Rucio-SWAN Integration Project

Google Summer of Code 2020 with CERN-HSF

Muhammad Aditya Hilmy
mhilmy@hey.com

THE BIG QUESTION

How can we help scientists work productively in the Exabyte-scale era?

Rucio

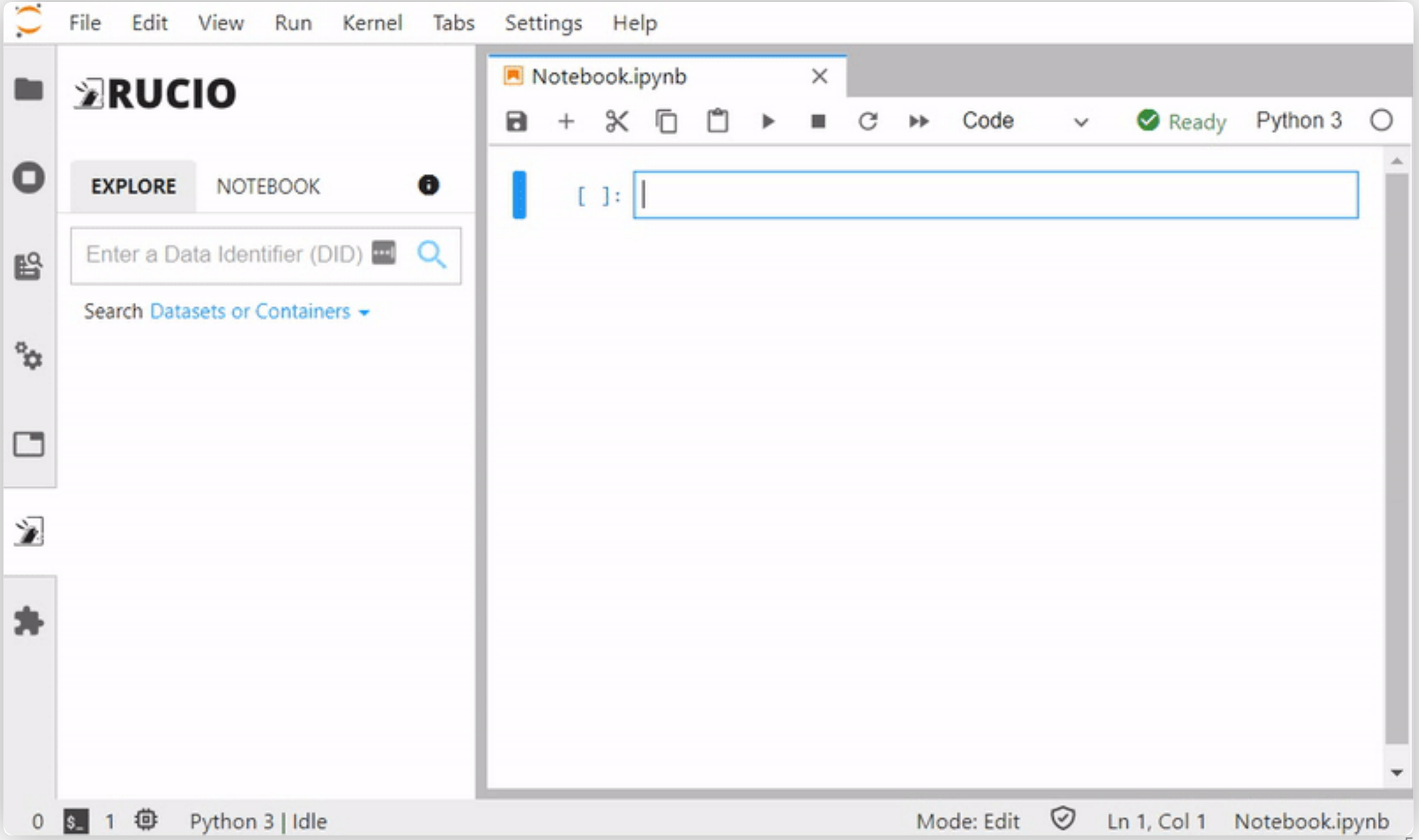
- Keeps track of data locations
- Moves data around as needed
- De facto standard for scientific data management

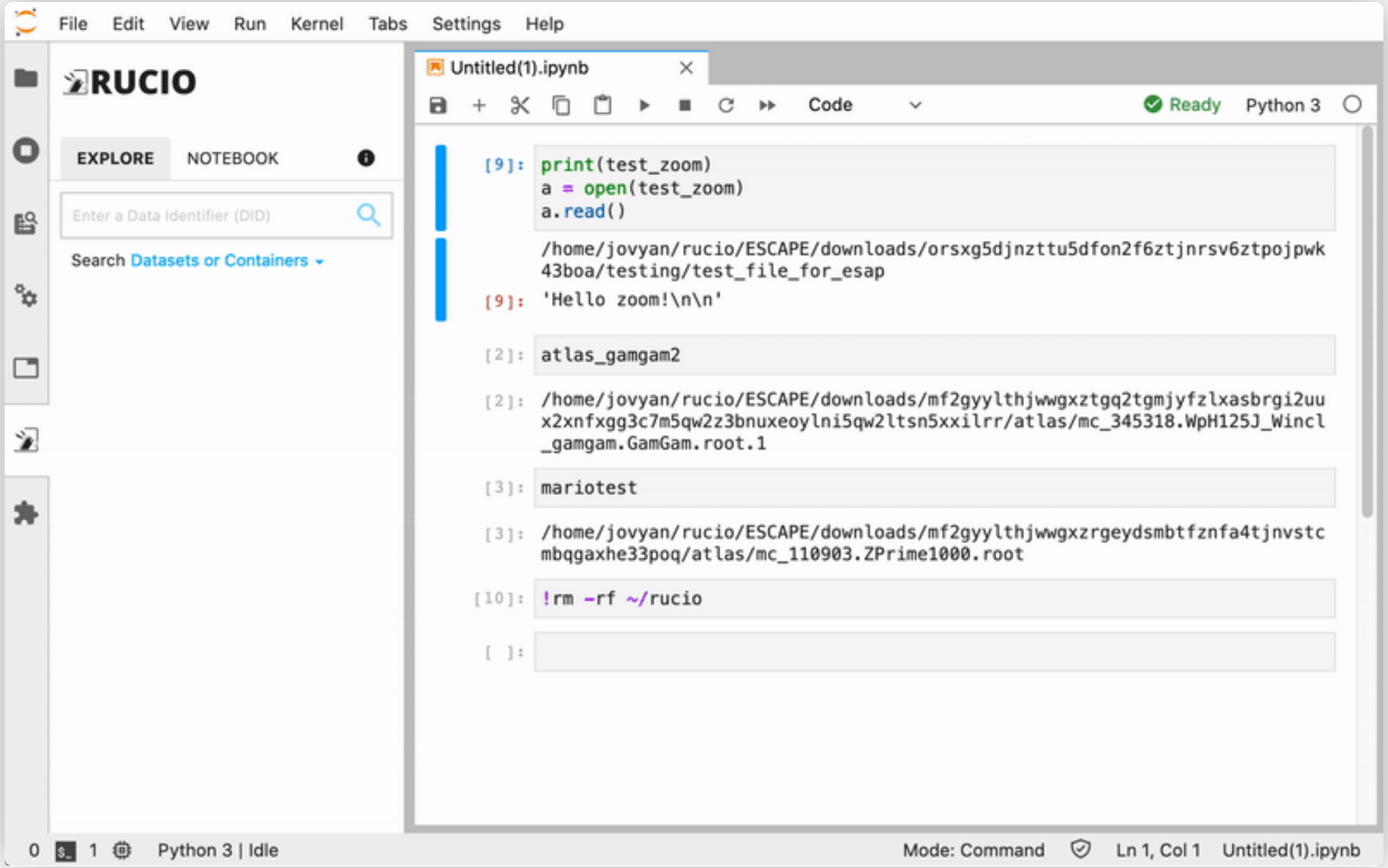
SWAN

- Online interactive Jupyter notebook
- No installation needed
- Enables collaboration through notebook sharing

Introducing, Rucio JupyterLab Extension.

