

WP4: Data Analysis Services – Introduction

16 June 2020

Hans Fangohr

European X-ray Free Electron Laser (EuXFEL)



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

Outline

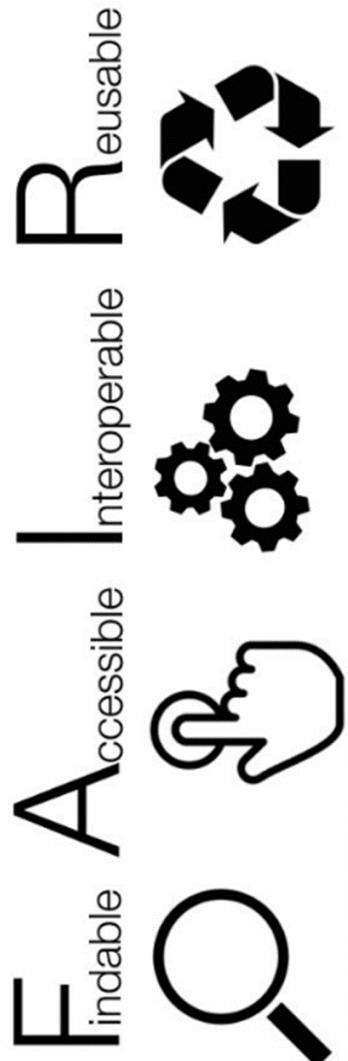
- Motivation and Objectives
- Programme of work (Tasks)
- Re-usable remote data analysis
- Status of remote data analysis services
- Towards a unification of data analysis services
- Relation to subsequent talks

WP4 Motivation for data analysis services

- **Data analysis extracts meaning from data**
 - non-trivial, embeds knowledge
 - need analysis software together with data

WP4 Objectives

- **Enable better reproducibility [1] and re-usability (faiR)**
- **Control and execute data analysis remotely**
 - Important for EOSC idea: data widely accessible (fAir)
 - Useful for facilities: data set size & COVID19



[1] Reproducibility crisis, Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a

Programme of work

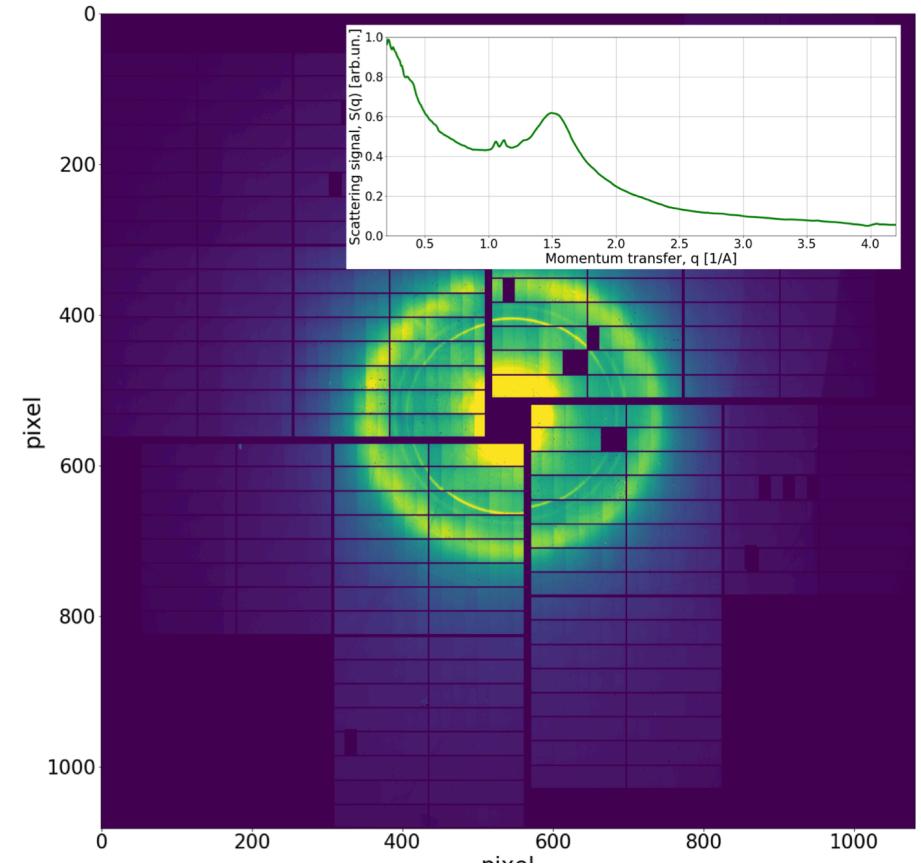
- Task 4.1: Survey (M1-13)
 - Submitted deliverable D4.1
- **Task 4.2 and 4.4 Remote Desktop & Jupyter based analysis services (M1-48)**
 - Submitted deliverable D4.2
 - 10:20 Demo (Thomas Vincent)
- **Task 4.3: Portal for remote data analysis (M13-36)**
 - 10:10 Demo (Jakub Grosz)
 - 11:25 Talk (Jamie Hall)
- **Task 4.5: Deployment of remote analysis services at facilities (M12-48)**
 - Submitted deliverable D4.2
- Task 4.6: Demonstrator (M36-48)

See also technical report M18 for further details.



Towards re-usable remote data analysis for FAIR data

- **Ingredients**
 - Data & Metadata
 - Data analysis procedure (scripts / procedure)
 - Data analysis software environment
 - Remote executability
- **Remote data analysis**
 - SSH
 - Remote Desktop (VNC, FastX, VISA, NoMACHINE)
 - Jupyter Notebook (JupyterHub, BinderHub)



Jupyter

Architecture

- Computation takes place on machine in facility / cloud
- Connect via https to user's web browser

Discussion

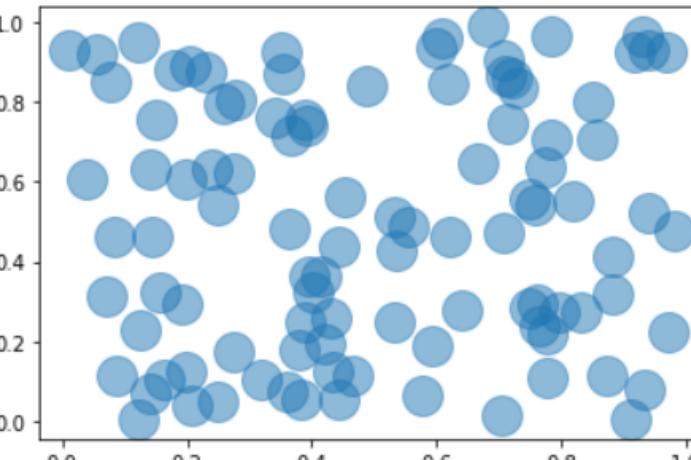
- “native web”
- open source & extensible
- emerging tool
- supports reproducibility

Jupyter-notebook-demo (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 O

In [1]: `import pylab, numpy`

In [2]: `n = 100
x = numpy.random.uniform(size=n)
y = numpy.random.uniform(size=n)
pylab.plot(x, y, 'o', alpha=0.5, markersize=20);`



Looks like it - at least they are not evenly distributed.

