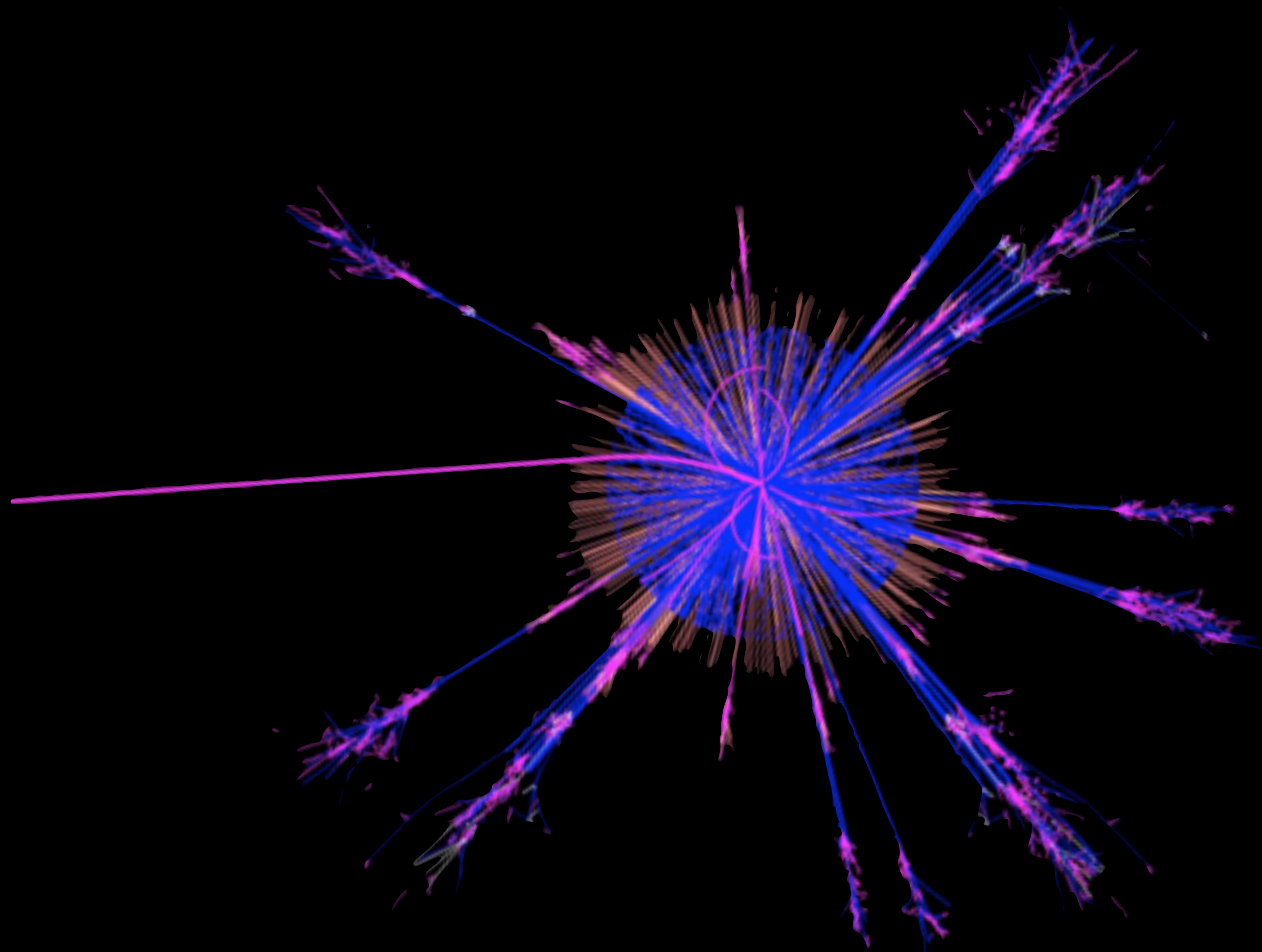
NYU Center for Data Science

Center for Cosmology and Particle Physics

LECTURES ON

STATISTICS

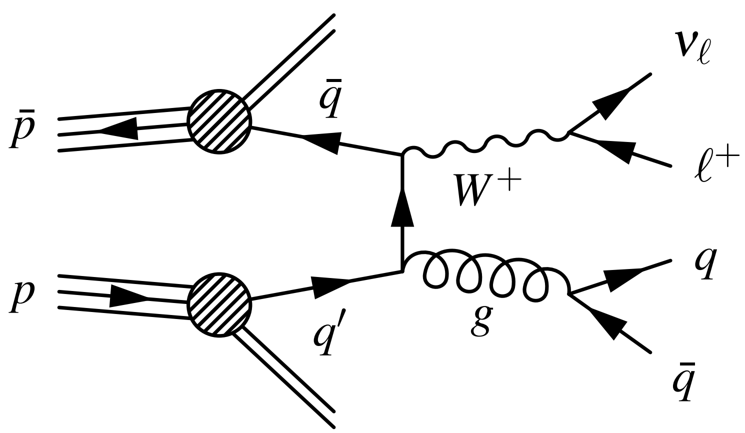


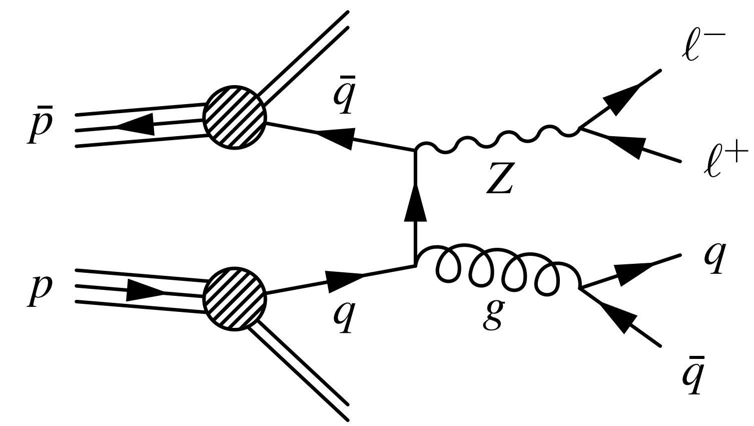
@KyleCranmer
New York University
Department of Physics
Center for Data Science

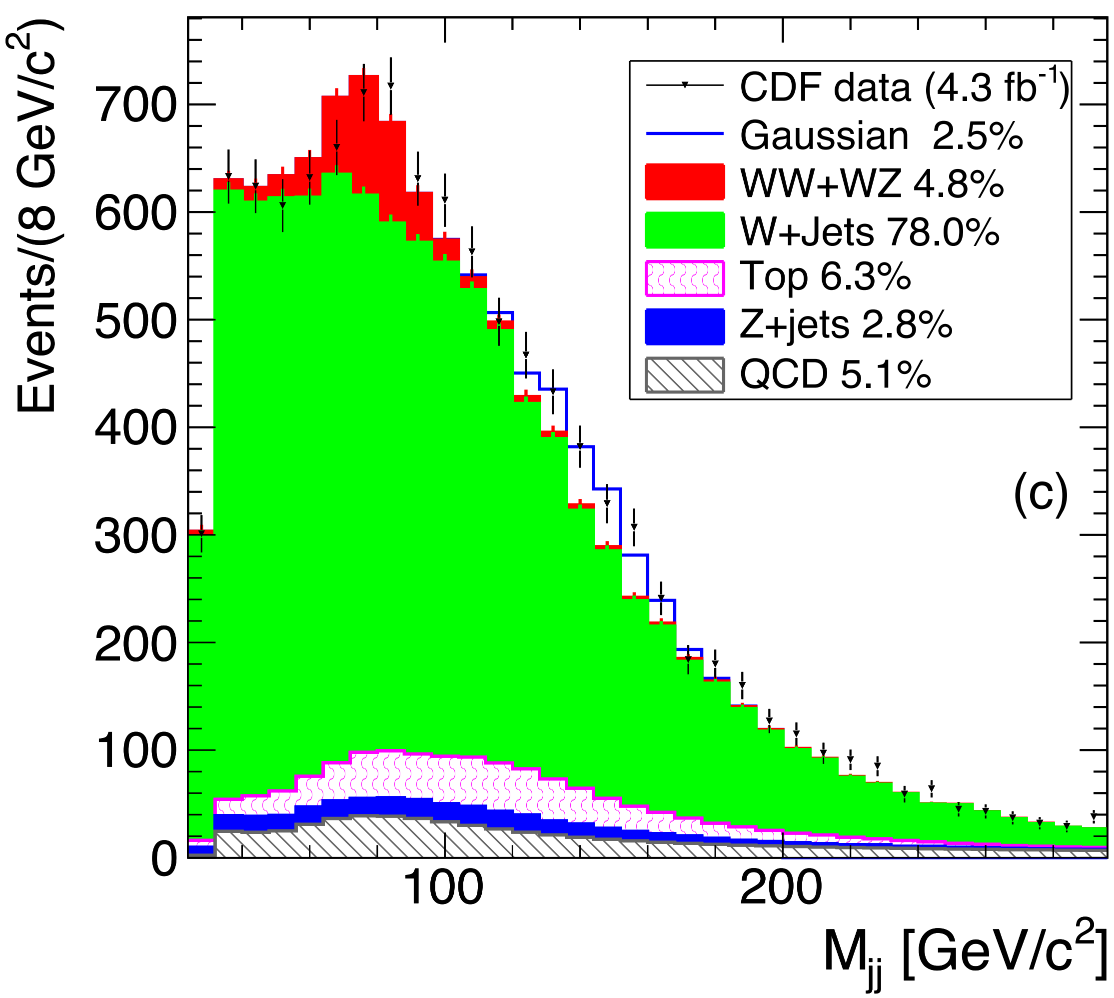
Modeling:
The Scientific Narrative

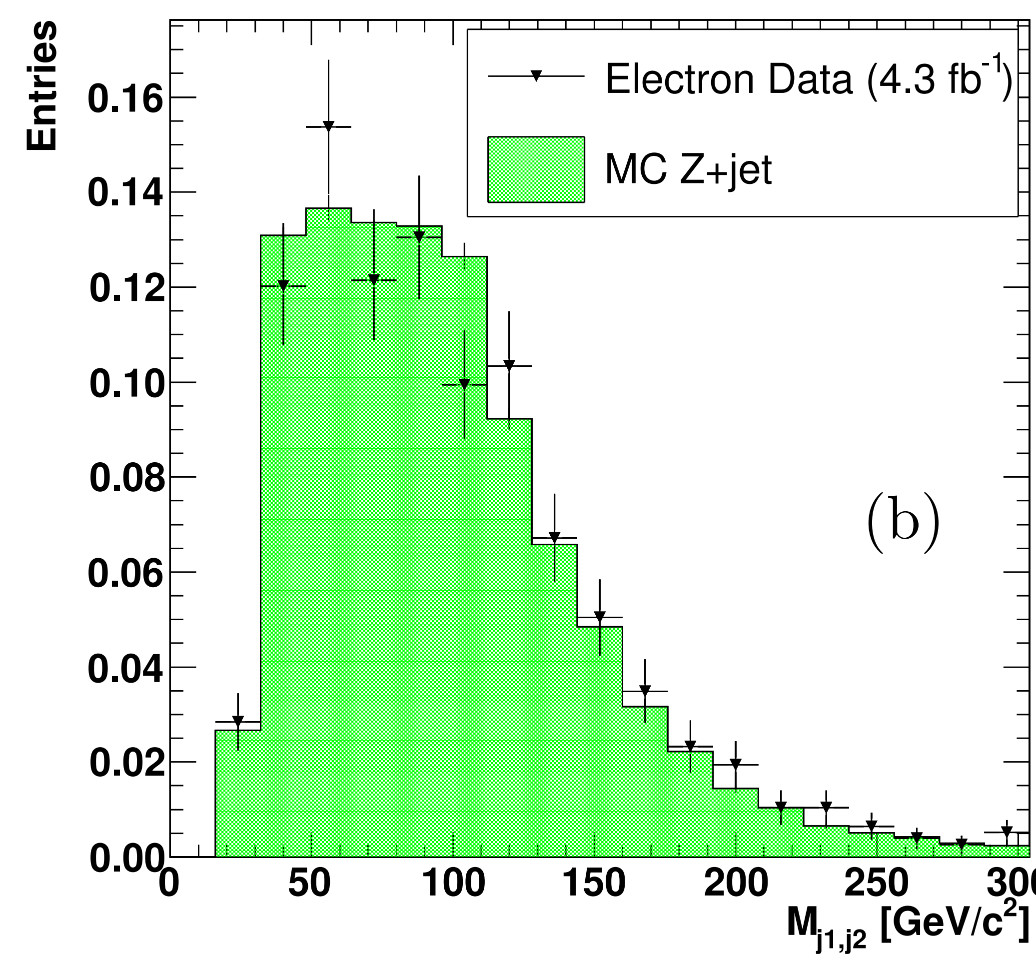
Choice: Data driven vs. Simulation

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.





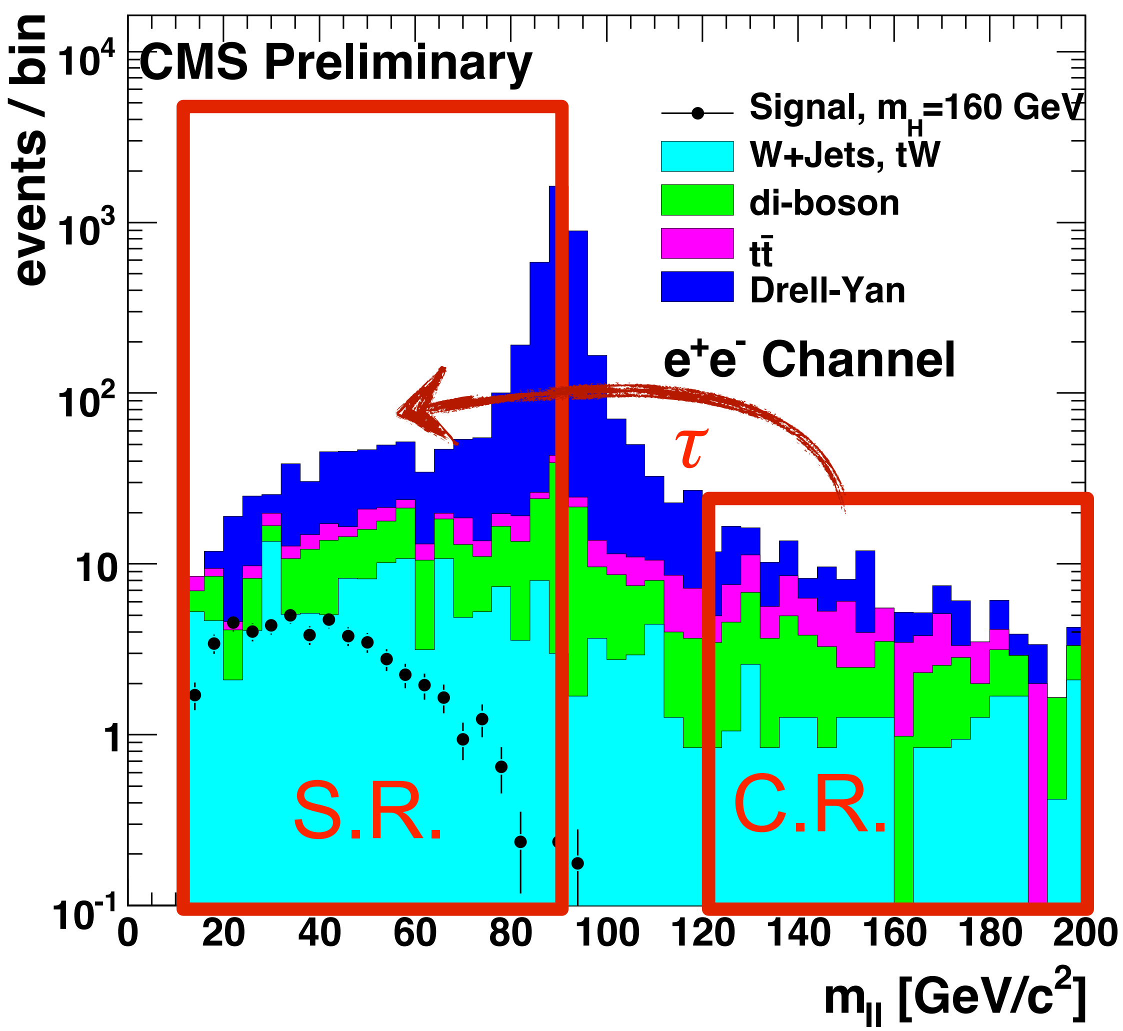




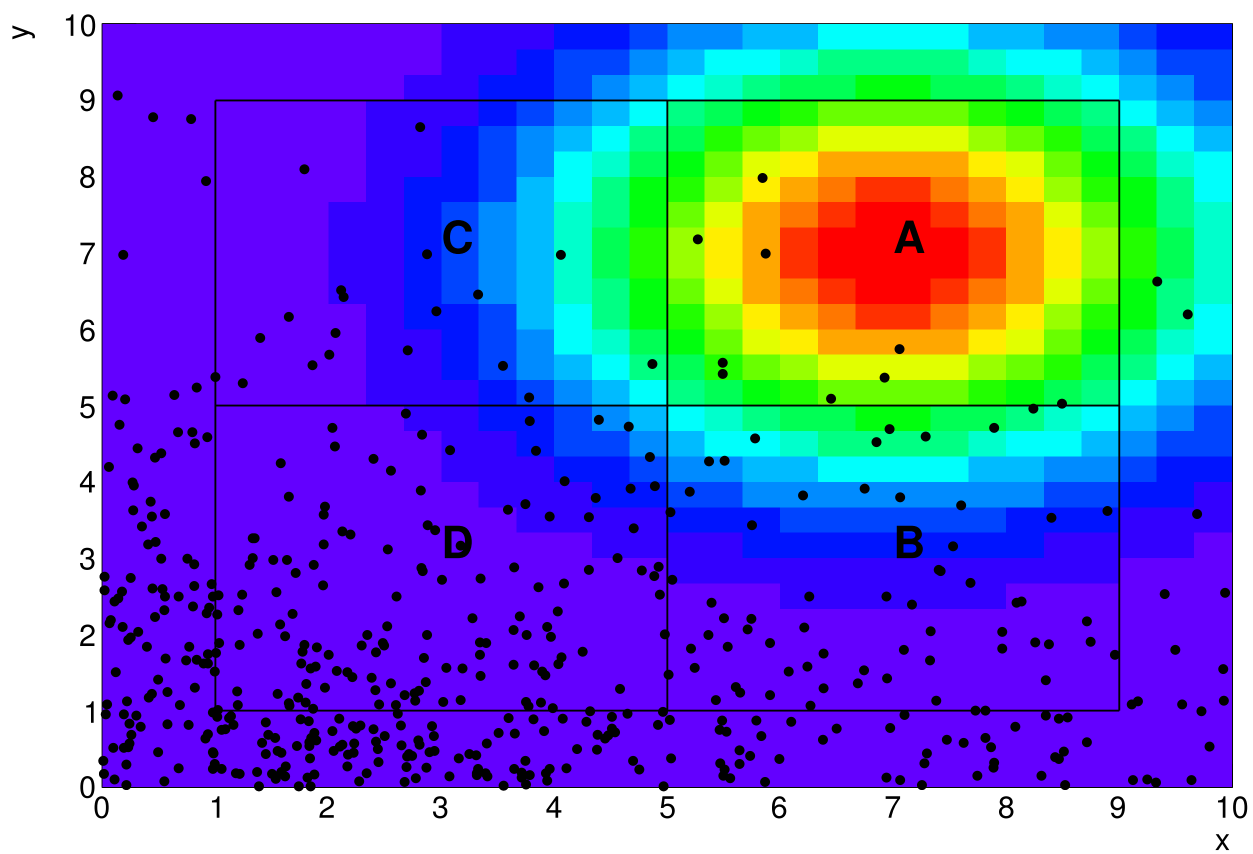
The data-driven narrative

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

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



"ABCD" method



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 "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

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

  - 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?

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

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

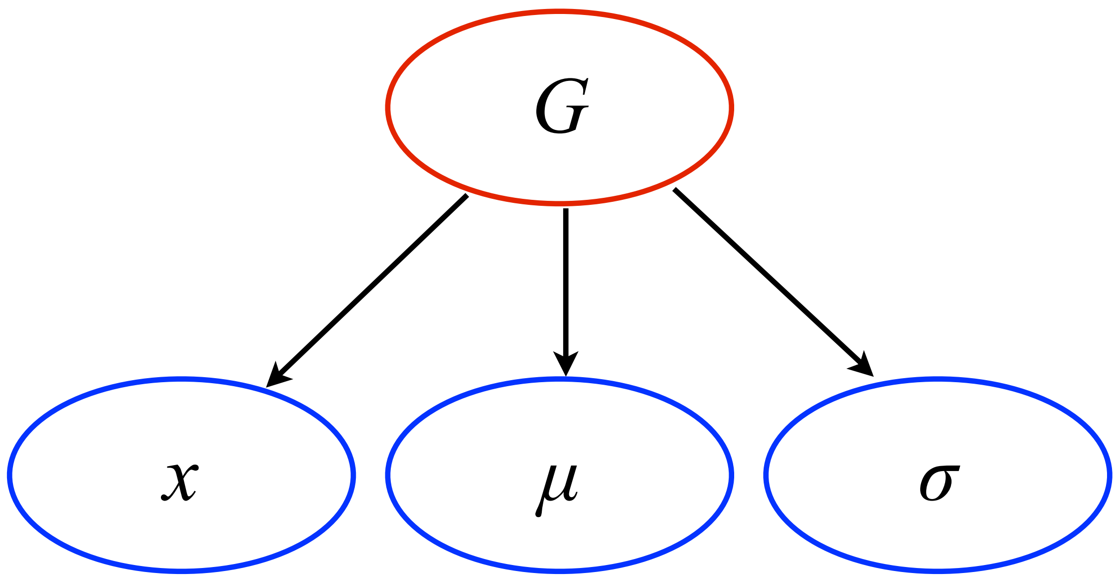
$$\pi(b) = P(b|n_{\mathrm{off}}) = \frac{P(n_{\mathrm{off}}|b)\eta(b)}{\int db P(n_{\mathrm{off}}|b)\eta(b)}.$$

