



**University of
Zurich^{UZH}**

Data (mis) modelling

PHYSTAT-Systematics, 2nd November 2021

A. de Wit

Introduction: an ideal world

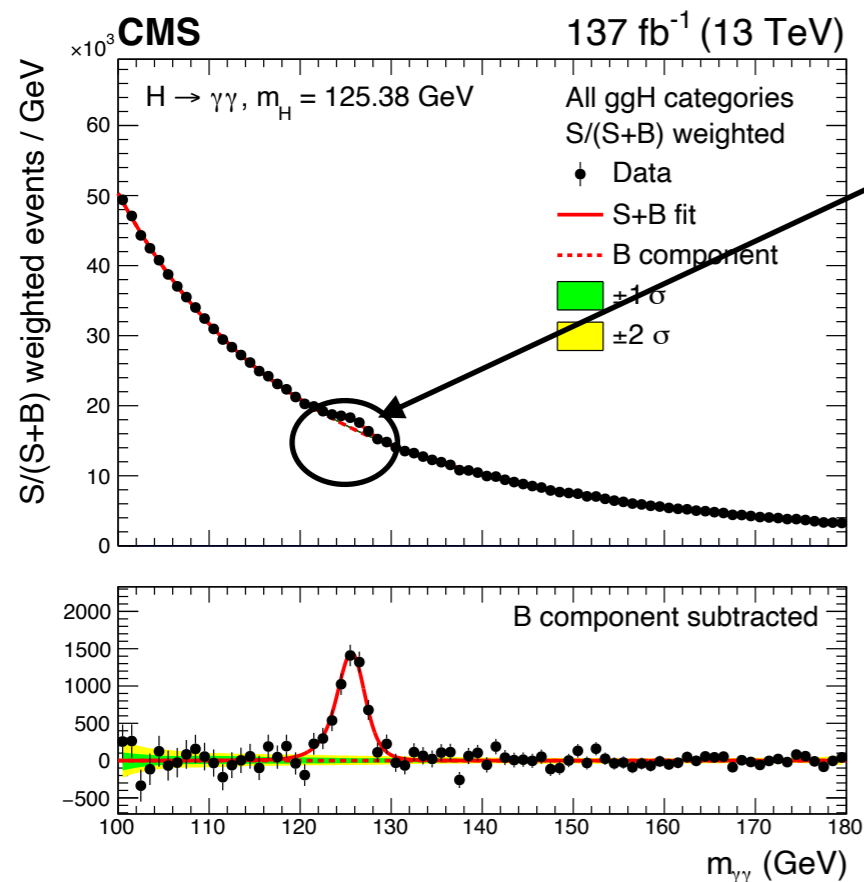
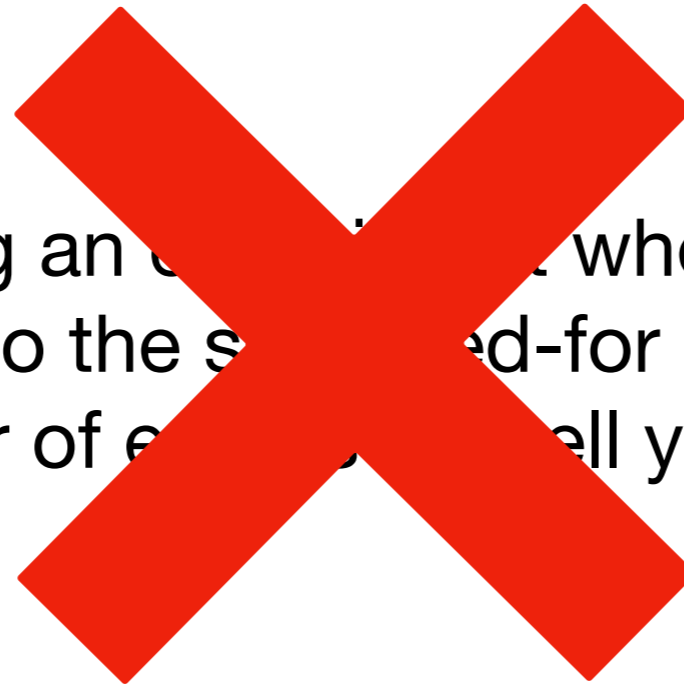
- Imagine you're doing an experiment where any data you collect is certain to be due to the searched-for signal and simply counting the number of events will tell you about the rate of this process

Introduction: an ideal world

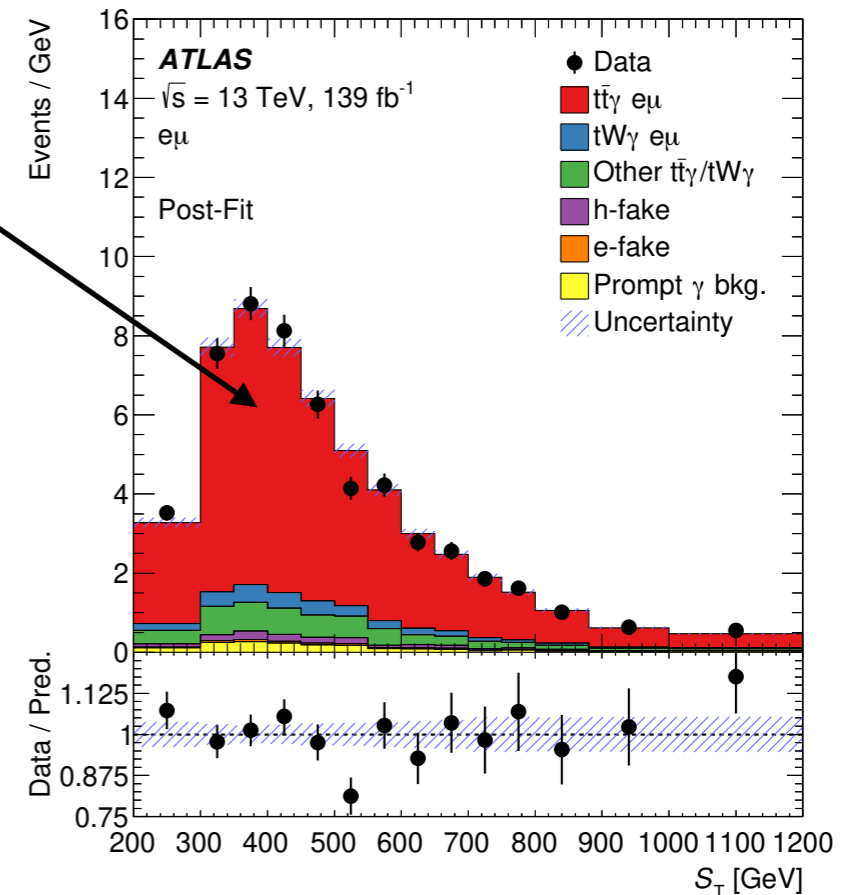
- Imagine you're doing an experiment where any data you collect is certain to be due to the signal and simply counting the number of events will tell you about the rate of this process

Introduction: an ideal world

- Imagine you're doing an experiment where any data you collect is certain to be due to the signal and simply counting the number of events will tell you about the rate of this process

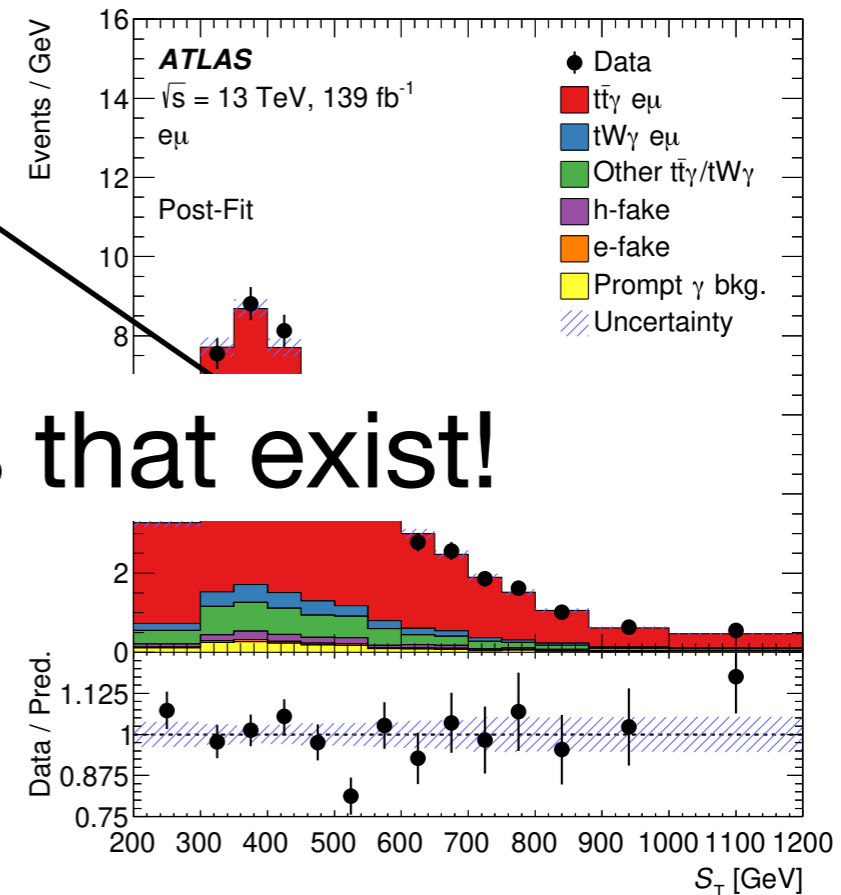
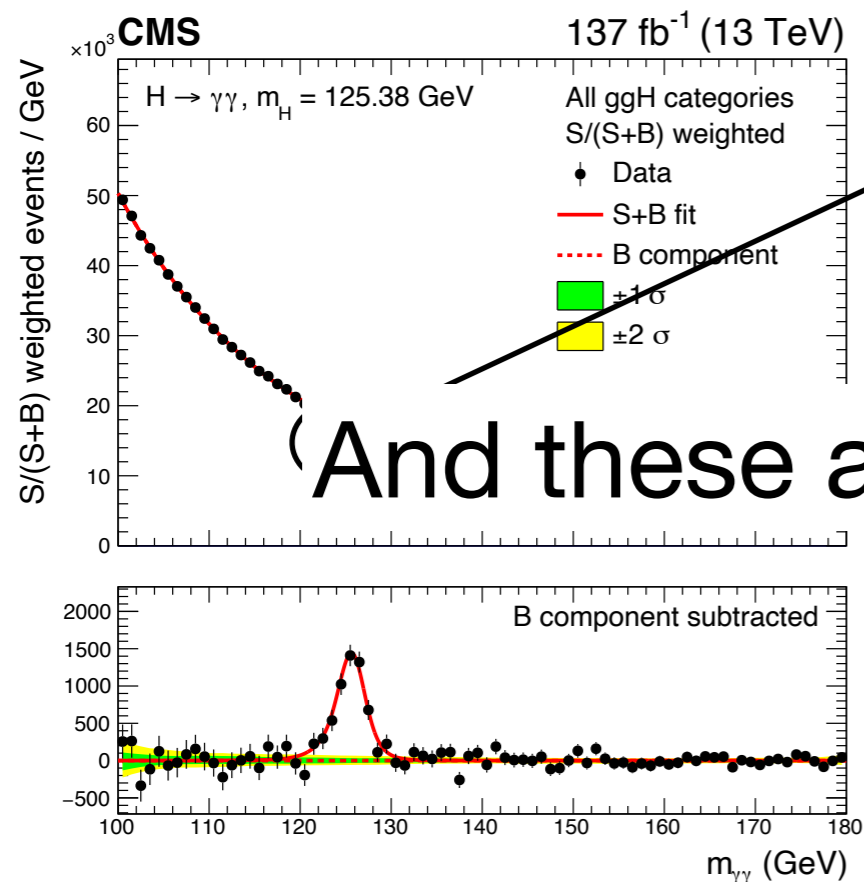
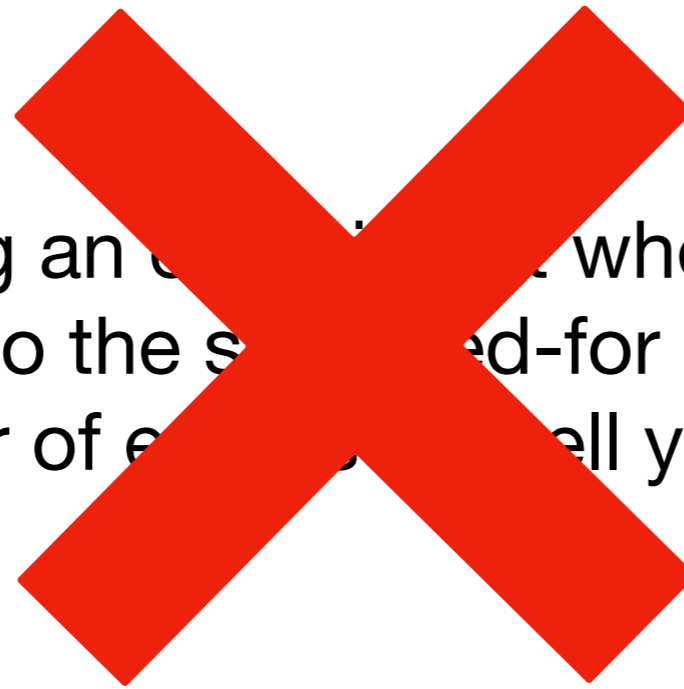


Signal



Introduction: an ideal world

- Imagine you're doing an experiment where any data you collect is certain to be due to the signal and simply counting the number of events will tell you about the rate of this process

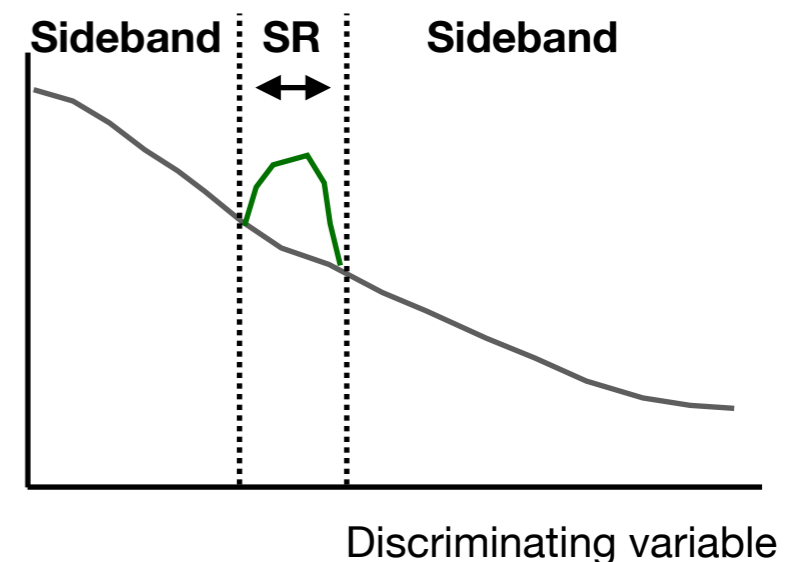


Signal

(And these are processes that exist!

