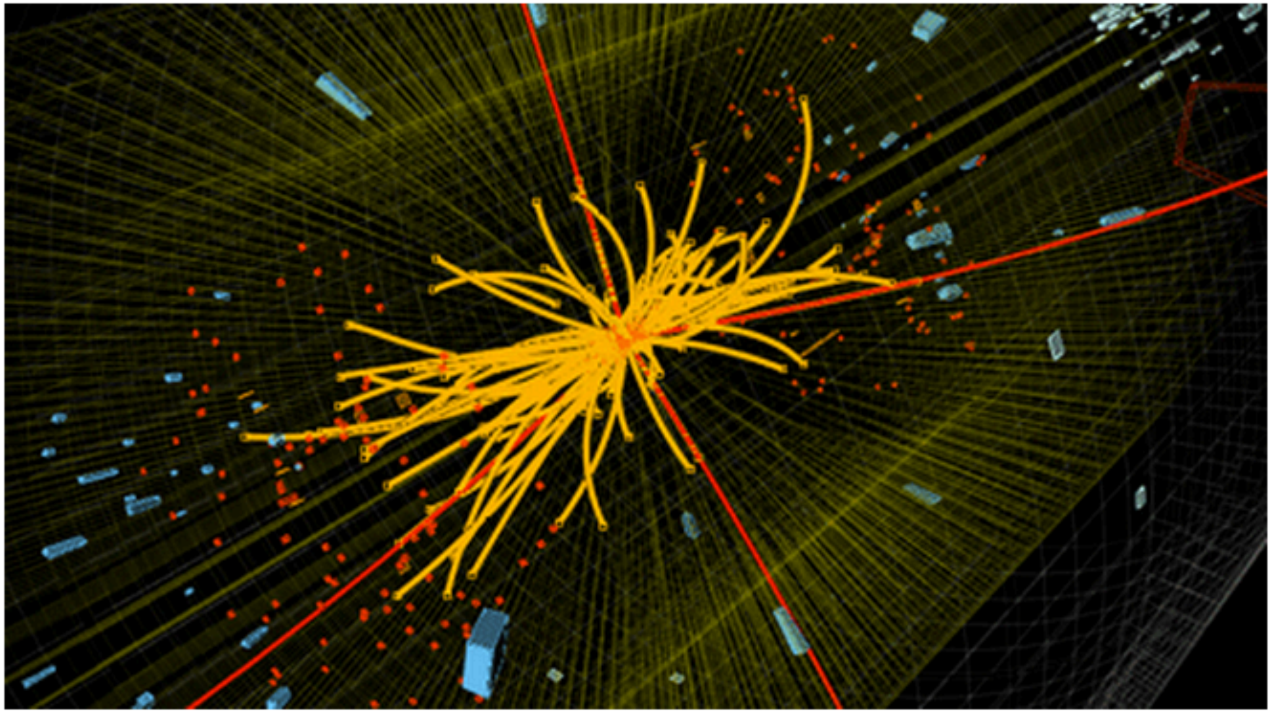
Data Analysis

Ivica Puljak
CERN and University of Split, FESB, Split, Croatia

Ivica.Puljak@cern.ch

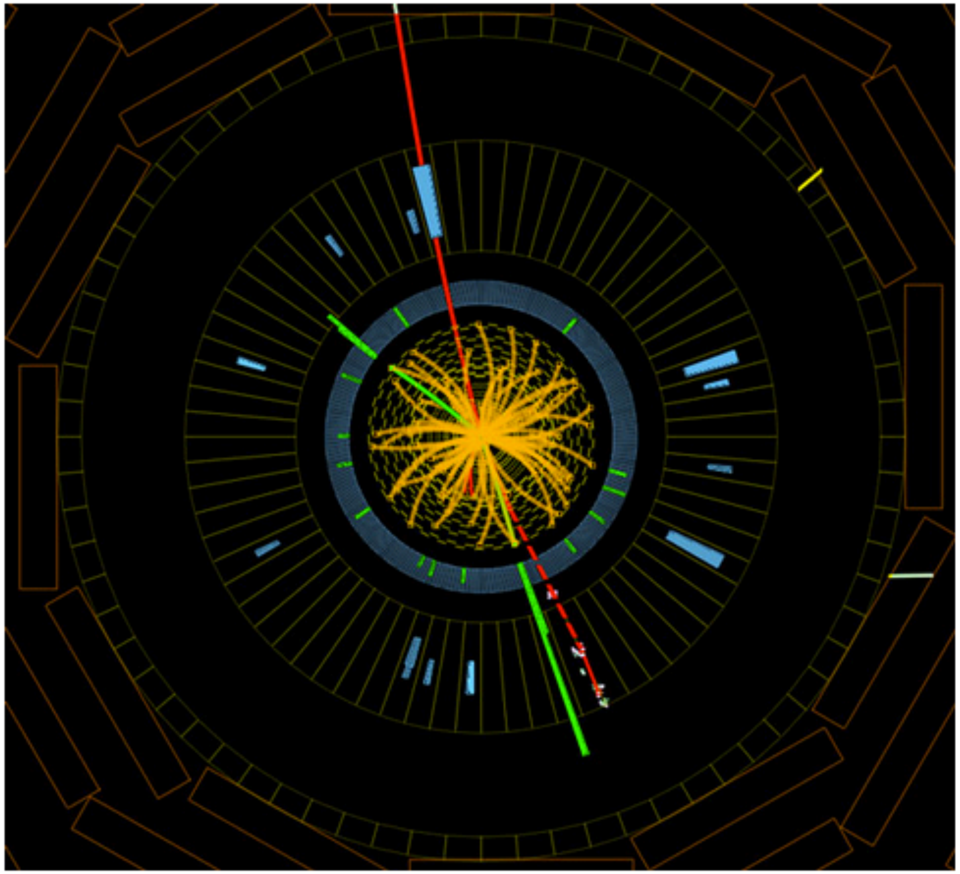
04.07.2012: Higgs within reach



Proton-proton collision in the CMS experiment producing four high-energy muons (red lines). The event shows characteristics expected from the decay of a Higgs boson but it is also consistent with background Standard Model physics processes (Image: CMS)

At a seminar on 4 July, the ATLAS and CMS experiments at CERN presented their latest results in the search for the long-sought Higgs boson. Both experiments see strong indications for the presence of a new particle, which could be the Higgs boson, in the mass region around 126 gigaelectronvolts (GeV).

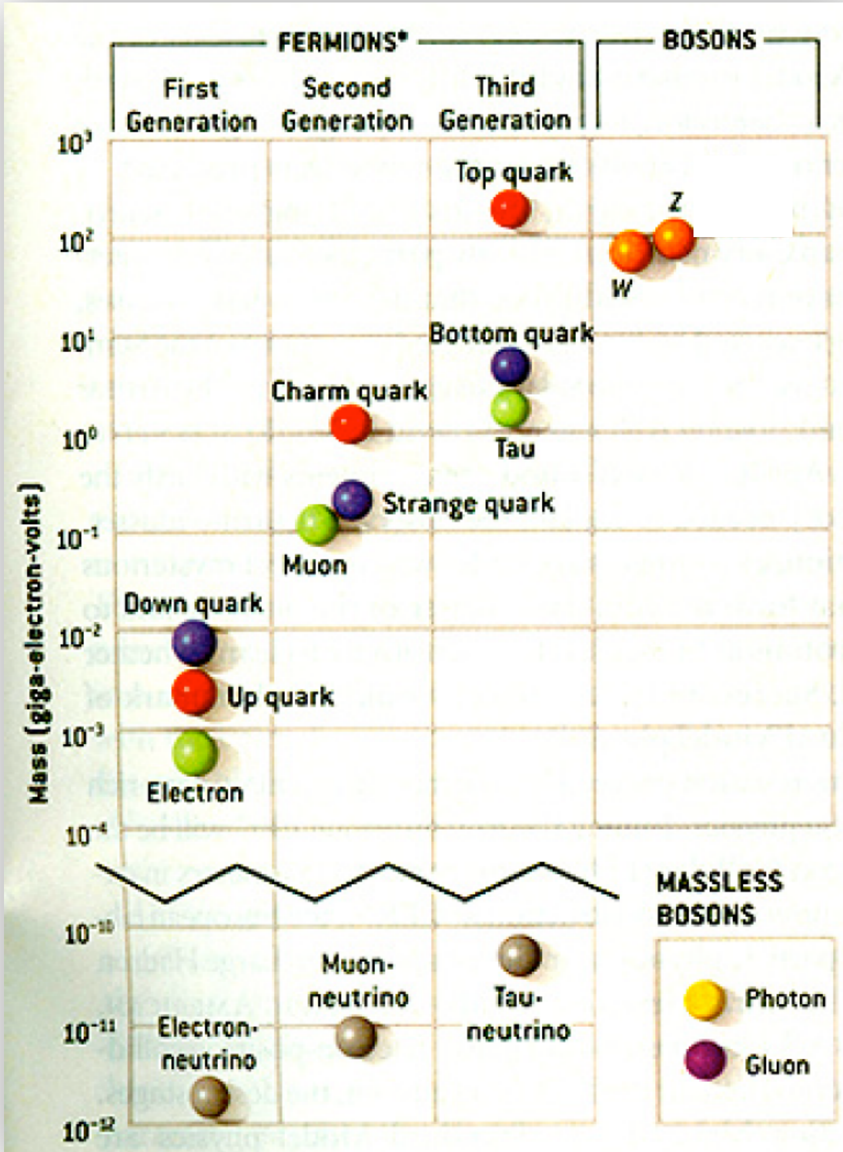
01.08.2012: ATLAS and CMS submit Higgs-search papers



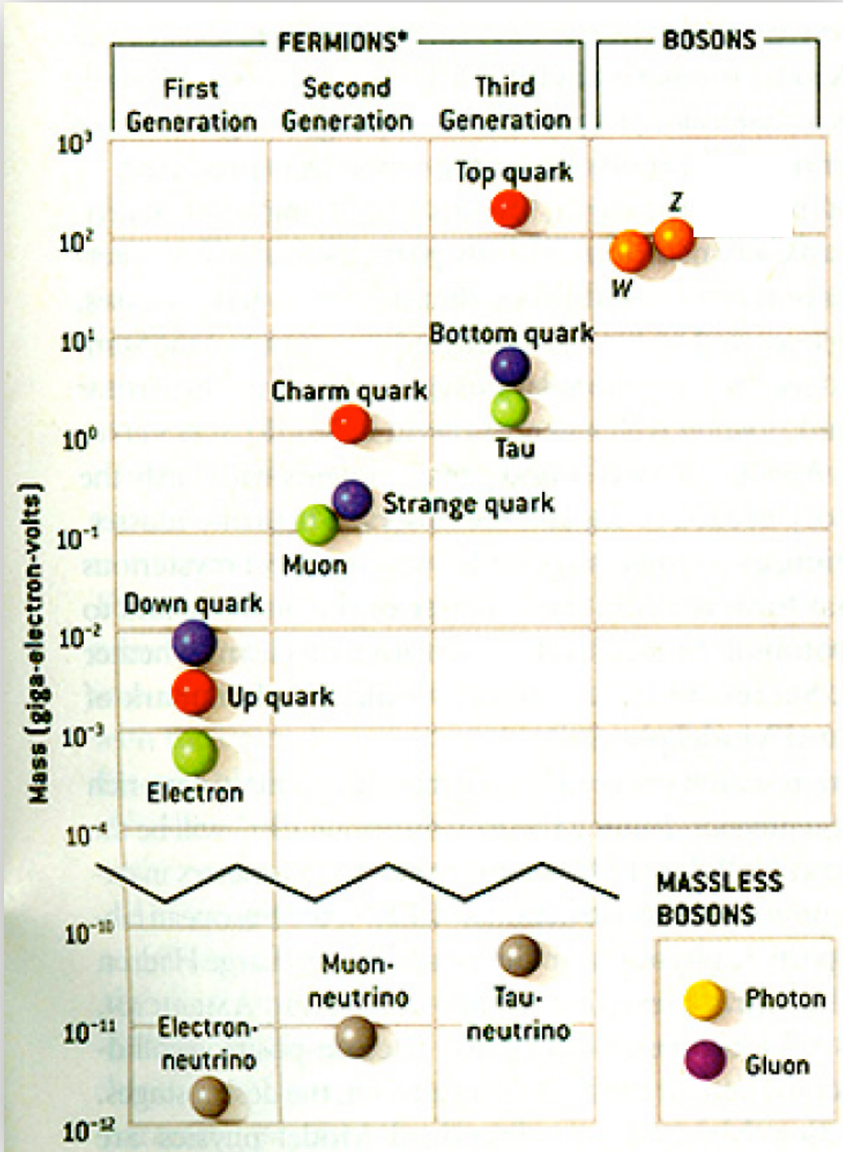
Protons collide in the CMS detector at 8 TeV, forming Z bosons which decay into electrons (green lines) and muons (red). Such an event is compatible with the decay of a Standard Model Higgs boson (Image: CMS)

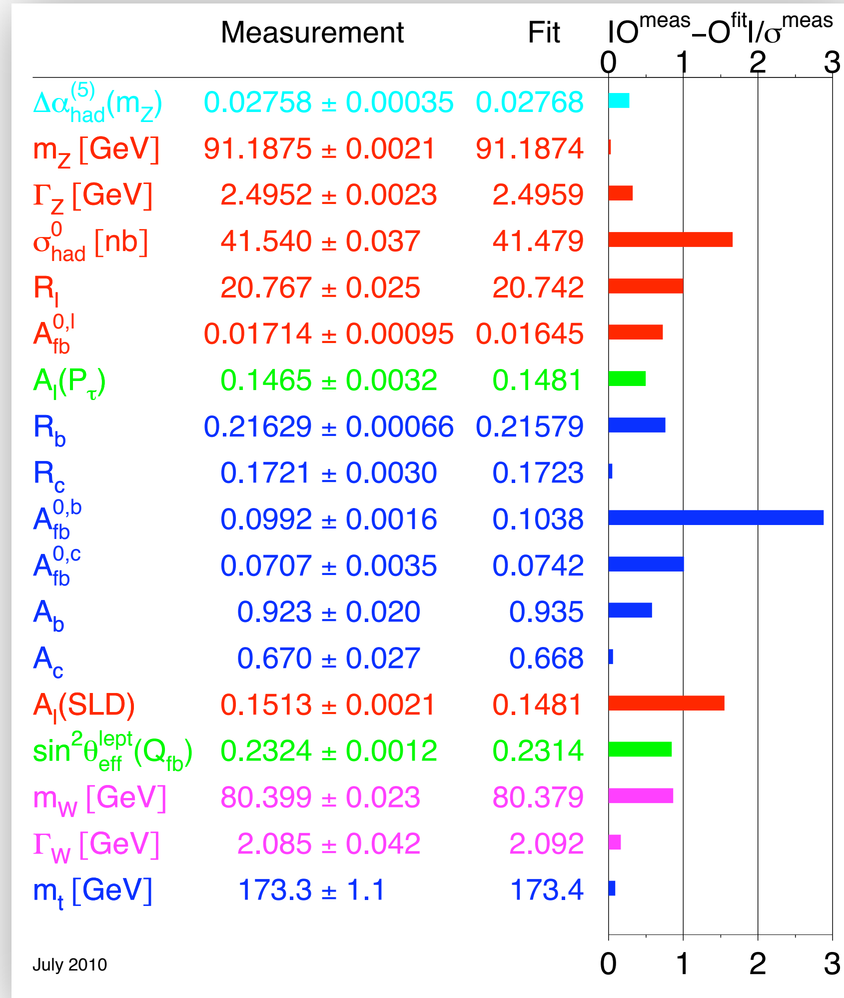
The ATLAS and CMS collaborations today submitted papers to the journal *Physics Letters B* outlining the latest on their searches for the Higgs boson. The teams report even stronger evidence for the presence of a new Higgs-like particle than announced on 4 July.

