Maintaining and improving of the training program on the analysis software in CMS

S Malik¹, F Hoehle², K Lassila-Perini³, A Hinzmann⁴, R Wolf⁵ and I Shipsey⁶

¹ University of Nebraska-Lincoln, Lincoln, NE, U.S.A.
² Rheinisch-Westfaelische Tech. Hoch (RWTH), Aachen, Germany
³ Helsinki Institute of Physics, Helsinki, Finland
⁴ CERN, Geneva, Switzerland
⁵ Massachusetts Institute of Technology, MA, U.S.A.
⁶ Purdue University, IN, U.S.A.

Email: malik@fnal.gov

Abstract. Since 2009, the CMS experiment at LHC has provided intensive training on the use of Physics Analysis Tools (PAT), a collection of common analysis tools designed to share expertise and maximize productivity in the physics analysis. More than ten one-week courses preceded by prerequisite studies have been organized and the feedback from the participants has been carefully analyzed. This note describes how the training team designs, maintains and improves the course contents based on the feedback, the evolving analysis practices and the software development.

1. Introduction
The large size of the CMS [1] community with 3500 members, its geographical spread over 5 continents, the distributed computing resources [2], and multiple time zones adds to the challenge of already complex analysis tools and software needed to successfully perform physics analysis. To tackle this challenge scenario CMS has instituted an extensive program of tutorials and training [3] under the CMS Physics Support [4]. Tutorials are organised on many software and physics analysis tools as well as hands on physics analysis example exercises.

• Physics Analysis Toolkit (PAT) [5] – collection of CMS specific analysis tools
• CMS Data Analysis School (CMSDAS) [6] – hands on learning experience in physics analysis from experts in the field
• Roostat and RooFit [7] - to learn and implement Frequentists, Bayesian and Likelihood based methods for statistical calculations
• Python language [8] - used to configure CMS Software (CMSSW)
• CRAB (CMS Analysis Remote Builder) [9] - a utility to create and submit CMSSW jobs to 4 distributed computing resources
• Fireworks [10] - 3D-accurate display to visualize proton-proton collisions

The tutorials provide hands on training experience from experts that are drawn from the collaboration itself. While some tutorials, like CRAB, provide one day training others like, PAT and CMSDAS, are spread over the duration of a week and, in addition, require pre-training before the actual tutorial happens. The PAT tutorial on physics analysis tools was started in 2008 and since then its structure has evolved and improved to adapt to the user needs and feedback. To maintain the rigor of the training and make it more complete the structure of the PAT tutorial has been adapted to the CMSDAS where one can apply the analysis tools learned to perform hands-on-training in physics analysis. The new paradigm in CMS where many scientists are not located at the host lab (CERN [11]) and the face-to-face interaction still remains indispensable, has lead to the establishment of LHC [12] regional centres that not only serve as a hub for the users but also as venues for the training. In this paper we look at the analysis tools training structure, its evolution and adaption to the CMSDAS and the feedback from users. The growth of LHC regional centres serving the CMS community is also described.

2. Organisation and Training Workflow of PAT tutorial
The requirements and the considerations that lead to establishing the PAT training program is described elsewhere [13]. Here we briefly discuss its structure. The tutorial is preceded by a 3-month preparatory phase that commences with identifying a person who is overall responsible for the tutorial. This leads to the setting up of a registration and a course website [14] and announcing the tutorial to the entire collaboration by the email forum called Hypernews. The announcement should be made as early as possible for the participants to make necessary travel arrangements. Due to high volume of meetings in the collaboration it is very important to book the physical rooms (capacity ~30 people) at CERN (where the tutorial is mostly held) early enough. The responsible person also sets up a detailed agenda for the entire week on the Indico [15]. A virtual room for any remote participant is also set up using EVO (Enabling Virtual Organization) [16].

The next step in the preparatory phase is very important as it involves identifying four categories of people who can perform the corresponding tasks. These tasks compose the tutoring material – lectures, exercises and twikis [17].

