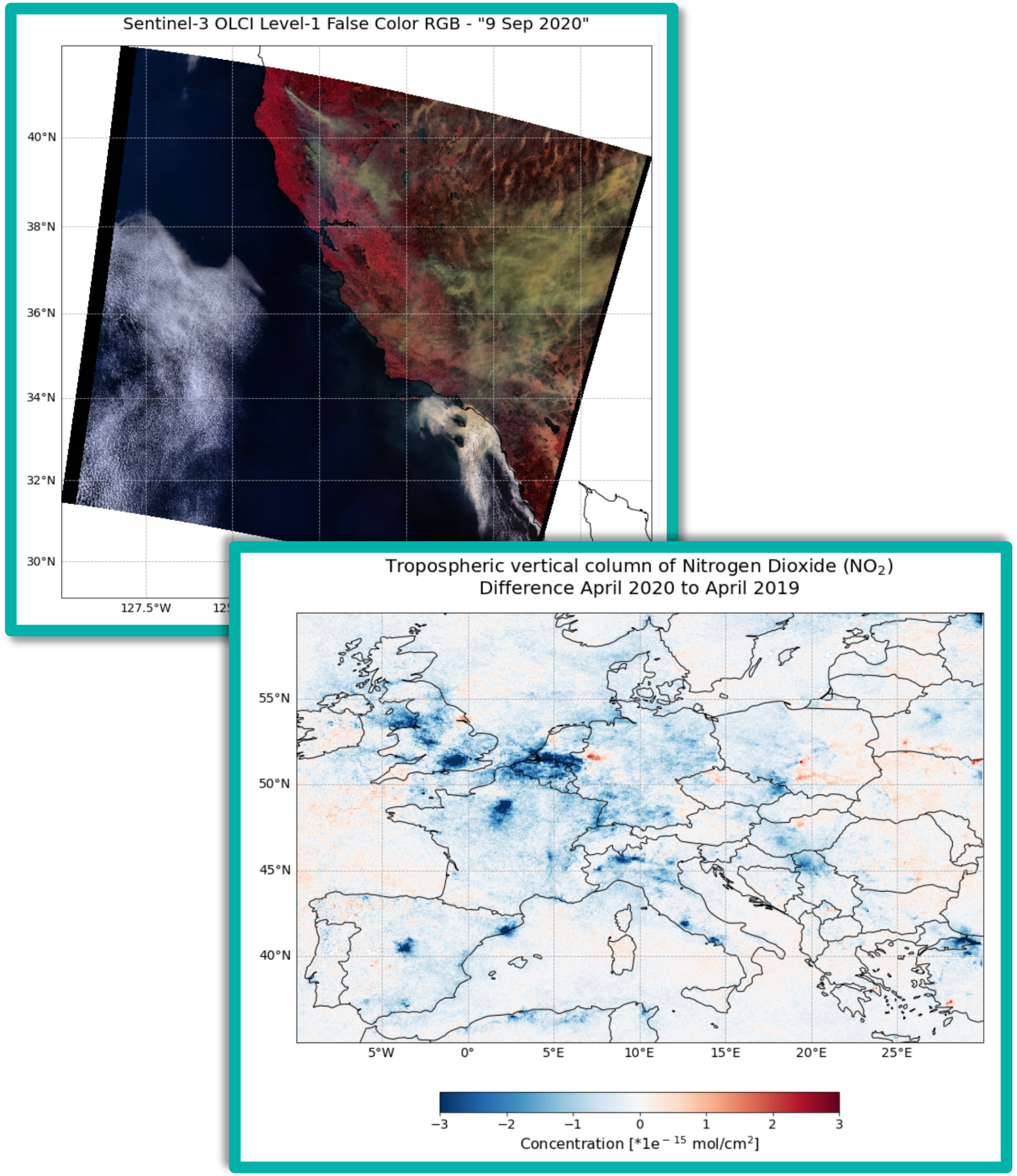
EUMETSAT Short Course #39 | 07 June 2023

How to develop impactful and educational notebooks?

Dr. Julia Wagemann | *MEEO s.r.l. for EUMETSAT*

Sabrina H. Szeto | *MEEO s.r.l. for EUMETSAT*
Simone Mantovani | *MEEO s.r.l. for EUMETSAT*
Dr. Federico Fierli | *EUMETSAT*

About


Sentinel-3 OLCI Level-1 False Color RGB - "9 Sep 2020"
40°N
38°N
36°N
34°N
32°N
30°N
127.5°W
125
Tropospheric vertical column of Nitrogen Dioxide (NO_2)
Difference April 2020 to April 2019
55°N
50°N
45°N
40°N
5°W
0°
5°E
10°E
15°E
20°E
25°E
−3
−2
−1
0
1
2
3
Concentration [*1e^{-15} mol/cm^2]

- Background in **Earth Observation / Remote Sensing / Climate Sciences**

- Use **Jupyter notebooks since 2014**

- Since 2019, I have developed **120+ educational notebooks on open Earth Observation data handling, access, visualisation and Machine Learning**

- Trainings range from **short webinars (1 to 1.5 hours) up to a weeklong intensive training schools**, but also Massive Open Online Courses

Particularities of Earth Observation data education outside academia

Limited time

With the instructor and to set up the programming environment

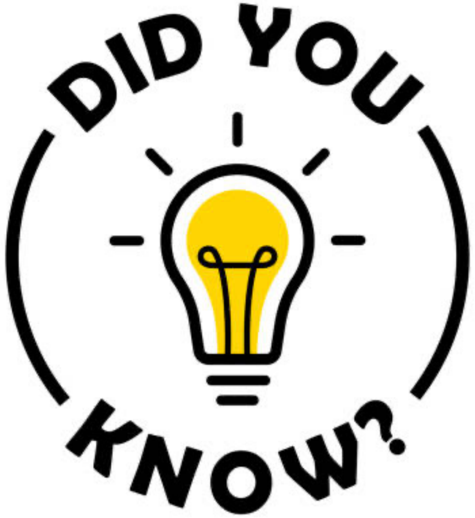
Diverse training audience

With regards to (EO) data, programming language and experience and thematic applications

