WP4 Data Analysis Services

Hans Fangohr
On behalf of WP4
Trieste, Italy

@ProfCompMod

04 November 2019

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Vision for European Open Science Cloud

One place to

- **F**ind existence of data sets
- **A**ccess data and metadata as
- **I**nteroperable data sets
- **R**eusable data and data analysis procedures
  - ▶ reproducibility
  - ▶ ability to re-use and extend studies

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Why data analysis for FAIR data?

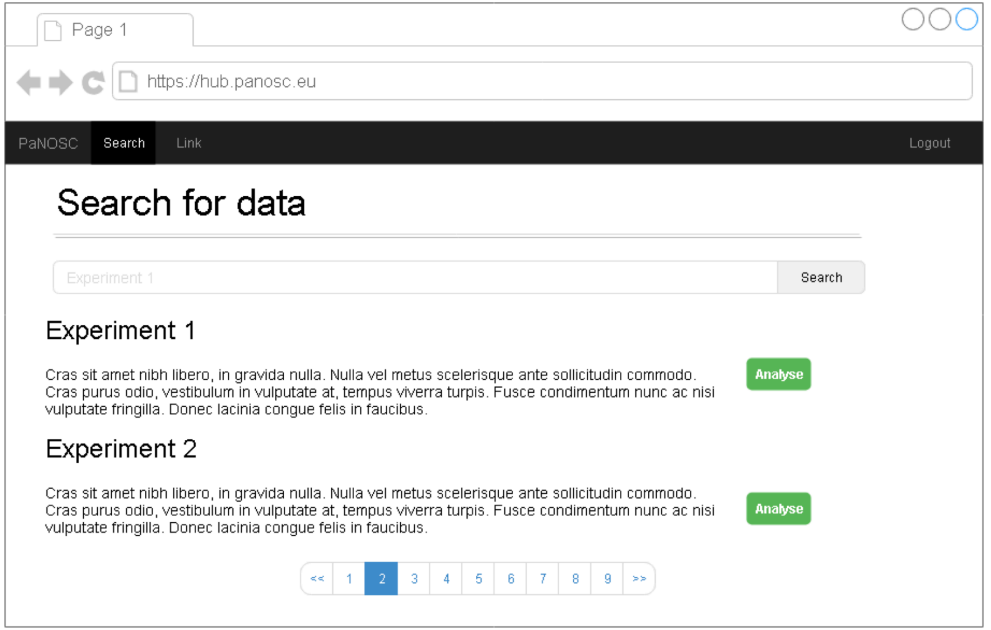
- To make the (raw) data usable and re-usable
  - ► significant post-processing required before knowledge can be extracted, including
  - ► calibrating the data from complex detectors
  - ► converting facility and hardware specific data into data representing the physics of the experiment
- Further data processing to extract insight from the data

- Ensure completeness of meta data
- If we have (working and meaningful) analysis routines, they contain all the required data
  - ► The source code is a complete description of the computation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Vision for data analysis services in PaNOSC

- Search and
- Select data set from portal
- Allow data analysis on data set from this portal (or some other instance that is invoked on the fly).



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Challenges and solution elements

- Required
  - Data sets and access for analysis
    - ► Either move data or move computation to data
  - Data set → data analysis procedure
    - ► Classification of data / experiment types
  - Metadata

  - Remote interactive analysis environment
    - ► Remote Desktop (VISA), JupyterHub
  - Analysis software environment
    - ► Containers (Docker, Singularity, Virtual Machines)
    - ► Specific for each data set?

- Desirable
  - Data analysis working across facilities
    - ► Agreed API / file format (NEXUS)?
    - ► Joint use of base libraries (Silx, pyFAI, h5py, …?)

- Additional challenges
  - EOSC and the technologies we need to use are developed during this project
    - ► Difficult to plan
    - ► Opportunities for flexible adjustments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Practical and historical aspects

- Facilities use different infrastructures
- Facilities in general support different types of experiments
- Where the experiments are of the same or similar type, the representation of the data and metadata is different
- Existing infrastructures that originate from hundreds of person years of effort are difficult to change
- To make our efforts sustainable, the PaNOSC developments must provide value to each facility (and their users)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

EOSC PaNOSC Data Analysis Portal vision

- Search for data set and select (WP3)
- For selected data set
  - Have metadata available (WP3)
  - Have choice of suggested data analysis procedures (WP3, WP4)
    - ► Could include simulation of experiment? (WP5)
    - ► Could provide link to tutorial for this type of experiment / analysis? (WP8)
  - Ability to trigger execution of these analysis procedures
    - ► Needs analysis code, software environment and data (WP4, WP6)
  - Ability to modify analysis interactively (WP4)
  - Allow download of new analysis results
- Focus on data sets connected with reproducible publications
  - Allow to view, re-execute (=reproduce) and modify (=re-use) code that created published results

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Why focus on reproducible publications as scientific use cases?

