Data Analysis Training in CMS

Gabriele Benelli (Brown University)

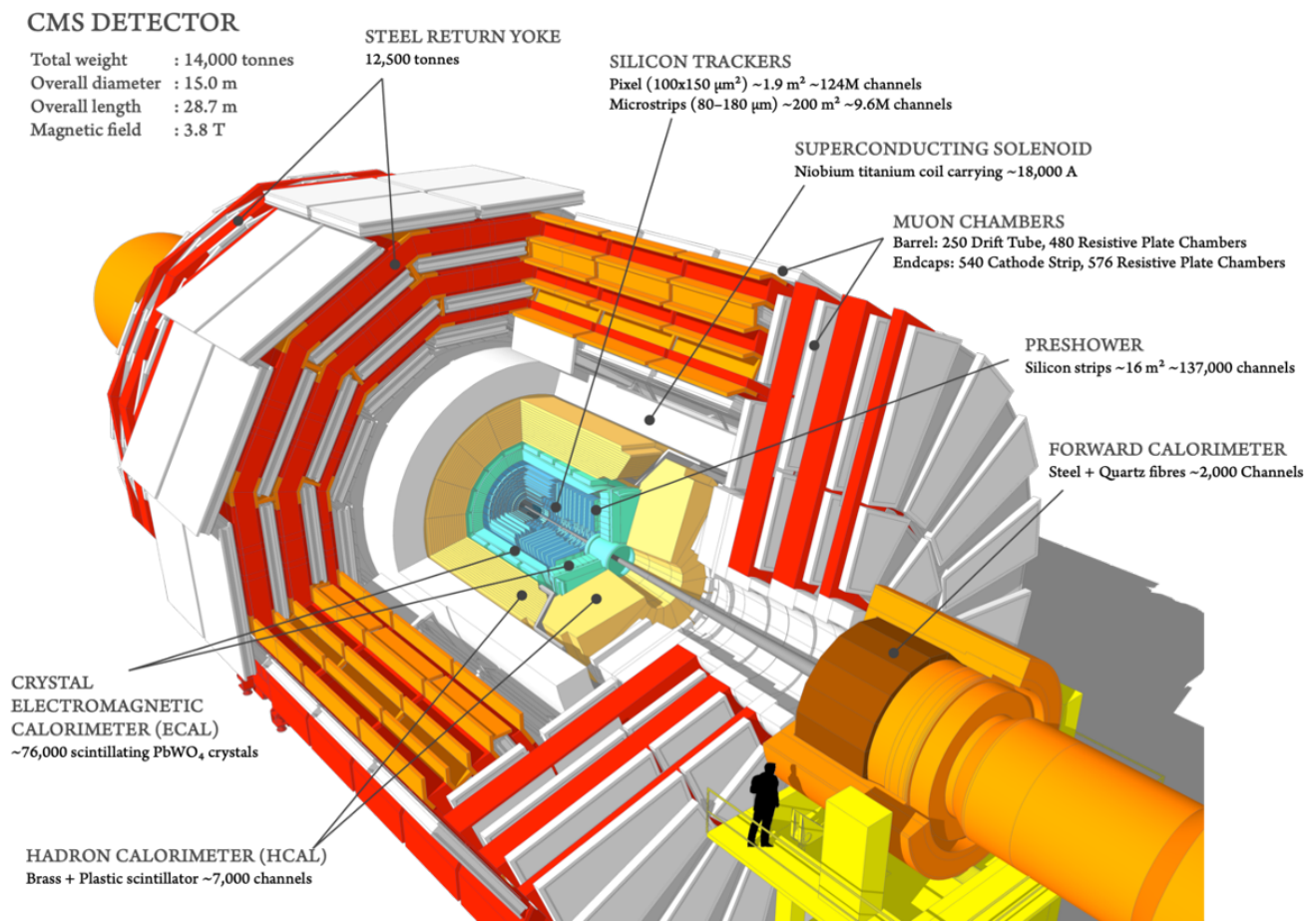
Data Analysis Training in HEP Experiments (HSF), Jul 27th 2022

Data Analysis Training in CMS

- A quick overview of CMS and the scope of the trainings
- Some details on our main training programs:
  - CMS Data Analysis Schools
  - Hands-on Advanced Tutorial Sessions
  - Topical workshops (Statistical tools, Machine Learning, Trigger)
- Some reflections along the training challenge survey questions based on our experience in CMS

