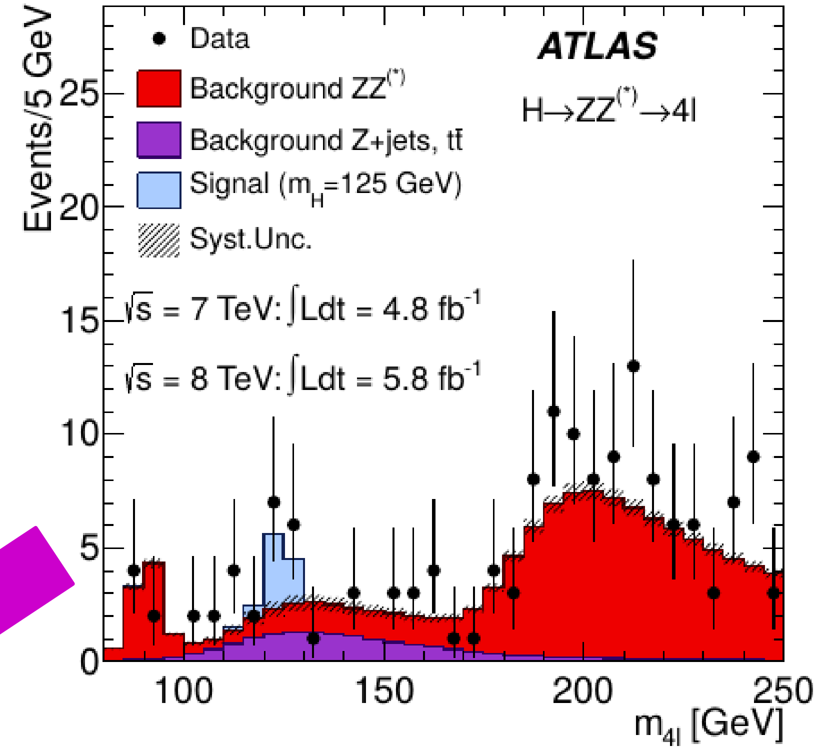
Statistical modeling and Systematic uncertainties in High Energy Physics

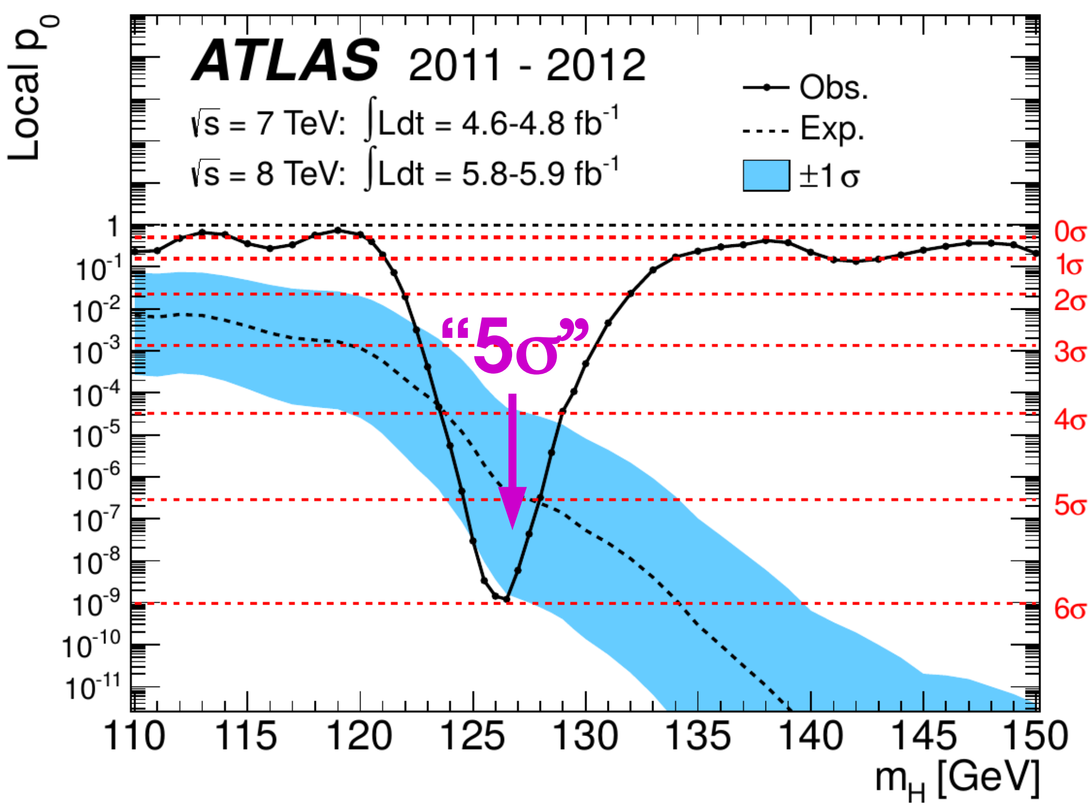
Nicolas Berger (LAPP Annecy)

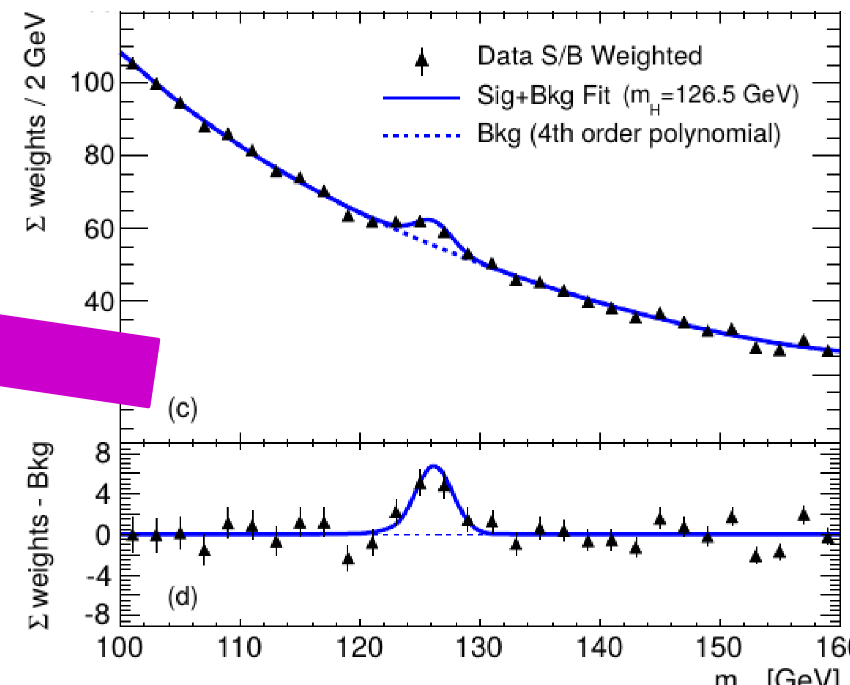
Introduction

Statistical methods play a critical role in many areas of physics

Higgs discovery : **“We have 5σ” !**



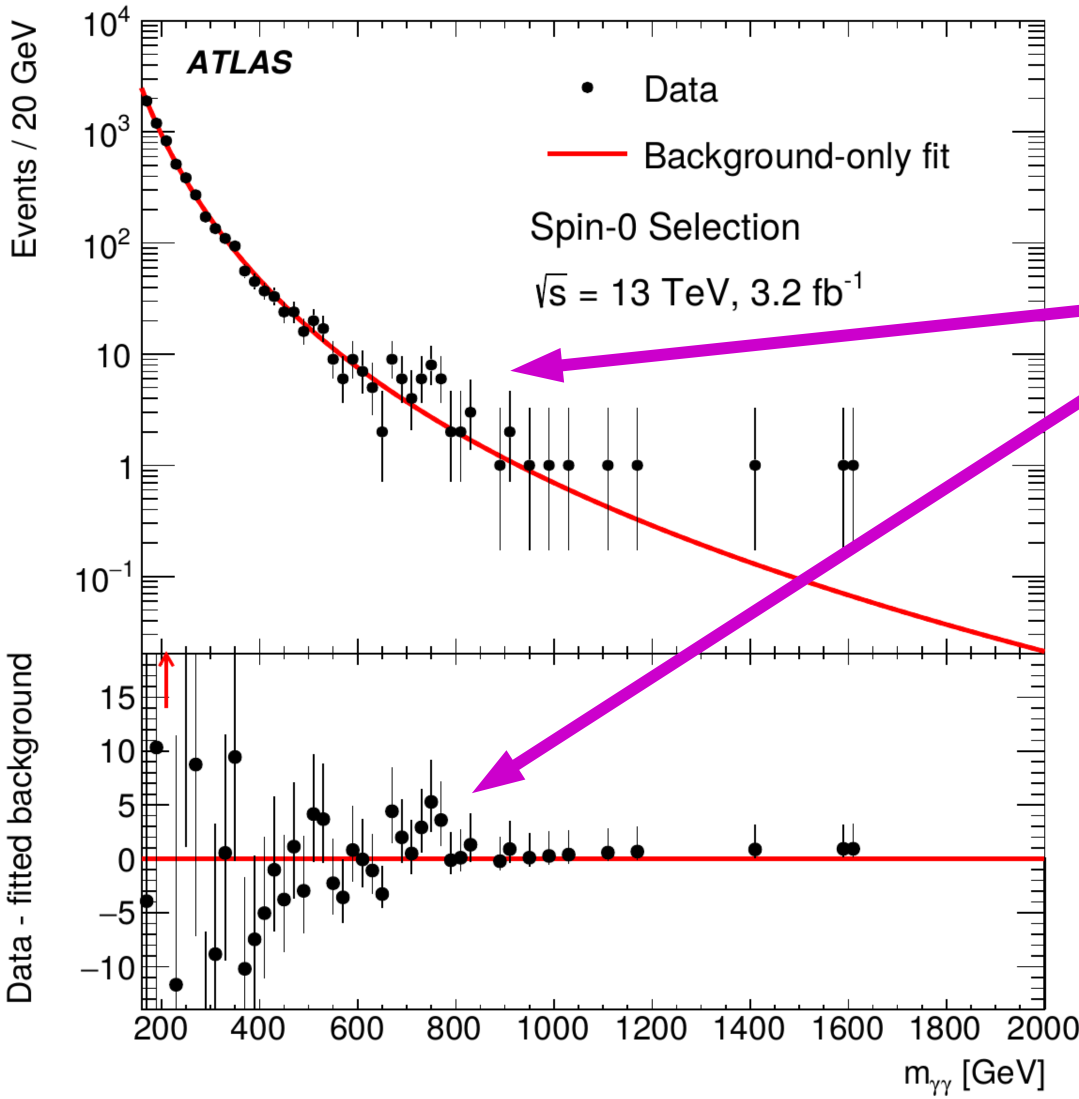




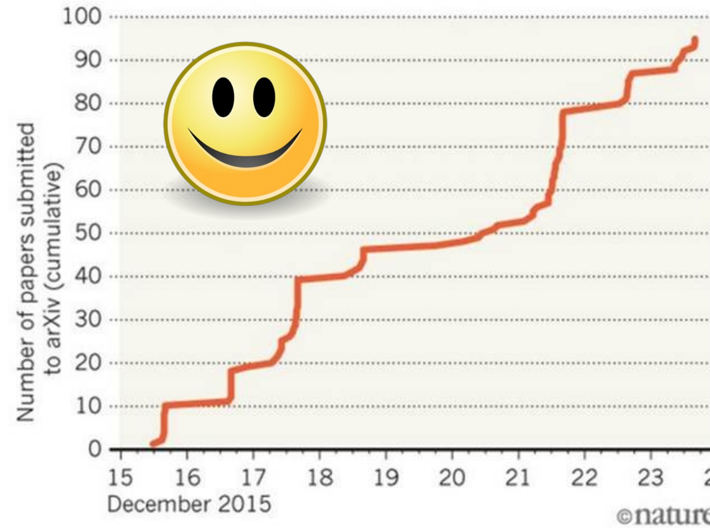
Phys. Lett. B 716 (2012) 1-29

Introduction

Sometimes difficult to distinguish a bona fide discovery from a **background fluctuation**...



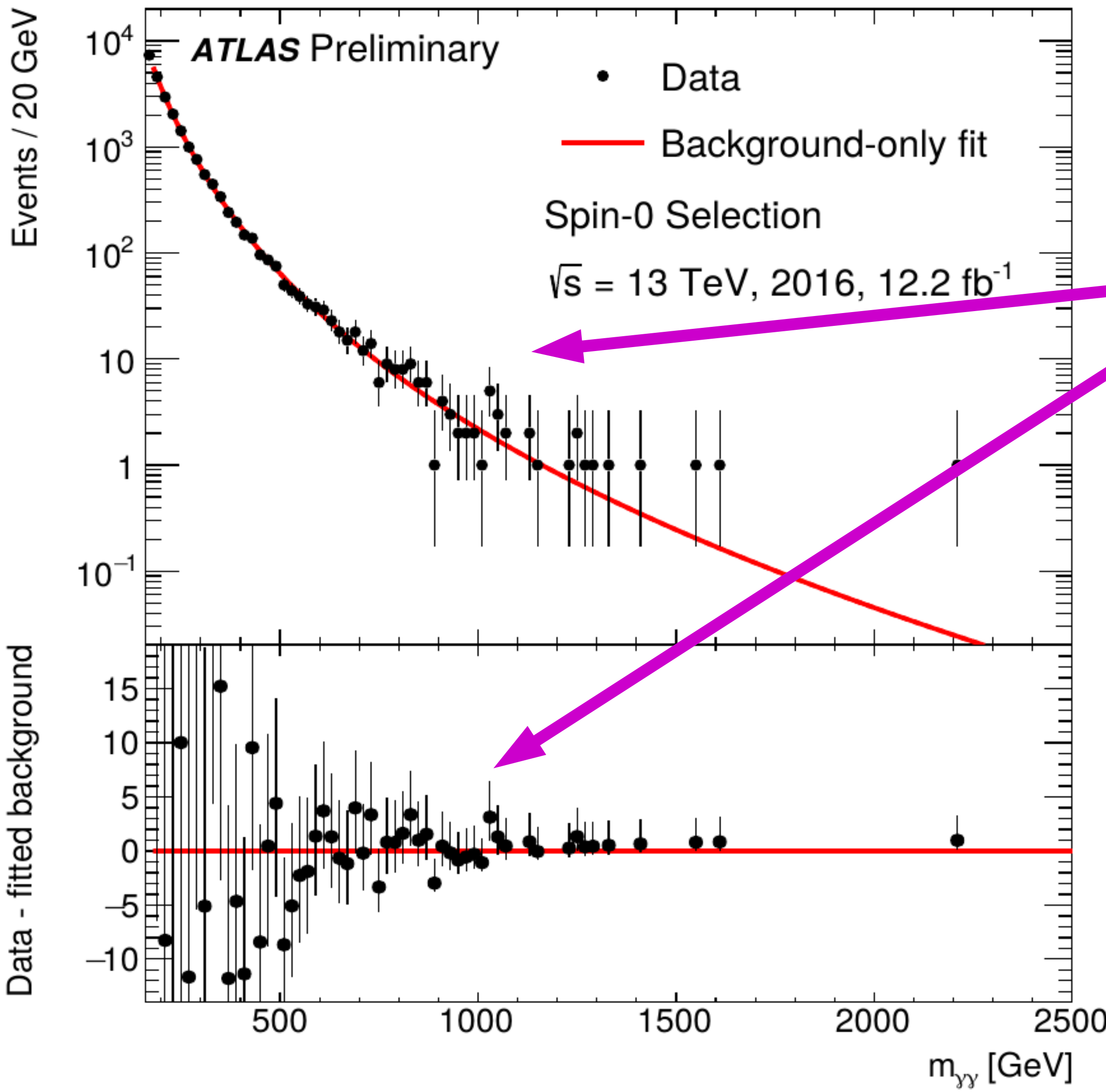
New Physics ?



JHEP 09 (2016) 1

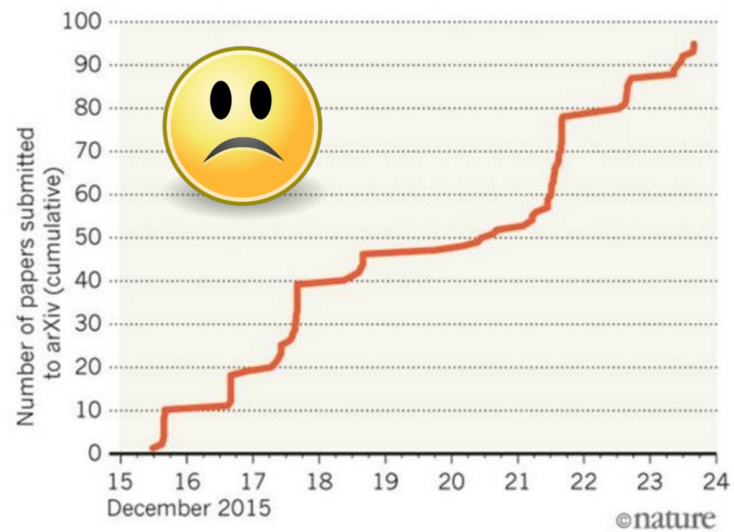
Introduction

Sometimes difficult to distinguish a bona fide discovery
from a **background fluctuation**...



A few months later...

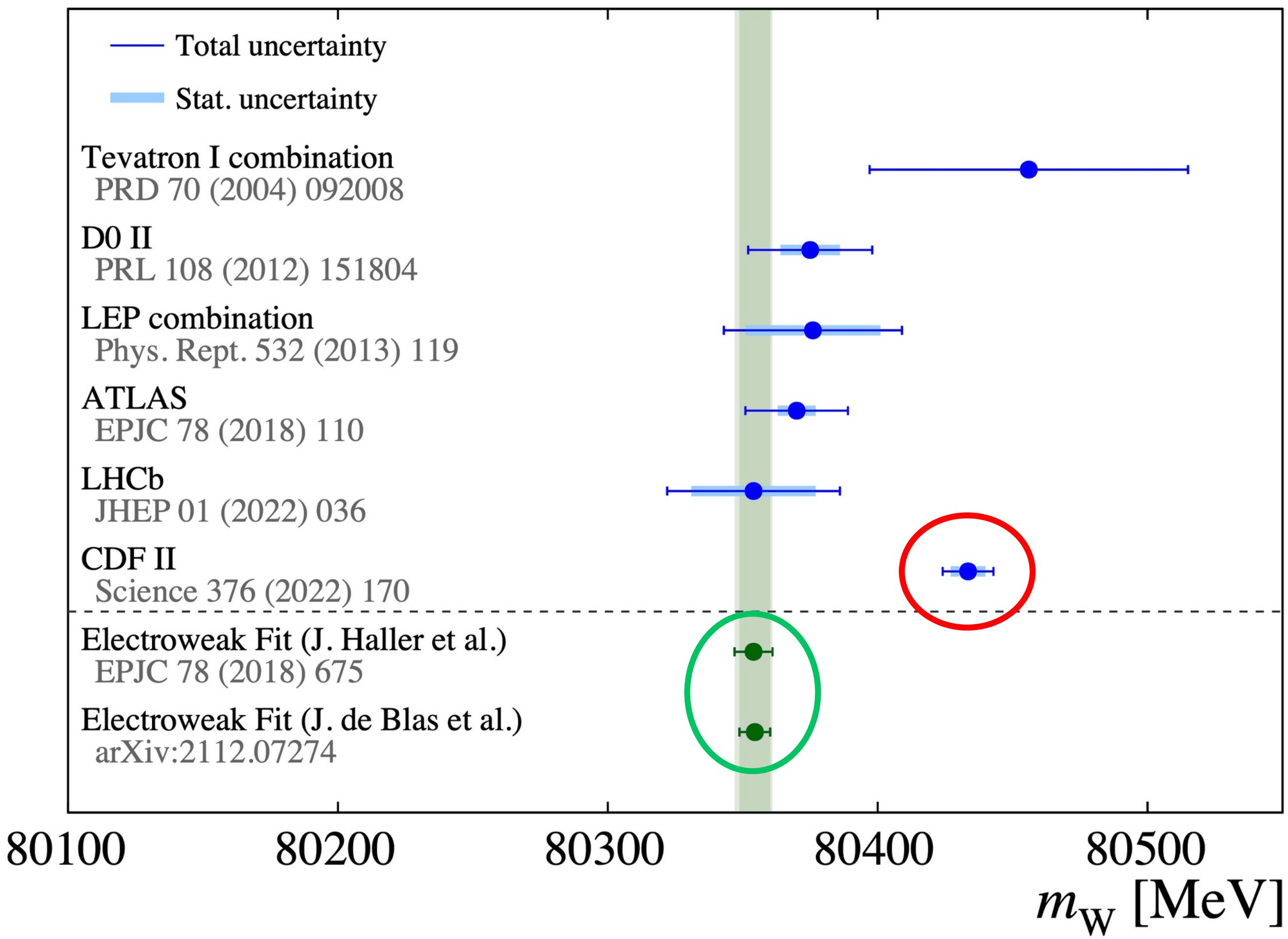
~~**New Physics ?**~~



JHEP 09 (2016) 1

Uncertainties

Many important questions answered by **precision measurements**,
Key point = determination of **uncertainties**

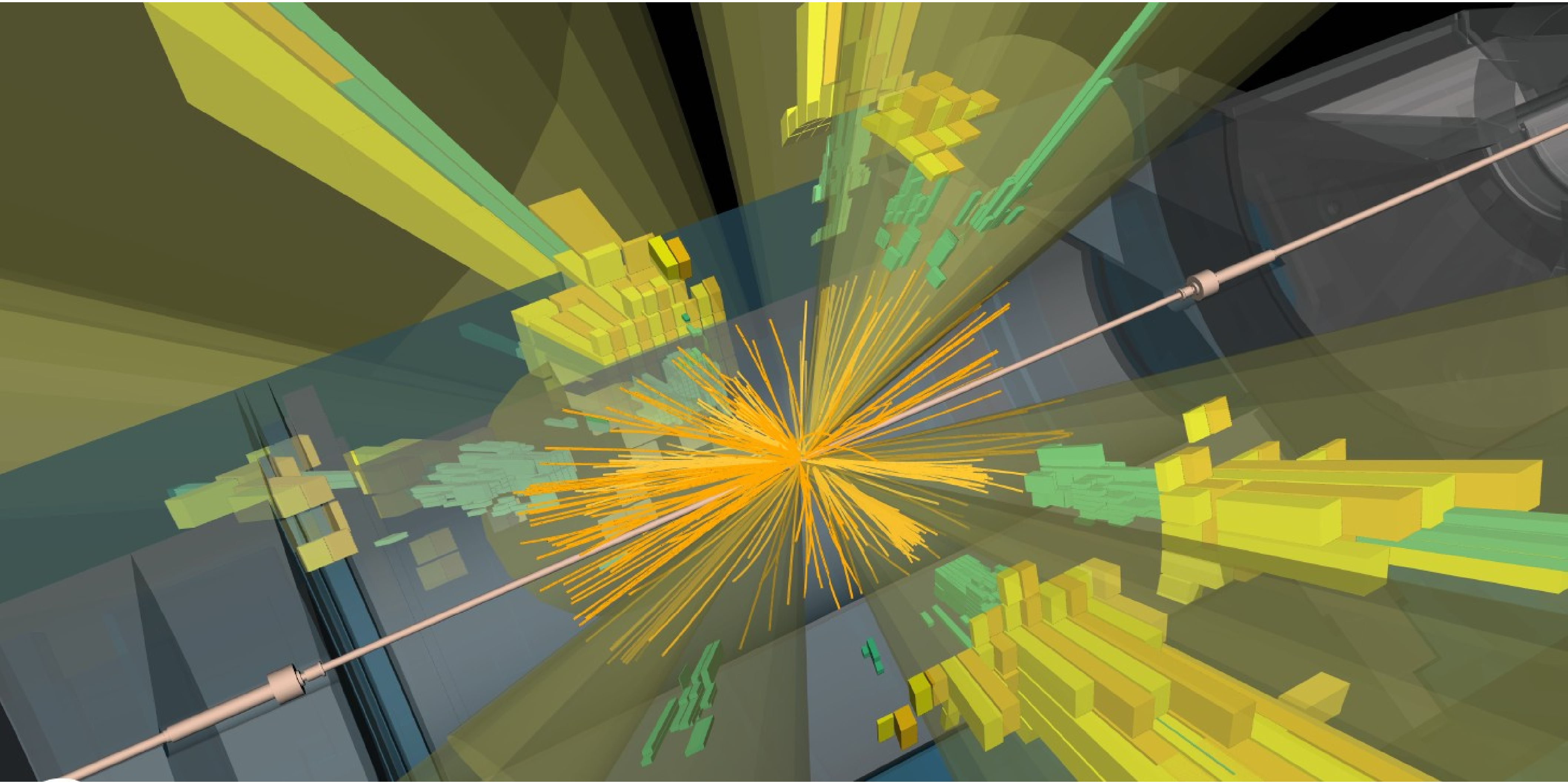


Science 376, 170–176 (2022), 8

$$M_W = 80,433.5 \pm 6.4_{\text{stat}} \pm 6.9_{\text{syst}} = 80,433.5 \pm 9.4\ \text{MeV}/c^2$$

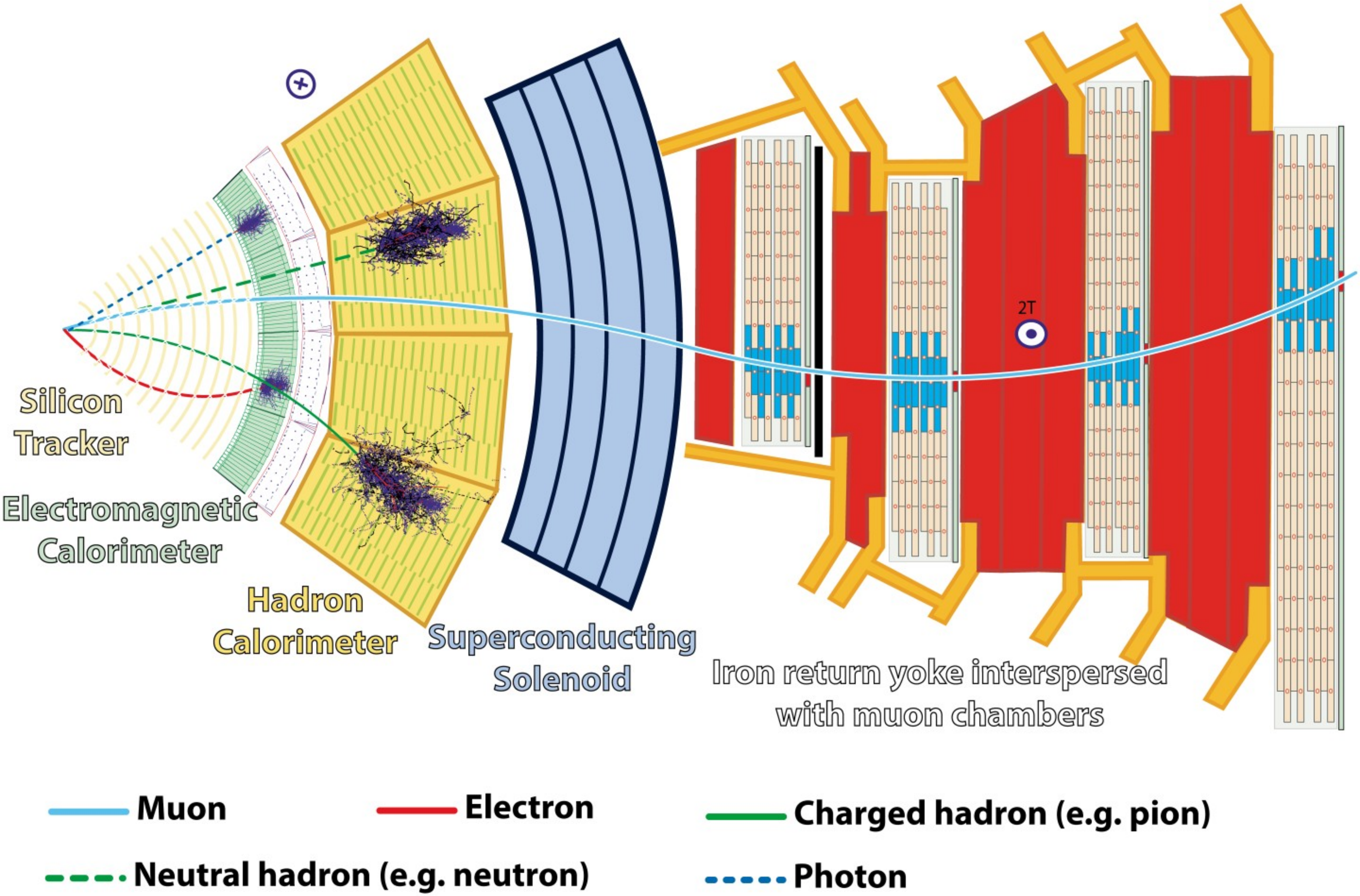
Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes



Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes



Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes

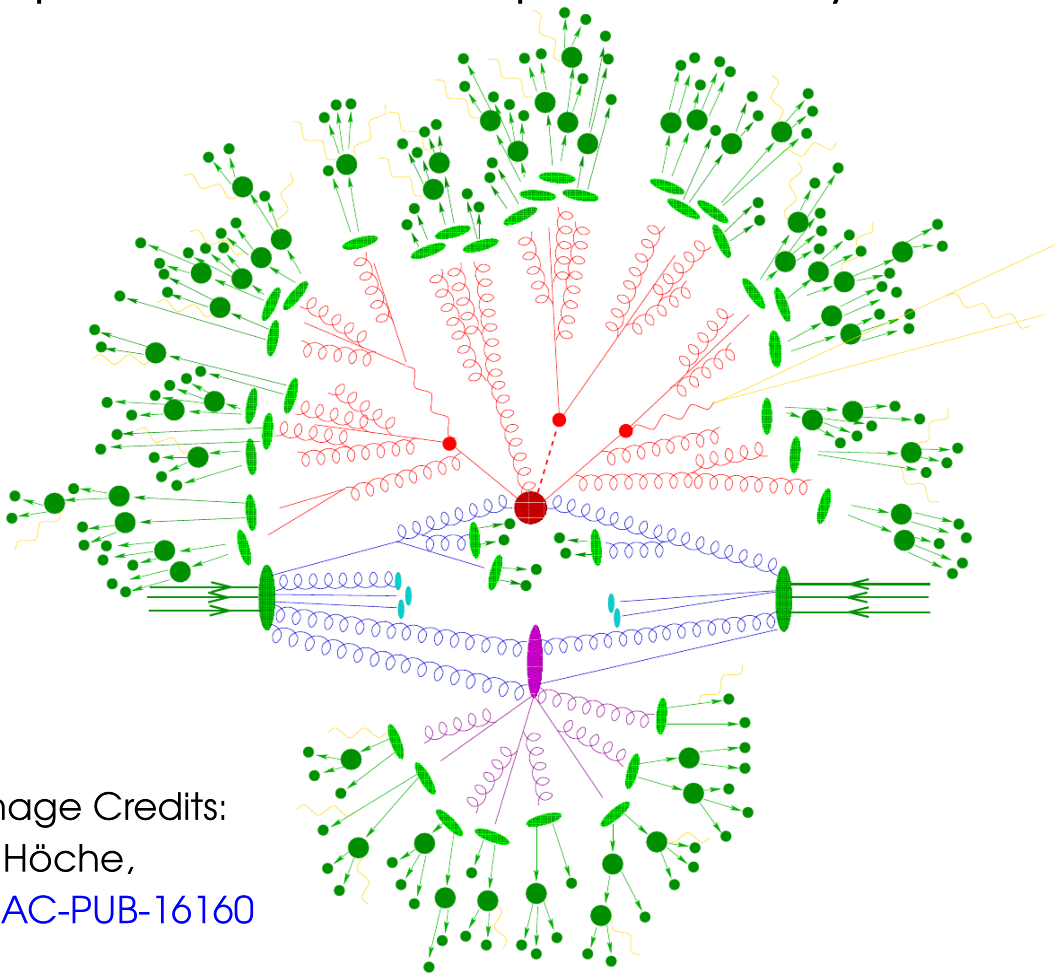


Image Credits:
S. Höche,
SLAC-PUB-16160

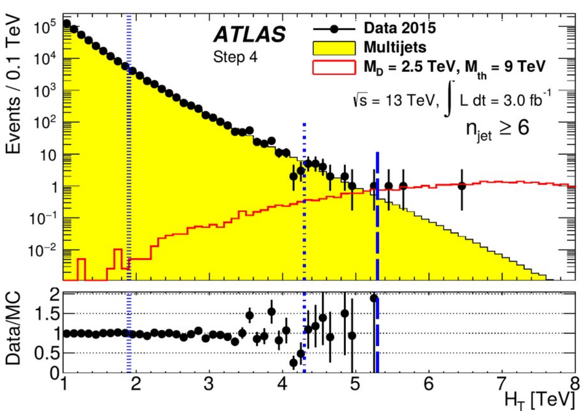
Hard scattering

PDFs, Parton shower, Pileup

Decays

Detector response

Reconstruction



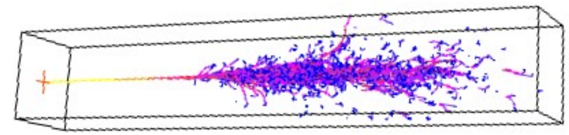
Randomness involved in all stages

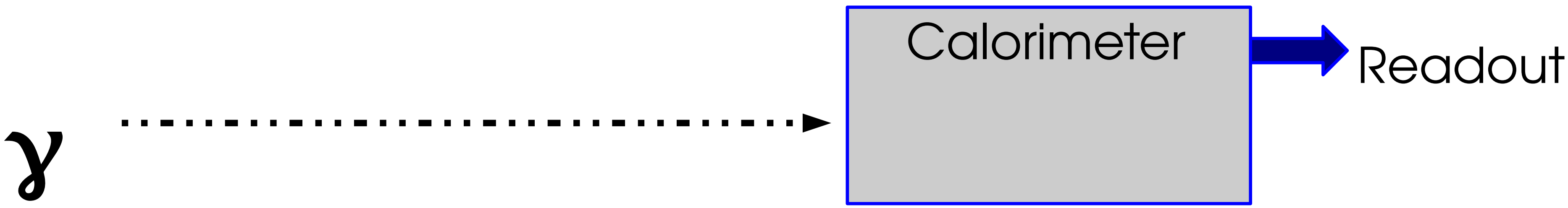
→ **Classical** randomness: detector reponse

→ **Quantum** effects in particle production, decay

Measurement Errors: Energy measurement

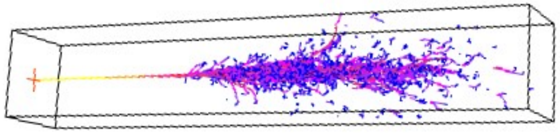
Example: measuring the energy
of a photon in a calorimeter

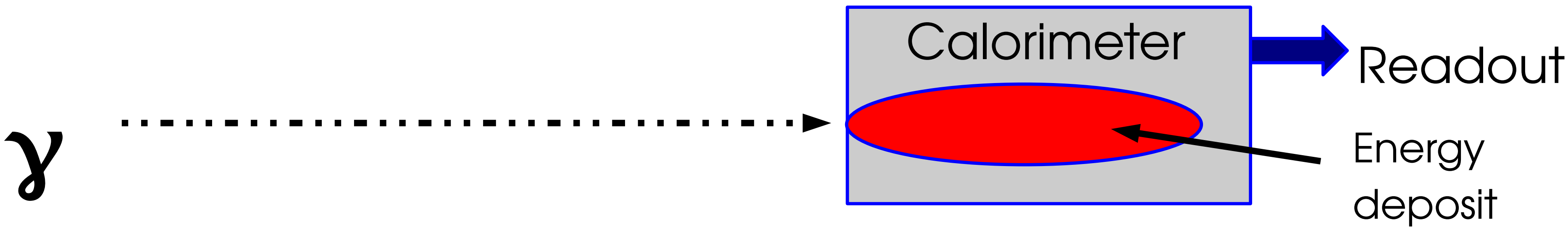


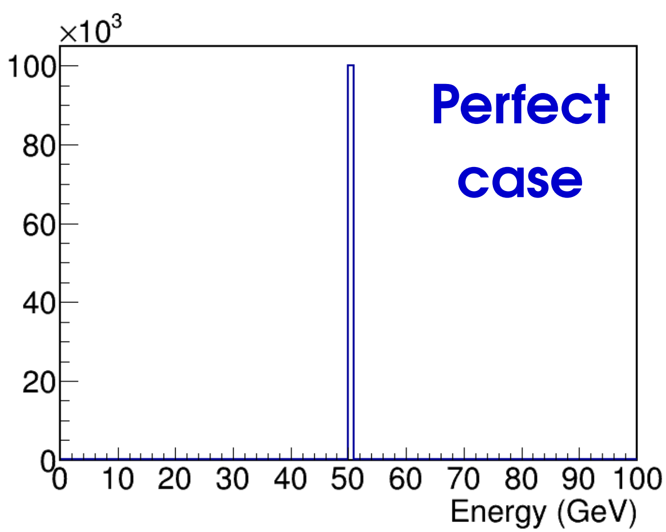


Measurement Errors: Energy measurement

Example: measuring the energy
of a photon in a calorimeter

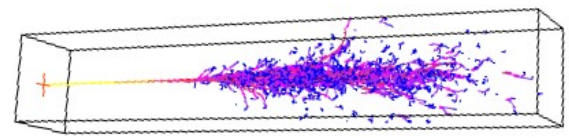


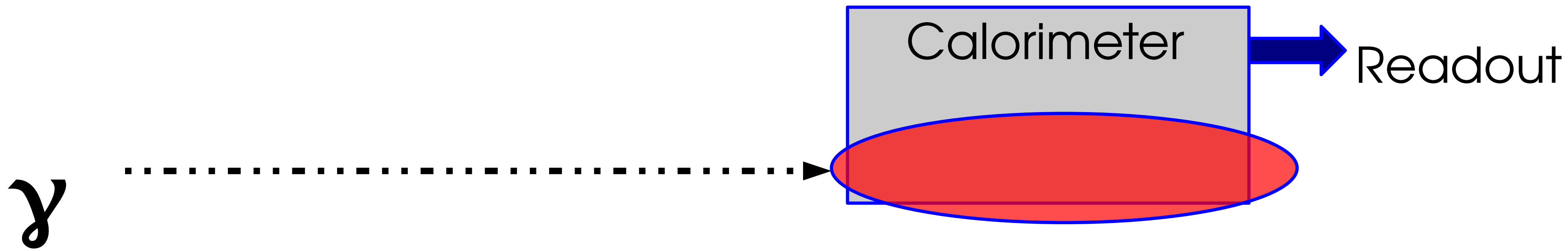


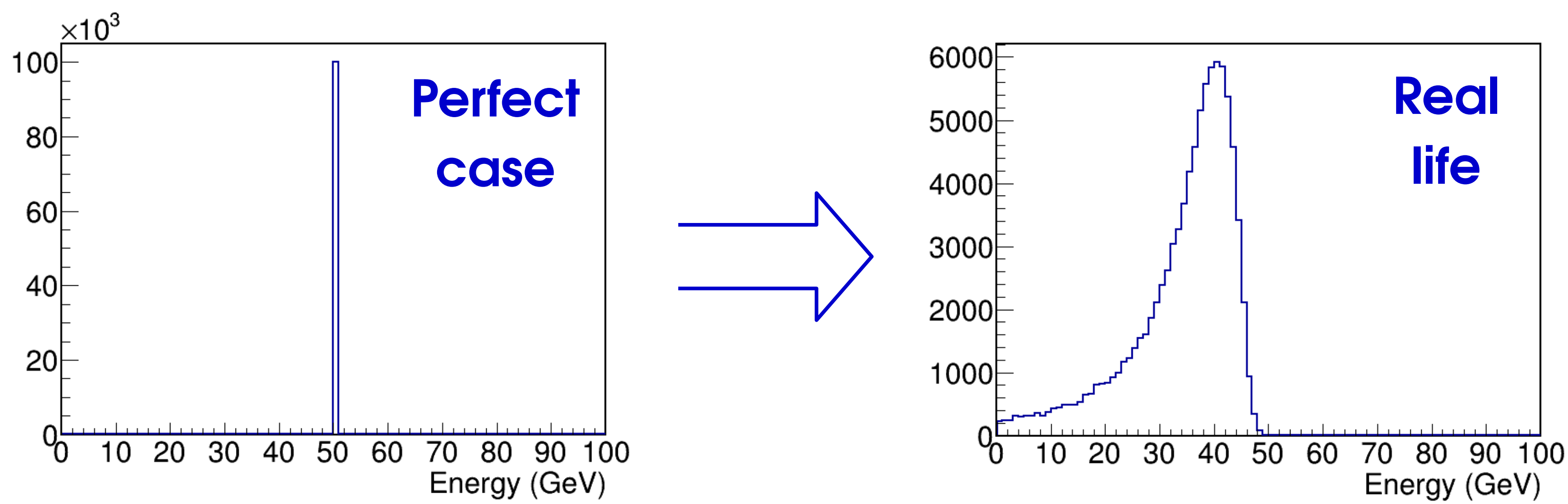


Measurement Errors: Energy measurement

Example: measuring the energy
of a photon in a calorimeter



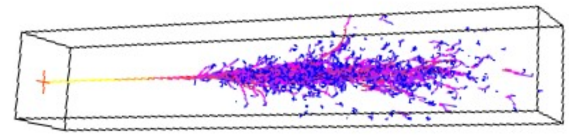


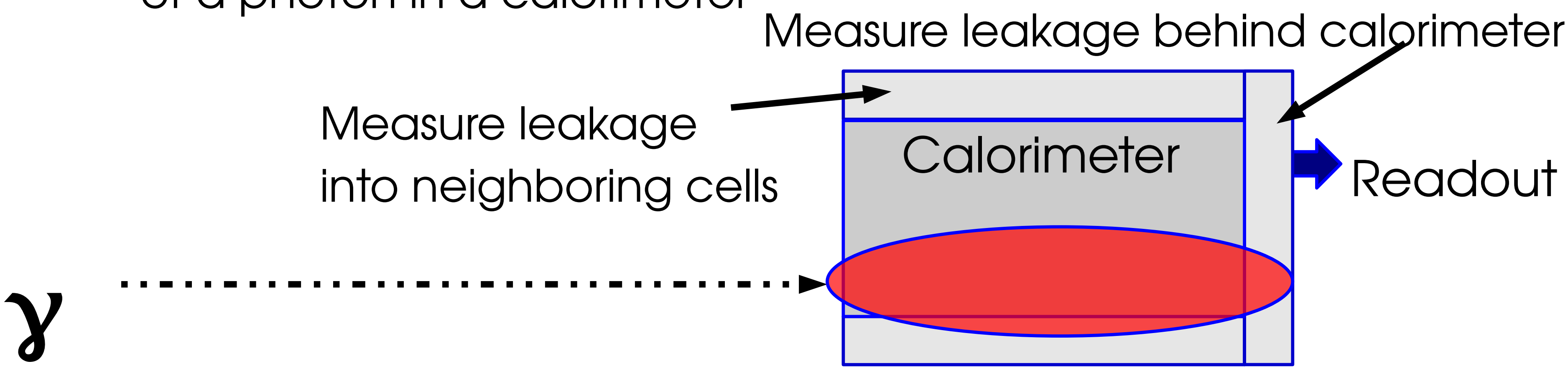


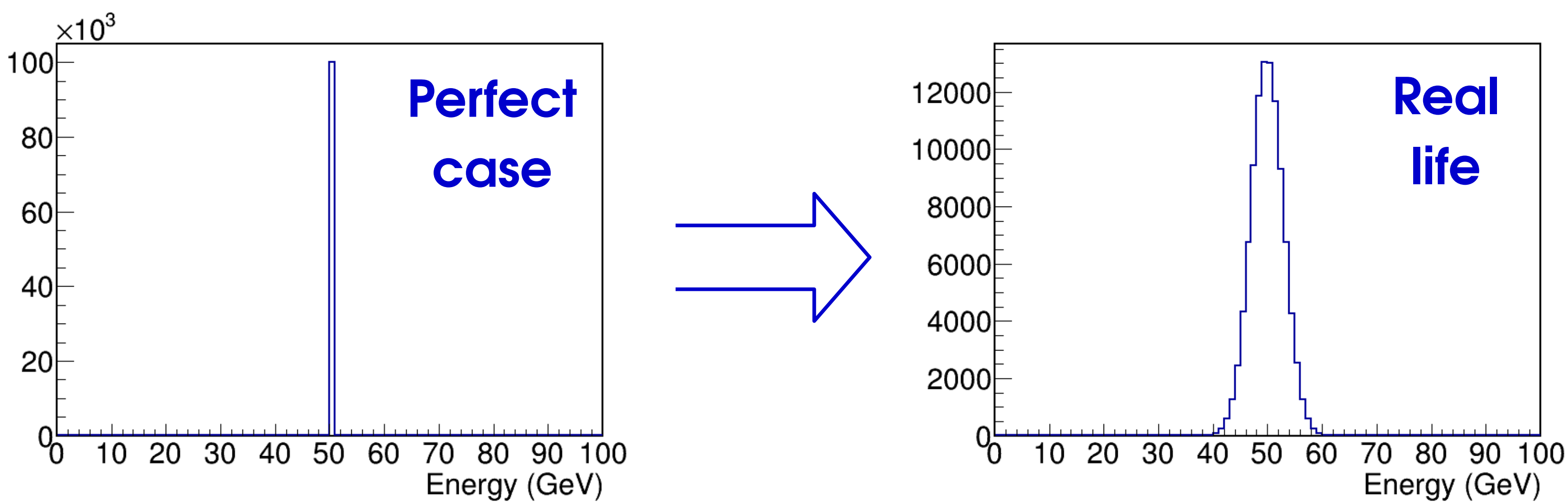
Measurement Errors: Energy measurement

Example: measuring the energy

of a photon in a calorimeter



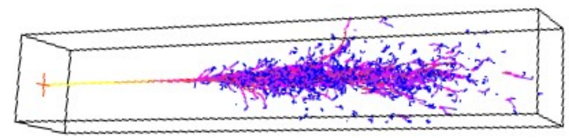


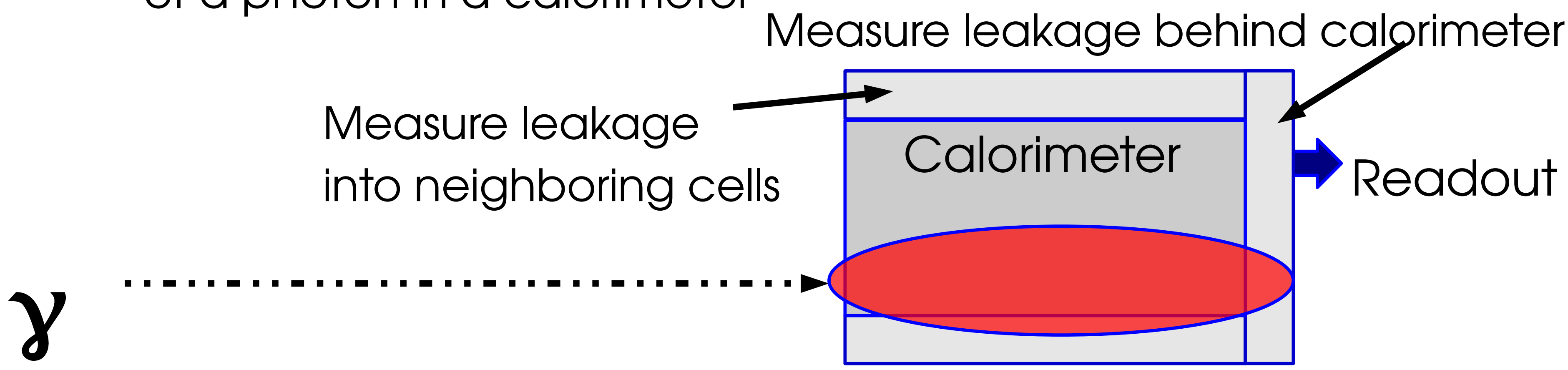


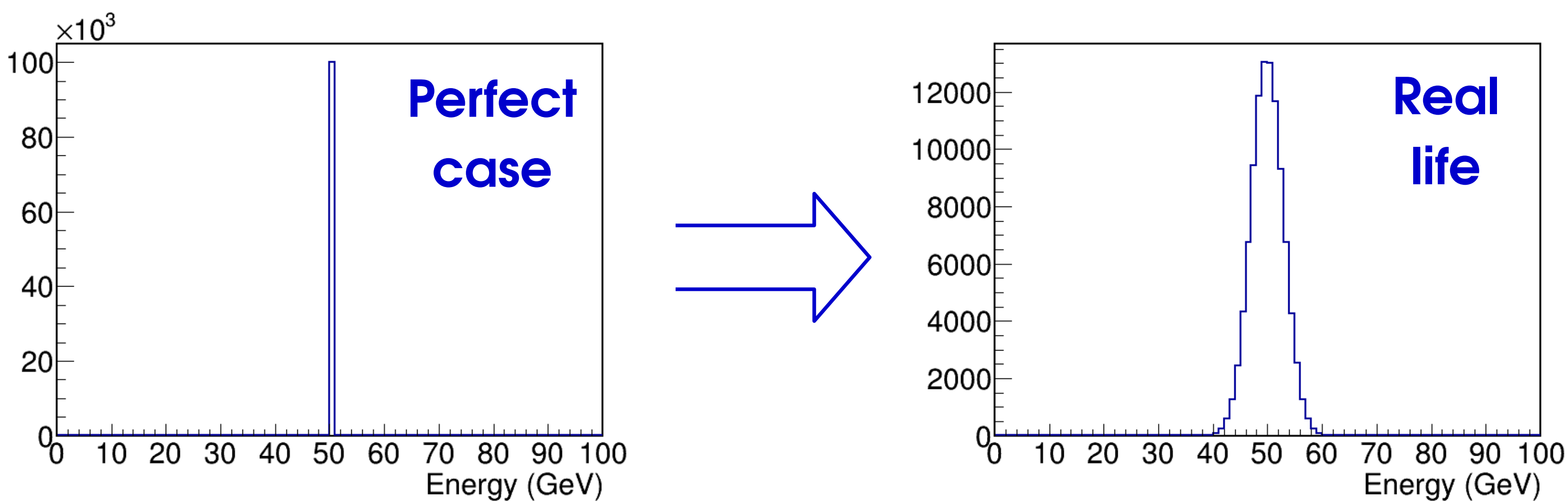
Measurement Errors: Energy measurement

Example: measuring the energy

of a photon in a calorimeter





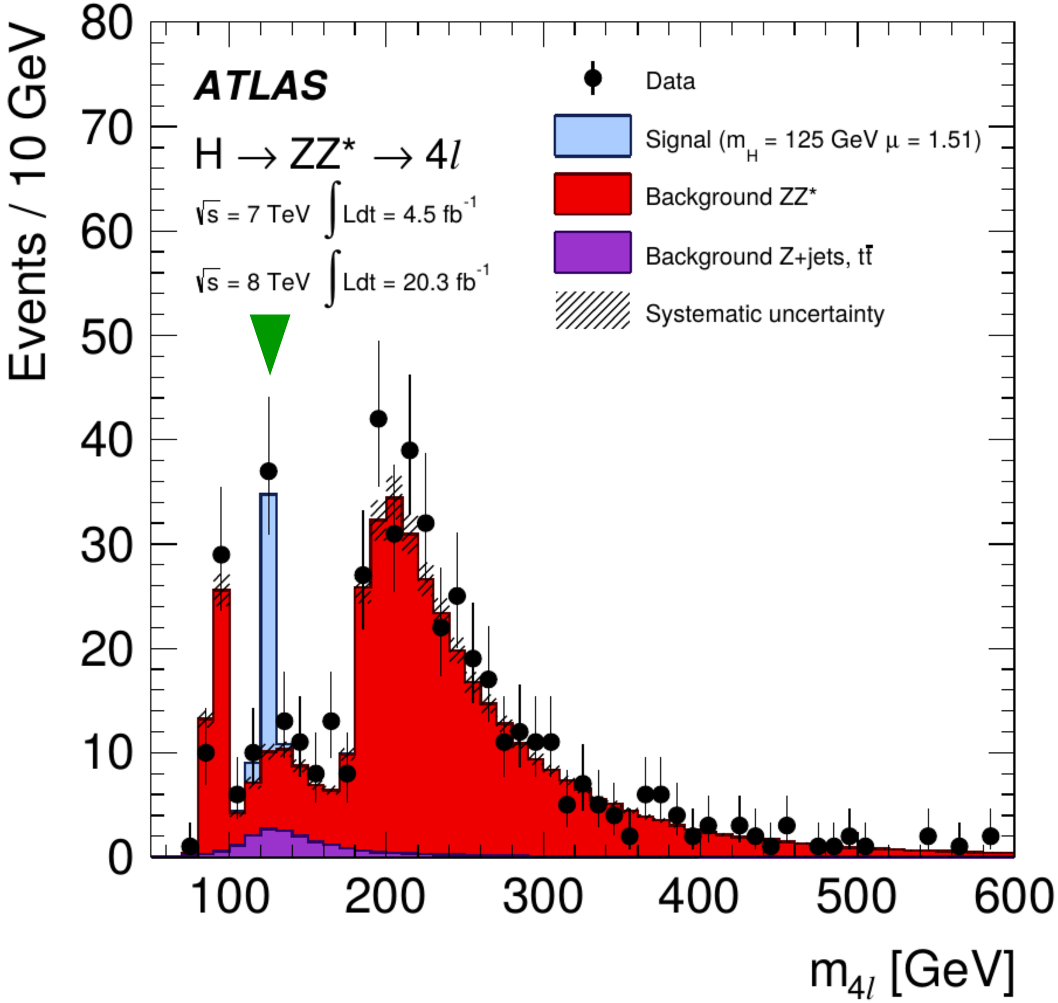


Cannot predict the measured value for a given event

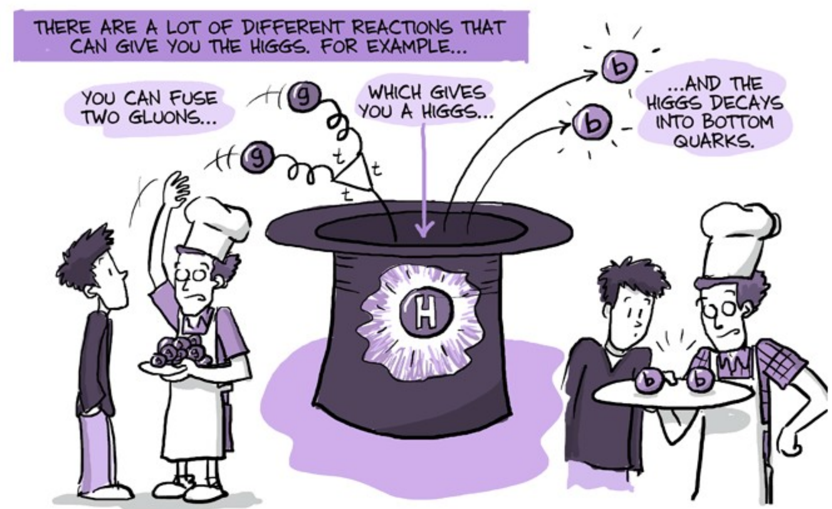
$\Rightarrow$ **Random process** $\Rightarrow$ **Need a probabilistic description**

