13 TeV ATLAS Open Data



- **How can we overcome geographical distances and allow anyone interested in experimental particle physics to learn remotely?**

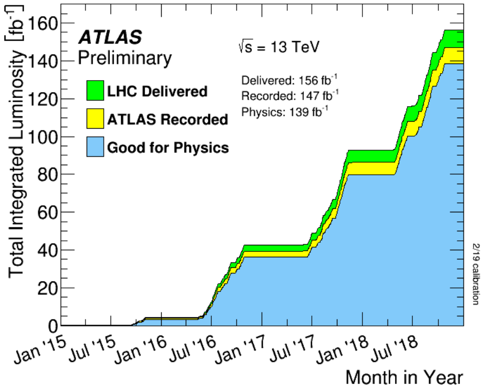
✓ ATLAS Collaboration launched a comprehensive educational platform to guide university-level students and teachers on how to use the data and analysis tools

✓ Provide a straightforward interface to replicate the procedures used by high-energy-physics researchers and enable users to experience the analysis of particle physics data in educational environments

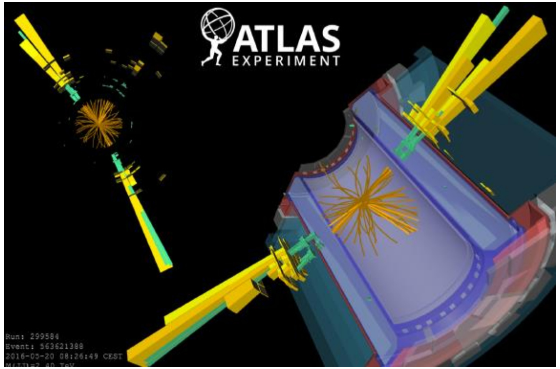
- **What is the aim of ATLAS Open Data?**
  - provide data and tools to high school, undergraduate and graduate students
  - help education in physics analysis techniques used in experimental HEP
  - ATLAS Open Data has been incorporated into curriculums of multiple universities in Belgium, Canada, Colombia, Greece, Germany, Norway, Poland, Portugal, Spain, Sweden, Switzerland, UK, USA, Venezuela and others
  - featured in ATLAS blog and news

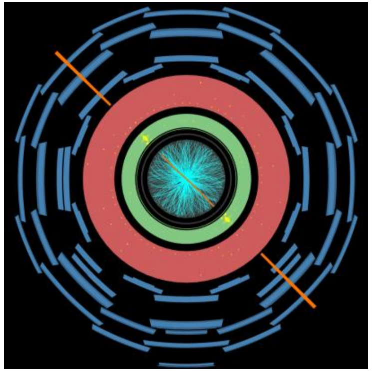
- **ATLAS Data Access Policy:**
  - ATL-CB-PUB-2015-001: sets out the guidelines regarding open access to ATLAS data by non-ATLAS members with a focus on education, training and outreach

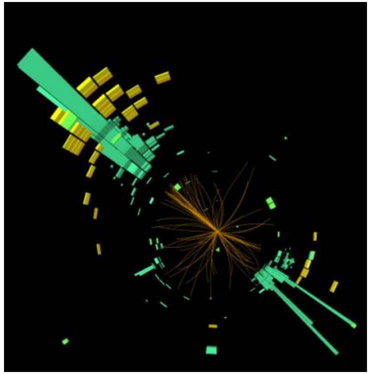
- **13 TeV ATLAS Open Data:**
  - ✓ 61 runs from the first 4 periods of the 2016 proton-proton data-taking
  - ✓ releasing to the public **10 fb^{-1}** of pp collision data (~ 270 million collision events)



- **Events are selected by applying several event-quality and trigger criteria, and classified according to the type and multiplicity of reconstructed objects**
  - ✓ subjected to a **loose** event preselection to reduce processing **time**







- **13 TeV ATLAS Open Data reconstructed objects contain:**
  - **electrons**, **muons**, **photons**, hadronically decaying **tau-leptons**, **small-R jet** and **large-R jet** candidates (and **MET**) reconstructed with the ATLAS detector

Electron (e)	Muon (μ)	Photon (γ)
InDet & EMCAL rec. loose identification loose isolation $p_T > 7$ GeV $\|\eta\| < 2.47$	InDet & MS rec. loose identification loose isolation $p_T > 7$ GeV $\|\eta\| < 2.5$	InDet & EMCAL rec. tight identification loose isolation $E_T > 25$ GeV $\|\eta\| < 2.37$