Modelling: essential

- LHC analyses always rely on good modelling
 - Backgrounds
 - Detector effects
- **How** to model the data?
 - Simulation → does not always represent the data well
 - From sidebands → extrapolation
 - Using histograms
 - With analytic functions



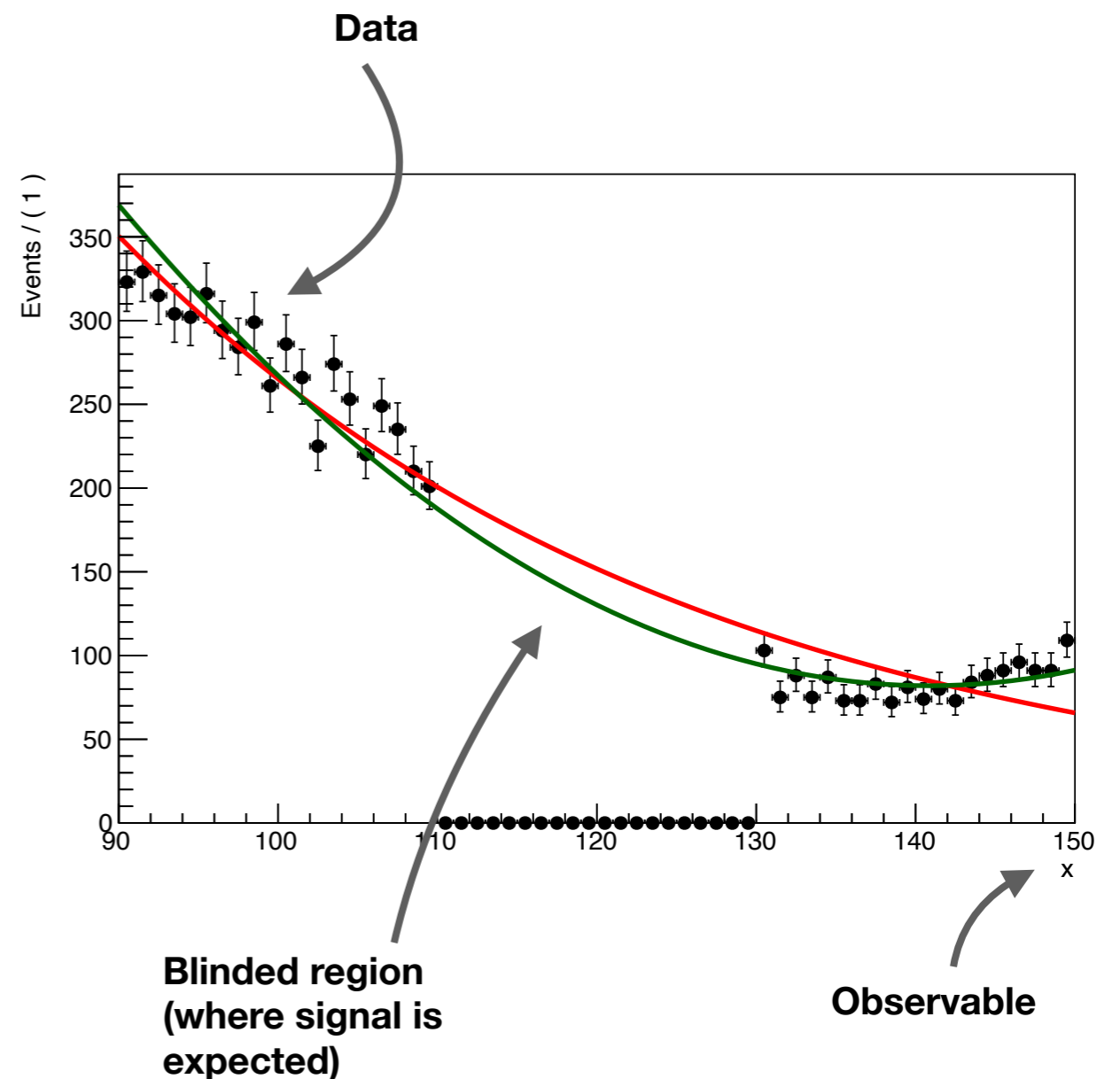
Disclaimer

- I'm an LHC physicist and spend a lot of my time doing likelihood fits
 - Apologies if your favourite topic is not included!
- This talk is split into two parts
 - Issues that arise when using **analytic functions** to model the data
 - Issues that arise when using **histogram templates** to model the data

Analytic functions

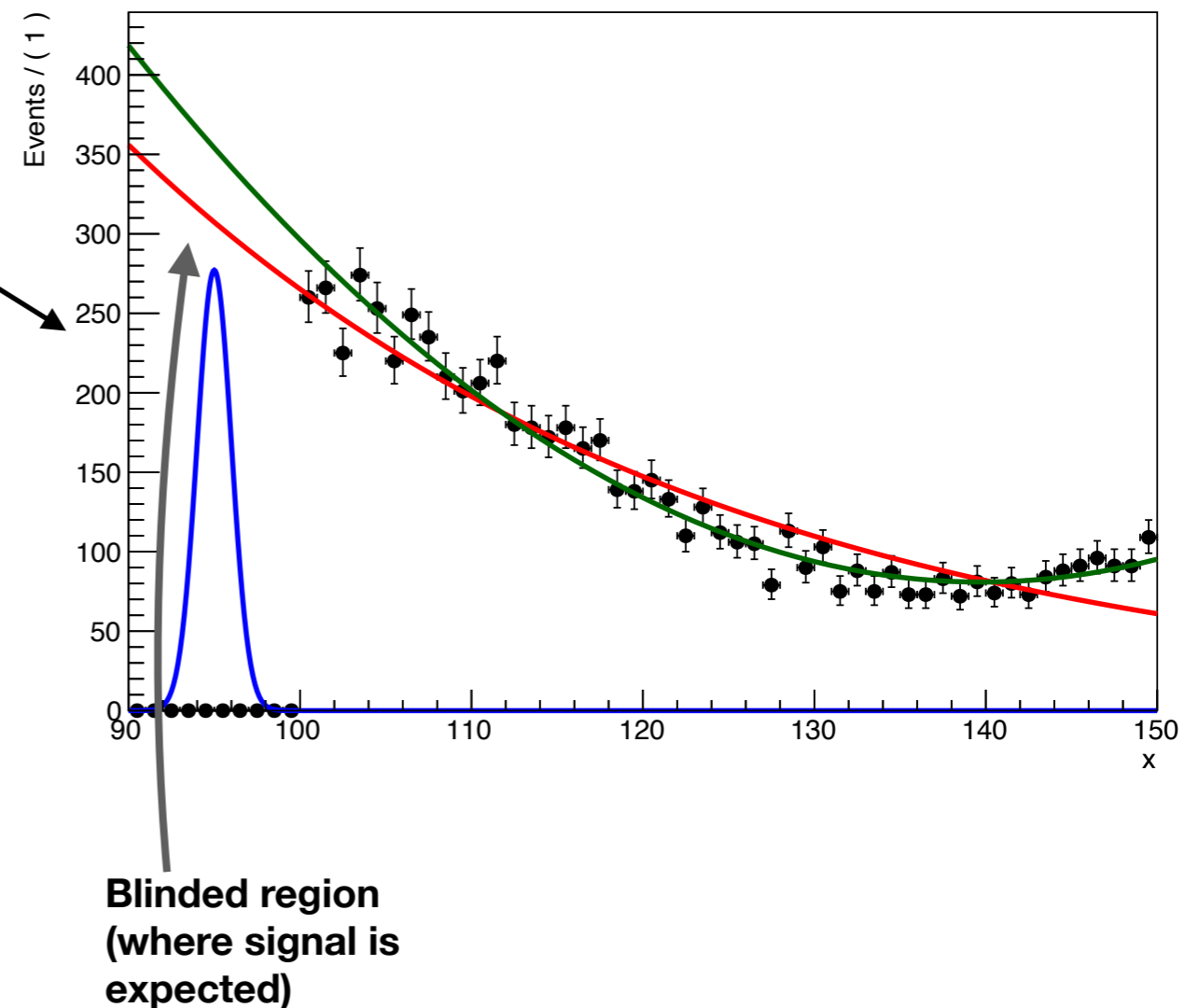
Determination of function

- Functional form that describes the data not known a-priori
- Usually use the **sidebands** to determine appropriate function(s) to fit the background
- Excluding the signal region - analyses tend to be **blind**



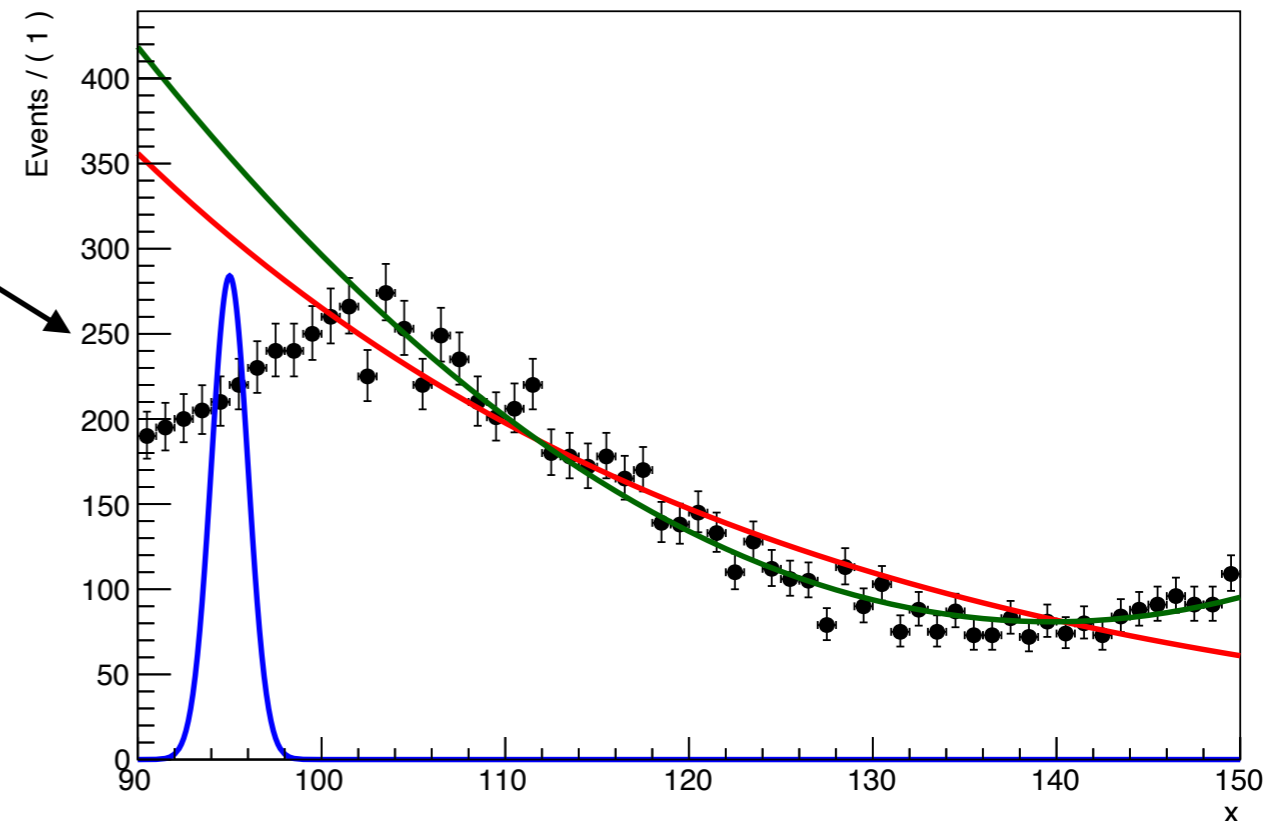
Interlude - sideband choice

- Sideband should be chosen carefully
- "turn-on" effects (from analysis selection ...)



