CMS Analysis School Model

S Malik[1], I Shipsey[2], R Cavanaugh[3,4], K Bloom[1], Kai-Feng Chan[5], J D'Hondt[6], B Klima[3], M Narain[7], F Palla[8], G Rolandi[9,10], T Schörner-Sadenius[11]

[1] University of Nebraska-Lincoln, Lincoln, NE, U.S.A.
[2] Purdue University, West Lafayette, IN, U.S.A.
[3] Fermilab, PO Box 500, Batavia, IL, U.S.A.
[4] University of Illinois, Chicago, IL, U.S.A.
[5] National Taiwan University, Taipei, Taiwan
[6] Vrije Universiteit Brussel, Brussels, Belgium
[7] Brown University, Providence, RI, U.S.A.
[8] Istituto Nazionale di Fisica Nucleare, Pisa, Italy
[9] Scuola Normale Superiore di Pisa, Pisa, Italy
[10] CERN-PH, 1211 Geneva 23, Switzerland
[11] DESY, Hamburg, Germany

E-mail: malik@fnal.gov

Abstract. To impart hands-on training in physics analysis, CMS experiment initiated the concept of CMS Data Analysis School (CMSDAS). It was born over three years ago at the LPC (LHC Physics Centre), Fermilab and is based on earlier workshops held at the LPC and CLEO Experiment. As CMS transitioned from construction to the data taking mode, the nature of earlier training also evolved to include more of analysis tools, software tutorials and physics analysis. This effort epitomized as CMSDAS has proven to be a key for the new and young physicists to jump start and contribute to the physics goals of CMS by looking for new physics with the collision data. With over 400 physicists trained in six CMSDAS around the globe, CMS is trying to engage the collaboration in its discovery potential and maximize physics output. As a bigger goal, CMS is striving to nurture and increase engagement of the myriad talents, in the development of physics, service, upgrade, education of those new to CMS and the career development of younger members. An extension of the concept to the dedicated software and hardware schools is also planned, keeping in mind the ensuing upgrade phase.

1. Introduction

The Large Hadron Collider (LHC) [1] at CERN (European Centre for Nuclear Research) is the largest and most powerful accelerator in the world that collides beams of protons or lead ions. These collisions recreate the conditions just after the Big Bang and are recorded by a set of four detectors around the collider. The data collected by these detectors is analysed by teams of physicists around the world. Physicists from over two hundred institutions in 5 continents have worked together to build these cathedral-sized state-of-the-art detectors. This global phenomenon has captivated the public

[1] Sudhir Malik, MS 121, Fermilab, PO Box 500, Batavia, IL 60510, U.S.A.

interest worldwide and our global collaborations are seen as a model for the world. It provides a wonderful and unique opportunity for the entire field of particle physics to explore and push the boundaries of technology in terms of hardware, computing and software. One of these detectors is called CMS (Compact Muon Solenoid) [2]. The CMS collaboration consists of 3500 members and is considered a star performer by the wider particle physics community in terms of producing new physics results, maintaining an excellent detector and a worldwide Grid computing network. However, in order to maximize the discovery potential and physics output it is very important to engage the entire collaboration in the physics discovery process. Today, many CMS colleagues are engaged and many more need to do so. CMS is striving to nurture and increase engagement of the myriad talents of CMS, in the development of physics, service, detector upgrade, education of those new to CMS and the career development of younger members.

2. Investing in the future

CMS experiment will likely remain a major productive scientific collaboration for at least two decades. The collaborating physicists will spend a majority of their careers as colleagues. Given this longevity, it is especially important to continue to develop a nurturing and sustainable collaboration where all can reach their full potential. It is critical that CMS provide effective collaboration by offering opportunities to aid the development of young member physicists in all aspects of experimental physics like training in data analysis, software and hardware, and provide career development paths. The large size of the CMS community, its geographical spread, the distributed computing resources spread over multiple time zones adds to the challenge of already complex analysis tools and software needed to successfully perform physics analysis. In the new paradigm where many scientists are not located at the host lab (CERN), the face-to-face interaction still remains indispensible. To tackle this challenging scenario, CMS had instituted an extensive program of tutorials and training [3]. A prime example of collaboration scale education is the CMS Data Analysis School (CMSDAS) [3].