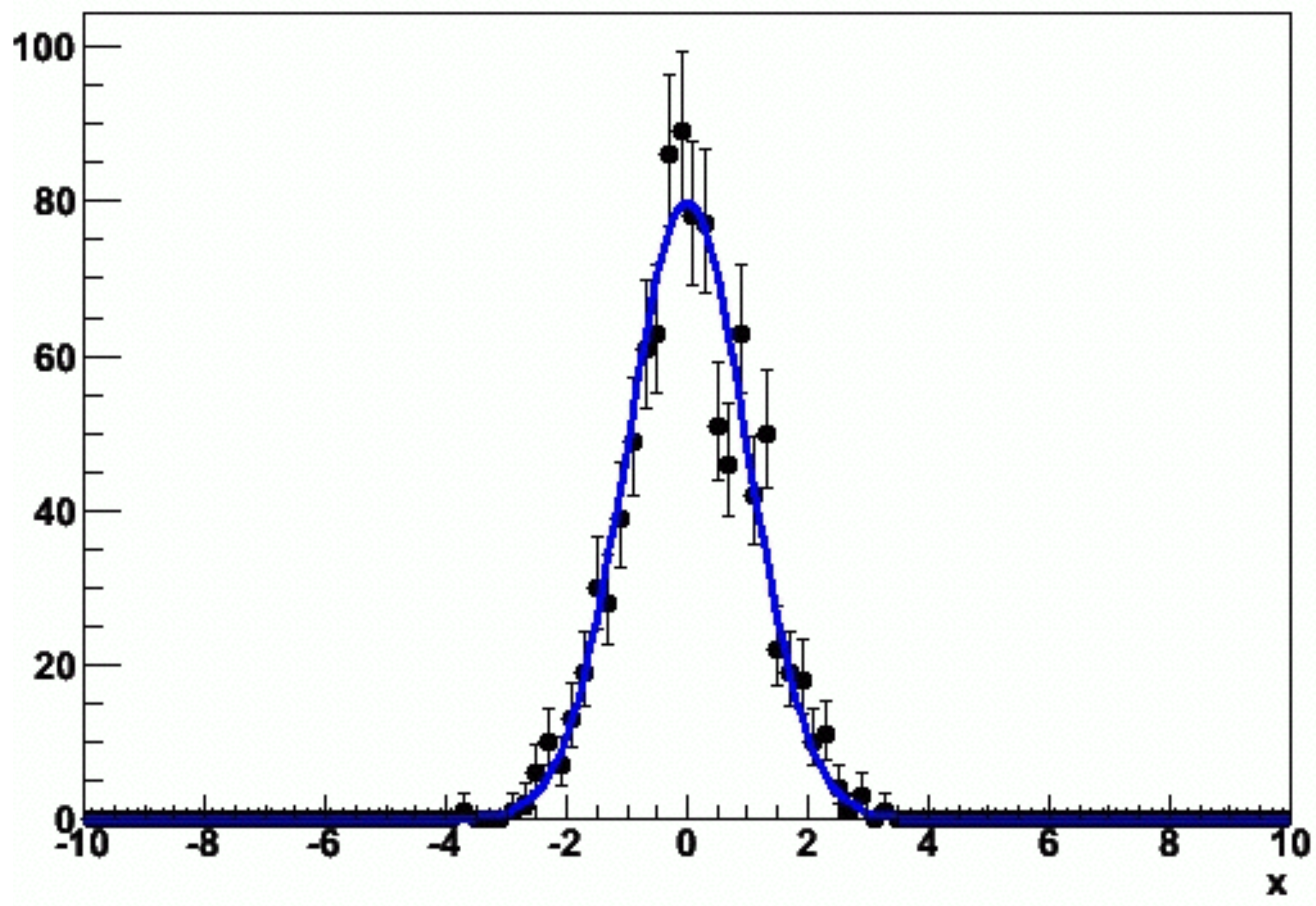
A GENERAL PURPOSE STATISTICAL MODEL

VISUALIZING PROBABILITY MODELS

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

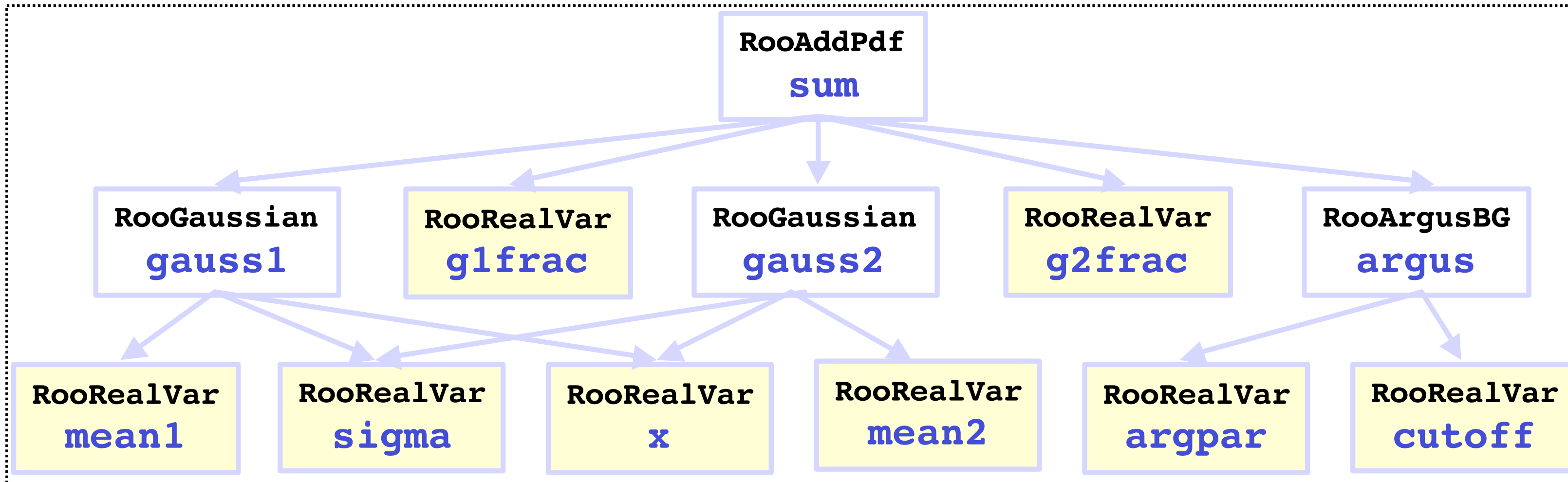
- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- every node is a real-valued function of the nodes below



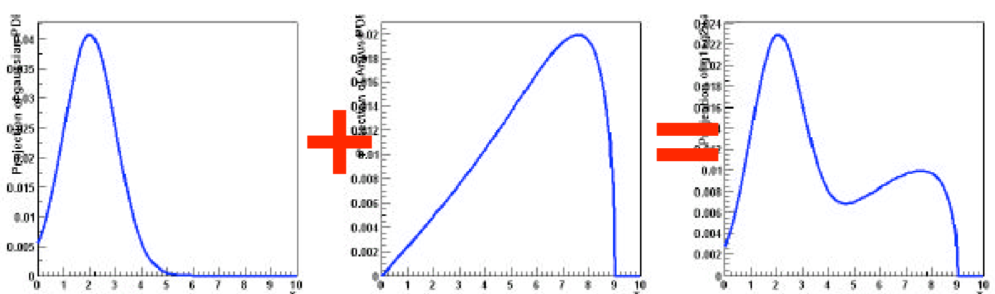


RooFit: A data modeling toolkit

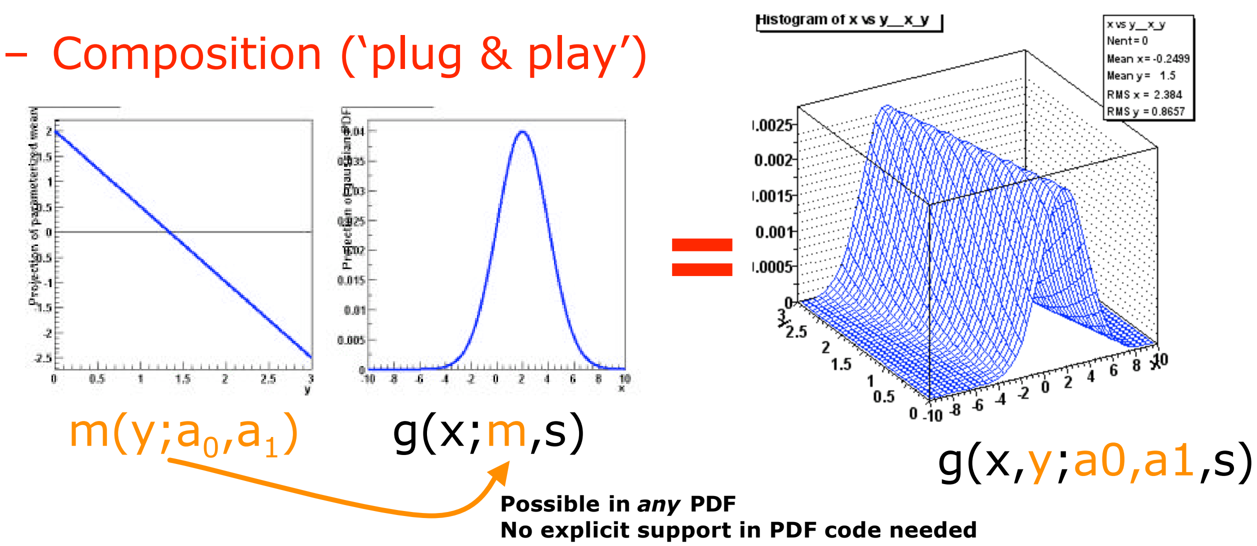
RooFit is a major tool developed at BaBar for data modeling.
RooStats provides higher-level statistical tools based on these PDFs.



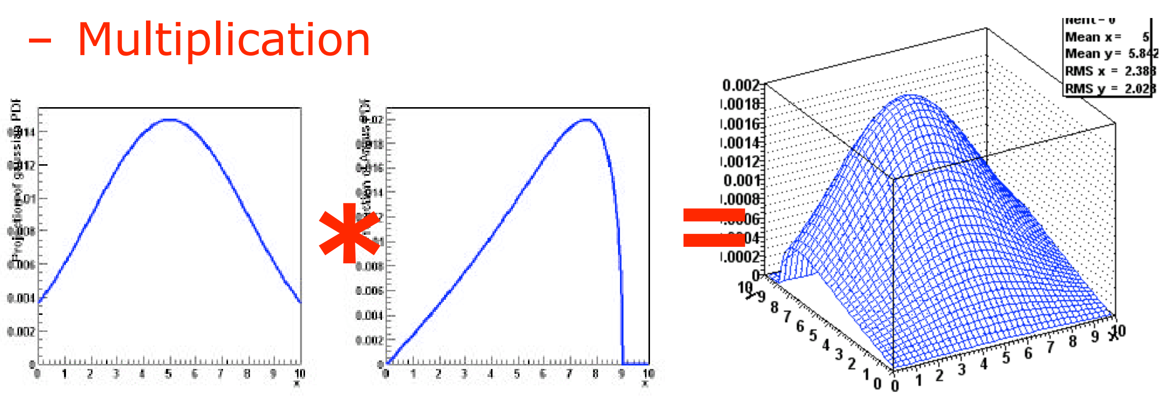
– Addition



– Composition ('plug & play')

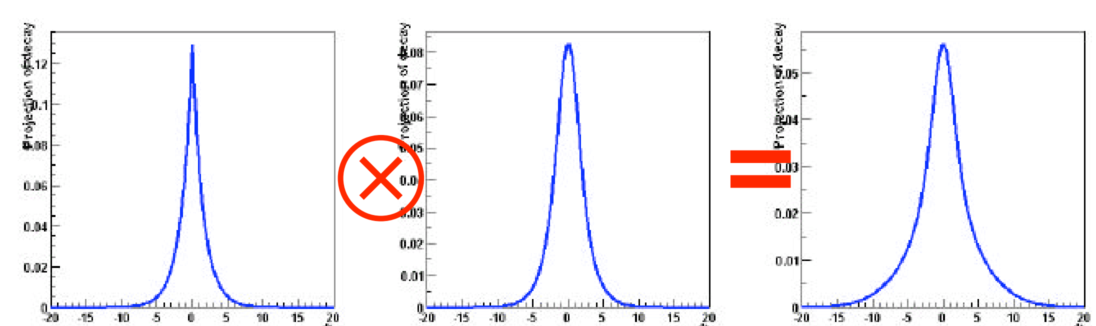


– Multiplication



Wouter Verkerke,

– Convolution



Wouter Verkerke, UCSB

Marked Poisson Process

Channel: a subset of the data defined by some selection requirements.

- eg. all events with 4 electrons with energy > 10 GeV
- n: number of events observed in the channel
- ν: number of events expected in the channel

Discriminating variable: a property of those events that can be measured and which helps discriminate the signal from background

- eg. the invariant mass of two particles
- $f(x)$: the p.d.f. of the discriminating variable x

$$\mathcal{D} = \{x_1, \ldots, x_n\}$$

Marked Poisson Process / Extended Likelihood:

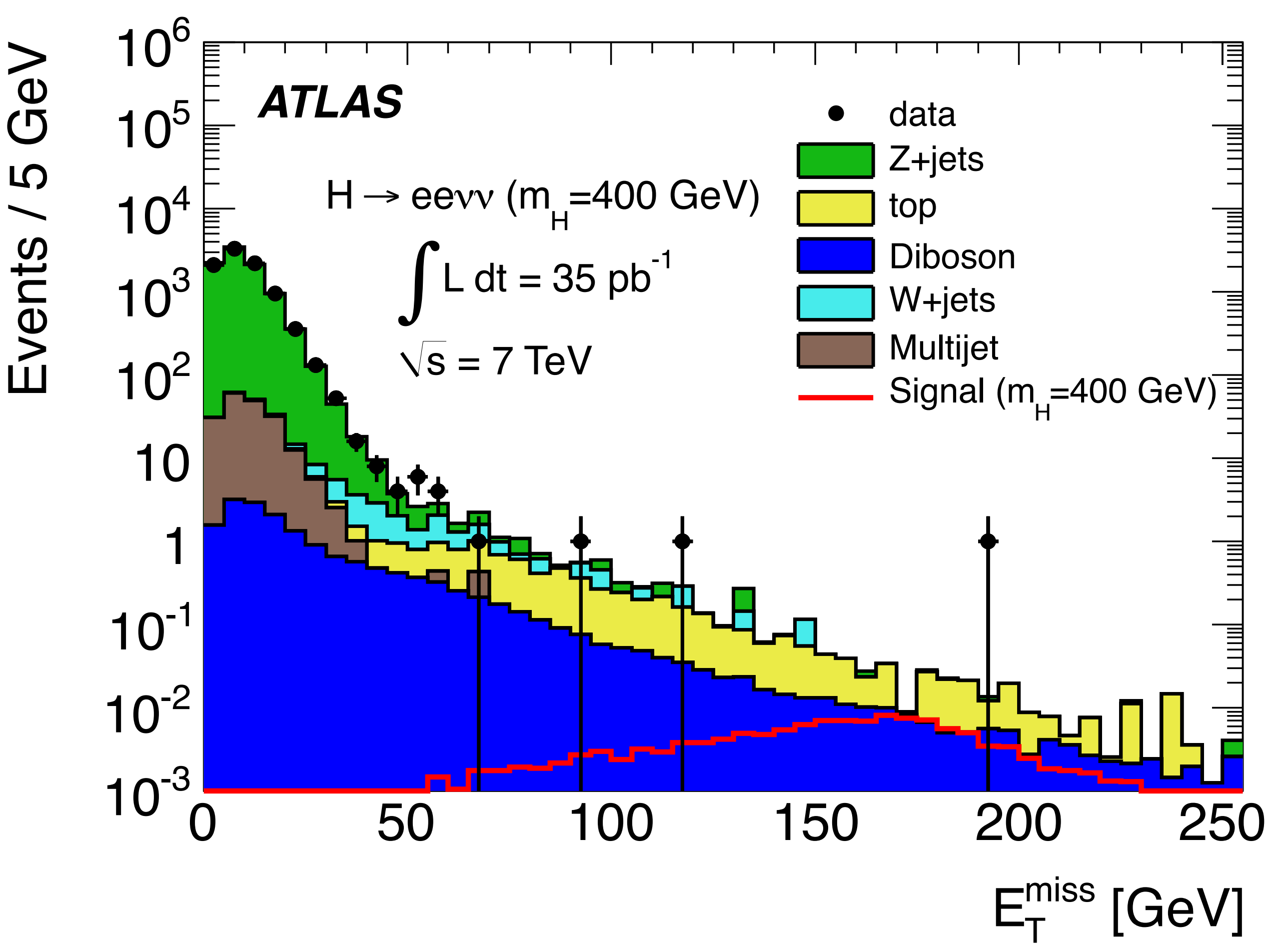
$$\mathbf{f}(\mathcal{D}|\nu) = \mathrm{Pois}(n|\nu) \prod_{e=1}^{n} f(x_e)$$

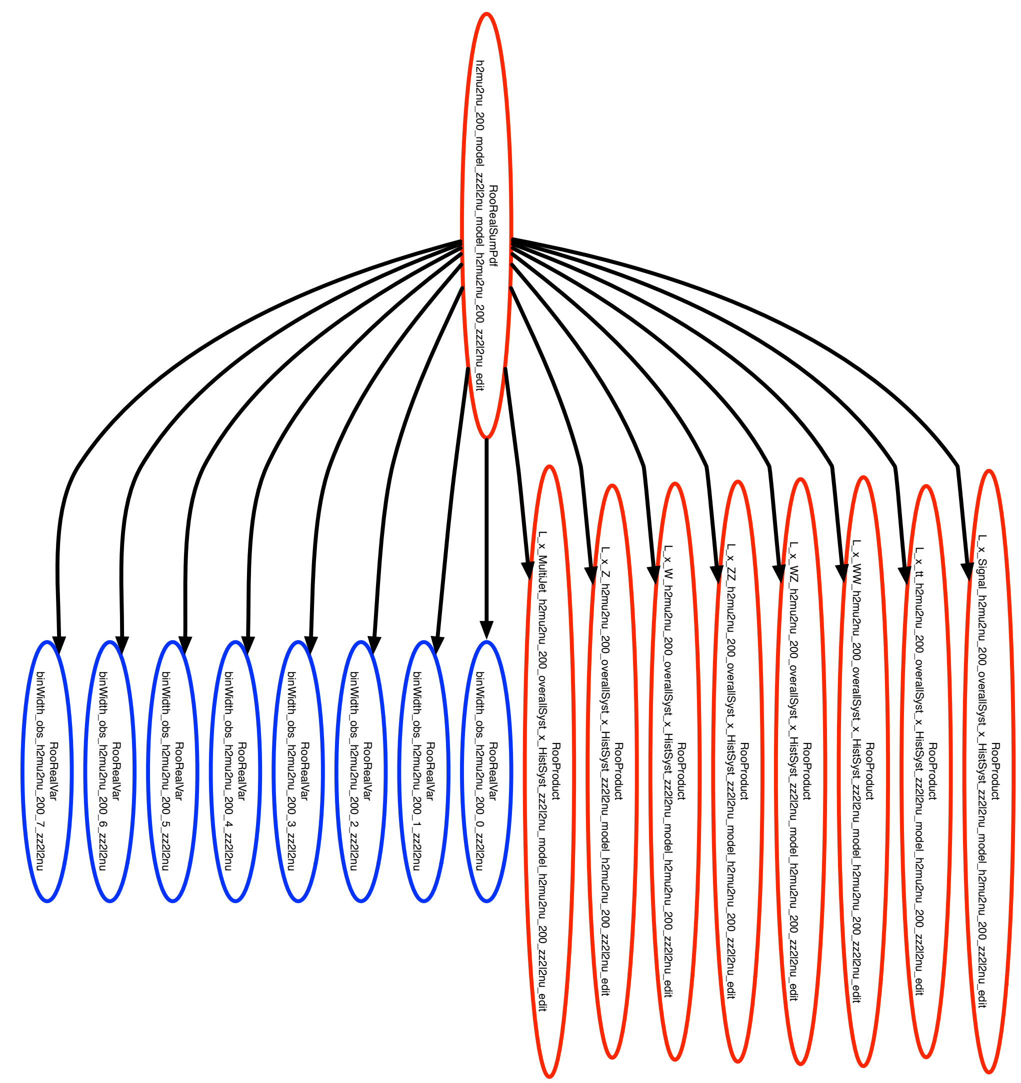
MIXTURE MODEL

Sample: a sample of simulated events corresponding to particular type interaction that populates the channel.

- statisticians call this a mixture model

$$f(x) = \frac{1}{\nu_{\mathrm{tot}}} \sum_{s \in \mathrm{samples}} \nu_s f_s(x) \,, \qquad \nu_{\mathrm{tot}} = \sum_{s \in \mathrm{samples}} \nu_s$$





Parametrizing the Model $\boldsymbol{\alpha} = (\mu, \boldsymbol{\theta})$

Parameters of interest (*μ*): parameters of the theory that modify the rates and shapes of the distributions, eg.

- the mass of a hypothesized particle
- the "signal strength" μ=0 no signal, μ=1 predicted signal rate

Nuisance parameters (*θ* or α_p): associated to uncertainty in:

- response of the detector (calibration)
- phenomenological model of interaction in non-perturbative regime

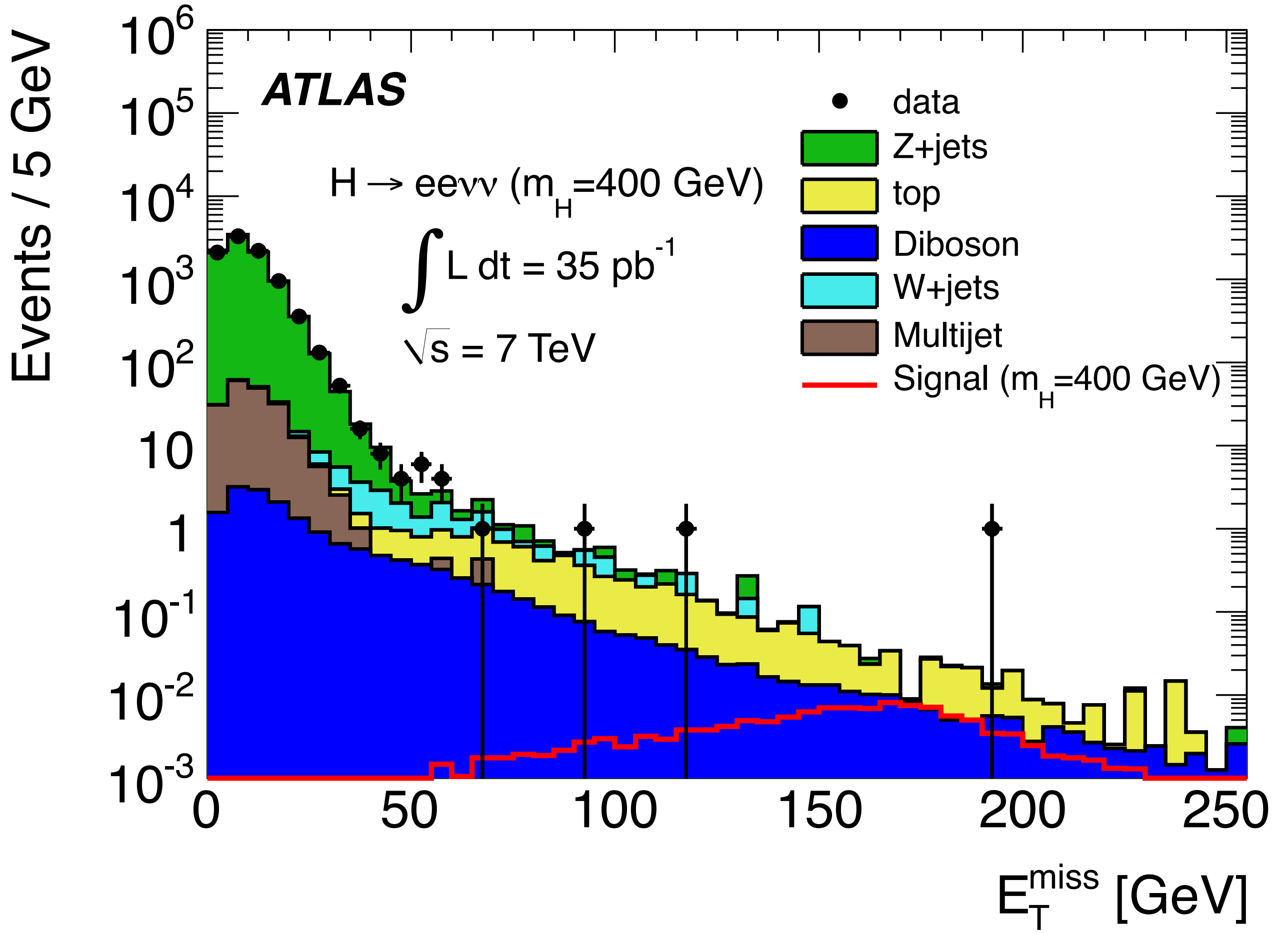
Lead to a parametrized model: $\nu \to \nu(\boldsymbol{\alpha}), f(x) \to f(x|\boldsymbol{\alpha})$

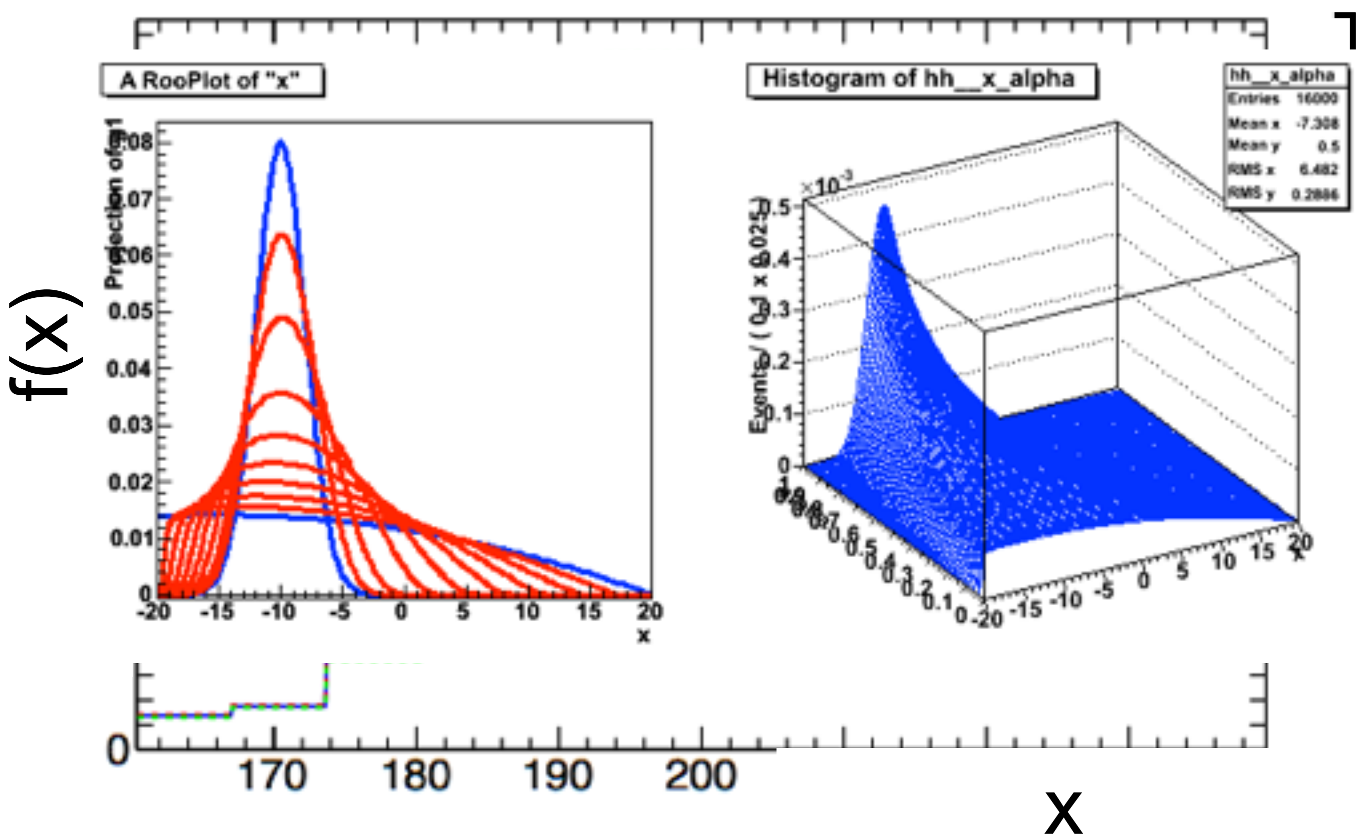
$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