Interlude - sideband choice

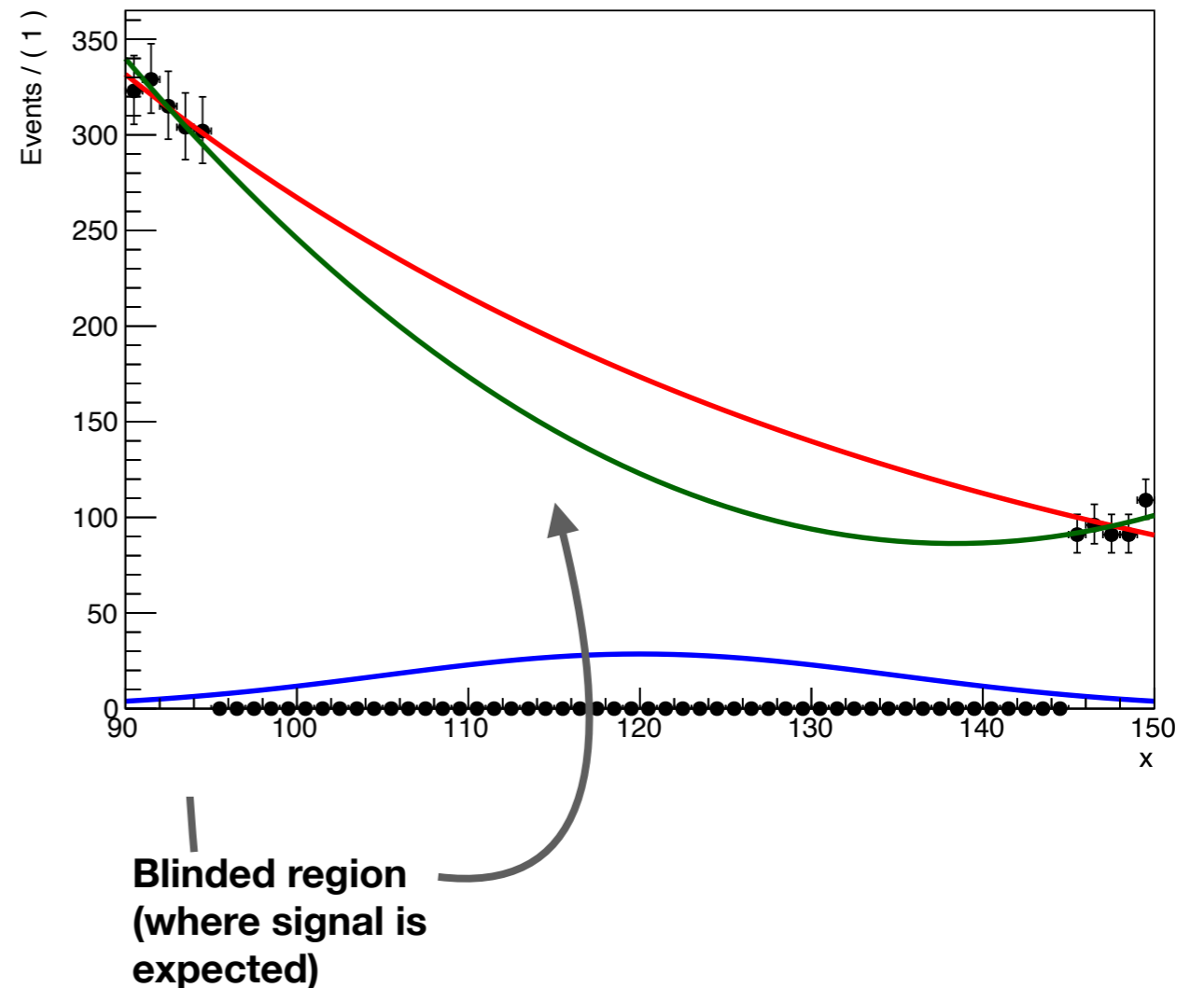
- Sideband should be chosen carefully
 - "turn-on" effects (from analysis selection ...)



Blinded region
(where signal is
expected)

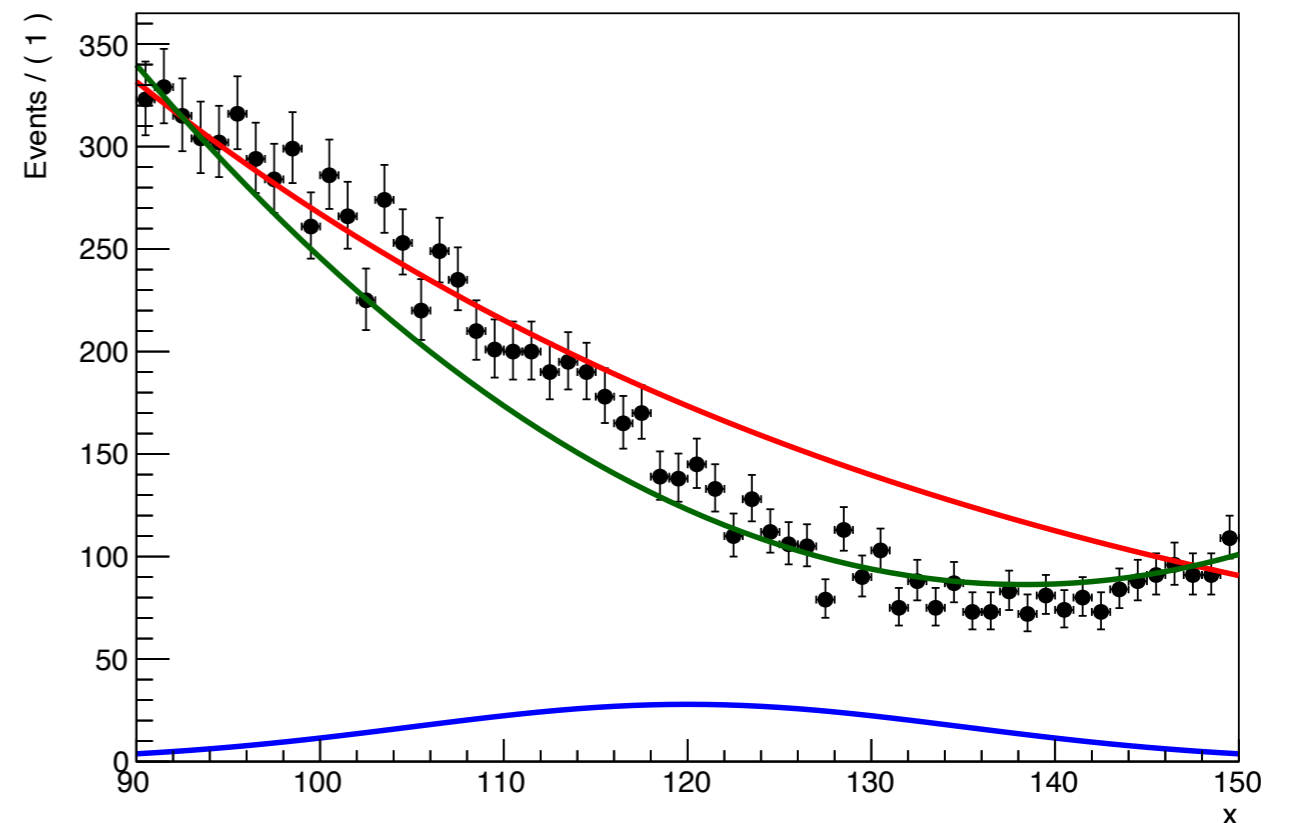
Interlude - sideband choice

- Sideband should be chosen carefully
 - "turn-on" effects (from analysis selection ...)
 - Wide signal peak
 - Need enough sideband!



Interlude - sideband choice

- Sideband should be chosen carefully
 - "turn-on" effects (from analysis selection ...)
 - Wide signal peak
 - Need enough sideband!

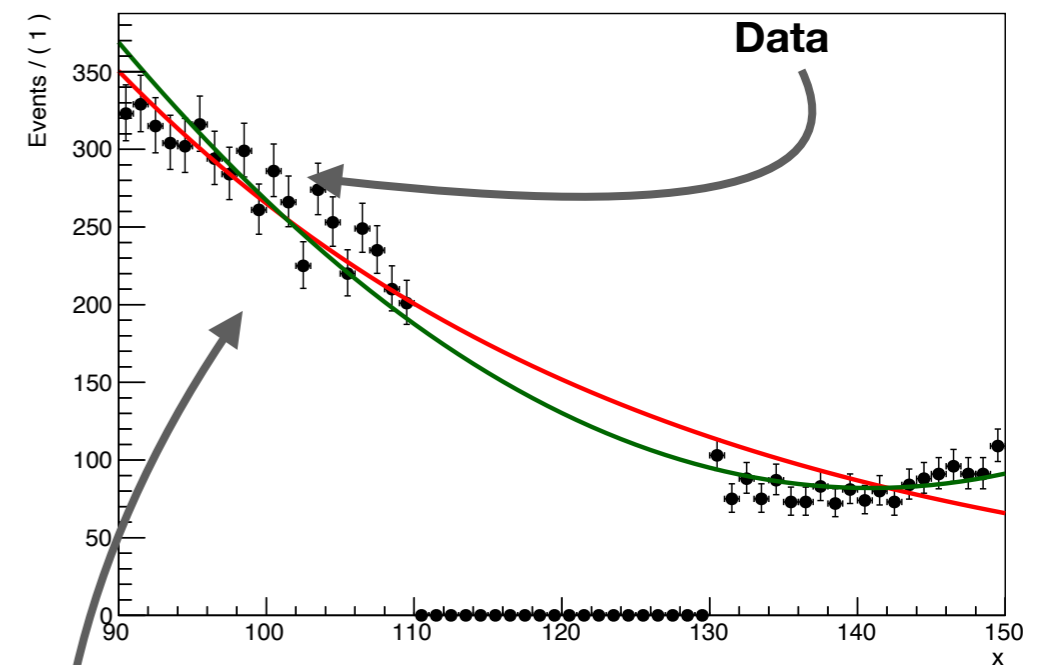
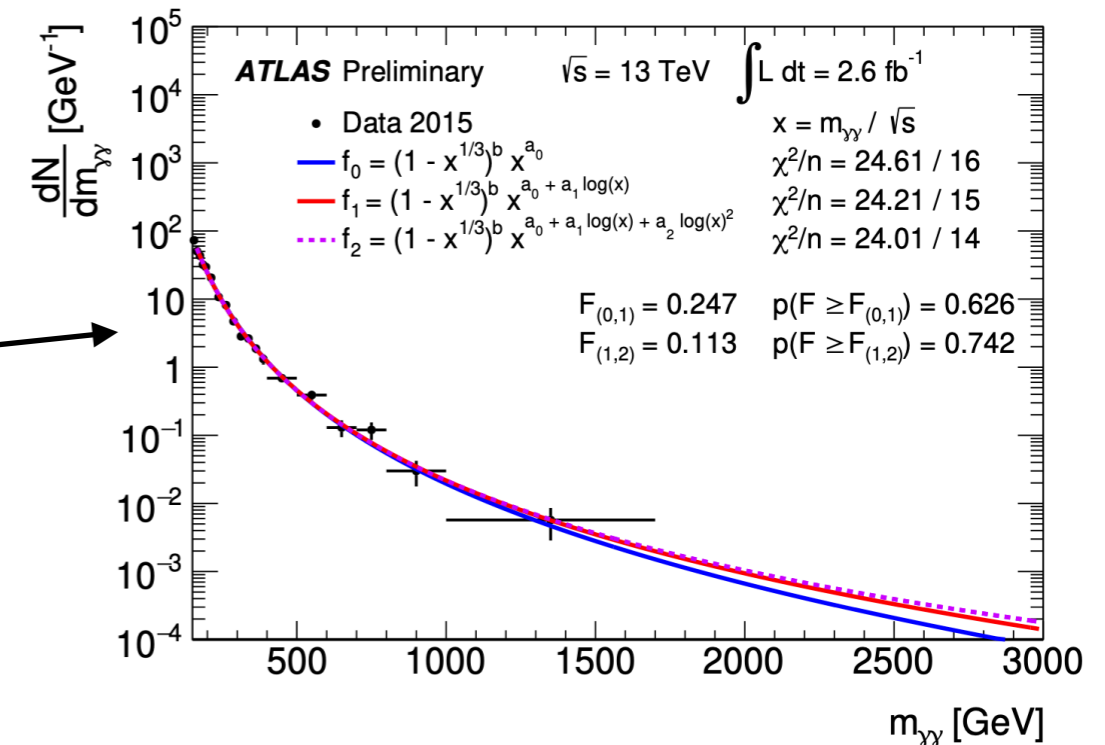


Back to the function determination

- Choose suitable function using
 - GoF
 - F-test

$$F = \frac{\frac{\chi_{\text{nom}}^2 - \chi_{\text{alt}}^2}{n_{\text{alt}} - n_{\text{nom}}}}{\frac{\chi_{\text{alt}}^2}{n - n_{\text{alt}}}}, \text{ follows F-distribution } F(n_{\text{alt}} - n_{\text{nom}}, n - n_{\text{alt}})$$

- Recall: no a-priori knowledge of the functional form
 - **Which one** do you pick?
 - How do you handle the **uncertainty** that arises from that choice?

