Giovanni Guerrieri, for the ATLAS Open Data team

03-07-2023

The ATLAS Open Data goal

- Accessibility
  - Make the data and the tools openly available for everyone to use, without technology, region, or knowledge restrictions.
- Transferable expertise
  - Along with particle physics analysis and ATLAS learning objectives, provide skills in programming, software and machine learning.
- Usability
  - Different target audiences, with different backgrounds and skills must be able to use the data and tools for a wide range of learning objectives.

Currently, ATLAS Open Data releases are being used by several schools, universities, interested individuals, as well as in public events, masterclasses and international workshops.

The datasets are used for an **educational purpose only**.

The ATLAS Open Data release

13 TeV — ATL-OREACH-PUB-2020-001

8 TeV — ATL-OREACH-PUB-2018-001

8 TeV — ATL-OREACH-PUB-2016-001

The ATLAS Open Data release

- Two campaigns
  - 8 TeV: 1fb^{-1} of data
  - 13 TeV: 10fb^{-1} of data

- Associated challenges
  - Create datasets and selections to account for different levels of complexity.
  - Include calibrated and simplified information about the reconstructed high-level objects, while containing the size of the datasets.
  - Adapt part of the ATLAS analysis framework to comply with our needs.
  - Provide useful tools and documentation to make data usable.
  - Maintaining and improving all the online resources to make sure that accessibility is always optimal.

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

More data have been released for specific purposes

- MC datasets for top tagging
  - https://opendata.cern.ch/record/15013
- MC datasets for fast calo simulation which were used as a part of the CaloChallenge:
  - https://opendata.cern.ch/record/15012
- MC datasets for the Higgs Learning challenge:
  - https://opendata.cern.ch/record/328
  - https://opendata.cern.ch/record/331
  - https://opendata.cern.ch/record/329
- And datasets for the TrackML challenge:
  - https://www.kaggle.com/c/trackml-particle-identification

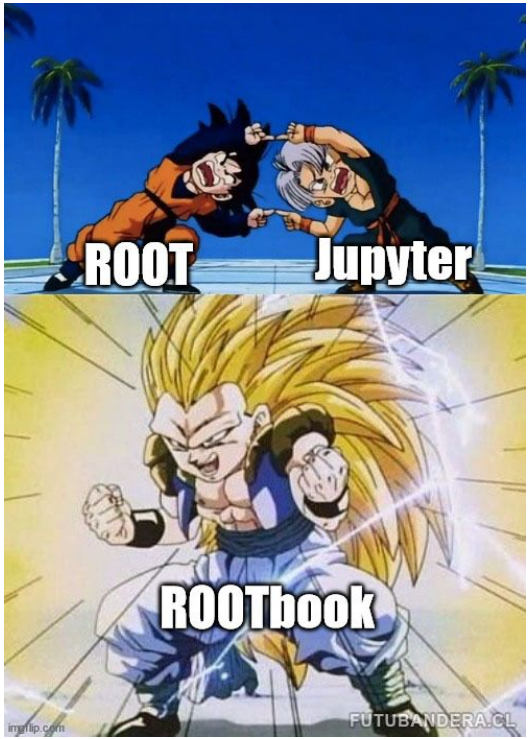
Tools for Open Data

The ATLAS Open Data comes with a set of Jupyter notebooks that allow data analysis to be performed directly in a web browser.

