Open Questions in Statistical Practice for Particle Physics

Francisco Matorras
Instituto de Física de Cantabria
Santander, Spain

Introduction

- Statistics in HEP a rich (and often non-trivial) topic
  - Lot of new techniques being explored
  - Better understanding of old techniques
- Will concentrate on few highlights
  - What's behind discovery statements
    - 5σ
    - *Local p-values, look elsewhere effect*
  - Data (re)interpretation
    - How to use published data for your interpretation
  - Machine learning vs statistics
- More on parallel session H, have a look to many interesting results and applications

Why?

Why statistics at a particle physics conference?

- Statistics is at the core of particle physics since quite many years
- Deal with data (often huge amounts) to produce:
  - observables (cross sections, masses, …) and uncertainties with proper statistical interpretation (does $m \pm \Delta m$ properly represent our 68% confidence interval?)
  - to set limits on our NP models (what a 90% exclusion is)
  - Or to establish discoveries (5σ !!!)
- More and more powerful statistical techniques used
- We want to get the best out of our data:
  - Can imply saving a lot of money (shorter running times)
  - Can imply reaching further away physics
  - Will help us to avoid embarrassing announcements

Statisticians & physicists

Trying to speak a common language

- Often physicists tend to reinvent existing methods
- And are not aware of recent (or not so recent) useful ones
- An effort ongoing of joining physicists and statisticians
- PHYSTAT founded in 2000 by L. Lyons
- many seminars, conferences and workshops to debate on relevant issues
- You are invited to attend and explore here or here (legacy page)

Many topics

- PHYSTAT-2sample: for 2 sample and GOF tests, 1-2 June 2023
- BIRS workshop (23w5096) (Banff) *"Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses"*, 23-28
- PHYSTAT- Gamma 2022: High Energy Gamma Ray Astronomy in a Multi-Wavelength Context, 27-30 Sep 2022
- PHYSTAT-Anomalies 2022: Model-independent searches for New Physics, 24th and 25th May 2022
- PHYSTAT-Systematics workshop 2021 1-3 Nov + 10 Nov 2021
- PHYSTAT-FLAVOUR 2020 virtual workshop 19-21 Oct 2020
- PHYSTAT-DM 2019 (Stockholm University) Jul 31 - Aug 2, 2019 "Statistical Issues in direct-detection Dark Matter search exp
- PHYSTAT-nu 2019 (CERN) Jan 22-25
- PHYSTAT-nu 2016 (FNAL) Workshop on Statistical Issues in Experimental Neutrino Physics September 19-21, 2016 Fermilab
- PHYSTAT-nu 2016 (Kavli, Japan)
- PHYSTAT 2011 (CERN) Proceedings *"Statistical issues related to discovery claims in search experiments, concentrating on those workshop"*
- BIRS workshop (10w5068) (Banff) "Statistical issues relevant to significance of discovery claims ", **11-16 Jul 2010**
- PHYSTAT 2007 (CERN) Link to proceedings *"Statistical issues for LHC physics."*
- BIRS workshop (06w5054) (Banff) "Statistical inference Problems in High Energy Physics and Astronmomy", **15-20 Jul 2006**

September 2024

- 09 Sept - 12 Sept PHYSTAT - Statistics meets ML

May 2024

- 15 May - 17 May PHYSTAT-SBI 2024 - Simulation Based Inference in Fundamental Physics
- 08 May Alexander Lincoln Read, Michael Evans, Tom Junk, "PHYSTAT informal review: CLs criterion for limit setting"

April 2024

- 10 Apr Hans Peter Dembinski, "Template fits: fitting non-parametric density models to data"

March 2024

- 13 Mar Charles Geyer, "PHYSTAT Seminar: An Introduction to the Nonparametric Bootstrap"

February 2024

- 28 Feb Alessandra Brazzale, Roger Barlow, "PHYSTAT informal review: Asymmetric Uncertainties"

January 2024

- 24 Jan Larry Wasserman, Robert Cousins Jr, "PHYSTAT informal review: Hybrid Bayesian-Frequentist approaches"
- 10 Jan Alan Heavens, "PHYSTAT Seminar: Extreme Lossless Data Compression for Likelihood-Free Inference"

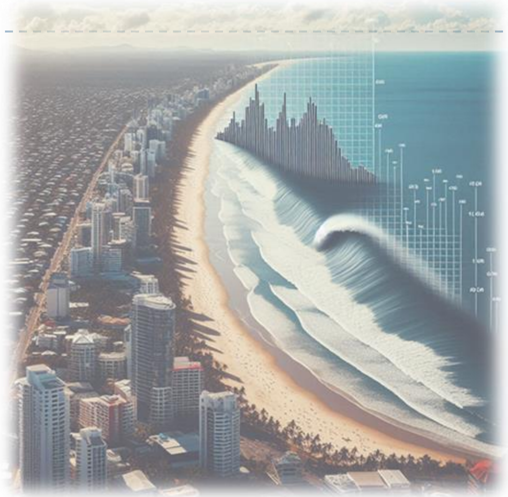
October 2023

- 25 Oct Lydia Brenner, "PHYSTAT Seminar: Comparison of Unfolding methods"