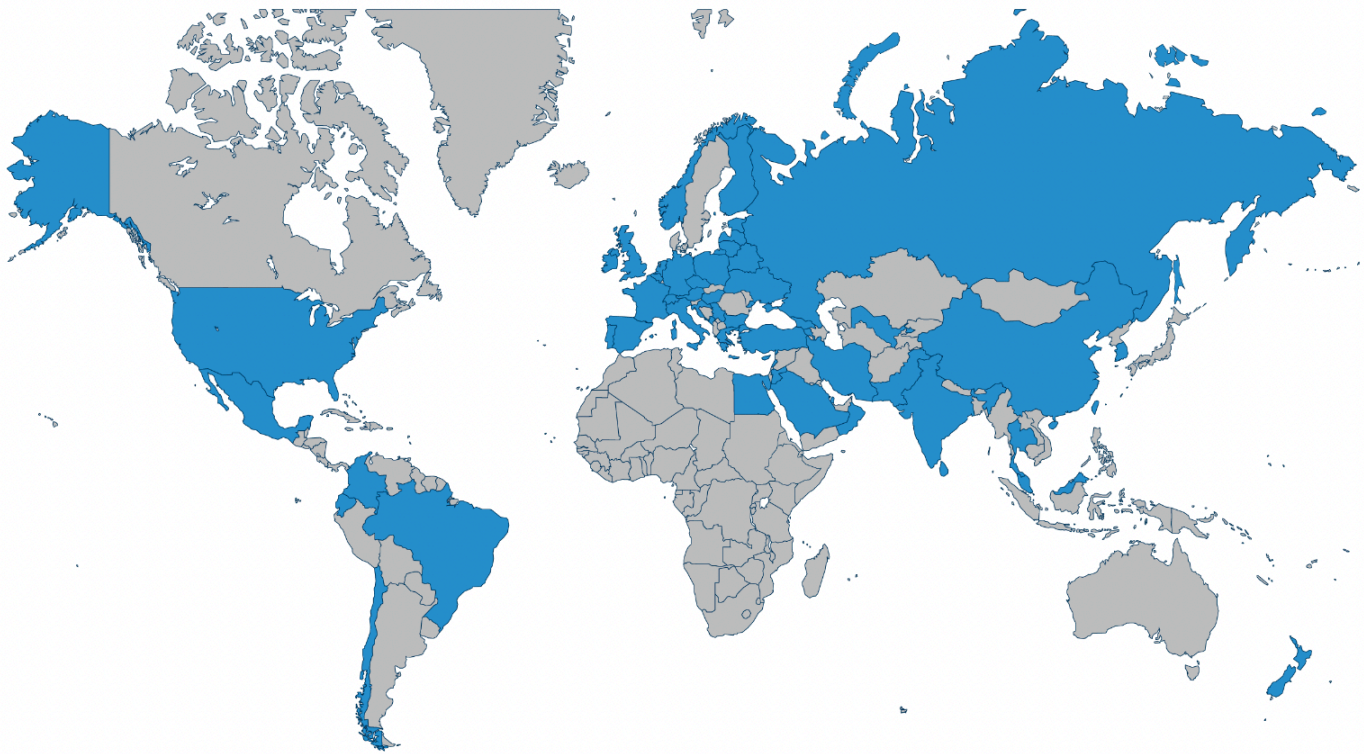
CMS

- Compact Muon Solenoid
- Large General-Purpose LHC experiment analyzing proton-proton collisions at the highest center-of-mass energies
- Running for over a decade, just entered Run 3 with 13.6 TeV collisions
- Over 1000+ publications



The CMS Collaboration

3103	1031	269	241	54
PHYSICISTS (1050 STUDENTS)	ENGINEERS	TECHNICIANS	INSTITUTES	COUNTRIES & REGIONS



- Large span of timezones (19hrs!)
- Large number of newcomers every year

Data Analysis Training in CMS

- CMS Induction Courses (once or twice a year since 2014, so far 13)
  - Newcomers introduction to the experiment and the collaboration
  - Talks/discussions (from Physics/Data Analysis to Tracking/Calorimetry/Trigger/Machine Learning/Offline and Computing, from Data Taking/Run Coordination to the various committees/institutions within CMS, Diversity, Collaboration Board, Secretariat, Communication, CMS visits etc.)
  - Split in two days, currently hybrid (in-person+Zoom), fully recorded
- **CMS Data Analysis Schools [DAS]** (2 or 3 a year since 2011, so far 25)
  - The prototype of Data Analysis Training in CMS
  - Taking newcomers from zero to a full physics analysis
  - From basics computing skills, through all ingredients for analysis, to actually participate in a physics analysis exercise, presenting results
- **Hands-on Advanced Tutorial Sessions [HATS]** (15/20 a year since 2013)
  - Focused tutorials on specific physics objects or tools for data analysis
  - Offered at Fermilab LPC (LHC Physics Center) in Spring/Summer

Data Analysis Training in CMS

- CMS Physics Object Schools [CMSPOS] (only twice)
  - Emphasis on training new contributors to Physics Object Groups (POGs, B-tagging, Tracking, EGamma, Muon etc) and Detector Performance Groups (DPGs, Tracker, ECAL, HCAL, CSC, DT, RPC, etc)
  - Statistical tools, GPU/heterogeneous computing, Trigger, Data Quality Monitoring
- CMS Upgrade Schools [CUPS] (only once)
  - Emphasis on training new contributors to Phase II/ HL LHC upgrades (test-beam, sensor characterization, thermal/mechanical lab measurements, tuning of operational parameters, test DAQ systems, design of tracking/muon systems)
- Graduate-level advanced courses (2 or 3 a year)
  - Hybrid (in person+remote) lectures with homework and exams
  - Wide range of analysis related topics: Computational Physics, Statistics, Detectors, Machine Learning, etc.
- Dedicated workshops/hackathons
  - Driven by DPG or POG or other groups in preparation to major developments or at critical times (before data-taking, after release of new tools etc)
  - Statistical tools, GPU/heterogeneous computing, Trigger, Data Quality Monitoring

CMS Data Analysis Schools (DAS)

- Concept started before actual data taking started, but first official CMSDAS was in January 2011
- Emphasis on teamwork and hands-on learning
- Full coverage of all data analysis aspects
- Evolved through the years and adapted to the Covid-19 era
- Offered yearly in January at Fermilab LPC, usually a couple more per year in other locations (Pisa, Taipei, DESY, Kolkata, Bari, Daegu, Beijing, CERN)
- Typical attendance 50 to 70 students (from undergraduates to faculty) and around 50 facilitators
- Major logistical and organizational effort

CMS Data Analysis Schools (DAS)

- Collaborative and active learning

(disclaimer pictures from the Before Times)






- From our Welcome to CMS DAS presentation:
- "A high-intensity, engaging camp, that will charge you up, change your mindset, and give you a reason to engage in cutting edge basic research and continue with it happily ever after (we hope)!"

CMS Data Analysis Schools (DAS)

- Over the years the structure has evolved slightly, but the basic structure:
  - Pre-exercises (mandatory with deadline before start of the school)
  - Lectures (Introduction to CMS Physics, LHC, CMS Detector, Software/ Analysis Tools, Diversity and Inclusion, Communications/Outreach)
  - Short Exercises (covering objects and basic analysis ingredients)
  - Writers PUB (going over the publication process of a paper in CMS)
  - Long Exercises (a set of complete physics analysis exercises)
  - Mini-Symposium (presentation of the results from all teams)
- Networking is a key ingredient of CMS DAS:
  - All participants are organized in small teams, based on the long exercise they are assigned to
  - Their commitment and responsibility to the team (and not to facilitators or organizers) is the key to each team (and team member) success
  - While all team members participate to the same long exercise, team members are assigned different short exercises so that the team as a whole can have full coverage of the needed tools for the long exercise analysis
  - All team members participate both in the slides preparation and in the final presentation at the Mini-Symposium at the end

CMS Data Analysis Schools (DAS)

- Networking is a key ingredient of CMS DAS:
  - All participants are organized in small teams, based on the long exercise they are assigned to

January 4					January 10	January 10-14	
Short Exercises						Long Exercise Assignment	
Period 1	Period 2	Period 3	Period 4	Period 5	Writers PUB		
Tuesday 10:45->11:30	Tuesday 11:35->12:20	Tuesday 13:20->14:05	Tuesday 14:10->14:55	Tuesday 15:10->15:55	Monday 19:00->20:45	Mon 11:00->Fri 12:00	7 short exercises
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Jets
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Statistics
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Tagging
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Machine Learning
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Generators
Statistics	Visualization	Tagging	Generators	Machine Learning	Sarah Eno	Double Higgs to 4b final state	Visualization
Statistics	PU/MET	Tagging	Jets	Machine Learning	Sarah Eno	Double Higgs to 4b final state	PU/MET
Statistics	PU/MET	Tagging	Jets	Machine Learning	Sarah Eno	Double Higgs to 4b final state	
Statistics	PU/MET	Tagging	Jets	Machine Learning	Sarah Eno	Double Higgs to 4b final state	
Statistics	PU/MET	Tagging	Jets	Machine Learning	Sarah Eno	Double Higgs to 4b final state	
Statistics	PU/MET	Tagging	Jets	Machine Learning	Sarah Eno	Double Higgs to 4b final state	

  - Their commitment and responsibility to the team (and not to facilitators or organizers) is the key to each team (and team member) success
  - While all team members participate to the same long exercise, team members are assigned different short exercises so that the team as a whole can have full coverage of the needed tools for the long exercise analysis
  - All team members participate both in the slides preparation and in the final presentation at the Mini-Symposium at the end

CMS Data Analysis Schools (DAS)

- Evolution during covid:
  - Anticipated the Pre-Exercises release (essential to make sure people can hit the ground running)
    - Mattermost support
  - Factored out some short exercises as Offline short exercises (ROOT, NanoAOD, PPD, Luminosity), fully remote some with video recordings, all with Mattermost support
  - Stretched the length of DAS to two full weeks
  - First week kick-off and asynchronous short exercises
    - Introductions, team building, plenaries
    - Short Exercises kick-off (live)
    - Asynchronous work through material with Mattermost support (some with extra live office hours)
    - Short Exercises wrap-up (live)
  - Second week more plenaries and live long exercises
  - Recordings of all live sessions
  - Heavy use of Mattermost understanding about support expectations from facilitators
  - Release logistical constraints
  - Social events replacement (Scribble I/O, Rubik's cube competition etc)

CMS DAS

- Last year, for example, we offered 13 short exercises:

January 4					January 12
Short Exercises					
Period 1	Period 2	Period 3	Period 4	Period 5	Writers PUB
Tuesday 10:45->11:30	Tuesday 11:35->12:20	Tuesday 13:20->14:05	Tuesday 14:10->14:55	Tuesday 15:10->15:55	Tuesday 19:00->20:00
Statistics	Forward Protons	Tagging	Jets	Machine Learning	Sarah Eno
Tracking/Vertexing	PU/MET	Muon	Generators	Electron/photon	Jacobo Konigsberg
Triggers	Visualization	Tau			

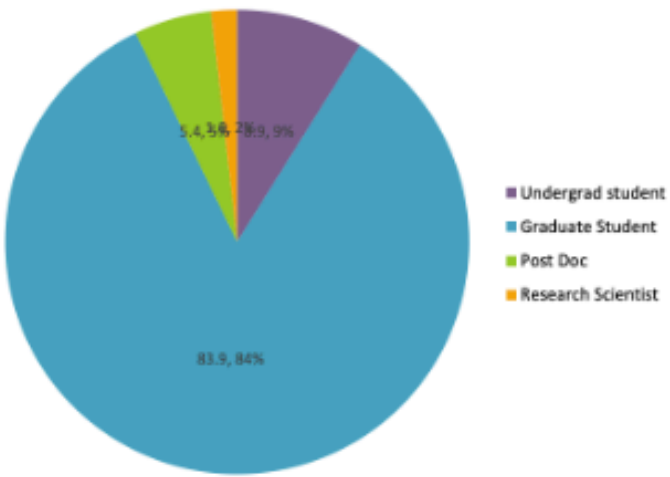
- And 8 long exercises (which resulted in as many teams):

Double Higgs to 4b final state
B2G Search for b*->tW all hadronic
TTGamma cross-section
Exclusive production of lepton pairs
Top quark mass at 13 TeV
SUSY hadronic with top tagging
Contact Interactions
Z -> tau tau cross-section at 13TeV

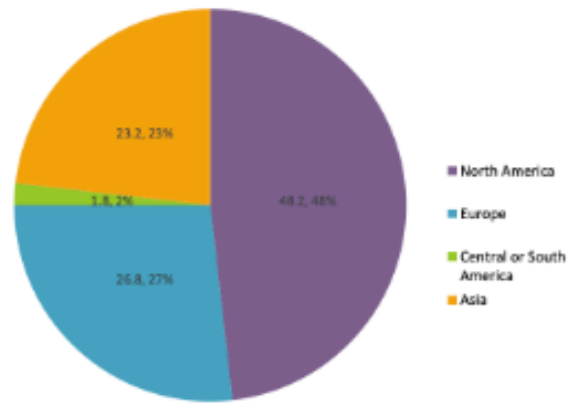
CMS DAS

- Feedback

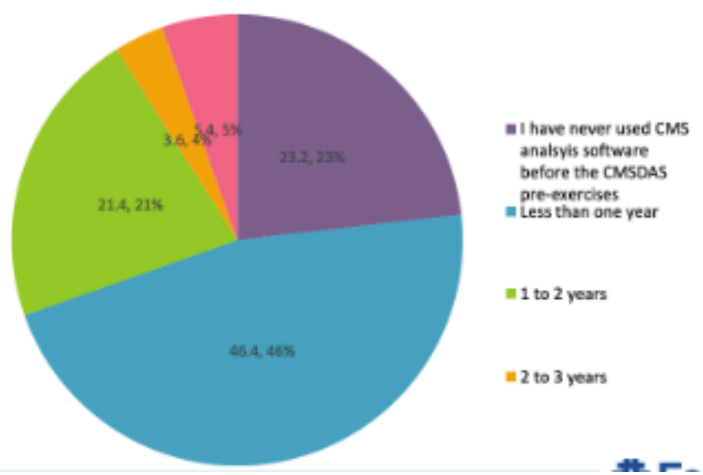
CMS DAS participation level



In which region is your institute located?



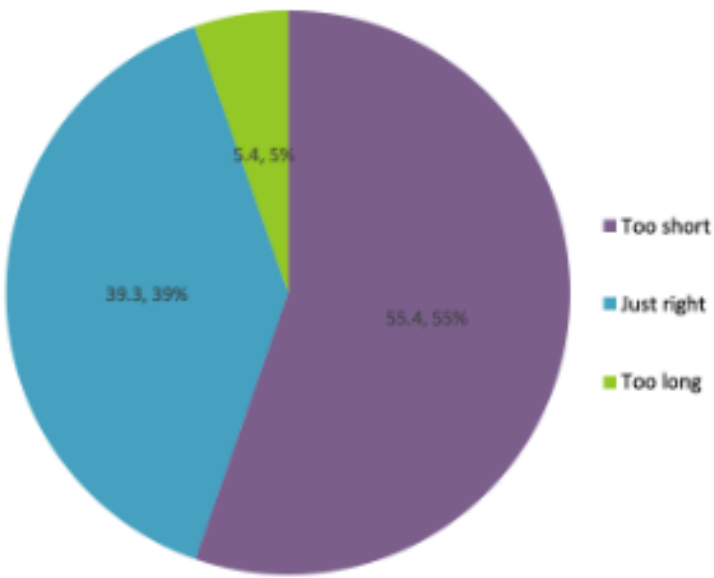
For how many years have you been using CMS software?



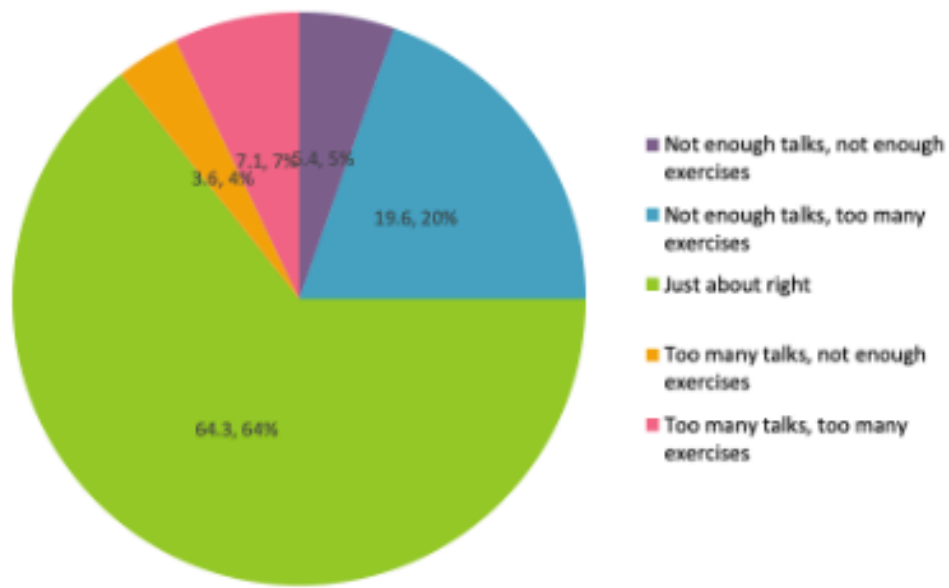
CMS DAS

- Feedback

Is the two weeks (including the asynchronous time) the right length of time for a virtual CMSDAS?



The number of talks and number of exercises at CMSDAS were



CMS DAS

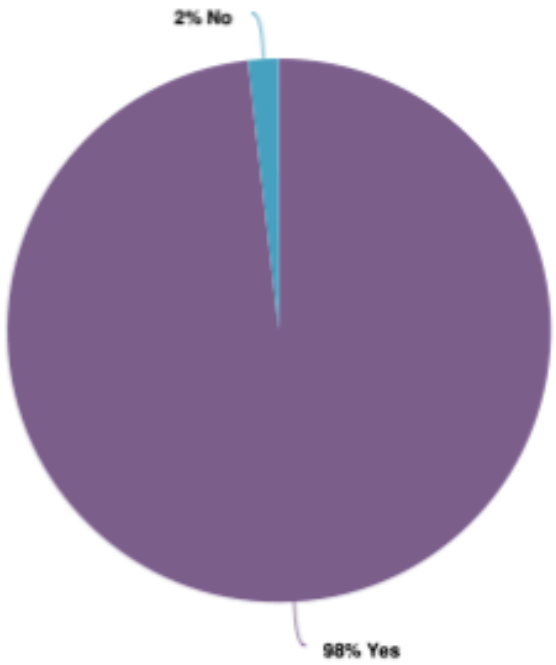
- Feedback

	Timing	Content/Support
Kick-off day (Tue Jan 4)	★★★★½ Count: 48 Not Applicable: 6	★★★★★ Count: 47 Not Applicable: 5
Plenaries (Tue Jan 4, Fri Jan 7, Mon Jan 10, Tue Jan 11)	★★★★½ Count: 51 Not Applicable: 4	★★★★★ Count: 54 Not Applicable: 1
Long Exercise Meet up (Tue Jan 4)	★★★★½ Count: 52 Not Applicable: 3	★★★★★ Count: 52 Not Applicable: 3
Short Exercise Kick-off (Tue Jan 4)	★★★★☆ Count: 48 Not Applicable: 4	★★★★½ Count: 50 Not Applicable: 4
Asynchronous short exercises (Tue - Wed Jan 5-6)	★★★★☆ Count: 53 Not Applicable: 1	★★★★½ Count: 55 Not Applicable: 1
Offline Short Exercises (released Mon Dec 13)	★★★★☆ Count: 43 Not Applicable: 11	★★★★½ Count: 44 Not Applicable: 11
Short Exercise Wrap-up (Fri Jan 7)	★★★★☆ Count: 53 Not Applicable: 2	★★★★½ Count: 52 Not Applicable: 3
Writers PUB (Mon Jan 10)	★★★★☆ Count: 40 Not Applicable: 13	★★★★★ Count: 36 Not Applicable: 18
Communication and Outreach (Tue Jan 12)	★★★★½ Count: 48 Not Applicable: 5	★★★★★ Count: 50 Not Applicable: 5
Optional Social events (Wed Jan 12, Thurs Jan 13)	★★★★½ Count: 39 Not Applicable: 14	★★★★½ Count: 41 Not Applicable: 14
Long Exercises (Mon-Thurs Jan 10-13)	★★★★☆ Count: 54 Not Applicable: 0	★★★★½ Count: 55 Not Applicable: 0

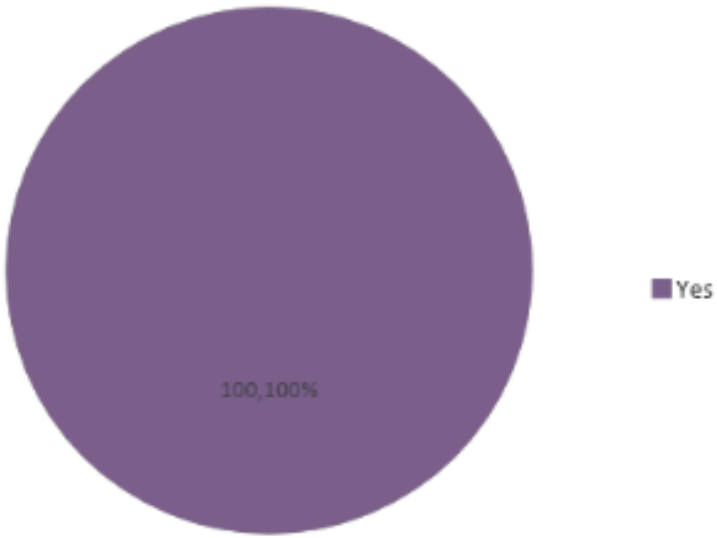
CMS DAS

- Feedback

Has CMS DAS been a valuable experience for you?



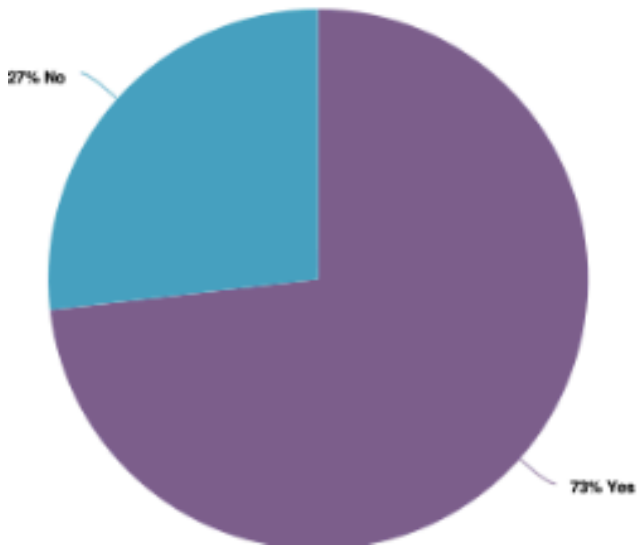
Would you recommend CMSDAS to your colleagues?



CMS DAS

- Feedback

Has CMSDAS enabled you to make new connections to CMS members or groups that you envisage will be helpful in your future analysis and other work for CMS?



Would you be willing to serve as a facilitator for future schools?

Hands-on Advanced Tutorial Sessions (HATS)

2021 HATS@LPC (and lectures)

- Statistics in Particle Physics (October 4, 2021 - December 1, 2021, Mondays & Wednesdays 3:00pm-4:30pm; Zoom) - Indico Agenda (CMS CERN account required)
  - Course recordings and materials (publicly available)

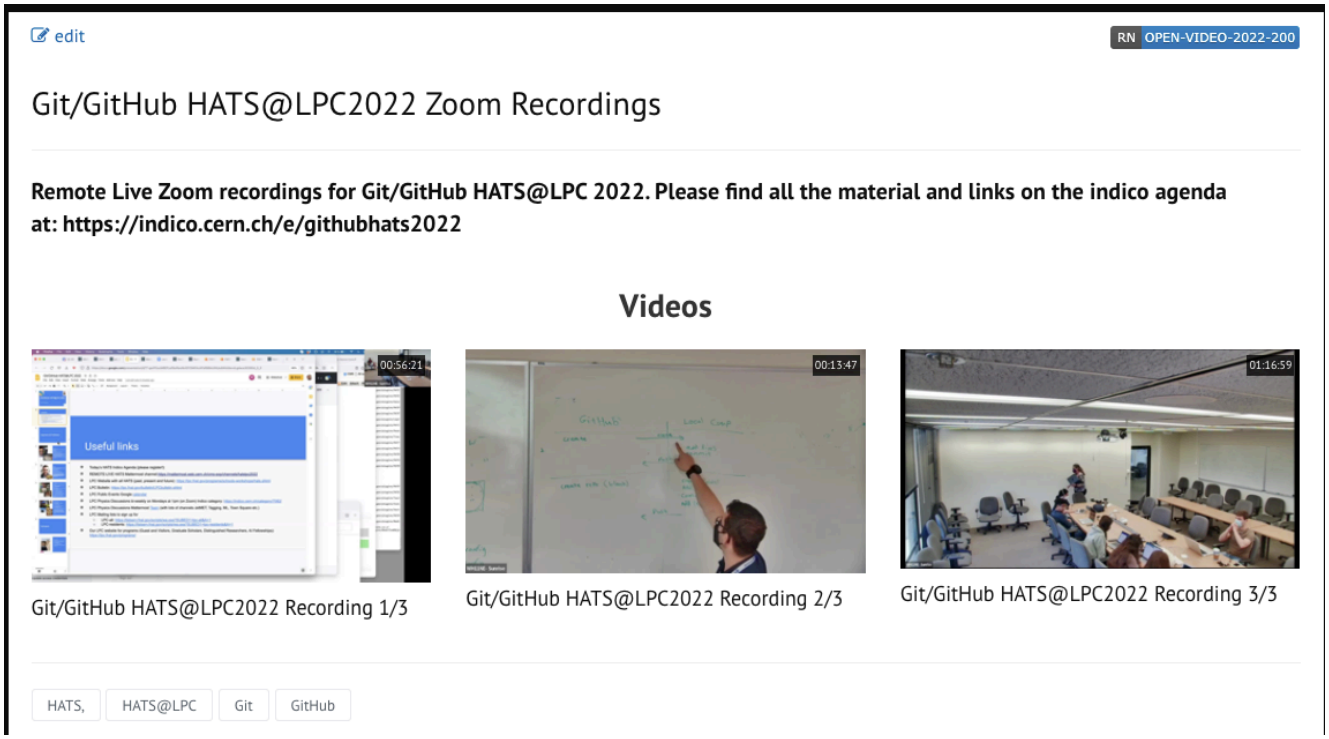
- Condor/CMSConnect/CRAB3 HATS@LPC (August 23, 2021) - Indico Agenda
- Visualization HATS@LPC (August 20, 2021) - Indico Agenda
- MET HATS@LPC (July 28, 2021) (Converted to OFFLINE) - Indico Agenda, Twiki
- Tau HATS@LPC (July 26, 2021) (Converted to OFFLINE) - Indico Agenda, Twiki
- Trigger HATS@LPC (July 22, 2021) - Indico Agenda
- Machine Learning HATS@LPC (July 12, 2021) - Indico Agenda
- Jets Energy Corrections and Pile-Up Mitigation HATS@LPC (June 29, 2021) - Indico Agenda
- Jet Algorithms and Substructure HATS@LPC (June 28, 2021) - Indico Agenda
- Effective Scale Out Techniques HATS@LPC (June 23, 2021) - Indico Agenda
- Columnar Analysis Tools HATS@LPC (June 16, 2021) - Indico Agenda
- Uproot and Awkward Array for columnar analysis HATS@LPC (June 14, 2021) - Indico Agenda
- Data and MC Preparation HATS@LPC (June 10, 2021) - Indico Agenda
- Container HATS@LPC (June 8, 2021) - Indico Agenda
- Git/GitHub HATS@LPC (June 2, 2021) - Indico Agenda

The following HATS will be provided as OFFLINE HATS, their Twikis will be updated in the next few weeks and users are encouraged to go through the exercises on their own reaching out through the dedicated Mattermost channels, provided for each of them.

- Jupyter and PyROOT HATS (OFFLINE) - last year's HATS
- Muon HATS@LPC (OFFLINE) - Twiki
- Electron and Photon HATS@LPC (OFFLINE) - Twiki
- (B,t,H,W,Z) Tagging HATS@LPC (OFFLINE) - Twiki
- Generators HATS@LPC (OFFLINE) - last year's HATS
- The Science and Lore of Instrumentation for Particle Physics Course Second Semester (Jan 25 - Dec 6, 2021) - Indico Agenda

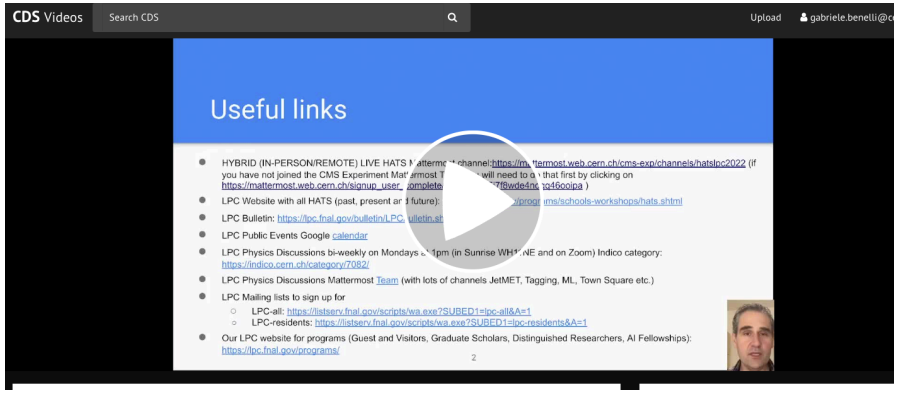
Hands-on Advanced Tutorial Sessions (HATS)

- Hands-on learning, typically with some pre-requisite training:
  - Partly lectures
  - Partly exercises
- Format has been evolving to adapt to audience in Covid-19 times, strictly connected to DAS
- All recorded (videos.cern.ch), with transcripts

