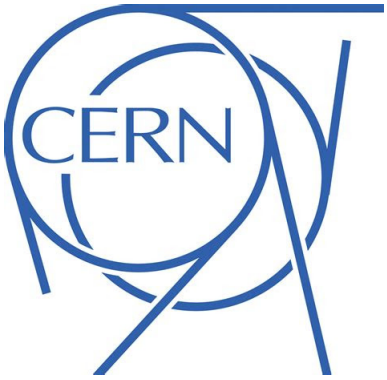


CERN Summer Student Project: CMS Open Data

Roxane Theriault

Supervisor: Kati Lassila-Perini

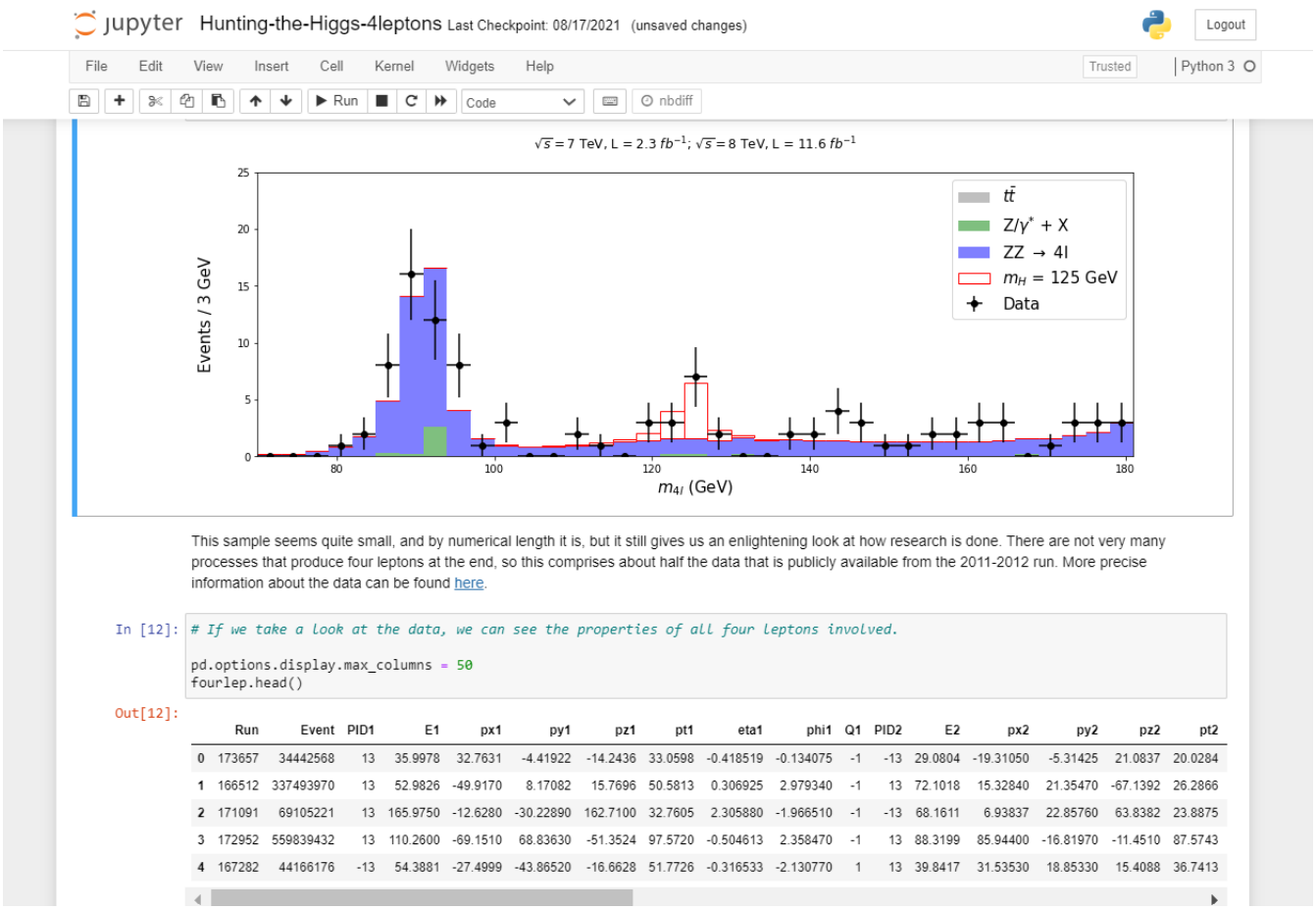
September 24, 2021

I worked with the Data Preservation and Open Access (DPOA) group under the supervision of Kati Lassila-Perini. My project can be broken down into two parts: website translation and running a workshop.

1 Website Translation

Part of my project this summer was to translate the Open Data in Education website [1], which was developed as part of the Helsinki Institute of Physics' project Education and Open Data [2]. The website provides exercises for high school students to become familiar with programming and data processing, as well as tools for teachers to help them create their own exercises.

The website was originally in Finnish and had already been translated to Swedish. Since I speak neither of these languages, I used Google Translate in order to translate the site into English. Of



	Run	Event	PID1	E1	px1	py1	pz1	pt1	eta1	phi1	Q1	PID2	E2	px2	py2	pz2	pt2
0	173657	34442568	13	35.9978	32.7631	-4.41922	-14.2436	33.0598	-0.418519	-0.134075	-1	-13	29.0804	-19.31050	-5.31425	21.0837	20.0284
1	166512	337493970	13	52.9826	-49.9170	8.17082	15.7696	50.5813	0.306925	2.979340	-1	13	72.1018	15.32840	21.35470	-67.1392	26.2866
2	171091	69105221	13	165.9750	-12.6280	-30.22890	162.7100	32.7605	2.305880	-1.966510	-1	-13	68.1611	6.93837	22.85760	63.8382	23.8875
3	172952	559839432	13	110.2600	-69.1510	68.83630	-51.3524	97.5720	-0.504613	2.358470	-1	13	88.3199	85.94400	-16.81970	-11.4510	87.5743
4	167282	44166176	-13	54.3881	-27.4999	-43.86520	-16.6628	51.7726	-0.316533	-2.130770	1	13	39.8417	31.53530	18.85330	15.4088	36.7413

Figure 1: A glimpse of one of the advanced exercises, Hunting the Higgs. It can be found here.

course, Google Translate does not always work that well, so in order to verify the translations, I would translate both the Finnish and Swedish versions into English. By comparing them, I was able to figure out what exactly was being said and write it out clearly in English.

Furthermore, the website is built from a GitHub repository [3], so part of my work was also to upload and organize the English files. I went over the English exercises and corrected a few things. Most of them are about particle physics, and I find they have a nice range of difficulties. One of the simpler ones involves importing a file and plotting histograms from the data, while one of the more advanced ones guides students through manipulations to "find" the Higgs boson (see Fig. 1). There are also a couple of examples about text analysis, which I found very interesting because I did not know Python had packages for this. One of them uses the library textblob to train a Naive Bayes Classifier to identify positive and negative sentences. Finally, I built a Jupyter Book to present the Physics open

data exercises in an organized way [4].

2 Running a Workshop

Another part of my project this summer was to organize a workshop about CMS Open Data and the uses of open data in education aimed at summer students. I worked on this with Veera Juntunen, a summer trainee at the Helsinki Institute of Physics and CERN.

Since the workshop was online and lasted two hours, we wanted a way to connect with the audience and keep them engaged. To that end, we decided to use Zoom Polls and a Kahoot quiz. We gave the first poll towards the beginning of the workshop, and its purpose was to get to know our audience a little better. Some of the questions were: "Are you a: Student/Researcher/Teacher/Other", "Have you heard about open data initiatives before?", "Do you know Python?", and "Have you used Jupyter Notebooks before?".