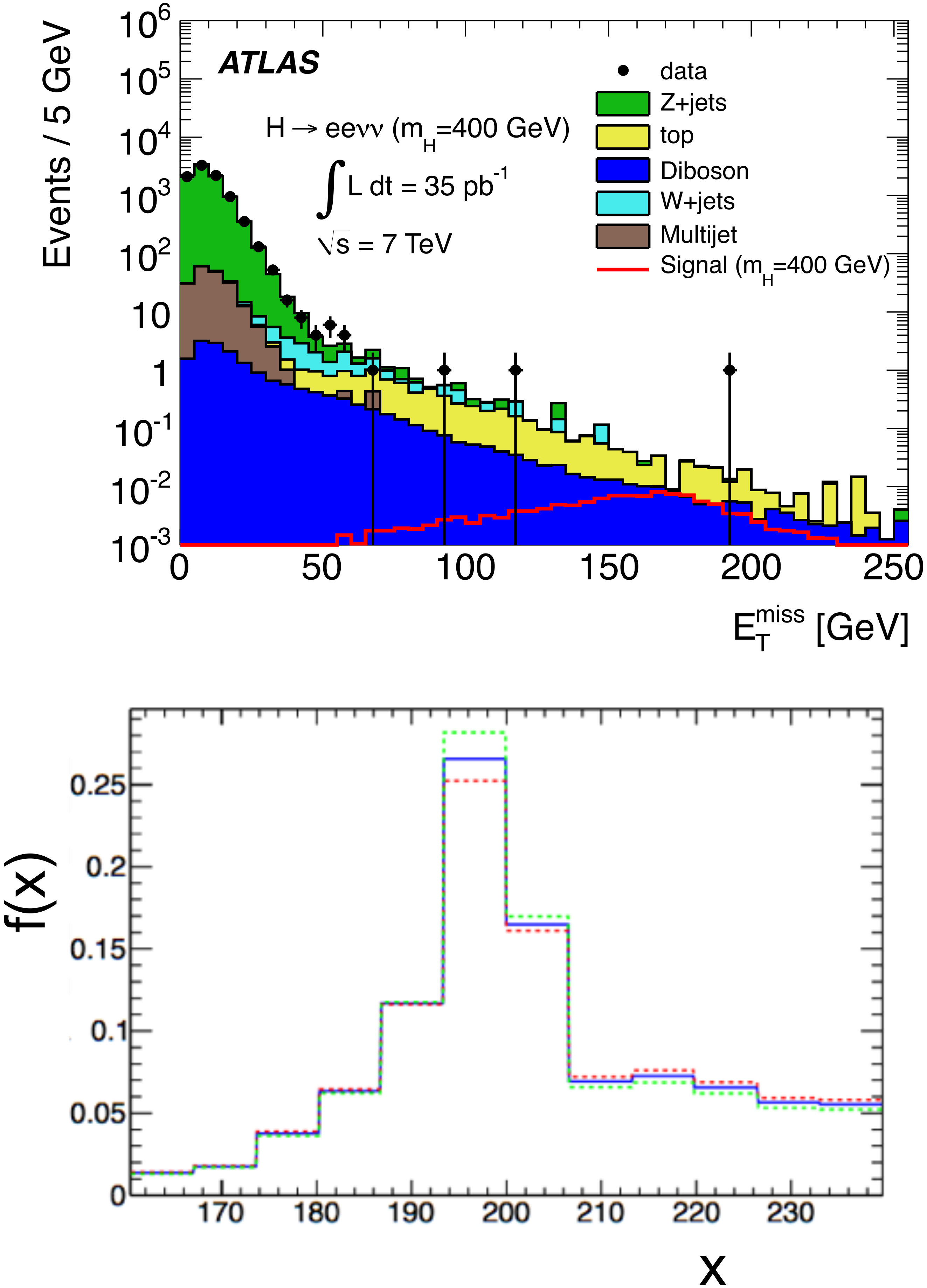
- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

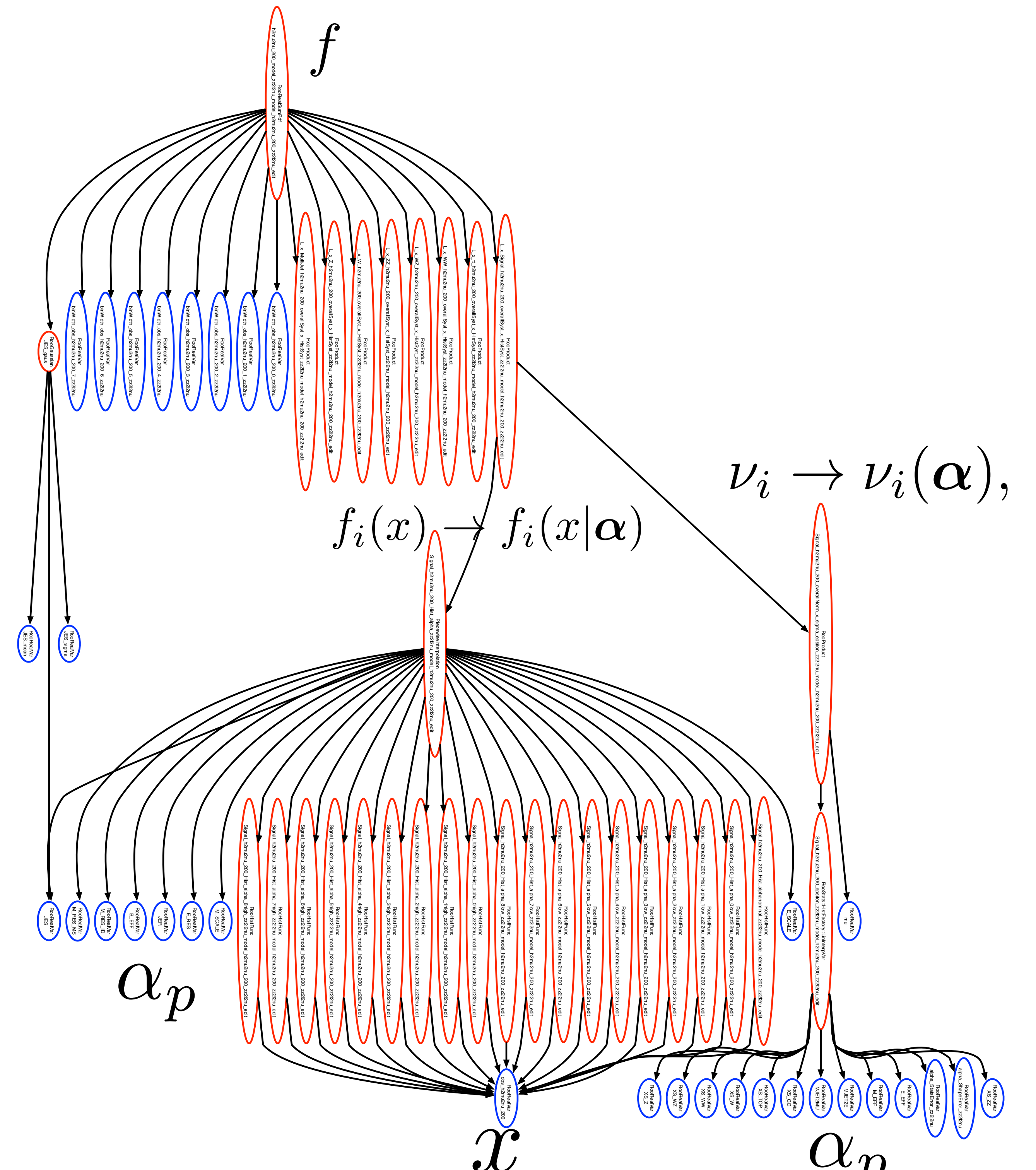




$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

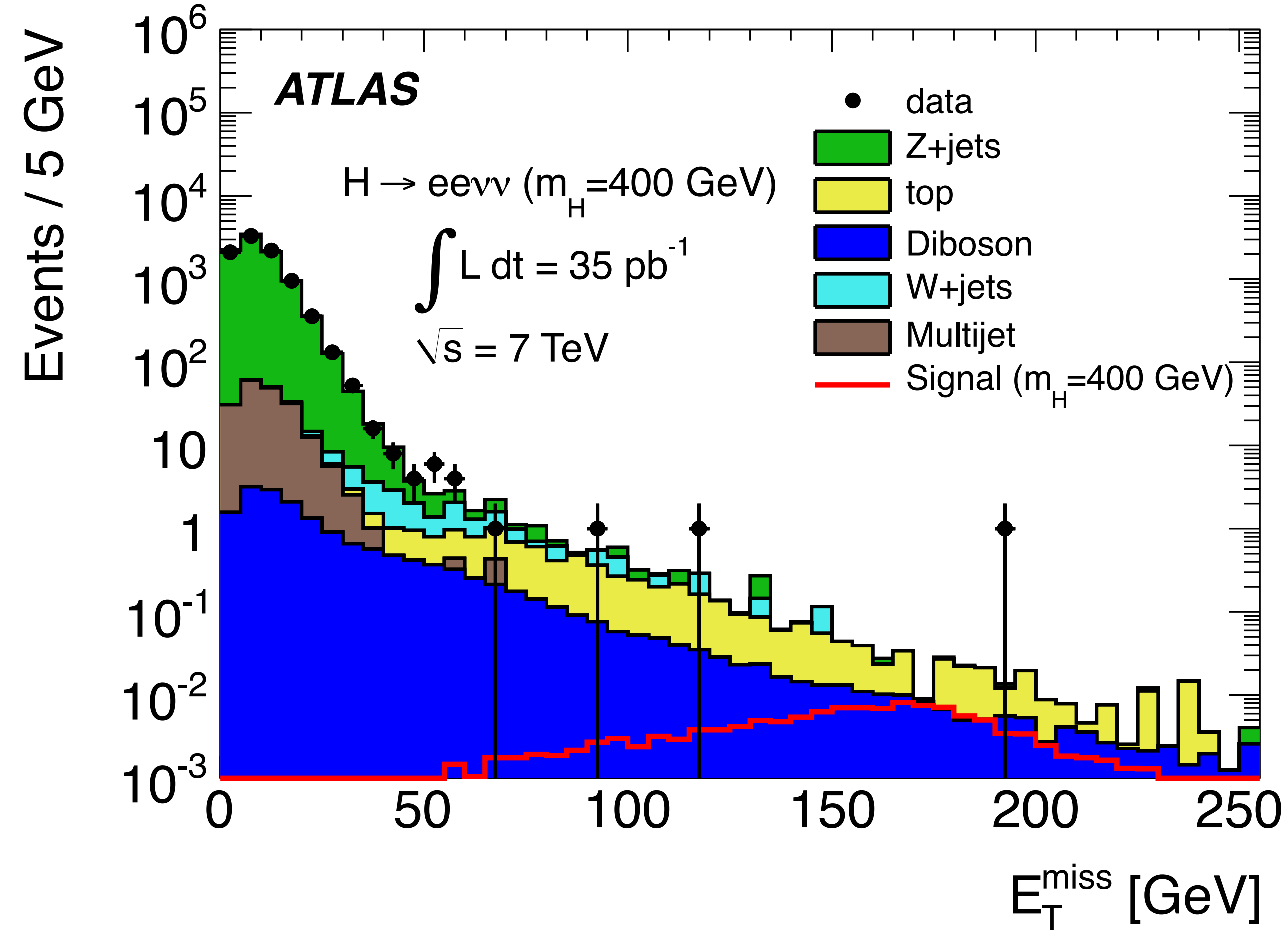
VISUALIZING THE MODEL FOR ONE CHANNEL

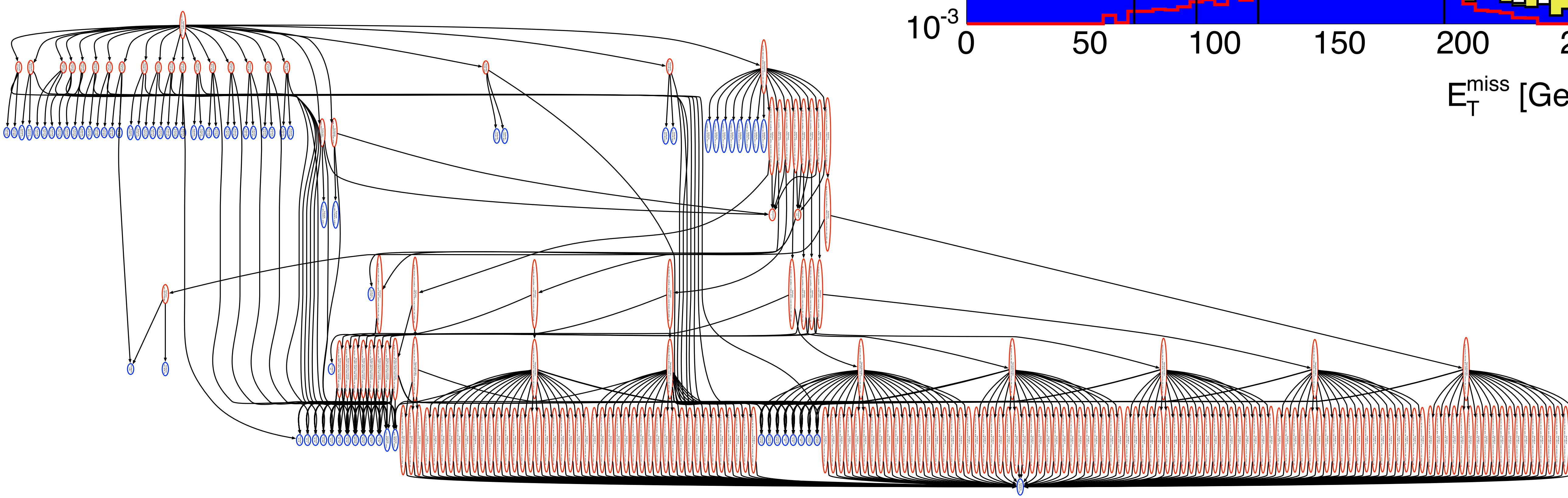




VISUALIZING THE MODEL FOR ONE CHANNEL

After parametrizing each component of the mixture model, the pdf for a single channel might look like this





Simultaneous Multi-Channel Model

Simultaneous Multi-Channel Model: Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

$$\mathbf{f}_{\rm sim}(\mathcal{D}_{\rm sim}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right]$$

where $\mathcal{D}_{\rm sim} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\rm max}}\}$

Control Regions: Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

- attempt to describe systematics in a statistical language
- Prototypical Example: "on/off" problem with unknown ν_b

$$\mathbf{f}(n,m|\mu,\nu_b) = \underbrace{\text{Pois}(n|\mu+\nu_b)}_{\text{signal region}} \cdot \underbrace{\text{Pois}(m|\tau\nu_b)}_{\text{control region}}$$

Constraint terms

Often detailed statistical model for auxiliary measurements that measure certain nuisance parameters are not available.

- one typically has MLE for α_p, denoted a_p and standard error

Constraint Terms: are idealized pdfs for the MLE.

$$f_p(a_p|\alpha_p) \quad \text{for} \quad p \in \mathbb{S}$$

- common choices are Gaussian, Poisson, and log-normal
- New: careful to write constraint term a frequentist way
- Previously: $\pi(\alpha_p|a_p) = f_p(a_p|\alpha_p)\eta(\alpha_p)$ with uniform η

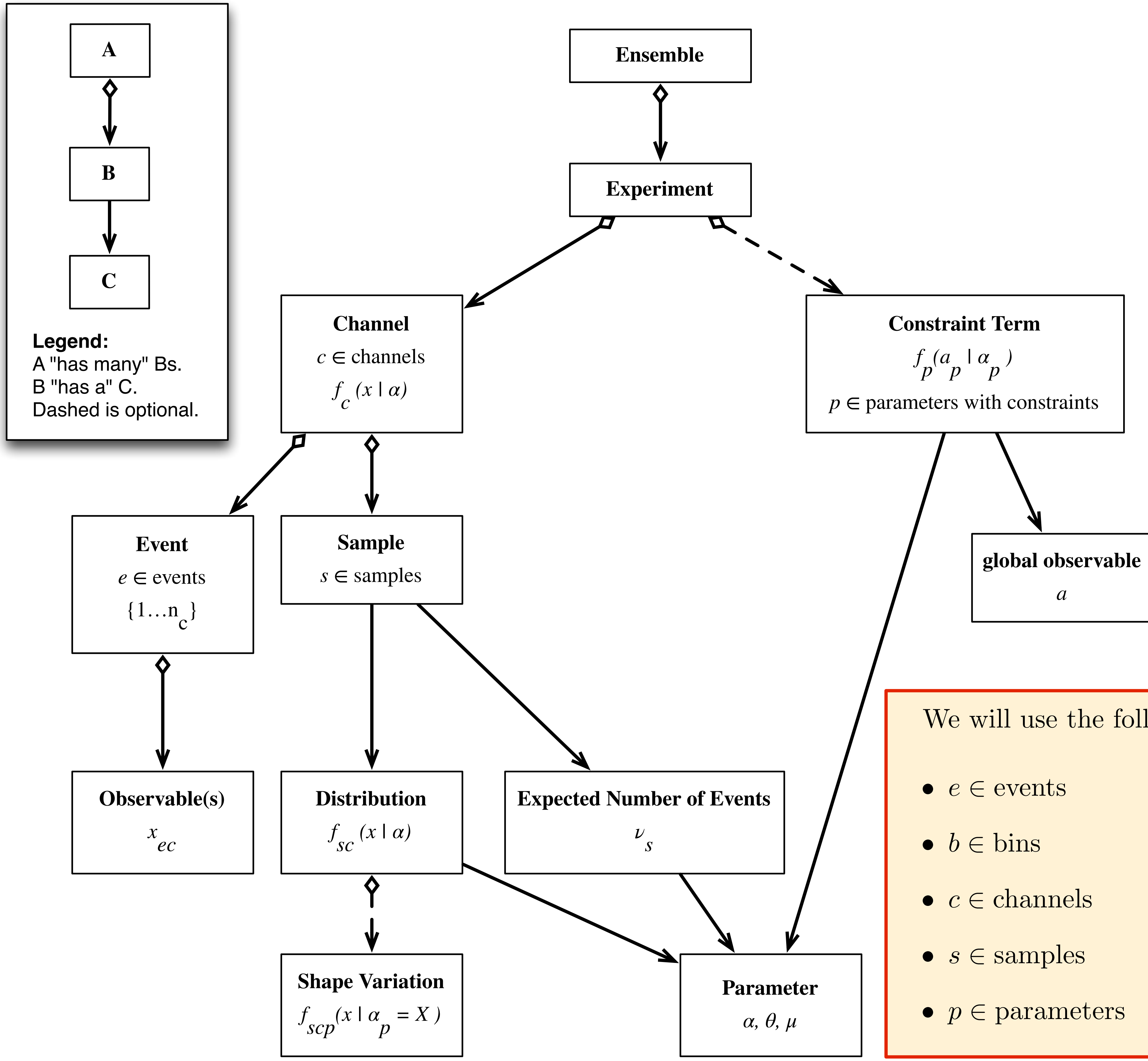
Simultaneous Multi-Channel Model with constraints:

$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c \in \text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right] \cdot \prod_{p \in \mathbb{S}} f_p(a_p|\alpha_p)$$

where

$$\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\max}}\}, \quad \mathcal{G} = \{a_p\} \quad \text{for} \quad p \in \mathbb{S}$$

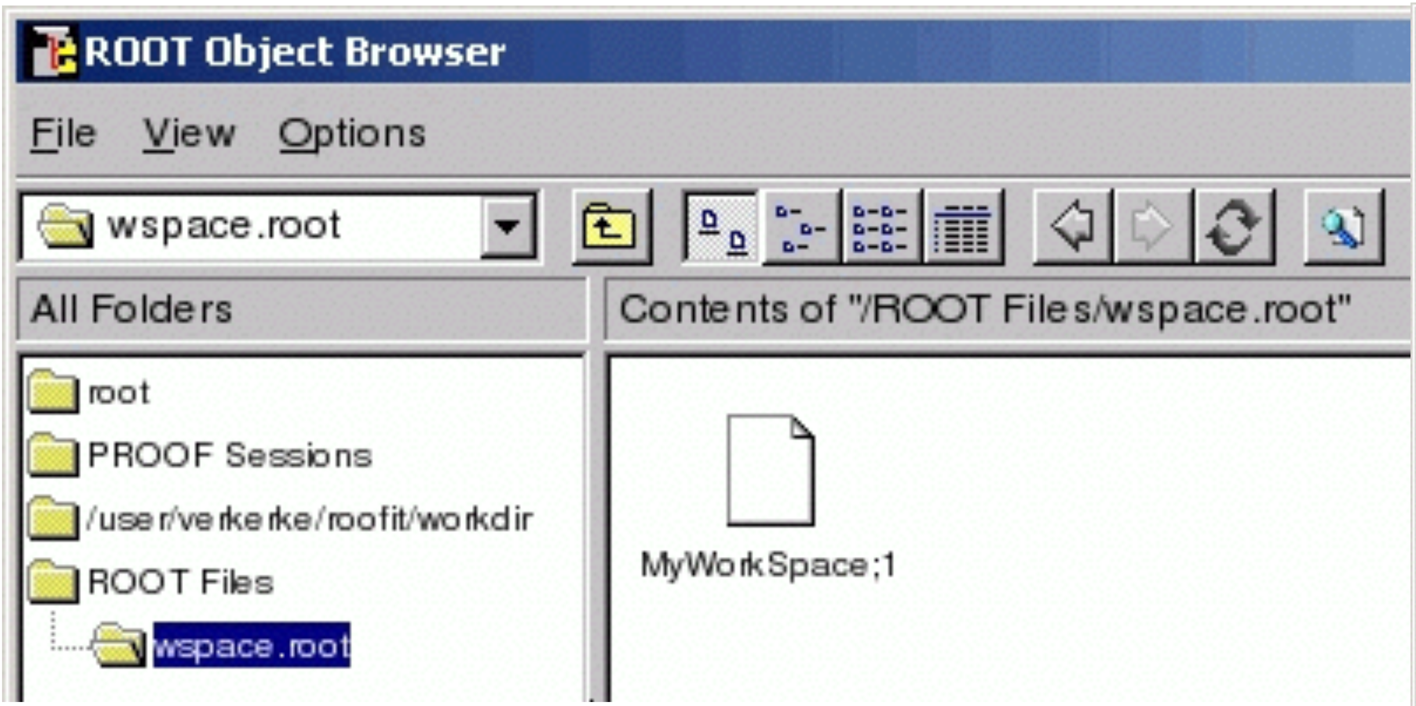
Conceptual building blocks

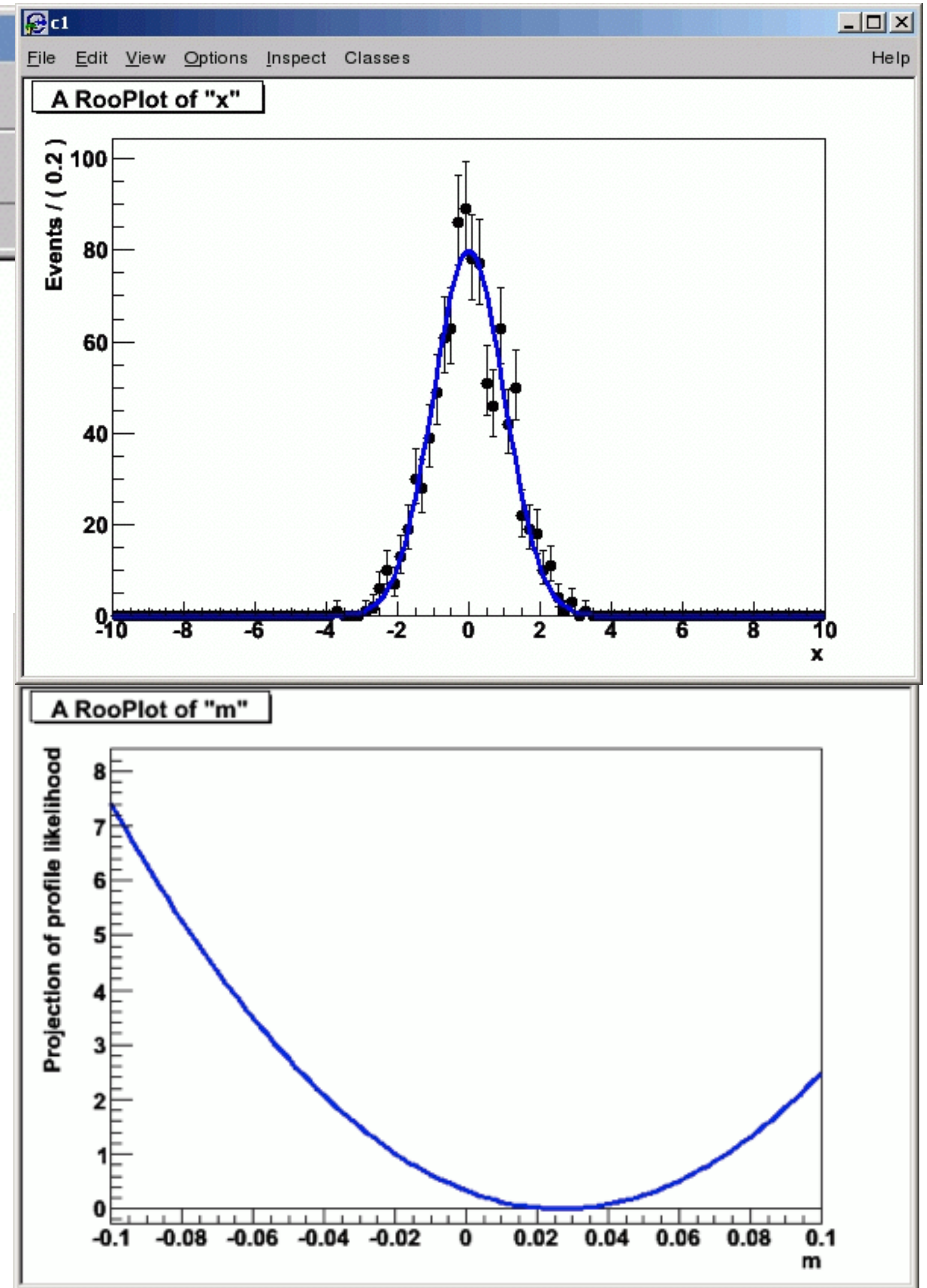


We will use the following mnemonic index conventions:

- $e \in$ events
- $b \in$ bins
- $c \in$ channels
- $s \in$ samples
- $p \in$ parameters

EXAMPLE OF DIGITAL PUBLISHING





RooFit's Workspace now provides the ability to save in a ROOT file the full likelihood model, any priors you might want, and the minimal data necessary to reproduce likelihood function.

