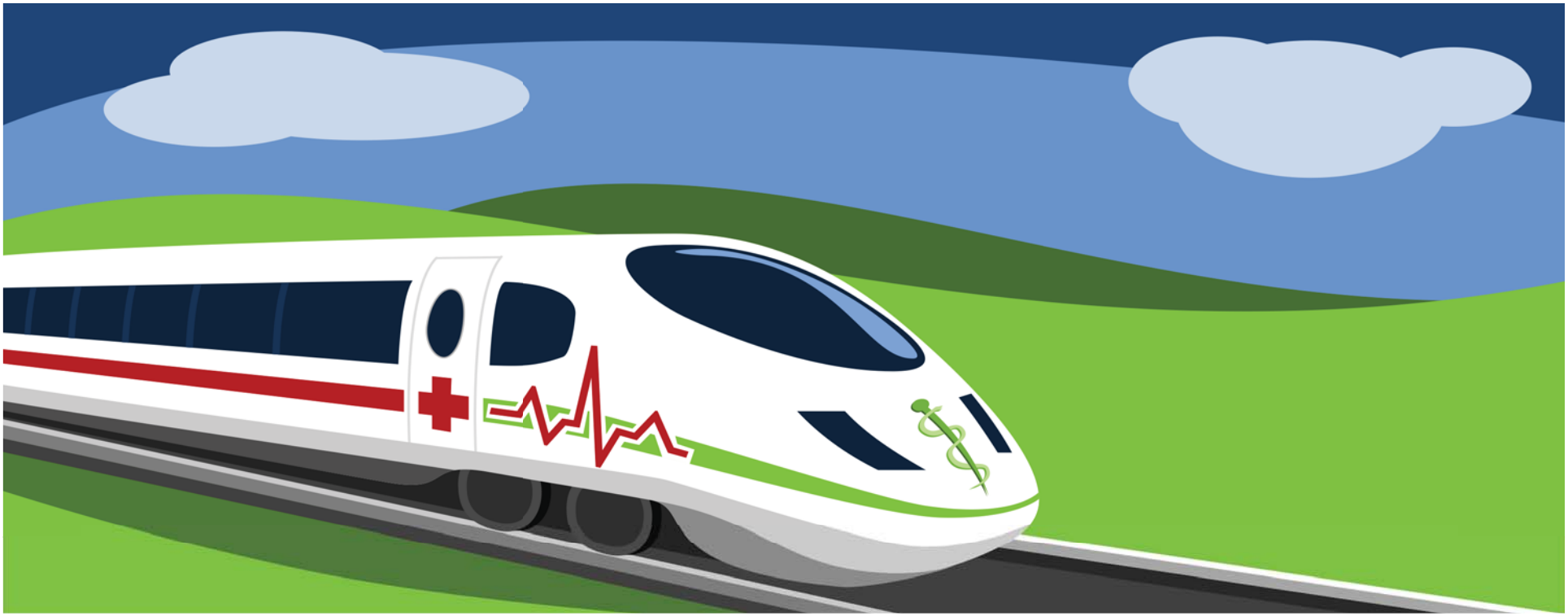
MEDIZIN INFORMATIK INITIATIVE

DIFUTURE
Data Integration for Future Medicine

UNIVERSITÄTS KLINIKUM TÜBINGEN

EBERHARD KARLS UNIVERSITÄT TÜBINGEN



Personal Health Trains

From Vision to Implementation

Lukas Zimmermann, Jörg Peter, Oliver Kohlbacher

University of Tübingen & University Hospital Tübingen

Cross-Consortium Collaboration

SMITH
Smart Medical Information Technology for Healthcare

Fraunhofer FIT / RWTH Aachen
Oya Beyan
Md. Rezaul Karim
Binh Phi Nguyen
Lars Gleim
Nils Lukas
Stefan Decker

Leipzig University
Matthias Löbe
Toralf Kirsten

DIFUTURE
Data Integration for Future Medicine

University of Tübingen
Lukas Zimmermann
Holger Stenzhorn
Oliver Kohlbacher

University of Augsburg
Fabian Rabe
Bernhard Bauer

GO FAIR

PHT Implementation Network

Bring the Algorithms to the Data

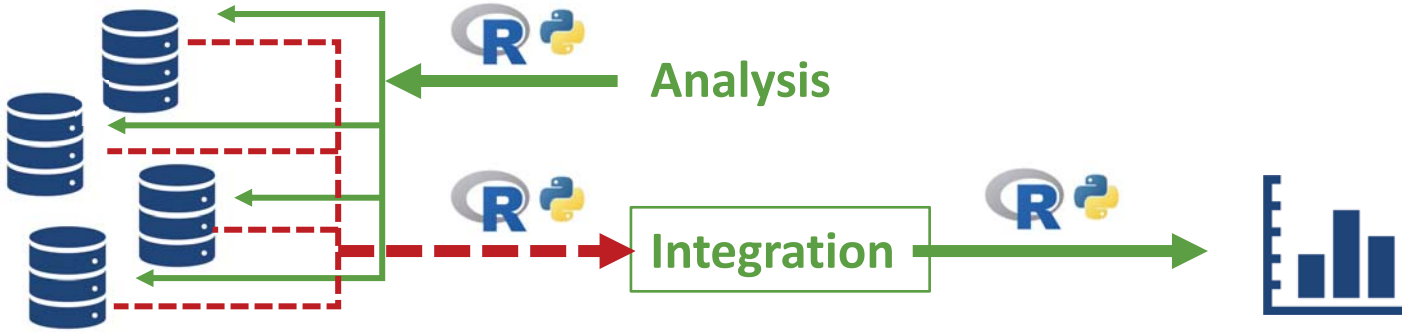
Data Pooling

- Requires transfer of all data to a centralized data pool
- Very little infrastructure required

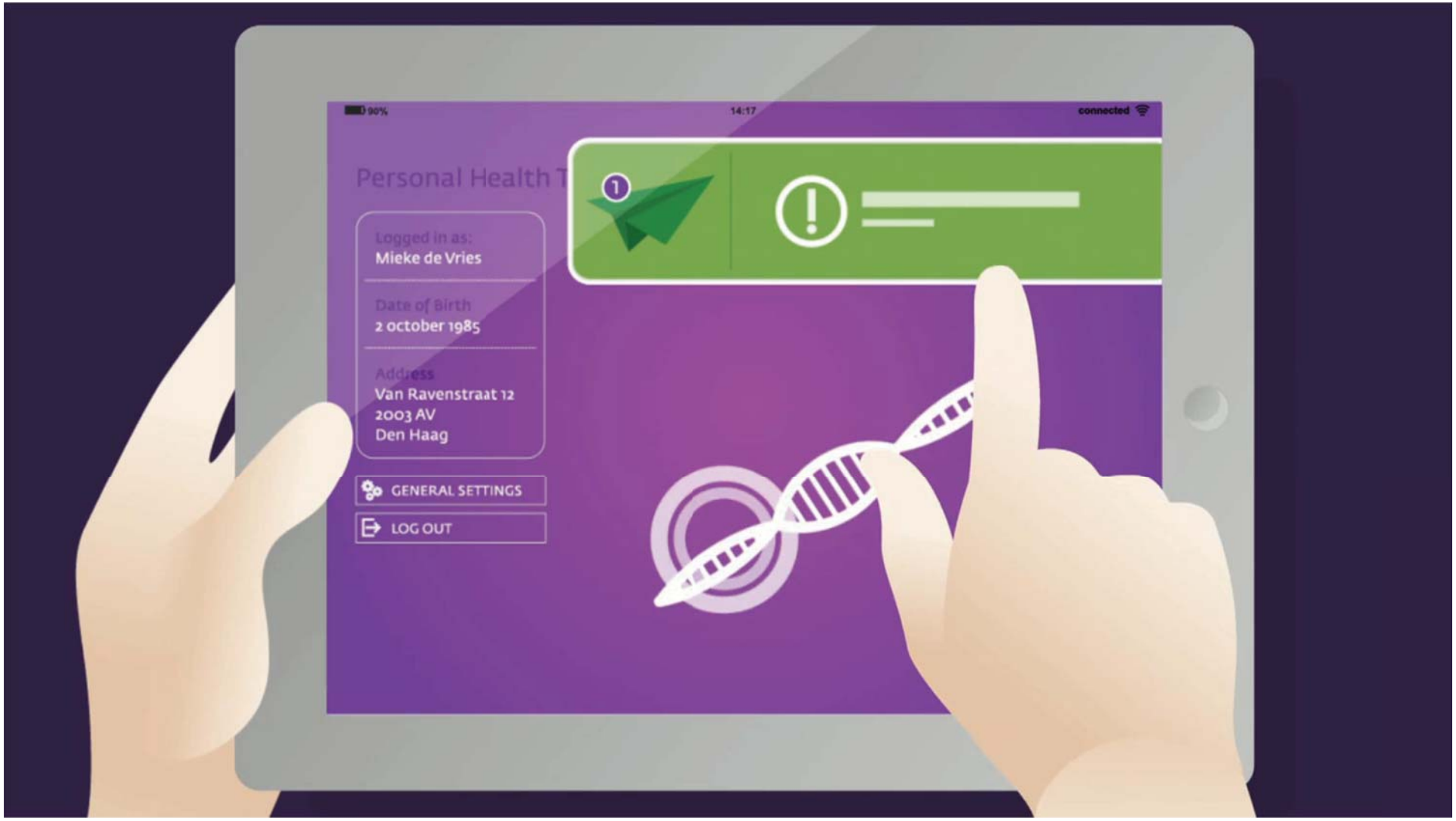
vs.

