

# The Story of an Open Science Experiment

**Sheeba Samuel**

Friedrich Schiller University, Jena, Germany

MPDL Open Science Days

19-20 October 2021



**FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA**



**MichaelStifelCenterJena**  
for Data-Driven and Simulation Science



# Introduction

- **2019-Present:** PostDoc Researcher, University of Jena, Germany
- **2016-2019:** PhD (A Provenance-based Semantic Approach to Support Understandability, Reproducibility, and Reuse of Scientific Experiments)



@sheebasamuel



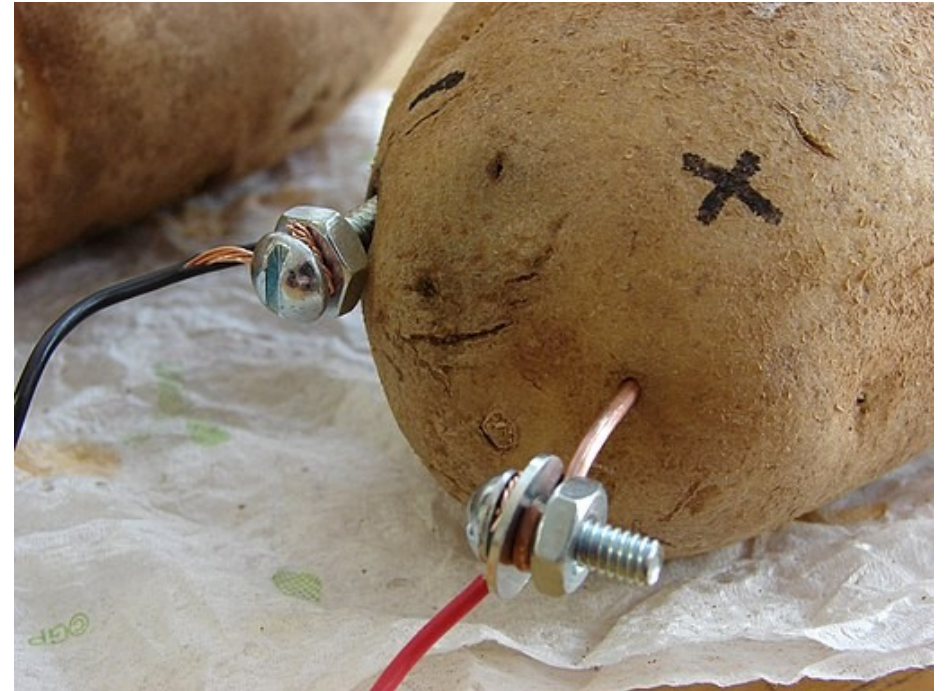
0000-0002-7981-8504

- <https://fusion.cs.uni-jena.de/>
- <https://w3id.org/reproduceme/research>
- <https://sheeba-samuel.github.io/>

# Story

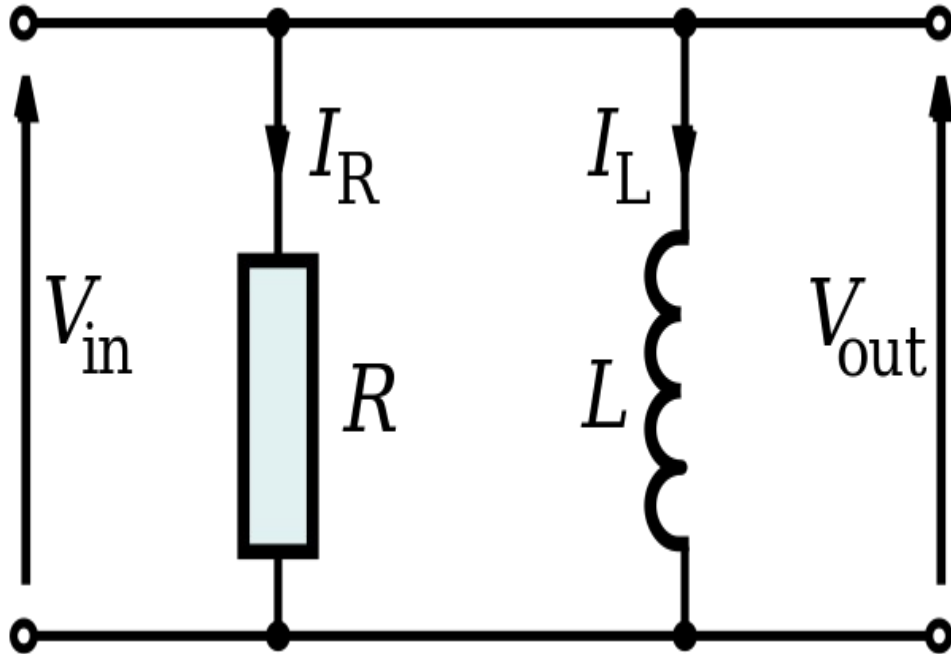


# Story of an Experiment



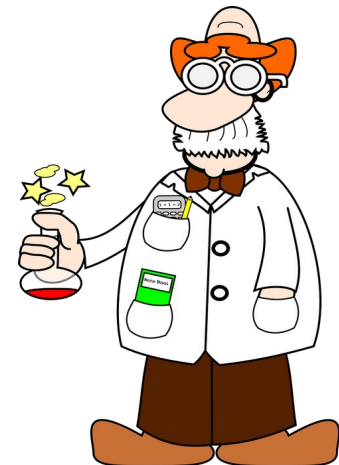
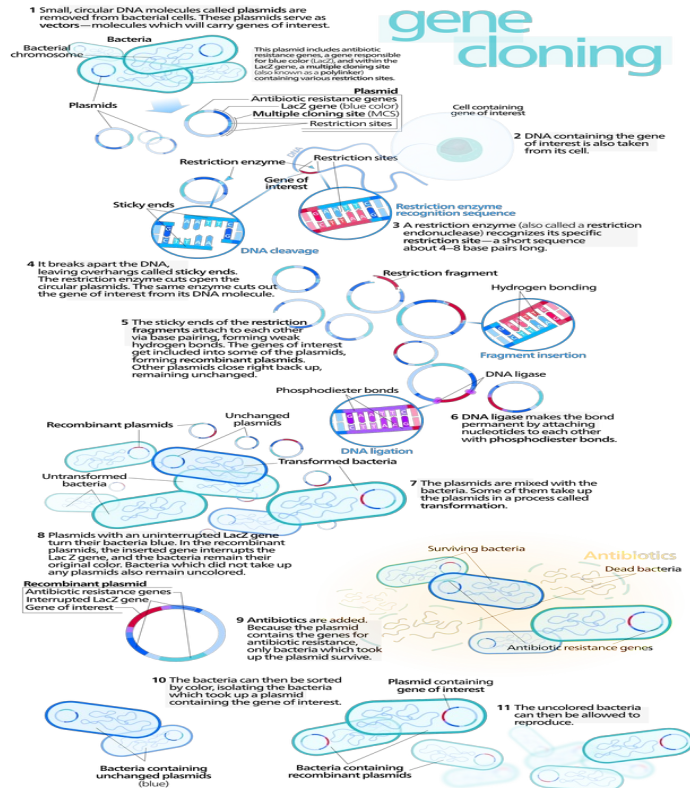
Source: Wikimedia Commons

# Story of an Experiment



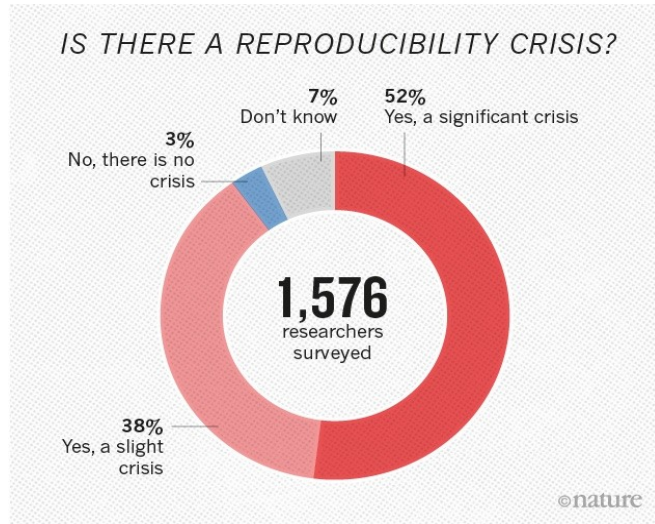
Source: Pixabay

# Story of an Experiment



# Introduction

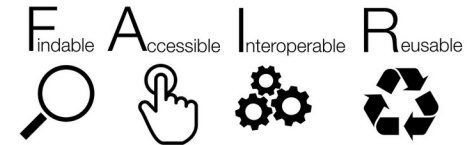
## Reproducibility Crisis\*



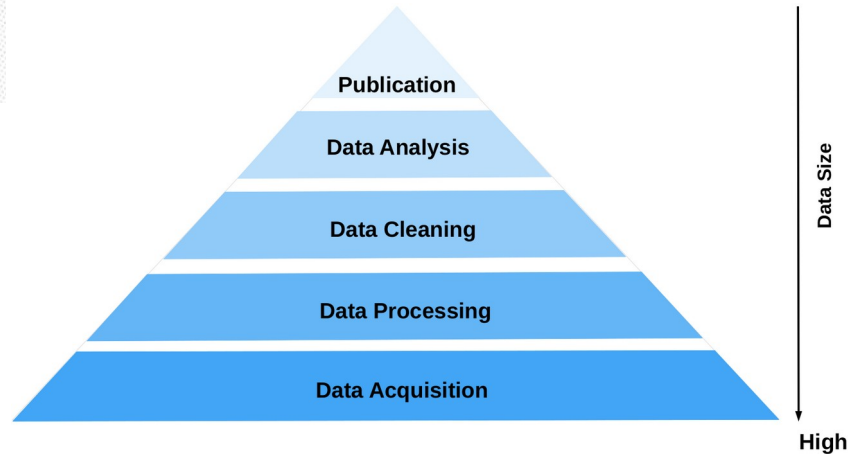
Source: 1,500 scientists lift the lid on reproducibility

## Reproducibility Measures

“Authors are required to make materials, data, code, and associated protocols promptly available to readers without undue qualifications.” - **nature**research

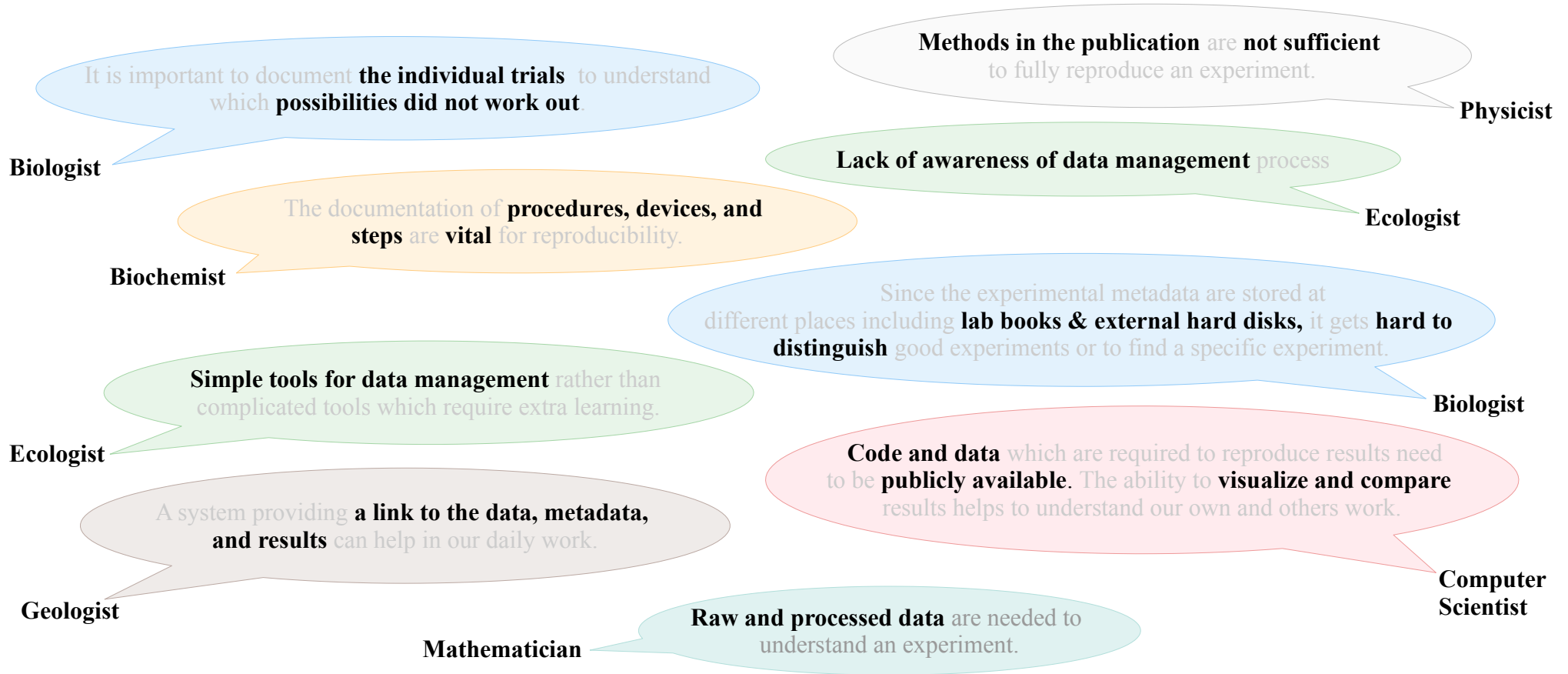


Source: Wikimedia Commons



\* [Kaiser 2015, Peng 2015, Begley and Ioannidis 2015, Baker 2016, Hutson 2018]

