Physics Seminar

DESY, December 13th 2016

Extraordinary Claims: the 0.000029% Solution,

or

Anomalies in Collider Data

Tommaso Dorigo
INFN Padova

Why This Seminar

- Driven by the Higgs search and discovery hype, in the last few years science outreach agents have busied themselves explaining to the public the idea that a scientific discovery in physics research requires that an effect be found with a statistical significance exceeding five standard deviations.

  → An **entirely arbitrary** convention, to be used with caution
  or substituted with something smarter

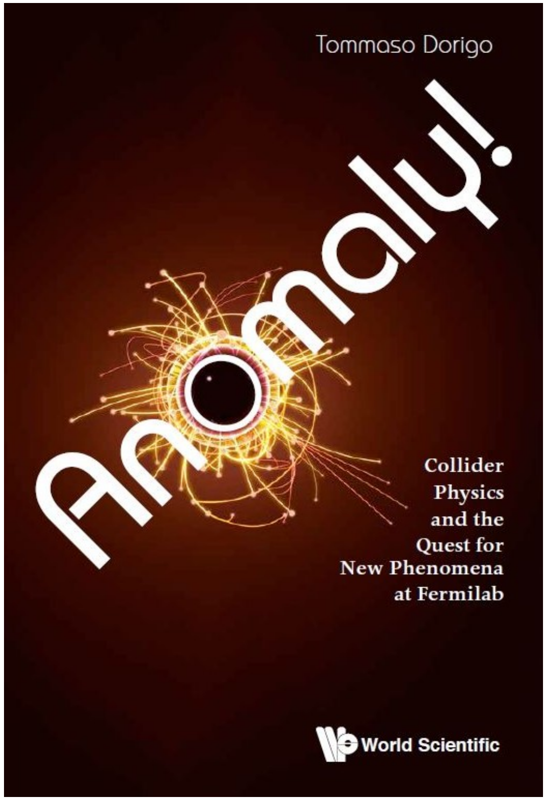
- Ultimately, conventions may still be a good thing provided one remembers their rationale – i.e. their roots

- One of the purposes of this seminar is to refresh our memory about where the five-sigma criterion comes from, what it was designed to address, where it may fail, and to consider its limitations and the need for good judgement when taking the decision to claim a discovery

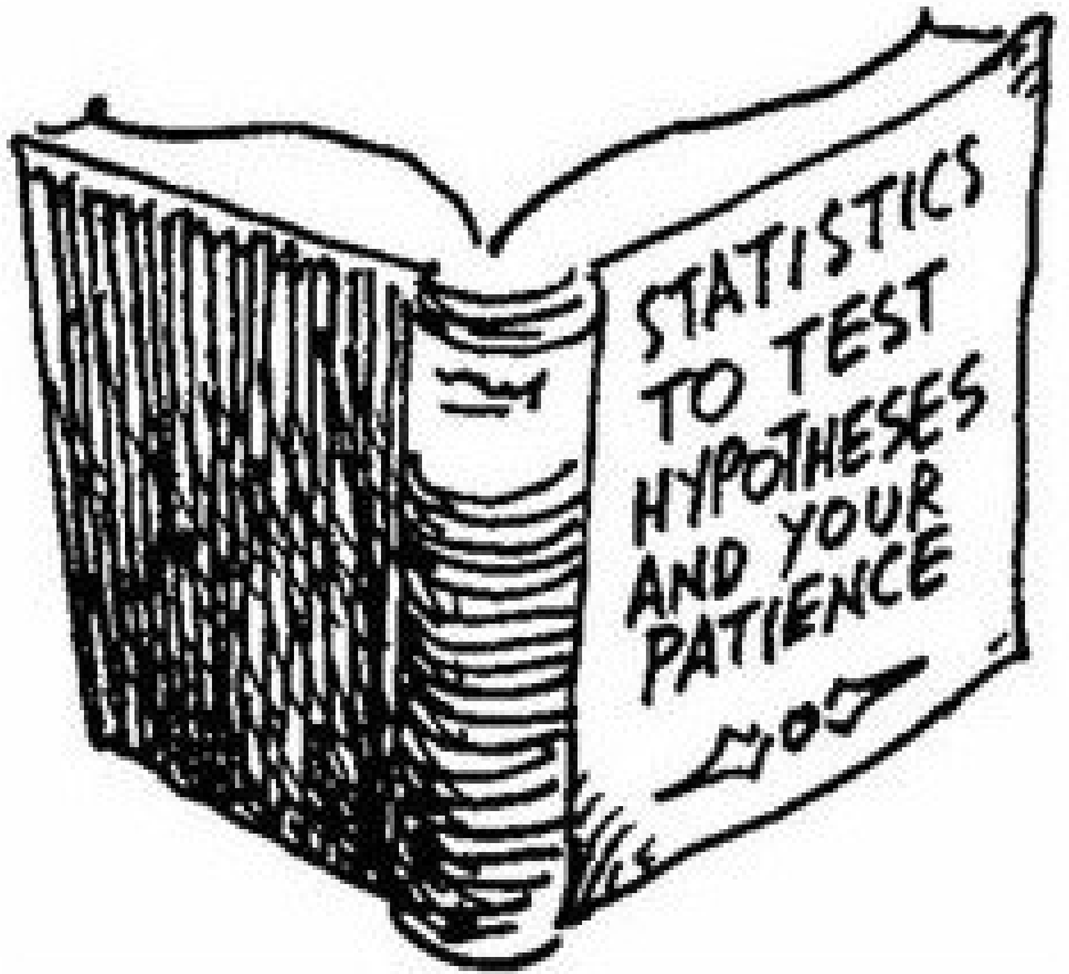
- In pursuit of that goal, we will examine several anomalous effects that surfaced in particle physics experiments – in search for insight, patterns, pitfalls of the blind sigma-counting.

Contents

- **Being on the same page**
  - Hypothesis testing in five slides: p-value, significance, Wilks' theorem, type-I and type-II error rates
- **The birth of the five-sigma criterion**
  - Rosenfeld on exotic baryons
  - Lynch and the GAME program
- **Anomalies** **in collider data**
  - CDF stories
  - Other successful and failed applications in recent times
- **The trouble with 5 σ**
  - Ill-quantifiable trial factors
  - Subconscious Bayes factors
  - Systematics
  - The Jeffrey-Lindley paradox
- **How to fix it ?**
  - Lyons' table
  - Agreeing on flexible thresholds



Hypothesis Testing in Five Slides



Statistical Significance: What it is

- Statistical significance is a way to report the probability that an experiment obtains data at least as discrepant as those actually observed, under a given "null hypothesis“ H_0

- In physics H_0 *usually* describes the currently accepted and established theory

- Given some data X and a suitable test statistic T (a function of X), one starts with the **p**-value, *i.e*. the probability of obtaining a value of T at least as extreme as the one observed, if H_0 is true.

  p can always be converted into the corresponding number of "sigma," *i.e.* standard deviation units from a Gaussian mean. This is done by finding **x** such that the integral from **x** to infinity of a unit Gaussian N(0,1) equals **p**:

$$\frac{1}{\sqrt{2\pi}}\int_x^\infty e^{-\frac{t^2}{2}}dt = p$$

- According to the above recipe, a **15.9%** probability is a one-standard-deviation effect; a **0.135%** probability is a three-standard-deviation effect; and a **0.0000285%** probability corresponds to five standard deviations - "**five sigma**" in jargon.

Notes

A few facts are worth noticing:

- the convention is to use a "one-tailed" Gaussian: we do not consider departures of x from the mean in the *un-interesting direction*
  - Hence "negative significances" are mathematically well defined, but we do not care about them

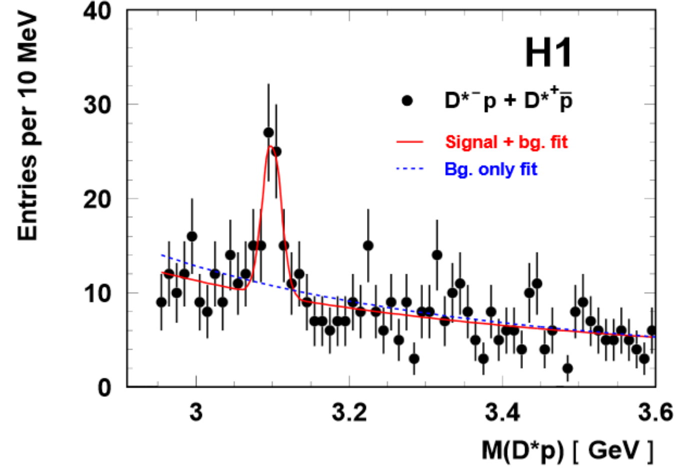
- the conversion of p into σ is fixed and independent of experimental detail. As such, using Nσ rather than p is **just a shortcut** to avoid handling numbers with many digits:

  we prefer to say "5σ" than "0.00000029" just as we prefer to say "a nanometer" instead than "0.000000001 meters" or "a Petabyte" instead than "1000000000000 bytes"

- The whole construction rests on a proper definition of the p-value. Any shortcoming of the properties of p (*e.g*. a tiny non-flatness of its PDF under the null hypothesis) totally invalidates the meaning of the derived Nσ
  - In particular, using "sigma" units does in no way mean we are espousing some kind of Gaussian approximation for our test statistic or in other parts of our problem. Care required here, as many are still led to confusion on this bit

- The "*probability of the data*" has no bearing on the concept, and is not used. What is used is the probability of a subset of the possible outcomes of the experiment, defined by the outcome actually observed (**as much** or **more** extreme)

An Important Ingredient: Wilks' Theorem



- A common method to derive a significance from a likelihood fit is the one of invoking **Wilks' theorem:**

- One has a likelihood under the null hypothesis, L_0 (e.g., a background-only fit), and a likelihood for an alternative, L_1 (a signal+background fit)

- One takes **$-2\ (\ln L_1 - \ln L_0) = -2\ \Delta\ (\ln L)$** and interprets it as a value sampled from a chisquare distribution

- $P(\chi^2, N_{dof})$ can then be obtained as a "**tail probability**", and from it a Z-value

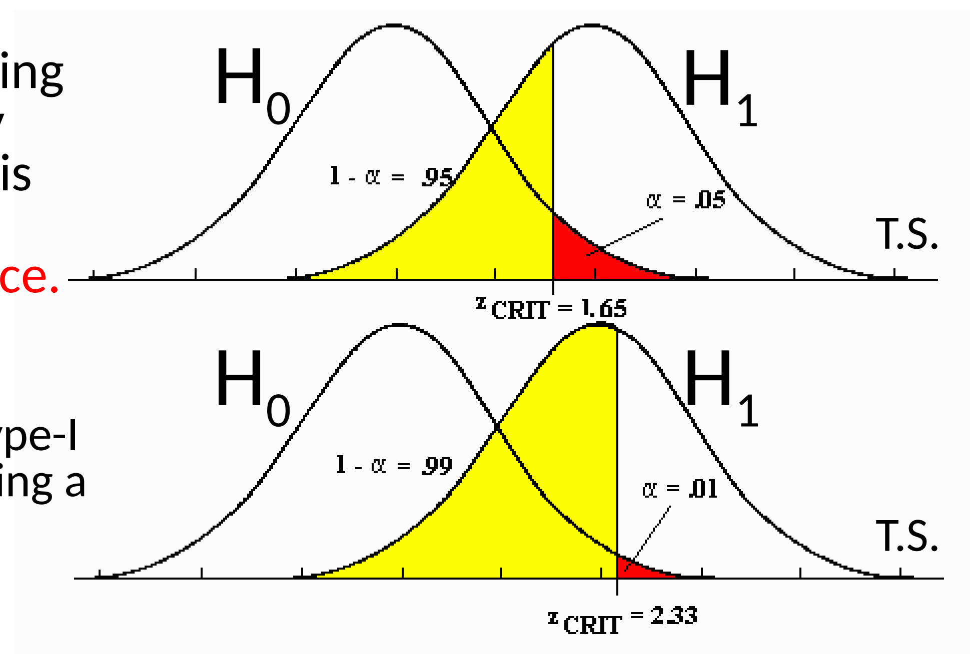
  – One should not forget that this is only applicable when the two hypotheses are connected by H_0 being a particular case of H_1 (i.e., $H_0 == H_1$ when some of the H_1

Type-I and Type-II Errors

n the context of hypothesis testing the type-I error rate α is the probability of rejecting the null hypothesis when it is true.

Strictly connected to α is the concept of "power" (1-β), where β is the type-2 error rate, defined as the probability of accepting the null when the alternative is instead true.

Once the test statistic is defined, by choosing α (*e.g.* to decide a criterion for a discovery claim, or to set a confidence interval) one is automatically also choosing β. In general there is no formal recipe to guide the choice.

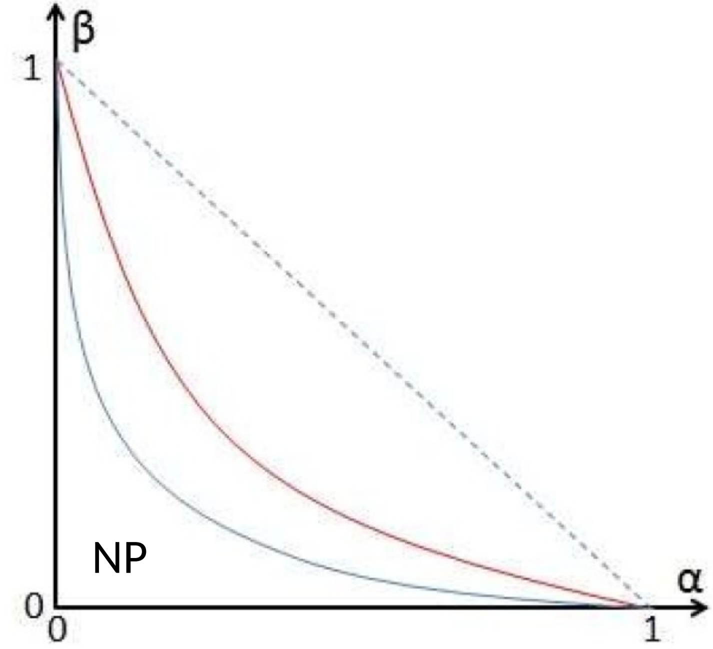


A stricter requirement for α (*i.e.* a smaller type-I error rate) implies a higher chance of accepting a false null (yellow region), *i.e.* smaller power.

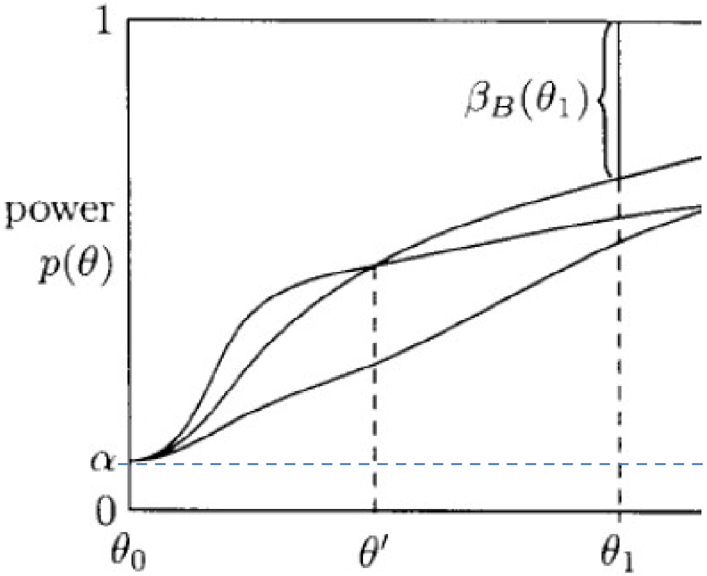
Alpha vs Beta and Power Graphs

- Where to stay in the curve provided by your analysis method highly depends on **habits in your field**
- What makes a difference is the test statistic.

The N-P likelihood-ratio test outperforms others for simple-vs-simple HT, as dictated by the Neyman-Pearsons lemma: highest power 1-β for any α.



As data size increases, the power curve (shown below) becomes closer to a step function



The **power 1-β of a test** usually depends on the parameter of interest: different methods may have best performance in different parameter space points

NB: for θ corresponding to the null hypothesis (θ_0), the power is by definition equal to α

The Birth of the Five-Sigma Criterion

Arthur H. Rosenfeld (Univ. Berkeley)

Far-Out Hadrons

- In 1968 Arthur Rosenfeld wrote a paper titled "*Are There Any Far-out Mesons or Baryons?*" **[1]**. In it, he demonstrated that the **number of claims** of discovery of such exotic particles published in scientific magazines agreed reasonably well with the **number of statistical fluctuations** that one would expect in the analyzed datasets.

  ("Far-out hadrons" are hypothetical particles which can be defined as ones that do not fit in SU(3) multiplets. In 1968 quarks were not yet fully accepted as real entities, and the question of the existence of exotic hadrons was important.)

- Rosenfeld examined the literature and pointed his finger at large trial factors coming into play due to the massive use of combinations of observed particles to derive mass spectra containing potential resonances:

  > *"[...] This reasoning on multiplicities, extended to all combinations of all outgoing particles and to all countries, leads to an estimate of 35 million mass combinations calculated per year. How many histograms are plotted from these 35 million combinations? A glance through the journals shows that a typical mass histogram has about 2,500 entries, so the number we were looking for, h is then 15,000 histograms per year."*

More Rosenfeld

“[…] Our typical 2,500 entry histogram seems to average 40 bins. This means that therein a physicist could observe 40 different fluctuations one bin wide, 39 two bins wide, 38 three bins wide... This arithmetic is made worse by the fact that when a physicist sees 'something', he then tries to enhance it by making cuts…”

(We shall get back to the last issue later)

“In summary of all the discussion above, I conclude that each of our 150,000 annual histograms is capable of generating somewhere between 10 and 100 deceptive upward fluctuations […]”.

That was indeed a problem! A comparison with the literature in fact showed a correspondence of his eyeballed estimate with the number of unconfirmed new particle claims.

Rosenfeld concluded:

*“To the theorist or phenomenologist the moral is simple: **wait for nearly 5σ effects**. For the experimental group who has spent a year of their time and perhaps a million dollars, the problem is harder... go ahead and publish... but they should realize that any bump less than about 5σ **calls for a repeat of the experiment**.”*

Gerry Lynch and GAME

Rosenfeld's article also cites the half-joking, half-didactical effort of his colleague Gerry Lynch at Berkeley:

> *"My colleague Gerry Lynch has instead tried to study this problem 'experimentally' using a 'Las Vegas' computer program called Game. Game is played as follows. You wait until a unsuspecting friend comes to show you his latest 4-sigma peak. You draw a smooth curve through his data (based on the hypothesis that the peak is just a fluctuation), and punch this smooth curve as one of the inputs for Game. The other input is his actual data. If you then call for 100 Las Vegas histograms, Game will generate them, with the actual data reproduced for comparison at some random page. You and your friend then go around the halls, asking physicists to pick out the most surprising histogram in the printout. Often it is one of the 100 phoneys, rather than the real '4-sigma' peak."*

Obviously particle physicists in the '60s were more "bump-happy" than we are today. The proposal to raise to 5-sigma of the threshold above which a signal could be claimed was an earnest attempt at reducing the flow of claimed discoveries, which distracted theorists and caused confusion.

Let's Play GAME

It is instructive even for a hard-boiled sceptical physicist raised in the years of Standard-Model-Precision-Tests Boredom to play GAME.

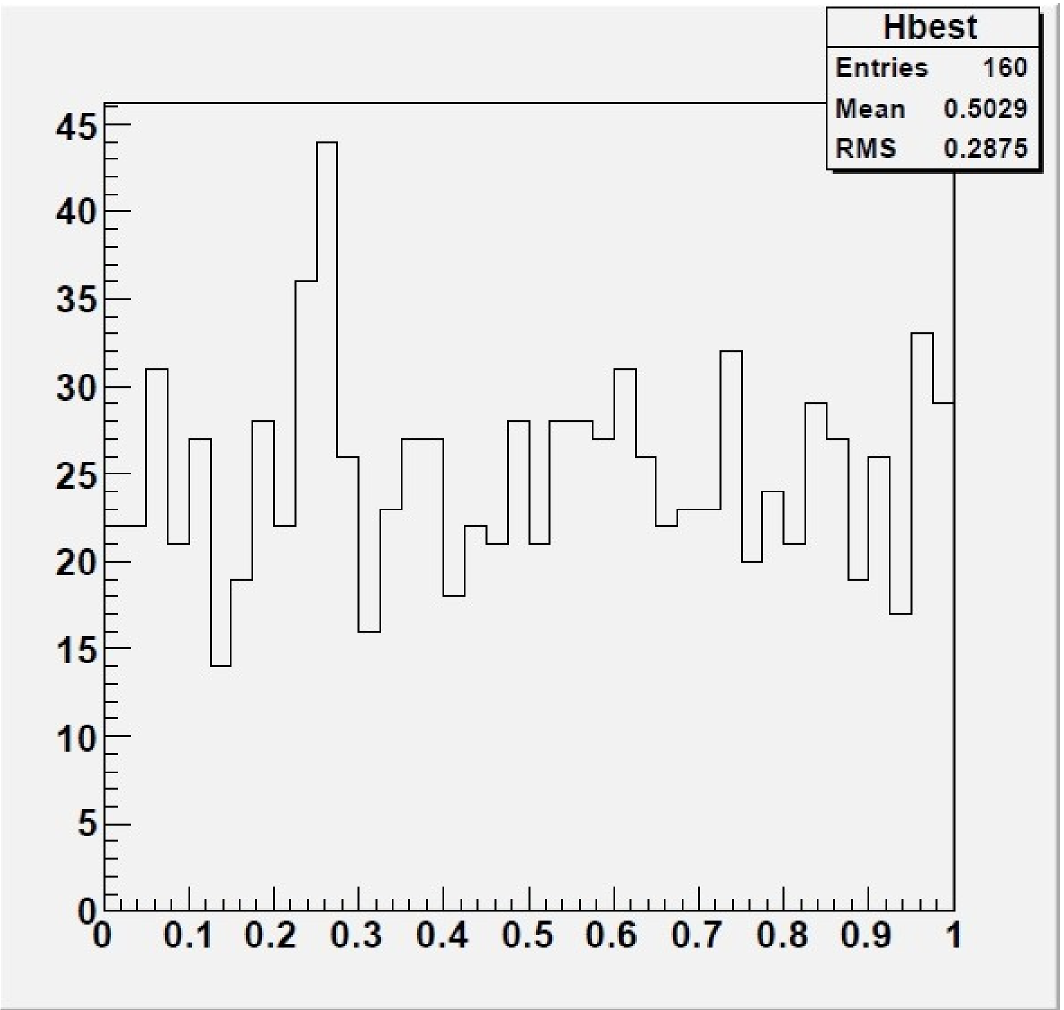
In the following slides are shown a few histograms, **each** selected by an automated procedure **as the one** containing "the most striking" peak **among a set of 100** drawn from a **uniform distribution.**

Details: 1000 entries; 40 bins; the "best" histogram in each set of 100 is the one with most populated adjacent pair of bins (in the first 4 slides) or triplets of bins (in the second set of 3 slides)

You are asked to consider **what you would tell your student if she came to your office with such a histogram**, claiming it is the result of an optimized selection for some doubly charmed baryon, say, that she has been looking for in her research project.

