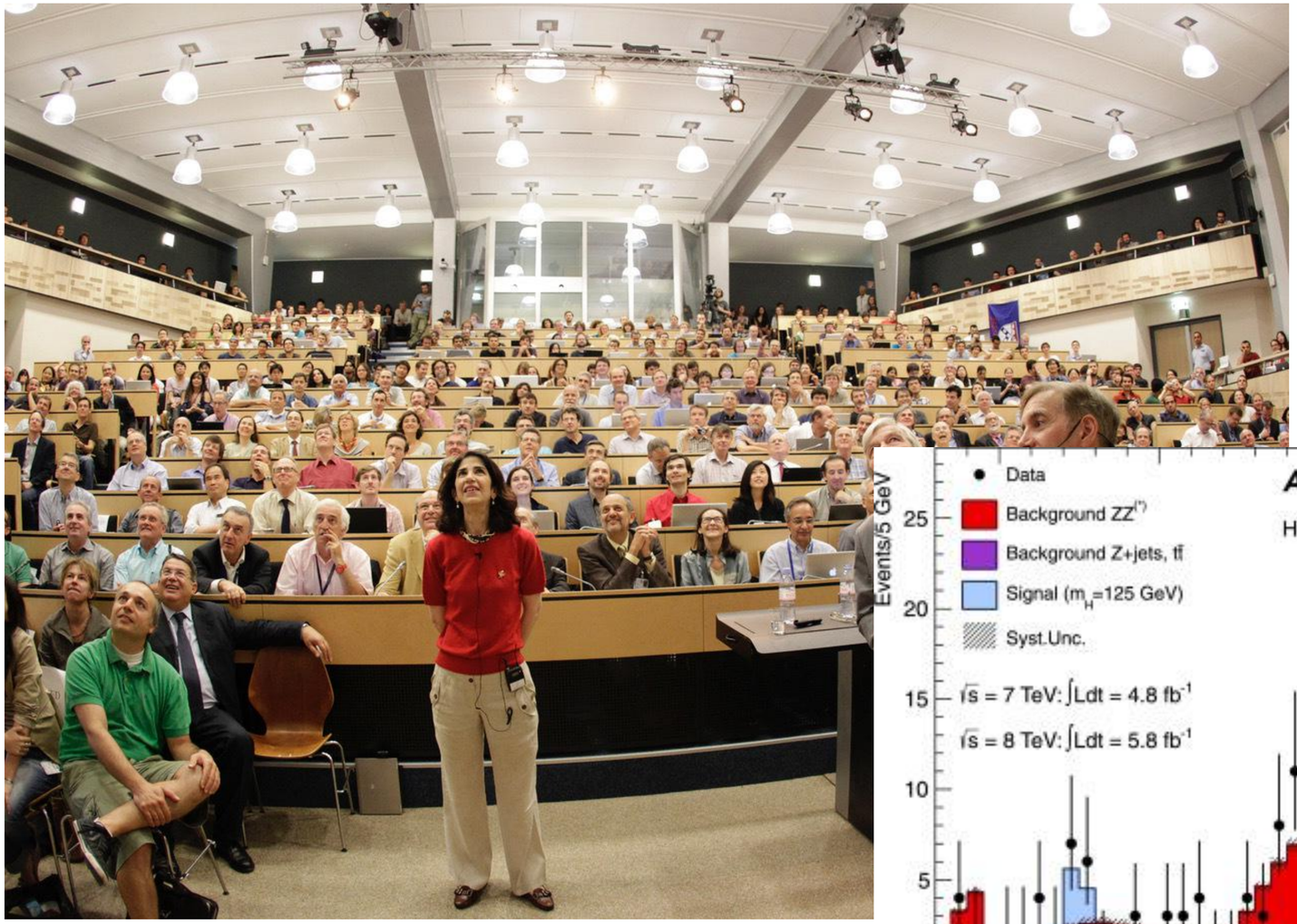
Systematics in HEP

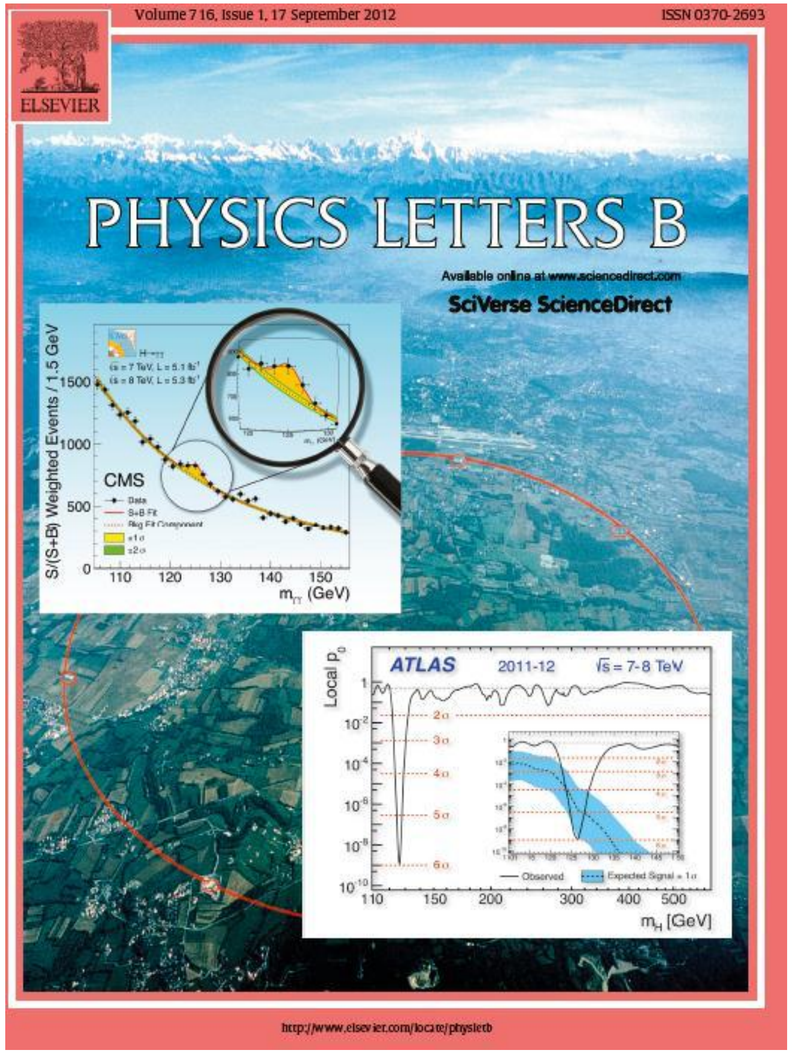
Event Selection, Limits, Discovery

Lukas Heinrich, TUM - STAMPS

10 years ago



Data
Background ZZ(*)
Background Z+jets, tt̄
Signal (m_H=125 GeV)
Syst.Unc.
√s = 7 TeV: ∫Ldt = 4.8 fb-1
√s = 8 TeV: ∫Ldt = 5.8 fb-1
ATLAS
H→ZZ(*)→4l
Events/5 GeV
m4l [GeV]

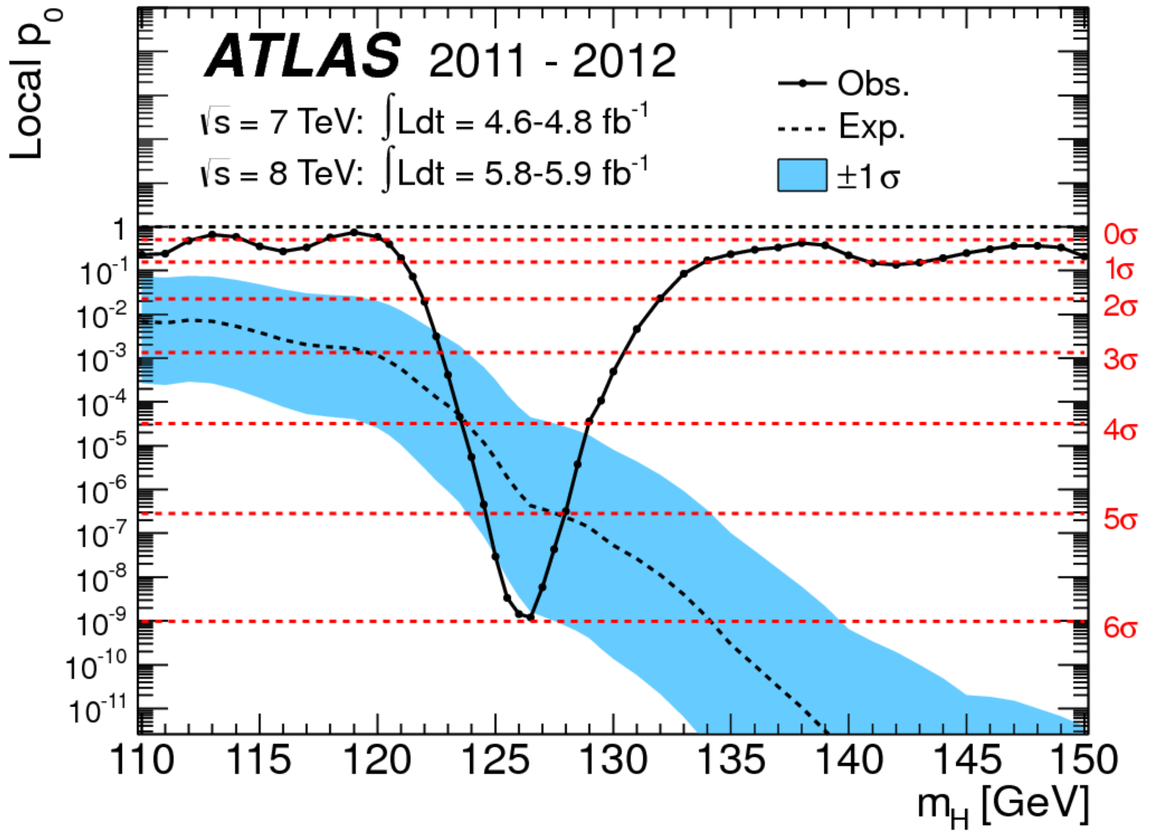

Volume 716, Issue 1, 17 September 2012
ISSN 0370-2693
ELSEVIER
PHYSICS LETTERS B
Available online at www.sciencedirect.com
SciVerse ScienceDirect
CMS
S/(S+B) Weighted Events / 1.5 GeV
mγγ (GeV)
ATLAS 2011-12 √s = 7-8 TeV
Local p0
mH [GeV]
http://www.elsevier.com/locate/physletb

10 years ago

Not only a significant moment for physics, but also:

LHC reached a **new level of statistical sophistication** in producing these results

This Talk: will try to give a feel for how these results come about



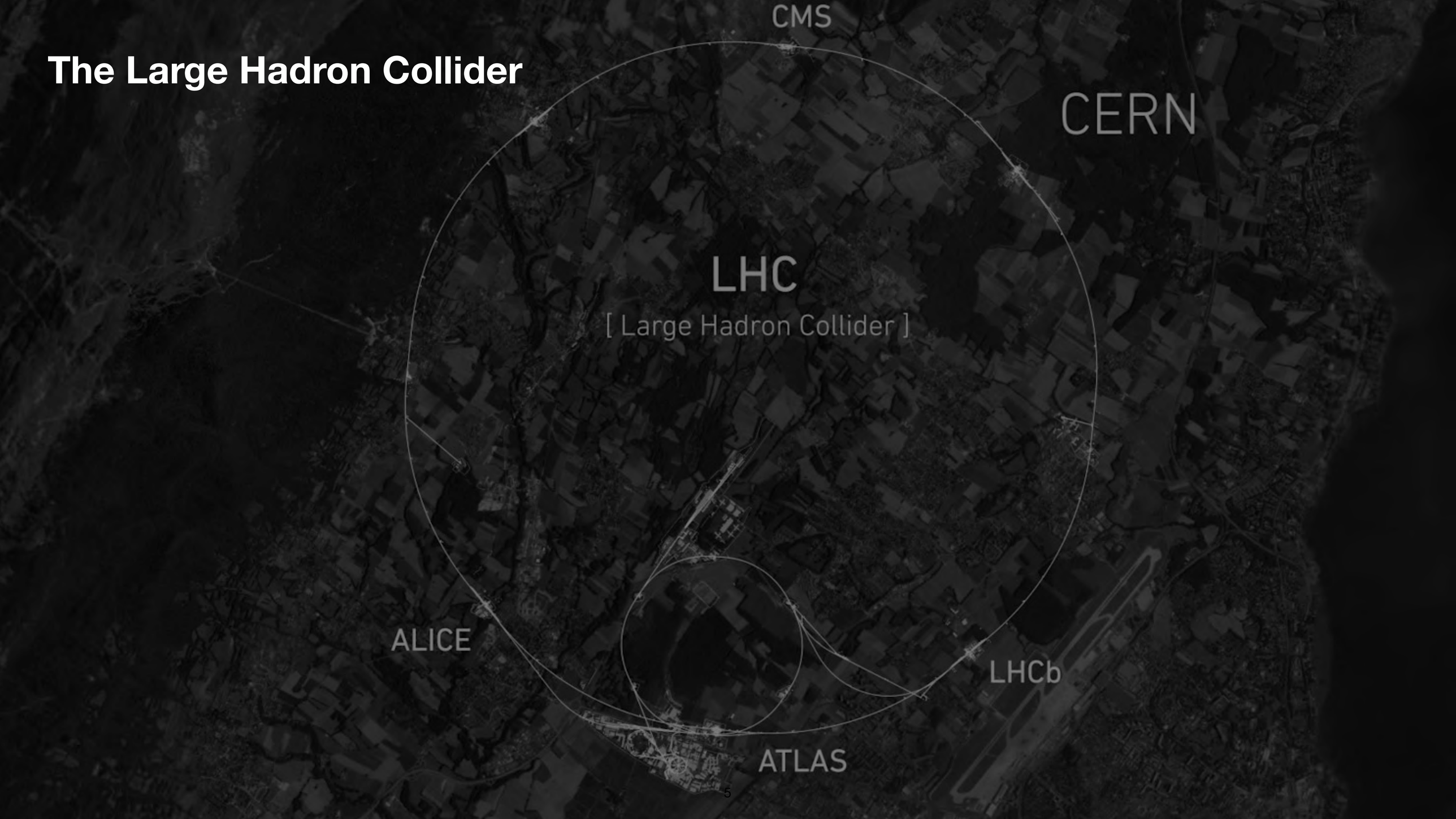
Goals for the talk

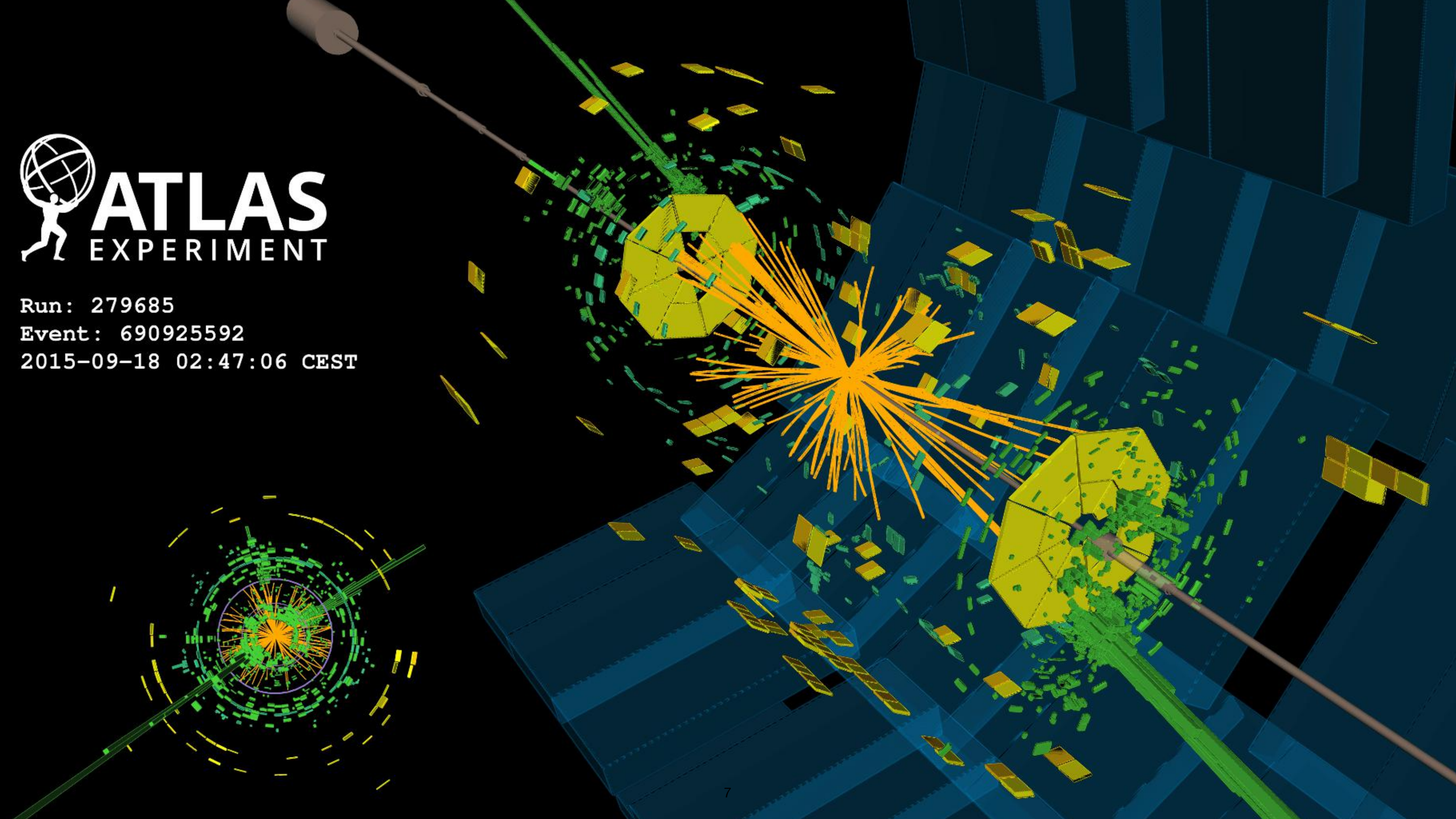
I will try to summarize

- how we build the statistical model at the LHC
  esp. how we incorporate nuisance parameters
- describe the current practice for statistical tests
- discuss some frequent assumptions made during inference
- focus on frequentist viewpoint as its dominant at LHC

Assuming familiarity w/ stats but less with HEP

The Large Hadron Collider



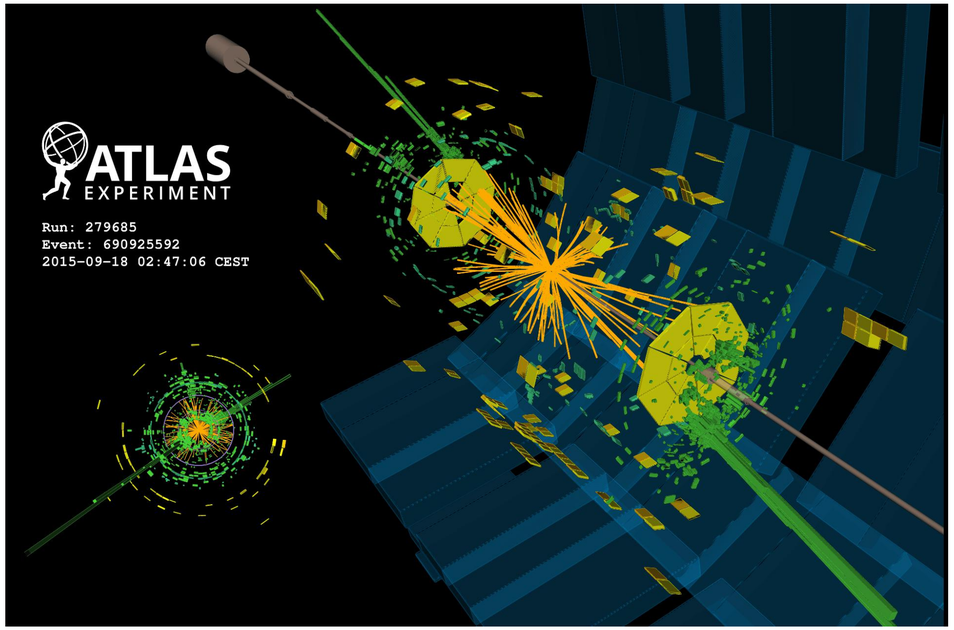

ATLAS
EXPERIMENT
Run: 279685
Event: 690925592
2015-09-18 02:47:06 CEST

What we would like to do

$$p(\text{theory}\,|\,\text{data}) = \frac{p(\text{data}\,|\,\text{theory})p(\text{theory})}{p(\text{data})}$$

How to describe the data?

The forward model of the data is pretty complicated



$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\Psi}\not{D}\Psi + h.c. + \bar{\Psi}_i y_{ij} \Psi_j \phi + h.c. + |D_\mu \phi|^2 - V(\phi)$$

data

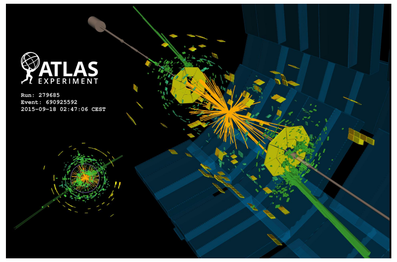
theory

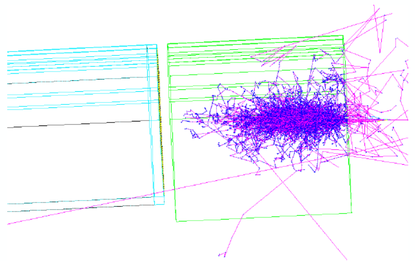
$$p(\text{data} \mid \text{theory})?$$

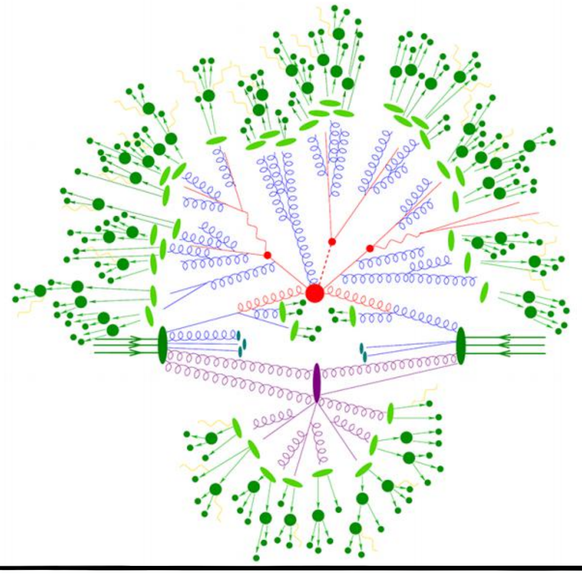
How to describe the data?

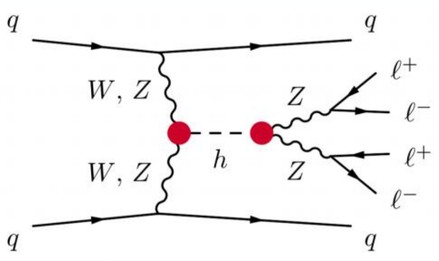
Fortunately: particle physics has very strong foundational modeling:

- can describe generative process as a (very deep) Markov Model









$$\mathcal{L} = -\tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\psi}\not{D}\psi + \text{h.c.} + \bar{\psi}_i y_{ij} \psi_j \phi + \text{h.c.} + |D_\mu \phi|^2 - V(\phi)$$

$p(x|z_d)$

observed voltages, etc in read-out electronics

$p(z_d|z_h)$

particle interaction with dense material

$p(z_h|z_p)$

particle evolution (decays, radiation)

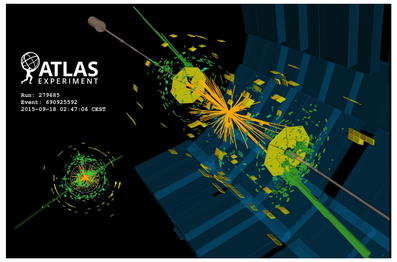
$p(z_p|\theta)$

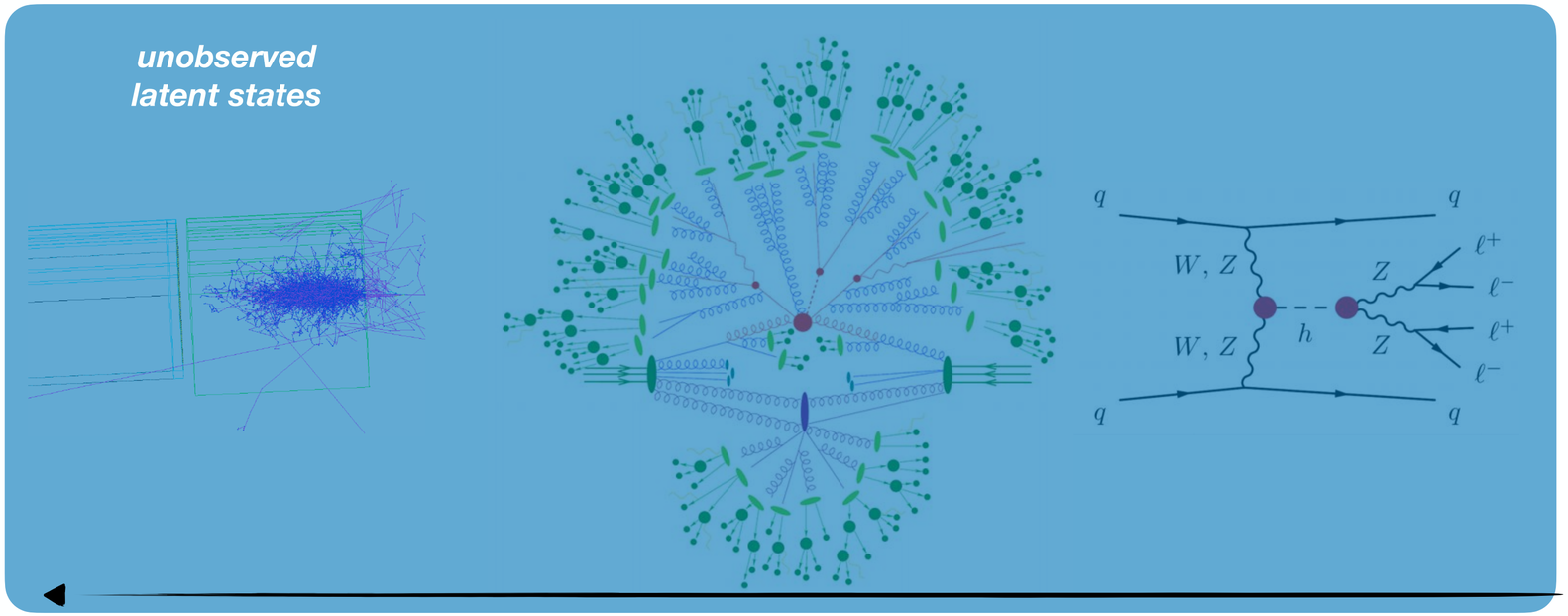
core physics process

How to describe the data?

Unfortunately, all this latent evolution is not observable

all we get is the final data x: ***integrate out all latent states***





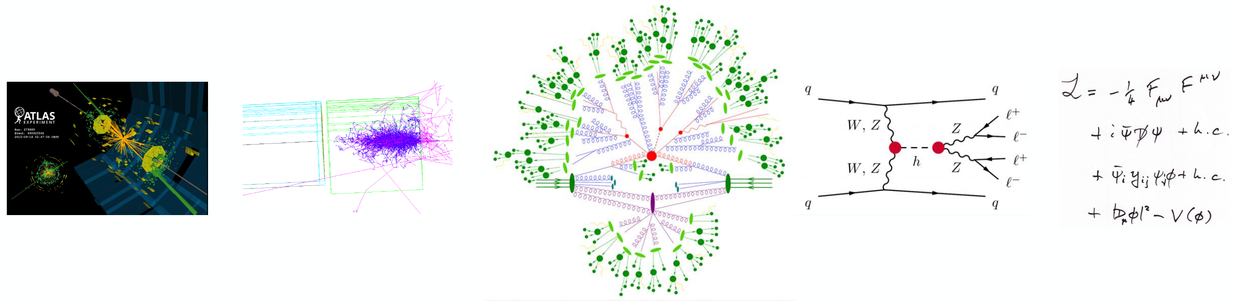
$$\mathcal{L} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\Psi}\not{D}\Psi + \text{h.c.} + \bar{\Psi}_i y_{ij} \Psi_j \phi + \text{h.c.} + |D_\mu \phi|^2 - V(\phi)$$

$$p(x|\theta) = \int \mathrm{d}z\; p(x,z) = \int \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p\; p(x|z_d) p(z_d|z_h) p(z_h|z_p) p(z_p|\theta)$$

The parameters θ

One more issue: the "theory" space is not the only thing effecting the data

- **every step of the forward process comes with its own parameters**
  (we understand the process generally but need additional knobs to model the data)



$$p(x|\theta) = \int \mathrm{d}z_d \mathrm{d}z_h \mathrm{d}z_p \quad p(x|z_d, \theta_x)\, p(z_d|z_h, \theta_d)\, p(z_h|z_p, \theta_h)\, p(z_p|\theta_p, \theta_{\mathrm{th}})$$

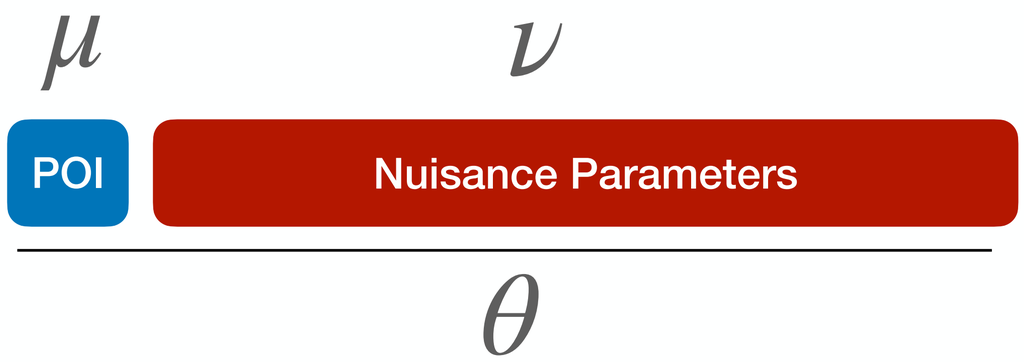
$p(x|z_d)$ $p(z_d|z_h)$ $p(z_h|z_p)$ $p(z_p|\theta)$

nuisance parameters

core "theory" parameters of inferest (e.g. "Higgs Mass"

Inference Goals

The nuisance parameters are necessary but seriously complexify the model



But not what we are interested in producing results on them.
Need to reflect this in inference methods

The Saving Grace: Simulators

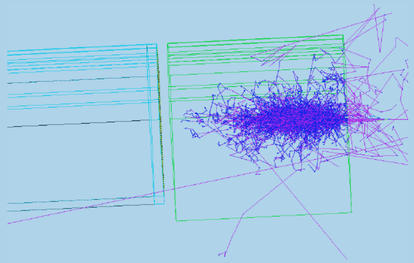
So far a pretty bleak picture: but we still do science, what gives?

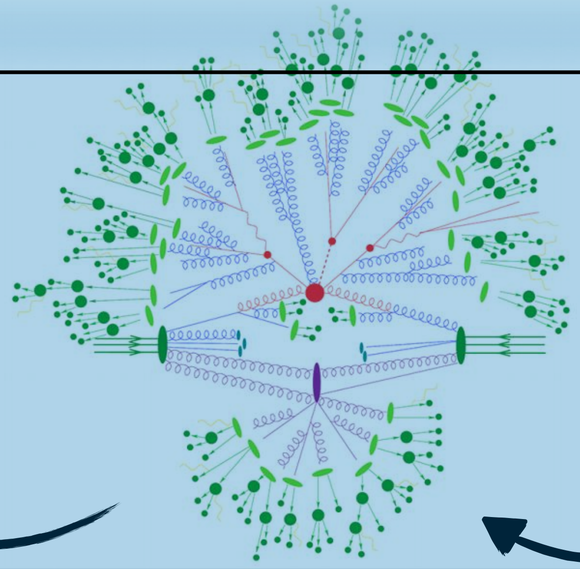
Simulators!

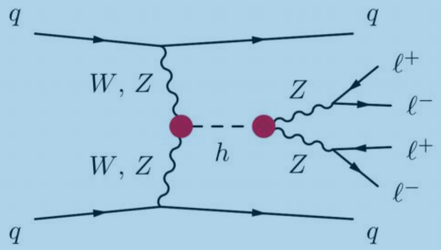
The Saving Grace: Simulators

While we cannot *evaluate the likelihood,* we can *sample it.*

$$x \sim p(x \mid \theta)$$







$$\mathcal{L} = -\tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} + i\bar{\Psi}\not{D}\Psi + \text{h.c.} + \bar{\Psi}_i y_{ij} \Psi_j \phi + \text{h.c.} + |D_\mu \phi|^2 - V(\phi)$$

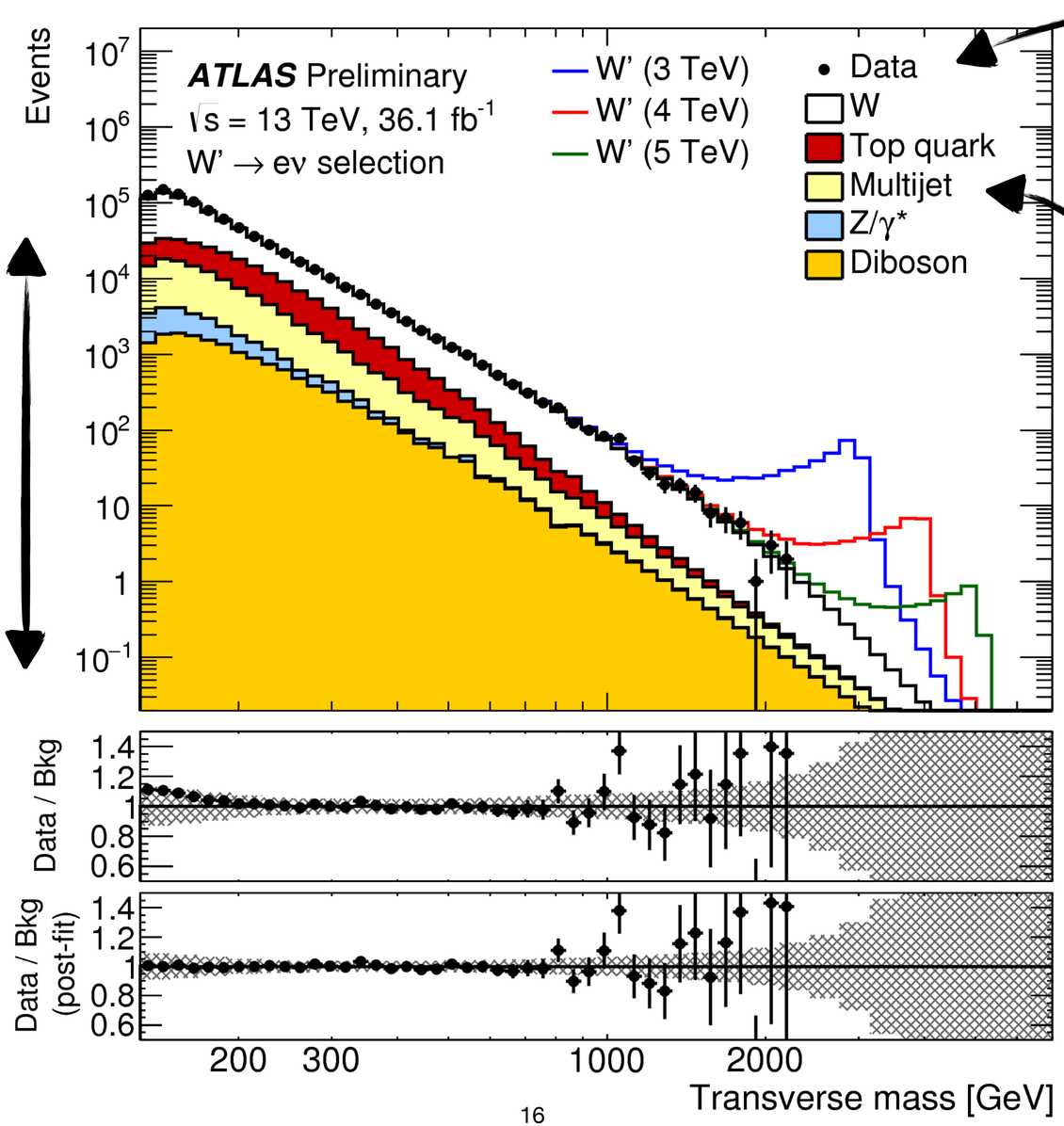
$$x \sim p(x|z_h, \theta_x) \qquad z_d \sim p(z_d|z_h, \theta_d) \qquad z_h \sim p(z_h|z_p, \theta_h) \qquad z_p \sim p(z_p|\theta, \theta_p)$$

x

θ

Extremely High-Quality Simulation

7 orders of magnitude in density



observed data

different theories

Frequentist Statistics and Likelihood-Free Inference

HEP one of major domains where frequentist statistics is used most results

One (possible) reason:
Freq. Stats is fundamentally more amenable to likelihood-free inference

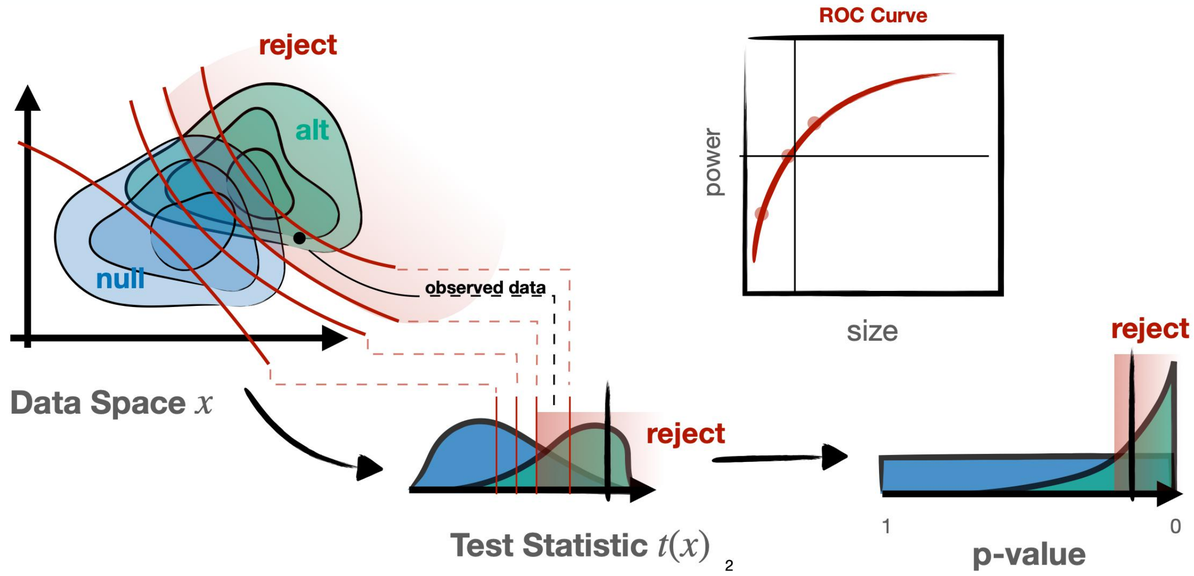
Example: Hypothesis Tests

- **just need be able to sample**

$$x \sim p(x \mid \theta)$$

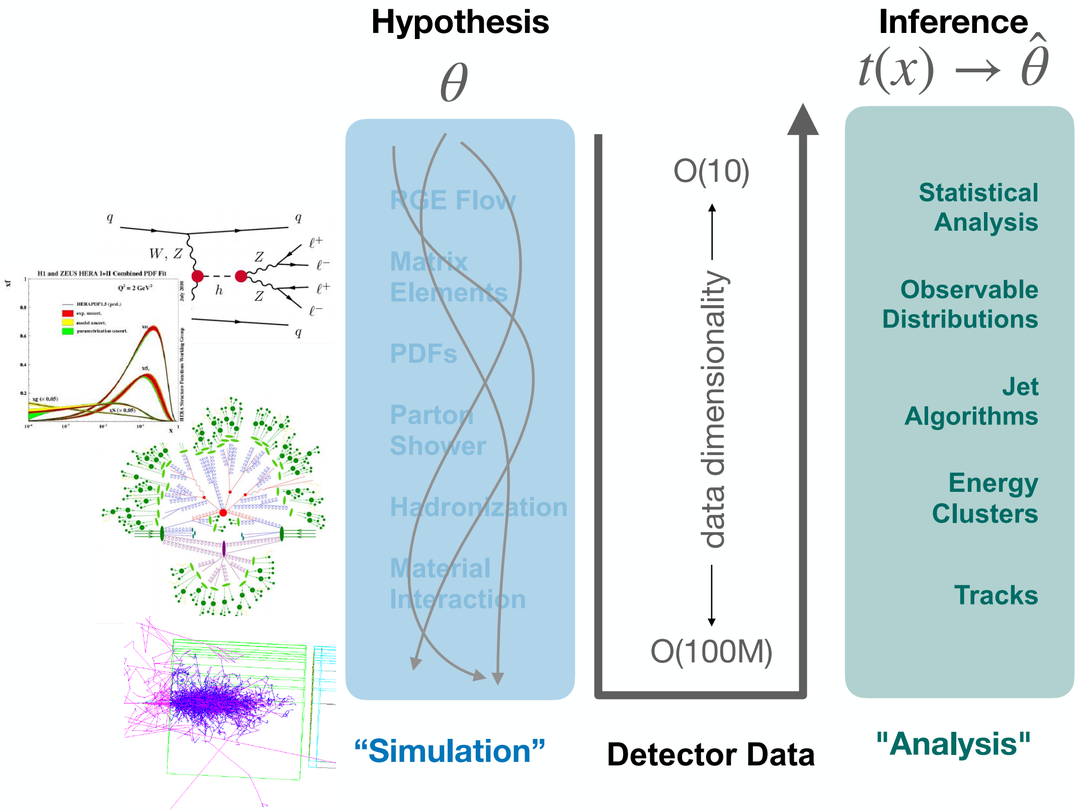
- **and be able to evaluate test statistic**

$$x \to t(x)$$



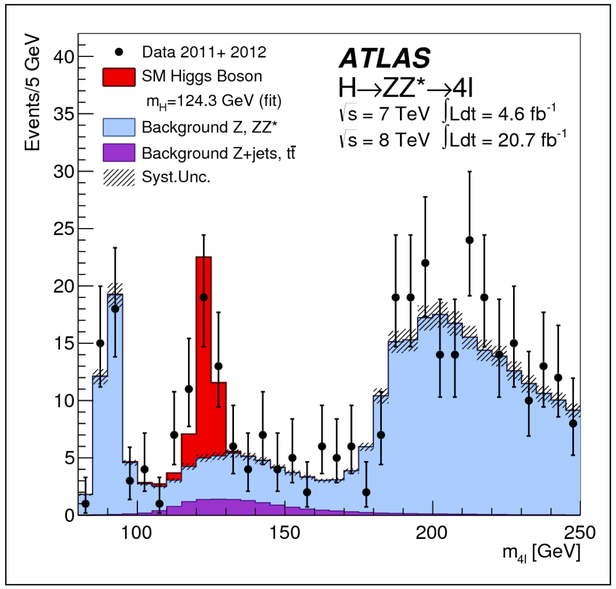
Compressing the data

Given the freedom of frequentist analysis, most of LHC analysis is about finding good data representations



Approximately follows a step-by-step inference / inversion of the data

In some sense just a very sophisticated way to build summary statistics of the data

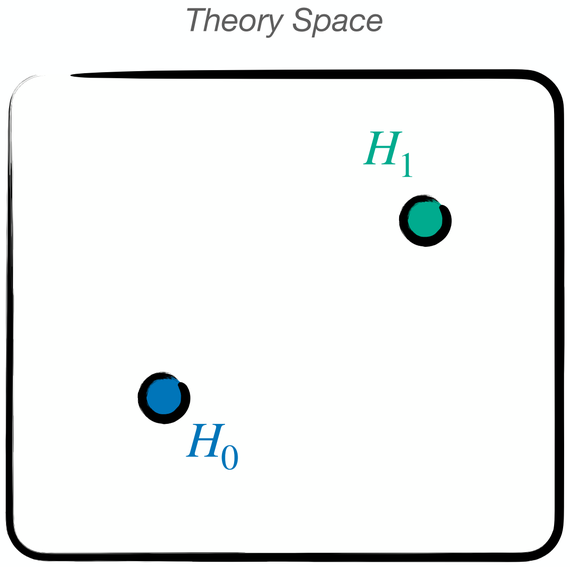


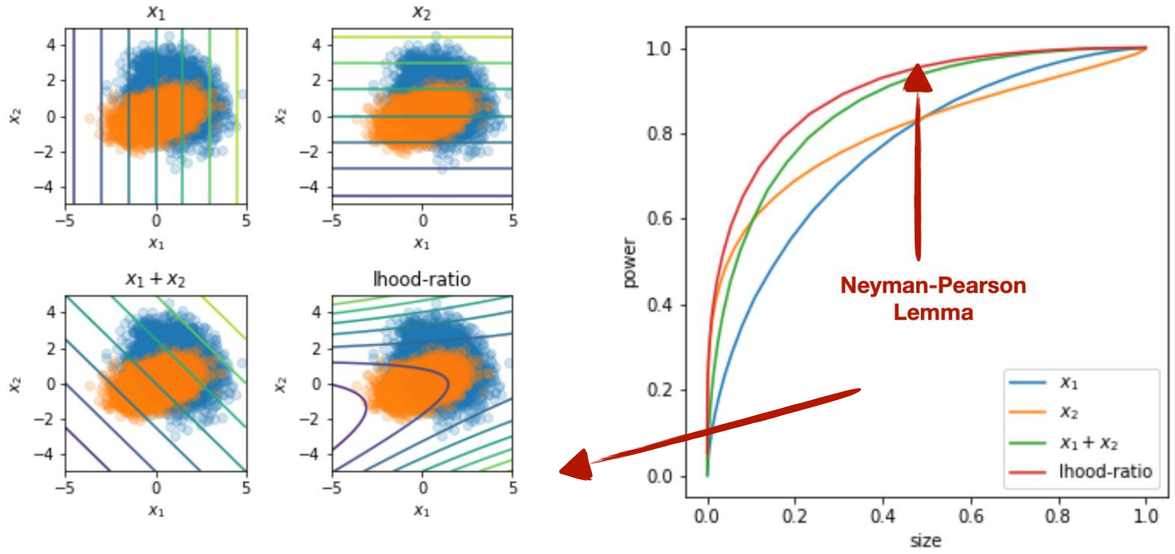
Likelihood are still useful

Even though frequentist analysis does not need likelihoods, they are still useful in order to find good test statistics

Classic Neyman-Pearson Lemma: optimal binary test
requires evaluating the likelihood (ratio)

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

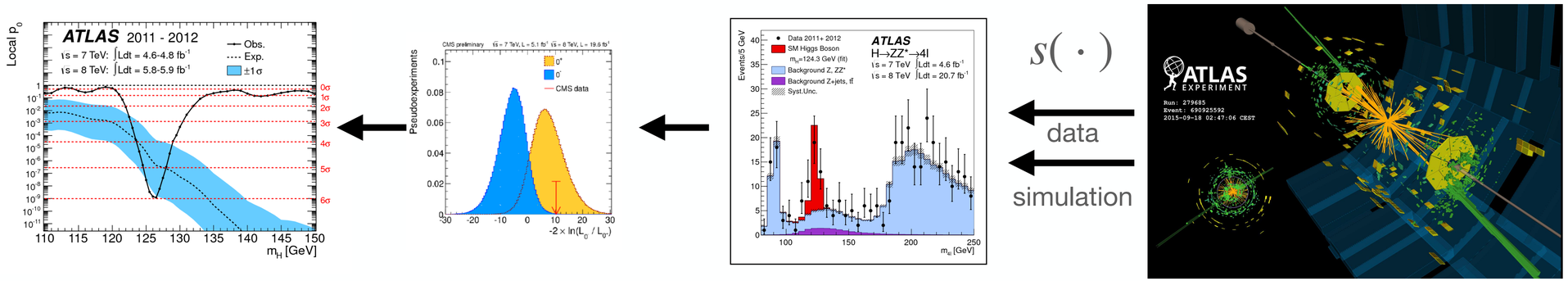




The usual workflow

In light of this, *the usual strategy in HEP* uses "approximate likelihoods" designed to model densities of summary statistics by using samples from simulation (e.g. histograms, kernel densities, …)

$$p(s \mid \theta) = p(s(x) \mid \theta)$$



"Likelihood-ful" Inference

test statistics for inference

summary statistics / density estimation

raw data

Most work goes into carefully building this "approximate likelihood"

(recently lots of interest in"more likelihood-free" methods, cf. later)

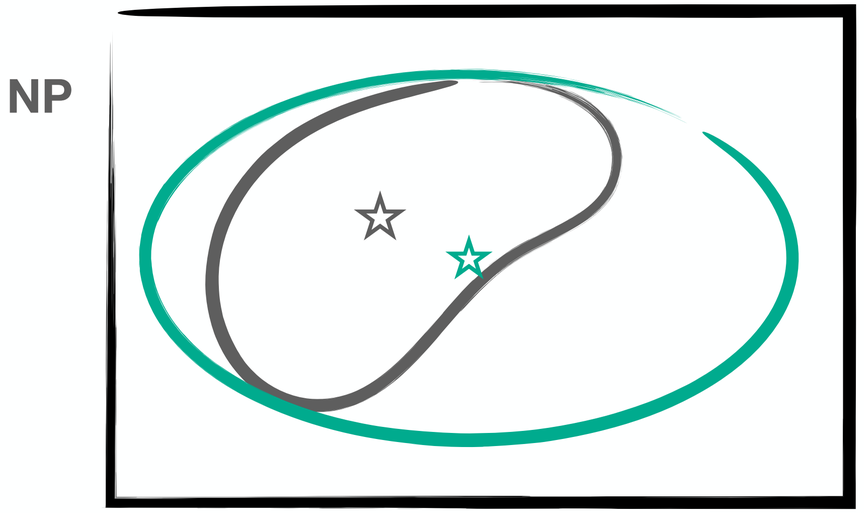
Advantages & Disadvantages of this workflow

Analysis via summary statistics: $p(x = t(\text{data}), \theta)$

Curse: by definition lose sensitivity (data processing inequality)

Blessing: by the same token gives freedom to tune for desiderata ("planned degradation")

- **e.g. design $t(x)$ to be robustness against NPs, remove correlations**

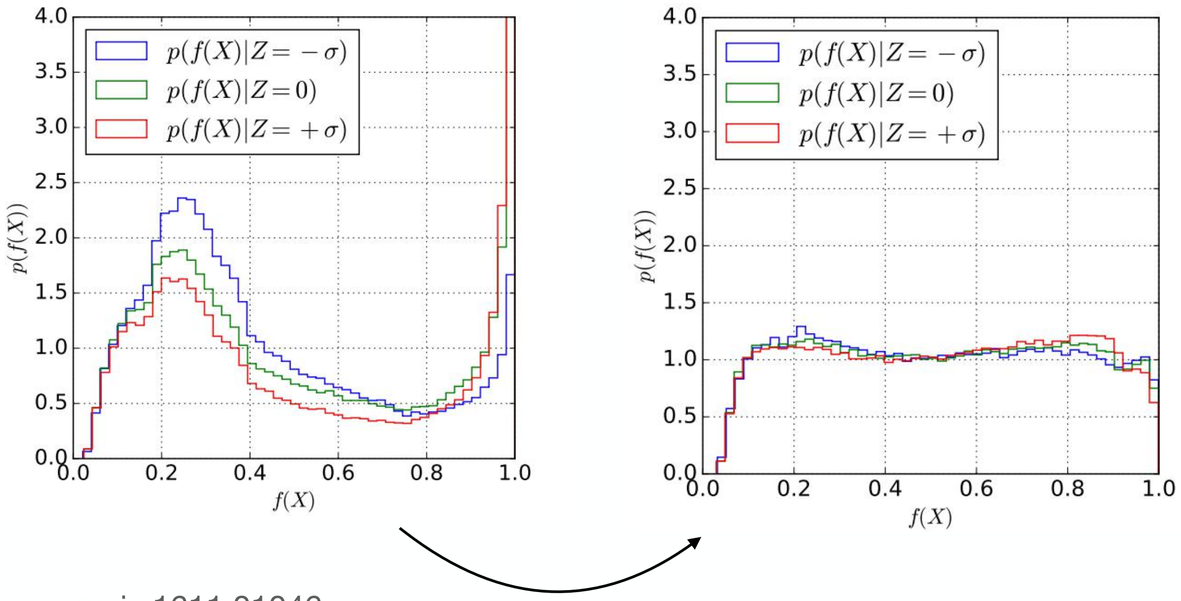


ML for controlling systematics

Using ML and “planned degradation” to achieve pivotal projections

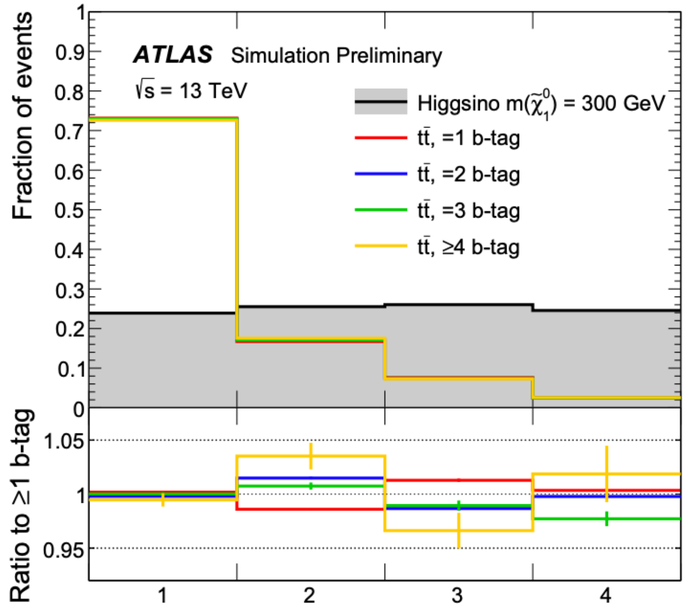
Goal: representations of the data that

- **good at** separating sig v. bkg
- **bad at** measuring NPs (i.e. invariant)



arxiv:1611.01046

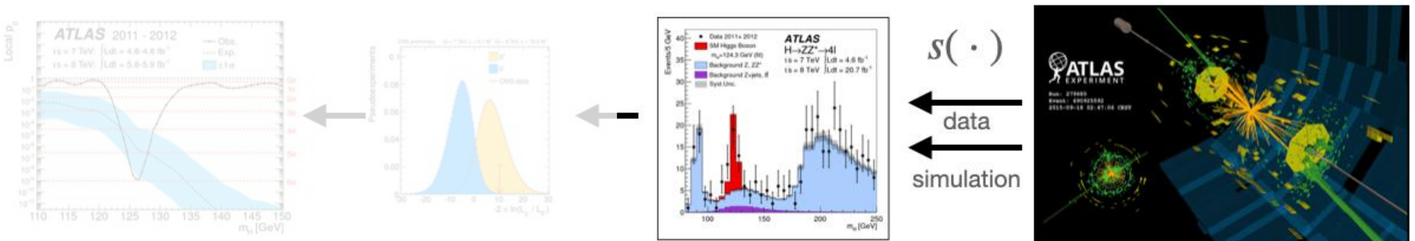
adversarial training



arxiv:2106.09609

arxiv:2001.05310 distance correlation penalty

Building the model



"Likelihood-ful" Inference

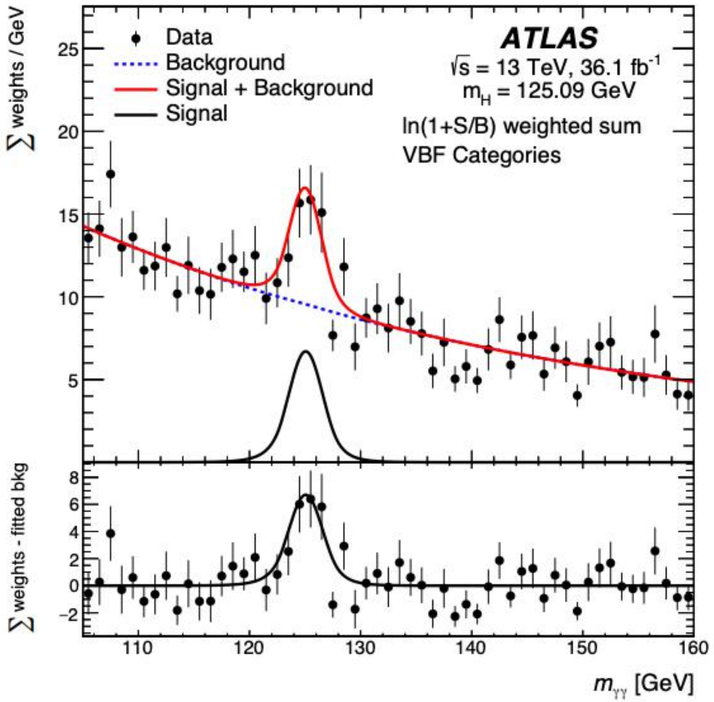
test statistics for inference

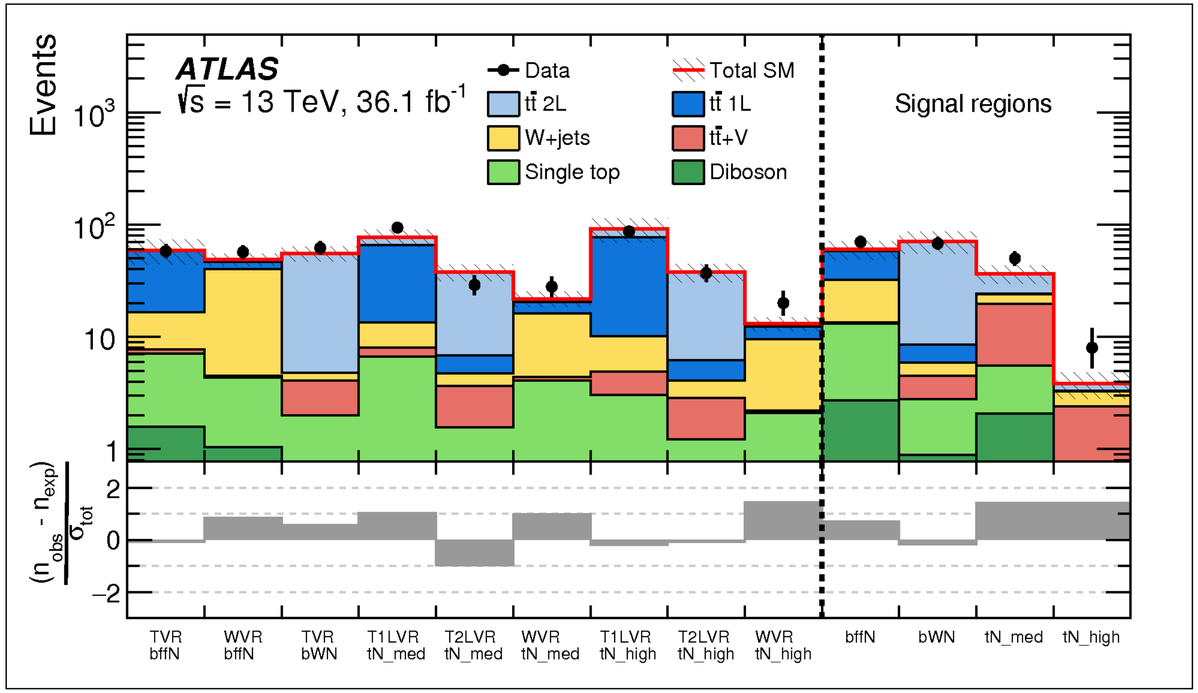
summary statistics / density estimation

raw data

Focus on building the low-dim. model

Two broad styles:





$$p(\{x_i\}\,|\,\theta) = \mathrm{Pois}(N\,|\,\lambda(\theta)) \prod_{i=0}^{N} p(x_i\,|\,\theta)$$

“Unbinned”

$$p(\{n_b\}\,|\,\theta) = \prod_{b} \mathrm{Pois}(n_b\,|\,\lambda_b(\theta))$$

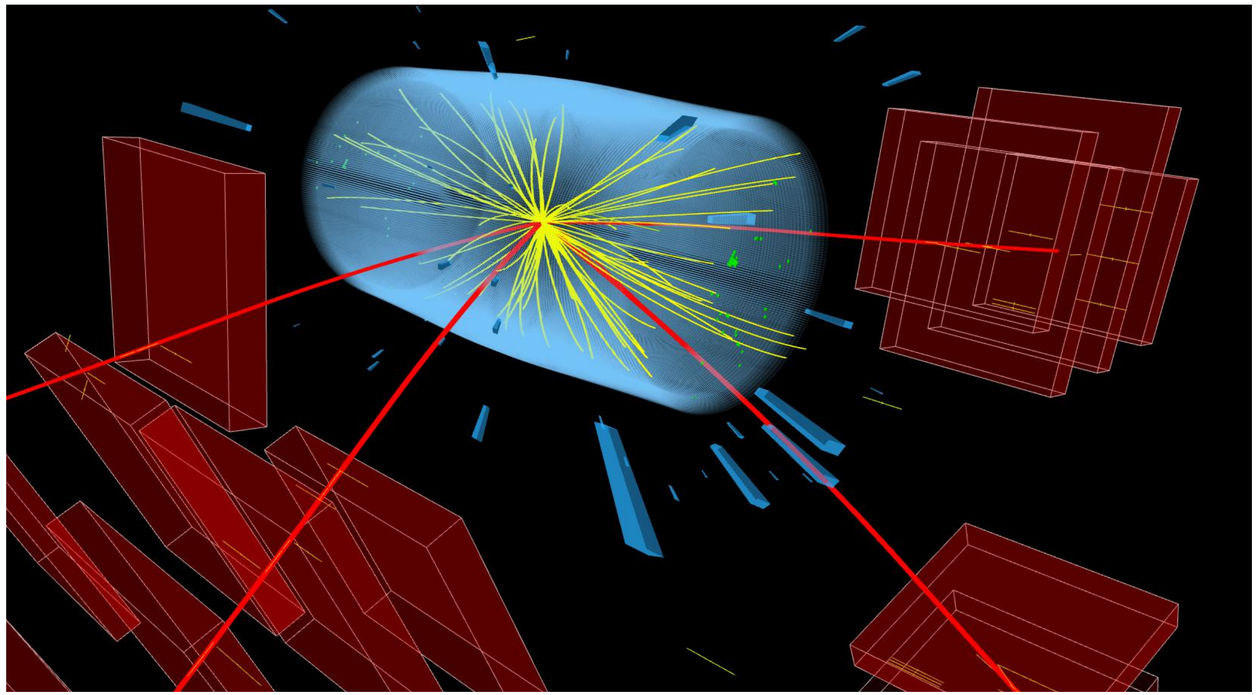
“Binned”

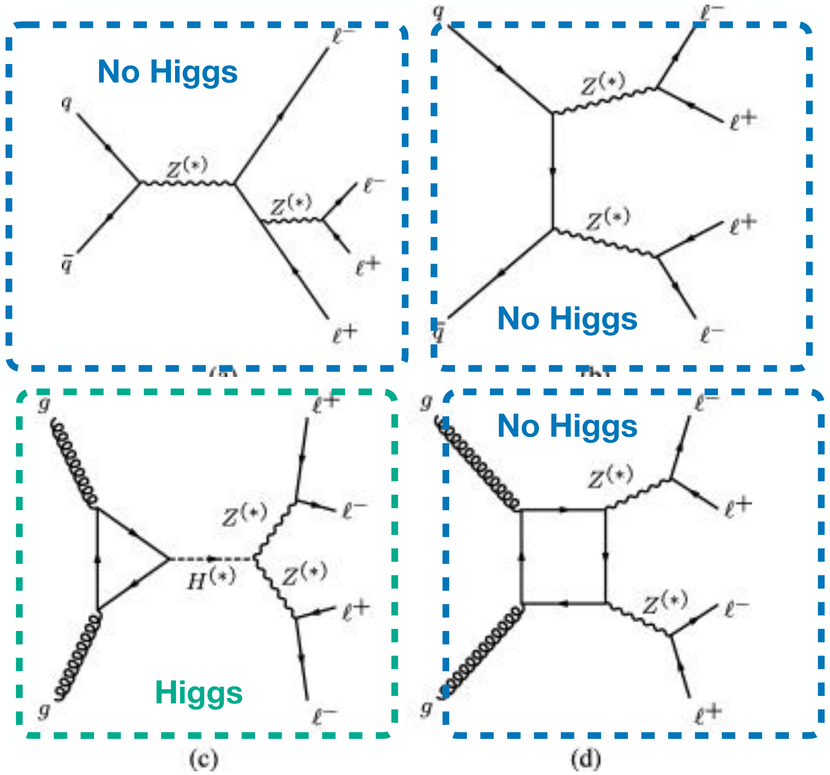
(equivalent to unbinned w/ step-wise constant)

Building the model: Mixture Models

Fundamentally, Quantum Mechanics doesn't allow us to predict, exactly what "process" happens: Particle physics will always be a mixture model.

$$p(x \mid \theta) = \sum_{\text{proc}} c_i(\theta) p(x \mid \text{proc}, \theta)$$





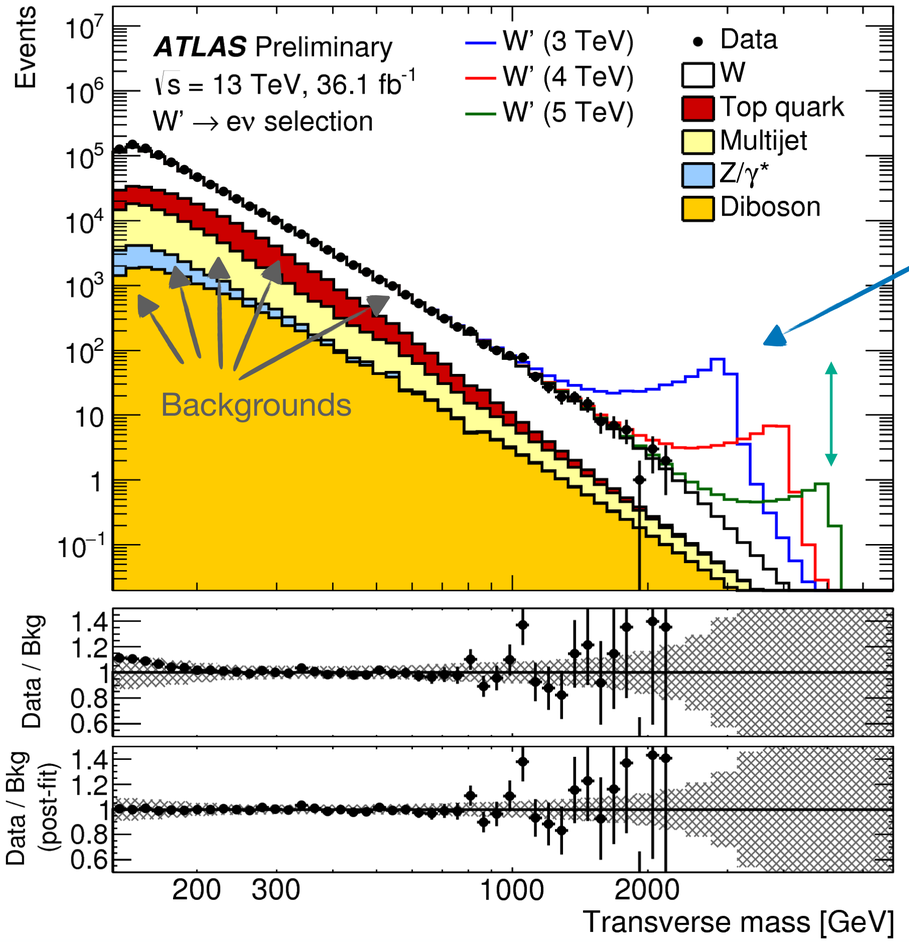
Building the model: Mixture Models

Useful categorization: “signal” and “background” components

$$\mathrm{All} = c_{\mathrm{B}}(\nu)\mathrm{Bkg}(\nu) + c_{\mathrm{S}}(\theta)\mathrm{Sig}(\theta, \nu)$$

Background:
only affected by nuisance parameter

Signal:
affected by nuisance and POIs

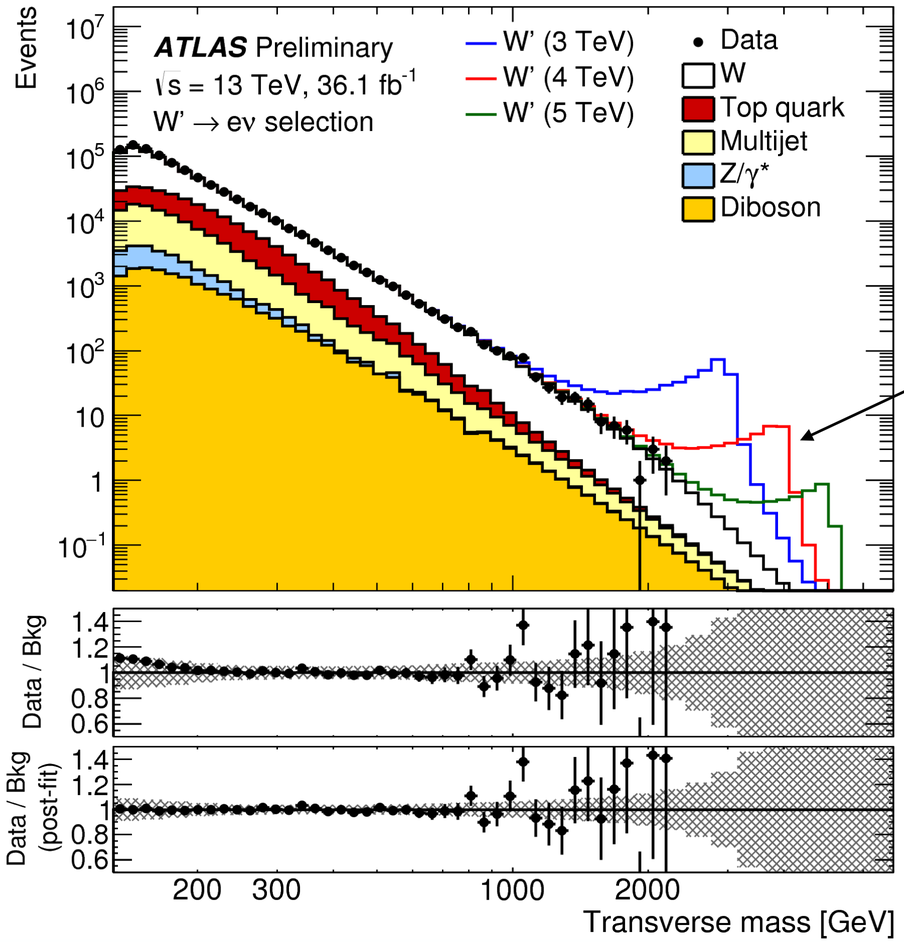


Mass Parameter (a POI) only affects signal

component size (also a POI) only affects signal

Interpolation

Simulators are very good, but very expensive as well



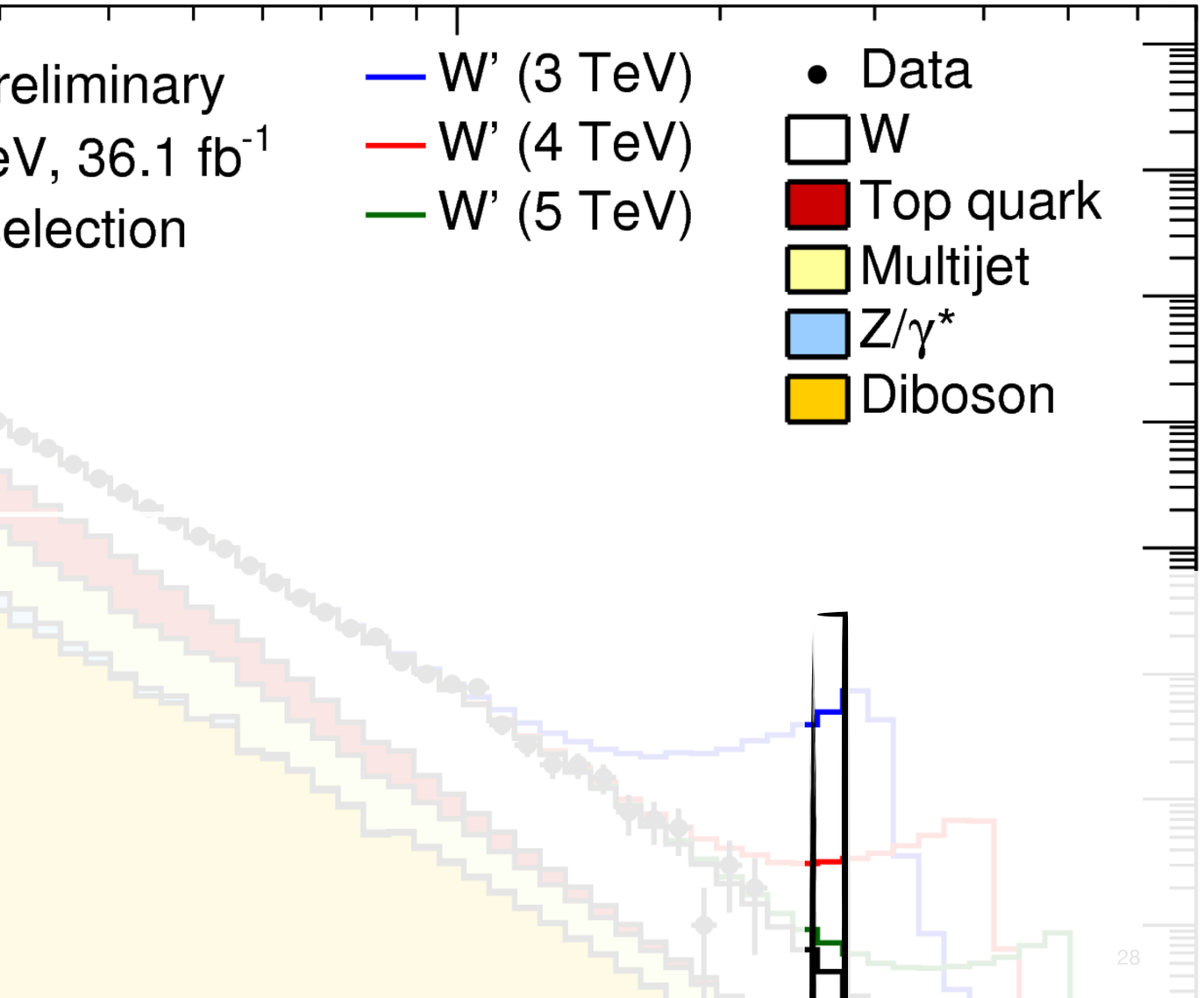
Getting expected rates $\lambda(\theta)$ O(10k) CPUh!

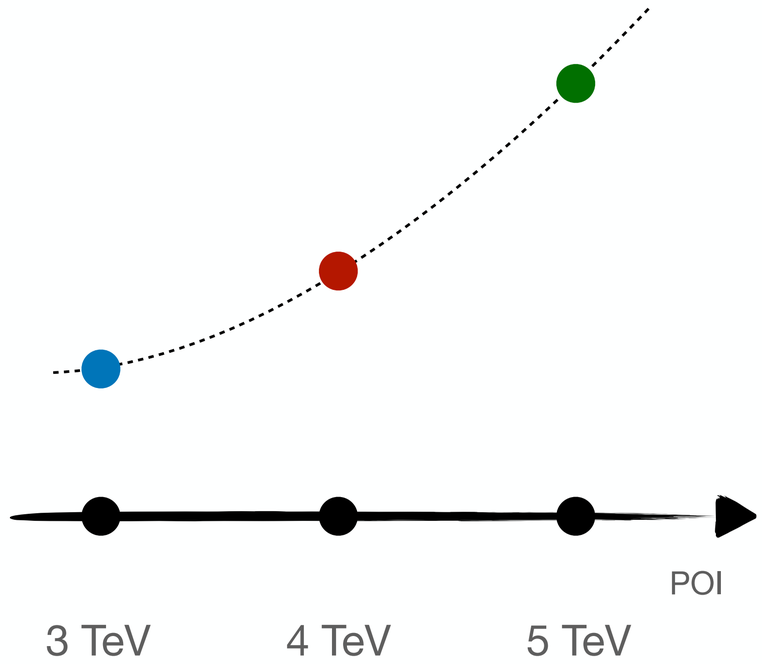
$$\mathrm{Pois}(n_i \mid \lambda_i(\theta))$$

Cannot be used to during inference (e.g. MLE fits). Need a faster option.

Interpolation

Simulators are very good, but very expensive as well





Incorporating Prior Information

As HEP analysis is so dominated by NPs, we must incorporate prior info.

- **But methodology designed to allow Bayesian OR Frequentist treatment**

where does this prior come from?

$$p(x \mid \mu, \nu) p(\nu)$$

Incorporating Prior Information

As HEP analysis is so dominated by NPs, we must incorporate prior info.

- **But methodology designed to allow Bayesian OR Frequentist treatment**

where does this prior come from?

probably from a separate measurement a

$$p(x \mid \mu, \nu) p(\nu \mid a)$$

Incorporating Prior Information

As HEP analysis is so dominated by NPs, we must incorporate prior info.

- **But methodology designed to allow Bayesian OR Frequentist treatment**

where does this prior come from?

probably from a separate measurement of data a

$$p(x \mid \mu, \nu) p(\nu \mid a)$$

via Bayes'

$$p(\nu \mid a) \sim p(a \mid \nu) p_{\mathrm{ur}}(\nu)$$

"subsidiary measurement"

Incorporating Prior Information

Can incorporate same prior information by collecting "subsidiary measurements" $p(a \mid \nu)$

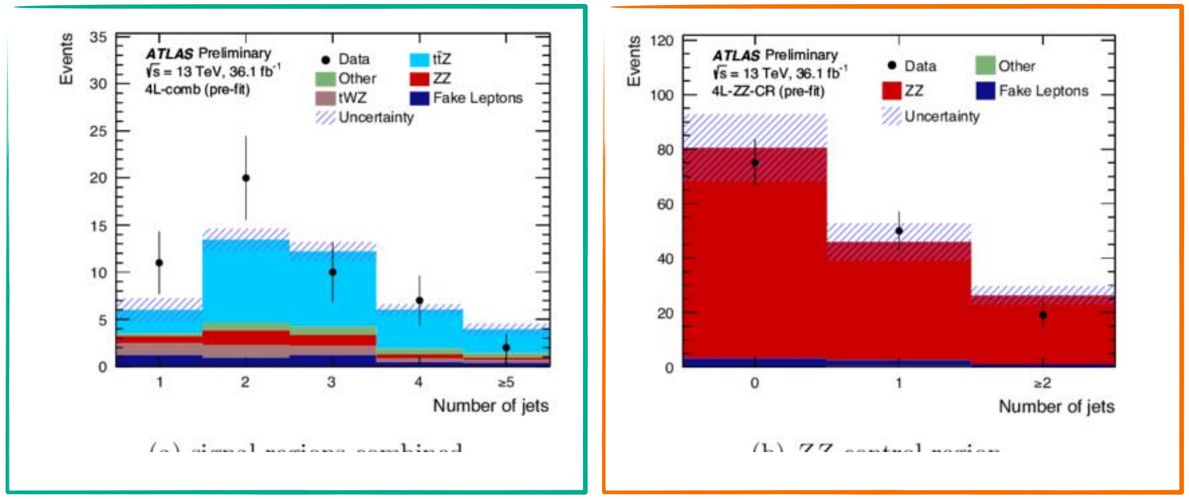
$$p(x \mid \theta) \quad \longrightarrow \quad p(x, a \mid \theta) = p(x \mid \mu, \nu) p(a \mid \nu)$$

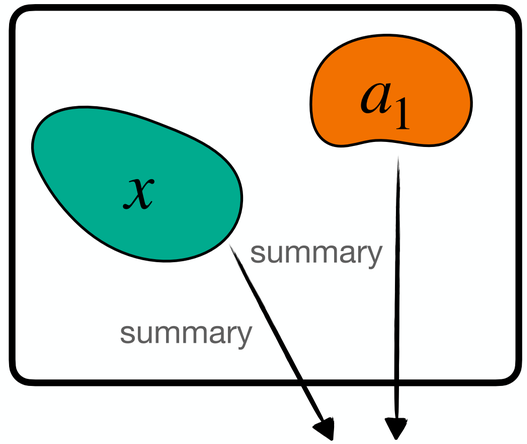
Next Question: what are these subsidiary likelihoods?

Option 1: “The Gold Standard”

Perform a full separate measurement (i.e. derive a full fledged model)

Example: constaining background by measuring a “control region”





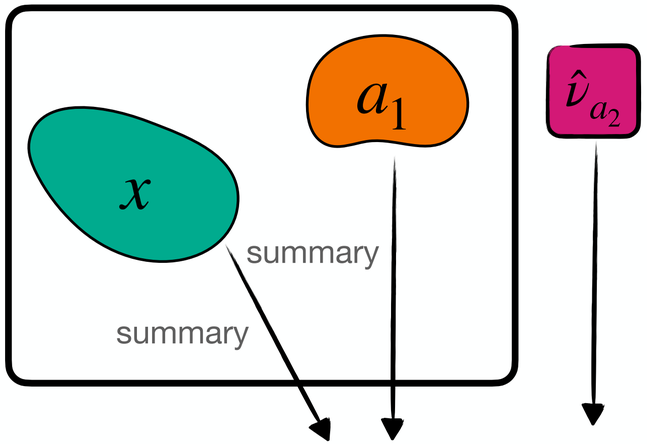
$$p(x, a \mid \theta) = p(x \mid \mu, \nu) p(a_1 \mid \nu)$$

A lot of work to get a full $p(a \mid \nu)$

Option 2: “Simplified Subsidiaries”

If no $p(a \mid \nu)$ is not available, or too complex, approximate contribution via

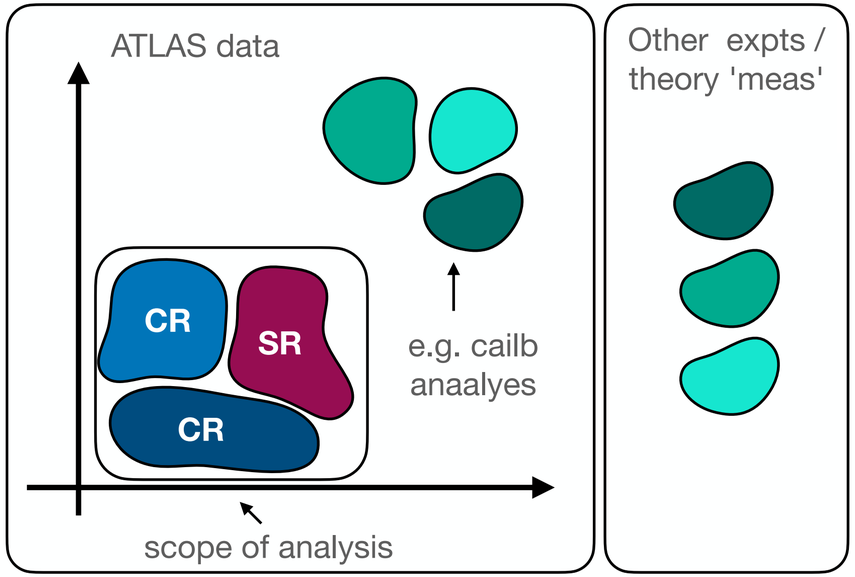
$$p(a \mid \nu) \sim p(\hat{\nu}(a) \mid \nu) \sim \text{Gaus}(\hat{\nu}_a \mid \nu, \hat{\sigma})$$



$$p(x, a \mid \theta) = p(x \mid \mu, \nu) p(a_1 \mid \nu) \text{G}(\hat{\nu}_{a_2} \mid \nu)$$

Putting it all together:

Realistic Analyses use a mixture. A lot of analyses covered by this approach.



$$f(\boldsymbol{n}, \boldsymbol{a} \mid \boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb} \mid \nu_{cb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{Simultaneous measurement of multiple channels}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_{\chi}(a_{\chi} \mid \chi)}_{\text{constraint terms for “auxiliary measurements”}},$$

nominal rates

$$\overline{\nu_{cb}\left(\boldsymbol{\phi}\right)} = \sum_{s \in \text{samples}} \nu_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) = \sum_{s \in \text{samples}} \underbrace{\left(\prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{multiplicative modifiers}} \underbrace{\left(\nu^{0}_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) + \sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{additive modifiers}}$$

sum over components in mixture mode

“parametrized interpolation"

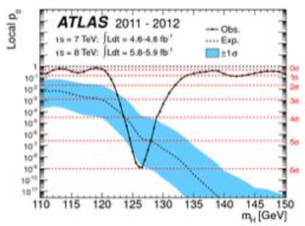
Public Tools available:

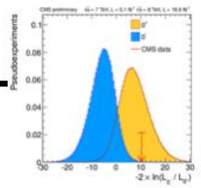
pyhf
∂ifferentiable
ℒikelihoods

github.com/scikit-hep/pyhf

[LH, M. Feickert, G. Stark]

Statistical Inference





$s(\cdot)$

data

simulation

"Likelihood-ful"
Inference

test statistics
for inference

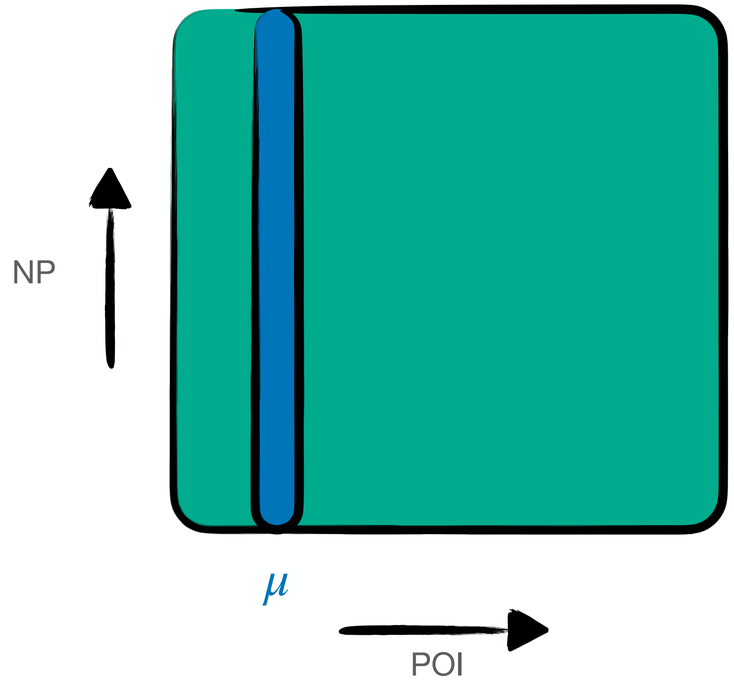
summary statistics /
density estimation

raw data

Nested Models

Split out parameter space into hypothesis spaces.

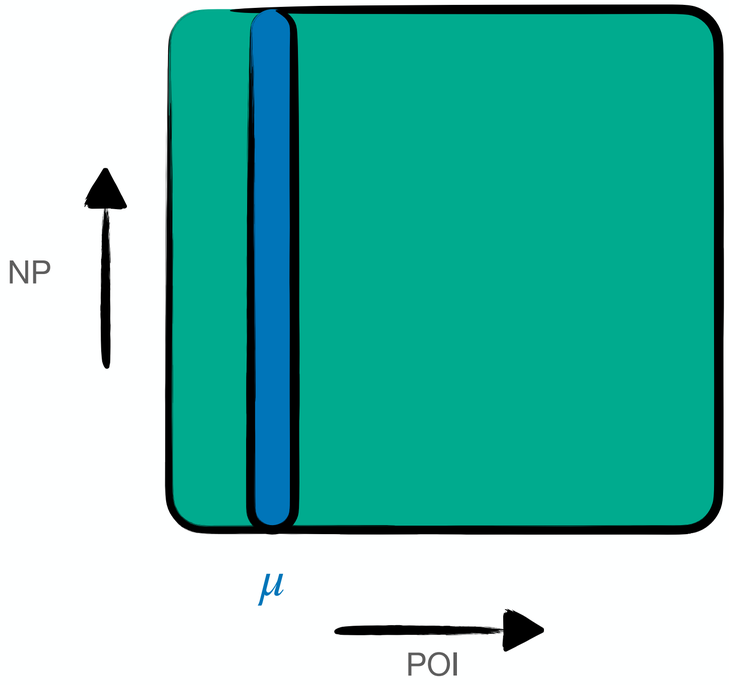
- **Reminder: not interested in inference on NPs**
- **Null hypotheses only indexed by POIs**

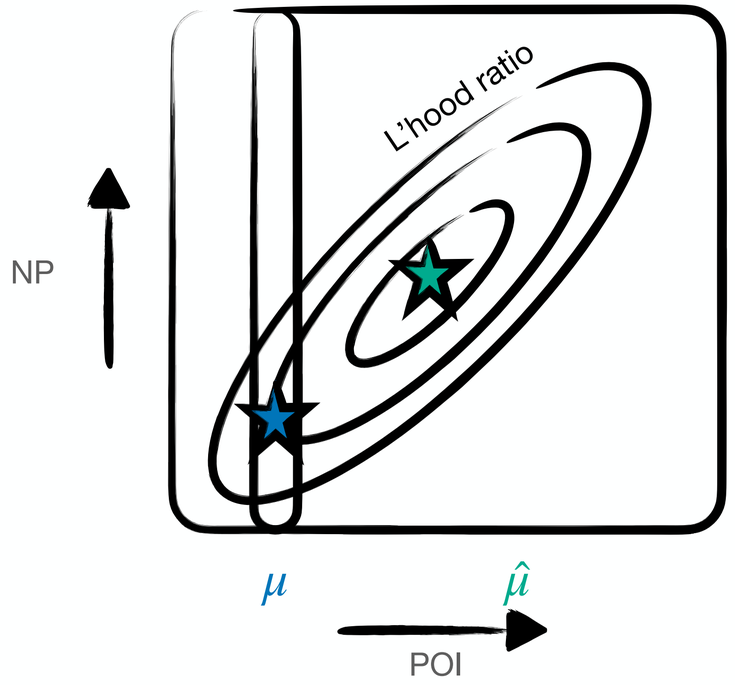


Nested Models

After decades of iteration, HEP landed on classic Likelihood Ratio Test

- **"profile likelihood ratio"**





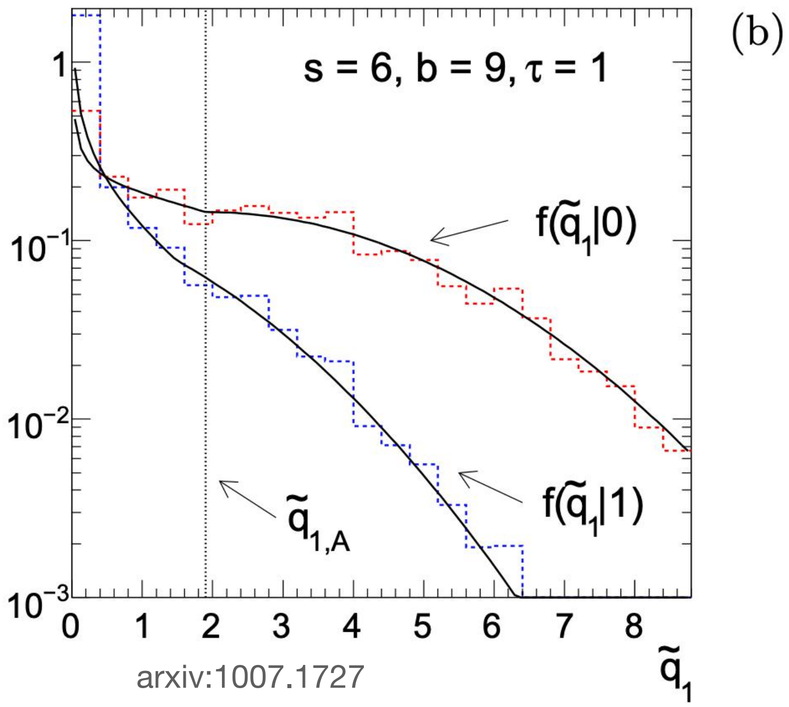
More Stats <> Physics connections would have helped as a lot!

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

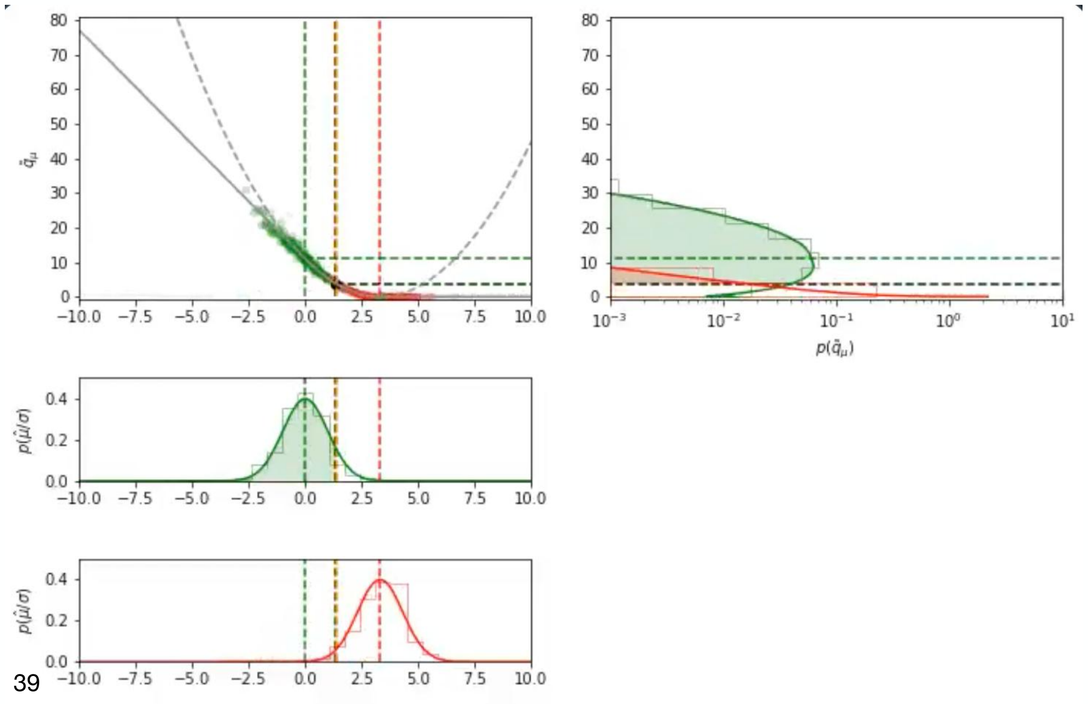
Use "best-in-class" parameter for each hypothesis set for ratio

Asymptotic Calculations

- Many analyses rely heavily on asymptotic shapes of test stat distributions
- Assumption: Wilks & Wald hold (cf. see next slide)



(b)



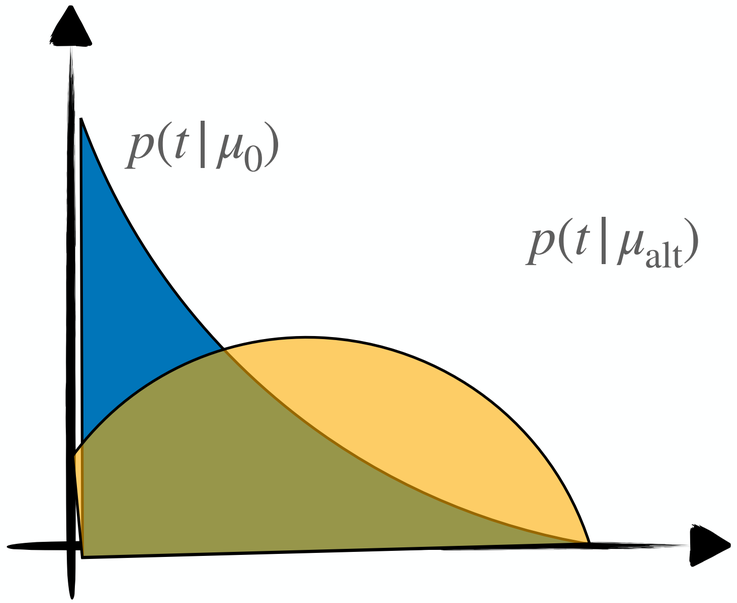
Interval construction

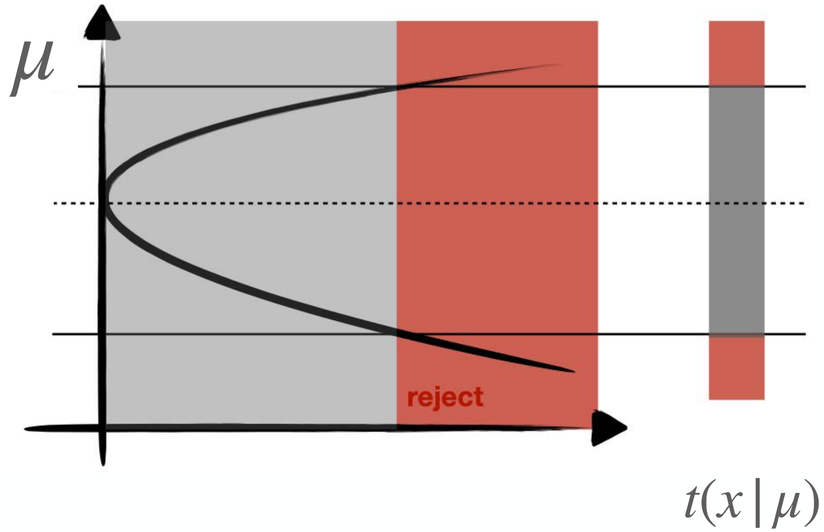
HEP fairly principled: Intervals strictly via exact Neyman-Construction.

Widely assumed, rarely checked $p(t \mid \mu, \nu) = p(t \mid \mu)$

If holds, can do Intervals only POI-space without loss of coverage

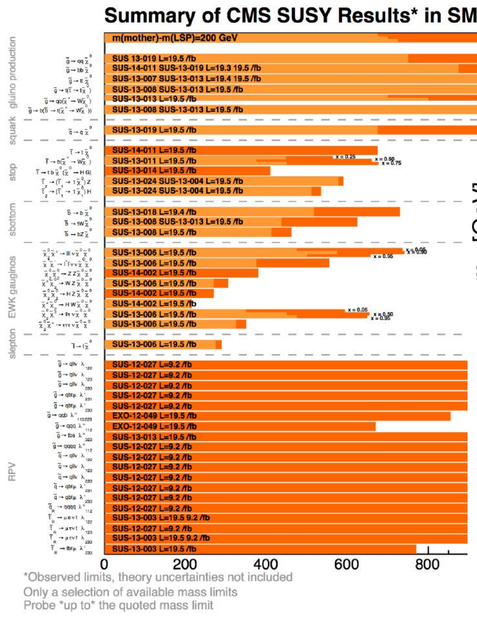
- **removed NP from quoted results**
- **saved a lot of compute O(1) << O(1000)**

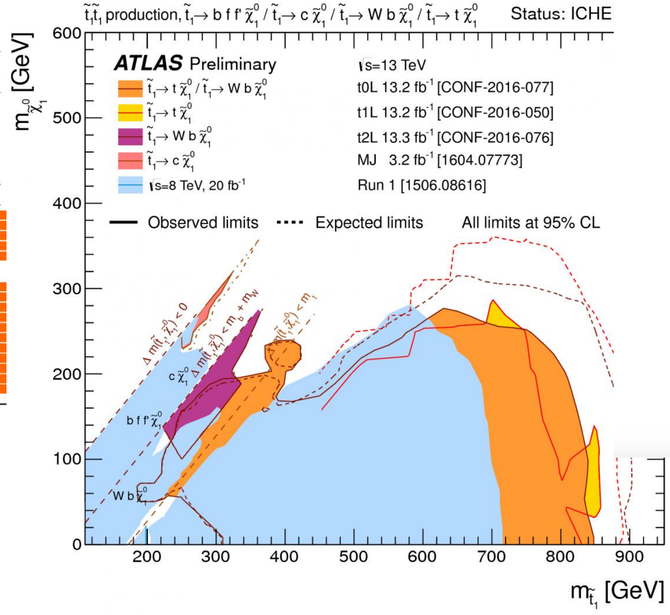


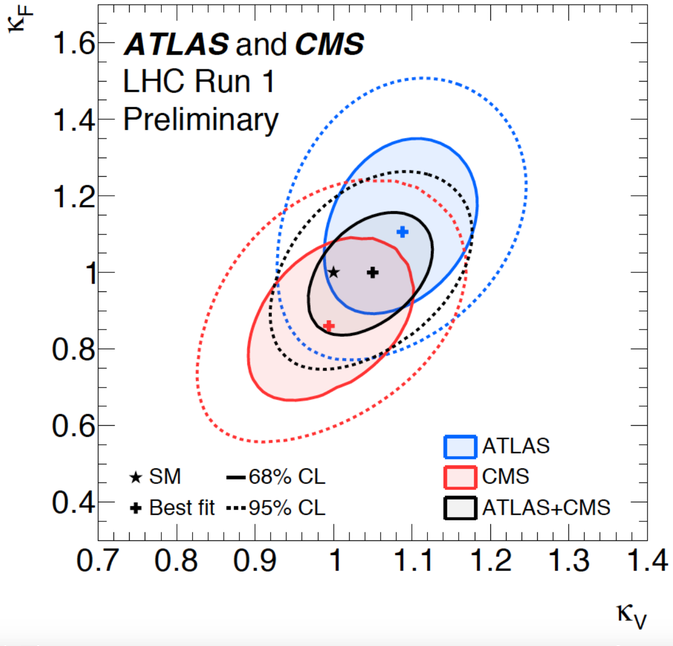


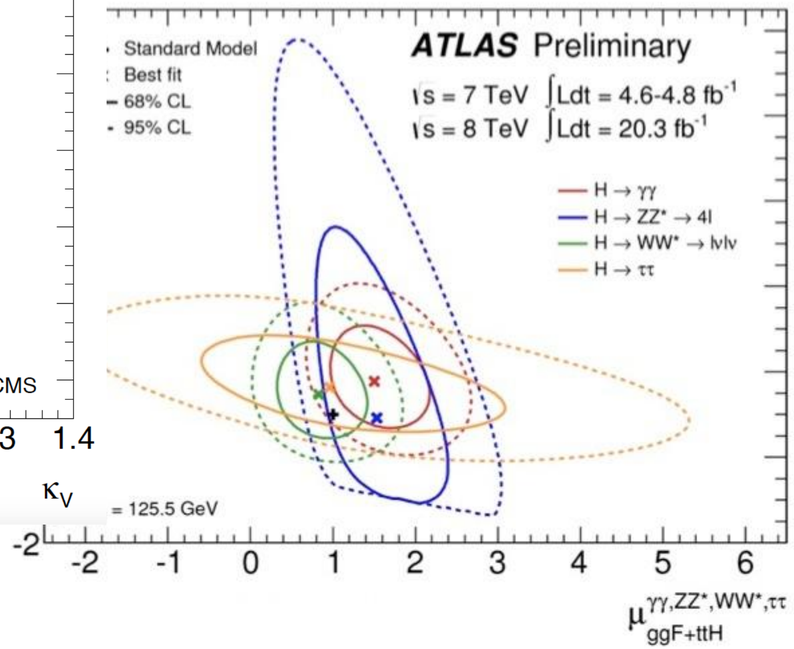
Looking Back and Forward

The methodology of simulation-driven frequentist inference is the result of many decades of development. Supported us for Higgs discovery and beyon



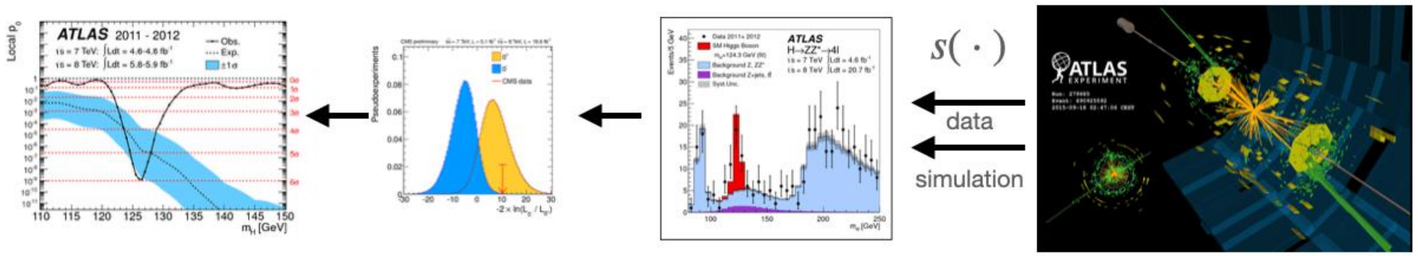






Are we done with Stats methodology?

Going beyond



"Likelihood-ful"
Inference

test statistics for inference

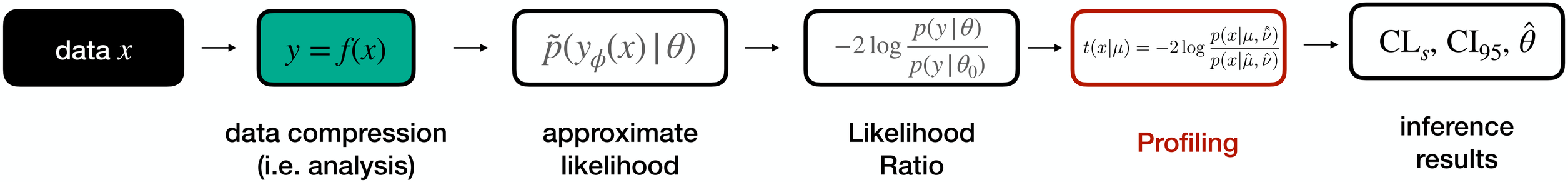
summary statistics / density estimation

raw data

Back to basics:

In frequentist analysis we do not need the likelihood

- **we *choose* to run likelihood-free inference with an approximate l'hood**



But ultimately all these intermediate step are optional

Recently: A lot of development in exploring ML-driven "direct" likelihood-free inference

What's the connection?

Key Stat. Concept: Likelihood-Ratios

Why? Among all functions of the data *they are optimal* when used as a test statistic

(Neyman-Pearson)

Machine Learning:

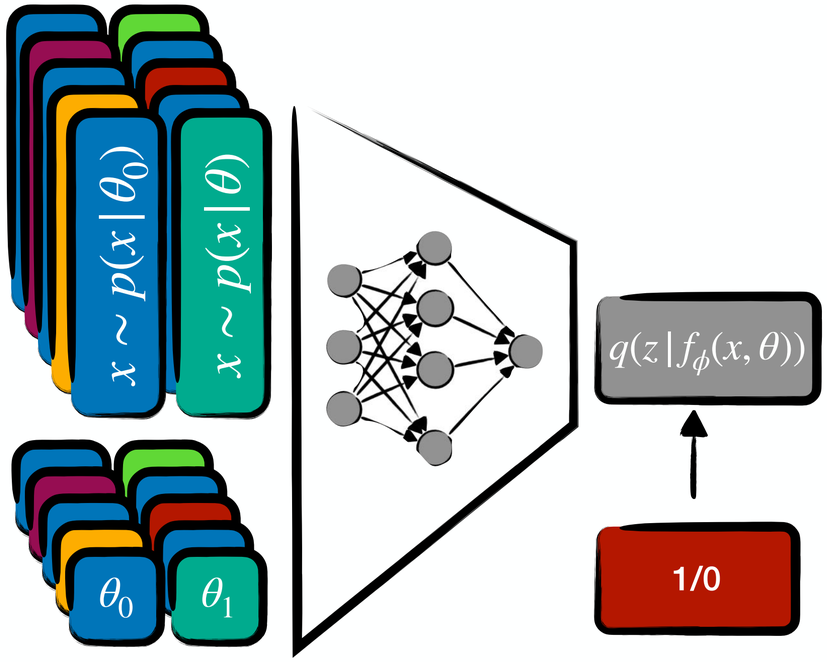
extremely good at searching through functions that are optimal in one sense of another

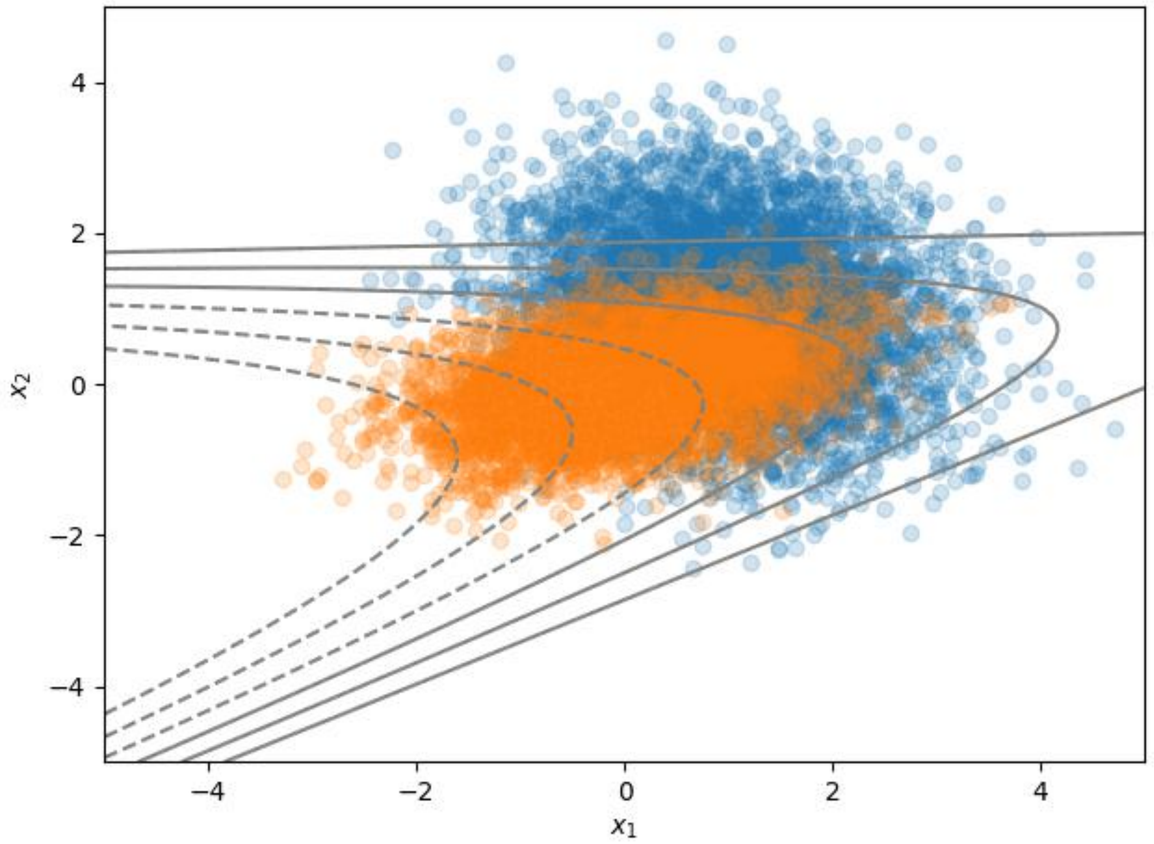
(based on samples)

seems like one should be able to use ML to do frequentist statistics

Likelihood Ratio Trick

Training neural networks for classification produces likelihood ratios, purely from samples without ever having to specify $p(x \mid \theta)$

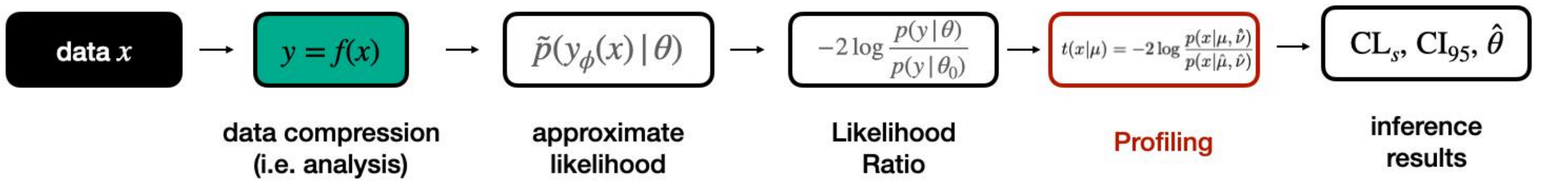


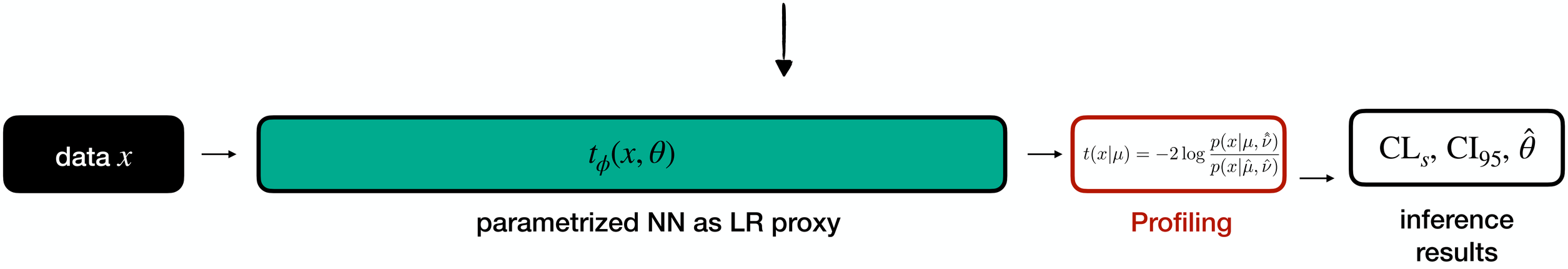


$$t_\phi(x, \theta) \to \frac{p(x \mid \theta)}{p(x \mid \theta_0)}$$

Back to basics:

We can start to replace a lot of our pipeline with ML components that directly output likelihood ratios

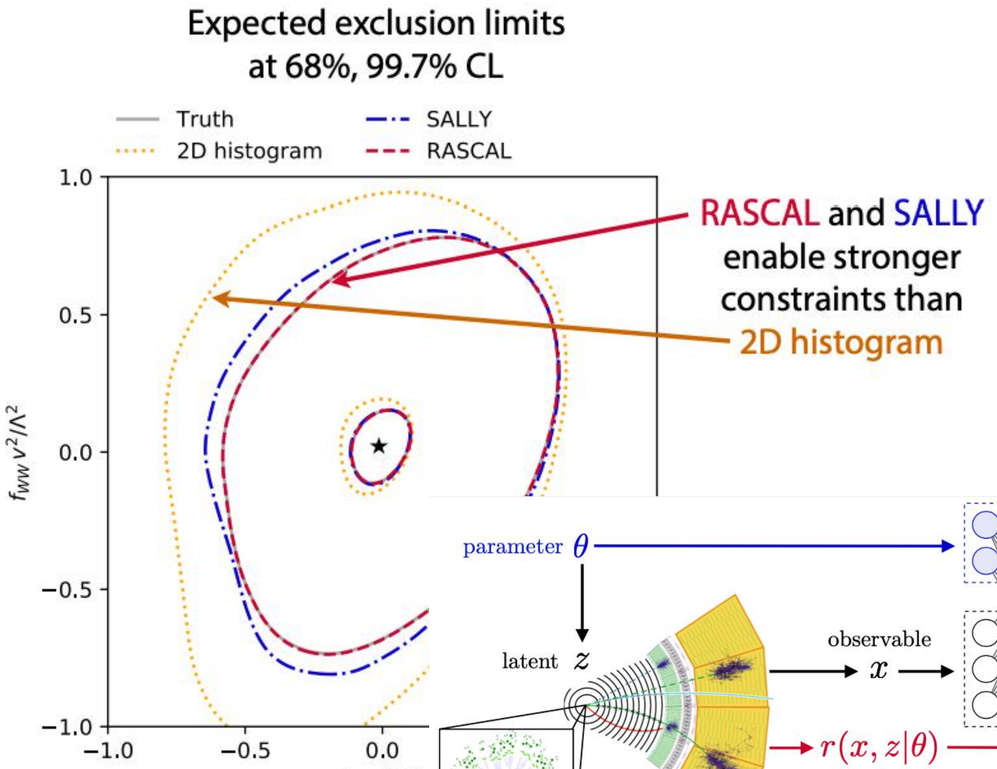


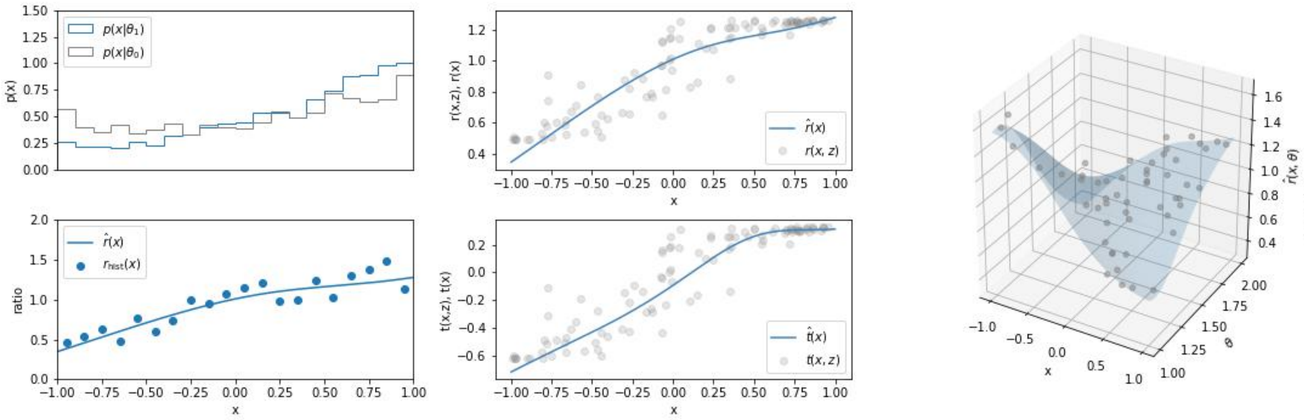


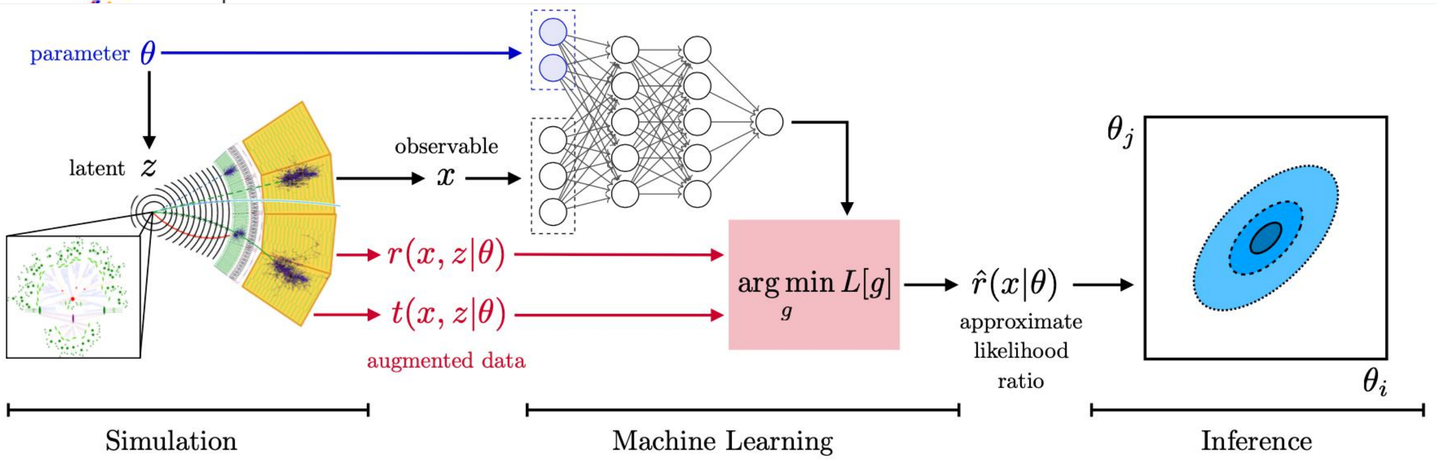
Key: this also removes the (lossy) compression we did to reduce dimensionality → increased sensitivity

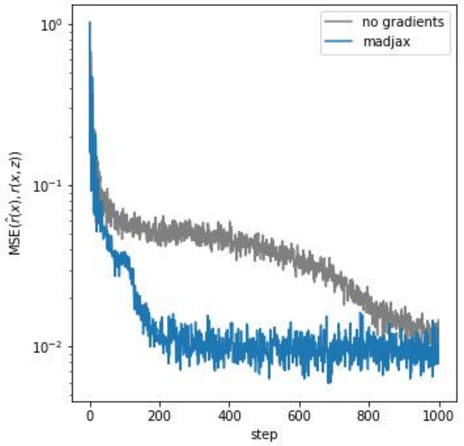
Back to basics:

Completely new workflow, but with promise: New generation of tooling to train these “likelihood ratio” networks









[Brehmer et al]

Going even Further

Now that we've stared? Just how much can we replace with ML?

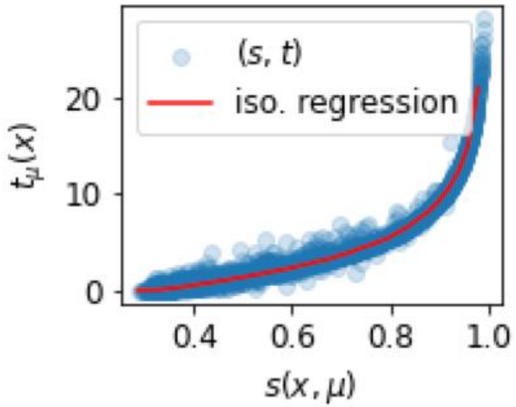
Profile Likelihood is <u>also</u> just used because it is optimal in a certain sense

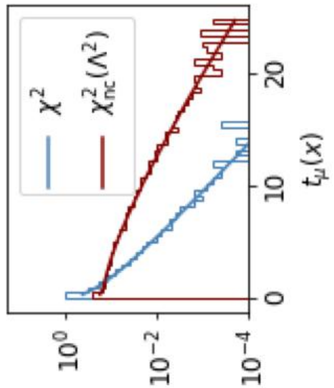
Can we learn a network that directly outputs profile likelihood?

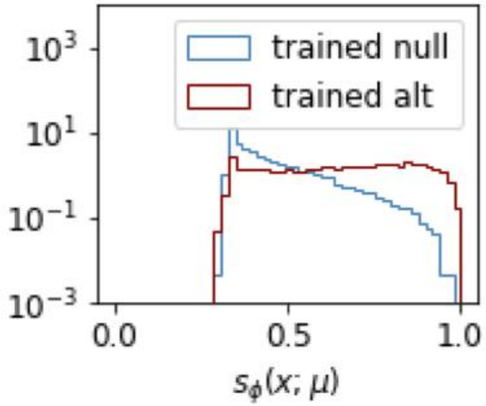
$$t(x|\mu) = -2 \log \frac{p(x|\mu, \hat{\hat{\nu}})}{p(x|\hat{\mu}, \hat{\nu})}$$

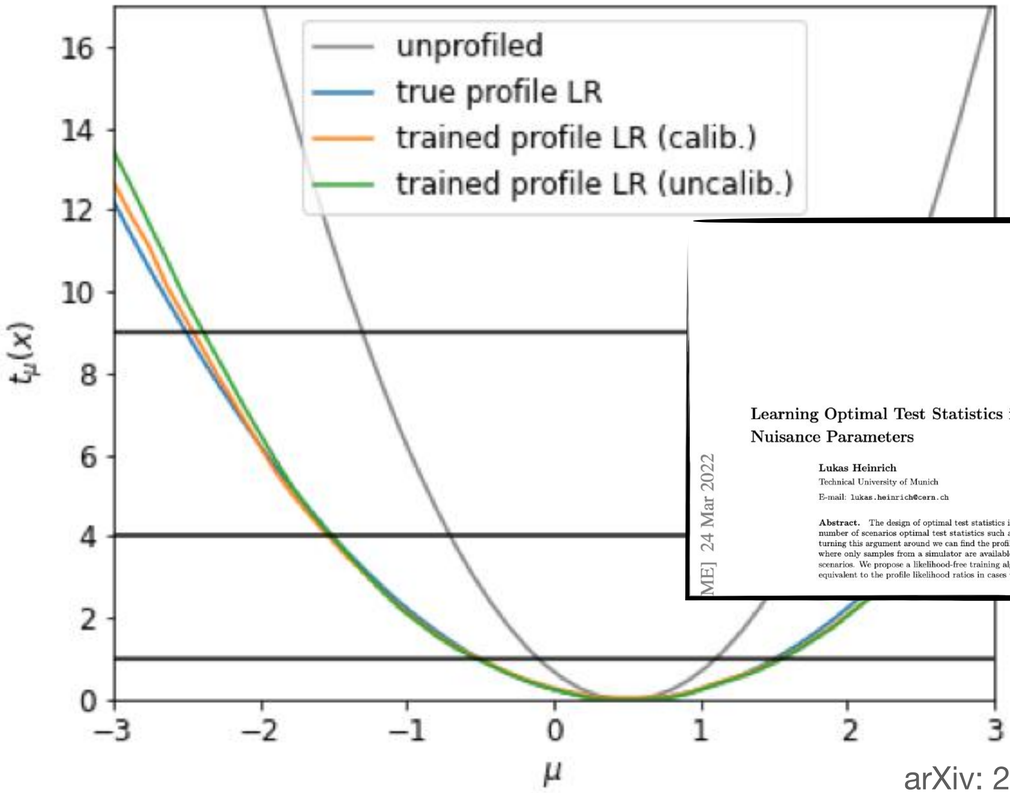
Going even Further

Yes, if the profile likelihood is optimal, we can find it through optimization









Learning Optimal Test Statistics in the Presence of Nuisance Parameters

Lukas Heinrich
Technical University of Munich
E-mail: lukas.heinrich@cern.ch

Abstract. The design of optimal test statistics is a key task in frequentist statistics and for a number of scenarios optimal test statistics such as the profile-likelihood ratio are known. By turning this argument around we can find the profile likelihood ratio even in likelihood-free cases, where only samples from a simulator are available, by optimizing a test statistic within those scenarios. We propose a likelihood-free training algorithm that produces test statistics that are equivalent to the profile likelihood ratios in cases where the latter is known to be optimal.

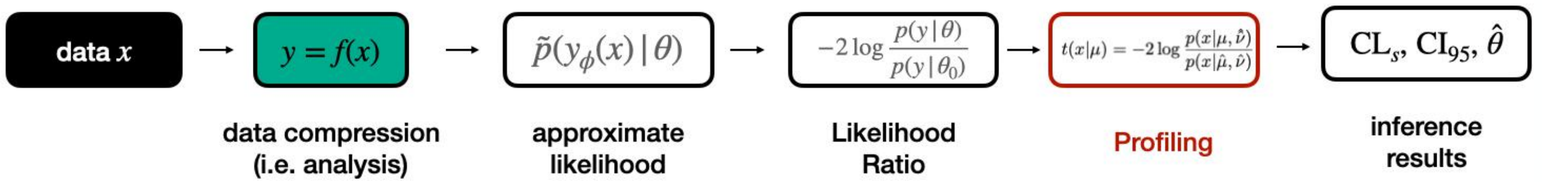
ME] 24 Mar 2022

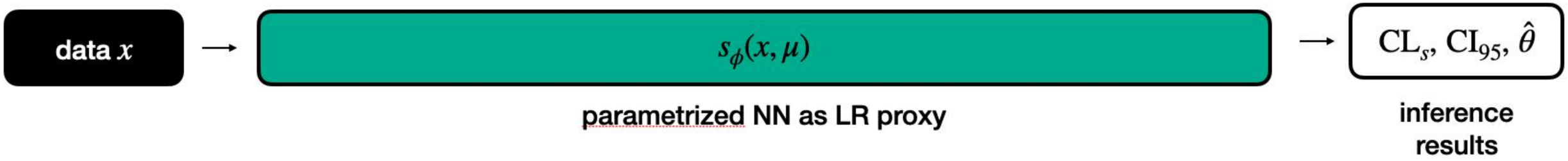
arXiv: 2203.13079

Active area of research: happy to chat more offline

Going even Further

It seems within reach, that one could target even more of the classical frequentist statistics pipeline with an end-to-end ML workflow





Summary

LHC Statistics is a poster child of frequentist analysis in a complex setting:

- likelihood-free (and even then expensive sampling)
- high-dimensional data and parameter spaces

Many decades of development culminated in a well-oiled machine with fairly principled approach to inference.

Classic Approach: proxy likelihood and focus on good modelling with incorporation of prior information for hundreds of nuisance parameters

Emerging: Machine Learning driving a resurgence in likelihood-free techniques that aim to eliminate some of the compromises we usually make

Backup

Summary: Many things work but also lots of interesting questions

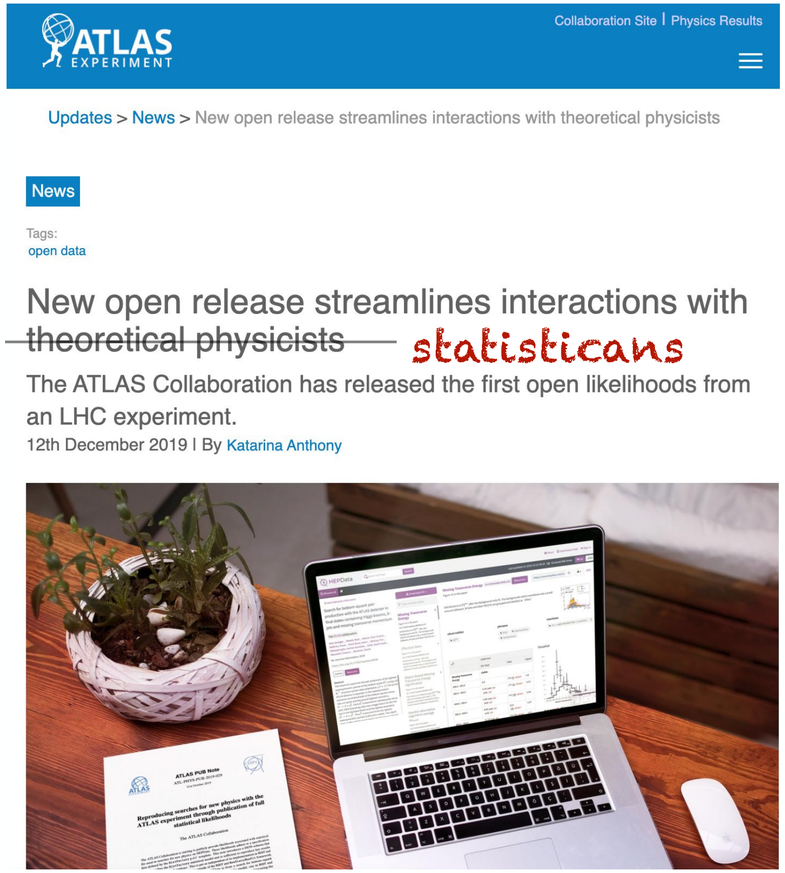
Exciting development: public real-world probability models

Not only opportunity to push new science but also new stats methodology / answer "what if"s

A new public workhorse / platform akin to e.g. on-off problem, $\mathrm{Pois}(n \mid \mu s + b)$, ...

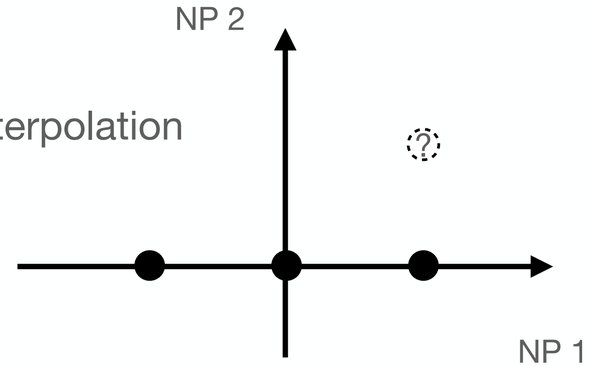
Examples:

- **Bayesian Workflow on LHC models**
- **Coverage & Asymptotics Studies**
- **Signal interpolation strategies**
- **Reinterpretation Tools**
- **Model destillation ("simplified likelihoods")**

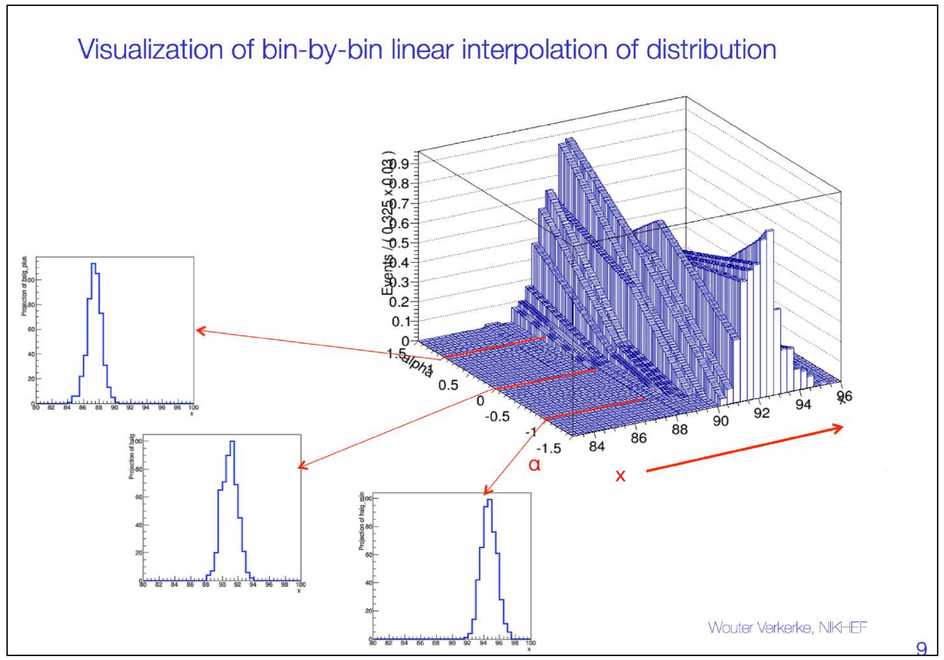


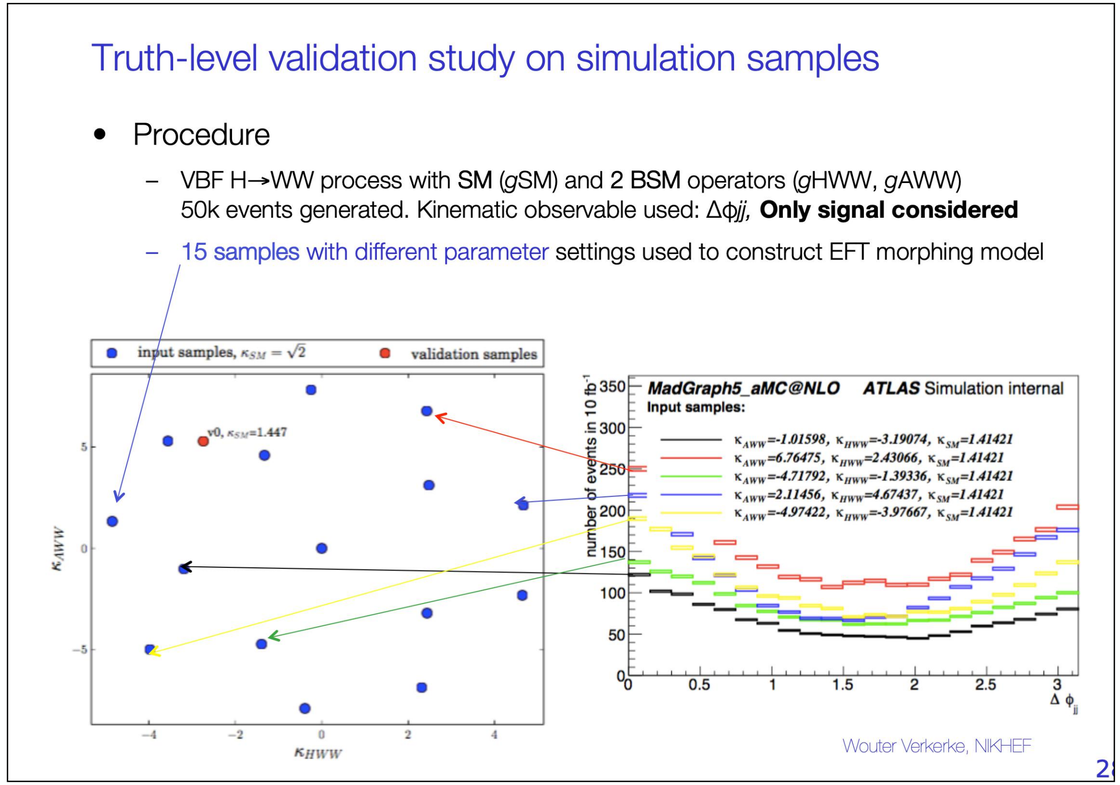
Example Interpolations

Widely used default:
“on-axis” Vertical Interpolation



More Complex Interpolation Schemes
with “off-axis” input distributions





Two-dimensional interpolation [W. Verkerke]

A problem for non-asymptotic cases

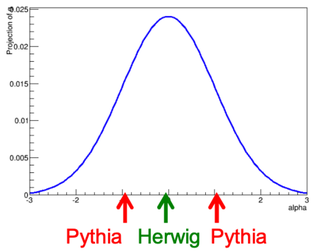
Sometimes, NPs relate to (categorial) choice of simulator itself

- lots of discussions on appropriate “simplified subsidiary”

Specific issues with theory uncertainties

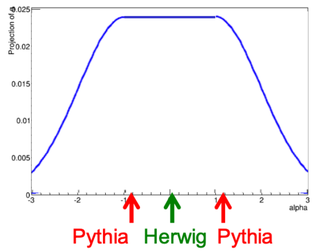
- Subsidiary measurement of a theoretical 2-point uncertainty effectively quantifies the ‘knowledge’ on these models
  - *Extra difficult to make meaningful statement about this*, since meaning of parameter is not well embedded in underlying theory model
  - But again, all procedures need to assume some distribution… Profiling requires you to spell it out
- Some options and their effects

Gaussian



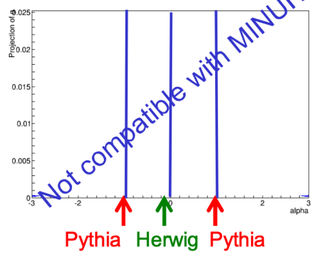
Prefers Herwig at 1σ

Box with Gaussian wings



All predictions ‘between’ Herwig and Pythia equally probable

Delta fuctions



Only ‘pure’ Herwig and Pythia exist

Wouter Verkerke, NIKHEF

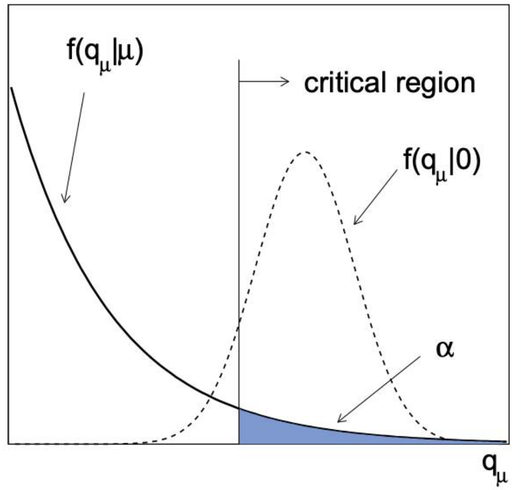
Decision Theory

Key point in inference (intervals, tests, …): reject / accept.

Here HEP differs from orthodox statistics.

Incorporate power of alternative into decision function of rejecting null → no power: never reject

Reason: huge consequence of rejecting null (only reject if good alternative available)

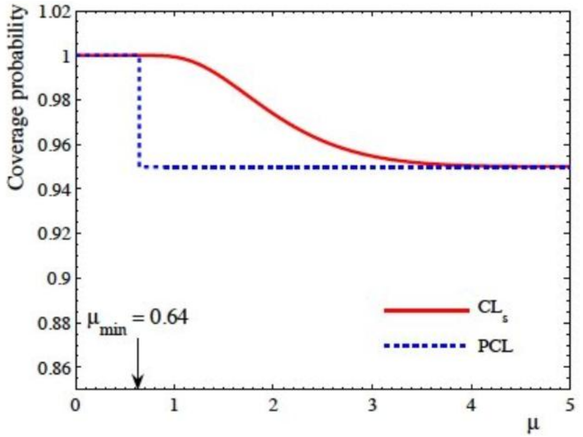


HEP Jargon: "CLs" method

$$\mathrm{CL_s} = p/\beta$$

p-value

power at p-value



Boundary Effects

Common occurence in HEP:

- **null is on boundary of parameter space: → adjust Wilks'**
- **extended region in alternative that's indistinguishable from null → input from statisticians interesting**