National Czech Programme | eosc

# Metadata, Incentives, Formats, and Accessibility guidelines for AI datasets for biological image analysis

Vladimír Ulman

May 22, 2025

Co-funded by the European Union

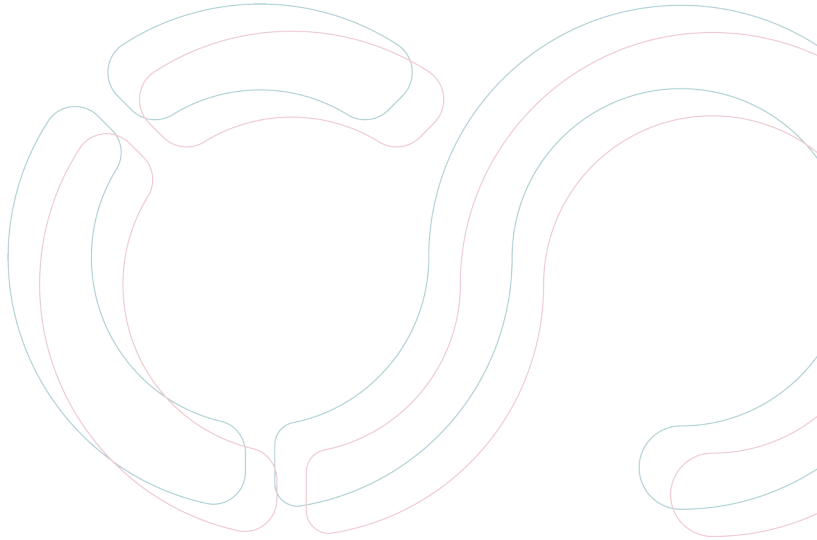
MINISTRY OF EDUCATION, YOUTH AND SPORTS

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

# MIFA: Metadata, Incentives, Formats, Accessibility

*„ ...recommendations will accelerate the development of AI tools for bioimage analysis by facilitating access to high-quality training and benchmarking data. “*

[Teresa Zulueta-Coarasa et al., 2025]



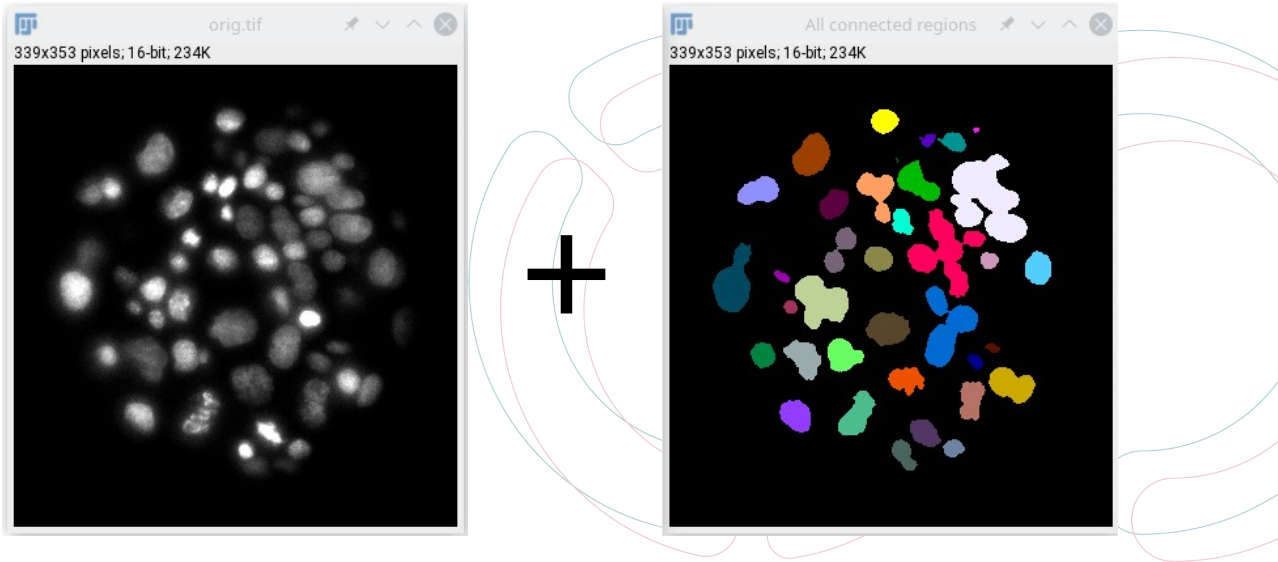
# MIFA: Metadata, Incentives, Formats, Accessibility

*„ ...recommendations will accelerate the development of AI tools for bioimage analysis by facilitating access to high-quality training and benchmarking data. “*

[Teresa Zulueta-Coarasa et al., 2025]

Training and Benchmarking data

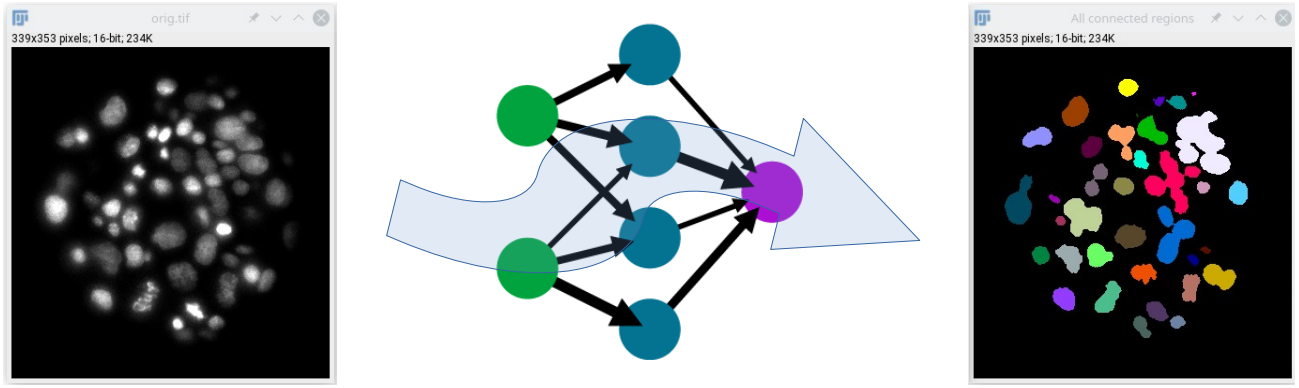
- No difference technically
- Pairs of:
  - Original (raw) image
  - Expected result --> **Annotation**



