# FROM RAW DATA TO PHYSICS RESULTS AT THE LHC

August 2022 PSI Particle Physics Summer School

Anna Sfyrla, UniGe

For any question don't hesitate to get in touch:

anna.sfyrla@unige.ch

### WHAT WILL THIS LECTURE BE ABOUT?

#### INTRODUCTION

• Definitions and basic concepts

#### INPUT TO THE PHYSICS

- The data: trigger, data preparation
- The theory: Monte carlo simulations
- Reconstruction, or how to translate detector signals to particles

#### **PHYSICS ANALYSES**

- Through example, step-by-step
- Discussion of analysis methods

#### MACHINE LEARNING IN HEP

• Just a teaser!

Is there a topic you would like to add to this material? If so: please let me know at the end of this lecture and I will see if I can add it!

# PART 2

### THE LIFETIME OF A COLLISION EVENT



# DATA PREPARATION





#### **WORLDWIDE LHC COMPUTING GRID** an international collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists.



- 161 sites, 42 countries
- 1 M CPU cores
- O 1 EB of storage
- o > 2 M jobs/day
- o > 100 PB moved/month
- o accessed by 10k users
- o 10-100 Gb links



Network proved better than anyone imagined: Any job can run anywhere



#### WORLDWIDE LHC COMPUTING GRID THE TIER SYSTEM

#### $\odot$ Tier-0 (CERN):

- Data recording, reconstruction and distribution
- $\circ$  Tier-1:
  - Permanent storage, re-processing, analysis

#### $\circ$ Tier-2:

• Simulation, end-user analysis



# ATLAS DATA MANAGEMENT

# RUCIO





<sup>600</sup> – ATLAS data volume managed by Rucio



# HARDWARE



Tape (at CERN) about 270 PB	<ul> <li>Most reliable and cost-effective technology for large-scale archiving</li> <li>Data stored there infinitely</li> </ul>	Magnetic tapes, retrieved by robotic arms, are used for long-term storage
Disk about 200 PB	<ul> <li>Data for initial processing</li> <li>Copies for further processing / user analysis</li> <li>Data in disks gets staged from tape, on demand</li> </ul>	

Mainly GRID • **CPUs** About 400k cores • Also considering for the future: Mostly for RnD ٠ **GPUs** FPGA accelerators Few 10s ۲ Online farm, 100k cores • Opportunistic High Performance Computers, primarily in the US ulletresources Volunteer computing •



100



Processing power

Storage

## SOFTWARE



-**o- 70,356** Commits 🗜 34 Branches 🔗 1,374 Tags 🗈 2.6 GB Files 🕞 2.6 GB Storage 🛷 124 Releases

The ATLAS Experiment's main offline software repository

 All software organized in packages in Git. For example: <u>https://gitlab.cern.ch/atlas/athena</u>



- All software open source, copyrighted and licenced (Apache 2)
  - "Copyright (C) 2002-2020 CERN for the benefit of the ATLAS collaboration"
  - For open use but also for crediting developers who move out of academia
- Thorough tracking of software developments a key of success
  - Via the Jira software, supported by CERN IT Jira Software
  - Multiple releases exist for merging of new code with existing one
  - Automated tools run nightly to verify code sanity & performance
  - Globally the software projects are coordinated with careful planning
- Software Tools
  - Databases
  - Analysis tools: ROOT is the workhorse!



• Analysis-specific software developed by teams available to whole collaboration!

101

# DATA PREPARATION

### THE LIFETIME OF A COLLISION EVENT



## THE EVENT AT TIER-0





## E.G. ALIGNMENT



105

Day-by-day value of the relative longitudinal shift between the two half-shells of the BPIX as measured with the primary vertex residuals, for the last month of pp data taking in 2012.

# DATA QUALITY

- ✓ The data we analyze have to follow norms of quality such that our results are trustable.
- Online: Fast monitoring of detector performance during data taking, using dedicated stream, "express stream".
- Offline: More thorough monitoring at two instances:
  - Express reconstruction; fast turn-around.
  - Prompt reconstruction: larger statistics.

#### What is monitored?

- Noise in the detector.
- Reconstruction (tracks, clusters, combined objects, resolution and efficiency).
- Input rate of physics.
- ◎ All compared to reference histograms of data that has been validated as "good".



# DATA QUALITY AND "GRL"



107

LUMINOSITY



**LUMINOSITY** 



### LUMINOSITY - THE FIGURE OF MERIT



More of less fixed parameters: Revolution frequency and Number of bunches

## LUMINOSITY - THE FIGURE OF MERIT



- The LHC is built to collide protons at 7 TeV per beam, which is 14 TeV centre of Mass
- In 2012 it ran at 4 TeV per beam, 8 TeV c.o.m.
- Since 2015 it runs at 6.5 TeV per beam, 13 TeV c.o.m
- In Run 3, starting this year, it will run at 6.8 TeV per beam, 13.6 TeV c.o.m



Figure from R. Steerenberg



URL: https://op-webtools.web.cern.ch/vistar/vistars.php?usr=LHC1

### LUMINOSITY - THE FIGURE OF MERIT

 $L = \int \mathcal{L} dt$ 

 $\sigma$ 

 $\frac{\text{N events}}{L}$ 



113

## LUMINOSITY DETERMINATION "FIGURE OF MERIT"

- ◎ A measurement of the number of collisions per cm<sup>2</sup> and second.
- Multiple methods used for determining luminosity: reducing uncertainties.
- Principle detectors for luminosity determination on ATLAS:
  - Beam Conditions Monitor (BCM)
    - Designed for beam abort system
    - Diamond Sensors,  $|\eta| \sim 4.2$



- LUCID
  - Oblicated Luminosity Monitor
  - <sup>(a)</sup> Cherenkov Tubes,  $5.6 < |\eta| < 6.0$





LUCID 2 installation in 2014



115

fast turn-around time.  $2^{n \ \square x} x^{\square y}$ 



#### **Standard Model Total Production Cross Section Measurements**

# A TINY BIT OF MONTE CARLO

## WHY DO WE NEED MONTE CARLO SIMULATION?

#### We only build one detector: how does this influence the physics we are doing?

- How do we compromise physics due to detector design?
- How would a different detector design affect measurements?
- How does the detector behave to radiation?
- In the detectors we only measure voltages, currents, times: how do we go from these to particles?
  - It's an interpretation to say that such-and-such particle caused such-and-such signature in the detector.
  - ◎ Simulating the detector behavior we correct for inefficiencies, inaccuracies, unknowns.
- We need a theory to tell us what we expect and to compare our data against.
- A good simulation is the way to demonstrate to the world that we understand the detectors and the physics we are studying.

### MONTE CARLO PRODUCTION CHAIN





### MONTE CARLO GENERATORS VARIOUS MODELS OF THE PHYSICS OF INTEREST

VARIOUS MODELS OF THE PHYSICS OF INTEREST



### OUR LHC SIMULATION: THE DREAM



121

### OUR LHC SIMULATION: THE REALITY?

#### THIS IS MOST PEOPLE'S VIEW OF THE CHAIN

and this is how we will treat it too, in lack of time...



### SIMULATION - FULL AND FAST



123

### SIMULATION - FULL AND FAST





# The **SATLAS** Open Data
#### Why? 🄊 Guarantee openness and preservation of experimental data

New open data policy in support of open science from CERN & the LHC experiments

#### PEER-REVIEWED PUBLICATIONS

- Open Access
- Followed by detailed data related to the results, available at hepdata.net
   Purpose: Communicate results and maximize their scientific value

#### **RECONSTRUCTED & CALIBRATED DATA**

- Followed by related metadata
- Accompanied by appropriate simulated data samples
- Purpose: Algorithmic, performance and physics studies

#### **DATA FOR OUTREACH AND EDUCATION**

- Selected and formatted ("light") datasets
- Examples available in Jupyter notebooks
- Used in university classes, in growing numbers
   Purpose: Maximize educational impact

More info: https://atlas.cern/resources/opendata



#### Searching for the Higgs boson in the $H{\rightarrow}\gamma\gamma$ channel

#### Python notebook example

Introduction Let's take a current ATLAS Open Data sample and create a histogram:

```
In [1]: import ROOT
from ROOT import TMath
import time
```

Welcome to JupyROOT 6.07/03

In [2]: start = time.time()