Hadronically decaying τ-leptons (τ_h)	Small-R jets	Large-R jets
InDet & EMCAL rec. medium identification $p_T > 20$ GeV $\|\eta\| < 2.5$ 1 or 3 associated tracks	EMCAL & HCAL rec.n anti-k_t, R = 0.4 $p_T > 20$ GeV $\|\eta\| < 2.5$	EMCAL & HCAL rec. anti-k_t, R = 1.0 $p_T > 250$ GeV $\|\eta\| < 2.0$ trimming: $R_{sub} = 0.2$, $f_{cut} = 0.05$

- **Selected events classified into separate final-state collections**
  - depending on the number of final state objects and their energy, and triggers used (single-lepton, diphoton,..)

Final-state categories	Leading object p_T (min) [GeV]	Collection name
$N_\ell = 1$	25	1lep
$N_\ell \geq 2$	25	2lep
$N_\ell = 3$	25	3lep
$N_\ell \geq 4$	25	4lep
$N_{\text{largeRjet}} \geq 1$ & $N_\ell = 1$	250 (large-R jet), 25 (lepton)	1largeRjet1lep
$N_{\tau-\text{had}} = 1$ & $N_\ell = 1$	20 (τ_h), 25 (lepton)	1lep1tau
$N_\gamma \geq 2$	35	GamGam

Also **MC simulation** samples describing several SM processes used to model the expected distributions of different signal and background events (top quark pair, single top quark, Z+jets, W+jets, WW/WZ/ZZ, SM Higgs and BSM signals)

Process	Unique "channelNumber"	Generator, hadronisation	Additional information
Top-quark production			
$t\bar{t}$+jets	410000	Powheg-Box v2 [68] + Pythia 8 [69]	only 1ℓ and 2ℓ decays of $t\bar{t}$-system
single (anti)top t-channel	(410012) 410011	Powheg-Box v1 + Pythia 6 [70]	
single (anti)top Wt-channel	(410014) 410013	Powheg-Box v2 + Pythia 6	
single (anti)top s-channel	(410026) 410025	Powheg-Box v2 + Pythia 6	
W/Z (+ jets) production			
$Z \to ee, \mu\mu, \tau\tau$	361106 – 361108	Powheg-Box v2 + Pythia 8	LO accuracy up to $N_{\text{jets}} = 1$
$W \to e\nu, \mu\nu, \tau\nu$	361100 – 361105	Powheg-Box v2 + Pythia 8	LO accuracy up to $N_{\text{jets}} = 1$
$W \to e\nu, \mu\nu, \tau\nu$ + jets	364156 – 364197	Sherpa 2.2 [71]	LO accuracy up to 3-jets final states
$Z \to ee, \mu\mu, \tau\tau$ + jets	364100 – 364141	Sherpa 2.2	LO accuracy up to 3-jets final states
Diboson production			
WW	363359, 363360	Sherpa 2.2	$qq'\ell\nu$ final states
WW	363492	Sherpa 2.2	$\ell\nu\ell'\nu'$ final states
ZZ	363356	Sherpa 2.2	$qq'\ell^+\ell^-$ final states
ZZ	363490	Sherpa 2.2	$\ell^+\ell^-\ell'^+\ell'^-$ final states
WZ	363358	Sherpa 2.2	$qq'\ell^+\ell^-$ final states
WZ	363489	Sherpa 2.2	$\ell\nu qq'$ final states
WZ	363491	Sherpa 2.2	$\ell\nu\ell^+\ell^-$ final states
WZ	363493	Sherpa 2.2	$\ell\nu\nu\nu'$ final states
SM Higgs production ($m_{\text{H}} = 125$ GeV)			
ggF, $H \to WW$	345324	Powheg-Box v2 + Pythia 8	$\ell\nu\ell'\nu'$ final states
VBF, $H \to WW$	345323	Powheg-Box v2 + Pythia 8	$\ell\nu\ell'\nu'$ final states
ggF, $H \to ZZ$	345060	Powheg-Box v2 + Pythia 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
VBF, $H \to ZZ$	344235	Powheg-Box v2 + Pythia 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
ZH, $H \to ZZ$	341947	Pythia 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
WH, $H \to ZZ$	341964	Pythia 8	$\ell^+\ell^-\ell'^+\ell'^-$ final states
ggF, $H \to \gamma\gamma$	343981	Powheg-Box v2 + Pythia 8	
VBF, $H \to \gamma\gamma$	345041	Powheg-Box v2 + Pythia 8	
$WH(ZH)$, $H \to \gamma\gamma$	345318, 345319	Powheg-Box v2 + Pythia 8	
$t\bar{t}H$, $H \to \gamma\gamma$	341081	aMC@NLO [72] + Pythia 8	
BSM production			
$Z' \to t\bar{t}$	301325	Pythia 8	$m_{Z'} = 1$ TeV
$\tilde{\ell}\tilde{\ell}' \to \ell\tilde{\chi}^0_1\ell'\tilde{\chi}^{0\prime}_1$	392985	aMC@NLO + Pythia 8	$m_{\tilde{\ell}} = 600$ GeV, $m_{\tilde{\chi}^0_1} = 300$ GeV