• Lecturers - give lectures/exercises session
• Tutors - assist participants in exercises
• People who prepare/update the material (on twikis)
• People who cross check and debug twikis

The PAT tutorial week is proceeded by a period of two weeks when the registered participants are required to complete a set of pre-exercises before they arrive for the training. This is to prepare them well for the intense one-week training period. The pre-exercises topics include basics of C++, computing account access, access to CMS code repository, access to data, run grid jobs, and python configuration. To make sure that all participants go through the pre-exercise material and answer questions, an answer form is required to be filled. This form is available at the course website. The course website, besides the link to the course material, also contains the feedback form as well as a discussion forum to post questions and answers during the course.
2.1. Feedback from the participants
The user feedback is the most important measure of success of any training. It is the key to further improvement. Besides logistics and suitable time to have the tutorial, some feedback related to the course material was as follows:

- Pre-exercises useful
- Course rated very good
- Need more pre-requisite exercises
- More guided exercises in class room
- Time needed for exercise not enough
- Some want to join the tutorial team (collaborative spirit)
- As material is too much to cover in 1-week, categorise into “must be done” and “nice if done”
- Guided physics analysis exercise to apply new methods learned

2.2. Evolution and improvements
Since 2009 12 PAT tutorials have been held at a frequency of 2-3/year. The feedback from each has been fed into the next leading to continuous improvement. The pre-exercises were established based on the practical experience of the first few PAT tutorials. It was observed that many participants instead of learning the actual PAT material were spending the precious time dealing with some basic issues. For example, some were not familiar with basic computing environment setup commands, how to check out the code from CVS, compile the code, understand python configuration language etc. This would waste a lot time and frustrate the tutors and those participants who were better prepared. To overcome this situation, pre-exercises were established and it was assumed that participants would come prepared with basic stuff so that they could get a jump-start on the real PAT material.

All of the exercises for each day are marked under three categories

- Basic exercise, which is obligatory for the PAT Tutorial.
- Continuative exercise, which is recommended for the PAT Tutorial to deepen what has been learned
- Optional exercise, which shows interesting applications of what has been learned.

During the morning session each day, lectures take place and the basic knowledge is spread. To keep everybody interested and to explain the methods with examples during these, the lectures are spiced with small exercises. These exercises or examples are built in a copy and paste style. So everybody is capable of solving them within a very short time. During the afternoons there are hands on sessions with tutors in the same room. These exercises try to connect the new learned methods to a physics/analysis case, so the students see, why and where this knowledge is needed. A physics exercise based on a toy semi-muonic sample has been coupled to each day’s material to build an analysis, which tries to measure the mass of the top quark. This should motivate them to invest some time in understanding the methods and tools and not to rush over them.

3. CMS Data Analysis School
The knowledge of the complex computing and software environment and complicated analysis tools used to successfully perform physics analysis at CMS, has made ample clear the need for a specialized physics analysis training. It was felt that a training structure should be designed to help CMS physicists from across the collaboration to learn, or to learn more, about CMS analysis and thereby participate in significant ways to the discovery and elucidation of the new physics. The challenge was
to develop innovative classes to allow the students, in some cases with zero experience, to have hands on experience with real data physics measurements that they can make even more precise by searching for new processes that the collaboration hasn't done yet.

3.1. Origin of CMSDAS

The LHC Centre (LPC) [18] at Fermilab has held week-long software and analysis tools workshop since 2006 called JTERM, held twice a year to bring the US physicists up to speed to contribute to physics analysis. This was a very innovative idea when started and the only one of its kind at that time. The workshop composed mostly of lectures and a couple of hour-long tutorial sessions on Grid computing and analysis tools. It lacked the hands-on analysis component mostly because the CMS experiment had not started data taking back then and the analyses were done using the simulated data.

Once LHC started taking data at proton-proton collisions in 2009, it was felt that there was an urgent need to have hands-on training so that physicists can jump-start and contribute to the physics program of CMS. By end of 2010, eight PAT tutorials had already been held and its hands-on style to learn analysis tools proved to be a great success. It was decided to adopt a similar concept to the JTERM and transform it into a hands-on physics analysis-training workshop. Thus was born the idea of CMS Data Analysis School in 2011. This task was by no means easy, as it would need an engagement of a mini-army of physicists unlike before.

