



DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS  
AND ARTIFICIAL INTELLIGENCE

# Research Data Management

## Robert Haase

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

# Recap quiz

- We write good documentation to enabling others to do an experiment. This is good for ...

Repeatability



Reproducibility



Replicability



Reliability



# Recap quiz

- “Resolution” in microscopy imaging describes

Camera  
pixel size



Screen  
pixel size



Size of  
differentiable  
objects



Objective  
magnification



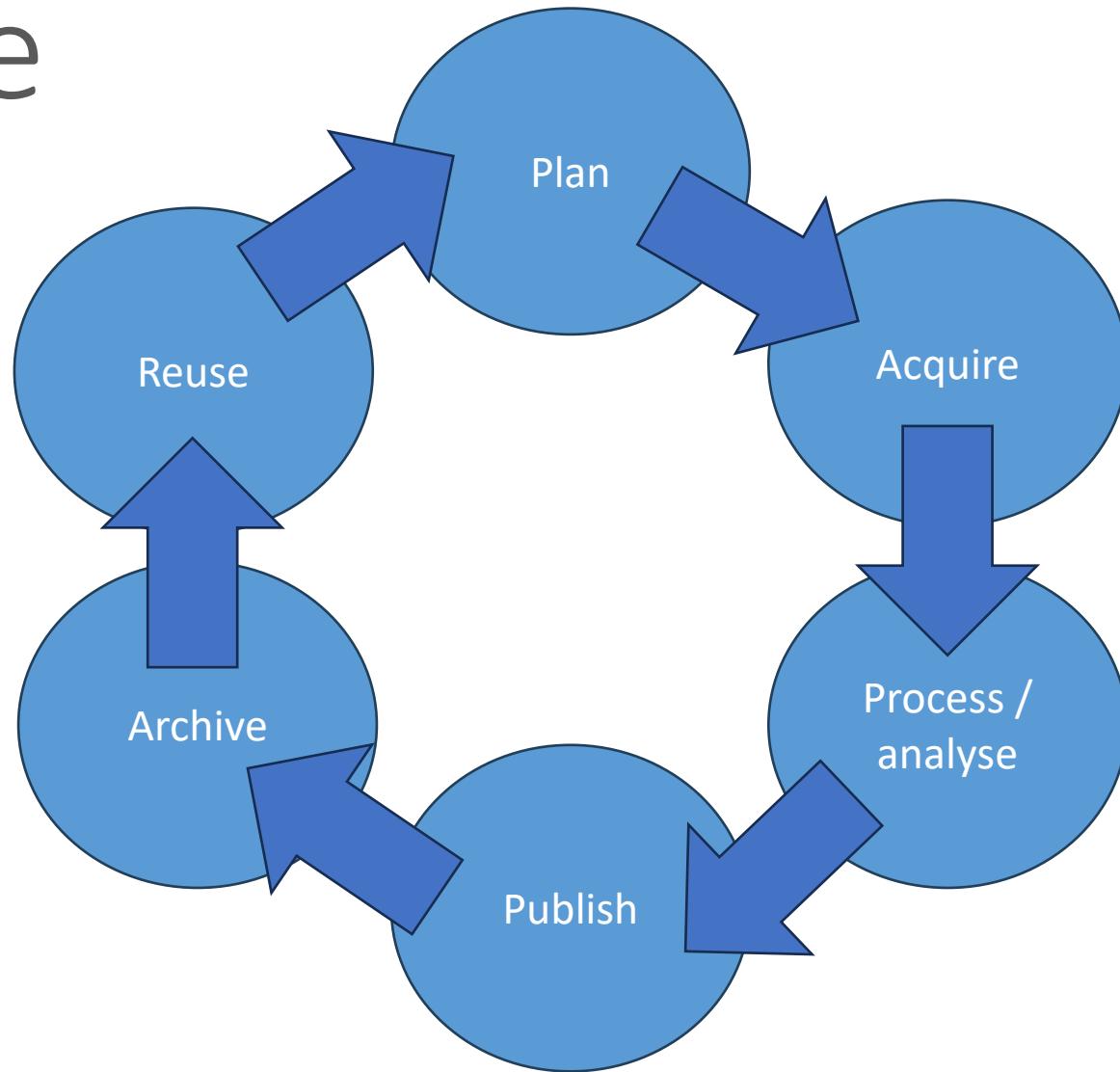
# Research Data Management (RDM)

- All activities, processes, terms, persons which have relationships with data
  - Processing
  - Storage
  - Organisation
  - Publication
  - ...
- In routine: working with data



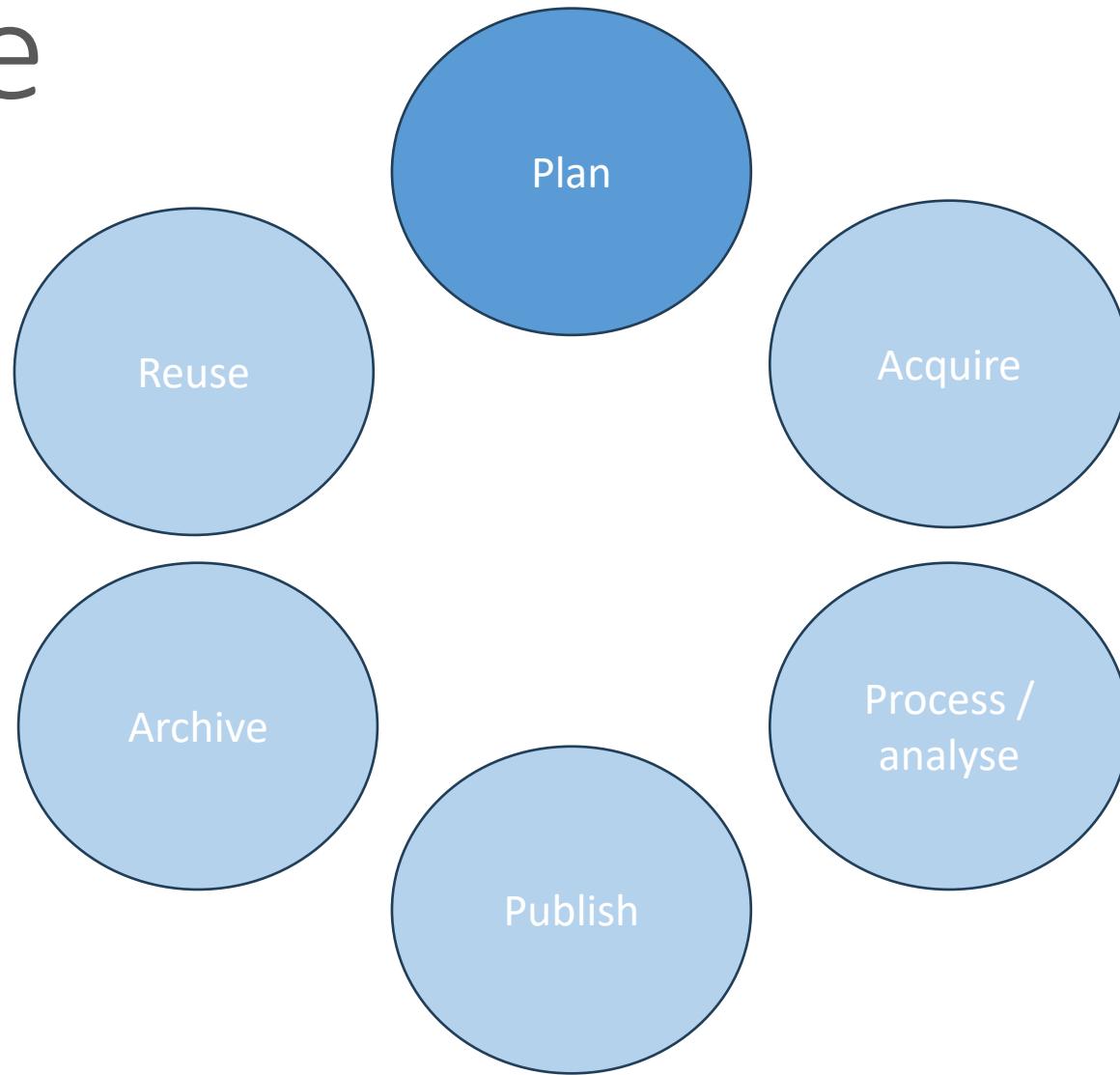
# RDM Life Cycle

- Processes are ideally cyclic



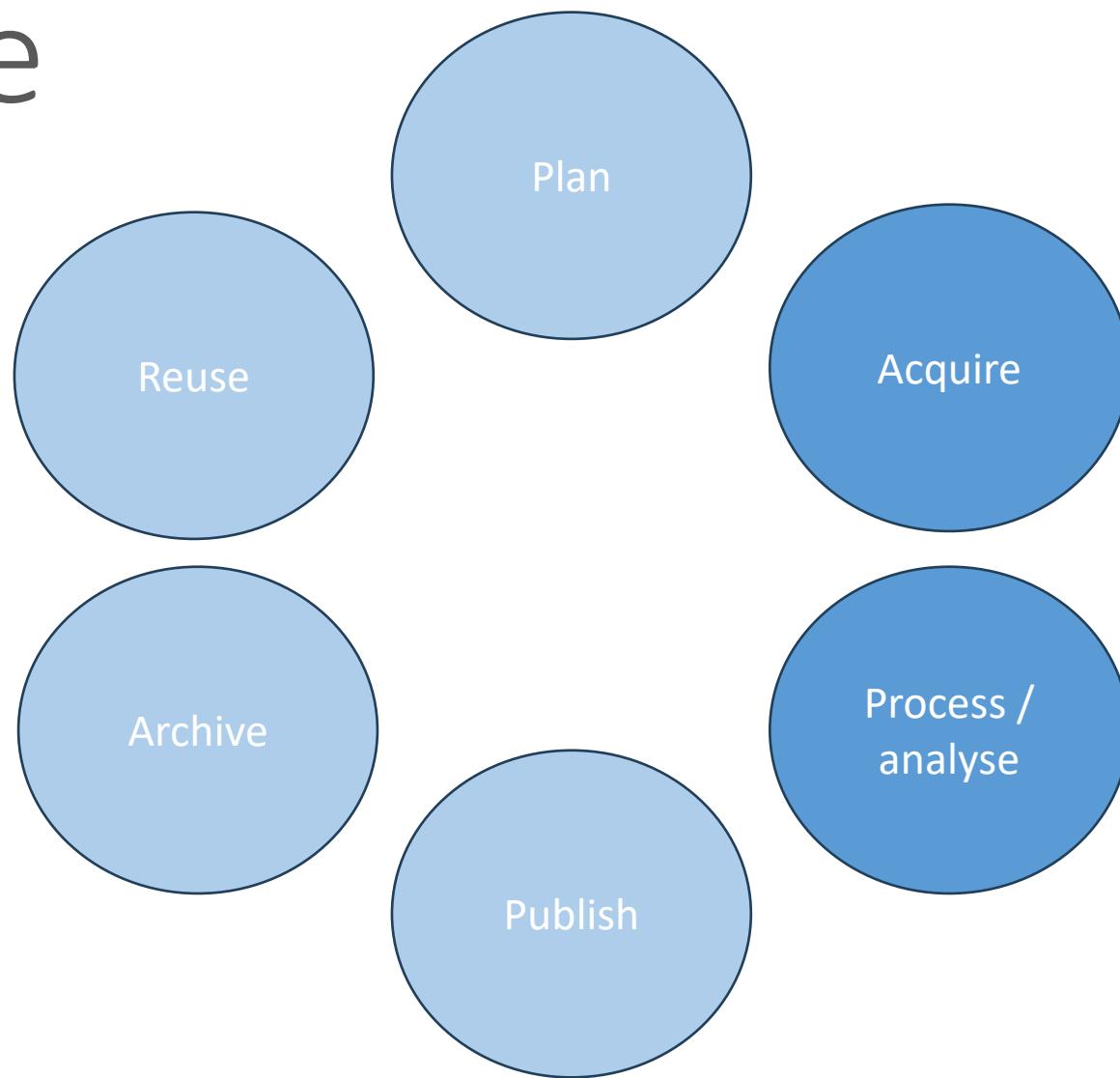
# RDM Life Cycle

- Cost
- Benefit
- Quality
- Strategic decisions



# RDM Life Cycle

- Types of data
- Terms and conditions
  - Usage rights
  - Copyright
- IT infrastructure
- Backup



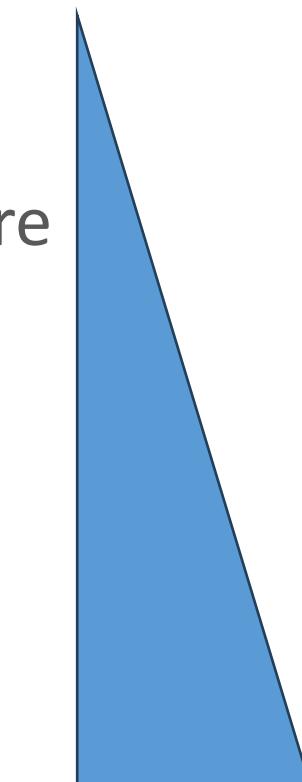
# Types of data

- Structured data
  - Tables, databases
- Unstructured data
  - Texte, emails, videos, pictures
- Semi-structured data
  - Fragebögen
  - Scientific images



# Types of data

- Openly accessible data
  - „open data“
  - „open source“ software
- Business data
- Research data
  - Hot / cold
- Personal data
- Secret data

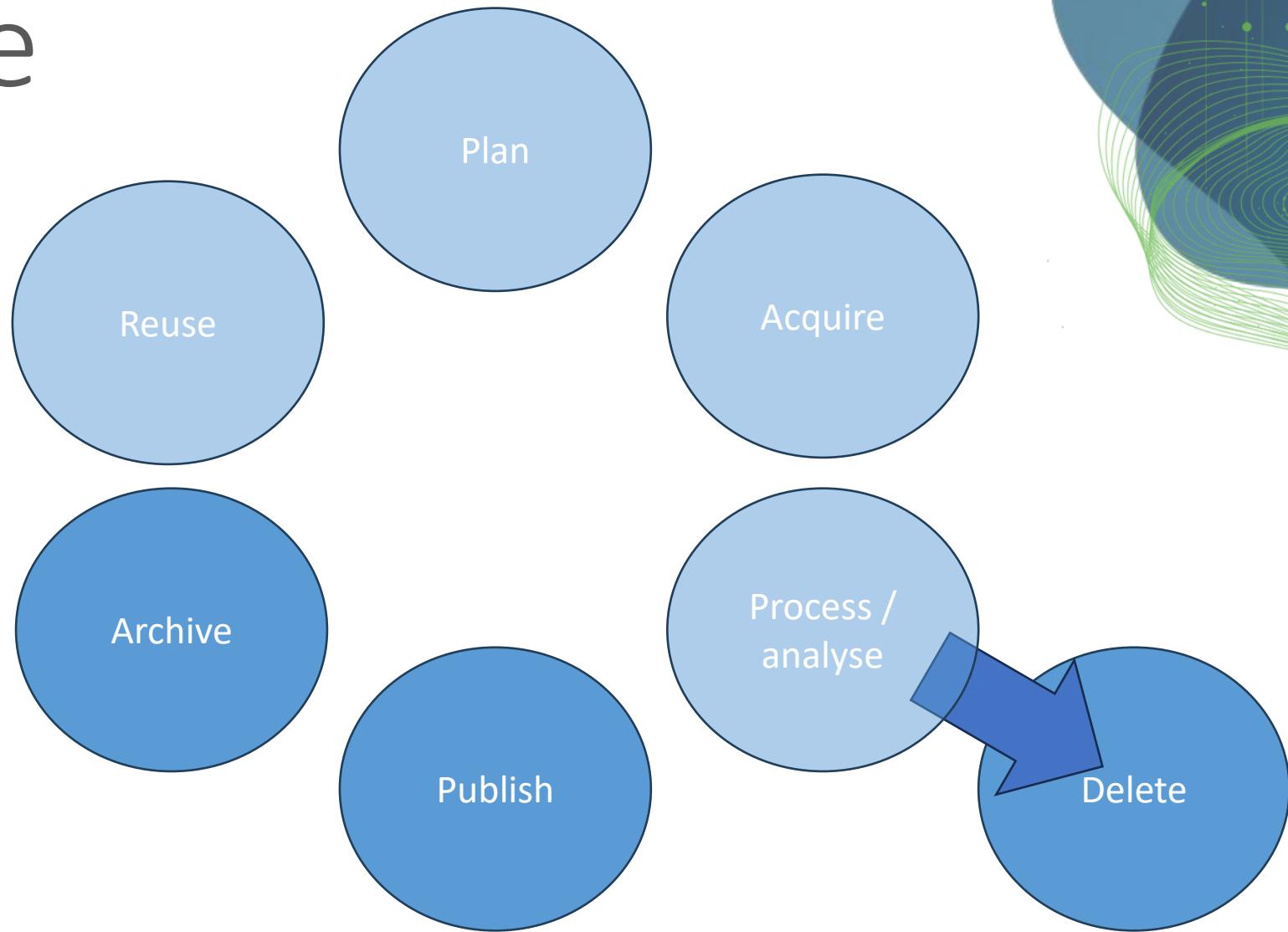


In need of  
protection  
(schutzbedürftig)



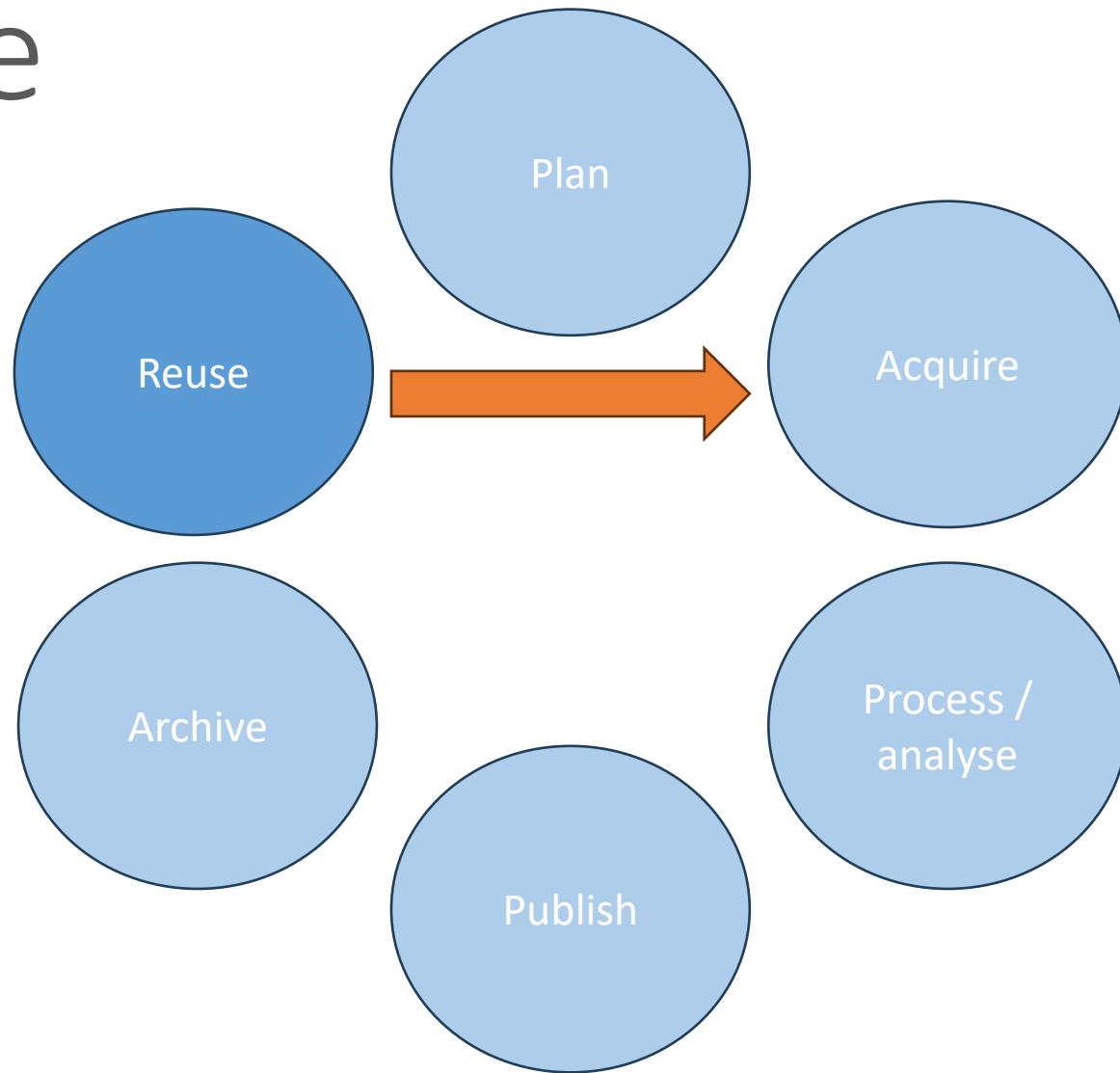
# RDM Life Cycle

- Right to publish
- Regulatory aspects
  - Research data: archive 15 years
- Authorship
- Registration (-> Findable)



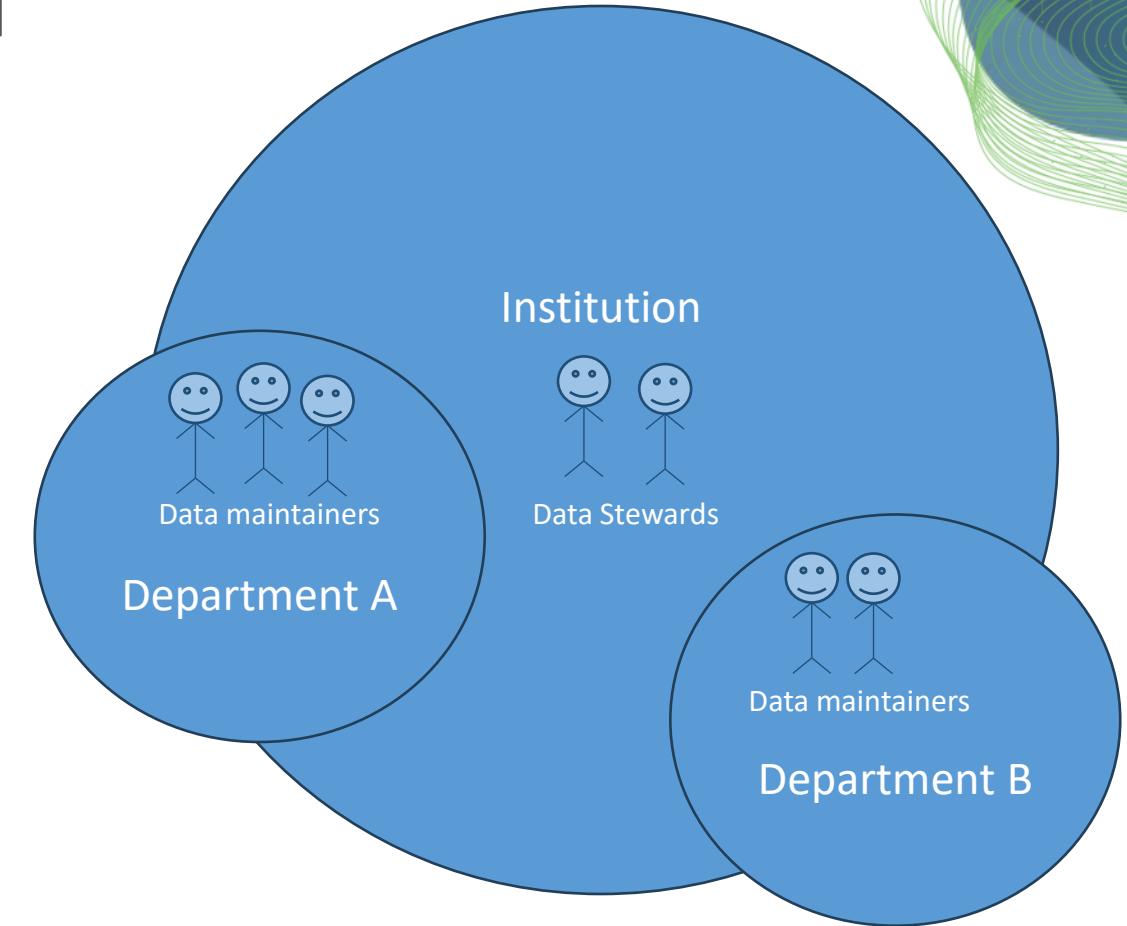
# RDM Life Cycle

- Potential future benefit
- Sustainability
- Important: **Licensing**
  - Has impact on next cycle / acquisition