Overview of the 13 TeV ATLAS Open Data

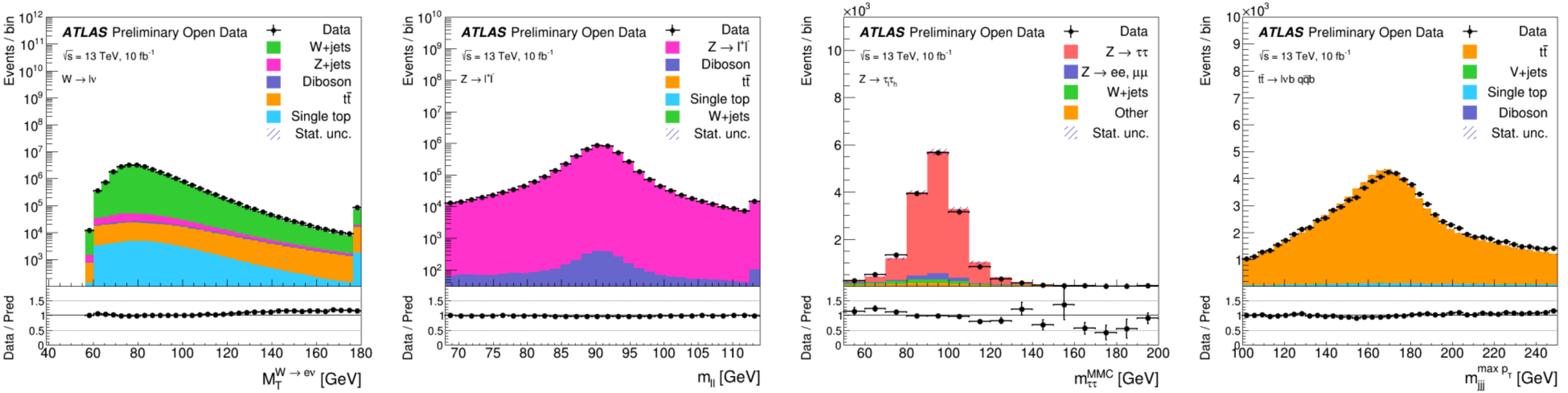
- both data and MC provided in a **simplified data format** reducing the information content of the original data analysis format used within ATLAS
- ROOT tuple with more than **80 branches**, optimised to reduce the complexities encountered in a full-scale analysis **(~150 GB** of storage)