Distributed Computing

- Analysis executed locally – data remains local
- Only aggregated results returned and integrated
- Requires complex infrastructure & modified analytics

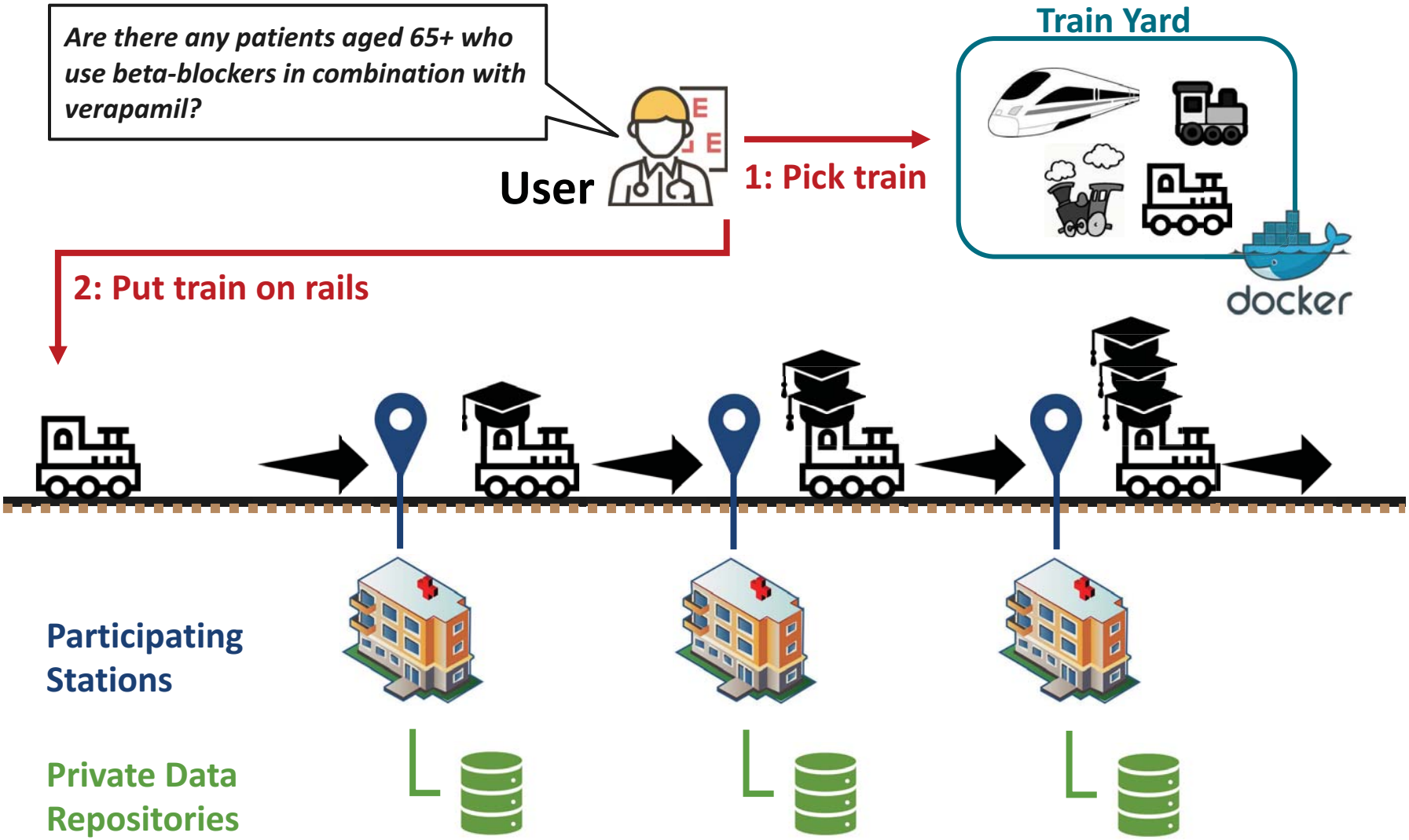


PHT – Personal Health Train



https://www.go-fair.org/implementation-networks/overview/go-fair-personal-health-train/

