News and data highlights

J. Wrigley **D. Ferreira de Lima**

European XFEL

Users' Meeting 2024

Data management, analysis, and reduction at European XFEL

January 26th 2024

Introduction

Things that we worked on this year:

- Secured access to **our systems**
- Improved tools to **track experiments**.
- Updates to the **correction system**.
- The **Scientific Data Policy** and the **Data Management Plan**.
- Updates in our **data analysis** tools.
- Automating *online* and *offline* beam characterization in SASE3 with the **virtual spectrometer**.
- Coming soon: **automating SFX data analysis**.

Multi-factor authentication

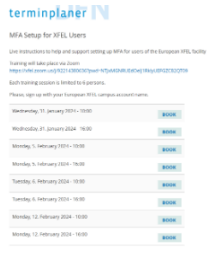
- Additional measures have either already been implemented or will take effect soon.
- Access to the data analysis facility (known as Maxwell cluster) will be protected by the 2nd authentication factor from **30.01.2024**.
  - ssh to Maxwell display nodes (e.g. `max-exfl-display.desy.de`)
  - FastX on the Maxwell display node
  - JupyterHub – `https://max-jhub.desy.de`

- Access to our systems requires at least one 2nd factor (e.g. TOTP). Consider setting up another (e.g. SMS or external email) for the password recovery service.

Setting up 2nd factor authentication remotely

- DESY-IT provides full documentation under:
  - `https://it.desy.de/services/mfa/naf__maxwell/index_eng.html`
- Email has been sent to `computing@xfel.eu`.
- Additional slides at the end of this presentation.
- We offer several Zoom sessions with live guidance to assist with the setup process. If you wish to use this opportunity, please register through our polling service.
  - `https://terminplaner6.dfn.de/en/b/d9cce8ad2a6325e63be64de22ca3fc98-555867`
- In case you have specific questions, you can contact us using `it-support@xfel.eu`.
- Sign up for the `computing@xfel.eu` mailing lists at the User Portal (UPEX): `in.xfel.eu/upex/user/mailinglist`.

Initially you can use UPEX password recovery, if your password expired or is forgotten.



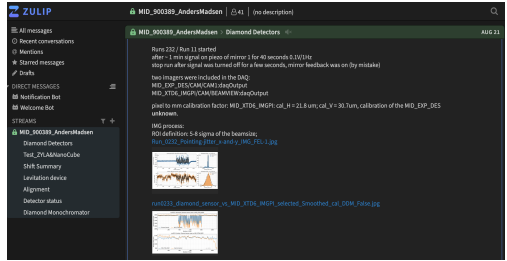
Change of Operating System

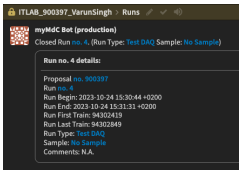
- The currently used OS for European XFEL offline and online analysis clusters is CentOS 7:
  - Maxwell cluster.
  - Online analysis clusters close to instruments.
- End of Life for CentOS 7 is 30th June 2024.
- All analysis clusters will be upgraded to RedHat EL9.
- Test login nodes and test partitions are available at the Maxwell cluster (maxcpu-el9, allcpu-el9).
  - We will create an upex equivalent partition with XFEL nodes (announcement via `computing@xfel.eu`).

```
~% ssh max-display-el9
[max-display003]~% cat /etc/redhat-release
Red Hat Enterprise Linux release 9.3 (Plow)
```

Experiment tracking services - myLog

- myLog electronic logbook officially released to all instruments!
  - Logbooks can be created in myMdC once proposal Arrival Form is submitted in UPEX.
  - Possible to configured desired sources of automatic metadata.
  - Logbook membership mirrors proposal team.
  - Markdown support makes writing messages extremely flexible.
- Integration architecture with Karabo and DAMNIT defined and standardized.
- Flexible configuration and RESTful APIs.



