FAIR Data Spaces Final Event

Activities in **Demonstrator 4.3:** Cross-Platform FAIR Data Analysis PADME PHT

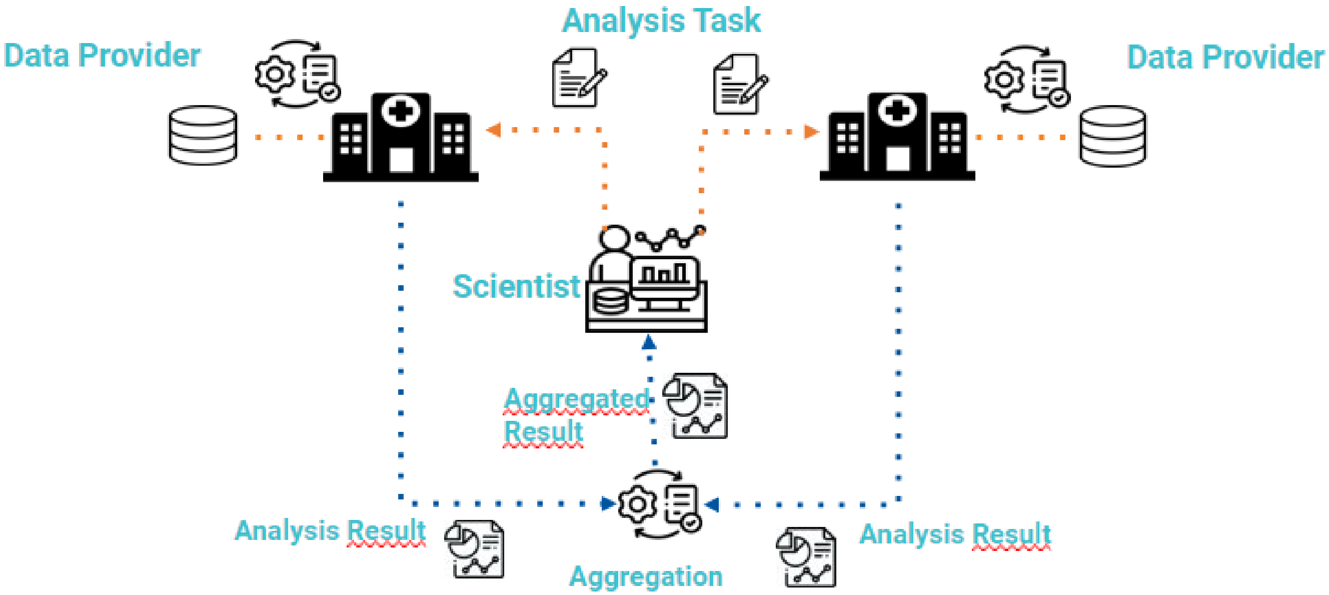
Yeliz Ucer Yediel, Muhammad Hamza Akhdar (Fraunhofer FIT),
Macedo Maia, Toralf Kirsten (University of Leipzig),
Mehrshad Jaberansary, Oya Beyan (University of Cologne)

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

FAIR Data Spaces

Cross-Platform FAIR Data Analysis PADME PHT

- Idea: “Bring the algorithms to the data” by using Distributed Analytics (DA)
- Benefits:
  - The data remains in the control of the data providers
  - Research can leverage otherwise inaccessible data
  - The results are made more robust by incorporating a variety of datasets.
- Provides ecosystem from the first idea to the analysis results
  - Central Components: Playground, Train Creator, Train Store House, Train Requester,
  - Client Software: PHT Station



PADME in a Nutshell

https://padme-analytics.de/ , https://docs.padme-analytics.de/

- Implementation of the PHT/FL concepts by using FAIR standards
- Result of a collaboration between four research institutes

Fraunhofer FIT · University of Cologne · Computer Science 5 - Information Systems and Databases | RWTH AACHEN UNIVERSITY · UNIVERSITÄT LEIPZIG

- Based on containerization technologies (www.docker.com), deployed on Kubernetes env.

docker · kubernetes

- Benefits:
  - Operating system agnostic
  - Data source and data structure agnostic
  - Programming-language agnostic

PHT PADME and EDC Integration

Overview

- Implement data-sharing capabilities within the PADME ecosystem using the EDC framework.
- Demonstrate two scenarios:
  - PADME as Data Provider.
  - PADME as Data Consumer.

