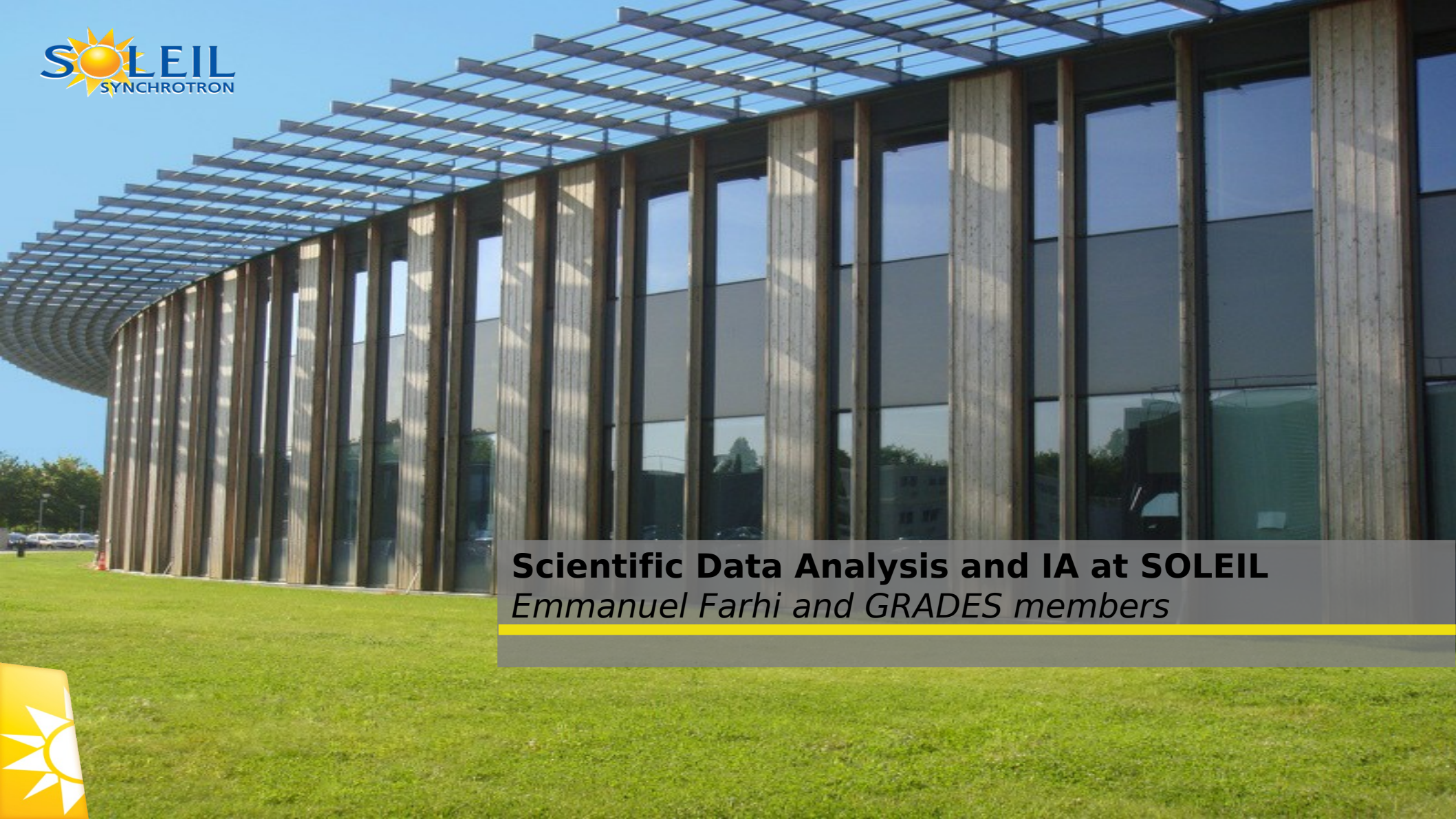


Scientific Data Analysis and IA at SOLEIL

Emmanuel Farhi and GRADES members

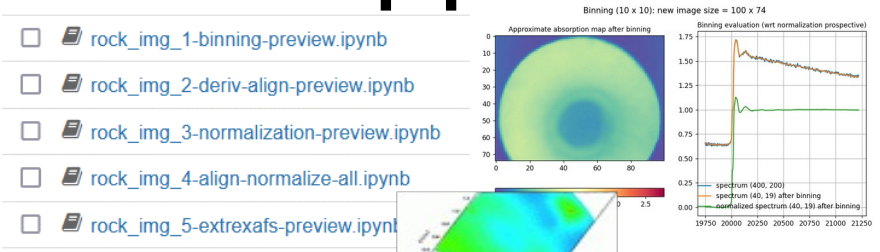
Our group

- *Groupe de Réduction et d'Analyse de Données Expérimentales de SOLEIL* (**GRADES**). Created in Feb 2020, in the EXP Div.
- Help beam-line users and scientists to better treat the data and increase the scientific output.

Emmanuel	**Frédéric**	**Olga**	**Anass**	**Stéphane**	**Majid**	***Gamil*** (LEAPS-INNOV)	***Lucia*** (BIG-MAP)	***Aurélien*** (Student)
Head, spectro, BL sim	Diff (XRD, MX, surf)	Abs, ML (GA)	AI/DL Tomo, ARPES	RIXS, instrum	EU projects	Data Compression	Abs/spectro	Req. capture doc
DARTS, McXtrace samples, MD/DFT, ...	BINoculars DIALS Debian, Notebooks,	Notebooks Dadimodo	PyTorch, TF	Igor, python, McXtrace	Meetings	LZ4/ ZSTD/ HDF5	Notebooks (batteries)	Django, optim.

Beamline support: take-away

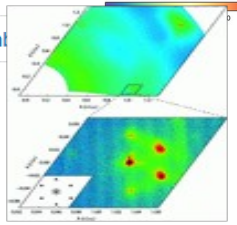
Absorption spectroscopy [Olga,Lucia]
- Notebooks (ROCK, SAMBA, LUCIA)



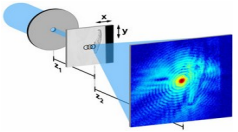
Diffraction [Fred, Olga]
- BINoculars, Notebooks (MARS, DIFFABS)

MX [Fred]
- DIALS, auto-processing/ISPYB (PX1/PX2a)

Ptychography [Fred, Emmanuel, Anass]
- PyNX (HERMES, SEXTANTS, CRISTAL), segmentation for HeloBio

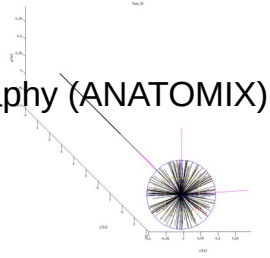


Daily support and advices (all members).



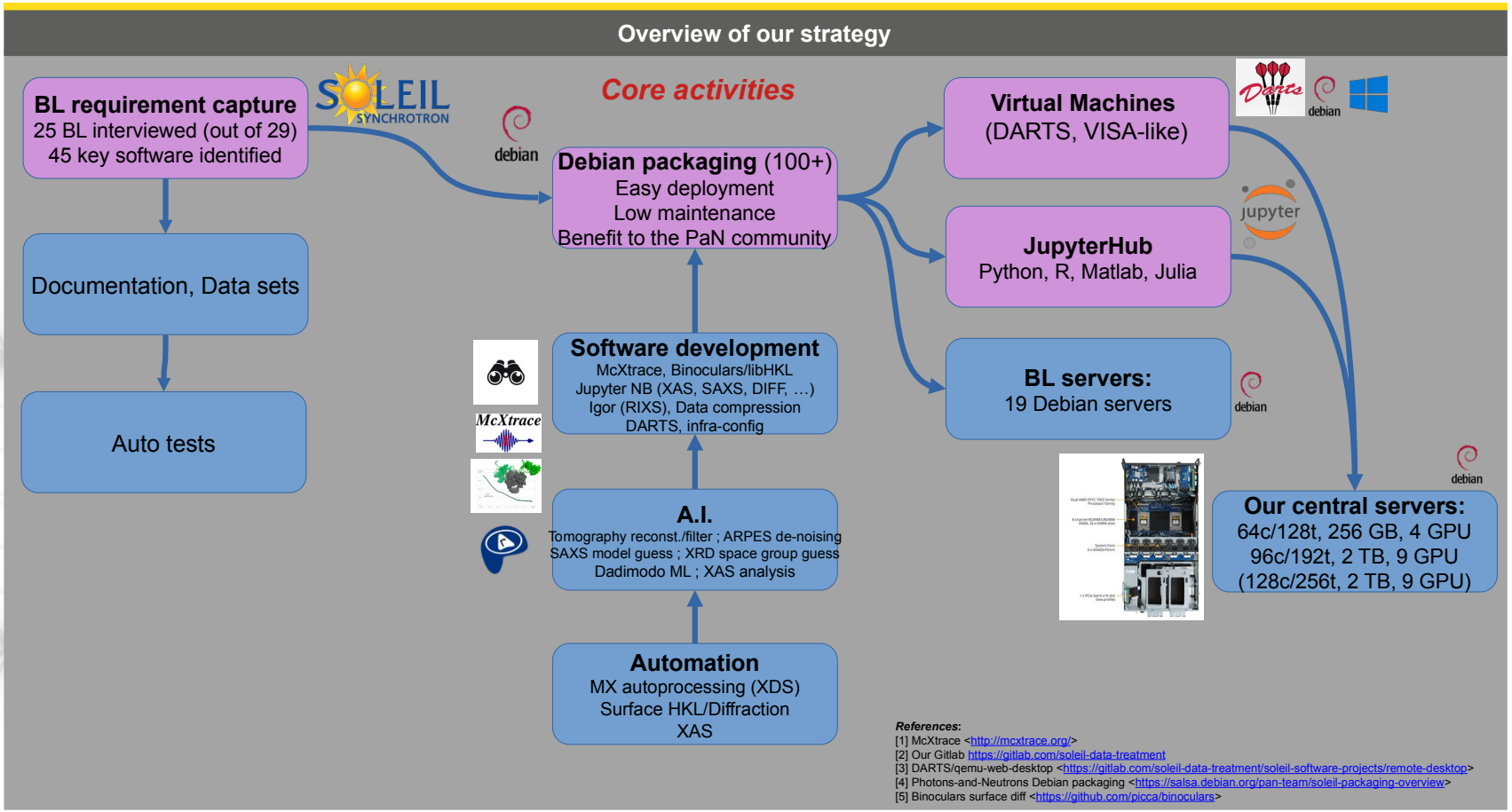
AI /ML [Anass, Olga]
- **Tomography** AI filtering/reconstruction/segmentation (ANATOMIX,PSICHE)
- **SAXS/XRD** AI classifier to guess best model and space-group resp. (SWING,MARS)
- **ARPES** denoising into NAVARP (CASSIOPEE)
- **SAXS** model tertiary structure with Dadimodo (SWING)