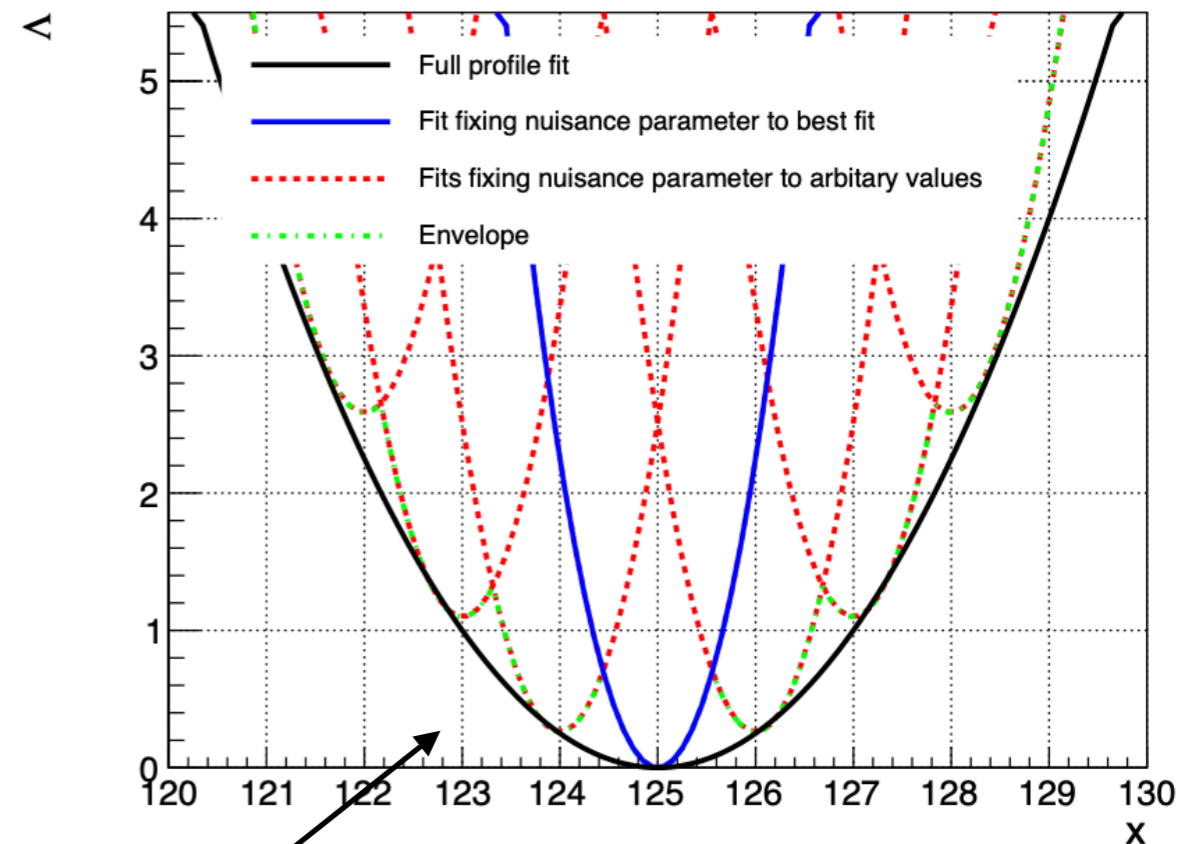


Systematics on analytic backgrounds

- **Two types** to consider:
- The parameters of the functional form
 - Usually freely floating
- The choice of functional form
 - **Discrete profiling**
 - **"Spurious signal" systematic**

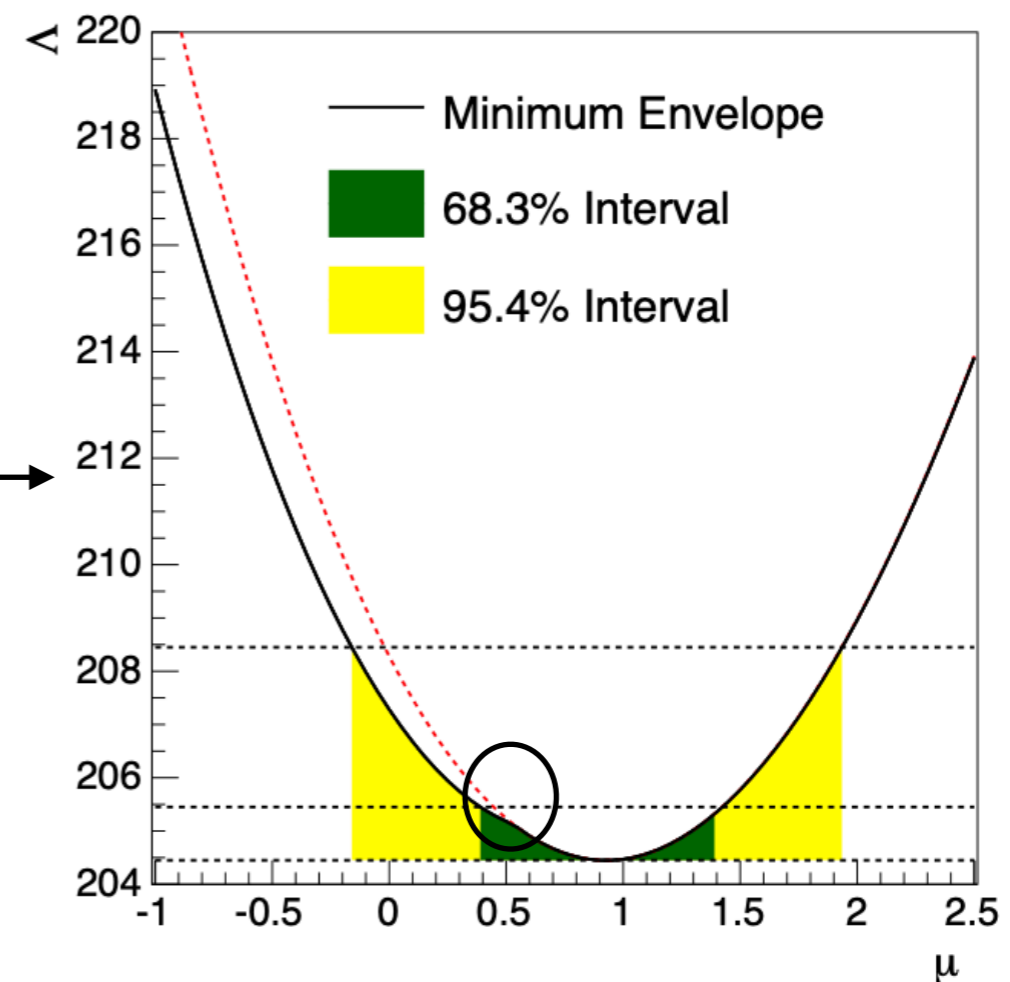
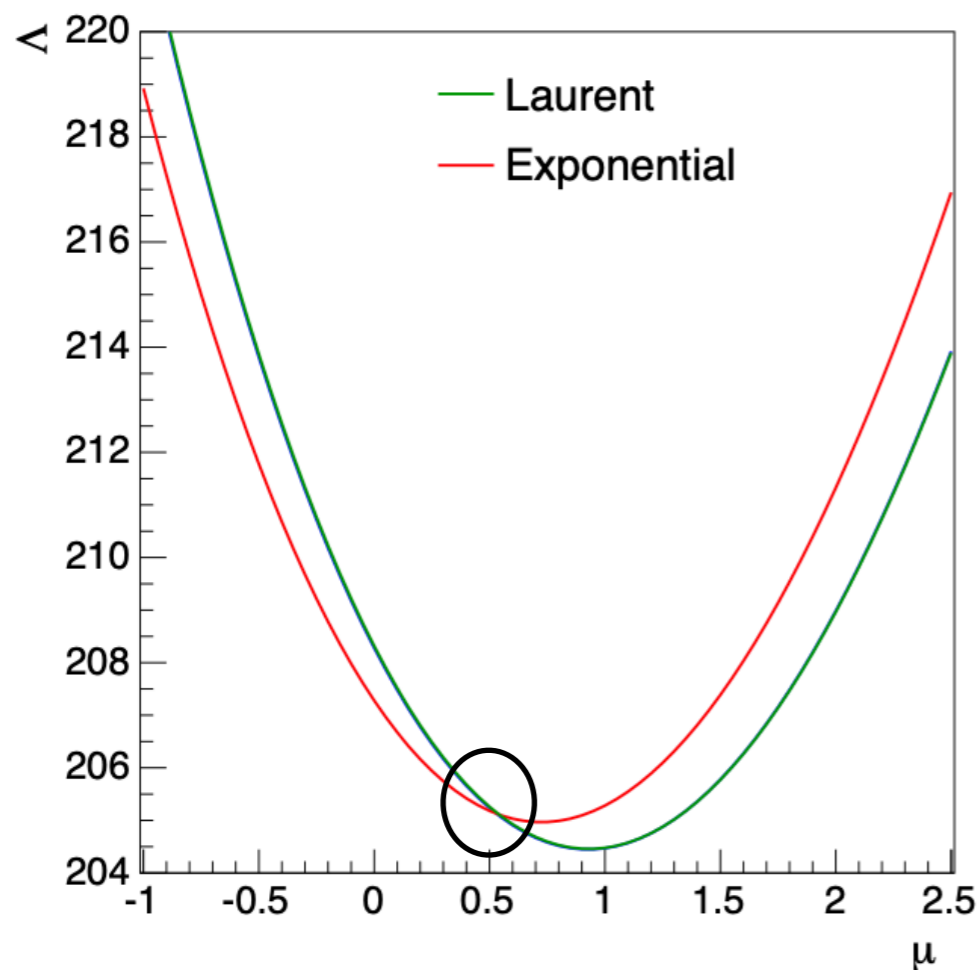
Discrete profiling

- Concept: treat choice of functional form as a nuisance parameter in the fit
 - Label each function with a discrete index and treat that index as an NP
 - "Normal" NPs are continuous - can minimization routine handle this?
 - → Exploit the fact that profile likelihood can be constructed by sampling arbitrary values of NPs



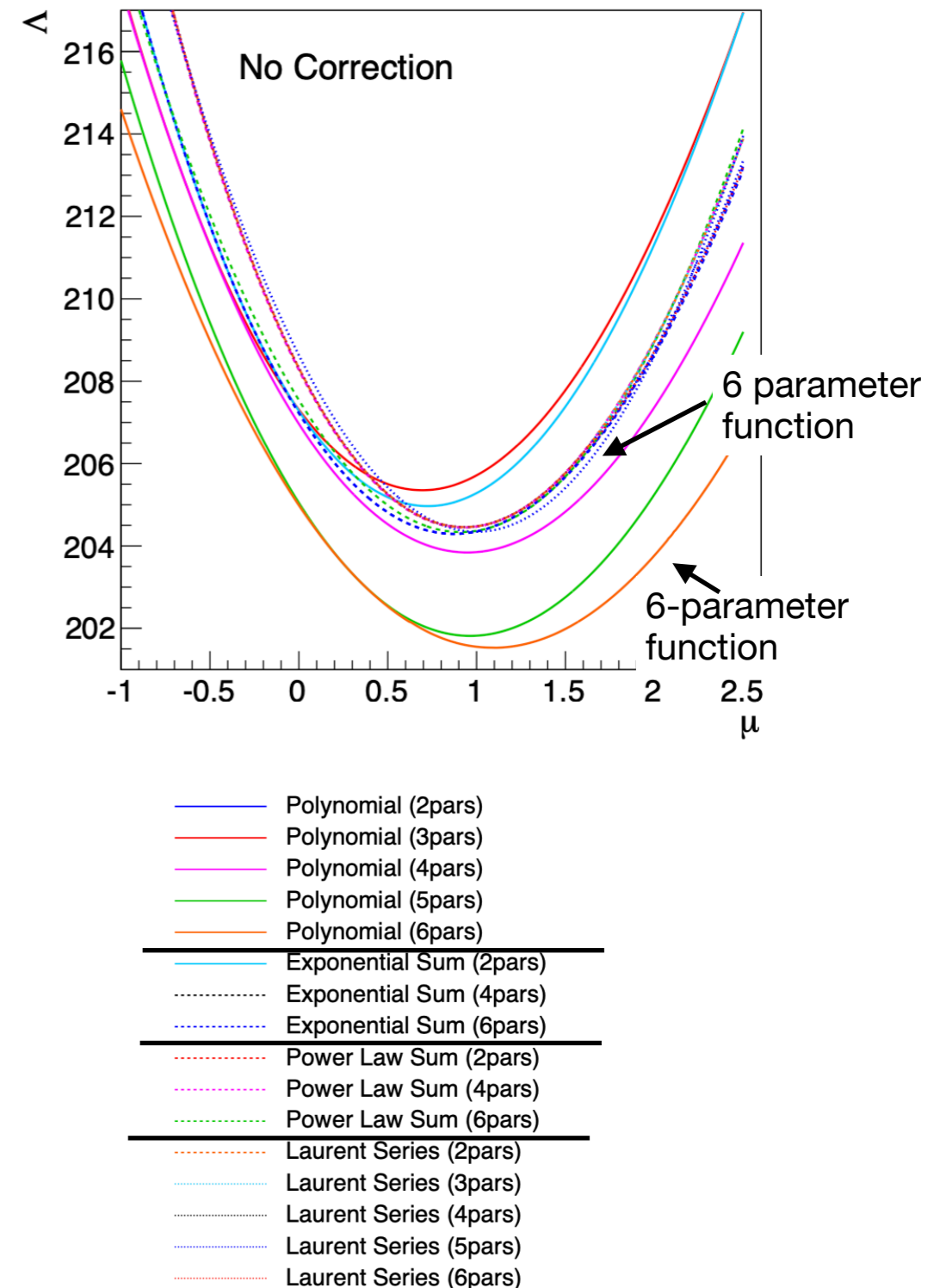
Discrete profiling - procedure

- Don't want to apply sampling to **all** NPs, only to the discrete indices
 - → Separate minimization for each function considered; construct profile likelihood from resulting curves
- NB this is the concept - in practice loop through the functions at each point to determine which gives the smallest NLL



