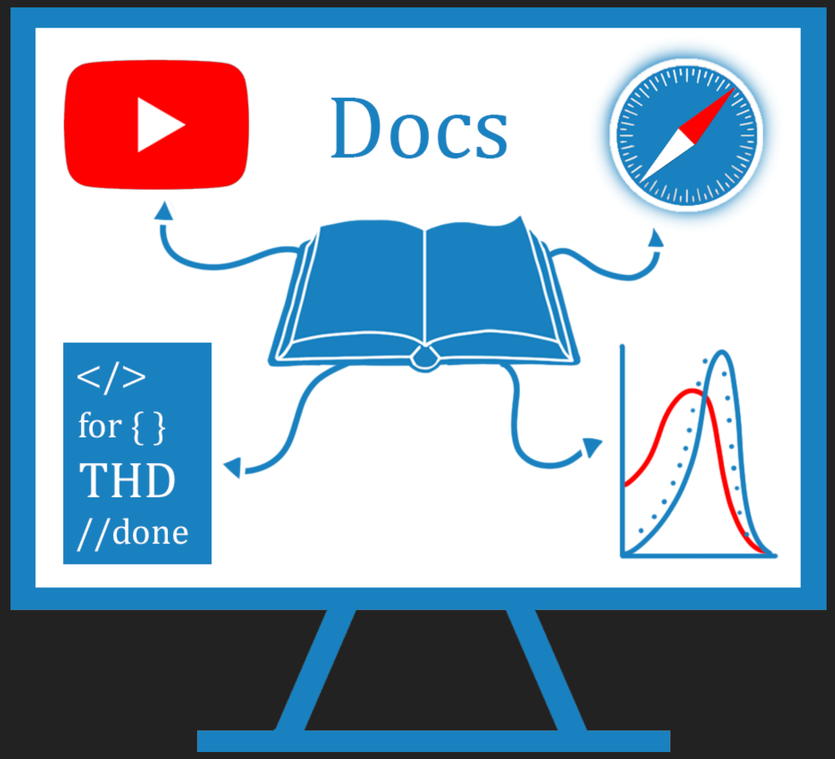


Meirin Oan Evans, Kate Shaw, Tom Stevenson

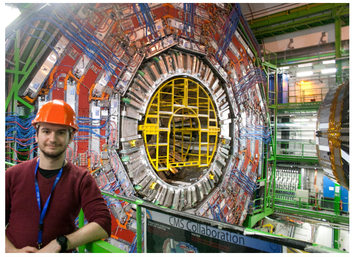
Data Intensive, AI & ML Summer School, Sussex

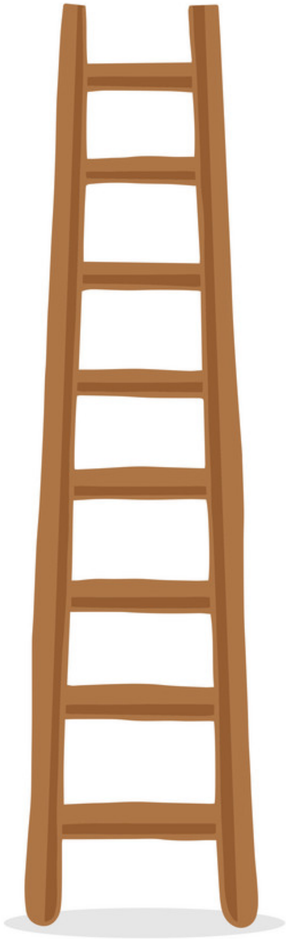
22nd July 2019

Optimising signal / background ratio in the search for the Higgs Boson using statistical techniques

Who are we?

Kate	Lecturer
Tom	Postdoc
Meirin	PhD student

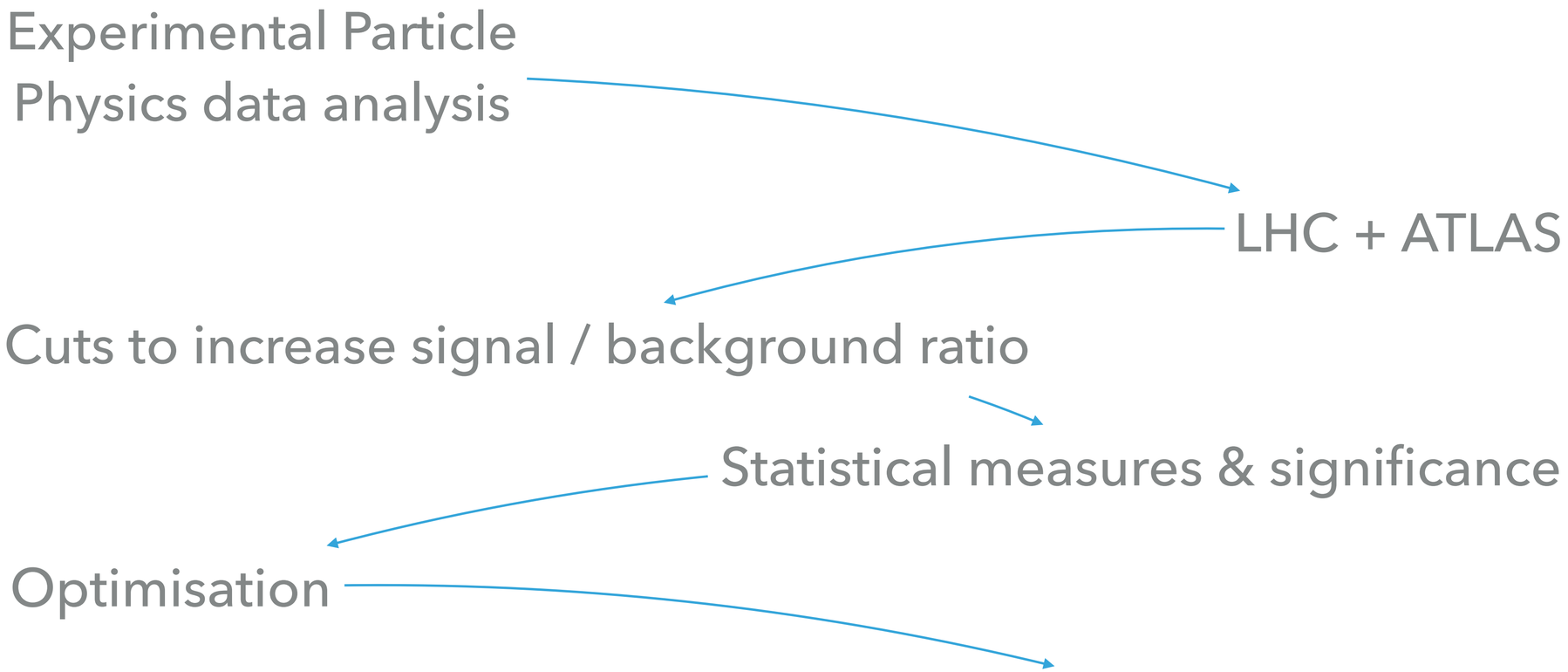




UNIVERSITY OF SUSSEX

open data ATLAS

What will we be doing today?



UNIVERSITY OF SUSSEX

open data ATLAS

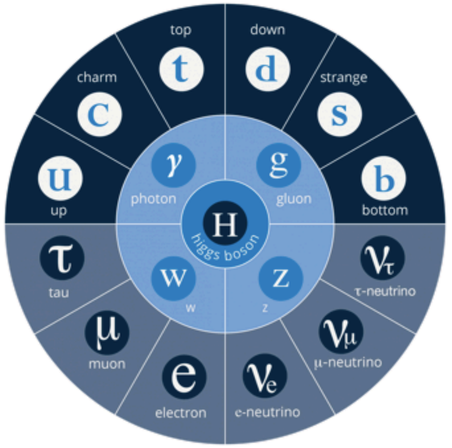
Prerequisites

- Python 3.6 (or above)
- Jupyter notebook
- Could be through an environment like Anaconda
  - Download Anaconda here if you need
- Or run entirely online using Binder
- Does anyone need help setting this up?

Experimental Particle Physics

What are the building blocks of matter?

What are the forces between them?



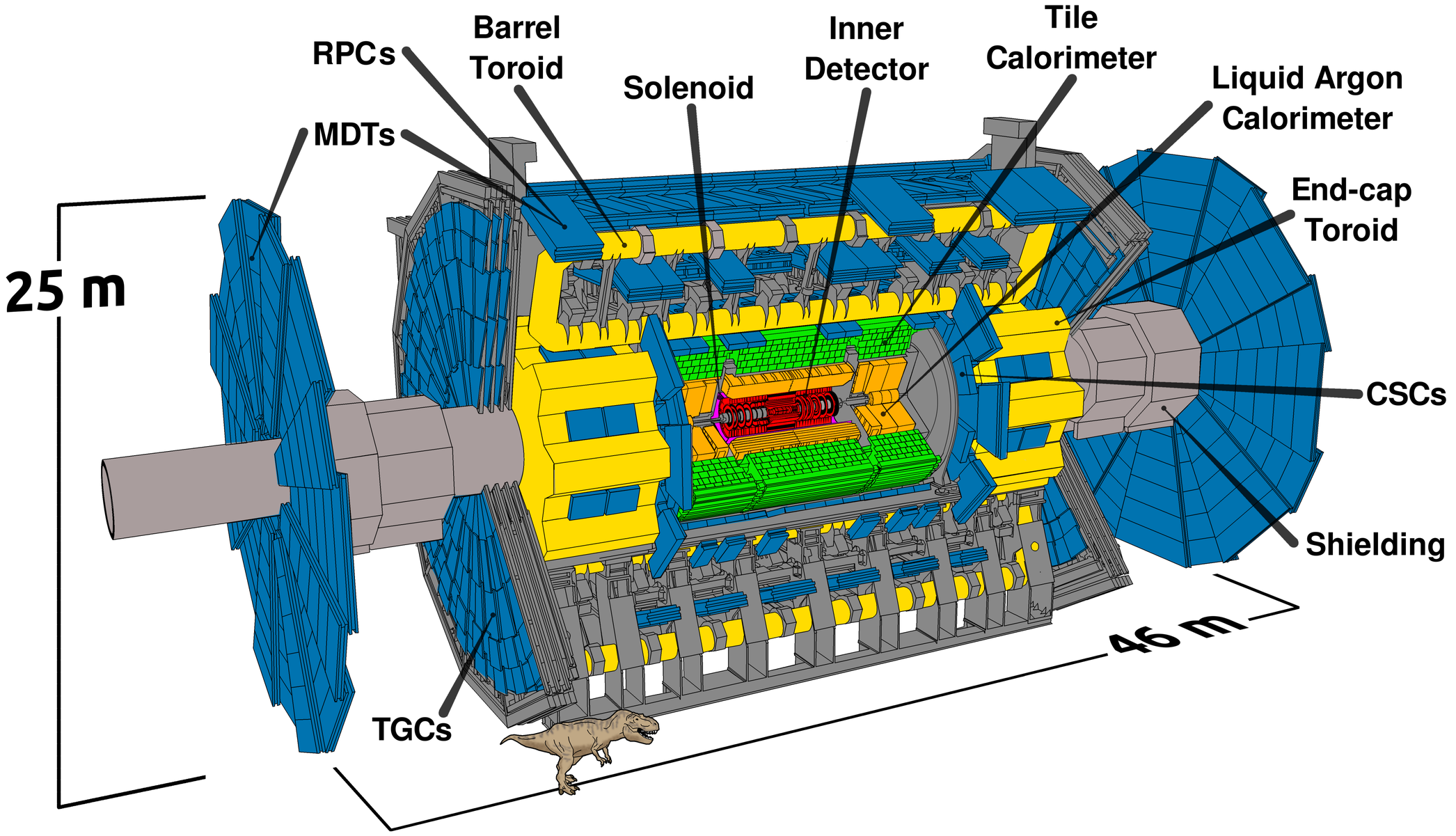
What happened to antimatter?

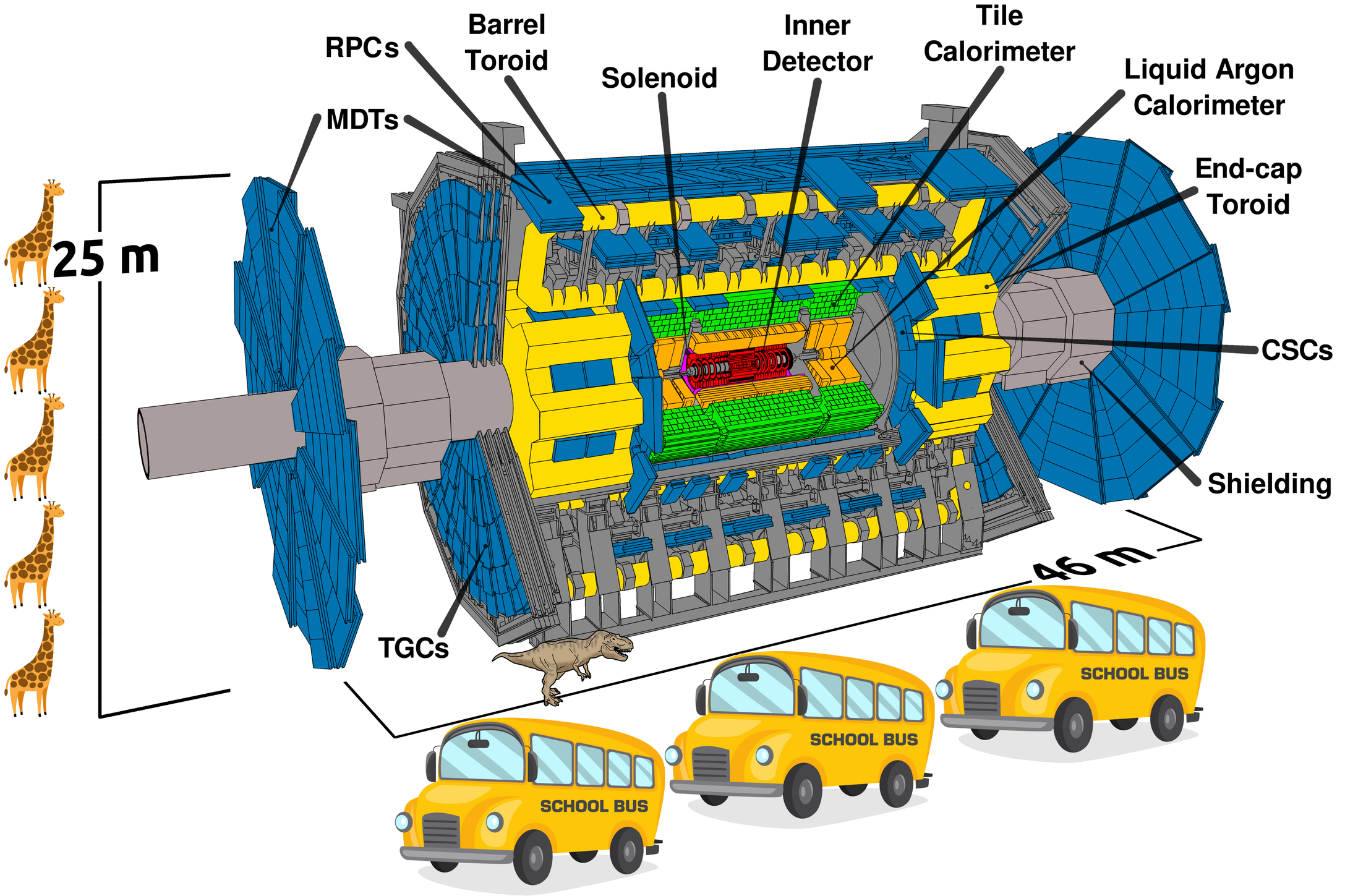
What's dark matter?

What was the early universe like & how did it evolve?

What about gravity?

Anything else?


RPCs
MDTs
Barrel Toroid
Solenoid
Inner Detector
Tile Calorimeter
Liquid Argon Calorimeter
End-cap Toroid
CSCs
Shielding
TGCs
25 m
46 m


RPCs
Barrel Toroid
Solenoid
Inner Detector
Tile Calorimeter
Liquid Argon Calorimeter
MDTs
End-cap Toroid
CSCs
Shielding
TGCs
25 m
46 m
SCHOOL BUS
SCHOOL BUS
SCHOOL BUS

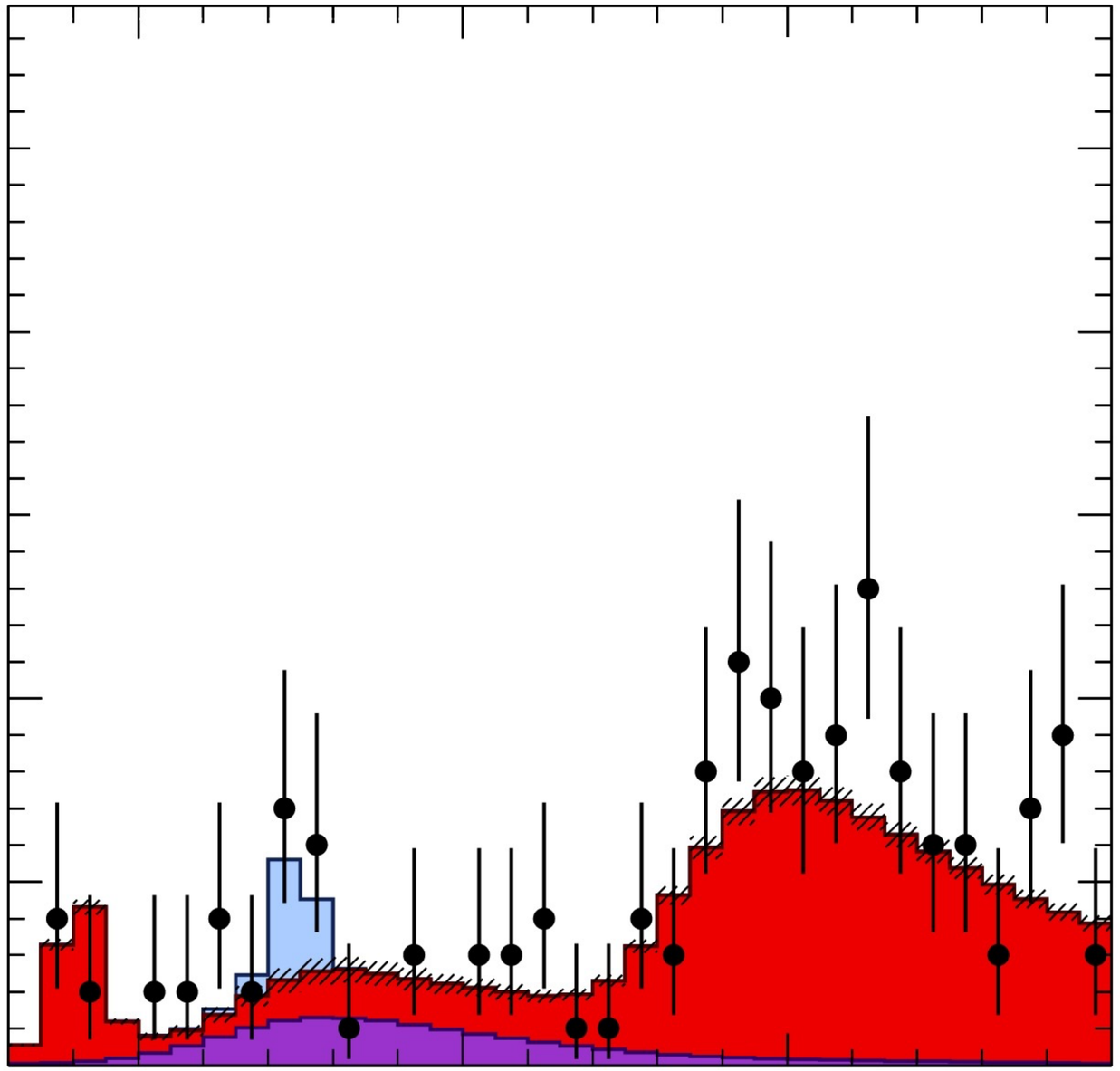
ATLAS data

- ATLAS is designed to observe up to 1.7 billion collisions a second
  - (> 60 million MB/s)
  - On a stack of standard 120mm 700MB CDs, it'd stretch from the Earth to the Sun every year!
- So a trigger system selects ~1000 of these collisions s^{-1}
  - Our CD stack now only stretches from Brighton to Oxford every year…