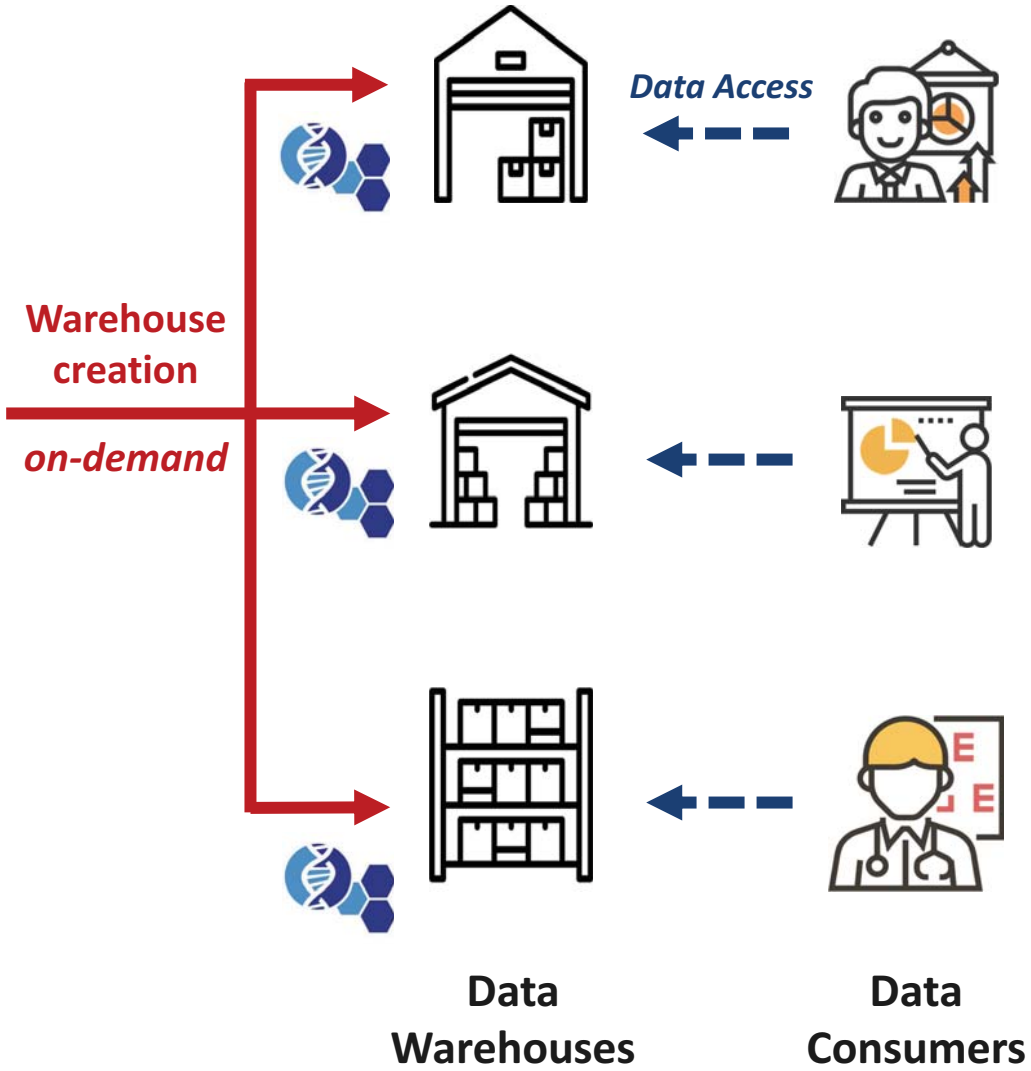
PHT – Concepts



From Data Lake to Data Warehouse

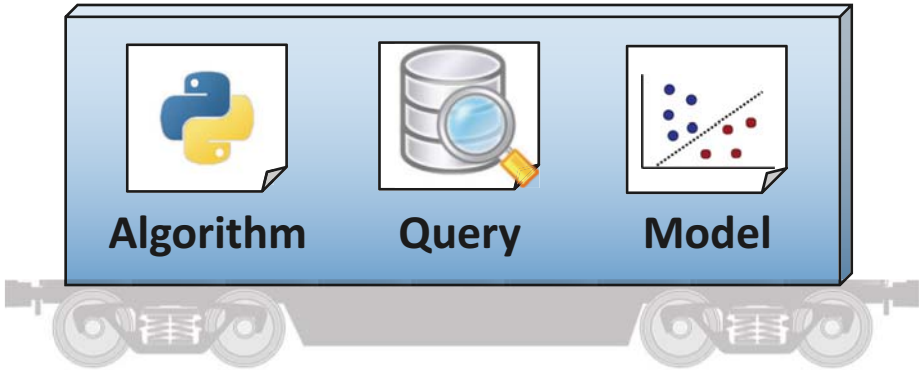
Data Lake

- Collect all data from various sources
- Schema-on-read



Personal Health Train - Key Ideas

Train



= docker

Containerized Algorithms

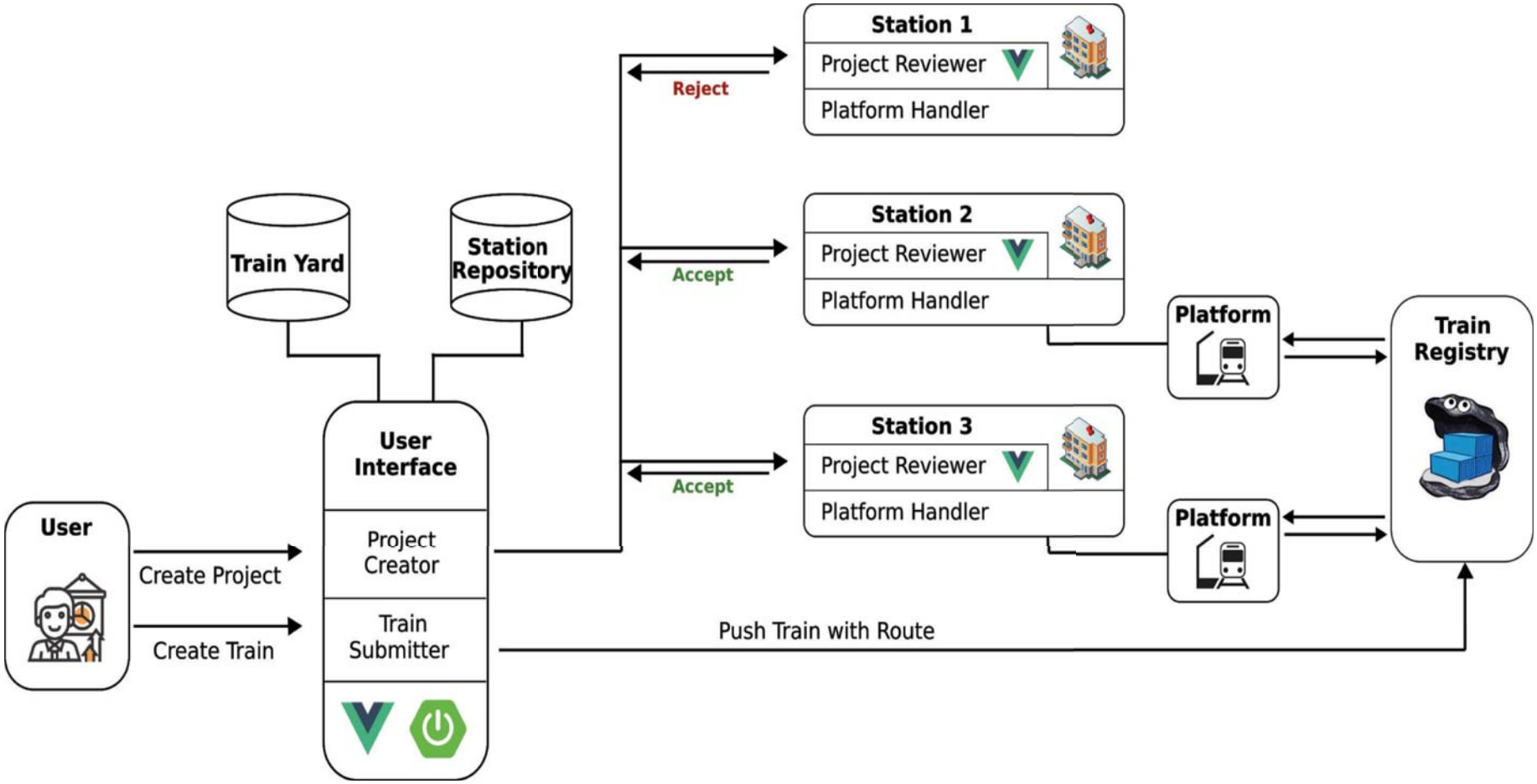
Station

Maintains private data repositories

Metadata

Describe Trains and Stations for interoperability

PHT Prototype – Implementation Overview



At the Platform

Platform

Resources:

I2B2 = http://resource-host:8080/transmart

Docker Network

```
docker run -e I2B2= http://resource-host:8080/transmart \
           --network station-network \
           train_demonstrator run_algorithm
```

docker
Network

Running Container for train

Data service

Algorithm execution initiated

Implementating a Train

1: Retrieve Data

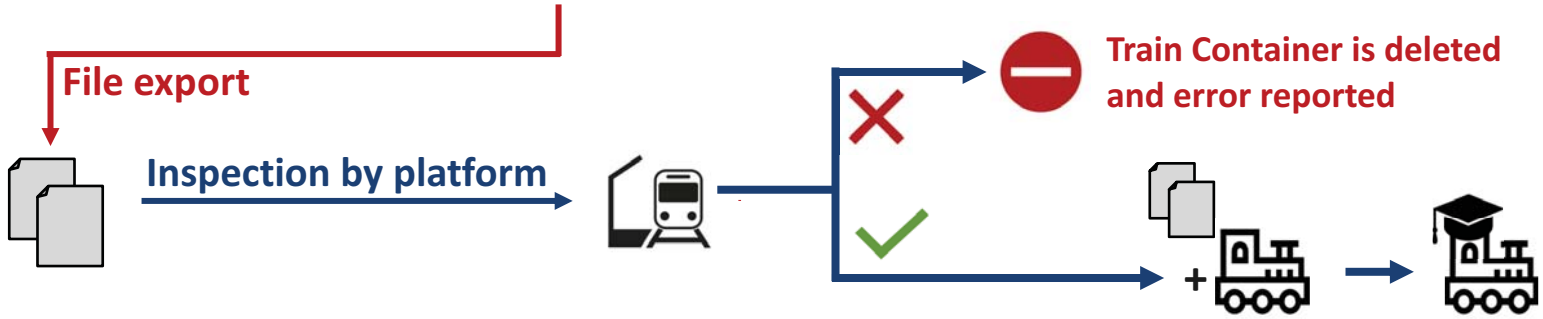
```
import requests
response = requests.get('http://resource-host:8080/transmart/')
data = response.text
```

2: Update Model

```
import sklearn
model = load_from_filesystem('/path/to/model')
model.partial_fit(data)
model.save_to_filesystem('/path/to/model')
```

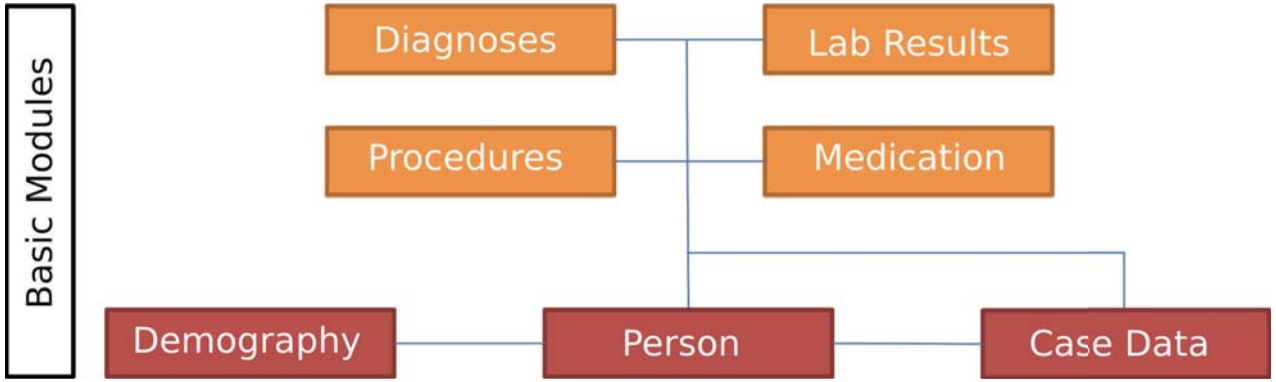
3: Declare Result of Algorithm Execution

```
return RunAlgorithmResponse(
      sucess=True,
      message='Execution Details',
      export_files=['/path/to/model'])
```



Demonstrator Study with the MII

Based on the National Core Dataset of the Medical Informatics Initiative



Source: Experiences from the National Demonstrator Study within the German Medical Informatics Initiative; Ganslandt et al. submitted to AMIA Clinical Informatics Conference 2019

Data available at each site, but cannot be pooled

→ Currently: set up DWH, e-mail analysis scripts, execute, e-mail aggregated results back

Clinical Use Cases

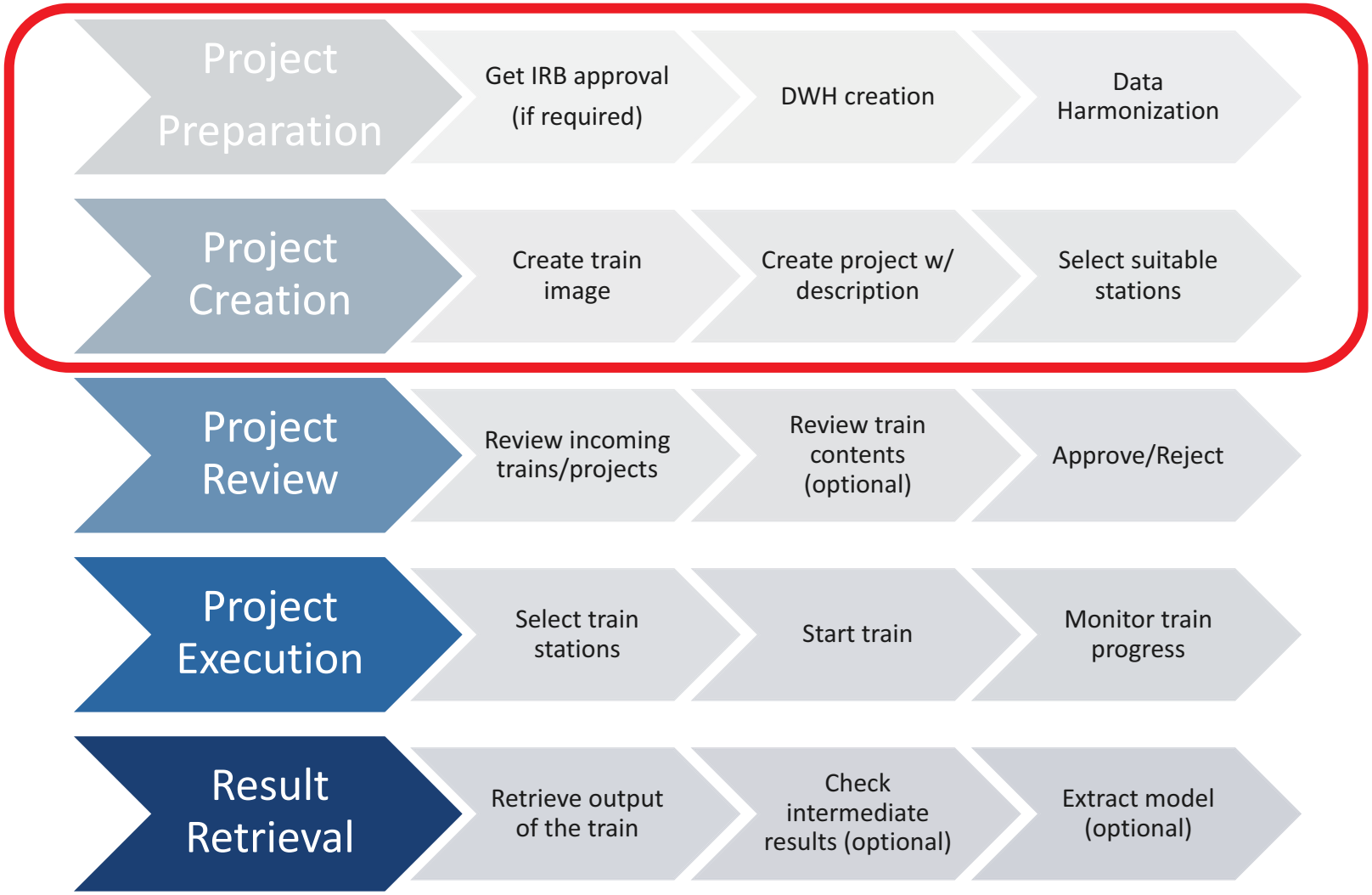
Rare Diseases

- Groups defined by ICD10
- Geo-visualization

Multimorbidity

- Charlson Index

Project Workflow



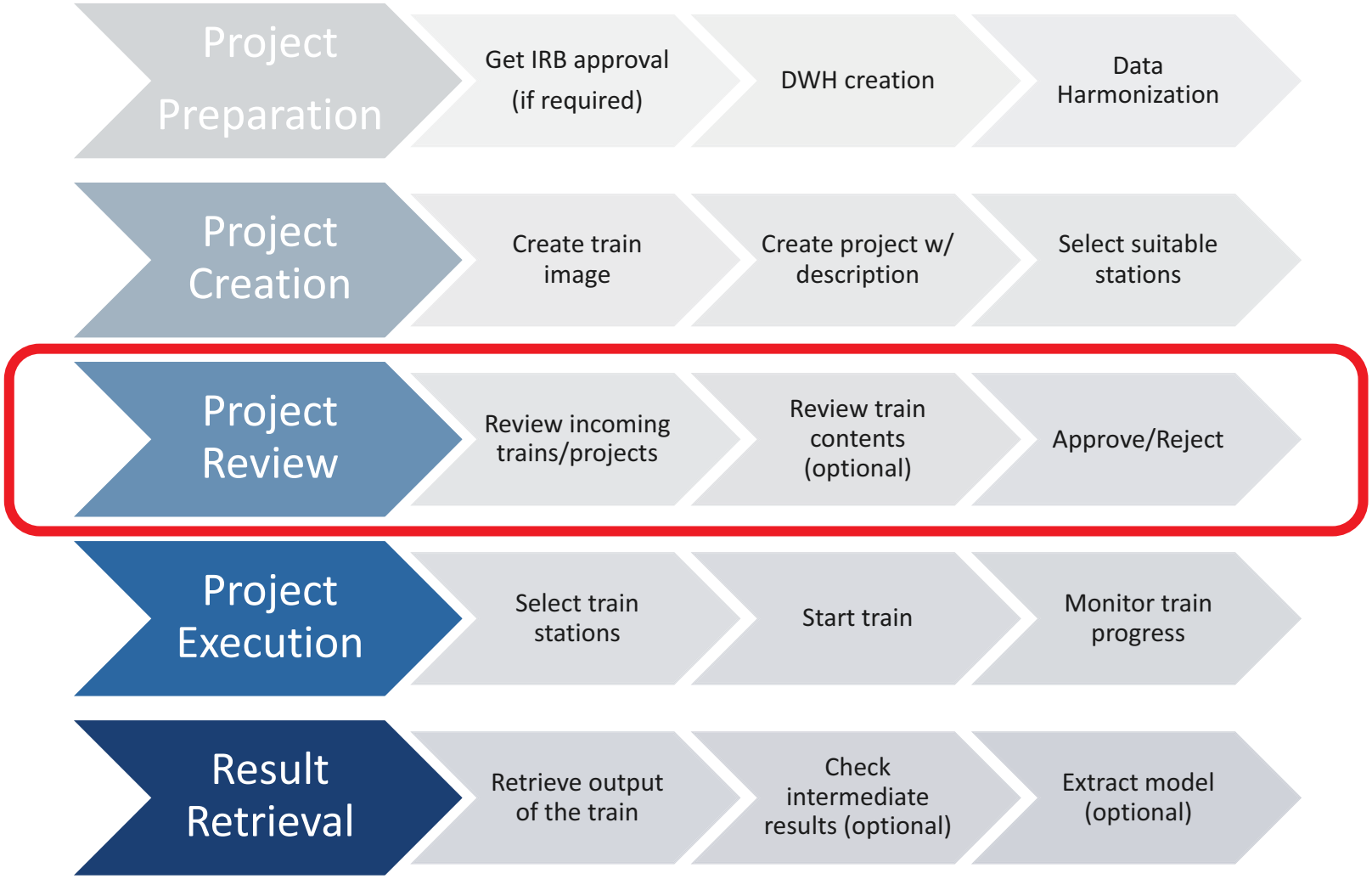
Train Station UI – Project Creation

Personal Health Train - DIFUTURE Project Management

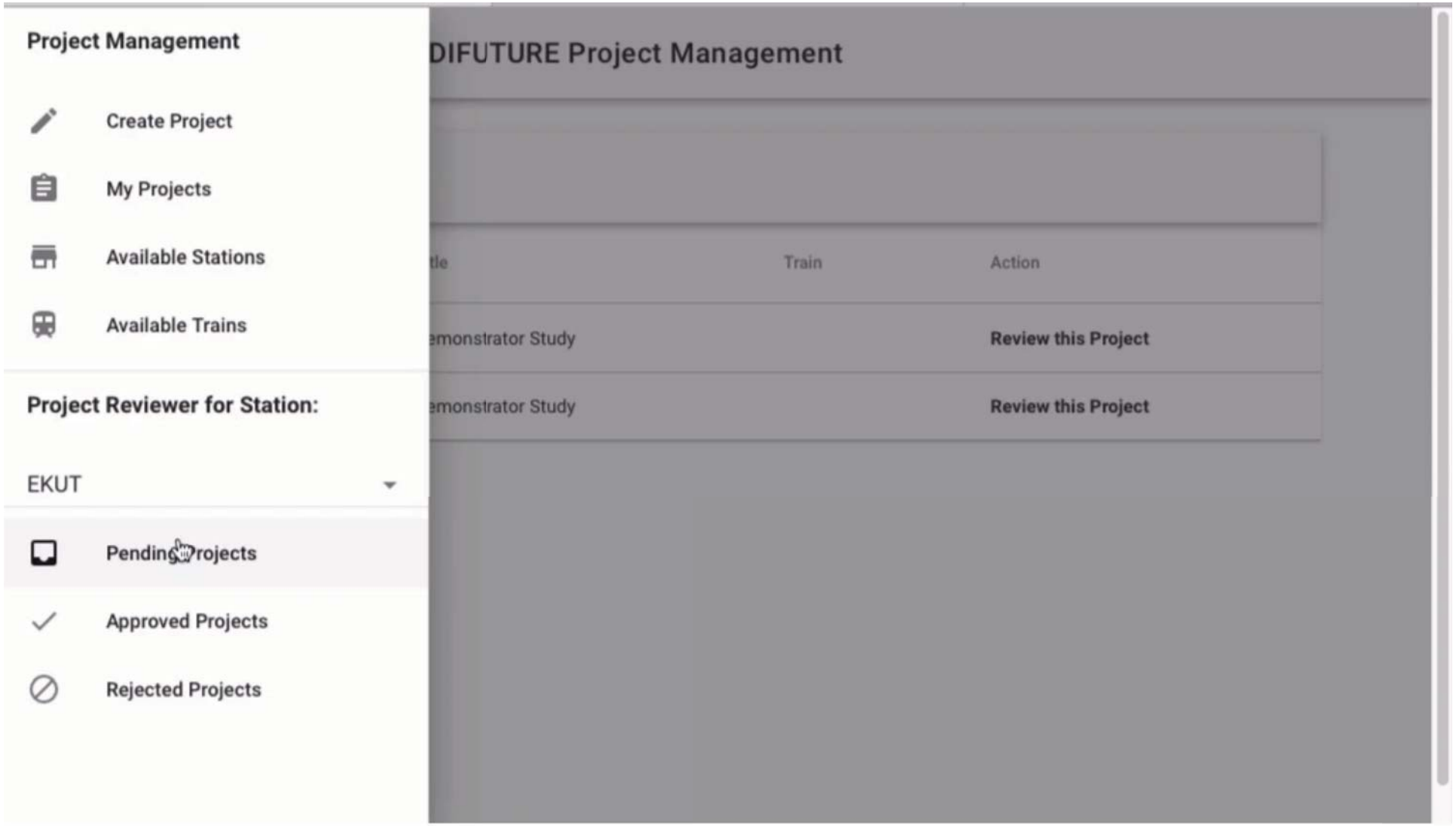
Train

Namespace ↑	Name	Last Updated
personalhealthtrain	train_demonstrator	2018-11-15T14:36:46.799096Z
personalhealthtrain	train_dummy	2018-11-07T14:21:04.983924Z
personalhealthtrain	train_statistics	2018-11-17T1127:05.984937Z

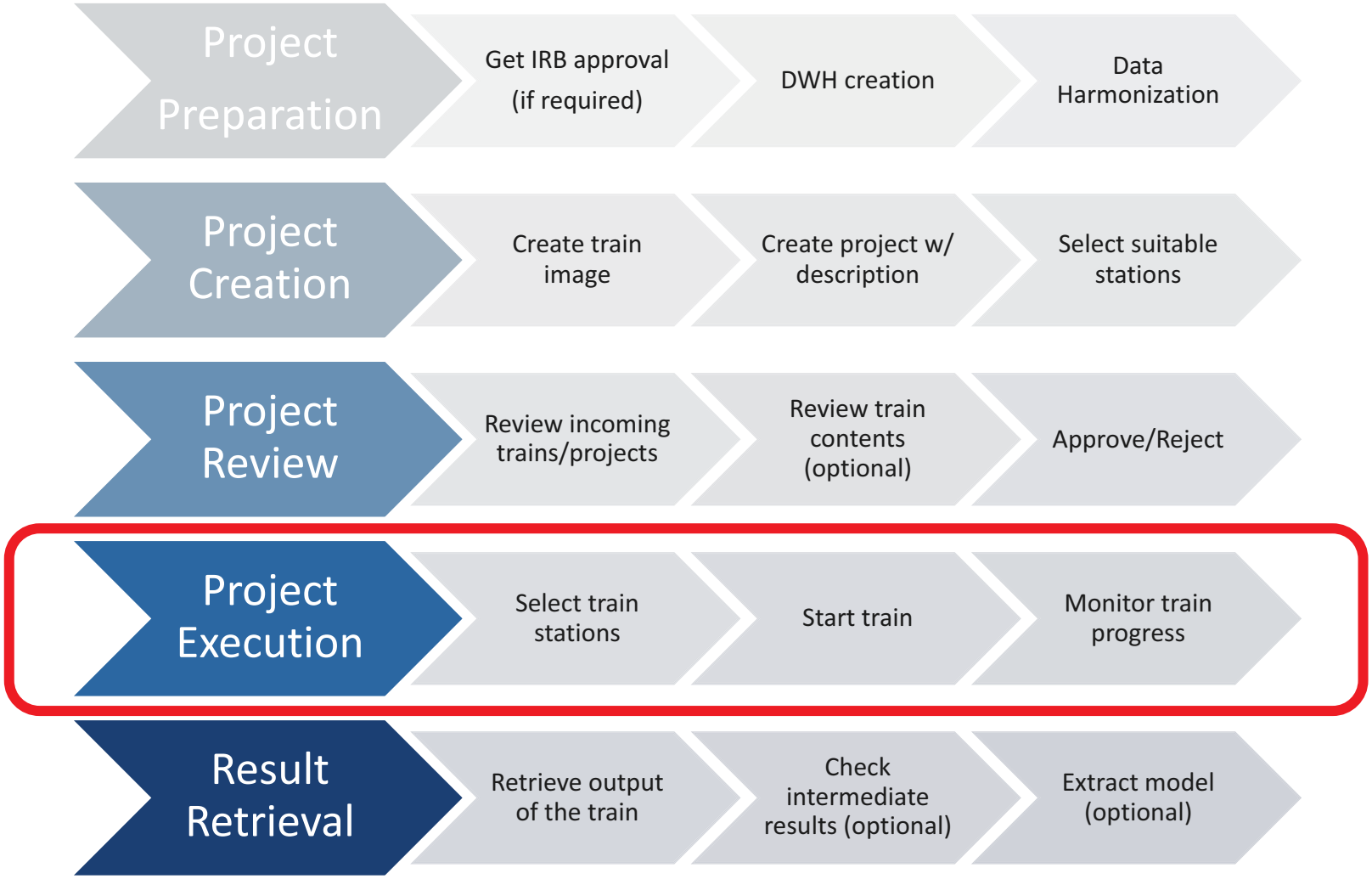
Project Workflow



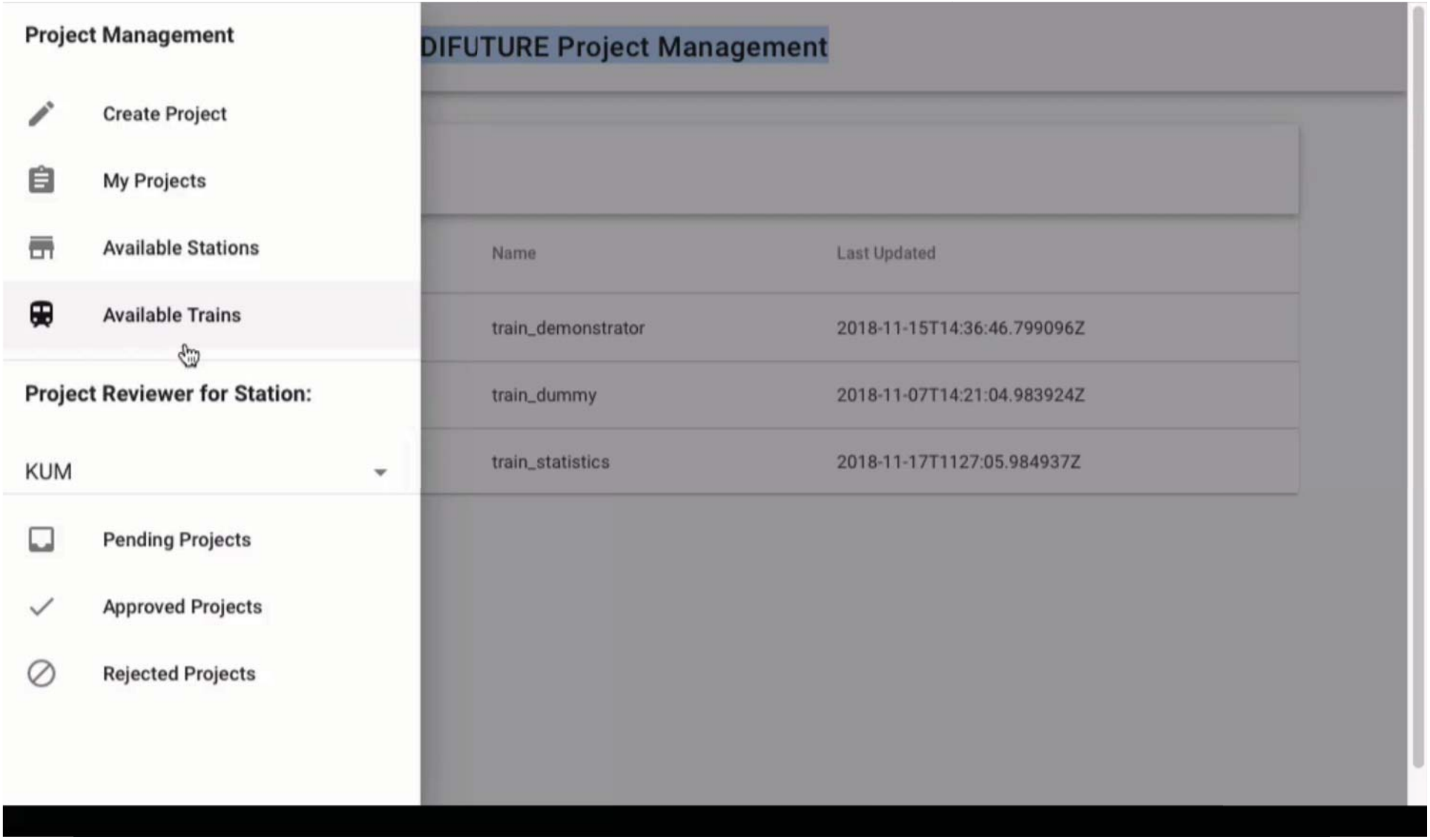
Train Station UI – Project Review



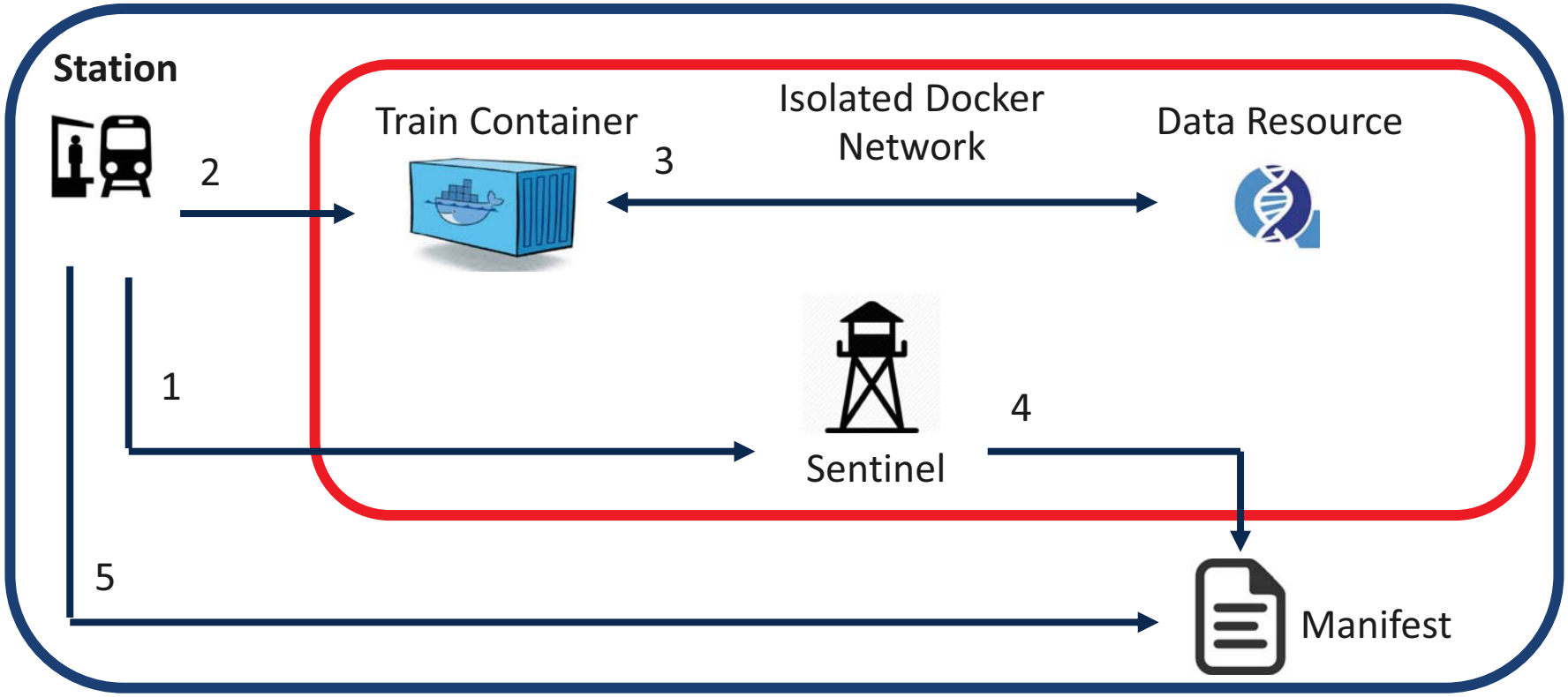
Project Workflow



PHT GUI – Starting the Train



Local Train Execution



Execution

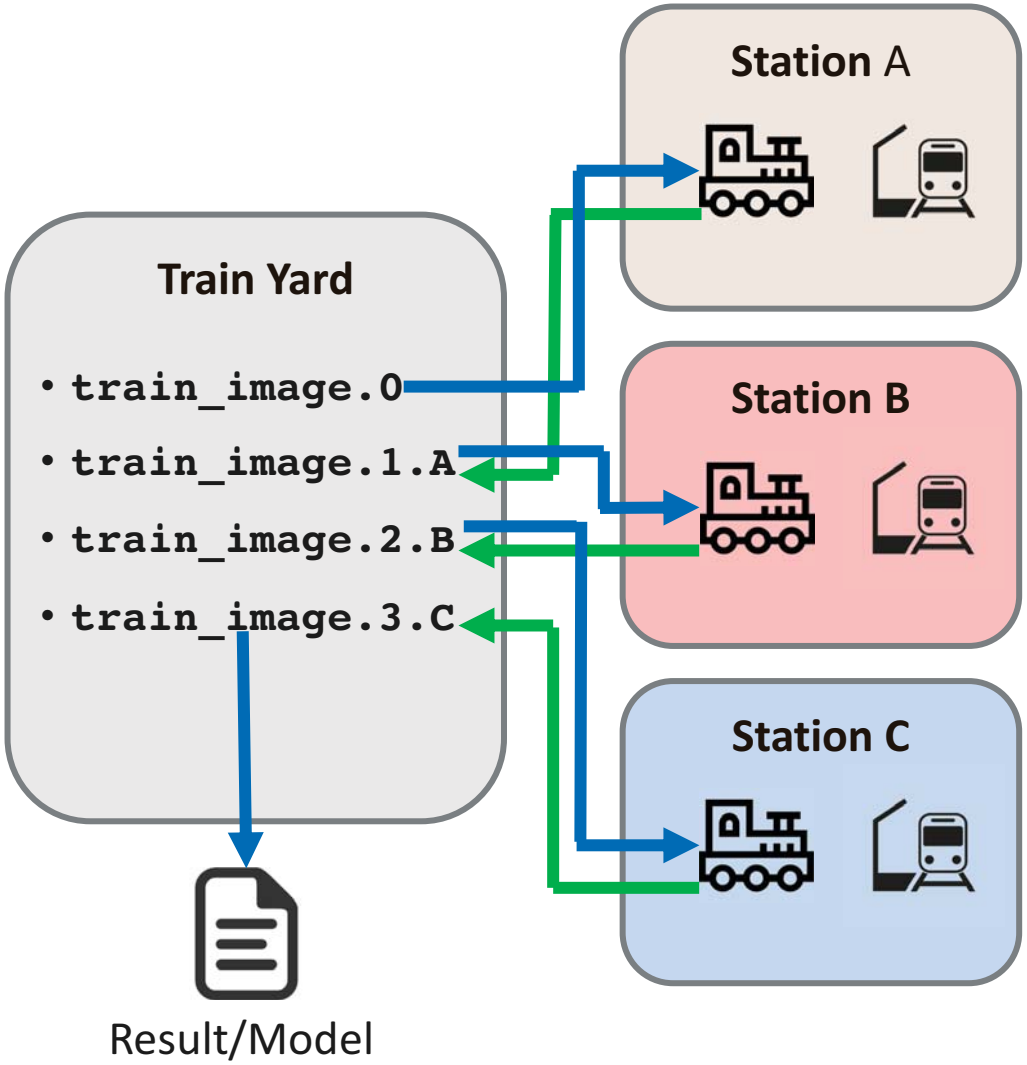
1. Start sentinel service & data res.
2. Excute train in isolated network
3. Manifest: log train's network traffic
4. Station inspects manifest

Train Evalution

- Delete container if
  1. Timeout or nonzero exit code
  2. Manifest suspicious
- Otherwise, push back to repository

Implementation of Train Routing

- Train yard implemented as private **docker registry**
- **Trains are stateful**
- Registry push/pull operations support versioning of trains
- Routing: visit all stations sequentially
  - Pull current state into station
  - Execute train and validate
  - Push back new state
  - Tags describe station
- **Result stored in last valid train**
- All intermediary results available as well



Project Workflow

Project Preparation	Get IRB approval (if required)	DWH creation	Data Harmonization
Project Creation	Create train image	Create project w/ description	Select suitable stations
Project Review	Review incoming trains/projects	Review train contents (optional)	Approve/Reject
Project Execution	Select train stations	Start train	Monitor train progress
Result Retrieval	Retrieve output of the train	Check intermediate results (optional)	Extract model (optional)

PHT GUI – Inspecting Results

Personal Health Train - DIFUTURE Project Management

Demonstrator Study

OVERVIEW | RUN NEW TRAIN

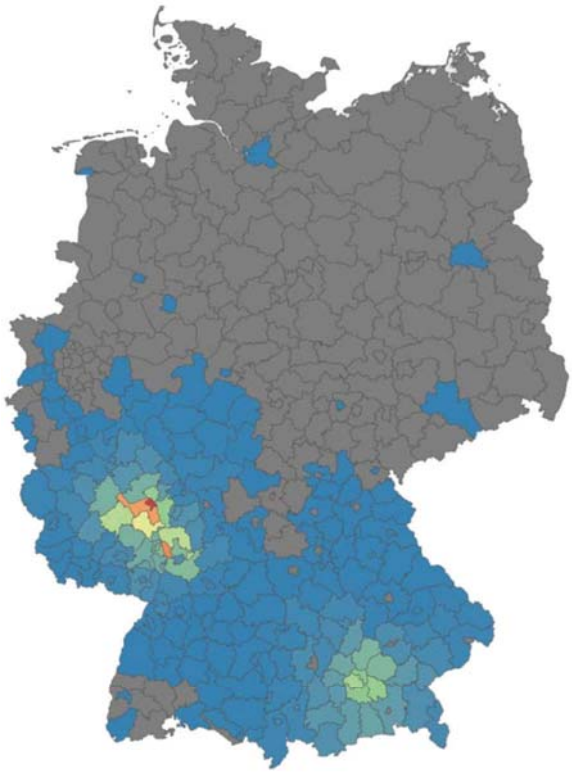
Select Train

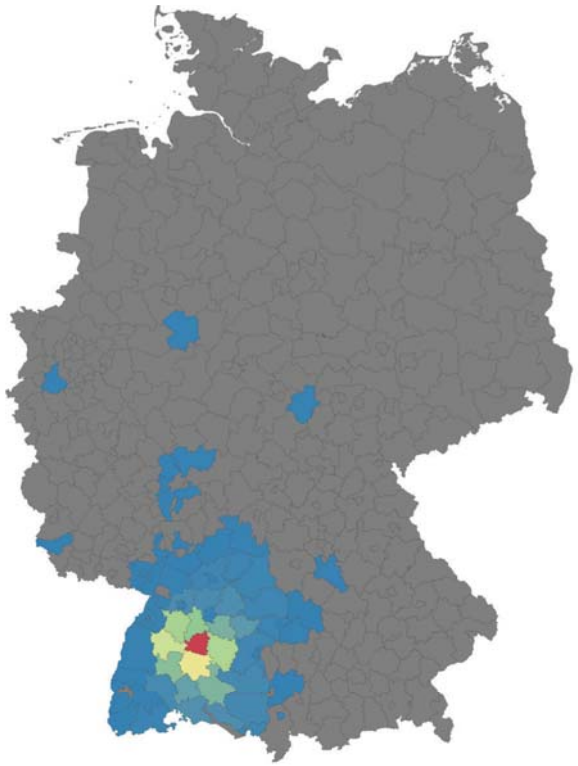
train_demonstrator

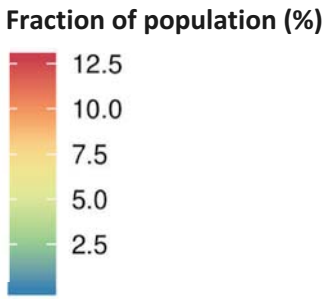
Use the station table below to construct a new route for the train. Click on a Station to append it to the current route, click again to remove the station from the route. If you are happy with the route, click on "Run Train". Only Stations that have approved the Project can be selected for creating the route.

Stop	Station ID	Name	
2	4	KUM	✓ Approved
3	1	EKUT	✓ Approved
4	2	UU	✓ Approved
1	3	MRI	✓ Approved

Geovisualization of Comorbidity







- Mannheim
- Mainz
- TU Munich

- Tübingen

Machine Learning – Predicting Phenotype from Genotype

- HIV can use different co-receptors for host cell entry (tropism)
- Usable co-receptors (CCR5, CXCR4) are defined by viral genome
- Sequencing yields patient-specific data sets (viral genomes)

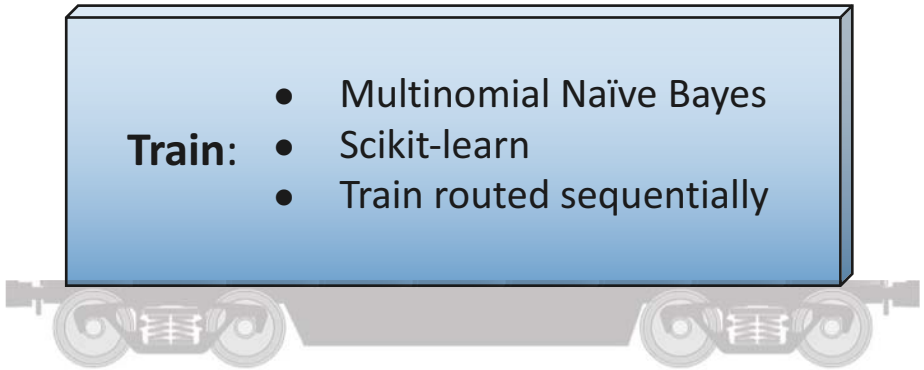
Problem

- Given partial virus sequence (V3 loop)
- Does it bind to co-receptor CCR5?

Binary Classification Problem

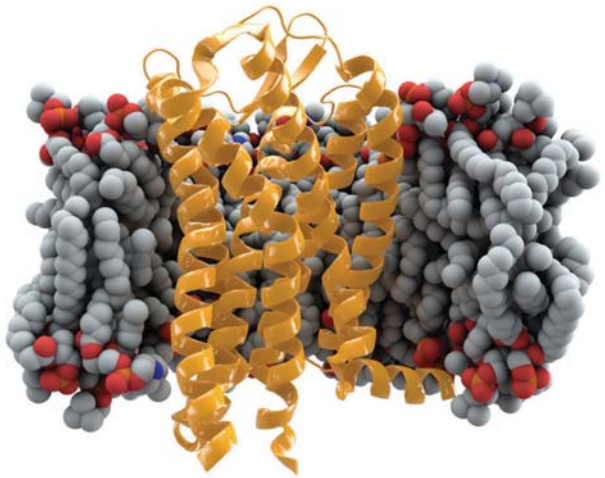
Train:

- Multinomial Naïve Bayes
- Scikit-learn
- Train routed sequentially



CIRPNNNTRKSIHIGPGR
AFYTTGDIIGDIRKAHC

CTRPNNYKRKRITTGPGK
VLYTTGQIIGDIRRAHC



Los Alamos National Laboratory (LANL) HIV Sequence Database
Kalinina OV, Pfeifer N, Lengauer T. *Retrovirology.* 2013;10:130. doi: 10.1186/1742-4690-10-130

Machine Learning – Prediction Accuracy

PHT Network

- 18 separate train stations spawned in the cloud

Training Data

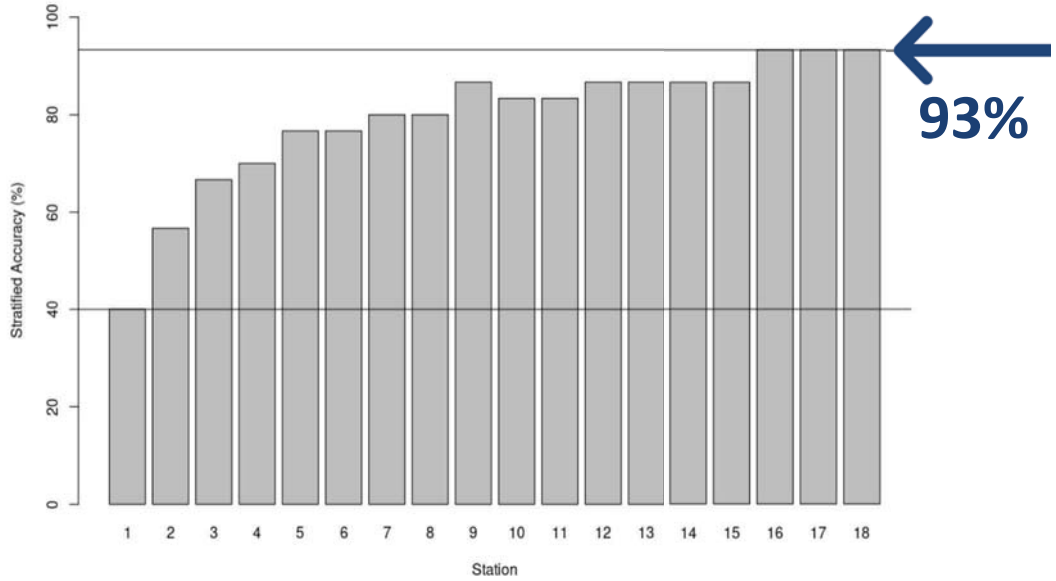
- Each station provides one positive and negative example

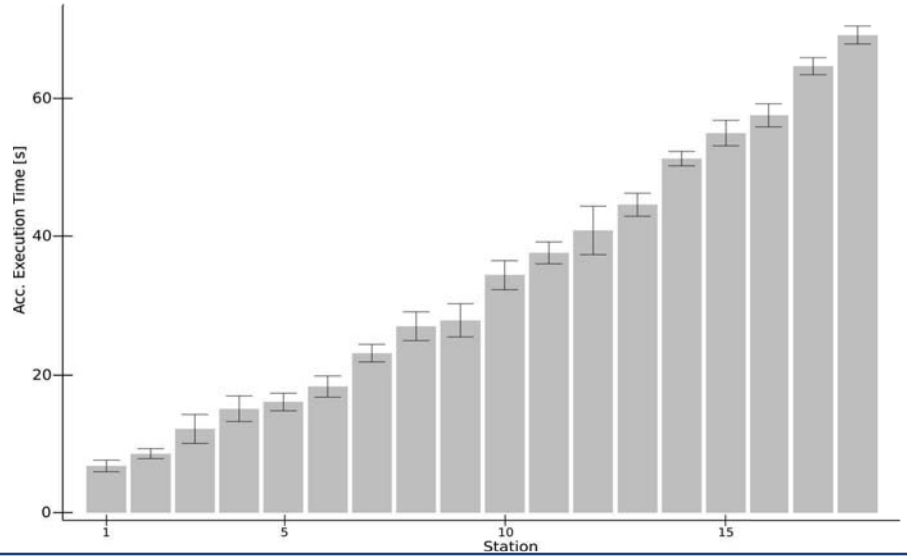
Validation Data

- 30 data points (balanced)
- Assess accuracy of the final model

Results

- Same prediction quality as on pooled data
- Fully automated analysis
- Startup & execution: ~ 2 s/station





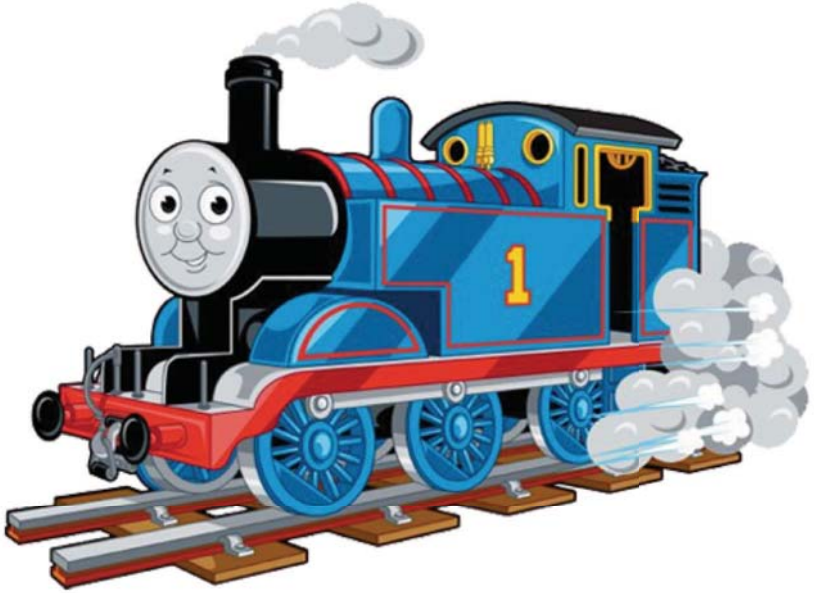
Current Status

Initial Prototype

- Not (yet) intended for production use, fully open source
- Proof-of-concept for
  - Container technologies
  - Train orchestration
  - API development
  - Scalability and performance test
  - Testbed for ML algorithm development
  - GUI specification development

Limitations

- Strong assumptions on data model
- No automated data discovery
- Security mechanisms missing
- Train review missing



Outlook

Data Discovery, Integration & Analysis

- Standards, mechanisms, and APIs for data discovery and mapping
- Which machine learning techniques are suitable to which project and to the infrastructure?
- How to orchestrate data discovery process?
- How to make analyses reproducible?

```
define "InDemographic":
  "ProstateCancerPositive" and "Type2DMPositive"

define "ProstateCancerPositive": exists (
      [Condition] C where ToCode(C.code.coding)~"
         CarcinomaOfProstate")

define "Type2DMPositive": exists (
      [Condition] C where ToCode(C.code.coding) ~ "
         DiabetesMellitusType2") and not exists (
      [Condition] C where ToCode(C.code.coding) ~ "
         DiabetesMellitus1") or exists ([
         MedicationRequest] MR where ToCode(MR.
         medication.coding[0]) ~ "Insulin")
  and ToQuantity (Last ([Observation] O where ToCode(O.
     code.coding) ~ "FastingGlucose"
         sort by effective.value).value as Quantity
    ).value > 200 and ToQuantity (Last ([Observation] O
         where ToCode(O.code.coding) ~ "HemoglobinA1C"
         sort by effective.value
      ).value as Quantity).value >= 6.5
```

Listing 1.2. Part-II: CQL query showing the InDemographic and statement definitions for population and condition

Karim et al., A Distributed Analytics Platform to Execute FHIR based Phenotyping Algorithms, SWAT4HCLS 2018 (accepted)

Outlook

Security

- How do we ensure that the train behaves according to the project proposal?
- How could an audit/review process be implemented?

Consent

- How do we reconcile train purpose and patient consent?
- Dynamic Consent?
- Consent encoding and inference? – Smart contracts

Privacy

- How to assess (and minimize) the re-identification risk for patients?
- Privacy-preserving machine learning methods

PHT Implementation Network Meeting
Berlin, Germany - 12 February 2019 - goo.gl/owvQVy

Acknowledgements

DIFUTURE
Data Integration for Future Medicine

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Lukas Zimmermann
Jörg Peter
Aydin Can Polatkan
Armin Roth
Holger Stenzhorn
Nico Pfeifer

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fabian Prasser
Klaus A. Kuhn

LMU

Daniel Schalk
Heidi Seibold
Ulrich Mansmann

Universität Augsburg University

Fabian Rabe
Bernhard Bauer

SMITH
Smart Medical Information Technology for Healthcare

RWTH AACHEN UNIVERSITY

Oya Beyan
Md. Rezaul Karim
Lars Gleim
Nils Lukas
Stefan Decker

Matthias Löbe
Frank Meineke
Toralf Kirsten

miracum
Medical Informatics in Research and Care in University Medicine

Thomas Ganslandt

UNI FREIBURG

Martin Boeker

GO FAIR

Rob Hooft

LUMC

Kees Burger
Luiz Olavo Bonino
Mark Thompson

UM

Johan van Soest

René Hietkamp
Marc Nieuwland
Mark D. Wilkinson

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung