GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

#1

Data standardization and linkage (Standards & Federation)

Vincent Planat- Dedalus
Davide Piscia- CNAG
Silvia Uribe- UPM

Agenda of the session

- ❑ [10 mn] Introduction
  - ○ Session introduction
  - ○ Rare disease data disparity: From problem statement to opportunities

GenoMed4All approach

- ❑ [15mn] data standardization
  - ○ The problem of data harmonization & the Common Data Model solution
- ❑ [15mn] Federated learning solution
  - ○ Platform design, architecture & development
- ❑ [15mn] Use of the platform
  - ○ Users, data flows & solution model

- ❑ [5mn] Conclusion & lesson learned
- ❑ (20 mn) Questions

Rare disease data disparity:

From problem statement to opportunities

Current situation

General Context

- ❑ Hematological area: most diseases have a genetic background
  - ○ Up to 450 variants, including oncological and non-oncological ones
- ❑ These diseases represent a growing public health challenge
  - ○ 5% of cancers, chronic health issues with life-threatening conditions…
- ❑ The application of precision medicine should be an optimal option in this context but…
- ❑ ..there is a lack of well-established international datasets to be used
  - ○ There are not centralized big data repositories

Current situation

Data disparity

- ❑ In the context of rare disease research (SCD, MM & MDS in GenoMed4ALL):
  - ○ There is a lot of small datasets (identification problem!!!)
  - ○ There are multiple data modalities to be considered (clinical, genomic, demographic, imaging, etc.)
  - ○ Different approaches for data standardization (OMOP, FHIR, Phenopackets, etc..)
- ❑ Clinical networks are needed to address this situation: ERN-EuroBloodNet clinical network (66 repositories in GenoMed4ALL)
- ❑ Optimal solution: the pooling and integration of multiple datasets from different centers but…
- ❑ ..there is a strong resistance about this approach in the healthcare context

Current situation

Why not sharing data?

Economic & Motivational

Legal & Ethics

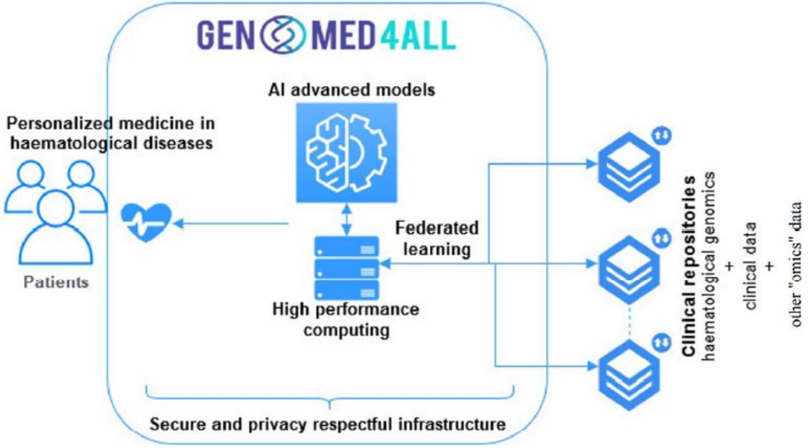
Technical

Political

Current situation

Thinking outside the box

- ❑ Federated Learning as a new paradigm
  - ○ Scalable and privacy-preserving approach to the join training of AI models across federated health data repositories



Main objective:

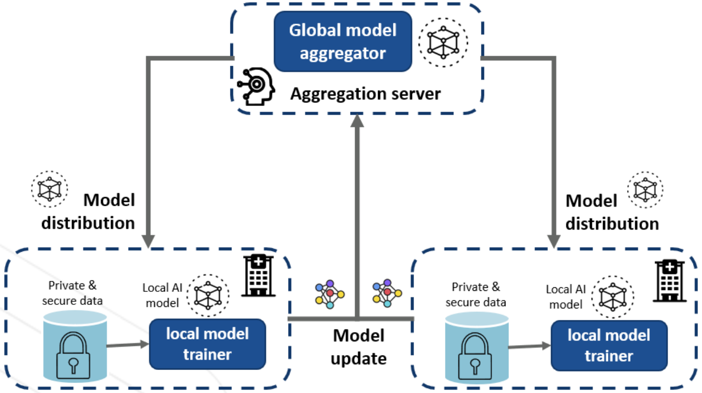
To allow AI model development without sharing data

GenoMed4All approach (I): data standardization

Why Harmonizingthe data forservingFederatedlearning architectureis a challenge ?

