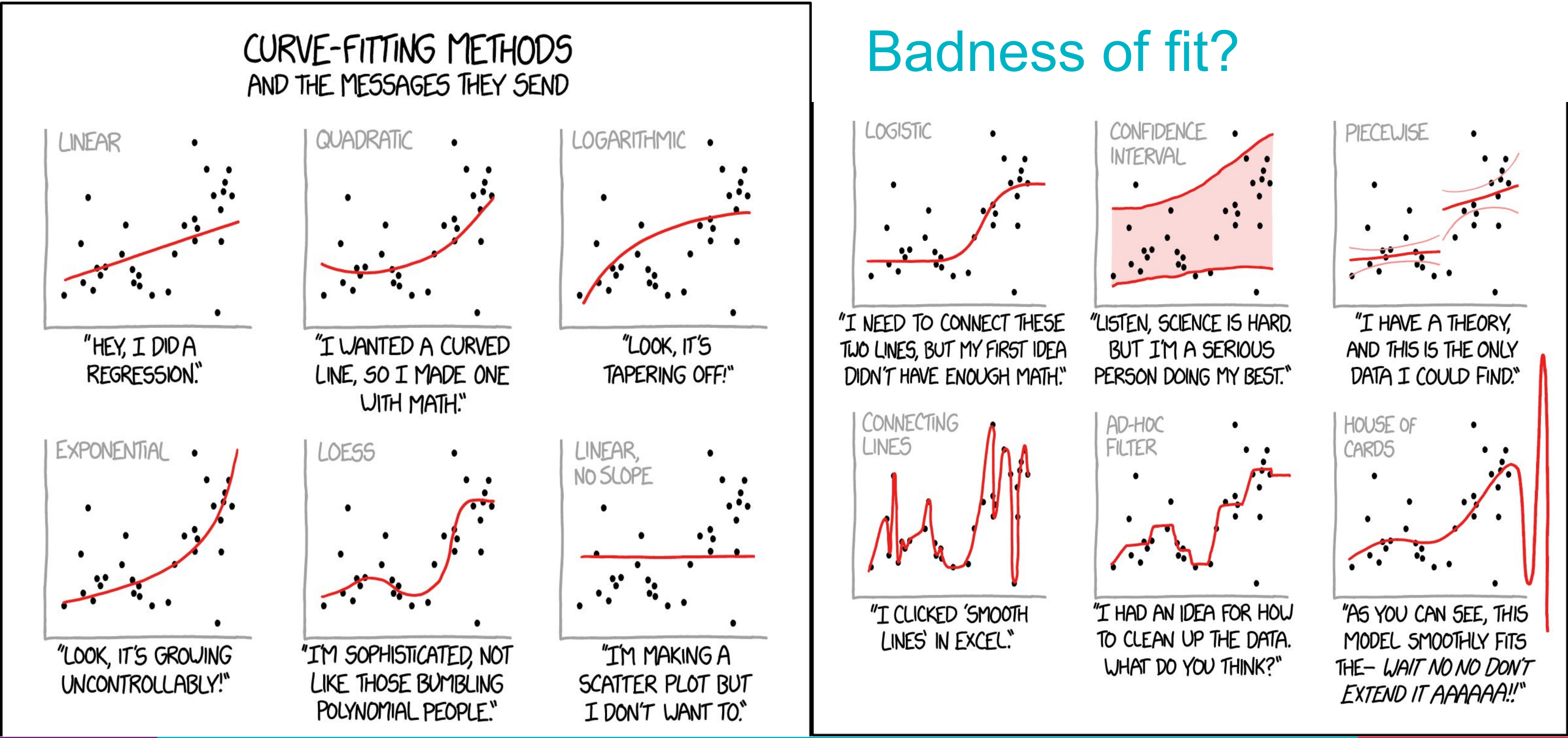
Validation & Evaluation in Physics: How we do statistics in HEP, how we need to change it for ML purposes

What we don't know and what we do wrong

Lydia Brenner

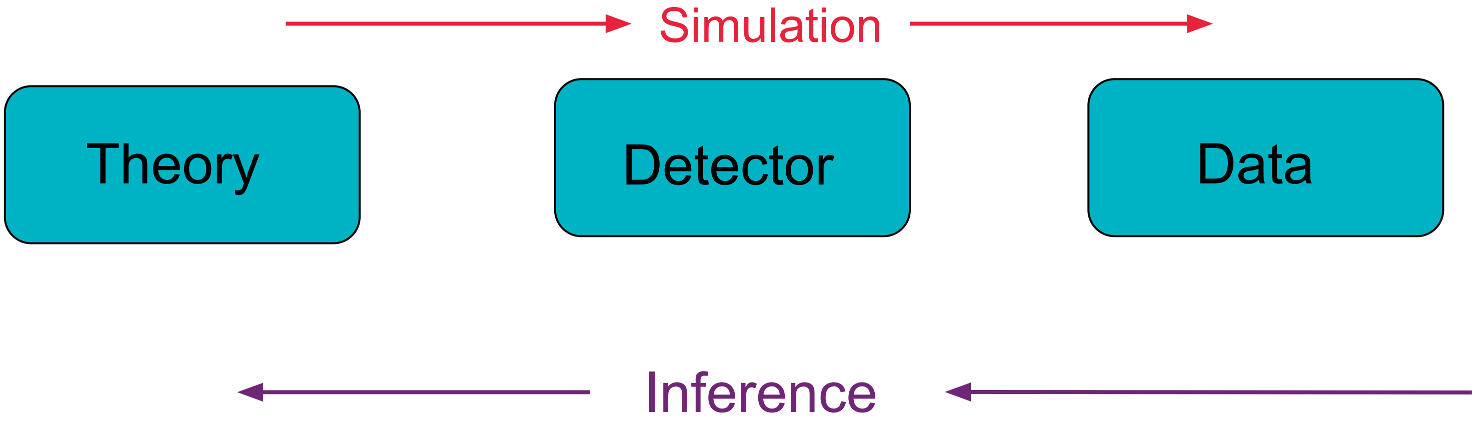
Nikhef

Badness of fit?



Physics analysis in a nutshell

Simulation based inference: the classical way



Physics analysis in a nutshell

Simulation based inference: the classical way

Monte Carlo Sampling

Observables

$$\boldsymbol{\mu} \longrightarrow f(x|\boldsymbol{\mu},\boldsymbol{\theta}) \longrightarrow \{x\} \longrightarrow \{\mathcal{Y}\}$$

Theory

Detector

Data

Standard Model

Events

$$\boldsymbol{\mu} \longleftarrow \mathcal{L}(\mathcal{Y}|\boldsymbol{\mu},\boldsymbol{\theta}) \longleftarrow \{\mathcal{Y}\}$$

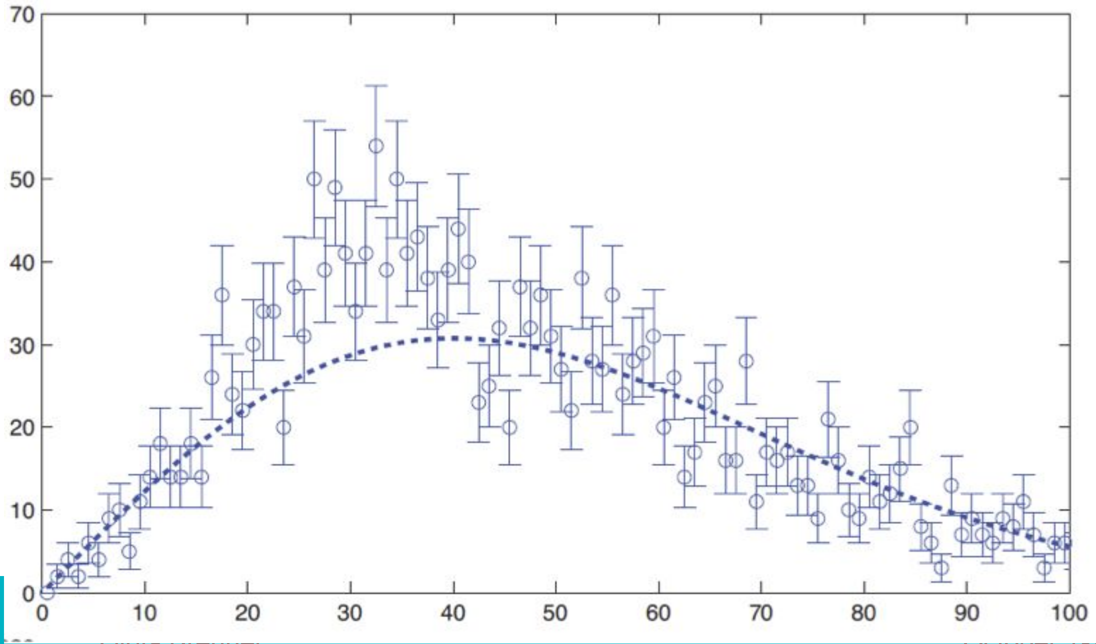
The question: Does my hypothesis describe the data

Basic question: what doe we want from the answer?

- → Would like a clearly understandable number
- → Would like it to match with visual input
- → Would like it to have a meaningful interpretation in terms of the likelihood

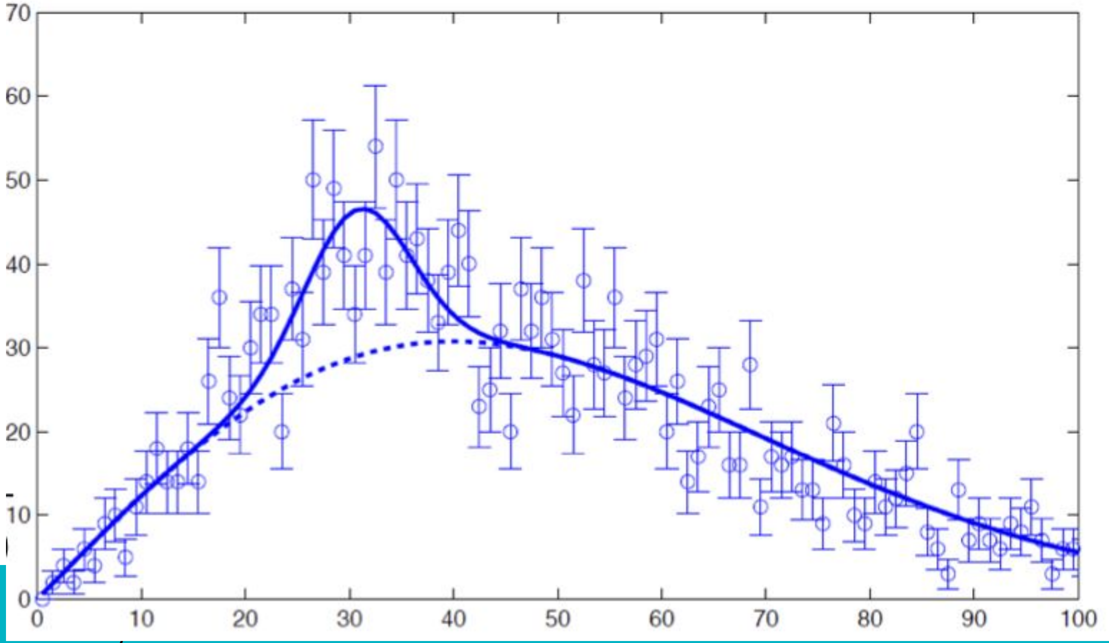
Is there a signal?

Try to distinguish background fluctuations from signals.



Is there a signal?

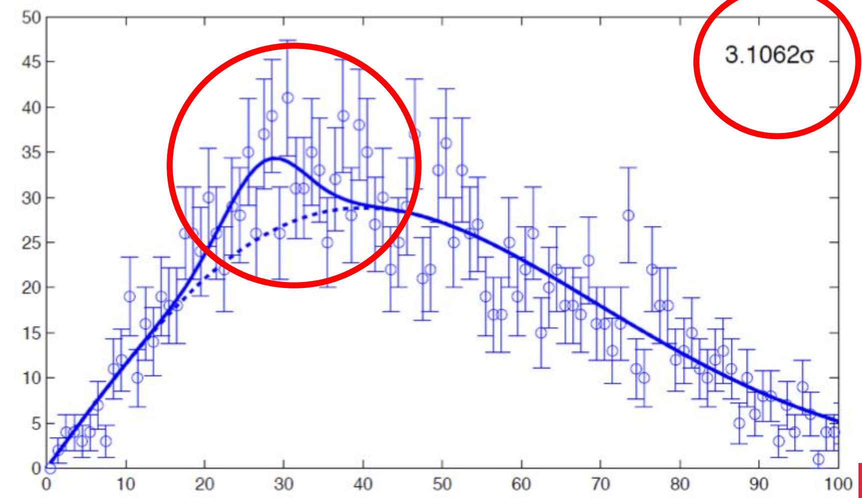
Looks like a signal around m=30 maybe



So is there a signal?

By eye; does not *really* look like a signal but: >3 sigma!

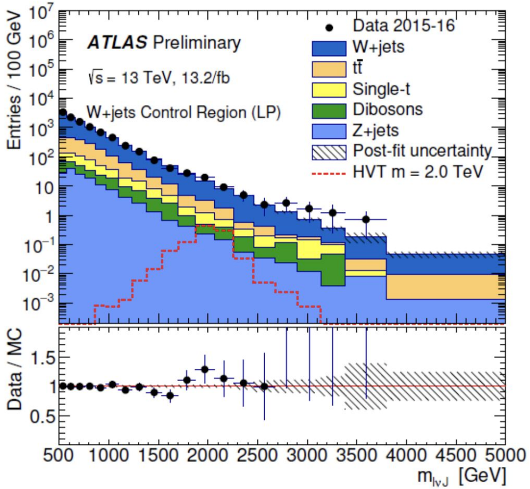
$$q_{fix,obs} = -2\ln\frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$$



Do we use by eye?

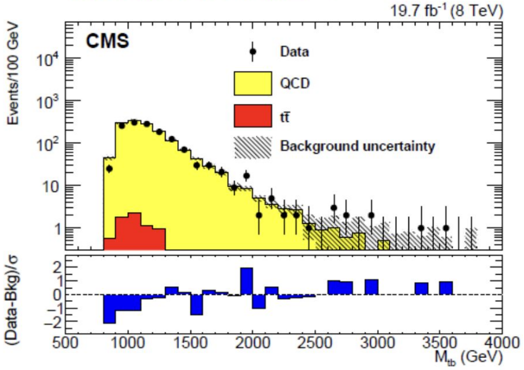
Yes!