*not to scale

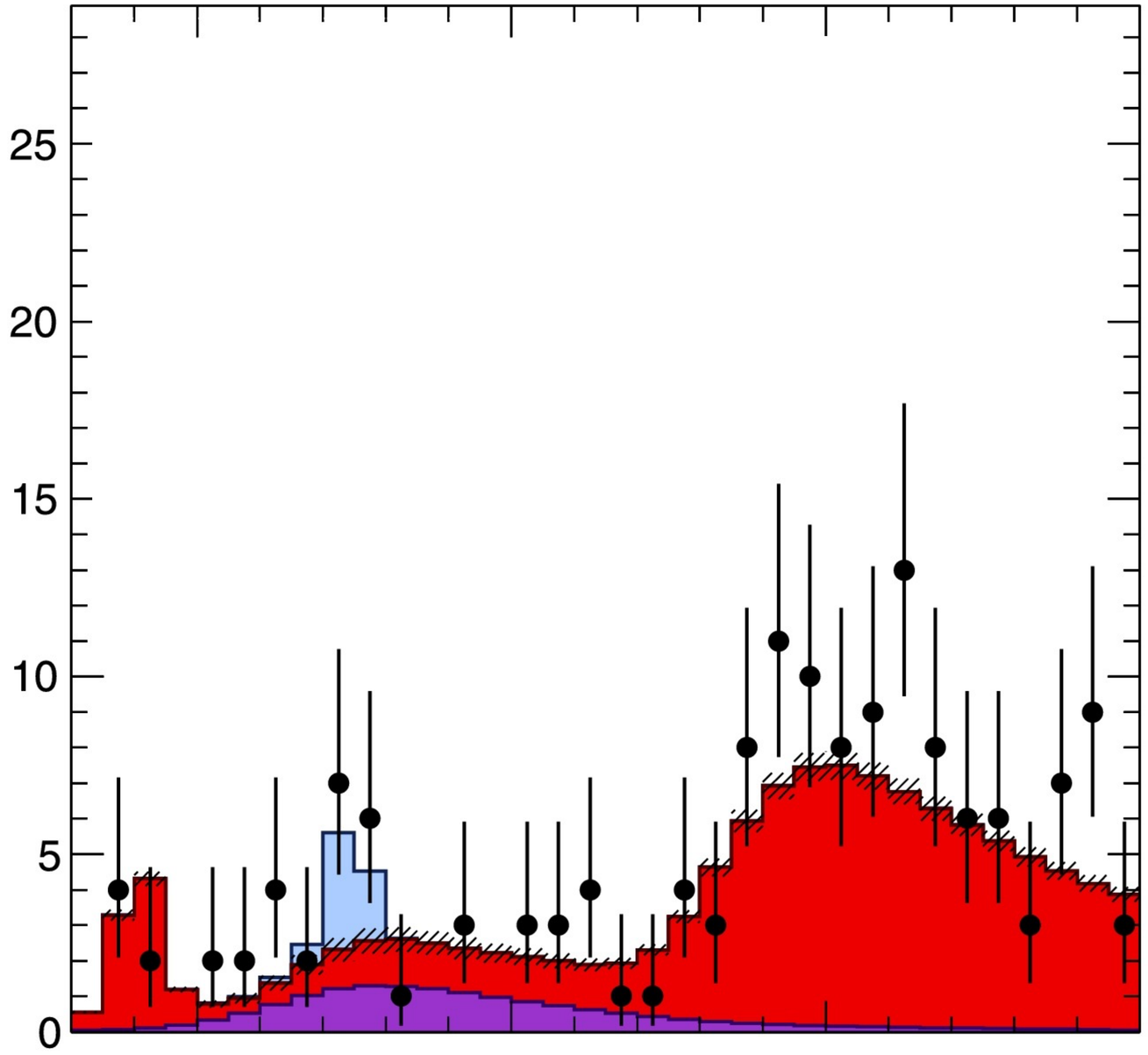
UNIVERSITY OF SUSSEX

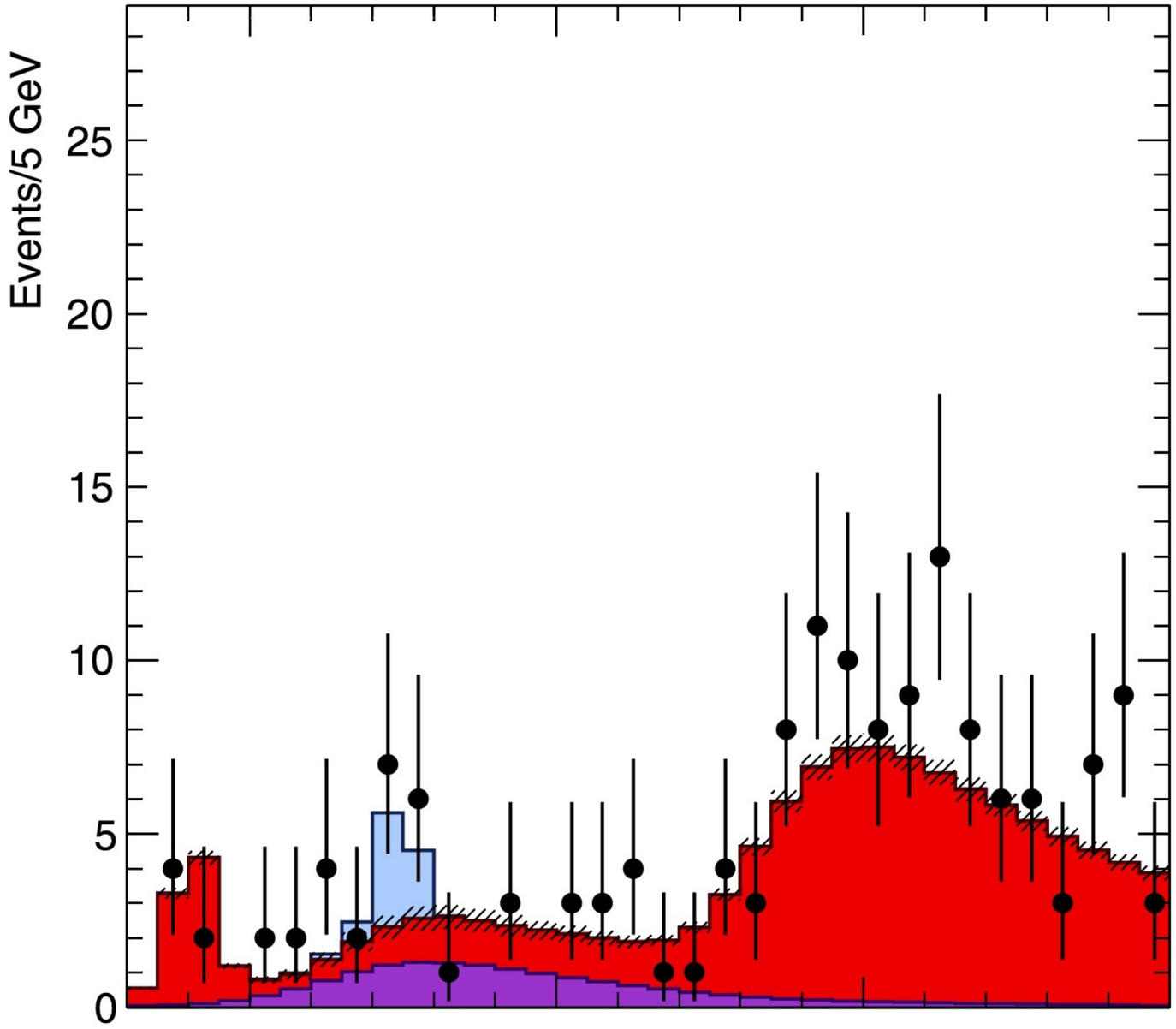
open data ATLAS

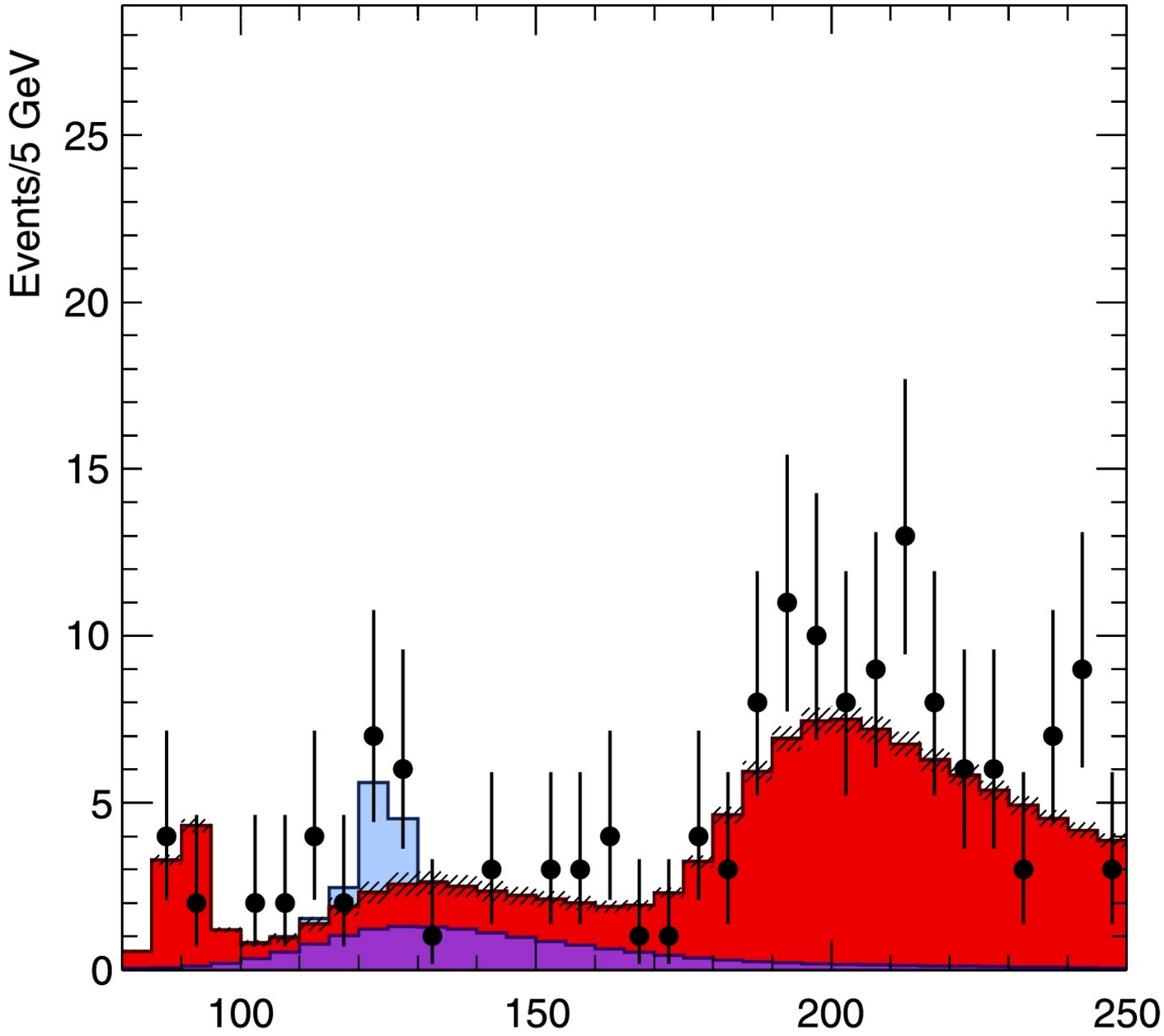


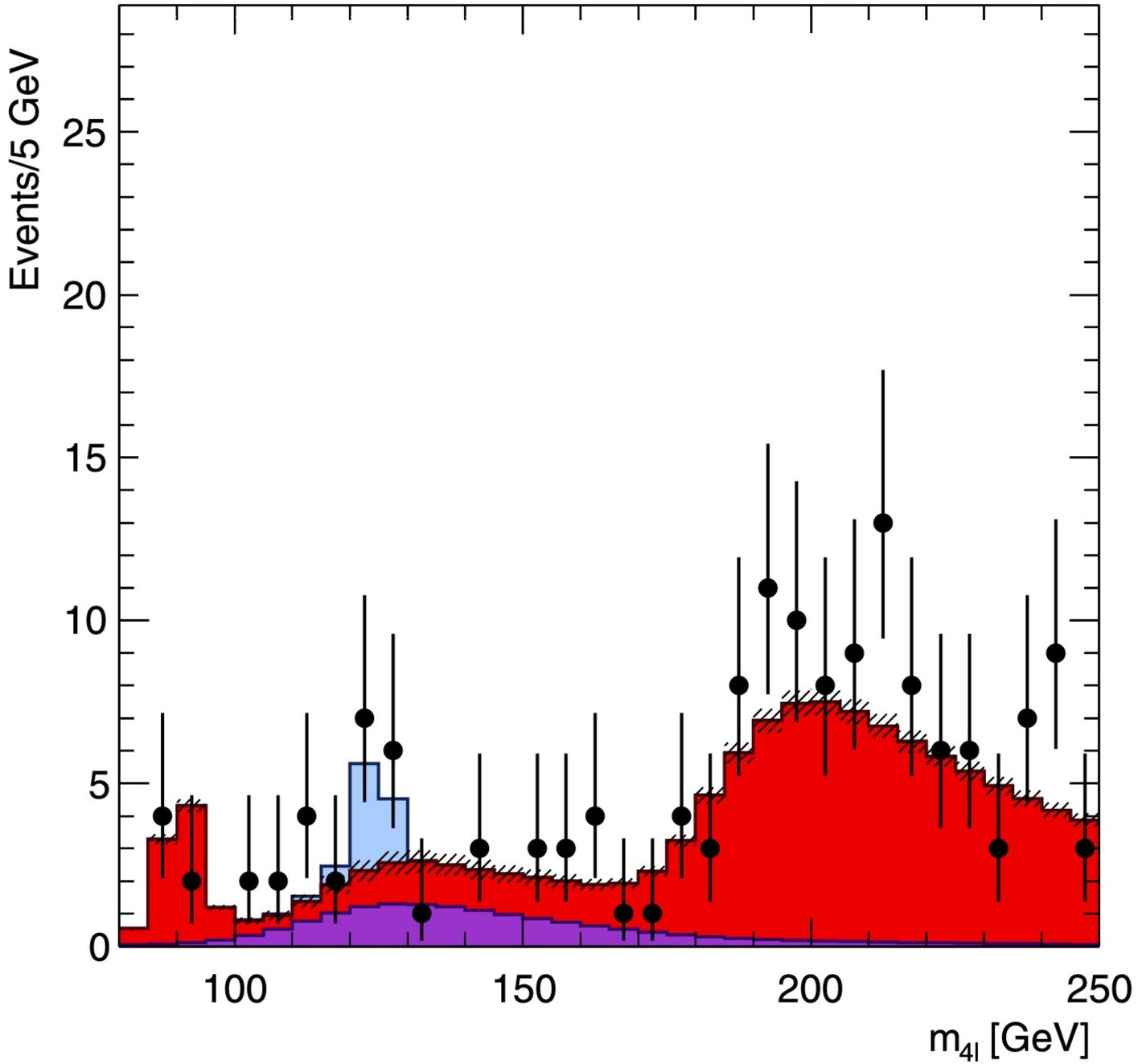
UNIVERSITY OF SUSSEX

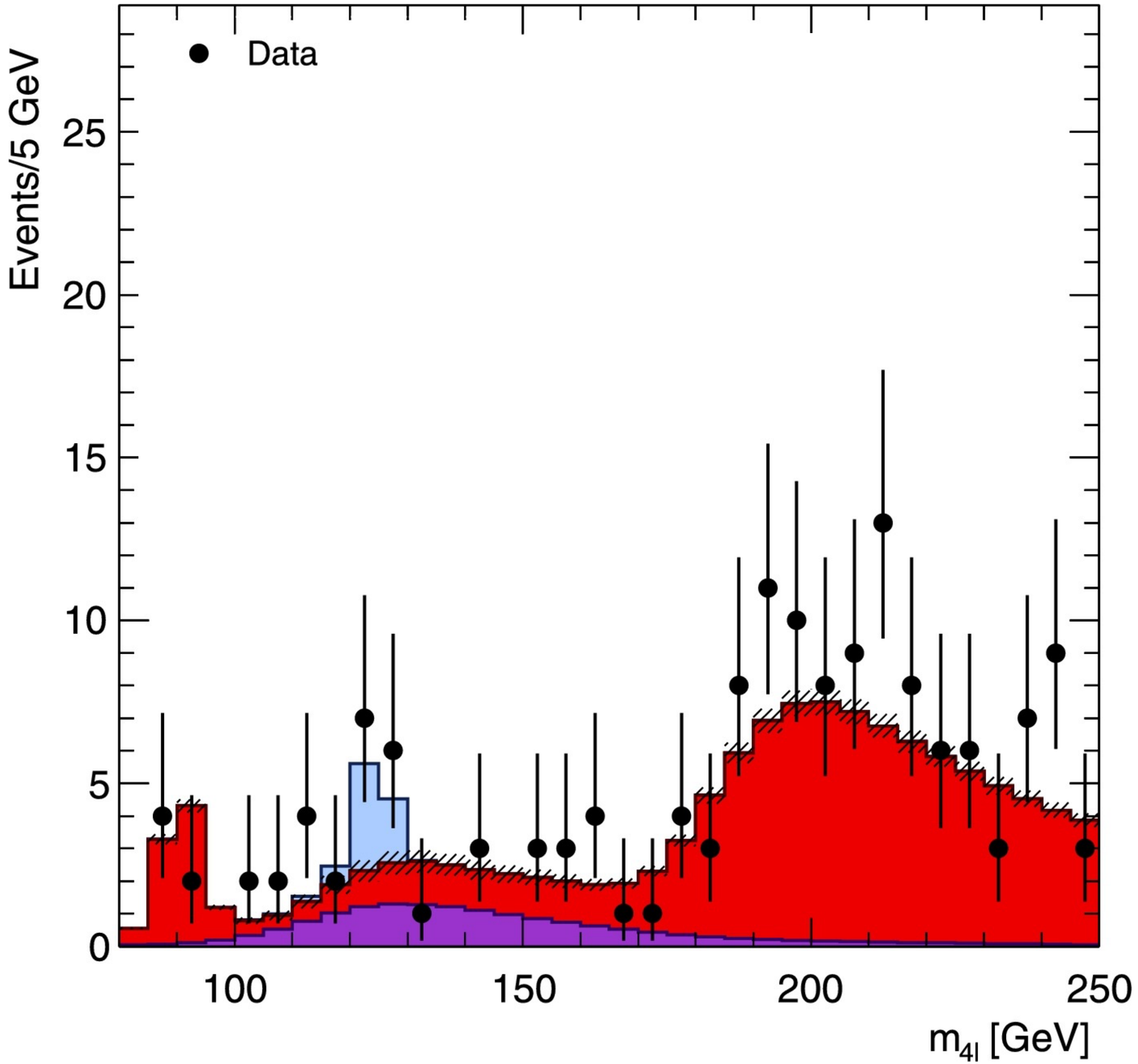
open data ATLAS

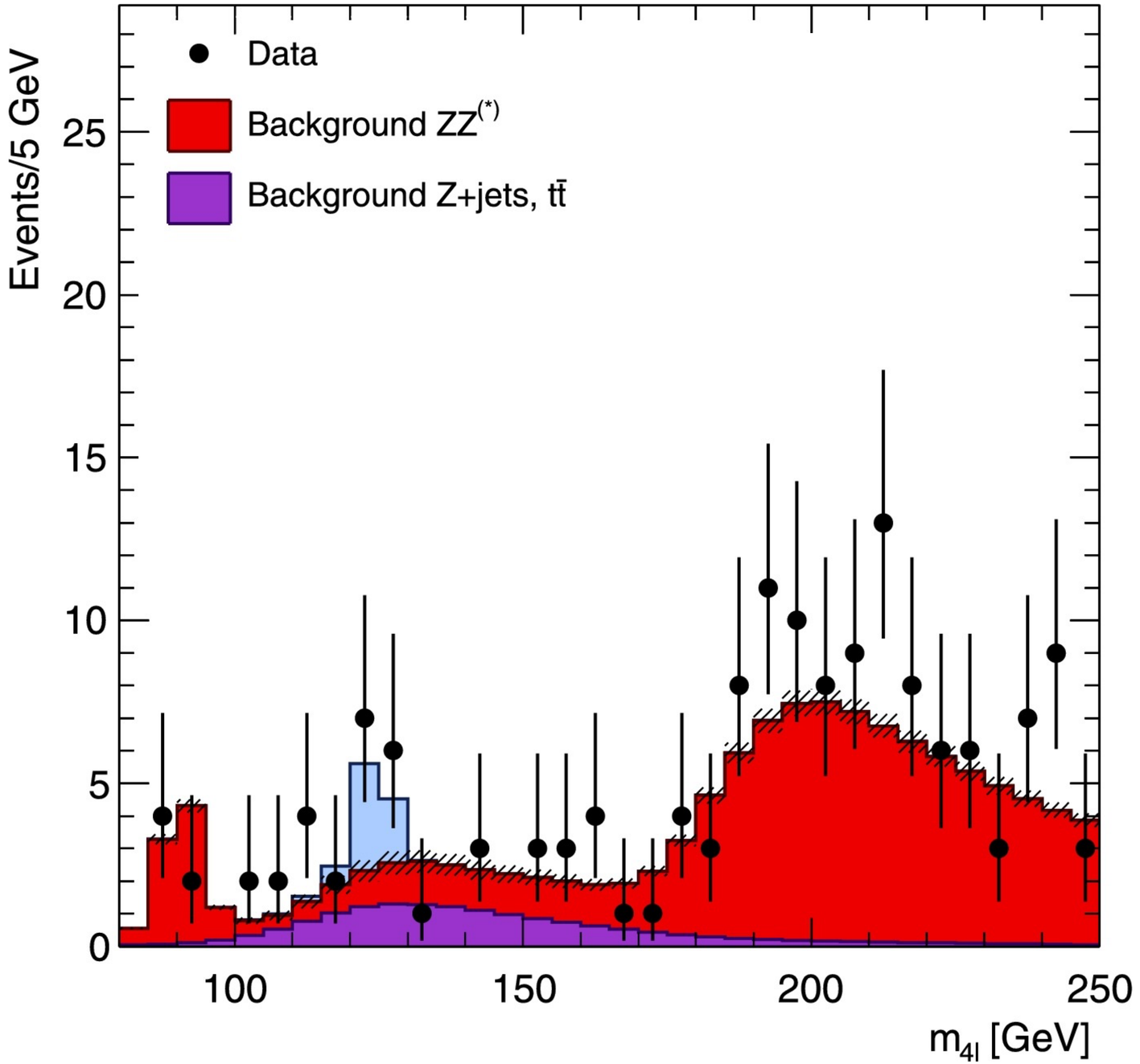

25
20
15
10
5
0

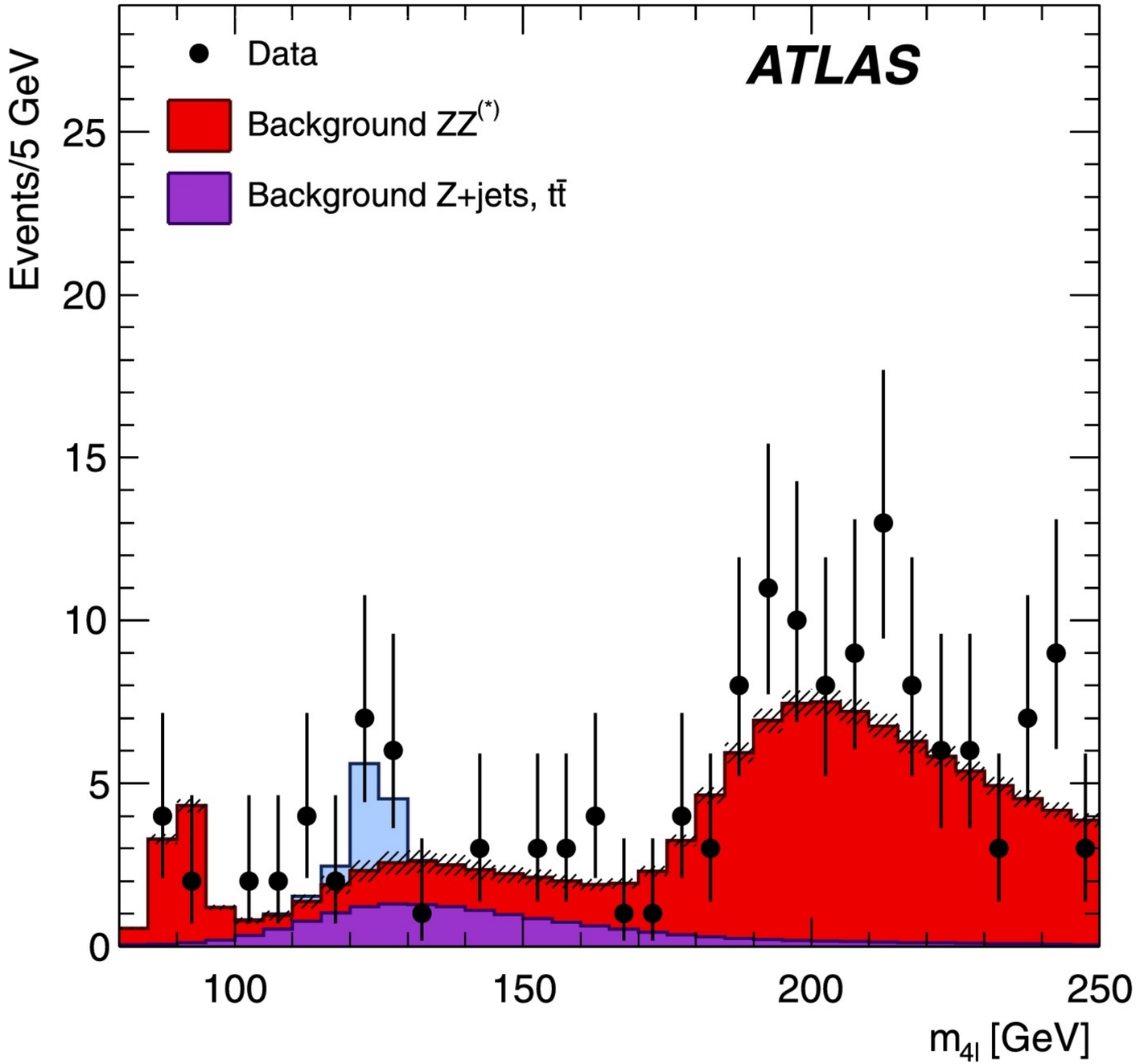

Events/5 GeV
0
5
10
15
20
25

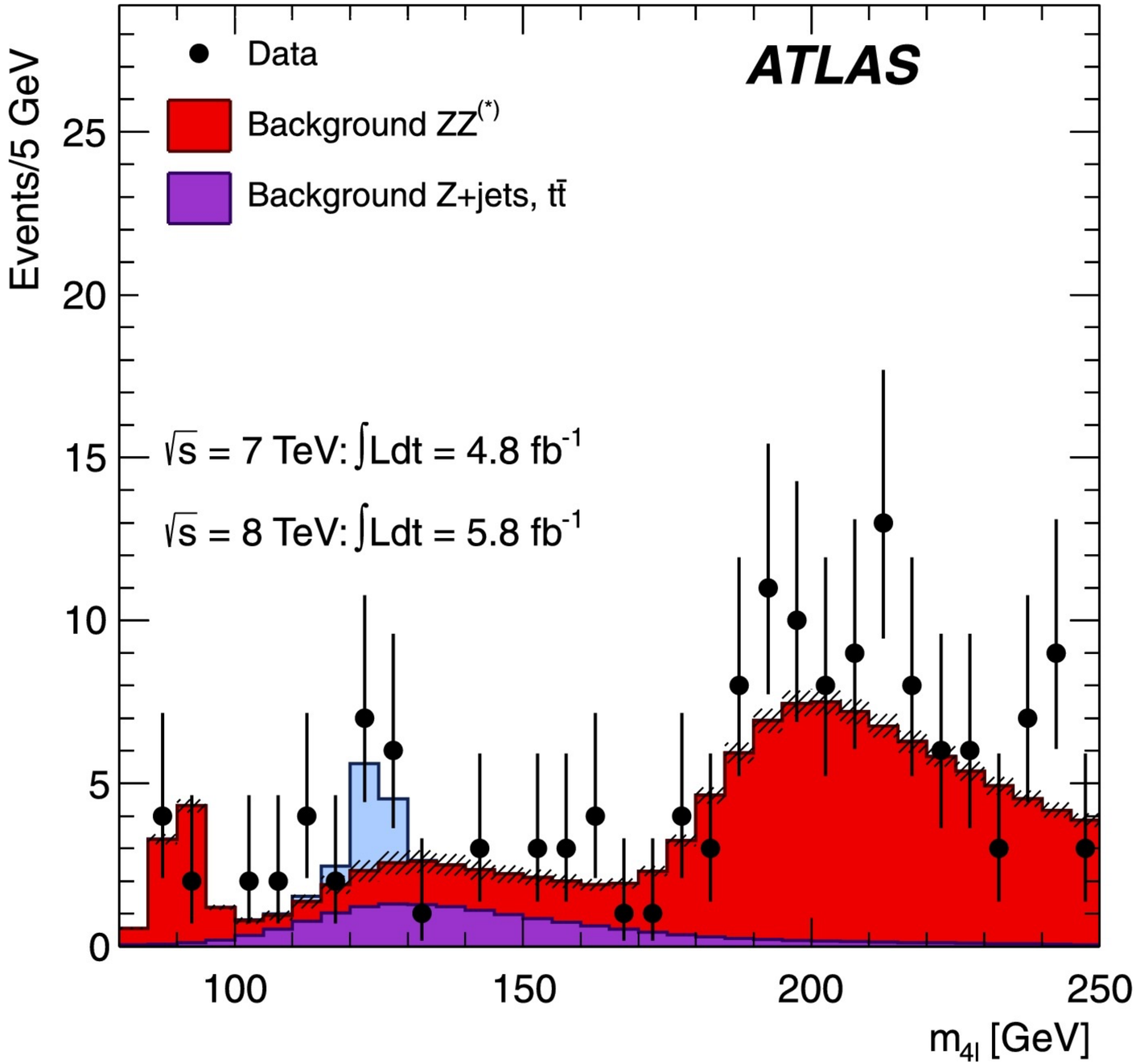

Events/5 GeV
0
5
10
15
20
25
100
150
200
250

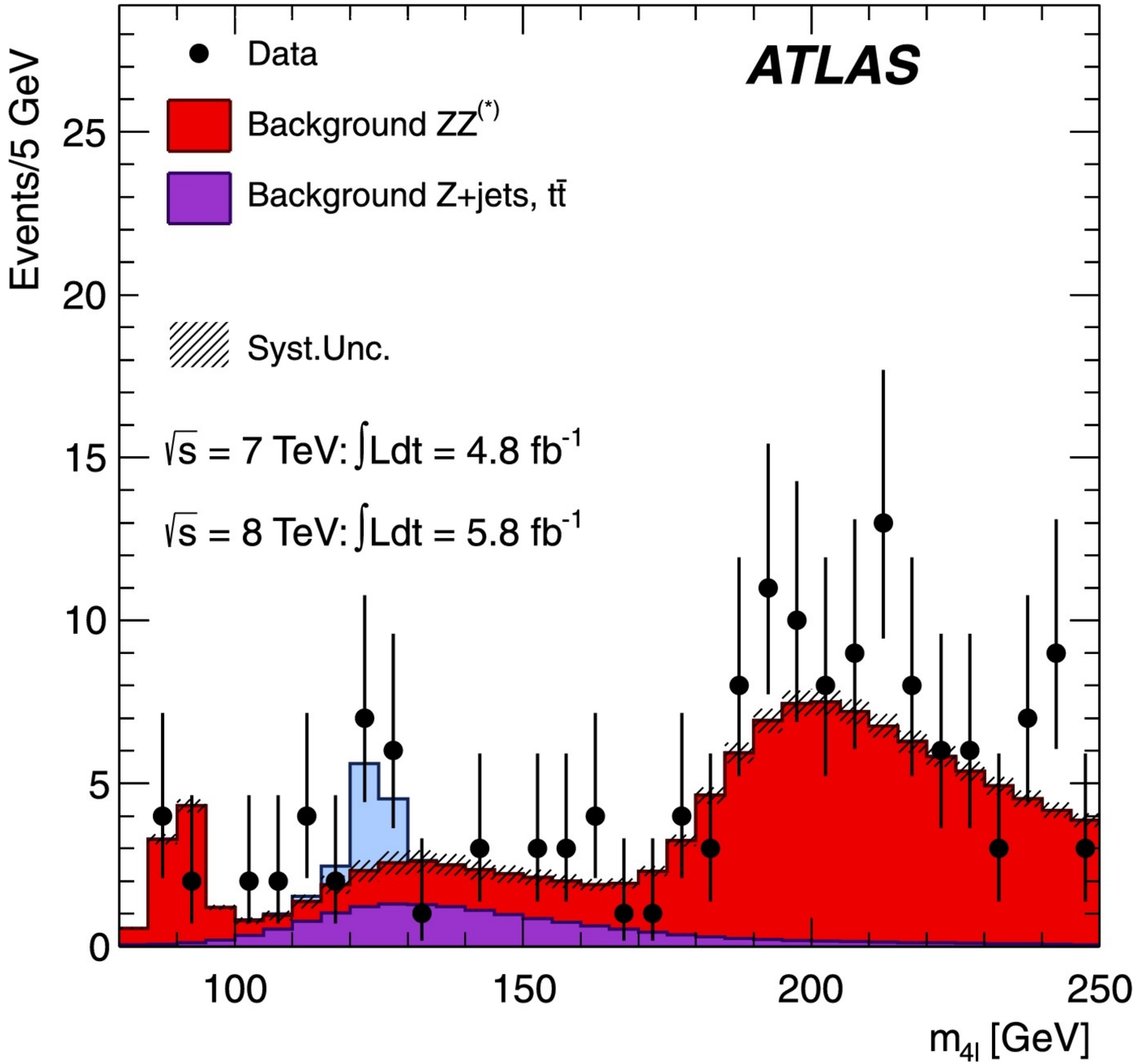

Events/5 GeV
25
20
15
10