- Need to narrow down scope of project somehow
  - Publications select and describe useful parts of data from experiment
  - An analysis method used in one publication is likely to be useful for a new study
    - ► Defines (minimum set of) analysis tools/services we support
    - ► Starting point for provision of this type of analysis at the facility where the data originates
  - Publication must provide relevant meta data
  - Publications are key performance indicator
- Data sets can (often) be made public as results are published (even before 3-year embargo expires)
- Related: *every* publication *should* be reproducible
  - Journals and funding councils increasingly push for reproducible publications
  - Reproducible publications are much more easily re-usable → more effective science

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Communication in WP4

- Communication and tool infrastructure
  - WP4 specific mailing list
  - Slack channel
  - Kick-off meeting in June at EuXFEL
  - WP4 subdirectory in panosc-eu repository on Github
    - With meeting agendas and minutes
    - "resources"
      - https://github.com/panosc-eu/panosc/tree/master/Work%20Packages/WP4%20Data%20analysis%20services
    - 2-weekly video meetings of (at least) site leaders
      - Tuesdays 15:00
  - Task tracking (https://github.com/orgs/panosc-eu/projects/2)
  - https://confluence.panosc.eu/ has WP4 area
  - Invited ExPaNDS project members to join meetings

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Ongoing activities in WP4

- Active projects
  - Recruitment of staff*
  - Research “literature”
    - ► computational workflows
    - ► and design remote analysis architecture
  - Portal design*
  - Remote analysis infrastructure*
  - JupyterHub*
  - Move analysis capabilities into Jupyter notebook*
  - Hdf5 and Visualisation in notebooks*
  - Questionnaire on data analysis for participating facilities (Task 4.1)
  - Identify reproducible scientific use cases at each institute

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

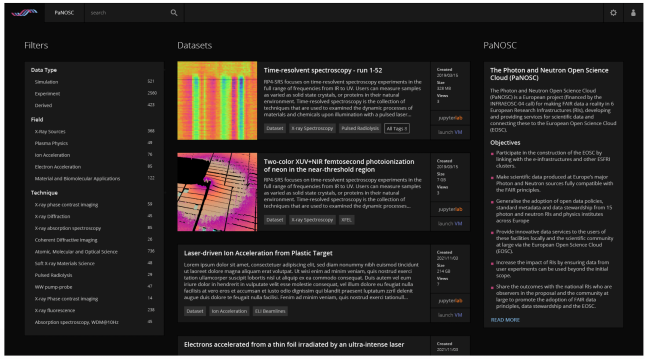
Recruited and contributing staff for PaNOSC WP4

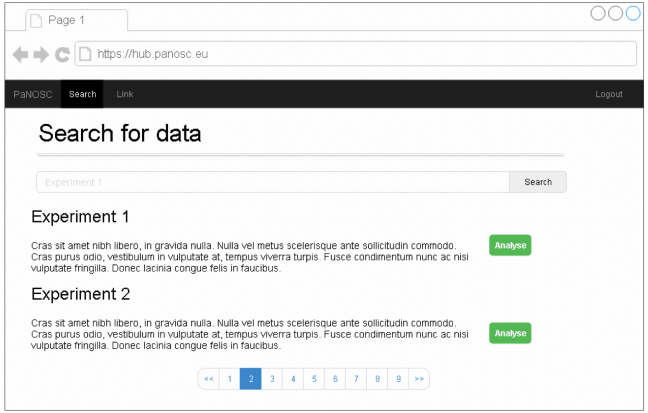
- CERIC-ERIC
  - *Carlos Reis*, Marco de Simone
- ELI
  - *Jakub Grosz* (Eli Beamlines), Mariana Danielova (Eli Beamlines), Rober Racz (Eli ALPS)
- ESRF
  - *Thomas Vincent, Andy Goetz*, B. Roussel, A. Roux, A. Campbell
- ESS
  - *Kareem Galal*, Jesper Rude Selknæs, Lottie Greenwood, Alexandre Stefanov and Ashok Nulguda, Torben Nielsen, Gareth Murphy
- European XFEL
  - *Robert Rosca, Thomas Kluyver*
- ILL
  - *Jamie Hall*, *William Turner*
- EGI
  - *Giuseppe La Rocca*, Enol Fernández

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Portal

- Design of Portal, integrating requirements from WP3 and WP4
  - https://confluence.panosc.eu/display/wp4
  - Presentation Jamie Hall Wednesday 9:30 (Bassovizza)
- Aim to deploy early prototype with minimum functionality at each facility in March 2020
  - Functionality
    - ► Start remote desktop instance
    - ► Start Jupyter Notebook instance
    - ► Will need local adaption
  - Gather feedback
  - Move towards federated instance iteratively



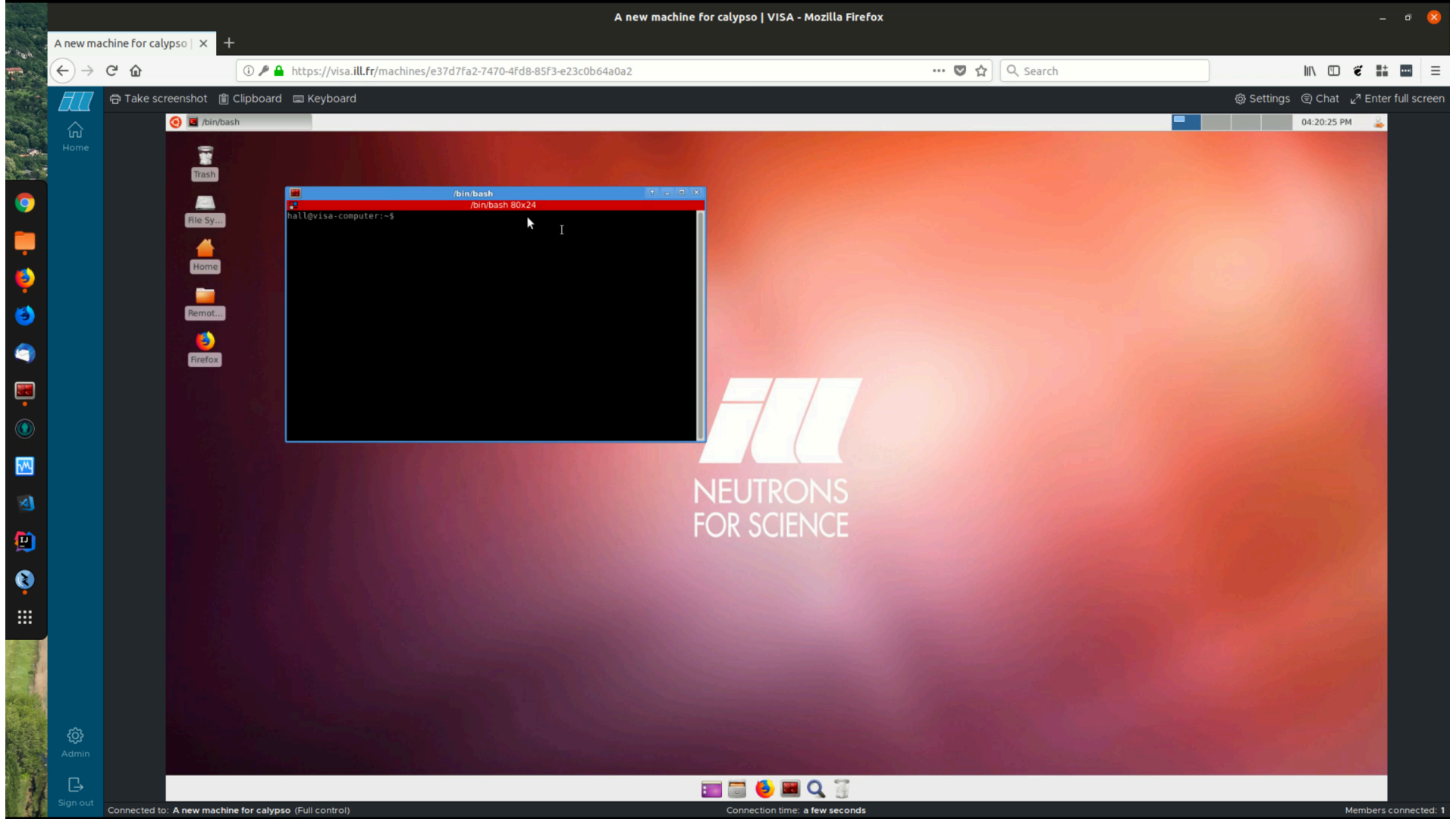


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

Remote data analysis technology 1: Remote Desktop and Virtual Machines

- Technology 1: VISA – Virtual Infrastructure for Scientific Analysis (ILL)
  - Connect Desktop GUI of Virtual Machine
  - Inside the web browser
  - Operate remotely
- Essentially creates a virtual machine for every analysis use case
  - Can support any data analysis software, any operating system.
  - Computation takes place where the data is stored
  - However:
    - ► Prescribes operating system, window manager, keyboard layout, …
    - ► Different environments for different analysis applications
- Related remote desktop technology used at other facilities (FastX).

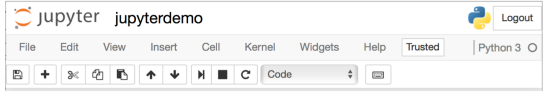
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852


A new machine for calypso | VISA - Mozilla Firefox
A new machine for calypso
https://visa.ill.fr/machines/e37d7fa2-7470-4fd8-85f3-e23c0b64a0a2
Search
Take screenshot
Clipboard
Keyboard
Settings
Chat
Enter full screen
Home
/bin/bash
04:20:25 PM
Trash
File Sy...
Home
Remot...
Firefox
/bin/bash
/bin/bash 80x24
hall@visa-computer:~$
NEUTRONS
FOR SCIENCE
Admin
Sign out
Connected to: A new machine for calypso (Full control)
Connection time: a few seconds
Members connected: 1

Remote data analysis technology 2: Jupyter Notebook for data analysis

- Jupyter Notebook
  - Combines commands
  - Processing outputs and
  - Interpretation / annotation
- Represents the whole research to publication life cycle
  - in one virtual research environment
  - in one document
    - ► better reproducibility
- Can be shared easily (ipynb, html, pdf)
- Tools from the Jupyter ecosystem allow remote execution of notebooks

Demo: https://github.com/fangohr/jupyter-demo/



Code cells show code input and output:

```
In [1]: 1 + 2
Out[1]: 3
```

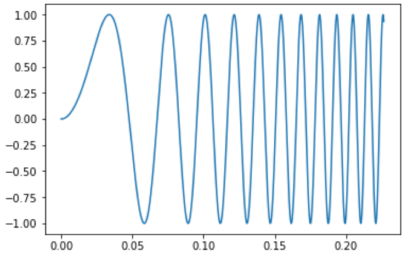
Cells can contain text and latex equations such as $f(x) = \sin(2\pi\omega t^2)$ and $\omega = 220$ Hz. We can use code to define the corresponding functions:

```
In [2]: import numpy as np
        def f(t):
            omega = 220
            return np.sin(2 * np.pi * omega * t**2)

In [3]: f(0)   # call the function
Out[3]: 0.0
```

Let's compute the data and plot the beginning of it:

```
In [4]: t = np.linspace(0, 2, 44100)
        y = f(t)
        ## Show plots inside the notebook
        %matplotlib inline
        import pylab
        pylab.plot(t[0:5000], y[0:5000])
Out[4]: [<matplotlib.lines.Line2D at 0x10a267898>]
```