List of notebooks

GitHub repository

- Several analysis examples targeting different users, with different expertise and interests.
- Different frameworks, to adapt to everyone's need:
  - C++
  - python
  - RDataFrame
  - uproot



ATLAS Open Data infrastructure

requiring only internet access	Online
requiring internet access and local resources	Hybrid
requiring only local resources*	Offline

*Internet connection is required in order to download material at the beginning

ATLAS Open Data infrastructure - Online

Histogram analyser: instructive and intuitive look into data

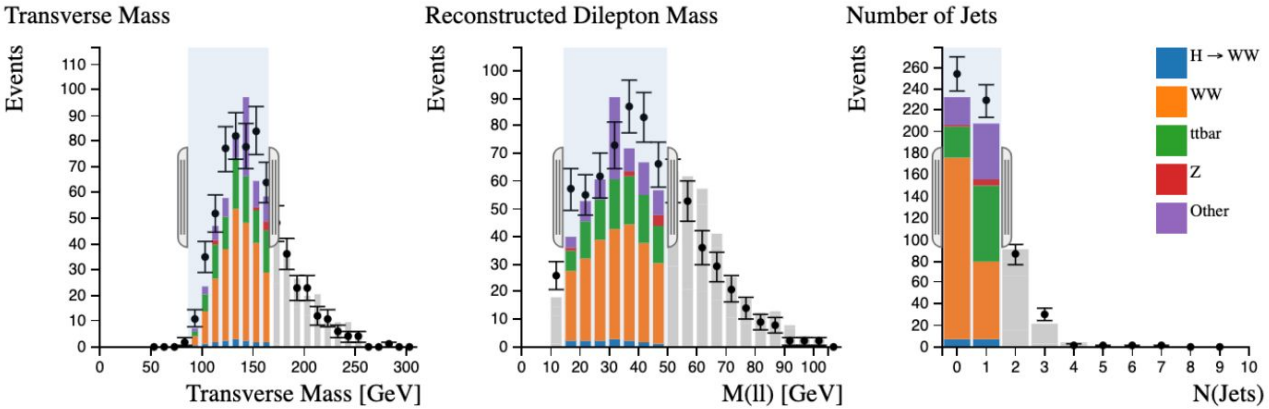
Example

No technical knowledge required ✅

Introduction to the studied process ✅

Step-by-step explanation ✅

Advanced contents ❌



ATLAS Open Data infrastructure - Online

Swan/Binder platforms: very useful for setting up a quick and individual workspace.

Available with the click of a button ✅

Data persistence ❌

Access to eos ✅

Do not need prerequisites ❌

No timeout time for sessions ✅

Spawn time <1min ✅

Software stack available ✅

I have a CERN account

SWAN

I don't have a CERN account

Binder

✅ Available with the click of a button

❌ Data persistence

❌ Access to eos

✅ Do not need prerequisites

❌ No timeout time for sessions

✅ Spawn time <1min

✅ Software stack available

ATLAS Open Data infrastructure - Hybrid

Docker containers: robust, replicable environment

No internet required (after pulling the container and the data) ✅

Data persistence ✅

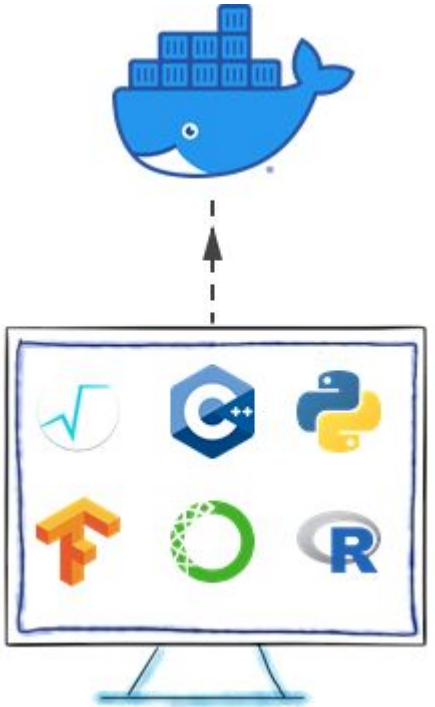
Do not need prerequisites ❌

No timeout time for sessions ✅

Spawn time <1min ✅

Software stack available ✅

Relies on local computational resources ⚠️



ATLAS Open Data infrastructure - Hybrid

Orchestrated docker containers

Set up a local cluster, based on a physical server or a cloud resource

Meant to help local experts, not users!

