Reproducible Research is more than Publishing Research Artefacts

A Systematic Analysis of Jupyter Notebooks from Research Articles

Problem Statement

With the advent of Open Science, researchers start to publish their research artefacts in order to allow others to reproduce their investigations. While this is beneficial for science in general, the publication often lacks a comprehensive documentation and completeness with respect to the artefacts, which in turn prevents the successful reproduction of the analyses.

- Jupyter notebooks (Fig. 1) recently gained increasing attention as a method for publishing investigations. These documents encapsulate both documentation and source code inside a single document. The corresponding web application enables to interactively view, re-execute, and extend the investigations.

- Despite the increasing use of jupyter notebooks for the publication of research investigations, their reproducibility is not automatically guaranteed.
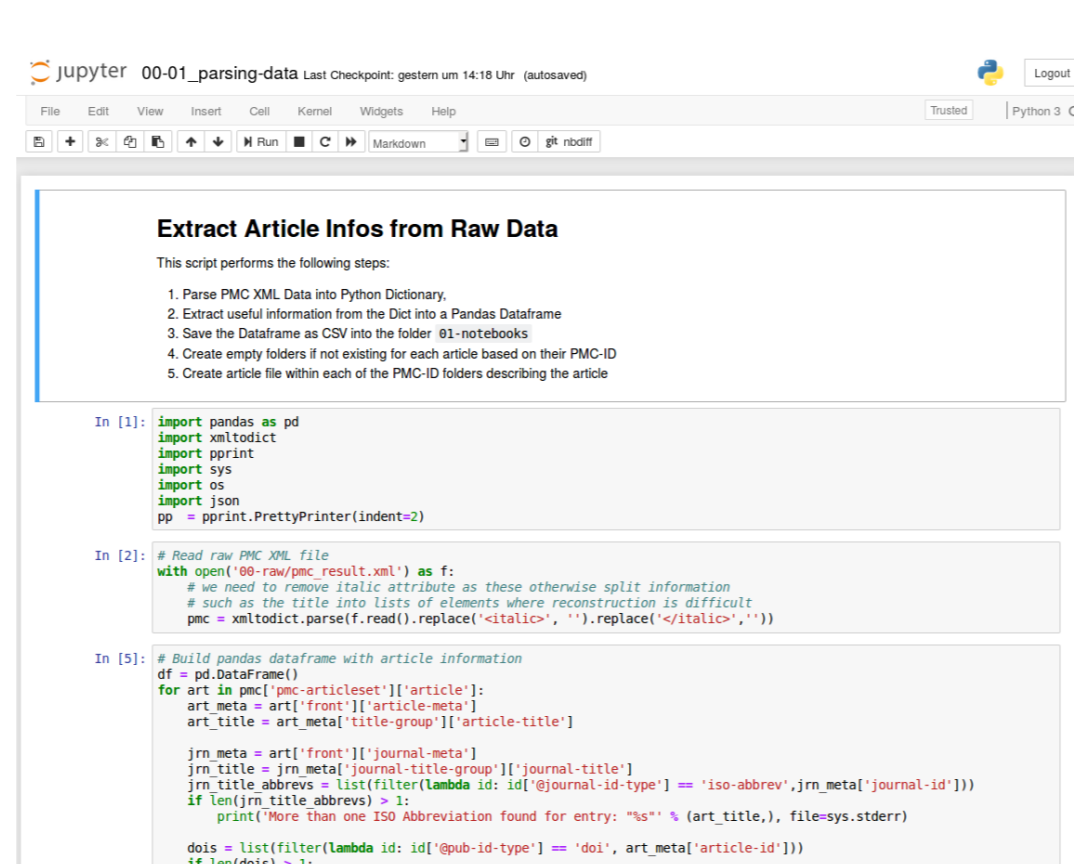

FIGURE 1: Screenshot of a jupyter notebook

Method

- We, systematically analysed jupyter notebooks mentioned in research articles with respect to their reproducibility.