ATLAS-CONF-2016-062 Diboson resonance

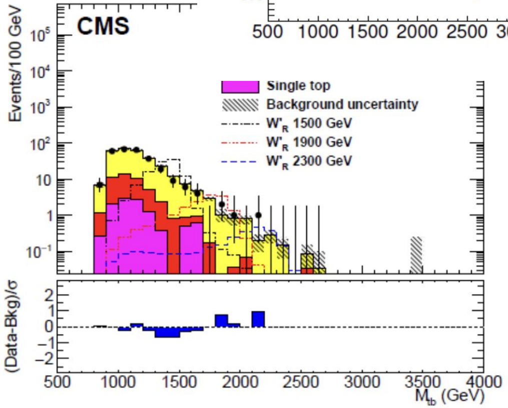


“Good agreement is observed (optical inspection of pulls) between the data and the background prediction”

arXiv:1509.06051v2 W′ → tb

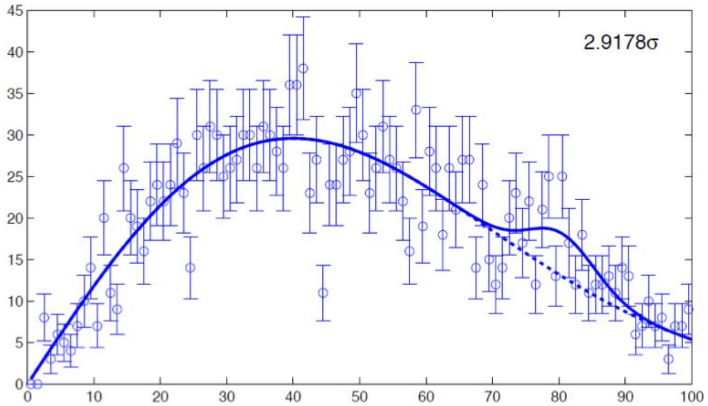


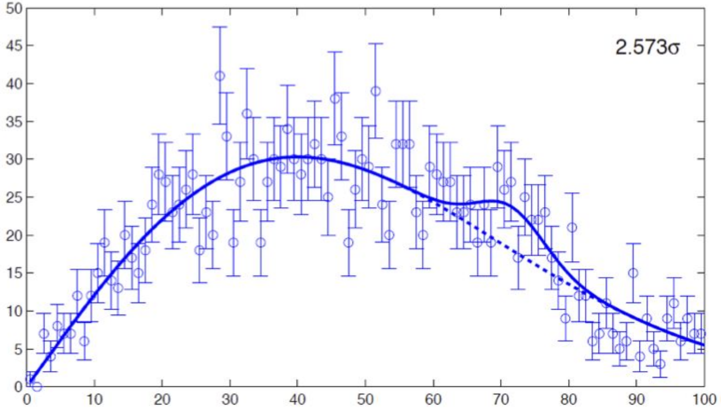
“This test (optical inspection) shows good agreement ”

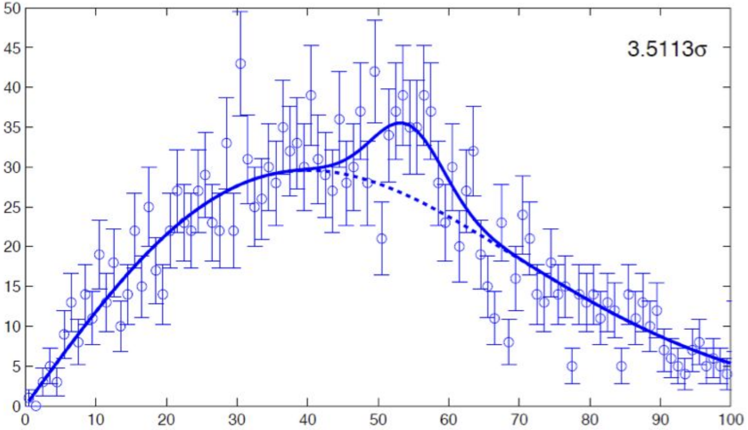


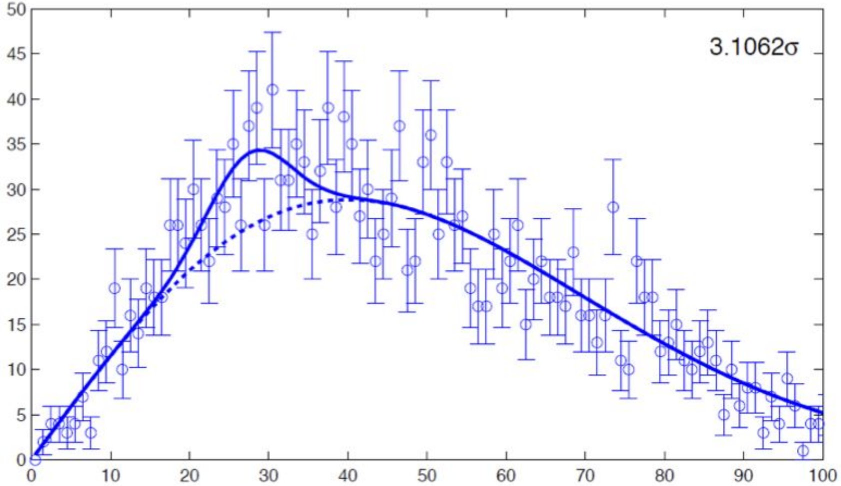
“Good agreement (optical inspection) between data and expectation from SM”

What about these signals?

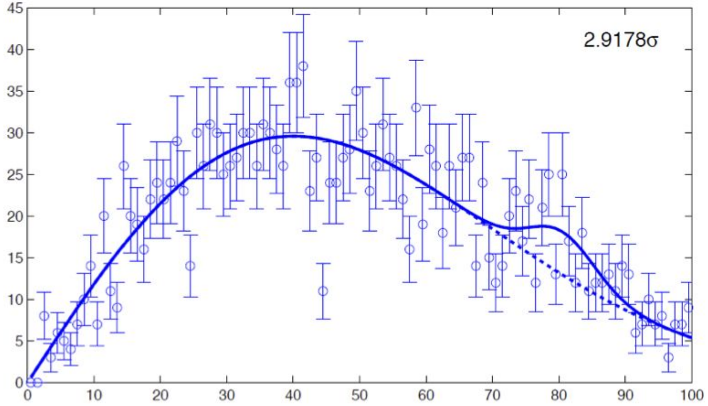


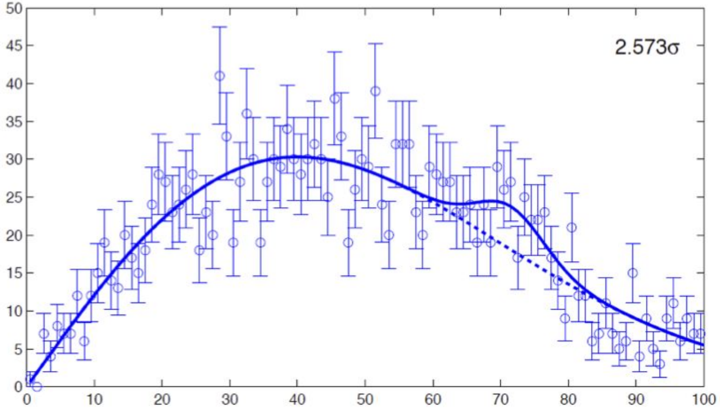




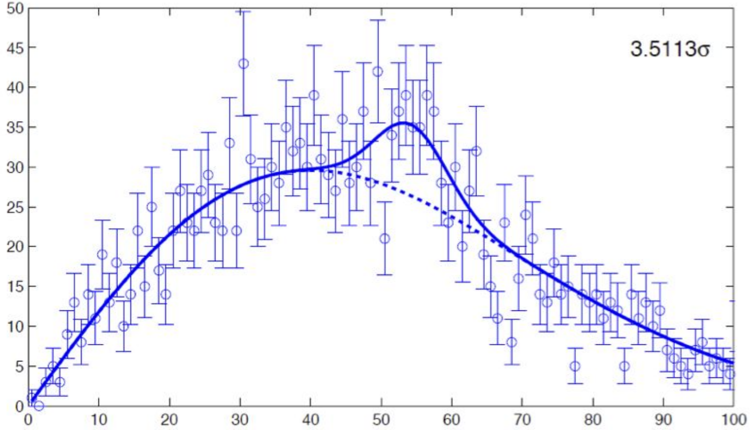


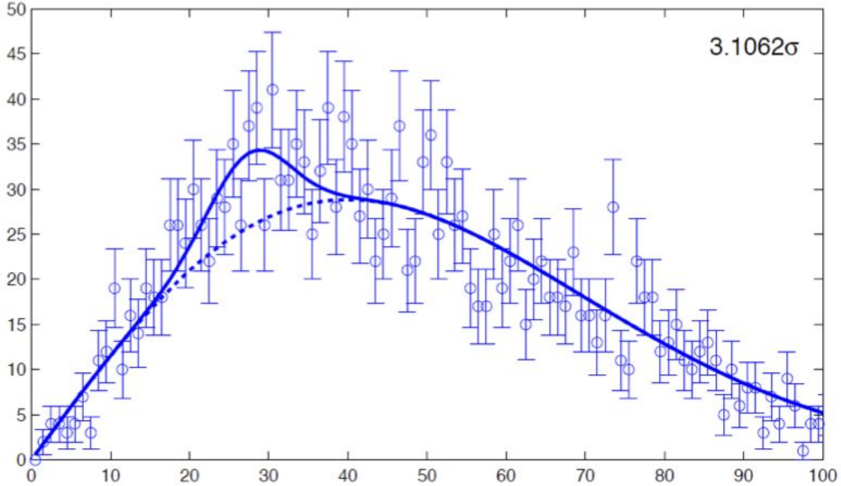
Answer: All background fluctuations





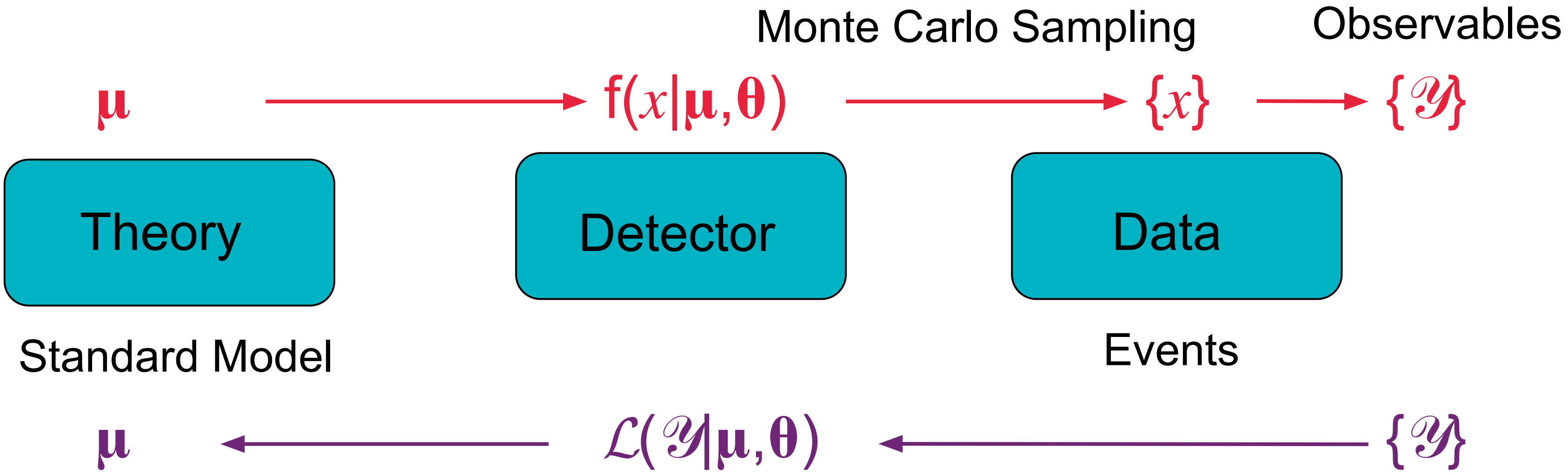
The look elsewhere effect





Physics analysis in a nutshell

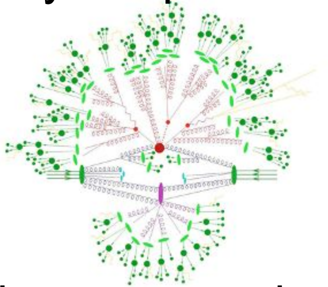
Simulation based inference



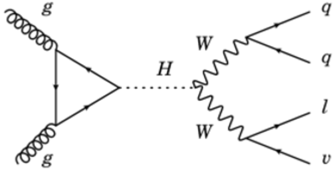
Sources of systematic uncertainties

Theory

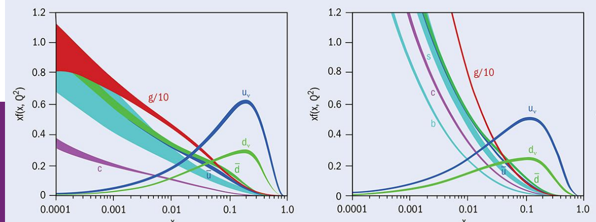
Soft physics process simulation



Physics process simulation



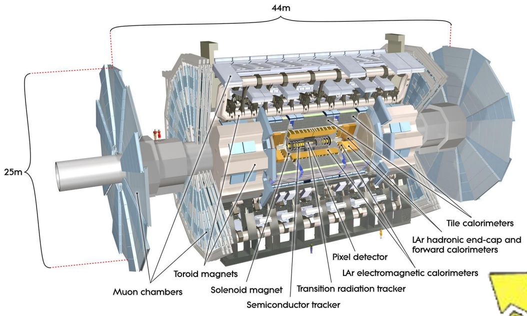
Proton structure functions



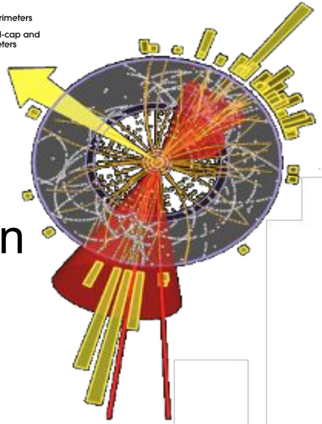
$\mathcal{L}(\mathcal{Y}|\mu,\theta)$

Experimental

Detector simulation



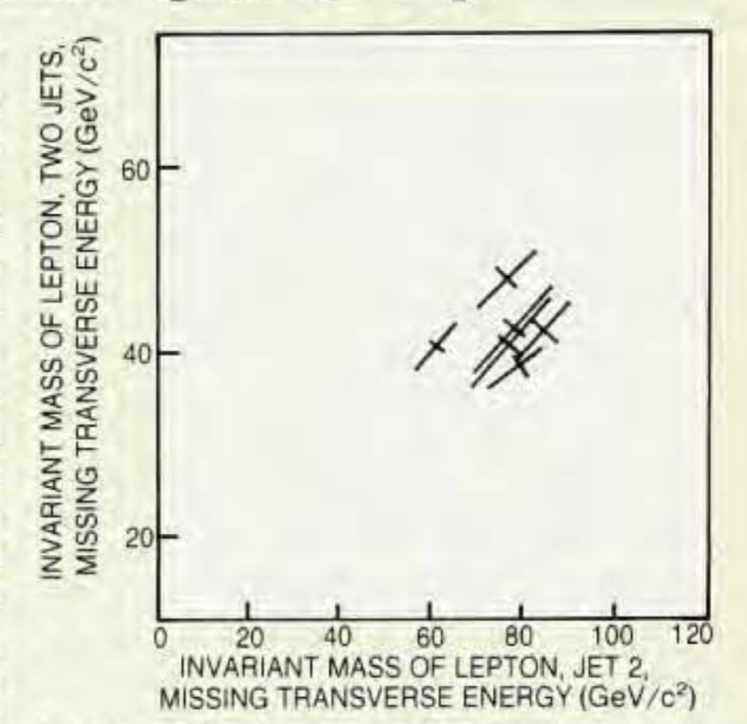
Detector reconstruction



The top quark 'discovery' at UA1

W → tb and t → bl±v
➔ 2 b jets, charged lepton
missing energy

Find 6 events.
➔ Plot total mass against bl±v mass
(v from missing energy/momentum)
➔ W mass in right place
➔ t mass around 40 GeV (correct mass approximately 173 GeV)
Turned out to be background - and very creative selection cuts



Samuel S. Wilks
(1906-1964)

Wilks theorem

- H0: Additional parameters (as predicted by H1) not needed
- If H0 correct then according to Wilks' theorem:
  $-\Delta\chi^2 = -2\ln[L(H1)/L(H0)]$ should follow for $n \rightarrow \infty$ χ^2 function with ndf = #added parameters

Wilks' theorem only applies for nested hypotheses:

H0: 1st order polynomial H1: 2nd order polynomial ✓

H0: 1st order polynomial H1: $a \cdot \exp(bx+cx2)$ ✕

Profiling over Nuisance Parameters ($\boldsymbol{\theta}$)

So if we are interested in the $\boldsymbol{\mu}$ that maximises $\mathcal{L}(\mathcal{Y}|\boldsymbol{\mu},\boldsymbol{\theta})$ we can use Wilk’s theorem and instead consider

$$\text{Max } \mathcal{L}(\boldsymbol{\mu},\boldsymbol{\theta}|\mathcal{Y}) = -2\ln\left(\mathcal{L}(\boldsymbol{\mu},\hat{\boldsymbol{\theta}}|\mathcal{Y}) / \mathcal{L}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\theta}}|\mathcal{Y}) \right) = \text{NLL}(\boldsymbol{\mu},\boldsymbol{\theta}|\mathcal{Y})$$

In the large N limit this is a chi-squared distribution with k degrees of freedom.

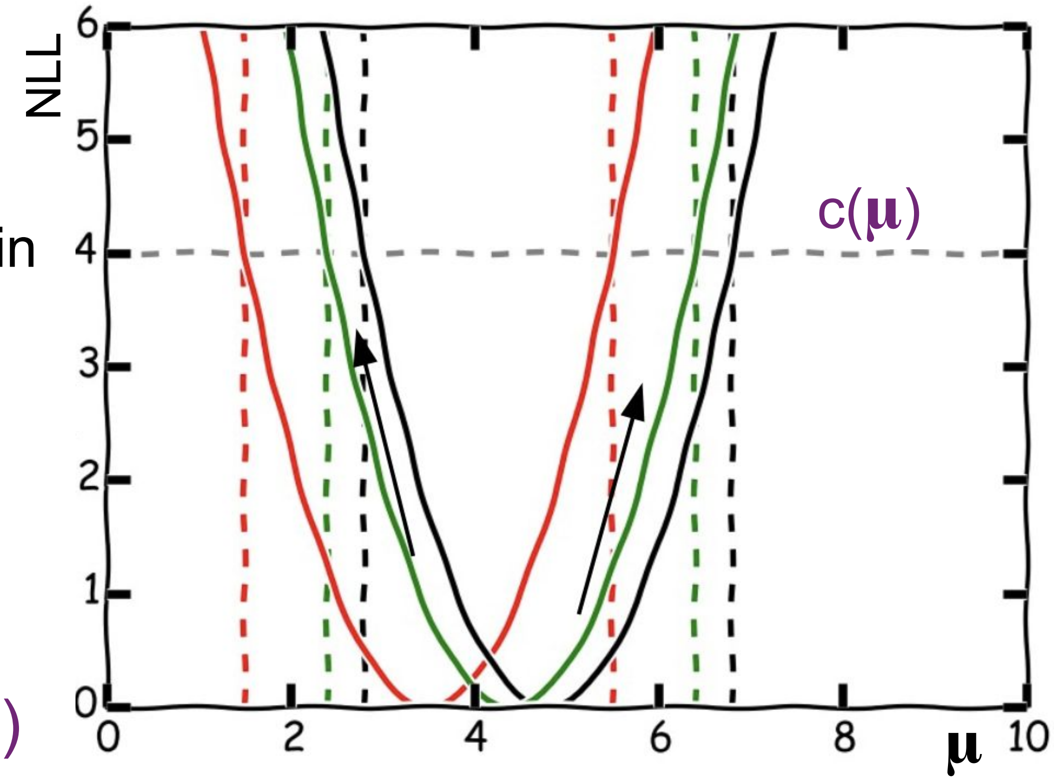
$$\text{Max } \mathcal{L}(\boldsymbol{\mu},\boldsymbol{\theta}|\mathcal{Y}) = \text{NLL}(\boldsymbol{\mu},\boldsymbol{\theta}|\mathcal{Y}) = \chi^2_k$$

Confidence intervals

Wilks' theorem (if it applies) makes it relatively simple to construct confidence intervals in a multi-dimensional model-parameter space.

Confidence interval is defined as

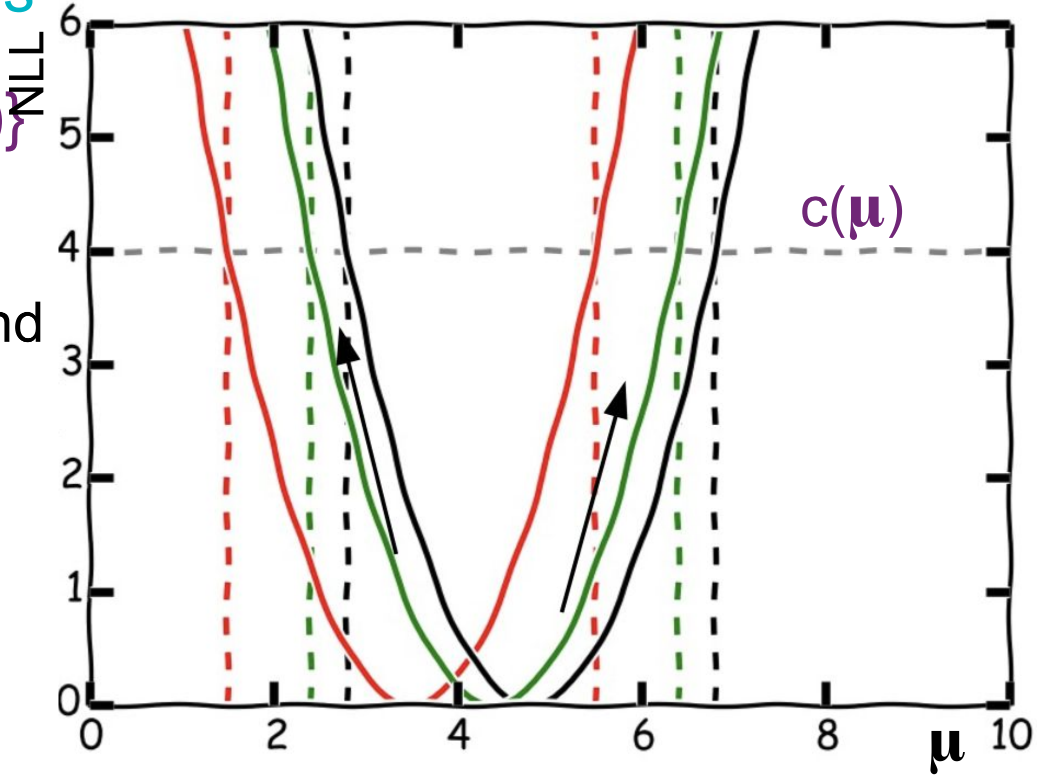
$I(\mathscr{Y}) = \{\mu \mid \mathrm{NLL}(\mu,\theta|\mathscr{Y}) < c(\mu)\}$



Multidimensional likelihoods

$I(\mathscr{Y}) = \{\mu \mid \mathrm{NLL}(\mu,\theta|\mathscr{Y}) < c(\mu)\}$

The threshold values c for different confidence level depend on the dimensionality of the confidence interval!

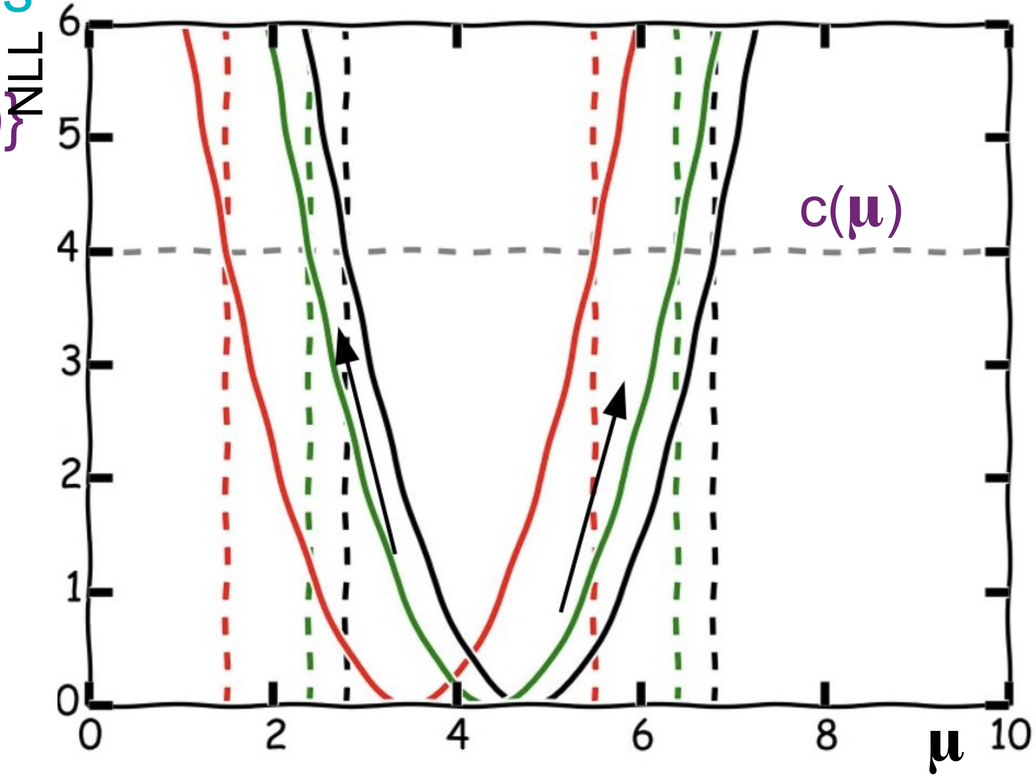


Multidimensional likelihoods

$I(\mathscr{Y}) = \{\mu \mid \mathrm{NLL}(\mu,\theta|\mathscr{Y}) < c(\mu)\}$

One dimension $\mathrm{NLL} = \chi^2_{k=1}$

68,3% CL	c=1
95,4%CL	c=4
99,7% CL	c=9



Multidimensional likelihoods

$I(\mathscr{Y}) = \{\mu \mid \mathrm{NLL}(\mu,\theta|\mathscr{Y}) < c(\mu)\}$

One dimension $\mathrm{NLL} = \chi^2_{k=1}$

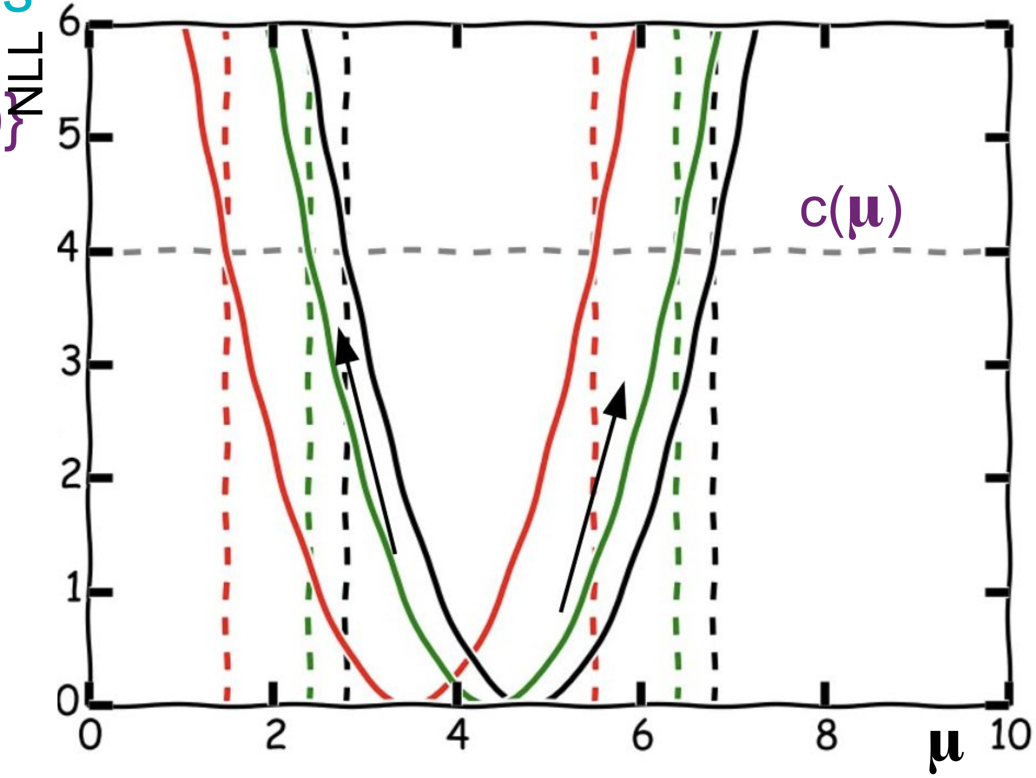
68,3% CL	c=1
95,4%CL	c=4
99,7% CL	c=9

$\mu=\{\mu_1\}$

Two dimensions $\mathrm{NLL} = \chi^2_{k=2}$

68,3% CL	c=2,3
95,4%CL	c=6,2
99,7% CL	c=11,8

$\mu=\{\mu_1,\mu_2\}$



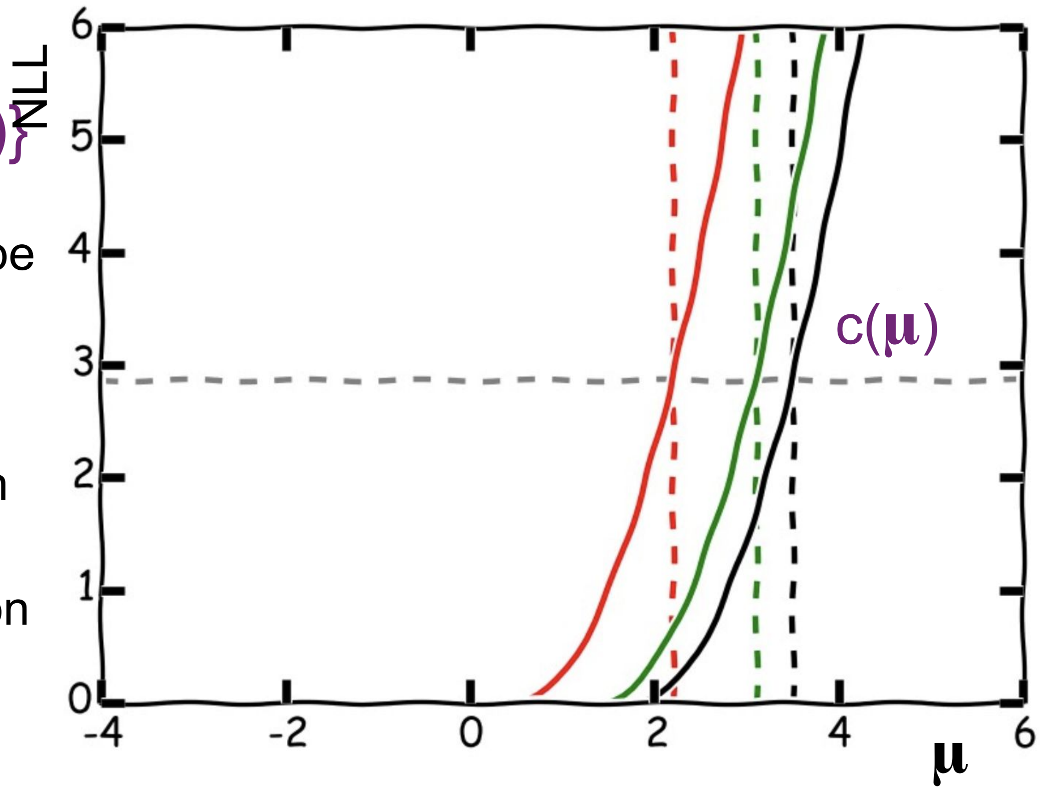
Setting limits

$I(\mathscr{Y}) = \{\mu \mid NLL(\mu,\theta|\mathscr{Y}) < c(\mu)\}$

Upper limits on a parameter (in contrast to two-sided intervals) can be obtained by setting the NLL to zero below the best-fit value.

The threshold c is a bit more tricky in this case, since the NLL follows a χ^2 distribution only half of the distribution In this case you can show that

95,4%CL	c=2,86

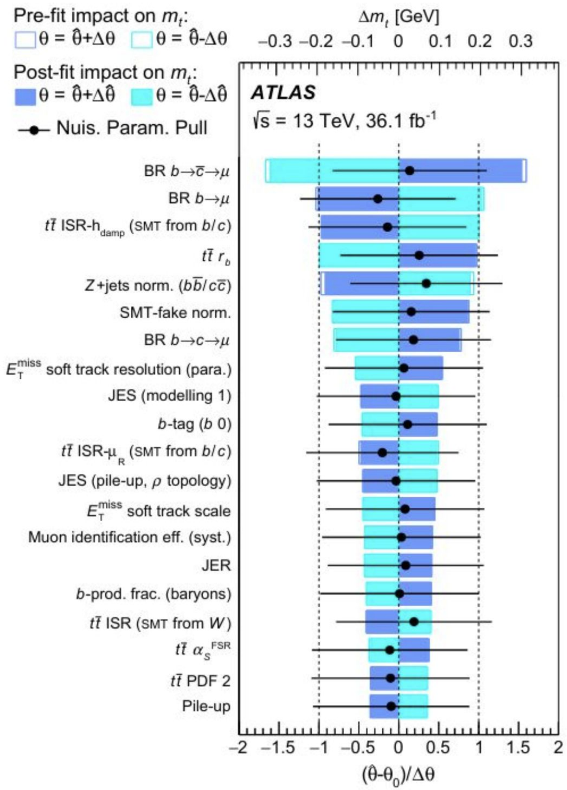


Pulls and rankings: A closer look at Nuisance Parameters

A way to look at the contribution to NLL from Nuisance Parameters (NPs) θ

Shows pre-fit and post-fit *impact* of individual NP on the determination of μ:

➔ Each NP fixed to ± 1 pre-fit and post-fit sigmas (Δθ and $\Delta\hat{\theta}$ = uncertainty on $\hat{\theta}$)
➔ Fit re-done with N-1 parameters
➔ Impact extracted as difference in central value of μ



Post-fit not smaller than pre-fit = no constraint

0 ±1= No pull

Pulls and rankings: A closer look at Nuisance Parameters

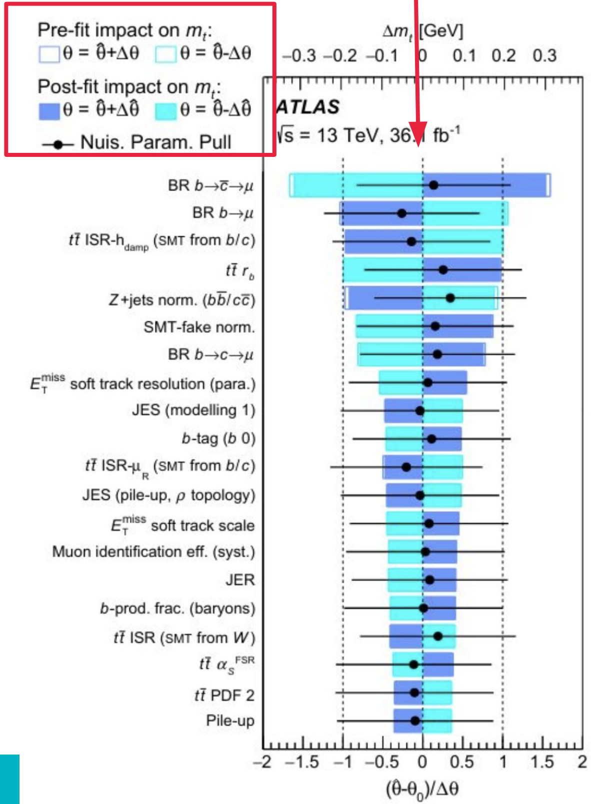
A way to look at the contribution to NLL from Nuisance Parameters (NPs)

Shows pre-fit and post-fit *impact* of individual NP on the determination of μ:

➔ Each NP fixed to ± 1 pre-fit and post-fit sigmas ($\Delta\theta$ and $\Delta\hat{\theta}$ = uncertainty on $\hat{\theta}$)

➔ Fit re-done with N-1 parameters

➔ Impact extracted as difference in central value of μ

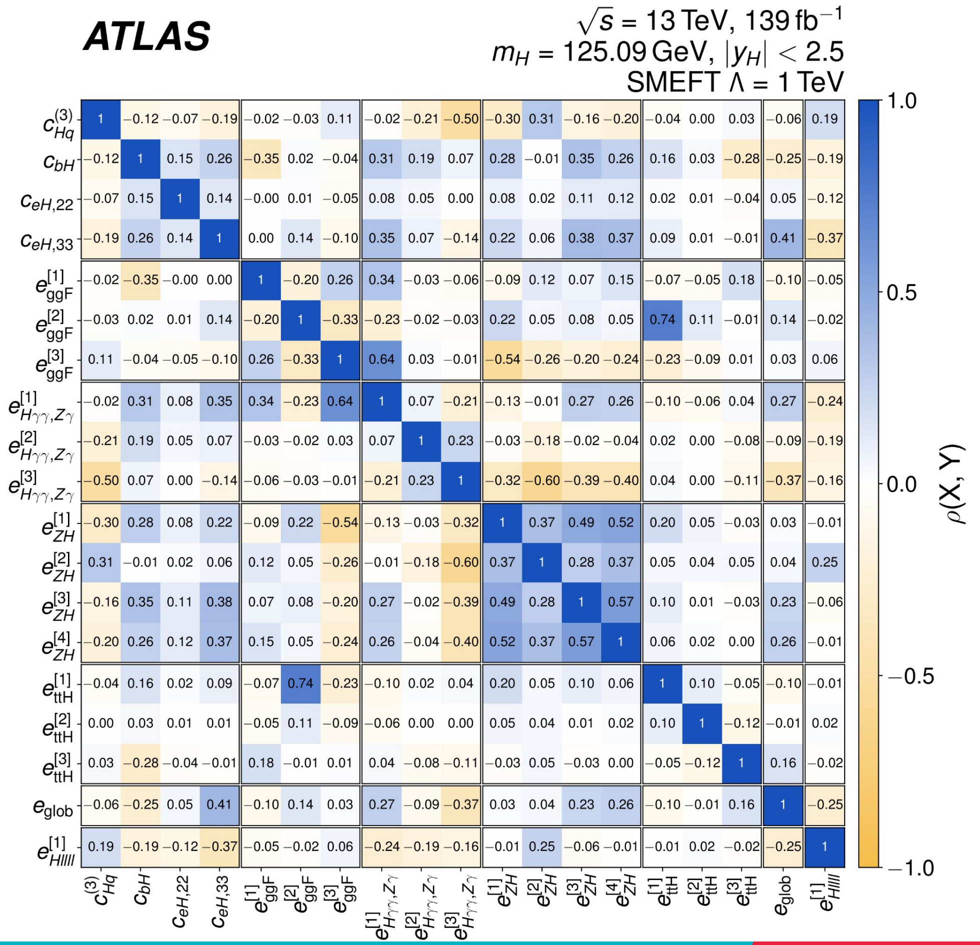


Nomenclature: fitting and parameter estimation

The correlation matrix

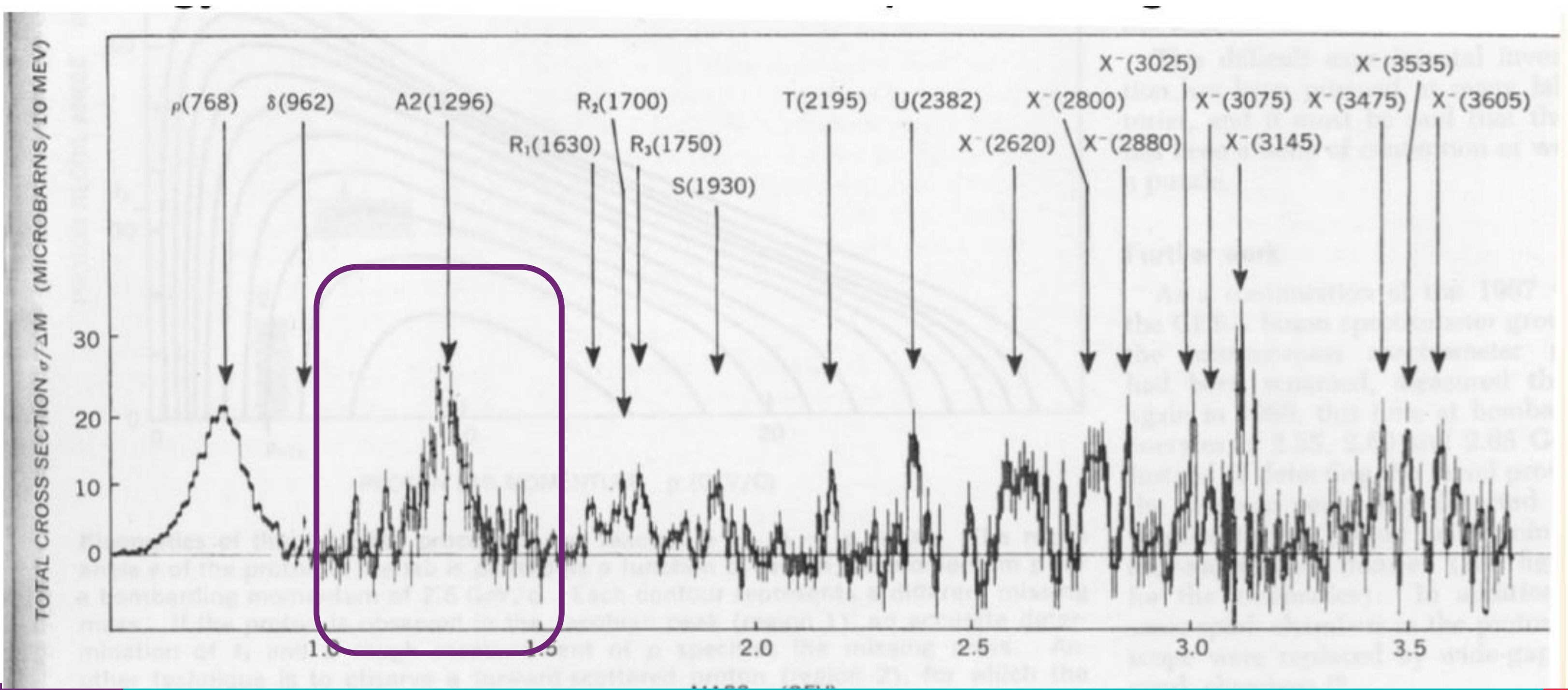
Large correlations can make the fit unstable

- ➔ Not necessarily a problem but worth investigating
- ➔ Change of basis can make constraints explicit
- ➔ Consider carefully when *reparametrizing*



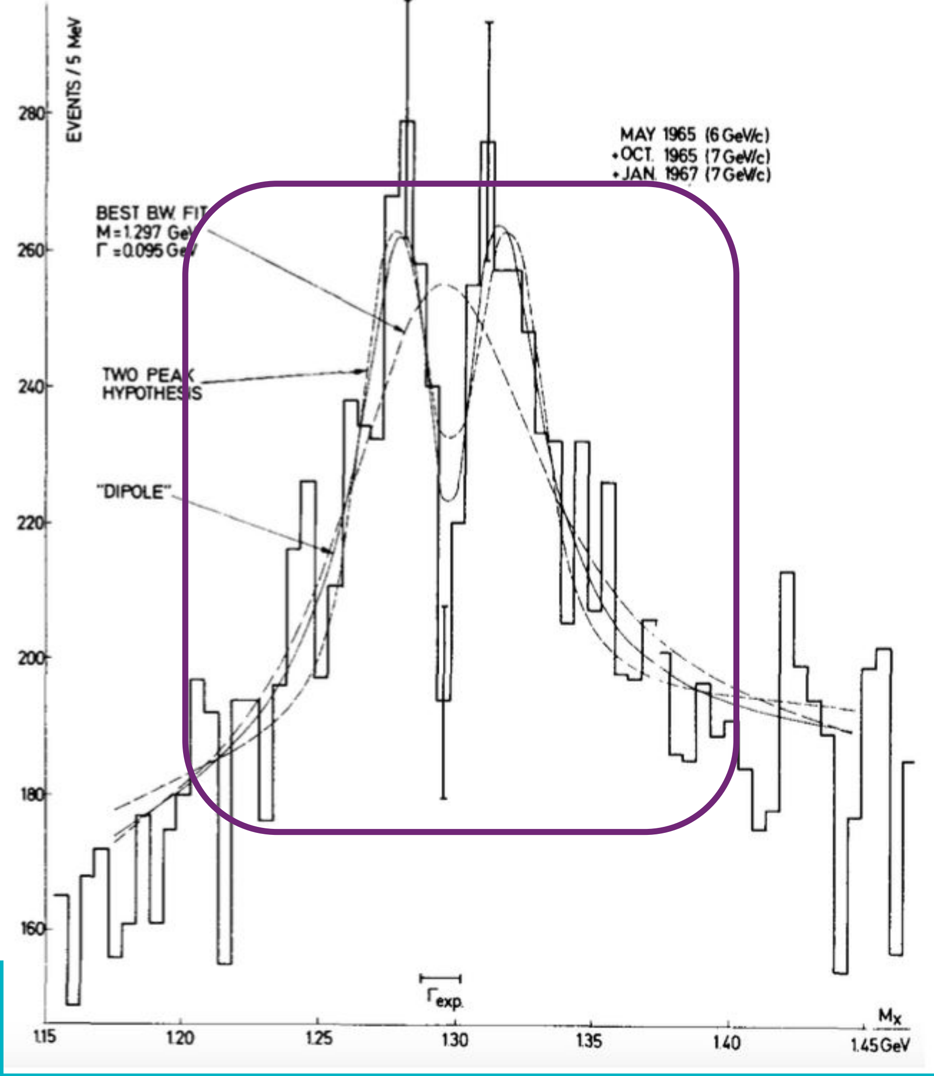
Nikhef

Splitting the A2 meson



Splitting the A2 meson

First reported 1967
More than 6σ effect!

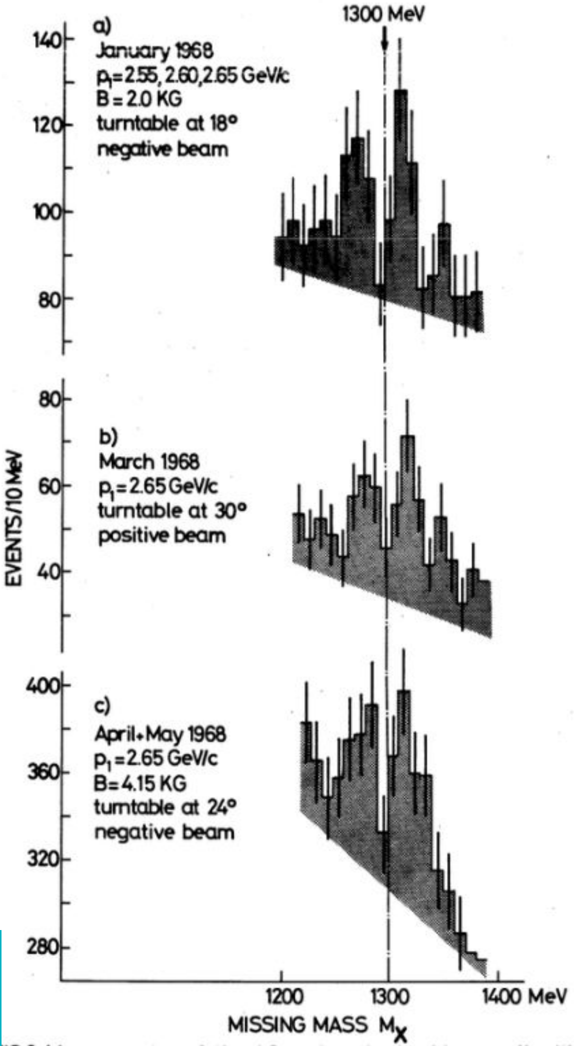


Splitting the A2 meson

First reported 1967
More than 6σ effect!

➔ Confirmed in 1969
➔ Went away 1971

1. “Creative” analysis cuts
2. Wanting to see it
   a. And not publishing when you don’t see it

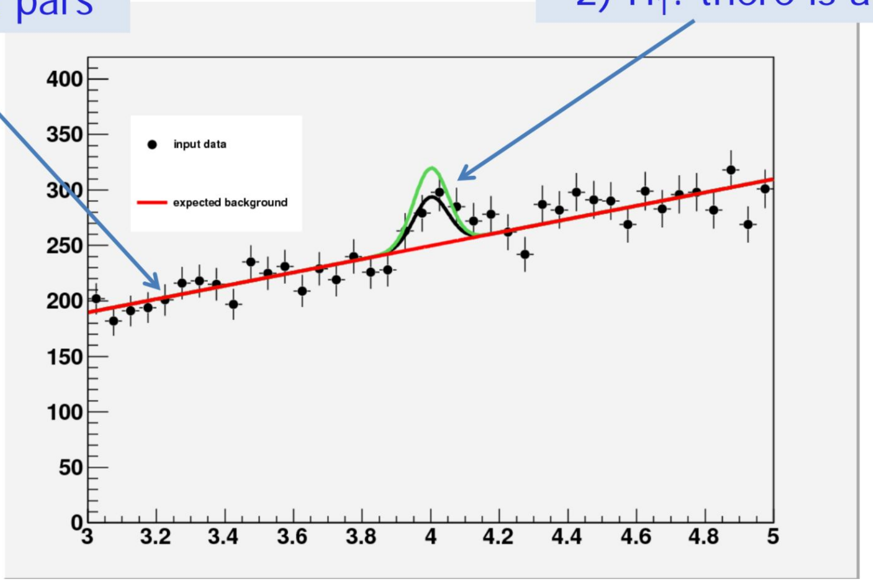


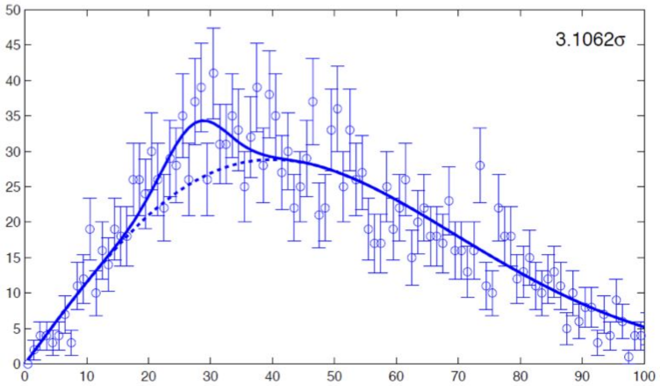
How many parameters do I need?