5
0
100
150
200
250
m_{4l} [GeV]

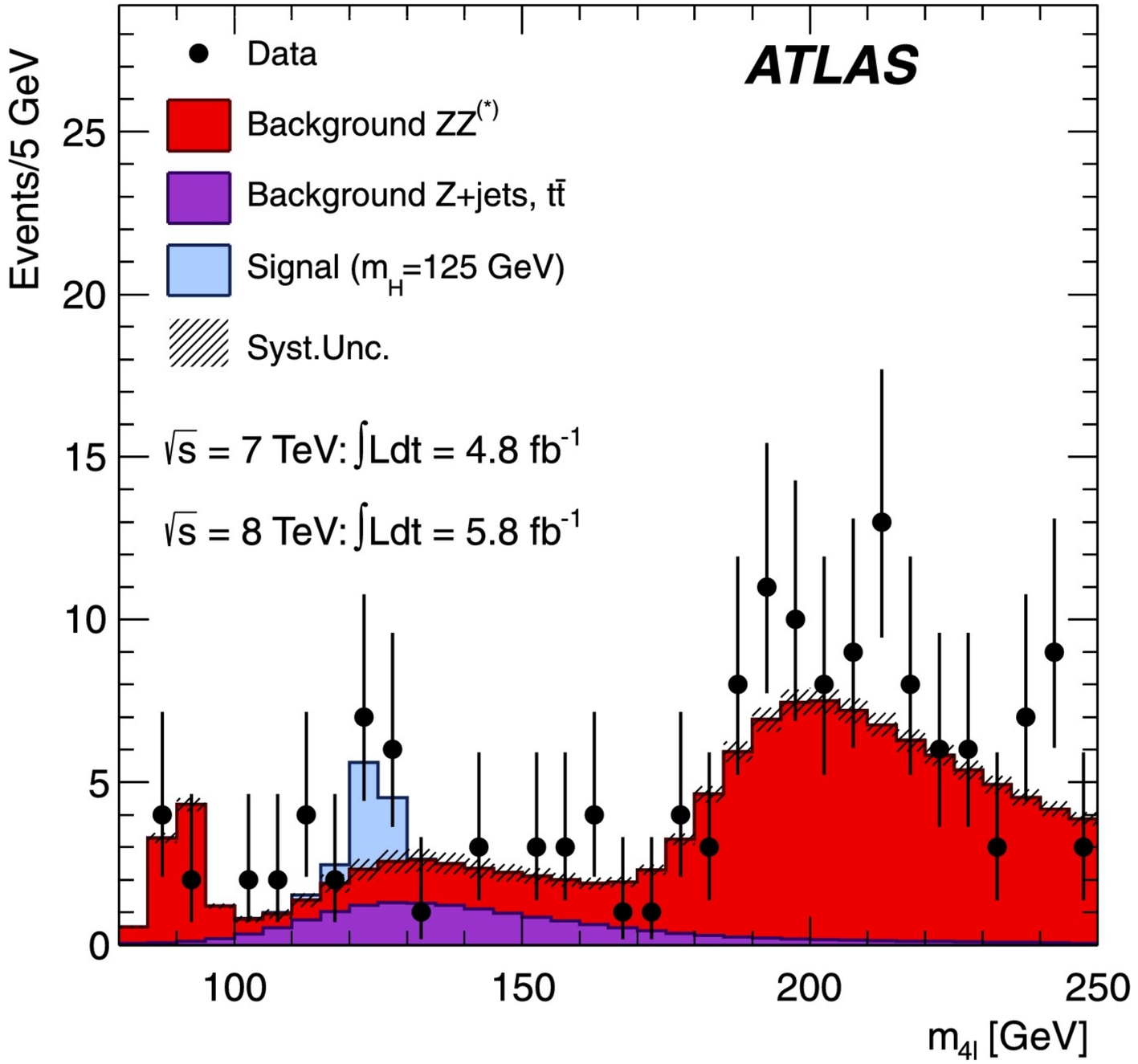

Data
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m_{4l} [GeV]

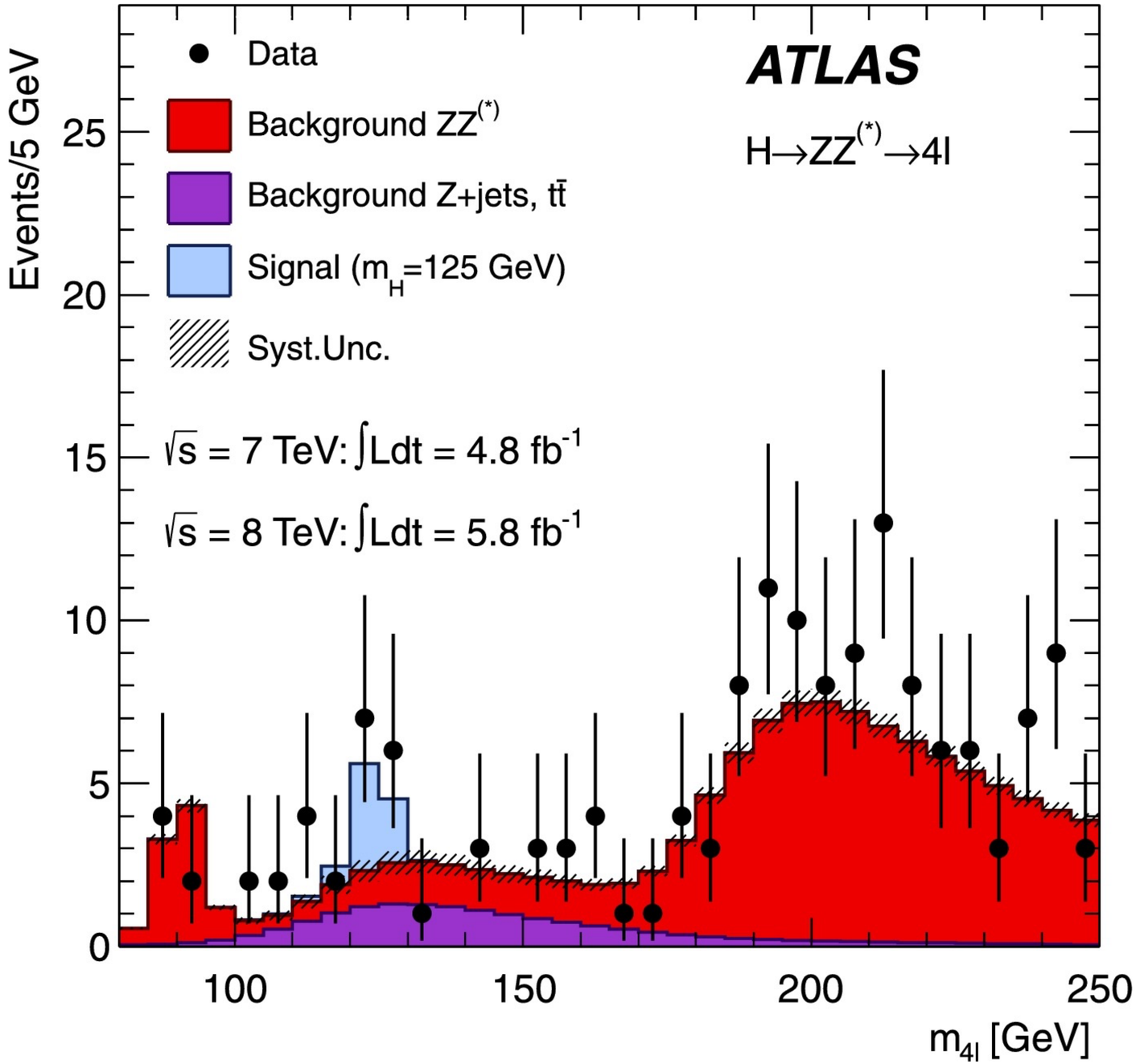

Data
Background ZZ(*)
Background Z+jets, tt̄
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m4l [GeV]


ATLAS
Data
Background ZZ(*)
Background Z+jets, tt̄
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m4l [GeV]


ATLAS
Data
Background ZZ(*)
Background Z+jets, tt̄
√s = 7 TeV: ∫Ldt = 4.8 fb-1
√s = 8 TeV: ∫Ldt = 5.8 fb-1
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m4l [GeV]

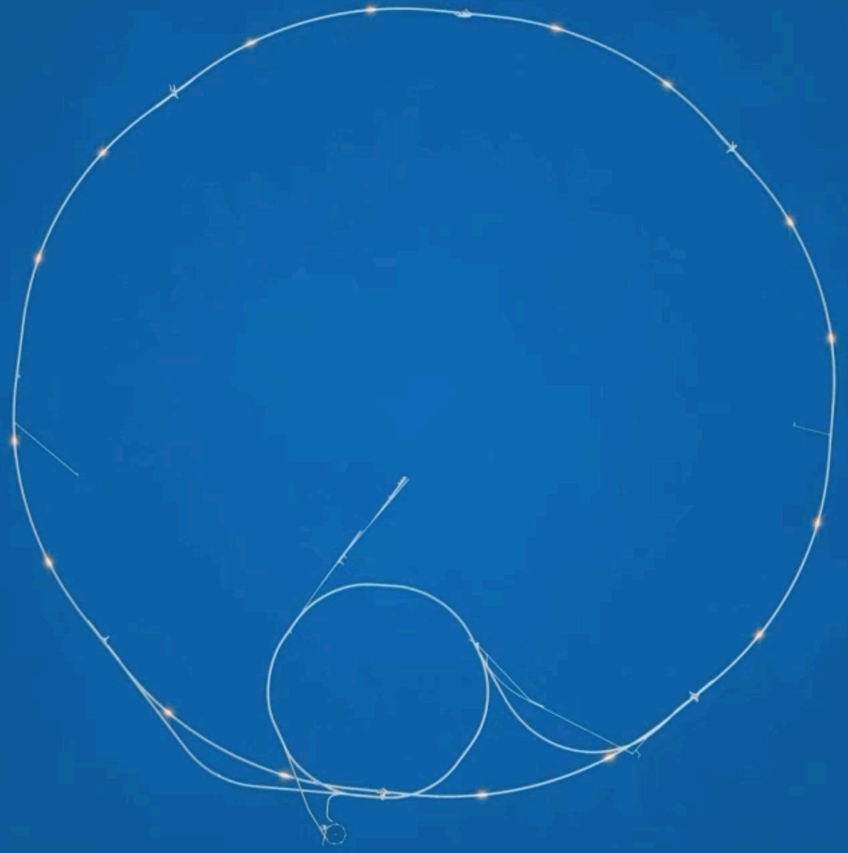

ATLAS
Data
Background ZZ(*)
Background Z+jets, tt̄
Syst.Unc.
√s = 7 TeV: ∫Ldt = 4.8 fb-1
√s = 8 TeV: ∫Ldt = 5.8 fb-1
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m4l [GeV]


ATLAS
Data
Background ZZ(*)
Background Z+jets, tt̄
Signal (m_H=125 GeV)
Syst.Unc.
√s = 7 TeV: ∫Ldt = 4.8 fb-1
√s = 8 TeV: ∫Ldt = 5.8 fb-1
Events/5 GeV
0
5
10
15
20
25
100
150
200
250
m4l [GeV]

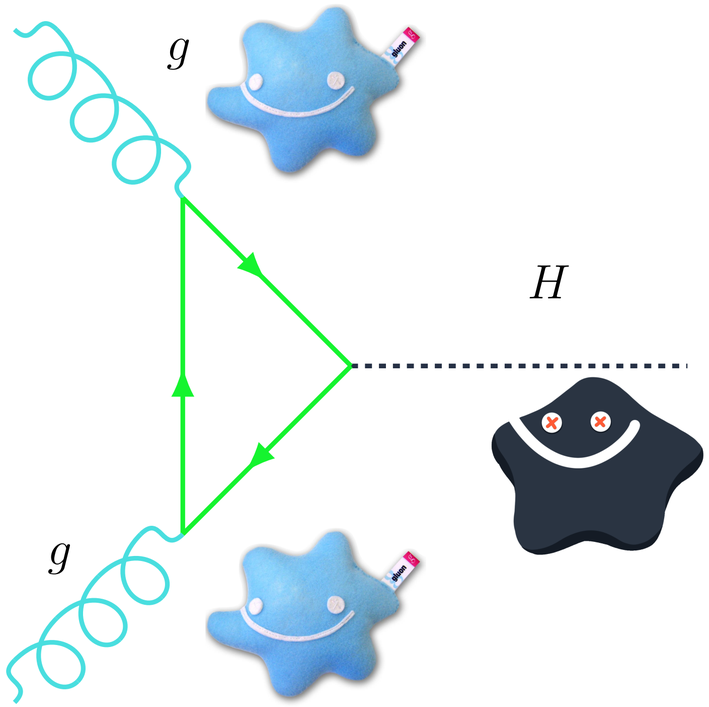


UNIVERSITY OF SUSSEX

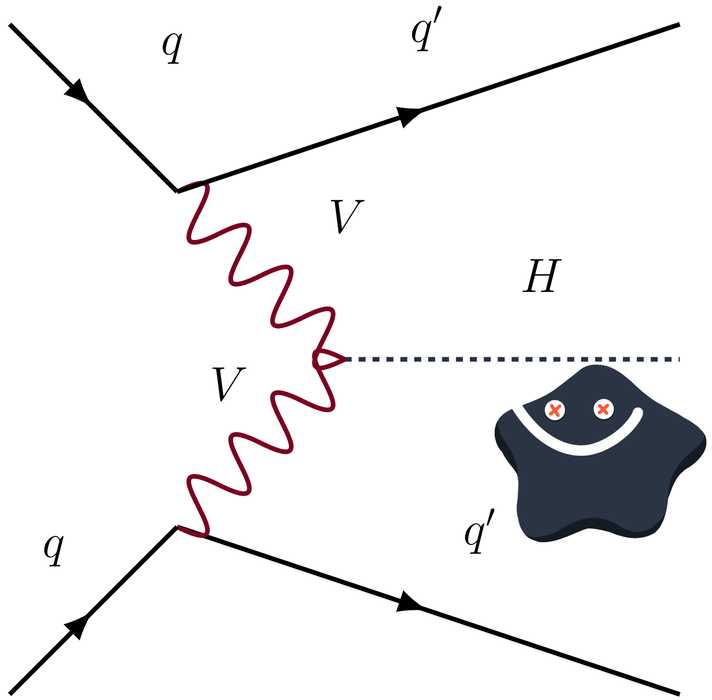
open data ATLAS



Higgs production

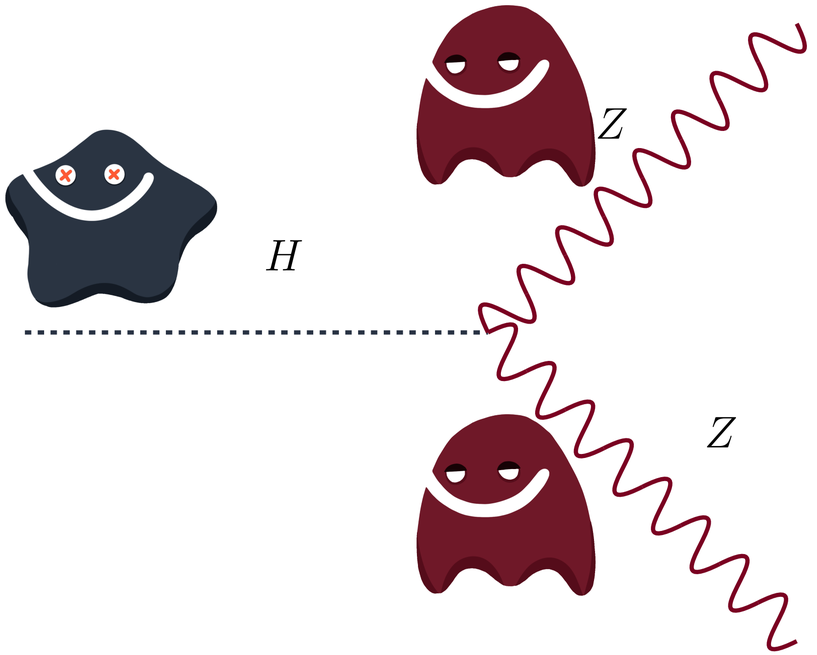


gluon-gluon (gg)

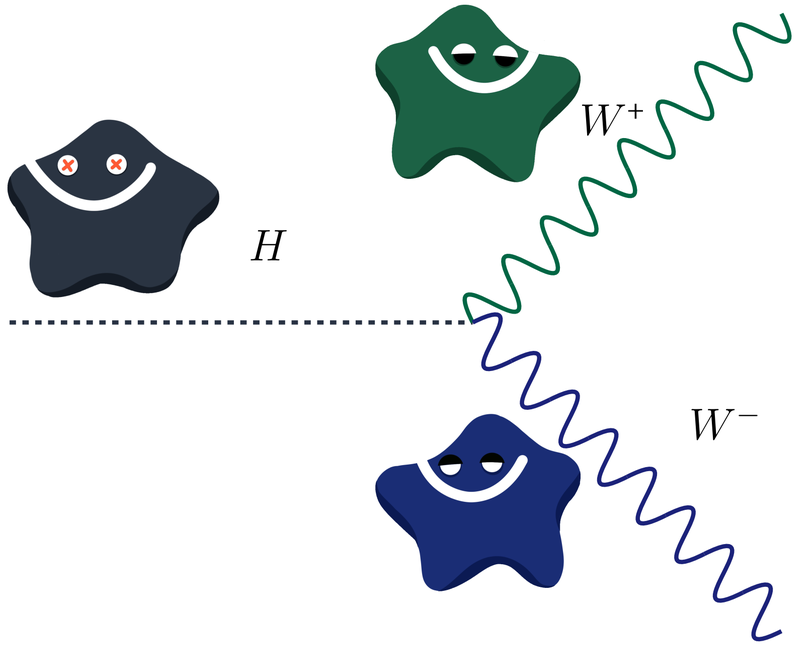


Vector Boson Fusion (VBF)

Higgs decay

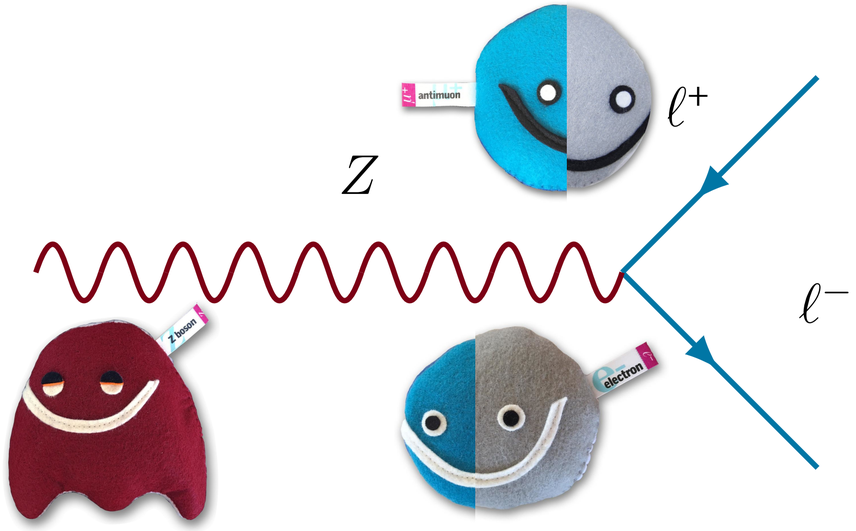


H → ZZ

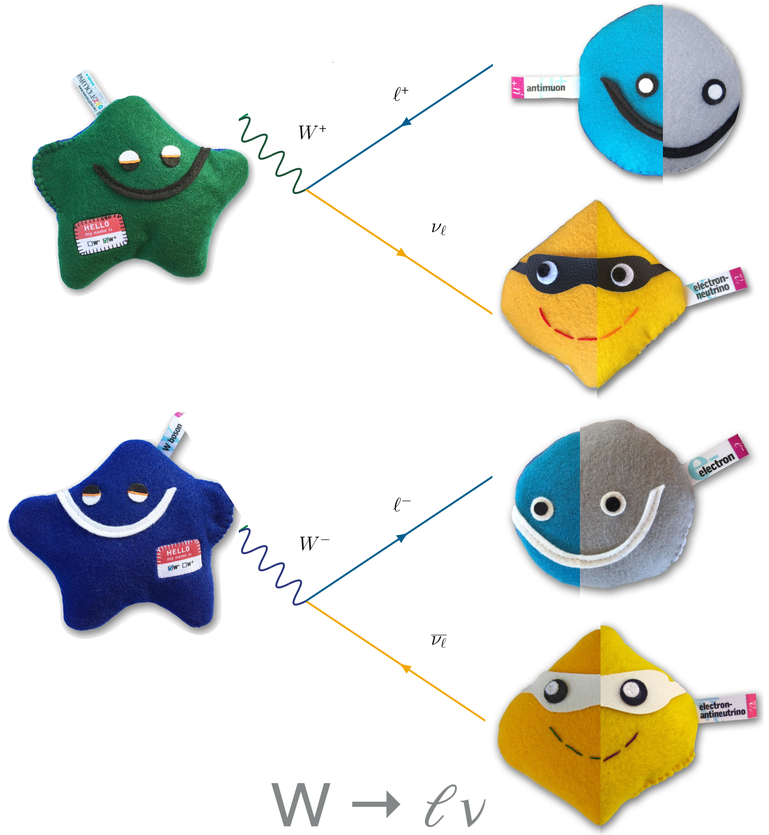


H → WW

Z & W decay



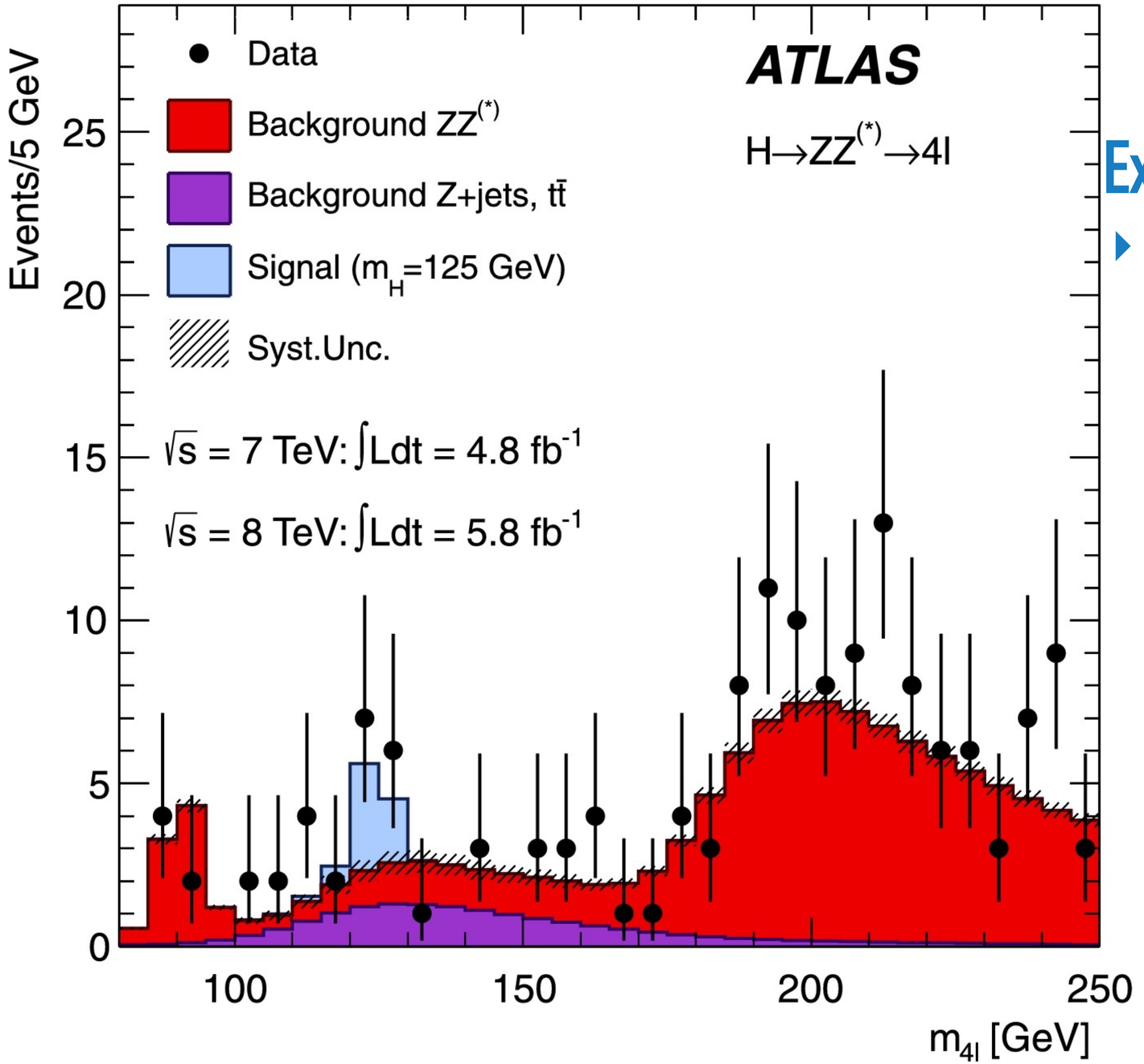
$Z \rightarrow \ell\ell$



$W \rightarrow \ell\nu$

Final state particles

- Of the particles we've talked about, only charged leptons ($\ell^{\pm}$) are directly measured by ATLAS
- Higgs, Z, W bosons have lifetimes < 10^{-22} s
  - They decay before being measured
  - They're measured indirectly by their final state particles

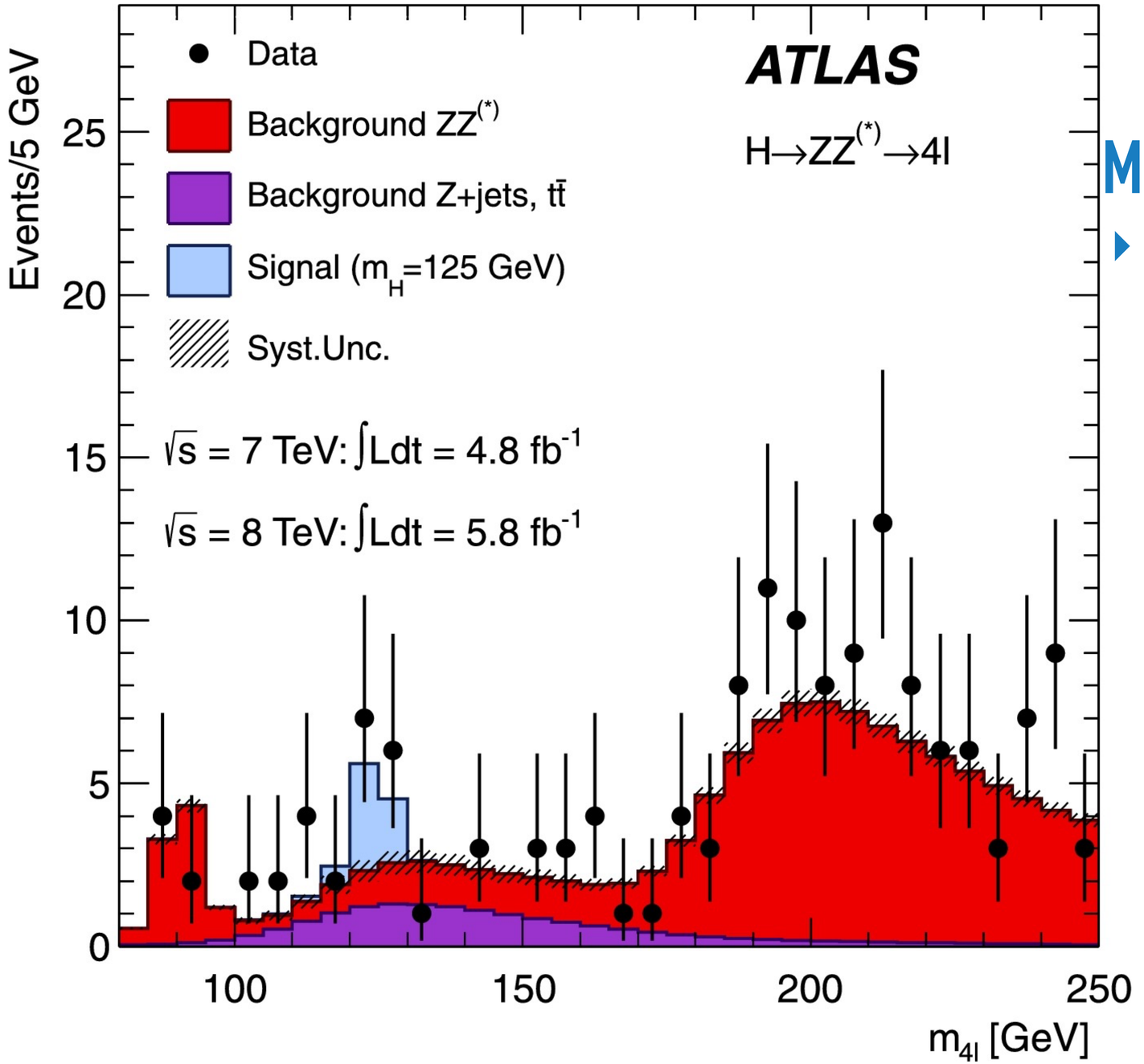


Experimental data

- Black points represent experimental data
- Statistical uncertainty represented by error bars on the data points