Discovery

A discovery from the point of view of statistics

Hypothesis testing in a nutshell



- What we usually do, *hypothesis test in stat words,* is confronting our data to a model against an alternative, H0 vs H1
  - H0, *null hypothesis*, is the model we want to **negate**, SM, SM without a given process, background only
  - H1 is the *alternative hypothesis*, our model with "new" physics, often depending on a parameter
  - We also might want to reverse the logic and use as H0 the NP model, to set limits
- Often based on likelihood ratio, built a test statistics that is the ratio of our best fit to H0 and best fit to H1 $q = -2\log(\mathcal{L}(H_0)/\mathcal{L}(H_1))$ where the likelihoods are the best fit to each hypothesis
  - If data produces a small q_0 it means it prefers H0, a large q rejects H0
  - We measure how big or small is q with the p-value:
    - If H0 is true, what is the probability that a fluctuation gives $q > q_0$?
    - And usually translate it to a significance, *(or z-score):* equivalent number of *gaussian* σ

5σ

When to claim a discovery? Aka 5σ

- It is well known that in HEP we have a (historical) convention that we cannot claim a discovery unless we have an excess of at least 5σ
- But are we aware what it means? Does it make sense?
- Some comments here, largely based on different pubs by Louis Lyons, nicely summarized in a recent CERN Courier article

Are 5σ enough? too much?

- 5σ is equivalent to p-value of **about $3 \cdot 10^{-7}$**
  - the probability of observing such an extreme event from a background fluctuation is smaller than $3 \cdot 10^{-7}$
- Note that statisticians usually consider 3σ ($p = 0.001$) enough to prevent fluctuations
- Is it worth struggling to make 5σ out of 4.9σ?
  - To me it is a bit naïve when we cannot control our systematics and asymptotic approximations to that level
  - Higgs discovery, Atlas and CMS

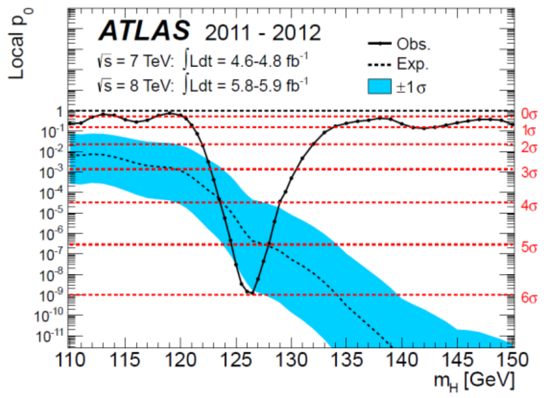


Figure 9: The observed (solid) local p_0 as a function of m_H in the low mass range. The dashed curve shows the expected local p_0 under the hypothesis of a SM Higgs boson signal at that mass with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances of 1 to 6 σ.

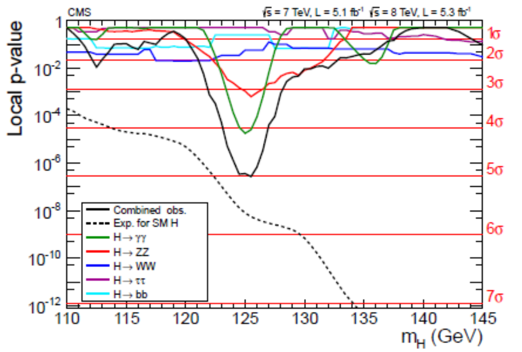


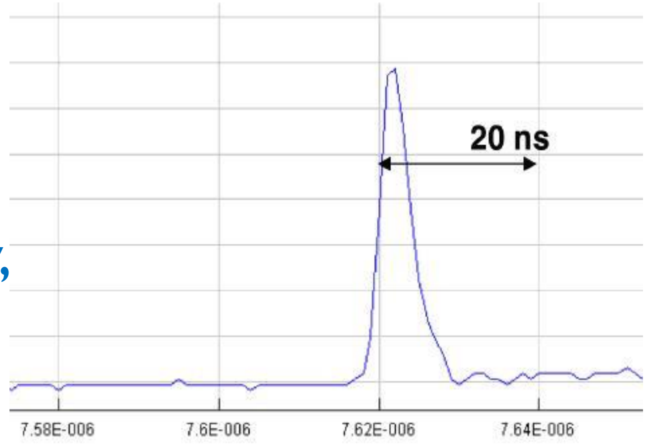
Figure 15: The observed local p-value for the five decay modes and the overall combination as a function of the SM Higgs boson mass. The dashed line shows the expected local p-values for a SM Higgs boson with a mass m_H.