3.2. Format and workflow

The CMSDAS is an intense effort that requires months of dedicated preparation on part of the host institute as well as the facilitators (physicists who prepare and teach various physics analysis exercises) who train the participants. The host institute not only has to ensure that it can run several analysis exercise sessions in parallel in several rooms with fast network connections but also has enough computing power to deal with several hundreds of analysis jobs submitted by the participants. The facilitators have to ensure in advance that the data and software used in their analysis exercises is in place and runs smoothly. To come well prepared to the school, with basics of CMS computing and software, the participants are also required to do extensive series of pre-exercises during the weeks preceding the actual school just like before a PAT tutorial but at a more rigorous level.

The international committee for the school discusses with the responsible people at the host institutions the competence of the computing and related infrastructure to conduct the school. For example, typically, multiple machines with 10 Gbit/s connection to the local storage cluster might be needed to meet the peak computing demand during the school. If the school is held at the CMS Tier-1 or Tier-2 computing centres [2], these requirements and expertise are taken for granted. However, in case it is not one of these places one has to ensure that these needs be met first before announcing the school. The times of the school is chosen carefully to avoid a time that either overlaps or closely precedes major physics conferences. In this way maximum availability of the facilitators and participants is ensured.

As part of the preparation, after choosing the site and the responsible people (called the chairs) for the school a main website is prepared containing basic information about the school and its guidelines. This includes software (CMSSW release version) and computing guidelines and naming convention to be followed when facilitators make twikis corresponding to their physics analysis exercises. The twikis made by the facilitators contain nuts and bolts and step-by-step instructions of their analysis exercise. This is followed by announcement of the school via email (Hypernews) to the entire collaboration indicating the Indico website for registration. The next single most important step is to identify the names of physics analysis exercises to be done and the names of the facilitators associated with them. This step takes a few weeks and multiple iterations as commitment is sought from the facilitators.
Next the series of pre-exercises are prepared followed by its announcement such that there is at least a month available to complete these. The answers to the per-exercises are submitted via an espace website similar to the one for the PAT tutorial.

After it is decided which physics analysis exercises will be taught, registered participants are asked to choose the ones they are interested in. This choice has to be made, as all analysis exercises cannot be taken and there is an upper limit to the number taken. The analysis exercises are divided into short and long exercises. The short exercises are two hour each and participants can take up to 6 out a total of 12. The group of short exercises last 2 days. The long exercises are 2-3 days long and a participant can choose only one. The actual number and names of the exercise can vary from one school to another. The school format also includes lectures on topics of interest from luminaries in the field. A typical list of topics for lectures and exercises is shown in Figure 1.

![Figure 1](image)

In the first part of a weeklong program, participants learn from the experts, how to identify physics objects in the CMS data, use statistical and simulation tools, particle flow techniques and deal with pile-up issues in doing physics analysis. In the later half, the participants are divided into several groups with each focusing on a particular physics analysis of their choice. Within this group they work on different aspects of an analysis under the personal guidance of the experts who stimulate their thinking to solve problems by discussion and example. While some apply this knowledge in doing min-analysis looking for Higgs boson in the low and high mass range, studying top quark properties and new physics with jets, others look for SUSY signatures and exotic particles that might produce vertices displaced from the primary collision point. This effort culminates in making a presentation of their findings in a competitive environment to a panel of distinguished physicists who judge them and choose a winner. This accelerated opportunity gives them skills to jump start with their own physics analysis which otherwise would take at least a couple of years of preparation.

3.3. Survey and Feedback
Feedback from the participants is invaluable as it is the only way to directly know what needs to be improved for next tutorial. We show below the results from a survey conducted about the CMSDAS held at LPC, Fermilab in January of 2012. The feedback is shown in Figure 2 and Figure 3. The participants as seen from Figure 2 include senior people besides students. The participant’s institute
indicate that the school, though open to entire collaboration worldwide majorly caters to the local region, North America in this case.

The participation across different analysis is uniform though some physics topics are more interesting than the others for an individual. Some participant did experience that the level of the exercises was too high for them.
The one aspect that students simply love about the school is the collaborative environment where they meet and interact with their counterparts facing similar challenges while dealing with the CMS data. Overall the feedback is very encouraging.

3.4. CMSDAS worldwide

After having 2 CMSDAS at LPC Fermilab with very good experience and feedback there was a very popular demand to hold it outside U.S.A. In 2012, CMSDAS was organised at INFN (National Institute of Nuclear Physics) Pisa in Italy. The noteworthy things are that INFN Pisa is a computing Tier-2 centre and being in Europe majority of participants were from Europe. The next CMSDAS will happen in Taiwan at the end of 2012 and the one after that is being planned in Brazil in 2013.

With 400 users trained at four schools so far and more to follow worldwide, CMSDAS is making a big difference in preparing CMS users in meeting the challenges of analysing the enormous amount of data that it will accumulate and eventually leading to discoveries.

4. LHC Physics Centres

CMS faces a new paradigm where many scientists in CMS are not located at the host lab (CERN). Though there are many tools that are powerful enablers to communicate remotely but face-to-face interaction with colleagues still remain indispensable. In this scenario, the role and efficacy of a remote regional centre can hardly be overemphasized. In this way a significant number of scientists can contribute to the CMS program by being based at regional centres. These centres serve as hub of expertise in physics, simulation and detector development. Three LHC regional centres [4] have come up in last few years, they are accessible to all CMS members while mostly serving their respective regions. These centres serve as main sites for CMS training and tutorials program.

• **LHC Physics Centre (LPC) at Fermilab** [18] – It is the regional centre for the entire US CMS collaboration of 700 members. It is a vibrant community of physics object, algorithms, simulation experts, theory experts, Tier-1 computing centre and Remote Operation Centre for real time connection to CMS. It hosts conferences, analysis schools (CMSDAS), tutorials (PAT) and workshops.

• **LPCC (LHC Centre) at CERN** [19] – It epitomizes initiatives to support LHC physics program. It organizes Workshops, leads in development of event generators, physics tools, and corresponding web pages. It also serves as a portal into LHC physics resources for the whole HEP community, holds EP/PP/LPCC seminar and LPCC students’ lecture.

• **Terascale Center at DESY** (Germany) [20]- Terascale bundles the German activities in high-energy collider physics. It encompasses 18 universities, two Helmholtz Centers and one Max Planck Institute. The alliance works on new accelerator and detector technologies, methods of data analysis and theoretical models and development of the relevant computing infrastructure.

5. Conclusion

The CMS training program is a success model and it is evolving and growing stronger. The PAT tutorials and CMSDAS epitomize these efforts. It has lead to motivated team formed around training and is demonstrative of a healthy collaborative spirit. Holding tutorials regularly has the side outcome to make the CMS documentation (twikis etc.) robust. This is key to the long-term preservation of valuable information that might be needed if one wants to re-analyse the data after a gap of few years. In the tutorials and schools conducted by CMS Physics Support 800-1000 people have been trained. This effort is making an impact in engaging the collaboration to contribute to physics. LHC centres act as training hubs, catalyst for physics learning and mutual exchange of new ideas. The synergy between
the training program and growth of LHC Physics Centres is very beneficial to the entire collaboration in meeting its goals for the future.

6. References

[1] The CMS Collaboration, “The Compact Muon Solenoid Technical Proposal”, CERN/LHCC 94-38, 1994
[2] http://lcg.web.cern.ch/LCG/
[3] Malik S 2011 J. Phys.: Conf. Ser. 331 082006
[4] Malik S and Lassila-Perini K arxiv.org/pdf/1110.0355
[5] Hegner B 2011 J. Phys.: Conf. Ser. 331 032014
[6] https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSDataAnalysisSchool
[7] http://root.cern.ch/root/
[8] https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideAboutPythonConfigFile
[9] Codispoti G et al 2009 IEEE TRANS NUCL SCI , vol. 56, no. 5, pp. 2850-2858
[10] Kovalskyi D 2010 J. Phys.: Conf. Ser. 219 032014
[11] http://public.web.cern.ch/public/
[12] http://hlc.web.cern.ch/lhc/
[13] Lassila-Perini K et al 2011 J. Phys.: Conf. Ser. 331 082010
[14] https://espace.cern.ch/learnecms/
[15] López J et al 2010 J. Phys.: Conf. Ser. 219 082002
[16] EVO stands for Enabling Virtual Organizations. More info on http://evo.caltech.edu
[17] https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookPATTutorial
[18] http://lpc.fnal.gov/
[19] http://lpcc.web.cern.ch/LPCC/
[20] http://www.terascale.de/