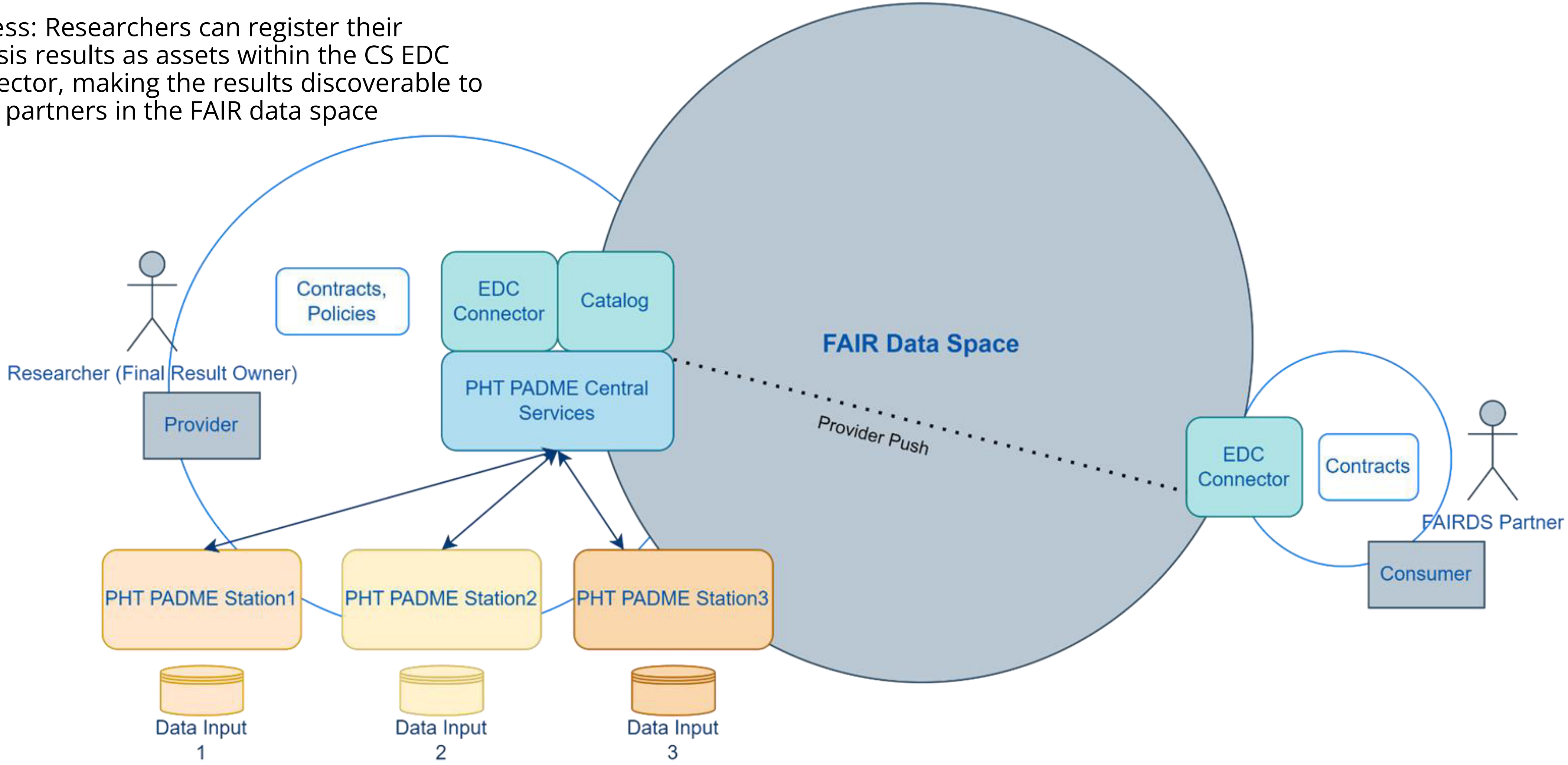
PHT PADME and EDC Integration (Cont‘d)

PADME as Data Provider (Scenario 1)

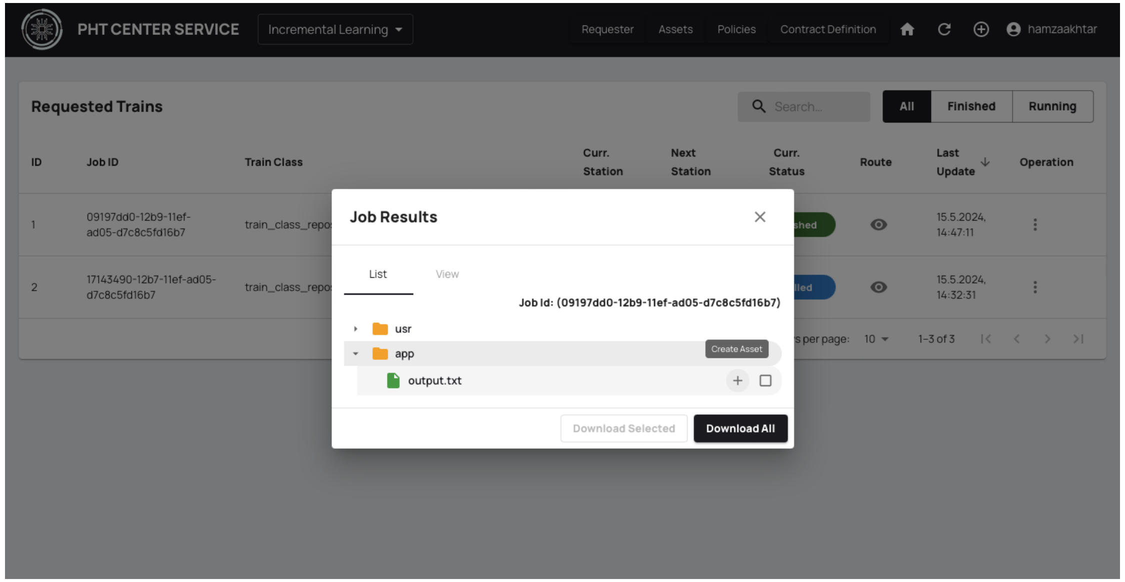
- Objective: Enable sharing of analysis results using EDC connectors.
- Solution:
  - Utilized sovity’s EDC Connector-as-a-Service.
  - Seamless integration facilitated by comprehensive documentation
- Workflow:
  - Researchers register the analysis results via a user-friendly interface in the Provider Connector.
  - Results are published in the data catalog, accessible by partners (e.g., Aruna) or via a federated catalog.
- Links:
  - Sovity Community Edition EDC
  - PADME as Data Provider Repository

EDC Integration into PADME - Provider Push

Process: Researchers can register their analysis results as assets within the CS EDC Connector, making the results discoverable to other partners in the FAIR data space



EDC Integration into PADME - Provider Push



Requested Trains

ID	Job ID	Train Class	Curr. Station	Next Station	Curr. Status	Route	Last Update	Operation
1	09197dd0-12b9-11ef-ad05-d7c8c5fd16b7	train_class_repo			shed		15.5.2024, 14:47:11	
2	17143490-12b7-11ef-ad05-d7c8c5fd16b7	train_class_repo			lled		15.5.2024, 14:32:31	

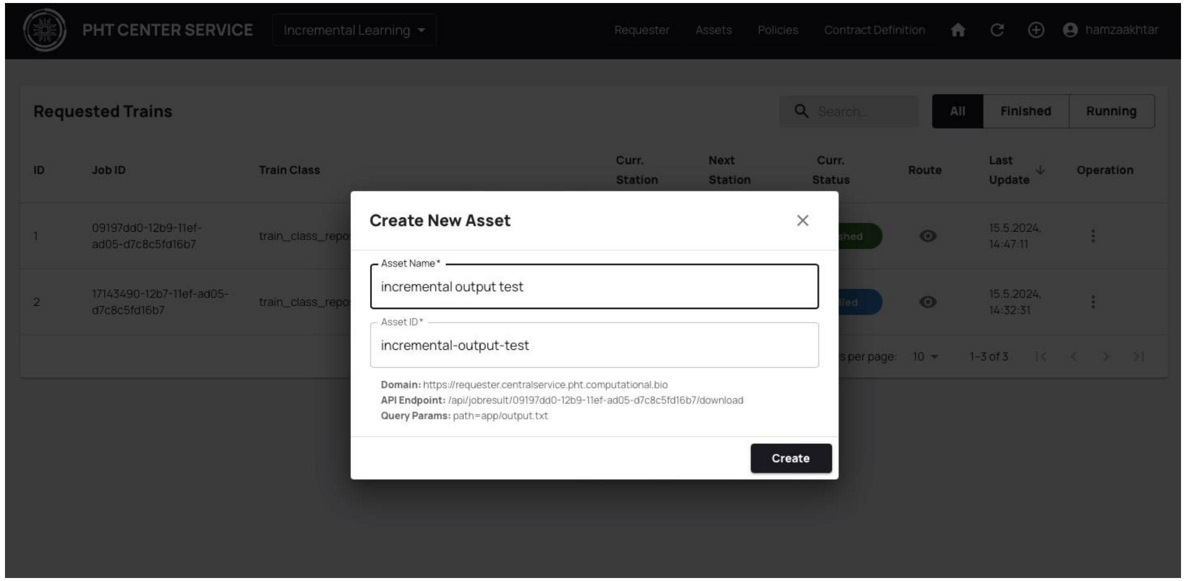
s per page: 10 1-3 of 3

Job Results

Job Id: (09197dd0-12b9-11ef-ad05-d7c8c5fd16b7)

- usr
- app
  - output.txt

Step 1.



Create New Asset

Asset Name*: incremental output test

Asset ID*: incremental-output-test

Domain: https://requester.centralservice.pht.computational.bio
API Endpoint: /api/jobresult/09197dd0-12b9-11ef-ad05-d7c8c5fd16b7/download
Query Params: path=app/output.txt

Step 2.

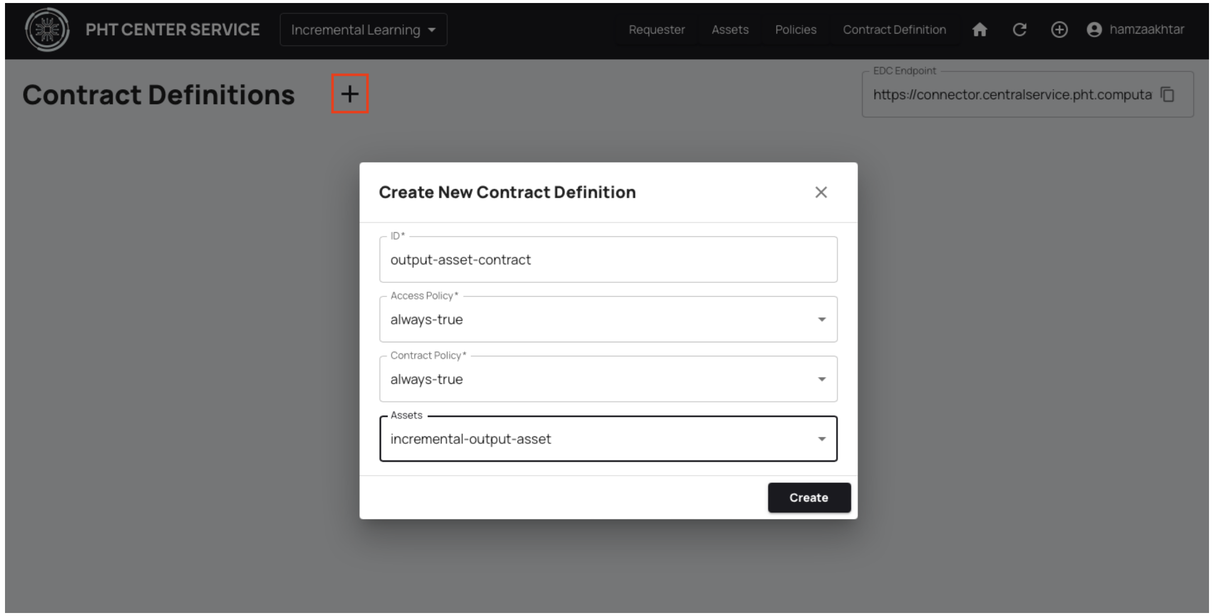
Steps:

1. Prepare analysis result to publish in PHT EDC Connector.
2. Add relevant metadata such as asset name and asset id.

EDC Integration into PADME - Provider Push

Steps (Cont’d):

3. Create Contract Definition to publish asset in EDC Catalog



ID*: output-asset-contract

Access Policy*: always-true

Contract Policy*: always-true

Assets: incremental-output-asset

Step 3.

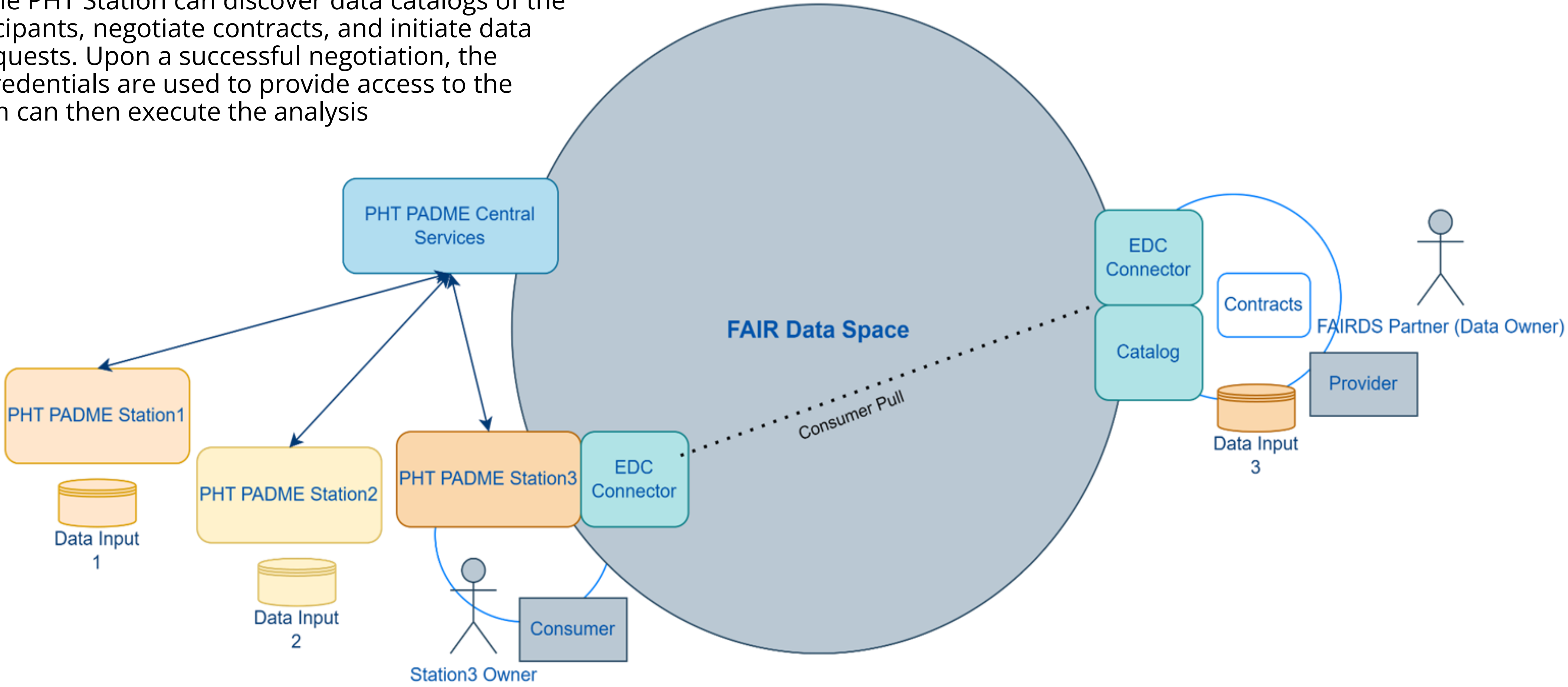
PHT PADME and EDC Integration (Cont‘d)

PADME as Data Consumer (Scenario 2)

- Objective: Enable researchers to consume data from external sources in the data space.
- Solution:
  - Opted for Eclipse Foundation’s EDC Connector due to limitations in sovity‘s EDC Connector free version.
  - Required additional integration effort but successfully set up the Consumer Connector.
- Workflow:
  - Consumer Connector parses external data catalogs, negotiates contracts, and consumes data.
  - Temporary credentials provided for data access after contract negotiations.
  - Credentials used for on-demand data access by analysis algorithms.
- Links:
  - Eclipse EDC Samples Repository
  - PADME as Data Consumer Repository

EDC Integration into PADME – Consumer Pull

Process: The PHT Station can discover data catalogs of the other participants, negotiate contracts, and initiate data transfer requests. Upon a successful negotiation, the provided credentials are used to provide access to the Train, which can then execute the analysis



EDC Integration into PADME – Consumer Pull

PADME as Data Consumer (Use Cases)

We developed two use cases as proof-of-concept to the consumer pull use case.

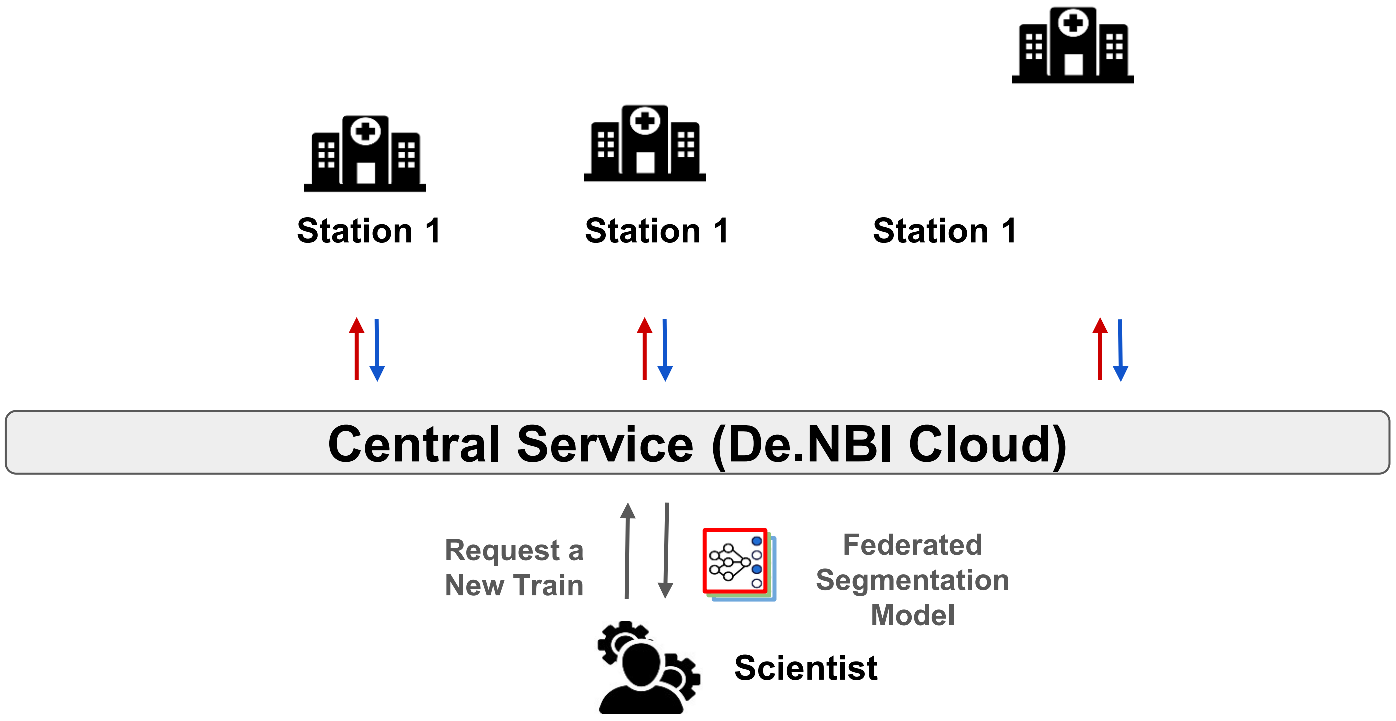
- Data Access from MinIO S3-compatible storage.
- Data Access from FHIR Server

Each use cases follows the same steps from analysis creation to execution.

1. Create the PADME PHT Train
2. Create Train Request
3. Get Contract Definitions
4. Negotiate Contract and Get Data Access Credentials
5. Execute Train
6. Check Results

PADME Use Cases

- During the FAIR Dataspaces project, we proposed two use cases in performing data analysis/machine learning training in distributed data:
  - Malaria cases prediction.
  - Federated Machine Learning on liver cancer segmentation based on 2D scans.