Need this for combinations, as p-value is not sufficient information for a proper combination.

HistFactory

32 page documentation of HistFactory tool + manual

- currently a “living document”

Information | Discussion (0) | Files | Linkbacks

http://cds.cern.ch/record/1456844

Preprint

Report number	CERN-OPEN-2012-016
Title	**HistFactory: A tool for creating statistical models for use with RooFit and RooStats**
Author(s)	Cranmer, Kyle (New York U.) ; Lewis, George (New York U.) ; Moneta, Lorenzo (CERN) ; Shibata, Akira (New York U.) ; Verkerke, Wouter (NIKHEF, Amsterdam)
Collaboration	ROOT Collaboration
Abstract	The HistFactory is a tool to build parametrized probability density functions (pdfs) in the RooFit/RooStats framework based based on simple ROOT histograms organized in an XML file. The pdf has a restricted form, but it is sufficiently flexible to describe many analyses based on template histograms. The tool takes a modular approach to build complex pdfs from more primative conceptual building blocks. The resulting PDF is stored in a RooWorkspace which can be saved to and read from a ROOT file. This document describes the defaults and interface in HistFactory 5.32.

Combined ATLAS Higgs Search

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

- Models for individual channels come from about 11 sub-groups performing dedicated searches for specific Higgs decay modes
- In addition low-level performance groups provide tools for evaluating systematic effects and corresponding constraint terms

Higgs Decay	Subsequent Decay	Additional Sub-Channels	m_H Range	L [fb^{-1}]
$H \to \gamma\gamma$	–	9 sub-channels ($p_{\mathrm{T}_t} \otimes \eta_\gamma \otimes$ conversion)	110-150	4.9
$H \to ZZ$	$\ell\ell\ell'\ell'$	$\{4e, 2e2\mu, 2\mu 2e, 4\mu\}$	110-600	4.8
	$\ell\ell\nu\nu$	$\{ee, \mu\mu\} \otimes$ {low pile-up, high pile-up}	200-280-600	4.7
	$\ell\ell qq$	{b-tagged, untagged}	200-300-600	4.7
$H \to WW$	$\ell\nu\ell\nu$	$\{ee, e\mu, \mu\mu\} \otimes$ {0-jet, 1-jet, VBF}	110-300-600	4.7
	$\ell\nu q\overline{q'}$	$\{e, \mu\} \otimes$ {0-jet, 1-jet}	300-600	4.7
$H \to \tau^+\tau^-$	$\ell\ell 4\nu$	$\{e\mu\} \otimes$ {0-jet} $\oplus$ {1-jet, VBF, VH}	110-150	4.7
	$\ell\tau_{\mathrm{had}} 3\nu$	$\{e, \mu\} \otimes$ {0-jet} $\otimes$ $\{E_T^{\mathrm{miss}} \gtrless 20$ GeV$\}$ $\oplus$ $\{e, \mu\} \otimes$ {1-jet, VBF}	110-150	4.7
	$\tau_{\mathrm{had}}\tau_{\mathrm{had}} 2\nu$	{1-jet}	110-150	4.7
$VH \to b\overline{b}$	$Z \to \nu\overline{\nu}$	$E_T^{\mathrm{miss}} \in \{120-160, 160-200, \geq 200 \text{ GeV}\}$	110-130	4.6
	$W \to \ell\nu$	$p_T^W \in \{< 50, 50-100, 100-200, \geq 200 \text{ GeV}\}$	110-130	4.7
	$Z \to \ell\ell$	$p_T^Z \in \{< 50, 50-100, 100-200, \geq 200 \text{ GeV}\}$	110-130	4.7

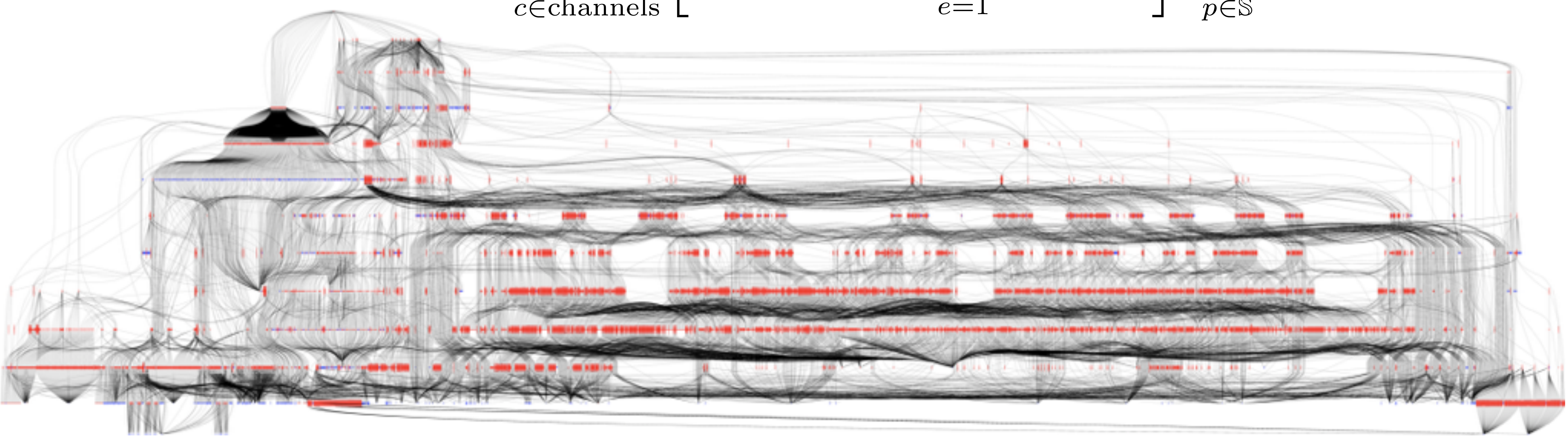
VISUALIZING THE COMBINED MODEL

State of the art: At the time of the discovery, the combined Higgs search included 100 disjoint channels and >500 nuisance parameters

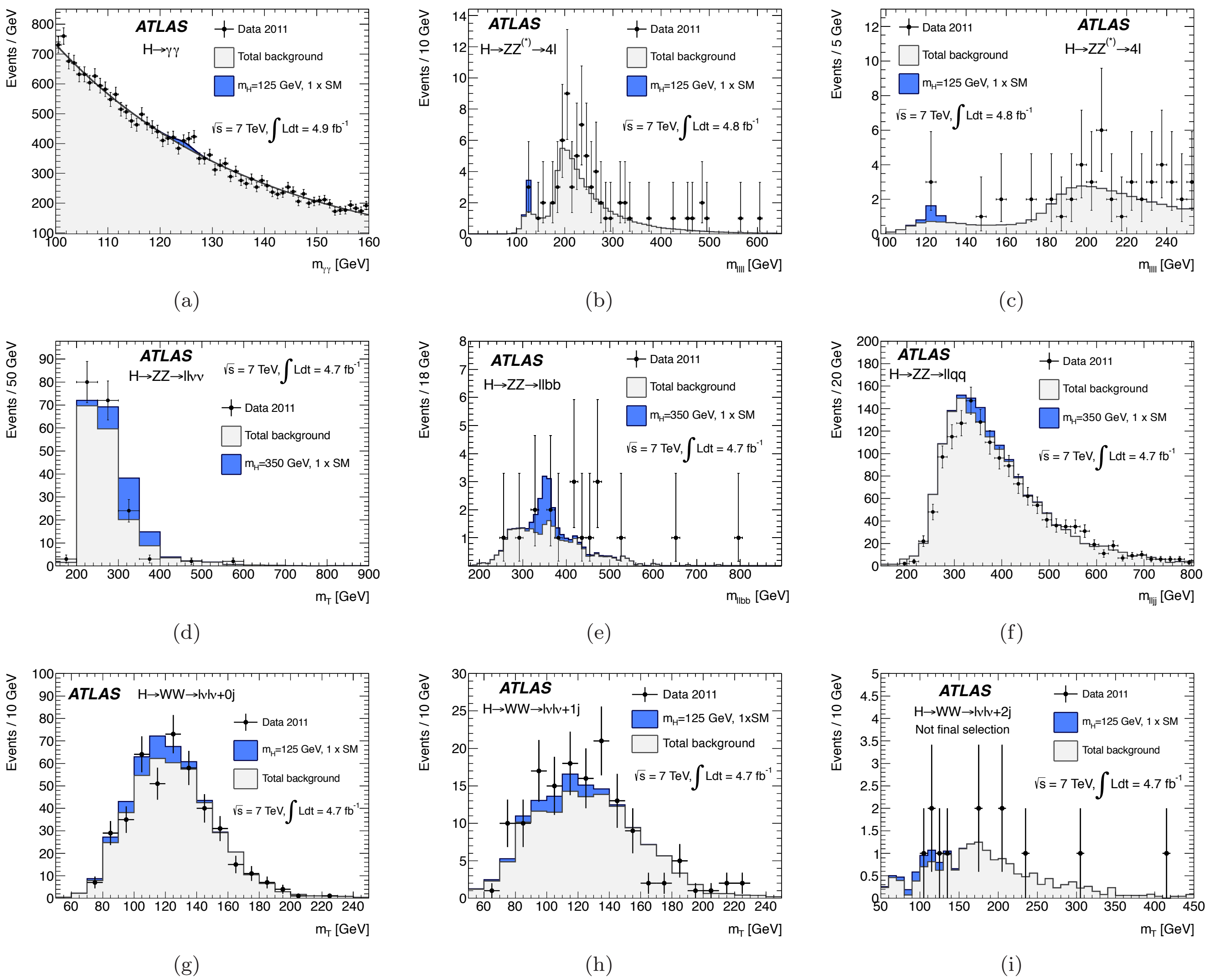
RooFit / RooStats: is the modeling language (C++) which provides technologies for collaborative modeling

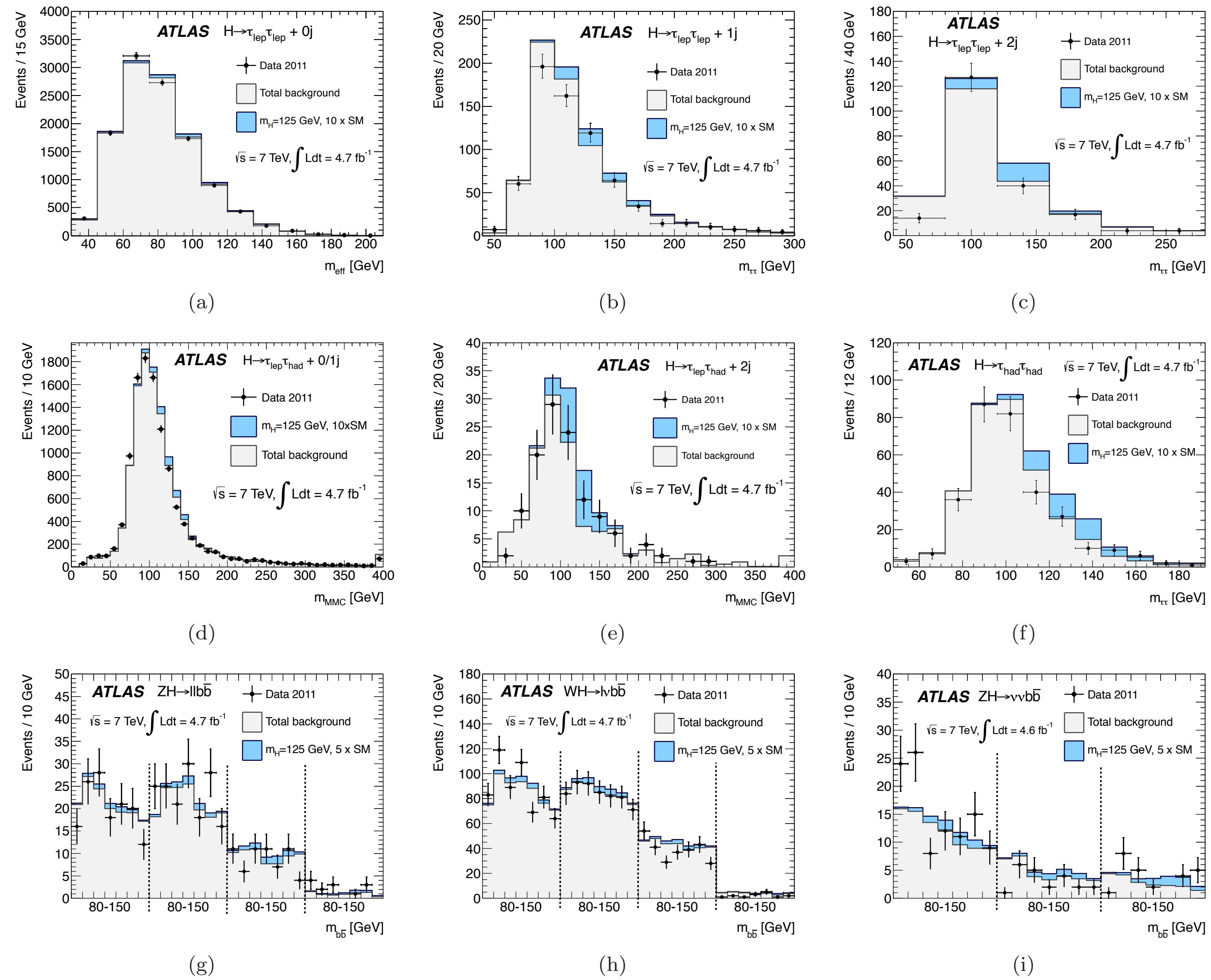
- provides technology to publish likelihood functions digitally
- and more, it's the full model so we can also generate pseudo-data

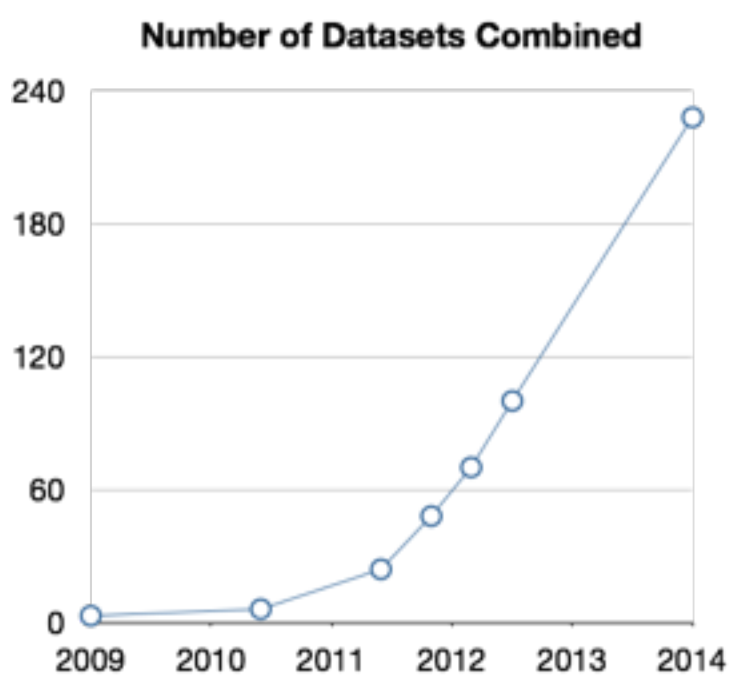
$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right] \cdot \prod_{p\in\mathbb{S}} f_p(a_p|\alpha_p)$$

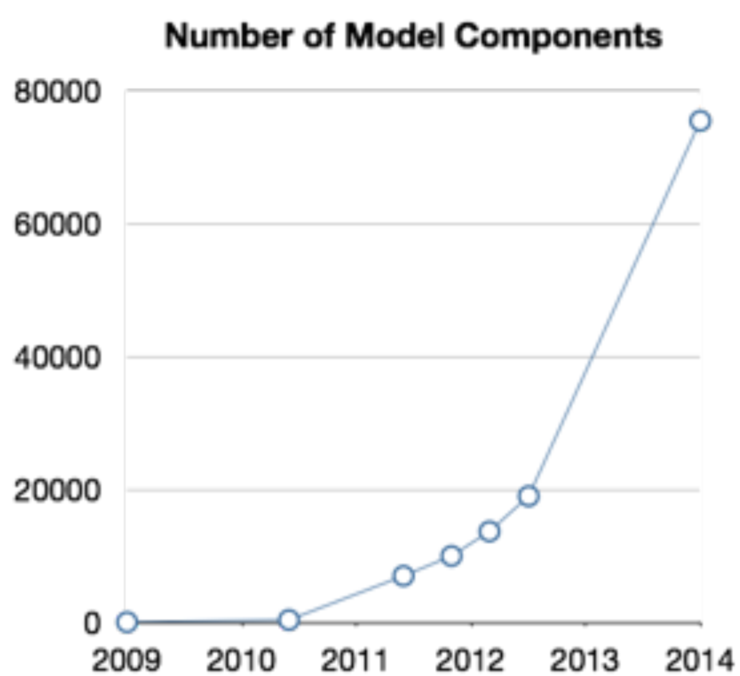


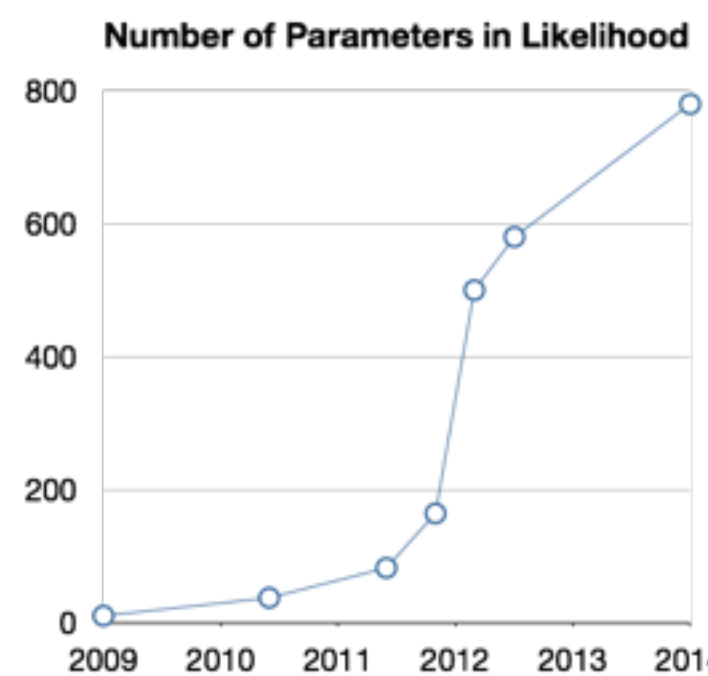
Evolution of Model Complexity







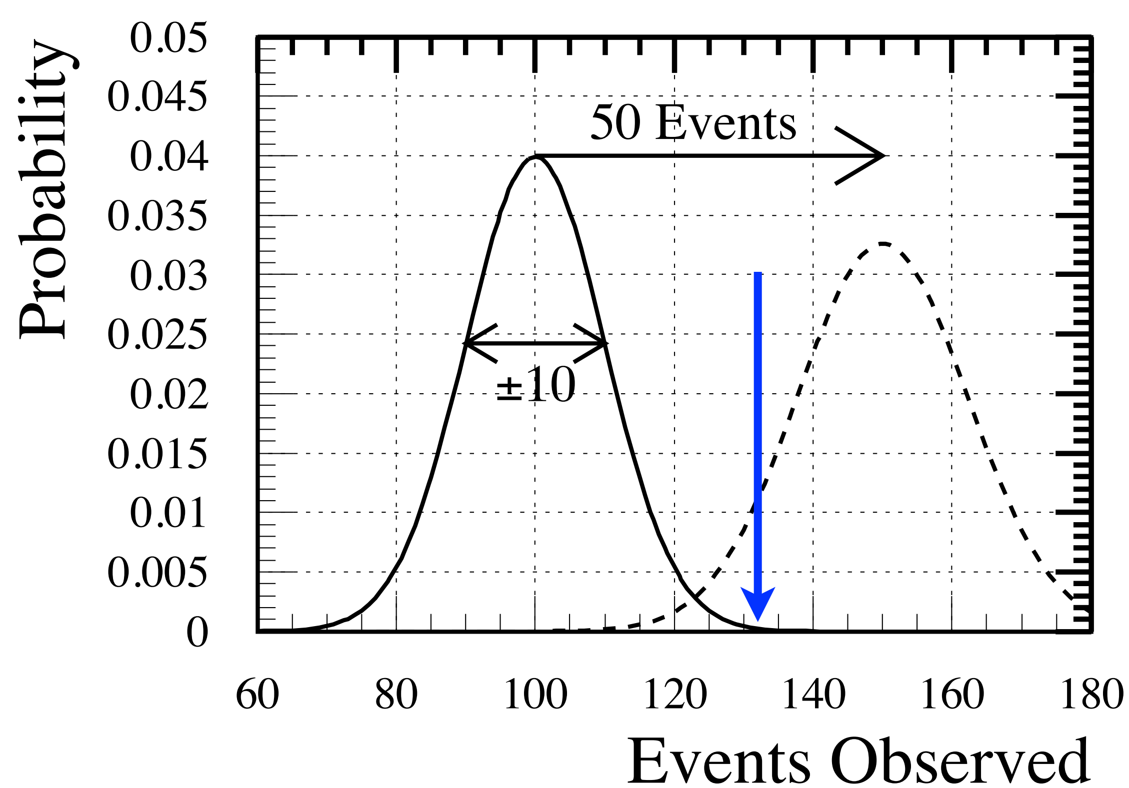




HYPOTHESIS TESTING

One of the most common uses of statistics in particle physics is Hypothesis Testing (e.g. for discovery of a new particle)

- assume one has pdf for data under two hypotheses:
  - Null-Hypothesis, H_0: eg. background-only
  - Alternate-Hypothesis H_1: eg. signal-plus-background
- one makes a measurement and then needs to decide whether to **reject** or **accept** H_0



Before we can make much progress with statistics, we need to decide what it is that we want to do.