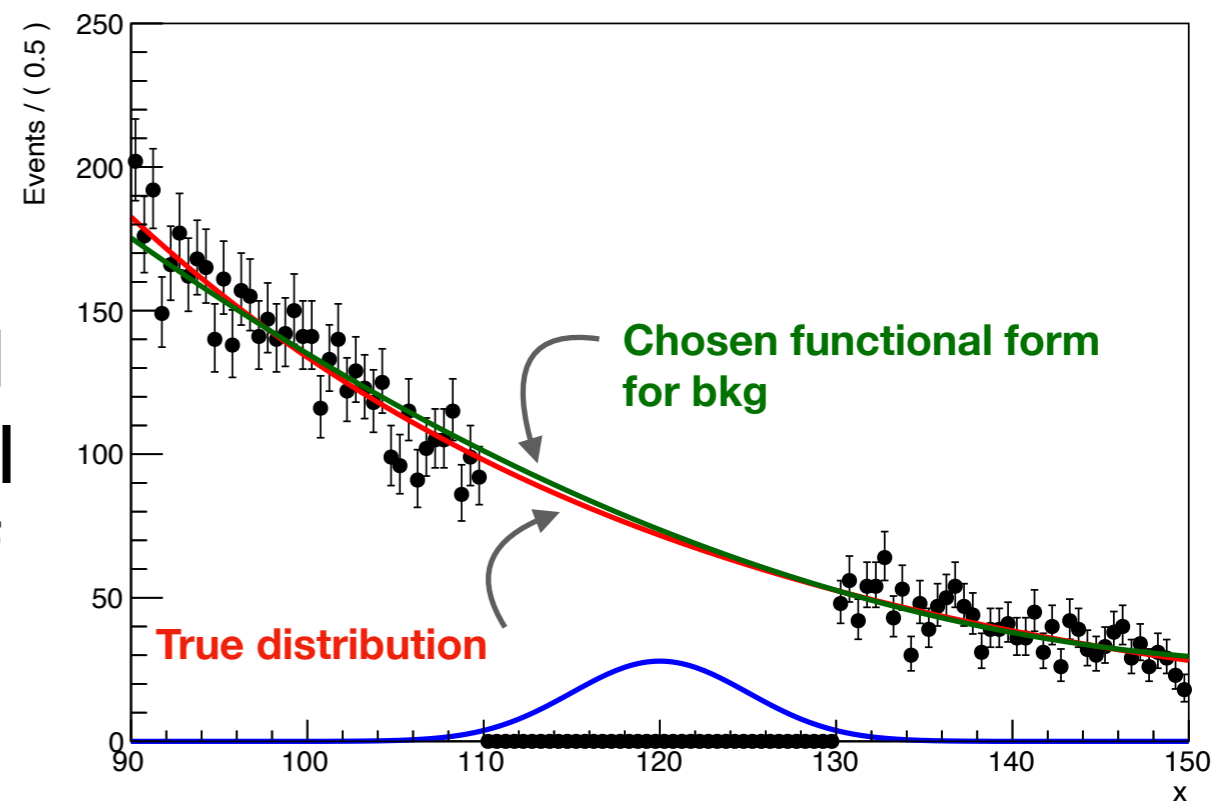
Functional form orders

- On previous slide, considered functions had the same number of parameters
- Nested function families \rightarrow highest considered order will give the minimum NLL
 - Usually solved using F-test!
 - \rightarrow apply correction to NLL to account for larger number of parameters
- Some degree of tradeoff - statistical power vs bias.
Default: $NLL \rightarrow NLL + 0.5 * N_{par}$



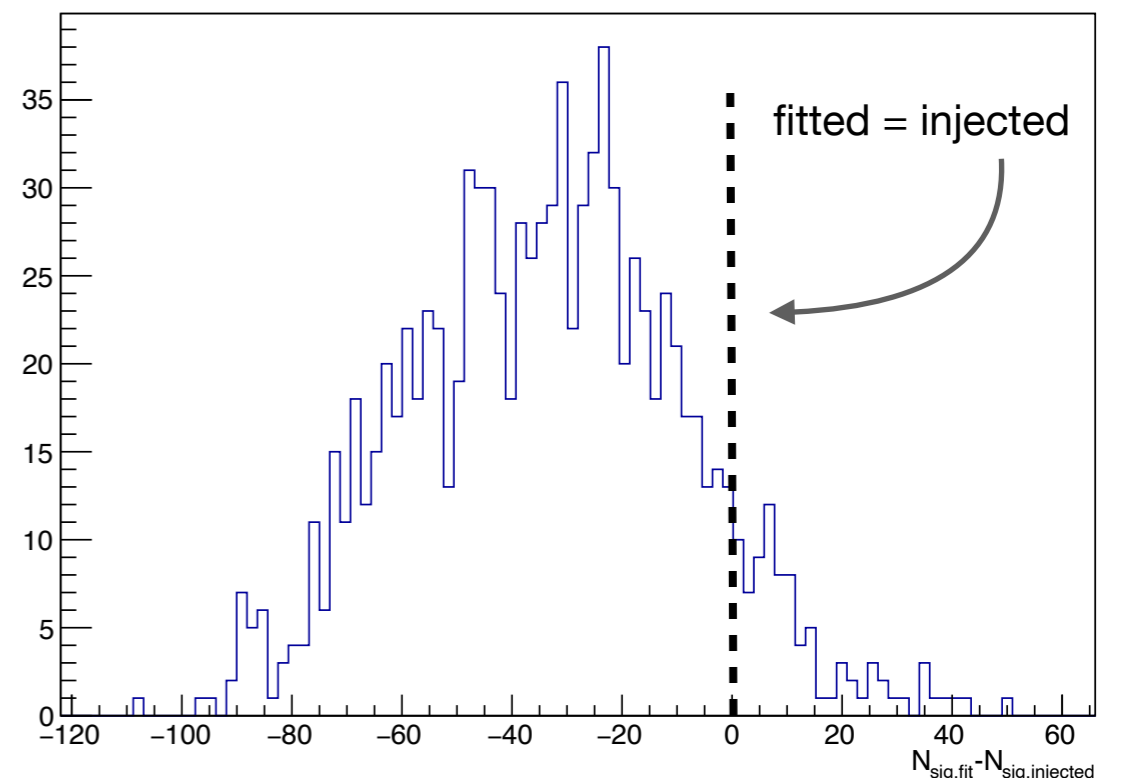
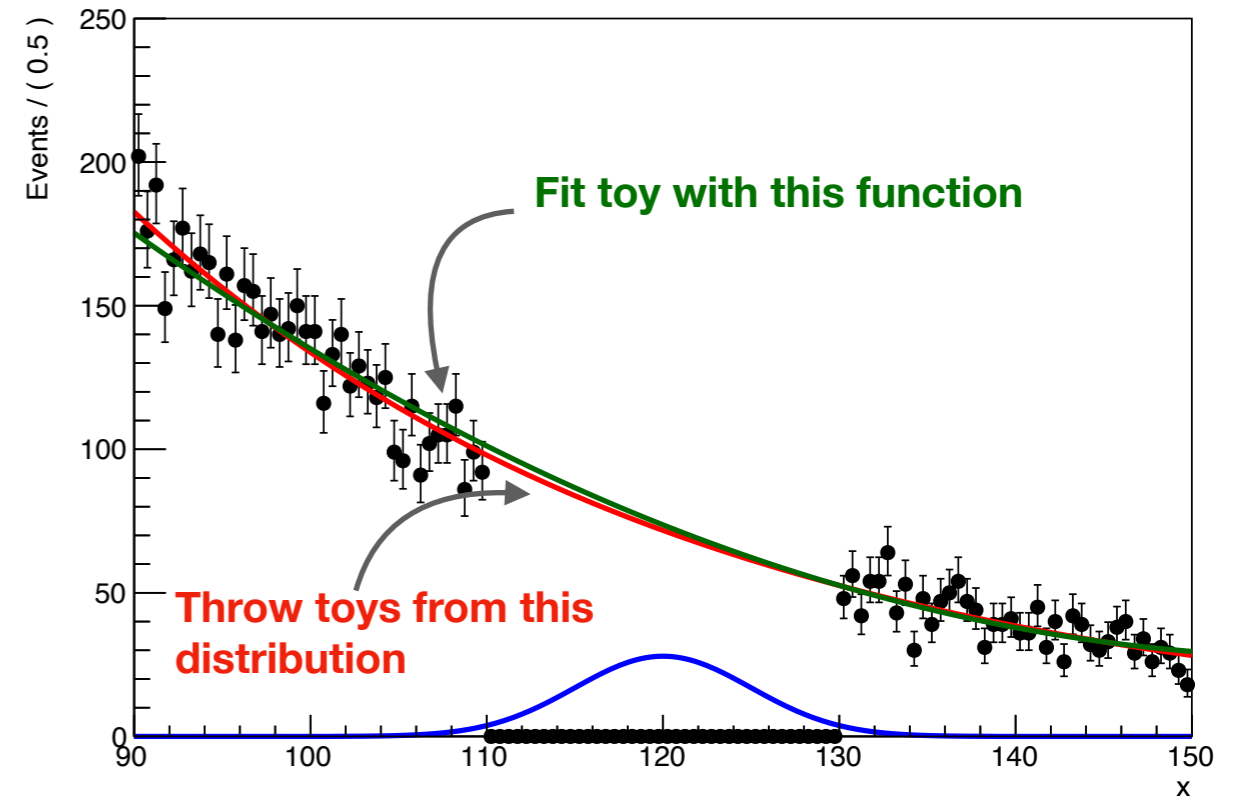
Spurious signal systematic

- A different approach to the same problem
- Choice of function \rightarrow impact on results
- Question to answer (example):
 - "How much does my measured amount of signal change when I fit with my chosen function A, if the true underlying distribution follows function B?"
- that is: how large is the **bias**?



Estimating the bias

- Generate toys from the **red** distribution with a known amount of **signal** added on top
- Fit those toys with the **green** background function + measuring the amount of **signal**
- Bias in the amount of fitted signal induced by fitting with the green function instead of the red: extracted from the distribution of fitted signal - injected signal



Estimating the bias

- Actually not just a single way in which this is done
 - Using a representative template
 - Data CR with large number of events
 - Simulation
 - **OR** Throwing toys from the template
 - → Requires some knowledge of the underlying distribution (from CR, or simulation)
- **OR** if no information is available:
 - Select the function with the smallest bias wrt alternatives
- Either way → include bias found from this procedure as 'spurious signal' in the likelihood

Including spurious signal in the likelihood

- Let's consider a binned likelihood

$$\mathcal{L}(\text{data} | \mu, \vec{\theta}) = \prod_i \text{Poiss}(n_i | (\mu \cdot s_i(\vec{\theta}) + b_i(\vec{\theta}))) \cdot p(\vec{\theta}_0 | \vec{\theta})$$

Parameter of interest (amount of signal, relative to exp.) Nuisance parameters Number of observed events Expected signal yield Background yield NP constraints

- Spurious signal modifies this:

$$\mu \cdot s_i(\vec{\theta}) \rightarrow \mu \cdot s_i(\vec{\theta}) + \text{bias} \cdot \theta_{\text{bias}}$$

N_{events} bias Unit gaussian

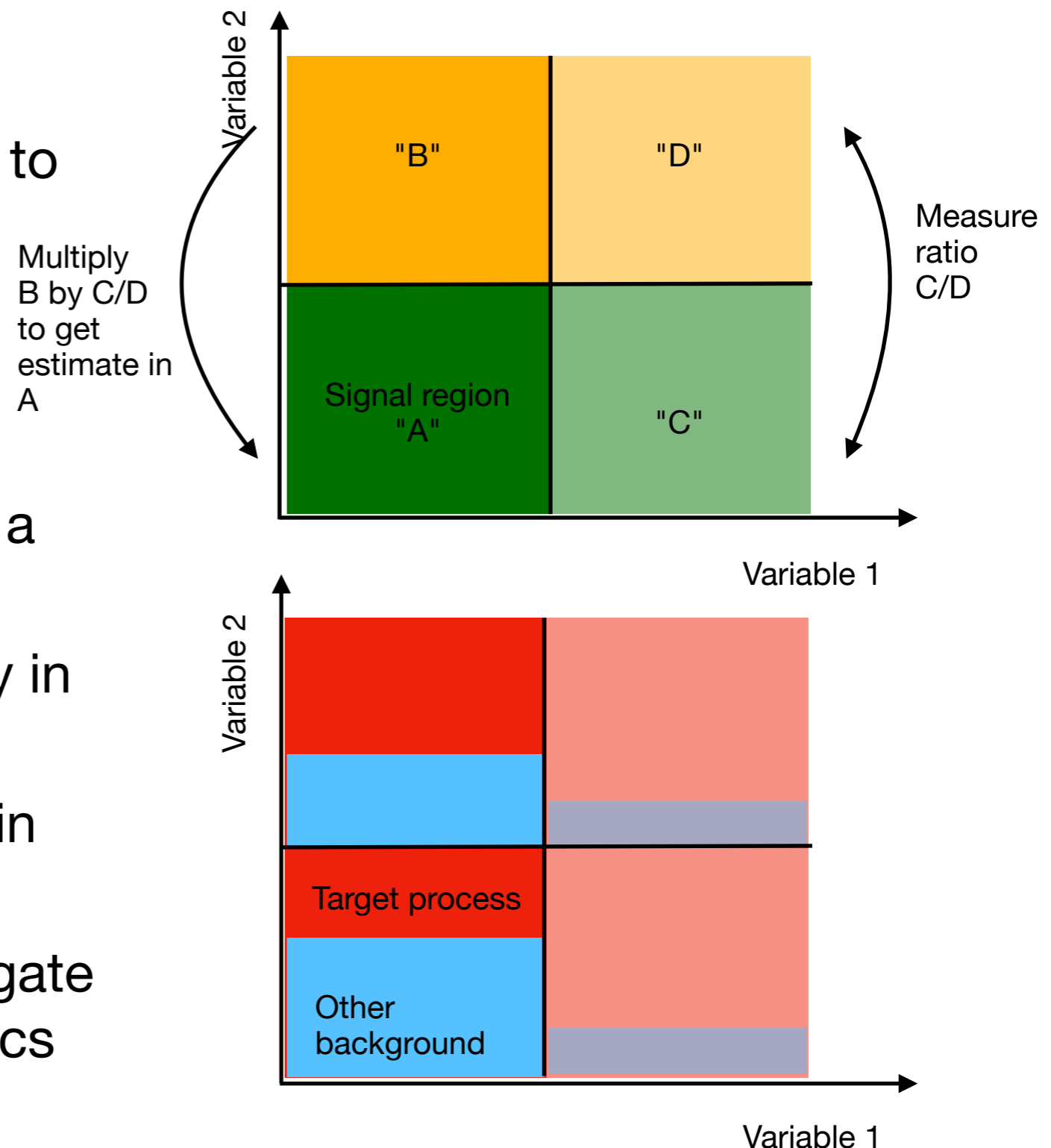
Histogram templates

Histogram templates

- Modelling data with histogram templates: can use simulation or control regions
- Will discuss a few topics related to this
 - Estimating backgrounds from data CRs
 - Morphing
 - Horizontal → signal morphing
 - Vertical → shape systematics
 - "Noisy templates"
 - Interpolation

Modelling background from CR's

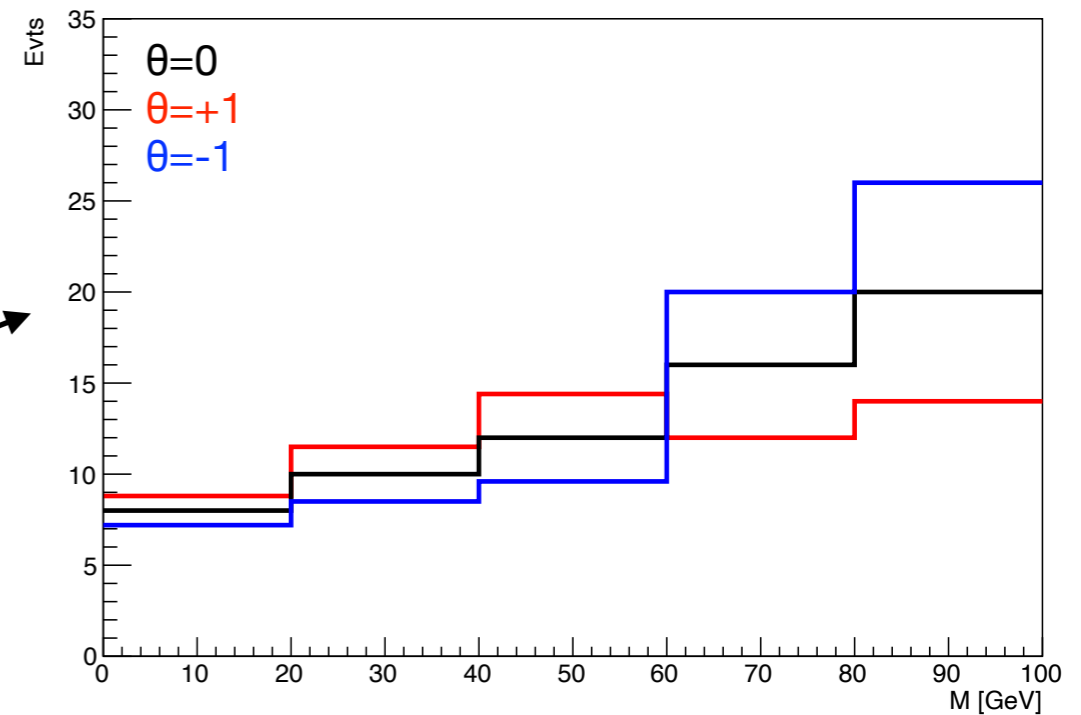
- ABCD method
 - Need 2 independent variables to define 3 control regions
 - Control regions should be enriched in background we're trying to estimate
- Could be applied for each bin of a template
 - Include control regions directly in the likelihood
 - Uncertainties taken care of in the likelihood
 - No need to manually propagate uncertainties on "other" procs



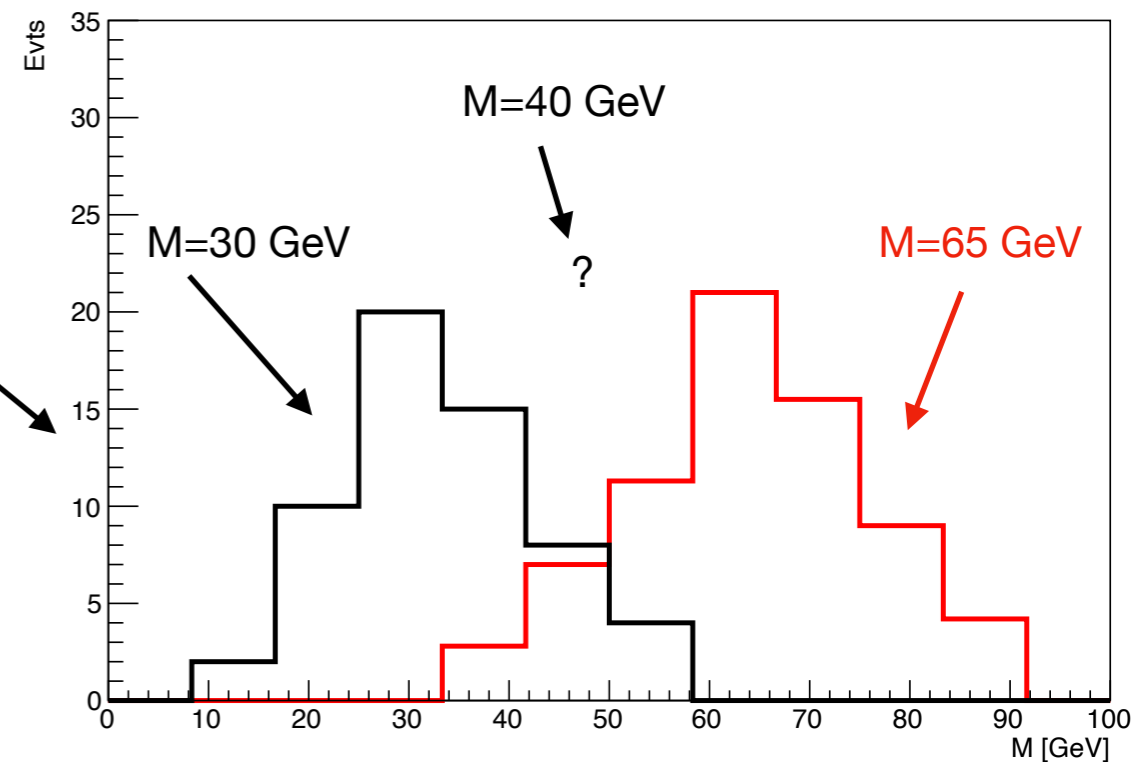
Morphing

- Where there are **templates** we need **morphing**

- "Vertical" template morphing for including systematic shapes

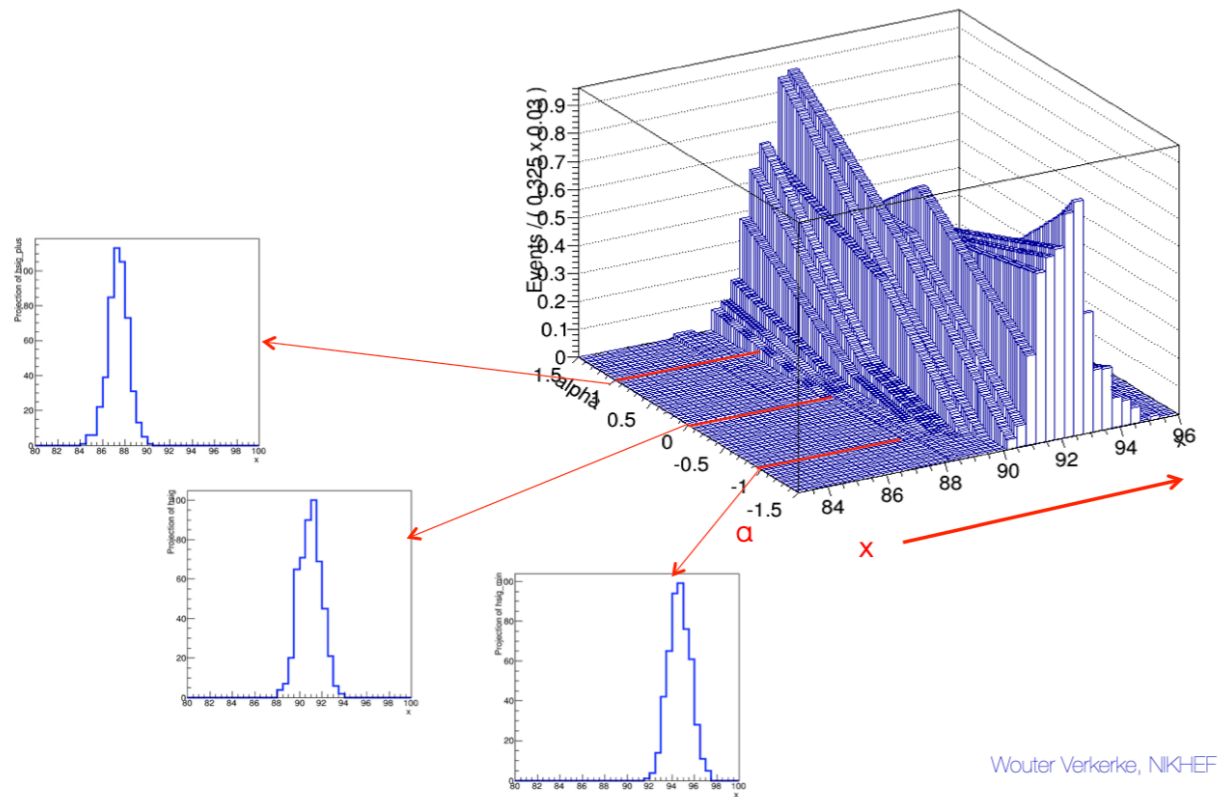


- "Horizontal" morphing e.g. for signal templates



Morphing limitations

Visualization of bin-by-bin linear interpolation of distribution

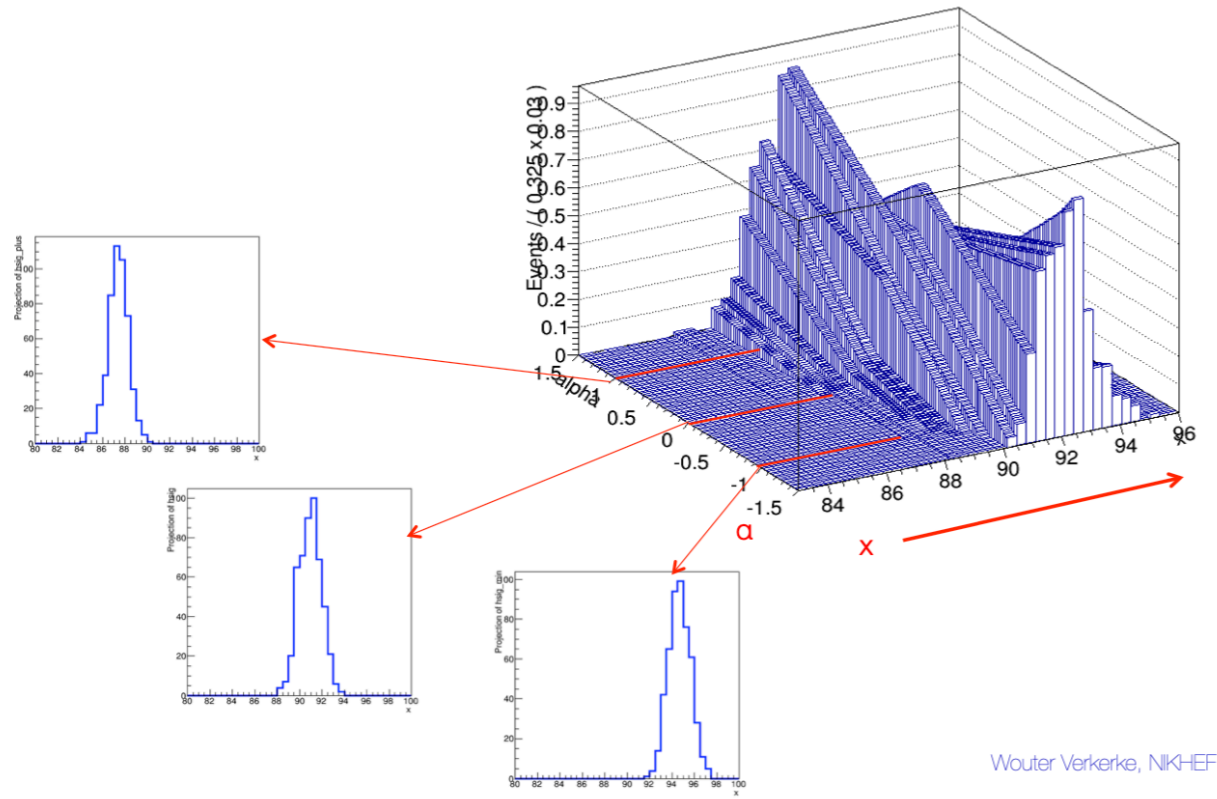


9

[*] From : <https://indico.cern.ch/event/507948/contributions/2028505/attachments/1262169/1866169/atlas-hcomb-morphwshop-intro-v1.pdf>

