PrivateAIM

Sichere verteilte Auswertung medizinischer Daten mit KI-Methoden

16.01.2024

Prof. Oliver Kohlbacher

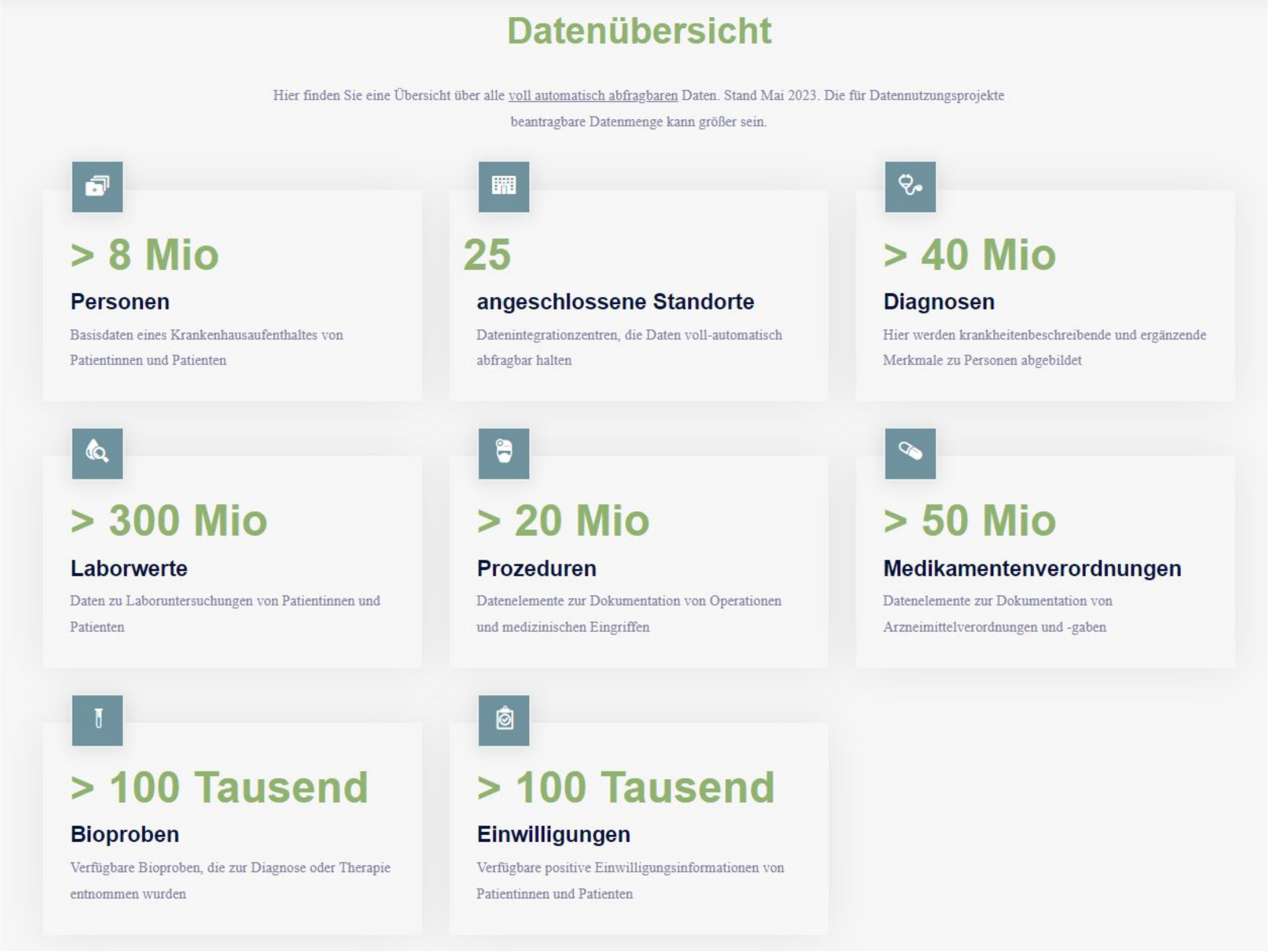
Dept. für IT und Angewandte Medizininformatik

Co-Coordinator PrivateAIM

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Universitätsklinikum
Tübingen

The German Research Data Portal for Health (FDPG)



Datenübersicht

Hier finden Sie eine Übersicht über alle voll automatisch abfragbaren Daten. Stand Mai 2023. Die für Datennutzungsprojekte beantragbare Datenmenge kann größer sein.

> 8 Mio
Personen
Basisdaten eines Krankenhausaufenthaltes von Patientinnen und Patienten

25
angeschlossene Standorte
Datenintegrationzentren, die Daten voll-automatisch abfragbar halten

> 40 Mio
Diagnosen
Hier werden krankheitenbeschreibende und ergänzende Merkmale zu Personen abgebildet

> 300 Mio
Laborwerte
Daten zu Laboruntersuchungen von Patientinnen und Patienten

> 20 Mio
Prozeduren
Datenelemente zur Dokumentation von Operationen und medizinischen Eingriffen

> 50 Mio
Medikamentenverordnungen
Datenelemente zur Dokumentation von Arzneimittelverordnungen und -gaben

> 100 Tausend
Bioproben
Verfügbare Bioproben, die zur Diagnose oder Therapie entnommen wurden

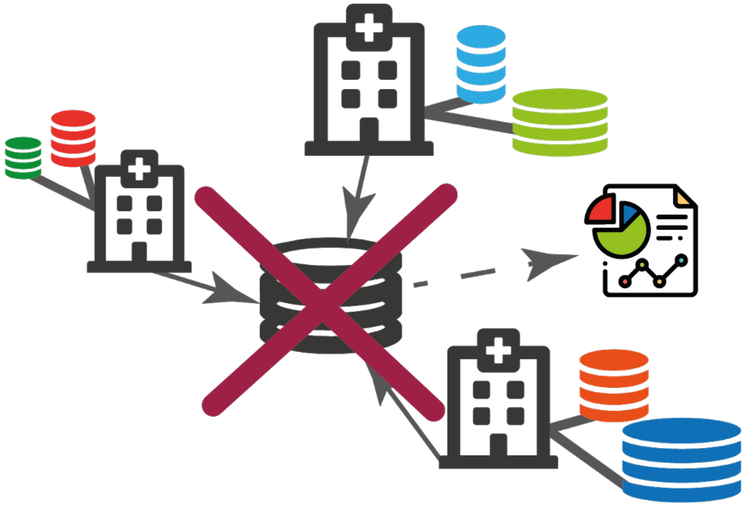
> 100 Tausend
Einwilligungen
Verfügbare positive Einwilligungsinformationen von Patientinnen und Patienten

- DICs in the MII provide data on millions of patients across dozens of sites
- Data is findable centrally through the FDPG portal
- Data can be either provided **centrally** (pooled) or for **federated analysis**

