nature physics

PERSPECTIVE
https://doi.org/10.1038/s41567-018-0342-2

OPEN

Open is not enough

Xiaoli Chen[1,2], Sünje Dallmeier-Tiessen[1]*, Robin Dasler[1,11], Sebastian Feger[1,3], Pamfilos Fokianos[1], Jose Benito Gonzalez[1], Harri Hirvonsalo[1,4,12], Dinos Kousidis[1], Artemis Lavasa[1], Salvatore Mele[1], Diego Rodriguez Rodriguez[1], Tibor Šimko[1]*, Tim Smith[1], Ana Trisovic[1,5]*, Anna Trzcinska[1], Ioannis Tsanaktsidis[1], Markus Zimmermann[1], Kyle Cranmer[6], Lukas Heinrich[6], Gordon Watts[7], Michael Hildreth[8], Lara Lloret Iglesias[9], Kati Lassili-Perini[4] and Sebastian Neubert[10]

The solutions adopted by the high-energy physics community to foster reproducible research are examples of best practices that could be embraced more widely. This first experience suggests that reproducibility requires going beyond openness.

Open science and reproducible research have become pervasive goals across research communities, political circles and funding bodies[1–3]. The understanding is that open and reproducible research practices enable scientific reuse, accelerating future projects and discoveries in any discipline. In the struggle to take concrete steps in pursuit of these aims there has been much discussion and awareness-raising, often accompanied by a push to make research products and scientific results open quickly.

Although these are laudable and necessary first steps, they are not sufficient to bring about the transformation that would allow us to reap the benefits of open and reproducible research. It is time to move beyond the rhetoric and the trust in quick fixes and start designing and implementing tools to power a more profound change.

Our own experience from opening up vast volumes of data is that openness cannot simply be tacked on as an afterthought at the end of the scientific endeavour. In addition, openness alone does not guarantee reproducibility or reusability, so it should not be pursued as a goal in itself. Focusing on data is also not enough: it needs to be accompanied by software, workflows and explanations, all of which need to be captured throughout the usual iterative and closed research lifecycle, ready for a timely open release with the results.

Thus, we argue that having the reuse of research results as a goal requires the adoption of new research practices during the data analysis process. Such practices need to be tailored to the needs of each given discipline with its particular research environment, culture and idiosyncrasies. Services and tools should be developed with the idea of meshing seamlessly with existing research procedures, encouraging the pursuit of reusability as a natural part of researchers' daily work (Fig. 1). In this way, the generated research products are more likely to be useful when shared openly.

In tackling the challenge of enabling reusable research, we keep these ideas as our guiding light when putting changes into practice in our community—high-energy physics (HEP). Here, we illustrate our approach, particularly through our work at CERN, and present our community's requirements and rationale. We hope that the explanation of our challenges and solutions will stimulate discussions around the practical implementation of workflows for reproducible and reusable research more widely in other scientific disciplines.

Approaching reproducibility and reuse in HEP

To set the stage for the rest of this piece, we first construct a more nuanced spectrum in which to place the various challenges facing HEP, allowing us to better frame our ambitions and solutions. We choose to build on the descriptions introduced by Carole Goble[4] and Lorena A. Barba[5] shown in Table 1.

These concepts assume a research environment in which multiple labs have the equipment necessary to duplicate an experiment, which essentially makes the experiments portable. In the particle physics context, however, the immense cost and complexity of the experimental set-up essentially make the independent and complete replication of HEP experiments unfeasible and unhelpful. HEP experiments are set up with unique capabilities, often being the only facility or instrument of their kind in the world; they are also constantly being upgraded to satisfy requirements for higher energy, precision and level of accuracy. The experiments at the Large Hadron Collider (LHC) are prominent examples. It is this uniqueness that makes the experimental data valuable for preservation so that it can be later reused with other measurements for comparison, confirmation or inspiration.

Our considerations here really begin after gathering the data. This means that we are more concerned with repeating or verifying the computational analysis performed over a given dataset rather than with data collection. Therefore, in Table 2 we present a variation of these definitions that takes into account a research environment in which 'experimental set-up' refers to the implementation of a computational analysis of a defined dataset, and a 'lab' can be thought of as an experimental collaboration or an analysis group.