# RECONSTRUCTION



### WHAT DO WE RECONSTRUCT?

• Tracks and clusters

Combining those:
"objects", i.e. "particles"



#### **Simplified Detector Transverse View**



### **RECONSTRUCTION - FIGURES OF MERIT**



### **RECONSTRUCTION - FIGURES OF MERIT**

	DEFINITION	EXAMPLE		NEEDS BE:
EFFICIENCY	how often do we reconstruct the object we are interested in	electron identification efficiency = (number of reconstructed electrons) / (number of true electrons) in bins of transverse momentum	$\begin{array}{c} 0.95\\ 0.95\\ 0.95\\ 0.8\\ 0.8\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.75\\ 0.8\\ 0.7\\ 0.7\\ 0.7\\ 0.7\\ 0.7\\ 0.7\\ 0.7\\ 0.7$	High
RESOLUTION	how accurately do we reconstruct the quantity	energy resolution = (measured energy – true energy)/(true energy)	$\sigma = (1.12 \pm 0.03)\%$	<b>Good</b> (a small number)
FAKE RATE	how often we reconstruct a different object as the object we are interested in	a jet faking an electron, fake rate = (Number of jets reconstructed as an electron) / (Number of jets) in bins of pseudorapidity	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$	Low

### **RECONSTRUCTION – GOALS**

- High efficiency
- Good resolution
- Low fake rate

Noise

Dead regions of the detector

Increased pile-up

#### **©** Computing-friendly-

CPU time per event
Memory use





### WHAT DO WE RECONSTRUCT?



- Combining those:
  - "objects", i.e. "particles"



#### **©** For a track we measure:

- Its momentum;
- Its direction;
- Its charge;
- Its "perigee": the closest point to a reference line, transverse ( $d_0$ ) or longitudinal ( $z_0$ ).

Tracks are key ingredients of most of particle reconstruction.

### TRACKING IN A NUTSHELL





### TRACKING IN A NUTSHELL - TRACK FITTING



Perfect measurement – ideal



Imperfect measurement – reality



**©** Small errors and more points help to constrain the possibilities



- **©** Quantitatively:
  - Parameterize the track;
  - Find parameters by Least-Squares-Minimization;
  - Obtain also uncertainties on the track parameters.

### TRACKING IN A NUTSHELL - TRACK FITTING

#### **©** For a track we measure:

- Its momentum;
- ◎ Its direction;
- Its charge;
- Its "perigee": the closest point to a reference line, transverse ( $d_o$ ) or longitudinal ( $z_o$ ).



**136** 

## TRACKING IN A NUTSHELL - TRACK FITTING

#### **©** For a track we measure:

- Its momentum;
- Its direction;
- Its charge;
- Its "perigee": the closest point to a reference line, transverse ( $d_0$ ) or longitudinal ( $z_0$ ).
- And their uncertainty

#### **Small uncertainties are required.**

- <sup>(©)</sup> δdo is < O(10 $\mu$ m) and δθ < O(0.1mrad).
- Allows separation of tracks that come from different particle decays (which can be separated at the order of mm).





#### Presence of Material

- Coulomb scattering off the core of atoms
- Energy loss due to ionization
- Bremsstrahlung
- Hadronic interaction

#### Misalignment

- Detector elements not positioned in space with perfect accuracy.
- Alignment corrections derived from data and applied in track reconstruction.



138

### **IMPACT OF GOOD ALIGNMENT**

Improving the tracker alignment description in the reconstruction gives better track momentum resolution which leads to better mass resolution.



- Can see the reconstructed Z width gets narrower if we use better alignment constants. Very important for physics analysis to have good alignment.
- Alignment of detector elements can change with time, for example when the detector is opened for repair, or when the magnetic field is turned on and off.

### **RECONSTRUCTION OF TRACKS AT DUNE'S NEAR DETECTOR PROTOTYPE**



140

Outside the LHC

# **RECONSTRUCTION OF TRACKS AT THE PROTODUNE DEMONSTRATOR**





### WHAT DO WE RECONSTRUCT?



- Combining those:
  - "objects", i.e. "particles"



### A CALORIMETER VIEW



### **CLUSTERING IN A NUTSHELL**



Reconstruct energy deposited in the calorimeter by charged or neutral particles;
 electrons, photons and jets.

#### For a cluster we measure:

- ◎ The energy;
- ◎ The position of the deposit;
- The direction of the incident particles;

#### Calorimeters are segmented in cells.

 Typically, a shower created by a particle interacting with the matter extends over several cells.

#### **•** Various clustering algorithms, e.g.:

- Sliding window. Sum cells within a fixed-size rectangular window.
- Topo-clustering. Start with a seed cell and iteratively add to the cluster the neighbor of a cell already in the cluster.

### CLUSTER FINDING - AN EXAMPLE

© CMS crystal calorimeter – ECAL clusters

 electron energy in central crystal ~80%, in 5x5 matrix around it ~96%.





### CLUSTER FINDING - AN EXAMPLE



#### **©** Simple example of an algorithm

Scan for seed crystals = local energy maximum above a defined seed threshold
 Starting from the seed position, adjacent crystals are examined, scanning first in φ and then in η

O Along each scan line, crystals are added to the cluster if

- <sup>©</sup> The crystal's energy is above the noise level (lower threshold)
- <sup>©</sup> The crystal has not been assigned to another cluster already

### CLUSTER FINDING - AN EXAMPLE: DIFFICULTIES

Oreful tuning of thresholds needed.

- needs usually learning phase;
- adapt to noise conditions;
- ◎ too low : pick up too much unwanted energy;
- ◎ too high : loose too much of "real" energy. Corrections/Calibrations will be larger.



### WHAT DO WE RECONSTRUCT?





Bosons

### **ELECTRONS / PHOTONS**

Final Electron momentum measurement can come from tracking or calorimeter information (or a combination of both)

- Often have a final calibration to give the best electron energy
- Working points define categories
  - E.g. loose, medium, tight
  - Trade-off: Efficiency vs Fakes
- Often want "isolated electrons"
  - Require little calorimeter energy or tracks in the region around the electron



### **ELECTRONS / PHOTONS**

Final Electron momentum measurement can come from tracking or calorimeter information (or a combination of both)

- Often have a final calibration to give the best electron energy
- Working points define categories
  - ◎ E.g. loose, medium, tight
  - Trade-off: Efficiency vs Fakes
- Often want "isolated electrons"
  - Require little calorimeter energy or tracks in the region around the electron



## ELECTRONS / PHOTONS - BACKGROUNDS

#### **Sources of backgrounds:**

Hadronic jets leaving energy in calorimeter

- While calorimeter clusters are much wider for jets than for electrons/photons there are many thousands more jets than electrons
   rate of jets faking an electron needs to be very small (~10<sup>-4</sup>)
- Complex identification algorithms are required to give the rejection whilst keeping a high efficiency

### **ELECTRONS / PHOTONS – IDENTIFICATION ALGOS**

Sional

Signal

Signal

Background

Background

0 900 10 Δ E. (MeV)



Information can be exploited using multi-variate techniques such as **likelihood discriminants** or **boosted decision trees** or **other machine learning methods**. 153

Example of different calorimeter shower shape variables used to distinguish electron showers from jets in ATLAS



 Combine the muon segments found in the muon detector with tracks from the tracking detector

- Momentum of muon determined from bending due to magnetic field in tracker and in muon system
  - Combine measurements to get best resolution
  - Need an accurate map of magnetic field in the reconstruction software
  - Alignment of the muon detectors also very important to get best momentum resolution



### MUONS ON ATLAS

**Simplified Detector Transverse View** 

"MS" - Muon Spectrometer



155

JETS



### **JET PRODUCTION PROCESSES**



#### Jets are produced:

- by fragmentation of gluons and (light) quarks in QCD scattering
- by decays of heavy Standard Model particles, e.g. W & Z
- in association with particle
   production in Vector Boson Fusion,
   e.g. Higgs
- In decays of beyond the Standard Model particles, e.g. in SUSY

JETS



At low energy, jets are more likely produced by gluon fusion.



#### JET ALGORITHMS - THEORY REQUIREMENTS

The final jet configuration should not change when

- adding extra soft particles (infrared safe) or when
- collinear splitting occurs (collinear safe)



Soft gluon radiation should not merge jets



Final jet should not depend on the ordering of the seeds...