After the poll, we talked about what open data is and what its uses are. We also gave some links to places where one can find open data. Then we gave some background information on the Compact Muon Solenoid (CMS), which led into talking about the CMS Open Data project. Finally, we discussed making material for high school education using open data. This involved presenting the website I translated, as well as some GitHub repositories with exercises. We also briefly explained Jupyter Notebooks (files that can have live code, text, visualizations, etc.) and Binder (a website where you can run Jupyter notebooks in your browser).

Then we had our Kahoot quiz. Kahoot is a website where you can make fun quizzes and people can join with a code, so you can run the quiz live and see the results in real-time. We did this because we thought it would be a nice break from us just talking and to promote audience involvement. The questions were lighthearted and not difficult. They pertained to what we had talked about during the first part of the workshop, so anyone who had been listening knew the answer (it was multiple choice). We had very good participation in this section, with approximately 27 out of the 30 participants answering.

The next part of the workshop consisted of example notebooks, which I went through while explaining what I was doing. One of the notebooks was a short introduction to using Jupyter Notebooks and showed the different types of cells, how to run them, and other things in that vein. Because of the

poll we did at the beginning, we knew that only a couple of people in the audience had never or hardly ever used Jupyter Notebooks, so we went over that example quickly. The other example was one of the exercises using CMS open data in which one reproduces the calculations that led to the discovery of the Higgs boson, albeit in a much more simplistic manner. After this, we gave a brief overview of the different exercises we provided for the practice session. At the end of this section, we gave another poll, this one asking questions like, "Do you know how to get started?" and "Do you have an idea which exercise you're going to start with?". This was to let us know if anyone needed any help, but were too shy to say something out loud or in the chat.

During the hands-on section, Veera and I stayed available to answer questions while everyone was working on the exercises they chose. We had a few exercises from the Education and Open Data project, ranging in difficulty from beginner to advanced. We also had a couple of ADL (Analysis Description Language) exercises, proposed to us by Sezen Sekmen [5]. ADL is a language created at CERN and focuses on event processing operations. It is meant to be simple and intuitive to use. Sezen, who was present at our workshop, gave a brief two-minute introduction to the language and the related exercises. The hands-on section went very well. We did not have many questions, but from the comments we received people seemed to be doing the exercises. The exercise that emerged as the favorite is Creating Sound from Data, in which you import CMS data and convert the invariant mass distribution into a noise.

At the end of the workshop, we sent out a Google Form for the participants to give us feedback. A couple of questions reiterated things that were asked in the introduction poll so we could have a record of the results. Out of the 20 people who stayed until the end of the workshop, 10 filled out the feedback form, all of whom were students. We also asked if they were already familiar with open data and if they think it is relevant to them. Their responses are shown in Fig. 2. The general conclusion is that most of them were familiar with open data and thought it relevant to them, with 80% of respondents answering "Yes, for work" or "Yes, for outreach or education".

Exactly half of the respondents had participated in some sort of outreach related to open data before. We also asked them if they used or intended to use open data for the following purposes: education, outreach, research, personal interests, or not at all (see Fig. 3).

Finally the last three questions on the survey were short-answer questions. The first one was, "What do you think of Jupyter Notebooks as a tool?". The four responses were all very positive, using

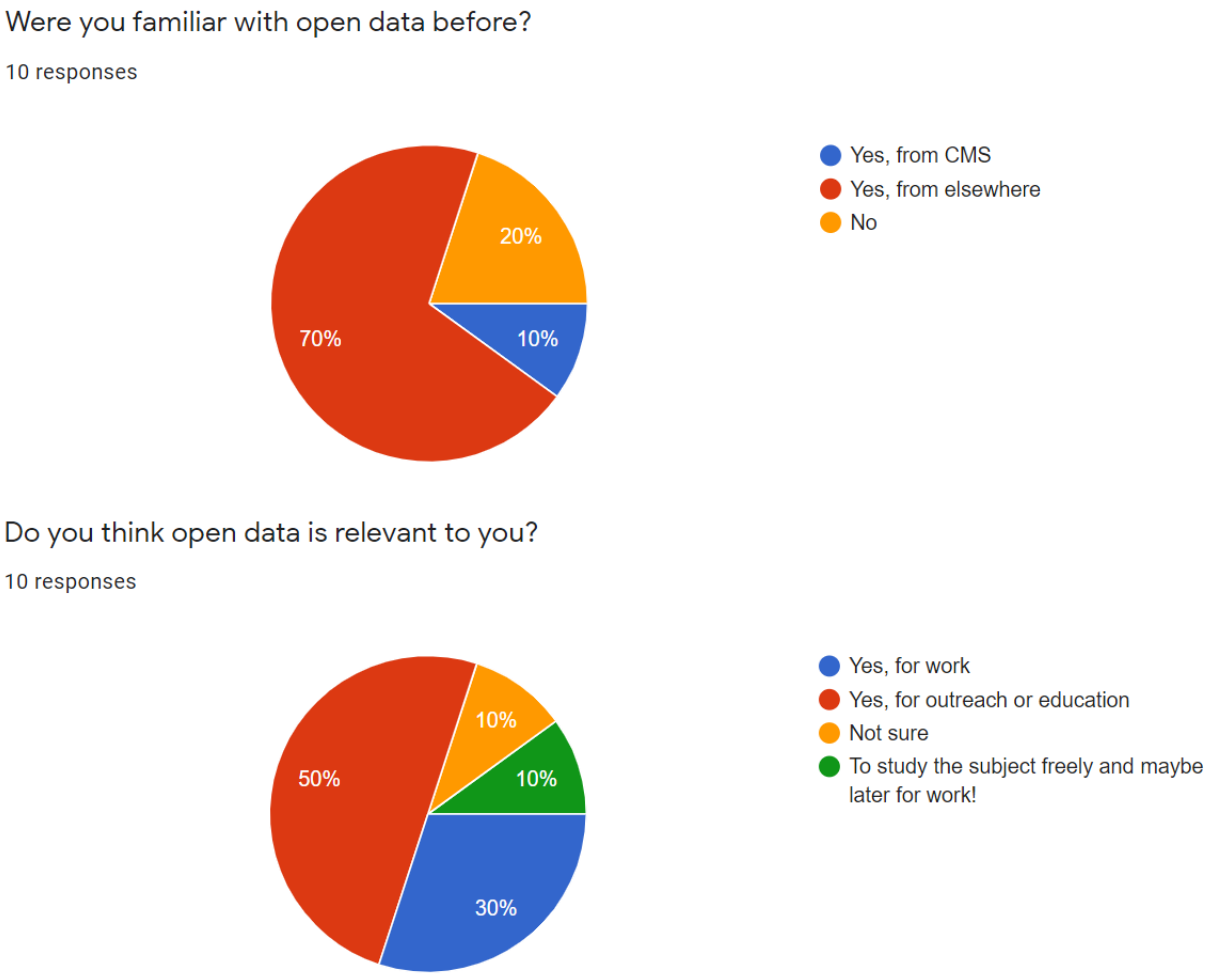


Figure 2: Graphical representations of the answers we received for two of the questions on our feedback form.

words like "useful" and "powerful". The next question asked, "What did you think of the format of this workshop? Were there any particular challenges due to it being online?". Again, most responses were positive, saying that the format worked well. One of the respondents thought that the duration was short for an online workshop. It was a 2 hour workshop, but I imagine that in the future it could be extended by adding more examples and exercises. Someone else said, "It is more difficult to ask questions but there were no questions so all good [sic]". It is true that there were very few questions and those we did get were asked in the chat. We also paused frequently to ask if there were any questions. I recognize that it is harder to ask questions online, which is why we made the polls. It is possible that we were asked so little questions because of the online format. In that case, more thought would be

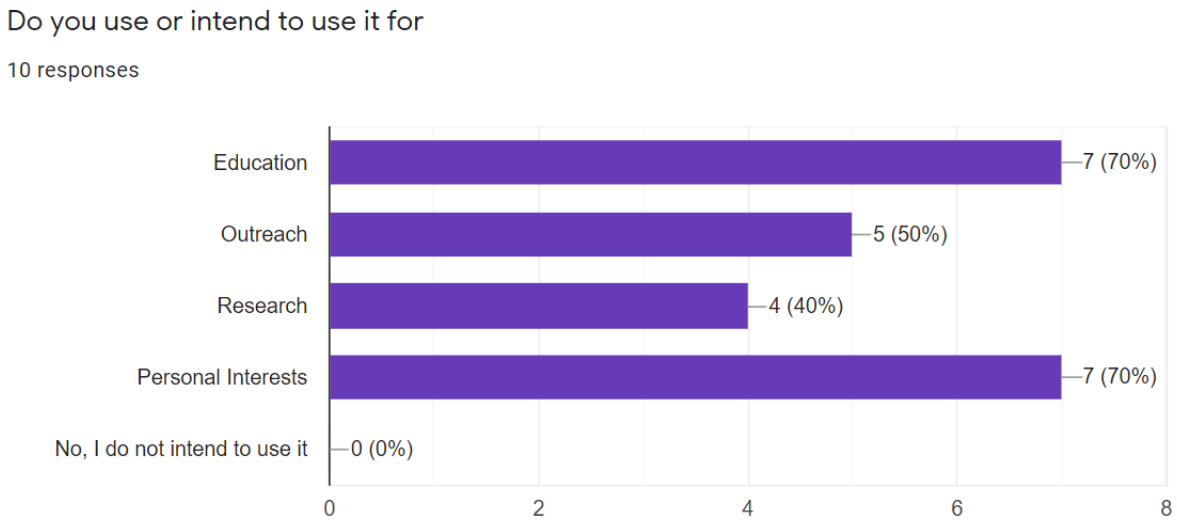


Figure 3: Chart showing the distribution of answers for the question"Do you use or intend to use open data for...?" Respondents could choose more than one answer.

needed on how to make asking questions easier in this format. The last question was simply, "Any other feedback or suggestions?" One person mentioned they enjoyed our attempts at engagement, especially the Kahoot. Someone else suggested that we should have a poll at the beginning of the example session to ask people if they want a demonstration on how to use Jupyter so it can be skipped if people already know how to use it. That was in fact a question in the poll we gave at the beginning of the workshop, and a couple of people had answered that they were not familiar with Jupyter. I went quickly over the Jupyter basics example because of that, but I could maybe have been clearer about my reasons for doing so.

Overall, the workshop seems to have been quite a success. The majority of our feedback, both during the workshop and from the survey, is positive. I learned a lot from running it, since I had never run one before.

Overall, I learned a lot about open data initiatives and how open data can be used for educational purposes during this project. It could certainly have been more challenging, but I think the work this group is doing is very meaningful, so I am glad to have contributed.

References

[1] HIP Education and Open Data Team. Open data in education. https://opendata-education.github.io/en/index, 2021.

[2] HIP Education and Open Data Team. Education and open data. https://www.hip.fi/research/education-and-open-data/, 2021.

[3] HIP Education and Open Data Team. opendata-education Github repository. https://github.com/opendata-education, 2021.

[4] HIP Education and Open Data Team. Physics exercises. https://opendata-education.github.io/en_Physics/intro.html, 2021.

[5] Sezen Sekmen. Adl: An analysis description language for the LHC. https://twiki.cern.ch/twiki/bin/view/LHCPhysics/ADL, 2021.