- first let us define a few terms:
  - Rate of Type I error α
  - Rate of Type II β
  - Power = $1 - \beta$

		Actual condition	
		Guilty	Not guilty
Decision	Verdict of 'guilty'	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Treat the two hypotheses asymmetrically

- the Null is special.
  - Fix rate of Type I error, call it "the size of the test"

Now one can state "a well-defined goal"

- Maximize power for a fixed rate of Type I error

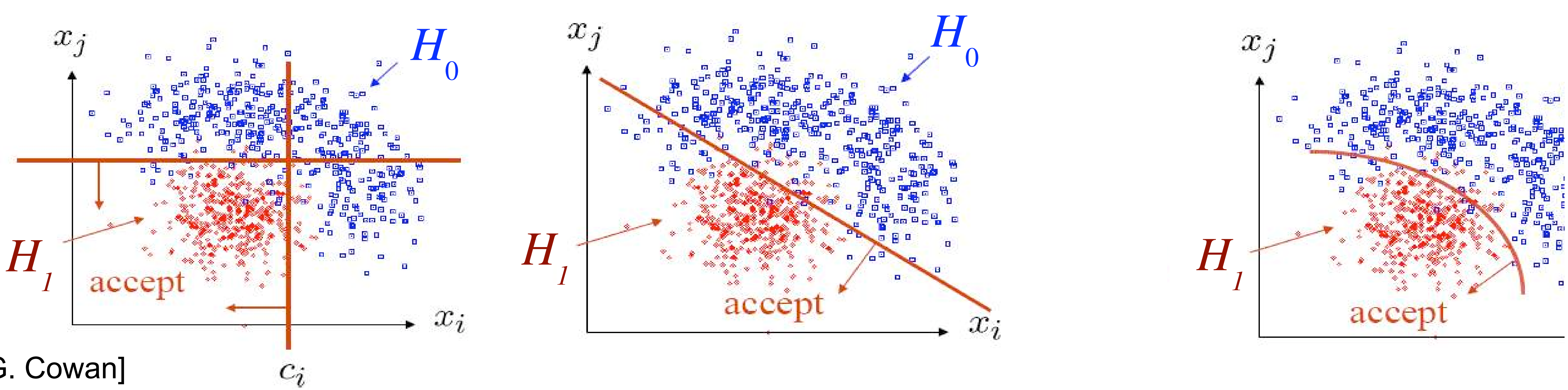
HYPOTHESIS TESTING

The idea of a “5σ” discovery criteria for particle physics is really a conventional way to specify the size of the test

- usually 5σ corresponds to $\alpha = 2.87 \cdot 10^{-7}$
  - eg. a very small chance we reject the standard model

In the simple case of number counting it is obvious what region is sensitive to the presence of a new signal

- but in higher dimensions it is not so easy



[G. Cowan]

The Neyman-Pearson Lemma

In 1928-1938 Neyman & Pearson developed a theory in which one must consider competing Hypotheses:

- the Null Hypothesis H_0 (background only)
- the Alternate Hypothesis H_1 (signal-plus-background)

Given some probability that we wrongly reject the Null Hypothesis

$$\alpha = P(x \notin W | H_0)$$

(Convention: if data falls in W then we accept H_0)

Find the region W such that we minimize the probability of wrongly accepting the H_0 (when H_1 is true)

$$\beta = P(x \in W | H_1)$$

The Neyman-Pearson Lemma

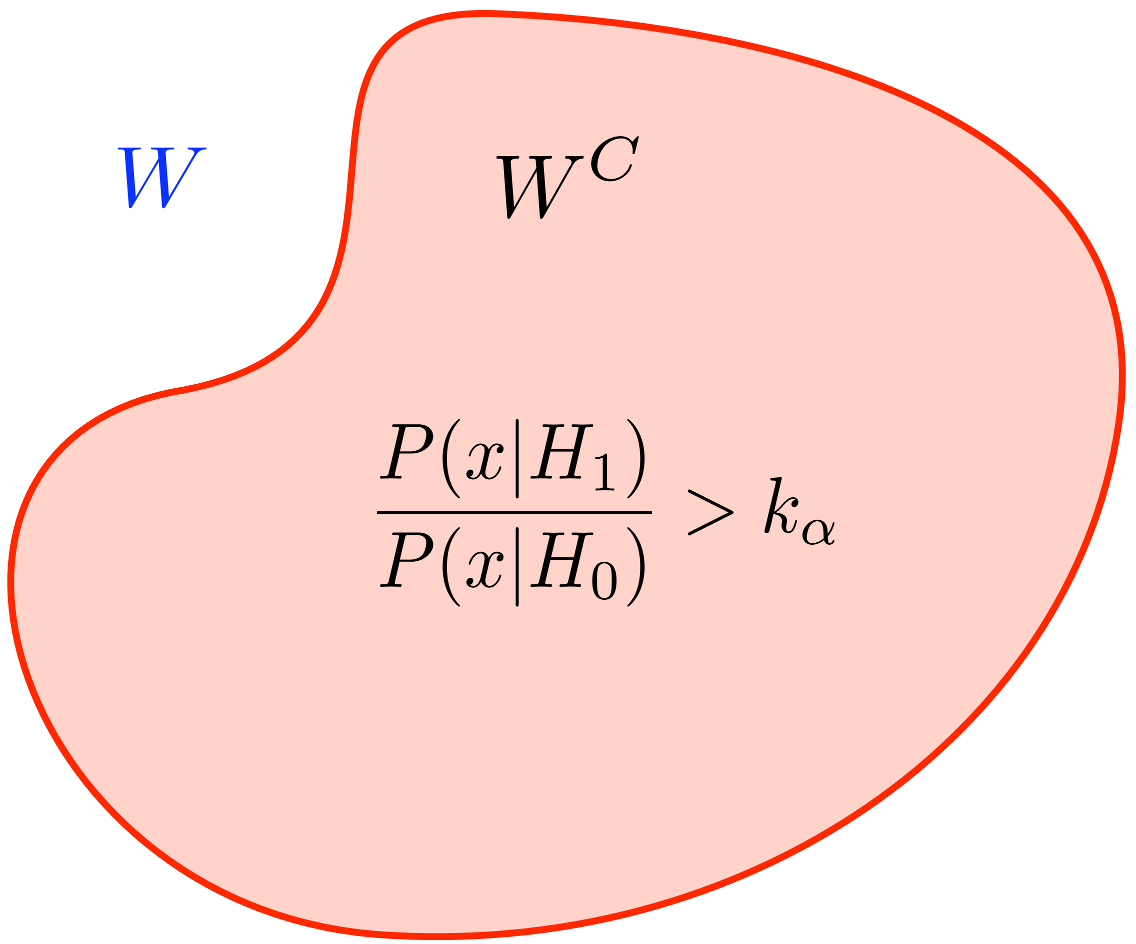
The region W that minimizes the probability of wrongly accepting H_0 is just a contour of the Likelihood Ratio

$$\frac{P(x|H_1)}{P(x|H_0)} > k_\alpha$$

Any other region of the same size will have less power

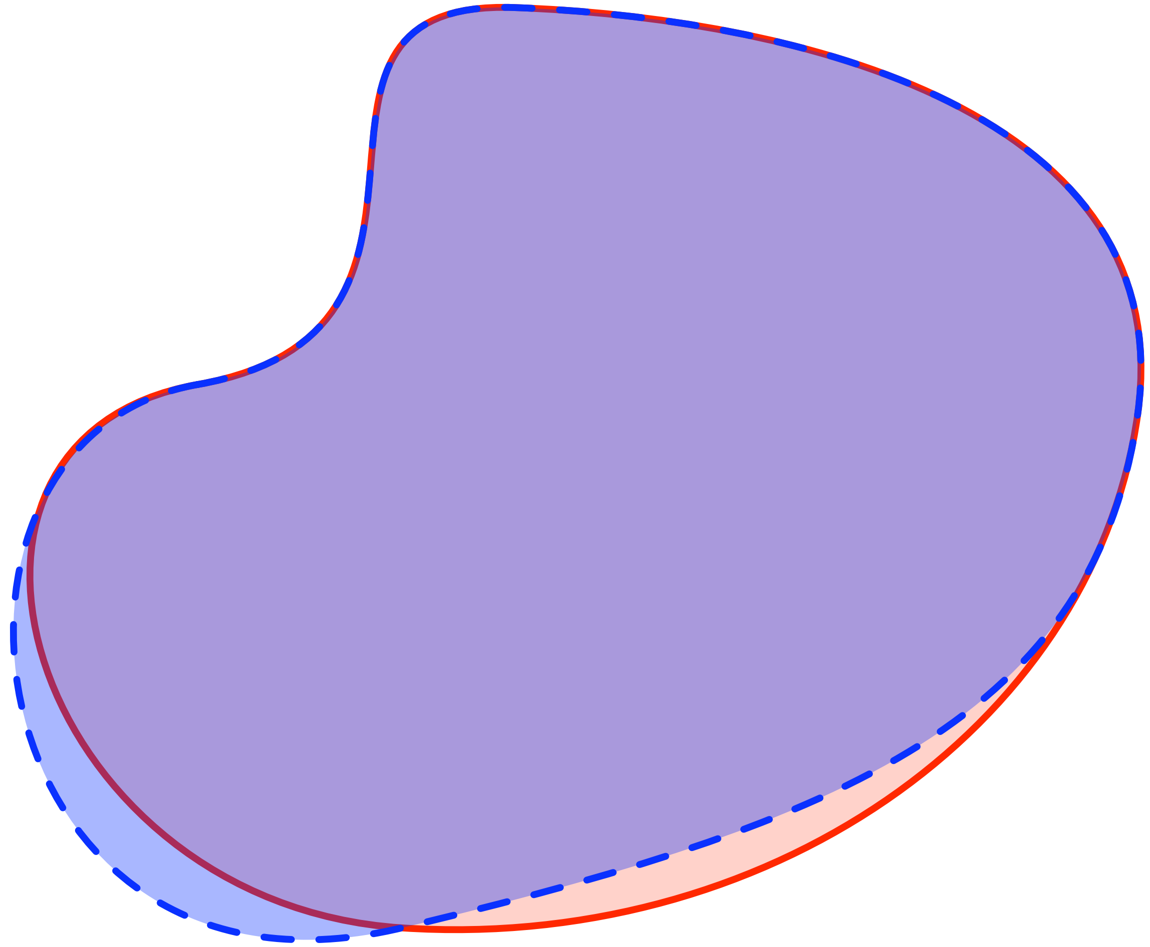
The likelihood ratio is an example of a **Test Statistic**, eg. a real-valued function that summarizes the data in a way relevant to the hypotheses that are being tested

A SHORT PROOF OF NEYMAN-PEARSON



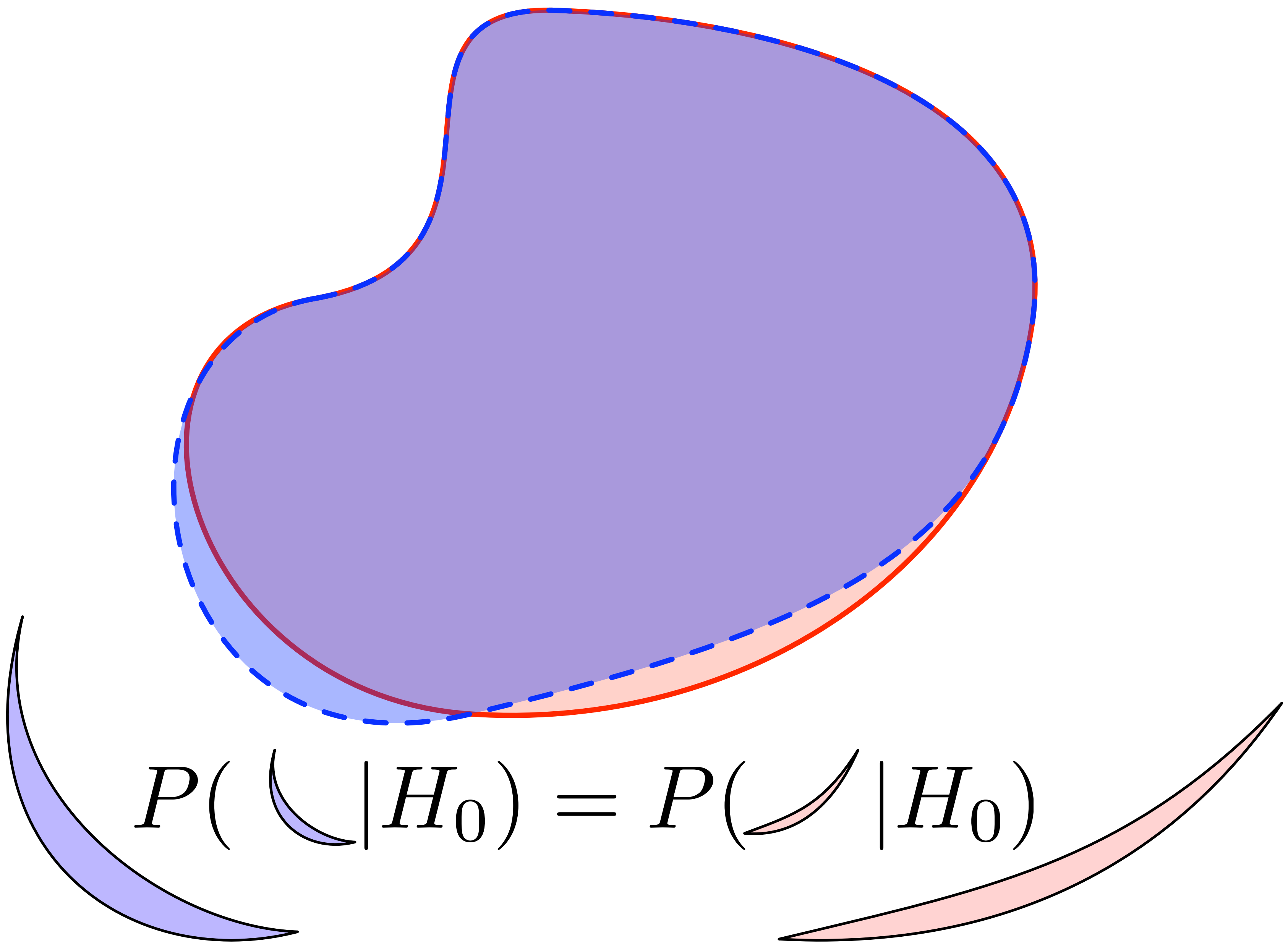
Consider the contour of the likelihood ratio that has size a given size (eg. probability under H_0 is 1-α)

A SHORT PROOF OF NEYMAN-PEARSON



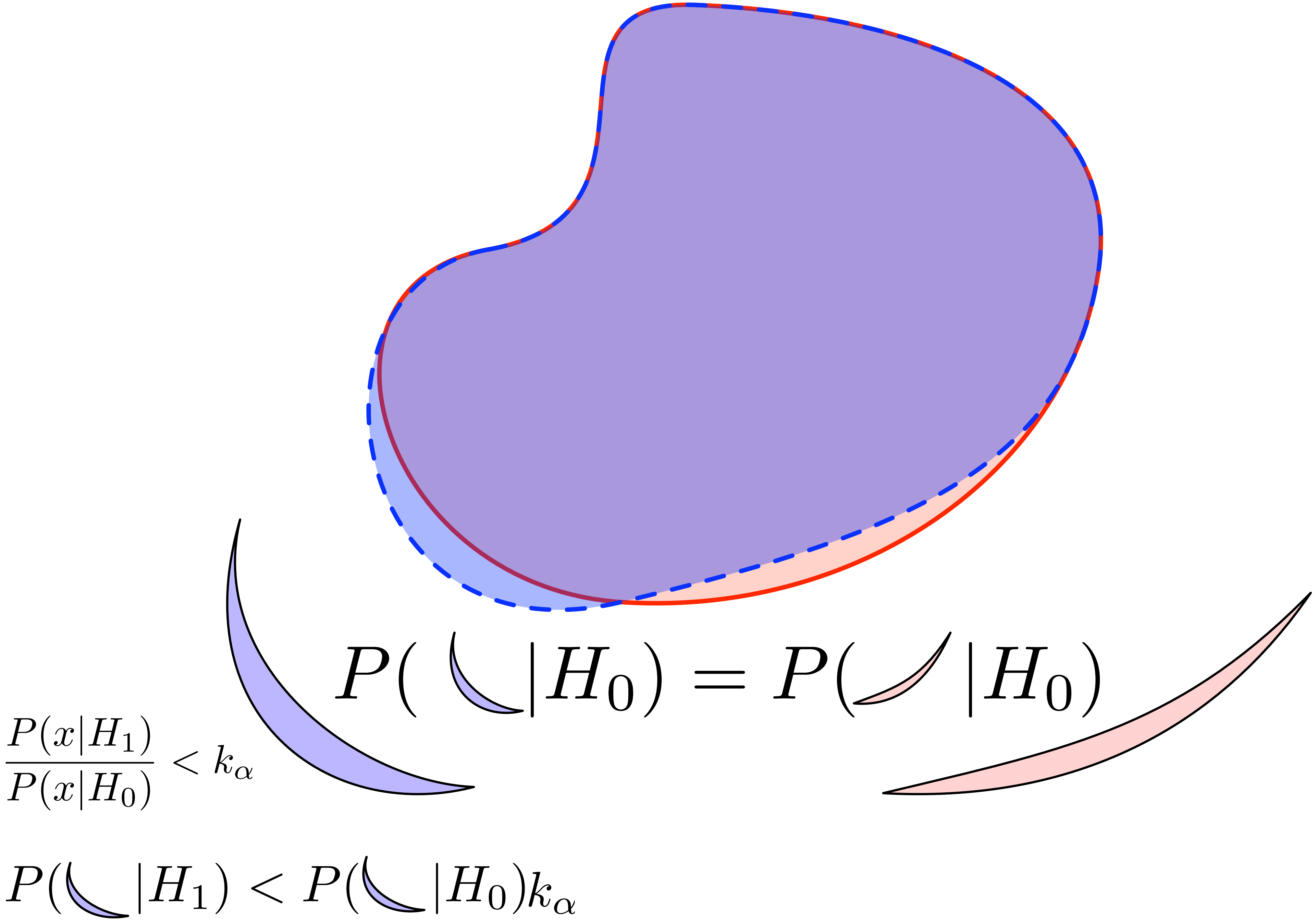
Now consider a variation on the contour that has the same size

A SHORT PROOF OF NEYMAN-PEARSON



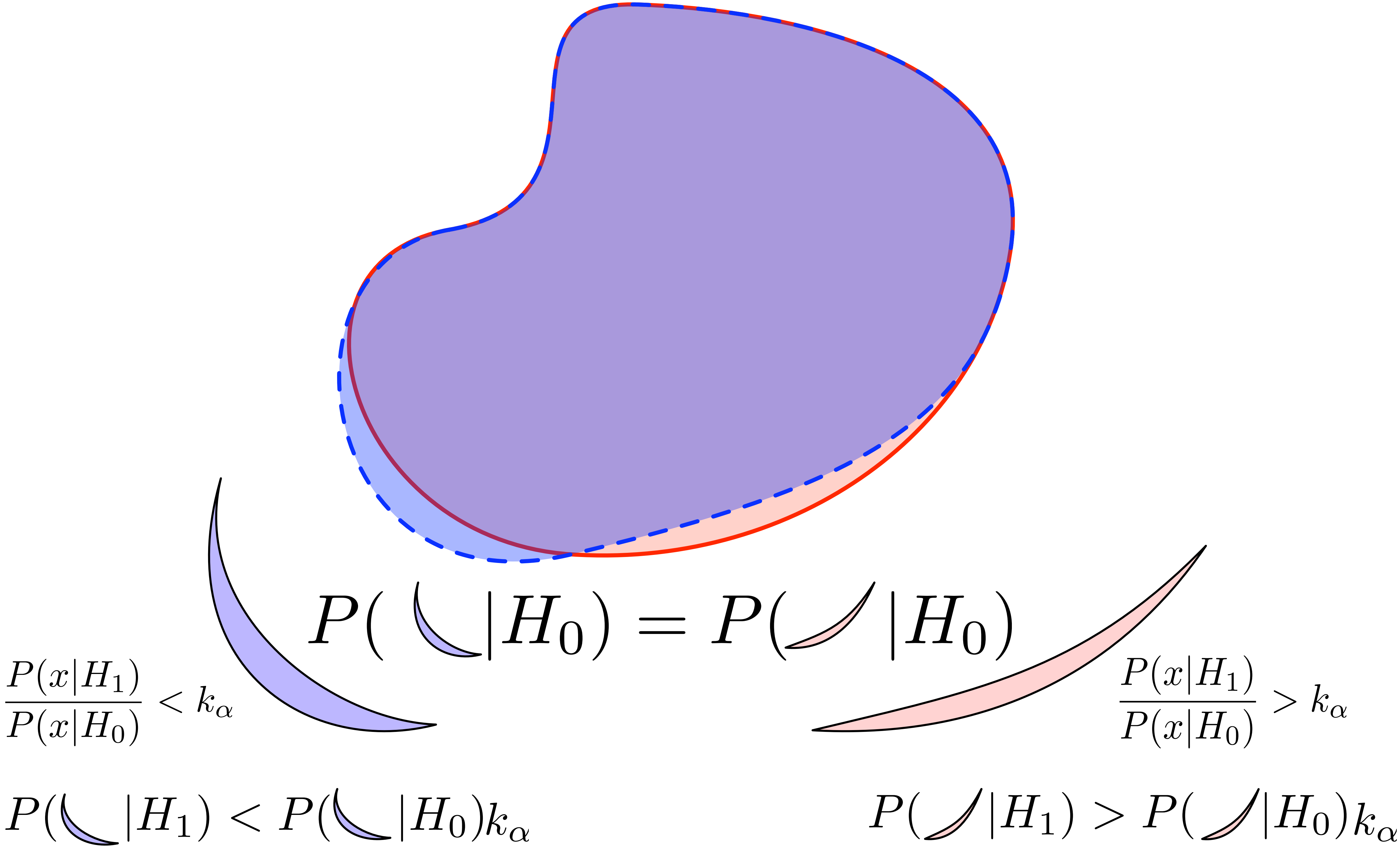
Now consider a variation on the contour that has the same size (eg. same probability under H_0)

A SHORT PROOF OF NEYMAN-PEARSON



Because the new area is outside the contour of the likelihood ratio, we have an inequality

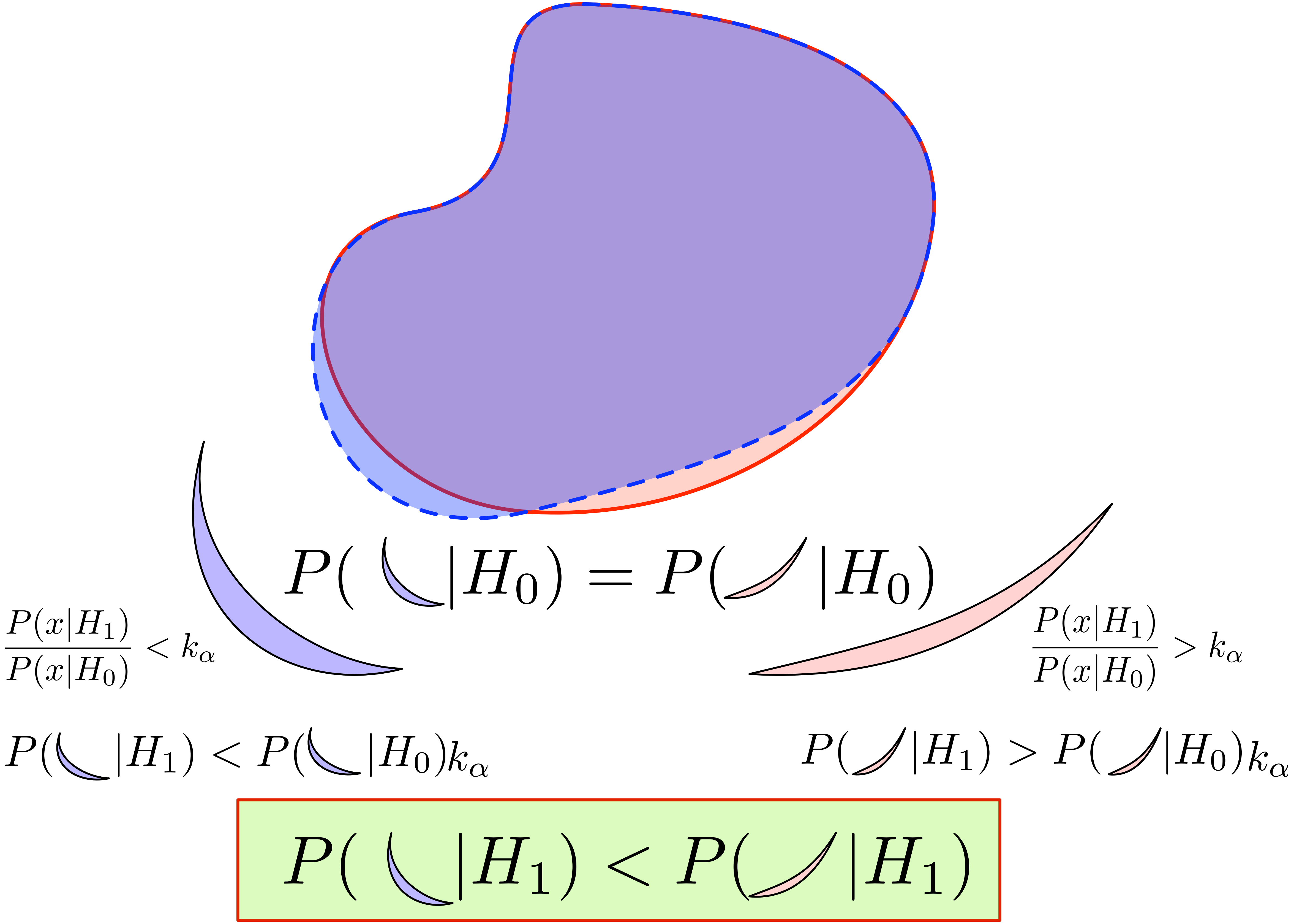
A short proof of Neyman-Pearson



And for the region we lost, we also have an inequality

Together they give...