...and on signal split in two possibly below threshold

### JET ALGORITHMS - EXPERIMENTAL REQUIREMENTS

Detector technology independent:

Insignificant effects of detector

- Noise
- Dead material
- Cracks

Data taking conditions independent:

Stability with

- Luminosity
- Pile-up
- Physics process

Easily implementable:

- Fully specified
- Fast

### JET ALGORITHM COMMONLY USED AT THE LHC

#### Algorithm

- Create a list of particles and produce all possible pairs (*i*, *j*)
- Calculate all distances between particle *i* and all other particles  $(d_{ij})$  & beam axis  $(d_{iB})$
- If  $\min(d_{ij}, d_{iB}) = d_{ij}$ , then combine *i* and *j* into a single "particle"
- If  $\min(d_{ij}, d_{iB}) = d_{iB}$ , then declare *i* as final state particle and remove it from the list of particles
- Repeat until no particles remain in the list

#### What is d<sub>ij</sub>?

- $d_{iB} = (p_{Ti}^2)^n$  and  $d_{ij} = \min[(p_{Ti}^2)^n, (p_{Tj}^2)^n] \Delta R_{ij} / R$
- For n = -1: <u>anti-k<sub>T</sub></u>.
   R: constant, the jet radius.

#### 'anti-k<sub>T</sub>'

- A 'recursive recombination' algorithm. Starts from (topo-)clusters
- Hard stuff clusters with nearest neighbor
- Various cone sizes (standard R=0.4/0.5, "fat" R=1.0)



161

### JET CALIBRATION

- Correct the energy and position measurement and the resolution.
- Account for:

Instrumental effects Detector inefficiencies 'Pile-up' Electronic noise Clustering, noise suppression Dead material losses Detector response Algorithm efficiency

#### Physics effects

Algorithm efficiency 'Pile-up' 'Underlying event'



162
#### JETS AND PILE-UP



Multiple interactions from pile-up



#### **B-JETS**

b-hadrons have a lifetime of ~ 10<sup>-12</sup> s.
They travel a small distance (fraction of mm) before decaying.
A "displaced vertex" creates a distinct jet, so b-jets can be tagged (b-tagged).
b-tagging uses sophisticated algorithms, mostly multi-variate (machine learning).





### MISSING TRANSVERSE MOMENTUM – ME<sub>T</sub>



In the transverse plane:

$$\Sigma_i \vec{p}_{T,i} = 0$$

So for what we can't directly measure (e.g. neutrinos)

$$E_{\rm T}^{\rm miss} = -\Sigma_i \vec{p}_{T,i}$$



### MISSING TRANSVERSE MOMENTUM – ME<sub>T</sub>



In the transverse plane:

$$\Sigma_i \vec{p}_{T,i} = 0$$

OR DARK MATTER CANDIDATES!

So for what we can't directly measure (e.g. neutrinos)

$$E_{\rm T}^{\rm miss} = -\Sigma_i \vec{p}_{T,i}$$

#### **Simplified Detector Transverse View Muon Spectrometer Toroids** HadCAL **EMCAL** photon Solenoid electrol TRT SCT **Pixels** muon κv

# PARTICLE FLOW FOR HADRONIC RECONSTRUCTION

#### PARTICLE FLOW



#### PARTICLE FLOW



### PARTICLE FLOW

- Reconstruct and identify all particles, photons, electrons, pions, …
- Use best combination of all subdetectors for measuring the properties of the particles.
- First used at LEP (ALEPH) and then at the LHC (CMS).



#### JETS IN PILE-UP



Multiple interactions from pile-up

#### JETS IN PILE-UP



173

Multiple interactions from pile-up



**Resolution**: the quality with which we measure the jet momentum.



**Resolution**: the quality with which we measure the jet momentum.



**Resolution**: the quality with which we measure the jet momentum.



Significant improvement for low-pT jets. Similar for MET.



In Jet Energy resolution and uncertainty, large improvements with respect to calo jets!

#### A COMPARISON



 PF jets (CMS) and calo jets (ATLAS) have similar performance.
 Particle reconstruction always needs to be optimized depending on the detector technologies and experimental requirements.