$$q_{fix,obs} = -2\ln\frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$$

1) H_1: add more background pars

2) H_1: there is a signal

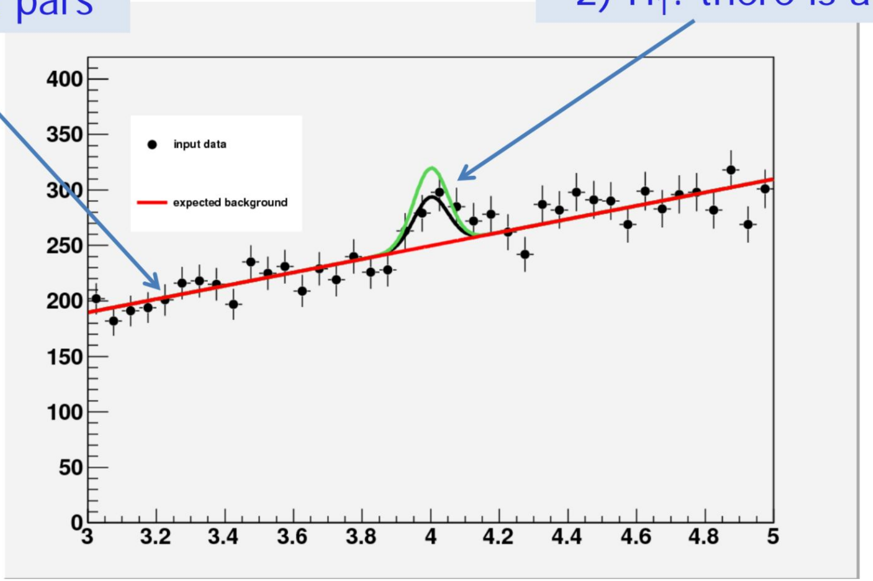


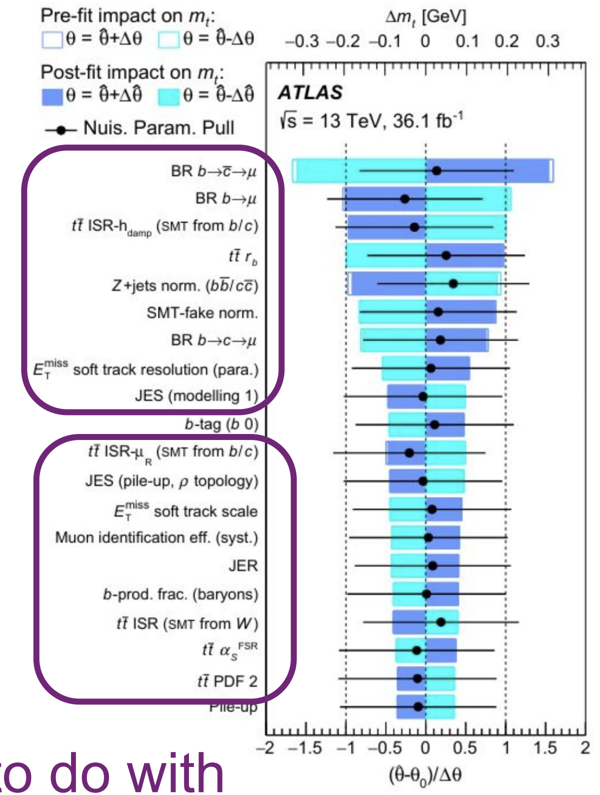


How many parameters do I need?

1) H_1: add more background pars

2) H_1: there is a signal





All to do with background estimation

The ζ(8.3)

"Discovered" in 1984 by the Crystal Ball experiment at DESY.

➔ Measure energy of photons

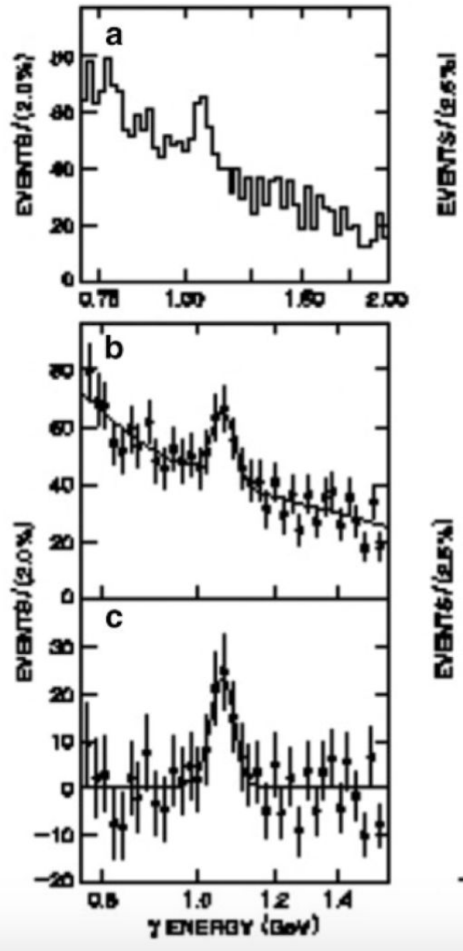
➔ Single energy peak seen!!

➔ Signals $e^+e^- \rightarrow \Upsilon \rightarrow \zeta\gamma$

➔ 4.2 sigma effect

➔ Plots show (a) raw data , (b) fit, and (c) background-subtracted fit

When more data was taken (in 1985) the peak went away.



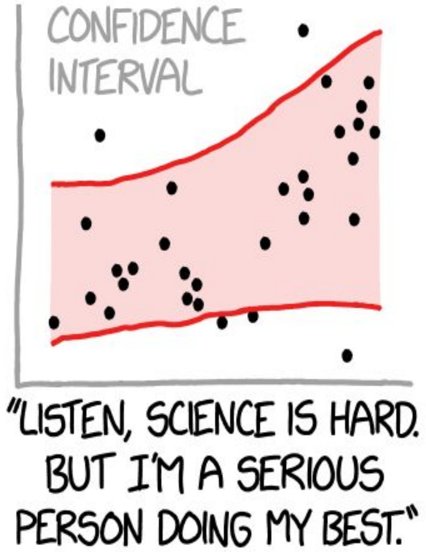
When things go wrong…

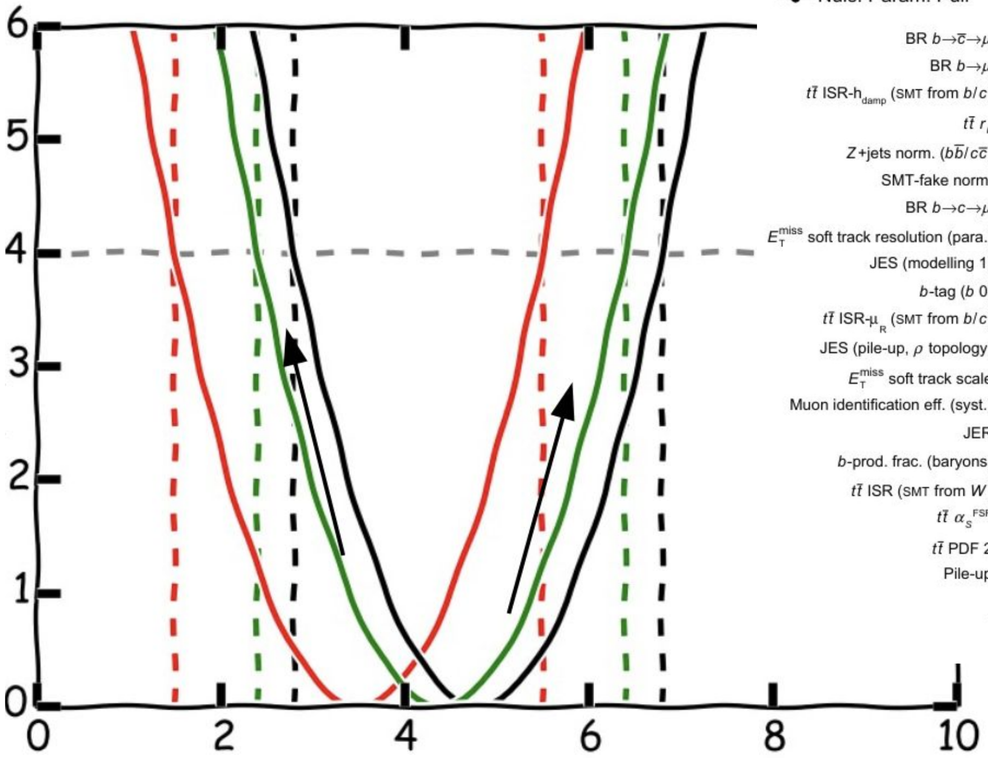
“It was easy - I just got a block of marble and chipped away anything that didn’t look like David.” - Michaelangelo Buonarotti(attrib.)

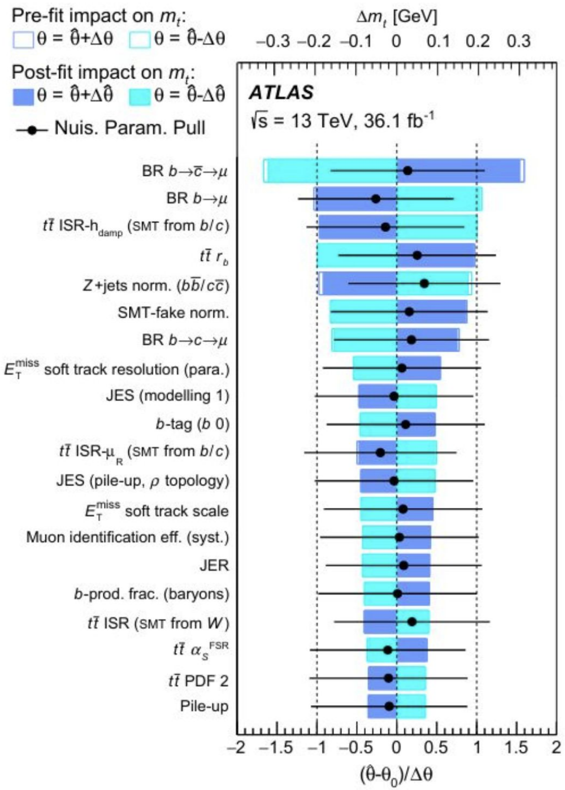
- ➔ Maybe good way of creating sculpture - but very bad way of doing physics
- ➔ To resist temptation: For searches, devise cuts before looking at the data. Use Monte Carlo simulations, and/or data in ‘sidebands’. Only when cuts are optimised do you ‘open the box’.
- ➔ For measurements, apply secret offset until selection frozen.

Partial summary

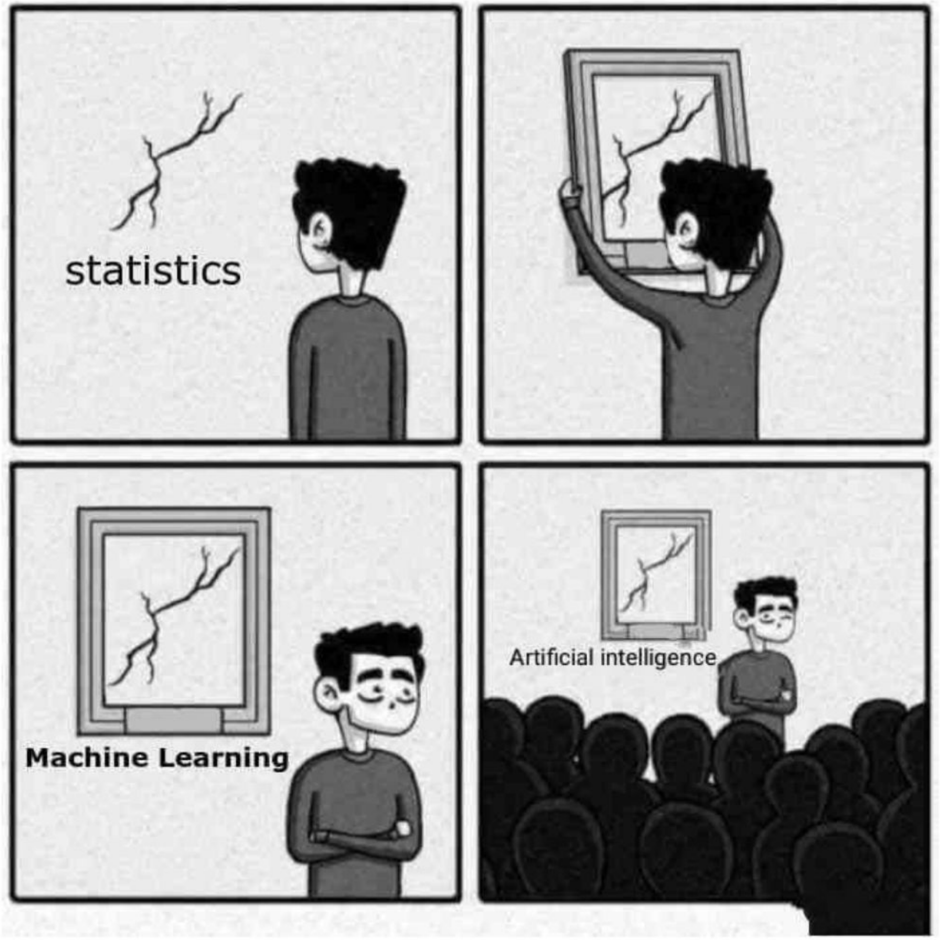
Not too bad…







Stats Meets ML?



What changes for ML?

Ask yourself: Why is my ML doing better?

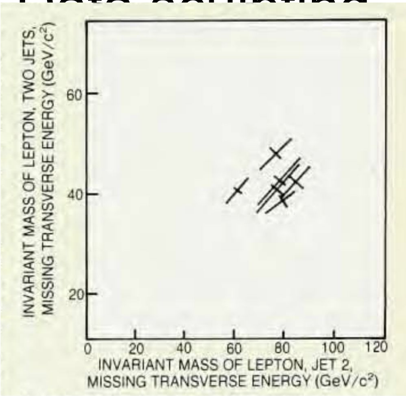
- Multidimensional
  - Does it have to do with correlated parameters?
  - Event-by-event
- Faster
  - Does it parallelise?
  - Simplify
- Wrong assumptions
  - Is there something we assume to be true that it doesn’t learn?

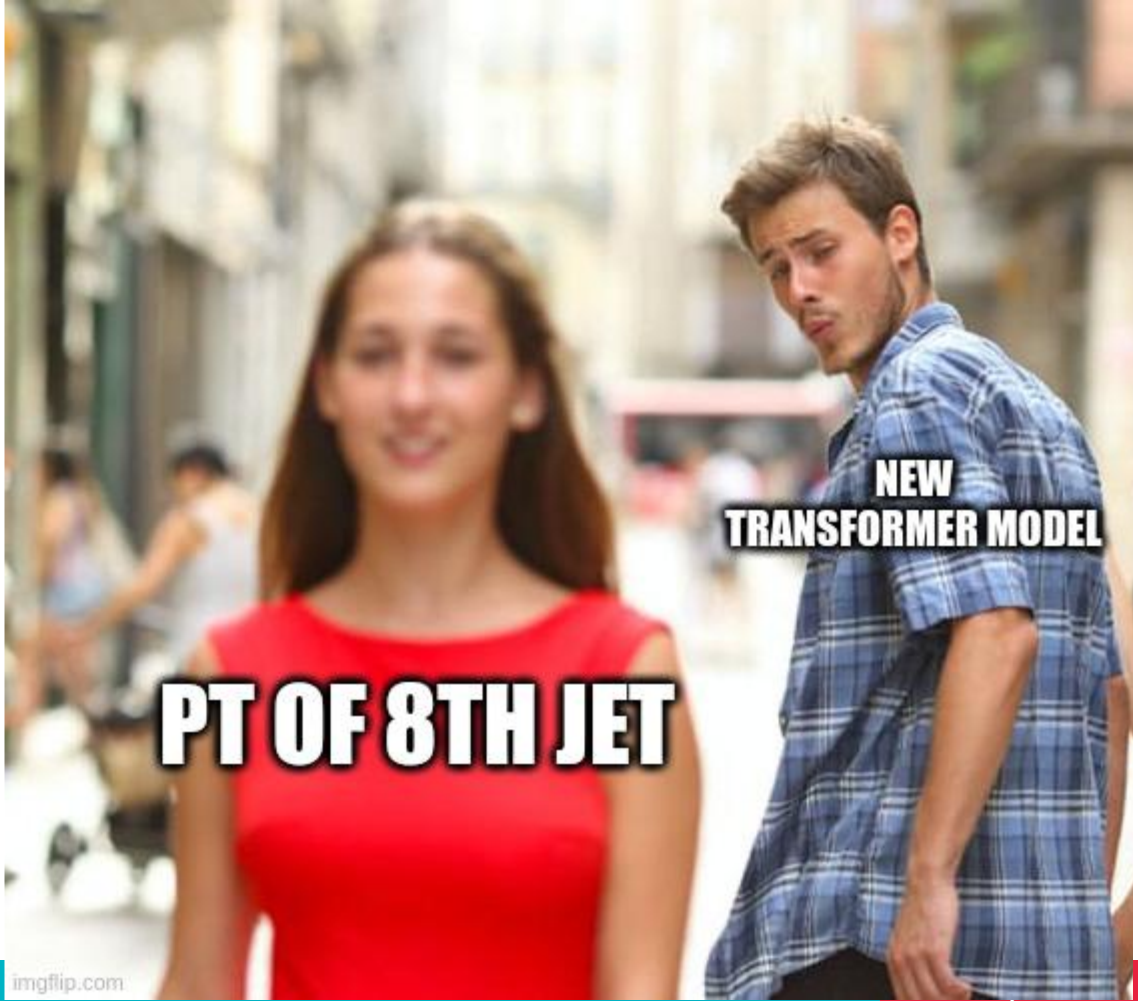
General types of ML used in physics

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Dimensionality reduction and compression
  - Outlier detection
- Weak and Semi-supervised Learning
  - Weak: Label smoothing or correction, Loss reweighting or Multiple-instance learning (MIL)
  - Semi: e.g. Pseudo-labeling with confidence thresholds

What could go wrong?

