Improving the sensitivity of Higgs Boson Searches

Kunal Kumar

Reference arxiv No. : 1108.2274

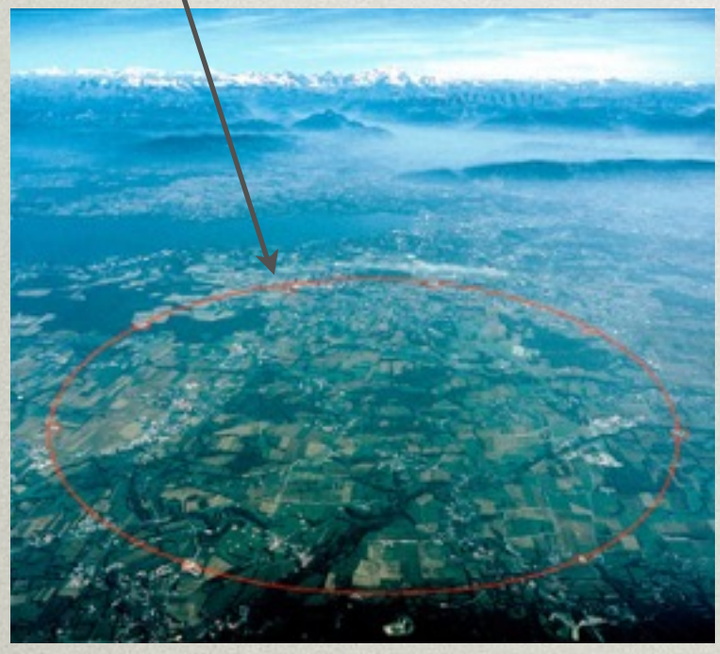
Setting the stage

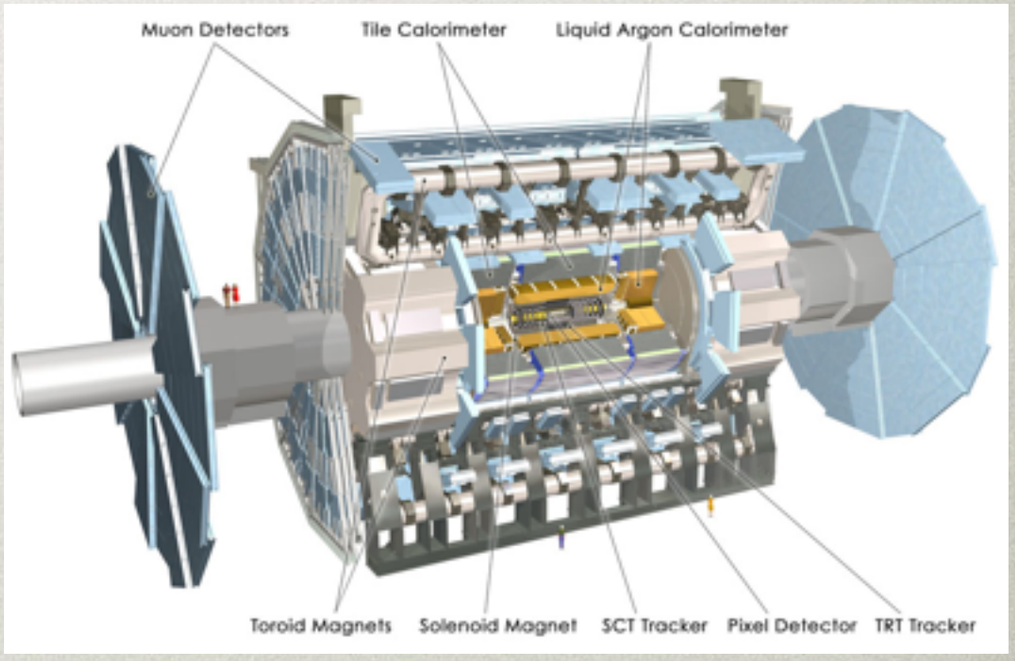
- Imagine looking for a particle that has been theorized to exist for 40 years
- It isn't found in everyday matter because it is too heavy and decays quickly
- The only way to find it is to try and create it at a magical machine called the Large Hadron Collider
- When created at the collider this particle lasts for 10^-22 seconds
- Imagine finding something that ephemeral in an environment where the amount of information you receive is 40 TB/sec.
- Chance of Higgs : Something Else :: 1 : 10^10
- With a problem of this magnitude you need 1000s of people contributing small steps in concert to make any progress
- I'll be talking about a project that was one such step. I worked on it in 2011, a few months before the Higgs was discovered.