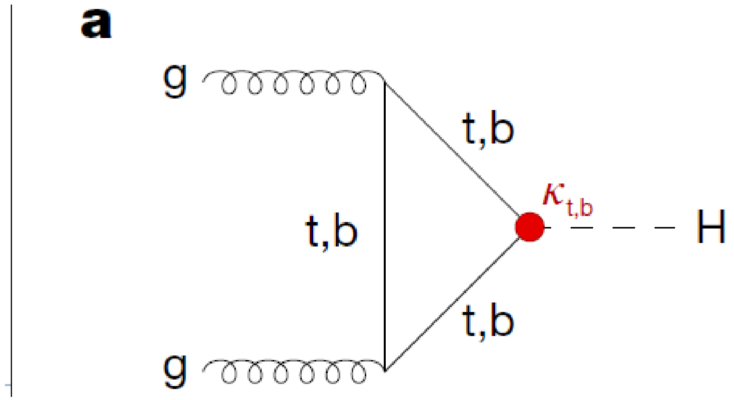
Why 5σ?

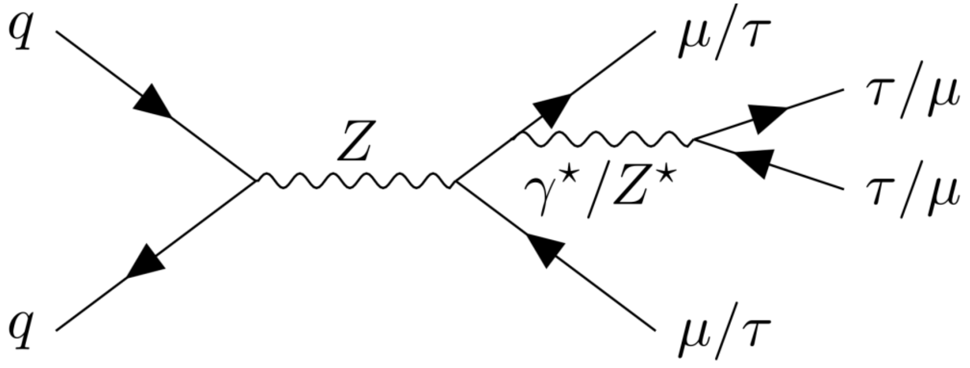
- Why? Historical reasons, probably:
  - To void embarrassing mistakes, apparently because at the time there were many 4σ observations that washed out with time
    - We still see 5σ *anomalies* disappearing!
  - Lack of confidence on systematic evaluation (we are above 3σ even if systematics doubled)
    - But what if your analysis is statistically dominated?
  - No chance of background fluctuation
    - Do we need to get to the $\mathbf{10^{-7}}$ level?
  - Safety margin? I might have missed some systematic uncertainty…
  - Look-elsewhere effect, probability to see a fluctuation as big anywhere in my spectra, in my analysis, in my experiment…
    - Will speak later, not all elsewheres are similar

Plausibility

- L. Lyons introduced the concept of "plausibility"
  - Should not use the same significance for Higgs discovery, leptoquarks, faster-than-light neutrinos, HH in pp, a given decay channel expected by SM, violation of lepton universality, or an anomaly not covered by any expected model
  - For some cases 3σ is enough, for others even 5σ not sufficient, suspicion beyond statistics…
- Should define a case-dependent threshold but not an easy task







Look elsewhere effect (LEE)

Look elsewhere

- What is the probability to get hit by a coconut falling in your hotel?

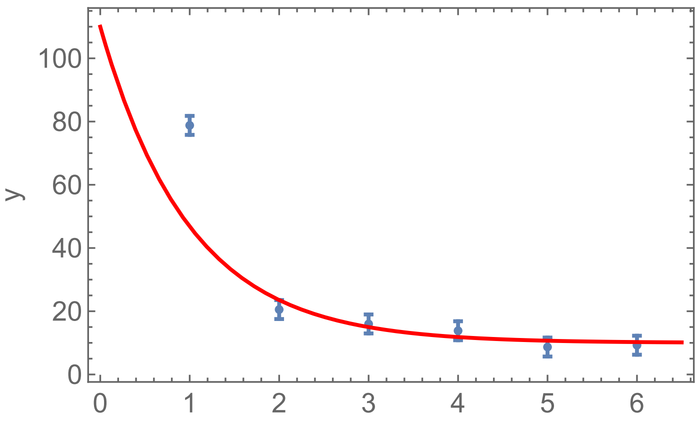
Apparently not that small if you account for many guests and many days

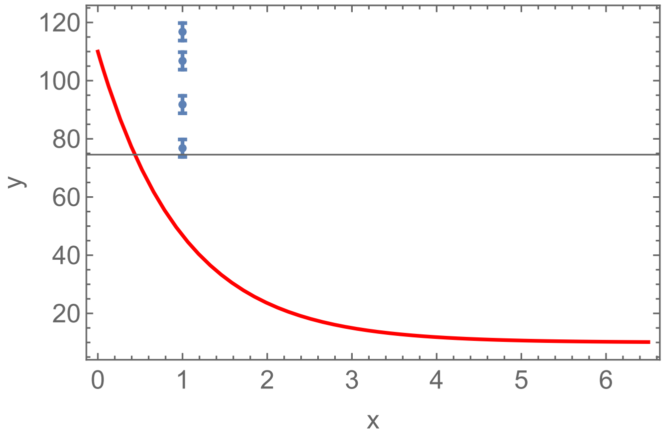
Look elsewhere effect

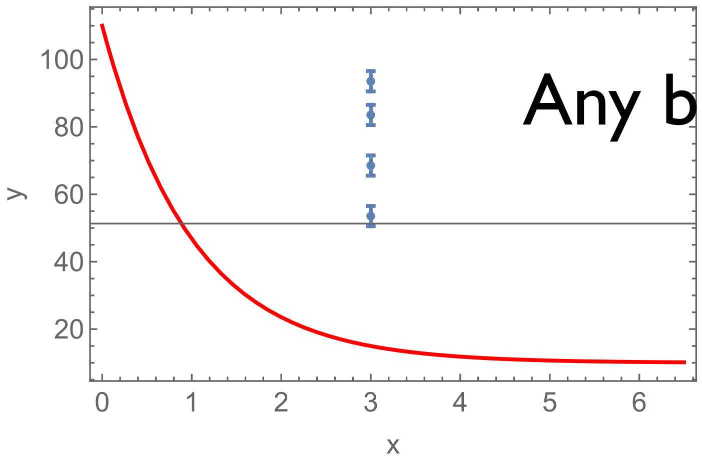
- A simple example: in one bin we see a big excess over background
  - We calculate the p-value, and it is very small.
- But one wonders:
  - I would have been equally surprised if the excess was in any other bin, what is the prob to see such a fluctuation in any bin? p increases..
- Maybe we have no information on the width, we would have been surprised by smaller fluctuations in two nearby bins, p grows more
- Why not smaller fluctuations in three consecutive bins? Or 4? Or…
- Even could accept cases where the excess oscillates, and look for excesses separated by one, two, three bins…
- We should account for all possibilities when calculating the p-value

Elsewhere?

- Depending on your model parameters, many elsewheres…

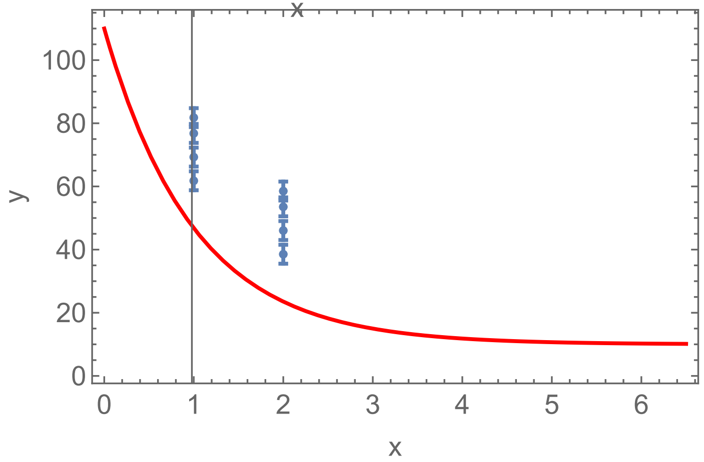


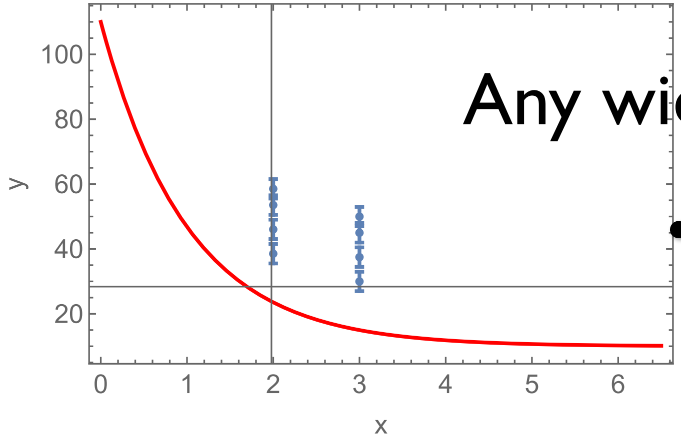




Any bin

• • •





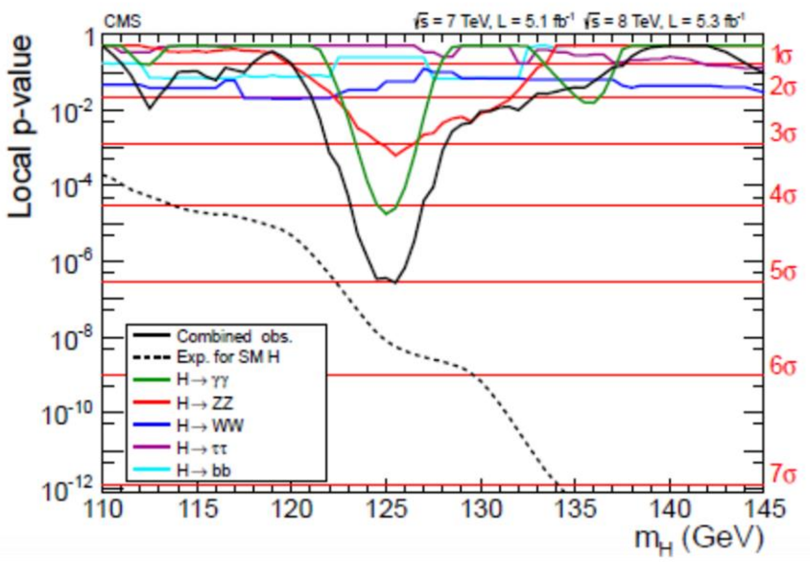
Any width

• • •

Going beyond local p-values

(ab)use of **local** p-values

- Usual procedure at LHC:
  - p-value calculated for a single free parameter (usually signal strength) as a function of other(fixed) parameters (ie mass and possibly width)
  - Quote as **local p-value** the smallest one in the scan
    - Note it is calculated assuming only one free parameter
    - i.e., Wilks with 1 dof (even if 2 or 3 free parameters)
- Decide if the excess is a discovery based on the local p-value
- Eventually quote a **global** p-value, based on estimation of LEE



- This is unfair, not all cases have the same LEE
  - Going to an extreme case, one can make a 5σ local excess from just a 1σ if one chooses a loose model, ie 20 parameters

Beyond local p-values

- Should move to **global** p-values to more properly quantify the significance
- Should be coupled with relaxing the 5σ requirement
  - Obviously experiments reluctant to *degrade* their discovery!
- Unfair to consider a discovery with 5σ local and 4σ global and reject a 4.5σ global
- All this is about the elsewhere in the studied spectrum, still there are unpredictable elsewheres, some safety margin is good 😉
- My personal opinion:
  - **Global p-values have to be used and 4σ should be enough**
  - **More emphasis on scrutinizing the systematic errors than having 4.9 or 5.0 σ**

Discoveries from measurements

Discoveries based on measurements: **caution**

- We find statements like:
  - I measure $x = 5 \pm 1$, SM predicts $x = 0$ $\Rightarrow$ I have 5σ $\Rightarrow$ I made a discovery
  - Well, that's not exactly true
    - **It is certainly an interesting result!**
    - But the uncertainties are calculated under assumptions, not necessarily valid at 5σ : gaussian behavior, error propagation (linearity), systematic tails not always checked
    - Should turn into a proper hypothesis test (do a proper hypothesis test)
- More worrying if the result coming from a combination

Combining for discovery

- Please do not confront your theory with ad-hoc combination of existing measurement
  - Experimental results are often correlated
  - Inside experiments, but also between experiments
  - Naïve combination will almost always give underestimated errors
- Be careful with PDG combinations
  - Much better but still incomplete for this pourpose
- BLUE combinations (Best Linear uncertainty estimator) way better
  - Account for correlations to some level
  - Still gaussian assumptions
  - Still linear