# Insights from Interviews



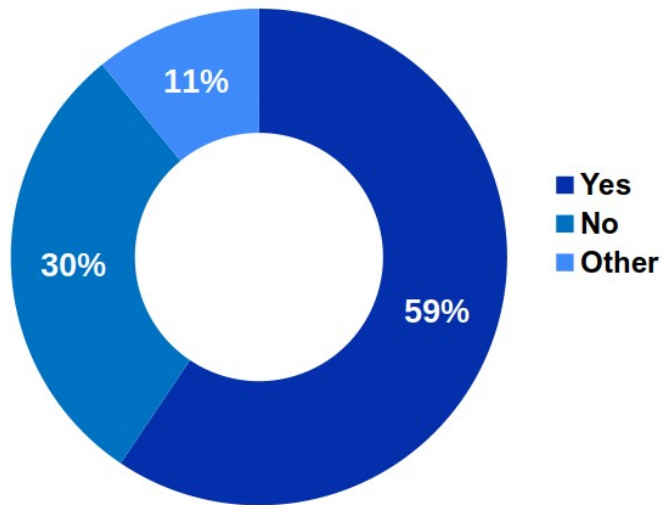


# Survey on experiments and research practices for reproducibility

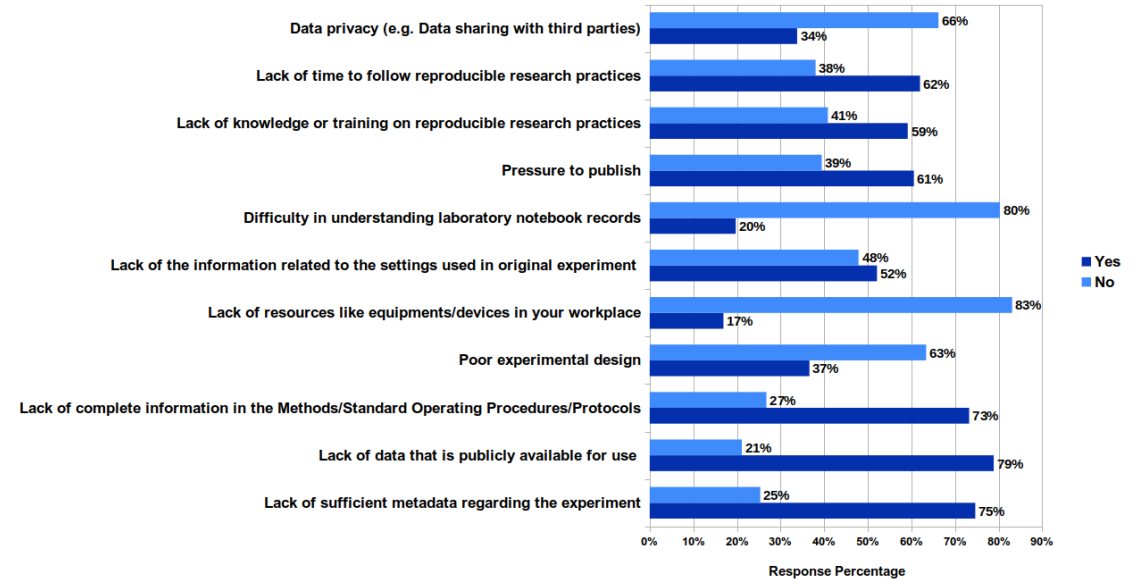
## Findings

- More than half (59%) of the participants think that there is a reproducibility crisis
- 54% of the participants had trouble reproducing other's published results

Do you think there is a reproducibility crisis in your field of research?

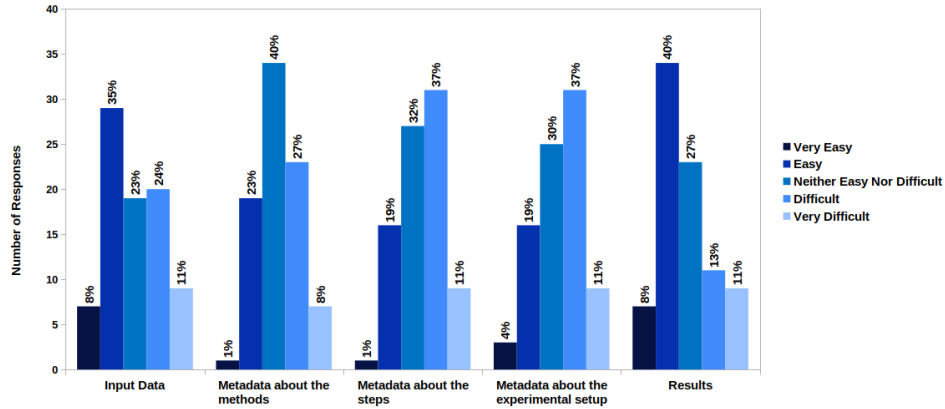


In your experience, what are the factors leading to poor reproducibility?

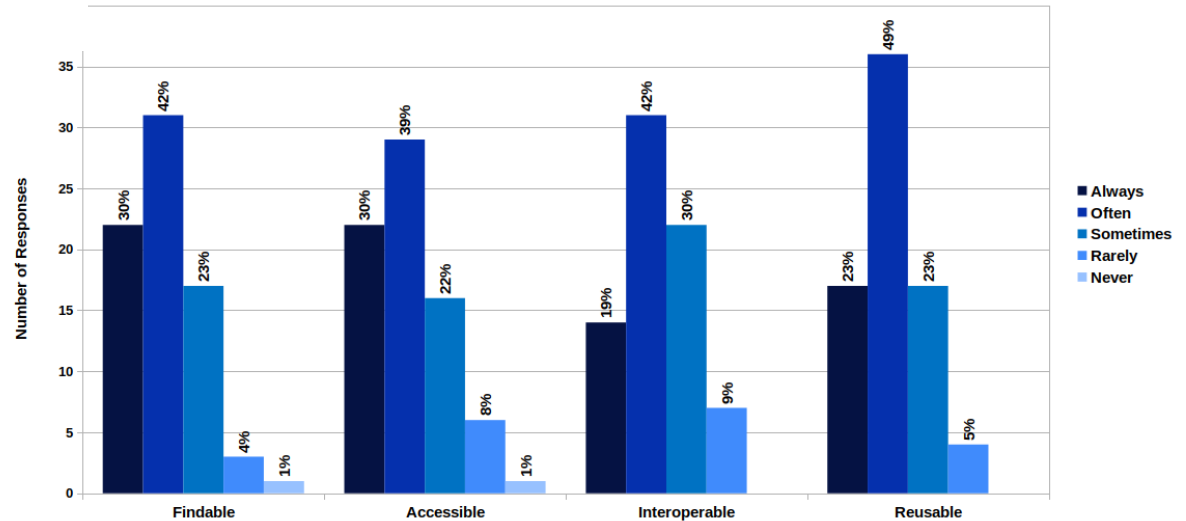


# FAIR data principles

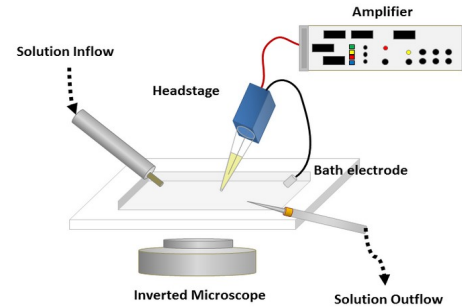
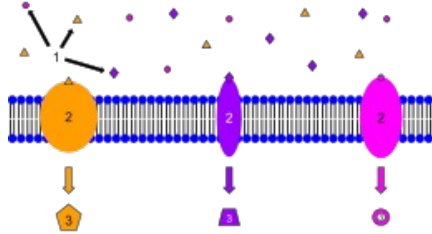
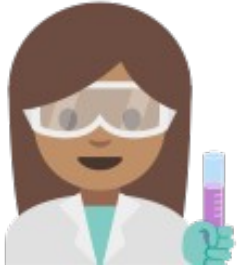
How easy would it be for a newcomer in your workplace to find all the experimental data related to your project/experiment without any/limited instructions from you?



Does your research follow the FAIR (Findable, Accessible, Interoperable, Reusable) principles?



# Scientific Experiments



**STEP 1** → **STEP 2** → **STEP 3**

**Patch-Clamp Fluorometry: Electrophysiology meets Fluorescence**

**Jana Klusák<sup>1</sup> and Giovanni Zampieri<sup>1</sup>**  
<sup>1</sup>Universitätsklinikum Köln, Institut für Physiologie I, Jena, Germany and <sup>2</sup>Natura 2000, Consiglio Nazionale delle Ricerche, Genoa, Italy

**ABSTRACT:** Ion channels and transporters are membrane proteins whose functions are driven by conformational changes. Classical biophysical techniques provide insight into either the structure or the function of these proteins, but a full understanding of their function requires a combination of both aspects to fully grasp subtle and subtle conformational changes and their electrogenic and electrophysiological consequences to simultaneously detect conformational changes and ionic currents across the membrane. Since its introduction, patch-clamp fluorometry has been responsible for numerous advances in our knowledge of ion channel biophysics. Over the years, the technique has been applied to many different ion channel families to address several biological questions with a variety of spectroscopic approaches and electrophysiological configurations. This review illustrates the strength and the flexibility of patch-clamp fluorometry, demonstrating its potential as a tool for future research.

**INTRODUCTION**

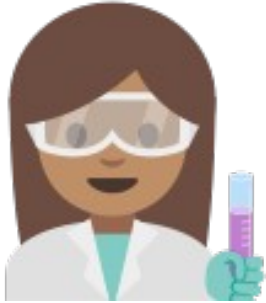
Ion channels are pore-forming membrane proteins, which allow ions to pass across the membrane. They are involved in many physiological processes such as the control of resting and action potentials, nerve transmission, excitation, and muscle contraction. Their activity is regulated by specific stimuli, including voltage, ligand binding, temperature, or mechanical stress. These stimuli cause conformational rearrangements that result in various gating

groups of cytosolic residues located at defined locations within the protein. Conformational changes in the region around the fluorophore lead to changes of its environment and thereby to changes of its emission spectra (1). In addition, VCF provides structural information (2) that is not available from other biophysical techniques. It does enable parallel recording of conformational changes and ion transport activity in a non-optimal membrane environment and thus provides the possibility

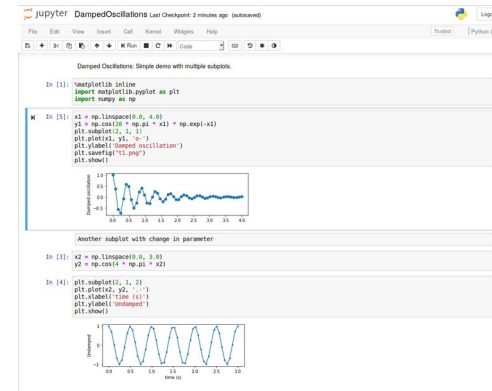
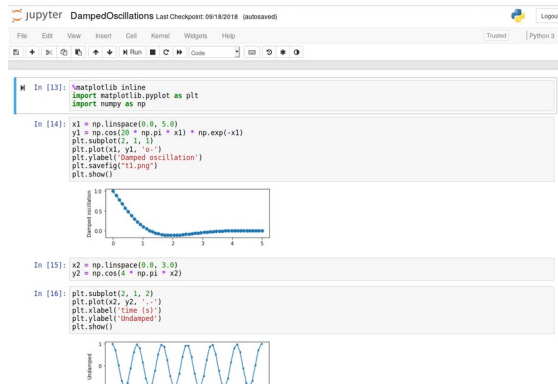
40°C