- **No consistent** & **harmonized datasets** = **No AI**
- **In Genomed4All, The scope of data types sourcing the AI model training is wide**
  - **Imaging**
  - **Clinical data (studies, tests, questionnaires, observations, tratement, diagnostic ...)**
  - **Genomic Data**

- **The Health Information Systems (HIS) sourcing the data are from multiple vendors**
- **Each vendors implements its own Database schema & structured data strategy**



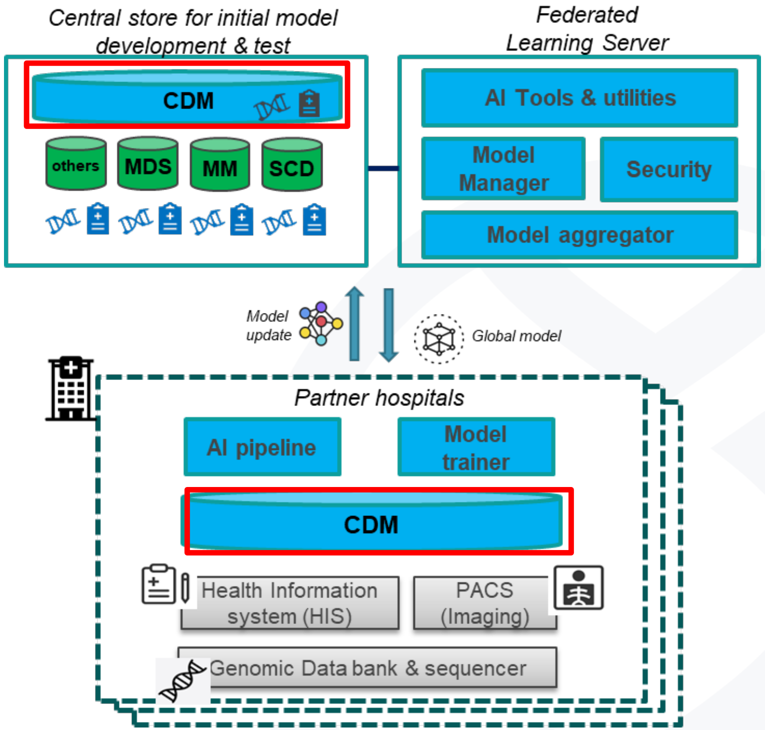
Federated learning model (adopted by Genomed4All)

We need a Common Data Model (CDM) even if we are federated

- We need a **data model reference** for the **initial** AI model development that will be subject to AI training federation

- The **AI federation** engage multiple Hospitals in the training porcess.
    - **The data interop**, to extract the training dataset, is **greatly facilitated** with a CDM (Common ETL applied to all contributors)
    - A CDM The model can be initially developed in the central server &

- To be trained at the edge the model needs a dataset extracted from EMR. The **ETL** is defined in the central server, referenced as part of the training plan & executed at the edge **on the same CDM**



Simplified platform design

A CDM … but which model ?

