True measurements…
false discoveries

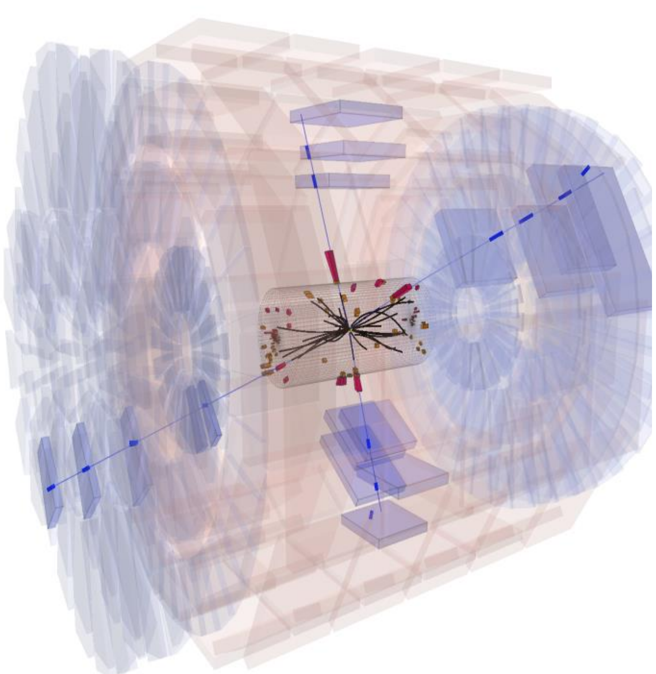
The history of HEP has seen extraordinary claims …an attempt to analyse why some were not true and what to do to avoid it

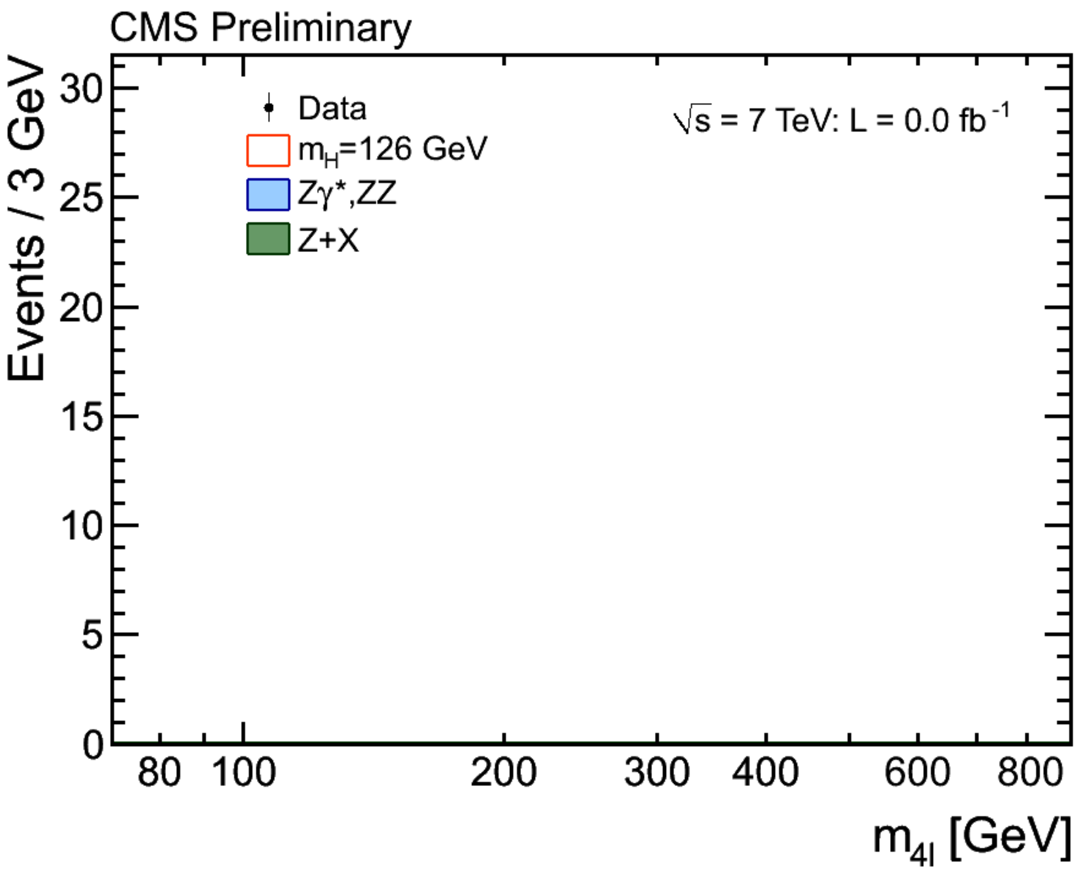
Extraordinary claims

- In the last years we have been confronted with extraordinary widely advertised discoveries
  - Higgs Boson 
  - Superluminal neutrinos 
  - Gravitational waves 
- We will attempt to go through the criterias we are using to make ‘discovery’ claims, a little history on how they have developed and see the possible limitations of our approaches

Success story (I)

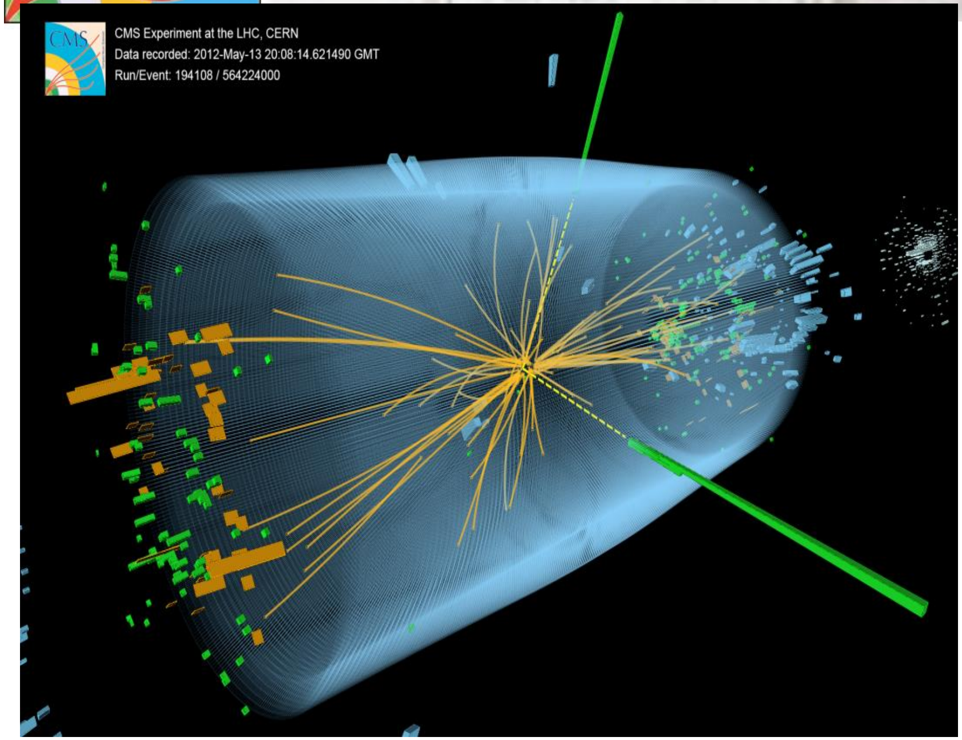
Higgs→ZZ→4μ



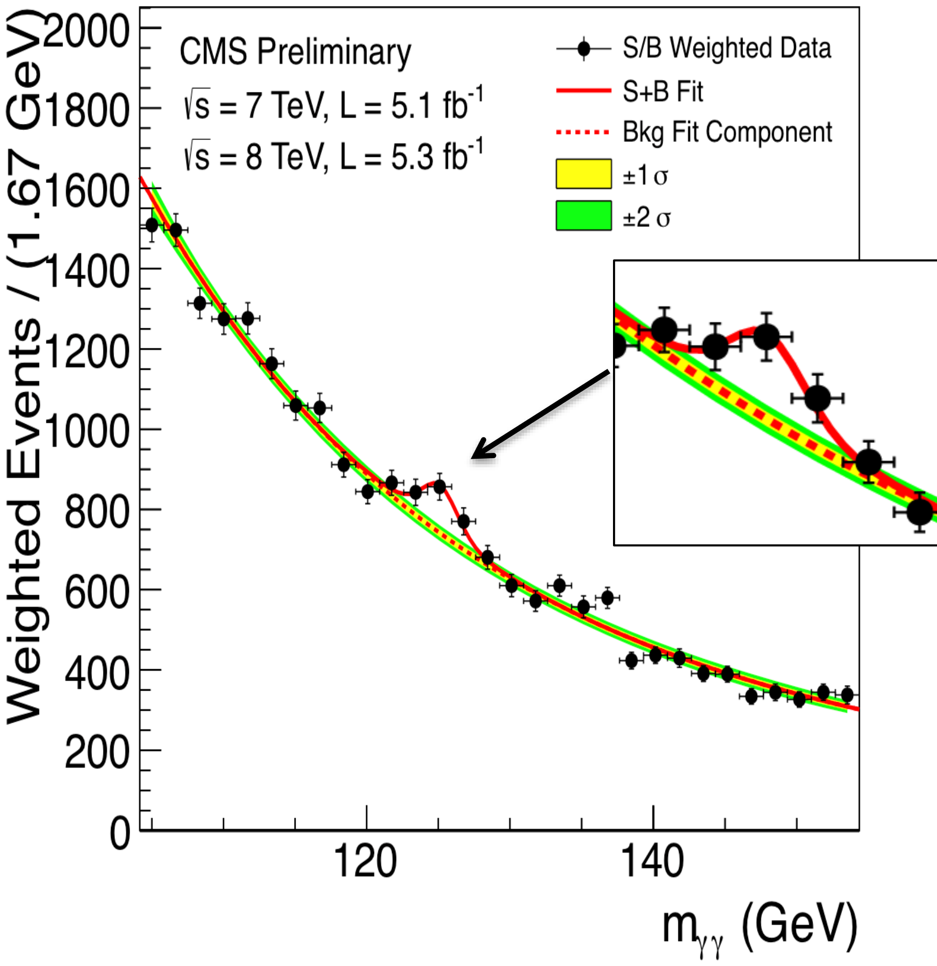


Probability to produce a Higgs boson in the 2 proton collision is 10^{10} times smaller than to produce any other final state: choice of decay channel determines S/B ratio and mass resolution

Success story (II)



pp → H → γγ



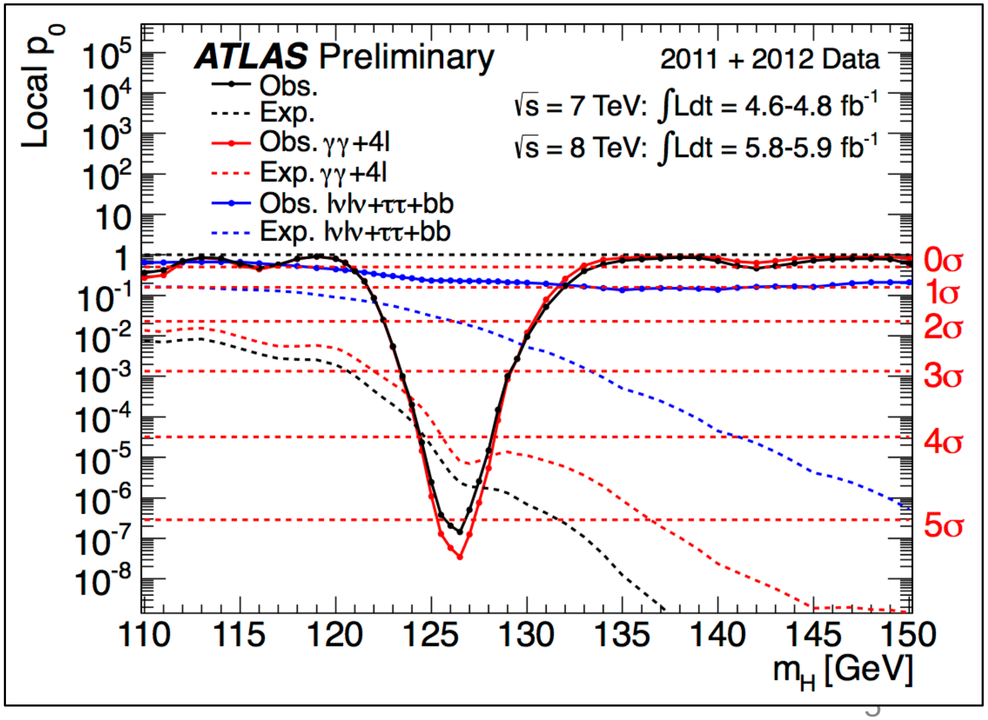
..and everybody was talking of '5σ'

The Number if sigmas are derived by:

1) making the ratio of the likelyhoods of the null (background only) hypothesis and the 'signal' hypothesis (background +signal) $\Lambda(\theta_0)/\Lambda(\theta_1)$
2) And using Wick's theorem (that for large stats the Log of the Likelyhood ratio is distributed like a χ^2) and then extract a **tail probability** which can be converted in number of Sigmas

$2\ \ln(\Lambda(\theta_0)/\Lambda(\theta_1)) =$
$-2\ (\ln\Lambda_1 - \ln\Lambda_0)$
is interpreted as a value sampled from a χ^2 distribution then $P(\chi^2, N_{dof})$ is the P value from which one extracts the N sigmas

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

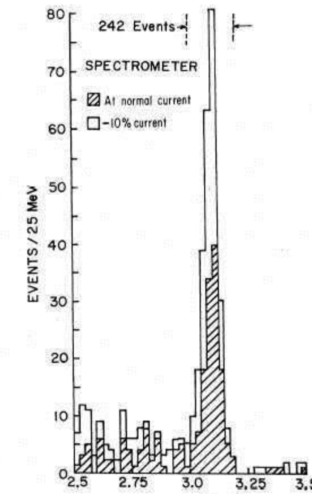


Some caveats

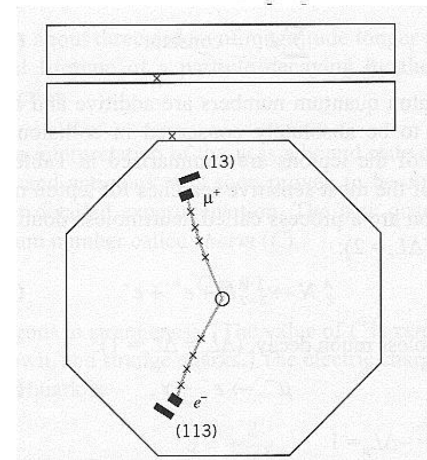
- The whole construction rests on a proper definition of the p-value. Any shortcoming of the properties of p (*e.g*. a tiny non-flatness of its PDF under the null hypothesis) totally invalidates the meaning of the derived Nσ
  - In particular, using "sigma" units does in no way mean we are implying some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

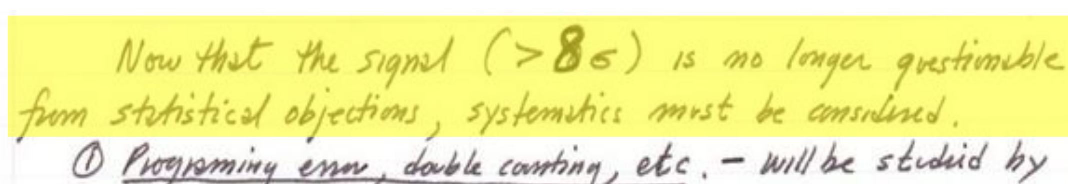
  we prefer to say "5σ" than "0.00000029"

Some history about major discoveries



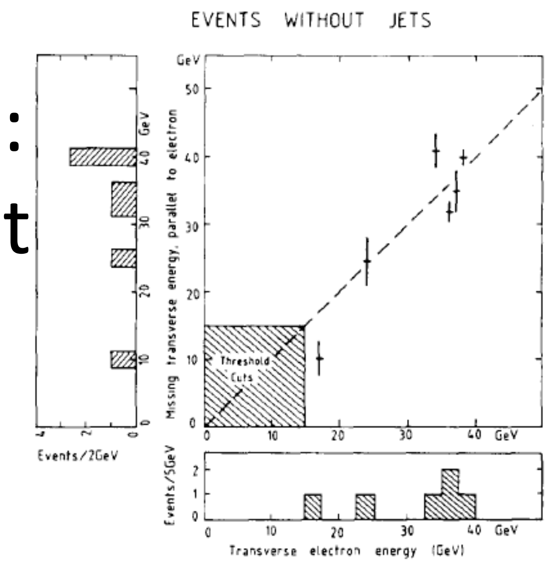
- A rigorous approach with respect to ‘discovery‘ was not always enforced
  - **The J/ψ discovery** (1974): no discussion of significance – the peaks were too big for even bothering discussing significance
  - The τ discovery: discussion about the excess of e-μ events were more about hadron backgrounds
  - The Upsilon discovery involved lots of statistical tests (mainly because of the ‘false’ evidence at 6 GeV –so called ‘Oops Leon’) even if the evidence exceed by far 5σ

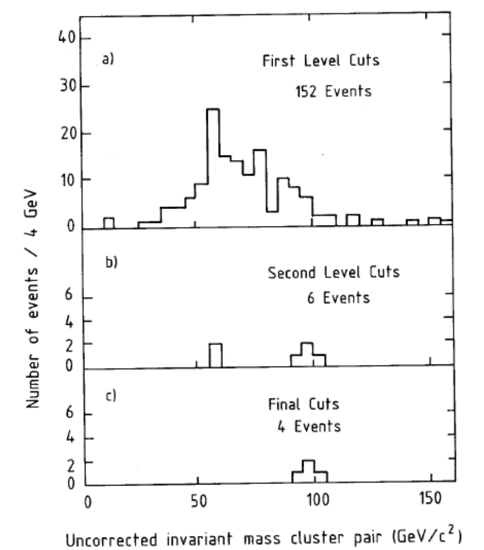



Now that the signal (> 8σ) is no longer questimable from statistical objections, systematics must be considered.
① Programming error, double counting, etc. – will be studied by

More history

- W boson discovery (January 1983): 6 events, no statistical analysis, but discussion about absence of background
- Z boson (May 1983) 4 events, also here discussion to show that backgrounds are negligible





Top: the first modern application of the 5σ criteria

- 1994 the CDF experiment publishes 'Evidence' based on a counting excess (2.7σ) in b-tagged single-lepton and di-lepton datasets accumulating in a mass peak which was over 3σ by itself

$M = 174 \pm 10^{+13}_{-12}$ GeV (now it is 173+-0.5

- One year later CDF and D0 (with 3 times more data) presented counting excesses at the level of 5σ and claimed 'Discovery' !

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.

The birth of the 5 sigma criteria

Read: exotic hadrons

"Are There Any Far-out Mesons or Baryons?" ,A.H.Rosenfeld in
Charles Baltay & Arthur H. Rosenfeld Meson Spectroscopy W.A. Benjamin Inc. 1968

Arthur H. Rosenfeld (Univ. Berkeley)

Table II.

Hydrogen bubble chamber events measured in U. S. in year ending August 1967 (excluding about 300,000 image-plane-digitizer measurements made to study Σ leptonic decay).

Outgoing prongs	Outgoing particles	Number of mass combinations	Events measured (thousands)	Mass combinations (millions)
2	2 or 3	1 or 3, avge = 2	500	1
4	4 or 5	10 or 25, avge = 17	1200	21
6	6 or 7	56 or 119, avge = 88	70	6
Total U. S. :			~1,700	~28
Assume 20% were remeasurements:			~1,400	~23
×1.5 (?) to include other countries:			~2	~35

Divide by 2500 events/histogram; yields 15,000 histograms.

Large trial factor

2. Number f/h of bumps/histogram. 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 t-cuts, looking both inside and outside

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; to be conservative, I used the number 10 for the number f/h.

Then, to repeat my warning at the beginning of this section; we are now generating at least 100,000 potential bumps per year, and should expect several 4σ and hundreds of 3σ fluctuations. What are the implications? To the theoretician or phenomenologist the moral is simple; wait for nearly 5σ effects. For the experimental group who have just spent a year of their time and perhaps a million dollars, the problem is harder. I suggest that they should go ahead and publish their tantalizing bump (or at least circulate it as a report.) But they should realize that any bump less than about 5σ constitutes only a call for a repeat of the experiment. If they, or somebody else, can double the number of counts, the number of standard deviations should increase by $\sqrt{2}$, and that will confirm the original effect.

A comparison with the literature in fact showed a correspondence of his estimate with the number of unconfirmed new particle claims.

First reason for 5σ: stat fluctuation

- Besides rendering pure statistical fluctuation unlikely, the 5σ criteria aims to protect from the fact that if we try hard enough we SHALL find a fluctuation
- The number of trials required to reach 10^{-7} probabilities is of course very large…on the other hand modern experiments are performing a large number of searches... so we tend to correct our significances by estimating the Look Elsewhere Effect which accounts for the reduction of significance due to the trials we made to find an excess
- The brute force way to estimate the LEE is to simulate a set of experiments under the null (background only) hypothesis and varying all the parameters within their precision and check for the likelyhood to have a significant fluctuation: in order to match 10^{-7} one has to 'simulate' order of 10^{7} experiments for each parameter set …
- When dealing with some of the searches at LHC this can be practically impossible (the Higgs search implied combination of dozens of deifferent channels with hundreds of nuisance parameters)
- Recently 'asymptotic' methods have been defined to evaluate the LEE
  E. Gross and O. Vitells, "*Trials factors for the Look-Elsewhere Effect in High-Energy Physics*", arxiv:1005.1891v3, Oct 7th 2010

Notes About the LEE Estimation

courtesy of Tommaso Dorigo

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. Different people 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.

The name of the game: systematic errors

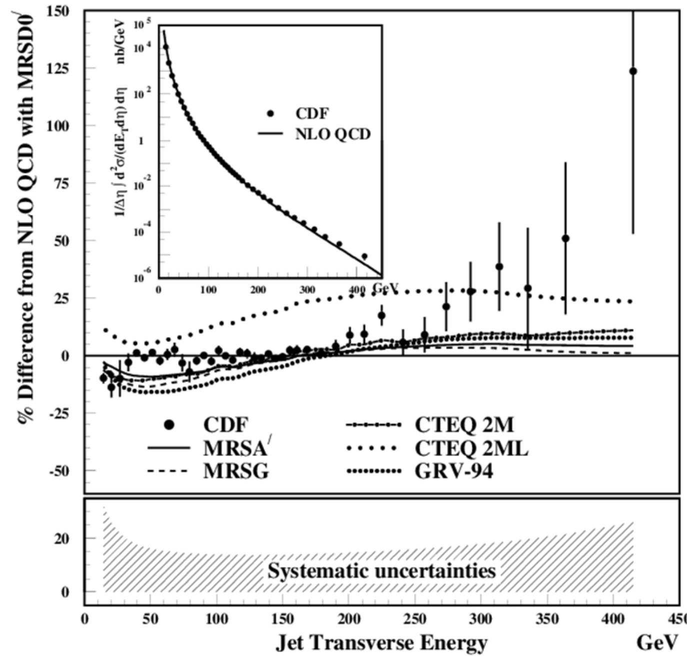
- The other reason for the 5σ criteria is to protect against problems with the modelling of the systematics behind a given measurement
- The evaluation of systematic errors is a challenging field:
  - The models used to evaluate the null hypothesis could be flawed/incomplete
  - The subjective prejudices on the way to evaluate them can play a significant role
  - The underlying assumption that the 'systematic' error is gaussian is often a rough approximation

Model inadequacy

- When looking for new phenomena the discovery *assumes* a correct estimation of the null-Hypothesis, i.e. showing that one sees an excess with respect to what is predicted by the model without the new phenomena.
- The limitation of the theoretical modelling (sometime its implementations and/or understanding by the experimental teams) have been the source of ‘false’ claims in the recent history

Examples: quark substructure

- Quark substructure: the imperfect knowledge of the Parton Distribution Functions inside the proton have led to some unjustified excitement




The New York Times

U.S.

Search All NYTimes.com

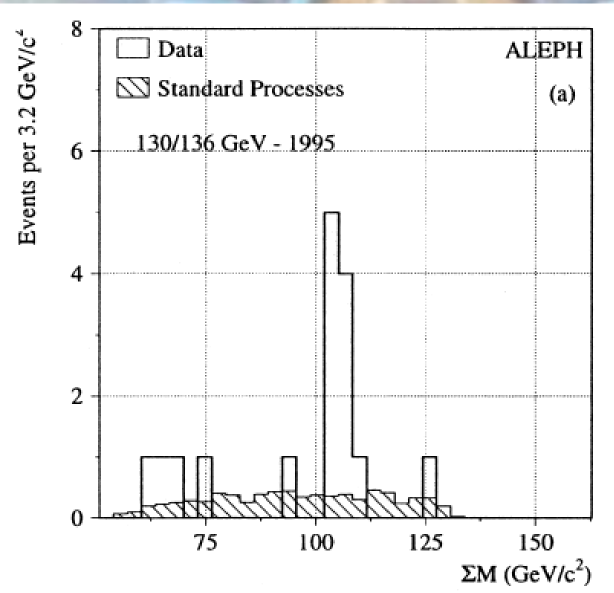
COLLECTIONS > FERMILAB

Tiniest Nuclear Building Block May Not Be the Quark

By MALCOLM W. BROWNE
Published: February 08, 1996

Example: new particle

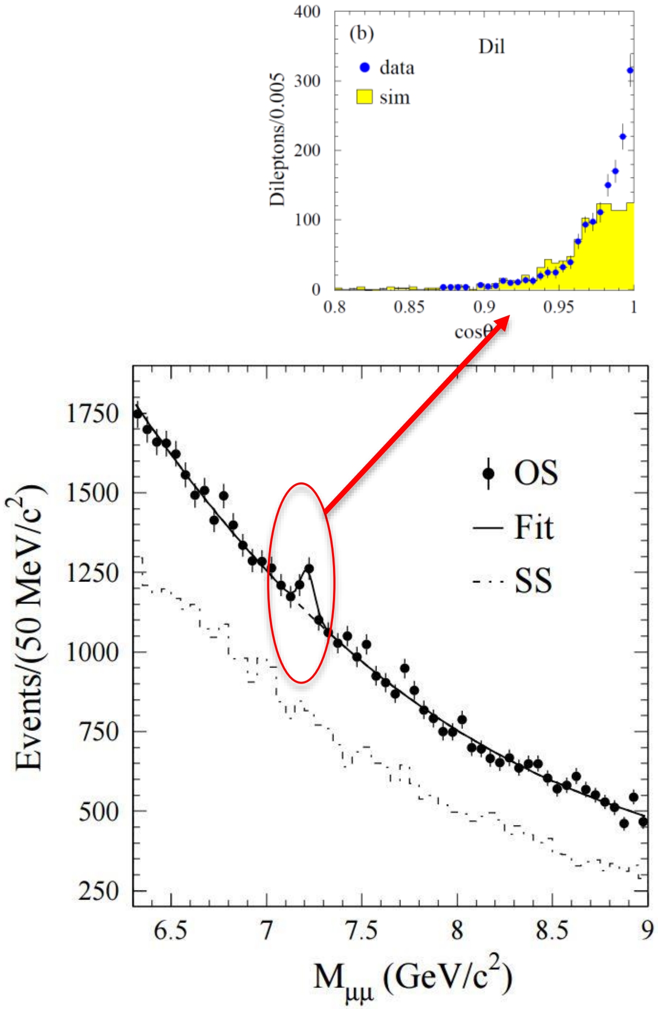
- ALEPH observed in 1996 a 4σ excess of Higgs-like events at 105 GeV in the 4-jet final state of electron-positron collisions at 130-136 GeV. They published the search:9 events in a narrow mass region with an expected background of 0.7
- None of the other LEP expts saw anything, but still the run at CM energy of 136 GeV was repeated..and the peak was not confirmed



In DELPHI we could see some events of this kind if we dropped the cut on possible radiative returns where the photon would 'hadronize to a ρ

Example : Sbottom 'discovery'

- CDF (1999) observed a significant excess of events with two or more leptons in dijet events
- …with characteristics different from B decays
- Evidence disappeared when inreasing by orders of magnitude the sample



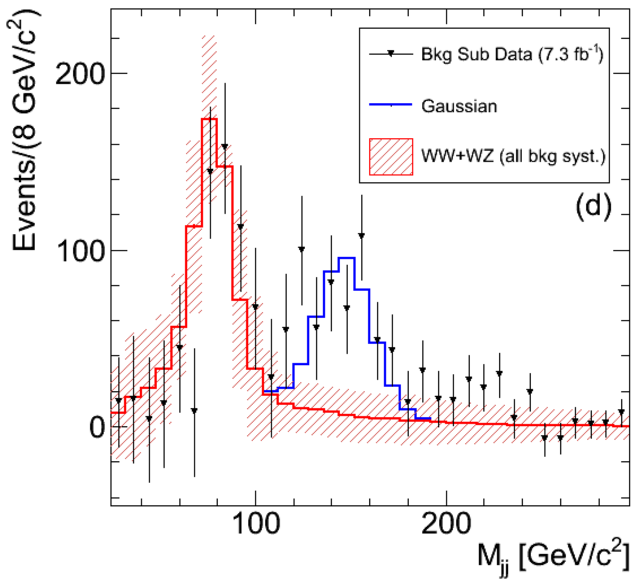
…but news spread

- Aleph informed about the CDF excess …found a 3σ effect in their data (LEPC, July 2000)
- DELPHI showed some problem with the MC simulation of ALEPH and that no excess was present in their data
- Later the signal was understood as an artifact of a wrong MC simulation and miscalibrated electron fake rates

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
205	94	4.7	3
Total	586	33.6	56

Example : exotic discovery

- 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.
- The effect grew with data size!
- It was eventually understood to be due to the combination of two nasty background contaminations



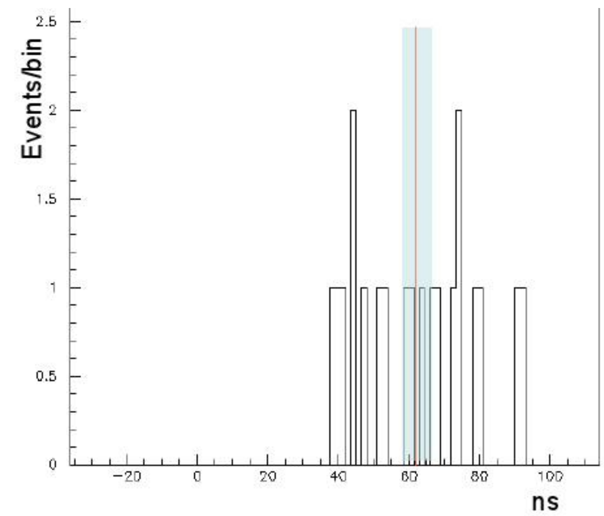
. Aaltonen et al., "*Invariant-mass distribution of jet pairs produced in association with a W boson in p pbar collisions at sqrt(s) = 1.96 TeV using the full CDF Run II data set*", Phys. Rev. D 89 (2014) 092001.

Subjective prejudice

- A certain level of 'personal' appreciation in the estimation of Systematic errors is almost unavoidable
- It can go both ways: adopting criteria which inflate the error (hence with the danger of preventing optimal extraction of information from the data) or having an attitude too optimistic about the understanding of possible systematics ..and so provoking false claims
- VERY dangerous is the "N" effect

Example: superluminal neutrinos

- In 2011 the OPERA collaboration produced a measurement of neutrino travel times from CERN to Gran Sasso which appeared to go faster than light in vacuum. The effect was at the level of 6σ It was finally understood to be due to a single large source of systematic uncertainty – a loose cable
- There have been conjectures that the haste with which the result was 'put out' was also due to the rumors about an imminent result from a 'competing' US collaboration



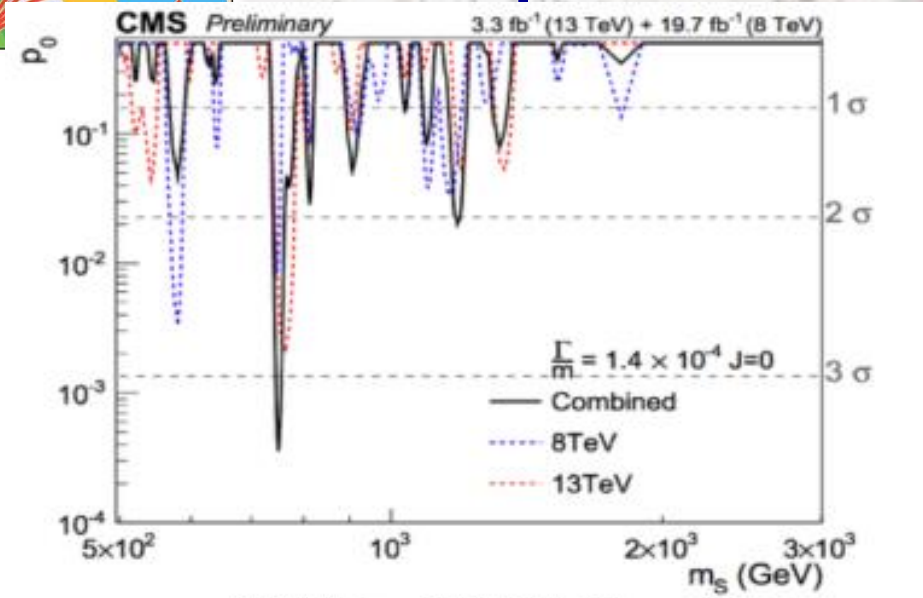
T. Adam et al., "*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam", *JHEP 10 (2012) 093.

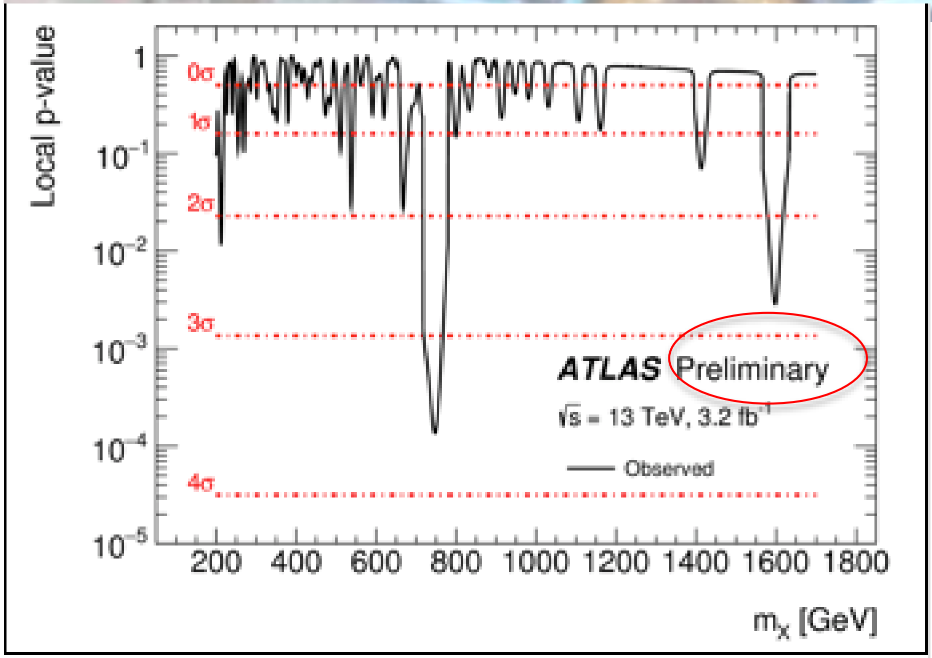
T. Adam et al., "*Measurement of the neutrino velocity with the OPERA detector in the CNGS beam using the 2012 dedicated data*", JHEP 01 (2013) 153.

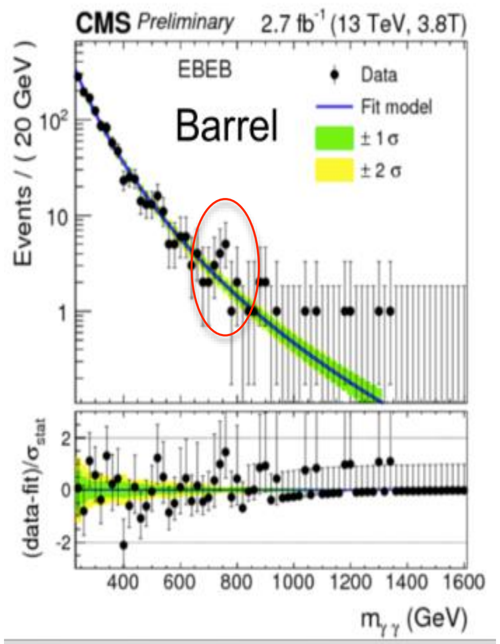
…the importance of 'preliminary'

- We have grown accustomed to have a 'quick' presentation of results at conferences or lab seminar which are labelled *preliminary* because not all the ultimate treatment of the data/sophistication of the analysis has been implemented..That often hides the fact that the systematic errors one quotes for these results might be rough estimations stemming sometime from 'subjective' judgement of what is still missing to achieve the ultimate result

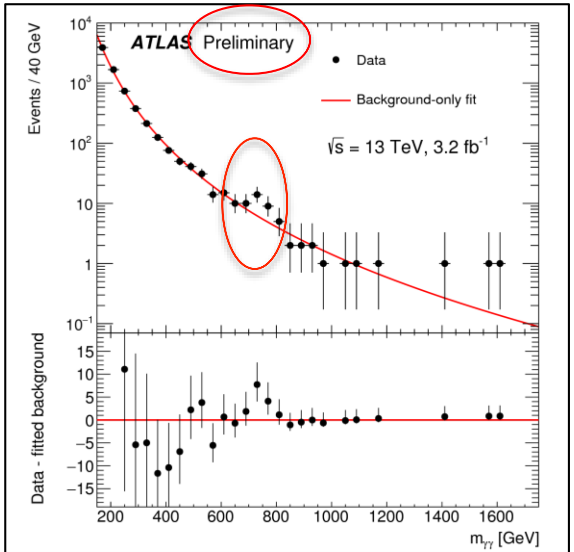
Example: the 750 Gev γγ bump







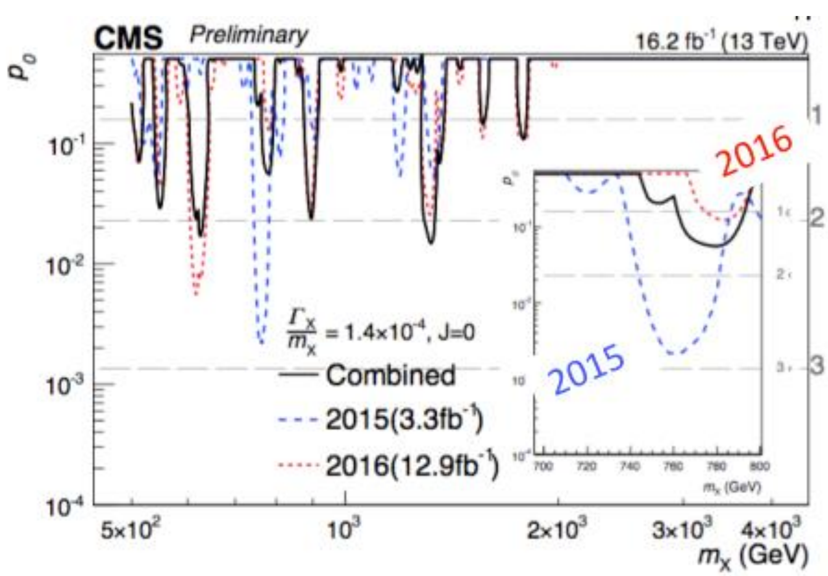
December 2015
CERN jamboree:
Striking coincidence of an excess at the same mass !
Refrained to make a combination which would have pushed things close to discovery level ..because of the preliminary nature



..and we know the story

- ...more than 400 theoretical papers in 3 months
- ... Excitement in the media
- ... But with 4 times more data ...looks like a fluctuation

CMS , ICHEP 2016



Atlas: D. Charlton, ICHEP 2016

2016 data: no clustering around 730-750 GeV, and 3.8x more data

- 2016 data consistent with 2015 at the 2.7σ level
- Appears that the 2015 excess was a statistical fluctuation

…but the real story is

- ..that there was never a match between the ATLAS and CMS excesses !

Diphoton Searches

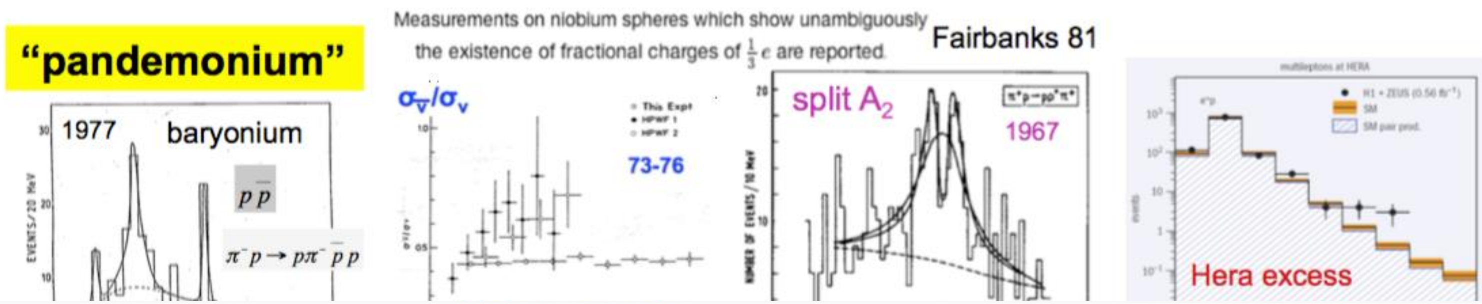
Localised excess seen in 2015 ATLAS data

- 2.1σ global (3.9σ local) significance at 750 GeV (spin-0 search), width ~50 GeV
- After reprocessing, new 2016 reconstruction → 3.4σ local, at ~730 GeV

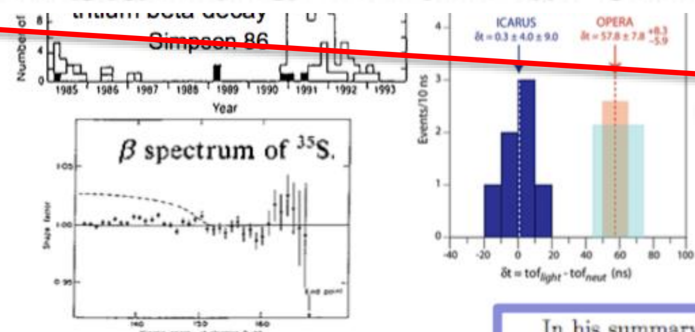
From D. Charlton presentation at ICHEP 2016, Chicago…and from discussion with the analyzers the major change was the calibration of the ECAL

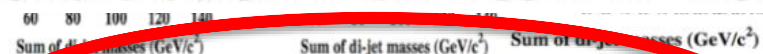
Importance of making sure that the detector response is fully understood: for example history of Higgs search at the endof LEP when DELPHI 'significant' candidates became perfect WW after final calibration/alignment

Moriond: the time for excesses



^{17}The Moriond Workshops play an extremely important role in speculative/controversial issues. They provide a forum for those working in the field to meet, present papers, and have both formal and informal discussions and criticism. For a dis-





^{17}The Moriond Workshops play an extremely important role in speculative/controversial issues. They provide a forum for those working in the field to meet, present papers, and have both formal and informal discussions and criticism. For a discussion of the role that the Moriond workshops played in another controversial episode, that of the fifth force, a proposed modification of Newton's law of gravity, see Franklin (1993a).

In his summary[3] of this conference one year ago, the first three experimental topics mentioned by John Collins were the excess of high Q^2 events at HERA, the excess of high E_T inclusive jets in CDF data, and the excess of $W+1$ jet events, relative to W's without jets, reported by DØ. At this meeting, important new results were presented that bear directly on these topics, to which I now turn.

Strovink 98

under scrutiny

- ✧ 3.5 keV line?
- ✧ DB Heidelberg?
- ✧ DAMA oscillation?

Courtesy of D. Treille

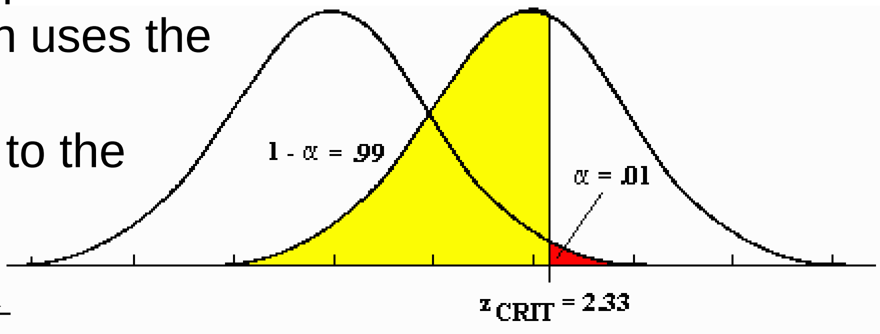
Subconscious 'expectations'

- Minds are 'bayesian' in nature: we have (most of the time subconscious) priors about the probabilities we assign to different hypothesis

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

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

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.

Example: new physics in Flavour-land

Trust in ’deviations’ depends on reliability of Theoretic al expectations

Modified Ligeti Plot from Gilad Perez (SEARCH 2016)

Ligeti: 1606.02756

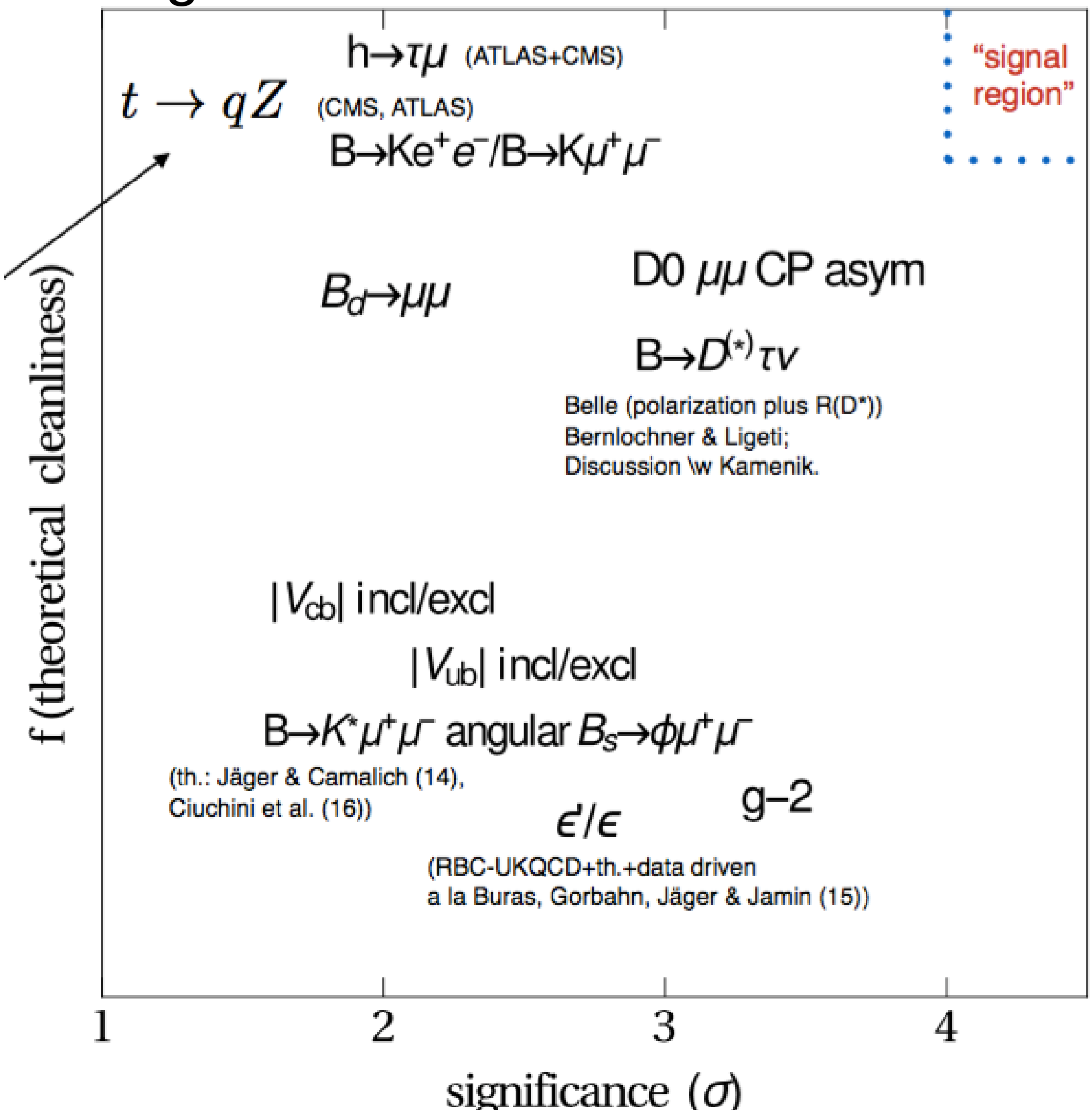


Table of Searches for New Phenomena and "Reasonable" Significance Levels

L. Lyons, "*Discovering the significance of 5σ*", arxiv:1310.1284v1

Search	Surprise level	Impact	LEE	Systematics	# of σ
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

An aside: Bayesian vs frequentists

- The approach to discovery is different depending on the approach one has:
  - Bayesian: compares posterior probability of a an assumed prior
  - Frequentist: uses P values
- The Jeffreys-Lindley paradox states that an assumed 'null' prior will always be favored when getting high statistics
- Frequentists and Bayesians draw **opposite conclusions** on large data when comparing a null-hypothesis to a composite alternative

example

	First data set	Second data set
H_0	Poisson, $\mu = 1.0$	Poisson, $\mu = 10.0$
H_1	Poisson, $\mu = 10.0$	Poisson, $\mu = 100.0$
n_{obs}	10	31
p_0	1.1×10^{-7} 5.2σ	0.8×10^{-7} 5.3σ
L_0/L_1	8×10^{-7} Strongly favours H_1	$1.2 \times 10^{+8}$ Strongly favours H_0

Table 2: Comparing p-values and likelihood ratios

Are we expecting new discoveries?

We have laid the Keystone of the Std Model Cathedral …

What we will do is to get a better ‘picture’ ie. measure better the characterisic of the Std Model

Is this all left to do ?

Not the first time that the issue is posed:
Lord Kelvin (1900)
There is nothing new to be discovered in physics now. All that remains is more and more precise measurement.

Some of the known unanswered questions:

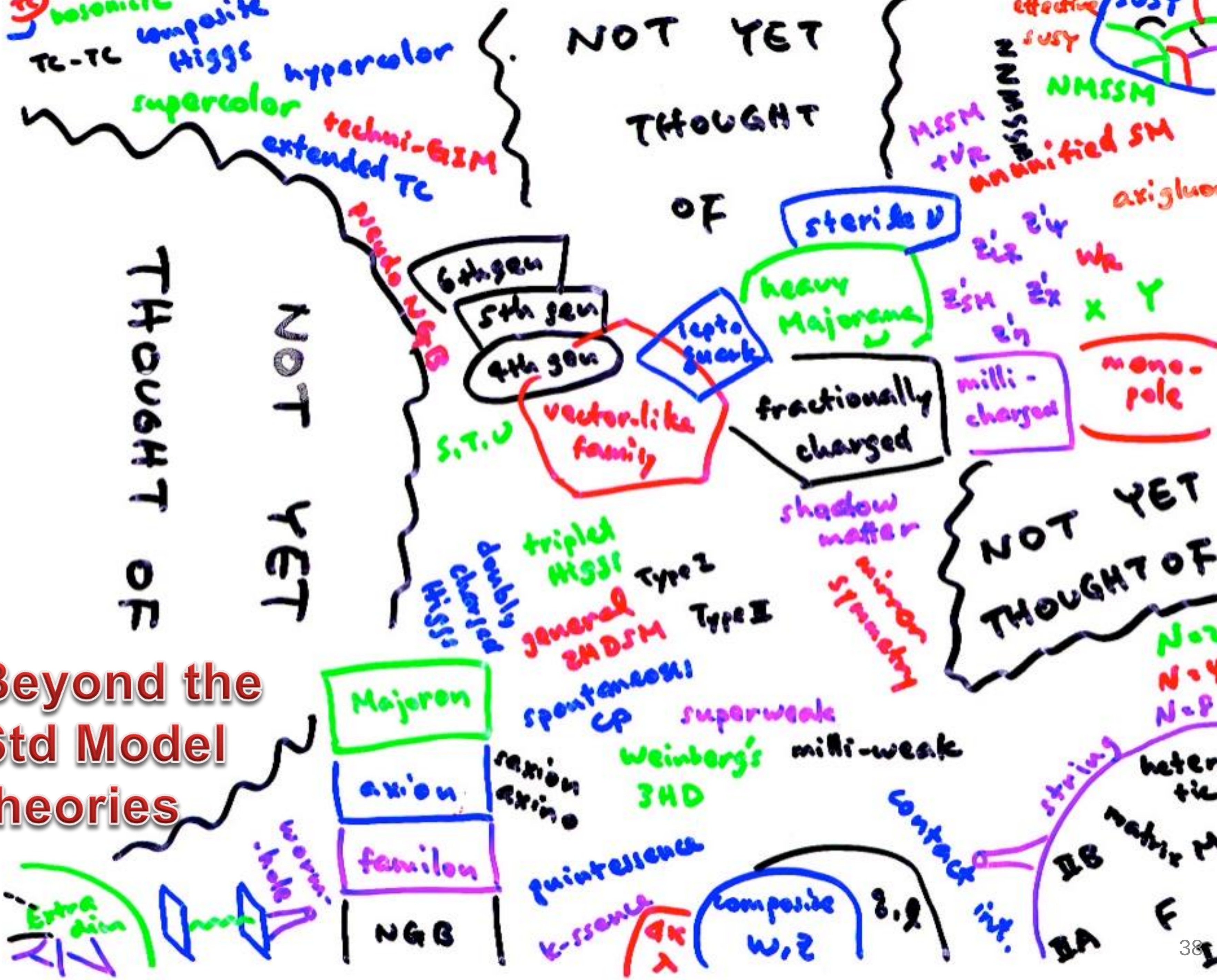
- The elephant in our 'research' room has been Gravity: the difficulty to reconcile Quantum Mechanics and Gravity has been a Theoretical Nightmare since ~ 100 years.
- Dark Matter is another cloud in the Standard Model sky (more on this later)
- Why we have essentially only Baryonic matter and not anti-baryons in the universe is another blemish on the Std Model
- …and we should be ready to deal with surprises: it would not be the first time in the field of High Energy Particle Physics that Nature has shown phenomena which we had not anticipated

..there must be more than the Std. Model!

- Why three families of Fundamental particles ?
- What is the structure of the Neutrino sector : is a signature for physics beyond the STD model hidden in the neutrino transformations?
- The standard model itself seems to indicate that something is missing :
  - What is allowing the Mass of the Higgs boson to be as low as measured ? If there is nothing
  - else the standard model has to be valid up to Planck Mass where Quantum Mechanics and Gravity HAVE to come together

There is more to Nature than the STD model construction

Beyond the Std Model theories



Relations between theory and experiment (as seen by theorists)

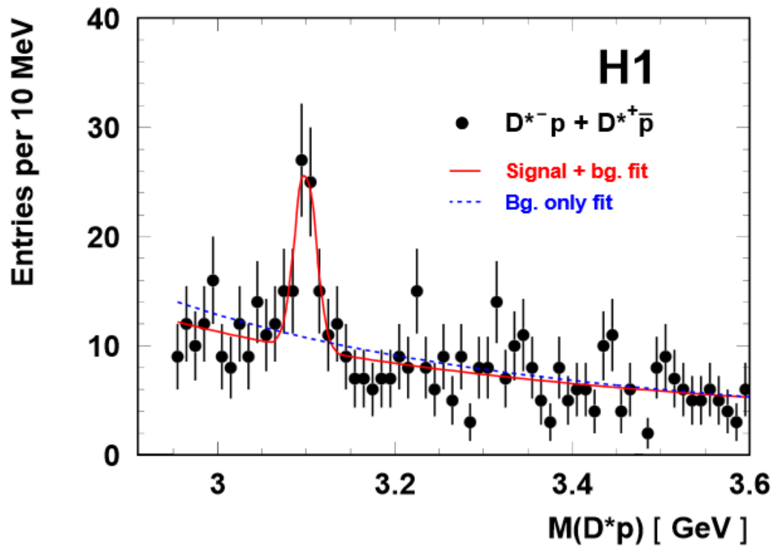
A defendable picture when you have very tight predictions: e.g. Higgs boson, rare decays rate

Courtesy of H. Yamamoto

..as seen by experimentalists

Are 5σ a safe 'bet'?

- Not really…example H1 'evidence' for pentaquark
- Despite the thing being quoted at the 6.2 Sigmas level they were smart enough to use the word 'Evidence' in the title
- …so be prepared for some possible fake peak at 5σ inthe future of the LHC running



A. Aktas et al., "*Evidence for a narrow anti-charm baryon state*", Phys. Lett. B588 (2004) 17.

Summary

Understanding of Nature behaviour has always required ever improved tools and measurement devices

The complexity of today's instruments and the sophistication of the measurements we are doing requires a rigorous approach to understand the detectors we are using and the backgrounds we are expecting

We have developed a deep understanding of the pit-falls to avoid from the errors of the past

And the fundamental principle of the necessity of having more than one experiment able to perform the same measurement has been proven over and over again.

We are ready to exploit fully our data and make discovery if nature will be kind to us!