- The PubMed Central Web Portal was used to systematically identify publications that mention jupyter notebooks and were recently (in 2018) published (Fig. 2 illustrates the workflow).

- Seven publications were excluded as the jupyter notebooks mentioned do not refer to the original analysis of the publication or are no longer available.
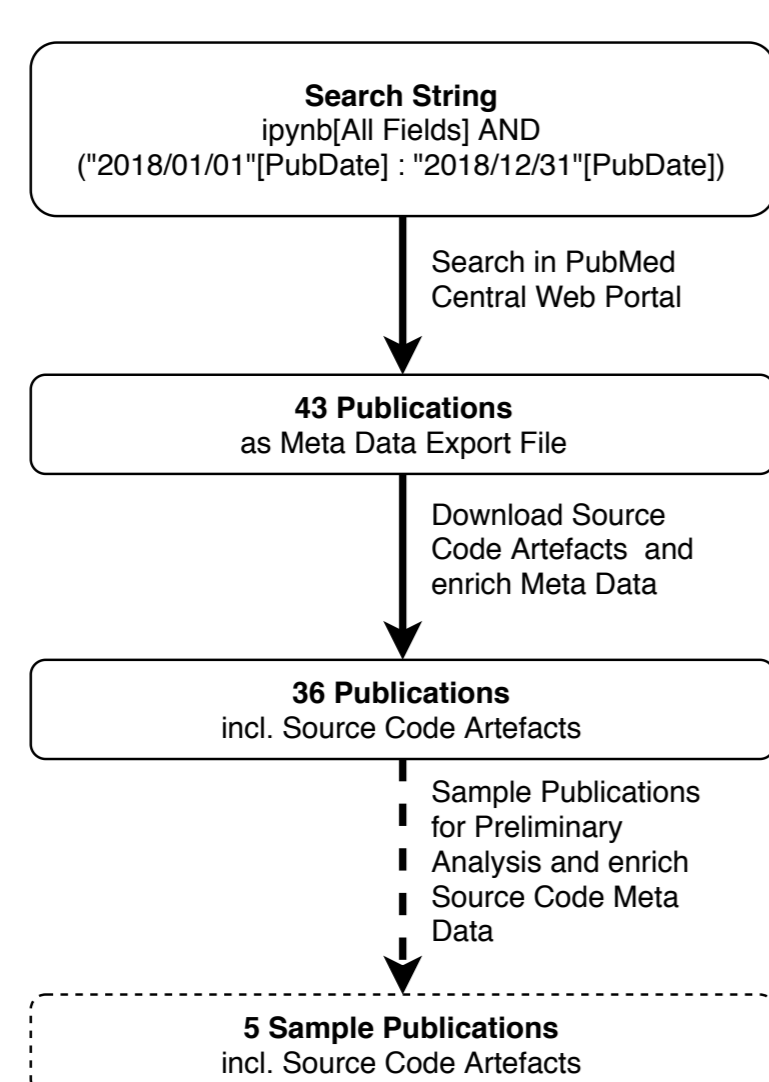
- Every identified publication will be analysed with respect to its reproducibility as specified in Fig. 3.

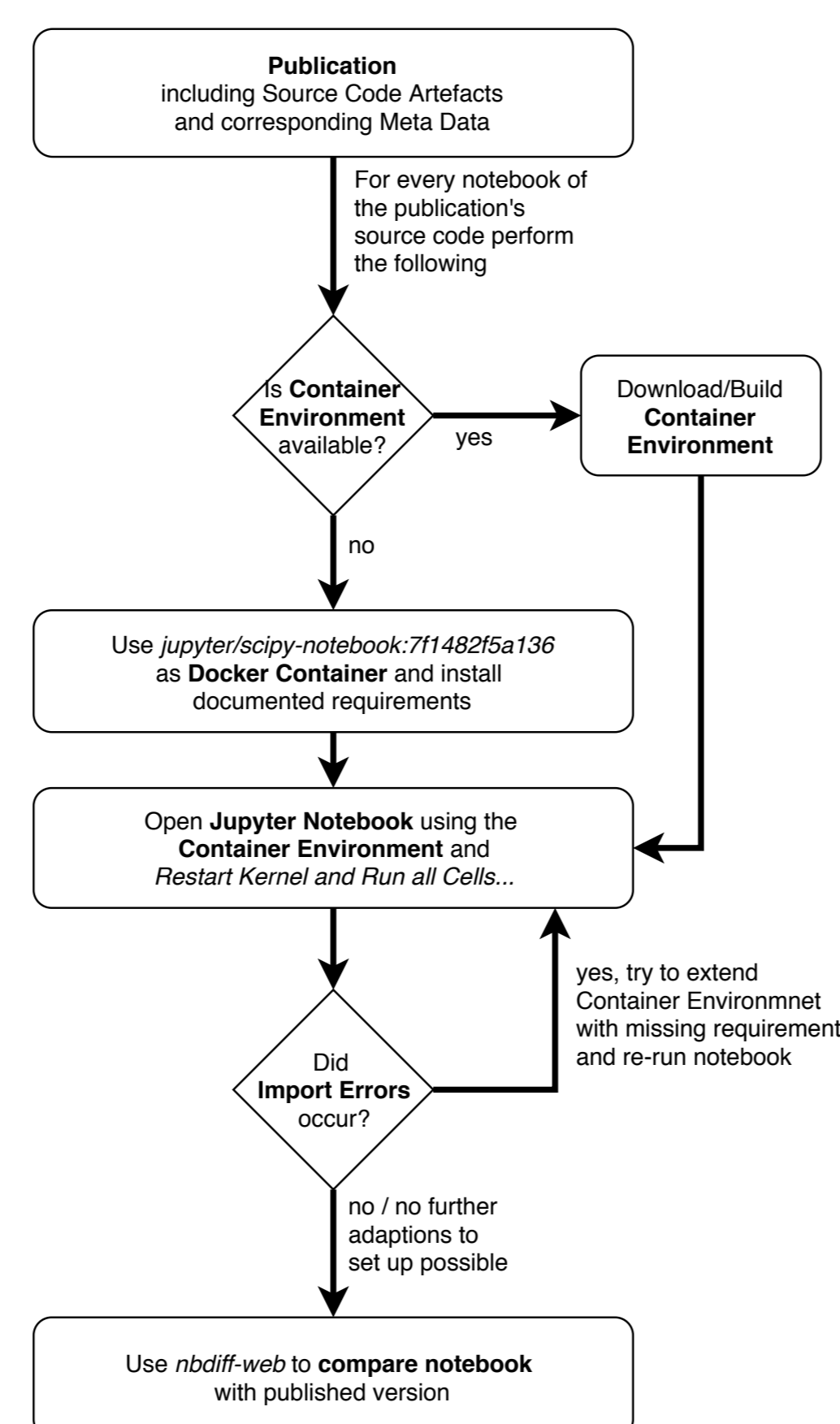FIGURE 2: Workflow used to identify relevant publications from the PubMed Central Web Portal.

FIGURE 3: Reproducibility analysis performed for every identified publication.

Results

Meta Data Analysis

The meta data has been analysed for the full set of identified publications (n=36):

- GitHub is by far the most frequently used repository for the upload of source code artefacts (see Fig. 4).

