Engaging Universities and Beyond – Open Data for Education and Research

Leonardo Toffolin[1,2], on behalf of the ATLAS Collaboration
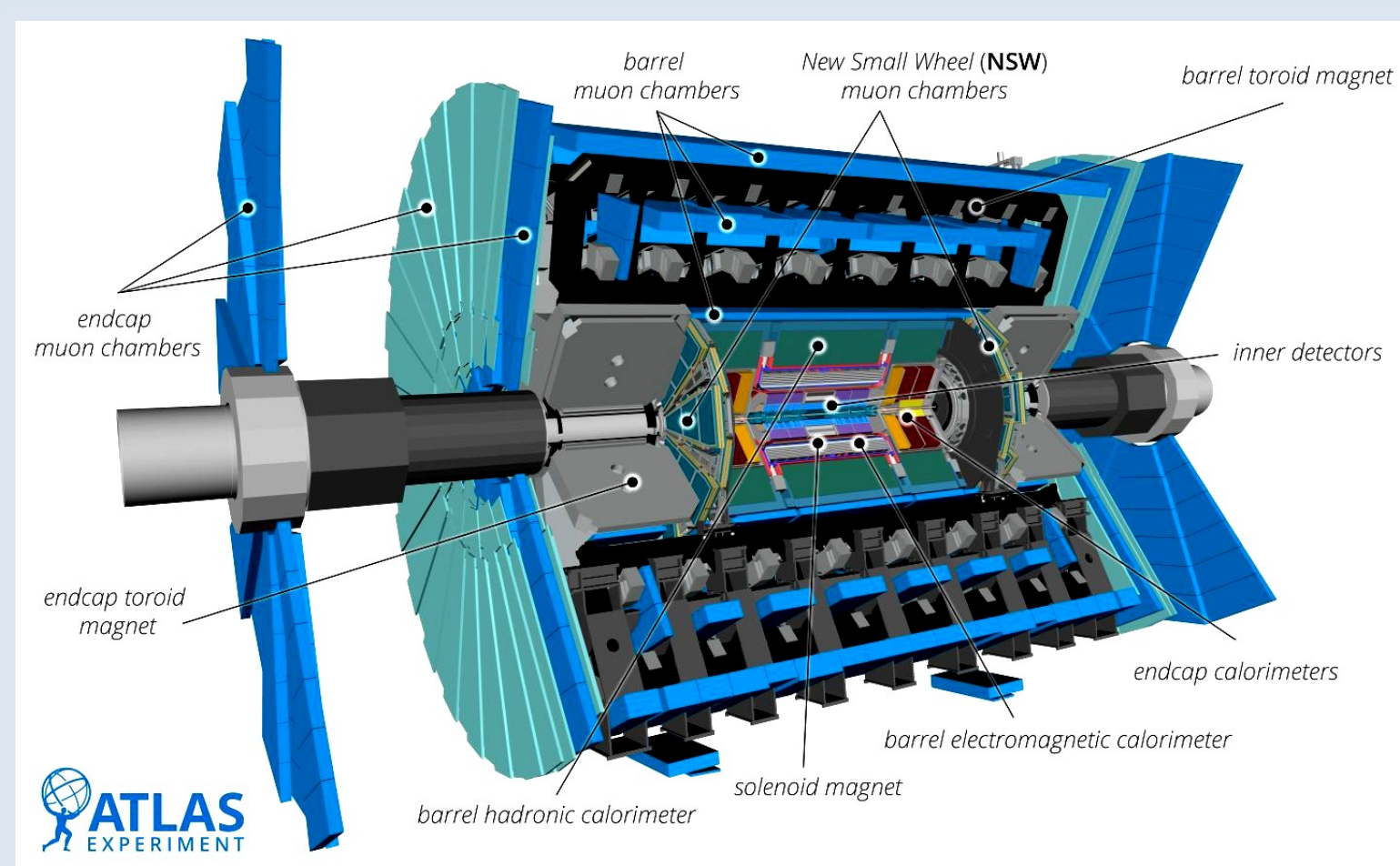
LHCP 2023, May 23rd 2023 (Belgrade, Serbia)

Introduction

- ATLAS is one of the two **general purpose experiments at the Large Hadron Collider** (LHC) at CERN
- Physics programme:
  - **Higgs Physics**
  - **QCD** and **EW** Sectors of the Standard Model
  - **Beyond Standard Model** physics
- ATLAS has collected **~4 PB of raw data** to date
- Reconstructed / pre-processed analysis data takes up >100 PB of space
- ATLAS reconstruction / simulation / common analysis software is ~4 million lines of C++
- Installed **size of the software**, with all of its "dependencies", is **O(10) GB**

