

# IRIS-HEP Fellowship Proposal

## Aaron Wang

**Duration: January 2020 - June 2020**

**WBS: Analysis Systems (AS)**

**Project: Jupyter Notebook Compatibility with ROB (Reproducible Open Benchmarks for Data Analysis Platform)**

**Funding Period: 5 months(20 weeks) of 1/4th FTE time commitment**

The *Reproducible Open Benchmarks for Data Analysis Platform (ROB)*[1] is a platform that allows for the evaluation of different data analysis algorithms in a controlled competition-style format. The benefit of *ROB* is that it allows for concrete comparison between neural networks, especially where the efficacy of neural networks are yet to be clearly compared, and are hard to reproduce: such as in particle jet tagging.

The *ROB* follows a simple four step workflow [2]. First, a common input data set used for benchmarks is imputed by the benchmark coordinator. Then the users provide code and prediction stages of the machine learning model. Lastly, metrics are evaluated with tables and plots as defined by the benchmark coordinator. Through this, different users are able to apply separate machine learning models on a common data set, in a controlled environment (*ROB*).

Although the *ROB* is already useful, it is still missing compatibility with the commonly used Jupyter Notebooks. Currently, the *ROB* only supports code for preprocessing and prediction stages in python files. Since machine learning models are often developed in Jupyter Notebooks, needing to convert the code into a compatible file before submitting it to the *ROB* can be an unnecessary step. Two possible ways that this problem can be solved is by either using PaperMill[4], a tool for parameterizing and executing Jupyter Notebooks, which transforms the notebooks into Docker containers and running them as workflows, or by extending the *ROB* client so that it can be used from within the Jupyter Notebook by developing a python library that runs *ROB* from python notebooks. Under the local supervision of Professor Shih-Chieh Hsu (UW), and the technical guidance of Heiko Müller (NYU), I will be exploring these two possibilities, and then engineering a way to make Jupyter Notebooks compatible with the *ROB*.

## Anticipated Schedule of Deliverables

### 1st Month(40 Hours)

#### Week 1 - 3:

- Familiarize with *ROB* by reading developer documentation and running the provided *ROB* demos
- Understand the input requirements of the code for the submitted pre-process and prediction files in *ROB*

### 2nd Month(40 Hours)

#### Week 4-7:

- Read PaperMill documentation and explore the suitability of using PaperMill as part of the compatibility project
- Explore integrating *ROB* clients into Jupyter notebook, and evaluate which method will work better.

#### Week 8:

- Start developing the *ROB* and Jupyter Notebook compatibility software

### 3rd to 5th Month(40 Hours per month)

#### Week 18:

- Have code finished and start running tests

#### Week 19:

- Finalize python library and start writing documentation

#### Week 20:

- Finalize documentation and upload code to a github repository available publically

**Academic Workload:** During the time, I will also be taking a full course load at the University of Washington

#### References:

[1] Müller, Heiko: <https://github.com/scailfin/flowserv-core>

[2] Müller, Heiko and Macaluso, Sebastian: *ROB: Reproducible Open Benchmarks for Data Analysis Platform*, <https://indico.cern.ch/event/809820/contributions/3632640/>

[3] Müller, Heiko and Macaluso, Sebastian: <https://github.com/scailfin/rob-demo-top-tagger>

[4] PaperMill: <https://papermill.readthedocs.io/en/latest/>