Quantum Randomness: H→ZZ*→4l

Phys. Rev. D **91**, 012006



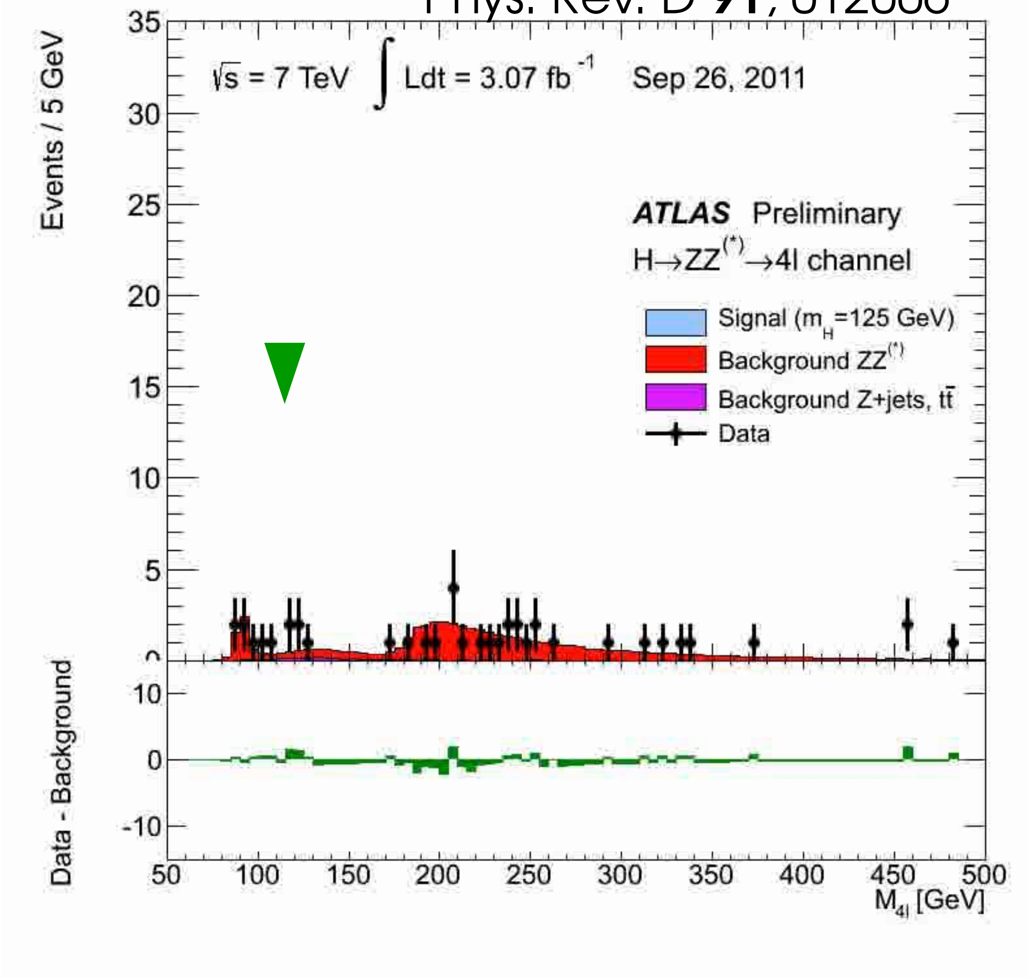
Rare process: Expect 1 signal event every **~6 days**



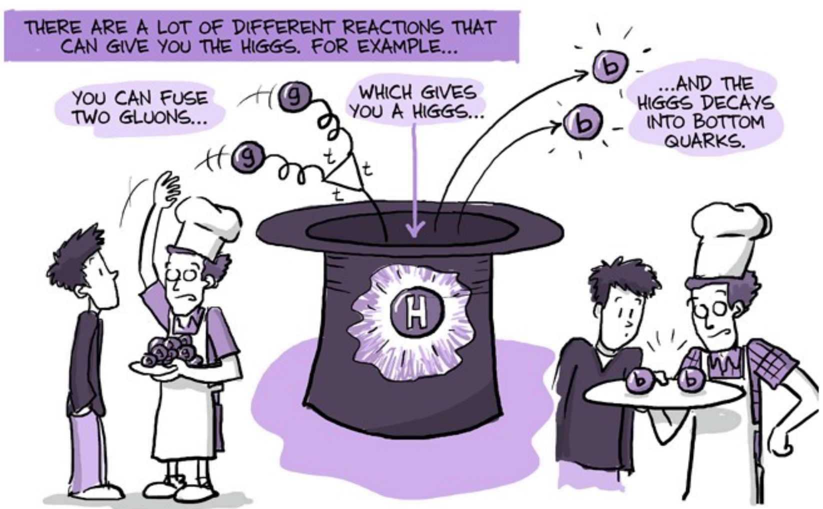
http://www.phdcomics.com/comics/archive.php?comicid=1489

Quantum Randomness: H→ZZ*→4l

Phys. Rev. D **91**, 012006



Rare process: Expect 1 signal event every **~6 days**

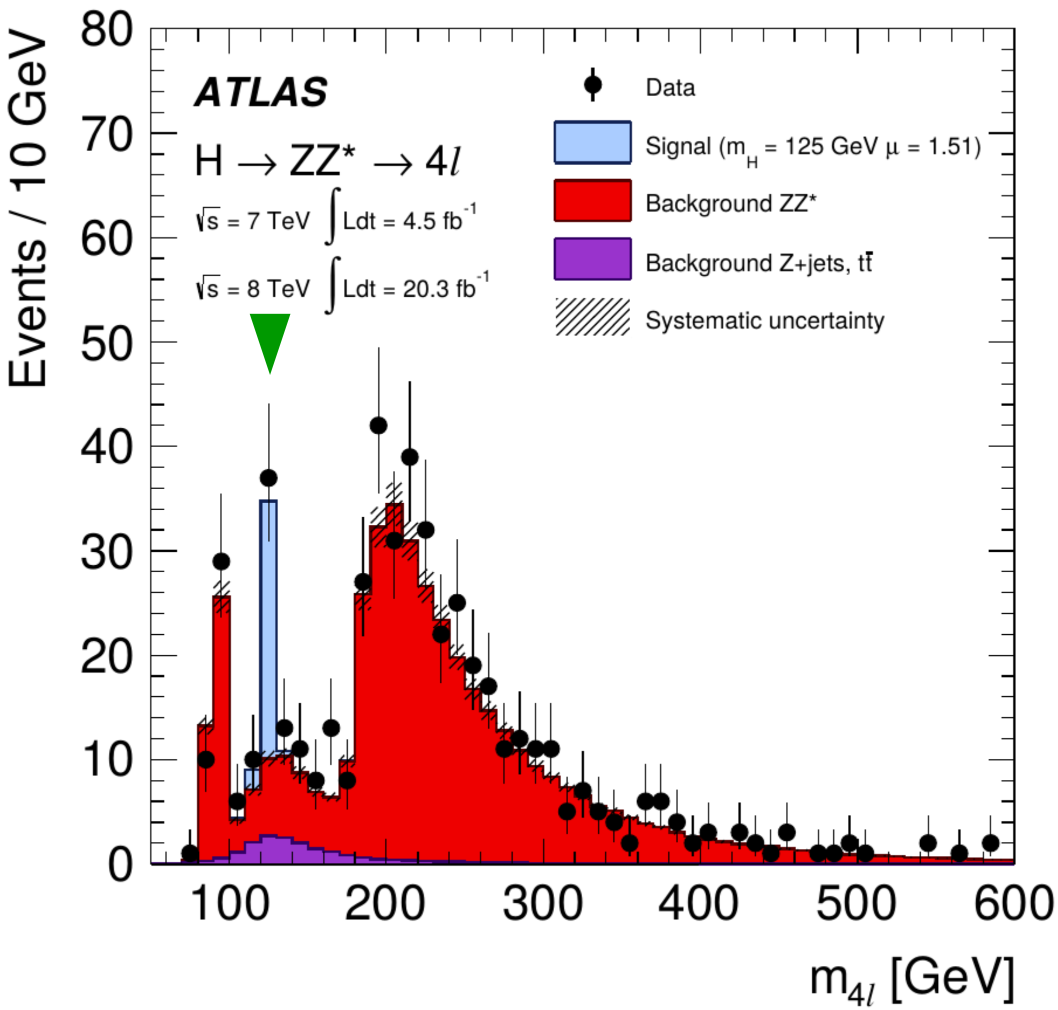


http://www.phdcomics.com/comics/archive.php?comicid=1489

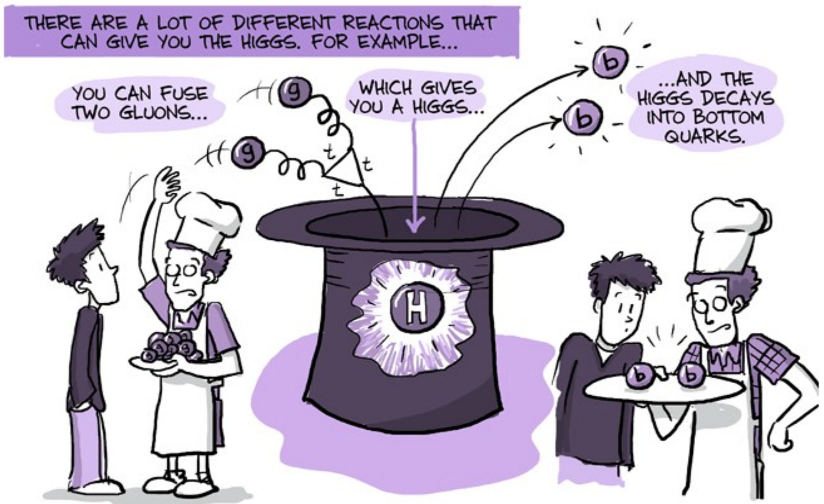
View online

Quantum Randomness: H→ZZ*→4l

Phys. Rev. D **91**, 012006



Rare process: Expect 1 signal event every **~6 days**



http://www.phdcomics.com/comics/archive.php?comicid=1489

"Will I get an event today ?" → only **probabilistic** answer

Performing a measurement

Phys. Lett. B 759 (2016) 601

Measure the cross-section (event rate) of the $Z \rightarrow ee$ process

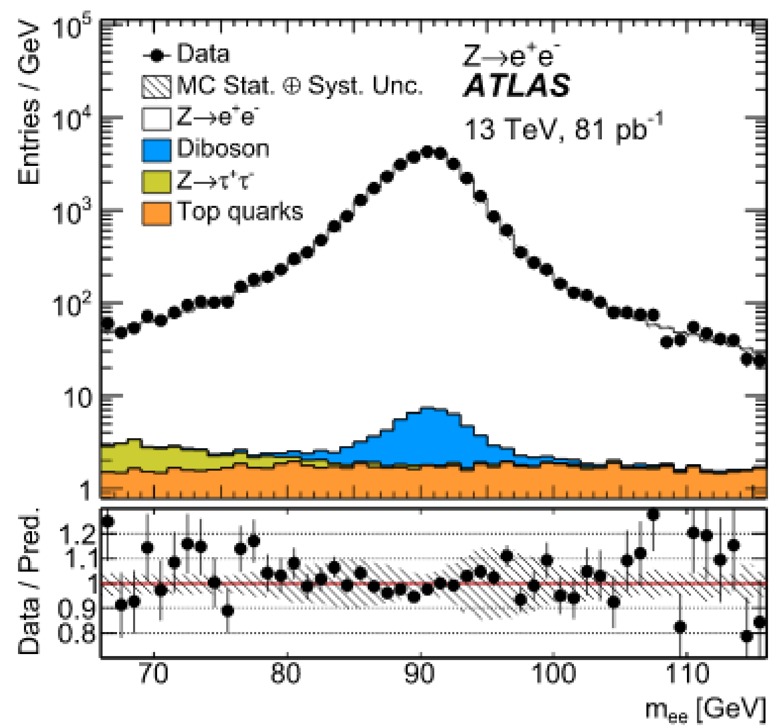
35000 ± 187

175 ± 8

$$\sigma^{fid} = \frac{n_{data} - N_{bkg}}{C_{fid}\, L}$$

(81 ± 2) pb^{-1}

0.552 ± 0.006



σ^{fid} = 0.781 ± 0.004 (stat) ± 0.018 (syst) nb

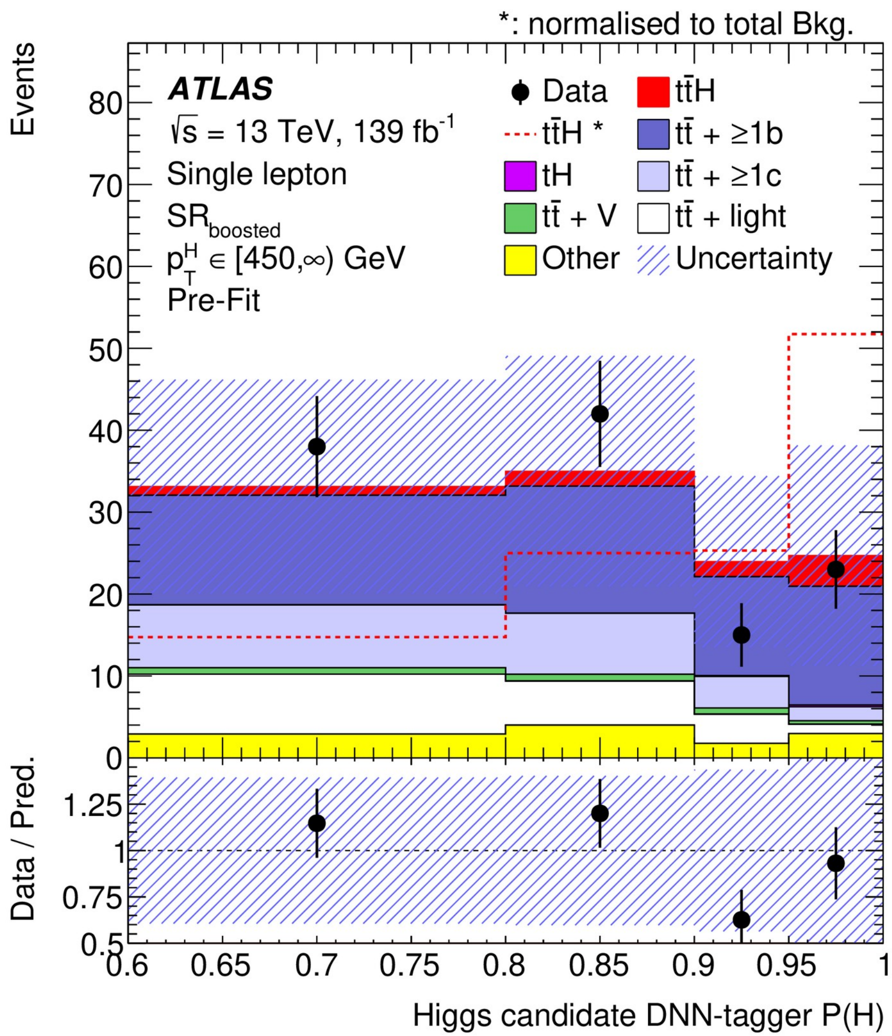
Fluctuations in the data counts

Other uncertainties (assumptions, parameter values)

"Single bin counting" : only data input is **N_{data}**.

Example 2: ttH→bb

arXiv:2111.06712



Event counting in different regions:

Multiple-bin counting

Lots of information available

→ Potentially higher sensitivity

→ How to make optimal use of it ?

HEP Statistical Modeling

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

→ In principle, binomial process

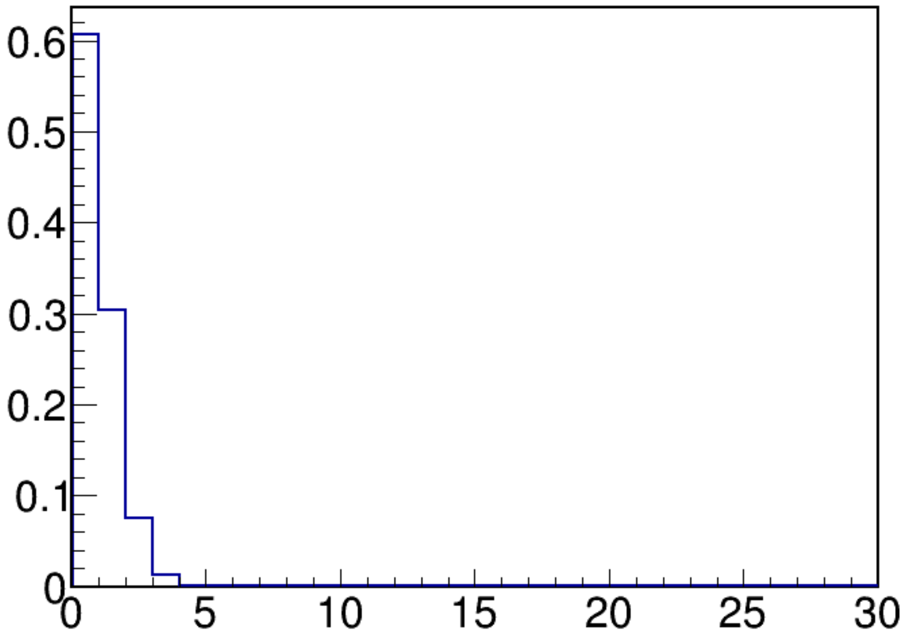
→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$

λ = 0.5



Mean = λ

Variance = λ

σ = $\sqrt{\lambda}$

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

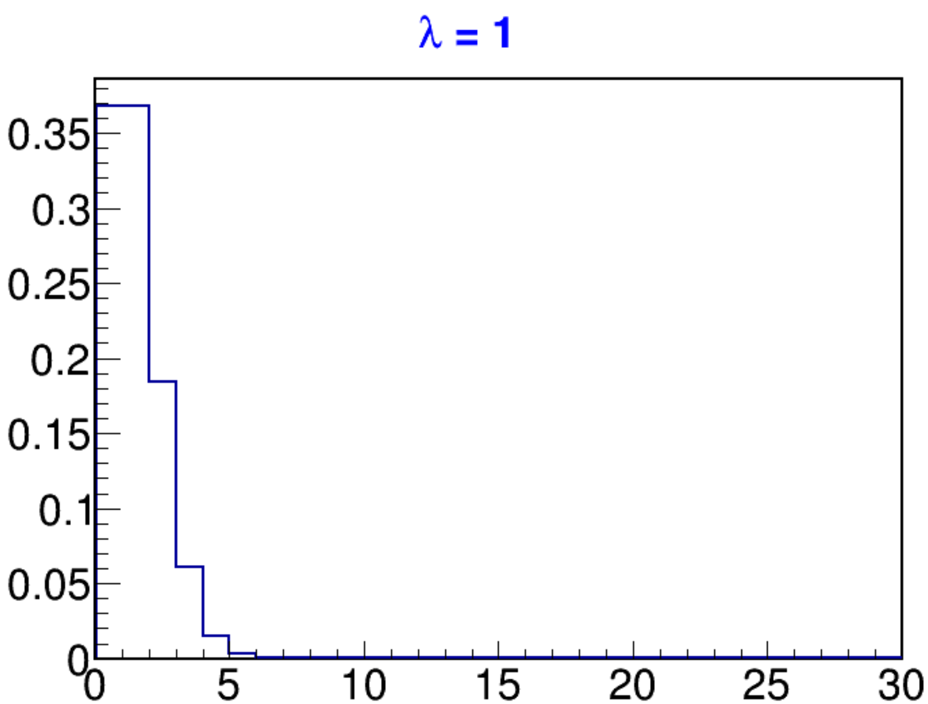
→ In principle, binomial process

→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = $\sqrt{\lambda}$

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

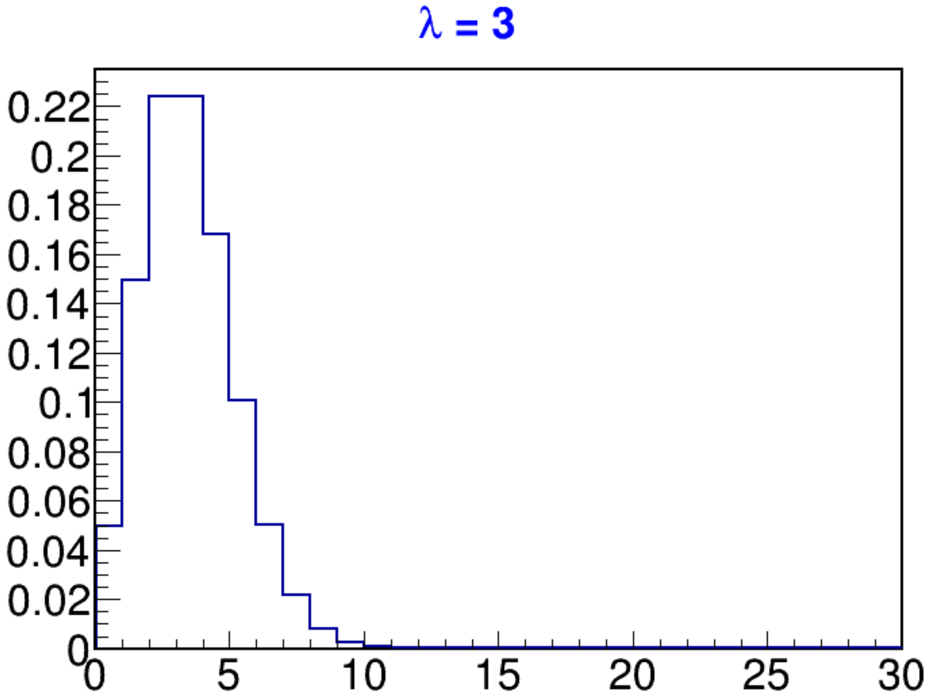
→ In principle, binomial process

→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

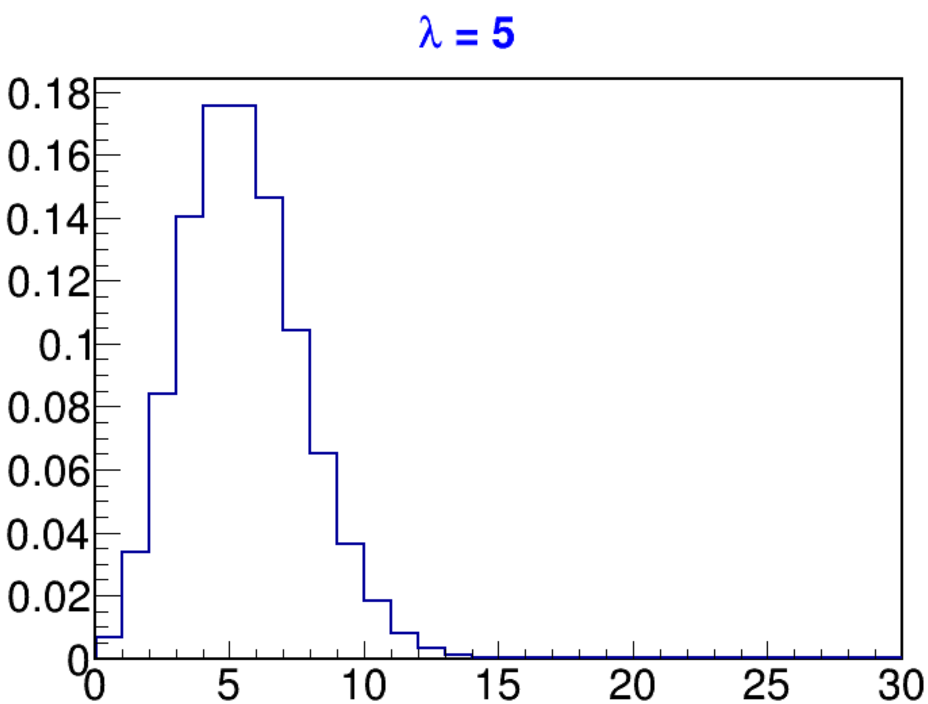
→ In principle, binomial process

→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = $\sqrt{\lambda}$

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \xrightarrow{\lambda \to \infty} G(\lambda, \sqrt{\lambda})$$

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

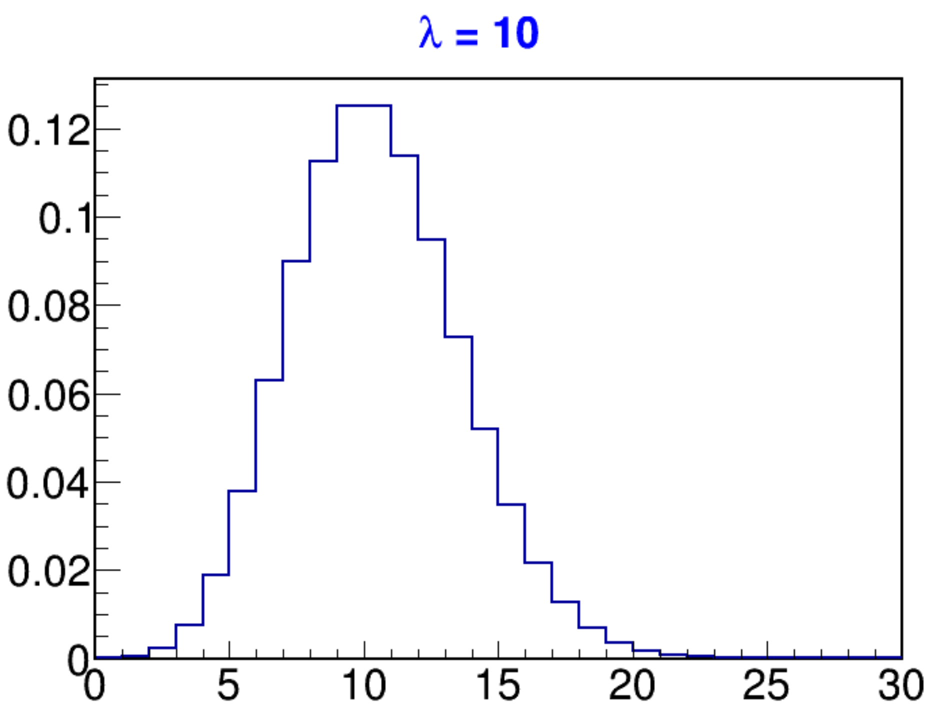
→ In principle, binomial process

→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\simeq} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\simeq} e^{-\lambda}$$



Mean = λ

Variance = λ

σ = √λ

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \overset{\lambda \to \infty}{\to} G(\lambda, \sqrt{\lambda})$$

How to count

Collider processes: produce (many) events N, select a (very) small fraction P

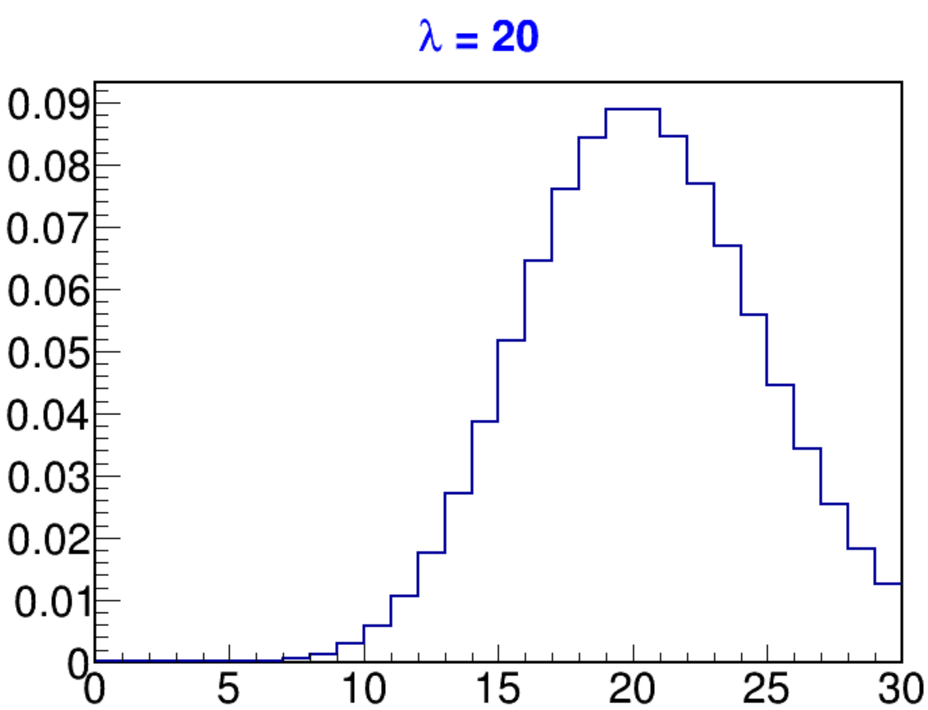
→ In principle, binomial process

→ In practice, **P ≪ 1, N ≫ 1**, ⇒ Poisson approximation.

