Data science in particle physics

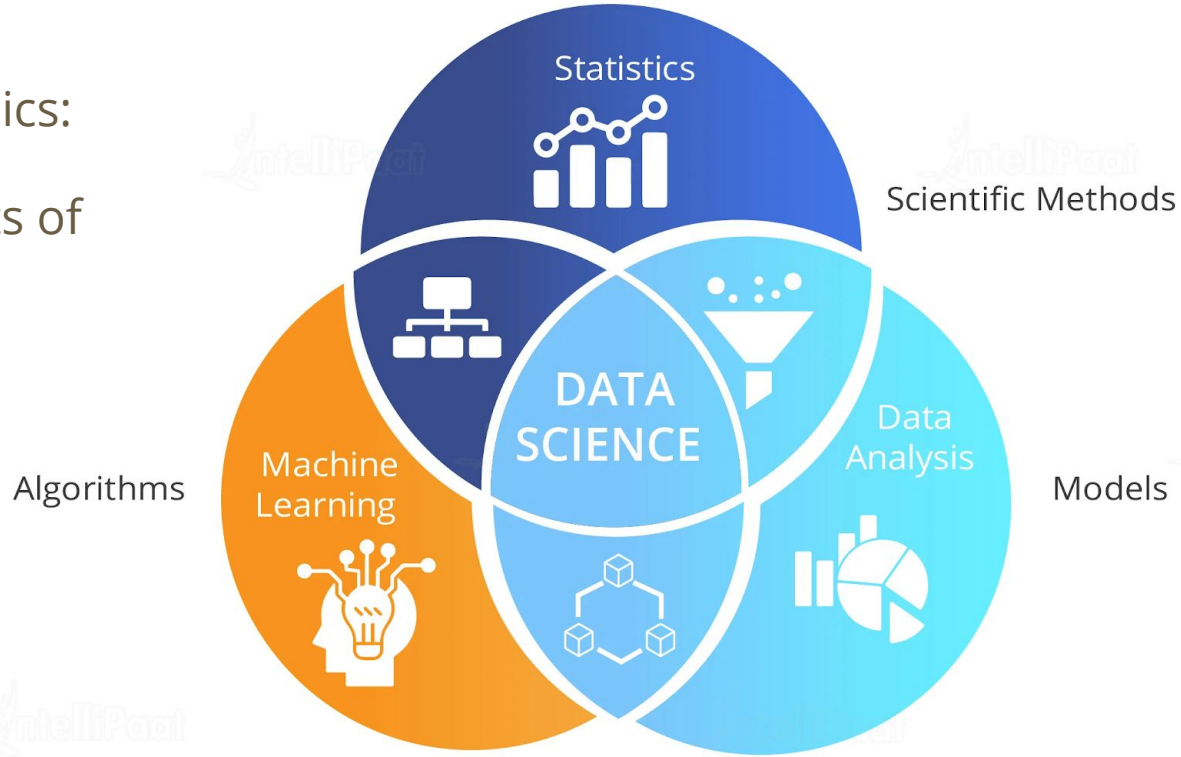
Lydia Brenner

Trieste, 16-18 November 2020

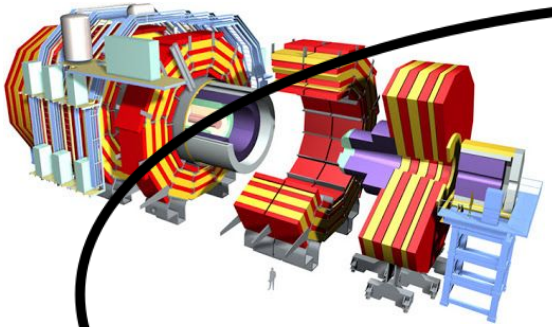
Introduction

Challenge in particle physics:

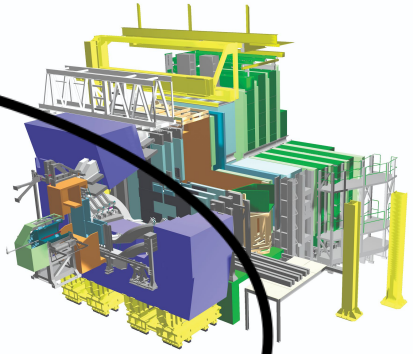
Handling of large amounts of data



The LHC and the large experiments



CMS



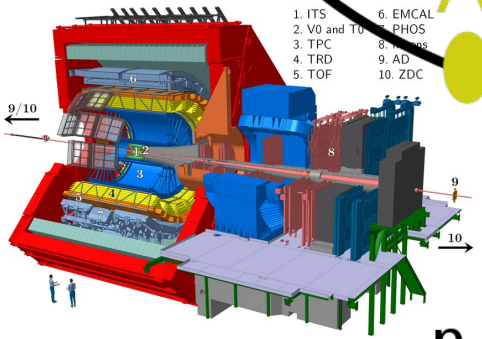
LHC

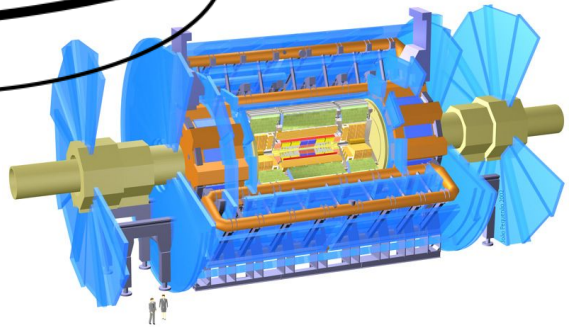
LHCb

ALICE

SPS

ATLAS



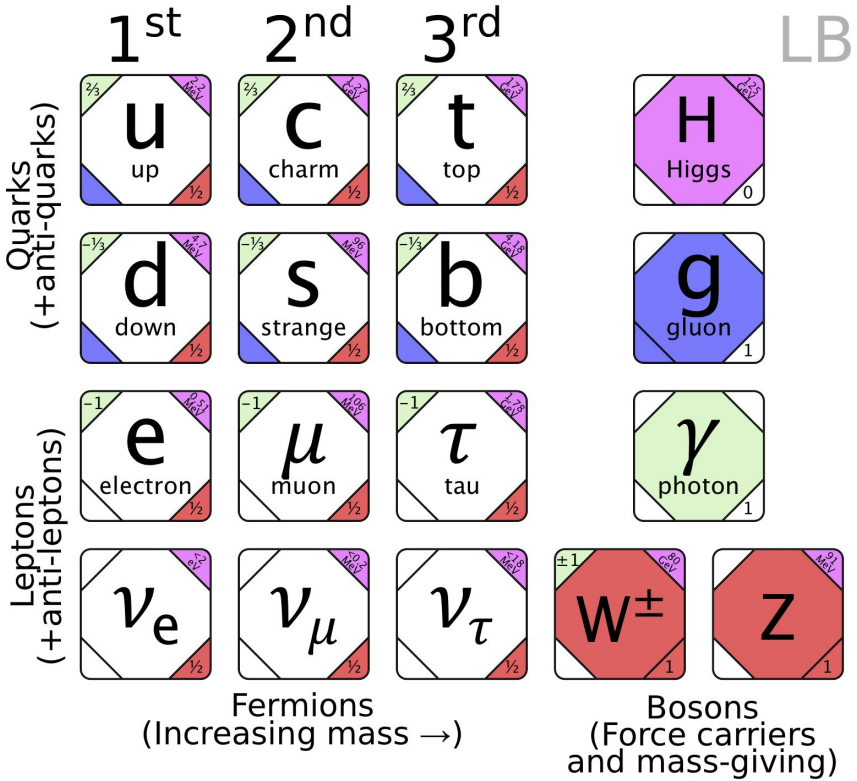


p

Pb

PS

Big data and the standard model



Some examples

Particle	Expected number of events in ATLAS in completed Run 2
Higgs Boson	7.7 million
Top quark	275 million
Z Boson	2.8 billion
W Boson	12 billion

Big data processing

The data flow from all four experiments for Run 2 is about 25 GB/s

- ALICE: 4 GB/s (Pb-Pb running)
- ATLAS: 800 MB/s
- CMS: 600 MB/s
- LHCb: 750 MB/s

Data selection in two stages

1) Hardware filter reduces 600 million events per second to 100 000 per second
2) Software filter uses algorithms to reduce to 100-200 events per second

End up with around 1050 MB/s of raw data

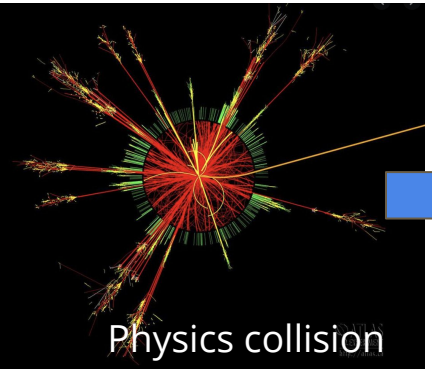
Grid computing

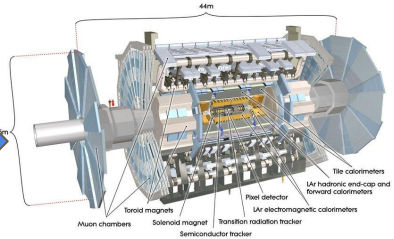
Using the Worldwide LHC Computing Grid (WLCG) to

- store,
- distribute
- process

More than 170 collaborating centres in 36 countries around the world, linked to CERN

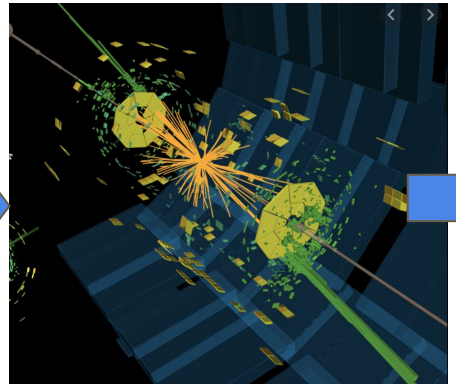
Measurement roadmap





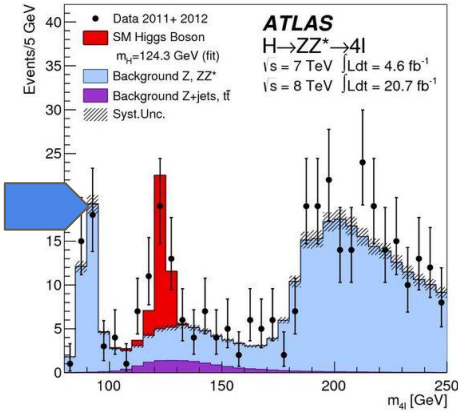
Detector interaction

Detector reconstruction

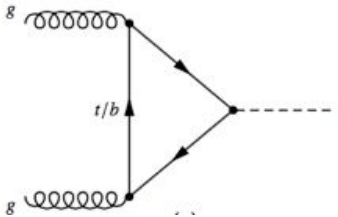


Event selection

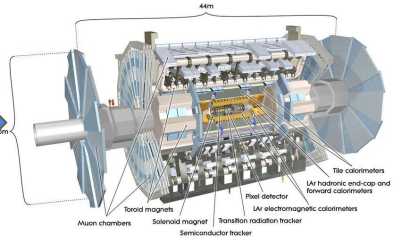
Analysis



Physics model



Detector simulation

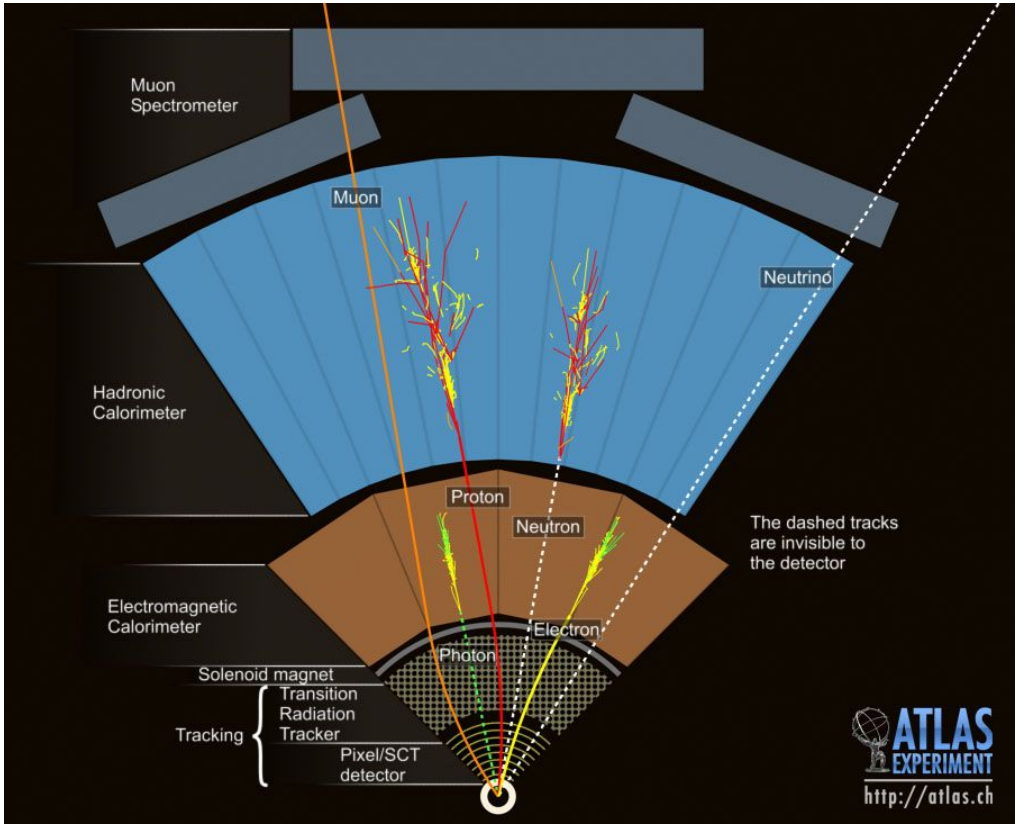


Particle identification

Event reconstruction based on

- Detector response
- Object classification
- Rejection of background processes

On real and simulated data

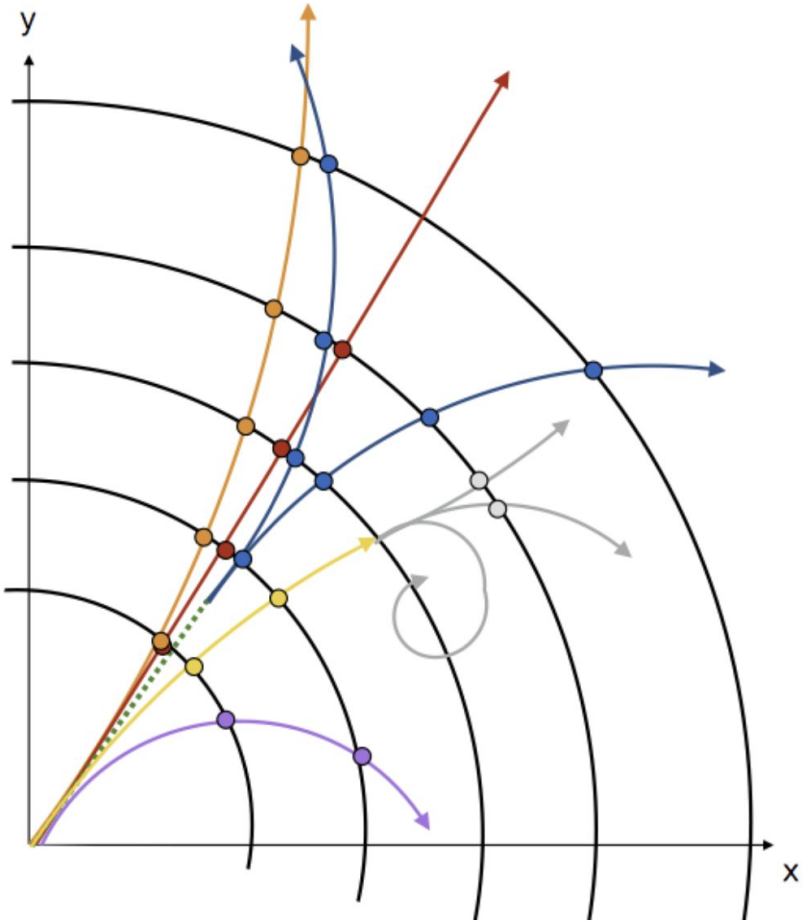


Detector response

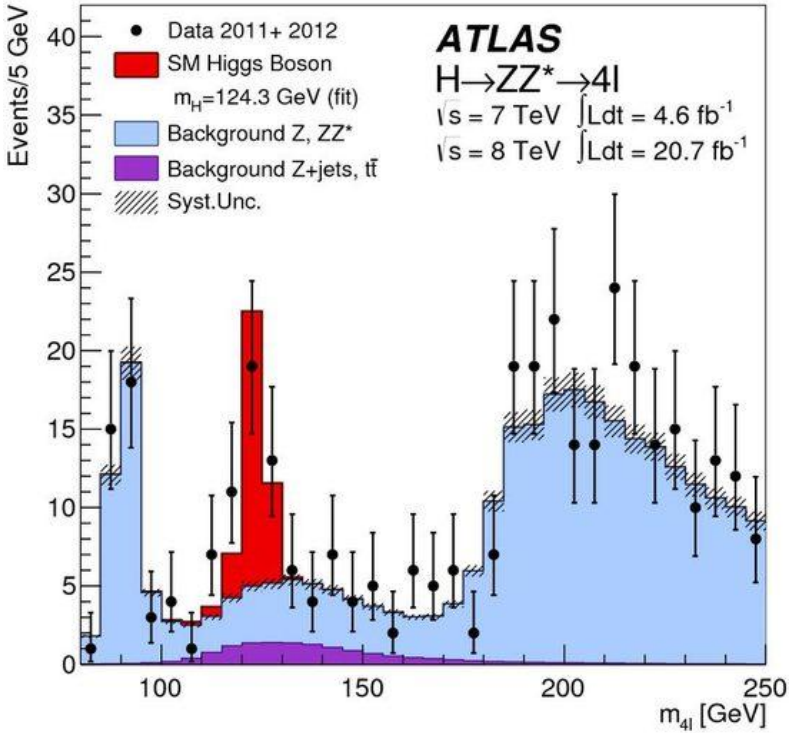
Event reconstruction considering insensitive areas in the detector due to

- Readout electronics
- Detector dead-time
- Detector efficiencies

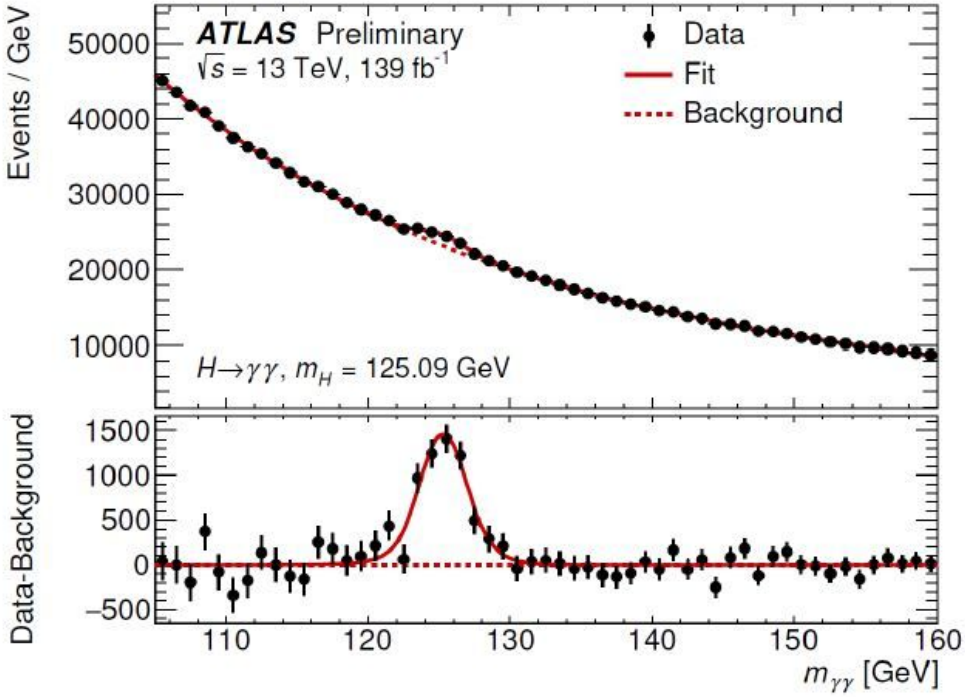
On real and simulated data



Background rejection



Remove predicted contribution in each bin

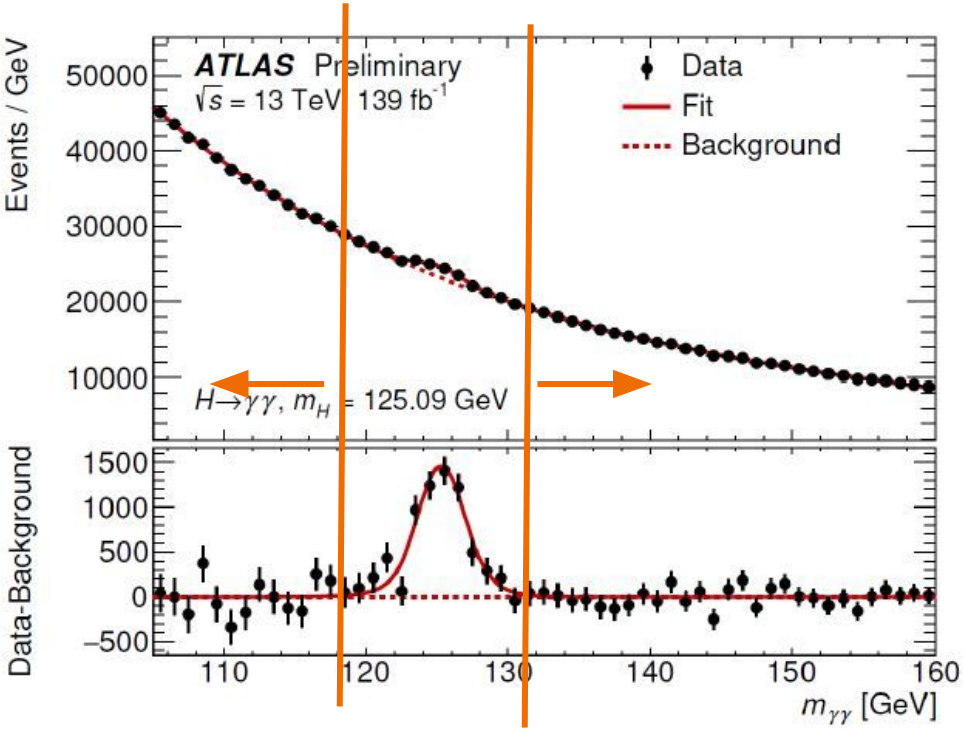


Fit analytical function to subtract

Background rejection

Use control regions to correct for simulation mismatch

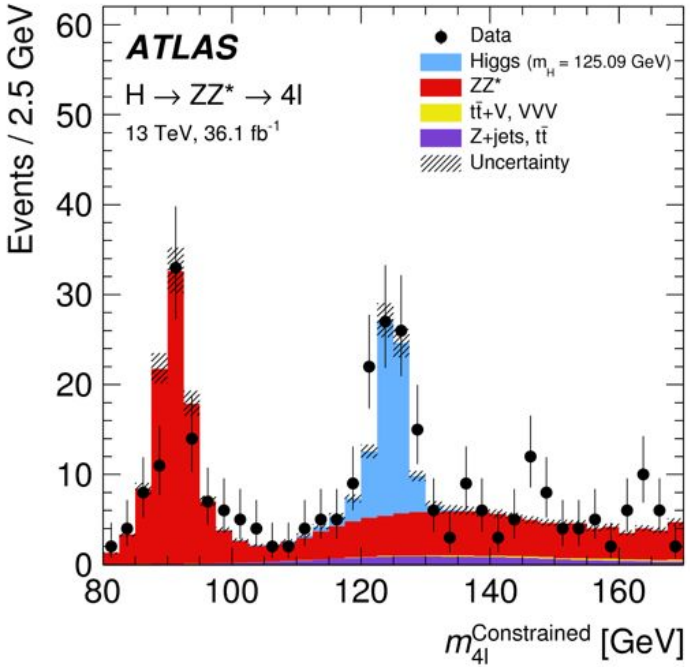
Needs region that is close to the signal region in terms of physics model, without signal

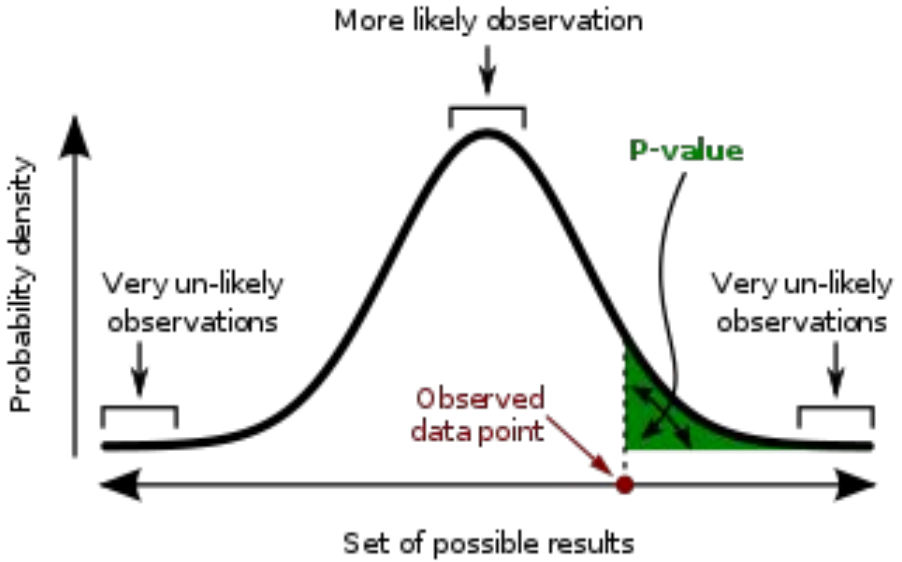


Fit analytical function to subtract

Particle discovery

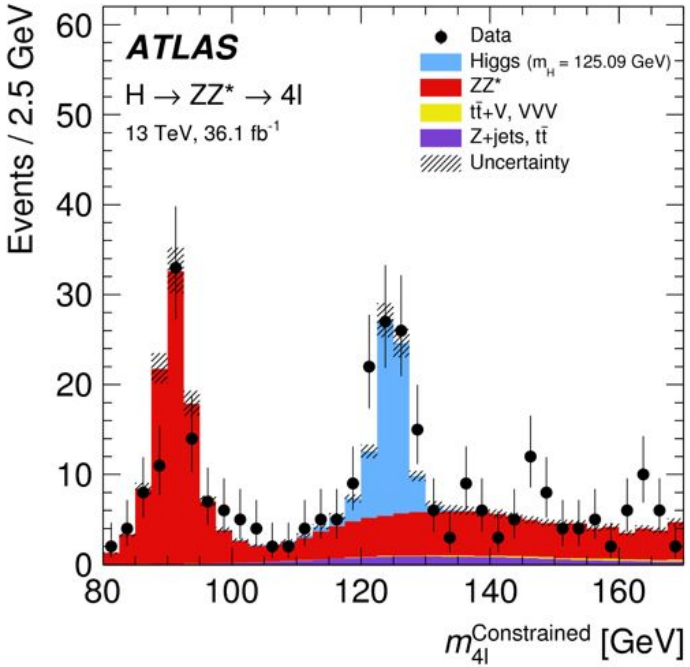
Determining significance

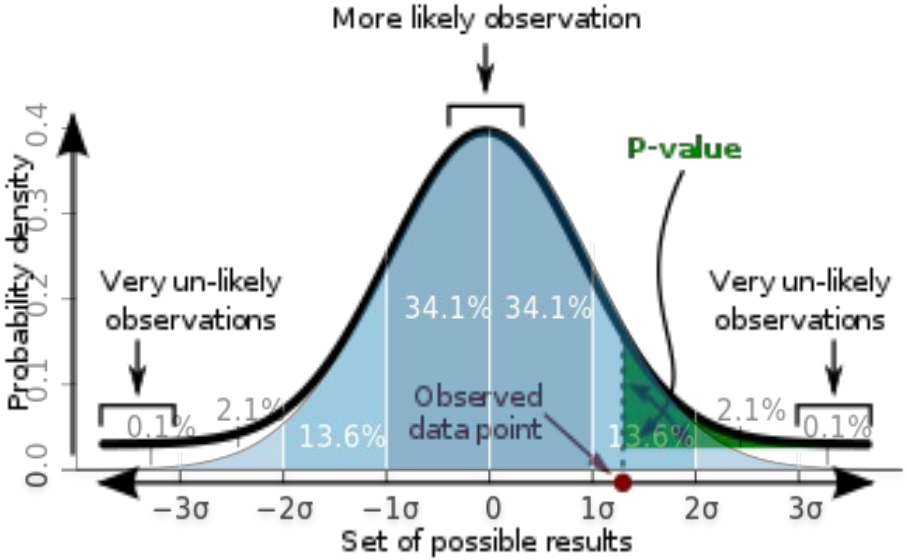




Particle discovery

Require 5 sigma for discovery

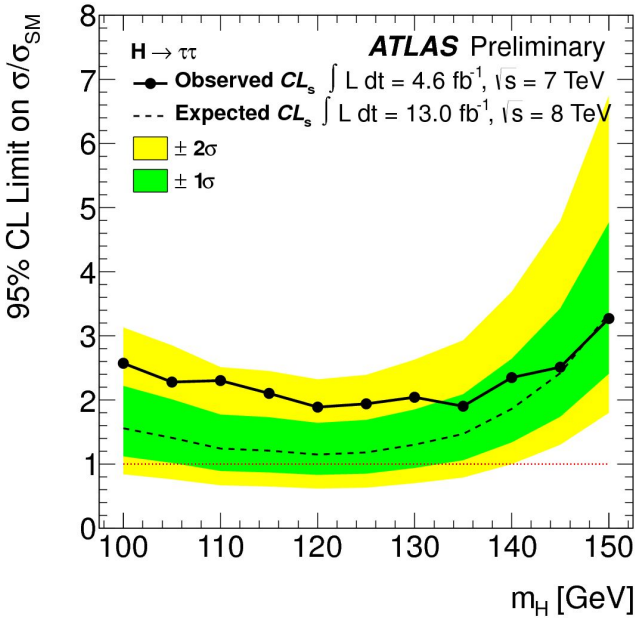


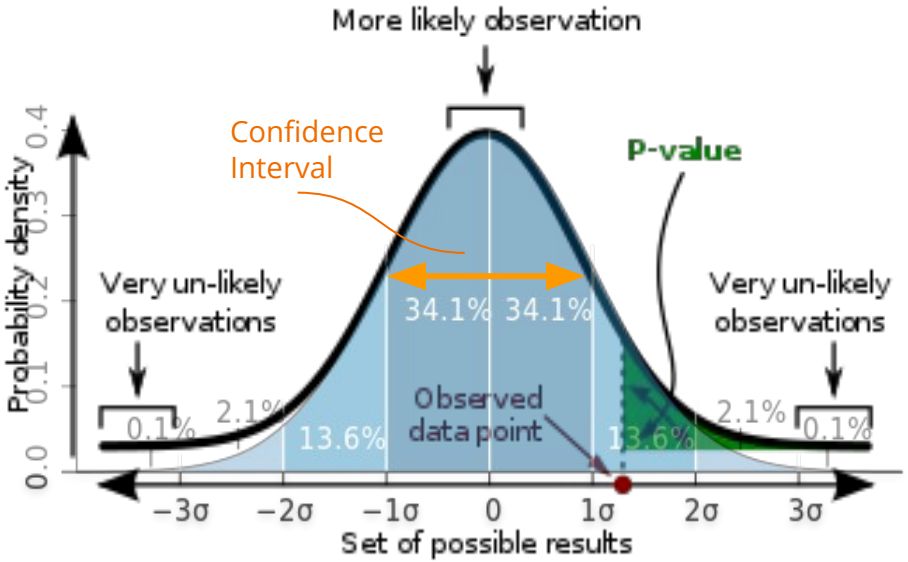


Limit setting

Confidence interval for searches

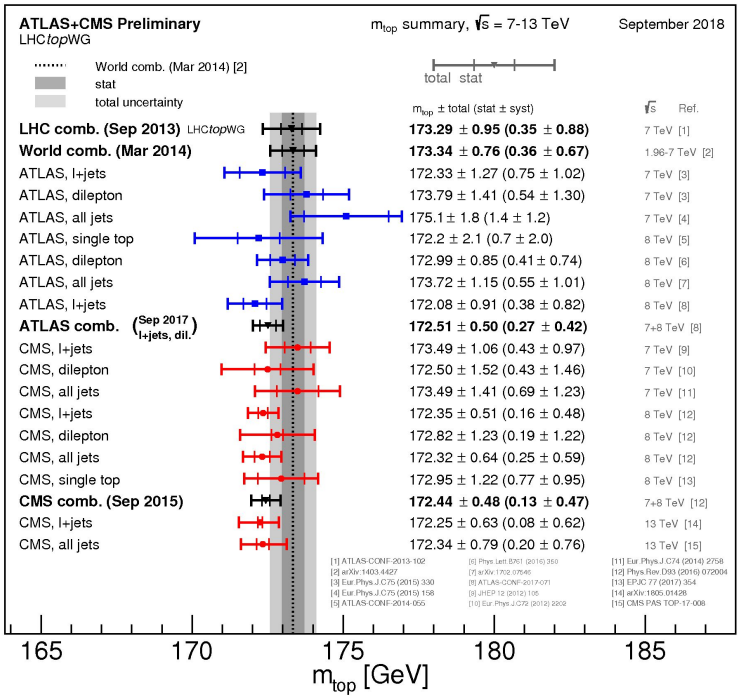
68% or 95%

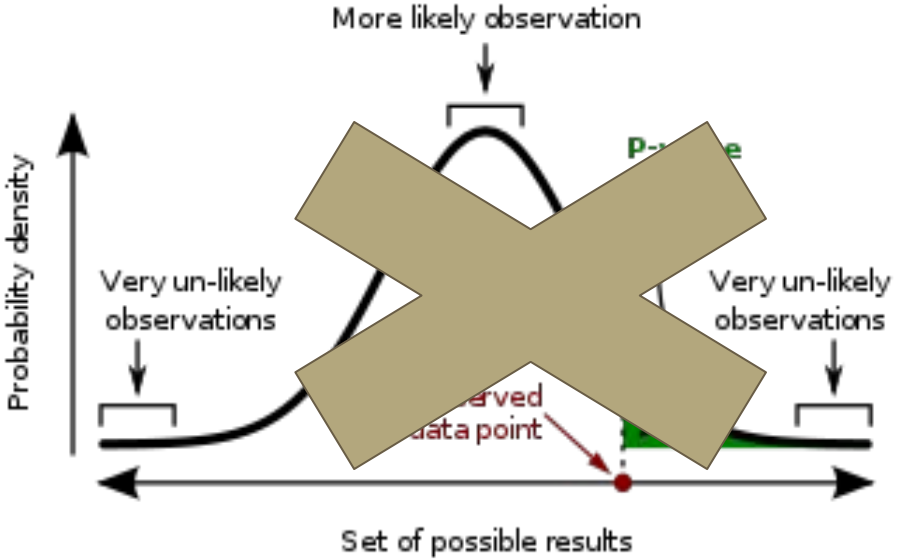




Precision measurements

Uncertainty bands

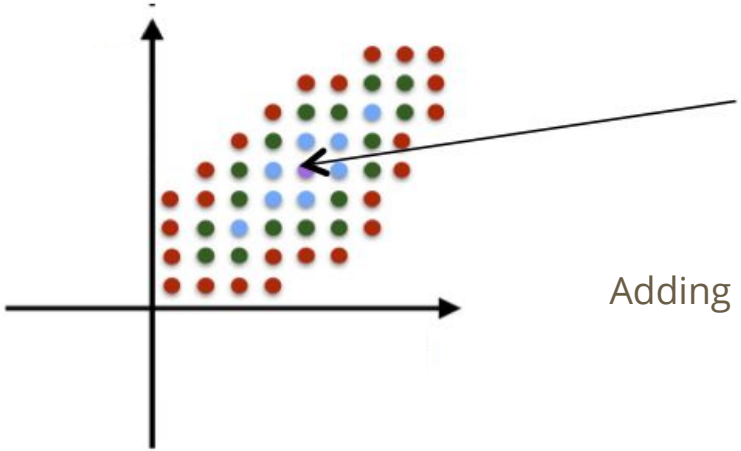




Construction of the likelihood

Start with the formulation of a likelihood: $L(\vec{x} \mid \vec{\mu}, \vec{\theta})$

- (B)SM physics model * Soft physics model * Detector description * Analysis reconstruction



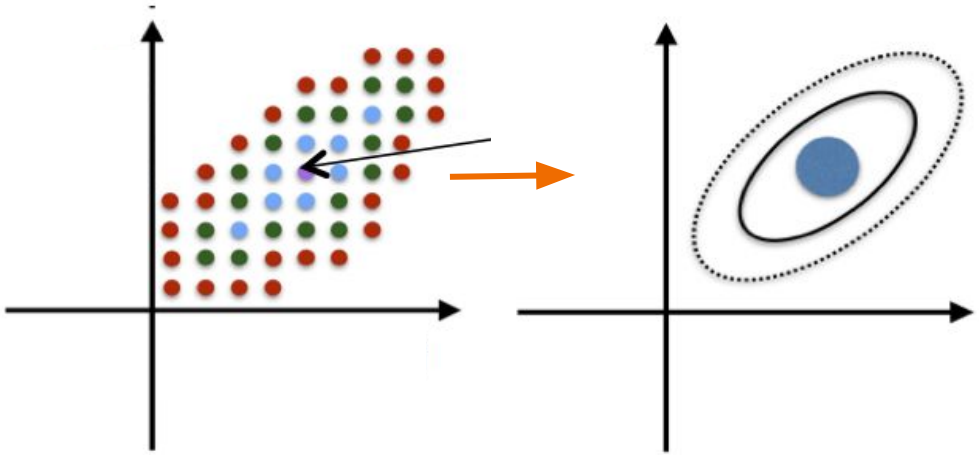
$$L_{ij}(\vec{\theta}) = S_{ij}(x \mid \tilde{R}_i, \tilde{I}_j, \vec{\theta}) + B(x \mid \vec{\theta})$$

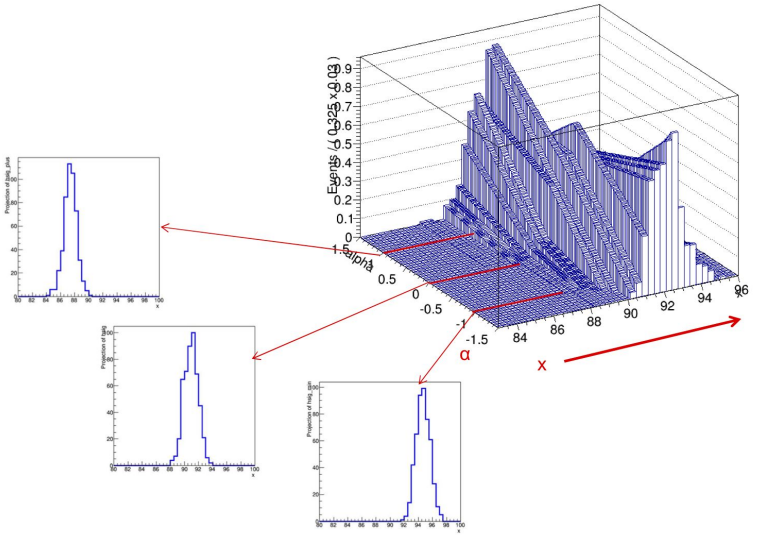
Adding additional terms to the likelihood for modelling uncertainties

Interpolation of uncertainties

Additional terms in the likelihood for modelling uncertainties only known for certain points in parameter phase-space

$$L(\vec{x} \mid \vec{\mu}, \vec{\theta})$$

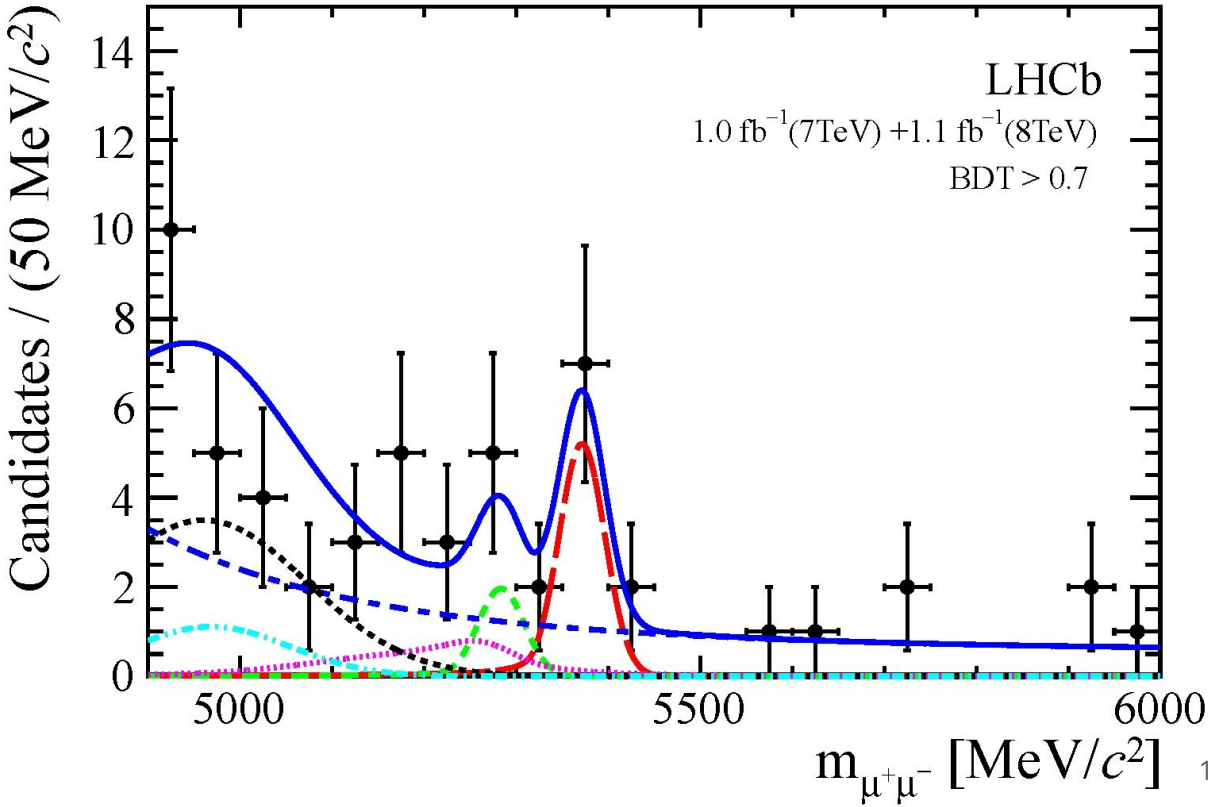




Fitting sum of analytical functions

Create a template based on physics model

- Create summed function for fitting
- Allow variations within modelling uncertainties



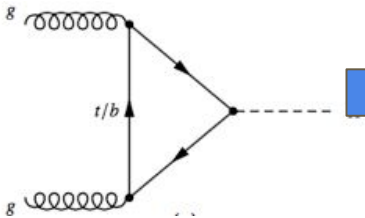
Histogramming

Monte carlo simulation created from physics model

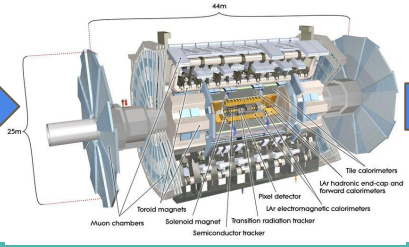
- Often impossible to create analytical function
- Simplify into counting-experiment by binning the distribution
- Single events saved to allow binning in the analysis step

Requires at least as much simulated data as expected data

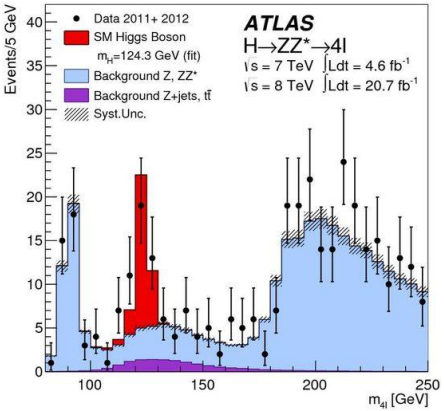
Physics model



Detector simulation



...



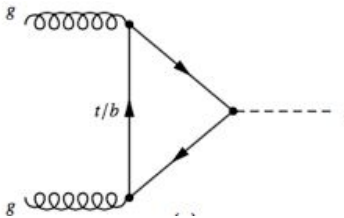
Classification

Use simulation to identify regions of interest based on

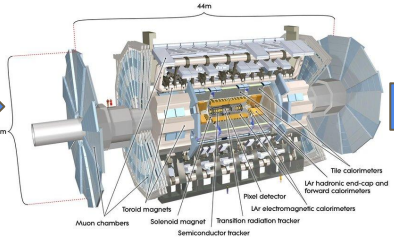
- Physics variables
- Phase-space regions

Multivariate analysis using simple cuts or machine learning

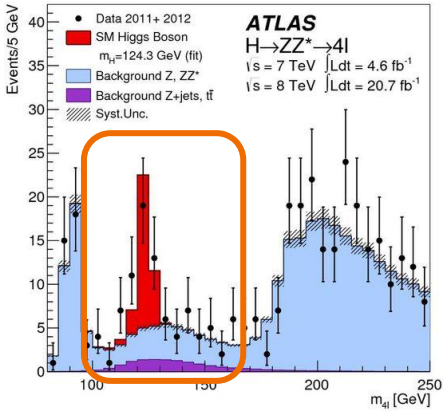
Physics model



Detector simulation

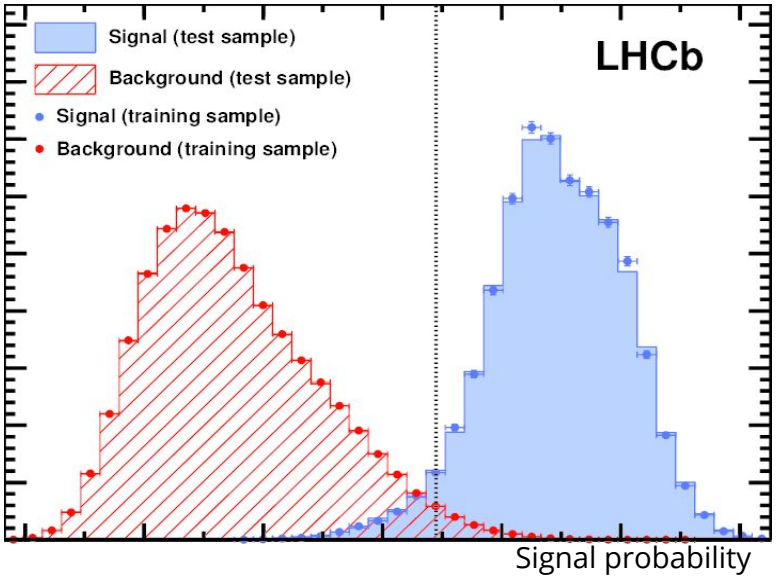


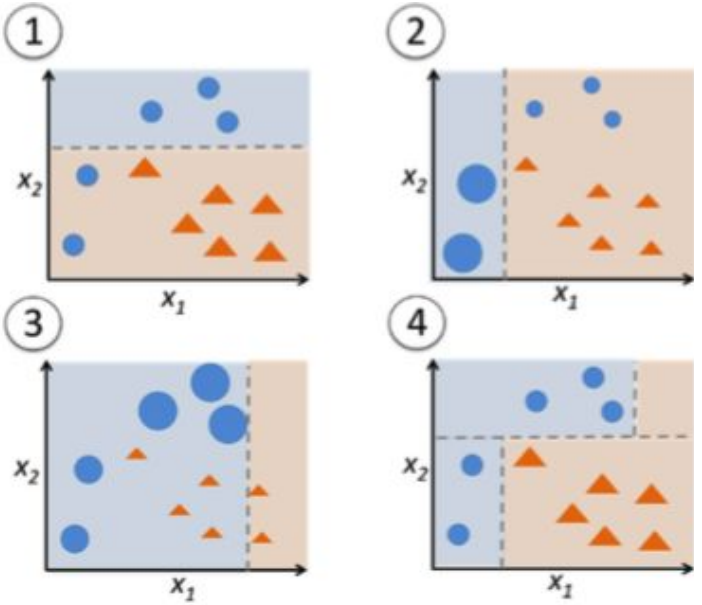
Event selection



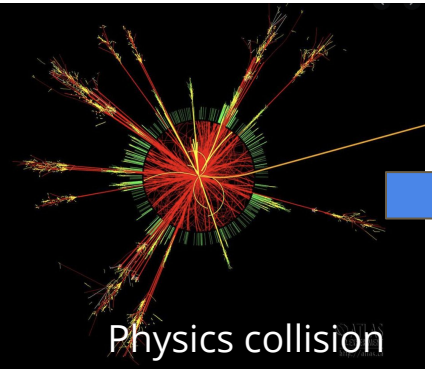
Machine learning

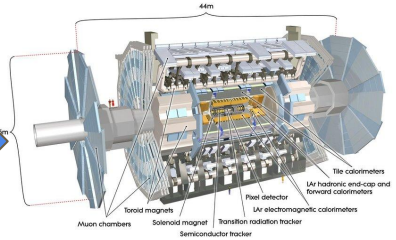
Used to differentiate between signal and background, different models, model detector responses or object reconstruction and particle identification





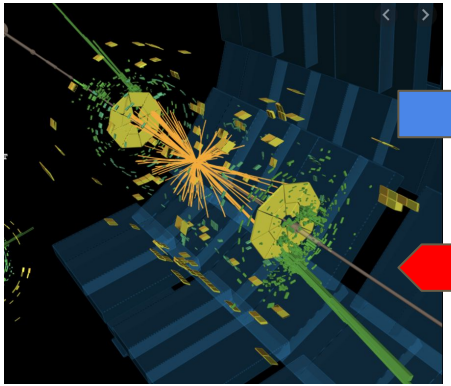
Alternative measurement roadmap: Unfolding





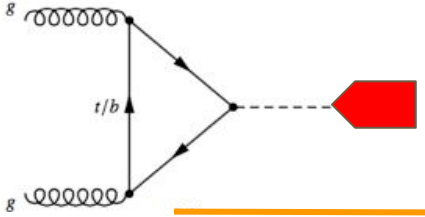
Detector interaction

Detector reconstruction

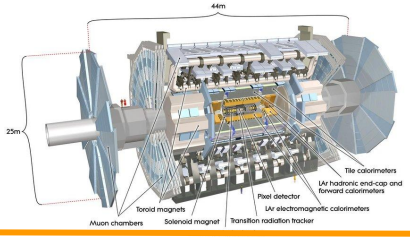


Event selection

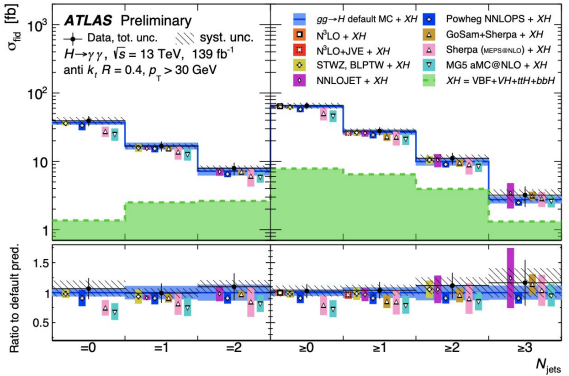
Physics model



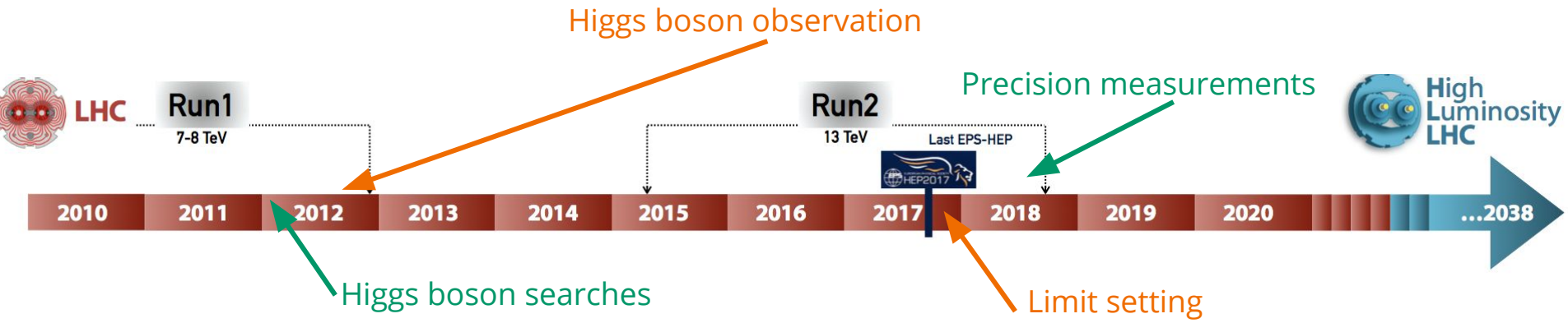
Detector simulation



Analysis



The future of data analysis at the LHC

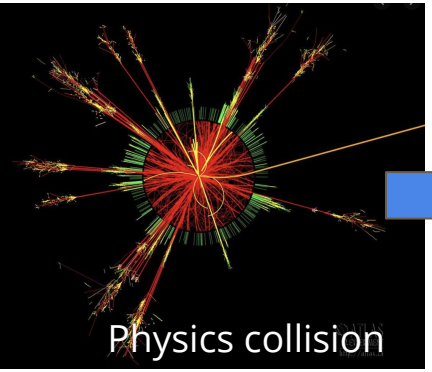


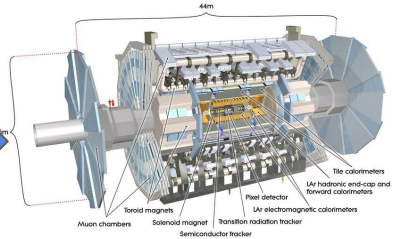
Amount of simulated data needed for the next run is too big

- Clever solutions needed for data manipulation

Moving from statistically limited measurements, to measurements limited by modelling of the physics model and the detector response

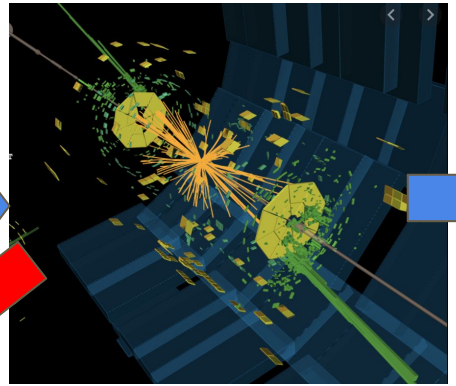
Summary





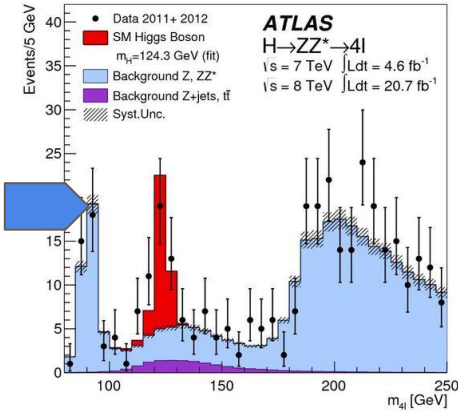
Detector interaction

Detector reconstruction

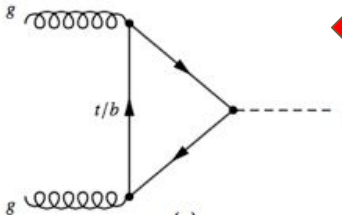


Event selection

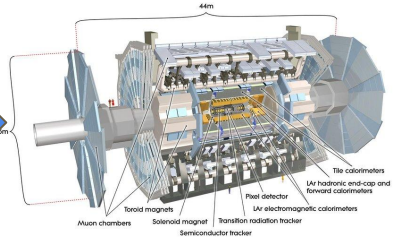
Analysis



Physics model



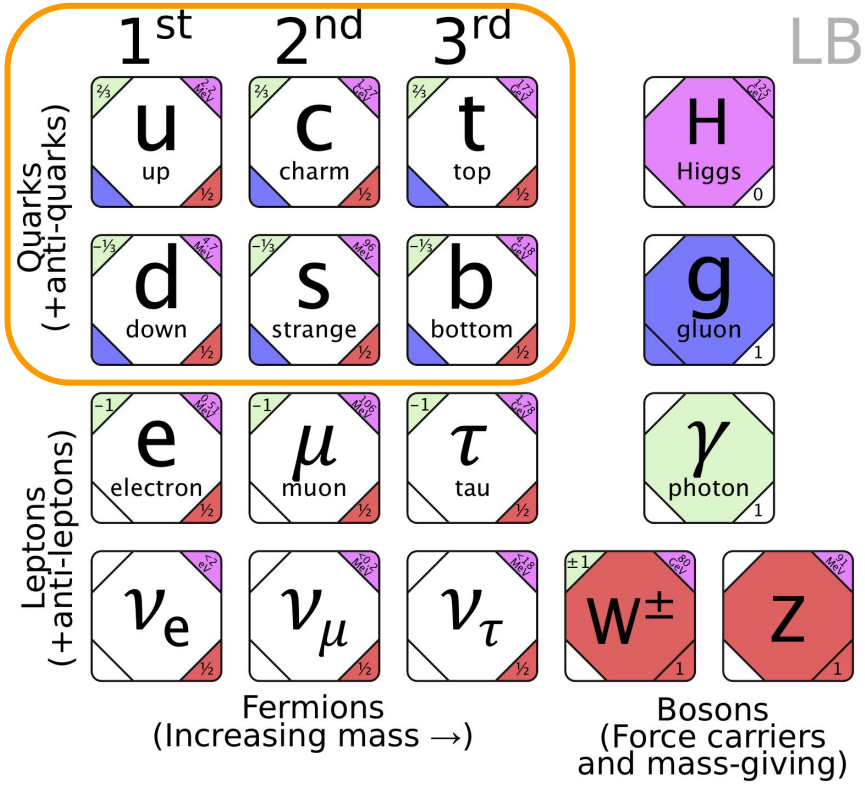
Detector simulation



Back-up

The standard model of particle physics

Hadron building blocks



Computing languages used

Special data-processing framework: ROOT

C++ for performance

More recently: python

Previously: Fortran

Others used: Java, Perl, html

Not used: R -> memory issues due to too large datasets

Morphing based on physics model

$$\mathcal{L}_0^V = \Big\{ c_\alpha \kappa_{SM} \left[\frac{1}{2} \tilde{g}_{HZZ} Z_\mu Z^\mu + \tilde{g}_{HWW} W_\mu^+ W^{-\mu} \right]$$
$$- \frac{1}{4} \left[c_\alpha \kappa_{H\gamma\gamma} \tilde{g}_{H\gamma\gamma} A_{\mu\nu} A^{\mu\nu} + s_\alpha \kappa_{A\gamma\gamma} \tilde{g}_{A\gamma\gamma} A_{\mu\nu} \tilde{A}^{\mu\nu} \right]$$
$$- \frac{1}{2} \left[c_\alpha \kappa_{HZ\gamma} \tilde{g}_{HZ\gamma} Z_{\mu\nu} A^{\mu\nu} + s_\alpha \kappa_{AZ\gamma} \tilde{g}_{AZ\gamma} Z_{\mu\nu} \tilde{A}^{\mu\nu} \right]$$
$$- \frac{1}{4} \left[c_\alpha \kappa_{Hgg} \tilde{g}_{Hgg} G^a_{\mu\nu} G^{a,\mu\nu} + s_\alpha \kappa_{Agg} \tilde{g}_{Agg} G^a_{\mu\nu} \tilde{G}^{a,\mu\nu} \right]$$
$$- \frac{1}{4} \frac{1}{\Lambda} \left[c_\alpha \kappa_{HZZ} Z_{\mu\nu} Z^{\mu\nu} + s_\alpha \kappa_{AZZ} Z_{\mu\nu} \tilde{Z}^{\mu\nu} \right]$$
$$- \frac{1}{2} \frac{1}{\Lambda} \left[c_\alpha \kappa_{HWW} W^+_{\mu\nu} W^{-\mu\nu} + s_\alpha \kappa_{AWW} W^+_{\mu\nu} \tilde{W}^{-\mu\nu} \right]$$
$$- \frac{1}{\Lambda} c_\alpha \left[\kappa_{H\partial\gamma} Z_\nu \partial_\mu A^{\mu\nu} + \kappa_{H\partial Z} Z_\nu \partial_\mu Z^{\mu\nu} + \kappa_{H\partial W} (W^+_\nu \partial_\mu W^{-\mu\nu} + h.c.) \right] \Big\} X_0$$

Used in Run 1

Plan Run 2

$$\mathcal{L}_0^f = - \sum_{f=t,b,\tau} \bar{\psi}_f \left(c_\alpha \kappa_{Hff} g_{Hff} + i s_\alpha \kappa_{Aff} g_{Aff} \gamma_5 \right) \psi_f X_0$$

Fermions...