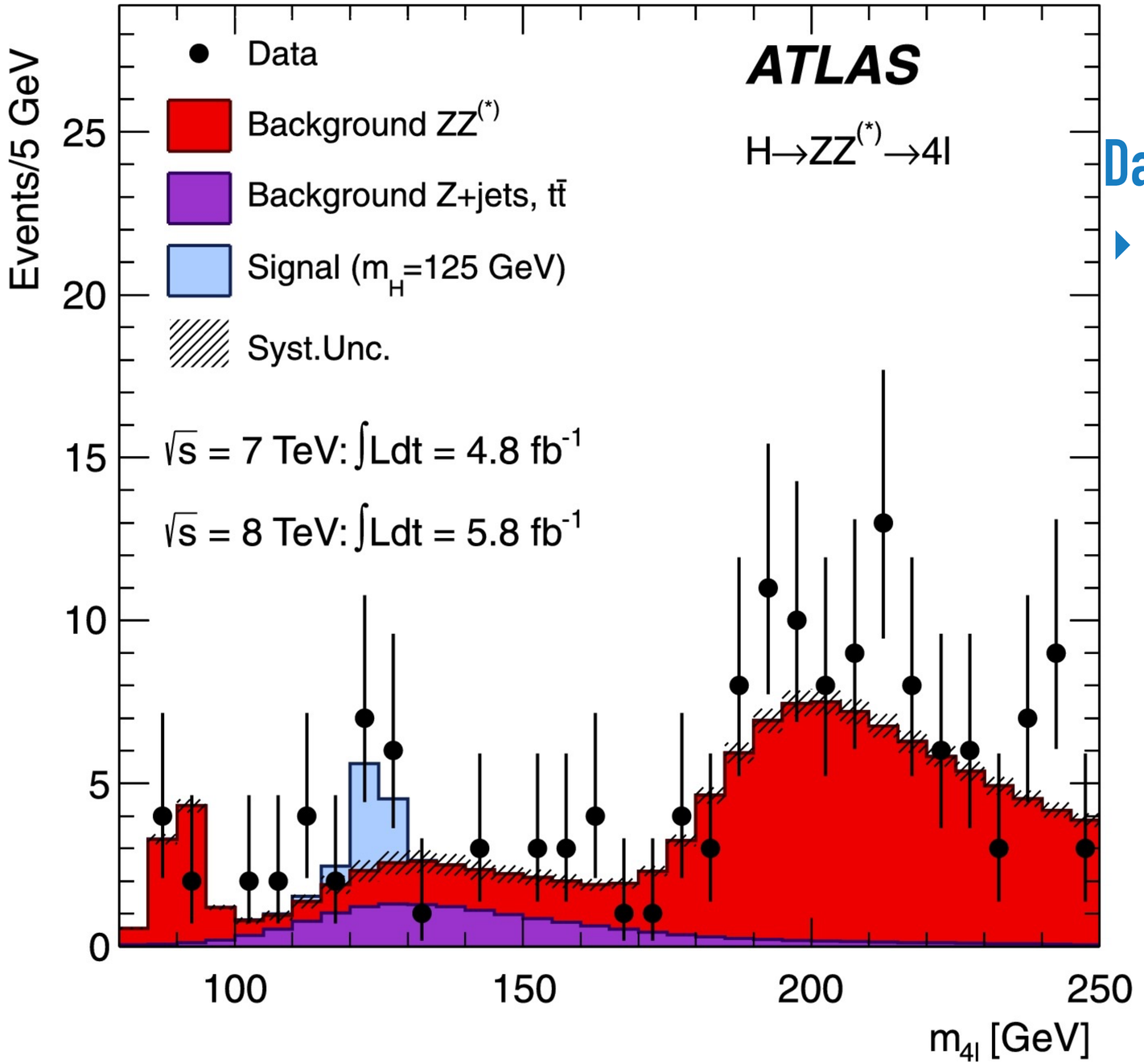
UNIVERSITY OF SUSSEX

open data ATLAS



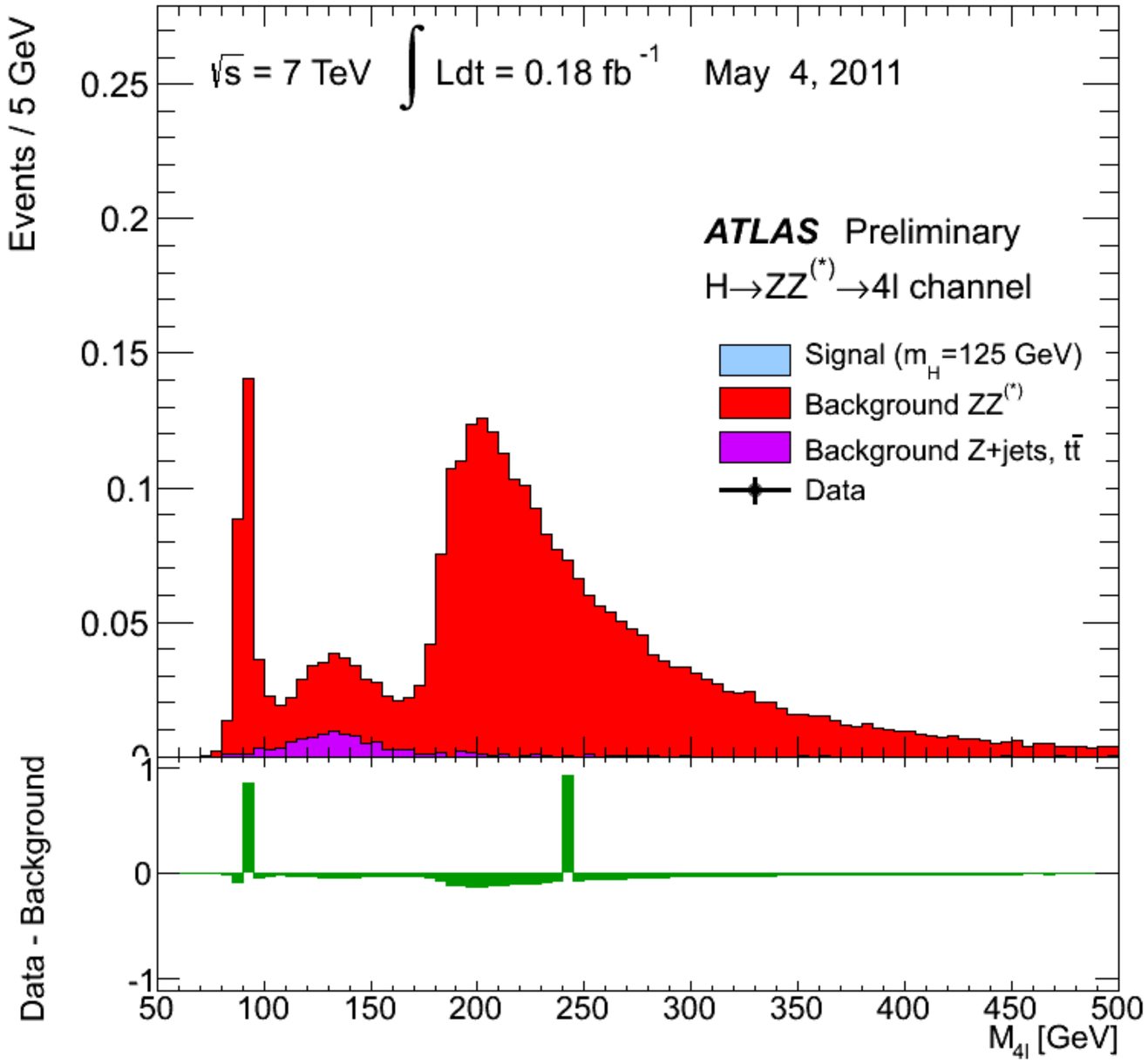
MC simulation

- Filled histograms show the prediction from different MC simulations
- The contributions are stacked

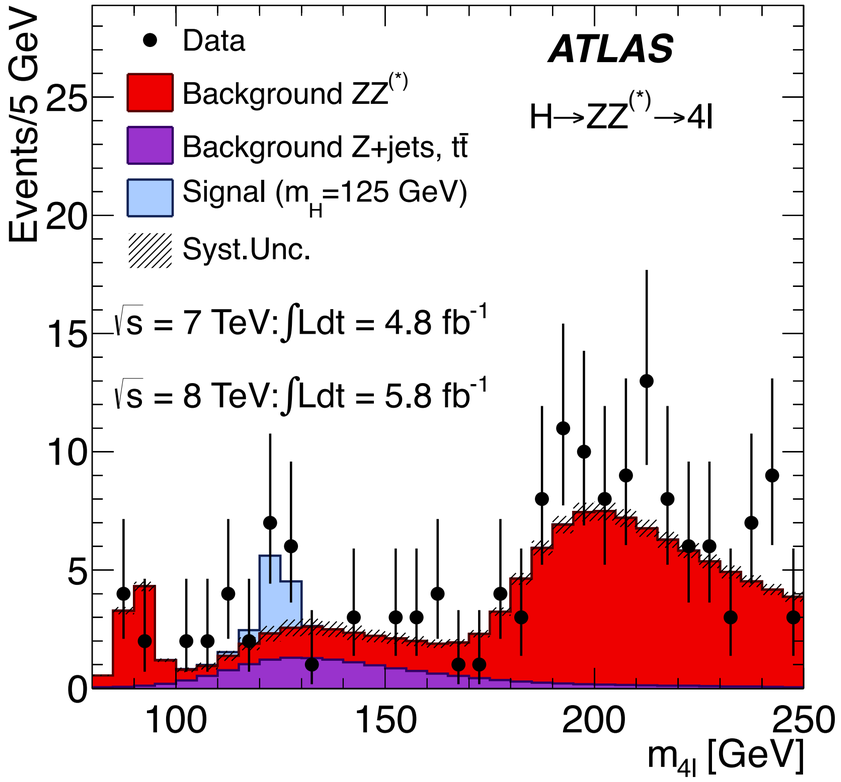


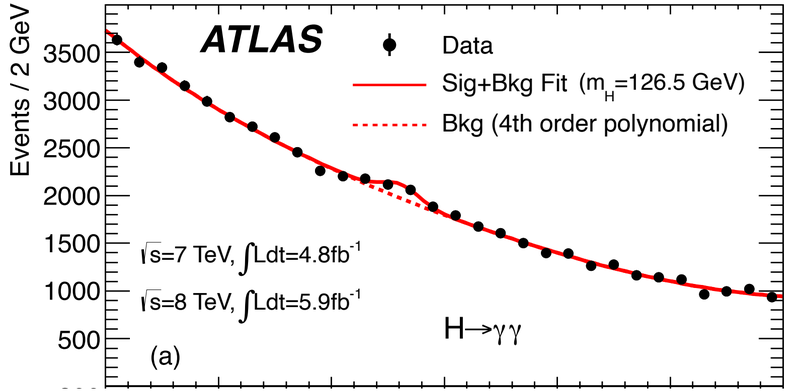
Data/MC discrepancy

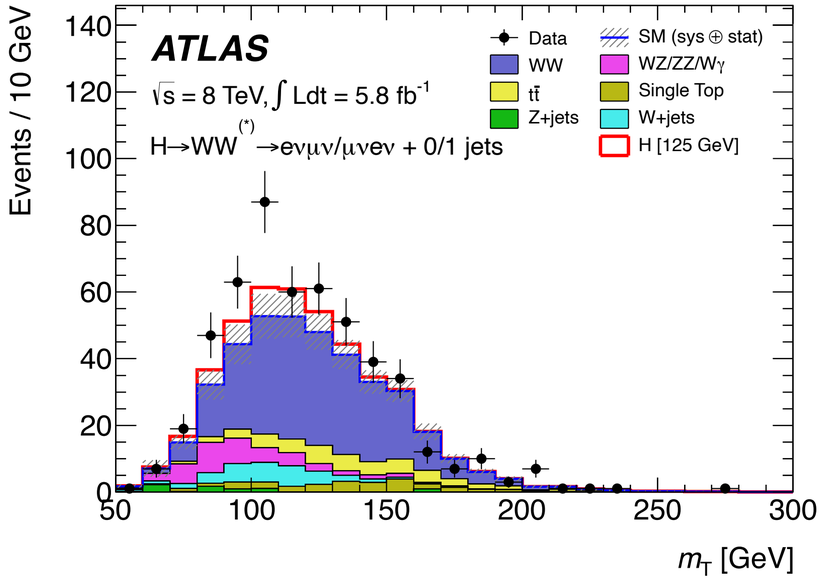
- Any discrepancy between Data & MC is statistical
  - Like line-of-best-fit only passing through 67% of 1σ error bars


√s = 7 TeV ∫ Ldt = 0.18 fb-1 May 4, 2011
ATLAS Preliminary
H→ZZ(*)→4l channel
Signal (mH=125 GeV)
Background ZZ(*)
Background Z+jets, tt̄
Data
Events / 5 GeV
Data - Background
M4l [GeV]

How did this lead to a discovery?





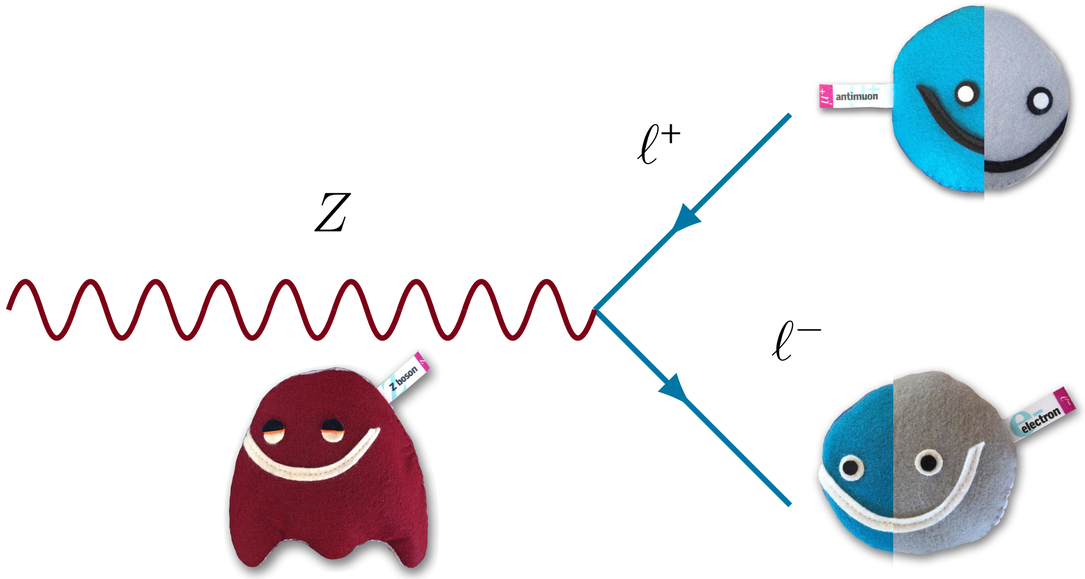


- Combining H → ZZ, H → $\gamma\gamma$, H → WW, gave total significance of 5σ in data over background-only