# Computational Experiments

## ➤ Repeatability

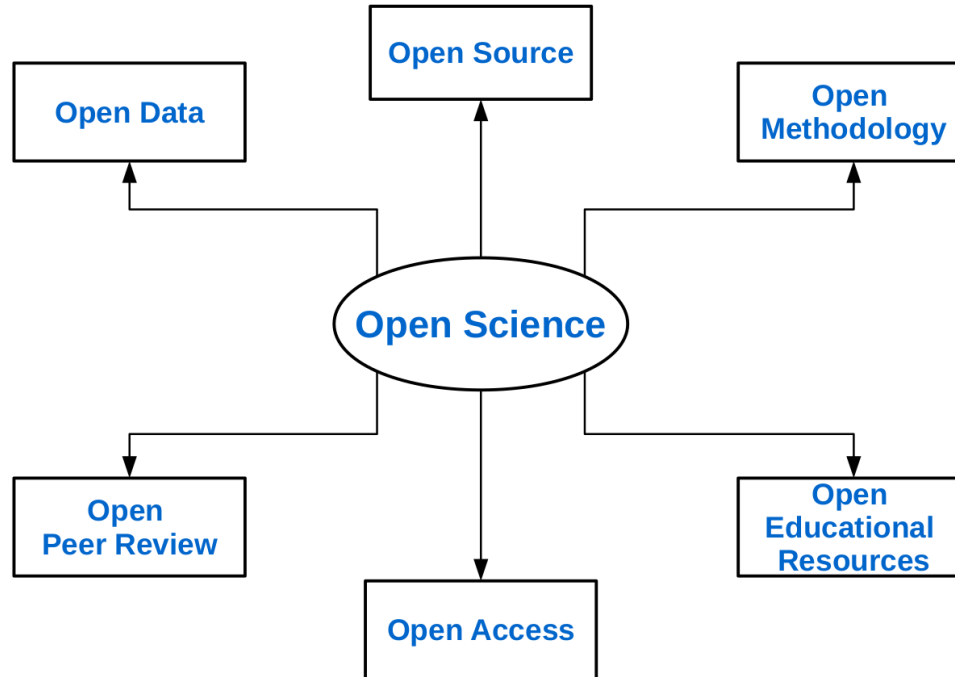


## ➤ Reproducibility



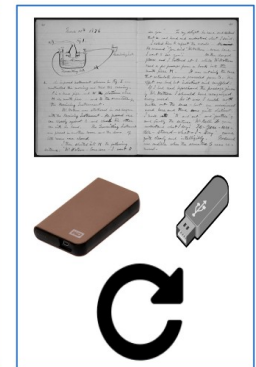
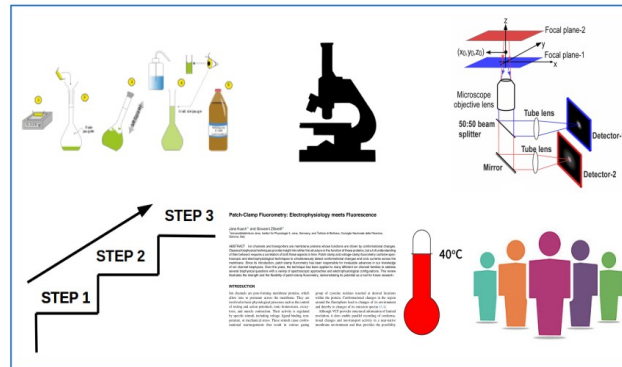
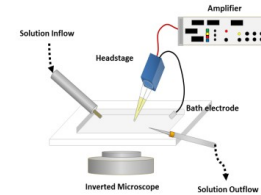
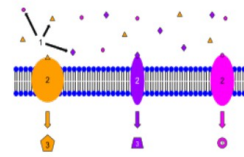
# Open Science

- Open science is the movement to make scientific research available to any member in society.



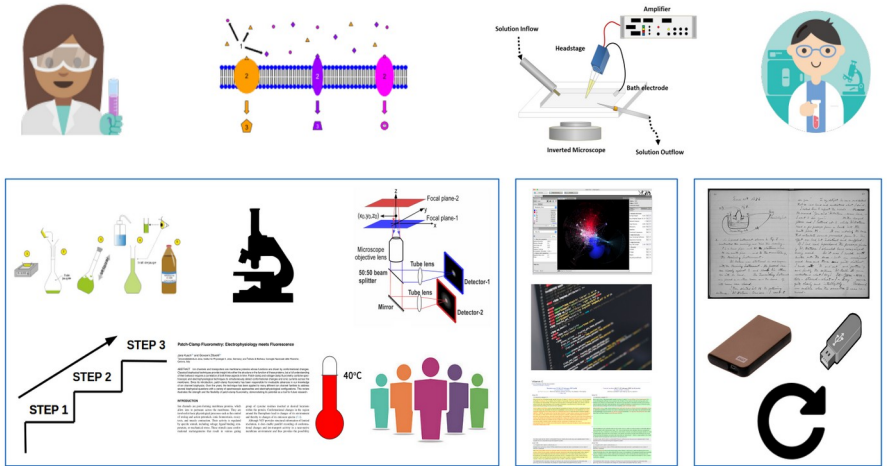
# Open Reproducible Scientific Experiments

- Open Data
- Open Source
- Open Publications
- Open Experiment Materials
- Open Hardware
- Open Instruments
- Open Notebook Science
- Open Methodology
- Open Steps
- Open Data Standards
- Open Workflows
- Open File Formats



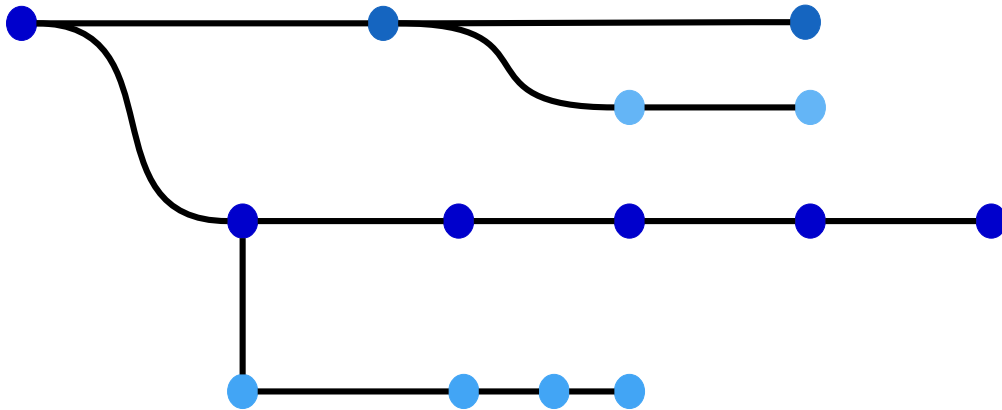
# Challenges: Reproducibility

- Proprietary File Formats
- Availability of hardware
- Proprietary Data Standards
- Unavailability of Experiment Materials
- Not enough description in methodology
- Missing steps



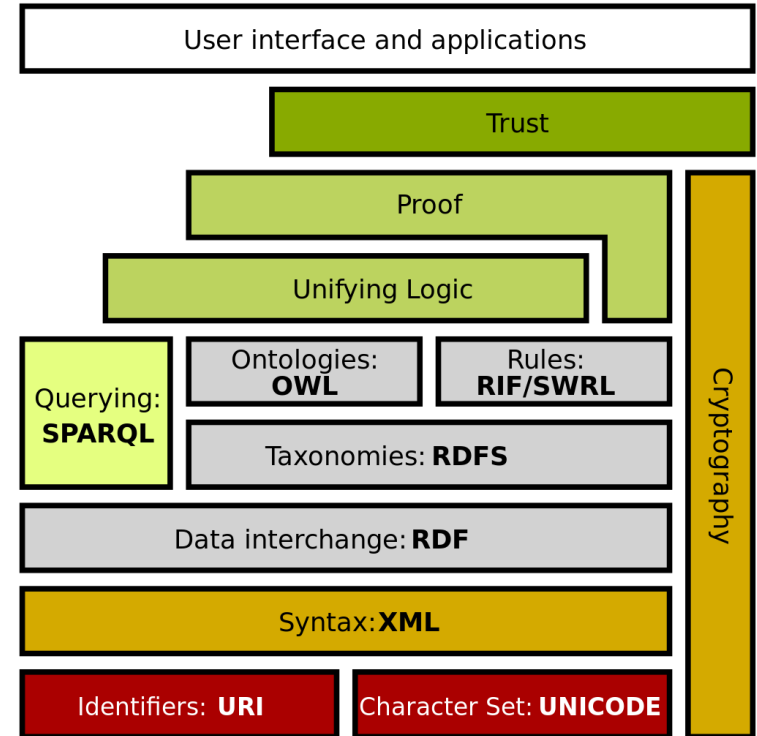
# Provenance

The source or origin of an object; its history



# Semantic Web

Machine Understandable





# Open Science Contributions for Supporting Reproducibility

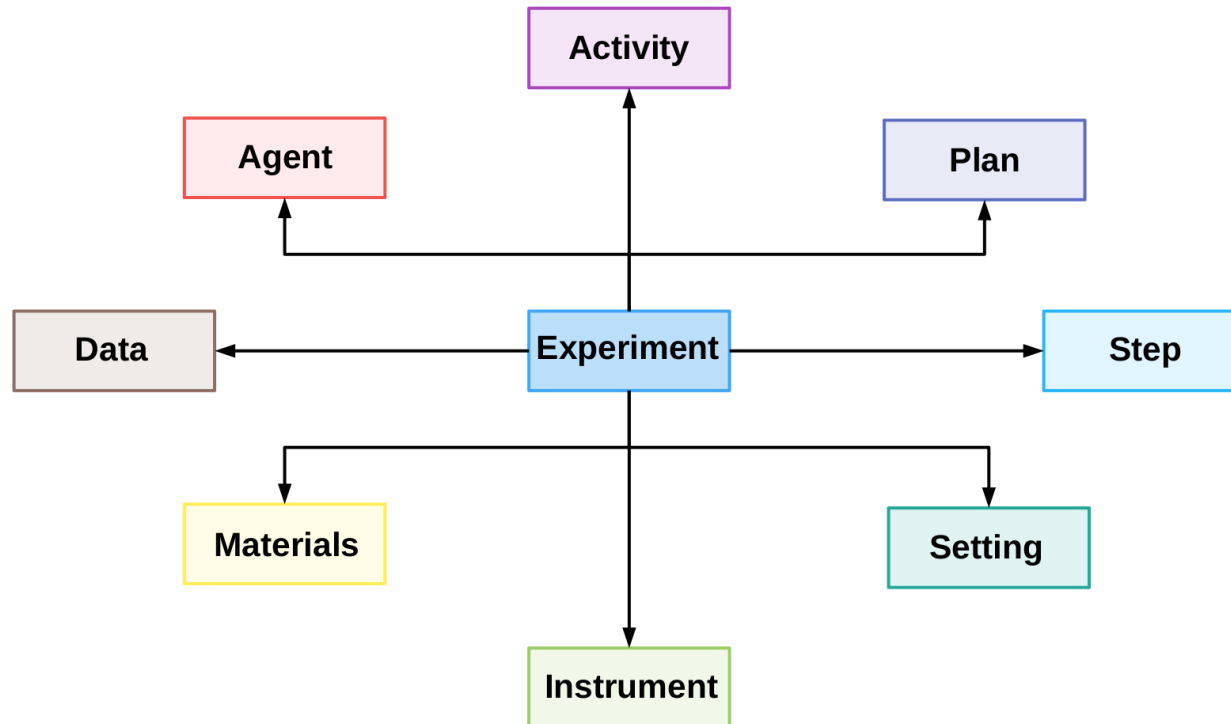
<p>1 Semantic Description of Scientific Experiments</p>	<ul style="list-style-type: none"><li>➤ <b>The REPRODUCE-ME Data Model &amp; Ontology</b> Semantic Description of Scientific experiments including Scripts and Computational Notebooks</li><li>➤ <b>The ReproduceMeON</b> An Ontology Network for Reproducibility of Scientific Studies</li></ul>
<p>2 Support of computational reproducibility</p>	<ul style="list-style-type: none"><li>➤ <b>ProvBook</b> Provenance capture, visualize, represent and difference of results</li><li>➤ <b>MLProvLab</b> Provenance capture, visualize, represent and difference of ML notebooks</li><li>➤ <b>ReproduceMeGit</b> A tool for analyzing the reproducibility of Jupyter Notebooks</li></ul>
<p>3 End-to-end provenance management of scientific experiments</p>	<ul style="list-style-type: none"><li>➤ <b>CAESAR</b> Support of a collaborative environment with visualization of the complete path of a scientific experiment</li></ul>

# The REPRODUCE-ME Data Model & Ontology

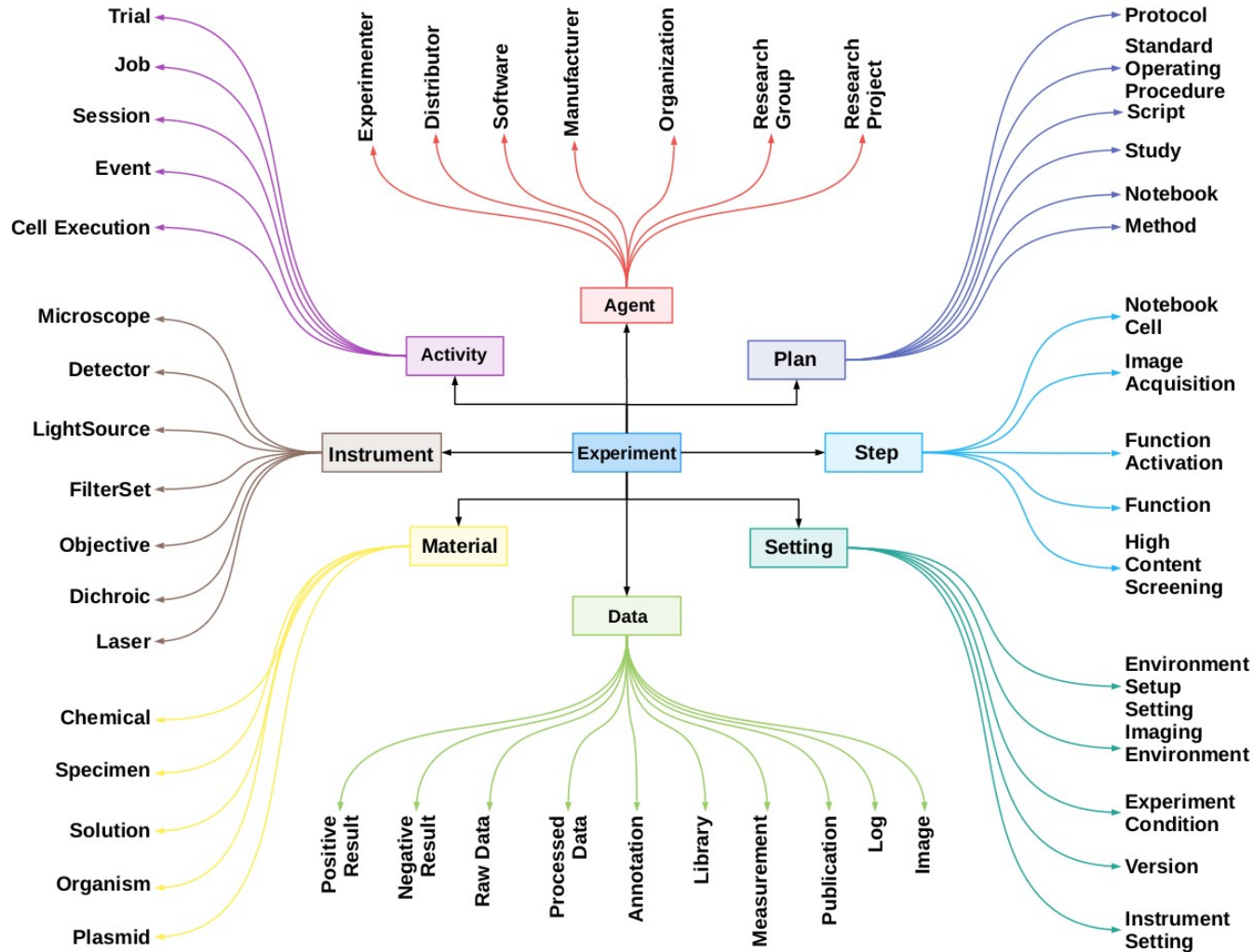
# The REPRODUCE-ME Data Model

Experiment is defined as an n-tuple

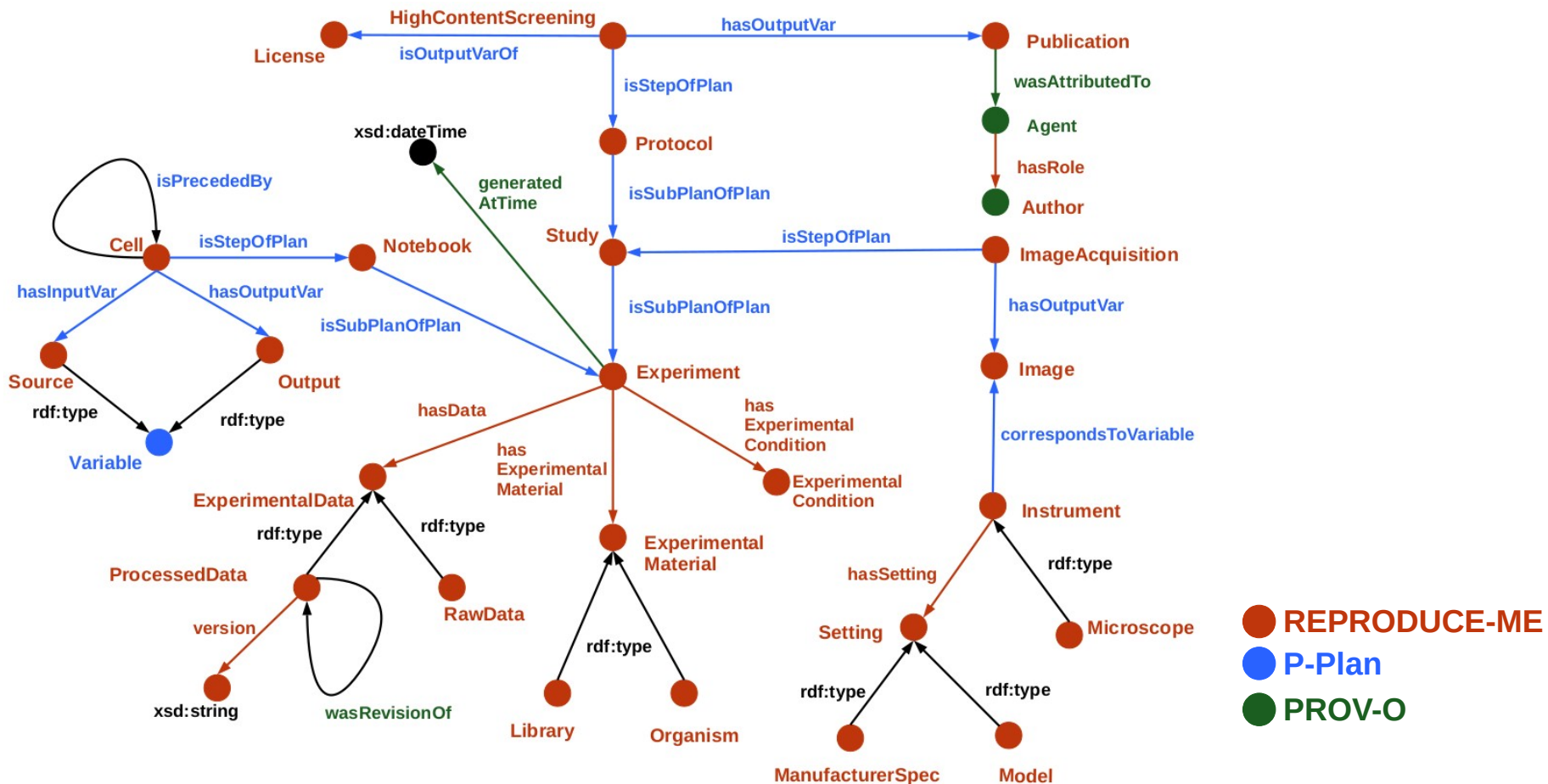
**$E = (\text{Data, Agent, Activity, Plan, Step, Setting, Instrument, Material})$**



# The REPRODUCE-ME Data Model



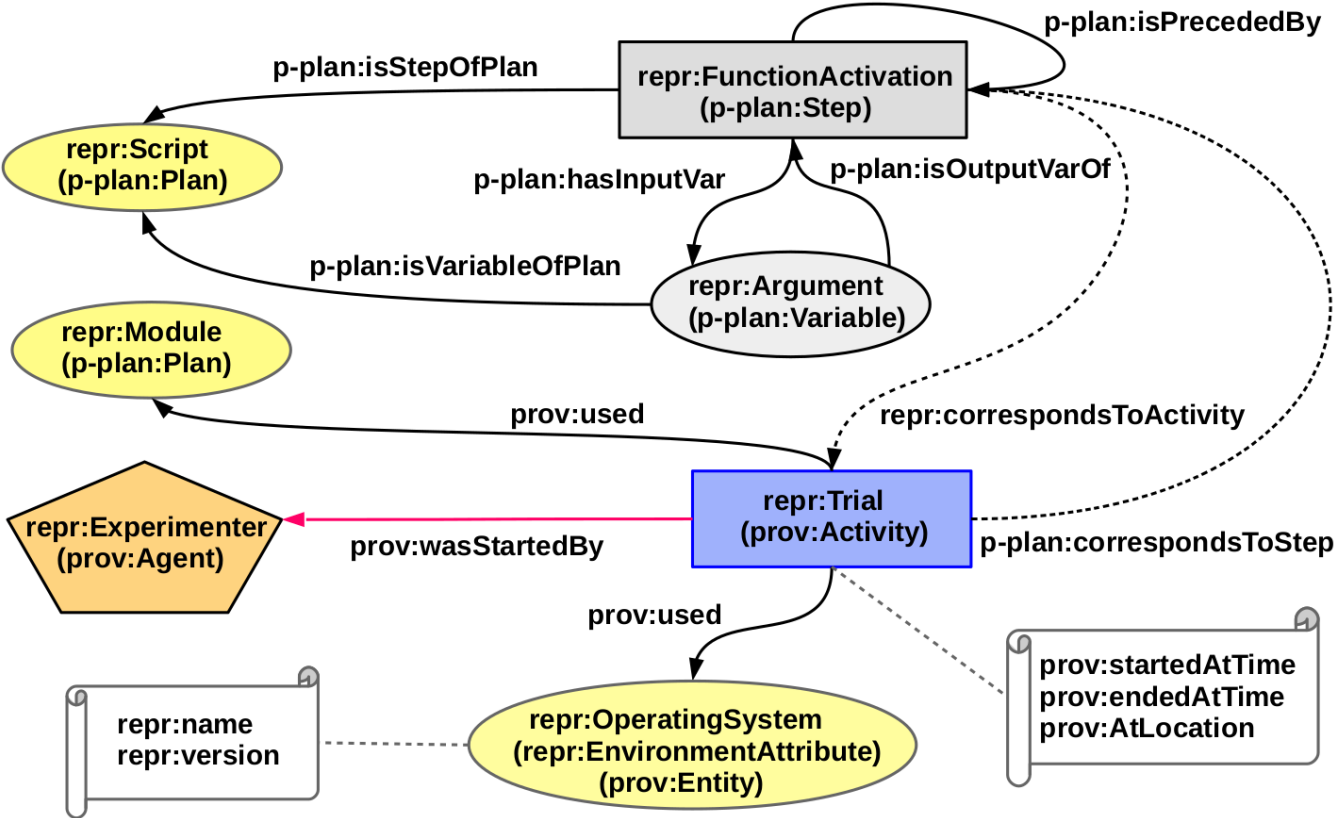
# The REPRODUCE-ME Ontology



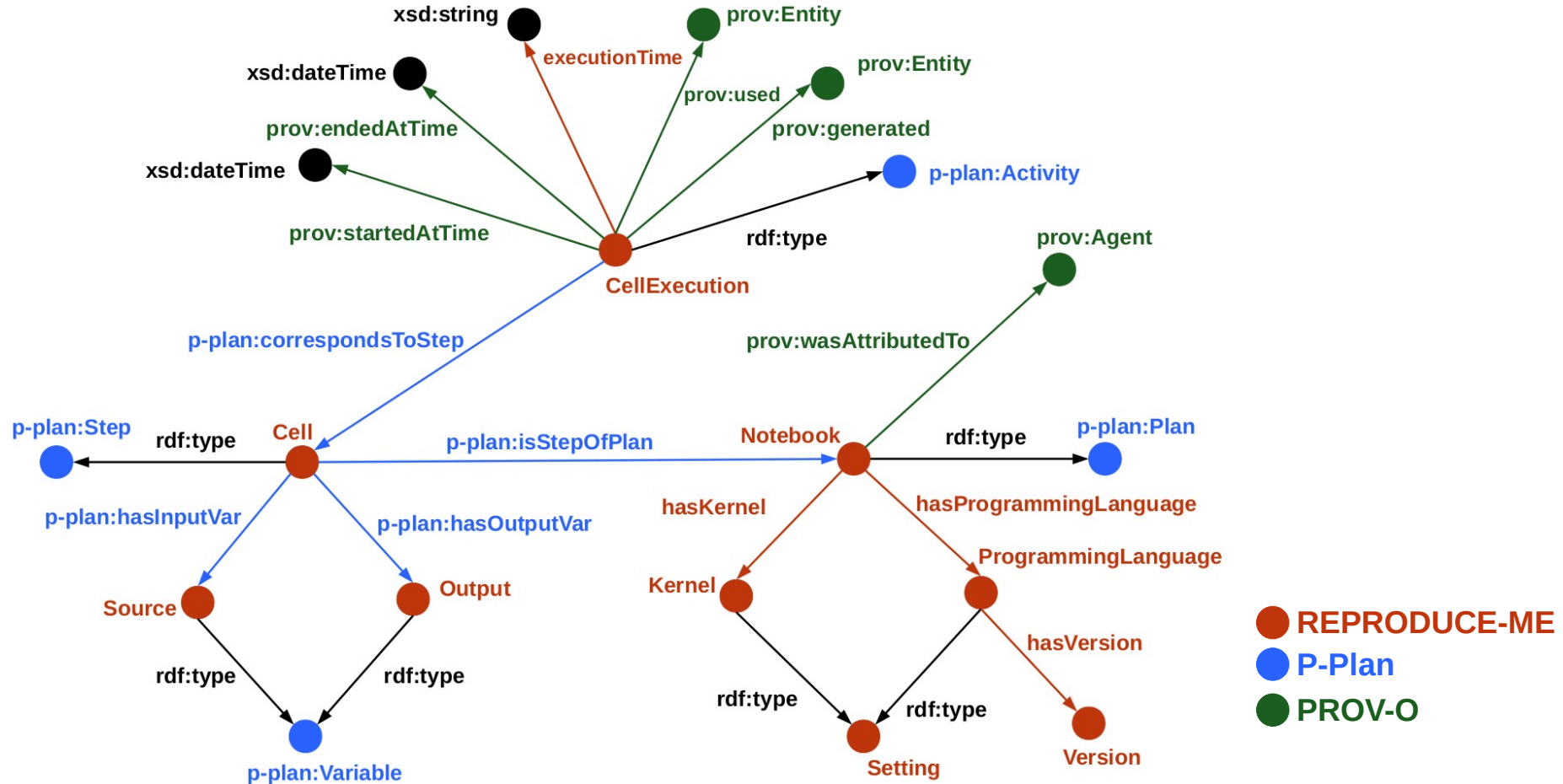
[Samuel and König-Ries 2017, Samuel and König-Ries 2018a, Samuel et al. 2018]

# Script Provenance

Ontology term	Remarks
<i>repr:Script</i>	<i>p-plan:Plan</i>
<i>repr:Function</i>	<i>p-plan:Plan</i>
<i>repr:Module</i>	<i>p-plan:Plan</i>
<i>repr:Version</i>	<i>repr:Setting</i>
<i>repr:Argument</i>	<i>p-plan:Variable</i>
<i>repr:Input</i>	<i>p-plan:Variable</i>
<i>repr:Output</i>	<i>p-plan:Variable</i>
<i>repr:FunctionActivation</i>	<i>p-plan:Step</i>
<i>repr:Trial</i>	<i>p-plan:Activity</i>
<i>repr:OperatingSystem</i>	<i>p-plan:Setting</i>
<i>repr:Author</i>	<i>prov:Person</i>
<i>prov:startedAtTime</i>	Data property
<i>prov:endedAtTime</i>	Data property
<i>p-plan:isPrecededBy</i>	Object property



# Computational Provenance

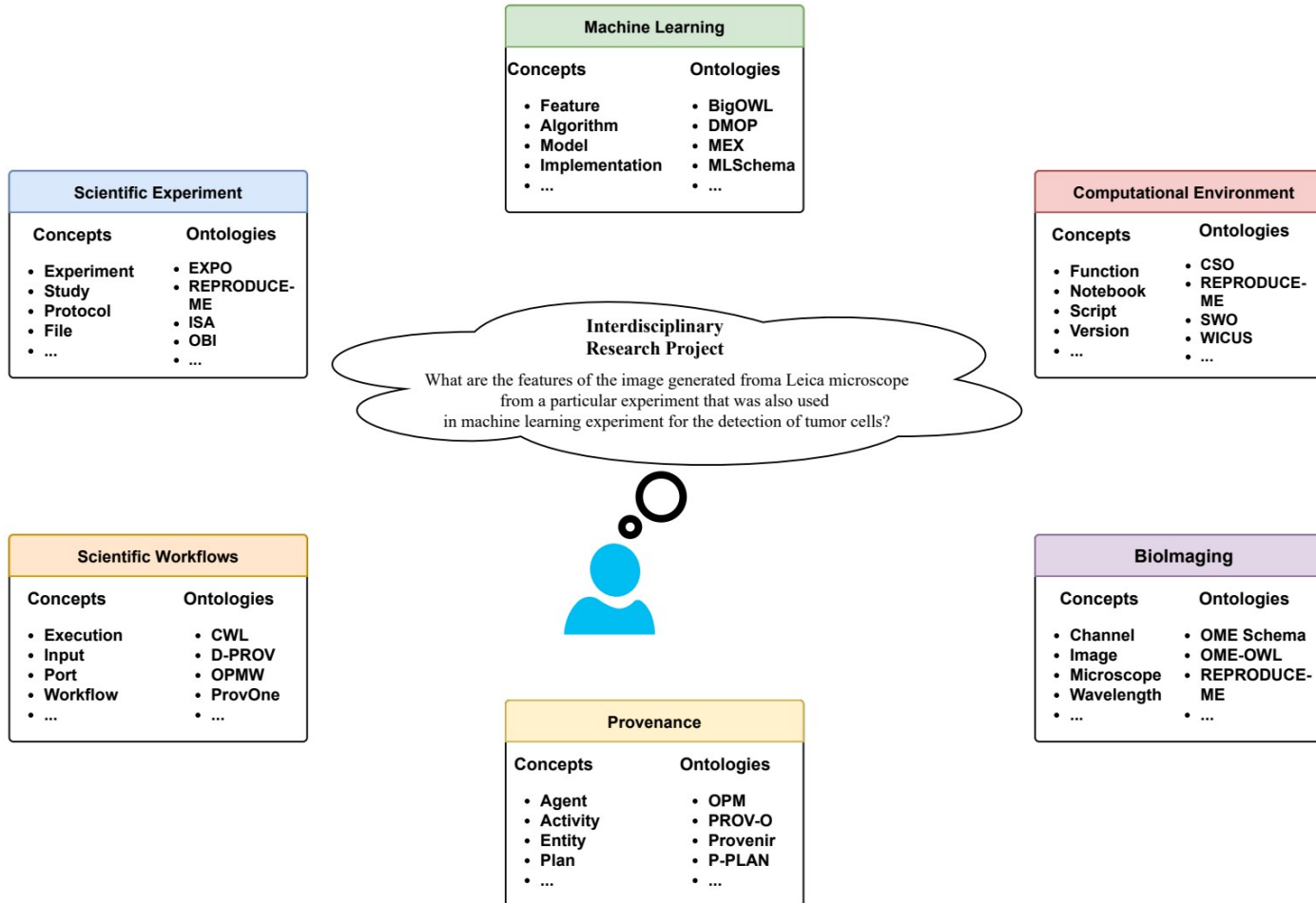


[Samuel and König-Ries 2018a, Samuel and König-Ries 2018b]

# The ReproduceMe Ontology Network (ReproduceMeON)

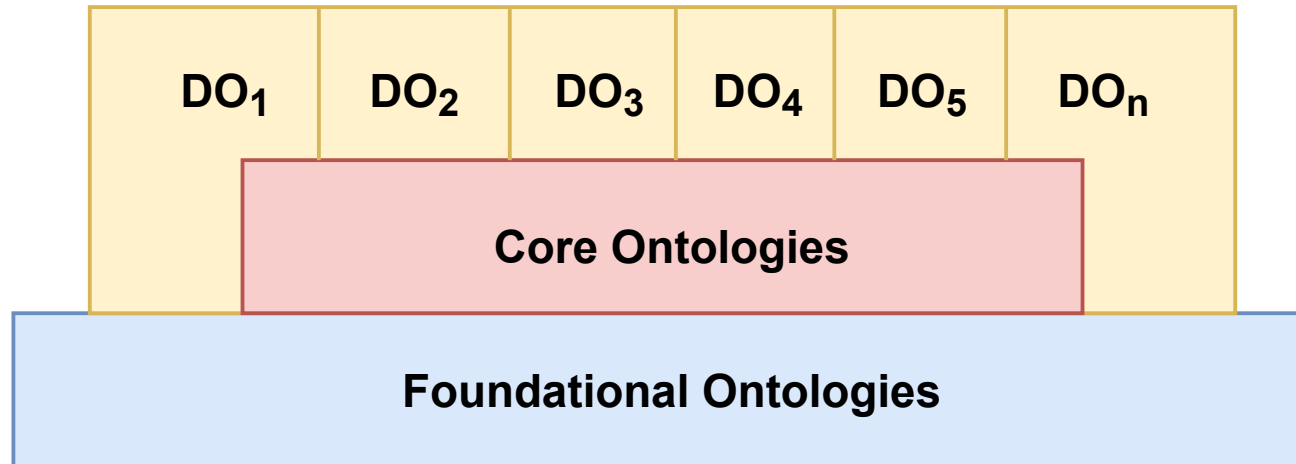


# Reproducibility related area assimilation



# ReproduceMeON

- A novel approach that brings together knowledge from several domains, such as ML, provenance, and scientific computing, based on the three-layered architecture



# Computational Reproducibility

# Computational Notebooks

- Share code along with documentation
  - Project Jupyter
  - RStudio

## Jupyter Notebook Facts

- Formerly known as *IPython* Notebook
- 1.7 million Jupyter notebooks on Github
- Millions of *users*
- Different *computational kernels* including Python, R, and MATLAB
- Export in different *formats* like HTML, LaTeX, PDF

## Structure of a Jupyter Notebook

The screenshot shows a Jupyter Notebook interface with the following components and labels:

- Markdown Cell:** A cell containing the text "Damped Oscillations: Simple demo with multiple subplots."
- Code Cell:** A cell containing Python code for plotting damped oscillations:

```
In [1]: %matplotlib inline
import matplotlib.pyplot as plt
import numpy as np

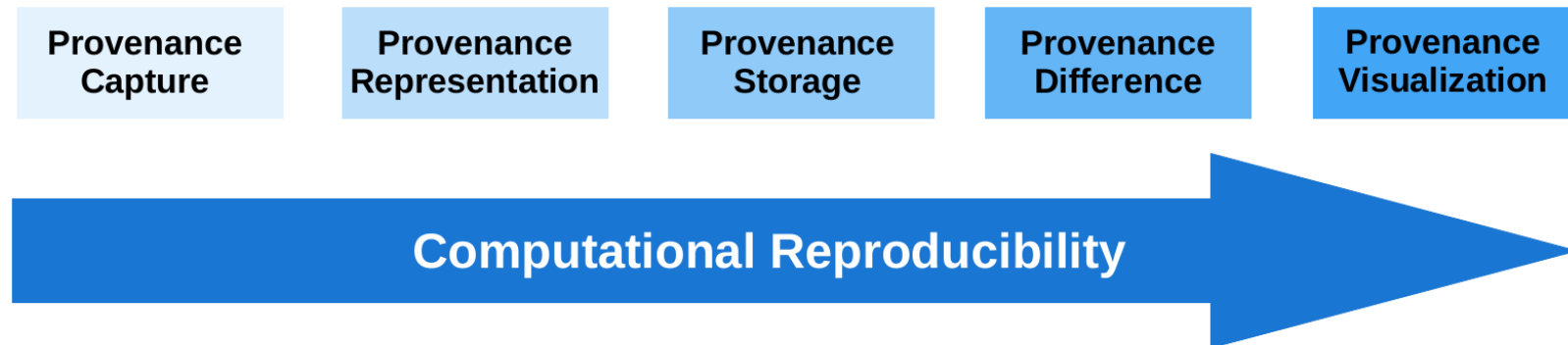
In [5]: x1 = np.linspace(0.0, 4.0)
y1 = np.cos(20 * np.pi * x1) * np.exp(-x1)
plt.subplot(2, 1, 1)
plt.plot(x1, y1, 'o-')
plt.ylabel('Damped oscillation')
plt.savefig("t1.png")
plt.show()
```
- Output:** A plot showing a damped oscillation with the y-axis labeled "Damped oscillation" and the x-axis ranging from 0.0 to 4.0.
- Raw Cell:** A cell containing Python code for plotting an undamped oscillation:

```
In [3]: x2 = np.linspace(0.0, 3.0)
y2 = np.cos(4 * np.pi * x2)

In [4]: plt.subplot(2, 1, 2)
plt.plot(x2, y2, '-.-')
plt.xlabel('time (s)')
plt.ylabel('Undamped')
plt.show()
```
- Execution Count:** A plot showing an undamped oscillation with the y-axis labeled "Undamped" and the x-axis labeled "time (s)" ranging from 0.0 to 3.0.

# Computational Reproducibility

- **Provenance support is limited** [Rule et al., 2018, Pimentel et al., 2019]
  - Tracking provenance when the cells are over-written and re-run
  - Track how exactly a final result has been achieved
  - Track of the experiments that have been attempted
- **“Record all intermediate results in a standardized format”**
  - One of the ten simple rules for computational reproducible research [Sandve et al., 2013]



The key components for the end-to-end provenance management for computational reproducibility

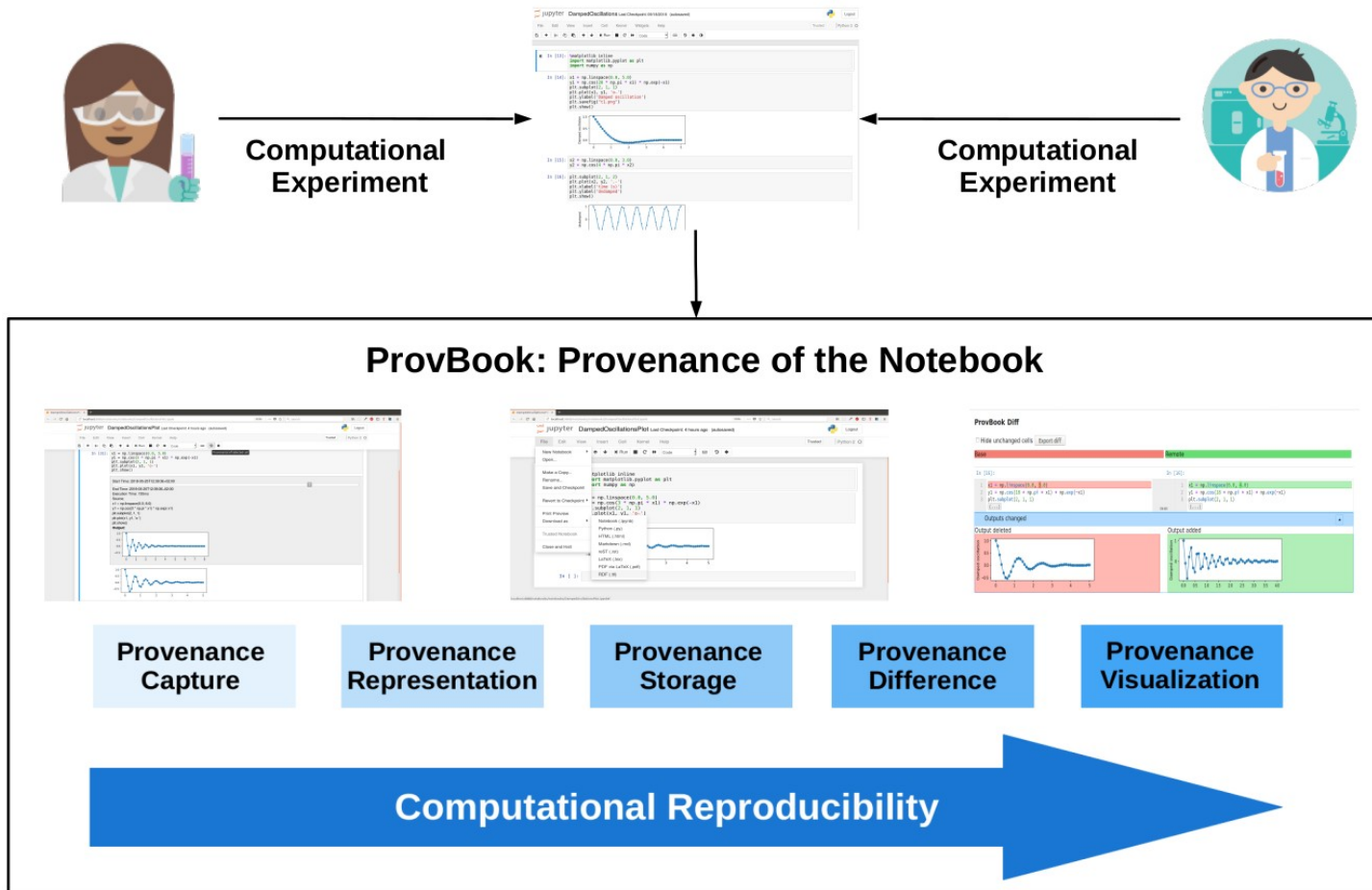
# Design Goals

- Support the provenance lifecycle
  - Tracking
  - Storage
  - Querying
  - Compare
  - Visualization

# Design Goals

- Support
  - Reproducibility
  - Collaboration
  - Semantic annotation and interoperability
  - Exporting provenance in different formats
  - Extensibility
- Ease of use

# ProvBook: Provenance of the Notebook





# ProvBook: Capture & Visualization

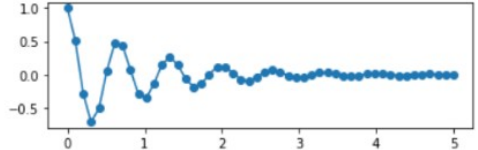
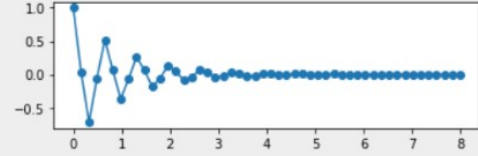
Jupyter DampedOscillationsPlot Last Checkpoint: 4 hours ago (autosaved) Python 2 Logout

File Edit View Insert Cell Kernel Help Trusted

```
In [31]: x1 = np.linspace(0.0, 5.0)
y1 = np.cos(3 * np.pi * x1) * np.exp(-x1)
plt.subplot(2, 1, 1)
plt.plot(x1, y1, 'o-')
plt.show()
```

Start Time: 2018-05-25T12:39:36+02:00  
End Time: 2018-05-25T12:39:36+02:00  
Execution Time: 135ms  
Source:  
x1 = np.linspace(0.0, 8.0)  
y1 = np.cos(3 \* np.pi \* x1) \* np.exp(-x1)  
plt.subplot(2, 1, 1)  
plt.plot(x1, y1, 'o-')  
plt.show()

Output:



Start Time  
End Time  
Execution Time  
Source  
Output

# ProvBook: Difference

- A provenance difference module to compare the different executions of a notebook
- Comparison of the input and the output
- Starting time to differentiate between two executions
- Extends the nbdime library from the Project Jupyter



## ProvBook Diff

Hide unchanged cells

Base

Remote

In [16]:

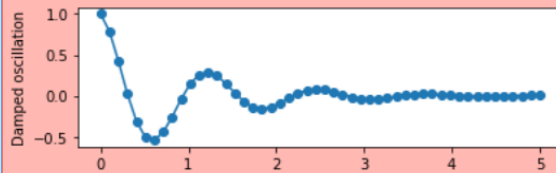
```
1 x1 = np.linspace(0.0, 5.0)
2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)
3 plt.subplot(2, 1, 1)
  (...)
```

In [16]:

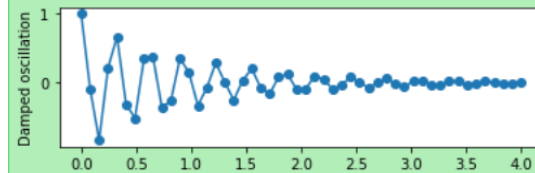
```
1 x1 = np.linspace(0.0, 4.0)
2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)
3 plt.subplot(2, 1, 1)
  (...)
```

Outputs changed

Output deleted

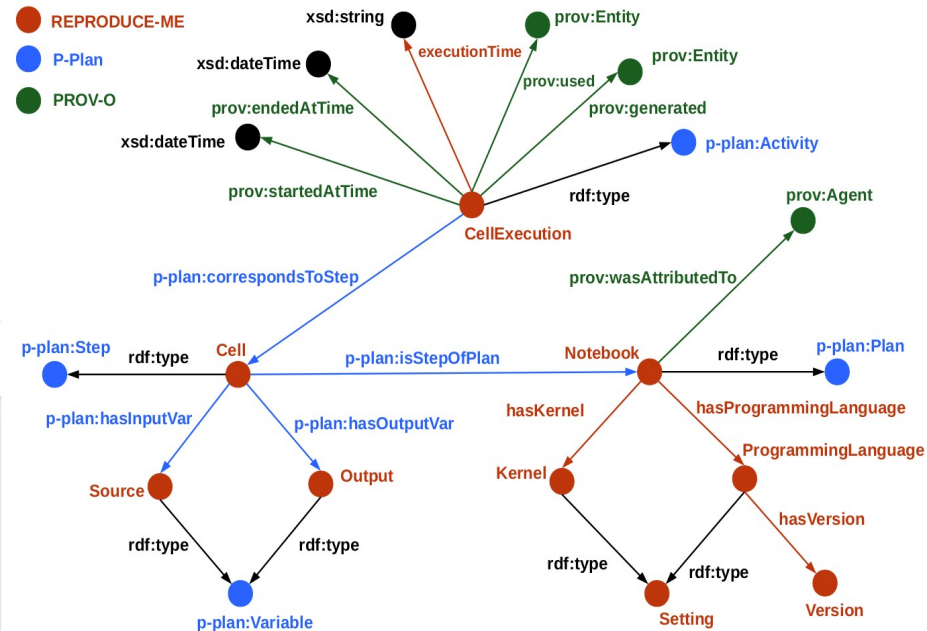
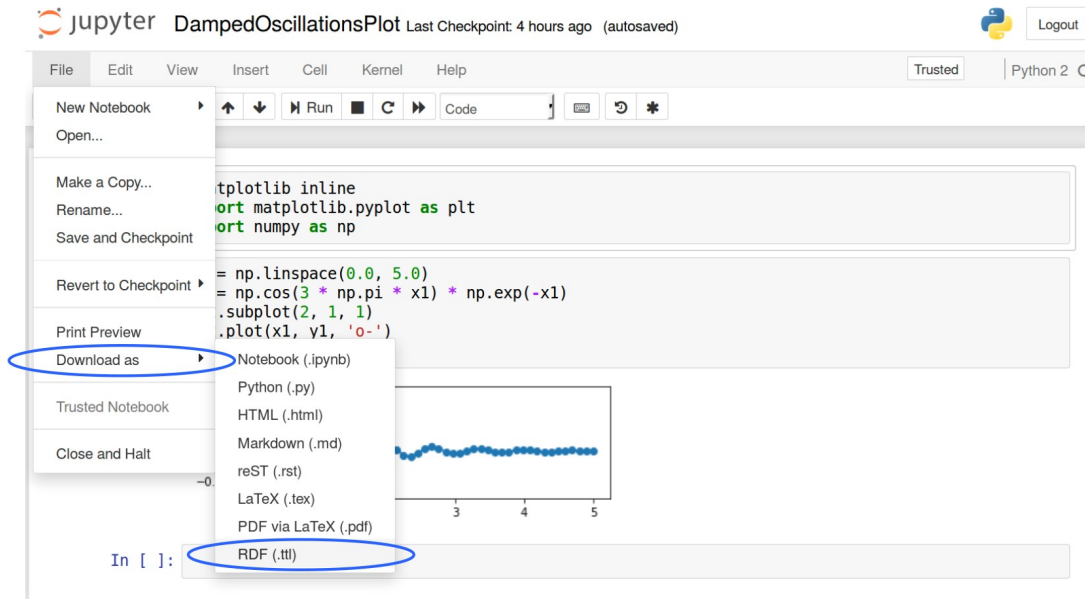