CONTEXT - Particle physics? collider physics?

- Study of matter at extremely small length scales (< 10^-15 m)
- At a collider charged particles like protons are accelerated to large velocities and smashed into each other
- Detector placed around the interaction point gathers data about each "event"

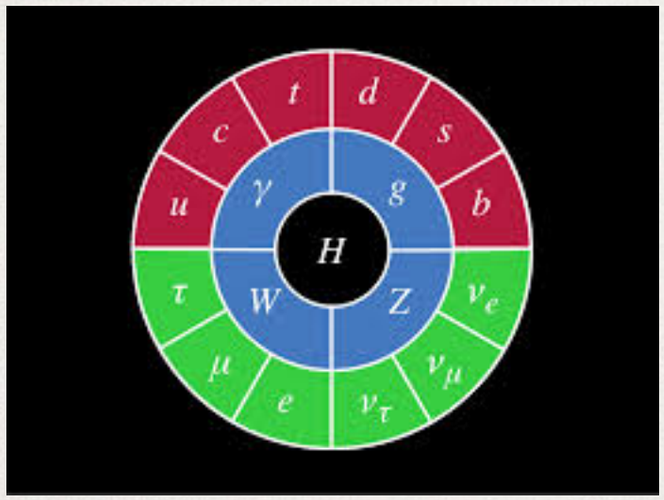
LHC circumference ~ 17 miles!!!





Why do we do this?

Finding the higgs boson

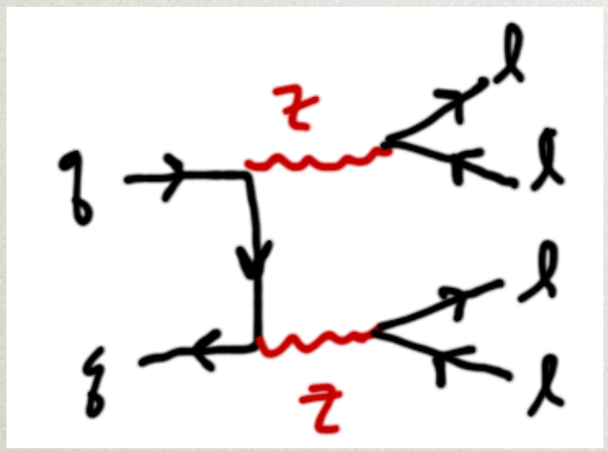


- The Higgs boson is critical to our understanding of nature because it provides mass to other elementary particles.
- Basic Idea : Produce the Higgs. It is heavy and interacts with other particles so it will decay. Study its decay products.
- Complications : Higgs decays can occur through many paths or “channels“
- Complications : Many other processes also produce the same final products. All these make it really hard to extract signal from large amount of background