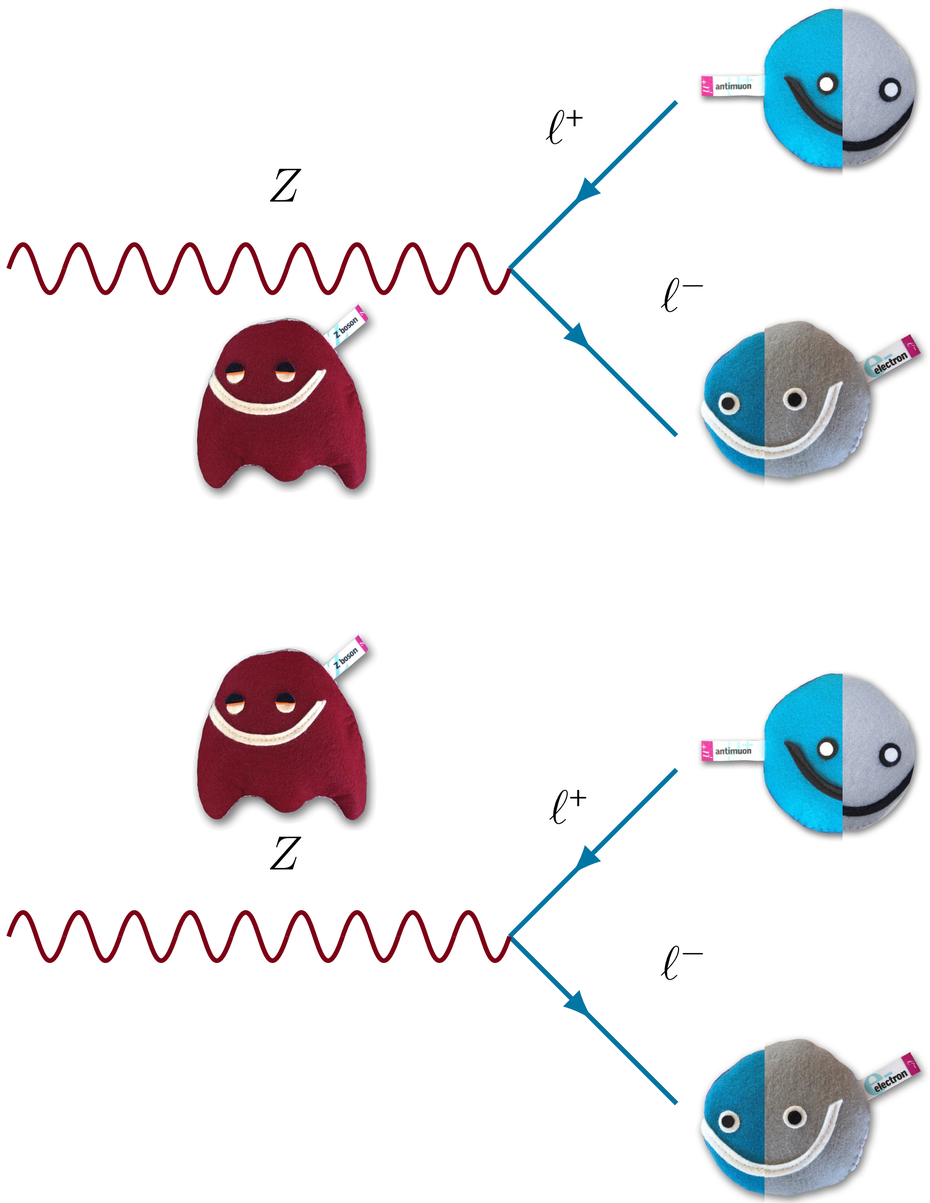
UNIVERSITY OF SUSSEX

open data ATLAS

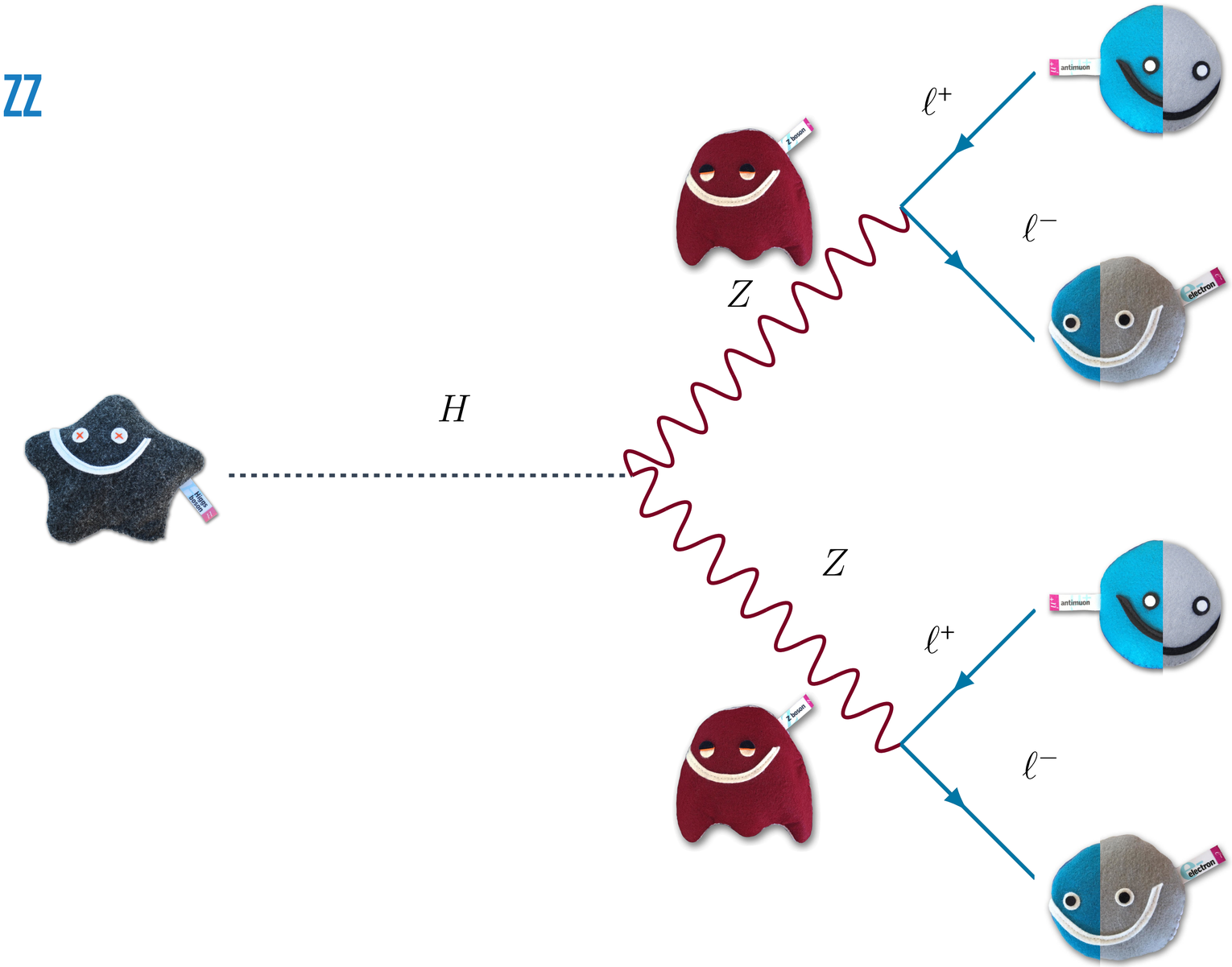
Z



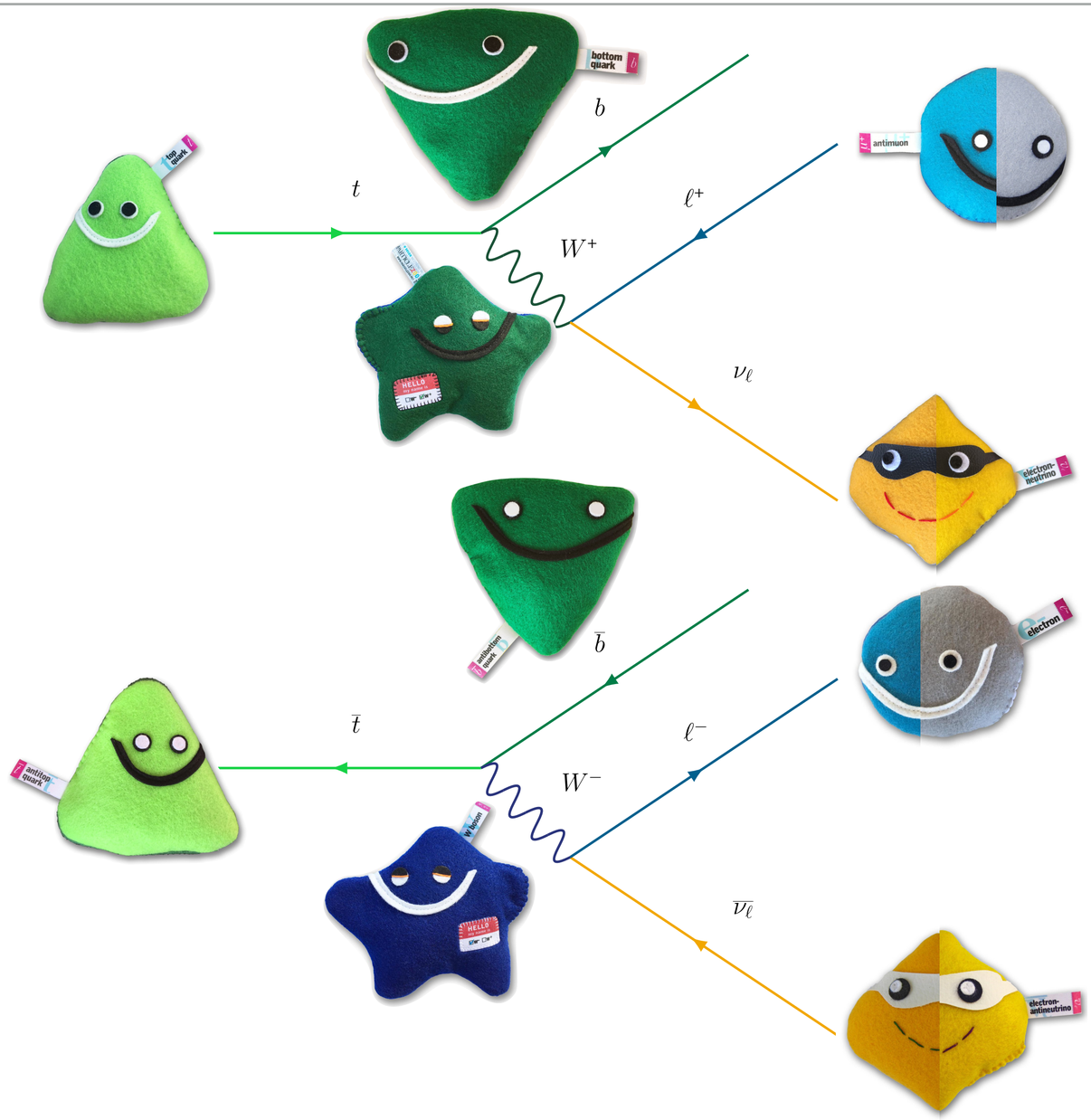
ZZ


ℓ^+
Z
ℓ^-
antimuon
Z boson
electron
Z boson
ℓ^+
Z
ℓ^-
antimuon
electron

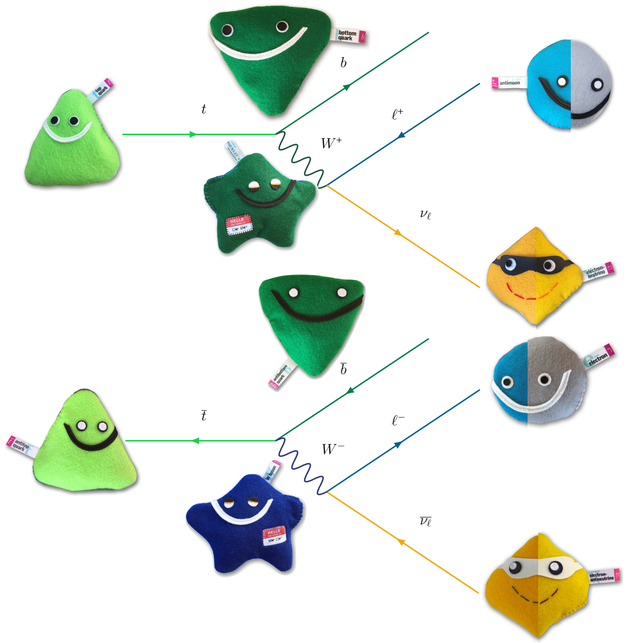
H → ZZ



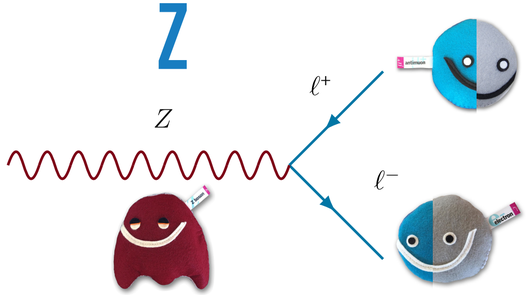
$t\bar{t}$


bottom quark
b
t
ℓ^+
antimuon
W^+
ν_ℓ
electron-neutrino
antibottom quark
$\bar{b}$
electron
$\bar{t}$
ℓ^-
W^-
$\overline{\nu_\ell}$
electron-antineutrino

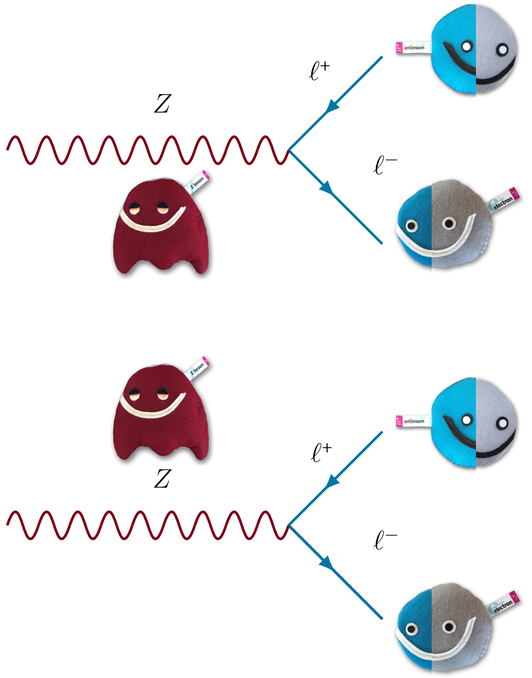
$t\bar{t}$



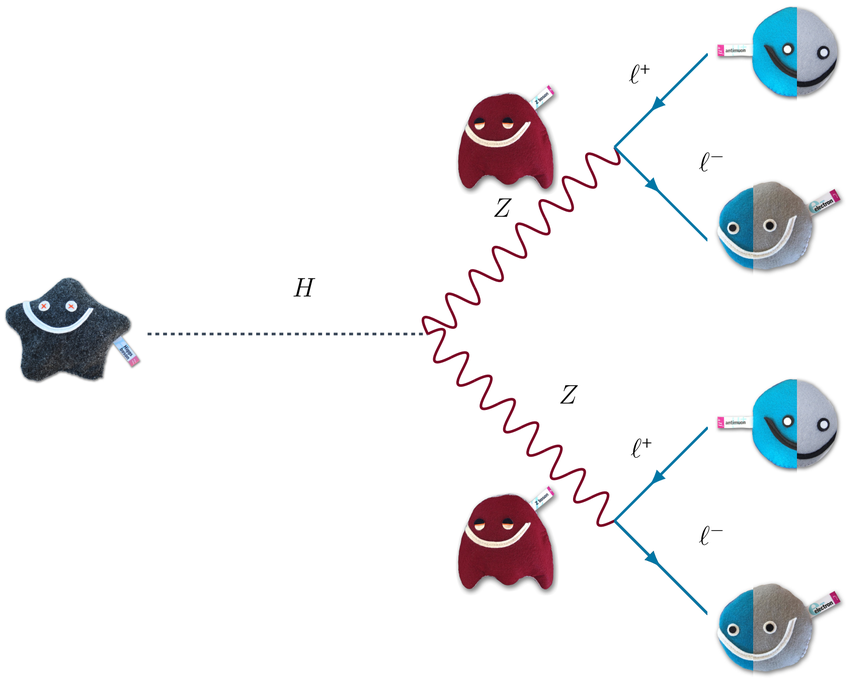
Z



ZZ



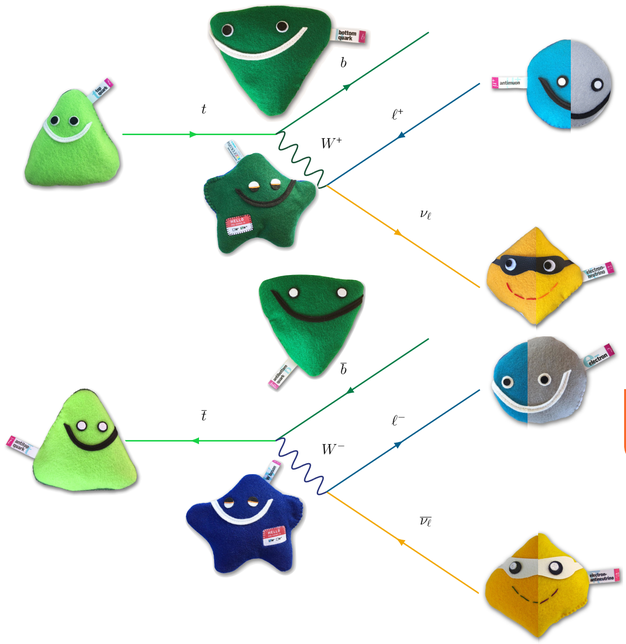
H → ZZ



UNIVERSITY OF SUSSEX

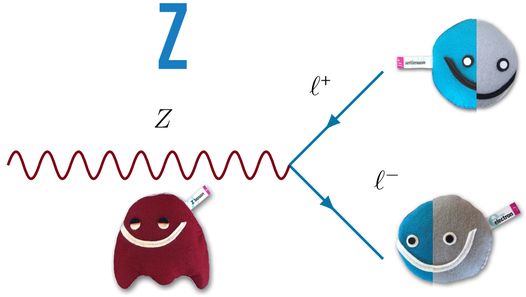
open data ATLAS

$t\bar{t}$



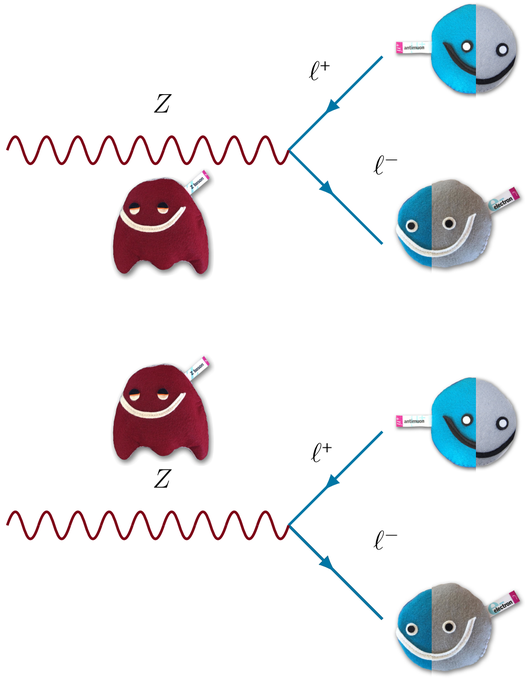
2 leptons

Z

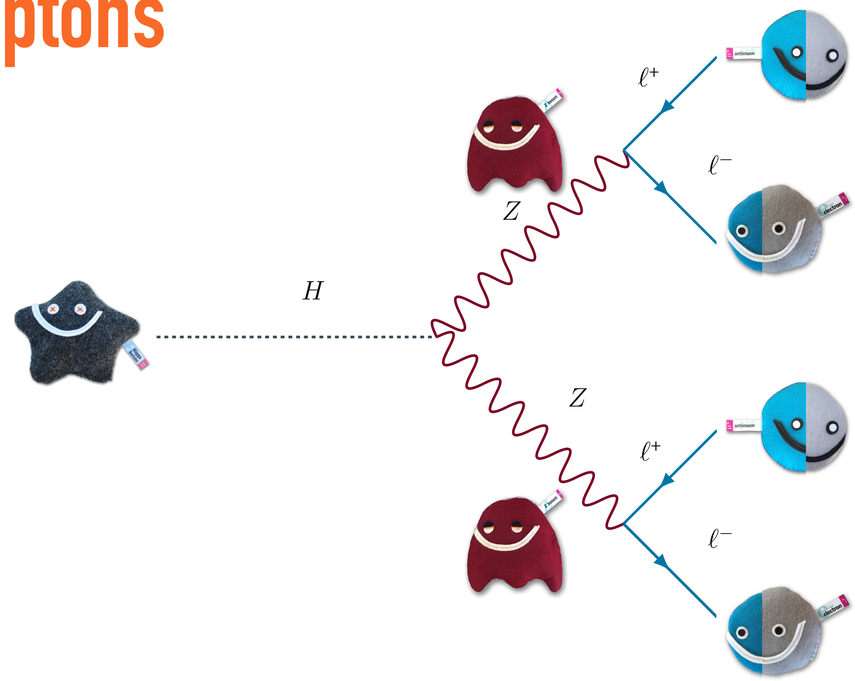


4 leptons

ZZ



H → ZZ



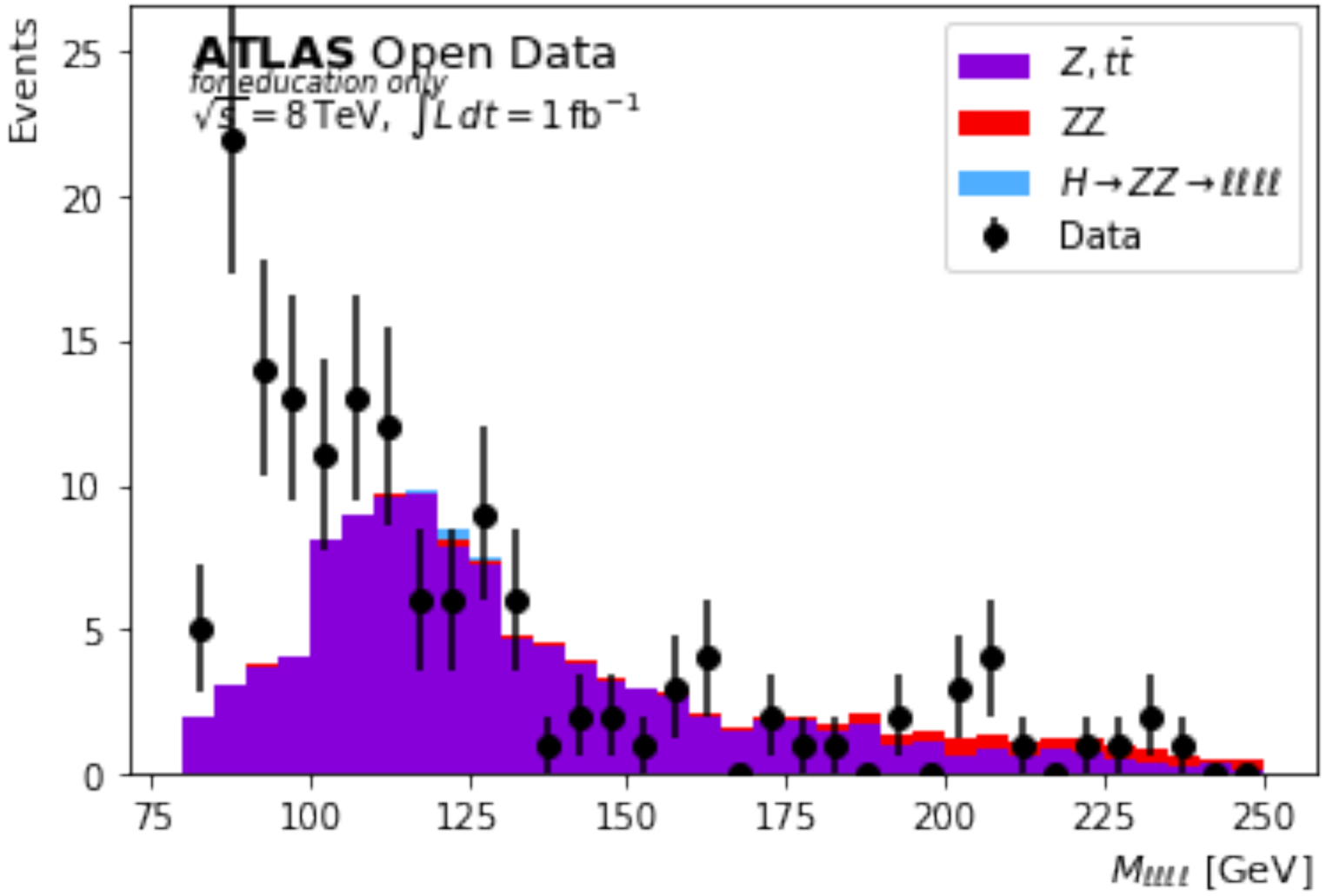
UNIVERSITY OF SUSSEX

open data ATLAS

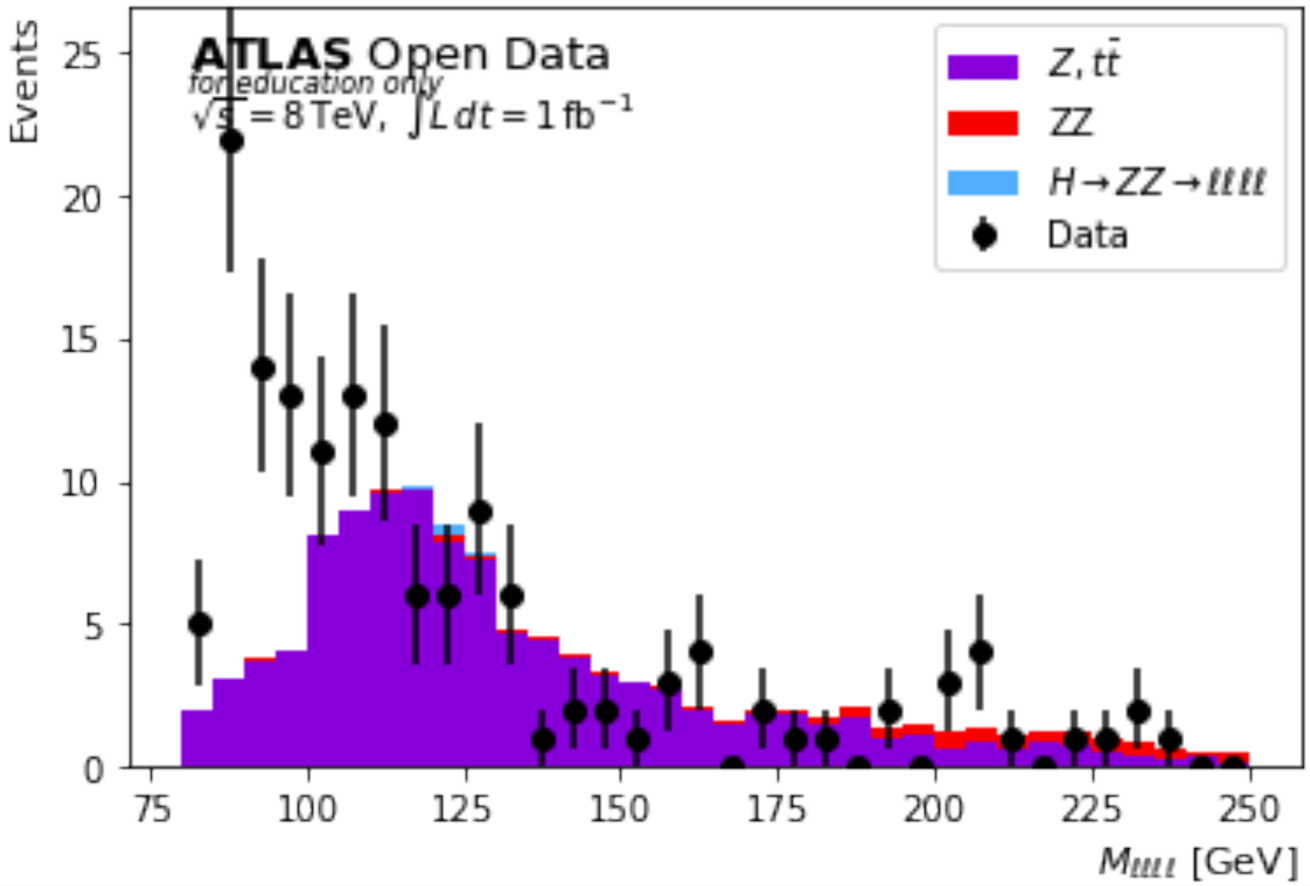
Run the template analysis

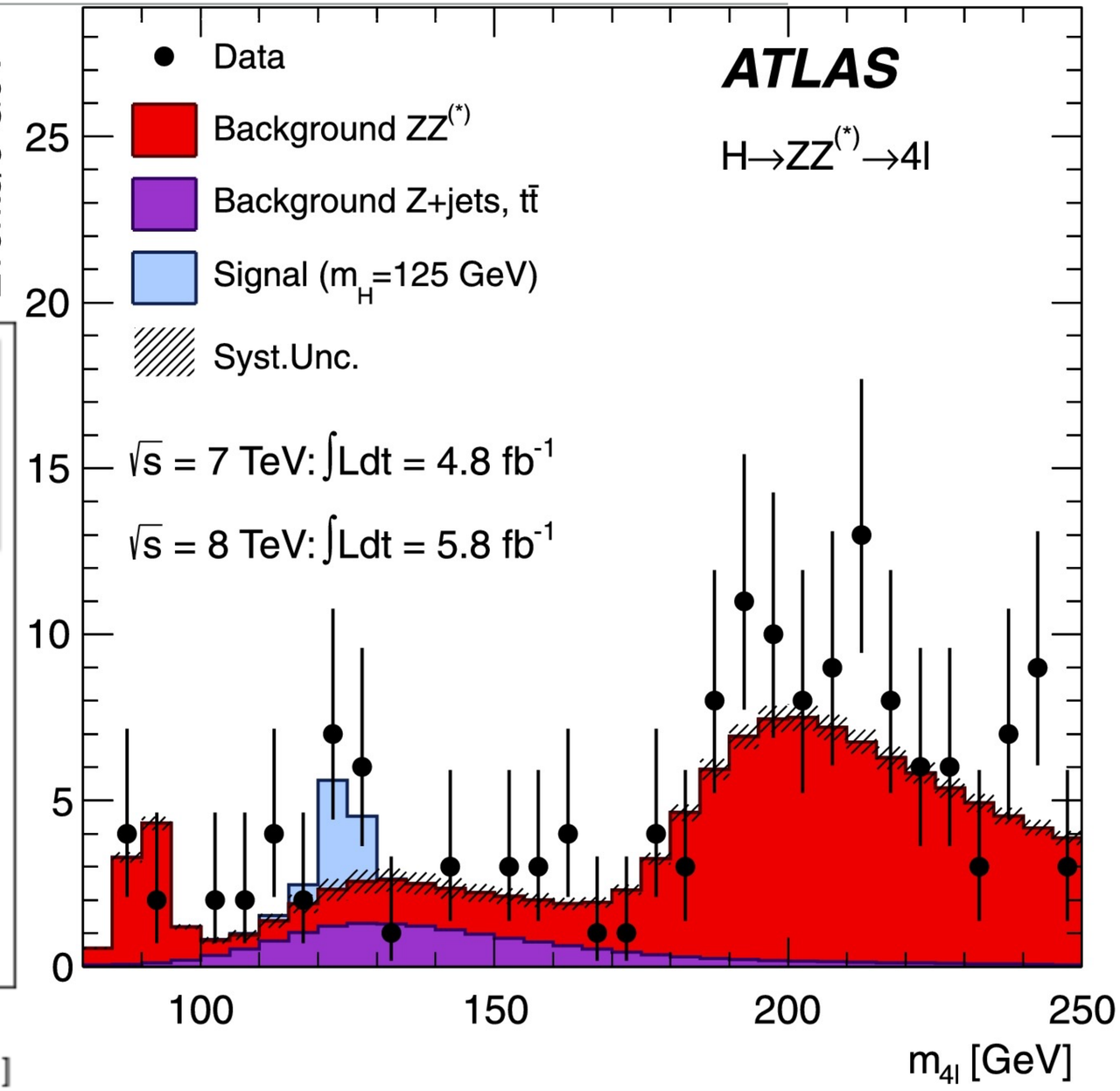
- Go to github.com/meevans1/How-to-rediscover-the-Higgs
- Follow the instructions to open the Jupyter notebook
- Cell -> Run All
- Let us know if you have any problems

Output from template analysis



Output from template analysis





Cut on more variables

- Variables described on the ATLAS Open Data website
- Try add cuts from the Higgs discovery paper to improve the signal / background ratio
- We'll be updating this live google doc with hints
- We haven't given you much info…
- But the idea is that you try things out yourself, discuss and ask questions!

UNIVERSITY OF SUSSEX

open data ATLAS

Implement more cuts

- What results did you get?
- How does the new result compare with previous results?
- How would you measure the success of these cuts?
- How might you improve?

Significance

- $\frac{\text{measurement}}{\text{error}}$
- Quote this as a number of "σ"
  - (Not to be confused with a 1σ error bar!)
- Error can usually be estimated using Poisson (square root)
- What measures of significance do you use in your work?

Significance measures

- $\frac{N_{\text{signal}}}{\sqrt{N_{\text{background}}}}$
- $\frac{N_{\text{data}} - N_{\text{background}}}{\sqrt{N_{\text{background}}}}$
- $\frac{N_{\text{signal}}}{\sqrt{N_{\text{signal}} + N_{\text{background}}}}$
- $\frac{N_{\text{data}} - N_{\text{background}}}{\sqrt{N_{\text{signal}} + N_{\text{background}}}}$
- $\frac{N_{\text{data}} - N_{\text{background}}}{\sqrt{N_{\text{data}}}}$
- $\sqrt{2((N_{\text{signal}} + N_{\text{background}})\ln(1 + \frac{N_{\text{signal}}}{N_{\text{background}}}) - N_{\text{signal}})}$

If $N_{\text{background}}$ known

- $$\frac{N_{\text{signal}}}{\sqrt{N_{\text{background}}}}$$
- Profile likelihood ratio test

$$\sqrt{2((N_{\text{signal}} + N_{\text{background}})\ln(1 + \frac{N_{\text{signal}}}{N_{\text{background}}}) - N_{\text{signal}})}$$

- Useful resource: [statistics presentation by Glen Cowan]

UNIVERSITY OF SUSSEX

open data ATLAS

Statistical fits

- Statistical fits are applied to the signal and background distributions to differentiate between them
- Starting point is Gaussian fit of signal

Optimisation

- Cuts used in Higgs discovery paper will have been optimised for that dataset
- We only have a subset (~10%) of that dataset
  - So the optimised cuts for our dataset might be different
- Modify some cuts and see what happens
- How might you optimise your cuts?

Optimisation clues

- Want optimum significance measure as a function of the applied cut
- x axis: applied cut value
- y axis: significance measure
- Find peak of distribution

Optimisation explanation

- Very loose cut will keep all signal but also all background
  - so ~constant value of significance
- Tightening the cut will start rejecting background
  - so significance starts increasing
- Tightening the cut too much will start rejecting signal
  - so significance starts decreasing
- There must exist an optimal cut

Machine Learning

- Recently, ML techniques have been playing an increasing role in optimising for signal / background ratio
  - Support Vector Machine (SVM)
  - Neural Network (NN)
  - Boosted Decision Tree (BDT)
  - Random Forest

Machine Learning optimisation

- Cut to be optimised can be the output of an ML technique
- So same optimisation principle applies if ML technique was used to separate signal and background

Hands on

- Implement examples
- Improve with new ideas
- What were your results now?
- How do they compare with others?
- If you could, how would you improve further?

Contribute to ATLAS Open Data

- Submit your notebook as a contribution to ATLAS Open Data
- If successful, your notebook will modified to use for the upcoming release of new 13 TeV data
- Instructions on the GitHub code page on how to enter

Judgement criteria

- In this order:
  - Some nice Machine Learning
  - Elegant code
  - Number of cuts correctly implemented
  - Reduce total elapsed time
  - Beautiful plots
  - Optimise for $\text{Signal}/\sqrt{\text{Background}}$

UNIVERSITY OF SUSSEX

open data ATLAS

Thanks!

- From this session we hope you've got an insight into:
  - How Python, Pandas, Numpy & Matplotlib can be used for (really) big data analysis
  - How interesting (& challenging) particle physics can be
  - How you can go onto rediscover the Higgs boson
  - How to (re-)win a Nobel Prize
  - How these techniques can be applied to your work!

UNIVERSITY OF SUSSEX

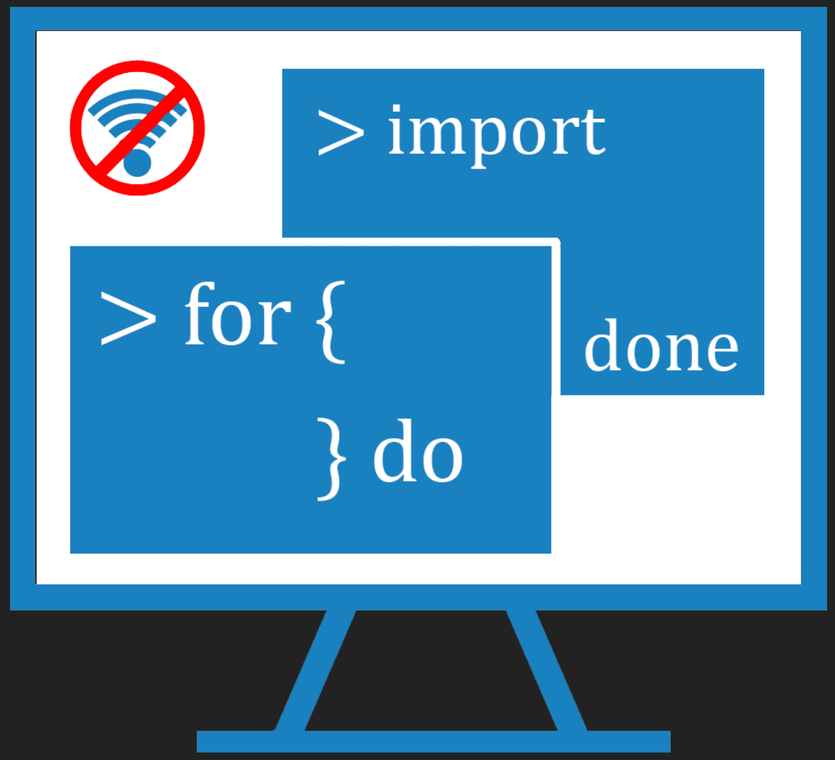
open data ATLAS

Feedback questionnaire

- This is our first time running such a workshop
- We'd love to hear your ideas / suggestions / comments on how to improve
- If you could fill in our questionnaire, it'd be much appreciated!
- Only a few minutes of your time!

UNIVERSITY OF SUSSEX

open data ATLAS



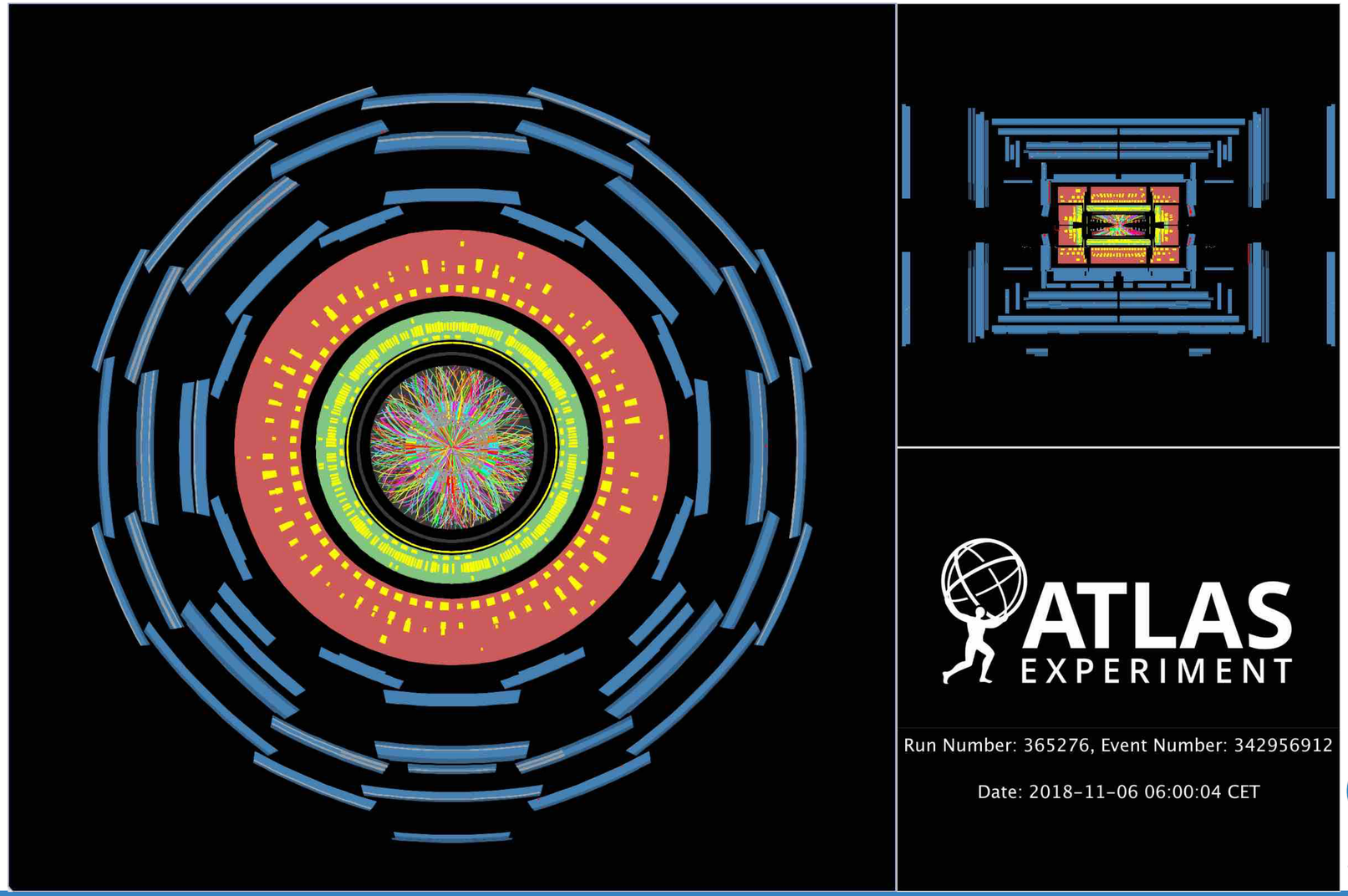
Meirin Oan Evans, Kate Shaw, Tom Stevenson

Data Intensive, AI & ML Summer School, Sussex

22nd July 2019

Backup

Transverse momentum? Why not momentum?

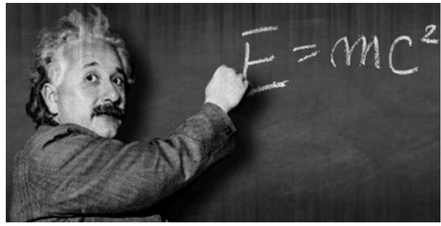


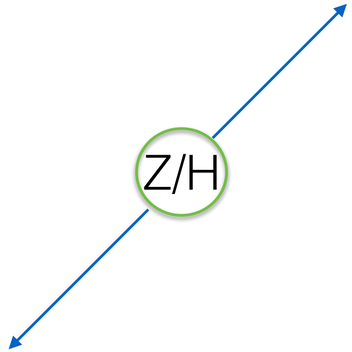
UNIVERSITY OF SUSSEX

open data ATLAS

Transverse momentum? Why not momentum?

- Z and Higgs bosons are ~100x heavier than protons
  - ➡ ~all proton momentum goes to making Z/Higgs rest mass
  - ➡ Z/Higgs produced almost at rest
  - ➡ decay products move ~back-to-back in random direction
  - ➡ products often oriented in transverse plane
- Leptons are ~1000x lighter than Z/Higgs

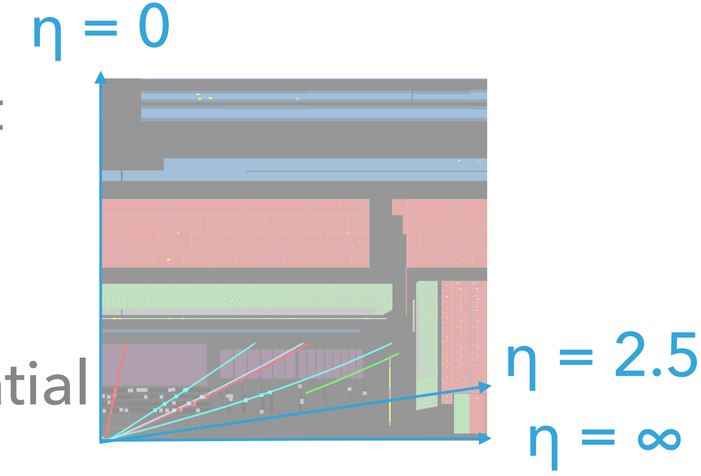




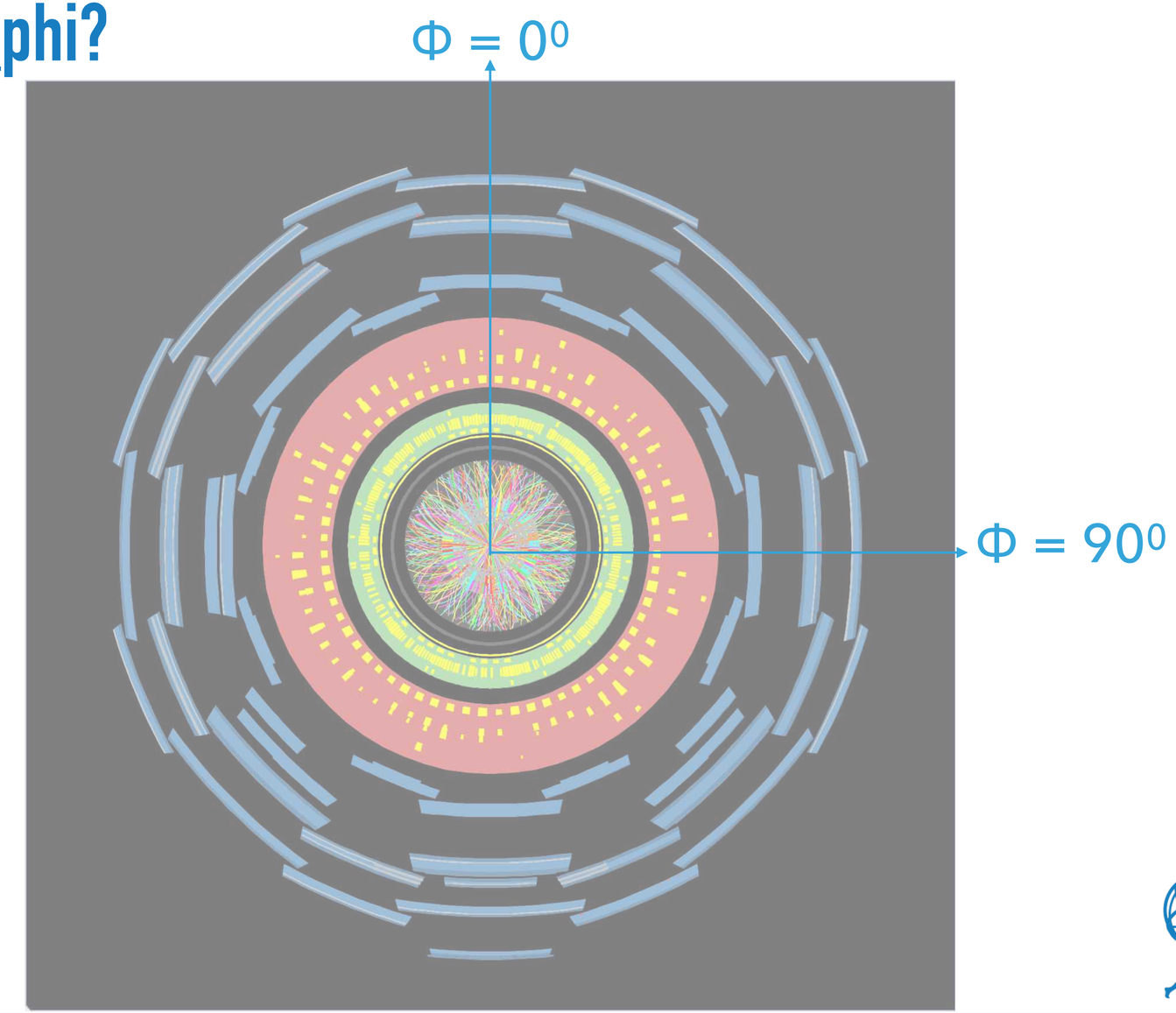
Pseudorapidity? Why not θ?

- $lep\_eta = -\ln\tan(\frac{\theta}{2})\,.$
- Differences in pseudorapidity are Lorentz invariant
- Differences in θ aren't

➡ difference in pseudorapidity gives handle on spatial separation

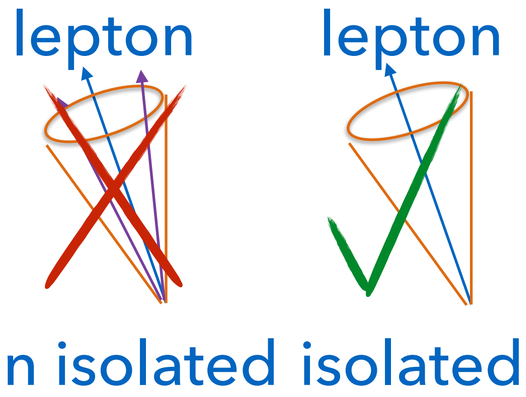


What's lep_phi?



What's lep_etcone20?

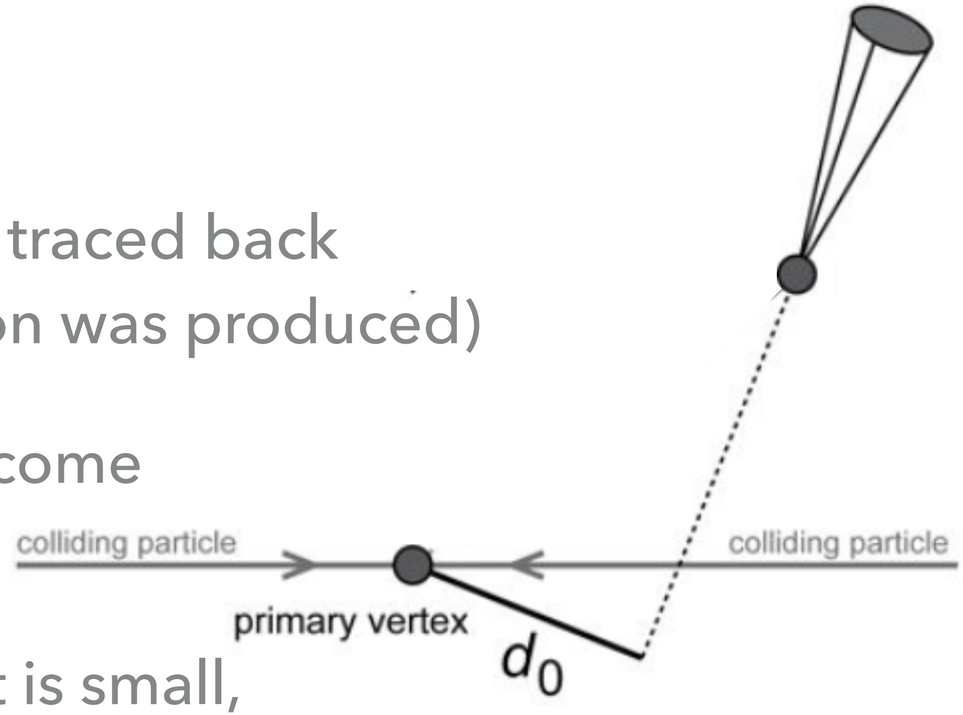
- A measure of the energy in a cone around the lepton, not including the lepton



- Leptons from boson decays are more likely to be isolated
- But we don't want to throw away leptons with really high pt, even if they aren't completely isolated
- So throw away leptons with etcone20/pt > 0.3 (e.g.)

What's lep_d0?

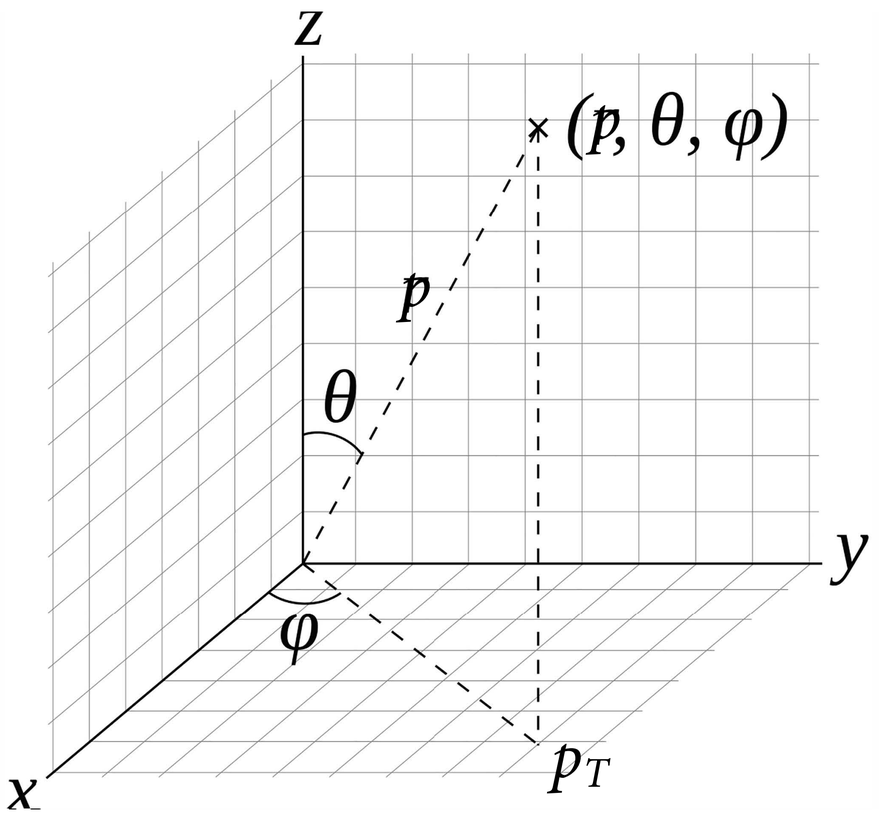
- Trace back the track of the lepton
- Take the minimum distance between the traced back track and primary vertex (where the boson was produced)
- Want to throw away leptons that haven't come from the primary vertex (large d0)
- If the significance of the d0 measurement is small, we can be confident of our d0 measurement
- So cut away leptons with d0/sigd0 > 6.5 (e.g.)



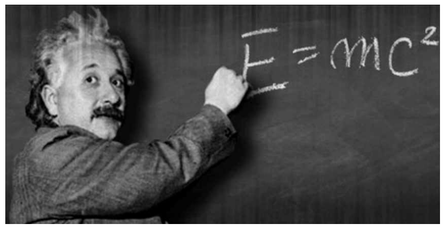
Momentum components

- $p_x = p_T \cos\phi$
- $p_y = p_T \sin\phi$
- $p_z = p\cos\theta, p_T = p\sin\theta$

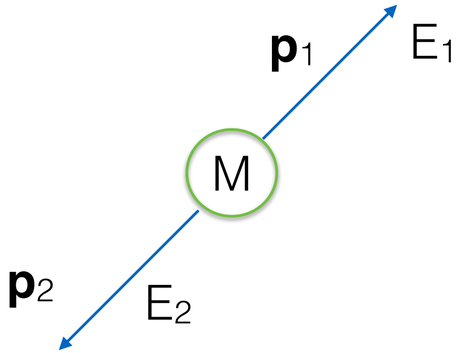
  ➡ $p_z = \dfrac{p_T}{\sin\theta}\cos\theta$



Invariant mass



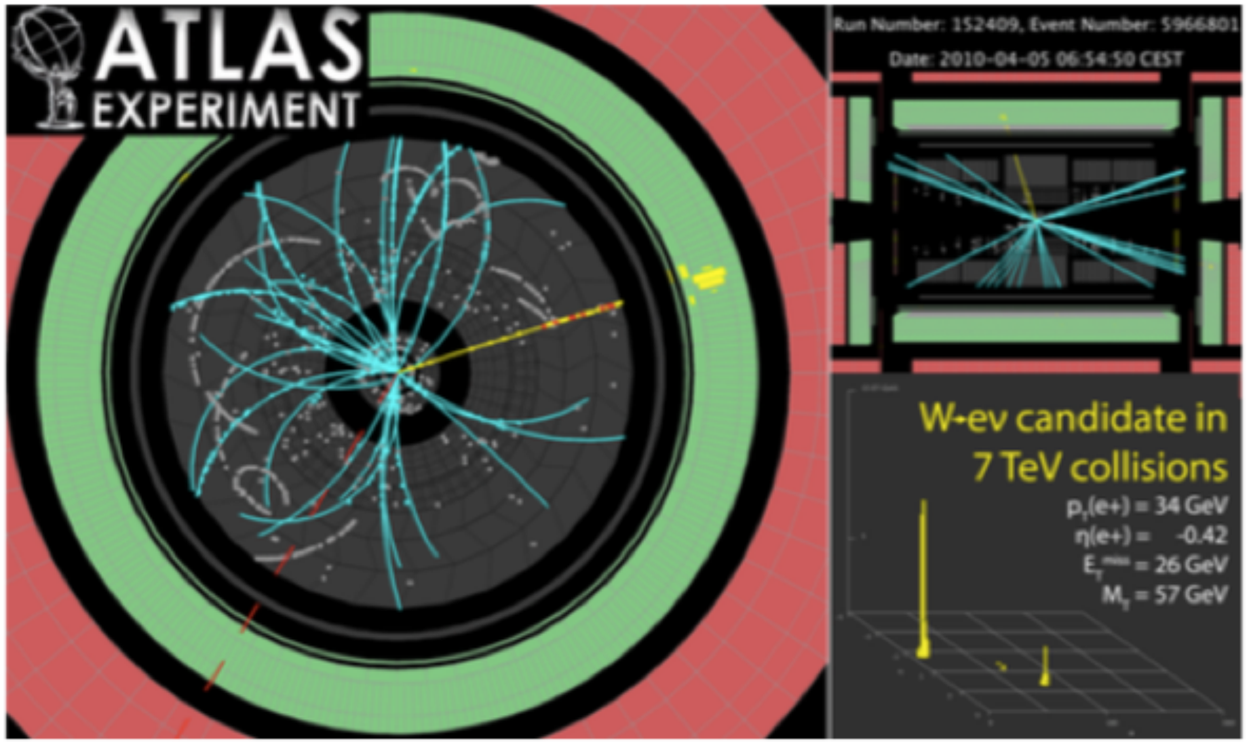
- $E^2 = p^2 + M^2$
- $M = \sqrt{E^2 - p^2}$
- $M = \sqrt{(E_1 + E_2 + \dots)^2 - (\overrightarrow{p_1} + \overrightarrow{p_2} + \dots)^2}$



- Use final state leptons to find invariant mass of the boson from which they decayed
- Throw away if invariant mass very different to prediction

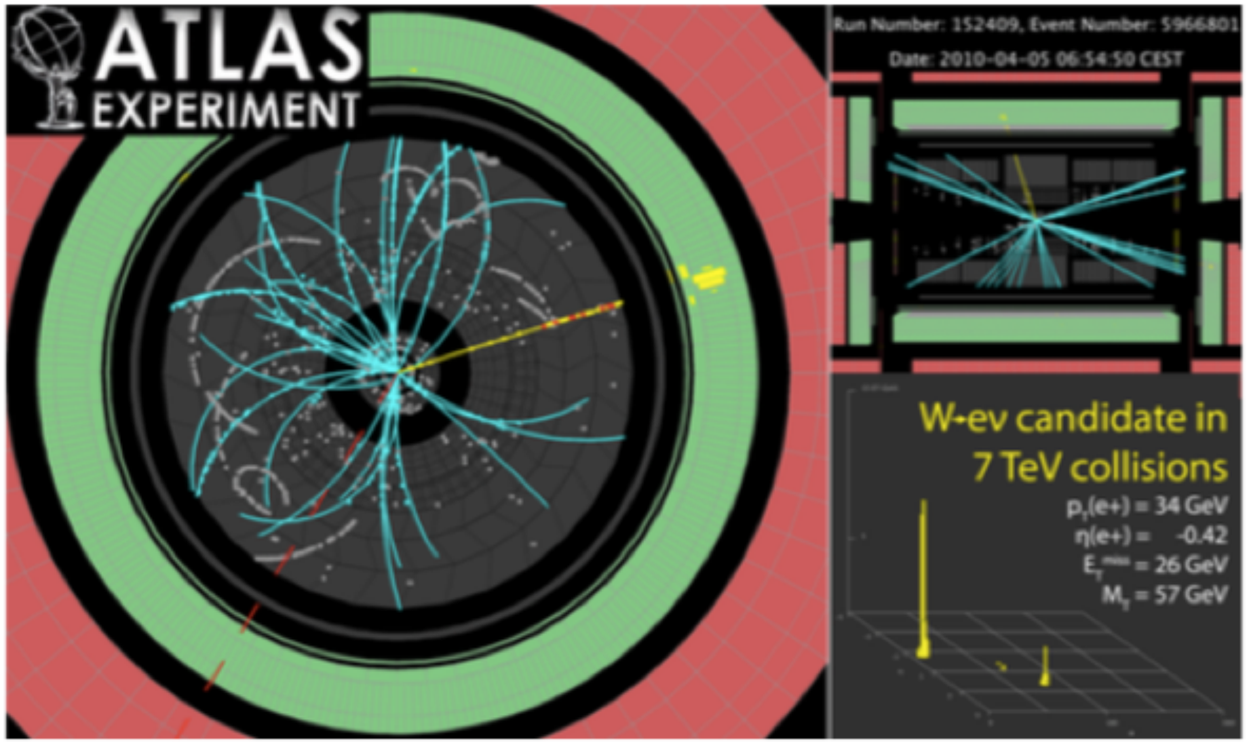
$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$$

- Don't want the tracks of leptons to overlap
- If they did overlap, it'd be hard to say whether it was 1 or 2 tracks
- So throw away leptons that are very close together in ΔR



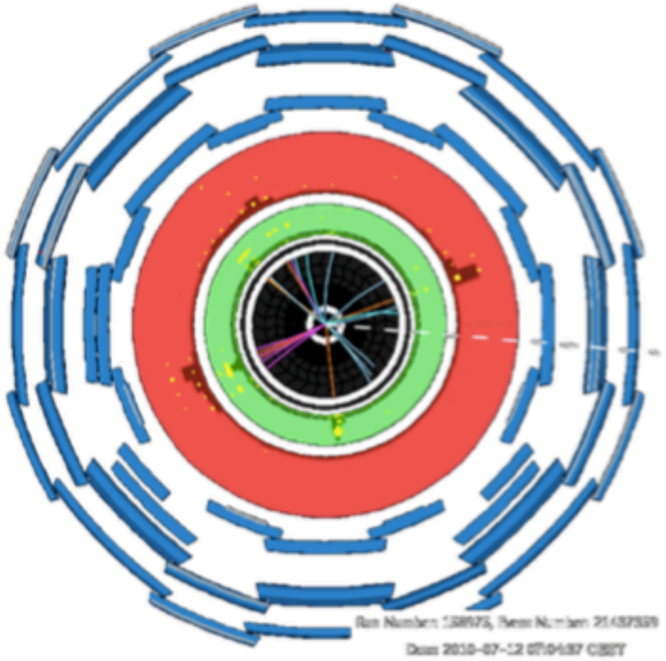
Why a stricter ΔR for leptons of same type?

- Electrons are detected by their energy deposited
- If electrons are really close together, you're not sure whether you're detecting 2 electrons, or a single high-energy electron
- Similar for muons



Why a lower p_T threshold for muons?

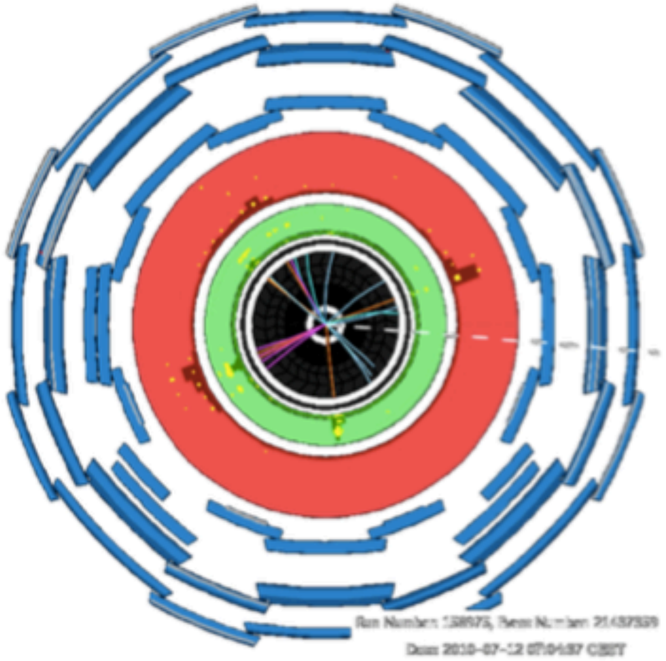
- Muons are the only particles detected in the outer part of ATLAS
- So if something is detected here, high chance it's a muon



- **Inner Detector:**
  - Tracks the path of charged particles.
  - Magnetic field bends the path of charged particles.
- **Electromagnetic Calorimeter:**
  - Particles that interact with the electromagnetic force leave energy deposits.
  - Absorbs Electrons and Photons.
- **Hadronic Calorimeter:**
  - Jets (from quarks) are absorbed.
- **Muon Spectrometer:**
  - Tracks the path of Muons.
  - Magnetic field bends the path of Muons.

Why a lower etcone20 threshold for electrons?

- Electrons are detected in a busier part of ATLAS than muons
- Need to account for other energy in Electromagnetic Calorimeter when measuring electrons



- **Inner Detector:**
  - Tracks the path of charged particles.
  - Magnetic field bends the path of charged particles.
- **Electromagnetic Calorimeter:**
  - Particles that interact with the electromagnetic force leave energy deposits.
  - Absorbs Electrons and Photons.
- **Hadronic Calorimeter:**
  - Jets (from quarks) are absorbed.
- **Muon Spectrometer:**
  - Tracks the path of Muons.
  - Magnetic field bends the path of Muons.

UNIVERSITY OF SUSSEX

open data ATLAS

$\text{Signal}/\sqrt{\text{Background}}$

- $\text{Signal}/\sqrt{\text{Background}}$ is a statistical measure of how well you're separating signal and background
- It's an example of "significance"
- This is taking the statistical uncertainty on the number of background events as a Poisson distribution
- Only background in denominator because when searching for a new particle, you have to reject the background-only hypothesis

Lepton p_T

- "the second (third) lepton in p_T order must satisfy $p_T > 15$ GeV ($p_T > 10$ GeV)"

Minimum opposite-charge-same-type lepton pair invariant mass

- "All possible lepton pairs in the quadruplet that have the same flavour and opposite charge must satisfy $m_{\ell\ell} > 5$ GeV in order to reject backgrounds involving the production and decay of J/ψ mesons"

Invariant mass of Z boson candidate 1

- "The same-flavour and opposite-charge lepton pair with an invariant mass closest to the Z boson mass (m_Z) in the quadruplet is referred to as the leading lepton pair. Its invariant mass, denoted by m_{12}, is required to be between 50 GeV and 106 GeV"

Invariant mass of Z boson candidate 2

- "The same-flavour and opposite-charge lepton pair with an invariant mass closest to the Z boson mass (m_Z) in the quadruplet is referred to as the leading lepton pair. The remaining same-flavour, opposite-charge lepton pair is the sub-leading lepton pair. Its invariant mass, m_{34} ,is required to be in the range $m_{\text{min}} < m_{34} < 115$ GeV, where the value of m_{min} depends on the reconstructed four-lepton invariant mass, $m_{4\ell}$. The value of m_{min} varies monotonically from 17.5 GeV at $m_{4\ell} = 120$ GeV to 50 GeV at $m_{4\ell} = 190$ GeV and is constant above this value."

UNIVERSITY OF SUSSEX

open data ATLAS

ΔR

- "leptons are required to be separated from each other by $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} > 0.1$ if they are of the same flavour and by $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} > 0.2$ otherwise."

Minimum lepton p_T

- "Each electron (muon) must satisfy $p_T > 7$ GeV ($p_T > 6$ GeV)"

Lepton etcone20

- "The … isolation for electrons is computed as the sum of the E_T of …energy … clusters … within a cone of size $\Delta R = 0.2$ around the candidate electron …, divided by the electron E_T. … The … energy of the cells assigned to the electron … is excluded,…The … isolation for electrons is required to be less than 0.20. The … isolation …for muons is defined by the ratio to the p_T of the muon of the E_T sum of the calorimeter cells inside a cone of size $\Delta R = 0.2$ around the muon direction minus the energy deposited by the muon. Muons are required to have a … isolation less than 0.30 "

Lepton d0

- "Non-prompt leptons from heavy flavour decays, electrons from photon conversions and jets mis-identified as electrons have broader … impact parameter distributions than prompt leptons from Z boson decays and/or are non-isolated. Thus, the Z and $t\bar{t}$ background contributions are reduced by applying a cut on the … impact parameter significance, defined as the transverse impact parameter divided by its uncertainty, d_0/σ_{d_0}. This is required to be less than 3.5 (6.5) for muons (electrons). The electron impact parameter is affected by bremsstrahlung and thus has a broader distribution."