Automated ✅

Data persistence ✅

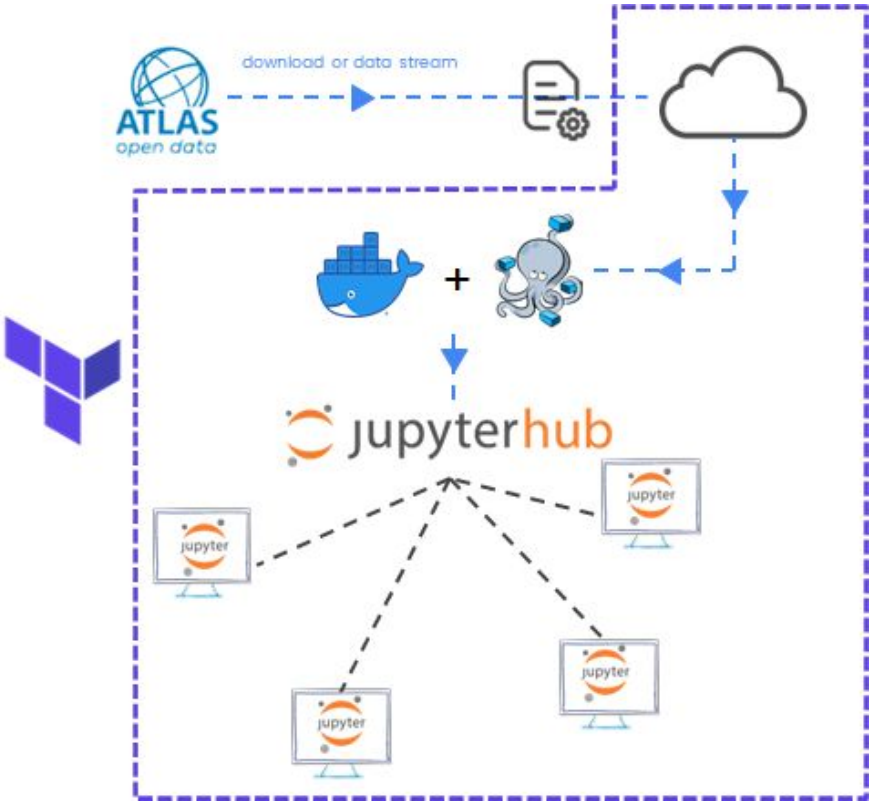
Data-Lake-like setup COOL

External/shared volumes mountable COOL

No timeout time for sessions ✅

Spawn time <1min ✅

Software stack available ✅



Virtual Machines

Download it and use it or put it in a USB key and take it where you want.

Plug ‘n play ✅

Data persistence ✅

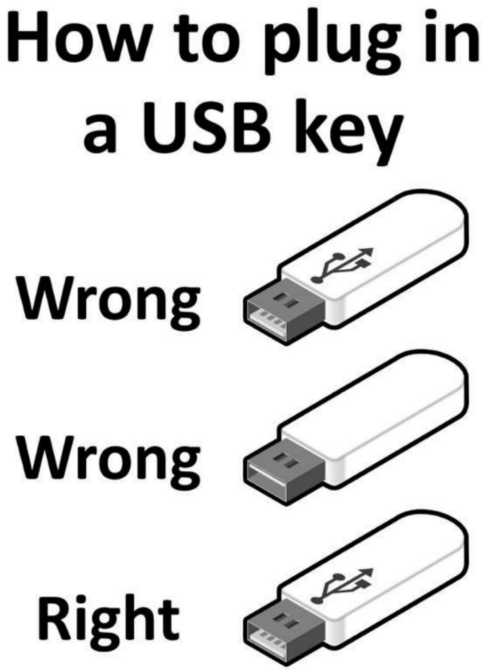
Do not need prerequisites ❌

Works even during a nuclear fallout ☢️

No timeout time for sessions ✅

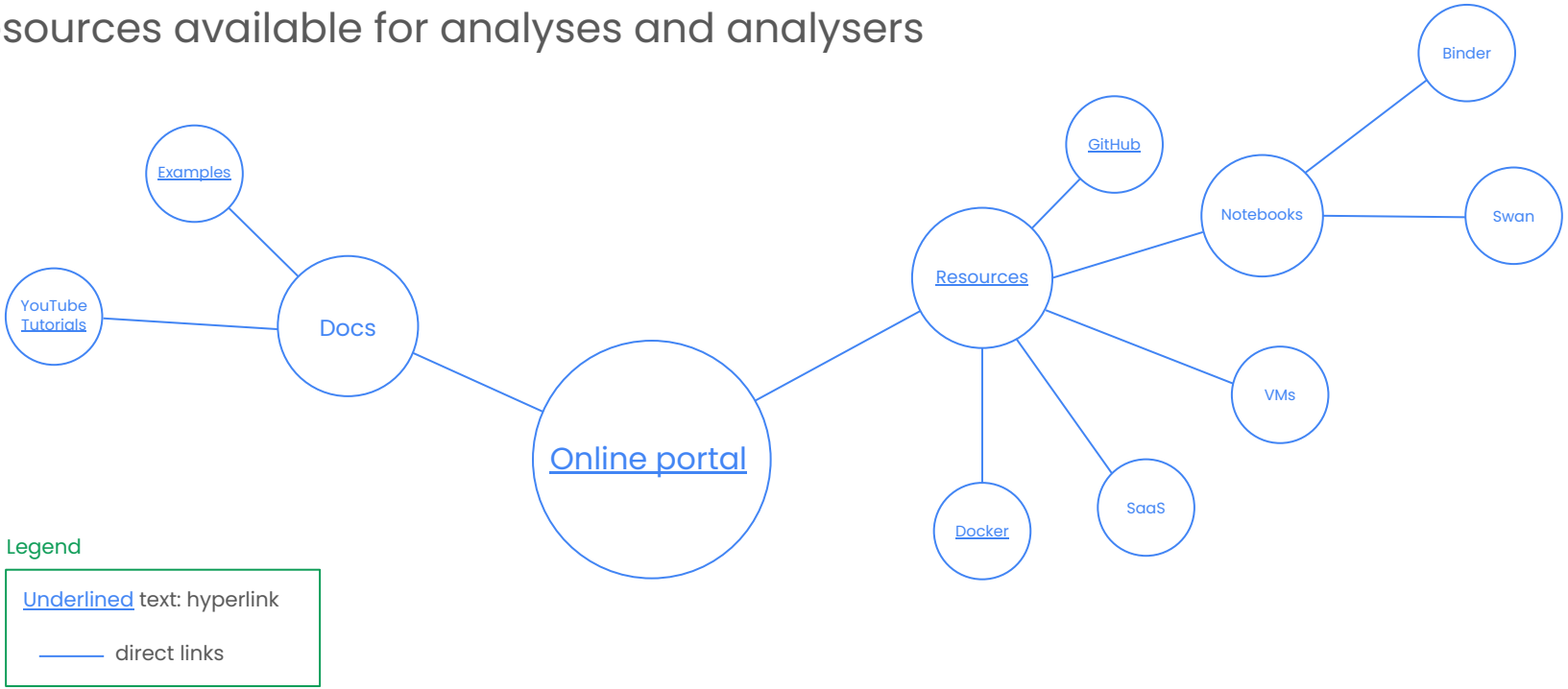
Spawn time <1min ✅

Software stack available 😐



Where do I find all of this?

Resources available for analyses and analysers



What's next?

- Improving what is there
  - Add more notebooks.
  - Enrich the documentation.
  - Maintain current infrastructure and add new resources.

- More data, less space!
  - Increase the amount of available data with a new release.
  - Provide agile and flexible formats for datasets (not only ROOT, not anymore)
  - Improved selection of physics objects (i.e. more analysis possibilities)

What's next?

Thanks!