Flexibility of how it is taught

(online vs. in-site, instructor-led vs. self-paced)

Jupyter notebooks are everywhere



There are **more than 10* million Jupyter notebooks available on Github**

Within 10 years, Jupyter became the **de-facto standard** for data exploration, analysis and training

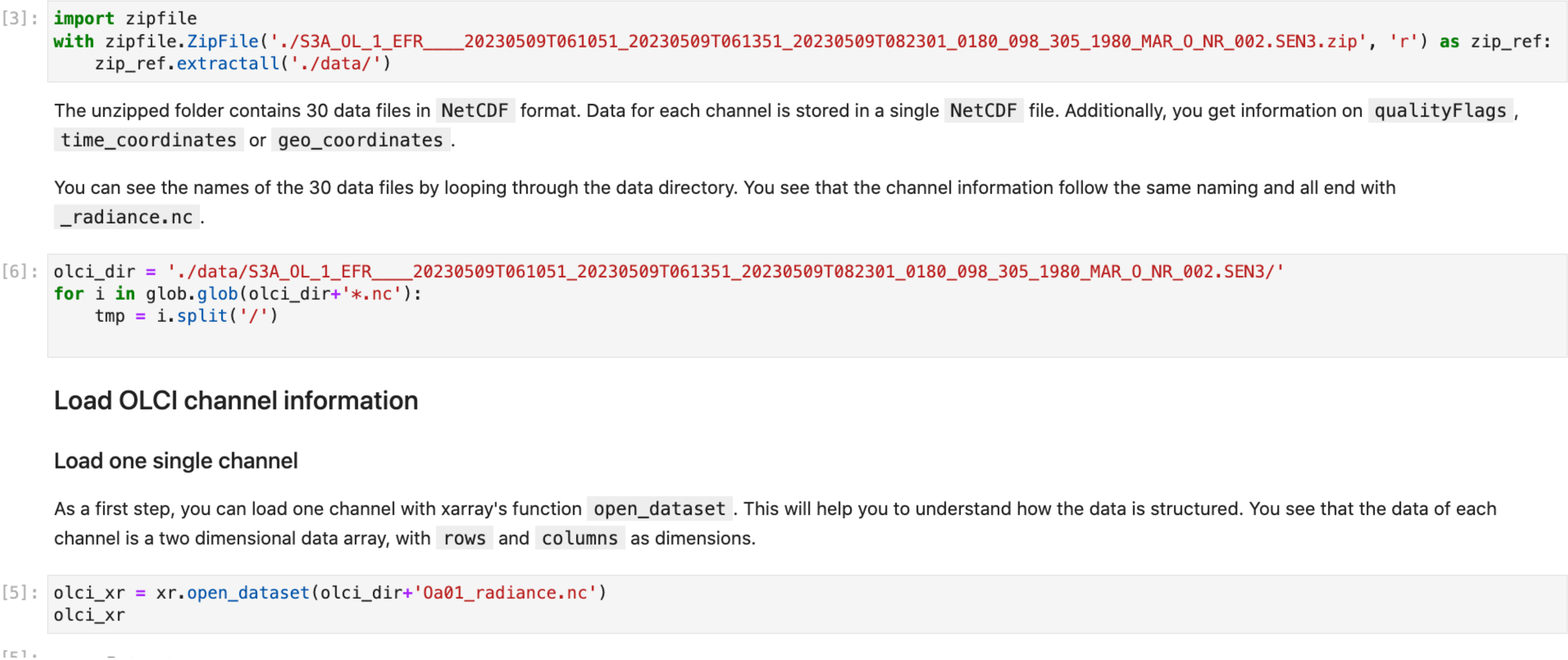
Mostly used for research experimentation, development of machine-learning pipelines **and education !!!**

* Perkel (2018): Why Jupyter is data scientists' computation notebook of choice. Nature.
* Perkel (2021): Ten computer codes that transformed science. Nature.

What’s the reality like? Some pitfalls and challenges, backed by research!

1

Out of order execution of code cells fosters poor coding practices



```
[3]: import zipfile
     with zipfile.ZipFile('./S3A_OL_1_EFR____20230509T061051_20230509T061351_20230509T082301_0180_098_305_1980_MAR_O_NR_002.SEN3.zip', 'r') as zip_ref:
         zip_ref.extractall('./data/')
```

The unzipped folder contains 30 data files in `NetCDF` format. Data for each channel is stored in a single `NetCDF` file. Additionally, you get information on `qualityFlags`, `time_coordinates` or `geo_coordinates`.

You can see the names of the 30 data files by looping through the data directory. You see that the channel information follow the same naming and all end with `_radiance.nc`.

```
[6]: olci_dir = './data/S3A_OL_1_EFR____20230509T061051_20230509T061351_20230509T082301_0180_098_305_1980_MAR_O_NR_002.SEN3/'
     for i in glob.glob(olci_dir+'*.nc'):
         tmp = i.split('/')
```

Load OLCI channel information

Load one single channel

As a first step, you can load one channel with xarray's function `open_dataset`. This will help you to understand how the data is structured. You see that the data of each channel is a two dimensional data array, with `rows` and `columns` as dimensions.

```
[5]: olci_xr = xr.open_dataset(olci_dir+'Oa01_radiance.nc')
     olci_xr
```

Grus, J. (2018): I don’t like Notebooks. JupyterCon 2018

Slido poll

2

Challenges to make notebooks **reproducible and reusable**

Slido.com

#EUMSC39

Pimentel et al. (2019): A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. IEEE

What’s the reality like? Some pitfalls and challenges, backed by research!

2 Challenges to make notebooks **reproducible and reusable**

- Only **1 out of 4** notebooks on Github could be executed

- Only **4 %** produced the same result

Pimentel et al. (2019): A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. IEEE

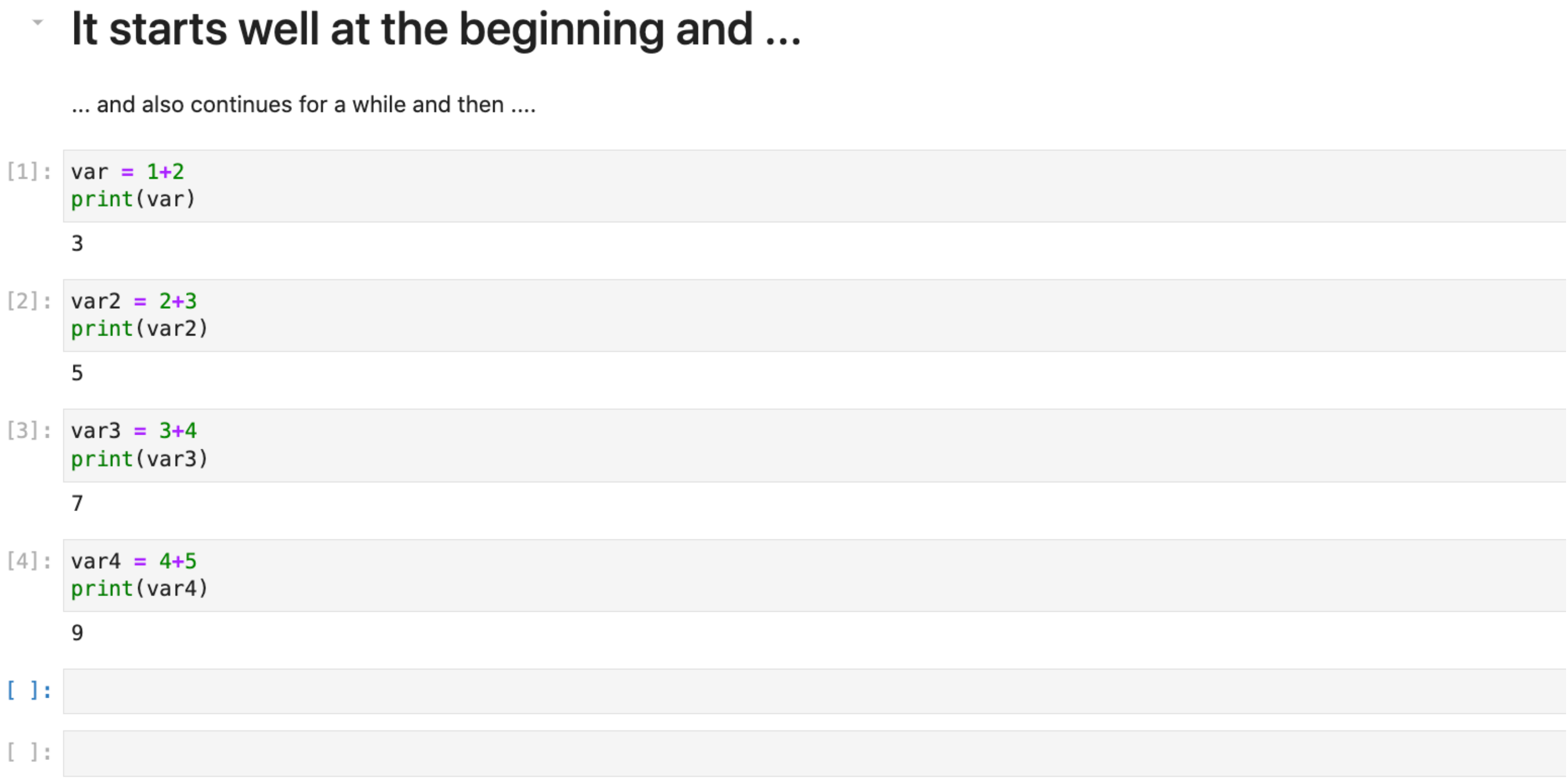
What’s the reality like? Some pitfalls and challenges, backed by research!

3 Annotations are not evenly distributed within a notebook

Most text at the beginning and hardly any text at the end

Resembles more a collection of lose scripts than a narrative

By far more code cells than descriptive text



It starts well at the beginning and ...

... and also continues for a while and then ....

```
[1]: var = 1+2
     print(var)

3

[2]: var2 = 2+3
     print(var2)

5

[3]: var3 = 3+4
     print(var3)

7

[4]: var4 = 4+5
     print(var4)

9

[ ]:

[ ]:
```

Rule et al. (2018): Exploration and Explanation in Computational Notebooks. In Proceedings of the 2018 CHI 1. Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018

Pimentel, J.F. et al. (2021): Understanding and Improving the Quality and Reproducibility of Jupyter Notebooks. Empir. Softw. Eng.

The need for best practices…

… how to **write and share Jupyter notebooks**

Rule, A. et al. (2019): Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks. PLoS Comput. Biol.

… how to **make notebooks reproducible**

Pimentel, J.F. et al. (2021): Understanding and Improving the Quality and Reproducibility of Jupyter Notebooks. Empir. Softw. Eng.

… how to **foster collaboration**

Quaranta, L. et al. (2022): Eliciting Best Practices for Collaboration with Computational Notebooks. Proc. ACM Hum. Comput. Interact.

