Submitted to the Proceedings of the US Community Study
on the Future of Particle Physics (Snowmass 2021)

Data Science and Machine Learning in Education

COMPF3 (Machine Learning)

Gabriele Benelli [1], Thomas Y. Chen [2], Javier Duarte [3], Matthew Feickert [4], Matthew Graham [5], Lindsey Gray [6], Dan Hackett [7], Phil Harris [7], Shih-Chieh Hsu [8], Gregor Kasieczka [9], Elham E. Khoda [8], Matthias Komm [10], Mia Liu [11], Mark S. Neubauer [4], Scarlet Norberg [12], Alexx Perloff [13], Marcel Rieger [10], Claire Savard [13], Kazuhiro Terao [14], Savannah Thais [15], Avik Roy [4], Jean-Roch Vlimant [5] Grigorios Chachamis [16]

1 Brown University, Providence, RI 02912, USA **2** Columbia University, New York, NY 10027, USA **3** University of California San Diego, La Jolla, CA 92093, USA **4** University of Illinois at Urbana-Champaign, Urbana IL 61801, USA **5** California Institute of Technology, Pasadena, California 91125, USA **6** Fermi National Accelerator Laboratory, Batavia, IL 60510, USA **7** Massachusetts Institute of Technology, Cambridge, MA 02139, USA **8** University of Washington, Seattle, WA 98195, USA **9** Universität Hamburg, Institut für Experimentalphysik, 22761 Hamburg, Germany **10** European Organization for Nuclear Research, Geneva, Switzerland **11** Purdue University, West Lafayette, IN 47907, USA **12** University of Puerto Rico Mayagüez, Mayagüez, Puerto Rico **13** University of Colorado Boulder, Boulder, CO 80309, USA **14** SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA **15** Princeton University, Princeton, NJ 08544, USA **16** Laboratório de Instrumentação e Física Experimental de Partículas (LIP), Lisboa, Portugal.

ABSTRACT

The growing role of data science (DS) and machine learning (ML) in high-energy physics (HEP) is well established and pertinent given the complex detectors, large data, sets and sophisticated analyses at the heart of HEP research. Moreover, exploiting symmetries inherent in physics data have inspired physics-informed ML as a vibrant sub-field of computer science research. HEP researchers benefit greatly from materials widely available materials for use in education, training and workforce development. They are also contributing to these materials and providing software to DS/ML-related fields. Increasingly, physics departments are offering courses at the intersection of DS, ML and physics, often using curricula developed by HEP researchers and involving open software and data used in HEP. In this white paper, we explore synergies between HEP research and DS/ML education, discuss opportunities and challenges at this intersection, and propose community activities that will be mutually beneficial.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

1 Introduction

The particle physics research community has a strong background and involvement in educational activities. Not only do many of its practitioners come from universities and centers for education, but the community also provides training and educational resources to facilitate our science and convey its importance to members of the public and policy makers.

Particle physics holds a prominent role within academic curriculum at institutions of learning. There are compelling reasons for this prominent role, such as the fundamental nature of our science, fascinating historical development of our field, theoretical research that applies (and often develops) advanced mathematics, powerful applications such as cancer treatment, and high-visibility spin-off technologies such as the World Wide Web.

Data science and machine learning have an increasingly prominent role in our science, as is evident from any recent particle physics conference and this Snowmass process. In recent years, machine learning techniques for detector and accelerator control [1], data simulation [2], parton distribution functions [3], reconstruction [4–6], anomaly detection [7, 8] and data analysis are increasing being applied to particle physics research. Recent reviews of these techniques applied to particle physics research can be found in [9–18]. A "living review" aiming to provide a comprehensive list of citations for those in the particle physics community developing and applying machine learning approaches to experimental, phenomenological, or theoretical analyses can be found in [19].

There is no consensus on the precise definitions of data science and machine learning. For our purposes, we consider *data science* to refer to scientific approaches, processes, algorithms and systems used to extract meaning and insights from data [20] and *machine learning* to refer to techniques used by data scientists that allow computers to learn from data. Machine learning is a subset of the field of artificial intelligence which aims to develop systems that can make decisions typically requiring human-level expertise, possessing the qualities of intentionality, intelligence and adaptability [21]. Figure 1 illustrates these relationships.

Particle physicists increasingly collaborate with computer scientists and industry partners to develop "physics-driven" or "physics-inspired" machine learning architectures and methods. However, the particle physics community in the U.S. has been generally slow to adopt data science and machine learning as formal components in educational curriculum. This situation is rapidly improving. The potential synergies between education and particle physics research in the areas of data science and machine learning motivates this study.

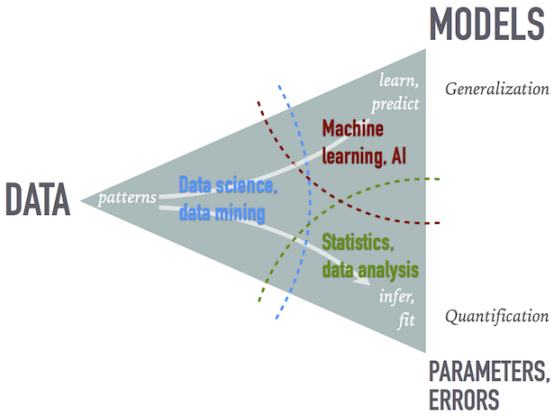

Figure 1: Simplified illustration of data science and machine learning in the context of data, models and statistical inference.