A SHORT PROOF OF NEYMAN-PEARSON

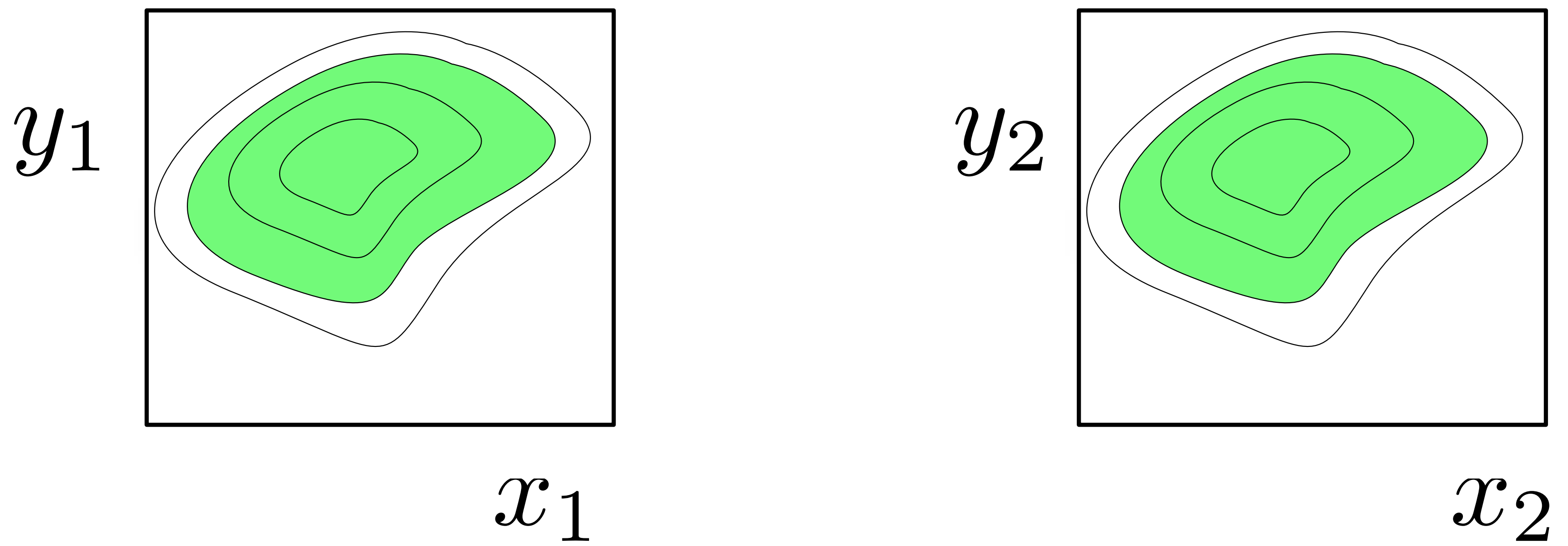


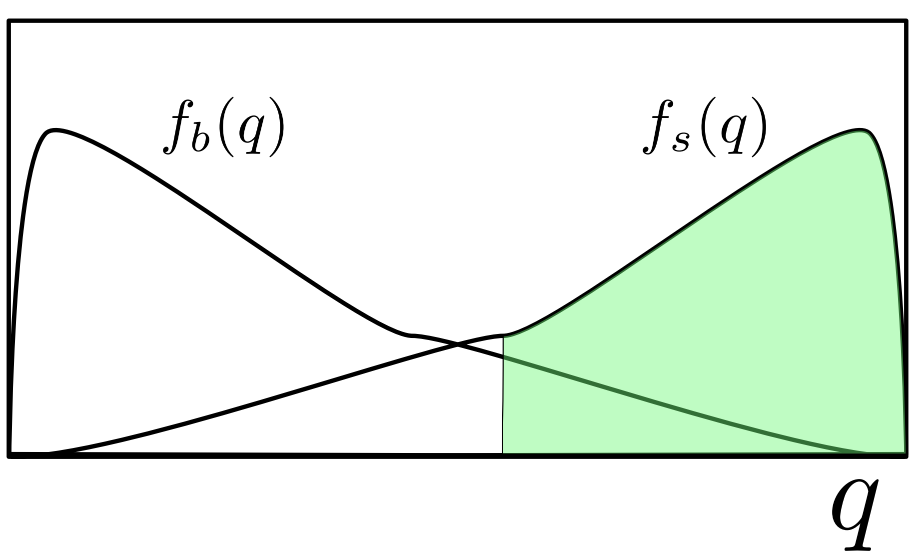
The new region region has less power.

2 DISCRIMINATING VARIABLES

Often one uses the output of a neural network or multivariate algorithm in place of a true likelihood ratio.

- That's fine, but what do you do with it?
- If you have a fixed cut for all events, this is what you are doing:





$$L_{tot} = L_1 \cdot L_2$$

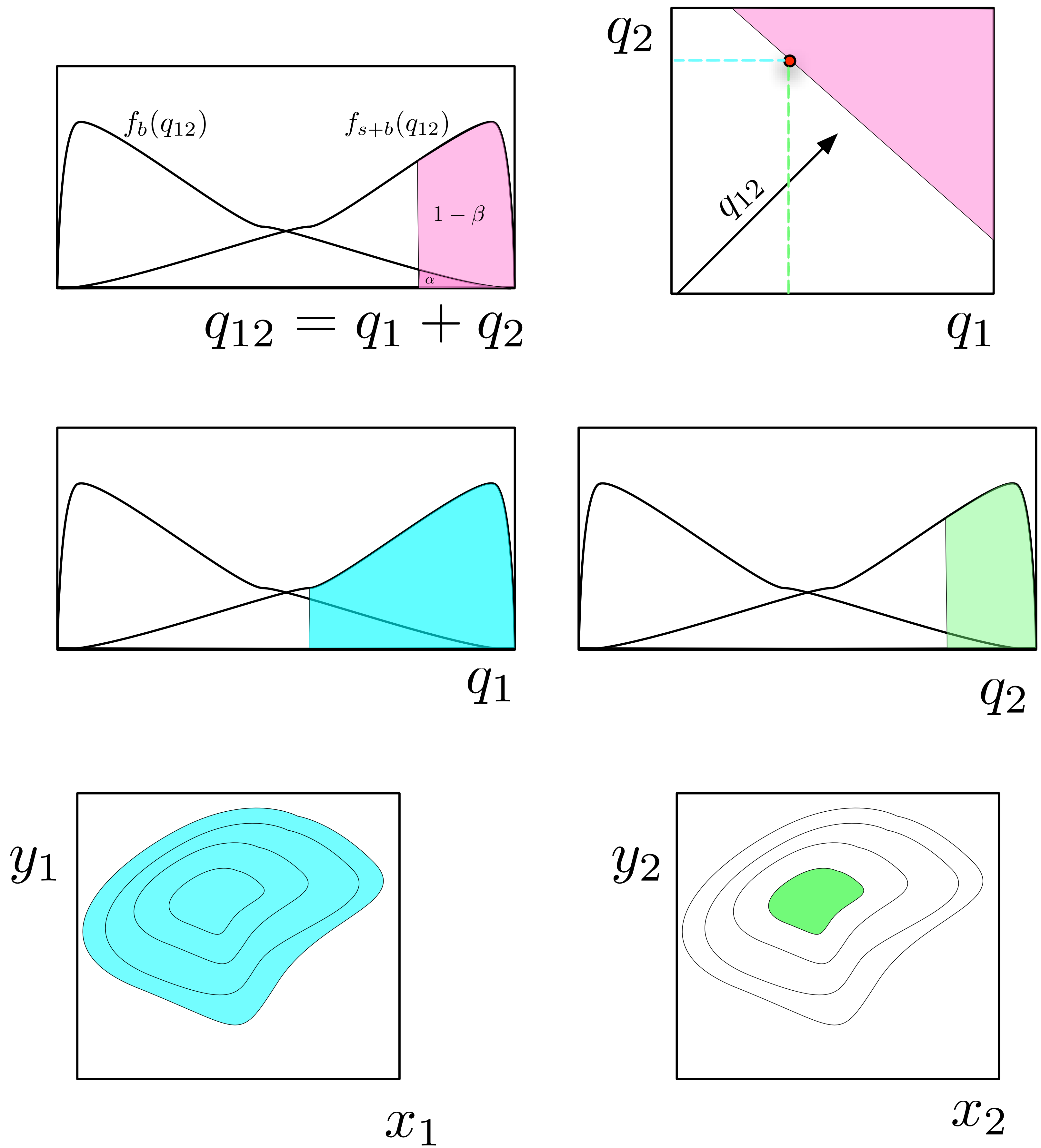
$$q_{12} = \ln L_{12} = \ln L_1 + \ln L_2 = q_1 + q_2$$

EXPERIMENTS VS. EVENTS

Ideally, you want to cut on the likelihood ratio for your **experiment**

- equivalent to a sum of log likelihood ratios

Easy to see that includes experiments where one event had a high likelihood and the other one was relatively small



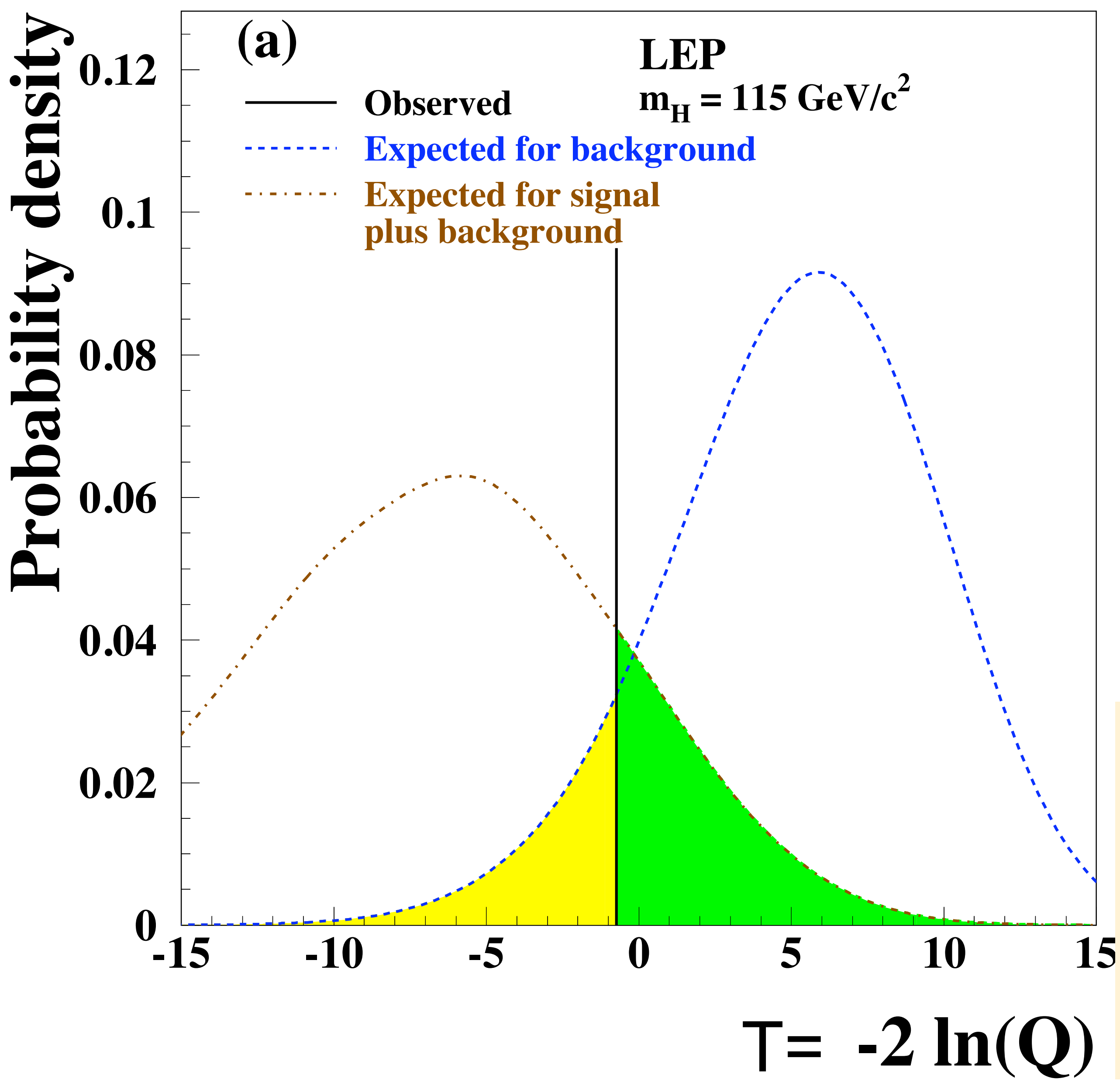
AN OPTIMAL WAY TO COMBINE

Special case of our general probability model

(no nuisance parameters)

$$Q = \frac{L(x|H_1)}{L(x|H_0)} = \frac{\prod_i^{N_{chan}} Pois(n_i|s_i+b_i) \prod_j^{n_i} \frac{s_i f_s(x_{ij}) + b_i f_b(x_{ij})}{s_i+b_i}}{\prod_i^{N_{chan}} Pois(n_i|b_i) \prod_j^{n_i} f_b(x_{ij})}$$

$$\ln Q = -s_{tot} + \sum_i^{N_{chan}} \sum_j^{n_i} \ln\left(1 + \frac{s_i f_s(x_{ij})}{b_i f_b(x_{ij})}\right)$$



Instead of simply counting events, the optimal test statistic is equivalent to adding events **weighted** by

ln(1+signal/background ratio)

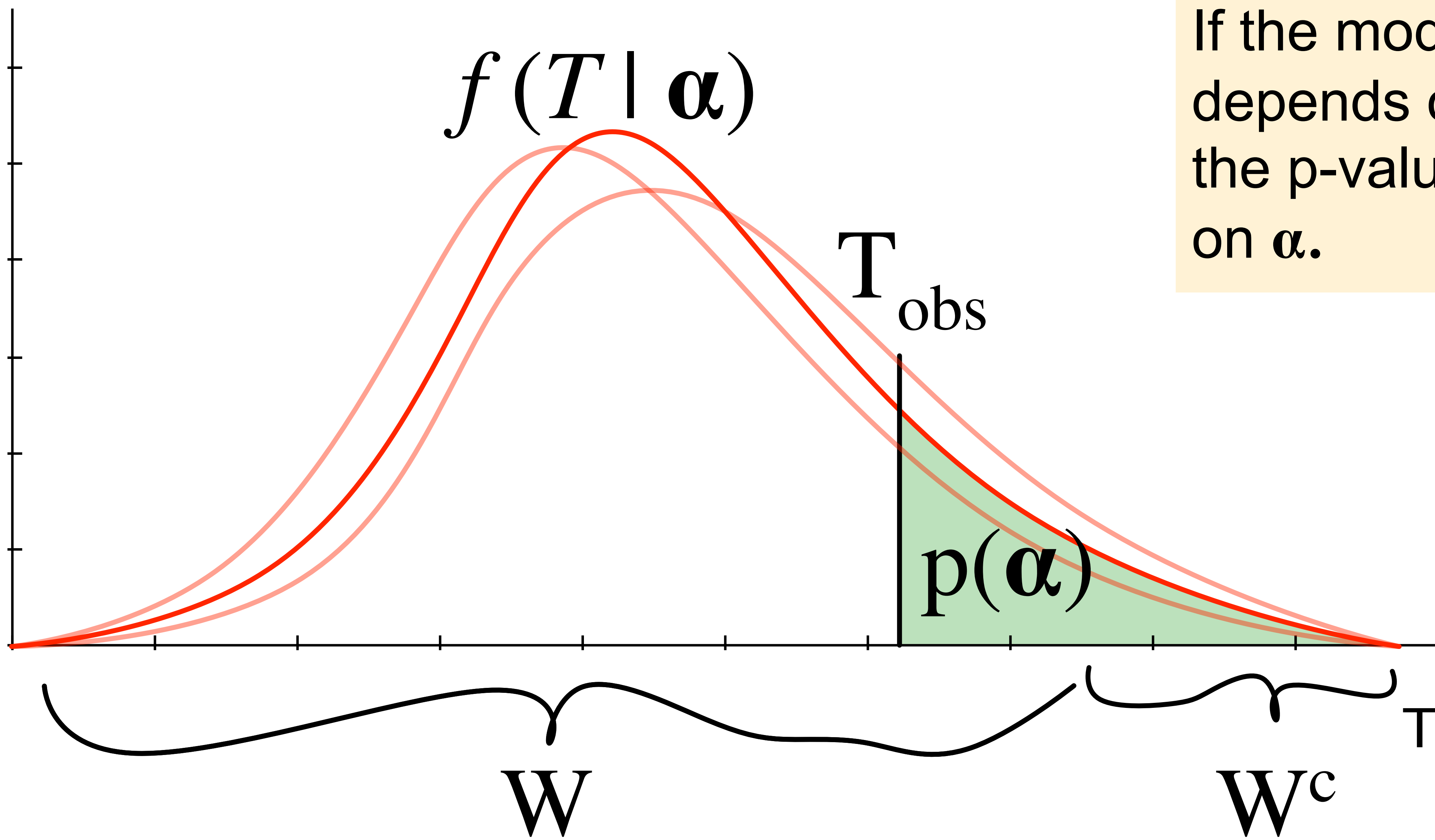
The test statistic is a map T:data $\rightarrow \mathbb{R}$

By repeating the experiment many times, you obtain a distribution for T

Instead of choosing to accept/reject H_0
one can compute the p-value

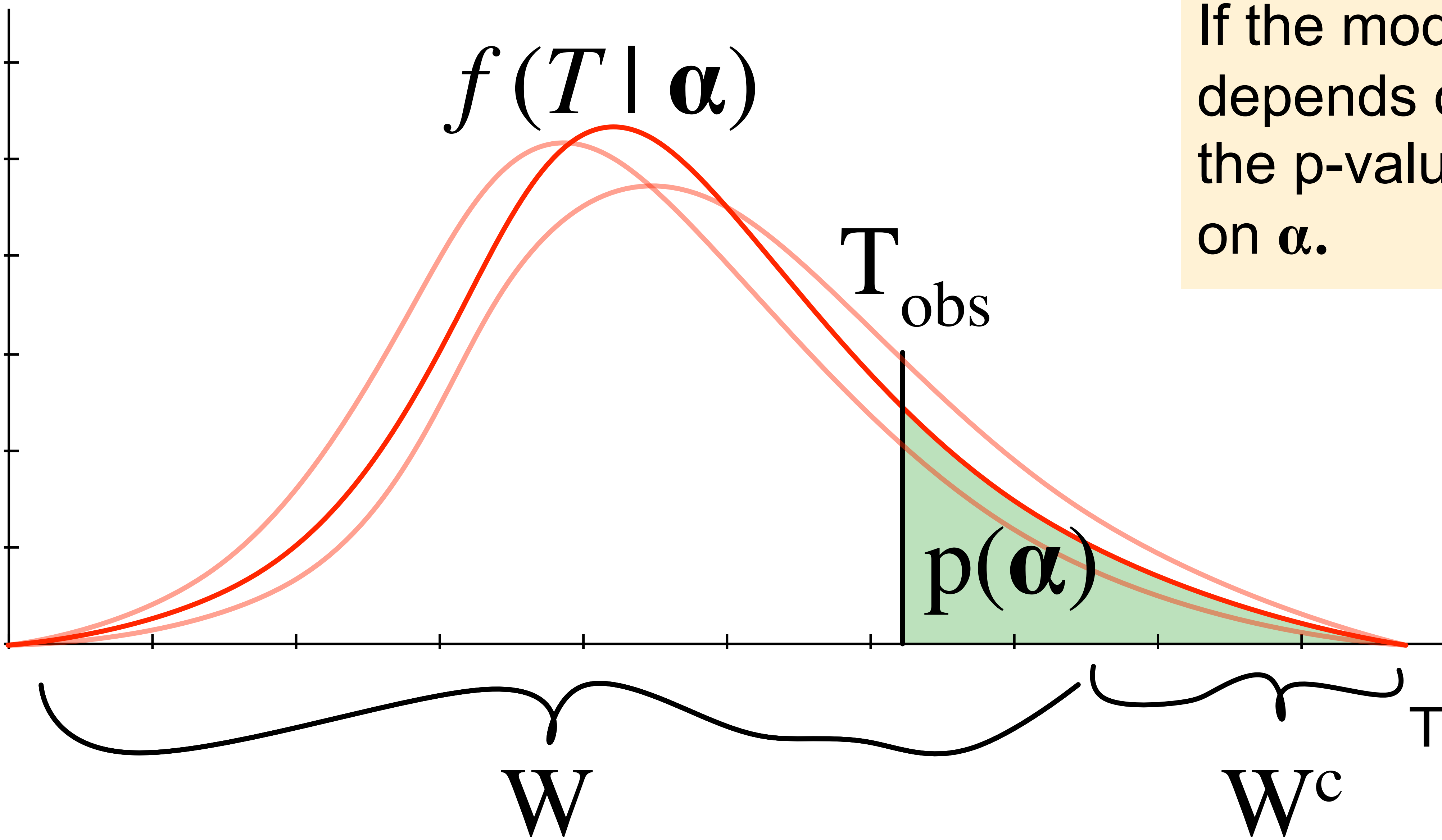
$$p = \int_{T_o}^{\infty} f(T|H_0)$$

If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.



$$p(\boldsymbol{\alpha}) = \int_{T_0}^{\infty} f(T|\boldsymbol{\alpha})dT = \int \mathbf{f}(\mathcal{D}|\boldsymbol{\alpha})\,\theta(T(\mathcal{D}) - T_0)\,d\mathcal{D} = P(T \geq T_0|\boldsymbol{\alpha})$$

When the model has nuisance parameters, only reject the null if $p(\boldsymbol{\alpha})$ sufficiently small **for all values** of the nuisance parameters.



If the model for the data depends on parameters $\boldsymbol{\alpha}$ the p-value also depends on $\boldsymbol{\alpha}$.

$$p(\boldsymbol{\alpha}) = \int_{T_0}^{\infty} f(T|\boldsymbol{\alpha}) dT = \int \mathbf{f}(\mathcal{D}|\boldsymbol{\alpha})\, \theta(T(\mathcal{D}) - T_0)\, d\mathcal{D} = P(T \geq T_0|\boldsymbol{\alpha})$$

The Profile Likelihood Ratio

Consider our general model with a single parameter of interest μ

- let μ=0 be no signal, μ=1 nominal signal

In the LEP approach the likelihood ratio is equivalent to:

$$Q_{\rm LEP} = \frac{L(\mu=1,\theta)}{L(\mu=0,\theta)} = \frac{f(\mathcal{D}|\mu=1,\theta)}{f(\mathcal{D}|\mu=0,\theta)}$$

- but this variable is sensitive to uncertainty on θ and makes no use of auxiliary measurements **a**

Alternatively, one can define **profile likelihood ratio**

$$\lambda(\mu) = \frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})} = \frac{f(\mathcal{D},\mathcal{G}|\mu,\hat{\hat{\theta}}(\mu;\mathcal{D},\mathcal{G})\,)}{f(\mathcal{D},\mathcal{G}|\hat{\mu},\hat{\theta})}$$

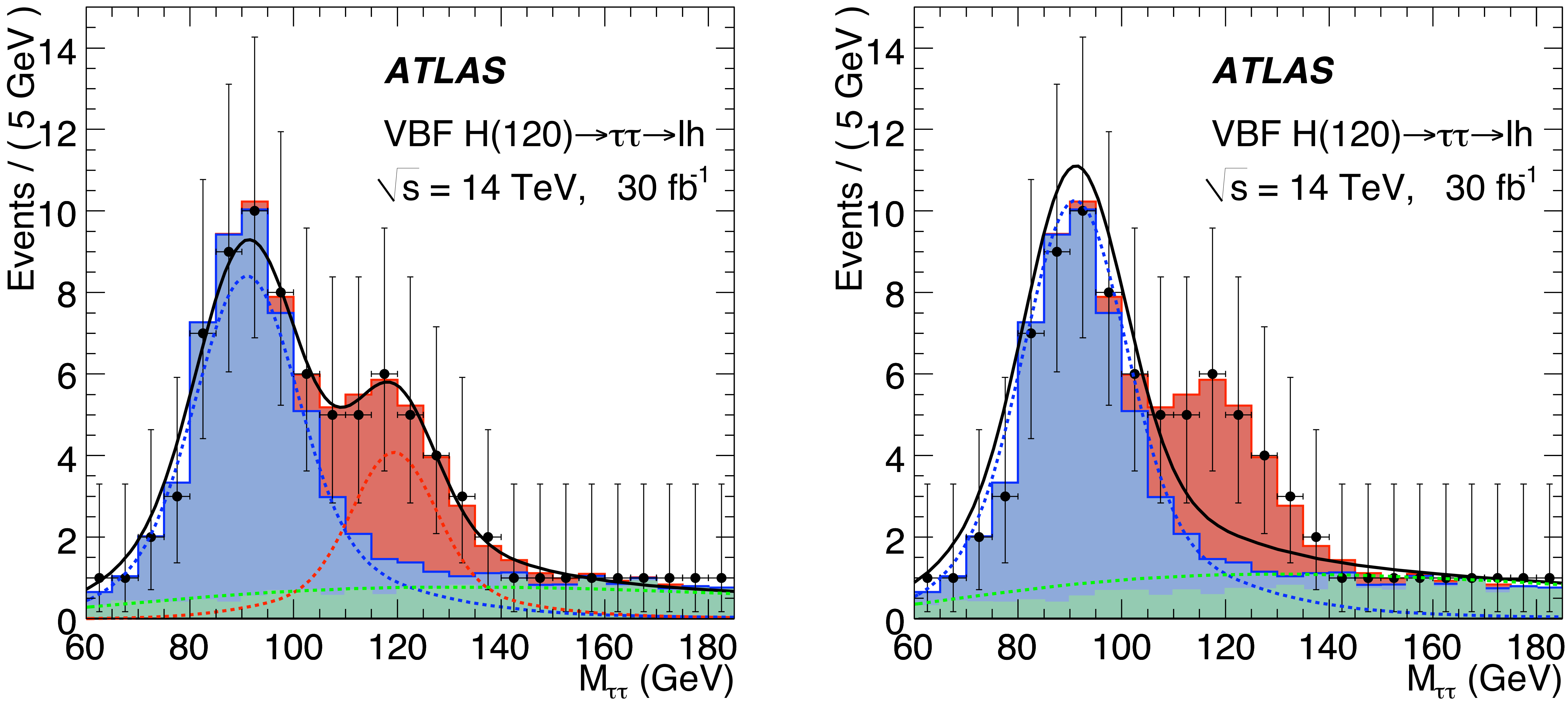
- where $\hat{\hat{\theta}}(\mu;\mathcal{D},\mathcal{G})$ is best fit with μ fixed (the constrained maximum likelihood estimator, depends on data)
- and $\hat{\theta}$ and $\hat{\mu}$ are best fit with both left floating (unconstrained)
- Tevatron used $Q_{\rm Tev} = \lambda(\mu=1)/\lambda(\mu=0)$ as generalization of $Q_{\rm LEP}$

An example

Essentially, you need to fit your model to the data twice:
once with everything floating, and once with signal fixed to 0

$$\lambda(\mu=0) = \frac{L(\mu=0, \hat{\hat{\theta}}(\mu=0))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D},\mathcal{G}|\mu=0, \hat{\hat{\theta}}(\mu=0;\mathcal{D},\mathcal{G})\,)}{f(\mathcal{D},\mathcal{G}|\hat{\mu},\hat{\theta})}$$

$f(\mathcal{D},\mathcal{G}|\hat{\mu},\hat{\theta})$ $\qquad$ $f(\mathcal{D},\mathcal{G}|\mu=0, \hat{\hat{\theta}}(\mu=0;\mathcal{D},\mathcal{G})\,)$



Properties of the Profile Likelihood Ratio

After a close look at the profile likelihood ratio

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})} = \frac{f(\mathcal{D}, \mathcal{G} | \mu, \hat{\hat{\theta}}(\mu; \mathcal{D}, \mathcal{G})\,)}{f(\mathcal{D}, \mathcal{G} | \hat{\mu}, \hat{\theta})}$$

one can see the function is independent of true values of θ

- though its distribution might depend indirectly

Wilks's theorem states that under certain conditions the distribution of $-2 \ln \lambda\ (\mu=\mu_0)$ given that the true value of μ is μ_0 converges to a chi-square distribution

- more on this later, but the important points are:
- "asymptotic distribution" is known and it is independent of θ !
  - more complicated if parameters have boundaries (eg. $\mu \geq 0$)

Thus, we can calculate the p-value for the background-only hypothesis without having to generate Toy Monte Carlo!

