TECHNISCHE UNIVERSITÄT
CHEMNITZ

Reproducible Research with Data Provenance

Responding to 6W and 1H Questions of Data Provenance

Sheeba Samuel

Chemnitz University of Technology, Chemnitz, Germany

Outline

- Introduction
- Definitions: Reproducibility & Repeatability
- Reproducible Research
- Data Provenance and its Semantics
- Solutions to 6W and 1H questions of Data Provenance

Story






Image Source: Wikimedia Commons: Snow White and the Seven Dwarfs

Story of an Experiment

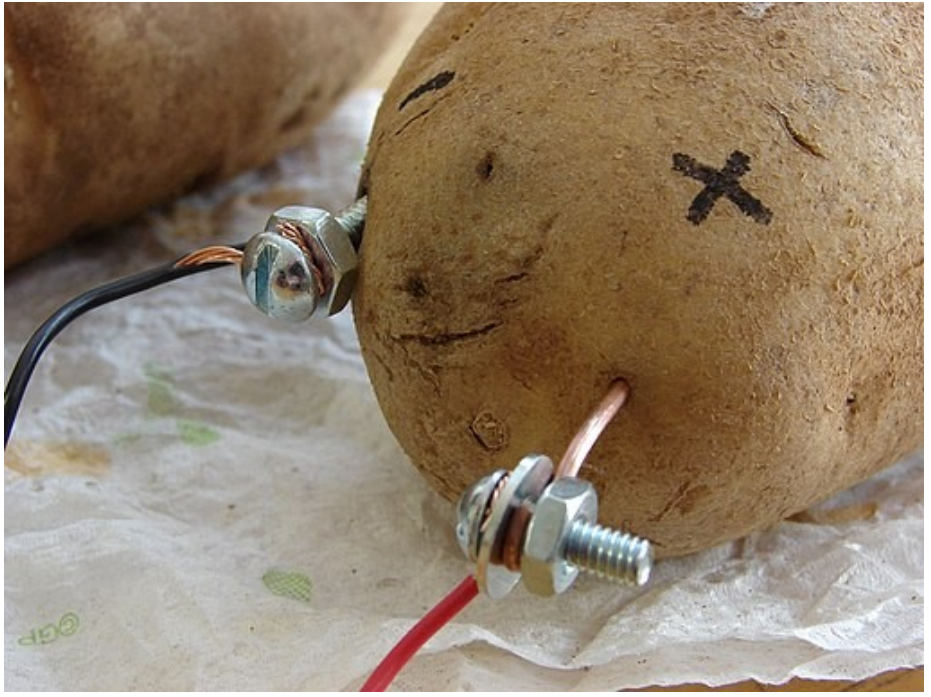


Image Source: Wikimedia Commons

Story of an Experiment

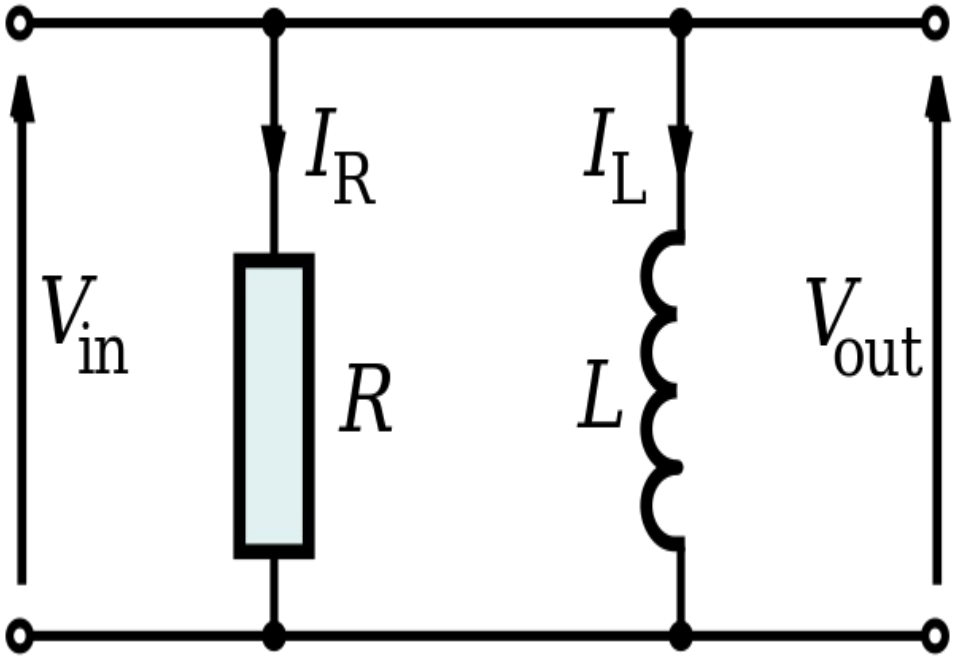
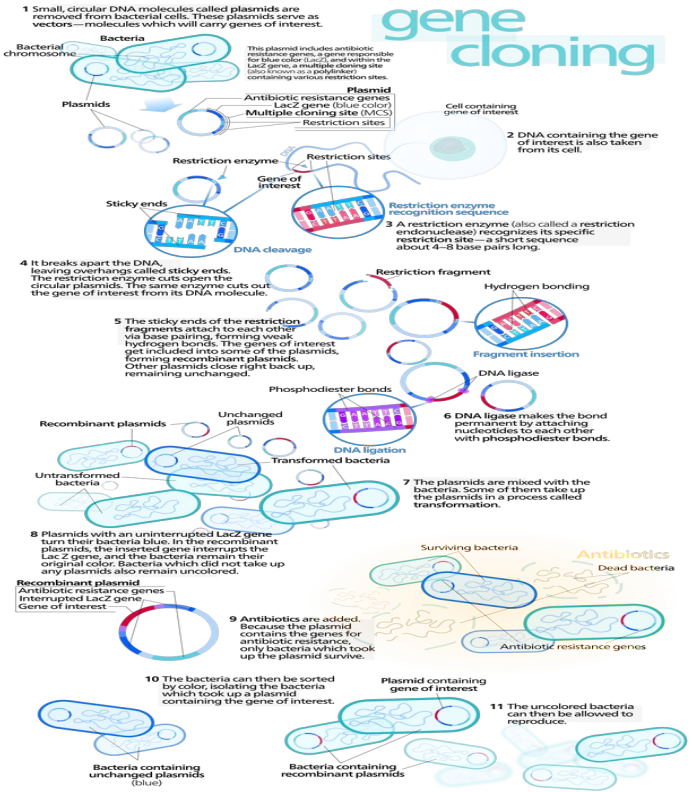


Image Source: Pixabay

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Story of an Experiment

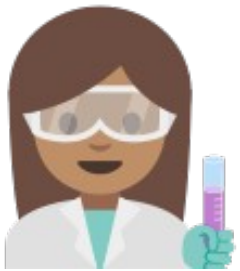


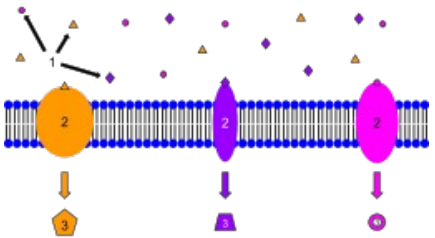
Source: Wikimedia Commons

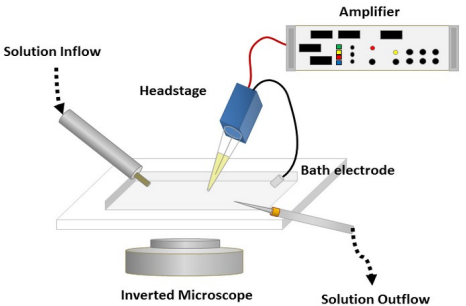
Image Source: Pixabay

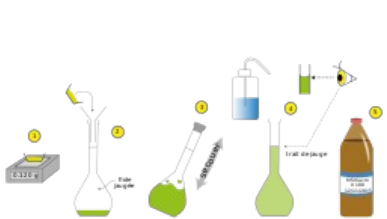
TECHNISCHE UNIVERSITÄT CHEMNITZ

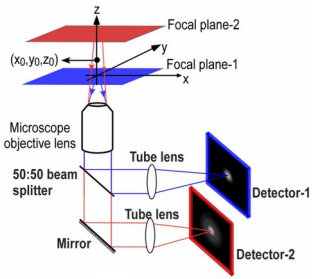
Scientific Experiments

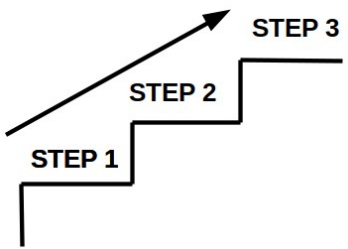


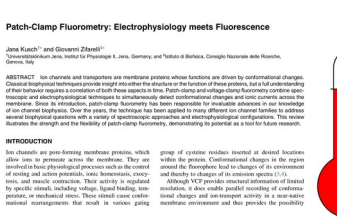


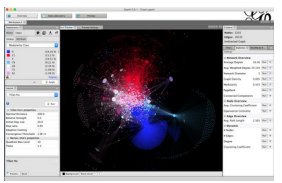


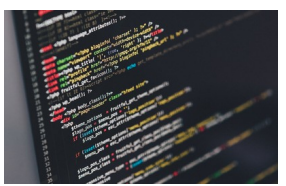


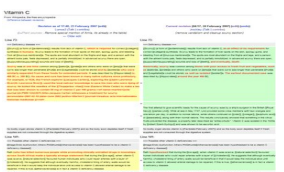


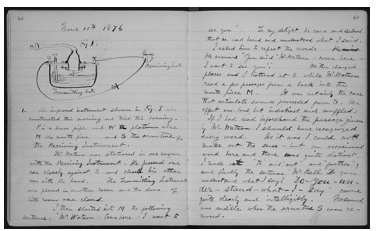








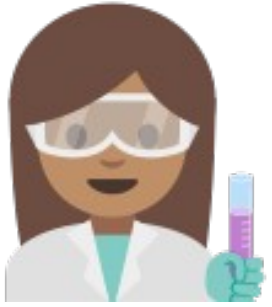


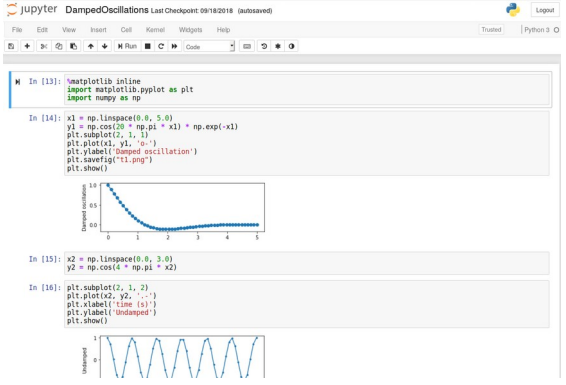


TECHNISCHE UNIVERSITÄT CHEMNITZ

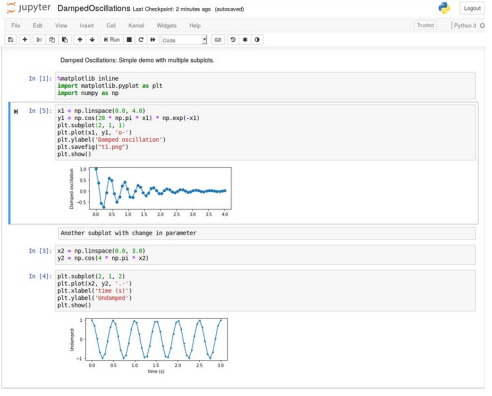
Computational Experiments

- Repeatability





- Reproducibility



[A provenance-based semantic approach to support understandability, reproducibility, and reuse of scientific experiments, Samuel S., 2019]

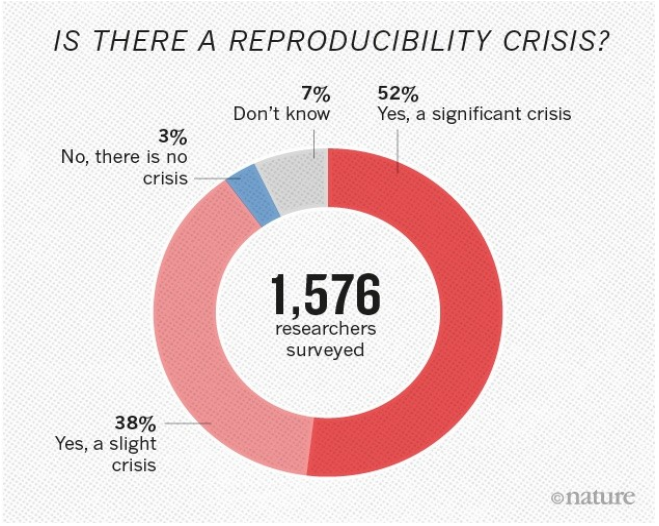
Definition of a Confused Terminology

	ACM Definition (August 24, 2020)	ACM Definition Revised (November 4, 2021)
Repeatability	Same team, same experimental setup	Same team, same experimental setup
Reproducibility	Different team, same experimental setup	Different team, different experimental setup
Replicability	Different team, different experimental setup	Different team, same experimental setup

*https://www.acm.org/publications/policies/artifact-review-badging

Reproducibility Crisis

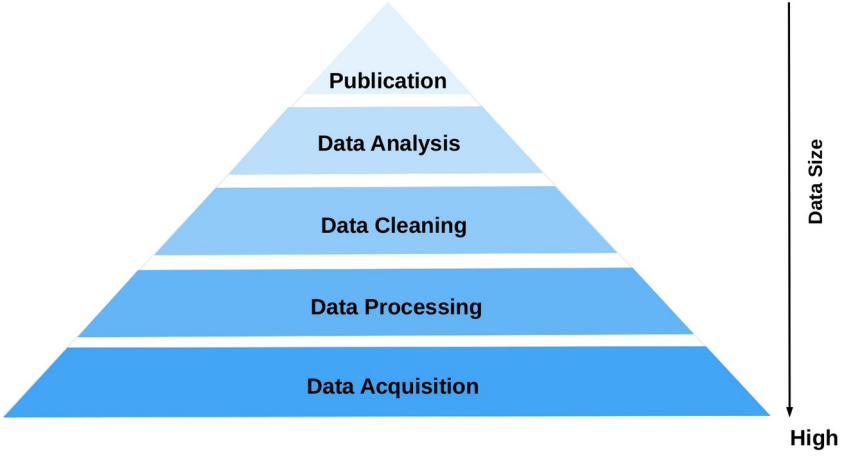
Reproducibility Crisis*



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." - **natureresearch**



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

TECHNISCHE UNIVERSITÄT CHEMNITZ

Insights from Interviews

It is important to document **the individual trials** to understand which **possibilities did not work out**.

Biologist

Methods in the publication are **not sufficient** to fully reproduce an experiment.

Physicist

Lack of awareness of data management process

Ecologist

The documentation of **procedures, devices, and steps** are **vital** for reproducibility.

Biochemist

Since the experimental metadata are stored at different places including **lab books & external hard disks,** it gets **hard to distinguish** good experiments or to find a specific experiment.

Biologist

Simple tools for data management rather than complicated tools which require extra learning.

Ecologist

Code and data which are required to reproduce results need to be **publicly available**. The ability to **visualize and compare** results helps to understand our own and others work.

Computer Scientist

A system providing **a link to the data, metadata, and results** can help in our daily work.

Geologist

Raw and processed data are needed to understand an experiment.

Mathematician

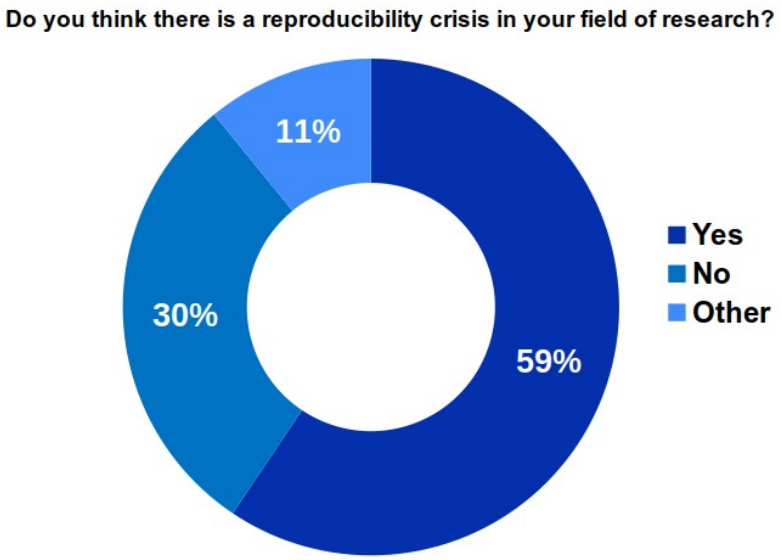
Source: CRC ReceptorLight & “Fostering reproducible science − What data management tools can do and should do for you” workshop (15 June 2017)

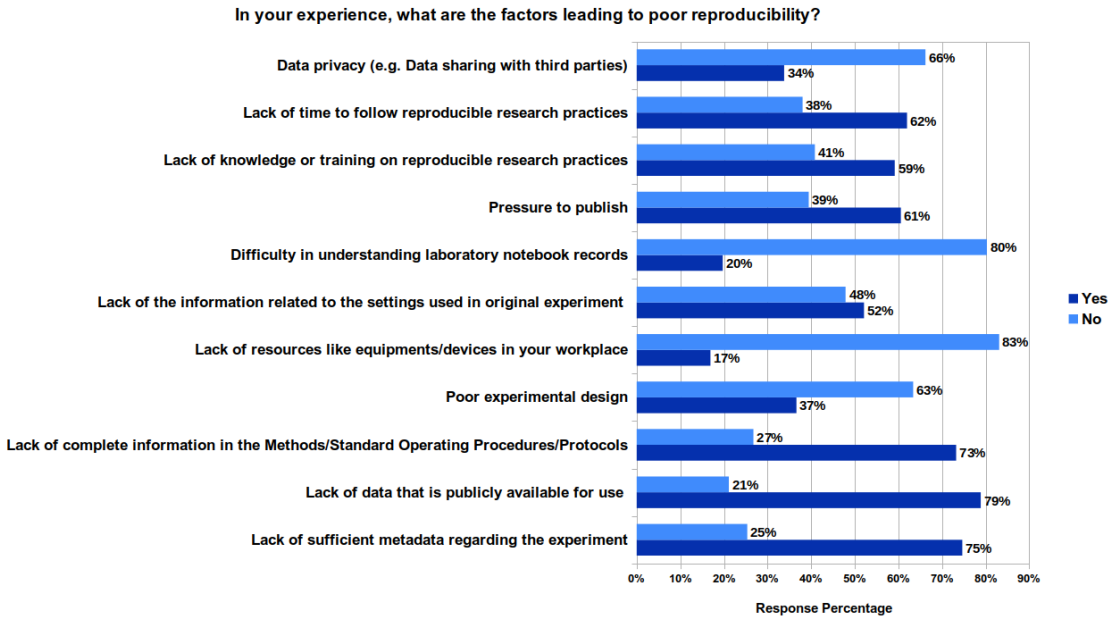
TECHNISCHE UNIVERSITÄT CHEMNITZ

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

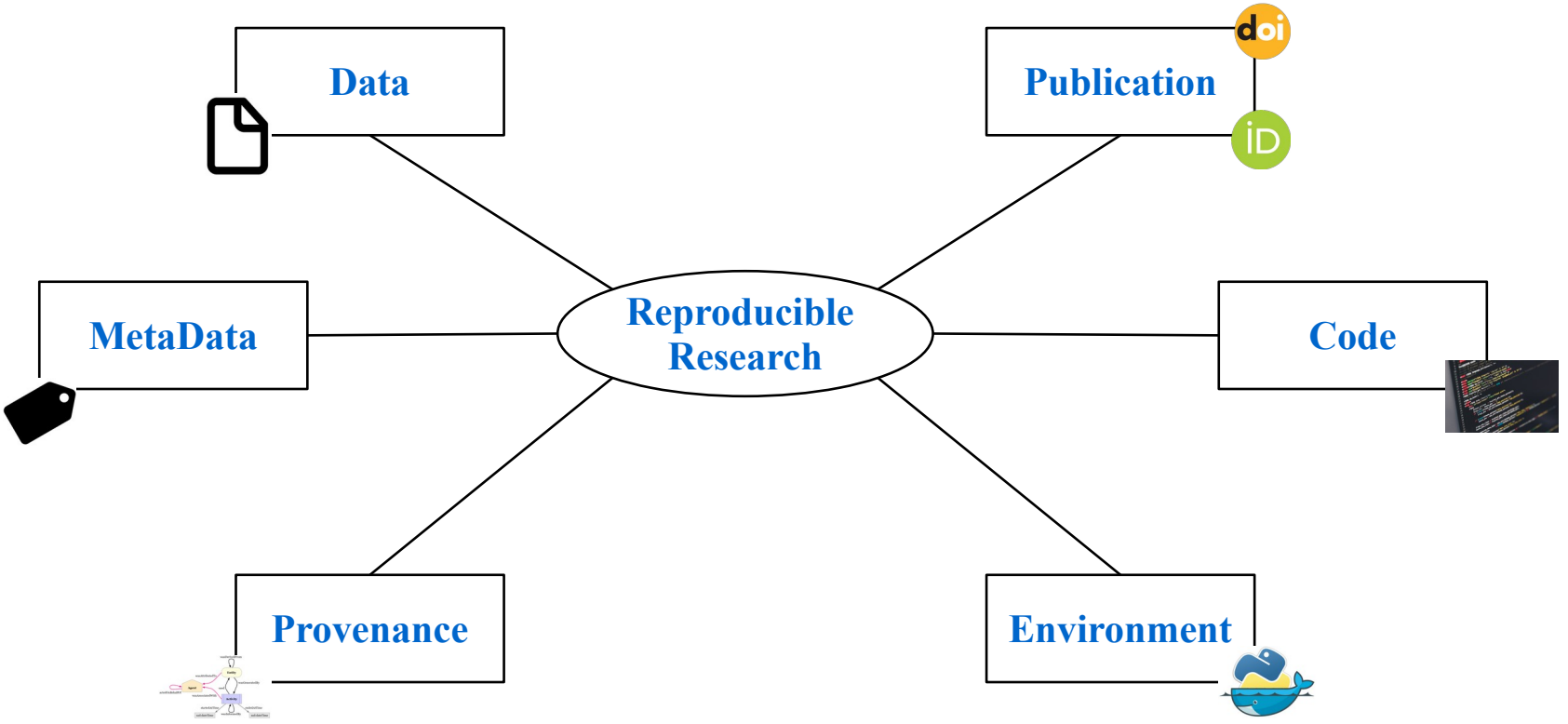




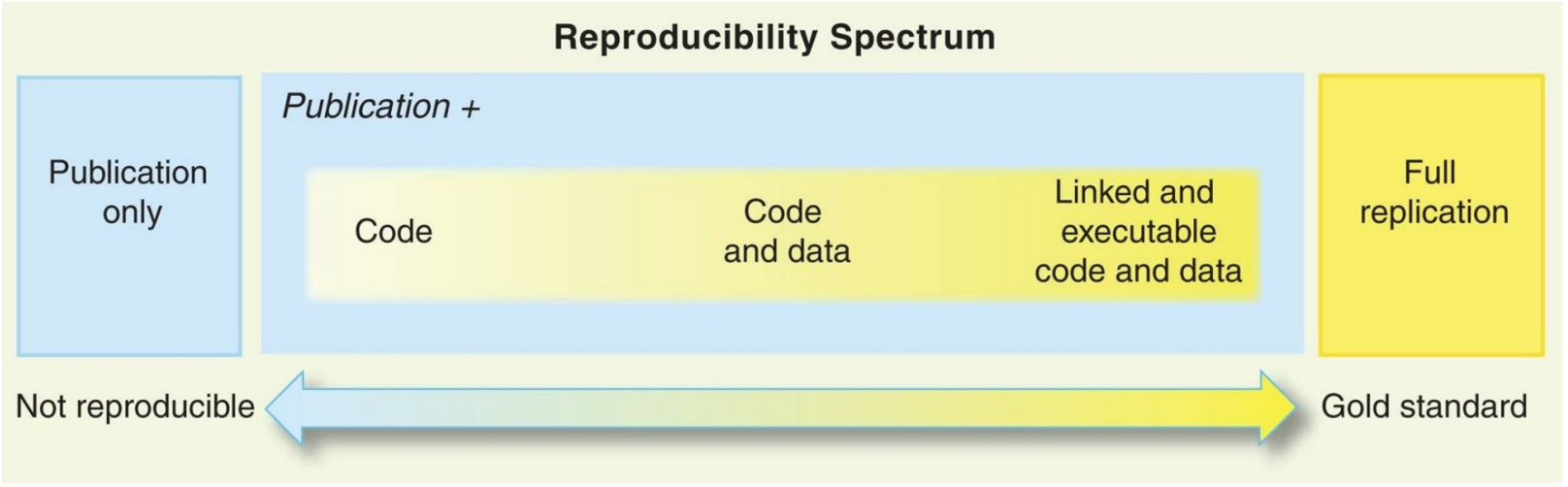
[Understanding experiments and research practices for reproducibility: an exploratory study, Samuel and König-Ries 2021]