Beam-line end-to-end simulation (McXtrace) [Steph, Emmanuel, Aurélien]
- Abs Edge modelling and glitches (ROCK), VUV spectrum (DISCO), IXS *S(q,w)* dispersion (GALAXIES), tomography (ANATOMIX), MX (PX2a)



Data Compression [Gamil, Majid]
- LEAPS-INNOV: data compression (lz4,zstd,hdf5plugin)
- ExPaNDS/PaNOSC: data management (within IT Strategy group)

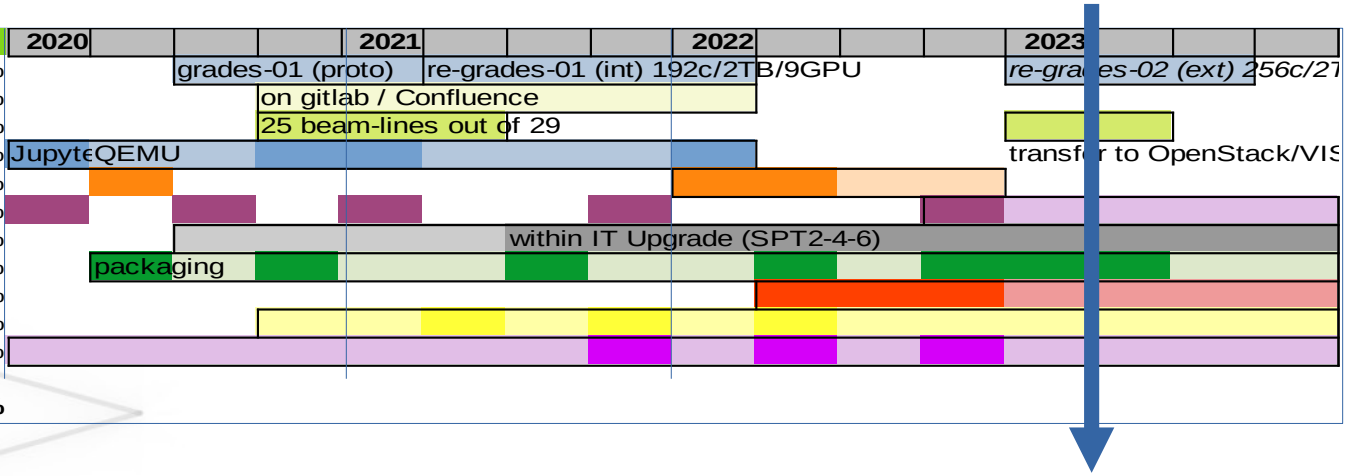
GRADES main tasks



GRADE-One 2020-2023

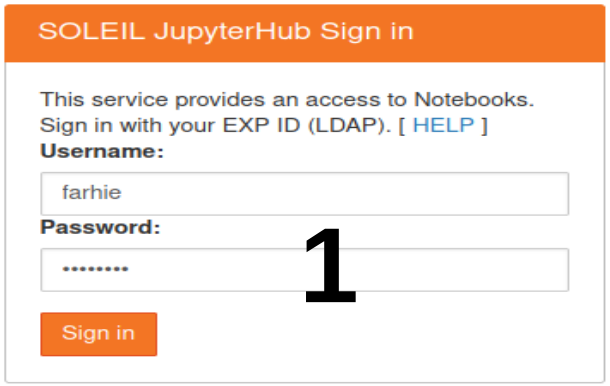
8.5 man.year effort initiated in Feb 2020 → 2023

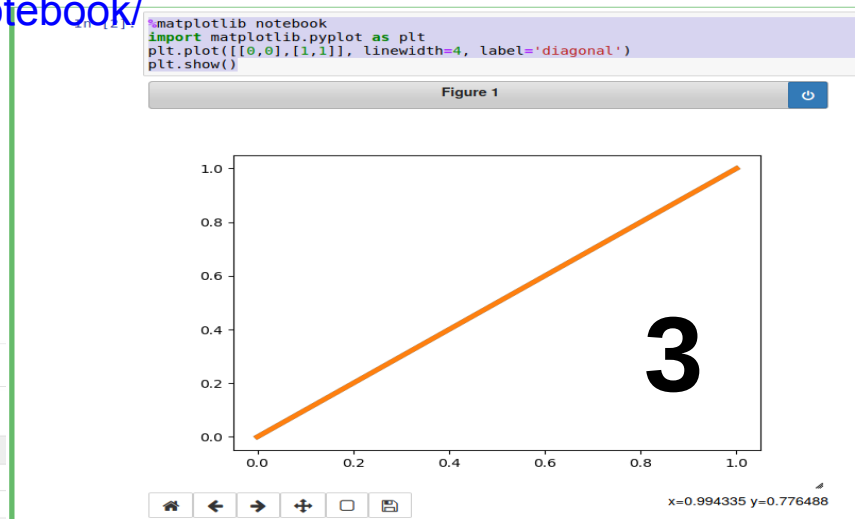
Task	Completion
Computing resources	100,00 %
Documentation	100,00 %
Scientific Apps (requirement capture)	86,00 %
Cloud infrastructure (web, VM)	100,00 %
Cloud services (Jupyter/DARTS, VISA)	100,00 %
Automatic processing	100,00 %
Treatment traceability	0,00 %
Maintenance	100,00 %
Collaborations (PaNOSC/ExPaNDS) VI	50,00 %
IA	50,00 %
Scientific Apps (coding)	100,00 %
Total	**80,55 %**



JupyterHub

https://re-grades-01.exp.synchrotron-soleil.fr/notebook
https://re-grades-02.exp.synchrotron-soleil.fr/notebook/





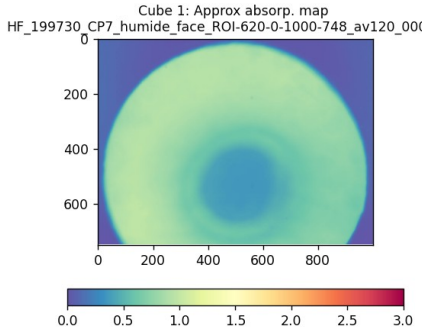
Deployment in minutes.
Access to GPUs.
Access to the Ruche.
Many libraries.

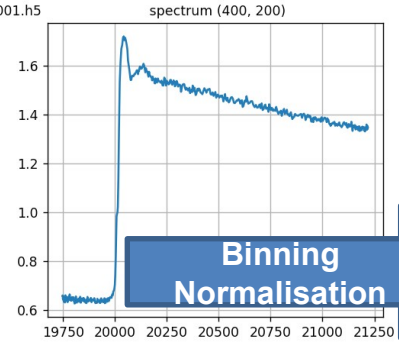
Jupyter Notebooks for ROCK

O. Roudenko

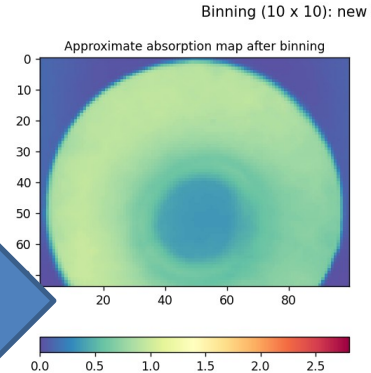
5 notebooks

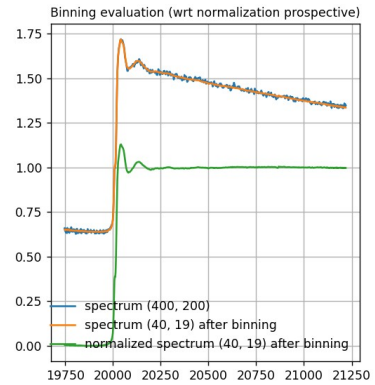
- rock_img_1-binning-preview.ipynb
- rock_img_2-deriv-align-preview.ipynb
- rock_img_3-normalization-preview.ipynb
- rock_img_4-align-normalize-all.ipynb
- rock_img_5-extrexafs-preview.ipynb



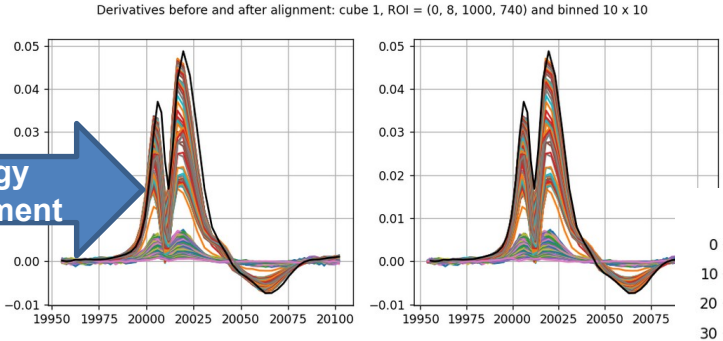


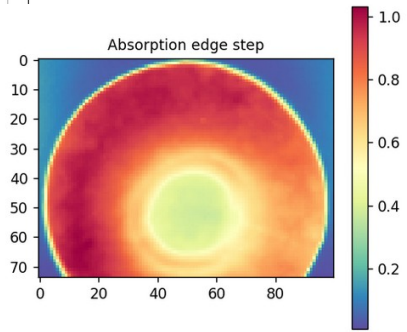
Binning Normalisation



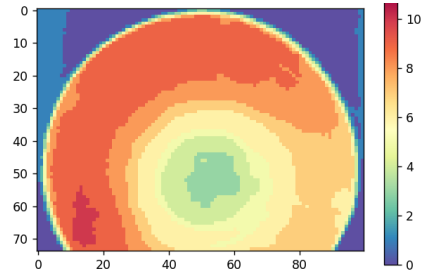


Energy Alignement



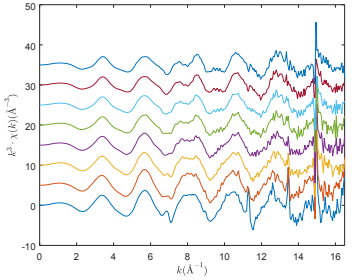


EXAFS



⇒ At each pixel

⇒ Clustering by edge step



Increasing Edge step

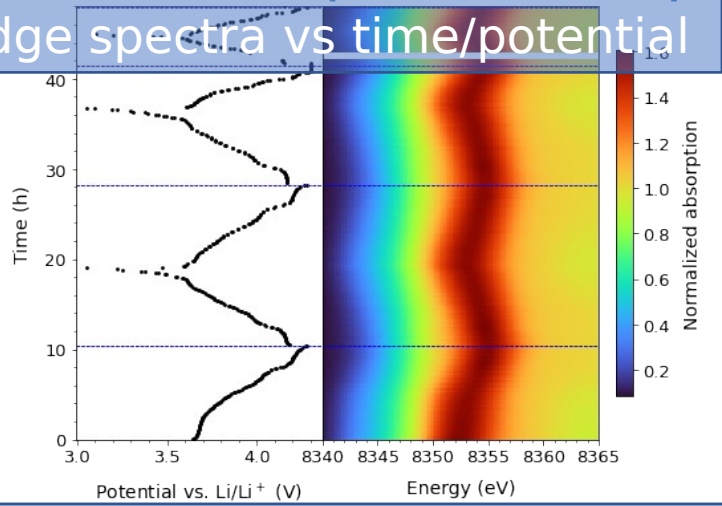
Jupyter Notebooks for BIG-MAP (batteries)