Morphing limitations

Visualization of bin-by-bin linear interpolation of distribution

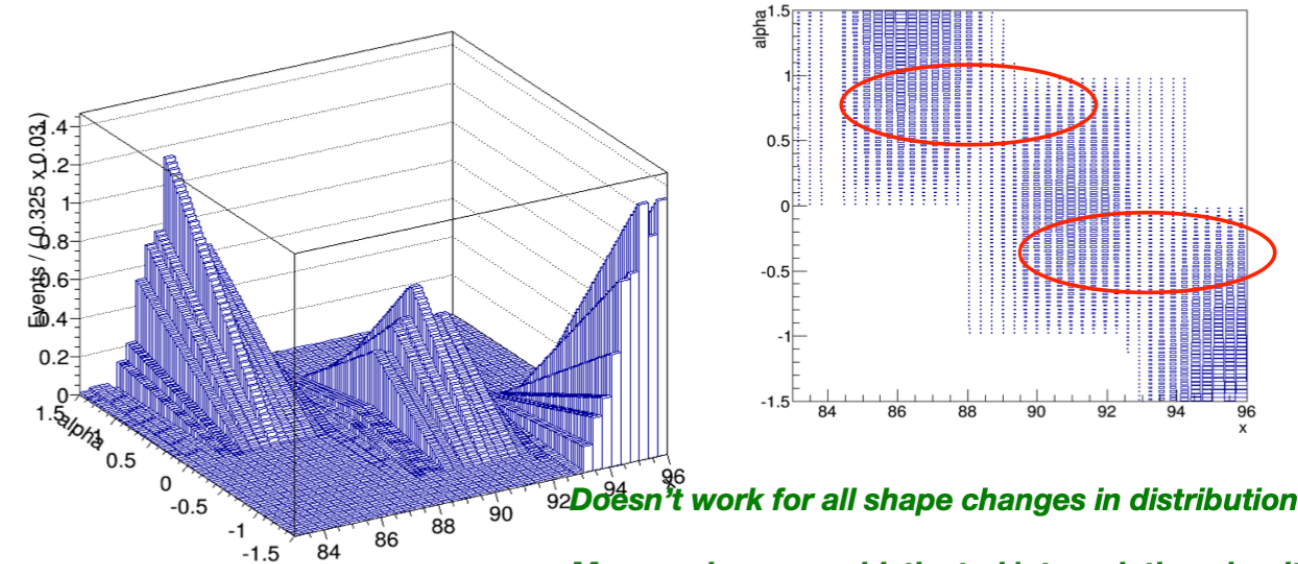


Wouter Verkerke, NIKHEF

Limitations of piece-wise linear interpolation

- Bin-by-bin interpolation looks spectacularly easy and simple, but be aware of its limitations
 - Same example, but with larger 'mean shift' between templates

Note double peak structure around $|\alpha|=0.5$

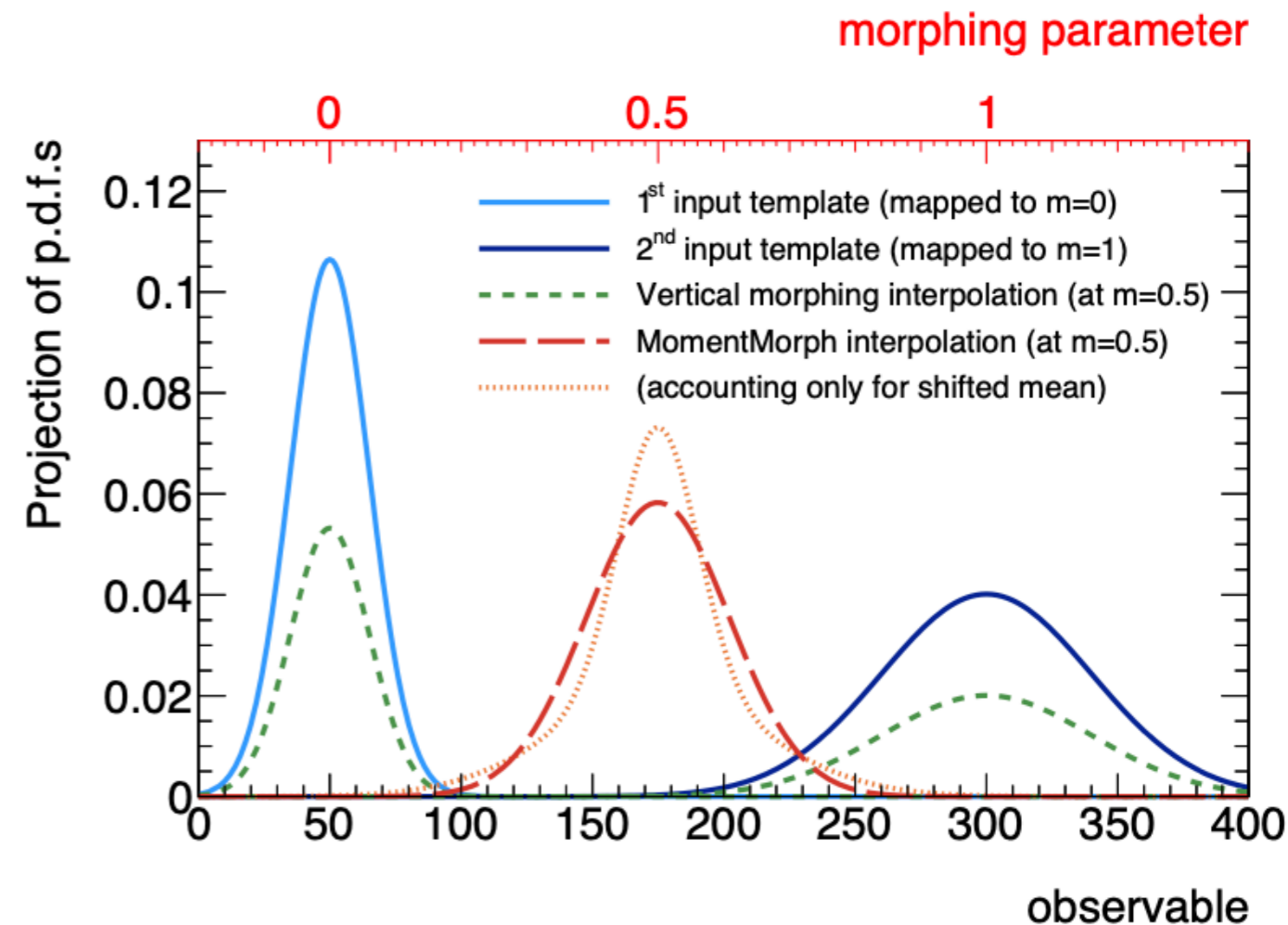


Doesn't work for all shape changes in distributions
May need more sophisticated interpolation algorithms
→ will show solutions later

[*] From : <https://indico.cern.ch/event/507948/contributions/2028505/attachments/1262169/1866169/atlas-hcomb-morphwshop-intro-v1.pdf>

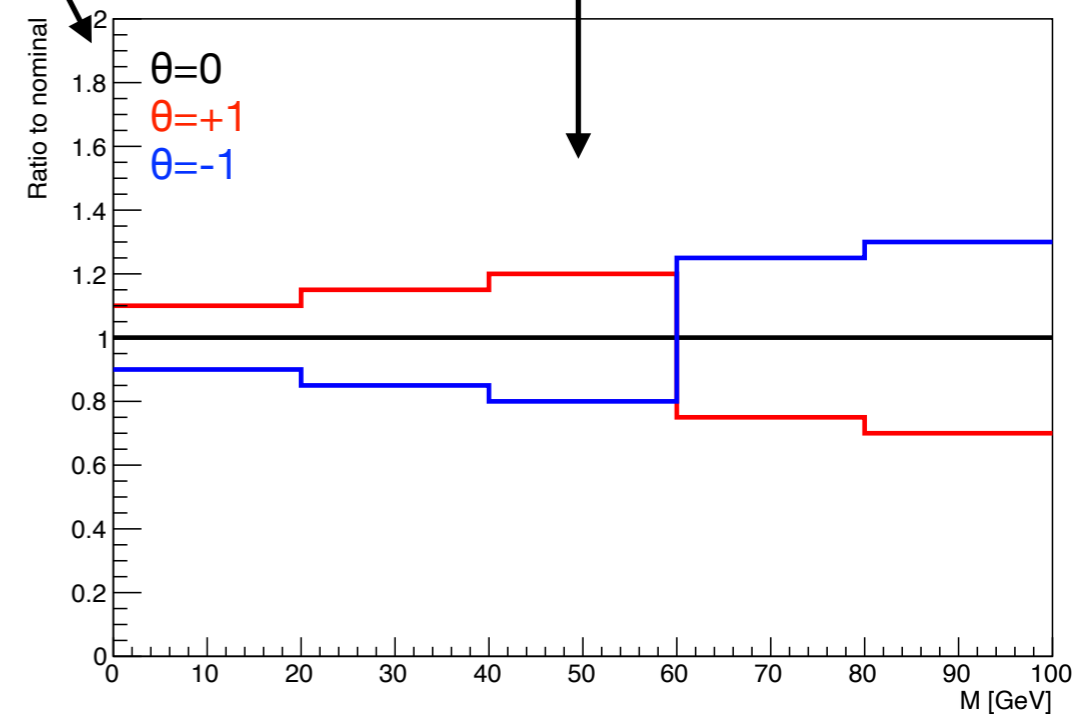
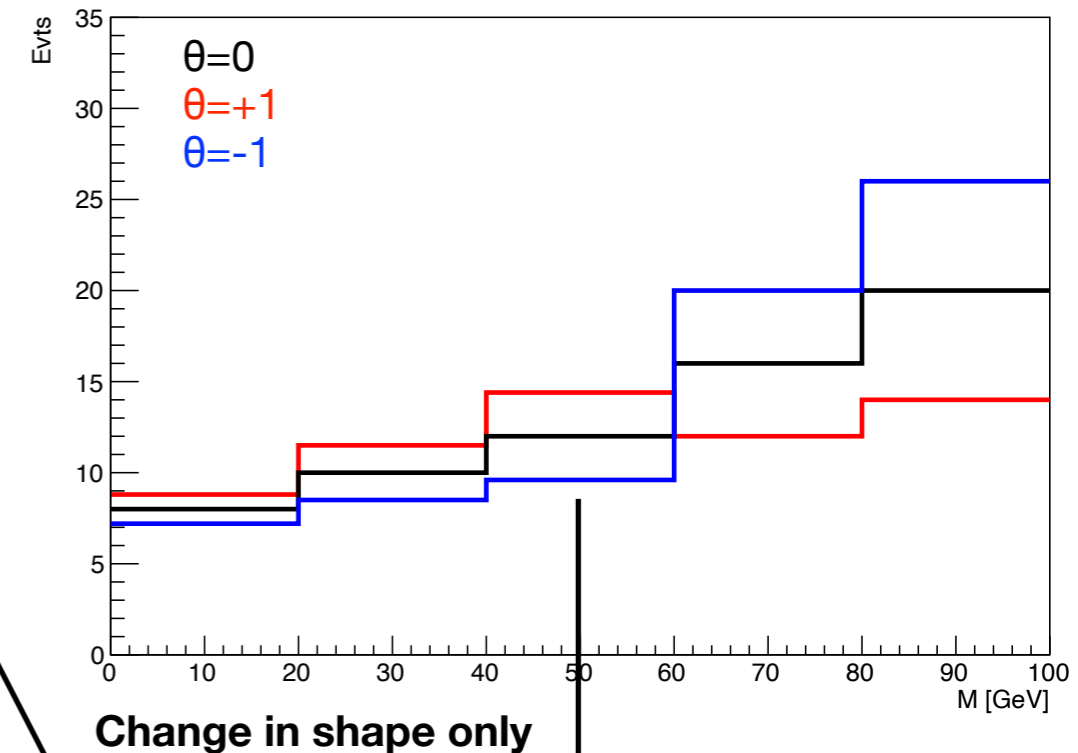
Moment morphing

- Instead of "pure" vertical or "pure" horizontal morphing, use vertical morphing & translation of input templates mean, rms to match at the morphed point
- Can be used for modelling shape-altering systematic uncertainties, complex signal models....