→ *i.e.* **very rare** process, but **very many trials** so still expect to see good events

Poisson distribution: $$P(n;\lambda) = e^{-\lambda}\frac{\lambda^n}{n!}$$

$$(1-P)^{N-n} \overset{n \ll N}{\sim} \left(1-\frac{\lambda}{N}\right)^N \overset{N \gg 1}{\sim} e^{-\lambda}$$



Mean = λ

Variance = λ

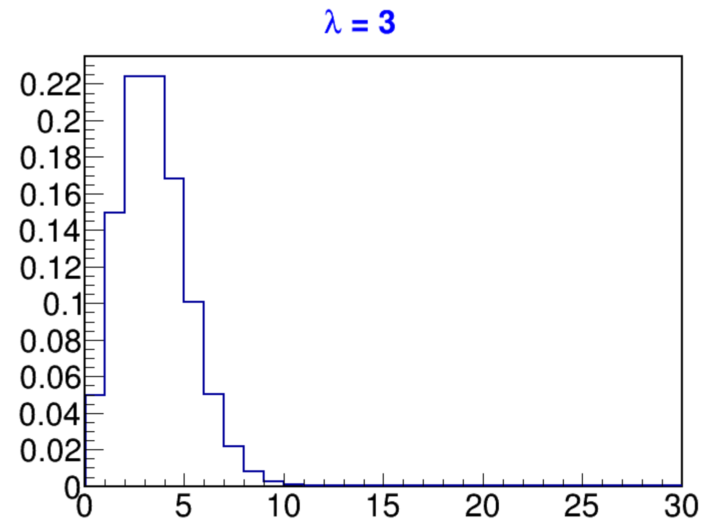
σ = $\sqrt{\lambda}$

Central limit theorem :

becomes **Gaussian for large λ** :

$$P(\lambda) \overset{\lambda \to \infty}{\to} G(\lambda, \sqrt{\lambda})$$

Statistical Model for Counting



Observable: **number of events n**

Typically both **S**ignal and **B**ackground present:

$$P(n; S, B) = e^{-(S+B)} \frac{(S+B)^n}{n!}$$

S : # of events from signal process

B : # of events from bkg. process(es)

Model has **parameters S** and **B.**

B can be known a priori or not (S usually not...)

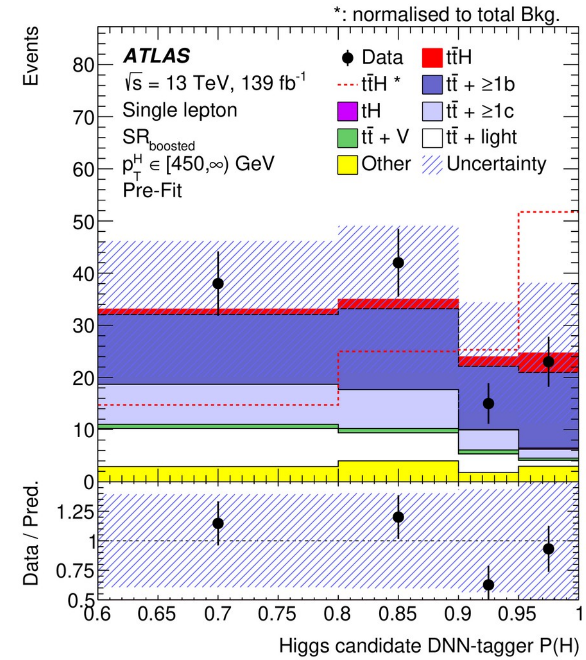
→ Example: **assume B is known,** use **measured n** to find out about **S.**

Multiple counting bins

Count in bins of a variable ⇒ *histogram* $n_1 \ldots n_N$.
(N : number of bins)

Per-bin fractions (=shapes) of Signal and Background

$$P(\{n_i\}; S, B) = \prod_{i=1}^{N} \underbrace{e^{-(S f_{S,i} + B f_{B,i})} \frac{(S f_{S,i} + B f_{B,i})^{n_i}}{n_i!}}_{\text{Poisson distribution in each bin}}$$



Shapes f typically obtained from simulated events (*Monte Carlo*)
→ HEP: typically excellent modeling from simulation, although some uncertainties need to be accounted for.

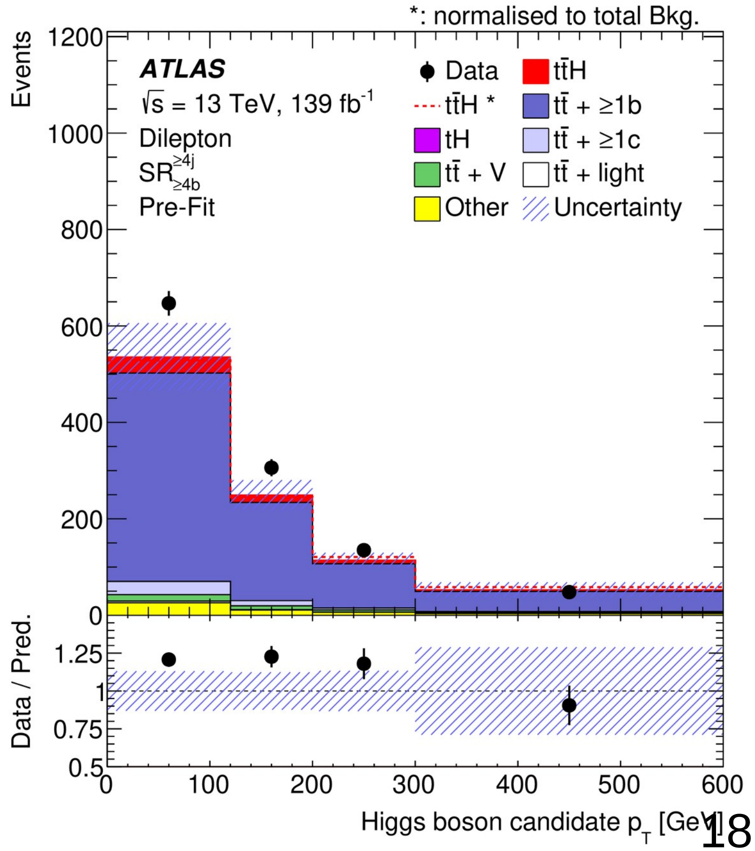
However not always possible to generate sufficiently large MC samples
MC stat fluctuations can create artefacts, especially for S ≪ B.

Model Parameters

Model typically includes:

- **Parameters of interest** (POIs) : what we want to measure
  → S, m_W, ...

- **Nuisance parameters** (NPs) : other parameters needed to define the model
  → Background levels (**B**)
  → For binned data, f^{sig}_i , f^{bkg}_i



NPs must be either:

→ **Known a priori** (within uncertainties) or

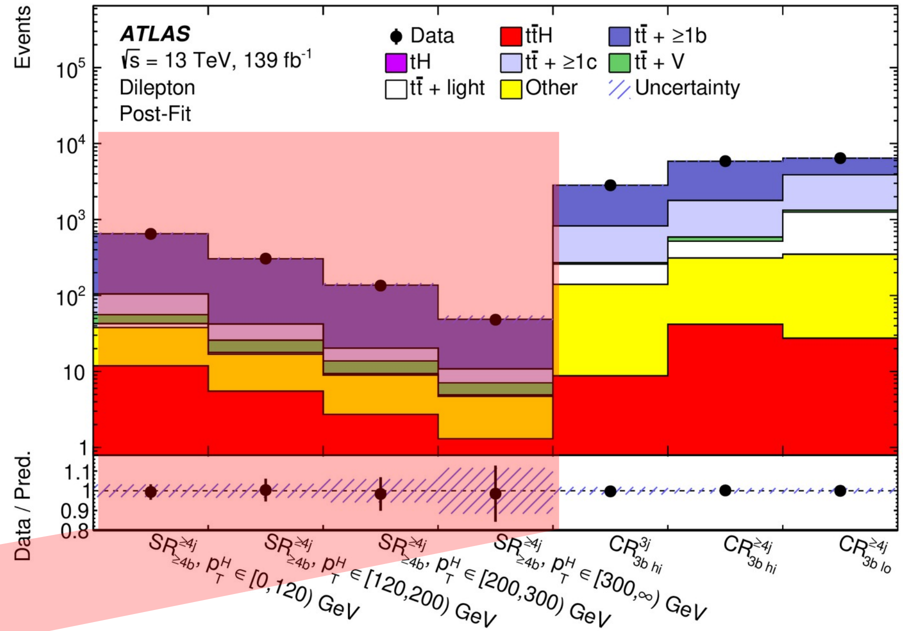
→ **Constrained by the data**

Categories

arXiv:2111.06712

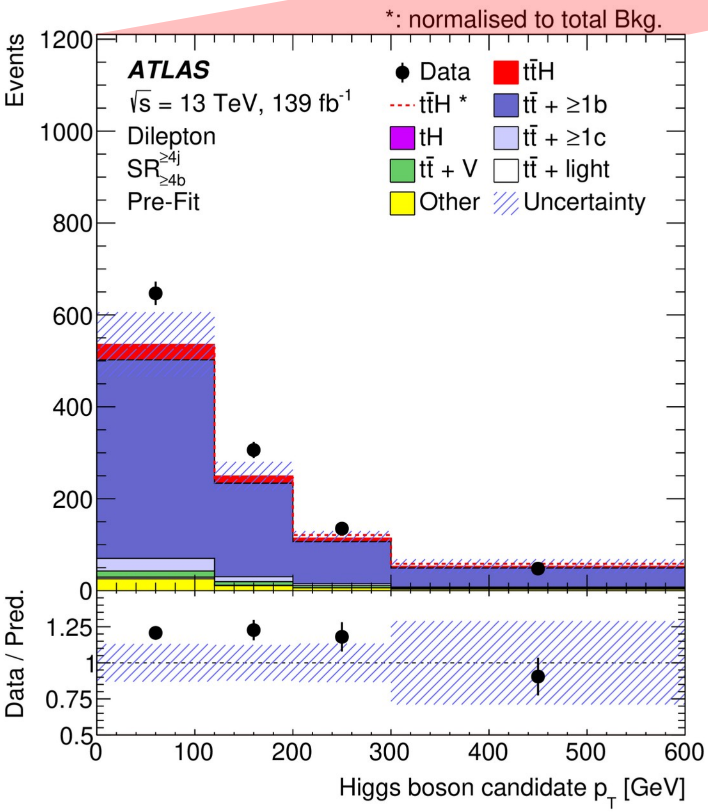
Multiple analysis regions often used.

→ Exploit better sensitivity in some regions



Here 7 regions:

→ 4 *Signal Regions* (**SR**) split in p_T(Higgs)



Better sensitivity at high p_T

→ lower B backgrounds, higher S/B

Backgrounds levels obtained from simulation here

→ Large uncertainties!

Categories

arXiv:2111.06712

Multiple analysis regions often used.

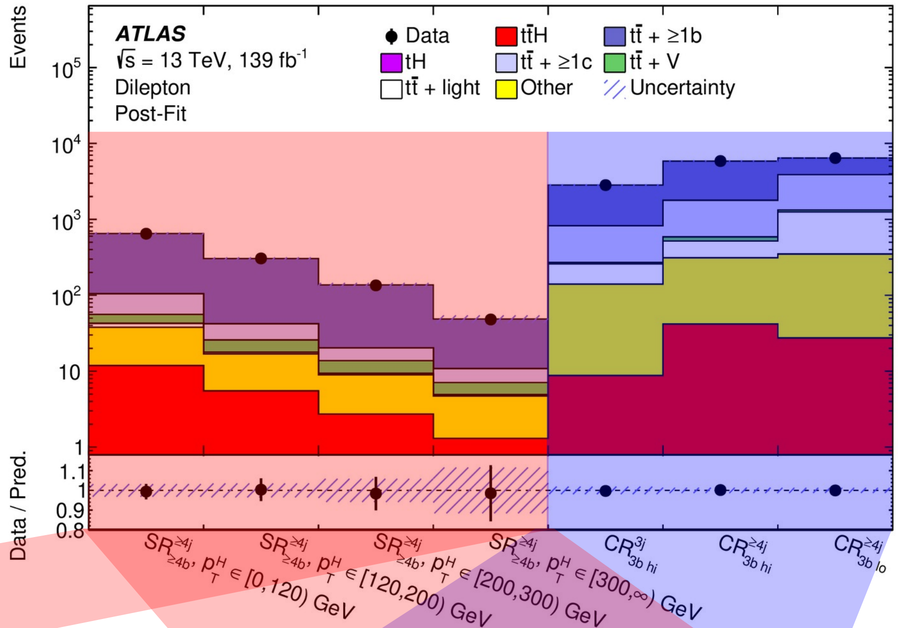
→ Exploit better sensitivity in some regions

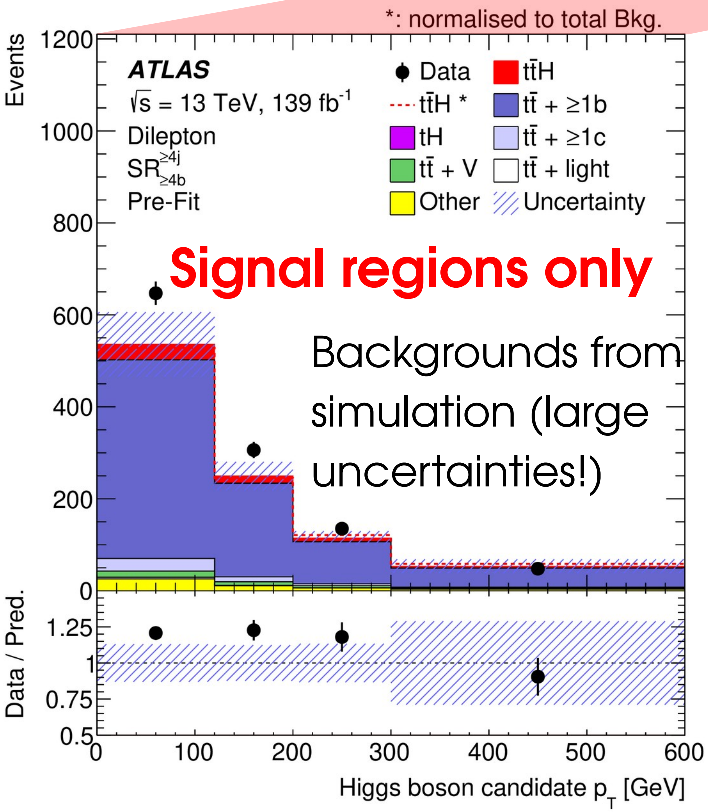
→ Constrain NPs: ***Control regions*** for bkgs

Here 7 regions:

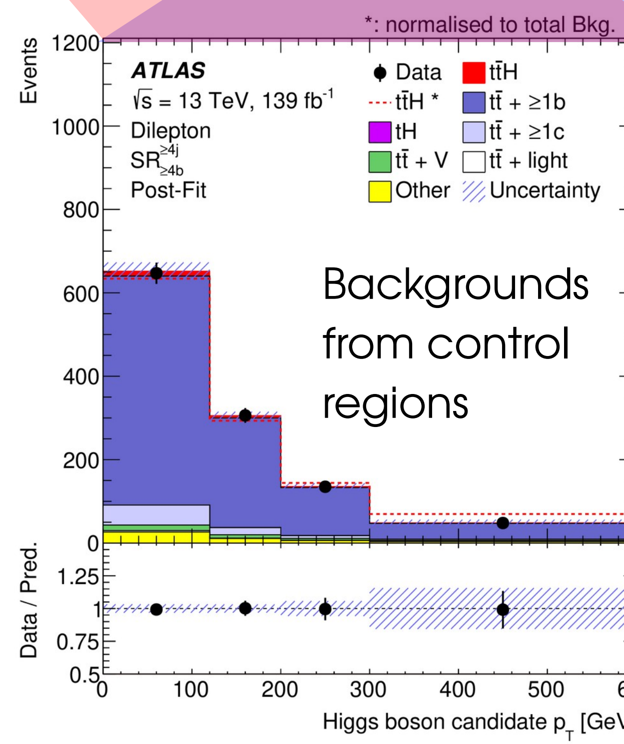
→ 4 *Signal Regions* (**SR**) split in p_T(Higgs)

→ 3 Background *Control Regions* (**CR**)





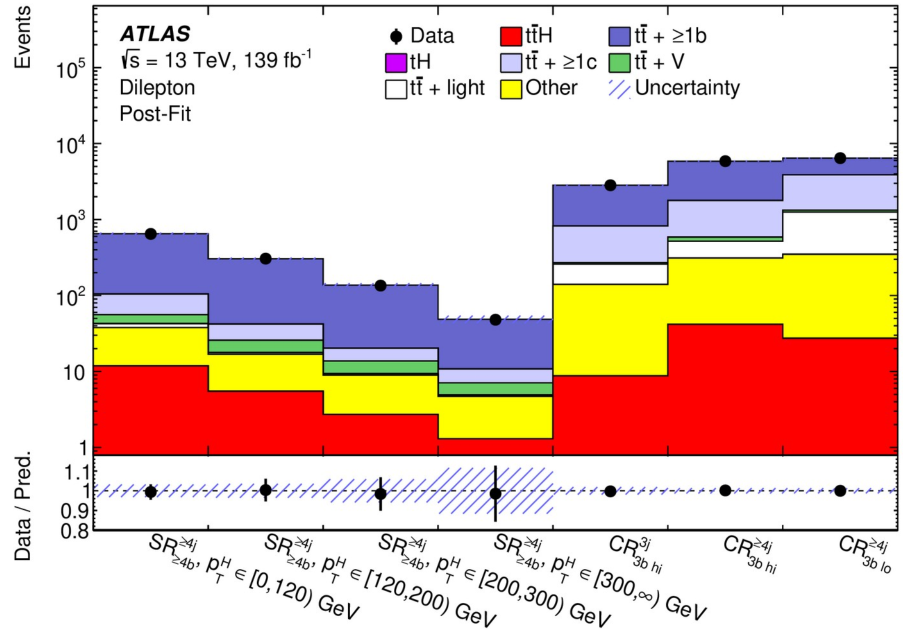
Include Background CRs



Signal + Bkg regions

Categories

arXiv:2111.06712



Multiple analysis regions often used.

→ Exploit better sensitivity in some regions

→ Constrain NPs: *Control regions* for bkgs

Here 7 regions:

→ 4 *Signal Regions* (**SR**) split in p_T(Higgs)

→ 3 Background *Control Regions* (**CR**)

⇒ ***Combined PDF*** :

PDF for category k

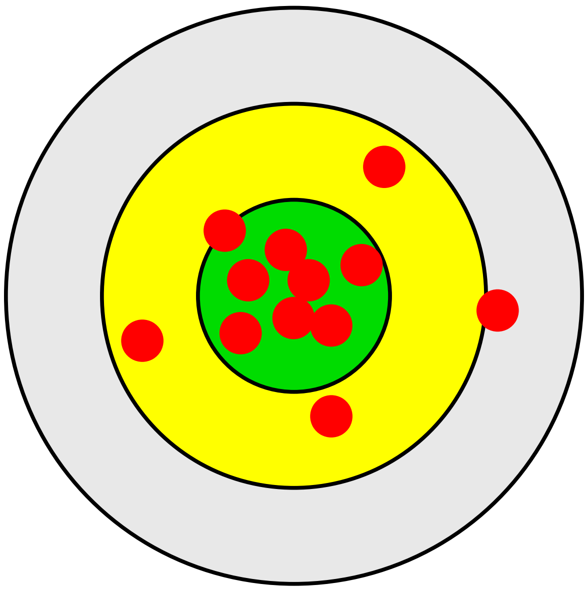
$$P\left(S;\{n_i^{(k)}\}_{i=1\ldots n_{\mathrm{evts}}^{(k)}}^{k=1\ldots n_{\mathrm{cats}}}\right) = \prod_{k=1}^{n_{\mathrm{cats}}} P_k\left(S;\{n_i^{(k)}\}_{i=1\ldots n_{\mathrm{evts}}^{(k)}}\right)$$

No overlaps between categories ⇒ No statistical correlations

⇒ can simply take product of individual PDFs.

Systematic Errors

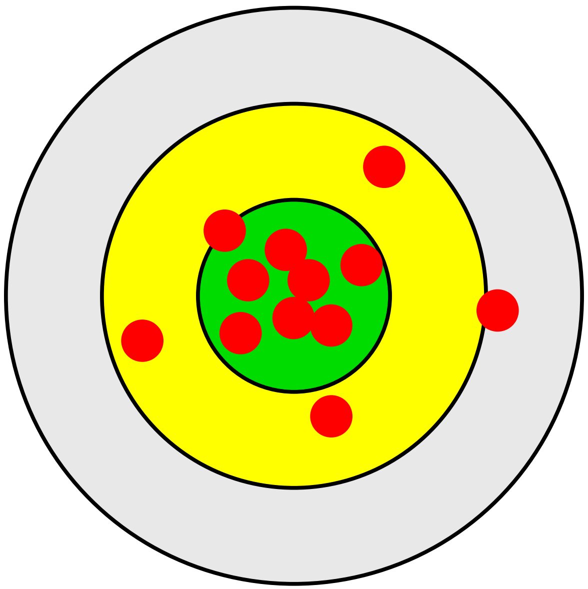
The statistical model (PDF) is a way to express **uncertainty** on the outcome of an experiment. e.g. 2D Gaussian :

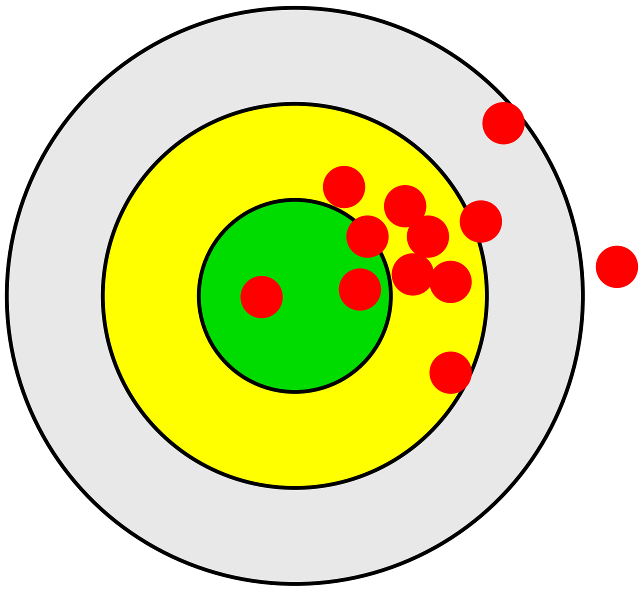


These uncertainties are also called **Statistical Uncertainties** – they are the ones encoded in the model PDF.

Systematic Errors

The statistical model (PDF) is a way to express **uncertainty** on the outcome of an experiment. e.g. 2D Gaussian :





These uncertainties are also called **Statistical Uncertainties** – they are the ones encoded in the model PDF.

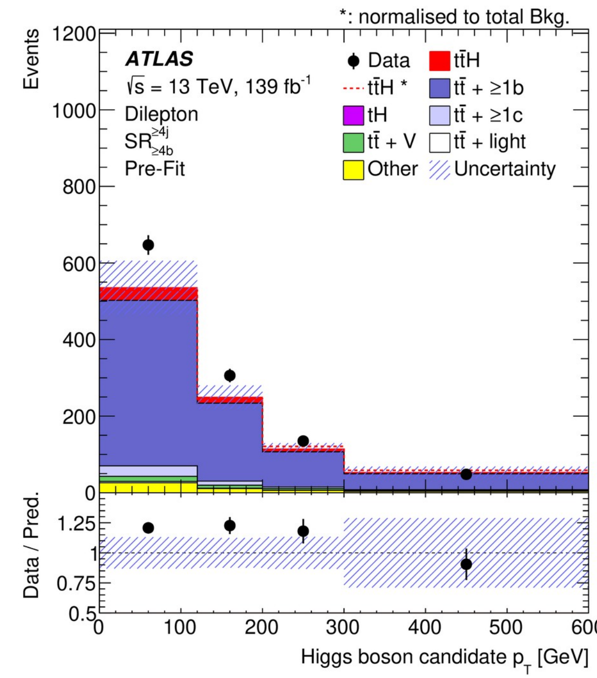
However **the model itself may be wrong** : this is a *systematic error*
→ To account for them, need a set of **Systematic uncertainties**

Systematics

Statistical models include:

- **Parameters of interest** (POIs) : **S**, **σ×B**, $\mathbf{m_W}$, ...
- **Nuisance parameters** (NPs) : other parameters needed to define the model

  → Ideally, **constrained by data** like the POI



And systematics ?

= Cover what we don't know about the random process.

⇒ **Parameterize using additional NPs**

→ Can't be constrained by the data ⇒ Add constraints in the likelihood

$$L(\mu,\theta;\mathbf{data}) = L_{\mathbf{measurement}}(\mu,\theta;\mathbf{data})\ C(\theta)$$

POI **Systematics NP** **Measurement Likelihood** **NP Constraint term**

"Systematic uncertainty is, in any statistical inference procedure, the uncertainty due to the incomplete knowledge of the probability distribution of the observables.
G. Punzi, ***What is systematics ?***

C(**θ**) represents **external knowledge** about the NP

Frequentist Systematics

Prototype: Systematics NP → measured in a separate ***auxiliary* experiment**
e.g. background levels

→ Build the combined PDF of the main+auxiliary measurements

$$P(\mu, \theta; \text{data}) = P_{\text{main}}(\mu, \theta; \text{main data})\ P_{\text{aux}}(\theta; \text{aux. data})$$

Independent measurements:
⇒ just a product

Gaussian form often used by default: $P_{\text{aux}}(\theta; \text{aux. data}) = G(\theta^{\text{obs}}; \theta, \sigma_{\text{syst}})$

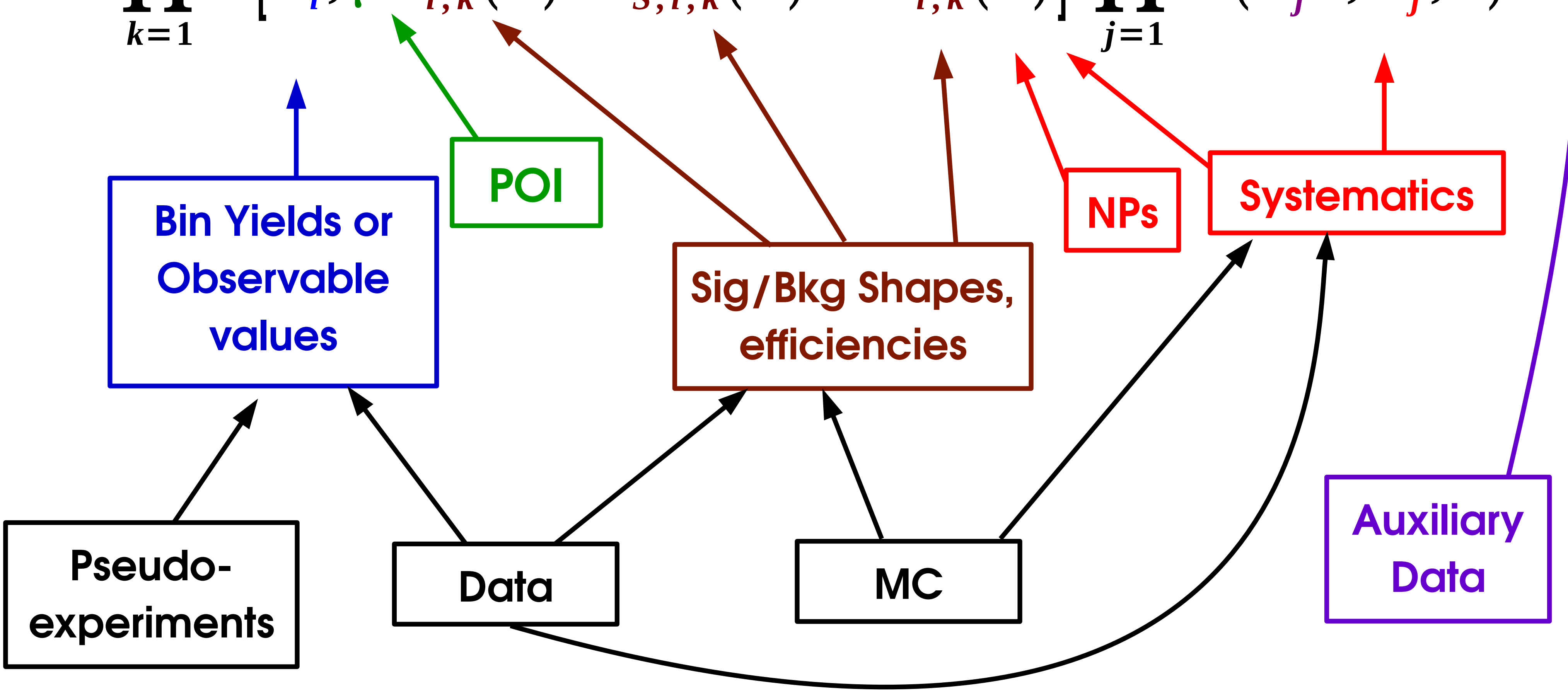
In the combined likelihood, **systematic NPs are constrained**
→ now same as NPs constrained in data.

