



Welcome to this survey. The purpose of this study is to evaluate MLProvLab, our extension for tracking provenance in Machine Learning Jupyter notebooks, and hence, help us improve upon it.

Please open the below link to launch MLProvLab in JupyterLab. This will take 1 minute.

https://mybinder.org/v2/gh/fusion-jena/MLProvLab/HEAD?urlpath=lab%2Ftree%2Fbinder%2Fevaluation_notebook.ipynb

There is a short introduction to MLProvLab which will help you to complete the survey. This survey should take around 10 minutes to complete. This survey is completely anonymous.

If you have any questions regarding this survey, please contact Dominik Kerzel (dominik.kerzel@uni-jena.de) or Dr. Sheeba Samuel (sheeba.samuel@uni-jena.de).

Thank you for your time!



Section A: Privacy Policy/Datenschutzerklärung

Dear scholar,

From 25th May 2018, the new General Data Protection Regulation (GDPR) (in German: Datenschutz-Grundverordnung, DSGVO) has come into effect. For compliance reasons, we are obliged to get your consent on the privacy policy before collecting any kind of personal information.

The following information applies as a supplement to the general privacy policy of the Friedrich Schiller University Jena. We request you to please kindly read both policies carefully and to agree.

Please note, as the Friedrich Schiller University Jena is headquartered in Germany, only the German version of this privacy policy is legally binding.

Liebe Wissenschaftlerinnen und Wissenschaftler,

Am 25. Mai 2018 trat die Datenschutz-Grundverordnung (DSGVO, <https://dsgvo-gesetz.de/>) in Kraft. Um Ihre personenbezogenen Daten erheben und verarbeiten zu können, benötigen wir daher aus rechtlichen Gründen zunächst Ihre Zustimmung zur Datenverarbeitung.

Die folgende Erklärung dient als Ergänzung zur Datenschutzerklärung der Friedrich-Schiller-Universität Jena. Wir möchten Sie bitten, beide Dokumente gründlich durchzulesen und mit Ihrer Einwilligung zu bestätigen, dass Sie der Datenverarbeitung zustimmen.

Privacy Policy (English) Datenschutzerklärung (Deutsch)

Collection and Use of Information

How do we collect personal information?

In this survey, we only collect data about your research contexts such as primary research fields and your research practice. All questions are not mandatory. You can omit questions you can not answer or you do not want to answer.

While browsing, some general information is stored in the server log files. We collect (1) the date and time of access to the Internet site (2) and the Internet protocol address (IP address).

All this information is needed to deliver correct website content and to optimize web content continuously. In the case of cyber-attacks, log files provide necessary information for criminal prosecution.

How do we use that information?

We use your answers to evaluate and improve MLProvLab.

The results of this study will help us in improving MLProvLab to support reproducibility. In turn, this will benefit the scientific community which uses Jupyter Notebooks.

Sharing We support the idea of generating only FAIR data. Thus, we intend to publish all answers as open data in a data repository with a digital identifier (e.g., DOI) and



Section B: General Questions

In this section, we would like to know about your research background.

B1. What is your current domain?

- Biology
- Chemistry
- Computer Science
- Physics
- Mechanics
- Environmental Sciences
- Health Sciences
- Other

Other

B2. What is your current position?

- Undergraduate Student
- Master Student
- PhD Student
- PostDoc
- Professor
- Company Employee
- Other

Other



B3. Do you use Jupyter Notebooks in your work?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Sometimes	<input type="checkbox"/>

B4. Do you use Machine Learning in your work?

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>
Sometimes	<input type="checkbox"/>

Section C: Testing MLProvLab

We would like you to test MLProvLab.

For this, please open this binder link if it is not open already at the beginning of the survey.

https://mybinder.org/v2/gh/fusion-jena/MLProvLab/HEAD?urlpath=lab%2Ftree%2Fbinder%2Fevaluation_notebook.ipynb

We provide an evaluation_notebook to answer the questions in our survey. Please do not run this Jupyter Notebook.

Before you start, you will get a brief introduction to MLProvLab in the next step.



Section D: MLProvLab Introduction

MLProvLab is an extension developed in JupyterLab to track the history (provenance) of user executions of notebooks. The tool is designed to help data scientists and ML practitioners track, capture, compare, and visualize the provenance of machine learning notebooks.

Please go below through the short introduction on MLProvLab features. You can also open it [here](#).

1. Main Menu

You can open the main menu by clicking the toolbar button.

With the sliders on the bottom, you can change the currently displayed epoch (user session) and the latest execution displayed. The epoch shows the number of times a user has created the session and Execution shows the number of times a user has executed the notebook in an epoch.

2. Execution Graph

The nodes in the execution graph denote the cells of a Jupyter Notebook and the edges are the dependencies between cells. Cells that are marked green denote if they have output, red cells denote error or deleted cells, and orange denotes if a data source has been used in the cell or edge.

You can move the nodes in the graph by dragging them to the desired position. By right-clicking and holding, you can open a context menu. You can select an option by moving your mouse to it and releasing the right button.

If you open the menu on a node, you can either open the execution info widget or the code difference widget. If you open it on an edge, you can open the execution info of the represented variable. If you click on a node group, you can focus the corresponding cell in the notebook.

3. Execution info widget

Here you can see all the information about the variables collected in an execution.



Section E: Evaluation Notebook Questions

E1. *Which version of the kernel was used in epoch '1'?*

You can use the Environment info widget.

- 7.22.0
6.33.1
5.45.3

E2. *Which external modules were used in epoch '1'?*

- matplotlib
keras
os
pandas
sklearn
numpy
seaborn
tensorflow

E3. *Are there any imported modules that were not used in epoch '3'?*

- numpy
os
pandas

E4. *Which one was the most used module in epoch '3'?*

- os
seaborn
numpy
sklearn

E5. *Which data sources were used in the notebook?*

- D:/Projects/mnist-evaluation/data/train.csv
D:/Projects/mnist-evaluation/data/test.csv
D:/Projects/mnist-evaluation/data



E6. In which execution and epoch the following figure got printed?

Use the code info widget to go through the executed code in each epoch.

Epoch: 2 | Execution: 6

Epoch: 1 | Execution: 10

Epoch: 3 | Execution: 9

E7. Which version of seaborn was used?

0.11.5

0.12.1

0.11.1

0.10.5

E8. Which versions of python were used in the notebook?

2.9.1

3.9.2

3.7.6

E9. When was the notebook last executed?

Thu, 24 Jun 2021 09:57:28 GMT

Thu, 24 Jun 2021 10:26:21 GMT

Thu, 24 Jun 2021 10:11:14 GMT

E10. Are there any differences in the python and kernel version used in the notebook in different executions?

Yes

No

E11. Which cells of the notebook in epoch '4' are dependent on the variable 'X_train'?

Execution Graph provides the dependencies of variables between cells.

Cell 1

Cell 13

Cell 14

Cell 15

Cell 16



E12. What is the accuracy score of the experiment in epoch '2' when RandomForestClassifier was used?

0.96	<input type="checkbox"/>
0.56	<input type="checkbox"/>
0.75	<input type="checkbox"/>

E13. Has the train-test split ratio for the dataset changed during different executions?

Use the code info widget to go through the executed code in each epoch. The train_test_split function provides the train-test split ratio.

Yes	<input type="checkbox"/>
No	<input type="checkbox"/>

Section F: General Impression of MLProvLab

F1. How important is each MLProvLab module for the provenance management of computational experiments?

	Not Important At All	Little Importance	Average Importance	Very Important	Absolutely Essential
Execution Graph	<input type="checkbox"/>				
Input-Output Difference	<input type="checkbox"/>				
Environment Information	<input type="checkbox"/>				
Code Information	<input type="checkbox"/>				
Import information	<input type="checkbox"/>				
General Information	<input type="checkbox"/>				
Export Module	<input type="checkbox"/>				

F2. How easy is it to find provenance information of data science scripts using MLProvLab?

	Very Easy	Easy	Neither easy nor difficult	Difficult	Very difficult
Inputs of previous executions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Outputs of previous executions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Datasets used	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Modules used	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dependencies between cells	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



	Very Easy	Easy	Neither easy nor difficult	Difficult	Very difficult
Deleted cells	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Execution Environment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Execution Date and Time	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

F3. Please rate the perceived usefulness of MLProvLab.

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
The system is easy to use	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The system is easy to navigate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The system is complex	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
It is easy to learn to use it	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
The system is important for provenance and metadata management of notebooks	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

F4. Would you like to use MLProvLab in your daily work ?

Yes
No
Maybe

F5. Select the word(s) that best describe MLProvLab.

Efficient

Comment

User Friendly

Comment



Provides data provenance management



Comment

Needs improvement



Comment

Other



Other

F6. What new features or changes would you like to see in MLProvLab ?