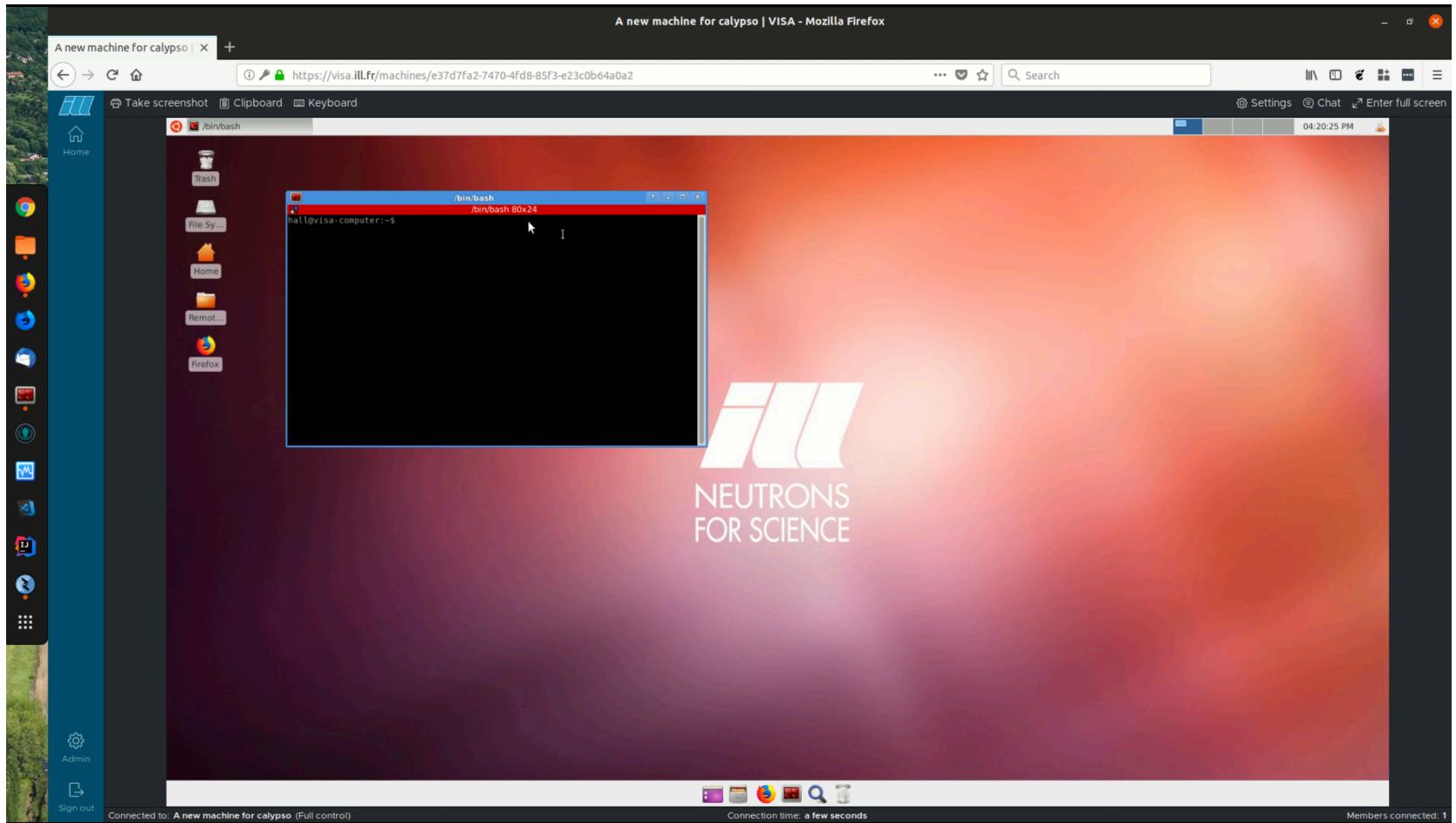
Remote desktop

Architecture

- Computation takes place on machine in facility / cloud
- Connect the machine's graphical desktop to user's web browser

Discussion

- Can make any existing analysis environment available remotely



Status remote data Analysis: Jupyter

Facility	JupyterHub	
	<i>Status</i>	<i>Spawner</i>
CERIC-ERIC	Pilot	Docker, Singleuser
ELI	Pilot	Docker (single test node)
ESS	Pilot	Kubernetes
ESRF	Production	Kubernetes, Slurm, Singleuser
ILL	Pilot	Sudo
EuXFEL	Production	Slurm, Singleuser



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



Status remote data Analysis: remote Desktop

Facility	Remote desktop	
	<i>Status</i>	<i>Technologies used</i>
CERIC-ERIC	Pilot	VNC, guacamole
ELI	NA	NA
ESS	Pilot	VNC
ESRF	Production	NoMachine
ILL	Production	VISA (XRDP, guacd)
EuXFEL	Production	FastX

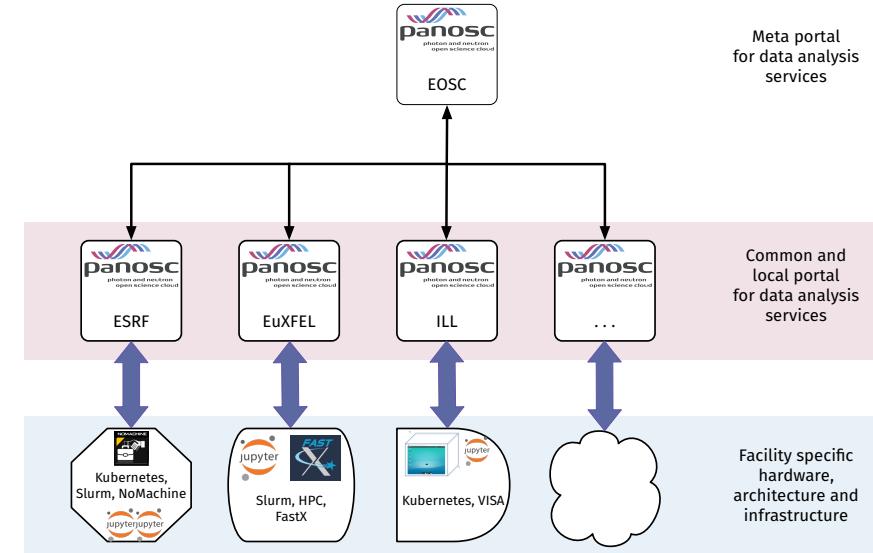


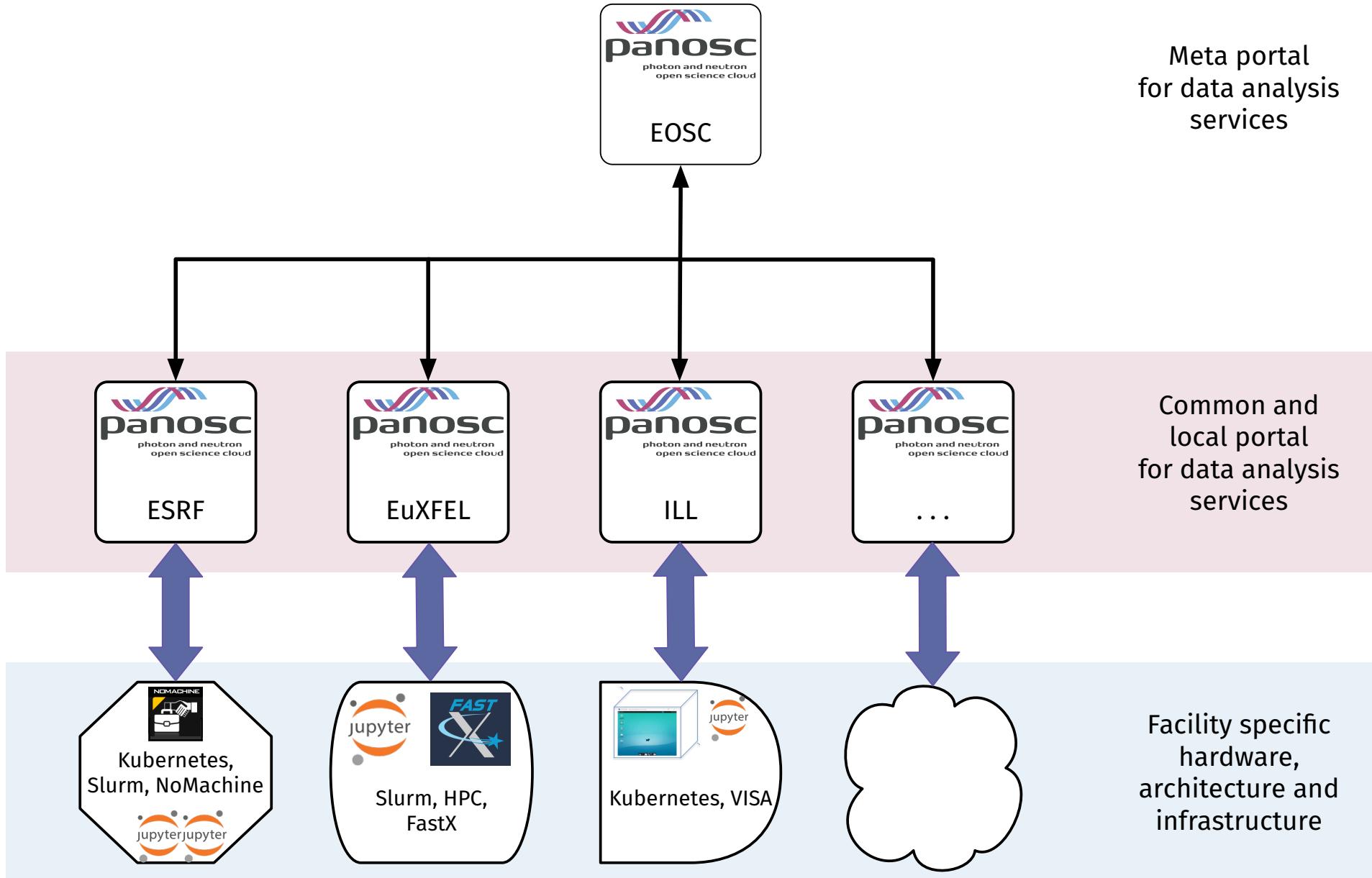
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852



Towards a unification of data analysis services

- **Constraints:**
 - Established hardware and infrastructures at each site
 - Need useful user experience for sustainability
- **Strategy**
 - Create common entry point (“Data search and analysis portal”) for each site as user interface
 - Portal translates request into local infrastructure
 - Potential to offer portal interface in EOSC



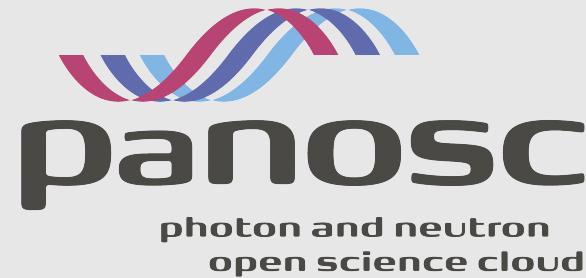


This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852



Programme today

- **10:10 Demo Portal (Jakub Grosz)**
 - User experience Portal
 - Data finding, selection, and triggering of analysis (FAiR)
- **10:20 Demo remote data analysis (Thomas Vincent)**
 - Example data analysis
 - Use case from scientist using ESRF
 - Remote analysis with Jupyter
 - Reproducible & re-usable work
- **11:25 Presentation Portal (Jamie Hall)**
 - Architecture of Portal



Thank you

hans.fangohr@xfel.eu



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