TheLion.AI
INTERDISCIPLINARY RESEARCH GROUP

UMIE datasets

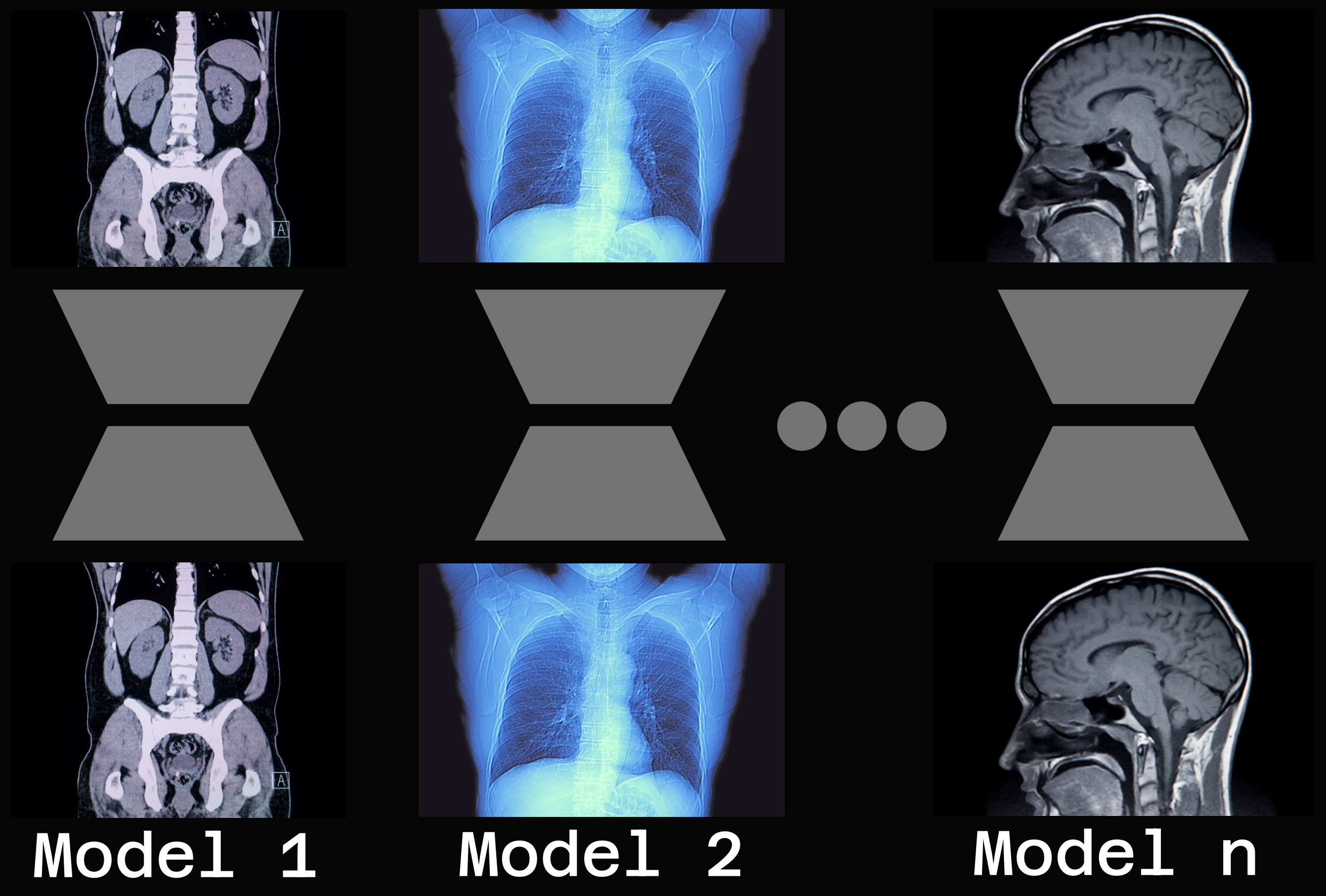
Towards medical foundation models. A unified dataset of medical images.

Bartłomiej Kuchnowski
TheLion.AI / Tooploox

Kacper Knitter
TheLion.AI / PJATK

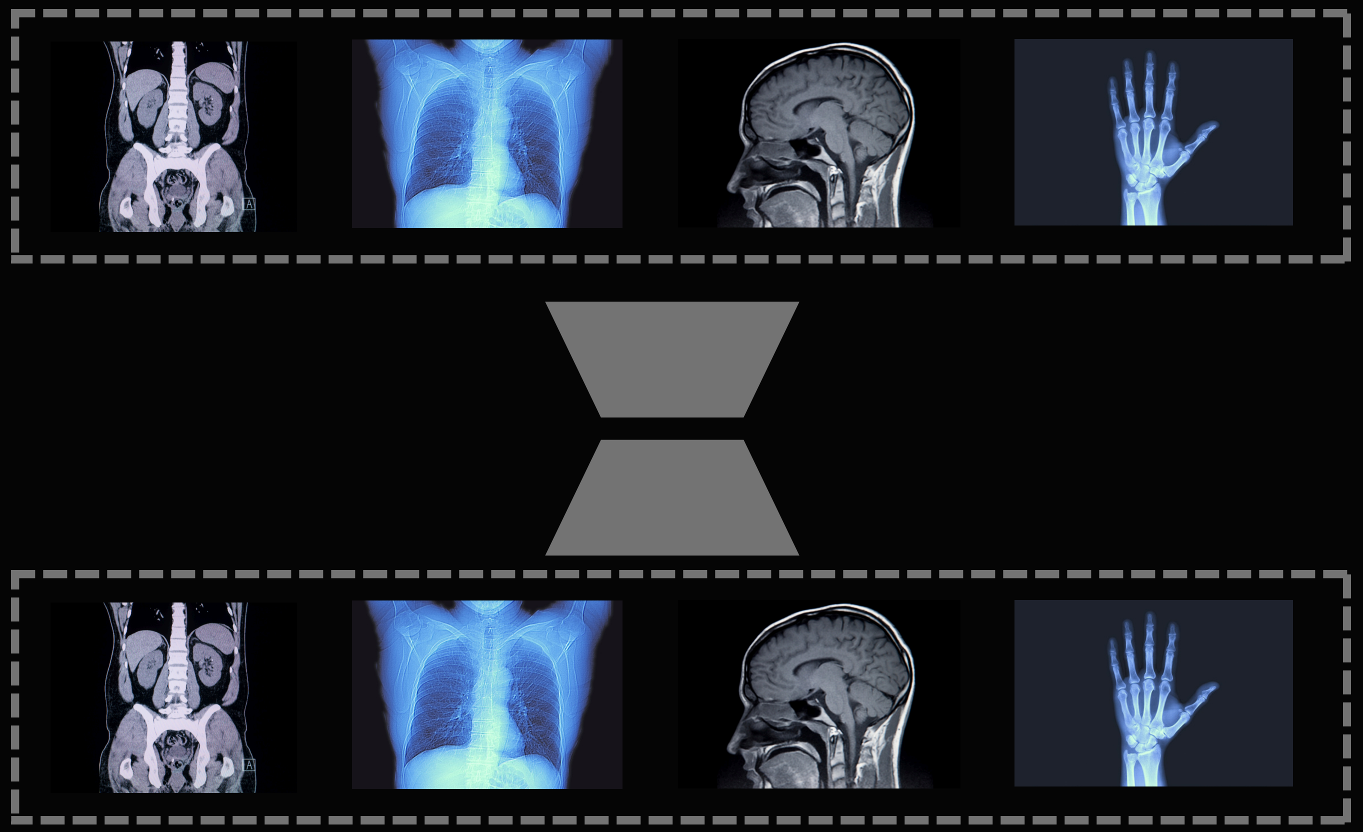
Barbara Klaudel
TheLion.AI

Traditional ML vs New ML



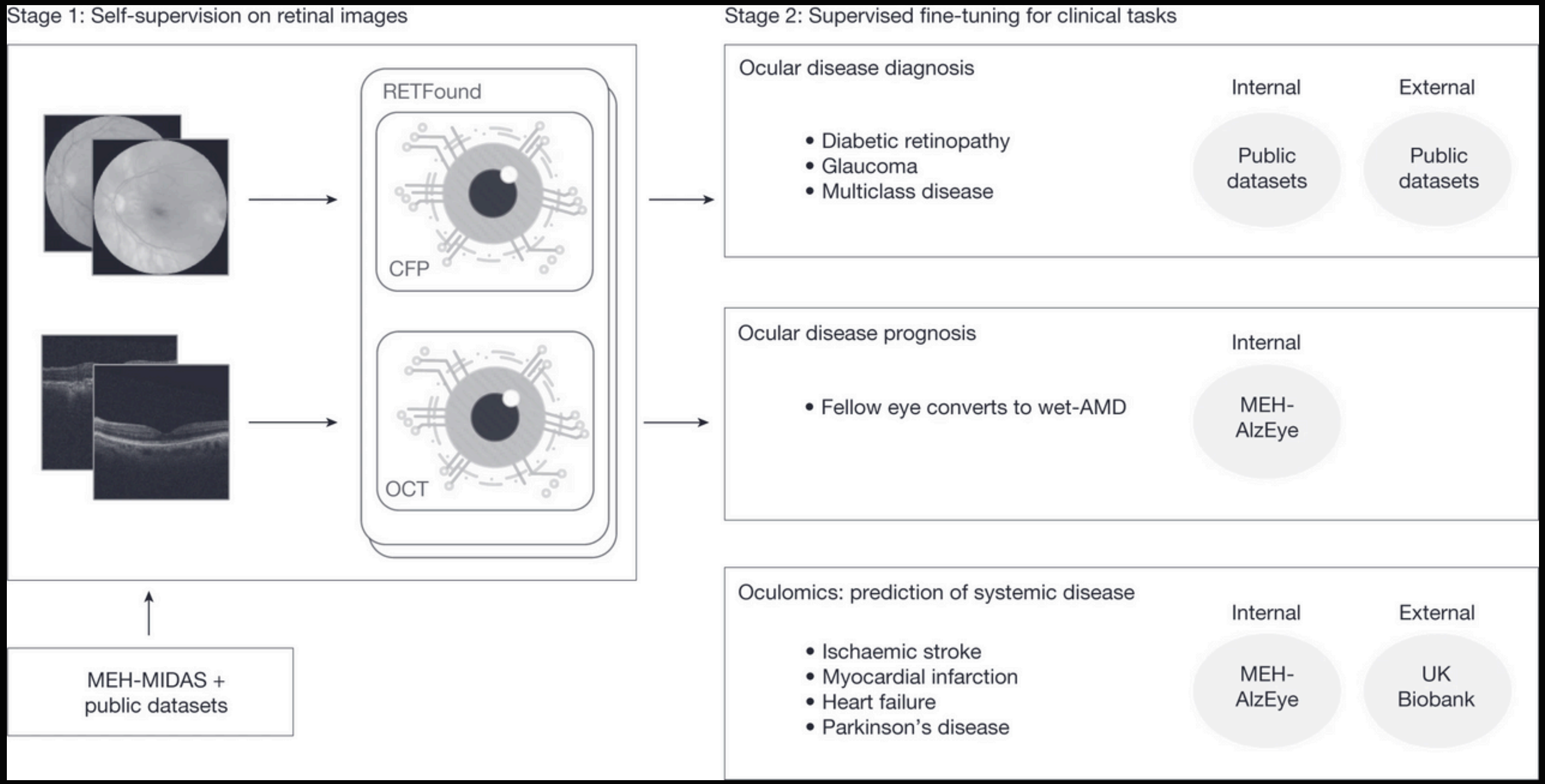
Old paradigm: One model per task

VS



New paradigm: One model for all

Healthcare “foundation models”



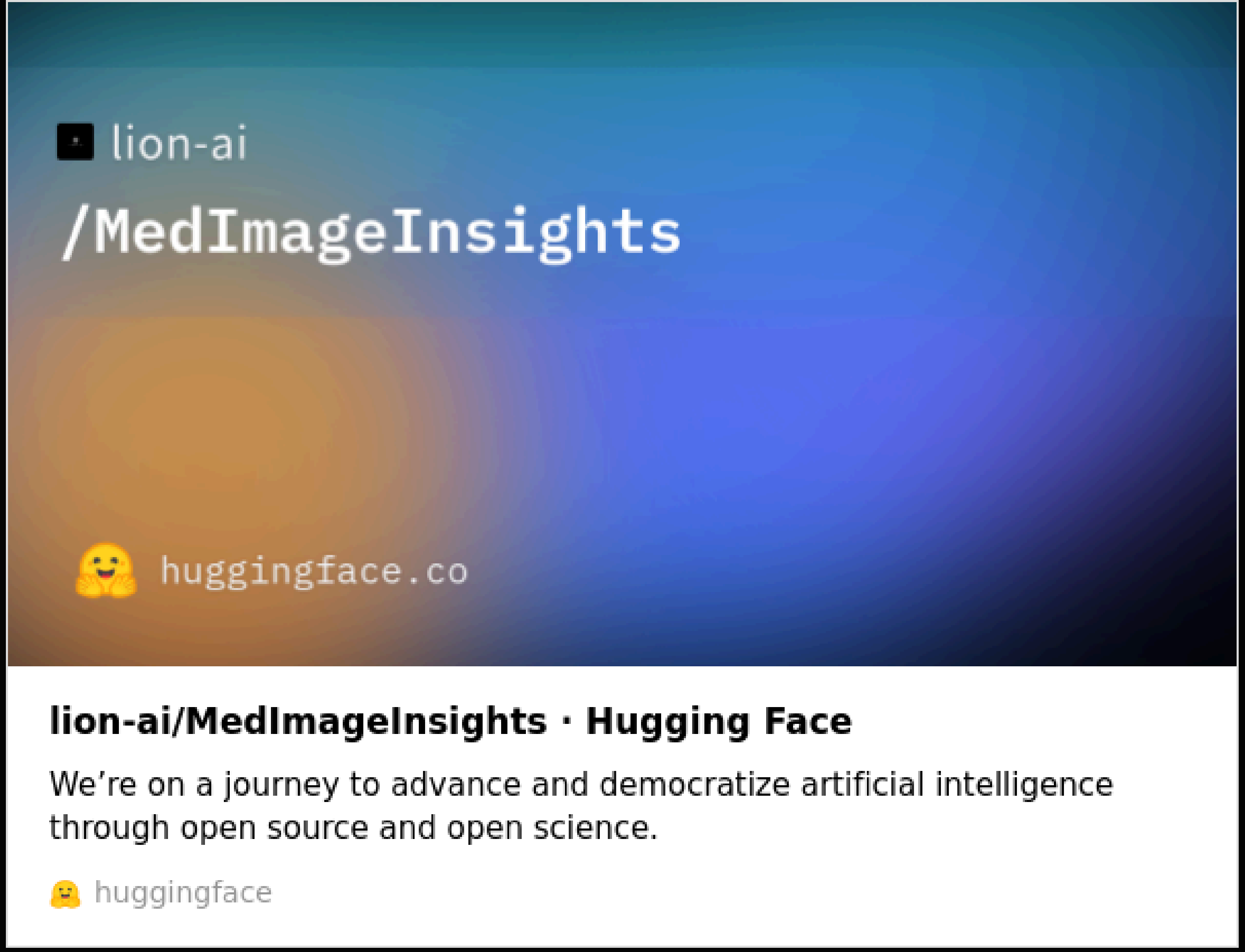
Zhou et al, Nature 2023

“We created “a foundation model” for OCT images that predicts 10 different diseases.”

Only 1 type of images

still a narrow task

Healthcare “foundation models”



lion-ai

/MedImageInsights

huggingface.co

lion-ai/MedImageInsights · Hugging Face

We’re on a journey to advance and democratize artificial intelligence through open source and open science.

huggingface

Data problems

- Research papers do not make their data public
- Not enough opensource datasets
- Lack of standards for storing data
- lack of standards for data labelling

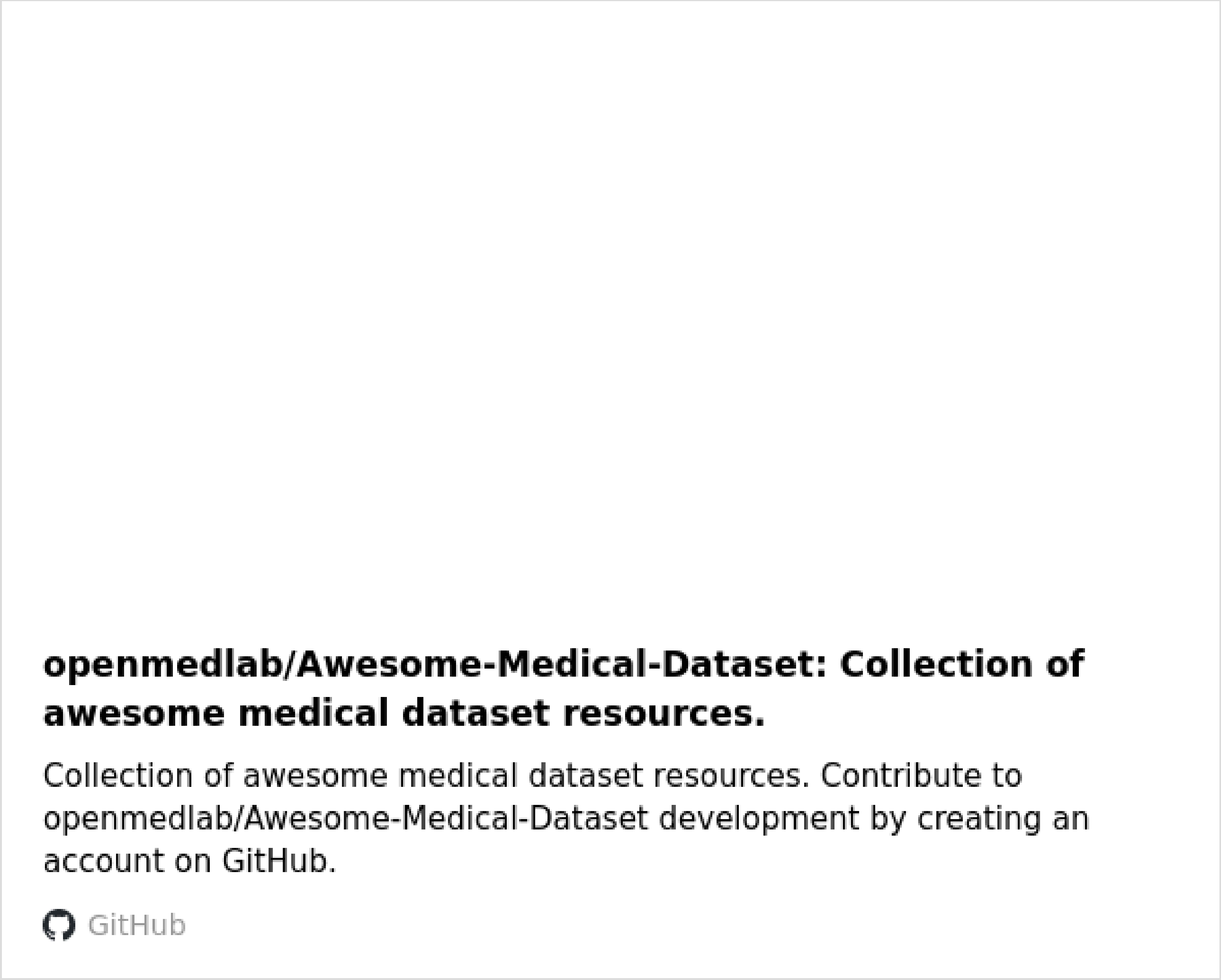
Lack of public datasets

Example: kidney tumor malignancy prediction

KITS 23	Private datasets	Altogether
12	24	30 research papers

Only 1 open-source dataset!

Not enough opensource datasets



openmedlab/Awesome-Medical-Dataset: Collection of awesome medical dataset resources.

Collection of awesome medical dataset resources. Contribute to openmedlab/Awesome-Medical-Dataset development by creating an account on GitHub.

GitHub

THE CANCER IMAGING ARCHIVE

kaggle

Stanford AIMI

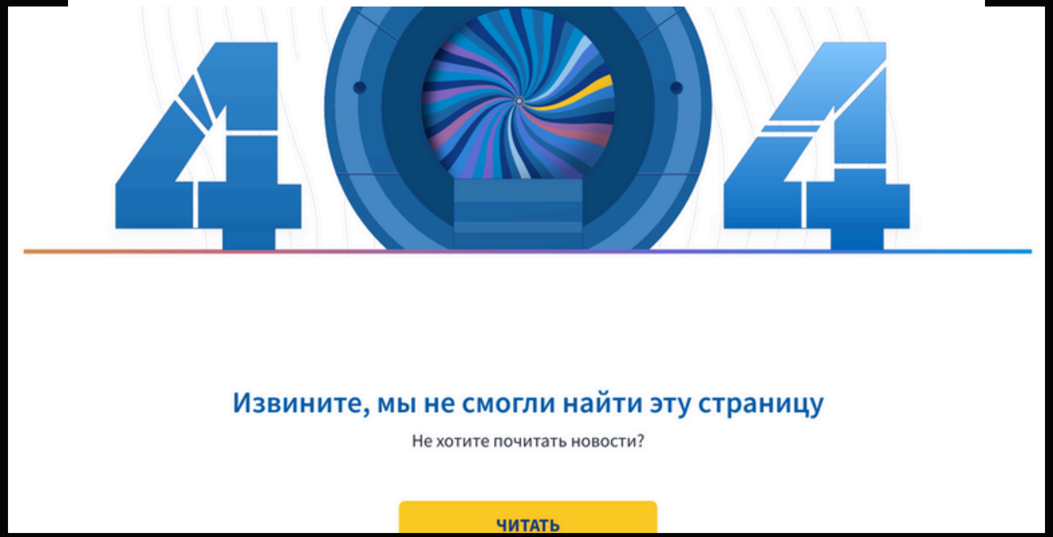
Radboudumc

Grand Challenge

A platform for end-to-end development of machine learning solutions in biomedical imaging.

104,000+ users 371 challenges 4,956 algorithms

Permanent link: https://mosmed.ai/datasets/covid19_1110.



Извините, мы не смогли найти эту страницу

Не хотите почитать новости?

ЧИТАТЬ

Lack of standards for storing data

Example 1: KITS 23

Image + mask

imaging.nii.gz
segmentation.nii.gz

Label - JSON

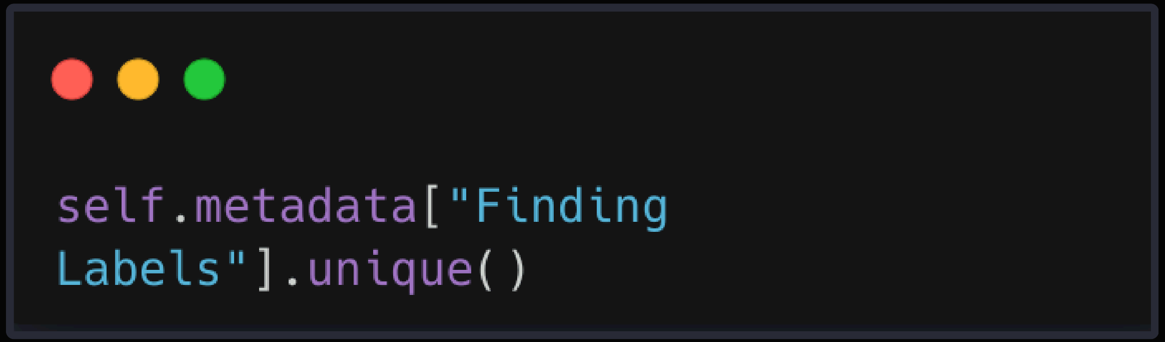
```
kits.json = {
"case_id": "case_00000",
"gender": "male",
"vital_status": "censored",
"vital_days_after_surgery": 1958,
"bmi": 29.47,
"tumor_histologic_subtype": "papillary_rcc",


}
```

Example 2: Stanford COCA

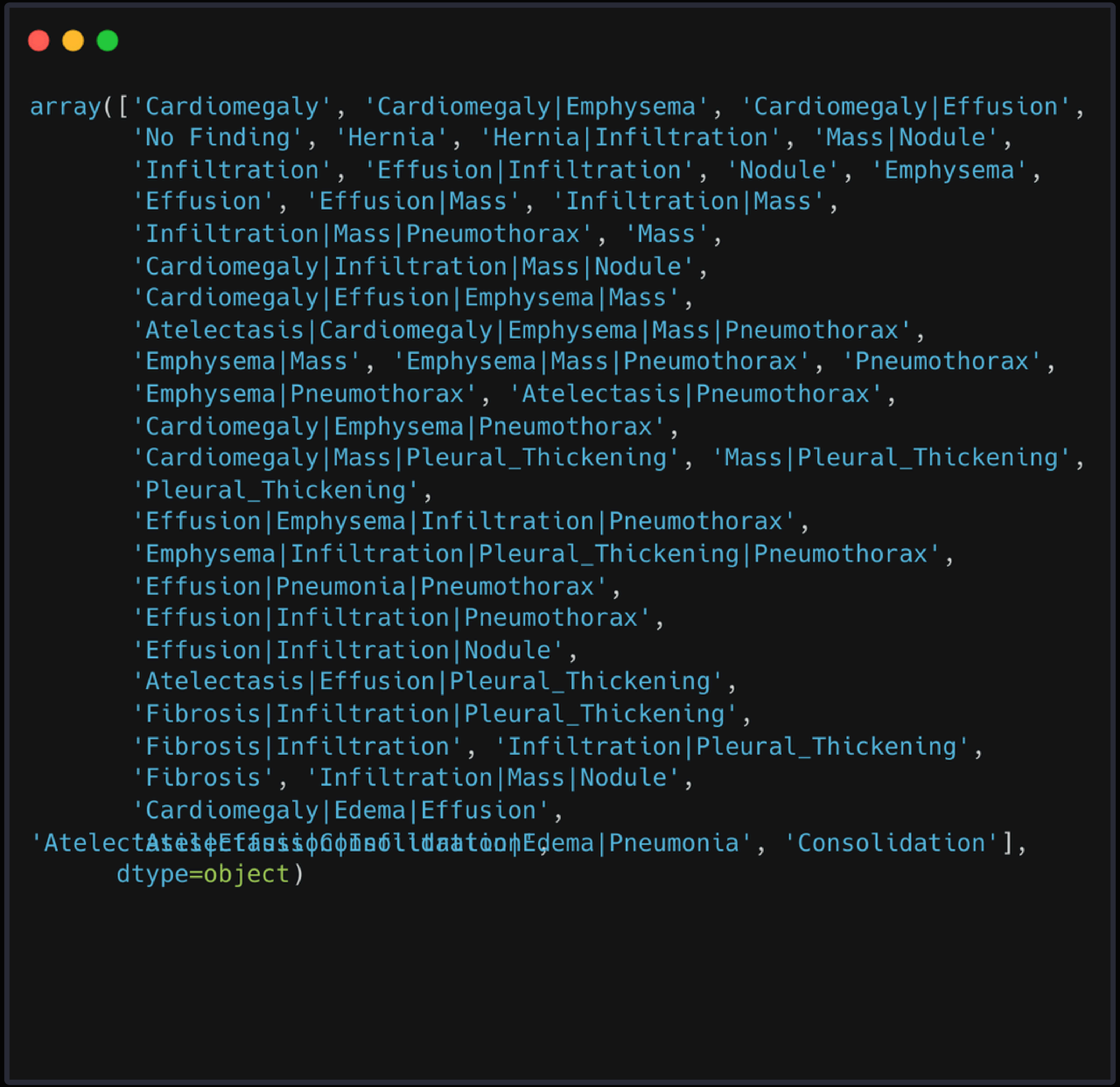
Mask - XML

```
   <key>ImageIndex</key>
   <integer>35</integer>
   <key>NumberOfROIs</key>
   <integer>1</integer>
   <key>ROIs</key>
   <array>
    <dict>
     <key>Area</key>
     <string>Right Coronary Artery</string>
     <key>NumberOfPoints</key>
     <integer>0</integer>
     <key>Point_mm</key>
     <array/>
     <key>Point_px</key>
     <array/>
     <key>Total</key>
     <real>0.0</real>
     <key>Type</key>
     <integer>20</integer>
    </dict>
   </array>
  </dict>
 </array>
</dict>
</plist>
```

Example 3: Chest Xray14 - multilabel classification in csv



```
self.metadata["Finding
Labels"].unique()
```



```
array(['Cardiomegaly', 'Cardiomegaly|Emphysema', 'Cardiomegaly|Effusion',
       'No Finding', 'Hernia', 'Hernia|Infiltration', 'Mass|Nodule',
       'Infiltration', 'Effusion|Infiltration', 'Nodule', 'Emphysema',
       'Effusion', 'Effusion|Mass', 'Infiltration|Mass',
       'Infiltration|Mass|Pneumothorax', 'Mass',
       'Cardiomegaly|Infiltration|Mass|Nodule',
       'Cardiomegaly|Effusion|Emphysema|Mass',
       'Atelectasis|Cardiomegaly|Emphysema|Mass|Pneumothorax',
       'Emphysema|Mass', 'Emphysema|Mass|Pneumothorax', 'Pneumothorax',
       'Emphysema|Pneumothorax', 'Atelectasis|Pneumothorax',
       'Cardiomegaly|Emphysema|Pneumothorax',
       'Cardiomegaly|Mass|Pleural_Thickening', 'Mass|Pleural_Thickening',
       'Pleural_Thickening',
       'Effusion|Emphysema|Infiltration|Pneumothorax',
       'Emphysema|Infiltration|Pleural_Thickening|Pneumothorax',
       'Effusion|Pneumonia|Pneumothorax',
       'Effusion|Infiltration|Pneumothorax',
       'Effusion|Infiltration|Nodule',
       'Atelectasis|Effusion|Pleural_Thickening',
       'Fibrosis|Infiltration|Pleural_Thickening',
       'Fibrosis|Infiltration', 'Infiltration|Pleural_Thickening',
       'Fibrosis', 'Infiltration|Mass|Nodule',
       'Cardiomegaly|Edema|Effusion',
'Atelectasis|Effusion|Consolidation',
       'Atelectasis|Consolidation|Edema|Pneumonia', 'Consolidation'],
      dtype=object)
```

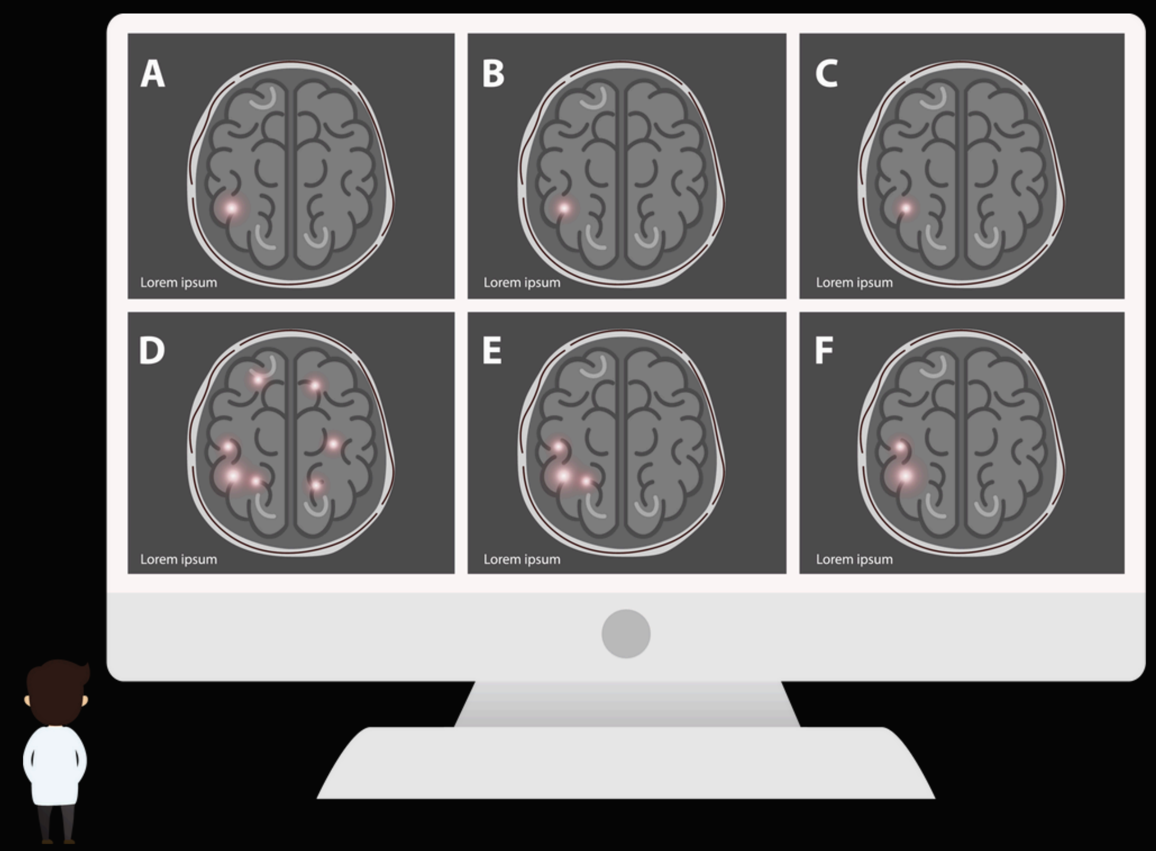
Lack of standards for data labelling

Example 1: mental shortcuts “cc_papillary”

Example 2: regionalisms: “serce podparte na przeponie”

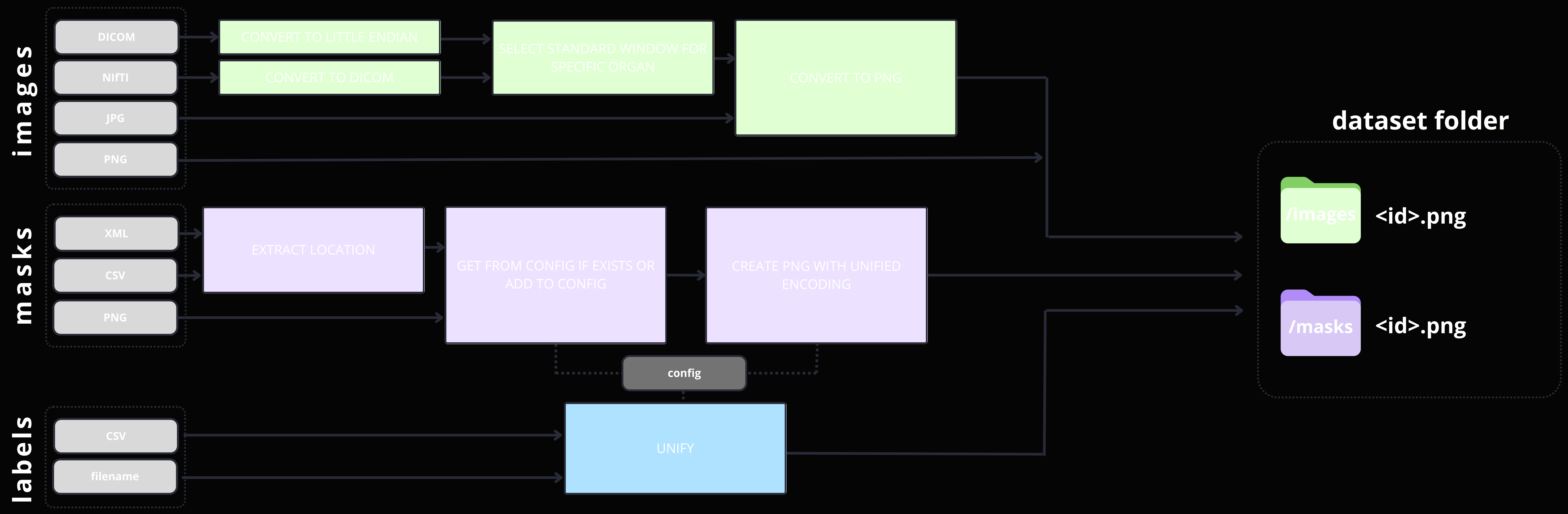
Universal Medical Image Encoding (UMIE)

UMIE dataset

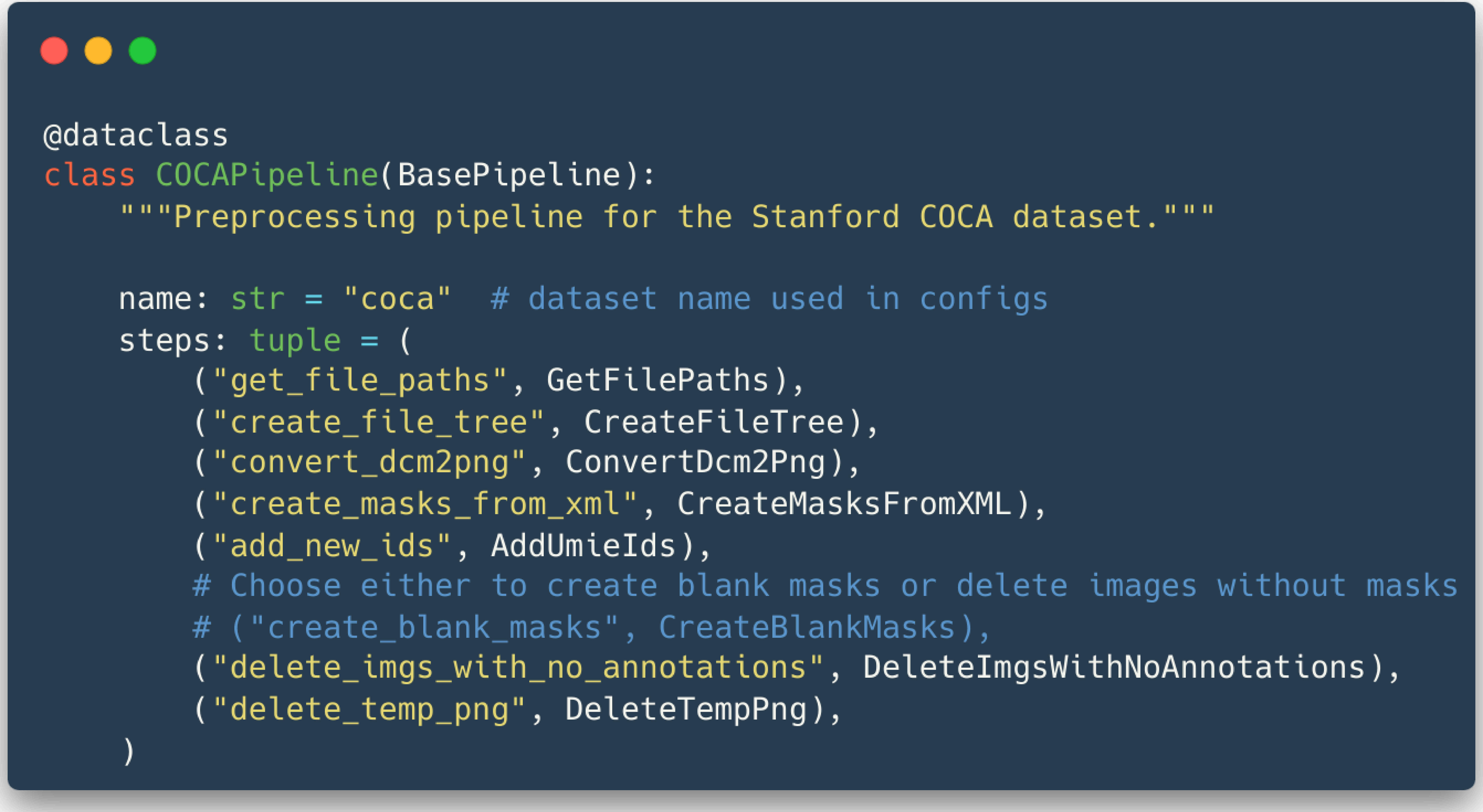


- 20+ datasets, 46 labels, 15 masks
- 1 million+ images
- CT, MRI, X-ray
- classification + segmentation

UMIE datasets

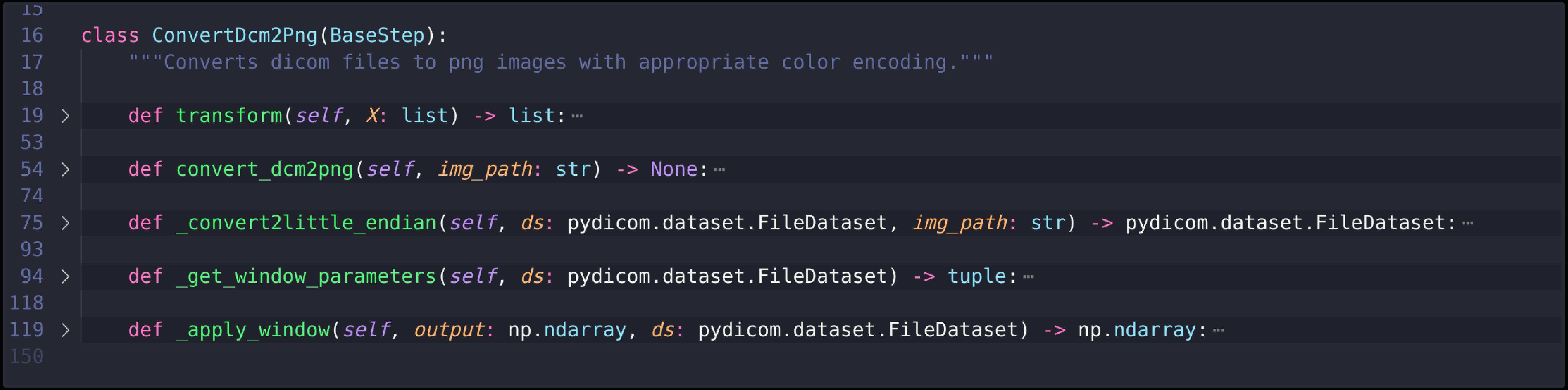


- Use one of 20+ pipelines for the supported datasets or construct a pipeline from ready-to-use steps.



```python
@dataclass
class COCAPipeline(BasePipeline):
    """Preprocessing pipeline for the Stanford COCA dataset."""

    name: str = "coca"  # dataset name used in configs
    steps: tuple = (
        ("get_file_paths", GetFilePaths),
        ("create_file_tree", CreateFileTree),
        ("convert_dcm2png", ConvertDcm2Png),
        ("create_masks_from_xml", CreateMasksFromXML),
        ("add_new_ids", AddUmieIds),
        # Choose either to create blank masks or delete images without masks
        # ("create_blank_masks", CreateBlankMasks),
        ("delete_imgs_with_no_annotations", DeleteImgsWithNoAnnotations),
        ("delete_temp_png", DeleteTempPng),
    )
```

sklearn.pipeline



```
class ConvertDcm2Png(BaseStep):
    """Converts dicom files to png images with appropriate color encoding."""

    def transform(self, X: list) -> list:⋯

    def convert_dcm2png(self, img_path: str) -> None:⋯

    def _convert2little_endian(self, ds: pydicom.dataset.FileDataset, img_path: str) -> pydicom.dataset.FileDataset:⋯

    def _get_window_parameters(self, ds: pydicom.dataset.FileDataset) -> tuple:⋯

    def _apply_window(self, output: np.ndarray, ds: pydicom.dataset.FileDataset) -> np.ndarray:⋯

```

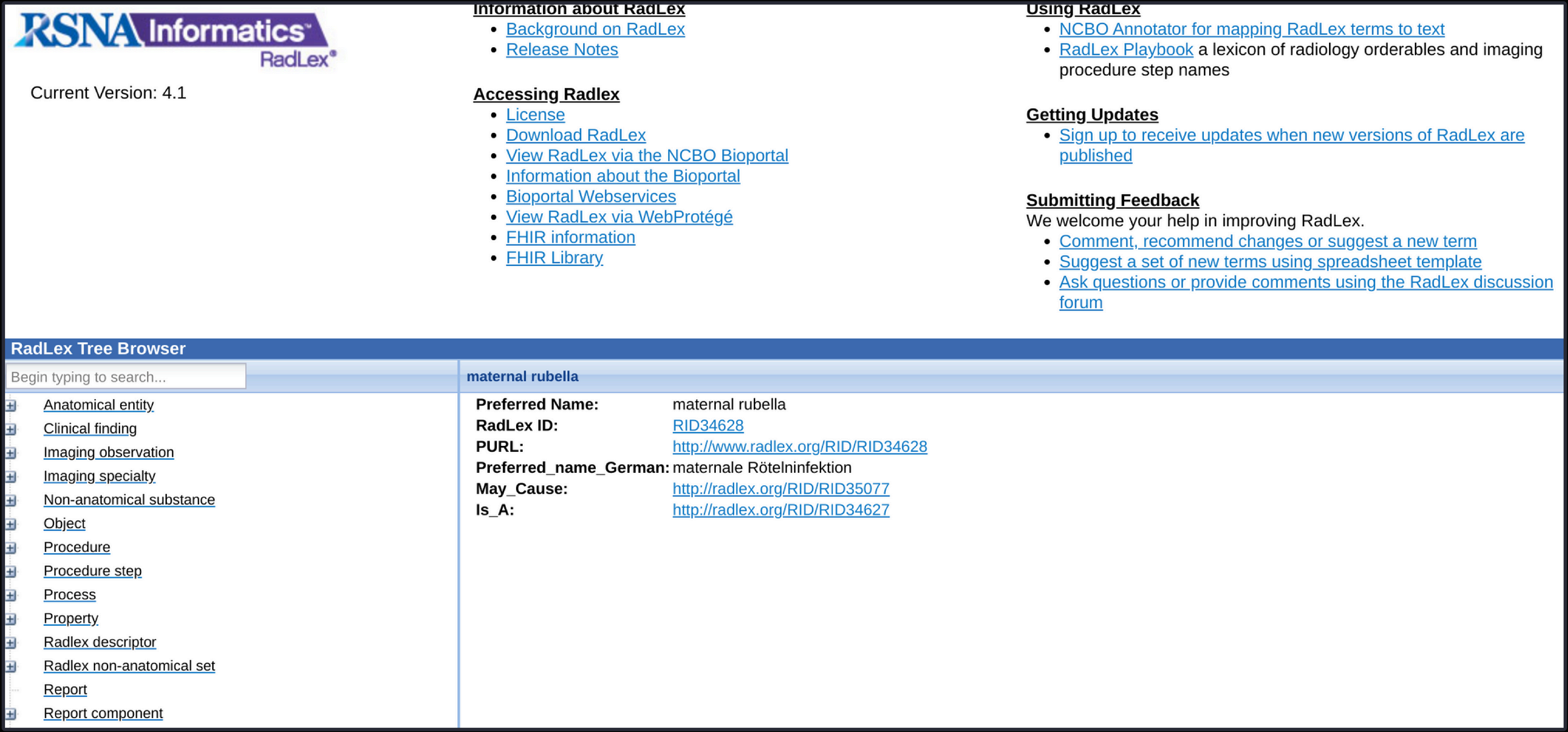
Mix’N Match

- Create File Tree
- Get File Paths
- Create Masks From XML
- Combine Multiple Masks
- Delete Imgs With No Annotations
- ...

20 reusable steps

CREATE YOUR OWN LOOKS!
10 PIECES!
20+ LOOKS!
Barbie
FASHION MIX'N MATCH
3+
WARNING: CHOKING HAZARD - Small parts. Not for children under 3 years.

Unified ontology



Current Version: 4.1

Information about RadLex

- Background on RadLex
- Release Notes

Accessing Radlex

- License
- Download RadLex
- View RadLex via the NCBO Bioportal
- Information about the Bioportal
- Bioportal Webservices
- View RadLex via WebProtégé
- FHIR information
- FHIR Library

Using RadLex

- NCBO Annotator for mapping RadLex terms to text
- RadLex Playbook a lexicon of radiology orderables and imaging procedure step names

Getting Updates

- Sign up to receive updates when new versions of RadLex are published

Submitting Feedback

We welcome your help in improving RadLex.

- Comment, recommend changes or suggest a new term
- Suggest a set of new terms using spreadsheet template
- Ask questions or provide comments using the RadLex discussion forum

RadLex Tree Browser

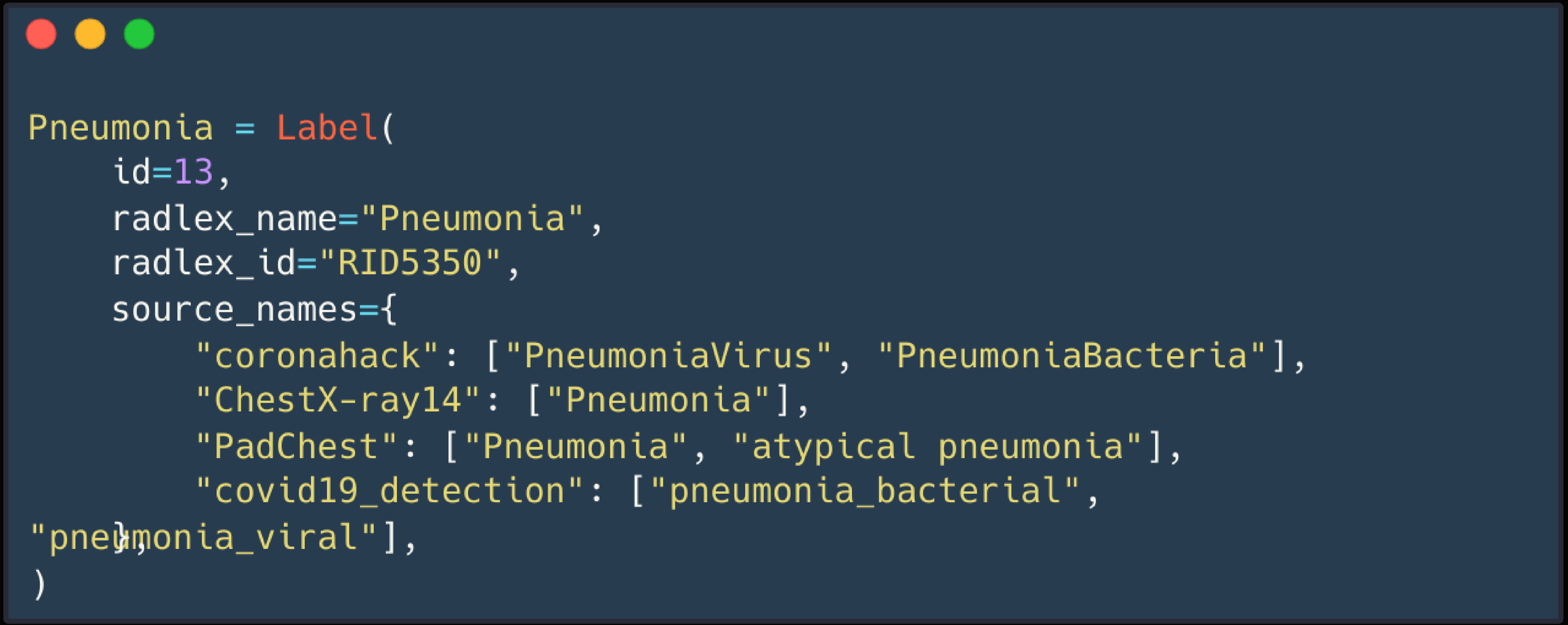
Begin typing to search...

- Anatomical entity
- Clinical finding
- Imaging observation
- Imaging specialty
- Non-anatomical substance
- Object
- Procedure
- Procedure step
- Process
- Property
- Radlex descriptor
- Radlex non-anatomical set
- Report
- Report component

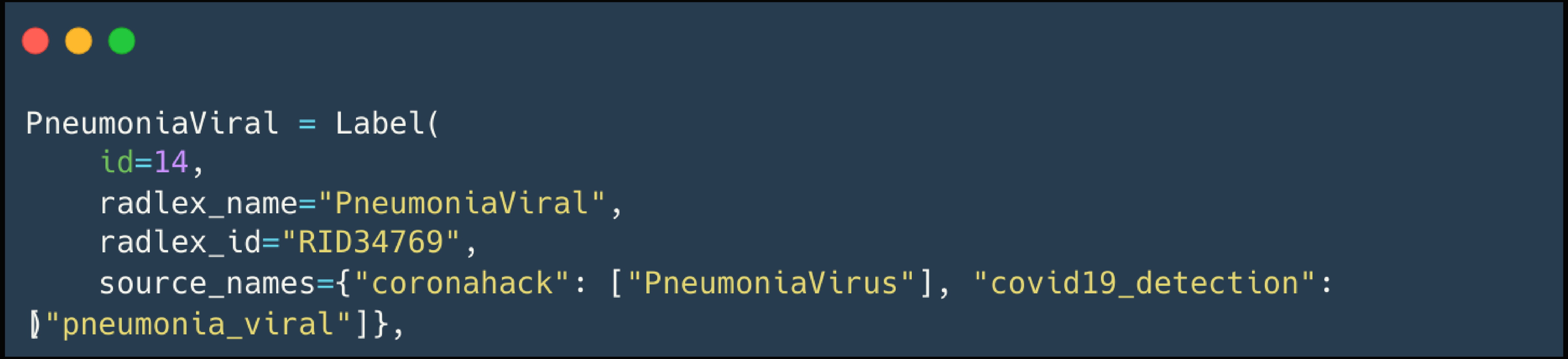
maternal rubella

Preferred Name:	maternal rubella
RadLex ID:	RID34628
PURL:	http://www.radlex.org/RID/RID34628
Preferred_name_German:	maternale Rötelninfektion
May_Cause:	http://radlex.org/RID/RID35077
Is_A:	http://radlex.org/RID/RID34627

Labels and masks



```
Pneumonia = Label(
    id=13,
    radlex_name="Pneumonia",
    radlex_id="RID5350",
    source_names={
        "coronahack": ["PneumoniaVirus", "PneumoniaBacteria"],
        "ChestX-ray14": ["Pneumonia"],
        "PadChest": ["Pneumonia", "atypical pneumonia"],
        "covid19_detection": ["pneumonia_bacterial",
"pneumonia_viral"],
    },
)
```



```
PneumoniaViral = Label(
    id=14,
    radlex_name="PneumoniaViral",
    radlex_id="RID34769",
    source_names={"coronahack": ["PneumoniaVirus"], "covid19_detection":
["pneumonia_viral"]},
```

Licences

- For each dataset, we provide preprocessing scripts.
- Selected datasets are going to be published on HuggingFace Datasets.

Why do we need a large-scale dataset of medical imaging?

Transfer learning

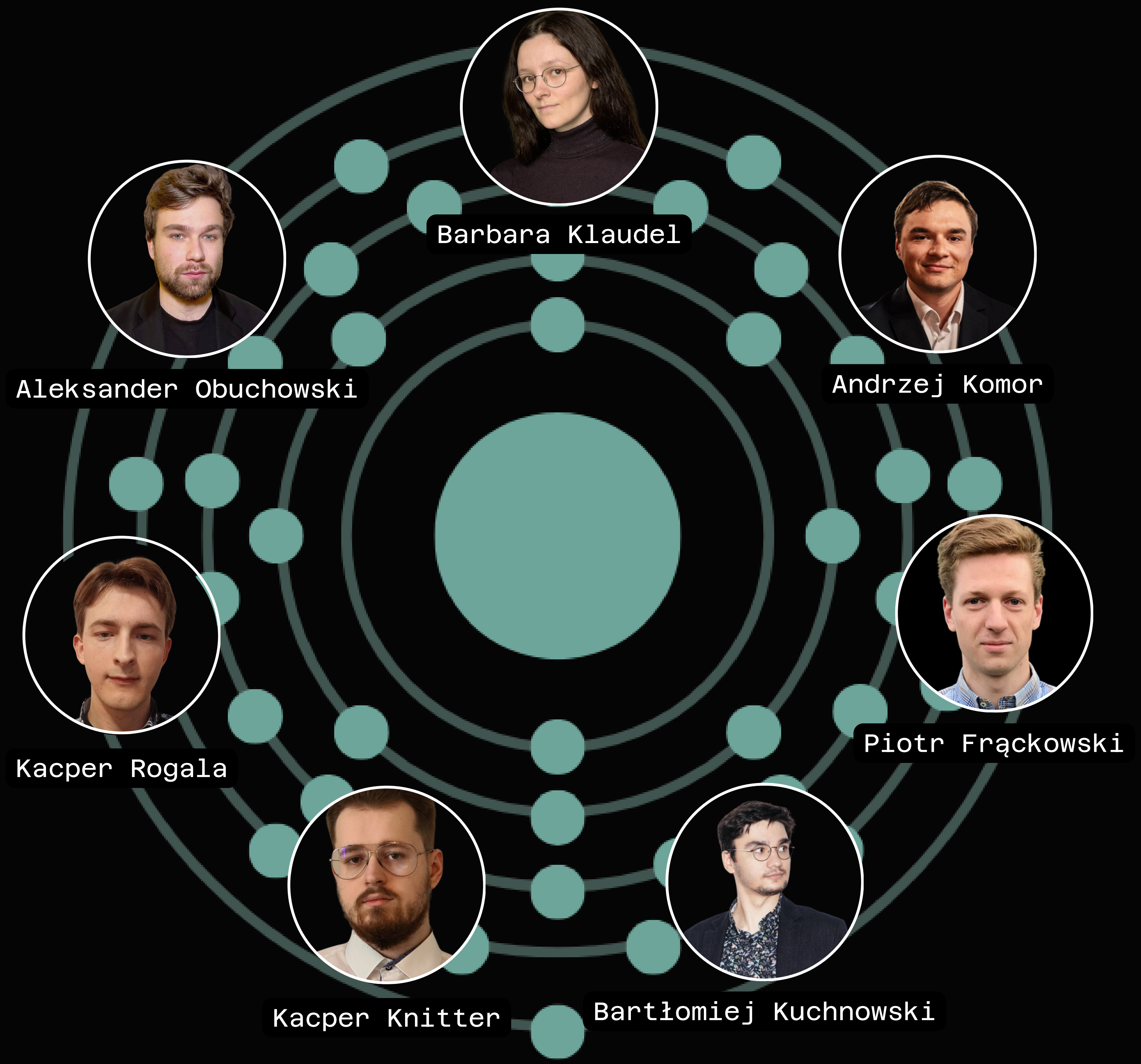
≠

The default strategy for training medical imaging models continues to be pretraining on ImageNet.

TBA

- ~~UMIE datasets on HuggingFace~~ ✓
- UMIE model

Towards Medical Foundation Model - a Unified Dataset for Pretraining Medical Imaging Models



Barbara Klaudel

Aleksander Obuchowski

Andrzej Komor

Kacper Rogala

Piotr Frąckowski

Kacper Knitter

Bartłomiej Kuchnowski

FRIDAY / 8 NOVEMBER @ 11:35 - 12:00
LECTURE HALL B

ML in PL
CONFERENCE 2024

UMIE datasets

GITHUB

HUGGINGFACE

RELEASED TODAY!

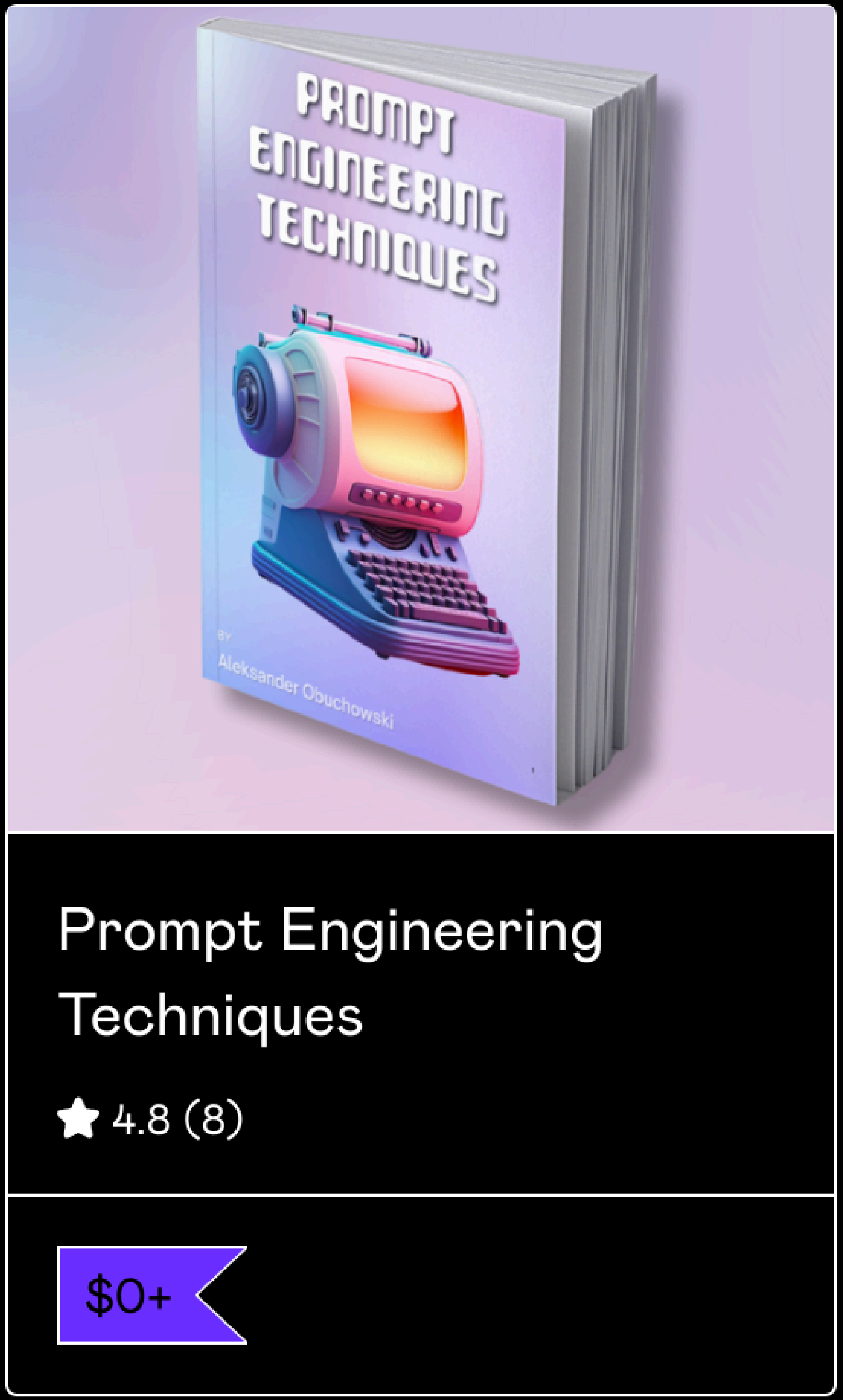
TheLion.AI

Universal
Medical
Image
Encoding

Eskulap
Polish Medical Large
Language Model

Interdisciplinary research group creating opensource AI-based solutions for healthcare

library.thelion.ai



Prompt Engineering Techniques

★ 4.8 (8)

$0+

Medical Computer Vision on Paper

★ 5.0 (2)

zł0+

Eskulap

Polish Medical Large Language Model

ALEKSANDER OBUCHOWSKI

MIKOŁAJ BADOCHA

KINGA MARSZAŁKOWSKA

MACIEJ GIERCZAK

BARBARA KLAUDEL

SATURDAY / 9 NOVEMBER
@ 10:30 - 11:30
POSTER SESSION