Comparing with theory

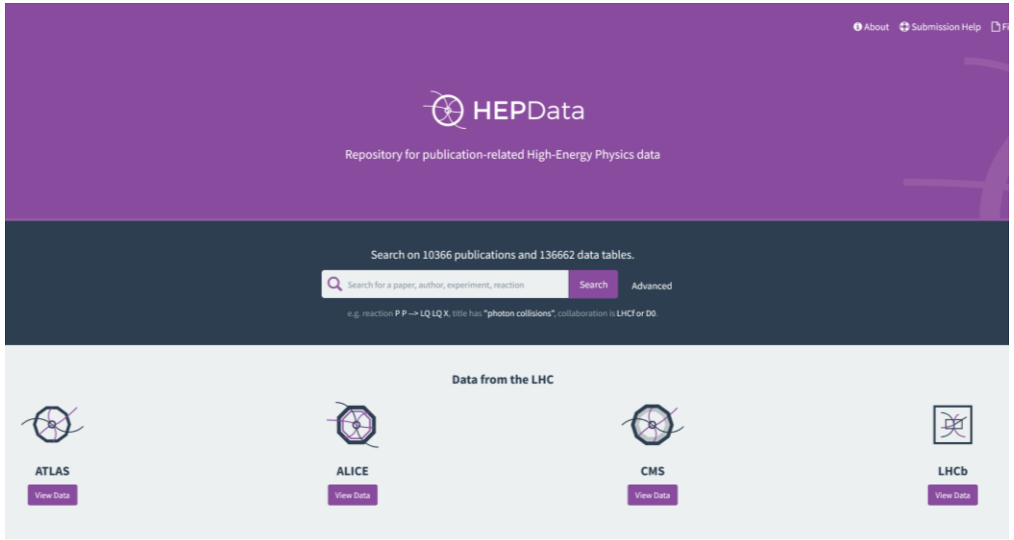
Data reinterpretation

Steps forward

- LHC experiments (and others) are aware of these limitations
- Experiments now tend to combine **data** and not results,
  - Build a global likelihood with all sets, include the systematics (nuisances) and their correlations to your best knowledge, **more than multiplying the likelihoods**
  - Do a single fit, you'll get the **most precise** measurement
- To combine experiments a bit harder…
- Even harder for theoreticians if they want to interpret their data
  - How can I test my physics model with that cross-section measurement (properly with all systematics and correlations)?

More steps

- HEPDATA
  - Most relevant information published in HEPDATA
  - 14000 data tables published!
- *Simplified likelihoods*
  - An attempt was made to make public a simplified version of the likelihood function
  - So that everyone could plug-in their data or theoretical model
- Publish the whole model
  - Publish full set of data, nuisances, statistical model…
  - **And tools**
  - Some results already available



Forum on the Interpretation of the LHC Results for BSM studies

- Please have a look to the twiki of this working group or to some of their publications
  - Les Houches guide to reusable ML models in LHC analyses
- Snowmass white paper on Data and Analysis Preservation, Recasting, and Reinterpretation
- White paper on Publishing statistical models: Getting the most out of particle physics experiments
- Reinterpretation of LHC Results for New Physics: Status and recommendations after Run 2

A few things you will be able to do

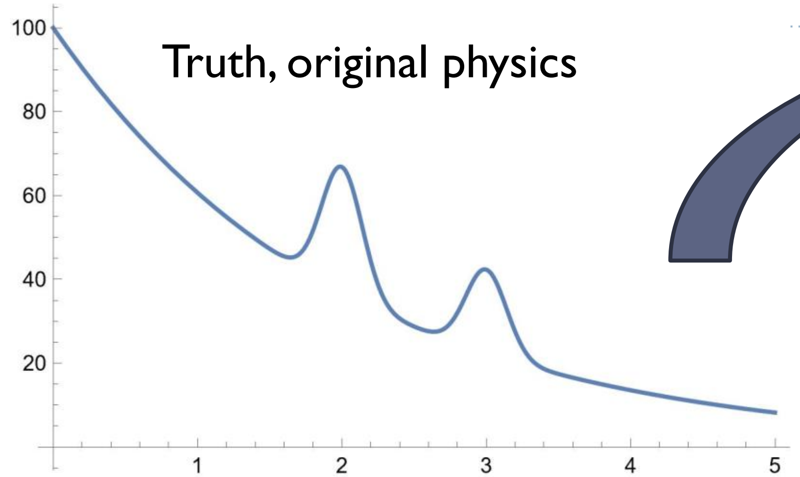
- Update existing analyses using
  - more precise theoretical calculations
  - improved experimental calibrations
  - different probability model…
- Kinematic reinterpretation considering a different physical process with a different phase space distribution, which might have different efficiencies
- Combinations of analyses or datasets in model surveys, global averages…
- Reuse of datasets for other studies such that the determination of parton distribution functions

unfolding

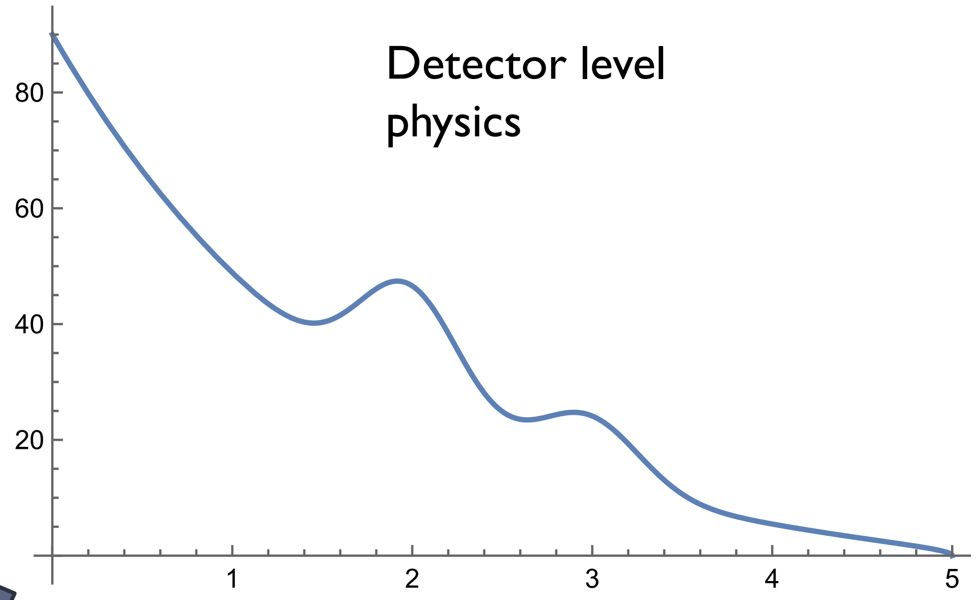
Correct your results for detector effects so it can directly be compared with theory

The problem

Truth, original physics



Detector level physics



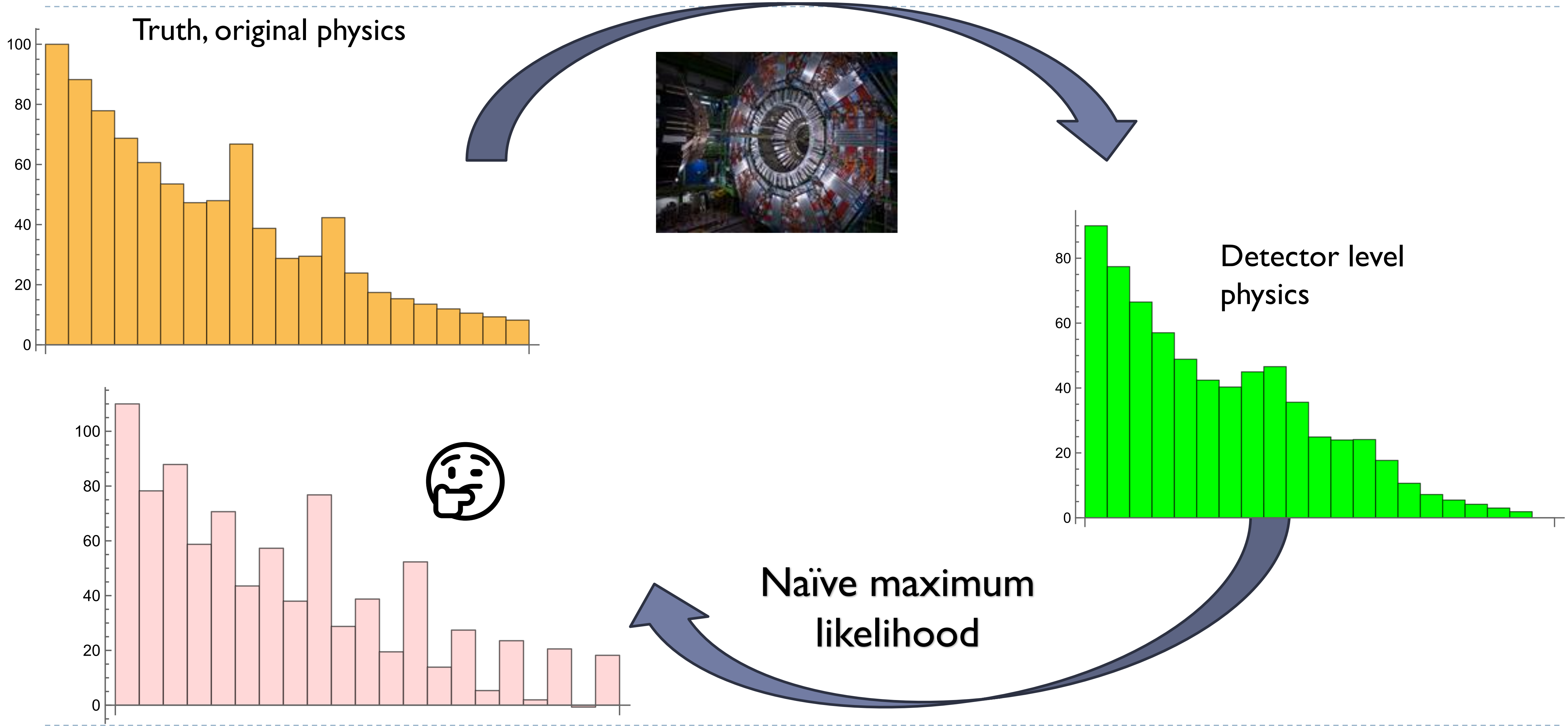
?

An inverse problem: basically, recover original quantities before suffering resolution and efficiency effects when going through a detector

A simple problem?

- So what?, I can:
  - Discretize my truth and experimental data in histograms, t_j, d_i
  - And use simulation to calculate the transfer (migration) matrix R_{ij}, connecting t_j *and* d_i
  - Then I expect $\vec{d} \leftarrow R\vec{t}$, I can do a MLE or maybe just $\vec{t} = R^{-1}\vec{d}$
- Unfortunately, does not usually work

Not so easy


Truth, original physics
Detector level physics
Naïve maximum likelihood

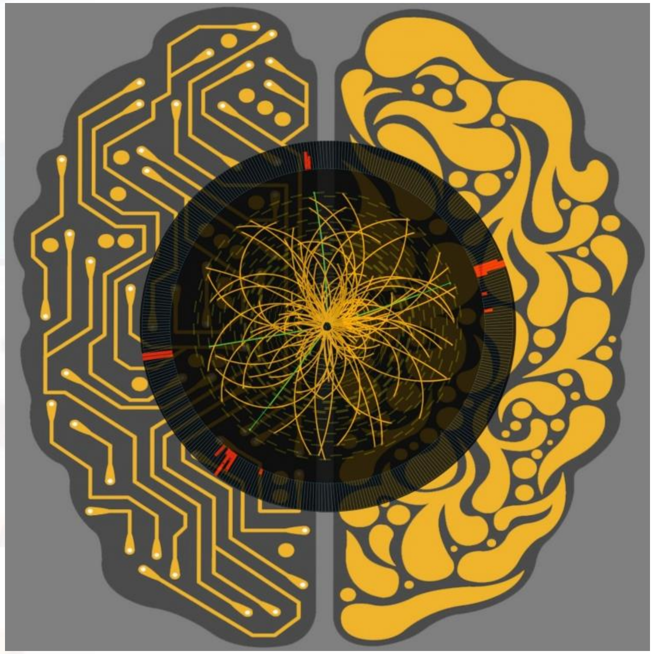
unfolding

- The problem is known to be ill-posed, with instabilities introducing high frequency terms
- Several techniques used to overcome the problem (See Mikael Kuusela talk for details): iterative, Tikhonov, NN, wide/narrow binning
- Include regularization, additional information to soften the fluctuations
- The challenge:
  - If you are doing the measurement, optimize the bias-uncertainty trade-off. Publish the whole information
  - If you are using the unfolded results, be aware that a bias exists, be aware that there are (potentially big) correlations
  - If you are doing combinations, consider possible correlations with older measurements

Machine learning

A whole new field of research

- Used in HEP since at least 20 years, but as in many other applications going through a big bang lately
- Initially only used for classification and (timidly) for regression (fits)
  - Particle id and calibration
  - Anomaly detection
  - Unfolding and other inverse problems
  - Simulation
  - Density estimation
  - Detector optimization
  - Reweighting MC
  - Theory: param tuning, lattice, nuclear…
- And many techniques: DNN, GAN, CNN, GNN…



Francisco Matorras, IFCA, Spain

Some challenges (from a statistics point of view)

- Overfitting and Generalization
  - Does it introduce a systematic?
- NN modeling uncertainty Quantification
- Bias and Systematic Errors
  - Systematics usually calculated one at a time, but ML power from combined separation
- Interpretability and Explainability
  - Where the power comes from, useful to understand systematics
- Handling Imbalanced Datasets
  - What if we look for a tiny signal?
- …

Summary and conclusion

- Statistics plays an important role in particle physics and is currently an active field
- It is probably time to revise de 5σ convention
- I suggest to move to global p-values and drop the concept of local p-value
- An important effort ongoing in LHC community to publish the whole data and statistical model to permit optimal and correct public (re)interpretation
- Machine learning is boiling (also here), lot of new techniques and challenges to accommodate them in our analyses

Thank you for your attention