Tuple branch name	C++ type	Variable description
runNumber	int	number uniquely identifying ATLAS data-taking run
eventNumber	int	event number and run number combined uniquely identifies event
channelNumber	int	number uniquely identifying ATLAS simulated dataset
mcWeight	float	weight of a simulated event
XSection	float	total cross-section, including filter efficiency and higher-order correction factor
SumWeights	float	generated sum of weights for MC process
scaleFactor_PILEUP	float	scale-factor for pileup reweighting
scaleFactor_ELE	float	scale-factor for electron efficiency
scaleFactor_MUON	float	scale-factor for muon efficiency
scaleFactor_PHOTON	float	scale-factor for photon efficiency
scaleFactor_TAU	float	scale-factor for tau efficiency
scaleFactor_BTAG	float	scale-factor for b-tagging algorithm @70% efficiency
scaleFactor_LepTRIGGER	float	scale-factor for lepton triggers
scaleFactor_PhotonTRIGGER	float	scale-factor for photon triggers
trigE	bool	boolean whether event passes a single-electron trigger
trigM	bool	boolean whether event passes a single-muon trigger
trigP	bool	boolean whether event passes a diphoton trigger
lep_n	int	number of pre-selected leptons
lep_truthMatched	vector<bool>	boolean indicating whether the lepton is matched to a simulated lepton
lep_trigMatched	vector<bool>	boolean indicating whether the lepton is the one triggering the event
lep_pt	vector<float>	transverse momentum of the lepton
lep_eta	vector<float>	pseudo-rapidity, η, of the lepton
lep_phi	vector<float>	azimuthal angle, ϕ, of the lepton
lep_E	vector<float>	energy of the lepton
lep_z0	vector<float>	z-coordinate of the track associated to the lepton wrt. primary vertex
lep_charge	vector<int>	charge of the lepton
lep_type	vector<int>	number signifying the lepton type (e or μ)
lep_isTightID	vector<bool>	boolean indicating whether lepton satisfies tight ID reconstruction criteria
lep_ptcone30	vector<float>	scalar sum of track p_T in a cone of R=0.3 around lepton, used for tracking isolation
lep_etcone20	vector<float>	scalar sum of track E_T in a cone of R=0.2 around lepton, used for calorimeter isolation
lep_trackd0pvunbiased	vector<float>	d_0 of track associated to lepton at point of closest approach (p.c.a.)
lep_tracksigd0pvunbiased	vector<float>	d_0 significance of the track associated to lepton at the p.c.a.
met_et	float	transverse energy of the missing momentum vector
met_phi	float	azimuthal angle of the missing momentum vector
jet_n	int	number of pre-selected jets
jet_pt	vector<float>	transverse momentum of the jet
jet_eta	vector<float>	pseudo-rapidity, η, of the jet
jet_phi	vector<float>	azimuthal angle, ϕ, of the jet
jet_E	vector<float>	energy of the jet
jet_jvt	vector<float>	jet vertex tagger discriminant [21] of the jet
jet_trueflav	vector<int>	flavour of the simulated jet
jet_truthMatched	vector<bool>	boolean indicating whether the jet is matched to a simulated jet
jet_MV2c10	vector<float>	output from the multivariate b-tagging algorithm [22] of the jet

Tuple branch name	C++ type	Variable description
photon_n	int	number of pre-selected photons
photon_truthMatched	vector<bool>	boolean indicating whether the photon is matched to a simulated photon
photon_trigMatched	vector<bool>	boolean indicating whether the photon is the one triggering the event
photon_pt	vector<float>	transverse momentum of the photon
photon_eta	vector<float>	pseudo-rapidity of the photon
photon_phi	vector<float>	azimuthal angle of the photon
photon_E	vector<float>	energy of the photon
photon_isTightID	vector<bool>	boolean indicating whether photon satisfies tight identification reconstruction criteria
photon_ptcone30	vector<float>	scalar sum of track p_T in a cone of R=0.3 around photon
photon_etcone20	vector<float>	scalar sum of track E_T in a cone of R=0.2 around photon
photon_convType	vector<int>	information whether and where the photon was converted
largeRjet_n	int	number of pre-selected large-R jets
largeRjet_pt	vector<float>	transverse momentum of the large-R jet
largeRjet_eta	vector<float>	pseudo-rapidity of the large-R jet
largeRjet_phi	vector<float>	azimuthal angle of the large-R jet
largeRjet_E	vector<float>	energy of the large-R jet
largeRjet_m	vector<float>	invariant mass of the large-R jet
largeRjet_truthMatched	vector<int>	information whether the large-R jet is matched to a simulated large-R jet
largeRjet_D2	vector<float>	weight from algorithm [57] for W/Z-boson tagging
largeRjet_tau32	vector<float>	weight from algorithm [57] for top-quark tagging
tau_n	int	number of pre-selected hadronically decaying τ-lepton
tau_pt	vector<float>	transverse momentum of the hadronically decaying τ-lepton
tau_eta	vector<float>	pseudo-rapidity of the hadronically decaying τ-lepton
tau_phi	vector<float>	azimuthal angle of the hadronically decaying τ-lepton
tau_E	vector<float>	energy of the hadronically decaying τ-lepton
tau_charge	vector<int>	charge of the hadronically decaying τ-lepton
tau_isTightID	vector<bool>	boolean indicating whether hadronically decaying τ-lepton satisfies tight ID reconstructio
tau_truthMatched	vector<bool>	boolean indicating whether the hadronically decaying τ-lepton is matched to a simulated
tau_trigMatched	vector<bool>	boolean signifying whether the τ-lepton is the one triggering the event
tau_nTracks	vector<int>	number of tracks in the hadronically decaying τ-lepton decay
tau_BDTid	vector<float>	output of the multivariate algorithm [24] discriminating hadronically decaying τ-leptons fr
ditau_m	float	di-τ invariant mass using the missing-mass calculator [54]
lep_pt_syst	vector<float>	single component syst. uncert. (lepton momentum scale and resolution [36,15]) affecting le
met_et_syst	float	single component syst. uncert. (E_T^{miss} scale and resolution [30]) affecting met_pt
jet_pt_syst	vector<float>	single component syst. uncert. (jet energy scale [37]) affecting jet_pt
photon_pt_syst	vector<float>	single component syst. uncert. (photon energy scale and resolution [16]) affecting photon_
largeRjet_pt_syst	vector<float>	single component syst. uncert. (large-R jet energy resolution [37]) affecting largeRjet_pt
tau_pt_syst	vector<float>	single component syst. uncert. (τ-lepton reconstruction and energy scale [24]) affecting ta