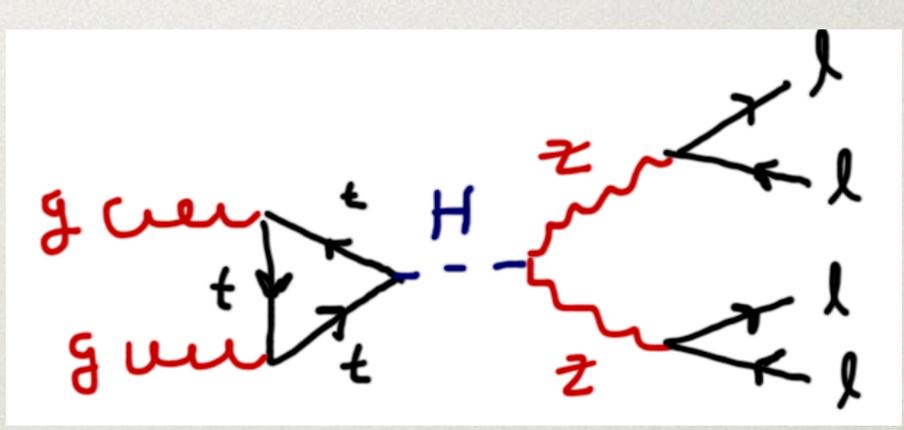
Looking back at 2011 ...

$$H \rightarrow ZZ \rightarrow 4l$$

- One of the promising channels (because of ease in identifying the final state particles) is the "Golden channel"



eg. of Background



eg. of Signal

- In this channel analysis by experiments typically used just one feature (or variable) of an event to distinguish signal from background
- In theory there are 10. More features means better discriminating power.

More Variables

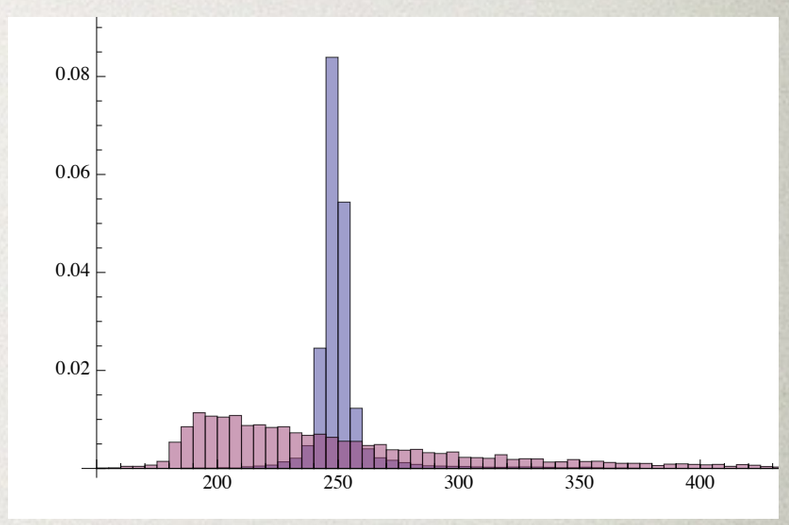
- Typically expected number of signal events for our study vary between 0.5 - 6. Background (after a LOT of filters) is roughly 5 - 10 times this.
- With so few events we should try to extract as much information as we can
- We set out to quantify the gain in significance of hypothesis discrimination from using a multivariate analysis instead of a univariate one