The Standard Model

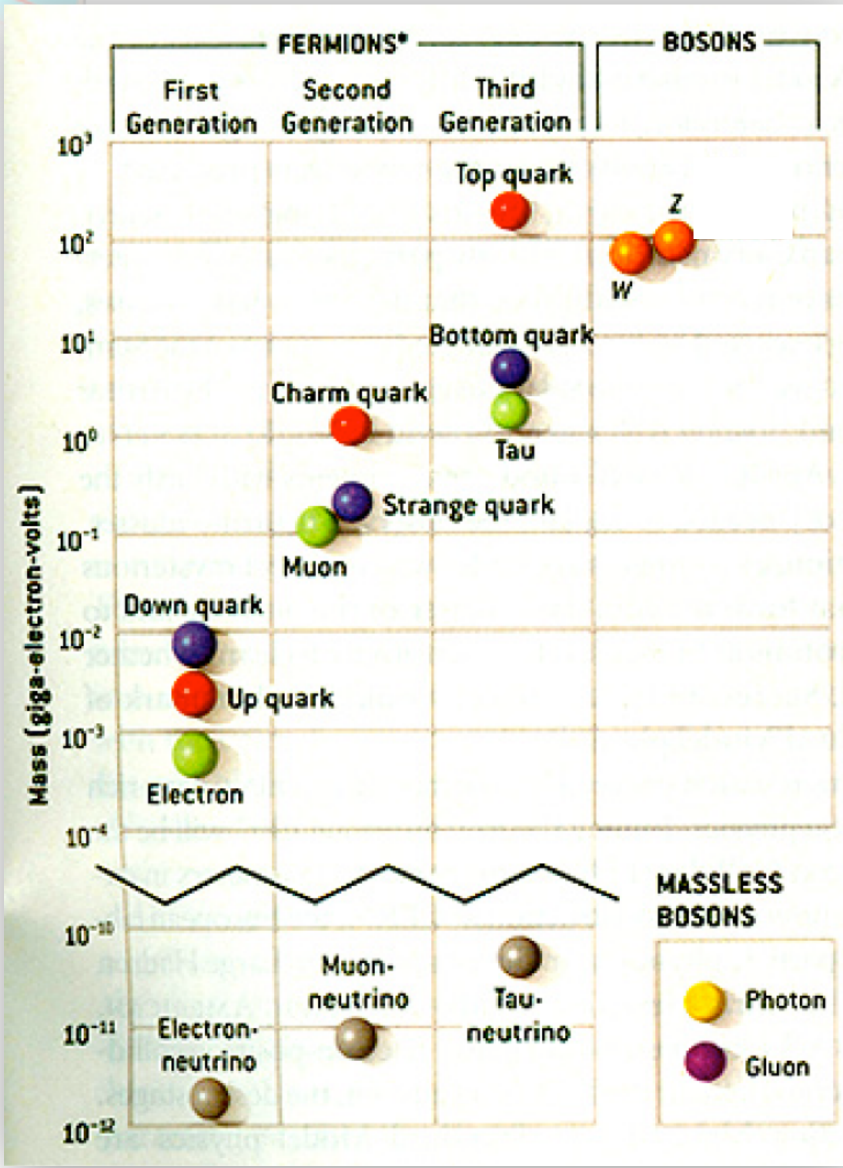


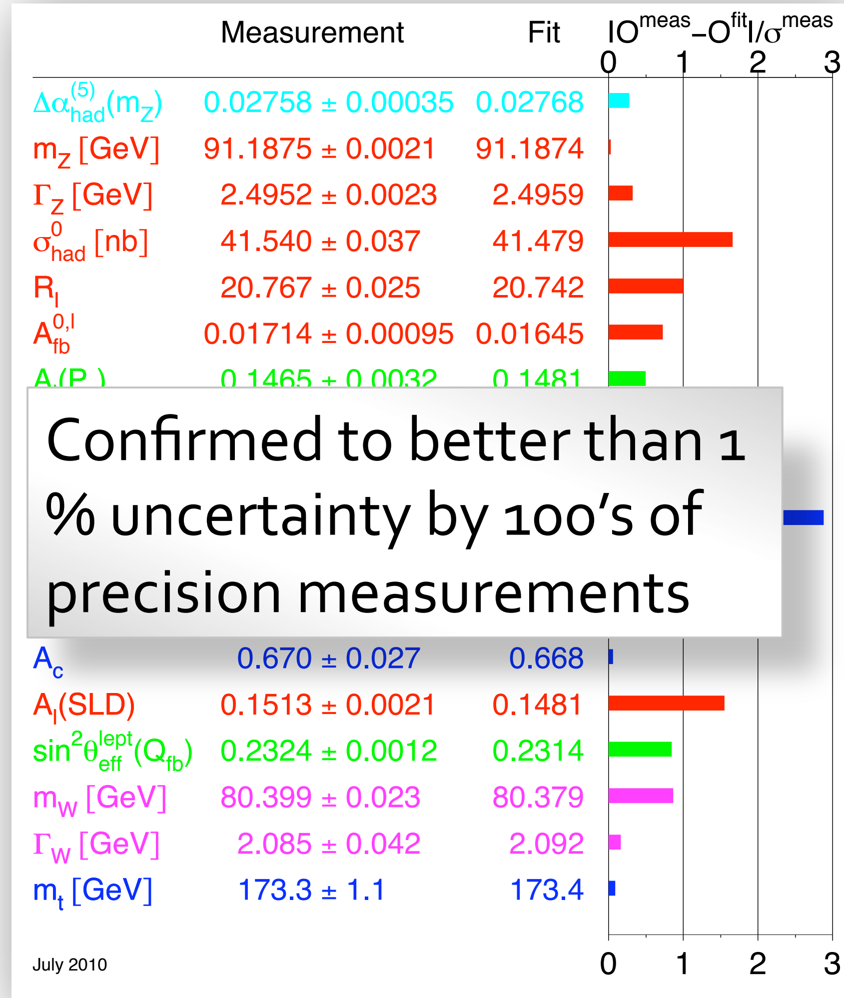
The Standard Model





The Standard Model

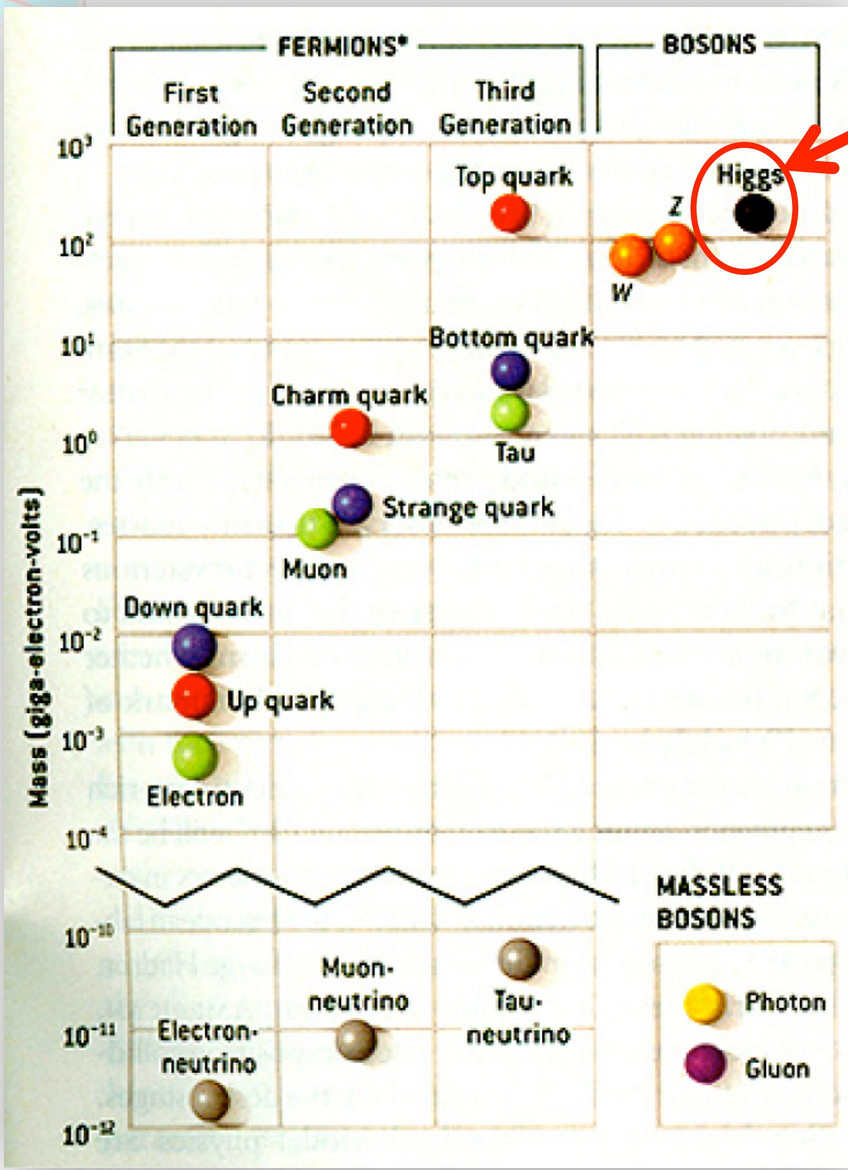


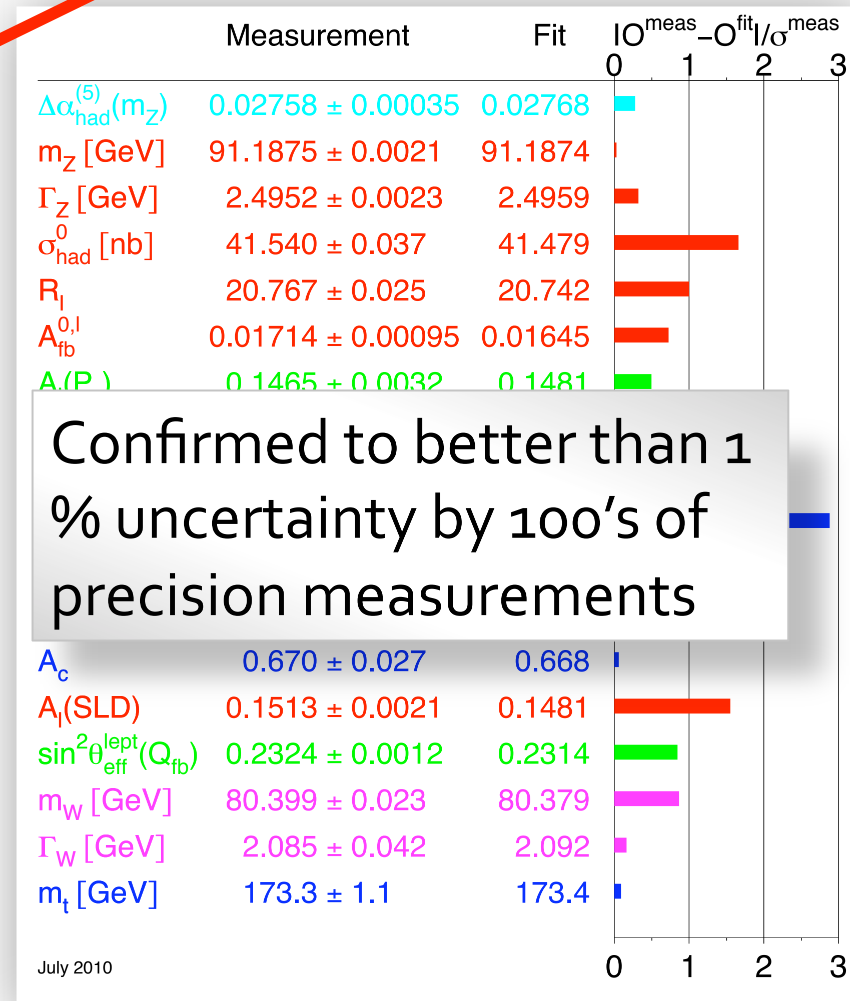


Confirmed to better than 1 % uncertainty by 100's of precision measurements

The Standard Model

1 Missing piece: Higgs





Confirmed to better than 1 % uncertainty by 100's of precision measurements

Higgs mass: theoretical constraints

- Problem: Higgs mass is free parameter

$$M_H^2 = 2\lambda v^2 \quad \ldots\ldots \quad v = 246\ \text{GeV}$$

- Theoretical constraints
  - Unitarity (no probabilities > 1)

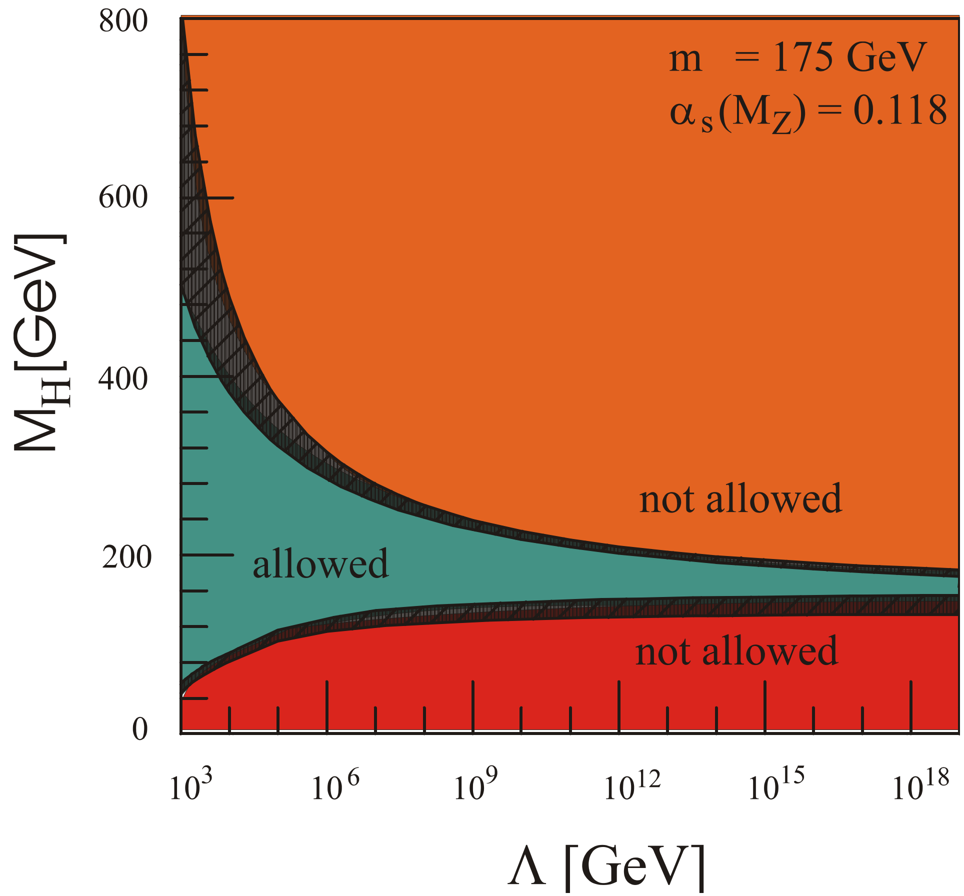
$$M_{\mathrm{H}} < 700 - 800\ \text{GeV}$$

  - Triviality
    (Higgs self coupling remains finite)

$$M_H^2 < \frac{4\pi v^2}{3\ln(\Lambda/v)}$$

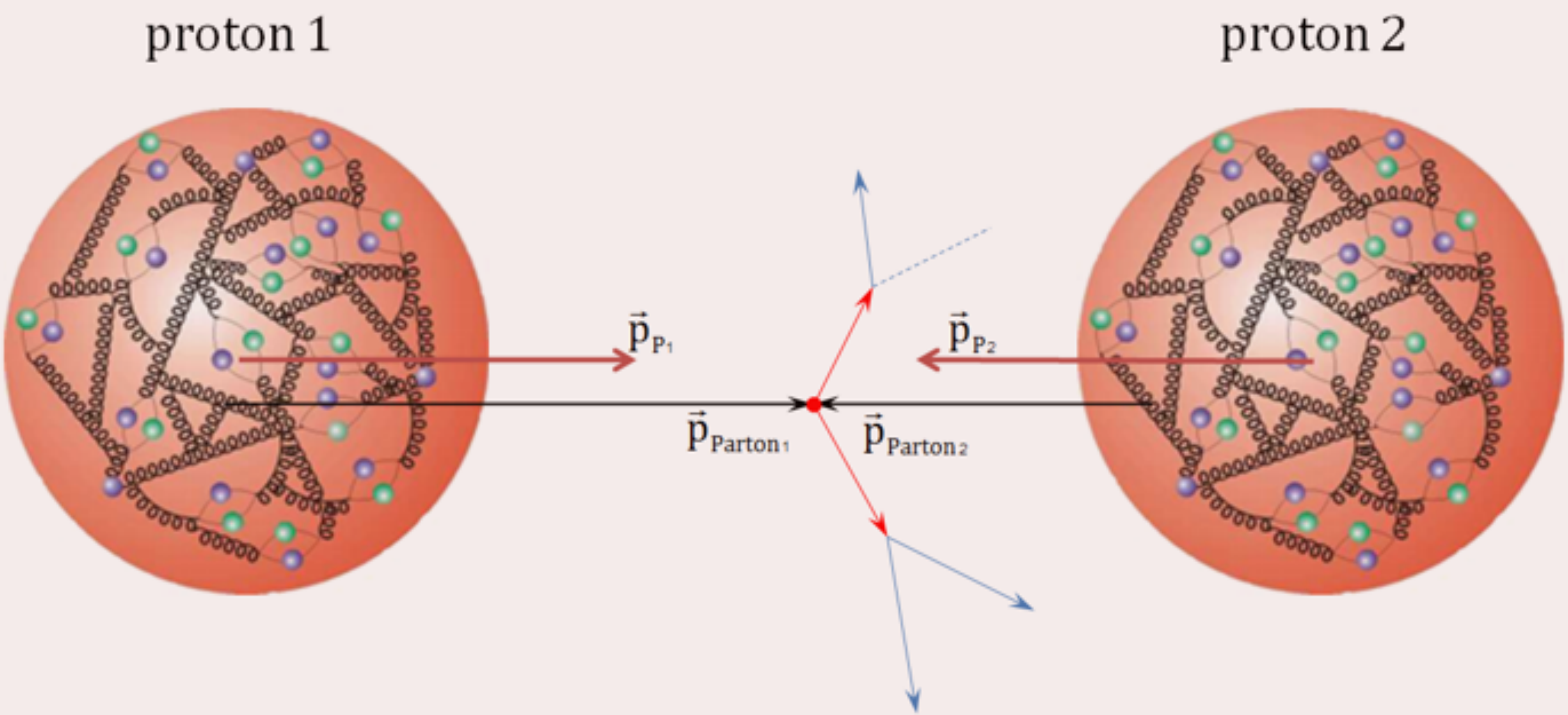
  - Stability (of vacuum)

$$M_H^2 > \frac{4 m_Z^4}{\pi^2 v^2}\ln(\Lambda/v)$$



cut-off scale

Interactions of constituents of the colliding protons, the so called partons (quarks, gluons)