3. Genesis of CMSDAS

The idea for CMSDAS came from two previous workshops, CLEO101 which was a hands on total immersion software and analysis tools workshop developed by the CLEO-c Collaboration in the early 2000s and JTERM developed at the LHC Centre (LPC) at Fermilab for the CMS experiment. These workshops were innovative. JTERM was composed of 90% lectures and a couple of hour-long tutorial sessions on Grid computing and analysis tools. However, it did not have the hands-on physics analysis component mainly because the CMS experiment had not started taking data back then and the analysis was done with the simulated data. Once LHC started taking proton-proton collisions data in 2009, it was felt that there was an urgent need to have hands-on training with physics analysis in order for users to jump-start and contribute to the physics program of CMS. Therefore, it was decided to adopt a concept similar to the JTERM and transform it into a hands-on physics analysis-training workshop. Thus was born the idea of CMSDAS. 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. The very first school was held in 2010 at LPC, Fermilab. A list of these schools can be found at [4].

4. Preparation for the school

Preparation for the school begins months in advance. It involves coordination between the host institute and the international advisory committee for the school. The time of the school is chosen carefully to avoid overlap with or closely preceding major physics conferences. The host institute must ensure a computing infrastructure be able to deal with hundreds of analysis jobs running in parallel and simultaneous availability of multiple conference rooms. An announcement to register for the school follows and the names of physics analysis exercises and the teachers (called facilitators) are decided. A website (internal to CMS collaboration) is prepared containing basic instructions and guidelines. The training material for computing and software tools, and physics analysis is prepared by the facilitators and is uploaded in the form of twikis[5] which are password protected. It contains nuts and bolts and step-by-step instructions on exercises. Facilitators have to ensure in advance that the data and software used in their analysis exercises is well tested and runs smoothly at the host institute. There have so far been more than 110 CMSDAS facilitators. Of these 40% are faculty, 40% are post docs, and 20% advanced graduate students. They are all eager volunteers. In order to arrive well prepared at the school with basics of CMS computing and software, the participants are required to do an extensive series of preparatory exercises (called Pre-exercises) preceding the school. A few weeks before the school, registered participants are asked to choose their first, second and third preference of a physics analysis exercise (called "Long Exercise") that are they are interested in. However, the format of the school is such that each student can only do one Long physics analysis exercise. Every effort is made to assign the first choice, keeping in mind a uniform distribution of students across all Long Exercises. Depending on the Long Exercise, students are assigned a set of up to six physics object exercises called Short Exercises. For example, for a choice of "Top quark cross section" Long Exercise, the Short Exercise essential to successfully complete this are "b-tagging", "Muons", "Particle flow" and "Roostat"

5. Format and workflow

The first half of the week-long program consists of half-day of lectures on HEP and CMS by experts, followed by two days of Short Exercises on topics like using statistical and simulation tools, particle flow techniques and deal with pile-up issues in doing physics analysis. In the second half of the week, the participants are divided into several groups with each group focusing on a particular Long Exercise. Within a group each student works on different aspects of an analysis under the personal guidance of experts. While some apply this knowledge in doing analysis looking for Higgs in the low and high mass range, Higgs properties, studying top quark properties and new physics with jets, others look for signatures of Super Symmetry in the data 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 at a mini-symposium on the last day.

5.1. Pre-exercises

A series of Pre-exercises prepared and tested by a team of facilitators are announced a month before the commencement of the school. The student's submit answers to the questions in the Pre-exercises via an online secure form in espace [6] (SharePoint Technology). The students are strongly recommended to complete these before their arrival to the school. The exercise completion rate is around 70%. The Pre-exercises cover basics of CMS Software, computer account access at the host institute, access to CMS code and collision data, running grid jobs, basics of physics analysis tools [7] specific to CMS such as PAT, python, ROOT and curve fitting to the data distributions.