What is Reproducible Research?

- Reproducible Research refers to the idea that scientific results are documented and published in a way that others may verify the findings and build upon them.

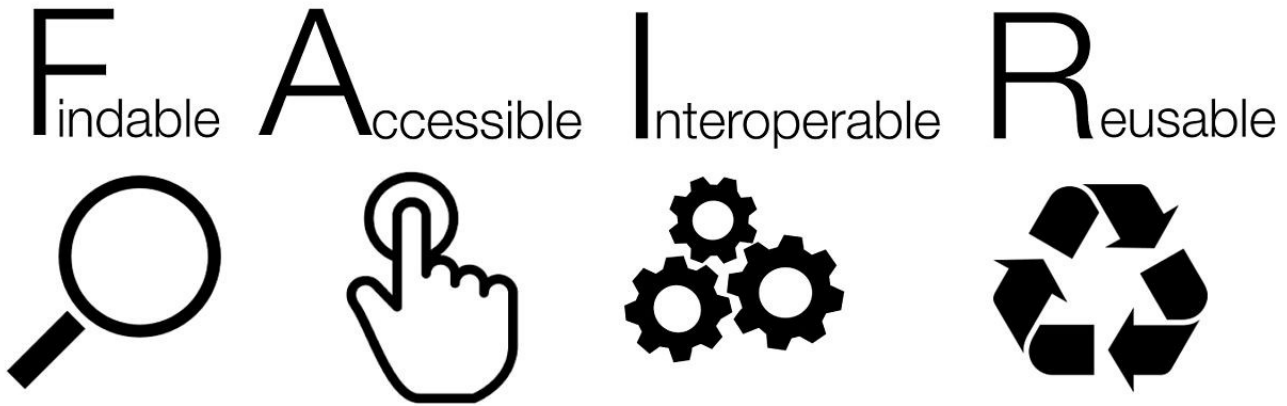


Reproducible Research in Computational Science



[Reproducible Research in Computational Science, Roger D. Peng, 2011]

FAIR Data Principles



The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al. 2016

Image Source: Wikimedia Commons

FAIR Data Principles:

- F1. (Meta)data are assigned a globally unique and persistent identifier
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al. 2016

Image Source: Wikimedia Commons

FAIR Data Principles: Accessible

- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available

The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al. 2016

Image Source: Wikimedia Commons

FAIR Data Principles: Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (Meta)data use vocabularies that follow FAIR principles
- I3. (Meta)data include qualified references to other (meta)data

The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al. 2016

Image Source: Wikimedia Commons

FAIR Data Principles: Reusable

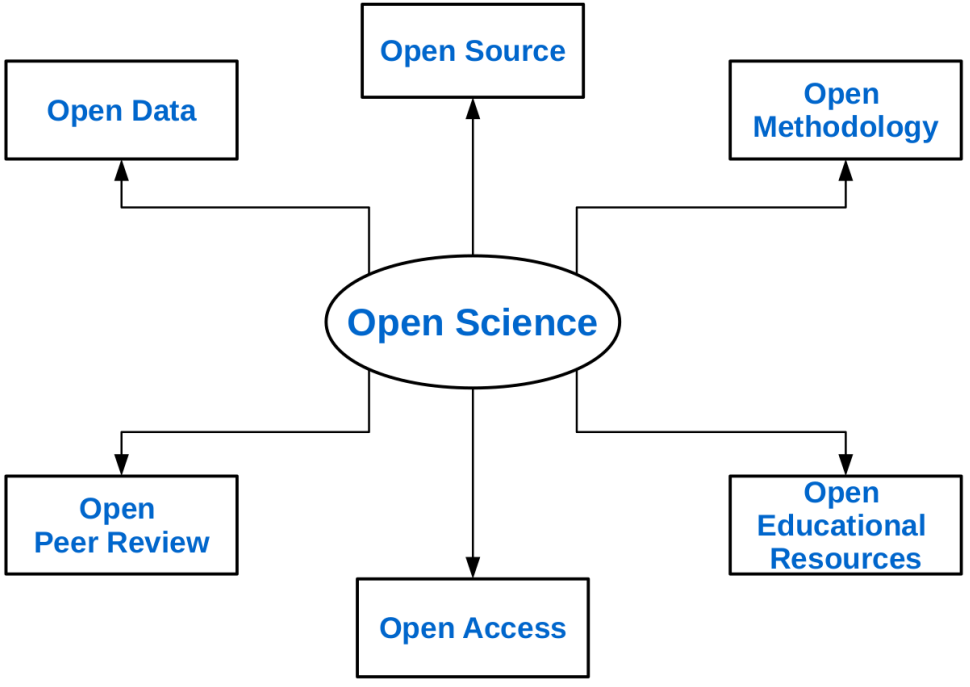
- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards

The FAIR Guiding Principles for scientific data management and stewardship, Wilkinson et al. 2016

Image Source: Wikimedia Commons

Open Science

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

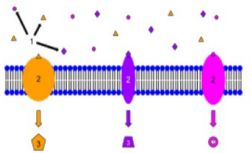


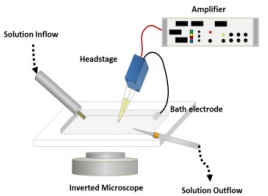
[The Open Traits Network: Using Open Science principles to accelerate trait-based science across the Tree of Life, Gallagher et al. 2020]

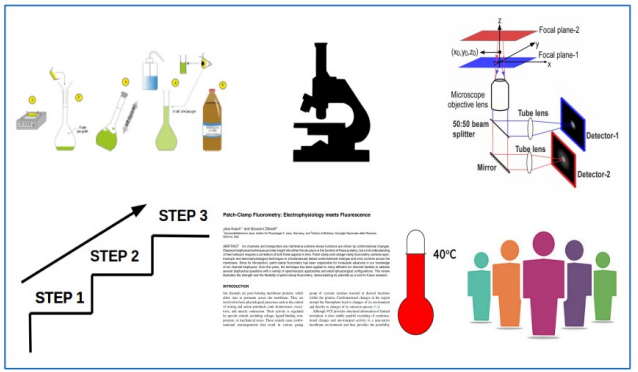
TECHNISCHE UNIVERSITÄT CHEMNITZ

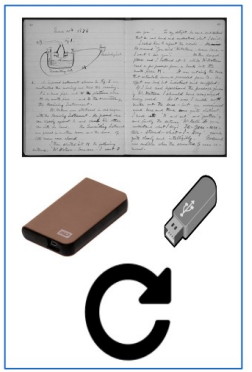
Open Reproducible Scientific Experiments

- 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









Factors behind irreproducible research

- Proprietary Data Standards and File Formats
- Availability of hardware
- Unavailability of Experiment Materials
- Not enough description in methodology
- Missing steps
- Human error
- Data and Software used to generate original results unavailable
- Difficulty in running computational steps
- Takes time

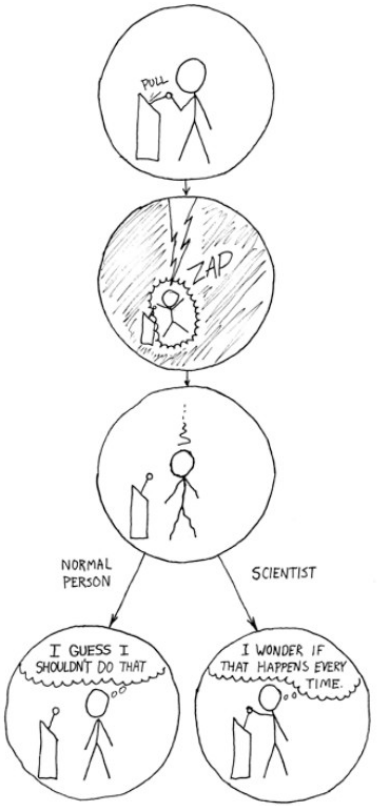


*Image Source: https://xkcd.com/242/

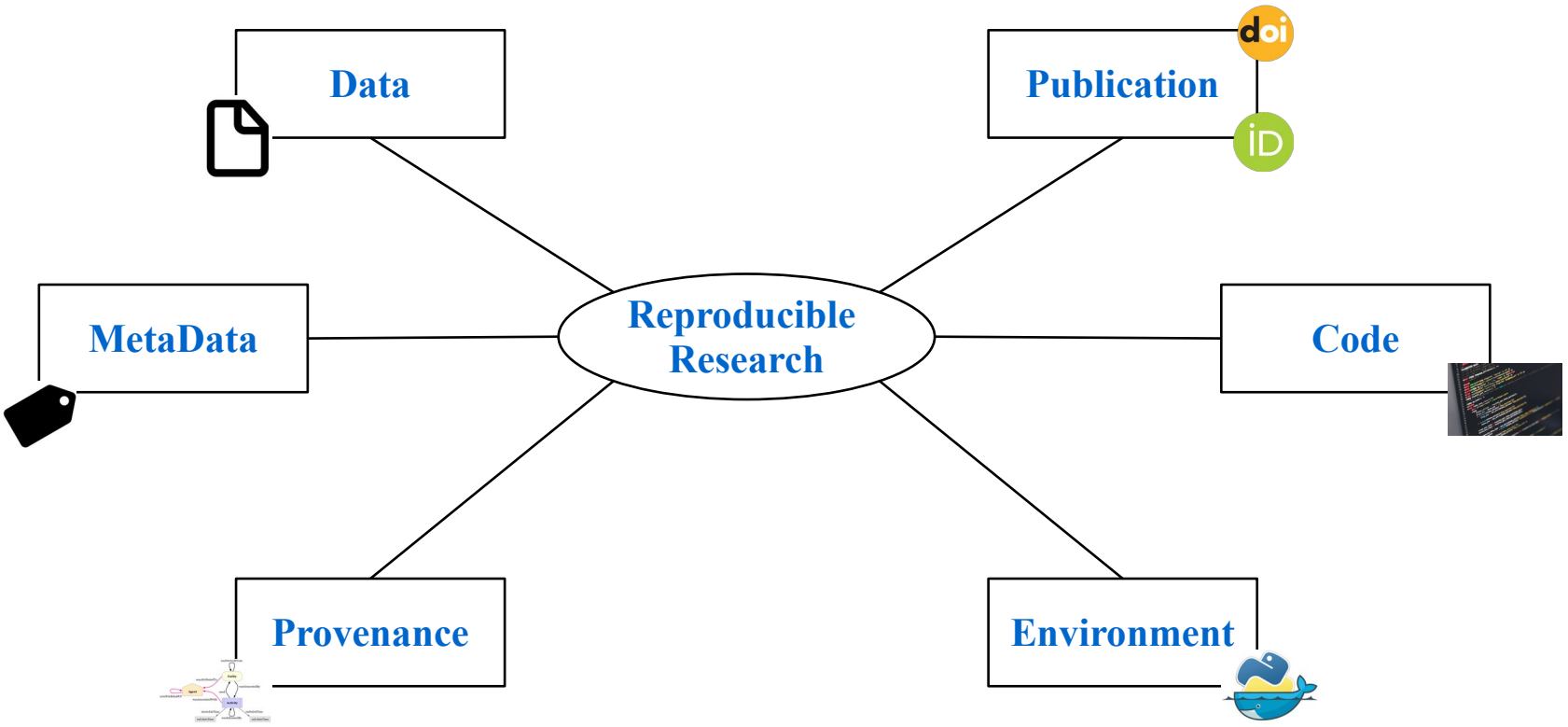
Why Reproducible Research?

- Reproducible Research benefits
  - YOU
  - The research community

Image Source: https://doi.org/10.6084/m9.figshare.5577340.v1

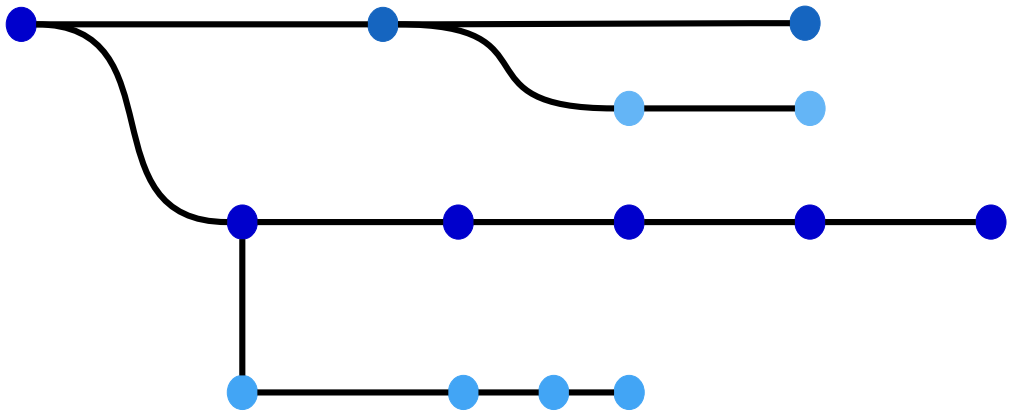
TECHNISCHE UNIVERSITÄT
CHEMNITZ

How do you make your work more reproducible?



Data Provenance

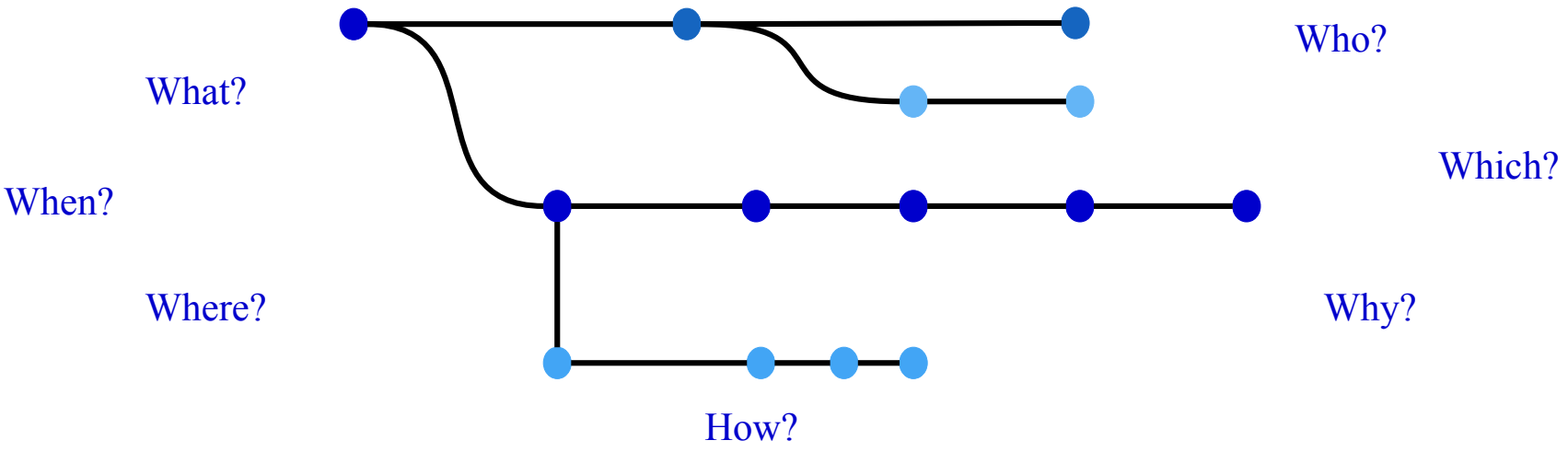
- The source or origin of an object; its history



The Semantics of Data Provenance

- W7 model: Provenance is defined as a n-tuple

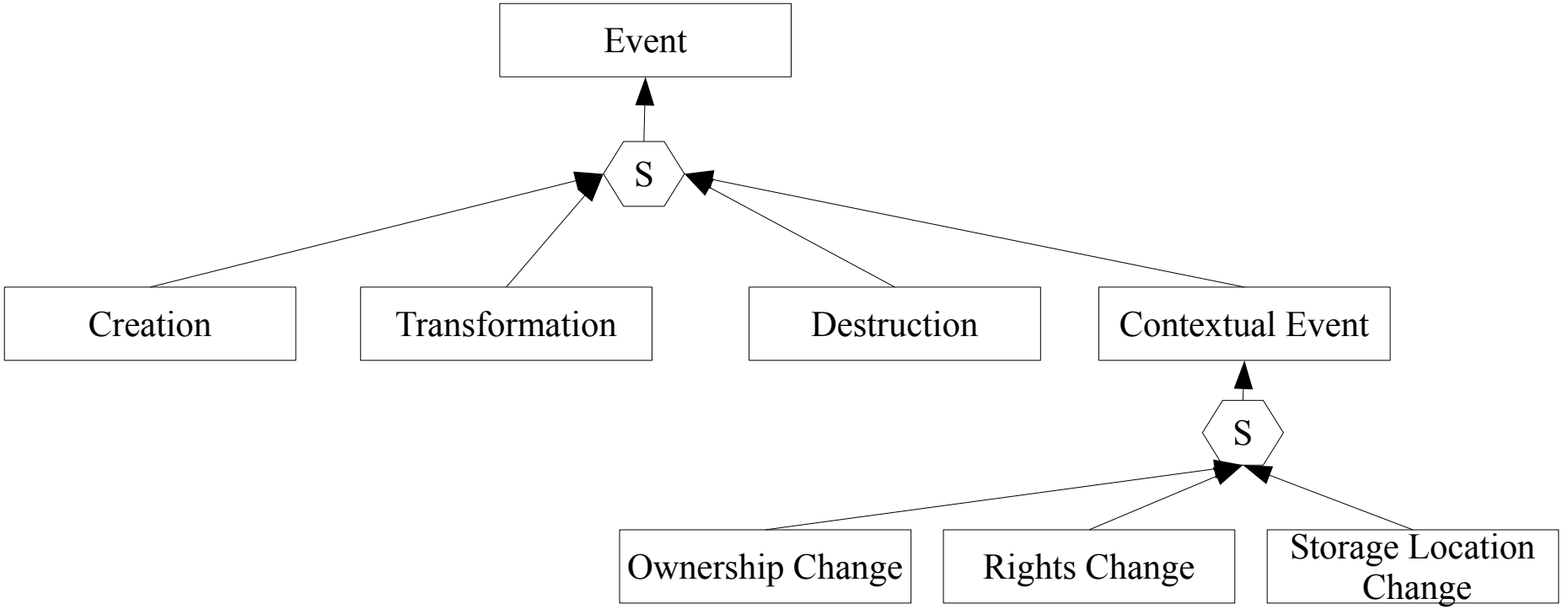
  P = (WHAT, WHEN, WHERE, HOW, WHO, WHICH, WHY)



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘What’ Provenance

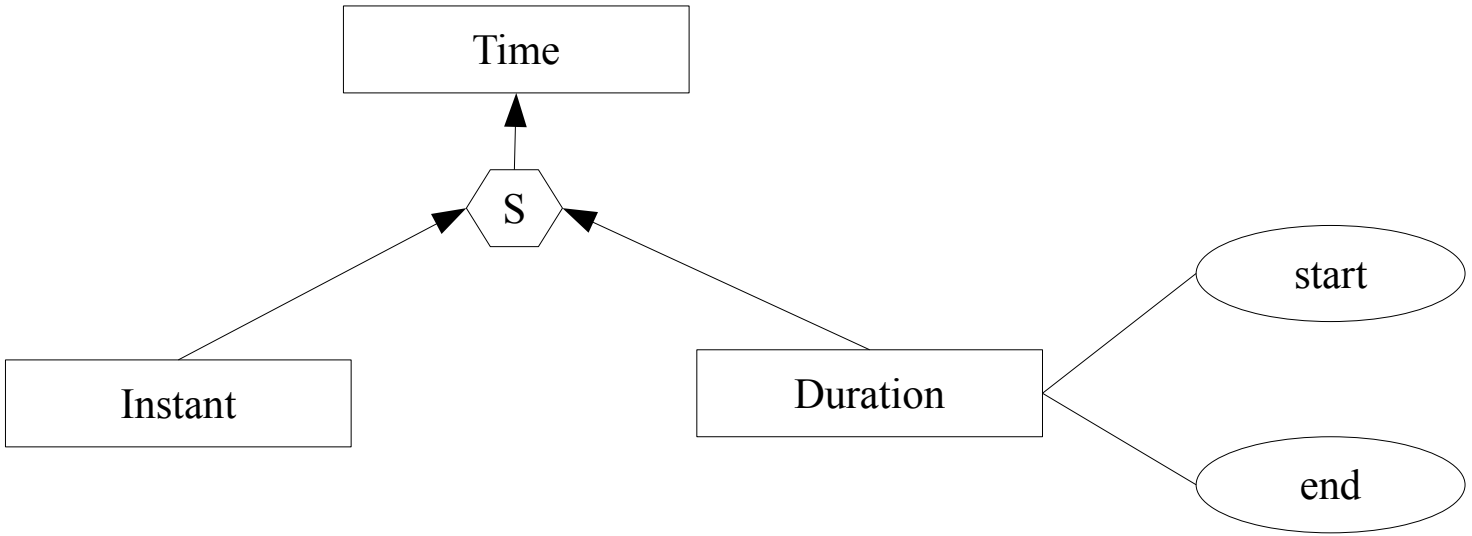
WHAT is a sequence of events that affect a data object during its life time.



