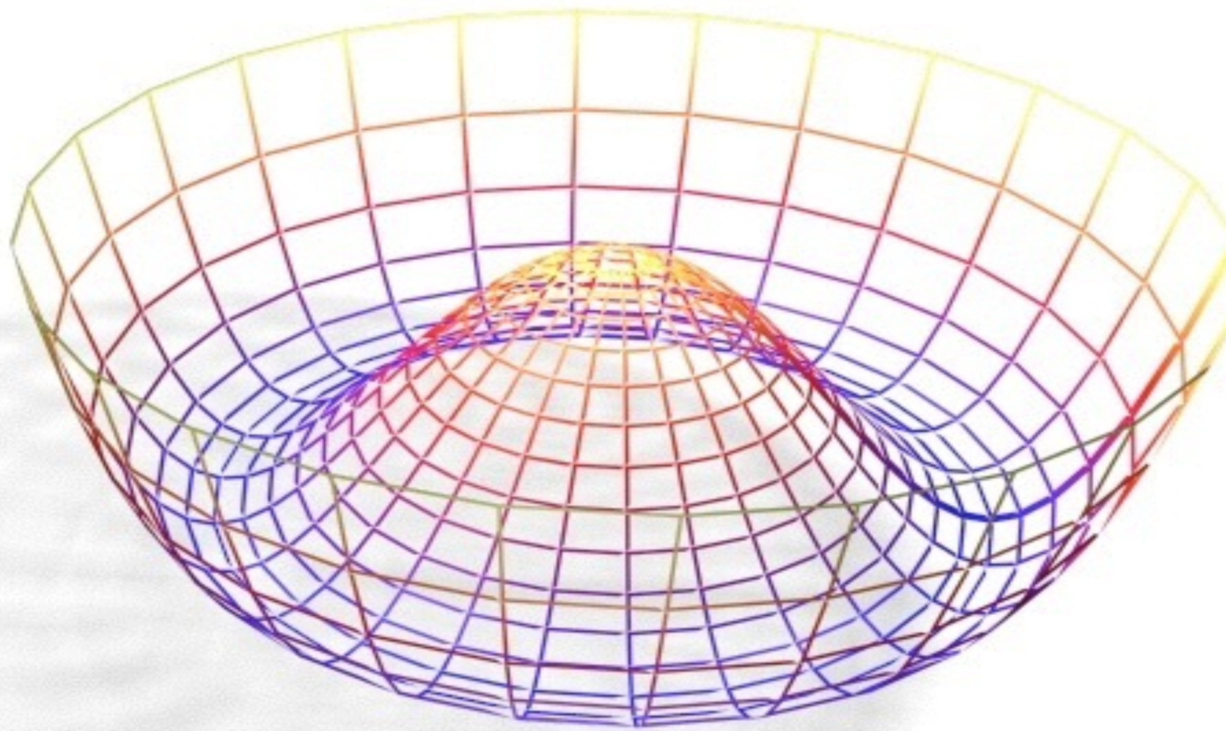
NYU Center for
Data Science

CENTER FOR
COSMOLOGY AND
PARTICLE PHYSICS

The Data Science Challenges of Particle Physics



Kyle Cranmer,
New York University

Paris-Saclay
Center for Data Science

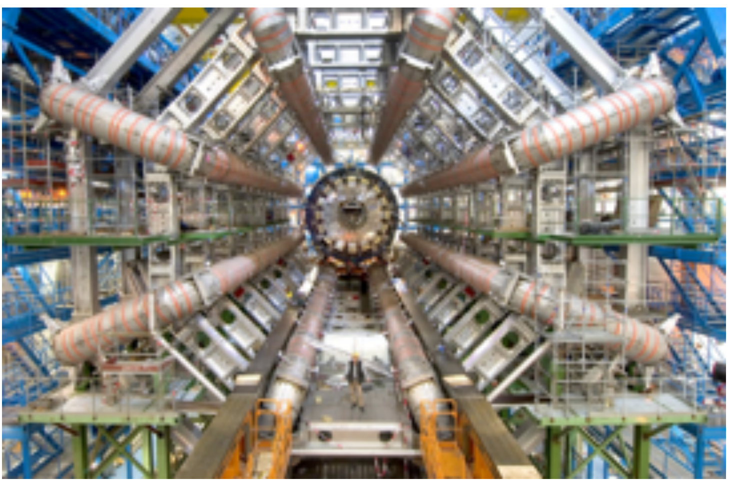
BIO

- Experimental Particle Physicist
- Statistics Convener of ATLAS experiment at LHC
- Founder of RooStats framework (used for Higgs discovery)
- Co-lead Open Science Working Group for Moore-Sloan Data Science Environment at NYU

A harbinger for things to come

Large, Distributed Collaborations

Big Science



Complicated Sensor Environment

Big Data

Big Simulation

$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G_a^{\mu\nu}}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$
$$+ \underbrace{\bar{L}\gamma^\mu(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)L + \bar{R}\gamma^\mu(i\partial_\mu - \frac{1}{2}g'YB_\mu)R}_{\text{kinetic energies and electroweak interactions of fermions}}$$
$$+ \underbrace{\frac{1}{2}\left|(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu)\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{ and Higgs masses and couplings}}$$
$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)G^a_\mu} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}$$

Scientifically Motivated Data Modeling

Big Simulation

Big Model

D S
R C

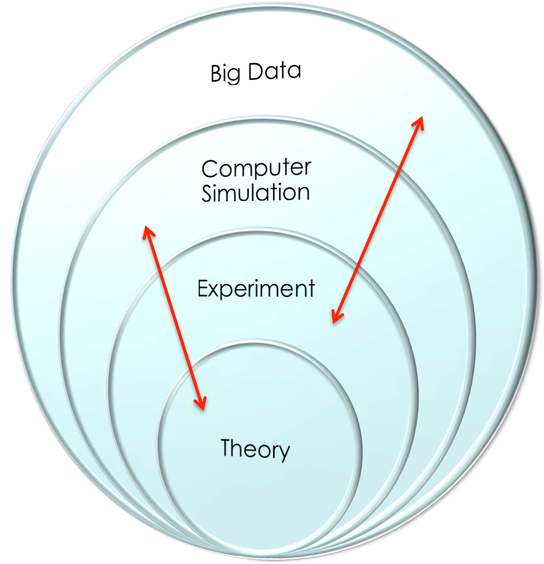
Data Science
Research Center

Complex Models for Big Data

Max Welling
UvA

D S
R C

The Four Paradigms



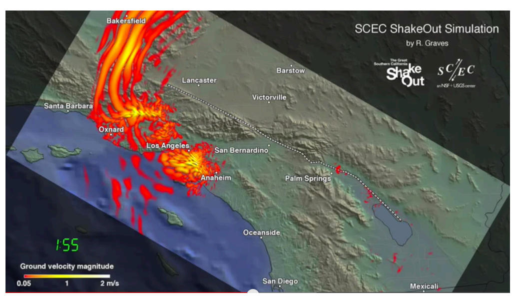
We have added big data to computer simulation, experiment and theory.

Not replaced it…

D S
R C

Big Simulation

Computer simulations have become increasingly complex (e.g. weather, earthquake models)



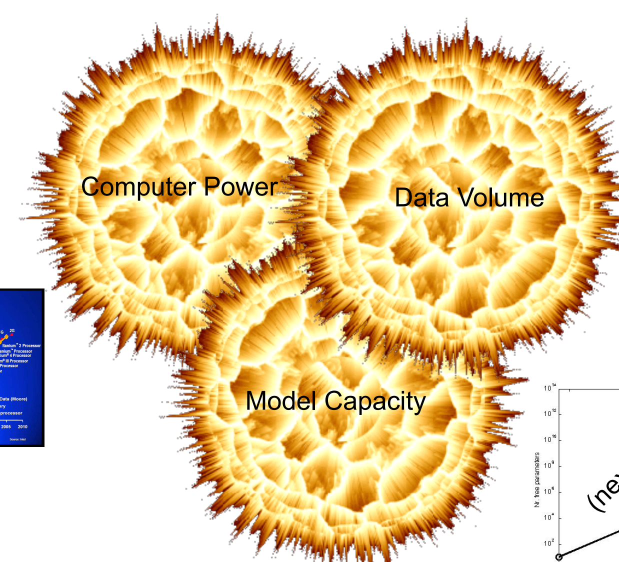
This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

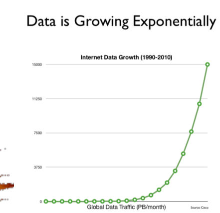
The Computational Wall: If a model has hundreds of parameters, how can we:

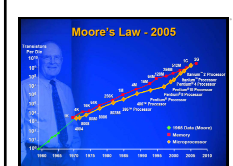
1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

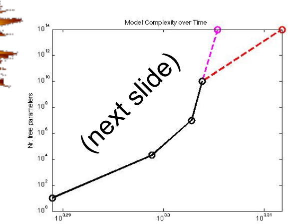
D S
R C

3x Exponential Growth in Machine Learning









Fundamental Particles & Interactions

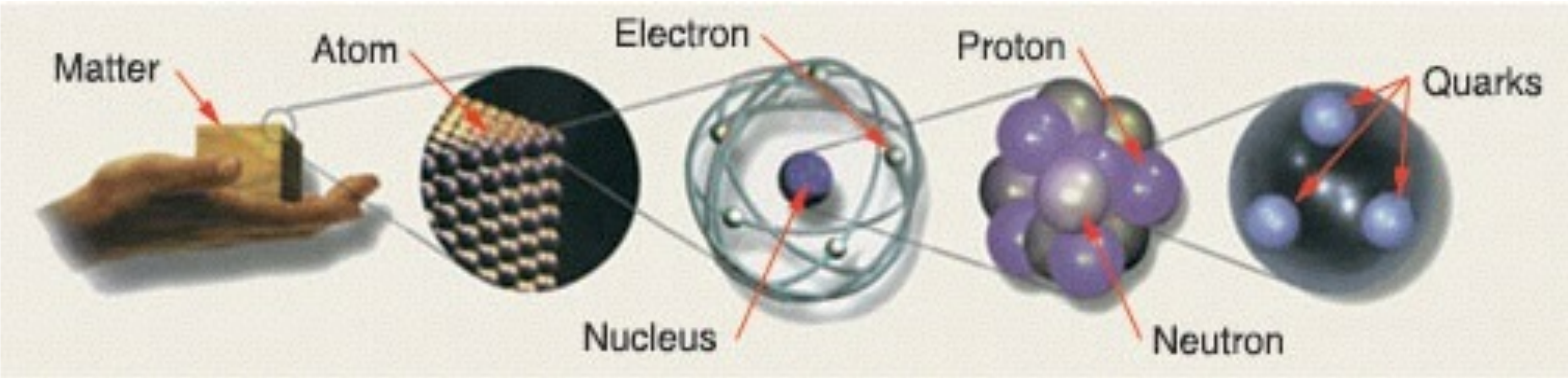
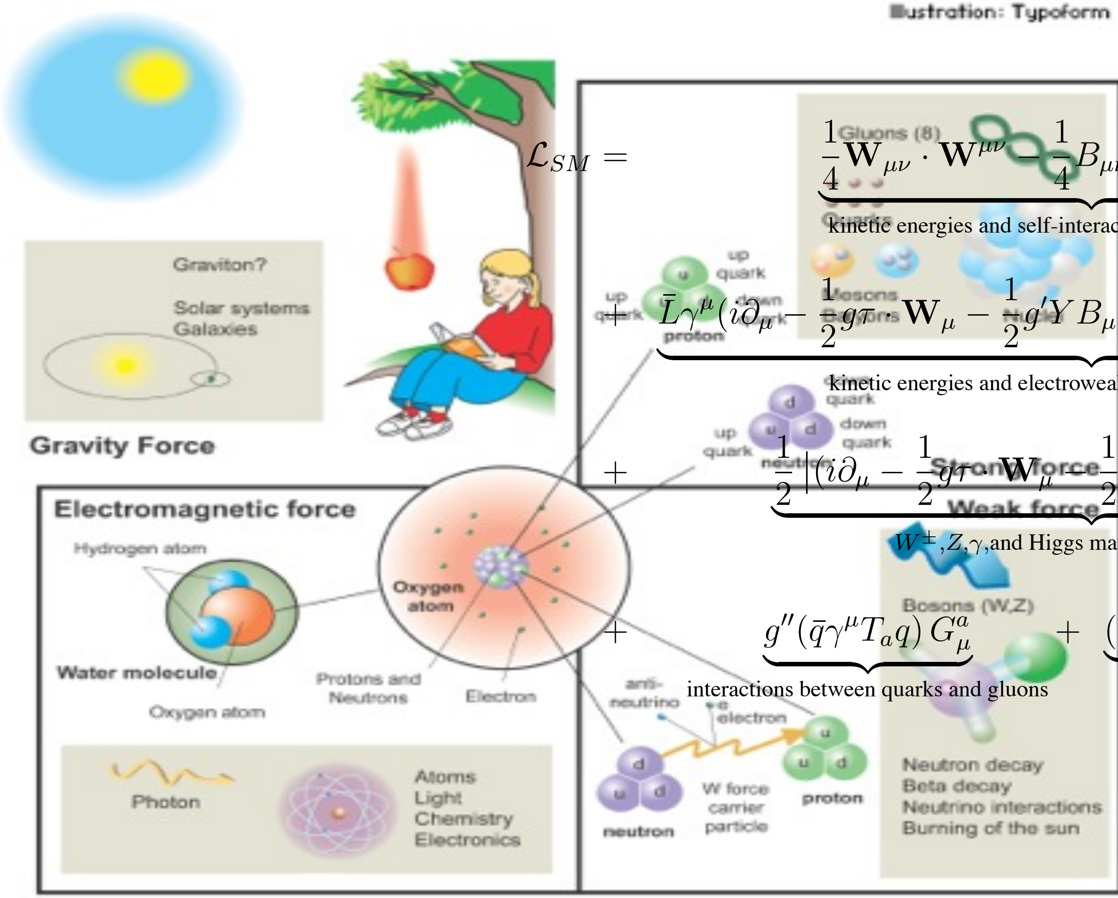


Illustration: Typoform

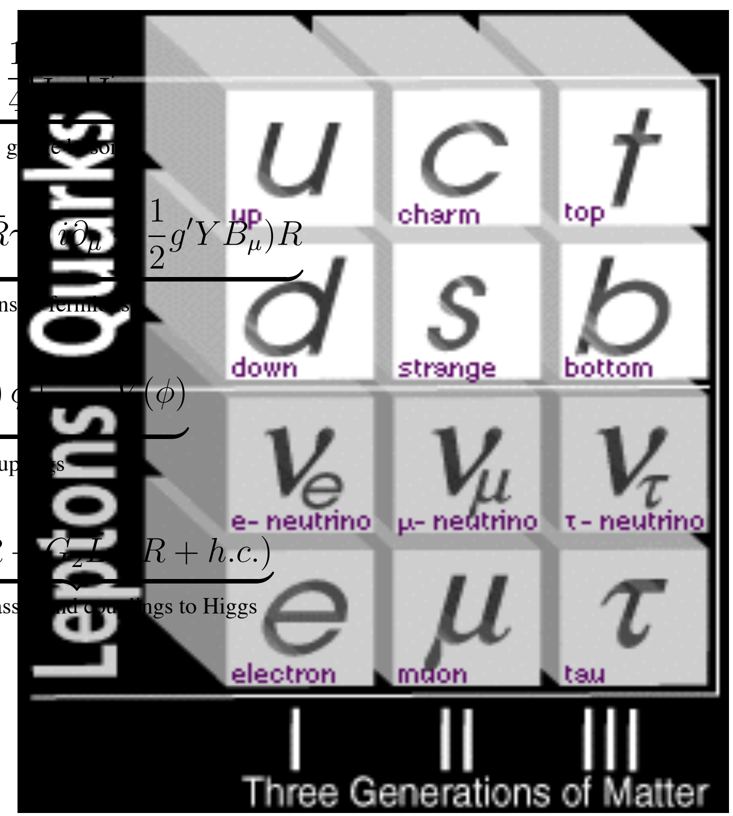


$$\mathcal{L}_{SM} = \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}}$$

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

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

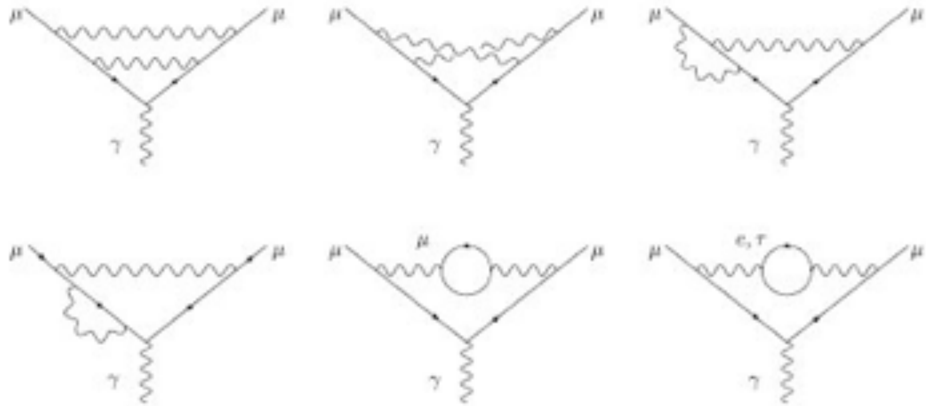
$$+ \underbrace{g''(\bar{q}\gamma^\mu T_a q)\,G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}$$



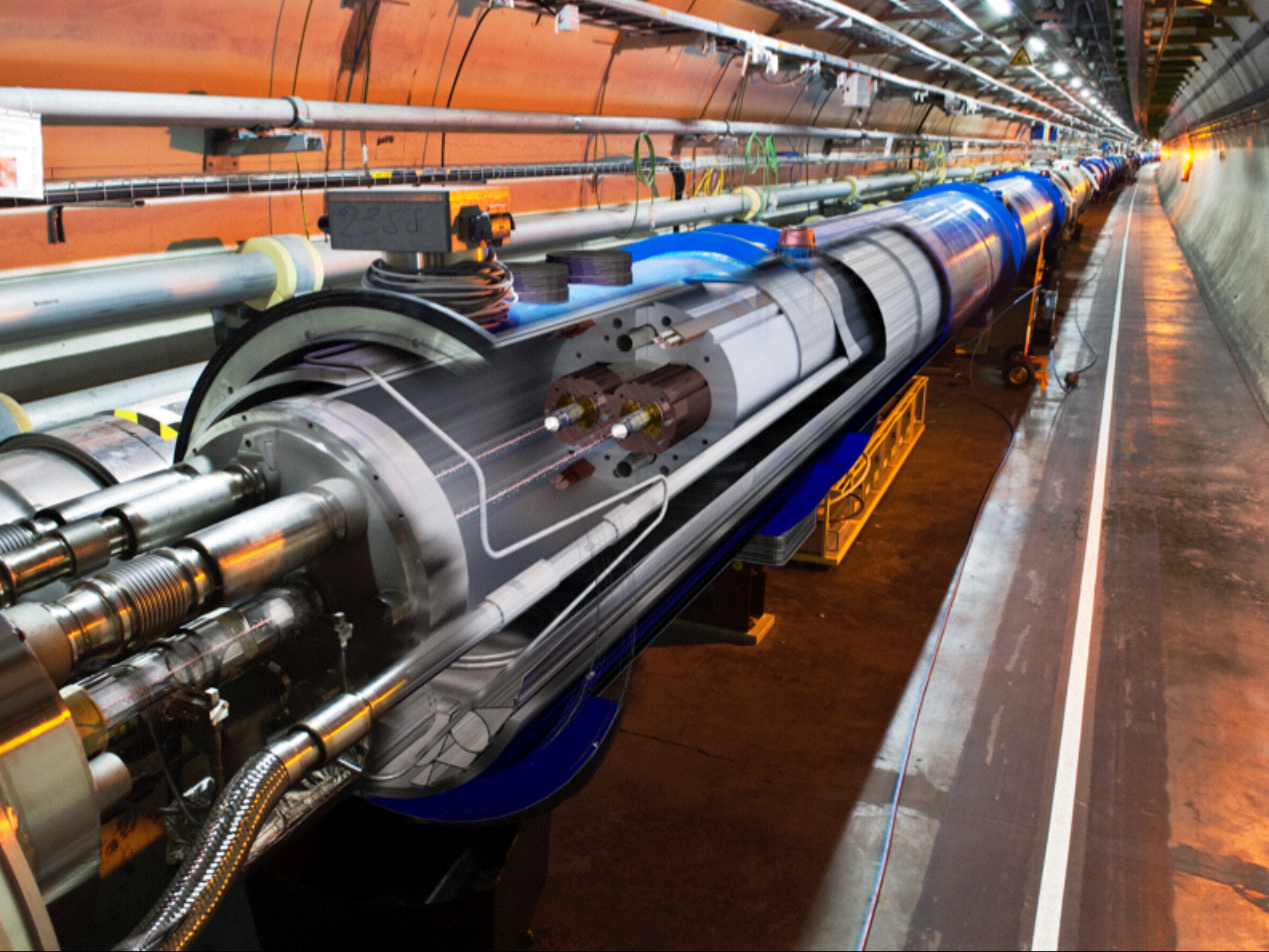
The Success of the Standard Model & QFT

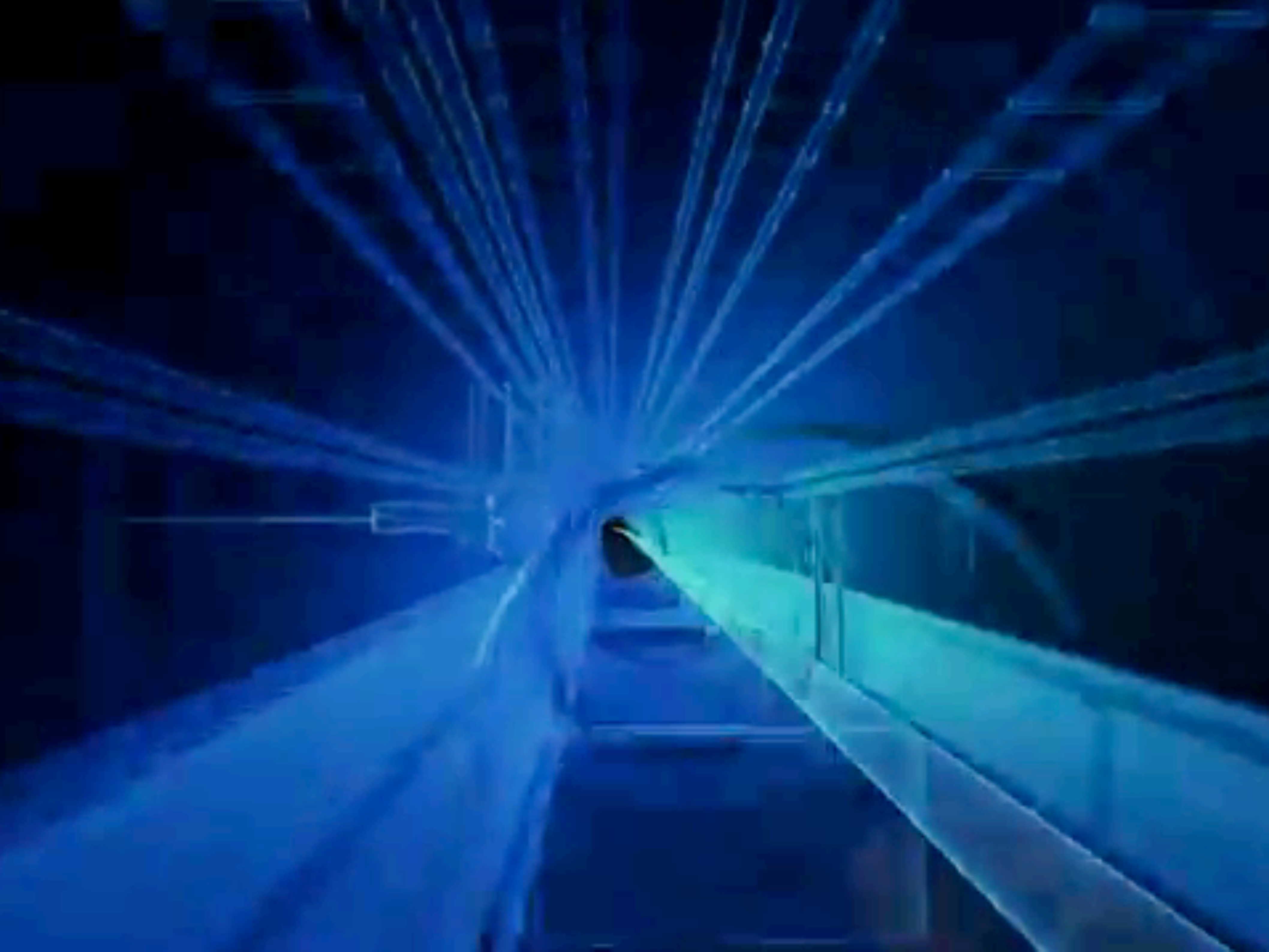
Non-trivial aspects of the theory have been tested to < 1 ppm

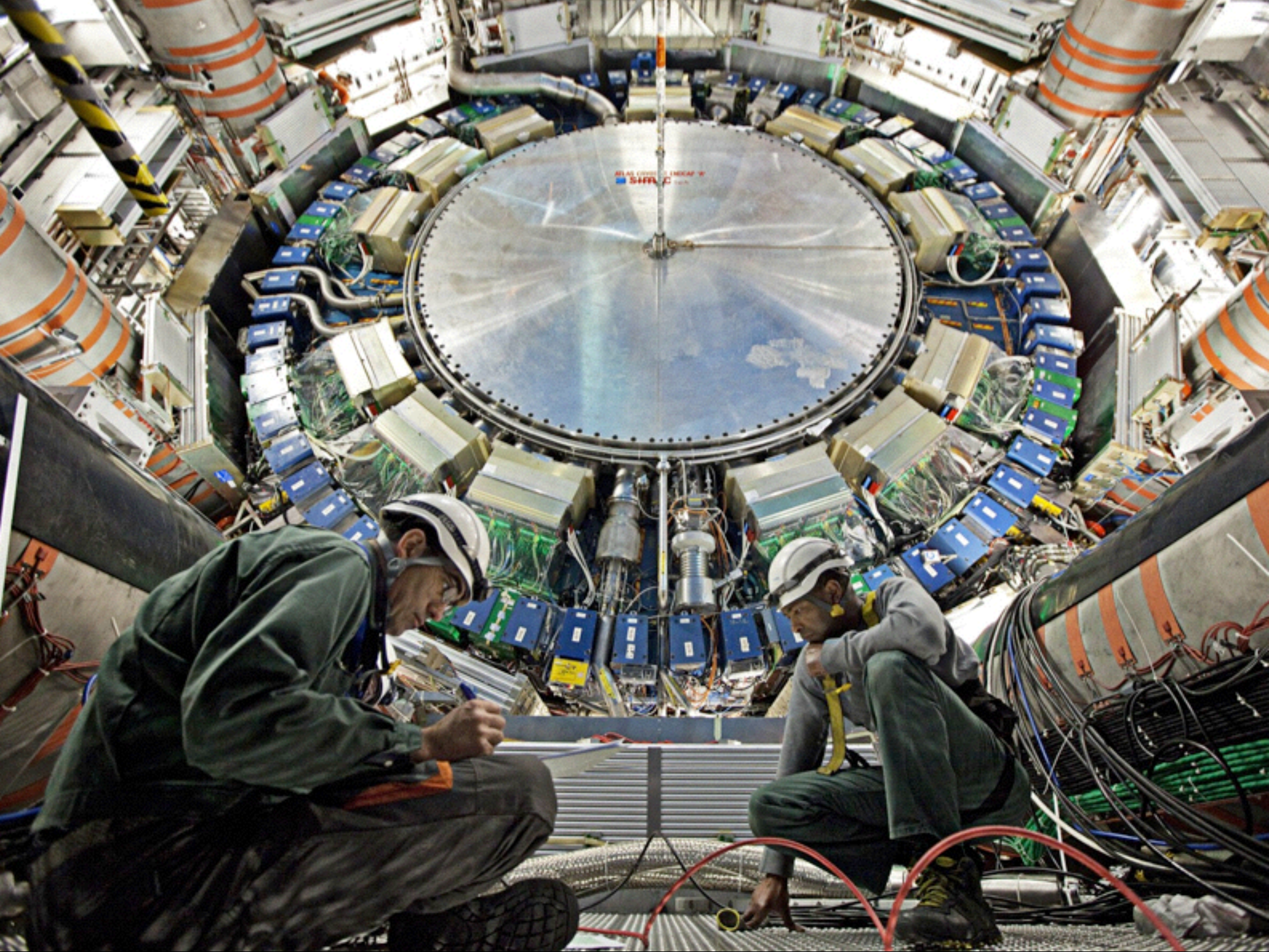
A unique realm for reasonable statistical exploration of a scientific theory



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

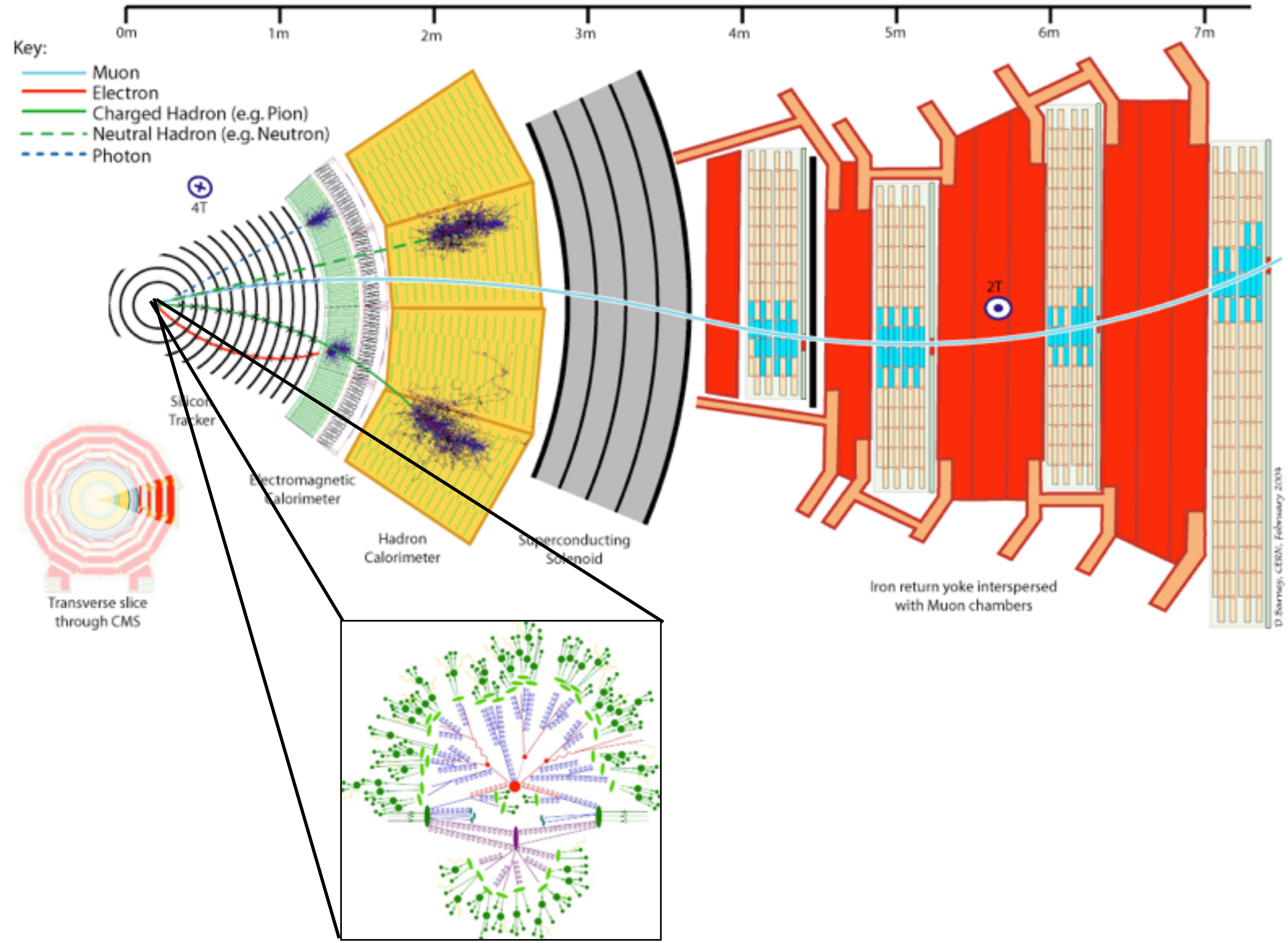







Détecteur "ATLAS"
PROFIL
SLICE of the ATLAS BARREL

>



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

Overview of Predictions

Center for Cosmology and Particle Physics

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

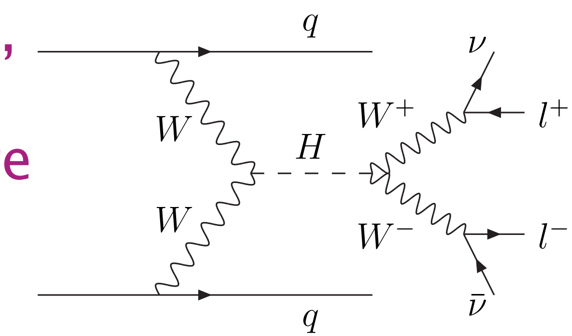
$$
+ \underbrace{\bar{L}\gamma^\mu\left(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu\right)L + \bar{R}\gamma^\mu\left(i\partial_\mu - \frac{1}{2}g'YB_\mu\right)R}_{\text{kinetic energies and electroweak interactions of fermions}}
$$

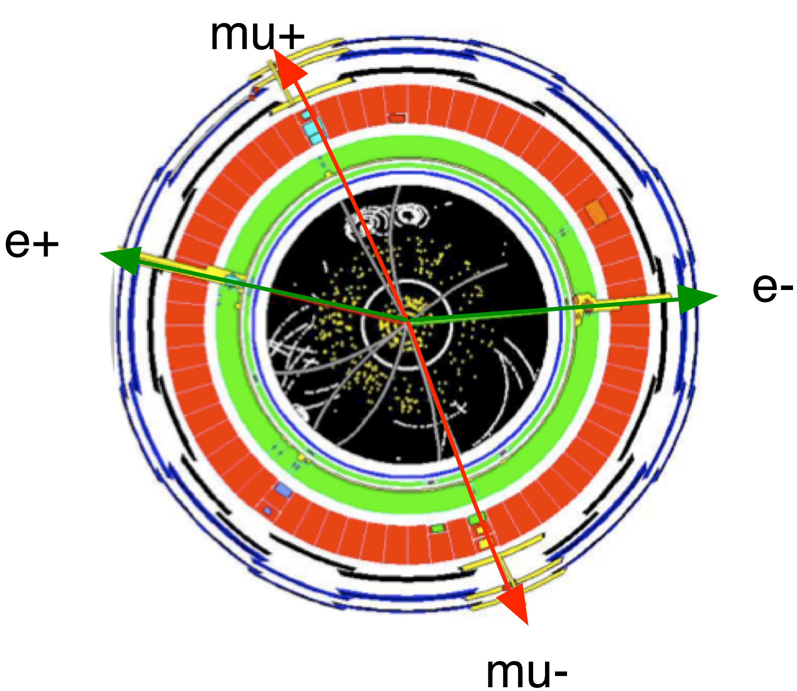
$$
+ \underbrace{\frac{1}{2}\left|\left(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu\right)\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}}
$$

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

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



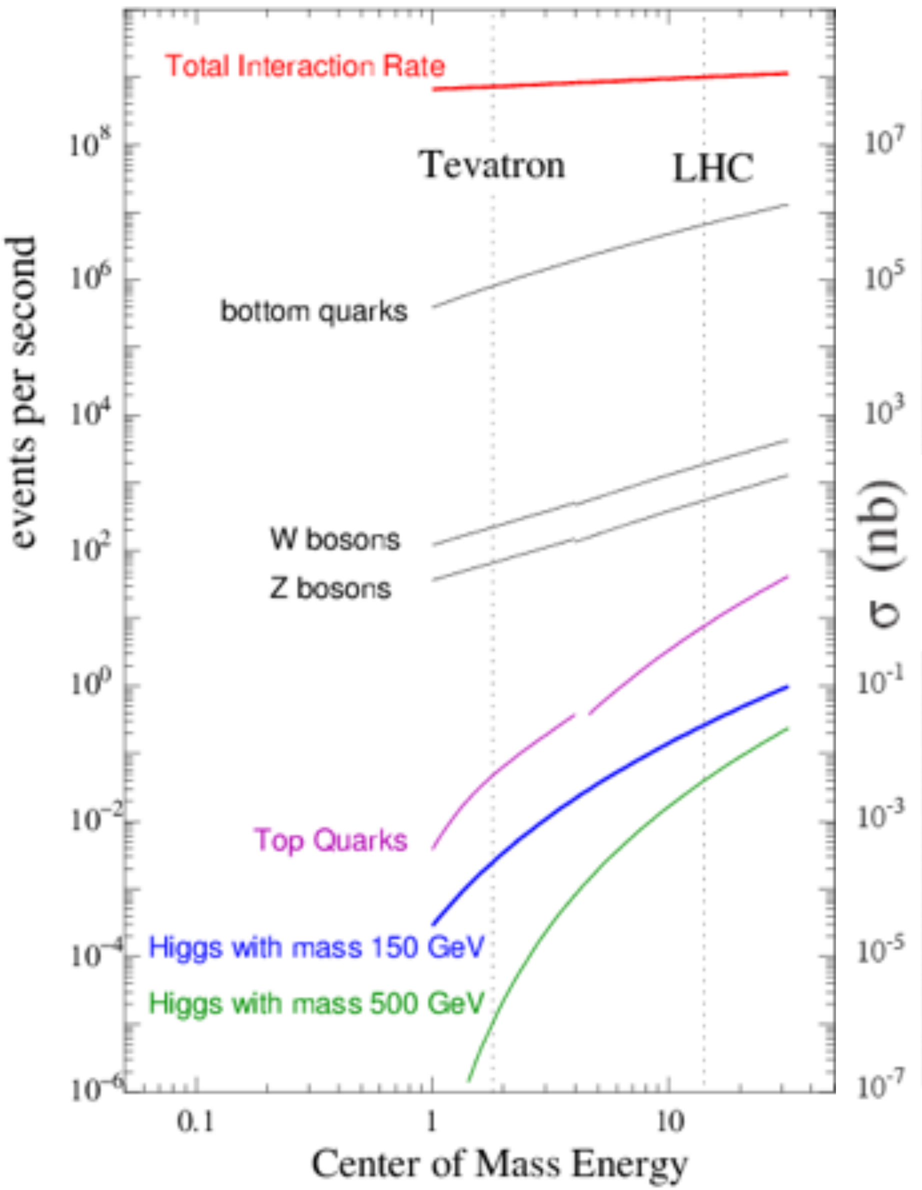


3) The interaction of outgoing particles with the detector is simulated.

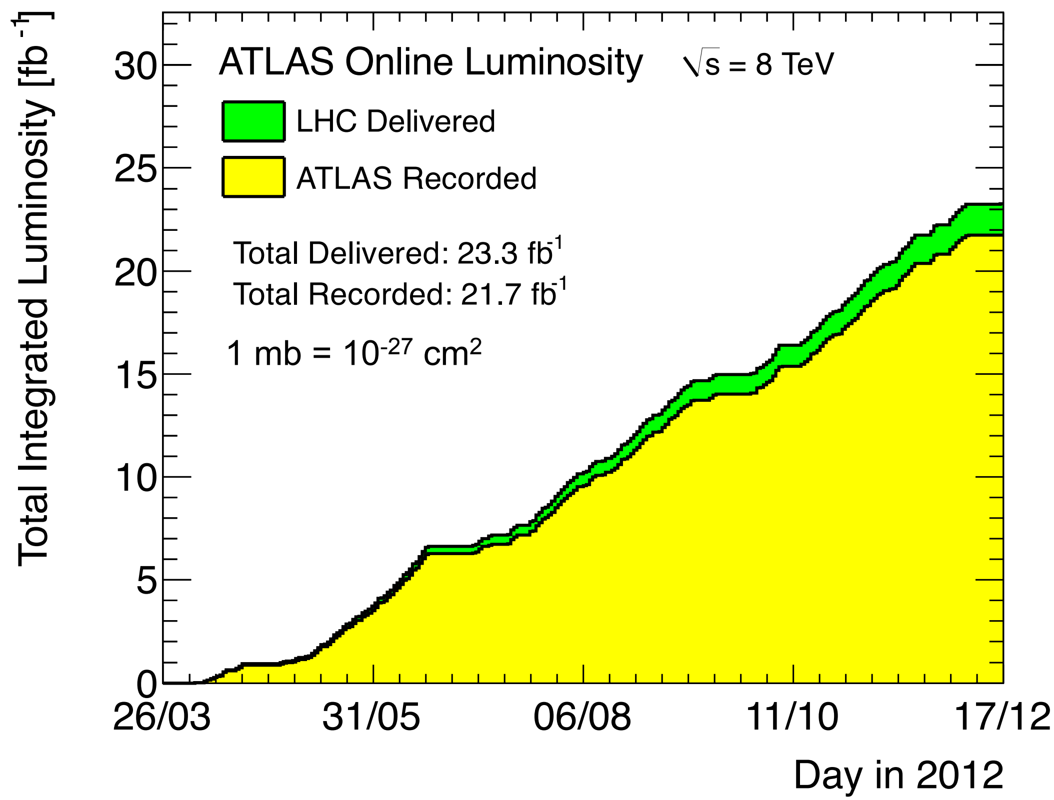
4) Finally, we run algorithms on the simulated data as if they were from real collisions.

Number of collisions

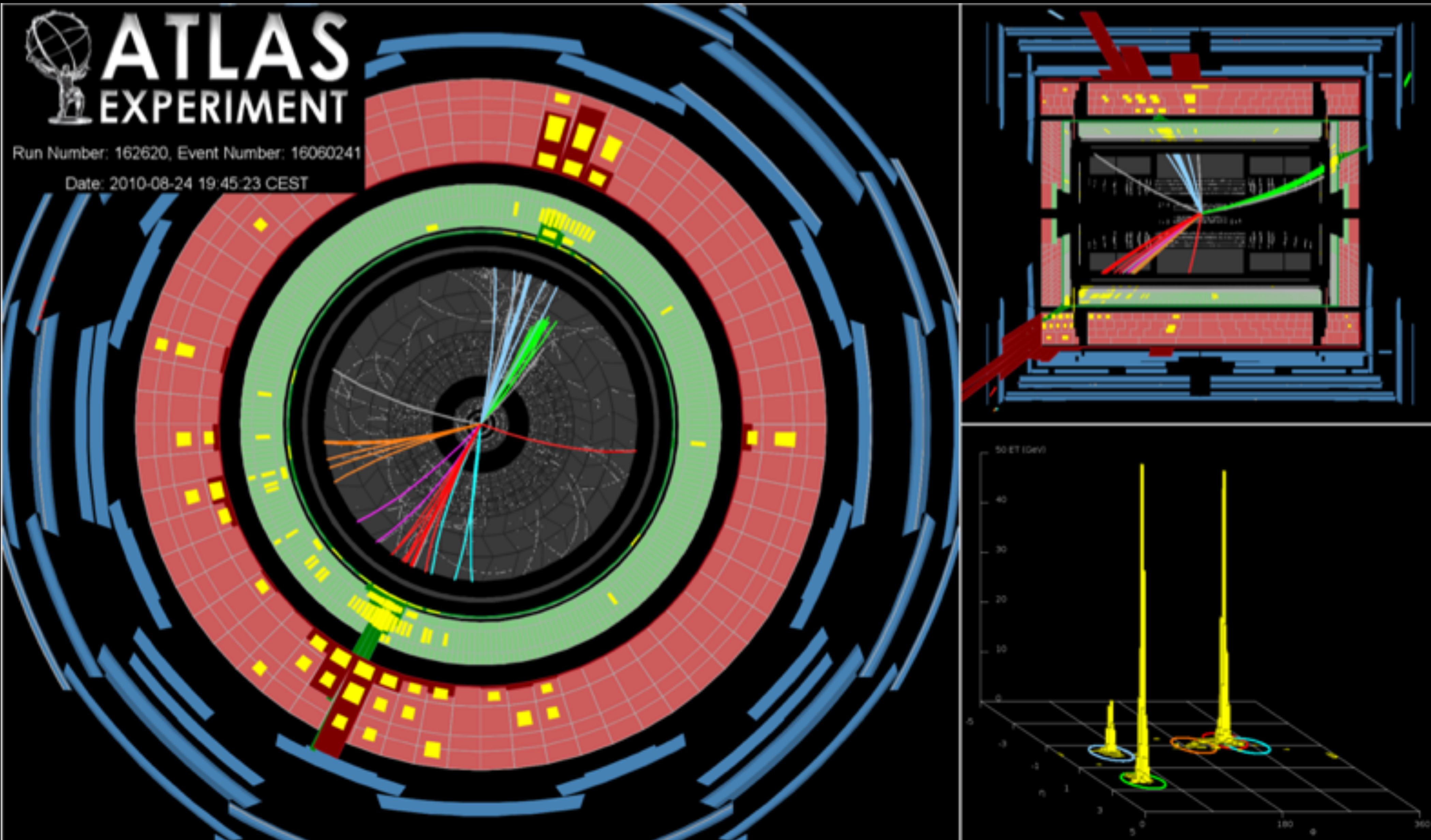
expected number of scatterings = cross section [cm^2] x Luminosity [$1/cm^2$]



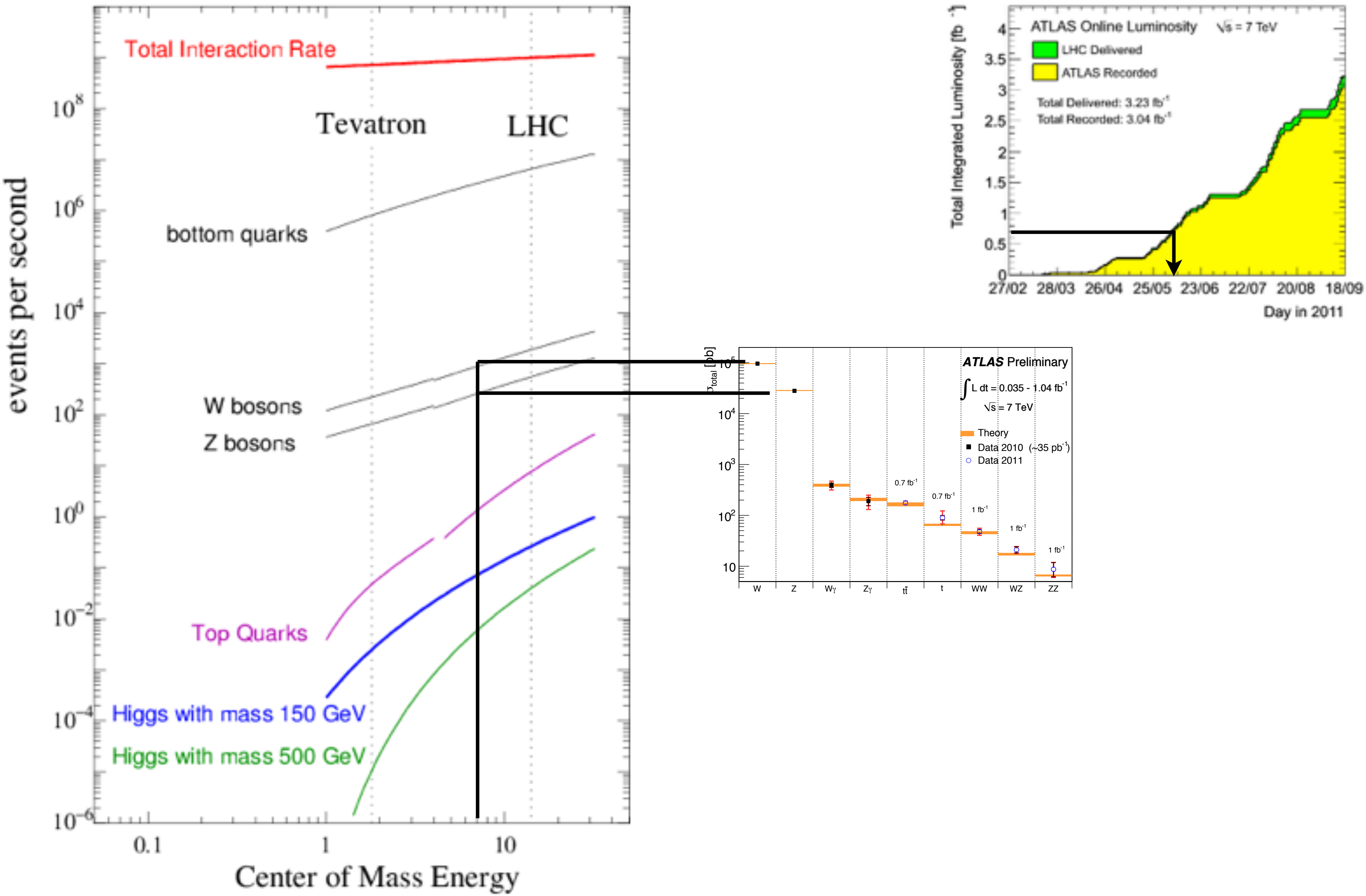
$80\ \text{mb} \cdot 25\ \text{fb}^{-1} = 2 \cdot 10^{15}$ collisions

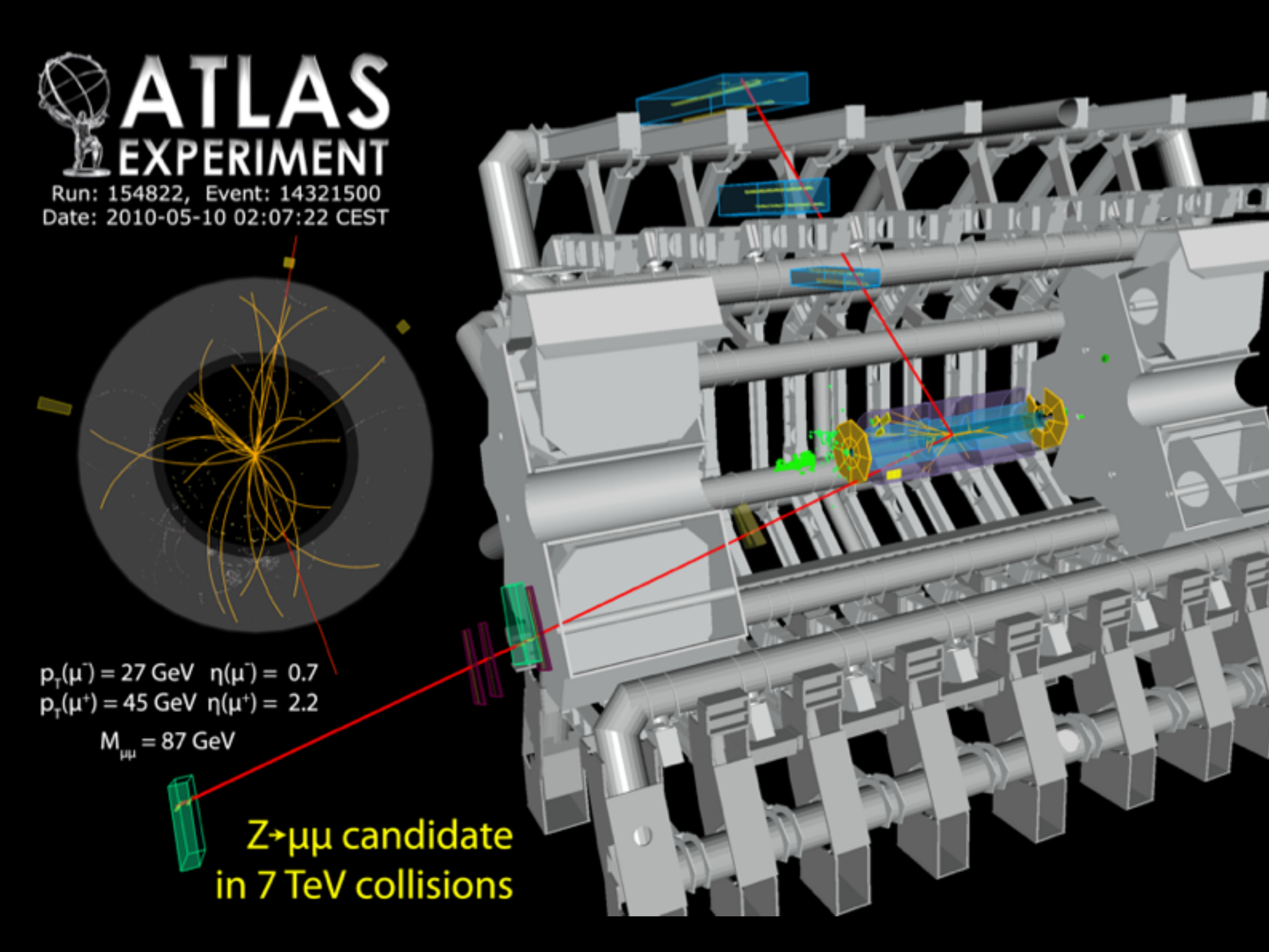


signal : background ~1 : 10^9

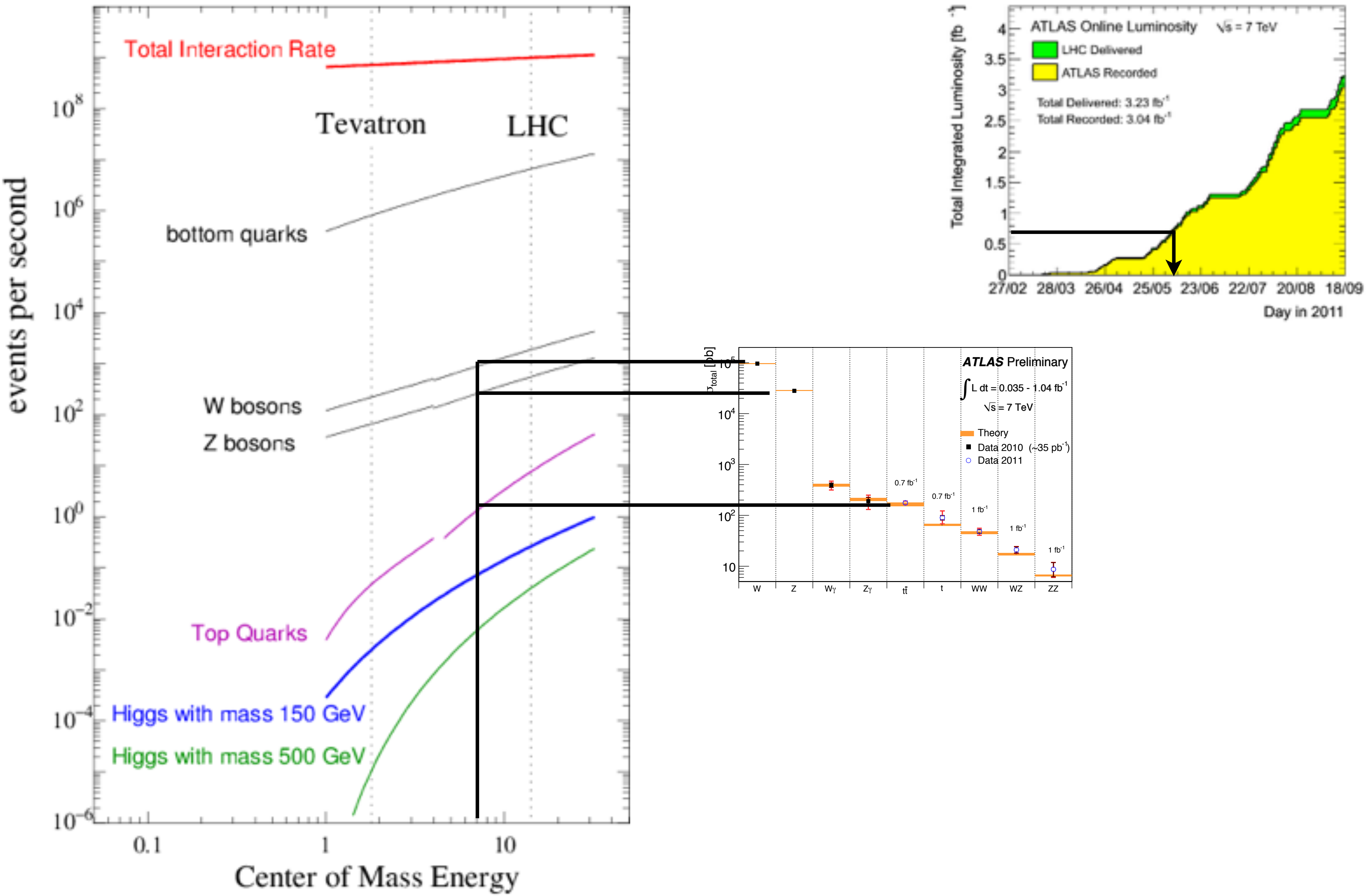

ATLAS
EXPERIMENT
Run Number: 162620, Event Number: 16060241
Date: 2010-08-24 19:45:23 CEST

The steady march of progress

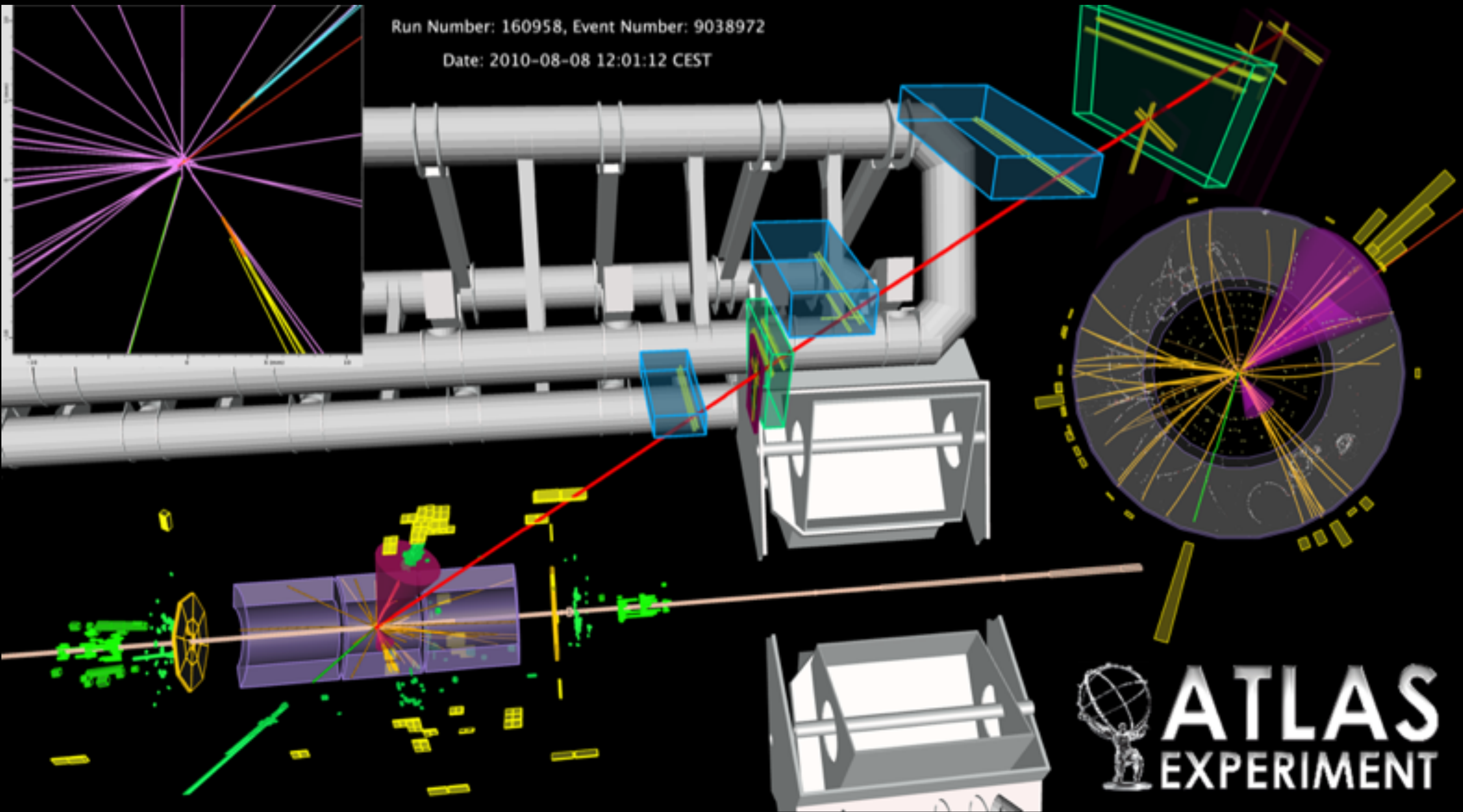



ATLAS EXPERIMENT
Run: 154822, Event: 14321500
Date: 2010-05-10 02:07:22 CEST
$p_T(\mu^-) = 27$ GeV $\eta(\mu^-) = 0.7$
$p_T(\mu^+) = 45$ GeV $\eta(\mu^+) = 2.2$
$M_{\mu\mu} = 87$ GeV
Z→μμ candidate
in 7 TeV collisions

The steady march of progress

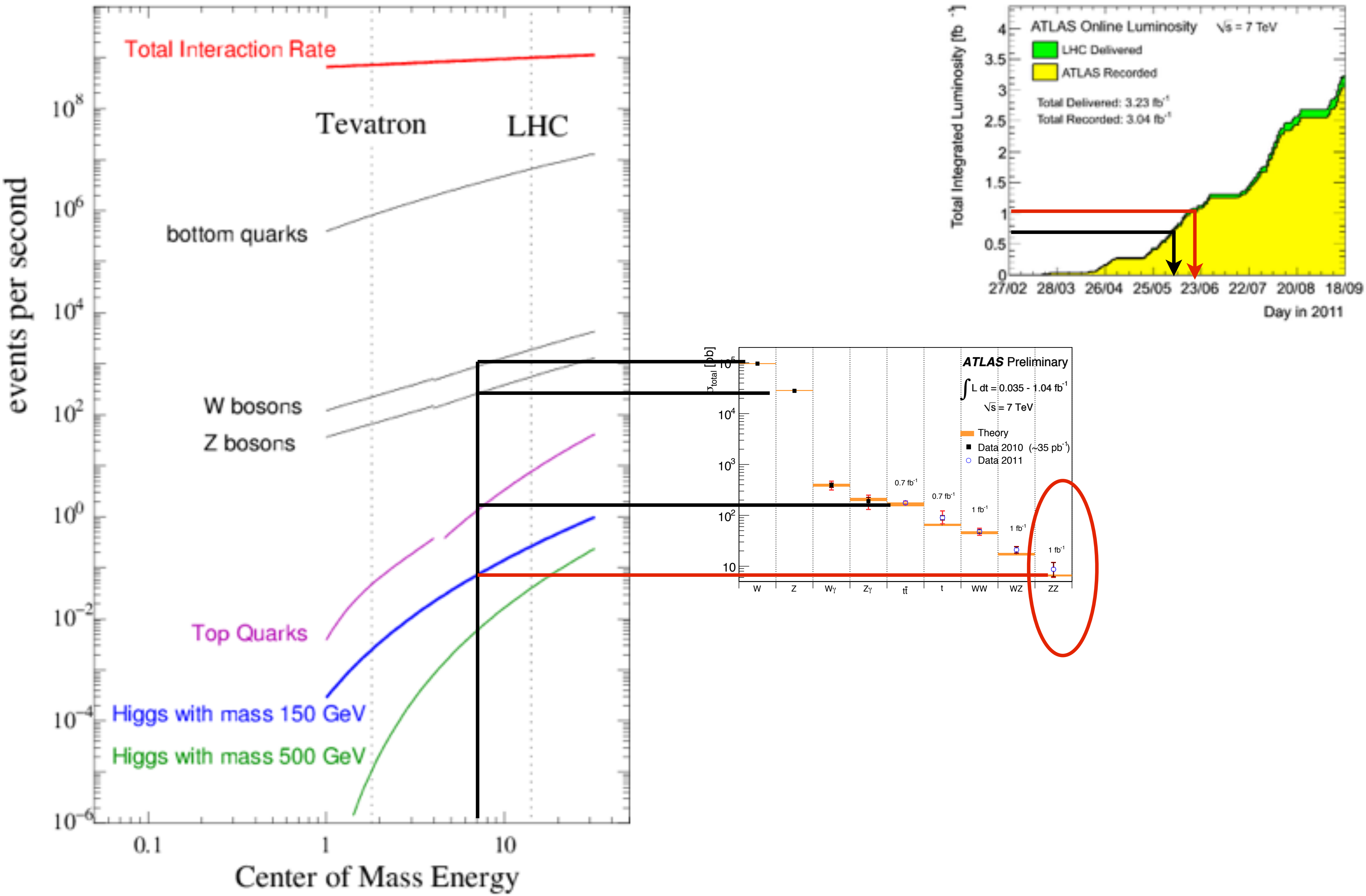


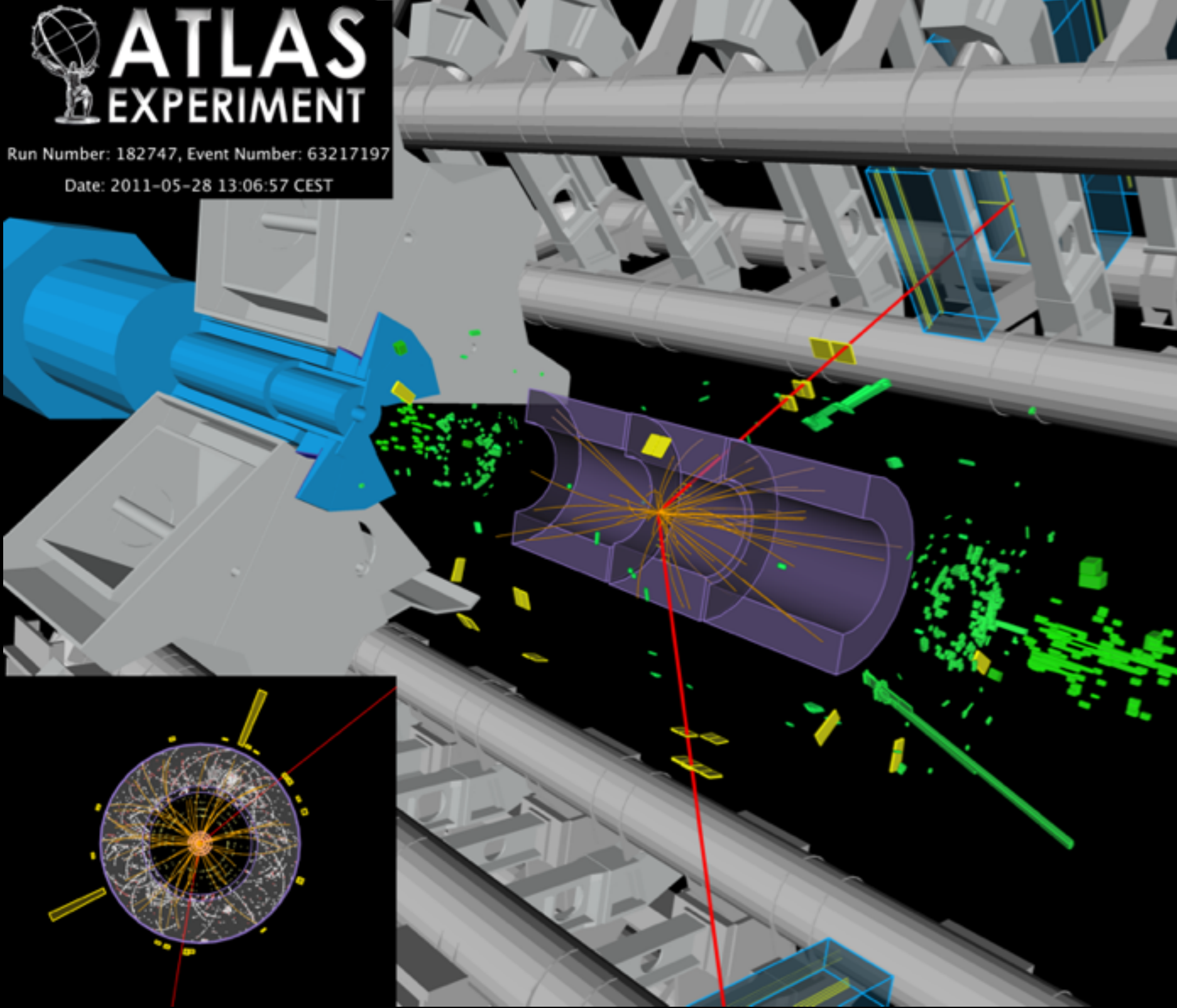
Top quark pair decaying to bb eμ $E_{T,miss}$