5.2. Lectures

The school opens with a registration session and half the first day is devoted to a series of lectures on topics of interest by luminaries in the field of HEP theory, experiment and accelerator physics. This covers a full spectrum of CMS relevant topic like the CMSDAS introduction, status of HEP research, LHC status and plans, summary of CMS physics results and future analyses, HEP analysis design, CMS detector and CMS software and tools.

5.3. Short Exercises

The lectures are followed typically by 6 periods of Short Exercises on several different topics. These topics include Roostats, Generators, Tracking & Vertexing, Electrons, Muons, Photons, Jets, b-tagging, particle flow, pile-up, event visualisation and Madgraph. These topics are essential to any kind of physics analysis that one may want to do at CMS. Each exercise session is 2 hour long. The class size for each exercise is typically 6-8 students taught by 2-3 facilitators. Each exercises begins with a brief presentation introducing the goals and overview of the material to be covered and tasks to be performed. While students work individually in a class like setting, facilitators are present to help and explain. A student can take only up to 4 Short Exercises.

5.4. Long Exercises
In the later half of the school each of the students performs one Long Exercise and works in teams of 5-10 with 3-4 facilitators. A full physics measurement is done in 2.5 days that could be a search of new physics or a particle or determination of cross section, and these typically include the determination of trigger and object efficiencies and acceptance. A fit to the data distributions is performed to extract the signal and the background yields. Methods that involve data driven estimation of the background determination are used and a range of systematic uncertainties are measured. The measurement frequently extends the state of the art on CMS either by adding or by modifying selection to improve the sensitivity, or it could be a new search, for example, 3 photon final state, not yet performed by CMS. The students decide on a division of labour to accomplish the tasks leading to a complete analysis.

5.5. Mini symposium
The last day of the school is dedicated to a competition in a mini-symposium setting. Each group (or team) that has performed a Long Exercise presents and defend their physics results. In the morning each team is given exactly 90 minutes to prepare slides summarizing physics motivation and measurement. Talks are required to be uploaded to the agenda by a specific time. Each team presents their results before a panel of physicists who serve as judges for the competition. The format of the symposium allows exactly 15 minutes for each presentation followed by exactly 5 minutes of questions from the judges. After all presentations are over, judges deliberate and select a winner. This experience gives them skills to jump start with their own physics analysis which otherwise would take at least a couple of years of preparation. It also trains them to make succinct presentations, preparing them for future conferences.

6. Survey and Feedback
Few weeks after the school, a feedback survey is sent to the students that they must respond to in a month. The feedback from the participants is invaluable as it is the only way to directly know what needs to be improved for the next one. The one aspect that students appreciate about the school is the collaborative environment where they meet and interact with their counter parts facing similar challenges while dealing with the CMS data. Overall the feedback is very encouraging. We show

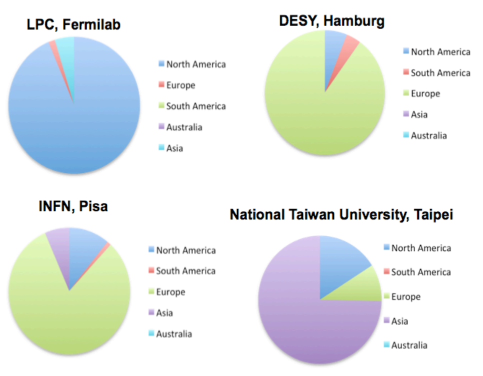

Figure 1. Most students are from the host region

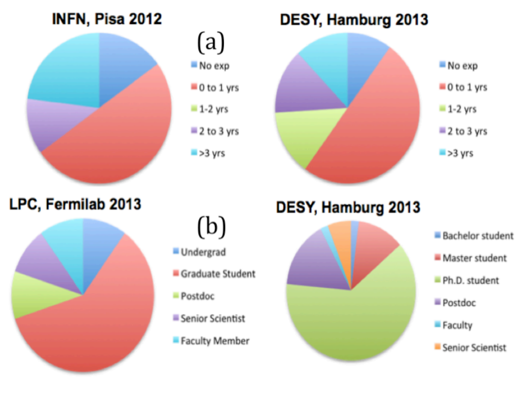

Figure 2 (a) CMS software experience of students
(b) Majority students are graduate level or postdocs

Response to Quality (%)

	Right	High	Easy	NR
DESY	68	8	14	10
LPC	71	14	10	5
INFN	92	4	0	4

Figure 3. Response to quality of Short Exercises (NR means No Response)

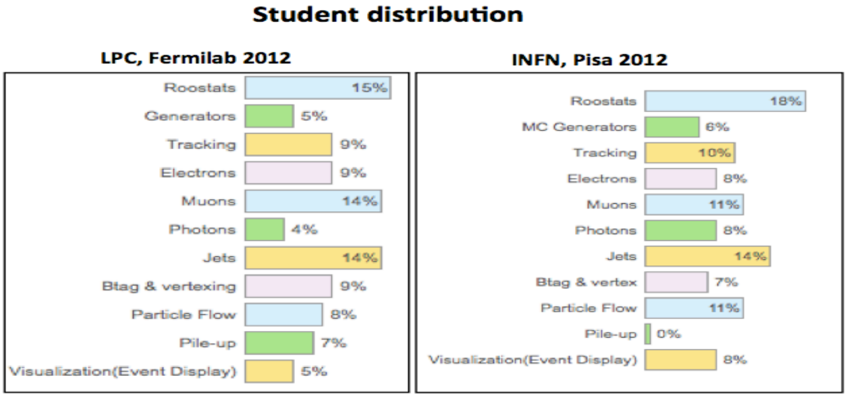

Figure 4. Distribution of students among different Short Exercises, in 2012

Response to Quality (%)

	Right	High	Easy	NR
DESY	78	6	6	10
LPC	90	5	5	0
INFN	81	4	4	11

Figure 5. Response to quality of Long Exercises (NR means No Response)

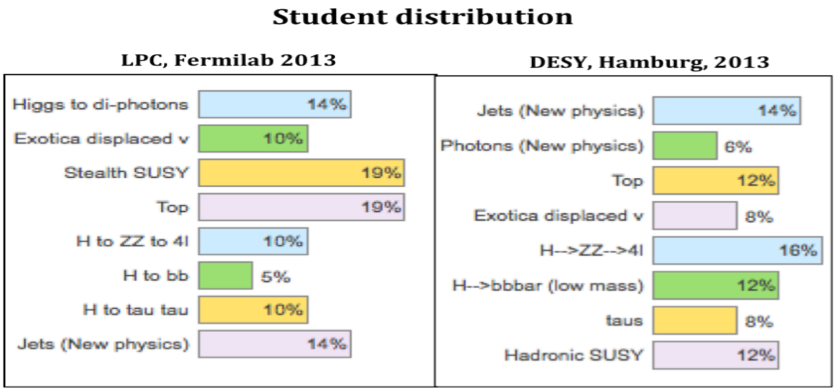

Figure 6. Distribution of students among different Long Exercises, in 2013

7. Conclusion

The CMSDAS training program is a successful model. It is growing stronger and evolving. With 400 users trained so far at seven schools and more to follow worldwide, CMSDAS is making a big difference to the CMS collaboration to prepare continuously in meeting the challenges of analysing the enormous data that it will accumulate in years to come and eventually lead to discoveries.

8. References

[1] *http://home.web.cern.ch/about/accelerators/large-hadron-collider*
[2] The CMS Collaboration, "*The Compact Muon Solenoid Technical Proposal*", CERN/LHCC 94-38, 1994
[3] Malik S et al 2012 *J. Phys.: Conf. Ser.* **396** 062013
[4] *http://lpc.fnal.gov/programs/schools-workshops/index.shtml*
[5] *https://twiki.cern.ch/twiki/bin/view/CMSPublic/CMSDataAnalysisSchool*
[6] Amran N et al 2010 J. Phys.: Conf. Ser. **219** 082008
[7] Malik S et al 2011 *J. Phys.: Conf. Ser.* **331** 082006