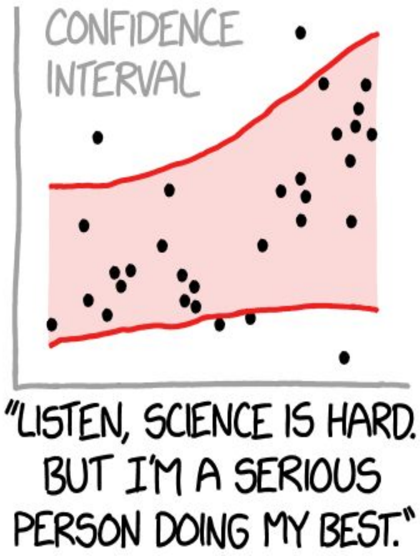
- Learning MC issues
- Finding the right answer, but no uncertainties
- Data sculpting





General types of ML used in physics: special cases

- Physics-informed ML
  - Grounding data-driven models in domain knowledge
  - Goal: accurate, robust, and interpretable outcomes
- Uncertainty aware ML
  - Helps look at intervals rather than best fit values



Interpretability: A warm fuzzy feeling…. Or more than that

When things go wrong… ML edition

“It was easy - I just got a block of marble and chipped away anything that didn’t look like David.” - Michaelangelo Buonarotti(attrib.)

- ➔ Often the way ML works! You just want to make sure you find a David not Goliath
- ➔ To resist temptation: For searches, devise cuts before looking at the data. Use Monte Carlo simulations, and/or data in ‘sidebands’. Only when cuts are optimised do you ‘open the box’.
  - Much harder to do for ML!!
- ➔ For measurements, apply secret offset until selection frozen.

Overfitting and techniques to prevent it

- Regularization: Adds a penalty term to the loss function to discourage large parameter values.
- Dropout: During training, randomly sets a fraction of neurons to zero at each iteration, which prevents co-adaptation of neurons and acts as an ensemble method.
- Early Stopping: Monitors performance on a validation set and halts training when the validation loss stops improving, reducing the risk of overfitting to training data.
- Data Augmentation: Increases the effective size of the training set by applying physically valid transformations

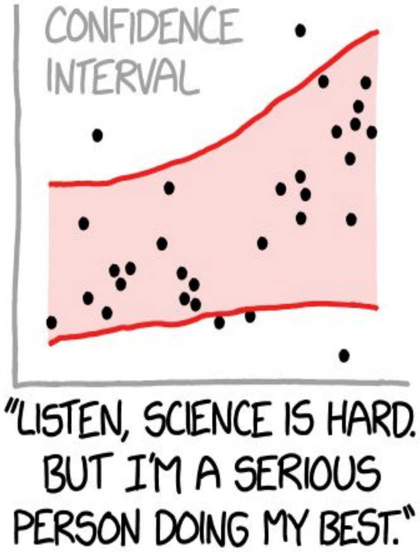
The difficult part

➔ To resist temptation: For searches, devise cuts before looking at the data. Use Monte Carlo simulations, and/or data in ‘sidebands’. Only when cuts are optimised do you ‘open the box’.

Possible checks:

- Injection test
- Smeared or adjusted input data
- Testing on MC from a different generator
- Stay blind

Summary

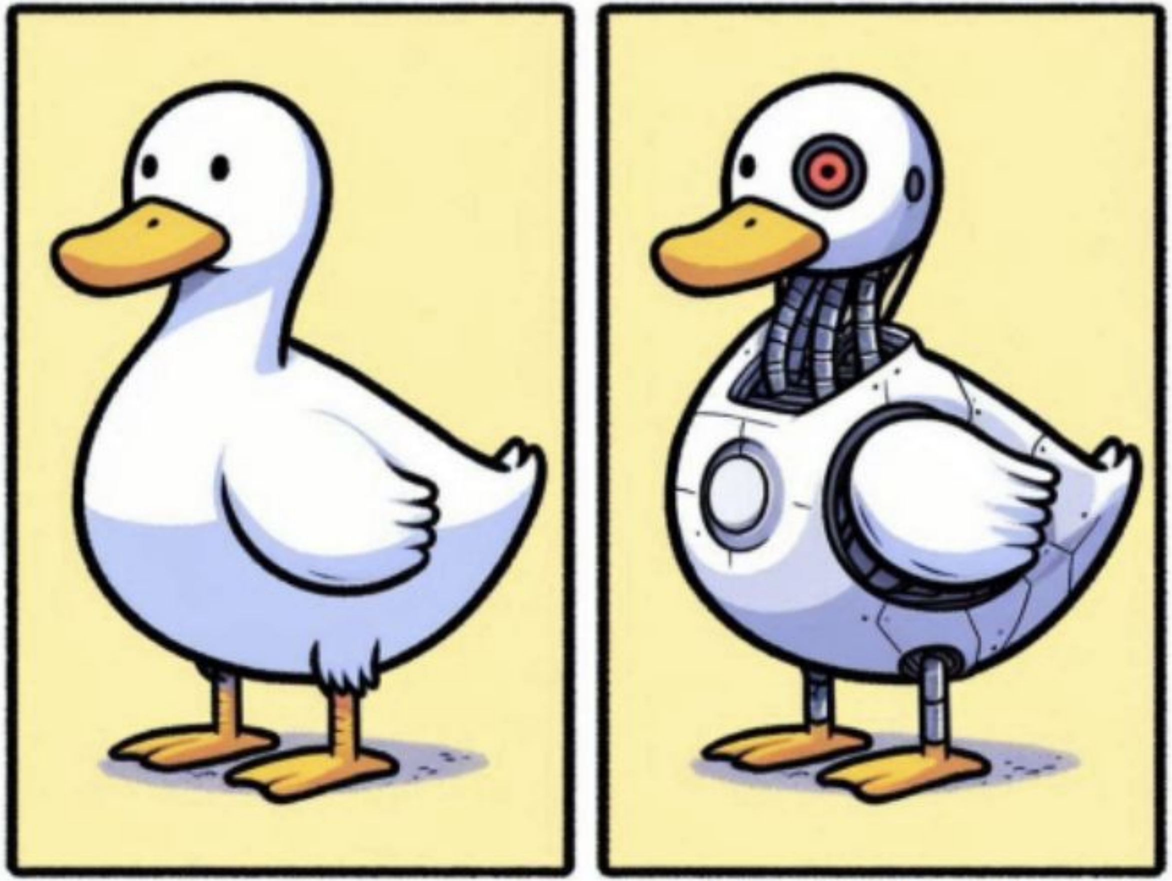


- ➔ Using a high level of significance (5σ)
  - ◆ History tells us this is necessary
  - ◆ High output of results
- ➔ Us "blind" analysis
- ➔ Focus in intervals
  - ◆ Rather than best estimated (fitted) values
- ➔ Independent validations
  - ◆ Rumours are hard to get rid off…
- ➔ Use common sense and healthy scepticism
- ➔ To learn more: https://indico.cern.ch/event/1407421

If it looks like a duck, swims like a duck, and quacks like a duck, then it probably is a duck.

[If it looks like data, it's a sufficiently good simulator?]

Thanks!

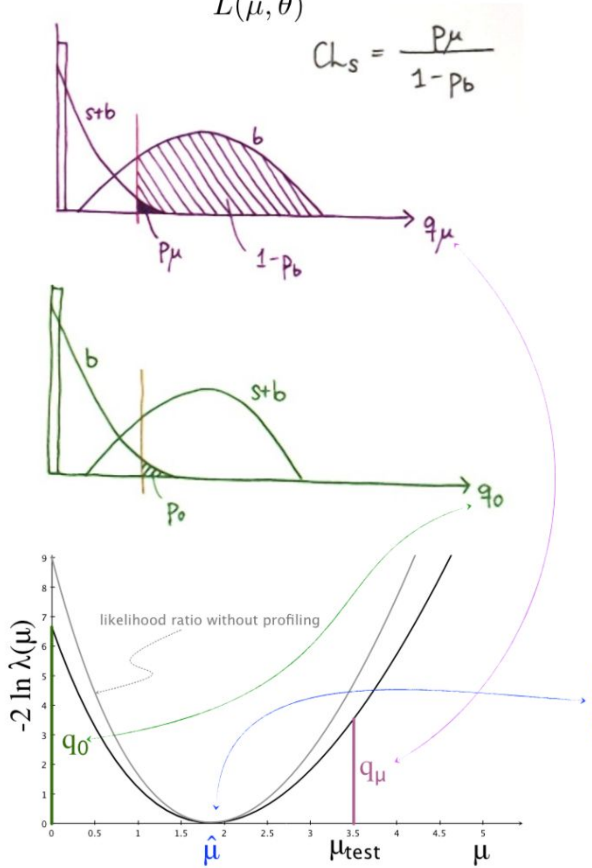


[DALL·E 3 take on the topic]

Back up

Thumbnail of the LHC statistical procedures

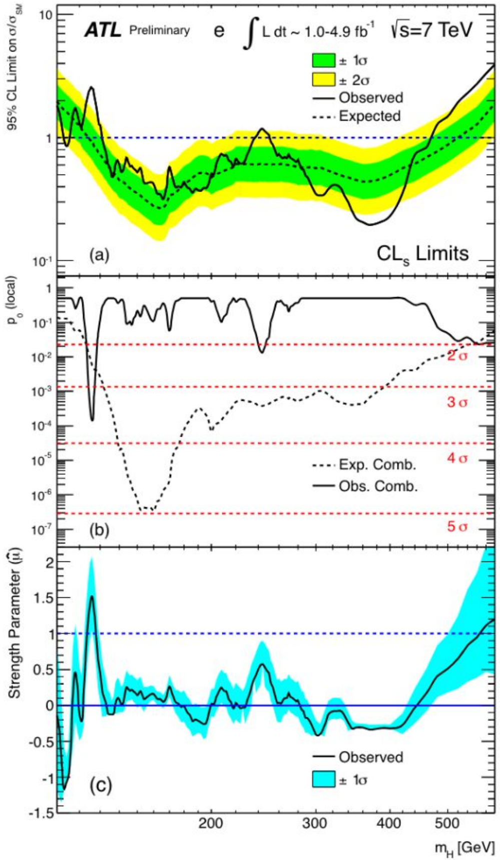
$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$



CL$_s$ to test signal hypothesis

p_0 to test background hypothesis

$\hat{\mu}$ to estimate signal strength



Parameter dependence

Say we want to model either $p(x\,|\,\theta)$ or $r(x\,|\,\theta) = \dfrac{p(x\,|\,\theta)}{p_{\text{ref}}(x)}$.

A few approaches that change structure of the model

- **Point-by-Point**: model $p(x\,|\,\theta_i)$ for a set of points $\{\theta_i\}$
  - Not explicitly parametrized in θ, no structure
- **Parametrized Network:** NN models both x-dependence and θ-dependence
  - Most flexible, but doesn't exploit any physics knowledge
- **Fixed Interpolation:** multiple NNs model x-dependence, but form of θ-dependence is fixed & defined by physicist
  - e.g. this is possible for EFT coefficients (exact)
  - This is what HistFactory etc. do for nuisance parameters but this makes assumptions

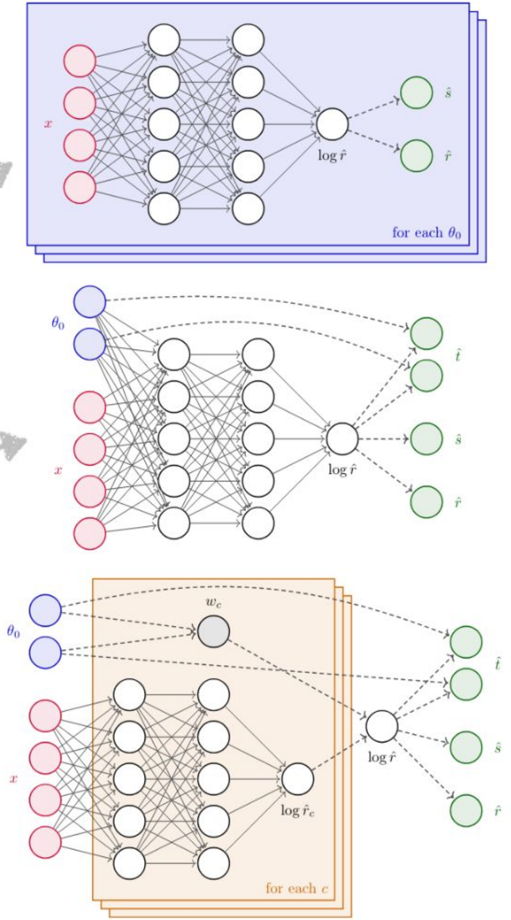
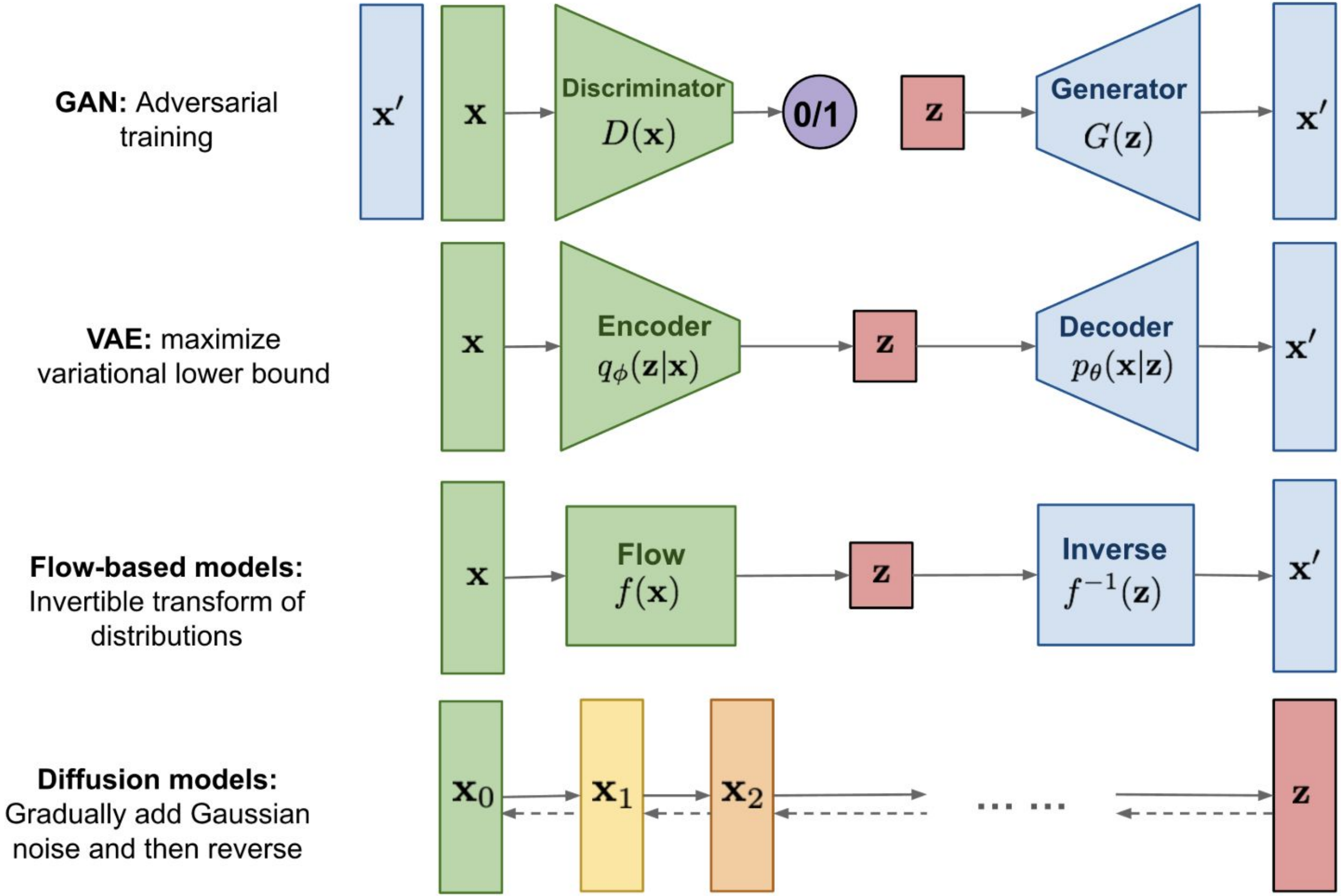


Figure 8: Schematic neural network architectures for point-by-point (top), agnostic parameterized (middle), and morphing-aware parameterized (bottom) estimators. Solid lines denote dependencies with learnable weights, dashed lines show fixed functional dependencies.