Station 1

Station 1

Station 1

Central Service (De.NBI Cloud)

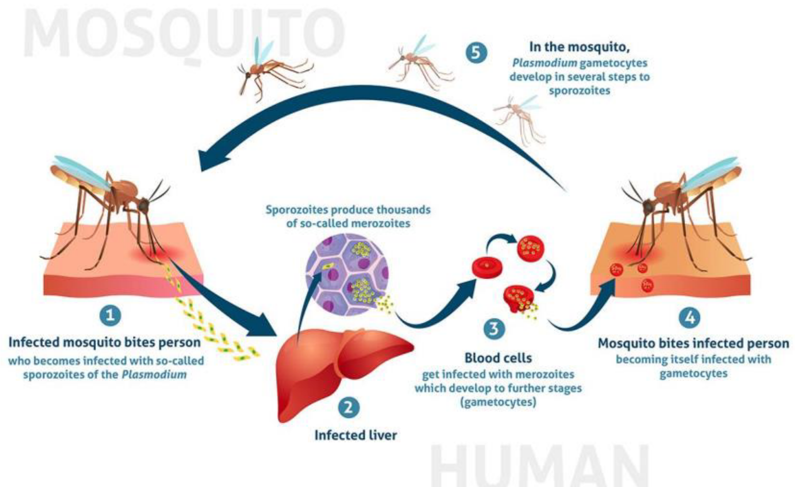
Request a New Train

Federated Segmentation Model

Scientist

Malaria Cases prediction use case

- Malaria is a life-threatening disease caused by parasites;
- Transmitted by the infected female Anopheles mosquitoes;
- WHO: 247 million malaria cases in 2021
- 87 malaria endemic countries
- Most of them in Central Africa
- ~409.000 malaria deaths


MOSQUITO

5 In the mosquito, *Plasmodium* gametocytes develop in several steps to sporozoites

1 Infected mosquito bites person who becomes infected with so-called sporozoites of the *Plasmodium*

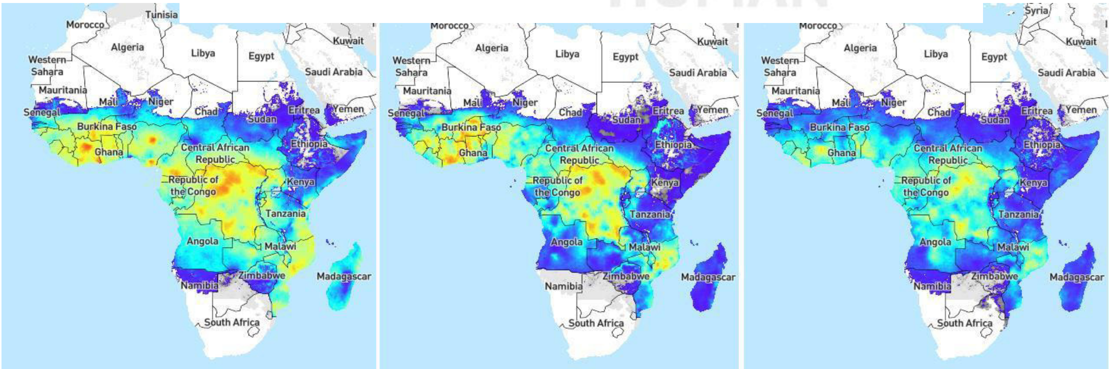
Sporozoites produce thousands of so-called merozoites