→ Often no clear setup for auxiliary measurements
(e.g. theory simulation uncertainties)
→ Define constraints "by hand" ("pseudo-measurement")

Statistical model, the full version

$$P\left(\mu,\{\theta_j\}_{j=1\ldots n_{NP}};\{n_i^{(k)}\}_{i=1\ldots n_{data}^{(k)}}^{k=1\ldots n_{cat}},\{\theta_j^{obs}\}_{j=1..n_{NP}}\right)=$$

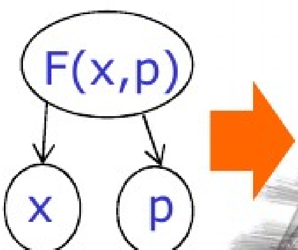
$$\prod_{k=1}^{n_{cats}} P\left[n_i;\overbrace{\mu\,\epsilon_{i,k}(\vec{\theta})\;N_{S,i,k}(\vec{\theta})+B_{i,k}(\vec{\theta})}^{\text{Expected bin yield}}\right]\prod_{j=1}^{n_{syst}} G(\theta_j^{obs};\theta_j;1)$$

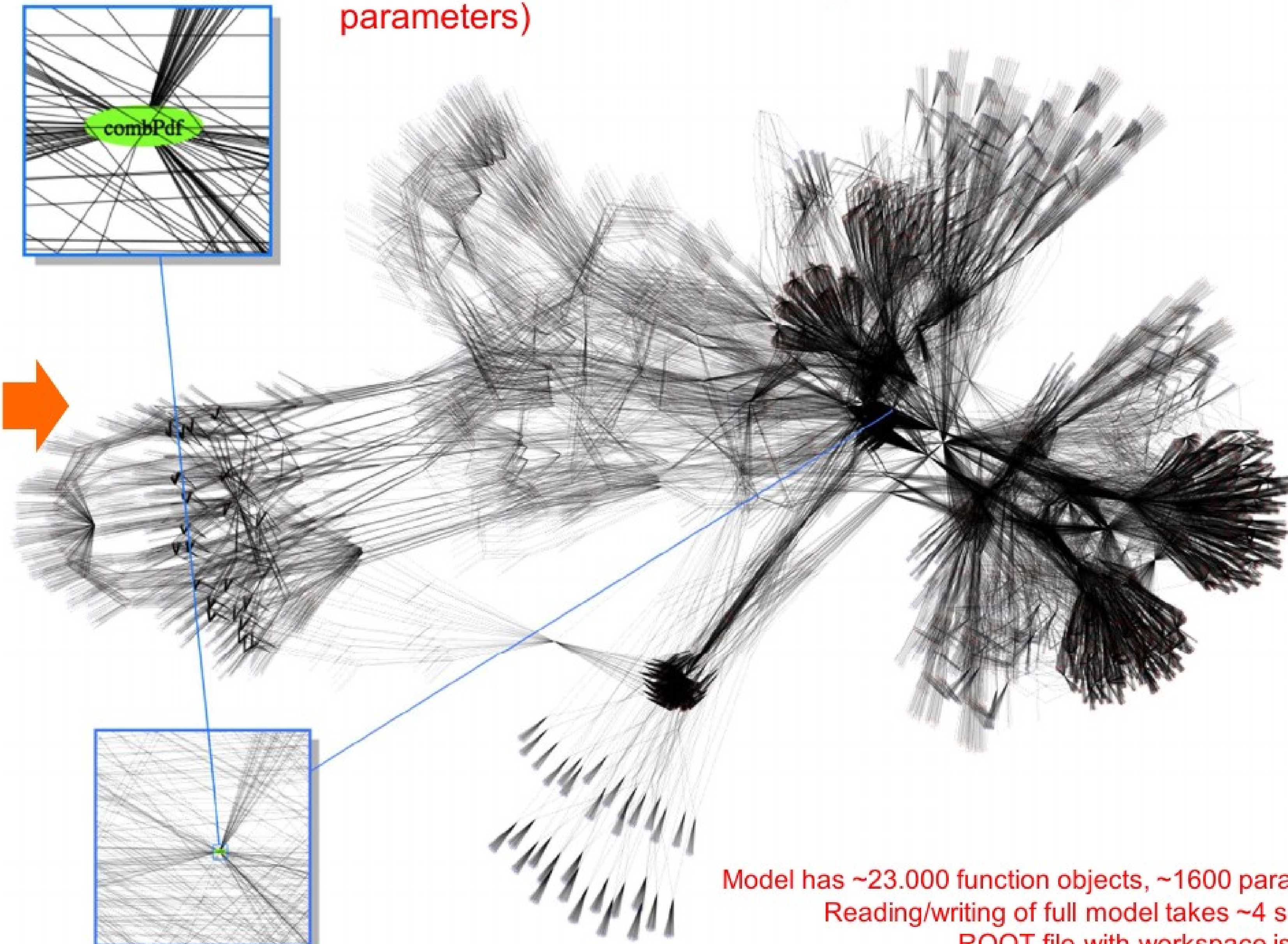


× number of categories!

ATLAS Higgs Run 1 Combination Model

Atlas Higgs combination model (23.000 functions, 1600 parameters)





Model has ~23.000 function objects, ~1600 parameters
Reading/writing of full model takes ~4 seconds
ROOT file with workspace is ~6 Mb

HEP Statistical Inference : Confidence Intervals

Using the PDF

Model describes the distribution of the observable: **P(data; parameters)**

$$P(S=5)$$

provide parameter values

Generate

2, 5, 3, 7, 4, 9, ...

Each entry = a separate pseudo-data "experiment"

We want the **other** direction: **use data to get information on parameters**

$$P(S=?)$$

Estimate

2

THE observed data

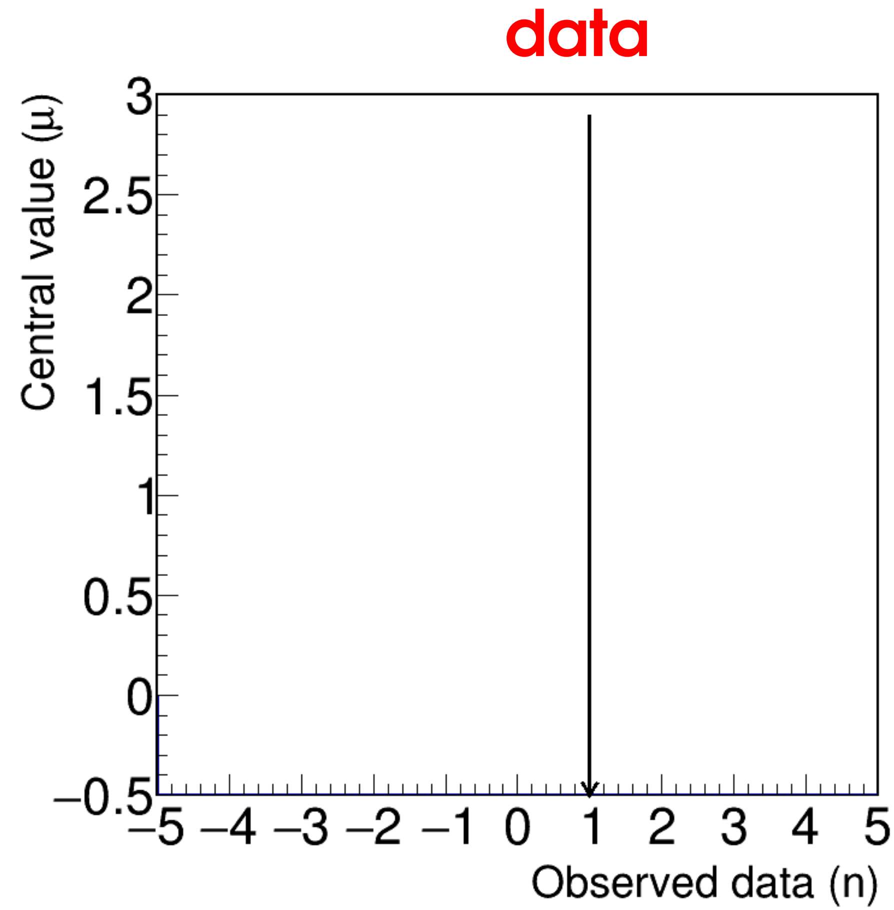
Maximum Likelihood

Define likelihood L(μ) = P(**data**; **μ**)
⇒ Implicitly a function of the data

Estimate μ as

$$\hat{\mu} = \mathrm{argmax}_{\mu}\, L(\mu)$$

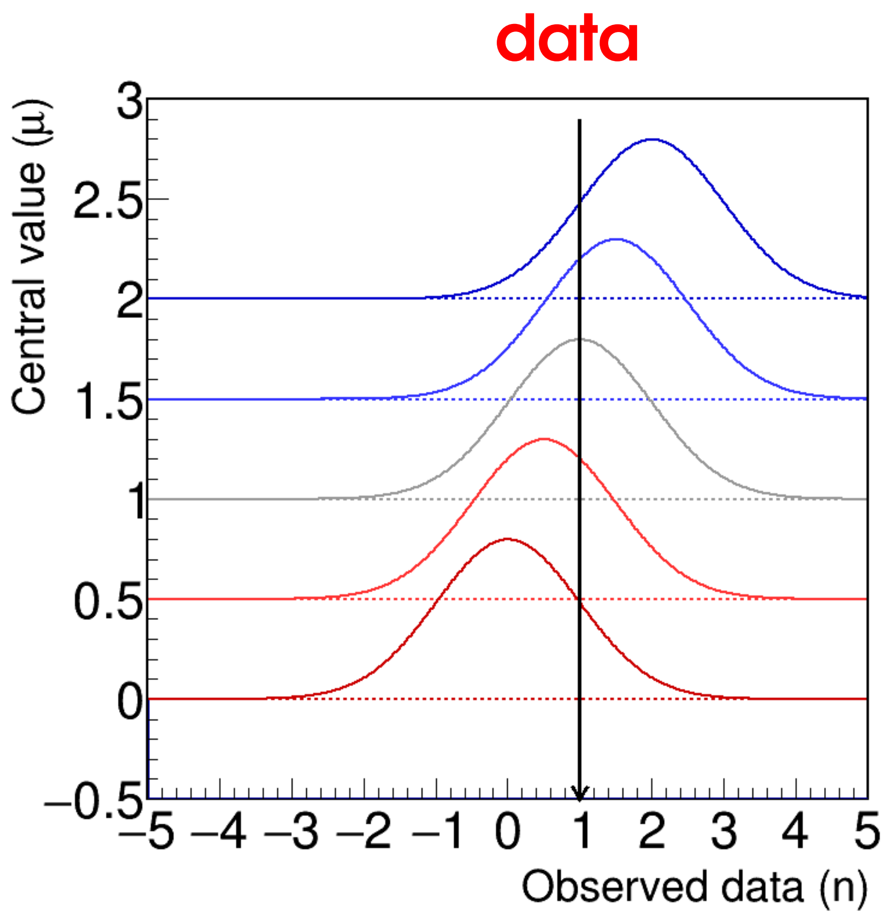
"Best fit" of model to data



Several good properties:

- Asymptotically **Gaussian** $\quad P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right) \quad$ for $n \rightarrow \infty$
- Asymptotically **Unbiased**
- Asymptotically **Efficient:** $\sigma_{\hat{\mu}}$ is the lowest possible
- Always **consistent** $\quad \hat{\mu} \xrightarrow{n \rightarrow \infty} \mu^*$

Maximum Likelihood



Define likelihood L(**μ**) = P(**data**; **μ**)

⇒ Implicitly a function of the data

Estimate μ as

$$\hat{\mu} = \mathrm{argmax}_{\mu}\, L(\mu)$$

"Best fit" of model to data

Several good properties:

- Asymptotically **Gaussian** $$P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right) \quad \text{for } n \to \infty$$
- Asymptotically **Unbiased**
- Asymptotically **Efficient:** $\sigma_{\hat{\mu}}$ is the lowest possible
- Always **consistent** $\hat{\mu} \xrightarrow{n \to \infty} \mu^*$

Maximum Likelihood

Define likelihood L(μ) = P(data; μ)

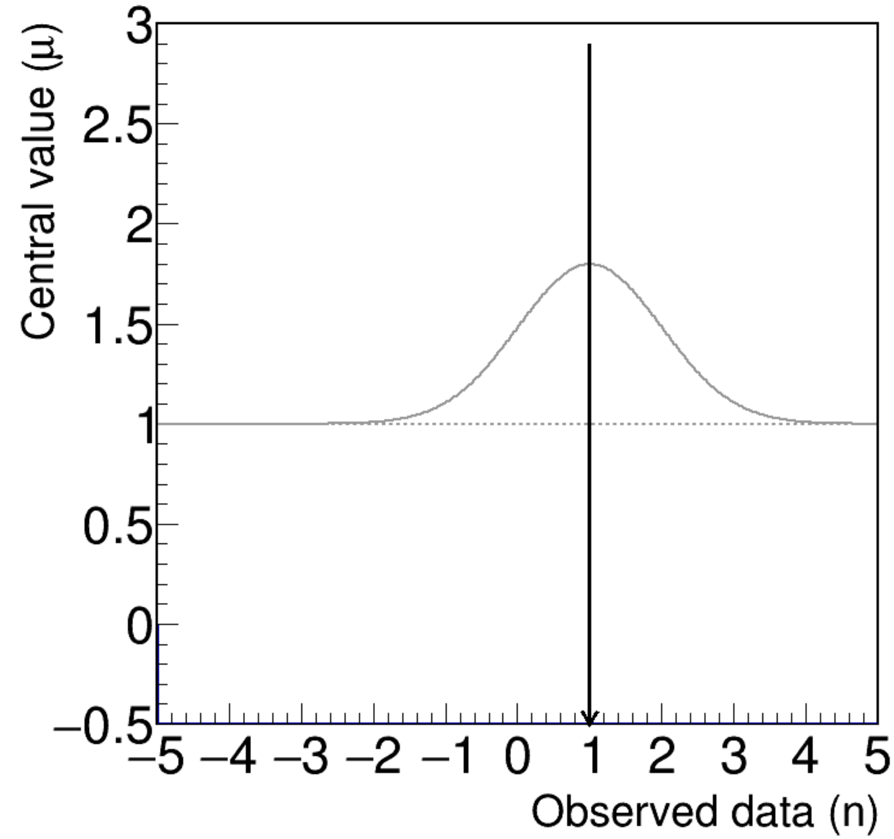
⇒ Implicitly a function of the data

Estimate μ as

$$\hat{\mu} = argmax_{\mu}\, L(\mu)$$

"Best fit" of model to data

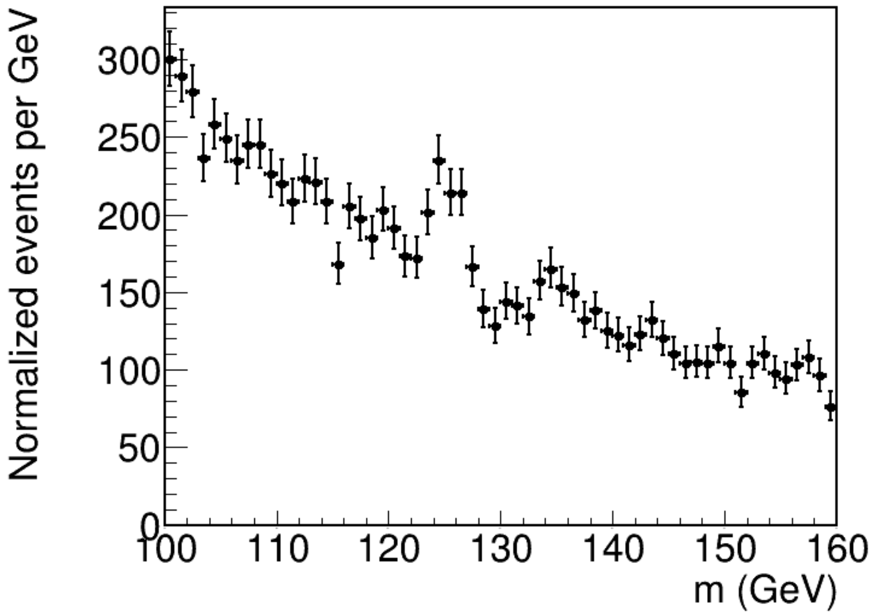
data



Several good properties:

- Asymptotically **Gaussian** $P(\hat{\mu}) \propto \exp\left(-\frac{(\hat{\mu}-\mu^*)^2}{2\sigma_{\hat{\mu}}^2}\right)$ for $n \to \infty$
- Asymptotically **Unbiased**
- Asymptotically **Efficient**: $\sigma_{\hat{\mu}}$ is the lowest possible
- Always **consistent** $\hat{\mu} \xrightarrow{n \to \infty} \mu^*$

Maximum Likelihood



Multiple Gaussian bins:

$$\lambda(\mu) = -2 \log L(\mu) = \sum_{i=1}^{N_{\mathrm{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood

⇔ Minimum χ^2

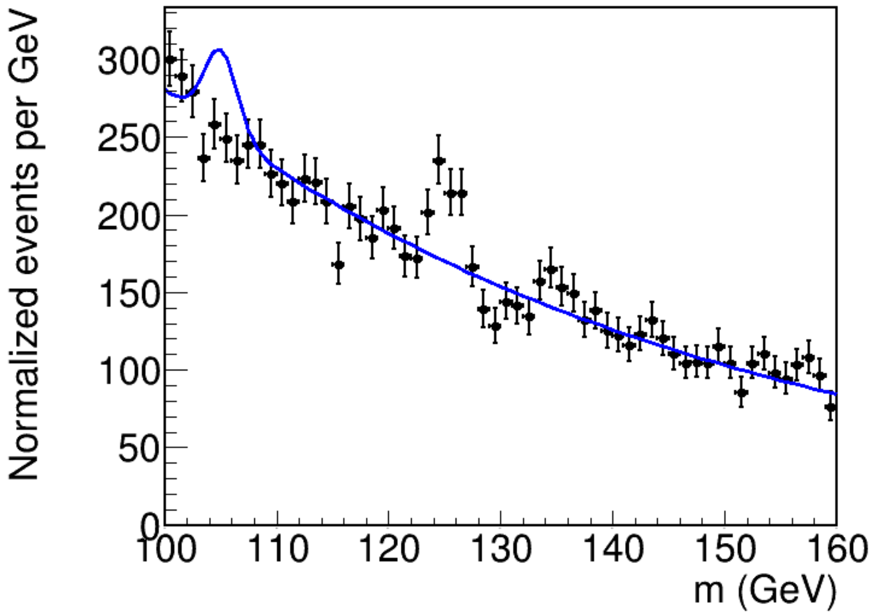
⇔ Least-squares minimization

However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends
  → Usual methods – gradient-based, etc.

Maximum Likelihood



Multiple Gaussian bins:

$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i}\right)^2$$

Maximum likelihood

⇔ Minimum χ^2

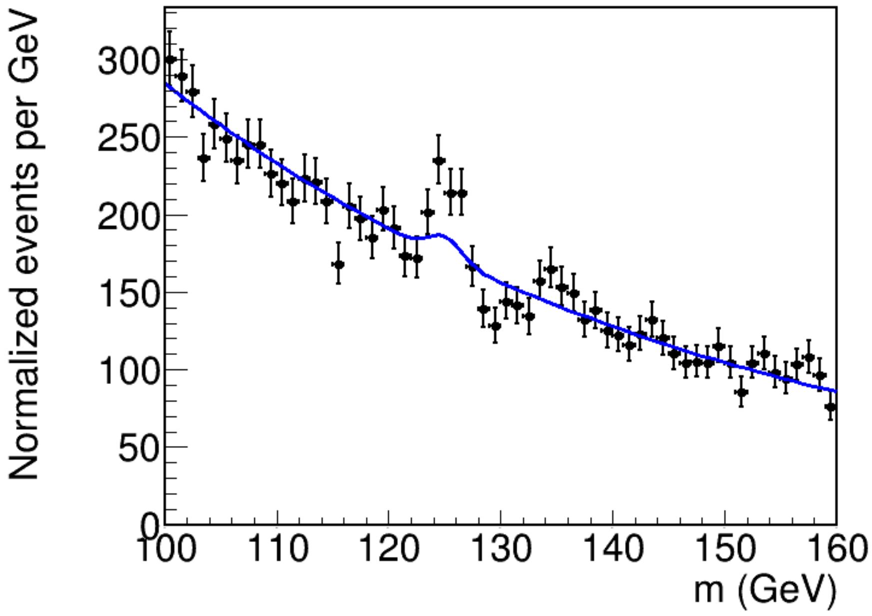
⇔ Least-squares minimization

However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends
  → Usual methods – gradient-based, etc.

Maximum Likelihood



Multiple Gaussian bins:

$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\mathrm{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i} \right)^2$$

Maximum likelihood

⇔ Minimum χ^2

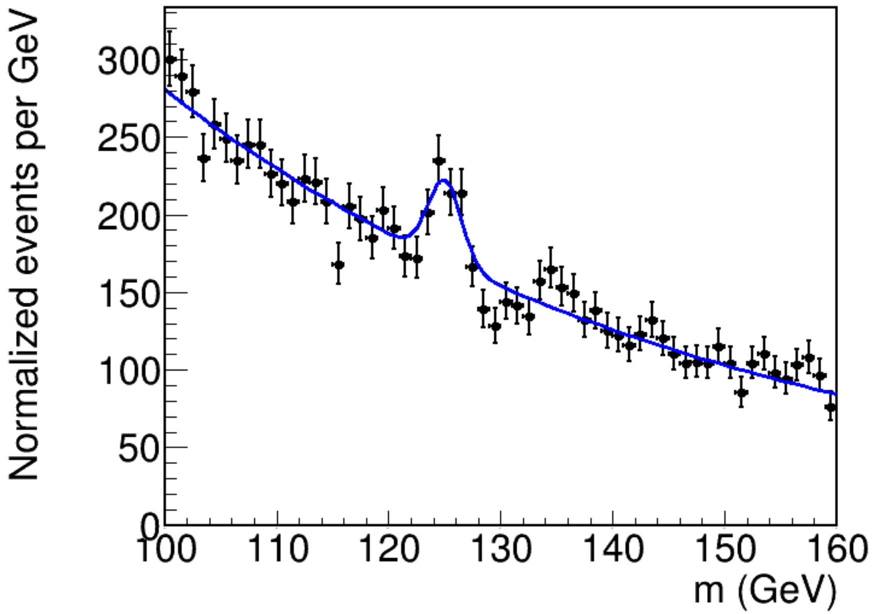
⇔ Least-squares minimization

However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends
  → Usual methods – gradient-based, etc.

Maximum Likelihood



Multiple Gaussian bins:

$$\lambda(\mu) = -2\log L(\mu) = \sum_{i=1}^{N_{\text{bins}}} \left(\frac{n_i - \mu_i}{\sigma_i}\right)^2$$

Maximum likelihood

⇔ Minimum χ^2

⇔ Least-squares minimization

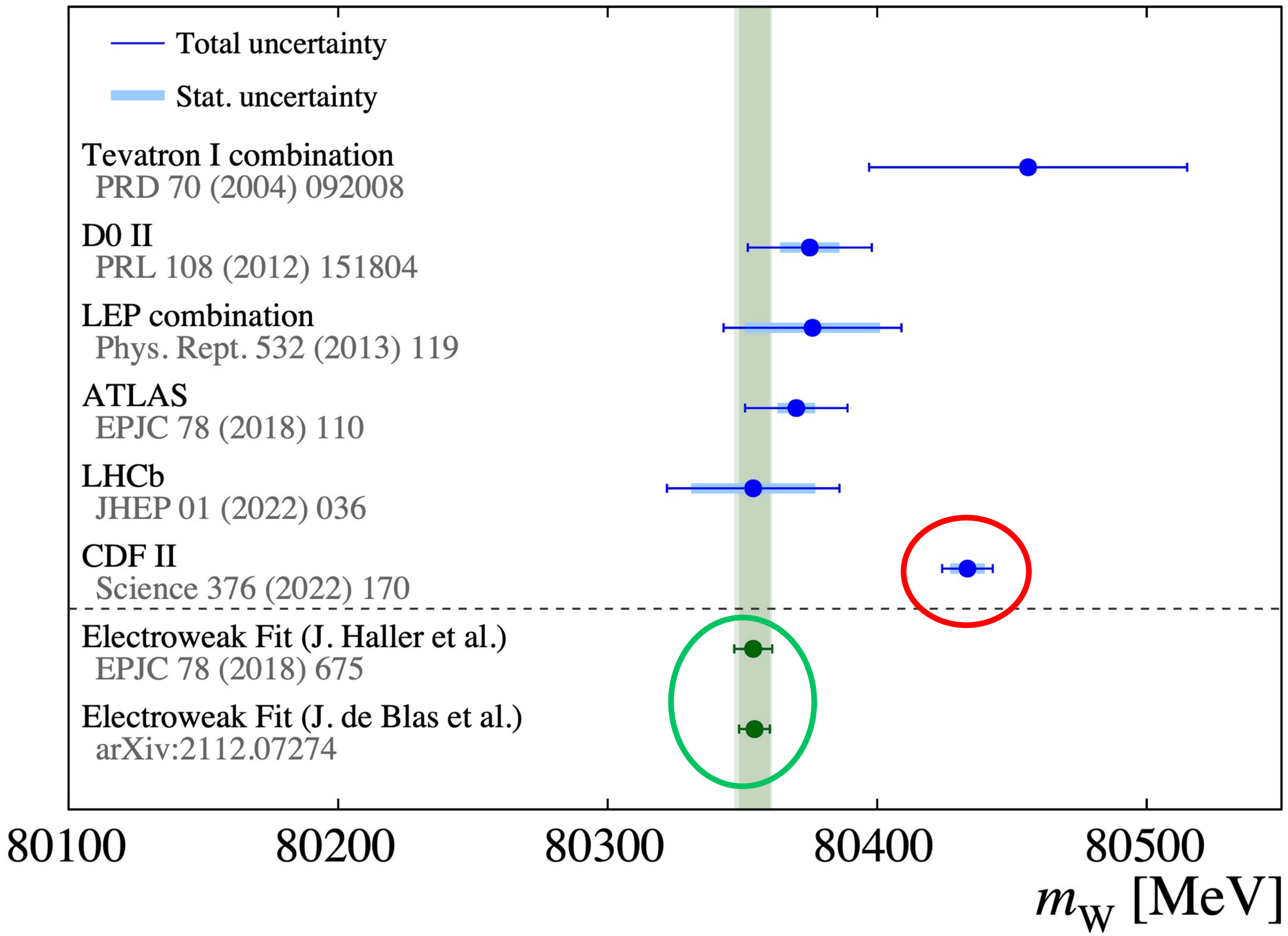
However typically need to perform non-linear minimization.

HEP practice:

- **MINUIT** (C++ library within ROOT, numerical gradient descent)
- **scipy.minimize** – using NumPy/TensorFlow/PyTorch/... backends
  → Usual methods – gradient-based, etc.