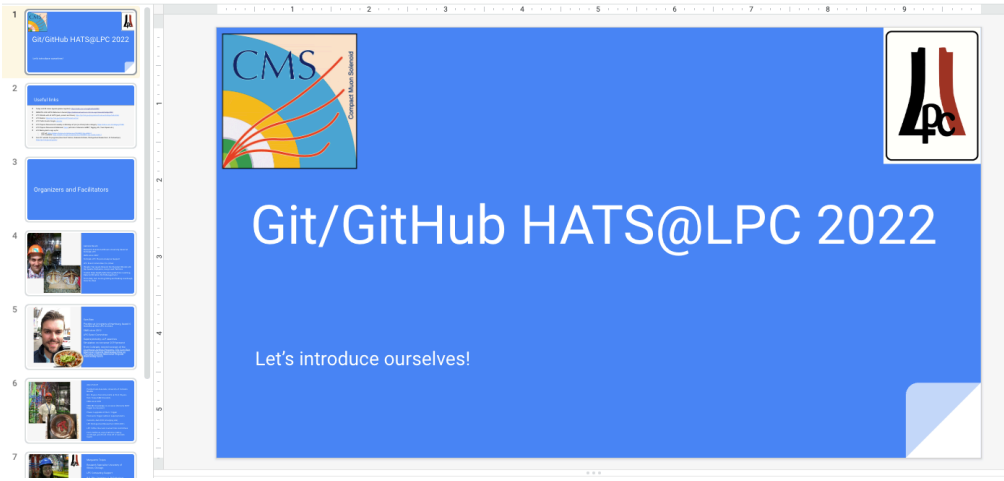


Hands-on Advanced Tutorial Sessions (HATS)

- Covid optimizations:
  - General remarks packaged into a video/set of slides



  - Google Slides for introductions



  - SliDo/Google Forms/Mattermost polling for interactive polling
  - Use of Carpentries format for pre-exercises simplifying maintenance (also for some of the short exercises and HATS)
- Challenge
  - Zoom fatigue, interactions, networking, collaboration spirit and continued interactions

HATS

- Special issues during the pandemic prevented the fully hands-on ones:
  - Particle Flow Lego HATS
    - Use a Lego physical model to do Particle Flow reconstruction by hand and identify an event
  - CMS Upgrade Detectors HATS
    - Experience several hardware hands-on exercises involving oscilloscopes etc (characterization of sensors, production of scintillating fibers, testing of readout electronics)
- Ideas to improve/try/add
  - Luminosity HATS
  - Higgs Combine (statistics tool) HATS
  - Generic Python/Jupyter (synergy with HSF curriculum)
  - Move to Software carpentry model for all to simplify maintenance
  - Automate some of the logistics

Reflections

Brainstorming and Networking

- The value of in-person interaction is made more evident by its lack in the past couple of years:
- Brainstorming and talking among colleagues in an informal setting where multiple conversations can happen in parallel and there is no meeting goal (having a coffee/tea, or eating lunch) is the seed of many important developments and ideas and it increases dramatically the efficiency of our collaborations
- Networking is a major benefit of in-person interactions, as one of the most important results of participating to DAS and HATS is to know who to ask questions about specific topics, mapping out a network within the collaboration
- Zoom and Mattermost while extremely precious, are not ideal to facilitate those kind of conversations
- The worst damage is to people who do not know what they lost with fully remote collaboration: newcomers who do not have a networking social capital as they navigate these hybrid and remote times.
- The interactions with facilitators, POG/PAG conveners for a remote center like the LPC are fundamental in establishing synergies and collaborations that extend beyond the training, a major benefit (including shifts-taking

Advanced Training for Developers

- A personal opinion on the concept of trainings for developers:
  - While covering the bases with technical knowledge is important (and the previous trainings provide those), ultimately “advanced” becomes very quickly very specific and sitting next to some expert to understand the scope of the problem, draft a solution and interact repeatedly with a few people to get feedback is what is needed to get to the last step and become a developer.
  - Without a specific project the training is not very effective risking to drain more effort from experts than real gain by new experts in training
  - The practical limit to what can be reasonably packaged/worked on as training for developers should be determined by the typical questions/issues that arise most frequently in developers communication channels

More Ideas for improvement

- Interaction with Physics Object Groups and Physics Analysis Groups to be strengthened to get more facilitators, and drive the content relevance
  - Analysis tools evolution
  - Data format evolution
  - Heterogenous resources
- Rely on the lesson learned to be more inclusive (time of day, asynchronous plus synchronous, asynchronous support on Mattermost)
  - But also encourage more in-person participation
    - More regional duplication
- Ideas about collaborating with tools like GatherTown (office hours?)
- Streamline the core curriculum

Reflections

- Challenges
  - Documentation is currently heterogeneous (transitioning to alternatives to traditional Twikis)
  - Training organization is extremely time consuming, lots of communications necessary
  - People's engagement online
  - Getting back to in-person interactions (in a hybrid scenario) after a long time of fully remote operations
- Reasons for hope
  - All of the work described is based on the good will and the collaboration spirit embodied by facilitators and organizers
  - Training, similarly to operations, highlights the collaborative nature of our endeavor, seeding the future of our experiments
  - Attention to diversity and inclusion is growing and definitely training is reflecting it, being much more accessible than in the past, reaching people previously not served.

Back-up