Toy Monte Carlo

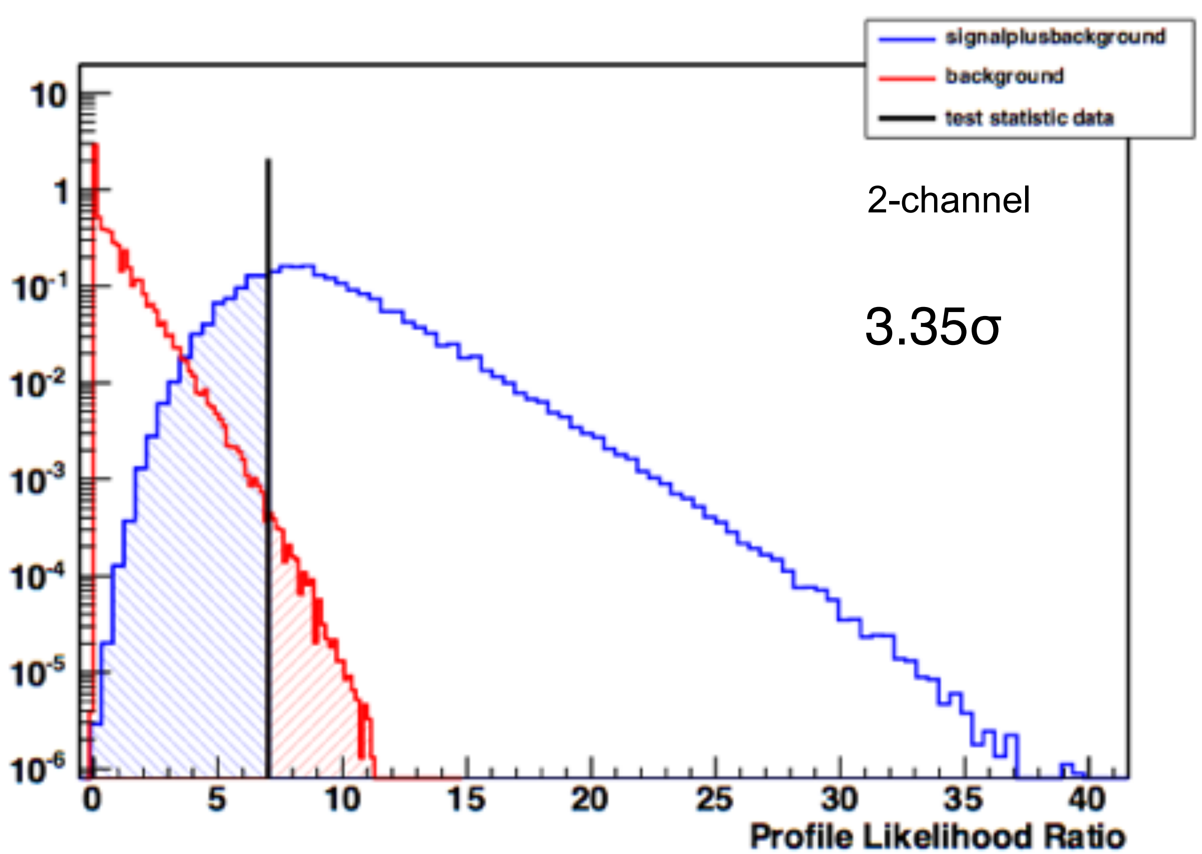
Explicitly build distribution by generating “toys” / pseudo experiments assuming a specific value of μ and ν.

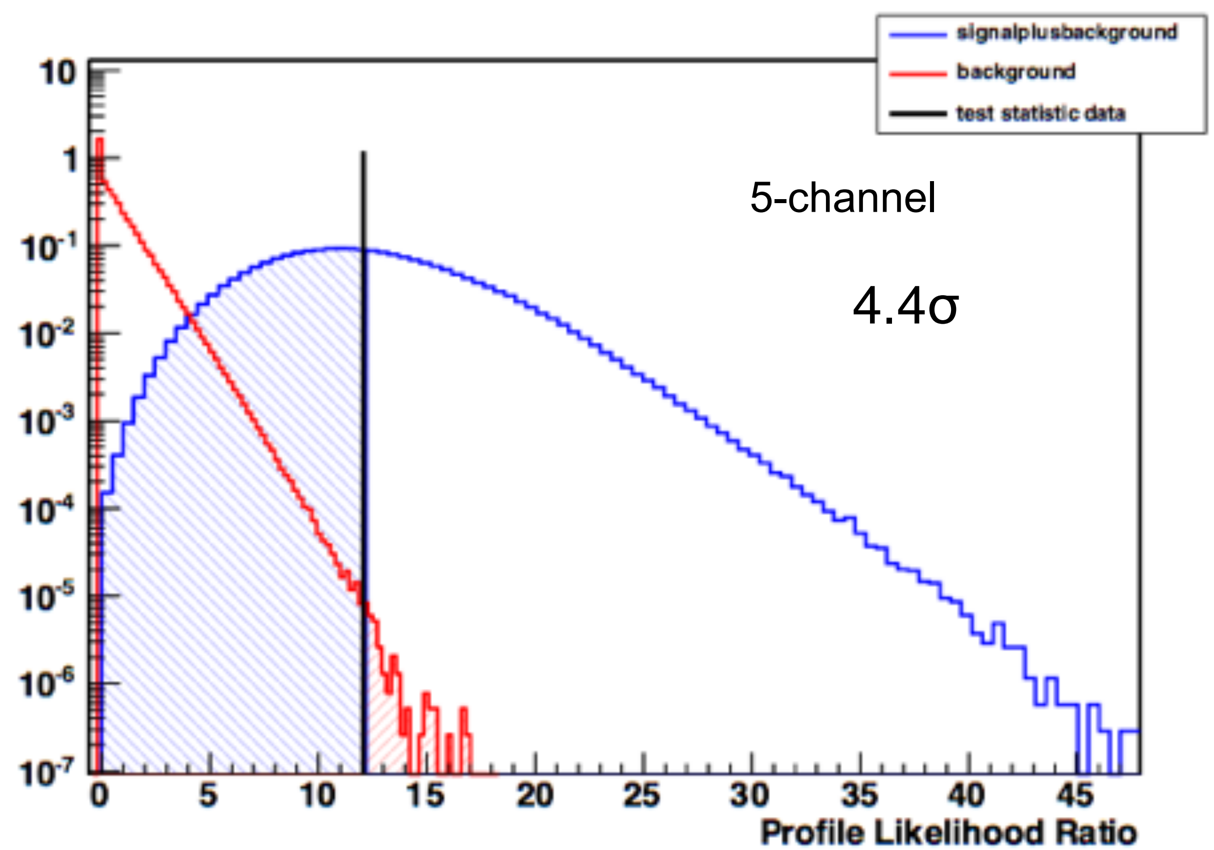
- randomize both main measurements $\mathcal{D}=\{x\}$ and auxiliary measurements $\mathcal{G}=\{\mathbf{a}\}$
- fit the model twice for the numerator and denominator of profile likelihood ratio
- evaluate $-2\ln\lambda(\mu)$ and add to histogram

Choice of μ is straight forward: typically $\mu=0$ and $\mu=1$, but choice of θ is less clear

- more on this later

This can be very time consuming. Plots below use millions of “toy” pseudo-experiments





"THE ASIMOV PAPER"

Recently we showed how to generalize this asymptotic approach

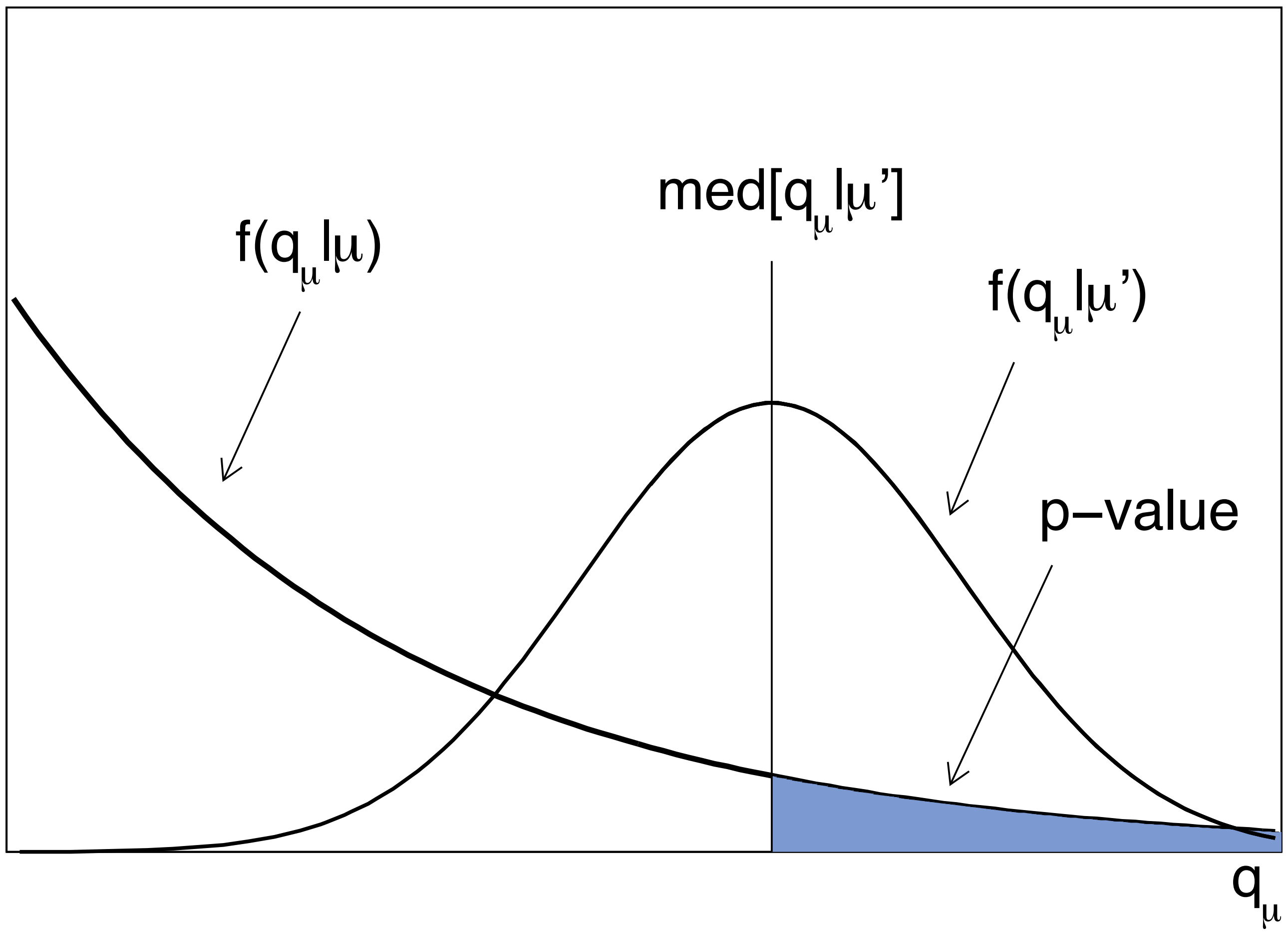
- generalize Wilks's theorem when boundaries are present
- use Wald's result for distribution for alternate hypothesis $f(-2\log\lambda(\mu) \mid \mu')$

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan, Kyle Cranmer, Eilam Gross, Ofer Vitells

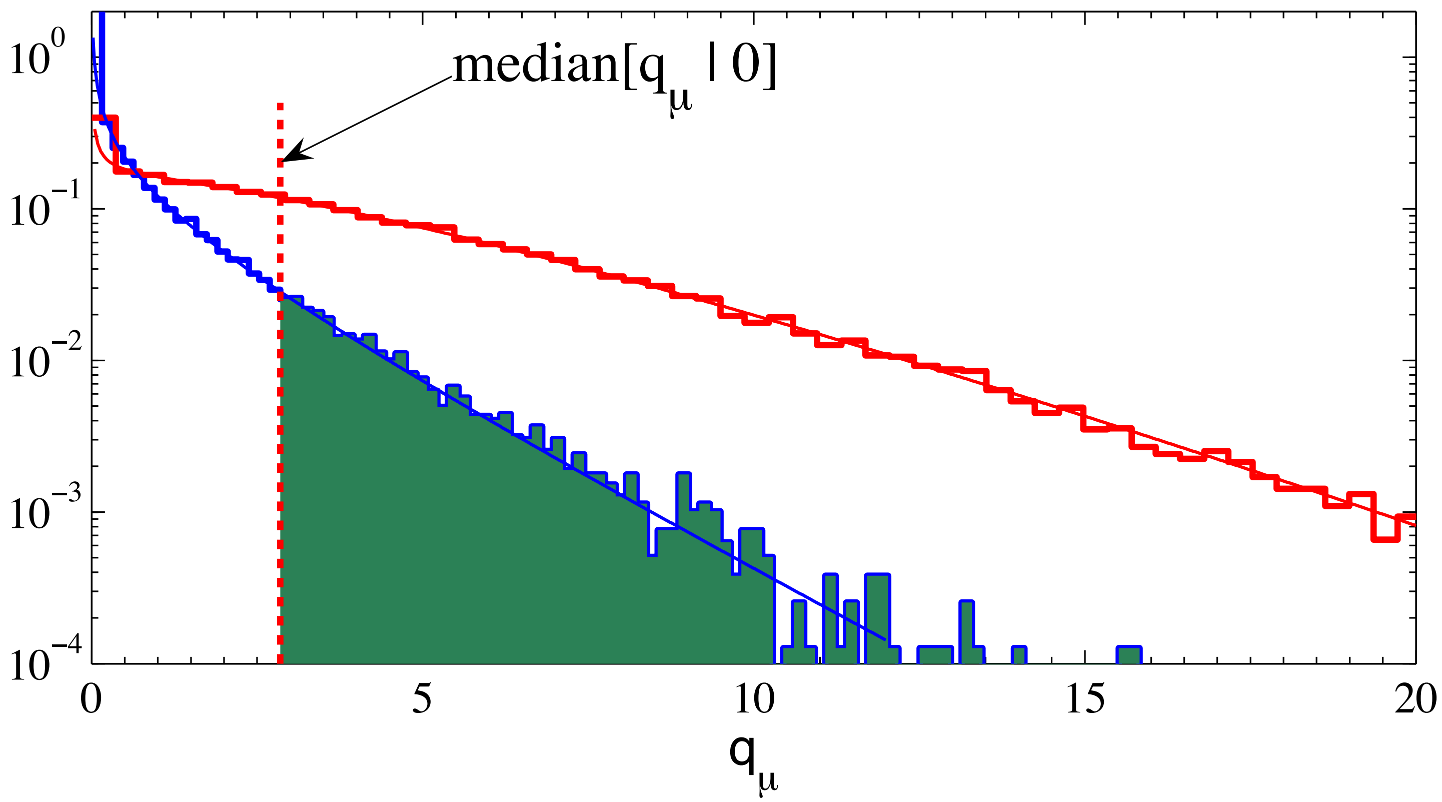
Eur.Phys.J.C71:1554,2011

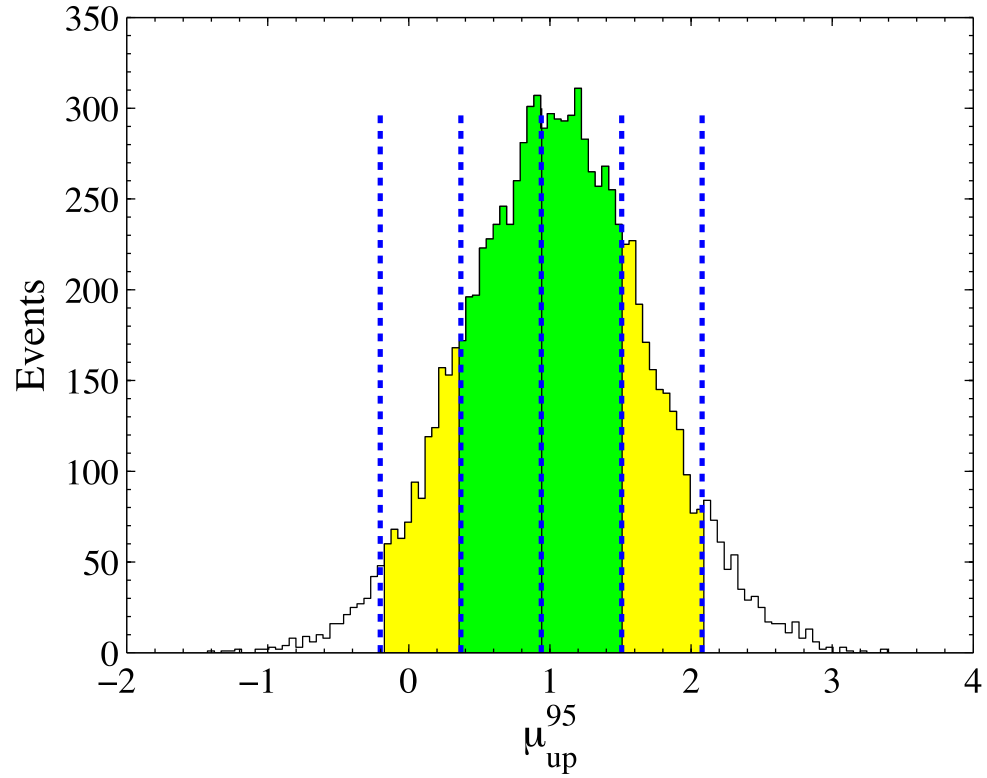
http://arxiv.org/abs/1007.1727v2



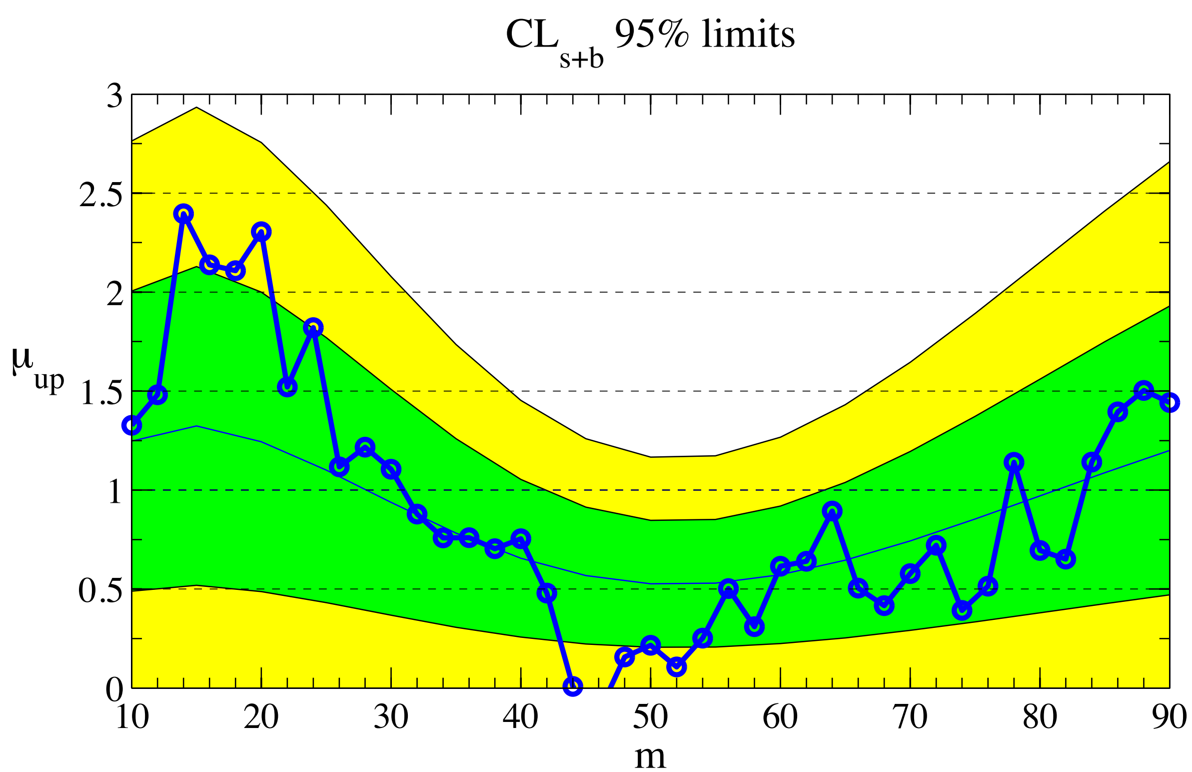
Comparison of asymptotic and ensembles

Compare asymptotic distributions to distributions obtained with large ensembles of pseudo-experiments generated with Monte Carlo techniques





This is a significant development as building this distribution from Monte Carlo approaches can take 100,000 CPU hours for Higgs search!