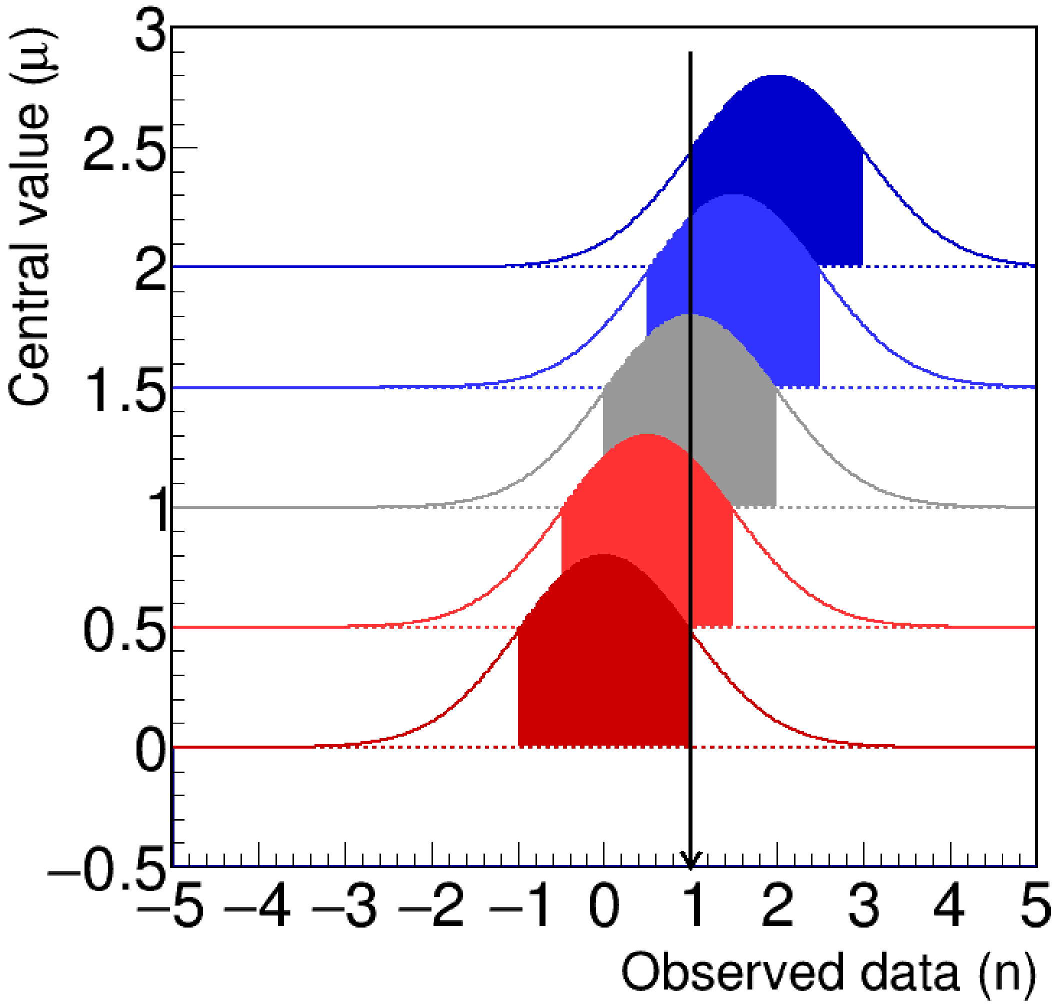
Uncertainties

Science 376, 170–176 (2022), 8



$$\boxed{M_W = 80,433.5 \pm 6.4_{\rm stat} \pm 6.9_{\rm syst} = 80,433.5 \pm 9.4\ \mathrm{MeV}/c^2}$$

Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

$$\Updownarrow$$

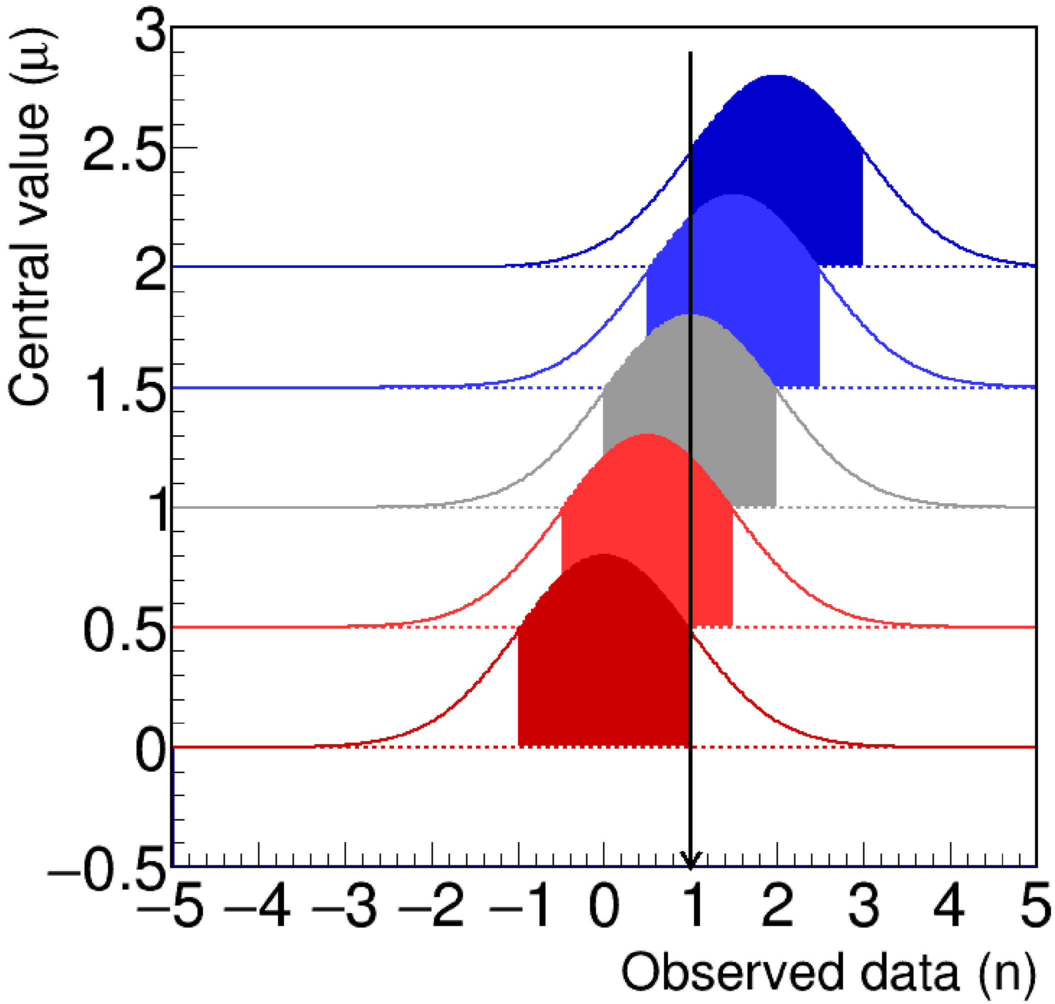
$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

$$\boxed{\mu = n \pm \sigma \quad \text{at} \quad 68.3\,\%\ \text{CL}(1\sigma)}$$

This interval will contain the true μ value 68.3% of the time ("1σ")

Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

$$\Updownarrow$$

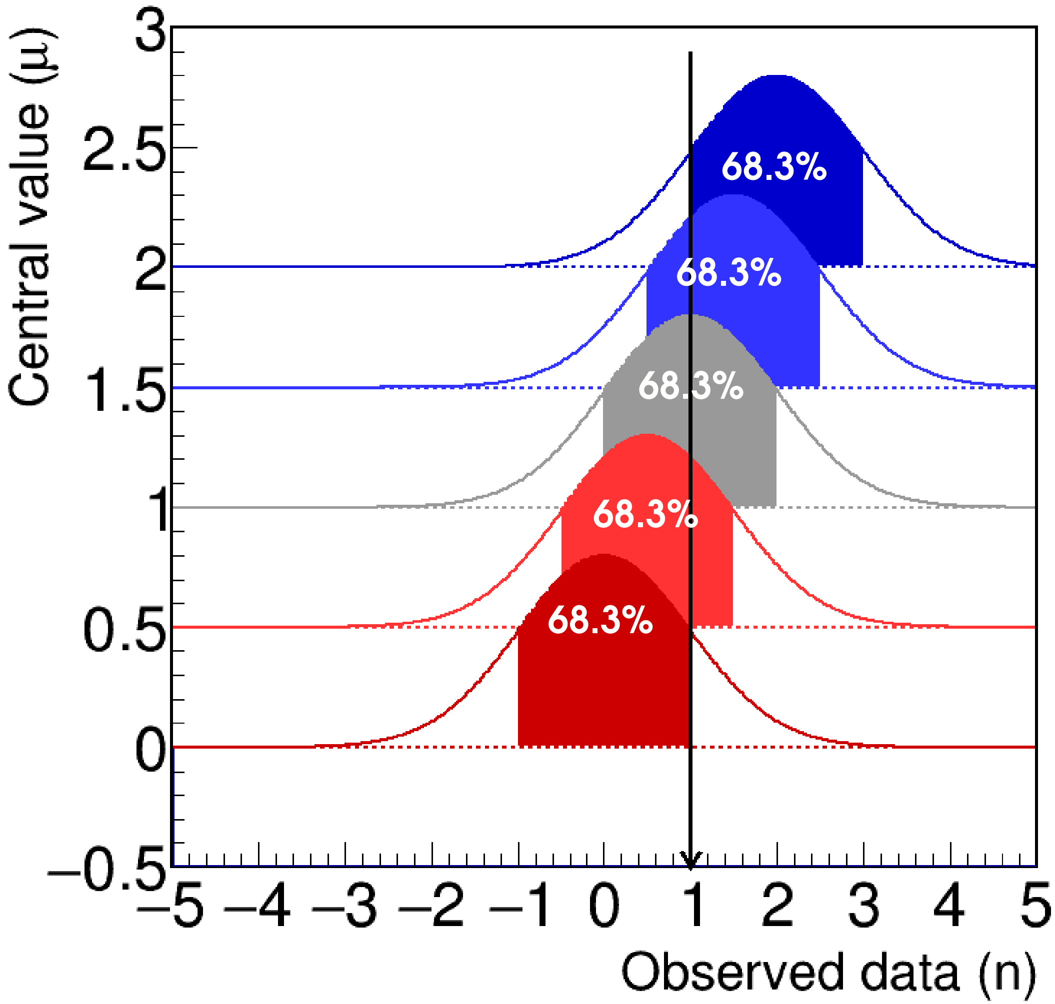
$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

$$\boxed{\mu = n \pm \sigma \quad \text{at} \quad 68.3\,\% \;\; \text{CL}(1\sigma)}$$

This interval will contain the true μ value 68.3% of the time ("1σ")

Gaussian confidence intervals



Consider a Gaussian likelihood:

$$L(\mu) = \exp\left[-\frac{1}{2}\left(\frac{n-\mu}{\sigma}\right)^2\right]$$

$$P(\mu - \sigma < n < \mu + \sigma) \geq 68.3\,\%$$

$$\Updownarrow$$

$$P(n - \sigma < \mu < n + \sigma) \geq 68.3\,\%$$

Still a statement on n!

$$\boxed{\mu = n \pm \sigma \quad \text{at} \quad 68.3\,\% \ \text{CL}(1\sigma)}$$

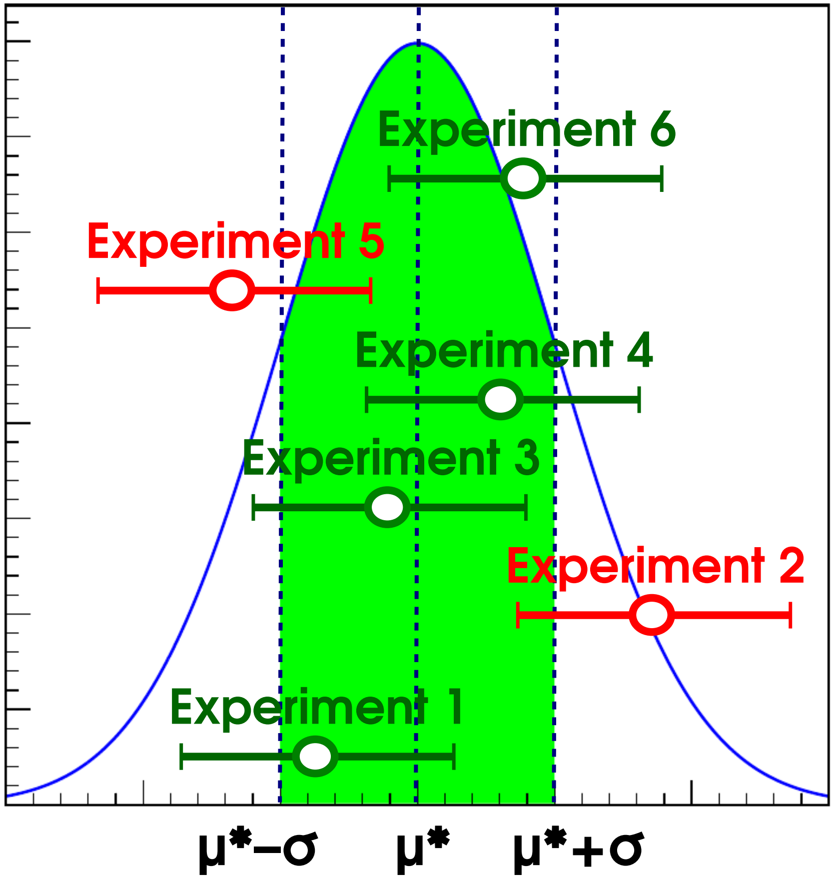
This interval will contain the true μ value 68.3% of the time (“1σ”)

Gaussian confidence intervals

Frequentist interpretation

If we would repeat the same experiment multiple times, with true value μ*, then 68.3% of the 1σ intervals would contain μ*.

→ Crucially, this works even if we do not know μ* !



For each experiment, get the interval

$$\boxed{\mu = n \pm \sigma \quad \text{at} \quad 68.3\,\% \ \ \text{CL}(1\,\sigma)}$$

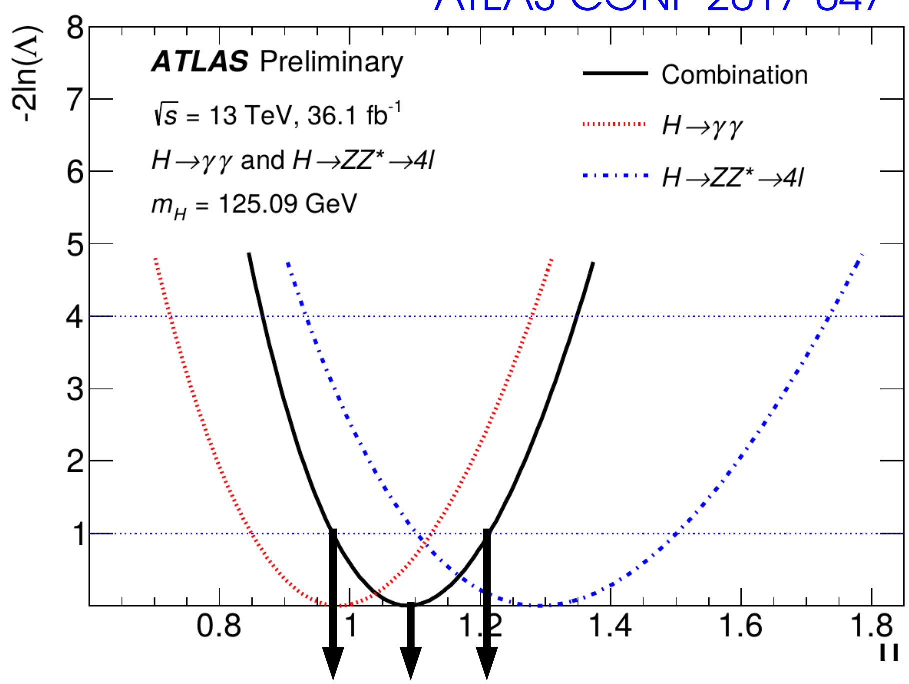
This interval will contain the true μ value 68.3% of the time ("1σ")

General case: Likelihood Intervals

Confidence intervals from L:

- Test various values μ using the **Profile Likelihood Ratio t(μ)**
- Minimum (=0) for $\mu=\hat{\mu}$
- **Likelihood ratio** universally most powerful test for simple hypotheses (no NPs, single POI values), also used in other cases

ATLAS-CONF-2017-047



$$\mu = 1.09 \pm 0.12$$

Probability to observe the data for a given μ. Use ***conditional best-fit*** $\boldsymbol{\theta(\mu)}$ of the NPs for this μ.

$$t(\mu) = -2\log\frac{L(\mu,\hat{\hat{\theta}}(\mu))}{L(\hat{\mu},\hat{\theta})}$$

Probability to observe the data for $\hat{\mu}$. Use ***best-fit θ*** for the NPs.

Gaussian L(μ):

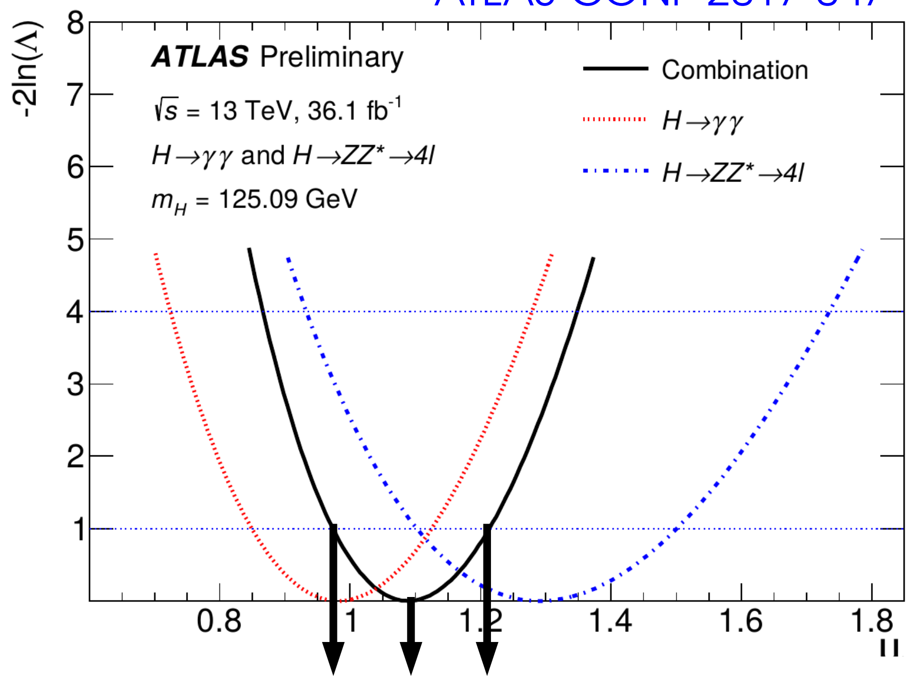
- Parabolic in μ
- Minimum occ=urs at $\boldsymbol{\mu = \hat{\mu}}$
- t(μ) distributed as a χ^2
- 1σ interval $[\mu_-,\mu_+]$ given by $\boldsymbol{t(\mu_\pm) = 1}$

General case: Likelihood Intervals

Confidence intervals from L:

- Test various values μ using the **Profile Likelihood Ratio t(μ)**
- Minimum (=0) for $\mu=\hat{\mu}$
- **Likelihood ratio** universally most powerful test for simple hypotheses (no NPs, single POI values), also used in other cases

ATLAS-CONF-2017-047



$\mu = 1.09 \pm 0.12$

$$t(\mu) = -2\log\frac{L(\mu, \hat{\hat{\theta}}(\mu))}{L(\hat{\mu}, \hat{\theta})}$$

General case:

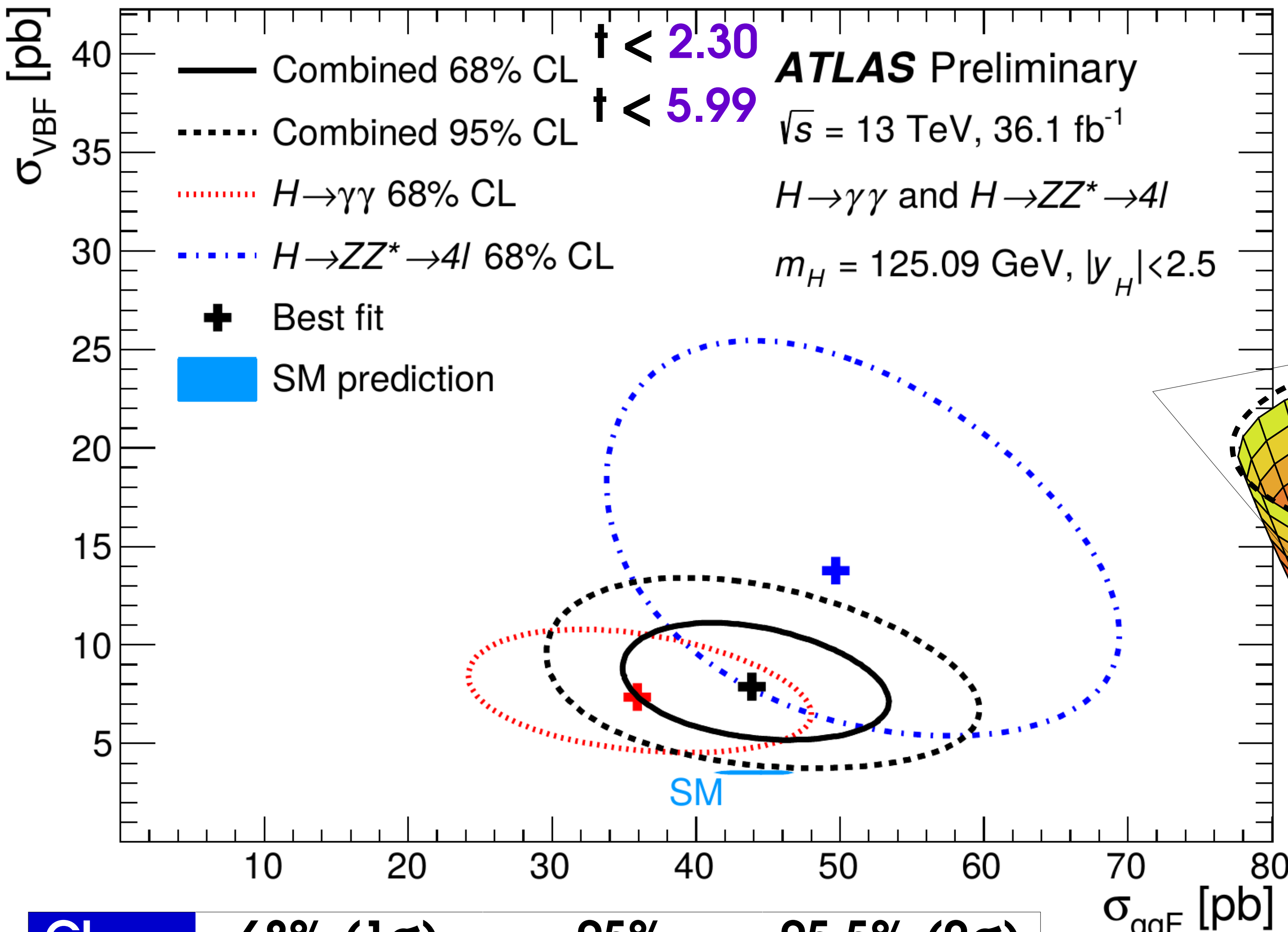
- Generally not a perfect parabola
- Minimum still at $\mu = \hat{\mu}$
- Distribution of t(μ) ?

Asymptotic approximation

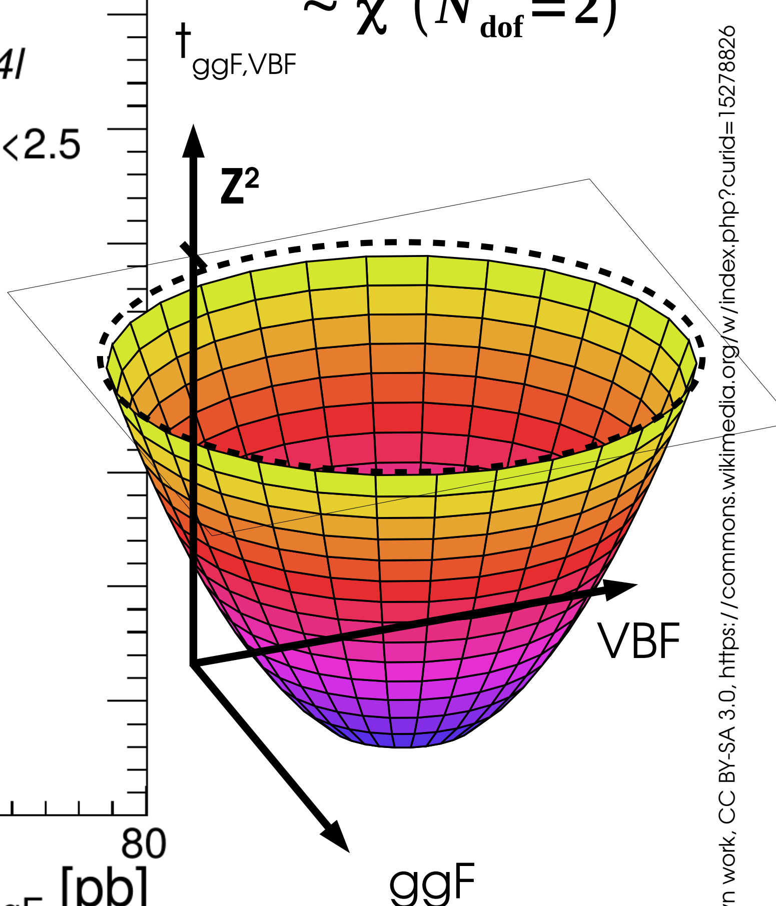
- Compute t(μ) using the exact L(μ)
- Assume $t(\mu) \sim \chi^2$ as for Gaussian (*"Wilks' Theorem"*)
- 1σ interval $[\mu_-,\mu_+]$ given by $\mathbf{t(\mu_\pm) = 1}$
- Can also obtain exact intervals using pseudo-dataset sampling ("toys"), but generally not needed and rarely done.