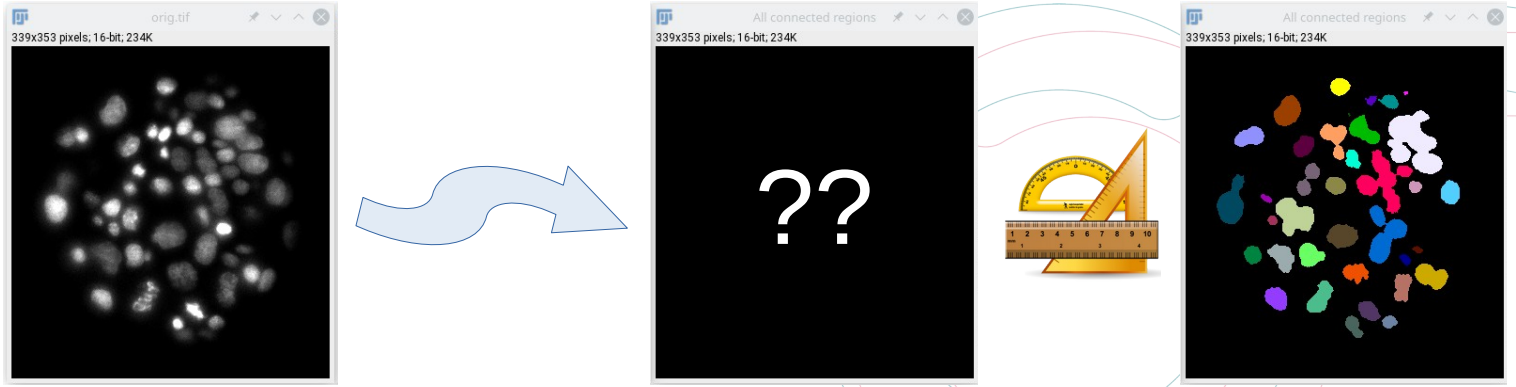
# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data applications

- Teaching AI:



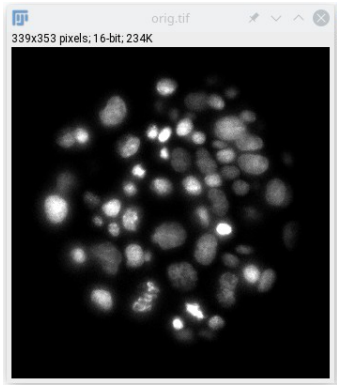

- Evaluate image processing:

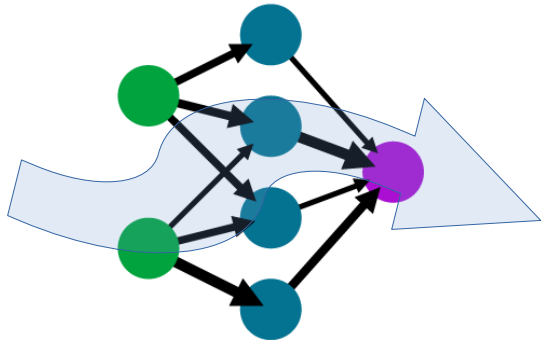


# MIFA: Metadata, Incentives, Formats, Accessibility

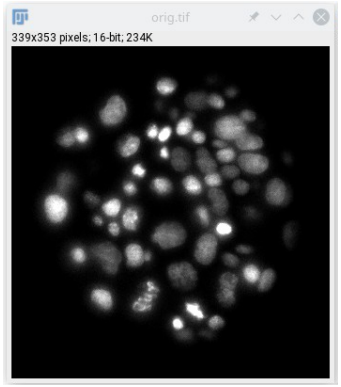
## Annotated data applications

- Teaching AI:



- Evaluate image processing:



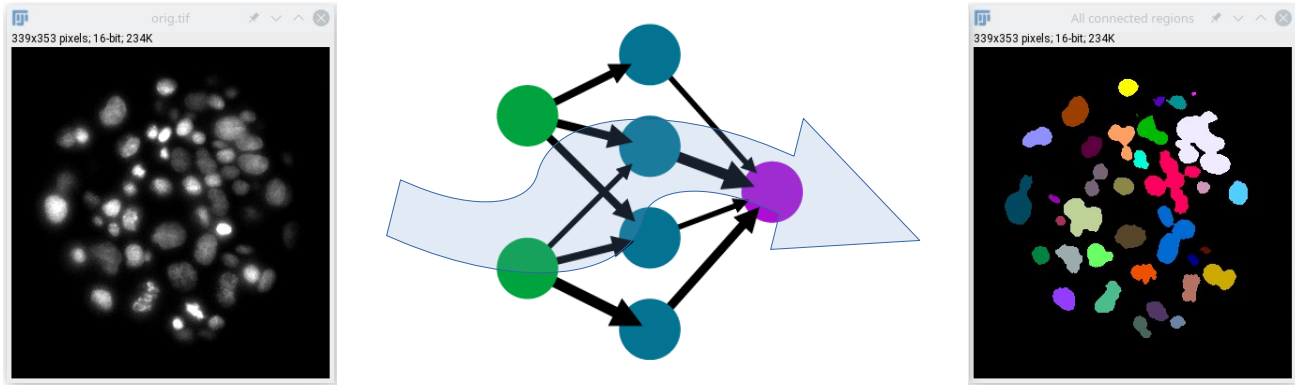


- Needs the annotations

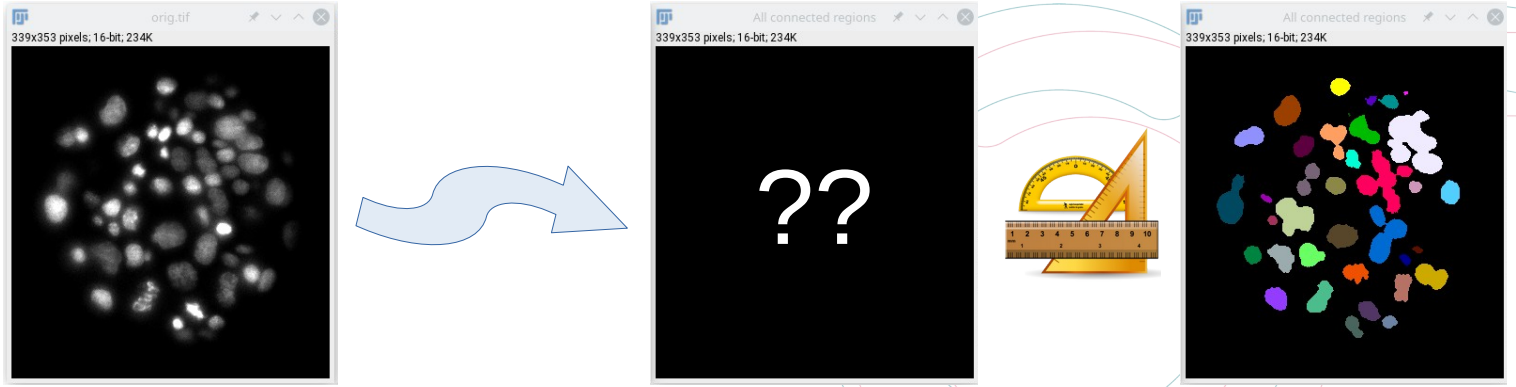
# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data applications

- Teaching AI:




- Evaluate image processing:




- Needs the annotations
- Nobody cared until AI came...

# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data publishing *(next to the models)*

- Overarching goal
  - Performant and universal method
  - SOTA:  AI-based method
  - For specific cases, good solutions exist
  - More cases needed

# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data publishing *(next to the models)*

- Overarching goal
  - Performant and universal method
  - SOTA: AI-based method

  - For specific cases, good solutions exist
  - More cases needed

  - Generally well-performant AI method
  - More cases needed („foundational“)

*„A strategy to cover a larger domain and to make models more reusable and generalizable is to use large, heterogeneous, annotated datasets for training.“*

# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data publishing *(next to the models)*

- Overarching goal
  - Performant and universal method
  - SOTA: AI-based method
  - For specific cases, good solutions exist
  - More cases needed
  - Generally well-performant AI method
  - More cases needed („foundational“)
- „More cases“ specifically
  - How to store and share annotations
  - Mainly for AI domain
  - Mainly biological imagery
  - From comp.sci. perspective

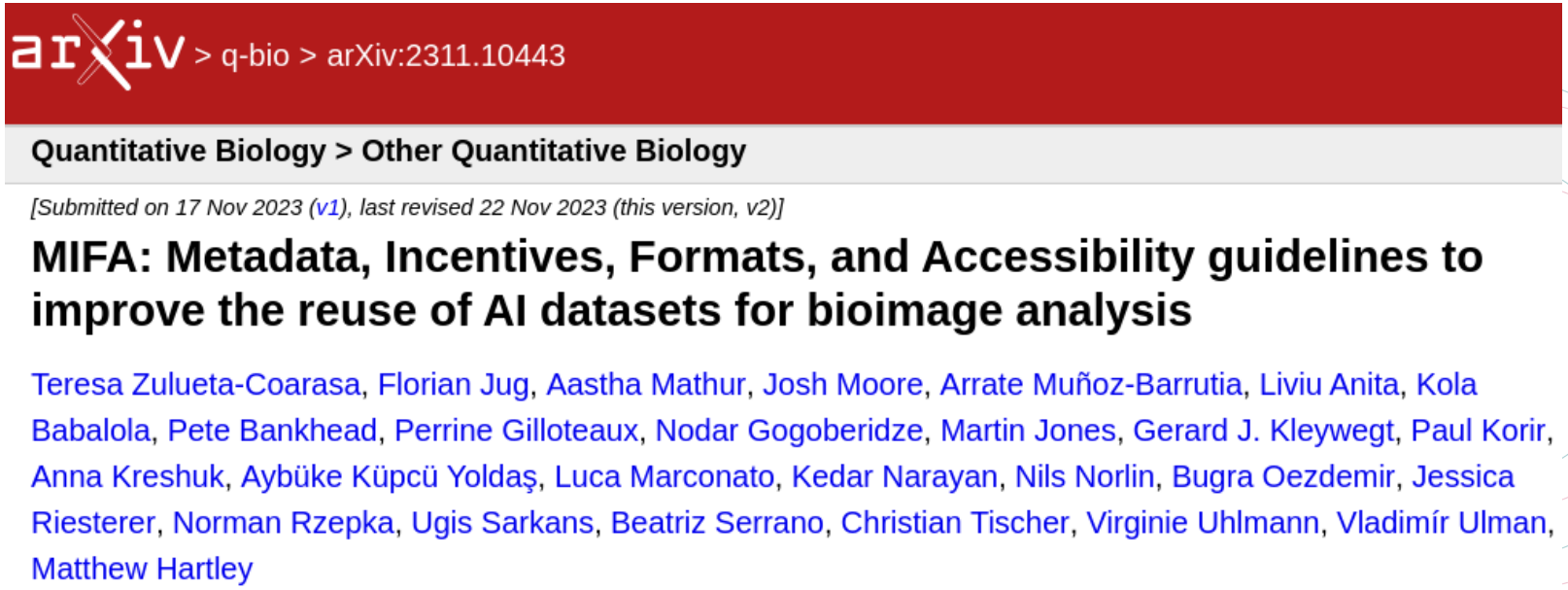
*„A strategy to cover a larger domain and to make models more reusable and generalizable is to use large, heterogeneous, annotated datasets for training.“*

# MIFA: Metadata, Incentives, Formats, Accessibility

## Annotated data publishing

- In 2023, 45 community (comp.sci. + biology) experts gathered
- MIFA recommendation is FAIR

- https://arxiv.org/abs/2311.10443
- Accepted now in Nature Methods



arXiv > q-bio > arXiv:2311.10443

Quantitative Biology > Other Quantitative Biology

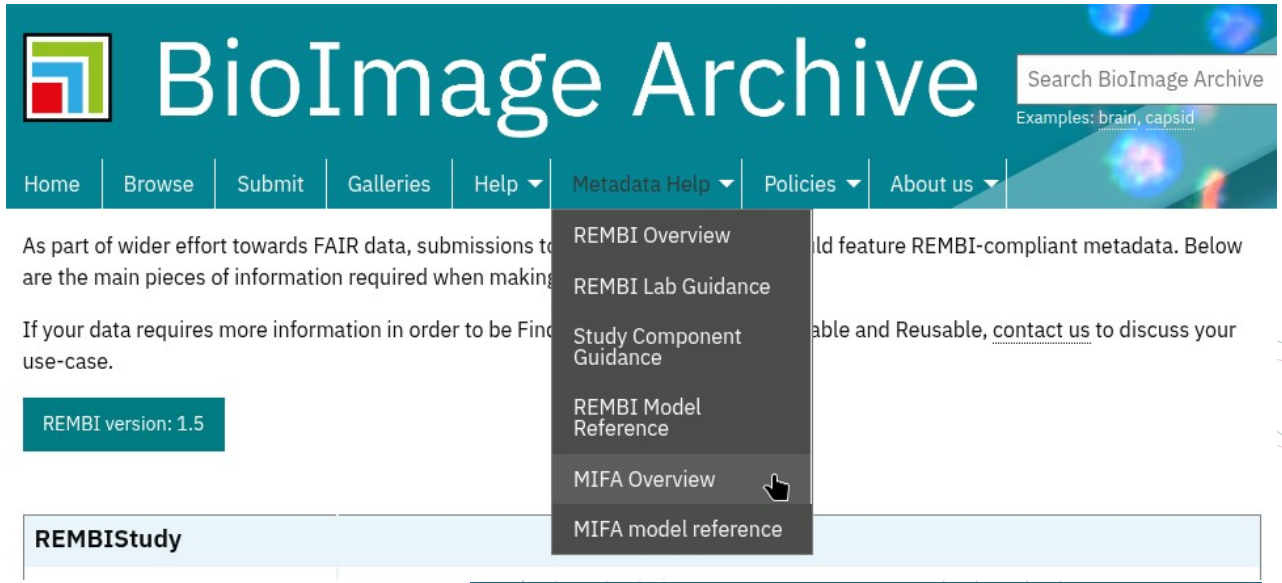
*[Submitted on 17 Nov 2023 (v1), last revised 22 Nov 2023 (this version, v2)]*

**MIFA: Metadata, Incentives, Formats, and Accessibility guidelines to improve the reuse of AI datasets for bioimage analysis**

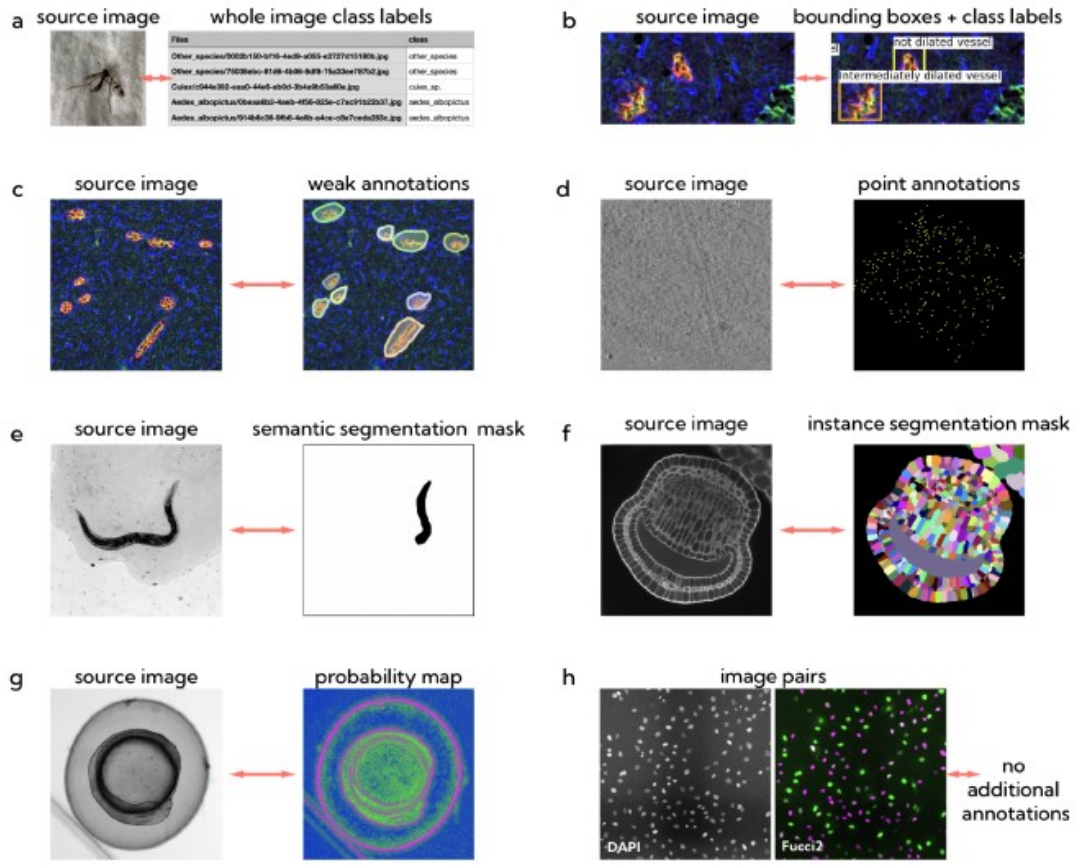
Teresa Zulueta-Coarasa, Florian Jug, Aastha Mathur, Josh Moore, Arrate Muñoz-Barrutia, Liviu Anita, Kola Babalola, Pete Bankhead, Perrine Gilloteaux, Nodar Gogoberidze, Martin Jones, Gerard J. Kleywegt, Paul Korir, Anna Kreshuk, Aybüke Küpcü Yoldaş, Luca Marconato, Kedar Narayan, Nils Norlin, Bugra Oezdemir, Jessica Riesterer, Norman Rzepka, Ugis Sarkans, Beatriz Serrano, Christian Tischer, Virginie Uhlmann, Vladimír Ulman, Matthew Hartley

# MIFA: Metadata, Incentives, Formats, Accessibility

- Hoped to become an *initiator of a standard*
  - „REMBI paper“ momentum
  - BioImage Archive people
  - BioImage Model Zoo (AI4Life) people
  - EuroBioimaging people
- To alert funders:
  - Creating annotations is Effort(!)
  - Creating annotations is Important




# MIFA: Metadata, Incentives, Formats, Accessibility




## Standardise METADATA

Authors
Annotation overview
Annotation type
Annotation method
Confidence level
Annotation criteria
Annotation coverage
Source image association
Transformations
Spatial information
Completion time

## Data production INCENTIVES

Preserve credit in different dataset versions

Acknowledge exceptional contributors

Organise annotation events

Clear contributor credit

## Reduce number of FORMATS

OME-Zarr

GeoJSON

COCO

EMDB-SFF

## Improve data ACCESSIBILITY

API

Data browsing

Allow versioning

Link to community tools

Metadata search

Reach out to journals

# MIFA: Metadata, Incentives, Formats, Accessibility

The devil is in the metadata: towards a recommended metadata standard

- Study-level metadata
  - Context/motivation to annotations
  - License
  - Connection to created models
- Images metadata
  - Raw images -> REMBI metadata model
  - Btw, EOSC.CZ bio. img. pilot repo extends CCMM with REMBI

# MIFA: Metadata, Incentives, Formats, Accessibility

The devil is in the metadata: towards a recommended metadata standard

- Annotations metadata
  - List of items with descriptions and examples, e.g.:
  - *Type, method, confidence level, criteria, coverage, source image*
- Versioning metadata
  - Annotations can always be improved...

ndi-wg-metadata, ndi-wg-ai-ml, ndi-wg-bio-health-food, Pilot Repo BioImg

# MIFA: Metadata, Incentives, Formats, Accessibility

Credit where credit's due: incentivising production and sharing of AI-ready datasets

- (Good) Annotation really is an effort
- Encourage sharing datasets
- Recognition for the work  (also quality-warranting effect)
- Impact factor for datasets, for annotators

- Support challenges, „clickathons“
- Tools development,  Training events

# MIFA: Metadata, Incentives, Formats, Accessibility

Formats: next stop, next generation

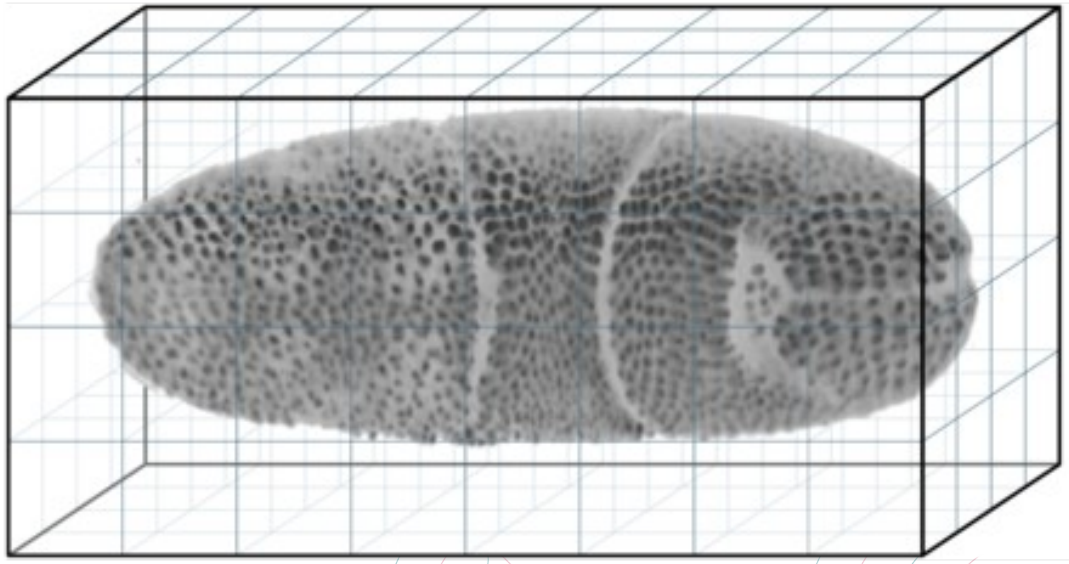
- Focus on limited number of them
- Pixel-based vs. Vector-based vs. Mixed
  - OME TIFF, OME Zarr
  - GeoJSON, EMDB-SFF
  - COCO JSON