L. Perez

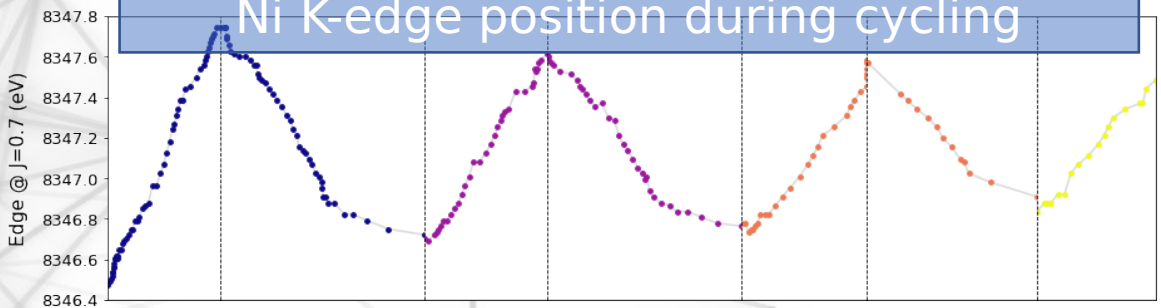
2 notebooks

Data aggregation and **visualisation** for multitechnique experiments (e.g. **XAS** + **electrochemical cycling**)
Chemometric methods (PCA and MCR-ALS) to resolve chemical constituents

