Data Exploration and Analysis with Jupyter Notebooks (TUCPR02)

Hans Fangohr Senior Data Analysis Scientist European X-ray Free Electron Laser (EuXFEL), Germany Professor of Computational Modelling University of Southampton, United Kingdom fangohr.github.io, hans.fangohr@xfel.eu, @ProfCompMod

ICALEPCS 2019, New York, US, 8 October 2019



H. Fangohr^{*1}, M. Beg, M. Bergemann, V. Bondar, S. Brockhauser^{2,3}, C. Carinan, R. Costa,
F. Dall'Antonia, C. Danilevski, J. C. E, W. Ehsan, S. G. Esenov, R. Fabbri, S. Fangohr, G. Flucke,
C. Fortmann⁴, D. Fulla Marsa, G. Giovanetti, D. Goeries, S. Hauf, D. G. Hickin, T. Jarosiewicz⁵,
E. Kamil, M. Karnevskiy, Y. Kirienko, A. Klimovskaia, T. A. Kluyver, M. Kuster, L. Le Guyader,
A. Madsen, L. G. Maia, D. Mamchyk, L. Mercadier, T. Michelat, J. Möller, I. Mohacsi, A. Parenti,
M. Reiser, R. Rosca, D. B. Rueck, T. Rüter, H. Santos, R. Schaffer, A. Scherz, M. Scholz, A. Silenzi,
M. Spirzewski⁵, J. Sztuk, J. Szuba, S. Trojanowski⁵, K. Wrona, A. A. Yaroslavtsev, and J. Zhu

European XFEL GmbH, Holzkoppel 4, 22869 Schenefeld, Germany ¹ also at University of Southampton, SO17 1BJ, Southampton, United Kingdom ² also at University of Szeged, Arpad ter 2, 6701 Szeged, Hungary ³ also at Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary ⁴ also at Max-Planck-Inst. for Evolutionary Biology, August-Thienem.-Strasse 2, 24306 Plön, Germany ⁵ also at NCBJ, Andrzeja Sołtana 7, 05-400 Otwock, Poland

> J. Reppin, F. Schlünzen, and M. Schuh DESY, Notkestr. 85, 22607 Hamburg, Germany

E. Fernandez-del-Castillo and G. Sipos EGI Foundation, 140 Science Park, 1098XG Amsterdam, Netherlands

> T. H. Rod, J. R. Selknaes, and J. W. Taylor ESS, Ole Maaløes Vej 3, Copenhagen, Denmark

A. Campbell, A. Götz, and J. Kieffer ESRF, 71 Avenue des Martyrs, 38000 Grenoble, France

J. Hall, E. Pellegrini, and J. F. Perrin ILL, 71 Avenue des Martyrs, 38000 Grenoble, France

Outline

Introduction Jupyter Notebook

Use cases

Summary

Data Analysis with Jupyter Notebooks (TUCPR02)

Jupyter Notebook

- Document hosted in web browser
- Combines
 - text (markdown with LaTeX support)
 computer code (Python)
 output from code
- Saved in one * . i pynb file (IPYthon NoteBook]
 combines input and output for each cell

Re-use

load, re-execute, modify

export to static formats (html, pdf, py, ...)

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

European XFEL

In [1]: # Import Python and libraries we need later
%matplotlib inline
from numpy import exp, cos, linspace
import pylab
from ipywidgets import interact

Mathematical model: We would like to understand $f(t, \alpha, \omega) = \exp(-\alpha t) \cos(\omega t)$

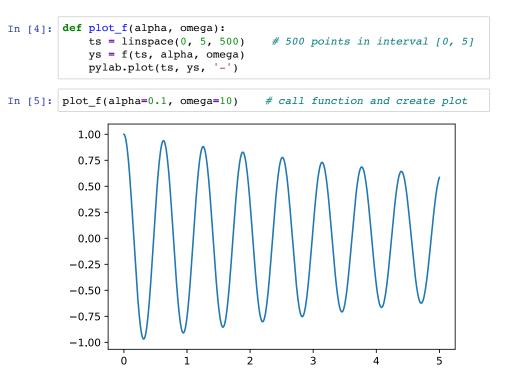
Code: Here is an implementation:

```
In [2]: def f(t, alpha, omega):
    """Computes and returns exp(-alpha*t) * cos(omega*t)"""
    return exp(-alpha * t) * cos(omega * t)
```

