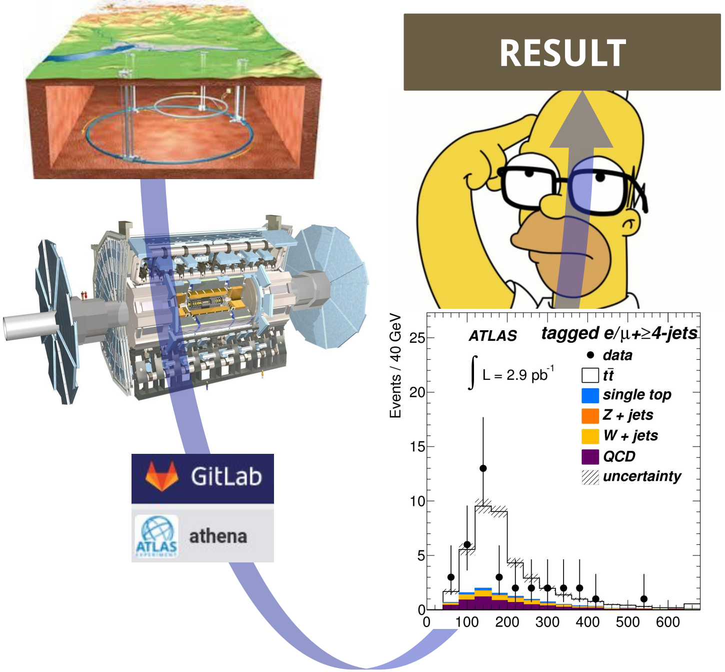
Statistical analysis tutorial

Michele Pinamonti *(INFN Trieste, ATLAS Udine Group)*
michele.pinamonti@ts.infn.it

CODATA-RDA Workshop - ICTP, Trieste 28/07/2022

Statistical analysis

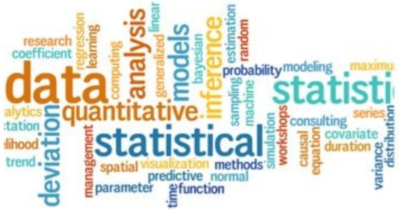
- **Why? What? ...**
  - ***why*** do we need statistical analysis?
  - ***what do we mean*** by statistical analysis?



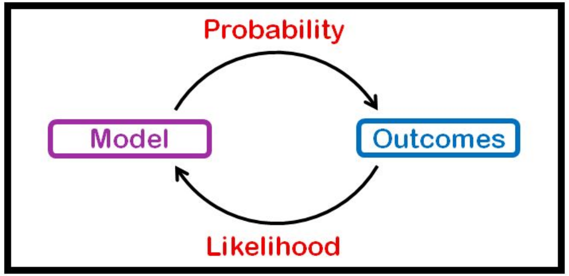
- **Statistical Analysis in particle-collider physics**:
  - the way to extract **quantitative information** from **collision data**

- ... and of course, what goes into the result section of your paper is the **quantitative information**:
  - we want to claim things like:
    - *"X = Y ± Z"*
    - *"X > Y excluded at 95% confidence level"*
    - *"X observed with a significance of 5 σ"*

Statistical Analysis Basics (for HEP)



Maximum likelihood and Fits



- **Likelihood**:
  - defined as **probability** of observing a certain set of **data** given a model / hypothesis (with certain **parameter** values)

$$L(\vec{\theta}) = Prob(\vec{x}|\vec{\theta}) = \prod_i Prob(x_i|\vec{\theta})$$

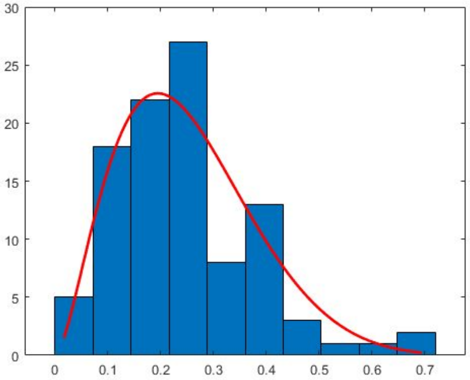
probability

data

parameters

if data points / measurements / observation are independent (i.e. uncorrelated)

- **Maximum Likelihood principle**:
  - ***estimated*** value(s) of parameter(s) = value(s) ***maximizing*** the Likelihood

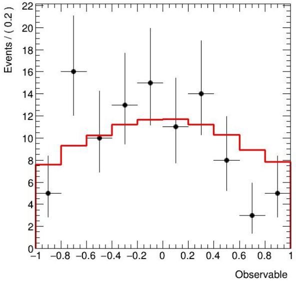


- "**Fit**":
  - **parameter estimation** procedure via Likelihood maximization

Types of likelihood

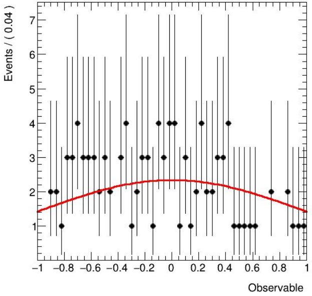
- **Binned Poisson Likelihood**:

$$\mathcal{L}(\vec{n}|\vec{\theta}) = \prod_{i \in bins} P(n_i|Y_i(\vec{\theta})) = \prod_{i \in bins} P(n_i|S_i(\vec{\theta}) + B(\vec{\theta}))$$



- **Unbinned Likelihood**:

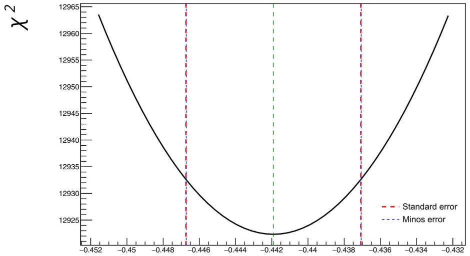
$$\mathcal{L}(\vec{m}|\vec{\theta}) = P(n_{obs}|S+B) \times \prod_{i=1}^{n_{obs}} \frac{S \cdot \mathcal{P}_S(m_i, \vec{\theta}) + B \cdot \mathcal{P}_B(m_i, \vec{\theta})}{S+B}$$



- **The χ^2 case**:
  - binned Poisson *L*, for large *n* ⇒ $\mathcal{L}(\vec{n}|\vec{\theta}) \simeq \prod_{i \in bins} \mathcal{G}(n_i|\mu_i = Y_i(\vec{\theta}), \sigma_i = \sqrt{n_i})) = \prod_{i \in bins} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{(n_i-\mu_i)^2}{2\sigma_i^2}}$

⇒

$$-2 \log \mathcal{L}(\vec{n}|\vec{\theta}) = \sum_{i \in bins} \frac{(n_i - \mu_i)^2}{\sigma_i^2} + Const. = \chi^2$$

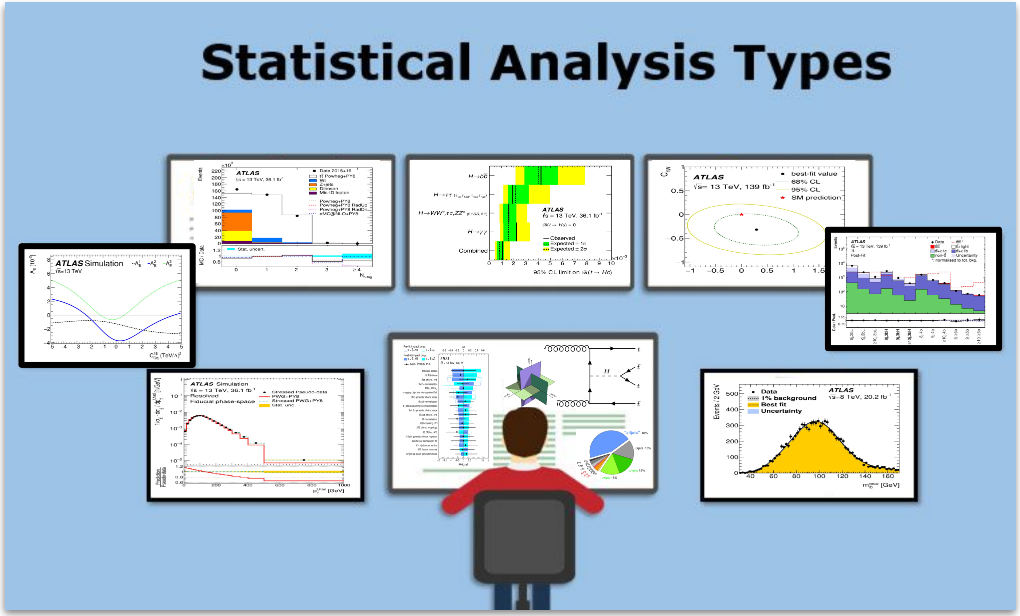


  - Maximizing Likelihood = Minimizing χ^2 or -2 log *L*

ATLAS EXPERIMENT

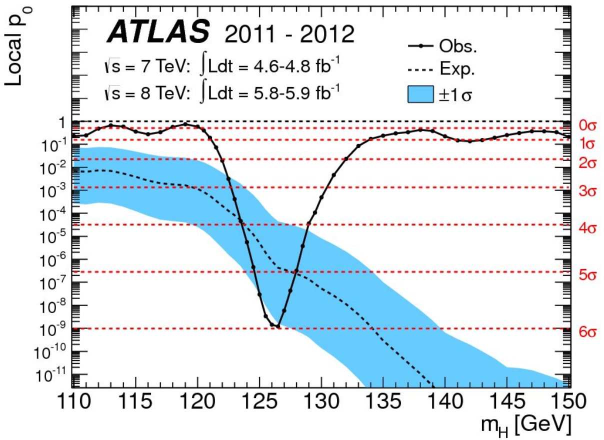
Types of measurements

- In HEP data analysis, **different types of measurement**:
  - **searches** for a new process
  - **cross-section** measurements:
    - *total cross-section*
      (full / fiducial phase-space...)
    - *differential cross-section*
    - *ratios of cross-sections...*
  - other **parameter estimation**:
    - *usually "shape analyses"*
      (e.g. top mass, top width...)
  - **EFT** fits / limits ...

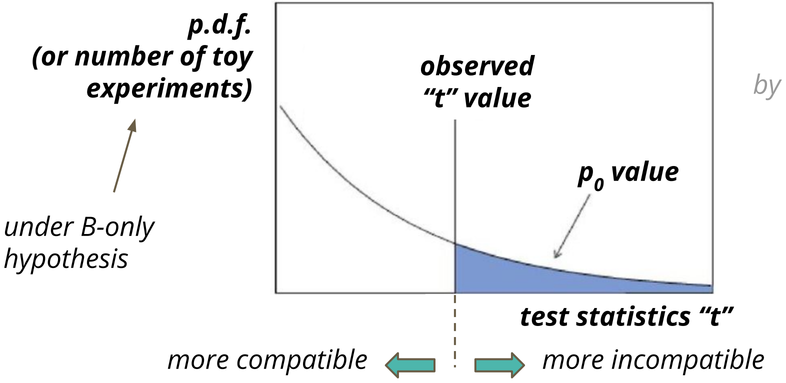


- Also, measurements can take as **inputs**:
  - **binned** data ⇒ histogram counts are the inputs
  - **unbinned** data ⇒ individual events as "input measurements"
  - other existing **measurements** / differential cross-section bins ⇒ "2 step" analysis

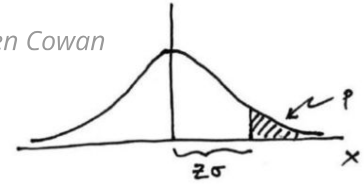
Searches: discovery significance



- **Observing** a new process (*in the Frequentist language*)
  = seeing data ***incompatible*** with **background-only** hypothesis
  (or "null hypothesis")
- **How to quantify it**?
  - define "**test statistics**", quantifying agreement btw. data and a prediction (e.g. likelihood, χ^2, LH-ratio...)
  - define **p_0-value** = probability of seeing ***worse agreement*** than observed one, in the background-only hypothesis
    - *i.e. "probability that what we see is a fake signal"*



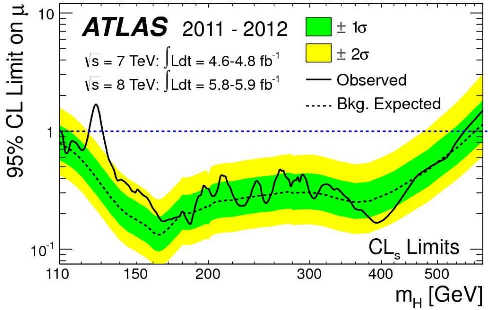
by Glen Cowan

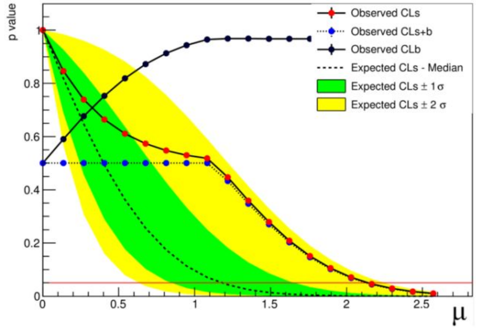


  - turn p_0 into number of Gaussian std.dev, define **significance** "Z" in terms of *number of sigmas*:
    - $5\sigma = \sim 3 \cdot 10^{-7}$