When looking at genomic Data standard

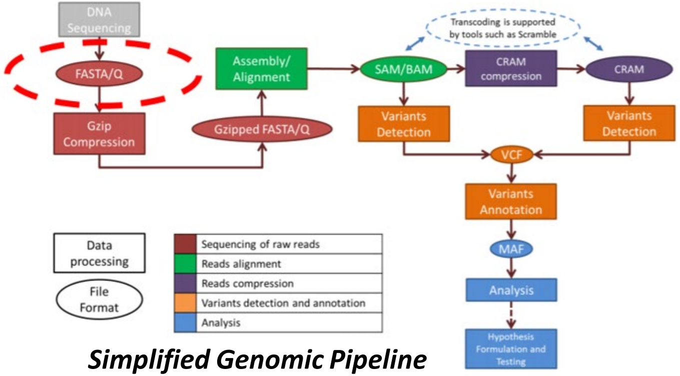
- Genomic data formats are already existing with **clear scope of applicability** (without overlapping)
  - The data format depends on the stage in the pipeline (FASTQ, BAM, VCF etc …)
  - All of these format do not contain any *clinical* information

What about clinical Data standard

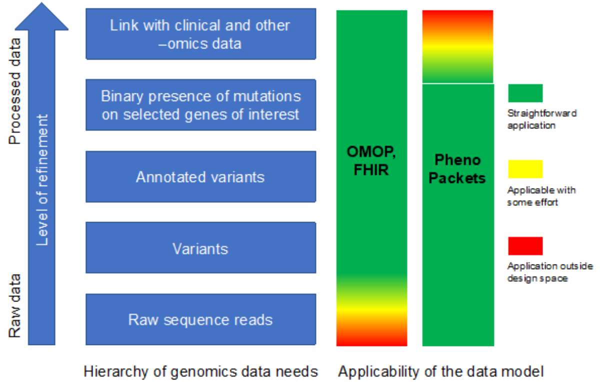
- **Clinical data space standard** is much more fragmented with strengths, weakness & **overlapping depending on the domain** of application: Research / Clinical, Care provider/Life-science etc …
- **No Stadnard CDM** currently cover the full spectrum (clincal & genotype). Existing initatives are (with +/-): **HL7, FHIR, OMOP, GA4G, ISO …**

➔**GenomedAll** conducted a deep comparison study through a research paper (*) & **chose the FHIR standard**

➔**FHIR** will ease Genomed4All **data interoperability** with other **European project**

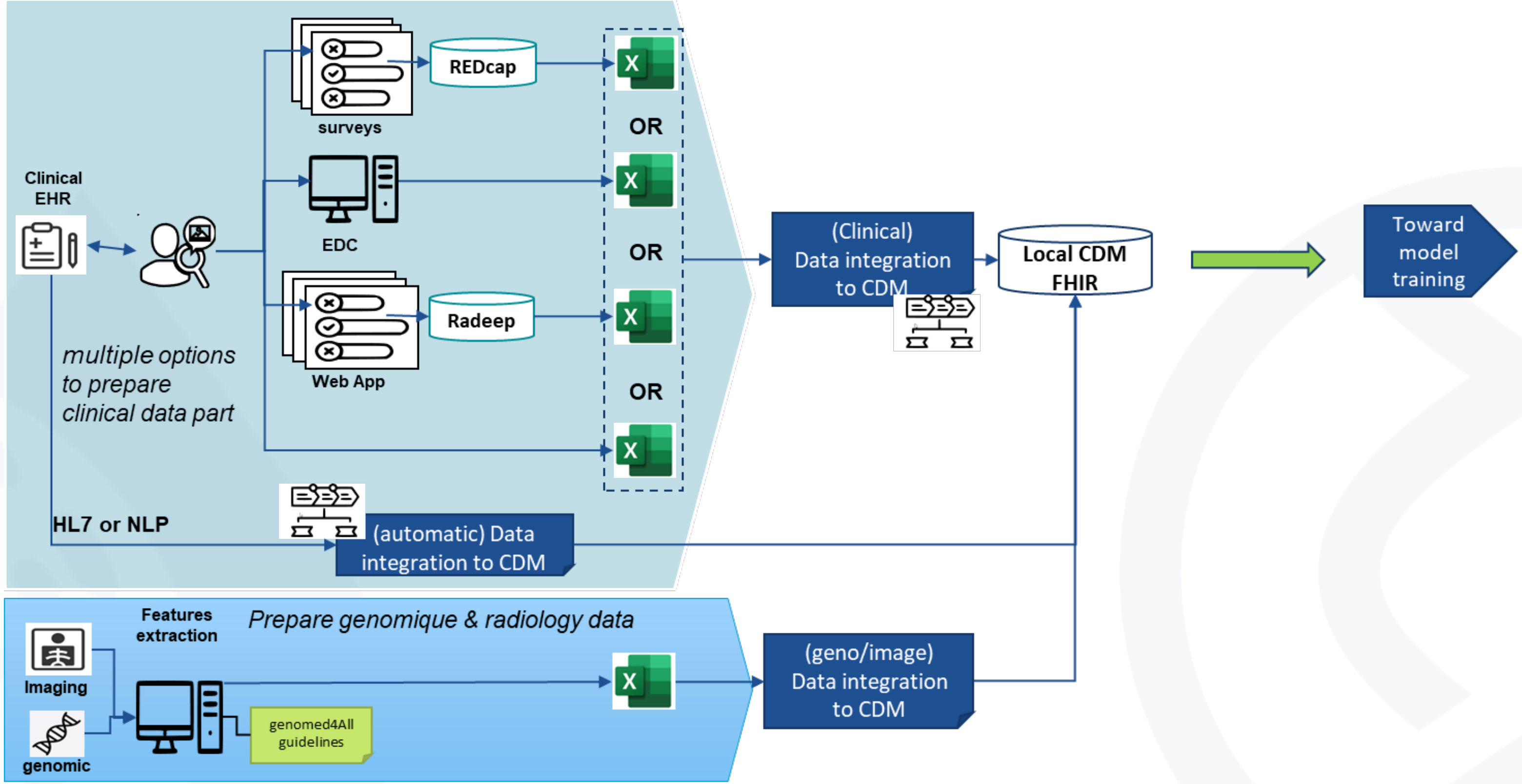


Simplified Genomic Pipeline


Processed data
Raw data
Level of refinement
Link with clinical and other –omics data
Binary presence of mutations on selected genes of interest
Annotated variants
Variants
Raw sequence reads
OMOP, FHIR
Pheno Packets
Straightforward application
Applicable with some effort
Application outside design space
Hierarchy of genomics data needs
Applicability of the data model

(*) The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform

Data provisioningis THE complex task…



Clinical EHR

surveys

REDcap

EDC

Web App

Radeep

OR

OR

OR

multiple options to prepare clinical data part

HL7 or NLP

(automatic) Data integration to CDM

(Clinical) Data integration to CDM

Local CDM FHIR

Toward model training

Features extraction

Prepare genomique & radiology data

Imaging

genomic

genomed4All guidelines

(geno/image) Data integration to CDM

GenoMed4All approach (II): federated learning solution

Architecture decision process

Building a platform for federated learning is complex

- ❑ Context:
  - ○ Distributed
  - ○ Asyncronous
  - ○ New concepts
  - ○ Requirements not fixed at the beginning
  - ○ Secure
- ❑ Objectives
  - ○ Flexible
  - ○ Sustainable in the future
- ❑ Solution
  - ○ Based on open source software
  - ○ Not reinventing the wheel
  - ○ Modular

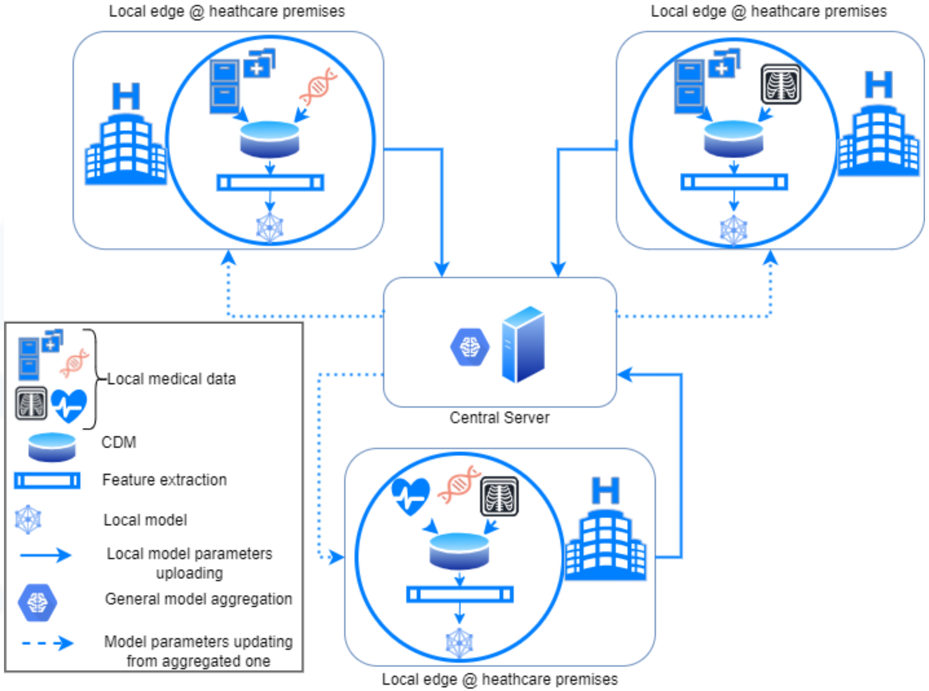
Architecture components

Federated learning library

- ❑ Core part is the federated learning process
  - ○ Reviewed different options with a set of criterias
  - ○ Flower
    - ■ Open source
    - ■ Minimal intrusion, just one thing and well done
    - ■ Flexible, different python based libraries and also different programming language
    - ■ Performant (based on gRPC and protobuffer)

GenoMed4ALL FL platform

Local and central edges structure


Local edge @ heathcare premises
Local edge @ heathcare premises
Central Server
Local edge @ heathcare premises
Local medical data
CDM
Feature extraction
Local model
Local model parameters uploading
General model aggregation
Model parameters updating from aggregated one

Operation	GenoMed4All
Local training	Local nodes
Parameter aggregation	Central node

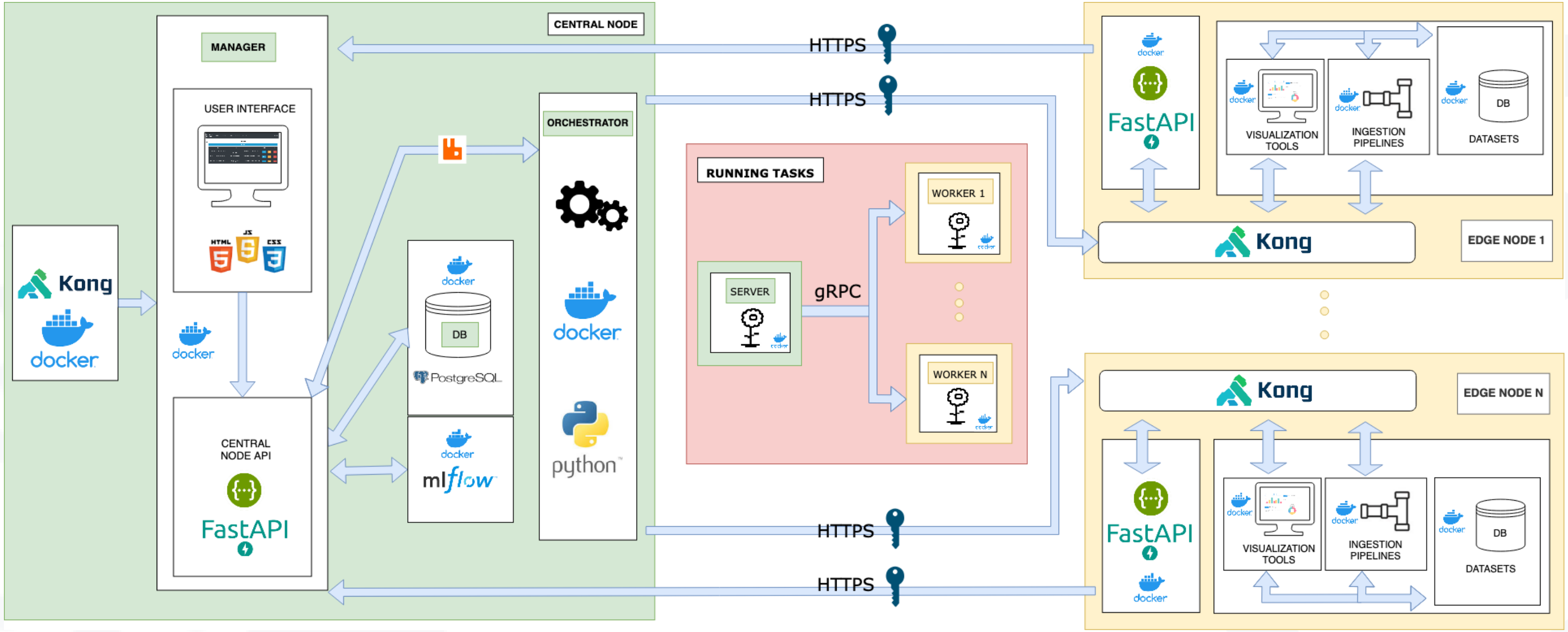
Architecture components

Building a platform for federated learning is complex

- ❑ Communications within the federated platform, for defining jobs, check status,etc.
  - ○ Fastapi (python web framework)
- ❑ User friendly and responsive UI for generating Single page application
  - ○ Vue
- ❑ Machine learning model management
  - ○ Mlflow
- ❑ Metadata storage
  - ○ Postgresql
- ❑ Easy to deploy in different scenario and contexts (Hospitals)
  - ○ Container (docker)
- ❑ Security for authentication
  - ○ Keycloak

GenoMed4ALL FL platform

Technical architecture



GenoMed4All approach (III):

use of the platform

Wrapping all together

Building a platform for federated learning is complex

- ❑ There are two main parts
  - ○ Data provisioning
  - ○ Machine learning training
- ❑ They run at different speed
  - ○ Data provisioning starts from the Raw data and it can take days to get processed and ready to be used by training
  - ○ Platform requires data to be immediately available for training
- ❑ Solution
  - ○ Link the two parts a data level, when data is ready is registered into the platform

Wrapping all together

Building a platform for federated learning is complex

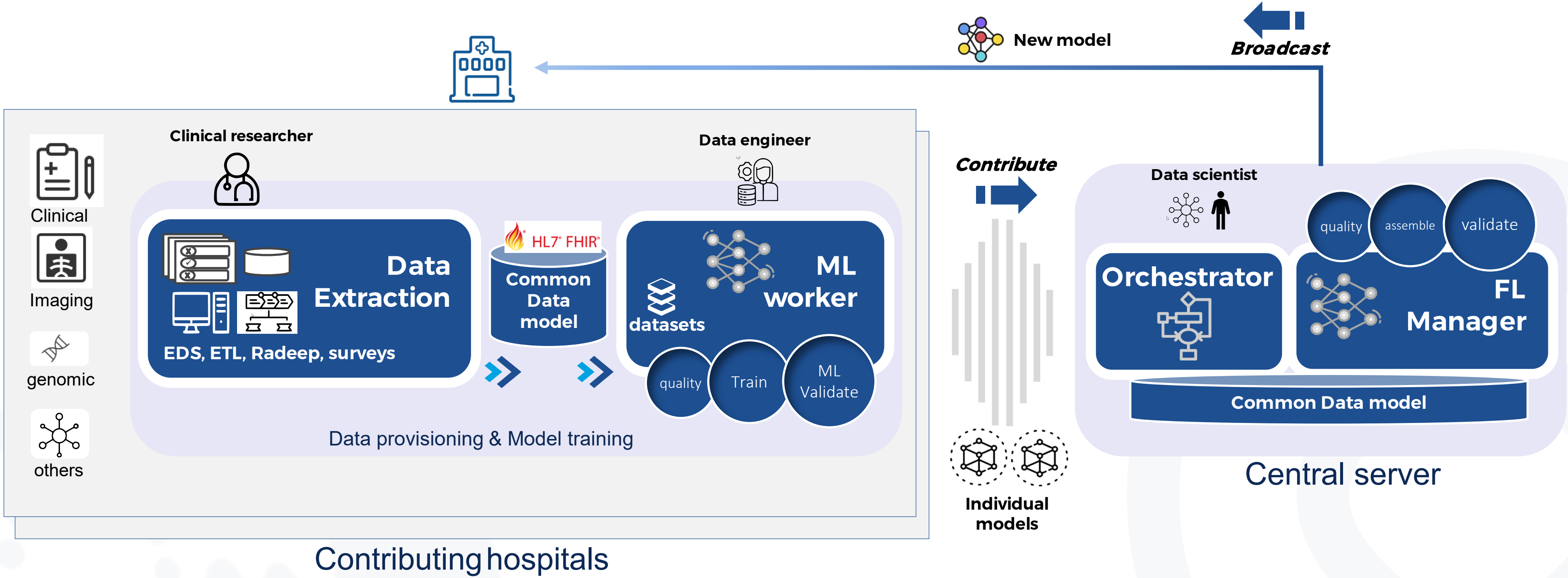
- ❑ 3 main users
    - ○ Data custodian/owner use case
        - Some datasets to share
        - Process the data by using the "same/standard" pipeline
        - Upload data to common data model
        - Export the data as file dataset (i.e csv)
        - Register the dataset
    - ○ Data scientist (outside the platform)
        - Get syntethic dataset or public dataset
        - Develop locally an algorithm
        - Federate the algorithm
        - Run algorithm and dataset through the validation protocol

Wrapping all together

Building a platform for federated learning is complex

- Data scientist (inside the platform)
  - Select algorithm
  - Select datasets
  - Run the training
  - Validate the model training
- Clinician
  - Use the model predictor with incoming data
  - ….

Genome4All solution– conceptualview



Conclusions

Lessons learnt

After some months we can say…

- ❑ Platform design
  - ○ Gather the requirements from different stakeholders is not easy
- ❑ Platform development &deployment
  - ○ Designing, implementing and deploying this type of solution is extremely complex as it is an asynchronous distributed system
  - ○ Development, integration, testing and deployment is facilitated if the necessary infrastructure is in place
  - ○ Close relationship &collaboration with Hospital IT department required to facilitate the deployment of the solution
- ❑ Data provisioning
  - ○ Collecting initial Dataset (for first model development) must be carefully anticipated
  - ○ Data transformation is a crucial stage: upgrade of the EHR to be included in the registries

Thanks!

Any questions?

GenoMed4All & ERN-EuroBloodNet

Educational Program on Artificial Intelligence for public-at-large

Follow us!

genomed4all.eu

@genomed4all

/genomed4all

eurobloodnet.eu

@ERNEuroBloodNet

/ERNEuroBloodNet

Acknowledgements

European Reference Network
for rare or low prevalence complex diseases

Network
Hematological
Diseases (ERN EuroBloodNet)

Co-funded by
the European Union

This project is supported by the European Reference Network on Rare Haematological Diseases (ERN EuroBloodNet) - Project ID No 101085717. ERN-EuroBloodNet is partly co-funded by the European Union within the framework of the Fourth EU Health Programme.

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them.

GENOMED4ALL

Funded by
the European Union

GenoMed4All has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017549.