Summary

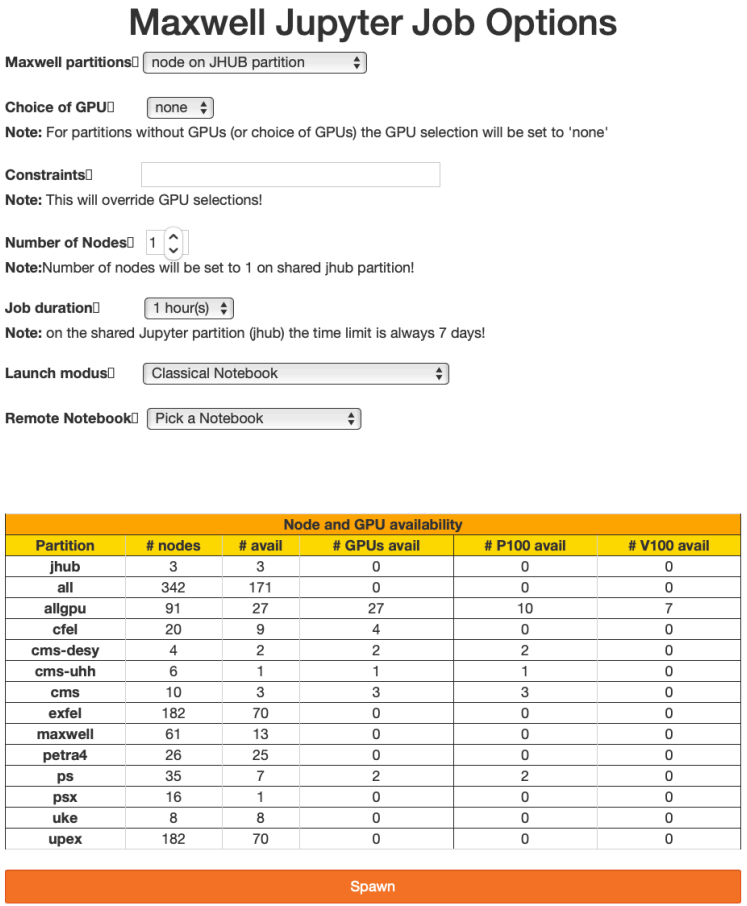
It is possible to combine equations, code, data processing commands, their output, including plots, and text in one document.

This project has received funding from the European Union's Horizon 2020 research and innovation pro

OSC

JupyterHub

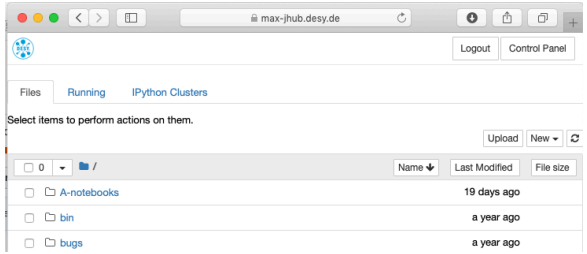
- JupyterHub: service that allows
  - remote execution of notebooks
  - Offering HPC and storage resources through webportal
- All facilities working towards JupyterHub instance offered to users
  - EuXFEL, ILL: in operation
  - ESRF, CERIC-ERIC: prototype deployed
- Welcome by users
  - Example: 150 users per week at EuXFEL
  - Anecdotal evidence that Jupyter Interface can be preferred over QT Interface

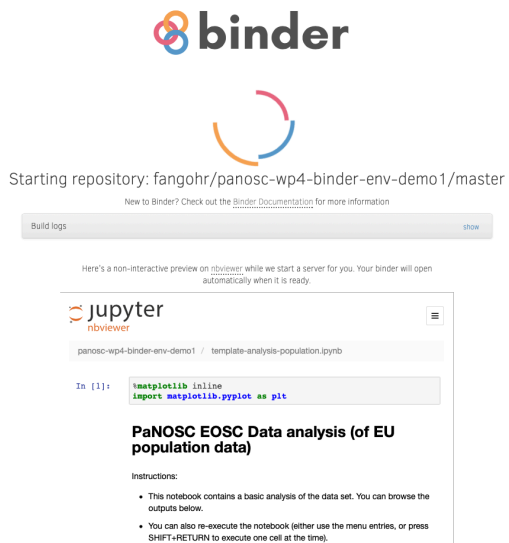


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

BinderHub

- JupyterHub: service that allows
  - remote execution of notebooks
  - Offering HPC and storage resources through webportal
- BinderHub (a JupyterHub with Binder-style software specification)
  - Tailored container is created based on ”binder software specifications”
    - ► `repo2docker`
      - `requirements.txt`, `environment.yml`, `Dockerfile`
  - Example: http://github.com/fangohr/panosc-wp4-binder-env-demo1
    - ► Uses global anonymous BinderHub called MyBinder. (https://mybinder.org)
  - Promising as environment that can be configured flexibly?

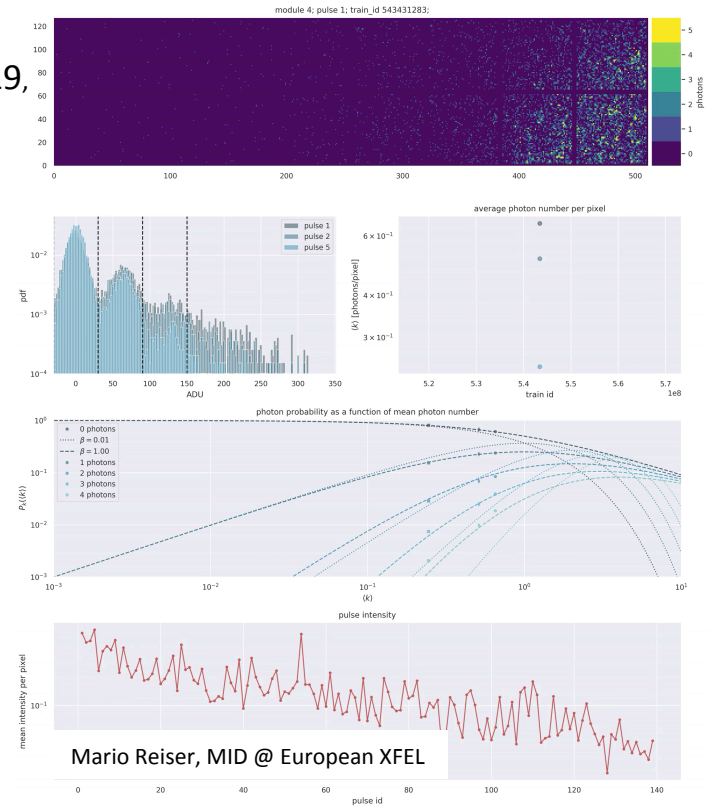




This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant

Moving data analysis into the Jupyter notebook

- Examples from ICALPECS 2019, preprint of slides at http://icalepcs2019.vrws.de/talks/tucpr02_talk.pdf
- Presentation Thomas Vincent (ESRF): *Reproducible science studies*
  Wednesday 11:30 (Bassovizza)



Mario Reiser, MID @ European XFEL

```
In [3]: import ToolBox as tb
```

X-ray Absorption Spectroscopy

Step 1: Load data and align them by train id and pulse id

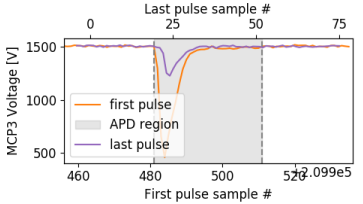
```
In [4]: proposalNB = 900074
        semesterNB = 201930
        runNB = 487
        topic = 'SCS'
        fields = ["SCS_photonFlux", "SCS_XGM", "MCP3apd", "nrj"]
        run = tb.load(fields, runNB, proposalNB, semesterNB, topic,
                      validate=True, display=False)
        nrun = tb.matchXgmTimPulseId(run)
```

```
Checking run directory: /gpfs/exfel/exp/SCS/201930/p900074/raw
/r0487/
No problems found
```

Step 2: check the pulse integration window

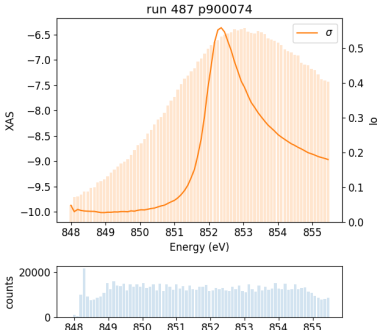
```
In [5]: tb.checkTimApdWindow(nrun, mcp=3)
```

```
no raw data for MCP3. Loading trace from MCP3
```



Step 3: bin the data and plot the XAS spectrum

```
In [6]: nrj = np.linspace(nrun.nrj.min(), nrun.nrj.max(), 80)
        xas = tb.xas(nrun, nrj, plot=True)
```



Activities on HDF5

- HDF5 used widely in facilities
- Multiple activities to improve data exploration capabilities for hdf5 files
  - Presentation Carlos Reis (CERIC ERIC), Thursday 9:40: *HDF5 Viewer – web and Jupyter Interface*
  - h5glance, silx, karabo-data-interactive, …


```
In [5]: import h5glance
        h5glance.install_ipython_h5py_display()

In [6]: f

Out[6]: example.h5
          group1
            subgroup1
              dataset1 [ ]: 200 entries, dtype: <u8
              dataset2 [ ]: 2 × 128 × 500 entries, dtype: <f4
            subgroup2

In [7]: f['group1']

Out[7]: /group1
          subgroup1
          subgroup2
```

panosc

Summary

- Many ideas and good initial progress
- Next steps
  - require more fine grained coordination
    - ► Within WP4
    - ► Within PaNOSC
  - Limit scope to be realistic
  - Deploy portal prototype (March 2020)
  - Involve all facilities
- Dedicated activities on Wednesday and Thursday for WP4
  - And joined workshops with WP3, WP6 and WP8

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852