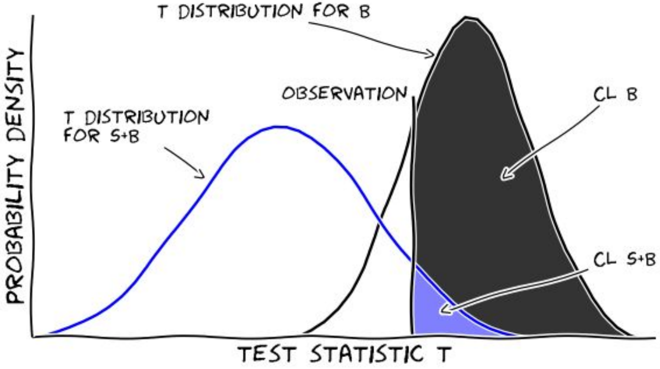
Searches: exclusion limits

- If no evidence for signal, setting **exclusion limits**
- Usually limits set on **signal strength** $\mu = \sigma^{obs} / \sigma^{theory}$:
  - values of μ > quoted value ***excluded at 95% confidence-level***
- <u>**Operationally**</u>:
  - define **test-statistics** (as before), ***t* (data,μ)**
  - **scan** values of μ, get t^{obs} for each μ
  - assign prob. of seeing worse t than t^{obs}, assuming that value of μ
  - **find μ** for which **prob. = 5%** *(i.e. 1 - 95%, corresponding to 2σ)*





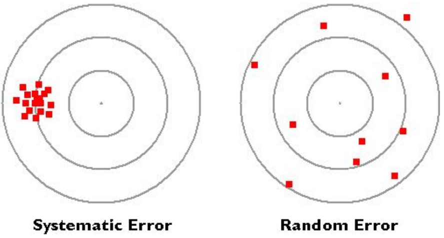
- What does CL_s mean?
  - description above defines "CL_{s+b}"
  - can then define "CL_b" as follows:
    - get t^{obs} for each μ (as before)
    - define CL_b as prob. of seeing worse t, in the **B-only hypothesis** (μ=0)
  - then define $CL_s = CL_{s+b} / CL_b$



see these slides

ATLAS EXPERIMENT

Systematic Uncertainties



Inclusion of systematic uncertainties

- In particle collision physics we distinguish:
  - **statistical uncertainty**:
    - result of **stochastic fluctuations** in data
    - consequence of limited size of analysed dataset

    ⇨ *Fully **uncorrelated** between subsequent measurements*
  - **systematic uncertainties**:
    - everything that is not a statistical uncertainty
    - uncertainties associated with measurement apparatus, assumptions made, or model used

    ⇨ *Fully **correlated** between subsequent measurements*

- ***Statistical*** uncertainty usually **intrinsically included** in inference method (e.g. in χ^2 fit)
- ***Systematic*** uncertainties: **non-obvious inclusion** in and **propagation** through statistical analysis

- **Side considerations**:
  - in our world, systematic uncertainties are uncertainties on Prob(x,θ), i.e. uncertainties **on *expected values*** (e.g. exp. S+B), ***not* on data** (!)
  - systematics divided into multiple independent / uncorrelated "**sources**"

The Profile Likelihood formalism

- More and more common approach for including systematics in HEP statistical analysis:
  - include systematic uncertainties as unknown parameters in the model
  - nuisance parameters modifying expectations in a parametric way
  - prior probabilities on values of nuisance parameters to reflect limited knowledge

- **The binned profile-likelihood**:

Gaussian
(or other pdf...)

data

Poisson

$$L(\vec{n} \mid \vec{\theta}, \vec{k}) = \prod_i P(n_i \mid \underbrace{S_i(\vec{\theta}, \vec{k}) + B_i(\vec{\theta}, \vec{k})}) \times \prod_j \underbrace{G(\theta_j)}$$

data events
in bin i

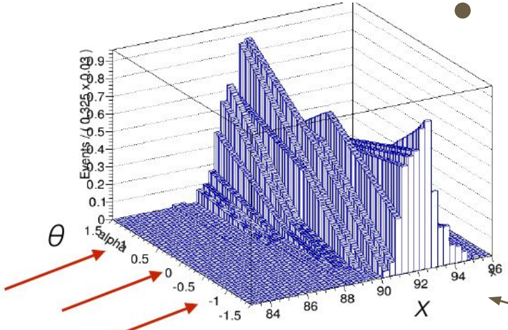
prediction in bin i
(*signal+background*)

constraint term
for nuisance
parameter j

constrained parameters:
nuisance parameters (**NPs**)
associated to systematic
uncertainties

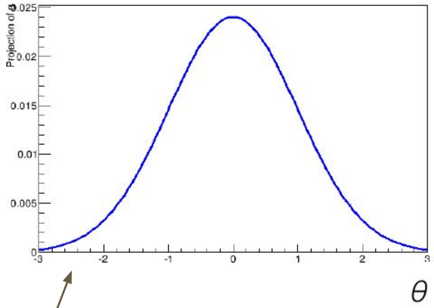
unconstrained parameters:
parameter of interest (**POI** or "μ") + unconstrained nuisance parameters (e.g. background normalization parameters)

Nuisance parameters and systematic uncertainties

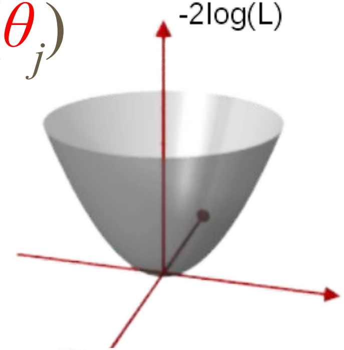


- Each (*independent*) source of **systematic uncertainty** included in the likelihood as constrained **nuisance parameters** (NPs):
  - affecting *S*+*B* prediction in a ***coherent* way**
  - effect ***interpolated*** and ***extrapolated*** from **3 discrete values** (0 = nominal, 1 = “up” var., -1 = “down” var.) to range of **continuous values**

normal distribution



$$L(\vec{n} \mid \vec{\theta}, \vec{k}) = \prod_i P(n_i \mid S_i(\vec{\theta}, \vec{k})+B_i(\vec{\theta}, \vec{k})) \times \prod_j G(\theta_j)$$