# What is good RDM?

- Clearly defined responsibilities and processes (Governance)
  - Data Management Plan (DMP)
- Communication of goals, metrics, responsibilities, processes
- Dedicated personnel
  - “Data maintainers”
  - IT infrastructure maintainers
- Expert consultants
  - “Data stewards”



# • Roles != Job profiles

## Domain specialist

- Focuses on scientific question, often related to the physical world
- Requires sound insights and sustainable solutions
- Examples:
  - Biologist
  - Geologist
  - City planner

## Data analyst

- Focuses on methods for data processing / visualization
- Gains sound insights
- Examples:
  - Statistician
  - Bioinformatician
  - Data Scientist

## IT specialist

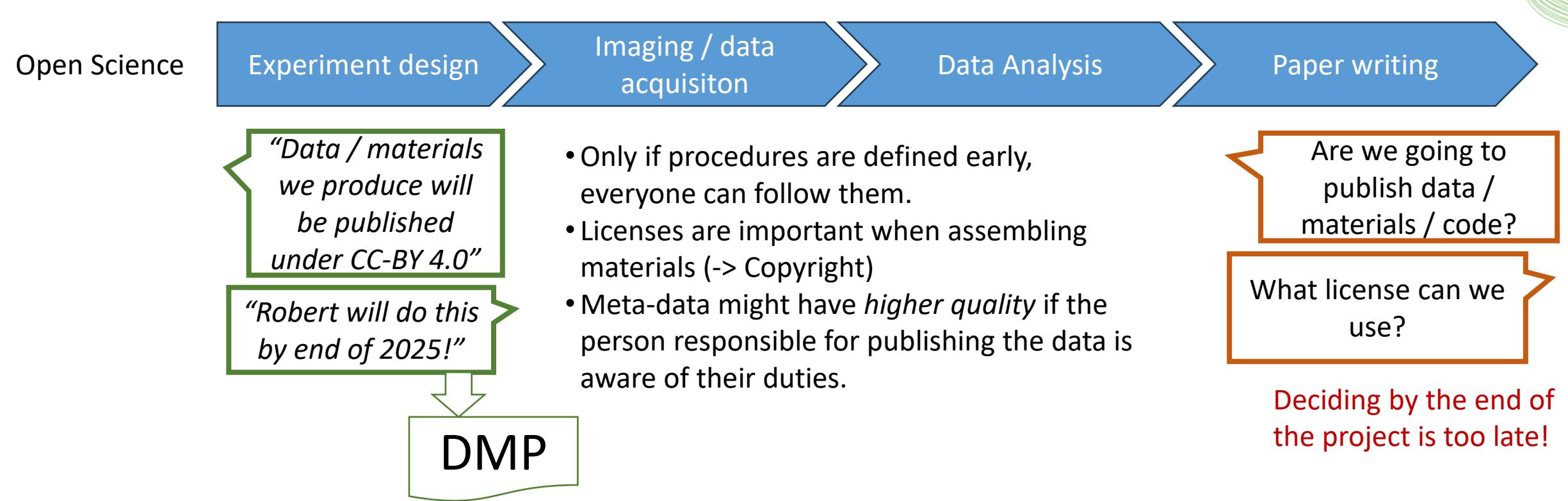
- Focuses on IT infrastructure
  - Hardware
  - Software
- Builds sustainable solutions
- Examples:
  - Computer scientist
  - IT specialist

# Data Management Plans (DMPs)

- Describes the IS-state of a data environment
  - Which data is acquired / processed?  
(content, format, amount)
  - What meta-data is collected?
  - Which quality standards are targeted?
  - How is data saved, archived, backed-up, shared, published...?
  - Who is responsible for what?
    - Roles, job-profiles
  - What does this all cost?  
(IT infrastructure + human resources)

# Data Management Plans (DMPs)

- Define responsibilities and procedures early!



# Quiz

- Regularly copying files to a remote place is ...

Archiving



Backup



# Quiz

- Data Scientists is a ...

Role



Job profile



# Quiz

- Data Steward is a ...

Role

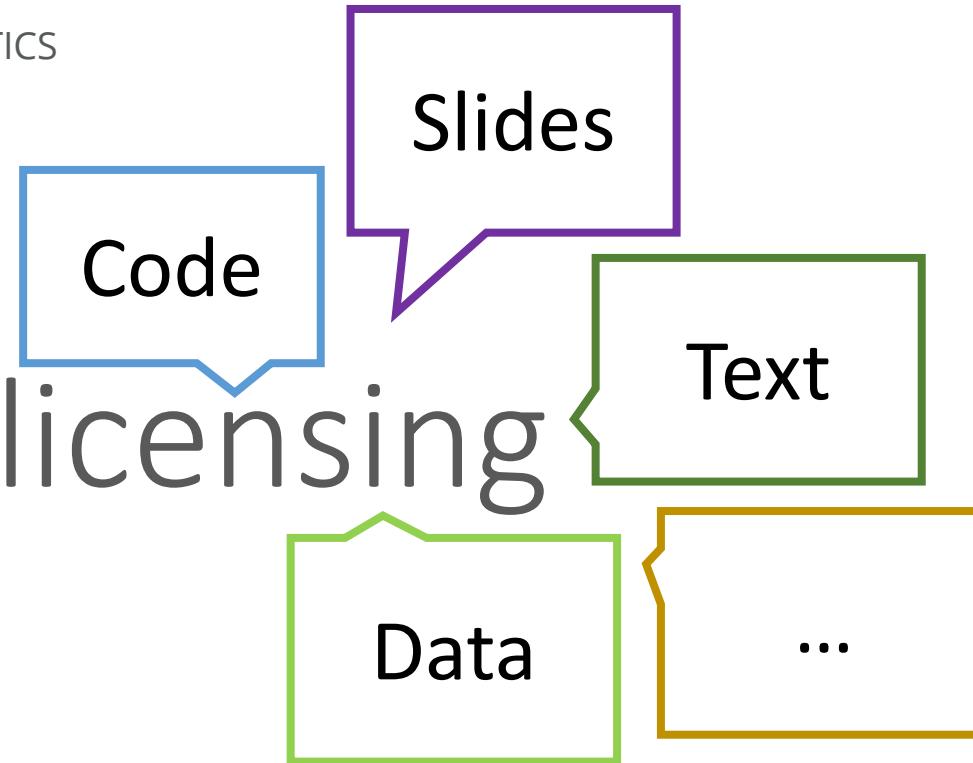


Job profile



# Sharing & licensing

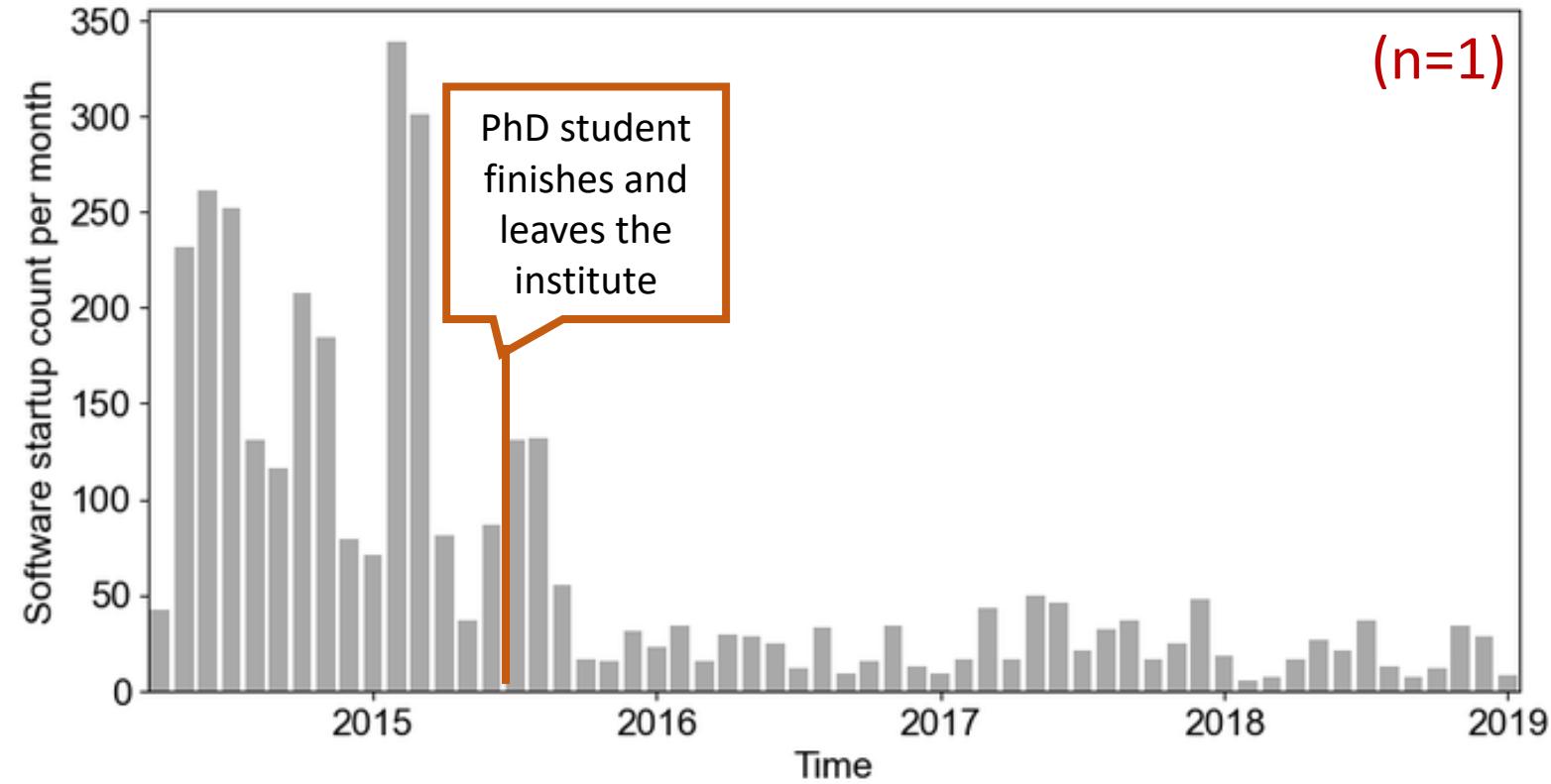
Robert Haase



Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages.  
Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

# Sustainability of my contribution to science

- What happens to research software once the PhD student leaves the institute / field?



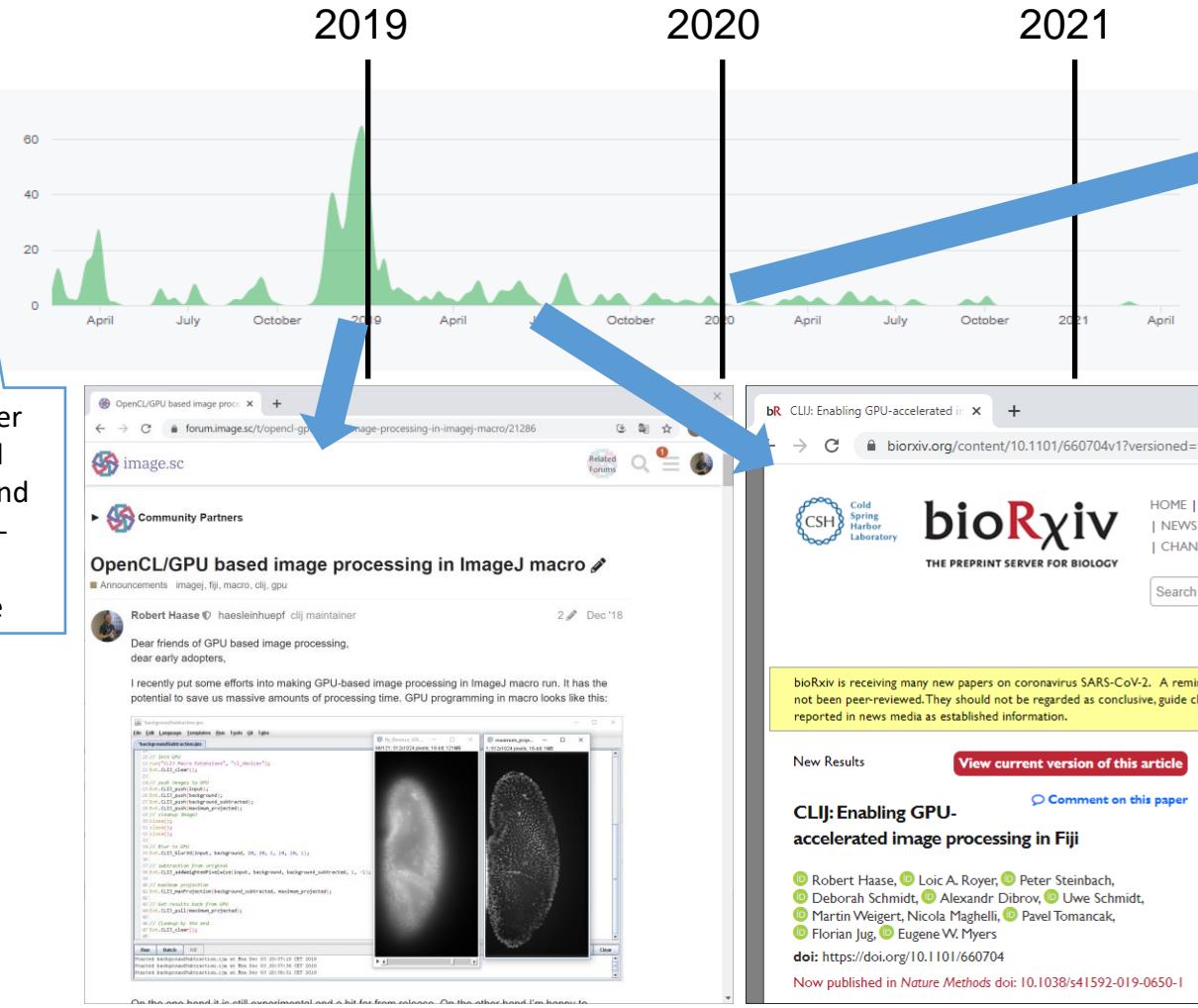
# Developing software in the open



Nov. 2017: I took over microscopy control software from Loic and “found” some GPU-accelerated image processing in there



Loic A. Royer  
(CZ Biohub)  
@loicaroyer



**nature > nature methods > correspondence > article**

## nature methods

Correspondence | Published: 18 November 2019

### CLIJ: GPU-accelerated image processing for everyone

Robert Haase , Loic A. Royer , Peter Steinbach, Deborah Schmidt, Alexander Dibrov, Uwe Schmidt, Martin Weigert, Nicola Maghelli, Pavel Tomancak, Florian Jug & Eugene W. Myers

Nature Methods 17, 5–6(2020) | Cite this article

Today: 134 citations  
(Google scholar, 2024-03-18)

Zitiert von: 134

The chart shows the citation count for the article over time. The counts are approximately: 2020 (~10), 2021 (~25), 2022 (~45), 2023 (~40), and 2024 (~10).

# Scientific culture

Public access to research results -> Reusability



## Guideline 13: Providing public access to research results

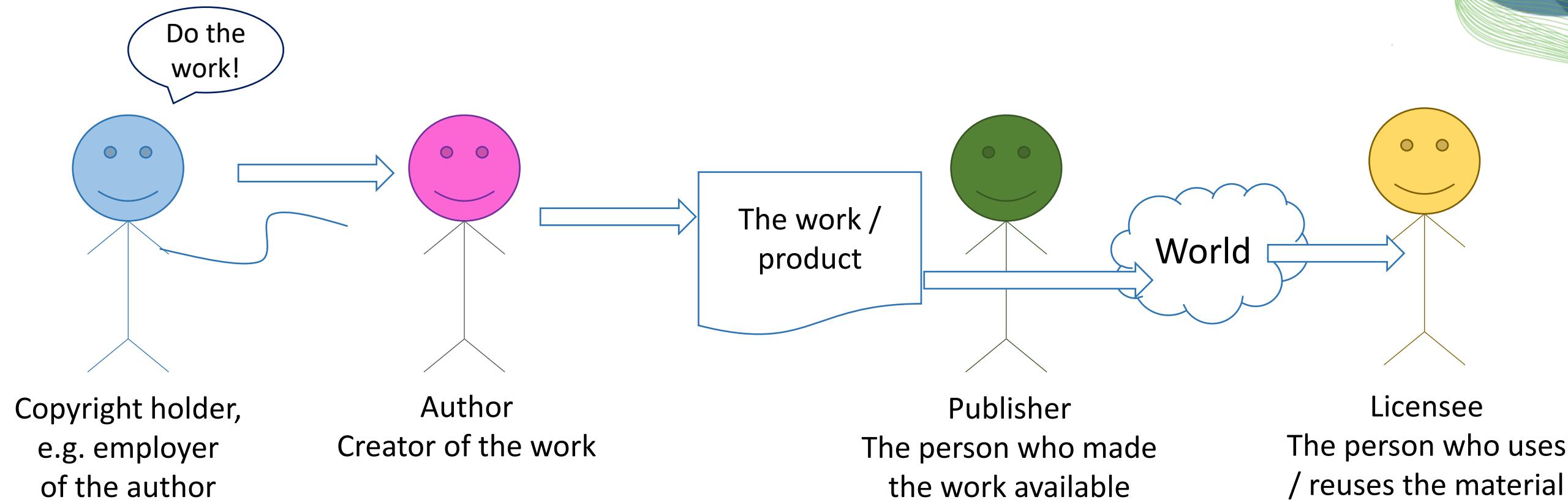
► As a rule, researchers make all results available as part of scientific/academic discourse. In specific cases, however, there may be reasons not to make results publicly available (in the narrower sense of publication, but also in a broader sense through other communication channels); this decision must not depend on third parties. Researchers decide autonomously – with due regard for the conventions of the relevant subject area – whether, how and where to disseminate their results. If it has been decided to make results available in the public domain, researchers describe them clearly and in full. Where possible and reasonable, this includes making the research data, materials and information on which the results are based, as well as the methods and software used, available and fully explaining the work processes. Software programmed by researchers themselves is made publicly available along with the source code. Researchers provide full and correct information about their own preliminary work and that of others.

### Explanations:

In the interest of transparency and to enable research to be referred to and reused by others, whenever possible researchers make the research data and principal materials on which a publication is based available in recognised archives and repositories in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable). Restrictions may apply to public availability in the case of patent applications. If self-developed

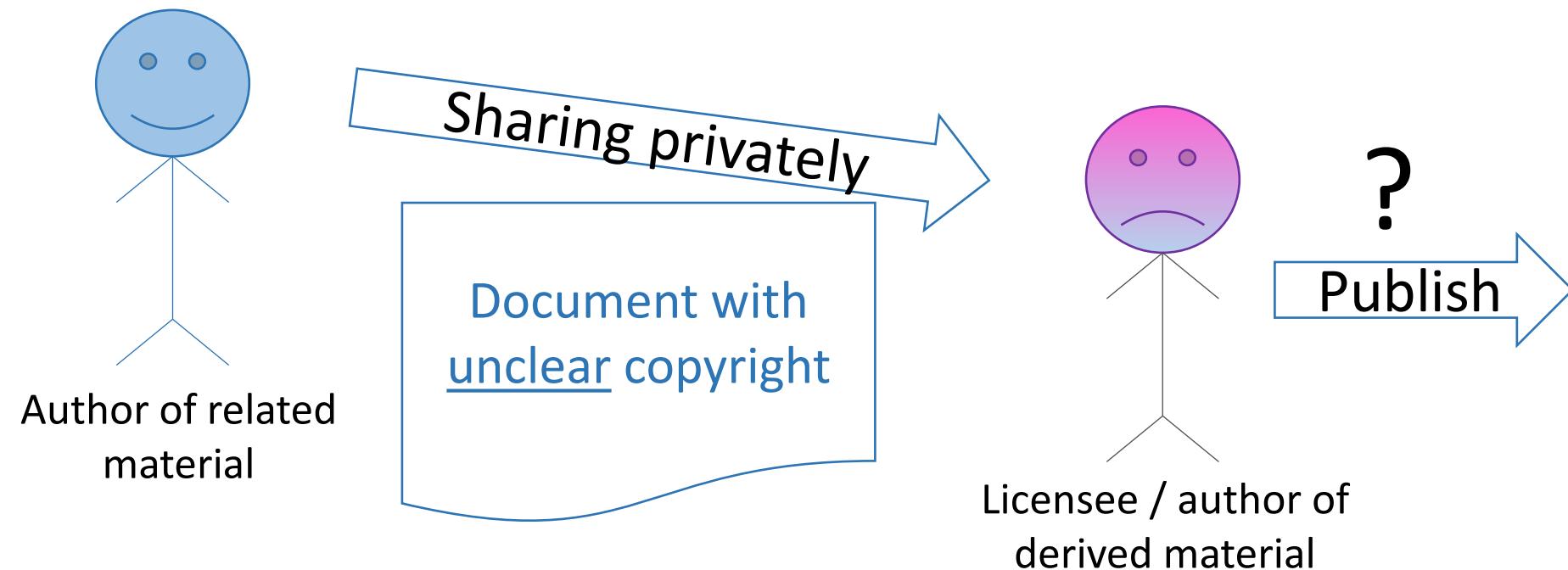
# Am I allowed to publish my stuff?

- ... it depends... on who is responsible



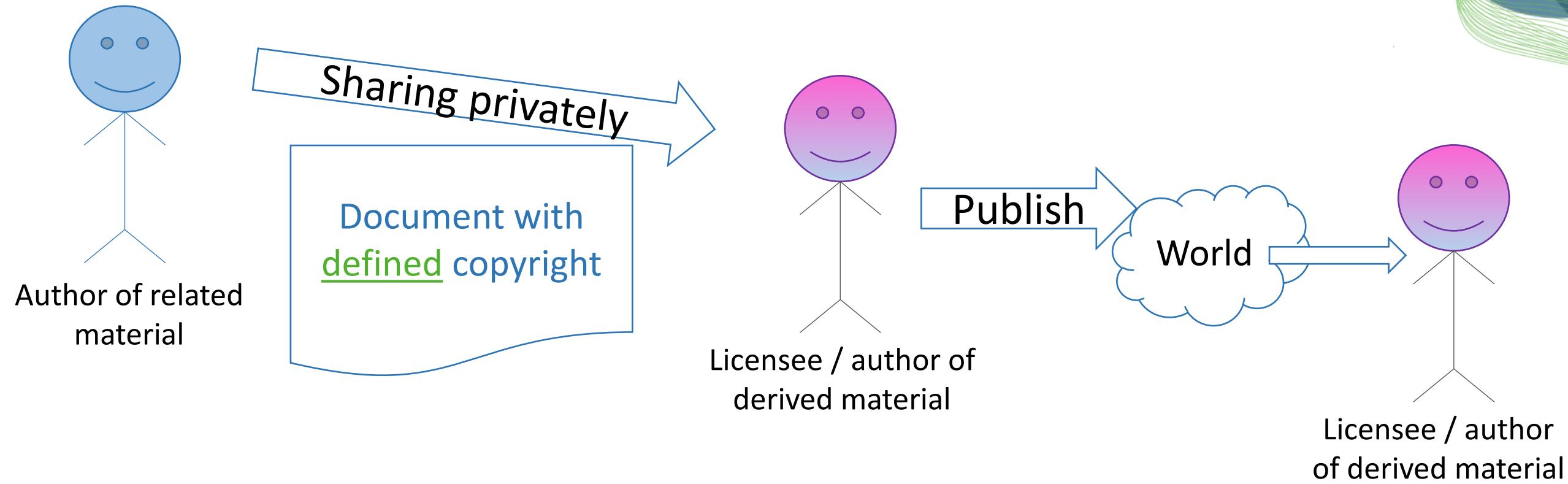
# Am I allowed to publish my stuff?

- ... it depends... on what materials served as basis



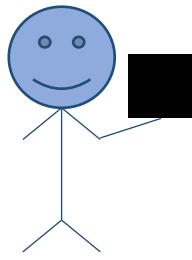
# Am I allowed to publish my stuff?

- ... it depends... on what materials served as basis



# Openness of software / projects

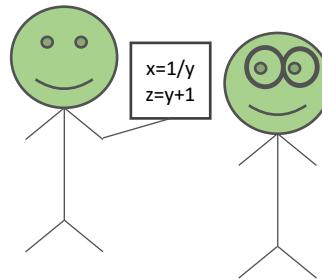
## Closed source



- Open to collaborations
- “Black box”
- Compiled code (e.g. C/C++)
- Good for protecting intellectual properties (\$\$\$)

Hardware device drivers

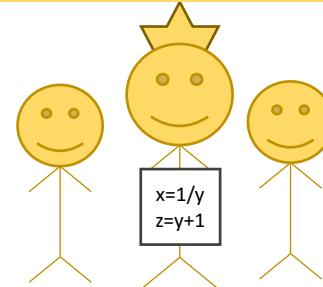
## Open source



- Code available to read
- Not necessarily executable code
- No maintenance / support efforts

Custom image analysis scripts

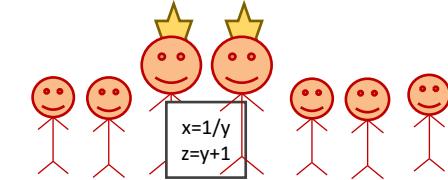
## Benevolent dictatorship



- Open to contributions
- Single maintainer, often overwhelmed
- Efficient decision making
- Bus factor ≈1

TrackMate, SNT, MorpholibJ, CLIJ

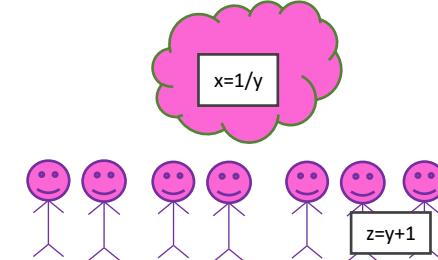
## Community driven



- Open to contributions
- Partially democratic
- Board of maintainers (core developers)
- Long-winded decision making

scikit-image, scipy, OpenCL

## Openly extensible



- Openly extensible; without maintainers involved
- Partially community driven

ImageJ, Python, numpy

# Quiz

- What is the role of Github in the context of publishing open-source code?

Copyright holder



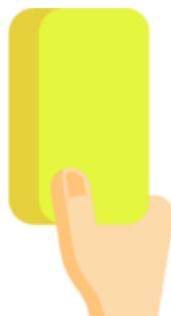
Author



Publisher



Licensee



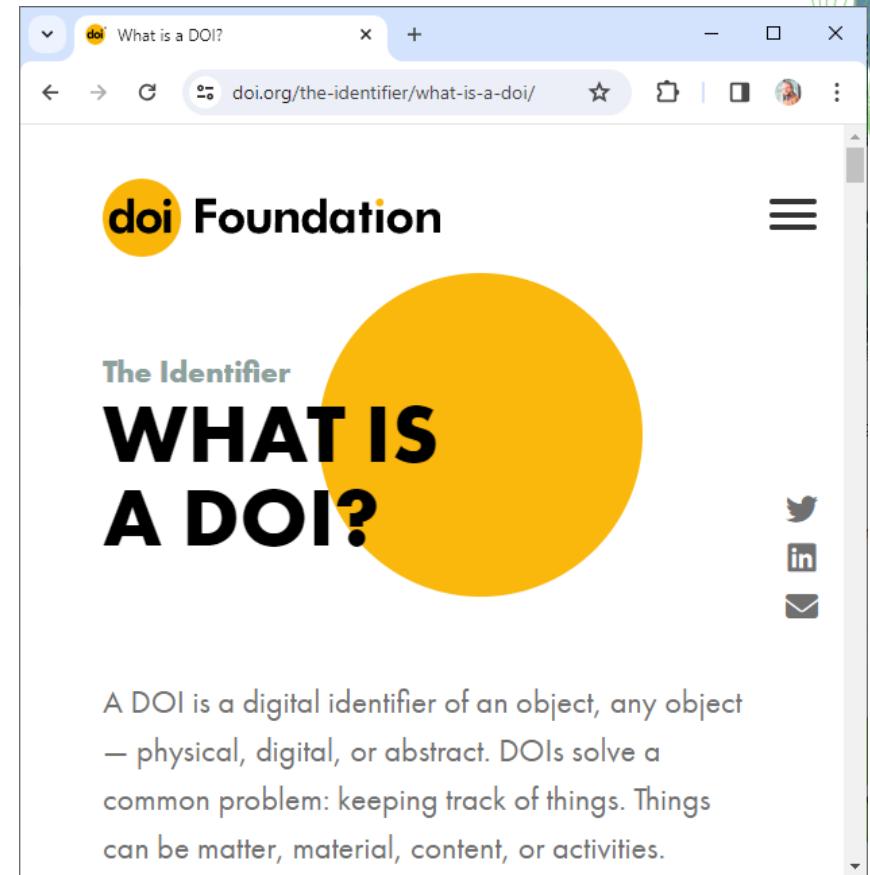
# Standard for sharing: The FAIR-principles

- Findable
- Accessible
- Interoperable
- Reusable



# The FAIR-principles

- Findable
- F1. (Meta)data are assigned a globally unique and persistent identifier
  - Universal Resource Identifier (URI)
  - Digital Object Identifier (DOI)
- F2. Data are described with rich metadata (defined by R1 below)
- F3. Metadata clearly and explicitly include the identifier of the data they describe
- F4. (Meta)data are registered or indexed in a searchable resource

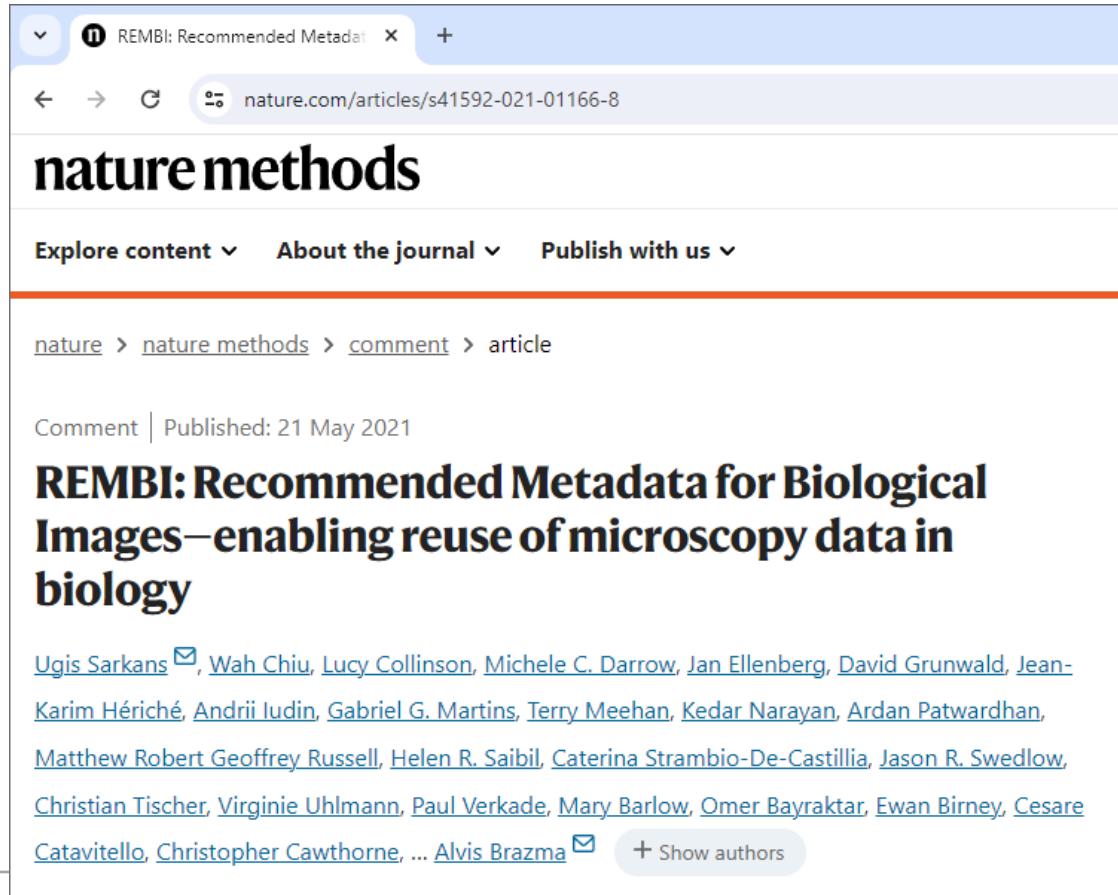


# Meta data

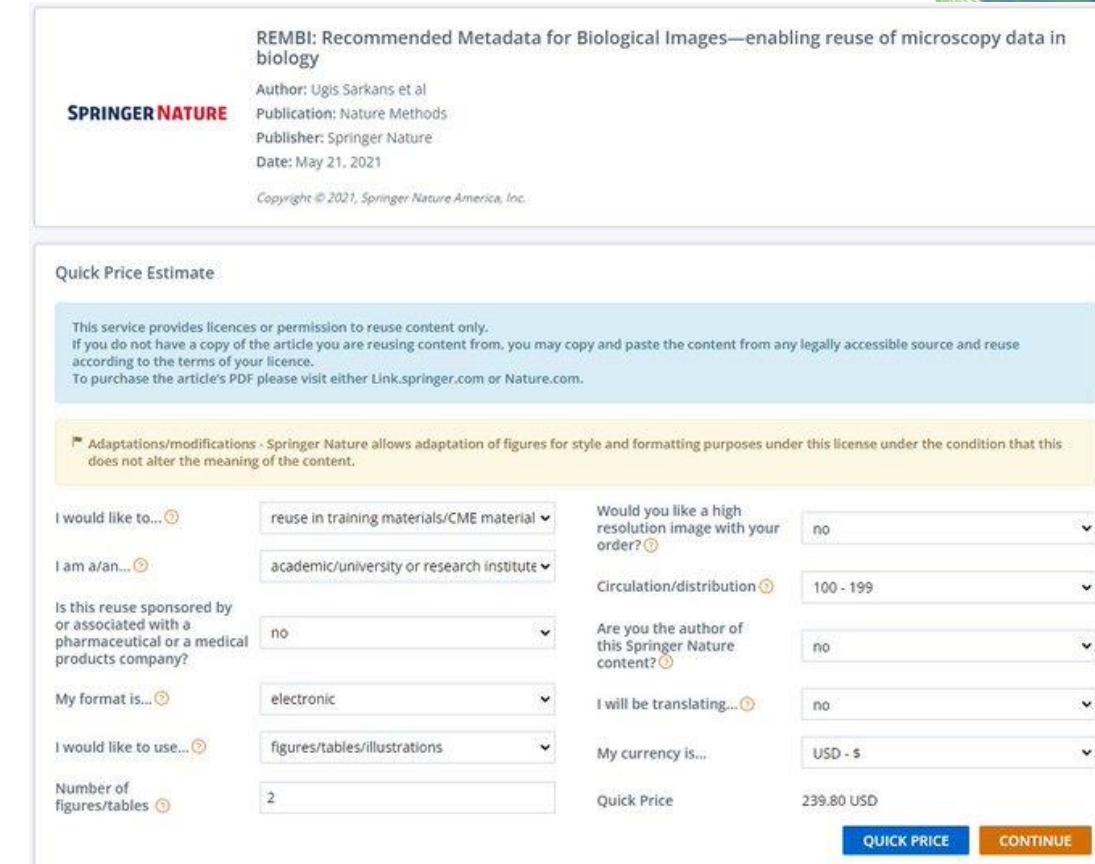
- Generic
  - Author
  - Usage license
  - Creation date
- Field-specific (microscopy)
  - Exposure time
  - Wavelength (colour)
  - Microscope type/vendor

# REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology

- Read more:



A screenshot of a web browser showing the REMBI article on nature.com. The title "REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology" is displayed prominently. Below the title, a list of authors is provided, including Ugis Sarkans, Wah Chiu, Lucy Collinson, Michele C. Darrow, Jan Ellenberg, David Grunwald, Jean-Karim Hériché, Andrii Iudin, Gabriel G. Martins, Terry Meehan, Kedar Narayan, Ardan Patwardhan, Matthew Robert Geoffrey Russell, Helen R. Saibil, Caterina Strambio-De-Castillia, Jason R. Swedlow, Christian Tischer, Virginie Uhlmann, Paul Verkade, Mary Barlow, Omer Bayraktar, Ewan Birney, Cesare Catavitello, Christopher Cawthorne, and Alvis Brazma. A "Show authors" button is visible at the bottom right of the author list.



A screenshot of the REMBI article page on Springer Nature's website. The title "REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology" is shown. Below the title, the Springer Nature logo is present. To the right, there is a "Quick Price Estimate" section with fields for reuse type, academic institution, company sponsorship, format, and currency. The price listed is 239.80 USD. At the bottom right are "QUICK PRICE" and "CONTINUE" buttons.

# Digital Object Identifiers (DOI)

- DOIs / URIs always point at the same data
- DOIs are centrally registers, URIs not
- Unified Resource Locators (URLs) may point at different things

The screenshot shows a web browser window for the 'Straßennetz, Stadt Leipzig' dataset on the [opendata.leipzig.de](https://opendata.leipzig.de/dataset/strassennetz-stadt-leipzig) website. The page includes:

- Formats for download:** CSV, GeoJSON, GeoPackage, WFS.
- Contact Information:** Ansprechpartner: Verkehrs- und Tiefbauamt, Stadt Leipzig; E-Mail: vta@leipzig.de.
- Administrative Details:** Gemeindenname: Leipzig, Stadt; Ausgestellt: 2021-08-20; Aktualisiert: 2024-01-17.

A blue callout box points to the administrative table with the text: "This no DOI, no URI, it's a URL".

# Unified Resource Identifiers

- Which of these are URIs?

<https://twitter.com/haesleinhuepf/status/891596662782779392>

<https://doi.org/10.5281/zenodo.28325>

<https://opendata.leipzig.de/dataset/vornamestatistik-2023>

<https://www.leipzig.de/>

# Digital Object Identifiers

- Which of these are DOIs?

<https://twitter.com/haesleinhuepf/status/891596662782779392>

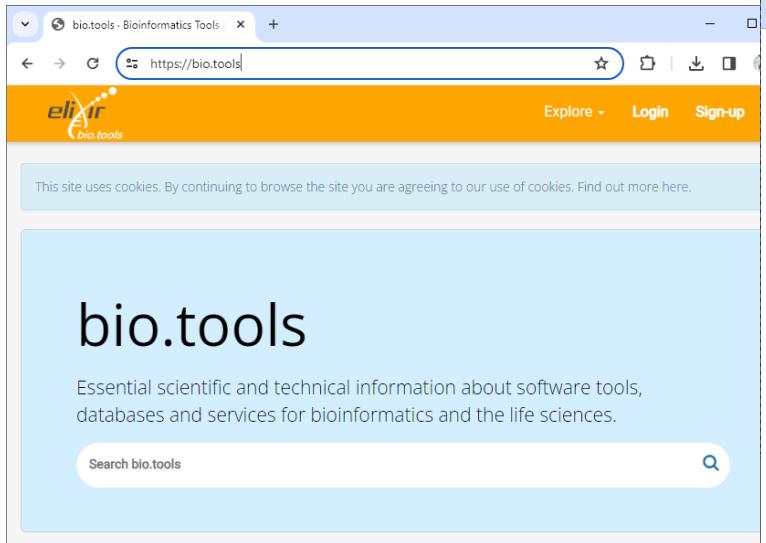
<https://doi.org/10.5281/zenodo.28325>

<https://opendata.leipzig.de/dataset/vornamestatistik-2023>

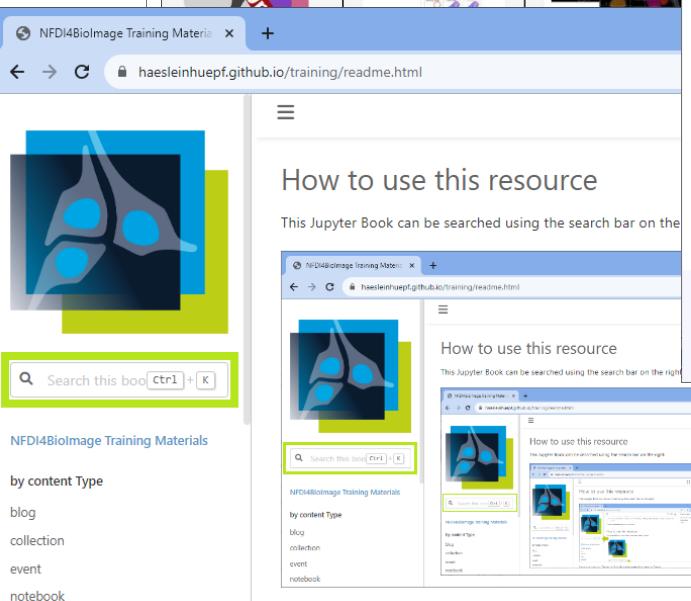
<https://www.leipzig.de/>

# Indexing

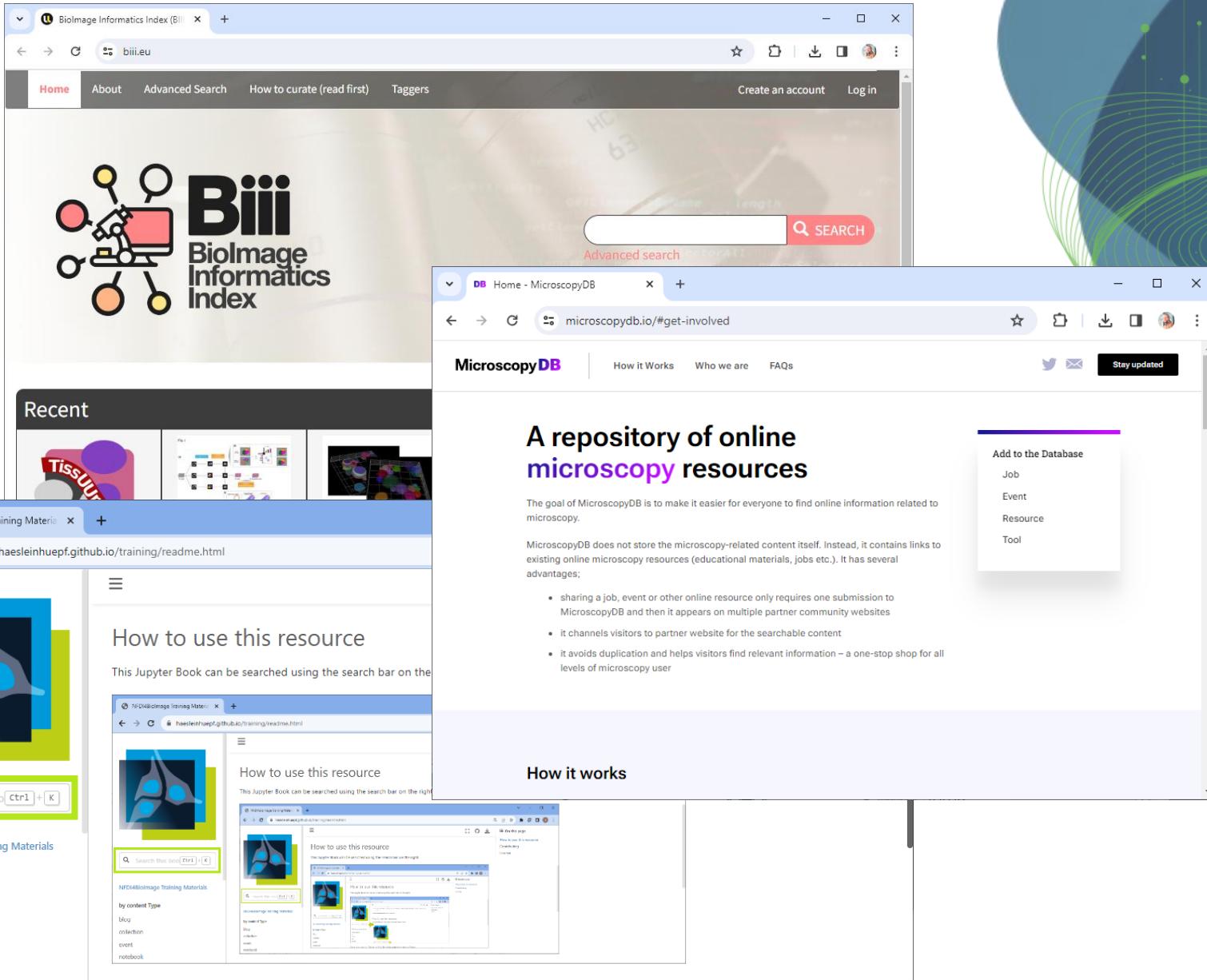
- Make sure your materials are listed in public search indices
- Do not trust google to make your stuff findable



The bio.tools homepage features the Elixir bio.tools logo at the top. Below it is a search bar and a large "bio.tools" title. A sub-header states: "Essential scientific and technical information about software tools, databases and services for bioinformatics and the life sciences." A sidebar on the right contains a search bar and a "by content Type" section with links for blog, collection, event, and notebook.



A screenshot of a Jupyter Book titled "NFDI4BioImage Training Materials". It shows a thumbnail of a brain image, a search bar with the placeholder "Search this book [ctrl + K]", and a sidebar with a "How to use this resource" section and a "by content Type" section with links for blog, collection, event, and notebook.

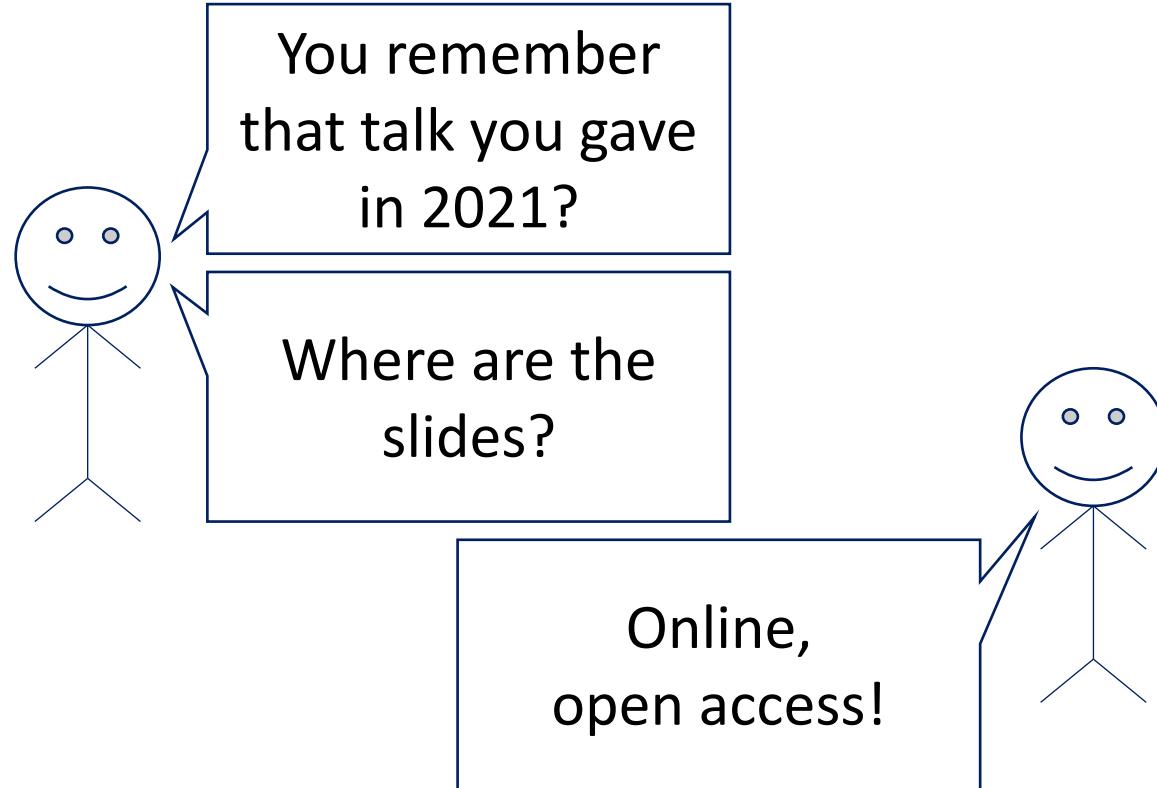


Three screenshots illustrating indexing and discoverability:

- BiiI (Biolimage Informatics Index):** Shows the BiiI logo and a search bar. The "Recent" section displays thumbnails for "Tissue" and other resources.
- MicroscopyDB:** Shows the MicroscopyDB homepage with a search bar and a "How it works" section. It highlights that MicroscopyDB is a repository of online microscopy resources.
- NFDI4BioImage Training Materials:** Shows a screenshot of the training materials page, demonstrating how the content can be indexed and searched across multiple platforms.

# Incentives: Findability

- Your *future-self* will thank you, because they will find your work



F1000Research

BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ABOUT BLOG MY RESEARCH SIGN IN

Sharing and licensing material | f1000research.com/slides/10-519

Home > Browse > Sharing and licensing material

NOT PEER REVIEWED

VIEW FULL SCREEN

SLIDES

PowerPoint P... 1 / 28 24%

Sharing and licensing material  
Robert Haase  
June 30<sup>th</sup> 2021

Code Slides Text Data ...

This material is licensed by Robert Haase, PoL Dresden under the CC-BY 4.0 license <https://creativecommons.org/licenses/by/4.0/>

Metrics | 411 Views | 60 Downloads

DOWNLOAD 30.92 MB

SHARE CITE

PART OF THE GATEWAY

neubias - the Bioimage Analysts Network

BROWSE BY RELATED SUBJECTS

Artificial intelligence

Computer and information sciences

Electrical engineering

# Incentives: Findability -> Visibility

- YouTube
- Github

Open & FAIR sharing  
is a PR instrument

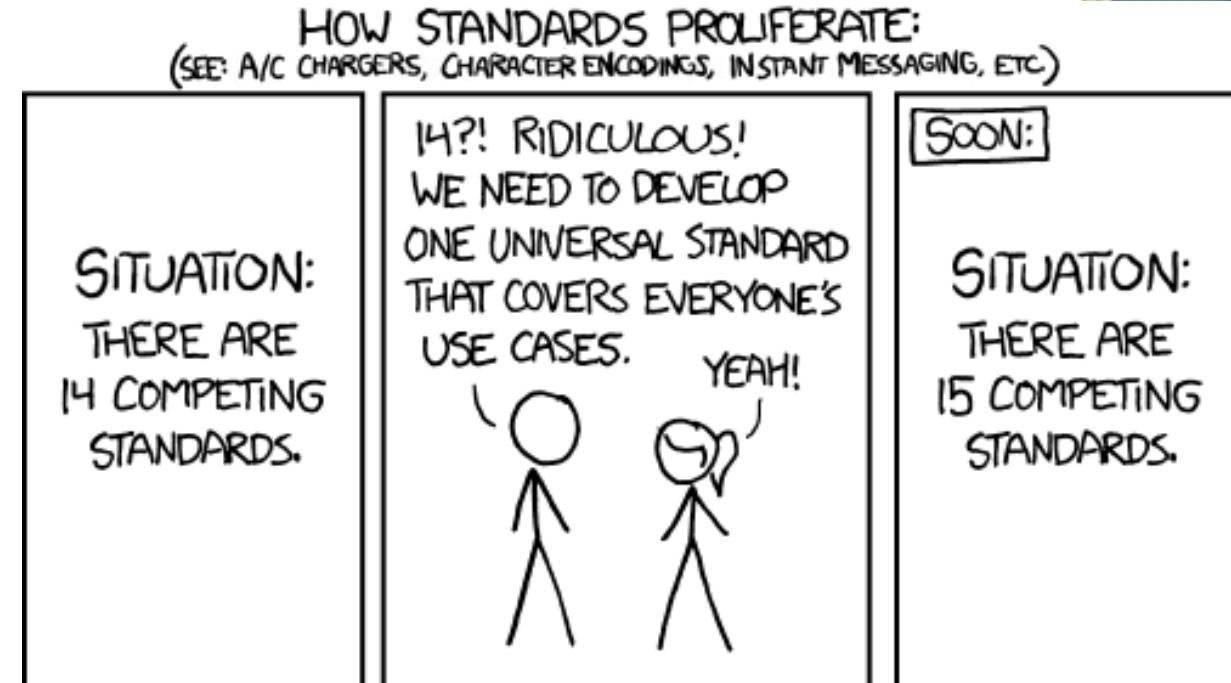
... leading to

- more software users
- new collaborations

The image displays three browser windows side-by-side. The left window shows the YouTube channel 'ScaaS-AI Living Lab' with several playlists listed under 'Created playlists'. The middle window shows the GitHub page for the 'Prompt Engineering Tutorial'. The right window shows a Jupyter Notebook titled '04\_generating\_images.ipynb' with Python code and generated image thumbnails.

# The FAIR-principles

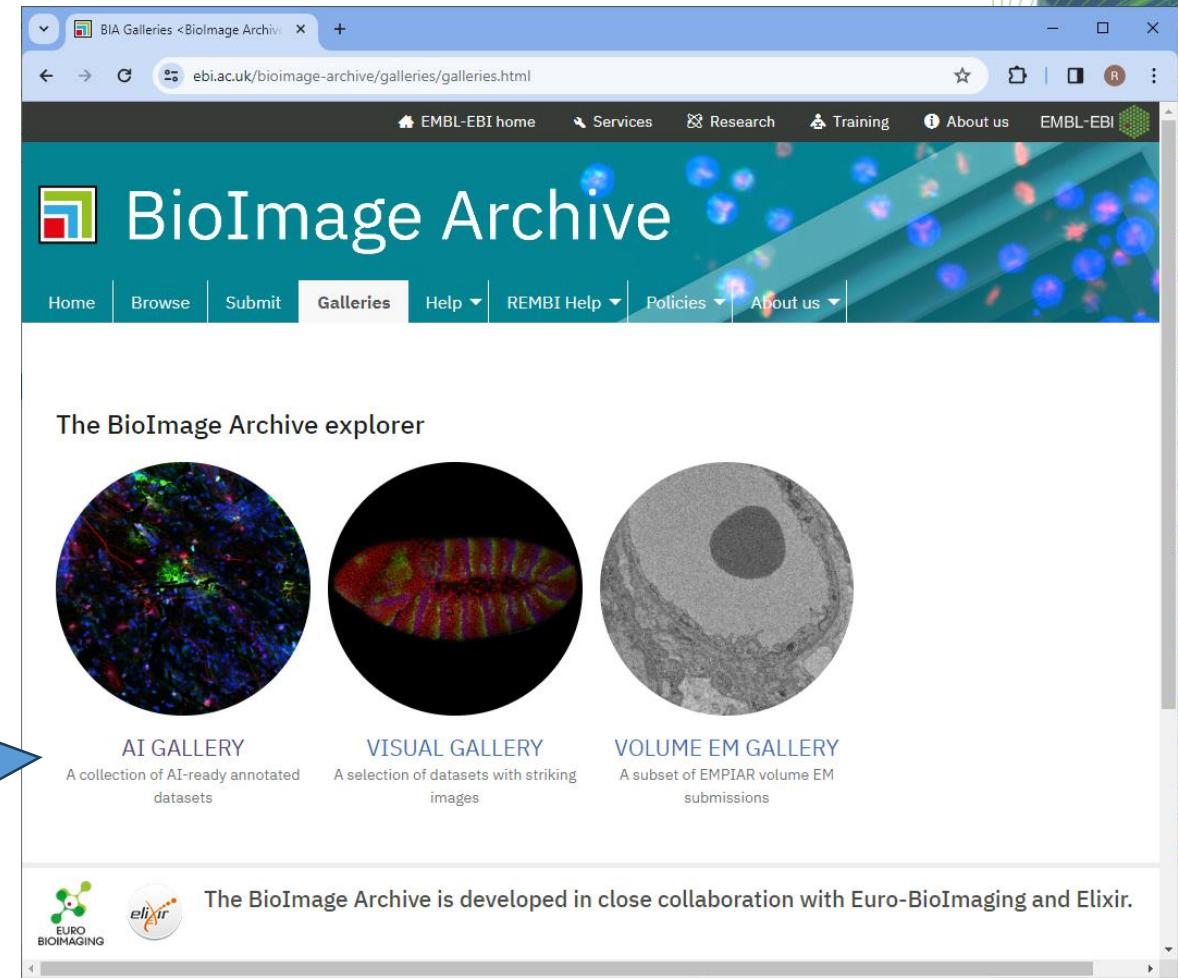
- Accessible
- A1. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - A1.1 The protocol is open, free, and universally implementable
  - A1.2 The protocol allows for an authentication and authorisation procedure, where necessary
- A2. Metadata are accessible, even when the data are no longer available



# Accessibility

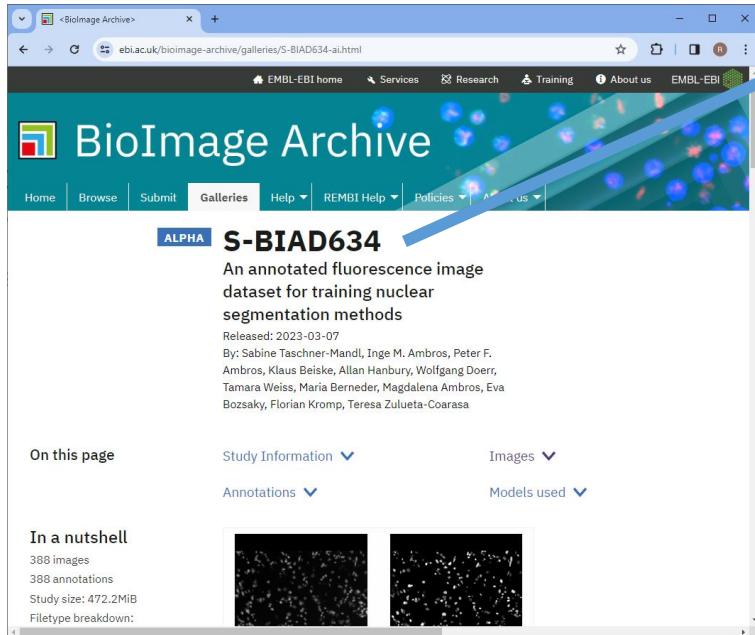
- The ability to download data, for humans and computers

Essential for AI  
developers =-)



# Accessibility

- The ability to download data, for humans and computers



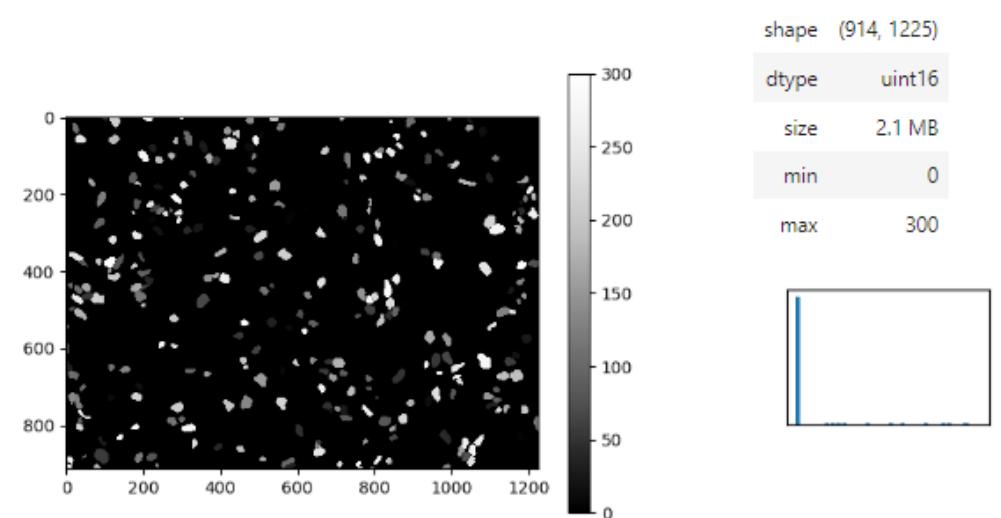
```
[1]: from bia_explorer import io, biostudies
from skimage.io import imread
import stackview

accession = 'S-BIAD634'
study = io.download_bia_study(accession)
image = study.images[0]
```

## Displaying images using stackview

```
[2]: uri = image.uri.replace("\\\\", "/")
image_data = imread(uri)
stackview.insight(image_data)
```

[2]:



# Restricted Access

- The A in FAIR does not necessarily stand for Open Access

The image displays two side-by-side screenshots of the Zenodo dataset page for the file 'blobs.tif'. Both screenshots show the same dataset information but differ in their access settings.

**Left Screenshot (Restricted Access):**

- The dataset is labeled as 'Restricted' in the top navigation bar.
- The 'Files' section shows a red box around the 'Restricted' status, with a blue arrow pointing to it from above.
- The 'Citations' section shows a red box around the 'Show' button.

**Right Screenshot (Open Access):**

- The dataset is labeled as 'Dataset' and 'Restricted' in the top navigation bar.
- The 'Edit' button is highlighted in orange.
- The 'New version' button is highlighted in green.
- The 'Share' button is highlighted in blue.
- The 'Files' section shows a pink box around the file thumbnail.

**Common Dataset Details:**

- Published March 18, 2024 | Version v1
- Haase, Robert<sup>1,2</sup> (DOI: 10.5281/zenodo.10829230)
- This dataset contains blobs.tif, which was published before as blobs.gif as part of ImageJ's example images. The dataset is public-domain, available online in png format as well: <https://samples.fiji.sc/blobs.png>
- This record in Zenodo serves demonstrating that data can be published with closed access.
- Views: 0, Downloads: 0
- Versions: Version v1 (Mar 18, 2024, DOI: 10.5281/zenodo.10829230)
- Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.10829229. This DOI represents all versions, and will always resolve to the latest one. [Read more](#).
- External resources: Indexed in OpenAIRE

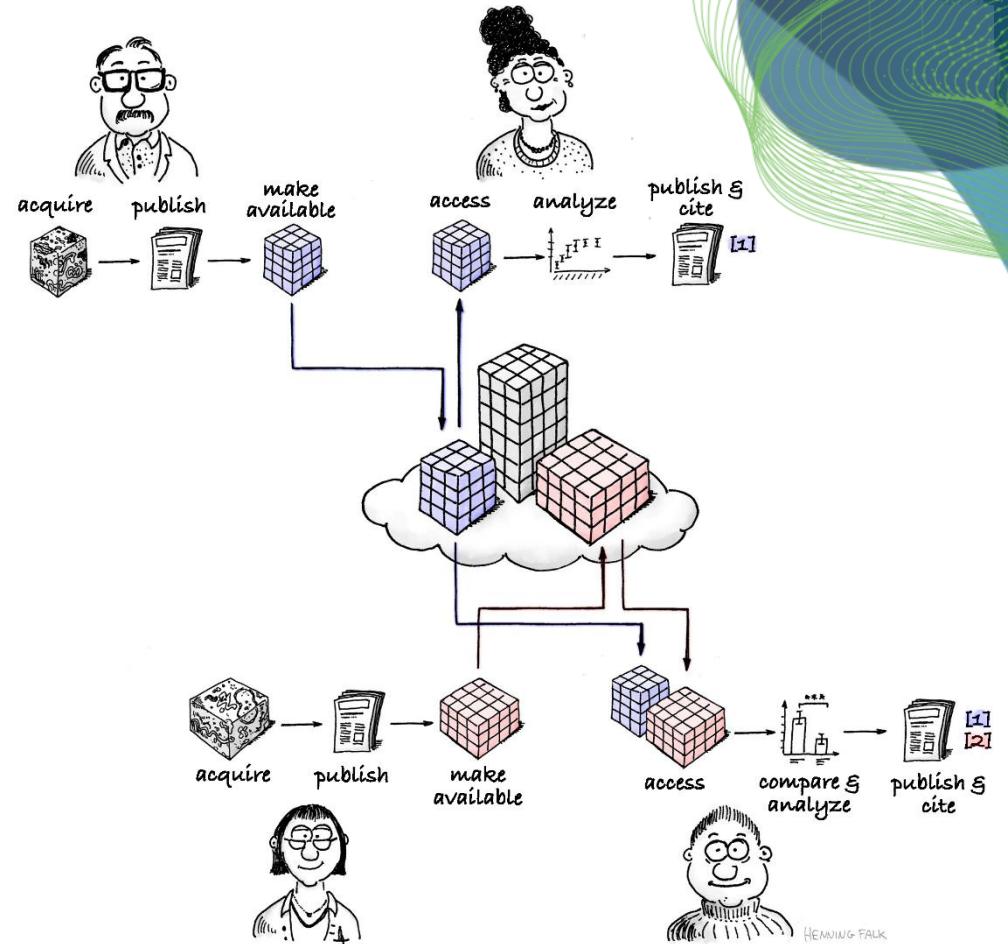
# The FAIR-principles

- Interoperable
  - I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - I2. (Meta)data use vocabularies that follow FAIR principles
  - I3. (Meta)data include qualified references to other (meta)data



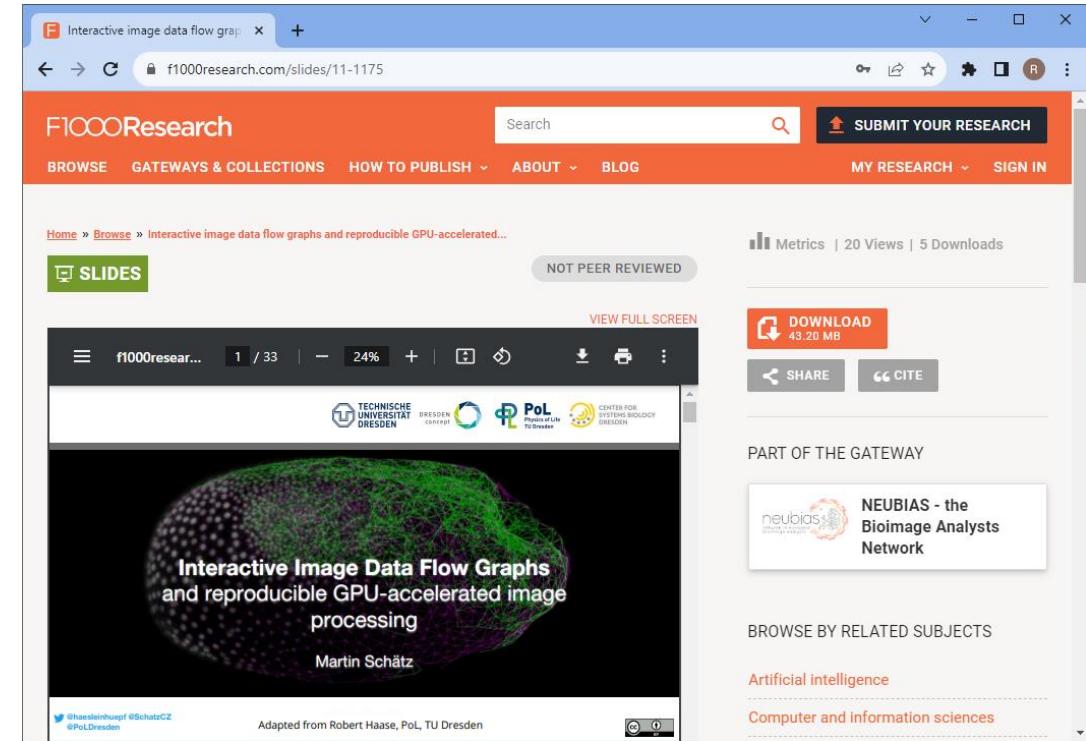
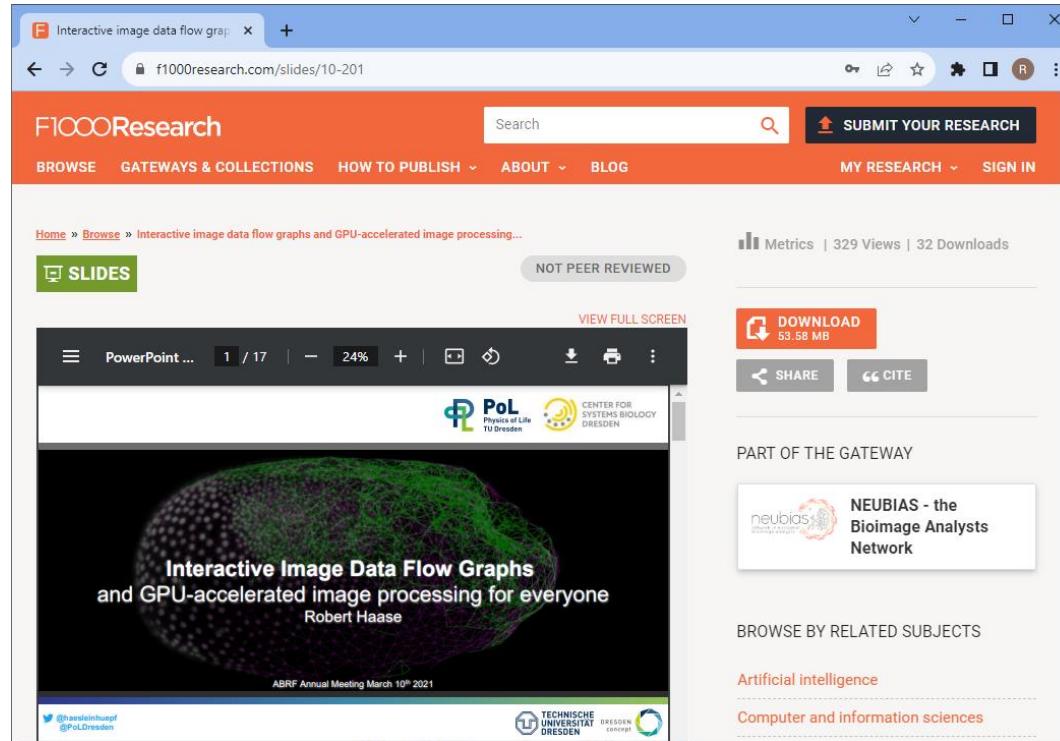
# The FAIR-principles

- Reusable
  - R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage license
  - R1.2. (Meta)data are associated with detailed provenance
  - R1.3. (Meta)data meet domain-relevant community standards



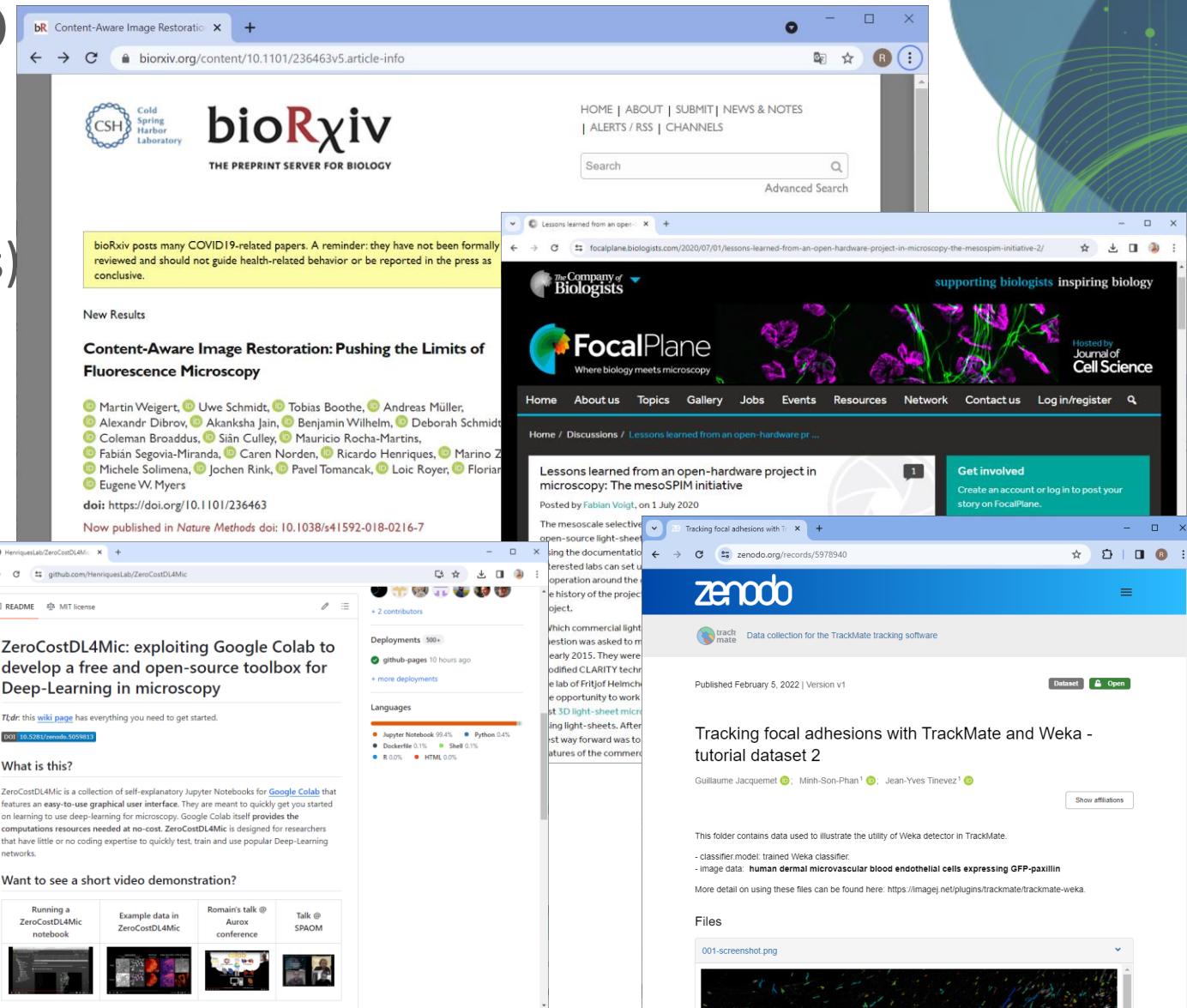
# Incentives: Reusability

- Open Access -> Others teach how to use your tools & methods



# Where to share?

- Open science related content
  - bioRxiv (manuscripts, no reviews)
  - Figshare
  - F1000
  - Bioimage Archive (data)
  - Github (code)
  - Zenodo
  - Focalplane
  - Institutional servers  
(if there is no alternative)



# Quiz

- Where might open source code be most *visible*?

Git server of the  
university



Zenodo.org



Github.com

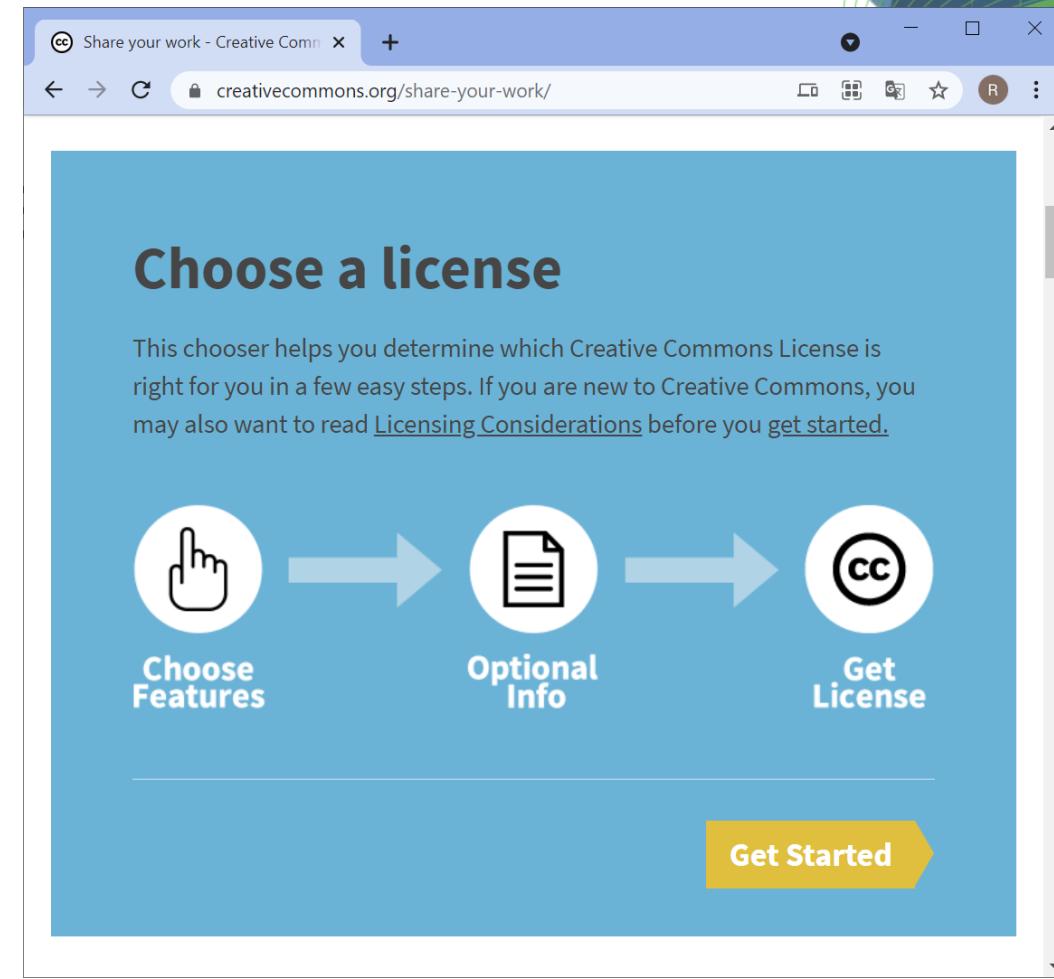


Group / institute /  
personal website



# Licensing: Creative Commons (CC)

- Public domain (CC0)
  - Attribution International (CC-BY)
  - Attribution ShareAlike Int. (CC-BY-SA)
  - Attribution Non-Commercial Int. (CC-BY-NC)
  - Attribution NoDerivatives Int. (CC-BY-ND)
- + Combinations, e.g. CC-BY-NC-ND



Content on this site is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)



- **Public domain (CC0)**
- Everyone can reuse without mentioning the source or author of the shared resource.
- Public domain licenses cannot be revoked.
- The author must own the right to copy (copyright) the resource.
  - If you authored work as part of your job, you may not be the copyright holder. (check your employers' guidelines)

Employers don't like this one because you give away the rights to *exploit* your work.

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)



“By attribution”

**CC BY:** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use.

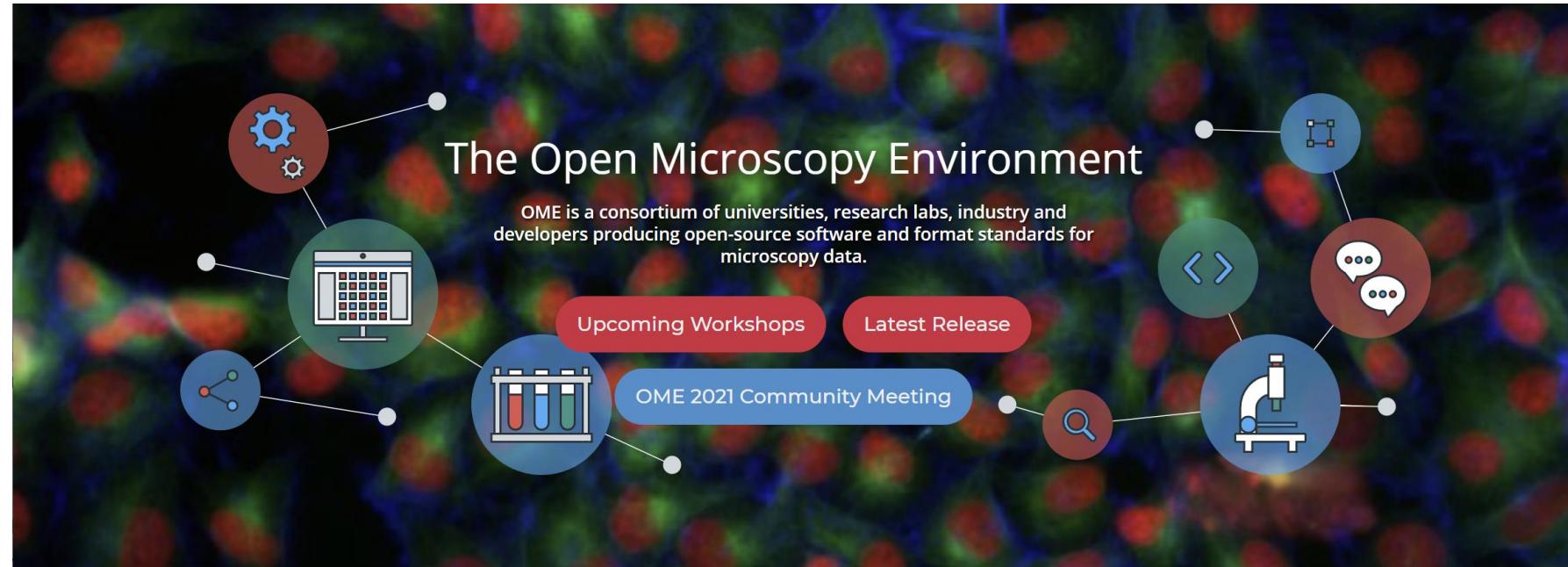
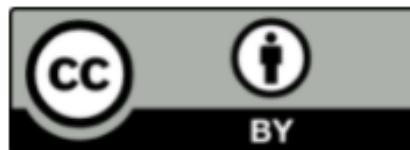
CC BY includes the following elements:

BY - Credit must be given to the creator

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)

## Example



You must put such a sentence and keep the link to CC-BY

Figure adapted from <https://www.openmicroscopy.org/> licensed by University of Dundee & Open Microscopy Environment under [Creative Commons Attribution 4.0 International License](#)

# Licensing: Creative Commons (CC)



“Share alike”

**CC BY-SA:** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, so long as attribution is given to the creator. The license allows for commercial use. If you remix, adapt, or build upon the material, you must license the modified material under identical terms.

CC BY-SA includes the following elements:

BY - Credit must be given to the creator

SA - Adaptations must be shared under the same terms

“Restrictive”  
licensing

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)



**CC BY-NC:** This license allows reusers to distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator.

It includes the following elements:

BY - Credit must be given to the creator

NC - Only noncommercial uses of the work are permitted

“Restrictive”  
licensing

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)



**CC BY-ND:** This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, and only so long as attribution is given to the creator. The license allows for commercial use.

CC BY-ND includes the following elements:

BY - Credit must be given to the creator

ND - No derivatives or adaptations of the work are permitted

“Restrictive”  
licensing

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Licensing: Creative Commons (CC)



to the creator.

**CC BY-NC-ND:** This license allows reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given

CC BY-NC-ND includes the following elements:

BY - Credit must be given to the creator

NC - Only noncommercial uses of the work are permitted

ND - No derivatives or adaptations of the work are permitted

“Restrictive”  
licensing

Content on this slide was adapted from <https://creativecommons.org/about/cclicenses/> which is licensed under a [Creative Commons Attribution 4.0 International license](#). Icons by The Noun Project.

# Quiz

- May I use one of the Figures from this preprint?
- May I download and redistribute this preprint to students of a course for free?

The image shows a screenshot of a bioRxiv preprint page. At the top left is the CSHL logo and the bioRxiv logo with the tagline "THE PREPRINT SERVER FOR BIOLOGY". A yellow banner at the top states: "bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive." Below this, the title "Content-Aware Image Restoration: Pushing the Limits of Fluorescence Microscopy" is displayed, along with the names of the authors and their ORCID IDs. The DOI is listed as <https://doi.org/10.1101/236463>. It also mentions that it was published in *Nature Methods* with DOI [10.1038/s41592-018-0216-7](https://doi.org/10.1038/s41592-018-0216-7). Below the title are social media sharing icons (1 comment, 0 tweets, 2 likes, 0 shares, 2 saves, 0 reads, 618 views) and navigation links for Abstract, Full Text, Info/History (which is selected), and Metrics. A "Preview PDF" link is also present. The "ARTICLE INFORMATION" section includes the DOI and a "History" entry for July 3, 2018. The "ARTICLE VERSIONS" section lists four previous versions and notes that the user is viewing Version 5. The copyright information at the bottom states: "Copyright: The copyright holder for this preprint is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license."

Yes



No



Yes

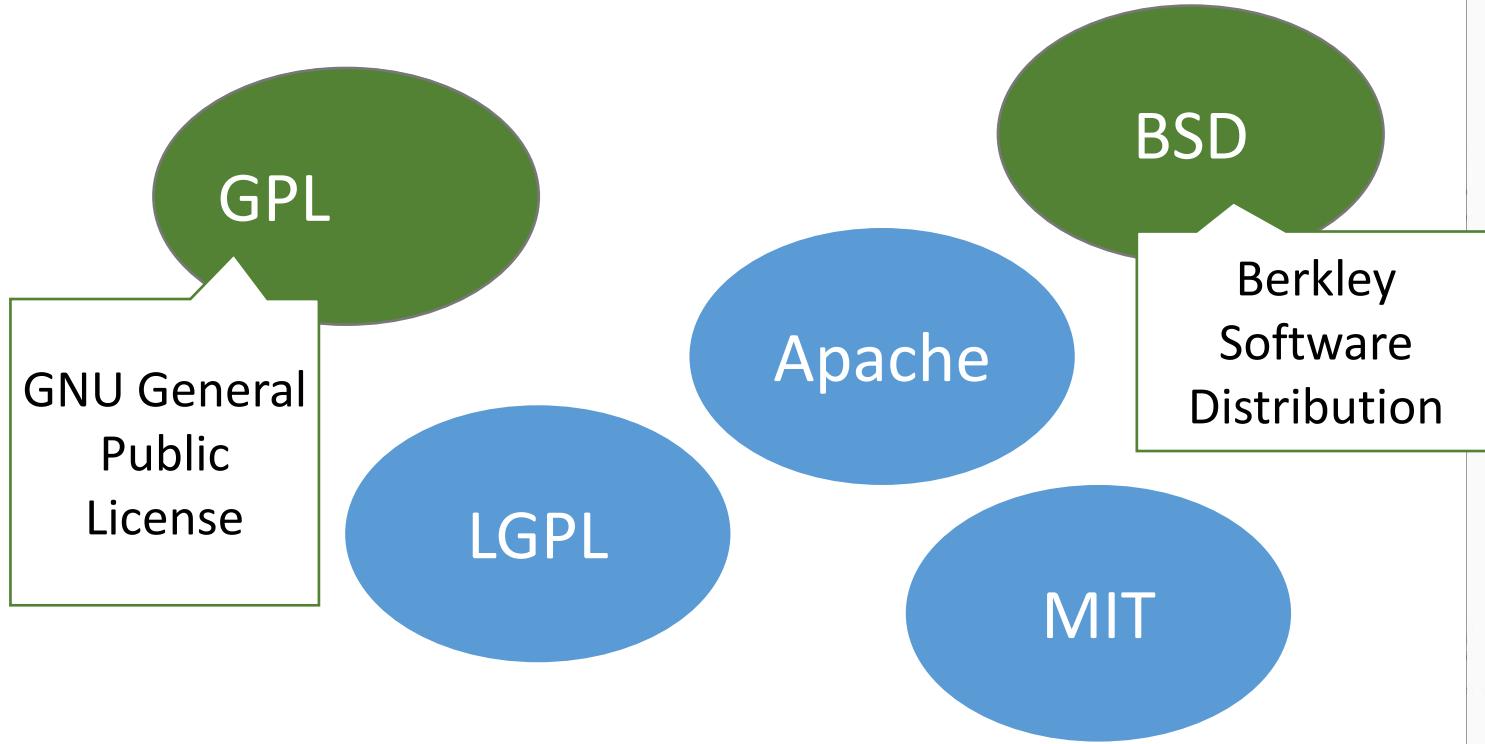


No



# Licensing Software

In the software world, other licenses are more popular, historically grown.



Choose an open source license

An open source license protects contributors and users. Businesses and savvy developers won't touch a project without this protection.

Which of the following best describes your situation?

I need to work in a community.

I want it simple and permissive.

I care about sharing improvements.

What if none of these work for me?

My project isn't software.

I want more choices.

I don't want to choose a license.

The content of this site is licensed under the Creative Commons Attribution 3.0 Unported License.

About Terms of Service Help improve this page Curated with ❤ by GitHub, Inc. and You!

# Licensing Software: GPL

GPL-derivatives must also  
be GPL-licensed

“Restrictive”  
licensing

The GNU General Public License is a free, copyleft license for software and other kinds of works.

The licenses for most software and other practical works are designed to take away your freedom to share and change the works. By contrast, the GNU General Public License is intended to guarantee your freedom to share and change all versions of a program--to make sure it remains free software for all its users. We, the Free Software Foundation, use the GNU General Public License for most of our software; it applies also to any other work released this way by its authors. You can apply it to your programs, too.

When we speak of free software, we are referring to freedom, not price. Our General Public Licenses are designed to make sure that you have the freedom to distribute copies of free software (and charge for them if you wish), that you receive source code or can get it if you want it, that you can change the software or use pieces of it in new free programs, and that you know you can do these things.

To protect your rights, we need to prevent others from denying you these rights or asking you to surrender the rights. Therefore, you have certain responsibilities if you distribute copies of the software, or if you modify it: responsibilities to respect the freedom of others.

For example, if you distribute copies of such a program, whether gratis or for a fee, you must pass on to the recipients the same freedoms that you received. You must make sure that they, too, receive or can get the source code. And you must show them these terms so they know their rights.

# Quiz

Can I build a commercial product on the basis of GPL-licensed code?

Yes



No



Do I have to release the code openly for this commercial product?

Yes



No



# Licensing Software: BSD0

Copyright (C) [year] by [copyright holder] <[email]>

Permission to use, copy, modify, and/or distribute this software for any purpose with  
or without fee is hereby granted.

} Similar to CC0  
(public domain)

THE SOFTWARE IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL  
WARRANTIES WITH REGARD TO THIS SOFTWARE INCLUDING ALL IMPLIED  
WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE  
AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR  
CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING  
FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF  
CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF  
OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE

# Licensing Software: BSD0

Copyright (C) [year] by [copyright holder] <[email]>

Permission to use, copy, modify, and/or distribute this software for any purpose with or without fee is hereby granted.

THE SOFTWARE IS PROVIDED "AS IS" AND THE AUTHOR DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY SPECIAL, DIRECT, INDIRECT, OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE

Disclaimer

Whatever you do with it, we [the authors] are not liable

# Licensing Software: BSD2

Copyright (c) <year>, <copyright holder>

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE

} Similar to  
CC-BY

# Licensing Software: BSD3

Copyright <year> <copyright holder>

Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE

} Similar to  
CC-BY

You must not use the  
copyright holder's name  
to endorse your  
derivative of the work.

# Licensing: Permissive versus restrictive

- Restrictive
  - You can reuse our stuff, but only if you ...
    - License your work with the same license we do
    - Make your stuff openly available
    - Make no money with derivatives of our work
  - Examples: **GPL, CC-BY-SA, CC-BY-NC, CC-BY-ND**
- Permissive licensing:
  - Do whatever you like with our stuff, just make sure to mention / cite us ...
  - Examples: **BSD, MIT, Apache, CC-BY**

I conclude,  
these are  
*less open* in  
a sense



# Quiz

May I reuse code  
from this repository  
in my own BSD-  
licensed work?

Yes



No



The screenshot shows the GitHub repository page for 'cnr-isti-vclab/meshlab'. The repository is public and has 150 watchers, 686 forks, and 3.3k stars. It features 124 issues, 2 pull requests, and 10,254 commits. The main branch is active, showing recent commits from alemuntoni. The repository is described as 'The open source mesh processing system' and includes links to its website ([www.meshlab.net](http://www.meshlab.net)) and various tags such as point-cloud, mesh, 3d-printing, 3d-scanning, 3d-reconstruction, 3d-models, mesh-processing, mesh-editing, mesh-simplification, and triangle-mesh.

Commit	Message	Date
alemuntoni bugfix while loading exif info	fix missing windeployqt in windows workflows	5 months ago
.github	Fix various typos	8 months ago
docs	icon on windows folder	5 months ago
resources	add gltf samples	16 months ago
sample	deploy all plugins in macos script	4 months ago
scripts	bugfix while loading exif info	22 days ago
src	moved textures folder outside distrib	22 days ago
textures	remove "vertex color noise" filter	3 years ago
unsupported	move build and install dirs on macos outside src	12 months ago
.aitianore	move build and install dirs on macos outside src	5 months ago

# Quiz

May I reuse code  
from this repository  
in my own GPL-  
licensed work?

Yes



No



A screenshot of a web browser displaying the GitHub repository for "napari/napari". The repository page shows a list of recent commits and pull requests. The commits listed are:

- Czaki set selection color for QListView item. (#5202) - 8196109 15 hours ago (2,552 commits)
- .devcontainer feat: add codespace (#4599) - 4 months ago
- .github Move docs to separate repo (#5216) - 5 days ago
- binder Drop python 3.7 (#4063) - 8 months ago
- examples Move docs to separate repo (#5216) - 5 days ago
- napari set selection color for QListView item. (#5202) - 15 hours ago
- napari\_builtins Split out builtins into another top-level module (#4706) - 3 months ago
- resources Re-add README screenshot (#5220) - 4 days ago
- tools Update some strings to be translated, some to be igno... - last month
- .env\_sample Add event.debugging tool (#3802) - 10 months ago

The repository page also includes sections for "About", "Readme", "BSD-3-Clause license", "Cite this repository", "1.5k stars", "45 watching", and "328 forks".



DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS  
AND ARTIFICIAL INTELLIGENCE

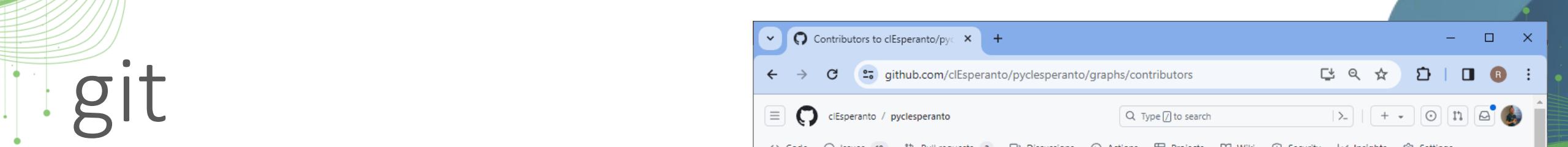
# Software environments

Robert Haase

GEFÖRDERT VOM

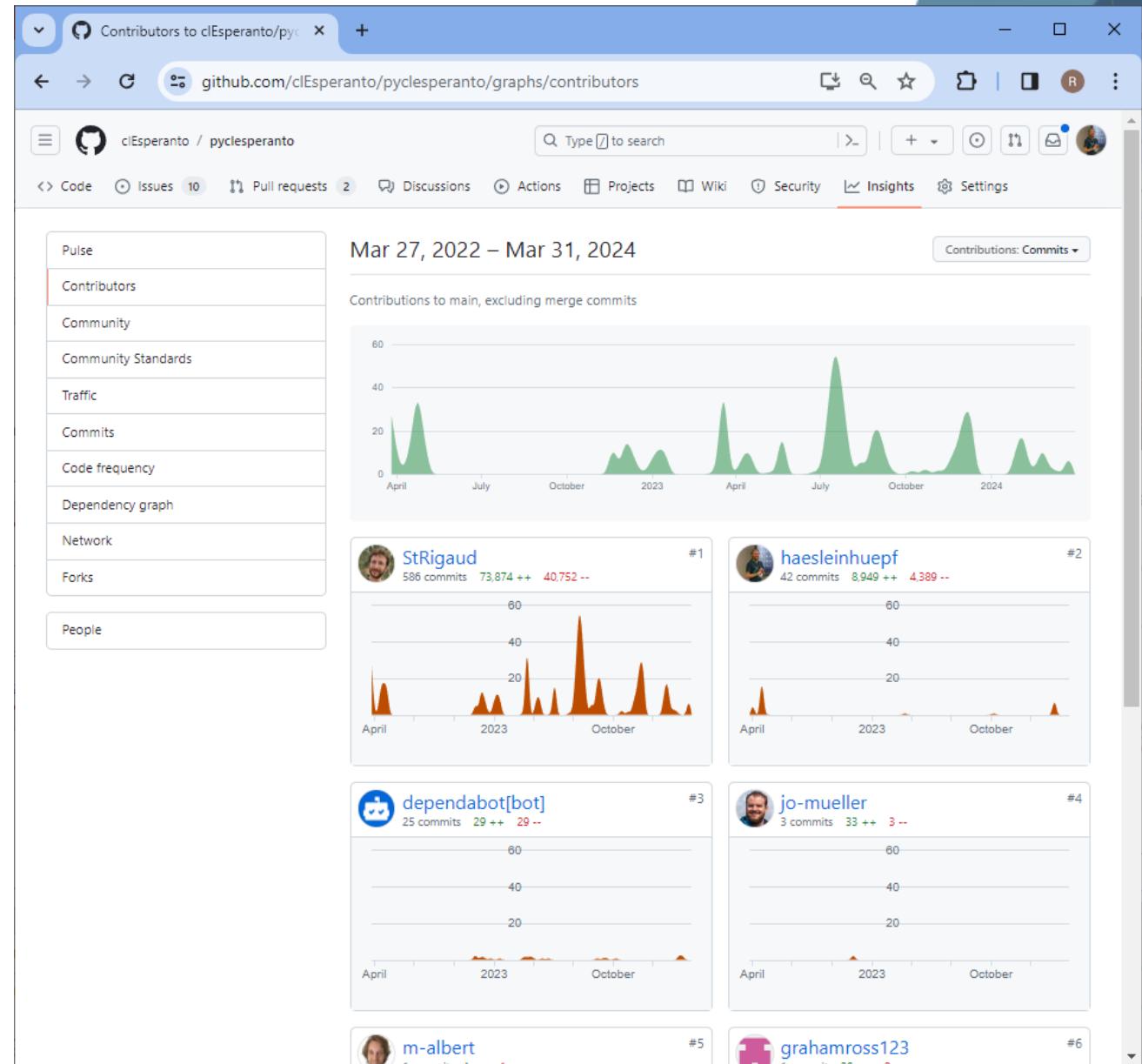


Bundesministerium  
für Bildung  
und Forschung



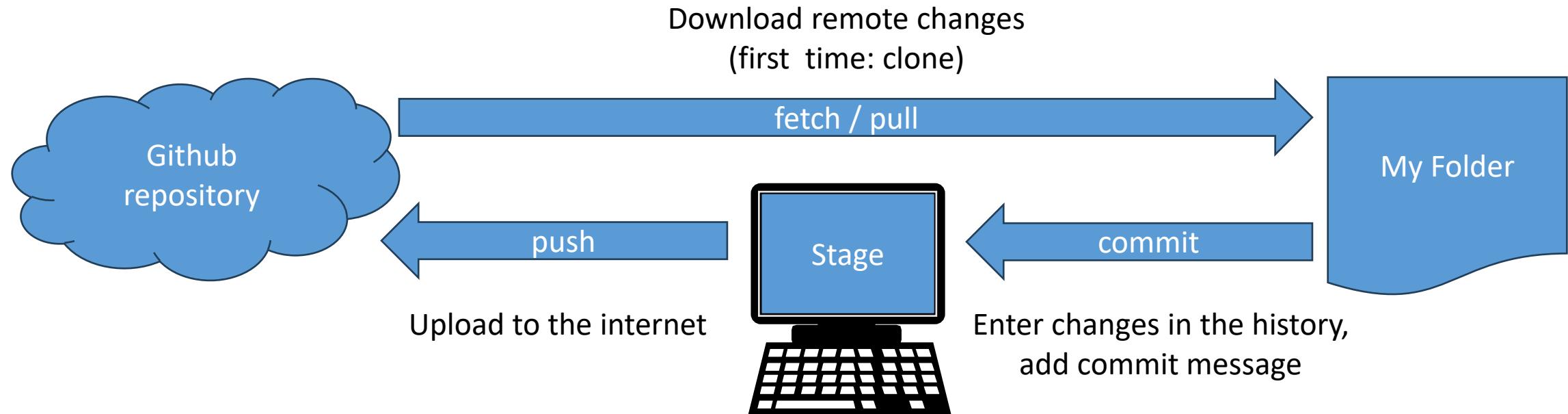
# git

- Version control is key element of data scientist's toolbox
- Distributed file system with sophisticated logging mechanisms
- Control about what becomes part of a repository and what not



# git

- Git makes file modifications a more active / involved process (making people think about)



# git

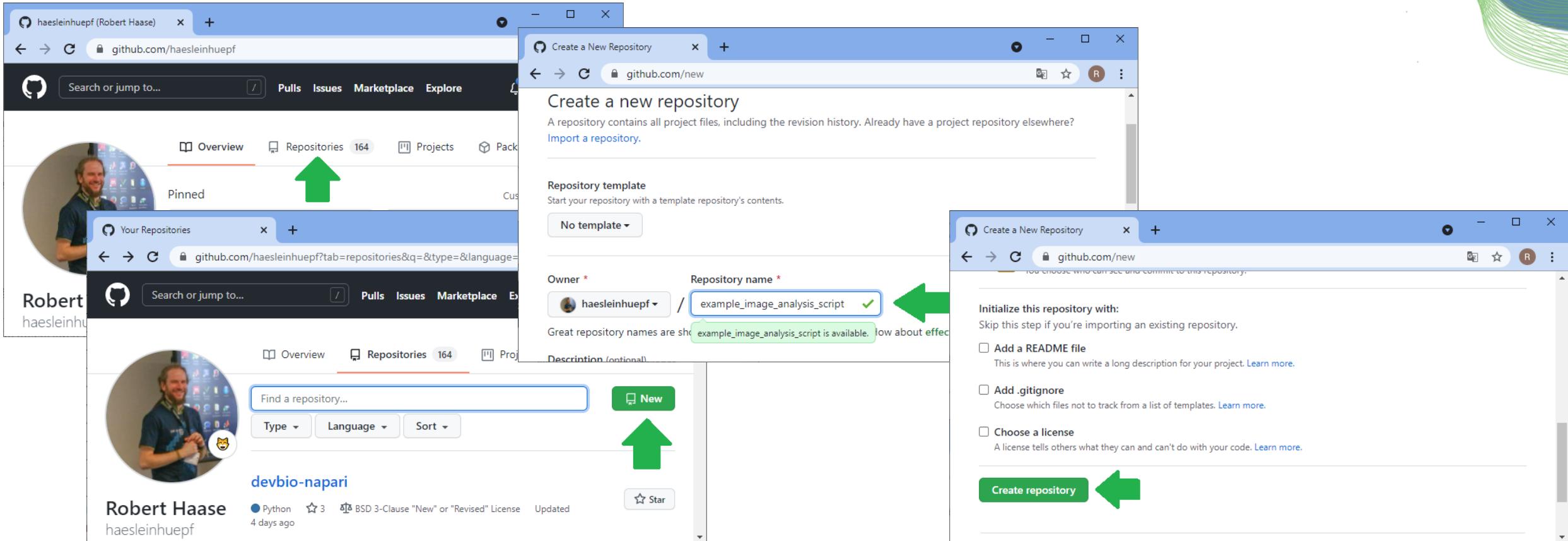
- Who wrote this code
- when and
- why?

The screenshot shows two browser tabs. The left tab displays the commit history for the repository `haesleinhuepf/example_image_analysis_script`. It lists several commits, with the commit titled "bugfix: threshold\_otsu" highlighted by an orange arrow. The right tab shows a detailed view of the changes made in that commit, specifically the file `my_library.py`. The diff view highlights the addition of the `threshold_otsu` function and its usage.

```
diff --git a/my_library.py b/my_library.py
index 3e3f3..65c07 100644
--- a/my_library.py
+++ b/my_library.py
@@ -6,7 +6,8 @@ def segment_image(image):
 6   6     blurred = gaussian(image, sigma=2)
 7   7
 8   8     # binarize the image
 9 - 9     binary = threshold_otsu(blurred)
 9 + 9     threshold = threshold_otsu(blurred)
10 + 10    binary = blurred > threshold
11 + 11
12 + 12     # label connected components
13 + 13     result = label(binary)
```

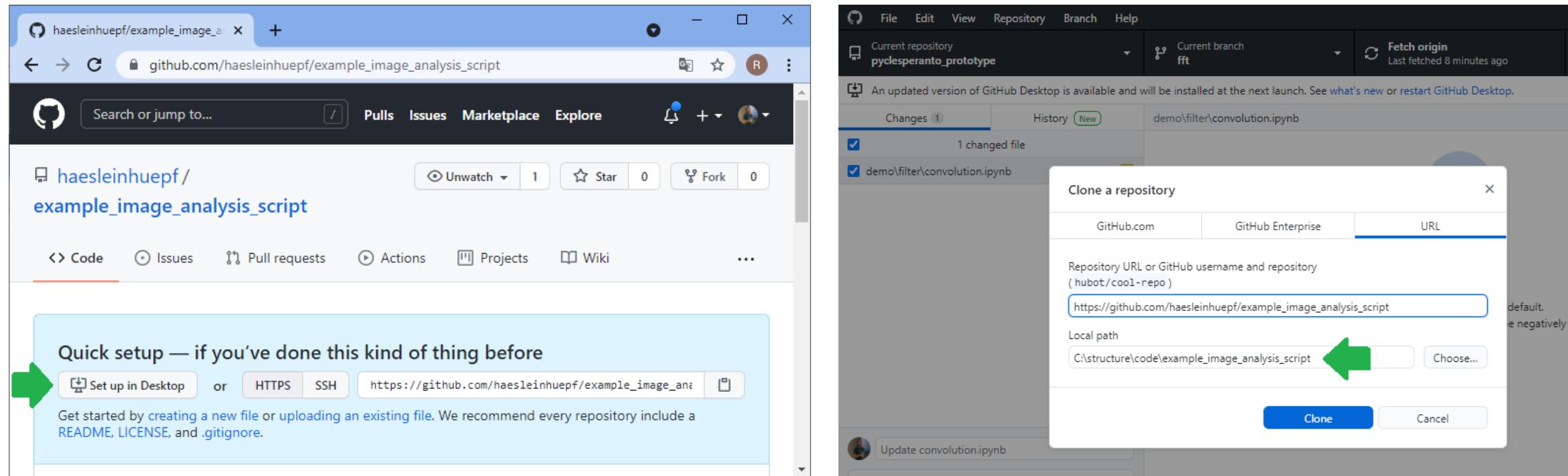
# github – creating repositories

- Add a new, empty repository



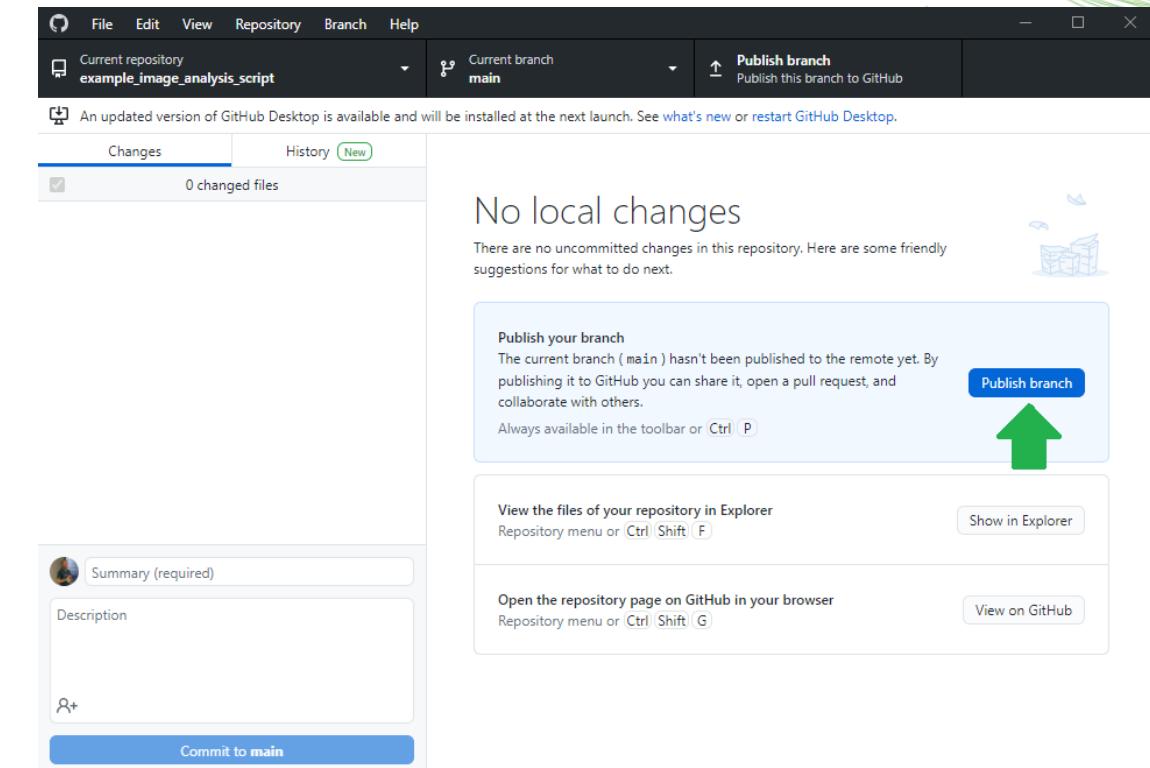
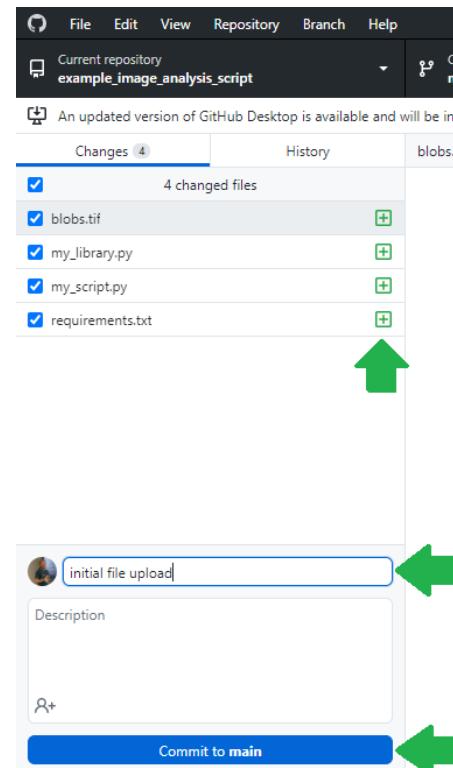
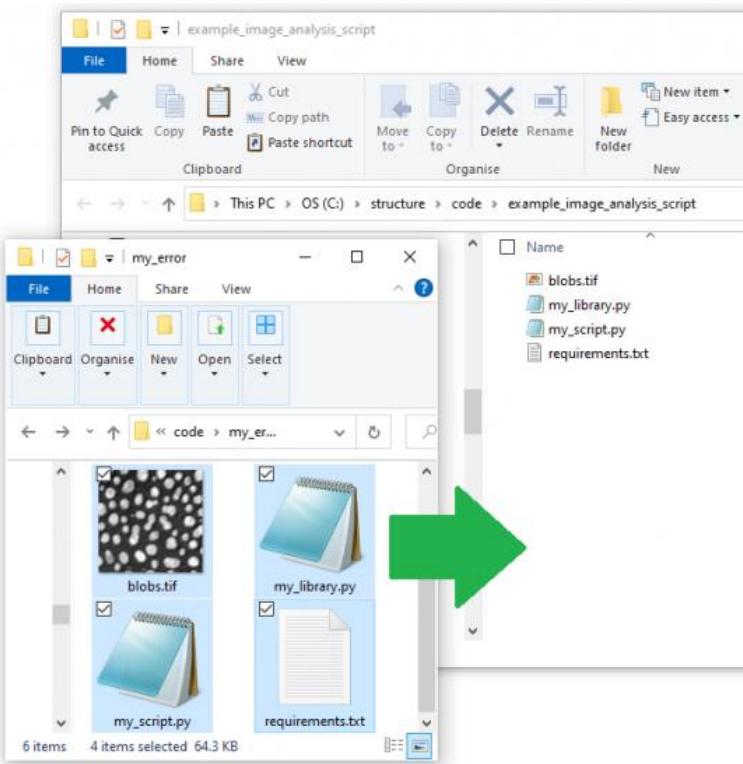
# github – clone repositories

- git clone <https://github.com/organization/repository>
- Or: Use the Github Desktop app



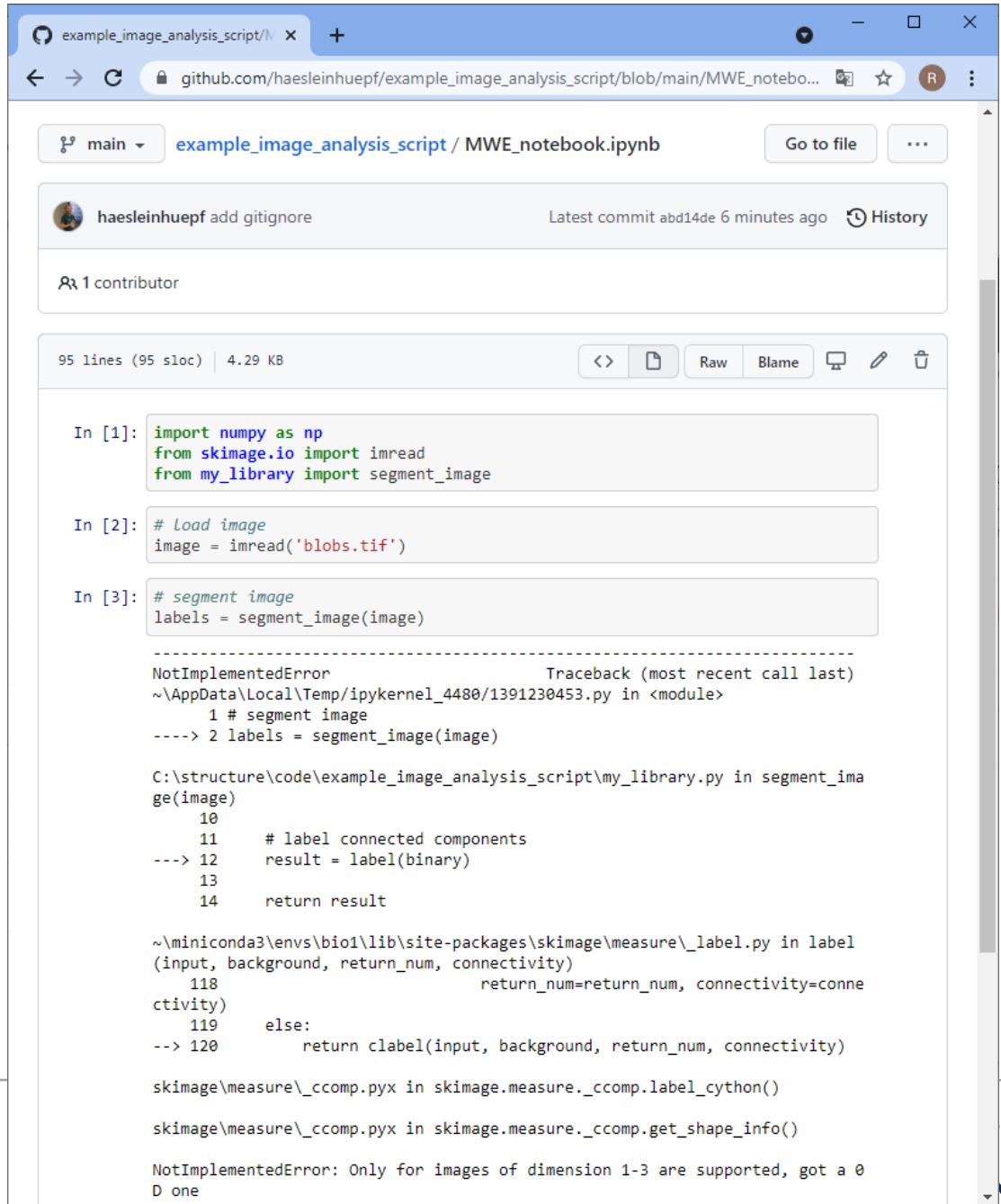
# github - uploading

- git [add], commit, push



# github

- Ease of reading notebooks online
- No need to download and execute code



The screenshot shows a GitHub browser window displaying a Jupyter notebook titled "example\_image\_analysis\_script / MWE\_notebook.ipynb". The notebook has 95 lines (95 sloc) and a size of 4.29 KB. It contains three code cells:

```
In [1]: import numpy as np
from skimage.io import imread
from my_library import segment_image

In [2]: # Load image
image = imread('blobs.tif')

In [3]: # segment image
labels = segment_image(image)

NotImplementedError Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_4480\1391230453.py in <module>
      1 # segment image
----> 2 labels = segment_image(image)

C:\structure\code\example_image_analysis_script\my_library.py in segment_image(image)
      10
      11     # label connected components
----> 12     result = label(binary)
      13
      14     return result

~\miniconda3\envs\bio1\lib\site-packages\skimage\measure\_label.py in label(input, background, return_num, connectivity)
      118                                         return_num=return_num, connectivity=connectivity)
      119     else:
--> 120         return clabel(input, background, return_num, connectivity)

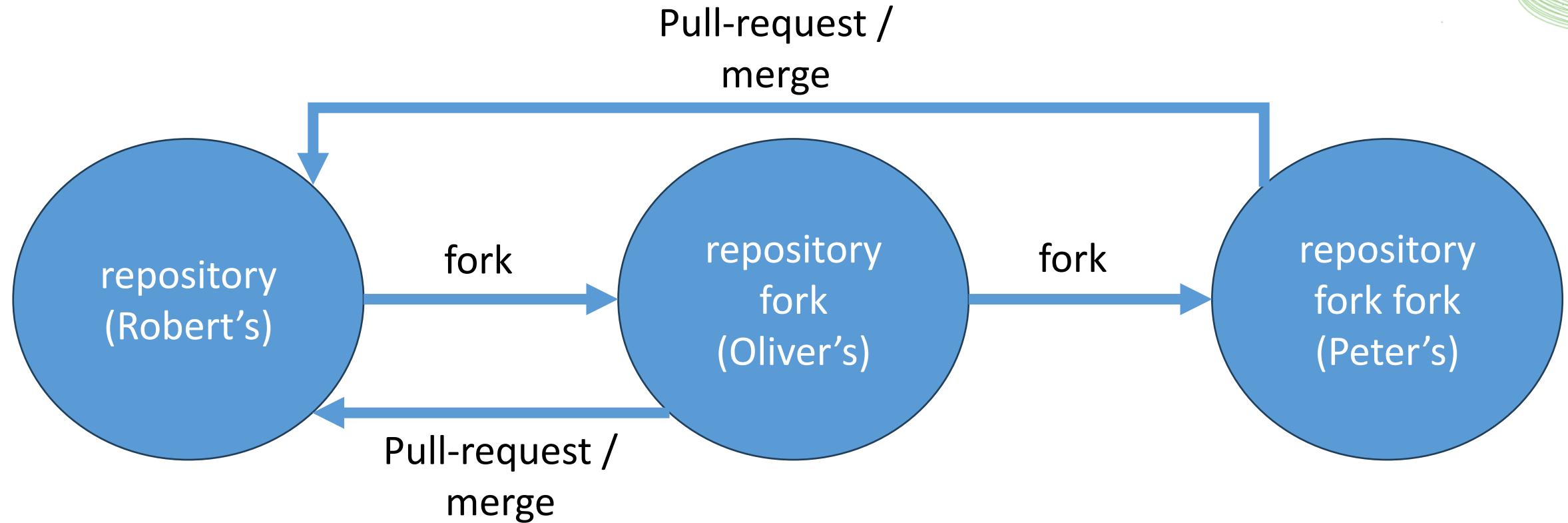
skimage\measure\_ccomp.pyx in skimage.measure._ccomp.label_cython()

skimage\measure\_ccomp.pyx in skimage.measure._ccomp.get_shape_info()

NotImplementedError: Only for images of dimension 1-3 are supported, got a 0D one
```

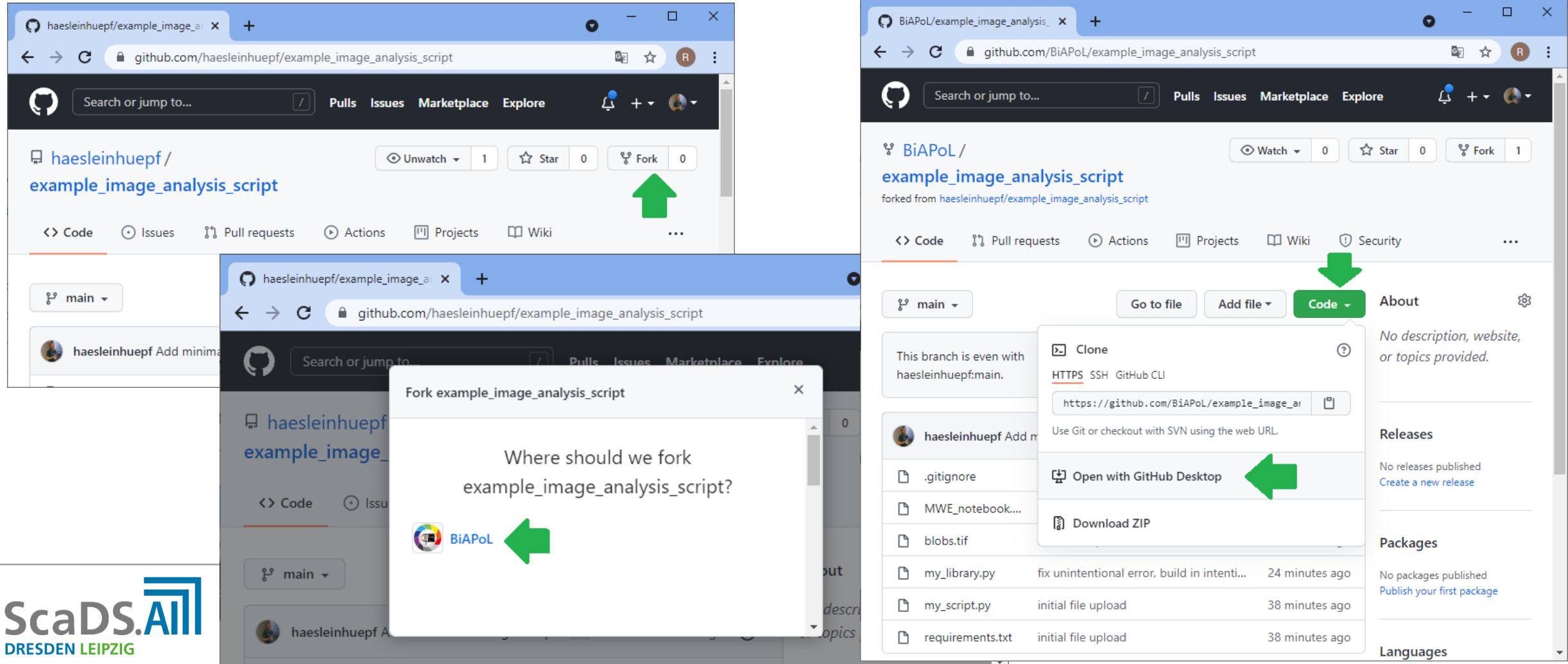
# git - forking

- Making a copy where we have edit rights



# github - forking

- Making a copy where we have edit rights



# github – uploading (again)

- After fixing a bug, we upload the changes to our fork

The screenshot shows the GitHub Desktop application interface. In the center, there is a code editor window displaying a diff for a file named `my_library.py`. The diff highlights a bug fix where the line `binary = threshold_otsu(blurred)` was replaced by `threshold = threshold_otsu(blurred)` and `binary = blurred > threshold`. A large green arrow points from the commit message area down to the `Commit to main` button at the bottom. In the bottom-left corner, a tooltip provides context about the bug fix.

Current repository: example\_image\_analysis\_script  
Current branch: main  
Fetch origin: Never fetched

An updated version of GitHub Desktop is available and will be installed at the next launch. See what's new or restart GitHub Desktop.

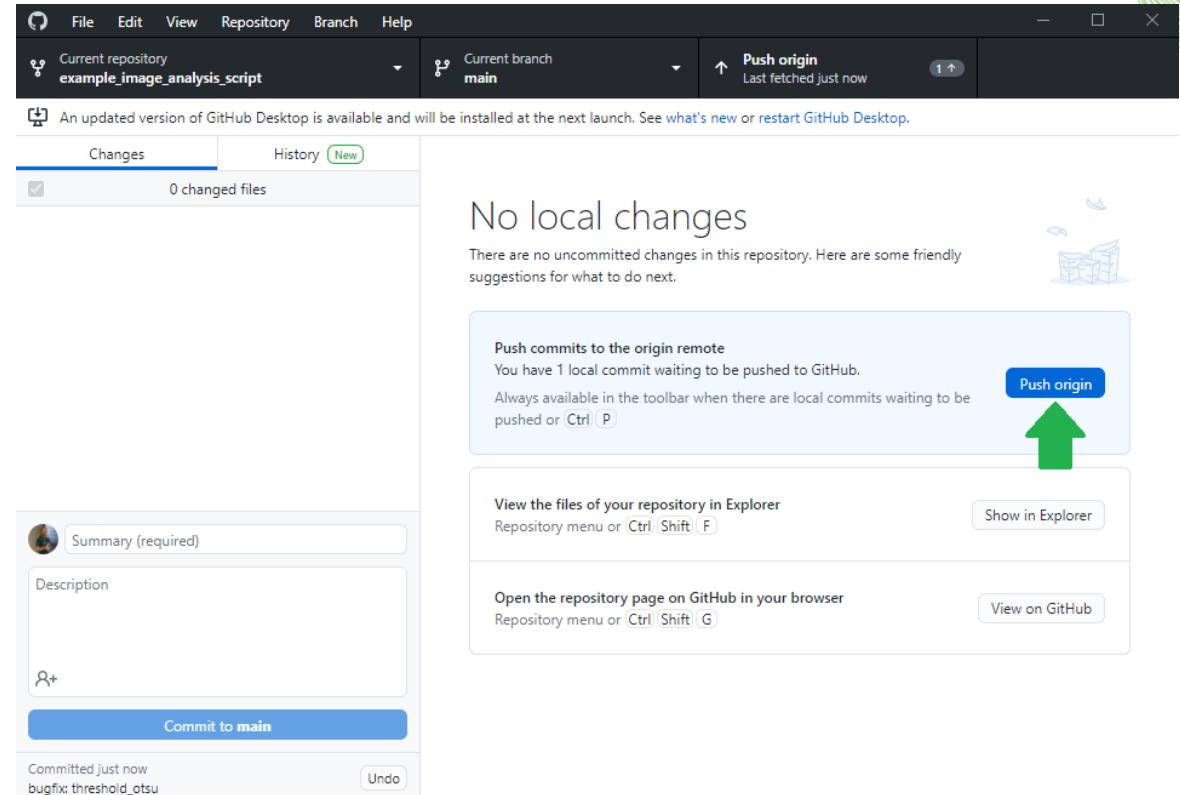
Changes (2) History (New)

my\_library.py

```
@@ -6,7 +6,8 @@ def segment_image(image):
    blurred = gaussian(image, sigma=2)
    # binarize the image
-   binary = threshold_otsu(blurred)
+   threshold = threshold_otsu(blurred)
+   binary = blurred > threshold
    # label connected components
    result = label(binary)
```

bugfix: threshold\_otsu  
threshold\_otsu delivers a number (the threshold), not a binary image. For thresholding the image, an additional step is necessary.

R+ Commit to main



# Github – pull requests

- Contribute to open-source projects

The image consists of two side-by-side screenshots of GitHub repositories.

**Left Screenshot (BiAPoL repository):**

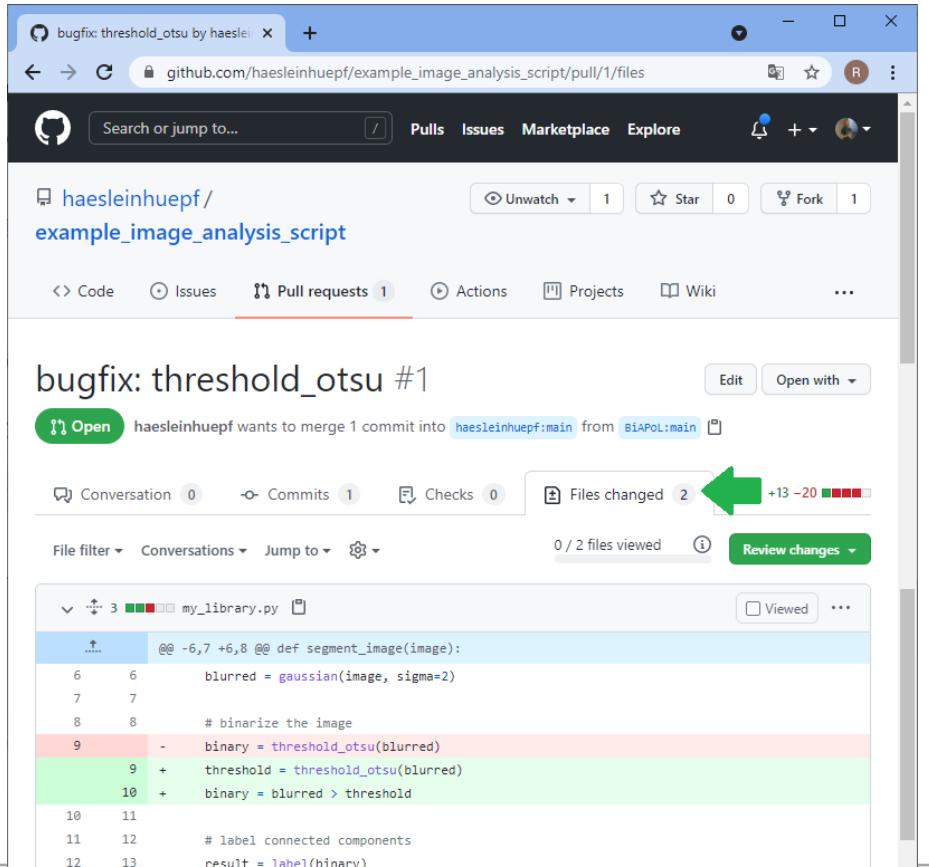
- The URL is [github.com/BiAPoL/example\\_image\\_analysis\\_script](https://github.com/BiAPoL/example_image_analysis_script).
- The repository name is **BiAPoL / example\_image\_analysis\_script**, forked from [haesleinhuepf/example\\_image\\_analysis\\_script](#).
- The main branch is **main**.
- A green arrow points to the **Open pull request** button at the bottom of the commit list.

**Right Screenshot (haesleinhuepf repository):**

- The URL is [github.com/haesleinhuepf/example\\_image\\_analysis\\_script/compare/main...BiAPoL%2Fmain](https://github.com/haesleinhuepf/example_image_analysis_script/compare/main...BiAPoL%2Fmain).
- The repository name is **haesleinhuepf / example\_image\_analysis\_script**.
- The base repository is **base repository: haesleinhuepf/example\_image...** and the head repository is **head repository: BiAPoL/example\_image\_analys...**.
- A green arrow points to the **Create pull request** button at the bottom of the pull request creation form.

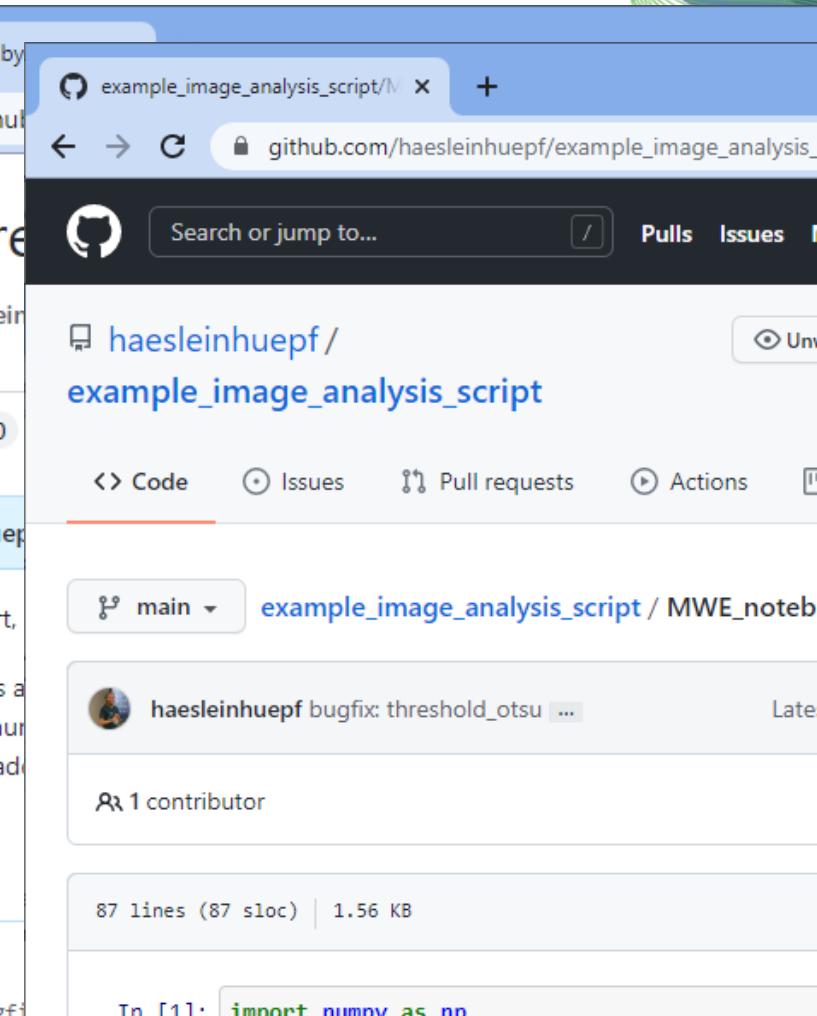
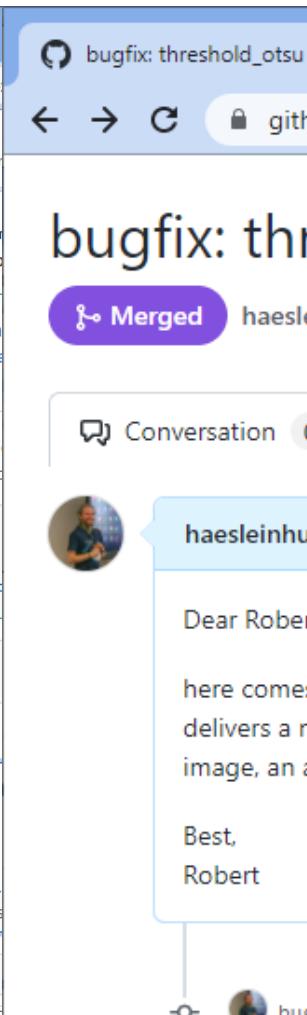
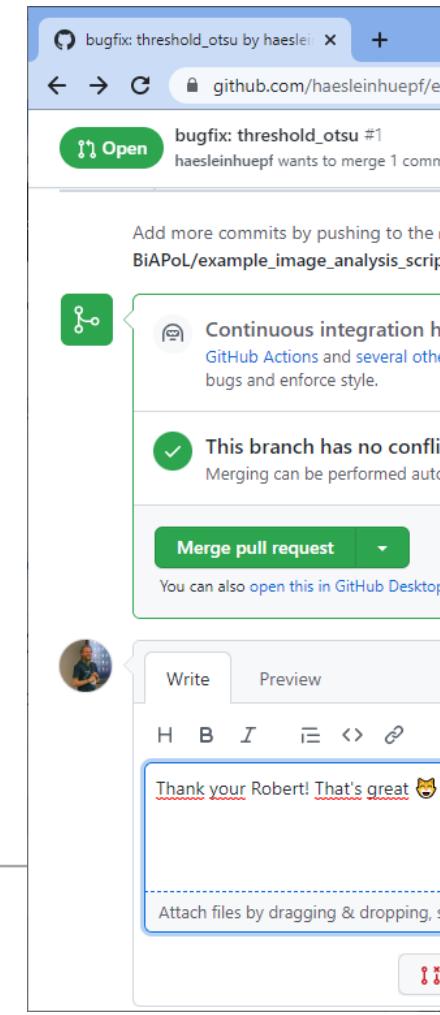
# Github – pull requests

- Reviewer perspective



A screenshot of a GitHub pull request page. The title is "bugfix: threshold\_otsu #1". It shows a diff of the file "my\_library.py". The diff highlights changes from line 9 to line 10. The commit message is "haesleinhuepf wants to merge 1 commit into haesleinhuepf:main from BiAPoL:main". There are 2 files changed, +13 -20 lines of code, and 2 commits.

```
diff --git a/my_library.py b/my_library.py
--- a/my_library.py
+++ b/my_library.py
@@ -6,7 +6,8 @@ def segment_image(image):
 6   6     blurred = gaussian(image, sigma=2)
 7   7
 8   8     # binarize the image
 9 - 9     binary = threshold_otsu(blurred)
 10 + 9     threshold = threshold_otsu(blurred)
 11 + 10    binary = blurred > threshold
 12   11
 13   12     # label connected components
 14     result = label(binary)
```



A screenshot of a Jupyter Notebook cell. The code `import numpy as np` has been run, and the output shows the version of numpy: "In [1]: import numpy as np" and "Out[1]: numpy 1.20.1".

# Github – pull requests

- Reviewer perspective

The screenshot shows a GitHub pull request page for a repository named 'example\_image\_analysis\_script'. The pull request is titled 'bugfix: threshold\_otsu #1' and has been merged. The commit message is 'bugfix: threshold\_otsu ...' with a SHA of '65c074a'. A review comment from 'haesleinhuepf' says: 'Dear Robert, here comes a bug fix for your image segmentation function. threshold\_otsu delivers a number (the threshold), not a binary image. For thresholding the image, an additional step is necessary.' Below this, another comment from 'haesleinhuepf' says: 'Thank you Robert! That's great 😊'. The right sidebar shows the repository details: 1 star, 1 fork, and 1 pull request.

The screenshot shows a GitHub repository page for 'haesleinhuepf/example\_image\_analysis\_script'. The repository has 1 star, 0 forks, and 1 pull request. A Jupyter notebook named 'MWE\_notebook.ipynb' is displayed. The code in the notebook is as follows:

```
In [1]: import numpy as np
from skimage.io import imread
from my_library import segment_image

In [2]: # Load image
image = imread('blobs.tif')

In [3]: # segment image
labels = segment_image(image)

In [4]: # count objects
number_of_objects = labels.max()
print('Number of objects', number_of_objects)
```

The output of the last cell is 'Number of objects 61'. To the right of the notebook, there is a large green box containing the text 'Problem solved :-)'.

# Github

- If this was too fast...

The screenshot shows a web browser window displaying a blog post from the FocalPlane website. The header features the FocalPlane logo and the text "Where biology meets microscopy". Below the header is a navigation bar with links to Home, About us, Topics, Gallery, Jobs, Events, Resources, Network, Contact us, and Log in/register. A search icon is also present. The main content area shows a breadcrumb trail: Home / How to / Collaborative bio-image analysis script ... The title of the post is "Collaborative bio-image analysis script editing with git". It was posted by Robert Haase on 4 September 2021. The post begins with a TL;DR summary: "I'm a computer scientist who often collaborates with biologists on bio-image analysis scripts. We are using more and more git, a version control program, for working on code collaboratively. When using git, we speak about repositories, commits and pushing to the origin. We also make forks, send pull-requests and merge code. This blog post explains these terms and demonstrates how a typical collaborative bio-image analysis scripting project looks like." The text continues with a personal anecdote about writing a script that counts cells and needing help from experts.

The screenshot shows a web browser window displaying a GitHub profile page for the user "haesleinhuepf". The profile page includes a bio image of a smiling man, a pinned item, and sections for Overview, Repositories (164), Projects, and Packages. A green arrow points upwards towards the "Repositories" tab, indicating where to click to demonstrate how to share code on GitHub.

# Quiz

It's ok to reuse this code if ...

haesleinhuepf / **imagej-run-async** Public

**Code** Issues Pull requests Actions Projects Wiki Security Insights Settings

master 1 branch 0 tags Go to file Add file Code

haesleinhuepf initial version 2f8c334 on 23 Jun 2019 1 commit

src/main/java/net/haeslein... initial version 3 years ago

.gitignore initial version 3 years ago

pom.xml initial version 3 years ago

Help people interested in this repository understand your project by adding a README. Add a README

About No description, website, or topics provided.

0 stars 1 watching 0 forks

Releases No releases published Create a new release



Mention author



Ask the authors



Link to the license



Copy the copyright statement



DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS  
AND ARTIFICIAL INTELLIGENCE

# Exercises

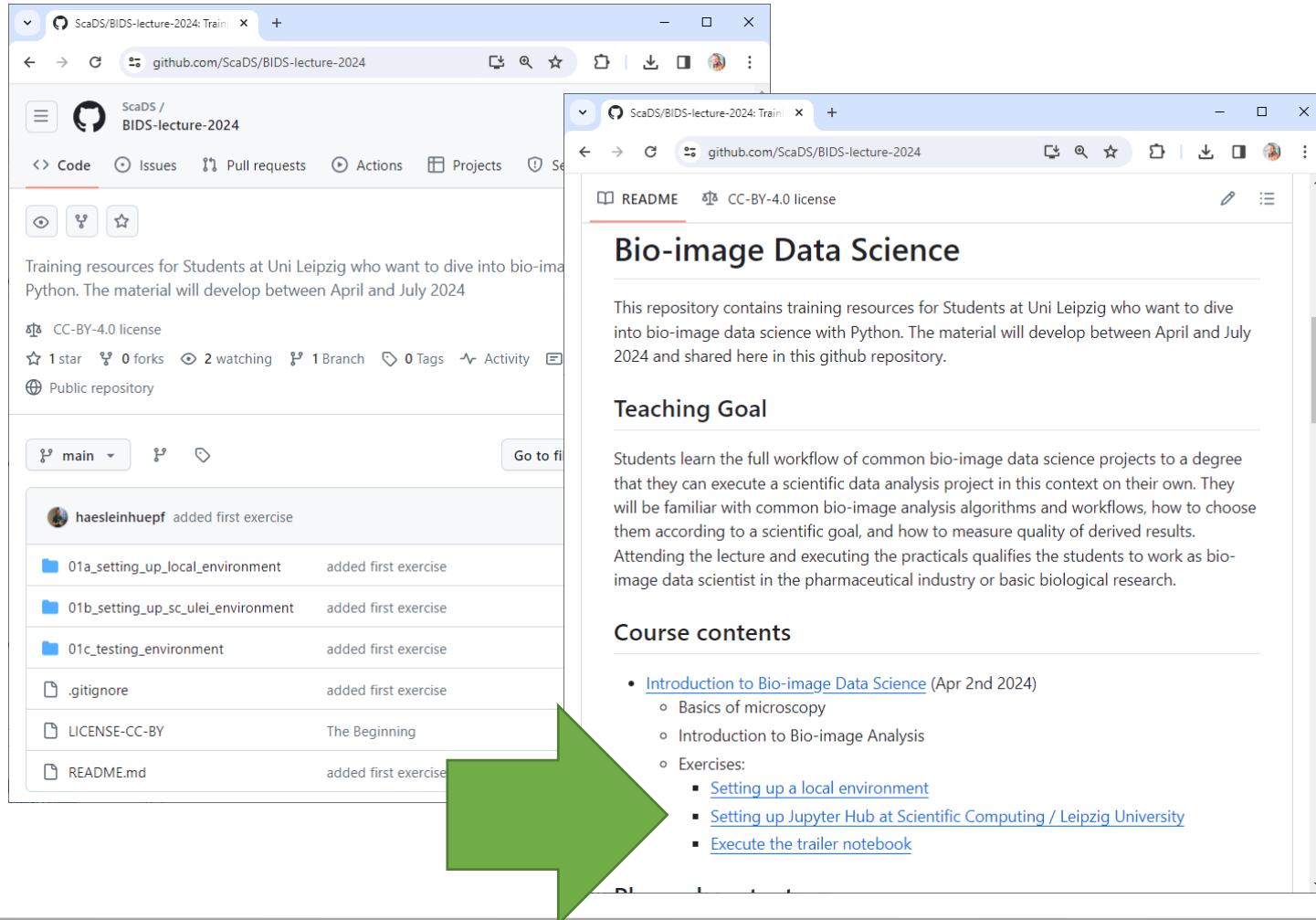
Robert Haase

GEFÖRDERT VOM



Bundesministerium  
für Bildung  
und Forschung

# Exercises



Training resources for Students at Uni Leipzig who want to dive into bio-image data science with Python. The material will develop between April and July 2024

CC-BY-4.0 license  
1 star 0 forks 2 watching 1 Branch 0 Tags Activity Public repository

main haesleinhuepf added first exercise  
01a\_setting\_up\_local\_environment added first exercise  
01b\_setting\_up\_sc\_ulei\_environment added first exercise  
01c\_testing\_environment added first exercise  
.gitignore added first exercise  
LICENSE-CC-BY The Beginning  
README.md added first exercise

## Bio-image Data Science

This repository contains training resources for Students at Uni Leipzig who want to dive into bio-image data science with Python. The material will develop between April and July 2024 and shared here in this github repository.

### Teaching Goal

Students learn the full workflow of common bio-image data science projects to a degree that they can execute a scientific data analysis project in this context on their own. They will be familiar with common bio-image analysis algorithms and workflows, how to choose them according to a scientific goal, and how to measure quality of derived results. Attending the lecture and executing the practicals qualifies the students to work as bio-image data scientist in the pharmaceutical industry or basic biological research.

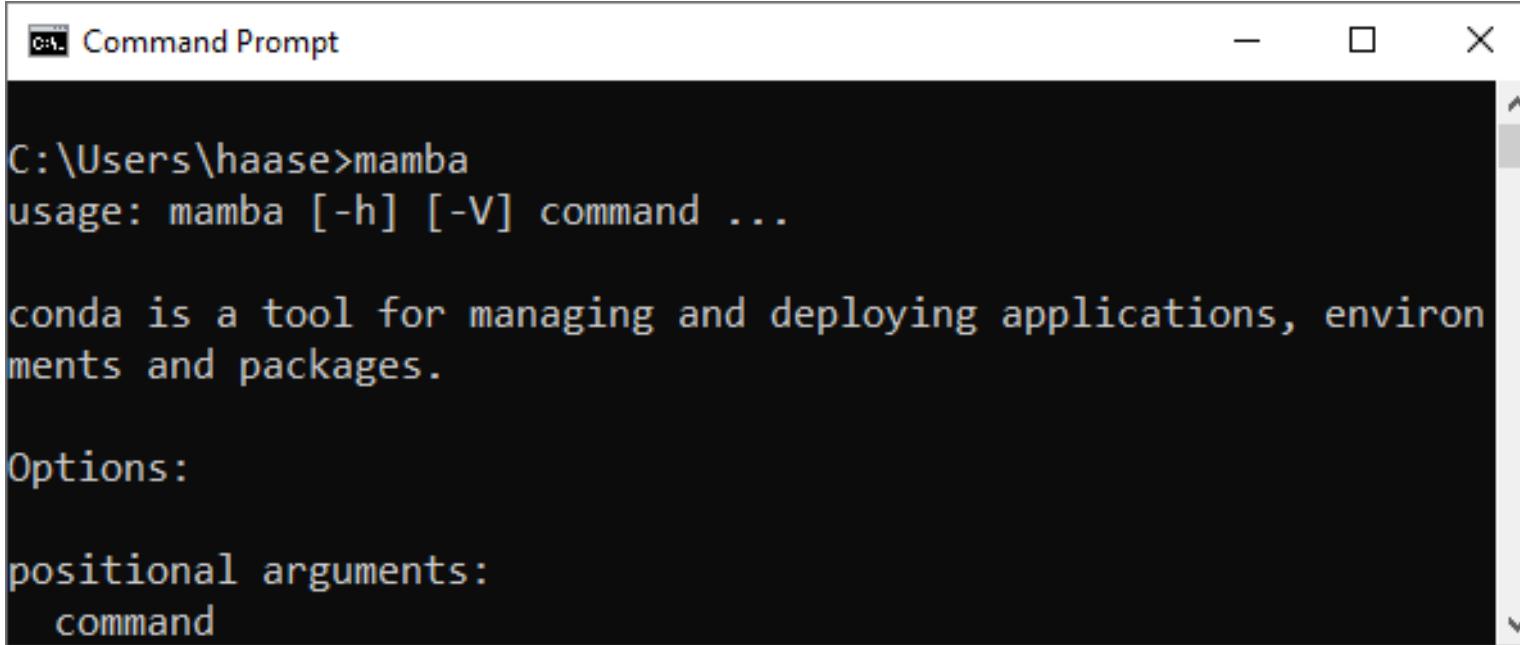
### Course contents

- Introduction to Bio-image Data Science (Apr 2nd 2024)
  - Basics of microscopy
  - Introduction to Bio-image Analysis
  - Exercises:
    - Setting up a local environment
    - Setting up Jupyter Hub at Scientific Computing / Leipzig University
    - Execute the trailer notebook



# Exercise (recap)

- Make sure mamba is installed on your computer  
(see instructions from last week)



The screenshot shows a Windows Command Prompt window titled "Command Prompt". The window contains the following text:

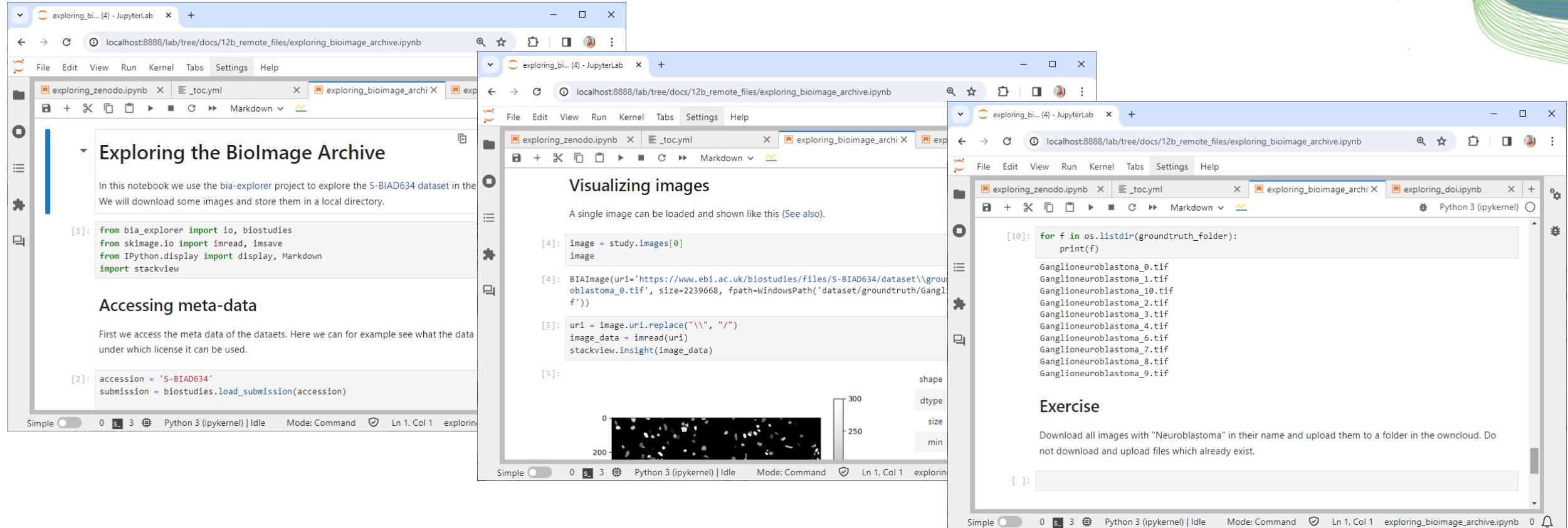
```
C:\Users\haase>mamba
usage: mamba [-h] [-V] command ...
conda is a tool for managing and deploying applications, environments and packages.

Options:

positional arguments:
    command
```

# Exercise (BioImage Archive)

- Download a dataset from the BioImage Archive



The image shows three side-by-side JupyterLab interfaces, each displaying a different notebook related to exploring the BioImage Archive.

- Left Notebook:** Titled "Exploring the BioImage Archive". It contains code to import necessary libraries and download the S-BIAD634 dataset. It also includes a section on "Accessing meta-data".
- Middle Notebook:** Titled "Visualizing images". It shows a single image of a tissue sample with white spots against a black background. Below the image, there are several numerical parameters: shape [300, 200], dtype <class 'numpy.uint8'>, size 60000, min 0, and max 255.
- Right Notebook:** Titled "Exercise". It lists file names for Ganglioneuroblastoma images: Ganglioneuroblastoma\_0.tif through Ganglioneuroblastoma\_9.tif. It also contains a code cell for listing files in a groundtruth folder.

# Exercise (nextcloud)

- Register at Speicherwolke @ Uni Leipzig,
- Upload the images from the BioImage Archive to a folder in the Speicherwolke.

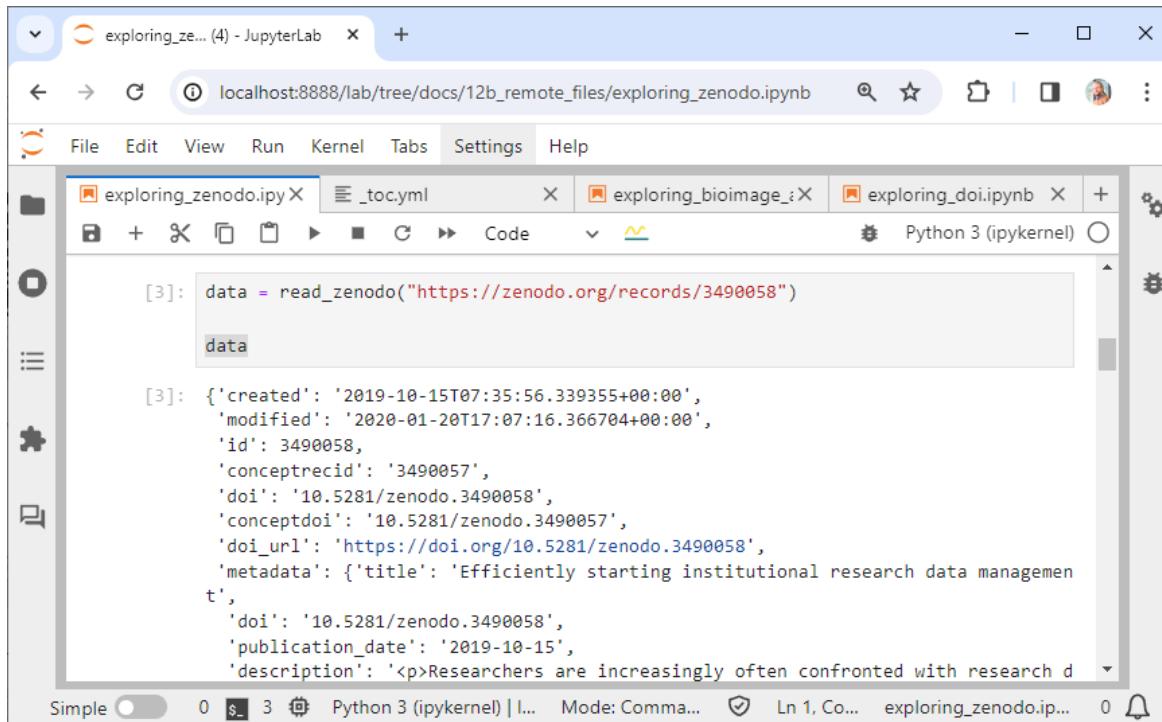
The screenshot shows the Universit t Leipzig website with a blue header. Below it, there's a section for 'EIGENER CLOUD-SPEICHER (SPEICHERWOLKE)'. It includes a brief description of the service, mentioning that users can store files and synchronize them with various devices. A note at the bottom states that currently, they are using the 'Speicherwolke' application.

The screenshot shows the BioImage Archive website with a banner for the 'S-BIAD634' dataset. The dataset is described as an annotated fluorescence image dataset for training nuclear segmentation methods. It was released on 2023-03-07 by a team of researchers. Below the description, there are sections for 'On this page', 'Study Information', 'Annotations', 'Images', and 'Models used'. There are also thumbnail images of the dataset.

The screenshot shows the Speicherwolke NextCloud interface. The left sidebar lists categories like 'Recent', 'Favorites', 'Shares', 'Group folders', 'Shared to Circles', 'Deleted files', and 'Files settings'. The main area shows a list of files and folders. At the top, it says '36 minutes (1/30)'. The list includes 'groundtruth' (0 KB, a minute ago), 'images' (0 KB, a minute ago), 'blobs.tif' (23 KB, a year ago), and 'blobs\_labels.tif' (254 KB, 22 minutes ago). A summary at the bottom indicates 2 folders and 2 files with a total size of 277 KB.

# Exercise (Zenodo and DOI)

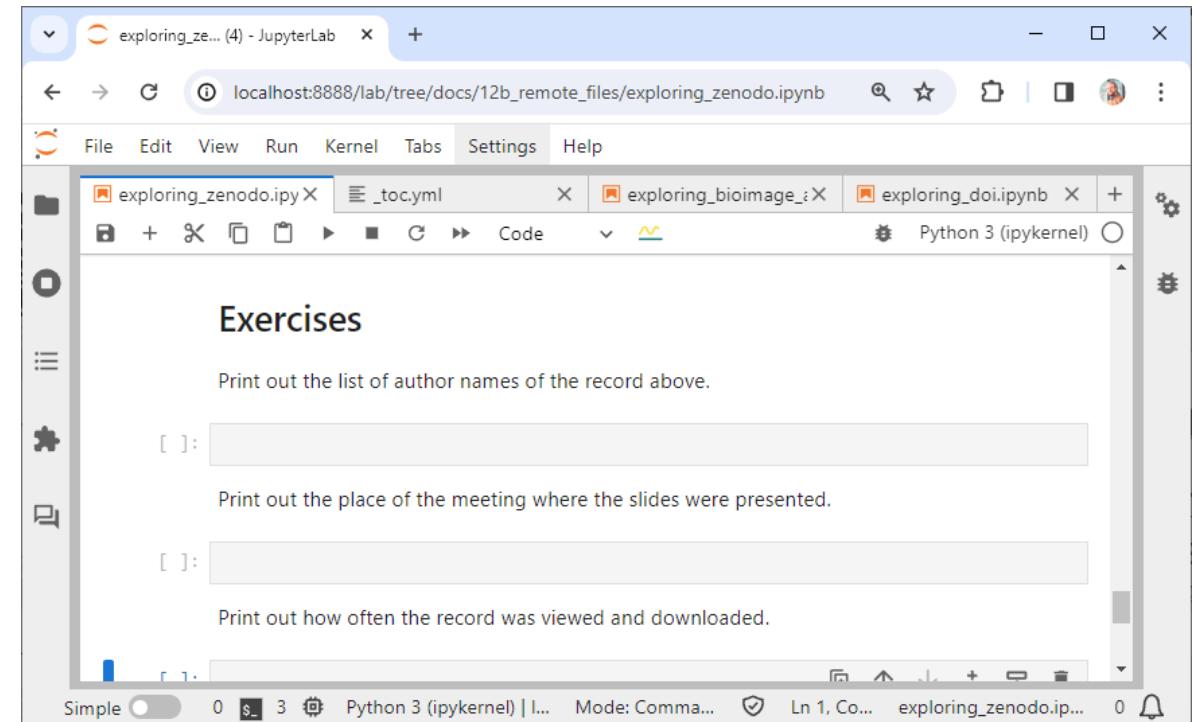
- Explore the DOI and Zenodo APIs to find out the author of online records



The screenshot shows a JupyterLab interface with two tabs open: "exploring\_zenodo.ipynb" and "exploring\_doi.ipynb". The "exploring\_zenodo.ipynb" tab contains the following code:

```
[3]: data = read_zendodo("https://zenodo.org/records/3490058")
data
```

The output cell [3] displays the JSON data of the Zenodo record, which includes fields like 'created', 'modified', 'id', 'doi', and 'description'.



The screenshot shows a JupyterLab interface with the "exploring\_zenodo.ipynb" tab selected. It displays the following exercises:

### Exercises

Print out the list of author names of the record above.

```
[ ]:
```

Print out the place of the meeting where the slides were presented.

```
[ ]:
```

Print out how often the record was viewed and downloaded.

```
[ ]:
```

# Exercise

- Clone the training materials repository
- Fix the typo on this page, send a pull-request