In the case of computational processes, physics analyses themselves are intrinsically complex due to the large data volume and algorithms involved[6]. In addition, the analysts typically study more than one physics process and consider data collected under different running conditions. Although comprehensive documentation on the analysis methods is maintained, the complexity of the software implementations often hides minute but crucial details,

[1]CERN, Geneva, Switzerland. [2]Sheffield University, Sheffield, UK. [3]Stuttgart University, Stuttgart, Germany. [4]Helsinki Institute of Physics, Helsinki, Finland. [5]Cambridge University, Cambridge, UK. [6]NYU, New York, NY, USA. [7]University of Washington, Seattle, WA, USA. [8]University of Notre Dame, Notre Dame, IN, USA. [9]Instituto de Física de Cantabria CSIC-UC, Santander, Spain. [10]Heidelberg University, Heidelberg, Germany. [11]Present address: DataCite, German National Library of Science and Technology, Hanover, Germany. [12]Present address: CSC, Espoo, Finland. *e-mail: sunje.dallmeier-tiessen@cern.ch; tibor.simko@cern.ch; ana.trisovic@cern.ch

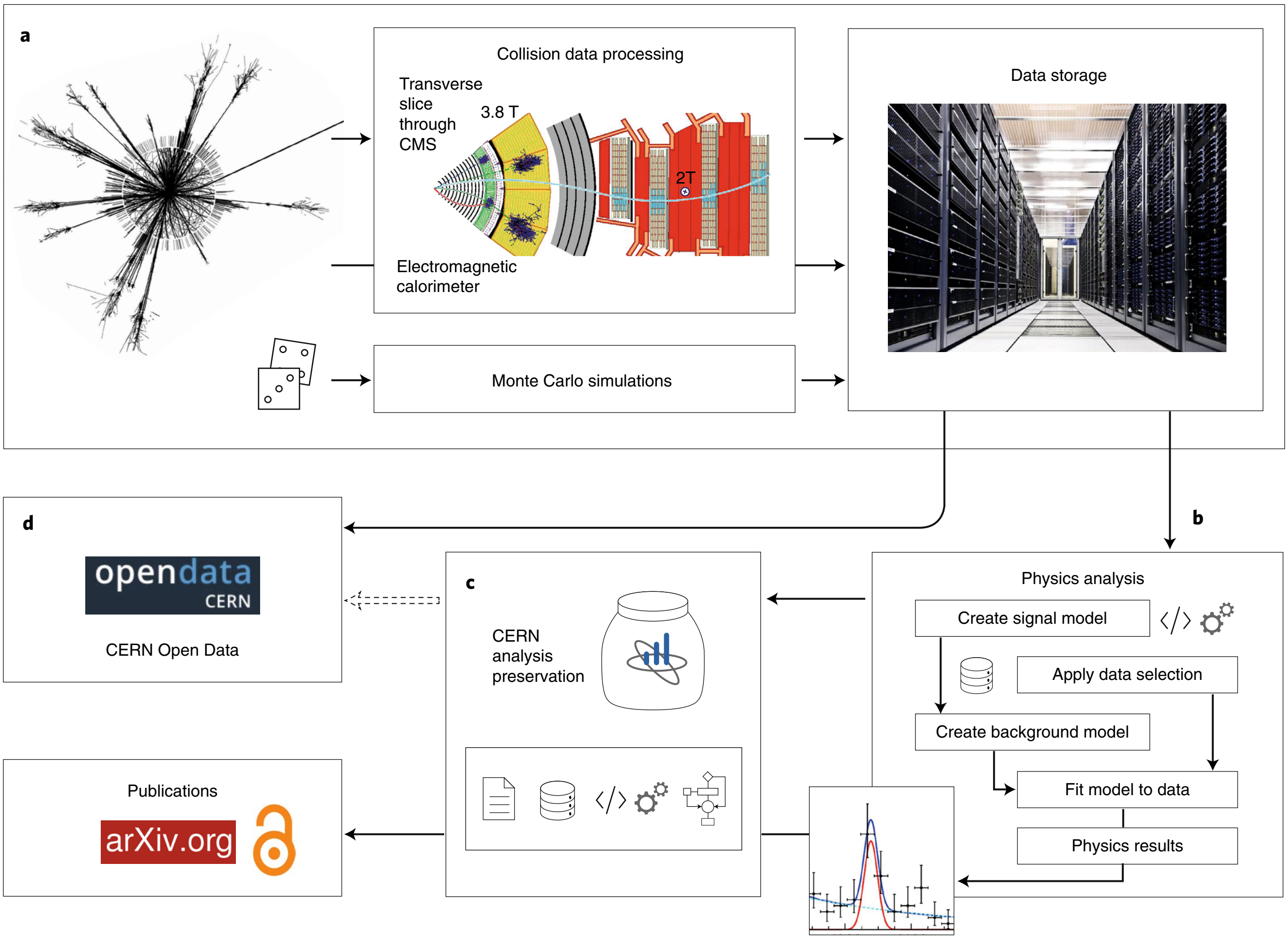

Fig. 1 | Data continuum in LHC experiments. a, The experimental data from proton–proton collisions in the Large Hadron Collider are being collected by particle detectors run by the experimental collaborations ALICE, ATLAS, CMS and LHCb. The raw experimental data is further filtered and processed to give the collision dataset formats that are suitable for physics analyses. In parallel, the computer simulations are being run in order to provide necessary comparison of experimental data with theoretical predictions. **b**, The stored collision and simulated data are then released for individual physics analyses. A physicist may perform further data reduction and selection procedures, which are followed by a statistical analysis on the data. Physics results are derived taking into account statistical and systematic uncertainties. The results often summarize which theoretical models have predictions that are consistent with the observations once background estimates have been included. The analysis assets being used by the individual researcher include the information about the collision and simulated datasets, the detector conditions, the analysis code, the computational environments, and the computational workflow steps used by the researcher to derive the histograms and the final plots as they appear in publications. **c**, The CERN Analysis Preservation service captures all the analysis assets and related documentation via a set of 'push' and 'pull' protocols, so that the analysis knowledge and data are preserved in a trusted long-term digital repository for preservation purposes. **d**, The CERN Open Data service publishes selected data as they are released by the LHC collaborations into the public domain after an embargo period of several years depending on the collaboration data management plans and preservation policies. Credit: CERN (**a**); Dave Gandy (**b**,**c**, code icon); SimpleIcon (**b**,**c**, gear icon); Andrian Valeanu (**b**,**c**, data icon); Umar Irshad (**c**, paper icon); Freepik (**c**, workflow icon).

Table 1 | Terminology related to reproducible research introduced by Carole Goble and Lorena A. Barba

Term	Purpose	Description
Rerun	Robust	Variations on experiment and set-up, conducted in the same lab
Repeat	Defend	Same experiment, same set-up, same lab
Replicate	Certify	Same experiment, same set-up, independent lab
Reproduce	Compare	Variations on experiment and set-up, independent labs
Reuse	Transfer	Different experiment

potentially leading to a loss of knowledge concerning how the results were obtained[7].

In absence of solutions for analysis capture and preservation, knowledge of specific methods and how they are applied to a given physics analysis might be lost. To tackle these community-specific challenges, a collaborative effort (coordinated by CERN, but involving the wider community) has emerged, initiating various projects, some of which are described below.