- The fit procedure becomes a **multi-dimensional Likelihood maximisation** problem
  - the fit **result** is not just the value (and uncertainty) on parameter(s) of interest (POI), but **a set of values** for all the parameters, including nuisance parameters: $(\hat{\mu}, \hat{\theta}_0, \ldots \hat{\theta}_{N-1}) : \mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}) = max$

Profile likelihood ratio and asymptotic regime

- **Neyman-Pearson lemma**:
  - the **likelihood ratio** $\lambda = L(H_1)/L(H_0)$ is the **optimal discriminator** when testing hypothesis H_1 vs. H_0 (e.g. H_1 = presence of signal (μ>0), H_0 no signal (μ=0))
  - in the case of our profile likelihood, can build profile likelihood ratio, as a function of POI:

Profile likelihood ratio only dependent on μ

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

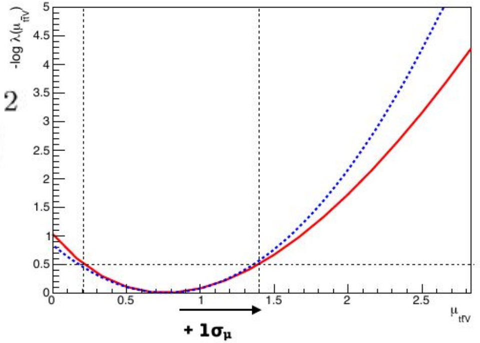
Maximize L for a given μ 'conditional' likelihood

Maximize L 'unconditional' likelihood

  - maximizing λ vs. μ = maximizing L vs. (μ,θ)

- **Wilks' Theorem**: in large statistics data samples, *λ distribution* **follows χ^2 distribution**:

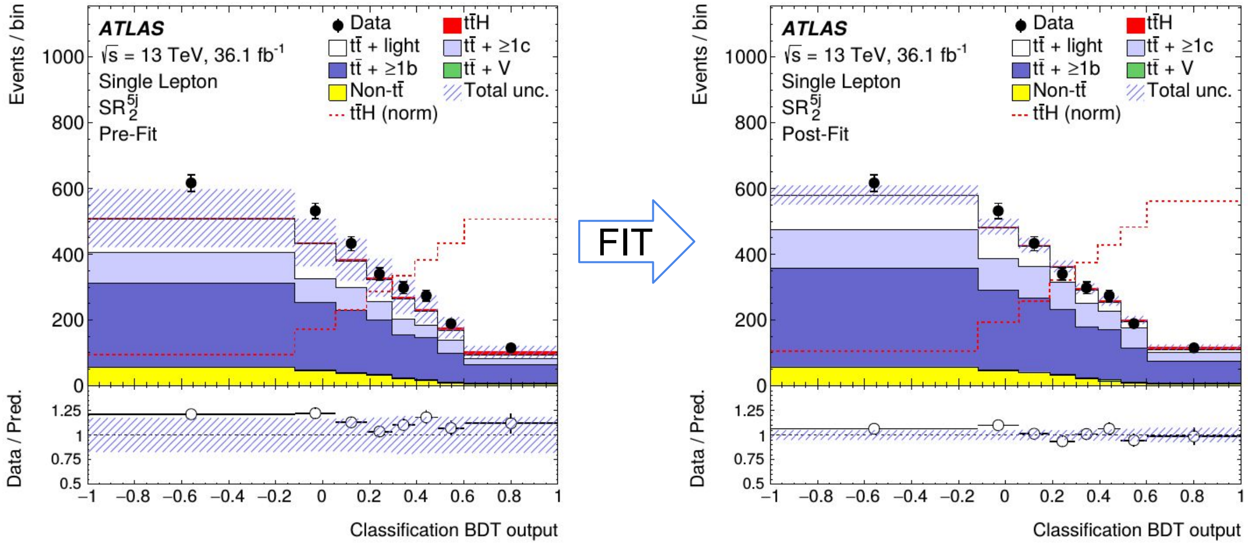
$$-2\log\lambda(\mu) = -2(\log L(\mu, \hat{\hat{\theta}}) - \log L(\hat{\mu}, \hat{\theta})) = \left(\frac{\mu - \hat{\mu}}{\sigma_\mu}\right)^2$$



⇒ can get the **uncertainty on *μ*** **(including effect of all systematics!!)**
  - large-statistics means > ~ O(10) events
  - saves from running very time consuming pseudo-experiments

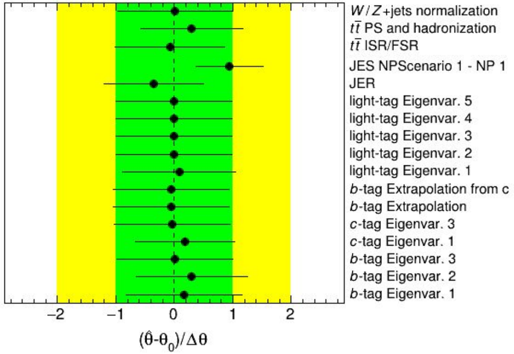
Profiling, pre-fit and post-fit

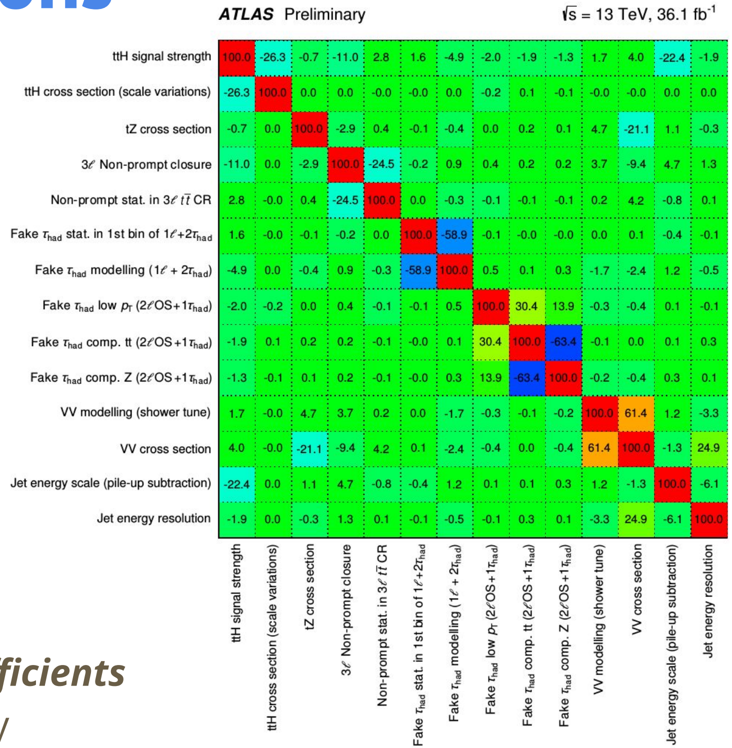
- Profile likelihood fit can:
  - **change background prediction**, if best-fit θ values different from θ_0
  - **reduce uncertainty** on background, through:
    - ***constraint*** of NPs
      (*"improved knowledge" of parameters that are affected by systematic uncertainties, i.e. data have enough statistical power to further constraint the NP*)
    - ***correlations***
      between NPs



NP pulls, constraints and correlations

- Useful to **monitor** NP **pulls** and **constraints**:
  - they are "*nuisance*", but they can be important!



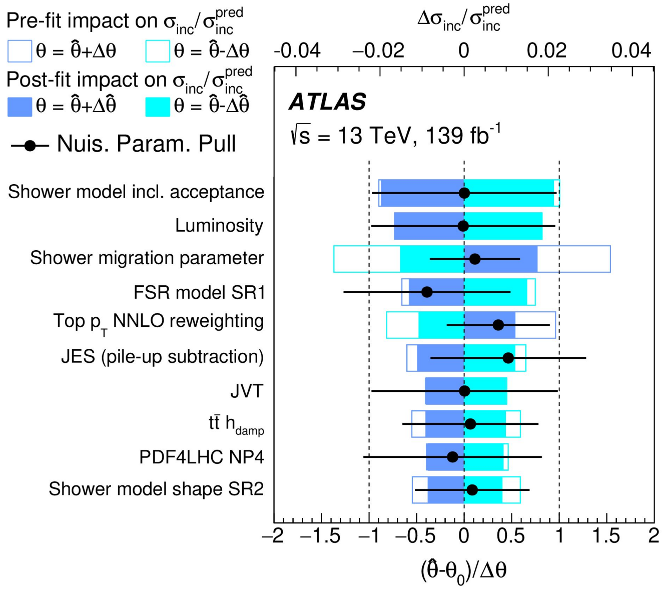


- Important to consider also NP **correlations**:
  - uncertainties on NPs (*and POI*) extracted from ***covariance matrix***, which includes ***correlation coefficients***
    - correlation **built by the fit**, even if completely independent / uncorrelated sources of uncertainty before the fit (*correlation in the improved knowledge of the parameters*)
    - (anti-)correlations can **reduce** total post-fit uncertainty!

Impact of systematics

"which systematics are more important?"

1. "**Ranking plot**" 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 error
    - fit re-done with *N-1* parameters
    - impact = difference in **central value** of μ



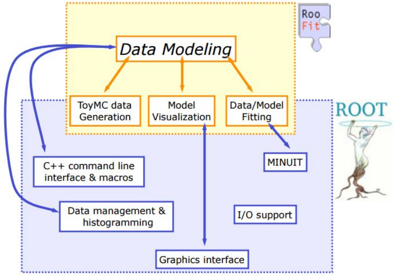
2. "**Grouped impact table**" reports *contributions* to *total uncertainty* from **groups** of syst.:
    - fix a group of NPs to post-fit values
    - repeat the fit, get reduced error on μ
    - impact = difference in quadrature btw. original and reduced error on μ
    - get stat. uncertainty by fixing all NPs

Category	$\frac{\Delta\sigma_{fid}}{\sigma_{fid}}$ [%]	$\frac{\Delta\sigma_{inc}}{\sigma_{inc}}$ [%]
Signal modelling		
$t\bar{t}$ shower/hadronisation	±2.8	±2.9
$t\bar{t}$ scale variations	±1.4	±2.0
...		
Total systematic uncertainty	±4.3	±4.6
Data statistical uncertainty	±0.05	±0.05
Total uncertainty	±4.3	±4.6

Tools for statistical analysis (with Profile Likelihood)

ATLAS EXPERIMENT

Profile likelihood - Implementation in ROOT



Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{min}}^{x_{max}} f(x)dx$	RooRealIntegral
list of space points		RooAbsData

- **RooFit:** toolkit to extend **ROOT** providing language to describe data models
  - model distribution of observable *x* in terms of parameters θ
    using probability density function PDF

- **RooStats:** project to provide advanced stat. techniques for LHC collaborations
  - built on top of **RooFit**

- **RooWorkspace:** generic container class for all **RooFit** objects, containing:
  - full model configuration
    (i.e. all information to run statistical calculations)
  - PDF and parameter/observables descriptions uncertainty/shape of nuisance parameters
  - (multiple) data sets

- **HistFactory:** tool for creating **RooFit** workspaces formatted for use with **RooStats** tools
  - meant for analyses based on template histograms

Practical part

Repository and environment

- GitHub repository: https://github.com/pinamont/statistics-tutorial

- The whole tutorial will be run through Jupyter notebooks (python and ROOT/C++ based)

- 2 available options:
    - **Binder**
    - **SWAN+cern-box**

- Goal: guide you through **what's *actually* done** to publish your results
    - with some **exercises** to get acquainted with the machinery
    - we'll choose ***dynamically*** what to cover (*raise your hands!*)
    - you may use the rests as a **reference** (*& feel free to contact us!*)

Setting up the environment

- Go to the [GitHub repository]

GitHub

- Choose one of the 2 options:
  - **Binder**:
    - no CERN account needed
    - could take more time to load...
  - **SWAN**:
    - CERN account needed (and cern-box / eos space set up)
    - should be faster to start
- Follow instructions on the **README file** for setting up environment, according to chosen option
- Once ready, try running the `hello_world.ipynb` notebook

README.md

Statistics Tutorial - ICTP ATLAS Open Data 2022

Authors

Michele Pinamonti - INFN Sezione di Trieste

credits to: Valerio Ippolito - INFN Sezione di Roma

Scope

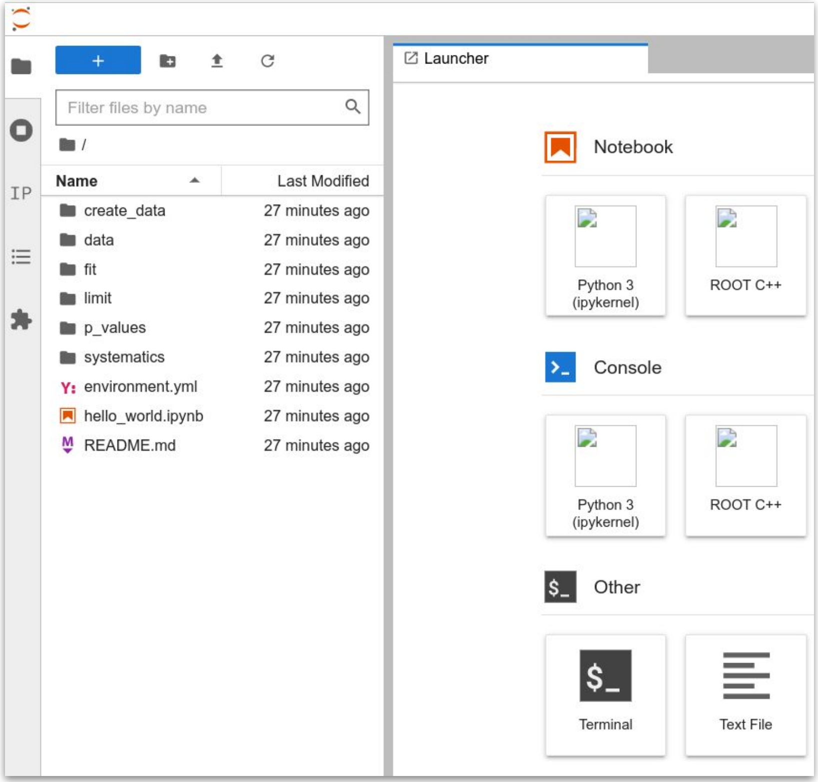
We will go through the typical steps of defining, filling up and analysing a workspace.

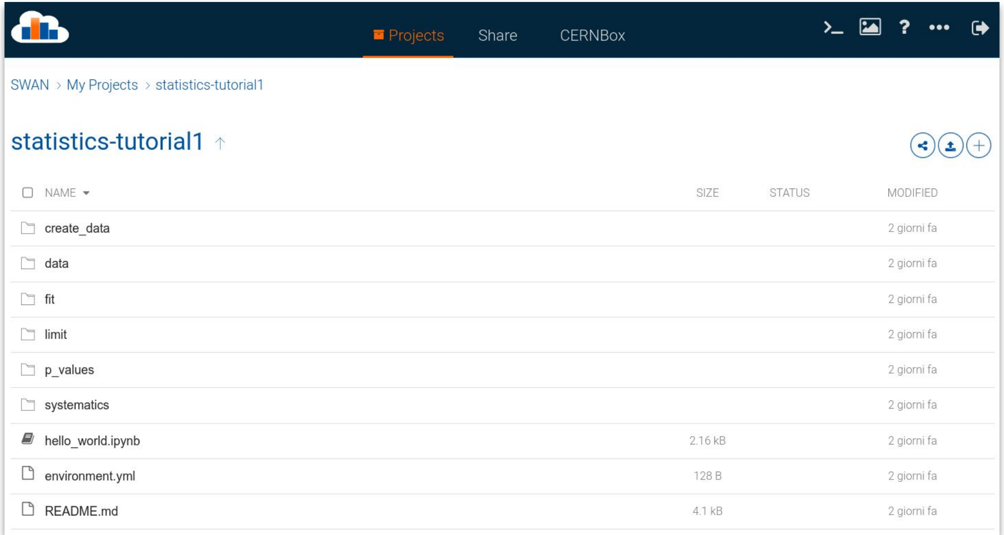
Preliminaries

There are two main ways to run this tutorial: Binder and SWAN.

- **Caveat:**
  - exercise doesn't seem to work with ROOT version 26.04 (set by default in SWAN)
  - setting-up ROOT version 24.06 in Binder
  - following instructions on README for SWAN should work as well (setting-up 24.06)

Binder and SWAN interfaces





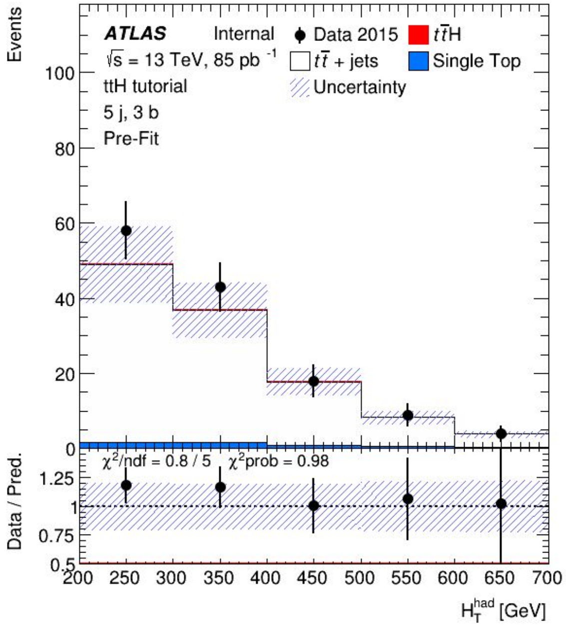
Tutorial

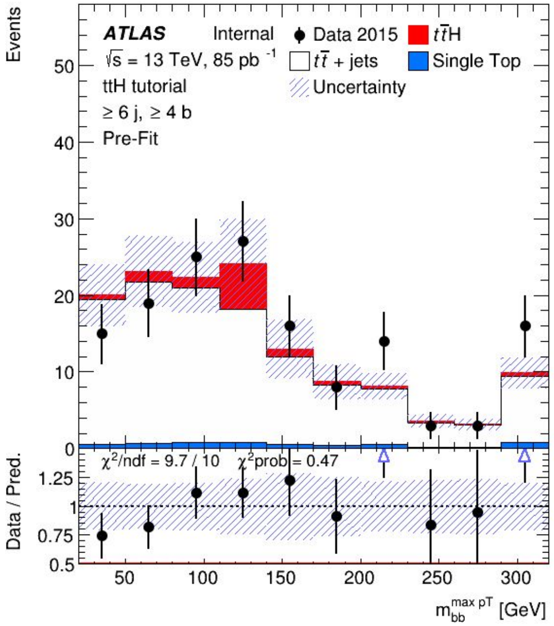
- Tutorial structured as a **set of notebooks**, each performing a single action:
    - `create_data/create_workspace.ipynb` → create a RooWorkspace from existing histograms
      → will use output of this notebook for all other operations
        - simplified version `create_workspace_minimal.ipynb` also available
    - `create_data/inspect_workspace.ipynb` → inspect what's inside the workspace we just created
    - `fit/simple_fit.ipynb` → perform a fit and print fit results
    - `fit/postfit_plots.ipynb` → visualize projection of fit results to expected distributions
    - `systematics/ranking.ipynb` → breakdown of impact of systematics - method 1
    - `systematics/impact_table.ipynb` → breakdown of impact of systematics - method 2
    - `limit/toys.ipynb` → perform exclusion limit extraction
    - `p_values/pvalues.ipynb` → p-value and significance calculation

Our example workspace

- We'll use as an exercise a set of inputs (histograms):
  - ATLAS ttH search (H → bb), part of real fitting exercise with very first 2015 data
  - tt̄+(b)-jets selection (1-lepton channel)

- Two statistically independent datasets ("regions" or "channels", as you wish):
  - "5 j, 3 b"
    → Control Region, enriched in tt̄ + (b)jets
  - "≥ 6 j, ≥ 4 b"
    → Signal Region





Backup

p_0-value and discovery significance

- **Observing** a new process
  = seeing data ***incompatible*** with **background-only** hypothesis ("null hypothesis")

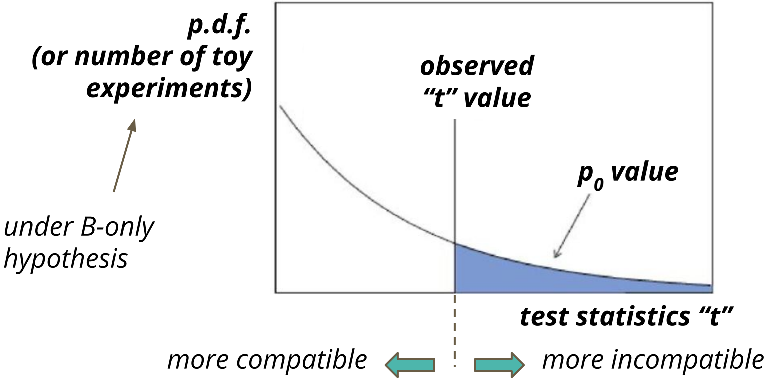
- <u>**How to quantify it**</u>?
  - define "**test statistics**", quantifying data-prediction agreement
  - define ***p_0*-value** = probability of seeing ***worse agreement*** (in B-only hypothesis)
  - turn p_0 into number of Gaussian std.dev, define **significance** "Z"
    in terms of *number of sigmas*

reminder:

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

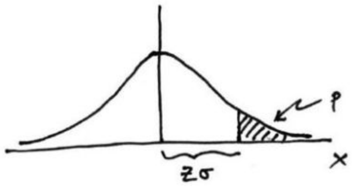
One-sided tests-statistics

- *in the case of profile-likelihood ratio:*



$$q_0 = \begin{cases} -2ln\lambda(0) & \hat{\mu} \geq 0 \\ 0 & \hat{\mu} < 0 \end{cases}$$

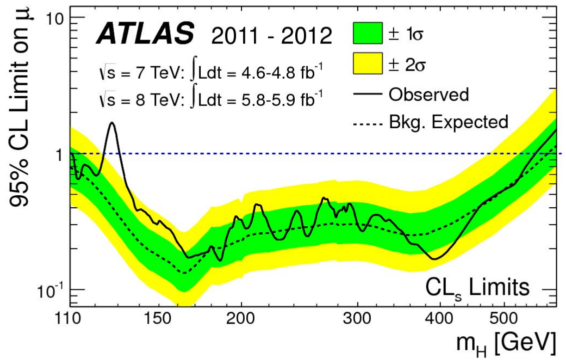
$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0)\, dq_0$$

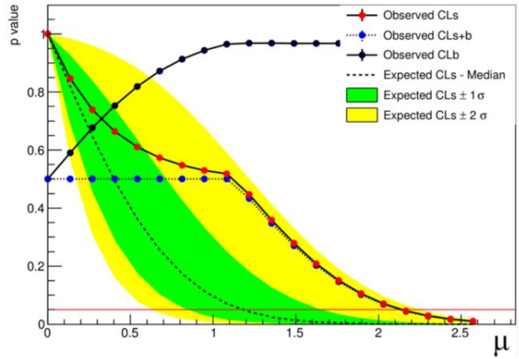


$$Z_0 = \Phi^{-1}(1 - p_0)$$

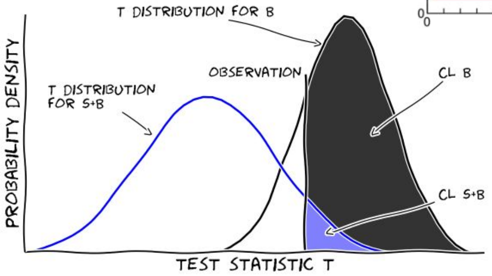
Exclusion limits

- No evidence ⇒ **exclusion limits**
  - usually on **signal strength** $\mu = \sigma^{obs} / \sigma^{theory}$
- Define **test-statistics** (as before), ***t* (data,μ)**
  - **scan** values of μ, get t^{obs} for each μ
  - assign prob. of seeing worse t than t^{obs}, assuming that value of μ
  - **find** μ for which **prob. = 5%** *(i.e. 1 - 95%, corresponding to 2σ)*

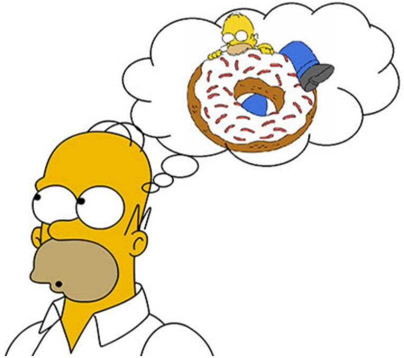




- What does CL_s mean?
  - description above defines "CL_{s+b}"
  - can then define "CL_b" as follows:
    - get t^{obs} for each μ (as before)
    - define CL_b as prob. of seeing worse t, in the **B-only hypothesis** (μ=0)
  - then define $CL_s = CL_{s+b} / CL_b$

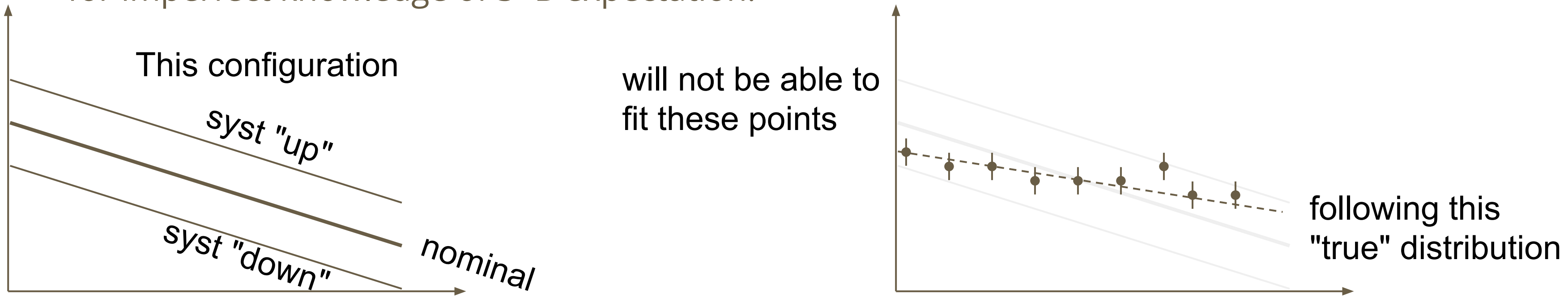


Profiling pitfalls



- The profile likelihood approach is **valid** with some **assumptions**
  - in particular, assumed that "*nature*" can be described by the model with ***a single combination of values*** for the parameters

- Cannot just take *large uncertainties* hoping that they are enough to cover for imperfect knowledge of S+B expectation!



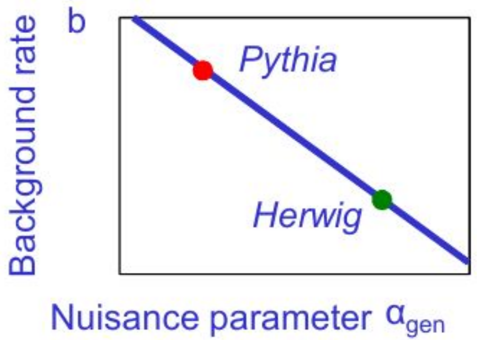
- "**Flexibility**" / "**granularity**" of the systematics model needs to be considered

Theory modeling systematics

- ***Experimental systematics*** nowadays often well suited for profile likelihood application:
  - come from calibrations ⇒ gaussian constraint appropriate
  - broken-down into several independent/uncorrelated components (JES, *b*-tagging...)
- Different situation for **theory systematics**:
  - **difficulty 1:** what is the **distribution** of the subsidiary measurement?
  - **difficulty 2:** what are the **parameters** of the systematic?
    - can a combination of the included parameters describe **any possible** configuration?
    - is **any allowed value** of the parameter physically meaningful?

See: https://indico.cern.ch/event/287744/contributions/1641261/attachments/535763/738679/Verkerke_Statistics_3.pdf

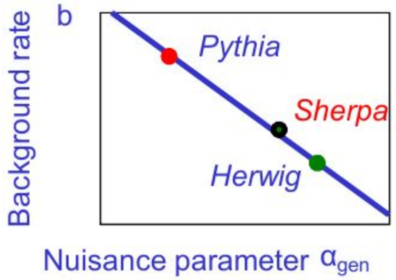
- ***The obviously tricky case:* "two point" systematics**
  - e.g. Herwig vs. Pythia as "parton shower and hadronization model uncertainty", as a single NP



Theory modeling systematics

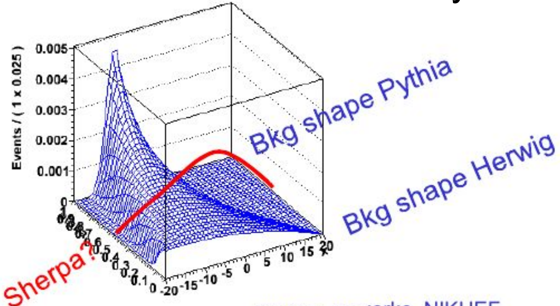
One-bin case:

- reasonable to think that "Sherpa" can be between Herwig and Pythia



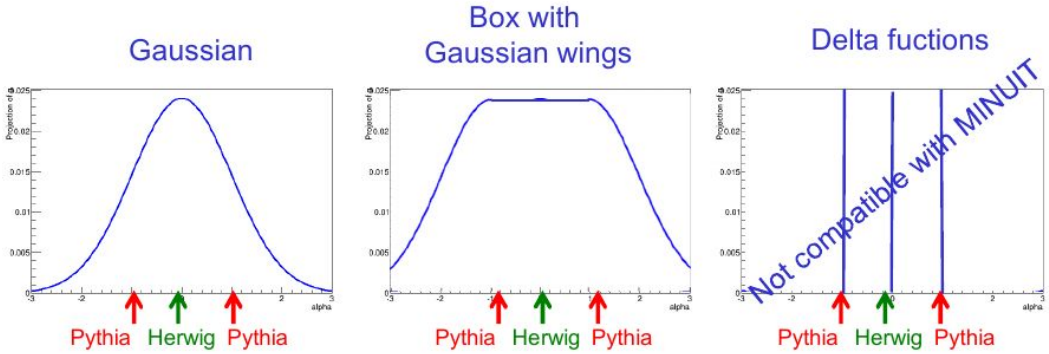
Shape case:

- Sherpa can be different from linear combination of Py and Her...

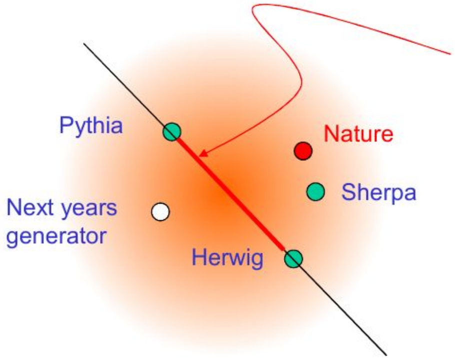


Wouter Verkerke, NIKHEF

Which prior?



Pre-fit / non-constrained NP could be fine to cover for all possible models...



... but is this level of constraint ok?