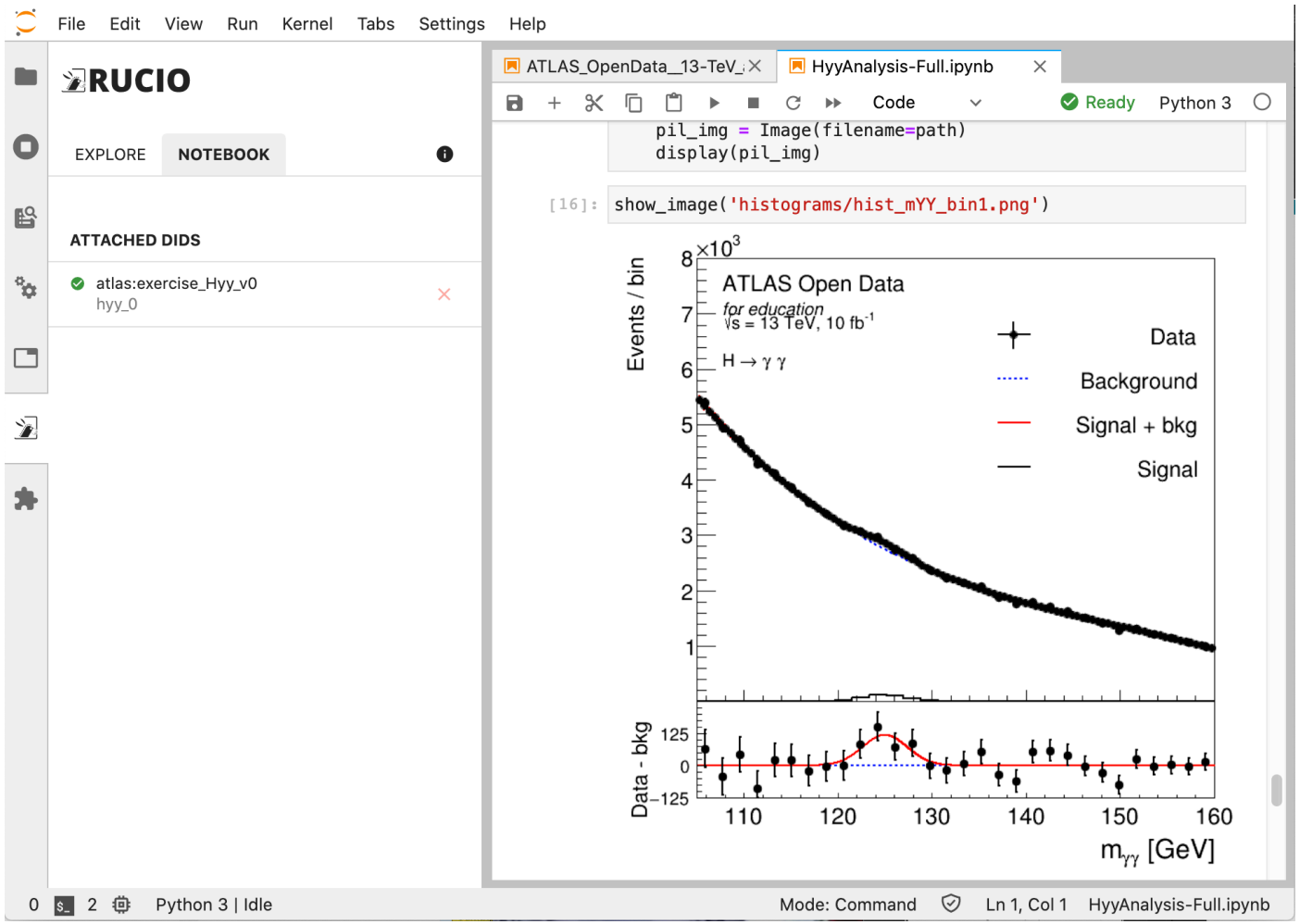
(I haven’t thought of a cool name for this project, so let’s stick to this extraordinarily ordinary name)


File
Edit
View
Run
Kernel
Tabs
Settings
Help
RUCIO
EXPLORE
NOTEBOOK
Enter a Data Identifier (DID)
Search Datasets or Containers
Notebook.ipynb
Code
Ready
Python 3
[ ]:
0
1
Python 3 | Idle
Mode: Edit
Ln 1, Col 1
Notebook.ipynb


File Edit View Run Kernel Tabs Settings Help
RUCIO
EXPLORE
NOTEBOOK
Enter a Data Identifier (DID)
Search Datasets or Containers
Untitled(1).ipynb
Code
Ready
Python 3
[9]: print(test_zoom)
a = open(test_zoom)
a.read()
/home/jovyan/rucio/ESCAPE/downloads/orsxg5djnzttu5dfon2f6ztjnrsv6ztpojpwk43boa/testing/test_file_for_esap
[9]: 'Hello zoom!\n\n'
[2]: atlas_gamgam2
[2]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxztgq2tgmjyfzlxasbrgi2uux2xnfxgg3c7m5qw2z3bnuxeoylni5qw2ltsn5xxilrr/atlas/mc_345318.WpH125J_Wincl_gamgam.GamGam.root.1
[3]: mariotest
[3]: /home/jovyan/rucio/ESCAPE/downloads/mf2gyylthjwwgxzrgeydsmbtfznfa4tjnvstcmbqgaxhe33poq/atlas/mc_110903.ZPrime1000.root
[10]: !rm -rf ~/rucio
[ ]:
0 1 Python 3 | Idle
Mode: Command
Ln 1, Col 1
Untitled(1).ipynb