Reuse and openness. The HEP experimental collaborations operate independently of each other, and they do not share physics results until they have been rigorously verified by internal review processes[8]. Because these reviews often involve the input of the entire collaboration, where the level of crosschecking is extensive, the measurements are considered trustworthy.

Table 2 | Terminology related to reproducible research from the angle of the particle physics environment

Term	Purpose	Description adapted to the HEP context	Example
Rerun Repeat[a]	Defend and make robust	Same implementation, same analysts	These actions are integral parts of HEP analyses. Using datasets recorded in slightly different conditions (for example, different energies or magnet orientation) to **verify** that the results are consistent across the data subsets.
		Variations in implementation, same analysts	Ensuring robustness by using the same analysis model (experiment simulation) to learn about the process or to verify the results (for example, checking the impact of slight model differences to the final result).
Replicate[b]	Certify	Same dataset, same implementation, independent analysts	**Replicate an analysis** Analysis certification includes a review using prepared written documentation, and in some scenarios even reproducing the entire (or some parts of the) analysis. Another common scenario of reproducing an analysis occurs when personnel changes.
Reproduce	Compare	Variations in implementation, independent analysts	**Reproduce a measurement** After the Higgs boson discovery was published by the CMS experiment, a group of independent researchers reproduced the measurement using CMS open data and obtained similar results (Fig. 3).
Reuse	Transfer, new (different) purpose	Original (or subset of) data, different implementation, independent analysts	**Reusing analysis data** Using data for a new independent discovery, such as the use of CMS data via the CERN Open Data portal by Thaler (MIT) and colleagues[31,32].
	Certify or transfer to a new (different) purpose	Same implementation, same or independent analysts	**Reusing analysis code** Using analysis elements for a new independent study. For example, redoing an analysis by reusing the code on a more recent dataset to improve the sensitivity (statistics) of a measurement (for example, 'evidence for' may become 'observation').
	Transfer	Same implementation, same or independent analysts	**Reusing complete analyses** Testing different analysis models using existing analyses' resources. This application can be adapted in some (more inclusive) HEP analyses, and it is captured by the RECAST project.

[a] Rerun and repeat are put together as integral parts of physics analyses. [b] It needs to be noted that in the community 'replicate' is often used interchangeably with 'reproduce'.

However, it is necessary to ensure the usability of the research in the long term. This is particularly challenging today, as much of the analysis code is available primarily within the small team that performs an analysis. We think that reproducibility requires a level of attention and care that is not satisfied by simply posting undocumented code or making data 'available on request'.

In the particular case of particle physics, it may even be true that openness itself, in the sense of unfettered access to data by the general public, is not necessarily a prerequisite for the reproducibility of the research. Take the LHC collaborations as an example: while they generally strive to be open and transparent in both their research and their software development[9,10], analysis procedures and the previously described challenges of scale and data complexity mean that there are certain necessary reproducibility use cases that are better served by a tailored tool rather than an open data repository.

Such tools need to preserve the expertise of a large collaboration that flows into each analysis. Providing a central place where the disparate components of an analysis can be aggregated at the start, and then evolve as the analysis gets validated and verified, will fill this valuable role in the community. Confidentiality might aid this process so that the experts can share and discuss in a protected space before successively opening up the content of scrutiny to ever larger audiences, first within the collaboration and then later via peer review to the whole HEP community.