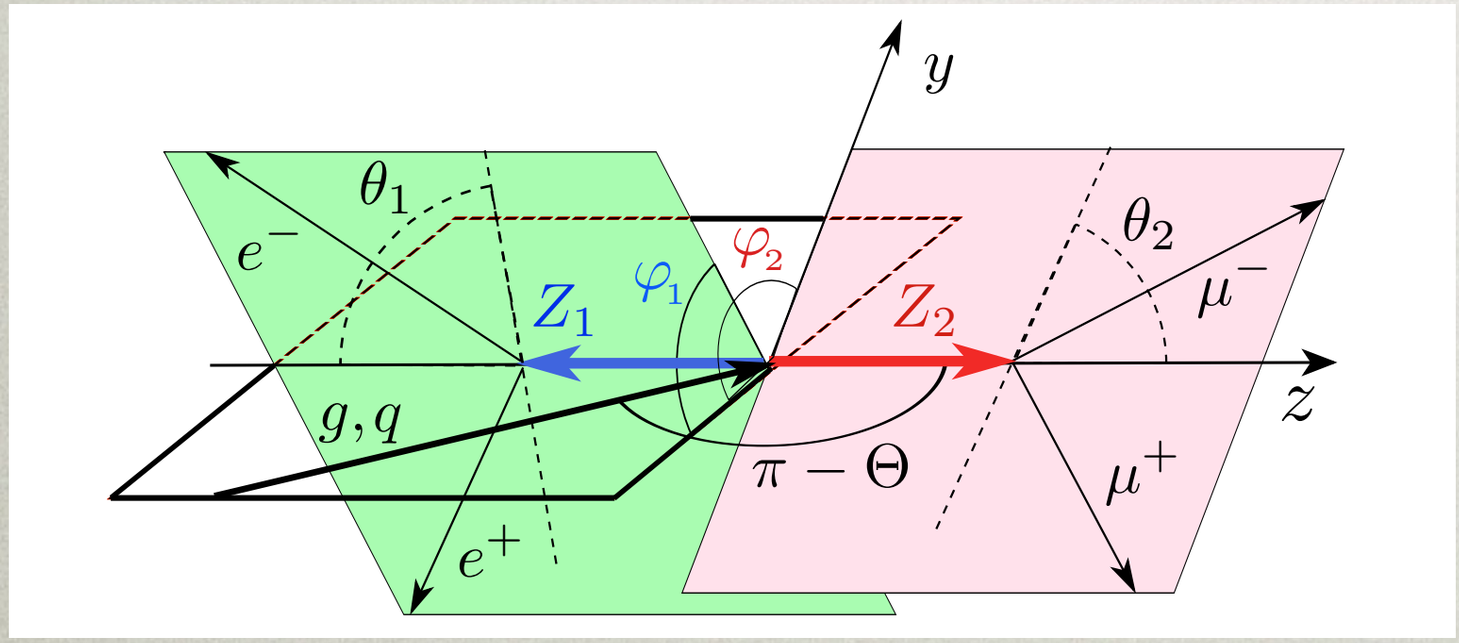
Univariate vs Multivariate

Univariate Case : Total energy in interaction

Multivariate Case : 10 features

$$x_i \equiv (x1, x2, M1, M2, \hat{s}, \Theta, \theta_1, \phi_1, \theta_2, \phi_2)$$





Analysis - Test Statistic

- Extended Maximum Likelihood Method to extract Higgs mass, signal fraction and significance of discriminating Higgs + background vs background only $\mathcal{S} = \sqrt{2 ln Q}$
- Likelihood ratio $Q = \frac{\mathcal{L}_{s+b}}{\mathcal{L}_b}$ (Never smaller than 1)
- Likelihood function : Ps and Pb are pdfs for the signal and background hypotheses and can be calculated from first principles

$$\mathcal{L}_{s+b} = e^{-\mathcal{N}_t} \mathcal{N}_t^N \prod_{i=0}^{N} [f_s \mathcal{P}_s(m_h, x_i) + (1 - f_s)\mathcal{P}_b(x_i)]$$

- You do need to go into the Avatar state to get Ps and Pb right
- For a given set of data (events) we can compute significance by maximizing each likelihood function with respect to the undetermined parameters

$$\mathcal{Q} = \frac{\mathcal{L}_{s+b}(\hat{N}t, \hat{f}s, \hat{m}h; x_i)}{\mathcal{L}_b(\hat{N}_t; x_i)}$$