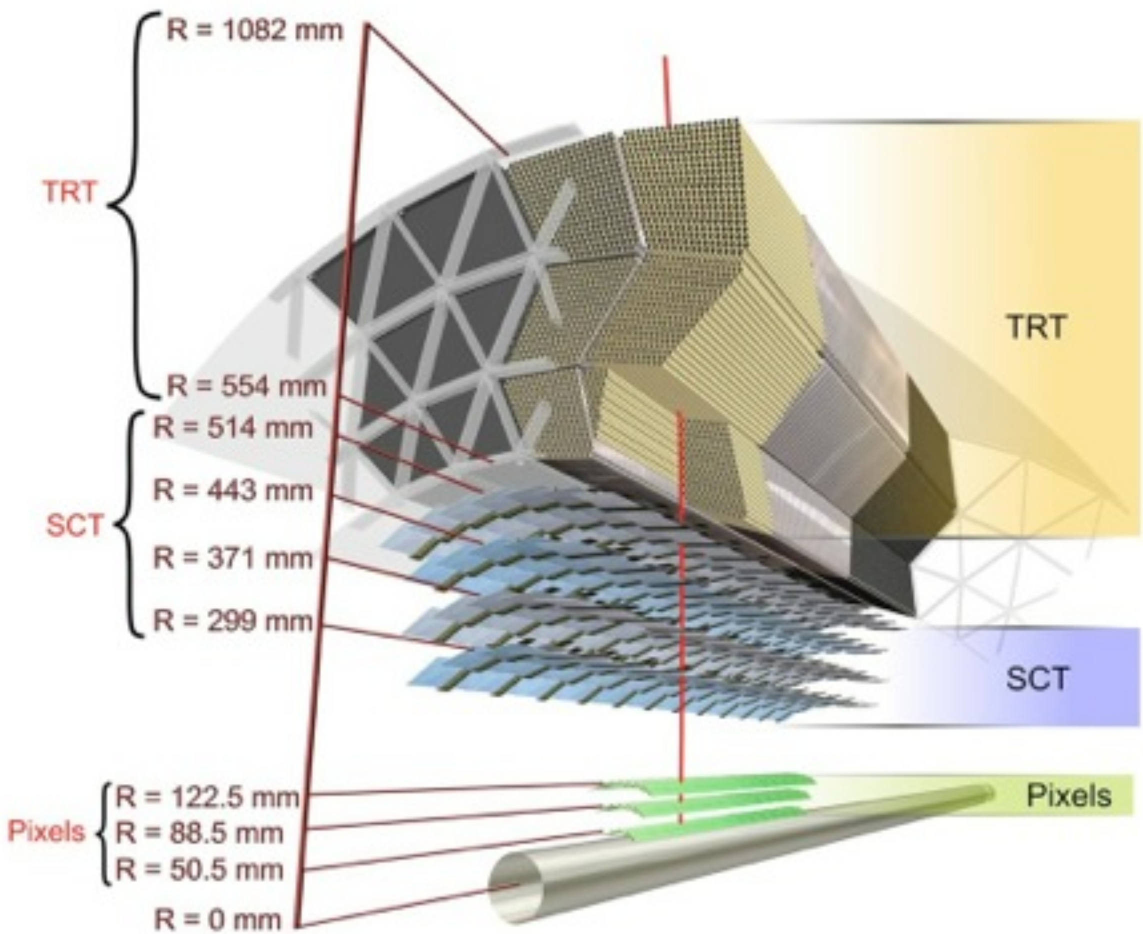
The steady march of progress

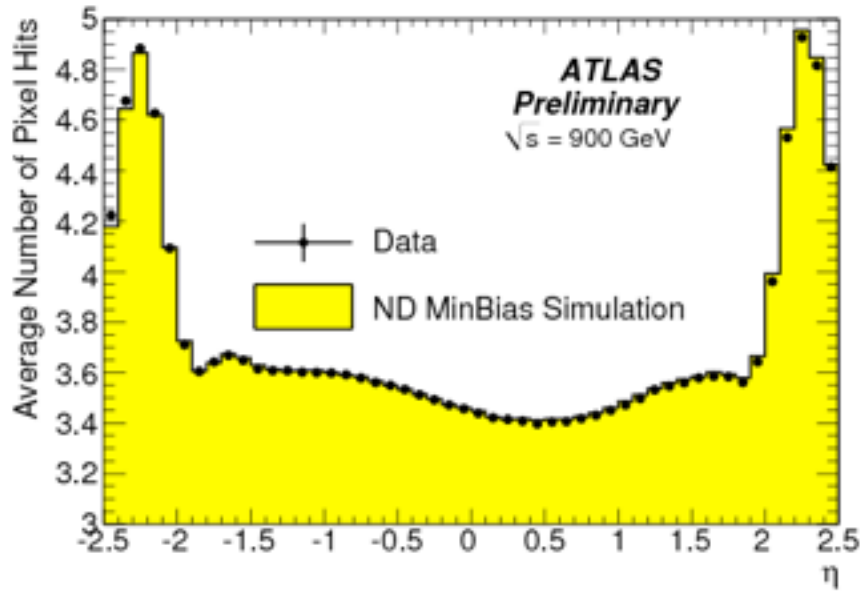
Center for Cosmology and Particle Physics

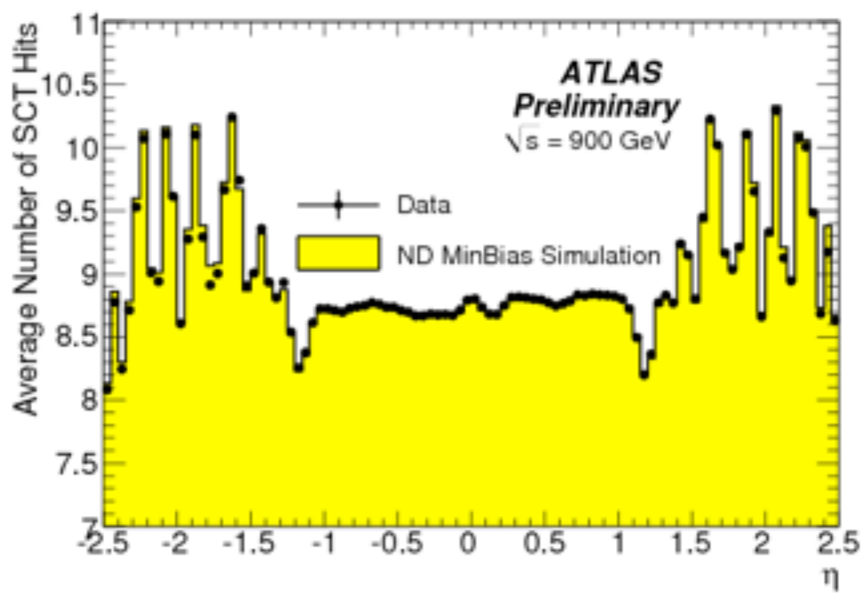


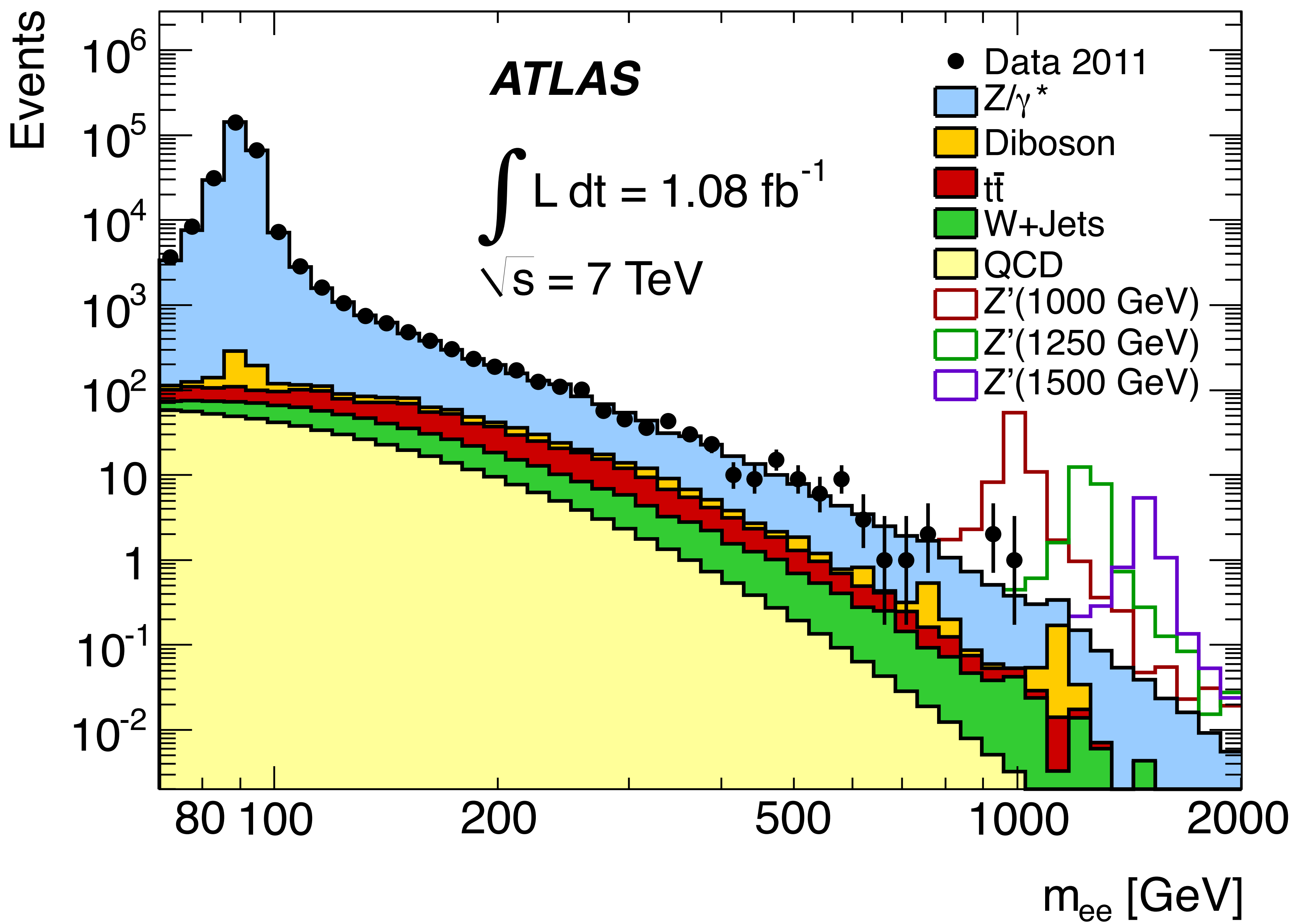

ATLAS
EXPERIMENT
Run Number: 182747, Event Number: 63217197
Date: 2011-05-28 13:06:57 CEST

How good is the modeling?









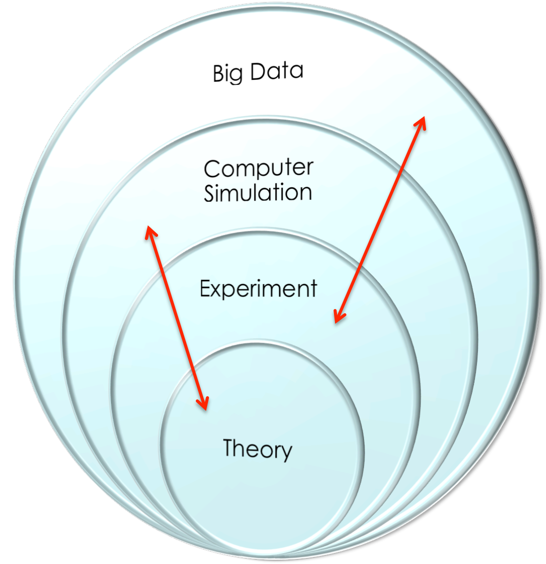
Events
ATLAS
$\int L\,dt = 1.08\ \mathrm{fb}^{-1}$
$\sqrt{s} = 7$ TeV
Data 2011
Z/γ^*
Diboson
$t\bar{t}$
W+Jets
QCD
Z'(1000 GeV)
Z'(1250 GeV)
Z'(1500 GeV)
10^{6}
10^{5}
10^{4}
10^{3}
10^{2}
10
1
10^{-1}
10^{-2}
80 100
200
500
1000
2000
m_{ee} [GeV]

Data Science Research Center

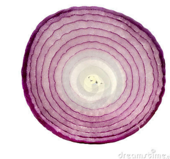
Complex Models for Big Data

Max Welling
UvA

The Four Paradigms



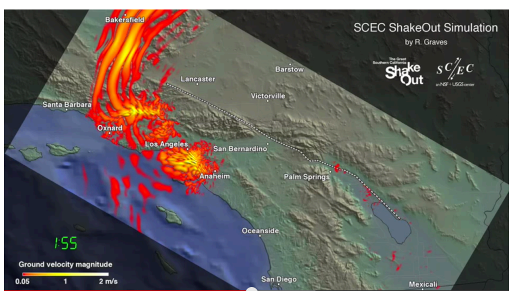
We have added big data to computer simulation, experiment and theory.



Not replaced it…

Big Simulation

Computer simulations have become increasingly complex (e.g. weather, earthquake models)

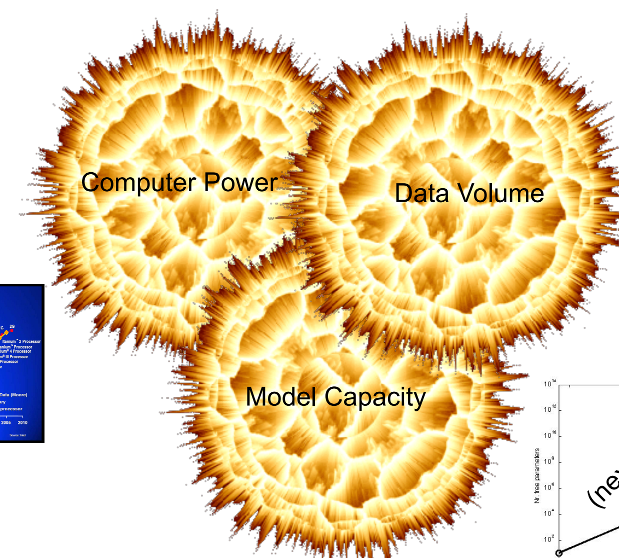


This production run producing 360 sec of wave propagation sustained 220 Tflop/s for 24 hours on NCCS Jaguar using 223,074 cores.

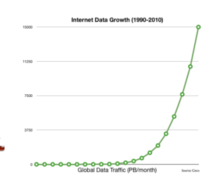
The Computational Wall: If a model has hundreds of parameters, how can we:

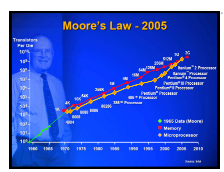
1) Find the parameter values that match the observations best?
2) Determine if we underfit (model too simple) or overfit (model too complex)?
3) Compare two models?

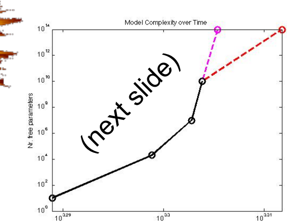
3x Exponential Growth in Machine Learning



Data is Growing Exponentially



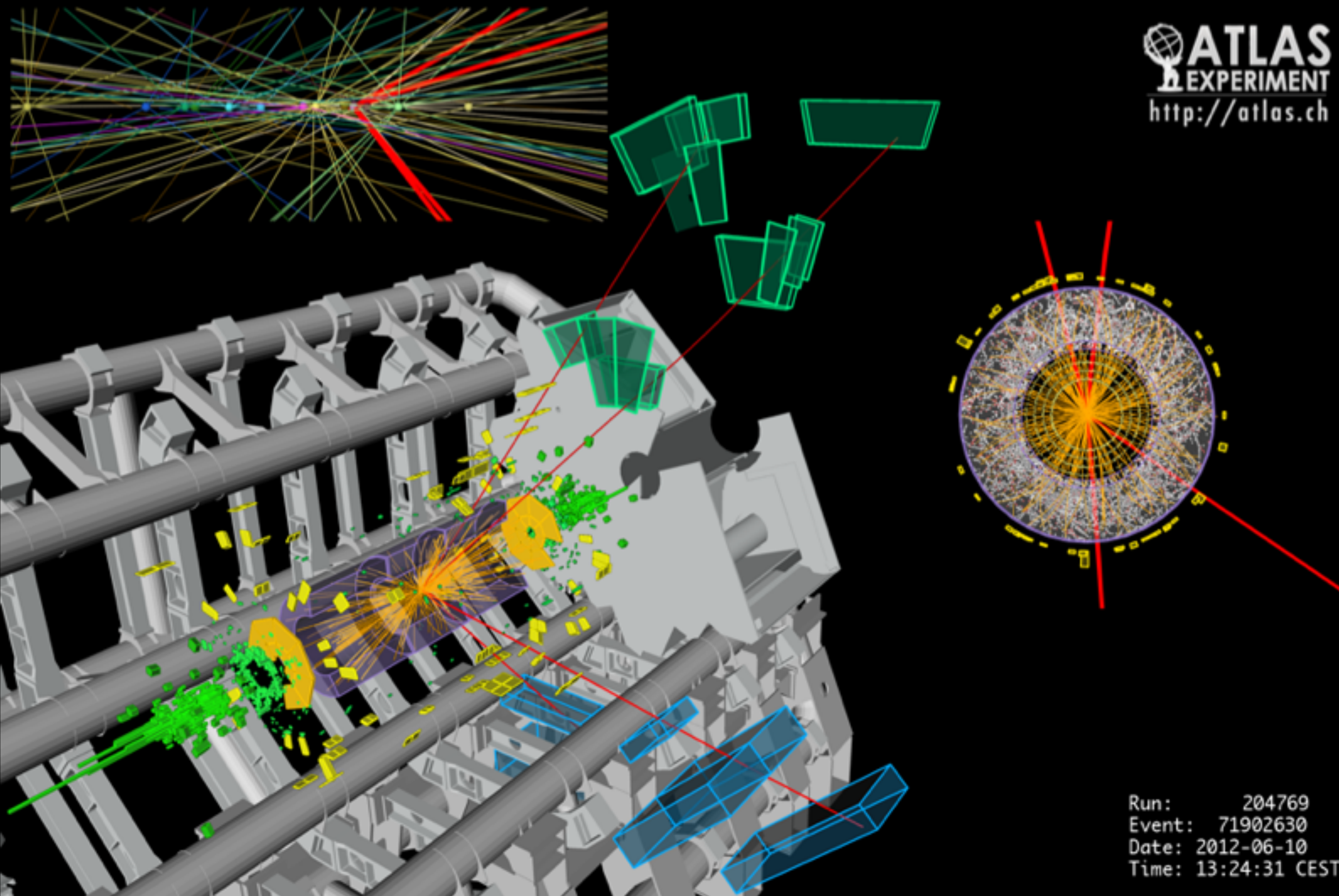




Use of Machine Learning:

Particle-Level
and
Event-Level

$H \to ZZ \to 4l$


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: $E \neq mc^2$

What Einstein really said:

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

Every physics student knows energy and momentum are conserved

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

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

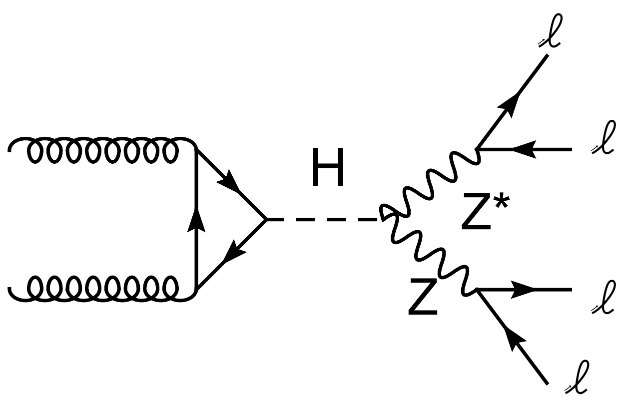
Thus, we can estimate the mass of the Higgs with

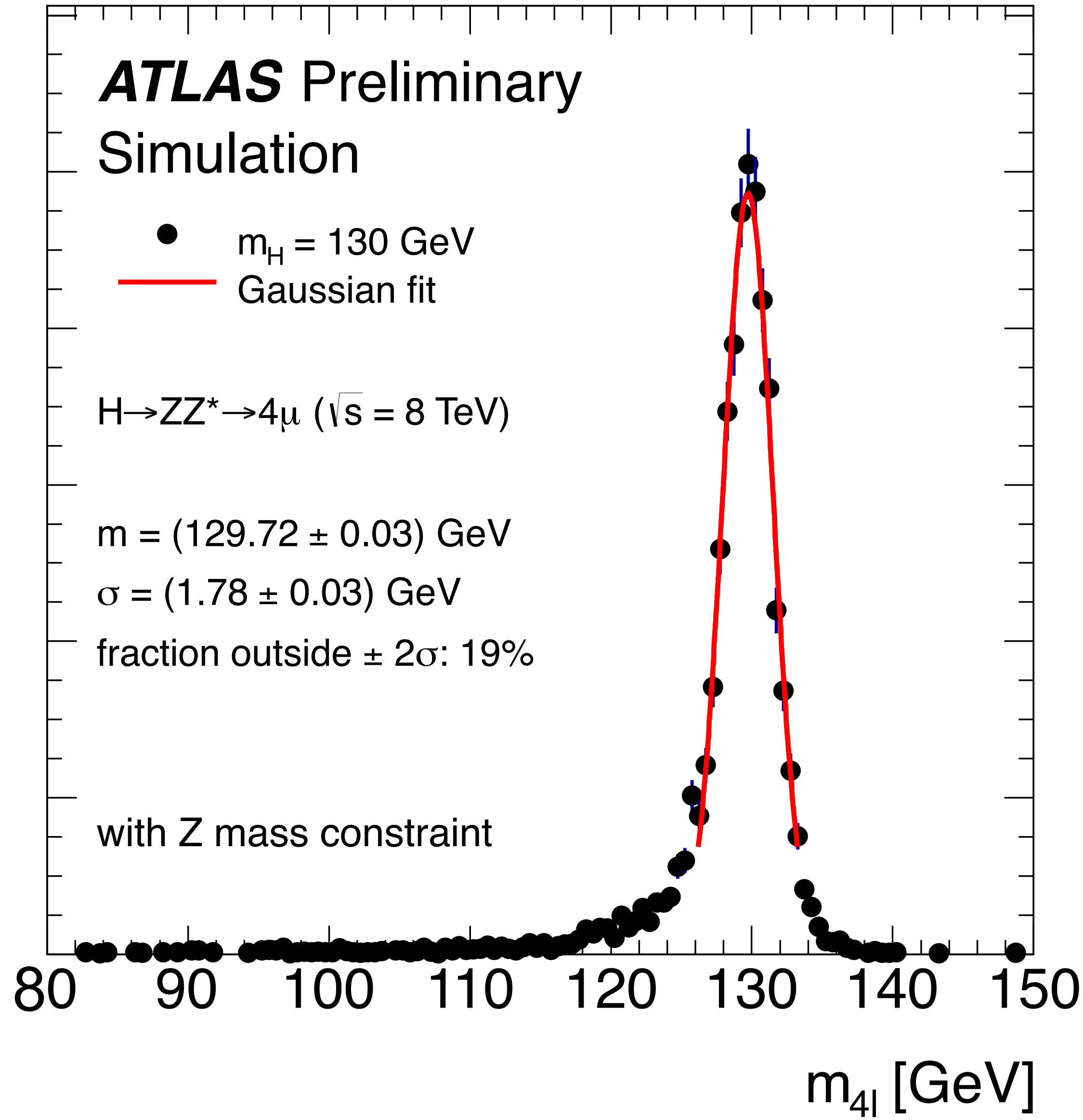
$$m_H = \sqrt{E^2_{\text{after}}/c^4 - |\vec{p}_{\text{after}}|^2/c^2}$$

An example high-level feature

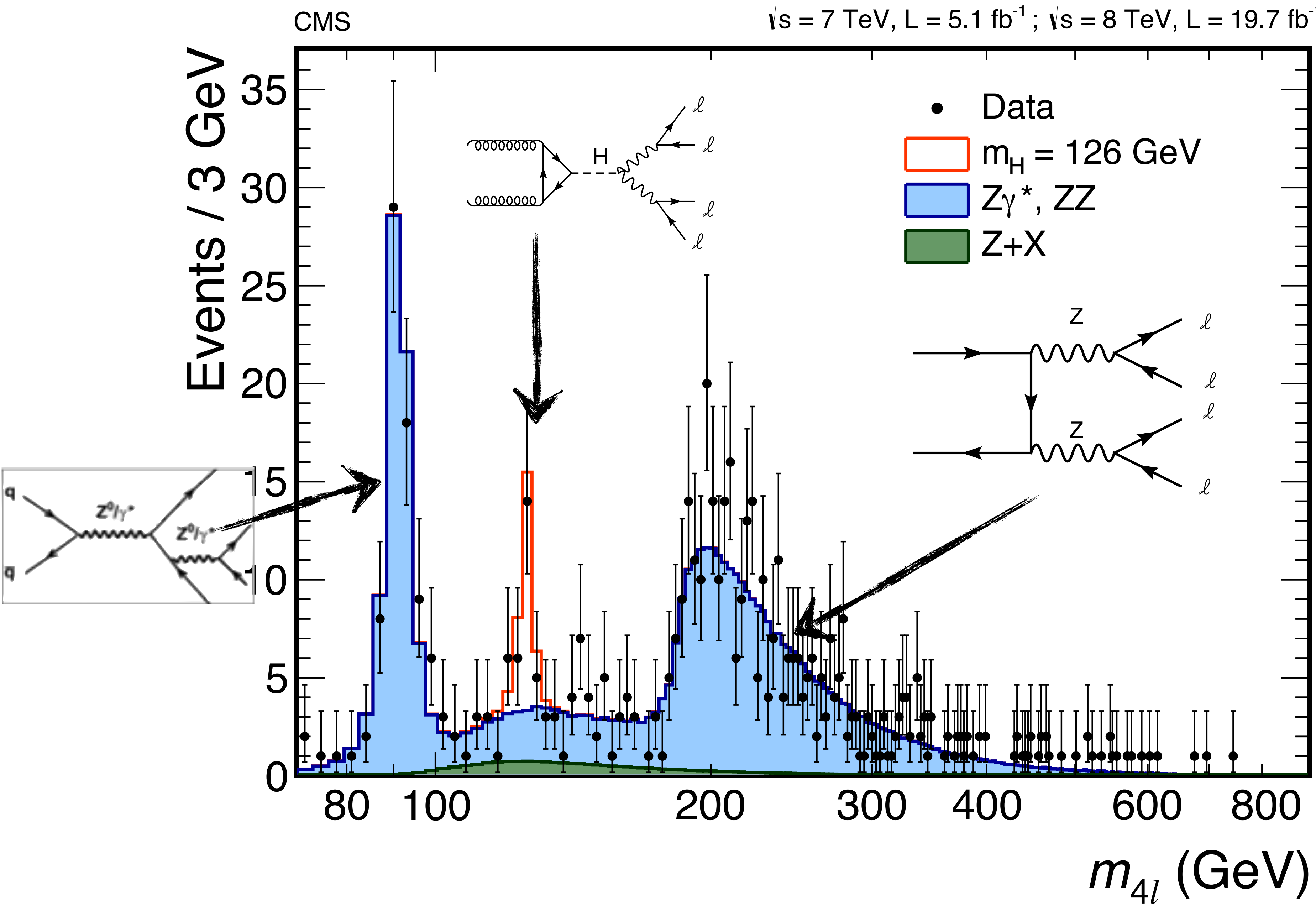
From the 16 energies and momenta measured in this system, this particular combination gives a very sharp feature.

~sufficient statistic



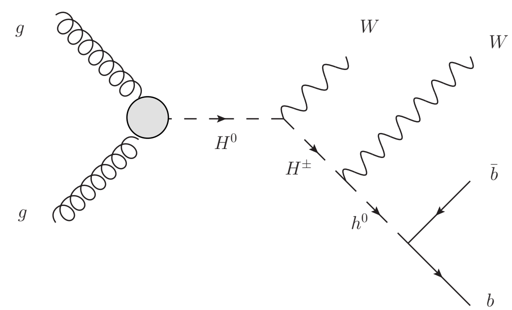


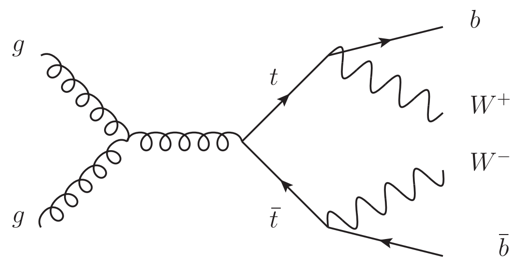
The observation in the 4l channel

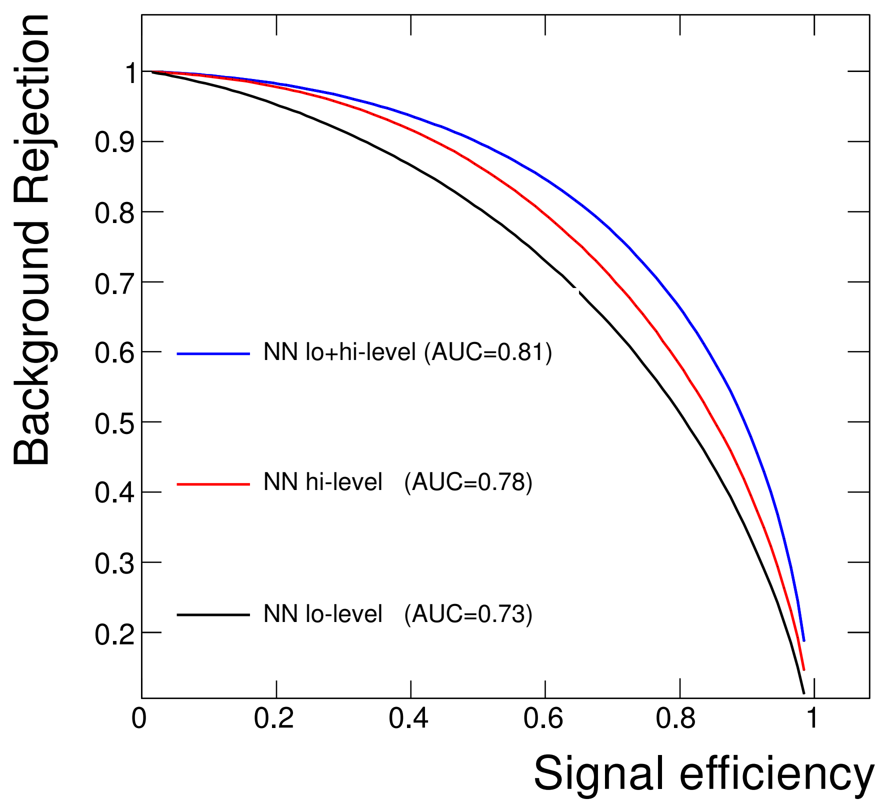


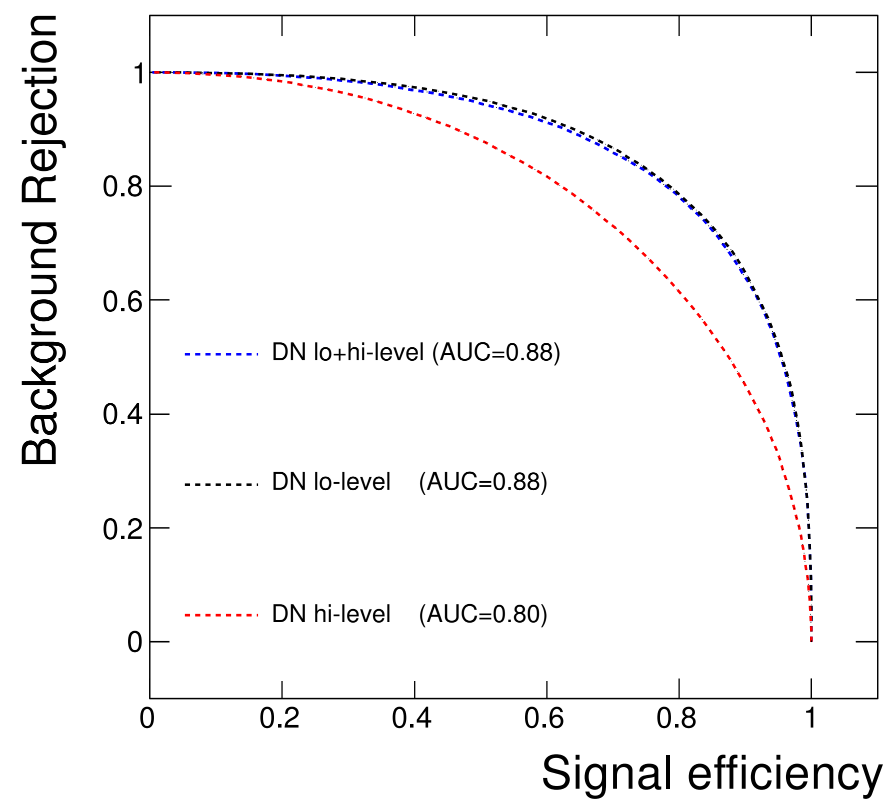
Deep Neural Network for event kinematics

ML techniques performed poorly unless these high-level features were supplied. Deep learning techniques can discover them.



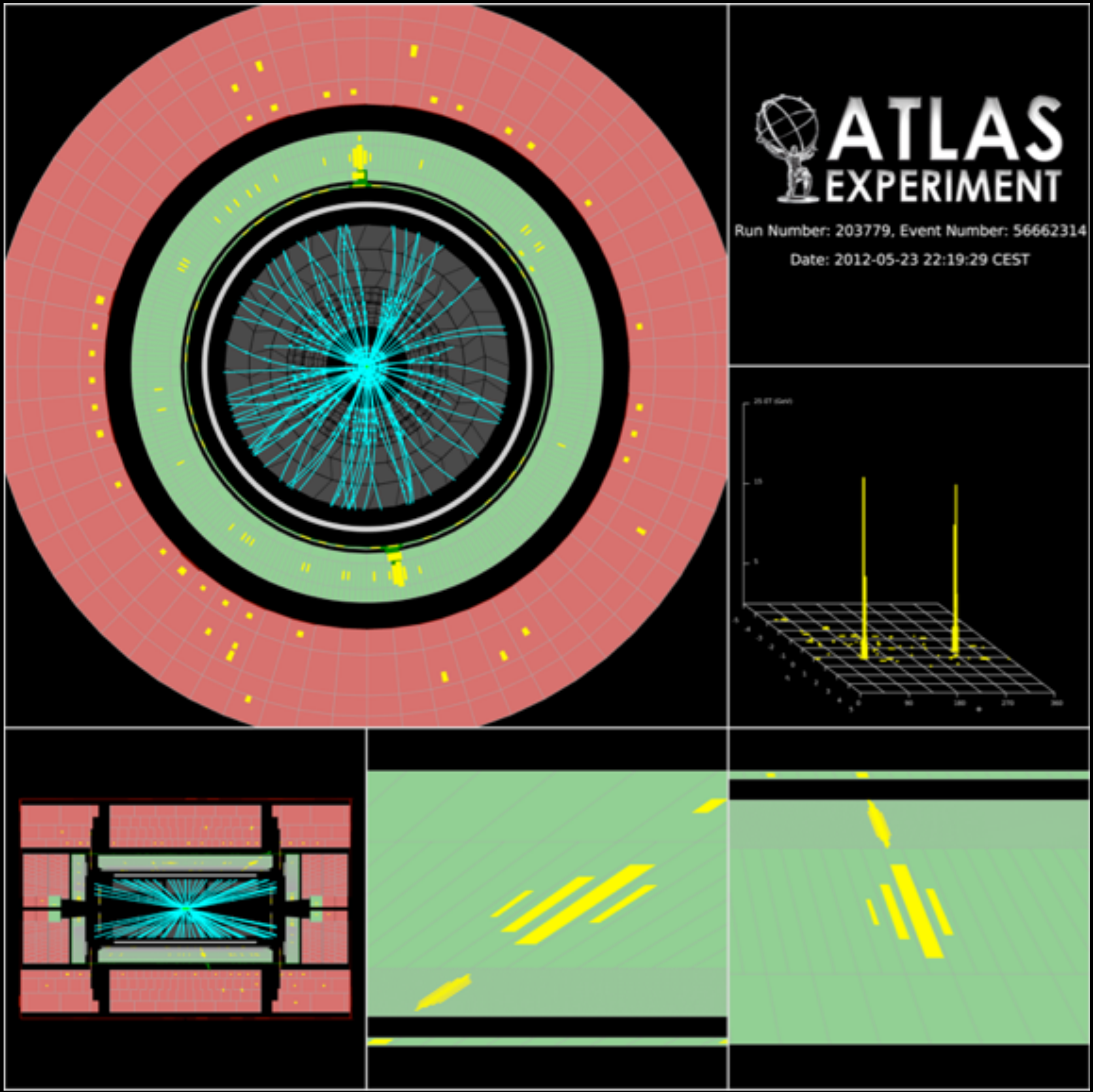




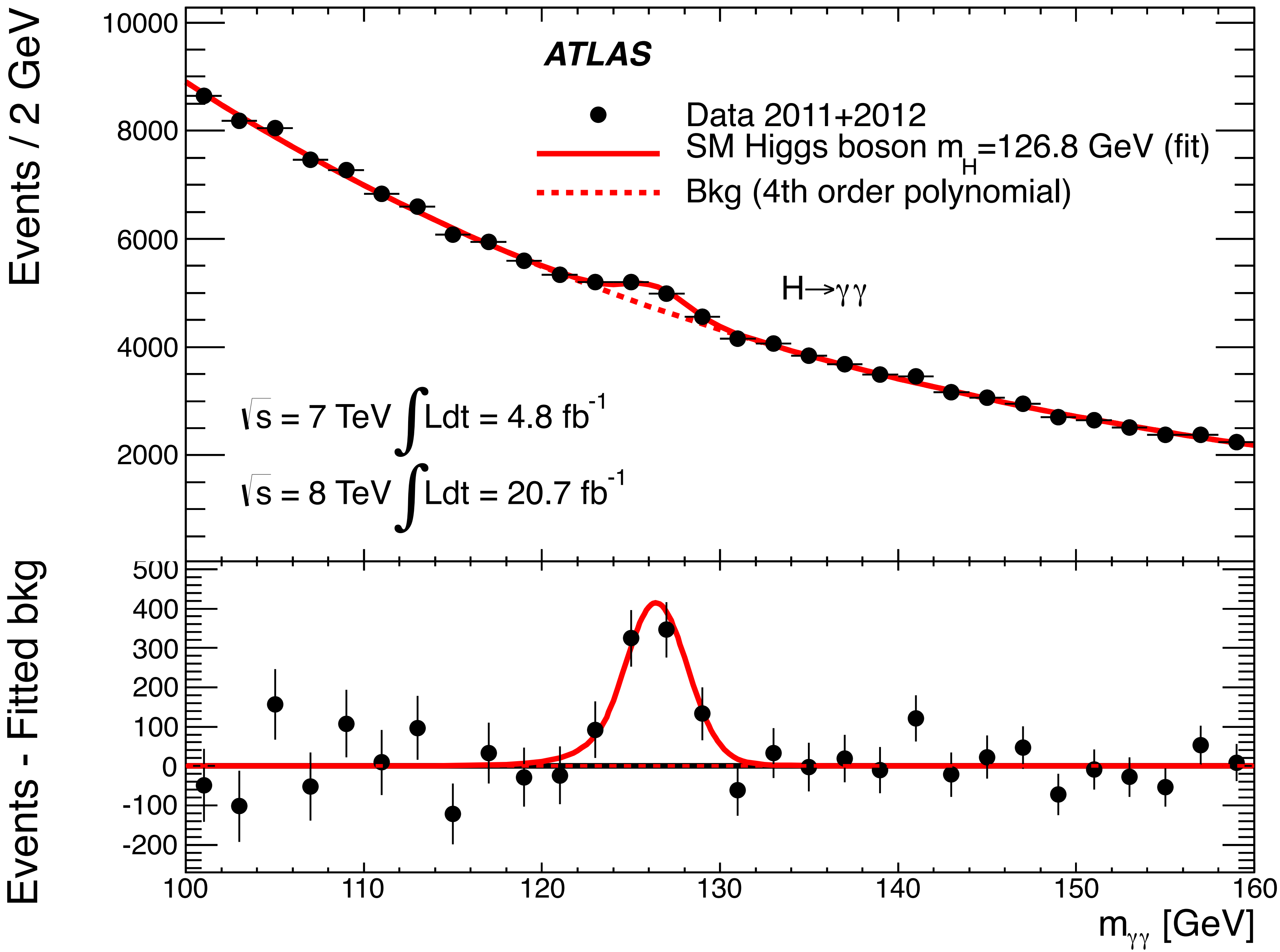


P. Baldi, P. Sadowski, and D. Whiteson [**arXiv:1402.4735**] GPU-accelerated Theano and Pylearn2 https://github.com/uci-igb/higgs-susy.

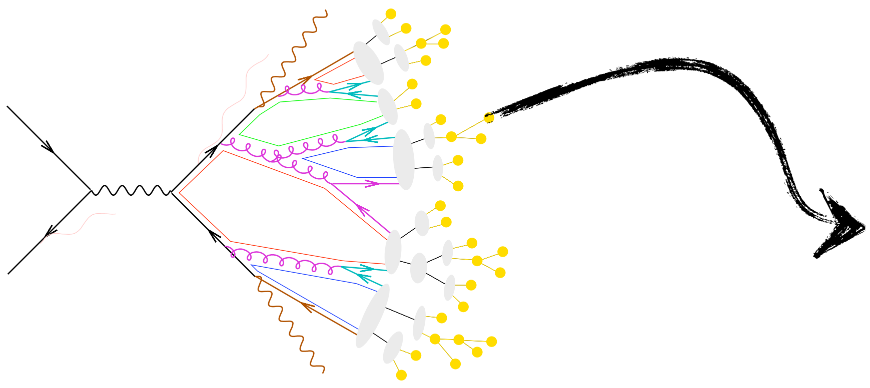
$H \to \gamma\gamma$

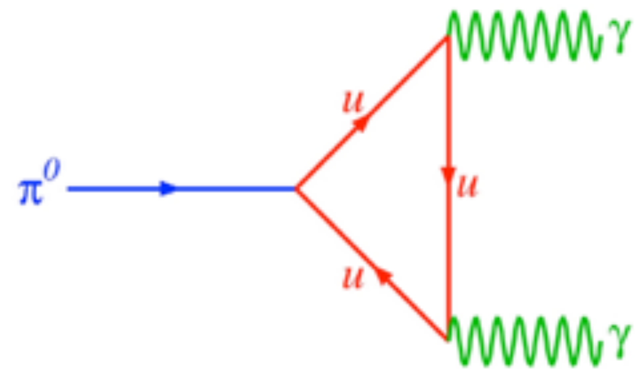


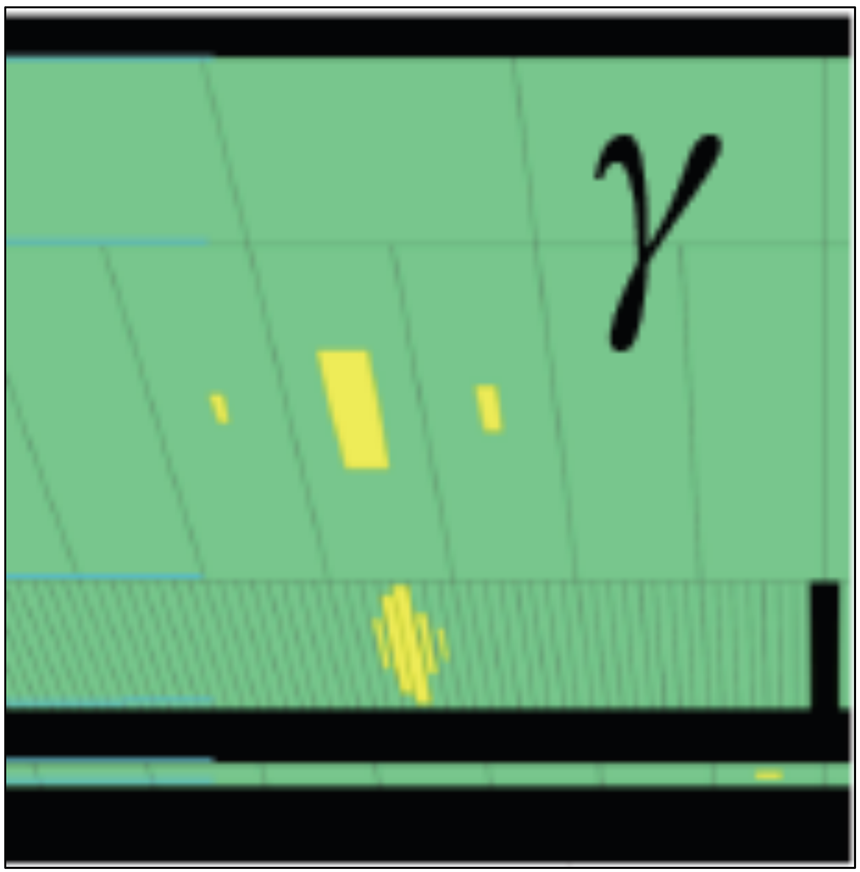
The observation in the 2 photon channel

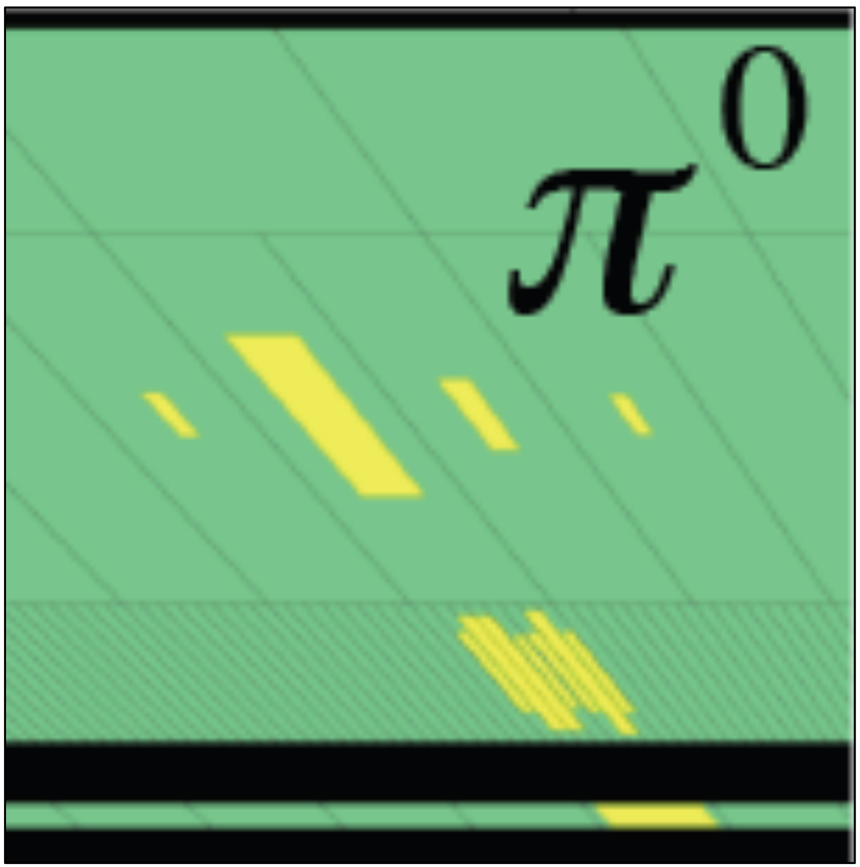


Particle Identification



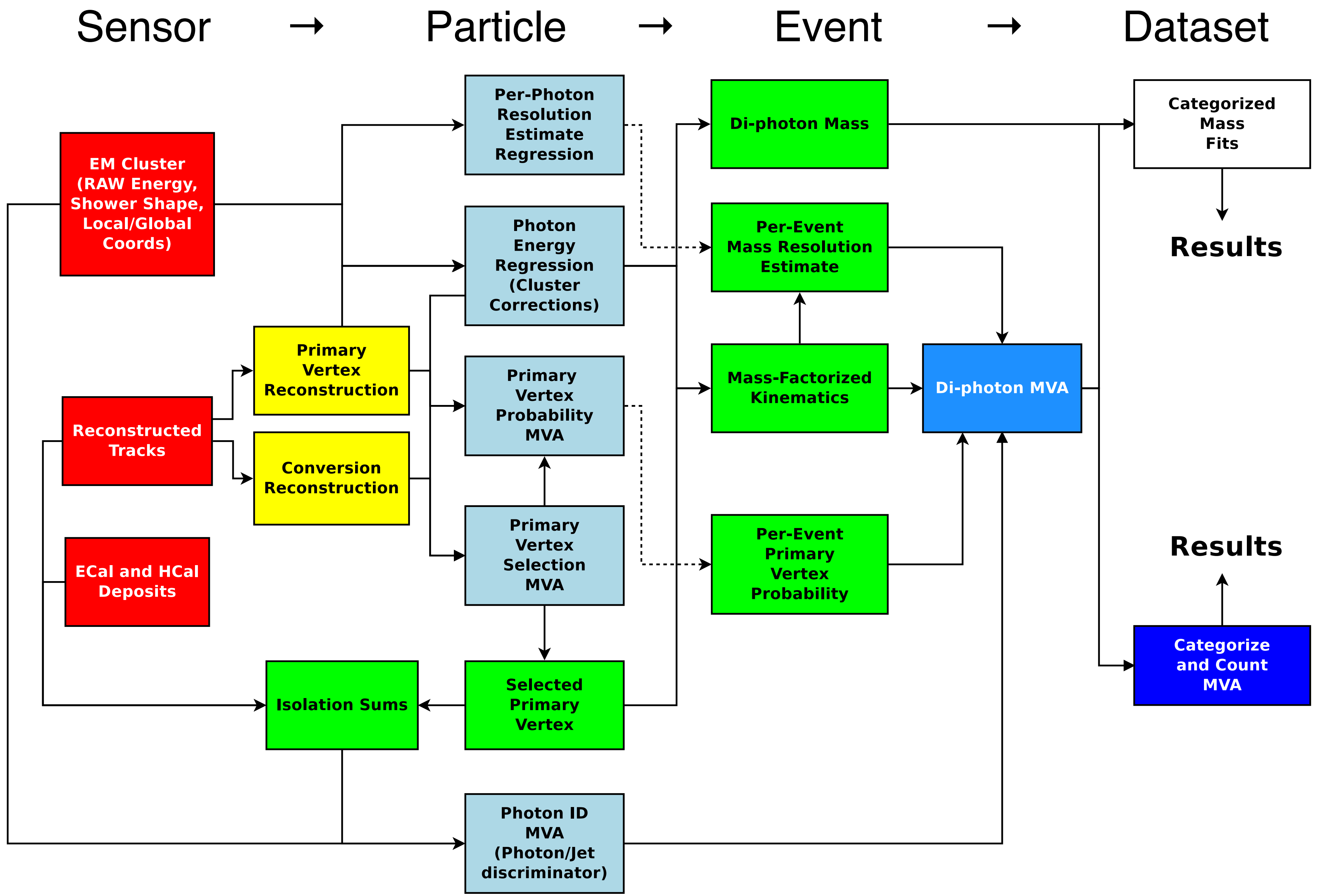






Machine Learning in H→γγ

Center for Cosmology and Particle Physics



***MVA = BDT implemented in TMVA**
(Deep networks being used for particle identification)

Reaching out to the ML community

Higgs challenge

Higgs Boson Machine Learning Challenge

2 months to go

Monday, May 12, 2014 — $13,000 • 837 teams — Monday, September 15, 2014

Dashboard

- Home
  - Data
  - Make a submission
- Information
  - Description
  - Evaluation
  - Rules
  - Prizes
  - About the Sponsors
  - Timeline

Competition Details » Get the Data » Make a submission

Evaluation

The evaluation metric is the *approximate median significance* (AMS):

$$\text{AMS} = \sqrt{2\left((s+b+b_r)\log\left(1+\frac{s}{b+b_r}\right)-s\right)}$$

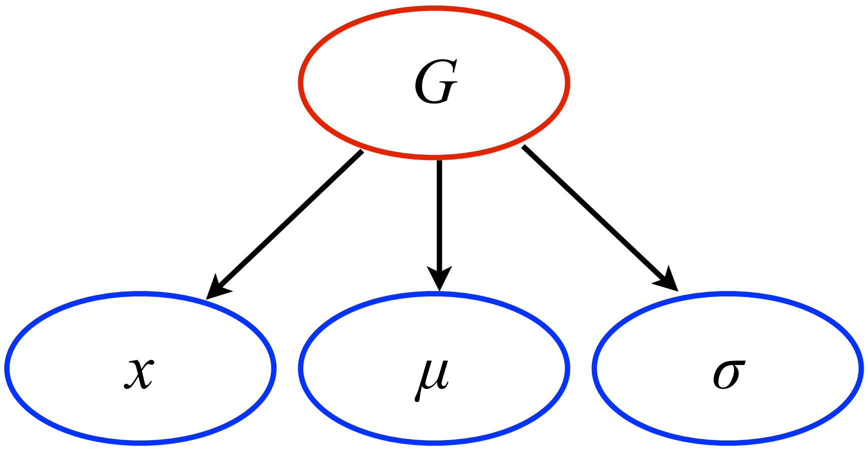
1	↑1	Gábor Melis *	3.80573	32	Thu, 26 Jun 2014 06:14:34 (-0.2h)
359	↓49	JeJe	3.25012	4	Sat, 21 Jun 2014 01:11:13
		simple TMVA boosted trees	3.24954		
360	↓49	Xiaohu SUN	3.24954	3	Tue, 03 Jun 2014 13:14:47

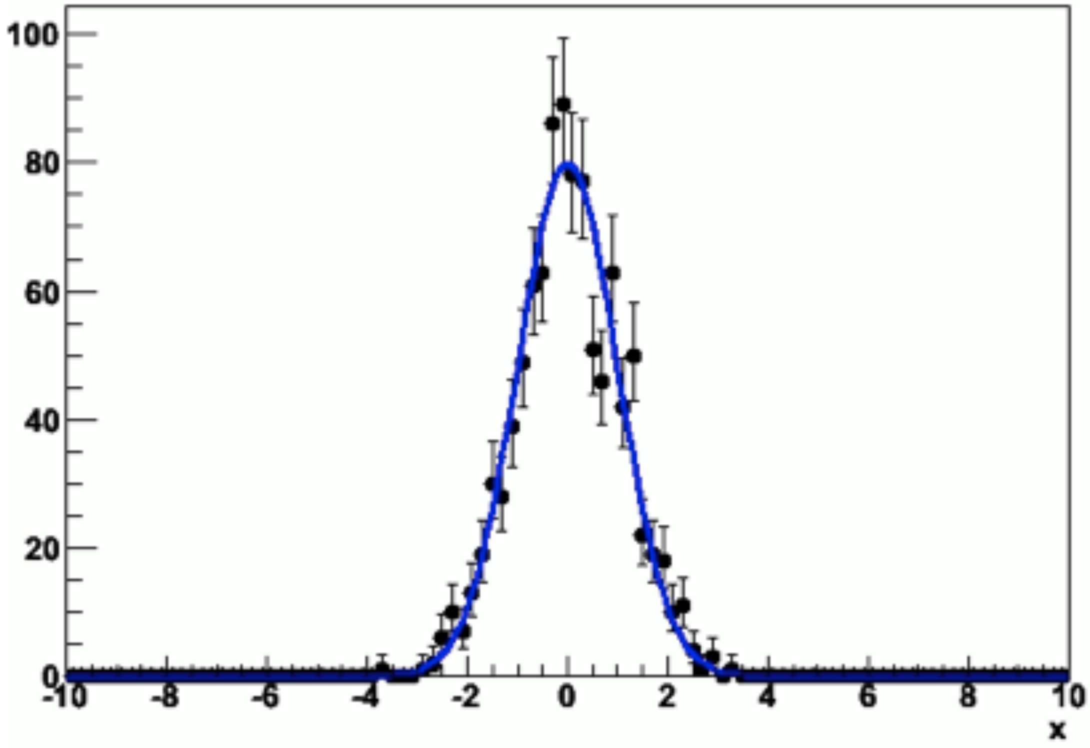
Statistical Modeling for Higgs Discovery

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



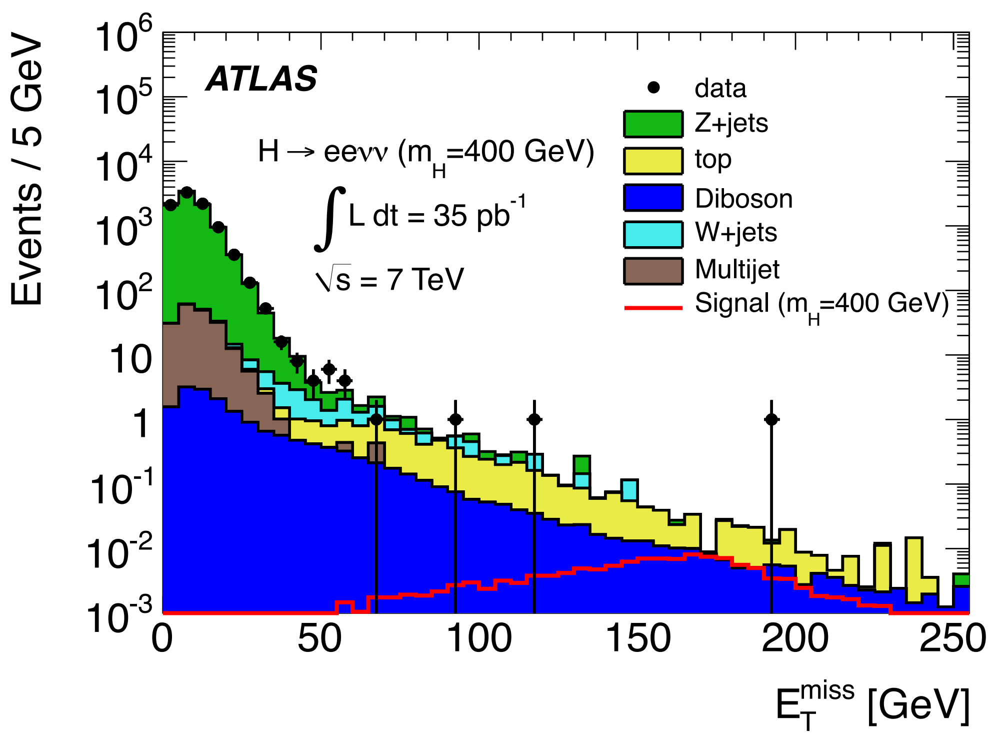


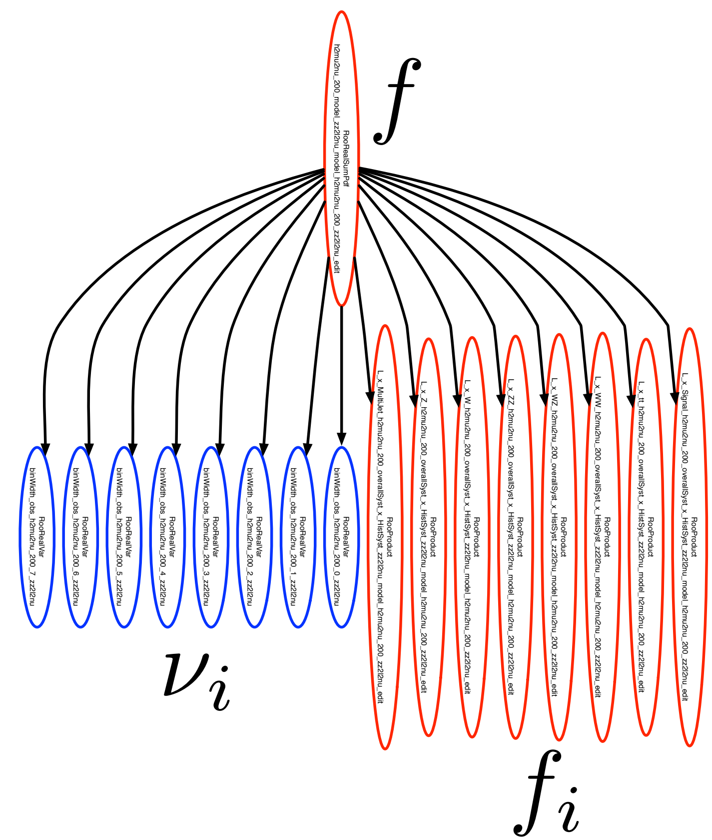
Clearly related to Graphical Models, but not the focus here.

Mixture model

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

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

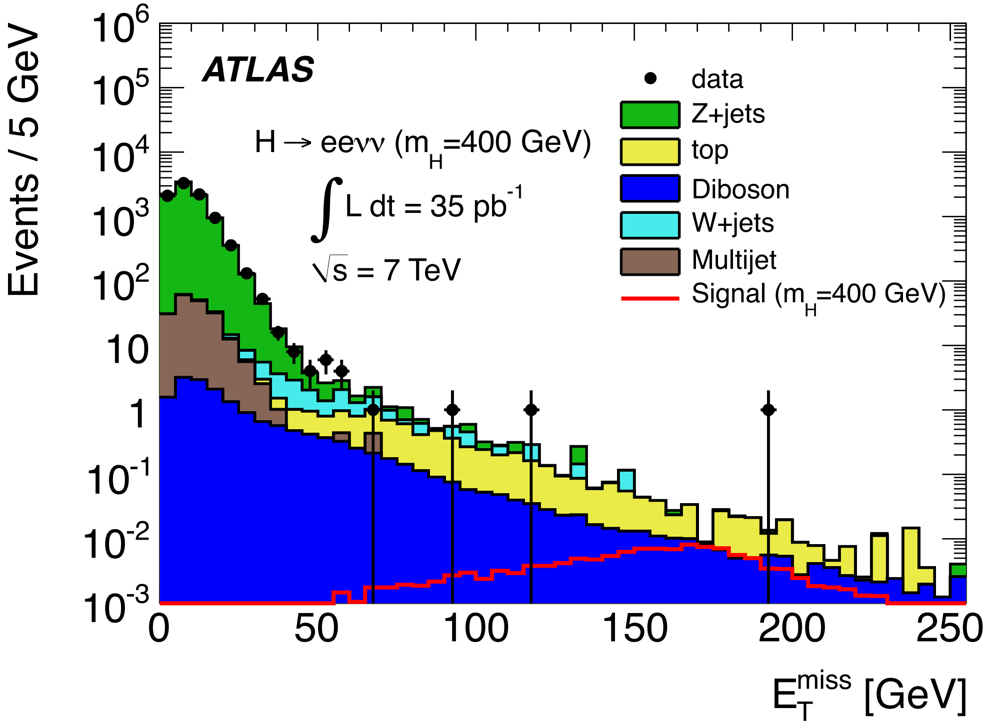


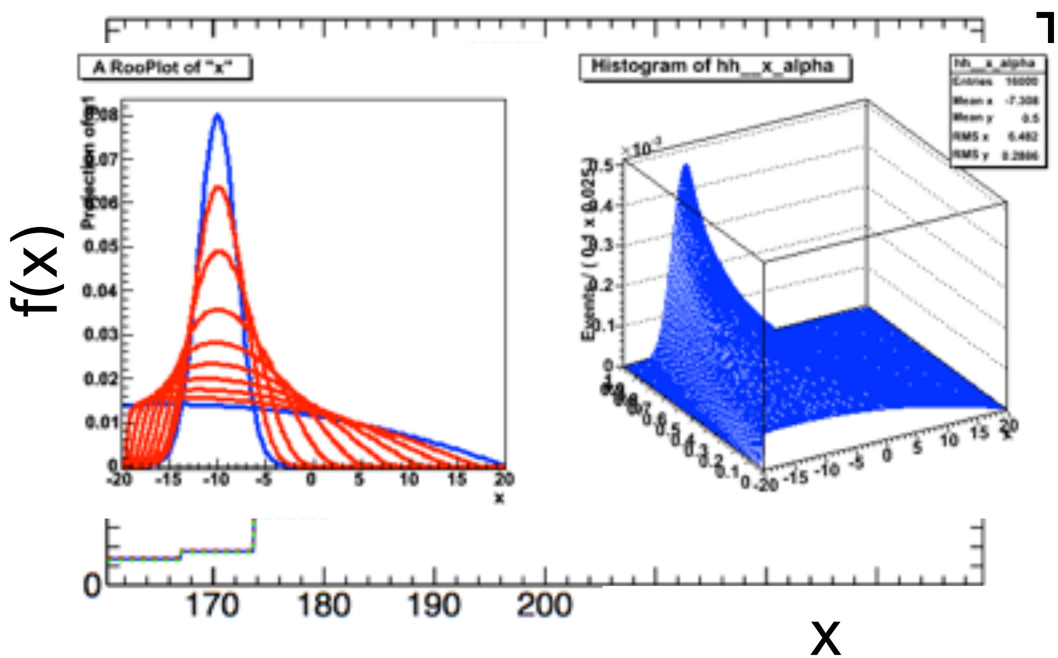


Incorporating Systematic Effects

Tabulate effect of individual variations of sources of systematic uncertainty

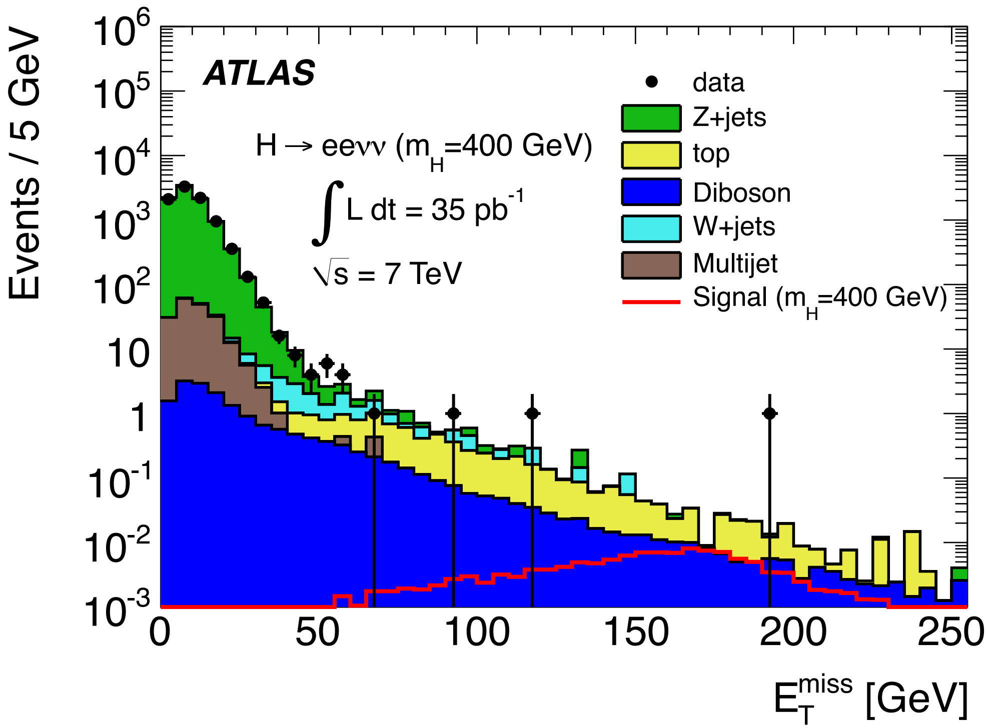
- typically one at a time evaluated at nominal and "± 1 σ"
- use some form of interpolation to parametrize p^{th} variation in terms of **nuisance parameter** α_p

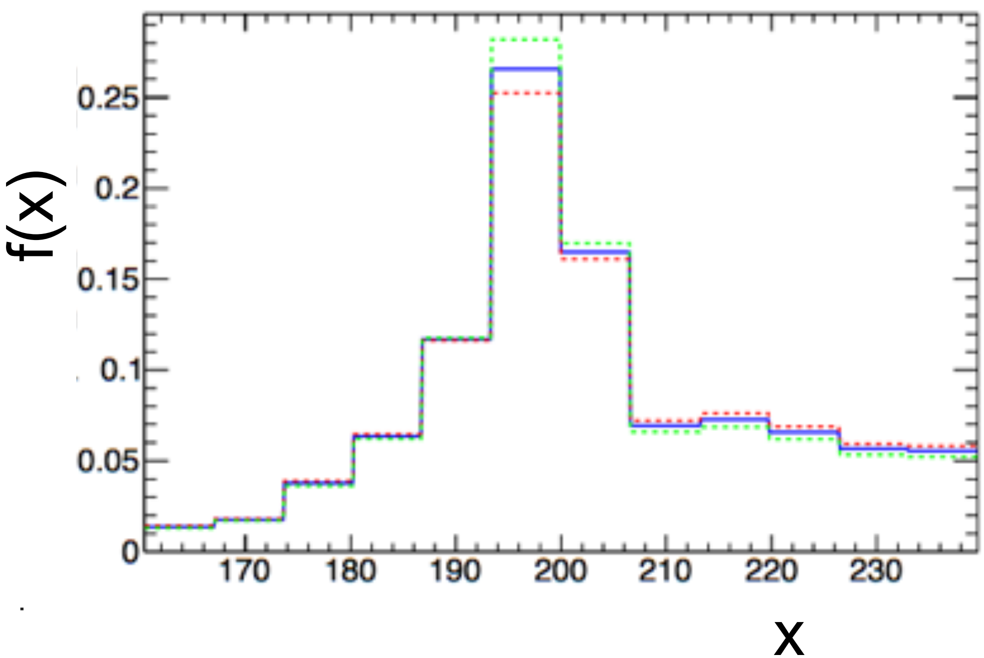


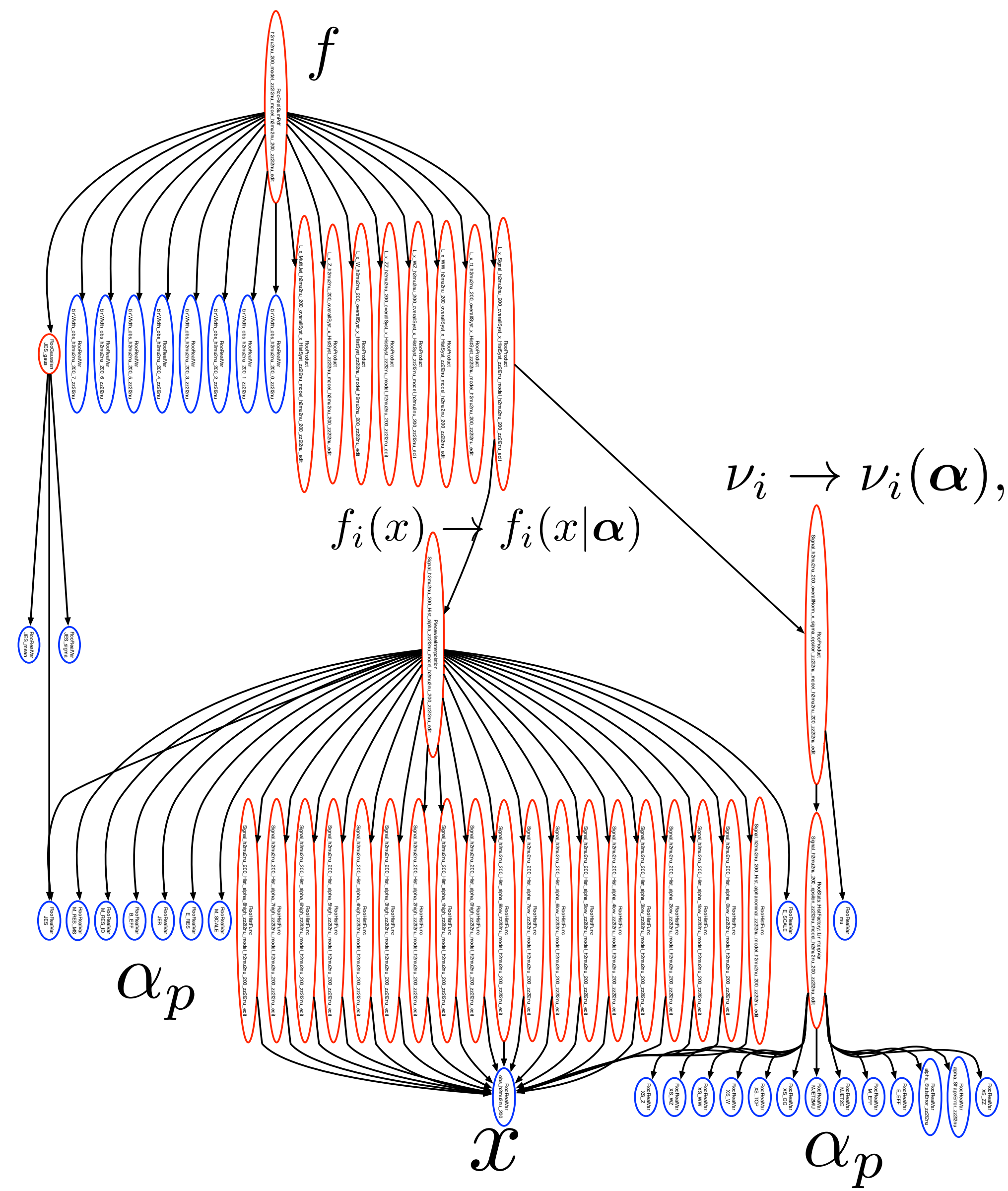


$$\mathbf{f}(\mathcal{D}|\boldsymbol{\alpha}) = \mathrm{Pois}(n|\nu(\boldsymbol{\alpha})) \prod_{e=1}^{n} f(x_e|\boldsymbol{\alpha})$$

Visualizing the model for one dataset

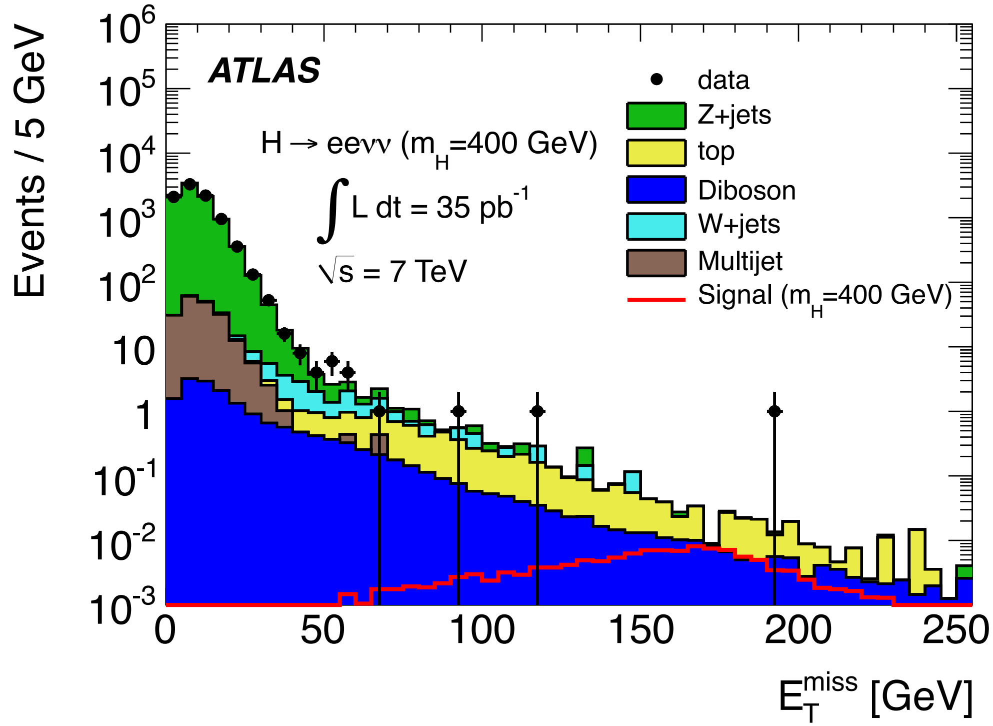


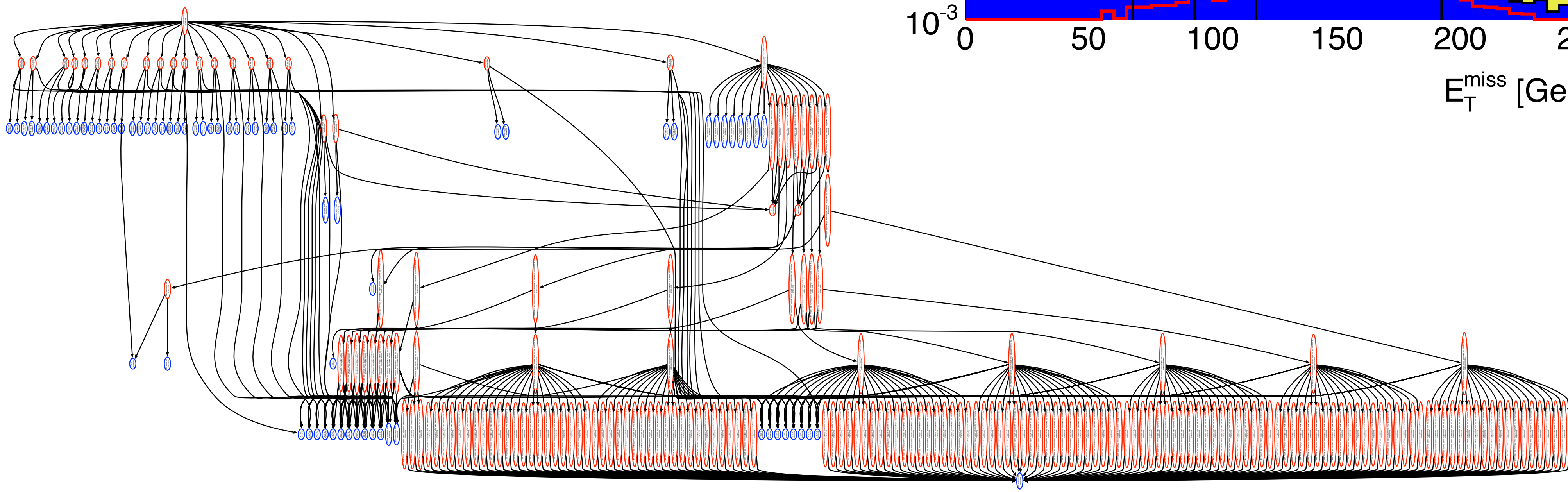




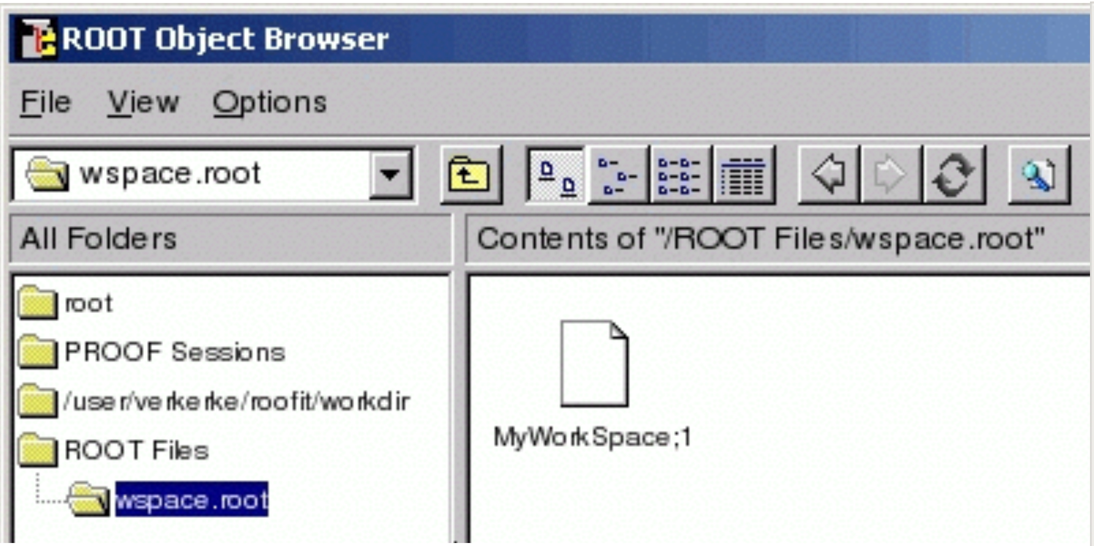
Visualizing the model for one dataset

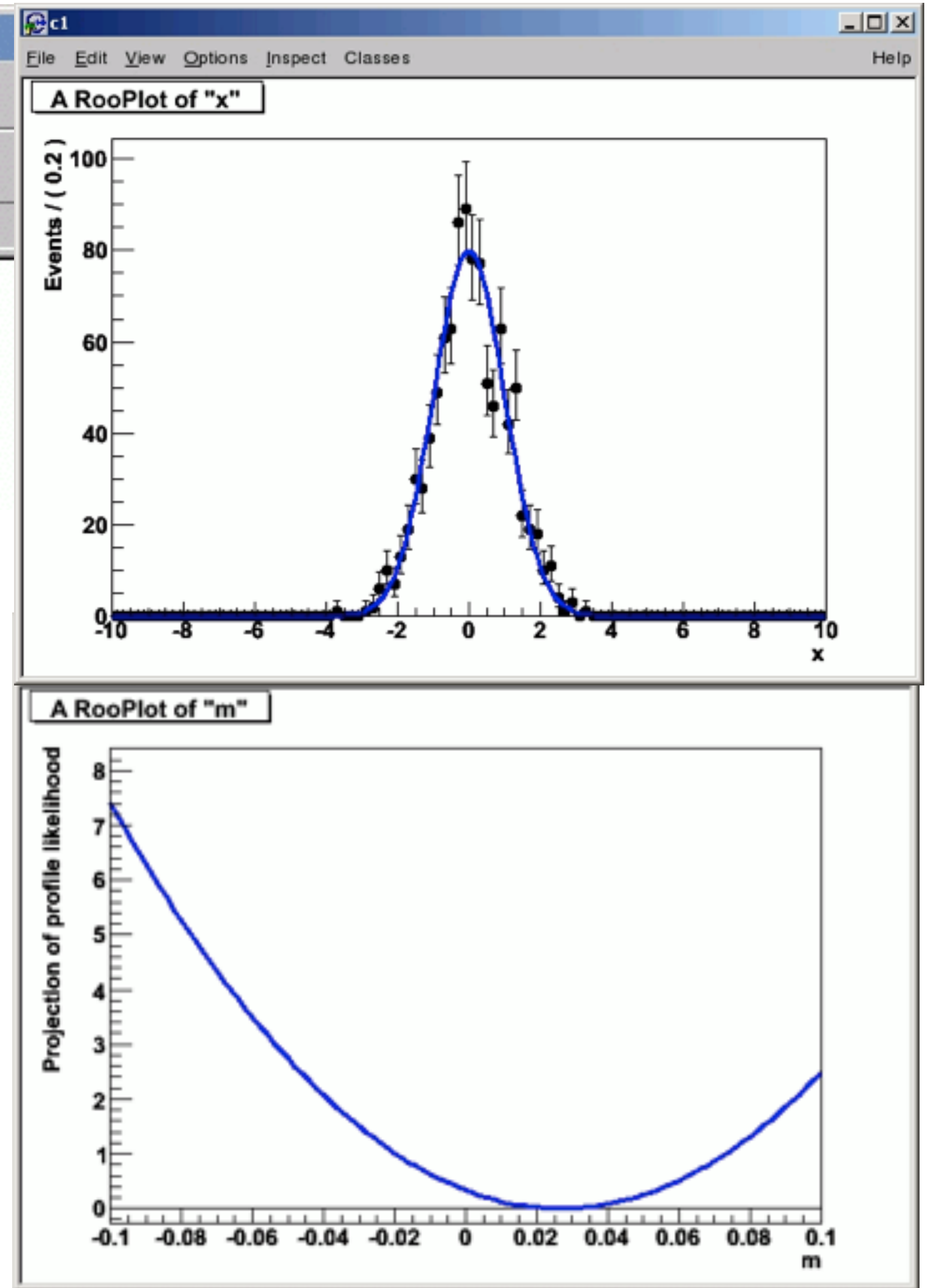
After parametrizing each component of the mixture model, the pdf for a single channel might look like this





Digital Publishing Statistical Models

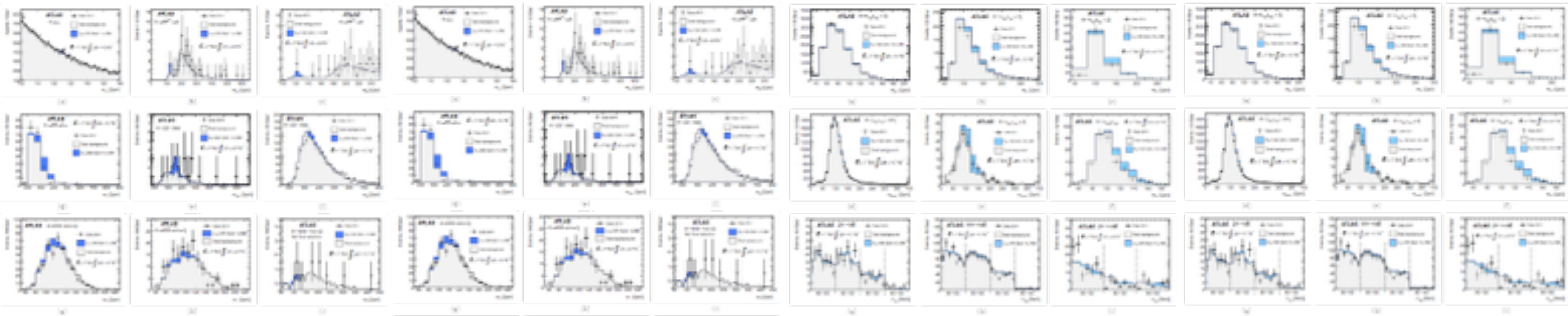


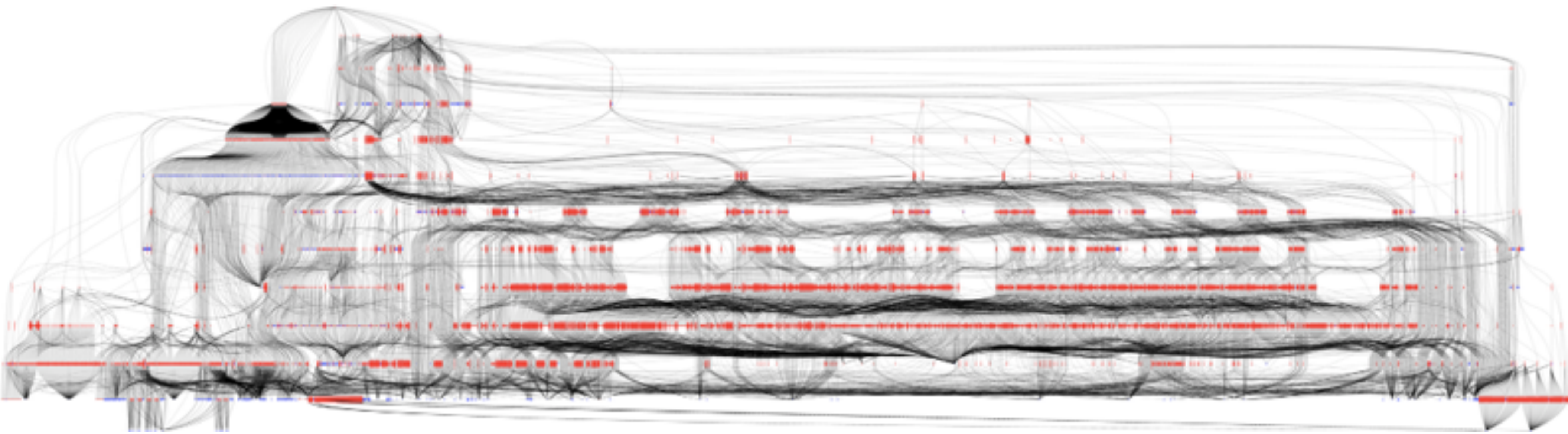


RooFit’s Workspace now provides the ability to save in a file the full likelihood model, any priors you might want, and the data necessary to reproduce likelihood function.

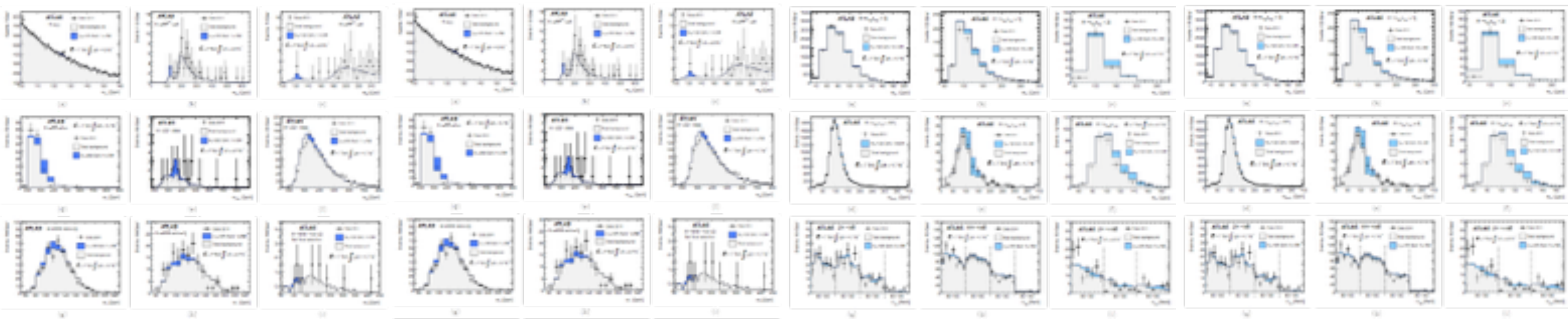
Gives flexibility in later statistical analysis (frequentist vs. bayesian) and handles for detailed meta-analysis

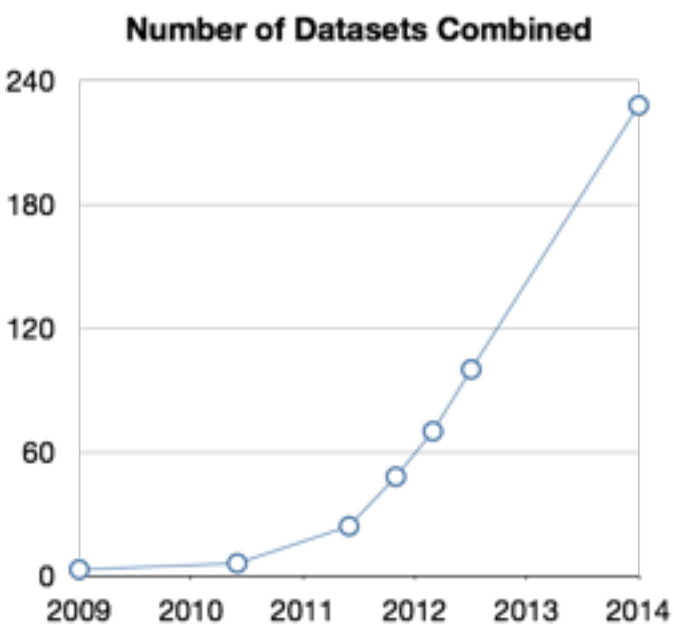
Collaborative Statistical Modeling

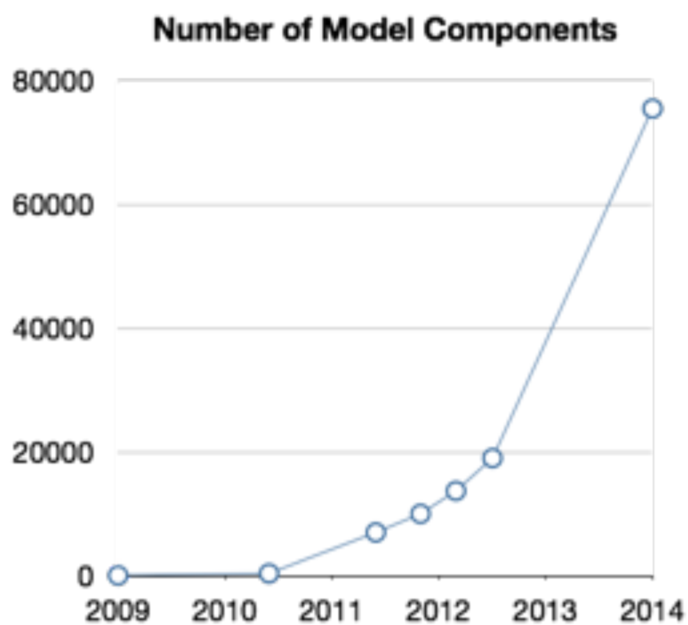


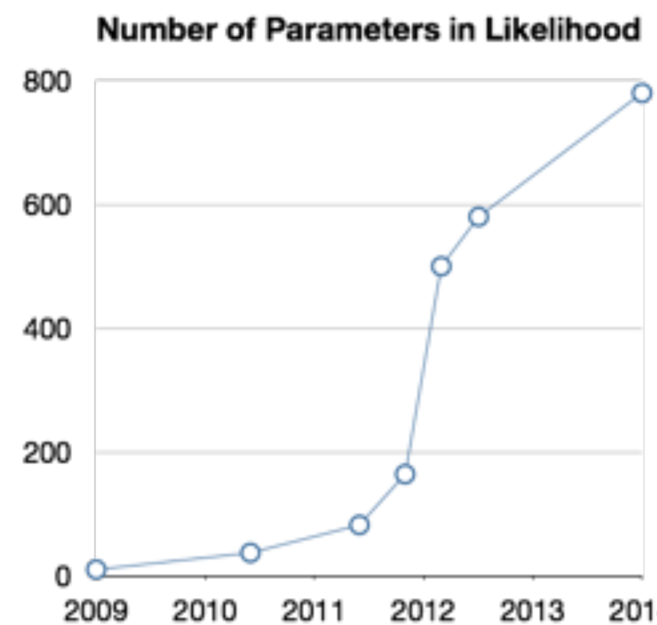


Collaborative Statistical Modeling



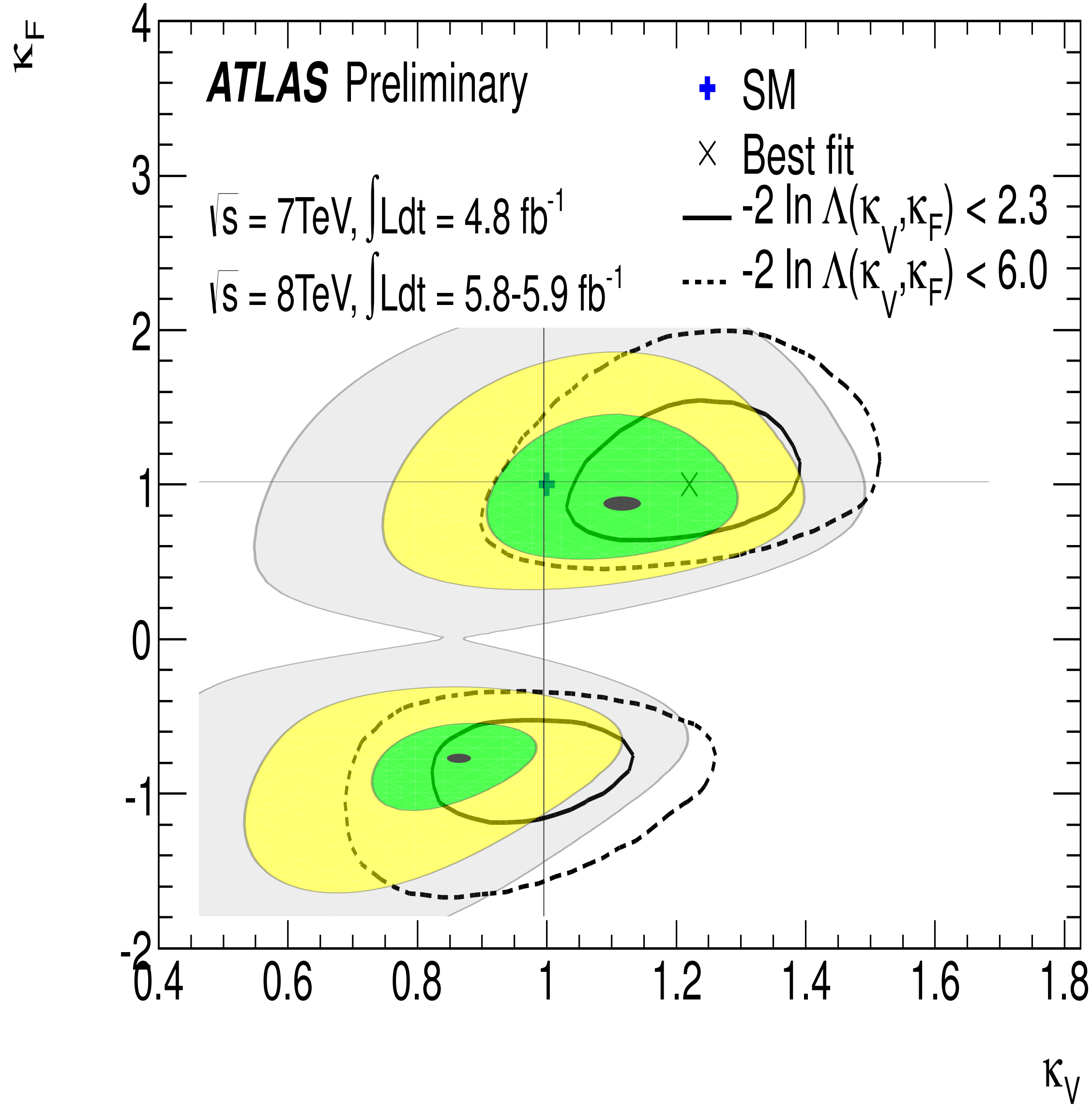






REPRODUCIBILITY PROBLEM

Not possible for others to reproduce results from paper.



PUBLISHING LIKELIHOODS

Information | References (121) | Citations (128) | Files | Plots | HepData

Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

ATLAS Collaboration (Georges Aad (Freiburg U.) *et al.*) *Show all 2923 authors*

Jul 4, 2013 - 32 pages

Phys.Lett. B726 (2013) 88-119
(2013)
DOI: 10.1016/j.physletb.2013.08.010

Note, data record itself has 4 citation

Information | Citations (4) | Files

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

ATLAS Collaboration (Aad, Georges (Freiburg U.) [...]) *Show all 2923 authors*

Cite as: ATLAS Collaboration (2013) HepData, http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

figshare
@figshare
Following

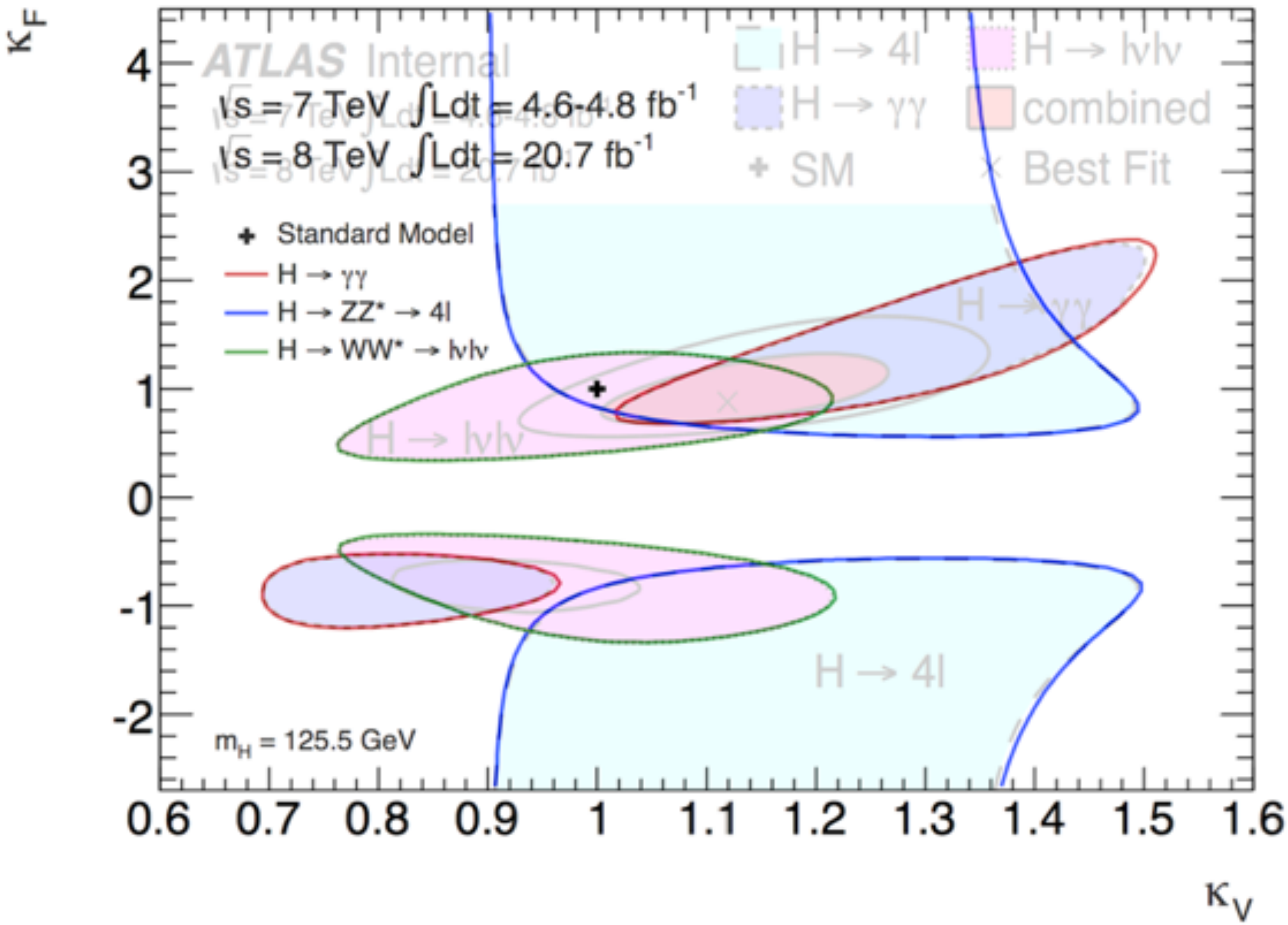
The Higgs Boson data is definitely the jewel in the #DataCite crown. Hopefully the first of many!

Reply Retweeted Favorite More

RETWEETS 7   FAVORITES 2

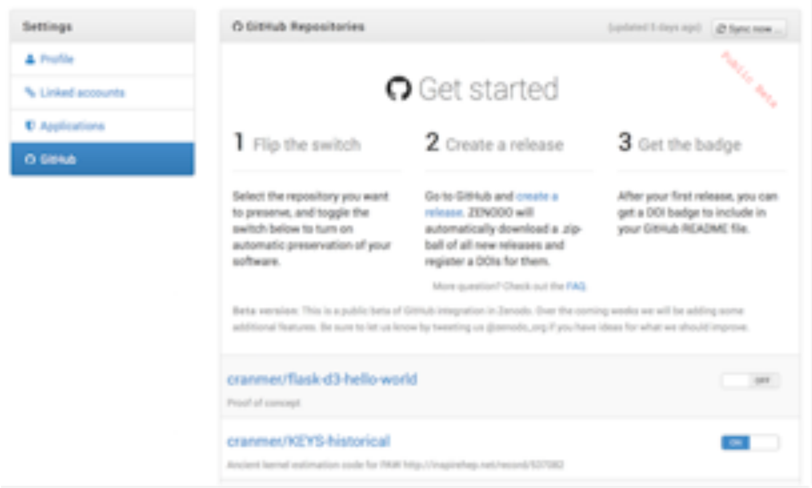
PUBLISHING LIKELIHOODS

Reproducing derived results from original paper!



CODE AS A RESEARCH PRODUCT

GitHub → Zenodo → INSPIRE



Mathematica → figshare → INSPIRE

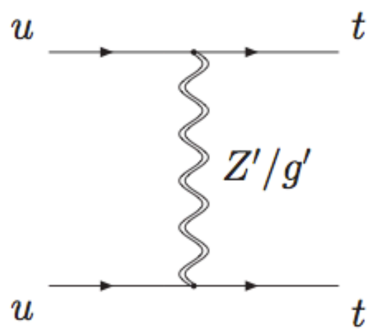
Supplementary Material for "A Novel Approach to Higgs Coupling Measurements"

(2013) *figshare.*

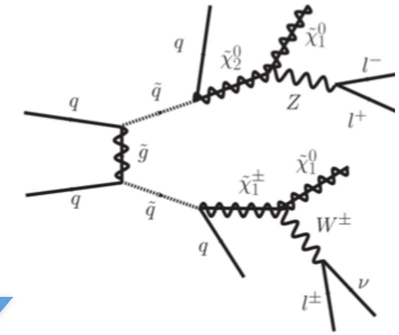
Impactstory

highly discussed by scholars	97 - 100 percentile of datasets published in 2013	3 figshare shares
highly viewed by scholars	97 - 100 percentile of datasets published in 2013	202 figshare views
highly viewed by scholars	97 - 100 percentile of datasets published in 2013	9 figshare downloads
highly discussed by public	97 - 100 percentile of datasets published in 2013	10 tweets

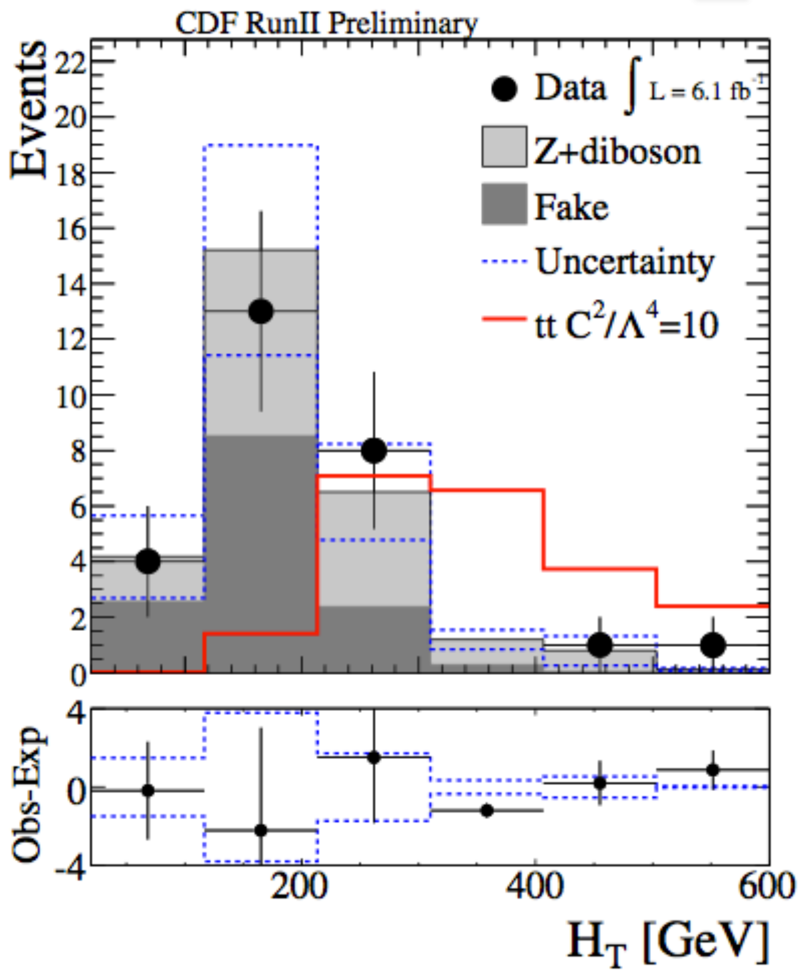
RECASTING



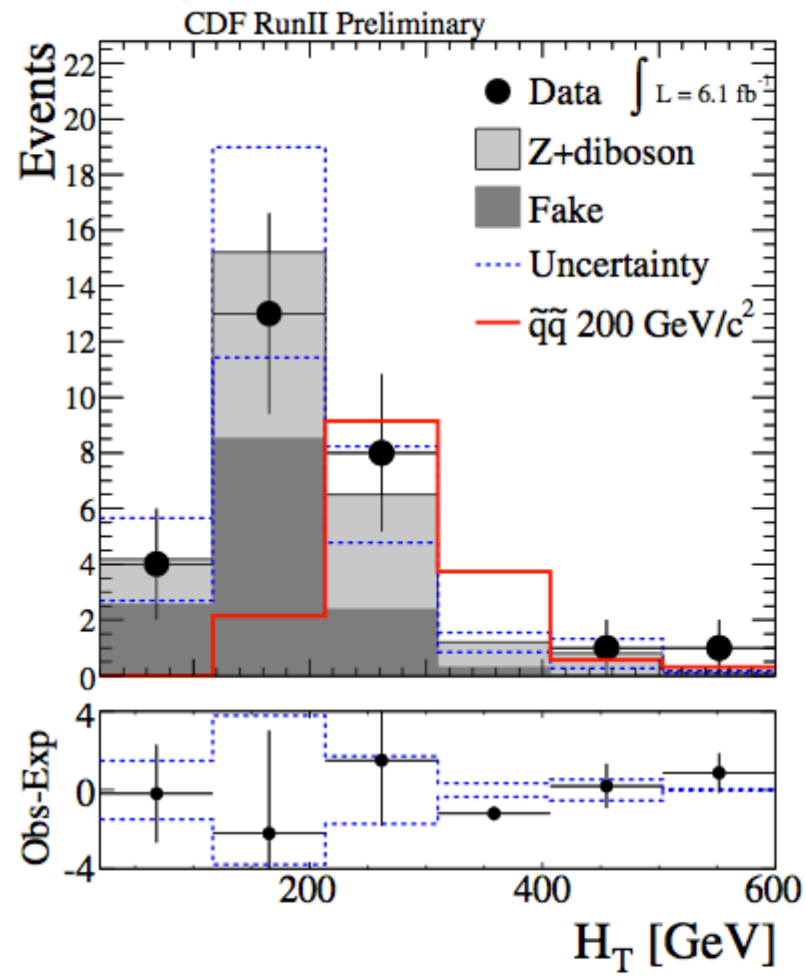
u
t
Z'/g'
u
t



q
q
q
q
W±
Z
ν
l±
l+
l−



CDF RunII Preliminary
Events
Data ∫ L = 6.1 fb⁻¹
Z+diboson
Fake
Uncertainty
tt C²/Λ⁴=10
Obs-Exp
H_T [GeV]



CDF RunII Preliminary
Events
Data ∫ L = 6.1 fb⁻¹
Z+diboson
Fake
Uncertainty
q̃q̃ 200 GeV/c²
Obs-Exp
H_T [GeV]

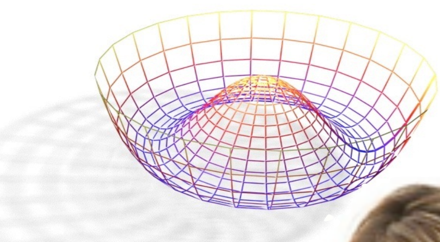
THEORY

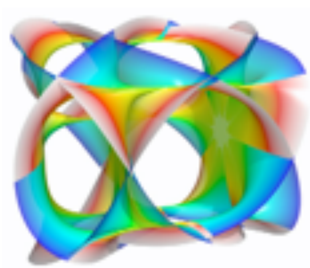
SERVICE

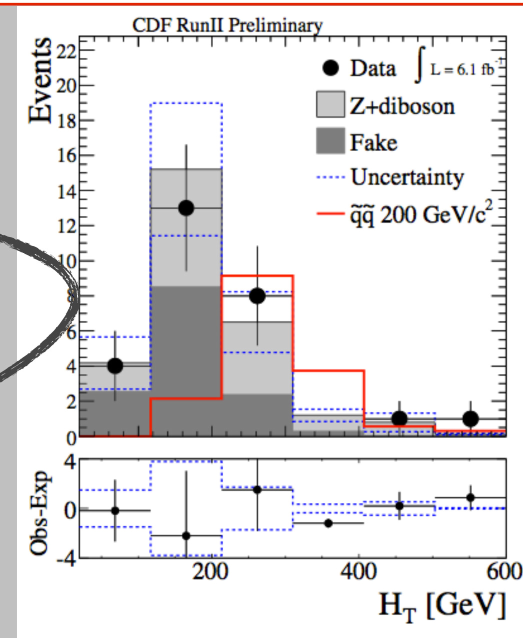
$$
\begin{aligned}
\mathcal{L}_{SM} = & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^{a}_{\mu\nu}G^{\mu\nu}_{a}}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})L + \bar{R}\gamma^{\mu}(i\partial_{\mu} - \frac{1}{2}g'YB_{\mu})R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
& + \underbrace{\frac{1}{2}\left|(i\partial_{\mu} - \frac{1}{2}g\tau\cdot\mathbf{W}_{\mu} - \frac{1}{2}g'YB_{\mu})\,\phi\right|^{2} - V(\phi)}_{W^{\pm},Z,\gamma,\text{and Higgs masses and couplings}} \\
& + \underbrace{g''(\bar{q}\gamma^{\mu}T_{a}q)\,G^{a}_{\mu}}_{\text{interactions between quarks and gluons}} + \underbrace{(G_{1}\bar{L}\phi R + G_{2}\bar{L}\phi_{c}R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
\end{aligned}
$$

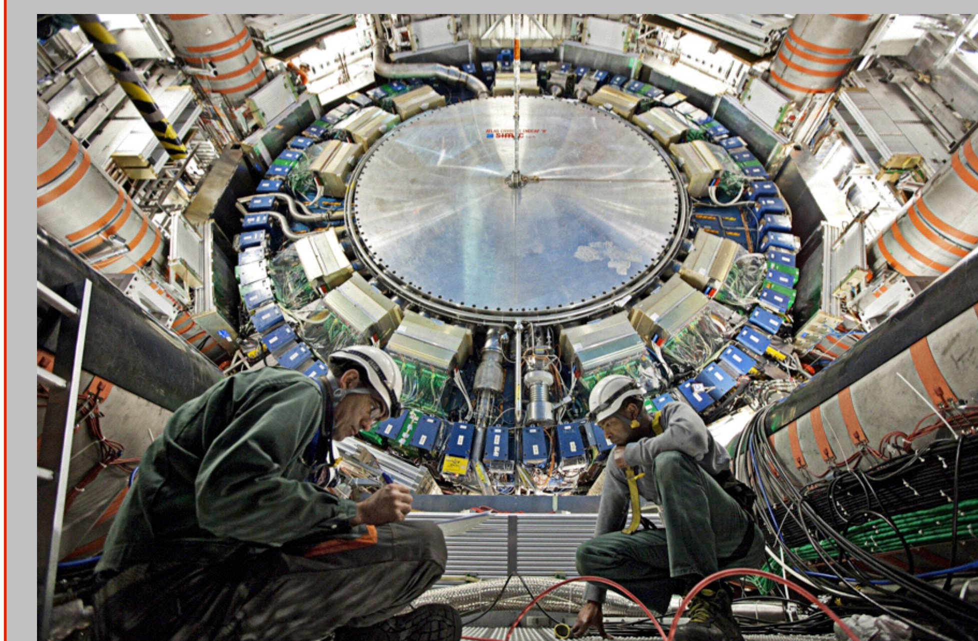
Q

A









Review of Challenges
and
Possible Research Topics

Challenges & possible research areas

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, automatic differentiation, distributed processing, …
- better optimization & sampling algorithms
- optimal statistical procedures subject to computational constraints (link)

Interpolation of distributions based on simulated samples with different parameter settings a weak point

- experimental design, response surface interpolation, Gaussian processes, …
- complication: samples often not statistically independent

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

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
- **progress:** movement towards open access (link to policy)

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 can do something smarter

Particle physics is a unique arena for data science

- well posed questions in an extreme setting
- lots of data, complicated sensor environment, strong theoretical basis

Congratulations and best wishes to Paris-Saclay Center for Data Science