$\vec{p}_{P_1}$ ... momentum proton 1

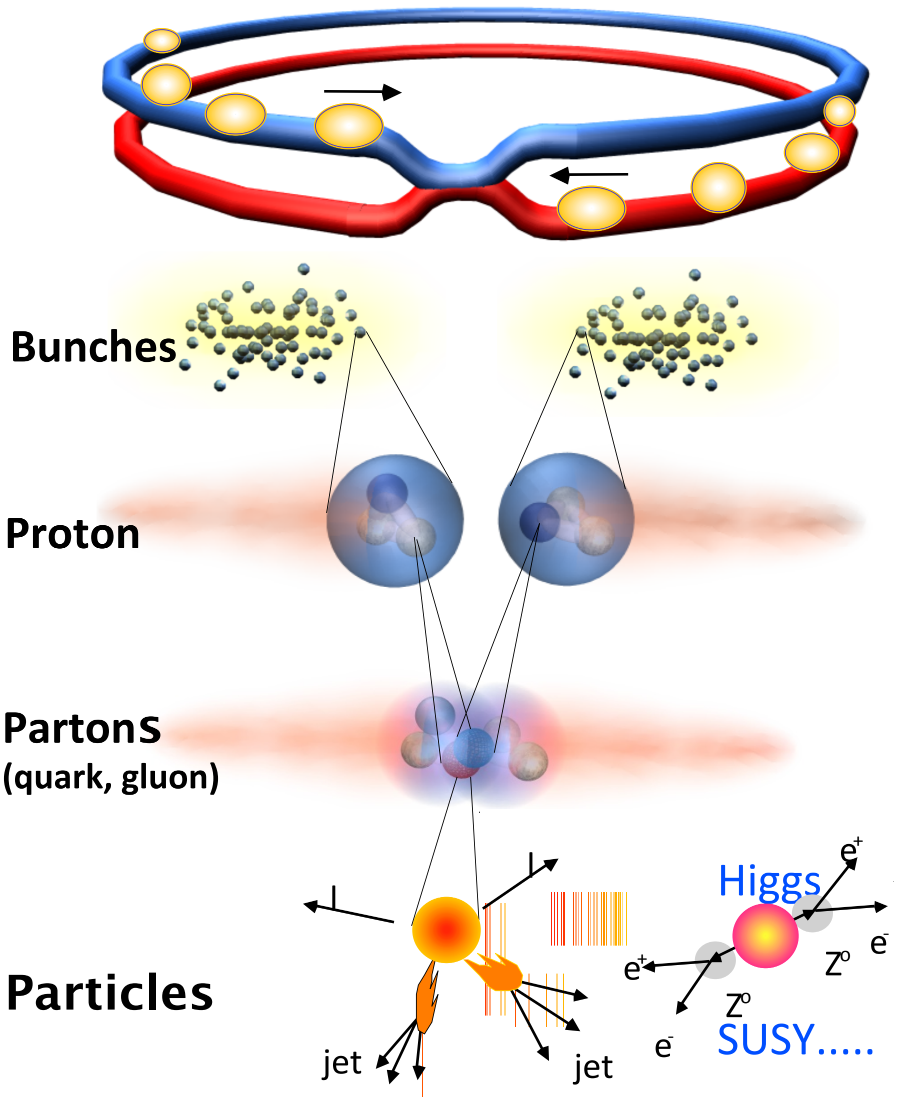
$\vec{p}_{P_1}$ ... momentum proton 2

• interaction vertex

$\vec{p}_{Parton_1}$ ... momentum parton 1

$\vec{p}_{Parton_2}$ ... momentum parton 2

Collisions in LHC

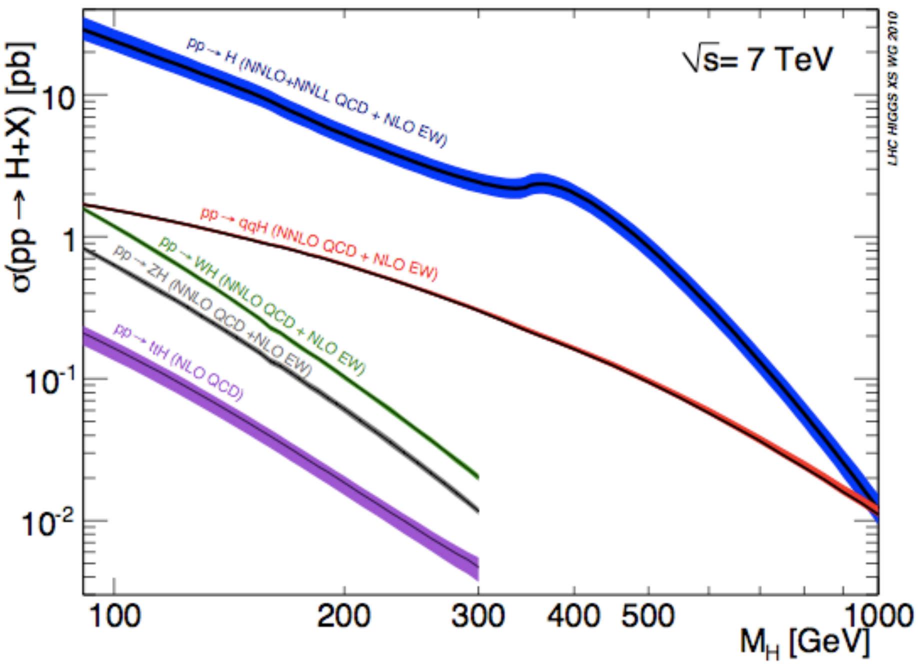


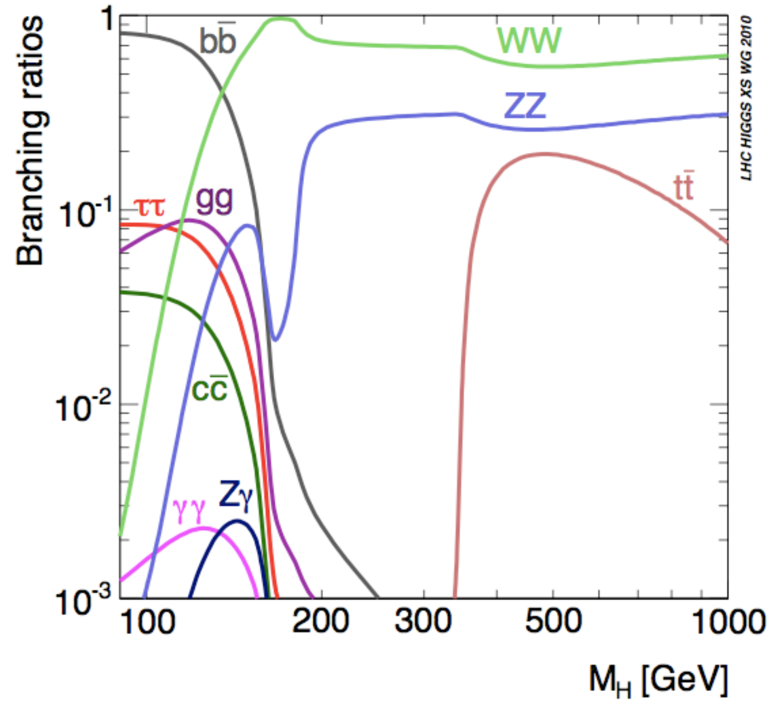
Proton - Proton	~1300 bunches/beam
Protons/bunch	10^{11}
Beam energy	4 TeV ($4x10^{12}$ eV)
Luminosity	10^{34} cm $^{-2}$ s $^{-1}$
Bunch collision frequency	20 MHz
Proton collision frequency	10^{7} - 10^{9} Hz

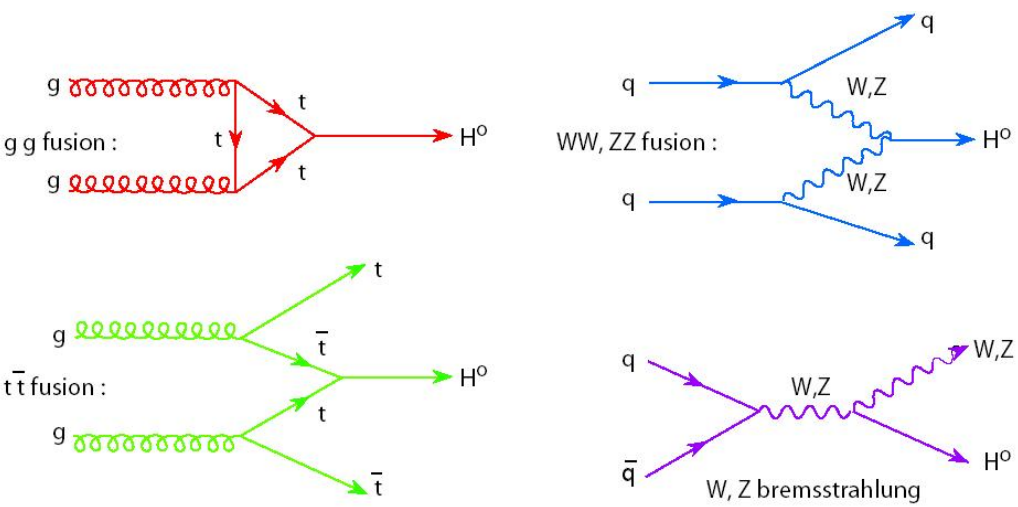
“New physics” frequency .00001 Hz

Event selection:
1 u 10 000 000 000 000

Higgs boson at LHC

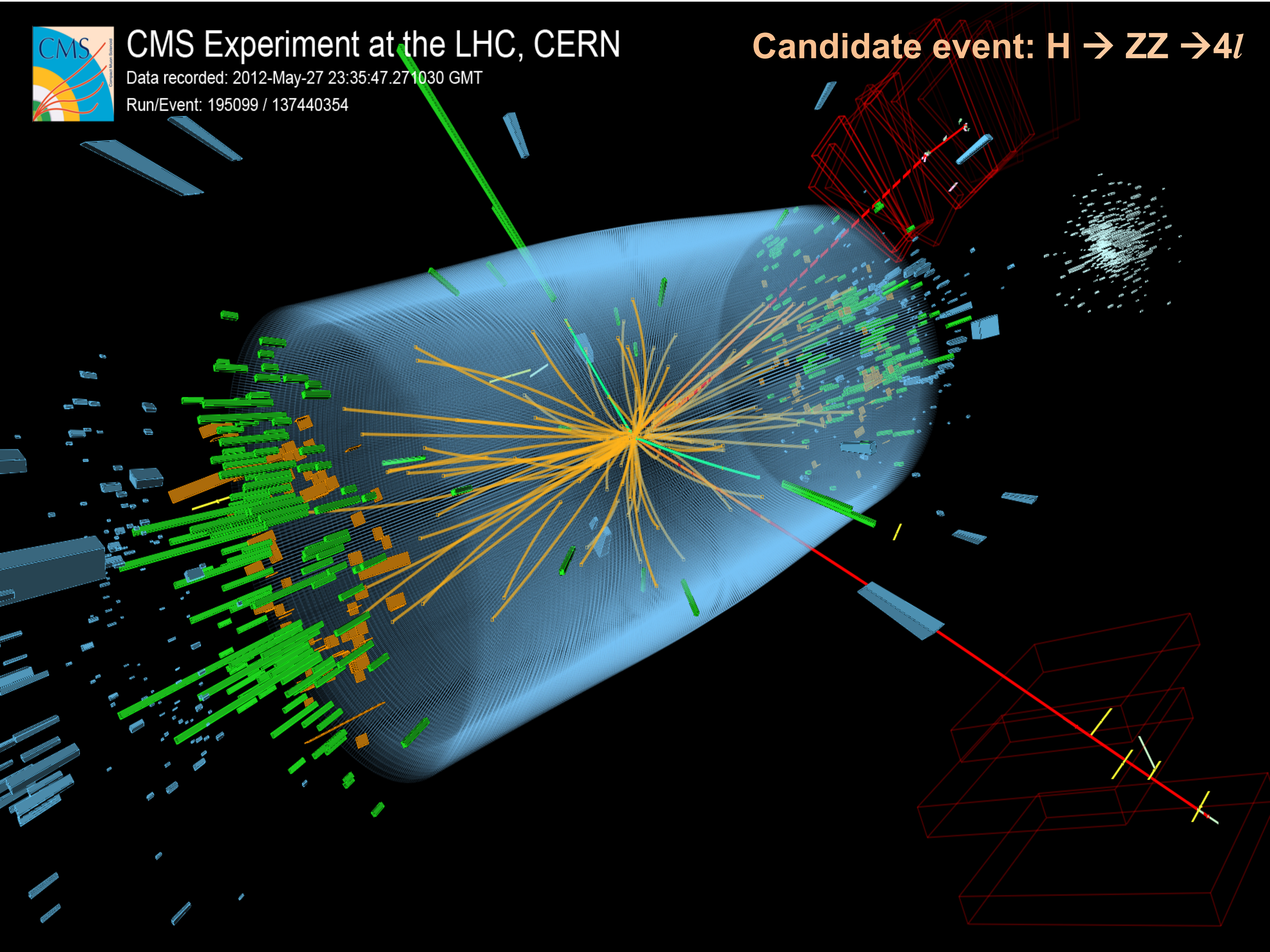




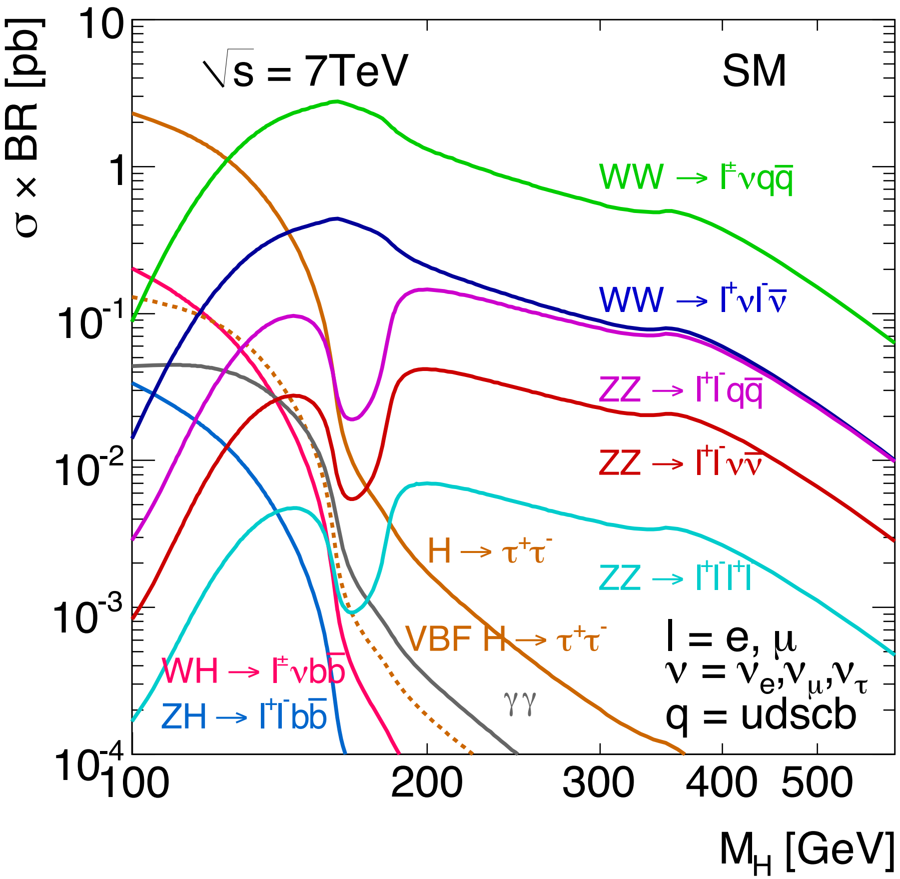


Higgs boson (M_H ~120 GeV) produced every **~10 seconds** @ L=5x10^{33} cm^{-2} s^{-1}

– If it exists ☺


CMS Experiment at the LHC, CERN
Data recorded: 2012-May-27 23:35:47.271030 GMT
Run/Event: 195099 / 137440354
Candidate event: H → ZZ →4l

Higgs boson: decay channels



Decay channel	Mass region
$H \rightarrow \gamma\gamma$	110-150
$H \rightarrow bb$	110-135
$H \rightarrow \tau\tau$	110-140
$H \rightarrow WW \rightarrow 2l\ 2\nu$	110-600
$H \rightarrow ZZ \rightarrow 4l$	110-600
$H \rightarrow ZZ \rightarrow 2l2\tau$	180-600
$H \rightarrow ZZ \rightarrow 2l2j$	226-600
$H \rightarrow ZZ \rightarrow 2l2\nu$	250-600

Signal at 1 fb^{-1}

m_H, GeV	$WW \to 2l2\nu$	$ZZ \to 4l$	$\gamma\gamma$
120	127	1.5	43
150	390	4.6	16
300	89	3.8	0.04

The most sensitive channels for low mass Higgs:

$H \rightarrow \gamma\gamma$

$H \rightarrow ZZ \rightarrow l^{-}l^{+}l^{-}l^{+}$

arXiv:1207.7235v1 [hep-ex] 31 Jul 2012

Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC

The CMS Collaboration*

Abstract

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to 5.1 fb^{-1} at 7 TeV and 5.3 fb^{-1} at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, WW, $\tau^+\tau^-$, and bb. An excess of events is observed above the expected background, a local significance of 5.0 standard deviations, at a mass near 125 GeV, signalling the production of a new particle. The expected significance for a standard model Higgs boson of that mass is 5.8 standard deviations. The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ; a fit to these signals gives a mass of $125.3 \pm 0.4\,(\text{stat.}) \pm 0.5\,(\text{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one.