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘When’ Provenance

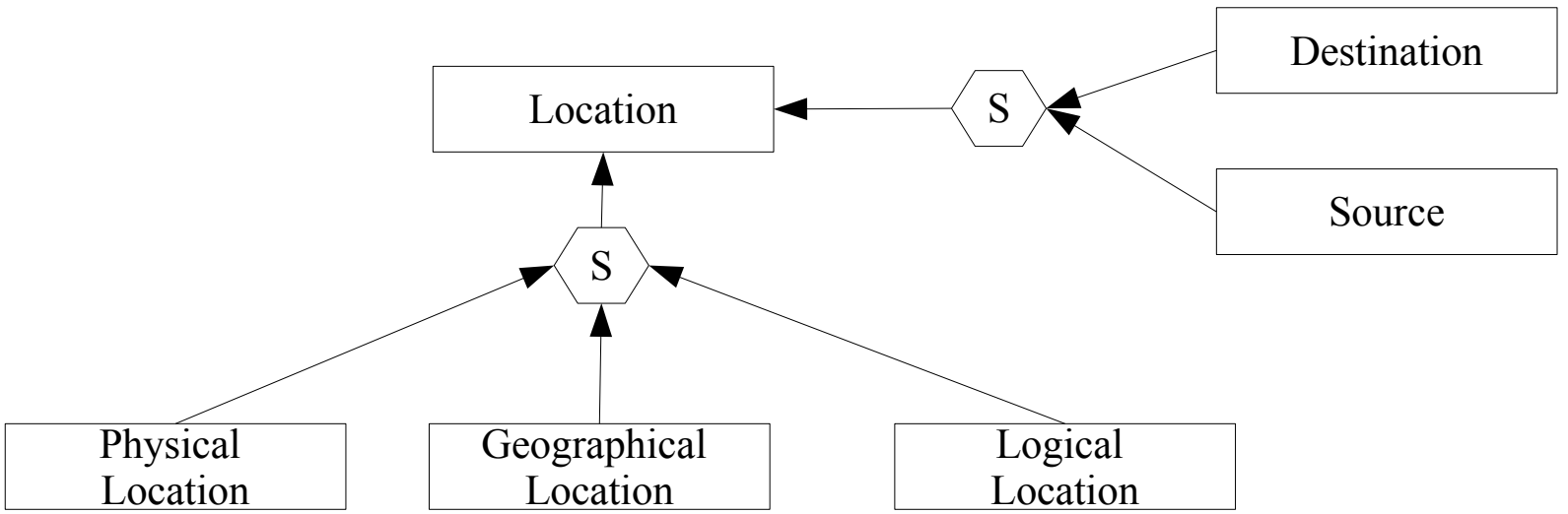
WHEN represents a set of timestamps associated with various provenance events.



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘Where’ Provenance

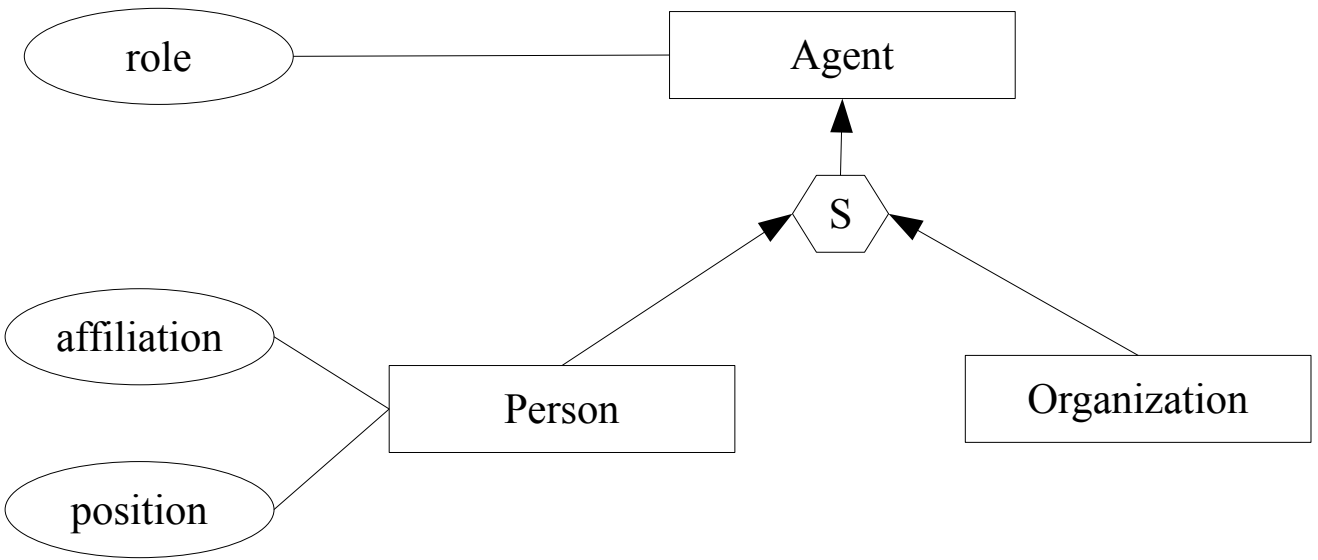
WHERE denotes a set of locations where various events happen.



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘Who’ Provenance

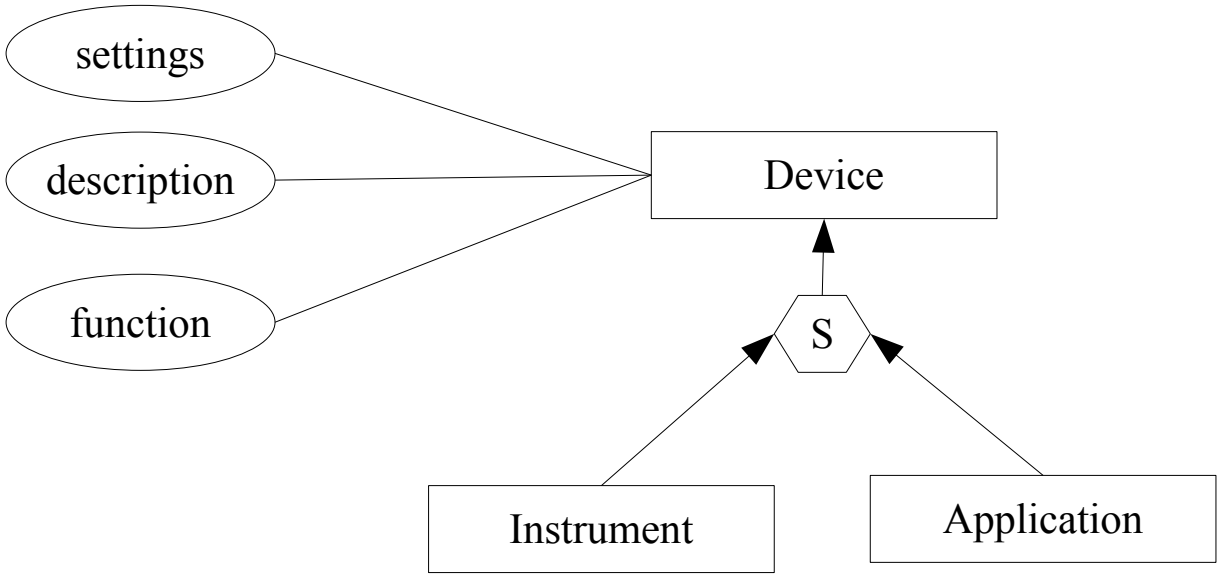
WHO is a triple (AGENT, ROLE, RL)



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘Which’ Provenance

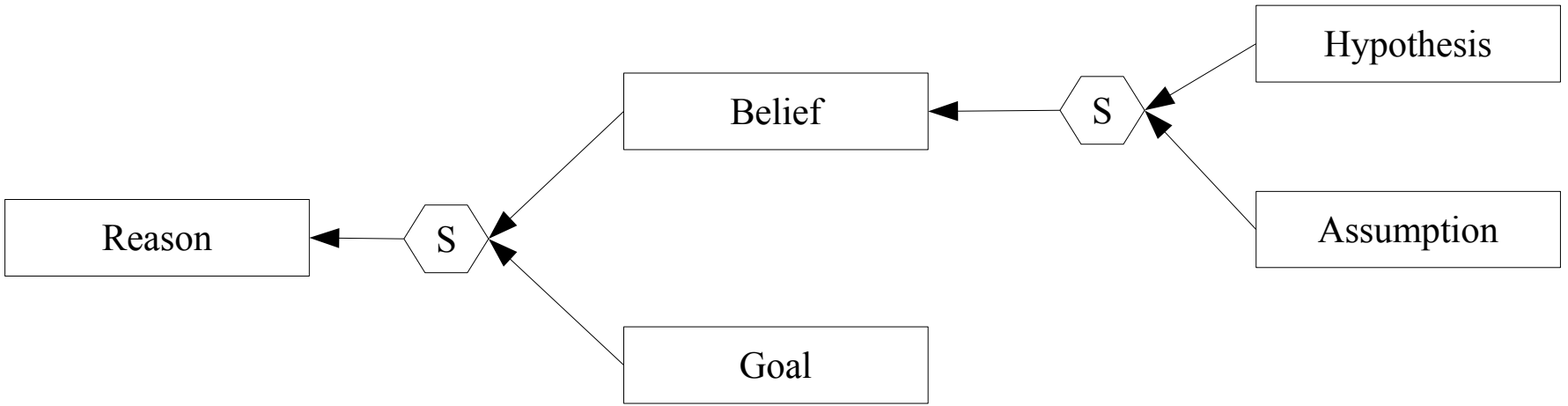
WHICH is a tuple (DEVICE, SETTINGS, DESCRIPTION, FUNCTION, S, D, F).



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘Why’ Provenance

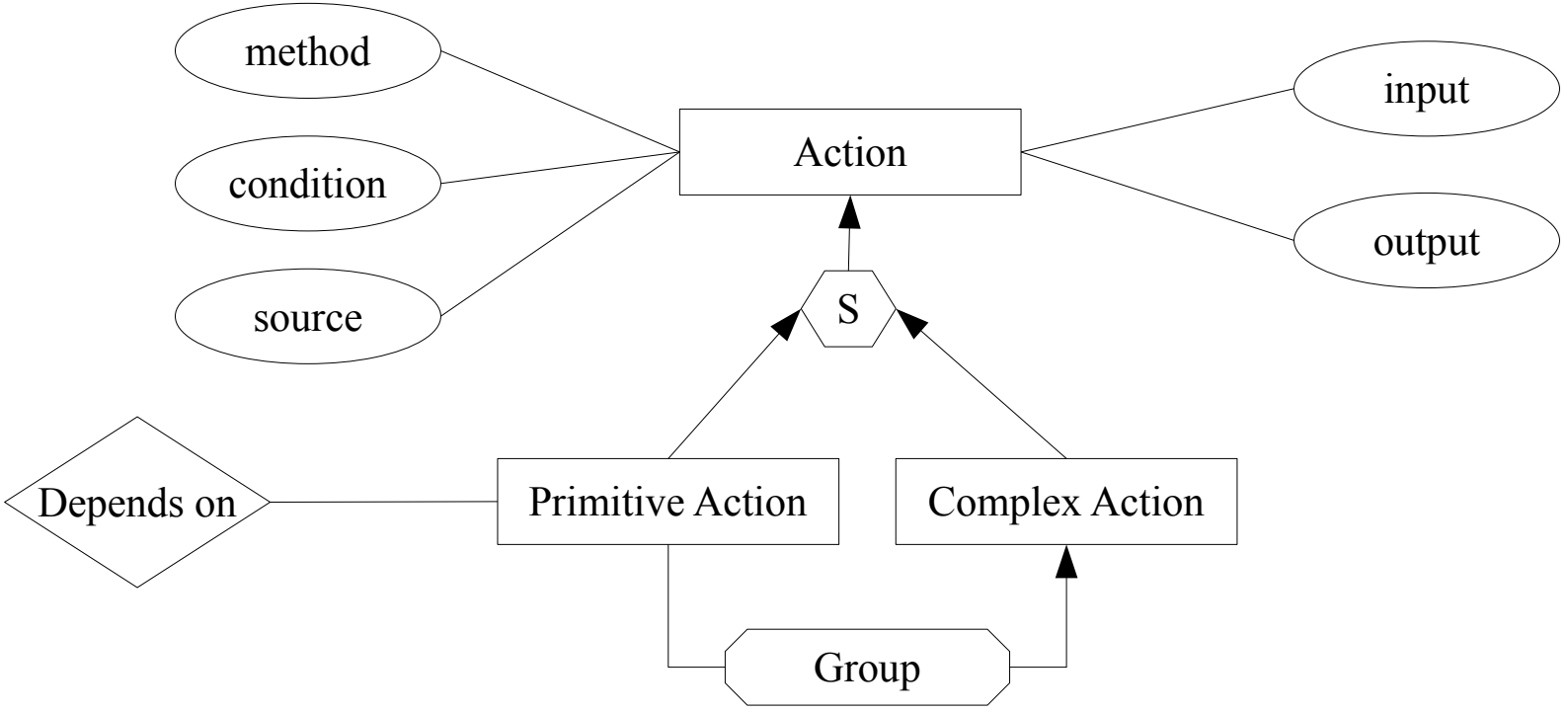
WHY represents a set of reasons for various provenance events



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Semantics of ‘How’ Provenance

HOW is defined as a tuple (ACTION, PRECONDITIONS, METHODS, INPUTS, OUTPUTS, SOURCES, P, M, I, O, S)



[Understanding the Semantics of Data Provenance to Support Active Conceptual Modeling, S Ram, J Liu, 2006]

Benefits of Provenance Tracking

- Ensuring transparency and accountability in research.
- Facilitating collaboration and knowledge sharing.
- Enhancing the credibility of research results.

Provenance Tracking Tools

- **Version Control Systems**
  - Tools: Git, SVN
  - Tracks changes to source code, documents, and other project files.
  - Enables collaboration and versioning.
  - Records commit history, including user, date, and changes made.
- **Workflow Management Tools**
  - Tools: Apache Taverna, Galaxy, Kepler
  - Manages and tracks the entire research workflow.
  - Captures dependencies, inputs, and outputs.
  - Provides a visual representation of the workflow.

Provenance Tracking Tools

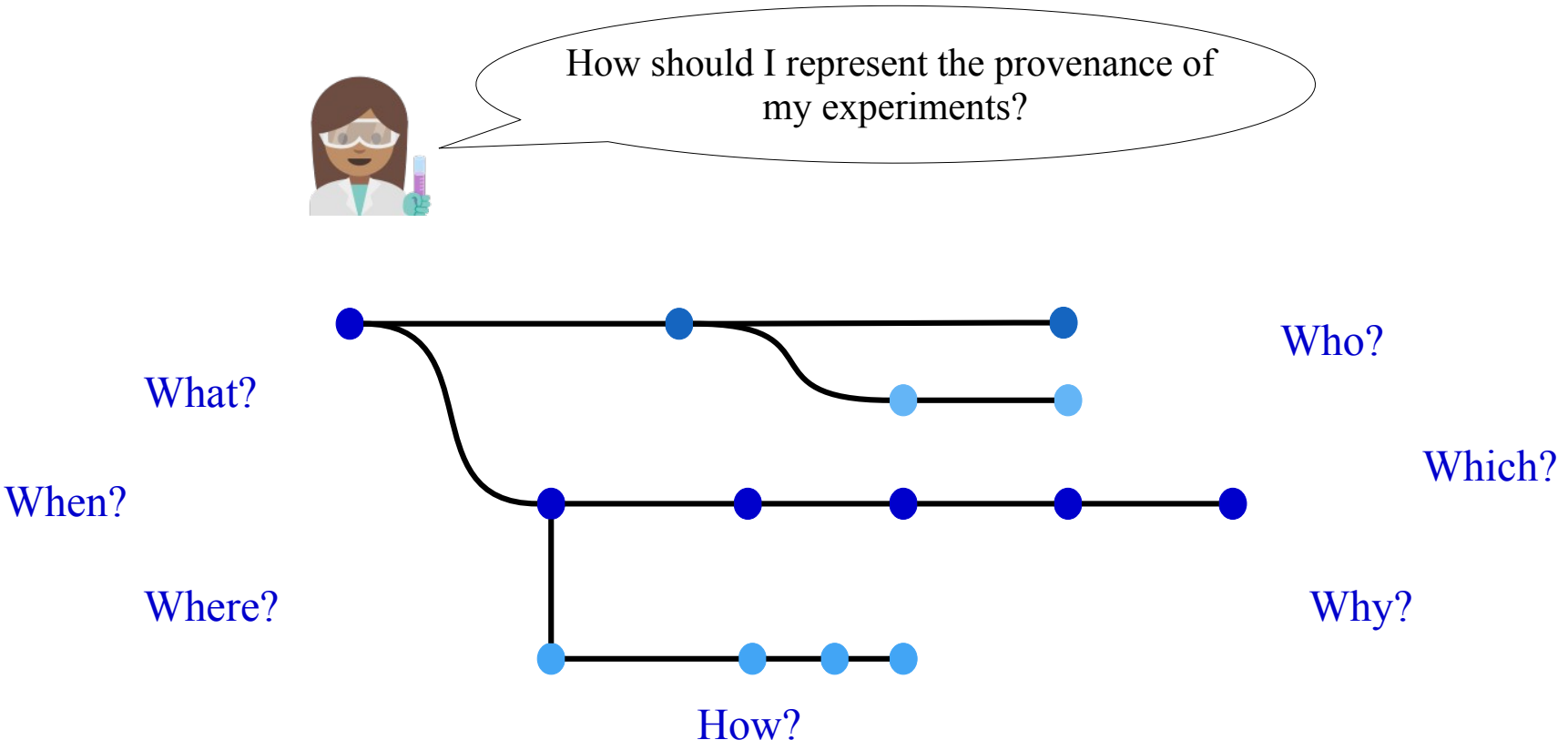
- **Provenance Representation**
  - Tools: W3C PROV, Open Provenance Model (OPM)
  - Standardizes representation of provenance information.
  - Captures data and user provenance in a machine-readable format.
  - Facilitates interoperability between different tools and systems.
- **Specialized Tools**
  - Tools: Docker (https://www.docker.com/), ReproZip (https://www.reprozip.org/)
  - Captures the computational environment and execution traces.
  - Creates reproducible packages containing code, data, and dependencies.
  - Ensures easy sharing and reproduction of computational experiments.

TECHNISCHE UNIVERSITÄT CHEMNITZ

Open science contributions: Responding to 6W1H questions of data provenance

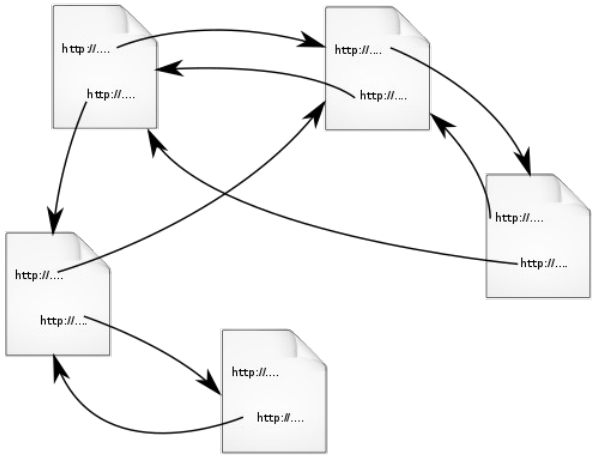
1 Semantic Description of Scientific Experiments	**The REPRODUCE-ME Data Model & Ontology** Semantic Description of Scientific experiments including Scripts and Computational Notebooks **The ReproduceMeON** An Ontology Network for Reproducibility of Scientific Studies
2 Support of computational reproducibility	**ProvBook** Provenance capture, visualize, represent and difference of results **MLProvLab** Provenance capture, visualize, represent and difference of ML notebooks **ReproduceMeGit** A tool for analyzing the reproducibility of Jupyter Notebooks **FAIR Jupyter** Knowledge Graph for the reproducibility of Jupyter Notebooks

Representing 6W1H Provenance of Experiments

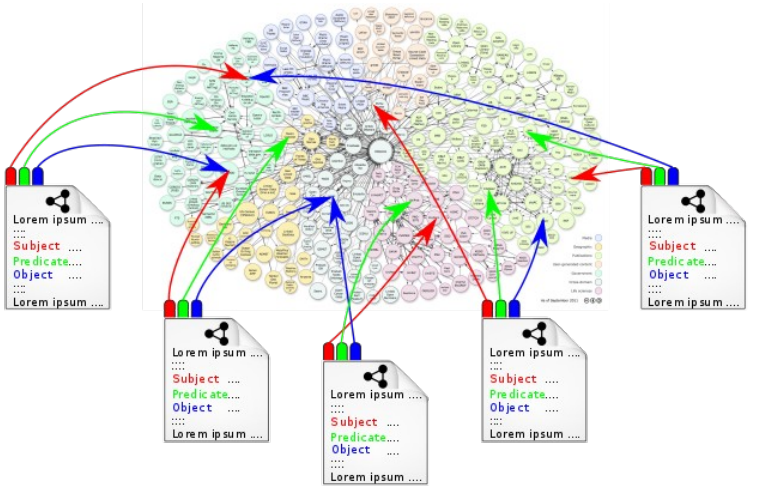


The REPRODUCE-ME Data Model & Ontology

Semantic Web



The traditional web -
A web of documents



The semantic web -
A web of human and machine readable content employing linked data

Image Source: Wikimedia Commons

Semantic Web

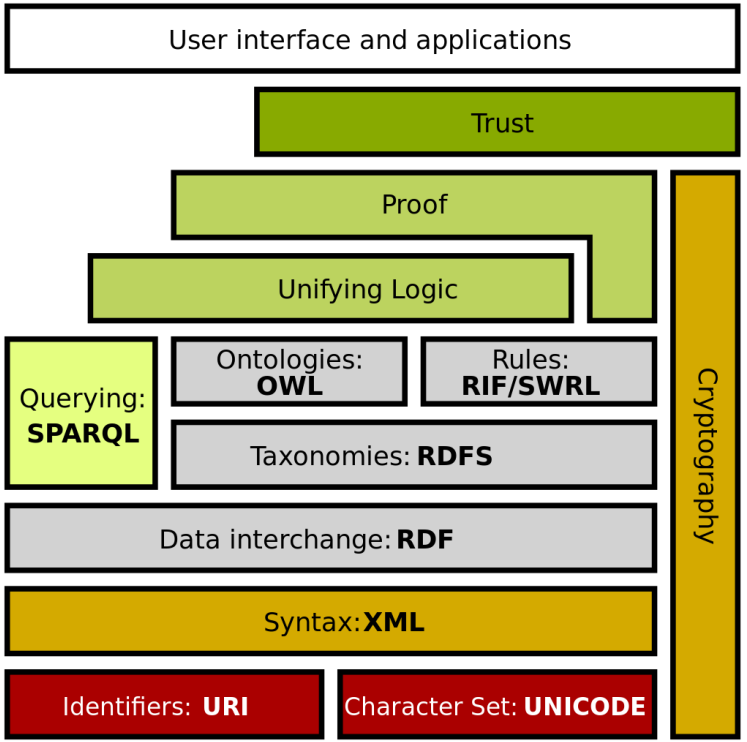


Image Source: Wikimedia Commons

Ontologies

- An ontology is a formal, explicit specification of a shared conceptualization.
  [Studer et al., 1998]

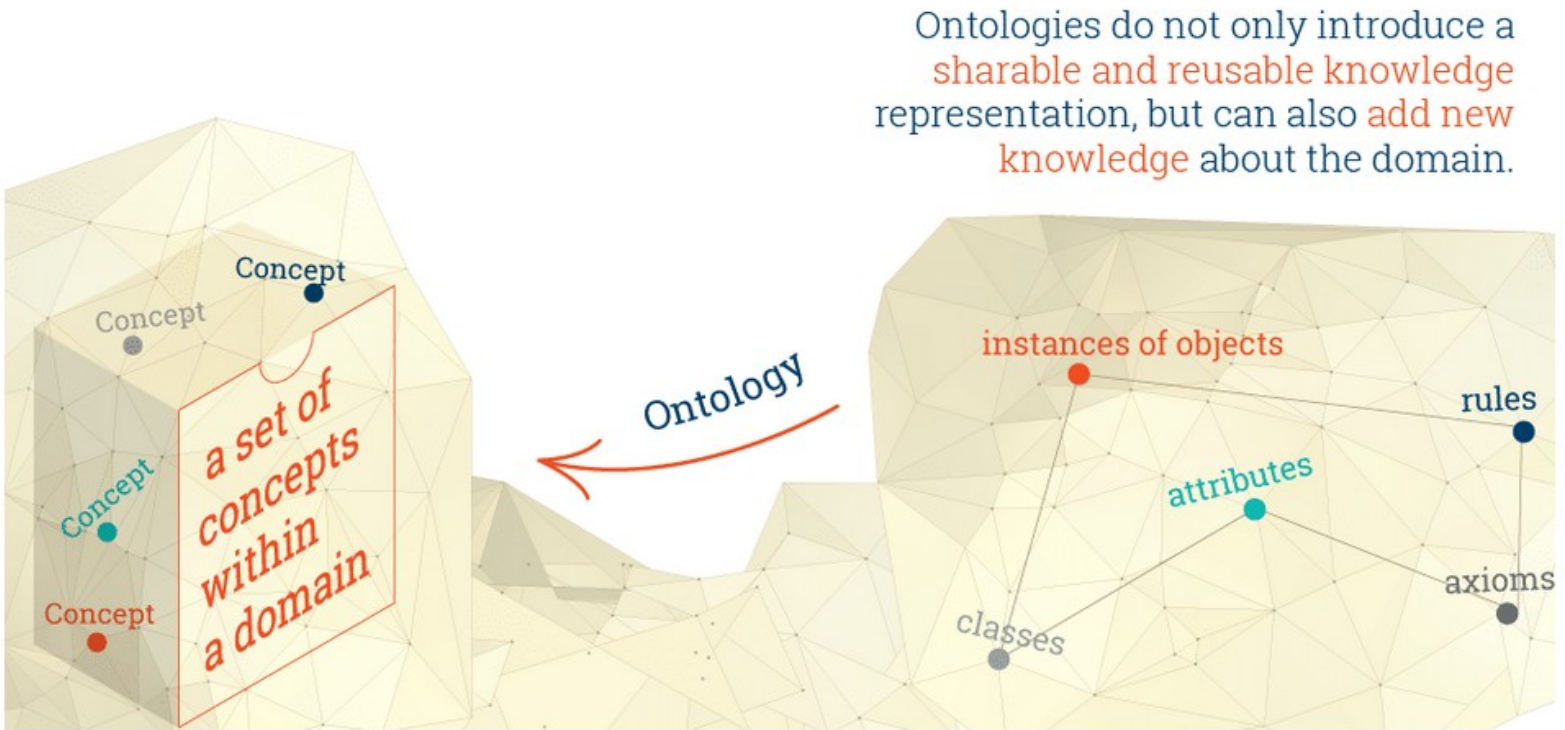
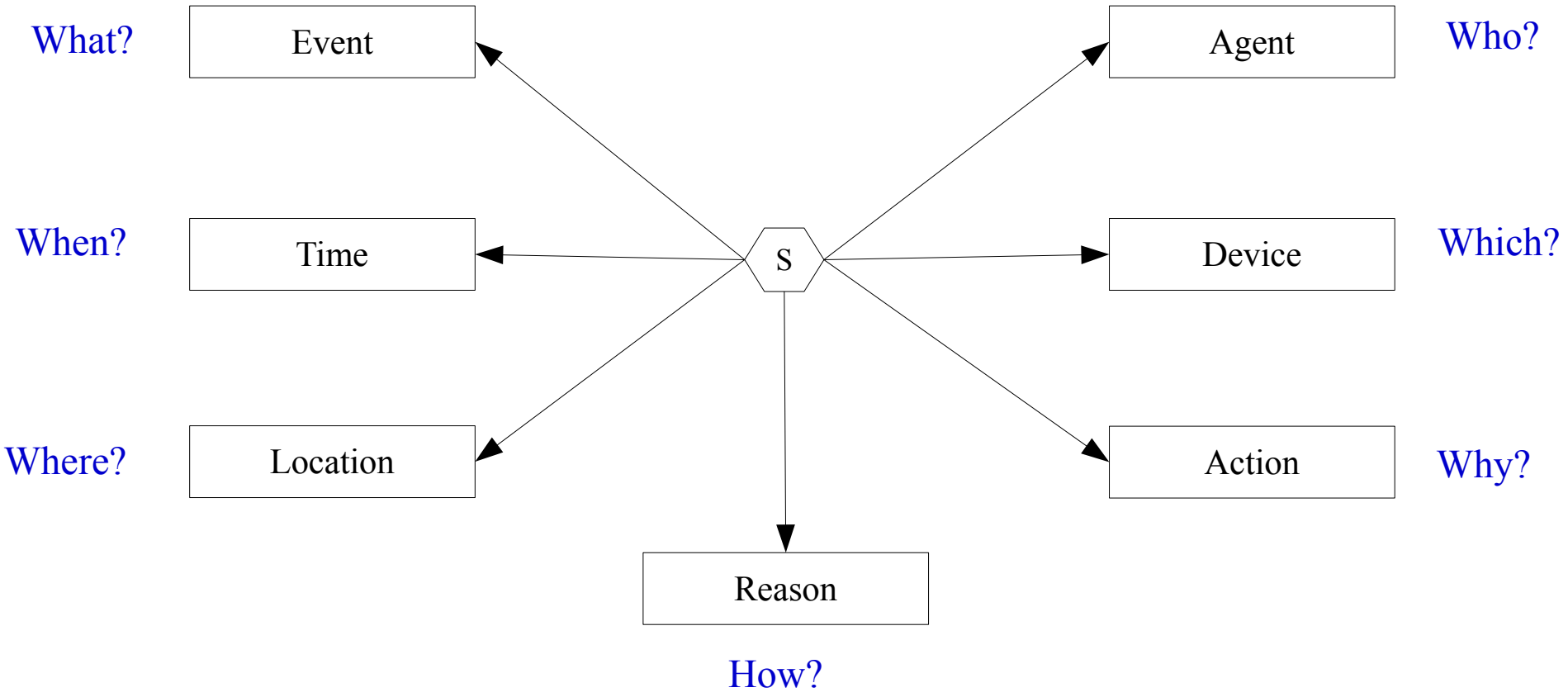


*Image Source: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

Representation of semantics of provenance

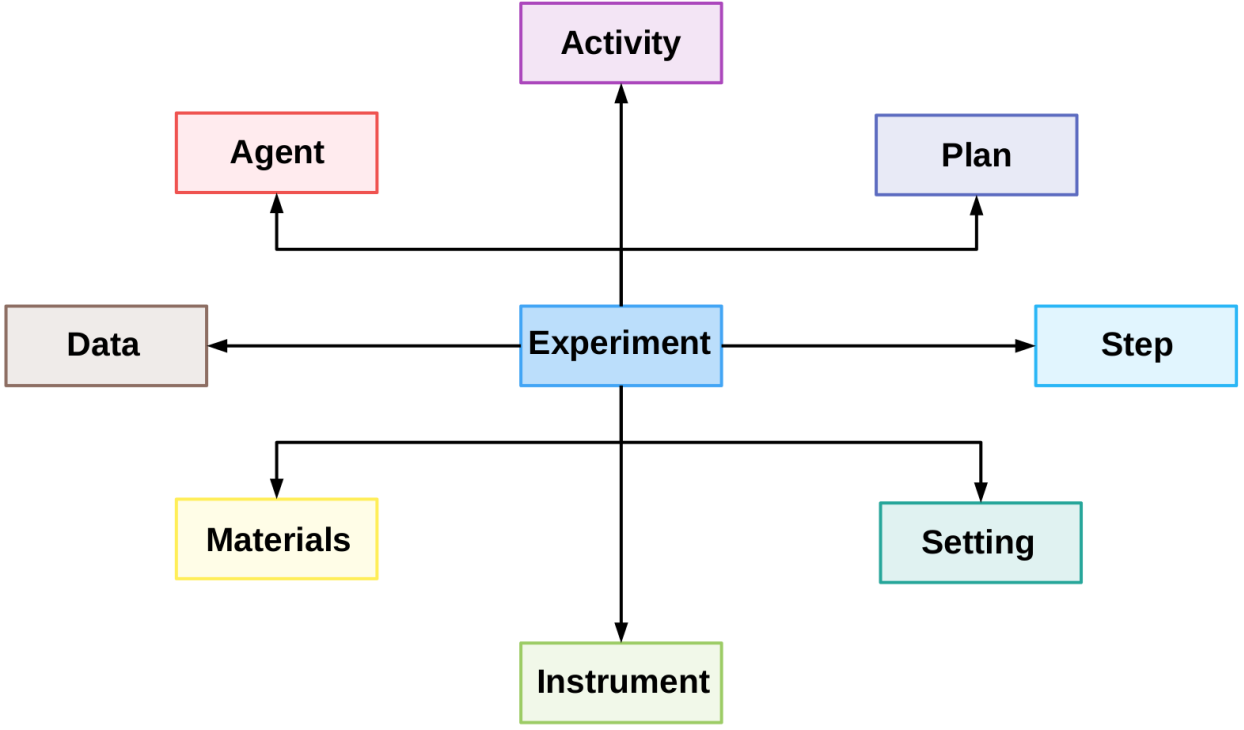


TECHNISCHE UNIVERSITÄT CHEMNITZ

The REPRODUCE-ME Data Model

Experiment is defined as an n-tuple

E = (Data, Agent, Activity, Plan, Step, Setting, Instrument, Material)



[A provenance-based semantic approach to support understandability, reproducibility, and reuse of scientific experiments, Samuel S., 2019]

PROV-O: Ontology to represent provenance

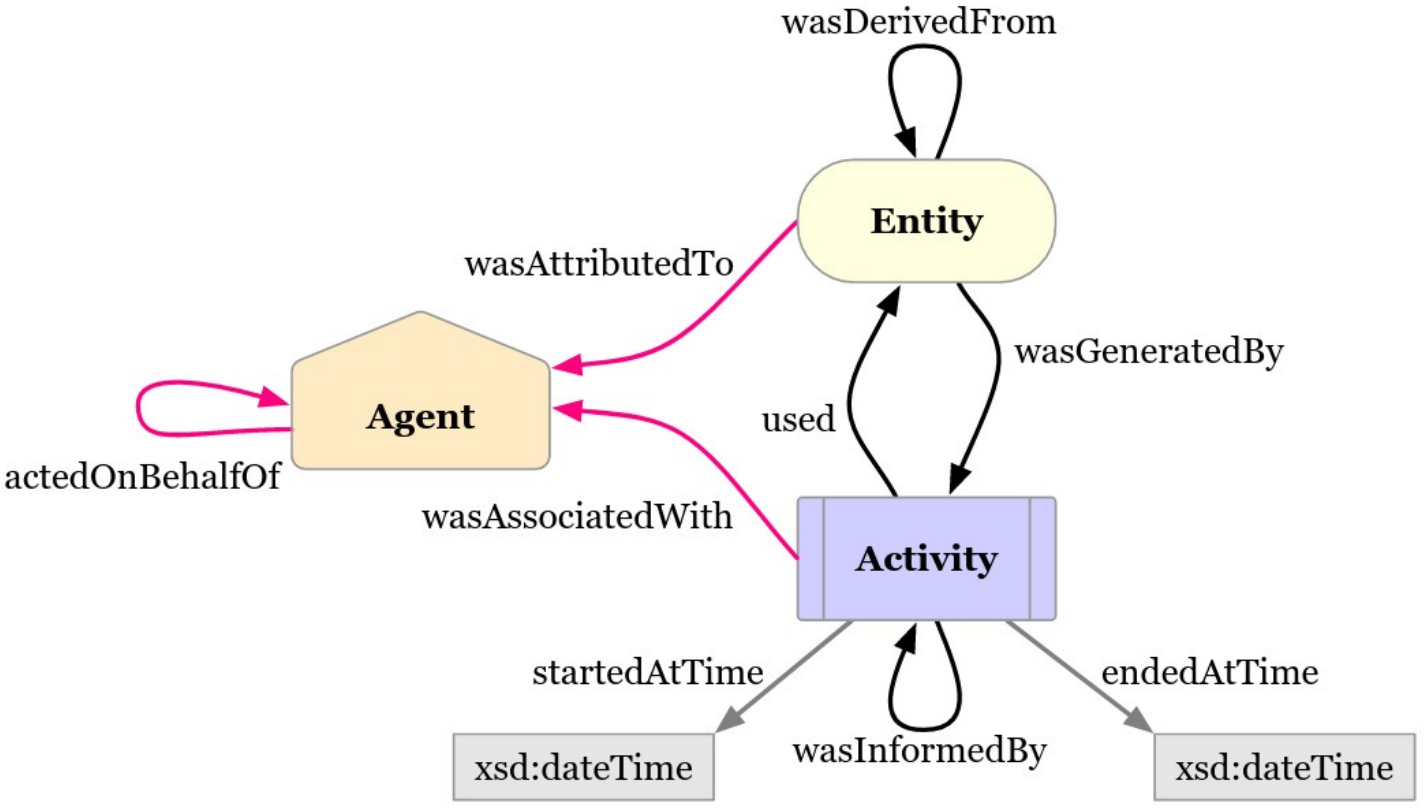


*Image Source: https://www.w3.org/TR/prov-o/

PROV-O: Ontology to represent provenance

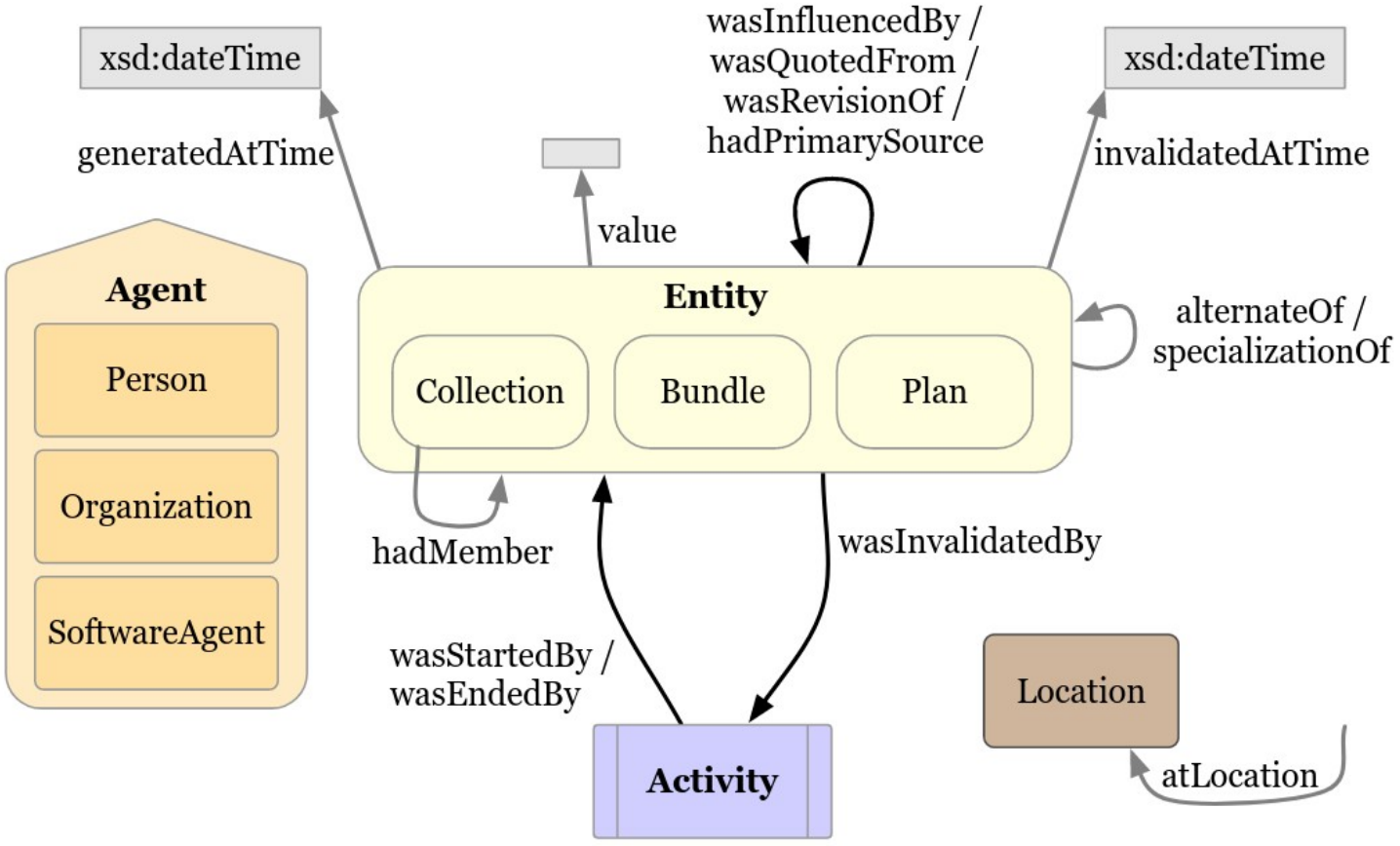


*Image Source: https://www.w3.org/TR/prov-o/

P-Plan: Ontology to represent plans

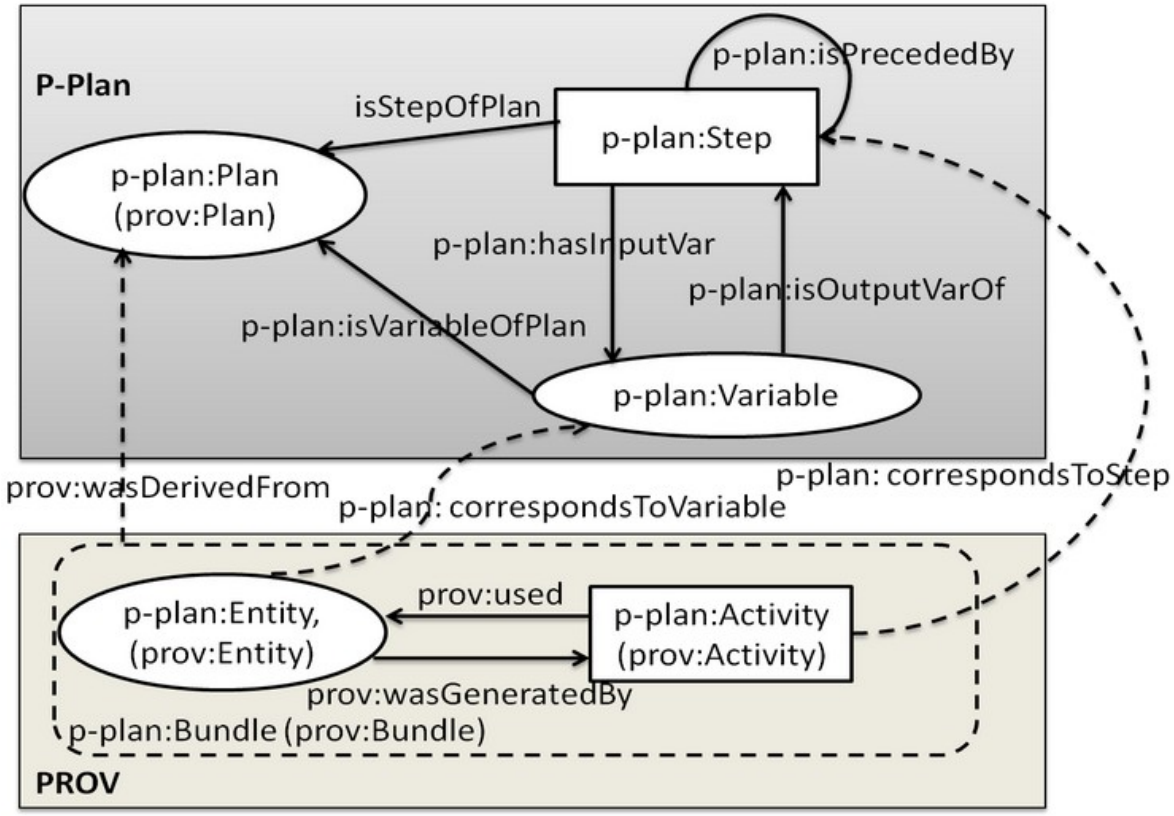
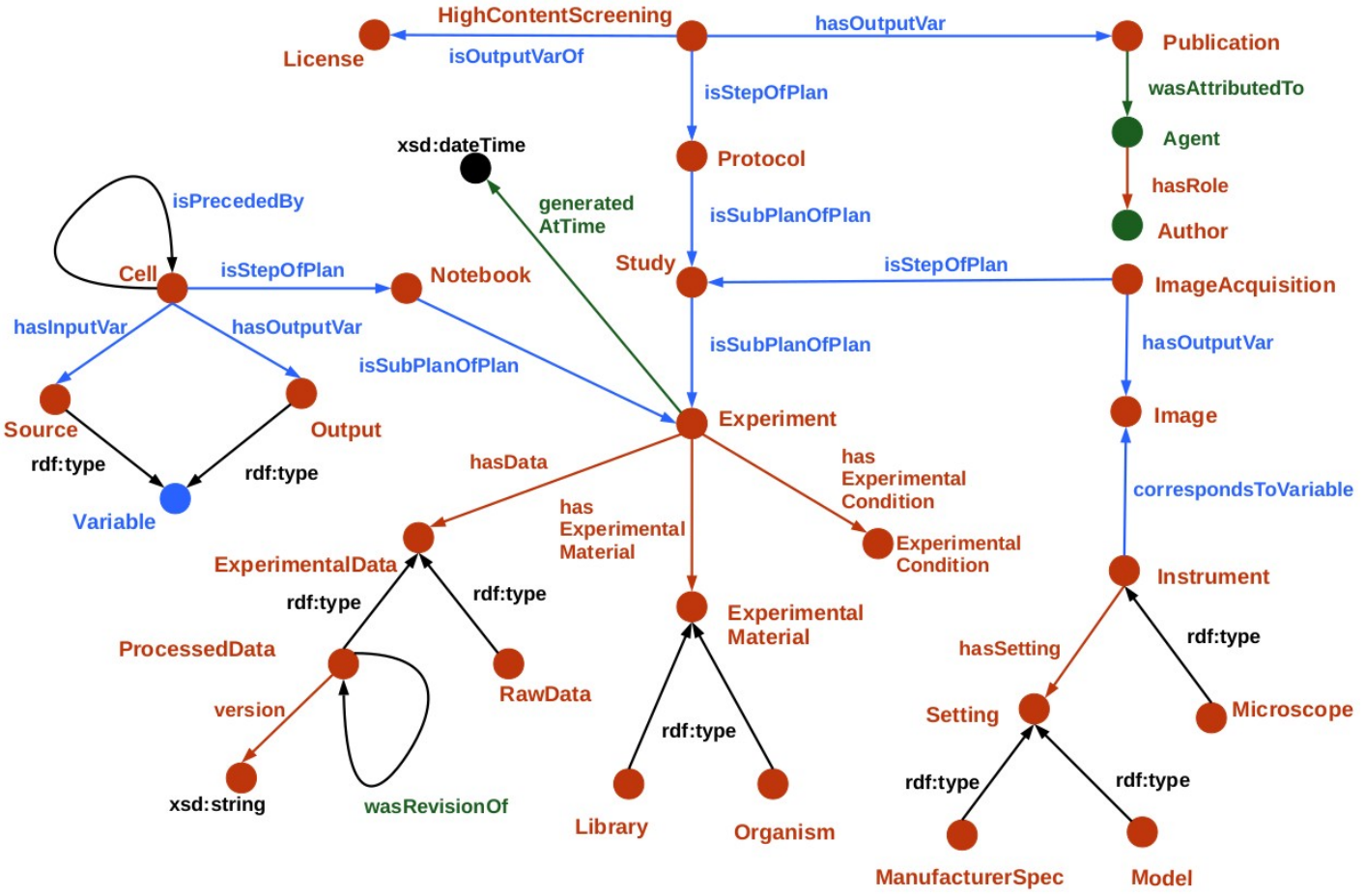


*Image Source: http://purl.org/net/p-plan

TECHNISCHE UNIVERSITÄT CHEMNITZ

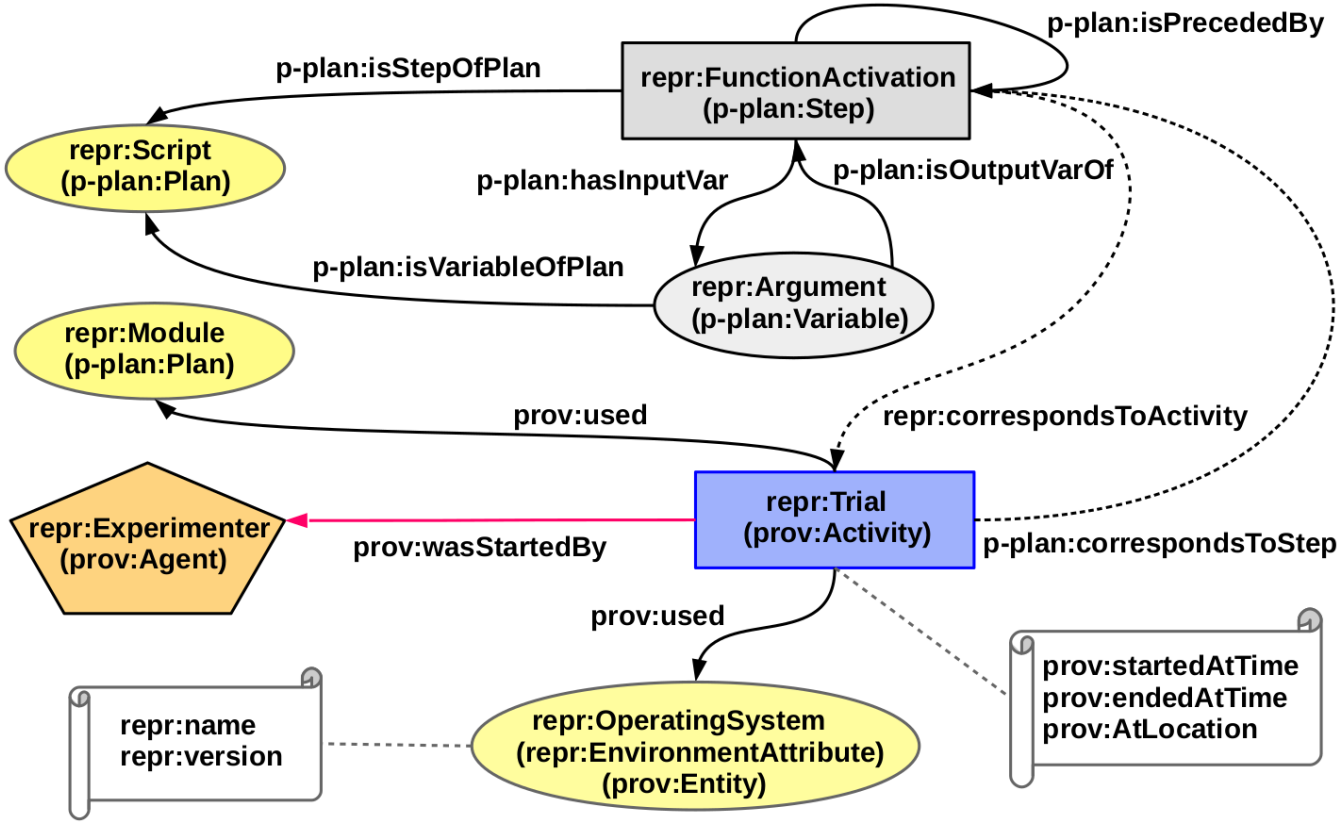
The REPRODUCE-ME Ontology



[Samuel and König-Ries 2017, Samuel and König-Ries 2018a, Samuel et al. 2018, Samuel et al. 2021]

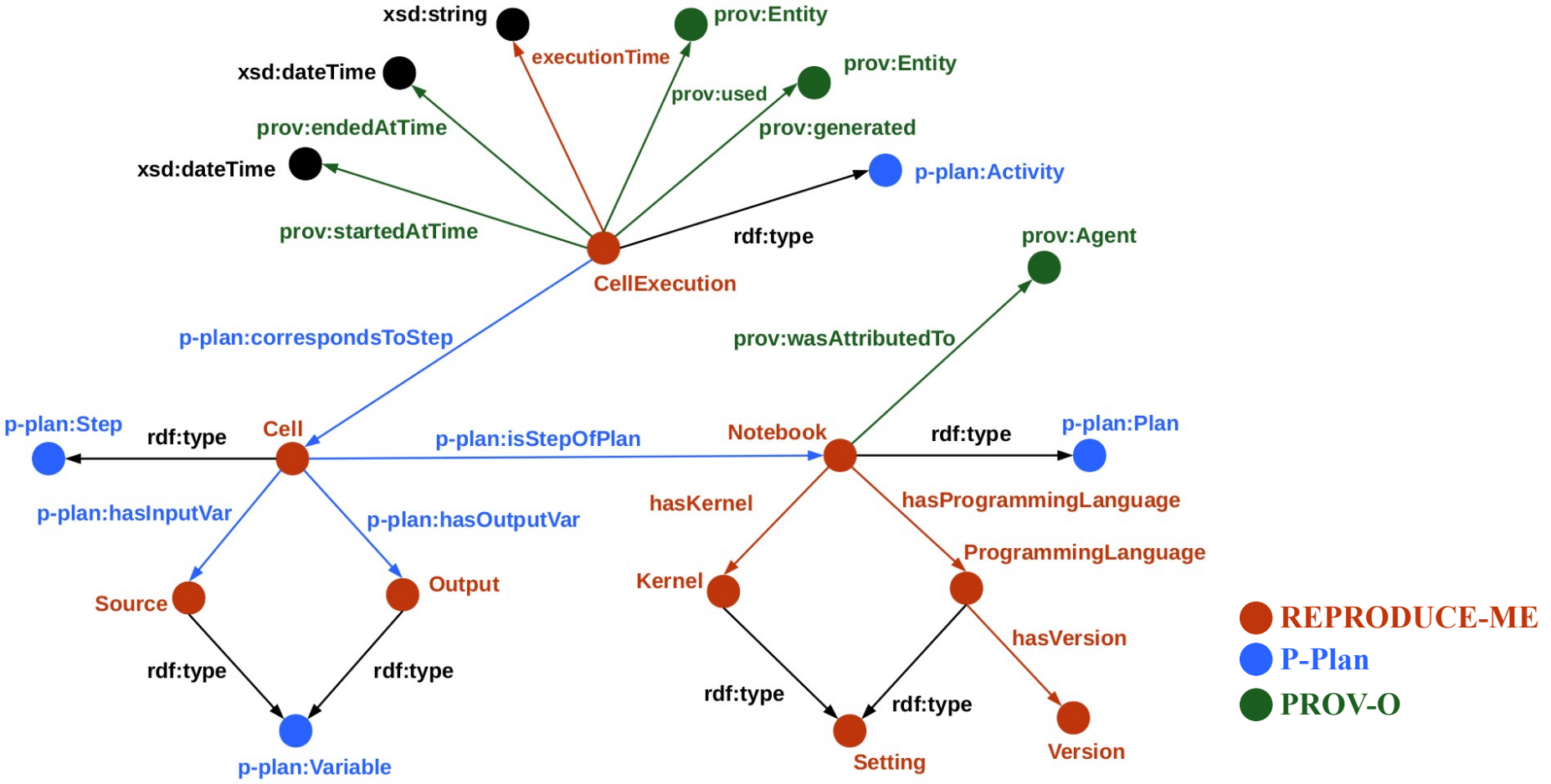
Script Provenance

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



Overview of the ontology terms to model script provenance

Computational Provenance

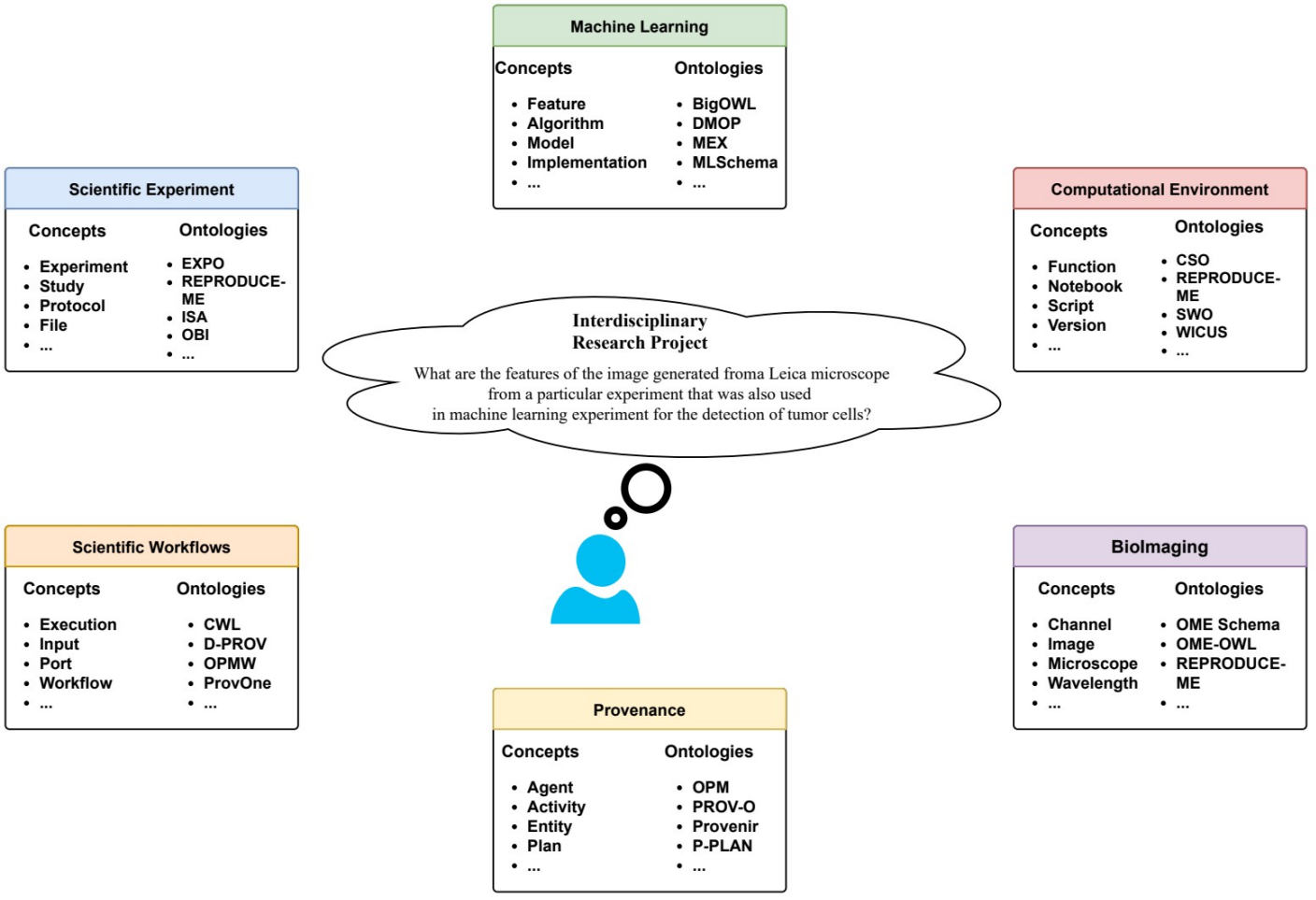


[Samuel and König-Ries 2018a, Samuel and König-Ries 2018b]

The ReproduceMe Ontology Network (ReproduceMeON)

TECHNISCHE UNIVERSITÄT CHEMNITZ

Reproducibility related area assimilation



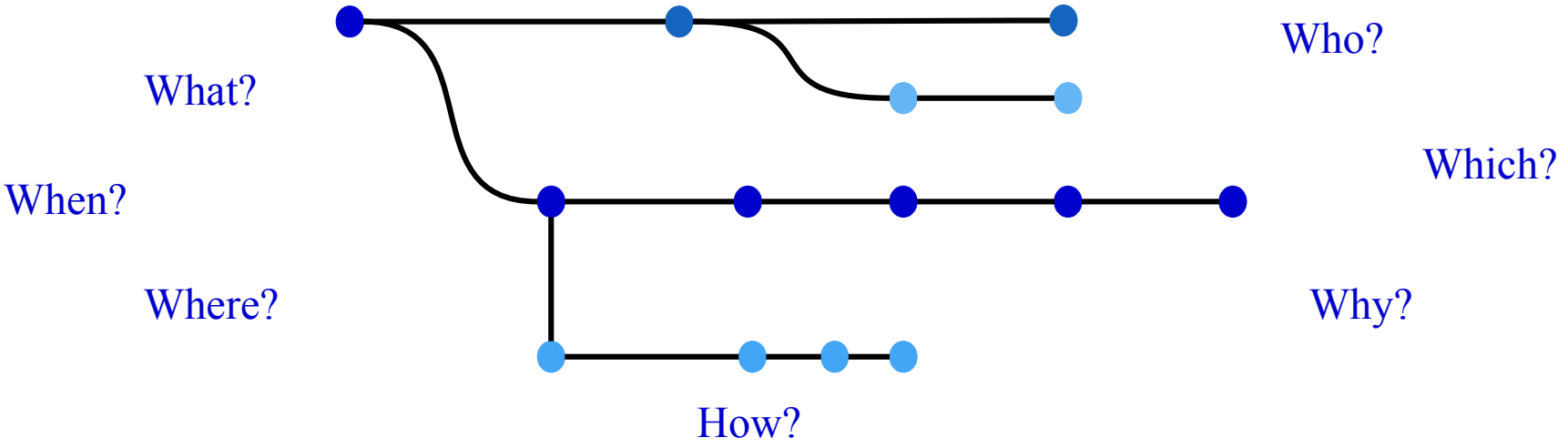
ReproduceMeON

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

[Towards an Ontology Network for the reproducibility of scientific studies., Samuel, Algergawy, König-Ries. 2021]

Capturing 6W1H Provenance of Computational Experiments

How should I capture and track the provenance of my computational experiments?



Computational Reproducibility

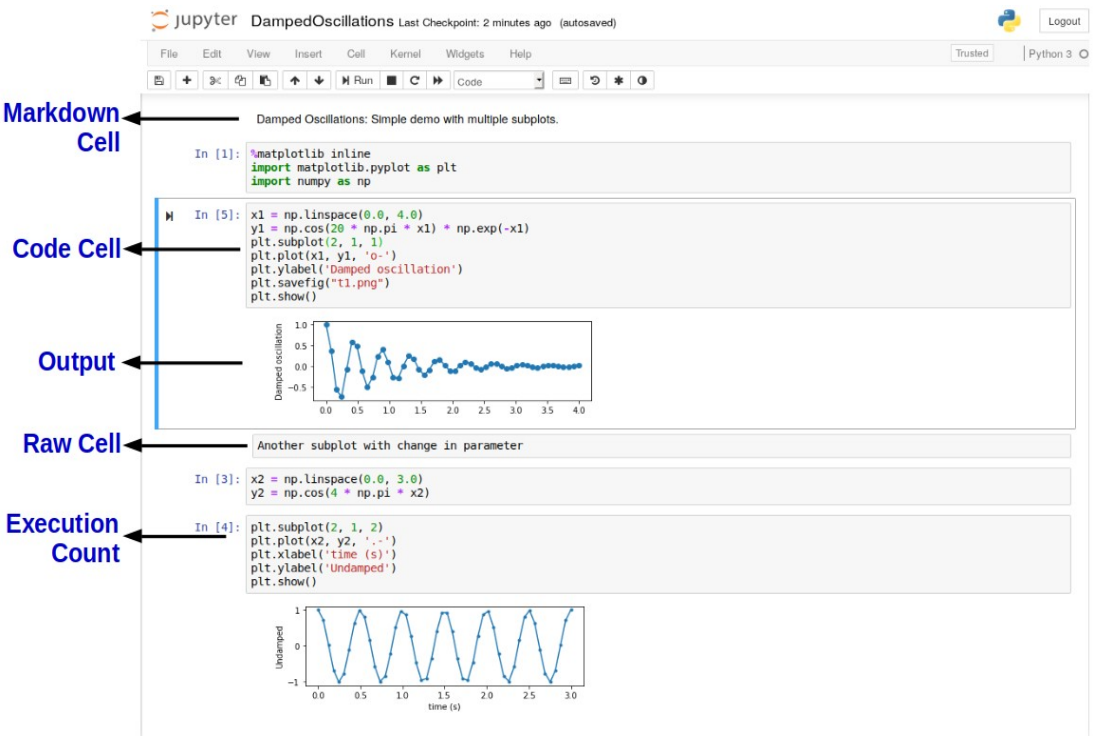
Computational Notebooks

- 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

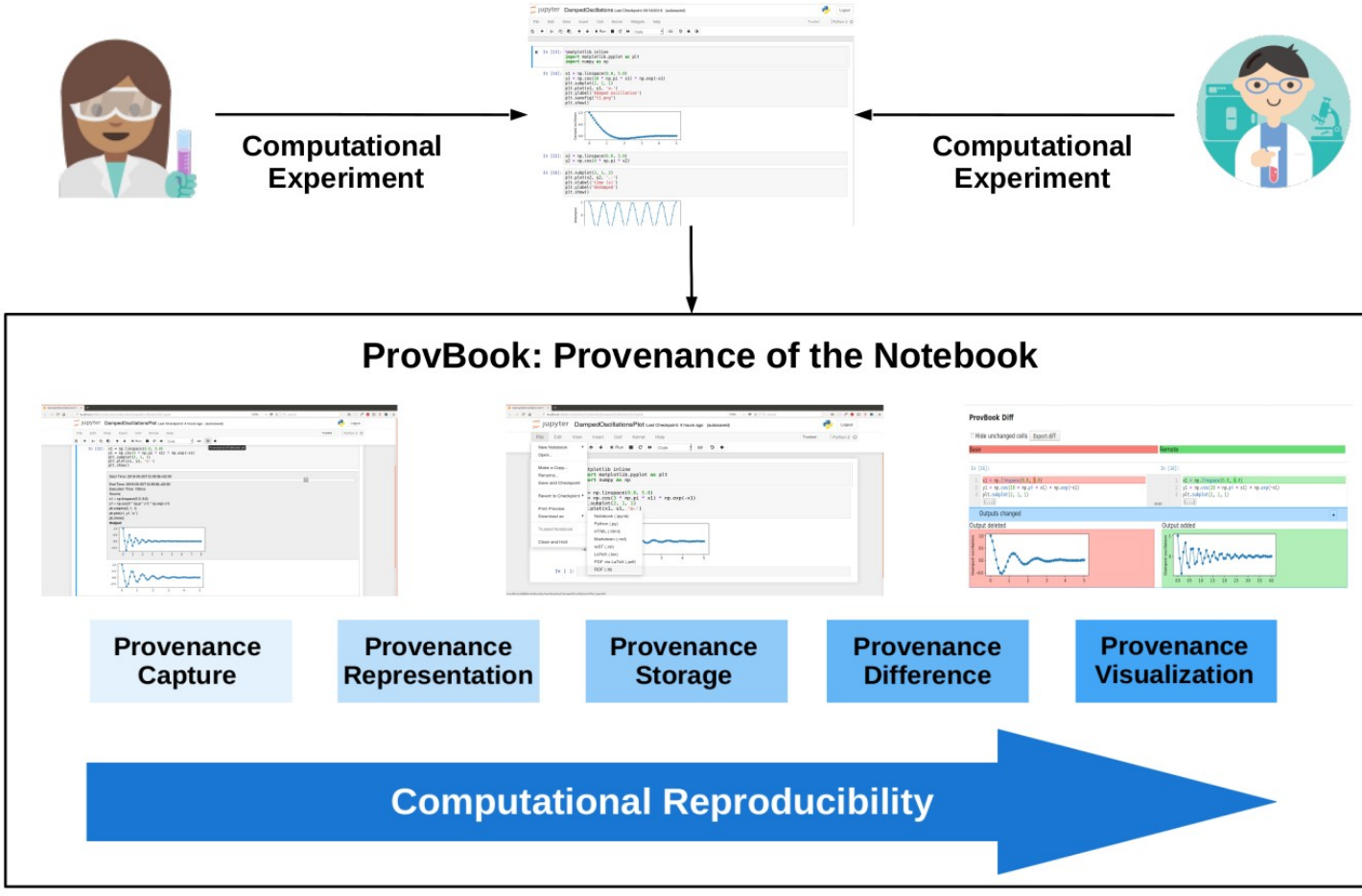


Computational Reproducibility

- 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

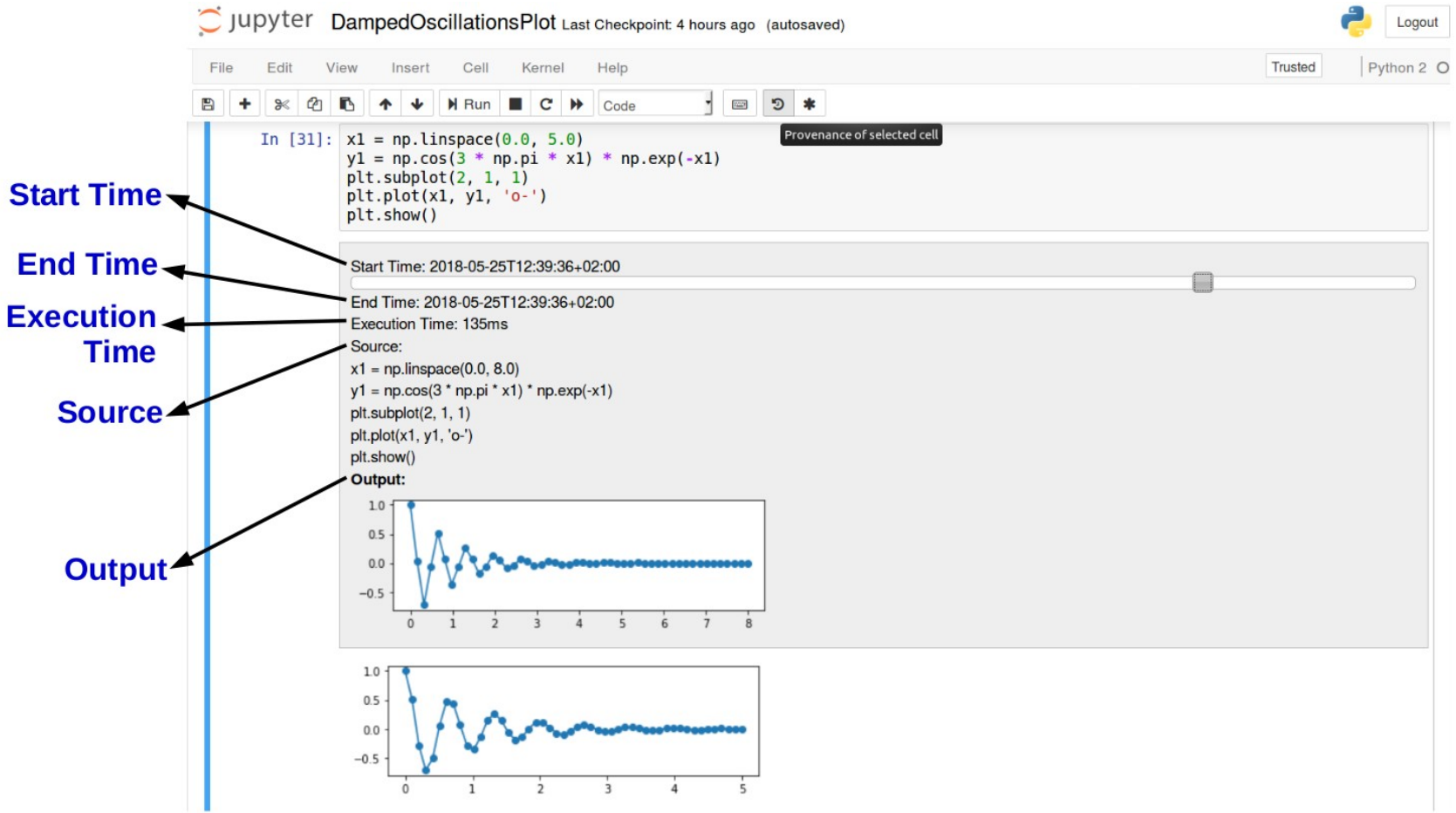
ProvBook: Provenance of the Notebook



[ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility, Samuel et al., 2018]

TECHNISCHE UNIVERSITÄT CHEMNITZ

ProvBook: Capture & Visualization



https://github.com/Sheeba-Samuel/ProvBook

TECHNISCHE UNIVERSITÄT CHEMNITZ

ProvBook: Difference

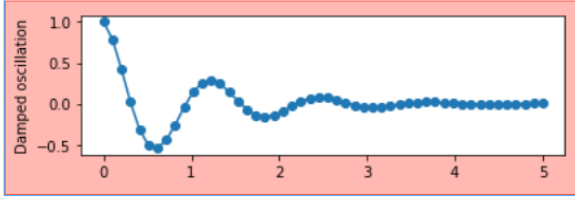
ProvBook Diff

☐ Hide unchanged cells | Export diff

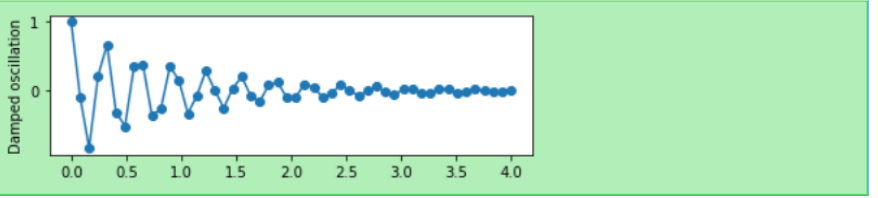
Base	Remote
In [16]:	In [16]:
1 x1 = np.linspace(0.0, 5.0)	1 x1 = np.linspace(0.0, 4.0)
2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)	2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)
3 plt.subplot(2, 1, 1)	3 plt.subplot(2, 1, 1)
(...)	(...)

Outputs changed

Output deleted



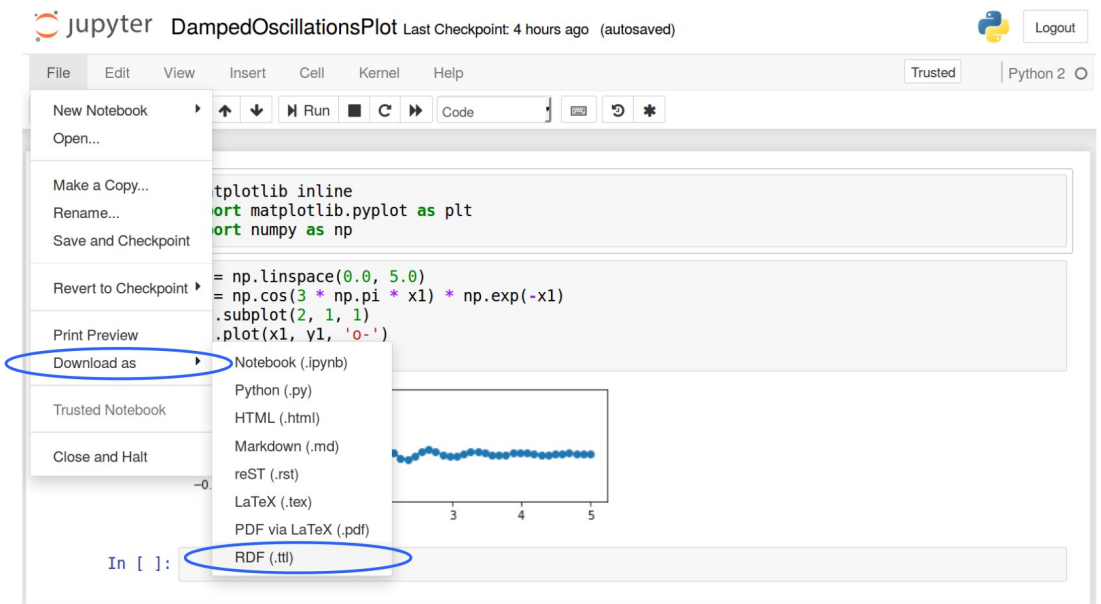
Output added

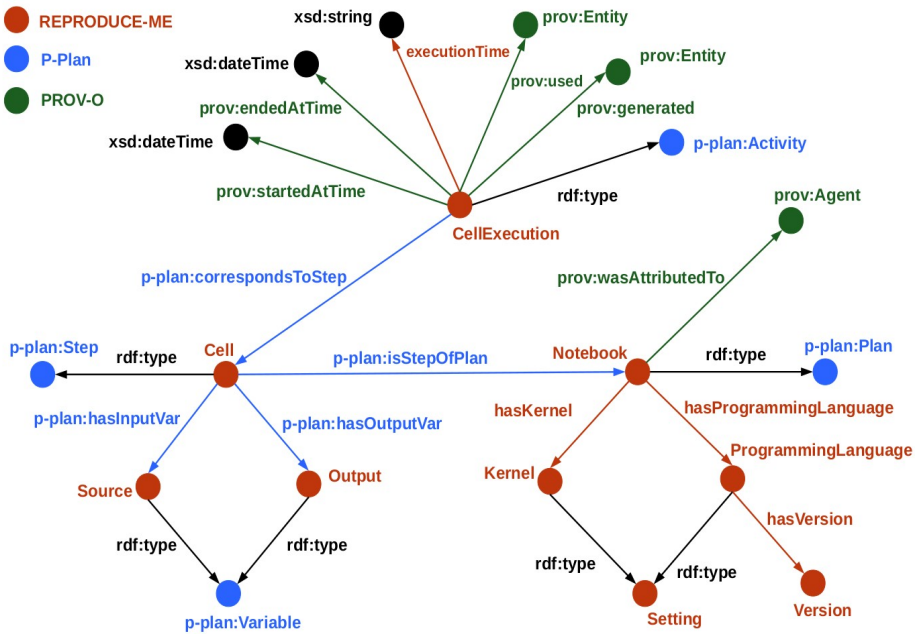


TECHNISCHE UNIVERSITÄT CHEMNITZ

ProvBook: Semantic Representation

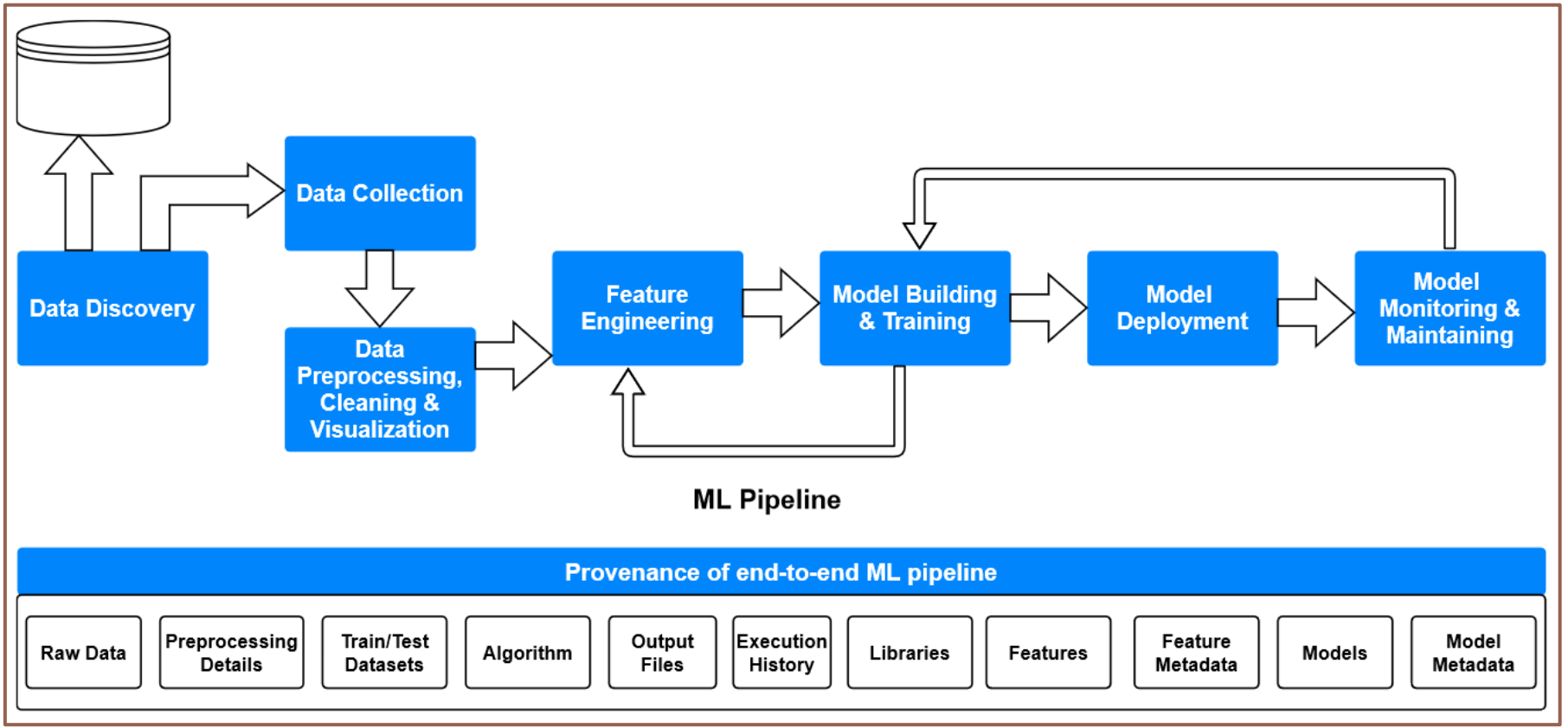
- Jupyter Notebook and its provenance described using the REPRODUCE-ME ontology
- **New extension**: Export in RDF from the user interface or command line





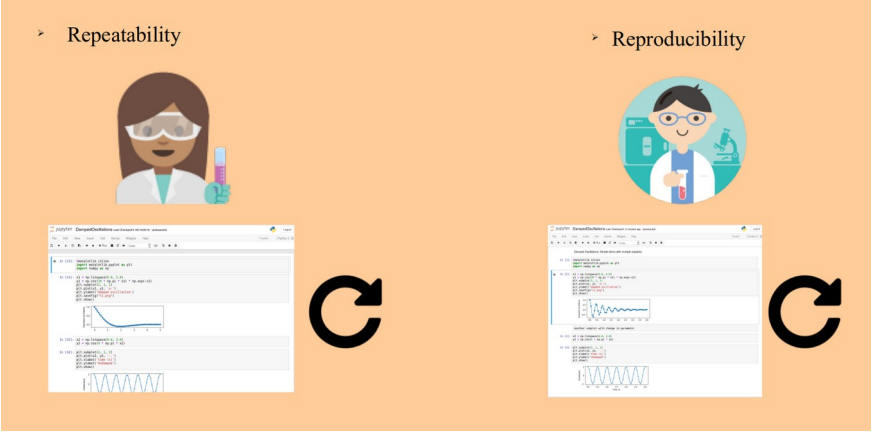
[Samuel and König-Ries 2018a]

Provenance of ML Pipelines



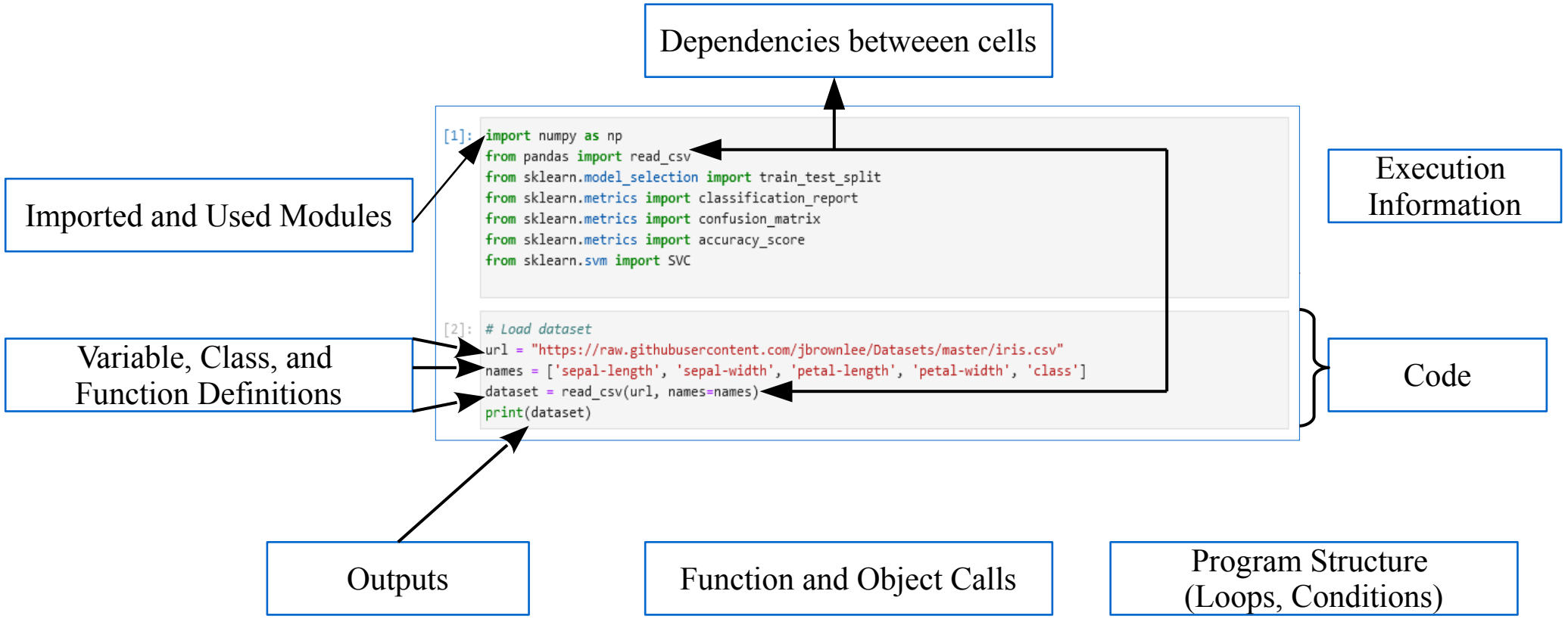
MLProvLab

- 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.



Towards Tracking Provenance from Machine Learning Notebooks, Kerzel, Samuel, and König-Ries, 2021

Provenance Capture



https://github.com/fusion-jena/MLProvLab

Modules and Libraries

evaluation_notebook.ipynb

Download MLProvLab

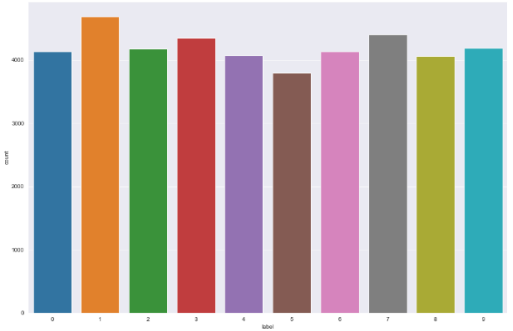
[5]:
```
X=df.drop("label",axis=1).values
y=df["label"].values
```

[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(
print(X_train.shape)
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(39900, 784)
(2100, 784)
```

Provenance: evaluation_not

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

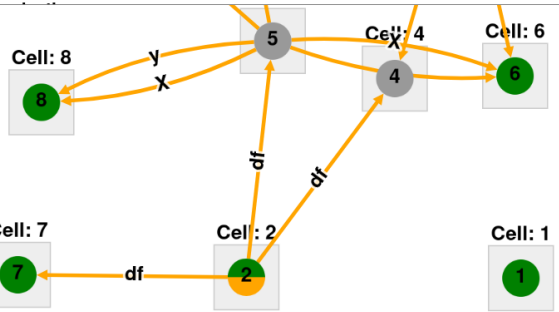
Import **matplotlib.pyplot** with alias **plt** was used **3** times

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

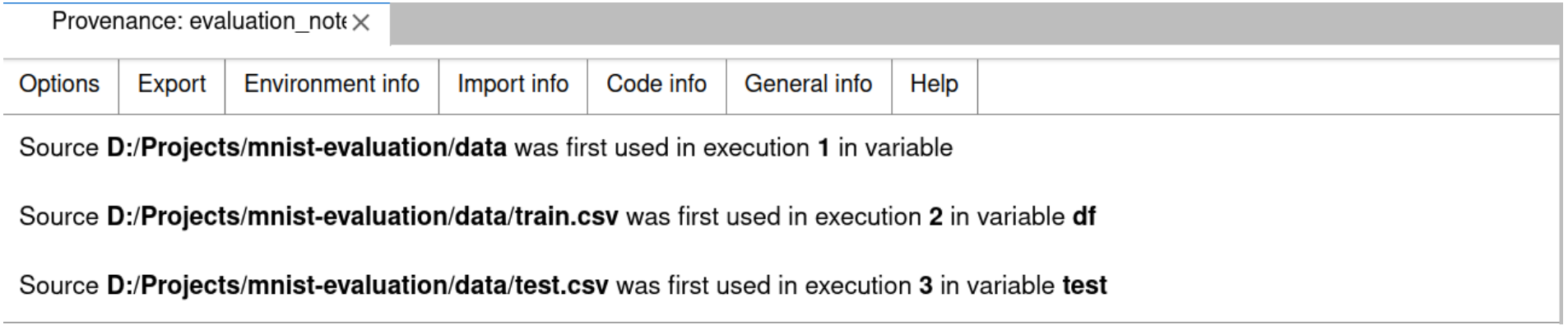


Epoch **4/4** Thu, 24 Jun 2021 10:19:30 GMT

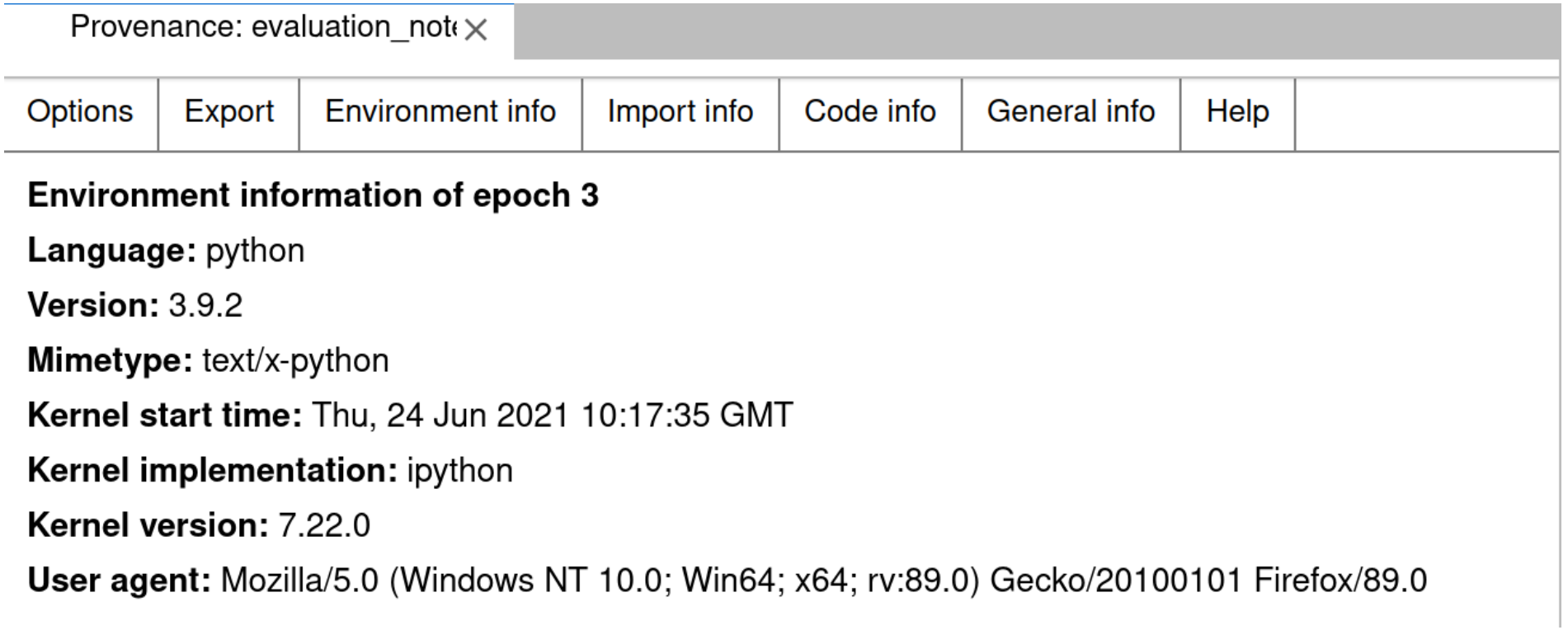
Execution **19/19** Thu, 24 Jun 2021 10:26:21 GMT

TECHNISCHE UNIVERSITÄT
CHEMNITZ

Datasets

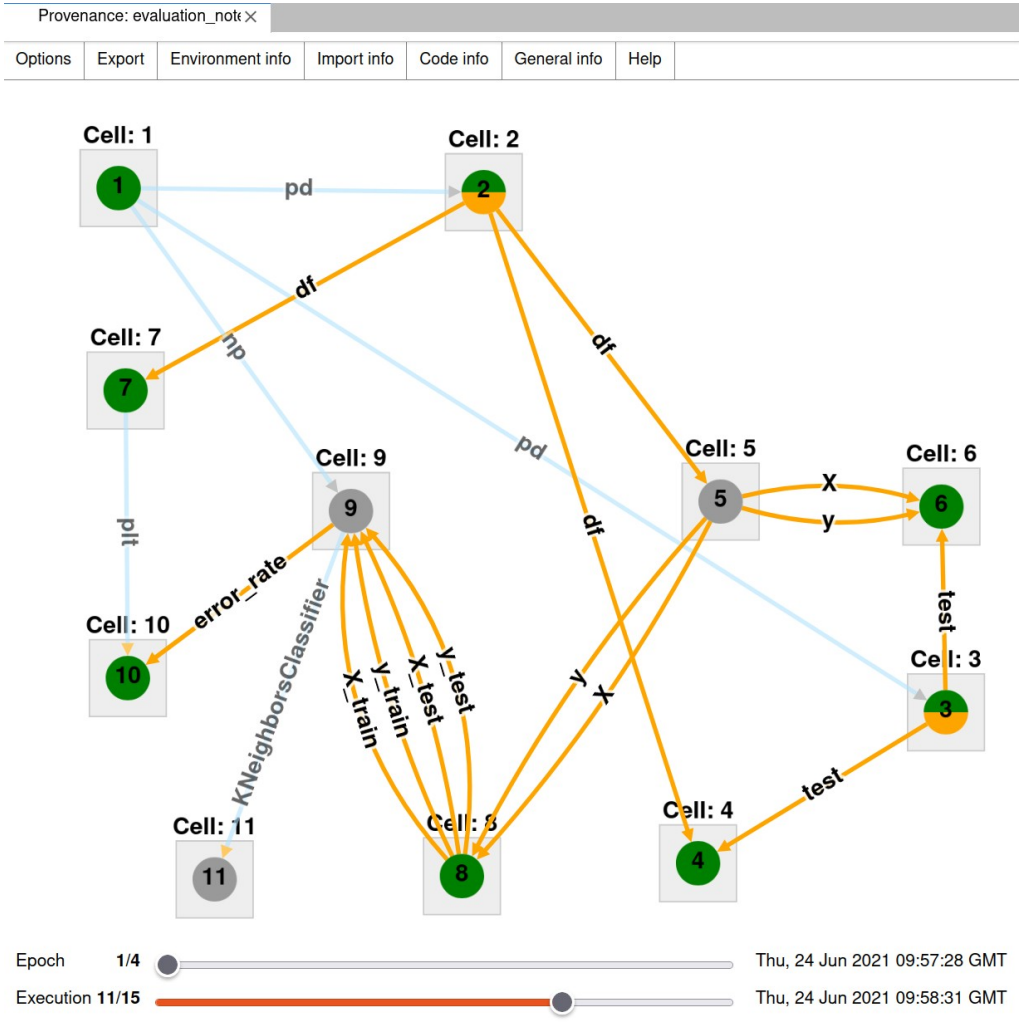


Execution Environment



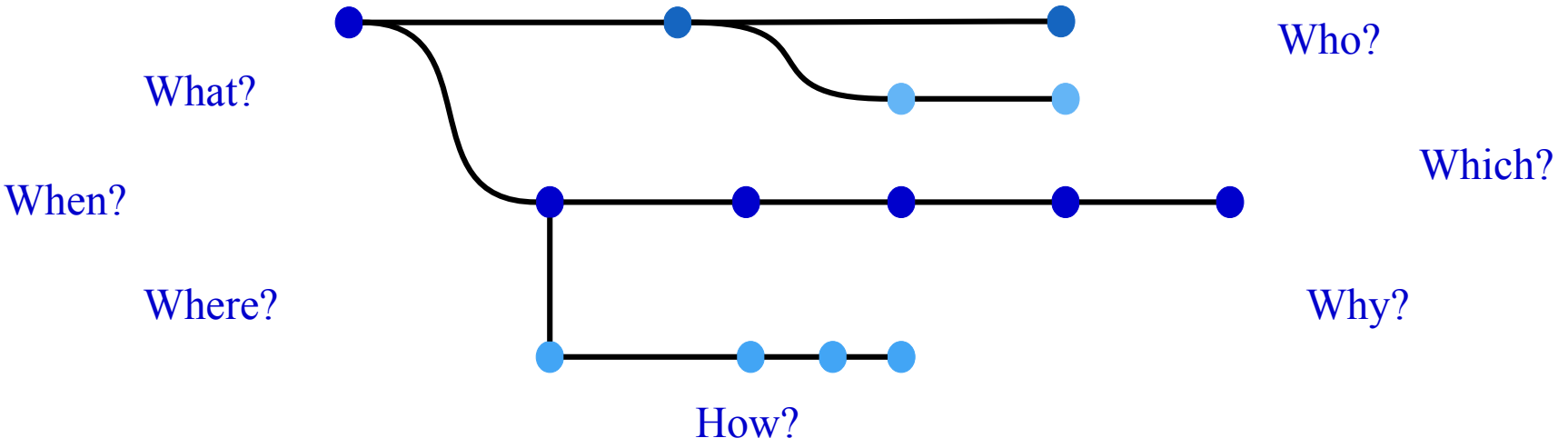
TECHNISCHE UNIVERSITÄT CHEMNITZ

MLProvLab: Provenance Visualization



Analyzing 6W1H Provenance of Experiments

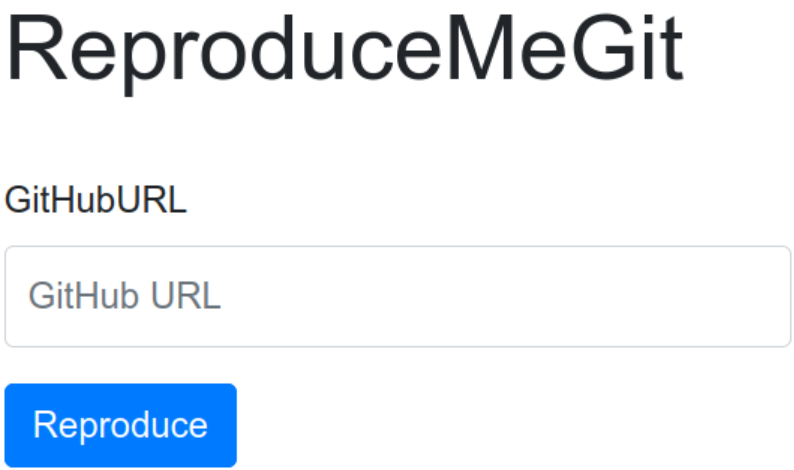
Is it possible to reproduce and analyze others experiments?



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



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

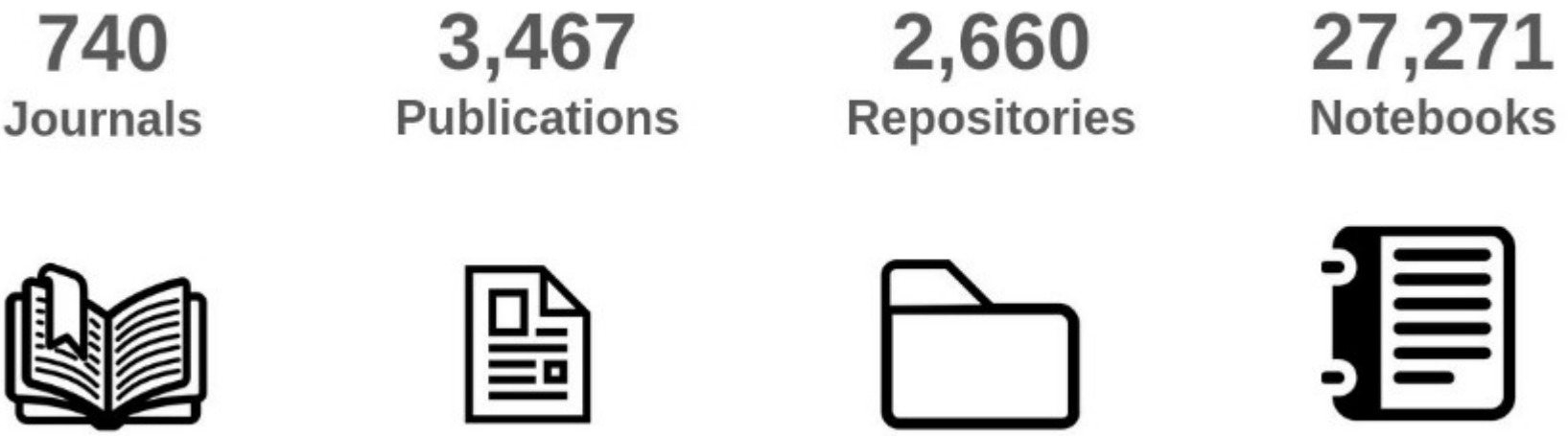
- Repository Overview
- Notebook Overview
- Cells Overview
- Modules Overview
- Distribution of programming Languages and the versions used

https://github.com/fusion-jena/ReproduceMeGit

TECHNISCHE UNIVERSITÄT CHEMNITZ

Computational reproducibility of Jupyter notebooks from biomedical publications

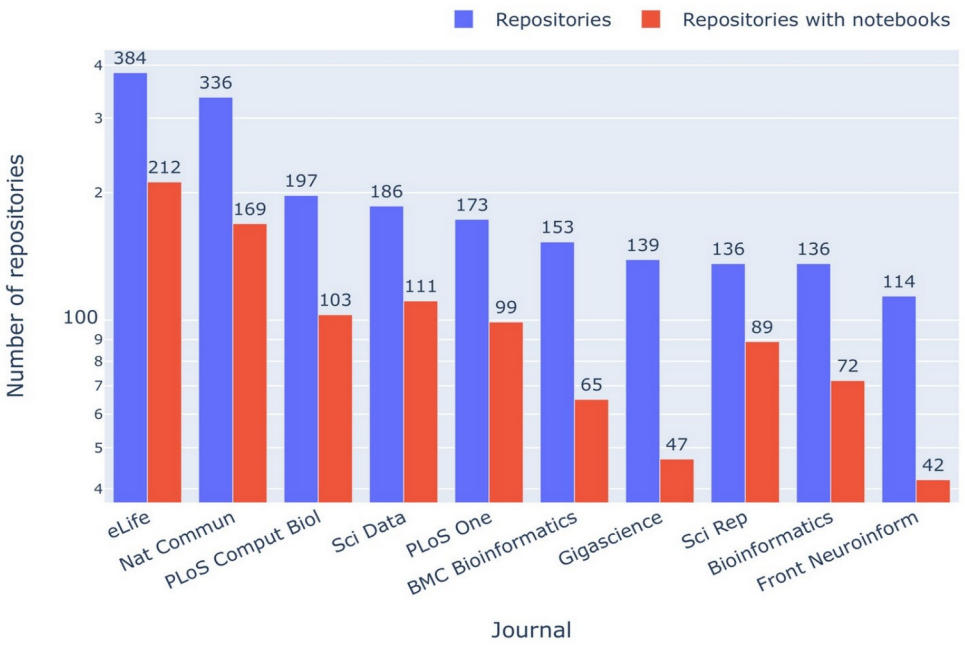
- Using fully automated workflows, we analyzed the computational reproducibility of Jupyter notebooks related to publications indexed in PubMed Central (Data till Mar 2023).



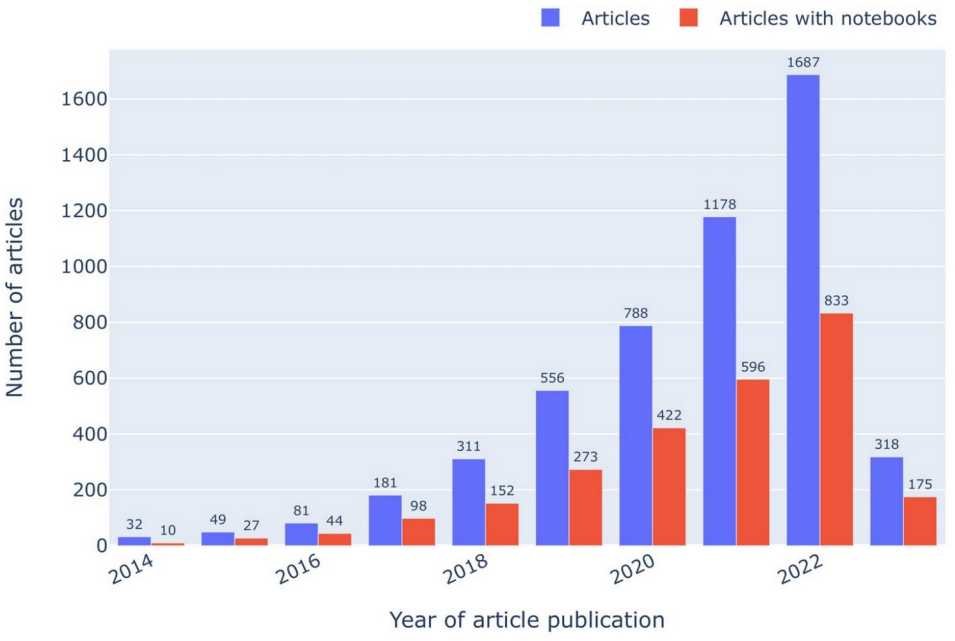
Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024. https://doi.org/10.1093/gigascience/giad113

TECHNISCHE UNIVERSITÄT CHEMNITZ

Results



(1)

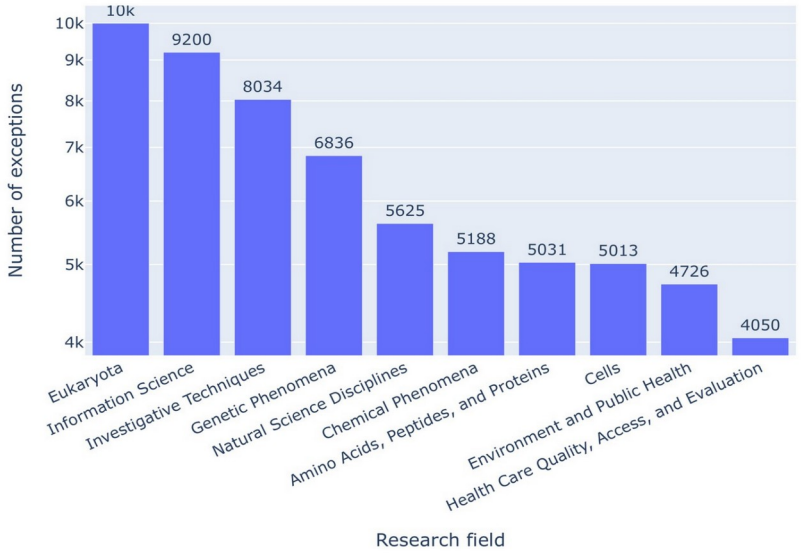


(2)

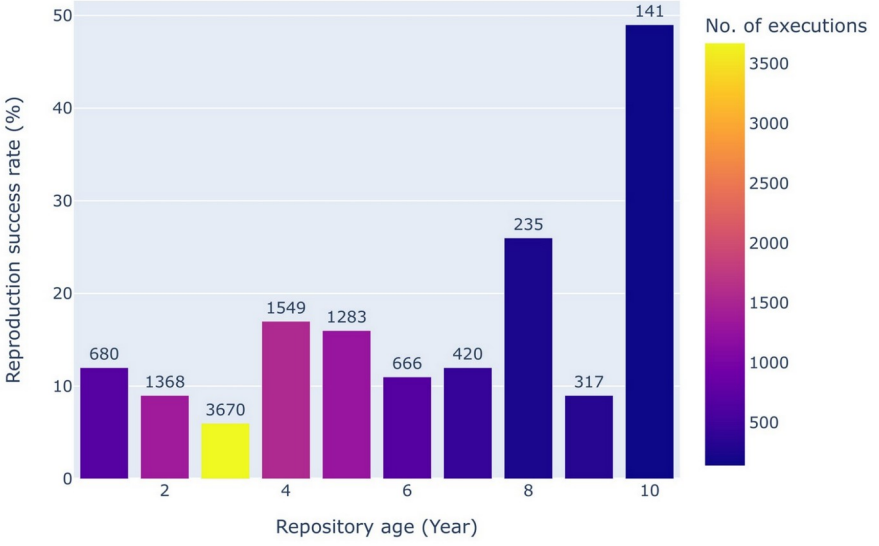
Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024 https://doi.org/10.1093/gigascience/giad113

TECHNISCHE UNIVERSITÄT CHEMNITZ

Results



(3)



(4)

Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024 https://doi.org/10.1093/gigascience/giad113

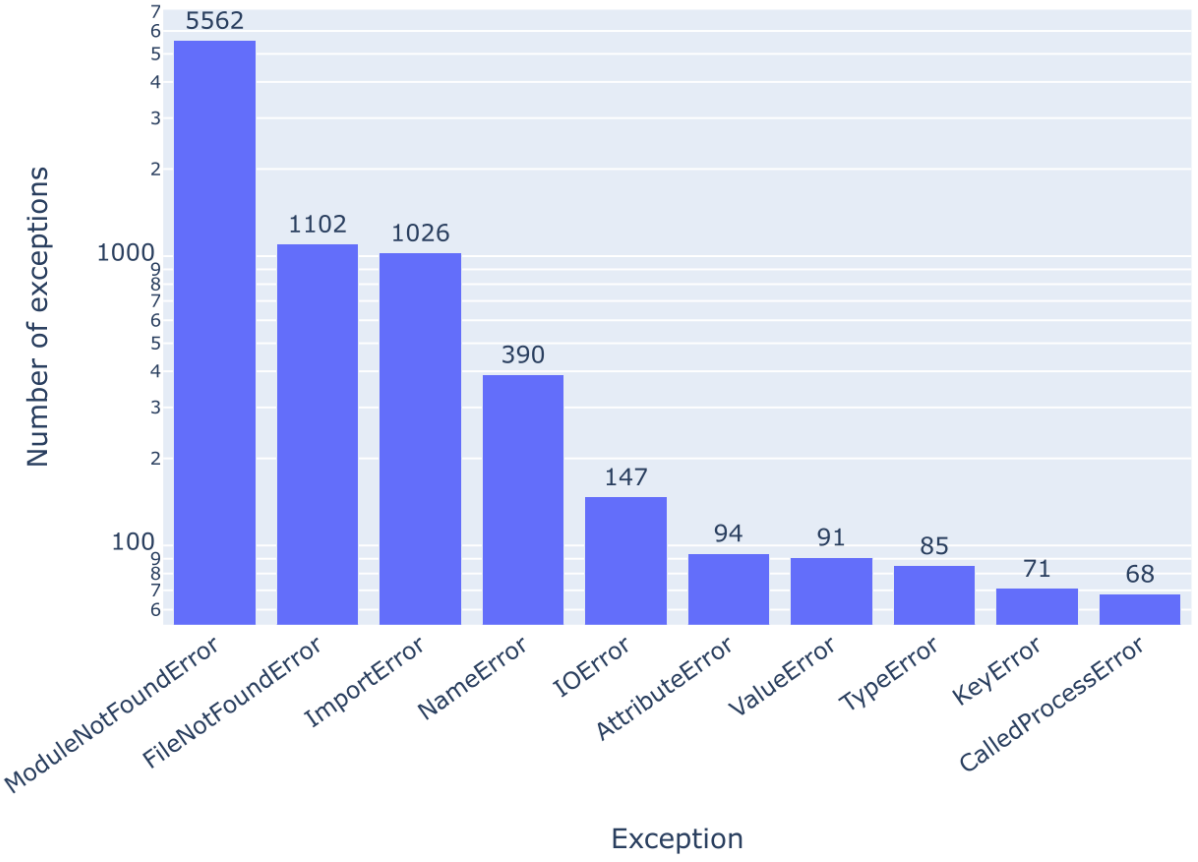
Results: Programming Languages

Notebook language	Rerun	Initial run	Ratio rerun/initial
Matlab	134	9	14.9
Julia	295	59	5.0
Unknown	3,112	720	4.3
Python	22,578	8,160	2.8
R	891	461	1.9
Bash	36	24	1.5
Sos	29	24	1.2
Groovy	95	95	1.0
Scala	29	29	1.0
Java	8	8	1.0

Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024 https://doi.org/10.1093/gigascience/giad113

TECHNISCHE UNIVERSITÄT CHEMNITZ

Results: Notebook Reproducibility



Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024 https://doi.org/10.1093/gigascience/giad113

TECHNISCHE UNIVERSITÄT CHEMNITZ

Results: Notebook Reproducibility

Features	Notebooks with `different` results	Notebooks with `identical` results
Number of notebooks	324 (⊠ 151)	879 (⊠ 245)
setup.py	0 (⊠ 0)	344 (⊠ 98)
requirement.txt	1 (⊠ 0)	353 (⊠ 107)
pipfile	0 (⊠ 0)	0 (⊠ 0)
Total cells	23 (⊠ 17.9)	19.6 (⊠ 17.1)
Code cells	15.4 (⊠ 12.3)	11.3 (⊠ 9.8)
Markdown cells	7.6 (⊠ 5.6)	8.3 (⊠ 6.7)
Ratio of Markdown vs. code cells	0.49 (⊠ 0.46)	0.73 (⊠ 0.68)
Empty cells	0.9 (⊠ 0.7)	0.7 (⊠ 0.7)
Differences	6.3 (⊠ 5.3)	0 (⊠ 0)
Execution time (s)	18.3 (⊠ 22.1)	57.6 (⊠ 16.4)
Execution time per code cell (s)	1.88 (⊠ 1.80)	5.09 (⊠ 1.67)

Computational reproducibility of Jupyter notebooks from biomedical publications, Sheeba Samuel and Daniel Mietchen 2024 https://doi.org/10.1093/gigascience/giad113

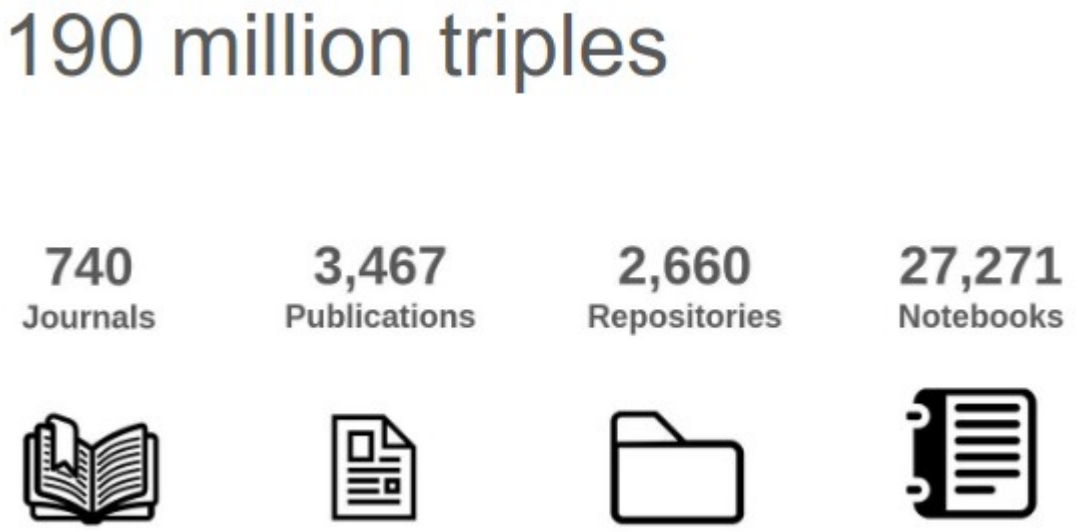
Resources

- Code : https://github.com/fusion-jena/computational-reproducibility-pmc
- Data : https://doi.org/10.5281/zenodo.8226725
- Paper : https://doi.org/10.1093/gigascience/giad113

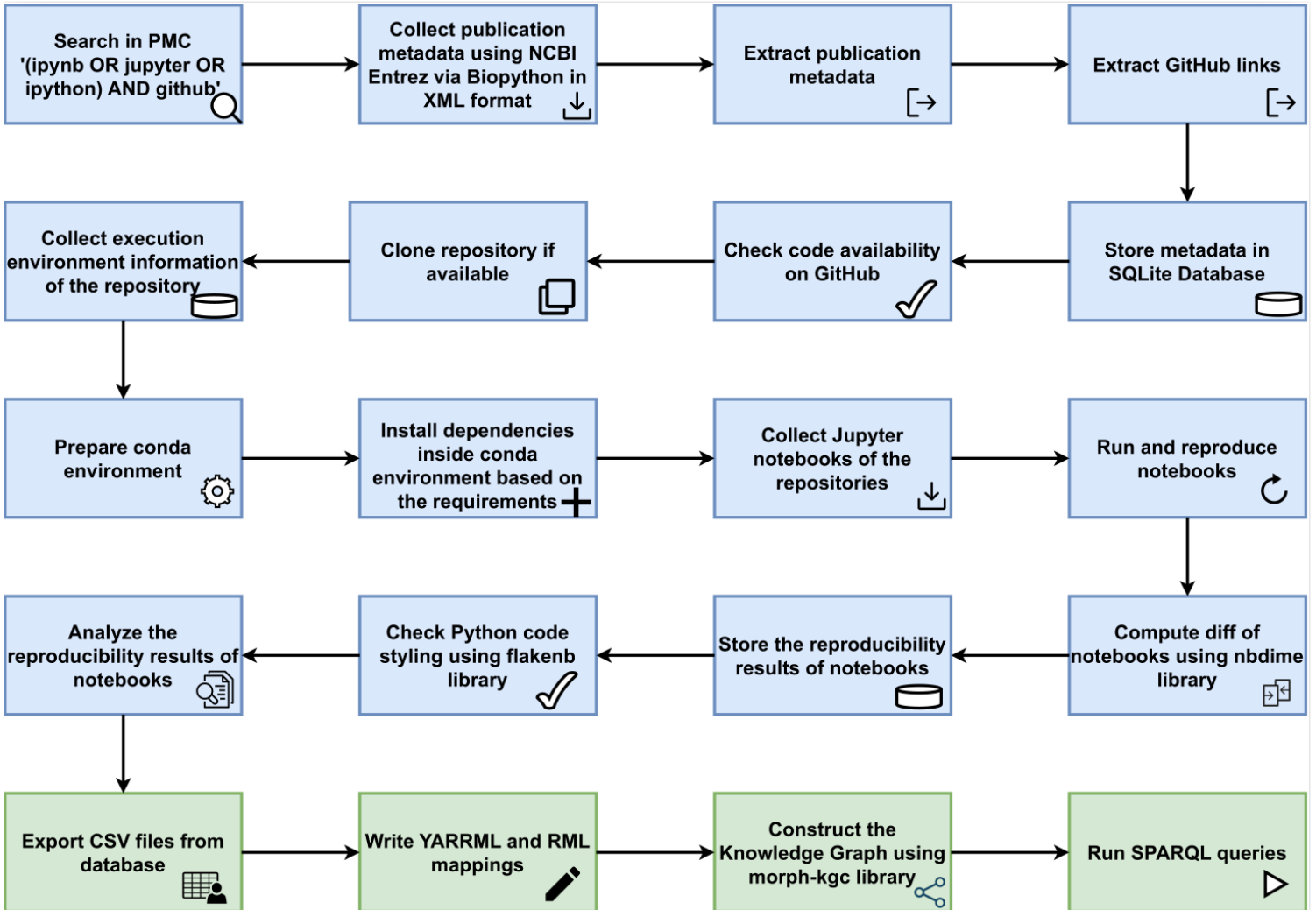
FAIR Jupyter

- A knowledge graph approach to semantic sharing and granular exploration of a computational notebook reproducibility dataset
  - Website: https://w3id.org/fairjupyter
  - SPARQL Endpoint: https://reproduceme.uni-jena.de/#/dataset/fairjupyter/query
  - Knowledge Graph Browser: https://reproduceme.uni-jena.de/fj/
  - Paper: https://doi.org/10.4230/TGDK.2.2.4

FAIR Jupyter Knowledge Graph

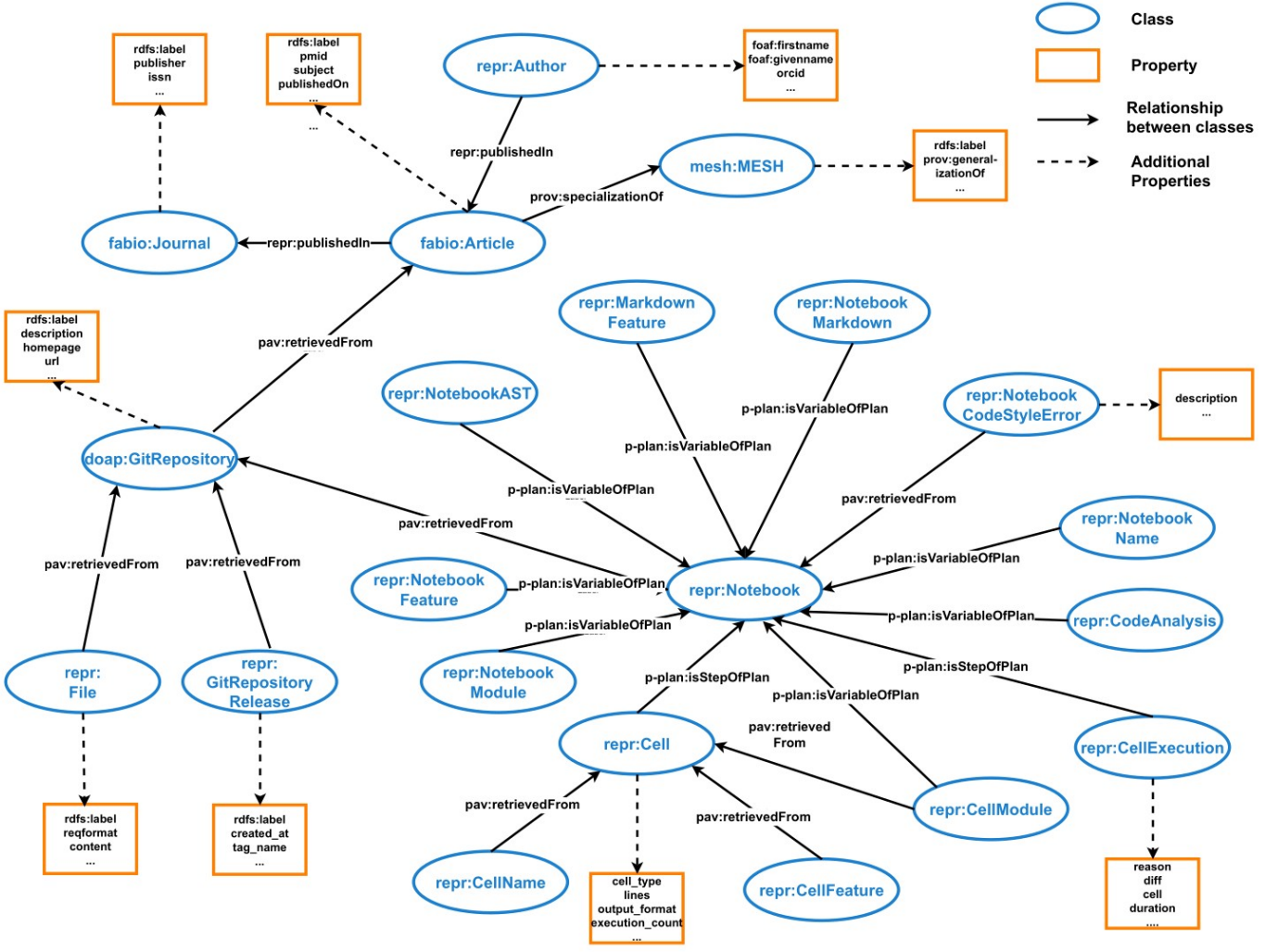


FAIR Jupyter Workflow



TECHNISCHE UNIVERSITÄT CHEMNITZ

FAIR Jupyter Ontology



TECHNISCHE UNIVERSITÄT CHEMNITZ

FAIR Jupyter SPARQL Queries

- Local
  - Exceptions occurring in Jupyter notebooks in our corpus.
  - Notebooks by search term: ‘immun’ AND (‘stem’ OR ‘differentiation’)
  - Article by keywords, e.g., ‘open source’
  - Most common errors in immunology
  - Most common errors in Nature journal
  - MeSH terms ranked by ‘ModuleNotFoundError’ frequency
  - GitHub repositories by their stargazers count
- Federated
  - Match articles between FAIR Jupyter and Wikidata via DOI

Environmental Footprint

- Simplified calculation: carbon footprint = energy needed * carbon intensity
- (http://calculator.green-algorithms.org/)

Future Trends and Challenges

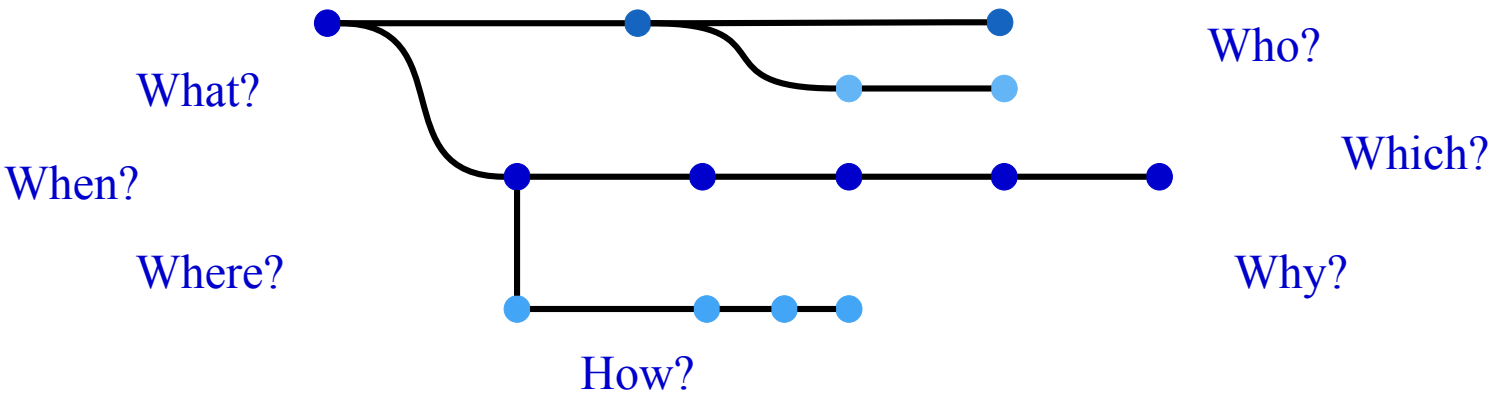
- Trends:
  - Integration with AI and Machine Learning
  - Enhanced Automation in Reproducibility Workflows
  - Reproducibility as a Metric

Future Trends and Challenges

- Challenges:
  - Interdisciplinary Research and Diverse Workflows
  - Long-Term Sustainability of Reproducible Artifacts
  - Ethical Considerations in Data Sharing

Summary

- Reproducible research benefits you and the research community
- The role of provenance, semantic web, FAIR data principles, and Open Science in reproducible research
- Solutions to 6W and 1H questions of data provenance for reproducible research
- Track, manage, and share the complete provenance of a research



Reproducible Research Resources

- Coursera Course on Reproducible Research: https://www.coursera.org/course/repdata
- Guide for Reproducible Research: https://the-turing-way.netlify.app/reproducible-research/
- ReproHack: https://www.reprohack.org/
- Awesome Reproducible Research: https://github.com/leipzig/awesome-reproducible-research

Thank you


sheeba.samuel@informatik.tu-chemnitz.de


@sheebasamuel


0000-0002-7981-8504

- https://sheeba-samuel.github.io/
- https://vsr.informatik.tu-chemnitz.de/about/people/samuel/