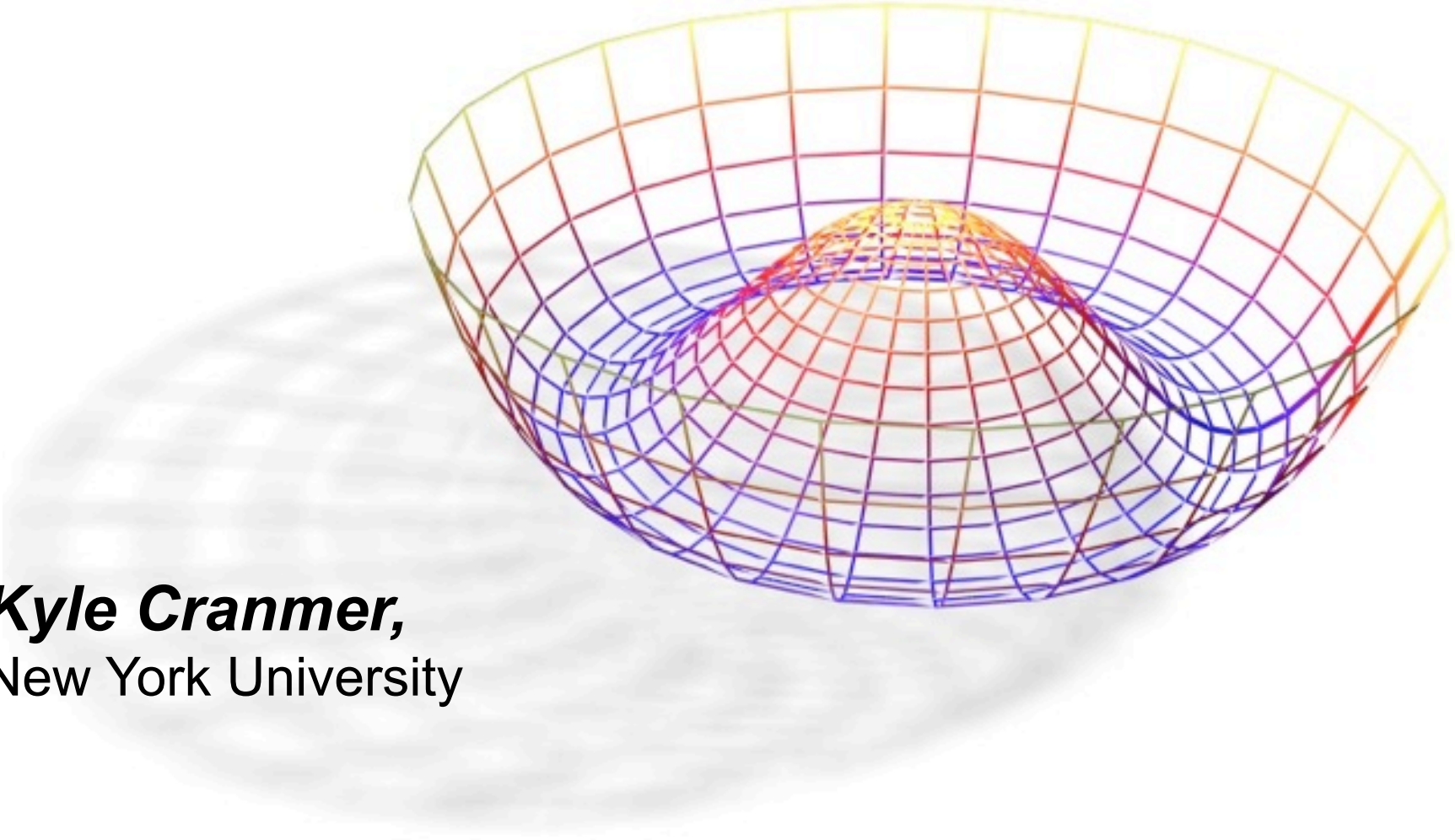
Center for
Cosmology and
Particle Physics

Statistical Aspects of the Discovery of the Higgs Boson at the Large Hadron Collider



Kyle Cranmer,
New York University

Introduction

Fundamental Particles & Interactions

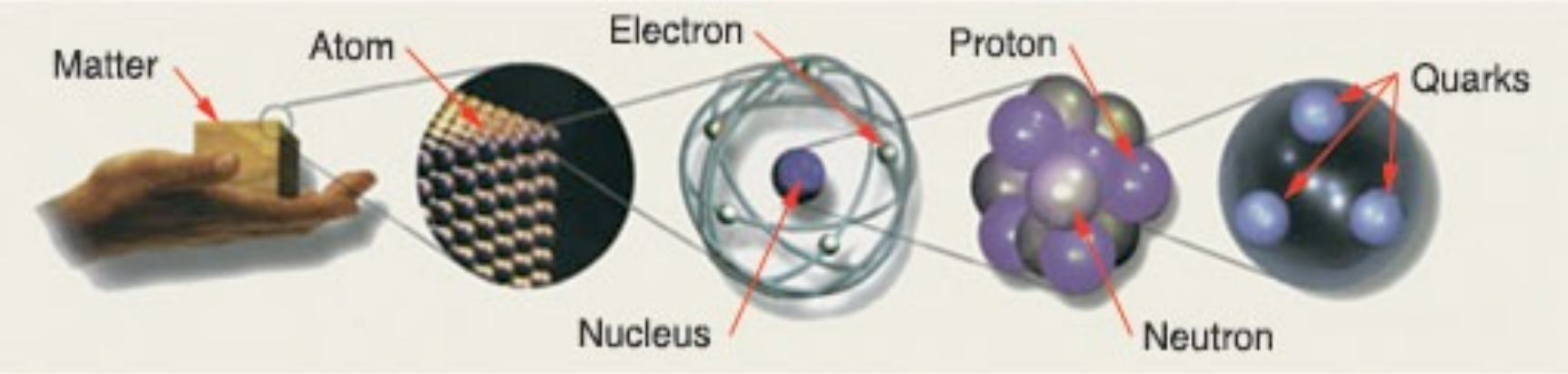
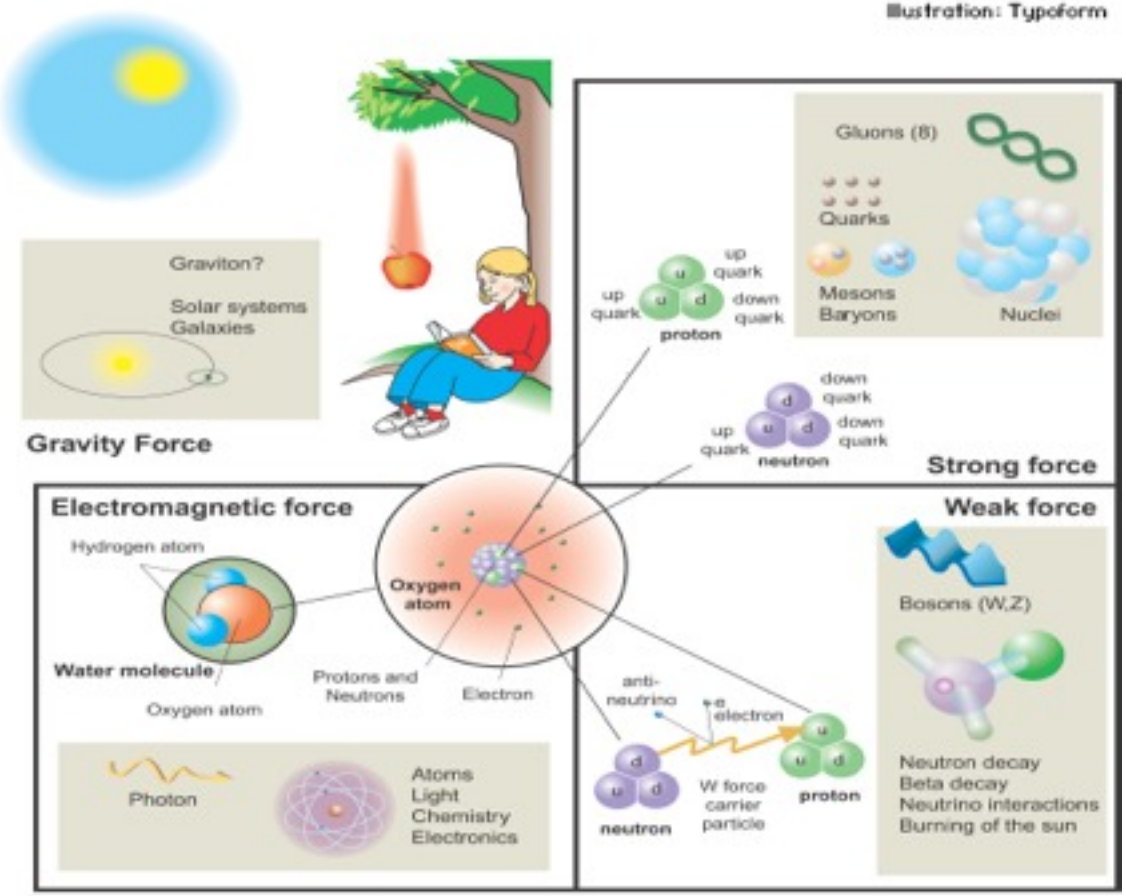
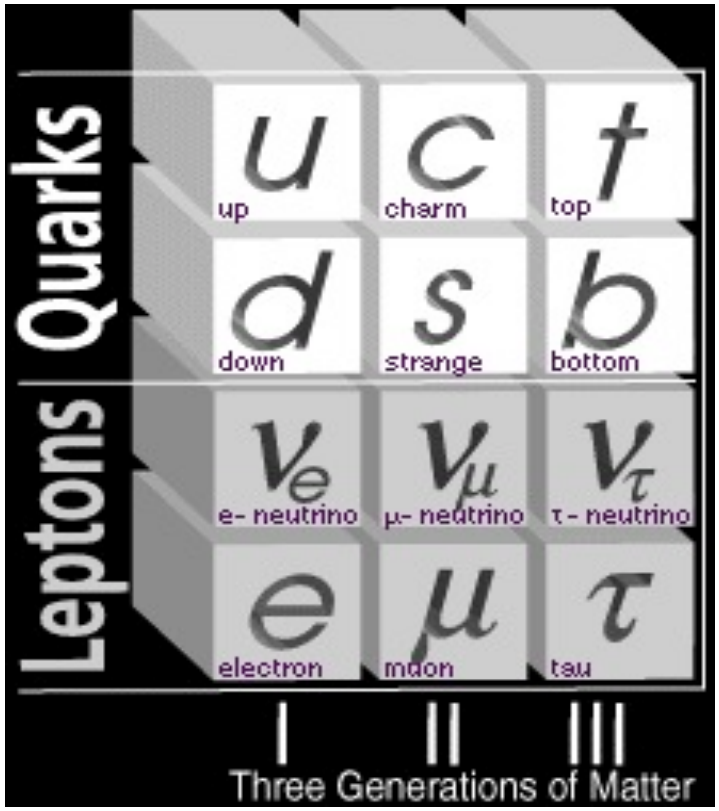
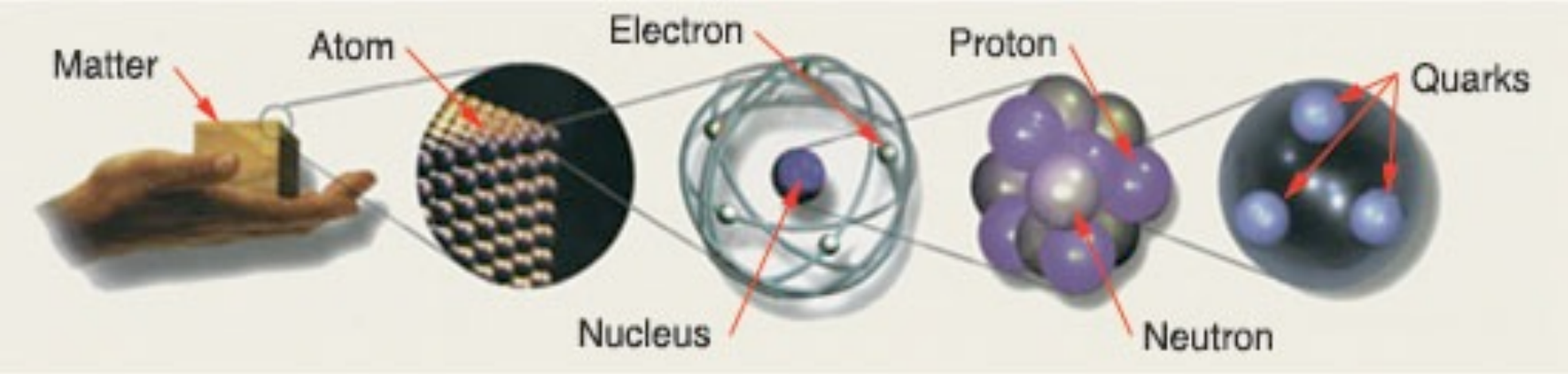


Illustration: Typoform





Fundamental Particles & Interactions

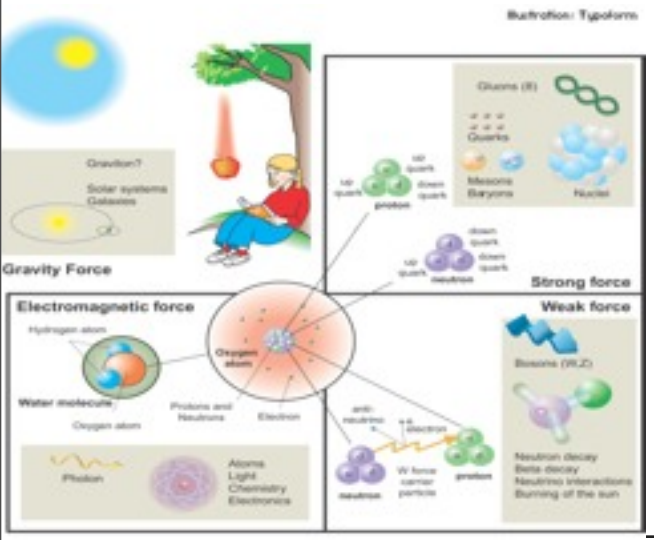


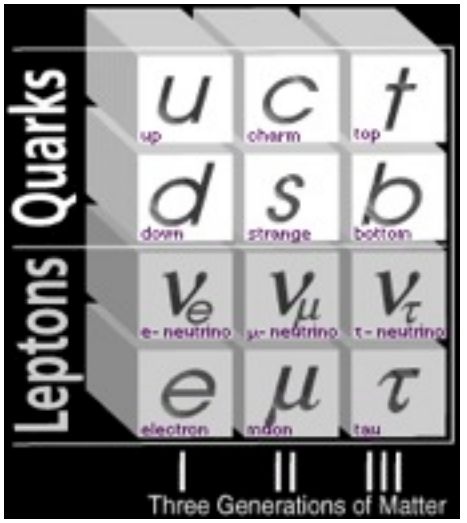
$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\,\phi\right|^2 - V(\phi)}_{W^\pm,Z,\gamma,\text{and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{R}\phi_c L + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

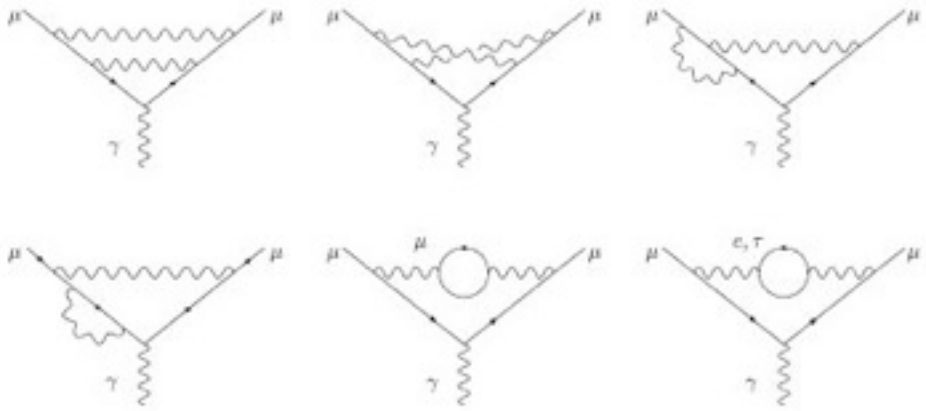




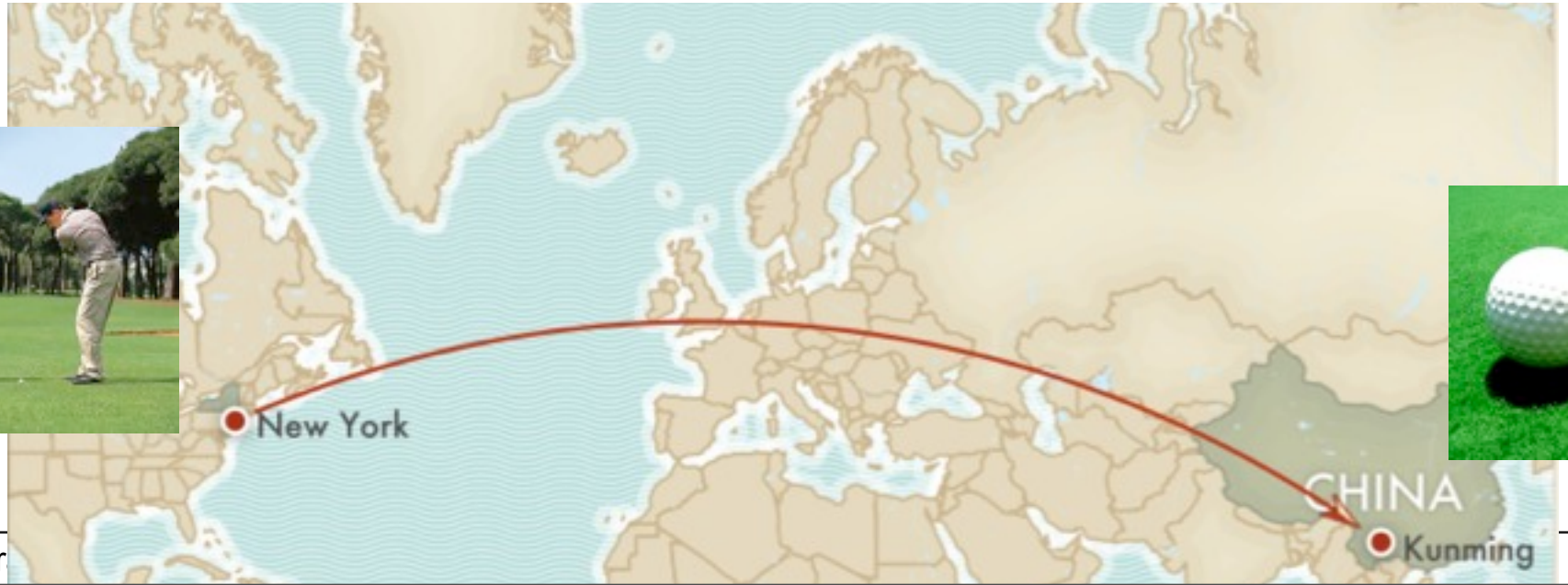
The Success of the Standard Model & QFT

Non-trivial aspects of the theory have been tested to < 1 ppm

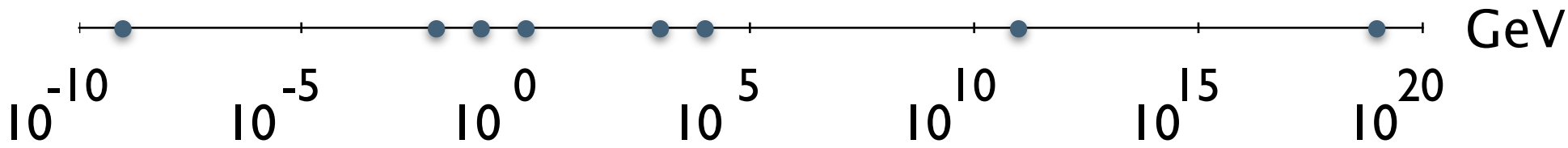
A unique realm for reasonable statistical exploration of a scientific theory



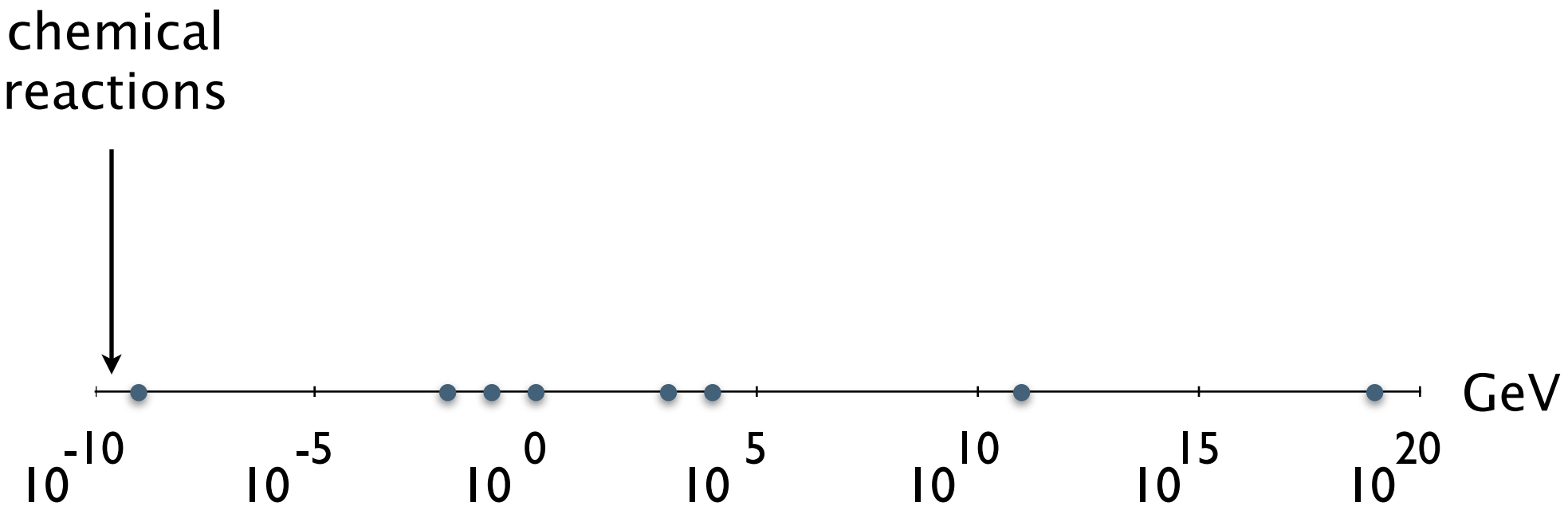
$$a_\mu\ (exp) = 11\ 659\ 208\ (6) \times 10^{-10}\ (0.5\ ppm)$$



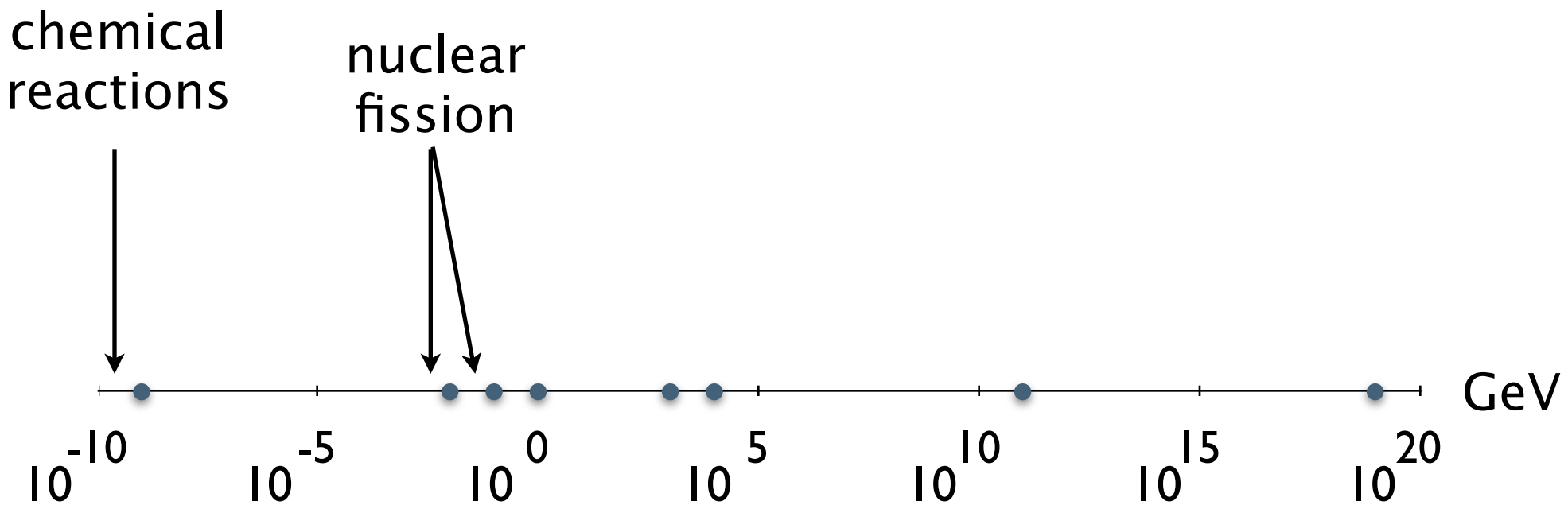
Energy scales in context



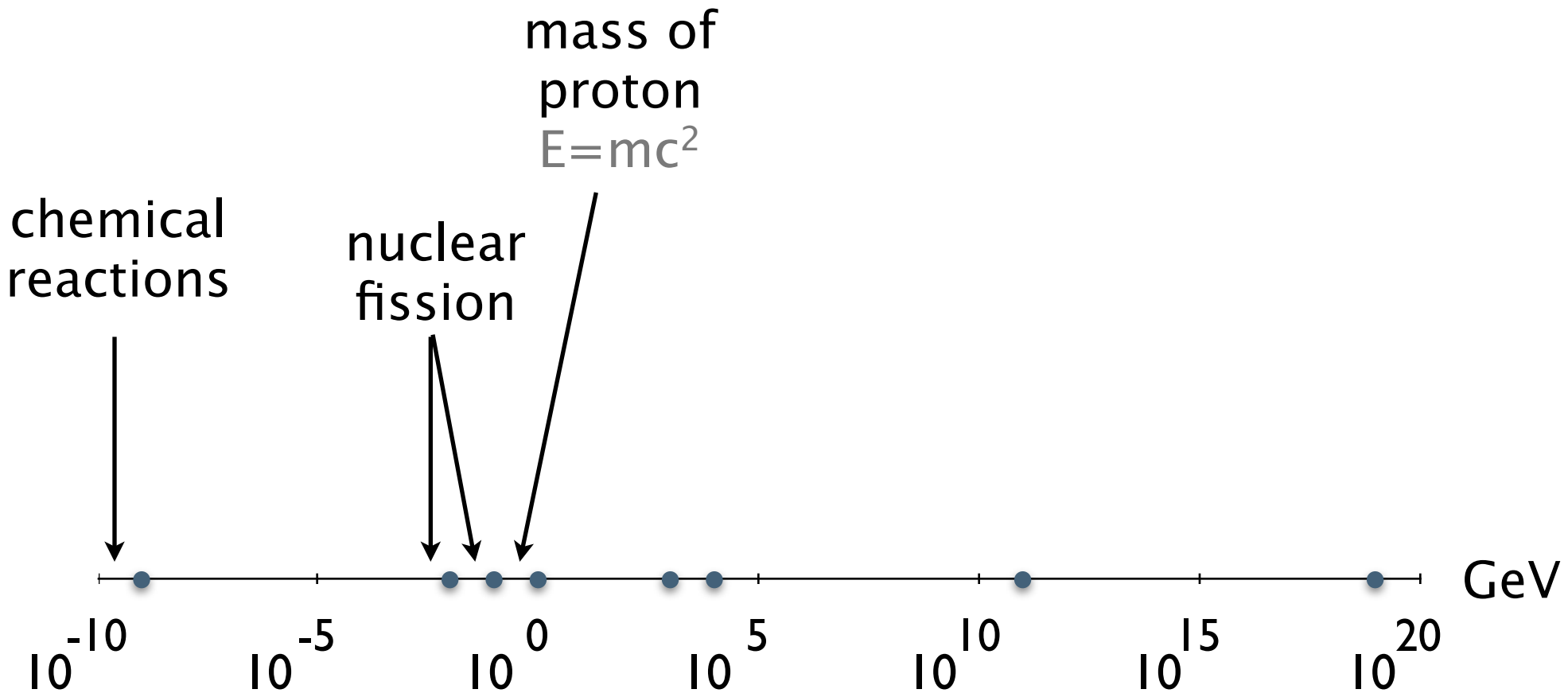
Energy scales in context



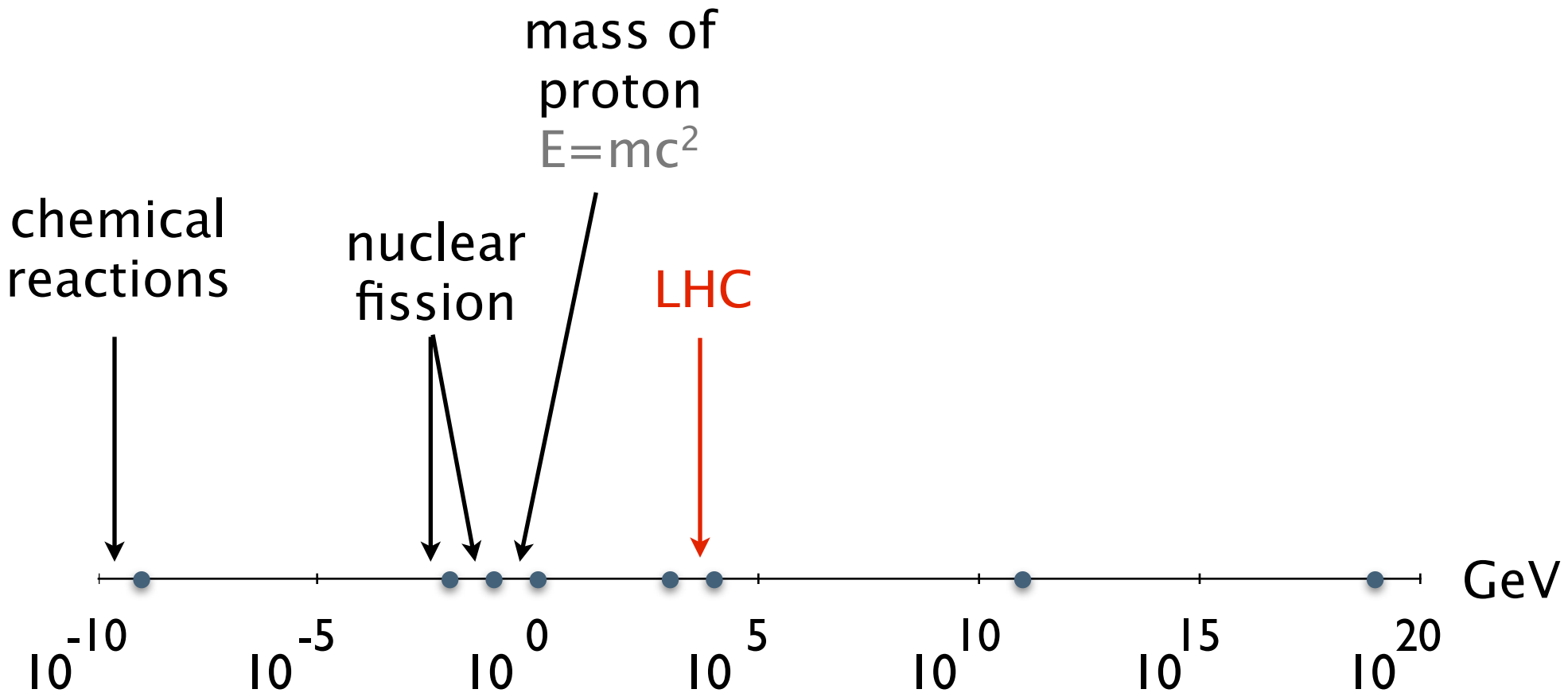
Energy scales in context



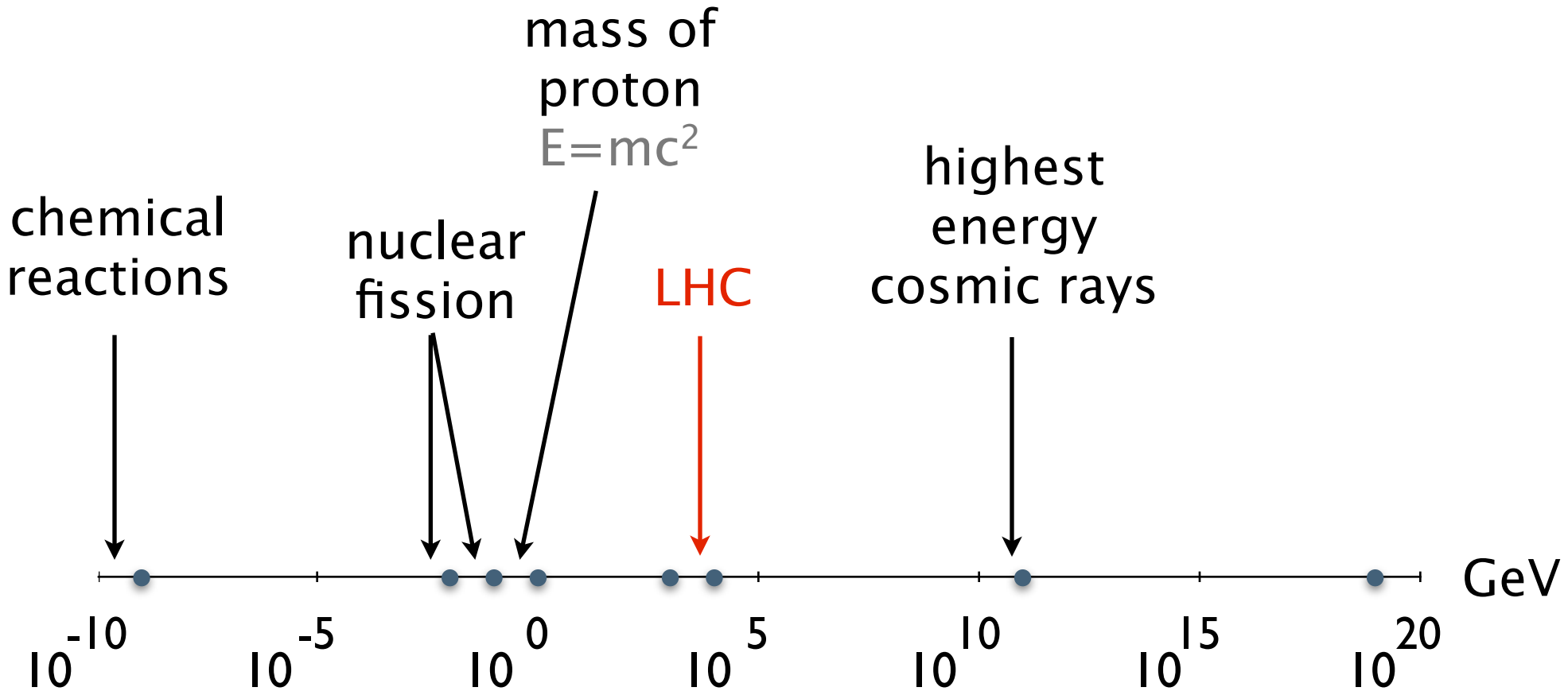
Energy scales in context



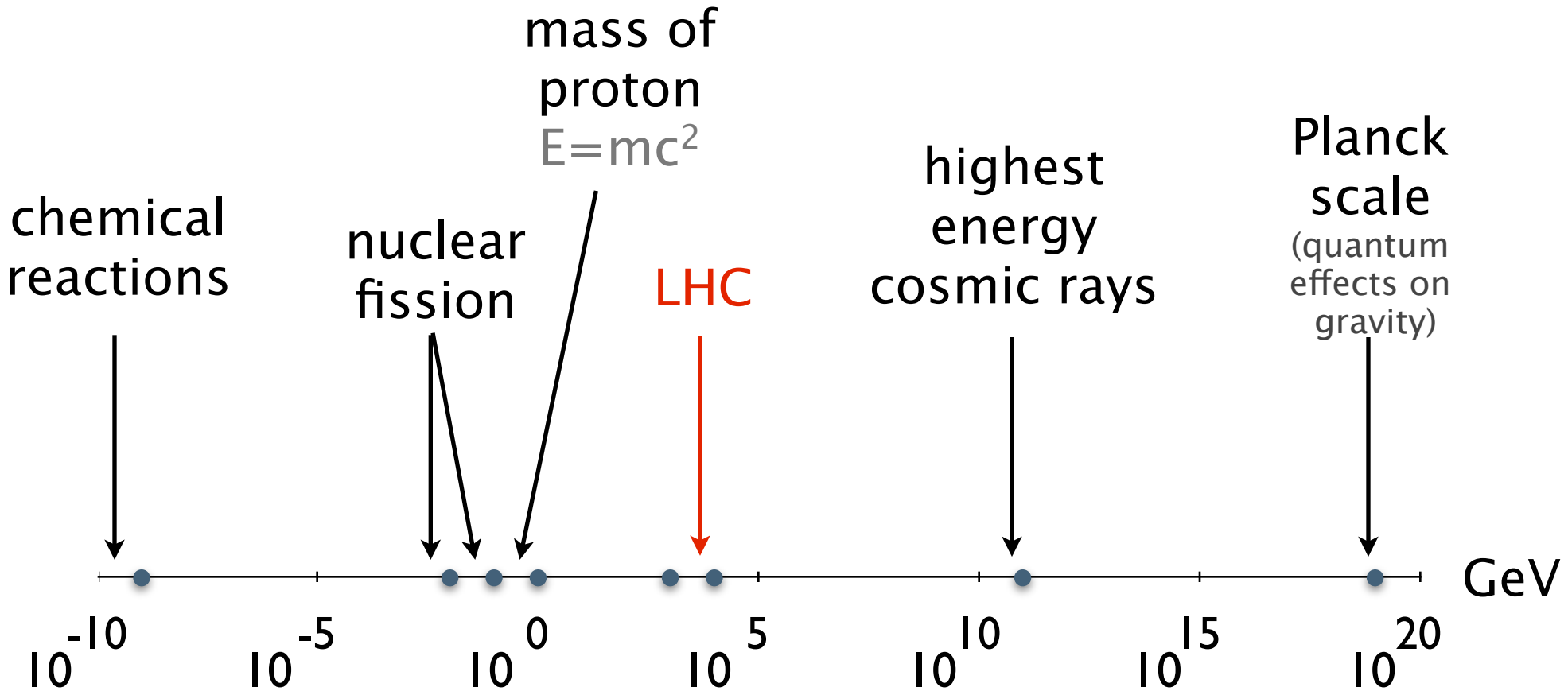
Energy scales in context



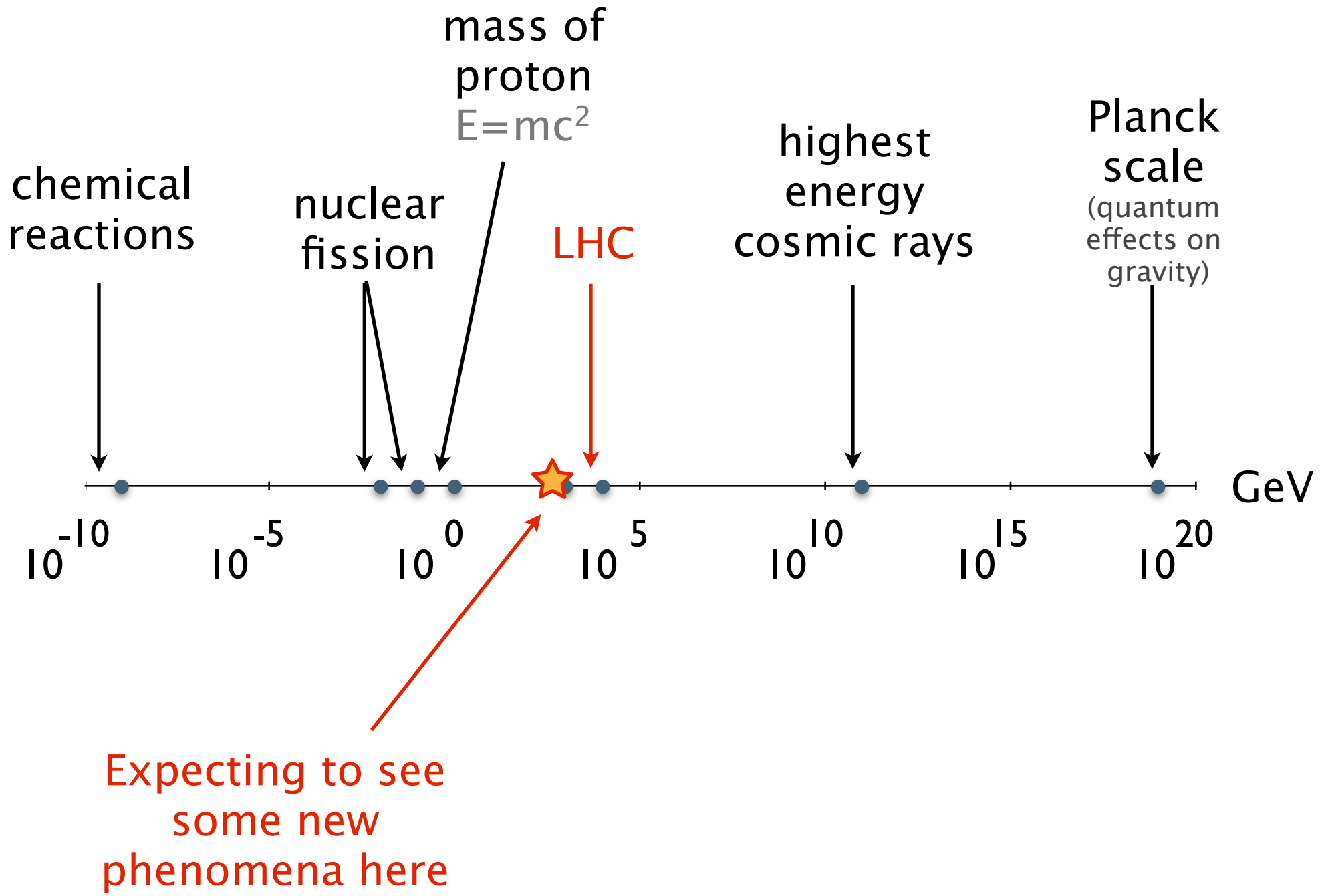
Energy scales in context



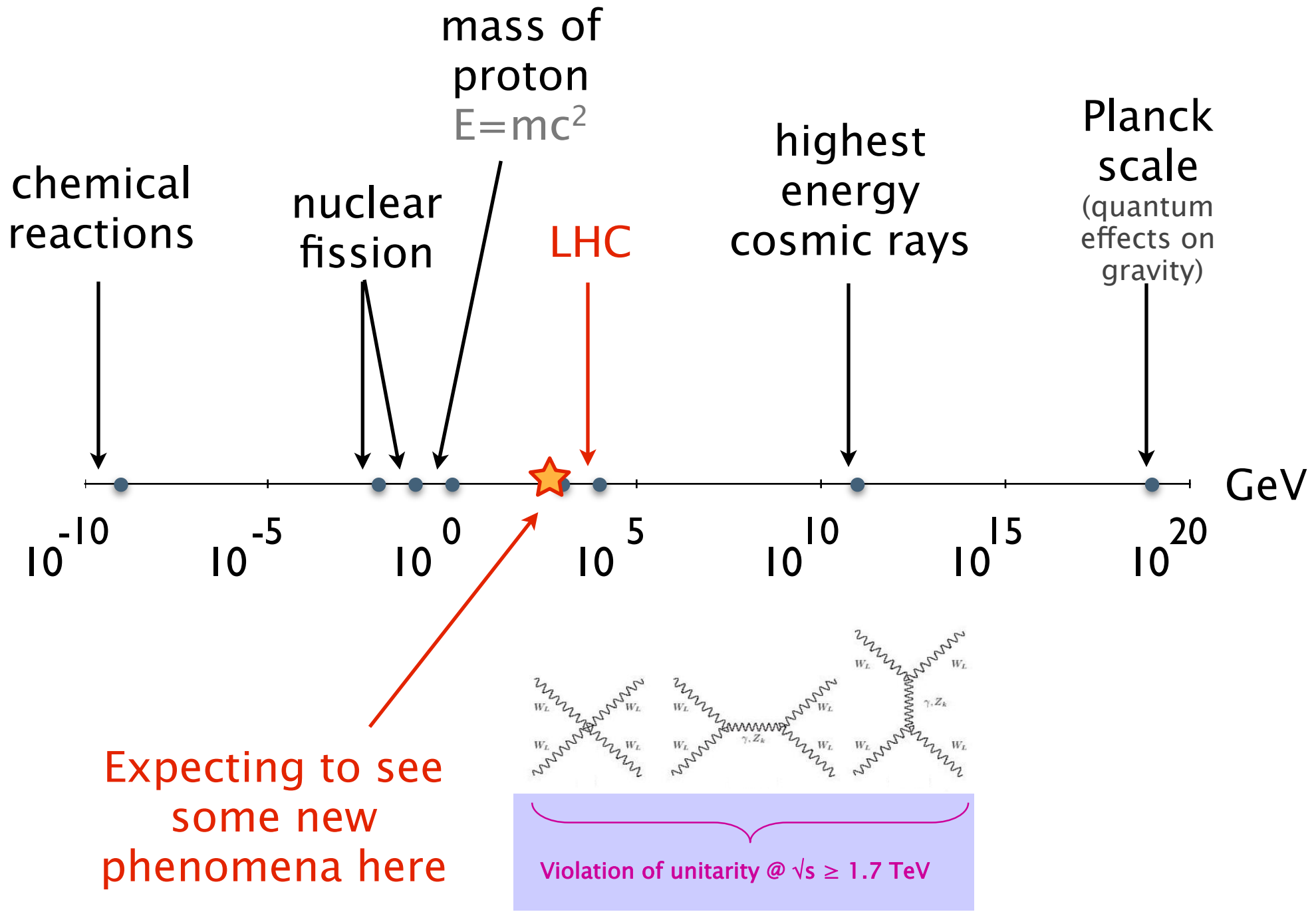
Energy scales in context



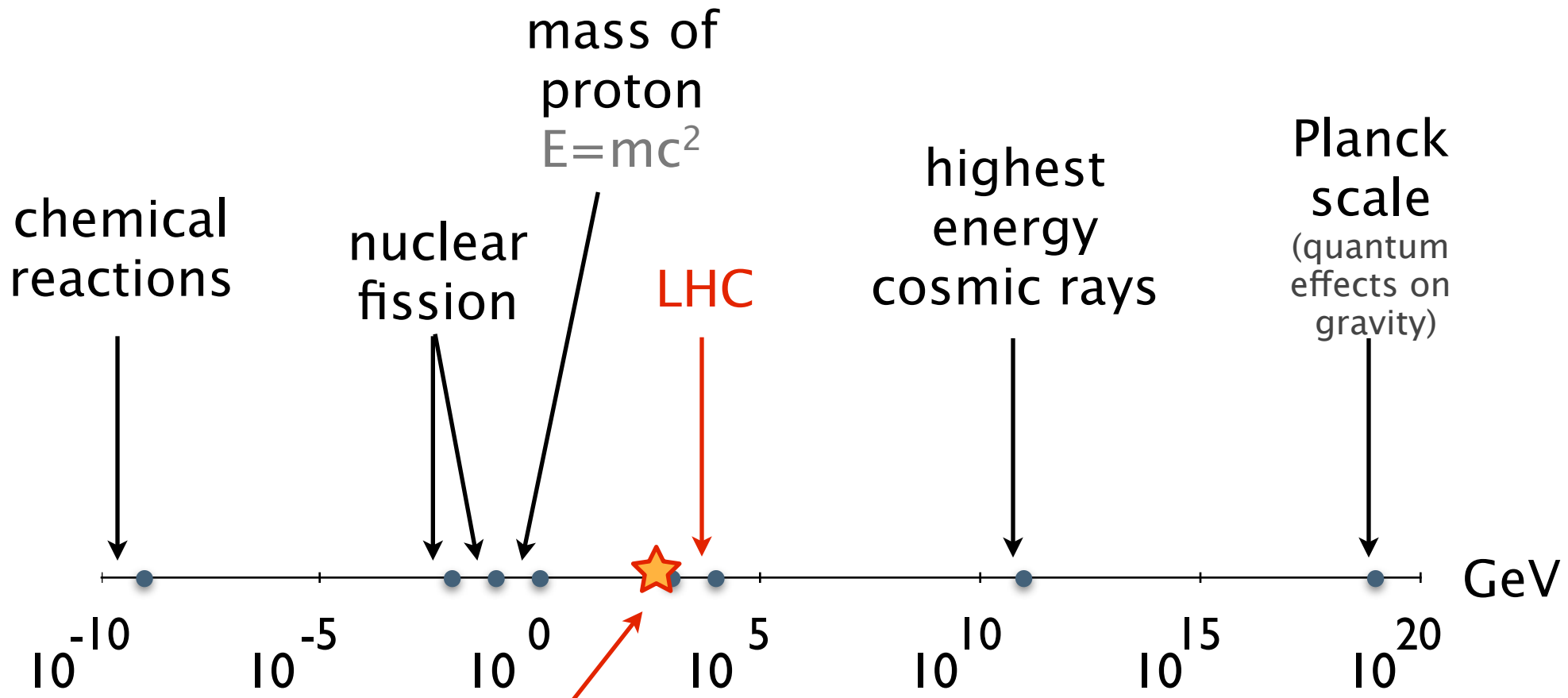
Energy scales in context



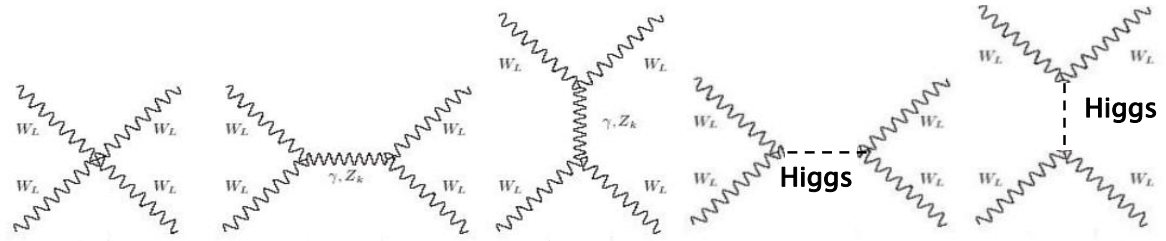
Energy scales in context



Energy scales in context

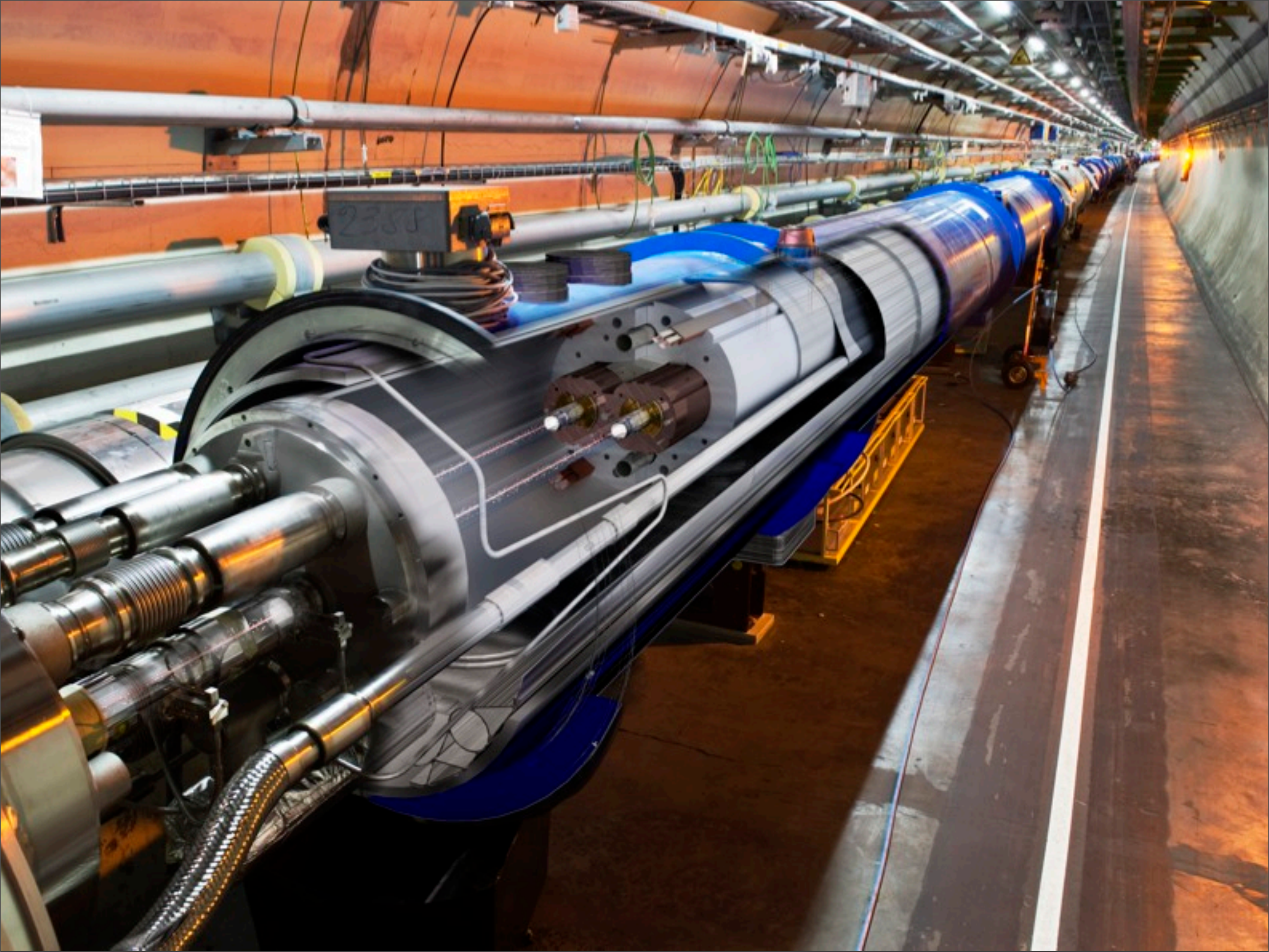


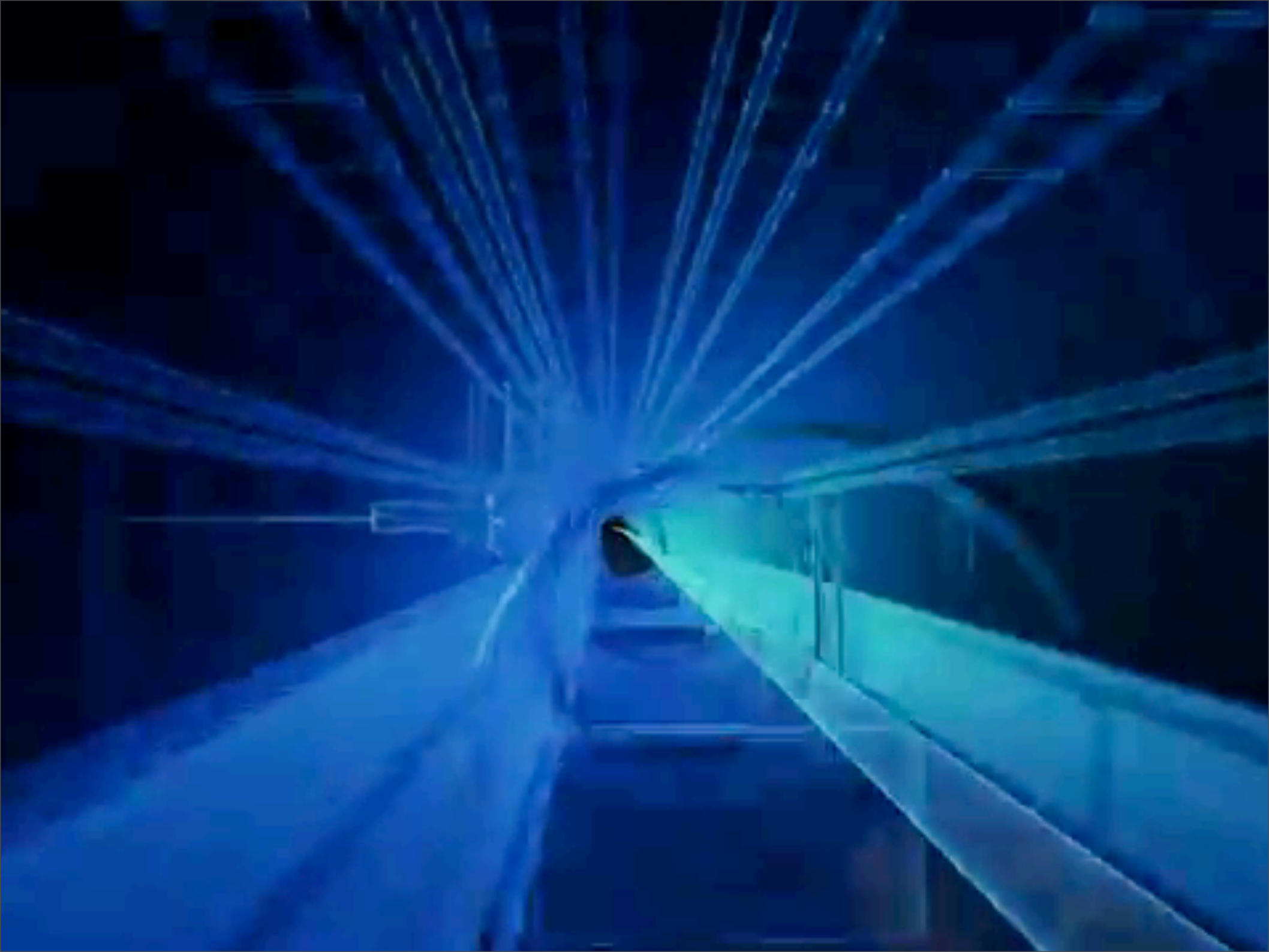
Expecting to see some new phenomena here

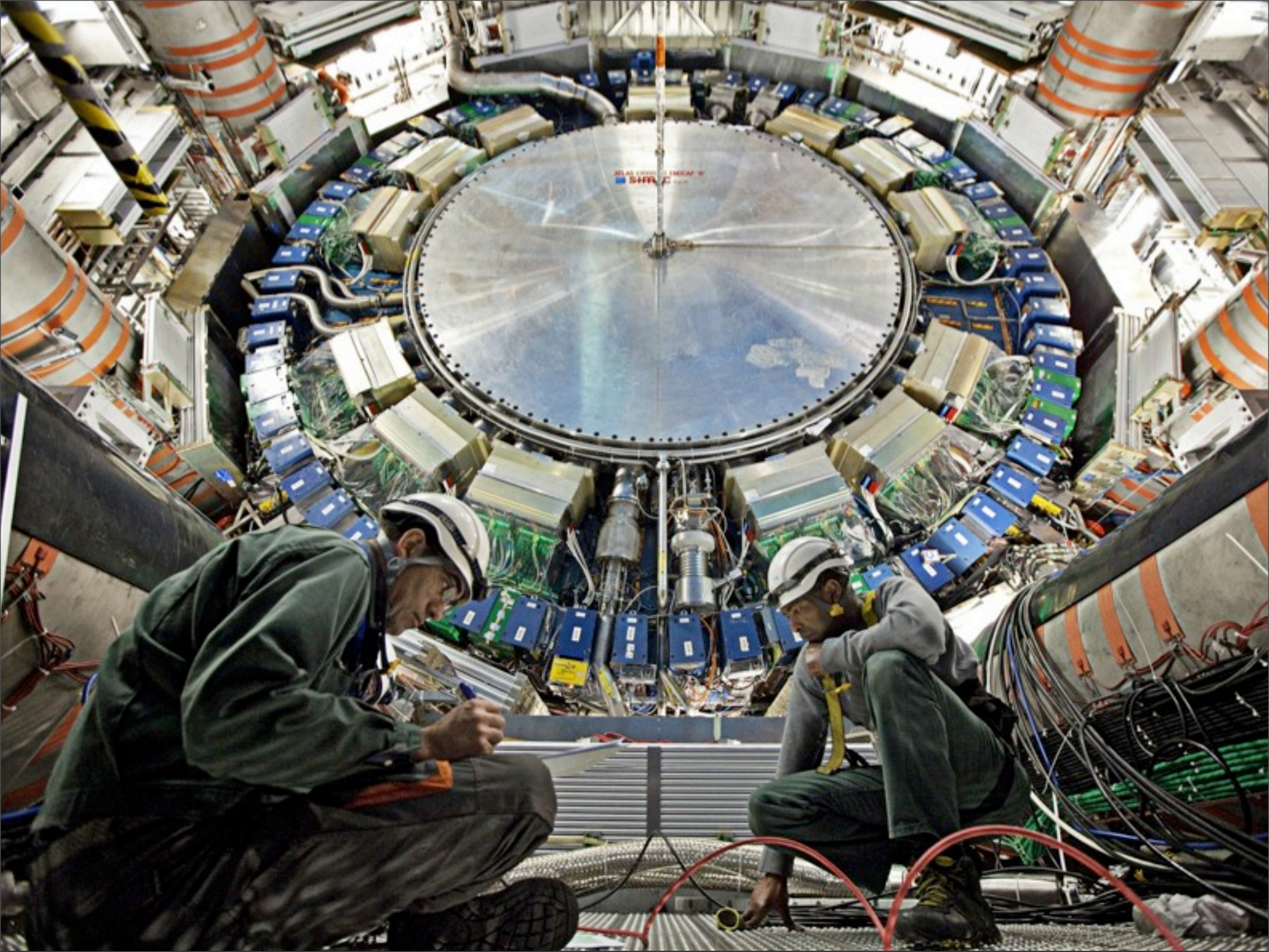


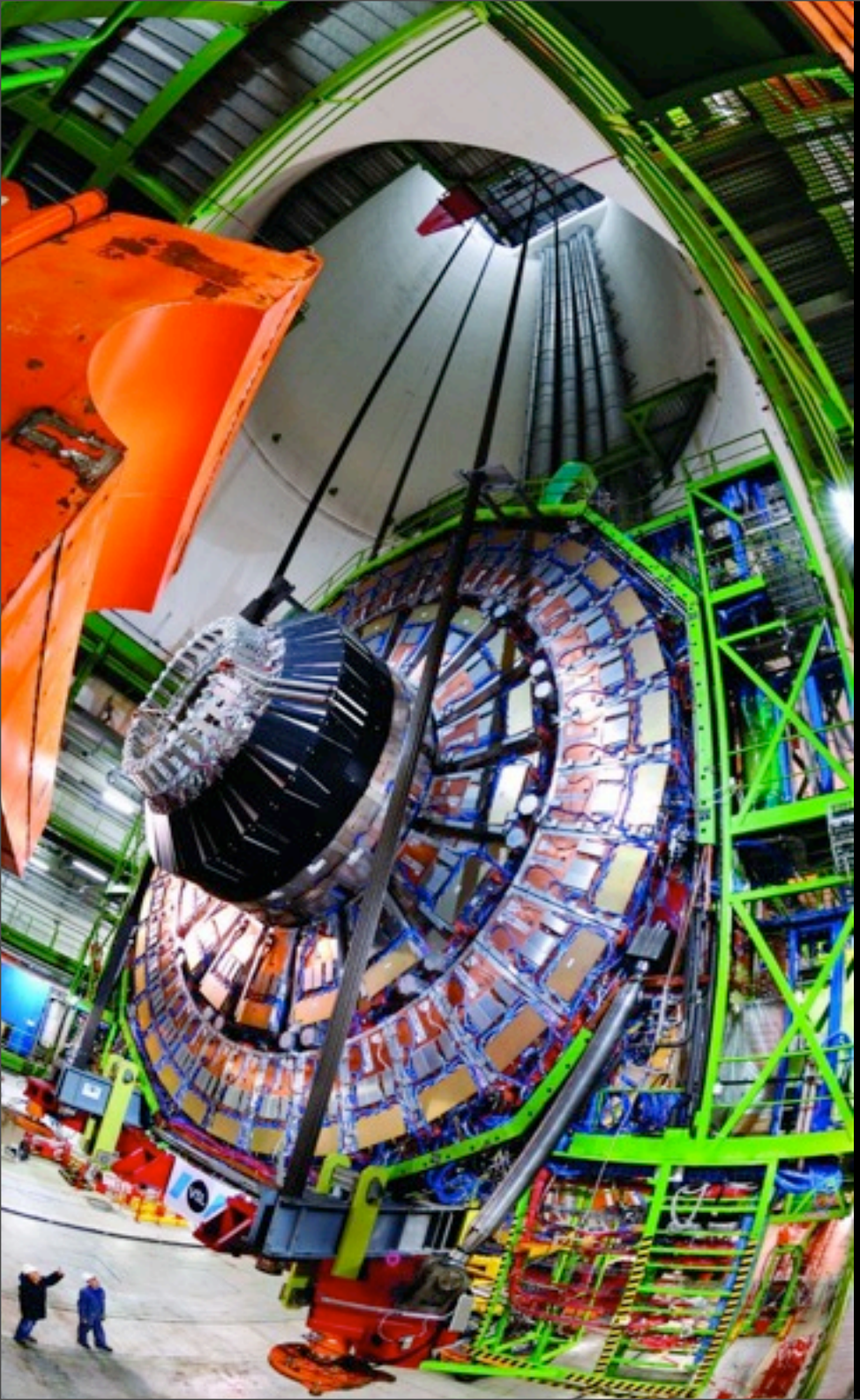
Violation of unitarity @ $\sqrt{s} \geq 1.7$ TeV

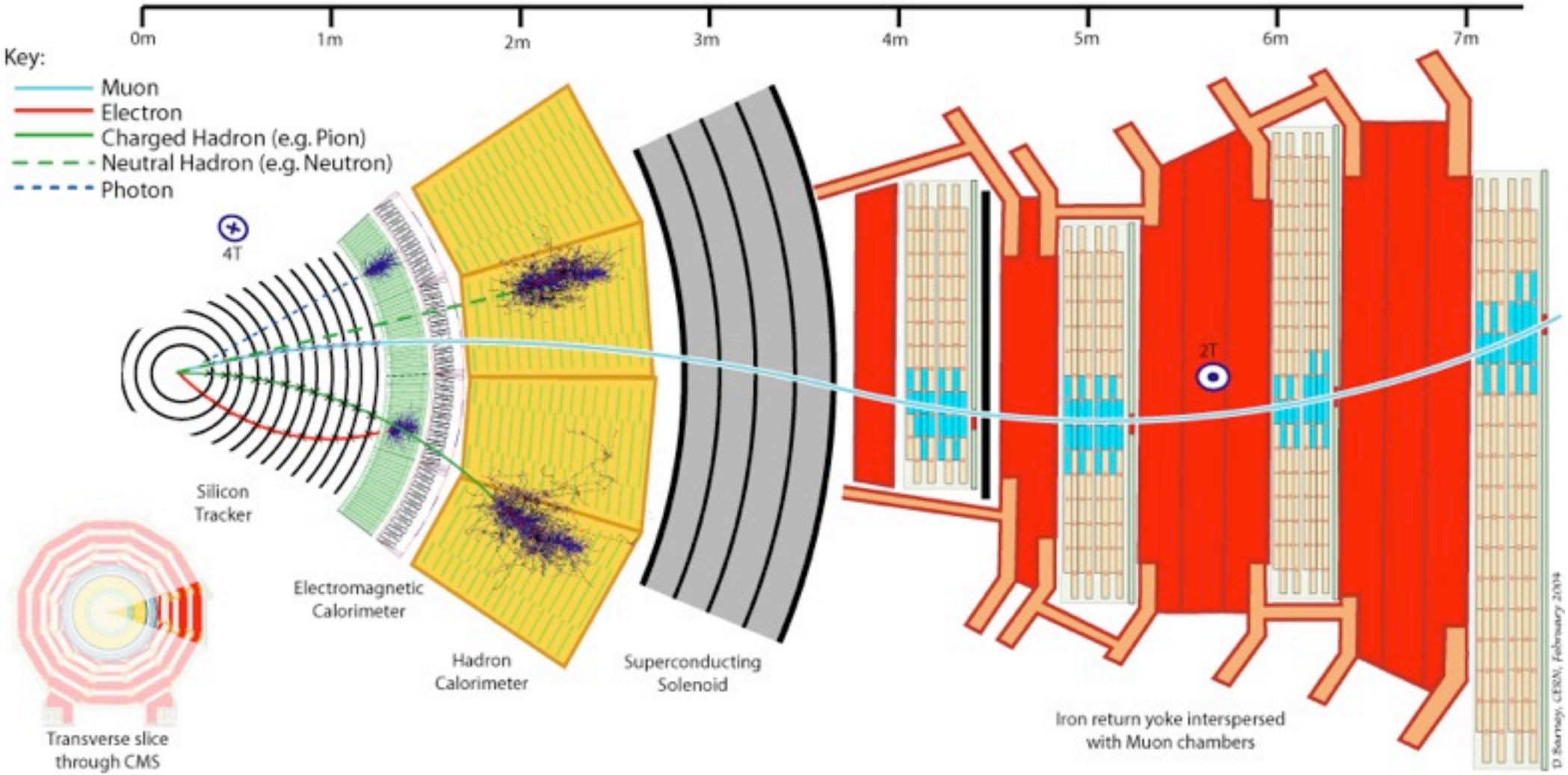
Restores unitarity if $m_H \lesssim 0.8$–1 TeV



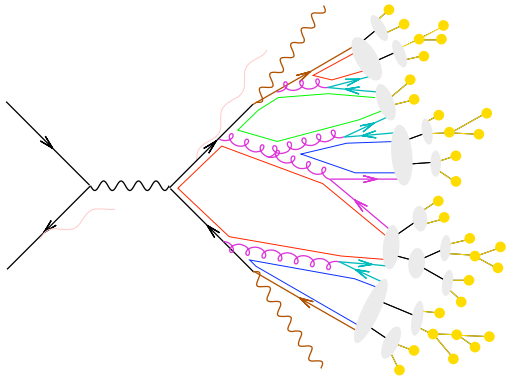


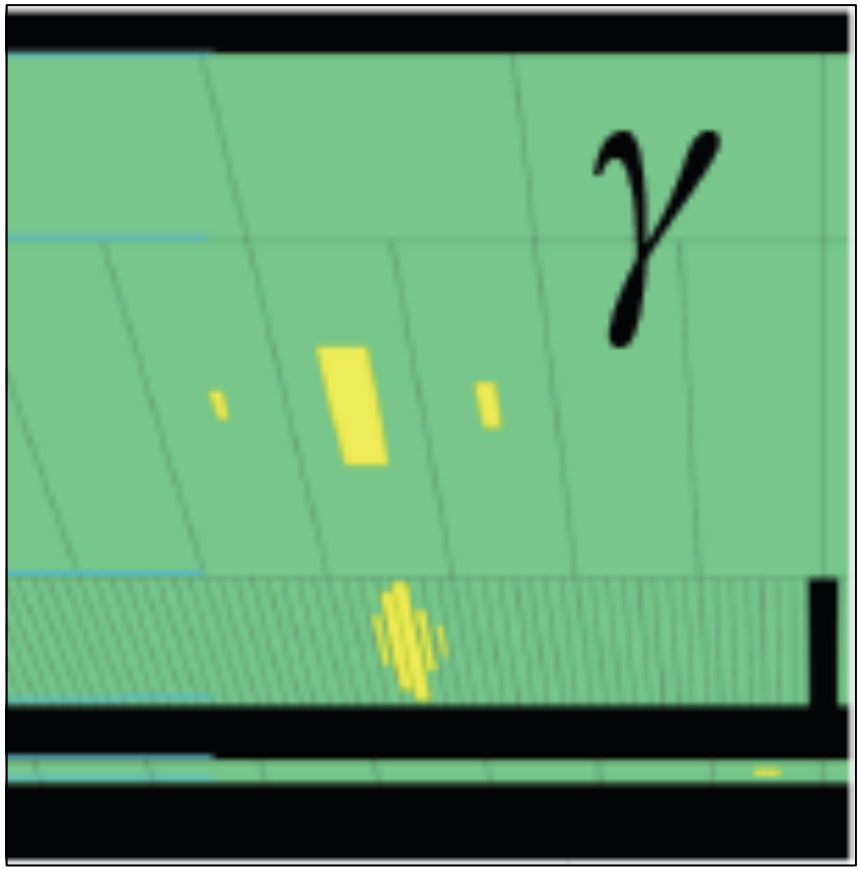


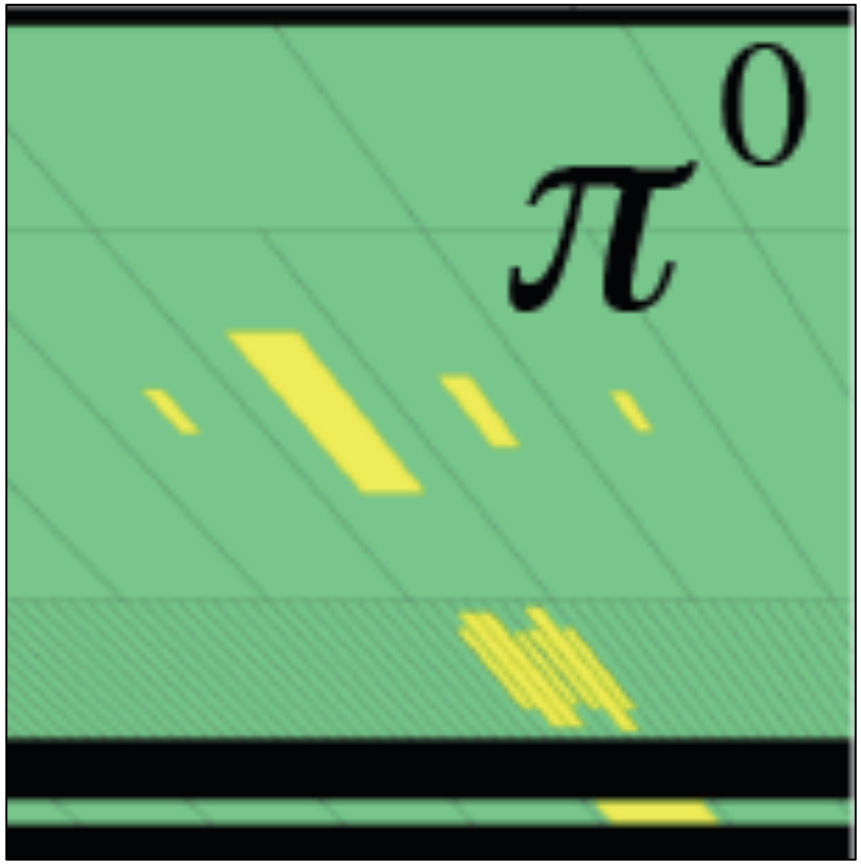



0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D Barney, CERN, February 2004

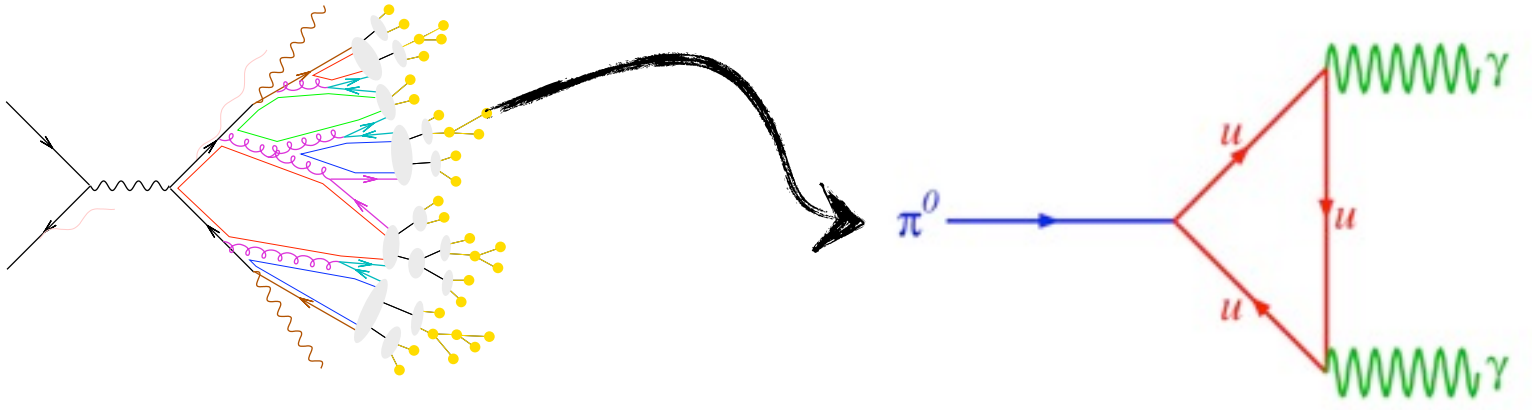
Backgrounds

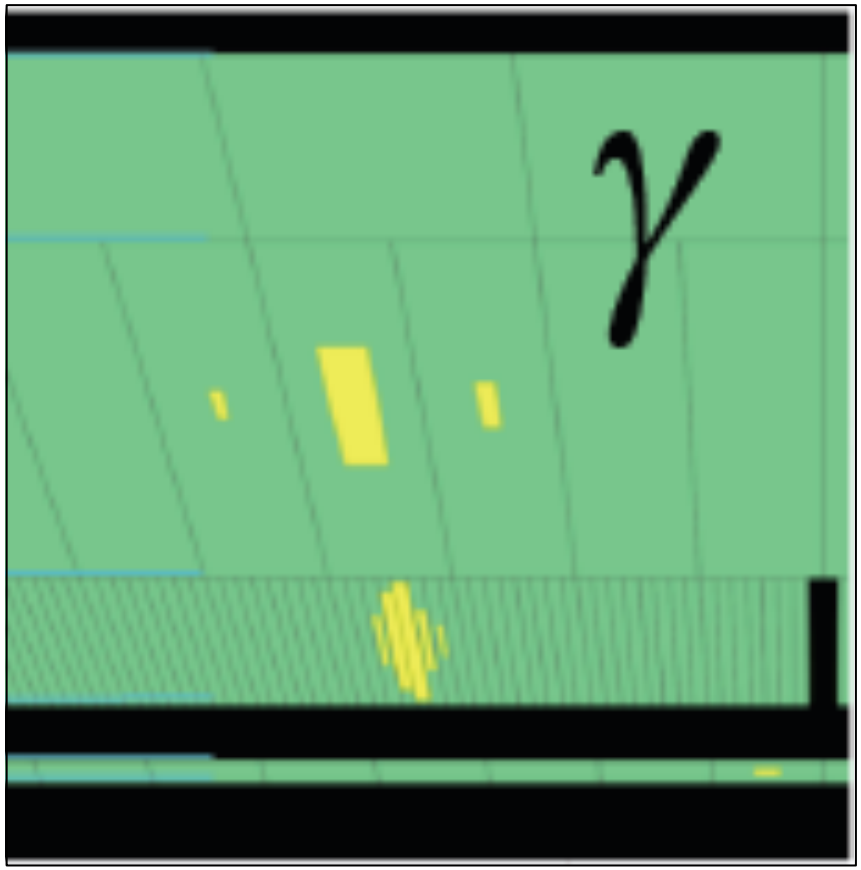


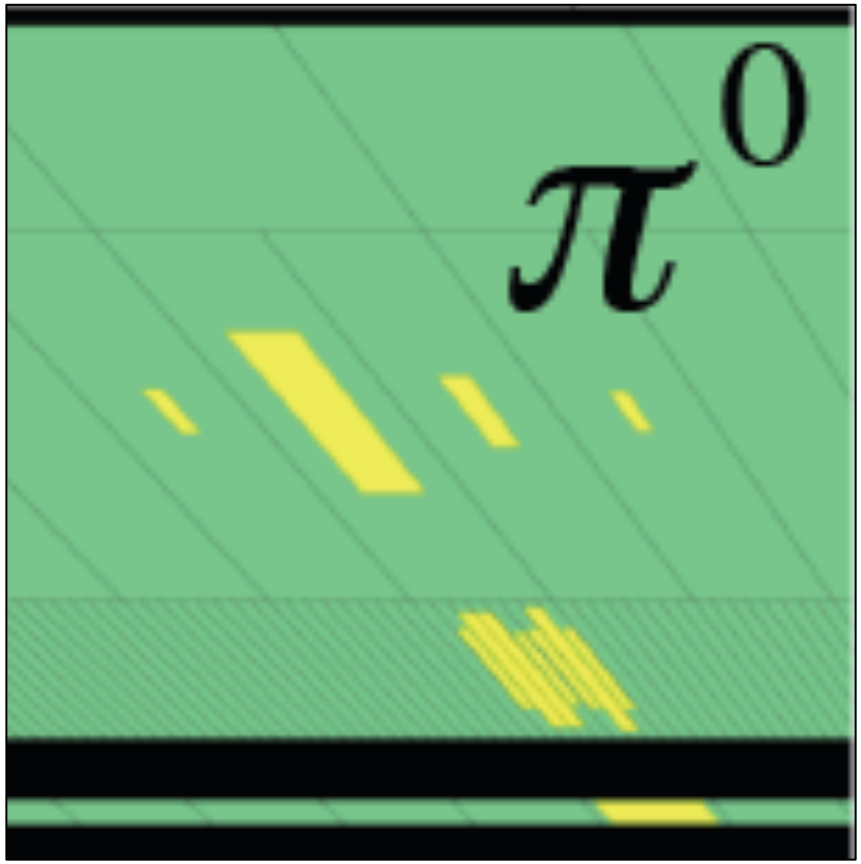




Backgrounds







Overview of Predictions

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma\text{ and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

1) The language of the Standard Model is Quantum Field Theory

Overview of Predictions

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

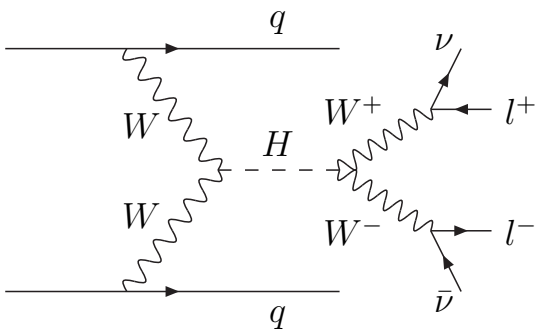
$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma\text{ and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

1) The language of the Standard Model is Quantum Field Theory

2) Perturbation Theory, Feynman Diagrams, and Factorization are used to construct Monte Carlo simulations of the interactions



Overview of Predictions

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

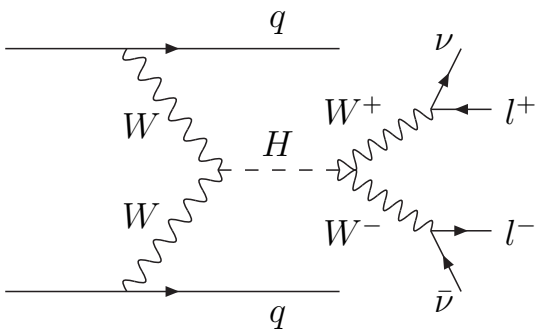
$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

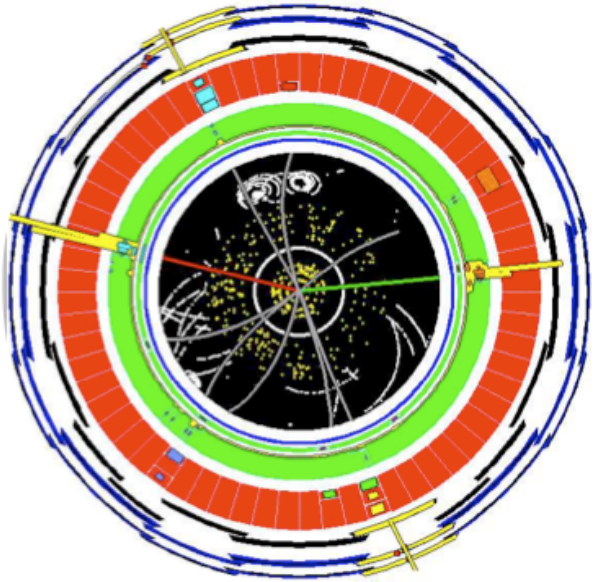
$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma \text{ and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

1) The language of the Standard Model is Quantum Field Theory

2) Perturbation Theory, Feynman Diagrams, and Factorization are used to construct Monte Carlo simulations of the interactions





3) The interaction of outgoing particles with the detector is simulated.

Overview of Predictions

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

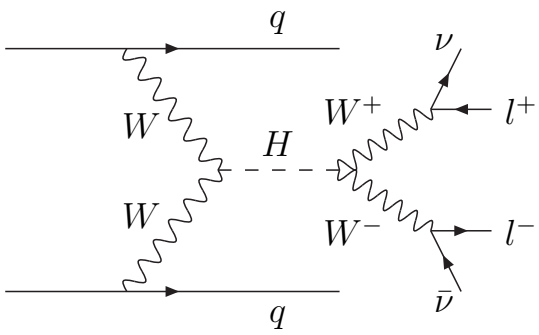
$$+ \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}}$$

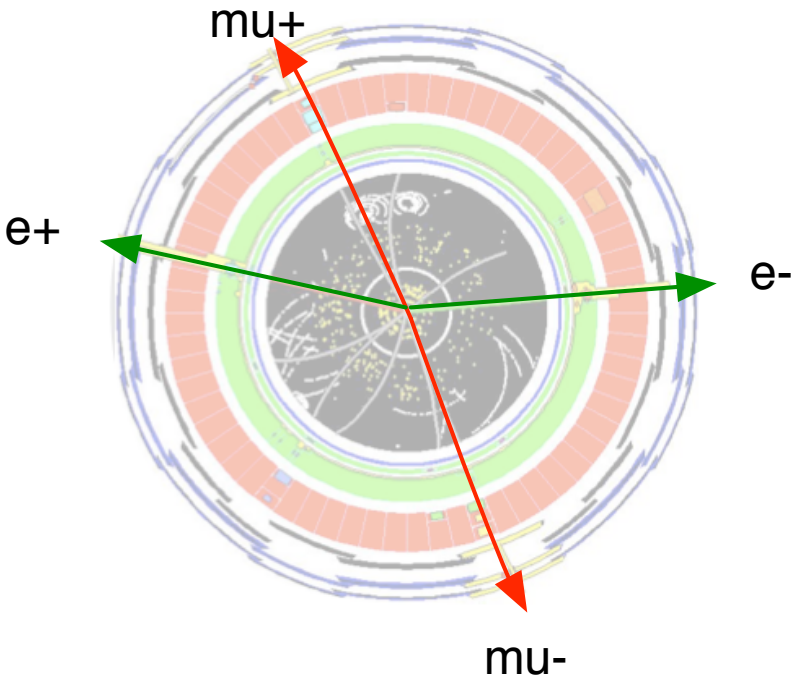
$$+ \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma\text{ and Higgs masses and couplings}}$$

$$+ \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$

1) The language of the Standard Model is Quantum Field Theory

2) Perturbation Theory, Feynman Diagrams, and Factorization are used to construct Monte Carlo simulations of the interactions





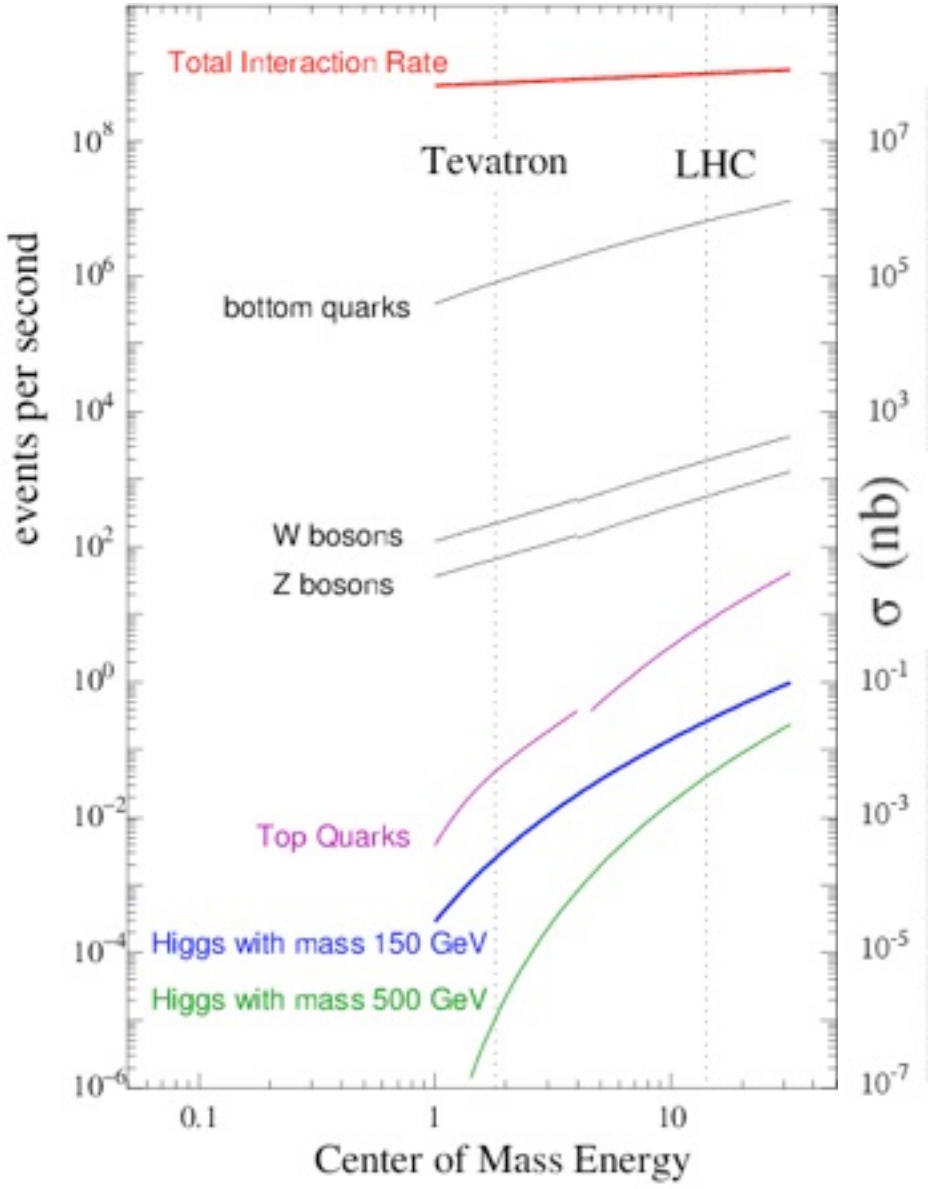
3) The interaction of outgoing particles with the detector is simulated.

4) Finally, we run algorithms on the simulated data as if they were from real collisions.

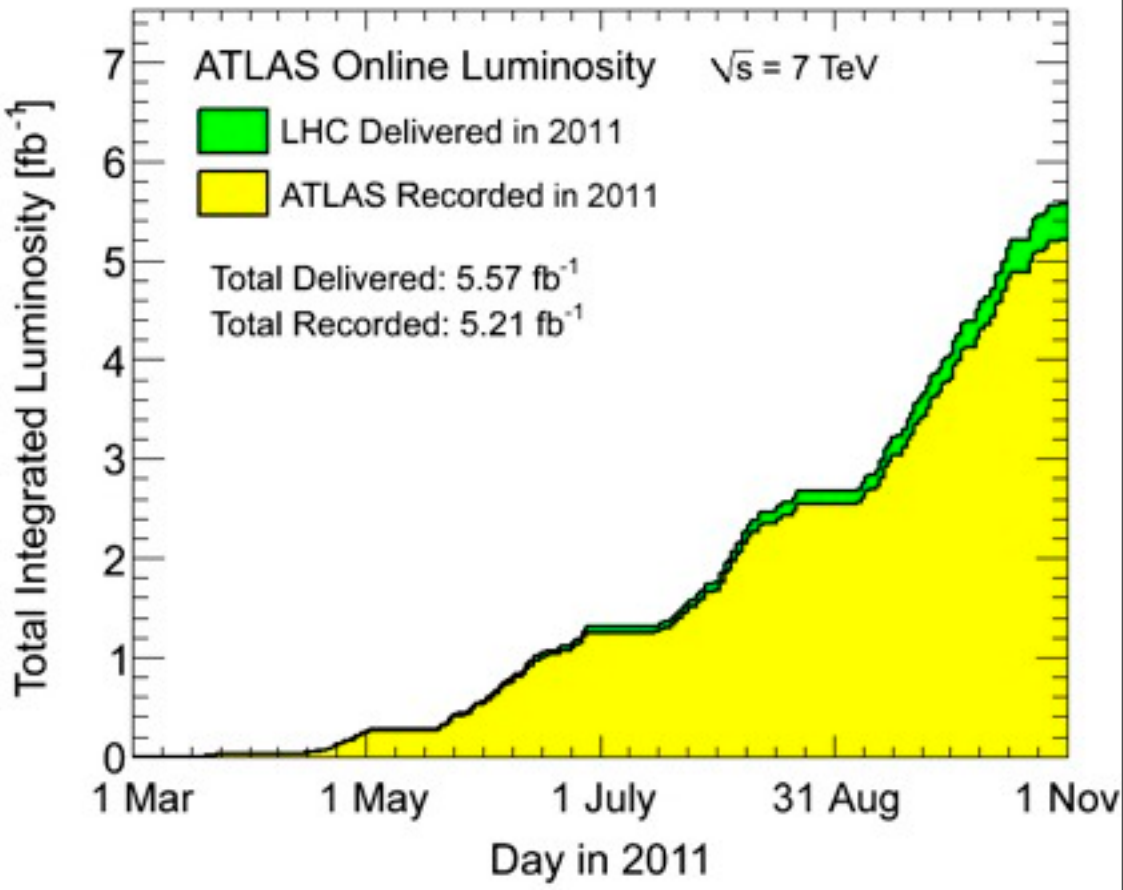
Number of collisions in 2011

expected number of scatterings = cross section [cm^2] x Luminosity [1/cm^2]

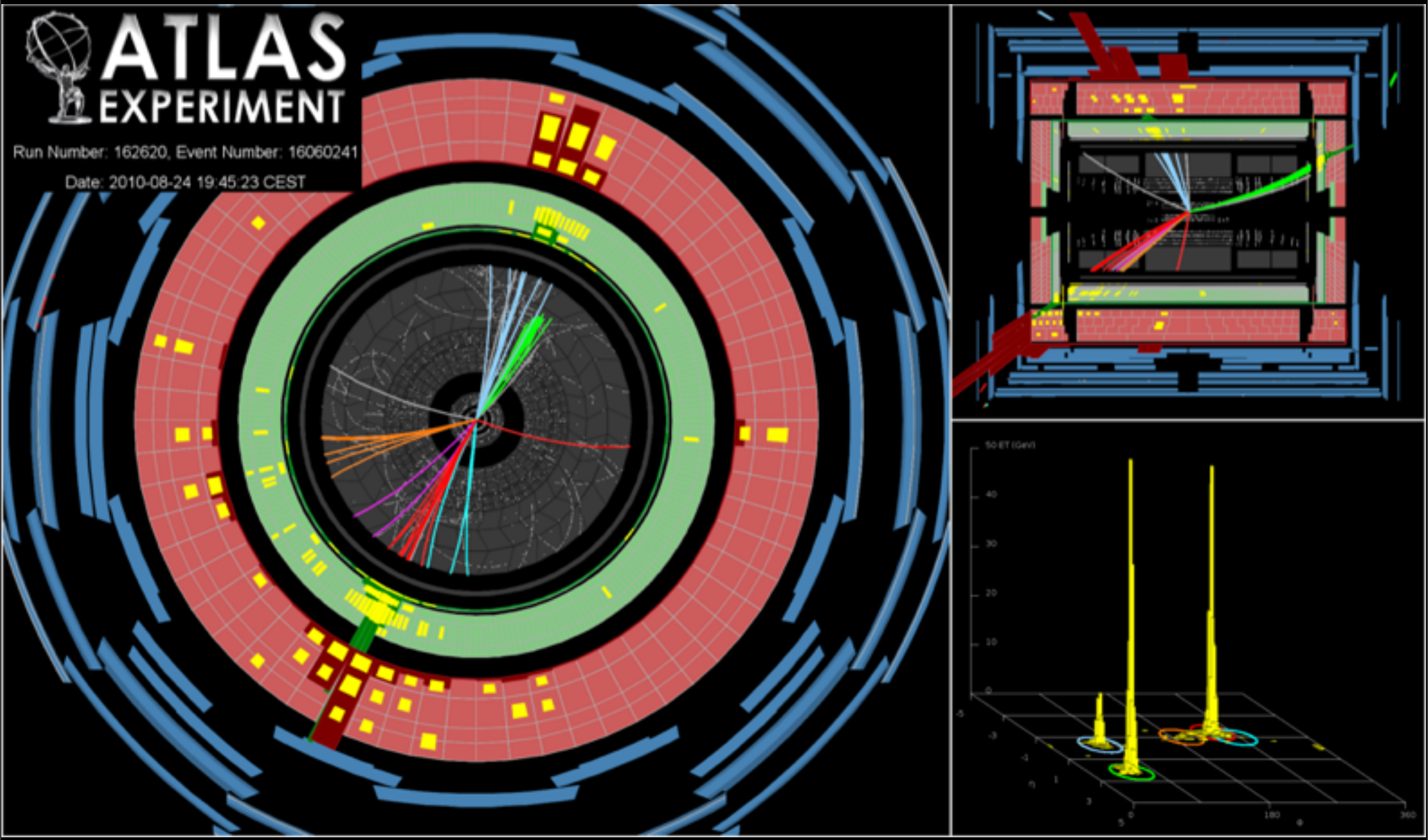
$$\langle N \rangle = \sigma L$$



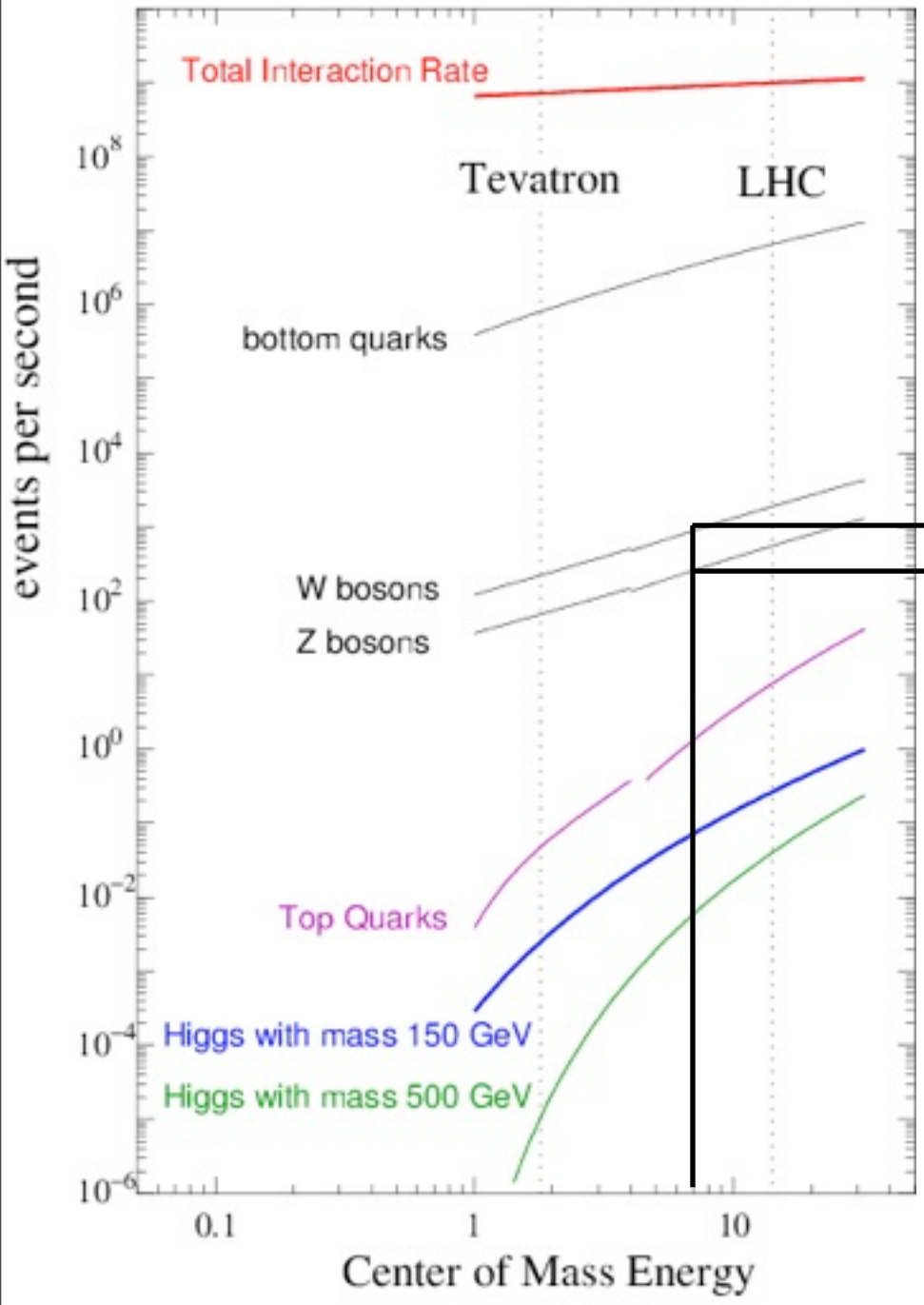
80 mb · 5 fb^{-1} = 4·10^{14} collisions

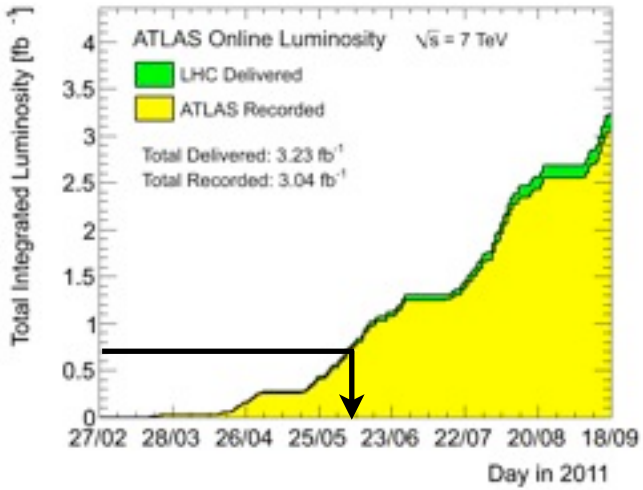


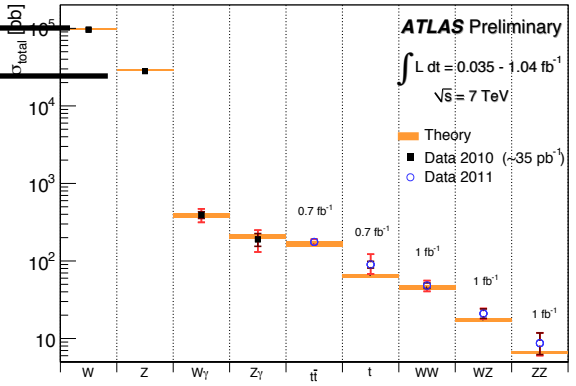
1 nb = 10^{-33} cm^2

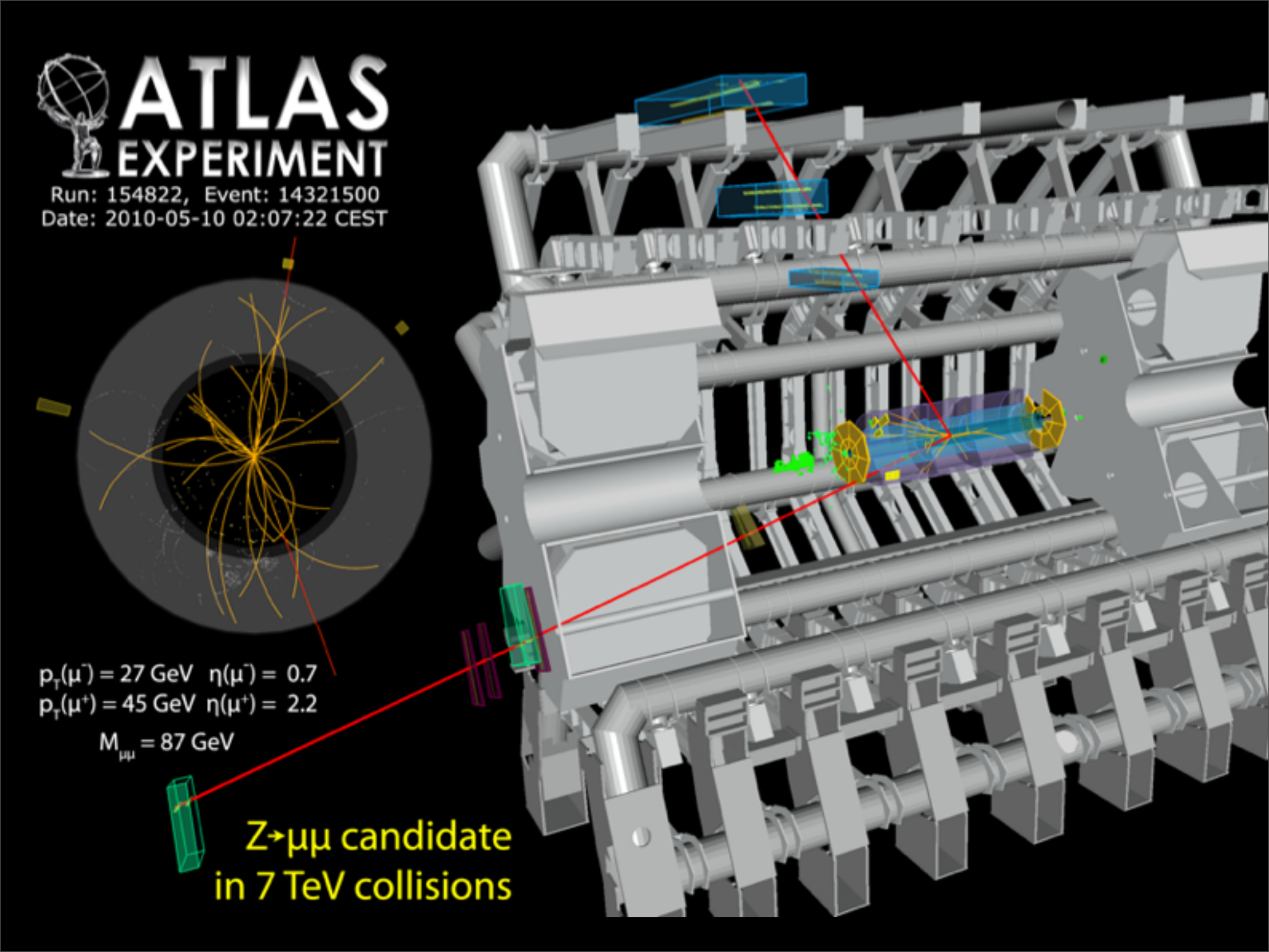

ATLAS
EXPERIMENT
Run Number: 162620, Event Number: 16060241
Date: 2010-08-24 19:45:23 CEST

The steady march of progress

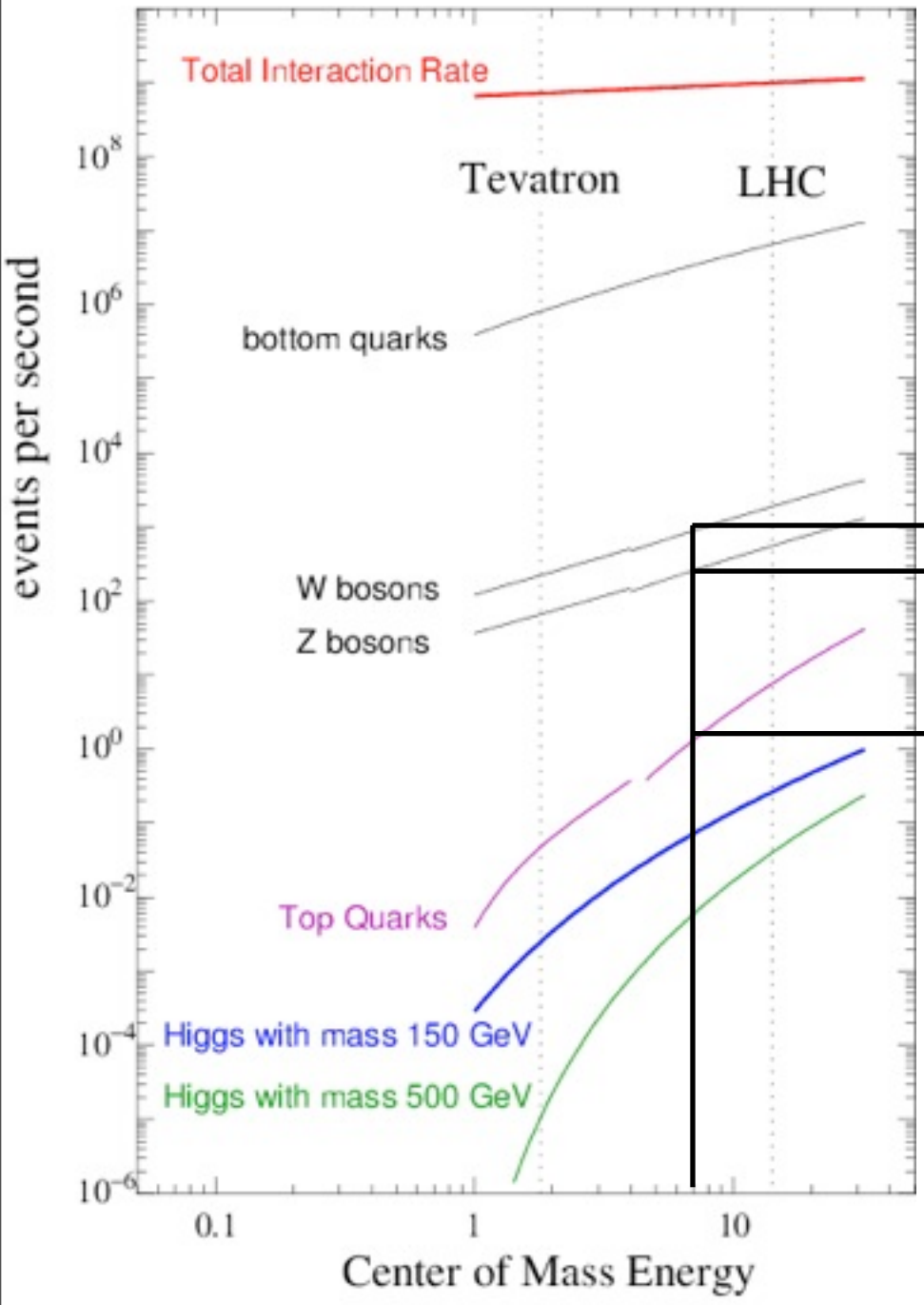


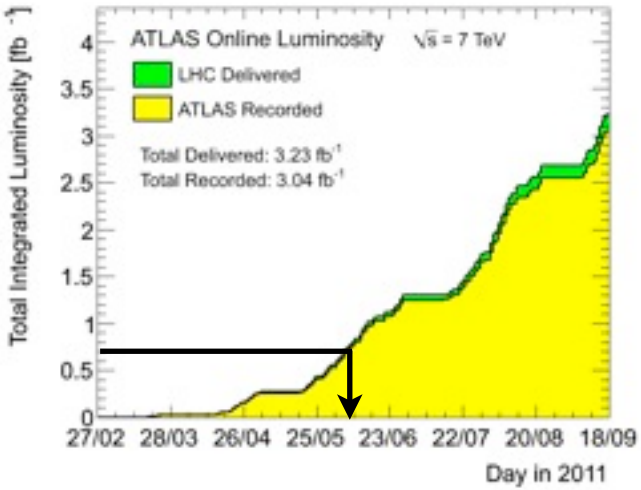


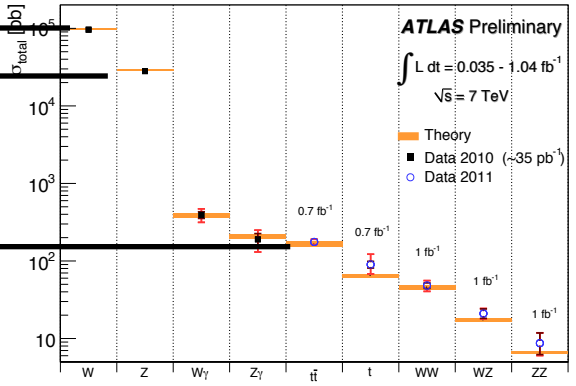



ATLAS
EXPERIMENT
Run: 154822, Event: 14321500
Date: 2010-05-10 02:07:22 CEST
$p_T(\mu^-) = 27$ GeV $\eta(\mu^-) = 0.7$
$p_T(\mu^+) = 45$ GeV $\eta(\mu^+) = 2.2$
$M_{\mu\mu} = 87$ GeV
Z→μμ candidate
in 7 TeV collisions

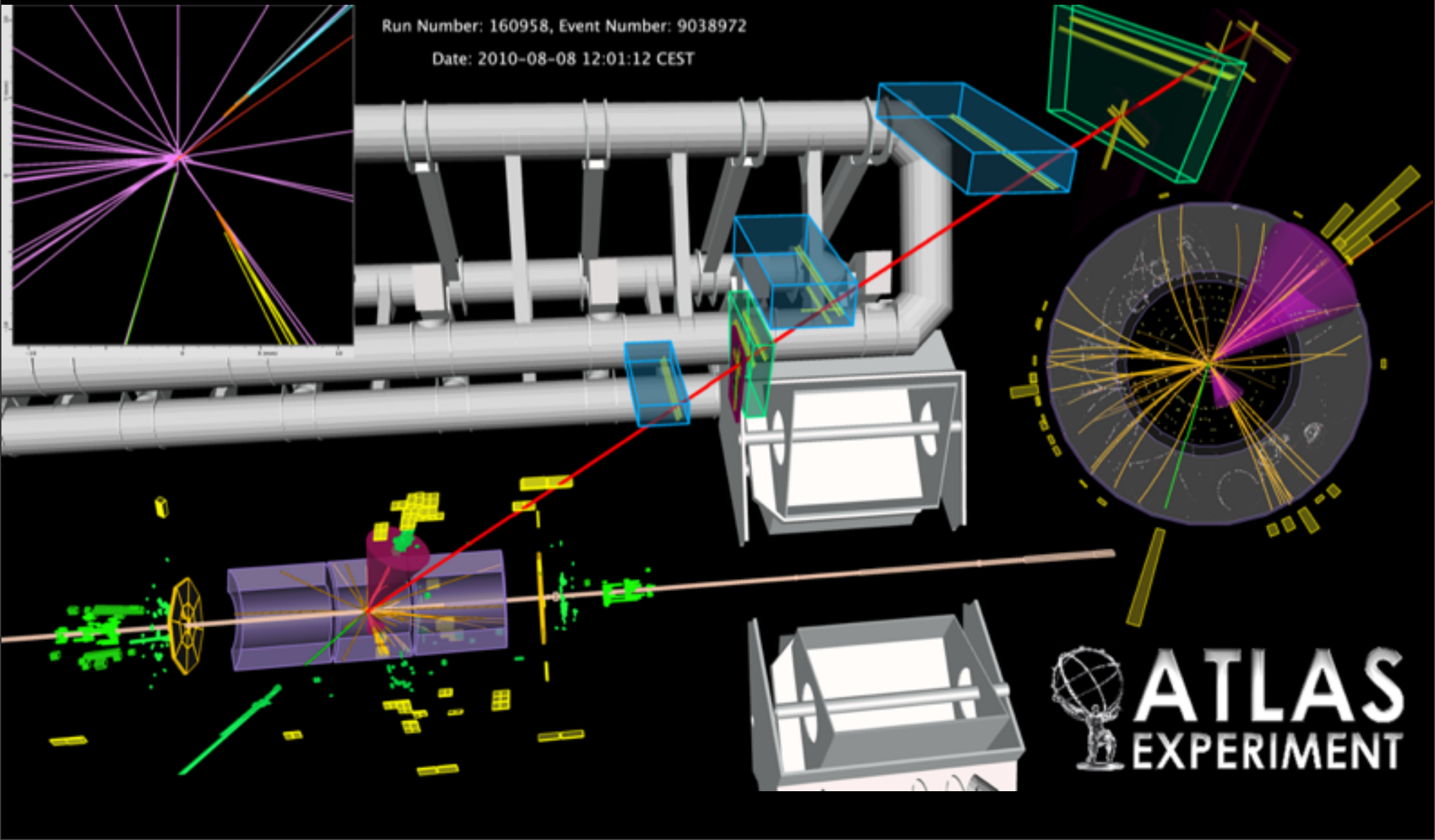
The steady march of progress



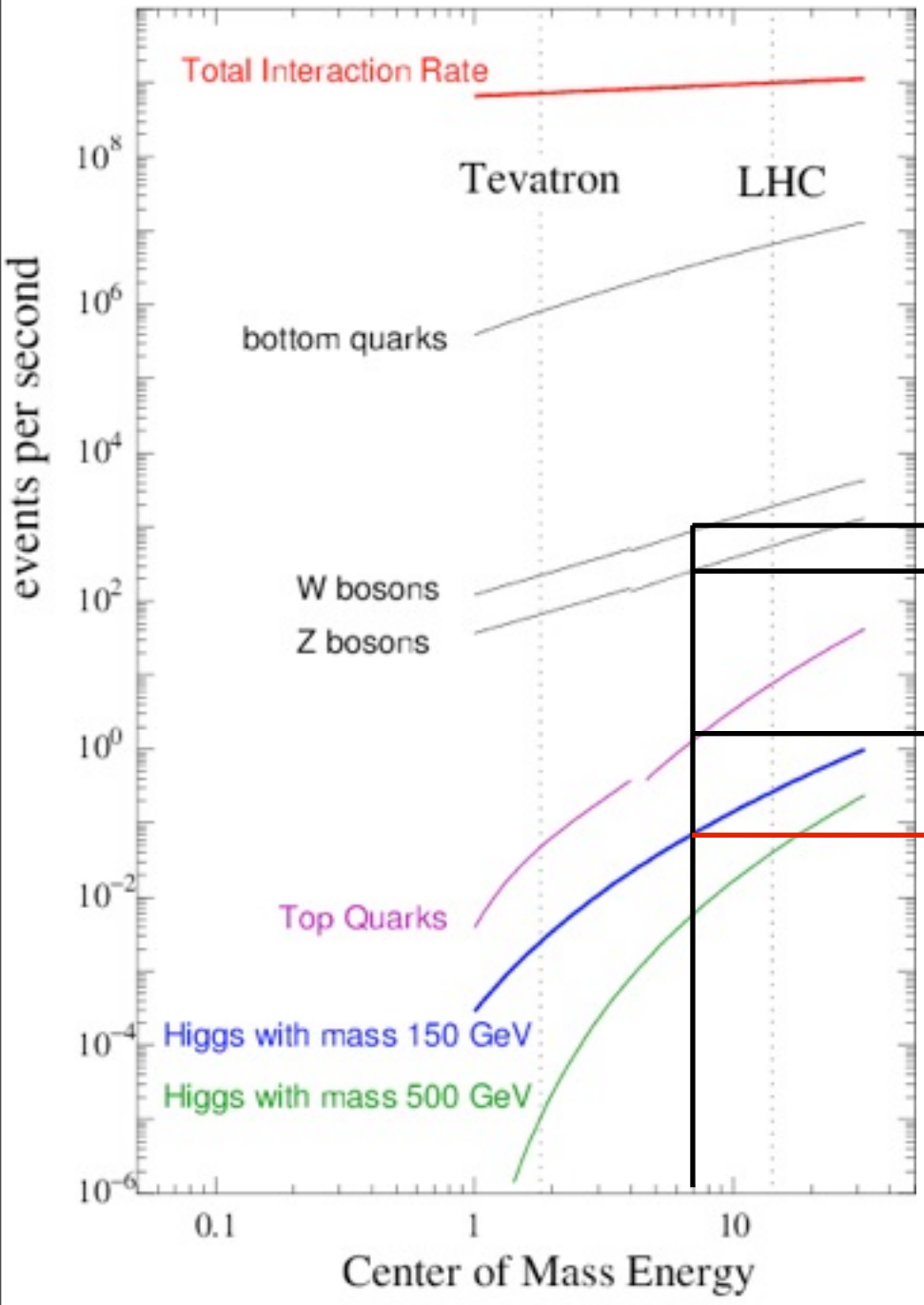


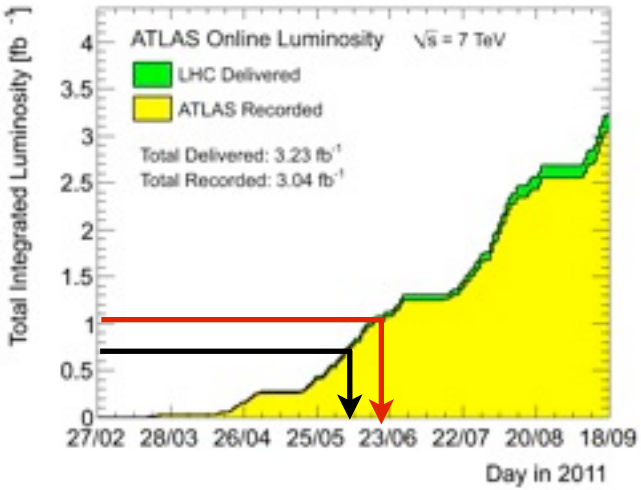


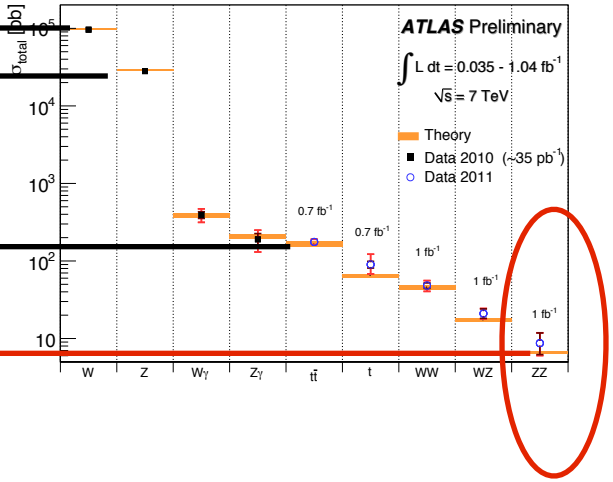
Top quark pair decaying to bb eμ $E_{T,miss}$

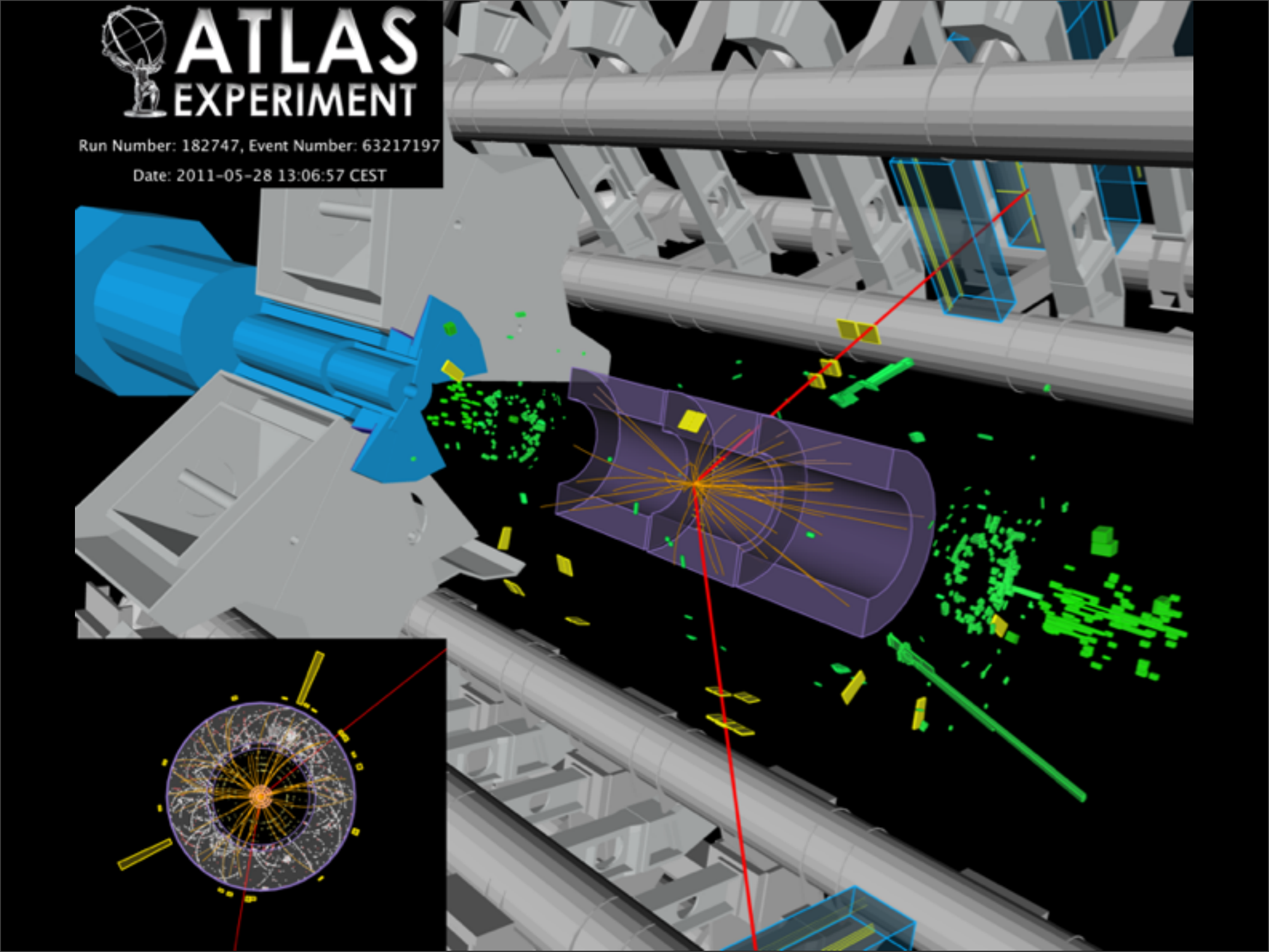


The steady march of progress

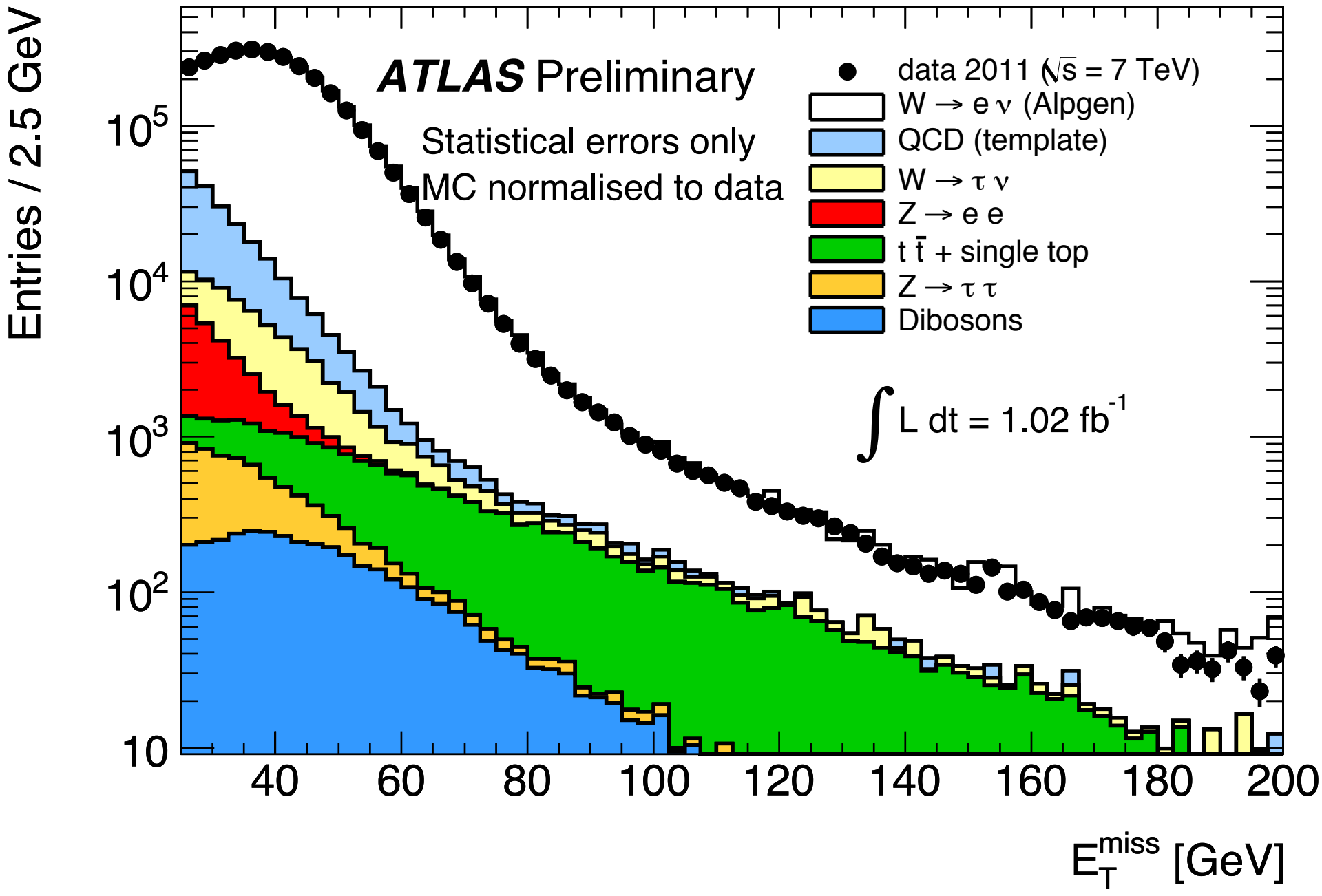




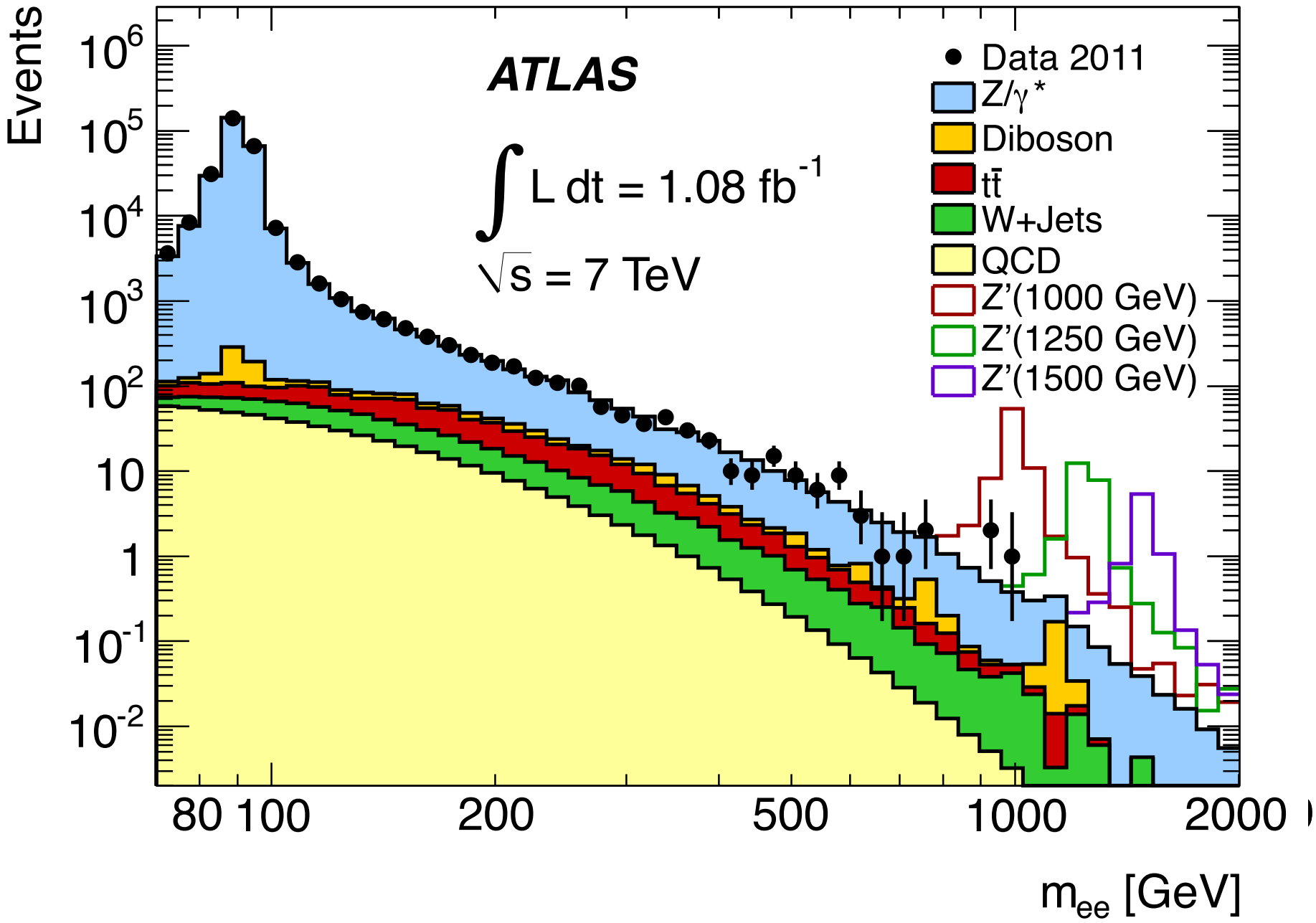



ATLAS
EXPERIMENT
Run Number: 182747, Event Number: 63217197
Date: 2011-05-28 13:06:57 CEST

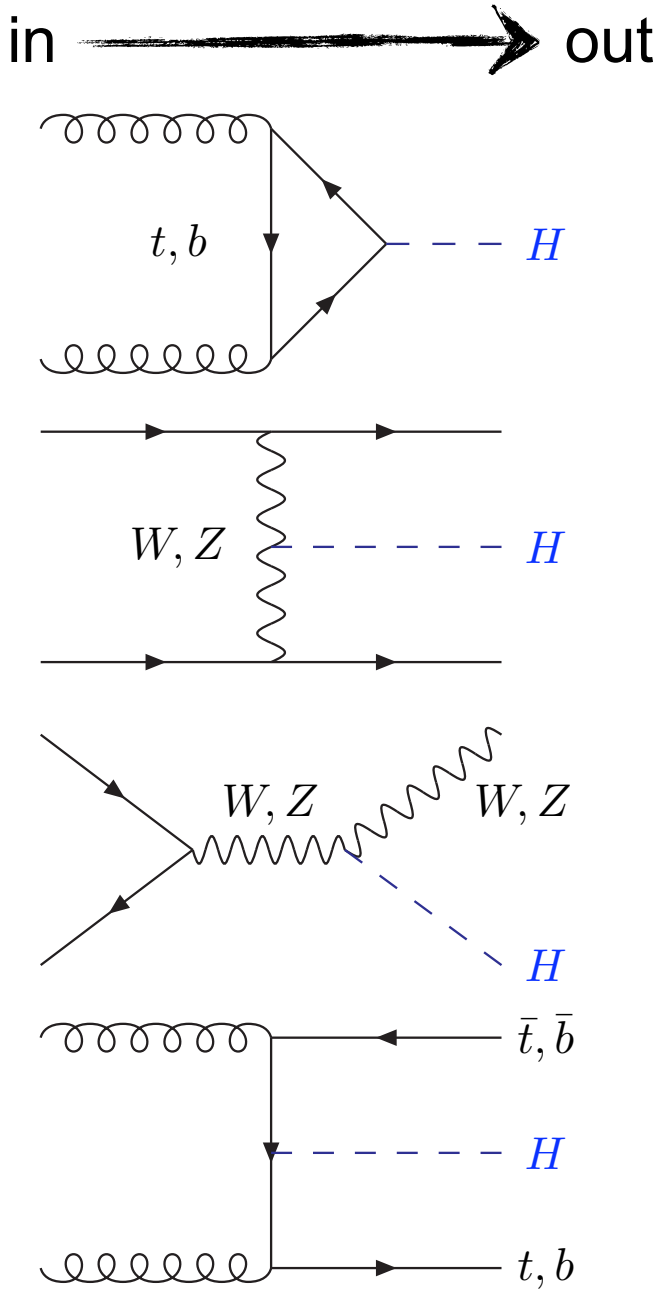
How good is the modeling?

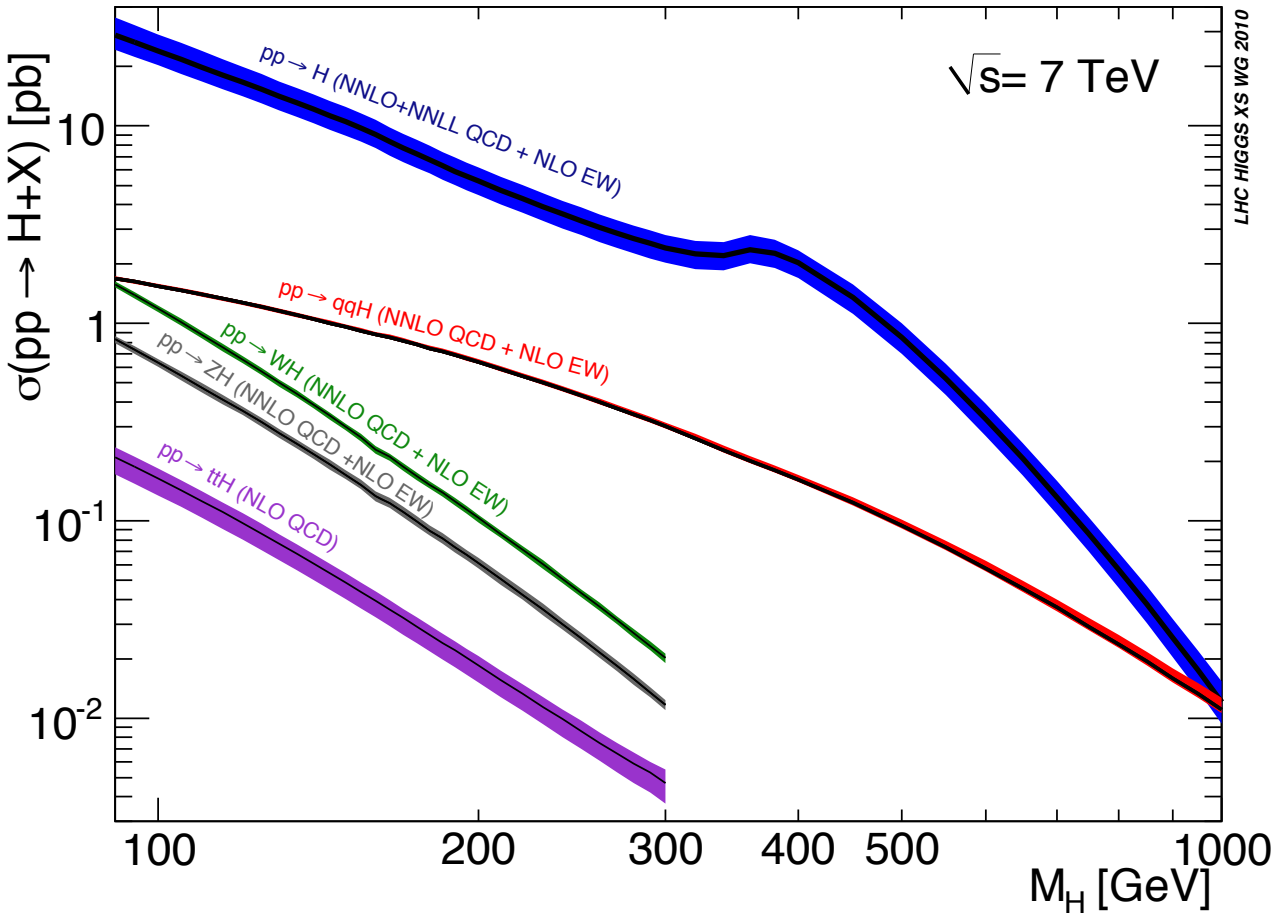


How good is the modeling?



Standard Model Higgs Properties



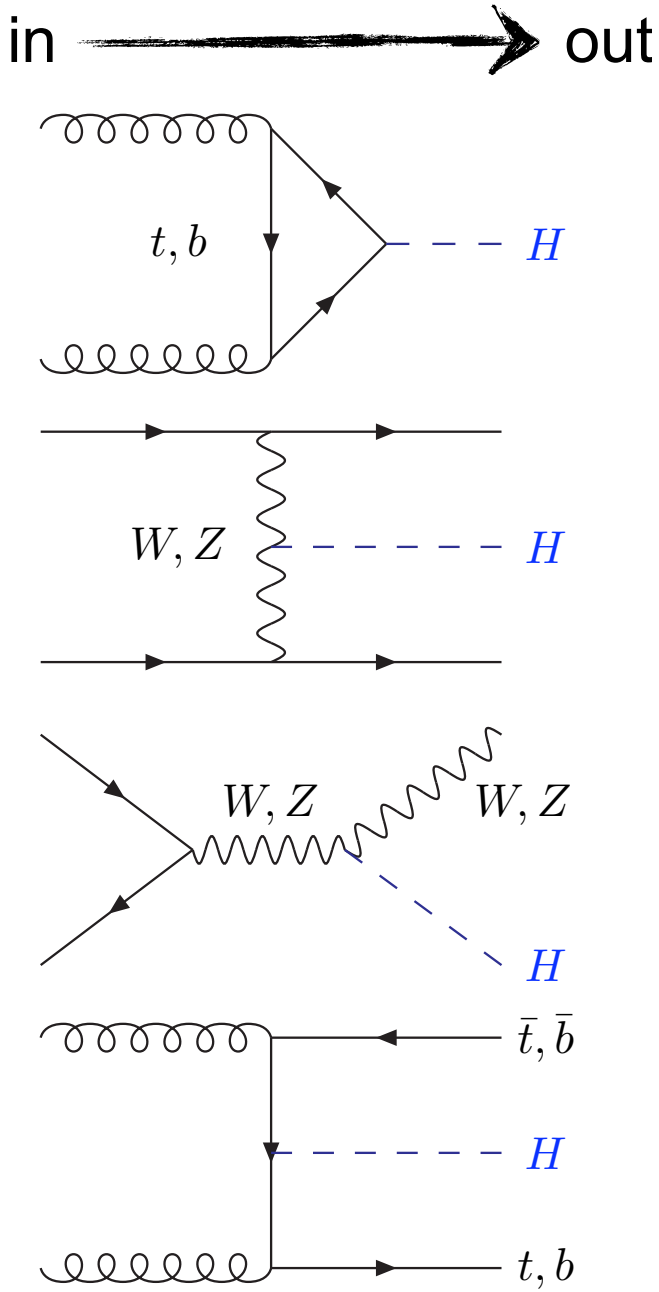


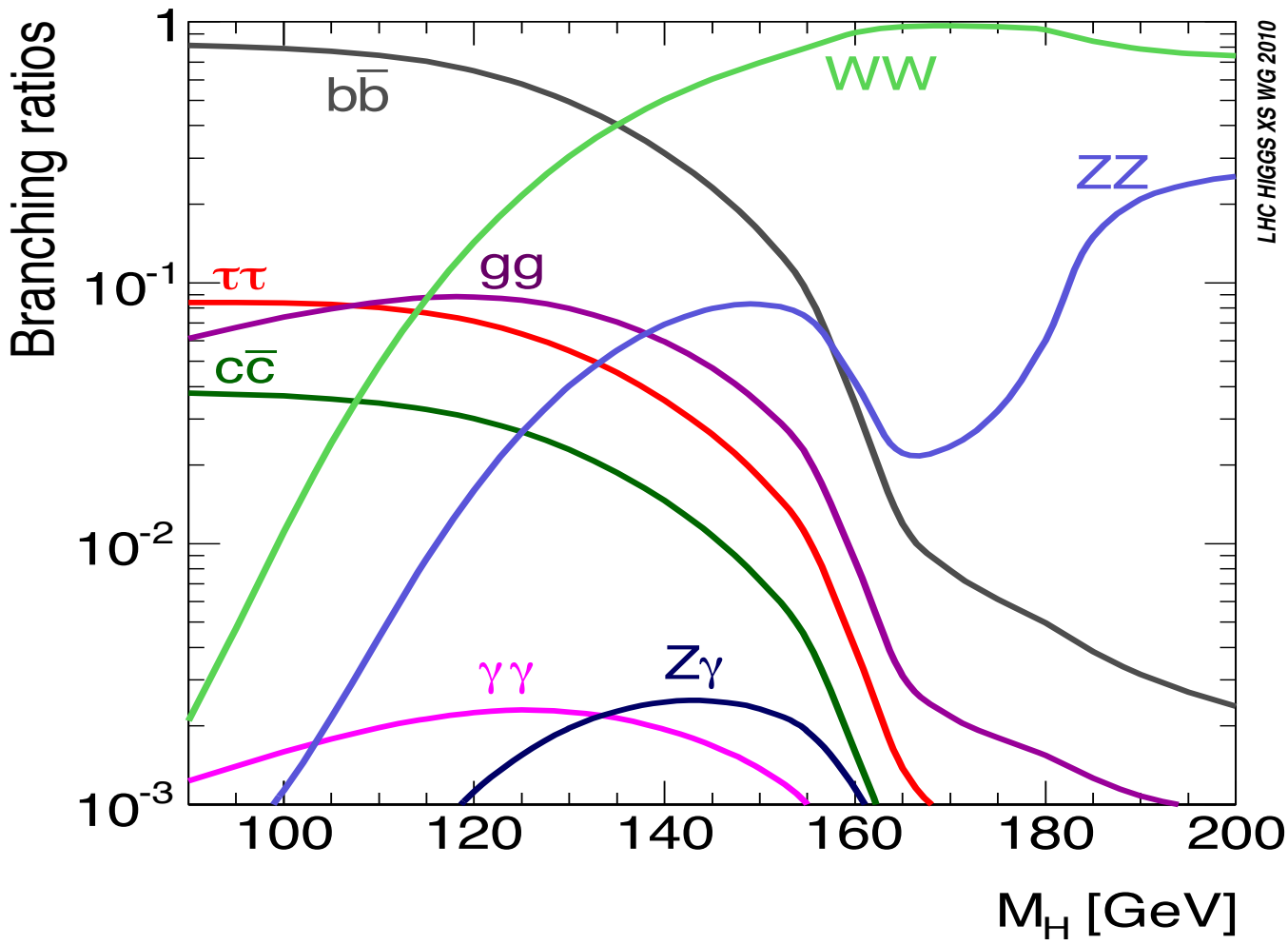
The Higgs boson can be produced via different interactions.

Production cross section σ depends on the unknown Higgs mass

Standard Model Higgs Properties

Center for Cosmology and Particle Physics

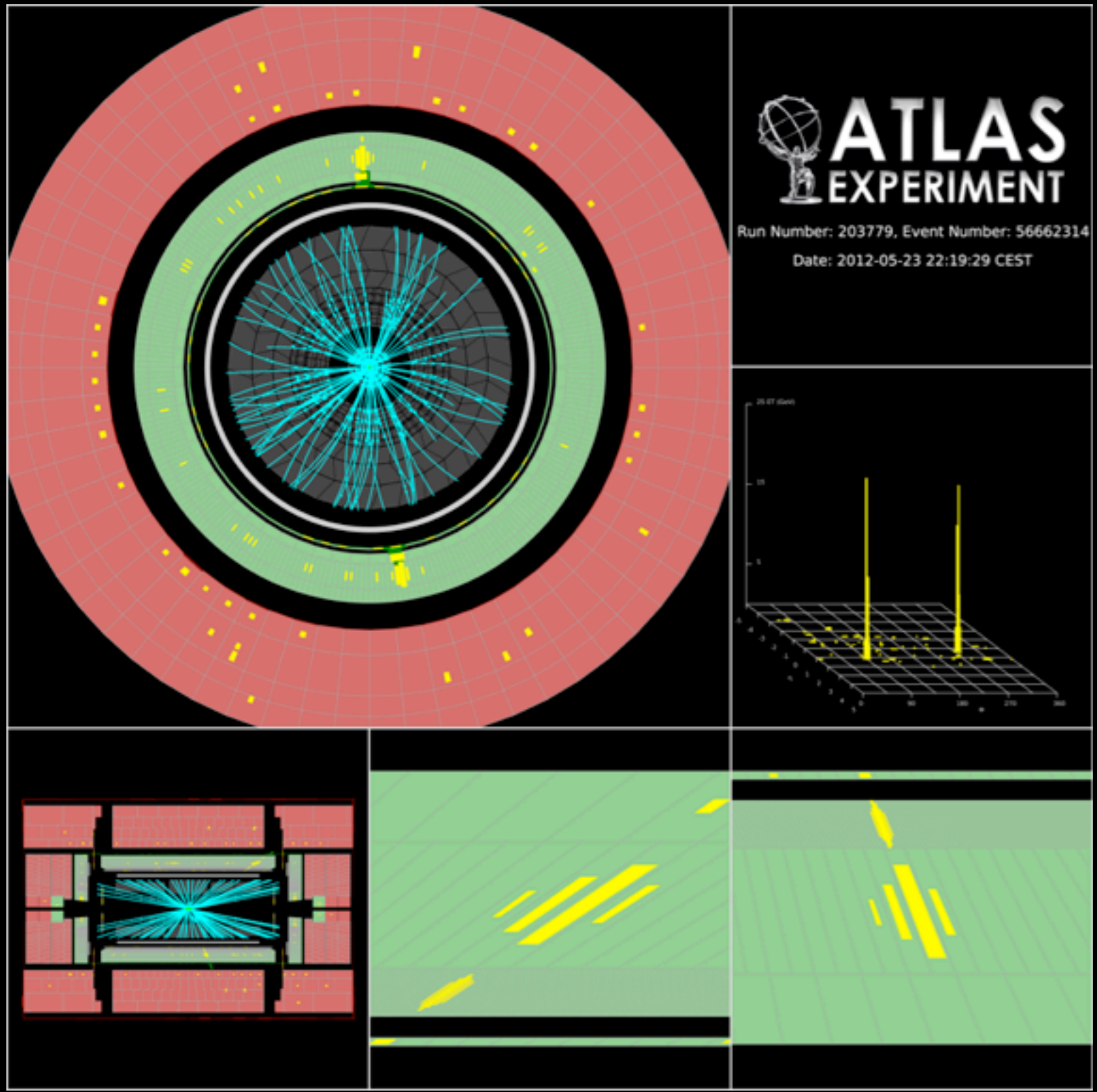




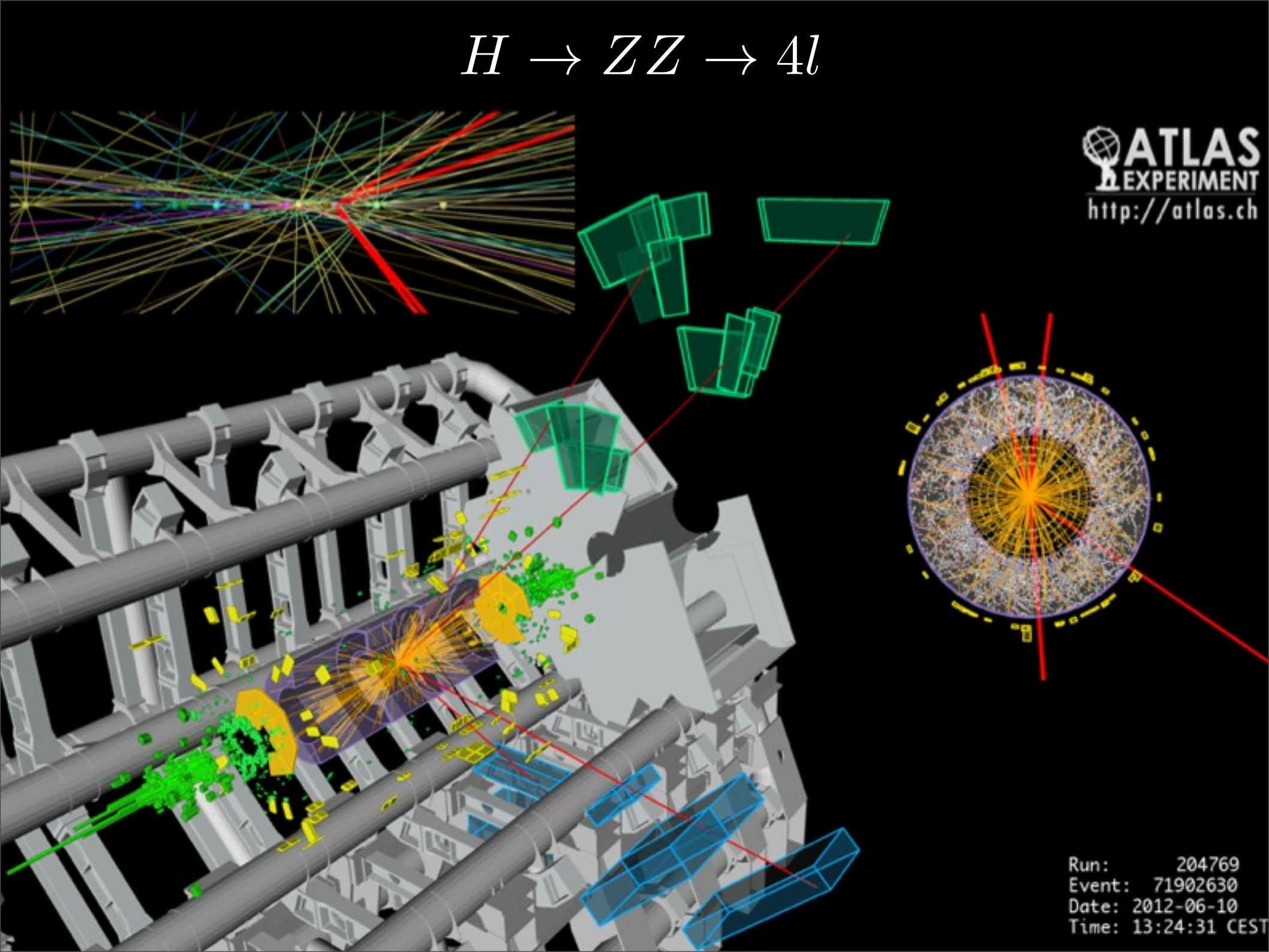
The Higgs boson then decays in one of several possible final states

The fraction of each decay mode also depends on the unknown Higgs mass

$H \to \gamma\gamma$



$H \to ZZ \to 4l$


ATLAS
EXPERIMENT
http://atlas.ch
Run: 204769
Event: 71902630
Date: 2012-06-10
Time: 13:24:31 CEST

Putting the Higgs back together again

Don't believe the media: $E \neq mc^2$

Putting the Higgs back together again

Don't believe the media: $E \neq mc^2$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Putting the Higgs back together again

Don't believe the media: $E \neq mc^2$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Every Physics I student knows energy and momentum are conserved

$$E_{\text{Higgs}} = E_{\text{before}} = E_{\text{after}} = \sum_i E_i$$

$$\vec{p}_{\text{Higgs}} = \vec{p}_{\text{before}} = \vec{p}_{\text{after}} = \sum_i \vec{p}_i$$

Putting the Higgs back together again

Don't believe the media: $E \neq mc^2$

What Einstein really said:

$$E^2 = (mc^2)^2 + (|\vec{p}|c)^2$$

Every Physics I student knows energy and momentum are conserved

$$E_{\text{Higgs}} = E_{\text{before}} = E_{\text{after}} = \sum_i E_i$$

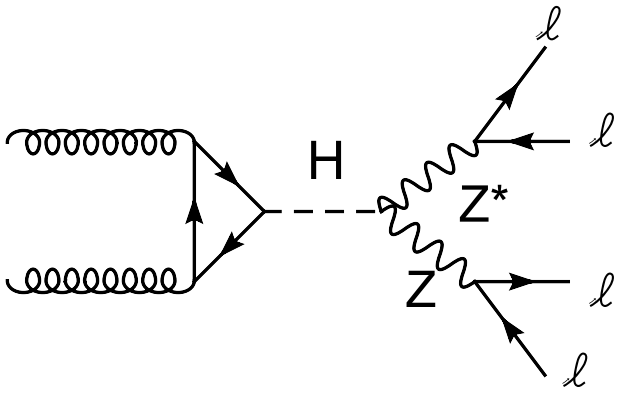
$$\vec{p}_{\text{Higgs}} = \vec{p}_{\text{before}} = \vec{p}_{\text{after}} = \sum_i \vec{p}_i$$

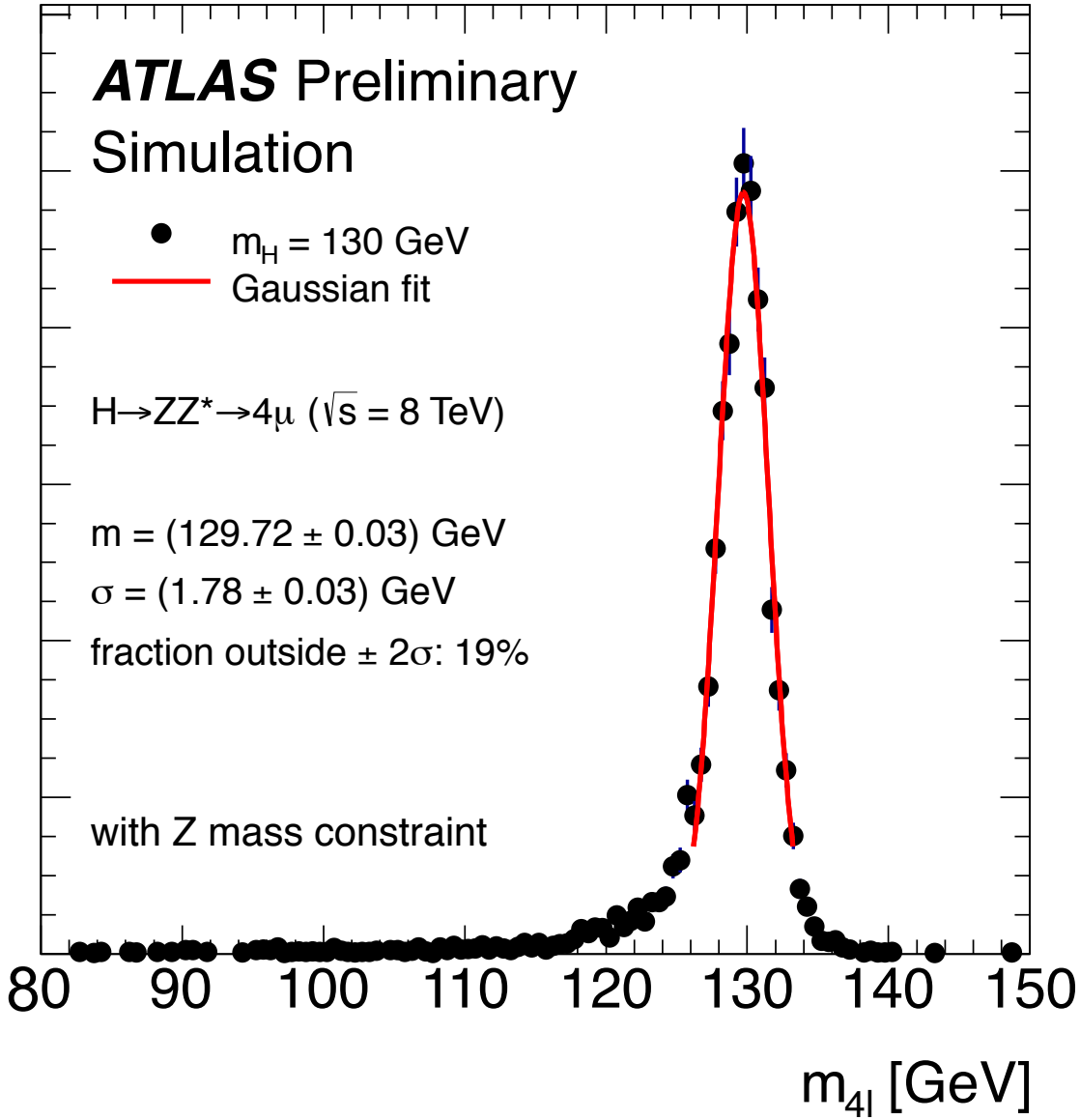
Thus, we can estimate the mass of the Higgs with

$$m_H = \sqrt{E^2_{\text{after}}/c^4 - |\vec{p}_{\text{after}}|^2/c^2}$$

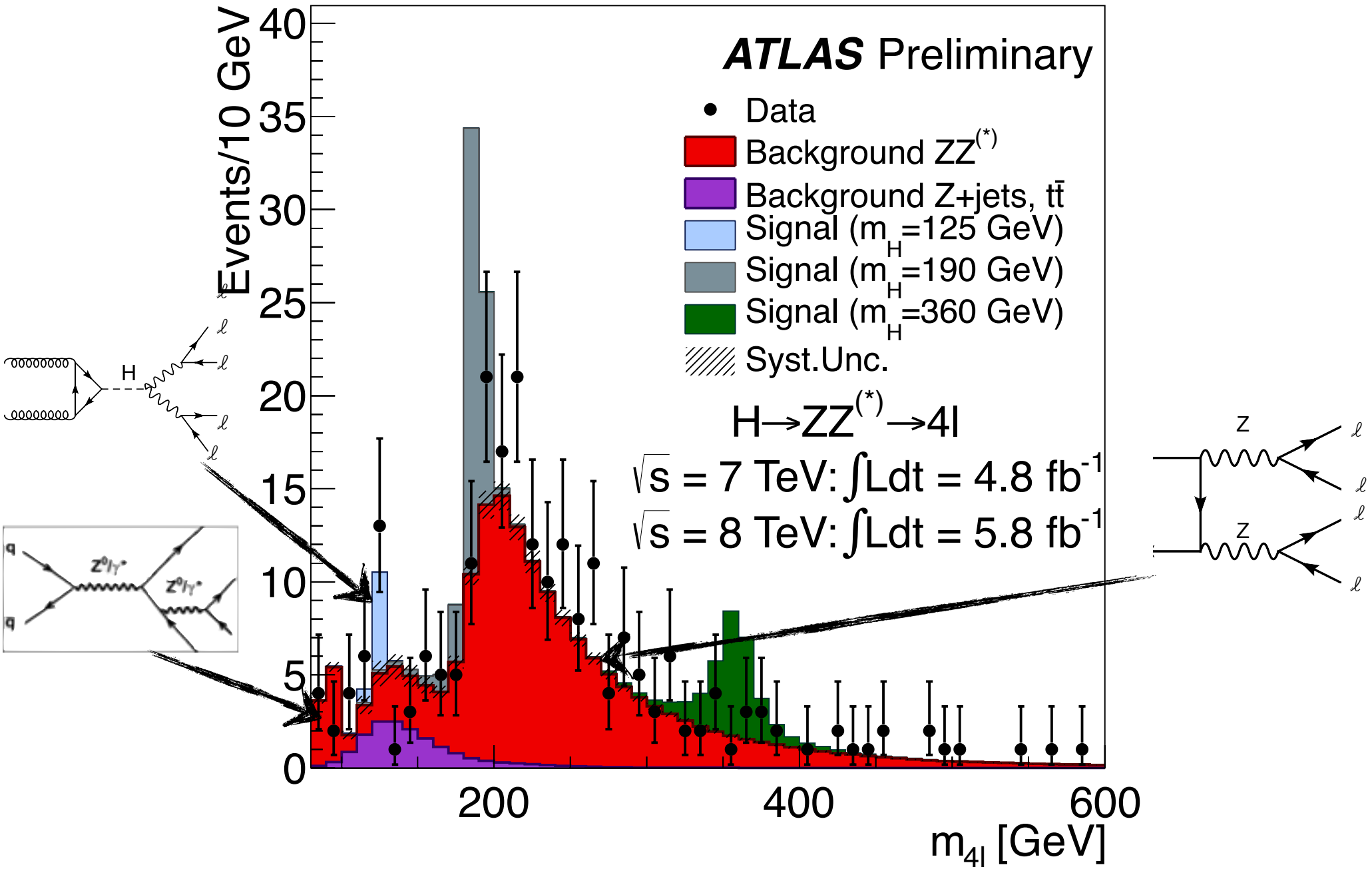
A simulated sample of signal events

The 4-lepton channel is called the 'golden' channel because it has very little background and a very sharp peak

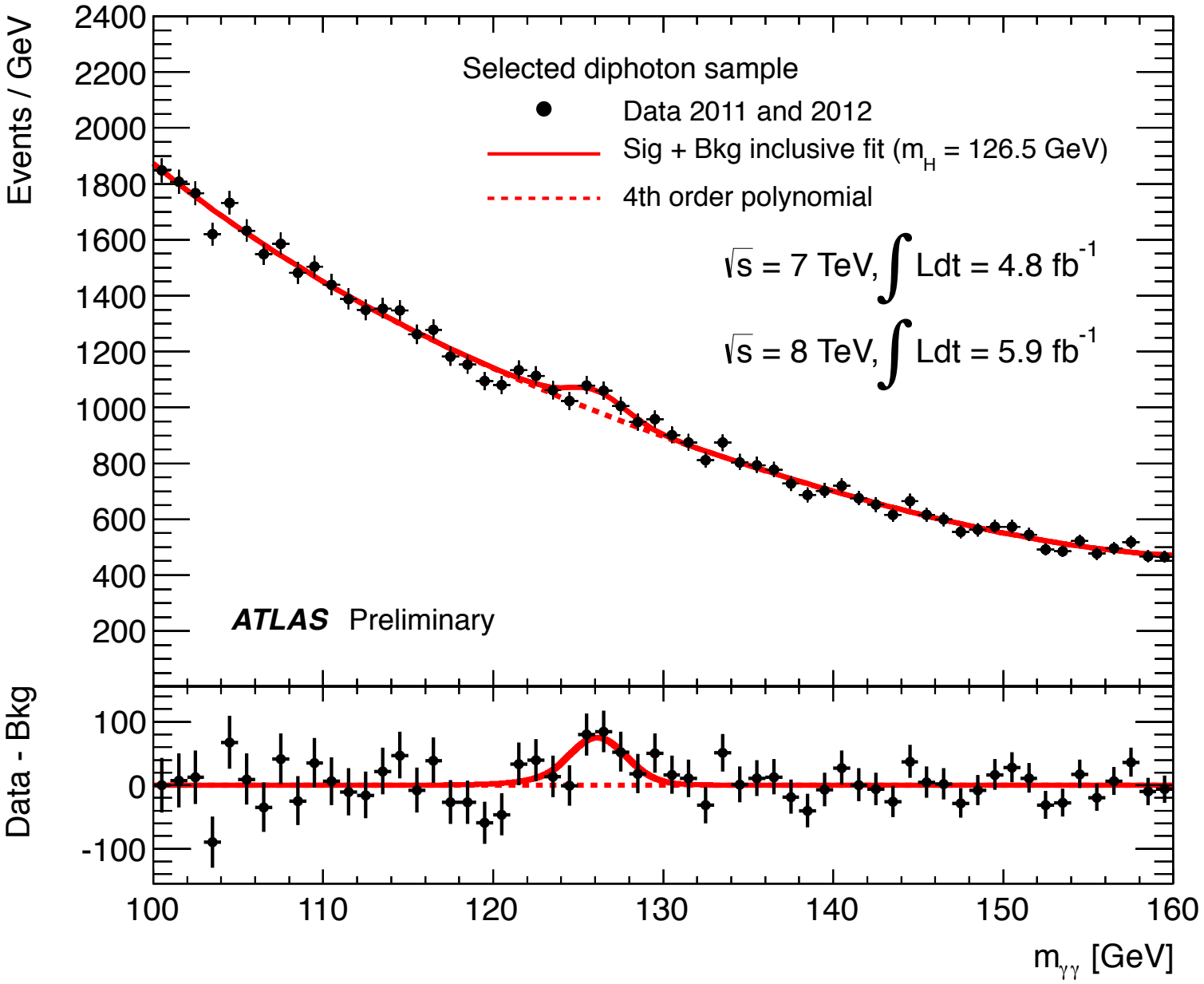




The observation in the 4l channel



The observation in the 2 photon channel

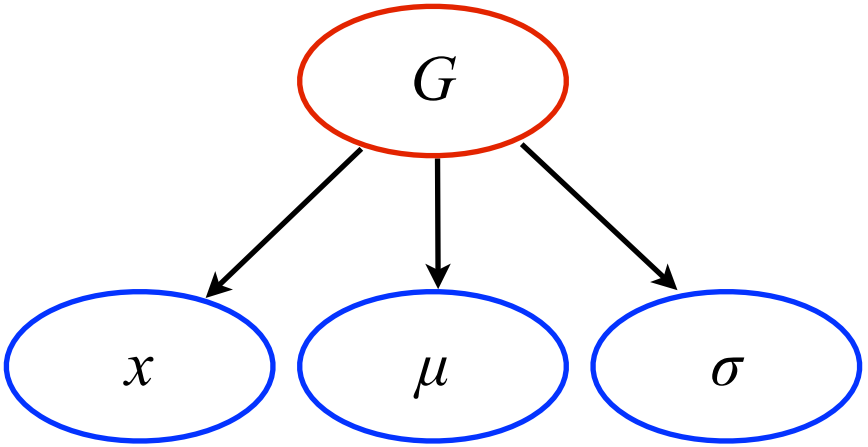


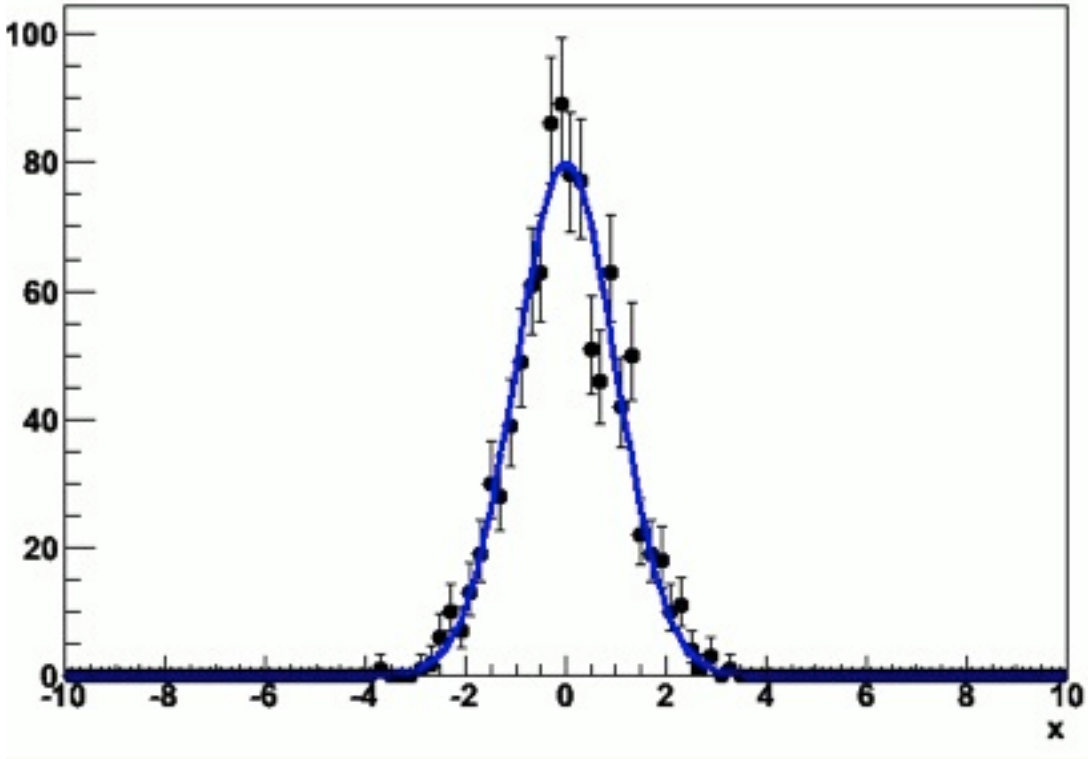
Modeling

Visualizing probability models

I will represent PDFs graphically as below (directed acyclic graph)

- eg. a Gaussian $G(x|\mu,\sigma)$ is parametrized by (μ,σ)
- every node is a real-valued function of the nodes below





Clearly related to Graphical Models, but not the focus here.

Marked Poisson Process

Channel: a subset of the data defined by some selection requirements.

- eg. all events with 4 electrons with energy > 10 GeV
- n: number of events observed in the channel
- ν: number of events expected in the channel

Discriminating variable: a property of those events that can be measured and which helps discriminate the signal from background

- eg. the invariant mass of two particles
- $f(x)$: the p.d.f. of the discriminating variable x

$$\mathcal{D} = \{x_1, \ldots, x_n\}$$

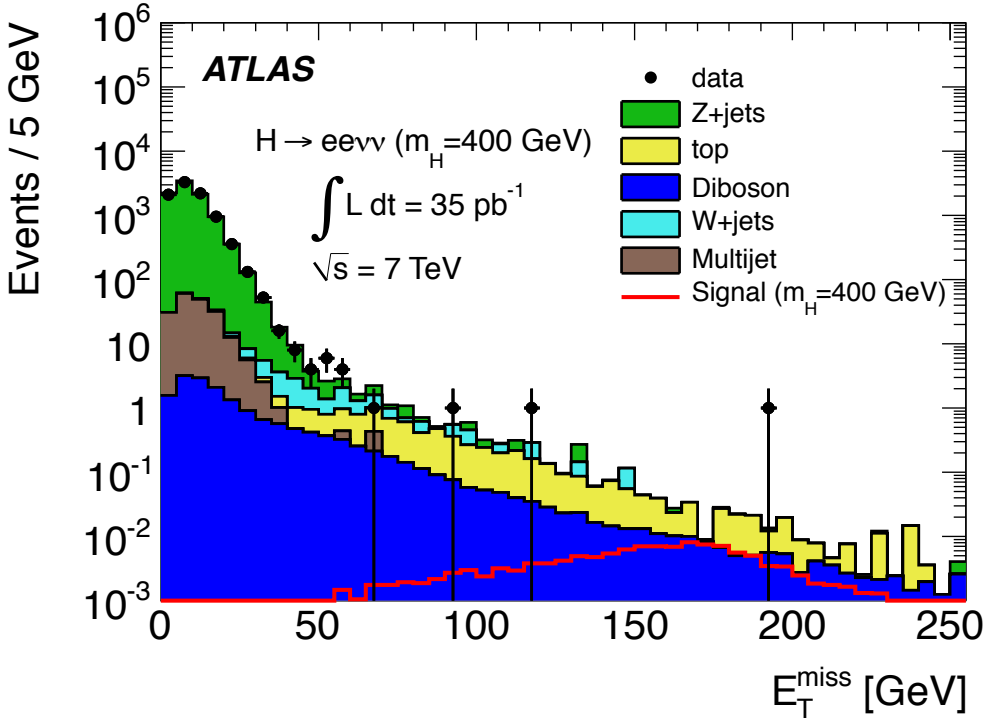
Marked Poisson Process:

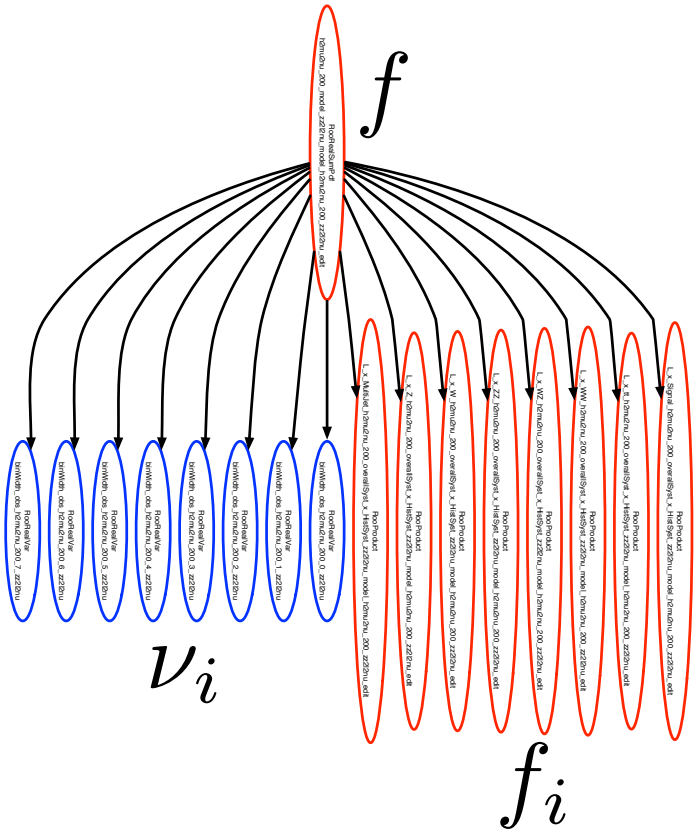
$$\mathbf{f}(\mathcal{D}|\nu) = \mathrm{Pois}(n|\nu) \prod_{e=1}^{n} f(x_e)$$

Mixture model

Total distribution is a mixture model with components corresponding to various signal and background interactions

$$f(x) = \frac{1}{\nu} \sum_{i \in \text{interactions}} \nu_i \, f_i(x) \,, \quad \nu = \sum_{i \in \text{interactions}} \nu_k$$





Parametrizing the model $\boldsymbol{\alpha} = (\mu, \boldsymbol{\theta})$

Parameters of interest (*μ*): parameters of the theory that modify the rates and shapes of the distributions, eg.

- the mass of a hypothesized particle
- the "signal strength" μ=0 no signal, μ=1 predicted signal rate

Nuisance parameters (*θ* or α_p): associated to uncertainty in:

- response of the detector (calibration)
- phenomenological model of interaction in non-perturbative regime

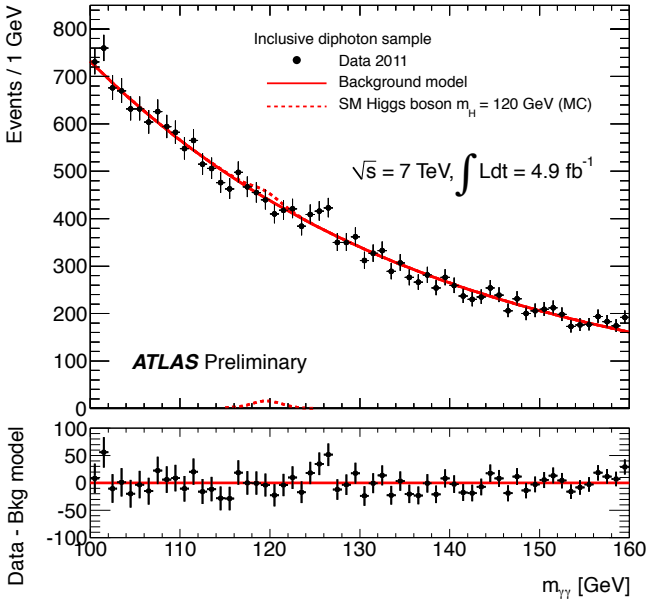
Lead to a parametrized model: $\nu \to \nu(\boldsymbol{\alpha}), f(x) \to f(x|\boldsymbol{\alpha})$

$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\boldsymbol{\alpha}) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Parametric vs. non-parametric

The densities $f_i(x|\boldsymbol{\alpha})$ for signal and background interactions are typically estimated either parametrically or non-parametrically

- in both cases simulated events are used for validation



Parametric modeling:

- simple exponential or polynomial
- parameters have no obvious interpretation

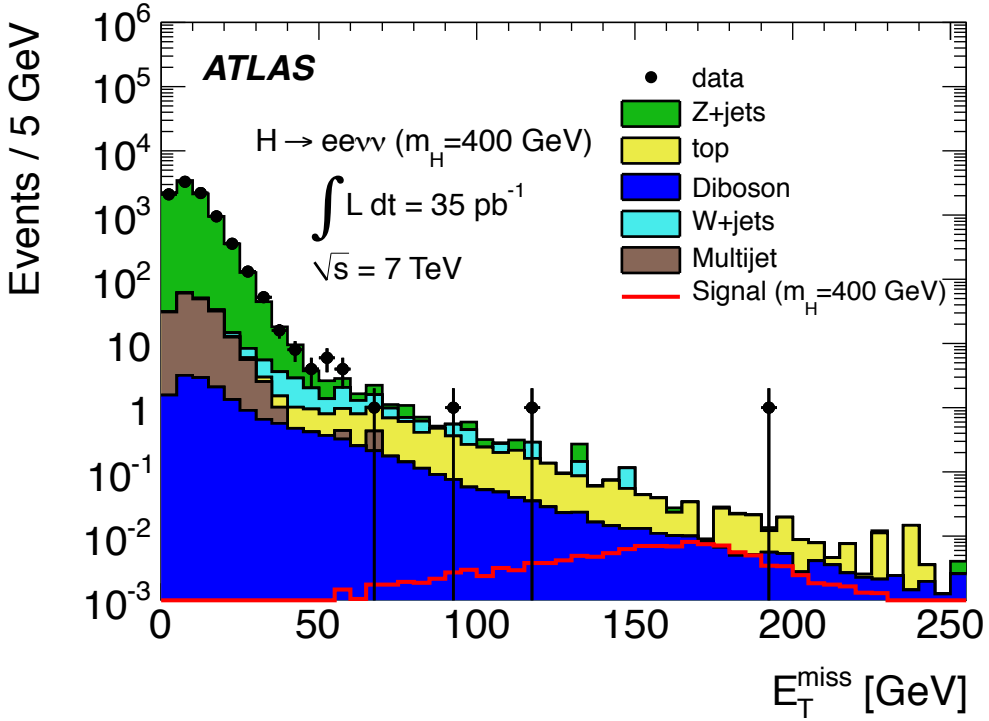
Non-Parametric modeling:

- densities estimated with histograms or kernel estimation based on simulated samples with specific settings $\{\{x_{ij}\}, \boldsymbol{\alpha}_j\}$
- parameters have clear physical interpretation
- difficulty is forming a parametrized model in many dimensions
- see contribution from Radford Neal, PhyStat 2007 [limited to likelihood]

Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p



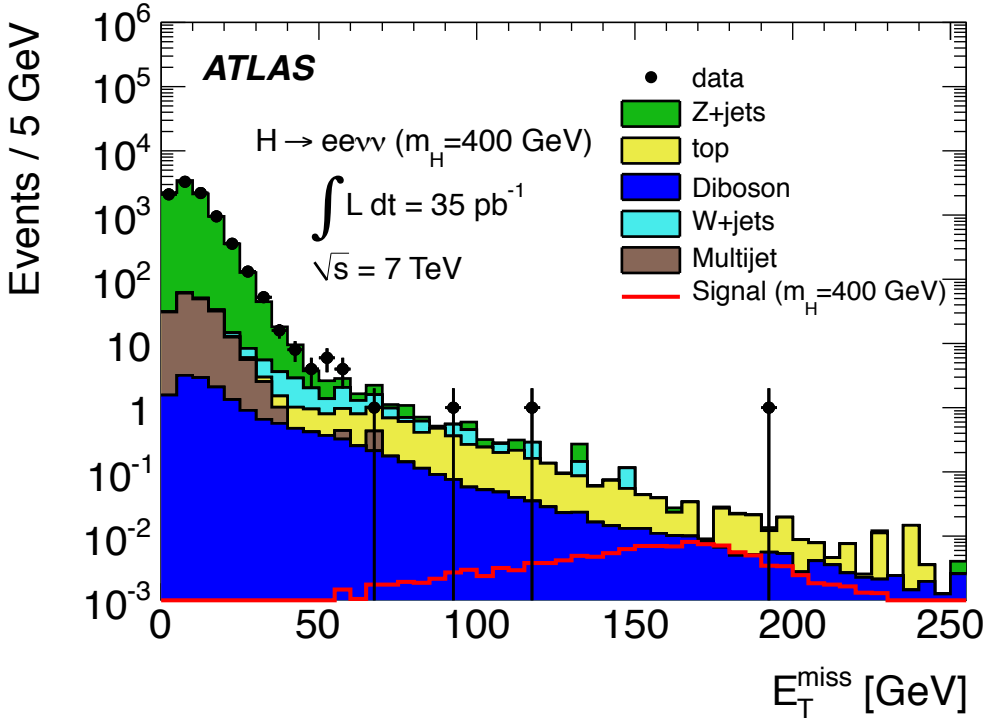
	Z+jets	top	Diboson	...
syst 1				
syst 2				
...				

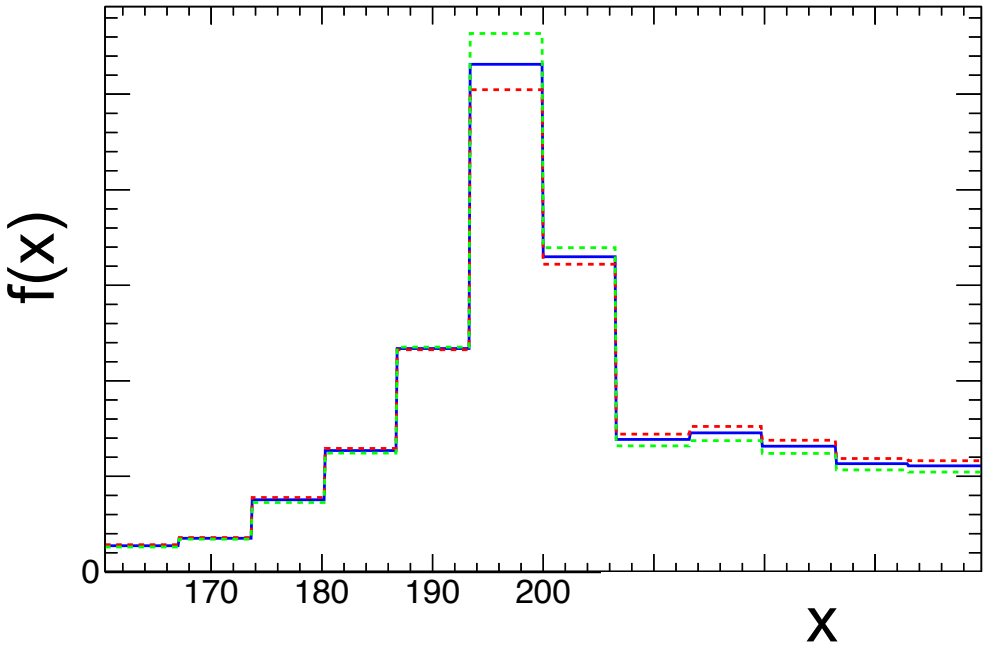
$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\boldsymbol{\alpha}) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p



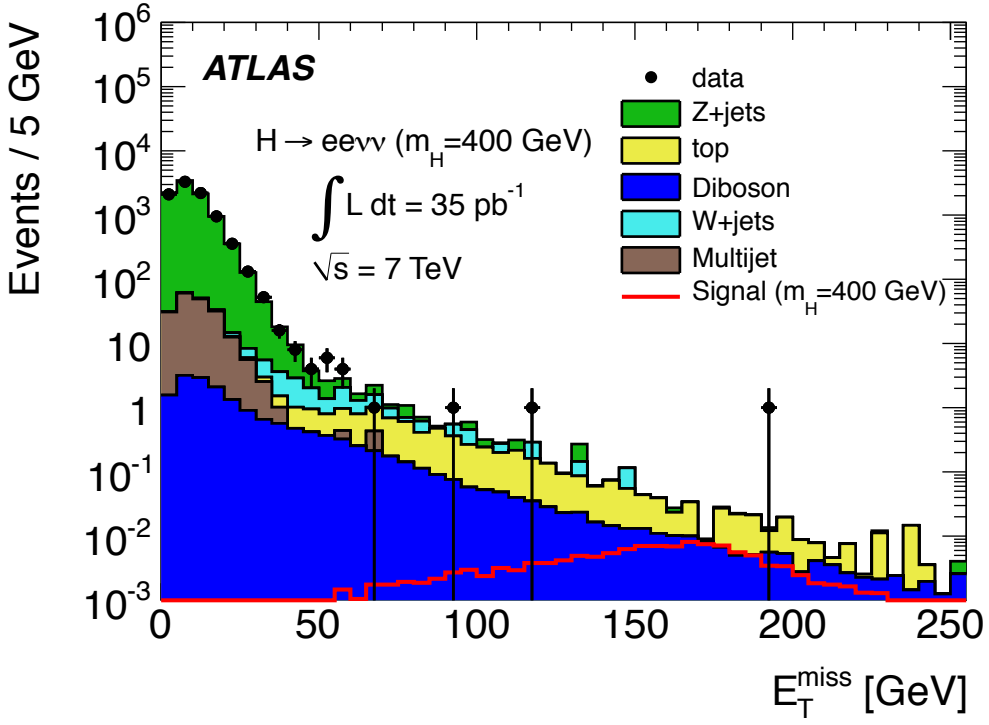


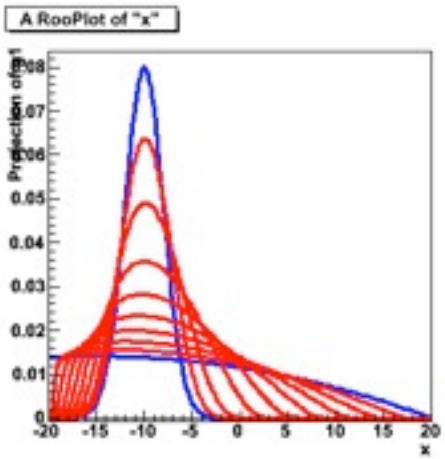
$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\boldsymbol{\alpha}) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

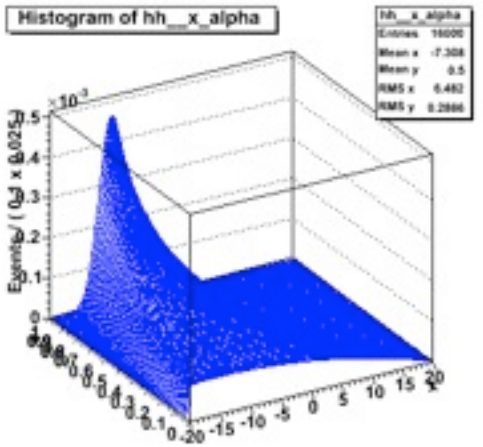
Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

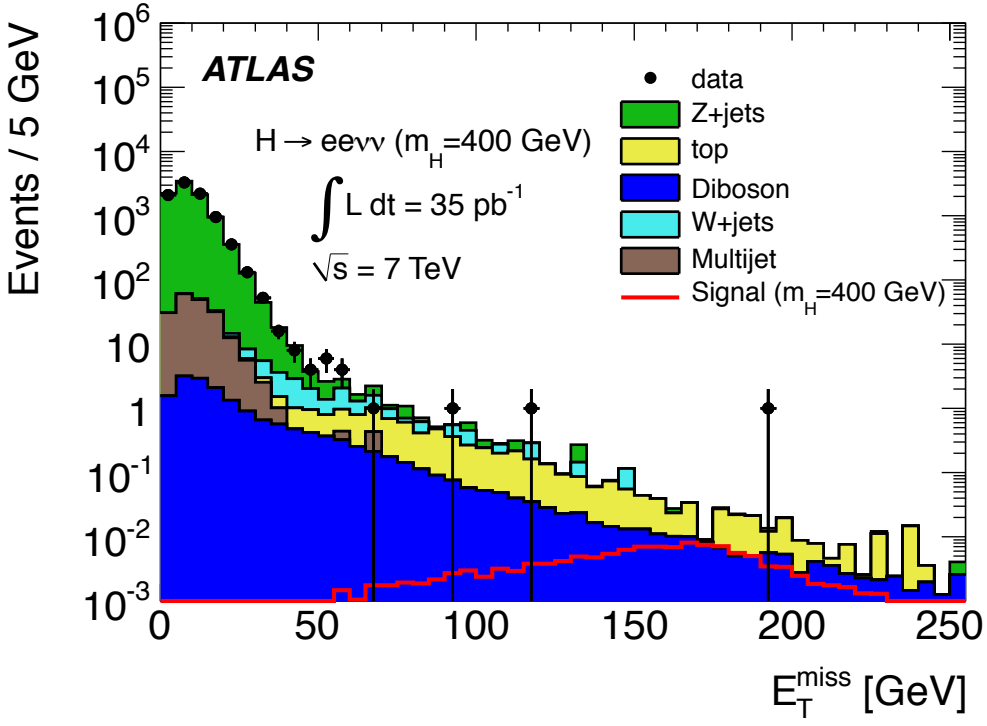


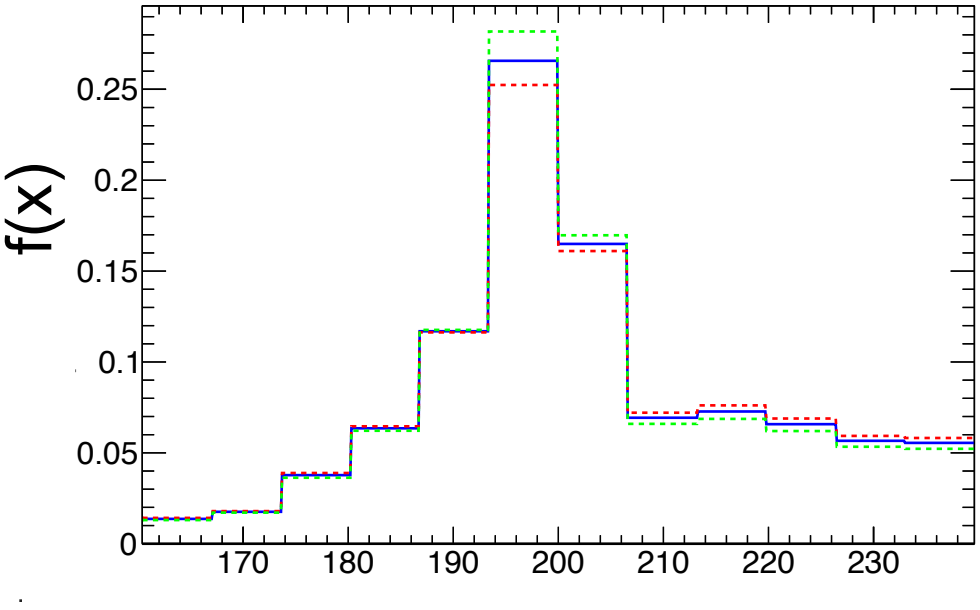


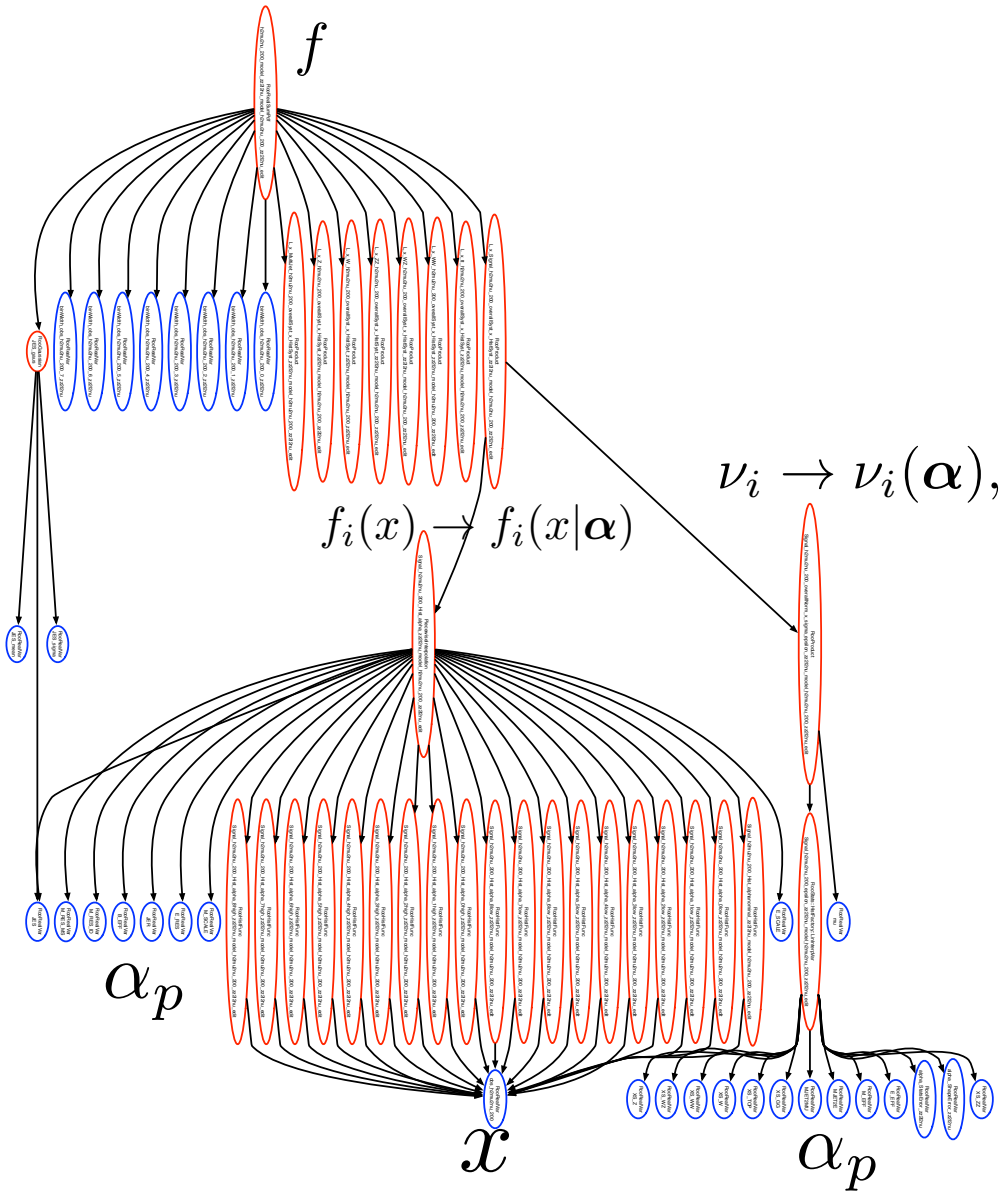


$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\boldsymbol{\alpha}) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Visualizing the model for one channel

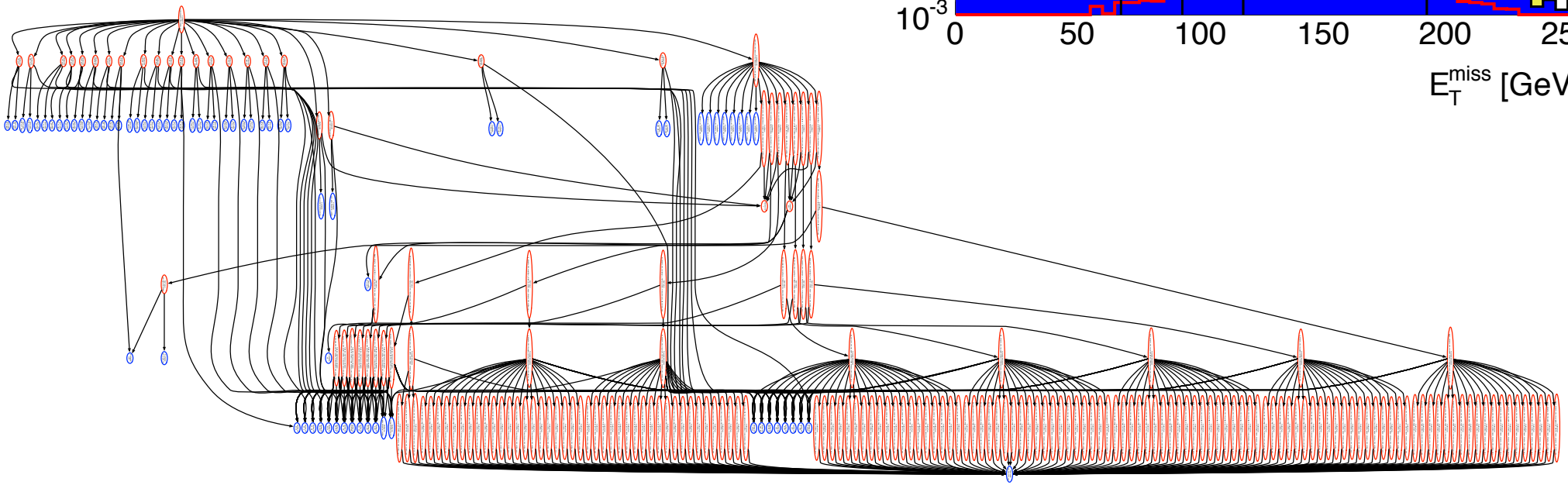






Visualizing the model for one channel

After parametrizing each component of the mixture model, the pdf for a single channel might look like this



Simultaneous multi-channel model

Simultaneous Multi-Channel Model: Several disjoint regions of the data are modeled simultaneously. Identification of common parameters across many channels requires coordination between groups such that meaning of the parameters are really the same.

$$\mathbf{f}_{\text{sim}}(\mathcal{D}_{\text{sim}}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha}) \right]$$

where $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\max}}\}$

Control Regions: Some channels are not populated by signal processes, but are used to constrain the nuisance parameters

- convert 'bad' systematics into 'good' systematics
- Prototypical Example: "on/off" problem with unknown ν_b

$$\mathbf{f}(n, m|\mu, \nu_b) = \underbrace{\text{Pois}(n|\mu + \nu_b)}_{\text{signal region}} \cdot \underbrace{\text{Pois}(m|\tau\nu_b)}_{\text{control region}}$$

Constraint terms

Often detailed statistical model for auxiliary measurements that measure certain nuisance parameters are not available.

- one typically has MLE for α_p, denoted a_p and standard error

Constraint Terms: are idealized densities for the MLE.

$$f_p(a_p|\alpha_p) \quad \text{for} \quad p \in \mathbb{S}$$

- common choices are Gaussian, Poisson, and log-normal
- New: careful to write constraint term a frequentist way.
- Previously: $\pi(\alpha_p|a_p) = f_p(a_p|\alpha_p)\eta(\alpha_p)$ with uniform η

Simultaneous Multi-Channel Model with constraints:

$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right] \cdot \prod_{p\in\mathbb{S}} f_p(a_p|\alpha_p)$$

where $\mathcal{D}_{\text{sim}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_{c_{\max}}\}$, $\mathcal{G} = \{a_p\}$ for $p \in \mathbb{S}$

Combined ATLAS Higgs Search

State of the art: Our most recent combined Higgs search includes 100 disjoint channels and >100 nuisance parameters

- Models for individual channels come from about 11 sub-groups performing dedicated searches for specific Higgs decay modes
- In addition low-level performance groups provide tools for evaluating systematic effects and corresponding constraint terms

Higgs Decay	Subsequent Decay	Additional Sub-Channels	m_H Range	L [fb^{-1}]
$H \to \gamma\gamma$	–	9 sub-channels ($p_{T_t} \otimes \eta_\gamma \otimes$ conversion)	110-150	4.9
$H \to ZZ$	$\ell\ell\ell'\ell'$	$\{4e, 2e2\mu, 2\mu2e, 4\mu\}$	110-600	4.8
	$\ell\ell\nu\nu$	$\{ee, \mu\mu\} \otimes$ {low pile-up, high pile-up}	200-280-600	4.7
	$\ell\ell qq$	$\{b$-tagged, untagged$\}$	200-300-600	4.7
$H \to WW$	$\ell\nu\ell\nu$	$\{ee, e\mu, \mu\mu\} \otimes$ {0-jet, 1-jet, VBF}	110-300-600	4.7
	$\ell\nu q\overline{q}'$	$\{e, \mu\} \otimes$ {0-jet, 1-jet}	300-600	4.7
$H \to \tau^+\tau^-$	$\ell\ell 4\nu$	$\{e\mu\} \otimes$ {0-jet} $\oplus$ {1-jet, VBF, VH}	110-150	4.7
	$\ell\tau_{\text{had}}3\nu$	$\{e, \mu\} \otimes$ {0-jet} $\otimes \{E_T^{\text{miss}} \gtrless 20$ GeV$\}$ $\oplus \{e, \mu\} \otimes$ {1-jet, VBF}	110-150	4.7
	$\tau_{\text{had}}\tau_{\text{had}}2\nu$	{1-jet}	110-150	4.7
$VH \to b\overline{b}$	$Z \to \nu\overline{\nu}$	$E_T^{\text{miss}} \in \{120-160, 160-200, \geq 200$ GeV$\}$	110-130	4.6
	$W \to \ell\nu$	$p_T^W \in \{<50, 50-100, 100-200, \geq 200$ GeV$\}$	110-130	4.7
	$Z \to \ell\ell$	$p_T^Z \in \{<50, 50-100, 100-200, \geq 200$ GeV$\}$	110-130	4.7

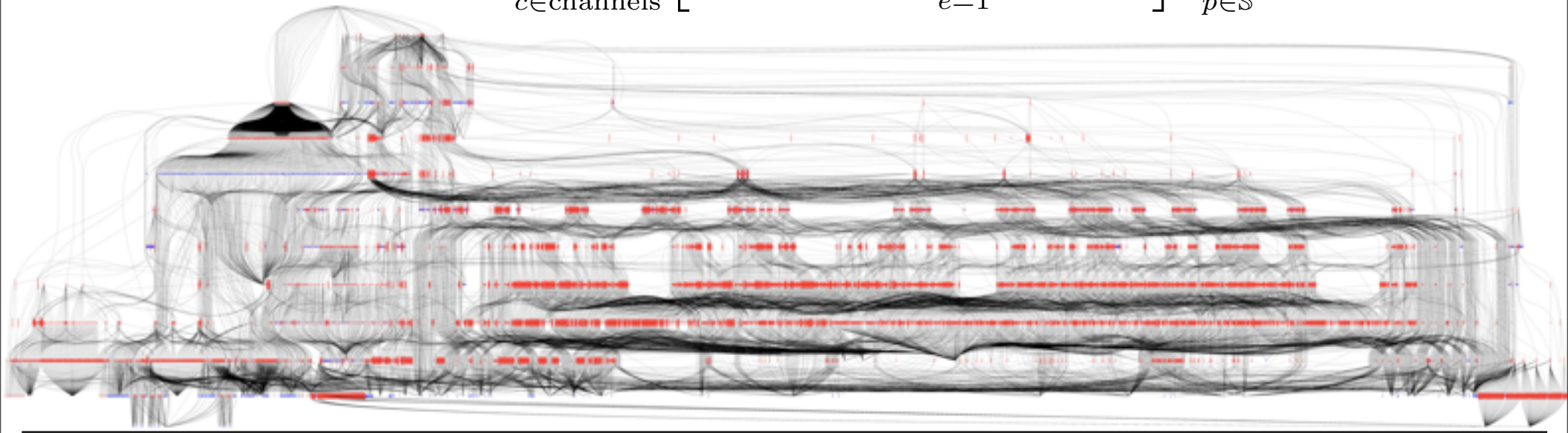
Visualizing the combined model

State of the art: Our most recent combined Higgs search includes 100 disjoint channels and >100 nuisance parameters

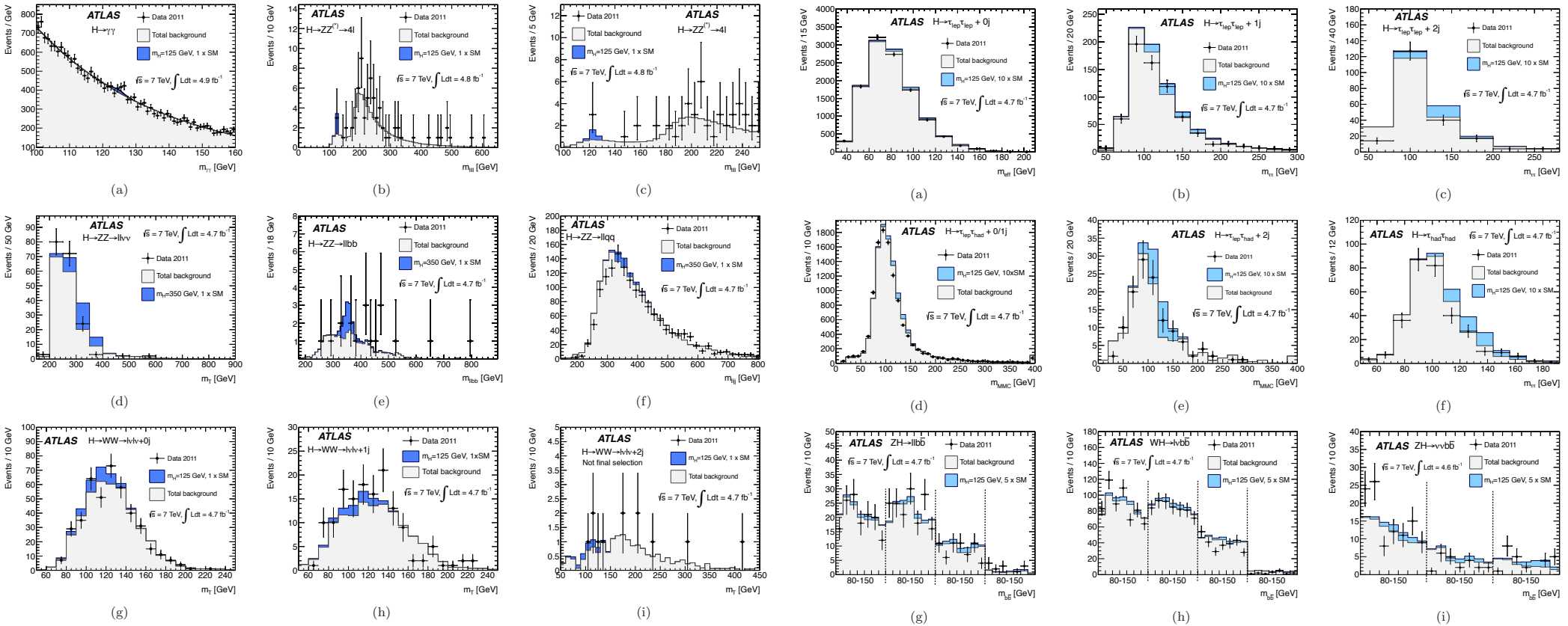
RooFit / RooStats: is the modeling language (C++) which provides technologies for collaborative modeling

- provides technology to publish likelihood functions digitally
- and more, it's the full model so we can also generate pseudo-data

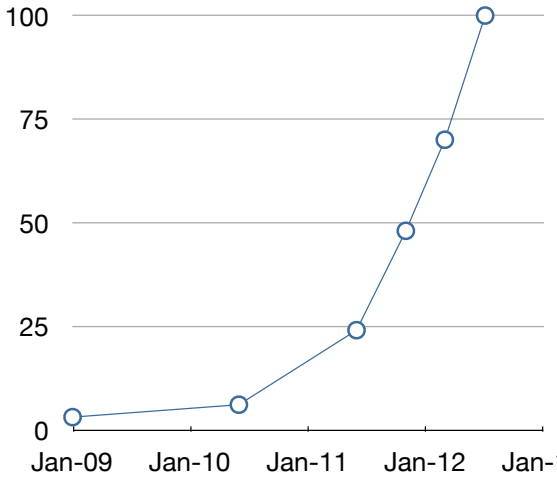
$$\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\boldsymbol{\alpha}) = \prod_{c\in\text{channels}} \left[\text{Pois}(n_c|\nu_c(\boldsymbol{\alpha})) \prod_{e=1}^{n_c} f_c(x_{ce}|\boldsymbol{\alpha})\right] \cdot \prod_{p\in\mathbb{S}} f_p(a_p|\alpha_p)$$



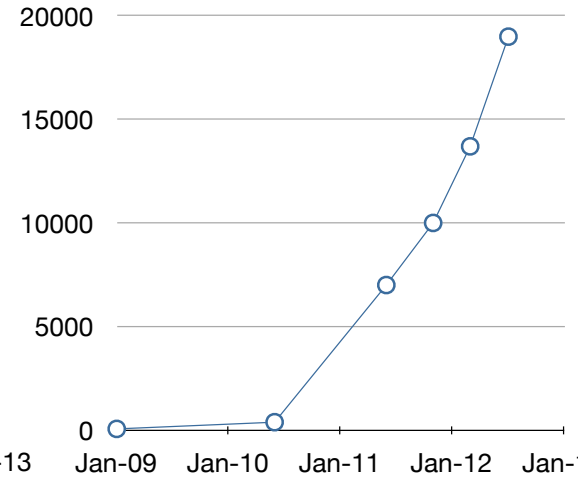
Evolution of Model Complexity



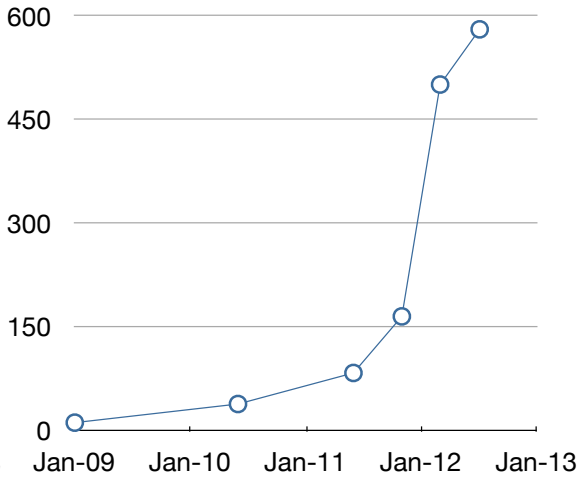
Number of Datasets Combined



Number of Model Components



Number of Parameters in Likelihood



Statistical tests

Types of statistical tests

Remember, we haven't discovered a new particle since 1995 and we have been searching for the Higgs boson for many decades

- particle physics takes a very conservative approach towards discovery and exclusion

Particle physicists are also 'frequentist dinosaurs' in that we worry a lot about coverage of our confidence intervals etc.

- We also use Bayesian techniques, but there is a general dislike for assigning prior probabilities to theoretical parameters like the mass of the Higgs boson or priors on models (ie. prior probability that the Higgs boson exists at all)

With the full model $\mathbf{f}_{\text{tot}}(\mathcal{D}_{\text{sim}}, \mathcal{G}|\mu, \boldsymbol{\theta})$ we can easily make likelihood function and Bayesian posterior, converse is not true.

Hybrid resampling / profile construction

Until recently, practice was to compute p-values with a Hybrid Bayesian-Frequentist procedure in which the nuisance parameters were marginalized with respect to some prior $\pi(\boldsymbol{\theta})$

$$p = \int_{t_{\rm obs}}^{\infty} dt \left[\int d\boldsymbol{\theta} f(t|\mu, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \right]$$

Frequentist p-values should be calculated for a specific point in $\mu, \boldsymbol{\theta}$

Confidence interval for parameter of interest, consists of μ such that $P(\lambda(\mu) < c_\alpha | \mu, \boldsymbol{\theta}) < \alpha$ for all $\boldsymbol{\theta}$

- checking condition for all $\boldsymbol{\theta}$ is computationally challenging!
- approximate by considering only $\hat{\theta}(\mu)$

C. Chuang and T. L. Lai. Hybrid resampling methods for confidence intervals. *Statist. Sinica*, 10:1–50, 2000. http://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n11.pdf.

Matthew Walker Bodhisattva Sen and Michael Woodroofe. On the unified method with nuisance parameters. *Statist. Sinica*, 19:301–314., 2009. http://www3.stat.sinica.edu.tw/statistica/oldpdf/A19n116.pdf.

Choice of test statistic

An important property of the profile likelihood ratio test statistic is that under certain conditions its asymptotic distribution is known

- Wilks's theorem [Ann. Math. Statist. 9 (1938)]:

$$f(-2\ln\lambda(\mu)|\mu,\boldsymbol{\theta})$$ chi-square, independent of θ

- Wald's theorem [Trans. Amer. Math. Soc. V54, No. 3 (1943)]

$$f(-2\ln\lambda(\mu)|\mu',\boldsymbol{\theta})$$ non-central chi-square

- with non-centrality parameter given by the Fisher information matrix

$$V_{ij}^{-1} = -E\left[\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right]$$

Our parameter of interest μ bounded physically to be non-negative

- fairly easy to incorporate the effect of boundary

Perhaps a clever trick for calculating the expected Fisher information matrix

- "Asimov data" is an artificial dataset with special properties
- observed Fisher information on Asimov = expected Fisher information

The Asimov dataset

The expected Fisher information is not trivial to compute for an arbitrary model

$$\left(\mathcal{I}(\theta)\right)_{i,j} = -\,\mathrm{E}\left[\left.\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ln f(X;\theta)\right|\theta\right] .$$

In the case of a binned distribution, the expected Fisher information is identical to the observed Fisher information calculated with an artificial dataset, referred to as the Asimov data

- follows from fact that $\partial^2[\ln \mathrm{Pois}(n|\theta)]/\partial\theta^2$ is linear in *n*
- The bin contents for the Asimov data are exactly the expectation of the model (non-integer in general)

This is a very convenient numerical trick to calculate:

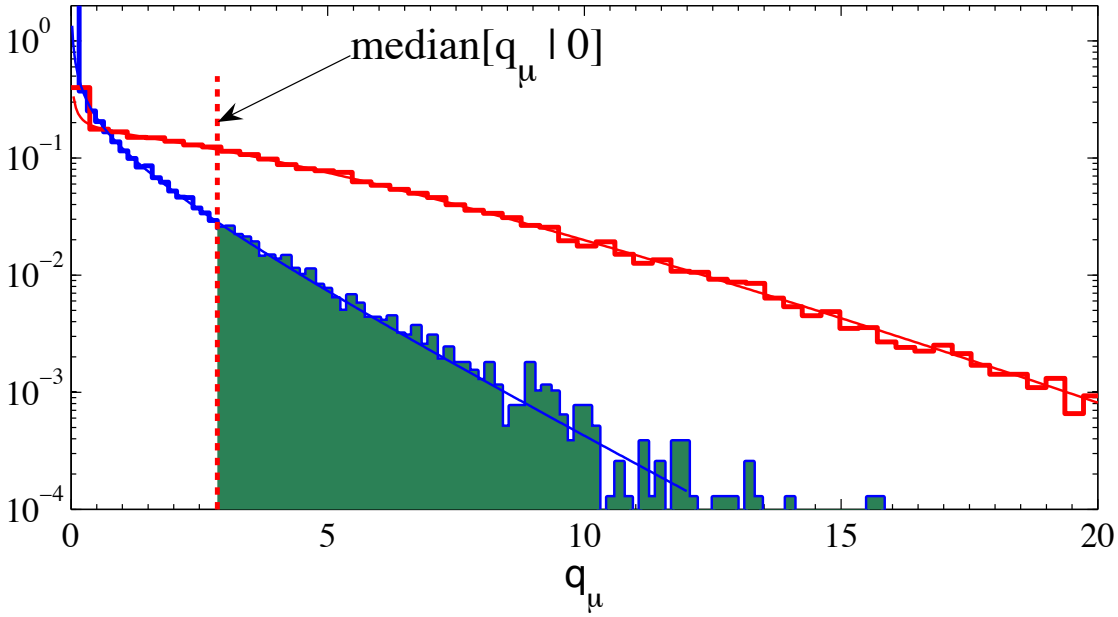
- expected Fisher information
- non-centrality parameter (Wald's theorem)
- Jeffreys's Prior (also helps with reference prior)
  - Recent progress in Reference analysis in particle physics:

L. Demortier, S. Jain, H. Prosper Phys.Rev. D82 (2010) 034002

G. Cowan, KC, E. Gross, O. Vitells
Eur.Phys.J. C71 (2011) 1554
[arXiv:1007.1727]

Comparison of asymptotic and ensembles

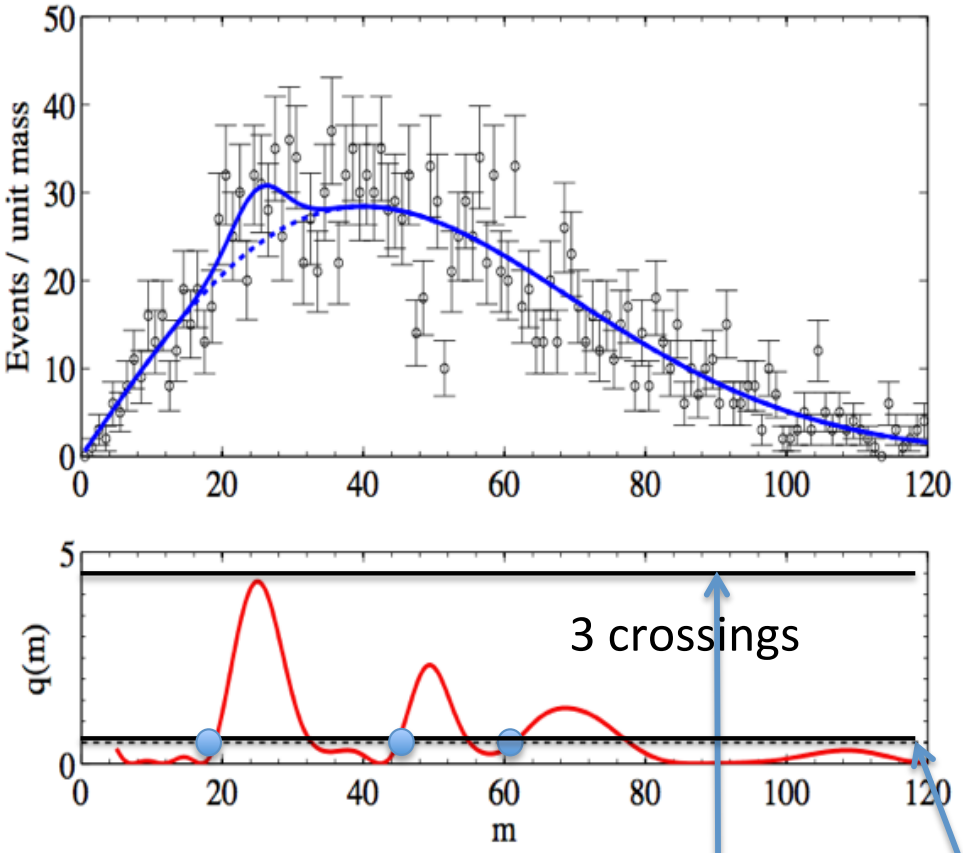
Compare asymptotic distributions to distributions obtained with large ensembles of pseudo-experiments generated with Monte Carlo techniques



This is a significant development as building this distribution from Monte Carlo approaches can take 100,000 CPU hours for Higgs search!

G. Cowan, KC, E. Gross, O. Vitells
Eur.Phys.J. C71 (2011) 1554
[arXiv:1007.1727]

Look-elsewhere effect



Typically our signal model has some location parameter, which does not affect the null.

This modifies the distribution of the likelihood ratio test statistic

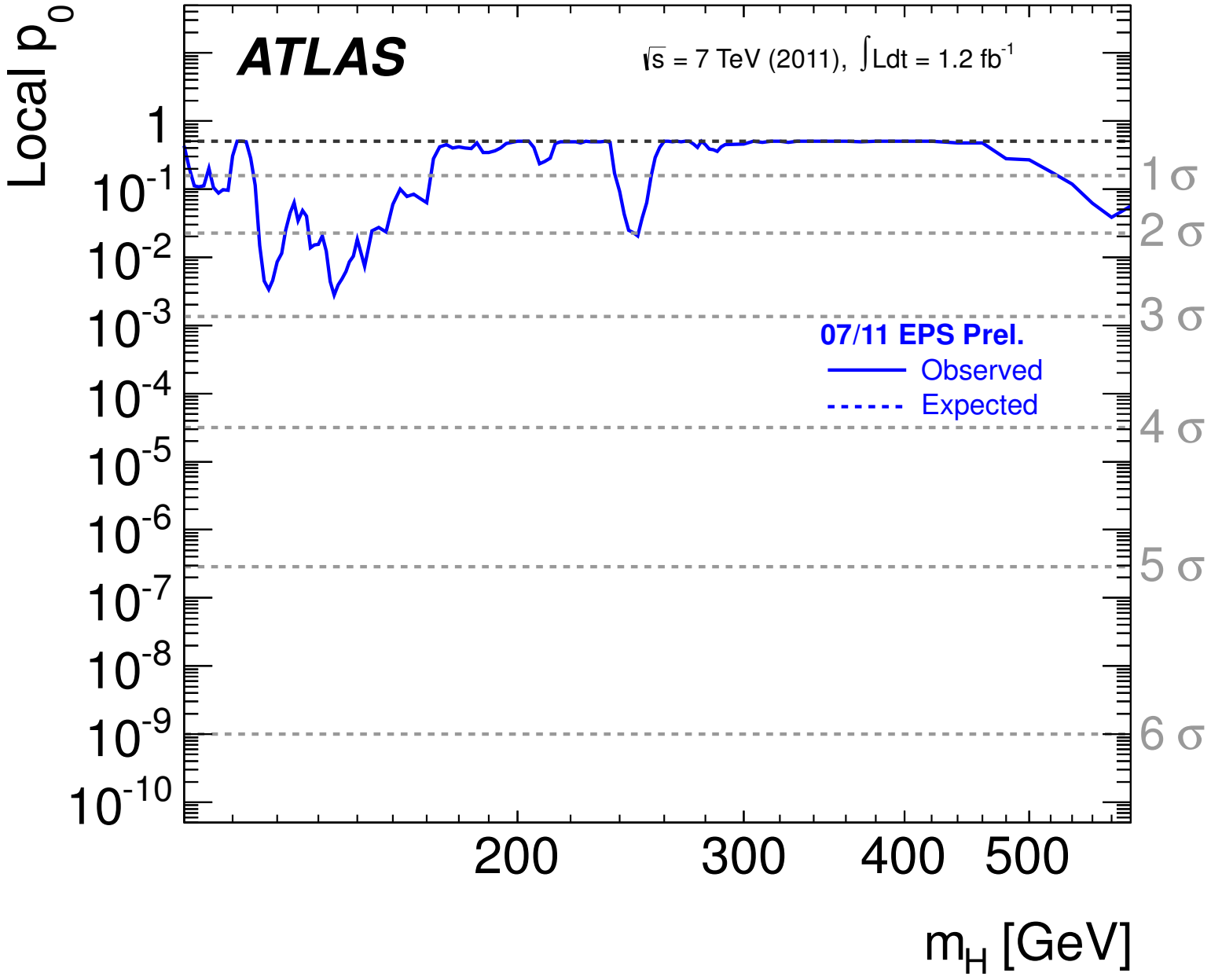
we call this the "look-elsewhere effect"

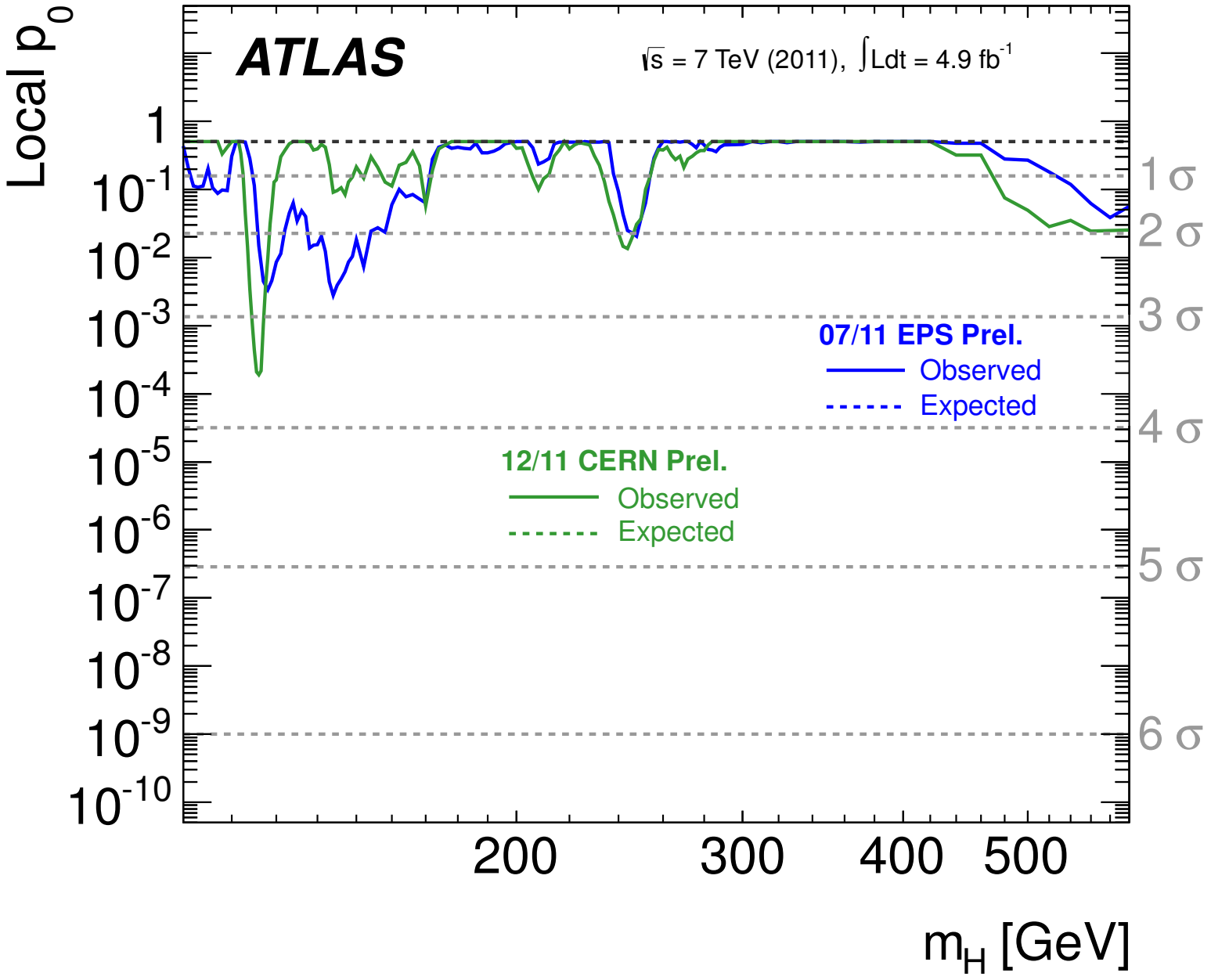
Recently Gross & Vitells found the results of Rice, Davies, and Leadbetter for a fast asymptotic approximation for the global p-value

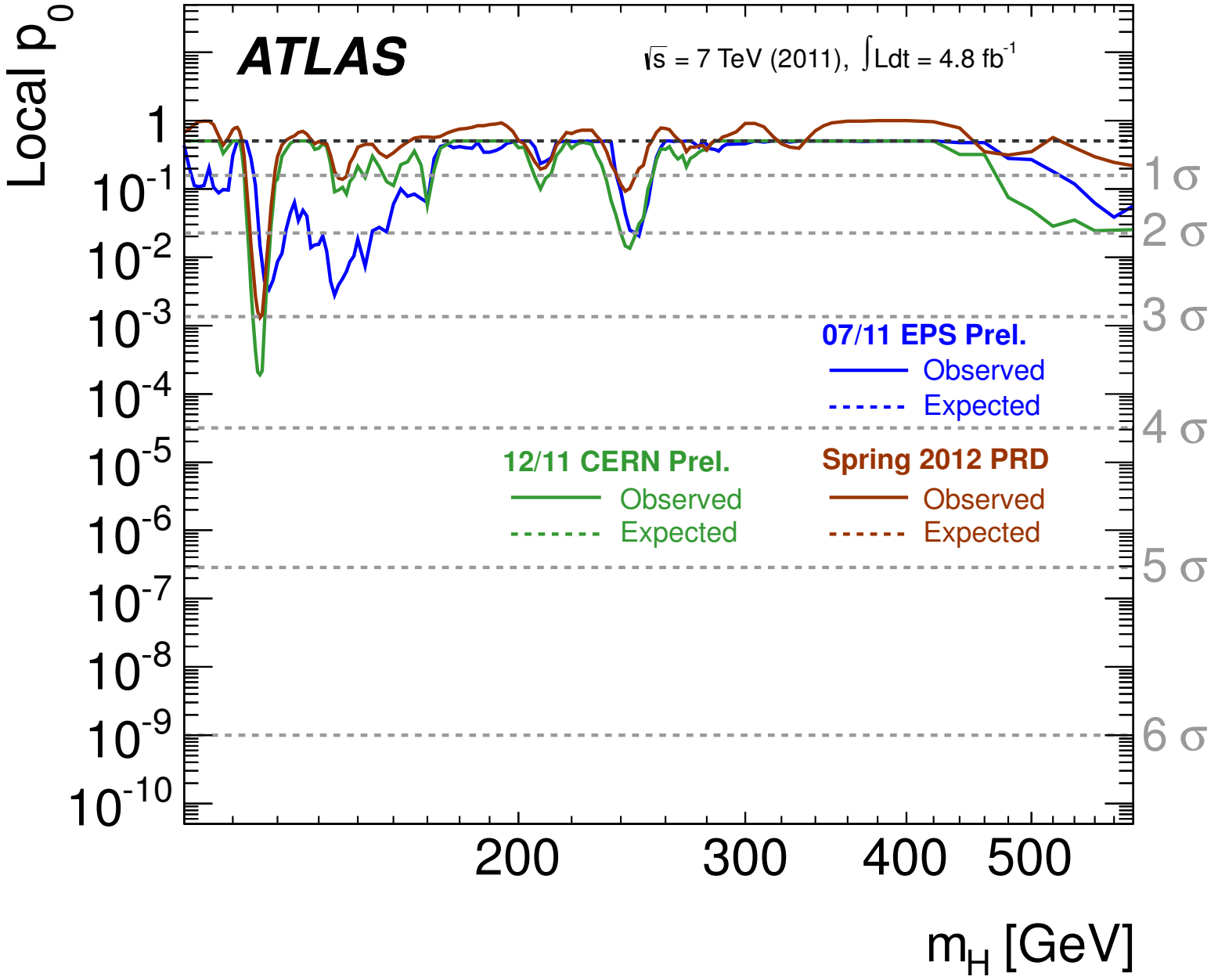
E. Gross & O. Vitells, **Eur.Phys.J. C70 (2010); Astropart.Phys. 35 (2011)**

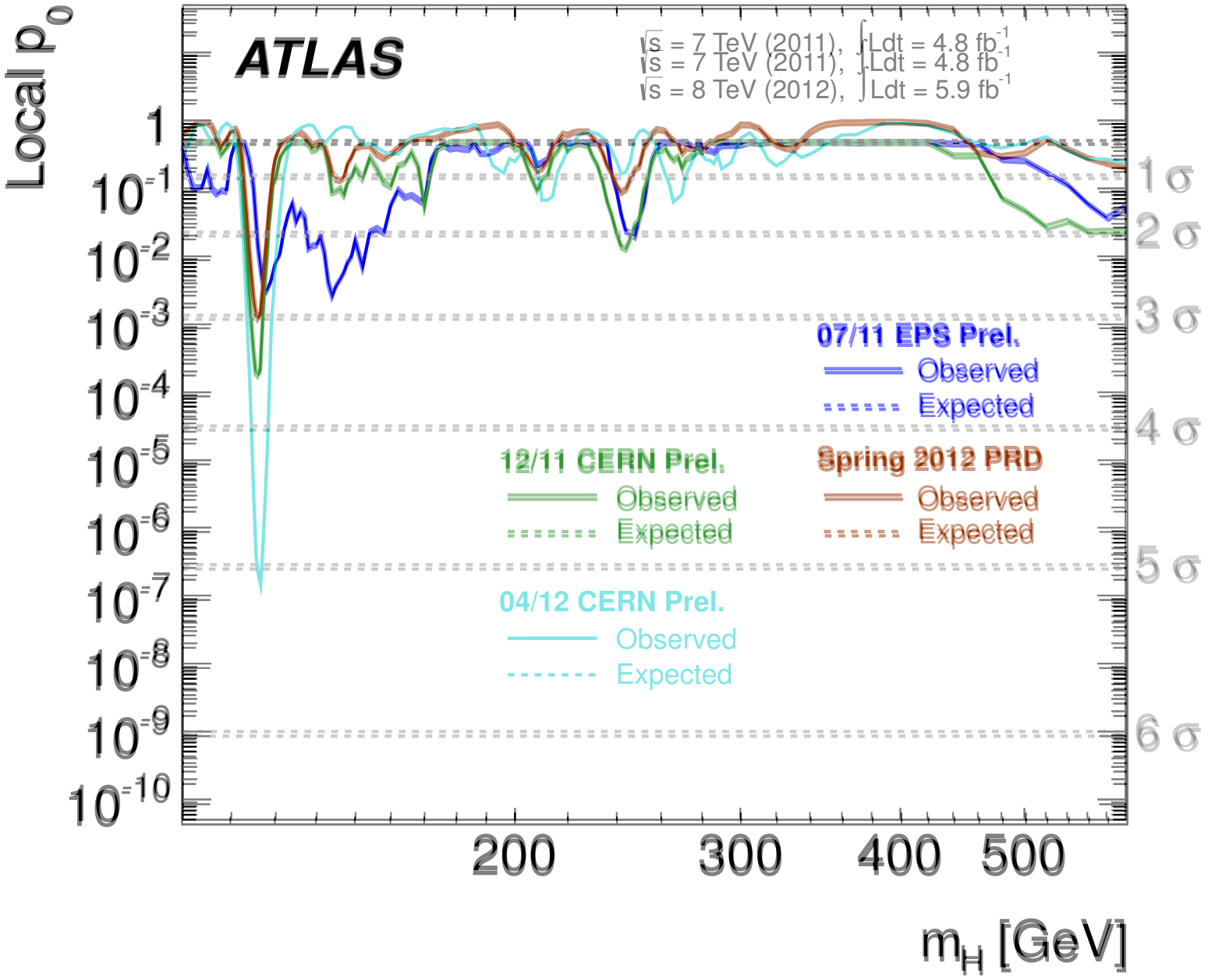
$$p_0^{global} \cong p_0^{local} + \left\langle N(q_{ref}) \right\rangle e^{-(q_{test}-q_{ref})/2}$$

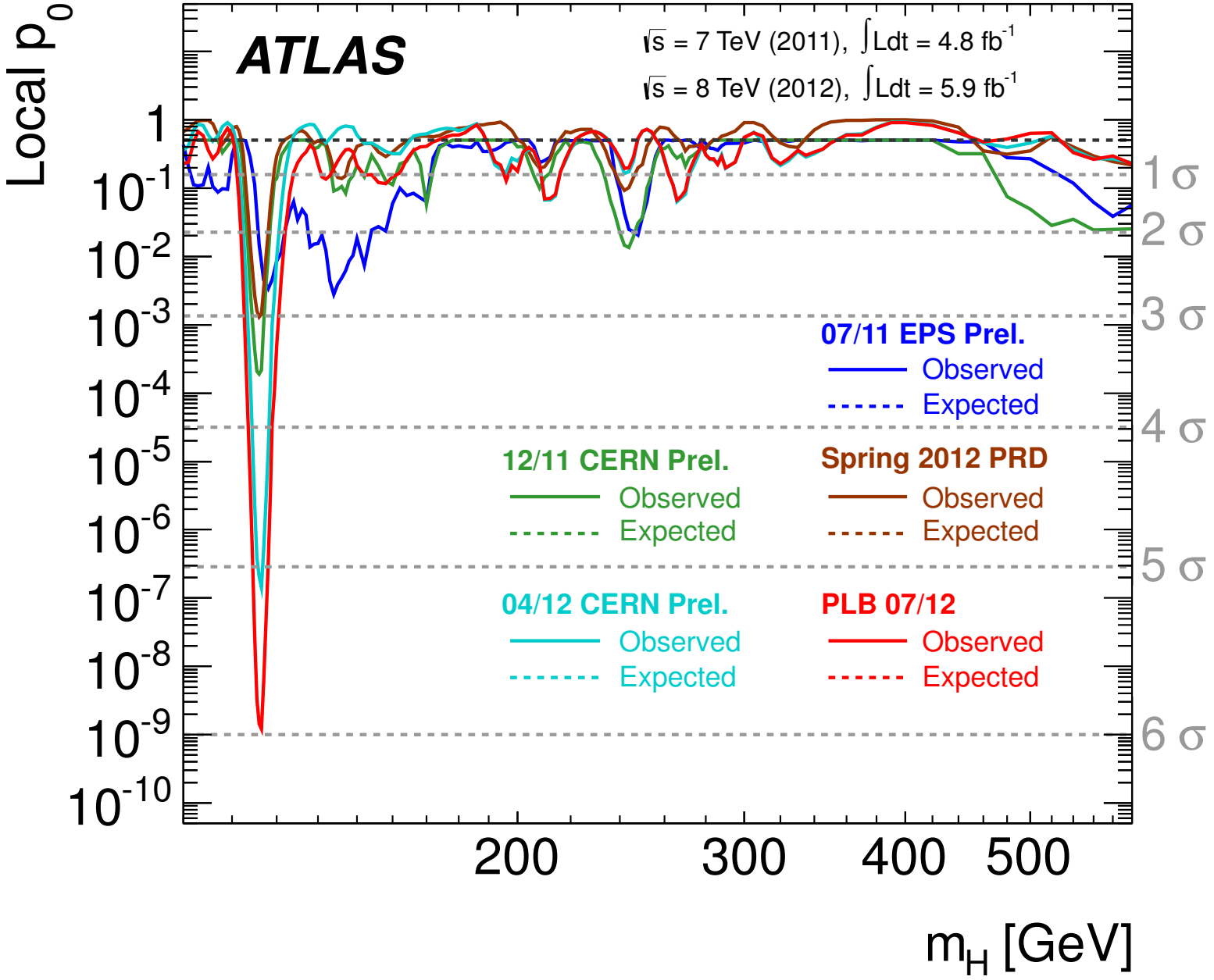
R. B. Davies,Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika 64 (1977); Biometrika 74 (1987).


ATLAS
√s = 7 TeV (2011), ∫Ldt = 1.2 fb-1
Local p0
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
07/11 EPS Prel.
Observed
Expected
200
300
400
500
mH [GeV]


ATLAS
√s = 7 TeV (2011), ∫Ldt = 4.9 fb-1
Local p0
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
07/11 EPS Prel.
Observed
Expected
12/11 CERN Prel.
Observed
Expected
200
300
400
500
mH [GeV]


ATLAS
√s = 7 TeV (2011), ∫Ldt = 4.8 fb-1
Local p0
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
07/11 EPS Prel.
Observed
Expected
12/11 CERN Prel.
Observed
Expected
Spring 2012 PRD
Observed
Expected
200
300
400
500
mH [GeV]

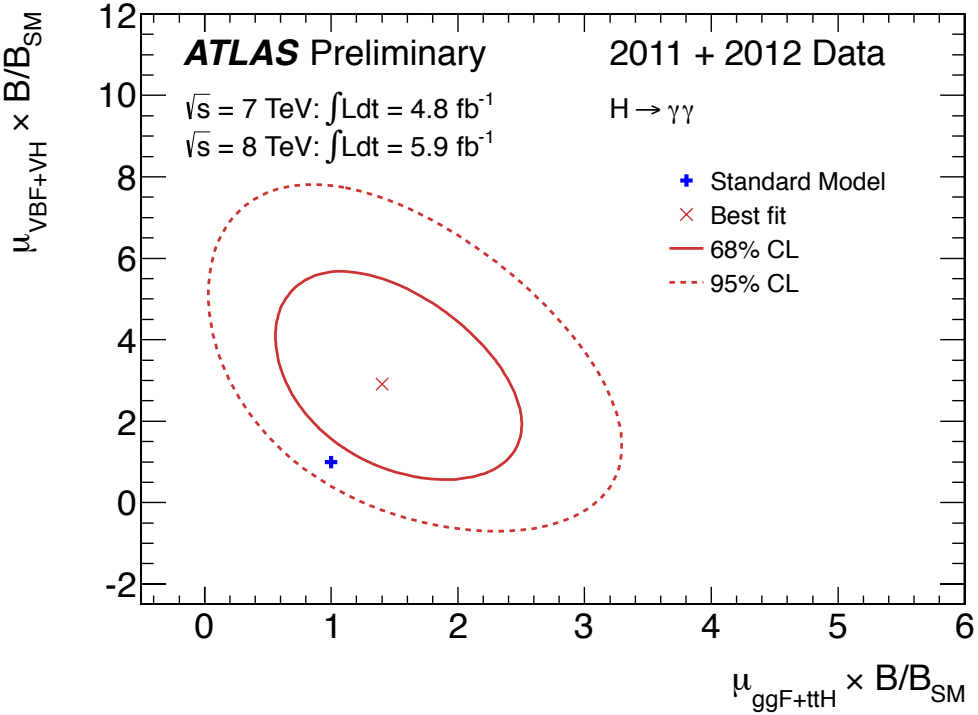

ATLAS
√s = 7 TeV (2011), ∫Ldt = 4.8 fb-1
√s = 7 TeV (2011), ∫Ldt = 4.8 fb-1
√s = 8 TeV (2012), ∫Ldt = 5.9 fb-1
Local p0
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
07/11 EPS Prel.
Observed
Expected
12/11 CERN Prel.
Observed
Expected
Spring 2012 PRD
Observed
Expected
04/12 CERN Prel.
Observed
Expected
200
300
400
500
mH [GeV]

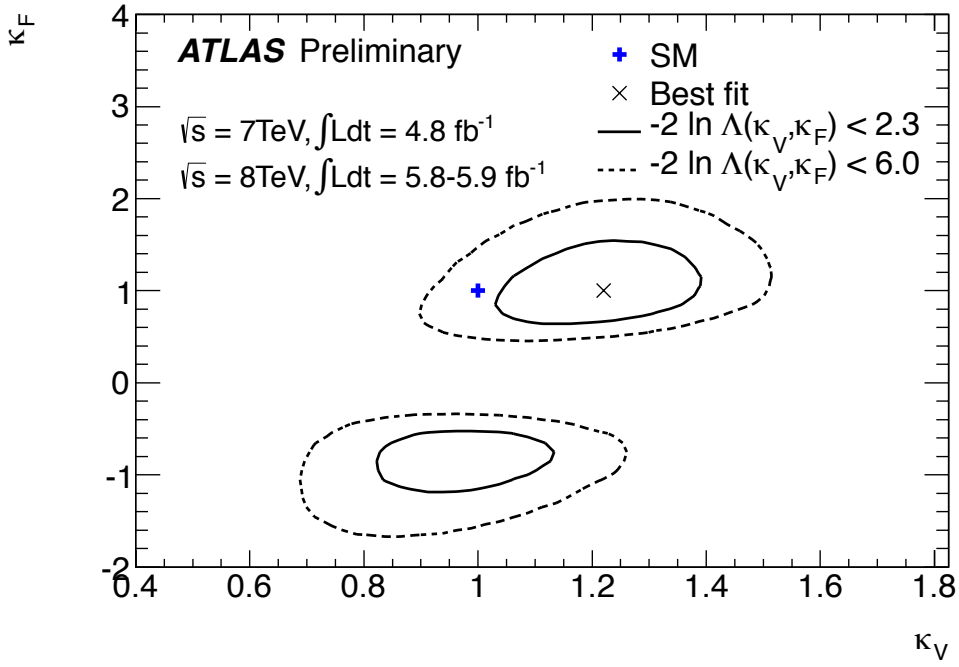

ATLAS
√s = 7 TeV (2011), ∫Ldt = 4.8 fb-1
√s = 8 TeV (2012), ∫Ldt = 5.9 fb-1
Local p0
1
10-1
10-2
10-3
10-4
10-5
10-6
10-7
10-8
10-9
10-10
1 σ
2 σ
3 σ
4 σ
5 σ
6 σ
07/11 EPS Prel.
Observed
Expected
12/11 CERN Prel.
Observed
Expected
Spring 2012 PRD
Observed
Expected
04/12 CERN Prel.
Observed
Expected
PLB 07/12
Observed
Expected
200
300
400
500
mH [GeV]

Is it the Higgs of the Standard Model

Once one measures the mass of the Higgs boson, all of it's properties are fixed in the Standard Model theory

- Just yesterday we released our first results on the properties of this new particle
- Mild tension, 3x more data by end of the year!





Review of Challenges
and
Ideas for Working Group

Challenges & possible research areas

Interpolation of distributions based on simulated samples with different parameter settings

- currently one-at-a-time variation of nuisance parameters
- need multivariate interpolation procedures
- clearly a weak point of our current modeling approach

As we improve treatment of uncertainties that are statistical in nature, attention turning towards truly systematic uncertainties

- Recently, Brad Efron's gave us a nice talk on accounting for model selection

The complexity of our statistical models is growing exponentially, starting to need new algorithms to deal with them or principles for simplifying them

- graphical models look promising

Publishing likelihood functions is technically possible, but significant sociological and psychological barriers

Challenges & possible research areas

Most discussion with statisticians has focused on hypothesis testing and confidence intervals for final results. Many interesting problems up-stream

- **exception:** machine learning for selecting candidate signal events
- **barriers:** collaborations do not openly share data, requires some semi-formal agreement

False Discovery Rate

- Number of discoveries is small, hasn't been used much
- Could be applied at the particle identification level

Importance sampling for rare events in simulation

- The simulation of our detectors is very computationally challenging and we use brute force to populate tails in cases where we might be able to do something smarter

Machine learning + computer simulations

- Most analyses either use computer simulations of the detector or ad hoc parametrized models.
- Little use of machine learning to learn the expensive computer simulation

Conclusions

Particle physics is a unique arena for scientific statistical computing

- relatively classical statistical questions in an extreme setting
- very precise theoretical predictions and lots of data

Collaborative statistical modeling is growing into a powerful scientific tool.

- we are seeing exponential increases in complexity in the models over the last year and a half
- quickly outstripping our tools for maximum likelihood and also our MCMC sampling
- a potential revolution in how we publish our results

Many areas where statisticians can contribute

- though, there are some formalities for accessing data & simulation

Thank you!