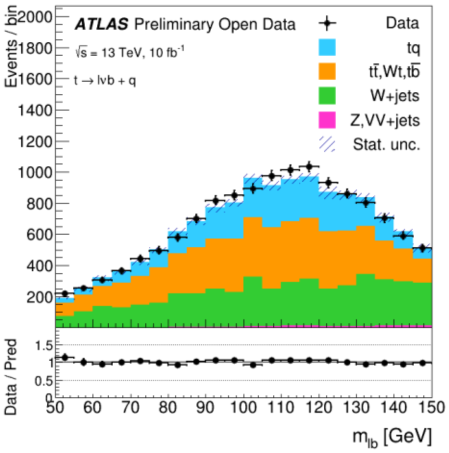
- **12 examples of physics analysis using 13 TeV ATLAS Open Data**
  - inspired and following as closely as possible the procedures and selections taken in already published ATLAS Collaboration results

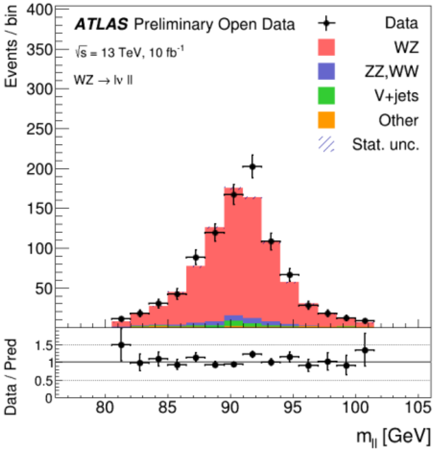
Analyses	Physics processes	Purpose
4 high statistics	W→lν, Z→(ee/μμ), ττ top-quark-pair	high event yields to study the SM processes in detail

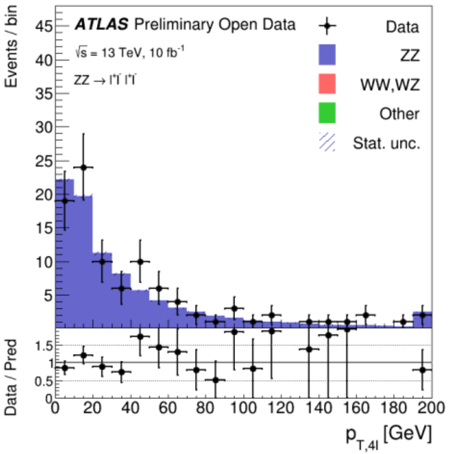


Using ATLAS Open Data, you can **re-create the major particle discoveries** of the late 20th century: the Z-boson, W-boson and top quark

Analyses	Physics processes	Purpose
3 low statistics	Single-top-quark, WZ and ZZ diboson	illustrate the statistical limitations of the released dataset

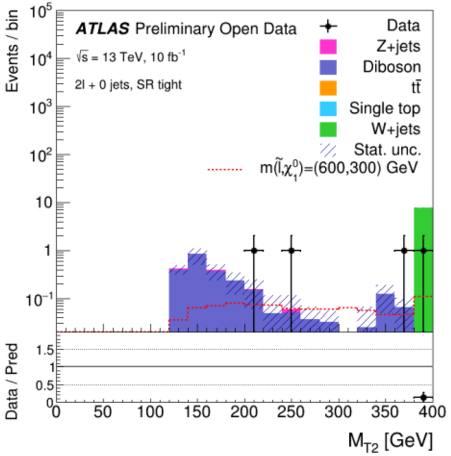


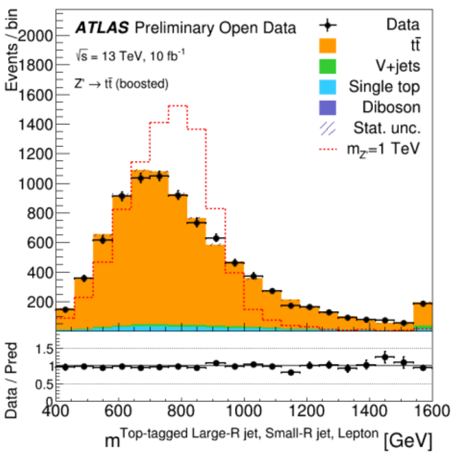




Educational test-bed for **new data-analysis techniques**, e.g. kinematic fitting, multivariate discrimination and machine learning tasks

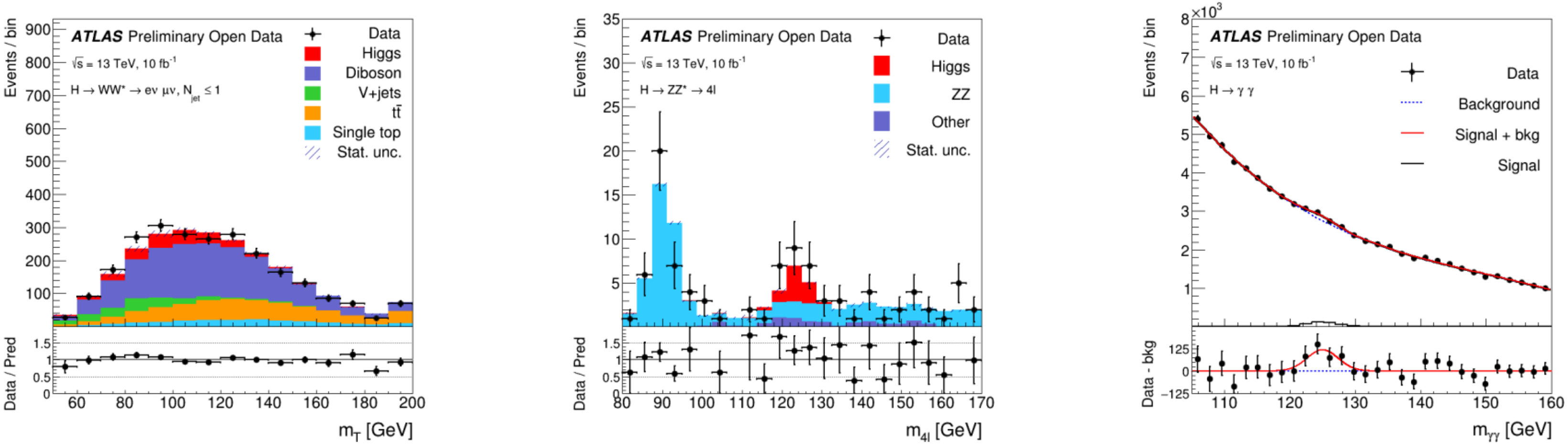
Analyses	Physics processes	Purpose
2 BSM physics	SUSY, heavy boson	searching for new physics using different physics objects





Evaluation of the impact of different sources of **systematic uncertainties** is one of the new tasks that is available with the 13 TeV datasets

Analyses	Physics processes	Purpose
3 Higgs boson	H→WW, H→ZZ H→γγ	“re-discover” the production of the SM Higgs boson



“Re-discover” the SM Higgs boson in **different final-state** scenarios!

- **But how?**

  - The 13 TeV ATLAS Open Data is hosted on the CERN Open Data online portal and ATLAS Open Data online portal

  - Is accompanied by a **set of analysis frameworks**, written in **C++** and interfaced with ROOT, **Python uproot and pandas/numpy**, **pyROOT** and **RDataFrame**, publicly available in a **GitHub repository**.

  - The frameworks implement the protocols needed for reading the datasets, making an analysis selection, writing out histograms and plotting the results.

  - **During this workshop, you will get familiar with all the frameworks, both written in C++, PyRoot and RootDataFrame**