#### **A COMPARISON**



PF jets (CMS) and calo jets (ATLAS) have similar performance.
 Particle reconstruction always needs to be optimized depending on the detector technologies and experimental requirements.





Objective:Trigger ("online") reconstruction same as "offline".Problem:Time. Trigger decision needs to be taken fast.Solution:Simplification.Challenge:Clever simplification = good performance.



E.g. track reconstruction in regions of interest and simplified MET calculation.

#### **ONLINE RECONSTRUCTION**

trigger efficiency =  $\frac{\# \text{ events passing offline selection \& trigger}}{\# \text{ events passing offline selection}}$ 



Clever ideas need to be deployed to bring online closer to offline, making efficiency curves **sharper** and **plateau closer to 1**.



• To profit fully from an improvement in reconstruction, the relevant algorithm has to be used at the relevant trigger selections to provide **optimal online-to-offline correlation**.



Variable A: e.g. leading jet pT

### **EFFICIENCY MEASUREMENTS**

Relevant beyond the trigger...

#### TAG AND PROBE

- Select events based on requirements on one object (tag) and study the response of the second object (probe), not used in the event selection, using some constraint such as the Z mass.
  - e.g.  $Z \rightarrow \tau \tau$  events.
  - Typically used for measurement of the identification efficiency

#### ORTHOGONAL SAMPL

- Measure directly the efficiency on an independent, orthogonal sample.
  - e.g. jet trigger efficiency on a sample triggered by muons,

#### BOOTSTRAP METHOD

• The efficiency,  $\varepsilon_B$ , of a selection B, inclusive compared to a selection A, can be determined in a sample of events passing selection A (provided that  $\varepsilon_{A}$  is measurable):  $\varepsilon_{B} = \varepsilon_{B|A} \times \varepsilon_{A}$ . ε ε<sub>ΒΙΑ</sub>

• e.g. trigger efficiencies, say B: tau50 loose & A: tau16 loose

## PHYSICS MENUS

Trigger selection	2015 offline threshold (GeV)	2016 offline threshold (GeV)	2017 offline threshold (GeV)	Representative physics case	
Peak Luminosity	5x10 <sup>33</sup> cm <sup>-2</sup> s <sup>-1</sup>	1.2X10 <sup>34</sup> cm <sup>-2</sup> s <sup>-1</sup>	1.7x10 <sup>34</sup> cm <sup>-2</sup> s <sup>-1</sup>		
isolated single e	25	27	27	"Main" triggers. Thrs driven by Higgs (ZH, WH), Top, SUSY.	
isolated single $\boldsymbol{\mu}$	21	27	27		
di-γ	40,30	40,30	40,30	Higgs (H→γγ, HH→bbγγ).	
di-τ	40,30	40,30	40,30	Higgs (H→ττ, HH→bbττ), SUSY.	
four-jet	45	45	45		
MET	180	200	200	SUSY, Higgs, exotics	

Offline selections from which the triggers are "usable", i.e. at efficiency plateau or highly efficient otherwise

#### **RECONSTRUCTING PARTICLES**





#### TAUS

Tau Decay Mode			B.R.
Leptonic		$\tau^{\pm} \rightarrow e^{\pm} + \nu + \nu$	17.8%
		$\tau^{\pm} \rightarrow \mu^{\pm} + \nu + \nu$	17.4%
Hadronic	1-prong	$\tau^{\pm} \rightarrow \pi^{\pm} + \nu$	11%
		$\tau^{\pm} \rightarrow \pi^{\pm} + \nu + n\pi^{\circ}$	35%
	3-prong	$\tau^{\pm} \rightarrow 3\pi^{\pm} + \nu$	9%
		$\tau^{\pm} \rightarrow 3\pi^{\pm} + \nu + n\pi^{\circ}$	5%
Other			~5%

Hadronic tau reconstruction extremely challenging
 Using multi-variate (machine learning) techniques
 based on track multiplicity and shower shapes



### TOP, W, Z



#### AND THE HIGGS!



#### HOW ABOUT NEW PARTICLES?

• These decay to Standard Model particles or create  $\ensuremath{\mathsf{ME}_{\mathsf{T}}}$ 



#### **PHYSICS ANALYSES**

