INTRODUCTION TO MACHINE LEARNING METHODS

Toni Šćulac

Faculty of Science, University of Split, Croatia
Corresponding Associate, CERN

tCSC Machine Learning 2024, Split, Croatia

LECTURES OUTLINE

1) Introduction to Statistics
2) Statistics and Machine Learning
3) Classical Machine Learning
4) Introduction to Deep Learning
5) Advanced Deep Learning

STATISTICS AND MACHINE LEARNING

INTRODUCTION TO HYPOTHESIS TESTING

- A key task in most of physics measurements is to discriminate between two or more hypotheses on the basis of the observed experimental data.
    - a new particle called the Higgs boson exists?
    - students cheated on the exam?
- This problem in statistics is known as **hypothesis test**, and methods have been developed to assign an observation considering the predicted probability distributions of the observed quantities under the different possible assumptions.
- A **hypothesis H** specifies the probability for the data, i.e., the outcome of the observation, here symbolically: x
- The probability for x given H is also called the **likelihood of the hypothesis**, written L(xlH).

TEST DEFINITION

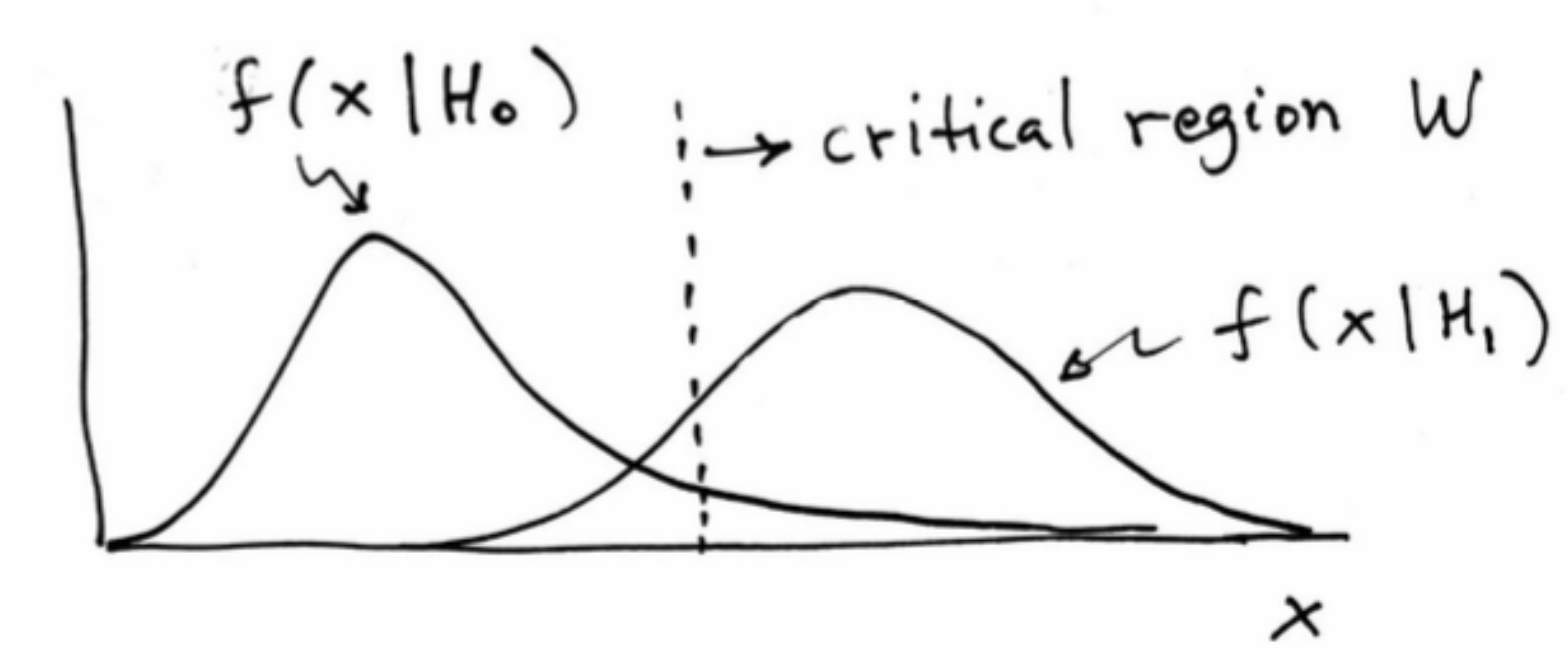
- Goal is to make some statement based on the observed data x as to the validity of the possible hypotheses.
- Consider e.g. a simple hypothesis H_0 and alternative H_1
  - In statistical literature when two hypotheses are present, these are called **null hypothesis** (H_0) and **alternative hypothesis** (H_1)
- A **test** of H_0 is defined by specifying a **critical region W** of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e.,

$$P(x \in W \,|\, H_0) \leq \alpha$$

  - If x is observed in the critical region, reject H_0.
- α is called the size or **significance level** of the test
- Critical region is also called “rejection” region; complement is acceptance region.

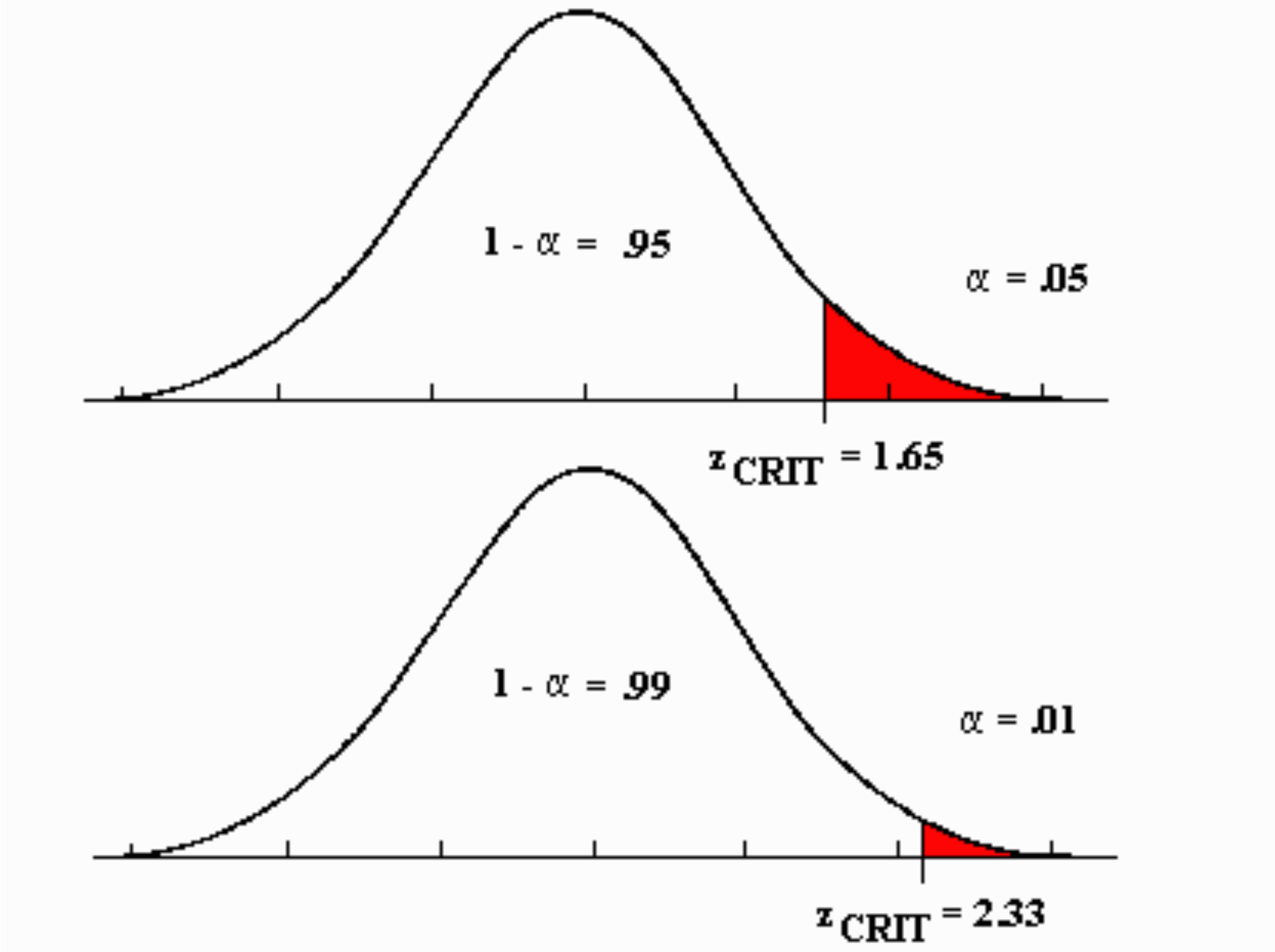
TEST DEFINITION

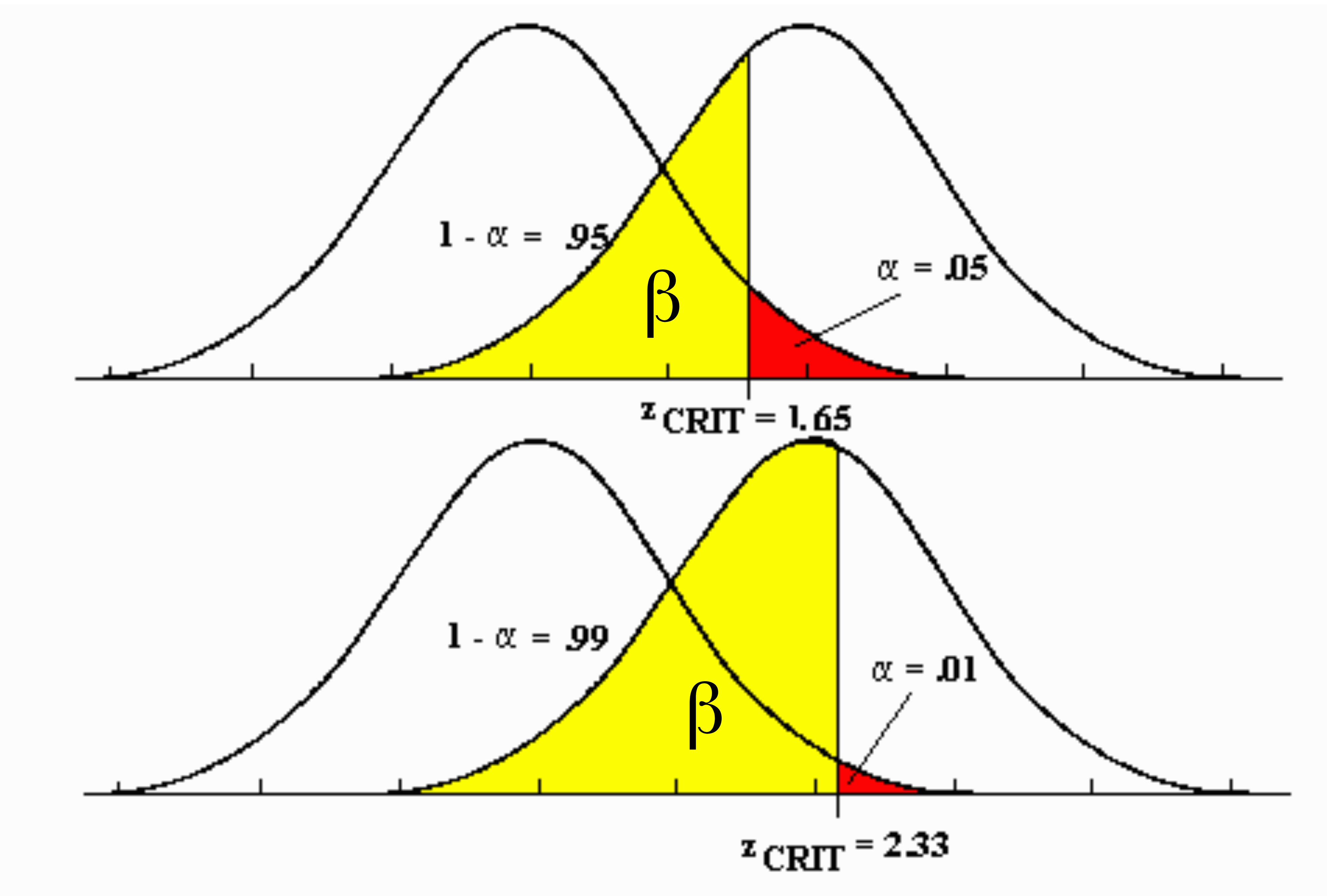
- In general there are an infinite number of possible critical regions that give the same significance level α
- The choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1
  - Roughly speaking, place the critical region where there is a low probability to be found if H0 is true, but high if H1 is true



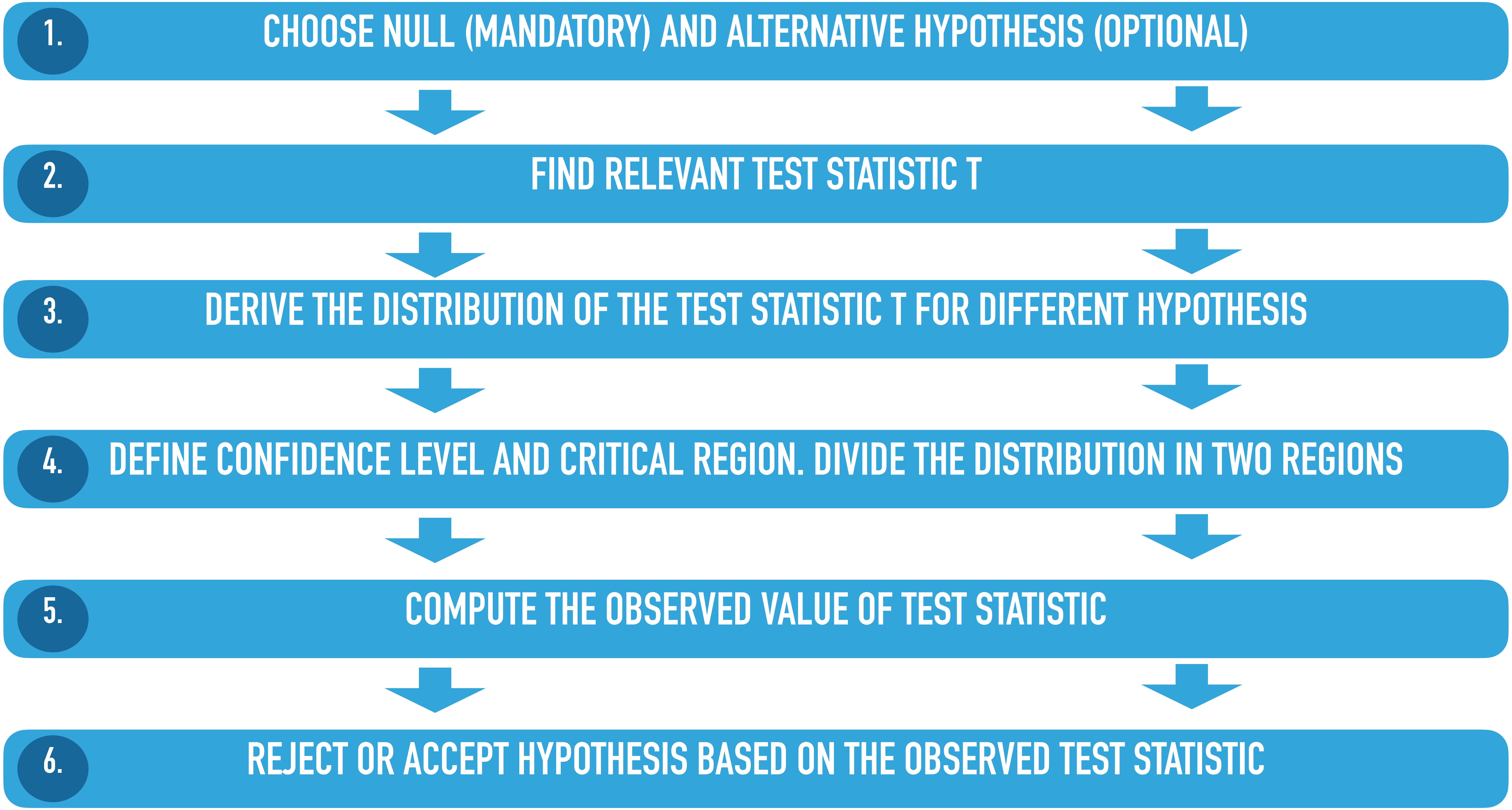
ERROR TYPES

		True state	
		H_0 is true	H_1 is true
Decision	Accept H_0	Right decision *Probability = 1-α* *(significance level)*	Wrong decison Type II error Probability = β
	Reject H_0	Wrong decision *Type I error* *Probability = α*	Right decision *Probability = 1-β* (power)





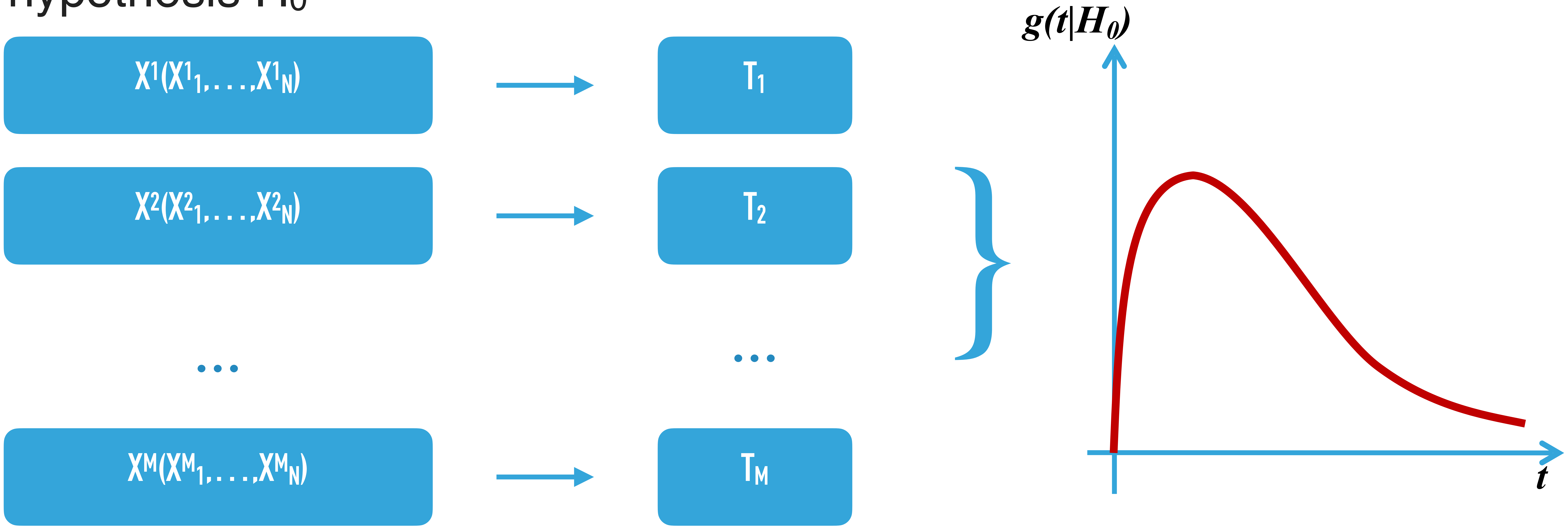
HYPOTHESIS TESTING PROCEDURE



1. CHOOSE NULL (MANDATORY) AND ALTERNATIVE HYPOTHESIS (OPTIONAL)
2. FIND RELEVANT TEST STATISTIC T
3. DERIVE THE DISTRIBUTION OF THE TEST STATISTIC T FOR DIFFERENT HYPOTHESIS
4. DEFINE CONFIDENCE LEVEL AND CRITICAL REGION. DIVIDE THE DISTRIBUTION IN TWO REGIONS
5. COMPUTE THE OBSERVED VALUE OF TEST STATISTIC
6. REJECT OR ACCEPT HYPOTHESIS BASED ON THE OBSERVED TEST STATISTIC

TEST STATISTIC PDF

- Using input data define a single **test statistic** $t(x_1,\ldots,x_N)$ whose value reflects the agreement between data and the hypothesis
- Using Monte Carlo simulate many (M) experiments trying to test the null hypothesis H_0

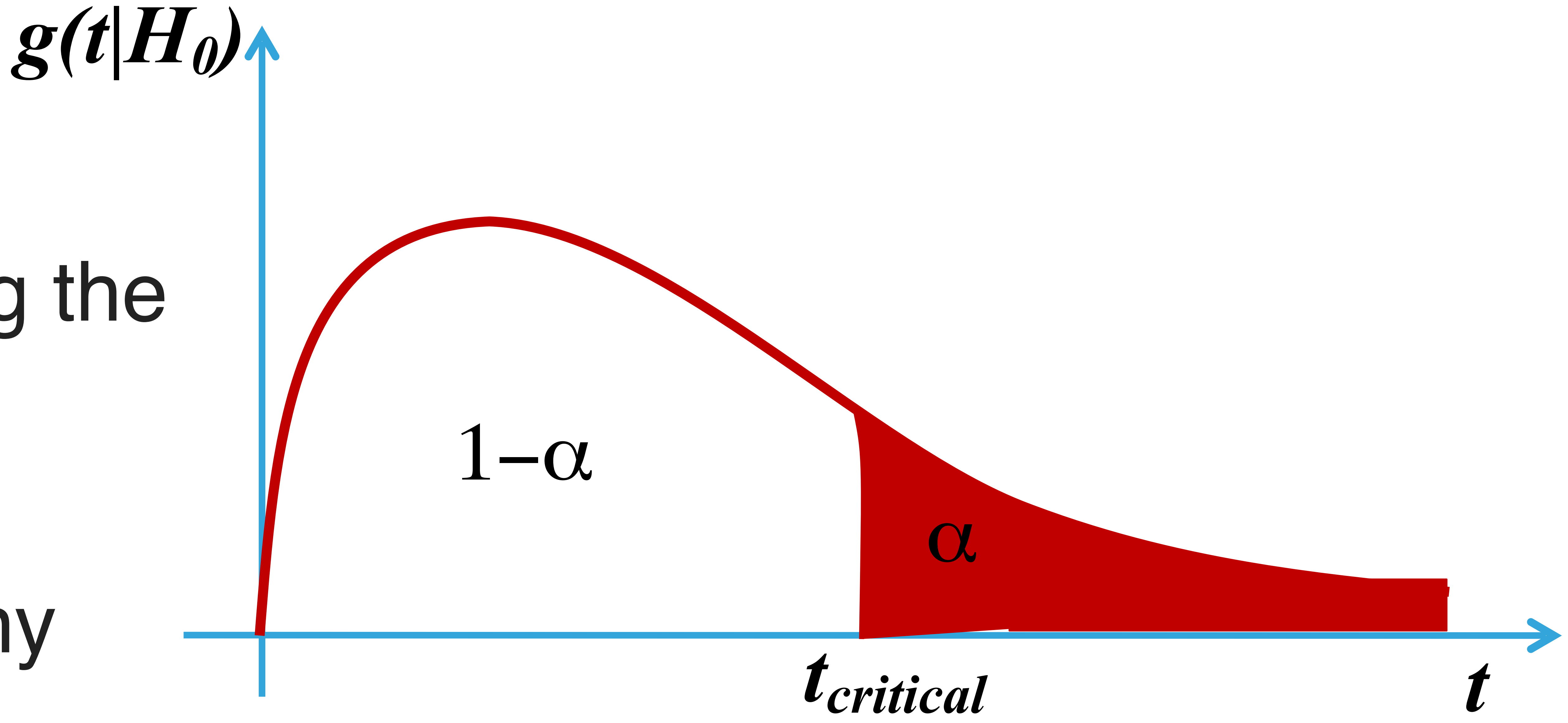


$X^1(X^1_1,\ldots,X^1_N)$
T_1
$X^2(X^2_1,\ldots,X^2_N)$
T_2
…
…
$X^M(X^M_1,\ldots,X^M_N)$
T_M
$g(t|H_0)$
t

- Obtain a probability density function (PDF) of the test statistic t, given null Hypothesis (H_0) is true, $g(t\,|\,H_0)$

OBSERVED TEST STATISTIC

- Now we have to divide the distribution in two regions:
  - where H_0 is rejected with CL α
  - where H_0 is not rejected with CL 1-α
- $t_{critical}$ is the value of test statistic diving the two regions
- We talk only about rejecting the null hypothesis H_0, not about accepting any other hypothesis

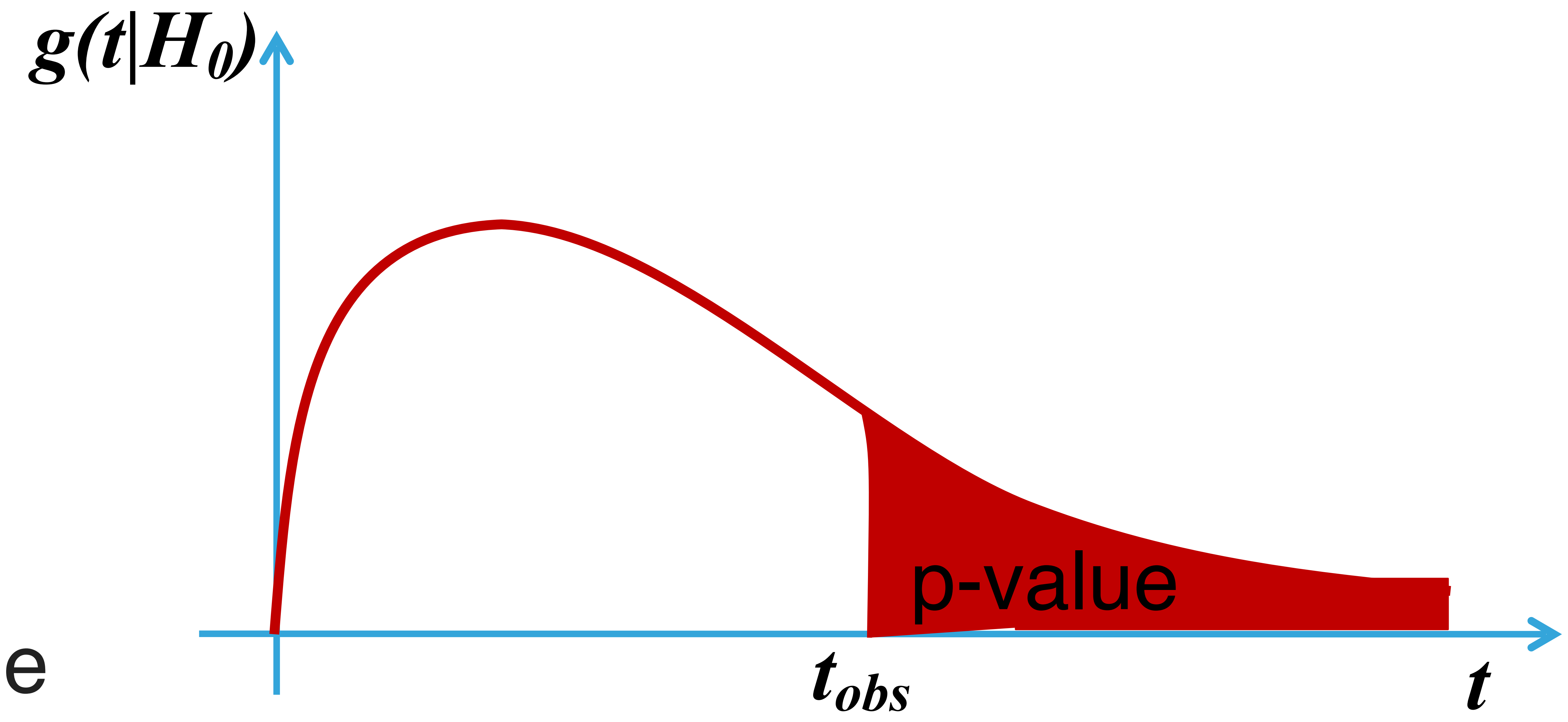


- **We should decide about two regions before looking at the observed value of the test statistics**
- Now we can calculate the observed test statistic t_{obs} and decide:
  - If $t_{obs} > t_{critical}$: reject H_0
  - If $t_{obs} < t_{critical}$: do not reject H_0

P - VALUE

- Knowing the PDF of our test statistic we can answer one important question:
- What is the probability to obtain the value of t equal or greater than the value t_{obs} we observed?

$$P(t \geq t_{obs}) = \int_{t_{obs}}^{\infty} g(t \mid H_0)dt$$



- This probability is the so-called p-value
- p-value is defined as the probability to find t in the region of equal and lesser compatibility with H_0 than the level of compatibility observed with actual data

SIGNIFICANCE

- For easier understanding p-values can be converted to **significance**

One tailed p-value	Significance	Gaussian area ±nσ	Probability of outcome: 1 in
0.159	1	0.68268949	6.3
0.023	2	0.95449974	44
0.00135	3	0.99730020	740
$3.17\cdot10^{-5}$	4	0.99993666	31,574
$2.87\cdot10^{-7}$	5	0.99999943	3,488,556

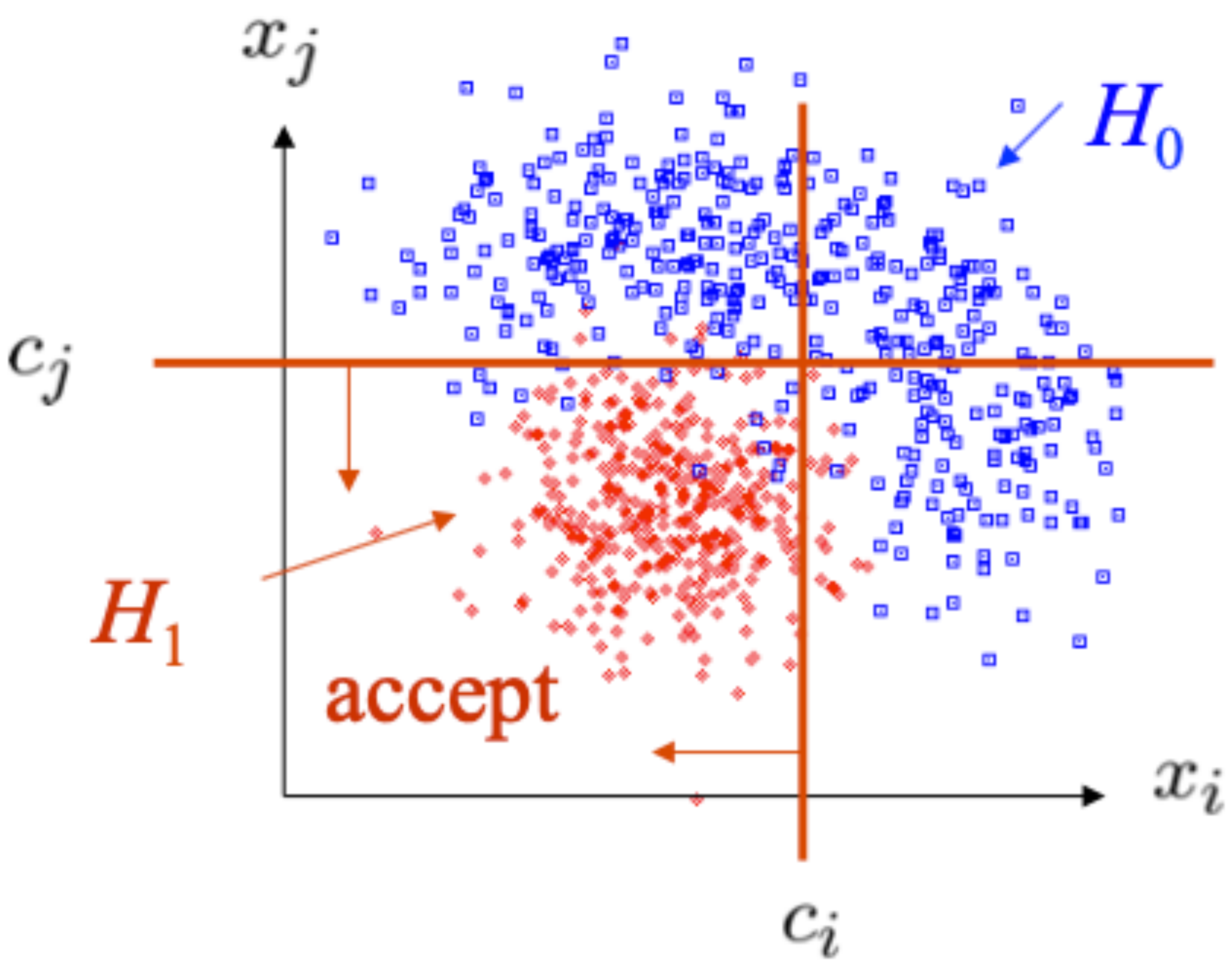
- For example: if you were to measure something with 5σ significance that means that either the null hypothesis is wrong (highly likely) or that due to statistical fluctuations your data sample corresponds to one in 3.5 million and the null hypothesis is correct (possible but extremely unlikely)

HYPOTHESIS TESTS EXAMPLE IN HEP

- Suppose the result of a measurement for an individual event is a collection of numbers x(x_1,…,x_N):
  - x_1 = number of muons
  - x_2 = mean pT of jets
  - x_3 = missing energy, ...
- x follows some N-dimensional joint PDF, which depends on the underlying particle process that produced final detected particles
- For each theory we consider we will have a hypothesis for the pdf of x, $f(x\,|\,H_0), f(x\,|\,H_1), \ldots$
  - We call H_0 the background hypothesis (the event type we want to reject) and it can be for example a hypothesis that particles are produced if SM is valid
  - H_1 is signal hypothesis (the type we want), and i this example it can be a hypothesis that particles are produced if SUSY is valid

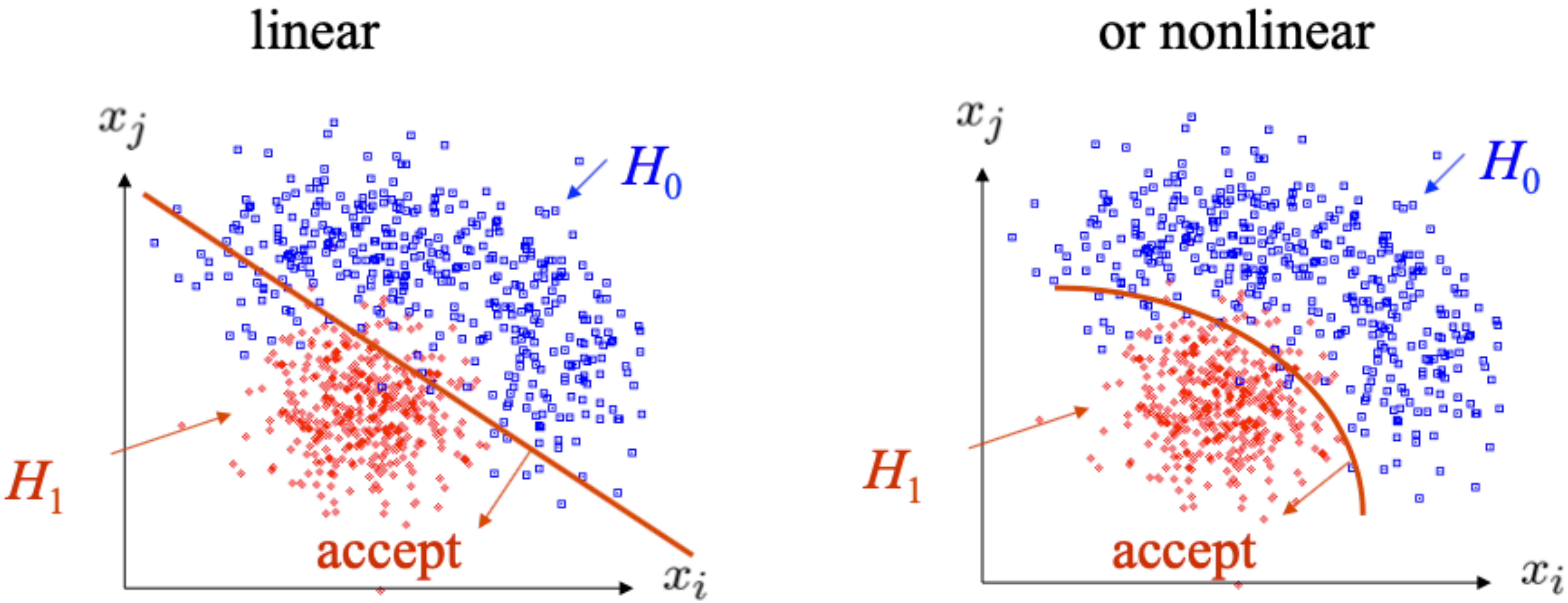
HYPOTHESIS TESTS IN HEP

- Suppose we have a data sample with two kinds of events, corresponding to hypotheses H_0 and H_1 and we want to select those of type H_1
- We can use Monte Carlo simulation to simulate events according to both hypothesis to better understand what are the similarities and differences and to understand how to define the test statistic
- How can we use Monte Carlo simulation to decide for what observed data we are going to accept/reject null/alternative hypothesis?
- Perhaps if events pass ‘cuts’:
  - $x_i < c_i$
  - $x_j < c_j$



HYPOTHESIS TESTS IN HEP

- Or maybe use some other sort of decision boundary:



linear

or nonlinear

- The big natural question that arises: Can we do this in an **optimal way**?

NEYMAN-PEARSON LEMMA

- How can we choose a test's critical region in an 'optimal way'?
- The performance of a selection criterion can be considered optimal if it achieves the smallest misidentification probability for a desired value of the selection efficiency
- A test statistic that ensures the optimal performance in this sense is provided by the **Neyman–Pearson lemma**:
- Optimal test statistic is defined as the **ratio of the likelihood functions** evaluated for the observed data sample x under the two hypotheses H_0 and H_1:

$$t(x) = \frac{L(x \mid H_1)}{L(x \mid H_0)} > c$$

- where c should be set in order to achieve the required Confidence Level (CL)

LIKELIHOOD RATIO DISCRIMINANT

- If the N variables $x(x_1, \ldots, x_N)$ that characterise our problem are independent, the likelihood function can be factorised into the product of 1D marginal PDFs:

$$t(x) = \frac{L(x(x_1, \ldots, x_N) \,|\, H_1)}{L(x(x_1, \ldots, x_N) \,|\, H_0)} = \frac{\prod_{i=1}^{N} L(x_i \,|\, H_1)}{\prod_{i=1}^{N} L(x_i \,|\, H_0)}$$

- This allows in many cases to simplify the computation
- Even if it is not possible to factorise the PDFs into the product of 1D marginal PDFs (i.e. the variables are not independent), the product can still be used as a discriminant
  - will differ from the exact likelihood ratio and hence it will correspond to worse performance
  - the simplicity of this method can justify its application in spite of the suboptimal performances

LIKELIHOOD RATIO AND ML

- The optimal test statistic for hypothesis testing:
  - $t(x) = \dfrac{L(x \,|\, H_1)}{L(x \,|\, H_0)} > c$
- Very often unobtainable in real-life cases. One of the biggest problem is that x^N is very often multidimensional (N >> 1)
- Machine Learning is our effort to approximate the likelihood ratio (LR)
  - One of key ideas is to build an algorithm that can “learn” the likelihood from training data and then apply the LR test statistic to distinguish data from different hypothesis with (close to) optimal performance

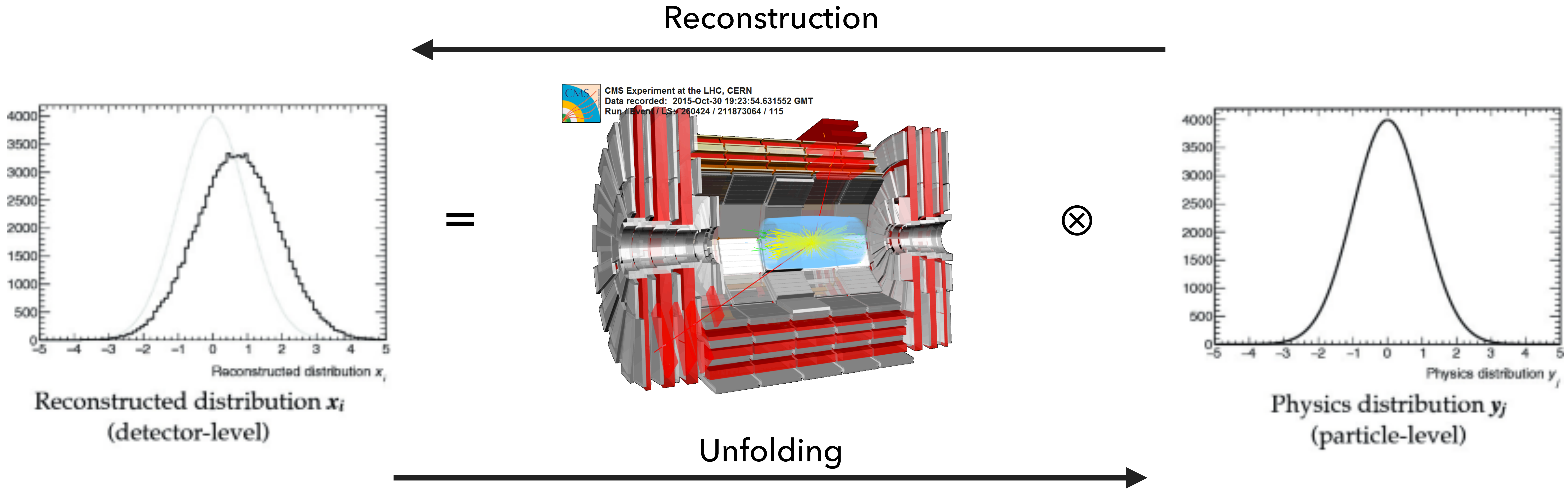
BLACK BOX

Machine Learning
AI
BDT
Neural Networks

1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
f(x|H₀)
critical region W
f(x|H₁)
x

UNFOLDING

- The **distortions** to distributions occur when the values of measured variables are subject to random fluctuations due to the limited resolution of the measuring device
- The procedure of correcting for these distortions is known as **unfolding**
  - Has applications in optical image reconstruction, radio astronomy, crystallography, medical imaging, particle physics, and many others…



Reconstruction

Reconstructed distribution x_i (detector-level) = ⊗ Physics distribution y_j (particle-level)

Unfolding

4000
3500
3000
2500
2000
1500
1000
500
0
−5 −4 −3 −2 −1 0 1 2 3 4 5
Reconstructed distribution x_i
CMS Experiment at the LHC, CERN
Data recorded: 2015-Oct-30 19:23:54.631552 GMT
Run / Event / LS: 260424 / 211873064 / 115
Physics distribution y_j

PROBLEM FORMULATION

- Consider you want to measure Higgs boson differential cross section as a function of its pseudorapidity:
  - you are measuring the number of Higgs boson candidates produced in different pseudorapidity bins
  - due to imperfect detector you will get a distorted shape
  - some events will migrate to different bins
  - some events won’t be reconstructed
  - some events without Higgs will be counted in
- In order to compare the result with theoretical prediction or with other experiments we need to somehow revert
- Several possible ways to do it

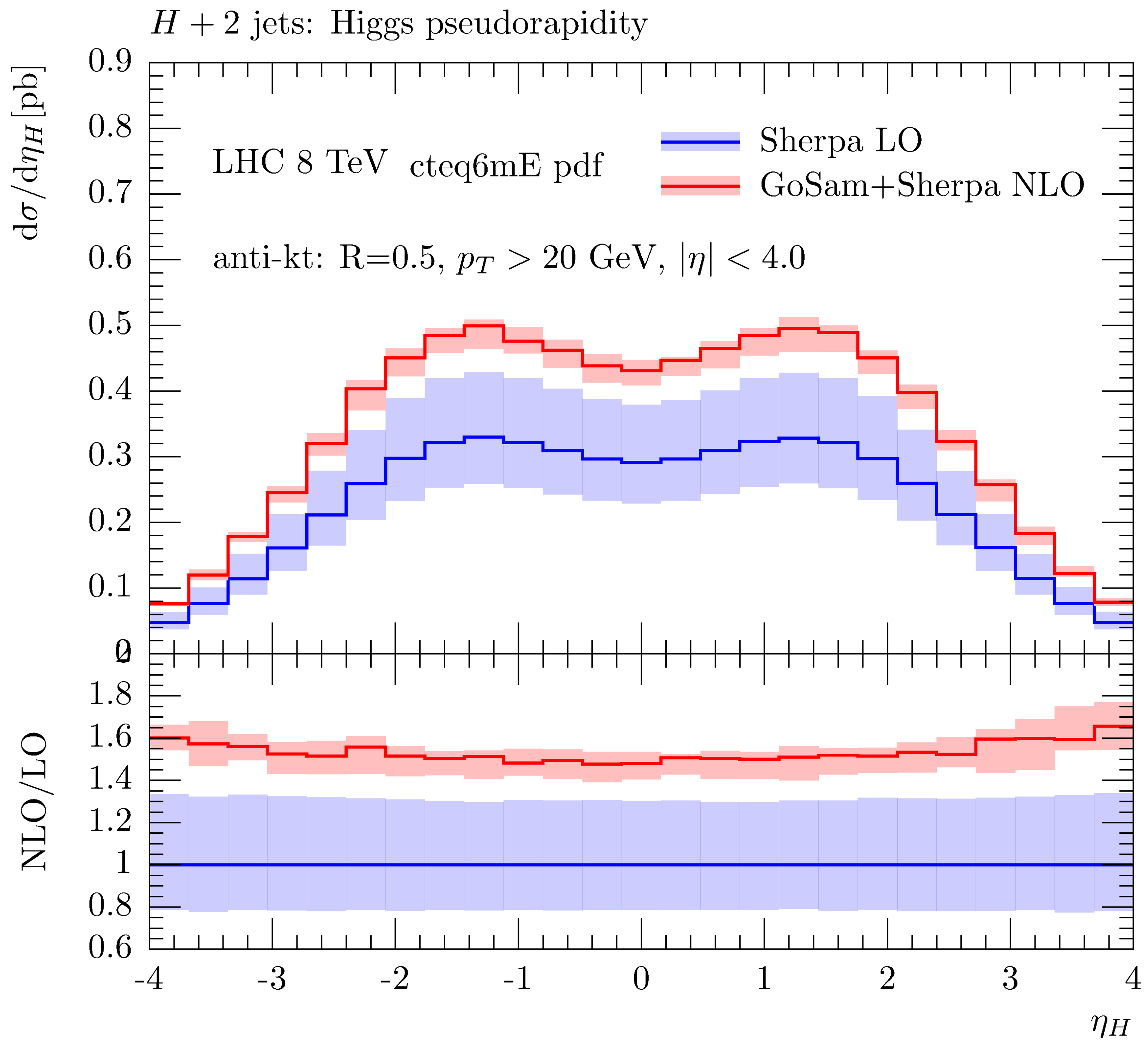


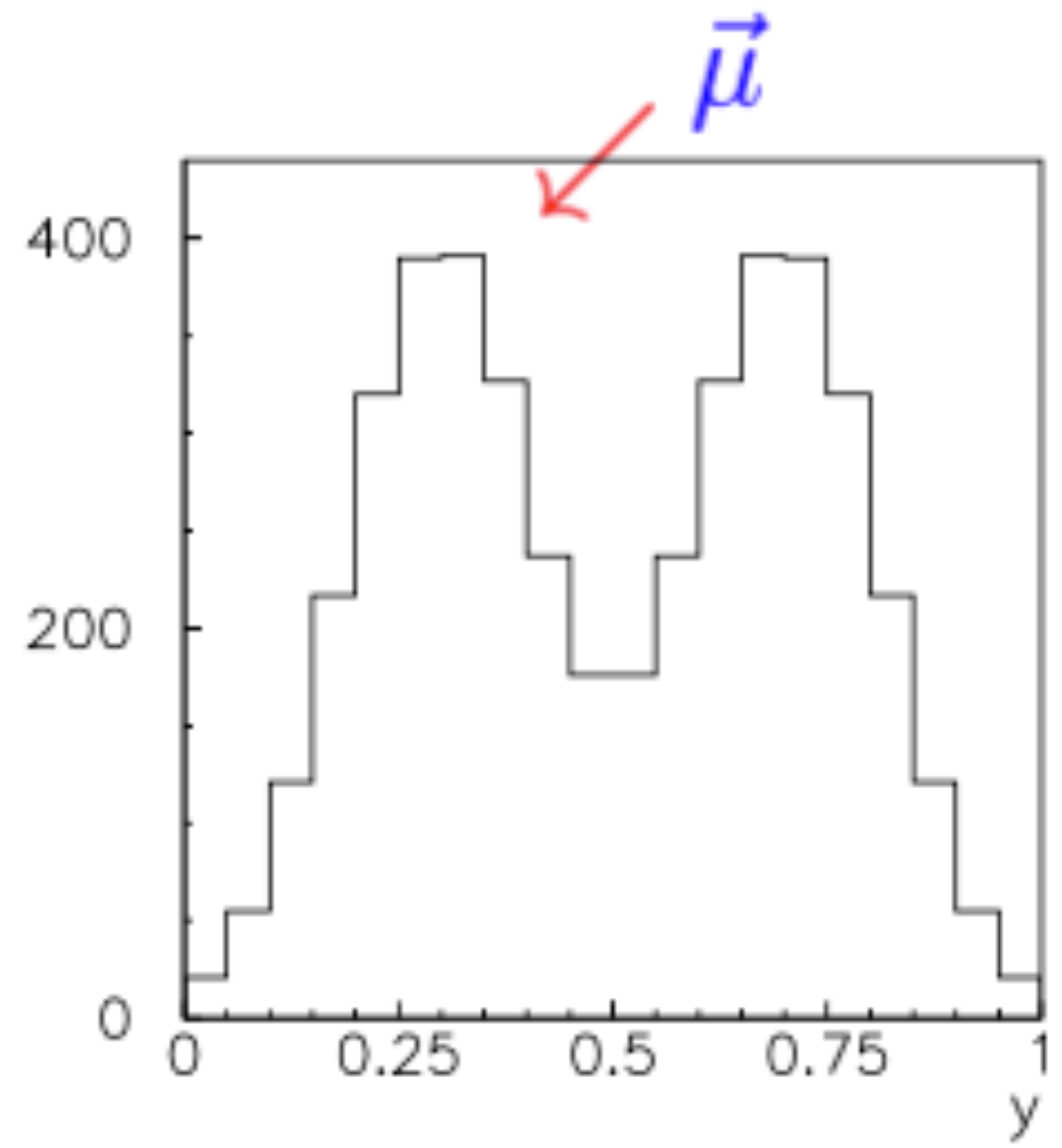
Figure 2: Pseudorapidity η of the Higgs boson.

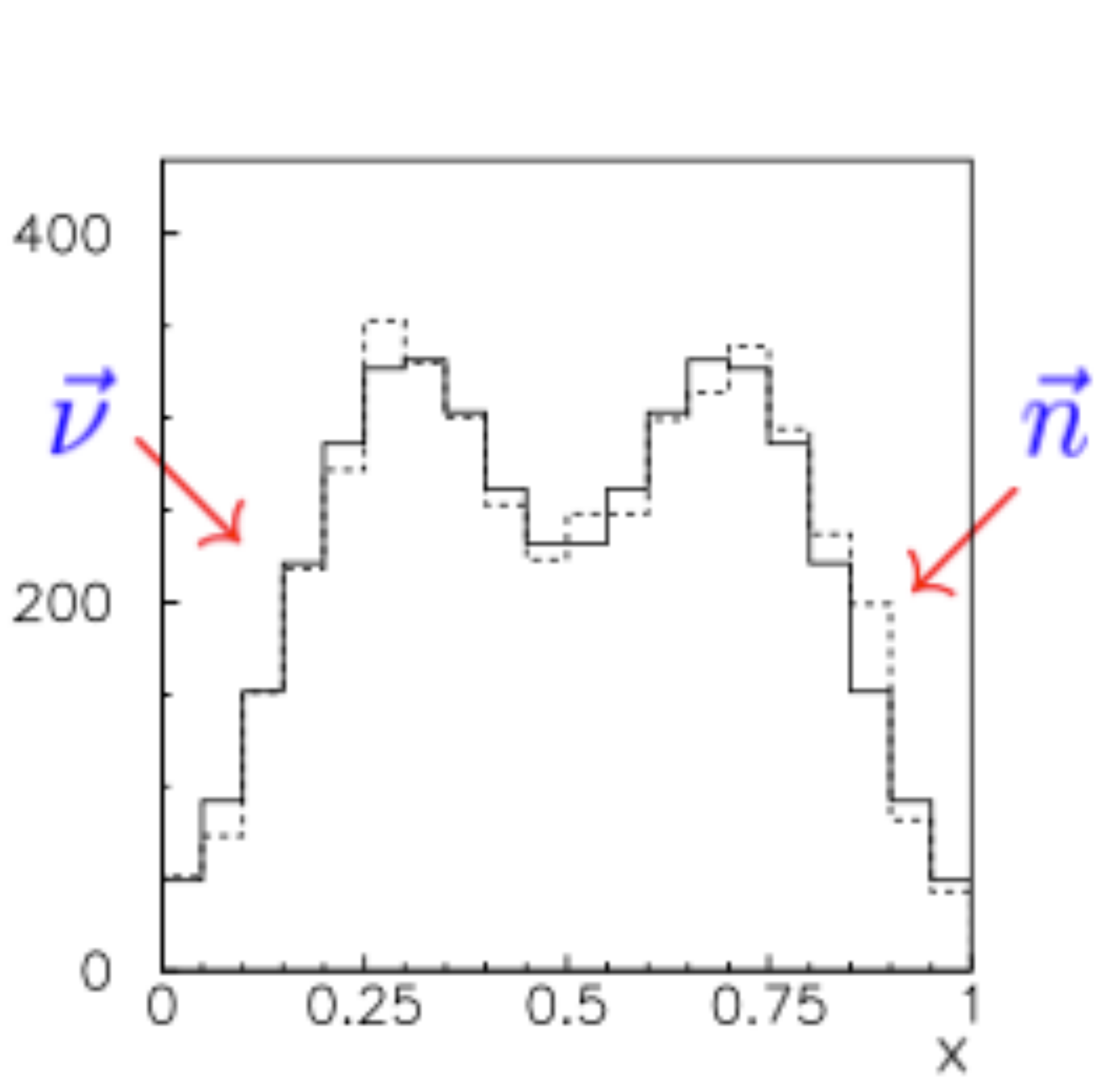
RESPONSE MATRIX

- We define a **response matrix** R_{ij} as:

$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j \ , \qquad R_{ij} = P(\text{observed in bin } i \,|\, \text{true value in bin } j)$$

- where ν_i is the expected observed value, and μ_j is the true value.
- We have observed data $\vec{n}(n_1, \ldots, n_N)$





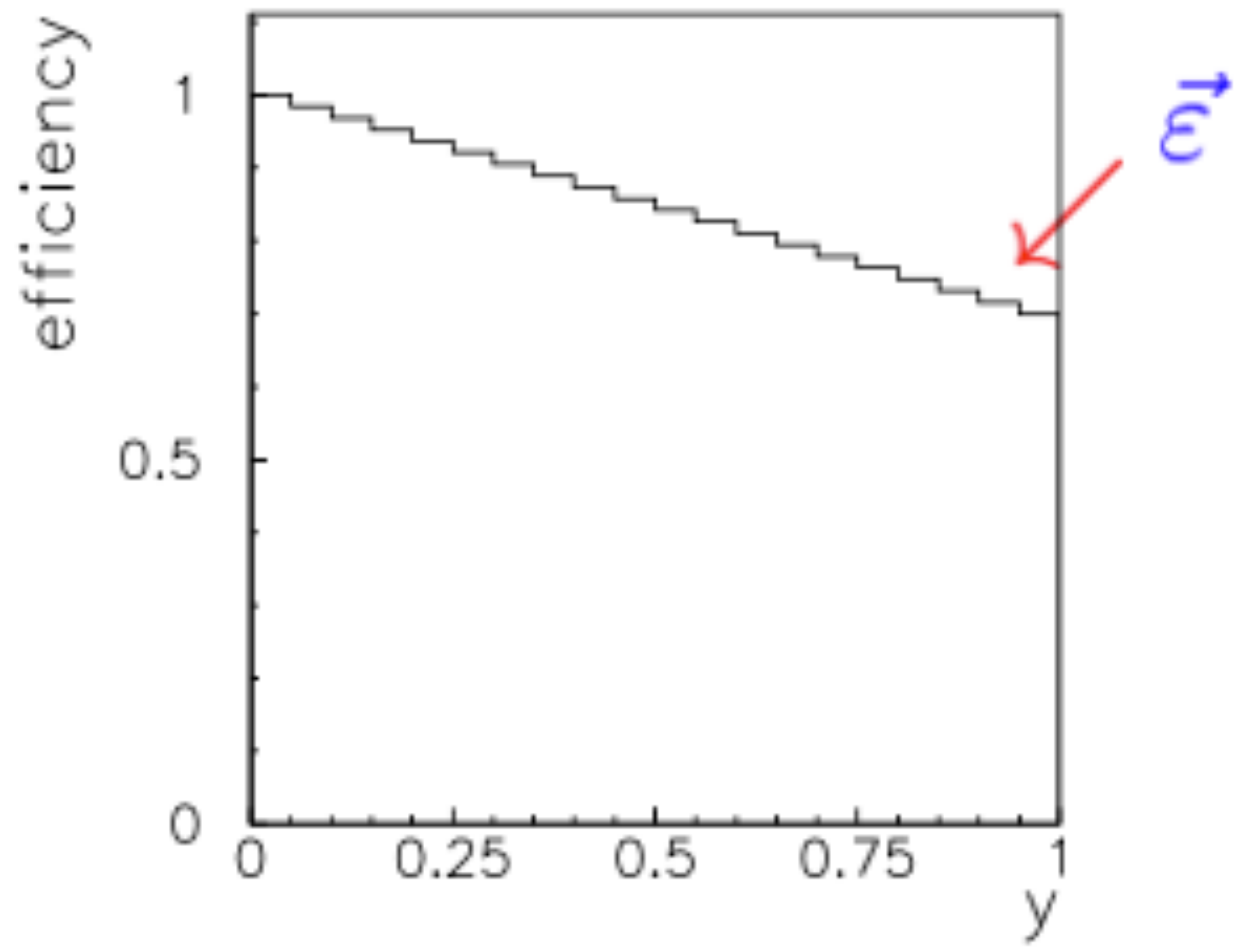
EFFICIENCY AND BACKGROUND

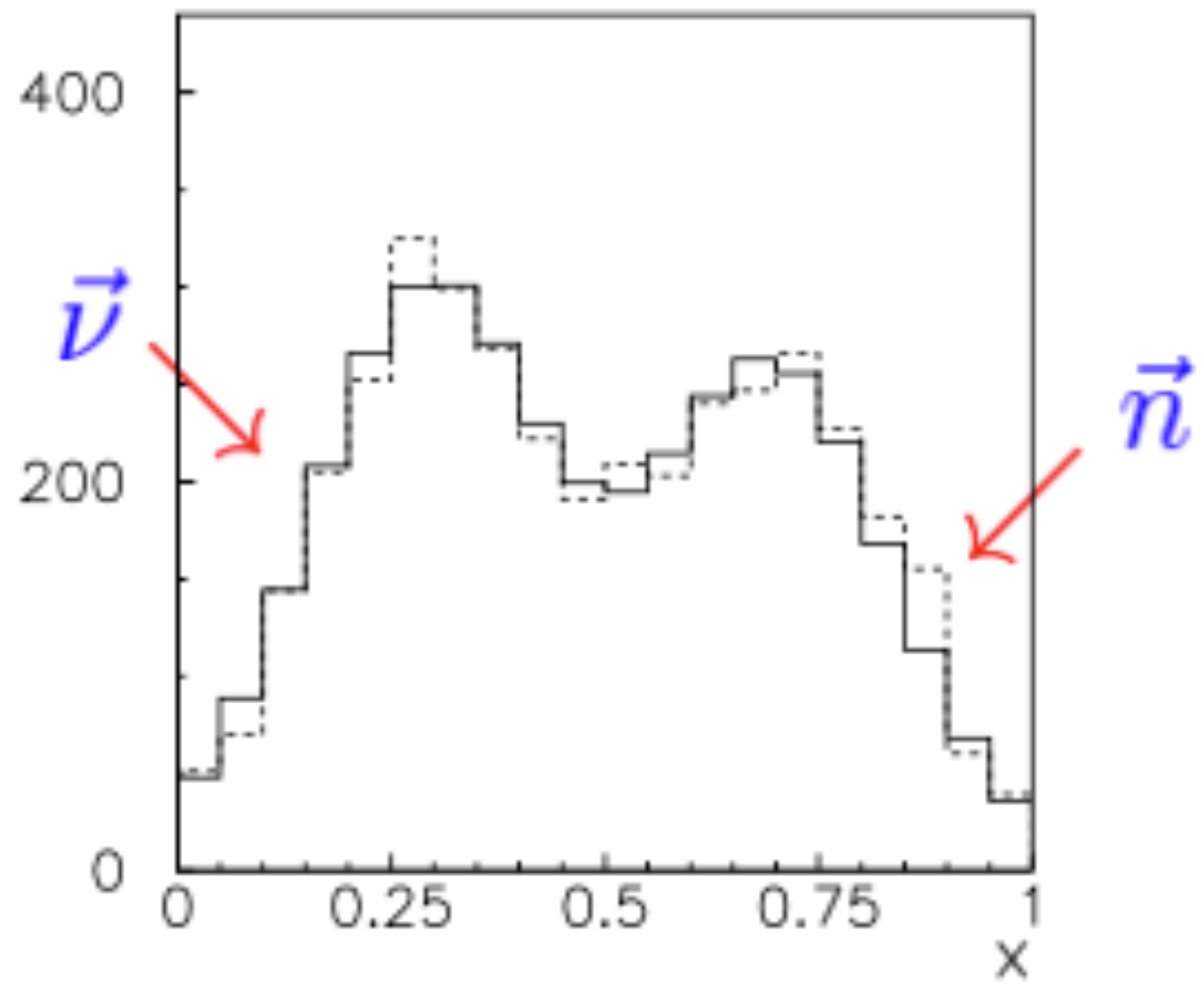
- To account for undetected events we introduce efficiency:

$$\sum_{i=1}^{N} R_{ij} = P(\text{observed anywhere}\,|\,\text{true value in bin } j) = \epsilon_j$$

- To account for observed events that come from background processes:

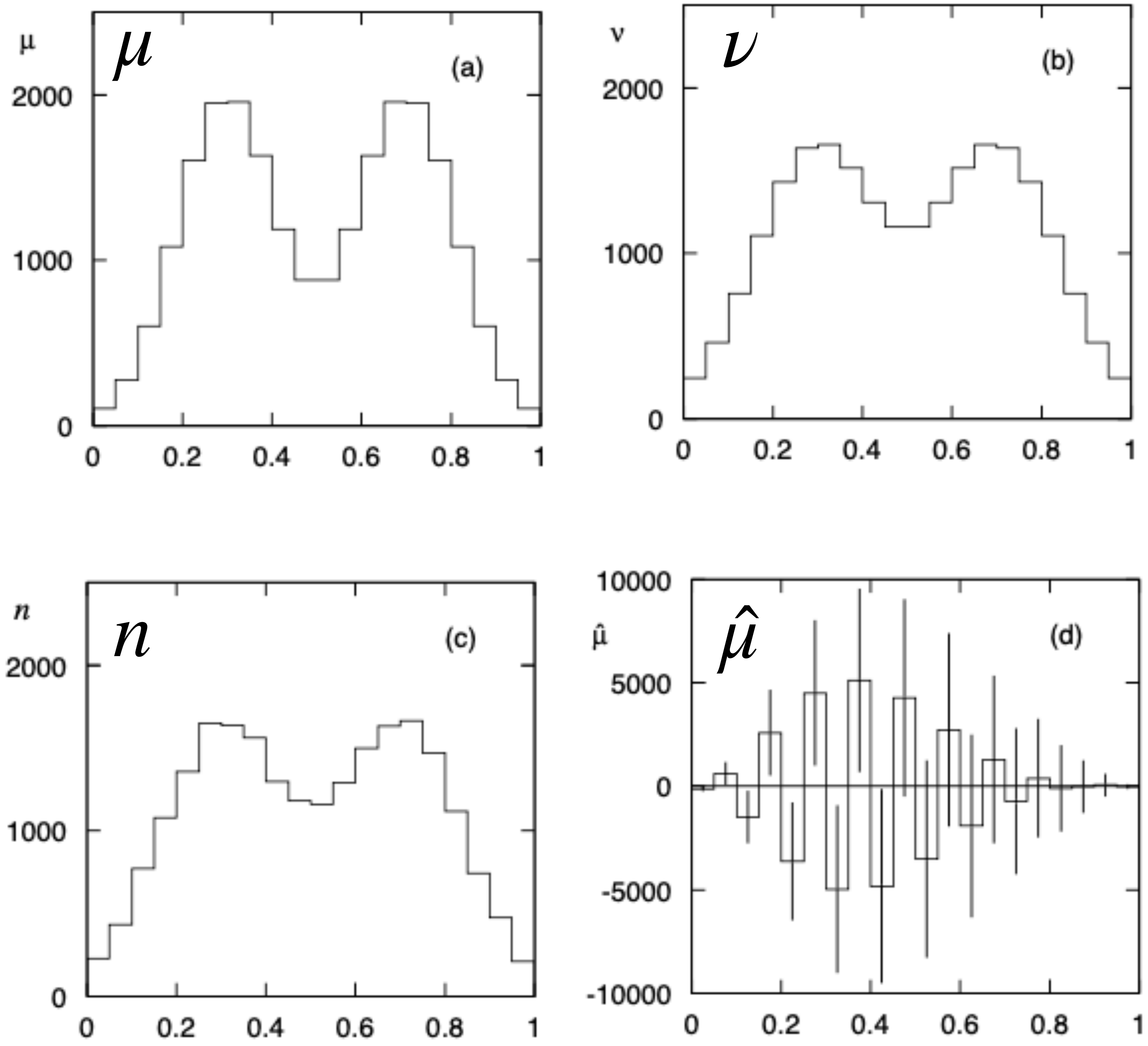
$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + b_i$$





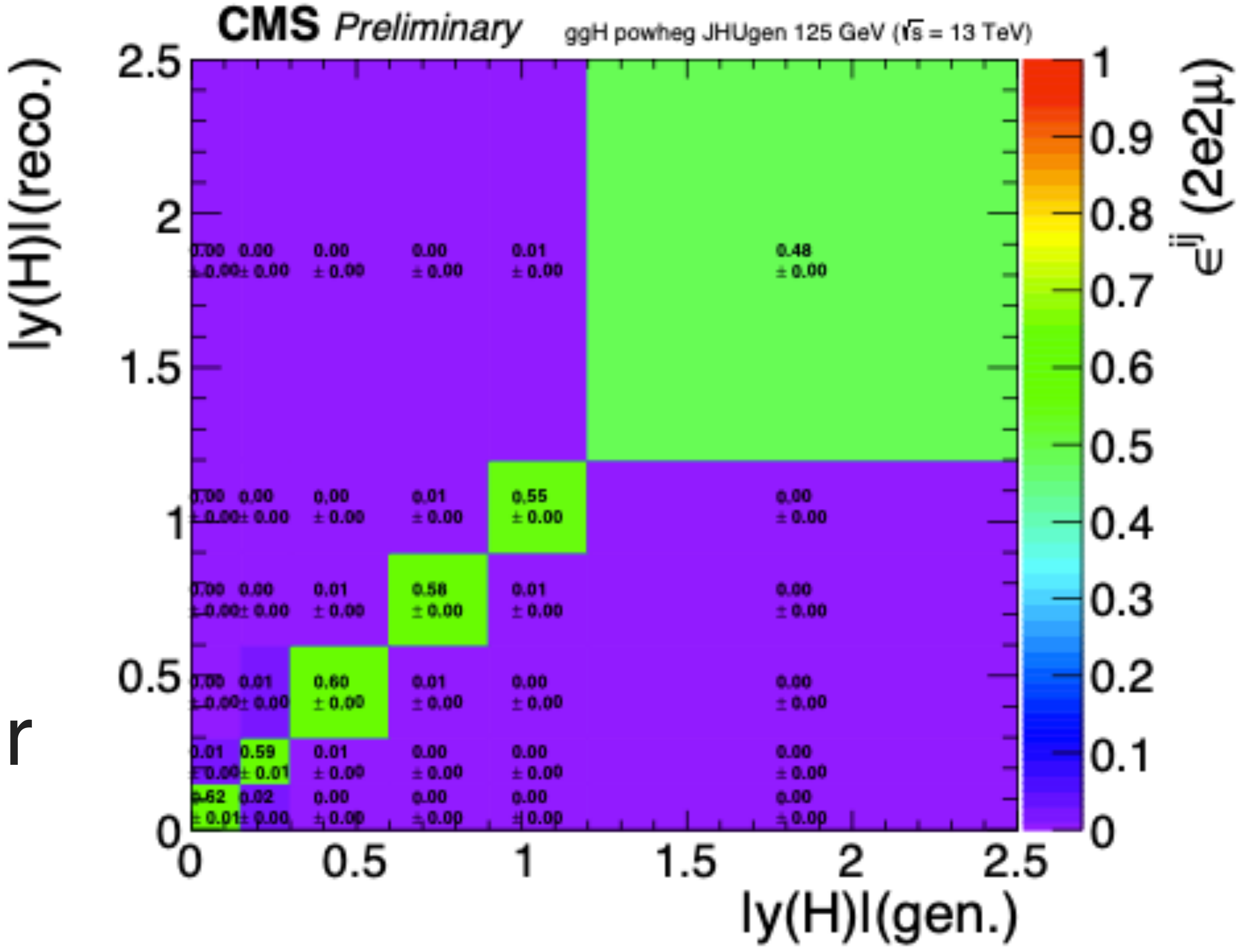
INVERTING THE RESPONSE MATRIX

- An obvious method for unfolding is to invert the matrix $\mu = R^{-1}(\nu - b)$ with an obvious choice for estimators $\hat{\mu} = R^{-1}(\hat{\nu} - b) \approx R^{-1}(n - b)$
- This estimator also comes from ML method:
  - $$\ln L(\mu) = \sum_{i}^{N} \ln P(n_i | \nu_i)$$
- In a simple example without background this method has a catastrophic failure
  - due to data being random variables and hence subject to statistical fluctuations this method is unreliable ($\hat{\nu} \neq n$)
- Method is mathematically correct but can lead to useless results
  - the idea is to improve the method uncertainty by sacrificing the correctness a bit (just like in the case of the Neyman-Pearson Lemma and LR)



THE METHOD OF CORRECTION FACTORS

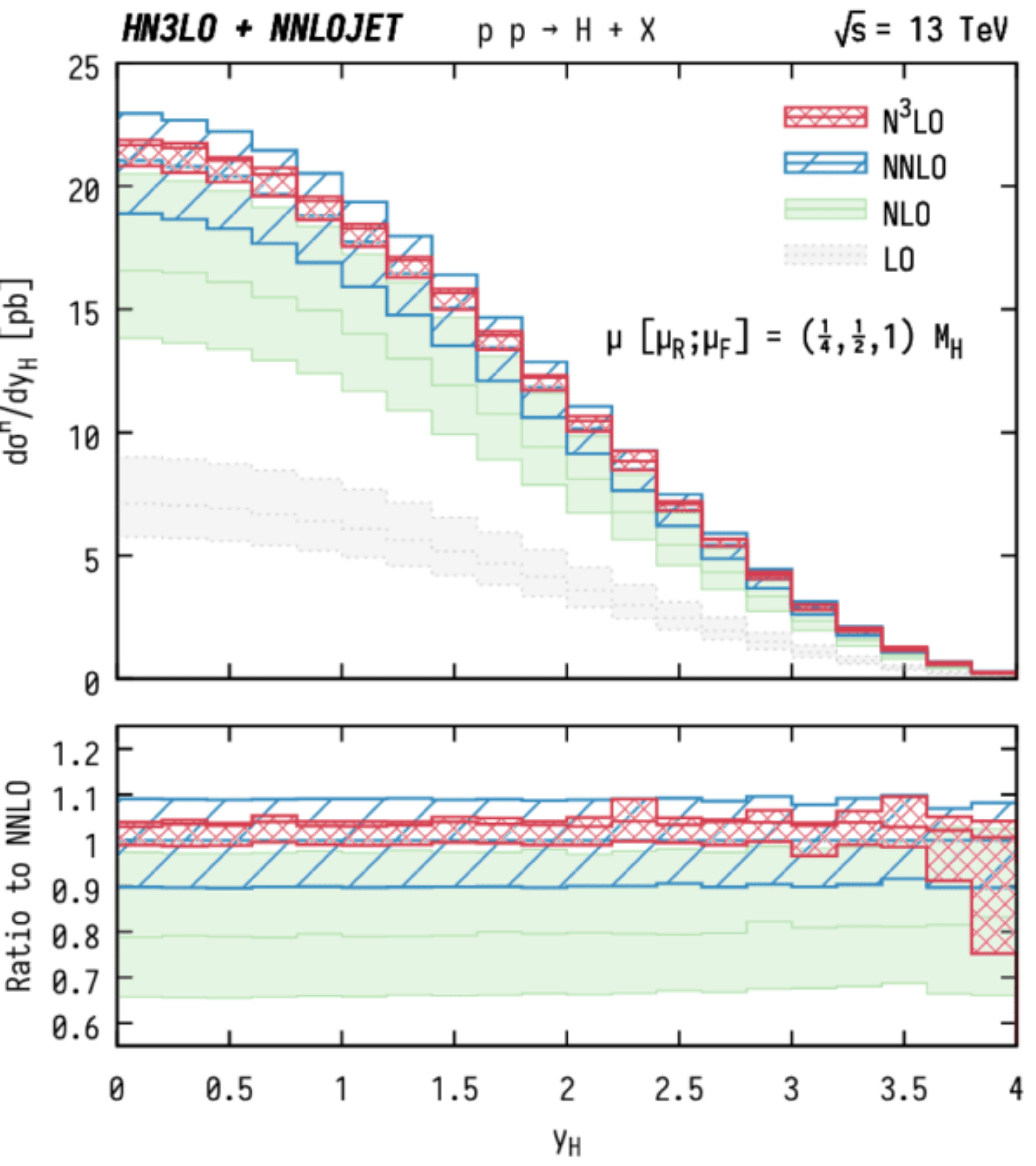
- There are several ways of improving the unfolding by matrix inversion
- Correction factors are defined as $\hat{\mu}_i = C_i(n_i - b_i)$
- The correction factors are determined by running the Monte Carlo program once with and once without the detector simulation, yielding model predictions for the observed and true values of each bin
  - $C_i = \dfrac{\mu_i^{MC}}{\nu_i^{MC}}$
  - statistical errors in the correction factors are negligible if it is possible to generate enough Monte Carlo data
- Problem: results are model dependent (because you are unfolding under the assumption that your Monte Carlo is the right description of nature)

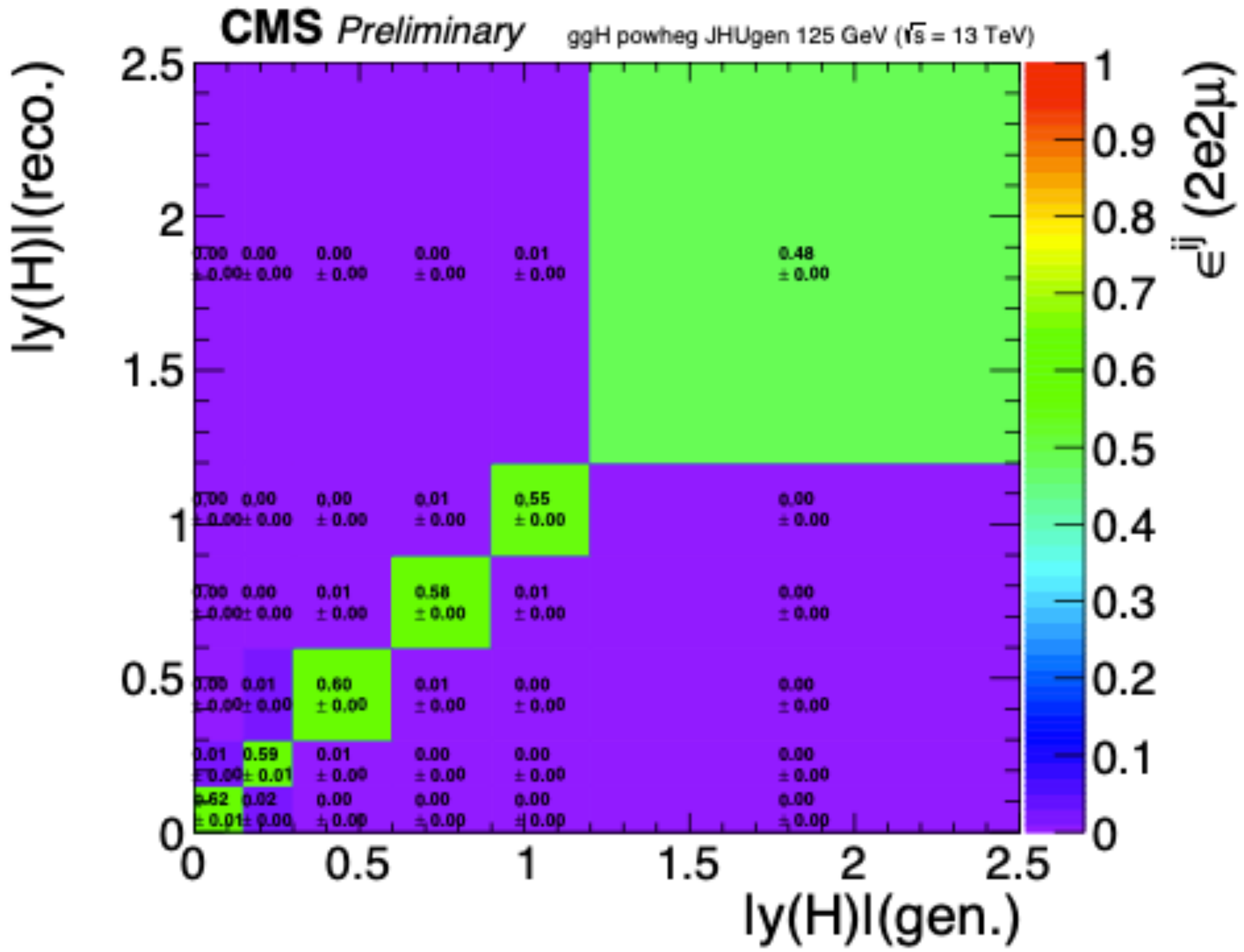


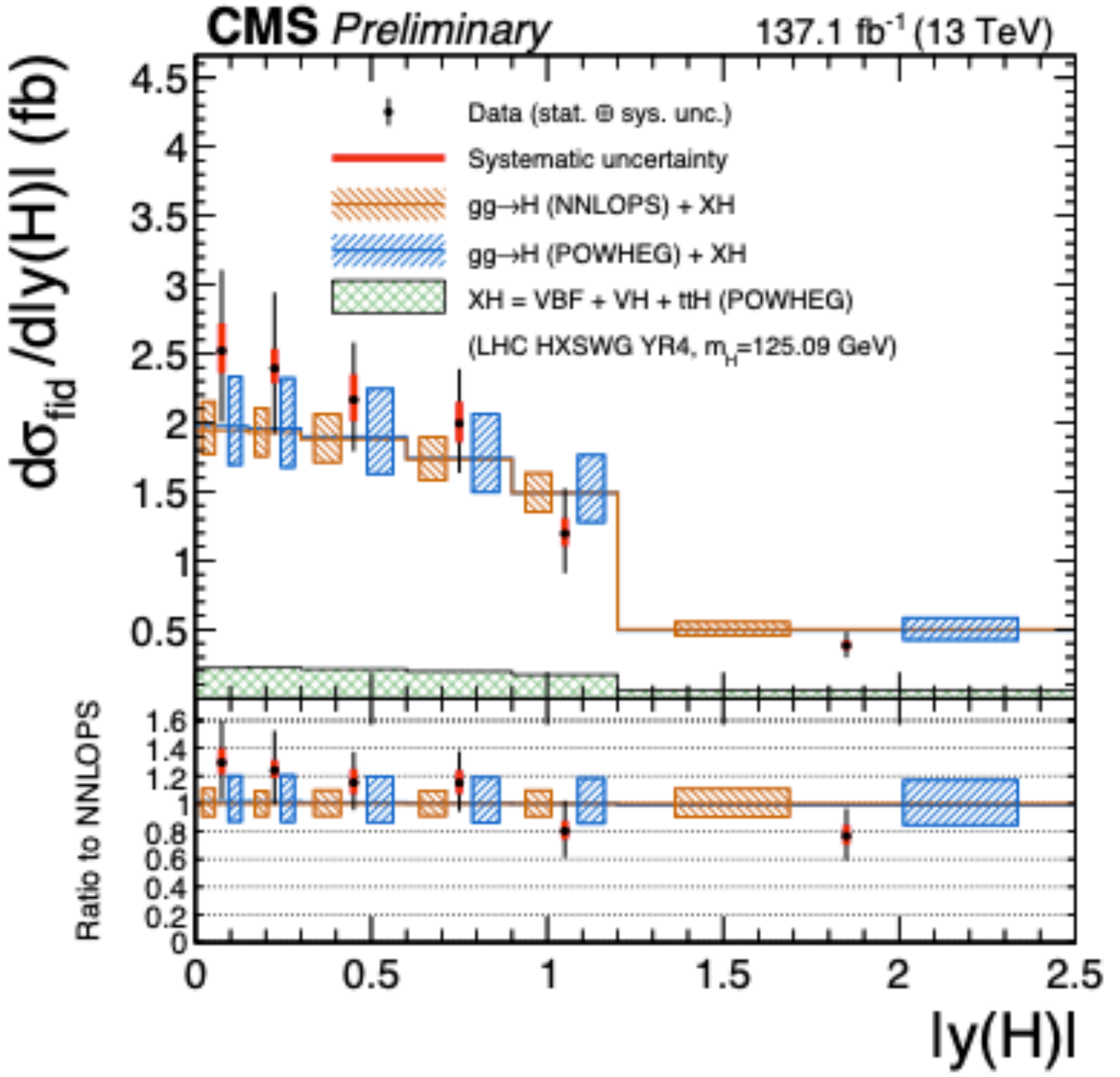
REGULARISED UNFOLDING

- An alternative approach is to impose in some way a measure of smoothness on the estimators for the true histogram. This is known as **regularisation** of the unfolded distribution.
- One considers some region around the Likelihood maximum
  - $\ln L(\mu) \geq \ln L_{max} - \Delta \ln L$
- We define a measure of smoothness by introducing a regularisation function $S(\mu)$
- Goal is to maximize $S(\mu)$ with the constraint that likelihood remains in the nearby maximum region that we defined. This is equivalent to maximizing:
  - $\tilde{L}(\mu, \alpha) = \alpha[\ln L(\mu) - (\ln L_{\max} - \Delta \ln L)] + S(\mu)$
  - $\Phi(\mu) = \alpha \ln L(\mu) + S(\mu)$
- Many different choices of regularisation functions $S(\mu)$

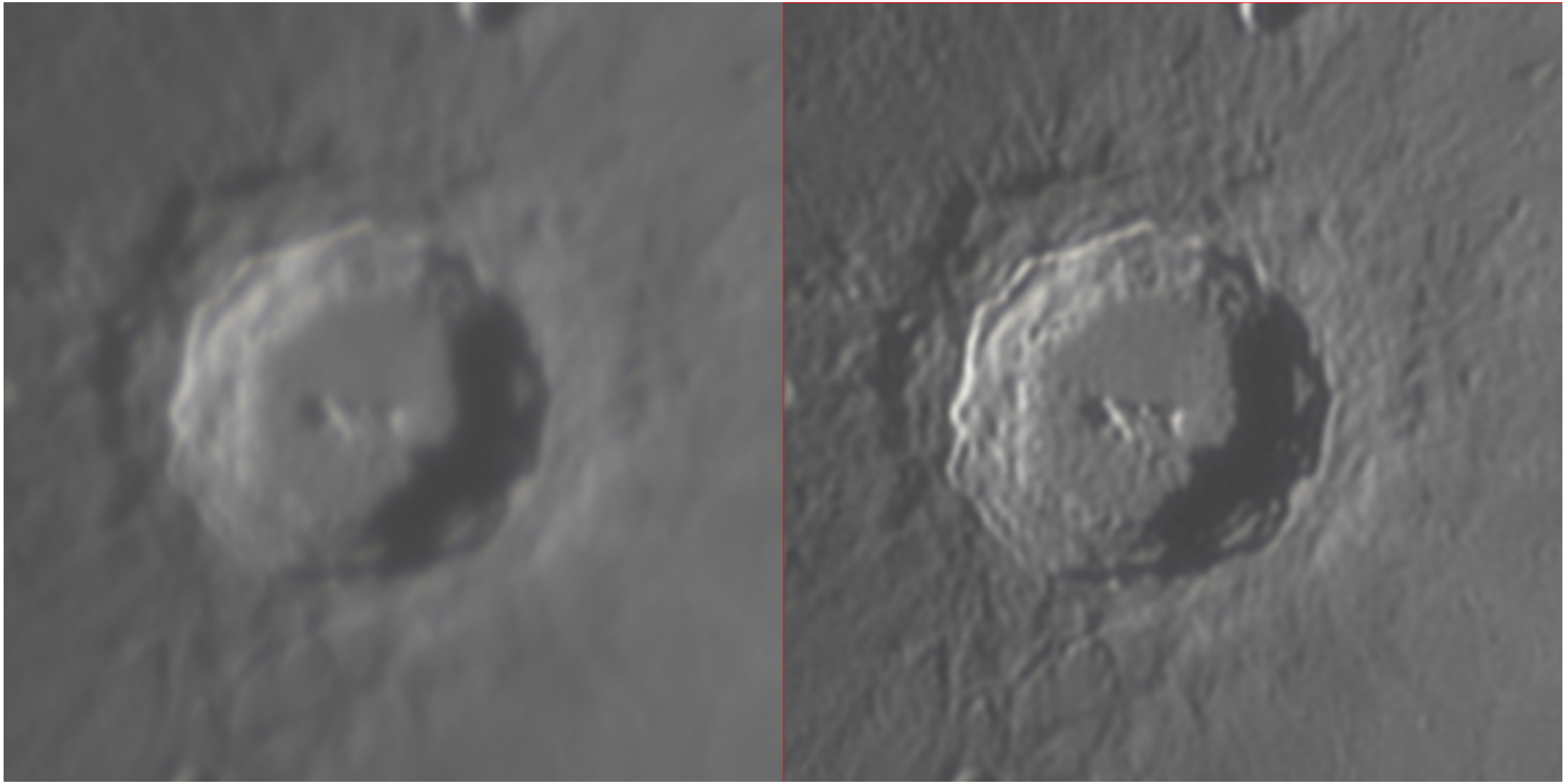
UNFOLDING EXAMPLE - HIGGS BOSON



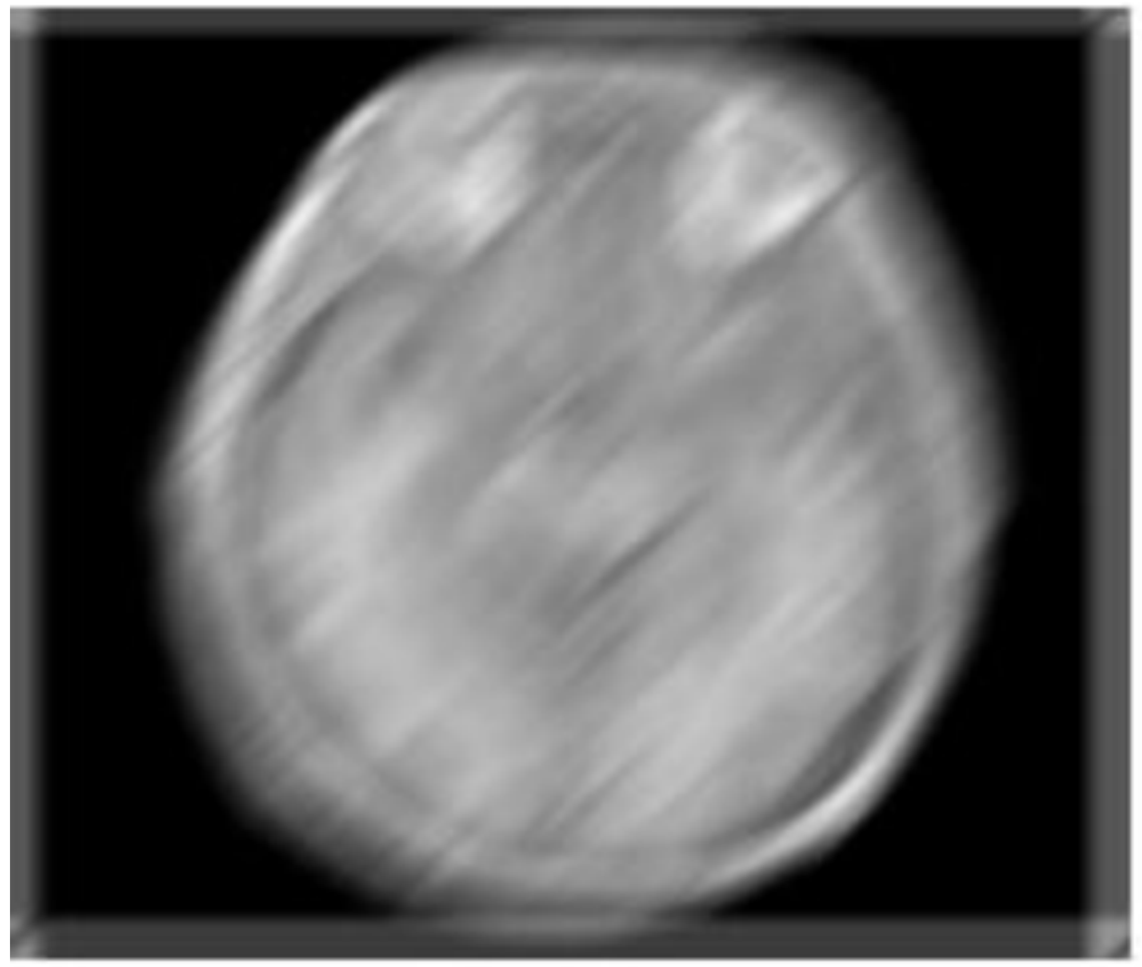




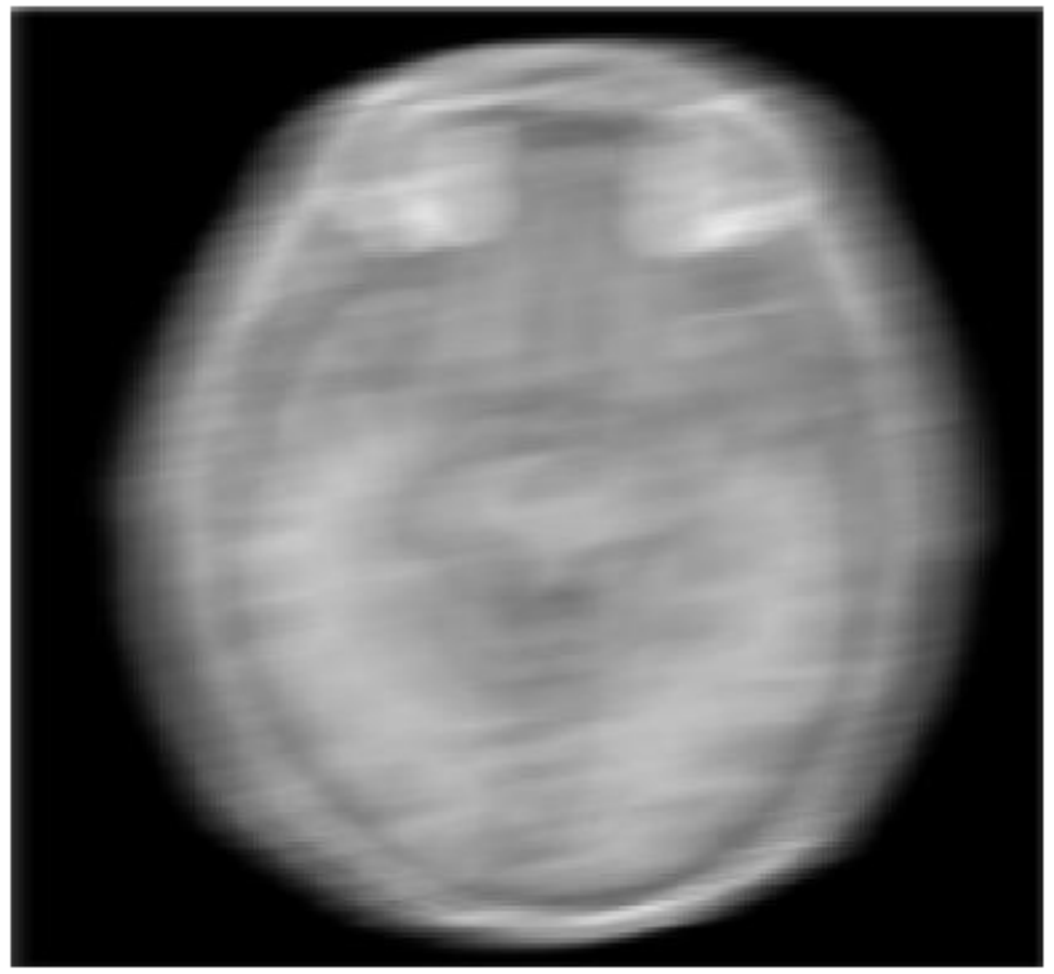
UNFOLDING EXAMPLE - IMAGING



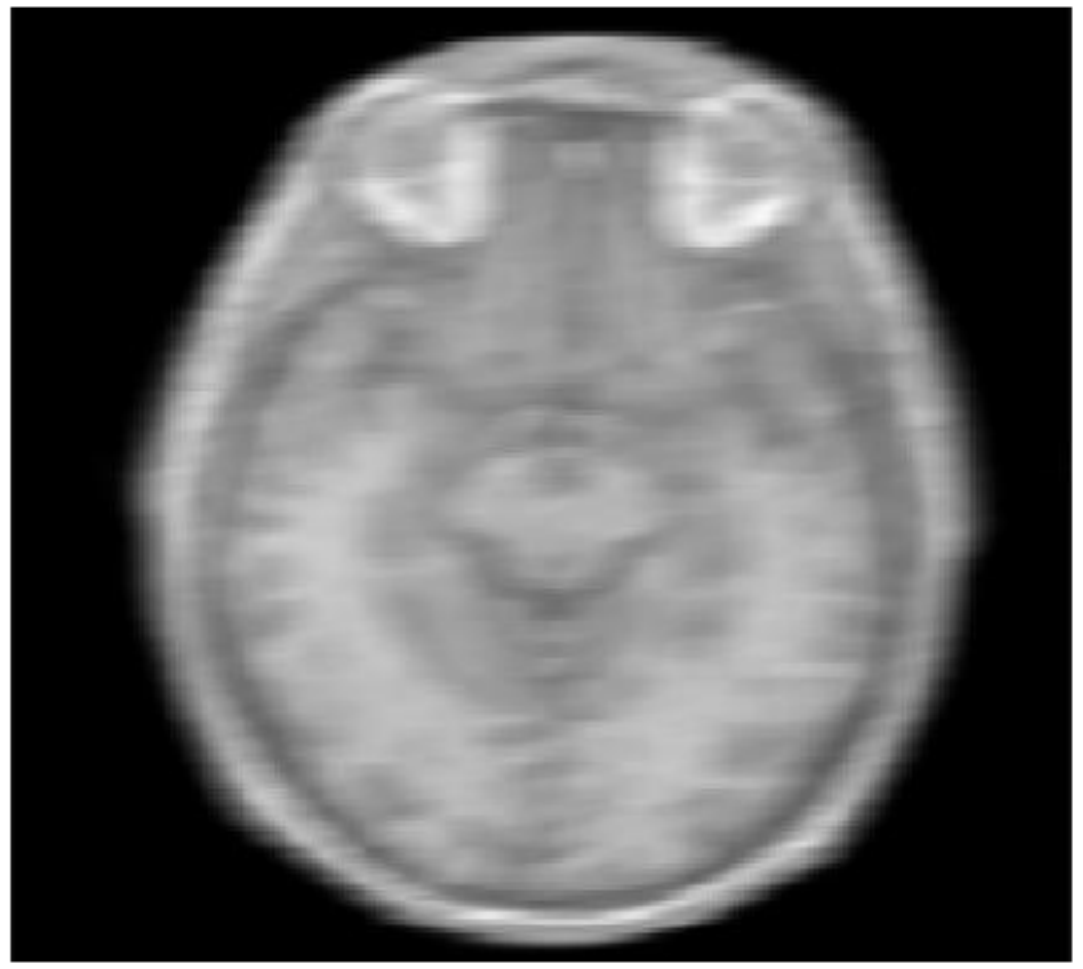
UNFOLDING EXAMPLE - MEDICAL IMAGING



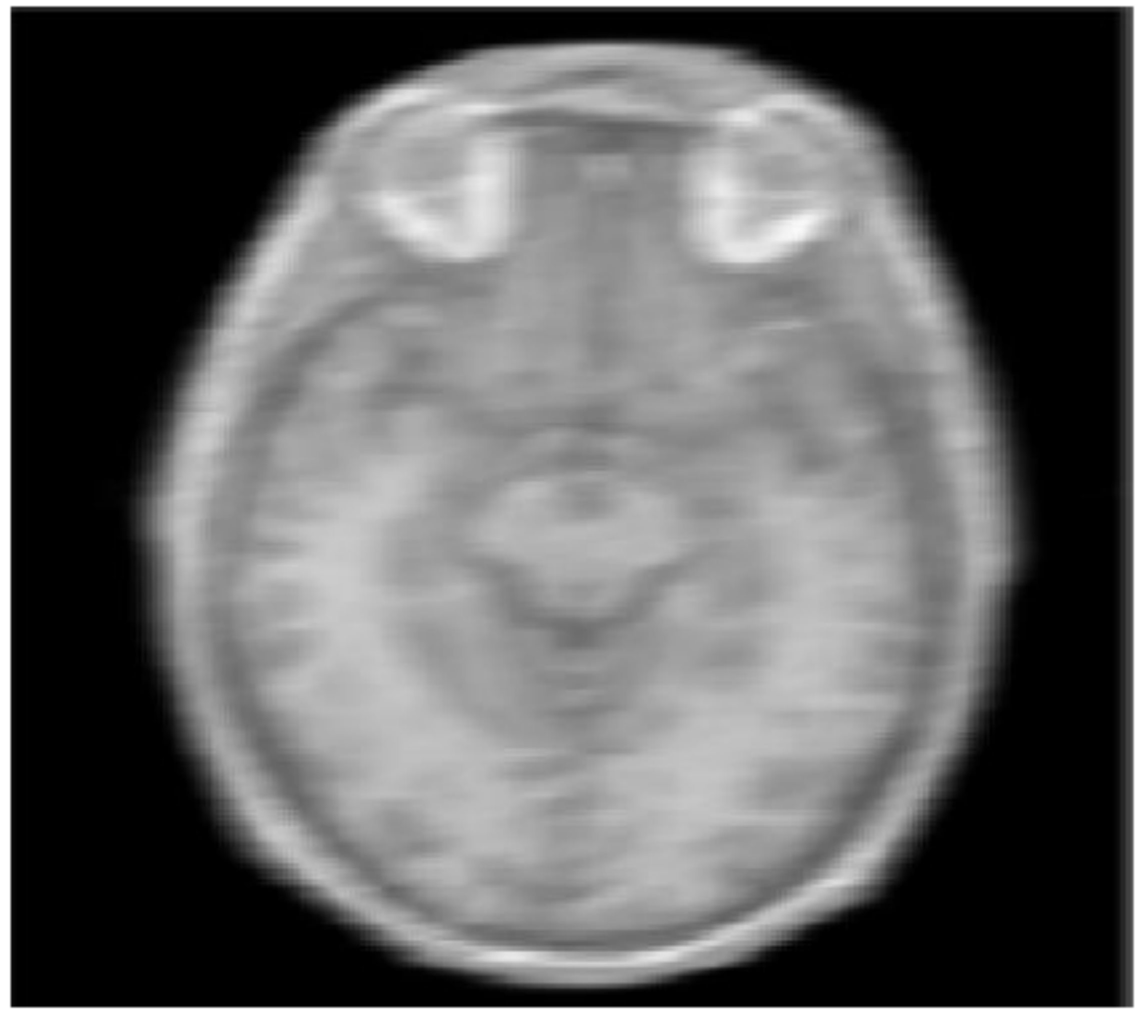
(a) Motion Blurred MRI Image



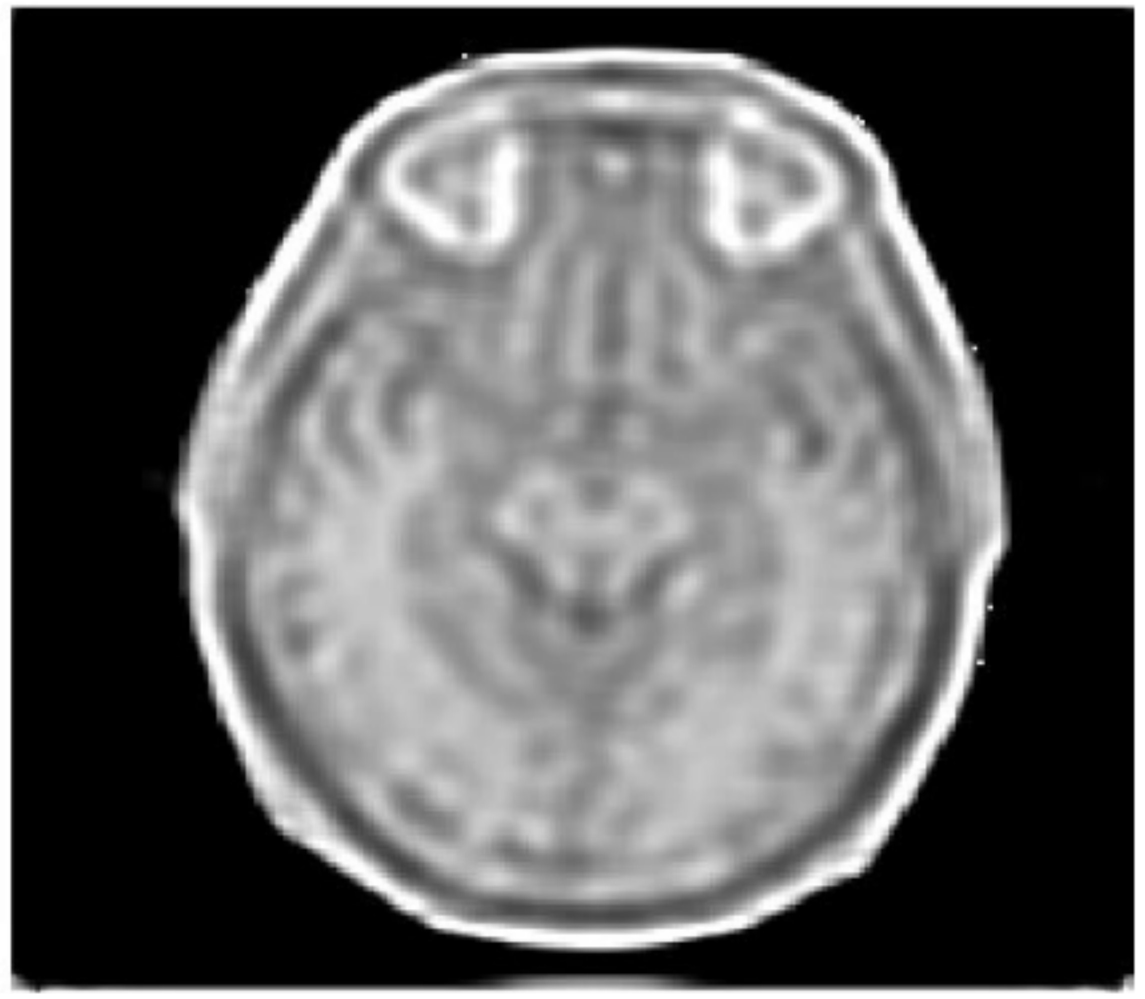
(b) Deblurring using Lucy-Richardson Algorith



(c) Deblurring using Regularized Filter

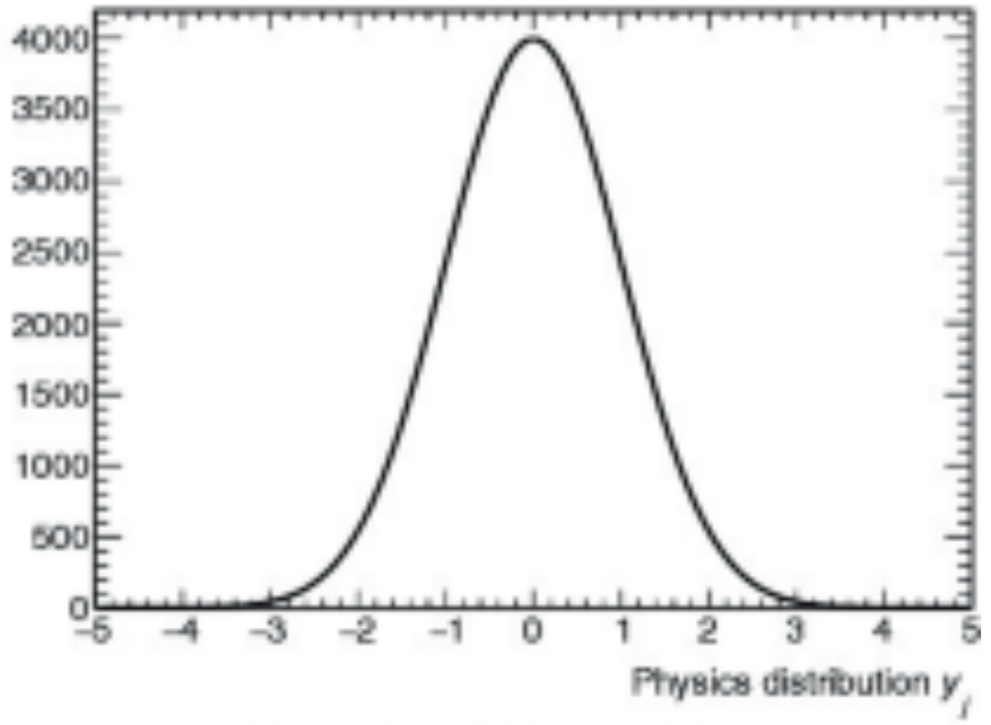


(d) Deblurring using Weiner Filter

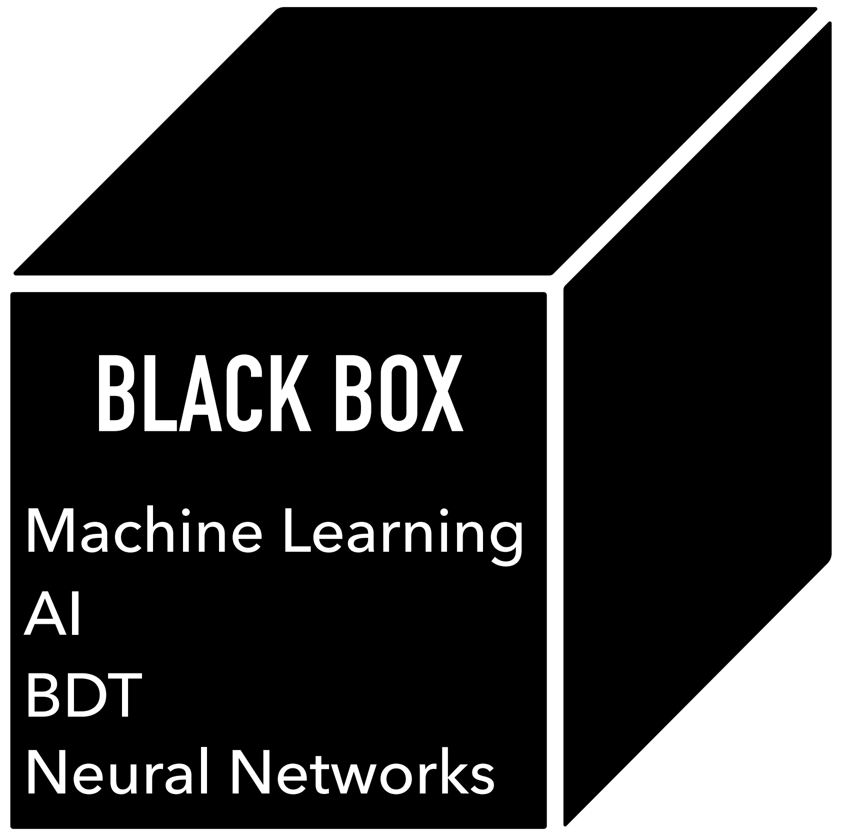


(e) Deblurring by Proposed method (Blind Deconvolution Algorithm

UNFOLDING AND ML

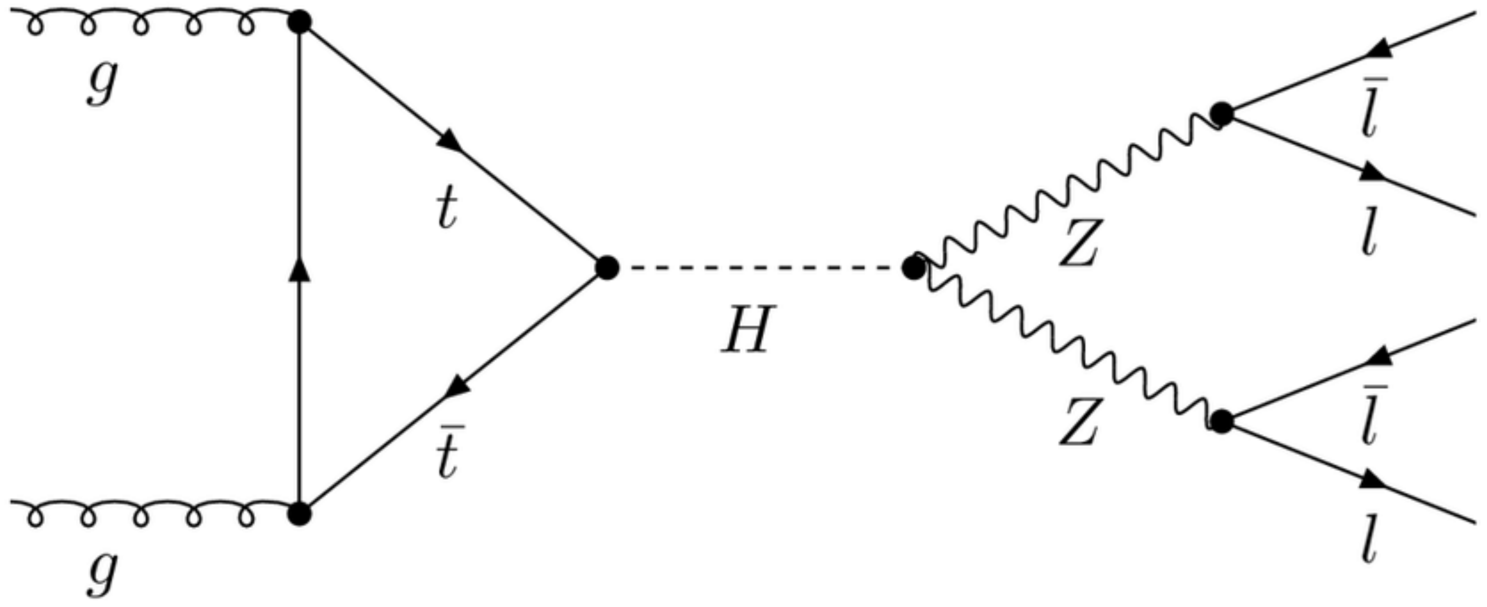


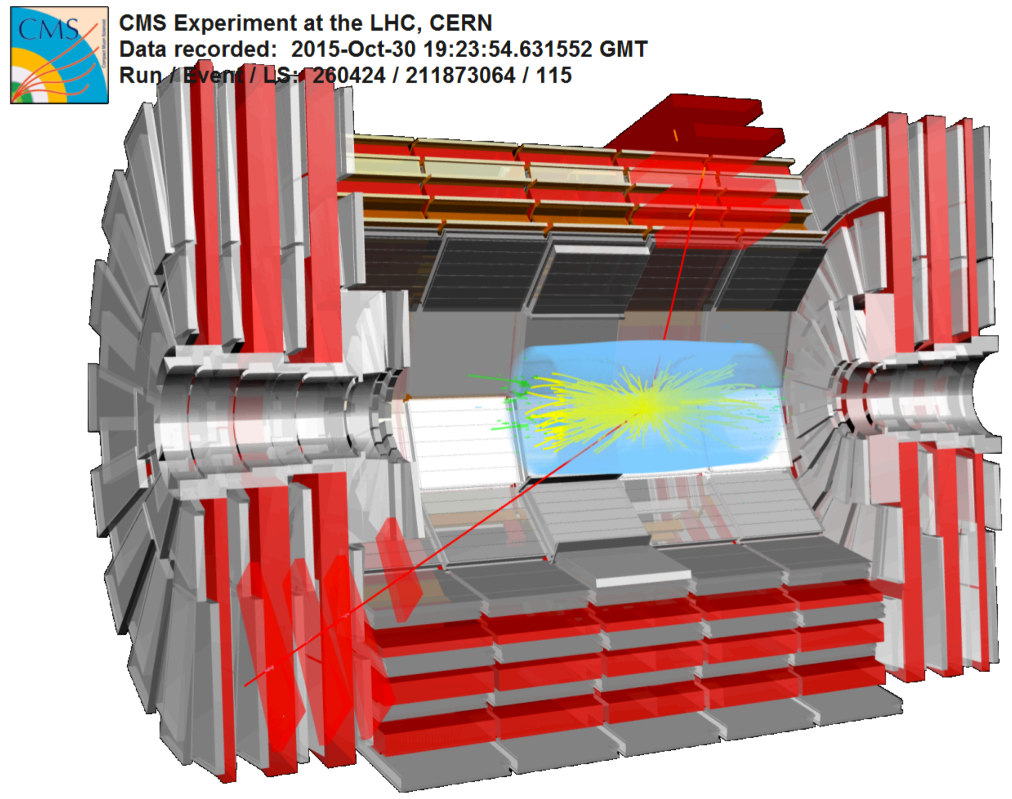
Physics distribution y_j
(particle-level)

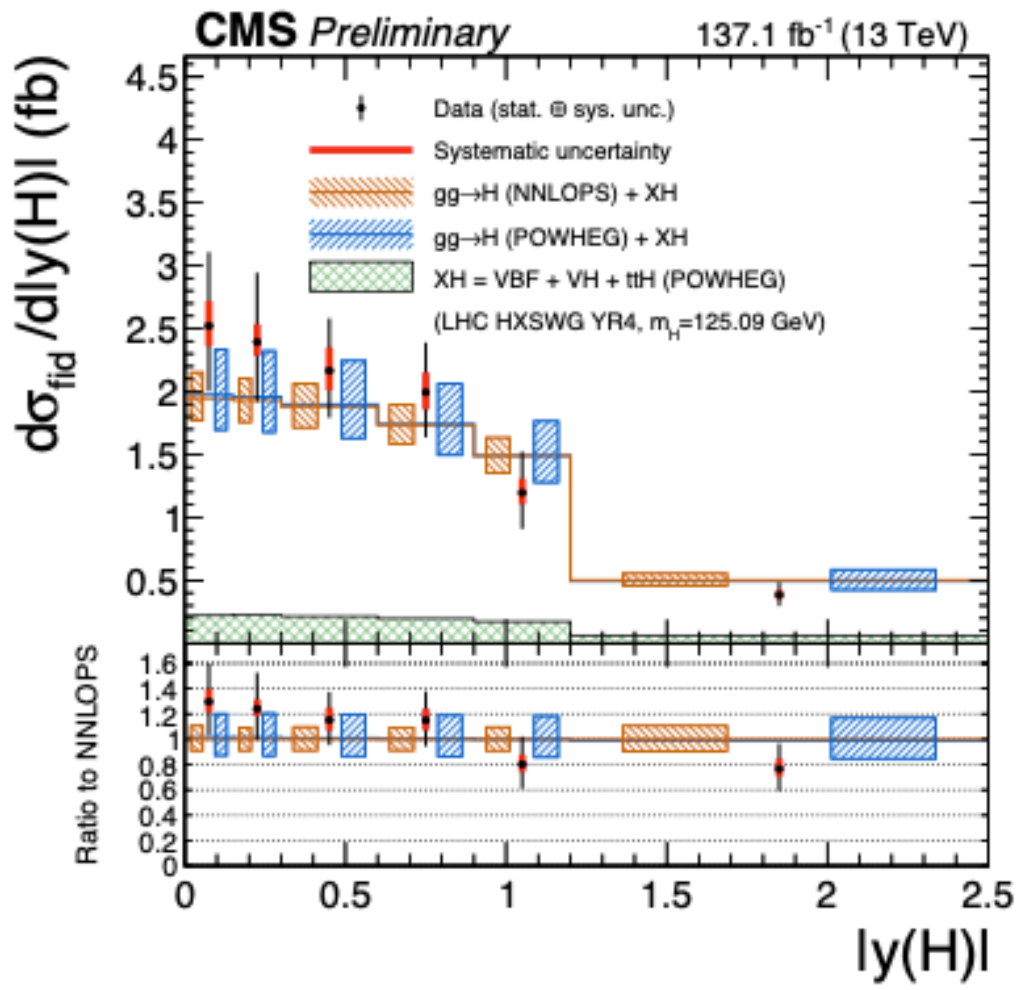


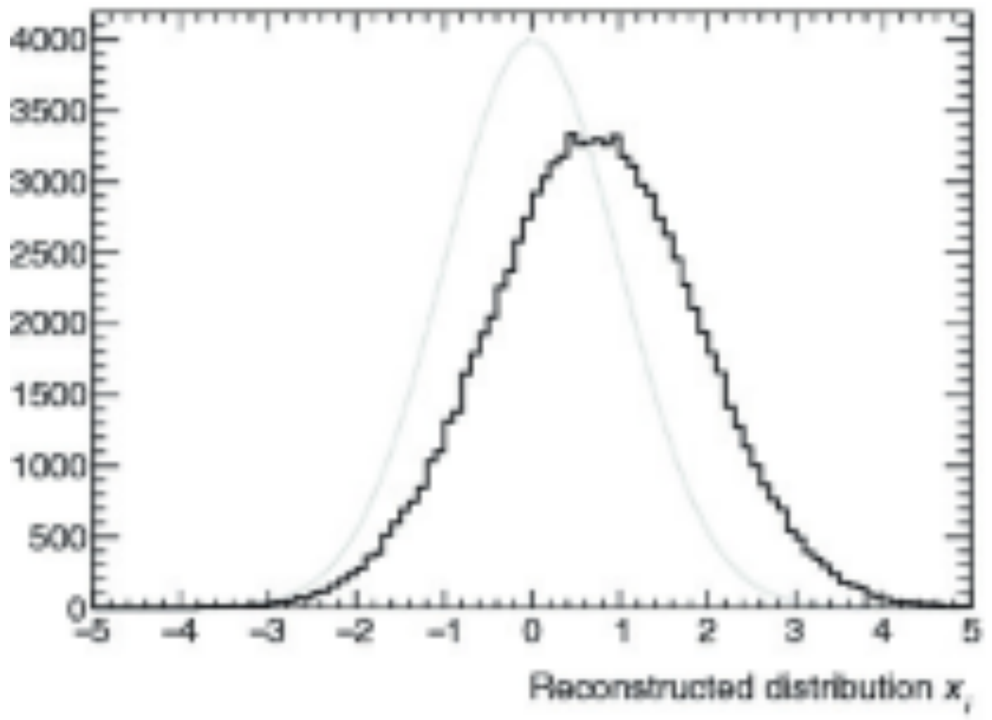
BLACK BOX

Machine Learning
AI
BDT
Neural Networks




CMS Experiment at the LHC, CERN
Data recorded: 2015-Oct-30 19:23:54.631552 GMT
Run / Event / LS: 260424 / 211873064 / 115





Reconstructed distribution x_i
(detector-level)