In this Snowmass white paper, we explore some of the challenges and opportunities for data science and machine learning in education and suggest future directions that could benefit the particle physics community.

2 Educational Pathways

There are many ways that particle physics researchers at all levels provide educational opportunities to students and other trainees. Mentoring in research activities is a key educational delivery method that provides an opportunity for professional development of early career individuals and bi-directional learning inherent in scientific research.

In addition to traditional advising of undergraduate and graduate in thesis research, there are dedicated programs such as the NSF Research Experience for Undergraduates [22], masterclasses (E.g. [23]), and capstone project-based courses (E.g. [24]). These programs are multidisciplinary and often involve direct involvement with industry which create strong opportunities for students to pursue careers within or beyond academia.

Particle physicists also develop curriculum at the intersection of data science, machine learning and physics for use in undergraduate and graduate-level courses in their home department(s). These curriculum might represent whole courses or specific teaching modules which could be used in multiple courses or other forms of educational delivery. A few examples are provided in Sec. 3. These materials are often drawn from and feed into educational and training materials from particle physics research, such as schools (e.g. [25–27]), training events (e.g. [28, 29]), workshops (e.g [30–34]) and bootcamps (e.g. [35, 36]).

3 Examples of Curriculum from HEP researchers

Particle physicists have been active in generating course curriculum and in many cases entire courses around the intersection of physics, data science and machine learning. Sec. 4.1 and Sec. 4.2 describe some of the motivating factors that compel particle physicists to devote time to educational aspects of data science and machine learning in physics.

In this section, we provide a few specific examples of courses developed by members of our community. This is a small sampling of courses and development efforts in data science and machine learning for education, included by the authors as a point of reference. There are many such courses developed by physicists outside of HEP for their respective physics departments (e.g. [37, 38]) from which we could learn a great deal in the spirit of interdisciplinary collaboration. In Sec. 6, we propose a more coordinated effort of discovery of such courses and the sharing of tools and experiences for those that have taught and/or developed such courses. Recently, several textbooks have also been developed specifically for teaching machine learning to (particle) physicists [39–41].

Particle Physics and Machine Learning (UCSD) [42]: This is a course on particle physics developed by Javier Duarte and Frank Würthwein as part of the UCSD Data Science Capstone sequence [24]. The course centers around applying modern machine learning techniques to particle physics data. The intended audience consists of fourth-year data science majors, who select among a set of topics for their capstone projects. The course is split into two quarters. The bulk of the first quarter focuses on the task of identifying Higgs boson decaying to bottom quarks. Specifically, the students are tasked with reproducing results in a recent ML and particle physics publication [43] on graph neural networks for the identification of boosted Higgs bosons decaying to bottom quarks. They are also provided with weekly lectures on data science topics (by the lecturer) as well as particle physics

topics (by the domain mentors). During the second quarter, students propose and execute a new project that extends the work of the previous quarter. Possible projects include studying the performance of different message passing and graph neural network structures, studying mass decorrelation strategies, applying explainable AI techniques (like layerwise relevance propagation) to the Higgs tagging task, comparing multiclassification to binary classification, and developing a network for Higgs boson jet mass regression.

Physics and Data Science (MIT) [44]: This is a course developed by Phil Harris and aims to present modern computational methods by providing realistic examples of how these computational methods apply to physics research. Topics include: Poisson statistics, error propagation, fitting, data analysis statistical measures, hypothesis testing, semi-parametric fitting, deep learning, Monte Carlo simulation techniques, Markov-chain Monte Carlo, and numerical differential equations. The class format is a mixture of lectures by course faculty (and guest speakers to highlight the relevance of this work towards department research), recitations run by undergraduate and graduate students, and completion of three projects with a final presentation based on an extension of any one of the three projects. The projects include data analysis in gravitational waves, cosmic microwave background, and LHC jet physics using open data.

Introduction to Machine Learning (Princeton) [45]: This is a course developed by Savannah Thais and is primarily intended for non-computer science students who want to understand the foundations of building and testing an ML pipeline, different model types, important considerations in data and model design, and the role ML plays in research and society. Topics covered in lectures and exercises include conceptual foundations of ML, artificial neural networks, convolutional and recurrent neural networks, unsupervised learning, generative models and topics in AI ethics such as data bias, algorithmic auditing, predictive policing, inequitable utilization of algorithms, proposed regulation.

Data Analysis and Machine Learning Applications for Physicists (Illinois) [46]: This is a course developed by Mark Neubauer which aims to teach the fundamentals of analyzing and interpreting scientific data and applying modern machine learning tools and techniques to problems commonly encounters in physics research. The class format is a combination of lectures, homework problems that elaborate on topics from the lectures that give students hand-on experience with data, and a final project that students can choose from in the areas of particle physics and astrophysics. Topics covered include handling, visualizing and finding structure in data, adapting linear methods to nonlinear problems, density estimation, Bayesian statistics, Markov-chain Monte Carlo, variational inference, probabilistic programming, Bayesian model selection, artificial neural networks and deep learning.

3.1 Tools and Techniques

Not surprisingly, the course examples just described utilize different tools and employ varied approaches to course delivery given independent and tailored development at their respective universities. In spite of this, there is a significant overlap in the tools and techniques used in these courses, which are briefly described in this section.

Software These courses make extensive use of open source software, especially Python as a core language. Python is a very popular language for education given it is a high-level

language with automatic memory management, simplicity of its syntax and readability of its code as compared with most other languages. Python is a binary platform-independent language such that it can be run on virtually any hardware platform and operating system. This aspect is important in an educational setting where students use a variety computing systems. Python is an open source project that is free to use, with an extensive ecosystem of code libraries for applications in science, engineering, data science and machine learning. The example courses described make extensive use of libraries used in scientific computing, mathematics, and statistics such as `NumPy` [47], `Pandas` [48, 49], `matplotlib` [50] and `seaborn` [51]. `SciPy` [52] provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and many other classes of problems. In terms of machine learning, these courses use the general purpose `scikit-learn` library [53], as well as deep learning frameworks such as `PyTorch` [54] and `TensorFlow` [55].

Python as a language for data science and machine learning has broad community support. Therefore, a key benefit of using Python in the classroom in terms of professional and skills development is that it is a language used extensively in real-world applications of data science and machine learning and widely used in industry. Of course, this is only the present landscape and it is anyone's guess as how it will change on the 5–10 year timescale.

Data The specific data used in the example courses described in this paper vary according to the exact lessons and projects being taught. However, the courses generally made use of open scientific data sources when appropriate. This is especially relevant for the project components of these courses. For example, open data resources at the UCI Machine Learning Repository [56], Galaxy Zoo challenge data [57], and CERN Open Data Portal [58] were used for HEP and astrophysics students projects in the course at Illinois.

Tools The use of Jupyter notebooks was a common aspect of the example course described in this section. Jupyter notebooks provide an interactive front-end to a rich ecosystem of python packages that support machine learning and data science tasks. They provide a means for students and instructors to create and share documents that integrate code, LaTeX-style equations, computational output, data visualizations, multimedia, and explanatory text formatted in markdown into a single document. When hosted by a cloud-based server resource such as JupyterLab, using these notebooks has huge benefits for teaching, including removing the need to install any software locally or require any specific machine to be used by students [59].

Course Materials The reference materials used in the courses were education and training materials that are widely available in the public domain and enhanced by a significant amount of supplementary resources linked on the course pages. In the Illinois course, all materials are managed through a dedicated Github Organization. The students and course staff are all members of this organization, with different access levels to material (repositories) according to their role. Students each create a private repository which is how they submit their homework and final projects for grading.

Infrastructure As with data used in the example courses, the infrastructure utilized for course delivery varied according to the specific needs of the courses and institutional arrangements. In general, the courses used open source software in the Python language to

implement scientific codes, and commonly used machine learning frameworks and libraries within Jupyter notebooks. The Python code that the students developed could be executed in a number of ways within these courses. For example, a common course environment could be generated by package management software (e.g. Anaconda [60]) or a Linux container service (e.g. Docker [61]). Another approach was to use an execution environment such as Google Colab [62] which allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs [62]. In the Illinois course example, a custom Docker container [63] maintained by the course staff is launched onto commercial cloud resources which used to serve notebooks for the students using JupyterLab and provide computational resources to execute the code.

Delivery The primary methods of content delivery and active student engagement varied by course but generally involved a mixture of lectures by course faculty that included physics and data science pedagogy demonstrated through in-class live examples in Jupyter notebooks, recitation/discussion style activities involving hands-on interactive exercises, and projects. In the MIT course, guest speakers were invited to highlight the relevance of the pedagogy with ongoing research in the department. In the UCSD course, the students were actively involved in proposing and executing a new project that extends the work of the previous quarter.

4 Opportunities

Physics departments are increasingly offering curriculum to their undergraduate and graduate students at the intersection of physics, data science and machine learning. Particle physicists are increasingly interested in developing new courses at this intersection. For those so inclined, these courses provide opportunities for particle physicists to (1) describe synergies between modern machine learning research and particle physics research, (2) make connections with colleagues from other departments, (3) make connections within their own department in other research domains, (4) recruit students interested research at the intersection of machine and particle physics, and (5) learn the tools and techniques from data science and machine learning that can be applied to particle physics research.

There are opportunities to take advantage of programs in education from federal agencies and engage with key organizations, such as the American Physical Society's (APS) Topical Group on Data Science (GDS) [64]. The APS GDS is focused on promoting research at the growing interface between physics and data science, spanning big data, machine learning, and artificial intelligence, with relevance to HEP and other scientific domains such as astronomy and materials science. The Data Science Education Community of Practice (DSECOP) [65], a program funded by the APS Innovation Fund and led by the APS Group on Data Science (GDS), seeks to support physics educators in integrating data science in their courses. DSECOP achieves this through

- A Slack community of physics educators and industry professionals to discuss data science education in physics courses. Specifically, conversation will be around challenges, opportunities, and cutting edge skills necessary for a wide-range of jobs.

- Workshops [66] to promote shared understanding and solidify the community
- Supporting DSECOP Fellows, a group of early career physicists (graduate students and postdocs) who receive modest stipends to develop and test data science education materials [67].

The DSECOP Fellow program is a good example of how researches at an early career stage can be strongly involved in curriculum development. In several of the example courses described in Sec. 3, students and postdocs from the instructors research group were involved in curriculum development and course delivery as part of their professional development. The same is true for several of the training and bootcamp events described in Sec. 2.

Large NSF Institutes such as the Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) [68], AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI) [69], and the Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery Institute (A3D3) [70] have HEP as a research driver and substantial efforts in education and training. The development of course curriculum for data science and machine learning is synergistic with particle physics research efforts within these and other institutes.

4.1 What does HEP research have to offer for ML/DS Education?

HEP research has much to offer education in data science and machine learning. Research in HEP has long required advanced, cutting-edge computing techniques, and physicists have historically contributed to the development of these methods. Over the past decades, there have been great advances in the data processing power of machine learning algorithms and these methods have quickly been adopted by physicists to address the unique timing, memory, and latency constraints of HEP experiments.

Our science typically involves analysis of large datasets generated by complex instruments at the frontier of scientific research. It is enabled by application of machine learning methods and data science tools which helps demonstrate the power and importance of these in a scientific research setting as well better understand their limitations. In recent times, cutting edge research in ML methods have been tested for their effectiveness and scalability in problems that interest high energy physics. For instance, generative models have found application in calorimeter simulation [71, 72], graph neural networks (GNNs) have been explored for particle flow reconstruction [73], and jet classification [74]. These applications serve as compelling evidence of wide range applicability of ML models for large, complicated datasets. Incorporating these exercises in ML pedagogy can enable the students to learn about the analytical and practical aspects of implementation of complex models that include hyperparameter optimization, data and model parallelization, uncertainty quantification, and model interpretation.

In short, in particle physics we have some of the most compelling scientific applications of data science and machine learning that involve very large and complex datasets.

Particle physics is also impacting machine learning research and therefore machine learning education. The constraints of HEP experiments and known symmetries of physical systems create a rich environment for the development of novel and physics-informed machine

learning (see for example [75–78]). There are even entire conferences and workshops dedicated to this intersection including the Microsoft Physics ∩ ML lecture series [32] and the ML and the Physical Sciences workshop at NeurIPS [33, 34].

Exploration of machine learning models within the domain of high energy physics goes beyond usual regression and classification problems. The pursuit of discovery in physics requires explicit understanding of the causal relationship between the inputs and outputs of an analysis model and classical ML techniques that help understand such relationships-polynomial regression models, decision trees, and random forests for instance, have found numerous applications in physics problems, including parameterized cross-section estimation for novel physics models [79] and event classification for $H \to \gamma\gamma$ channel in search of Higgs boson at the LHC [80]. However, deep neural networks are becoming increasingly popular and showing improved performance over simpler ML techniques (see for instance [81] for comparison of classical and deep neural techniques for identifying longitudinal polarization fraction in same-sign WW production). These complex, highly nonlinear models often comprising $\mathcal{O}(100k)$ parameters are extremely difficult to explain and often regarded as black box surrogates. Recent literature has focused towards explainable AI (xAI) and a number of methods [82–87] have been explored to identify importance of features and intermediate hidden layers in the context of a wide range of deep neural network models. Application of these methods in HEP research [88,89] is quickly becoming popular, as model interpretability remains a crucial aspect to determine the relevance of physics insights for ML models. With a long history of using interpretable models for physics research, HEP research allows validation of xAI techniques for mainstream ML research and pedagogy.

4.2 What does ML/DS Education have to offer HEP research?

From the very beginning of high energy collider experiments, data analysis and statistical techniques have been integral for HEP research that have led to important discoveries, precision measurements, and setting limits on search for physics beyond the standard model. Frequentist and Bayesian statistical models are ubiquitous in our treatment of uncertainties associated with finite size of MC simulation, detector response, particle and jet reconstruction [90]. HEP research has, therefore, significantly benefited from pedagogical introduction to data science and statistics. As a result, recent particle physics workshops have arranged for dedicated theoretical and hands-on sessions to introduce concepts of statistics that play important role in everyday HEP research [91, 92]. A number of tutorials [93, 94] and textbooks [95, 96] have been written on this subject. With the overwhelming emergence of ML techniques over the past decade in order to solve important questions in HEP research, this trend can be extrapolated to include ML training in the standard curriculum of HEP curricula and workshops.

Materials developed to familiarize students with ML methods can benefit the community in two different ways. First, they enable the researchers of the future to make use of state-of-the art tools in data analysis techniques. Second, instructors and professors who develop these materials can benefit from these ventures by learning about the most recent developments in different areas of ML and CS and then applying those methods in their own research. This also opens up opportunities for multi-disciplinary projects such as FAIR4HEP [97] and A3D3 [70], that can symbiotically benefit research across different

fields of study such as HEP, multimessenger astronomy and computational neuroscience. Such endeavors help make HEP research more visible to multiple communities, developing a broader interest among students and researchers in learning about and solving problems in HEP.

Finally, as all instructors know, from teaching assistant to professor levels, teaching a course really helps one understand at a deeper level the material being taught. This is true for traditional physics curriculum as well as courses at the intersection of physics, data science and machine learning. Teaching in this broader space makes us better researchers in particle physics as practitioners of these tools and methods, and keeping up with current developments (to some limited degree).

5 Challenges

There a numerous challenges confronting particle physicists incorporating data science and machine learning into the physics curriculum. We touch on a few of them in this section and include some ideas for mitigating the risks associated with these challenges.

5.1 Student (and Instructor) Preparation

It may be surprising to some HEP researchers, but coding is new for many physics students at the undergraduate or even graduate level. Many are unfamiliar with languages such as Python and tools such as Git and Jupyter notebooks. Also, certain subjects like linear algebra are required to understand machine learning even at an introductory level. Lack of preparation can lead to frustration, stress and possibly failure in these type courses. Fortunately, foresight by the instructor on enforcing appropriate prerequisites and the providing of ample resources (with examples) for coding in Python such as *A Whirlwind Tour of Python* by Jake VanderPlas (O'Reilly Media, Inc) can go a long way to providing an platform for student success. On the point of prerequisites however, one must realize that there are also equity, diversity, and inclusion considerations, in that underrepresented minority students may have less exposure to the necessary coding/software background.

5.2 Rapid development

It is difficult enough for HEP researchers to keep up with current developments and literature in one's own field of research let alone two or three. Data science and machine learning are currently moving at a very rapid pace which means that some tools and approaches become obsolete on timescale of years or even between semesters. It is important to keep up on literature within the *HEP domain* on applications of data science and machine learning to HEP research. It is also useful to read CS and ML papers to understand the some of the current developments and penetrate domain jargon (as people outside of HEP try to do with our published work), but the field is moving too fast to keep current with all developments, some of which do not have an obvious application to our science. This is not to say we should look out for potential HEP applications of ML developments, but at least for undergraduate curriculum it is best to focus primarily reasonably established applications.

5.3 Distinction and Adoption

We want to make sure that courses developed in the physics department distinguish themselves from other courses and are valued by the students they are designed to teach. Two suggestions are as follows:

1. **Not trying to do too much**. Our strengths in HEP lie in the analysis and interpretation of large scientific data sets and physics-inspired AI. Leave the foundational AI pedagogy to the CS courses.

2. **Balancing physics and ML pedagogy**. Remember that its a physics course taught in the Physics Department. It's best to use as many physics examples and datasets to support your instruction as possible. I.e. Classification of jets and galaxies over cats and dogs.

5.4 Career Development at the HEP/DS/ML Intersection

Despite the demonstrated criticality and vibrancy of HEP research, from a physics background the path to a sustainable research career at the intersection of physics and ML is unclear at best. For early career researches interested in this intersection, this research interest breadth needs to be affirmed and nurtured. It is good to help make connections between early career HEP researchers with this interest and those that have transitioned from academic to industry or other pursuits.

5.5 Tools and Infrastructure

We have discussed in this paper several successful examples of tools and infrastructure utilized in courses at the intersection of physics, data science and machine learning. Some challenges around tools and infrastructure include (1) uniformity of software environment for students, (2) keeping the primary focus on physics by minimizing prerequisite on coding skills, as discussed previously, (3) providing sufficient GPU resources to train deep networks often of of most interest, (4) hosting/caching of datasets for use in course projects and (5) maintaining a working software environment with modern DS/ML tools between course transitions. On the point of maintenance, it is not uncommon for the API for some software package to change between semesters and break notebooks that worked before, requiring care for maintenance and evolution of course materials.

6 Future Directions and Recommendations

In this section, we describe some future directions based on the experiences just described that would help HEP researchers interested in physics education that includes data science and machine learning pedagogy.

6.1 Tools and Infrastructure

To address prerequisites, a basic Python-based programming course for physicists can go a long way toward improving the foundation for students.

The most important consideration in terms of tools and infrastructure is to have these element work well without detracting from the learning environment. As a simplistic example, if I want to use a lamp to illuminate a room, I just want the lamp to be functional and the electrical infrastructure to work without being distracted with all the details of how electrical current arrives at the outlet (of course, that is interesting in other contexts). The same is true for a course in physics and machine learning.

Software and tools to launch student code within notebooks on CPUs, GPUs, and other resources to study aspects of physics should be open, portable, robust and easy to make physics education using DS/ML most effective.

Containers are a great technology to provide custom, course-specific software and data environments for use by student on infrastructure. These custom containers can be hosted externally and launched on cloud-based services to execute student notebooks. Students just write code in notebooks with a common software environment and the backend resources are provisioned for cell executions. Further customization of the run-time environment is of course possible with the appropriate instructions in the notebook (e.g. via `pip install`).

All of these functions are available with existing technologies. However, it is important for universities to provide support to educators, who are most often not experts in these technologies, in maintaining a working environment of tools and infrastructure. Increased sharing of experiences with the tools and infrastructure used for education in data science and machine learning among researchers in HEP and other fields is strongly encouraged.

6.2 Role of Open Data and AI Models Adhering to FAIR Principles

The FAIR principles (findable, accessible, interoperable, and reusable) were originally proposed to inspire scientific data management for reproducibility and maximal reusability [98]. These principles can be extended as guidelines to explore management and preservation of digital objects like research software [99] and AI models [100]. If data and models are preserved in accordance with the FAIR principles, they can be reliably reused by researchers and educators to reproduce benchmark results for both research and pedagogical purposes. For instance, the detailed analysis of FAIR and AI-readiness of the CMS $H(b\bar{b})$ dataset in Ref. [101] has explained how the FAIR readiness of this dataset has been useful in building ML exercises for the course [42] offered at UCSD. In fact, making data and models FAIR helps understand their context, content, and format, enabling transparent provenance and reproducibility [102]. Such practices help with interpretation of AI models by allowing comparison of benchmark results across different models [100] and application of *post-hoc* xAI methods. FAIR can facilitate education in ML in numerous ways, like interpretability, uncertainty quantification, and easy access to data/models.

The use of real data sets communicates to students that they are doing publication-quality work at the interface of machine learning and physics [103]. Course-based undergraduate research experiences (CUREs) in DS/ML and physics can also help foster a sense of identity and belonging in the field for students [104].

6.3 Curation and Coordination of Educational Materials

We recommend careful curation of materials for data science and machine learning in physics education and the open sharing of these materials. We also encourage a forum to openly

share the experiences of this type of teaching and general discovery of who has developed and taught such courses at their institutions. Answering questions like:

- What courses have been developed and delivered by our community?
- What open data is available for possible use in ML/DS education?
- What training/education/bootcamp/hackathon materials already exist?