This paper is dedicated to the memory of our colleagues who worked on CMS but have since passed away.

In recognition of their many contributions to the achievement of this observation.

Submitted to Physics Letters B

CERN-PH-EP-2012-218

Submitted to: Physics Letters B

arXiv:1207.7214v1 [hep-ex] 31 Jul 2012

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\mathrm{stat}) \pm 0.4\,(\mathrm{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Expectations vs measurements

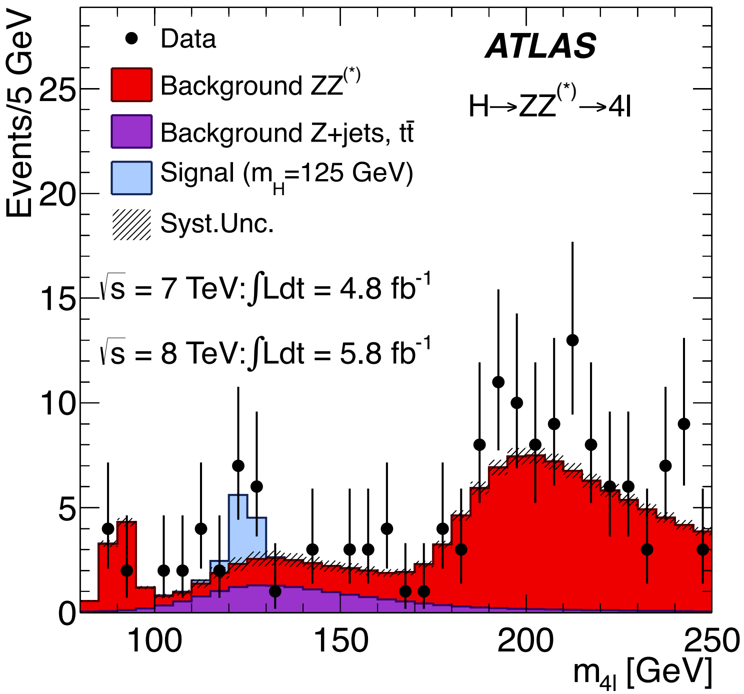


Figure 2: The distribution of the four-lepton invariant mass, $m_{4\ell}$, for the selected candidates, compared to the background expectation in the 80–250 GeV mass range, for the combination of the $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV data. The signal expectation for a SM Higgs with $m_H = 125$ GeV is also shown.

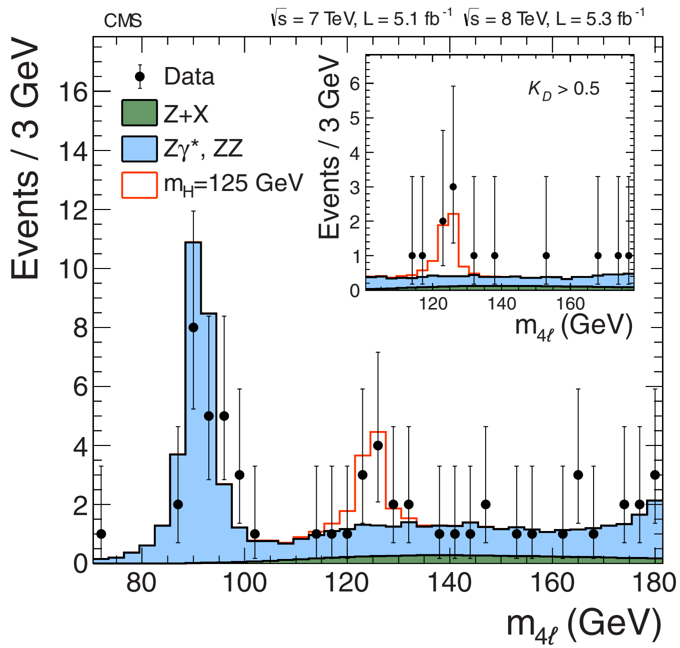
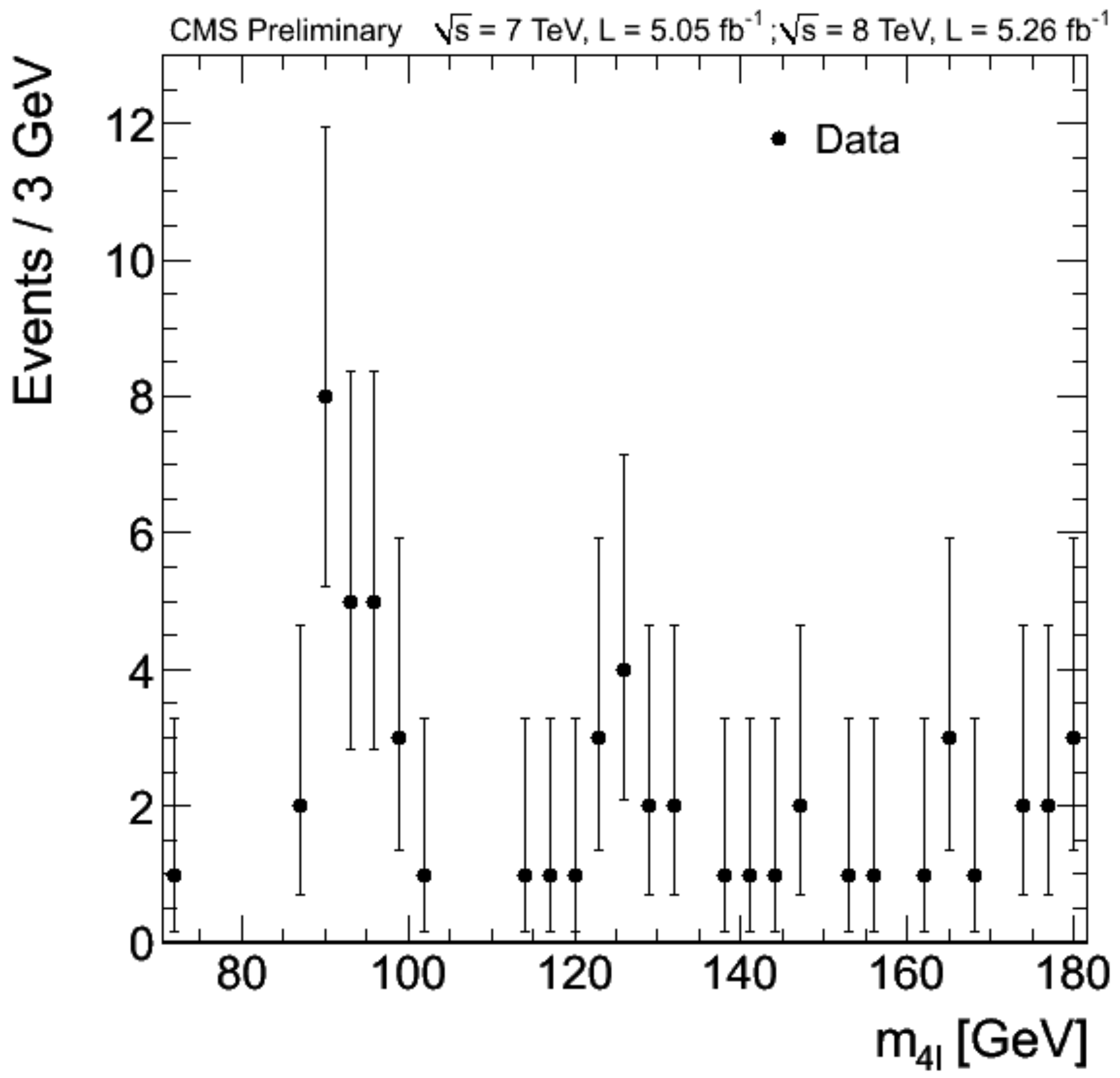
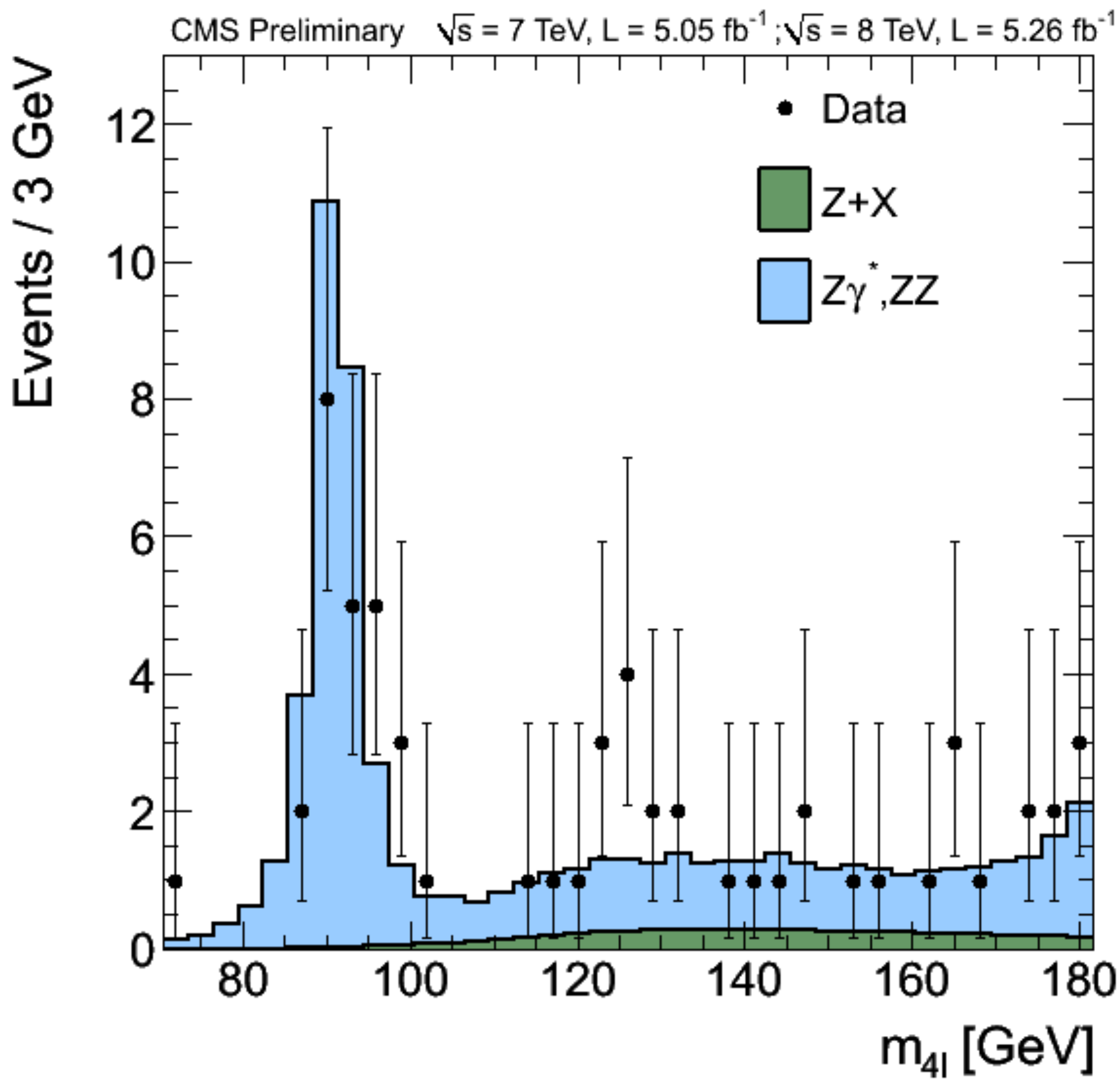


Figure 4: Distribution of the four-lepton invariant mass for the ZZ $\to 4\ell$ analysis. The points represent the data, the filled histograms represent the background, and the open histogram shows the signal expectation for a Higgs boson of mass $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, added to the background expectation. The inset shows the $m_{4\ell}$ distribution after selection of events with $K_D > 0.5$, as described in the text.

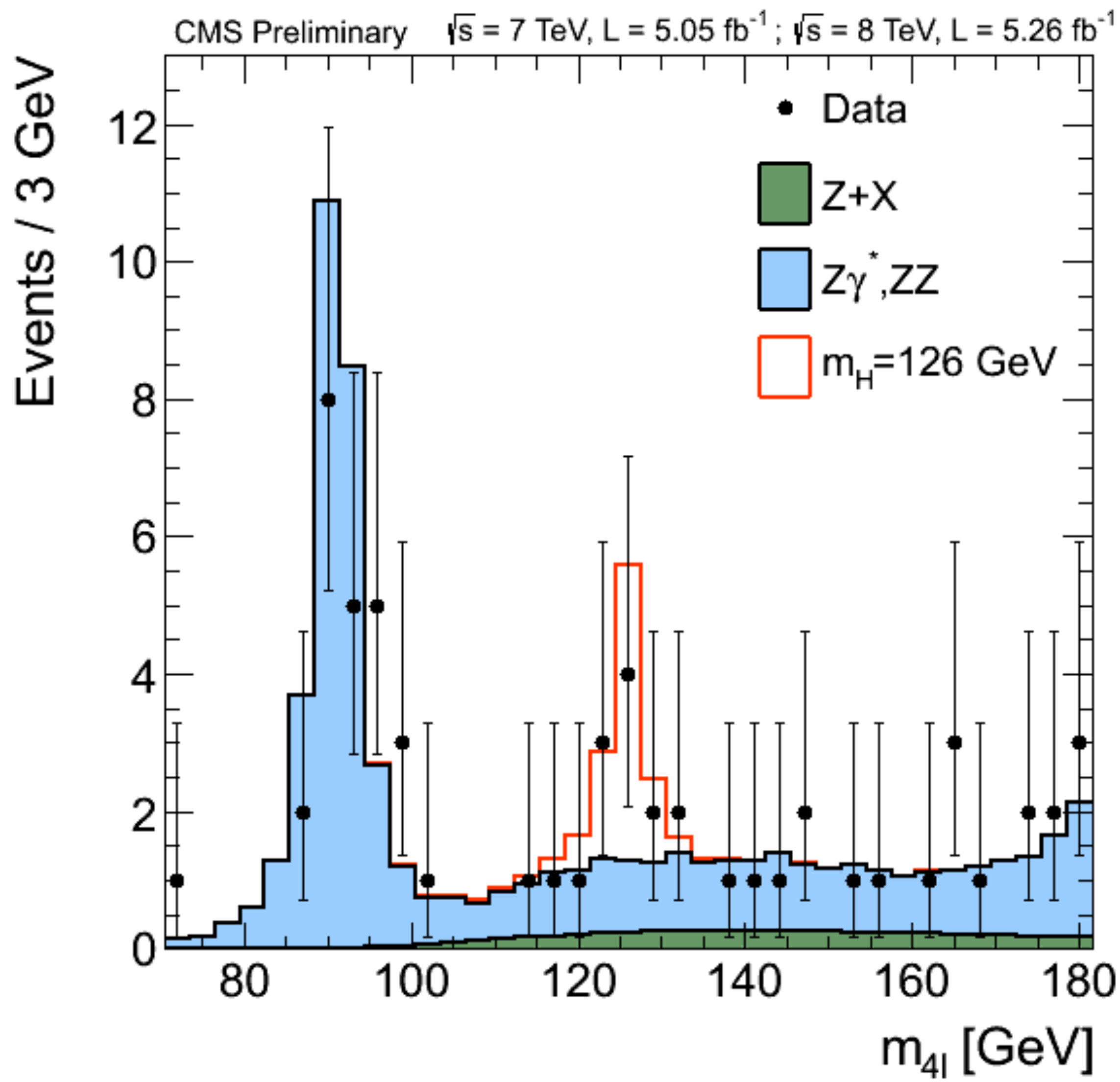
H → ZZ → $l^- l^+ l^- l^+$ events distribution



H → ZZ → $l^-l^+l^-l^+$ events distribution



H → ZZ → $l^- l^+ l^- l^+$ events distribution



H→γγ: Example of fitting

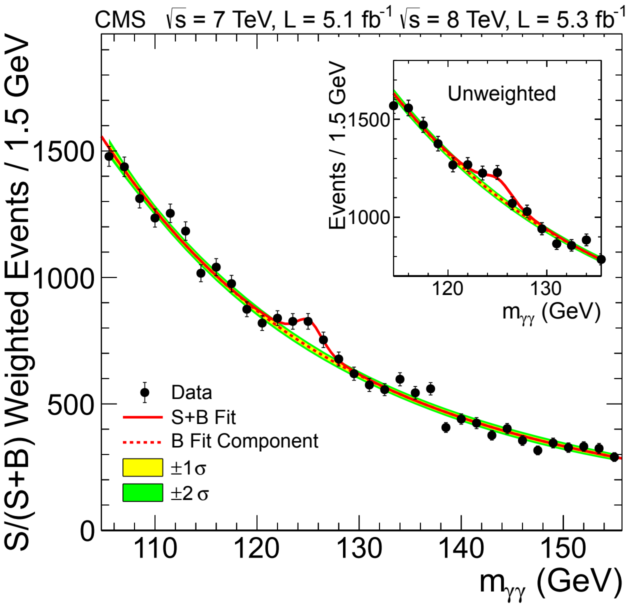


Figure 3: The diphoton invariant mass distribution with each event weighted by the $S/(S+B)$ value of its category. The lines represent the fitted background and signal, and the coloured bands represent the ± 1 and ± 2 standard deviation uncertainties on the background estimate. The inset shows the central part of the unweighted invariant mass distribution.

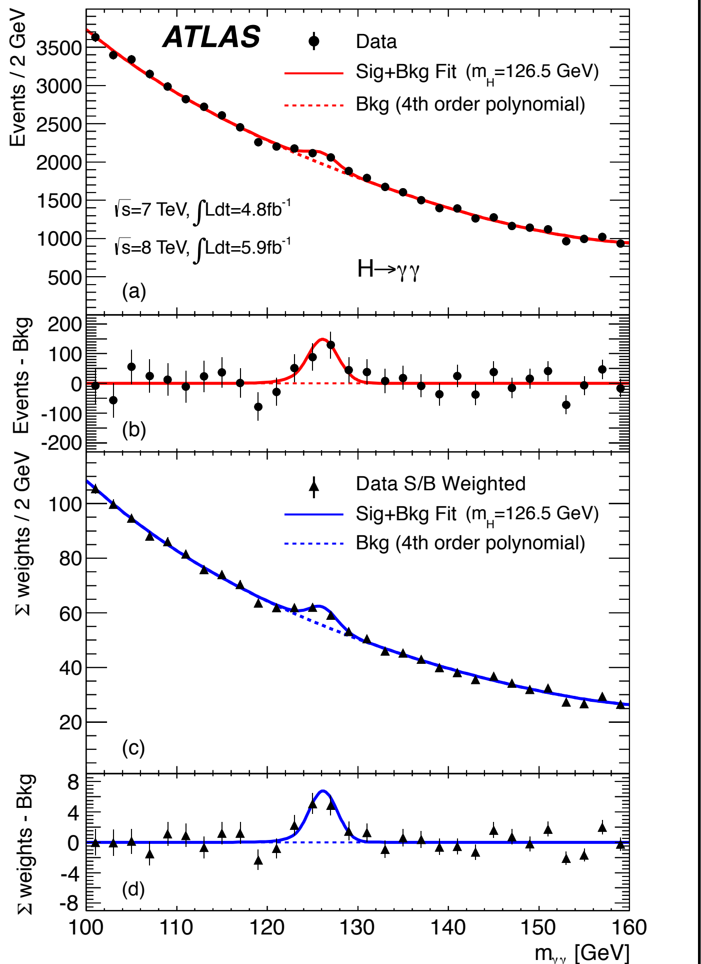


Figure 4: The distributions of the invariant mass of diphoton candidates after all selections for the combined 7 TeV and 8 TeV data sample. The inclusive sample is shown in a) and a weighted version of the same sample in c); the weights are explained in the text. The result of a fit to the data of the sum of a signal component fixed to $m_H = 126.5$ GeV and a background component described by a fourth-order Bernstein polynomial is superimposed. The residuals of the data and weighted data with respect to the respective fitted background component are displayed in b) and d).

H→bb: example of Multivariate analysis (MVA)

For the multivariate analysis, a boosted decision tree (BDT) [115, 116] is trained to give a high output value (score) for signal-like events and for events with good diphoton invariant mass resolution, based on the following observables: (i) the photon quality determined from electromagnetic shower shape and isolation variables; (ii) the expected mass resolution; (iii) the per-event estimate of the probability of locating the diphoton vertex within 10 mm of its true location along the beam direction; and (iv) kinematic characteristics of the photons and the diphoton system. The kinematic variables are constructed so as to contain no information about the invariant mass of the diphoton system. The diphoton events not satisfying the dijet selec-

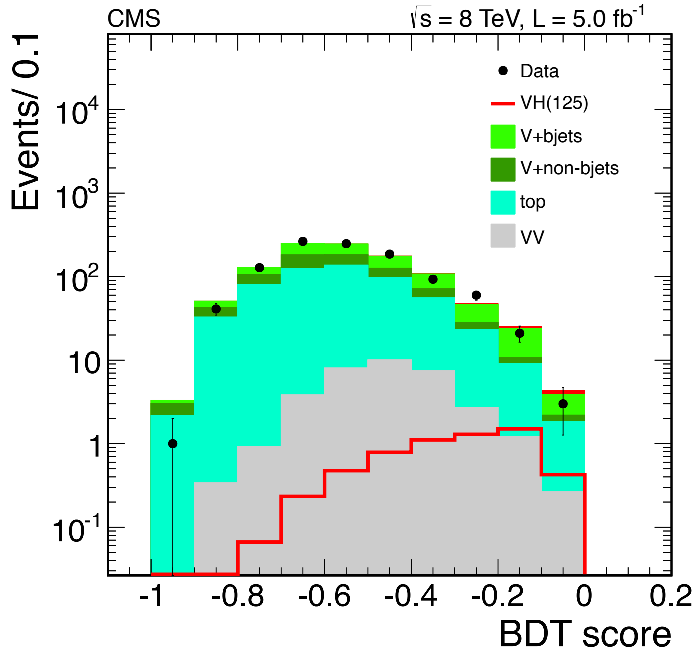


Figure 11: Distribution of BDT scores for the high-p_T subchannel of the $Z(\nu\nu)H(bb)$ search in the 8 TeV data set after all selection criteria have been applied. The signal expected from a Higgs boson (m_H = 125 GeV), including $W(\ell\nu)H$ events where the charged lepton is not reconstructed, is shown added to the background and also overlaid for comparison with the diboson background.

Example of limits

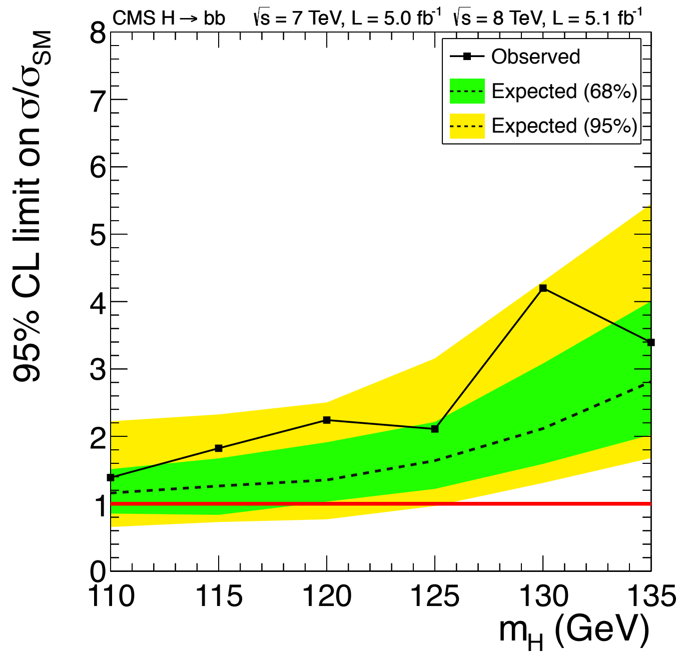


Figure 12: The 95% CL limit on the signal strength $\sigma/\sigma_{\mathrm{SM}}$ for a Higgs boson decaying to two b quarks, for the combined 7 and 8 TeV data sets. The symbol $\sigma/\sigma_{\mathrm{SM}}$ denotes the production cross section times the relevant branching fractions, relative to the SM expectation. The background-only expectations are represented by their median (dashed line) and by the 68% and 95% CL bands.

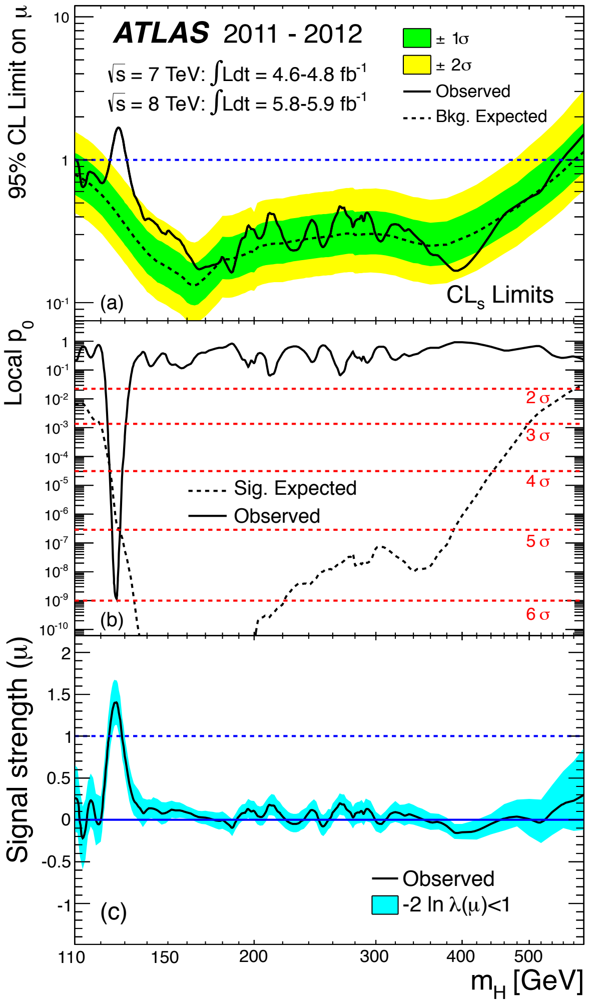


Figure 7: Combined search results: (a) The observed (solid) 95% CL limits on the signal strength as a function of m_H and the expectation (dashed) under the background-only hypothesis. The dark and light shaded bands show the $\pm1\sigma$ and $\pm2\sigma$ uncertainties on the background-only expectation. (b) The observed (solid) local p_0 as a function of m_H and the expectation (dashed) for a SM Higgs boson signal hypothesis ($\mu = 1$) at the given mass. (c) The best-fit signal strength $\hat{\mu}$ as a function of m_H. The band indicates the approximate 68% CL interval around the fitted value.

p-value and hypothesis testing

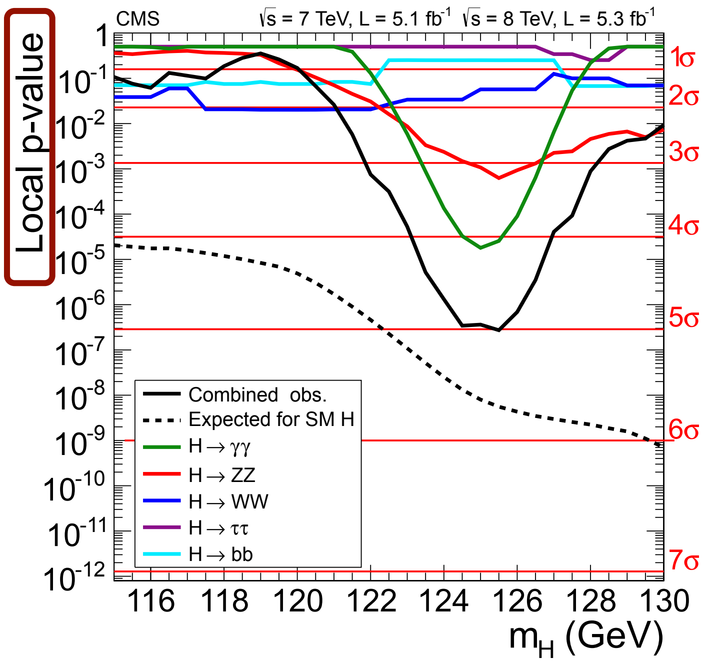


Figure 15: The observed local p-value for the five decay modes and the overall combination as a function of the SM Higgs boson mass. The dashed line shows the expected local p-values for a SM Higgs boson with a mass m_{H}.

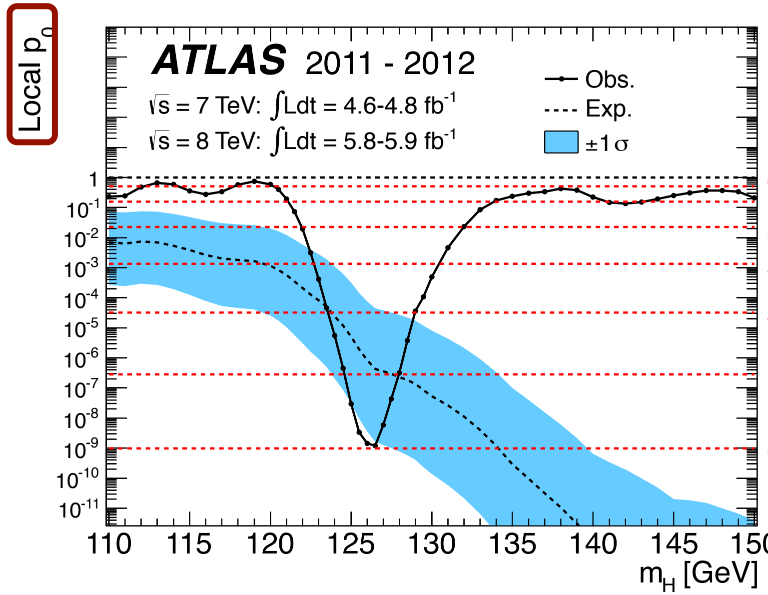


Figure 9: The observed (solid) local p_0 as a function of m_H in the low mass range. The dashed curve shows the expected local p_0 under the hypothesis of a SM Higgs boson signal at that mass with its $\pm 1\sigma$ band. The horizontal dashed lines indicate the p-values corresponding to significances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

Measuring properties

Asymptotically, the test statistic $-2\ln\lambda(\mu, m_H)$ is distributed as a χ^2 distribution with two degrees of freedom. The resulting 68% and 95% CL contours for the $H\to\gamma\gamma$ and $H\to WW^{(*)}\to \ell\nu\ell\nu$ channels are shown in

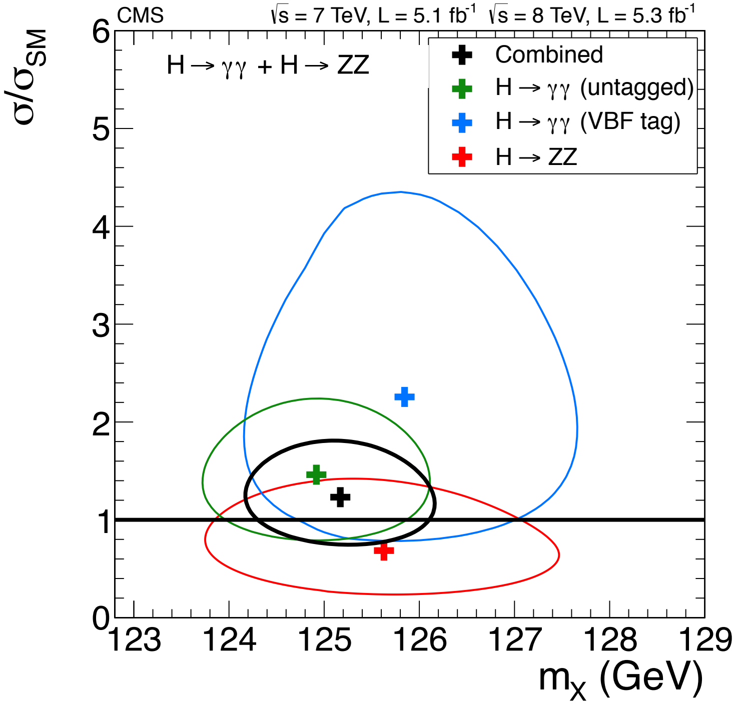


Figure 17: The 68% CL contours for the signal strength $\sigma/\sigma_{\rm SM}$ versus the boson mass m_X for the untagged $\gamma\gamma$, $\gamma\gamma$ with VBF-like dijet, 4ℓ, and their combination. The symbol $\sigma/\sigma_{\rm SM}$ denotes the production cross section times the relevant branching fractions, relative to the SM expectation. In this combination, the relative signal strengths for the three decay modes are constrained by the expectations for the SM Higgs boson.

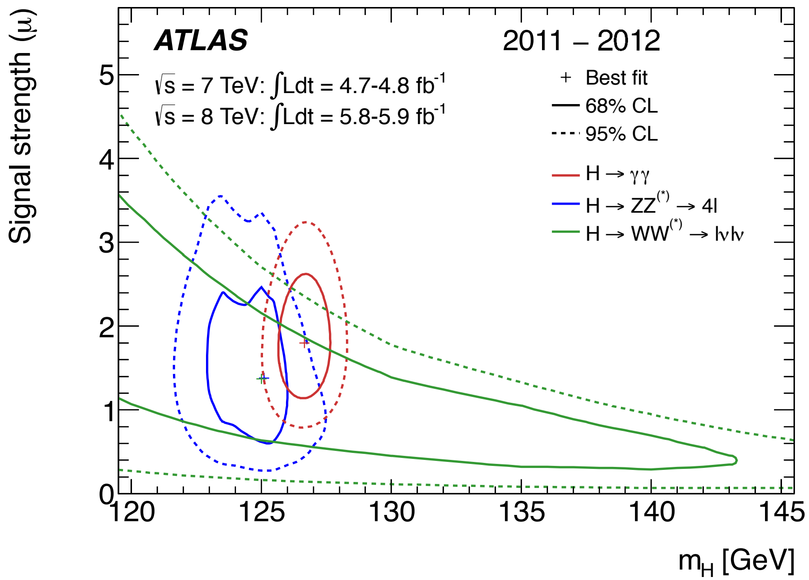
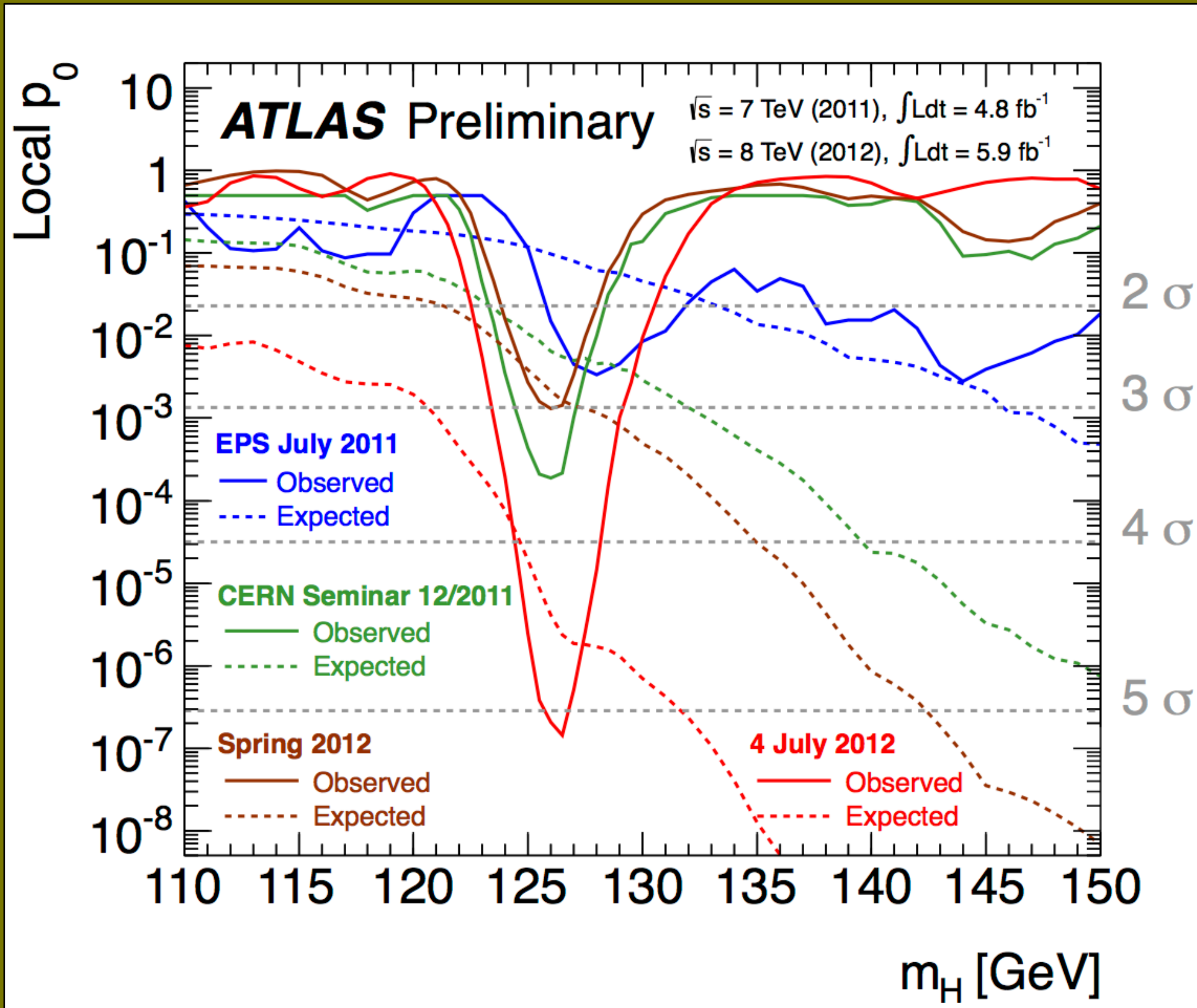


Figure 11: Confidence intervals in the (μ, m_H) plane for the $H\to ZZ^{(*)}\to 4\ell$, $H\to\gamma\gamma$, and $H\to WW^{(*)}\to\ell\nu\ell\nu$ channels, including all systematic uncertainties. The markers indicate the maximum likelihood estimates $(\hat{\mu}, \hat{m}_H)$ in the corresponding channels (the maximum likelihood estimates for $H\to ZZ^{(*)}\to 4\ell$ and $H\to WW^{(*)}\to\ell\nu\ell\nu$ coincide).

Evolution of the excess with time



Energy-scale systematics not included

Conclusions of papers – ATLAS

10. Conclusion

Searches for the Standard Model Higgs boson have been performed in the $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ channels with the ATLAS experiment at the LHC using 5.8–5.9 fb^{-1} of pp collision data recorded during April to June 2012 at a centre-of-mass energy of 8 TeV. These results are combined with earlier results [17], which are based on an integrated luminosity of 4.6–4.8 fb^{-1} recorded in 2011 at a centre-of-mass energy of 7 TeV, except for the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels, which have been updated with the improved analyses presented here.

The Standard Model Higgs boson is excluded at 95% CL in the mass range 111–559 GeV, except for the narrow region 122–131 GeV. In this region, an excess of events with significance 5.9 σ, corresponding to $p_0 = 1.7 \times 10^{-9}$, is observed. The excess is driven by the two channels with the highest mass resolution, $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$, and the equally sensitive but low-resolution $H \to WW^{(*)} \to \ell\nu\ell\nu$ channel. Taking into account the entire mass range of the search, 110–600 GeV, the global significance of the excess is 5.1 σ, which corresponds to $p_0 = 1.7 \times 10^{-7}$.

These results provide conclusive evidence for the discovery of a new particle with mass 126.0 $\pm$ 0.4 (stat) $\pm$ 0.4 (sys) GeV. The signal strength parameter μ has the value 1.4 $\pm$ 0.3 at the fitted mass, which is consistent with the SM Higgs boson hypothesis $\mu = 1$. The decays to pairs of vector bosons whose net electric charge is zero identify the new particle as a neutral boson. The observation in the diphoton channel disfavours the spin-1 hypothesis [140, 141]. Although these results are compatible with the hypothesis that the new particle is the Standard Model Higgs boson, more data are needed to assess its nature in detail.

Conclusions of papers – CMS

Results are presented from searches for the standard model Higgs boson in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV in the CMS experiment at the LHC, using data samples corresponding to integrated luminosities of up to $5.1\,\mathrm{fb}^{-1}$ at 7 TeV and $5.3\,\mathrm{fb}^{-1}$ at 8 TeV. The search is performed in five decay modes: $\gamma\gamma$, ZZ, $\mathrm{W^+W^-}$, $\tau^+\tau^-$, and $\mathrm{b\bar{b}}$. An excess of events is observed above the expected background, with a local significance of $5.0\,\sigma$, at a mass near 125 GeV, signalling the production of a new particle. The expected local significance for a standard model Higgs boson of that mass is $5.8\,\sigma$. The global p-value in the search range of 115–130 (110–145) GeV corresponds to $4.6\,\sigma$ ($4.5\,\sigma$). The excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ, and a fit to these signals gives a mass of $125.3 \pm 0.4\,(\mathrm{stat.}) \pm 0.5\,(\mathrm{syst.})$ GeV. The decay to two photons indicates that the new particle is a boson with spin different from one. The results presented here are consistent, within uncertainties, with expectations for a standard model Higgs boson. The collection of further data will enable a more rigorous test of this conclusion and an investigation of whether the properties of the new particle imply physics beyond the standard model.

We'll come back to this at the end of lectures

Outline of Lecture Series

1. Introduction to data analysis
2. Monte Carlo methods
3. Distributions and estimators
4. Confidence intervals
5. Hypothesis testing

Data Analysis

Lecture 1: Introduction to data analysis

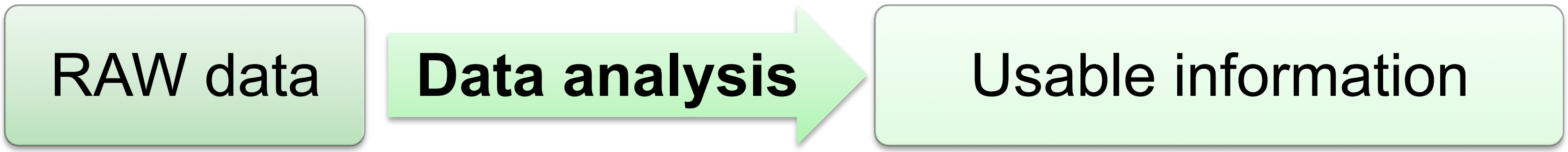
August 18, 2012

In this lecture

- **Introduction to data analysis**
  - Confirmatory and exploratory data analysis
  - Quantitative vs graphical techniques
  - Experimental vs observational studies
  - Exploring the data

Data analysis, statistics and probability

- **Data analysis** is the process of transforming raw data into usable information



- Data analysis uses **statistics** for presentation and interpretation (explanation) of data
  - ***Descriptive statistics***
    - Describes the main features of a collection of data in quantitative terms
  - ***Inductive statistics***
    - Makes ***inference*** about a random process from its observed behavior during a finite period of time
- A mathematical foundation for statistics is the **probability theory**

Confirmatory and exploratory data analysis

- **Confirmatory** data analysis = Statistical **hypothesis testing**
  - A method of making statistical decisions using experimental data
  - Two main methods
    - **Frequentist** hypothesis testing
      - Hypothesis is either true or not
    - **Bayesian** inference
      - Introduces a “degree of belief”
- **Exploratory** data analysis
  - Uses data to suggest hypothesis to test
  - Complements confirmatory data analysis
  - Main objectives:
    - Suggest hypothesis about the causes of observed phenomena
    - Asses assumptions on which statistical inference will be based
    - Select appropriate statistical tools and techniques
    - Eventually suggest further data collection

Quantitative vs graphical techniques

- **Quantitative techniques** yield numeric or tabular output
  - Hypothesis testing
  - Analysis of variance
  - Point estimation
  - Interval estimation
- **Graphical techniques**
  - Used for gaining insight into data sets in terms of testing assumptions, model selection, estimator selection ...
  - Provide a convincing mean of presenting results
  - Includes: graphs, histograms, scatter plots, probability plots, residual plots, box plots, block plots, biplots
  - Four main objectives:
    - Exploring the **content** of a data set
    - Finding **structure** in data
    - Checking **assumptions** in statistical models
    - **Communicate** the results of an analysis

Experimental vs observational studies

- **Experimental studies**

  Measure the system → Manipulate the system → Measure again and compare

  - Example: Study of whether and how much a free coffee would improve working performace of scientists in Building 40 at CERN

- **Observational studies**
  - No experimental manipulation
  - Data are gathered and analysed
  - Example:
    - Study of correlation between number of beers drunk in a pub on Wednesday evening on performance on the exam the day after
    - Be careful who pays! → see later
    - One could discuss whether to manipulate or not the system ☺

Experiments – basic steps

Planning	• Select subject to study • Select an information source
Design and Building	• Design an experiment • Build and test a model (f.g. MC simulation) • Once happy with the model build the experiment
Collecting data	• Employ descriptive statistics to summarize data • Suppres details • Early exploratory analysis
Analysing data	• **Statistical inference** • **Reach a consensus what observations tell about an underlaying reality**
Presenting Documenting	• Publish article and disseminate results • Enjoy in the fruits of the hard work!

LHC experiments – basic steps

Planning	• Started ~ 20 years ago (Aachen 1989) • Core teams from previous experiments UA1&2
Design Building	• 'Best' experimental design chosen (CMS, ATLAS, ALICE and LHCb) • Detailed MC simulations performed before started to build
Collecting data	• Trigger and DAQ carefully planed and built • MC simulation used for optimization
Analysing data	• **Statistical inference → a part of work done at this school too (learning methods&tools)** • **For the consensus → let's see ☺**
Presenting Documenting	• Many articles published • And first discoveries announced and published!

What we (will) measure at LHC?

Something we already know

- At the very beggining of the LHC operation
- For example: production of W and Z bosons

Something that (probably) exists but wasn't measured yet

- Simply because we are exploring new energy domain
- Standard Model processes
- But surprises are always possible

Hopefully something new but reasonably expected

- Altought "reasonably" is not very well defined ☺
- For example we all expect to find the Higgs boson
- Heavy neutrions?