Structure and strategy

- Consists of **2 "campaigns"**, both with accompaining Monte Carlo datasets
  - **1 fb^{-1} of 8 TeV data** [1]
  - **10 fb^{-1} of 13 TeV data** [2]
- Real and simulated data stored in a **"ROOT ntuple format"** [4]
  - Calibrated and simplified information about the **reconstructed high level objects**
  - **Simplified information** for taking systematic uncertainties into account on the properties of the objects
    - Dataset identifiers
    - Trigger (boolean) variables
    - Kinematic variables
    - Sources of systematic uncertainties
  - Generated with the help of one of **ATLAS's "analysis frameworks"**
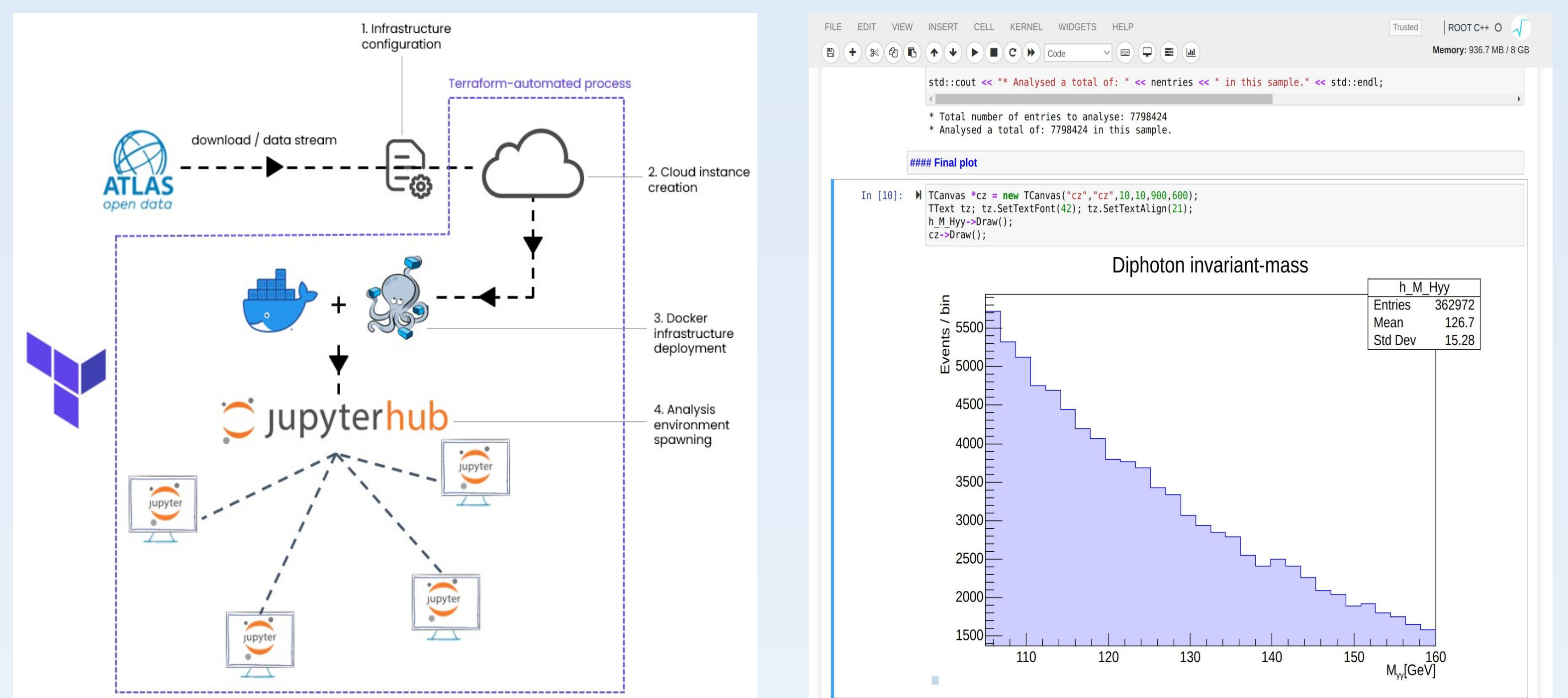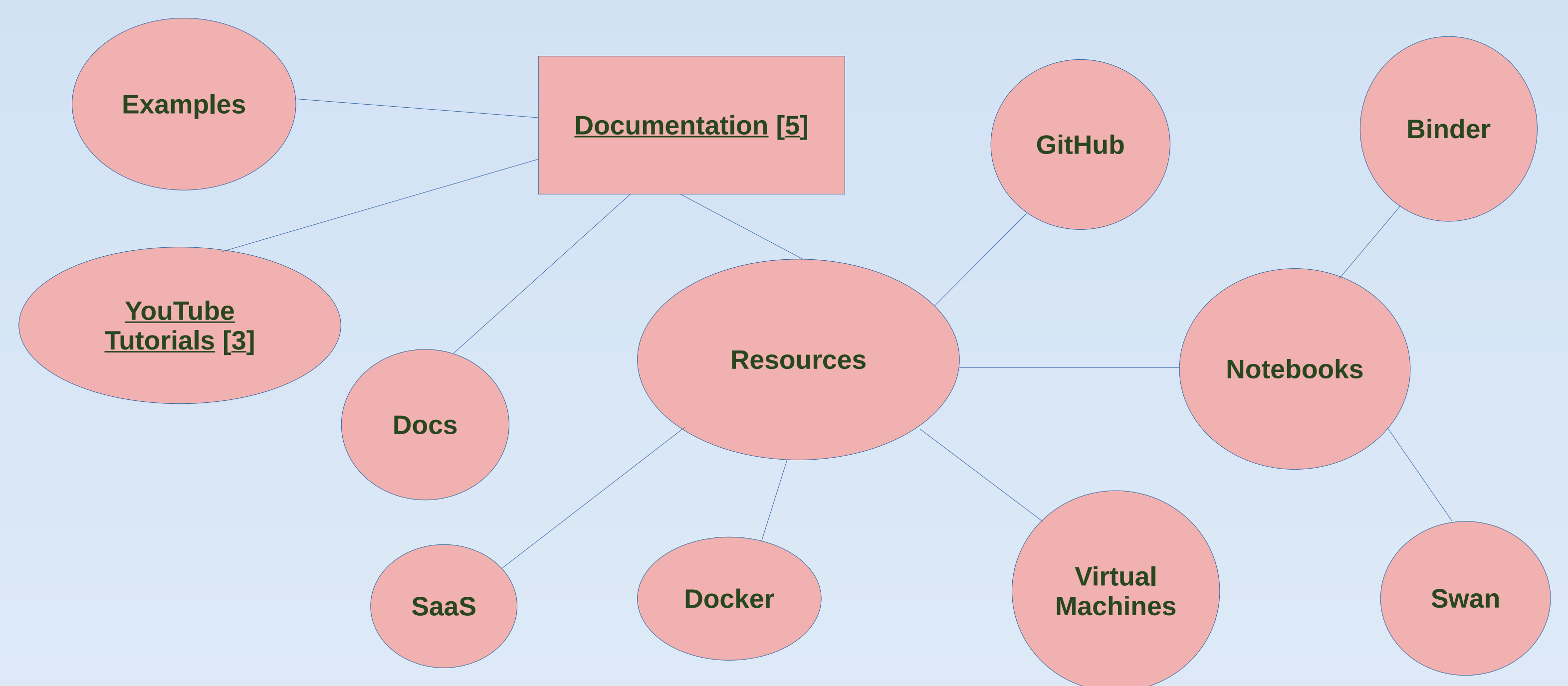
Tuple branch name	C++ type	Variable description
runNumber	int	number uniquely identifying ATLAS data-taking run
eventNumber	int	event number and run number combined uniquely identifies event
channelNumber	int	number uniquely identifying ATLAS simulated dataset
mcWeight	float	weight of a simulated event
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