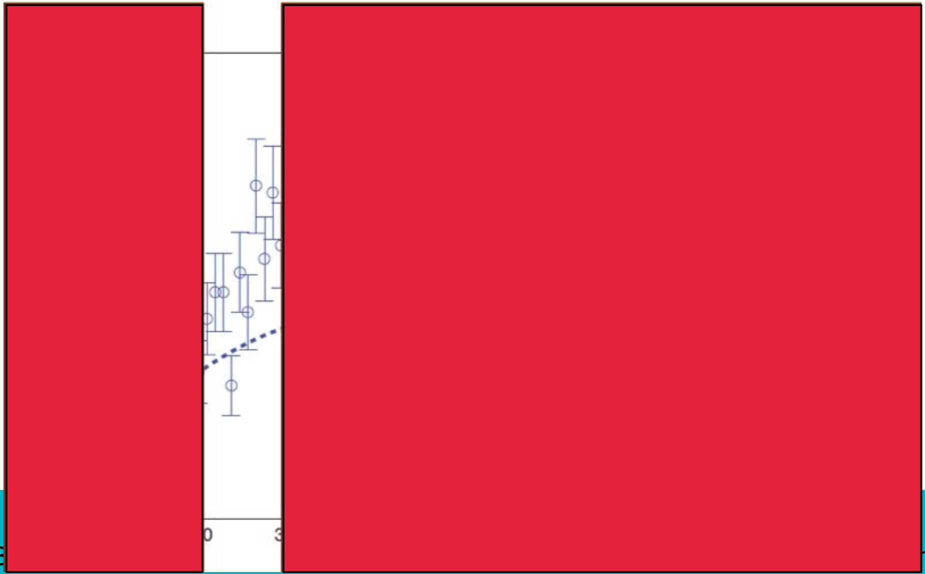
Fig from: https://arxiv.org/abs/1805.00020



https://lilianweng.github.io/posts/2021-07-11-diffusion-models/

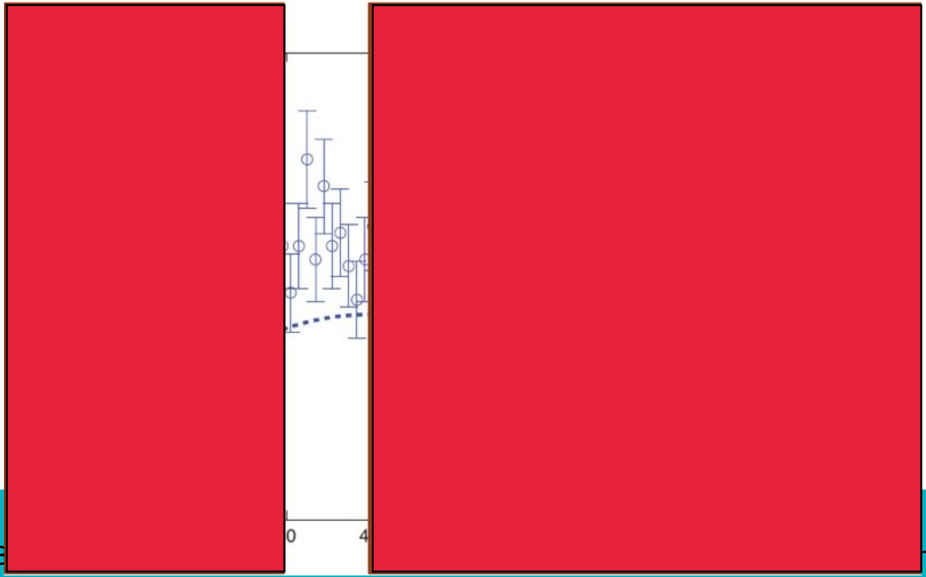
Look elsewhere effect: Option 1 Fixed Mass

Scan mass range in predefined steps.
Perform a fixed mass analysis at each point $q_{fix,obs} = -2\ln\frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$



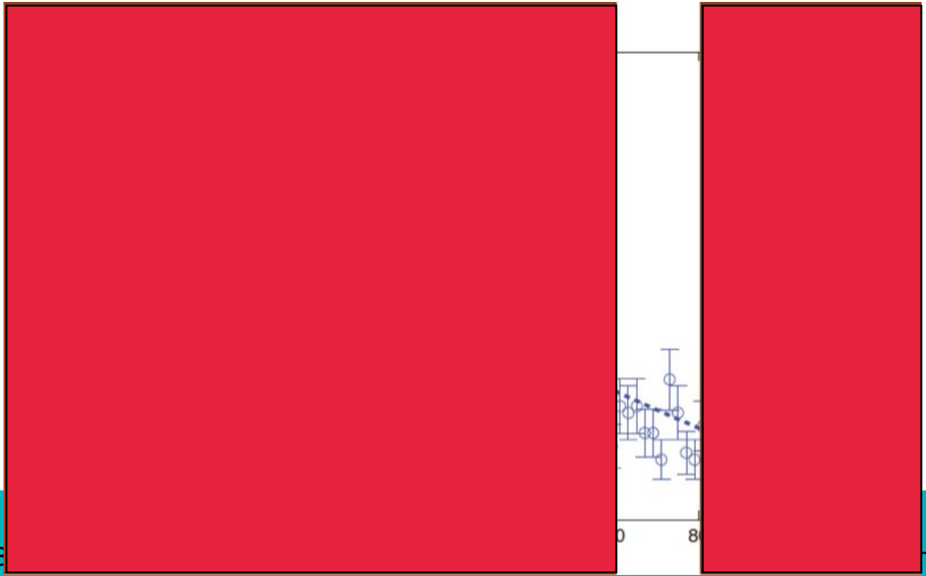
Look elsewhere effect: Option 1 Fixed Mass

Scan mass range in predefined steps.
Perform a fixed mass analysis at each point $q_{fix,obs} = -2\ln \frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$

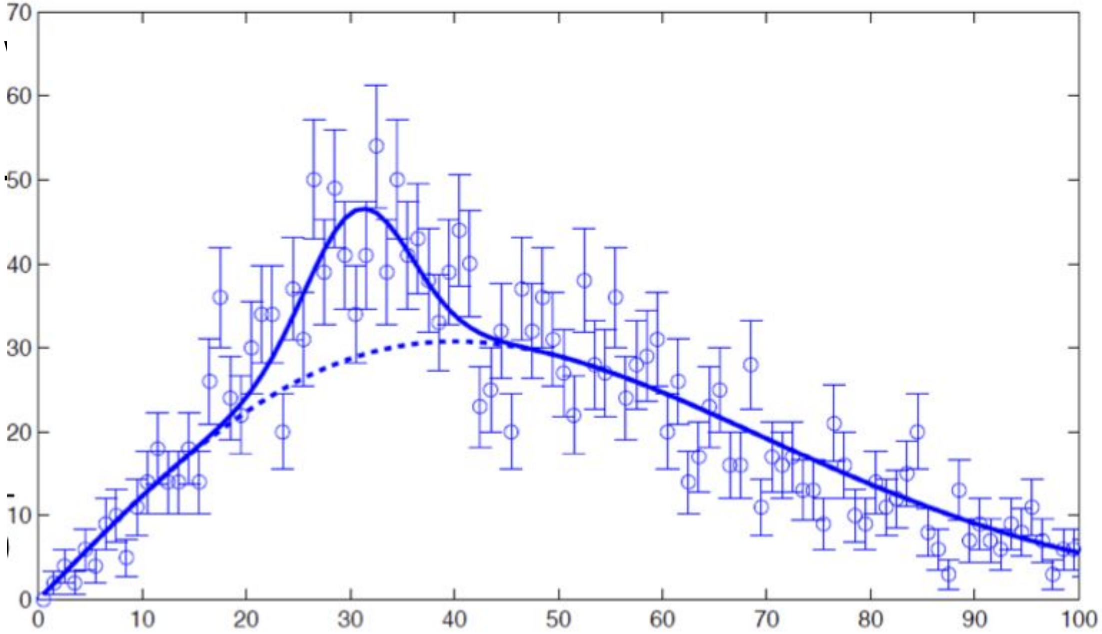


Look elsewhere effect: Option 1 Fixed Mass

Scan mass range in predefined steps.
Perform a fixed mass analysis at each point $q_{fix,obs} = -2\ln\frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$



Is there a signal?



$$q_{fix,obs} = -2\ln\frac{L(b)}{L(\hat{\mu}s(m=30)+b)}$$

Is there a signal?

For GOF tests with binned data:

➔ compare observed event numbers n_i with expectation values f_i

Since no H1 specified there are many different GOF tests possible

$$\chi^2 = \Sigma_i (f_i - n_i)^2/\sigma^2$$

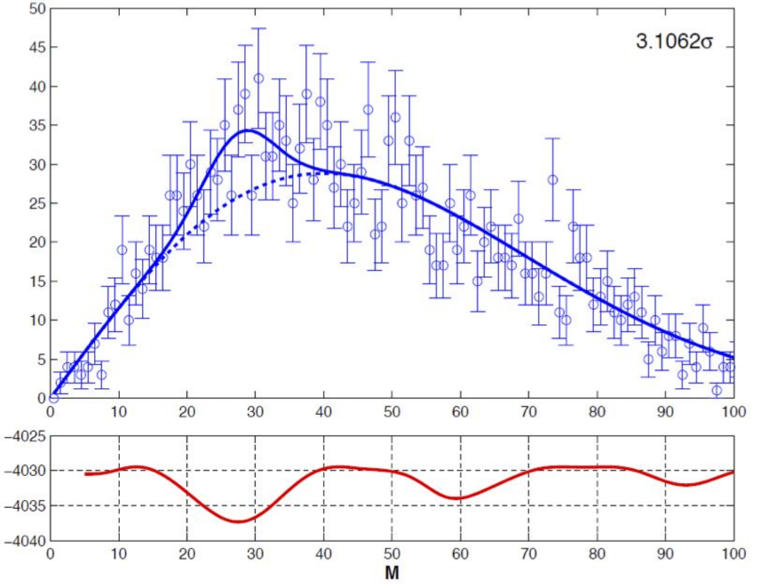
➔ Basically does what you do by eye; Minimise distance from hypothesis to the data points

Look elsewhere effect: Option 2 Floating mass

Leave the mass floating

Use a modified test statistic

$$q_{float,obs}(\hat{\mu},\hat{m}) = -2\ln\frac{L(b)}{L(\hat{\mu}s(\hat{m})+b)}$$



Look elsewhere effect: Option 3 Trials factor

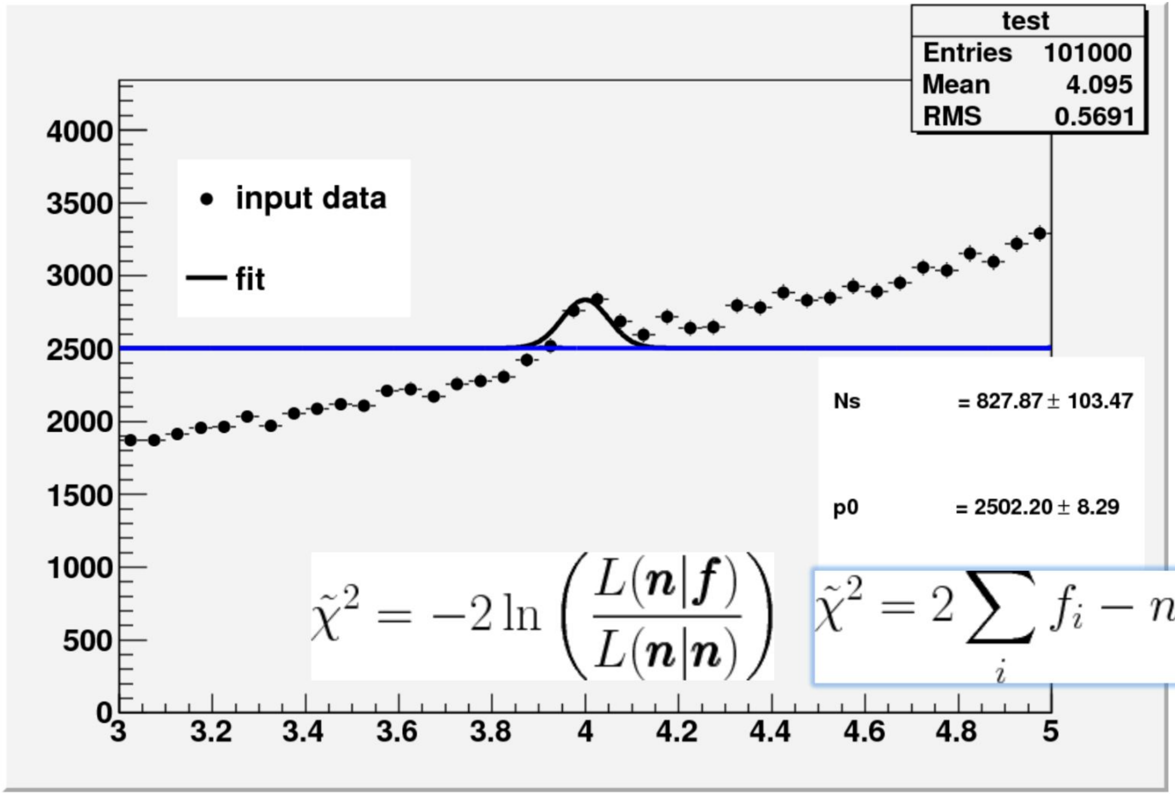
Define your significance of a local excess of events by taking into account the probability of observing an excess anywhere in the range. Quantify in terms of a trial factor

- Ratio between the probability of observing the excess at some fixed mass point to the probability of observing it anywhere in the range.
- The trial factor grows linearly with the (fixed mass) significance

$$trial\# = \frac{\int_{q_{obs}} f(q_{float} \mid H_0)\, dt_{float}}{\int_{q_{obs}} f(q_{fix} \mid H_0)\, dt_{fix}} = \frac{p_{float}}{p_{fix}} > 1$$

Recommended

How many parameters do I need? – Example



Fit function:
gauss+p0

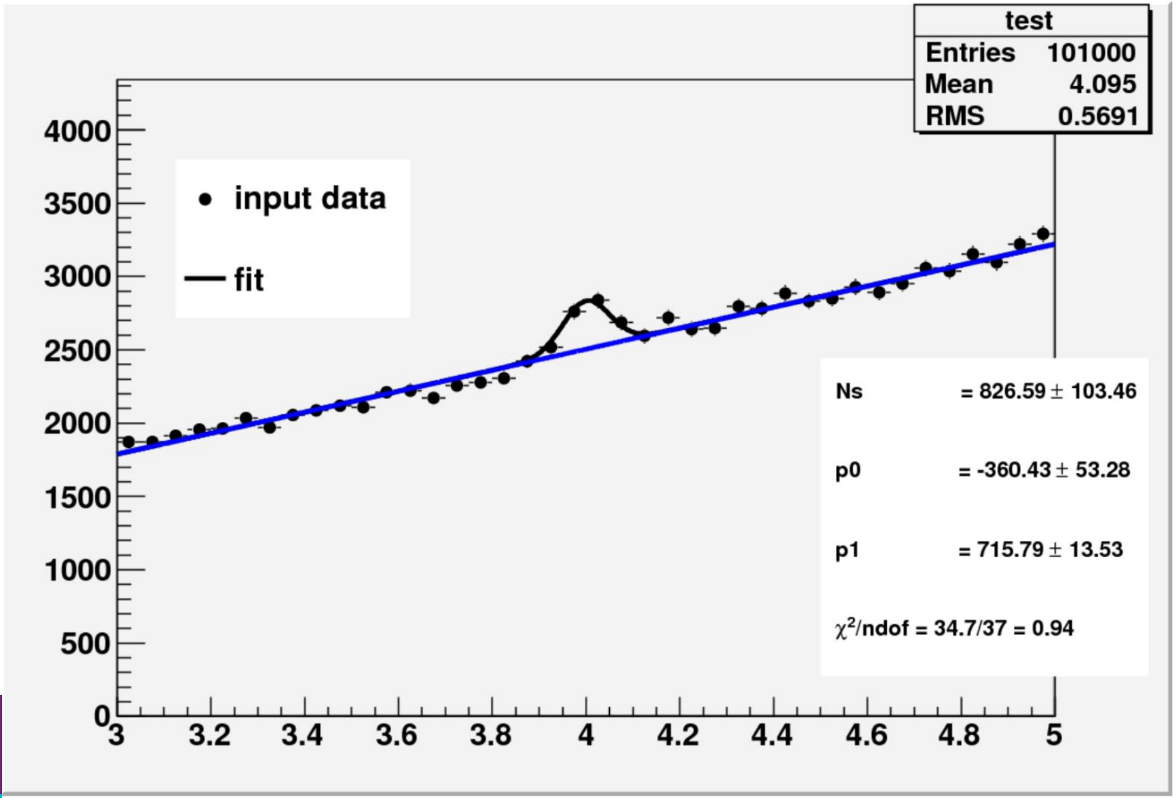
$$\chi^2 = \Sigma_i (f_i - n_i)^2/\sigma^2$$

$$\tilde{\chi}^2 = -2\ln\left(\frac{L(\boldsymbol{n}|\boldsymbol{f})}{L(\boldsymbol{n}|\boldsymbol{n})}\right)$$

$$\tilde{\chi}^2 = 2\sum_i f_i - n_i + n_i \ln(n_i/f_i)$$

$$\tilde{\chi}^2 = 2880$$

How many parameters do I need? – Example



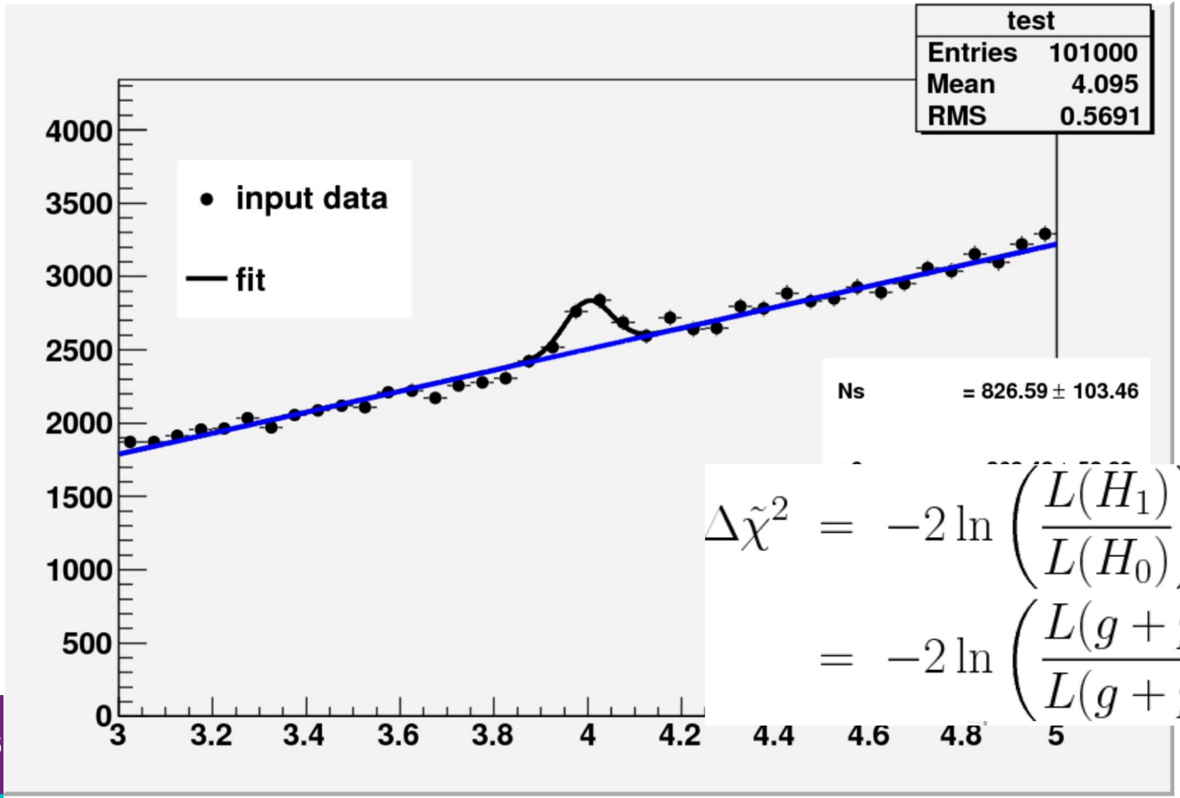
Fit function:
gauss+p1

$$\tilde{\chi}^2 = 34.7$$

Should we stop here?

Nikhef

How many parameters do I need? – Example



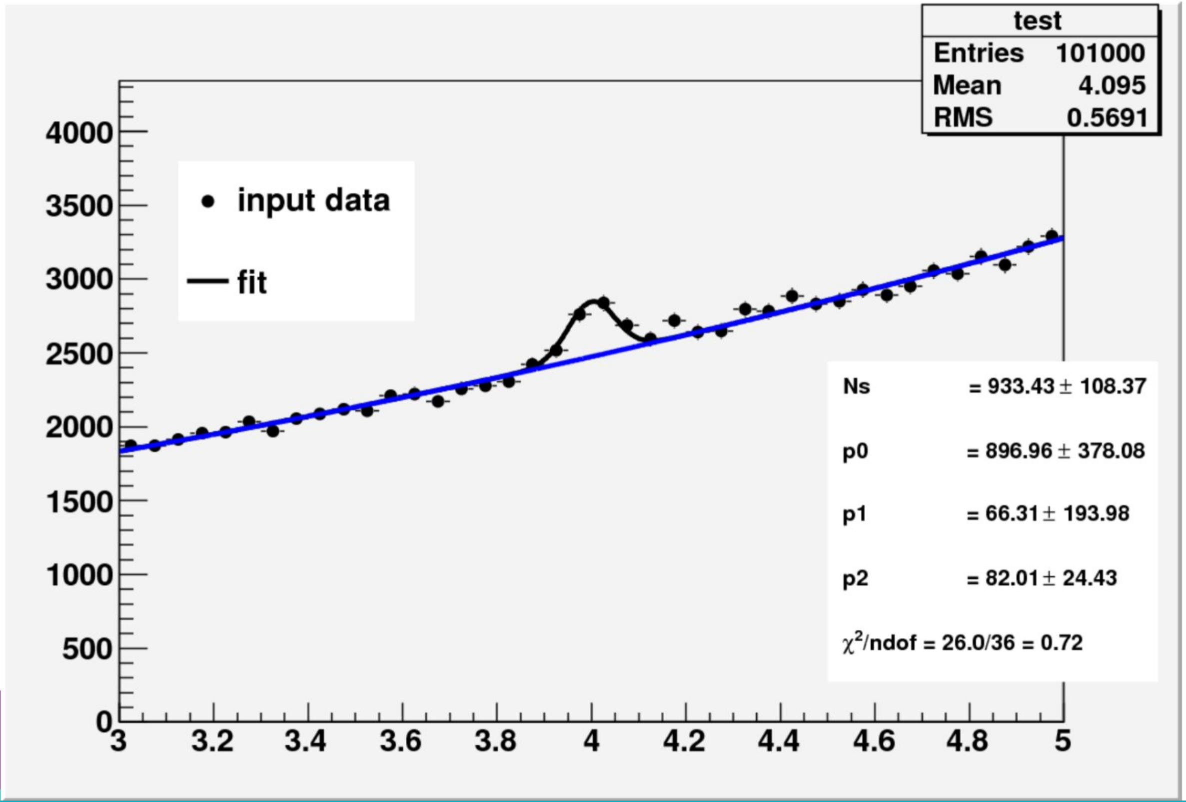
Fit function:
gauss+p1

$\tilde{\chi}^2 = 34.7$

Should we stop here?

$$\Delta\tilde{\chi}^2 = -2\ln\left(\frac{L(H_1)}{L(H_0)}\right) = -2\ln\left(\frac{L(g+p_1)}{L(g+p_0)}\right) = -2845.3$$

How many parameters do I need? – Example



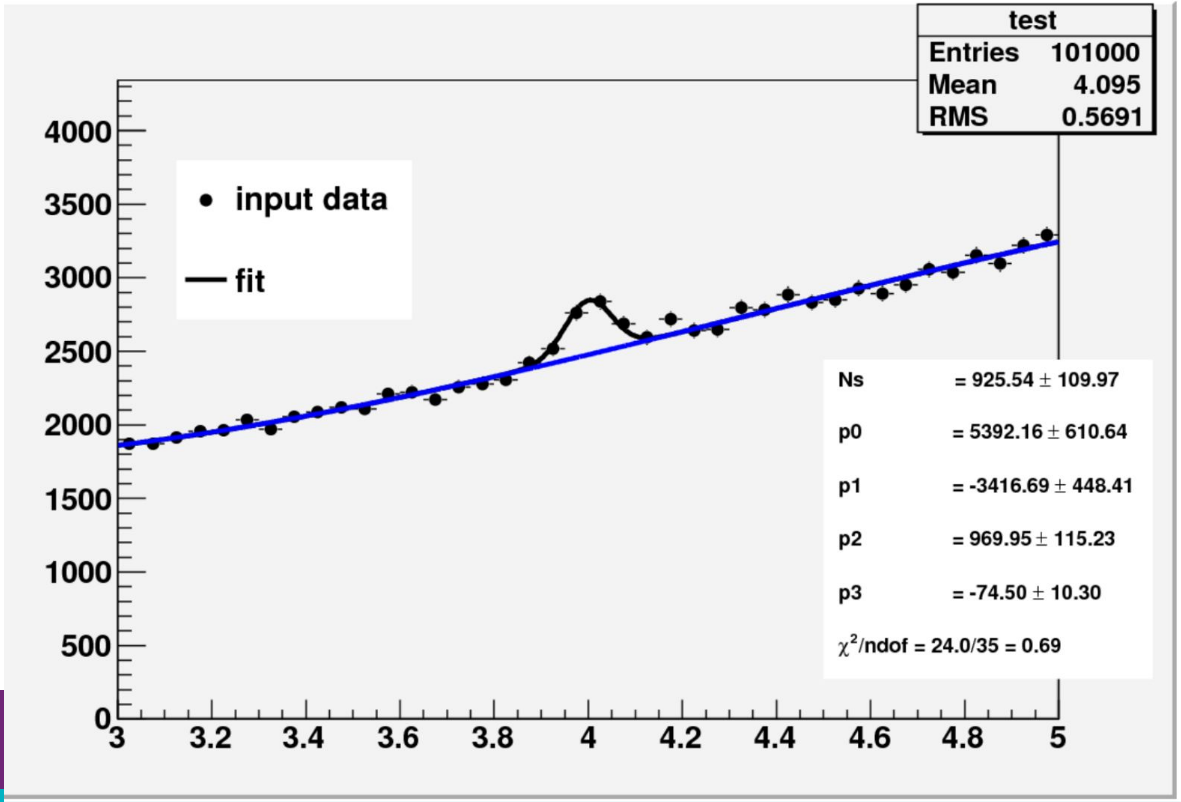
Fit function: gauss+p2

$$\tilde{\chi}^2 = 26.0$$

$$\Delta\tilde{\chi}^2 = -8.7$$

Should we stop here?

How many parameters do I need? – Example



Fit function: gauss+p3

$$\tilde{\chi}^2 = 24.0$$

$$\Delta\tilde{\chi}^2 = -2.0$$

Please stop!

How many parameters do I need? – Example

If H0 correct then according to Wilks' theorem: $-\Delta\chi^2$ should follow a χ^2 function with ndf=1 (in asymptotic regime of large n)

Note: p-values for χ^2: TMath::Prob(χ2 obs,ndf)

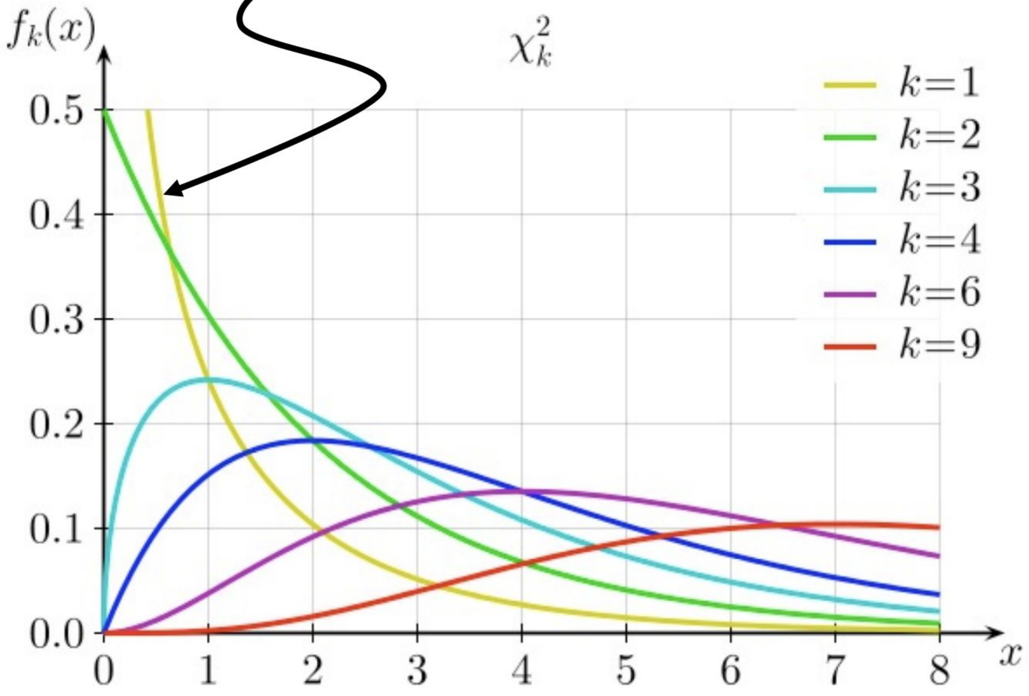
g+p0	$\tilde{\chi}^2 = 2880$		
g+p1	$\tilde{\chi}^2 = 34.7$	$\Delta\tilde{\chi}^2 = -2845.3$	
g+p2	$\tilde{\chi}^2 = 26.0$	$\Delta\tilde{\chi}^2 = -8.7$	P value = 0.003 Favoured over g+p1
g+p3	$\tilde{\chi}^2 = 24.0$	$\Delta\tilde{\chi}^2 = -2.0$	P value = 0.15 Not favoured over g+p2

χ^2

Remember:

What does a χ^2 distribution look like for n=1?

- Note that it for n=1, it does not peak at 1, but rather at 0…



Wouter Verkerke, NIKHEF

Improving on χ^2

Likelihood ratio is an improved χ^2 – S. Baker & R.D. Cousins, NIM 221 (1984) 437

- ➔ Still a single number
- ➔ “Optimal estimator” e.g. parameter estimation
- ➔ Requires a second hypothesis
- ➔ Allows taking uncertainties into account systematically!!
  - ◆ What if there’s a correlation?
  - ◆ What if there’s a systematic uncertainty

Alternatives to Likelihoods and χ^2

Kolgomorov-Smirnov test

F_c: Cumulative distribution function

F_e: Empirical distribution function

$$q_{GoF,KS} = \sup \left| F_c(x) - F_e(x) \right|$$

Anderson-Darling test

$$q_{GoF,AD} = n \cdot \int dF_e(x) \frac{\left(F_c(x) - F_e(x)\right)^2}{F_e(x) \cdot \left(1 - F_e(x)\right)}$$