How can we better collect and expose the above information for use in our community? The information is available but diffuse and therefore some coordination would make the sharing of knowledge and experiences much more efficient. This type of coordinated effort could make the sharing and improvement of projects utilizing HEP data more efficient.

7 Conclusions

We believe it will become ever more crucial that both our young and experienced researchers have a working understanding of data science and machine learning tools. We would like to see a continuation of community efforts towards raising he level of ML proficiency among current researchers by providing in-depth and innovative schools and other training events. It would also be advantageous to see more movement towards the addition of DS and ML studies within the physics curriculum at our educational institutions. All of this will take cooperation from the entire HEP community. We have described in this white paper some of the experiences from HEP researchers in ML education, outlined opportunities and challenges, and recommended future directions to make this area more efficient and effective for the HEP community.

References

[1] J. St. John et al., *Real-time artificial intelligence for accelerator control: A study at the Fermilab Booster*, *Phys. Rev. Accel. Beams* **24** (2021) 104601 [2011.07371].

[2] A. Butter and T. Plehn, *Generative Networks for LHC events*, 2008.08558.

[3] S. Forte and S. Carrazza, *Parton distribution functions*, 2008.12305.

[4] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014 [1902.09914].

[5] J. Duarte and J.-R. Vlimant, *Graph neural networks for particle tracking and reconstruction*, in *Artificial Intelligence for High Energy Physics*, P. Calafiura, D. Rousseau and K. Terao, eds., p. 387, World Scientific (2022), DOI [2012.01249].

[6] Z.A. Elkarghli, *Improvement of the NOvA Near Detector Event Reconstruction and Primary Vertexing through the Application of Machine Learning Methods*, Master's thesis, Wichita State U., 2020, [2112.01494].

[7] B. Nachman, *Anomaly Detection for Physics Analysis and Less than Supervised Learning*, 2010.14554.

[8] Y. Alanazi, N. Sato, P. Ambrozewicz, A.N.H. Blin, W. Melnitchouk, M. Battaglieri et al., *A survey of machine learning-based physics event generation*, 2106.00643.

[9] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, *J. Phys. Conf. Ser.* **1085** (2018) 022008 [1807.02876].

[10] D. Guest, K. Cranmer and D. Whiteson, *Deep Learning and its Application to LHC Physics*, *Ann. Rev. Nucl. Part. Sci.* **68** (2018) 161 [1806.11484].

[11] Alexander Radovic, Mike Williams, David Rousseau, Michael Kagan, Daniele Bonacorsi, Alexander Himmel, Adam Aurisano, Kazuhiro Terao, and Taritree Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics*, *Nature* **560** (2018) 41.

[12] D. Bourilkov, *Machine and Deep Learning Applications in Particle Physics*, *Int. J. Mod. Phys. A* **34** (2020) 1930019 [1912.08245].

[13] F. Psihas, M. Groh, C. Tunnell and K. Warburton, *A Review on Machine Learning for Neutrino Experiments*, *Int. J. Mod. Phys. A* **35** (2020) 2043005 [2008.01242].

[14] A.J. Larkoski, I. Moult and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, *Phys. Rept.* **841** (2020) 1 [1709.04464].

[15] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby et al., *Machine learning and the physical sciences*, *Rev. Mod. Phys.* **91** (2019) 045002 [1903.10563].

[16] M.D. Schwartz, *Modern Machine Learning and Particle Physics*, *Harv. Data Sci. Rev.* **3** (2021) [2103.12226].

[17] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman and D. Shih, *Machine Learning in the Search for New Fundamental Physics*, *Nature Rev. Phys.* **4** (2022) 399 [2112.03769].

[18] A.M. Deiana et al., *Applications and techniques for fast machine learning in science*, *Front. Big Data* **5** (2022) 787421 [2110.13041].

[19] M. Feickert and B. Nachman, *A Living Review of Machine Learning for Particle Physics*, 2102.02770.

[20] V. Dhar, *Data science and prediction*, *Communications of the ACM* **56** (2013) 64.

[21] S. Shukla Shubhendu and J. Vijay, *Applicability of artificial intelligence in different fields of life*, *International Journal of Scientific Engineering and Research* **1** (2013) 28.

[22] "NSF Research Experience for Undergraduates program." https://www.nsf.gov/crssprgm/reu.

[23] "International Masterclasses hands-on particle physics." https://physicsmasterclasses.org.

[24] "UCSD Data Science Capstone." https://dsc-capstone.github.io.

[25] CMS Collaboration, "2020 CMS Data Analysis School." (accessed on August 28, 2020) https://lpc.fnal.gov/programs/schools-workshops/cmsdas.shtml.

[26] "2020 Hands-on Advanced Tutorial Sessions at the LPC." (accessed on August 28, 2020)https://lpc.fnal.gov/programs/schools-workshops/hats.shtml.

[27] "CERN Summer School." https://home.cern/summer-student-programme.

[28] HEP Software Foundation Software Training Center website. https://hepsoftwarefoundation.org/training/curriculum.html.

[29] "Computational and data science training for high energy physics." https://codas-hep.org.

[30] D.S. Katz et al., *Software Sustainability & High Energy Physics*, in *Sustainable Software in HEP*, 10, 2020, DOI [2010.05102].

[31] "2020 ML4Jets workshop 2020." (accessed on August 28, 2020)https://iris-hep.org/projects/ml4jets.html.

[32] Microsoft, "*Physics Meets ML Lecture Series*." http://physicsmeetsml.org.

[33] "*2020 Machine Learning and the Physical Sciences Workshop*." https://ml4physicalsciences.github.io/2020.

[34] "*2021 Machine Learning and the Physical Sciences Workshop*." https://ml4physicalsciences.github.io/2021.

[35] The 2019 US-ATLAS Computing Bootcamp website. https://sammeehan.com/2019-08-19-usatlas-computing-bootcamp.

[36] The 2020 US-ATLAS Computing Bootcamp website. https://indico.cern.ch/event/933434.

[37] Pankaj Mehta, "*Machine Learning for Physicists*." http://physics.bu.edu/~pankajm/PY895-ML.html.

[38] Michael Coughlin, "*Big Data in Astrophysics*." https://github.com/mcoughlin/ast8581_2022_Spring.

[39] M. Erdmann, J. Glombitza, G. Kasieczka and U. Klemradt, *Deep Learning for Physics Research*, World Scientific (2021), 10.1142/12294, [https://www.worldscientific.com/doi/pdf/10.1142/12294].

[40] P. Calafiura, D. Rousseau and K. Terao, *Artificial Intelligence for High Energy Physics*, World Scientific (2022), 10.1142/12200.

[41] P. Mehta, M. Bukov, C.-H. Wang, A.G. Day, C. Richardson, C.K. Fisher et al., *A high-bias, low-variance introduction to machine learning for physicists*, *Phys. Rep.* **810** (2019) 1 [1803.08823].

[42] Javier Duarte, "*Particle Physics and Machine Learning*." https://jduarte.physics.ucsd.edu/capstone-particle-physics-domain. 10.5281/zenodo.4768815.

[43] E.A. Moreno, T.Q. Nguyen, J.-R. Vlimant, O. Cerri, H.B. Newman, A. Periwal et al., *Interaction networks for the identification of boosted $H \to b\bar{b}$ decays*, *Phys. Rev. D* **102** (2020) 012010 [1909.12285].

[44] Phil Harris, "*Data Science for Physics*." https://github.com/MIT-8s50/course.

[45] Savannah Thais, "*Introduction to Machine Learning*." https://github.com/savvy379/intro_to_ml.

[46] M. Neubauer, "PHYS398: Data Analysis and Machine Learning Applications." https://illinois-mla.github.io/syllabus/#course-description.

[47] C.R. Harris, K.J. Millman, S.J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau et al., *Array programming with NumPy*, *Nature* **585** (2020) 357.

[48] The pandas development team, *pandas-dev/pandas: Pandas*, Feb., 2020. 10.5281/zenodo.3509134.

[49] Wes McKinney, *Data Structures for Statistical Computing in Python*, in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, eds., pp. 56 – 61, 2010, DOI.

[50] J.D. Hunter, *Matplotlib: A 2d graphics environment*, *Computing in Science & Engineering* **9** (2007) 90.

[51] M.L. Waskom, *seaborn: statistical data visualization*, *Journal of Open Source Software* **6** (2021) 3021.

[52] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau et al., *SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python*, *Nature Methods* **17** (2020) 261.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *J. of Mach. Learn. Res.* **12** (2011) 2825.

[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds., pp. 8024–8035, Curran Associates, Inc. (2019), http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[55] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean et al., *Tensorflow: A system for large-scale machine learning*, in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[56] "UCI machine learning repository." https://archive.ics.uci.edu/ml/datasets.

[57] "Galaxy zoo challenge data." https://www.kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge/data.

[58] "CERN Open Data Portal." https://opendata.cern.ch.

[59] E. Van Dusen, *Jupyter for teaching data science*, in *SIGCSE '21: Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, p. 1359, 2021.

[60] "Anaconda package management." https://www.anaconda.com.

[61] "DockerHub." https://hub.docker.com.

[62] "Google Colaboratory." https://research.google.com/colaboratory/faq.html.

[63] "Phys398 mla docker image." https://github.com/illinois-mla/phys-398-mla-image.

[64] "Aps topical group on data science." https://engage.aps.org/gds.

[65] "Data science education community of practice (dsecop)." https://www.aps.org/programs/innovation/fund/dsecop.

[66] "Data science education community of practice (dsecop) workshops." https://dsecop.org/workshops.

[67] "Data science education community of practice (dsecop) github repository." https://github.com/GDS-Education-Community-of-Practice/DSECOP.

[68] "*Institute for Research and Innovation in Software for High-Energy Physics*." https://iris-hep.org.

[69] "*AI Institute for Artificial Intelligence and Fundamental Interactions*." https://iaifi.org.

[70] "*Accelerated Artificial Intelligence Algorithms for Data-Driven Discovery Institute*." https://a3d3.ai.

[71] M. Paganini, L. de Oliveira and B. Nachman, *CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, *Phys. Rev. D* **97** (2018) 014021 [1712.10321].

[72] M. Paganini, L. de Oliveira and B. Nachman, *Accelerating science with generative adversarial networks: an application to 3d particle showers in multilayer calorimeters*, *Physical review letters* **120** (2018) 042003.

[73] J. Pata, J. Duarte, J.-R. Vlimant, M. Pierini and M. Spiropulu, *MLPF: Efficient machine-learned particle-flow reconstruction using graph neural networks*, *Eur. Phys. J. C* **81** (2021) 381 [2101.08578].

[74] E.A. Moreno, T.Q. Nguyen, J.-R. Vlimant, O. Cerri, H.B. Newman, A. Periwal et al., *Interaction networks for the identification of boosted $H \to b\bar{b}$ decays*, *Phys. Rev. D* **102** (2020) 012010 [1909.12285].

[75] M. Cranmer, S. Greydanus, S. Hoyer, P. Battaglia, D. Spergel and S. Ho, *Lagrangian neural networks*, *arXiv preprint arXiv:2003.04630* (2020) .

[76] Z. Liu, Y. Chen, Y. Du and M. Tegmark, *Physics-augmented learning: A new paradigm beyond physics-informed learning*, *CoRR* **abs/2109.13901** (2021) [2109.13901].

[77] Z. Liu and M. Tegmark, *Machine-learning hidden symmetries*, 2109.09721.

[78] A. Maiti, K. Stoner and J. Halverson, *Symmetry-via-Duality: Invariant Neural Network Densities from Parameter-Space Correlators*, 2106.00694.

[79] A. Roy, N. Nikiforou, N. Castro and T. Andeen, *Novel interpretation strategy for searches of singly produced vectorlike quarks at the lhc*, *Physical Review D* **101** (2020) 115027.

[80] The CMS Collaboration, *Observation of a new boson with mass near 125 gev in pp collisions at $\sqrt{s} = 7$ and 8 tev*, *Journal of High Energy Physics* **2013** (2013) 1.

[81] C.W. Murphy, *Class imbalance techniques for high energy physics*, *SciPost Phys.* **7** (2019) 76.

[82] M.T. Ribeiro, S. Singh and C. Guestrin, *Model-agnostic interpretability of machine learning*, *arXiv preprint arXiv:1606.05386* (2016) .

[83] S.M. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, *Advances in neural information processing systems* **30** (2017) .

[84] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller and W. Samek, *On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation*, *PloS one* **10** (2015) e0130140.

[85] Z. Ying, D. Bourgeois, J. You, M. Zitnik and J. Leskovec, *Gnnexplainer: Generating explanations for graph neural networks*, *Advances in neural information processing systems* **32** (2019) .

[86] A. Shrikumar, P. Greenside and A. Kundaje, *Learning important features through propagating activation differences*, in *International conference on machine learning*, pp. 3145–3153, PMLR, 2017.

[87] M.S. Schlichtkrull, N. De Cao and I. Titov, *Interpreting Graph Neural Networks for NLP With Differentiable Edge Masking*, in *International Conference on Learning Representations*, 2020 [2010.00577].

[88] D. Turvill, L. Barnby, B. Yuan and A. Zahir, *A survey of interpretability of machine learning in accelerator-based high energy physics*, in *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, p. 77, IEEE, 2020.

[89] F. Mokhtar, R. Kansal, D. Diaz, J. Duarte, J. Pata, M. Pierini et al., *Explaining machine-learned particle-flow reconstruction*, in *4th Machine Learning and the Physical Sciences Workshop at the 35th Conference on Neural Information Processing Systems*, 2021 [2111.12840].

[90] L. Lyons, *Bayes and frequentism: a particle physicist's perspective*, *Contemporary Physics* **54** (2013) 1.

[91] "*2021 CERN-Fermilab HCP Summer School*." https://indico.cern.ch/event/1023573/.

[92] "*Theoretical Advanced Study Institute Summer School 2018 "Theory in an Era of Data"*." https://sites.google.com/a/colorado.edu/tasi-2018-wiki/.

[93] R.J. Barlow, *Practical statistics for particle physics*, *arXiv preprint arXiv:1905.12362* (2019) .

[94] G. Cowan, "Statistics for Particle Physicists." https://cds.cern.ch/record/2773595, 2021.

[95] O. Behnke, K. Kröninger, G. Schott and T. Schörner-Sadenius, *Data analysis in high energy physics: a practical guide to statistical methods*, John Wiley & Sons (2013).

[96] L. Lista, *Statistical methods for data analysis in particle physics*, vol. 941, Springer (2017).

[97] "FAIR4HEP: Findable, Accessible, Interoperable, and Reusable Frameworks for Physics-Inspired Artificial Intelligence in High Energy Physics." https://fair4hep.github.io/.

[98] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak et al., *The FAIR guiding principles for scientific data management and stewardship*, *Sci. Data* **3** (2016) 1.

[99] A.-L. Lamprecht, L. Garcia, M. Kuzak, C. Martinez, R. Arcila, E. Martin Del Pico et al., *Towards FAIR principles for research software*, *Data Sci. J.* **3** (2020) 37.

[100] D.S. Katz, F. Psomopoulos and L. Castro, *Working towards understanding the role of FAIR for machine learning*, *DaMaLOS@ ISWC* (2021) 1.

[101] Y. Chen, E. Huerta, J. Duarte, P. Harris, D.S. Katz, M.S. Neubauer et al., *A fair and ai-ready higgs boson decay dataset*, *Sci. Data* **9** (2022) 1 [2108.02214].

[102] S. Samuel, F. Löffler and B. König-Ries, *Machine learning pipelines: provenance, reproducibility and fair data principles*, in *Provenance and Annotation of Data and Processes*, p. 226, Springer (2020).

[103] V. Acquaviva, *Teaching machine learning for the physical sciences: A summary of lessons learned and challenges*, in *Teaching ML workshop at the European Conference of Machine Learning 2021*, 2021 [2108.08313].

[104] E. Stanfield, C.D. Slown, Q. Sedlacek and S.E. Worcester, *A course-based undergraduate research experience (CURE) in biology: Developing systems thinking through field experiences in restoration ecology*, *CBE—Life Sciences Education* **21** (2022) ar20.