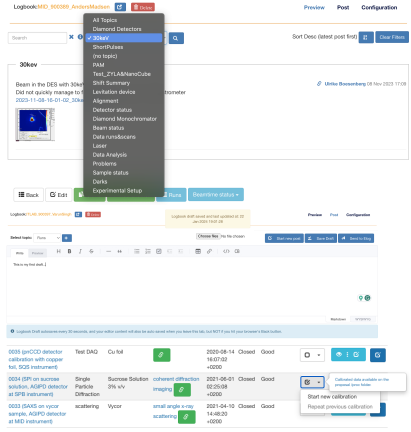


See slides from L. Maia and R. Rosca

▸ Basics: services, resources, and data access

Experiment tracking services - myMdc

- Enhancing myLog integration features:
  - Improved GUI and search.
  - Show entries history.
  - DAMNIT bot generation.
  - Allow drafting HTML/Markdown messages and posting to myLog from myMdC (with auto-save).
- Ability to assign PanNET techniques to proposals so they can be associated with DAQ runs.
- Calibration runs report available in myMdC.
- Enhancement of myMdC flexible configurations and RESTful APIs.
- Support for the latest version of the Globus API.

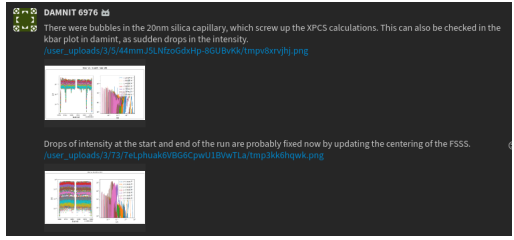


See slides from L. Maia and R. Rosca

▸ Basics: services, resources, and data access

Experiment tracking services - DAMNIT

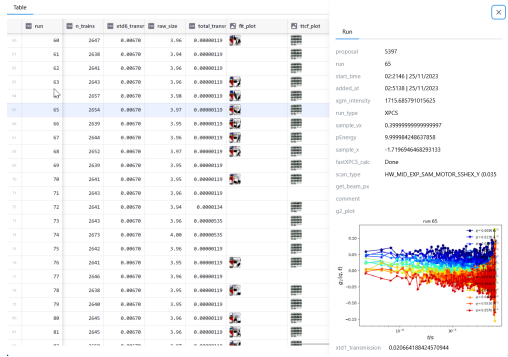
- New breaking release this week:
  - Improvements to the internal data format.
  - Full support for `DataArray`'s and `Dataset`'s.
  - Automatic migration in the next couple of weeks, or contact `da-support@xfel.eu`.
- Open-sourced:
  `github.com/european-xfel/damnit`
- Integration with myLog.
- Web interface coming Soon™.
- Other features: displaying logs in the GUI, export to CSV/Excel, saving matplotlib `Figure`'s.



See slides from T. Michelat ▸ Experiment overview with DAMNIT

Experiment tracking services - DAMNIT

- New breaking release this week:
  - Improvements to the internal data format.
  - Full support for `DataArray`'s and `Dataset`'s.
  - Automatic migration in the next couple of weeks, or contact `da-support@xfel.eu`.
- Open-sourced: `github.com/european-xfel/damnit`
- Integration with myLog.
- Web interface coming Soon™.
- Other features: displaying logs in the GUI, export to CSV/Excel, saving matplotlib `Figure`'s.



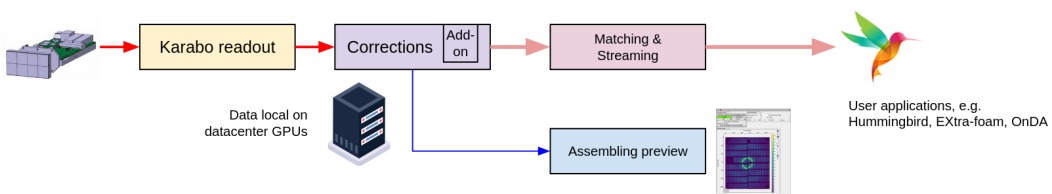
See slides from T. Michelat ▸ Experiment overview with DAMNIT

Detector corrections and automatic processing

- Convenient access to facility calibration data.

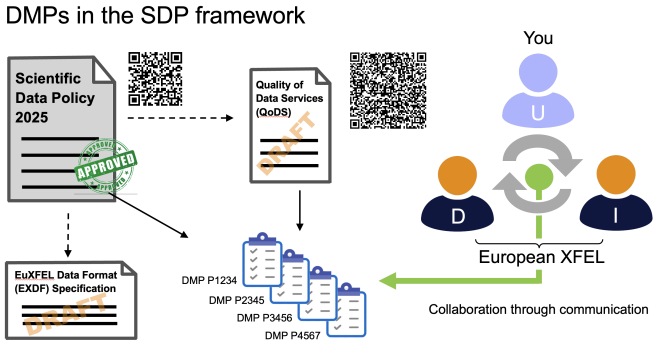
```
from extra.calibration import AGIPDConditions, CalibrationData
cond = AGIPDConditions(memory_cells=352, integration_time=18)
caldata = CalibrationData.from_condition(cond, 'SPB_DET_AGIPD1M-1')
caldata['Offset'].ndarray()  # Obtain current calibration constants
```

- Integrate your own online analysis code early in high-performance stages.



Scientific Data Policy (SDP) and Data Management Plan (DMP)

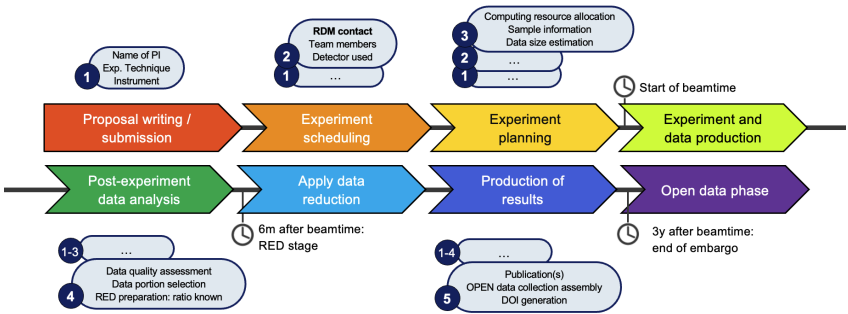
- The new SDP has been approved and will come into effect in 2025:
  `https://www.xfel.eu/users/policies/index_eng.html`
- Supplementary documents define the Quality of Data Services (QoDS) and the EuXFEL data format.
- DMPs are anchored in the SDP and regulate practical aspects of the QoDS in a proposal-specific way.
- DMPs are means to improve services.



See slides from K. Wrona ▸ The future scientific data policy

Data Management Plans in the proposal lifecycle

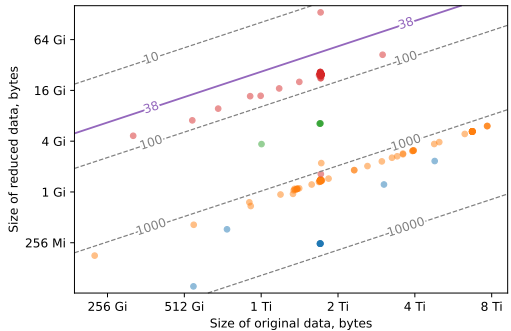
- DMPs will be integrated to the proposal workflow, incrementally adding information.
- Updates will be possible any time.
- A DMP pilot phase with voluntary user participation will start in the first half of 2024.
- One main aim of our DMP implementation will be to make users' life as easy as possible, and not to impose another bureaucratic burden.



See slides from F. dall'Antonia ▸ The data management plan

Data Reduction

- The first data reduction methods are already routinely applied.
- More methods are under development and testing.
- The data reduction based on users selection is testing with volunteered user groups.



See slides from E. Sobolev ▸ Data reduction activities at European XFEL

Data access and analysis tools

- EXtra library as single entry point for all facility-provided analysis tools.
  - EXtra-data, -geom, karabo-bridge and more in the future.
  - Remaining available under their separate package.
- Components for high-level data access.
  - Wrap specific Karabo data sources, e.g. XGM, Scantool, ADQ digitizer, ...
  - Interpret facility-specific information, e.g. pulse pattern, timing, ...
  - Generic functionality, e.g. Scan, ...