2D Example: Higgs σ_{VBF} vs. σ_{ggF}

ATLAS-CONF-2017-047



$$t = -2\log\frac{L(X_{0}, Y_{0})}{L(\hat{X}, \hat{Y})} \sim \chi^2(N_{dof}=2)$$



CL	68% (1σ)	95%	95.5% (2σ)
1D Z²	1	3.84	4
2D Z²	2.30	5.99	6.18

Gaussian case: elliptic paraboloid surface

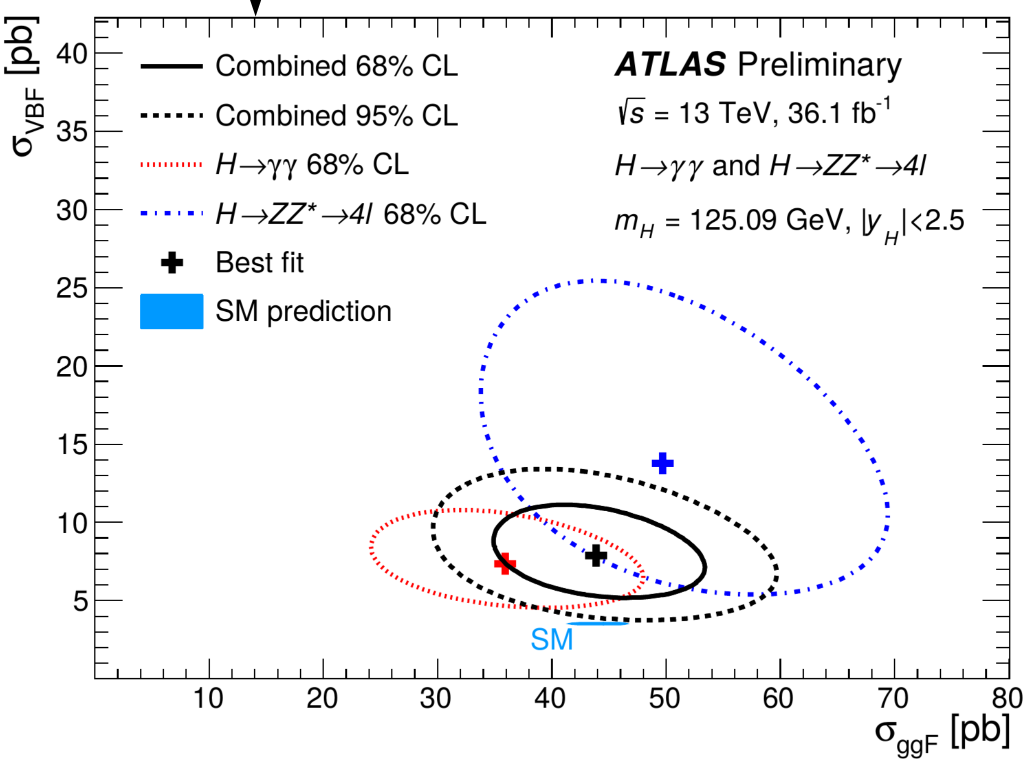
Reparameterization

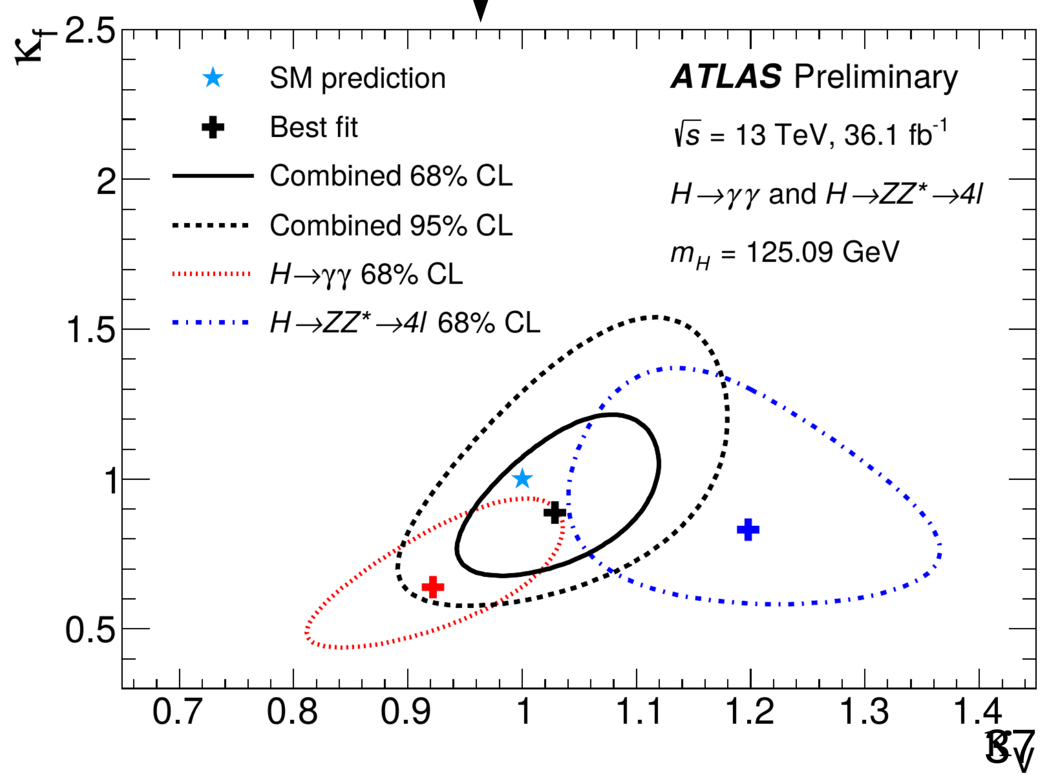
Start with basic measurement in terms of e.g. **σ×B**

→ How to measure derived quantities (couplings, parameters in some theory model, etc.) ? **→ just reparameterize the likelihood:**

e.g. Higgs couplings: σ_{ggF}, σ_{VBF} sensitive to Higgs coupling modifiers κ_V, κ_F.

$$L(\sigma_{ggF}, \sigma_{VBF}) \xrightarrow[\sigma_{VBF} \to \sigma_{VBF}(\kappa_V, \kappa_F)]{\sigma_{ggF} \to \sigma_{ggF}(\kappa_V, \kappa_F)} L(\sigma_{ggF}(\kappa_V, \kappa_F), \sigma_{VBF}(\kappa_V, \kappa_F)) \equiv L'(\kappa_V, \kappa_F)$$





Example: Gaussian Profiling

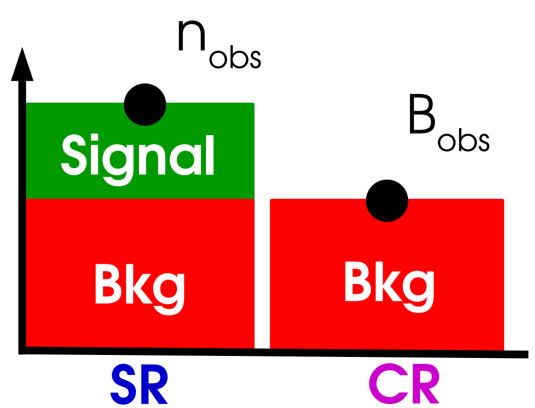
Counting experiment with background uncertainty: **n = S + B** :

→ **Signal region (SR)**: $n_{obs} \sim G(S + B, \sigma_{stat})$

→ **Control region (CR)**: $B_{obs} \sim G(B, \sigma_{bkg})$

$$L(S, B) = G(n_{obs}; S + B, \sigma_{stat})\, G(B_{obs}; B, \sigma_{bkg})$$

Recall: Signal region only (fixed B):
$$t_S = \left(\frac{S - n_{obs}}{\sigma_{stat}}\right)^2 \qquad S = (n_{obs} - B) \pm \sigma_{stat}$$



→ Compute the best-fit (MLEs) for S and B

→ Show that the conditional MLE for B is

$$\hat{\hat{B}}(S) = B_{obs} + \frac{\sigma_{bkg}^2}{\sigma_{stat}^2 + \sigma_{bkg}^2}(\hat{S} - S)$$

→ Compute the profile likelihood t_S

→ Compute the 1σ confidence interval on S

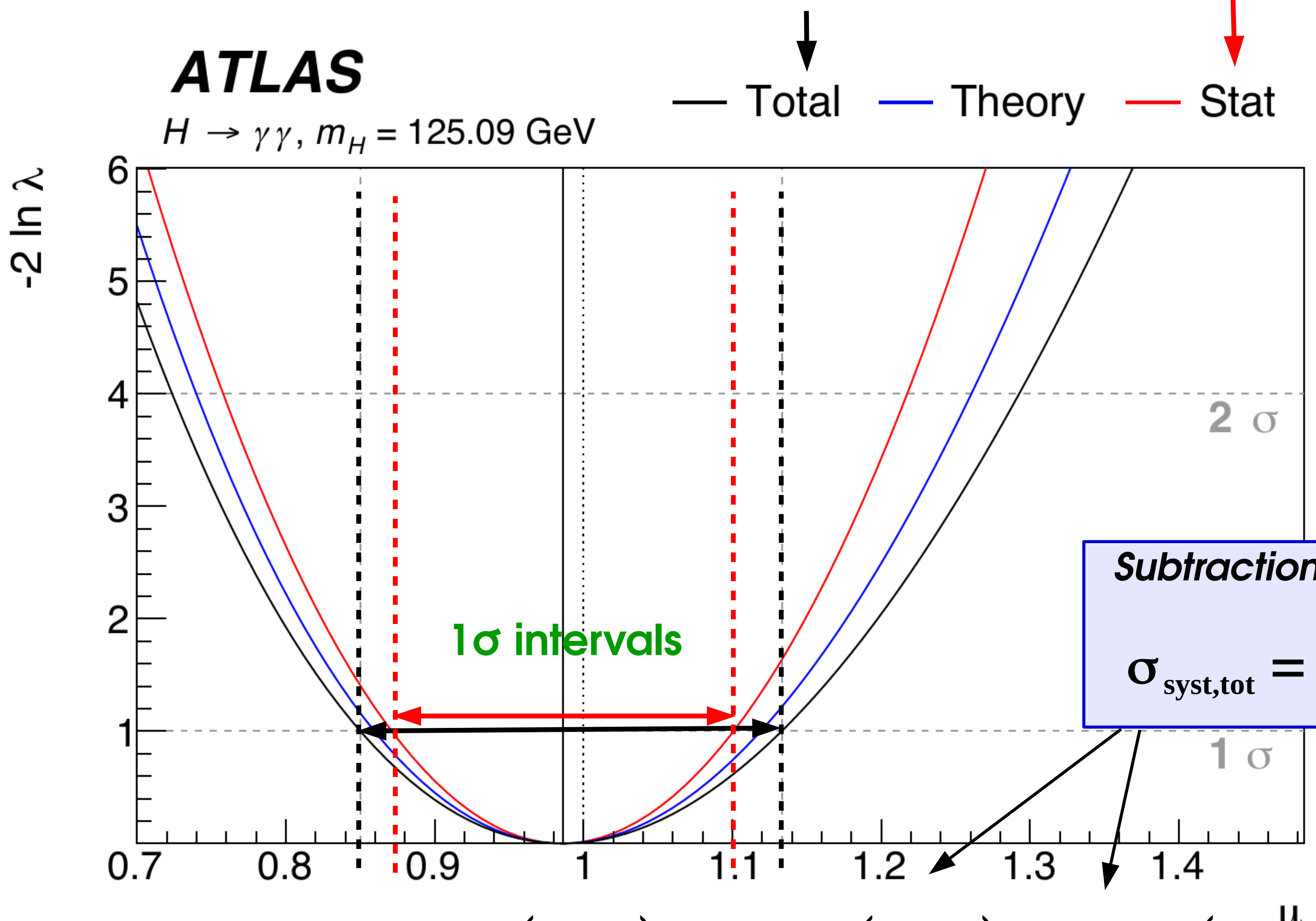
$$S = (n_{obs} - B_{obs}) \pm \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2} \qquad \sigma_S = \sqrt{\sigma_{stat}^2 + \sigma_{bkg}^2}$$

Stat uncertainty (on n) and systematic (on B) add in quadrature

Uncertainty decomposition

No systematics NPs included : statistical uncertainty only

All systematics NPs included: stat+syst uncertaintes



Subtraction in quadrature

$$\sigma_{\text{syst,tot}} = \sqrt{\sigma^2_{\text{total}} - \sigma^2_{\text{stat}}}$$

$$\mu = 0.99 \pm 0.12\,(\text{stat}) \pm 0.06\,(\text{syst}) \pm 0.06\,(\text{theo})$$

Profiling issues

ATLAS-CONF-2016-058

Systematics are described by NPs included in the fit. Define **pull** as

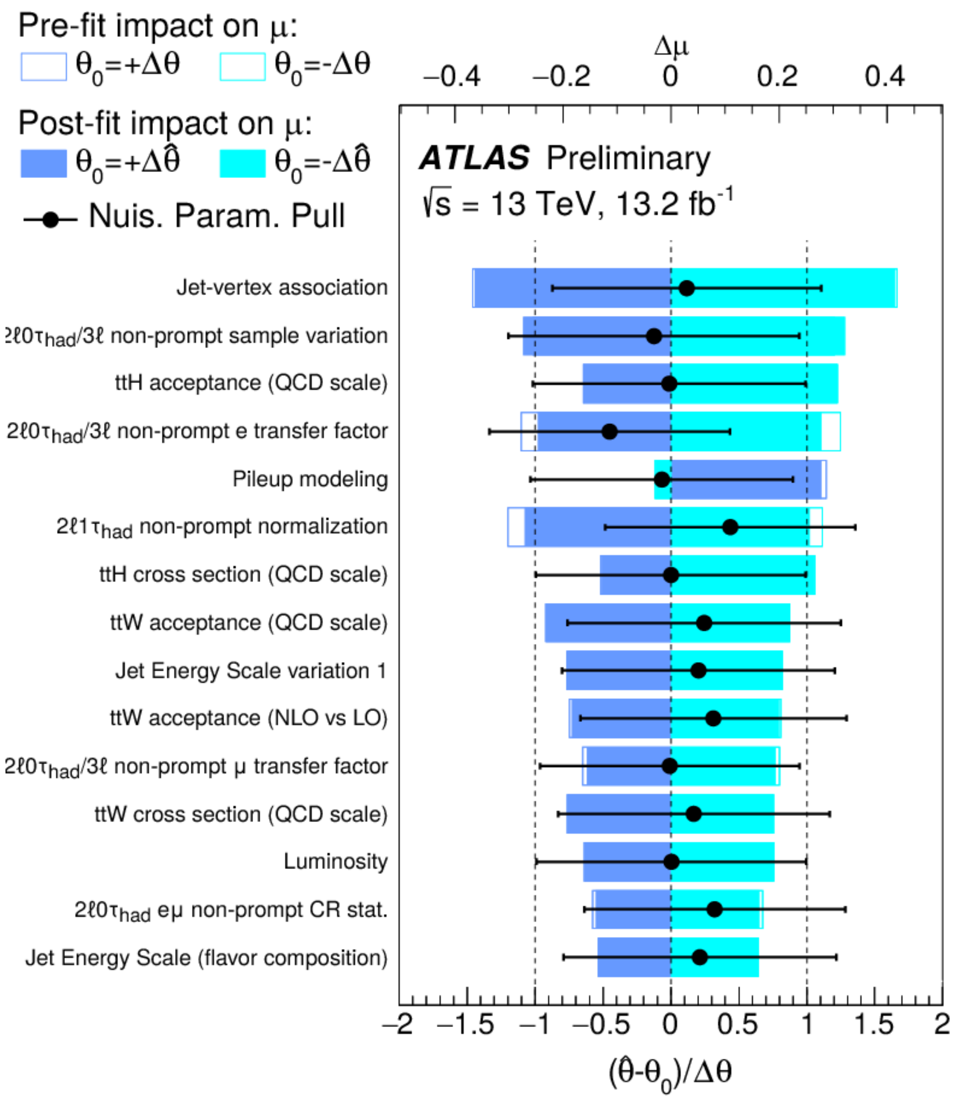
$$\left(\hat{\theta} - \theta_0\right) / \sigma_\theta$$

Nominally:

- **Pull = 0** : i.e. the pre-fit expectation
- **Pull uncertainty = 1** : from the Gaussian

However fit results may be different:

- **Central value ≠ 0**: some data feature differs from MC expectation
  ⇒ Need investigation if large
- **Uncertainty < 1** : effect is *constrained* by the data ⇒ Needs checking if this legitimate or a modeling issue



→ **Impact on result** of $\pm 1\sigma$ shift of NP allows to gauge which NPs matter most .

Profiling issues

Systematics are described by NPs included in the fit. Define **pull** as

$$\left(\hat{\theta}-\theta_0\right) / \sigma_\theta$$

Nominally:

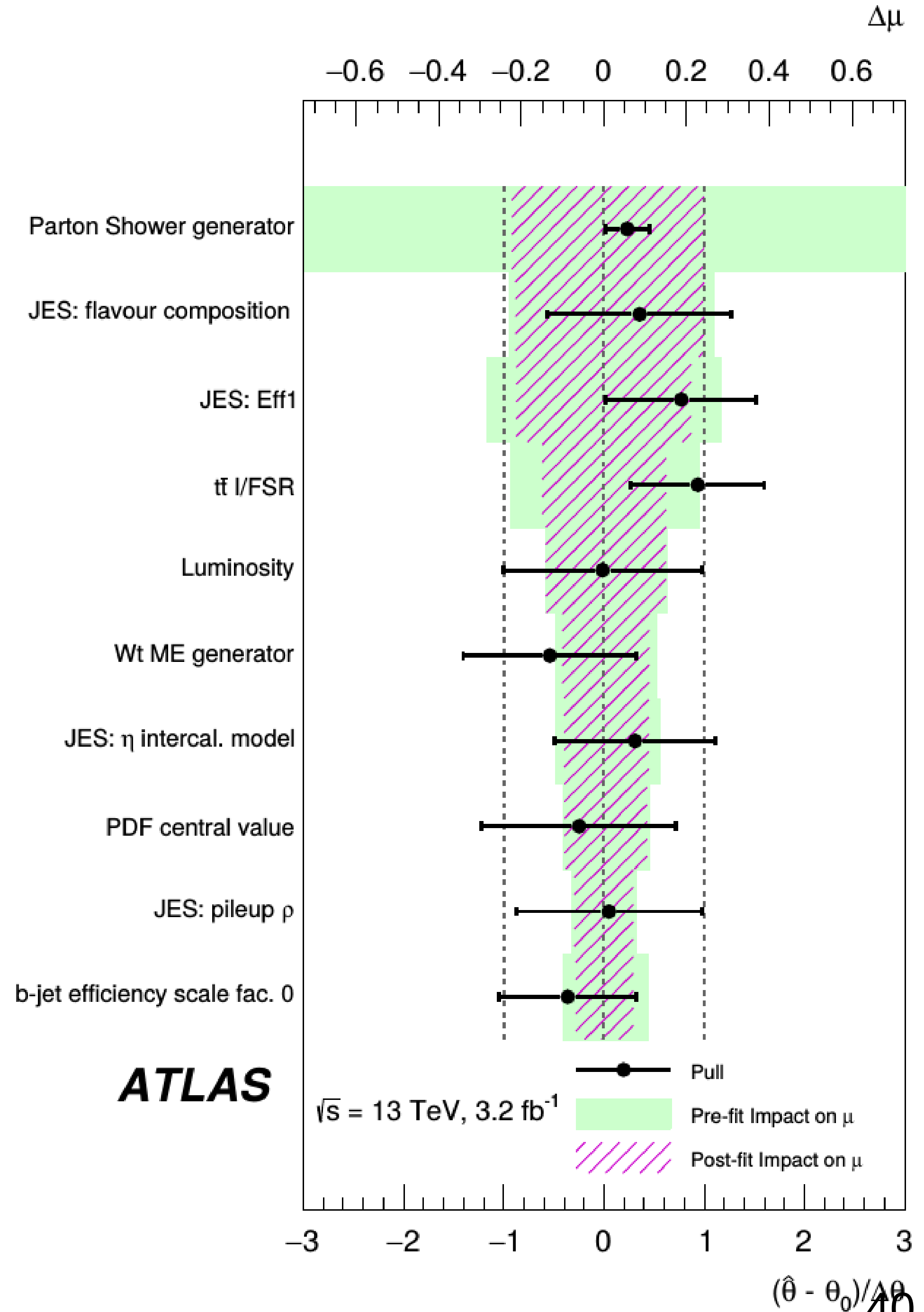
- **Pull = 0** : i.e. the pre-fit expectation
- **Pull uncertainty = 1** : from the Gaussian

However fit results may be different:

- **Central value ≠ 0**: some data feature differs from MC expectation
  ⇒ Need investigation if large
- **Uncertainty < 1** : effect is *constrained* by the data ⇒ Needs checking if this legitimate or a modeling issue

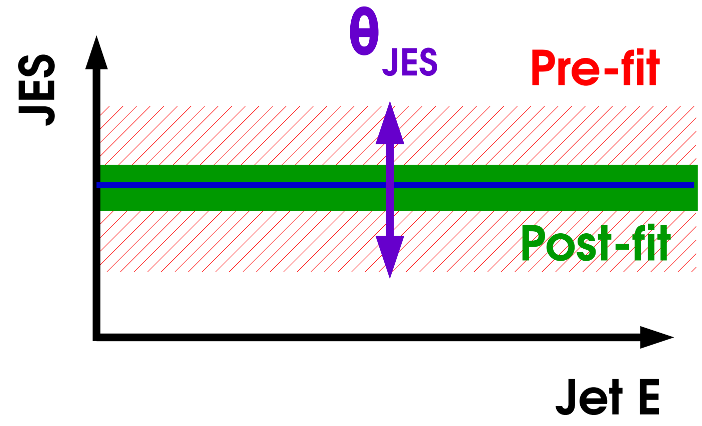
→ **Impact on result** of $\pm1\sigma$ shift of NP allows to gauge which NPs matter most .

13 TeV single-t XS (arXiv:1612.07231)



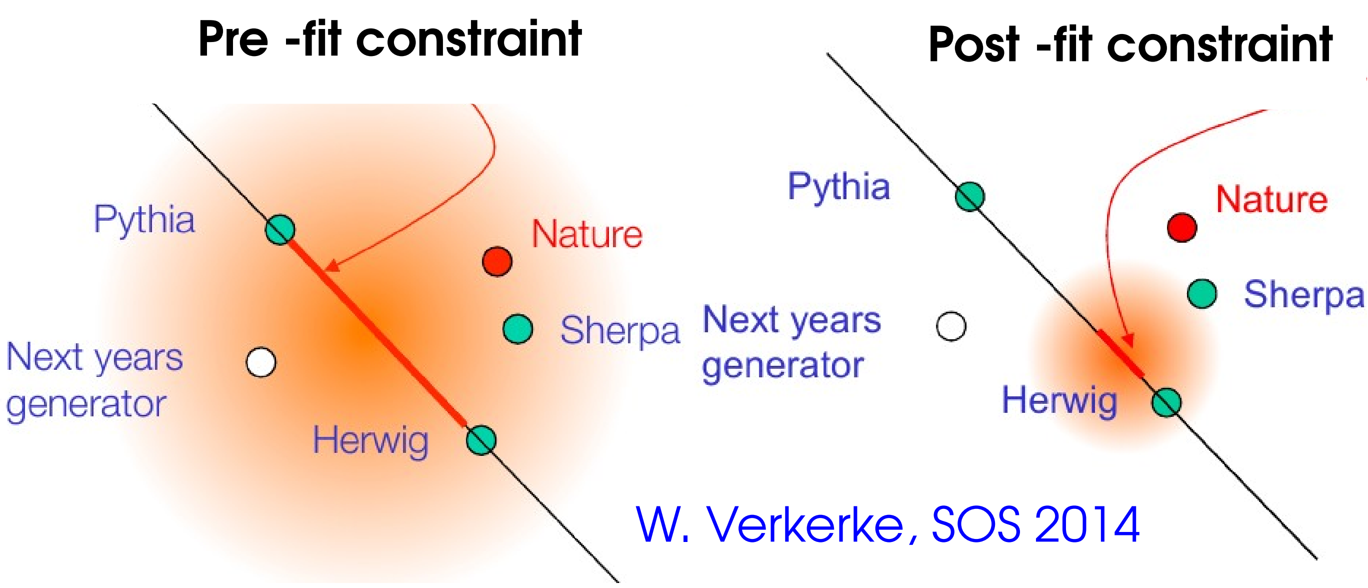
Profiling issues

Too simple modeling can have unintended effects
→ e.g. single Jet E scale parameter:
⇒ Low-E jets calibrate high-E jets – intended ?



***Two-point* uncertainties**:
→ Interpolation may not cover full configuration space
⇒ Can lead to too-strong constraints

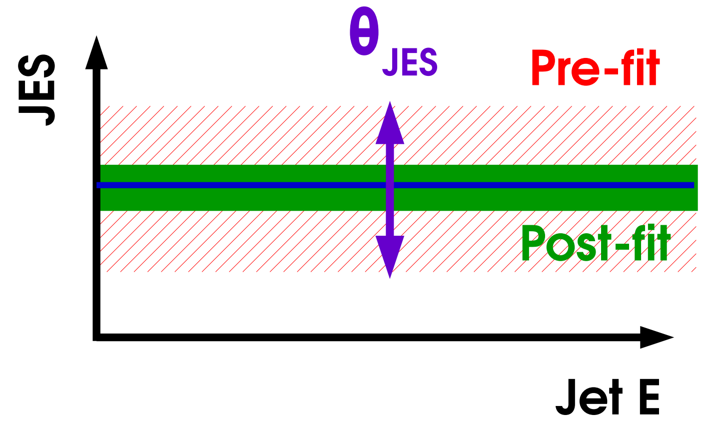
Typical examples: simulation uncertainties ("PYTHIA vs. HERWIG")



W. Verkerke, SOS 2014

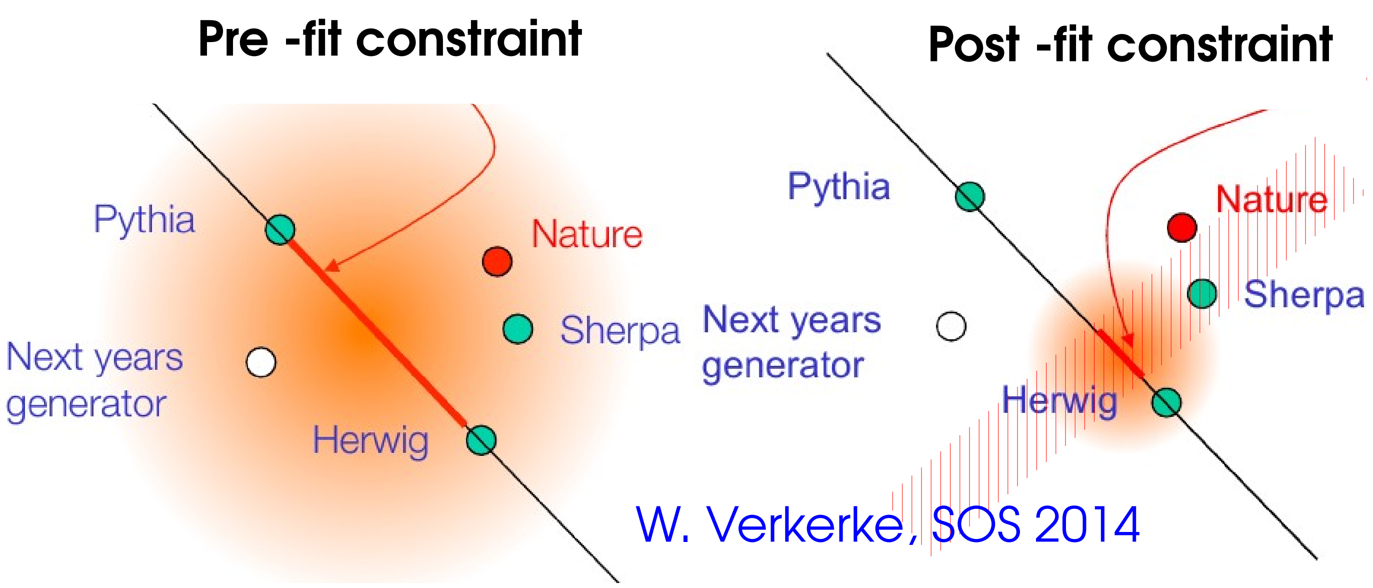
Profiling issues

Too simple modeling can have unintended effects
→ e.g. single Jet E scale parameter:
⇒ Low-E jets calibrate high-E jets – intended ?



***Two-point* uncertainties**:
→ Interpolation may not cover full configuration space
⇒ Can lead to too-strong constraints

Typical examples: simulation uncertainties ("PYTHIA vs. HERWIG")



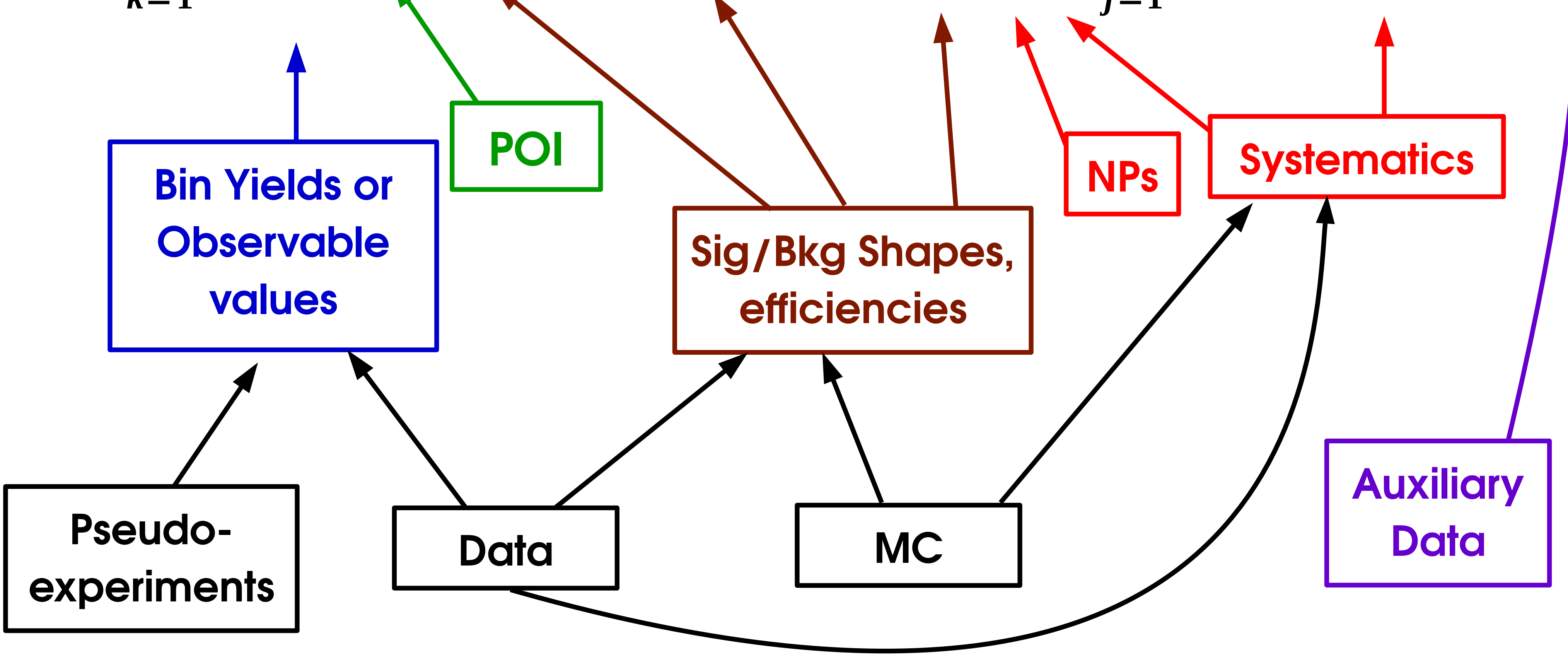
W. Verkerke, SOS 2014

Systematics

Impact of Systematics

$$L\left(\mu, \{\theta_j\}_{j=1\ldots n_{NP}}; \{n_i^{(k)}\}_{i=1\ldots n_{data}^{(k)}}^{k=1\ldots n_{cat}}, \{\theta_j^{obs}\}_{j=1..n_{NP}}\right) =$$

$$\prod_{k=1}^{n_{cat}} P\left[n_i; \overbrace{\mu\, \epsilon_{i,k}(\vec{\theta})\, N_{S,i,k}(\vec{\theta}) + B_{i,k}(\vec{\theta})}^{\text{Expected bin yield}}\right] \prod_{j=1}^{n_{syst}} G(\theta_j^{obs}; \theta_j; 1)$$



× number of categories!

Impact of Systematics

Expected bin yield

$$L\left(\mu,\{\theta_j\}_{j=1\ldots n_{NP}};\{c\,n_i^{(k)}\}_{i=1\ldots n_{data}^{(k)}}^{k=1\ldots n_{cat}},\{e\,\theta_j^{obs}\}_{j=1..n_{NP}}\right)=$$

$$\prod_{k=1}^{n_{cat}} P\left[n_i;\mu\,\epsilon_{i,k}(\vec{\theta})\,N_{S,i,k}(\vec{\theta})+B_{i,k}(\vec{\theta})\right]\prod_{j=1}^{n_{syst}} G(\theta_j^{obs};\theta_j;1)$$

Bin Yields or Observable values

POI

Systematics

Key ingredient: impact of systematics on signal and background yields

$$N_{S,i,k}(\theta_j)=N^0_{S,i,k}\prod_j(1+\delta_{i,j,k}\theta_j)$$

Pseudo-experiments

Data

MC

Auxiliary Data

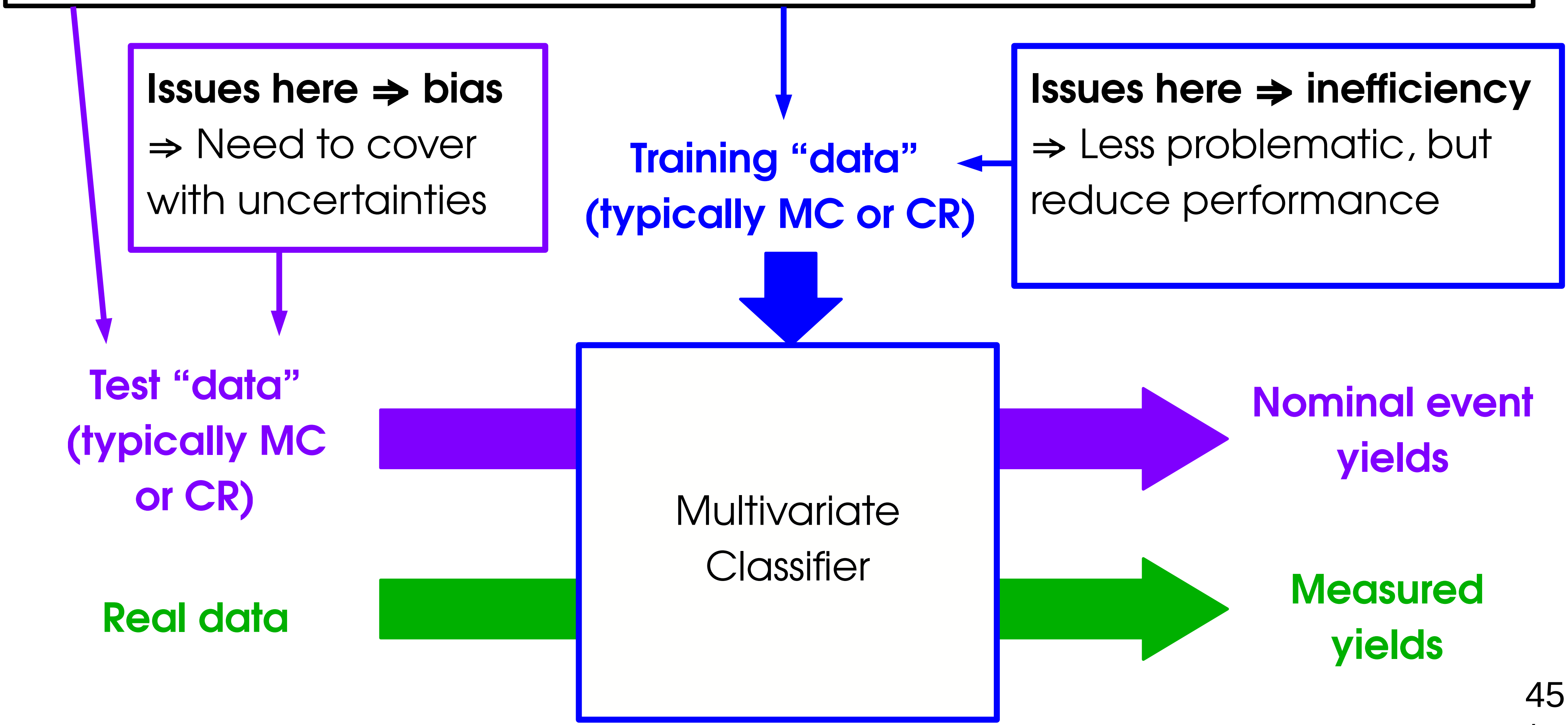
× number of categories!

Accounting for Systematics

SciPost Phys. 8, 090 (2020)

Effects to account for:

- **Training/test data not representative of real data (mismodeling)**
  → Typically covered by MC systematics: variations in the MC sample describing a range of possible models.
- **Limited size of training dataset** → Covered by *"MC stat"* uncertainties



Randomness in High-Energy Physics

Experimental data is produced by incredibly complex processes

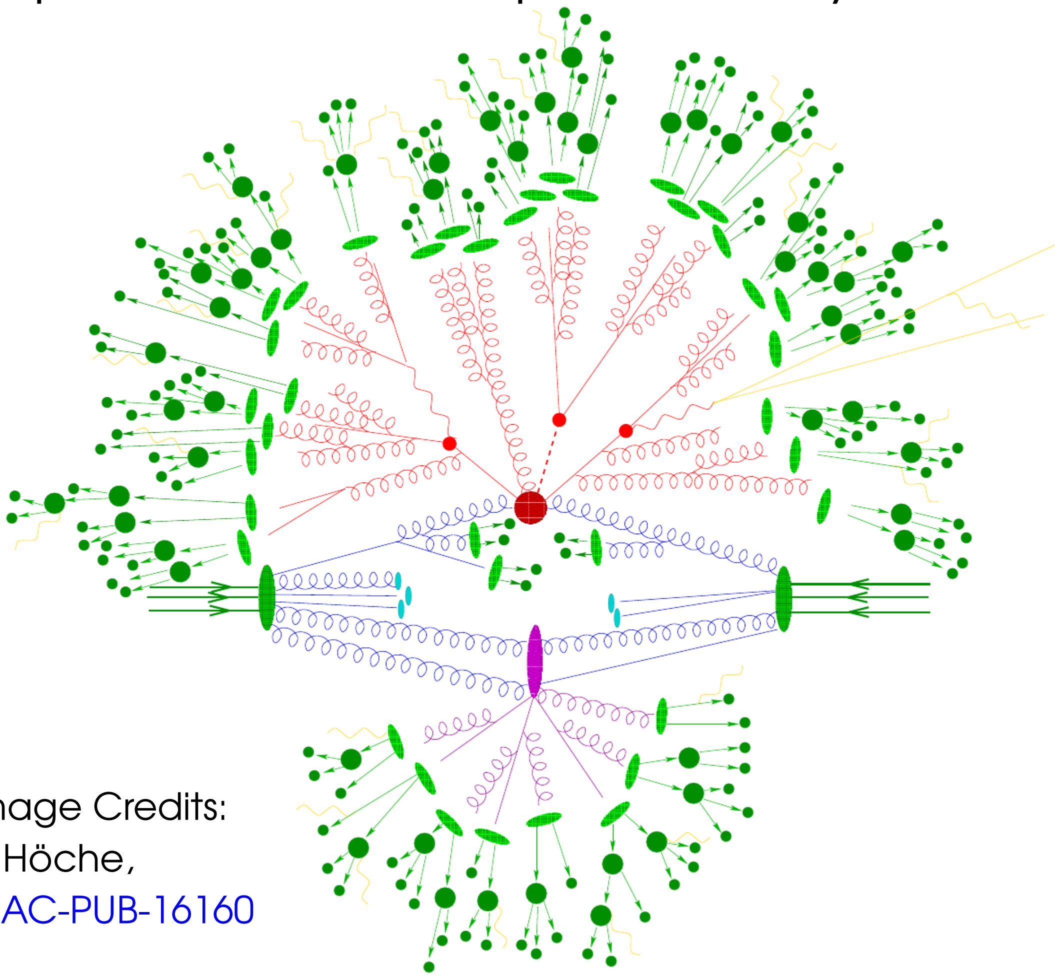


Image Credits:
S. Höche,
SLAC-PUB-16160

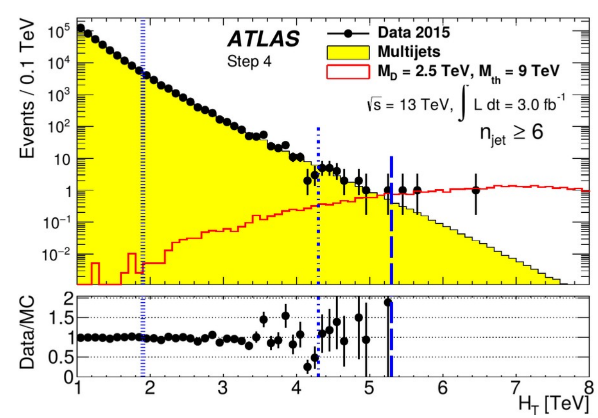
Hard scattering

PDFs, Parton shower, Pileup

Decays

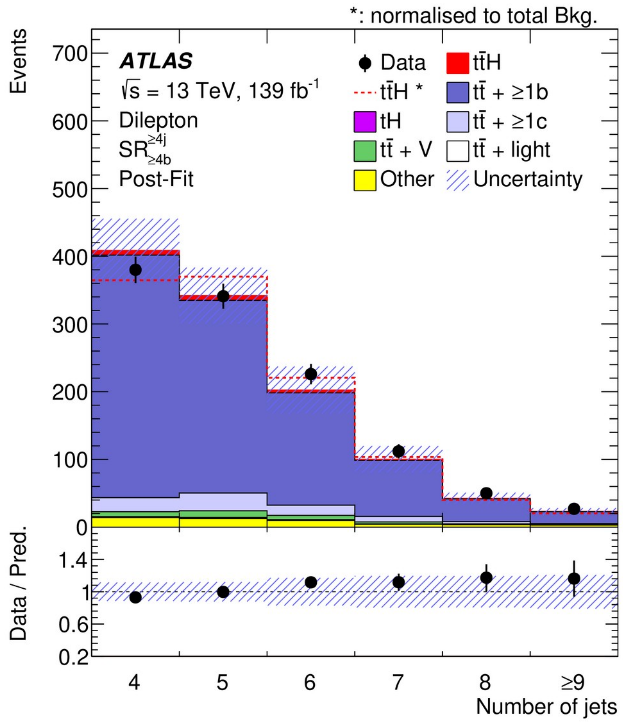
Detector response

Reconstruction



Randomness involved in all stages
→ **Classical** randomness: detector reponse
→ **Quantum** effects in particle production, decay

Modeling Systematics



Some distributions not predicted with sufficient accuracy:

- MC modeling
- Detector response
- CR statistics, CR→SR extrapolation

Error band: combination of above
Typically described by many NPs

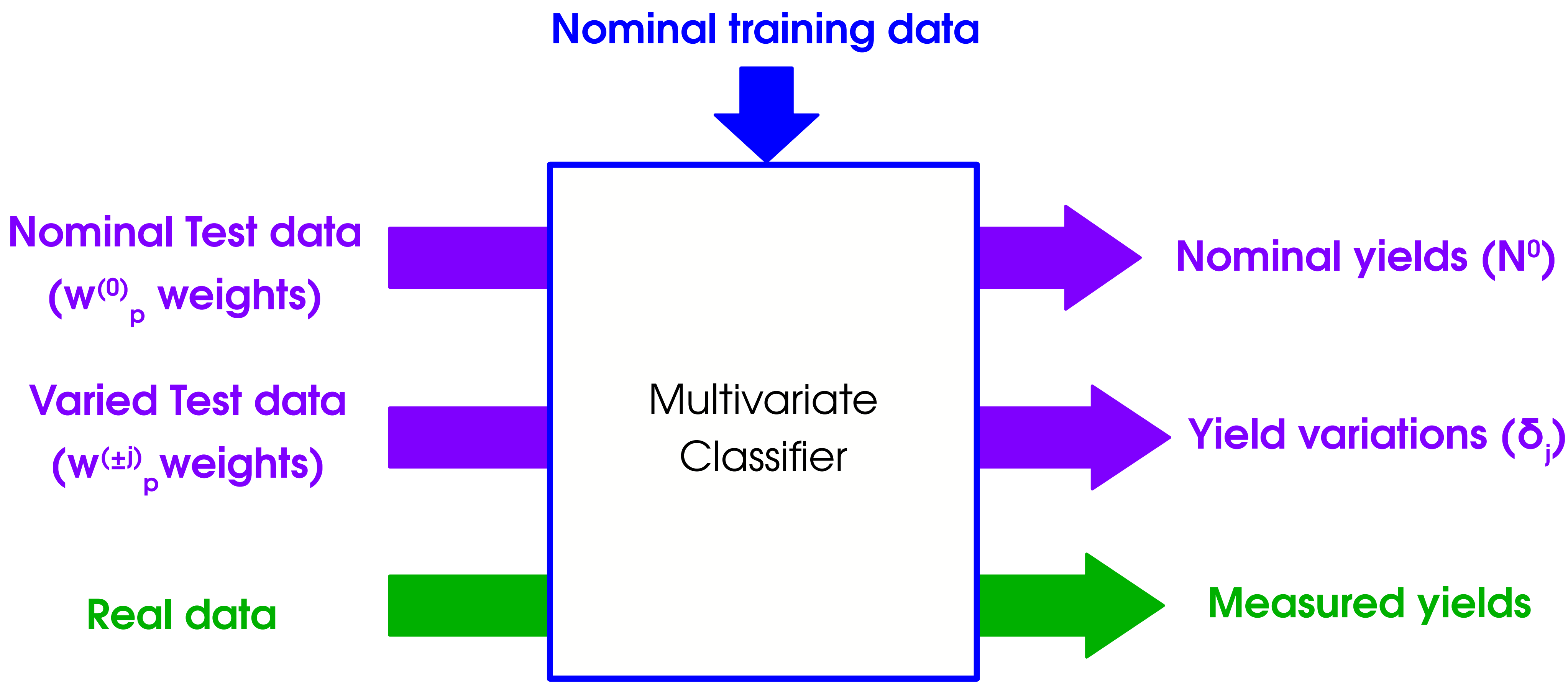
Modeling variations typically implemented through event weights:

- **Nominal modeling** → nominal event weight $w^{(0)}_p$.
- **Each variation $\theta_j = \pm 1$** → associated event weight $w^{(\pm j)}_p$.

Distributions for each case obtained by applying the appropriate weights.

Ultimately, need impact on yields: $$N_{s,i}(\theta_j) = N^0_{s,i} \prod_j (1 + \delta_{i,j}\theta_j)$$

Treating ML Systematics



- **“Propagate uncertainties through the DNN”**
- MC stat uncertainties can be treated similarly using resampling

→ Allows to properly cover for uncertainties, but optimal performance only in nominal case (since used in training).

Parameterized classifiers

Eur. Phys. J. C (2016) 76:235

Optimize a classifier for each value of a model parameters (e.g. a syst NP)

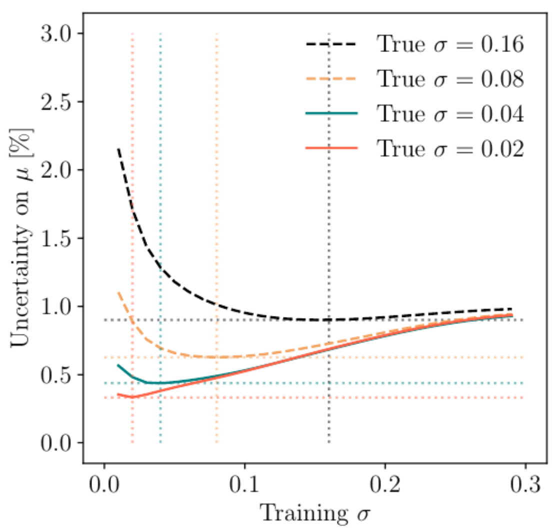
Training data (θ)

Nominal Test data ($w^{(0)}_p$ weights) → DNN(θ_j=0) → Nominal yields (N^0)

Varied Test data ($w^{(\pm j)}_p$weights) → DNN(θ_j=±1) → Yield variations (δ_j)

Real data → DNN(θ) → Measured yields

SciPost Phys. 8, 090 (2020)



- ⊕ Use the optimal classifier for each NP value
  ⇒ Retain optimal performance in each case
- ⊖ Scaling with with |θ| ?
  (in practice |θ| can be $10^{2\text{-}3}$...)

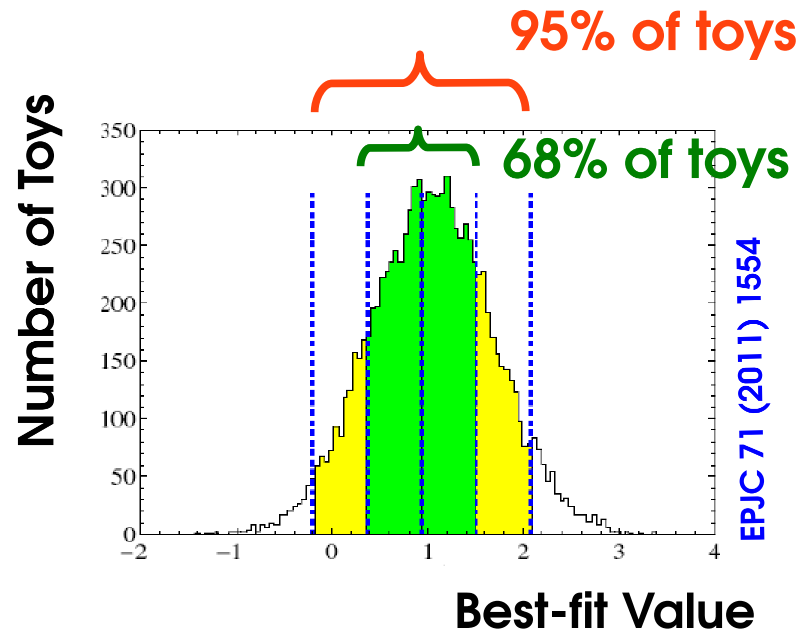
Brute force approach

Generate pseudo-experiments ("toys") and repeat best-fit for each case

→ **Statistics**: resample observed dataset

→ **Systematics**: randomize auxiliary obs. θ_j^{obs}

Obtain intervals from quantiles of the distibution of results



$$\prod_{k=1}^{n_{cat}} P\left[n_i; \mu\, \epsilon_{i,k}(\vec{\theta})\, N_{S,i,k}(\vec{\theta}) + B_{i,k}(\vec{\theta})\right] \boxed{\prod_{j=1}^{n_{syst}} G(\theta_j^{obs}; \theta_j; 1)}$$

- ⊕ No reliance on asymptotic formulas

- ⊖ High CPU requirements (need a fit for each of O(1000) toys)
- ⊖ As before, changing syst NPs ⇒ non-optimal classifier performance
- ⊖ **Optimal case**: need to retrain classifier for each toy ?

Other approaches

- **Inference-aware NN** (De Castro, Dorigo, Com. Phys. Comm. 244 (2019), 170-179)
  → Design a NN to directly minimize the width of the confidence interval on the target POI

- **Likelihood-free inference** (Cranmer, Pavez, Louppe, arXiv:506.02169).
  – Typically, trained classifiers asymptotically learn the likelihood ratio p(x|S)/p(x|B), e.g. when using cross-entropy loss.
  – Parameterized classifiers can estimate POIs without computing L.
  ⇒ Bypass the profile likelihood construction, get intervals from toys ?

- … ?

(Further) Discussion,
Questions, Comments ?

Backup

Collider processes

HEP : Poisson approximation almost always valid:

ATLAS :

- **Event rate ~ 1 GHz**
  (L~10^{34} cm^{-2}s^{-1}~10 nb^{-1}/s, σ_{tot}~10^8 nb,)
- **Trigger rate ~ 1 kHz**

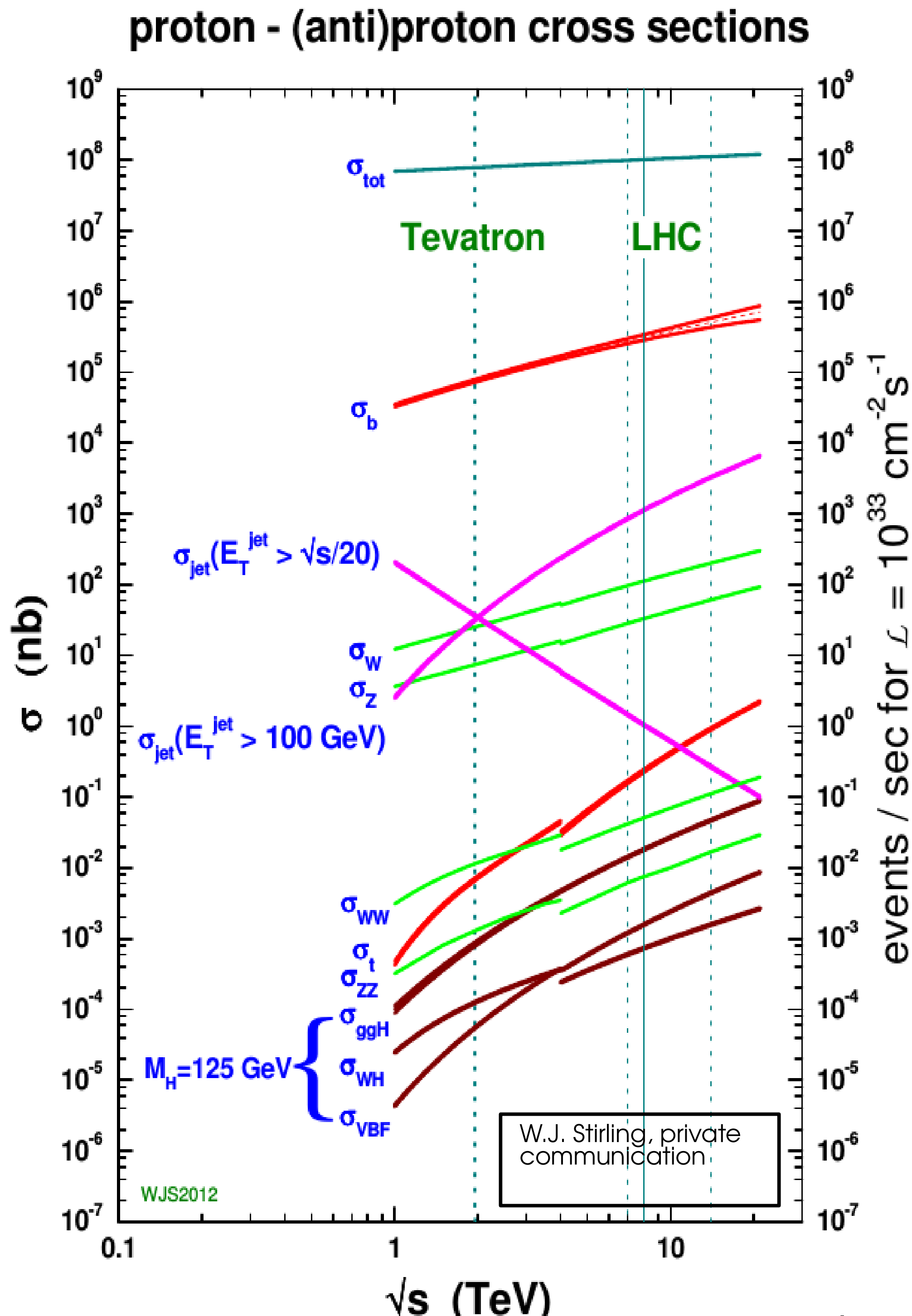
  (Higgs rate **~ 0.1 Hz**)

⇒ p ~ 10^{-6} ≪ 1 ($p_{H\to\gamma\gamma}$ ~ **10^{-13}**)

A day of data: **N ~ 10^{14} ≫ 1**

⇒ Poisson regime! Similarly true in many other physics situations.

Large N = design requirement, to get not-too-small λ=Np...



Bayesian methods

Probability distribution (= likelihood) :

→ Same as frequentist case, but treat systematics by **marginalization**, i.e. **integrating over priors**, instead of profiling:

→ Integrate out θ to get P(μ) : $P(\mu) = \int P(\mu, \theta) C(\theta)\, d\theta$

→ Use probability distribution P(μ) directly for limits & intervals

e.g. 68% CL ("Credibility Level") interval [A, B] is: $\int_A^B P(\mu)\pi(\mu)\, d\mu = 68\,\%$

where π(μ) is the prior on μ. Uses **Bayes' Theorem**: $P(\mu \mid n) = P(n \mid \mu) \dfrac{P(\mu)}{P(n)}$

- No simple way to test for discovery
- Integration over NPs can be CPU-intensive (but can use MCMC methods)

Priors : most analyses use flat priors in the analysis variable(s)

⇒ **Parameterization-dependent**: if flat in σ×B , them not flat in couplings....

→ Can use the Jeffreys' or reference priors, but difficult in practice