Maybe something new but less likely

- New heavy bosons (Z′, W′)
- Micro black holes
- Extra dimensions

Something completely unexpected

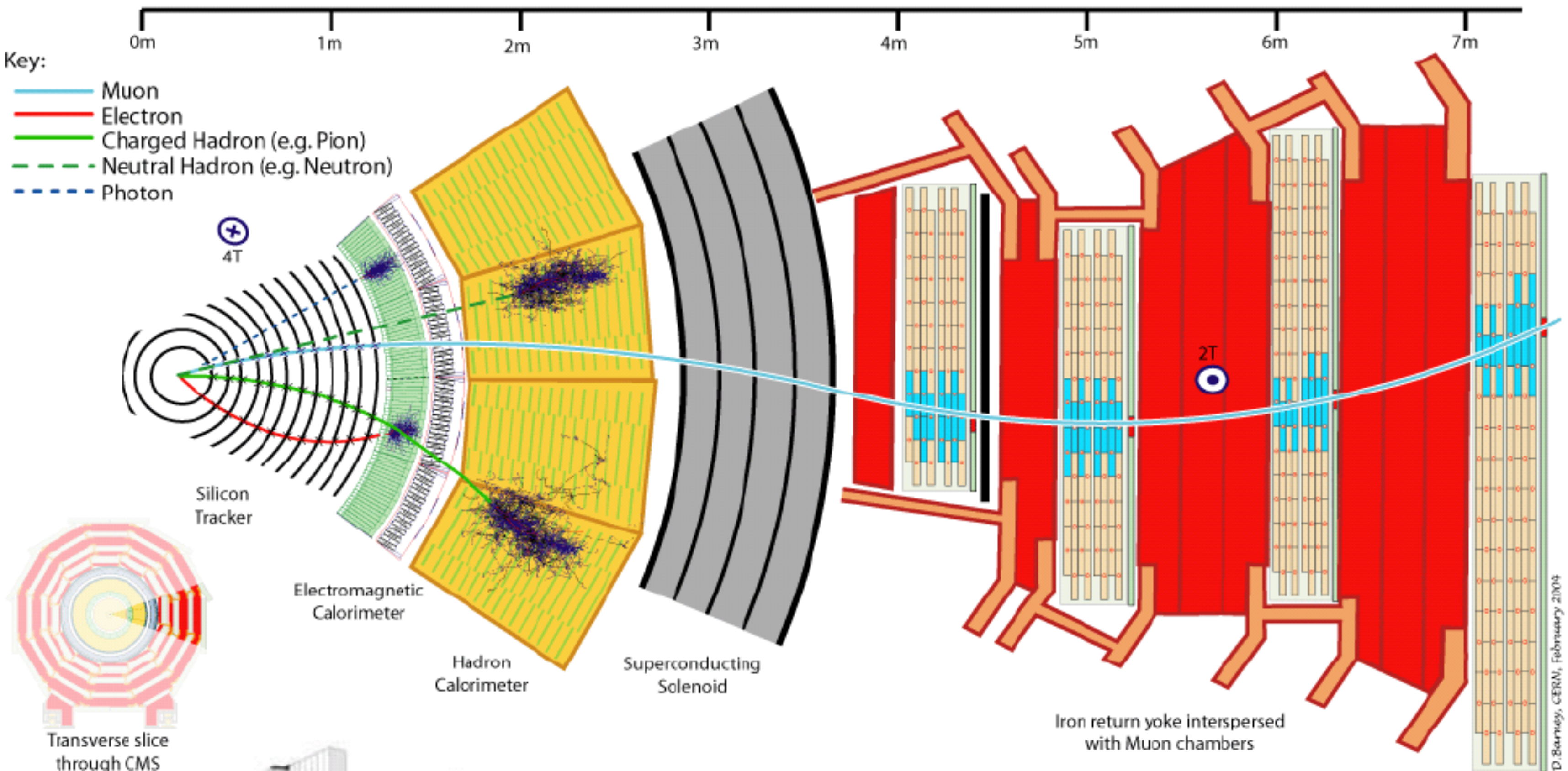
- Well, it's hard to look for unexpected ☺

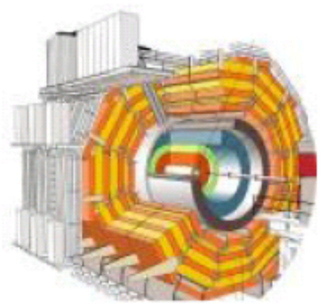
Some of the physicists' jargon

- **Cross section** (σ)
  - A measure of 'frequency' of the physical process
  - Units: barns (10^{-28} cm^2)
    - Typical values: femtobarns (fb), picobarns (pb)
- **Luminosity** (L)
  - Or *instantenous luminosity*
  - A measure of collisions 'frequency'
    - Typical (at Tevatron/Early LHC): $L = 10^{32}$ cm^{-2}s^{-1}
- **Integrated luminosity** ($\mathcal{L} = \int L dt$)
  - A measure of number of accumulated collisions after a certain time period
  - Units: (cross section)$^{-1}$ .... E.g. 1 fb^{-1} = 1000 pb^{-1}
    - Typical (Tevatron/Early LHC): few fb^{-1}
- **Number of events** (N)
  - Number of (expected) events (N) after a certain time of running

$$N = \sigma \cdot \mathcal{L}$$

Measuring physical objects





Data analysis – general picture

1 **Physical phenomena**

Described by a theory

2 **Sampling a reality**

Experiment

3 **Data sample**

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$$

4 **Data analysis**

5 **Results**

- **parameter estimates**
- **confidence limits**
- **hypothesis tests**

Described by PDFs, depending on p <u>uknown</u> parameters with <u>true values</u>

$$\boldsymbol{\theta}^{true} = (\theta_1^{true}, \theta_2^{true}, \ldots, \theta_p^{true})$$

For example:

$$\boldsymbol{\theta}^{true} = (m_H^{true}, \Delta m_s^{true}, \ldots, \sigma_{tot}^{true})$$

For example:

$$\boldsymbol{x} = (event_1, \ldots, event_N)$$

In statistics x is a multivariate random variable *(each event has many properties, all potential variables)*

Data analysis – general picture

The main goal:
learn more about NATURE

Make an experiment and obtain a DATA SAMPLE

For example, let's suppose the TRUE state of nature is:
Higgs boson exists with the mass of m_H(true) = 134.26 GeV

Use data sample to examine this!

Events collected after some time of LHC running
Event 1
Event 2
...
Event N

Event 1
Object 1
Object 2
...
Object k

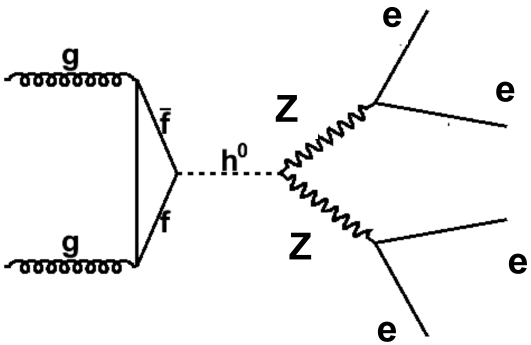
If Object 1 == electron
p_x
p_y
p_z
E
...

N ~ 100/s x 10^7 s/year
N ~ 10^9 events per year

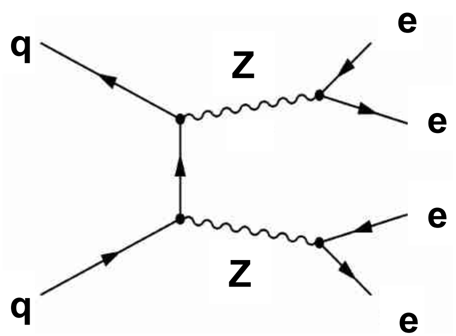
Objects ≡ reconstructed objects
i. e. electrons, photons, jets, muons ...

Signal vs background(s)

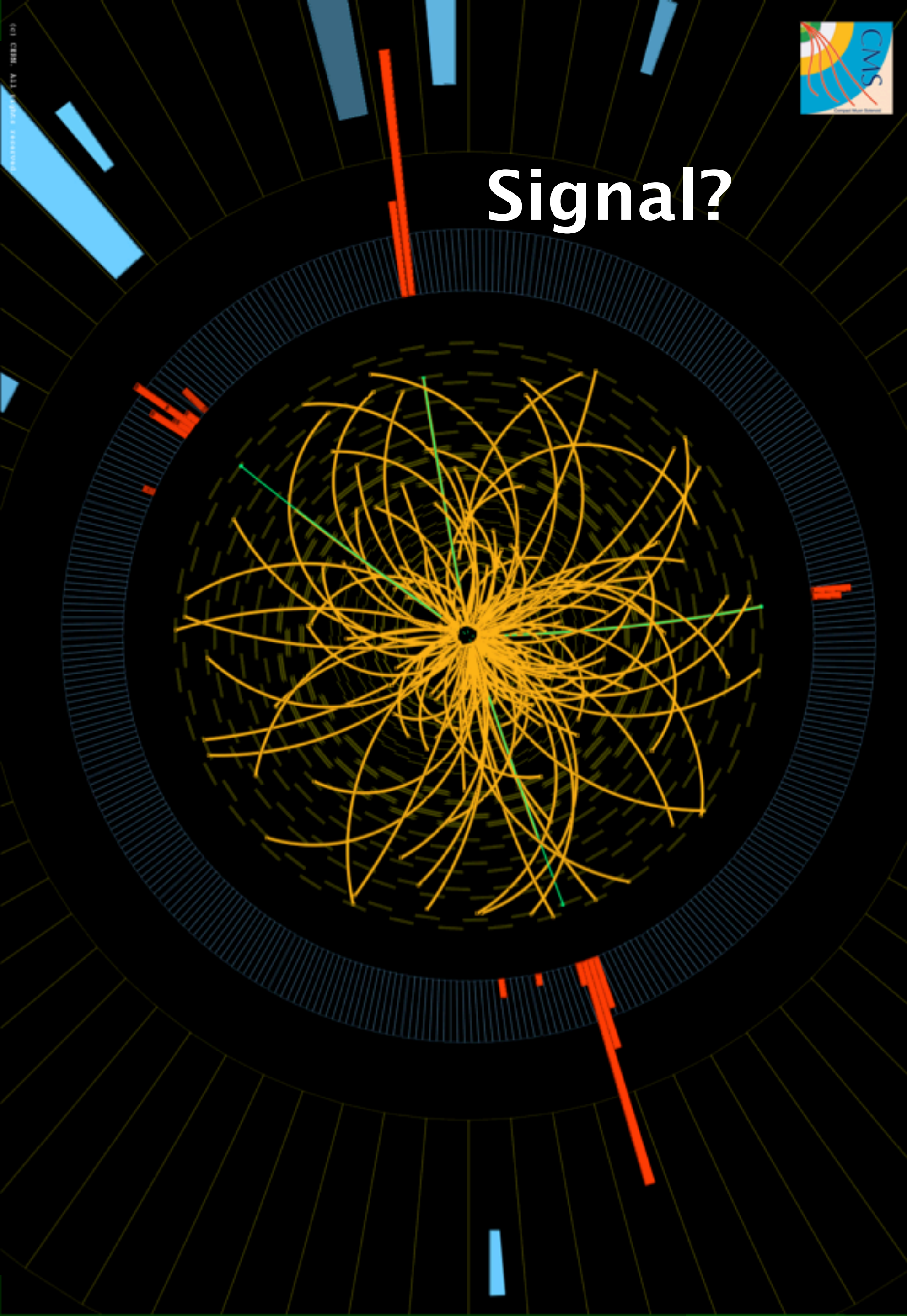
- **Signal**: an event coming from the physical process under study
  - Example: $H \rightarrow ZZ \rightarrow e^+e^-e^+e^-$ (henceforth both e^+ and e^- are '*electron*')
- **Background**: any other event
  - 'Dangerous' background is any other process giving at least 4 electrons in the final state
    - But be careful: electrons seen by detector are reconstructed objects and in some cases when some other objects (f.g. jets) are misreconstructed as electrons
  - 'Trivial' backgrounds are all other backgrounds and are easily *rejected* by a simple requirement of having at least 4 electrons in the final state

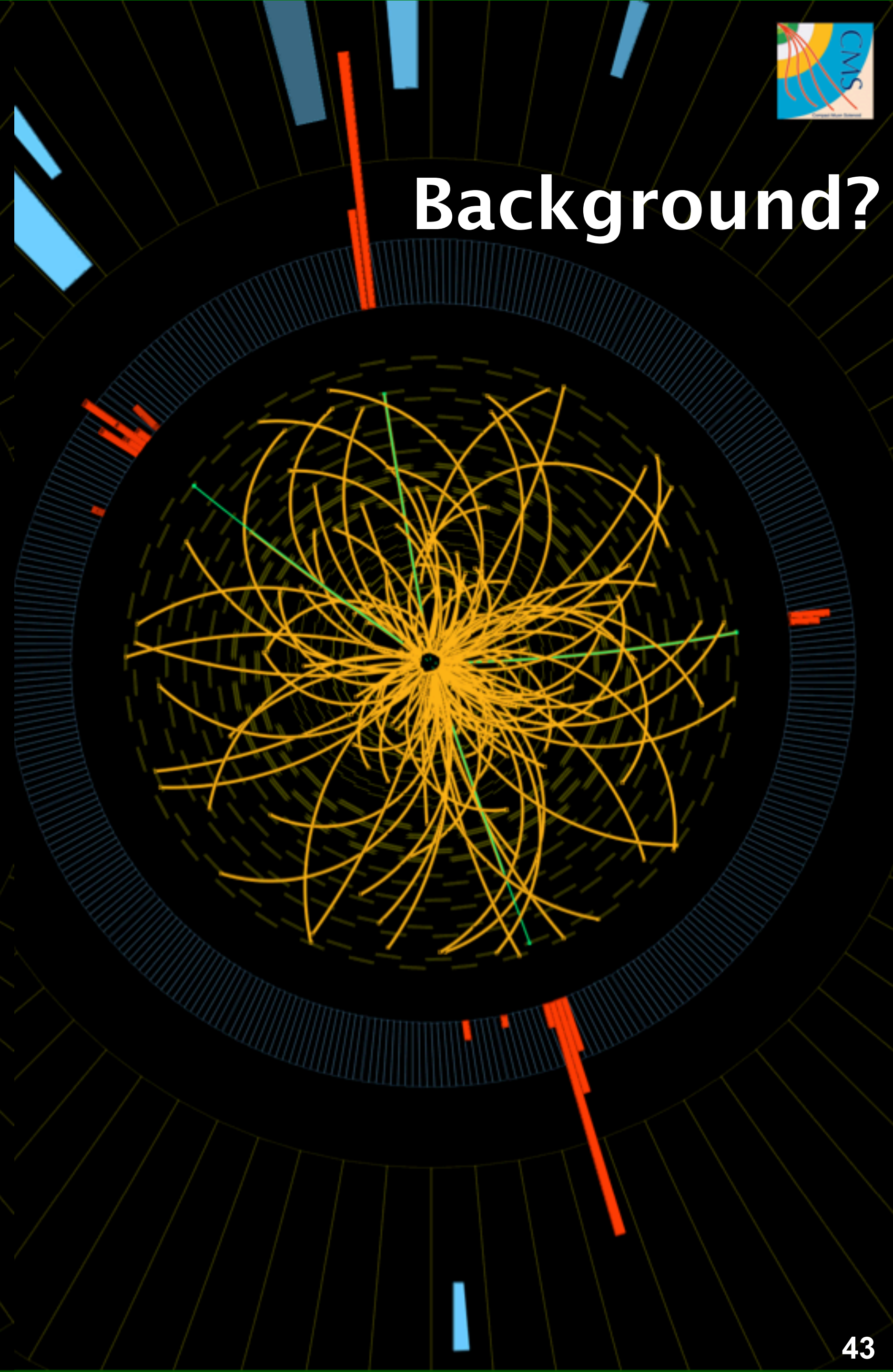


Signal: pp→H→ZZ→4e



'Dangerous' background: pp→ZZ→4e


CMS
Signal?


CMS
Background?

Separating signal and background

- Ultimate goal of the analysis: separate as much as possible signal from background events to obtain a **reduced sample** as clean as possible
  - This is usually obtained in several steps

Trigger → Skimming and preselection → Selection

  - Usually all these steps have substeps
  - More in example on the next page
- Be aware:
  - Nature is probabilistic, i.e. for a given event it'll never be possible to tell whether it's signal or background!
  - We can only make an educated guess → attribute probabilities that the observed event comes from signal or background

p(event|signal) and ***p(event|background)***

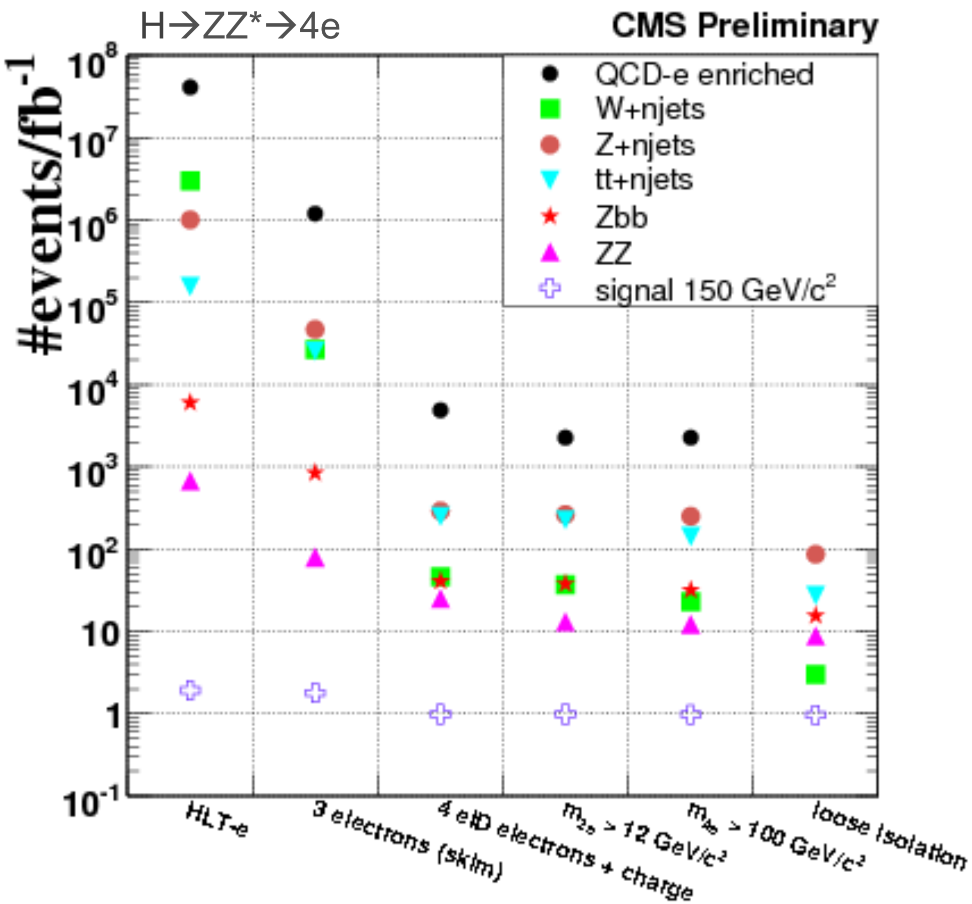
- Very often we have to solve the following statistical problem: **maximum reduction of the background for a given signal acceptance**