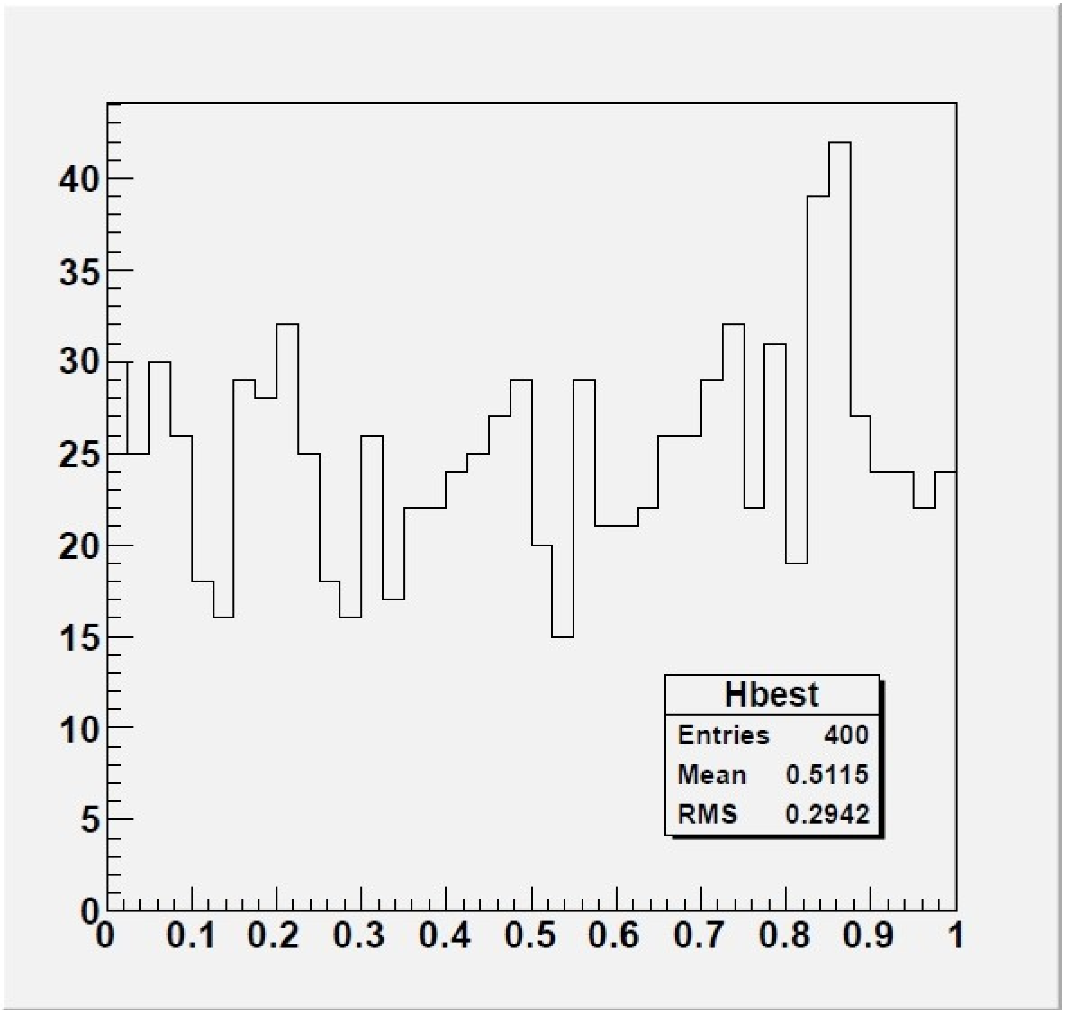
2-Bin Bumps

- Here are the outputs of the most significant 2-bin bumps in five 100-histogram sets: #1



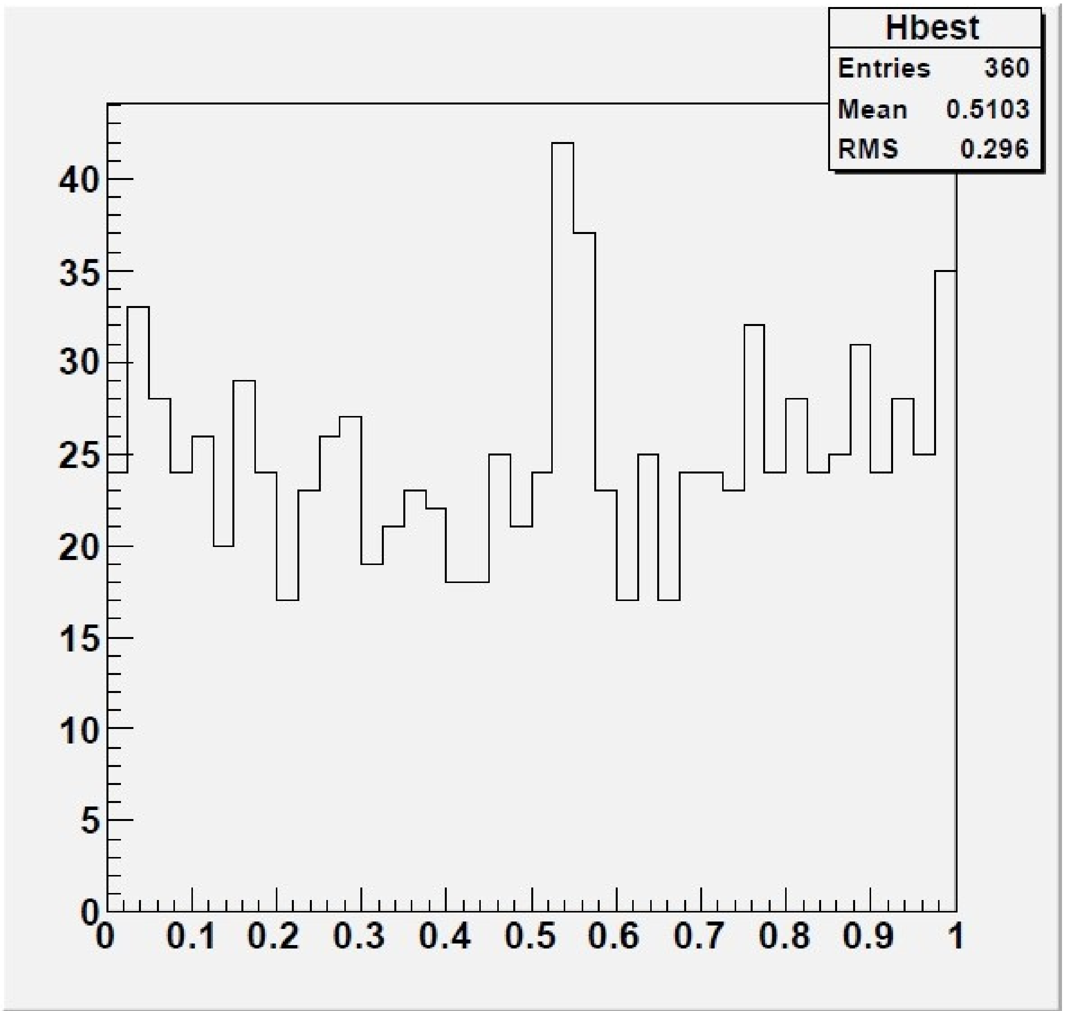
2-Bin Bumps

- Here are the outputs of the most significant 2-bin bumps in five 100-histogram sets: #2



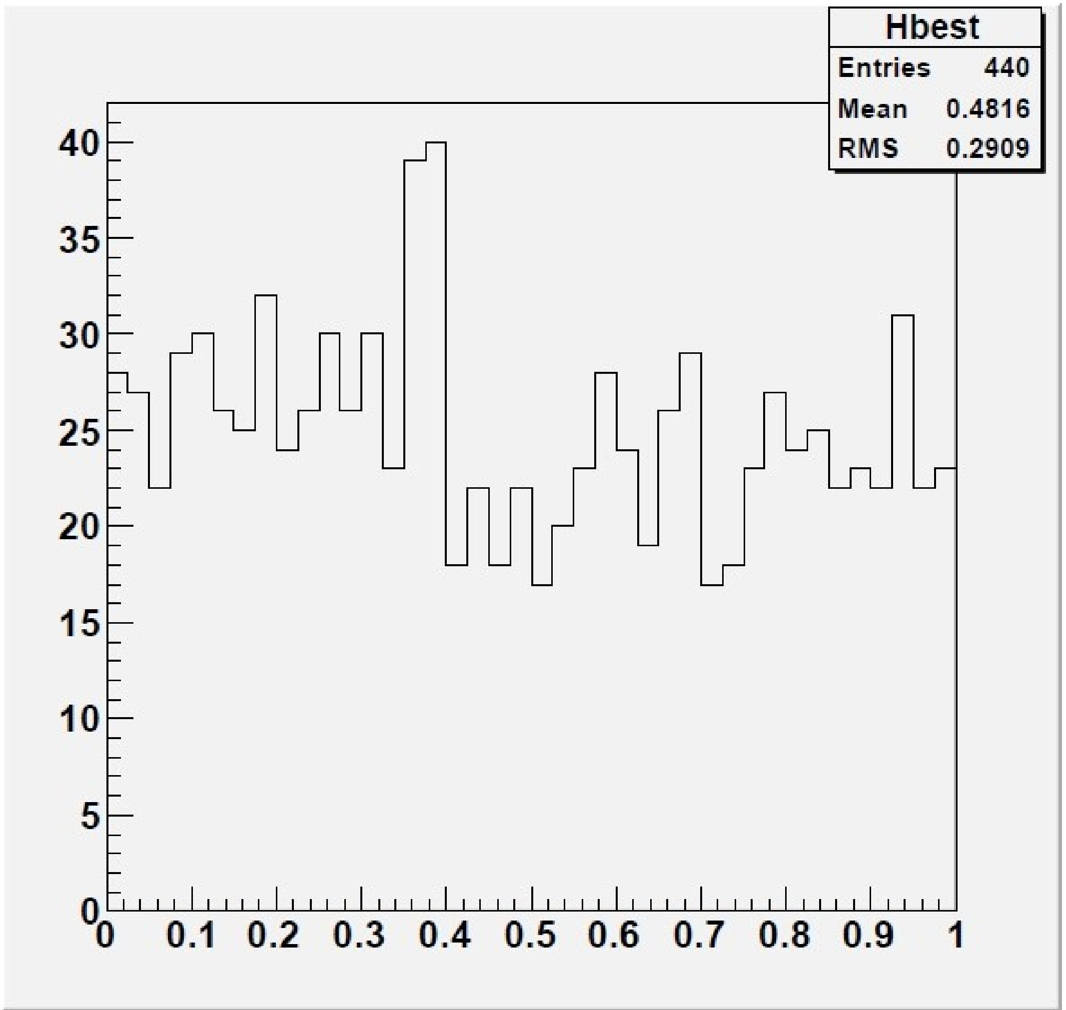
2-Bin Bumps

- Here are the outputs of the most significant 2-bin bumps in five 100-histogram sets: #3



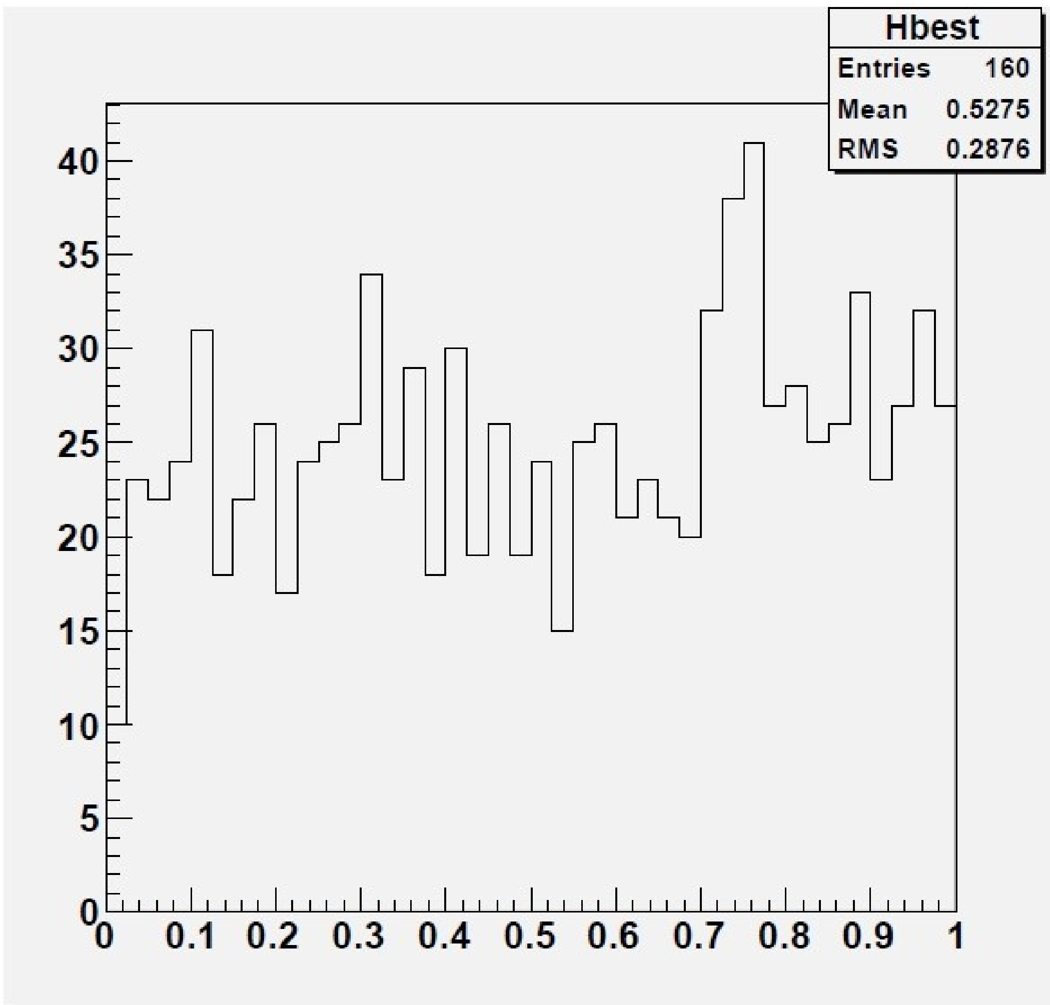
2-Bin Bumps

- Here are the outputs of the most significant 2-bin bumps in five 100-histogram sets: #4



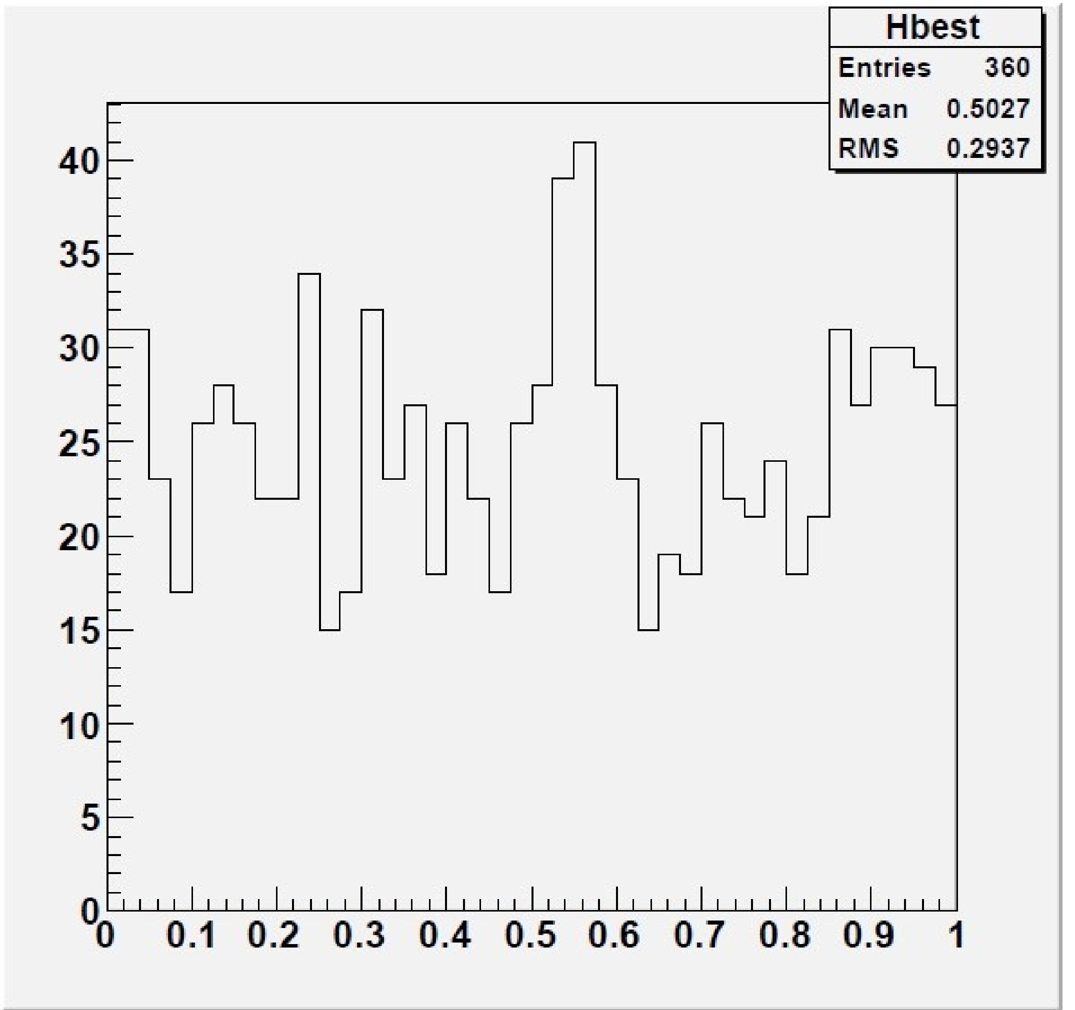
3-Bin Bumps

- Here are the outputs of the most significant 3-bin bumps in five 100-histogram sets: #1



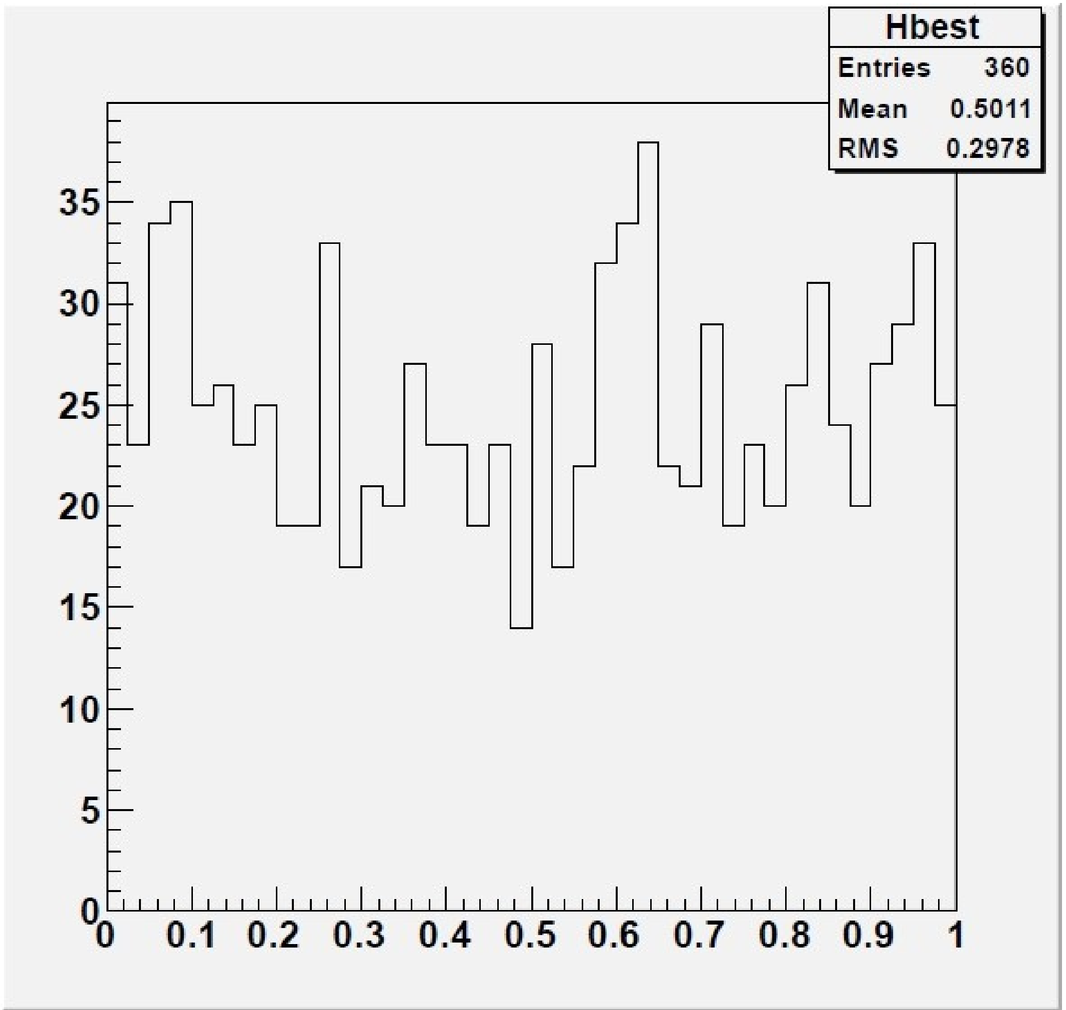
3-Bin Bumps

- Here are the outputs of the most significant 3-bin bumps in five 100-histogram sets: #2



3-Bin Bumps

- Here are the outputs of the most significant 3-bin bumps in five 100-histogram sets: #3



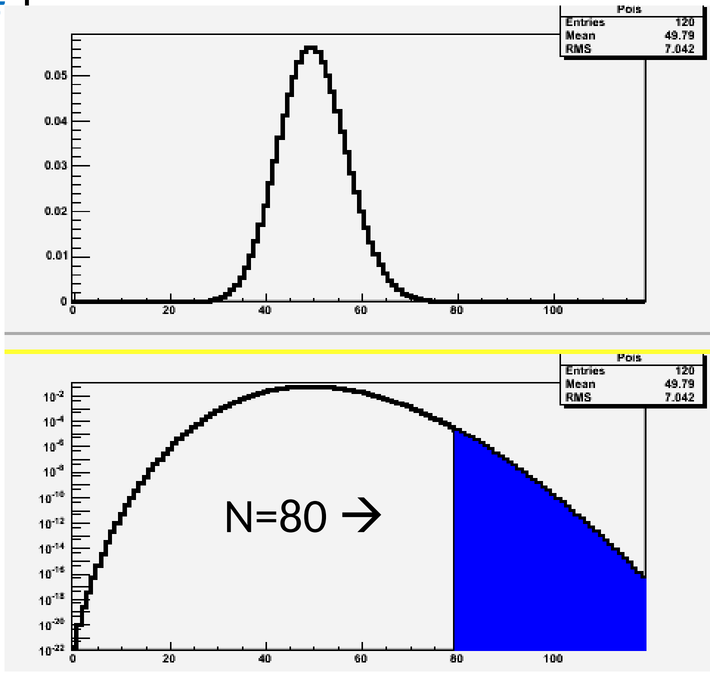
Notes on GAME

Each of the histograms in the previous slides is the best one in a set of a hundred; yet some of the isolated signals have p-values in the 0.0001 range, corresponding to 3.5σ - 4σ effects

[*As the 2-bin bumps contain N=80 events with an expectation of* $\mu=2*1000/40=50$, *and* $p_{Poisson}(\mu=50;N>=80) = 5.66*10^{-5}$ → $Z = 3.86\,\sigma$]

Why so large significance?