References:

[1] http://opendata.atlas.cern/samples-8tev/
[2] http://opendata.atlas.cern/samples-13tev/
[3] https://www.youtube.com/playlist?list=PL1qU3k-RDRsvy3jhxUTmq7ZJQTdFvjLJn
[4] https://cds.cern.ch/record/2707171
[5] http://opendata.atlas.cern/release/2020/documentation/

Motivation

- To **expose students and the general public to real ATLAS proton-proton collision data**, and method and techniques in data analysis and computational skills
- To help **raise interest in Particle Physics** and STEM
- To support the **long-term improvement of the scientific literacy of the public**
- To provide resources to teach **transferable skills in programming** and analysis techniques
- Possibility to **help teachers with data and software** when dealing with modern high-energy physics
- **To meet high expectations from national governments and scientific funders** for the effort made by our experiment to make real data and educational resources available to the public

Software tools and infrastructure

- Analysis code provided with the data to help with its usage
  - **Jupyter Notebooks** (Swan, Binder)
  - **Virtual Machines**
  - **Docker containers**
    - **Notebook interface**
    - **Locally running web service**
  - **Downloadable dataset** and analysis code

Summary and future plans

- Change of ATLAS analysis data formats
  - **DAOD_PHYSLITE format**, meant for LHC Run3 and HL-LHC ATLAS analyses, investigated as the starting point for the next ATLAS Open Data release
  - An **upcoming data release is in planning**
- Data **formats other than ROOT ntuples** being considered
- To **benefits institutions that have computing facilities** and just need infrastructures
- Both open releases of ATLAS's data proved very successful in outreach
  - Many Institutions used it over the years to engage high-school and university students
  - Great help in involving students from countries in various geographic locations
    - Also countries that are not part of the ATLAS Collaboration
- To **embrace new technologies with the software (Jupyter notebooks, Docker containers)**, which are not even used by all ATLAS analyzers at this point

[1] INFN Trieste, Gruppo Collegato di Udine
[2] Università degli Studi di Trieste