Exploring the data

- Once data are collected → exploratory data analysis
  - Heavily use of graphical techniques
- Example: **data reduction =** skimming [+ preselection]
  - Goal: **getting rid of all unuseful events**
  - Unusefullness is not uniquely defined:
    - We have a certain interest to keep some background events for better control and its measurement from data
  - Some numbers:
    - ~ 10^9 events collected per year (after trigger)
    - ~ 1 MB event size on a tape (rought estimate)
    - ⇒ ~ 1 PB of data collected per year → non manageable at once
  - Interested physical processes are rare
    - F.g. just a handful (~10) H→ZZ→4e events per year
    - So be careful when choosing criteria for data reduction not to lose too many signal events

Example: H→ZZ→4e in CMS

- **Skimming cuts:** High Level Trigger+ **≥ 3 electrons, any charge and $p_T^{1,2,3}$ > 10, 10, 5 GeV/c**
- Preselection cuts:
  - ≥ 2 *ee* pairs of identified, opposite charge and same flavor leptons with
    - p_T > 5 GeV/c; $|\eta|$ < 2.5
  - At least two m_{ee} > 12 GeV/c^2
  - At least one m_{4e} > 100 GeV/c^2
  - Loose track based isolation
- After these steps
  - Some background gone
  - Some heavily reduced
  - Some still resisting
- Full **selection** needed for the final analysis



Probability

Random variables

Probability – basic concepts

- Definitions of probability
  - **Mathematical probability**
    - Probability is a basic and an abstract concept
  - **Frequentist probability**
    - Using only measured frequencies
  - **Bayesian probability**
    - Based on a *degree of belief*

Mathematical probability

- Developed in 1933 by Kolmogovor in his "*Foundations of the Theory of Probability*"
- Define Ω as an exclusive set of all possible elementary events x_i
  - Exclusive means the occurence of one of them implies that none of the others occurs
- We define the probability of the occurency of x_i, $P(x_i)$ to obey the **Kolmogorov axioms**:

$$
\begin{aligned}
&(a)\ P(x_i) \geq 0 \quad \text{for all } i \\
&(b)\ P(x_i \text{ or } x_j) = P(x_i) + P(x_j) \\
&(c)\ \sum_{\Omega} P(x_i) = 1
\end{aligned}
$$

- From these properties more complex probability expressions can be deduced
  - For non-elementary events, i.e. set of elementary events
  - For non-exclusive events, i.e. overlapping sets of elementary events

Frequentist probability

- Experiment:
  - N events observed
  - Out of them n is of type x
- **Frequentist probability** that any single event will be of type x

$$P(x) = \lim_{N \to \infty} \frac{n}{N}$$

- Important restriction: such a probability can only be applied to repeatable experiments
  - For example one can't define a probability that it'll snow tomorrow
  - Altough this seems to be a serious problem, a job of scientist is to try to get as close as possible to repeatable experiments and produce reproducible results
- Frequentist statistics is often associated with the names of *Jerzy Neyman* and *Egon Pearson*

Bayesian probability

- Based on a concept of “degree of belief”
- An operational definition of belief is based on coherent bet by Finneti
  - **What’s amount of money one ‘s willing to bet based on her/ his belief on the future occurence of the event**
- Bayesian inference uses Bayes’ formula for conditional probability:

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}$$

- *H* is a **hypothesis**, and *D* is the **data**.
- *P*(*H*) is the **prior probability** of *H*: the probability that *H* is correct before the data *D* was seen.
- *P*(*D*|*H*) is the **conditional probability** of seeing the data *D* given that the hypothesis *H* is true. *P*(*D*|*H*) is called the **likelihood**.
- *P*(*D*) is the **marginal probability** of *D*.
  - *P*(*D*) is the prior probability of witnessing the data *D* under all possible hypotheses
- *P*(*H*|*D*) is the **posterior probability**: the probability that the hypothesis is true, given the data and the previous state of belief about the hypoth.

Example: Who will pay the next round?

You meet an old fried at Göttingen in a pub. He proposes that the next round should be payed by whichever of the two extracts the card of lower value from a pack of cards.
This situation happens many times in the following days. What is the probability that your friend cheats if you end up paying *wins* consecutive times[2]
You assume:

- $P(cheat) = 5\%$ and $P(honest) = 95\%$. (Surely an old friend is an unlikely cheater ...)
- $P(wins|cheat) = 1$ and $P(wins|honest) = 2^{-wins}$

Bayesian solution:

$$P(cheat|wins) = \frac{P(wins|cheat)P(cheat)}{P(wins|cheat)P(cheat) + P(wins|honest)P(honest)}$$

$$P(cheat|0) = \frac{1P(cheat)}{1P(cheat) + 2^{-0}P(honest)} = \frac{0.05}{0.05 + 0.95} = 5\%$$

$$P(cheat|5) = \frac{1P(cheat)}{1P(cheat) + 2^{-5}P(honest)} = \frac{0.05}{0.05 + 0.03} = 63\%$$

[2] Adapted from G. D'Agostini, *Bayesian Reasoning in High-Energy Physics: Principles and Applications*, CERN-99-03, 1999

Example: Learning by experience

The process of updating the probability when new experimental data becomes available can be followed easily if we insert

- $P(cheat) = P(cheat|wins-1)$ and $P(honest) = P(honest|wins-1)$, where $wins-1$ indicate the propability assigned after *the previous win*
- $P(wins=1|cheat) = P(win|cheat) = 1$ and $P(wins=1|honest) = P(win|honest) = \frac{1}{2}$

Iterative aplication of the Bayes formula for $P(cheat|wins)$=

$$\frac{P(win|cheat)P(cheat|wins-1)}{P(win|cheat)P(cheat|wins-1) + P(win|honest)P(honest|wins-1)}$$

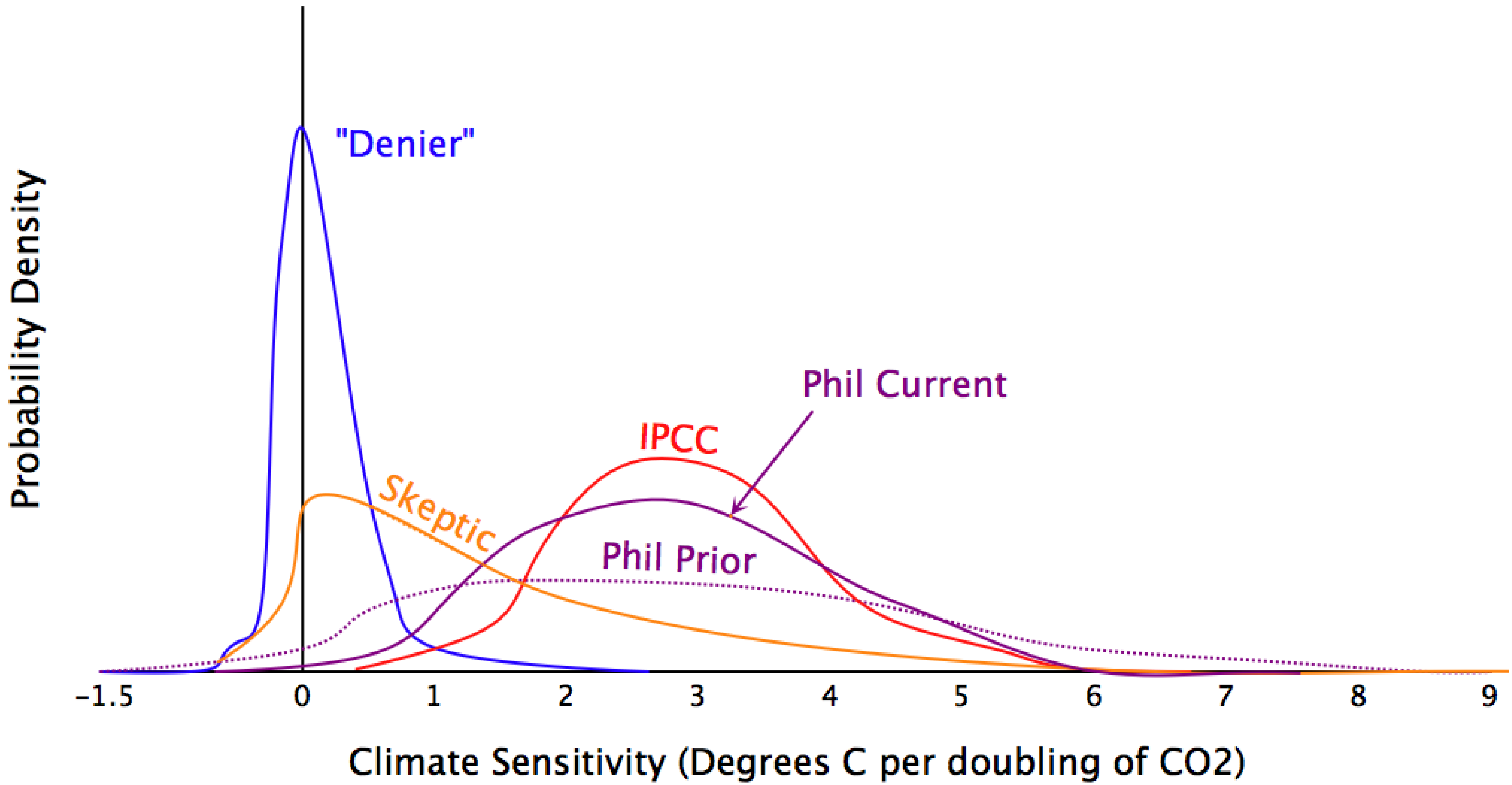
$$= \frac{P(cheat|wins-1)}{P(cheat|wins-1) + \frac{1}{2}P(honest|wins-1)}$$

P(cheat) %	$P(cheat\|wins)$ wins=5	10	15
1	24	91	99.7
5	63	98	99.94
50	97	99.9	99.997

When you learn from the experience, your conclusions no longer depend on the initial assumptions.

Example: Priors and posteriors – expressing degree of belief

Phil is learning from experience:



(From discussion of climate change on Andrew Gelman's blog.)

Random variables

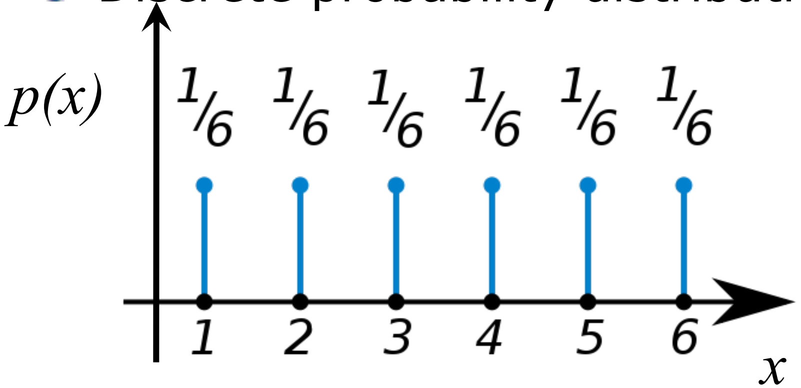
- **Random event**: event having more than one possible outcome
  - Each outcome may have associated probability
  - Outcome not predictible, only the probabilities known
- Different possible outcomes may take different possible numerical values $x_1, x_2, \ldots$ → **random variable** $\boldsymbol{x}$
  - The corresponding probabilities $P(x_1), P(x_2), \ldots$ form a ***probability distribution***
- If observations are **independent** the distribution of each random variable is unaffected by knowledge of any other observation
- When an experiment consists of N repeated observations of the same random variable x, this can be considered as the single observation of a random vector $\boldsymbol{x}$, with components $x_1, \ldots, x_N$

Random variables: discrete

- Rolling a die:
  - Sample space = {1,2,3,4,5,6}
  - Random variable x is the number rolled

$$x = \begin{cases} 1 & \text{if a 1 is rolled} \\ 2 & \text{if a 2 is rolled} \\ 3 & \text{if a 3 is rolled} \\ 4 & \text{if a 4 is rolled} \\ 5 & \text{if a 5 is rolled} \\ 6 & \text{if a 6 is rolled} \end{cases}$$

  - Discrete probability distribution



Random variables: continuous

- A spinner
  - Can choose a real number from [0,2n]
  - All values equally likely
  - X = the number spun
  - Probability to select any real number = 0
  - Probability to select any **range** of values > 0
    - Probability to choose a number in [0,n] = 1/2
  - Now we say that probability **density** $p(x)$ of x is $1/2n$
  - Probability to select a number from any range Δx is $\Delta x/2n$
  - More general

$$P(A < x < B) = \int_A^B p(x)dx$$

Probability density function

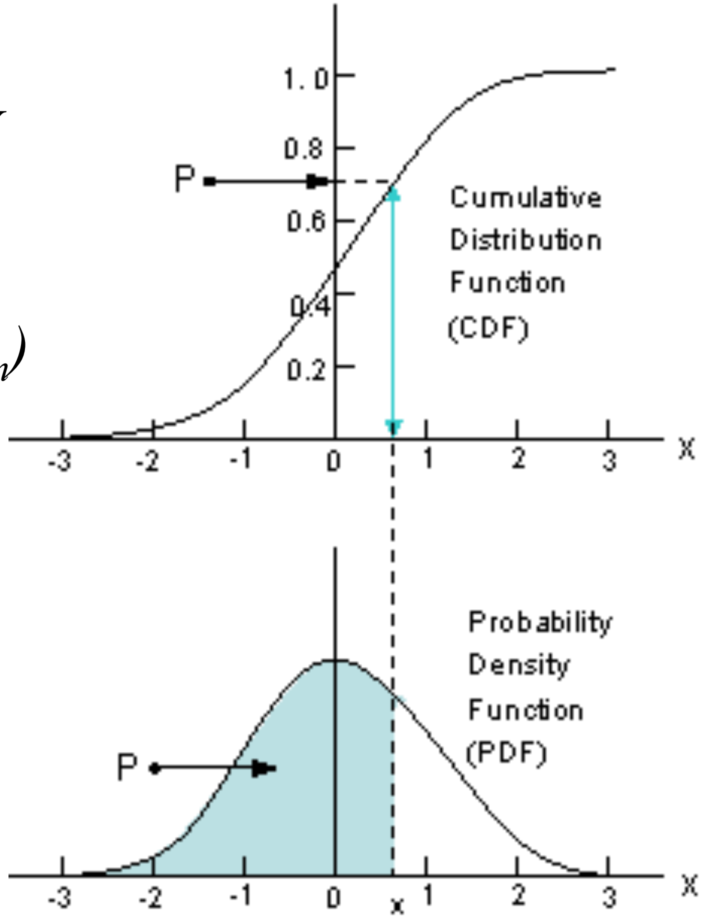
- Let x be a possible outcome of an observation and can take any value from a continuous range
- We write $f(x;\theta)dx$ as the probability that the masurement's outcome lies betwen x and $x + dx$
- The function $\boldsymbol{f(x;\theta)dx}$ is called the **probability density function *(PDF)***
  - And may depend on one or more parameters θ
- If $f(x;\theta)$ can take only **discrete values** then $\boldsymbol{f(x;\theta)}$ is itself a **probability**
- The p.d.f. is always normalized to unit area (unit sum, if discrete)
- Both x and θ may have multiple components and then written as vectors
- If θ is unkown we may wish to estimate its value from a set of measurements of x → Parameter estimation in Lecture 2

Cumulative and marginal distributions

- **Cumulative distribution function, CDF**
  - For every real number Y, the CDF of Y is equal to the probability that the random variable x takes a value less or equal to Y

$$F(Y) = P(x \le Y) = \int_{x_{\min}}^{Y} f(x)dx$$

  - If x restricted to $x_{min} < x < x_{min}$ then $F(x_{min}) = 0$, $F(x_{max}) = 1$
  - $F(x)$ is a monotonic function of x



- **Marginal density function**
  - Is the projection of multidimensional density
  - Example: if $f(x,y)$ is two-dimenisonal PDF the marginal density $g(x)$ is

$$g(x) = \int_{y_{\min}}^{y_{\max}} f(x,y)dy$$