- Publications often lack mentioning the version of the source code that is used for the investigations. Instead, only the repository base url is mentioned.
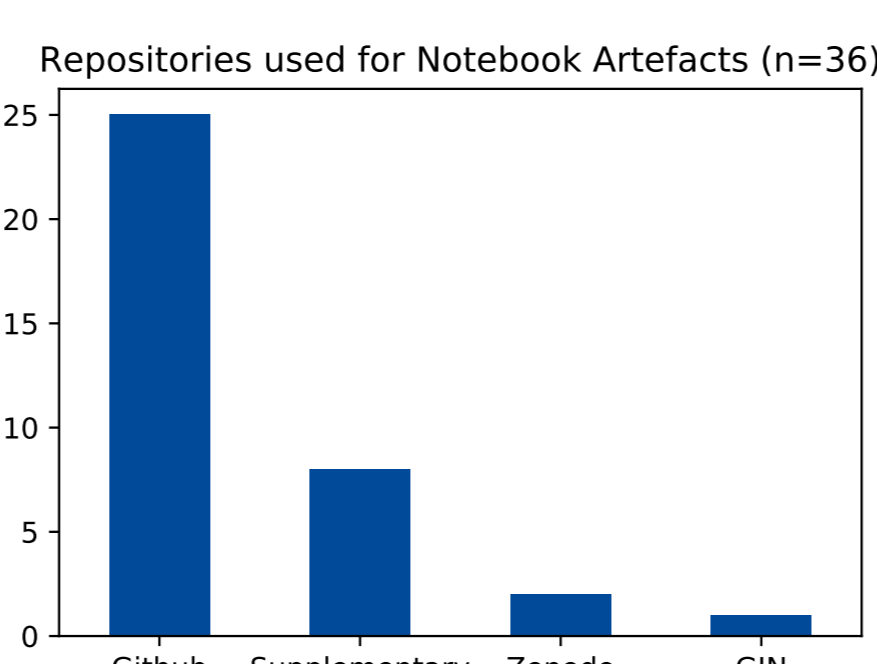

FIGURE 4: Which repositories are used to publish source code artefacts i.e., jupyter notebooks?
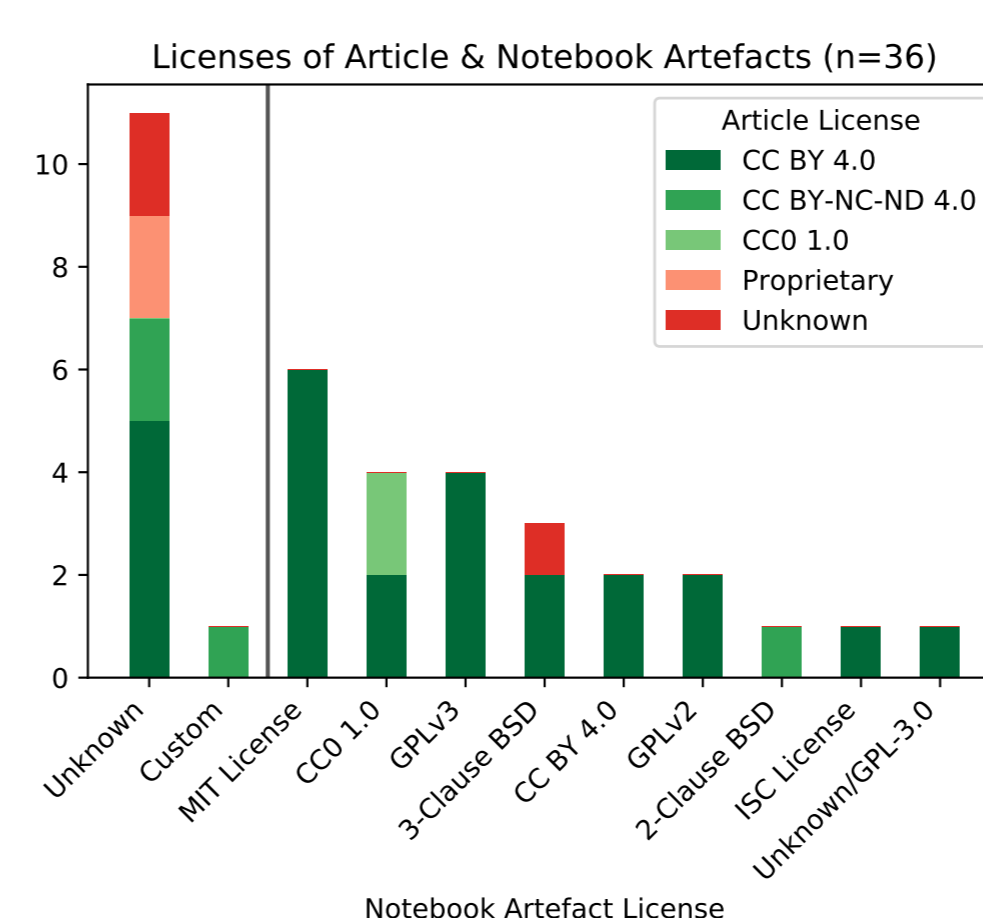

FIGURE 5: Shows the publishing license of the jupyter notebooks w.r.t. the publishing license of the article.

- Almost one third of the source code artefacts lack in providing a license for their usage (see Fig. 5) and, thus, preventing other researchers from re-using their investigations.

- Furthermore, even some of the articles are not open accessible.

Jupyter Notebook Analysis

For this preliminary analysis, we sampled five publications from the overall set. Each of these publications, was analysed as illustrated in Fig. 3:

- Only a single article referencing a repository with eleven jupyter notebooks provided a container environment in the form of a Docker description as well as a Travis-CI configuration (see Fig. 6).

- Documentation about requirements often lack in providing information about packages and specifically package versions.

- Not every jupyter notebook in the repositories is mentioned in the corresponding article.

Ref.	Ment.	Documentation	Requirements	Req. Problems
[1]	1 / 2	notebook, readme	no documentation	—
[2]	1 / 11	html, notebook, readme, readthedocs	docker, notebook, readme, readthedocs, travisci	docker fails build, custom image did not install successfully
[3]	1 / 2	mkdocs, notebook, readme	mkdocs	missing versions
[4]	4 / 6	notebook, notes, readme	notebook, readme	missing python packages
[5]	1 / 1	notebook, readme	no documentation	—

FIGURE 6: Summary of the Notebook Meta Data analysis for the five sample publications: 'Ment.' refers to the number of jupyter notebook mentions within the publication compared to the number of jupyter notebooks within the source code publication; 'Req. Problems' refers to problems with documented requirements.

- None of the source code artefacts provided a ready-to-use computing environment. However, one publication provided the description for the containerisation service docker.

- A computing environment for more than half of the jupyter notebooks (part of two articles) could not be created so that no import error occurs (see Fig. 7).

- Only three notebooks could be successfully reproduced from which one contained only definitions of functions and no calculations at all. Most of the other notebooks lack in providing complete data, threw errors during run and generate different results than the published version.
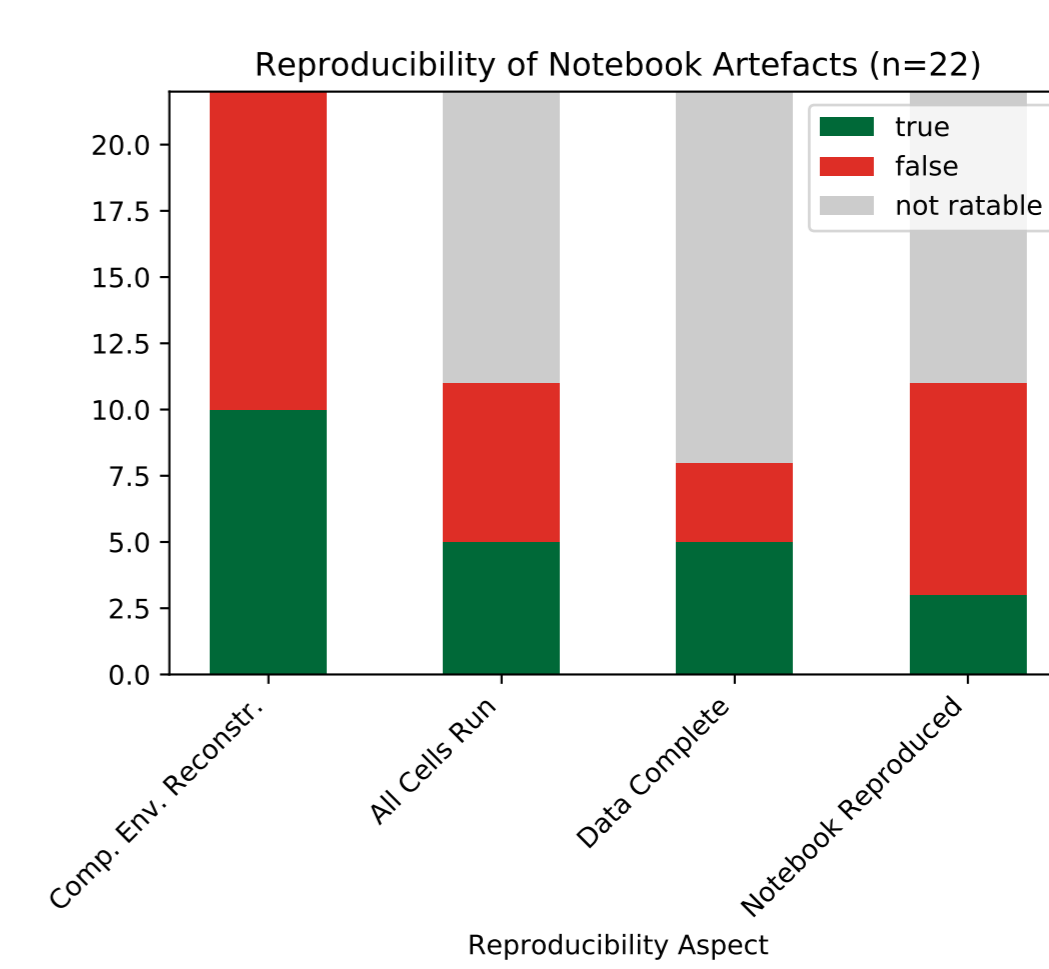

FIGURE 7: Illustration of the reproducibility analysis showing several aspects that were analysed for every notebook from the five sample publications. Some aspects could not be rated as prerequisites are not satisfied e.g., a computing environment exists.

References

[1] A. Tambe, A. East-Seletsky, G. J. Knott, J. A. Doudna, and M. R. O'Connell, "RNA binding and HEPN-nuclease activation are decoupled in CRISPR-cas13a," *Cell Reports*, vol. 24, no. 4, pp. 1025–1036, jul 2018.

[2] B. Cummins, T. Gedeon, S. Harker, and K. Mischaikow, "DSGRN: Examining the dynamics of families of logical models," *Frontiers in Physiology*, vol. 9, may 2018.

[3] J. Yang, H. Zhu, and X. Tian, "Group-level multivariate analysis in EasyEEG toolbox: Examining the temporal dynamics using topographic responses," *Frontiers in Neuroscience*, vol. 12, jul 2018.

[4] T. Collier and N. Manoukis, "Evaluation of predicted medfly (ceratitis capitata) quarantine length in the united states utilizing degree-day and agent-based models," *F1000Research*, vol. 6, p. 1863, mar 2018.

[5] K. H. Fisher-Wellman, M. T. Davidson, T. M. Narowski, C.-T. Lin, T. R. Koves, and D. M. Muoio, "Mitochondrial diagnostics: A multiplexed assay platform for comprehensive assessment of mitochondrial energy fluxes," *Cell Reports*, vol. 24, no. 13, pp. 3593–3606.e10, sep 2018.

Max Schröder [a,b], **Frank Krüger** [a], **Sascha Spors** [a]
(OpenScience.elaine@uni-rostock.de)

[a] INSTITUTE OF COMMUNICATIONS ENGINEERING | **University of Rostock**, Germany
[b] UNIVERSITY LIBRARY | **University of Rostock**, Germany