Analysis - Expected Significance

- 5 different signal scenarios (Higgs mass is different in each and in the range 175 GeV - 350 GeV). For each scenario 10000 pseudo-experiments.
- Pseudo-experiments are vital because you need a thorough strategy before you spend millions of dollars and years collecting data
- Generate two pools of a million signal events and 10 million background events for 4 lepton final states using MadGraph (a collider event generator) and a cluster at Northwestern U.

```
<event>
 7   0  0.3787848E-06  0.1001269E+02  0.7957747E-01  0.1778220E+00
        1   -1    0    0  501    0  0.00000000000E+00  0.00000000000E+00  0.62500000000E+02  0.62500000000E+02  0.00000000000E+00 0. -1.
       -1   -1    0    0    0  501 -0.00000000000E+00 -0.00000000000E+00 -0.62500000000E+02  0.62500000000E+02  0.00000000000E+00 0.  1.
       23    2    1    2    0    0  0.14249644531E+02  0.13380203193E+02 -0.13920202990E+02  0.10025392849E+03  0.97339590715E+02 0.  0.
       11    1    3    3    0    0 -0.41452935425E+00 -0.32145848096E+02  0.23253765328E+02  0.39677008284E+02  0.00000000000E+00 0. -1.
      -11    1    3    3    0    0  0.14664173885E+02  0.45526051289E+02 -0.37173968318E+02  0.60576920210E+02  0.00000000000E+00 0.  1.
       13    1    1    2    0    0 -0.17300302216E-01  0.42351377367E+00  0.98638495762E+00  0.10736007179E+01  0.00000000000E+00 0.  1.
      -13    1    1    2    0    0 -0.14232344229E+02 -0.13803716966E+02  0.12933818032E+02  0.23672470788E+02  0.00000000000E+00 0. -1.
</event>
```

- Number of signal and background events for each experiment picked from Poisson distributions (number of expected events is small)
- Events that are part of an experiment are chosen randomly from the pool
- Run the analysis for each pseudo-experiment using Mathematica for both univariate and multivariate case. Compare improvement.
- We know underlying parameters so we can check validity of the analysis

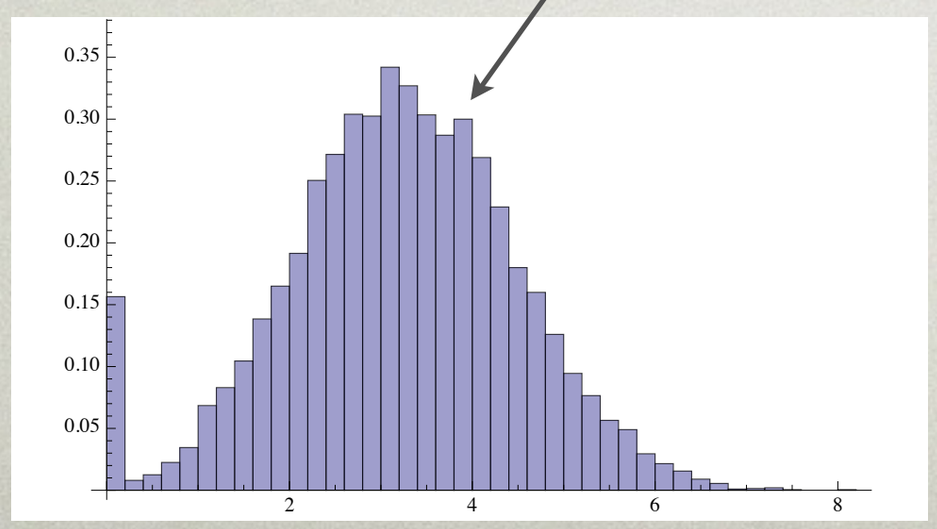
Analysis - Expected Significance

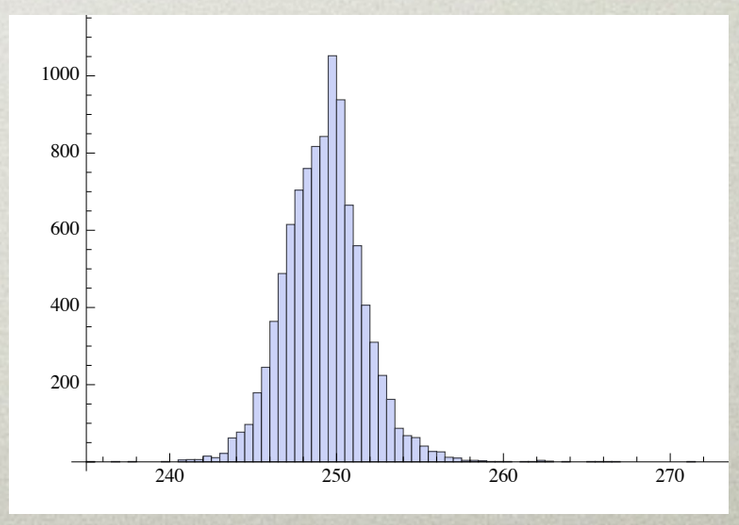
- For each experiment we calculate the significance (for univariate and multivariate case) and extract underlying parameters