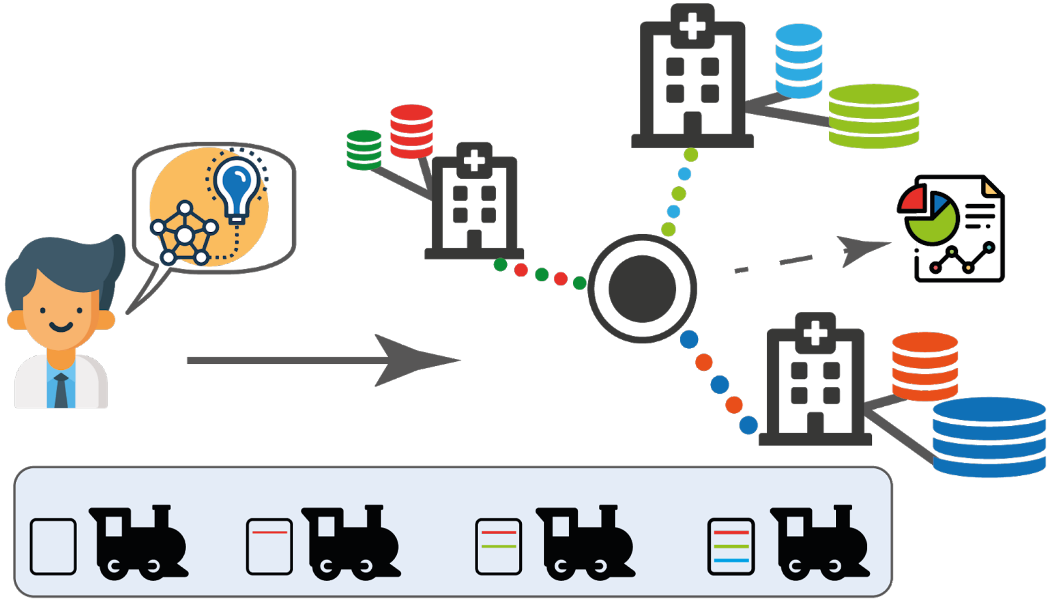
FORSCHUNGS-
DATENPORTAL
FÜR GESUNDHEIT

Federated Analysis – Key Idea

Data Pooling



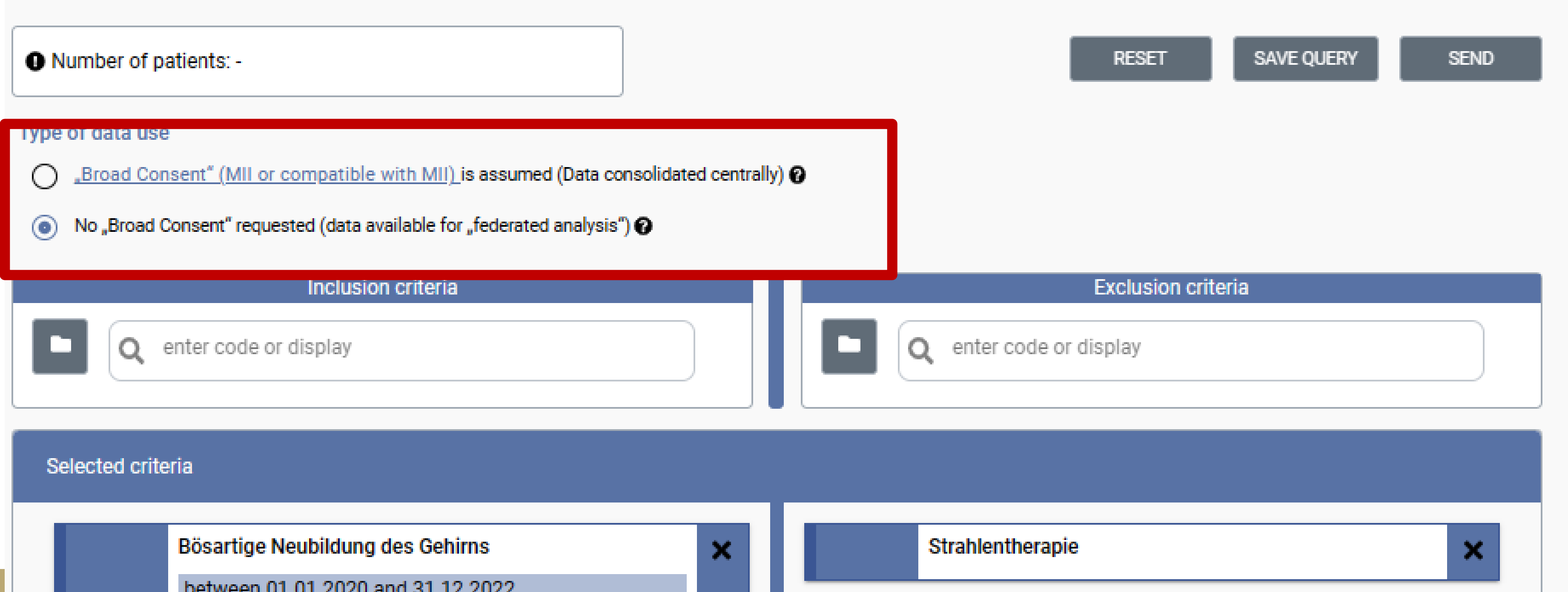
Federated Analysis



- **Data**: Stays at the hospital (station)
- **Analysis**: Represented by trains. Travel from one station to another to iteratively build and update a model

Central vs. Federated

- Legal basis for the central pooling is the national broad consent (MII BC)
- Vast majority of the data (98+%) are not accessible under the MII BC
- In this case, federated analysis is a **technical solution to a legal problem**



Number of patients: -

Type of data use

○ „Broad Consent“ (MII or compatible with MII) is assumed (Data consolidated centrally)

● No „Broad Consent“ requested (data available for „federated analysis“)

Bösartige Neubildung des Gehirns

between 01.01.2020 and 31.12.2022

Strahlentherapie

Personal Health Train (PHT) - Background

Manifesto of PHT from DTL

- advance healthcare and biomedical science through shared infrastructure
- keep control over data at each local site
- machine-readability at the core
- advance data analysis and medical decision making

DTL
DUTCH TECHCENTRE FOR LIFE SCIENCES

GO FAIR PHT in Implementation Network: German Chapter

Goals

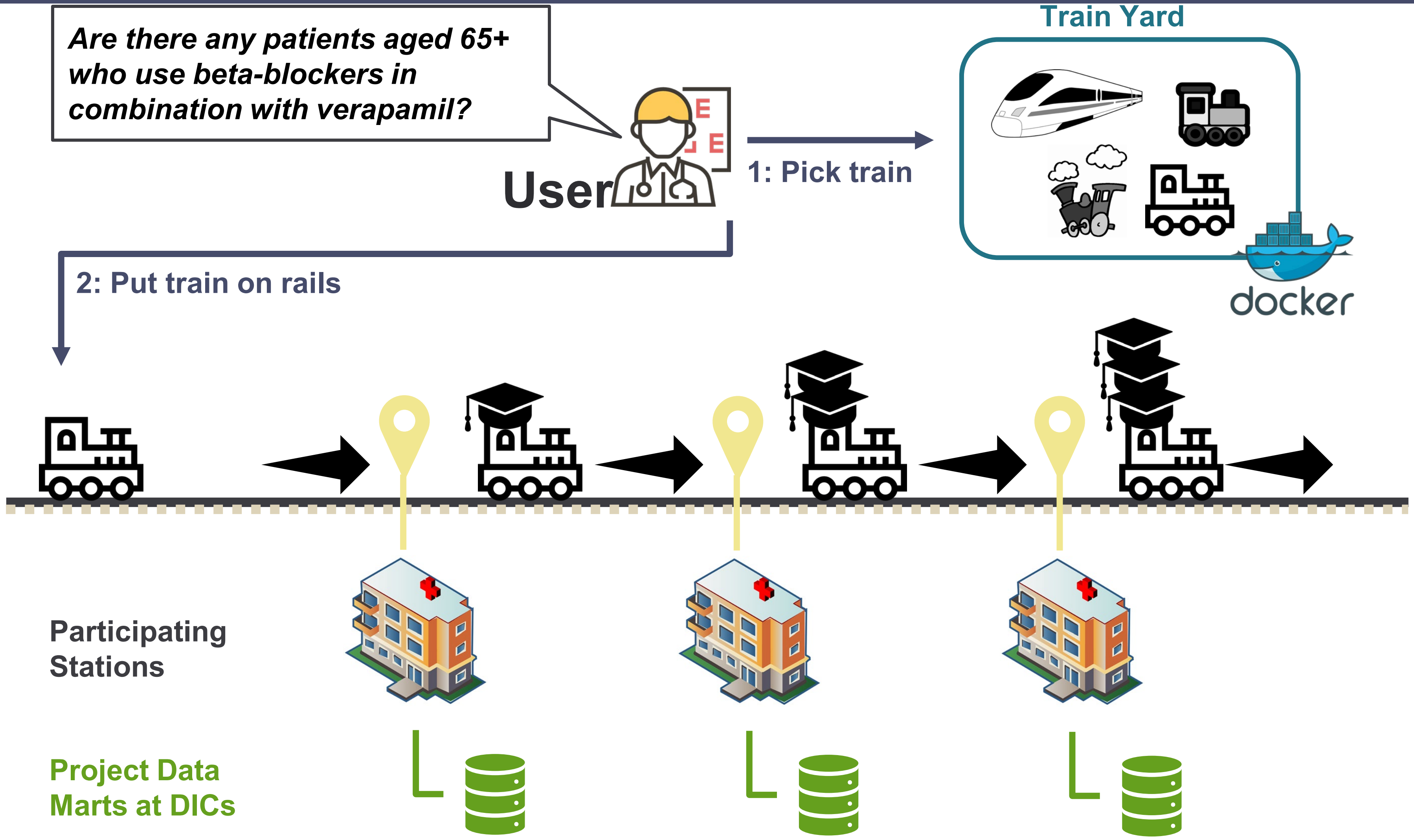
- a common core infrastructure
- set of standards, guidelines, specifications
- reference implementations

