

Learning Optimal Test Statistics

in Presence of Nuisance Parameters

Lukas Heinrich

Technische
Universität
München

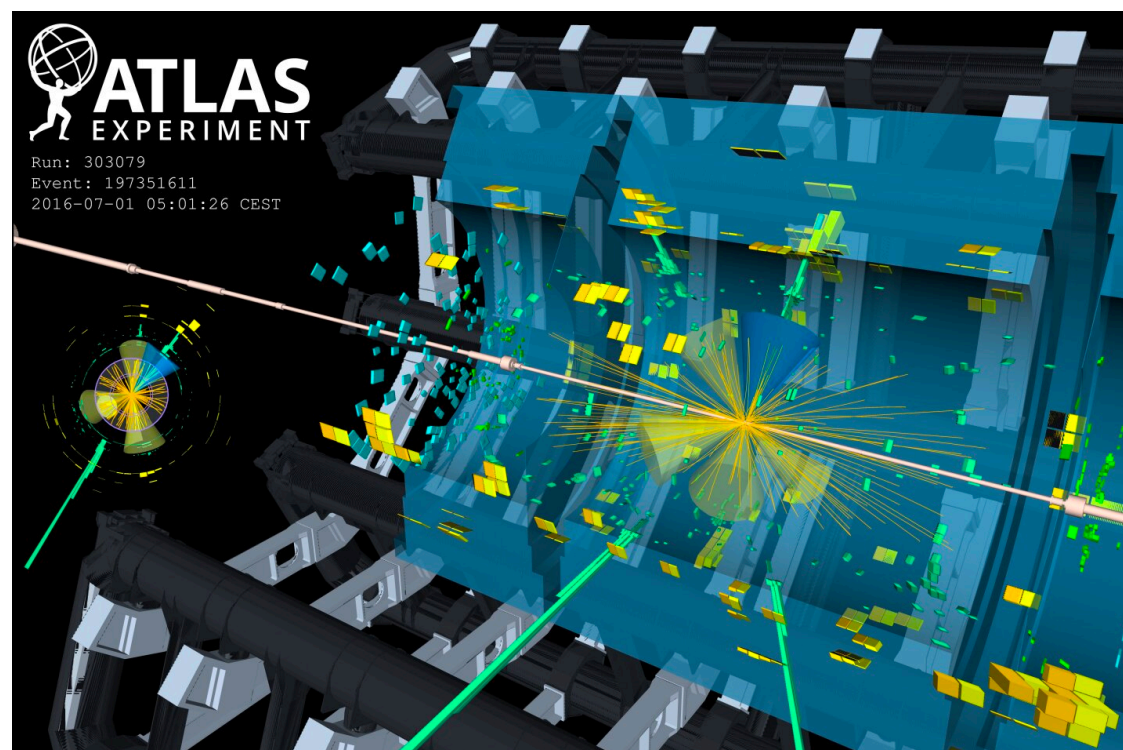


[arXiv:2203.13079](https://arxiv.org/abs/2203.13079)

Introduction

Our main goal as physicists is to make **inferences about nature** in light of the data we collect

$$\begin{aligned}\mathcal{L} = & -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} \\ & + i \bar{\psi} \not{D} \psi + h.c. \\ & + \bar{\psi}_i y_{ij} \psi_j \phi + h.c. \\ & + |\mathcal{D}_\mu \phi|^2 - V(\phi)\end{aligned}$$



Theory θ

Inference

Results

Prior Data

Our Data

The Textbook way

The way we do this usually is through *statistical inference* by formulating a *data-generating process* $p(x | \theta)$

When we say $p(x | \theta)$ (or “likelihood”) we actually mean two things:

- ability to generate data for a given theory: $x \sim p(x | \theta)$
- ability to evaluate the probability under a given theory: $L(\theta) = p(x | \theta)$

Bayesian and Frequentist Inference

With a likelihood in hand, we can follow inference procedures

Bayesians: Let's update our priors!

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

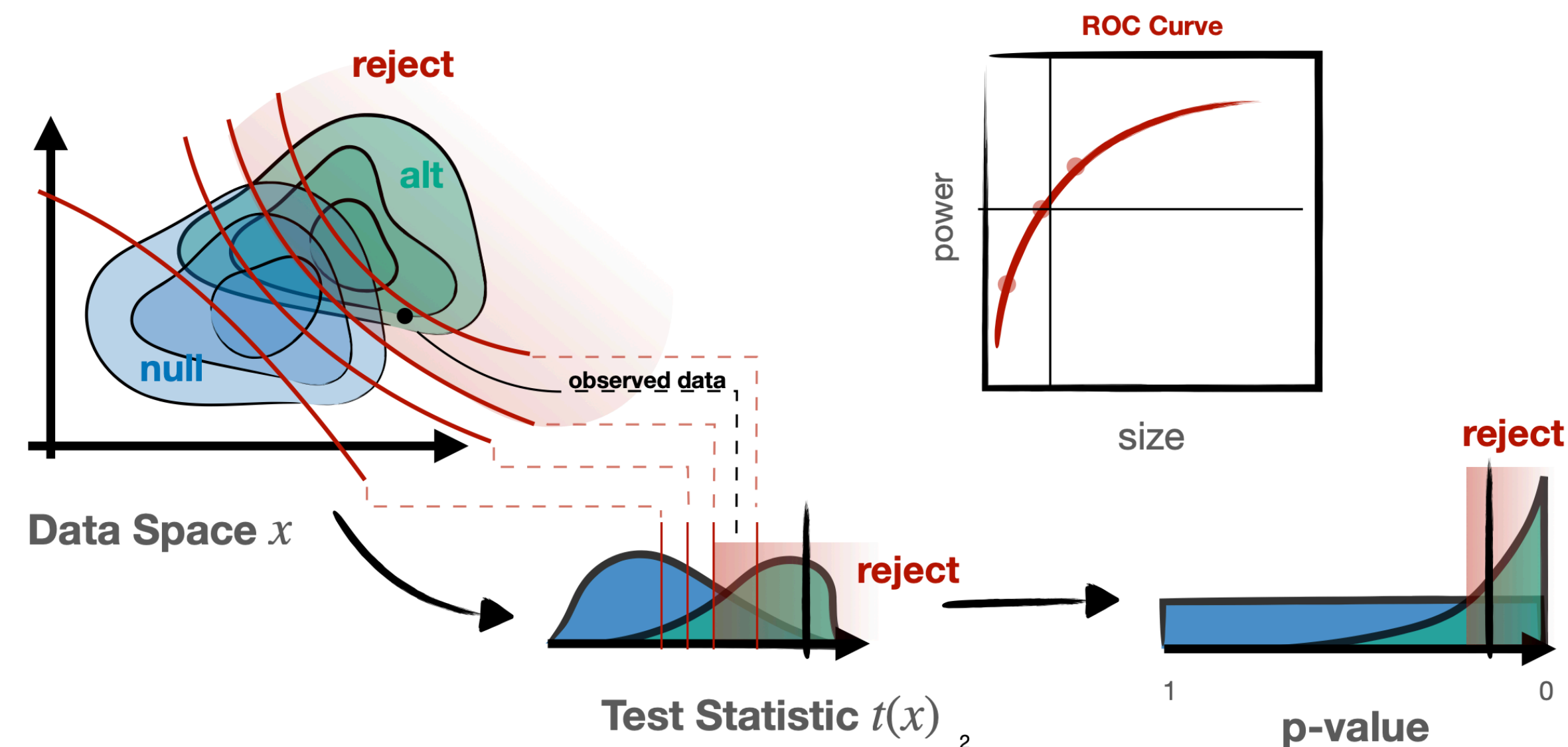
Note 1: requires ability to compute $p(x | \theta)$

Note 2: subjective choice on your priors $p(\theta)$

Bayesian and Frequentist Inference

Frequentists: let's look at the data distribution!

- **ideally** in a way that **accentuates the difference** between theories
i.e. through a scalar “test statistic” $t(x) : \mathbb{R}^N \rightarrow \mathbb{R}$



Note 1: in principle only requires ability to sample $x \sim p(x | \theta)$ and compute $t(x)$

Note 2: subjective choice of which test statistic $t(x)$ to use

Optimal Test Statistics

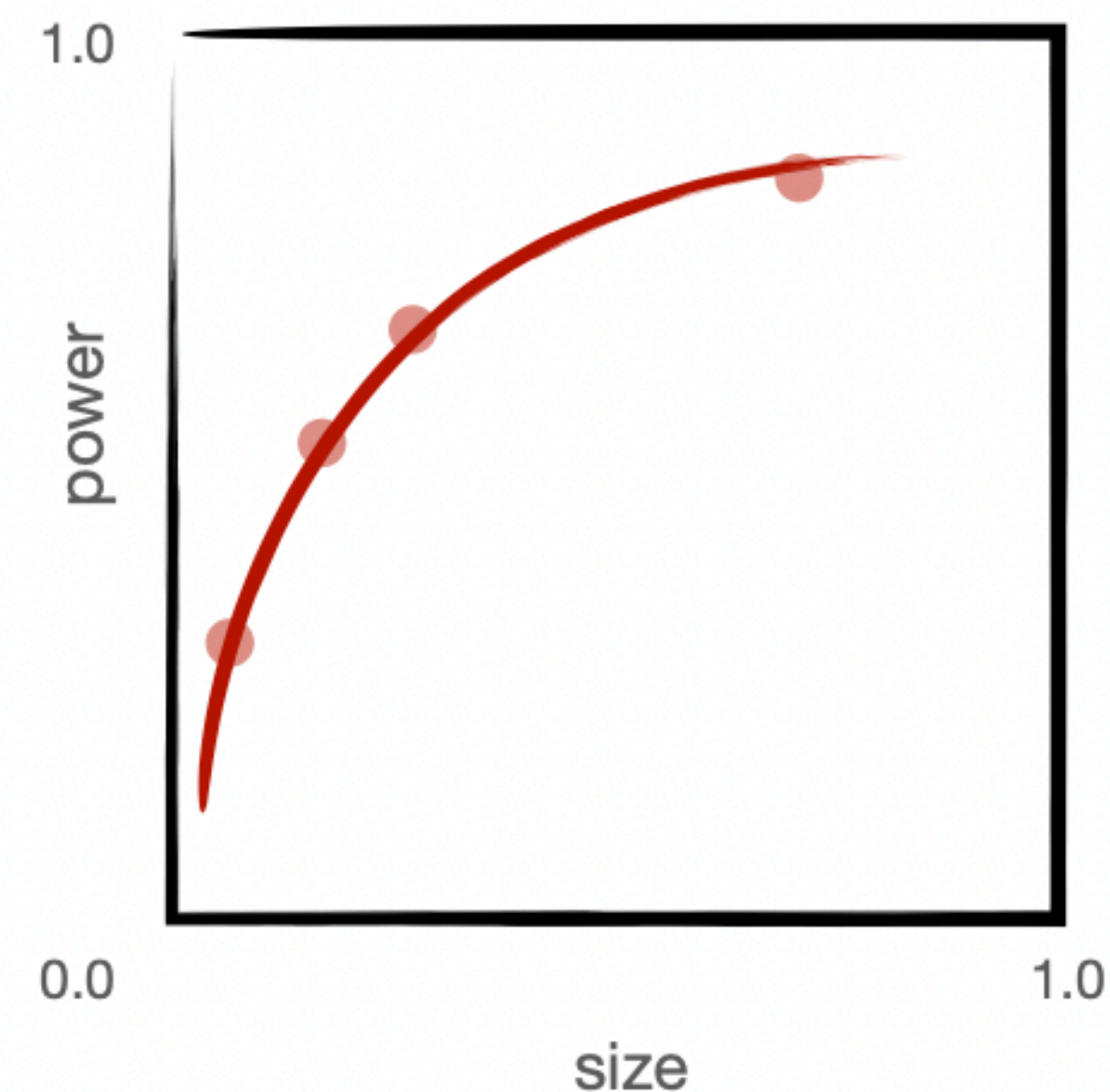
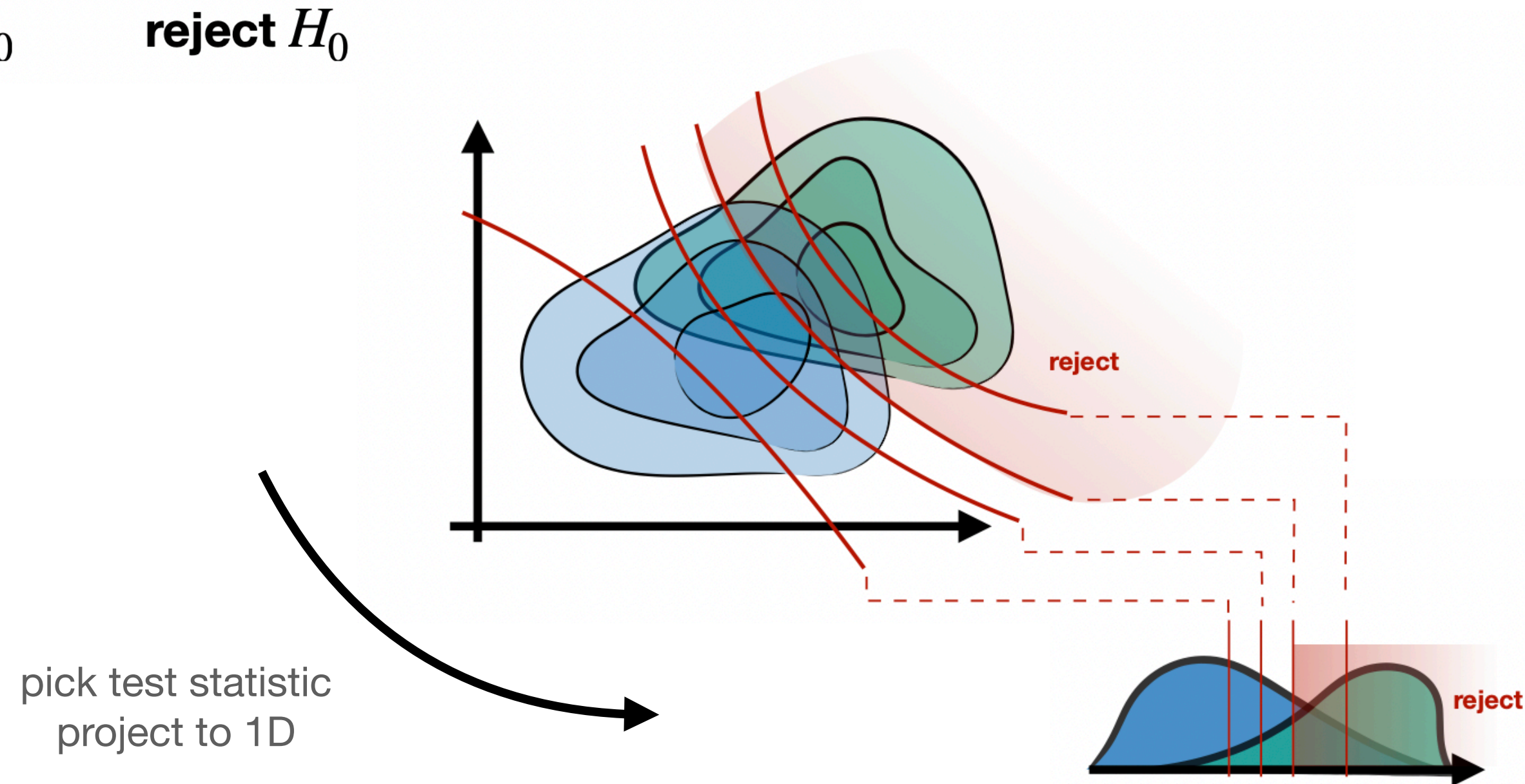
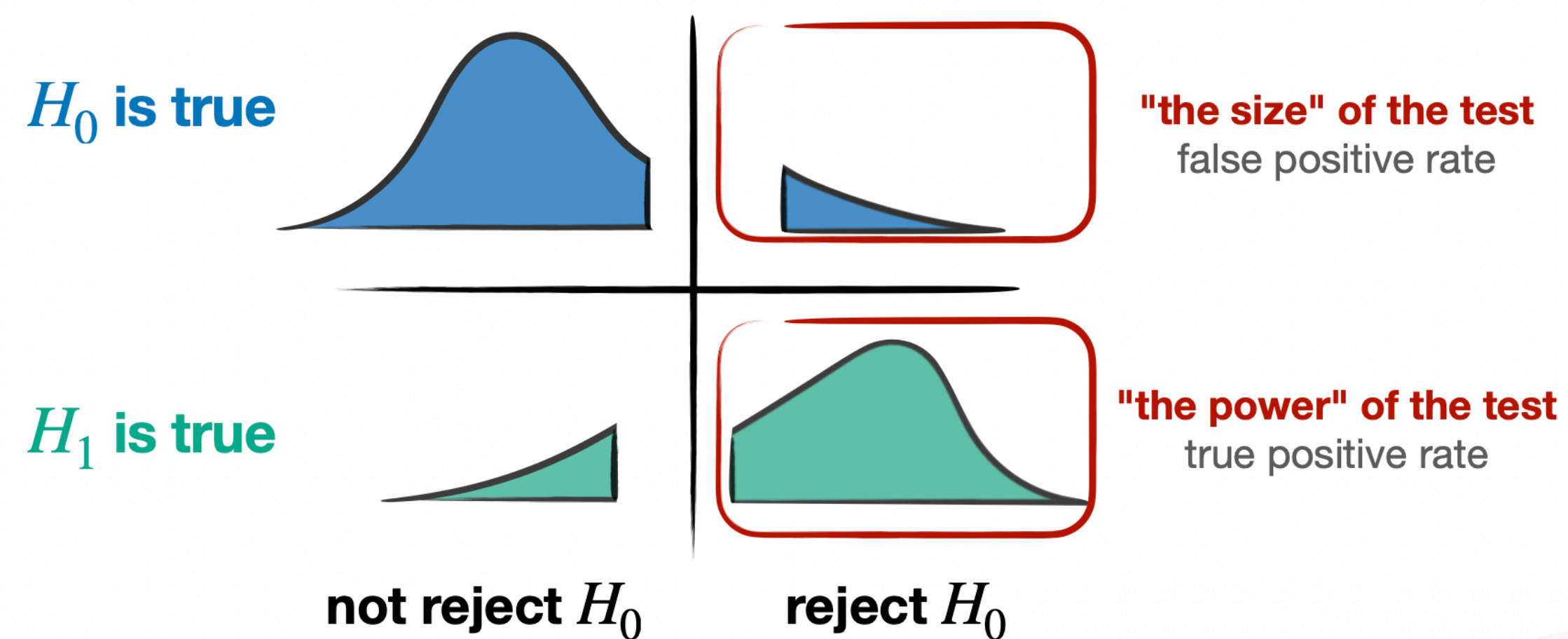
In reality, we often **do want evaluate the likelihood** $p(x | \theta)$

Why? Because it leads to the optimal test statistic!

Neyman-Pearson Lemma: The most powerful test is the Likelihood Ratio Test

$$t(x) = -2 \log \frac{p(x|H_0)}{p(x|H_1)}$$

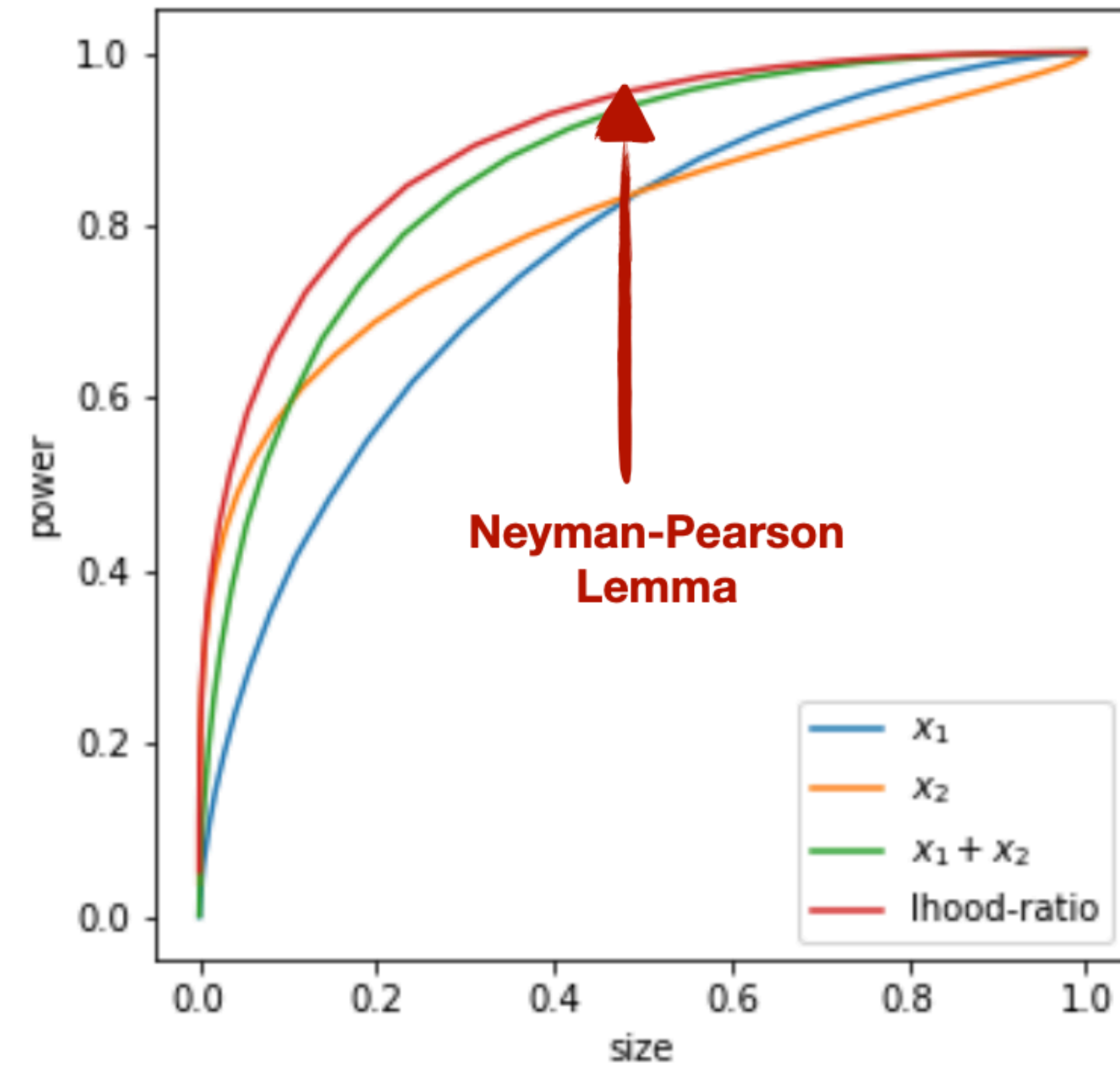
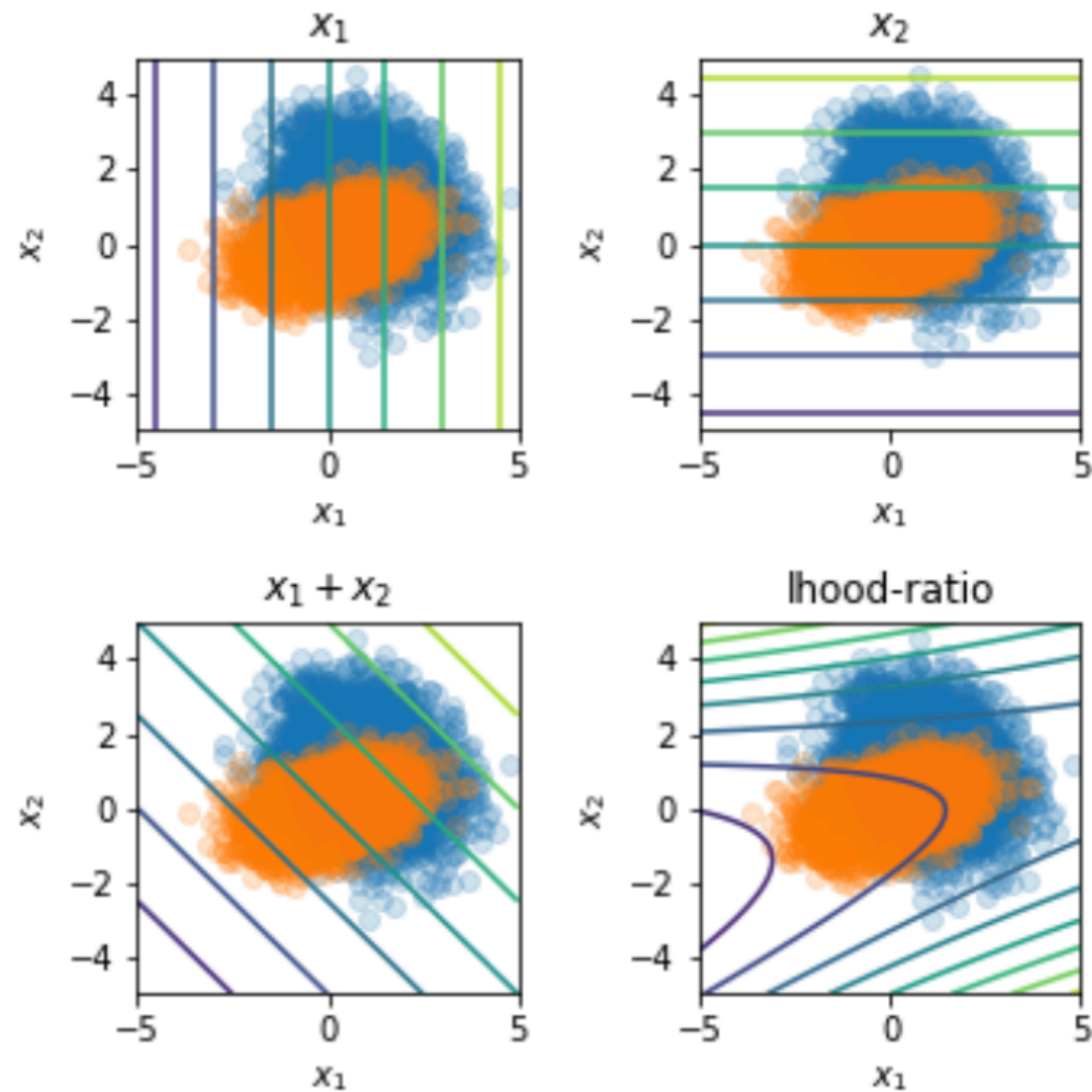
In what sense is it optimal?



a given threshold defines power & size

Optimal Test Statistics

Likelihood-Ratio is best test *for any size!*



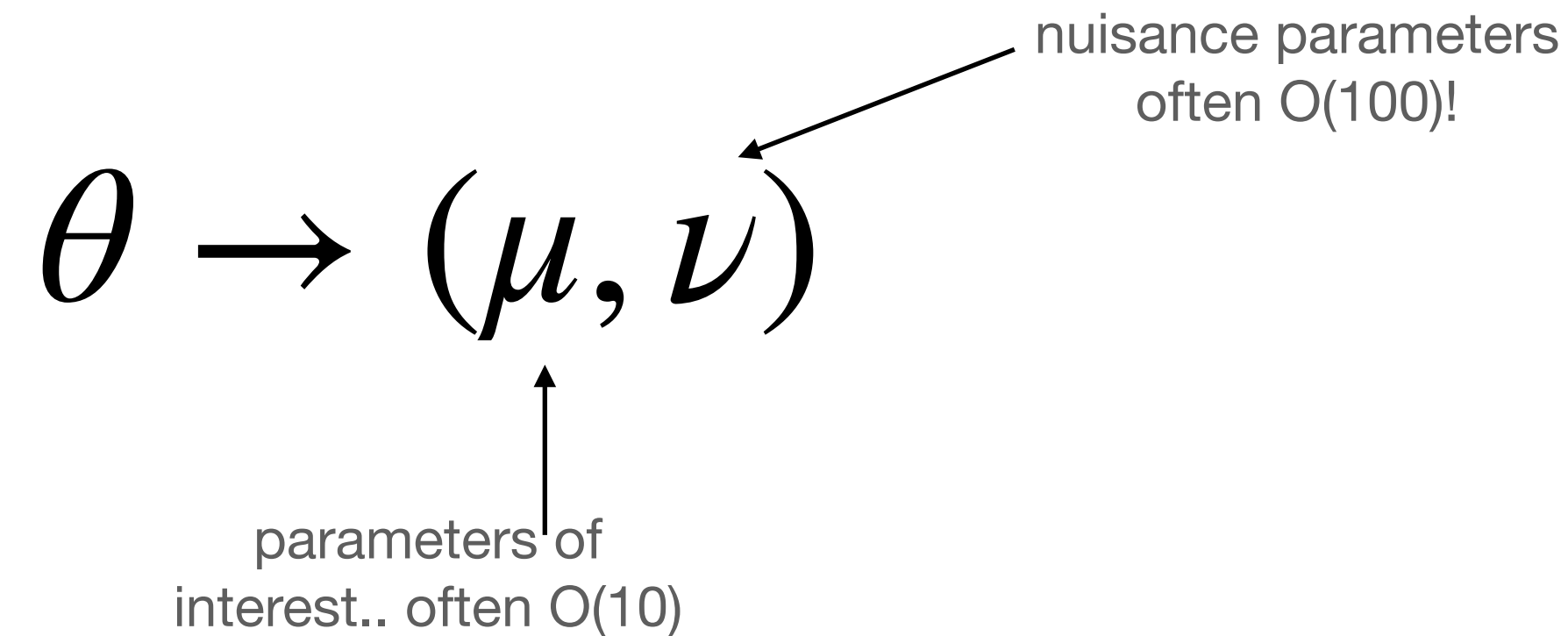
Adding Nuisance Parameters

For realistic models we often seriously expand the parameter space

$$\theta \rightarrow (\mu, \nu)$$

parameters of
interest.. often O(10)

nuisance parameters
often O(100)!



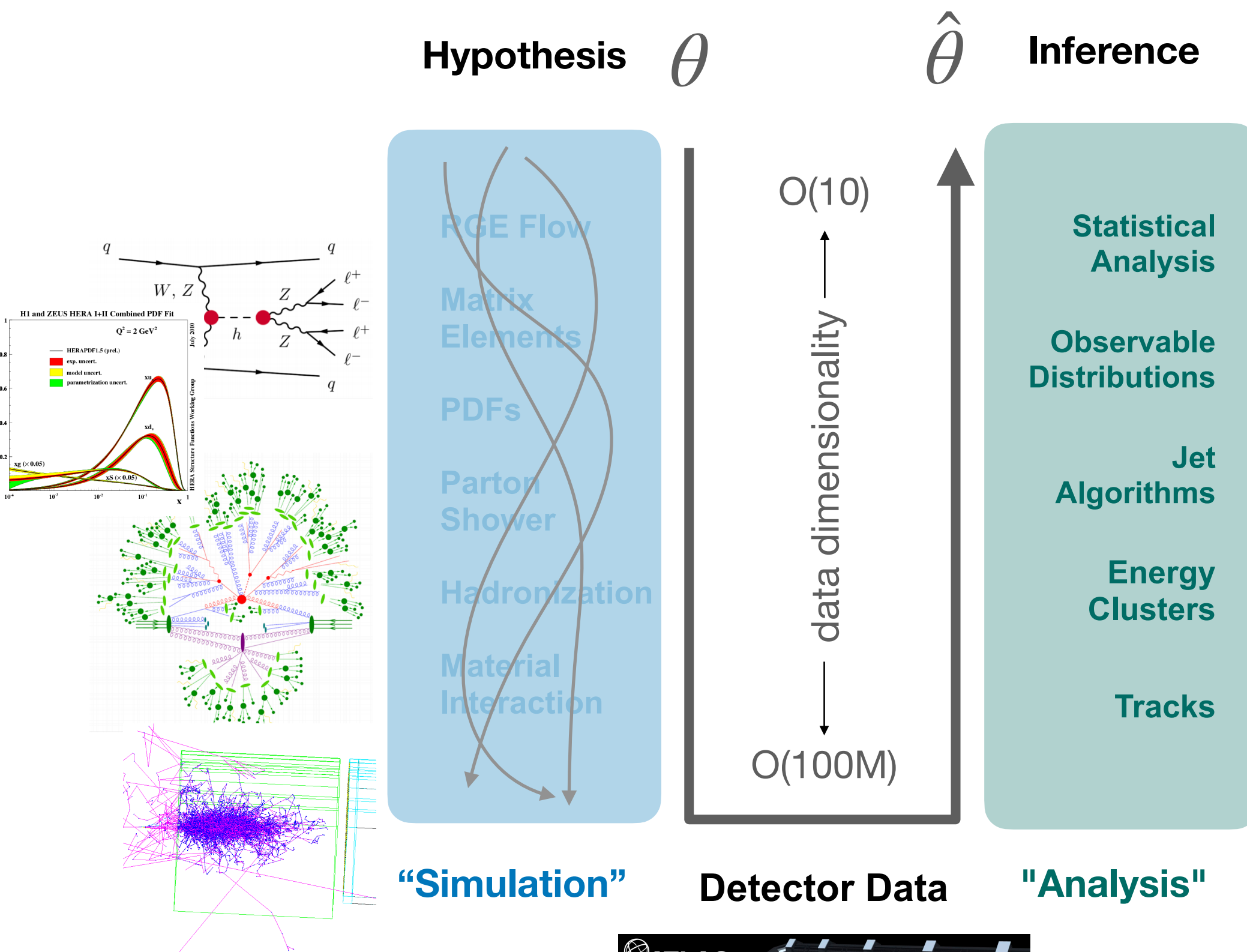
The Profile Likelihood:

$$t_{\mu}(x) = -2 \log \frac{p(x|\mu, \hat{\hat{\nu}})}{p(x|\hat{\mu}, \hat{\nu})}$$

... proven by A. Wald in 1943 to be optimal in the sense of having *optimal average power*

A slight problem

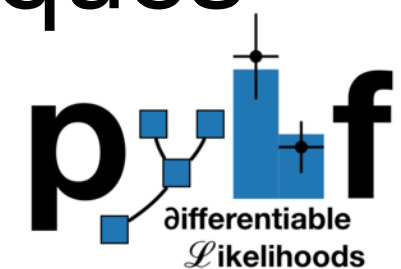
Unfortunately in HEP we cannot evaluate $p(x | \theta)$ - it's likelihood-free!



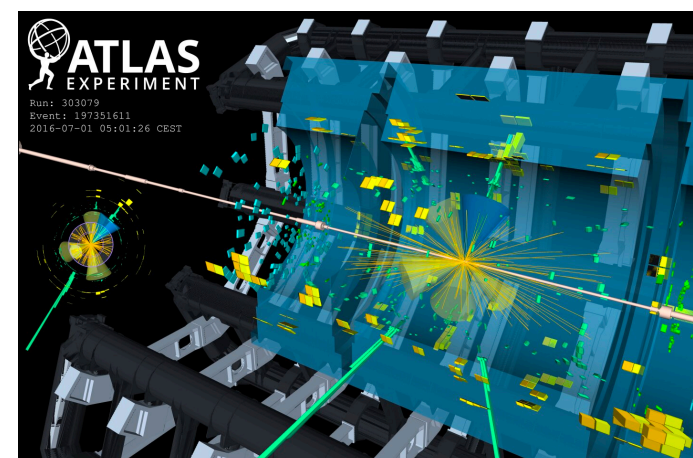
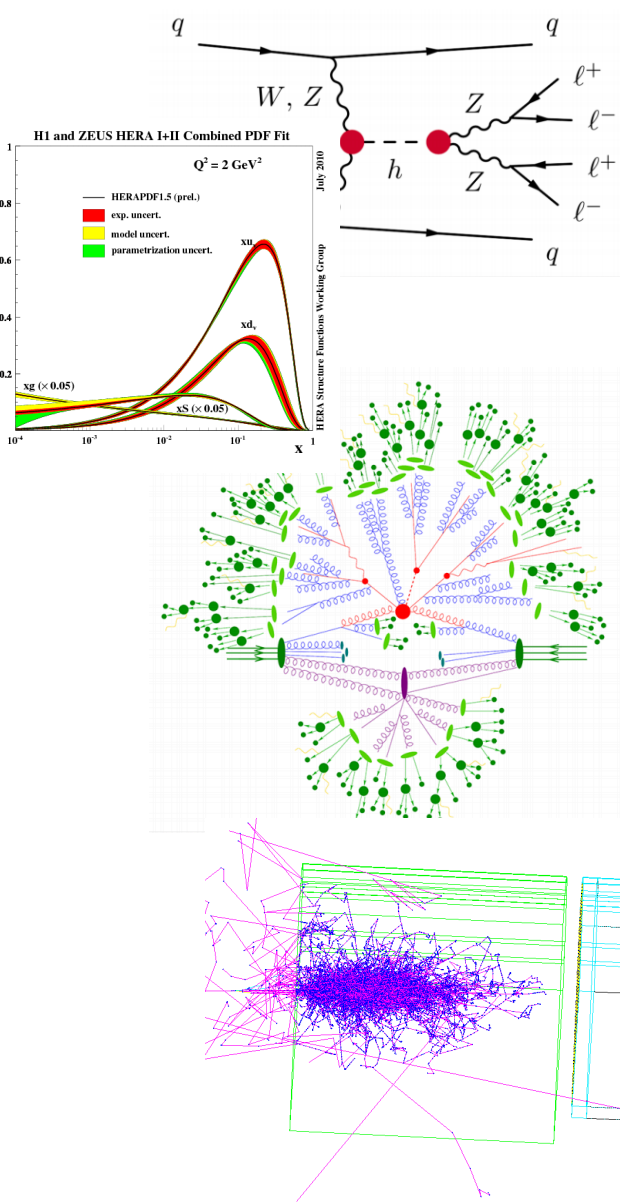
often try to at least build an approximate likelihood using smart dim. reduction
e.g. reconstruction & analysis

$$p(x | \theta) \approx p(f_{\text{ana}}(x) | \theta)$$

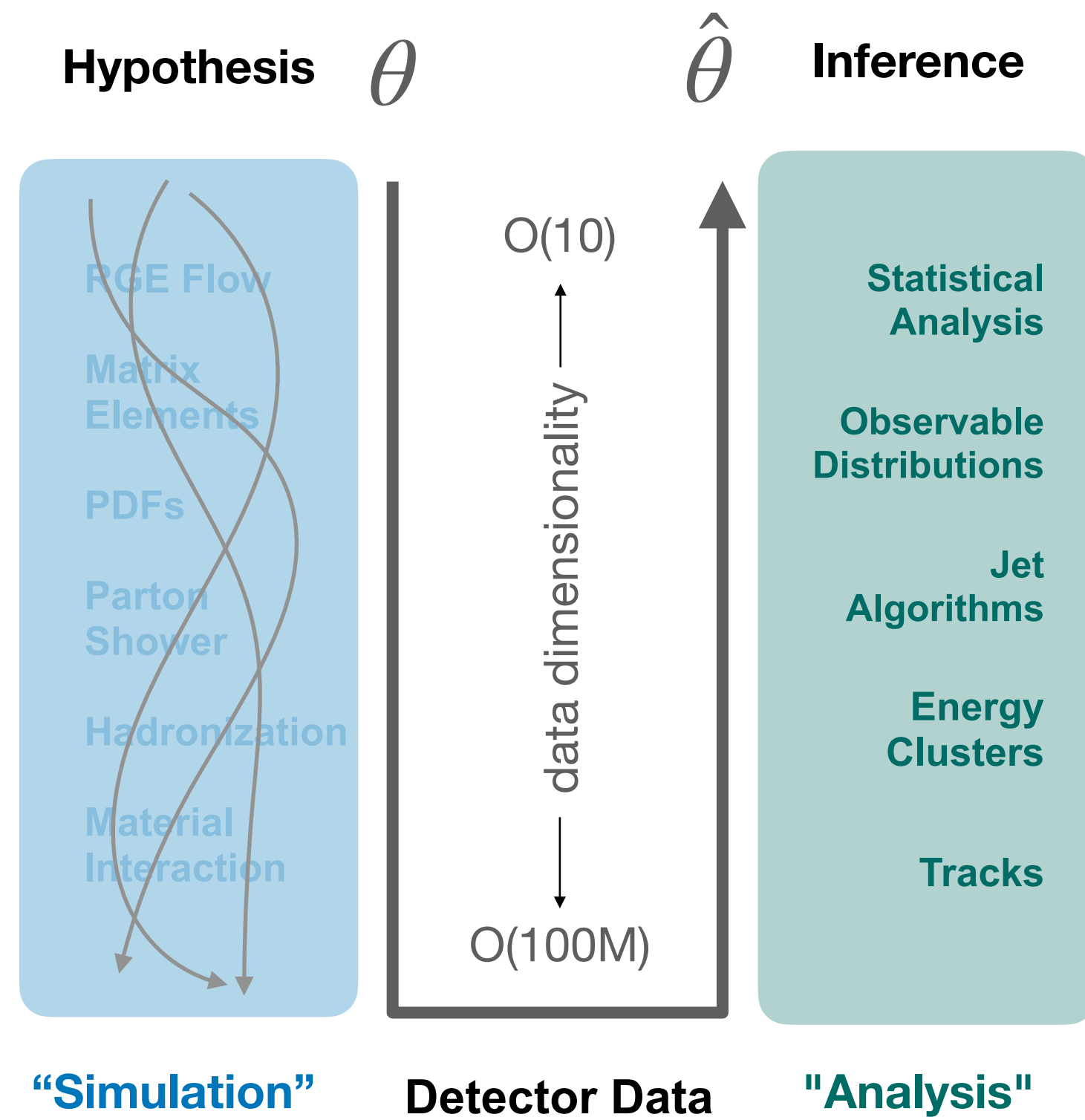
proceed using standard techniques
e.g. via pyhf-based models



derived $t(x)$ (e.g. approx. LR) may not be optimal, but inference will never be wrong



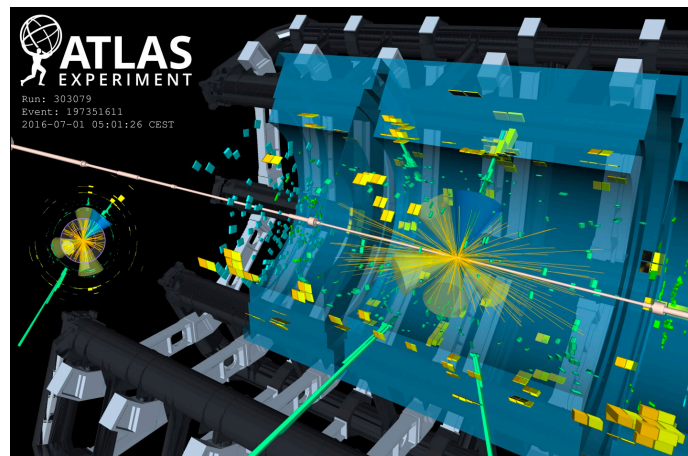
As discussed:



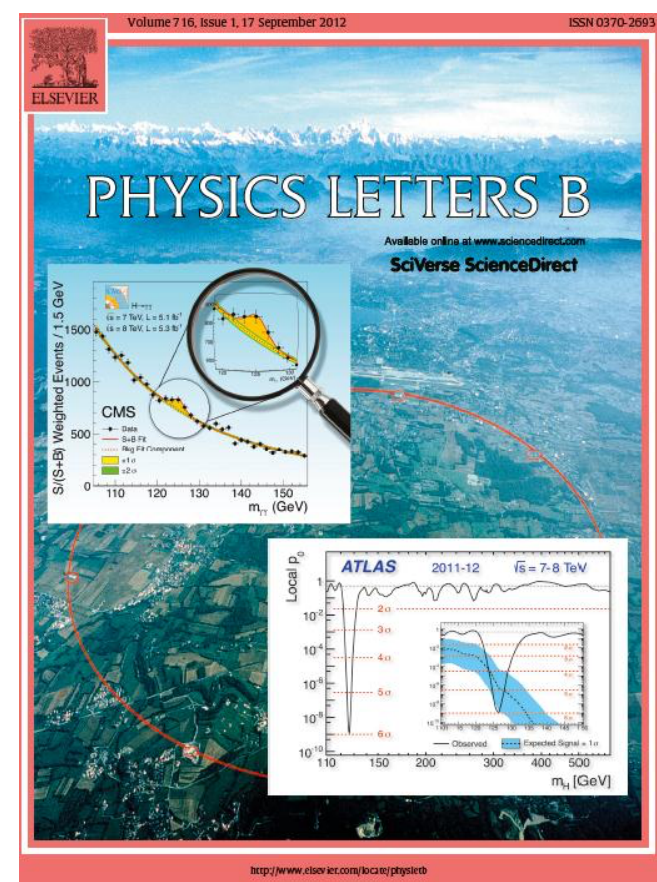
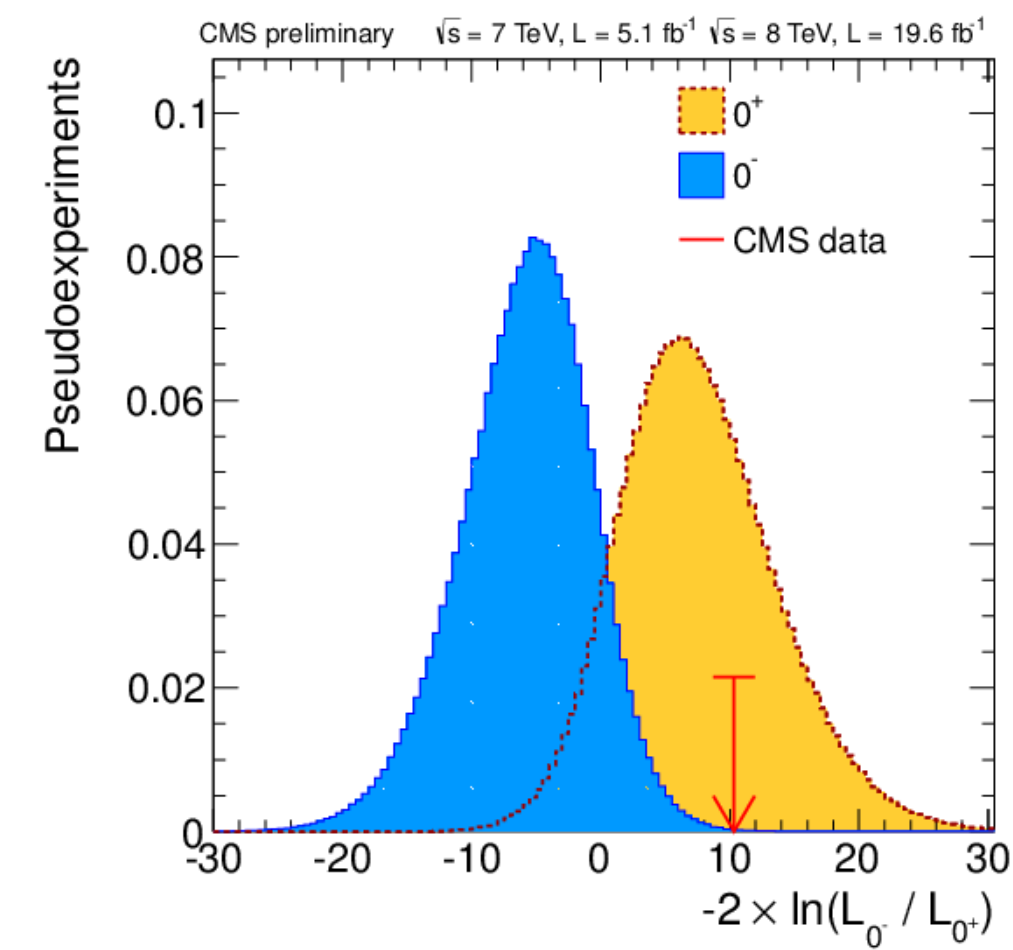
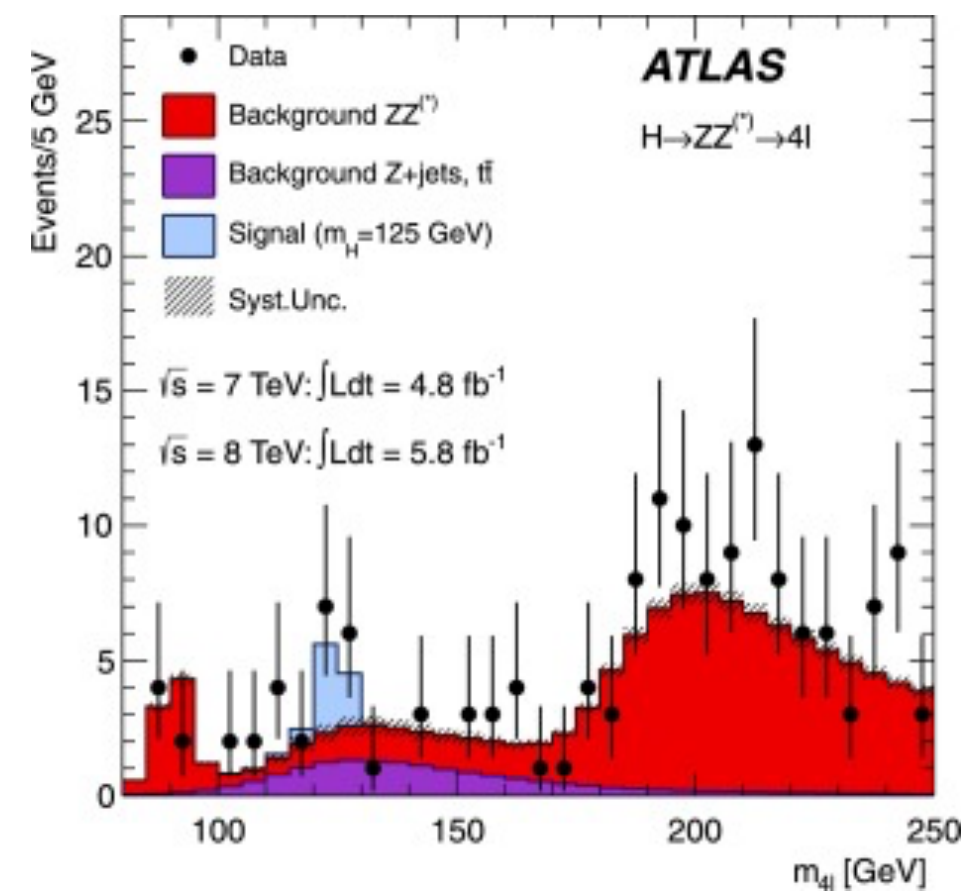
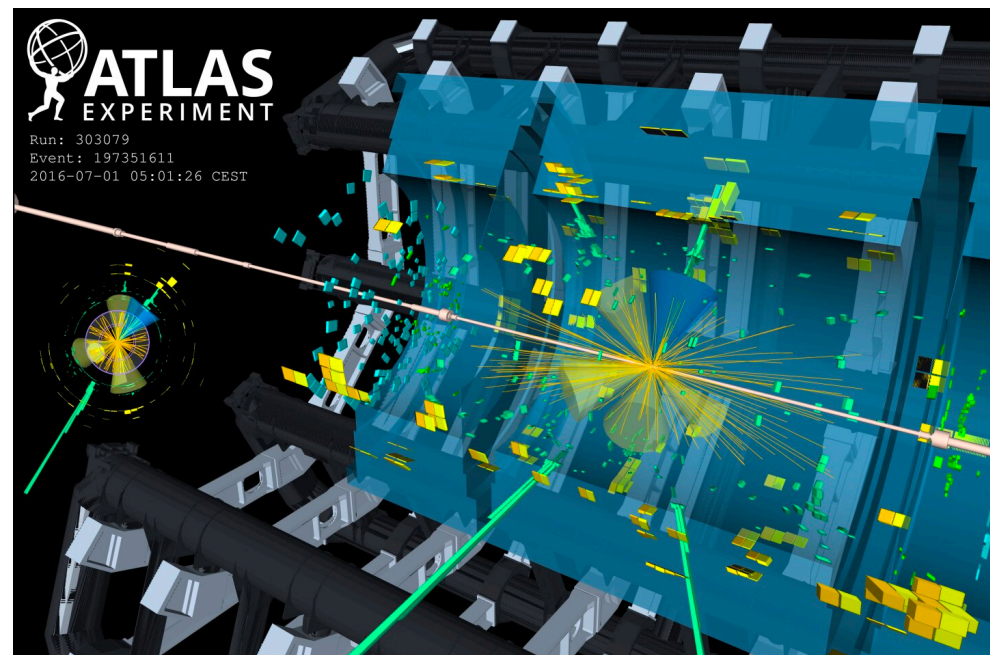
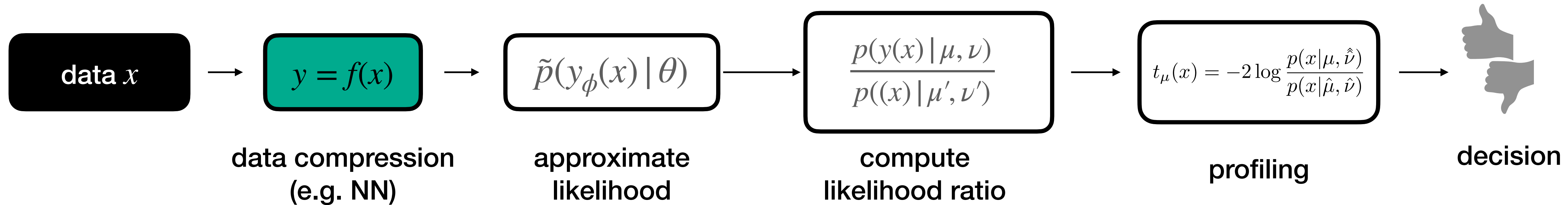
**Likelihood-free Inference and ML
are a match made in heaven**

**But key question: ML depends on training data,
which depends on nuisance parameters**

What does uncertainty-aware ML look like



A typical workflow



if $y = f(x)$ is well-chosen, $\tilde{t}(y(x)) \rightarrow t(x)$

A simple Idea

There is a path to likelihood-free frequentist inference by exploiting the optimality properties of the test statistic we seek

If we're using $t(x)$ (e.g. likelihood ratio or profile likelihood ratio) because it is optimal....

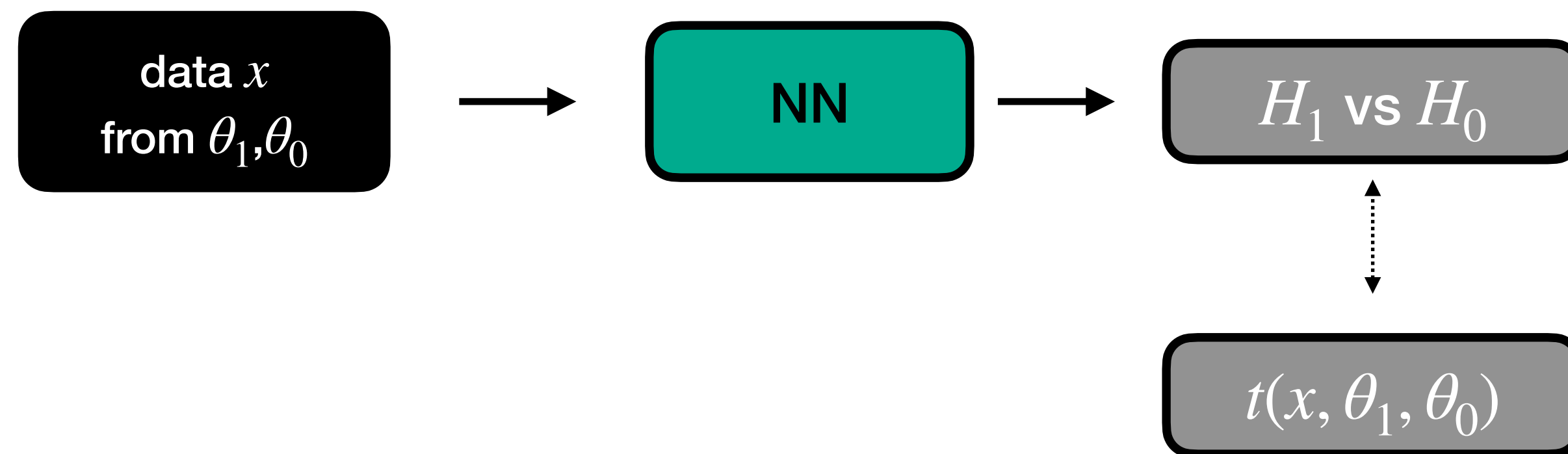
... that just means that **we can find $t(x)$ through optimization in function space, a.k.a. Machine Learning**

- **just requires samples from $p(x | \theta)$, not the likelihood**

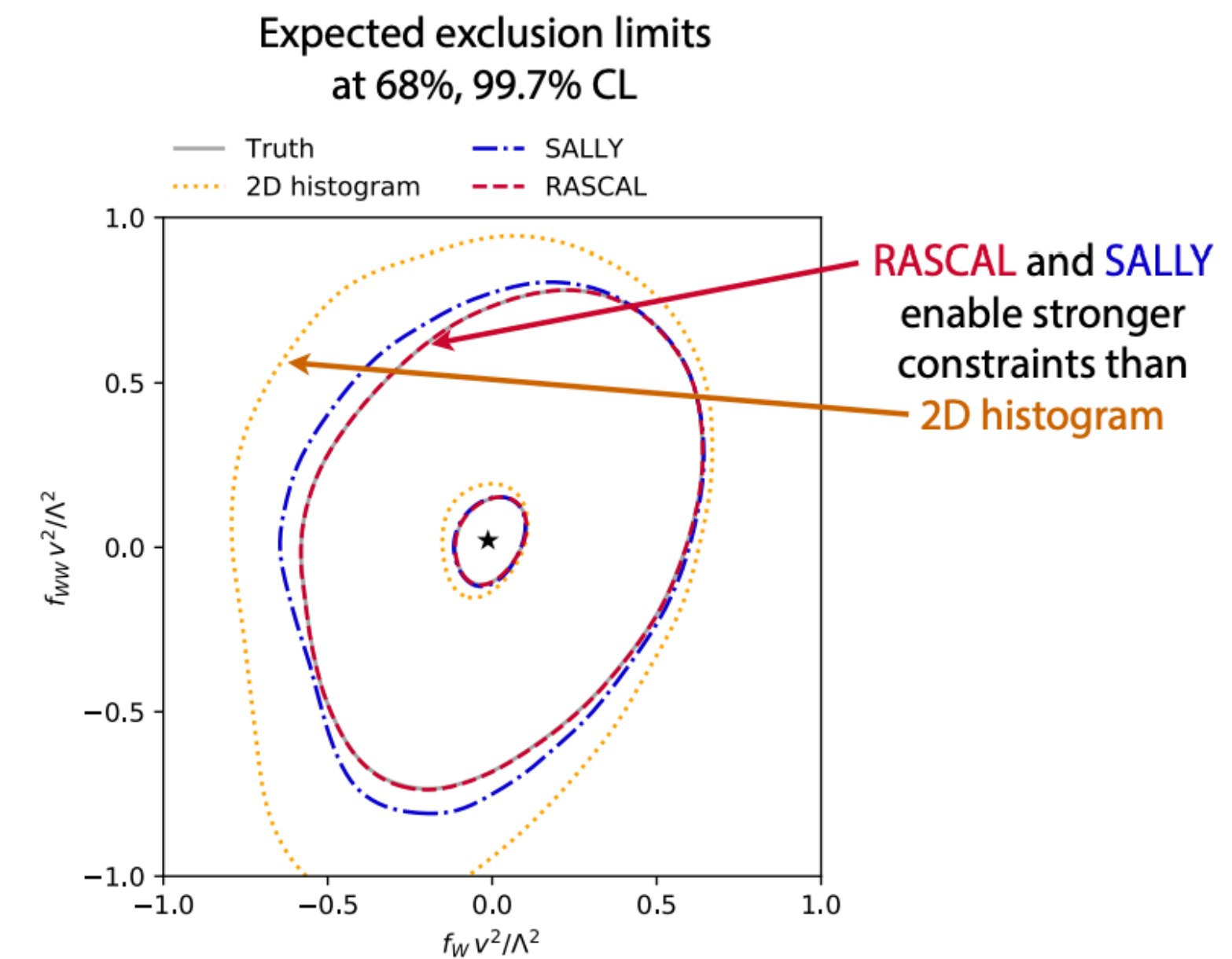
Likelihood Ratio Trick

For the non-nuisance case this is the “likelihood ratio trick”

Training to discriminate H_0 v. H_1 will converge to a function $f(x)$ that is $1 \leftrightarrow 1$ to the exact Likelihood Ratio $t(x)$ instead of



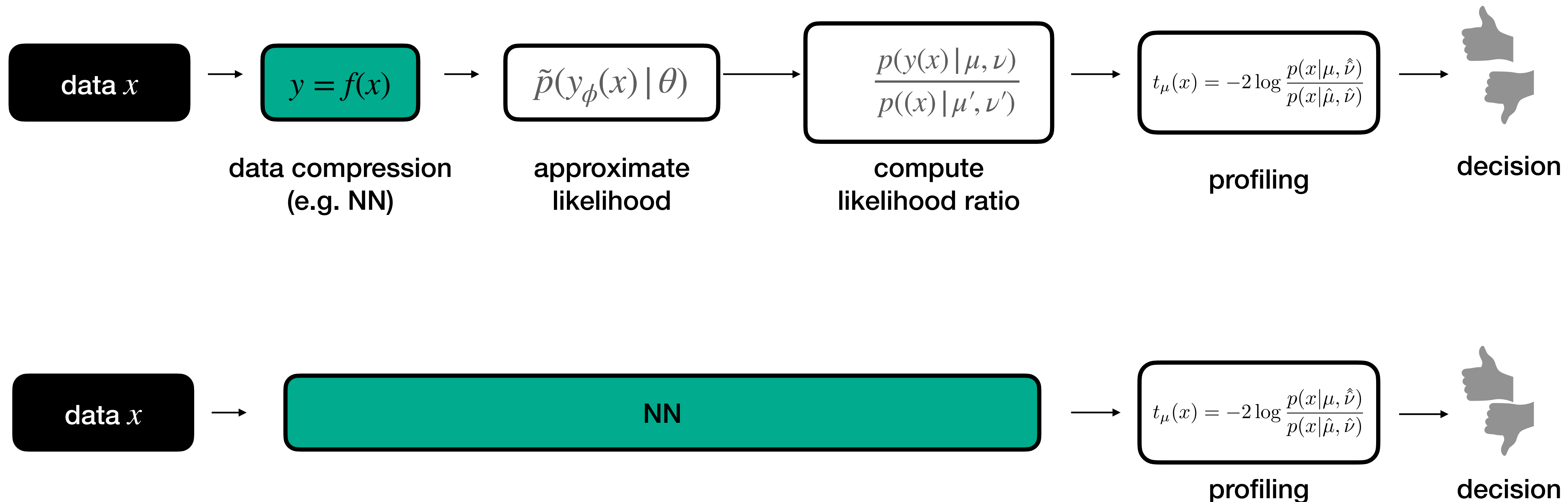
Avoid “degradation” of intermediate compression $x \rightarrow y = f(x)$



[Brehmer et al]

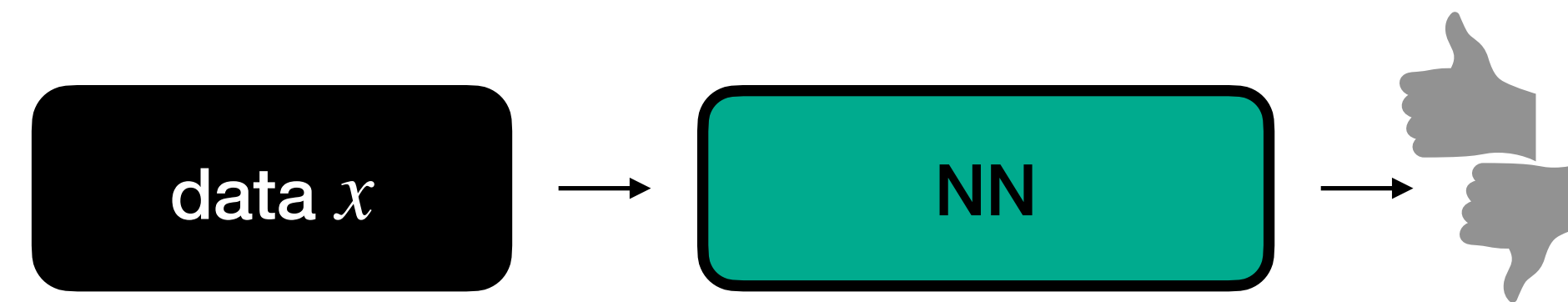
Likelihood Ratio Trick

What's happening? We're replacing a big chunk of the workflow with a NN with a clever training objective that asymptotes to the target

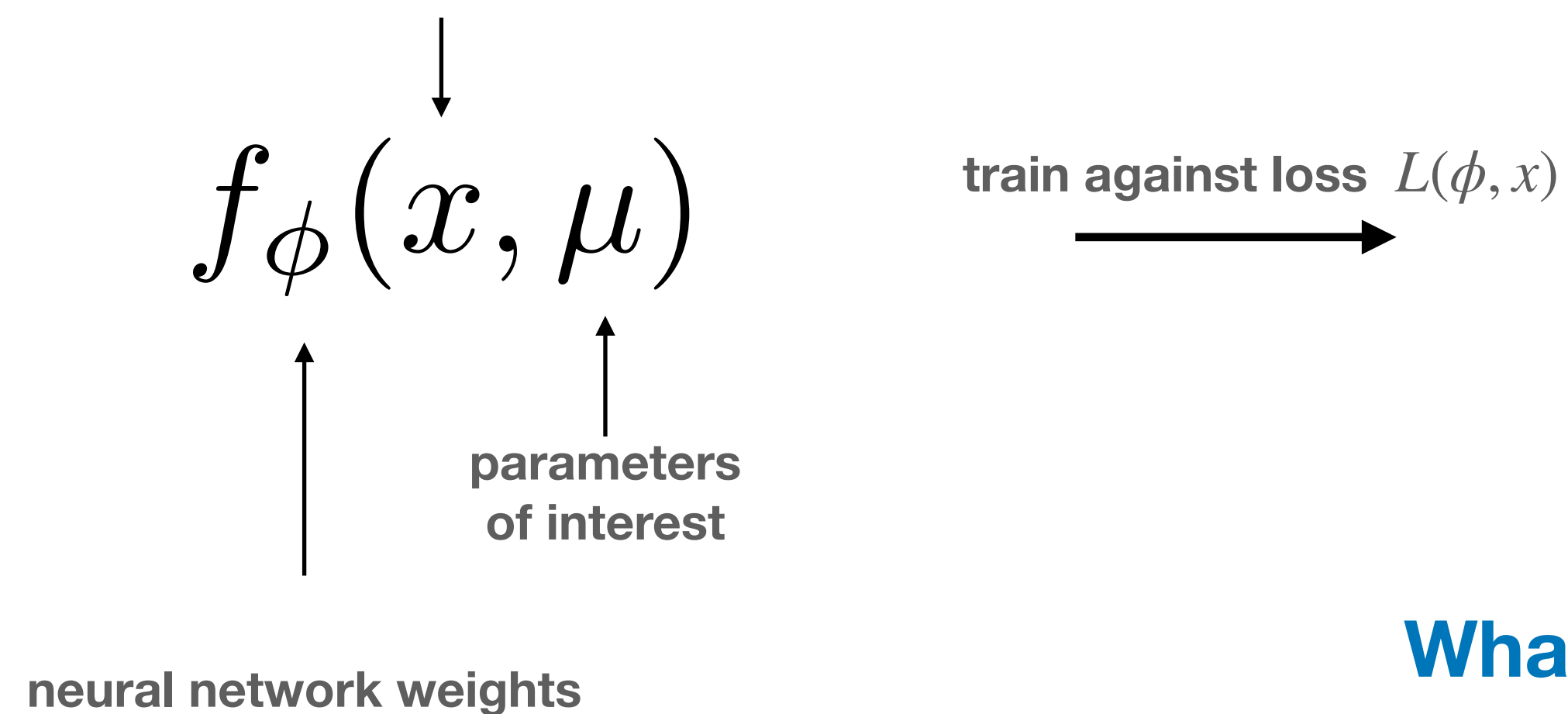


With Nuisance Parameters

Can we extend go all the way?



i.e. train a neural network $f(x, \mu)$ such that it converges to the profile likelihood (or a function that is $1 \leftrightarrow 1$ with it)



$$t_\mu(x) = -2 \log \frac{p(x|\mu, \hat{\hat{\nu}})}{p(x|\hat{\mu}, \hat{\nu})}$$

What is the appropriate training procedure?

Go back to Wald 1943

To find appropriate training procedure to optimize $f(x, \mu) \rightarrow t_\mu(x)$ we need to recall in what sense the profile likelihood is optimal

TESTS OF STATISTICAL HYPOTHESES CONCERNING SEVERAL PARAMETERS WHEN THE NUMBER OF OBSERVATIONS IS LARGE⁽¹⁾

BY
ABRAHAM WALD

TABLE OF CONTENTS

1. Introduction.....	426
2. Assumptions on the density function $f(x, \theta)$	428
3. The joint limit distribution of $\hat{\theta}_n$	429
4. Reduction of the general problem to the case of a multivariate normal distribution..	433
5. Tests of simple hypotheses which have uniformly best average power over a family of surfaces.....	445
6. Tests of simple hypotheses which have best constant power on a family of surfaces...	450
7. Most stringent tests of simple hypotheses.....	451
8. Definitions of "best" tests of composite hypotheses.....	453
9. Tests of linear composite hypotheses which have uniformly best average power over a family of surfaces.....	455
10. Tests of linear composite hypotheses which have best constant power on a family of surfaces.....	461
11. Most stringent tests of linear composite hypotheses.....	461
12. The general composite hypothesis.....	463
13. Optimum properties of the likelihood ratio test.....	470
14. Large sample distribution of the likelihood ratio.....	478
15. Summary.....	481

1. **Introduction.** In this paper we shall deal with the following general problem: Let $f(x^1, x^2, \dots, x^r, \theta^1, \dots, \theta^k)$ be the joint probability density function of the variates (chance variables) x^1, \dots, x^r involving k unknown parameters $\theta^1, \dots, \theta^k$. Any set of k values $\theta^1, \dots, \theta^k$ can be represented by a point θ in the k -dimensional Cartesian space with the coordinates $\theta^1, \dots, \theta^k$. We shall denote the set of all possible parameter points by Ω . The set Ω is called parameter space. The parameter space Ω may be the whole k -dimensional Cartesian space, or a subset of it. For any subset ω of Ω , we shall denote by H_ω the hypothesis that the parameter point lies in ω . If ω consists of a single point, H_ω is called a simple hypothesis, otherwise H_ω is called a composite hypothesis. In this paper we shall discuss the question of an appropriate test of the hypothesis H_ω based on a large number of independent observations on x^1, \dots, x^r .

For simplicity we shall introduce the following notations: The letter θ or θ_i for any subscript i will denote a point in the parameter space Ω . The letter x

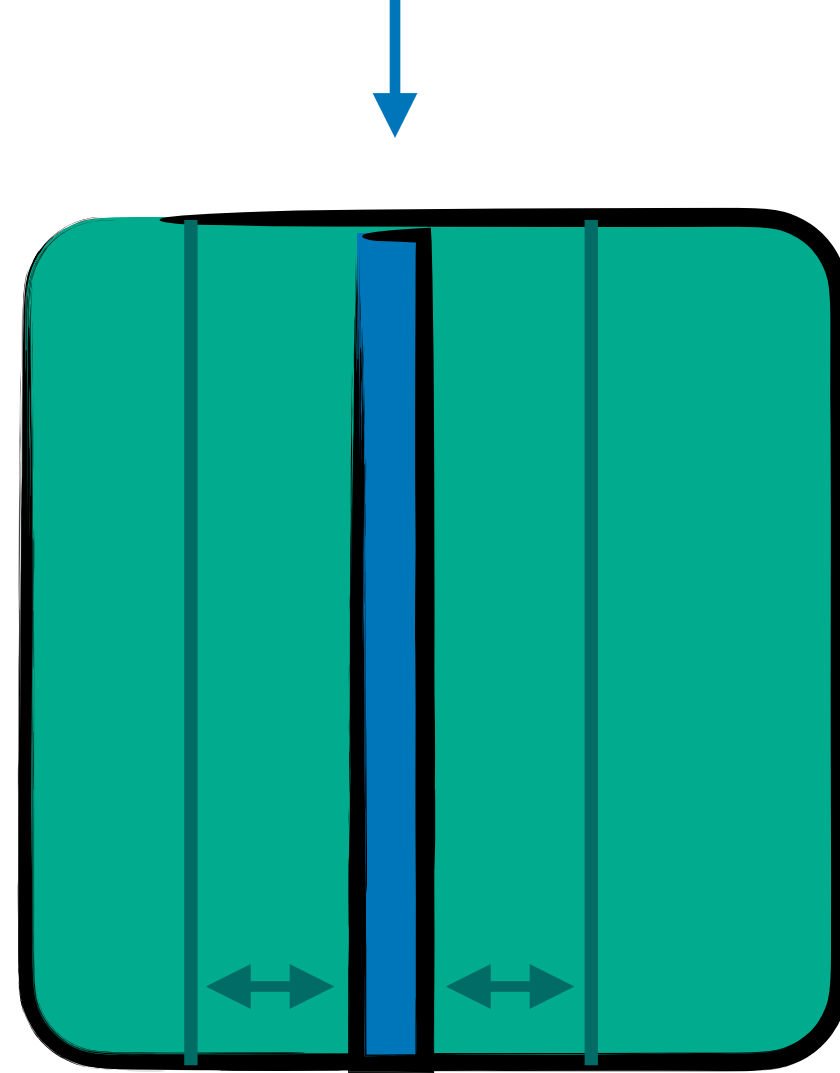
Some of the results contained in this paper were presented to the Society, February 22, 1941 and September 2, 1941; received by the editors March 31, 1943.

⁽¹⁾ Research under a grant-in-aid from the Carnegie Corporation of New York.

Best Average Power

Wald defines optimality as a test having best average power against alternatives “equally distant” from subspace defined by the parameters of interest

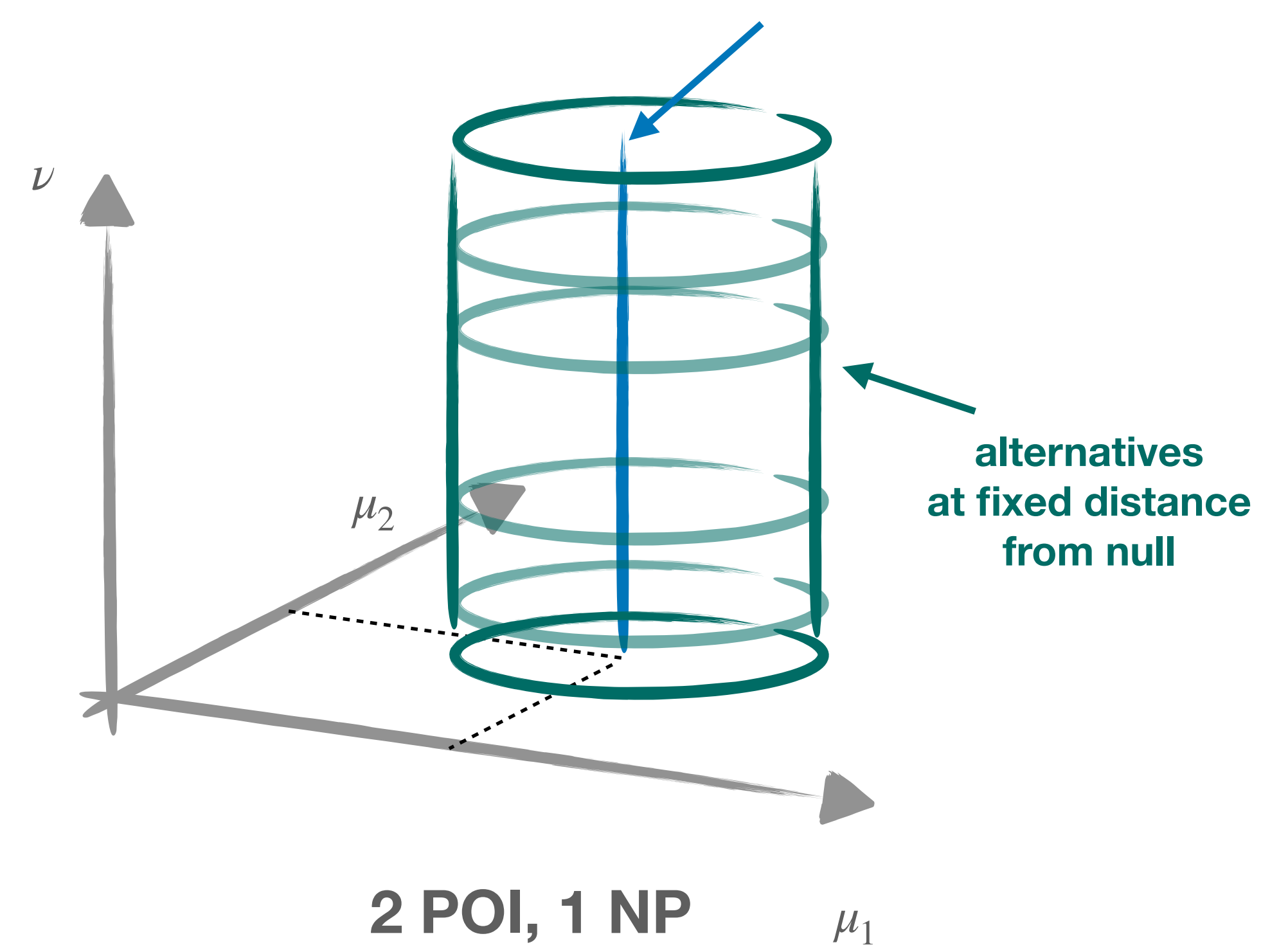
subspace of null hypo $\mu = \mu_0$



alternatives at fixed distance from null

1 POI, 1 NP

subspace of null hypo $\mu = \mu_0$



2 POI, 1 NP

Best Average Power

Gives us a clear recipe on what loss to train our network on.

- LR trick: binary cross entropy optimizes for best power for fixed alternative
- Wald: Profile LR will emerge from optimized for best average power
- ▶ **optimize on best average BXE by sampling fixed-distance alternatives and average over them. Then watch $f(x, \mu) \rightarrow t_\mu(x)$**

Algorithm 1 Training a Test Statistic with Best Average Power

Require: η : learning rate

Require: ϕ_0 : initial parameters

Require: $\theta \sim p(\theta)$, $\theta \sim p(\theta, S_c|\theta_0)$: sampling routines

1: **while** not converged **do**

2: $\theta_0 = (\mu_0, \nu_0) \sim p(\theta)$

▷ sample null

3: $\theta_i = (\mu_i, \nu_i) \sim p(\theta, S_c|\theta_0)$

▷ sample alternatives

4: $(x_i, y_i) \sim p(x|\theta_0), p(x|\theta_i)$

▷ null: $y_i = 0$, all alternatives have $y_i = 1$

5: $p_i \leftarrow s_\phi(x_i; \mu_0)$

6: $L = \sum_{\text{null,alts}} L_{\text{BXE}}(y_i, p_i)$

7: $\phi_{i+1} \leftarrow \phi_i - \eta \nabla_\phi L$

8: **end while**

9: **return** ϕ_N

Does this work?

Check on a well-known example from HEP stats: **the on-off problem**

$$p(x_1, x_2 | \mu, \nu) = \text{Pois}(x_1 | \mu s + \nu b) \text{Pois}(x_2 | \nu \tau b),$$

In this case we can solve for the true profile likelihood analytically

$$\begin{aligned} \hat{\mu} &= \frac{n - m/\tau}{s}, \\ \hat{b} &= \frac{m}{\tau}, \\ \hat{\hat{b}} &= \frac{n + m - (1 + \tau)\mu s}{2(1 + \tau)} + \left[\frac{(n + m - (1 + \tau)\mu s)^2 + 4(1 + \tau)m\mu s}{4(1 + \tau)^2} \right]^{1/2}. \end{aligned}$$

can check, whether this idea works

1007.1727v3 [physics.data-an] 24 Jun 2013

Asymptotic formulae for likelihood-based tests of new physics

Glen Cowan¹, Kyle Cranmer², Eilam Gross³, Ofer Vitells³

¹ Physics Department, Royal Holloway, University of London, Egham, TW20 0EX, U.K.
² Physics Department, New York University, New York, NY 10003, U.S.A.
³ Weizmann Institute of Science, Rehovot 76100, Israel

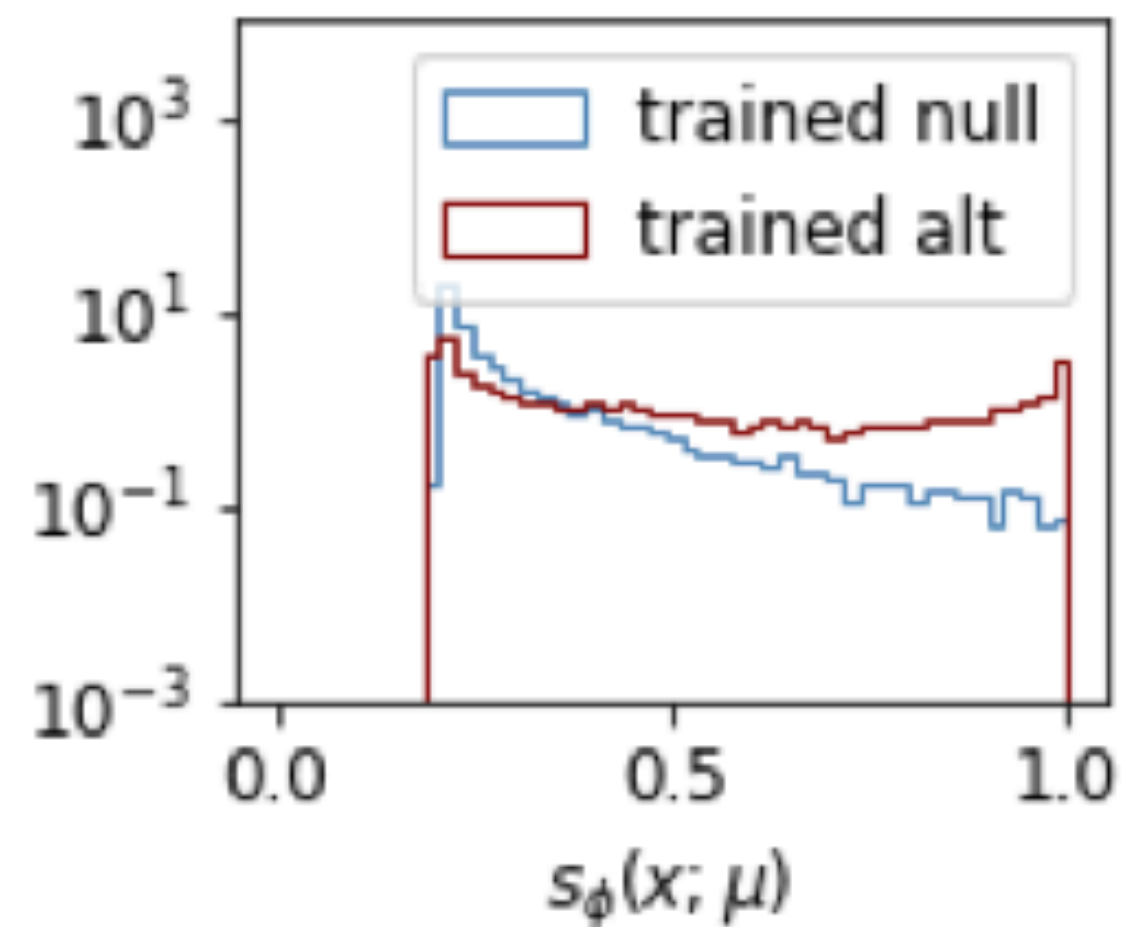
Abstract

We describe likelihood-based statistical tests for use in high energy physics for the discovery of new phenomena and for construction of confidence intervals on model parameters. We focus on the properties of the test procedures that allow one to account for systematic uncertainties. Explicit formulae for the asymptotic distributions of test statistics are derived using results of Wilks and Wald. We motivate and justify the use of

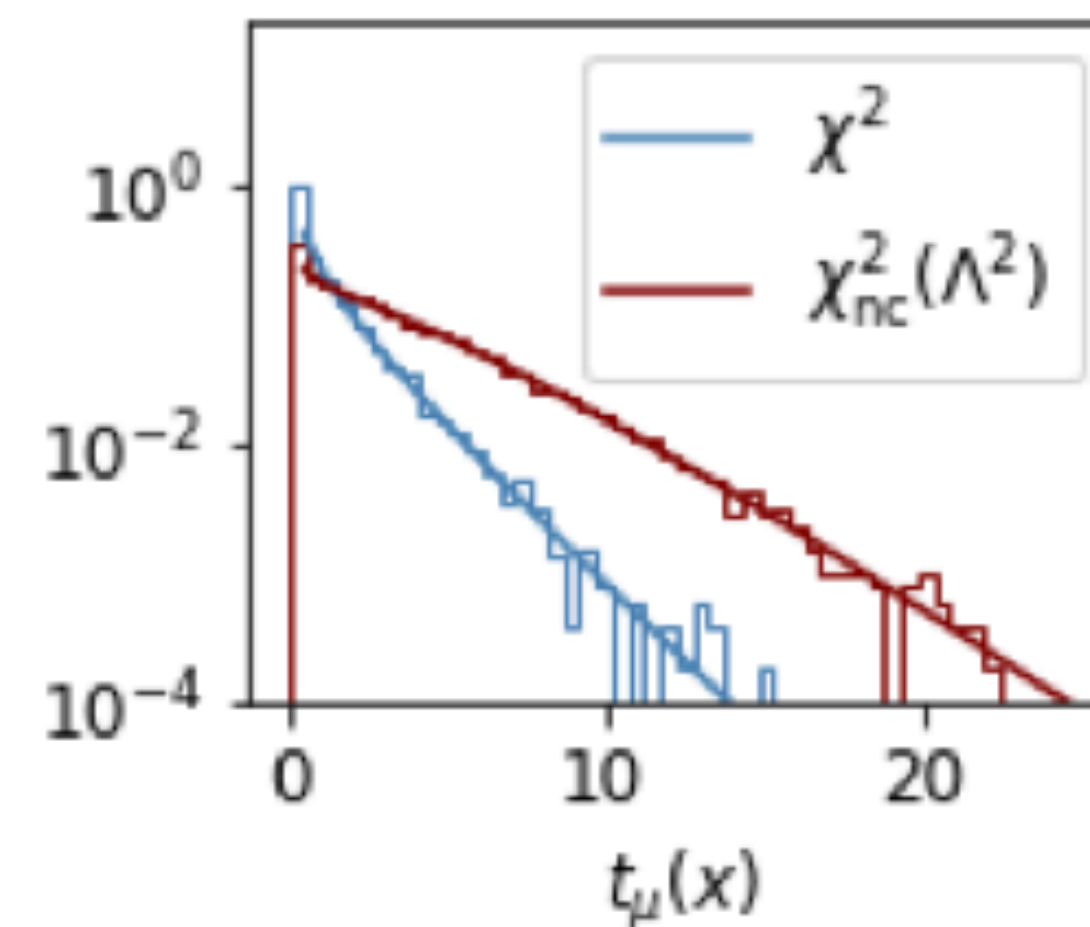
Examples

Does this work? Check on a well-known example from HEP Stats

$$p(x_1, x_2 | \mu, \nu) = \text{Pois}(x_1 | \mu s + \nu b) \text{Pois}(x_2 | \nu \tau b),$$



Neural Network Training



Analytic Result well inside
the asymptotic regime

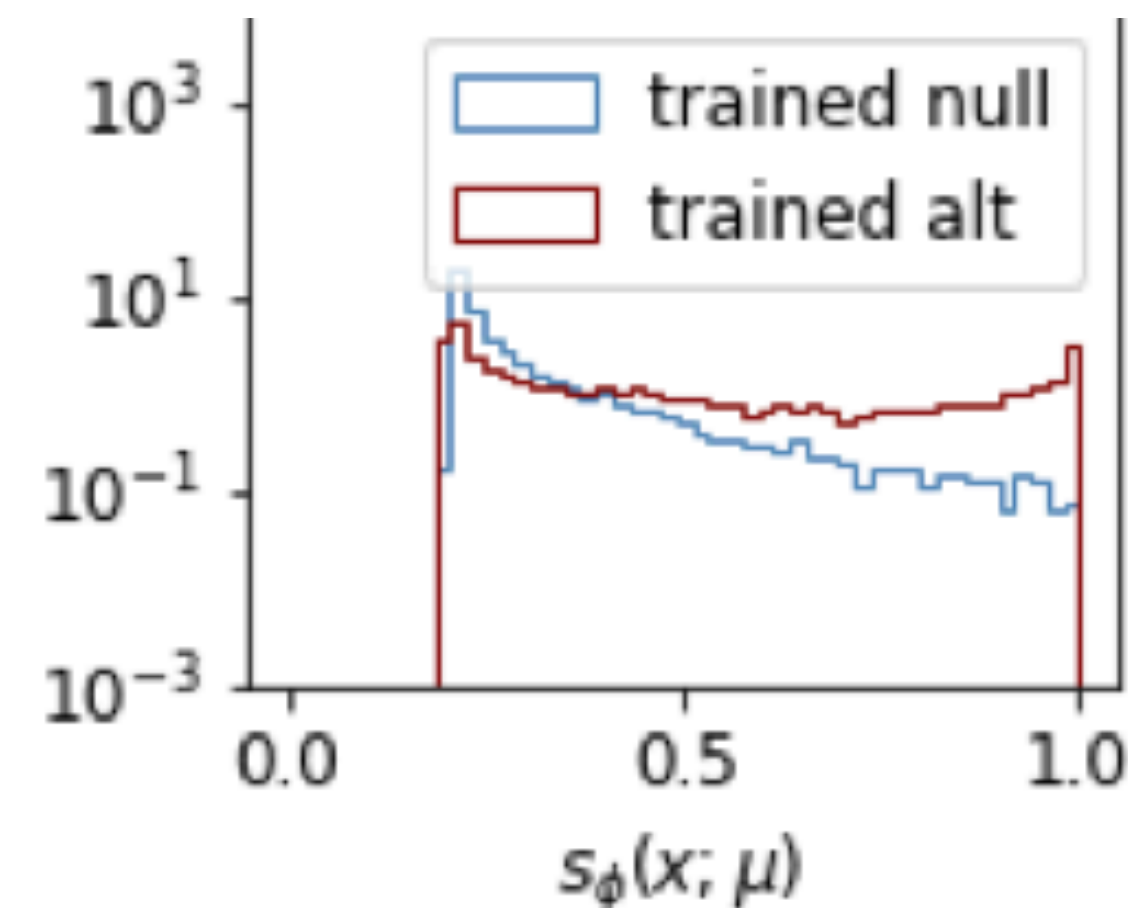
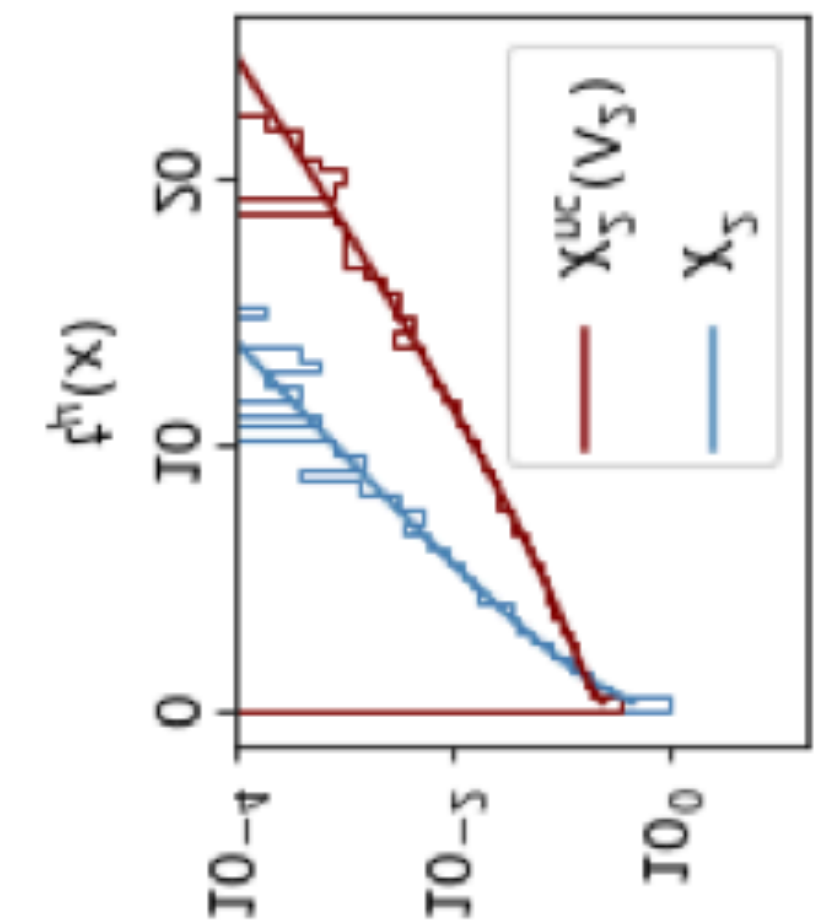
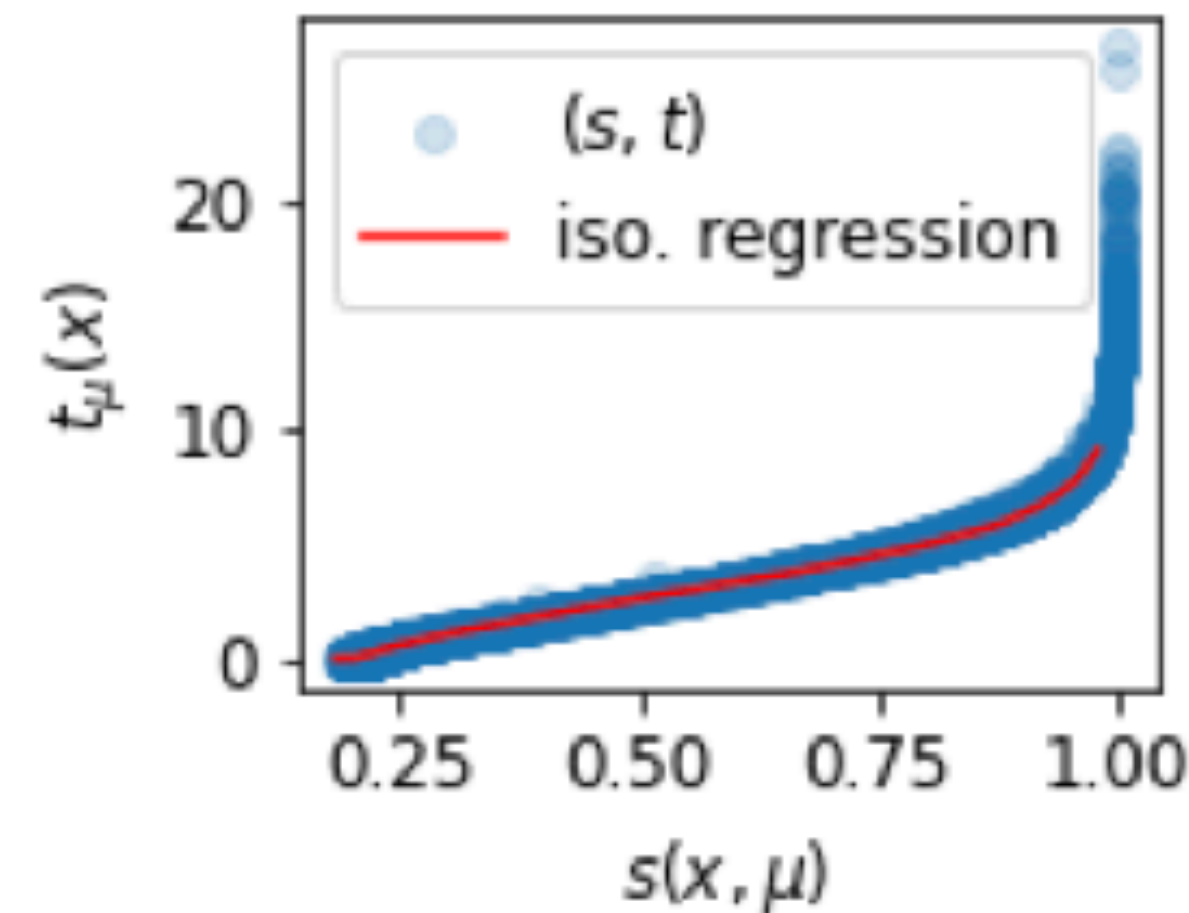
Examples

Are these two test statistics related?

Yes: they're $1 \leftrightarrow 1$

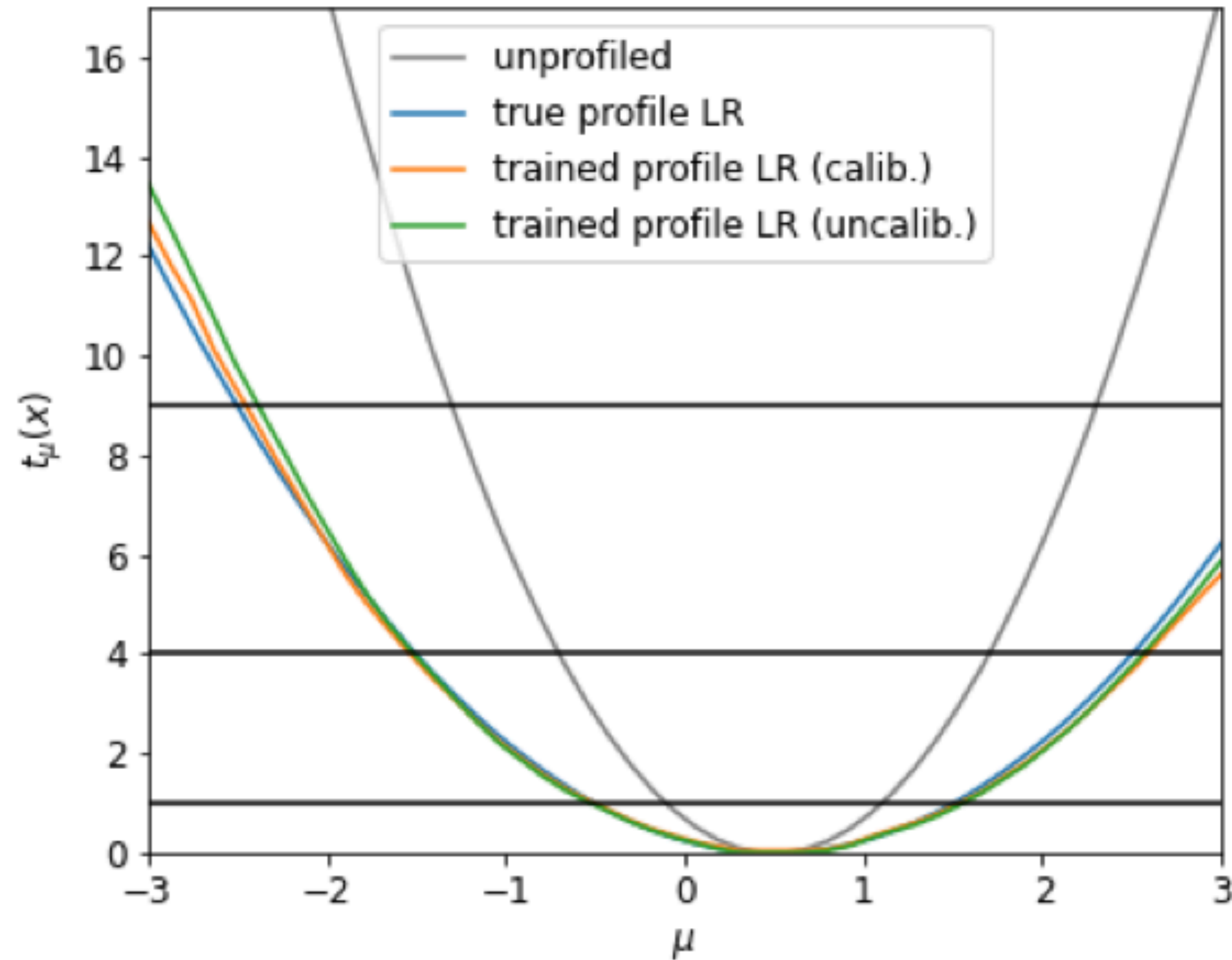
We can transform to standard χ^2 -type units of just do inference in the learned test statistic

Both will produce the same results.



Examples

We can recover the “profile likelihood” in a fully likelihood free way



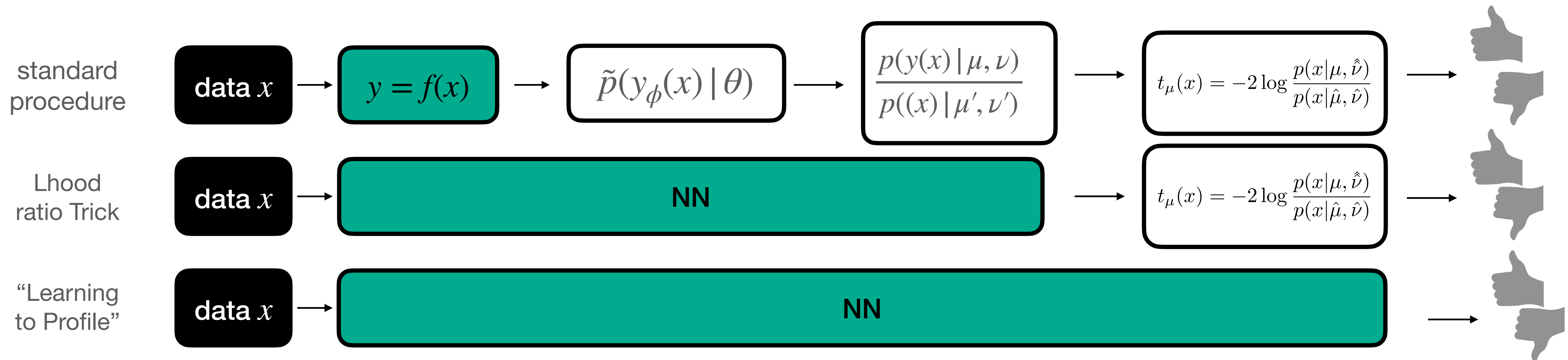
Summary

Described method to compute

$$t_{\mu}(x) = -2 \log \frac{p(x|\mu, \hat{\nu})}{p(x|\hat{\mu}, \hat{\nu})}$$

- without evaluating $p(x|\theta)$
- without running any optimiation to find $\hat{\mu}, \hat{\nu}, \hat{\hat{\nu}}$
- just using samples from $p(x|\theta)$

by choosing appropriate training procedure.



Summary

Taking Wald's optimality criterion seriously and using it as an optimization objective: extension of LR trick to case of nuisance parameters

Larger Question: As physicists, we often talk about adding knowledge to ML but open question where to add it / how much is needed

e.g. this approach manages to effectively “shortcircuit” a lot of steps we usually associate with data analysis (fitting, model building, ...), while retaining some nice properties (robustness to NPs)