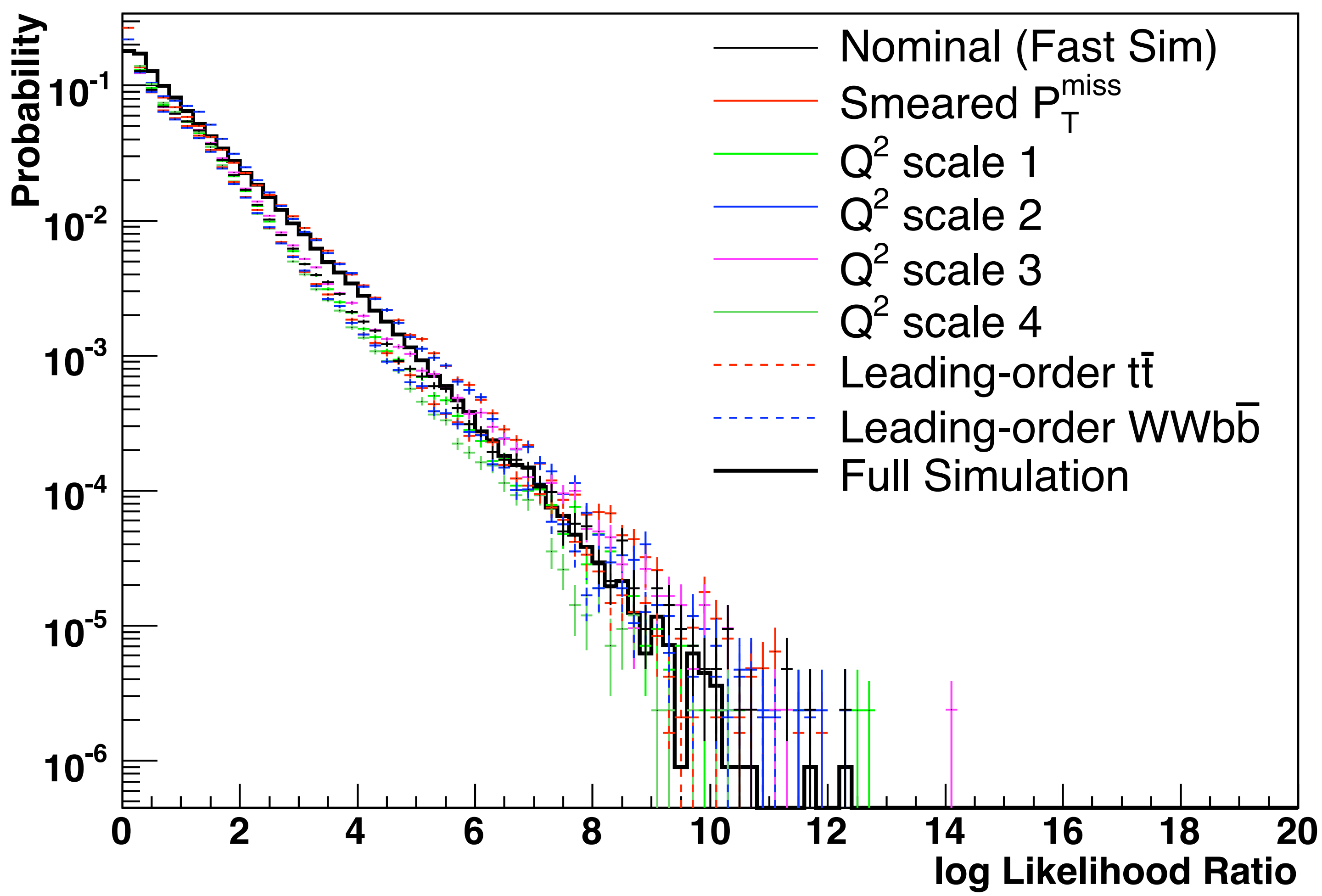
G. Cowan, KC, E. Gross, O. Vitells
Eur.Phys.J. C71 (2011) 1554
[arXiv:1007.1727]

Experimentalist Justification

So far this looks a bit like magic. How can you claim that you incorporated your systematic just by fitting the best value of your uncertain parameters and making a ratio?

It won't unless the the parametrization is sufficiently flexible.

So check by varying the settings of your simulation, and see if the profile likelihood ratio is still distributed as a chi-square



Here it is pretty stable, but it's not perfect (and this is a log plot, so it hides some pretty big discrepancies)

For the distribution to be independent of the nuisance parameters your parametrization must be sufficiently flexible.

A VERY IMPORTANT POINT

If we keep pushing this point to the extreme, the physics problem goes beyond what we can handle practically

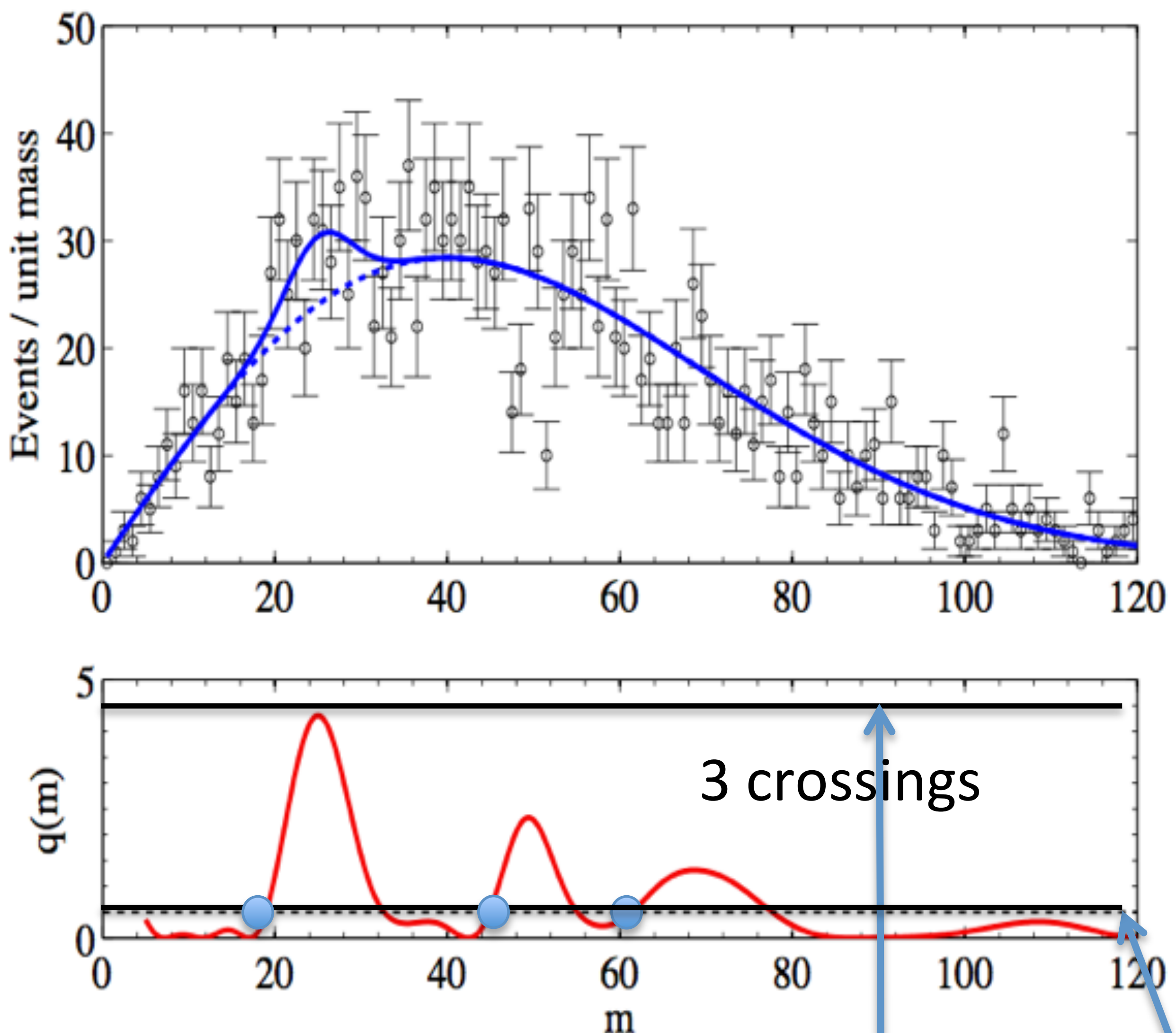
The p-values are usually predicated on the assumption that the **true distribution** is in the family of distributions being considered

- eg. we have sufficiently flexible models of signal & background to incorporate all systematic effects
- but we don't believe we simulate everything perfectly
- ..and when we parametrize our models usually we have further approximated our simulation.
  - nature -> simulation -> parametrization

At some point these approaches are limited by honest systematics uncertainties (not statistical ones). Statistics can only help us so much after this point. Now we must be physicists!

LOOK-ELSEWHERE EFFECT

Approximation best above 3σ



Typically our signal model has some parameter (eg. m_H), which does not affect the null (background only).

This modifies the distribution of the likelihood ratio test statistic

we call this the “look-elsewhere effect”

Recently Gross & Vitells found the results of Rice, Davies, and Leadbetter for a fast asymptotic approximation for the global p-value

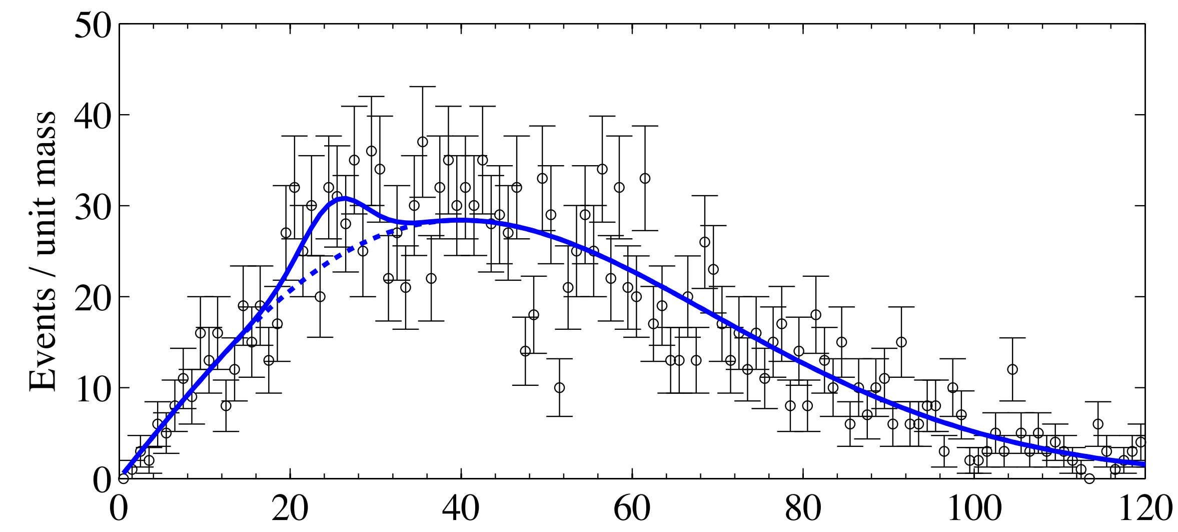
E. Gross & O. Vitells, **Eur.Phys.J. C70 (2010); Astropart.Phys. 35 (2011)**

$$p_0^{global} \cong p_0^{local} + \left\langle N(q_{ref}) \right\rangle e^{-(q_{test}-q_{ref})/2}$$

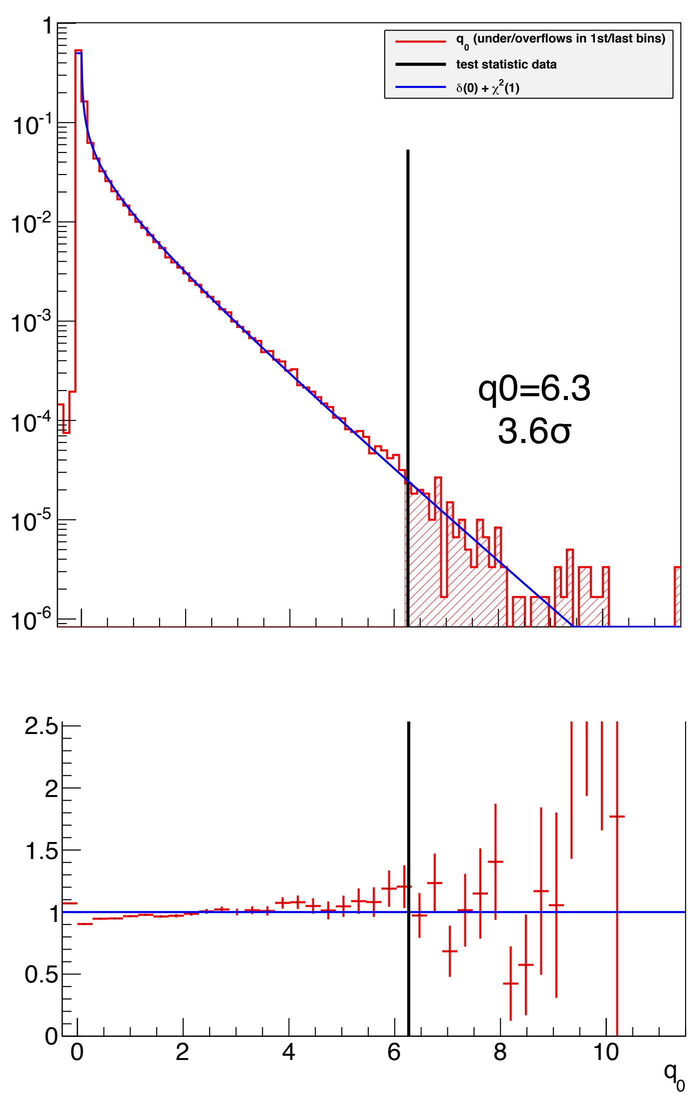
R. B. Davies,Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 64 (1977); Biometrika 74 (1987).

Deviations from the Asymptotic Distributions

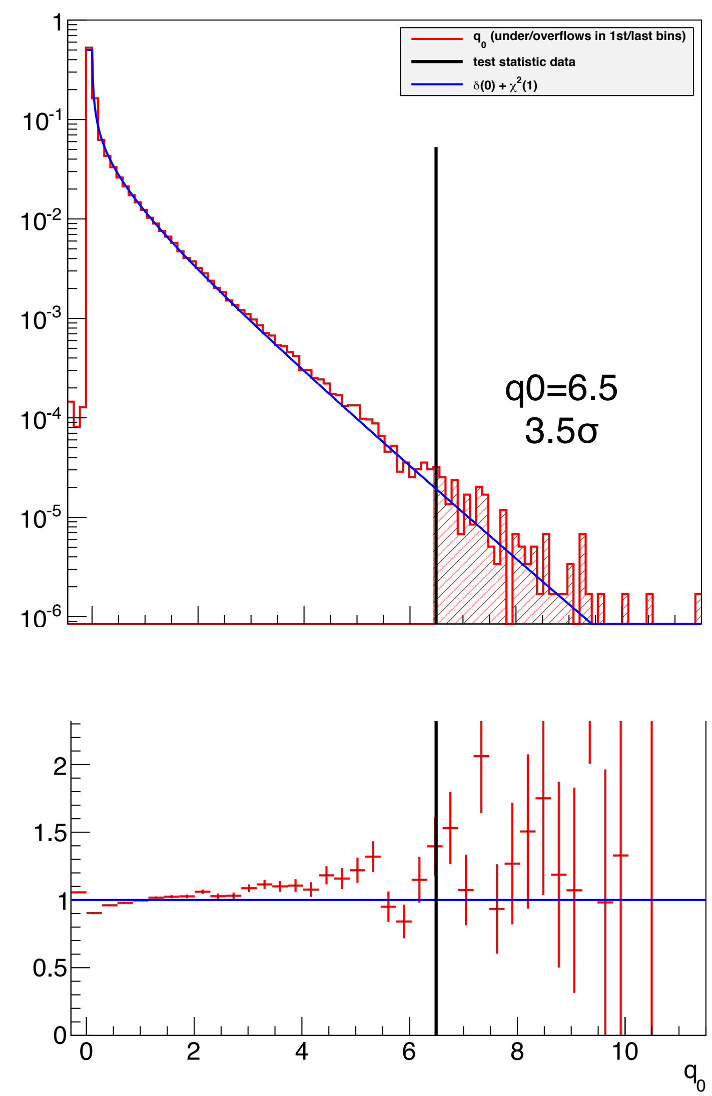
Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).



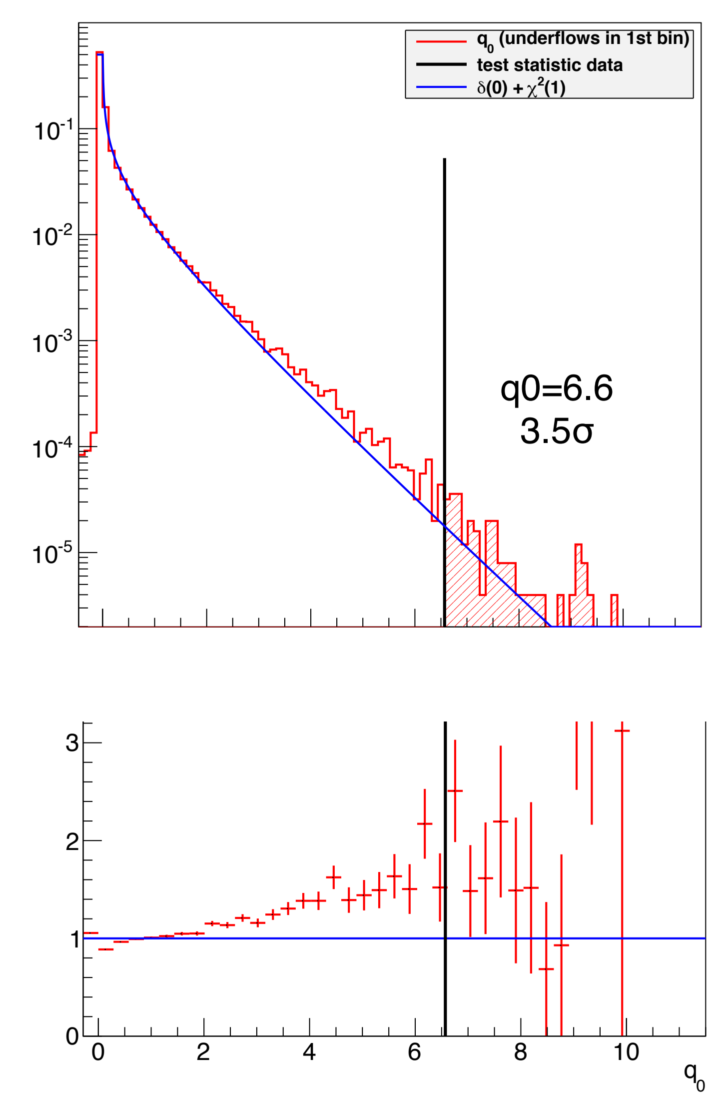
Without energy scale uncertainty
Without mass resolution uncertainty



Without energy scale uncertainty
With mass resolution uncertainty



With energy scale uncertainty
With mass resolution uncertainty

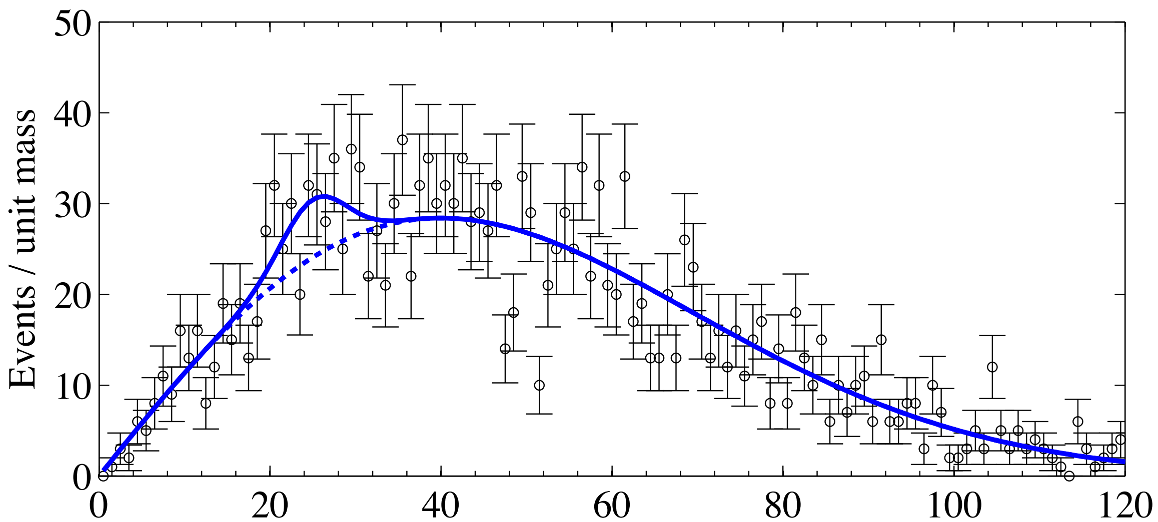


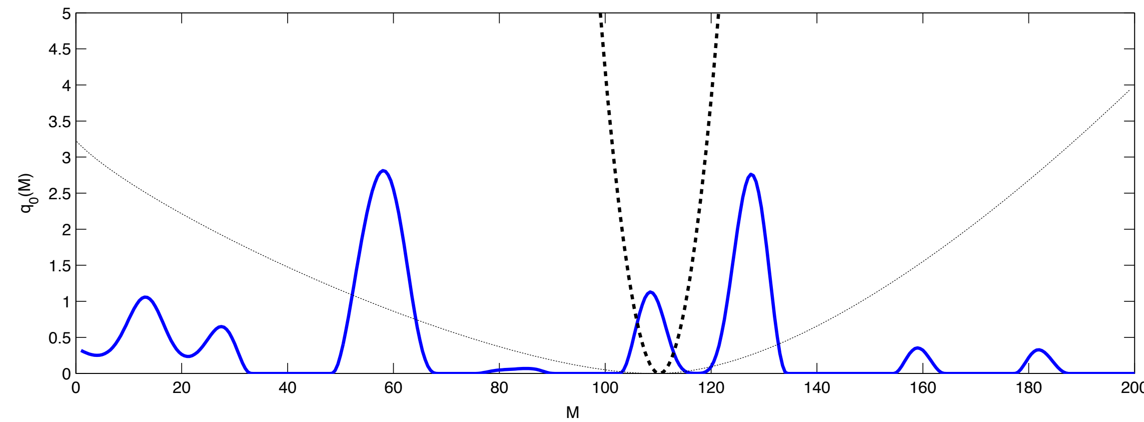
A MORE SUBTLE EFFECT

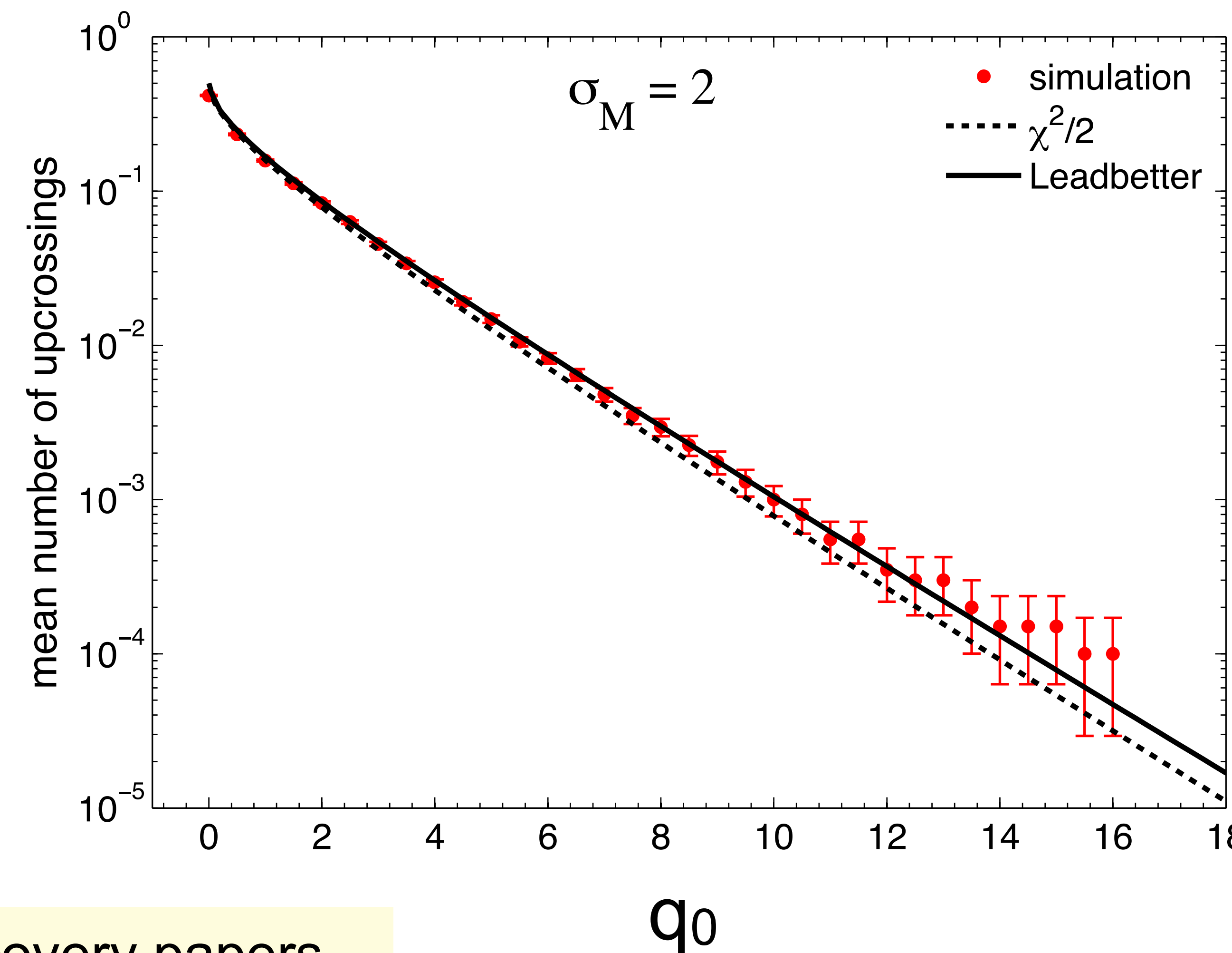
Even if we fix the location of the signal some systematic effects are equivalent to small uncertainty in the location (e.g. energy calibration).

(H.R. Leadbetter, 1965

$$\mathbb{E}[N_u] = \sigma_2 \int \phi(u(M)) \left[\phi\left(\frac{u'(M)}{\sigma_2}\right) + \frac{u'(M)}{\sigma_2} \left\{ \Phi\left(\frac{u'(M)}{\sigma_2}\right) - \frac{1}{2} \right\} \right] dM$$







These parameters are slowing convergence to the asymptotic distribution and variance may not reduce with more data.

O. Vitells found exact solution by Leadbetter for the case of only one such nuisance parameter

note: used in Higgs discovery papers