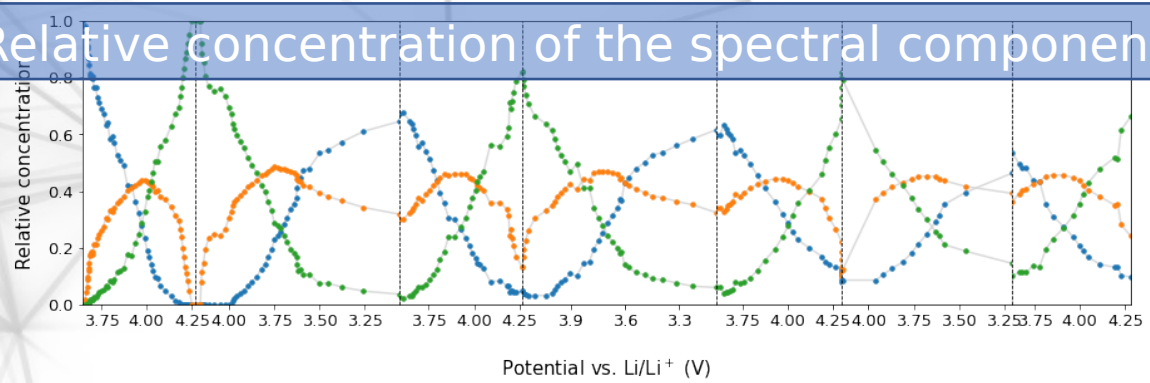
Ni K-edge spectra vs time/potential



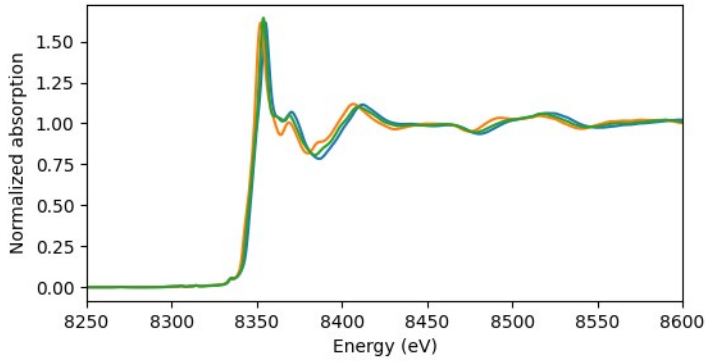
Ni K-edge position during cycling



Relative concentration of the spectral components

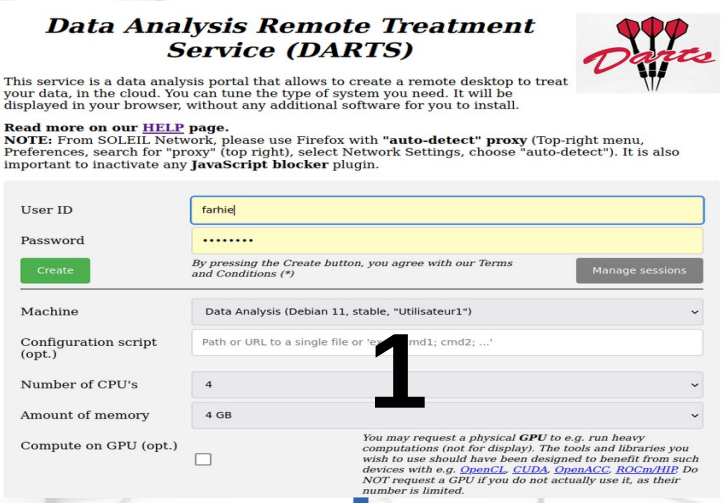


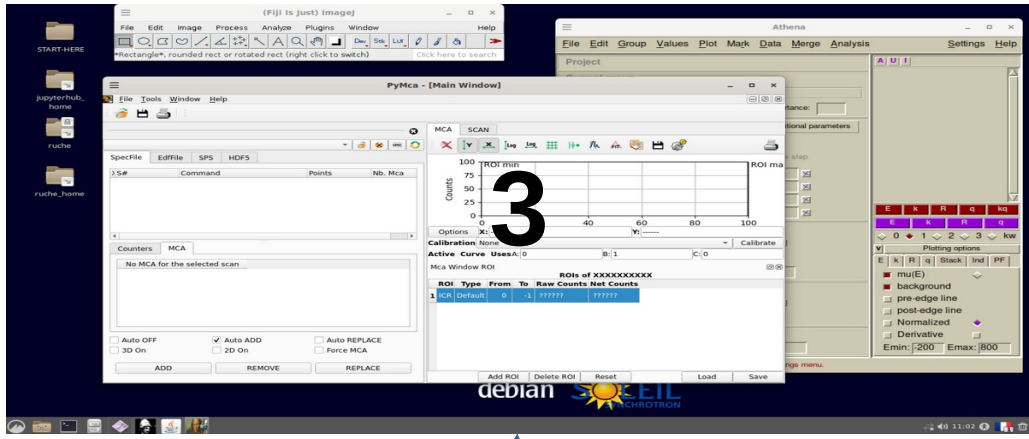
Spectral components obtained by MCR-ALS

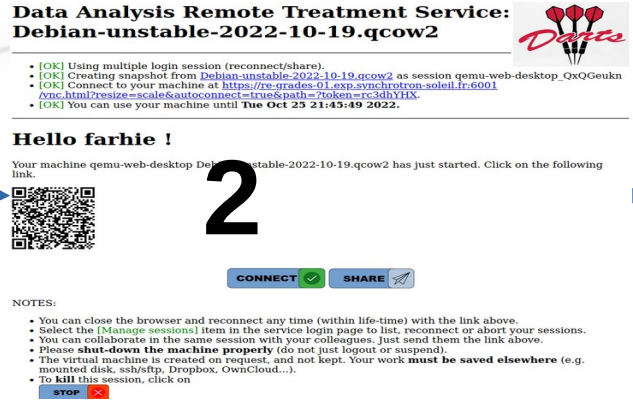


Data Analysis Remote Treatment Service

https://re-grades-01.exp.synchrotron-soleil.fr/desktop
https://re-grades-02.exp.synchrotron-soleil.fr/desktop/



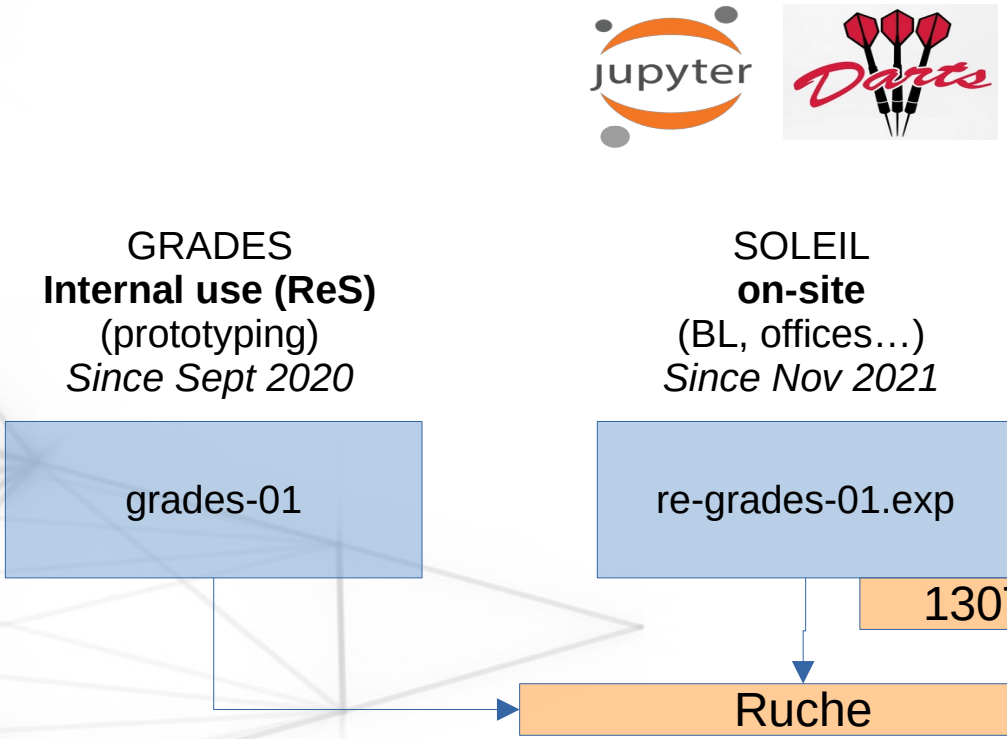




DARTS is simple and efficient.
Minimal maintenance, fully configurable.

Deployment in minutes.
Scalable.
QR code.
Access to GPUs.
Access to the Ruche.
Access to JupyterHub area.
Open-source/Debian.
100+ of sci software in VM

GRADES services availability



576 CPU threads, 4 TB mem, ~200 TB disk, 17 GPU's

Our current resources are dimensioned for up to ~50 users.