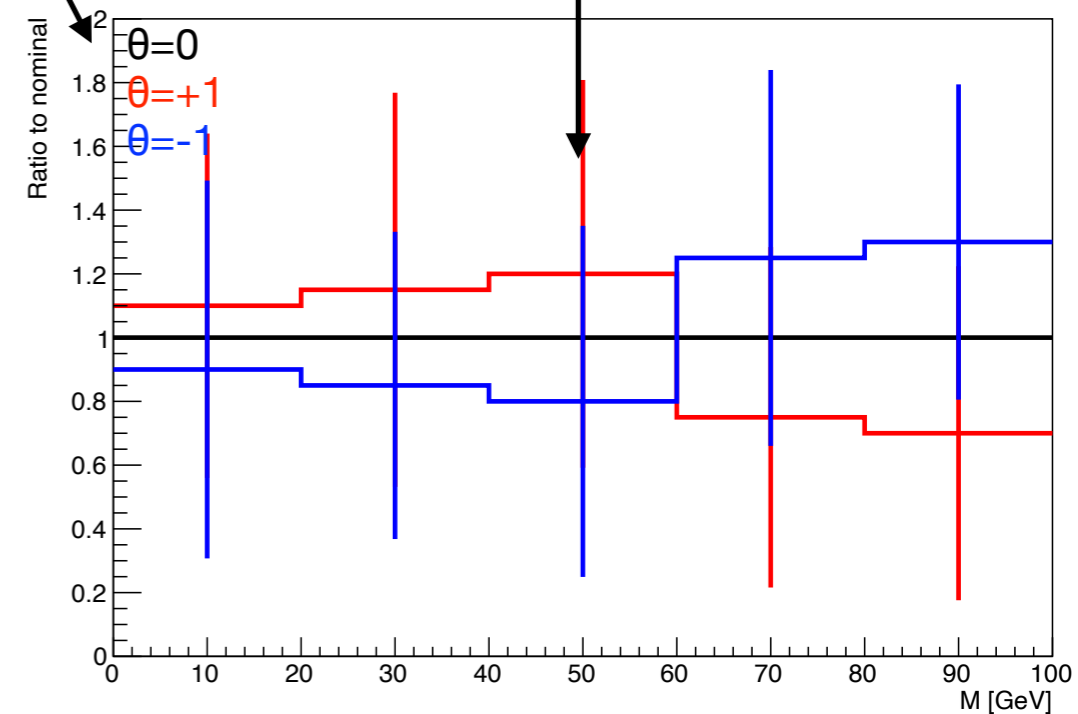
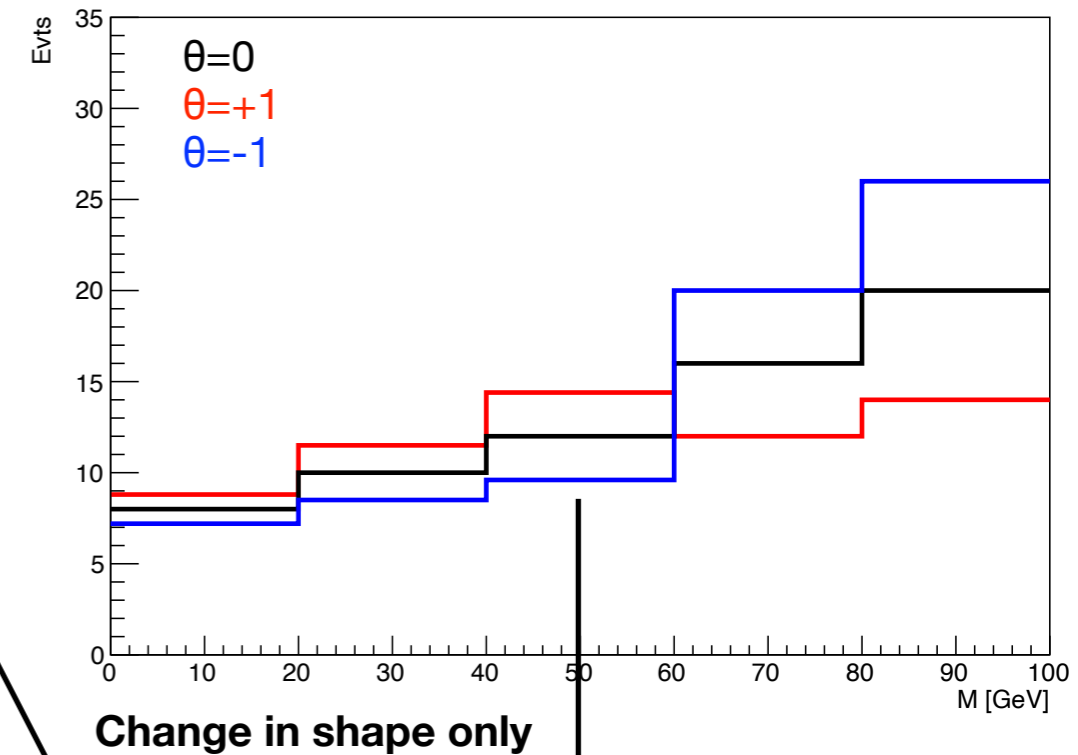
On shape-changing systematics

- Are shape-changing effects **genuine**?
 - Don't want to model "noise"
- Consider the example given earlier
 - Looks like the changes are large!



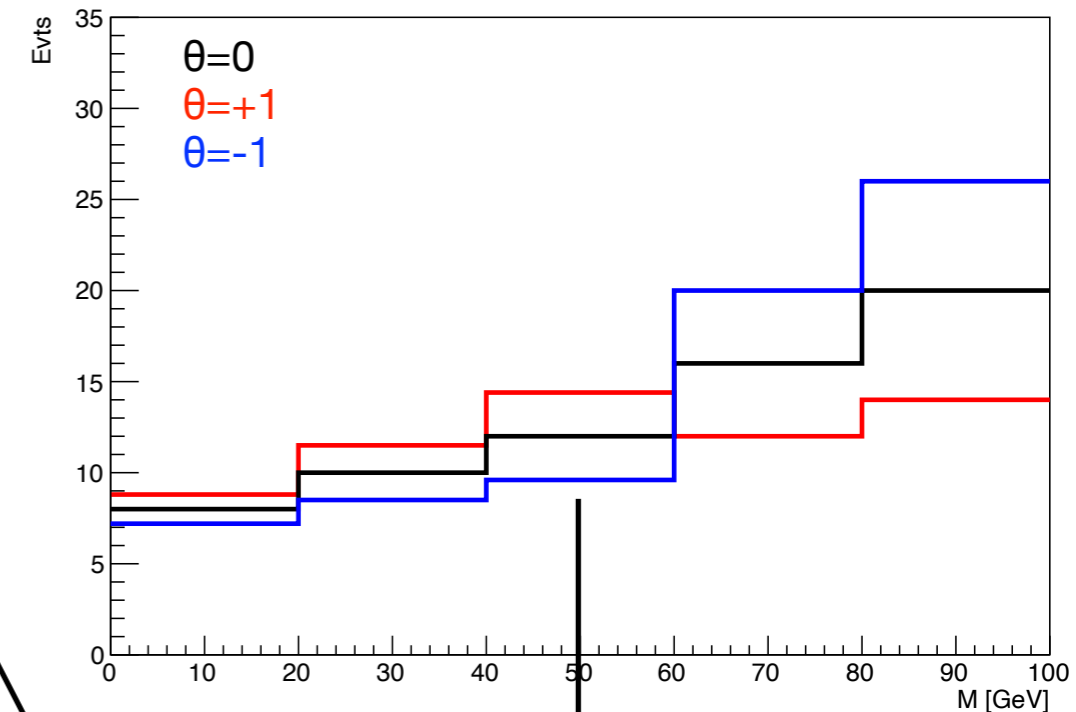
On shape-changing systematics

- Are shape-changing effects **genuine**?
 - Don't want to model "noise"
- Consider the example given earlier
 - Looks like the changes are large!
- Now let's also look at the uncertainties

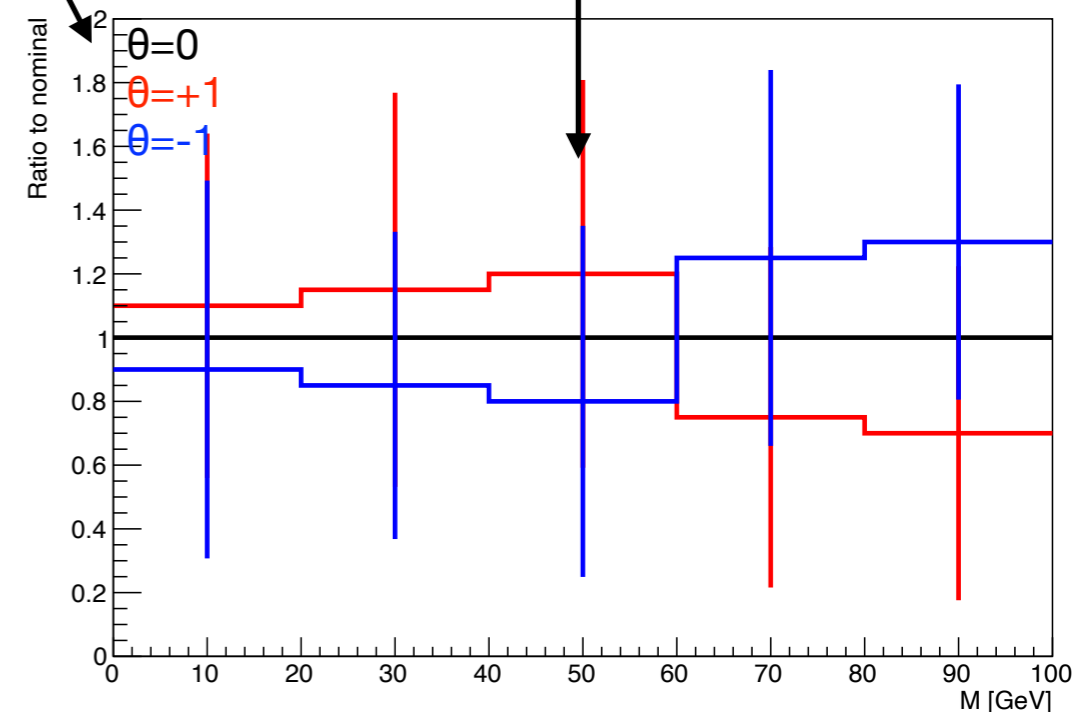


On shape-changing systematics

- Are shape-changing effects **genuine**?
 - Don't want to model "noise"
- Consider the example given earlier
 - Looks like the changes are large!
- Now let's also look at the uncertainties
- Extreme example
 - Can end up in such a situation, unfortunately → not easy to take these uncertainties into account

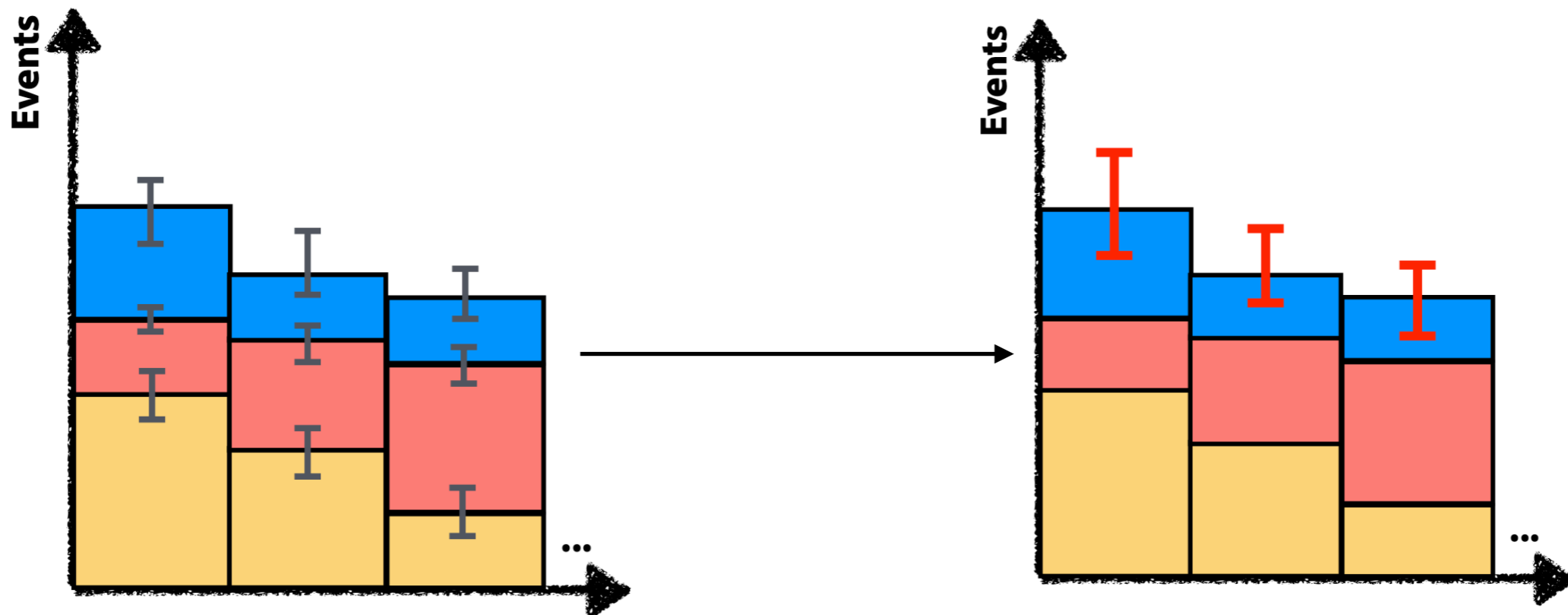


Change in shape only



Accounting for limited sample size

- Can't account for "noisy" shape systematics, but we **can** account for uncertainties due to the limited sample size used for creating the template
- Barlow-Beeston method: M processes, N bins \rightarrow $M \times N$ nuisance parameters
 - **Lite** version: assuming process uncertainties can be modelled with a Gaussian pdf, sum of M gaussians is also a gaussian \rightarrow 1 nuisance parameter per bin!
- NB: Gaussian pdf only a good approximation for large event counts



Barlow-Beeston method
[Comput. Phys. Commun 77 \(1993\) 219](#)

Barlow-Beeston-**lite** method
[arXiv:1103.0354](#)

Summary

- Modelling the data well is important for physics analyses
- Have shown techniques used for including modelling systematics in the likelihood
- Shown some issues that can appear when using these methods