See tutorial from T. Kluyver

▸ Introduction to offline data analysis

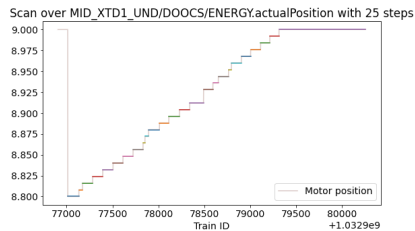
```
from extra.data import open_run
from extra.geom import AGIPD_1MGeometry
from extra.karabo_bridge import Client

data = open_run(proposal=3333, run=10)
geom = AGIPD1MGeometry()
```

```
from extra.components import Scan
from extra.components import XrayPulses

# Obtain current calibration constants
pulses = XrayPulses(data)
pulses.pulse_ids(); pulse.search_pulse_patterns()

# Detect scans and separate into parts.
scan = Scan(data.alias['jet-motor'])
scan.info(); scan.positions; scan.steps
```



Data access with EXtra-data

- Access to physical units in data saved in 2023 and later, and many more new APIs across `DataCollection`, `SourceData` and `KeyData`.
- Source and key aliases defined interactively or through files `KeyData`-like API for multi-module detectors.
- Visualize trains and pulses that have data or are missing it.

See tutorial from T. Kluyver

▸ Introduction to offline data analysis

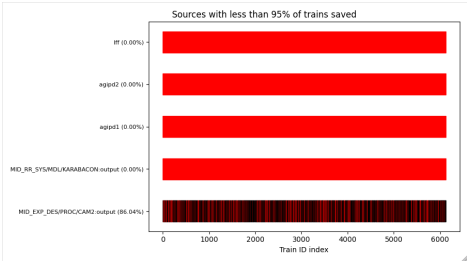
```
data = open_run(...)
data['SA3_XTD10_XGM/XGM/DOOCS',
     'pulseEnergy.photonFlux'].units
> 'μJ'
```

```
data = open_run(...).with_aliases('p3333.yaml')
jet_pos = data.alias['jet-motor',
                     'encoderPosition'].as_single_value()
pulse_energies = data.alias['pulse-energy'].ndarray()
```

```
det = AGIPD1M(open_run(...))
# Accessing all modules at the same time.
im = det['image.data']
im.dimensions, im.shape, im.ndarray(), im.dask_array(), ...
```

Data access with EXtra-data

- Access to physical units in data saved in 2023 and later, and many more new APIs across `DataCollection`, `SourceData` and `KeyData`.
- Source and key aliases defined interactively or through files `KeyData`-like API for multi-module detectors.
- Visualize trains and pulses that have data or are missing it.



See tutorial from T. Kluyver

▸ Introduction to offline data analysis

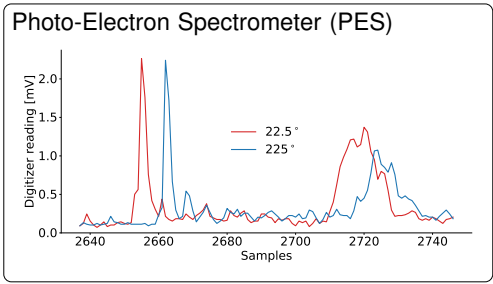
Software environments

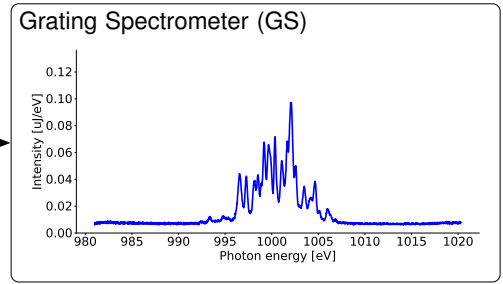
- Per-cycle environments in use since ∼ May last year.
  - `module load exfel exfel-python` → loads the current environment.
  - `module load exfel exfel-python/202301` → previous cycle environments remain accessible.
- Moved to Python 3.10 for 202401.
- Environment definitions are on Zenodo for easy citation.
  - `https://zenodo.org/records/10548701` DOI 10.5281/zenodo.10548701
- Improved content of automated environment documentation, lists:
  - Major version changes for packages.
  - Additions/removals.
- *Federated* documentation: environments docs integrated into user docs: `rtd.xfel.eu/docs/data-analysis-user-documentation/en/latest`

See slides from L. Maia and R. Rosca

▸ Basics: services, resources, and data access

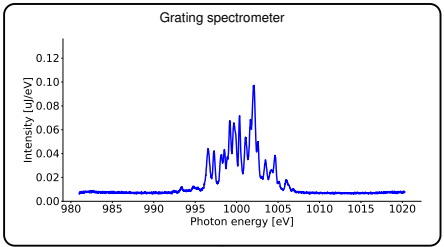
Data quality improvements: Virtual Spectrometer

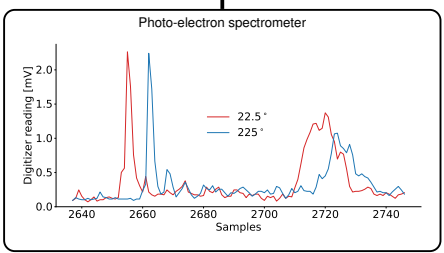




- Low resolution.
- Complex calibration.
- Non-invasive.
- Pulse-resolved.

- High resolution.
- Simple calibration
- Invasive.
- Train-resolved.

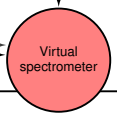


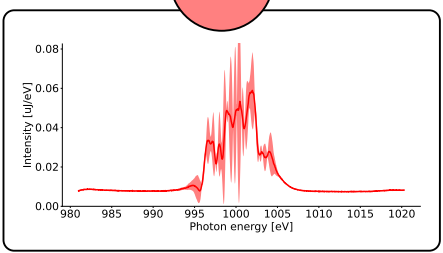


Training and inference

Training

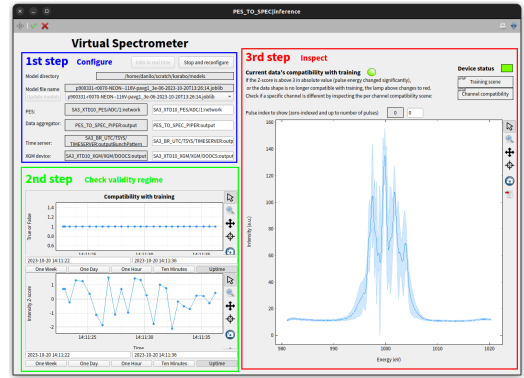
Training and inference





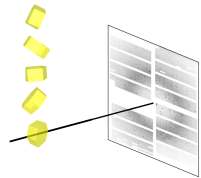
Virtual Spectrometer's status

- Deployed for automated PES calibration online and offline.
- Characterised resolution and observed significant improvement relative to the PES.
- Uncertainty informs on model validity.
- Operates as a *virtual spectrometer*: during inference collects data from XGM and PES to provide a processed output on-the-fly.
- Also works as an offline tool (in Jupyter notebooks) for user analysis.
- Further extensions for hard x-rays and extropolating to other PES settings.



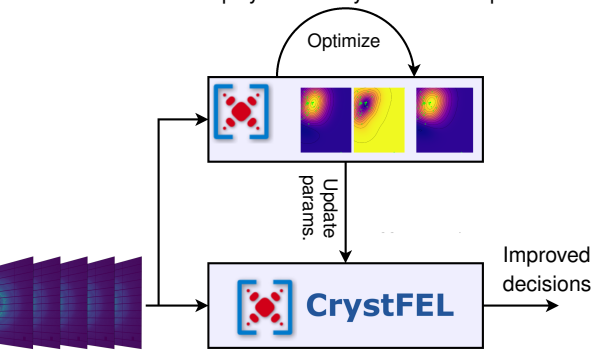
Coming soon: Streamlining data analysis

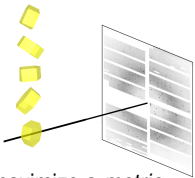
- Often data analysis pipelines have parameters.
- **Idea:** Simplify data analysis for non-experts.



Coming soon: Streamlining data analysis

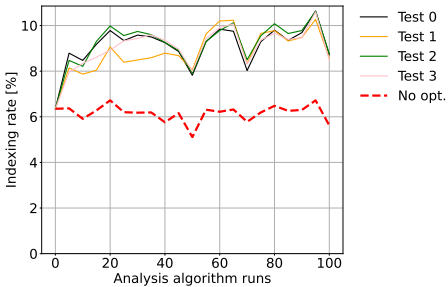
- Often data analysis pipelines have parameters.
- **Idea:** Simplify data analysis for non-experts.

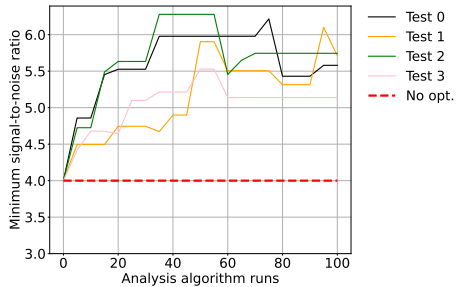




- **Goal:** Tune parameters to maximize a *metric*.
- This example: maximize the fraction of indexed frames in SFX.
- *Online*: fast feedback, higher success chances.
- *Offline*: improved scientific findings.

Testing streamlined data analysis

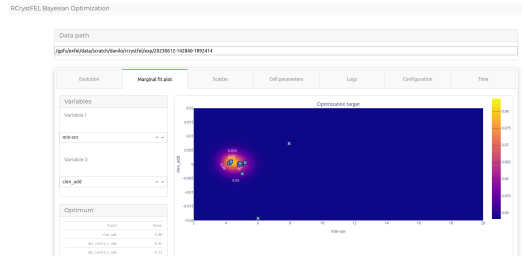




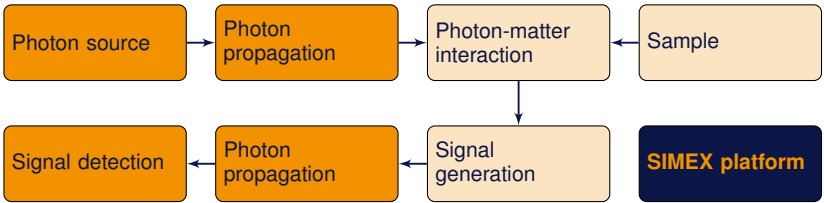
- Hen Egg-White (HEW) Lysozyme.
- AGIPD detector at EuXFEL SPB/SFX.

Automated SFX optimization status

- Automated online analysis.
- Allows for quasi-online data quality assessment in SFX.
- Tested online and deployment on-going in the online cluster.



Simulation of experiments



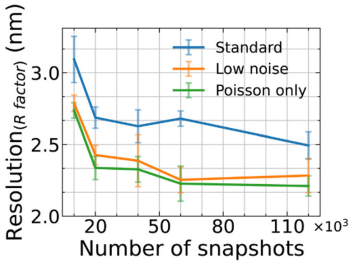
The SIMEX platform is being developed to simulate experiments at advanced laser and X-ray light sources.

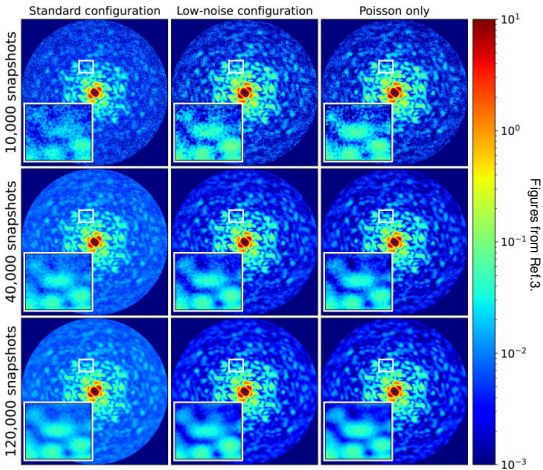
- SIMEX is a modular framework integrating advanced simulations tools,
- which are exposed through a Python API available at `https://github.com/PaNOSC-ViNYL/SimEx-Lite`.

SIMEX for Single particle imaging

SIMEX has been used so far mostly to study the effect of

- X-ray pulse duration[1],
- radiation damage and inelastic scattering[2], and
- detector (AGIPD) noise[3].





1. C. Fortmann-Grote et al. 2017. IUCrJ 4: 560
2. J. E et al. 2021. Sci. Rep. 11: 17976
3. J. E et al. 2022. Struct. Dyn. 9: 064101

SIMEX for you

Software integrated and actively developed into SimEx.

Module	Target	Tool
Photon source	FEL process	FAST-XPD
Photon propagation	Coherent wavefront propagation	WPG/SRW
Sample	Molecules, soft matter	GROMACS
X-ray photon-matter interaction	Atoms, molecules and clusters	XMDYN, XATOM
Signal generation	Scattering: molecules and clusters	SingFEL
	Crystal diffraction	CrystFEL pattern_sim
	Powder-like scattering	DebyeCalculator

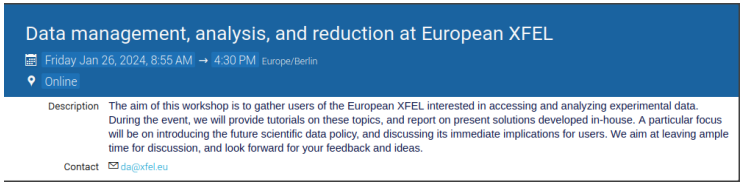
Current and future activities:

- Integration of more tools,
- development of a web interface to offer simulations as a service, primarily to improve experiments planning.

Contact: Juncheng E <juncheng.e@xfel.eu>

Summary

- Lot of developments on IT, Data Science and Simulation.
- Next presentations go into details on many of those topics.
- You are always welcome to contact us in case of doubts.
- All material is available at `https://indico.desy.de/event/42904/timetable/`.

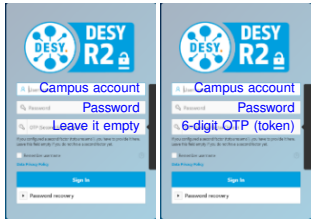
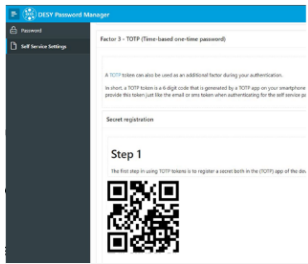


Supplementary slides

Setting up 2nd factor authentication remotely

- DESY-IT provides full documentation under:
  - `https://it.desy.de/services/mfa/naf__maxwell/index_eng.html`
- Access to this documentation requires login with your XFEL campus account and password.
  - If you do not remember your campus account password or it expired, you can still use the UPEX service to reset it.
  - The UPEX service for resetting passwords will be deprecated at some point.
- The term “DESY” used in the documentation also refers to the European XFEL.
- You may be asked to confirm the condition of use and Central Works’ agreement before you can set the 2nd authentication factor.

Setting up 2nd factor authentication remotely by users

- Use the following link to set the 2nd authentication factor:
  - `https://passwd.desy.de`
- For remote access to the data analysis facility (Maxwell), you must set the TOTP method
- We also recommend setting up SMS and alternative email for password recovery and login to passwd.desy.de in the
- To set the TOTP, you will need a smartphone and an Authenticator app (see documentation)
- You will need to use the TOTP as 2nd-factor authentication when accessing Maxwell by:
  - ssh to Maxwell display node (e.g. `max-exfl-display.desy.de`, `max-display.desy.de`)
  - Using FastX on the Maxwell display node.
  - Using the JupyterHub service at `https://max-jhub.desy.de`.

Setting up 2nd FA. To change your password.