SHOWCASE

ATLAS Open Data



- Hyy analysis using ATLAS Open Data
- No hardcoded path to file

SHOWCASE

ATLAS Open Data (2)

```
chain_data = ROOT.TChain("mini")
chain_paths = hyy_0[0:4]
for path in chain_paths:
    chain_data.AddFile(path)

chain_ggH125 = ROOT.TChain("mini")
chain_ggH125.AddFile(hyy_0[5])

chain_VBFH125 = ROOT.TChain("mini")
chain_VBFH125.AddFile(hyy_0[6])

chain_WH125 = ROOT.TChain("mini")
chain_WH125.AddFile(hyy_0[7])

chain_ZH125 = ROOT.TChain("mini")
chain_ZH125.AddFile(hyy_0[8])

chain_ttH125 = ROOT.TChain("mini")
chain_ttH125.AddFile(hyy_0[4])
```

- This is the code to load the ROOT files (in PyROOT).
  - No need to know the file paths
- `hyy_0` is an array of paths to files in dataset `atlas:exercise_Hyy_v0` in ESCAPE datalake.
  - The paths are injected by the extension automatically.

Notebook preview on
https://nbviewer.jupyter.org/gist/didithilmy/28400804ed55b1e4ff683902fa1cc58d

Key Features

- Browse Rucio data from the Lab sidebar
- Replicate data with just one click
- Resolves file path automagically
- Inject path to notebook as a variable
- Supports two methods of authentication (currently):
  - Username & Password
  - X.509 User Certificate
- Supports two modes of operation:
  - Replica mode: uses network-attached storage as an RSE, utilizes Rucio's file transfer capability.
  - Download mode: downloads data directly to the user's directory using Rucio clients.
- Remote configuration

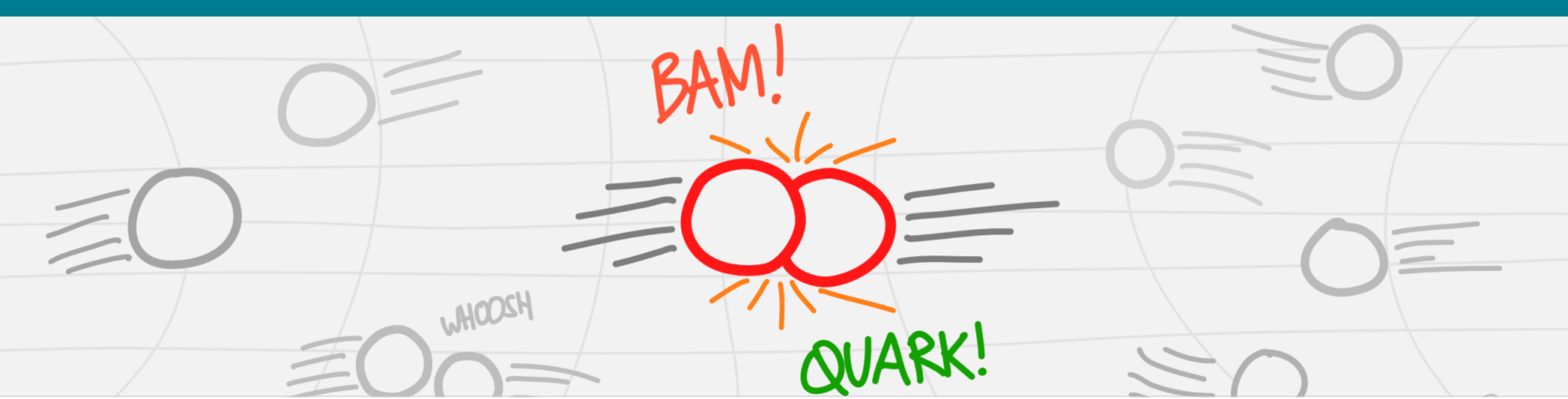
Future Developments

- More Kernel compatibility
  - Octave, R, ROOT C++
- More authentication methods
  - OAuth/OpenID Connect
- Share notebooks across JupyterLab installations
  - Allows any JupyterLab instance to connect to publicly-accessible Rucio installations and their RSEs
  - Fetches Rucio configuration on-the-fly, URL known from notebook metadata
- (If you have other ideas, please let me know)

Acknowledgements

Huge thanks to my mentors:
- Aris
- Riccardo
- Martin
- Diogo
- Mario
- Enric
- Enrico

..and an unofficial mentor
- Thomas



Thank you.

Attributions:
CERN-HSF logo courtesy of hepsoftwarefoundation.org
Rucio logo courtesy of rucio.github.io
SWAN logo courtesy of swan.web.cern.ch

Muhammad Aditya Hilmy
mhilmy@hey.com
didithilmy