- Vector-based:  More efficient & expressive
- Pixel-based:  Easily „cloud-ready“

# MIFA: Metadata, Incentives, Formats, Accessibility

Formats: next stop, next generation

- Cloud-ready also means
  - HPC-ready
  - Collaboration-ready
  - Paradigm change in tools
    - Missing pixels in RAM
    - ROI is not Full image



….relevant mainly to image (tensor) data

# MIFA: Metadata, Incentives, Formats, Accessibility

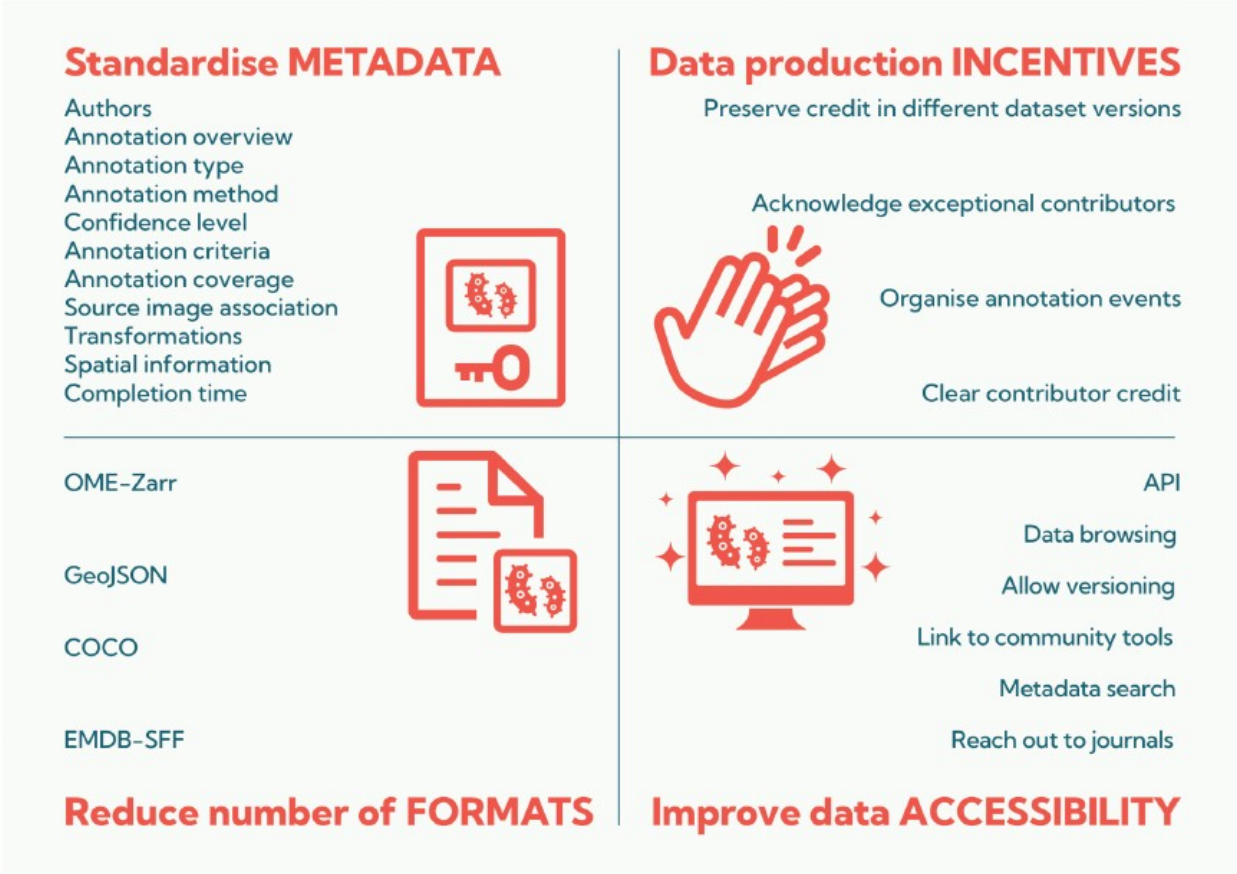
Improving findability, accessibility and presentation of AI-ready datasets

- Build repositories
  - Host data *and* models
  - Figure-out relevant ontologies
- Develop API and Ready-made Tools
  - For data searching
  - For remote and local viewing, accessing ROIs
- Manage quality: Reviews and Versions

# MIFA: Metadata, Incentives, Formats, Accessibility

- Overview of main *issues*
- Offers answers, or directions
- Various stand-points of various stakeholders
- Details missing…

- Overlap with EOSC-CZ
- Define the missing details?



National Czech Programme | eosc

# Thank you.



vladimir.ulman@vsb.cz

Co-funded by the European Union

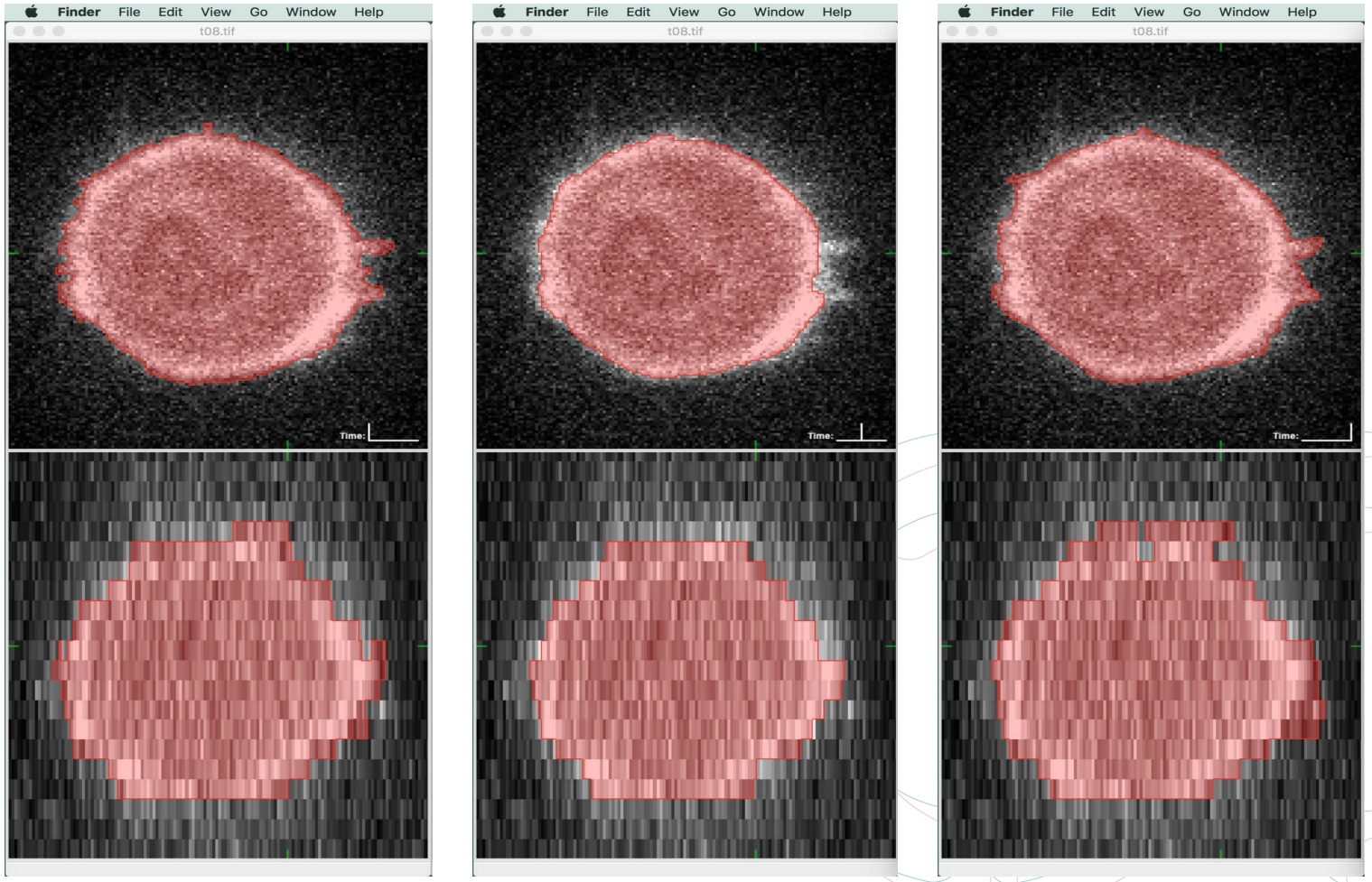
MINISTRY OF EDUCATION, YOUTH AND SPORTS

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

# Segmentation Annotation Creation (before Sharing)

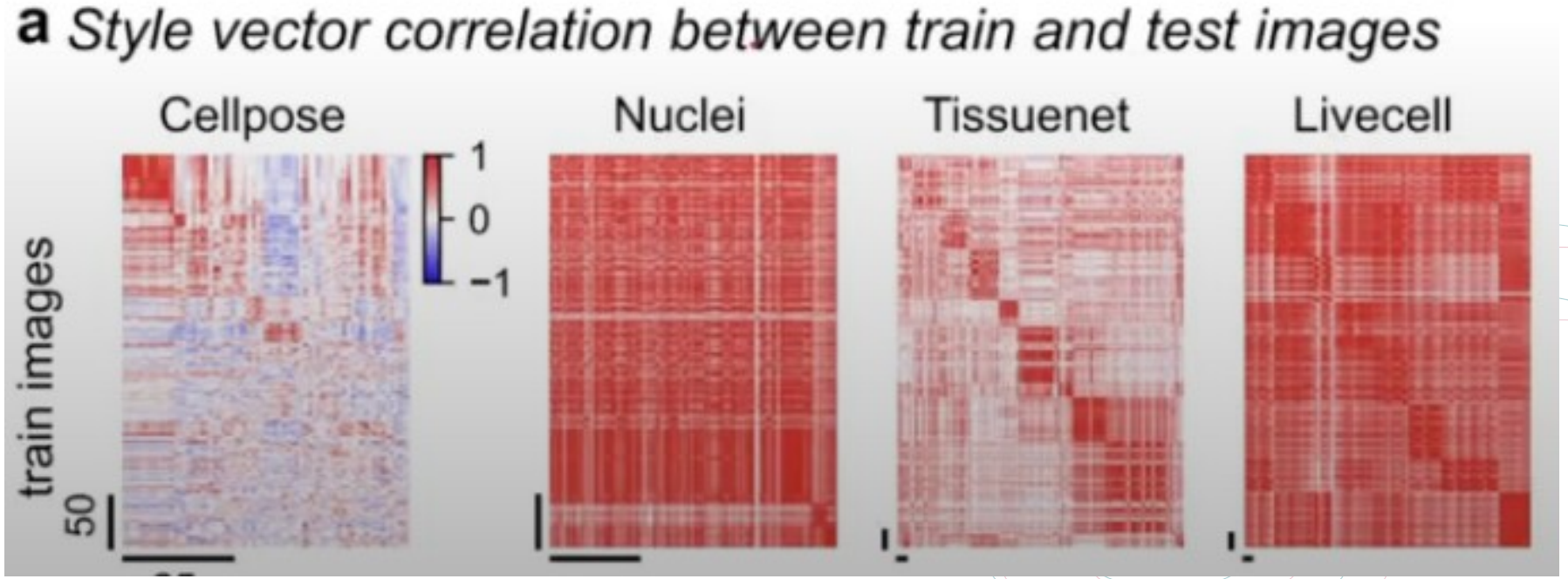
Comparison of segmentation annotation between three annotators

Implication on model performance, and user satisfaction



# Segmentation Annotation Creation (before Sharing)

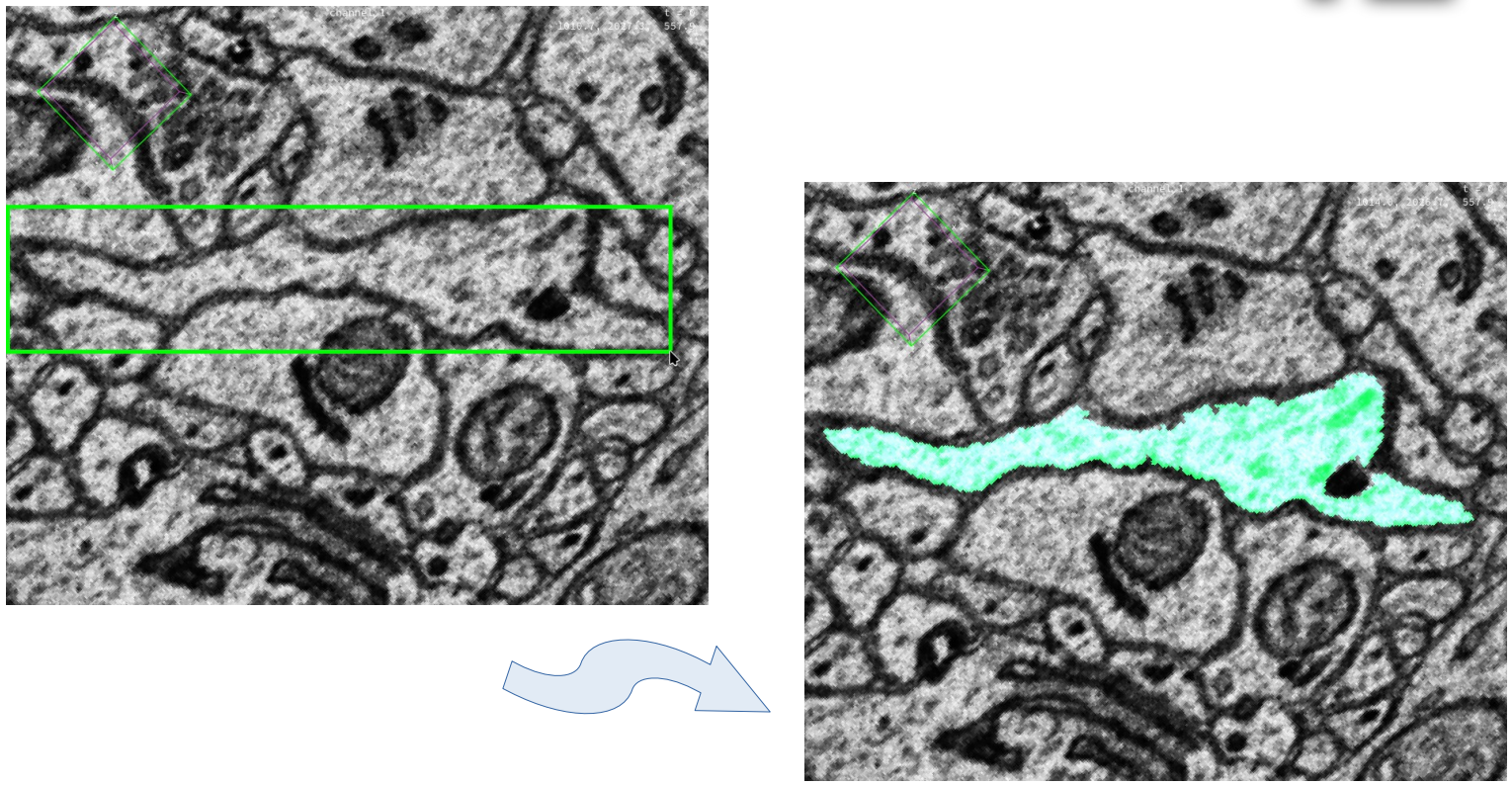
Visualization of „self similarity“ in a training dataset




Courtesy of: Marius Pachitariu, HHMI Janelia Research Campus

# Segmentation Annotation Creation (before Sharing)

Annotation „AID“ for 2D and 3D images in



# Segmentation Annotation Creation (before Sharing)

Annotation „AID“ for 2D and 3D images in