Interactive exploration: We can execute the function for values of *t*, α and ω :

- In [3]: f(t=0.1, alpha=1, omega=10)
- Out[3]: 0.48888574340060287

Or produce a plot (in a function plot_f so it can be re-used for different parameters):



Jupyter Notebook

- Supports a wide range of languages
 - Julia
 - Python (\rightarrow JuPyTer)
 - Shell, R, Matlab
 - C++
 - https://github.com/jupyter/jupyter/wiki/Jupyt er-kernels

A number of useful tools and extensions available Some mentioned in talk: nbconvert, JupyterHub, Binder, nbval,

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

European XFEL

In [1]: # Import Python and libraries we need later
%matplotlib inline
from numpy import exp, cos, linspace
import pylab
from ipywidgets import interact

Mathematical model: We would like to understand $f(t, \alpha, \omega) = \exp(-\alpha t) \cos(\omega t)$

Code: Here is an implementation:

```
In [2]: def f(t, alpha, omega):
    """Computes and returns exp(-alpha*t) * cos(omega*t)"""
    return exp(-alpha * t) * cos(omega * t)
```

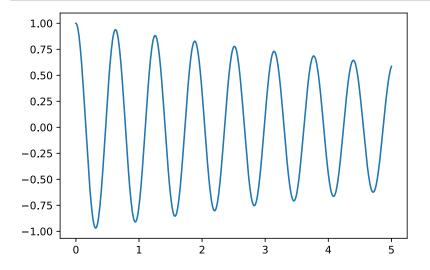
Interactive exploration: We can execute the function for values of *t*, α and ω :

- In [3]: f(t=0.1, alpha=1, omega=10)
- Out[3]: 0.48888574340060287

Or produce a plot (in a function plot_f so it can be re-used for different parameters):

```
In [4]: def plot_f(alpha, omega):
    ts = linspace(0, 5, 500)  # 500 points in interval [0, 5]
    ys = f(ts, alpha, omega)
    pylab.plot(ts, ys, '-')
```

In [5]: plot_f(alpha=0.1, omega=10) # call function and create plot



Use case 1: data analysis in notebook

- Explorative data analysis
- Convenient combination of processing, results and interpretation
- Complete capture of all computational steps
 good record for *reproducibility* and *re-use* FAIR data
- Through export to HTML, easy to share with collaborators & supervisors
- Scientists are confident drivers of this example on the right from SCS instrument

X-ray Absorption Spectroscopy

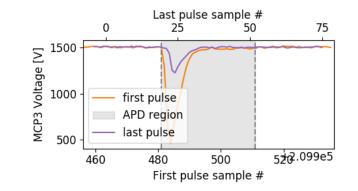
Step 1: Load data and align them by train id and pulse id

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)



Use case 2: notebooks as recipes

- Pre-populate notebook with cells to carry out a particular type of data analysis
 - provide a directory full of such recipes to users
 users execute cells during beamtime and later
- Convenient compromise between
 - static recipe (=script)interactive exploration

Experience

- keep code in notebook cells short and
 move functionality into library (here "ToolBox")
- archive directory of modified recipes with data

X-ray Absorption Spectroscopy

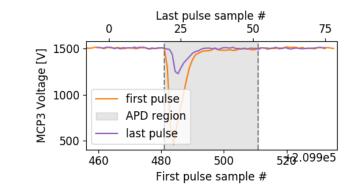
Step 1: Load data and align them by train id and pulse id