Cases in point are the CERN Analysis Preservation (CAP) and Reusable Analyses (REANA), which will be described in more detail below. Their key feature is that they leave the decision as to when a dataset or a complete analysis is shared publicly in the hands of the researchers. Open access can be supported, but the architecture does not depend on either data or code being publicly available. This gives the experimental collaborations full control over the release procedure and thus fully supports internal processing, review protocols and possible embargo periods. Hence, the service is accessible to the thousands of researchers who need the information it contains in order to replicate or reuse results, but the public-facing functions in HEP are better served by other services, such as CERN Open Data[11], HEPData[12] and INSPIRE[13].

The standard data deluge in particle physics is another challenge that calls for separate approaches for reproducibility, reusability and openness. As we do not have the computational resources to enable open access and processing of raw data, there needs to be a decision on the level at which the data can meaningfully be made open to allow valuable scrutiny by the public. This is governed by the individual experiments and their respective data policies[14–17].

Enabling open and reusable research at CERN

The CERN Analysis Preservation and reuse framework[18,19] consists of a set of services and tools, sketched in Fig. 1, that assist researchers in describing and preserving all the components of a physics analysis such as data, software and computing environment—addressing the points discussed earlier. These, along with the associated documentation, are kept in one place so that the analysis, or parts of it, can be reused even several years after the publication of the original scientific results.

The CERN Analysis Preservation and reuse framework relies on three pillars:

1. Describe: adequately describe and structure the knowledge behind a physics analysis in view of its future reuse. Describe all the assets of an analysis and track data provenance. Ensure sufficient documentation and capture associated links.

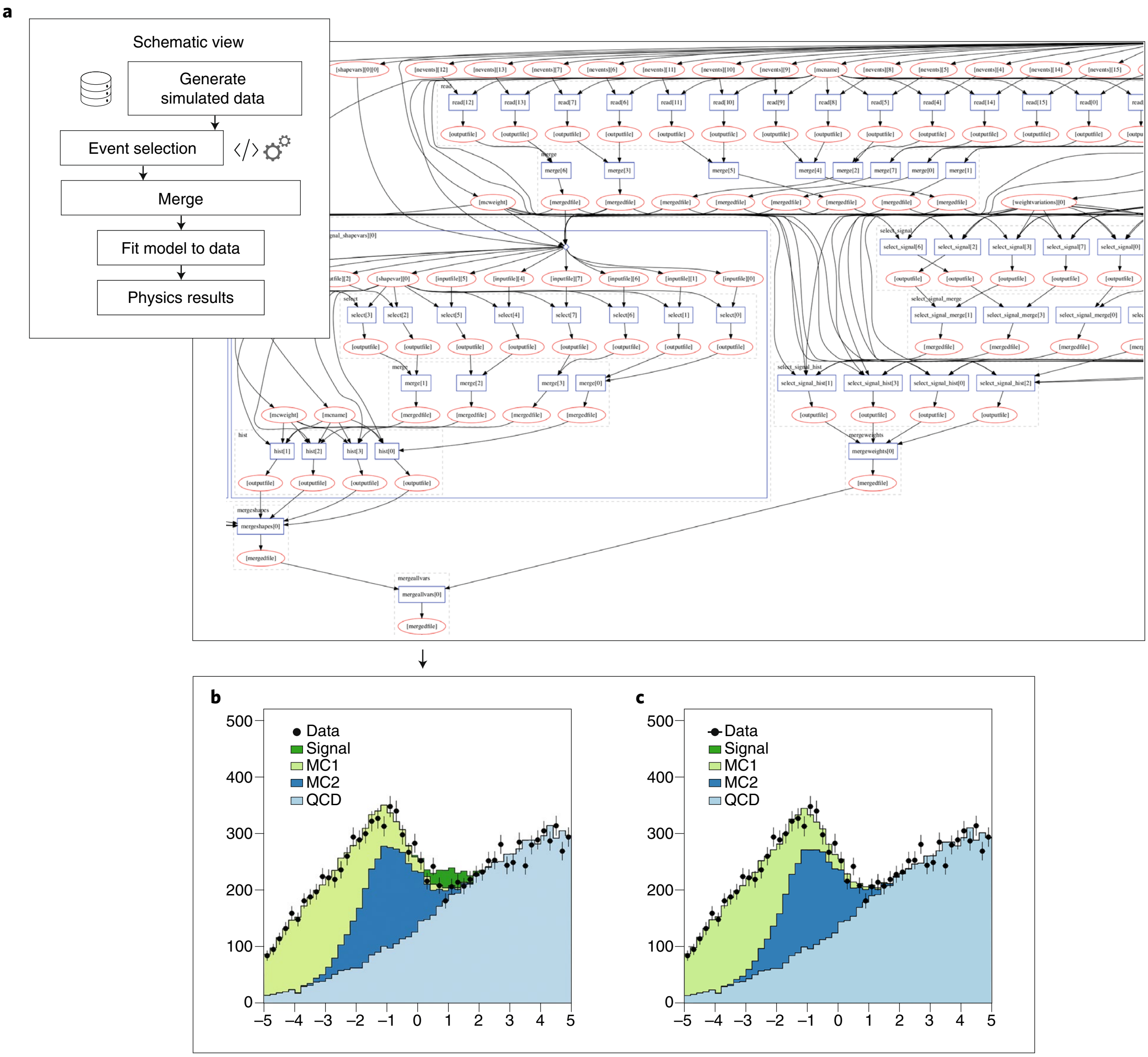

Fig. 2 | Example of a complex computational workflow on REANA mimicking a beyond the standard model (BSM) analysis . This figure shows an example where the experimental data is compared to the predictions of the standard model with an additional hypothesized signal component. The example permits one to study the complex computational workflows used in typical particle physics analyses. **a**–**c**, The computational workflow (**a**) may consist of several tens of thousands of computational steps that are massively parallelizable and run in a cascading 'map-reduce' style of computations on distributed compute clusters. The workflow definition is modelled using the Yadage workflow specification and produces an upper limit on the signal strength of the BSM process. A typical search for BSM physics consists of simulating a hypothetical signal process (**c**), as well as the background processes predicted by the standard model with properties consistent with the hypothetical signal (marked dark green in (**b**)). The background often consists of simulated background estimates (dark blue and light green histograms) and data-driven background estimates (light blue histogram). A statistical model involving both signal (dark green histogram) and background components is built and fit to the observed experimental data (black markers). **b**, Results of the model in its pre-fit configuration at nominal signal strength. We can see the excess of the signal over data, meaning that the nominal setting does not describe the data well. The post-fit distribution would scale down the signal in order to fit the data. This REANA example is publicly available at ref. [35]. For icon credits, see Fig. 1.

2. Capture: store information about the analysis input data, the analysis code and its dependencies, the runtime computational environment and the analysis workflow steps, and any other necessary dependencies in a trusted digital repository.
3. Reuse: instantiate preserved analysis assets and computational workflows on the compute clouds to allow their validation or execution with new sets of parameters to test new hypotheses.

All of these services, developed through free and open source software, strive to enable FAIR compliant data[20] and can be set up for other communities as they are implemented using flexible data models. For all these services, capturing and preserving data provenance has been a key design feature. Data provenance facilitates reproducibility and data sharing as it provides a formal model for describing published results[7].

CERN Analysis Preservation. The CERN Analysis Preservation (CAP) service is a digital repository instance dedicated to describing and capturing analysis assets. The service uses a flexible metadata structure conforming to JavaScript Open Notation (JSON)

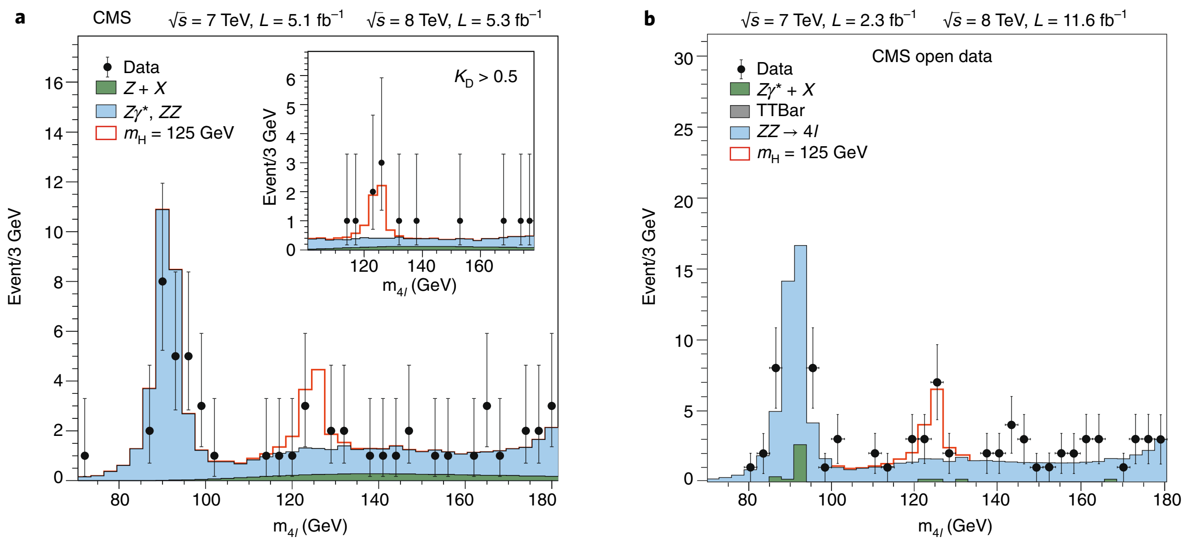

Fig. 3 | Observing the Higgs boson with CMS open data. The CMS collaboration released over one petabyte of research-grade collision and simulated datasets with associated computing tools such as the virtual machine, the analysis software and the examples of physics analyses. This permits independent researchers to understand and study the data in a way similar to how CMS physicists perform research. A characteristic example presented here is the analysis of the 'Higgs-to-four-lepton' decay channel that led to the Higgs boson experimental discovery in 2012. The Higgs boson produced in proton–proton collisions is short-lived and transforms almost instantaneously into other particles that may live longer and that are subsequently observed (directly or indirectly) by particle detectors. There are several Higgs decay modes or transformation channels possible; the present example studies the Higgs transformation into four leptons (electrons or muons) in the final state. **a**, The official CMS result[36] as it was presented during the announcement of the Higgs boson discovery in 2012. **b**, Plot produced by Nur Zulaiha Jomhari and colleagues[34] using CMS open data from 2011 and 2012 that are available on the CERN Open Data portal. The analysis using CMS open data is simplified and not scrutinized by the wider community of CMS experts. Nevertheless, it permits to run a realistic particle physics analysis example and learn about the Higgs decay physics using the same data formats, the same software tools and computational techniques that are remarkably close to procedures being used by CMS experimental physicists. Reproduced from ref. [36], CERN (**a**).

schemas that describe the analysis in order to help researchers identify, preserve and find the information about components of analyses. These JSON components define everything from experimental configurations to data samples and from analysis code to links to presentations and publications. By assembling such schemas, we are creating a standard way to describe and document an analysis in order to facilitate its discoverability and reproducibility.

The CAP service features a 'push' protocol that enables individual researchers to deposit material either by means of a user interface or with an automated command-line client. In the case of primary data, it can store links to data deposited in trusted long-term preservation stores used by the HEP experiments. For software and intermediate datasets, it can also completely ingest the material referenced by the researcher.

The CAP service can also 'pull' information from internal databases of LHC collaborations, when such information exists. Aggregating various sources of information from existing databases, source code repositories and data stores is an essential feature of the CAP service, helping researchers find and manage all the necessary information in a central place. Such an aggregation and standardization of data analysis information offers advanced search capabilities to researchers, facilitating discovery and search of high-level physics information associated with individual physics analyses.

REANA. We argue that physics analyses ideally should be automated from inception in such a way that they can be executed with a single command. Automating the whole analysis while it is still in its active phase permits to both easily run the 'live' analysis process on demand as well as to preserve it completely and seamlessly once it is over and the results are ready for publication. Thinking of restructuring a finished analysis for eventual reuse after its publication is often too late. Facilitating future reuse starts with the first commit of the analysis code.

This is the purpose served by the Reusable Analyses service, REANA: a standalone component of the framework dedicated to instantiating preserved research data analyses on the cloud. While REANA was born from the need to rerun analyses preserved in the CERN Analysis Preservation framework, it can be used to run 'active' analyses before they are published and preserved.

Using information about the input datasets, the computational environment, the software framework, the analysis code and the computational workflow steps to run the analysis, REANA permits researchers to submit parameterized computational workflows to run on remote compute clouds (as shown in Fig. 2). REANA leverages modern container technologies to encapsulate the runtime environment necessary for various analysis steps. REANA supports several different container technologies (Docker[21], Singularity[22]), compute clouds (Kubernetes[23]/OpenShift[24], HTCondor[25]), shared storage systems (Ceph[26], EOS[27]) and structured workflow specifications (CWL[28], Yadage[29]) as they are used in various research groups.

RECAST. RECAST[30] is a notable example of an application built around reusable workflows, which targets a specific particle physics use case. In particular, RECAST provides a gateway to test alternative physical theories by simulating what those theories predict and then running the simulated data through the analysis workflow used for a previous publication. The application programming interface exposes a restricted class of trustworthy, high-impact queries on the data. The experiment's data and the data processing workflow need not be exposed directly. Furthermore, the experimental collaborations can optionally maintain an approval process for the new result.

Box 1 | Guiding principles towards reproducibility

From what we have learned and discussed in this article, we distil a few general principles that may be applicable to other disciplines and individual researchers or research groups.

Define your reproducibility goals
The definition of reproducibility goals early on is essential for ensuring future reusability of scientific results. Questions to consider are: what do you produce? What is the amount of collaborative work and personnel turnover? Would you like to achieve reproducibility and reuse internally or even externally? Choosing an appropriate and balanced reproducibility strategy for an analysis can involve a number of considerations, such as the available resources, the required level of detail, the reuse value of the processed data, the analysis results and so on. Many funding agencies tend to demand data management plans and it is worthwhile if we can use this requirement to our advantage by including it in our daily routine[37].

Incorporate best practices early in your research
Adopting preservation and reproducibility practices and tools early in the research development process benefits the project and the research proponents. Invest time at the beginning of a project to do the groundwork, and document the planned outputs and how they could be organized, preserved and shared in order to support your reproducibility goals. Mentor good practice and demonstrate its usefulness. For example, use your bespoke preservation and reproducibility practices to familiarize new people with ongoing and past analyses in your team. Ensure verification and validation of your code before running new analyses. Use a version control system and continuous integration. Follow the reproducibility manifesto[38] and similar guidelines[39].

Build on what is there
There are many dependable tools available, such as data and code repositories, and methods to facilitate computational reproducibility. Do not reinvent the wheel by creating new solutions from scratch unless really necessary. Use existing tools that are already popular and available, tailor them to your needs and extend them if necessary. Opt for open source solutions with large user communities.

Structure your knowledge
What would it take to preserve and understand the knowledge associated with a research analysis? Will others be able to understand what you shared? Structure your knowledge to be both human and machine readable. Using descriptive 'readme' files is good; using a structured JSON format to describe knowledge and make it searchable is even better. Use standard vocabularies existing in your community.

Capture your content
What are the core elements that need to be included in a reproducible analysis package? What needs to be documented? Capture your analysis assets if they are located in volatile places or the location of your assets if they are stored safely. Think about input data, configurations and parameters, as well as analysis software and its dependencies. Make sure to preserve the computational environment and runtime external dependencies. Use established community platforms or talk to teams in your institution to check how you can safeguard your research.

Capture your workflows
How did you arrive at the results? Preserve your computational workflow steps. Automate your analysis and make it scriptable instead of using interactive user interfaces. Use a structured computational workflow engine that can run your analysis in a suitable computational environment.

Raise awareness
Care about the longevity of scientific results. Whether you are a professor, funding body, research associate or a graduate student, ask and discuss with your collaborators if your and their results are preserved and reusable in the long term. Can the next generation of PhD students build on top of your work? Think about publishing code, data and workflow recipes in trusted repositories.

Embrace openness whenever possible
Identify materials that can be shared publicly, publish them in trusted repositories, link and reference them to contextualize them. Consider using embargo periods for sensitive datasets or materials. Share restricted data or work in progress among your collaborators.

Enable liberal and fair reuse
Licensing and crediting is crucial. All our open scholarly materials (data, code, documentation, papers) are accompanied by liberal licenses and data/software citation recommendations and we do encourage other scientific disciplines to adopt the same principle.

The system has been used internally to streamline the reinterpretation of several experiments, and ultimately could be opened to independent researchers outside of the LHC collaborations.

CERN Open Data. The CERN Open Data portal was released in 2014 amid a discussion as to whether the primary particle physics data, due to its large volume and complexity, would find any use outside of the LHC collaborations. In 2017, Thaler and colleagues[31,32] confirmed their jet substructure model predictions using the open data from the Compact Muon Solenoid (CMS) experiment that were released on the portal in 2014, demonstrating that research conducted outside of the CERN collaborations could indeed benefit from such open data releases.

From its creation, the CERN Open Data service has disseminated the open experimental collision and simulated datasets, the example software, the virtual machines with the suitable computational environment, together with associated usage documentation that were released to the public by the HEP experiments. The CERN Open Data service is implemented as a standalone data repository on top of the Invenio digital repository framework[33]. It is used by the public, by high school and university students, and by general data scientists.

Exploitation of the released open content has been demonstrated both on the educational side and for research purposes. A team of researchers, students and summer students reproduced parts of published results from the CMS experiment using only the information that was released openly on the CERN Open Data portal. The developed code produced plots comparable to parts of the official CMS Higgs-to-four-lepton analysis results[34] (Fig. 3).

This shows that the CERN Open Data service fulfils a different and complementary use case to the CERN Analysis Preservation framework. The openness alone does not sufficiently address all the required use cases for reusable research in particle physics that is naturally born 'closed' in experimental collaborations before the analyses and data become openly published.

Challenging, but possible

In this paper we have discussed how open sharing enables certain types of data and software reuse, arguing that simple compliance with openness is not sufficient to foster reuse and reproducibility in particle physics. Sharing data is not enough; it is also essential to capture the structured information about the research data analysis workflows and processes to ensure the usability and longevity of results.

Research communities may start by using open data policies and initiating dialogues on data sharing, while embracing the reproducibility and reuse principles early on in the daily research processes. We compiled a few guiding principles that could support such dialogues (Box 1). In particle physics, the possibility of actual internal or external reuse of research outputs is an intrinsic motivation for taking part in these activities; one could assume the same for many other scientific communities.

Using computing technologies available today, solving the challenges of open sharing, reproducibility and reuse seems more feasible than ever, helping to keep research results viable and reusable in the future.