Output added



# ProvBook: Semantic Representation

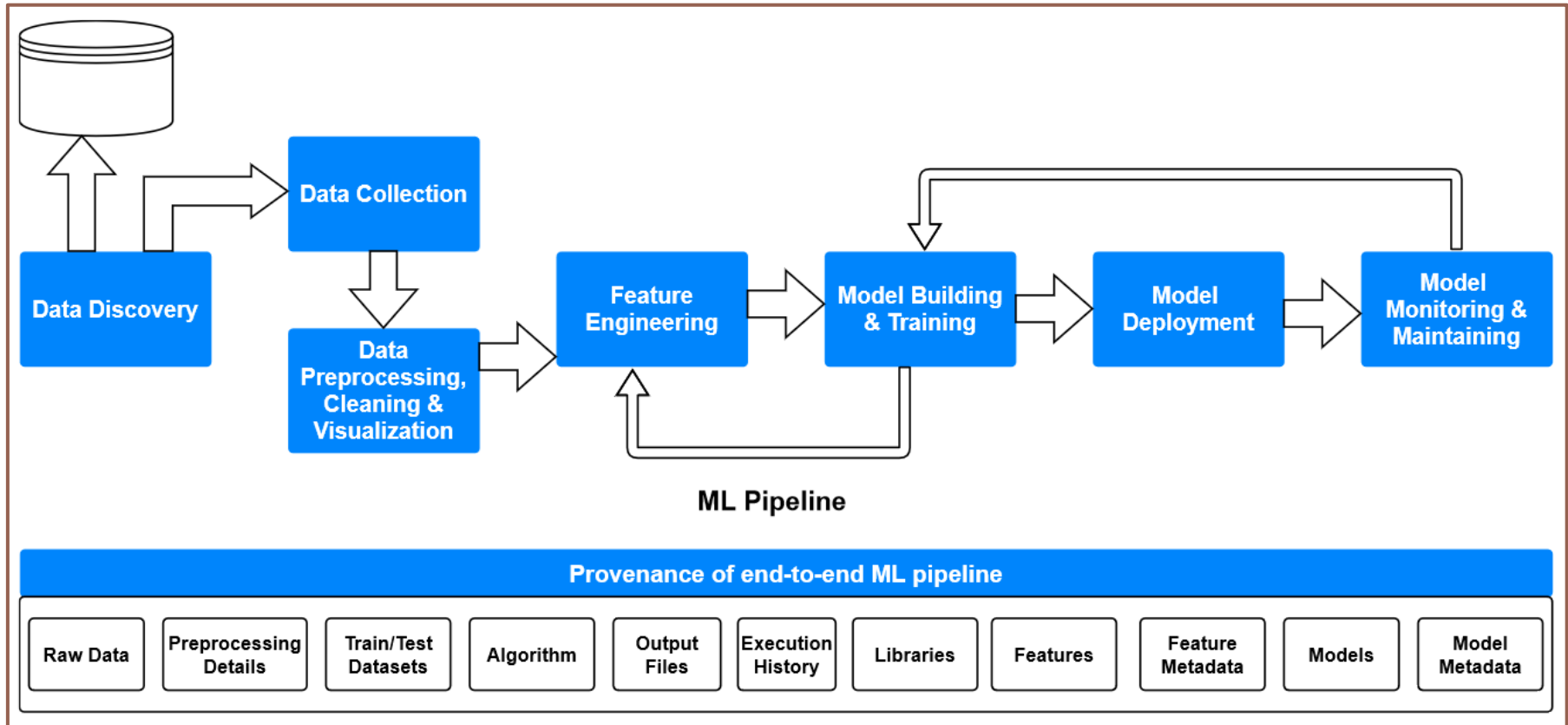
- Jupyter Notebook and its provenance described using the REPRODUCE-ME ontology
- **New extension:** Export in RDF from the user interface or command line



[Samuel and König-Ries 2018a]

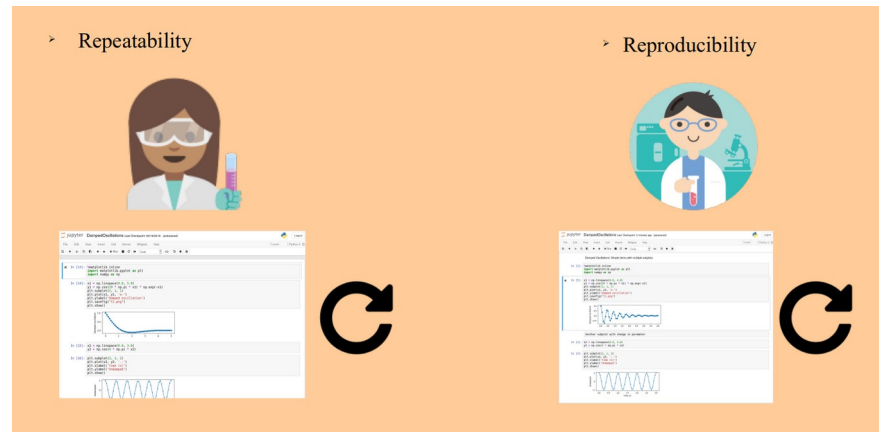


# Provenance of ML Pipelines



# MLProvLab

- Provenance management of end-to-end ML pipelines in a notebook environment
- An extension to JupyterLab
- This tool tracks, compares, manages, and visualizes provenance information
- Track, at runtime, the datasets, variables, libraries, and functions used in the notebook and their dependencies between cells.



# Provenance Capture

Dependencies between cells

Imported and Used Modules

Execution Information

Variable, Class, and Function Definitions

Code

Outputs

Function and Object Calls

Program Structure  
(Loops, Conditions)

```
[1]: import numpy as np
      from pandas import read_csv
      from sklearn.model_selection import train_test_split
      from sklearn.metrics import classification_report
      from sklearn.metrics import confusion_matrix
      from sklearn.metrics import accuracy_score
      from sklearn.svm import SVC

[2]: # Load dataset
      url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/iris.csv"
      names = ['sepal-length', 'sepal-width', 'petal-length', 'petal-width', 'class']
      dataset = read_csv(url, names=names)
      print(dataset)
```

# Modules and Libraries

```
evaluation_notebook.ipynb x
```

[5]: `X=df.drop("label",axis=1).values`  
`y=df["label"].values`

[6]: `print(X.shape)`  
`print(y.shape)`  
`print(test.shape)`

(42000, 784)  
(42000,)  
(28000, 784)

[7]: `import matplotlib.pyplot as plt`  
`import seaborn as sns`  
`plt.figure(figsize=(15,10))`  
`sns.set_style("darkgrid")`  
`sns.countplot(x="label",data=df)`

[7]: `<AxesSubplot:xlabel='label', ylabel='count'>`

[8]: `from sklearn.model_selection import train_test_split`  
`X_train, X_test, y_train, y_test = train_test_split(X_train, X_test, y_train, y_test)`  
`print(X_train.shape)`  
`print(X_test.shape)`  
`print(y_train.shape)`  
`print(y_test.shape)`

(39900, 784)  
(2100, 784)

Provenance: evaluation\_noto x

Options | Export | Environment info | Import info | Code info | General info | Help

Info about imports and modules in epoch 4

Imports from module **keras.preprocessing.image**:

Import **ImageDataGenerator** was used 1 times

Imports from module **keras.utils.np\_utils**:

Import **to\_categorical** was used 1 times

Import **matplotlib.pyplot** with alias **plt** was used 3 times

Import **numpy** version 1.19.5 with alias **np** was used 0 times

Import **os** was used 2 times

Import **pandas** version 1.2.3 with alias **pd** was used 3 times

Import **seaborn** version 0.11.1 with alias **sns** was used 2 times

Epoch 4/4

Execution 19/19

Thu, 24 Jun 2021 10:19:30 GMT

Thu, 24 Jun 2021 10:26:21 GMT

# Datasets

Provenance: evaluation\_note ✕

Options

Export

Environment info

Import info

Code info

General info

Help

Source **D:/Projects/mnist-evaluation/data** was first used in execution **1** in variable

Source **D:/Projects/mnist-evaluation/data/train.csv** was first used in execution **2** in variable **df**

Source **D:/Projects/mnist-evaluation/data/test.csv** was first used in execution **3** in variable **test**



# Execution Environment

Provenance: evaluation\_note X

Options

Export

Environment info

Import info

Code info

General info

Help

## Environment information of epoch 3

**Language:** python

**Version:** 3.9.2

**Mimetype:** text/x-python

**Kernel start time:** Thu, 24 Jun 2021 10:17:35 GMT

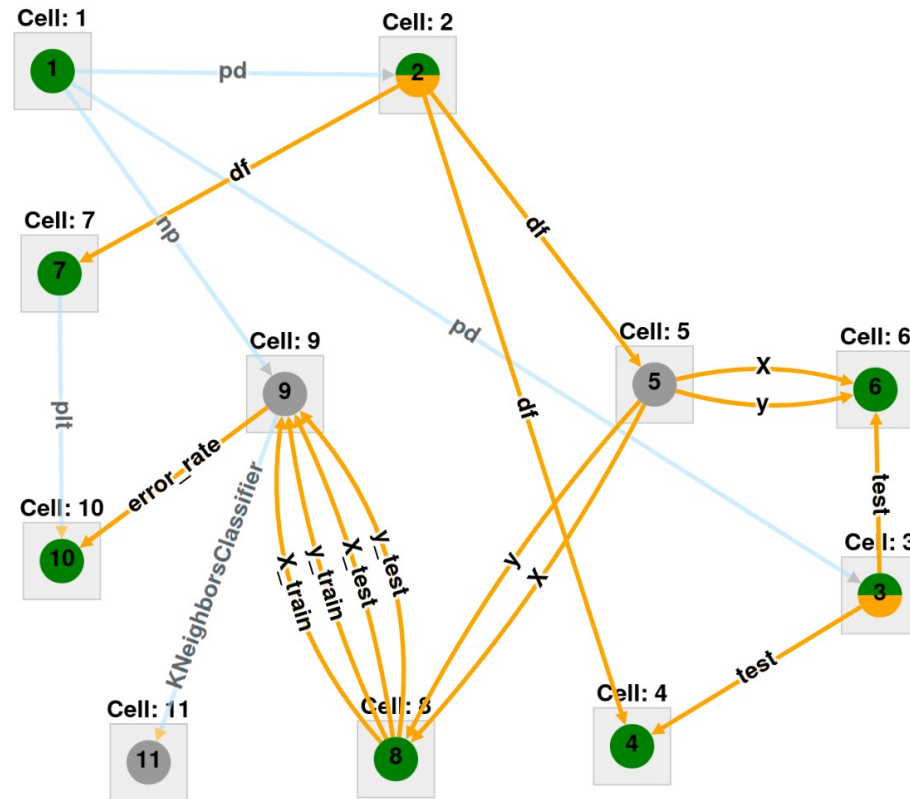
**Kernel implementation:** ipython

**Kernel version:** 7.22.0

**User agent:** Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:89.0) Gecko/20100101 Firefox/89.0

# MLProvLab: Provenance Visualization

Provenance: evaluation_not x						
Options	Export	Environment info	Import info	Code info	General info	Help



Epoch 1/4  Thu, 24 Jun 2021 09:57:28 GMT  
Execution 11/15  Thu, 24 Jun 2021 09:58:31 GMT

# ReproduceMeGit

# ReproduceMeGit

- A visualization tool for analyzing the reproducibility of Jupyter Notebooks.
- Goals:
  - Help repository users and owners to reproduce, directly analyze and assess the reproducibility of notebooks
  - Get information on notebooks
    - that were successfully reproducible
    - that resulted in exceptions during runs
  - Analyze the notebooks:
    - the difference in the results from the original notebooks
    - provenance history of runs

## ReproduceMeGit

GitHubURL

Reproduce

# An Overview

## Reproducibility Study

Notebooks (un-)successfully finishing the executions

Notebooks with same or different results compared to the original.

Exceptions occurred in the runs  
ImportError, ModuleNotFoundError  
FileNotFoundError, IOError, SyntaxError

Provenance History in RDF  
using REPRODUCE-ME ontology

Direct access to  
Binder and ProvBook

**ReproduceMeGit**

## Structure & Usage

Repository Overview

Notebook Overview

Cells Overview

Modules Overview

Distribution of programming  
Languages and the versions used

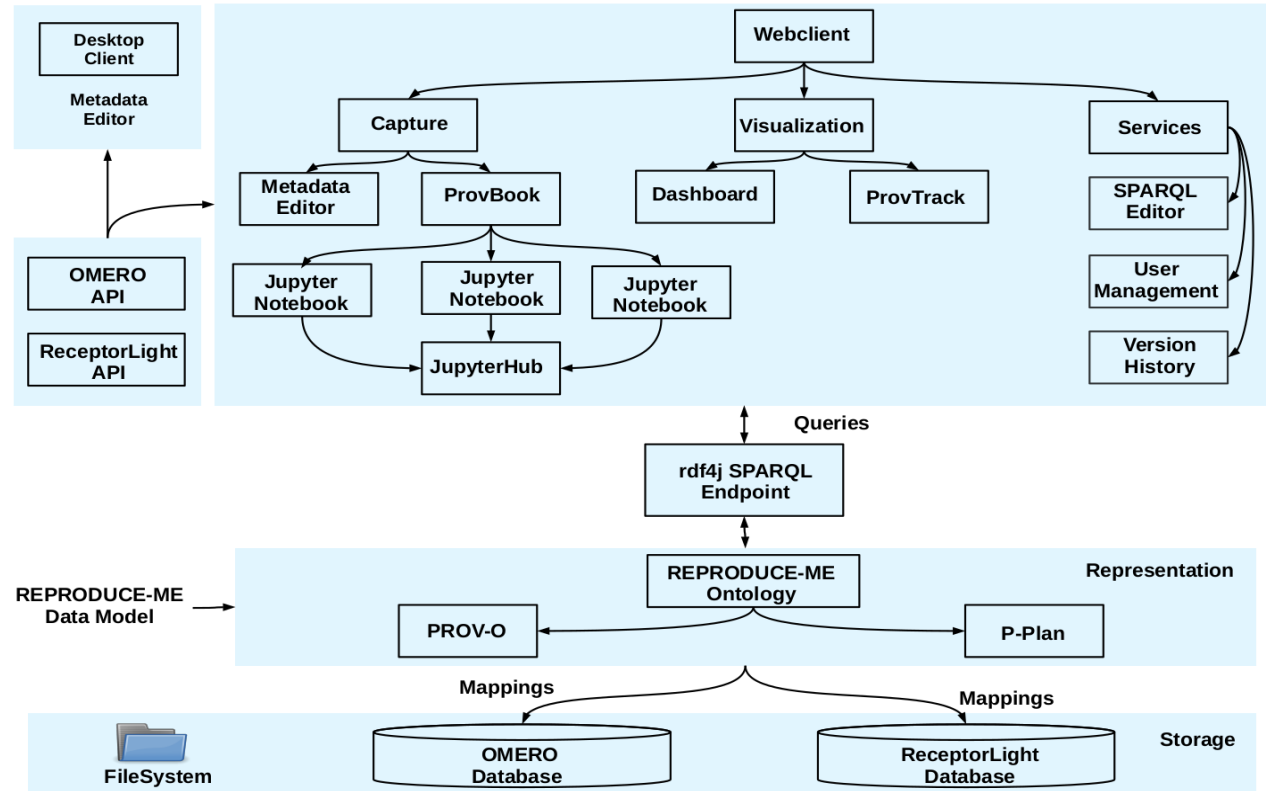
# CAESAR

A Collaborative Environment  
for Scientific Analysis with  
Reproducibility

# CAESAR

## Collaborative Environment for Scientific Analysis with Reproducibility

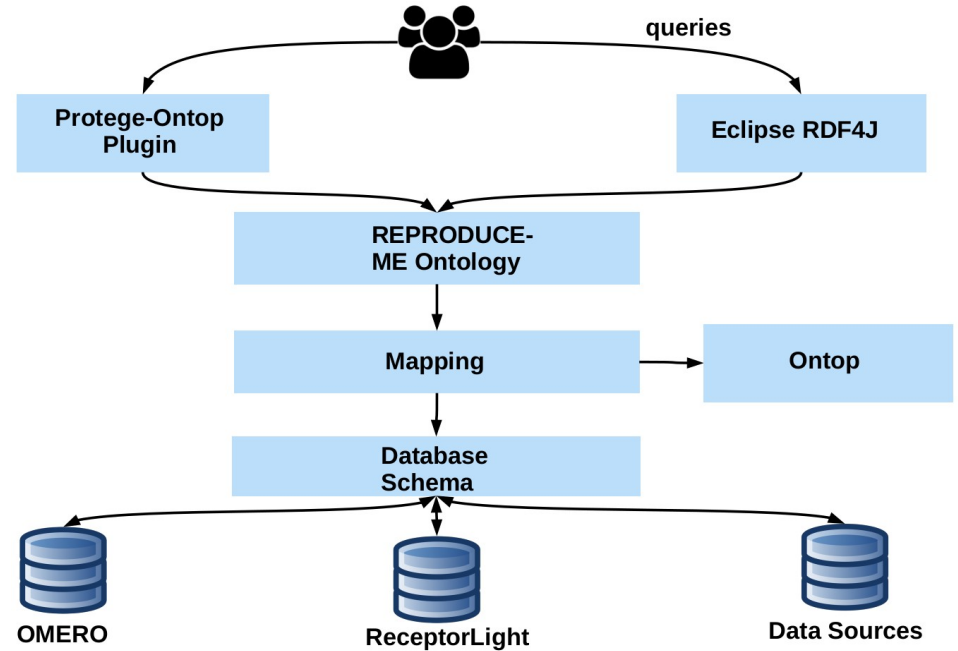
- › Extends OMERO
- › Experimental data management with images.
- › Form-based provenance capture system
- › Link Experiments with its associated variables
- › User and Group management
- › Suggestions on other user's experimental data.
- › Version history of an experiment
- › Reuse of experiments



[Samuel et al. 2017, Samuel et al., 2018]

# CAESAR: Provenance Representation

- Ontology-based data access
  - Data Sources
  - Ontologies
  - Federation
  - Mappings
- Around 800 mappings to create the virtual RDF graph



[Samuel and König-Ries 2017]

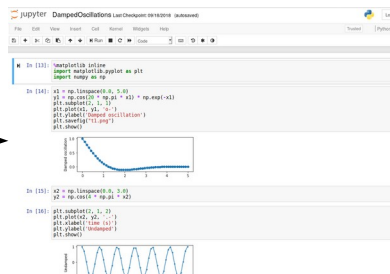


# CAESAR: Computational Reproducibility

- Integration of JupyterHub (<https://jupyter.org/hub>)
  - A distributed, collaborative and multi-user research environment
  - Directly work with the images and other datasets linked to an experiment in CAESAR
- Integration of ProvBook
  - The Notebook is linked to the experiment that used them
- Create a knowledge graph of the provenance of experiments with their computational and non-computational steps.



work



work



# Project Dashboard: Provenance Visualization

## Features

- Visualization of provenance data at the project level
- A panel for each component of a story

## Panels

- External Resources
- Steps
- Devices
- Settings
- Jupyter Notebooks
- Results

The screenshot displays a project dashboard with the following panels:

- Explore:** A file tree showing project components like 'Binding via FRET 8', 'dose-binding A1-617-GFP', 'dose-response A1-617-GFP', 'dose-responses A1', 'fcGMP affinity to CNGA1 6', '...finity to CNGA1-617-GFP', 'fcGMP efficiency', '...ficity to ligand binding 33', 'GFP bleaching', 'Example Data 3', and 'Orphaned Images'.
- The Plot:** A table with columns: startedAtTime, Experiment, AgentRole, AgentName. Data rows include: 2018-02-28T... A1+fcGMP Project fcGMP affir; 2017-02-28T... fcGMP Disp... Research G... ReceptorLi; 2018-02-28T... A1+fcGMP Research G... ReceptorLi; 2017-02-28T... fcGMP Disp... Project FRET speci.
- The Characters:** A table with columns: rsonName, Experiment, Plan, PersonRole. Data rows include: fcGMP Disp... Solution Pr... Aliquots Re...; A1+fcGMP Solution Pr... Aliquots Re...; fcGMP Disp... Solution Pr... Aliquots Re...; A1+fcGMP Solution Pr... Aliquots Re...; fcGMP Disp... Solution Pr... Aliquots Re...
- Materials:** A table with tabs for Vector, Plasmid, Protein, Chemical, Solution (selected), DNA, RNA, Restriction Enzyme. Sub-tabs for Fluorescent Protein and Oligonucleotide. Data rows include: -4°C A1+fcGMP 150mM KCl + 1µ... KCl 150mM KCl; 4°C fcGMP Disp... 150mM KCl + 1µ... KCl 150mM KCl.

# ProvTrack: Provenance Visualization

## ProvTrack: Tracking Provenance of Scientific Experiments

Search

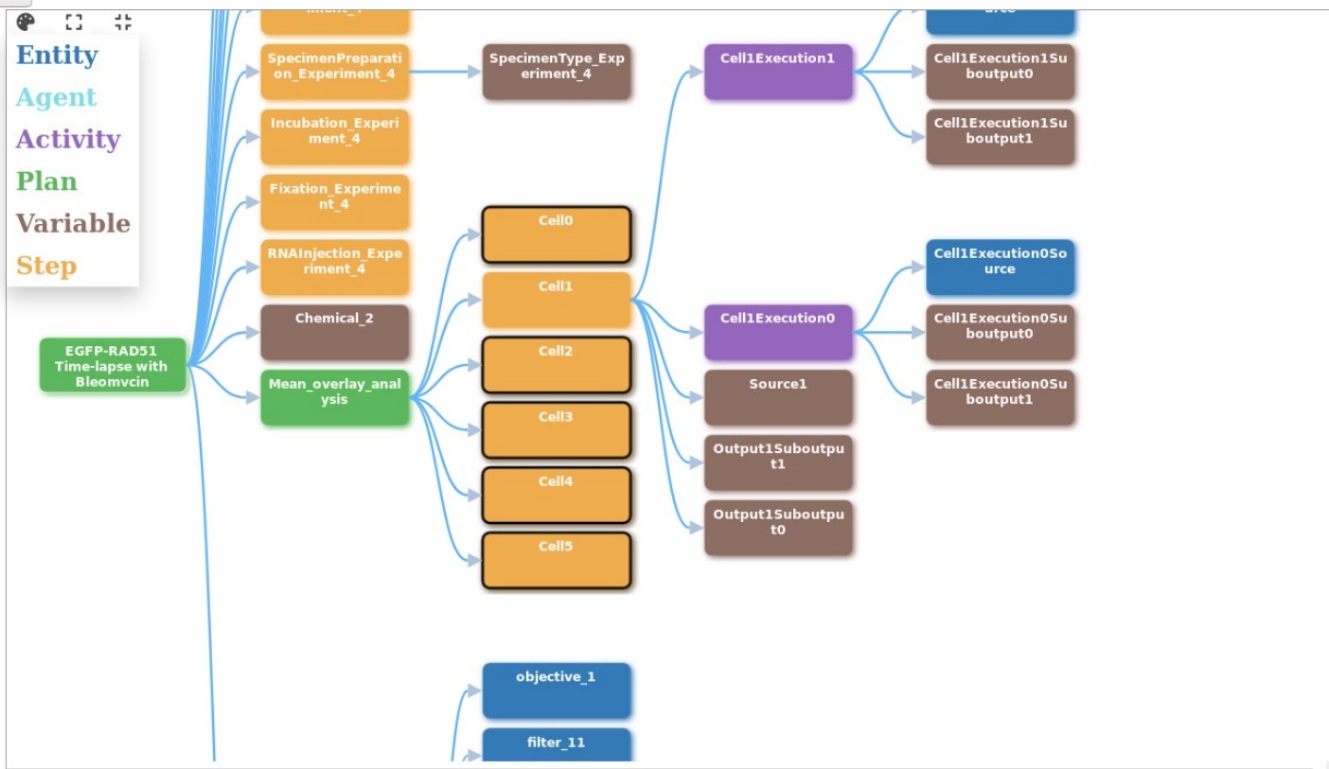
Select an Experiment to track its provenance.

EGFP-RAD51 Time-lapse with Bleomycin

**Path**  
[EGFP-RAD51 Time-lapse with Bleomycin](#)

**Infobox**

Key	Value
<a href="https://w3id.org/reproduce#description">https://w3id.org/reproduce#description</a>	EGFP-RAD51-transfected, S-phase synchronised U2OS cells in HEPES-1 DMEM (without phenolred, without are subjected to bleomycin. Fluoresc imaging of EGFP in time lapse with a frame taken every 10 min the course of 4 h.
<a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">http://www.w3.org/1999/02/22-rdf-syntax-ns#type</a>	<a href="https://w3id.org/reproduce#Experiment">https://w3id.org/reproduce#Experiment</a> <a href="http://www.opmw.org/model/p-plan/">http://www.opmw.org/model/p-plan/</a> <a href="http://www.w3.org/ns/prov#Plan">http://www.w3.org/ns/prov#Plan</a> <a href="http://www.w3.org/ns/prov#Activity">http://www.w3.org/ns/prov#Activity</a> <a href="http://www.w3.org/ns/prov#Entity">http://www.w3.org/ns/prov#Entity</a>
<a href="https://w3id.org/reproduce#hasDataset">https://w3id.org/reproduce#hasDataset</a>	<a href="https://w3id.org/reproduce#dataset">https://w3id.org/reproduce#dataset</a>
<a href="https://w3id.org/reproduce#status">https://w3id.org/reproduce#status</a>	1
<a href="http://www.w3.org/2000/01/rdf-schema#label">http://www.w3.org/2000/01/rdf-schema#label</a>	Experiment
<a href="http://www.w3.org/ns/prov#startedAtTime">http://www.w3.org/ns/prov#startedAtTime</a>	2018-10-26T00:00:00
<a href="https://w3id.org/reproduce#name">https://w3id.org/reproduce#name</a>	EGFP-RAD51 Time-lapse with Bleom
<a href="https://w3id.org/reproduce#id">https://w3id.org/reproduce#id</a>	4
<a href="https://w3id.org/reproduce#isAccessibleTo">https://w3id.org/reproduce#isAccessibleTo</a>	<a href="https://w3id.org/reproduce#ExperimenterGroup">https://w3id.org/reproduce#ExperimenterGroup</a> <a href="https://w3id.org/reproduce#Researchgroup_Exp">https://w3id.org/reproduce#Researchgroup_Exp</a> <a href="https://w3id.org/reproduce#ContactPerson_Exp">https://w3id.org/reproduce#ContactPerson_Exp</a>



InfoBox

Provenance Graph

# Conclusion

- Reproducibility: an important concern and needs much attention
- Open Science for Reproducibility
- Different tools to support reproducibility with provenance and semantic web

# Thank you for your attention

 @sheebasamuel

 0000-0002-7981-8504

- <https://fusion.cs.uni-jena.de/>
- <https://w3id.org/reproduceme/research>
- <https://sheeba-samuel.github.io/>



FRIEDRICH-SCHILLER-  
UNIVERSITÄT  
JENA



MichaelStifelCenterJena  
for Data-Driven and Simulation Science