… how to **use notebooks in academic classrooms**

Johnson, J.W. (2020): Benefits and Pitfalls of Jupyter Notebooks in the Classroom. In Proceedings of the 21st Annual Conference on Information Technology Education

Basis of 5 guiding principles for EO data education

Principles are **founded in recognized best practices** from the fields of scientific computing and Jupyter notebook research

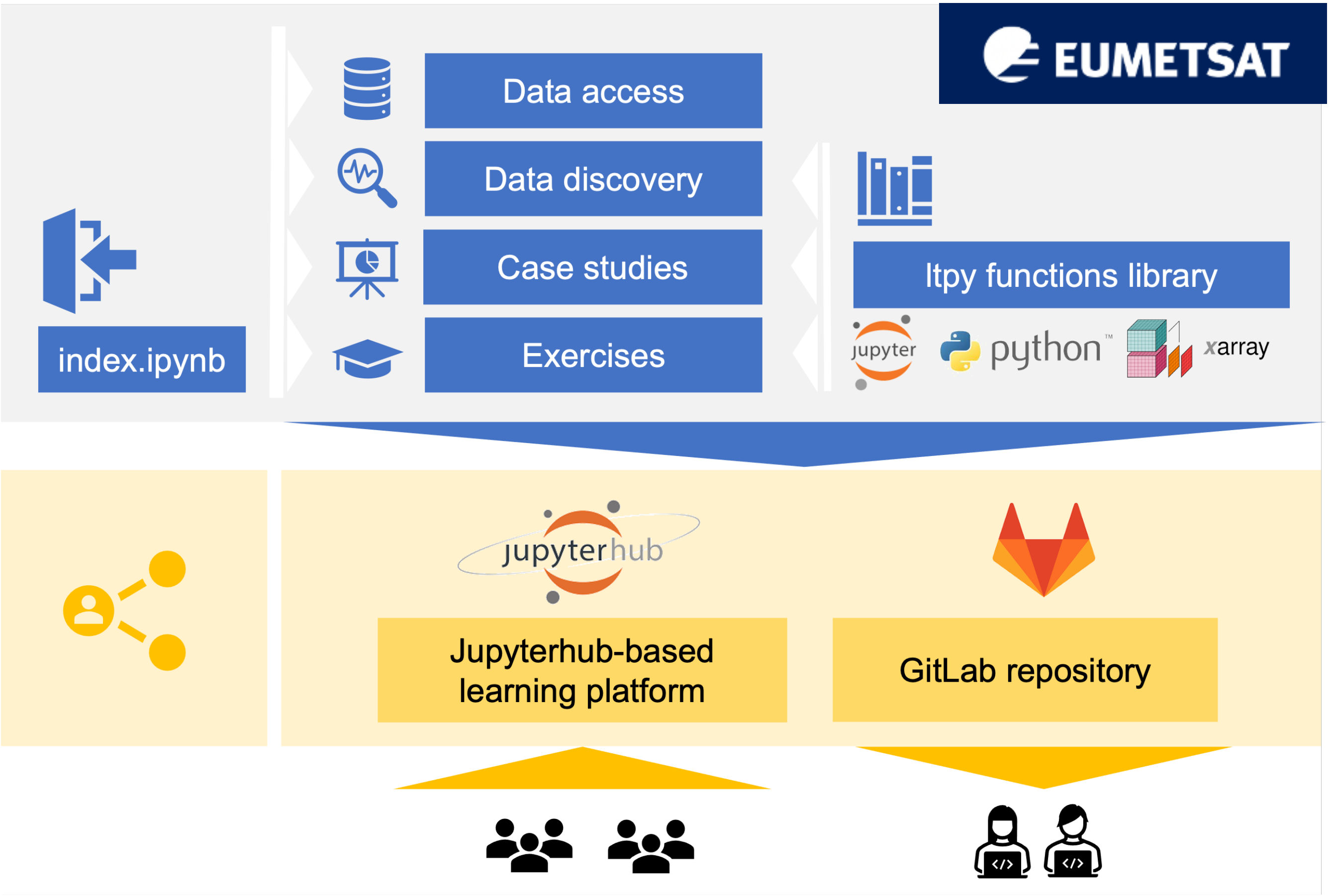
Were selected based on their **applicability for training and capacity-building**

Learning Tool for Python (LTPy) on Atmospheric composition

Data from **6 different satellites** and **5 different model-based products**

Over 70 notebooks related to (i) data access, (ii) data discovery, (iii) case studies and (iv) exercises

A **collection of 14 reusable functions** for effective visualization and data handling



Over 2000 learners trained in 41 training events

https://ltpy.adamplatform.eu
https://gitlab.eumetsat.int/eumetlab/atmosphere/atmosphere