Received: 6 April 2018; Accepted: 10 September 2018;
Published online: 15 November 2018

References

1. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature News* **533**, 452–454 (2016).
2. Boulton, G. Reproducibility: International accord on open data. *Nature* **530**, 281 (2016).
3. Goodman, S. N., Fanelli, D. & Ioannidis J. P. A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).
4. Goble, C. What is reproducibility. *SlideShare* https://www.slideshare.net/carolegoble/what-is-reproducibility-gobleclean (2016).
5. Barba, L. A. Terminologies for reproducible research. Preprint at https://arxiv.org/abs/1802.03311 (2018).
6. Brun, R. in *From the Web to the Grid and Beyond* (eds Brun, R., Carminati, F. & Galli-Carminati, G.) 1–53 (Springer, Berlin, Heidelberg, 2011).
7. Pasquier, T. et al. If these data could talk. *Sci. Data* **4**, 170114 (2017).
8. Boisot, M., Nordberg, M., Yami, S. & Nicquevert, B. *Collisions and Collaboration: The Organization of Learning in the ATLAS Experiment at the LHC* (Oxford Univ. Press, Oxford, 2011).
9. Albrecht, J. et al. A roadmap for HEP software and computing R&D for the 2020s. Preprint at https://arxiv.org/abs/1712.06982 (2017).
10. Elmer, P., Neubauer, M. & Sokoloff, M. D. Strategic plan for a scientific software innovation institute (S2I2) for high energy physics. Preprint at https://arxiv.org/abs/1712.06592 (2017).
11. CERN Open Data portal; http://opendata.cern.ch/
12. HEPData; https://hepdata.net/
13. INSPIREHEP; http://inspirehep.net/
14. ATLAS Collaboration. ATLAS data access policy. *CERN Open Data Portal* https://doi.org/10.7483/opendata.atlas.t9yr.y7mz (2014).
15. Clarke, P. & LHCb Collaboration. LHCb external data access policy. *CERN Open Data Portal* https://doi.org/10.7483/opendata.lhcb.hkjw.twsz (2013).
16. CMS Collaboration. CMS data preservation, re-use and open access policy. *CERN Open Data Portal* https://doi.org/10.7483/opendata.cms.udbf.jkr9 (2012).
17. ALICE Collaboration. ALICE data preservation strategy. CERN Open Data Portal https://doi.org/10.7483/opendata.alice.54ne.x2ea (2013).
18. CERN Analysis Preservation. *GitHub* https://github.com/cernanalysispreservation (2018).
19. REANA; http://reana.io/
20. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
21. Docker; https://www.docker.com/
22. Singularity. *GitHub* https://github.com/singularityware (2018).
23. Kubernetes; https://kubernetes.io/
24. OpenShift; https://www.openshift.com/
25. HTCondor; https://research.cs.wisc.edu/htcondor/
26. Ceph; https://ceph.com/ceph-storage/
27. EOS service. *CERN* http://information-technology.web.cern.ch/services/eos-service (2018).
28. Common workflow language. *GitHub* https://github.com/common-workflow-language/common-workflow-language (2018).
29. Cranmer, K. & Heinrich, L. Yadage and Packtivity – analysis preservation using parameterized workflows. *J. Phys. Conf. Ser.* **898**, 102019 (2017).
30. Cranmer, K. & Yavin, I. RECAST — extending the impact of existing analyses. *J. High Energy Phys.* **2011**, 38 (2011).
31. Larkoski, A., Marzani, S., Thaler, J., Tripathee, A. & Xue, W. Exposing the QCD splitting function with CMS open data. *Phys. Rev. Lett.* **119**, 132003 (2017).
32. Tripathee, A., Wei, X., Larkoski, A., Marzani, S. & Thaler, J. Jet substructure studies with CMS open data. *Phys. Rev. D* **96**, 074003 (2017).
33. Invenio Software; http://invenio-software.org/
34. Jomhari, N. Z., Geiser, A. & Bin Anuar, A. A. Higgs-to-four-lepton analysis example using 2011–2012 data. *CERN Open Data Portal* https://doi.org/10.7483/opendata.cms.jkb8.rr42 (2017).
35. REANA example: BSM search. *GitHub* https://github.com/reanahub/reana-demo-bsm-search (2018).
36. Chatrchyan, S. et al. (CMS Collaboration) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B* **716**, 30–61 (2012).
37. Schiermeier, Q. Data management made simple. *Nature* **555**, 403–405 (2018).
38. Barba, L. A. Reproducibility PI manifesto. *Figshare* https://figshare.com/articles/reproducibility_pi_manifesto/104539 (2012).
39. Goodman, A. et al. Ten simple rules for the care and feeding of scientific data. *PLoS Comput. Biol.* **10**, e1003542 (2014).

Acknowledgements

Considerations and services described here benefit from funding and discussions from the EC funded projects FREYA (777523), THOR (654039) and CRISP (283745). Furthermore, we acknowledge support for the DASPOS (https://daspos.crc.nd.edu/) and Diana/HEP projects (http://diana-hep.org/).

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to S.D. or T.Š. or A.T.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.