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

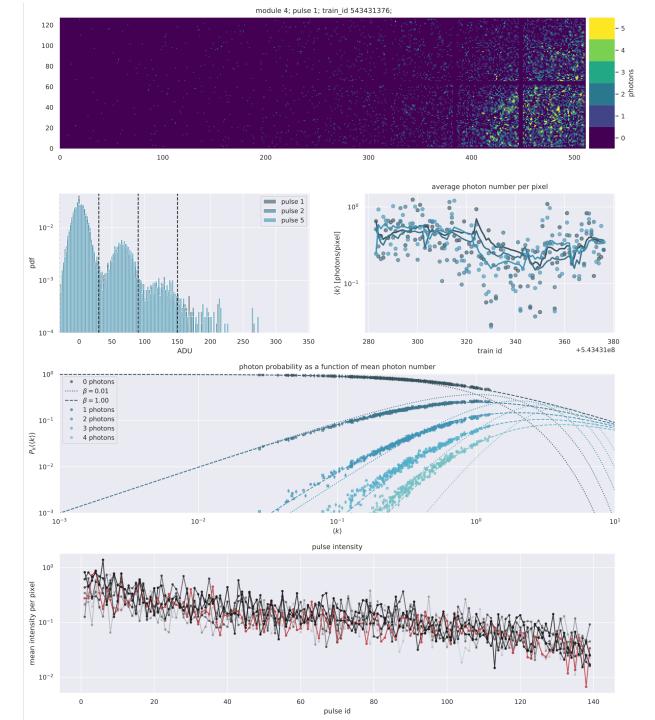
In [6]: nrj = np.linspace(nrun.nrj.min(), nrun.nrj.max(), 80)
xas = tb.xas(nrun, nrj, plot=True)



Use case 3: online visualisation

- Use Jupyter Notebook to
 fetch live experiment data from network
 carry out some processing and update plots
- "plots" are chosen from library of processing and visualisation units
- Selection of plots, and distribution in rows and columns can be done by instrument scientist
 Attractive because it provides flexibility
- Piloted at MID instrument at EuXFEL (Mario Reiser, Johannes Möller)

European XFEL



Data Analysis with Jupyter Notebooks (TUCPR02)

Use case 4: notebooks as a script

Use Jupyter Notebook as a script can execute using nbconvert to take commands in notebook, execute them, save resulting notebook

can create data files and plots in process

- Use case: detector calibration pipeline use of nbparametrize to insert run parameters into notebook before execution executes on HPC facility
 - after execution, turn notebook into pdf report
 - error messages embedded in output

Cjupyter LPDChar_Darks_NBC (autosaved)	Cogout
File Edit View Insert Cell Kernel Widgets Help	usted Python 3 O
E + S 2 E ↑ ↓ NRun ■ C > Markdown ↓ ■	

Offset, Noise and Dead Pixels Characterization Author: M. Karnevskiy, S. Hauf This notebook performs re-characterize of dark images to derive offset, noise and bad-pixel maps. All three types of constants are evaluated per-pixel and per-memory cell. The notebook will correctly handle veto settings, but note that if you veto cells you will not be able to use these offsets for runs with different veto settings -The evaluated calibration constants 'Offset, Noise and Dead Pixels Characterization' Documentation, Release The offset (O) is defined as the med cell (c). 4.9 Variation of offset and noise across Tiles and ASICs The noise N is the standard deviati The following plots show a standard deviation σ of the calibration constant. The plot of standard tiles show pixels of one tile (128×32) . Value for each pixel shows a standard deviation across 1 The bad pixel mask is encoded as a deviation across ASICs are shown overall tiles. The plot shows pixels of one ASIC (16×32) , w a standard deviation across all ACIS of the module. "OFFSET_OUT_OF_THRESHOLD": Offset outside of bounds: 4.9.1 Variation of offset and noise across ASICs - High gain $M(O)_{X,V} - \sigma(O)_{X,V}$ thre or offset outside of hard limits Media thres ____ Max "NOISE_OUT_OF_THRESHOLD": Noise outside of bounds: $M(N)_{X,V} - \sigma(N)_{X,V}$ three or noise outside of hard limits 3 8 8 thre σ Offset [ADU] "OFFSET_NOISE_EVAL_ERROR": 2.0 2.5 σ Offset [ADU 7 Noise (ADU Offset and Noise both not nan value Values: thresholds_offset_sigma 4.9.2 Variation of offset and noise across ASICs - Medium gain given as parameters. In []: cluster profile = "noDB" # Median in folder = "/gpfs/exfel/ex - Min

- Max

out folder = "/gpfs/exfel/d

capacitor_setting = 5 # cap run high = 112 # run number run med = 113 # run number run low = 114 # run number

sequences = [0] # sequence modules = [-1] # list of mo

Use case 5: (remote) data analysis environment (JupyterHub)

JupyterHub

users connect through browser and https
serve notebooks on facility hardware
use existing authentication systems
connect to users' file storage

Example: JupyterHub at EuXFEL & DESY
 sses Maxwell HPC cluster

Popular with users:

no software installation & browser of choice

works locally and remotely the same

Model for European Open Science Cloud?

European XFEL

Maxwell Jupyter Job Options Maxwell partitions (dedicated node on ALL partition (preemption) ~ Choice of GPU[®] none ~ Note: For partitions without GPUs (or choice of GPUs) the GPU selection will be set to 'none' Constraints(i) V100&INTEL Note: This will override GPU selections! Number of Nodes 🛈 🏼 4 🖨 Note:Number of nodes will be set to 1 on shared jhub partition! Job duration ① 1 hour 🗸 Note: on the shared Jupyter partition (jhub) the time limit is always 7 days! Launch modus Launch JupyterLAB **Remote Notebook** PyFAI Tutorials (Collection)

Node and GPU availability						
Partition # nodes # avail # GPUs avail # P100 avail # V100					# V100 avail	
jhub	3	3	0	0	0	
maxwell	61	46	0	0	0	
maxgpu	19	12	12	1	10	
all	327	188	0	0	0	
allgpu	88	67	67	48	10	

Spawn

[7]: demo.interactive()

^ |

Image Type:

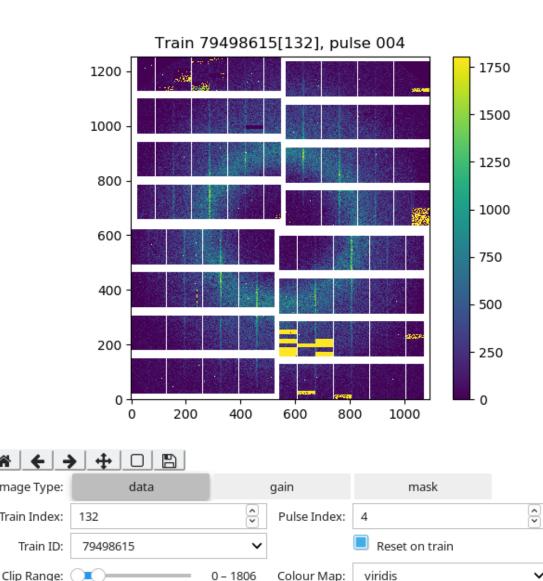
Train Index:

Train ID:

< Prev Good

Next Good >

Figure 1



Status:

Use case 6: blending GUI and script

- JupyterWidgets provide graphical control elements in notebook
 - buttons, sliders etc trigger code execution and update of plot

Useful for

data analysis of fixed type data exploration of data sets

Discussion

- less powerful than, for example, QT GUI
- popular with users due to
- being embedded in notebook
- no software installation (via JupyterHub)

V

European XFEL

Build and launch a repository

GitHub repository name or URL

GitHub repository name or URI

Git branch, tag, or commit

Git branch, tag, or commit

••• • • •

Binder project

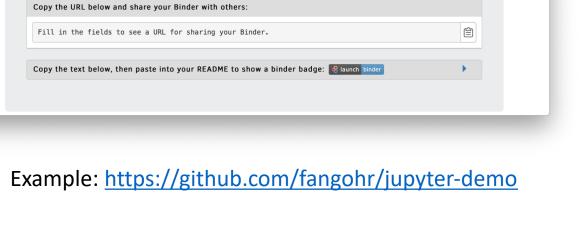
Given a (github) repository with
 Jupyter notebooks
 software requirements (Dockerfile, requirements.txt, environment.yml)

Binder service

 builds a container with the required software
 starts Jupyter notebook server in that container offering the notebooks

Binder project provides free pilot at
 <u>https://mybinder.org</u>

Institutional Binder instances are being deployed



Δ D

GitHub 👻

C

mybinder.org

8 binder

Turn a Git repo into a collection of interactive

notebooks

Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an

executable environment, making your code immediately reproducible by anyone, anywhere.

Path to a notebook file (optional)

Path to a notebook file (optional)

Data Analysis with Jupyter Notebooks (TUCPR02)

	🚍 🔒 github.com	Ċ	Ê
	Pull requests Issues Marketplace Ex	plore	🔹 +-
fangohr / jupyter-demo		⊙ Watch -	0 ★ Star 0 8 Fork 1
<>Code ① Issues 0 ↑↑ Pull req	quests o 🔲 Projects o 📰 Wiki 🗊 Securi	ity 🔟 Insights 🔅	Settings
ntroducing basics of Jupyter Noteb lanage topics	book (intended to be used in presentation / ch	at)	Edit
25 commits	2 branches 🛇 1 release	🎎 1 contributor	ររ្វុំរ BSD-3-Clause
P 25 commits		La 1 contributor	화 BSD-3-Clause Find File Clone or download -
	Create		
Branch: master - New pull request	Create		Find File Clone or download +
Branch: master - New pull request	Create		Find File Clone or download + Latest commit 85f2283 7 days ago
Branch: master New pull request fangohr Add Zenodo badge to version executed-notebooks	Create 1.0 ··· Save widget state		Find File Clone or download + Latest commit 85f2283 7 days ago 17 days ago
Branch: master New pull request fangohr Add Zenodo badge to version executed-notebooks	Create 1.0 ··· Save widget state git to ignore checkpoint files		Find File Clone or download - Latest commit 85f2283 7 days ago 17 days ago 9 months ago
Branch: master New pull request fangohr Add Zenodo badge to version executed-notebooks g.gitignore 1-basics.ipynb	1.0 Save widget state git to ignore checkpoint files simplify example 1		Find File Clone or download - Latest commit 85f2283 7 days ago 17 days ago 9 months ago 9 months ago
Branch: master New pull request fangohr Add Zenodo badge to version executed-notebooks .gitignore 1-basics.ipynb 2-widgets.ipynb	Create 1.0 ··· Save widget state git to ignore checkpoint files simplify example 1 remove output from this notebook		Find File Clone or download - Latest commit 85f2283 7 days ago 17 days ago 9 months ago 9 months ago 17 days ago

launch binder DOI 10.5281/zenodo.3463132

Most recent version of this repository is located at https://github.com/fangohr/jupyter-demo

Jupyter notebook demo repository

https://mybinder.org/v2/gh/fangohr/jupyter-demo/master

Prof Hans Fangohr, 8 October 2019

Thanks to Google Cloud and OVH for sponsoring our computers 🎉!

mybinder.org

Ċ



Starting repository: fangohr/jupyter-demo/master

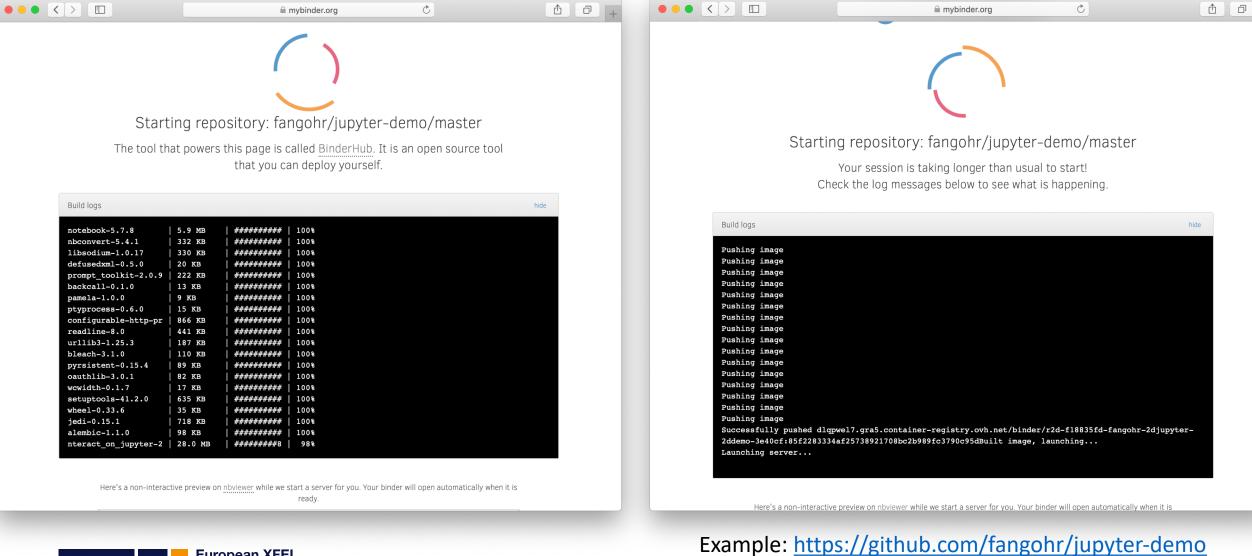
The tool that powers this page is called BinderHub. It is an open source tool that you can deploy yourself.

Build logs hide Waiting for build to start... Picked Git content provider. Cloning into '/tmp/repo2dockerjaic8b8i'... HEAD is now at 85f2283 Add Zenodo badge to version 1.0 Building conda environment for python=3.7Using PythonBuildPack builder Building conda environment for python=3.7Building conda environment for python=3.7Step 1/51 : FROM buildpack-deps:bionic Fetching base image...

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

ê o I

14



Data Analysis with Jupyter Notebooks (TUCPR02)

	🔒 hub-binder.mybinder.ovh	Ċ		Ċ
💭 Jupyter				Quit
Files Running Clusters				
Select items to perform actions on them.			Upload	New - 2
0 - 0		Name 🕹	Last Modified	File size
executed-notebooks			9 minutes ago	
1-basics.ipynb			9 minutes ago	2.9 kB
2-widgets.ipynb			9 minutes ago	3.48 kB
			9 minutes ago	1.58 kB
README.md			9 minutes ago	1.68 kB
□ □ requirements.txt			9 minutes ago	17 B

	hub-binder.mybinder.ovh ♂ 1
💭 Jupyter	2-widgets (unsaved changes)
File Edit	View Insert Cell Kernel Widgets Help Not Trusted Python 3 O
In []:	<pre>%matplotlib inline from numpy import exp, cos, linspace import pylab from ipywidgets import interact, interact_manual</pre>
	Title of investigationMathematical modelWant to understand $f(t) = exp(-\alpha t) cos(\omega t)$
	Code / Data
In []:	<pre>def f(t, alpha, omega): """Computes and returns exp(-alpha*t) * cos(omega*t)""" return exp(-alpha * t) * cos(omega * t)</pre>
	Interactive exploration We can execute the function for valuel of α and ω :
In []:	f(t=0.1, alpha=1, omega=10)
In []:	f(t=0.0, alpha=1, omega=10)

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

Use case 7: documenting software library

Use notebook as chapter in documentation
■ supported by sphinx → html, pdf as usual

Documentation easy to create:
 enter commands in notebook
 output is produced automatically

updating docs means re-running notebook

Can run regression test on documentation notebooks using NoteBook VALidate (nbval)

Can make documentation interactive using Binder ☆ European XFEL Python data tools Search docs Reading data files AGIPD, LPD & DSSC data Streaming data over ZeroMQ Checking data files AGIPD, LPD & DSSC Geometry Command line tools Data files format Performance notes Reading data with karabo_data Accessing LPD data Assembling detector data into images

Examining detector geometry

Detector geometry for AGIPD

DSSC detector geometry Working with non-detector

Comparing fast XGM data from two simultaneous

suppression ratio (with error)

 \odot

DigitalOcean

New: DigitalOcean Marketplace Deploy your

favorite dev tools with 1-Click

Apps.

Sponsored • Ads served ethically

Parallel processing with a virtual dataset

Overall comparison of

data

recordings

Release Notes

P

Assembling detector data into images

The X-ray detectors at XFEL are made up of a number of small pieces. To get an image from the data, or analyse it spatially, we need to know where each piece is located.

C Edit on GitHub

This example reassembles some commissioning data from LPD, a detector which has 4 quadrants, 16 modules, and 256 tiles. Elements (especially the quadrants) can be repositioned; talk to the detector group to ensure that you have the right geometry information for your data.

[1]: %matplotlib inline import numpy as np import matplotlib.pyplot as plt import h5py

from karabo_data import RunDirectory, stack_detector_data
from karabo_data.geometry2 import LPD_1MGeometry

[2]: run = RunDirectory('/gpfs/exfel/exp/FXE/201830/p900020/pro run.info()

of trains: 513
Duration: 0:00:51.200000
First train ID: 54861753
Last train ID: 54862265

14 detector modules (FXE_DET_LPD1M-1)
 e.g. module FXE_DET_LPD1M-1 0 : 256 x 256 pixels
 128 frames per train, 39040 total frames

0 instrument sources (excluding detectors):

0 control sources:

[3]: # Find a train with some data in

empty = np.asarray([])
for tid, train_data in run.trains():
 module_imgs = sum(d.get('image.data', empty).shape[0]
 if module_imgs:
 print(tid, module_imgs)
 break

54861797 1792

[4]: tid, train_data = run.train_from_id(54861797)
print(tid)
for dev in sorted(train_data.keys()):
 print(dev, end='\t')
 try:
 print(train_data[dev]['image.data'].shape)

Use case 7: documenting software library

Use notebook as chapter in documentation supported by sphinx \rightarrow html, pdf as usual

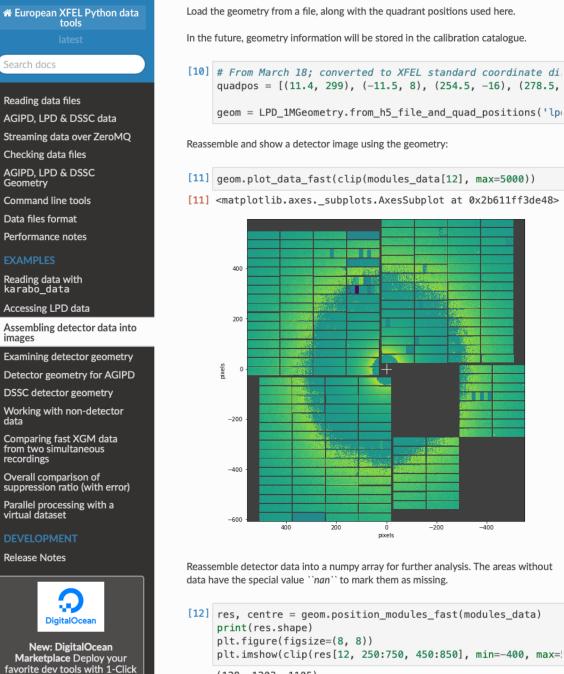
Documentation easy to create: enter commands in notebook output is produced automatically

updating docs means re-running notebook

Can run regression test on documentation notebooks using NoteBook VALidate (nbval)

Can make documentation interactive using Binder

European XFEL



P

(128, 1203, 1105)

Apps.

Sponsored · Ads served ethically

[12] <matplotlib.image.AxesImage at 0x2b60ec9f4160>

-400



Use case 8: reproducible publication

Create github repository to complement publication
 create one notebook per figure / main result
 define software environment using Binder syntax

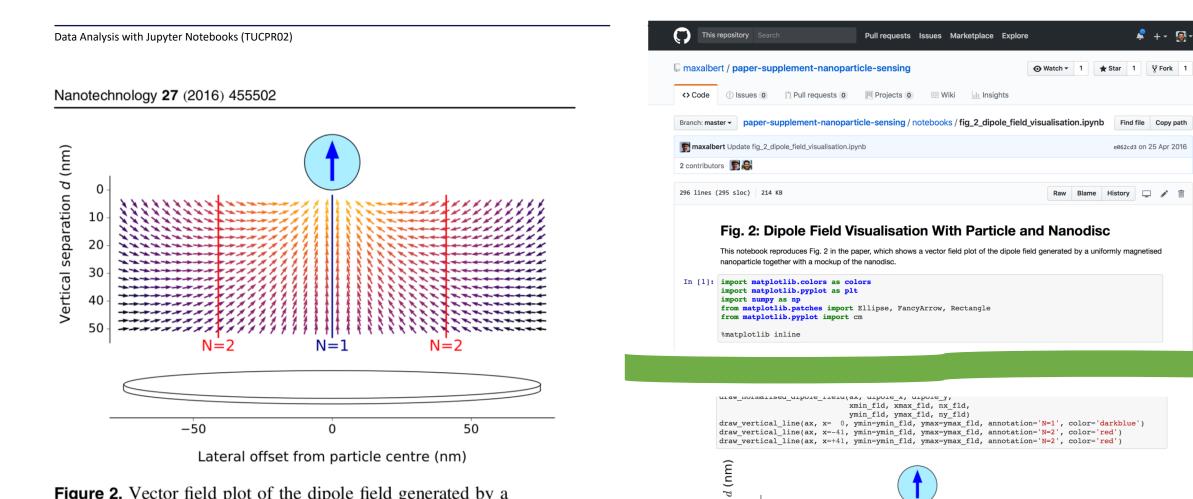
Close to reproducible publication:
 Interproducible fully specified software environment
 Interproduction fully specified data analysis
 Interproduction data analysis
 Interproduction data analysis
 Interproduction data analysis

Zenodo for long term preservation
 create Zenodo deposit for repository
 cite Zenodo DOI in publication

gohr, 8 Oc	tober 2019						10
earch or ju	ımp to		/ Pu	II requests	Issues	Marketplace	Explore
	📮 maxalbei	rt / paper-si	upplem	ent-nanop	particle	-sensing	
	<> Code	Issues 0	ឿ Pull r	equests 0	💷 Proj	ects 0	Wiki 🕕
	Branch: mast	er • paper-	supplem	ient-nanop	oarticle-	sensing / not	ebooks /
	🛐 maxalbei	rt Add labels to n	nain figure a	and inset indic	ating that	the plots show da	a •••
	style_she	ets		Add mis	ssing style	esheets.	
	🖹 explanati	on_of_the_data	_format.ip	Add not	tebook ex	plaining the da	ta format.
	fig_2_dip	ole_field_visual	isation.ipy	Update	fig_2_dip	ole_field_visual	isation.ipyr
	fig_4_free	quency_depend	lence_on_	Replace	e pickled o	data frame with	text-base
	fig_7_frec	uency_change	_vs_latera	Simplify	/ construc	ction of style cy	cle for Fig.
	fig_8_free	quency_change	_vs_partic	I Add lab	els to ma	in figure and in	set indicati
	🖹 fig_9a_de	ependence_of_f	requency	Replace	e pickled o	data frame with	n text-base
	🖹 fig_9b_de	ependence_of_f	requency	Replace	e pickled o	data frame with	n text-base
	<pre> fig_9c_cc </pre>	mparison_of_fr	equency_	Replace	e pickled o	data frame with	n text-base
	style_hel	pers.py		Simplify	/ construc	ction of style cy	cle for Fig.

Example:

Prof Hans Fangohr. 8 October 2019



paration

se 30

Vertical

10

20

40

50

N=2

-50

Figure 2. Vector field plot of the dipole field generated by a uniformly +z-magnetised MNP. The vectors are scaled to uniform length with their colour indicating the field strength (orange is high and violet/black is low). The vertical lines correspond to the xvalues where modes 1 and 2 have maxima in their spin precession amplitude (see figures 3(a)–(b)). A schematic of the nanodisc is shown at the bottom.

0 Lateral offset from particle centre (nm)

N=1

N=2

50

Frequency-based nanoparticle sensing over large field ranges using the ferromagnetic resonances of a magnetic nanodisc: supplementary material

20

DOI 10.5281/zenodo.60605 preprint arxiv:1604.07277 launch binder license MIT

This repository accompanies the paper "*Frequency-based nanoparticle sensing over large field ranges using the ferromagnetic resonances of a magnetic nanodisc*", published in *Nanotechnology*, Volume 27, Number 45. It provides the data underlying the figures in the paper, as well as Jupyter notebooks to reproduce those figures.

The latest version of this repository can be found at https://github.com/maxalbert/paper-supplement-nanoparticle-sensing

Authors: Maximilian Albert, Marijan Beg, Dmitri Chernyshenko, Marc-Antonio Bisotti, Rebecca L. Carey, Hans Fangohr and Peter Metaxas.

Contents

The directory notebooks/ contains Jupyter notebooks for the relevant figures in the paper. On Github you can view them directly in the browser:

- Fig. 2: Dipole field visualisation
- Fig. 4: Frequency dependence on external field strength
- Fig. 7: Frequency change vs. lateral particle position

Summary

Jupyter Notebook and ecosystem increasingly popular in academia and industry

Welcome by users, and driven by users

- Combines code, results and interpretation
- Strong technology candidate for European Open Science Cloud and remote analysis portals?

A quickly moving field

- significant potential to enable better science
- many open questions
- This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654220 (PaNOSC).

Contact: Hans Fangohr, hans.fangohr@xfel.eu, http://fangohr.github.io, @ProfCompMod