Guiding principles to make educational and reusable notebooks

1 Leverage the ‘Literate Programming Paradigm’

Use of instructional design elements 2

3 Follow best practices for scientific computing

Take advantage of the full Jupyter Ecosystem 4

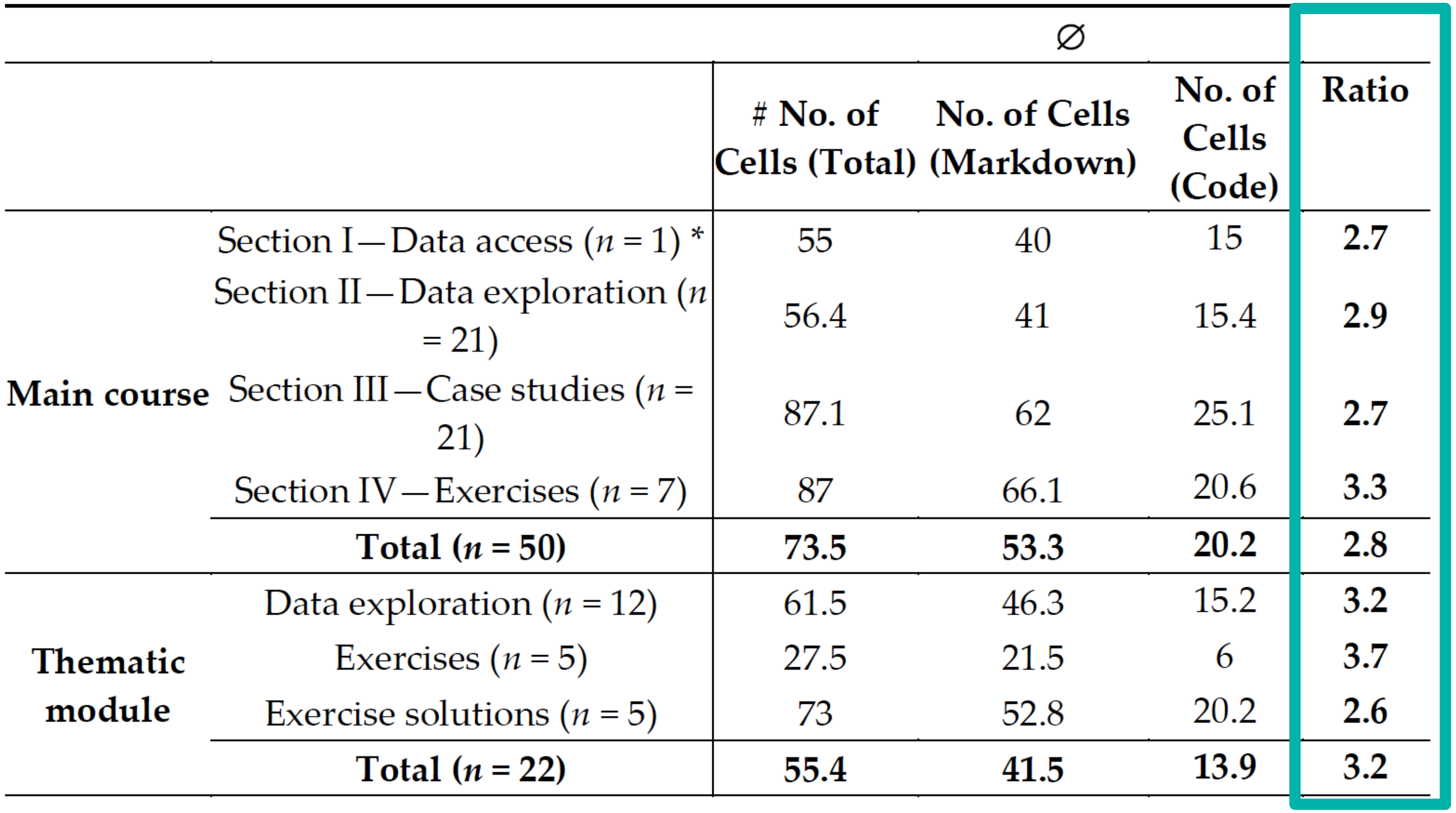
5 Aim for Reproducibility

Wagemann, J., Fierli, F., Mantovani, S., Siemen, S., Seeger, B. and J. Bendix (2022): Five Guiding Principles to Make Jupyter Notebooks Fit For Earth Observation Data Education. *Remote Sensing 2022, 14(14), 3359.*

1. Leverage the literate programming paradigm

Rule et al. (2018) → analysed > 1 Mio. Notebooks → **1 out of 4 had no text at all**

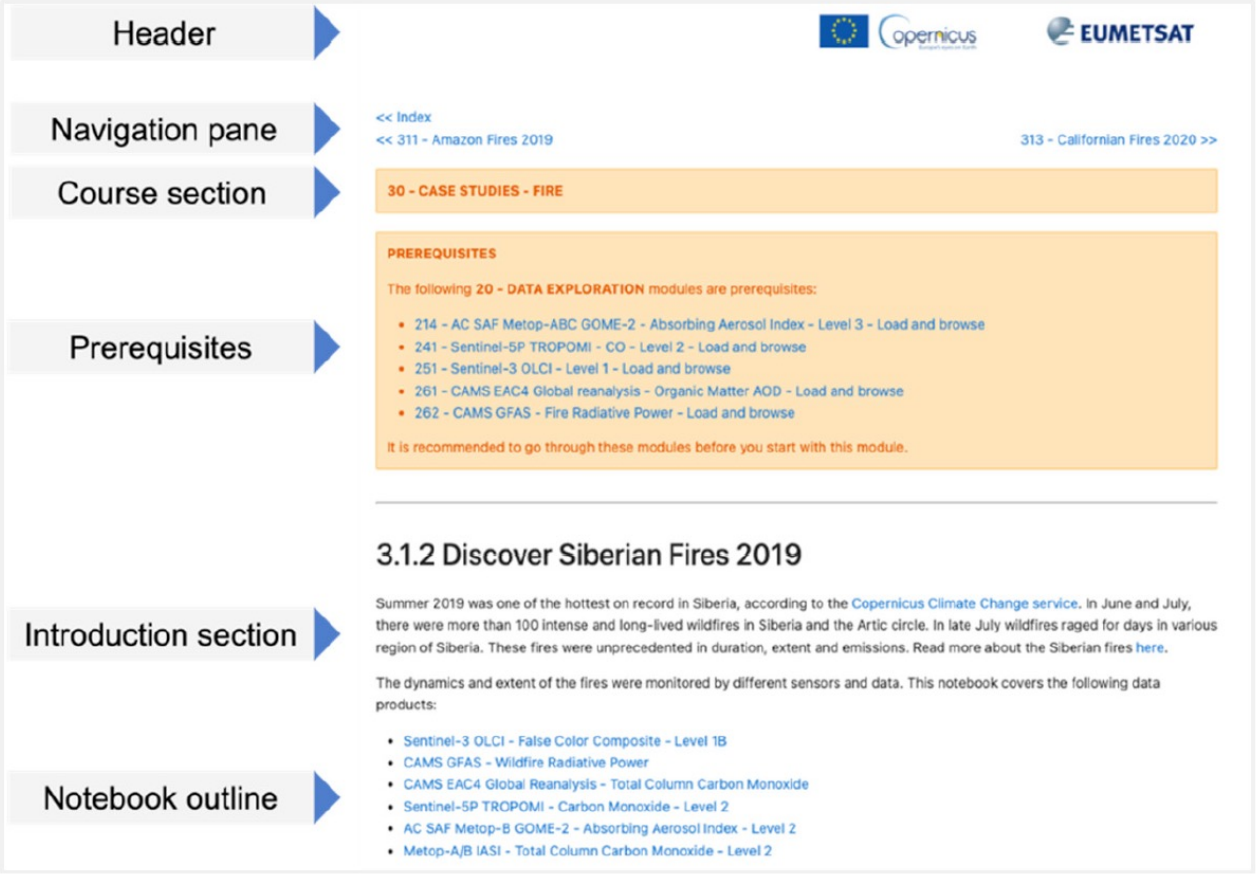
Quaranta et al. (2022) → analysed > 1000 notebooks → **Median text/code ratio of 0.4**



		∅			
		# No. of Cells (Total)	**No. of Cells (Markdown)**	**No. of Cells (Code)**	**Ratio**
Main course	Section I—Data access (*n* = 1) *	55	40	15	**2.7**
	Section II—Data exploration (*n* = 21)	56.4	41	15.4	**2.9**
	Section III—Case studies (*n* = 21)	87.1	62	25.1	**2.7**
	Section IV—Exercises (*n* = 7)	87	66.1	20.6	**3.3**
	Total (*n* = 50)	**73.5**	**53.3**	**20.2**	**2.8**
Thematic module	Data exploration (*n* = 12)	61.5	46.3	15.2	**3.2**
	Exercises (*n* = 5)	27.5	21.5	6	**3.7**
	Exercise solutions (*n* = 5)	73	52.8	20.2	**2.6**
	Total (*n* = 22)	**55.4**	**41.5**	**13.9**	**3.2**

3 times more text cells than code cells

2. Use of instructional design elements



<< Index
<< 311 - Amazon Fires 2019 313 - Californian Fires 2020 >>

30 - CASE STUDIES - FIRE

PREREQUISITES

The following **20 - DATA EXPLORATION** modules are prerequisites:

- 214 - AC SAF Metop-ABC GOME-2 - Absorbing Aerosol Index - Level 3 - Load and browse
- 241 - Sentinel-5P TROPOMI - CO - Level 2 - Load and browse
- 251 - Sentinel-3 OLCI - Level 1 - Load and browse
- 261 - CAMS EAC4 Global reanalysis - Organic Matter AOD - Load and browse
- 262 - CAMS GFAS - Fire Radiative Power - Load and browse

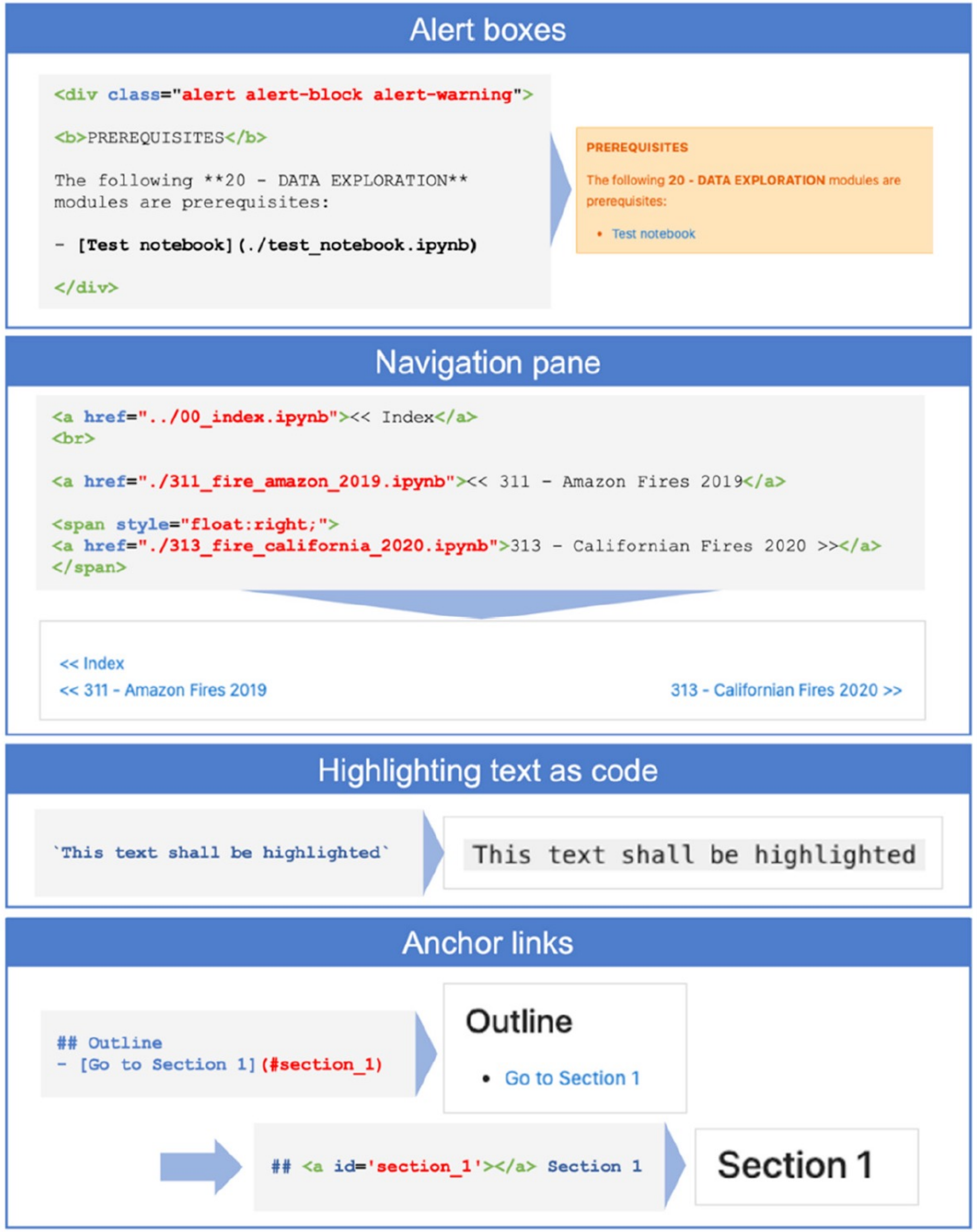
It is recommended to go through these modules before you start with this module.

3.1.2 Discover Siberian Fires 2019

Summer 2019 was one of the hottest on record in Siberia, according to the Copernicus Climate Change service. In June and July, there were more than 100 intense and long-lived wildfires in Siberia and the Artic circle. In late July wildfires raged for days in various region of Siberia. These fires were unprecedented in duration, extent and emissions. Read more about the Siberian fires here.

The dynamics and extent of the fires were monitored by different sensors and data. This notebook covers the following data products:

- Sentinel-3 OLCI - False Color Composite - Level 1B
- CAMS GFAS - Wildfire Radiative Power
- CAMS EAC4 Global Reanalysis - Total Column Carbon Monoxide
- Sentinel-5P TROPOMI - Carbon Monoxide - Level 2
- AC SAF Metop-B GOME-2 - Absorbing Aerosol Index - Level 2
- Metop-A/B IASI - Total Column Carbon Monoxide - Level 2



Alert boxes

```
<div class="alert alert-block alert-warning">

<b>PREREQUISITES</b>

The following **20 - DATA EXPLORATION**
modules are prerequisites:

- [Test notebook](./test_notebook.ipynb)

</div>
```

PREREQUISITES

The following **20 - DATA EXPLORATION** modules are prerequisites:

- Test notebook

Navigation pane

```
<a href="../00_index.ipynb"><< Index</a>
<br>

<a href="./311_fire_amazon_2019.ipynb"><< 311 - Amazon Fires 2019</a>

<span style="float:right;">
<a href="./313_fire_california_2020.ipynb">313 - Californian Fires 2020 >></a>
</span>
```

<< Index
<< 311 - Amazon Fires 2019 313 - Californian Fires 2020 >>

Highlighting text as code

```
`This text shall be highlighted`
```

`This text shall be highlighted`

Anchor links

```
## Outline
- [Go to Section 1](#section_1)
```

Outline

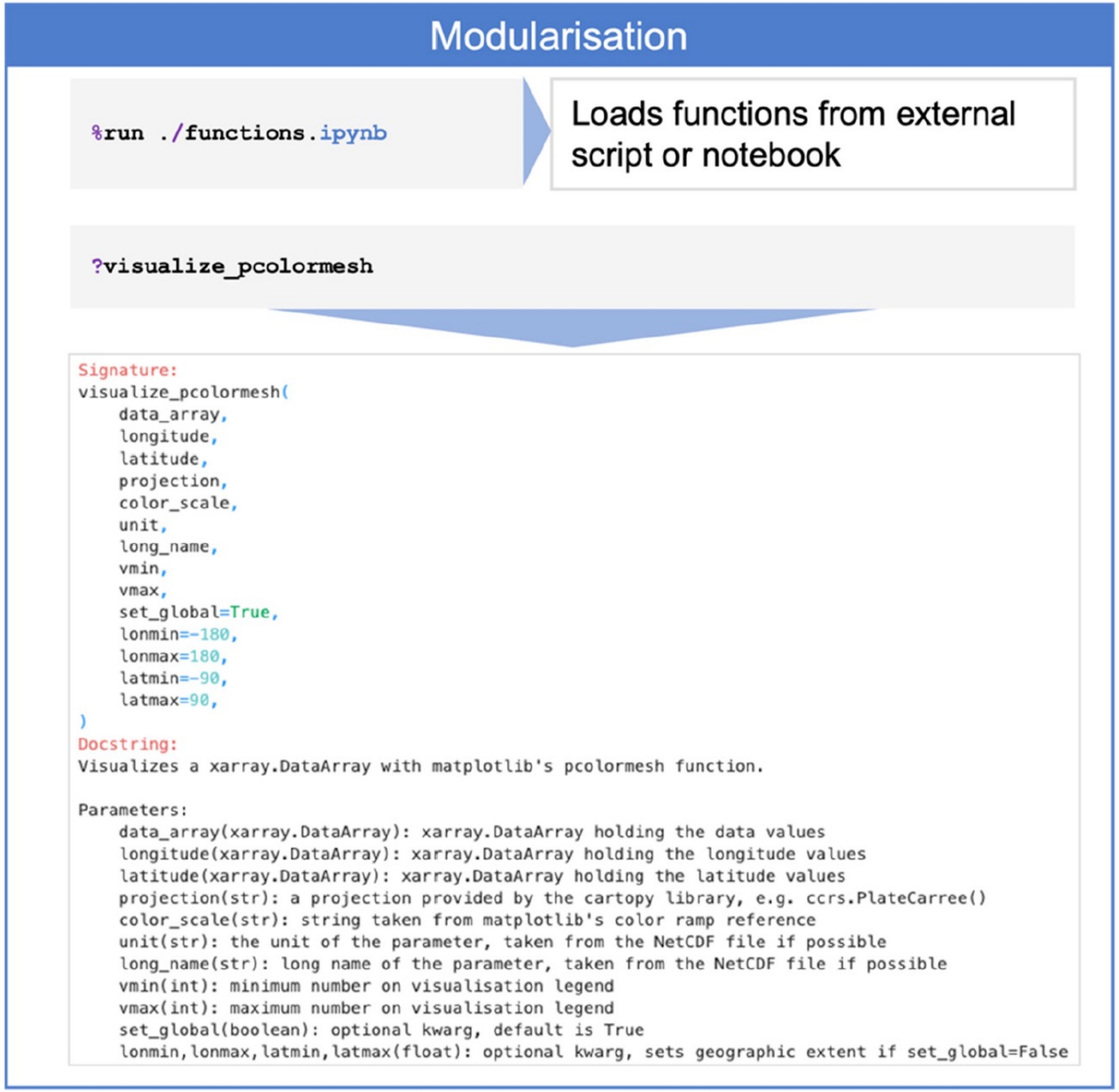
- Go to Section 1

```
## <a id='section_1'></a> Section 1
```

Section 1

3. Follow best practices for scientific computing

- Modularisation of code
- Import libraries at the beginning of a workflow
- Making code style and formatting consistent
- Using meaningful names for variables



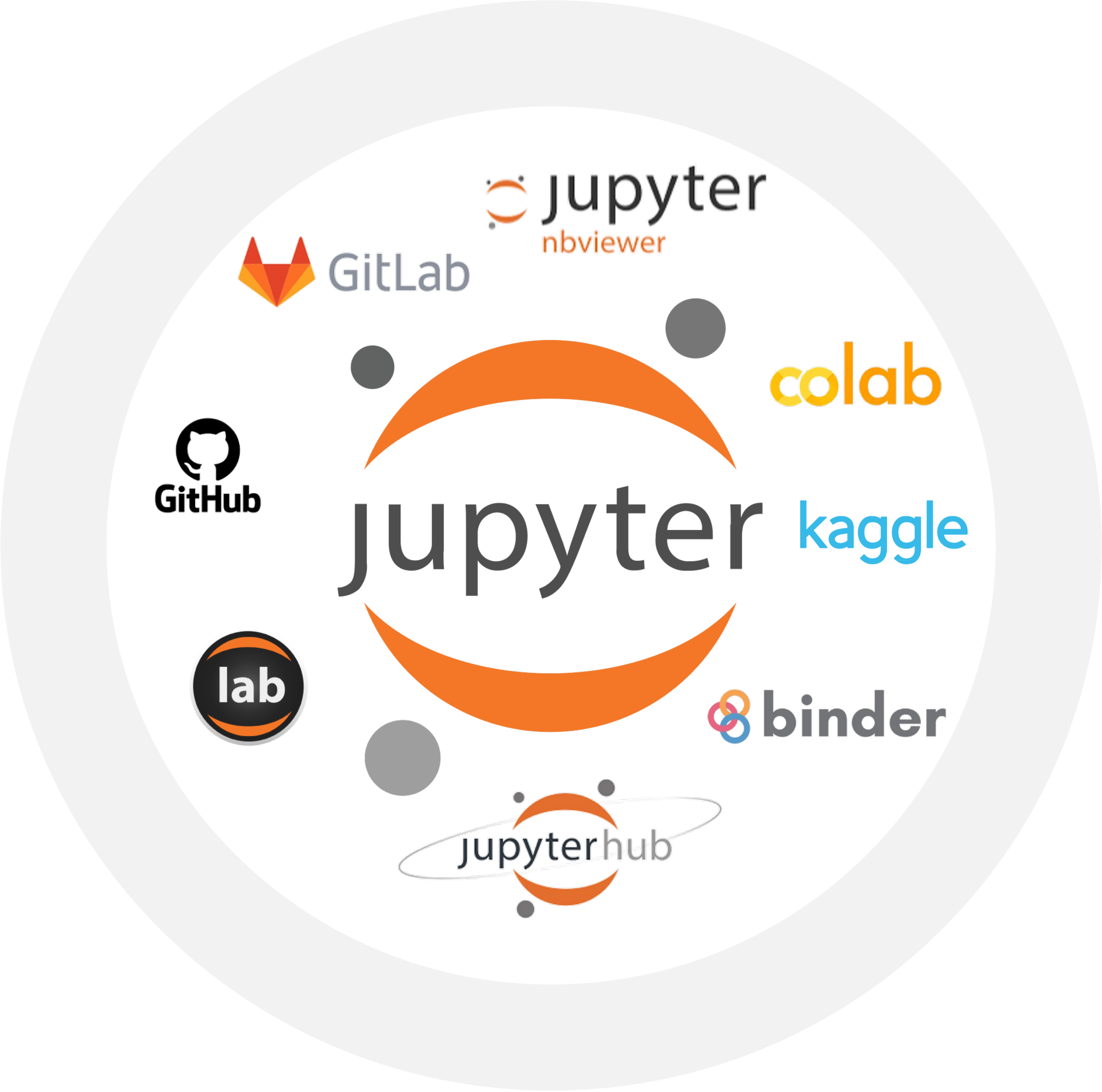
```
%run ./functions.ipynb
```

```
?visualize_pcolormesh
```

```
Signature:
visualize_pcolormesh(
    data_array,
    longitude,
    latitude,
    projection,
    color_scale,
    unit,
    long_name,
    vmin,
    vmax,
    set_global=True,
    lonmin=-180,
    lonmax=180,
    latmin=-90,
    latmax=90,
)
Docstring:
Visualizes a xarray.DataArray with matplotlib's pcolormesh function.

Parameters:
    data_array(xarray.DataArray): xarray.DataArray holding the data values
    longitude(xarray.DataArray): xarray.DataArray holding the longitude values
    latitude(xarray.DataArray): xarray.DataArray holding the latitude values
    projection(str): a projection provided by the cartopy library, e.g. ccrs.PlateCarree()
    color_scale(str): string taken from matplotlib's color ramp reference
    unit(str): the unit of the parameter, taken from the NetCDF file if possible
    long_name(str): long name of the parameter, taken from the NetCDF file if possible
    vmin(int): minimum number on visualisation legend
    vmax(int): maximum number on visualisation legend
    set_global(boolean): optional kwarg, default is True
    lonmin,lonmax,latmin,latmax(float): optional kwarg, sets geographic extent if set_global=False
```

4. Take advantage of the full Project Jupyter ecosystem



- Do not rely solely on Github rendering
- Make notebooks available as static and executable content

5. Aim for reproducibility

- Greatly increases the ‘usability’ of notebooks
- In particular relevant when notebooks are used in an educational context
- Includes data, instructions for environment settings, package versions, dependencies, execution from top to bottom, remove empty code cells

Reproducibility is *‘going the extra mile’*

Thank you!

@JuliaWagemann

Wagemann, J., Fierli, F., Mantovani, S., Siemen, S., Seeger, B. and J. Bendix (2022): Five Guiding Principles to Make Jupyter Notebooks Fit For Earth Observation Data Education. *Remote Sensing 2022, 14(14), 3359.*