$$\mathcal{S} = \sqrt{2 ln \mathcal{Q}}$$

$$\mathcal{Q} = \frac{\mathcal{L}_{s+b}(\hat{N}t, \hat{f}s, \hat{m}h; x_i)}{\mathcal{L}_b(\hat{N}_t; x_i)}$$

- With 10000 such pseudo experiments we get a distribution of significances which is close to a normal distribution

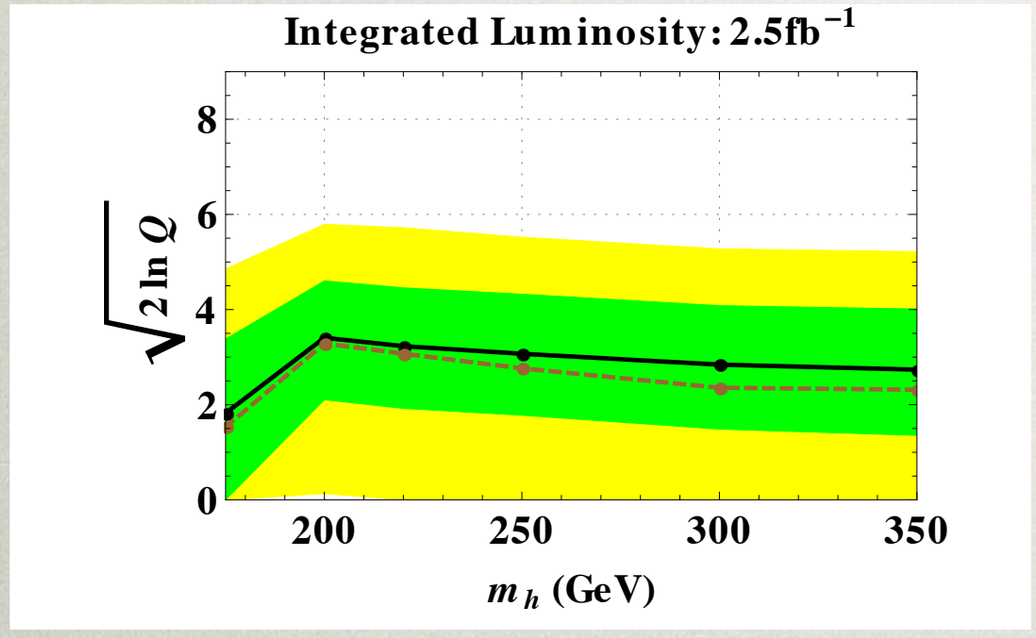


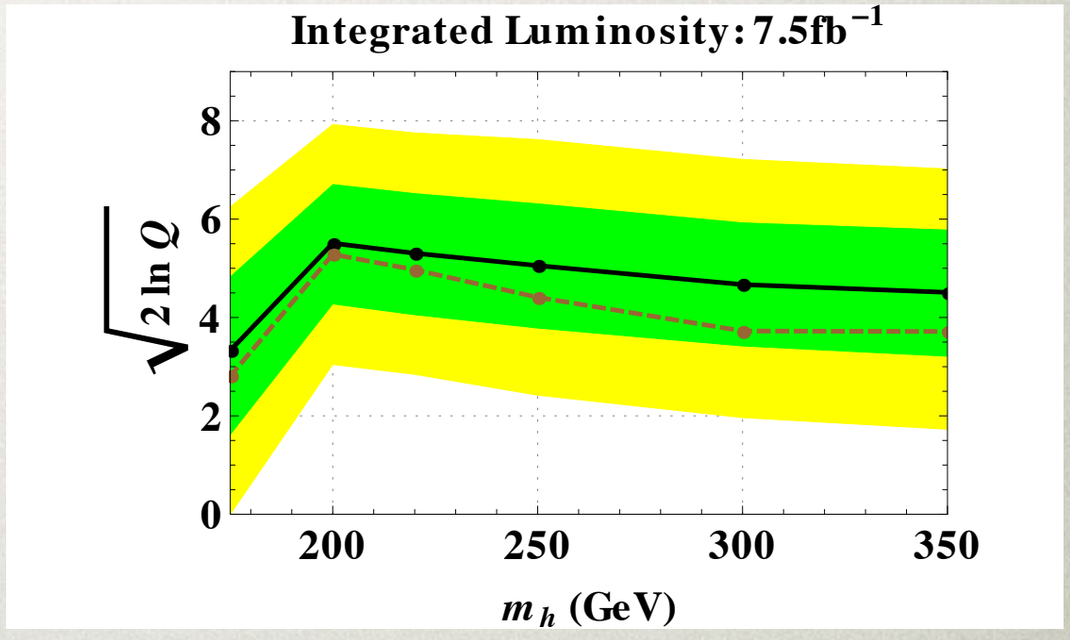


Extracted Higgs Mass

Analysis - Expected Significance

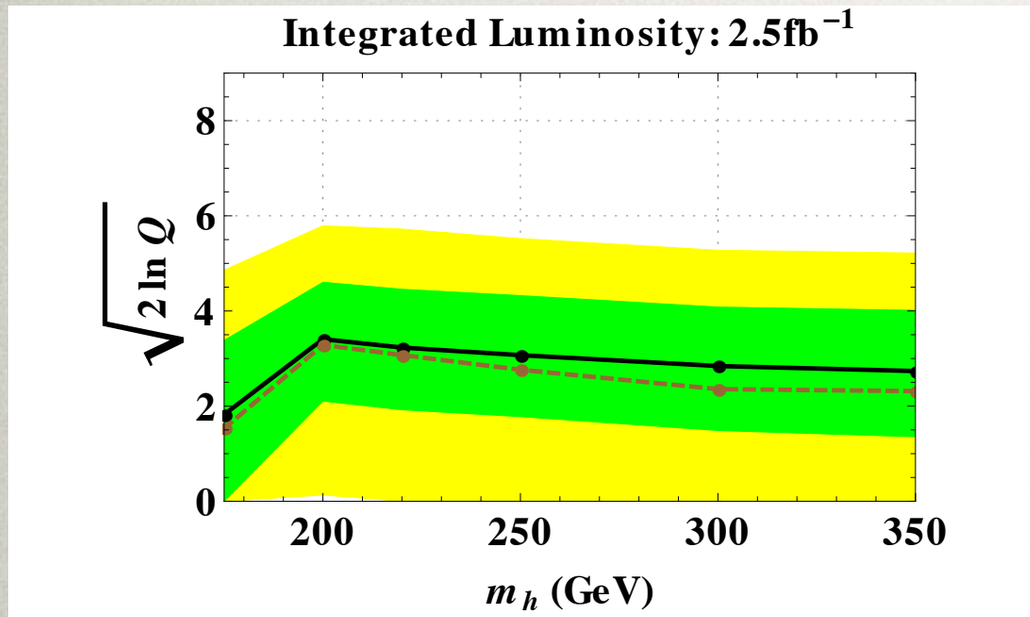
- We choose median as the expected value and compute the 1 and 2 sigma deviations of the expected significance
- Compare univariate and multivariate case for different amounts of gathered data

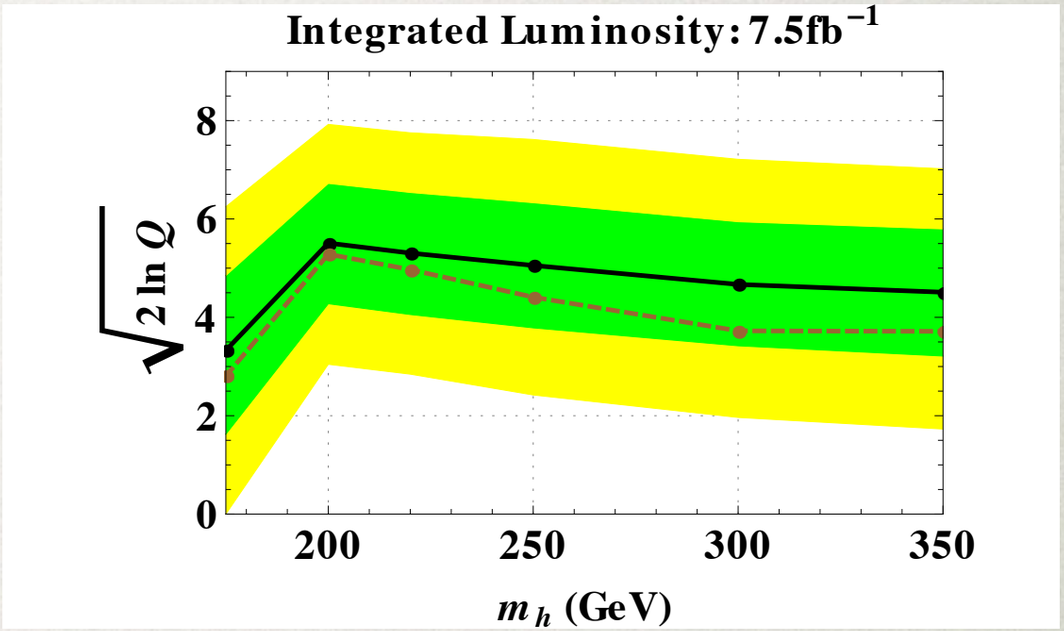




- 1 and 2 sigma bounds for univariate case not shown (same as the size of multivariate case)

Analysis : Uni- vs Multivariate





~ 1.5 yrs of data gathering

~ 2.5 yrs of data gathering

- At higher masses we can gain 10 - 15% in discovery significance
- We published this in Nov 2011
- The Higgs boson discovery was announced on Jul 4, 2012.

Data Driven Decisions

- The mass of the Higgs boson is 125 GeV but in principle it could have been anywhere between 100 - 1000 GeV
- Our analytically calculated pdf for the background was used by experimentalists to help rule out the Higgs in the mass range above 175 GeV

Evidence for a new state in the search for the standard model Higgs boson in the $H \to ZZ \to 4\ell$ channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV

The CMS Collaboration

July 9, 2012

We use a matrix element likelihood analysis (MELA). We construct a kinematic discriminant (KD) based on the probability ratio of the signal and background hypotheses, MELA KD $= \mathcal{P}_{\rm sig}/(\mathcal{P}_{\rm sig} + \mathcal{P}_{\rm bkg})$, as described in Ref. [32]. The likelihood ratio is defined for each value of $m_{4\ell}$. The signal and $q\bar{q} \to ZZ$ background analytical parametrisations are taken from Refs. [32] and [70], respectively, and include the phase-space and Z propagator terms. When $m_{4\ell}$ is above

[70] J. S. Gainer, K. Kumar, I. Low et al., "Improving the sensitivity of Higgs boson searches in the golden channel", *JHEP* **1111** (2011) 027, doi:10.1007/JHEP11(2011)027, arXiv:1108.2274.

Summary

- Our work helped quantify how much significance we gain in moving from univariate to multivariate analysis
- We provided analytical expressions that helped the experimentalists to discover the Higgs
- Along with other papers we helped to push for more multivariate analyses using analytical expressions of pdfs
- This method has the added advantage (mainly for theorists) that the physics is clear at each step (as compared to using Neural Networks for eg. where you use physics to make sense of why certain features have bigger weights)

THANKS!