DARTS/JupyterHub at SOLEIL

We pushed ~90 scientific software packages into Debian/Ubuntu.
https://salsa.debian.org/pan-team/soleil-packaging-overview

SAXS: Foxtrot, SasView, PySAXS, DENSS, BioXTAS-RAW, ATSAS , cSAXS Ptycho, SAXSGUI(matlab)
XRD Diffraction: Pyfai, Dioptas, FullProf, XSOCS
MX: ADXV, Relion, AlphaFold, XDS
Cryo-EM: IMOD, SIMPLE
Surface/Reflectivity: GenX, BornAgain, Binoculars
Ptychography: PyNX, Gwaihir/BCDI, cSAXS PtychoShelves/Matlab
Absorption: Fastosh, Xray LARCH , Artemis/Athena , FEFF/iFEFFit, FDMNES, MCR-ALS (matlab)
XRF fluo: pymca, Mantis-XRAY, XrayLib
Multi-spectral: HyperSpy, Mantis-XRAY, ParaView, Mayavi2
ARPES/XPS/PEEM: Axis2000, NavARP, ARPYS, PyImageTool
RIXS/IXS: XRSTools, codes DFT, Quanty/Crispy
IR spectroscopy: Orange/Quasar, SPView
Imaging: Fiji/ImageJ, Mantis-XRAY, silx view, pymca, Gwyddion
Tomography: PyHST2, UFO, Orange/Tomwer/Nabu, astra-toolbox, IMOD
Simulation: McXtrace, Orange/OASYS , XRT, MuMax3, GEANT4, codes DFT/MD
For all: Jupyter, pip, conda, Silx, Fiji, Matlab, Labplot (Origin), Code::Blocks, VSCode, ...

DARTS software

We also provide other environments :
- **Reduced env as take-away**
- **Windows 10 (with Igor pro, Crysalis, CasaXPS, OPUS, ...)**

Available for all beam-lines at SOLEIL Synchrotron.

What is AI ?

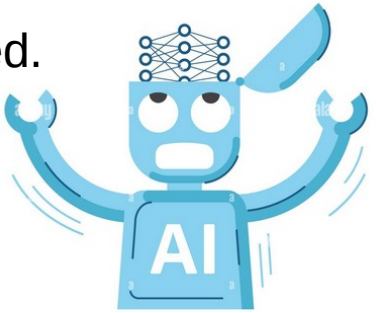
Any methods that mimic an “intelligent” behaviour (AI = Artificial Idiot).

Old-style AI/ML:
problem optimisation, decision tree (PCA, k-means), advanced signal processing.
New-style AI:
Deep learning → neural networks with billions of parameters, to “fit” a problem.

- Training = store knowledge in the neural link weights (the “memory”), output → input
- Inference = use the neural network, input → output

Implications:
1) Everything outside the “training” will fail.
2) The training data set must be huge and the “output” must be well defined.

Where neural networks really work:
- Image processing (image → image, *e.g.* denoising, segmentation)
- Classification (image → ID of a class, *e.g.* “this is a cat”)
- Decision tree (data correlation → “close” objects, e.g. chatGPT)



AI applicability to X-rays

AI methods can be used for some automatic pre-processing tasks.

Denoising (remove Poisson-noise), help further data treatment and compression.

Beam-time optimisation (find best configuration).

Classify data sets – requires large training sets and associated classes.
e.g. estimate space-group, XAS/fluo element, SAXS model
can also use simple clustering (PCA, k-means, …)

Correlate some data sets with others (similarity, extrapolation).

Chatbot to help scientists at 3 am, or provide tips for on-going experiments.

A. Bellachehab

AI/ML activities (1/2)

Evaluate the applicability of modern AI techniques for our data sets.

GPU

Tomography

De-noise tomogram

3D vol reconstruction

3D vol segmentation

PyHST2

GPU

SAXS

DL/CNN

Predict initial SASView model from exp. Data *I(q)*

SasView

GPU

ARPES

DL/CNN

De-noise *(hk, ω)*

Integration into **NAVARP**

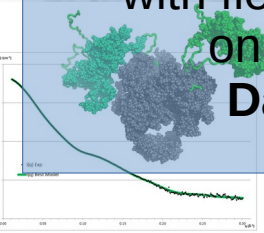
PyTorch

SAXS

Genetic Algorithm

Refine tertiary model with flexible proteins on exp. Data

Dadimodo

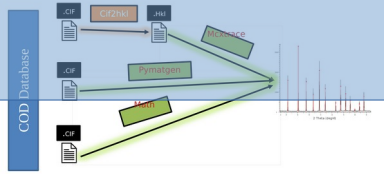


GPU

XRD

DL/CNN

Predict space group from exp. data



GPU

Accelerator

Build approx stability model from MOGA-ELEGANT

To speed-up Upgrade est.

AI/ML activities (2/2)

Evaluate the applicability of modern AI techniques for our data sets.

PyTorch
Generic AI

GPU

Scikit-learn
Generic ML

GPU

AlphaFold
Protein tertiary structure

GPU

ChatGPT / LLaMa
AI chatbot
for late evening help
no one around ?

GPU

DragonFly
Image segmentation
3D volume analysis

GPU

Aydin
Image denoising
(any 2D detectors)

https://confluence.synchrotron-soleil.fr/display/EG/Denoising+methods

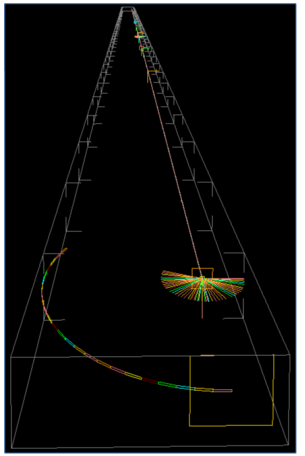
Train AI with digital twins

Getting training data sets is an issue.
We have plenty of experimental data sets, not annotated.

A solution:
1) build a digital twin,
2) use it to generate annotated data sets
3) train a CNN with that data

Example:
- A basic SAXS (McXtrace) with a set of sample models (SASView)
- An XRD BL (MARS, McXtrace) with a set of diffractograms (CIF)

McXtrace 2023 school
https://indico.synchrotron-soleil.fr/e/mcxtrace2023



Upgrade TDR IT / IA

SPT6.1: Étude des applications IA pour SOLEIL
- État de l'art.
- Recherche de besoins des lignes/accélérateur pouvant être réalisés.

BL : Le GRADES expérimente en continu des méthodes/logiciels et les met à disposition.

SPT6.4: Exploitation des données scientifiques
- Recensement des usages sur les lignes.
- Plan de déploiement de logiciels pour les lignes.

BL : Le GRADES continuera ces actions.

SPT6.5: Simulations / jumeaux numériques
- Développements d'outils de simulation "bout-en-bout" (assemblages).

BL : Le GRADES développe McXtrace (ech. absorption, fluorescence, inélastique, diff…).
BL : Le GRADES met à disposition des moyens de calcul et des logiciels de simulation.

Things to improve

Our services have a limited bandwidth: 400 MB/s ← → Ruche
- IT will somehow remedy to this.

DARTS was not as stable as today mid-2022.
- Now solved.
- But as any IT service, we sometimes need to perform maintenance operations.

Computing capacity was limited.
- Now provide more computing power (incl. A6000 48 GB GPU).

Some software may not work properly, or is unavailable.
- We maintain 100+ software. We do our best. Please tell us !

Perspectives

The GRADES will continue to deploy the software required by the BL.

This includes providing hardware resources (GPU), libraries, services, fed by continuous BL requirement capture and support. AI is part of our job.
AI tools are spreading among our lives. Soon, ChatGPT will be everywhere.
Beware of the limitation

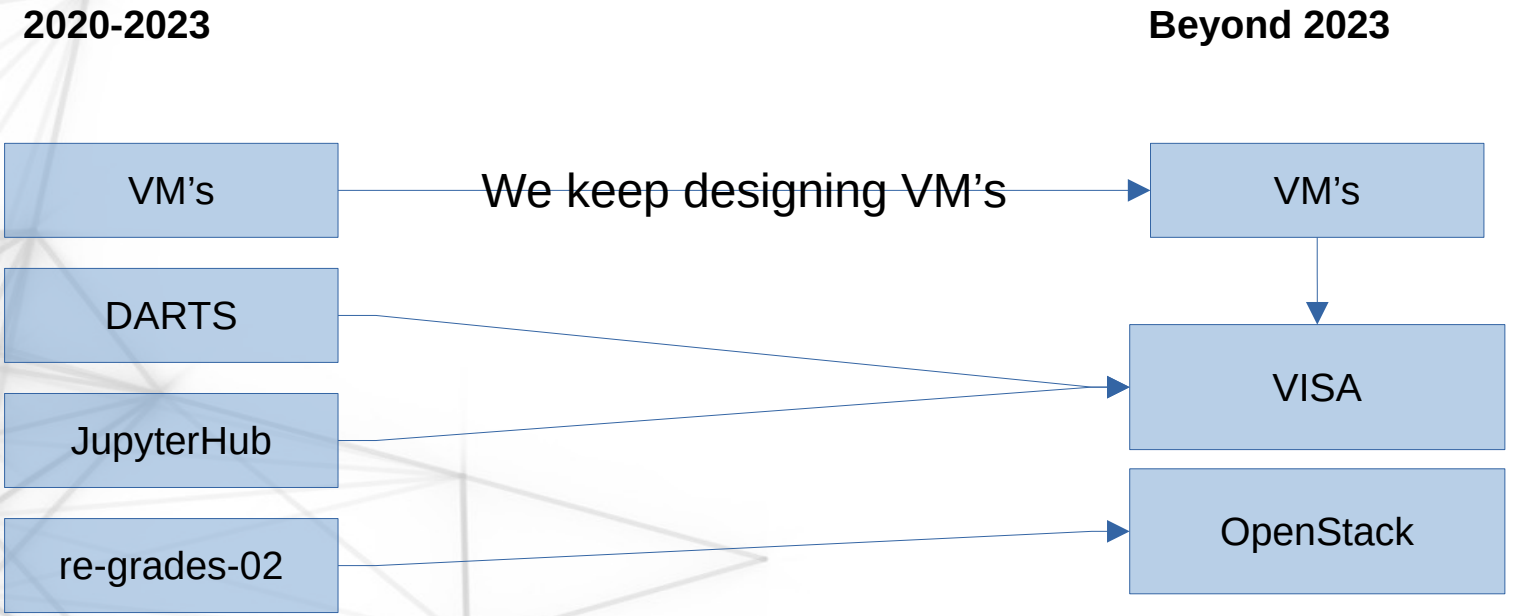
Soon to arrive in DARTS:

- TensorFlow
- Weka
- Scipion
- *DragonFly (→ in W10 VM)*
- *SAMSON-Connect (→ in W10 VM)*

And more SOLEIL beam-line McXtrace models with samples:

- 4D crystal IXS sample (currently only 2D) ← DFT/MD
- ARPES/XPS sample approximate (2D/4D) ← DFT/MD

DARTS and VISA (PaNOSC/ExPaNDS)



IT services will provide:

- Authentication
- Access to data (Ruche and Homes)
- Network/storage performance

Software packaging is essential: minimize maintenance, ease deployment, benefit to the World.