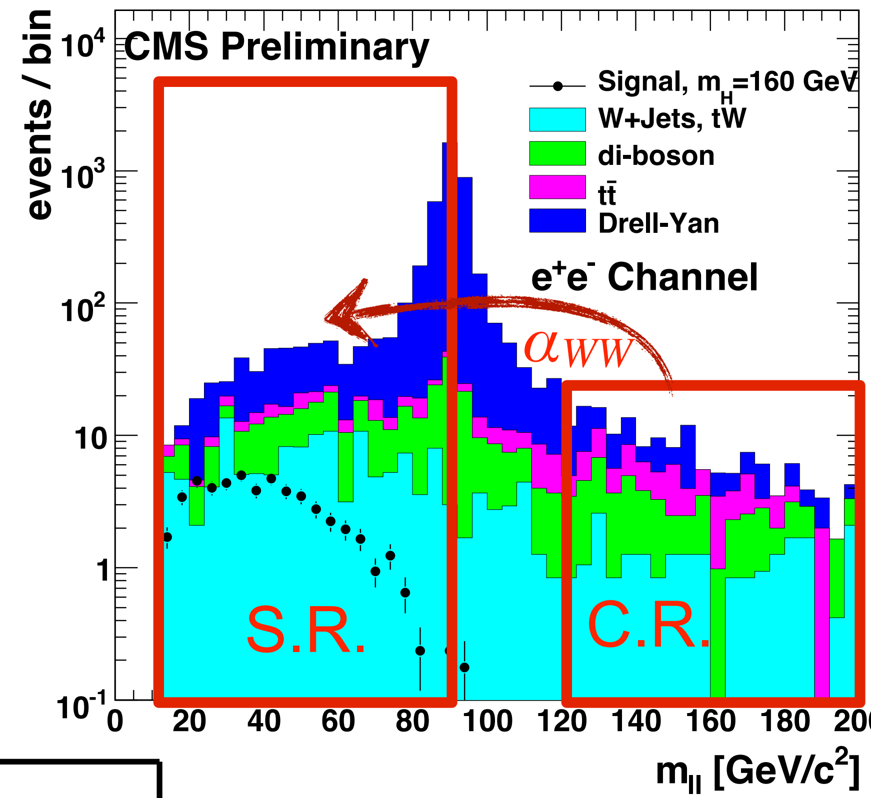


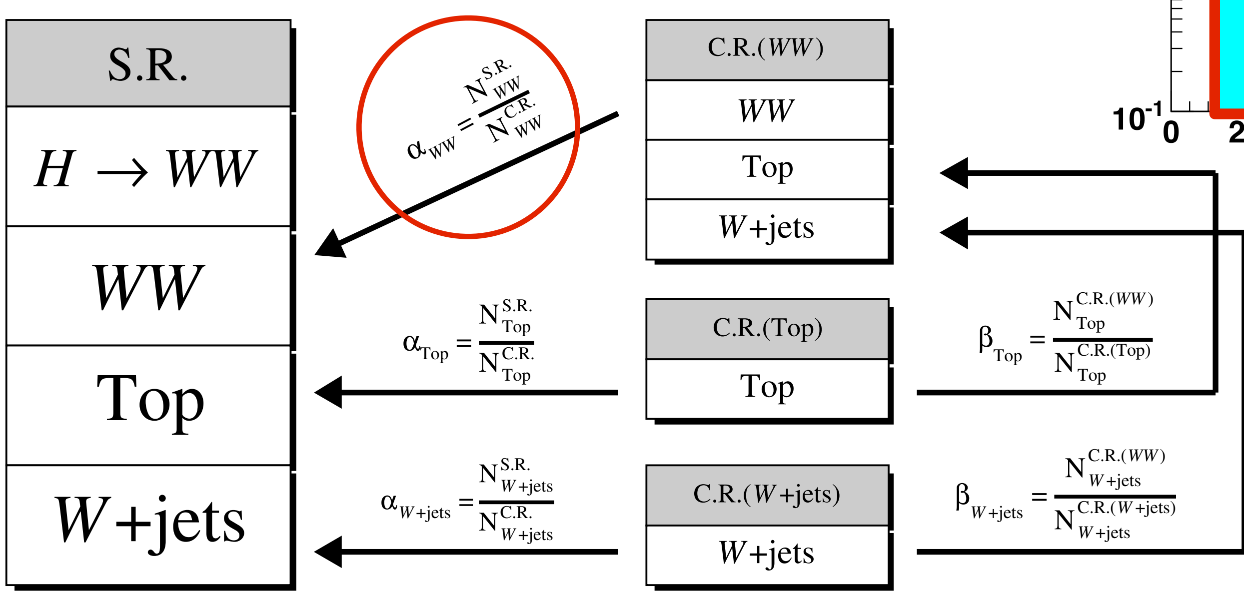
Figure 10: Flow chart describing the four data samples used in the $H \to WW^{(*)} \to \ell\nu\ell\nu$ analysis. S.R and C.R. stand for signal and control regions, respectively.

The Data-driven narrative

Regions in the data with negligible signal expected are used as control samples

- simulated events are used to estimate extrapolation coefficients
- extrapolation coefficients may have theoretical and experimental uncertainties





Notation for next slides:

\# in S.R. $\to n_{\text{on}}$

\# in C.R. $\to n_{\text{off}}$

$\alpha_{WW} \to \tau$

Figure 10: Flow chart describing the four data samples used in the $H \to WW^{(*)} \to \ell\nu\ell\nu$ analysis. S.R and C.R. stand for signal and control regions, respectively.

The "on/off" problem

Now let's say that the background was estimated from some control region or sideband measurement.

- We can treat these two measurements simultaneously:
  - main measurement: observe n_{on} with $s+b$ expected
  - sideband measurement: observe n_{off} with τb expected

$$\underbrace{P(n_{\rm on}, n_{\rm off}|s,b)}_{\text{joint model}} = \underbrace{\text{Pois}(n_{\rm on}|s+b)}_{\text{main measurement}}\ \underbrace{\text{Pois}(n_{\rm off}|\tau b)}_{\text{sideband}}$$

  - In this approach "background uncertainty" is a statistical error
  - justification and accounting of background uncertainty is much more clear

How does this relate to the smearing approach?

$$P(n_{\rm on}|s) = \int db\, \text{Pois}(n_{\rm on}|s+b)\, \pi(b),$$

- while $\pi(b)$ is based on data, it still depends on some original prior $\eta(b)$

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Separating the prior from the objective model

Recommendation: where possible, one should express uncertainty on a parameter as a statistical (random) process

- explicitly include terms that represent auxiliary measurements in the likelihood

Recommendation: when using a Bayesian technique, one should explicitly express and separate the prior from the objective part of the probability density function

Example:

- By writing $P(n_{\rm on}, n_{\rm off}|s,b) = {\rm Pois}(n_{\rm on}|s+b)\,{\rm Pois}(n_{\rm off}|\tau b).$
  - the objective statistical model is for the background uncertainty is clear
- One can then explicitly express a prior $\eta(b)$ and obtain:

$$\pi(b) = P(b|n_{\rm off}) = \frac{P(n_{\rm off}|b)\eta(b)}{\int db P(n_{\rm off}|b)\eta(b)}.$$

Constraint terms for our example model

For each systematic effect, we associated a nuisance parameter α

- for instance electron efficiency, JES, luminosity, etc.
- the background rates, signal acceptance, etc. are parametrized in terms of these nuisance parameters

These systematics are usually known ("constrained") within ± 1σ.

- but here we must be careful about Bayesian vs. frequentist
- Why is it constrained? Usually b/c we have an **auxiliary measurement** *a* and a relationship like:

$$G(a|\alpha, \sigma)$$

- Saying that α has a Gaussian distribution is Bayesian.
  - has form "Probability of parameter"
- The frequentist way is to say that *a* fluctuates about α

While *a* is a measured quantity (or "observable"), there is only one measurement of *a* per experiment. Call it a "**Global observable**"

Common Constraints Terms

Many uncertainties have no clear statistical description or it is impractical to provide

Traditionally, we use Gaussians, but for large uncertainties it is clearly a bad choice

- quickly falling tail, bad behavior near physical boundary, optimistic p-values, ...

For systematics constrained from control samples and dominated by statistical uncertainty, a Gamma distribution is a more natural choice [PDF is Poisson for the control sample]

- longer tail, good behavior near boundary, natural choice if auxiliary is based on counting

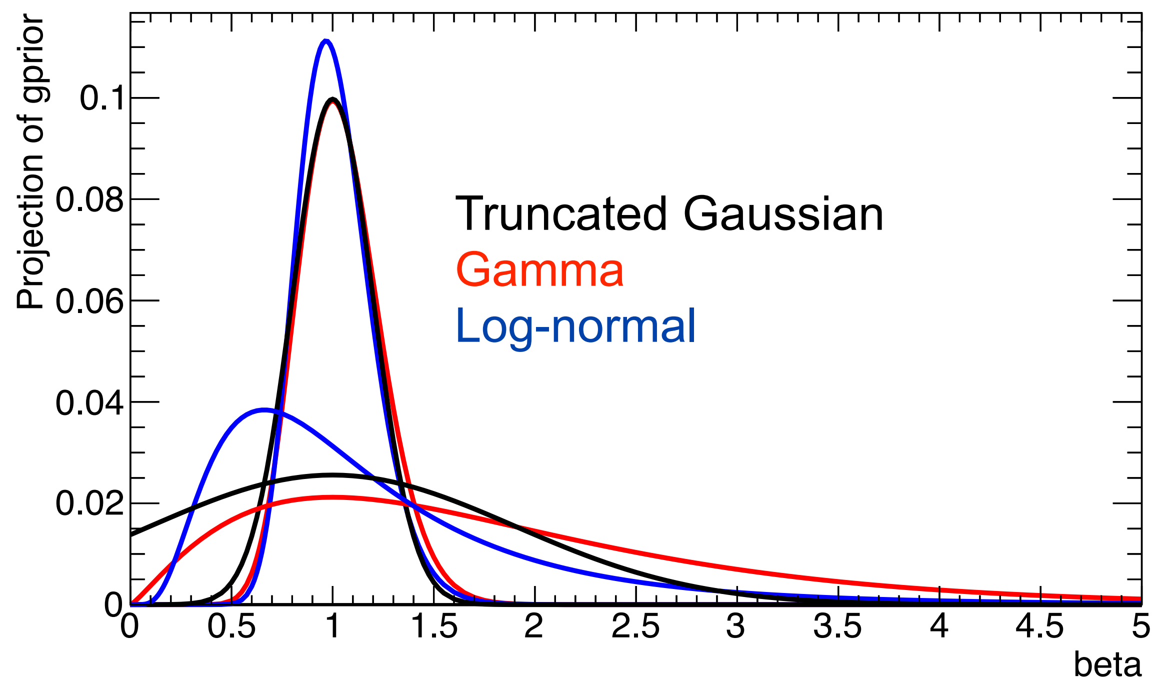
For "factor of 2" notions of uncertainty log-normal is a good choice

- can have a very long tail for large uncertainties

None of them are as good as an actual model for the auxiliary measurement, if available

To consistently switch between frequentist, Bayesian, and hybrid procedures, need to be clear about prior vs. likelihood function

PDF(y| β)	Prior(β)	Posterior(β|y)
Gaussian	uniform	Gaussian
Poisson	uniform	Gamma
Log-normal	1/β	Log-Normal



Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II - "The Bad"

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don't necessarily scale with luminosity
  - eg: "shape" systematics

Classification of Systematic Uncertainties

Taken from Pekka Sinervo's PhyStat 2003 contribution

Type I - "The Good"

- can be constrained by other sideband/auxiliary/ ancillary measurements and can be treated as statistical uncertainties
  - scale with luminosity

Type II - "The Bad"

- arise from model assumptions in the measurement or from poorly understood features in data or analysis technique
  - don't necessarily scale with luminosity
  - eg: "shape" systematics

Type III - "The Ugly"

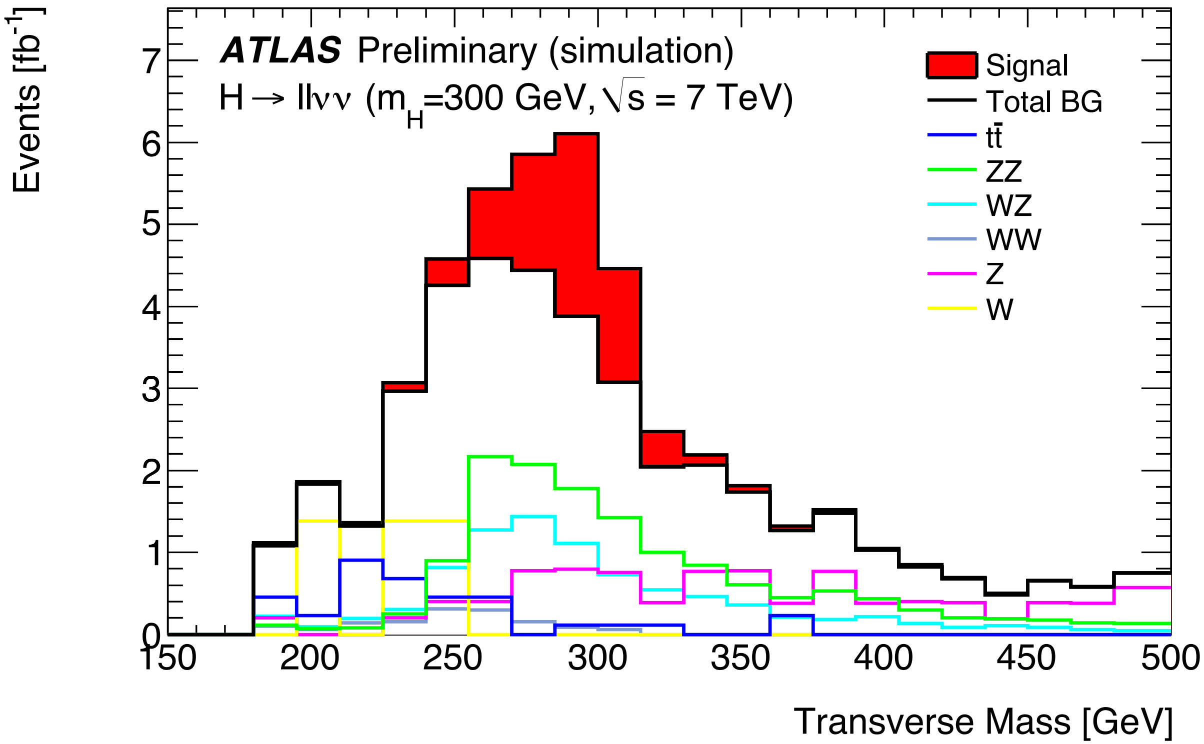
- arise from uncertainties in underlying theoretical paradigm used to make inference using the data
  - a somewhat philosophical issue

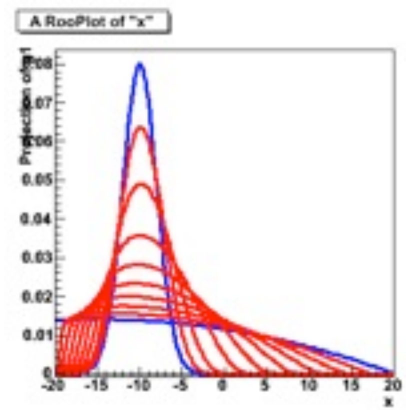
Modeling:
The Scientific Narrative
(continued)

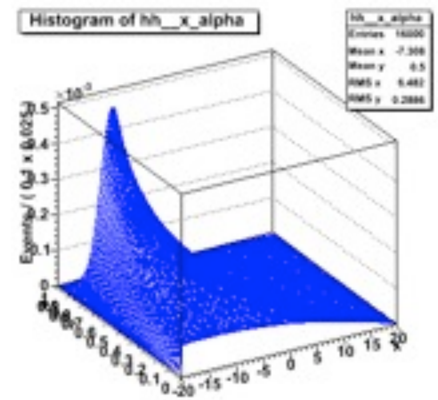
Constraint terms for our example model

Something must 'constrain' the nuisance parameters α

- the data itself: sidebands; some control region
- "**constraint terms**" are added to the model... this part is subtle.







$$P(\mathbf{m}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha}))\prod_j^n \frac{s(\boldsymbol{\alpha})f_s(m_j|\boldsymbol{\alpha})+b(\boldsymbol{\alpha})f_b(m_j|\boldsymbol{\alpha})}{s(\boldsymbol{\alpha})+b(\boldsymbol{\alpha})}$$

$$\boxed{\times\prod_i G(a_i|\alpha_i,\sigma_i)}$$

Building the model: HistFactory (RooStats)

Center for Cosmology and Particle Physics

Several analyses have used the tool called **hist2workspace** to build the model (PDF)

- command line: **hist2workspace myAnalysis.xml**
- construct likelihood function below via XML + histograms

$$\mathscr{L}(\mu, \alpha_i) = \prod_{m \in \text{bins}} \text{Pois}(n_m | \nu_m) \prod_{i = \in \text{Syst}} N(\alpha_i)$$

$$\nu_m = \mu L \eta_1(\alpha)\, \sigma_{1m}(\alpha) + \sum_{j \in \text{Bkg Samp}} L \eta_j(\alpha)\, \sigma_{jm}(\alpha),$$

interpolation convention

$$\eta_j(\alpha) = \prod_{i \in \text{Syst}} I(\alpha_i; \eta_{ij}^+, \eta_{ij}^-)$$

$$\sigma_{jm}(\alpha) = \sigma_{jm}^0 \prod_{i \in \text{Syst}} I(\alpha_i; \sigma_{ijm}^+/\sigma_{jm}^0, \sigma_{ijm}^-/\sigma_{jm}^0)$$

$$I(\alpha; I^+, I^-) = \begin{cases} 1 + \alpha(I^+ - 1) & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ 1 - \alpha(I^- - 1) & \text{if } \alpha < 0 \end{cases}$$

```xml
<!DOCTYPE Channel  SYSTEM 'Config.dtd'>

 <Channel Name="channel1" InputFile="./data/example.root" HistoName="" >
  <!--<Data Name="data" InputFile="" HistoPath="" HistoName=""/>-->
  <Sample Name="signal" HistoPath="" HistoName="signal">
    <OverallSys Name="syst1" High="1.05" Low="0.95"/>
    <NormFactor Name="SigXsecOverSM" Val="1" Low="0.5" High="1.8" Const="True" />
  </Sample>
  <Sample Name="background1" HistoPath="" NormalizeByTheory="True" HistoName="background1">
    <OverallSys Name="syst2" Low="0.95" High="1.05"/>
  </Sample>
  <Sample Name="background2" HistoPath="" NormalizeByTheory="True" HistoName="background2">
    <OverallSys Name="syst3" Low="0.95" High="1.05"/>
    <!-- <HistoSys Name="syst4" HistoPathHigh="" HistoPathLow="histForSyst4"/>-->
  </Sample>
 </Channel>
```

CMS Higgs example

Center for
Cosmology and
Particle Physics

The CMS input:

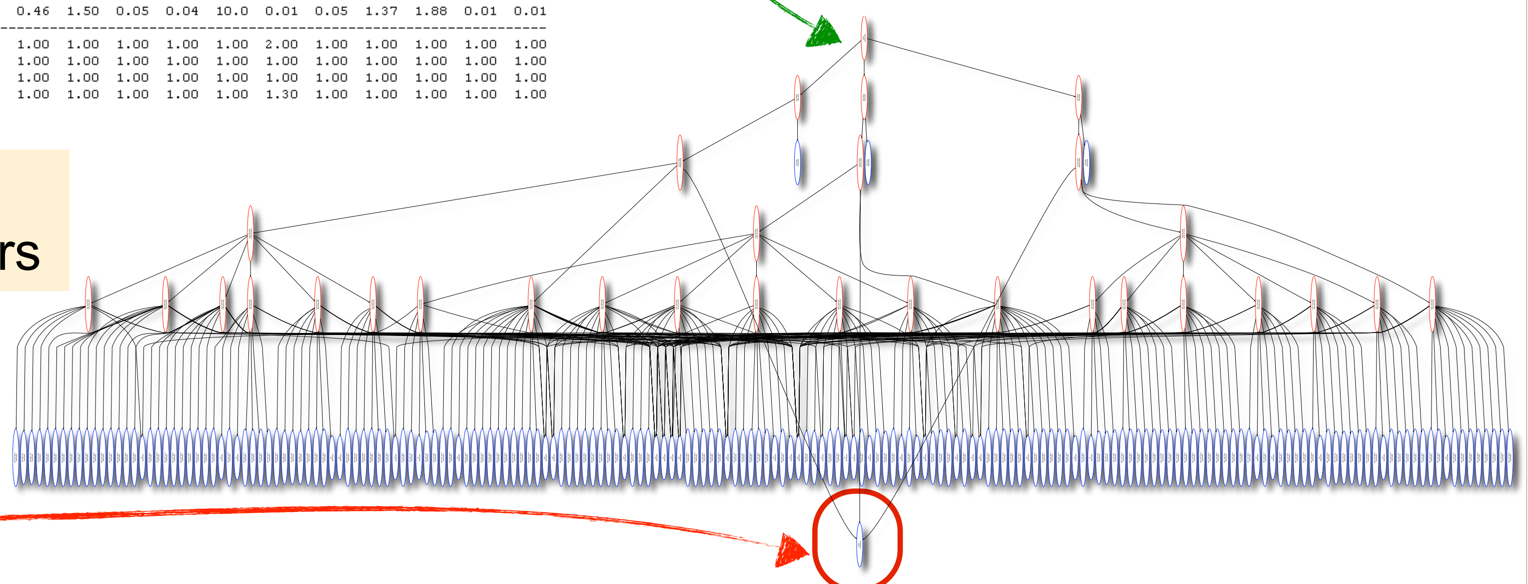
- cleanly tabulated effect on each background due to each source of systematic
- systematics broken down into uncorrelated subsets
- used lognormal distributions for all systematics, Poissons for observations

Started with a txt input, defined a mathematical representation, and then prepared the RooStats workspace

```
Date: June 22, 2010
Description: HWW-->2l2v, 0jets, cut-and-count for 3 channels: mumu, ee, emu; made-up numbers for a ATLAS+CMS combination exercise
mH    160  Higgs mass hypothesis
comE  7.0   center of mass energy
lumi  1  luminosity in fb-1
----------------------------------------------------------------------------------------------------------------------------------------
imax   3  number of channels
jmax   6  number of backgrounds
kmax  37  number of nuisance parameters
----------------------------------------------------------------------------------------------------------------------------------------
Observation   15  7  13
----------------------------------------------------------------------------------------------------------------------------------------
bin         1     1     1     1     1     1     1     2      2     2     2     2     2     2     3     3     3     3     3     3     3
process     H     Wj    Zj    tX    WW    WZ    ZZ    H      Wj    Zj    tX    WW    WZ    ZZ    H     Wj    Zj    tX    WW    WZ    ZZ
process     0     1     2     3     4     5     6     0      1     2     3     4     5     6     0     1     2     3     4     5     6
----------------------------------------------------------------------------------------------------------------------------------------
rate      10.5  0.01  0.05  0.94  3.39  0.01  0.01  5.39  0.01  0.05  0.46  1.50  0.05  0.04  10.0  0.01  0.05  1.37  1.88  0.01  0.01
----------------------------------------------------------------------------------------------------------------------------------------
 1  lnN   1.00  2.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  2.00  1.00  1.00  1.00  1.00  1.00
 2  lnN   1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  2.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
 3  lnN   1.00  1.30  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00
 4  lnN   1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.00  1.30  1.00  1.00  1.00  1.00  1.00  1.00  1.30  1.00  1.00  1.00  1.00  1.00
```

$$L_{b+rs} = \prod_i \left[\frac{\left(\sum_{j=0,1,..} \tilde{n}_{ij} \cdot \kappa_{ijk}^{\theta_k}\right)^{N_i}}{N_i!} \cdot \exp\left(-\sum_{j=0,1,..} \tilde{n}_{ij} \cdot \kappa_{ijk}^{\theta_k}\right) \right] \cdot \prod_k f(\theta_k)$$

3 observables and
37 nuisance parameters



$$n = \mu L \epsilon \sigma_{SM}$$

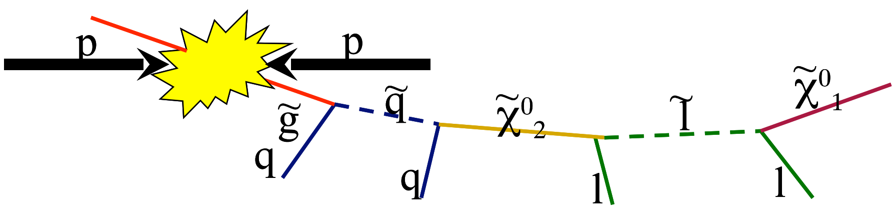
The Data-Driven narrative

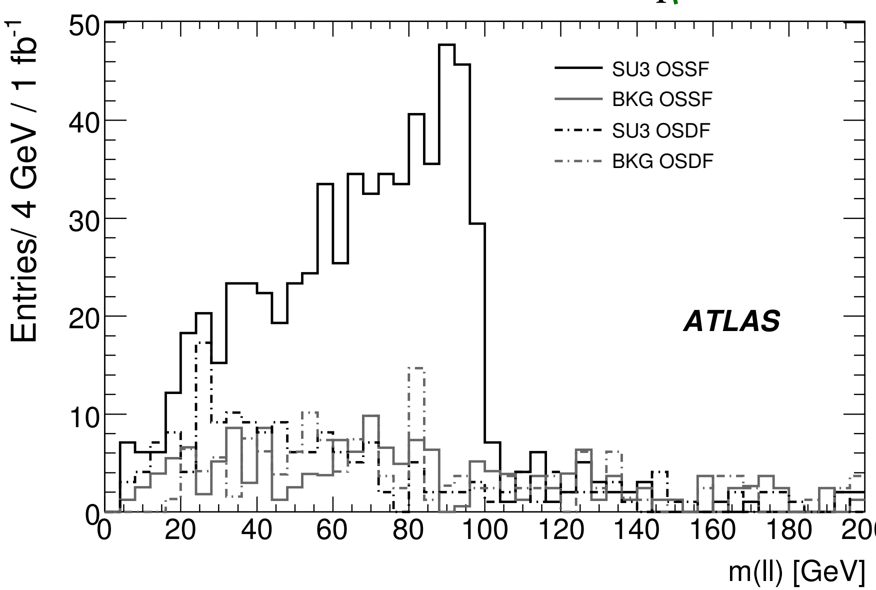
In the data-driven approach, backgrounds are estimated by assuming (and testing) some relationship between a control region and signal region

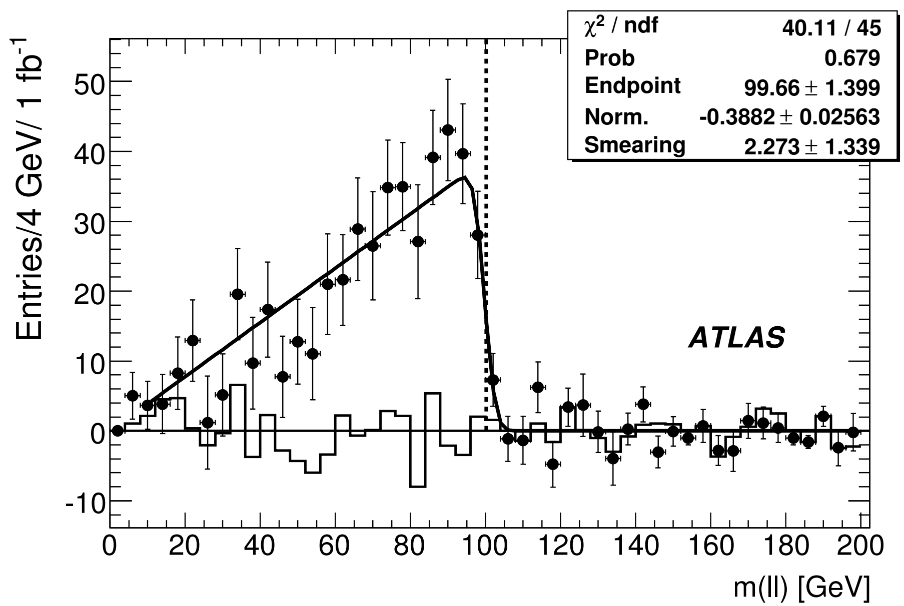
- flavor subtraction, same-sign samples, fake matrix, tag-probe, ...

Pros: Initial sample has “all orders” theory :-) and all the details of the detector

Cons: assumptions made in the transformation to the signal region can be questioned







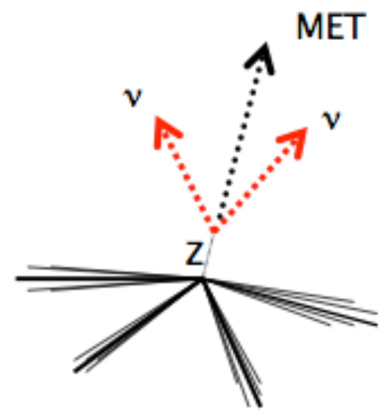
All-hadronic searches with MHT

Search for high pT jets, high HT and high MHT (= vector sum of jets)

3 jets, E_T>50 $|\eta|$<2.5

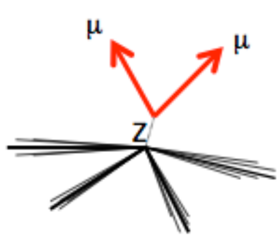
HT > 350 and MHT > 150

Event cleaning cuts.



Predict each bkgd separately

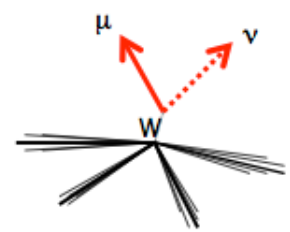
- QCD: rebalance & smear
- W & ttbar from μ control
- $Z \to \nu\nu$ from γ+jets and $Z \to \mu\mu$



$Z \to ll$ + jets

Strength: very clean

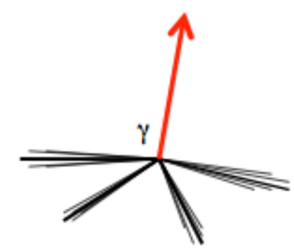
Weakness: low statistics



$W \to l\nu$ + jets

Strength: larger statistics

Weakness: background from SM and SUSY



γ + jets

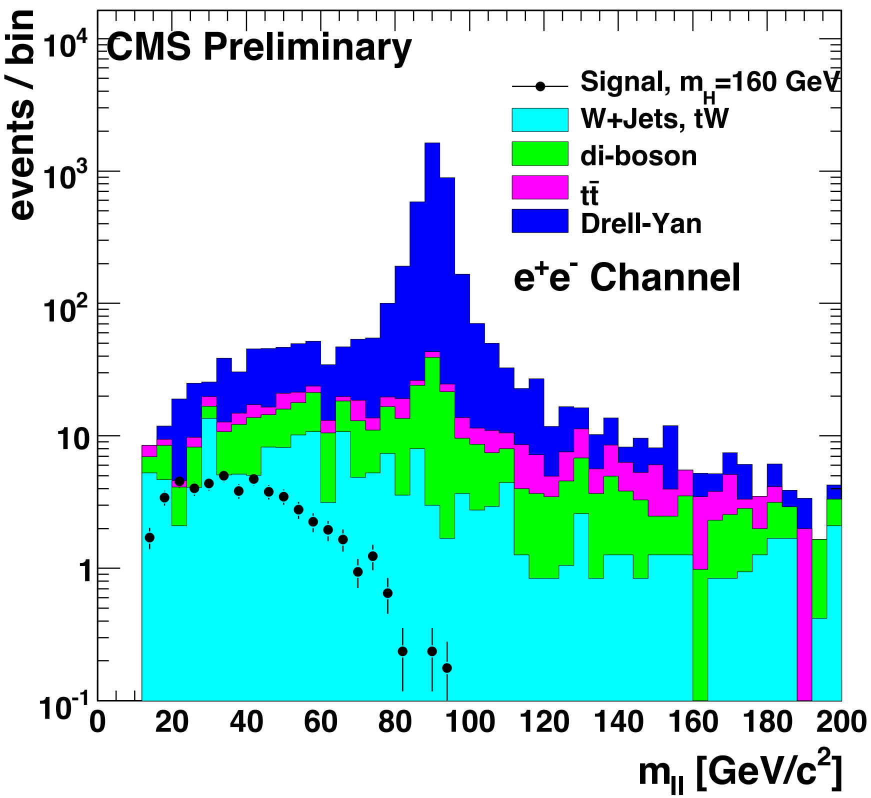
Strength: large statistics and clean at high E_T

Weakness: background at low E_T, theoretical errors

Going beyond on/off

Often the extrapolation parameter has uncertainty

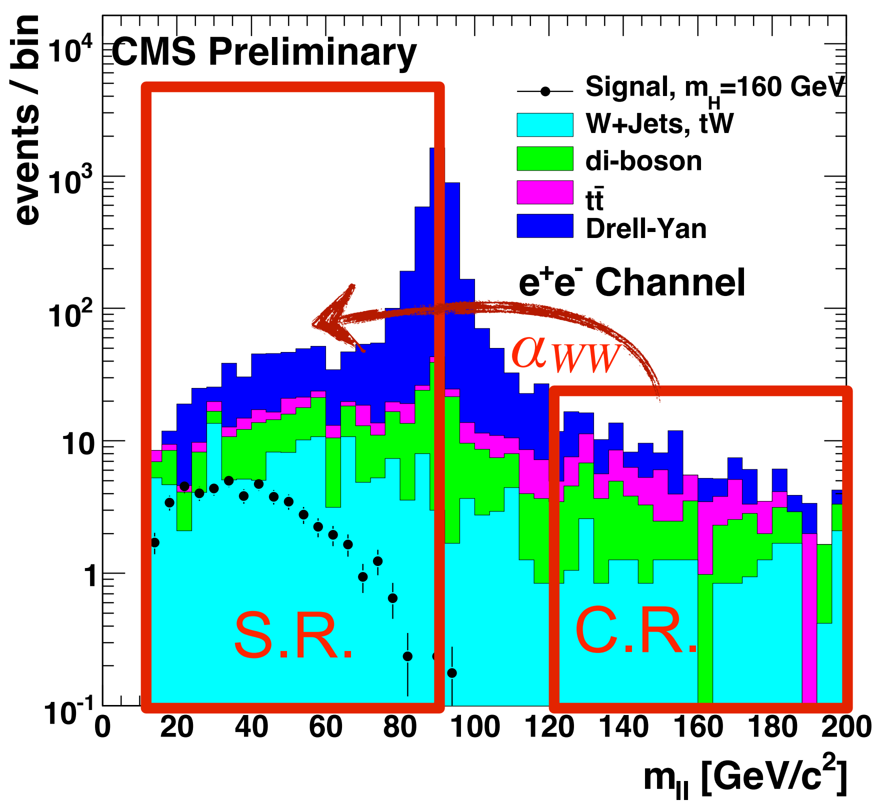
- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...



Going beyond on/off

Often the extrapolation parameter has uncertainty

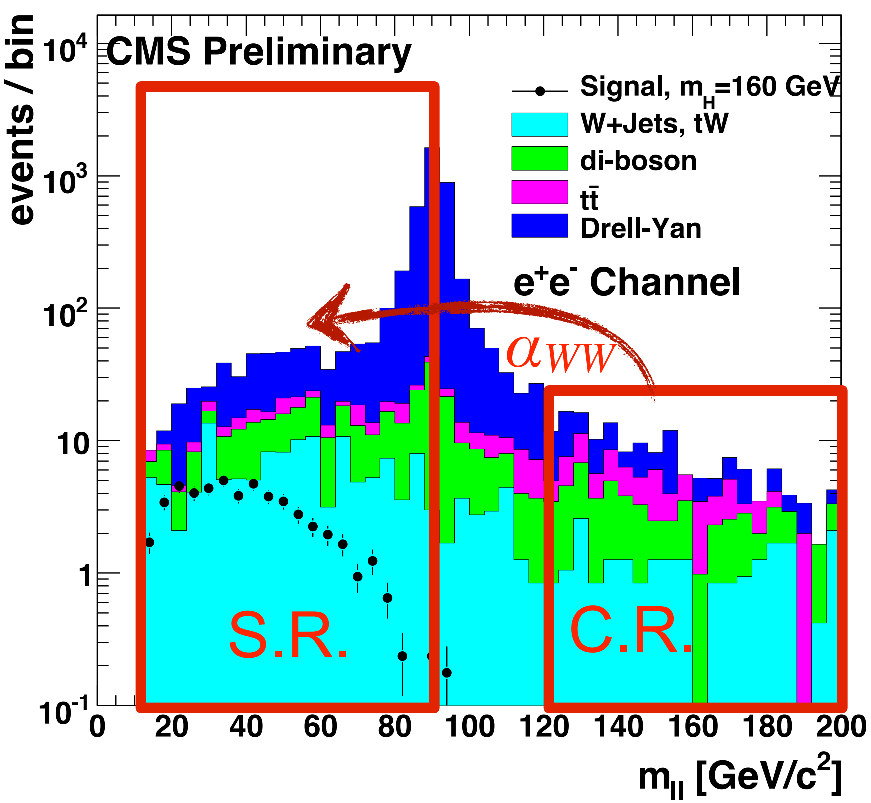
- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...



Going beyond on/off

Often the extrapolation parameter has uncertainty

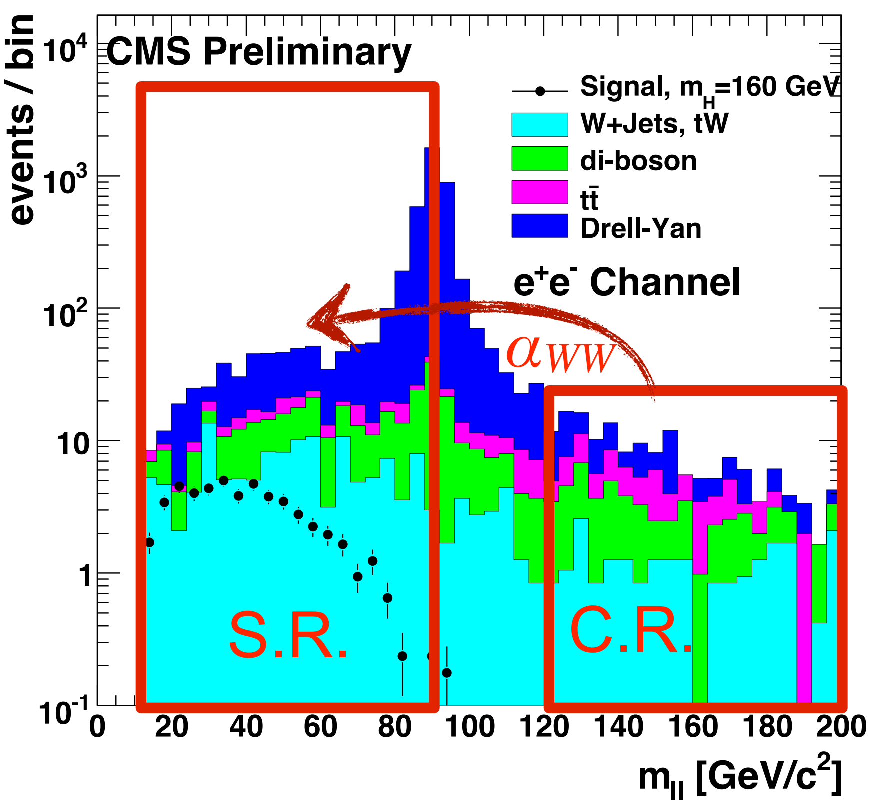
- introduce a new measurement to constrain it as in the ABCD method

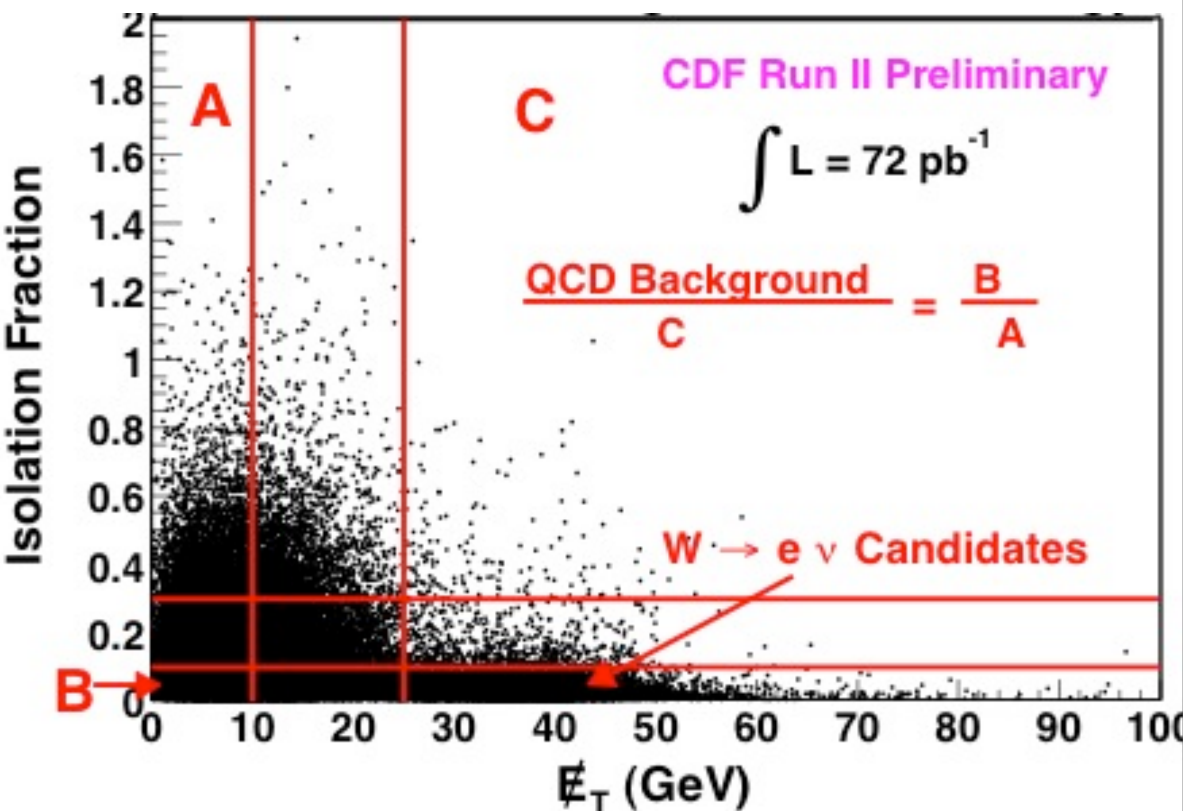


Going beyond on/off

Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method

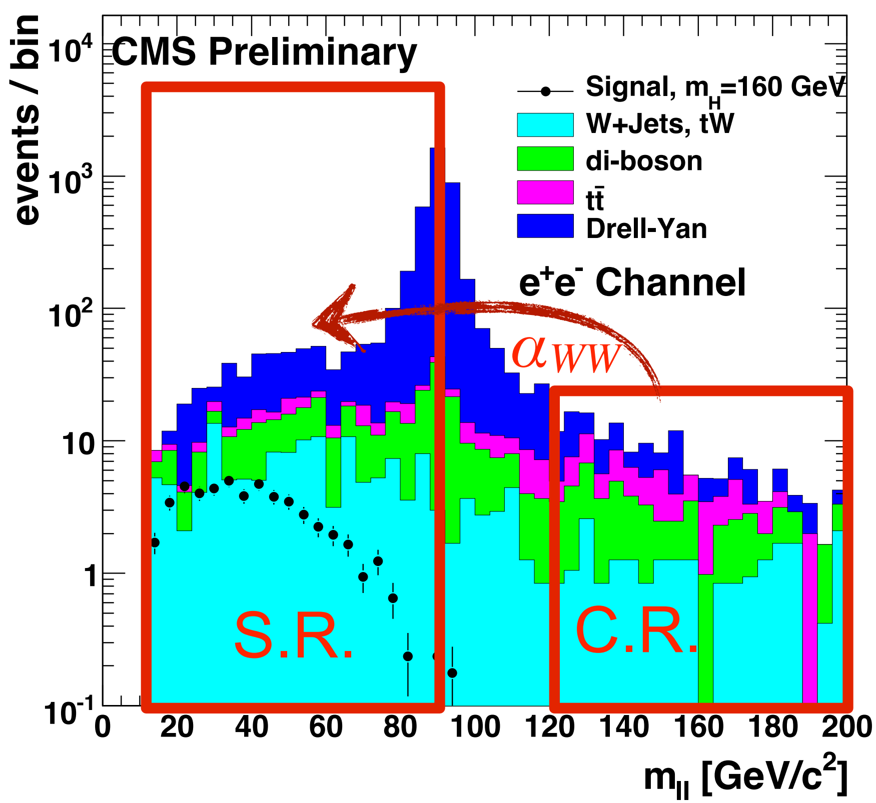


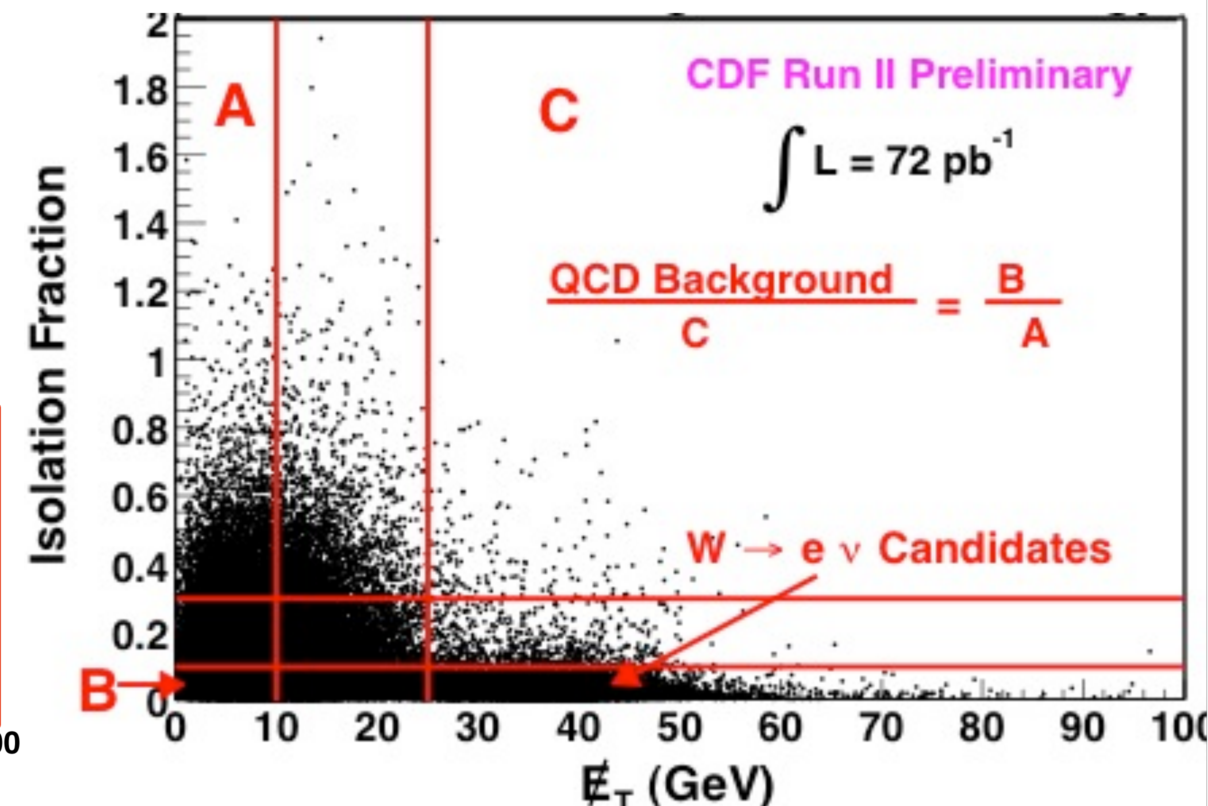


Going beyond on/off

Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...





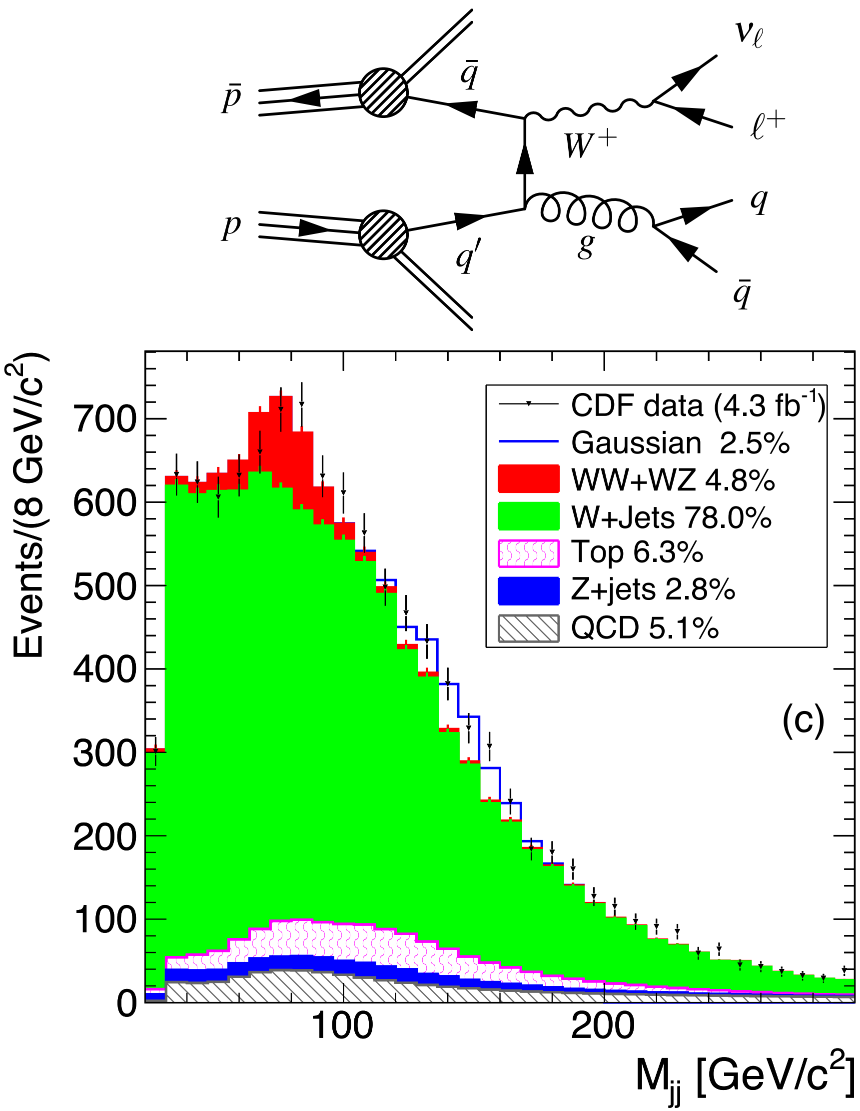
Going beyond on/off

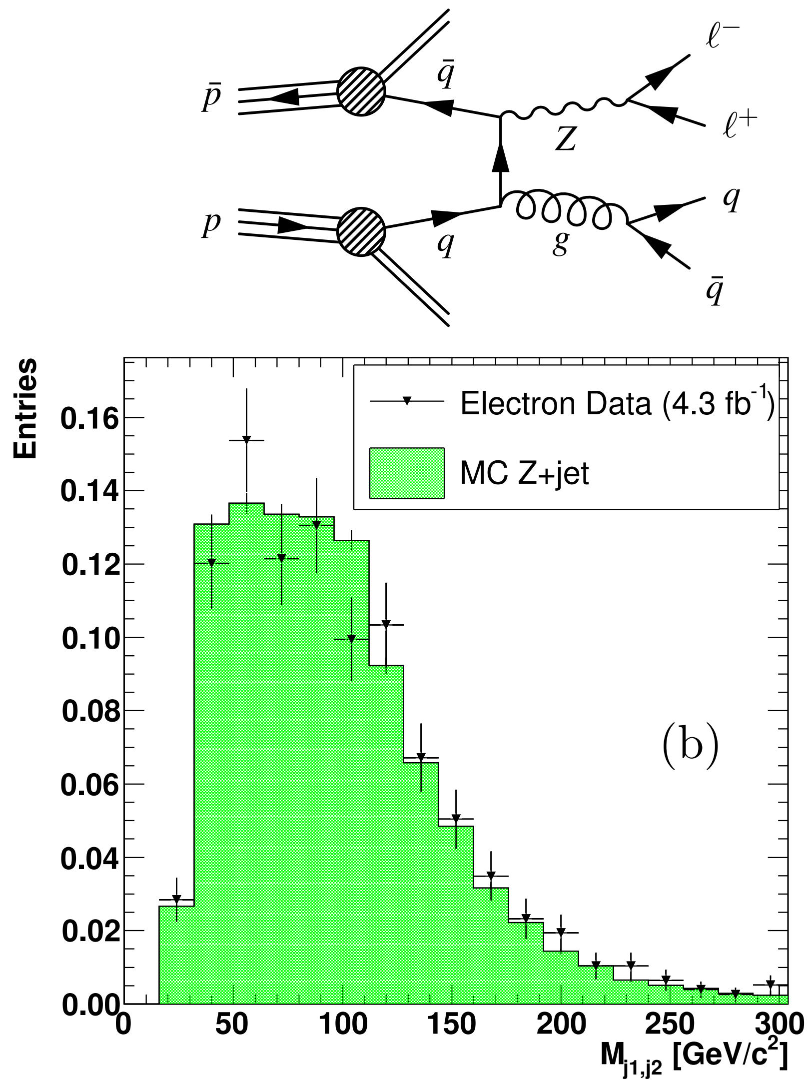
Often the extrapolation parameter has uncertainty

- introduce a new measurement to constrain it as in the ABCD method
- what if..., what if ..., what if..., what if ..., what if..., what if ...

Data driven estimates

In the case of the CDF bump, the Z+jets control sample provides a data-driven estimate, but limited statistics. Using the simulation narrative over the data-driven is a **choice**. If you trust that narrative, it's a good choice.



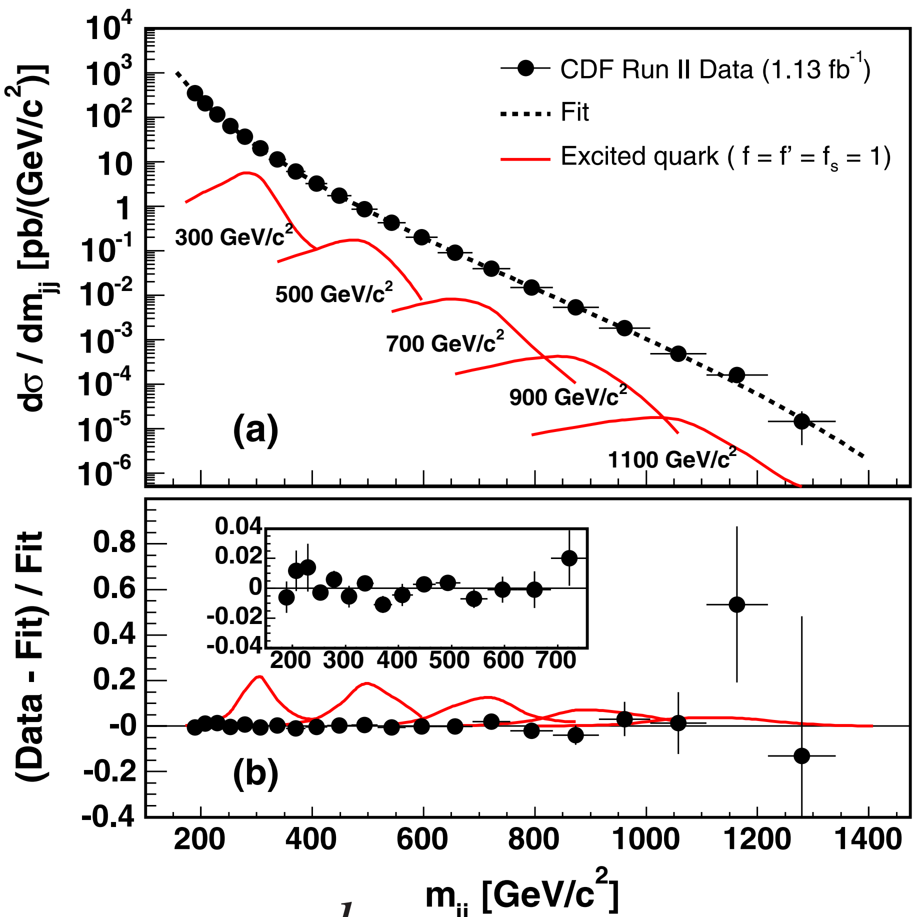


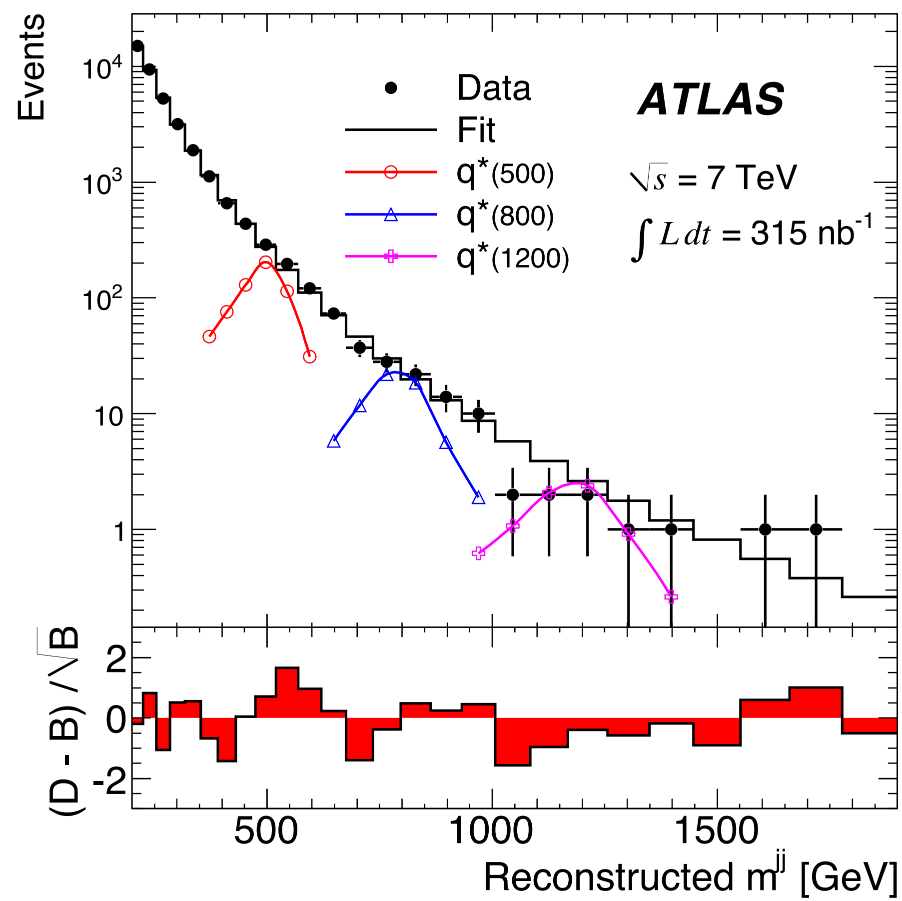
The Effective Model Narrative

It is common to describe a distribution with some parametric function

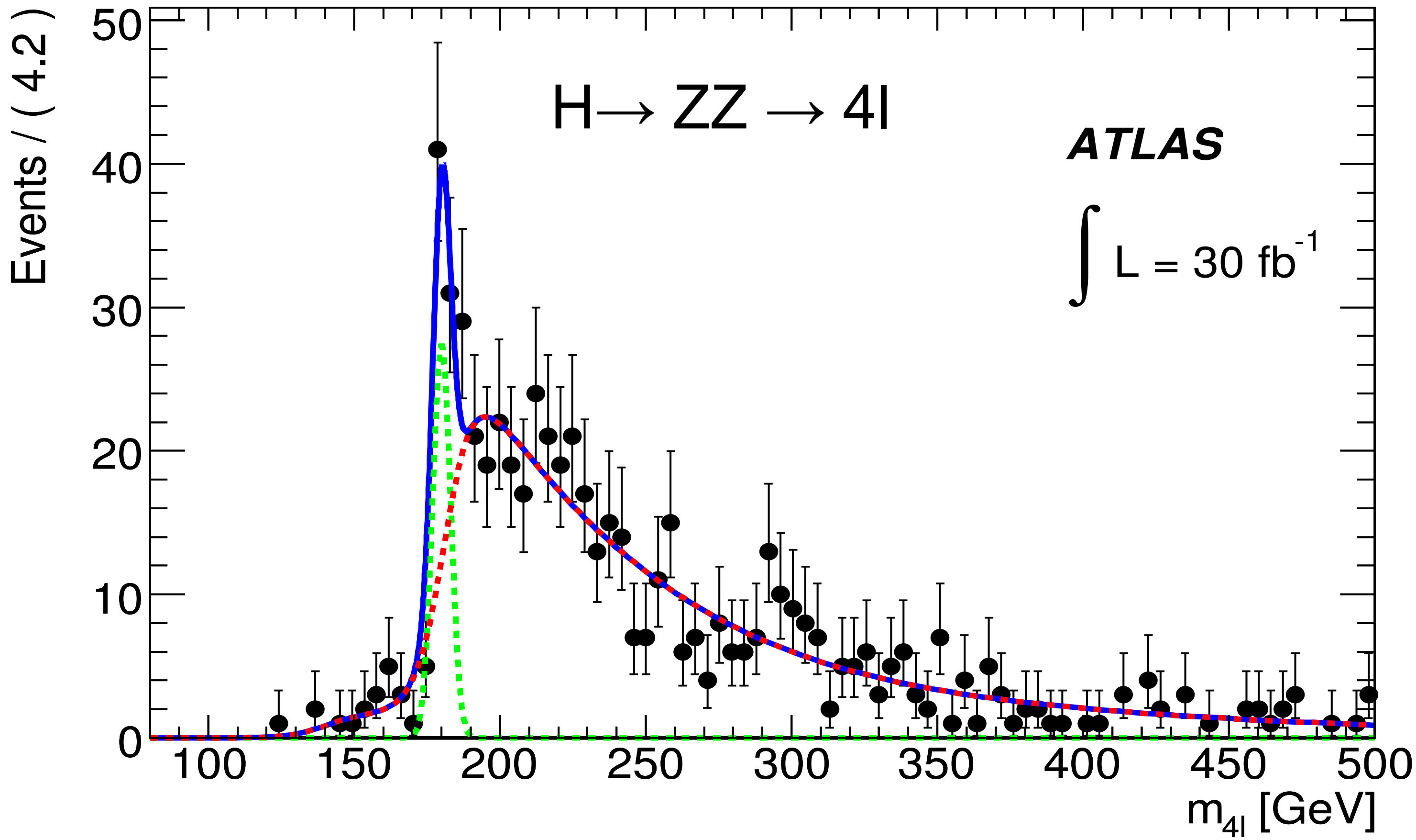
- “fit background to a polynomial”, exponential, ...
- While this is convenient and the fit may be good, the narrative is weak

PHYSICAL REVIEW D **79,** 112002 (2009)





$$\frac{d\sigma}{dm_{jj}} = p_0(1-x)^{p_1}/x^{p_2+p_3\cdot\ln(x)}, \qquad x = m_{jj}/\sqrt{s},$$

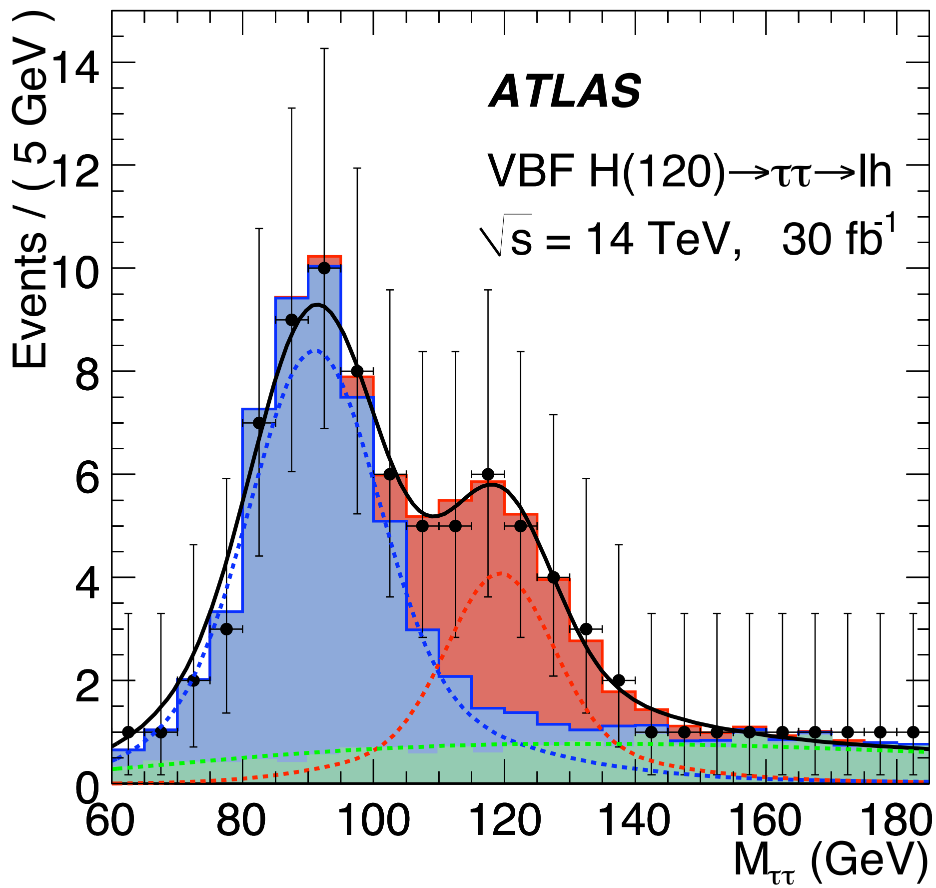


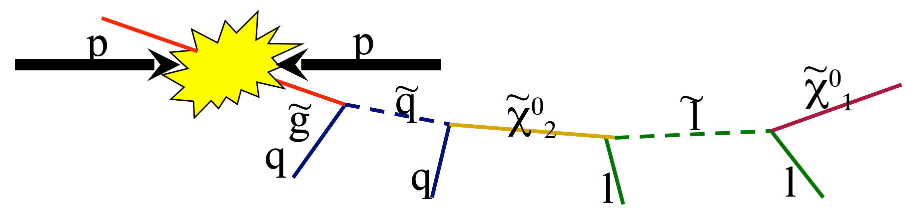
$$f(m_{ZZ}) = \frac{p0}{(1+e^{\frac{p6-m_{ZZ}}{p7}})(1+e^{\frac{m_{ZZ}-p8}{p9}})} + \frac{p1}{(1+e^{\frac{p2-m_{ZZ}}{p3}})(1+e^{\frac{p4-m_{ZZ}}{p5}})}$$

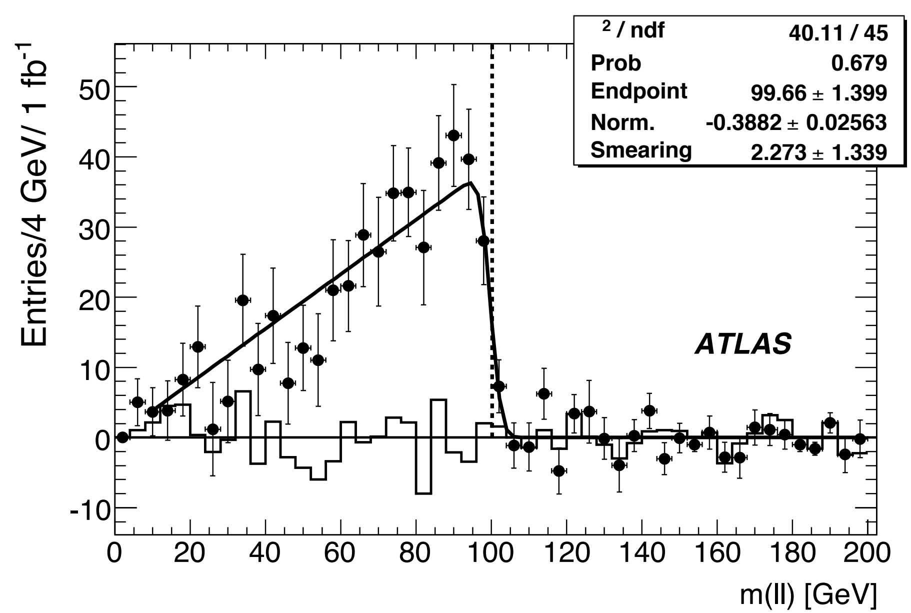
The Effective Model Narrative

Sometimes the effective model comes from a convincing narrative

- convolution of detector resolution with known distribution
  - Ex: MissingET resolution propagated through $M_{\tau\tau}$ in collinear approximation
  - Ex: lepton resolution convoluted with triangular M_{ll} distribution



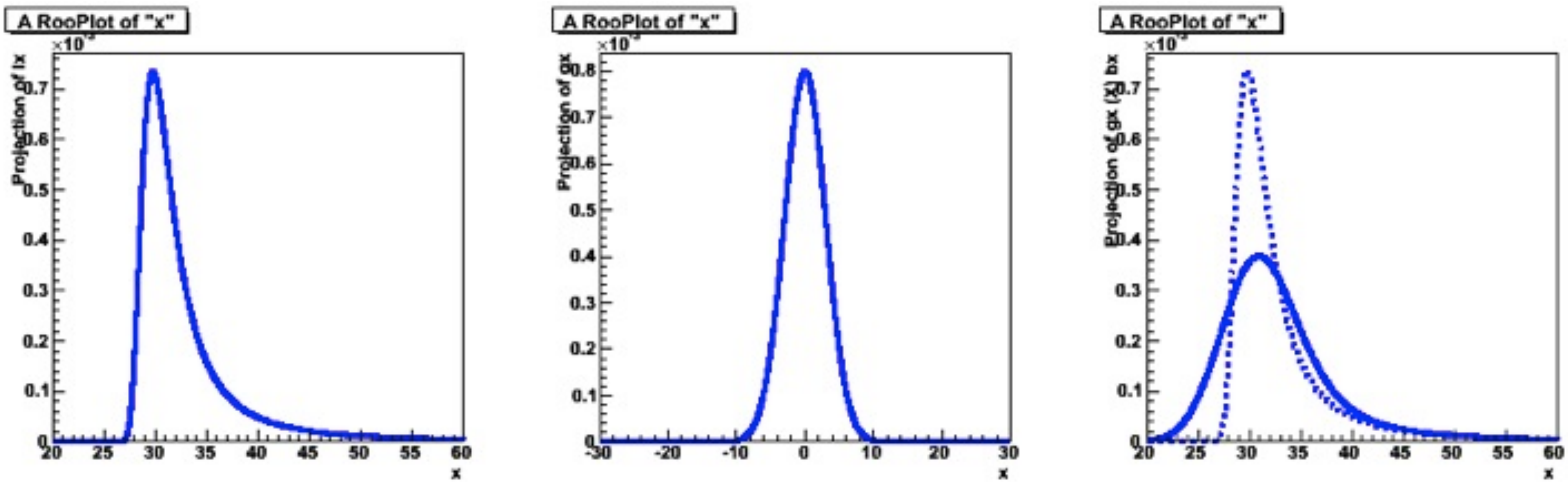




Tools for building effective models

- RooFit's convolution PDFs can aid in building more effective models with a more convincing narrative

```
// Construct landau (x) gauss (10000 samplings 2nd order interpolation)
t.setBins(10000,"cache") ;
RooFFTConvPdf lxg("lxg","landau (X) gauss",t,landau,gauss,2) ;
```

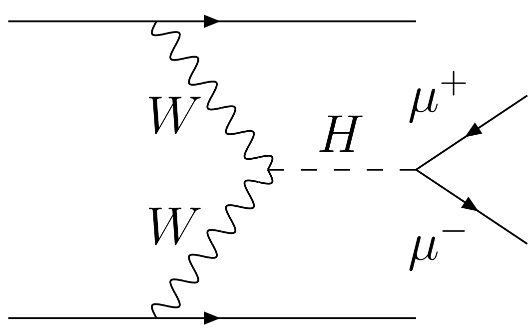


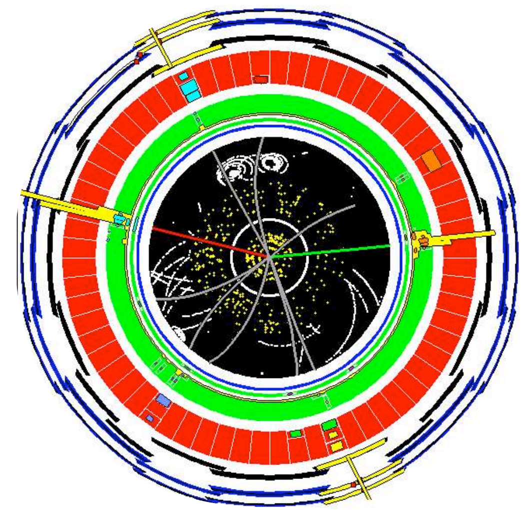
Wouter Verkerke, NIKHEF

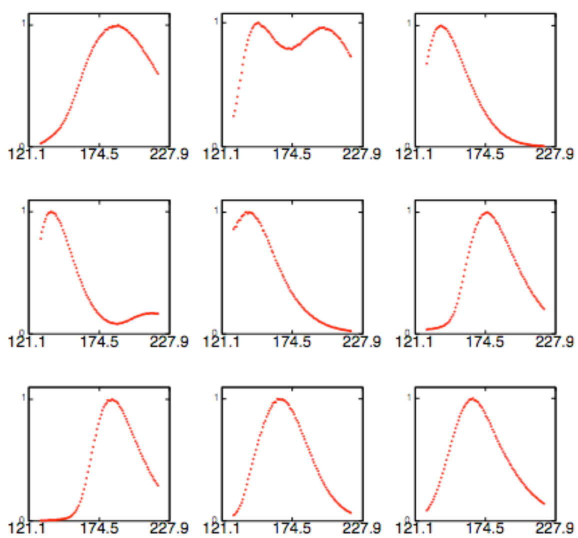
The parametrized response narrative

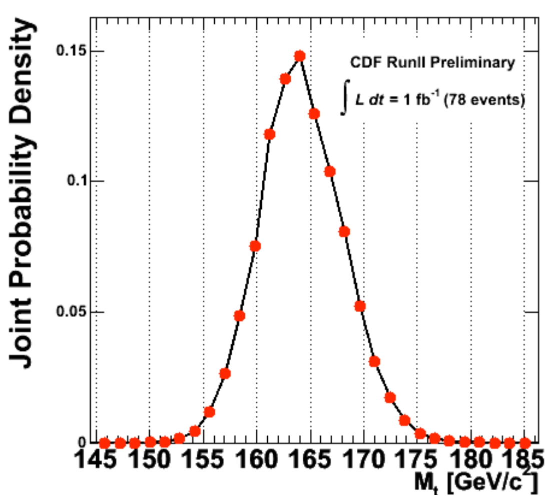
The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

- Doesn't require building parametrized PDF by interpolating between non-parametric templates.

$L(x|H_0) =$ 

$\oplus$ 





The parametrized response narrative

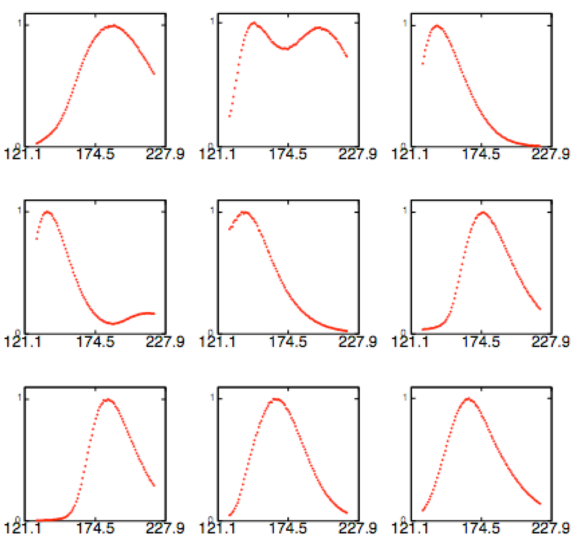
Center for Cosmology and Particle Physics

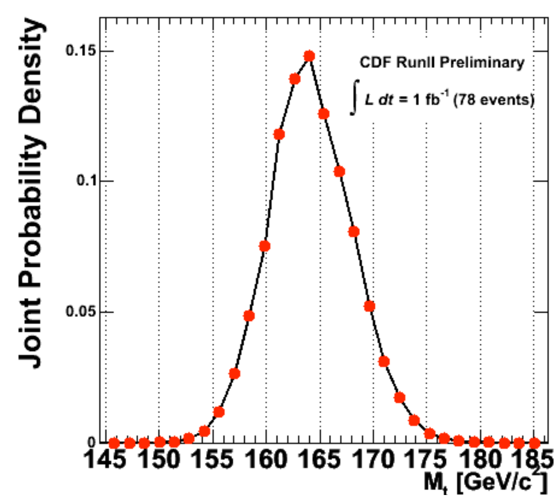
The Matrix-Element technique is conceptually similar to the simulation narrative, but the detector response is parametrized.

- Doesn't require building parametrized PDF by interpolating between non-parametric templates.

$$P(\mathbf{x}|M_t) = \frac{1}{N}\int d\Phi\ |\mathcal{M}_{t\bar{t}}(p;M_t)|^2 \prod_{jets} f(p_i, j_i) f_{PDF}(q_1) f_{PDF}(q_2)$$

Phase-space Integral — **Matrix Element** — **Transfer Functions**



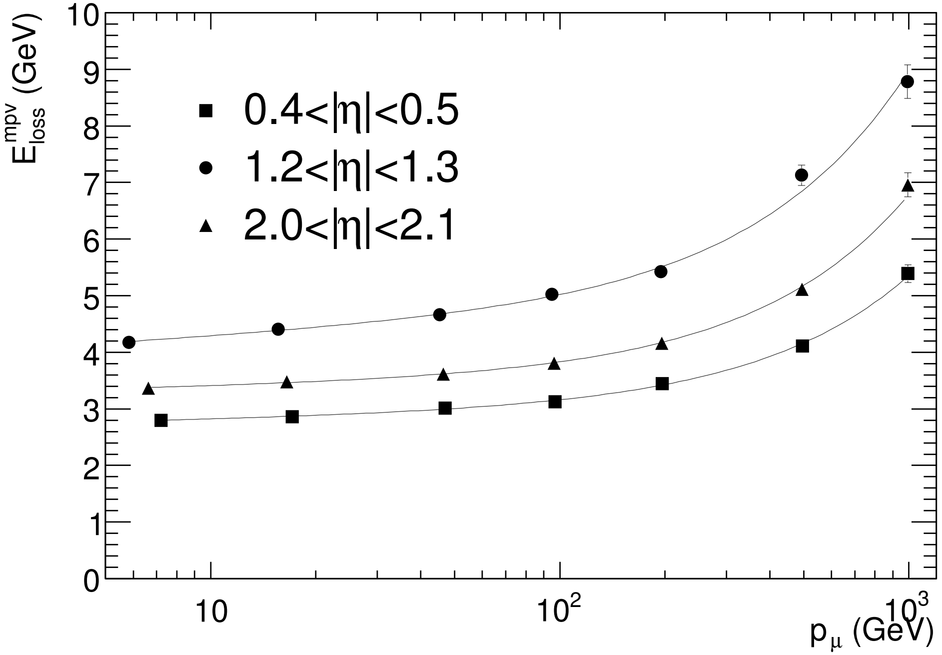


Examples of parametrized response

While we often see the parametrized response as overly simplistic, the parametrizations are often based on some deeper understanding

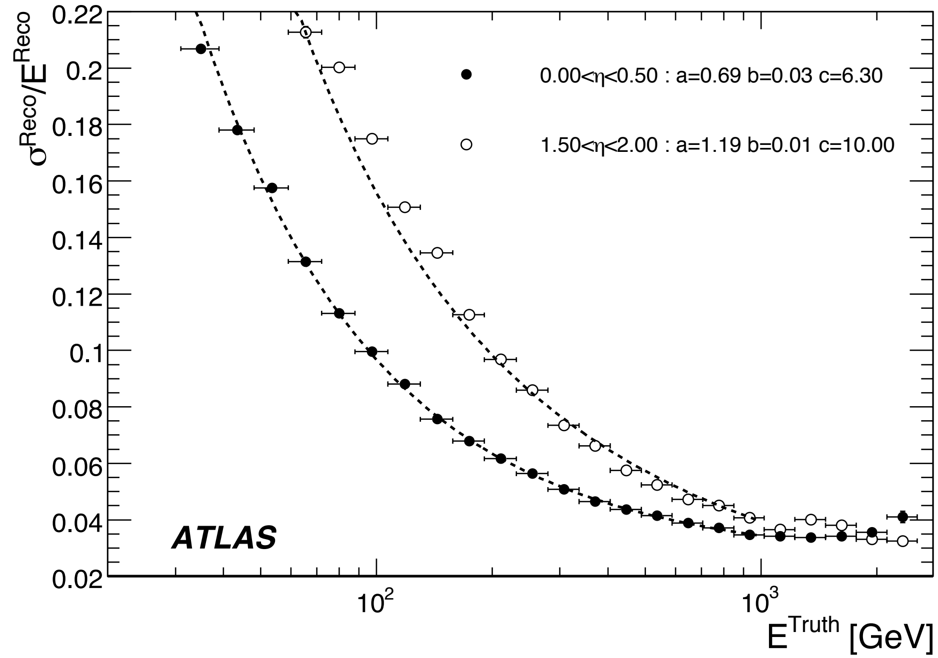
- and parameters can often be measured in data with in situ calibration strategies. No reason we can't propagate uncertainty to next stage.

Muon Energy Loss (Landau)



$$E_{\text{loss}}^{\text{mpv}}(p_\mu) = a_0^{\text{mpv}} + a_1^{\text{mpv}} \ln p_\mu + a_2^{\text{mpv}} p_\mu$$

Jet Resolution



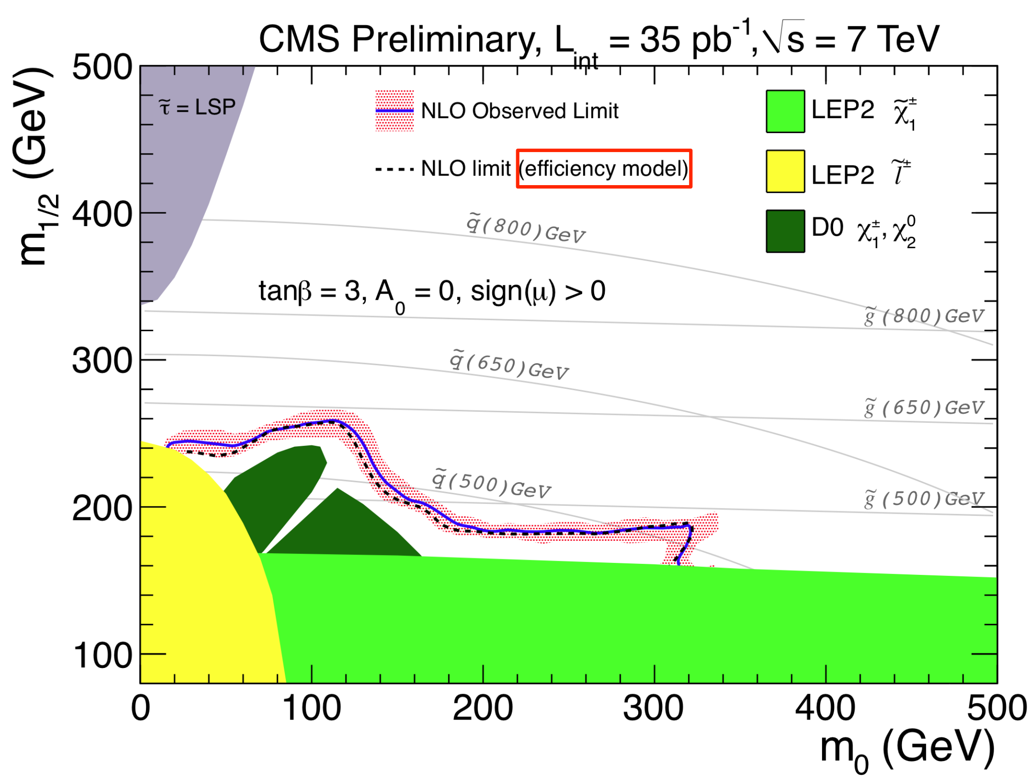
$$\frac{\sigma}{E} = \frac{a}{\sqrt{E\ (\text{GeV})}} \oplus b \oplus \frac{c}{E}.$$

Fast Simulation

Fast simulations based on parametrized detector response are very useful and can often be tuned to perform quite well in a specific analysis context

- For example: tools like PGS, Delphis, ATLFAST, ...

Same sign di-lepton + jets + MET search



Paper includes a simple efficiency model (i.e. for PGS calibrations) and compares full limit to limit with simple model.

CMS SUSY Results, D. Stuart, April 2011, SUSY Recast, UC Davis 36

Fast Simulation

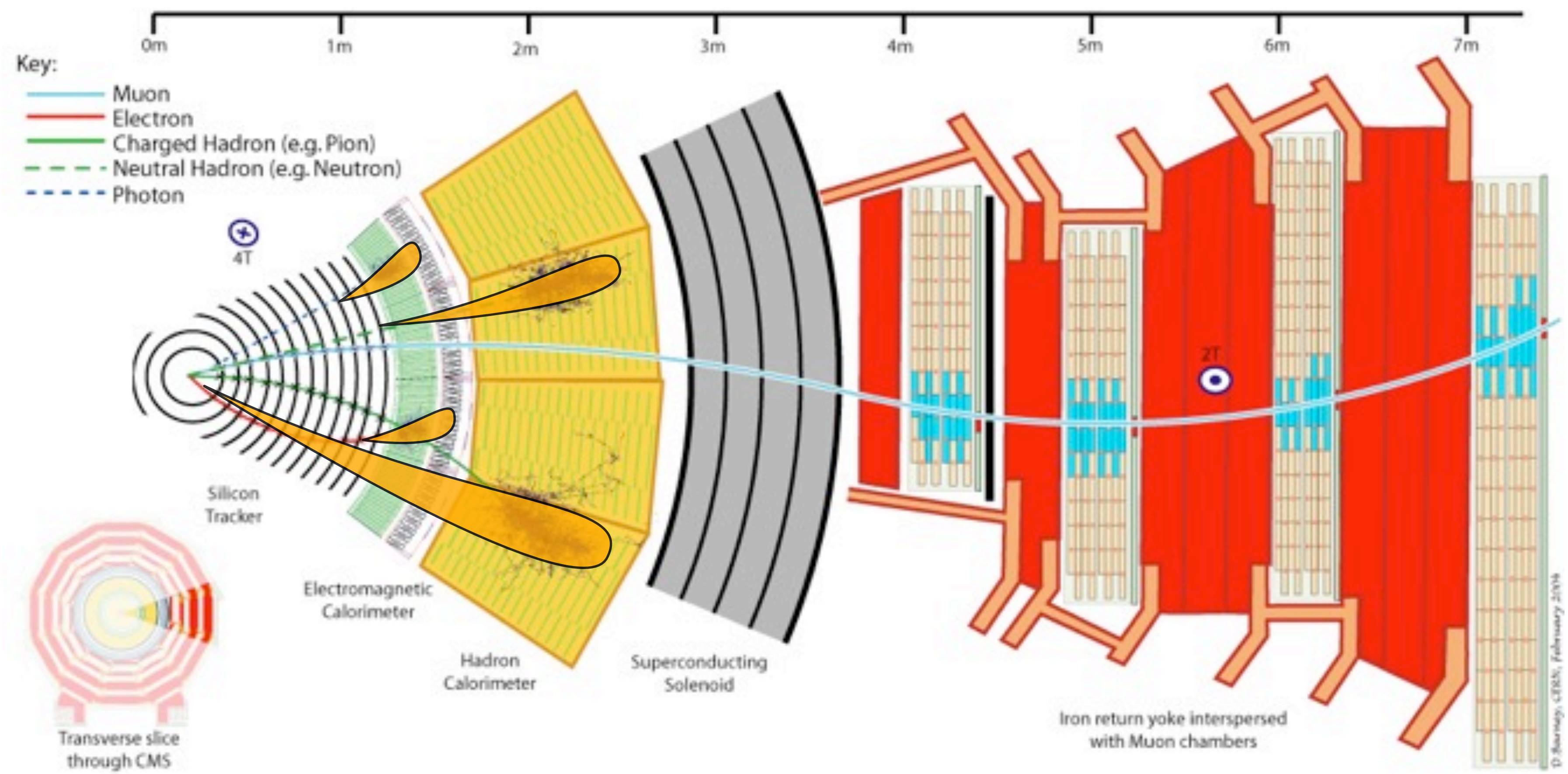
CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

Fast simulations based on parametrized detector response are very useful and can often be tuned to perform quite well in a specific analysis context

- For example: tools like PGS, Delphis, ATLFAST, ...

But these tools still use accept/reject Monte Carlo.

- Would be much more useful if the parametrized detector response could be used as a transfer function in Matrix-Element approach



Narrative styles

The Monte Carlo Simulation narrative (MC narrative)

- each stage is an accept/reject Monte Carlo based on P(out|in) of some microscopic process like parton shower, decay, scattering
- PDFs built from non-parametric estimator like histograms or kernel estimation
  - need to supplement with interpolation procedures to incorporate systematics
  - smearing approach fundamentally Bayesian
- **pros:** most detailed understanding of micro-physics
- **cons:** computationally demanding, loose analytic scaling properties, relies on accuracy of simulation
- **new ideas:** improved interpolation, Radford Neal's machine learning, "design of experiments"

The Data-driven narrative

- independent data sample that either acts as a proxy for some process or can be transformed to do so
- **pros**: nature includes "all orders", uses real detector
- **cons**: extrapolation from control region to signal region requires assumptions, introduces systematic effects. Appropriate transformation may depend on many variables, which becomes impractical

Narrative styles

Center for Cosmology and Particle Physics

Effective modeling narrative

- parametrized functional form: eg. Gaussian, falling exponential para polynomial fit to distribution, etc.
- **pros**: fast, has analytic scaling, parametric form may be well justified (eg. phase space, propagation of errors, convolution)
- **cons**: approximate, parametric form may be ad hoc (eg. polynomial from)
- **new ideas:** using non-parametric statistical methods

Parametrized detector response narrative (eg. kinematic fitting, Matrix-Element method, ~fast simulation)

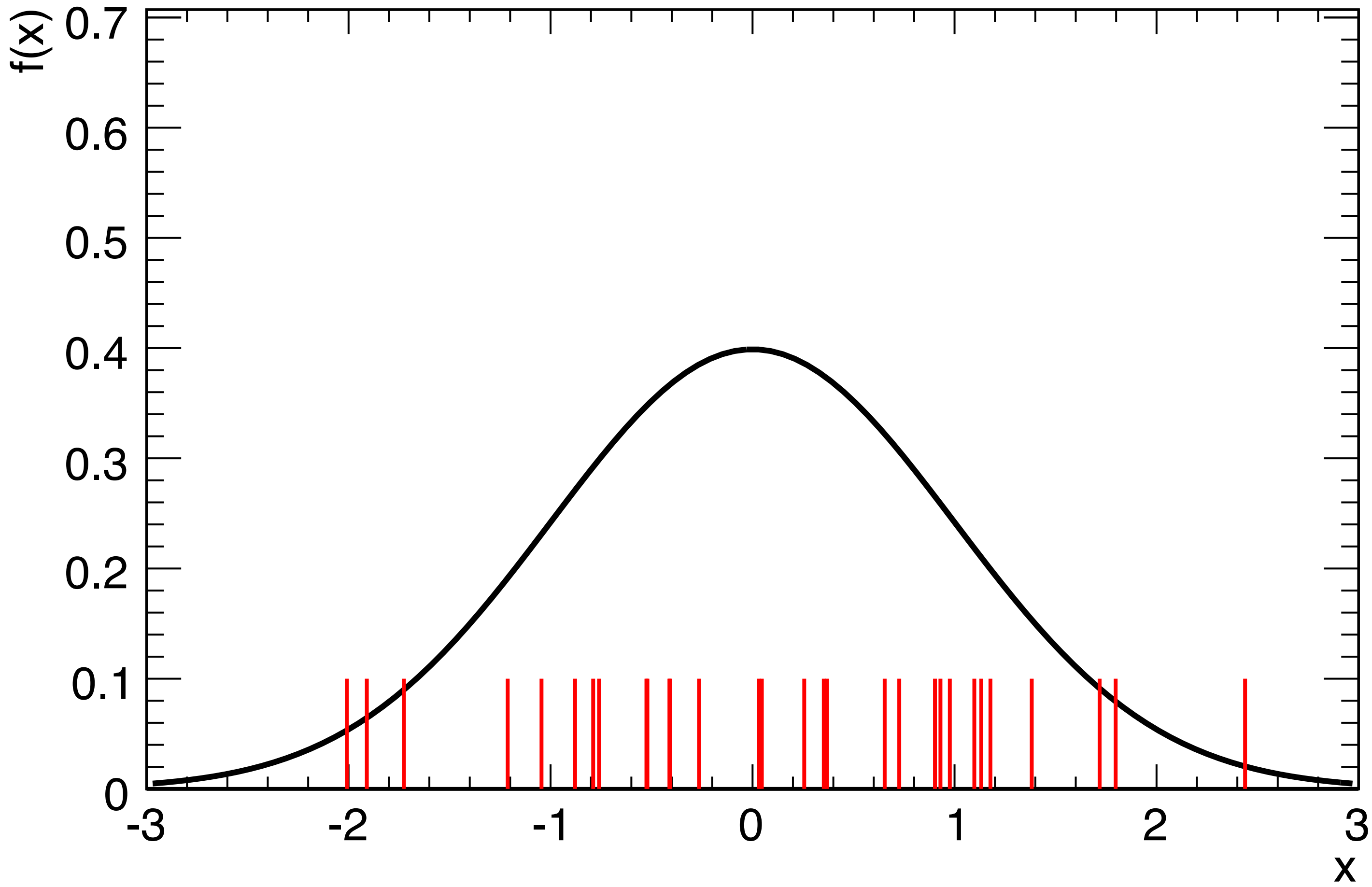
- **pros**: fast, maintains analytic scaling, response usually based on good understanding of the detector, possible to incorporate some types of uncertainty in the response analytically, can evaluate P(out|in) for arbitrary out,in.
- **cons**: approximate, best parametrized detector response is often not available in convenient form
- **new ideas:** fast simulation is typically parametrized, but we use it in an accept/reject framework (see Geant5)

Parametric vs. Non-Parametric PDFs

No parametric form, need to construct **non-parametric** PDFs

From Monte Carlo samples, one has empirical PDF

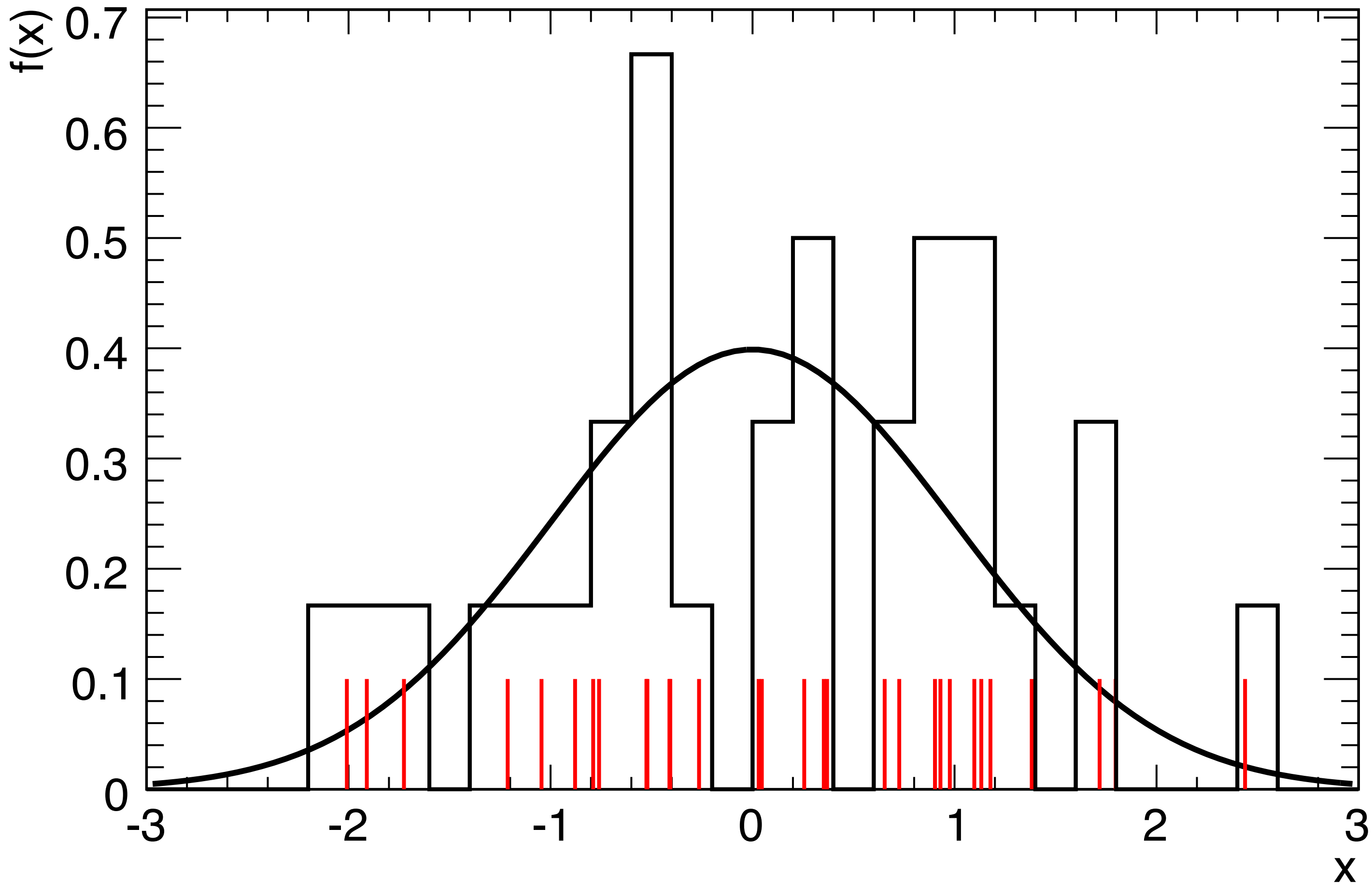
$$f_{emp} = \frac{1}{N}\sum_{i}^{N}\delta(x - x_i)$$



Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

$$f^{w,s}_{hist}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

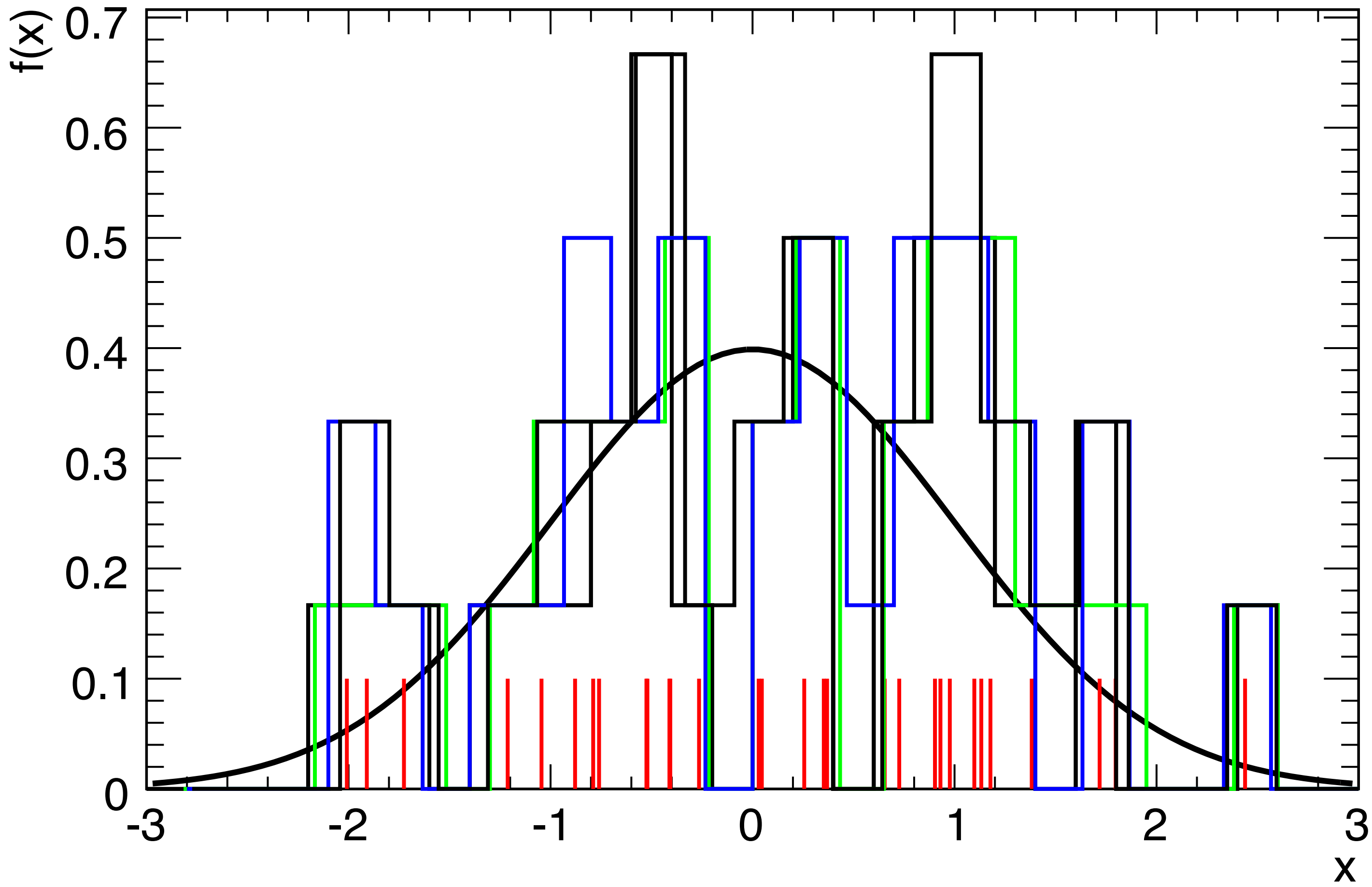


Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

but they depend on bin width and starting position

$$f_{hist}^{w,s}(x) = \frac{1}{N}\sum_i h_i^{w,s}$$

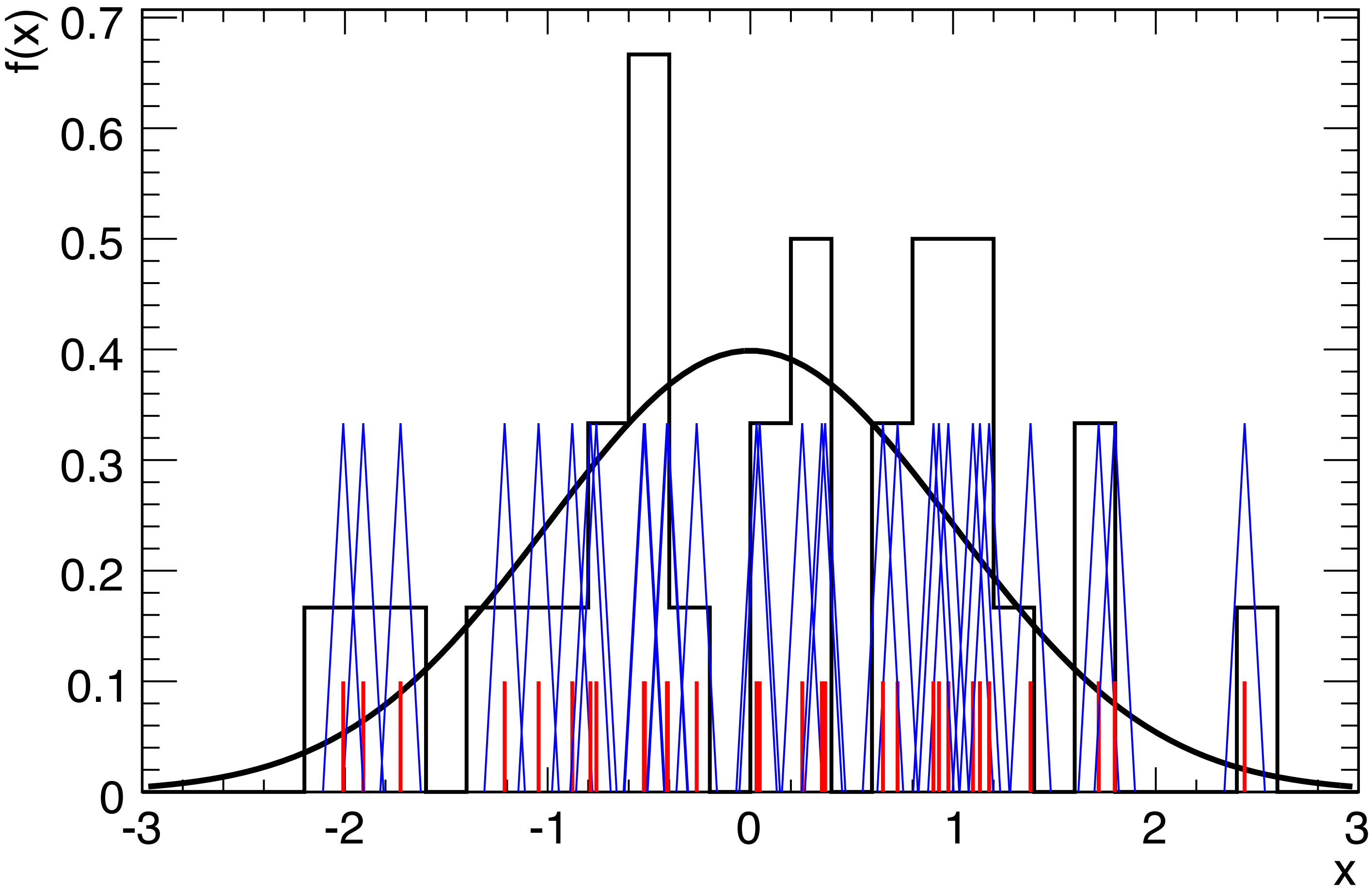


Parametric vs. Non-Parametric PDFs

Classic example of a **non-parametric** PDF is the histogram

"Average Shifted Histogram" minimizes effect of binning

$$f^w_{ASH}(x) = \frac{1}{N}\sum_i^N K^w(x - x_i)$$

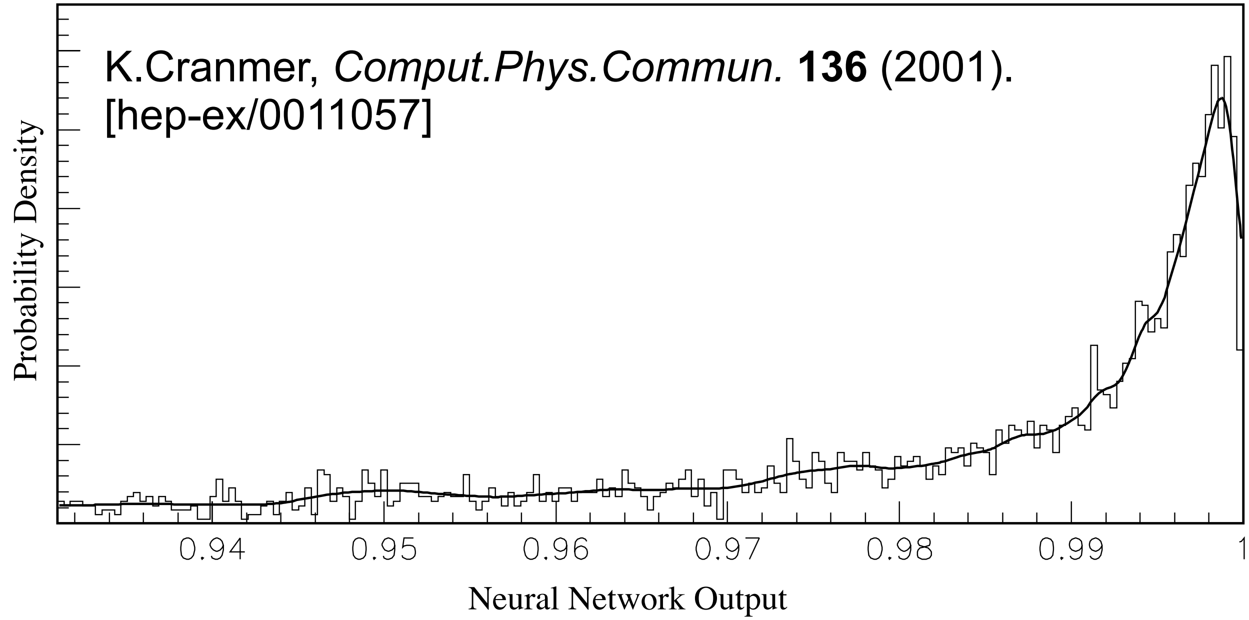


Kernel Estimation

Kernel estimation is the generalization of Average Shifted Histograms

$$\hat{f}_1(x) = \sum_i^n \frac{1}{nh(x_i)} K\left(\frac{x - x_i}{h(x_i)}\right)$$

$$h(x_i) = \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{\hat{f}_0(x_i)}}\, n^{-1/5}$$



“the data is the model”

Adaptive Kernel estimation puts wider kernels in regions of low probability

Used at LEP for describing pdfs from Monte Carlo (KEYS)

Multivariate, non-parametric PDFs

Kernel Estimation has a nice generalizations to higher dimensions

- practical limit is about 5-d due to curse of dimensionality

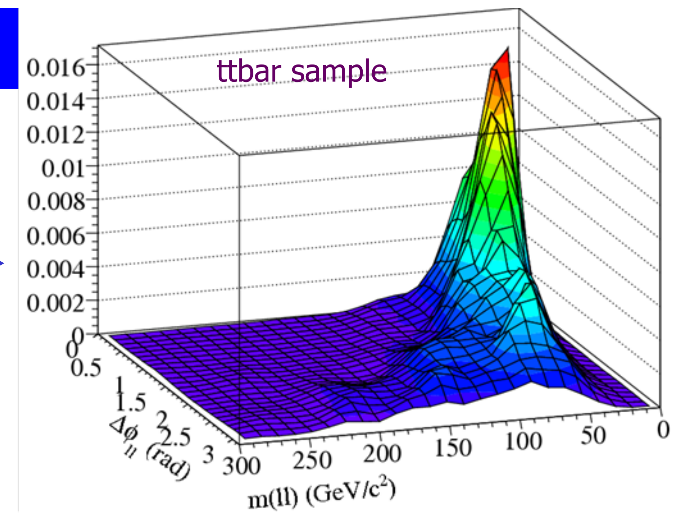
Max Baak has coded N-dim KEYS pdf described in Comput.Phys.Commun. **136** (2001) in RooFit.

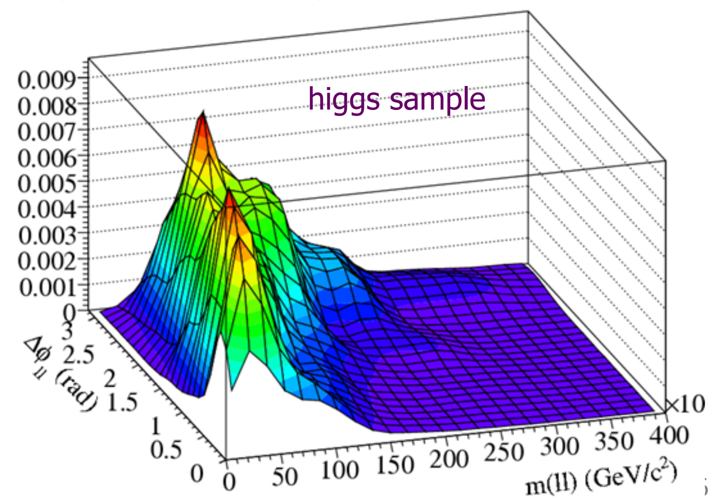
These pdfs have been used as the basis for a multivariate discrimination technique called "PDE"

$$D(\vec{x}) = \frac{f_s(\vec{x})}{f_s(\vec{x}) + f_b(\vec{x})}$$

Correlations

- 2-d projection of pdf from previous slide.
- RooNDKeys pdf automatically models (fine) correlations between observables ...





Max Baak