GO FAIR

DIFUTURE
Data Integration for Future Medicine

SMITH
Smart Medical Information Technology for Healthcare

PHT – Concepts



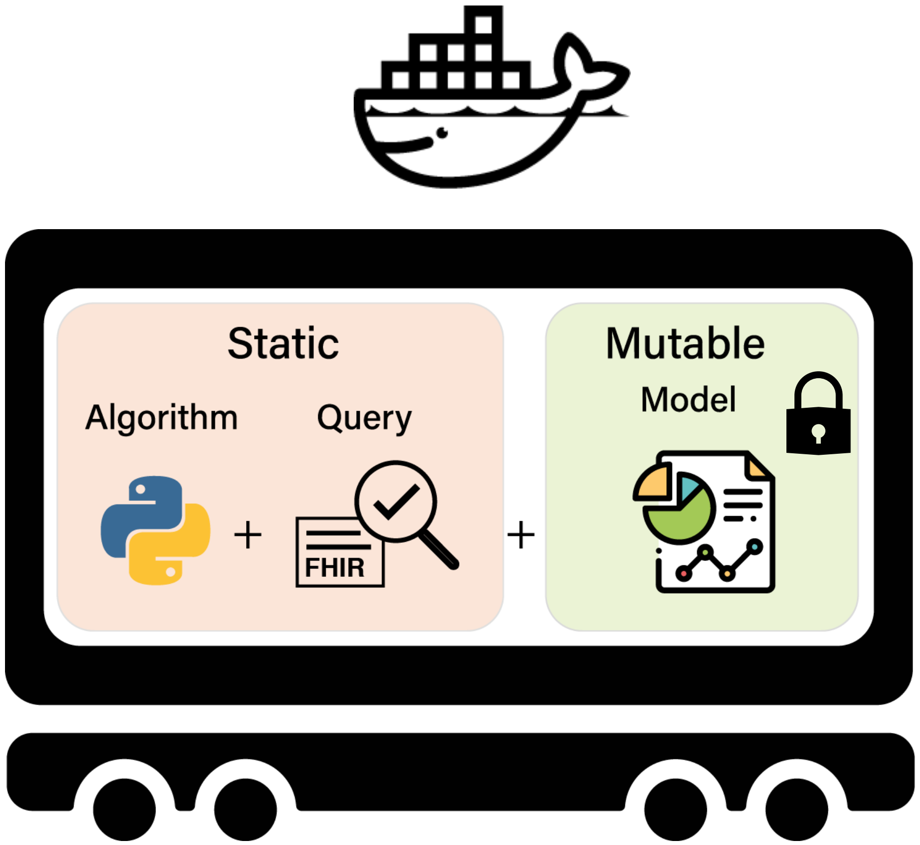
Are there any patients aged 65+ who use beta-blockers in combination with verapamil?

What is a train?

- Trains are container images (Docker), which include the following:

- operating system structures (Linux) and runtimes for scripts
- Train Library (FHIR access, train endpoints)
- algorithm and query (static during execution)
- results, created during execution (dynamic)

- **Privacy of data and results**
- Stations detect manipulation of trains, access to data with train library, results and models are encrypted using envelope encryption
- Paillier cryptosystem for adding and multiplying sensitive data.
- Methods such as k-anonymity, differential privacy as extension of train

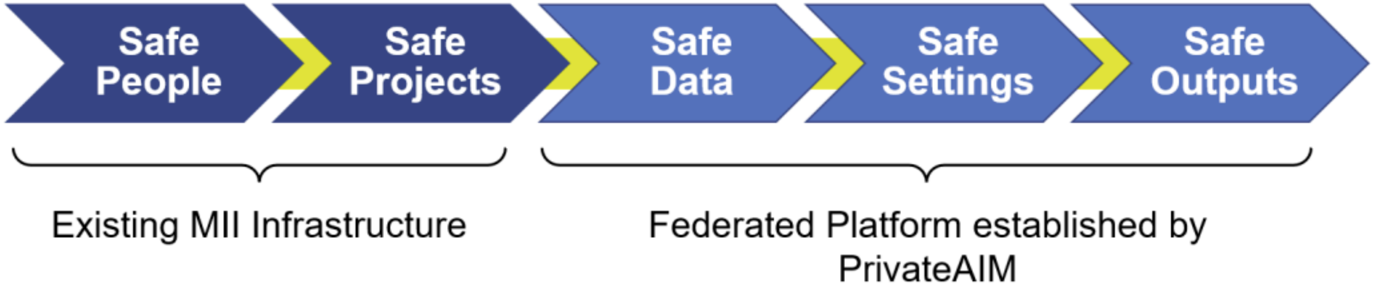


Project Overview

- Module 3 – Method Platform within the coming phase of the MII
- Goal of the project:

> *„The goal of PrivateAIM is to develop a federated machine learning (ML) and data analytics platform for the Medical Informatics Initiative (MII), where analyses come to the data instead of data coming to the analyses."*

- "Code to data" paradigm – data remains where it is to reduce potential privacy impact and resolve legal issues



Safe People

Safe Projects

Safe Data

Safe Settings

Safe Outputs

Existing MII Infrastructure

Federated Platform established by PrivateAIM

PrivateAIM

PrivateAIM - Consortium

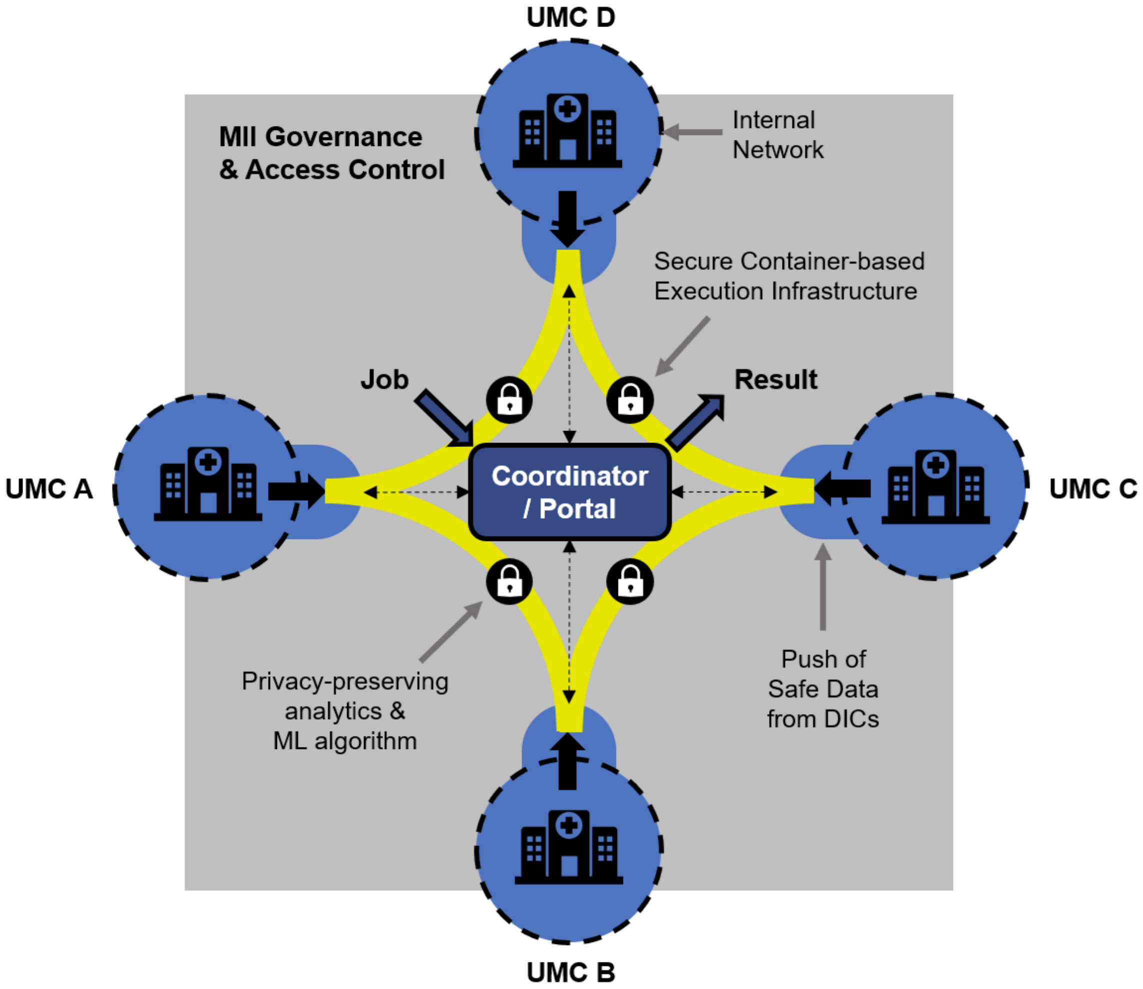
PrivateAIM

- 15 Participants from all four MII consortia (and beyond)
- Coordinators
  - Oliver Kohlbacher (U Tübingen)
  - Fabian Prasser (Charité)
  - Daniel Rückert (TU Munich)
- Three associated junior research groups
  - Mete Akgün - Medical Data Privacy and Privacy-Preserving ML on Healthcare Data (MDPPML) (Tübingen)
  - Michael Kamp – Trustworthy Machine Learning (Essen)
  - Björn Schreiweis – Medical Informatics (Kiel)

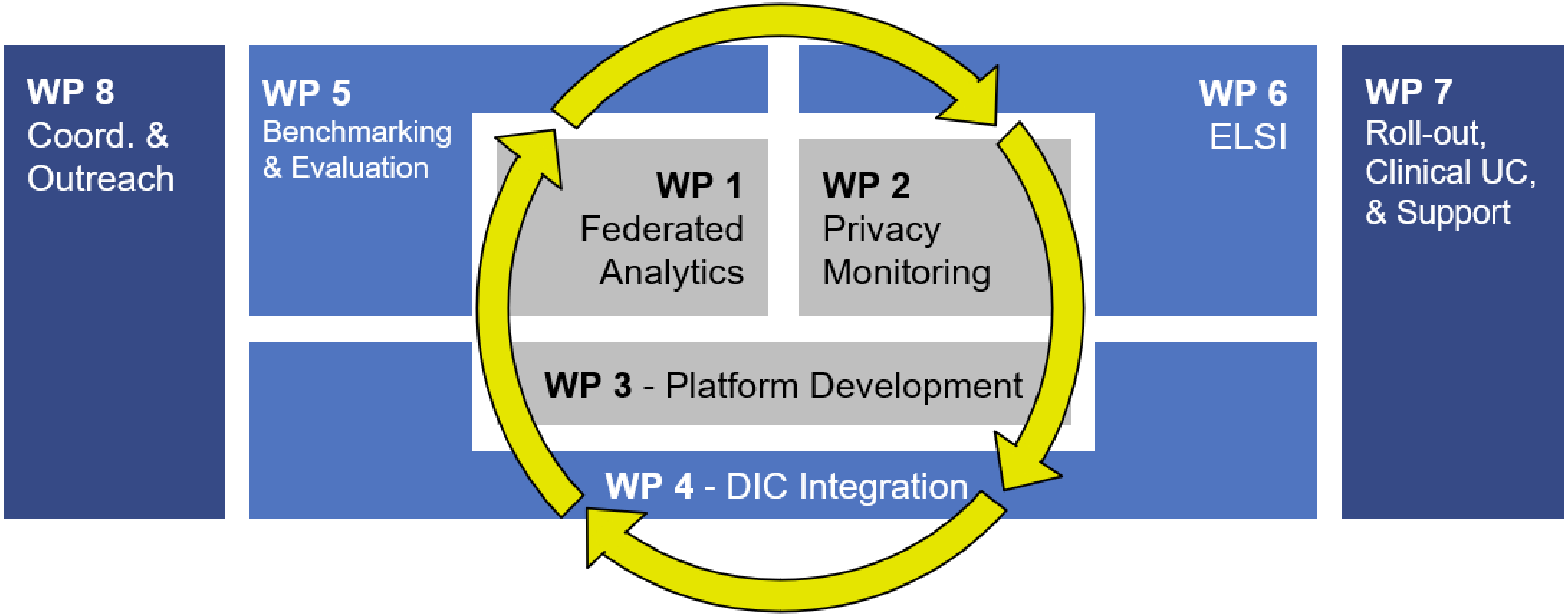
Charité - Universitätsmedizin Berlin (Charité)	Prof. Dr. Fabian Prasser
Helmholtz Center for Information Security (CISPA)	Prof. Dr. Mario Fritz
Deutsches Krebsforschungszentrum (DKFZ)	Dr. Ralf Omar Floca
University of Tübingen (EKUT)	Prof. Dr. Nico Pfeifer
Ludwig-Maximilians-Universität München (LMU)	Prof. Dr. Ulrich Mansmann
Technology, Methods, and Infrastructure for Networked Medical Research (TMF)	Dr. Sebastian C. Semler
Technische Universität München (TUM)	Prof. Dr. Daniel Rückert
Friedrich-Alexander-Universität Erlangen-Nürnberg (UKER)	Prof. Dr. Thomas Ganslandt
University of Freiburg (UKFR)	Prof. Dr. Harald Binder
University Hospital Heidelberg (UKHD)	Prof. Dr. Christoph Dieterich
University of Cologne (UKK)	Prof. Dr. Oya Beyan
Leipzig University Medical Center (UKL)	Prof. Dr. Toralf Kirsten
University Hospital Tübingen (UKT)	Prof. Dr. Oliver Kohlbacher
Ulm University (UKU)	Prof. Dr. Hans Kestler
Medical Faculty Mannheim, Heidelberg University (UMM)	Prof. Dr. Martin Lablans

PrivateAIM – Key Ideas

- Make major contributions in
  - Methods for federated machine learning
  - Privacy guarantees for federated analytics
  - Real-world platform for privacy-preserving analytics
- Deploy these ideas in a consistent platform across the MII sites
- Support other (clinical) use cases within the MII with the platform


UMC D
MII Governance & Access Control
Internal Network
Secure Container-based Execution Infrastructure
Job
Result
UMC A
Coordinator / Portal
UMC C
Privacy-preserving analytics & ML algorithm
Push of Safe Data from DICs
UMC B

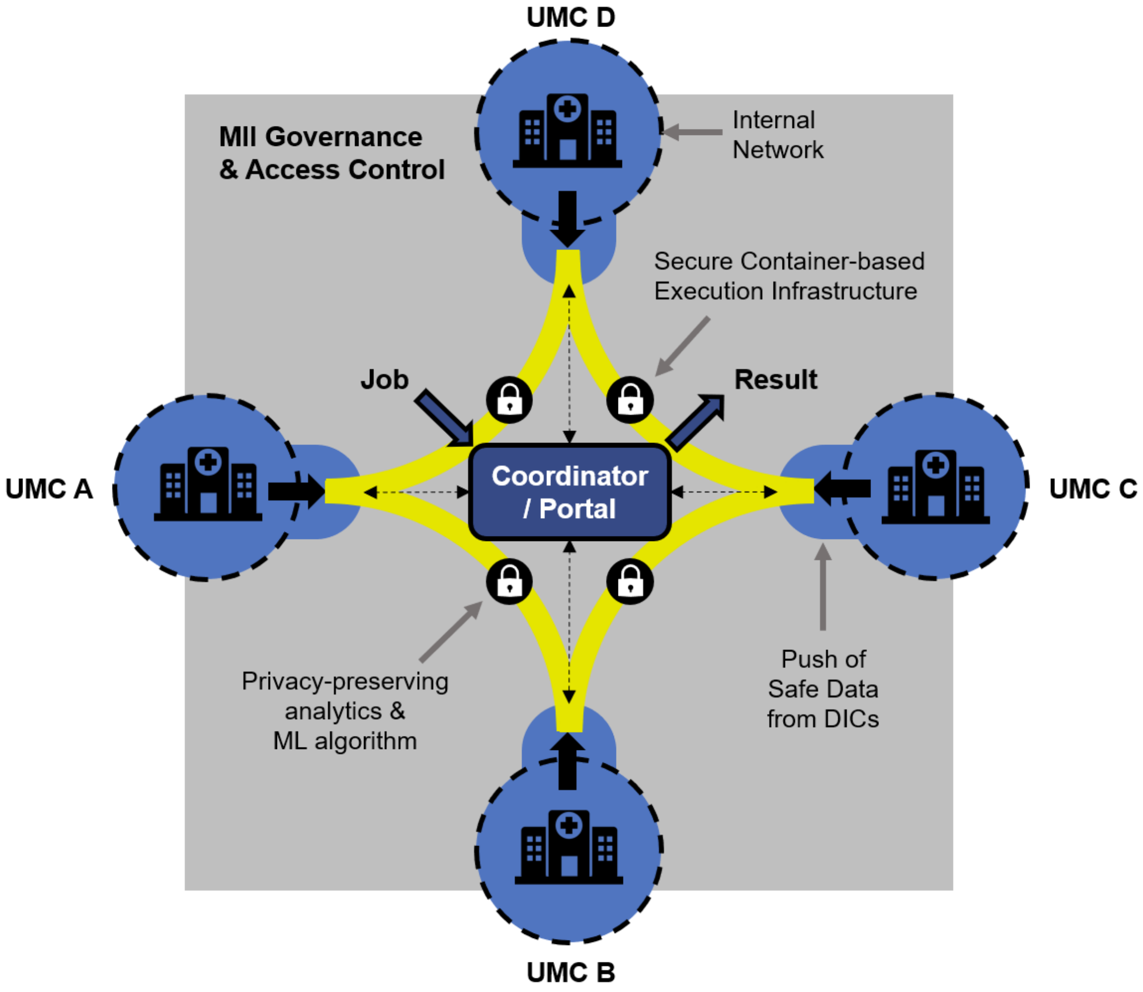
PrivateAIM – Work Packages



PrivateAIM – WP1 – Federated Analytics and ML

Federated Analytics and Machine Learning (WP1)

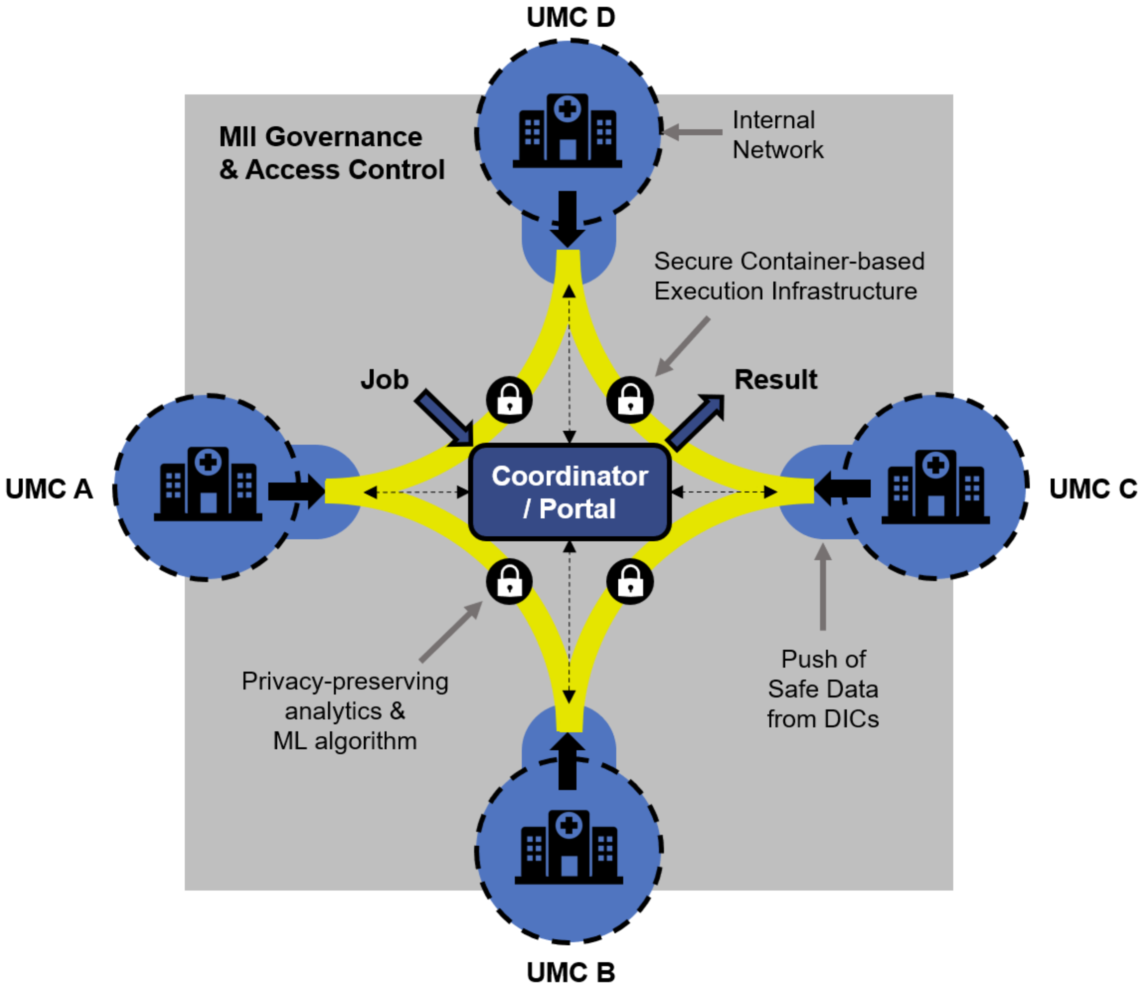
- Federated machine learning approaches that can deal with vertically and horizontally partitioned data (including non-independent and identically distributed data)
- Methods for zero/few shot learning approaches for domain adaptation/transfer learning on multi-modal medical data
- Approaches for balancing and resolving trade-offs between privacy and utility as well as privacy, fairness and robustness of ML models.


UMC D
MII Governance & Access Control
Internal Network
Secure Container-based Execution Infrastructure
Job
Result
UMC A
Coordinator / Portal
UMC C
Privacy-preserving analytics & ML algorithm
Push of Safe Data from DICs
UMC B

PrivateAIM – WP2 – Federated Analytics and ML

Privacy Monitoring and Guarantees (WP2)

- Privacy frameworks combining guarantees for combinations of different types of data and multi-modal datasets.
- Translation of privacy accounting, enforcement and monitoring techniques into practical biomedical settings while maintaining utility.
- Certification of components and their privacy properties to improve trust and make transparent privacy protection under the honest-but-curious model.

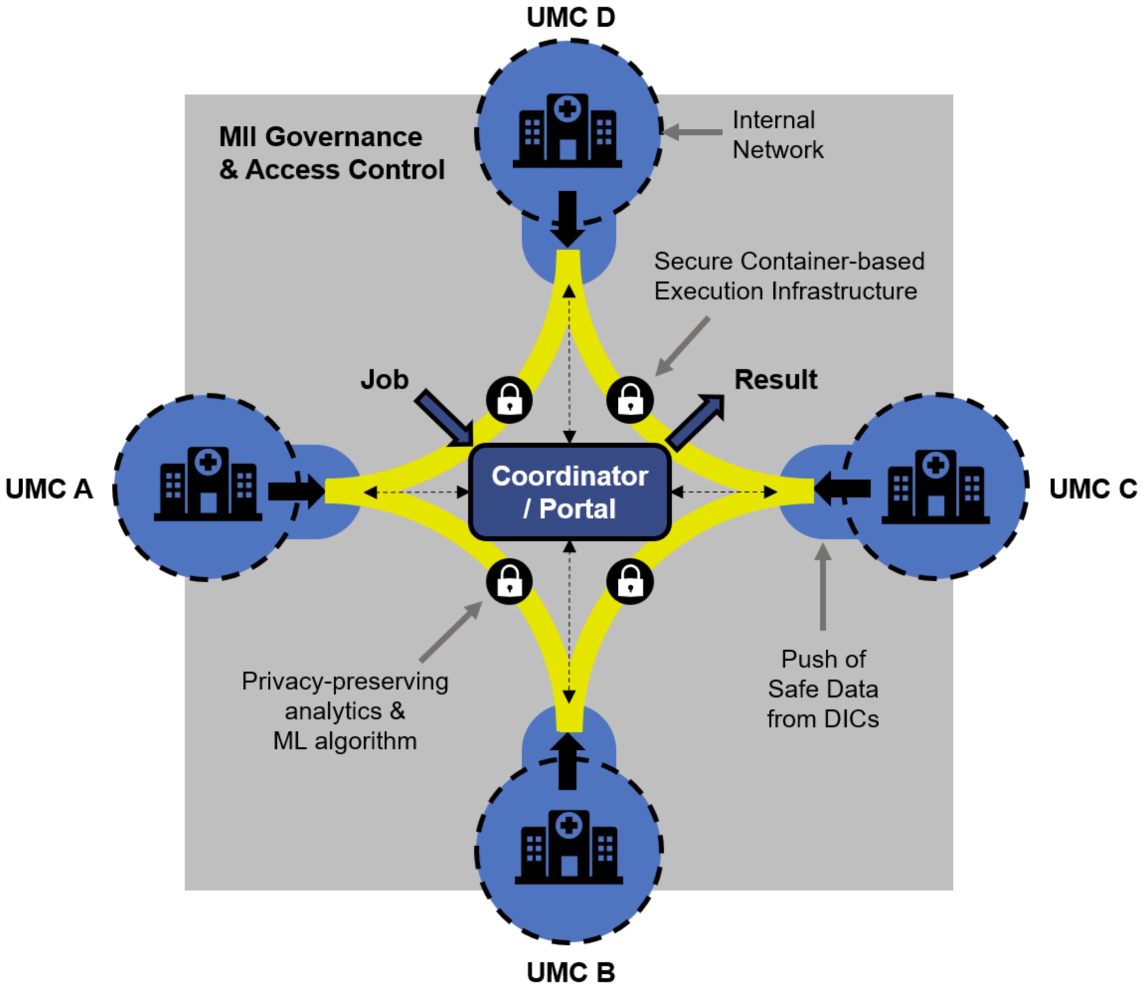

UMC D
MII Governance & Access Control
Internal Network
Secure Container-based Execution Infrastructure
Job
Result
UMC A
Coordinator / Portal
UMC C
Privacy-preserving analytics & ML algorithm
Push of Safe Data from DICs
UMC B

PrivateAIM – WP3 – Federated Analytics Platform

Federated Analytics Platform (WP3)

- Scalable infrastructures for federated and reproducible processing of clinical, omics, and imaging data with light-weight containerized components.
- Design and integrate interoperability mechanisms with other distributed analysis platforms

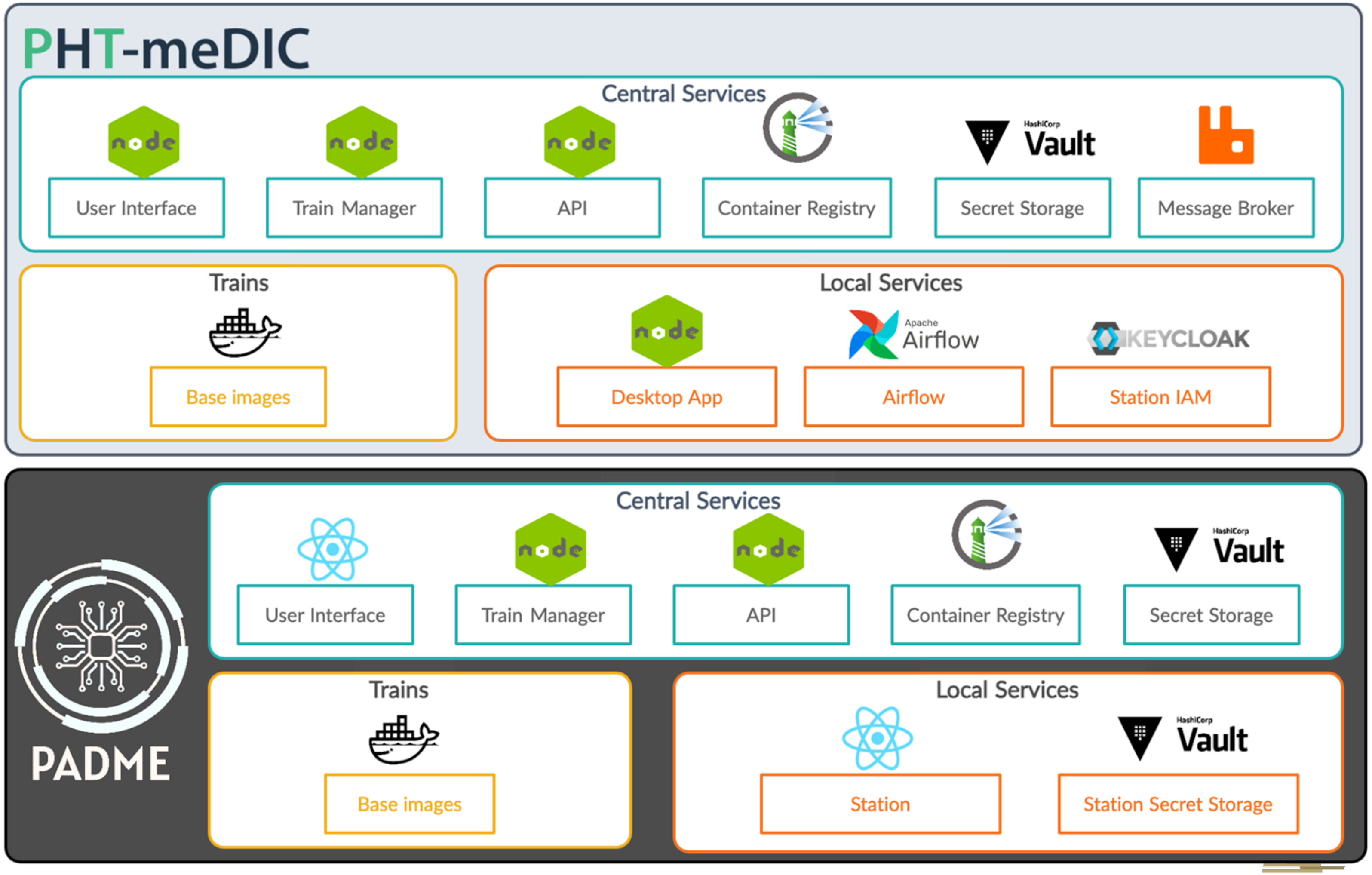
The **Federated Learning and Analytics in MEdicine (FLAME) platform**, developed within this project and compiled in WP3, forms the core of the project and brings together the fundamental research questions addressed in WP1 and WP2 with the practical application within the MII in WP4.


UMC D
MII Governance & Access Control
Internal Network
Secure Container-based Execution Infrastructure
Job
Result
UMC A
Coordinator / Portal
UMC C
Privacy-preserving analytics & ML algorithm
Push of Safe Data from DICs
UMC B

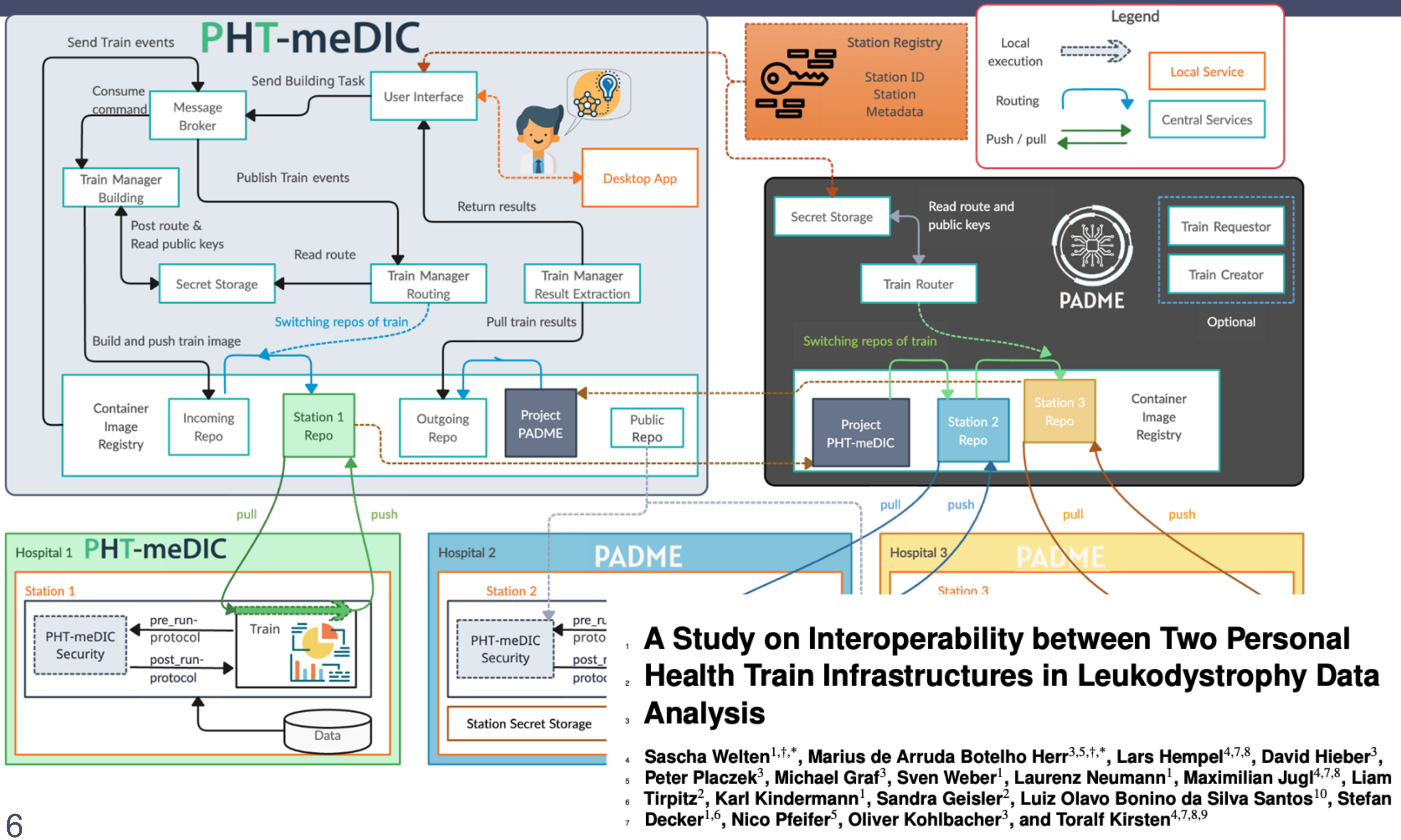
PHT – Two Implementations

- Both DIFUTURE and SMITH have been developing software inspired by the PHT concept: PHT-meDIC and PADME
- We have been working on integrating the two platforms (and others) over the past years
- While the core ideas (trains = containers, secrets in a Vault instance, …) are the same, the implementations differ

- PHT-meDIC: https://personalhealthtrain.de/
- PADME: https://websites.fraunhofer.de/PersonalHealthTrain/

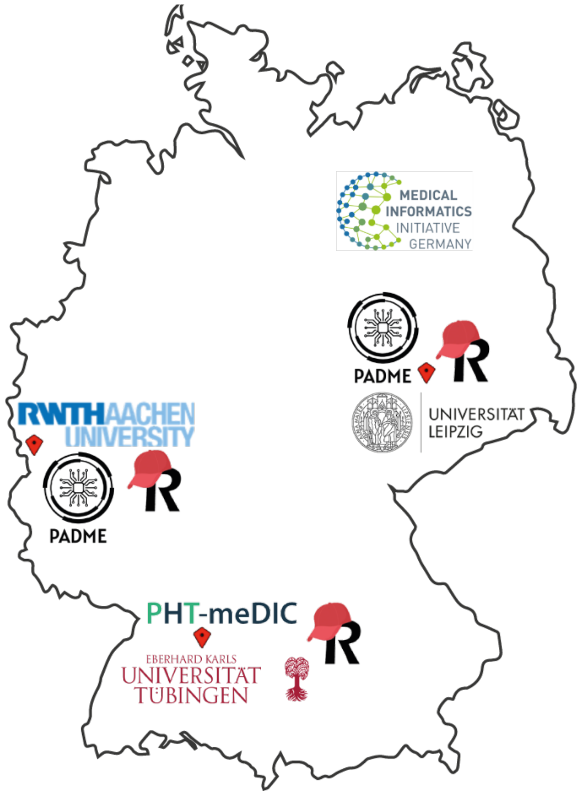


Interoperability



1 **A Study on Interoperability between Two Personal**
2 **Health Train Infrastructures in Leukodystrophy Data**
3 **Analysis**

4 **Sascha Welten[1,†,*], Marius de Arruda Botelho Herr[3,5,†,*], Lars Hempel[4,7,8], David Hieber[3],**
5 **Peter Placzek[3], Michael Graf[3], Sven Weber[1], Laurenz Neumann[1], Maximilian Jugl[4,7,8], Liam**
6 **Tirpitz[2], Karl Kindermann[1], Sandra Geisler[2], Luiz Olavo Bonino da Silva Santos[10], Stefan**
7 **Decker[1,6], Nico Pfeifer[5], Oliver Kohlbacher[3], and Toralf Kirsten[4,7,8,9]**


MEDICAL INFORMATICS INITIATIVE GERMANY
PADME
R
UNIVERSITÄT LEIPZIG
RWTH AACHEN UNIVERSITY
PADME
R
PHT-meDIC
R
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

FLAME

The FLAME Platform

Federated
Learning and
Analytics in
MEdicine

PrivateAIM

Current Progress

- Initial code bases of PADME and PHT-meDIC exist - but we stopped coding in the existing projects
- In a workshop in Tübingen we made a few decisions on how to go forward
  - Moratorium on coding
  - Focus on writing specification for the platform first
  - Initial agreement on core assumptions on the platform
  - Discuss these assumptions with other WPs
- In several face-to-face meetings we
  - Discussed architectural options
  - Made the final decisions on these options
- Now: Coding! - Deployable prototype should be available in a few months

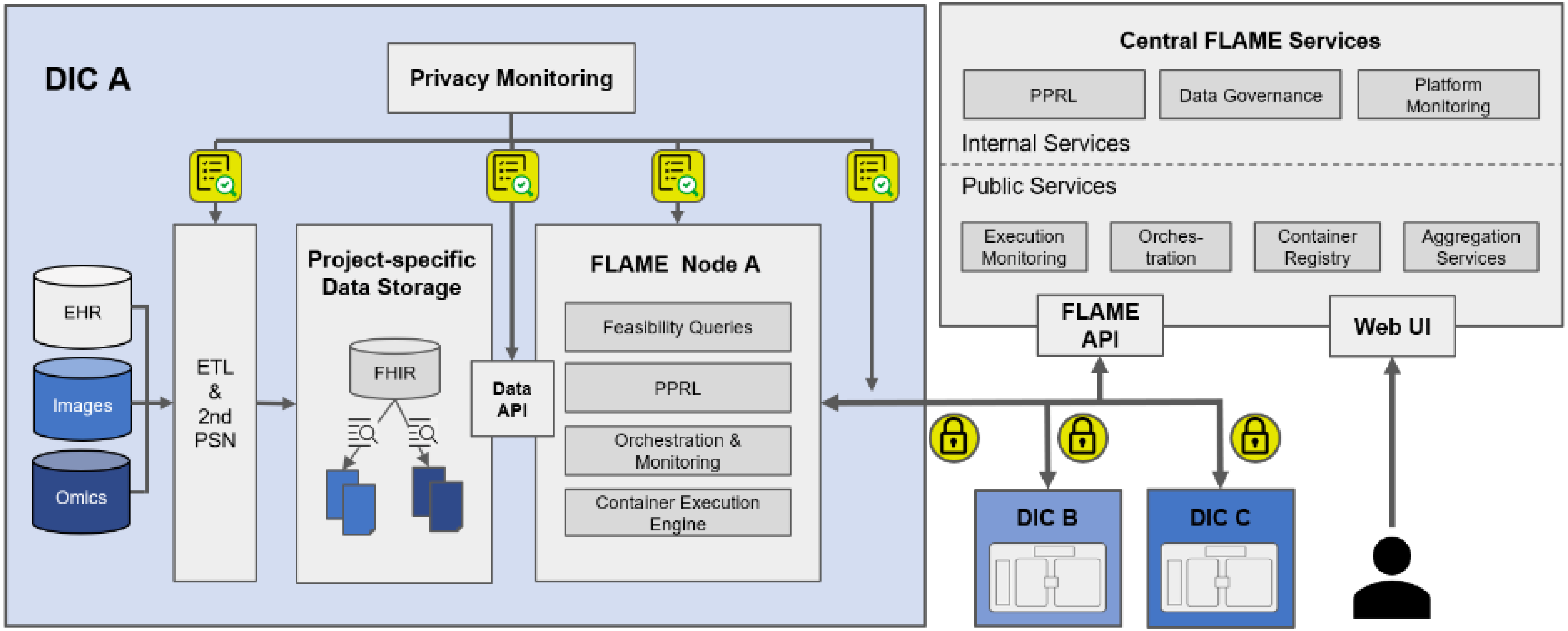
Core Assumptions

- **Results of an analysis should be transmitted via an API** on the central side and should not be transmitted back as an (docker) image.
- On the Node side, the executing analysis container should only have access to local data through an a **single API**. The API should support different data modalities (structured data - FHIR; genomes/images - object storage).
- **Communication** between the Node and Central Services **will always be initialized from the node**; the Central Service cannot instantiate a connection to the nodes - reduces the attack surface.
- There are **no direct point-to-point connections** between nodes
- To prevent the analysis from being altered in any way, the **analysis image will be signed** during the creation, the signature will be checked before the execution on the node and the node will consider the image to be immutable.

Core Assumptions

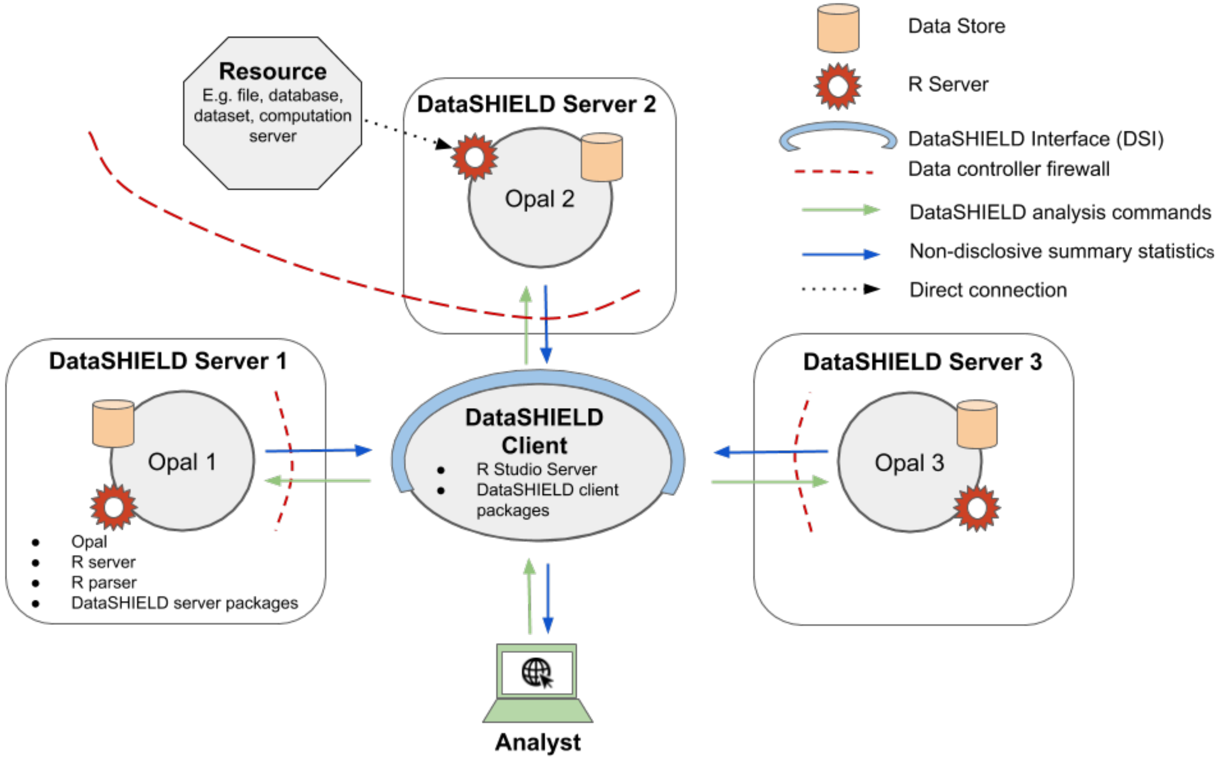
- **Data Sources are immutable**. They may only be read from by the analysis, not written to. Intermediate results may be stored to and fetched from by a separate source that is not managed by the DIC.
- An **analysis** is based on **execution** of the same container **at all sites ONCE**
- Local **scratch storage is available in each node** and will be deleted after execution of the analysis container
- Every analysis comes with a manifest providing a description of their needs
- We assume that if a Node doesn't approve, we'll skip the analysis (for now)
- We assume that the **final result is encrypted** with the public key of each participating node of the proposal (aggregation node -> central & aggregation -> central)
- We assume that we **do not install any software on** the analyst's **local computer.**

Overview of the Platform



What about DataSHIELD?

- DataSHIELD is a federated analytics platform already deployed at some DICs
- Key differences between FLAME and DataSHIELD (DS)
  - DataSHIELD
    - requires transformation of the data into an OPAL DWH
    - Suitable only for structured clinical data, but not for volume data
    - Restricted to certain R packages, use of complex (bioinformatics, image analysis) pipelines not possible
  - FLAME
    - Uses a FHIR server with MII Core Data Set instead of separate DWH – tailor-made for ready integration into DIC infrastructure w/o further transformations
    - Containers can transport arbitrarily complex analysis pipelines (deep learning, bioinformatics, imaging, ...)
    - Access to volume data designed to handle large image and genomics data sets



https://www.datashield.org/about/technical-overview-of-datashield

Outlook

- PrivateAIM started April 2023 as one of the first method platforms
- We have a strong interdisciplinary team working on all aspects of federated privacy-preserving analytics
- The FLAME platform developed within the project will be the successor of the PHT meDIC infrastructure current deployed within the MII
- We expect PrivateAIM to form the foundation for federated AI within the MII and support other clinical use cases

More information on the project: https://PrivateAIM.de

MEDICAL INFORMATICS INITIATIVE GERMANY

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Universitätsklinikum Tübingen