Because the bump can appear anywhere (x39) in the spectrum – we did not specify beforehand where we would look because we admit 2- as well as 3-bin bumps as "interesting"



P(N|μ=50) in linear (top) and semi-log scale (bottom)

What 5σ May Do For You

- Setting the bar at 5σ for a discovery claim undoubtedly removes the large majority of spurious signals due to statistical fluctuations
  - The trials factor required to reach 10^{-7} probabilities is of course very large, but in today's experiments we do perform a large number of searches!

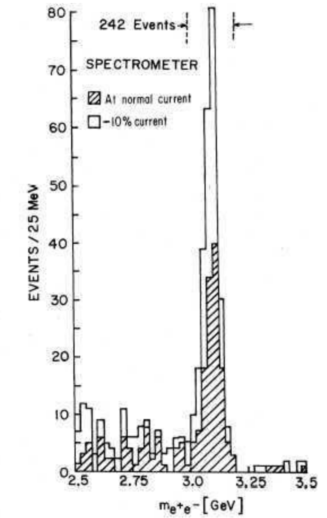
- Nowadays we call this "**LEE**", for "look-elsewhere effect".

- The other reason at the roots of the establishment of a high threshold for significance has been the ubiquitous presence in our measurements of unknown, or ill-modeled, systematic uncertainties
  - To some extent, a 5σ threshold protects systematics-dominated results from being published as discoveries

Protection from trials factor and unknown or ill-modeled systematics is the rationale behind the 5σ criterion

It is to be noted that the criterion has no basis in professional statistics literature, and is considered **totally arbitrary** by statisticians, no less than the 5% threshold commonly used for HT in medicine, biology, social sciences, *et cetera*. As shown before, the type-1 error rate is an arbitrary choice.

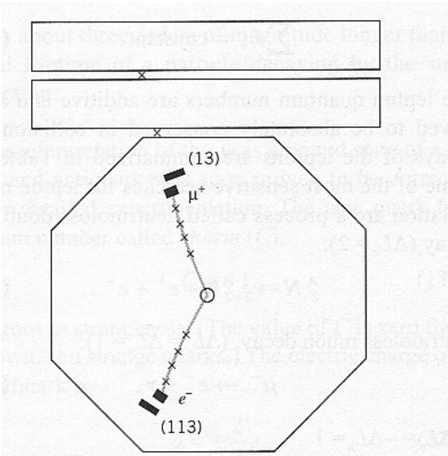
How 5σ Became a Standard in HEP: 1 - the Seventies



In the seventies the gradual consolidation of the SM shifted the focus from random bump hunting to more targeted searches

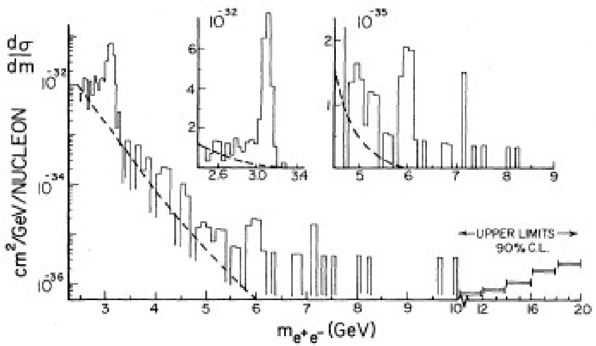
Let us have a look at a few important searches to understand how the 5σ criterion gradually became a standard

- **The J/ψ discovery** (1974): no question of significance – the bumps were too big for anybody to bother fiddling with statistical tests
- **The τ discovery** (1975-1977): no mention of significances for the excesses of (eμ) events; rather a very long debate on hadron backgrounds.



- **The Oops-Leon(1976)**: "*Clusters of events as observed occurring anywhere from 5.5 to 10.0 GeV appeared less than 2% of the time*[8]*. <u>Thus the statistical case for a narrow (<100 MeV) resonance is strong</u> although we are aware of the need for a confirmation.*"**[2]**

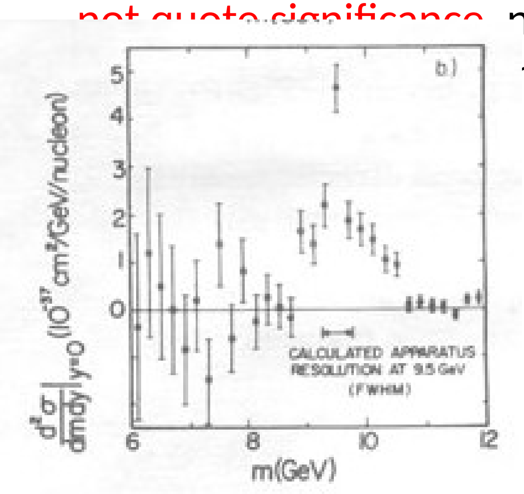
  In footnote 8 they add: "*An equivalent but cruder check is made by noting that the "continuum" background near 6 GeV and within the cluster width is 4 events. The probability of observing 12 events is again <=2%*"
  Note that $P(\mu=4;N>=12) = 0.00091$, so this does include a x20 trials factor.



The Real Upsilon

The Upsilon discovery (1977): burned by the Oops-Leon, the E288 scientists waited more patiently for more data after seeing a promising 3σ peak at 9.5 GeV

- They did exensive statistical tests to account for the trials factor (comparing MC probability to Poisson probability)
- Even after obtaining a peak with very large significance (>>5σ) they continued to investigate systematical effects
- Final announcement claims discovery but does not quote significance, noting however that the …ificant”[3]



Nov 19th 1976

I determined this factor by monte carlo. I threw 30 events over 100 bins (expectation is 2 for 6 bins) and searched for clusters of 10 in 6 bins. I found 15 successes in 40000 trys or CL = 3.75×10^{-4}. The poisson probability for ≥ 10 for an expectation of 2 is 1.94×10^{-5}. Thus bin counting factor is 19.3. JKY assumption would say 94 and 100/6 would say 17.

Nov 21st 1976

CONCLUSION: μμ I data is consistent with a narrow resonance.

So, to reiterate: ① PROBABILITY THAT THE 9.6 IS SMOOTH CONTINUUM ~1 in 1-2000 – i.e. ~3σ

② μμ I DATA CONSISTANT WITH APPARATUS RESOLUTION.

June 6th 1977

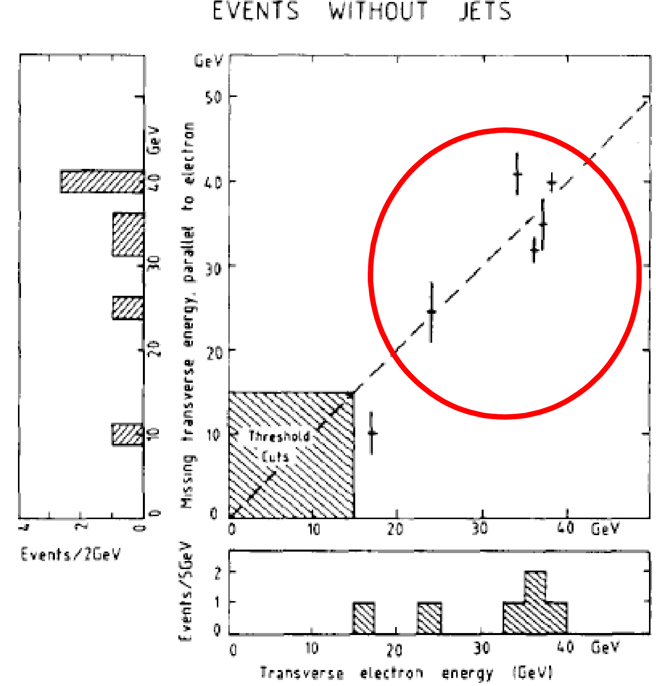
Now that the signal (>8σ) is no longer questionable from statistical objections, systematics must be considered.

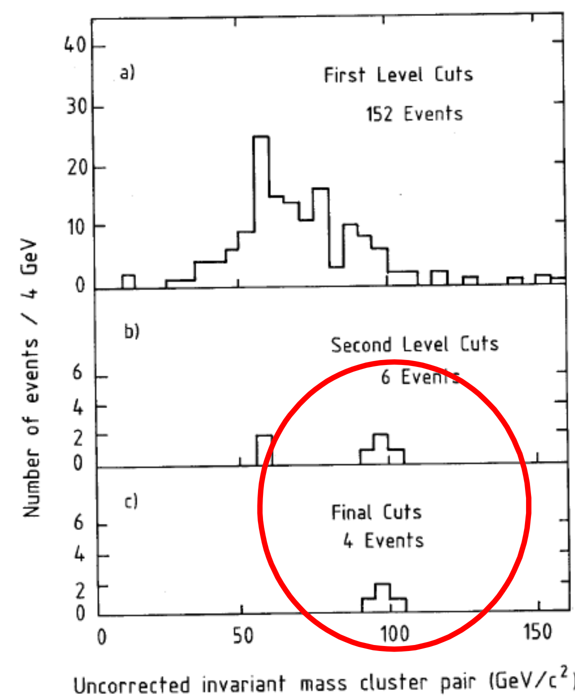
① Programming error, double counting, etc. – will be studied by

The W and Z Bosons

- The W discovery was announced on January 25th 1983 based on 6 electron events with missing energy and no jets. No statistical analysis is discussed in the discovery paper[4], which however tidily rules out backgrounds as a source of the signal
  - Note that there was no trials factor to account for: the signature was unique and predetermined; further, theory prediction for the mass (82+-2 GeV) matched well with measurement (81+-5 GeV).

- The Z was “discovered” shortly thereafter, with an official CERN announcement made in May 1983 based on 4 events.
  - Also for the Z no trials factor was applicable
  - No mention of statistical checks in the paper[5], except for notes that the various background sources were negligible.





The Top Quark Discovery

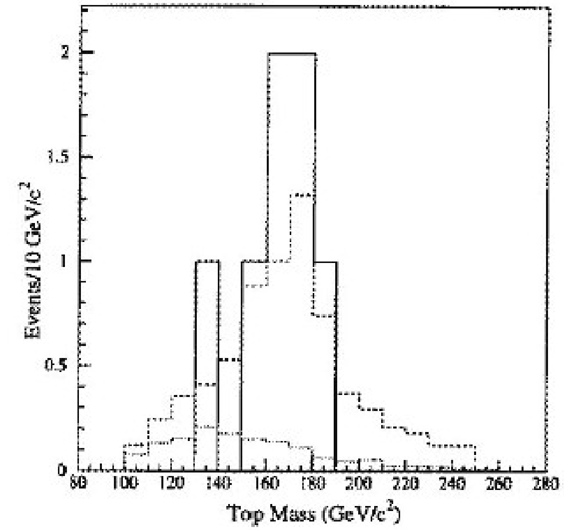
- In 1994 the CDF experiment had a serious counting excess (2.7σ) in b-tagged single-lepton and dilepton datasets, plus a towering mass peak at a value compatible with theory predictions
  - the mass peak, or corresponding kinematic evidence, was over 3σ by itself

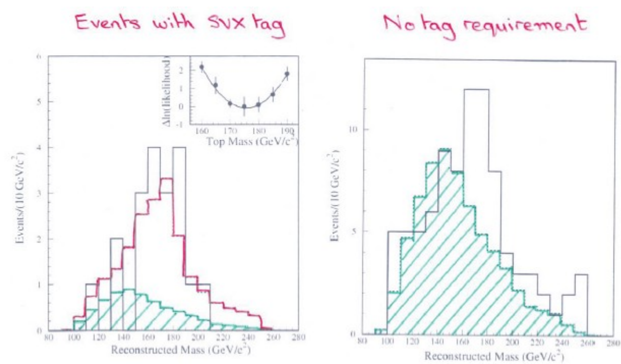
    M = 174 +- 10^{+13}_{-12} GeV (now it is 173+-0.5 GeV !)

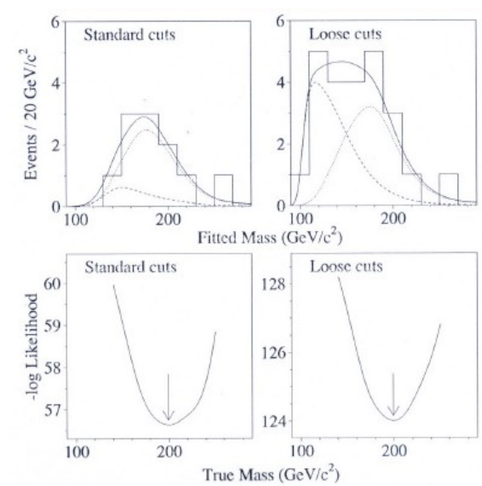
The paper describing the analysis (120-pages long) spoke of "**evidence**" for top quark production[6]

- One year later CDF and DZERO[7] both presented 5σ significances based on their counting experiments, obtained by analyzing 3x more data

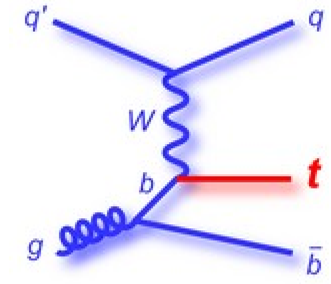
The top quark was thus the first particle discovered by a willful application of the "5σ" criterion





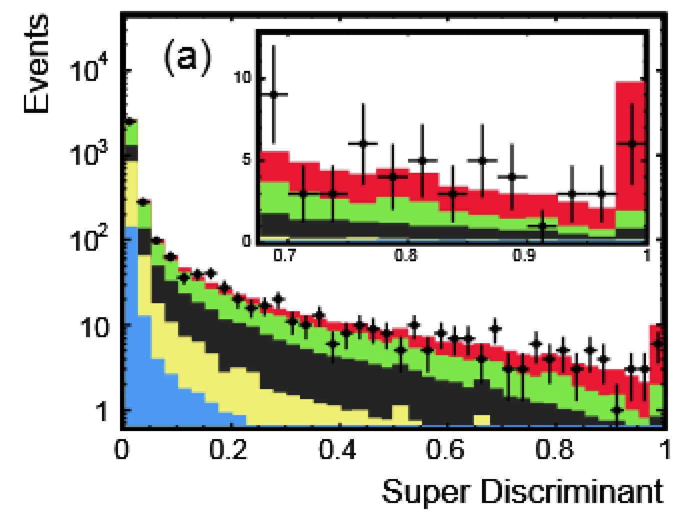


Following the Top Quark…

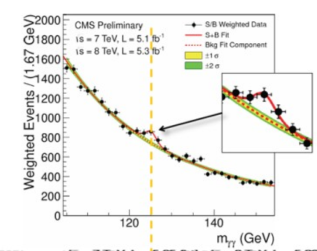


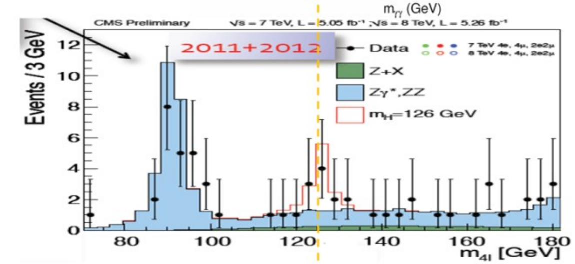
- Since 1995, the requirement of a p-value below $3*10^{-7}$ slowly but steadily became a standard. Two striking examples of searches that diligently waited for a 5-sigma effect before claiming discovery are:

  – **Single top quark production**: harder to detect than strong pair-production processes; it took 14 more years to be claimed. CDF and DZERO competed for a decade, resolving to claim observation in 2009 [8], when clear 5-sigma effects had been observed.



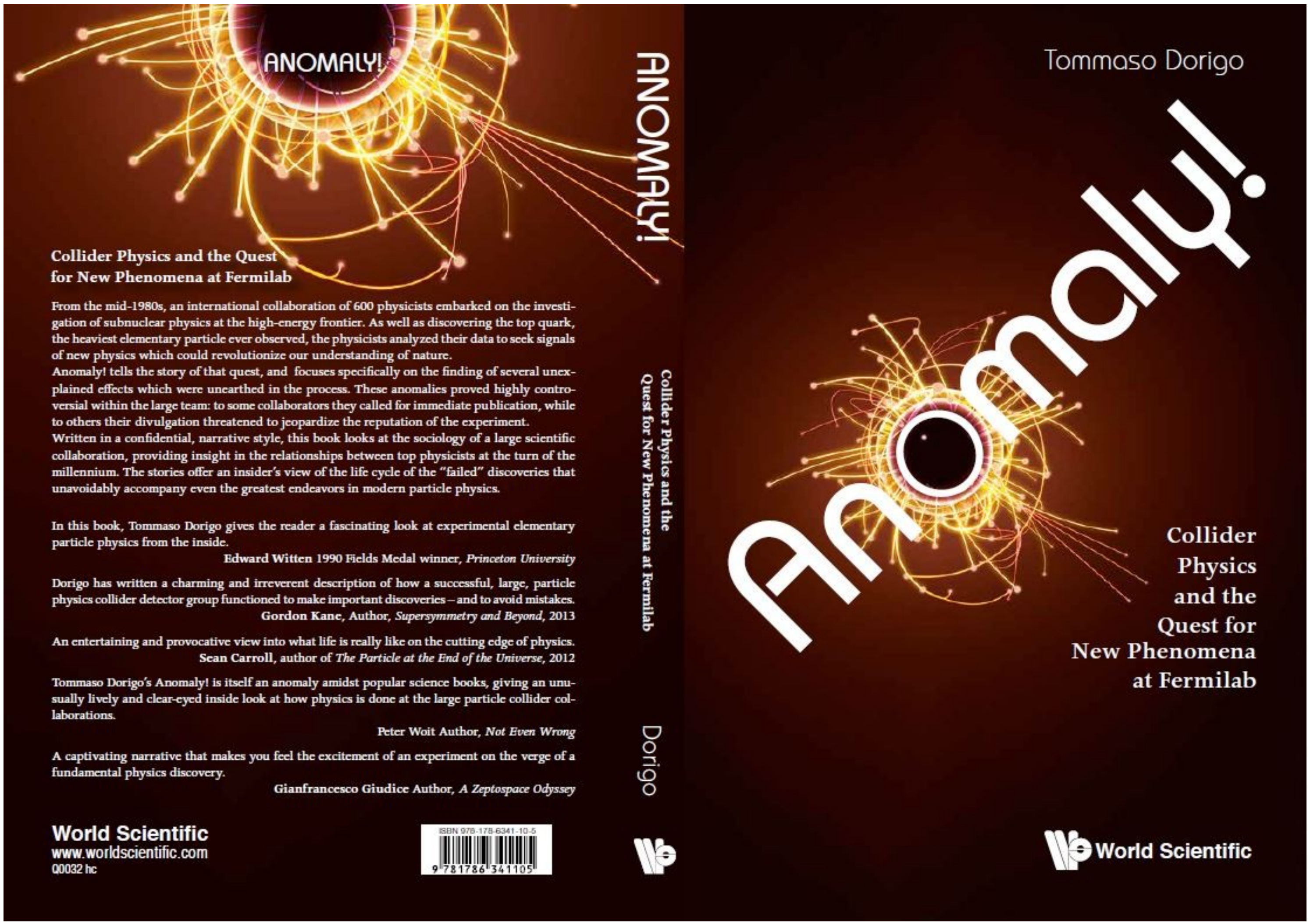
  – In 2012 the **Higgs boson** was claimed by ATLAS and CMS [9]. Note that the two experiments had mass-coincident >3σ evidence in their data 6 months earlier, but the 5σ recipe was followed diligently.



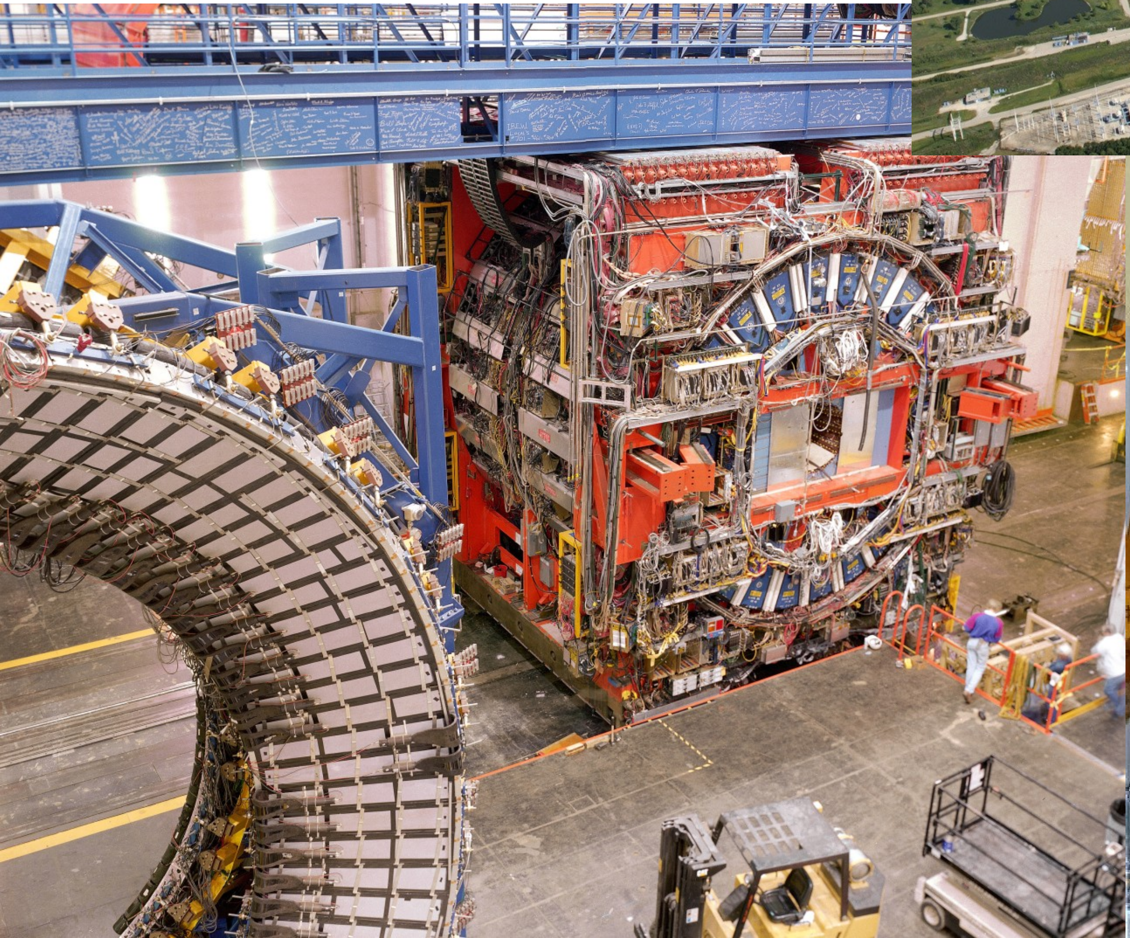


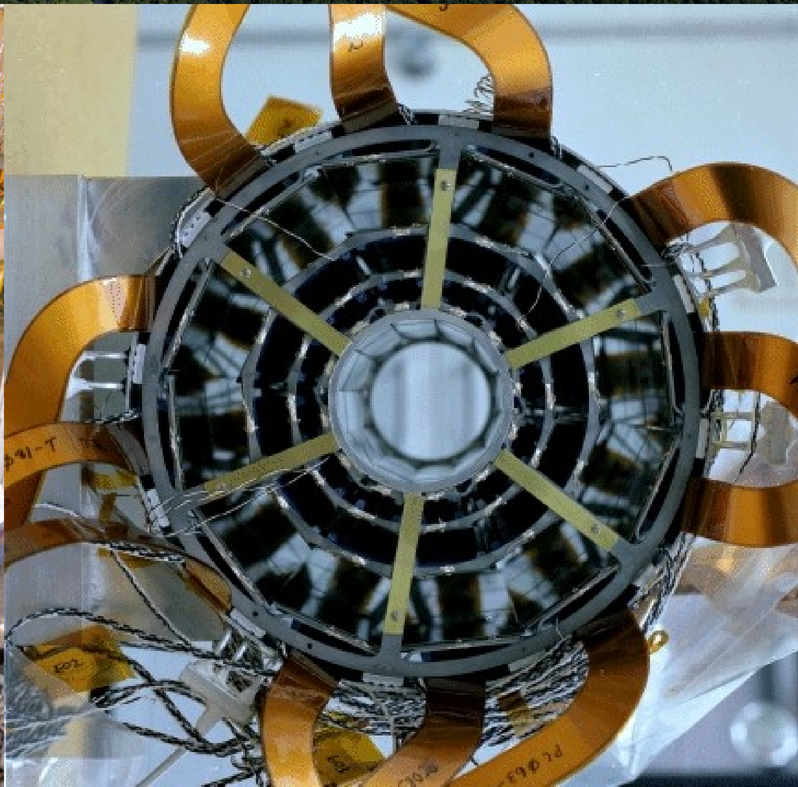
It is precisely the Higgs search what brought the five-sigma criterion to the attention of media

ANOMALIES in Collider Data



The Stage: CDF and the Tevatron





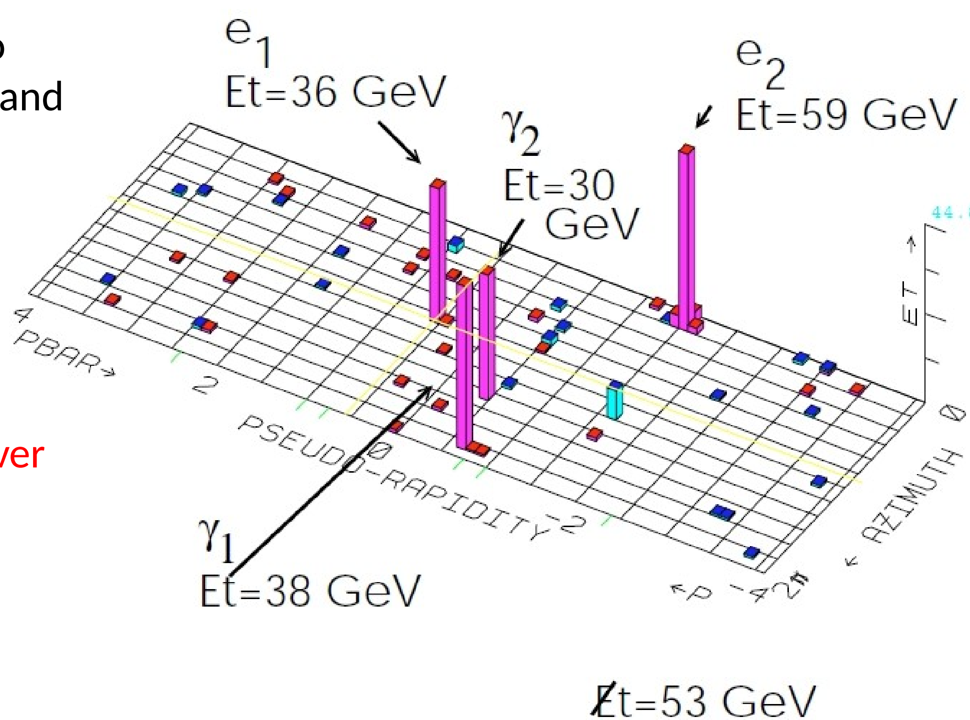
The Impossible Event

In April 1995 CDF collected an event which fired four distinct "alarm bells" by a monitoring trigger. It featured two clean electrons, two clean photons, large missing transverse energy, and *nothing else*

It could be nothing! No SM process appeared to come close to explain its presence. Possible backgrounds were estimated below 10^{-7}, a <u>6-sigma</u> find

– The observation[10] caused a whole institution to dive in a 10-year-long campaign to find "cousins" and search for an exotic explanation; it also caused dozens of theoretical papers and revamping or development of SUSY models

– In Run 2 no similar events were found; DZERO never saw anything similar either



The Fat-Jets Bump

While in the process of searching for "cousins" of the $ee\gamma\gamma+ME_T$ event, In 1996 CDF found a clear resonance structure of b-quark jet pairs at 110 GeV, produced in association with photons

The signal [11] had almost 4σ significance and looked quite good – but there was no compelling theoretical support for the state, no additional evidence in orthogonal samples, and the significance did not pass the threshold for discovery.

In addition, it was only significant when using a wide R=1.0 clustering radius…

Nothing similar resurfaced in Run 2 data, and the effect was archived.

CDF PRIVATE

Background-subtracted mass distribution of b-tagged jet pairs in photon events

The Higgs Wannabe

- The dijet bump in bbγ events was not the only one to keep CDF researchers excited. In the winter of 1996 another similar bump surfaced in W+jj events with b-tags
- Two different groups eyed the anomaly and a fierce "CDF notes" fight ensued
- The signal was again hard to explain, and suggestive of an anomalous Higgs production, but there was no way to confirm it.
- Upon closer inspection it turned out that some of the jets were not of good quality, that the event selections were somewhat fine-tuned, etcetera. The effect was finally archived
- Interestingly, the events that made it up ended up being part of a bigger controversy later on

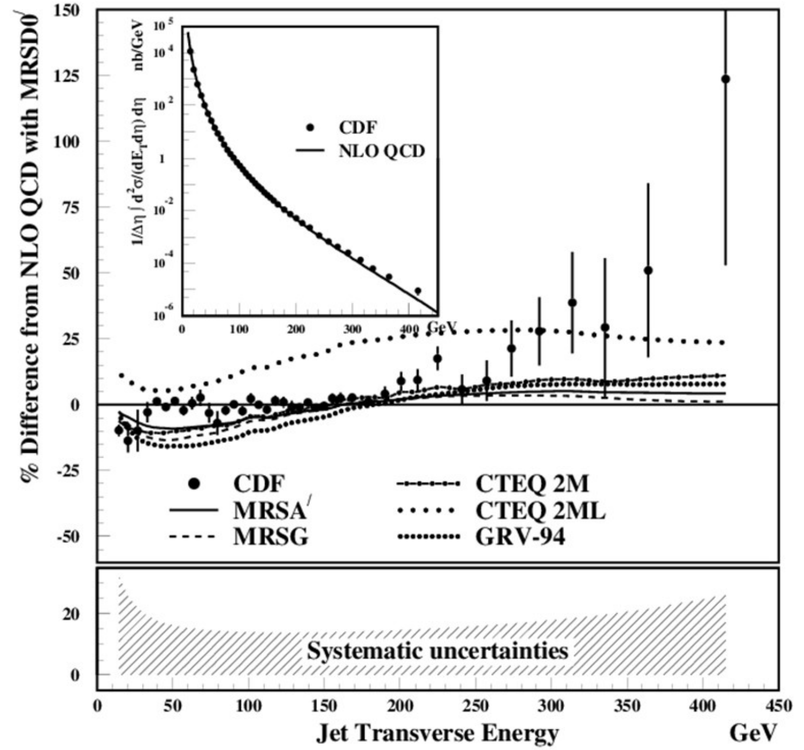
CDF PRIVATE

CDF PRIVATE

Preon Dreams

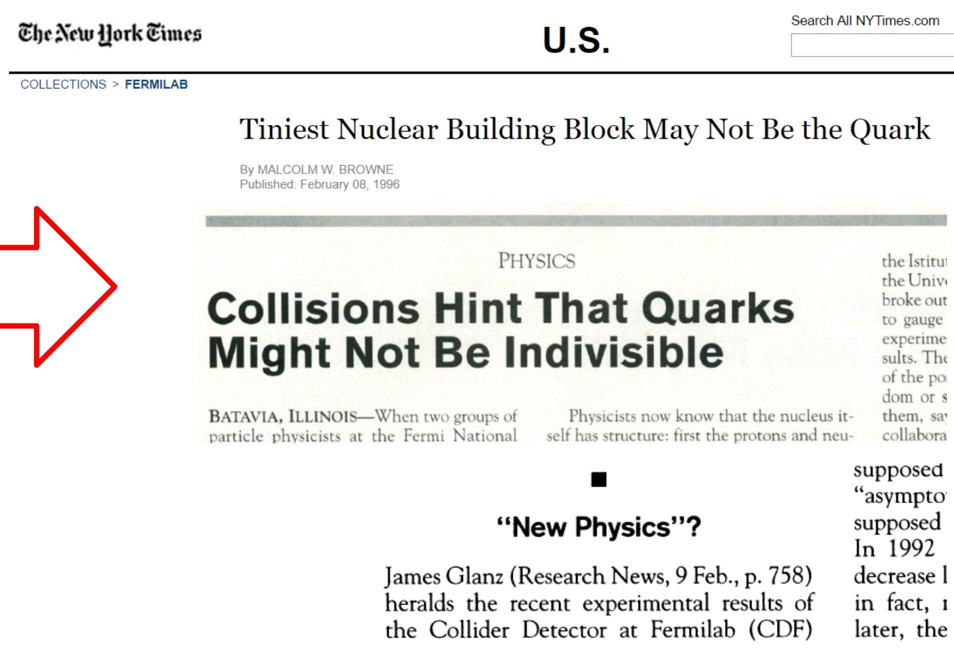
- In 1996 CDF published a jet E_T-differential cross section measurement which appeared to support **quark compositeness**

- That was preceded by endless internal discussions on how to estimate the significance of the effect.
- Estimates went from p=0.01 to significances of over 3-sigma



- A media storm hit the experiment as reporters spun the story evidencing the "New Physics" interpretations

- Soon a theoretical reanalysis showed how it was possible to tweak the "parton distribution functions" in the proton to accommodate the observed effect


The New York Times U.S.

Search All NYTimes.com

COLLECTIONS > FERMILAB

Tiniest Nuclear Building Block May Not Be the Quark

By MALCOLM W. BROWNE
Published: February 08, 1996

PHYSICS

Collisions Hint That Quarks Might Not Be Indivisible

BATAVIA, ILLINOIS—When two groups of particle physicists at the Fermi National

Physicists now know that the nucleus itself has structure: first the protons and neu-

the Istitu
the Univ
broke out
to gauge
experime
sults. The
of the po
dom or s
them, sa
collabora

"New Physics"?

James Glanz (Research News, 9 Feb., p. 758) heralds the recent experimental results of the Collider Detector at Fermilab (CDF)

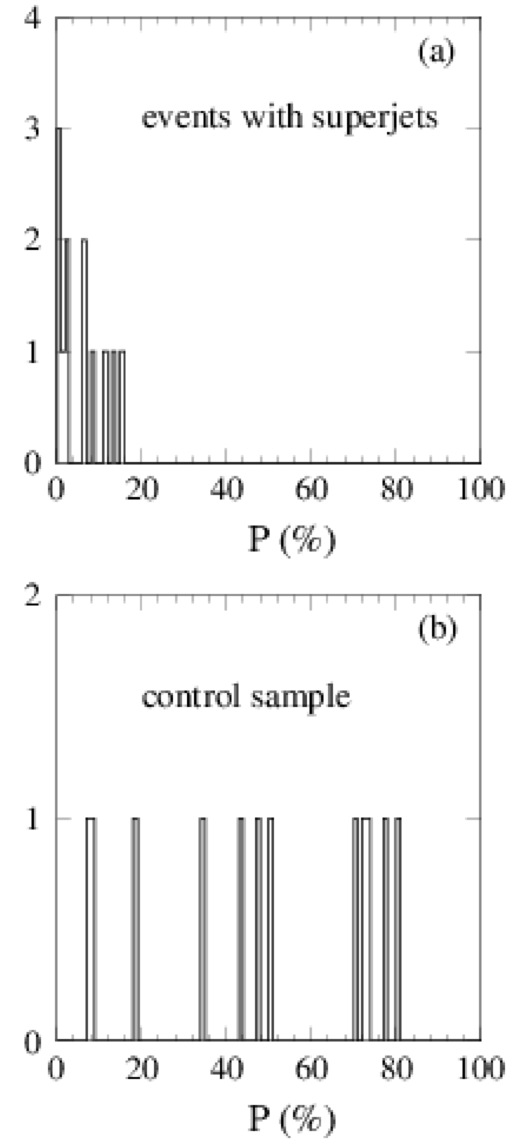
supposed
"asympto
supposed
In 1992
decrease l
in fact, i
later, the

The Superjets

As a spin-off of the top discovery and cross section measurement, in 1998 CDF observed 13 “superjet” events in the W+2,3-jet sample; a 3σ excess from background expectations (4+-1 events) but **weird kinematics in addition**

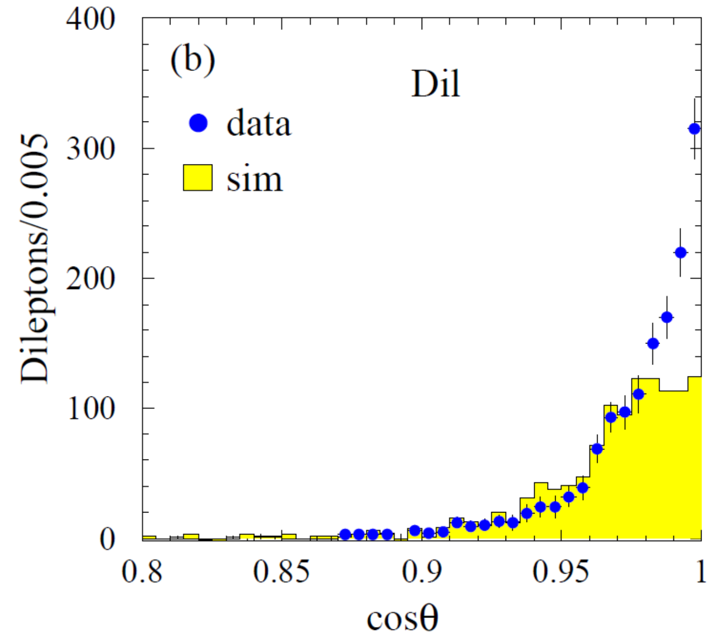
Checking a “complete set” of kinematical variables yielded a combined significance in the 6σ ballpark

The analysis was published [12]only after a fierce, three-year-long fight within the collaboration; no similar excess appeared in the x100 statistics of Run II.



The Sbottom

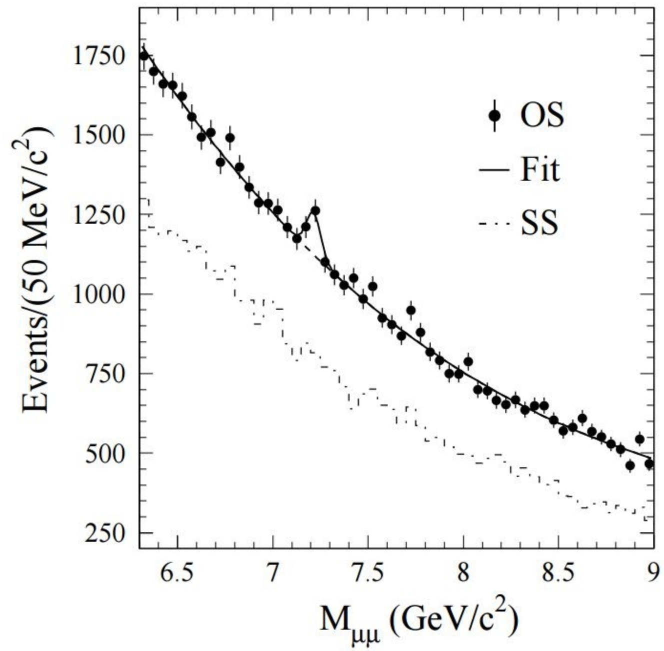
- While the battle over what to do with the 13 superjet events raged, authors found additional anomalies in an orthogonal data sample of inclusive lepton data, which fit a common interpretation: a bottom squark could be causing all effects
- The first effect was a significant excess of events with two or more leptons in dijet events
- The kinematics of same-jet leptons were strikingly different from B decay expectations (right)



- A sbottom quark with mass in the 3.5-4 GeV range could be hypothesized to be a cause of the excess of superjets, with an odd mechanism producing the squark in association with W bosons

- Such a squark would make a spin-0 bound state. It would decay to muon pairs at a smaller rate than vector mesons, but estimates predicted that 250 events could be seen in dimuon data

- Incredibly, a bumplet was seen with the right size and a compatible mass in a third dataset (dimuon-triggered events).. After LEE this was however only a 2.5-sigma effect...



Sbottom Quarks in LEP II Data

ALEPH

ELEP (GeV)	L(pb^{-1})	#exp	# obs
161	11	0.7	1
172	11	0.7	4
183	59	3.6	9
189	174	10.7	19
192	29	1.7	1
196	80	4.4	6
200	86	4.6	8
202	42	2.5	5

- In the summer of 2000 ALEPH researchers were informed of the CDF lepton excess and the sbottom quark interpretation. They looked for dijets with leptons and found a 3-sigma effect in their own data!
- The signal was shown at a and then at a conference i
- DELPHI (see plot, right) firs thrust distribution for the and that no excess was pre
- Later the signal was under artifact of a wrong MC sim miscalibrated electron fake disproven by the other LEF CLEO

Light s-bottom Search

At LEPC on July 20, ALEPH presented a fresh analysis with a possible excess for:

b-jets with leptons: 56 obs. / 33.6 exp. for 580 pb^{-1}

(39 obs. / 23.0 exp. 411 pb^{-1})

- **n-tuple of preliminary analysis contained lepton-id for isolated leptons.**
- **new study using e.g. heavy flavour lepton identification, more adequate for leptons in jets yields no excess**

⇒ 24 obs. / 20. exp. for 411 pb^{-1}

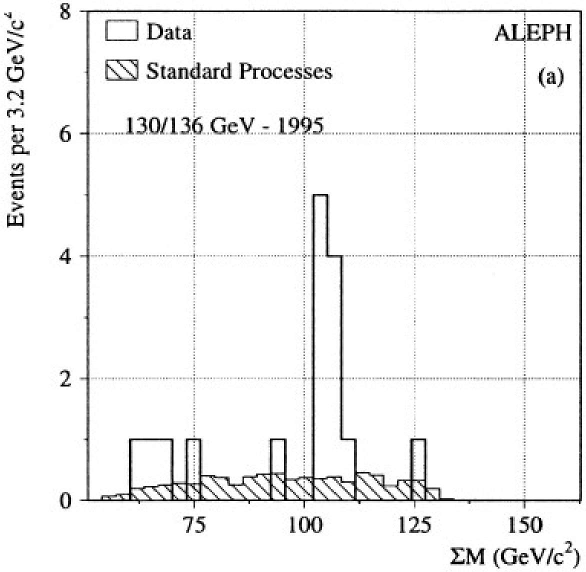
excess is NOT confirmed

Notable Anomalies in Other Experiments

1996 was a prolific year for particle ghosts in the 100-110 GeV region.

ALEPH also observed a 4σ-ish excess of Higgs-like events at 105 GeV in the 4-jet final state of electron-positron collisions at 130-136 GeV. They published the search[13], which found 9 events in a narrow mass region with a background of 0.7, estimating the effect at the 0.01% level

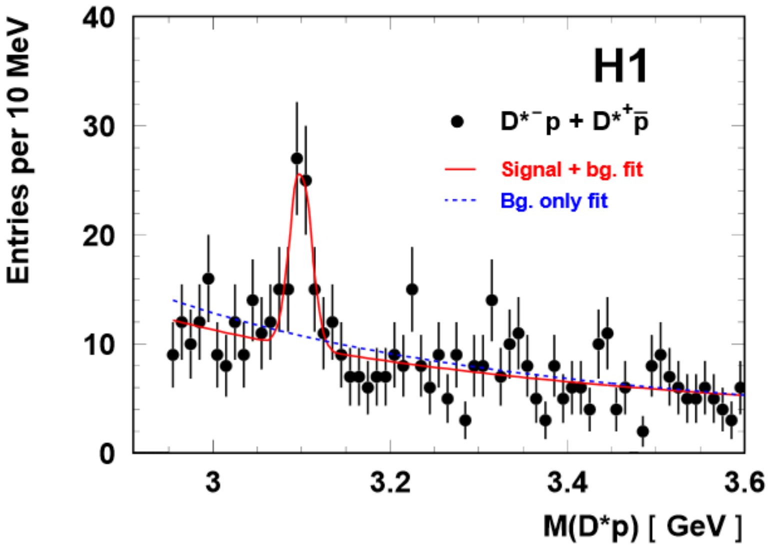
→ later it was understood to be a fluctuation



In 2004 H1 published a pentaquark signal of 6 sigma significance[14]. The prominent peak at 3.1 GeV was indeed suggestive, however it was not confirmed by later searches.

In the paper they write that "From the change in maximum log-likelihood when the full distribution is fitted under the null and signal hypotheses, corresponding to the two curves shown in figure 7, the statistical significance is estimated to **be p=6.2σ**"

Note: H1 worded it "Evidence" in the title ! This was a (wise) departure from blind application of the 5-sigma rule…

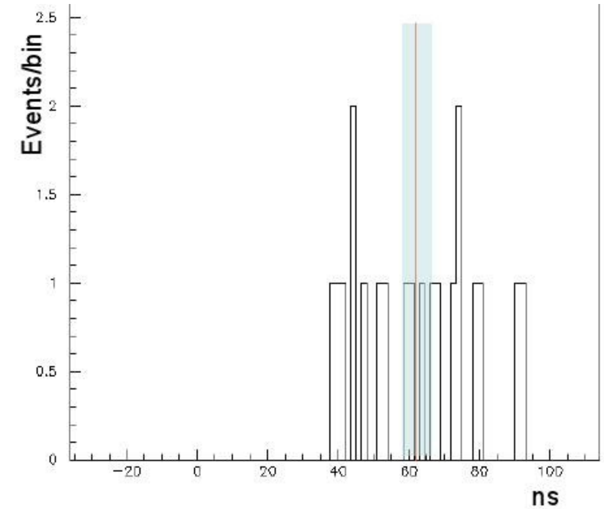


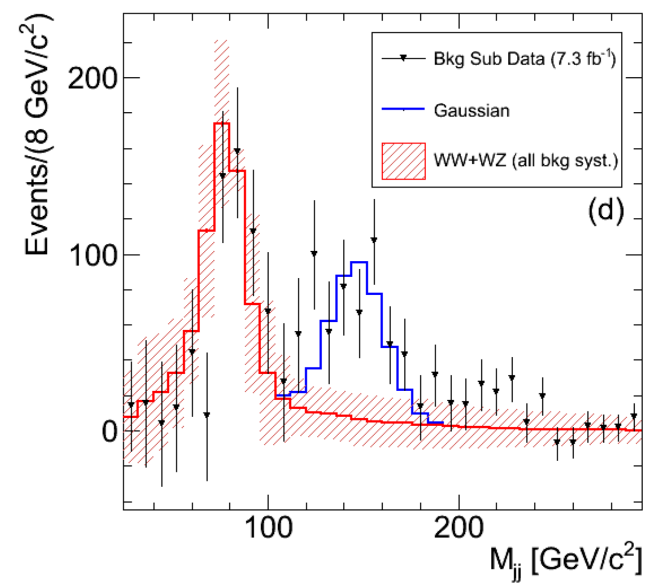
Other Notable Anomalies - 2

A mention has also to be made of a few more recent, striking examples:

- In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared smaller by 6σ than the travel time of light in vacuum[15]. The effect spurred lively debates, media coverage, checks by the nearby ICARUS experiment and dedicated beam runs. It was finally understood to be due to a single large source of systematic uncertainty – a loose cable[16]

- Also in 2011 the CDF collaboration showed a large, 4σ signal at 145 GeV in the dijet mass distribution of proton-antiproton collision events producing an associated leptonic W boson decay[17]. The effect grew with data size and was systematical in nature; indeed it was later understood to be due to the combination of two nasty background contaminations[18].





An Interesting Pattern Emerges…

Claim	Claimed Significance				Verified or Spurious
Top quark evidence					
Top quark observation					
CDF bbγ signal					
CDF eeggMEt event					
CDF superjets					
B_s oscillations					
Single top observation					
HERA pentaquark					
ALEPH 4-jets					
LHC Higgs evidence					
LHC Higgs observation					
OPERA v>c neutrinos					
CDF Wjj bump					

An Interesting Pattern Emerges…

Claim	Claimed Significance				Verified or Spurious
Top quark evidence					
Top quark observation					
CDF bbγ signal		4			False
CDF eeggMEt event				6	False
CDF superjets				6	False
Bs oscillations					
Single top observation					
HERA pentaquark				6	False
ALEPH 4-jets		4			False
LHC Higgs evidence					
LHC Higgs observation					
OPERA v>c neutrinos				6	False
CDF Wjj bump		4			False

An Interesting Pattern Emerges…

Claim	Claimed Significance				Verified or Spurious
Top quark evidence	3				True
Top quark observation			5		True
CDF bbγ signal		4			False
CDF eeggMEt event				6	False
CDF superjets				6	False
Bs oscillations			5		True
Single top observation			5		True
HERA pentaquark				6	False
ALEPH 4-jets		4			False
LHC Higgs evidence	3				True
LHC Higgs observation			5		True
OPERA v>c neutrinos				6	False
CDF Wjj bump		4			False

A Look Into the Look-Elsewhere Effect

- The discussion above clarifies that a compelling reason for enforcing a small test size as a prerequisite for discovery claims is the presence of large trials factors, aka LEE

- The LEE was a concern 50 years ago, but nowadays we have enormously more CPU power. Yet the complexity of our analyses has also grown considerably
  - Take the Higgs discovery: CMS combined dozens of final states with hundreds of nuisance parameters, partly correlated, partly constrained by external datasets, often non-Normal.

    → we still occasionally cannot compute the trials factor by brute force!

  - A further complication is that in reality **the trials factor also depends on the significance of the local fluctuation,** adding dimensionality to the problem.

- A study by E. Gross and O. Vitells[19] demonstrated in 2010 how it is possible to estimate the trials factor in most experimental situations, without resorting to simulations

Trials Factors

In statistics literature the situation in which one speaks of a **trials factor** is the one of a hypothesis test when a nuisance parameter is present only under the alternative hypothesis. The regularity conditions under which Wilks' theorem applies are then **not satisfied.**

Let us consider a particle search when the mass is unknown. We measure masses **x**.

The null hypothesis is that the data follow the background-only model **b(x)**, and the alternative hypothesis is that they follow the model **b(x)+ μs(x|m$_H$)**, with **μ** a signal strength parameter and **m$_H$** the particle's true mass, which here acts as a nuisance parameter only present in the alternative.

μ=0 corresponds to the null, μ>0 to the alternative.

One then defines a test statistic encompassing all possible particle mass values,

$$q_0(\hat{m}_H) = \max_{m_H} q_0(m_H)$$

This is the maximum of the test statistic defined above for the bgr-only, across the many tests performed at the various possible masses being sought. The problem consists in assigning a p-value to the maximum of q(m$_H$) in the entire search range.

One can use an asymptotic "regularity" of the distribution of q to get a global p-value by using the technique of Gross and Vitells.

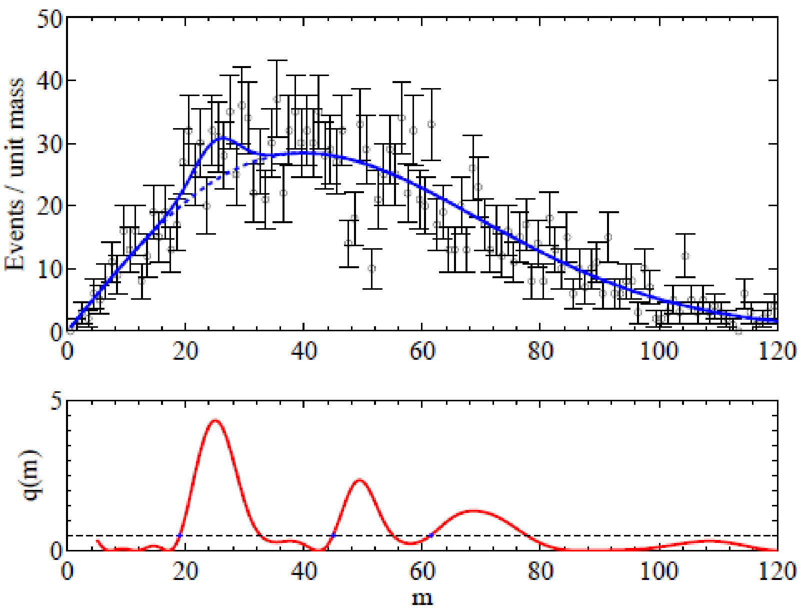
Local Minima and Upcrossings

One counts the number of “upcrossings” of the distribution of the test statistic, as a function of mass. Its wiggling tells how many independent places one has been searching in.

The number of local minima in the fit to a distribution is closely connected to the freedom of the fit to pick signal-like fluctuations in the investigated range

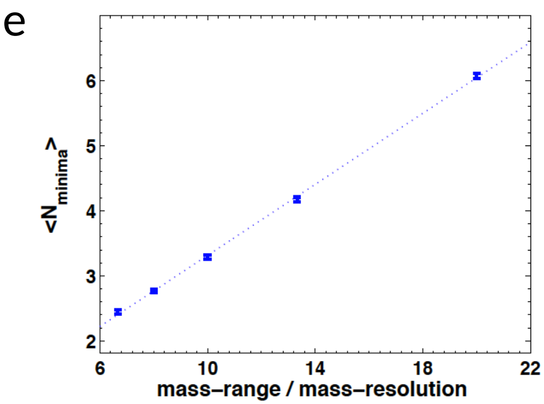
The number of times that the test statistic (below, the likelihood ratio between H_1 and H_0) crosses some reference line can be used to estimate the trials factor. One estimates the global p-value with the number N_0 of upcrossings from a minimal va rmula

$$p_b^{global} = P(q_0(\hat{m}_H) > u) \leq \langle N_u \rangle + \frac{1}{2} P_{\chi_1^2}(u)$$



The number of upcrossings can be best estimated using the data themselves at a low value of significance, as it has been shown that the dependence on Z is a simple negative exponential:

$$\langle N_u \rangle = \langle N_{u_o} \rangle e^{-(u-u_o)/2}$$



Notes About the LEE Estimation

Even if we can usually compute the trials factor by brute force or estimate with asymptotic approximations, there is a degree of uncertainty in how to define it

If I look at a mass histogram and I do not know where I try to fit a bump, I may consider:

1. the location parameter and its freedom to be anywhere in the spectrum
2. the width of the peak: is that really fixed *a priori* ?
3. the fact that I may have tried different selections before settling on the one I actually end up presenting
4. the fact that I may be looking at several possible final states and mass distributions
5. My colleagues in the experiment can be doing similar things with different datasets; should I count that in ?
6. There is ambiguity on the LEE depending who you are (grad student, experiment spokesperson, lab director…)

Also note that Rosenfeld considered the whole world's database of bubble chamber images in deriving a trials factor

The bottomline is that while we can always compute a local significance, it may not always be clear what the true global significance is.

Systematic Uncertainties

- Systematic uncertainties (a.k.a. "nuisance parameters") affect any physical measurement and it is sometimes quite hard to correctly assess their impact.

  Often one sizes up the typical range of variation of an observable due to the imprecise knowledge of a nuisance parameter **at the 1-sigma level**; then one stops there and assumes that the probability density function of the nuisance be Gaussian.

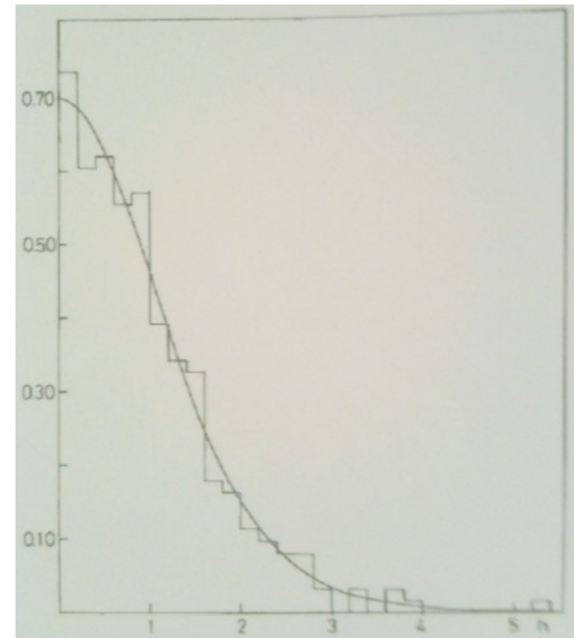
  → if however the PDF has larger tails, it makes the odd large bias much more frequent than estimated

- Indeed, the potential harm of large non-Gaussian tails of systematic effects is one arguable reason for sticking to a 5σ significance level even when the LEE is not a concern. However, the safeguard that the criterion provides to mistaken systematics is not always sufficient.

- One quick example: if a 5σ effect has uncertainty dominated by systematics, and the latter is underestimated by a factor of 2, the 5σ effect is actually a 2.5σ one (a p=0.006 effect): in p-value terms this means that the size of the effect is **overestimated by a factor 20,000**!

A Study of Residuals

The distribution of residuals of 306 measurements in [20]

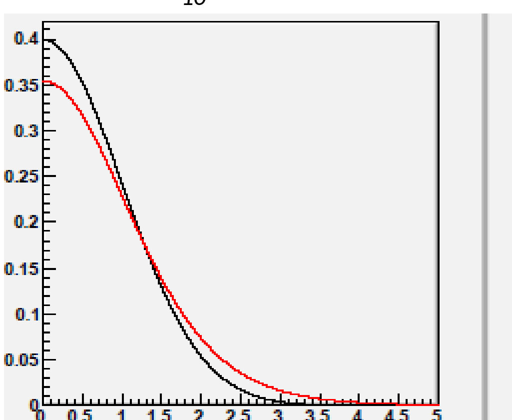


A study of the residuals of particle properties in the RPP in 1975 revealed that they were **not Gaussian in fact**. Matts Roos *et al.* [20] considered residuals in kaon and hyperon mean life and mass measurements, and concluded that these seem to all have a similar shape, well described by a Student distribution $S_{10}(h/1.11)$:

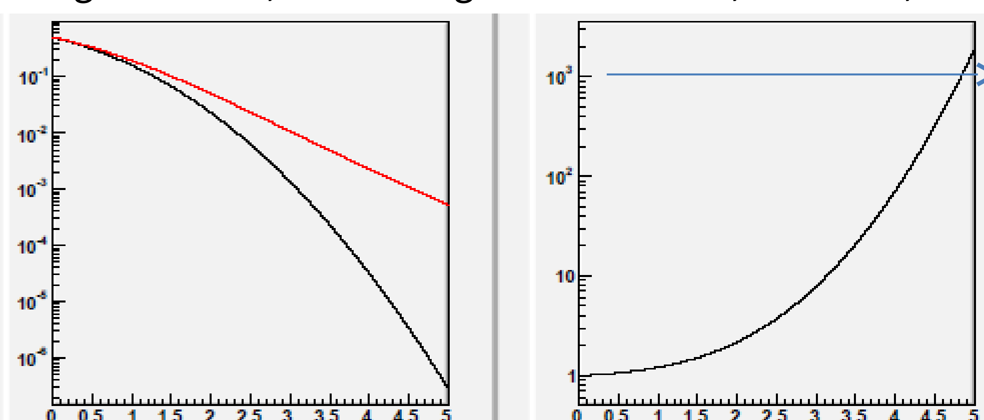
$$S_{10}\left(\frac{x}{1.11}\right)=\frac{315}{256\sqrt{10}}\left(1+\frac{x^2}{12.1}\right)$$

Of course, one should not extrapolate to 5-sigma the behaviour observed by Roos and collaborators in the bulk of the distribution; however, one may consider this as evidence that the uncertainties evaluated in experimental HEP may have a significant non-Gaussian component

Black: a unit Gaussian; red: the $S_{10}(x/1.11)$ function



Left: 1-integral distributions of the two functions. Right: ratio of the 1-integral values as a function of z



x1000!

The "Subconscious Bayes Factor"

Louis Lyons calls this way [21] the ratio of prior probabilities we subconsciously assign to the two hypotheses

When comparing a "background-only" H_0 hypothesis with a "background+signal" one H_1 one often uses the likelihood ratio $\boldsymbol{\lambda = L_1/L_0}$ as a test statistic

– The $p<0.000029\%$ criterion is then applied to the distribution of λ under H_0 to claim a discovery

However, what would be more relevant to the claim would be the ratio of the probabilities:

$$\frac{P(H_1 \mid data)}{P(H_0 \mid data)} = \frac{p(data \mid H_1)}{p(data \mid H_0)} \times \frac{\pi_1}{\pi_0} = \lambda \frac{\pi_1}{\pi_0}$$

where p(data|H) are the likelihoods, and π are the priors of the hypotheses

In that case, if our prior belief in the alternative, π_1, were low, we would still favor the null even with a large evidence λ against it.

- The above is a Bayesian application of Bayes' theorem, while HEP physicists prefer to remain in Frequentist territory. Lyons however notes that "*this type of reasoning does and should play a role in requiring a high standard of evidence before we reject well-established theories: there is sense to the oft-quoted maxim 'extraordinary claims require extraordinary evidence'* ".

The Jeffreys-Lindley Paradox

So what happens if one tries to move to Bayesian territory ?

The issue involves the existence of a null hypothesis, H_0, on which we base a strong belief. In physics **we do believe in our "point null"** – a theory which works for a specific value of a parameter, known with arbitrary accuracy; in other sciences a true "point null" hardly exists

When we compare a point null hypothesis to an alternative which has a **continuous support** for the parameter under test, we need to suitably encode this in a prior belief for the parameter. Bayesians speak of a "probability mass" at $\theta=\theta_0$.

The use of probability masses in priors in a simple-vs-composite test throws a monkey wrench in the Bayesian paradigm, as it can be proven that no matter how large and precise is the data, Bayesian inference **strongly depends** on the scale over which the prior is non-null – that is, on the **prior belief** of the experimenter.

The Jeffreys-Lindley paradox [22] is that frequentists and Bayesians draw **opposite conclusions** on large data when comparing a point null to a composite alternative.

let us give it a look.

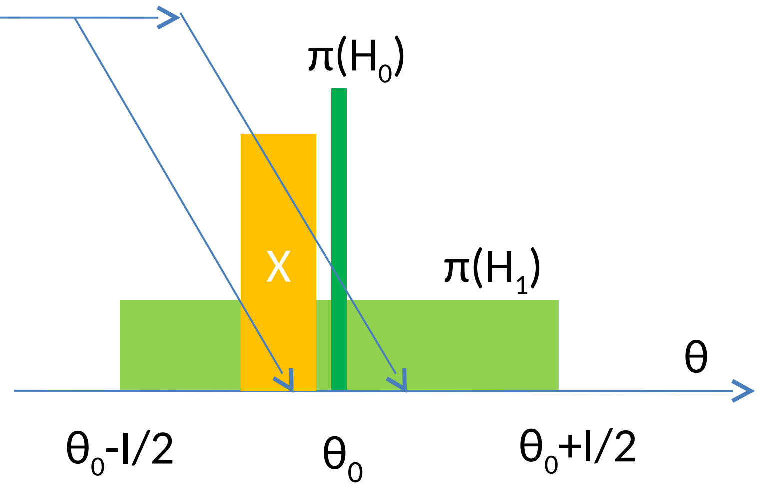
The Paradox

Take $X_1 \ldots X_n$ i.i.d. as $X_i|\theta \sim N(\theta,\sigma^2)$, and a prior belief on θ constituted by a mixture of a point mass **p** at θ_0 and **(1-p)** uniformly distributed in $[\theta_0-l/2, \theta_0+l/2]$.

In classical hypothesis testing the "critical values" of the sample mean delimiting the rejection region of **H_0: θ = θ_0** in favor of **H_1: θ <> θ_0** at significance level α are

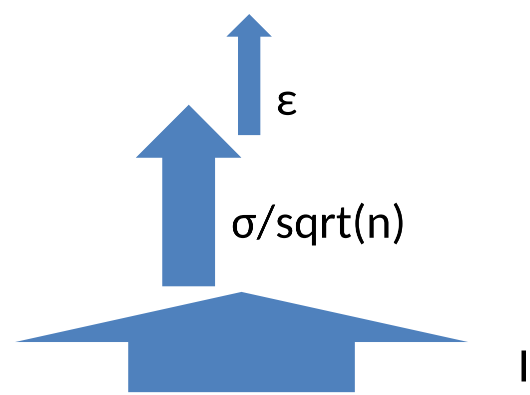
$$\bar{X} = \theta_0 \pm (\sigma/\sqrt{n}) z_{\alpha/2}$$

where $z_{\alpha/2}$ is the significance corresponding to test size α for a two-tailed normal distribution



The **paradox** is that it can be proven that the **posterior probability that H_0 is true** conditional on seeing data in the *critical region* (i.e. ones which exclude H_0 in a classical α-sized test) **approaches 1 as the sample size becomes arbitrarily large**.

As evidenced by Bob Cousins[23], the paradox arises if there are three independent scales in the problem, **ε << σ/sqrt(n) << l**, i.e. the width of the point mass, the measurement uncertainty, and the scale **l** of the prior for the alternative hypothesis



This is a common situation in HEP!!

JLP Example: Charge Bias of a Tracker

- Imagine you want to investigate whether your tracker has a bias in reconstructing positive versus negative curvature. Say we work with a zero-charge initial state at a lepton collider (e^+e^-). You take a unbiased set of collisions, and count how many positive and negative curvature tracks you have reconstructed, say, in a set of n=1,000,000.
- You get n^+=498,800, n^-=501,200. You want to test the hypothesis that R=0.5 with a size **α=0.05**.
- Bayesians will **need a prior** to make a statistical inference: their typical choice would be to **assign equal probability** to the chance that R=0.5 and to it being different (R<>0.5): a "*point mass*" of p=1/2 at R=0.5, and a uniform distribution of the remaining p=1/2 in [0,1]
- We are in high-statistics regime and away from 0 or 1, so Gaussian approximation holds for the Binomial. The probability to observe a number of positive tracks n^+ can then be written, with $x=n^+/n$, as $N(x,\sigma)$ with $\sigma^2=x(1-x)/n$.

  The posterior probability that R=0.5 is then

$$P(R=\frac{1}{2}\,|\,x,n) \approx \frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \Bigg/ \left[\frac{1}{2}\frac{e^{-\frac{(x-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \frac{1}{2}\int_0^1 \frac{e^{-\frac{(x-R)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}\,dR\right] = 0.97816$$

from which a Bayesian concludes that there is **no evidence against R=0.5,** and actually the data strongly supports the null hypothesis (P>>α)

JLP Charge Bias: Frequentist Solution

Frequentists will not need a prior, and just ask themselves how often a result "at least as extreme" as the one observed arises by chance, if the underlying distribution is N(R,σ) with R=1/2 and $\sigma^2=x(1-x)/n$ as before.

One then has

$$P(x \leq 0.4988 \mid R=\frac{1}{2}) = \int_0^{0.4988} \frac{e^{-\frac{(t-\frac{1}{2})^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dt = 0.008197$$

$$\Rightarrow P'(x \mid R=\frac{1}{2}) = 2 * P = 0.01639$$

(we multiplied by two since we would be just as surprised to observe an excess of positives as a deficit).

From this, frequentists conclude that the tracker is biased, since there is a less-than 5% probability, P'<α, that a result as the one observed could arise by chance!

A frequentist thus draws the **opposite conclusion** of a Bayesian from the same data !

Notes on the JL Paradox

- Bayesians have used the JLP to criticize the way inference is drawn by frequentists:
  - Jeffreys: "*What the use of [the p-value] implies, therefore, is that a hypothesis that may be true may be rejected because it has **not** predicted observable results that have **not** occurred*" **[24]**

- Unfortunately, the Bayesian approach offers no clear substitute to the Frequentist p-value for reporting experimental results
  - Bayes factors, which describe by how much prior odds are modified by the data, are **not factorizing out the subjectivity** of the prior belief when the JLP applies: even asymptotically, **they retain a dependence on the scale of the prior of H_1.**

- In their debates on the JL paradox, Bayesian statisticians have blamed the concept of a "point mass", as well as suggested n-dependent priors. There is a large body of literature on the subject
  - As the source of the problem is assigning to the null hypothesis a non-zero prior, statisticians tend to argue that "the precise null" is never true.

However, we do believe our point nulls in HEP and astro-HEP!!

In summary, the issue is an active research topic and is not resolved. I have brought it up here to show how the trouble of defining a test size α in classical hypothesis testing is not automatically solved by moving to Bayesian territory

So What to Do With 5σ ?

To summarize the points made so far:

- the LEE can be estimated analytically as well as computationally; experiments in fact now routinely produce "global" and "local" p-values and Z-values
  - What is then the point of protecting from large LEE ?
  - Sometimes the trials factor is 1 and sometimes it is enormous; a one-size-fits-all is hardly justified – **it is illogical to penalize an experiment for the LEE of others**

- the impact of systematic uncertainties varies widely from case to case; *i.e.* sometimes one has control samples (*e.g.* particle searches), sometimes one does not (*e.g.* OPERA's neutrinos speed measurement)
- The cost of a wrong claim, as image damage or backfiring of media hype, can vary dramatically
- Some claims are intrinsically less likely to be true, hence we have a subconscious Bayes factor at work.

So why a fixed discovery threshold ?

- One may take the attitude that any claim is subject to criticism and independent verification, and the latter is always more rigorous when the claim is steeper; and it is good to just have a "**reference value**" for the level of significance of the data – a «tradition», a useful standard

Lyons' Table

My longtime CDF and CMS colleague Louis Lyons considered several known searches in HEP and astro-HEP, and produced a table where for each effect he listed several "inputs":

1. the degree of surprise of the potential discovery
2. the impact for the progress of science
3. the size of the trials factor at work in the search
4. the potential impact of unknown or ill-quantifiable systematics

He could then derive a "reasonable" significance level that would account for the different factors at work, for each considered physics effect [21]

The approach is of course only meant to provoke a discussion, and the numbers in the table entirely debatable. The message is however clear:

we should beware of a "one-size-fits-all" standard.

I have slightly modified his original table to reflect my personal bias

Table of Searches for New Phenomena and "Reasonable" Significance Levels

Search	Surprise level	Impact	LEE	Systematics	Z-level
Neutrino osc.	Medium	High	Medium	Low	4
Bs oscillations	Low	Medium	Medium	Low	4
Single top	Absent	Low	Absent	Low	3
$B_s \rightarrow \mu\mu$	Absent	Medium	Absent	Medium	3
Higgs search	Medium	Very high	Medium	Medium	5
SUSY searches	High	Very high	Very high	Medium	7
Pentaquark	High	High	High	Medium	7
G-2 anomaly	High	High	Absent	High	5
H spin >0	High	High	Absent	Low	4
4th gen fermions	High	High	High	Low	6
V>c neutrinos	Huge	Huge	Absent	Very high	THTQ
Direct DM search	Medium	High	Medium	High	5
Dark energy	High	Very high	Medium	High	6
750 GeV boson	High	High	High	Low	6
Grav. waves	Low	High	Huge	High	7

Conclusions

- 48 years after the first suggestion of a 5-sigma threshold for discovery claims, and 22 years after the start of its consistent application, the criterion appears inadequate
  - It does not protect from steep claims that later peter out
  - It delays acceptance of uncontroversial finds
  - It is arbitrary and illogical in many aspects
- Bayesian hypothesis testing does not offer a robust replacement, due to hard-to-circumvent prior dependence of conclusions
- A single number never summarizes the situation of a measurement
  - experiments have started to publish their likelihoods, so combinations and interpretation get easier
- My suggestion is that for each considered (relevant) search the community should seek a consensus on what could be an acceptable significance level for a media-hitting claim
- For searches of unknown effects and fishing expeditions, the **global** p-value is the only real weapon – but in most cases the trials factor is hard to quantify
- Probably 5-sigma are insufficient for unpredicted effects, as large experiments look at thousands of distributions, multiple times, and the experiment-wide trials factor is extremely high

Expect some spurious 5-sigma effect from the LHC soon!

Thank you for your attention!

References

[1] A. H. Rosenfeld, “*Are there any far-out mesons and baryons?*,” In: C.Baltay, AH Rosenfeld (eds) Meson Spectroscopy: A collection of articles, W.A. Benjamin, New York, p.455-483.

[2] D. C. Hom et al., “*Observation of High-Mass Dilepton Pairs in Hadron Collisions at 400 GeV*”, Phys. Rev. Lett. 36, 21 (1976) 1236

[3] S. W. Herb et al., “*Observation of a Dimuon Resonance at 9.5-GeV in 400-GeV Proton-Nucleus Collisions*”, Phys. Rev. Lett 39 (1977) 252.

[4] G. Arnison et al., “Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at sqrt(s)=540 GeV, Phys. Lett. 122B, 1 (1983) 103.

[5] G. Arnison et al., “Experimental Observation of Lepton Pairs of Invariant Mass Around 95 GeV/c2 at the CERN SpS Collider”, Phys. Lett. 126B, 5 (1983) 398.

[6] F. Abe et al., “Evidence for Top Quark Production in p anti-p Collisions at s**(1/2) = 1.8 TeV”, Phys. Rev. D50 (1994) 2966.

[7] F. Abe et al., “*Observation of Top Quark Production in p anti-p Collisions with the Collider Detector at Fermilab*”, Phys. Rev. Lett. 74 (1995) 2626; S. Abachi et al., “*Observation of the Top Quark*”, Phys. Rev. Lett. 74 (1995) 2632.

[8] V.M. Abazov et al., “Observation of Single Top-Quark Production”, Phys. Rev. Lett. 103 (2009) 092001; T. Aaltonen et al., “Observation of Electroweak Single Top Quark Production”, Phys. Rev. Lett. 103 (2009) 092002.

[9] J. Incandela and F. Gianotti, “Latest update in the search for the Higgs boson”, public seminar at CERN. Video: http://cds.cern.ch/record/1459565; slides: http://indico.cern.ch/conferenceDisplay.py?confId=197461.

[10] S. Park, “Searches for New Phenomena in CDF: Z’, W’ and leptoquarks”, Fermilab-Conf-95/155-E, July 1995.

[11] J. Berryhill et al., “Search for new physics in events with a photon, b-tag, and missing Et”, CDF/ANAL/EXOTIC/CDFR/3572, May 17th 1996.

[12] D. Acosta et al., “*Study of the Heavy Flavor Content of Jets Produced in Association with W Bosons in p anti-p Collisions at s**(1/2) = 1.8 TeV*”, Phys. Rev. D65, (2002) 052007.

[13] D. Buskulic et al., “*Four-jet final state production in e^e collisions at centre-of-mass energies of 130 and 136 GeV*”, Z. Phys. C 71 (1996) 179.

[14] A. Aktas et al., “*Evidence for a narrow anti-charm baryon state*”, Phys. Lett. B588 (2004) 17.

[15] T. Adam et al., “*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam*”, JHEP 10 (2012) 093.

[16] T. Adam et al., “*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam using the 2012 dedicated data*”, JHEP 01 (2013) 153.

[17] T. Aaltonen et al., “*Invariant Mass Distribution of Jet Pairs Produced in Association with a W Boson in p anti-p Collisions at s**(1/2) =1.96 TeV*”, Phys. Rev. Lett. 106 (2011) 71801.

[18] T. Aaltonen et al., “*Invariant-mass distribution of jet pairs produced in association with a W boson in p pbar collisions at sqrt(s) = 1.96 TeV using the full CDF Run II data set*”, Phys. Rev. D 89 (2014) 092001.

[19] E. Gross and O. Vitells, “*Trials factors for the Look-Elsewhere Effect in High-Energy Physics*”, arxiv:1005.1891v3, Oct 7th 2010

[20] M. Roos, M. Hietanen, and M.Luoma, “*A new procedure for averaging particle properties*”, Phys.Fenn. 10:21, 1975

[21] L. Lyons, “*Discovering the significance of 5σ*”, arxiv:1310.1284v1, Oct 4th 2013

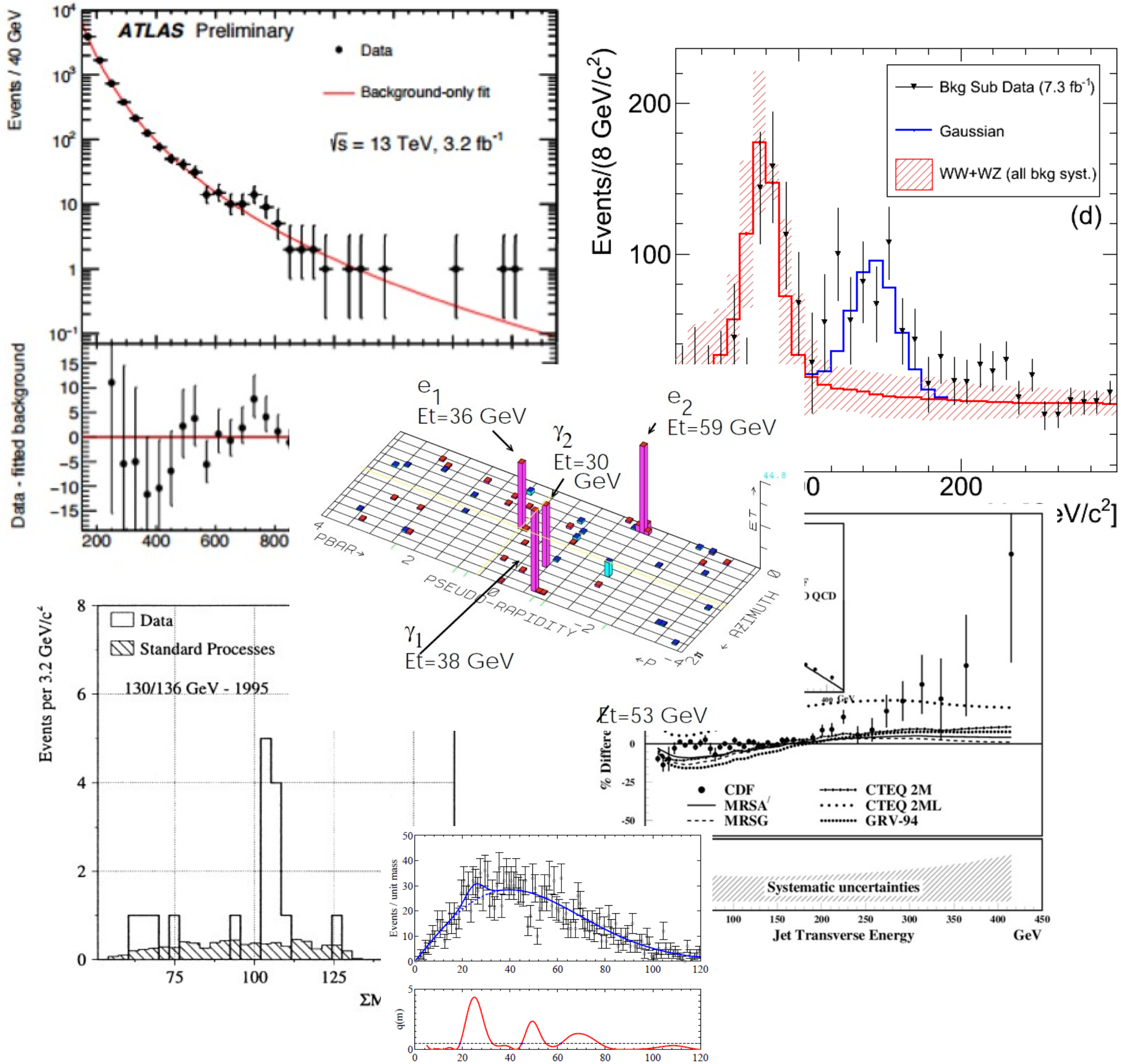
[22] D.V. Lindley, ”*A statistical paradox*”, *Biometrika, 44 (1957) 187-192.*

[23] R. D. Cousins, “*The Jeffreys-Lindley Paradox and Discovery Criteria in High-Energy Physics*”, arxiv:1310.3791v4, June 28th 2014, to appear in a special issue of Synthese on the Higgs boson

[24] H. Jeffreys, “*Theory of Probability*”, 3rd edition Oxford University Press, Oxford, p.385.

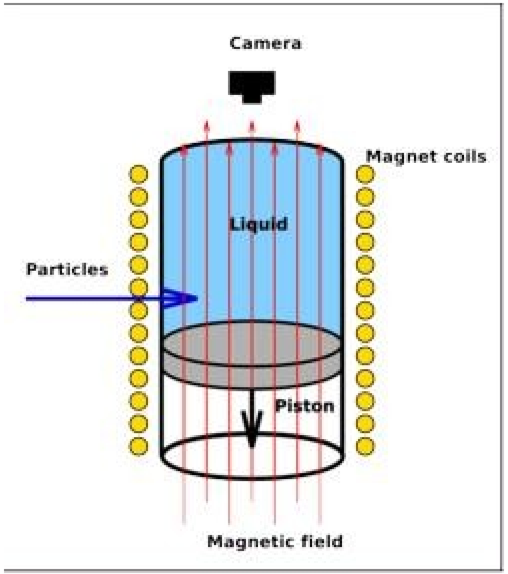
[25] G. K. Karagiannidis and A. S. Lioumpas, A. S., “*An improved approximation for the Gaussian Q-function*.” Communications Letters, IEEE, 11(8), (2007), 644

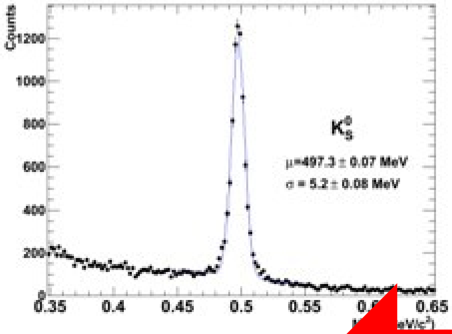
Backup slides

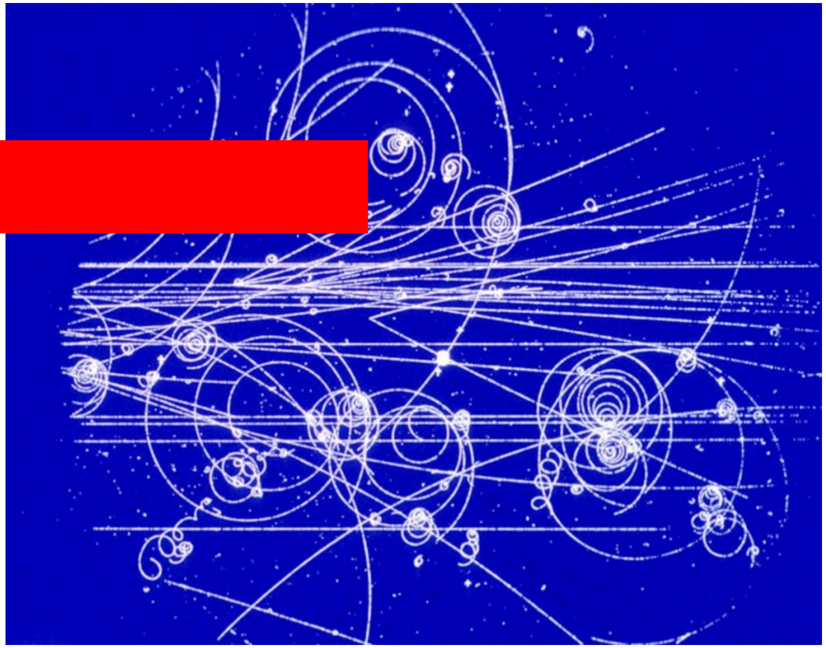

ATLAS Preliminary
Data
Background-only fit
√s = 13 TeV, 3.2 fb⁻¹
Events / 40 GeV
Data - fitted background
Events/(8 GeV/c²)
Bkg Sub Data (7.3 fb⁻¹)
Gaussian
WW+WZ (all bkg syst.)
(d)
e1
Et=36 GeV
e2
Et=59 GeV
γ2
Et=30 GeV
γ1
Et=38 GeV
Et=53 GeV
PBAR
PSEUDO-RAPIDITY
AZIMUTH
ET
Data
Standard Processes
130/136 GeV - 1995
Events per 3.2 GeV/c²
% Difference
CDF
MRSA′
MRSG
CTEQ 2M
CTEQ 2ML
GRV-94
Systematic uncertainties
Jet Transverse Energy
GeV
Events / unit mass
q(m)
m

Bubble Chamber Physics

A bubble chamber is a vessel filled with a gas in a phase of superheating. The passage of charged particles ionizes the gas and bubbles are formed along the path







By measuring the tracks in a magnetic field, one determines their momentum. The mass of a particle decaying into others can be determined from the daughters' momenta

The Standard Model

A misnomer – it is not a model but a full-blown theory which allows us to compute the result of subatomic processes with high precision

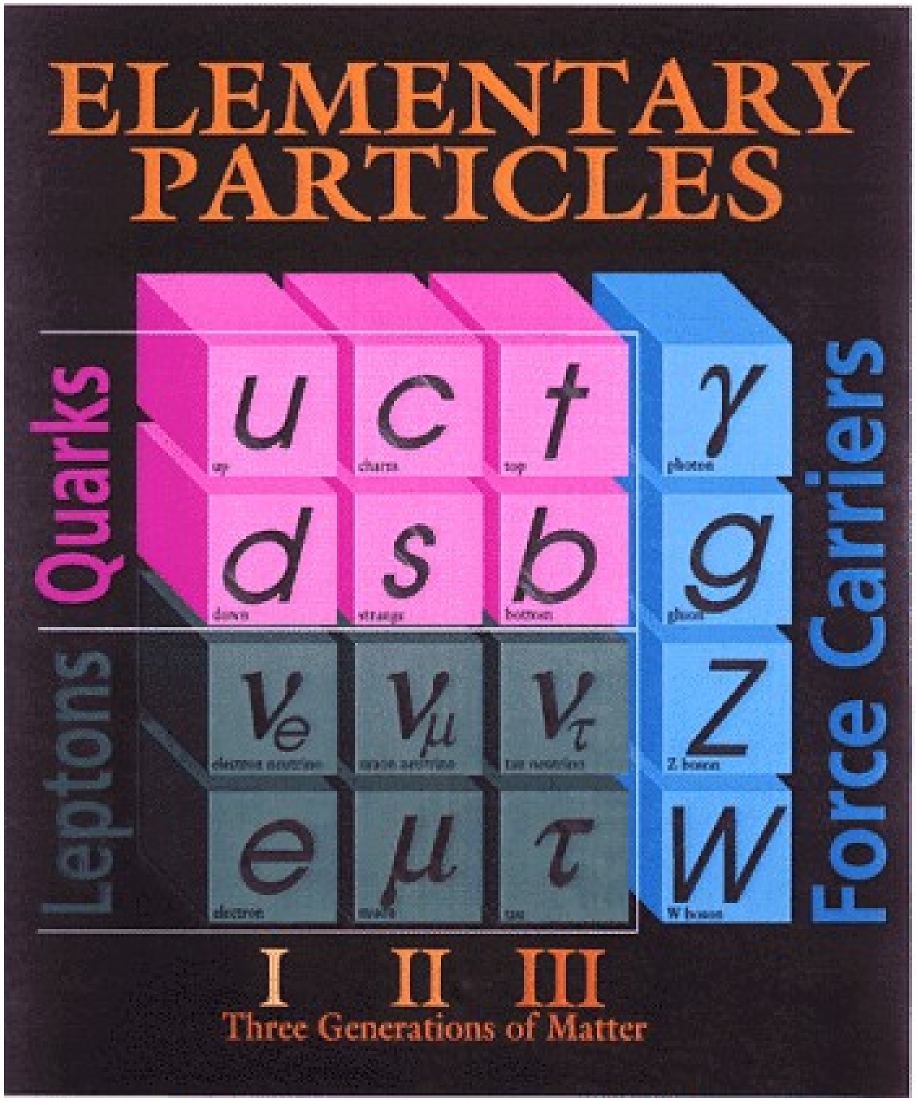
Three families of quarks, and three families of leptons, are the matter constituents

Strong interactions between quarks are mediated by 8 gluons, g

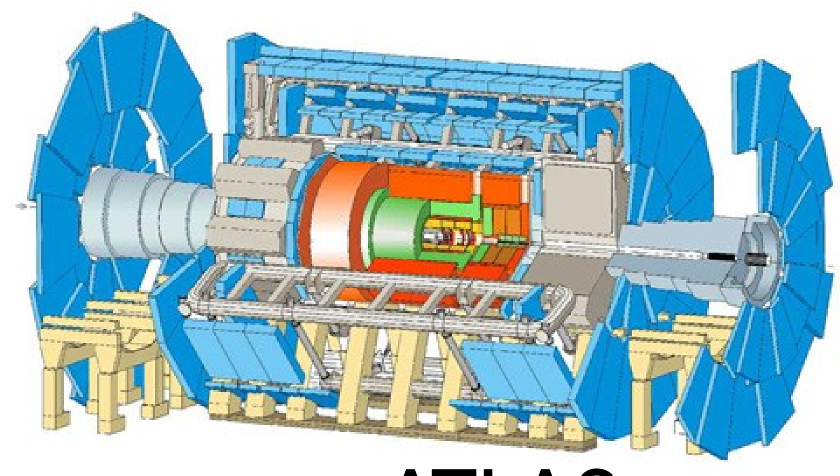
Electromagnetic interactions between charged particles are mediated by the photon, γ

The weak force is mediated by W and Z

<u>*Gravity is not included in the model*</u>



The LHC



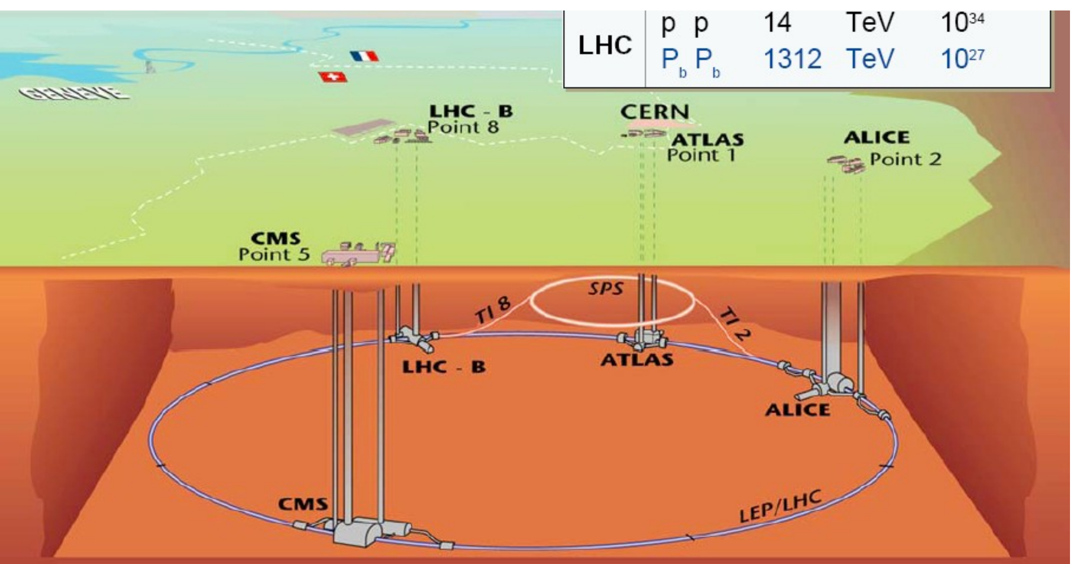
ATLAS

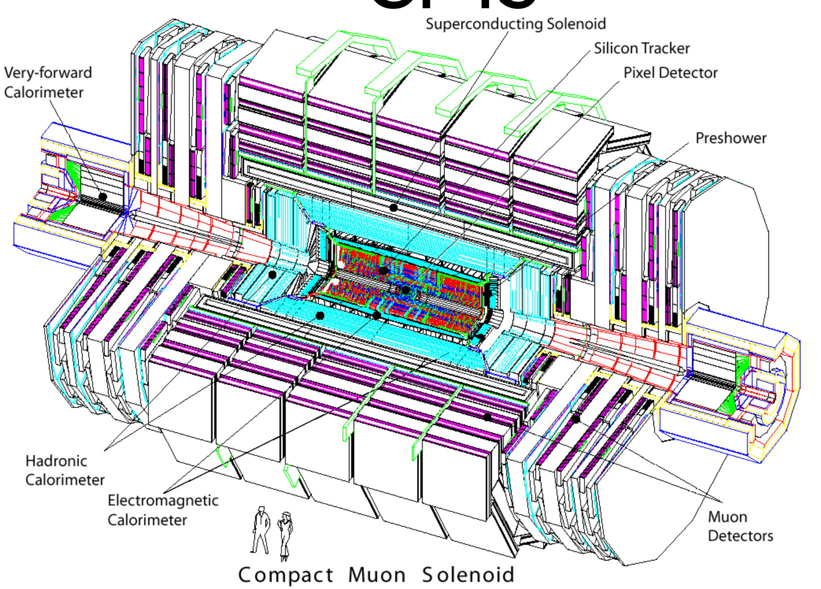
LHC is the largest and most powerful particle accelerator, built to investigate matter at the shortest distances

It resides in a 27km long tunnel 100 meters underground near Geneva

Collisions between protons are created where the beams intersect: the caverns are equipped with huge detectors. Two of these are multi-purpose «electronic eyes» that try to detect everything that comes out of the collision

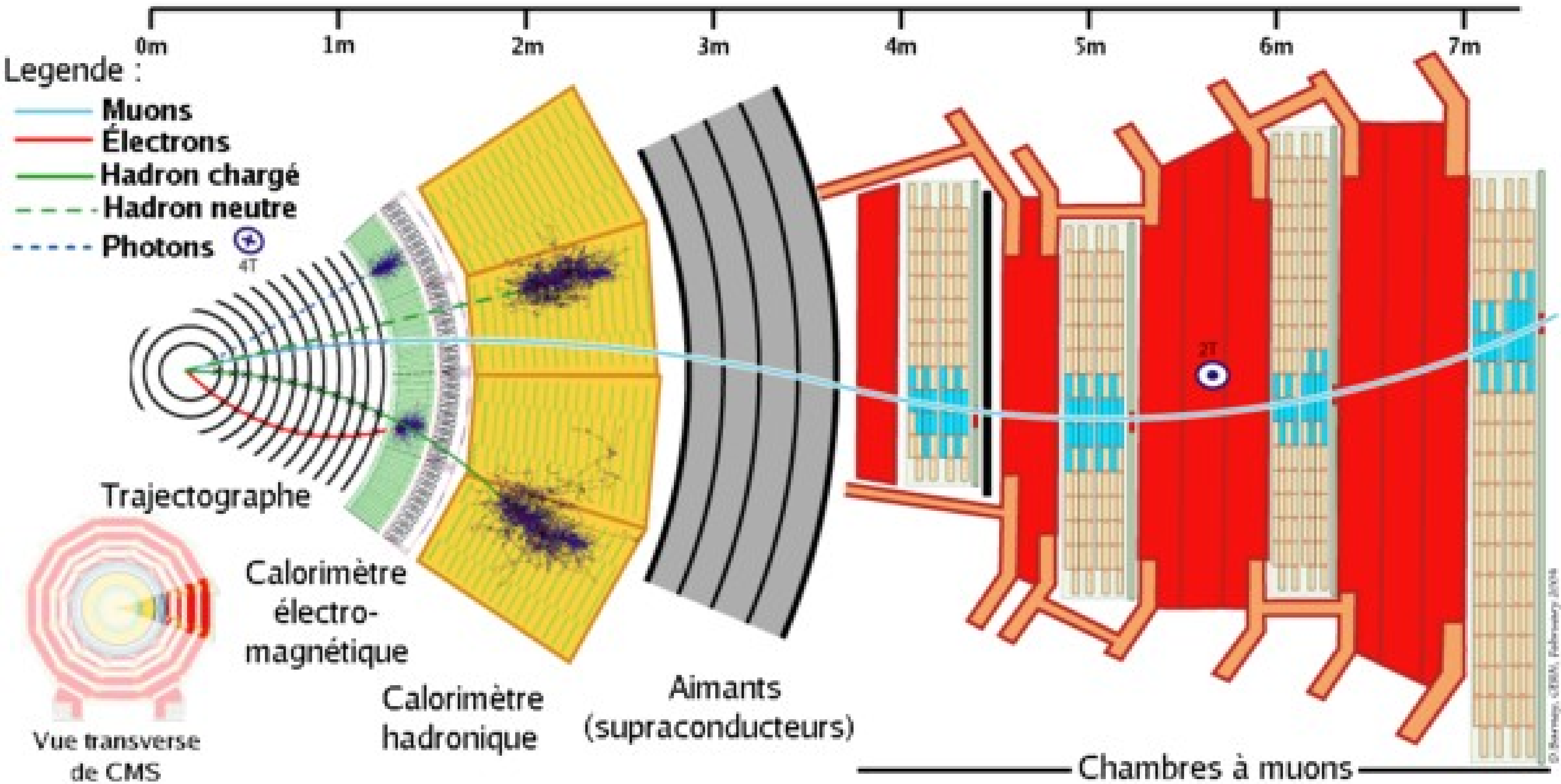
CMS





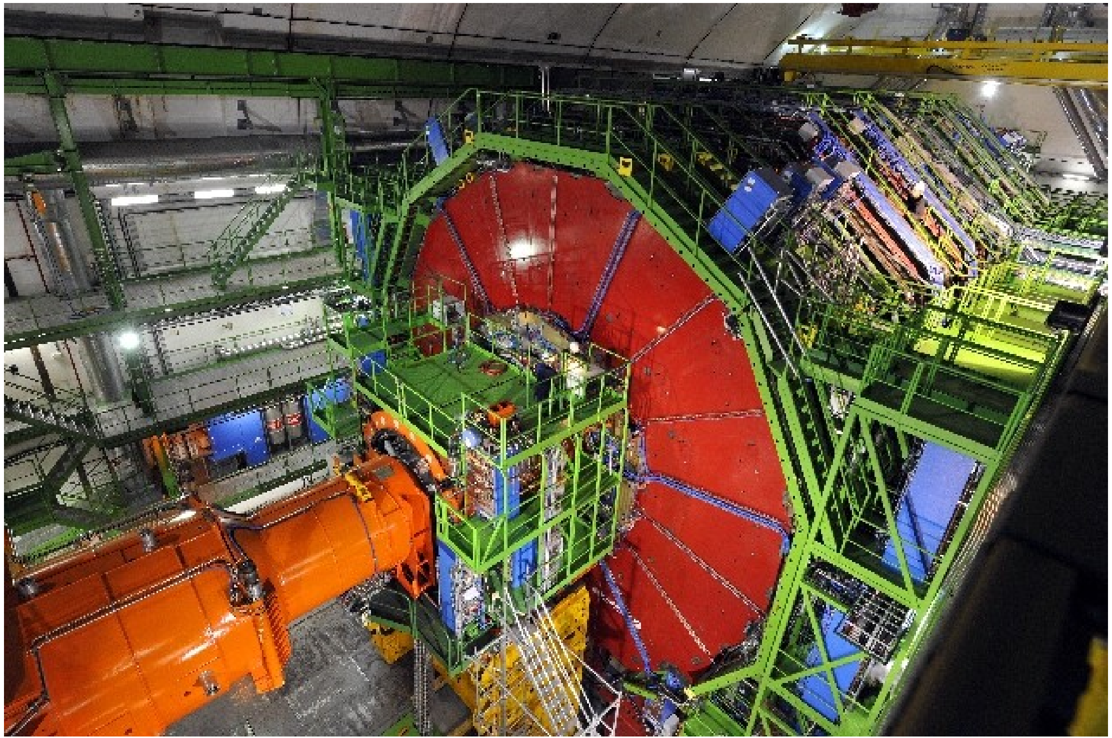
How we detect particles

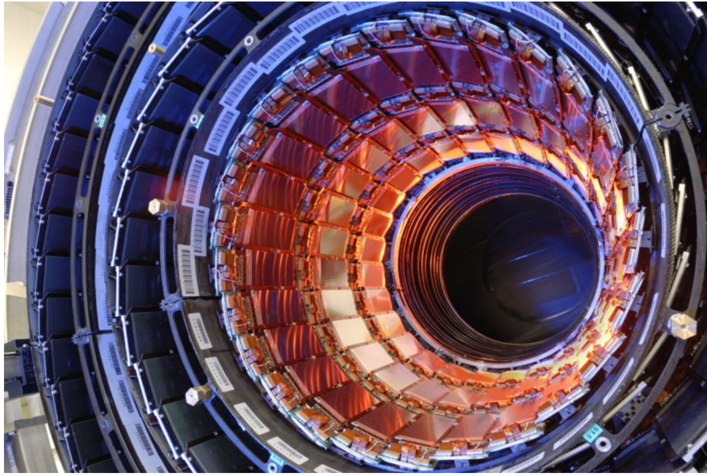
Charged particles are tracked in the inner section, through the ionization they leave on silicon; a powerful magnet bends their trajectories, allowing a measurement of their momentum
Then calorimeters destroy both charged and neutral ones, measuring their energy
Muons are the only particles that can traverse the dense material and get tracked outside

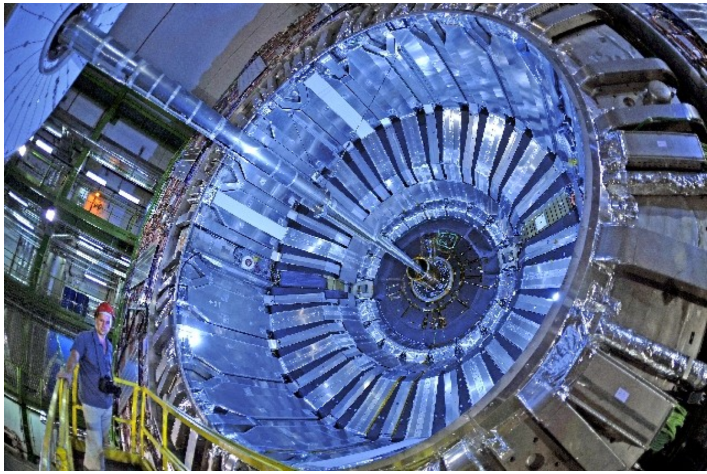


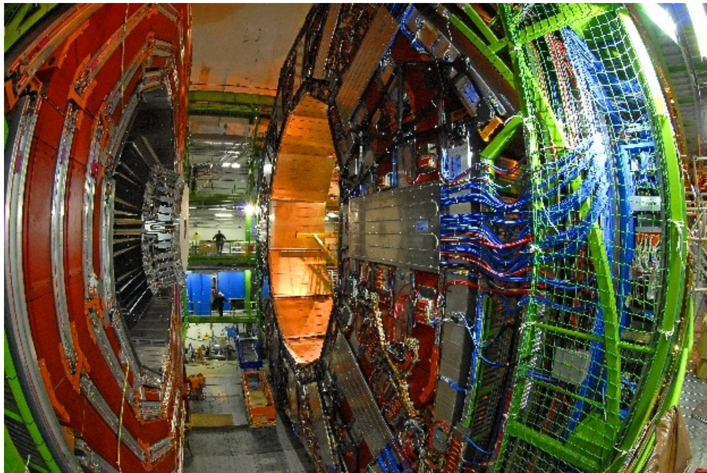
CMS

- CMS (Compact Muon Solenoid) was built with the specific goal of finding the Higgs boson
- Along with ATLAS, it is arguably the most complex machine ever built by mankind
- Hundreds of millions collisions take place every second in its core, and each produces signals in hundreds of millions of electronic channels. These data are read out in real time and stored for offline analysis



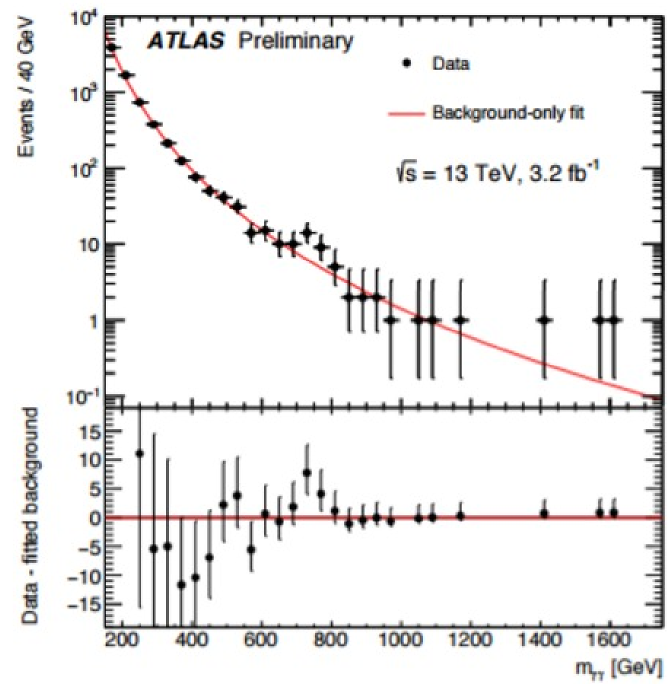


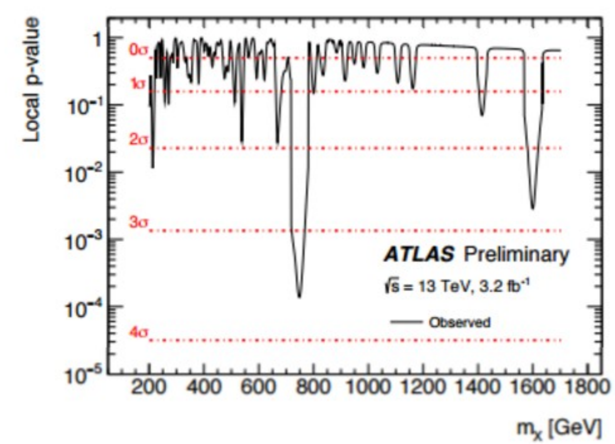




The Case Of The Photon Pairs

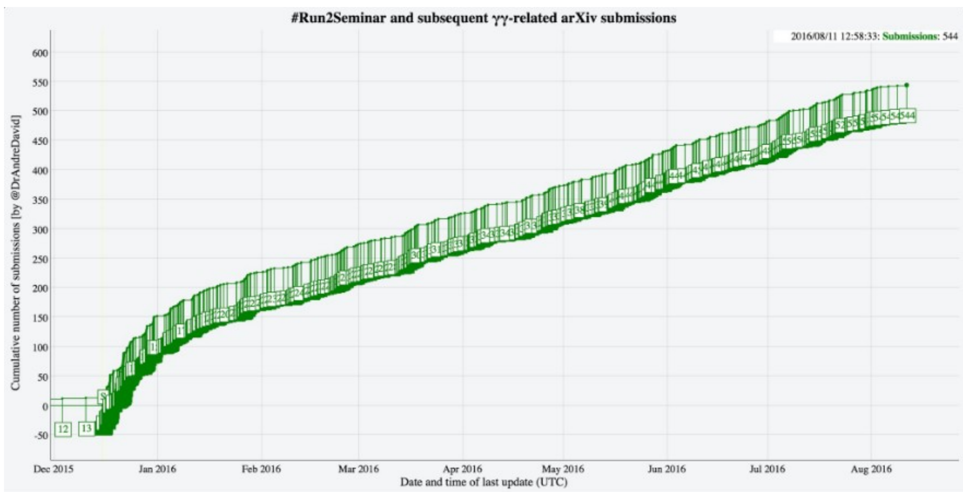
- Last December, ATLAS and CMS announced evidence for a 750 GeV particle decaying to photon pairs
  - Significance in the 4-sigma ballpark
    - ATLAS 3.6σ alone, CMS 2-sigmaish evidence
    - Conflicting evidence on width
  - Theorists jumped at it, proposing interesting and less interesting scenarios to fit it in
  - Experiments set out to search for it in other ways and with additional data





The pheno feeding frenzy

In the matter of 8 months the Cornell arxiv got flooded with over **550 new papers** that tried to explain the diphoton excesses of ATLAS and CMS



Bets were offered and accepted on the nature of the new particle, with various odds

In the process, we learned that finding new physics will not teach us much per se – one needs to then characterize it quite well to sort out what underlying theory can be responsible for it!

Some of the proposed explanations:

Two higgs doublets
Seesaw vectorlike fermions
Closed strings
Neutrino-catalyzed
Indirect signature of DM
Colorful resonances
Resonant sneutrino
SU(5) GUT
Inert scalar multiplet
Trinification
Dark left-right model
Vector leptoquarks
D3-brane
Deflected-anomaly SUSY breaking
Radion candidate
Squarkonium-Diquarkonium
R-parity violating SUSY
Gravitons in multi-warped scenario

750-GeV Bump Interpretation Summary

1 - It seems quicker to say what a 750 GeV bump cannot be:

Not the Lochness monster, which has an evident 3-bump structure

Not Mickey Mouse, who clearly has a non-Gaussian tail

2 – The signal clearly inspired the creativity of theorists: best title In arXiv paper for a while -

"**How the gamma-gamma Resonance Stole Christmas**"

J.L. Paradox: Proof

$$P \quad (H_0|\bar{X} = \bar{x} = \theta_0 + (\sigma/\sqrt{n}) z_{\alpha/2}) \quad = \quad \frac{P(H_0)P(data|H_0)}{P(H_0)P(data|H_0)) + P(H_A)P(data|H_A)}$$

$$= \frac{p \, \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \, e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta_0)]^2\}}}{p \, \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \, e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta_0)]^2\}} + (1-p) \int_{\theta_0 - I/2}^{\theta_0 + I/2} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \, e^{\{(-1/2)[(\sqrt{n}/\sigma)(\bar{x}-\theta)]^2\}} \, \frac{1}{I} \, d\theta}$$

$$= \frac{p \, e^{\{-(1/2) z^2_{\alpha/2}\}}}{p \, e^{\{-(1/2) z^2_{\alpha/2}\}} + \frac{(1-p)}{I} \int_{\theta_0 - I/2}^{\theta_0 + I/2} e^{\{(-1/2)[(\sqrt{n}/\sigma)(\theta-\bar{x})]^2\}} d\theta}$$

$$\geq \frac{p \, e^{\{-(1/2) z^2_{\alpha/2}\}}}{p \, e^{\{-(1/2) z^2_{\alpha/2}\}} + \frac{(1-p)}{I} \frac{\sqrt{2\pi}\sigma}{\sqrt{n}}} \quad \rightarrow \quad 1 \quad as \quad n \quad \rightarrow \quad \infty$$

In the first line the posterior probability is written in terms of Bayes' theorem;
in the second line we insert the actual priors p and (1-p) and the likelihood values in terms of the stated Normal density of the iid data X;
in the third line we rewrite two of the exponentials using the conditional value of the sample mean in terms of the corresponding significance z, and remove the normalization factors sqrt(n)/sqrt(2π)σ;
in the fourth line we maximize the expression by using the integral of the Normal.

THTQ: One Last Note on Very High Nσ

Recently heard claim from respected astrophysicist "**The quantity has been measured to be non-zero at 40σ level**", referring to a measurement quoted as 0.110+-0.0027.

That is a silly statement! As N goes above 7 or so, we are rapidly losing contact with the reality of experimental situations

To claim *e.g.* a 5σ effect, one has to be reasonably sure to know the p-value **PDF to the 10^{-7} level**

Remember, Nσ is just as femtobarns or or attometers: a useful placeholder for small numbers

– Hence before quoting high Nσ blindly, one shoud **think at what they really mean**

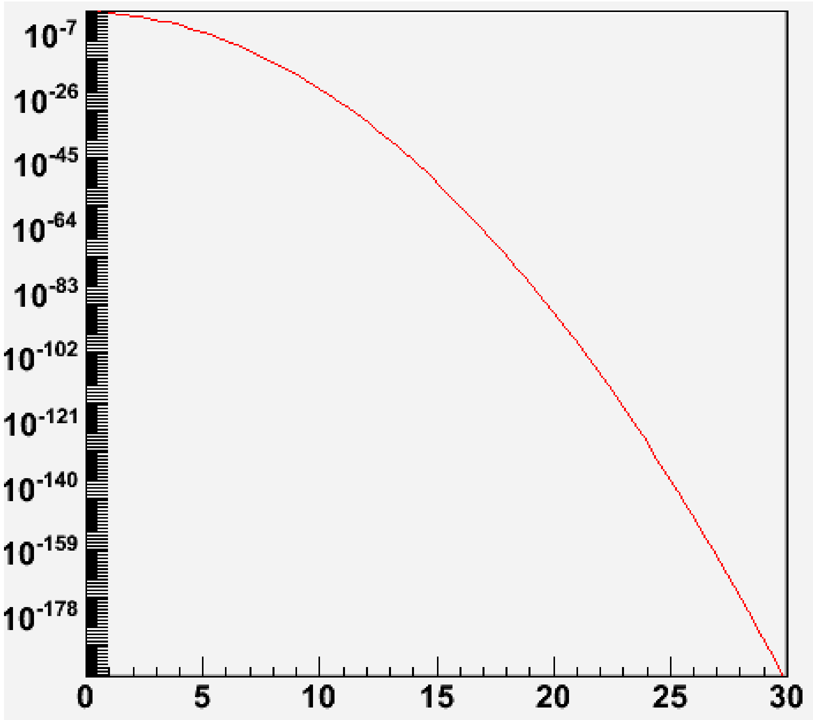
In the case of the astrophysicist, it is not even easy to directly make the conversion, as ErfInverse() breaks down above 7.5 or so. I resorted to a good approximation by Karagiannidis and Lioumpas [25],

$$Q(x) \approx \frac{\left(1-e^{-1.4x}\right)e^{-\frac{x^2}{2}}}{1.135\sqrt{2\pi}x}, x > 0$$

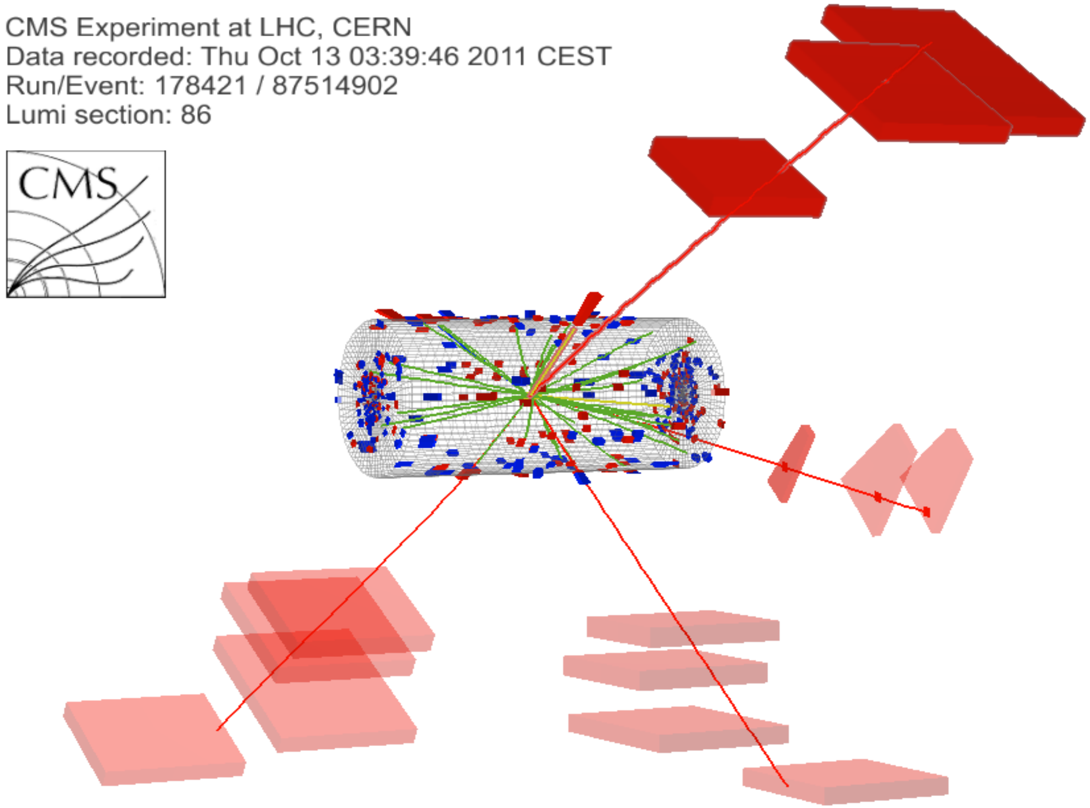
For N=40 my computer still refuses to give anything above 0, but for N=38 it gives p=2.5*10^{-316}

– so he was basically saying that the data had a probability of **less than a part in 10^316** of being observed if the null hypothesis held.

That is **beyond ridiculous** ! We will never be able to know the tails of our systematic uncertainties to something similar.



Higgs Discovery: a case study



Nuts and Bolts of Higgs Combination

The recipe must be explained in steps. The first one is of course the one of writing down extensively the likelihood function!

1) One writes a global likelihood function, whose parameter of interest is the strength modifier μ. If s and b denote signal and backg ... ite for a single channel:

$$\mathcal{L}(\mathrm{data}\,|\,\mu,\theta) = \mathrm{Poisson}\left(\,\mathrm{data}\,|\,\mu\cdot s(\theta)+b(\theta)\,\right)\cdot p(\tilde{\theta}|\theta)$$

Note that θ has a "prior" coming from a hypothetical auxiliary measurement.
In the LHC combination of Higgs searches, nuisances are treated in a frequentist way by taking for them the likelihood which would have produced as posterior, given a flat prior, the PDF one believes the nuisance is distributed f

In L one may combine many different search chan ... is performed as the product of their Poisson factors:

$$\prod_i \frac{(\mu s_i + b_i)^{n_i}}{n_i!}\,e^{-\mu s_i - b_i}$$

or from a unbinned likelih

$$k^{-1}\prod_i(\mu S f_s(x_i)+B f_b(x_i))\cdot e^{-(\mu S+B)}$$

2) One then constructs a profile likelihood test statistic q_μ as

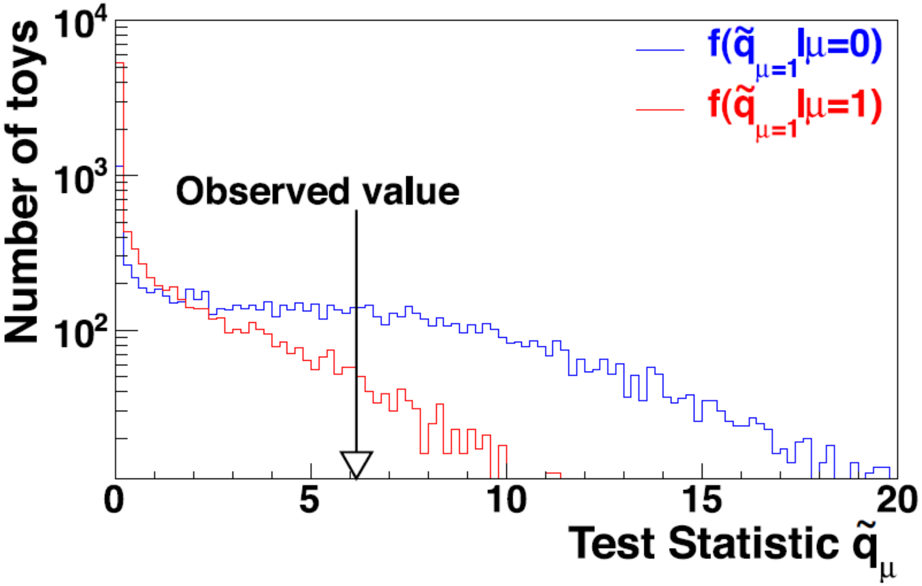
$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data}|\hat{\mu}, \hat{\theta})}$$

Note that the denominator has L computed with the values of μ^ and θ^ that globally maximize it, while the numerator has $\theta=\theta^{\wedge}_\mu$ computed as the conditional maximum likelihood estimate, given μ.

A constraint is posed on the MLE μ^ to be confined in 0<=μ^<=μ: this avoids negative solutions for the cross section, and ensures that best-fit values *above* the signal hypothesis μ are not counted as evidence against it.

3) ML values θ_μ^ for H_1 and θ_0^ for H_0 are then computed, given the data and μ=0 (bgr-only) and μ>0

4) Pseudo-data is then generated for the two hypotheses, using the above ML estimates of the nuisance parameters. With the data, one constructs the pdf of the test statistic given a signal of strength μ (H_1) and μ=0 (H_0). This way has good coverage properties.



5) With the pseudo-data one can then compute the integrals defining p-values for the two hypotheses. For the signal plus background hypothesis H_1 one has

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{signal+background}) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | \mu, \hat{\theta}_\mu^{\text{obs}}) \, d\tilde{q}_\mu$$

and for the null, background-only H_0 one has

$$1 - p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{obs} \,|\, \text{background-only}) = \int_{q_0^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | 0, \hat{\theta}_0^{\text{obs}}) \, d\tilde{q}_\mu$$

6) Finally one can compute the value called CL_s as

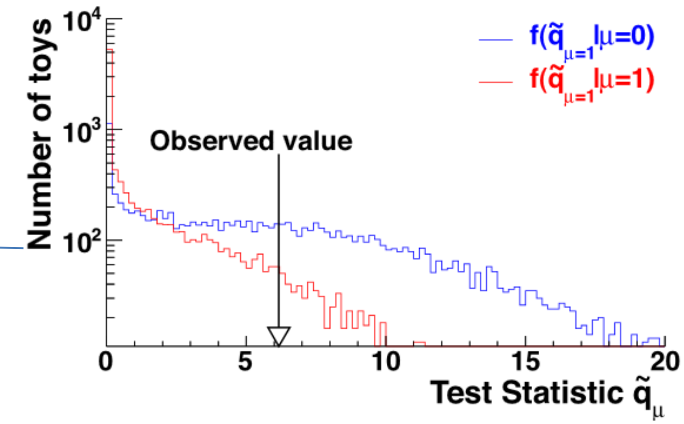
$CL_s = p_\mu/(1-p_b)$

CL_s is thus a "modified" p-value, in the sense that it describes how likely it is that the value of test statistic is observed under the alternative hypothesis by also accounting for how likely the null is: the drawing incorrect inferences based on extreme values of p_μ is "damped", and cases when one has no real discriminating power, approaching the limit $f(q|\mu)=f(q|0)$, are prevented from allowing to exclude the alternate hypothesis.

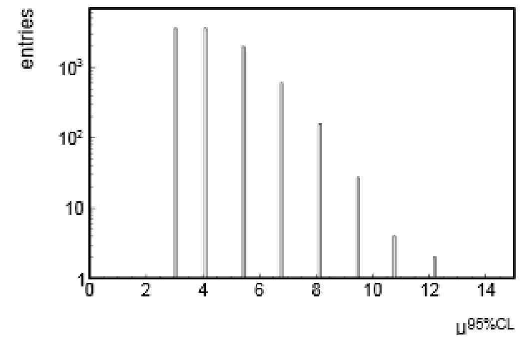
7) We can then **exclude H_1 when $CL_s < \alpha$**, the (defined in advance !) *size* of the test. In the case of Higgs searches, all mass hypotheses $H_1(M)$ for which $CL_s<0.05$ are said to be excluded (one would rather call them "disfavoured"…)

Derivation of expected limits

One starts with the **background-only hypothesis μ=0**, and determines a distribution of possible outcomes of the experiment with toys, obtaining the CLs test statistic distribution for each investigated Higgs mass point

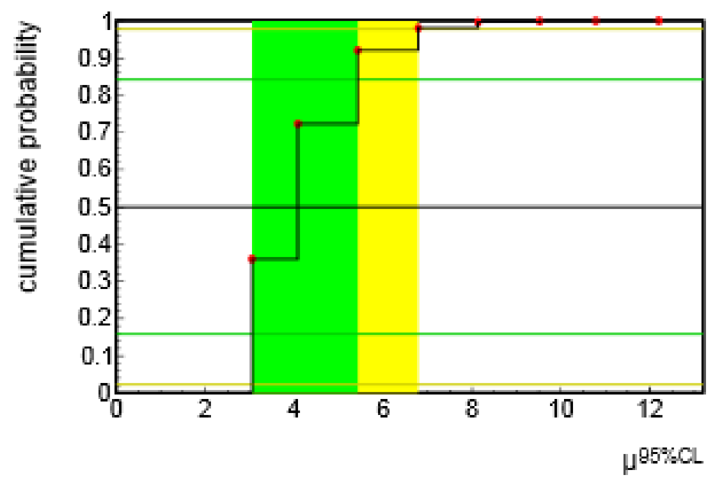


From CLs one obtains the PDF of upper limits μ^{UL} on μor each M_h. [*E.g. on the right we assumed b=1 and s=0 for μ=0, whereas μ=1 would produce <s>=1*]



Then one computes the cumulative PDF of μ^{UL}

Finally, one can derive the median and the intervals for μ which correspond to 2.3%, 15.9%, 50%, 84.1%, 97.7% quantiles. These define the “expected-limit bands” and their center.



Significance in the Higgs search

- To test for the significance of an excess of events, given a M_h hypothesis, one uses the bgr-only hypothesis and constructs a modified version of the q test statistic:

$$q_0 = -2\ln\frac{\mathcal{L}(\text{data}|0,\hat{\theta}_0)}{\mathcal{L}(\text{data}|\hat{\mu},\hat{\theta})} \quad \text{and } \hat{\mu} \geq 0.$$

- This time we are testing any $\mu>0$ versus the H_0 hypothesis. One builds the distribution $f(q_0|0,\theta_0^{\wedge obs})$ by generating pseudo-data, and derives a p-value corresponding to a given observation as

$$p_0 = P(q_0 \geq q_0^{obs}) = \int_{q_0^{obs}}^{\infty} f(q_0|0,\hat{\theta}_0^{\text{obs}})\, dq_0.$$

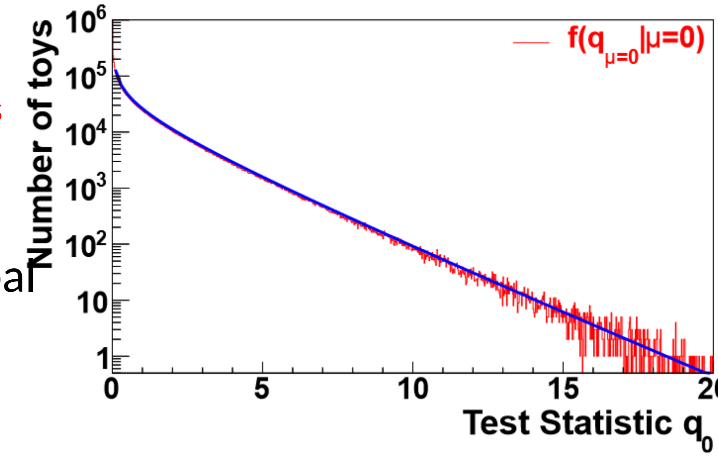
One then converts p into Z using the relation

$$p = \int_Z^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)\,dx = \frac{1}{2}P_{\chi_1^2}(Z^2)$$

where $p_\chi{}^2$ is the survival function for the 1-dof chi2.

Often it is impractical to generate large datasets given the complexity of the search (dozens of search channels and sub-channels, correlated among each other). One then relies on a very good asymptotic approximation:

$$p^{estimate} = \frac{1}{2}\left[1-\text{erf}\left(\sqrt{q_0^{\text{obs}}/2}\right)\right]$$



The derived p-value and the corresponding Z value are "local": they correspond to the specific hypothesis that has been tested (a specific M_h) as q_0 also depends on M_h (the search changes as M_h varies)

When dealing with many searches, one needs to get a global p-value and significance, i.e. **evaluate a trials factor.** This can be done using the techniques discussed earlier.