2 Infected liver

3 Blood cells get infected with merozoites which develop to further stages (gametocytes)

4 Mosquito bites infected person becoming itself infected with gametocytes

HUMAN

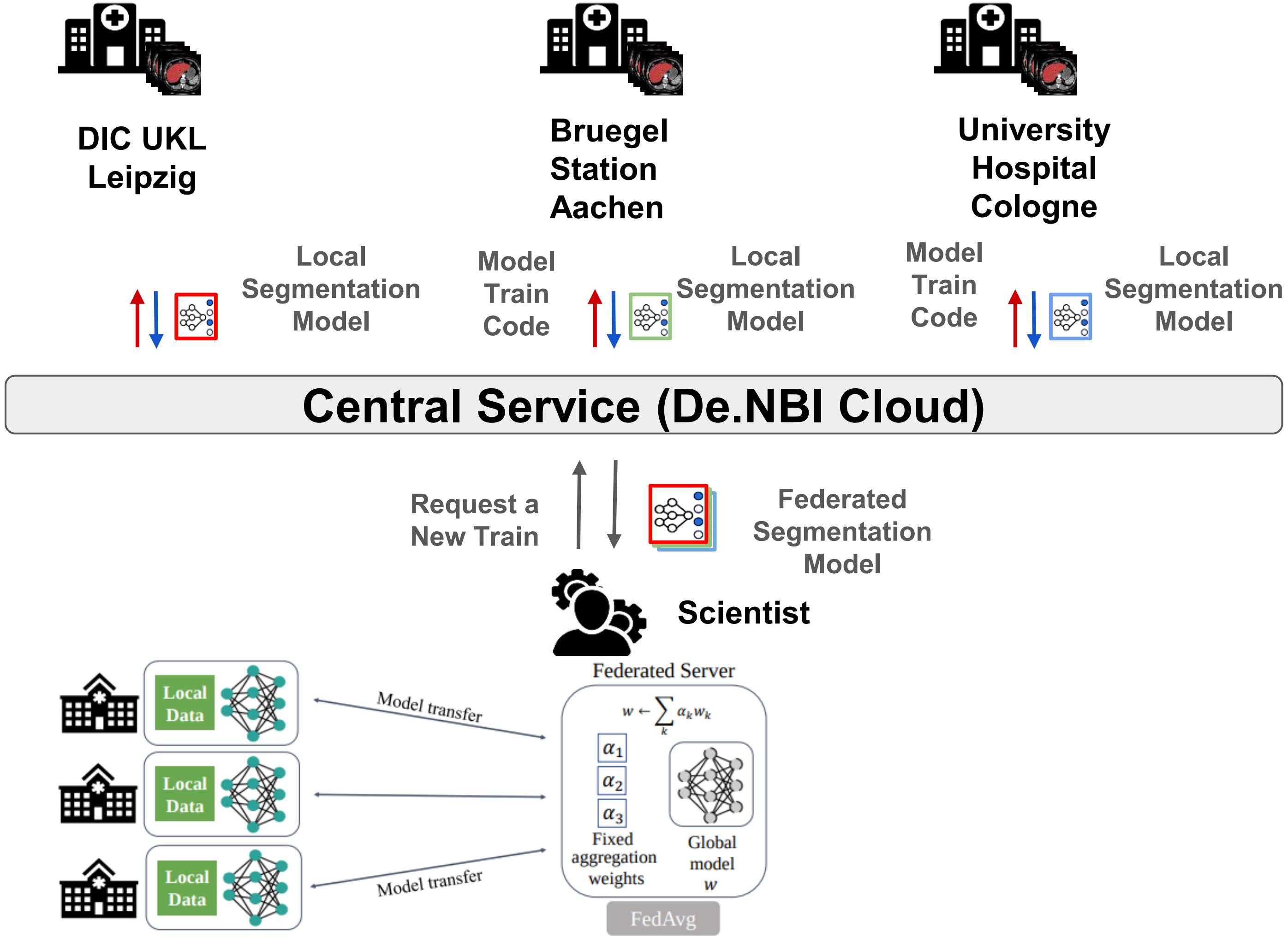


Malaria Cases prediction in PADME

- Data are distributed in multiple sources
- Integrating fragmented data to analyze differences at the country or larger regional levels poses a significant challenge.
- The study introduces a predictive analysis using incremental regression models:
  - To predict the number of malaria cases
  - Incremental models are designed
  - The models take into account time series data from previous years to enhance predictive accuracy.
- Three distinct sources categorize the data by the WHO region attribute.
  - **Source 1 (Leipzig University):** Malaria data from Eastern Mediterranean and Africa regions.
  - **Source 2 (DeNBI Cloud):** Contains Malaria data from the Americas and Europe.
  - **Source 3 (University of Cologne):** Contains Malaria data from Southeast Asia, Western Pacific.
- Each data source is equipped with a PHT Station, facilitating data inclusion in the analysis.

Federated Learning over Multiple PADME PHT Stations

- Federated learning involves training a central model using data distributed across multiple Stations (Client/Provider) and Central Service (Server/Consumer)
- Local models are trained on each PHT Station
- Each local model are sending to Central Service
- The application of a aggregation function over each local model weights determines a federated learning
- The federated model are sending back to each Station and retrained in the next round



DIC UKL Leipzig

Bruegel Station Aachen

University Hospital Cologne

Local Segmentation Model

Model Train Code

Local Segmentation Model

Model Train Code

Local Segmentation Model

Central Service (De.NBI Cloud)

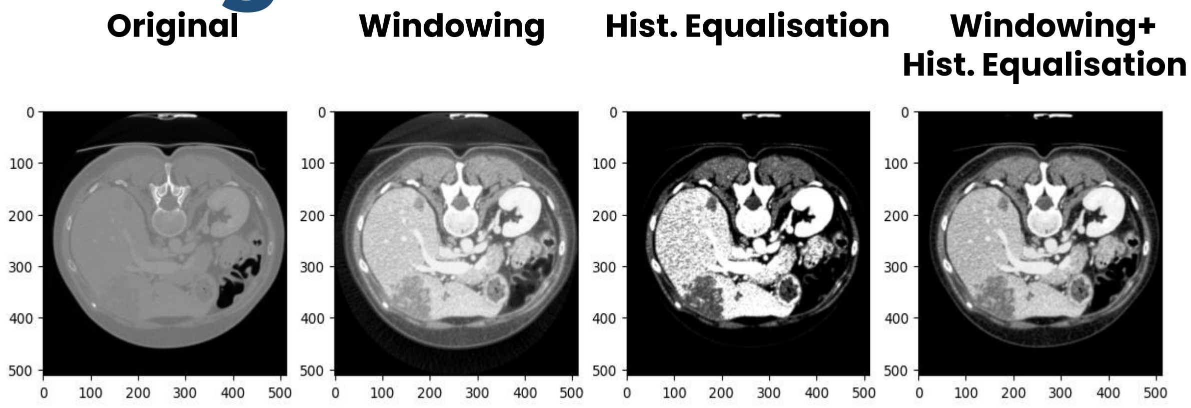
Request a New Train

Federated Segmentation Model

Scientist

Use case: Liver Tumours Segmentation

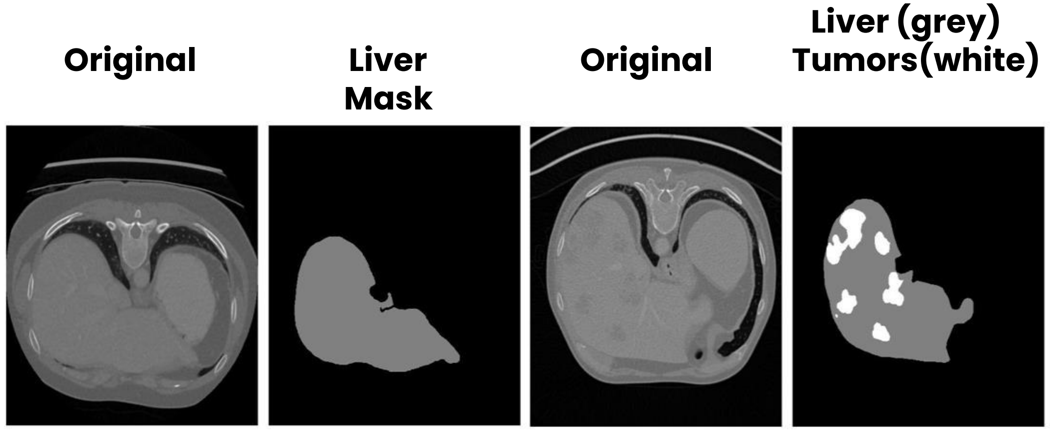
- Problem Statement
  - Based on The American Cancer Society’s:
    - About 41,630 new liver cancer cases in US were diagnosed in 2023
    - About 29,840 people have died of these cancers
  - AI-based approaches helps to early detect tumours
  - However, the data can be distributed in different sources (e.g., hospitals)
  - Data access depends on distinct rules or regulations from each data provider
- Possible Solution
  - Explore Computed Tomography (CT) scans for image segmentation
  - Federated learning models over data from multiple data providers


Original

Windowing

Hist. Equalisation

Windowing+ Hist. Equalisation


Original

Liver Mask

Original

Liver (grey) Tumors(white)

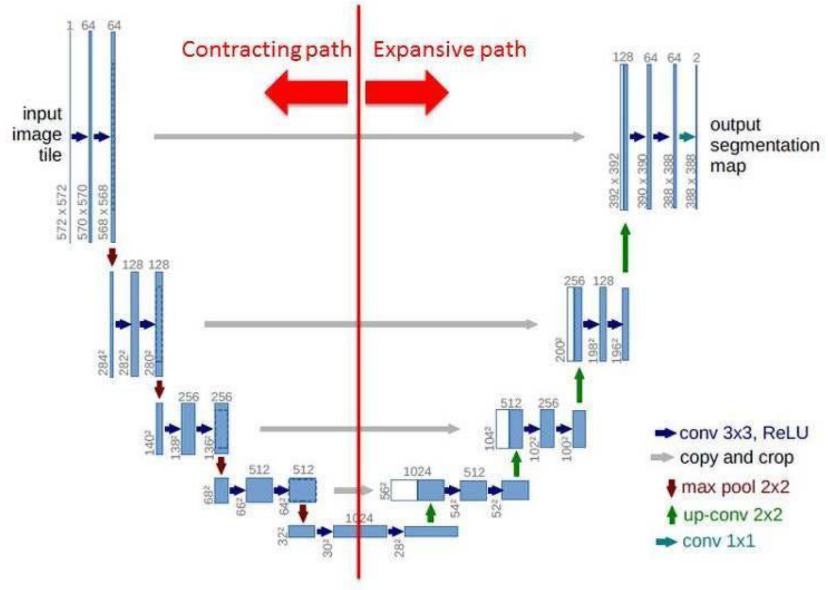
No cancer

With cancer

Image Segmentation Models based on UNET

- **Encoder-Decoder Architecture:** U-Net uses a symmetrical structure with an encoder for downsampling (capturing context) and a decoder for upsampling (reconstructing spatial details).
- **Encoder or Contracting path (Downsampling)**
- **Decoder or Expansive path (Upsampling)**
- **Medical Image Segmentation:** Originally designed for medical imaging, U-Net is widely used for segmenting organs, tumors, and other structures in MRI, CT, and microscopy images.

U-NET Network Architecture



Use case: Liver Tumours Segmentation

- Liver CT Scan Data for Segmentation:
  - The CT Liver dataset consists of 3D NIFTI images or 2D DICOM scans
  - Segmentation masks are the labels
  - Segmentation models for medical scans:
    - UNET
    - nn-UNET
    - Dense-UNET

Papers & Presentations

- Maia, M.; Jaberansary, M.; Ucer, Y.; Beyan, O.; Kirsten, T. Providing Publicly Available Medical Data Access under FAIR Principles for Distributed Analysis. In Proceedings of the 67. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 13. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF), 2022.
- Mehrshad Jaberansary, Macedo Maia, Yeliz Ucer Yediel, Oya Beyan, and Toralf Kirsten. 2023. Analyzing distributed medical data in FAIR data spaces. In Companion Proceedings of the ACM Web Conference 2023(WWW '23 Companion). ACM, 1480–1484. https://doi.org/10.1145/3543873.3587663
- Maia, M.; Jaberansary, M.; Ucer, Y.; Beyan, O.; Kirsten, T. Incremental Machine Learning using Distributed Data Processing Techniques for Malaria Data Across Multiple Online Sources. In Proceedings of the 68. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 14. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF), 2023.
- Maia, M.; Jaberansary, M.; Hamza Akhtar, Muhammad; Ucer, Y.; Beyan, O.; Kirsten, T. Training Federated Liver Cancer Image Segmentation Models using PHT Infrastructure. In Proceedings of the 69. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS), 15. Jahreskongress der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e. V. (TMF), 2024.

Thank you

Thank you for your interest!

FAIR Data Spaces

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung