Dario Gregori

dario.gregori@unipd.it

Unit of Biostatistics, Epidemiology, and Public health

University of Padova

Università degli Studi di Padova

Dipartimento di Scienze Cardio-Toraco-Vascolari e Sanità pubblica

UBEP

Unità di Biostatistica, Epidemiologia e Sanità Pubblica

Unit of Biostatistics, Epidemiology and Public Health

UBEP

efsa

ecdc
EUROPEAN CENTRE FOR DISEASE PREVENTION AND CONTROL

IASLC

Faculty members:

- 11 (Full, Associate and Assistants)
- 22 Medical Residents in Medical Statistics
- 35 PhD Students
- 47 Fellows, Scholars, Visiting

LeSexp

Statistics in Experimental Sciences – LeSEXP

Epidemiological Methods and Biostatistics – LEMBS

Clinical Epidemiology and Digital Health – LACED

LAIMS

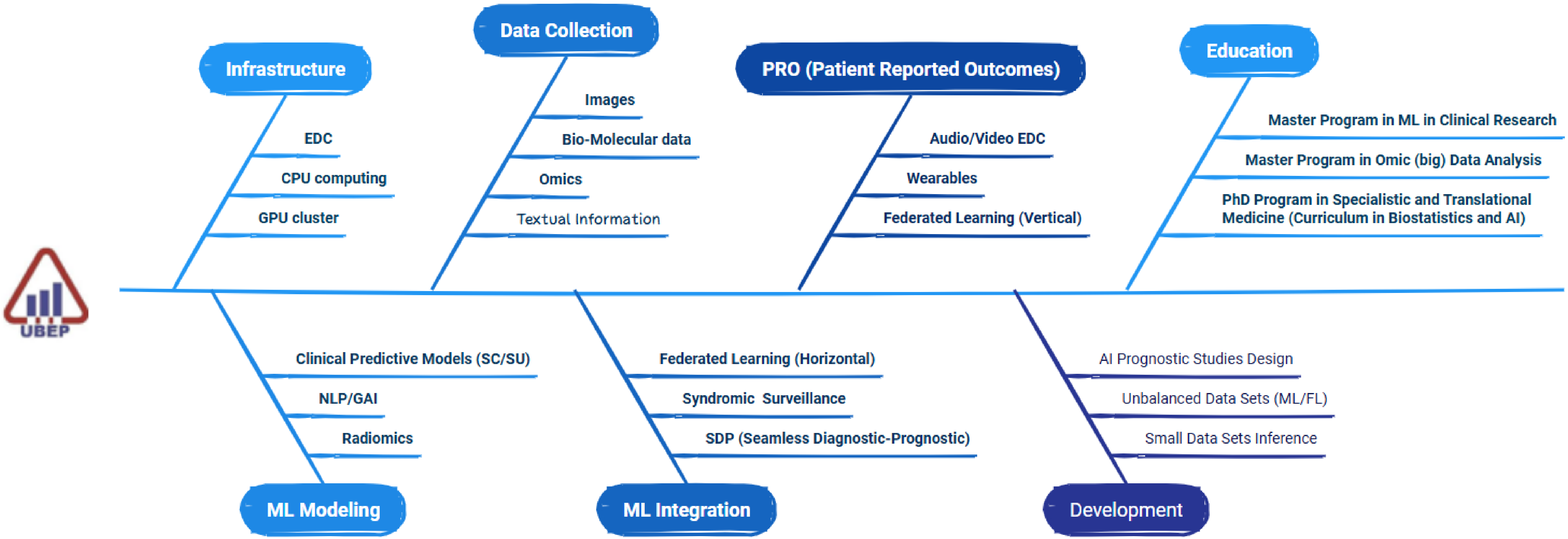
Artificial Intelligence for Medical Sciences – LAIMS

LAboratory of Health SErvice Research

LaSeR – Laboratory of Health Service Research
Unit of Biostatistics, Epidemiology and Public Health
Department of Cardiac Thoracic Vascular Sciences and Public Health

Health Service Research – LaSER

Ecological Statistics and Environmetrics – LESE


Infrastructure
EDC
CPU computing
GPU cluster
Data Collection
Images
Bio-Molecular data
Omics
Textual Information
PRO (Patient Reported Outcomes)
Audio/Video EDC
Wearables
Federated Learning (Vertical)
Education
Master Program in ML in Clinical Research
Master Program in Omic (big) Data Analysis
PhD Program in Specialistic and Translational Medicine (Curriculum in Biostatistics and AI)
UBEP
Clinical Predictive Models (SC/SU)
NLP/GAI
Radiomics
ML Modeling
Federated Learning (Horizontal)
Syndromic Surveillance
SDP (Seamless Diagnostic-Prognostic)
ML Integration
AI Prognostic Studies Design
Unbalanced Data Sets (ML/FL)
Small Data Sets Inference
Development

Past/Current

- GWAS statistical methods review
- LLM Applications to GWAS data
- Food identification with smatwatch sensors
- Genetic predictors and flux (animal to human) of antibiotic resistance
- Novel genetic modifier of DMD identified through GWAS and WES approaches

Current

- Spatial single cell transcriptomics data simulation
- Multilevel data optimal regression strategies identification
- Multi-omics + radiology integration in mesothelioma (TRASCAN EU Grant)
- ML-based, Pubmed text-to-gene in silico analysis of ALS gene expression in the heart
- Whole-genome sequencing analysis in DMD patients with neurodevelopmental impairment

Current

- Spatial single cell clustering benchmark
- LLM automatic classification for RNAseq experiments
- Hystopathology slides ML analysis
- Single-cell analysis of muscle biopsies from DMD patients

Omics-Driven Precision Medicine & Federated Learning Synergies

Integrated Research Axes		FL Adds Unique Value
Multi-omics Integration	• Transcriptomics, proteomics, metabolomics • Bulk & single-cell RNA-seq, spatial transcritomics • Whole-gehome/exome sequencing, GWAS	→ Privacy-breserying, cross-cohort data fusion FL lets hospitals & labs co-train mo-dels on sensitive omics without
Biomarker Discovery & Outcome Prediction	• Mass-spectrometry biomarkers (proteo- & metabomics) • Genetic modifiers in neuromuscular disease • ML & Bayesian risk models for cardiothorarcic-& neuro outcomes	**Faster validation across sites** federated feature-selection & rist-scoring ploelines iden/throbust biumarkers on larger, distributed samples.
AI / LLM-Enhanced Systems Biology	• Large-language-model integration for pathway insight • Tidymodels, mixOmics, Seurat, PLINK, GATK, DESeq... (R/Python)	**Continual learning** FL stream model updates from each nods, keeping LLMs & prediction models current as new data arrive.
Data Ecosystem	• GEO, ArrayExpress, PRIDE, MetaboLights, Metabolomics WB • GnomAD, 1000 Genomes, GTEx, ClinVar, OMIM • Clinical registries (Euro-Fontan, Euro-AAOCA), rare-disease cohorts	Scalable multi-centre trials Distributed GWAS, omics-guided RCTstratification, and lederated adaptive designs accelerate evidence generation.
Applied Fields	• Congenital heart disease & pediatric surgery • Neonatal neurodeveloomen & lung pathology	

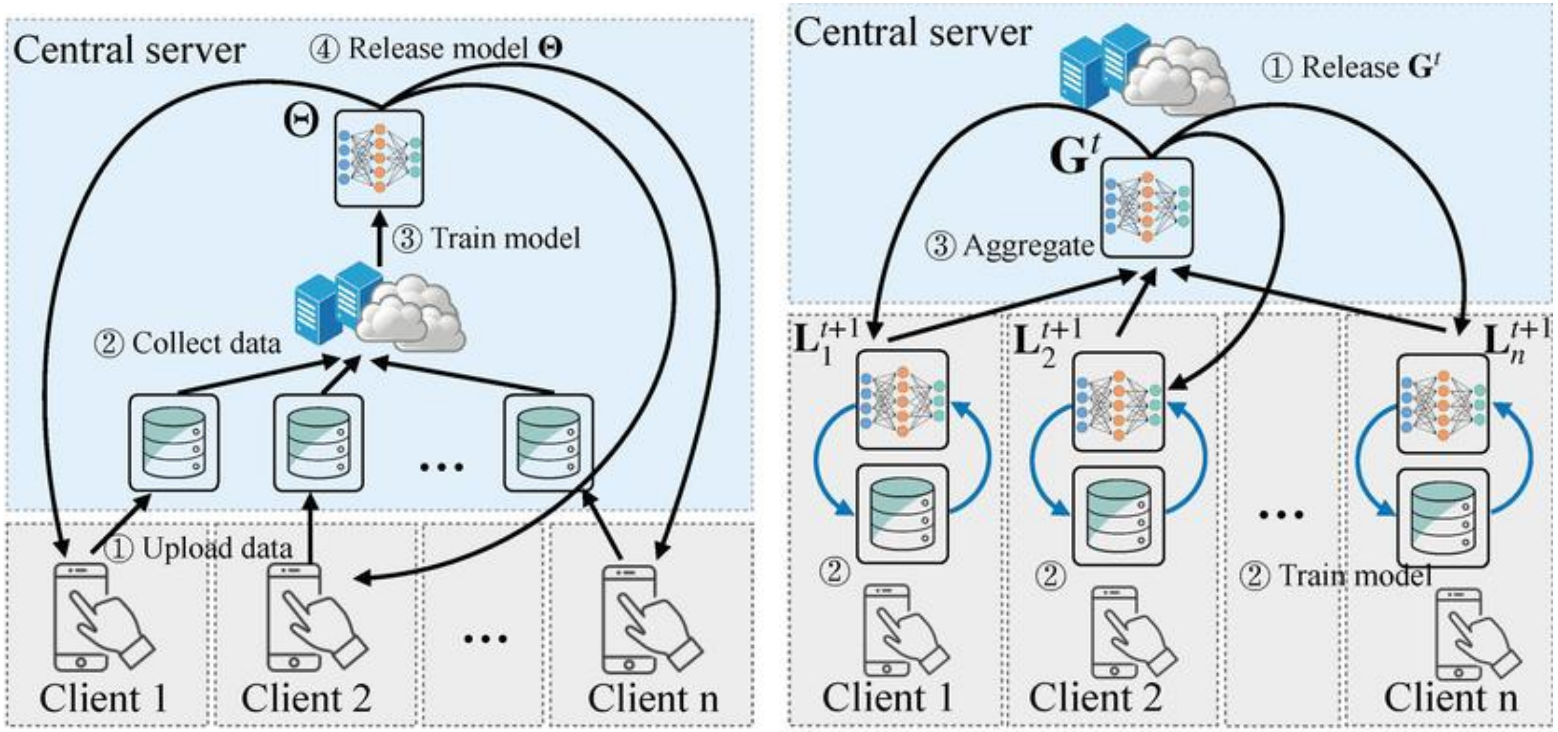
Federated Learning

ntroduction and definition of Federated earning

Federated learning (FL): distributed machine learning approach to address the problem of data governance and privacy in multiple institutions or devices, collaboratively.
No raw data exchange, only learned model parameters or updates, sensitive data on-premise.

General formula: global loss function via w_k weighted combination of K local losses computed from private data X_k:

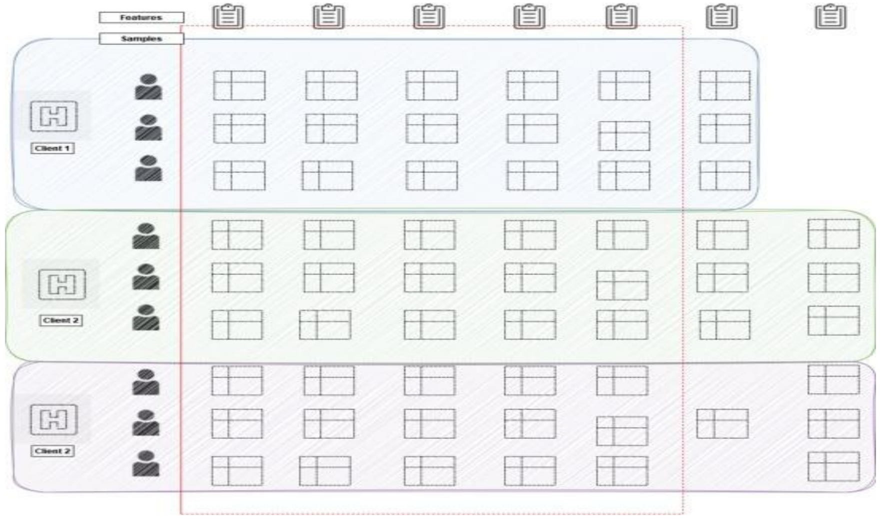
$$\min_{\phi} \mathcal{L}(X;\phi) \quad \text{with} \quad \mathcal{L}(X;\phi) = \sum_{k=1}^{K} w_k \, \mathcal{L}_k(X_k;\phi),$$

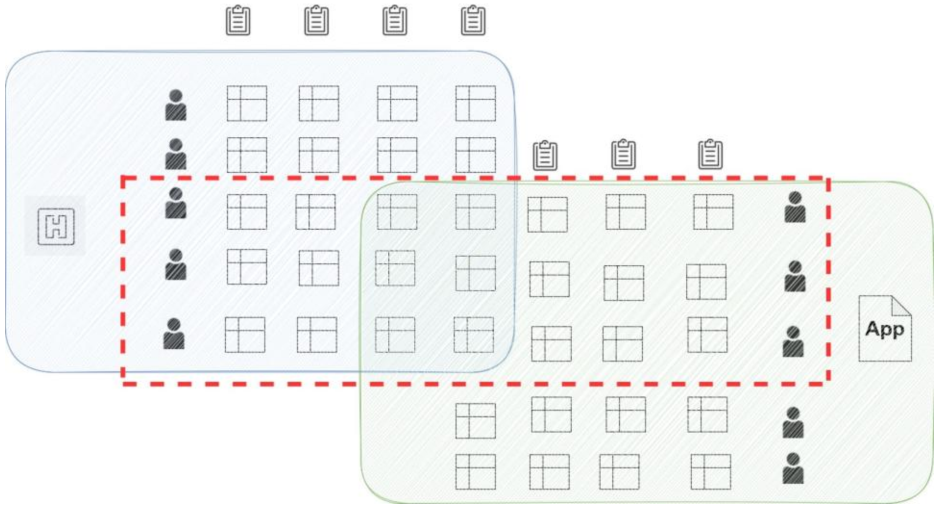


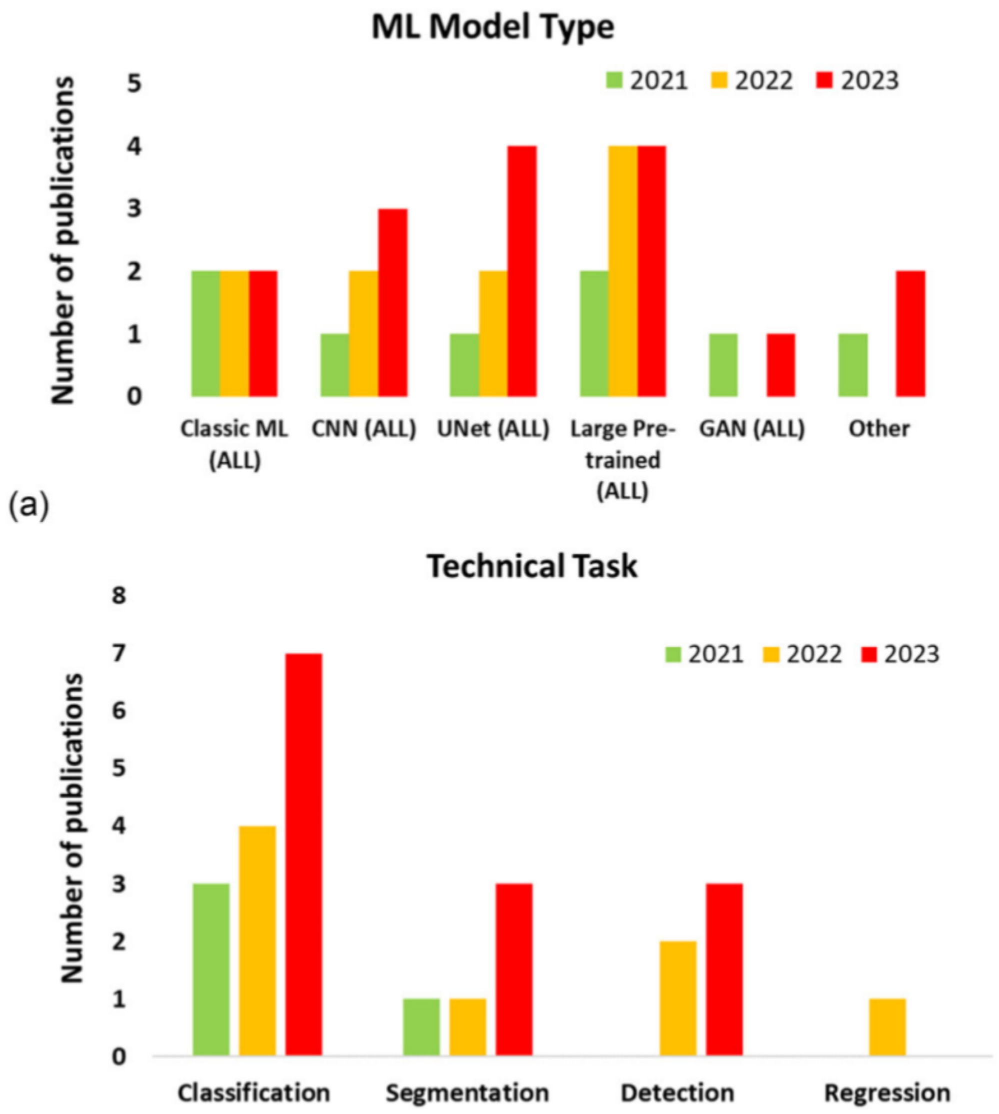
(a) Centralized learning (b) Federated learning

Categorization of Federated Learning

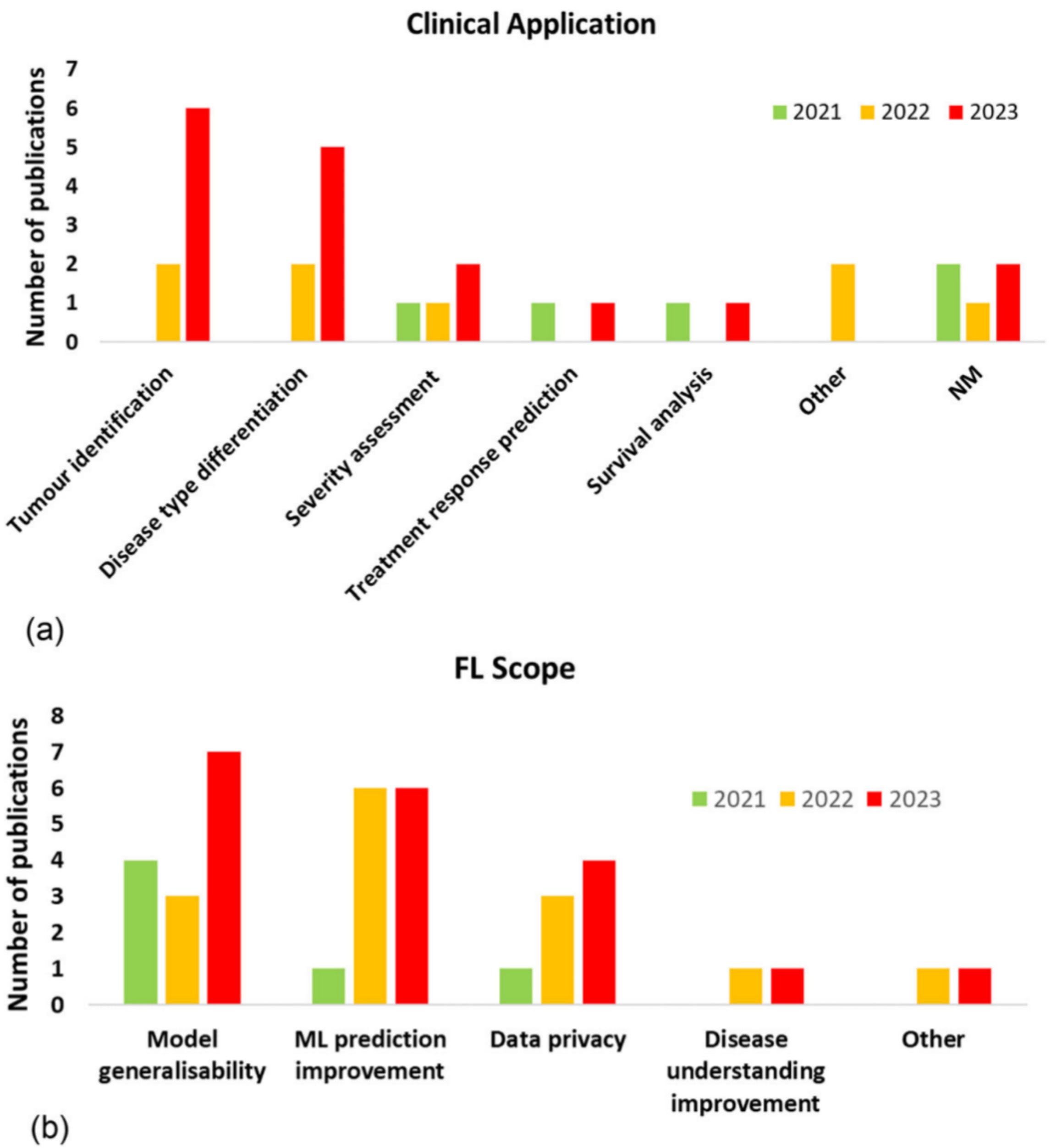
- ***Horizontal FL***:
  Different records (Patient ID)
  Same features (variable)
- ***Vertical FL***:
  Different features
  Same people.
- ***Federated Transfer Learning***:
  Different data and feature spaces
  (or very little overlap)



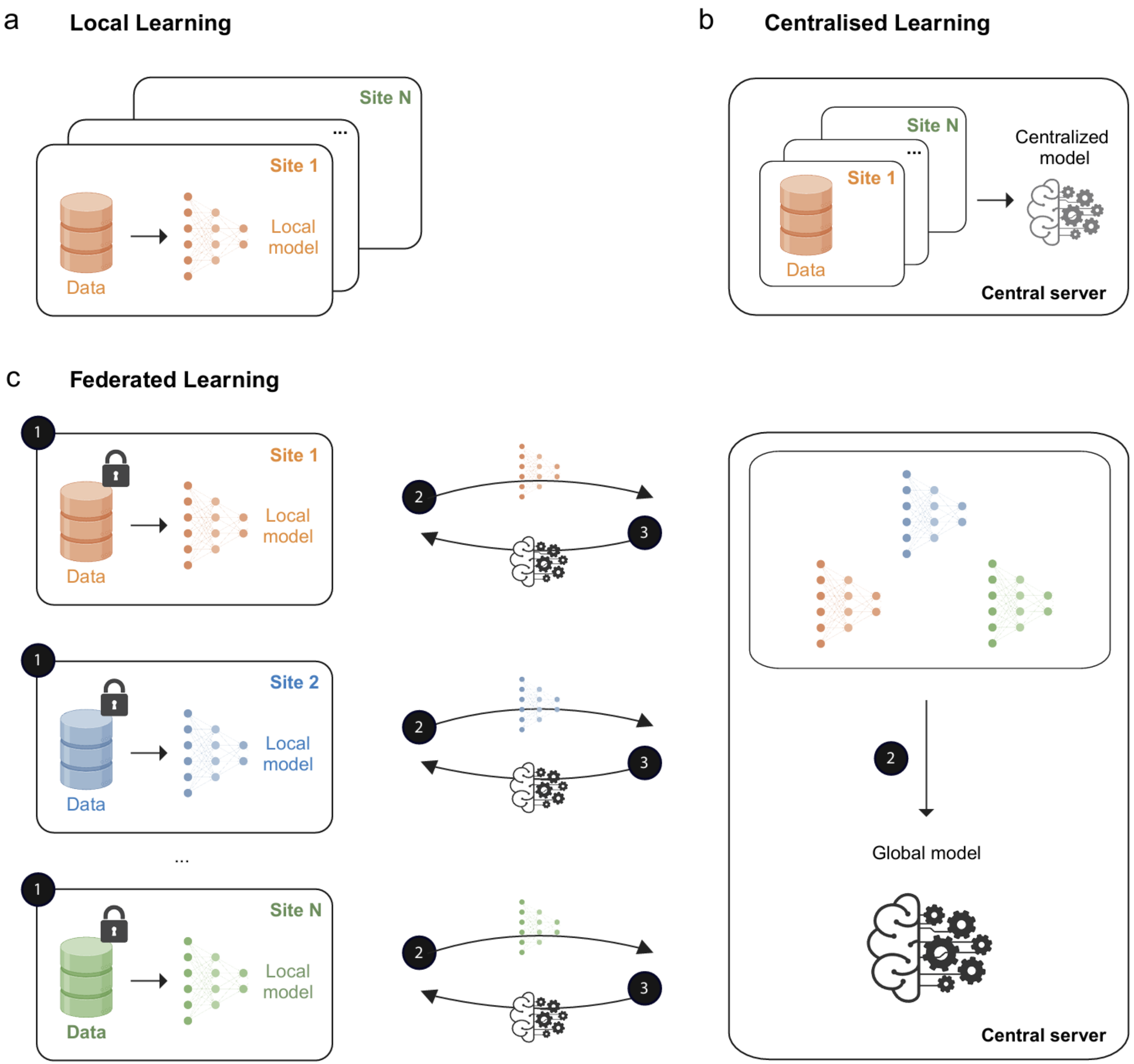


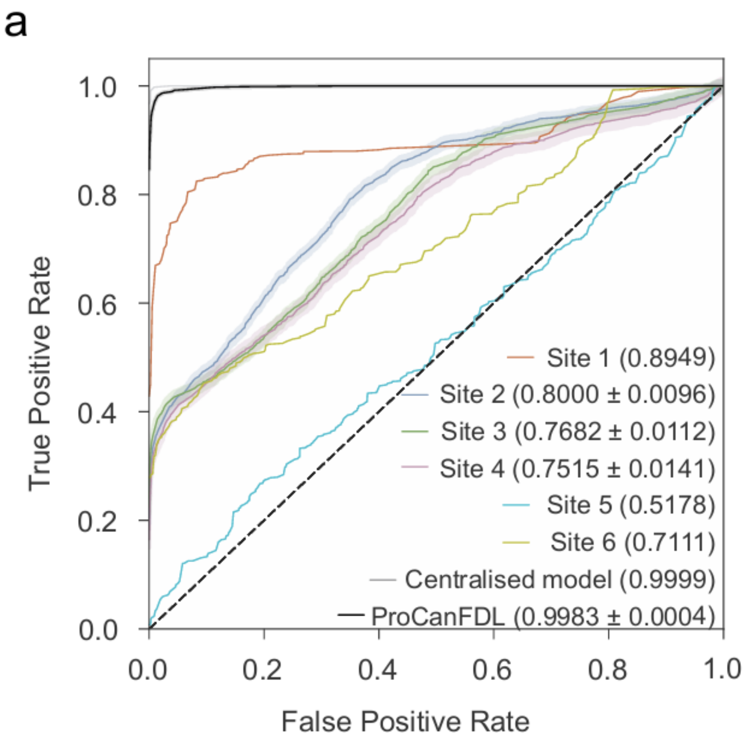


Ankolekar, A., Boie, S., Abdollahyan, M. et al. Advancing breast, lung and prostate cancer research with federated learning. A systematic review. npj Digit. Med. 8, 314 (2025). https://doi.org/10.1038/s41746-025-01591-5



Ankolekar, A., Boie, S., Abdollahyan, M. et al. Advancing breast, lung and prostate cancer research with federated learning. A systematic review. npj Digit. Med. 8, 314 (2025). https://doi.org/10.1038/s41746-025-01591-5





Zhaoxiang Cai, Emma L. Boys, Zainab Noor, Adel T. Aref, Dylan Xavier, Natasha Lucas, Steven G. Williams, Jennifer MS. Koh, Rebecca C. Poulos, Yangxiu Wu, Michael Dausmann, Karen L. MacKenzie, Adriana Aguilar-Mahecha, Carolina Armengol, Maria M. Barranco, Mark Basik, Elise D. Bowman, Roderick Clifton-Bligh, Elizabeth A. Connolly, Wendy A. Cooper, Bhavik Dalal, Anna DeFazio, Martin Filipits, Peter J. Flynn, J Dinny. Graham, Jacob George, Anthony J. Gill, Michael Gnant, Rosemary Habib, Curtis C. Harris, Kate Harvey, Lisa G. Horvath, Christopher Jackson, Maija R J. Kohonen-Corish, Elgene Lim, Jia (Jenny) Liu, Georgina V. Long, Reginald V. Lord, Graham J. Mann, Geoffrey W. McCaughan, Lucy Morgan, Leigh Murphy, Sumanth Nagabushan, Adnan Nagrial, Jordi Navinés, Benedict J. Panizza, Jaswinder S. Samra, Richard A. Scolyer, John Souglakos, Alexander Swarbrick, David Thomas, Rosemary L. Balleine, Peter G. Hains, Phillip J. Robinson, Qing Zhong, Roger R. Reddel; Federated deep learning enables cancer subtyping by proteomics. Cancer Discov 2025; https://doi.org/10.1158/2159-8290.CD-24-1488

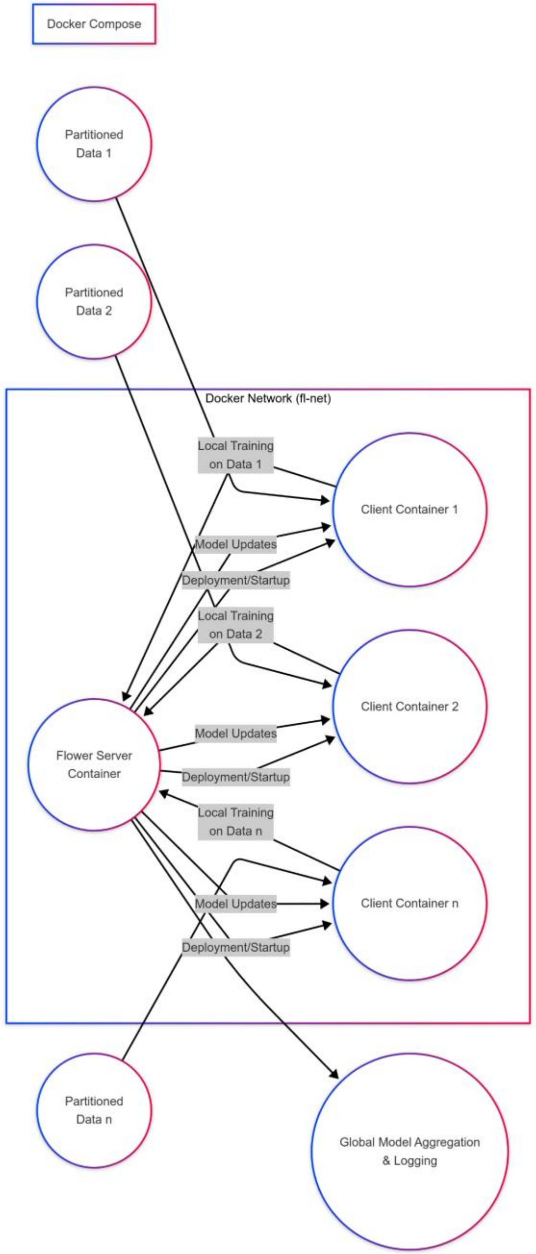
GDPR Requirements for Observational Research

- **Legal Framework**
  - GDPR (EU 2016/679): Governs personal data processing throughout the EU.
- **Key Principles**
  - Lawfulness, Fairness, and Transparency: Participants must be informed clearly about how data is collected and used.
  - Data Minimization: Collect only what is strictly necessary for research.
  - Purpose Limitation: Use personal data exclusively for the stated research goals.
- **Consent & Informational Notice**
  - Informed Consent: If identifiable data is collected, ensure participants explicitly consent (where applicable).
  - Privacy Notice: Provide clear information on data handling, retention, and subject rights.
  - Anonymization / Pseudonymization
  - Scientific Validity: Balance privacy measures with the scientific objectives of observational research.
- **Data Processing & Storage Security Measures:**
  - Implement robust physical and digital safeguards.
  - Retention Policies: Store data only as long as necessary for the research purpose.

Aspect	Traditional Centralised ML	Federated Learning (FL)
Privacy & Security	Raw data pooled ⇒ single-point breach risk	Data stay local; share only model weights
Regulatory fit	Cross-border transfers onerous; high liability	Aligns with GDPR/EHDS; lower legal burden
Infrastructure & Cost	Large central storage & high-bandwidth uploads	Minimal central storage; lightweight updates; distributed compute
Implementation	Simple tech stack, heavy data-harmonisation logistics	Orchestrator needed, but **plug-and-play** for new sites
Model performance	Often limited by incomplete pooling	Near-pooled accuracy on full virtual dataset
Generalisability	Prone to site bias & domain shift	Robust across diverse hospitals/patient groups
Collaboration & Scale	Hard to add sites; localisation barriers	Scales to 100 + nodes; fosters global consortia

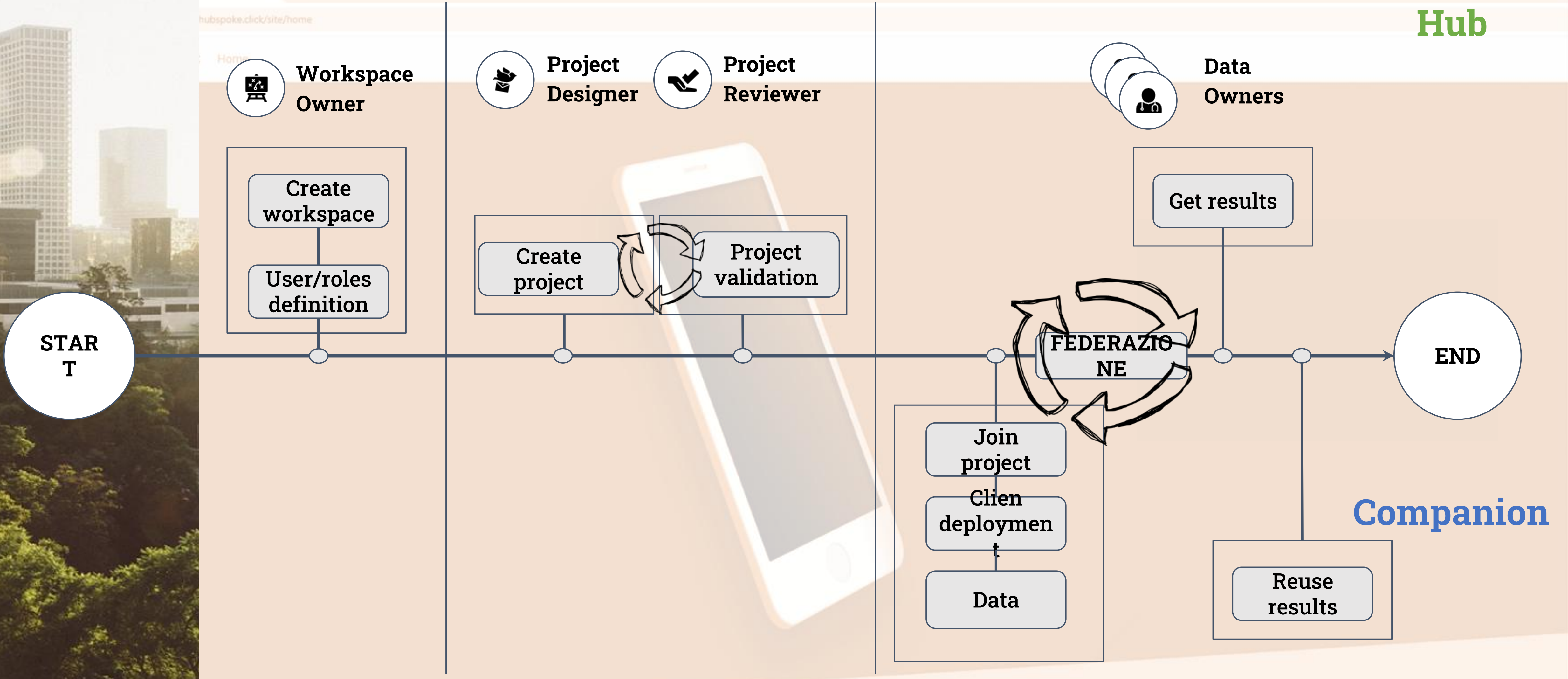
FL implementing platform @UBEP

- **Horizontal** Federated Learning Framework (**Flower**)
  - Facilitates multi-node experimentation and addresses client heterogeneityLanguage-agnostic, enabling easy integration of diverse platforms
  - Supports cross-device federation across multiple clients
- Containerization with **Docker**:
  - Each client runs in a lightweight Alpine-based image containing:
    - TensorFlow 2.12
    - Flower 1.4
    - Pre-processed medical center dataset
- Server container handles:
  - Global model aggregation
  - Logging and orchestration
- Networking and Deployment:
  - A custom Docker network (fl-net) secures communication between server and clients
  - Docker Compose automates deployment and scaling:
    - Initializes server and clients with predefined FL settings (rounds, epochs, batch size)
    - Isolates the entire FL system from external traffic for data privacy

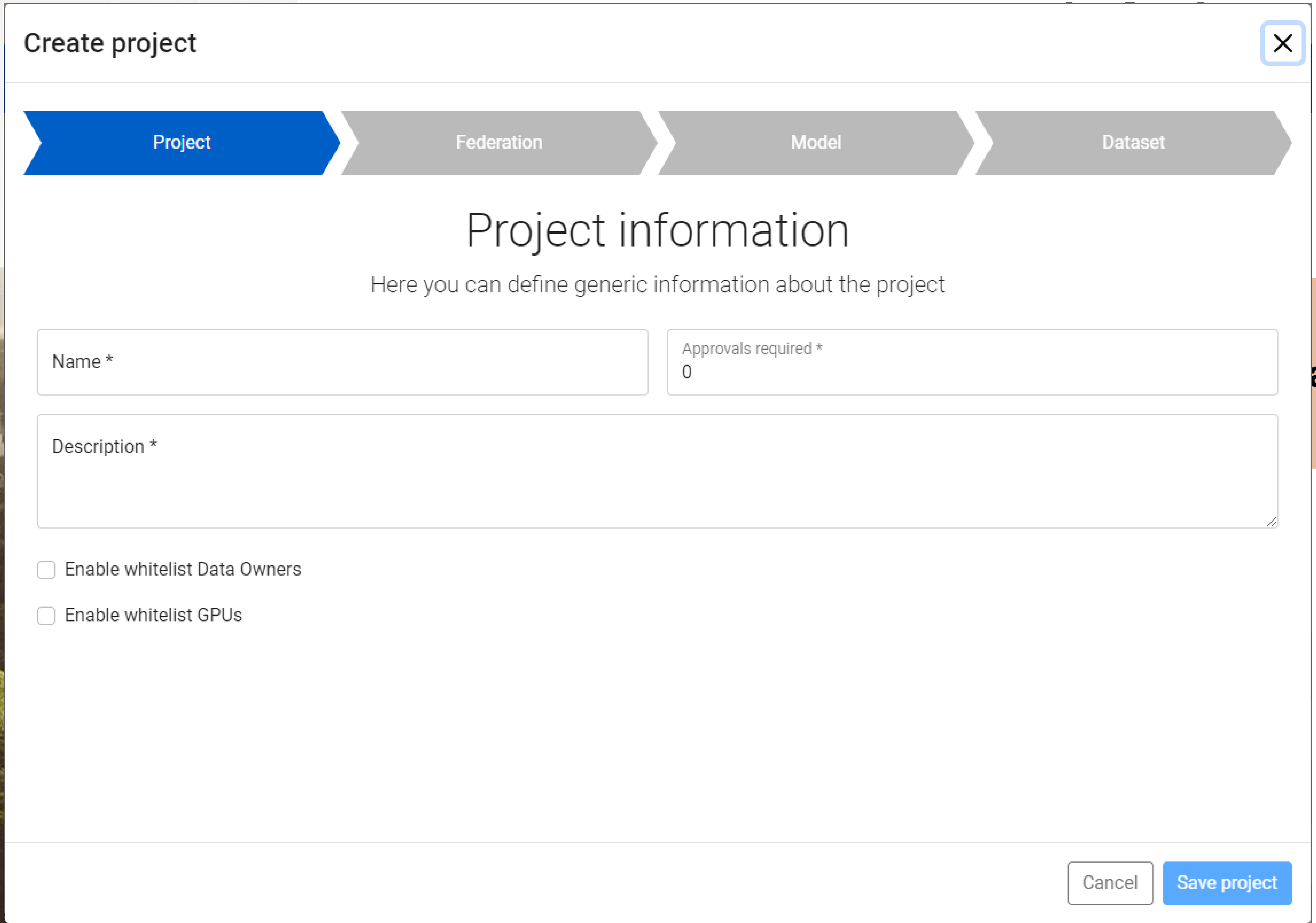


Federated Biomedical Research

User Flow



Federated Biomedical Research

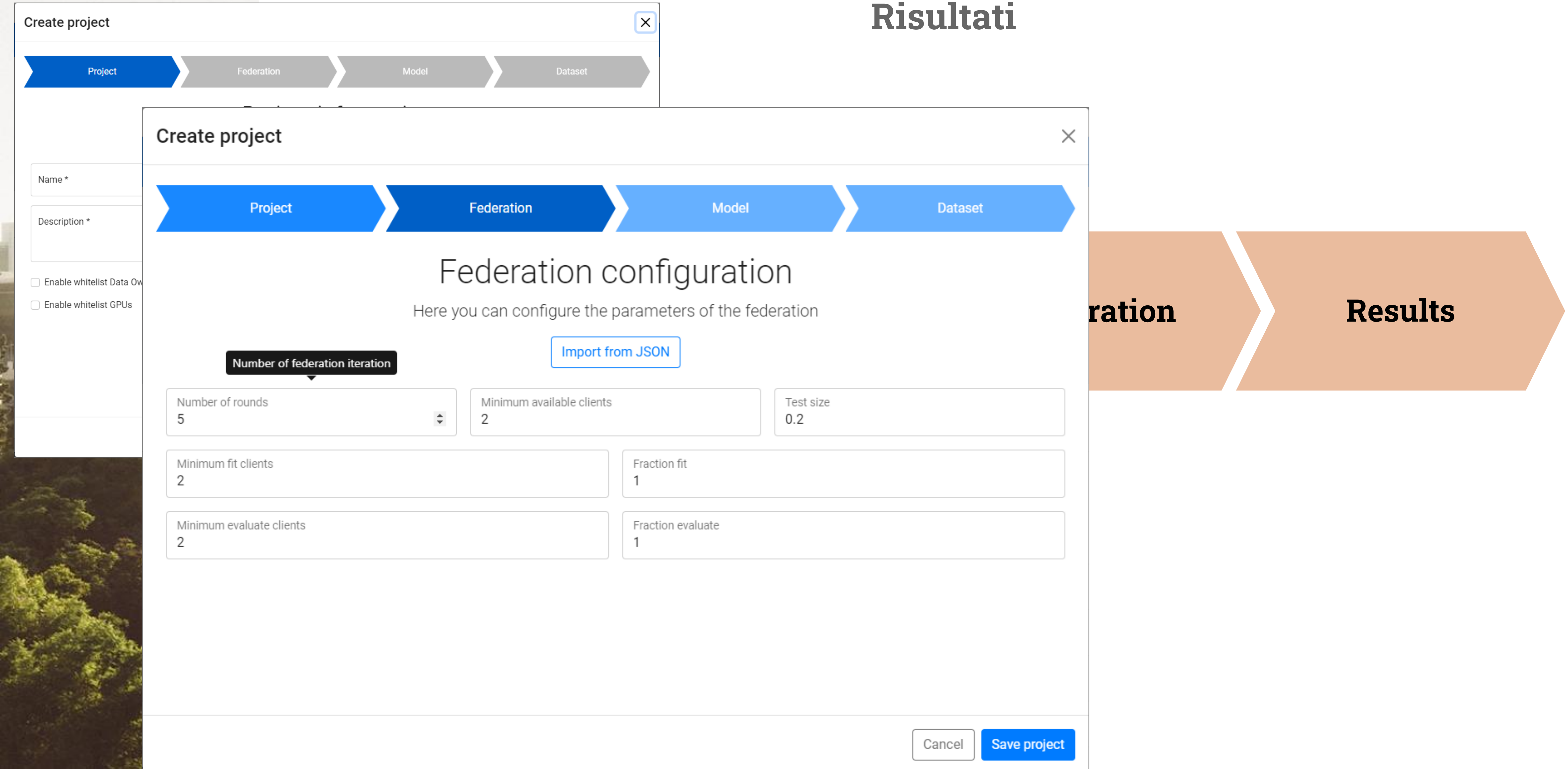


ation

Results

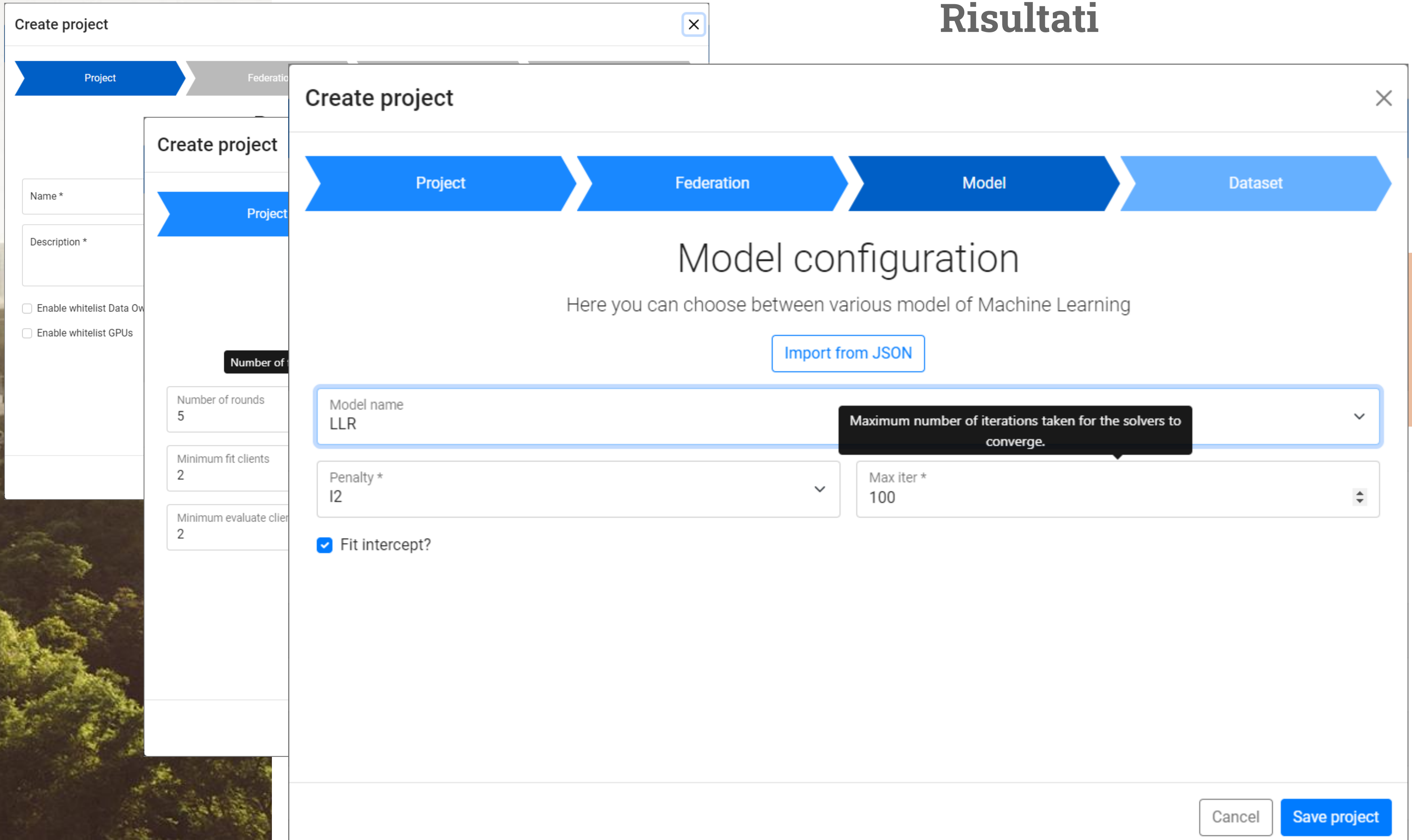
Federated Biomedical Research

Risultati



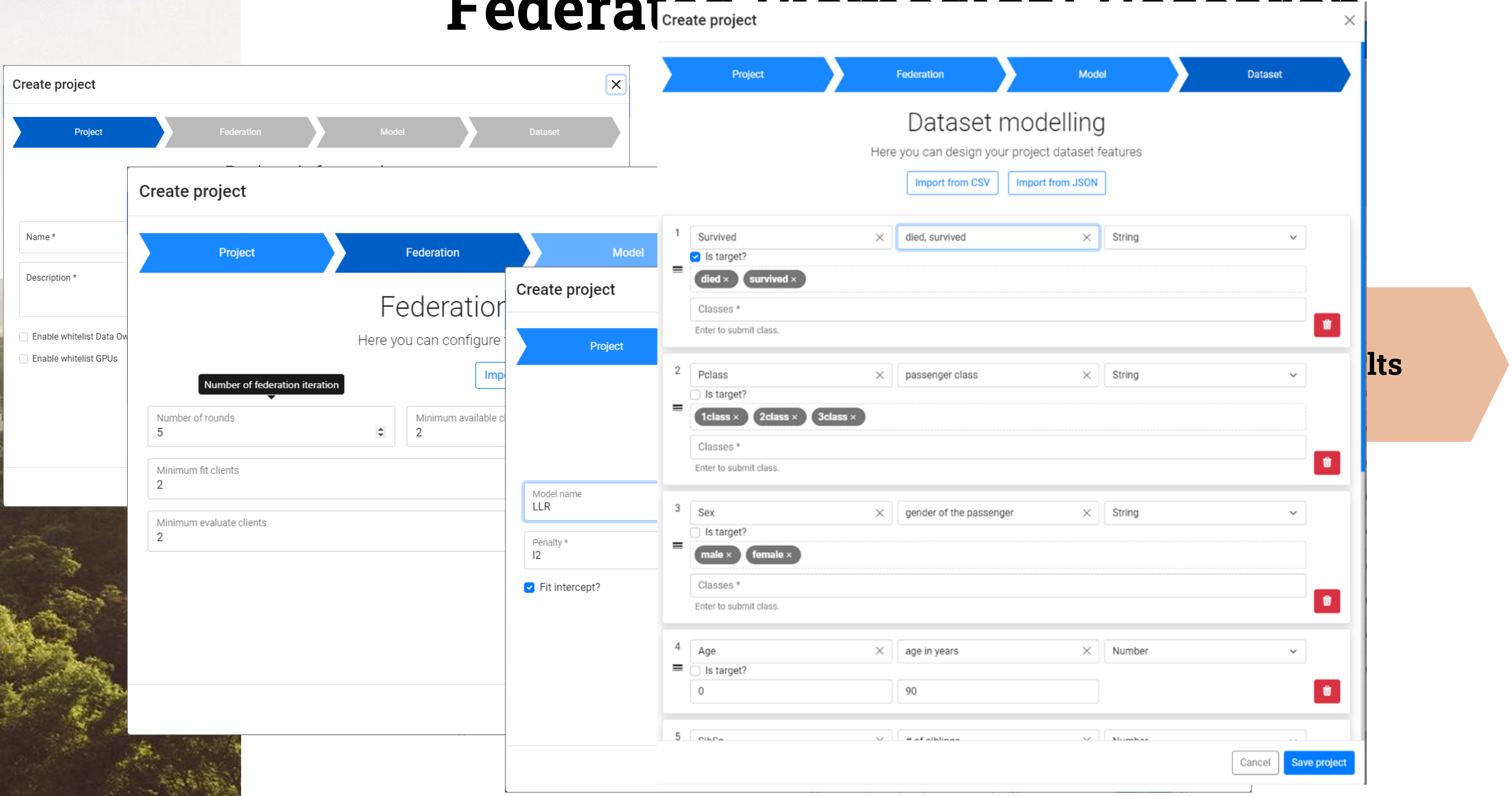
Federated Biomedical Research

Risultati



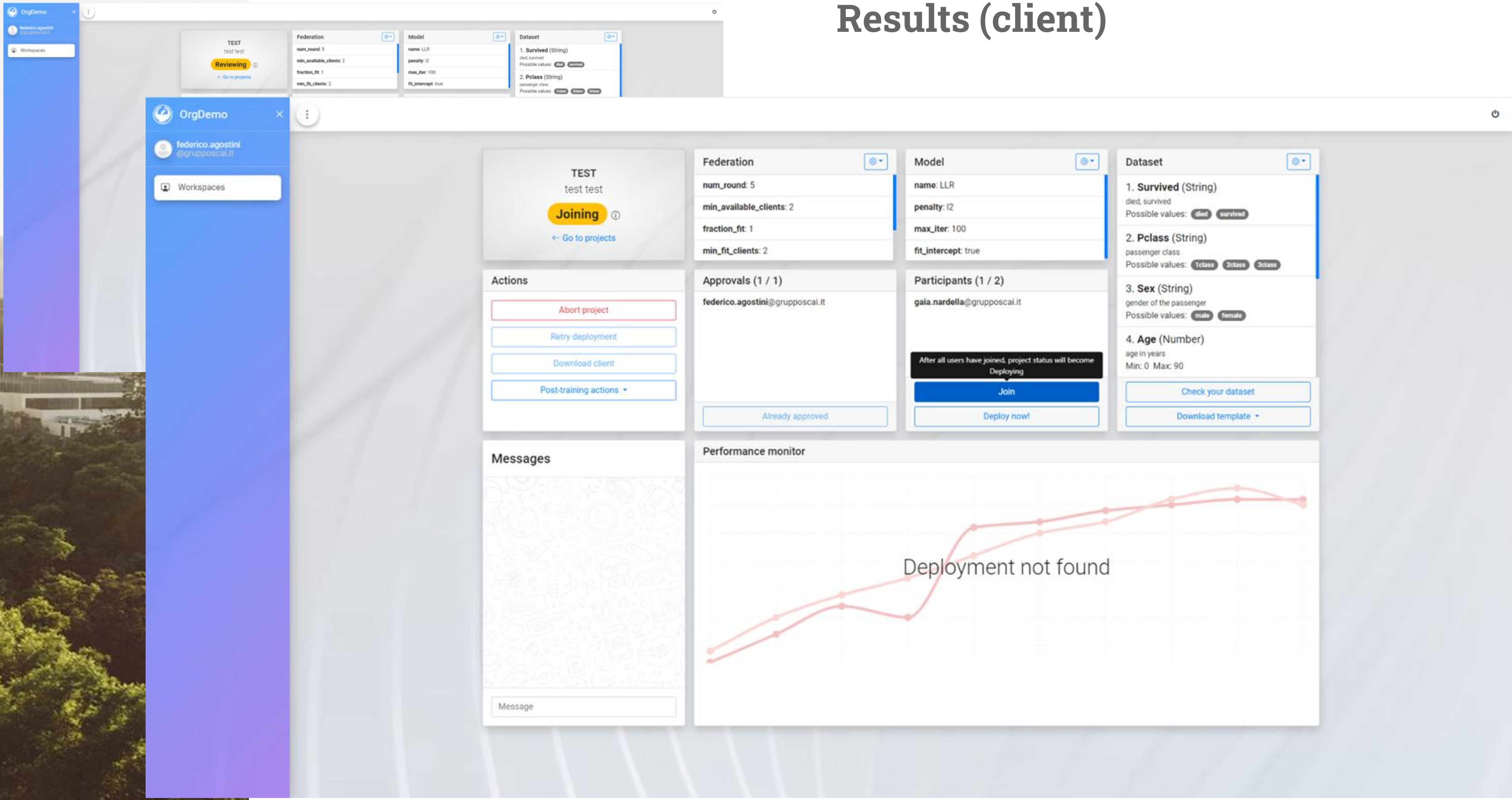
Results

Federated Biomedical Research



Federated Biomedical Research

Results (client)



Federated Biomedical Research

Results (Hub)

Case Study: IRP

Istituto di Ricerca Pediatrica (IRP) Padova

- The “Torre della Ricerca” hosts the Istituto di Ricerca Pediatrica (IRP) Città della Speranza, Europe’s largest paediatric research centre, with > 350 researchers working in ten storey laboratories devoted to oncology, gene therapy, regenerative medicine and rare disease genomics.
- IRP comprises ≈ 30 research groups and 200 investigators who generate high dimensional omics datasets (WGS, RNA seq, single cell, proteomics, metabolomics) for precision medicine projects in childhood leukaemia and other diseases.
- In 2024 the institute sequenced ≈ 500 whole genomes, 1000 bulk RNA seq samples and 300 quantitative proteomics runs for its flagship AI powered projects on JMML and BALL stratification.

Dimension	Traditional Centralised ML	Federated Learning (FL) – Torre della Ricerca
Annual data moved	80 TB raw uploads lab → core	10.8 GB weight updates (-99.9 %)
Central storage (5 yr)	400 TB RAID + backup	≈ 0 TB (only models/logs)
New hardware	Existing € 200 k HPC	30 edge GPUs + hub € 195 k
Annual O&M	€ 50 k (power, cooling, IT)	€ 40 k (lower central, + local power)
Expected breach loss / yr	€ 0.18 M (single-point risk)	€ 0.05 M (encrypted, segmented)
GDPR posture	High-risk: minors' genomes centralised; DPIA/GDPR heavy	Data stay in-lab ⇒ "data-minimising"; DPIA/GDPR light
Model freshness	Annual batch retrain	Quarterly / rolling without shipping data
Scalability & partners	Limited; external centres wary	Plug-and-play nodes (Heidelberg, IOV, …)
Strategic upside	—	PIs keep IP; boosts trust & EHDS funding fit
Key challenges	Warehouse bottleneck; single-point failure	Secure gradient privacy; ops for 30 nodes

Volumes below use public file size benchmarks

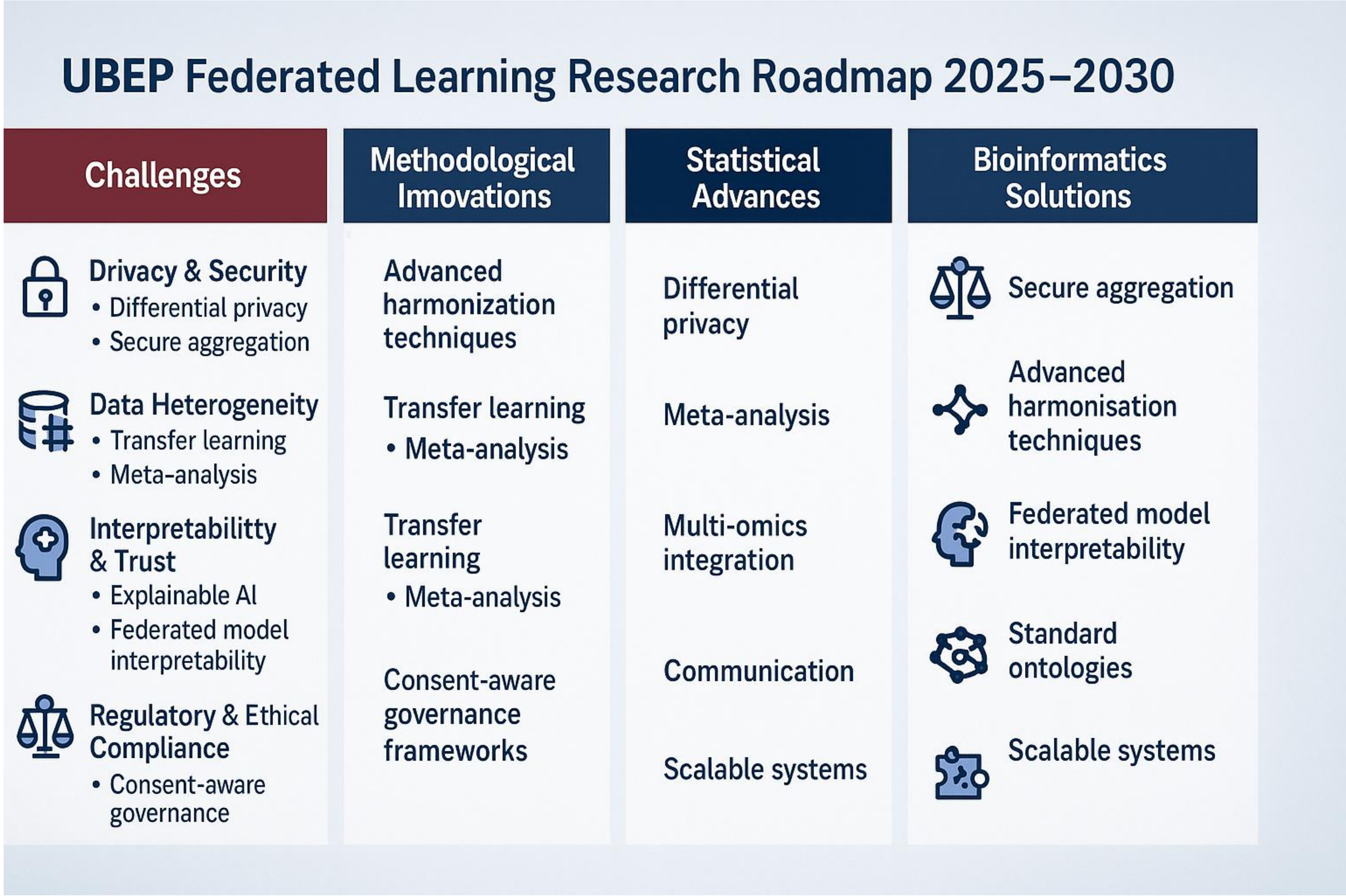
Connection of FL with Horizon calls

Horizon Call (ID & Title)	Partner Institution & Lab Focus	Research Domain	Why a Fit for FL
HORIZON-INFRA-2025-01-SERV-01: Research infrastructure services to enable R&I addressing main challenges and EU priorities related to the health domain	KISTI – DataON platform team (national research data platform for health)	Health data infrastructure	FL enables secure, distributed analysis of health data across multiple research centers (no central data pooling needed).
HORIZON-HLTH-2025-01-DISEASE-04: Leveraging artificial intelligence for pandemic preparedness and response	Seoul National University – Health System Data Science Lab (large-scale epidemiology & AI)	Infectious disease / AI	FL allows cross-border learning from diverse epidemiological datasets, improving pandemic response models while preserving data privacy.
HORIZON-HLTH-2025-01-CARE-01: End user-driven application of Generative Artificial Intelligence models in healthcare (GenAI4EU)	KAIST – BioImaging & Signal Processing Lab (medical generative AI research)	Healthcare AI (Generative models)	FL enables **collaborative training of generative AI models on hospital data** (e.g. medical images, health records) without sharing sensitive patient data.
HORIZON-HLTH-2025-01-TOOL-03: Leveraging multimodal data to advance Generative Artificial Intelligence applicability in biomedical research (GenAI4EU)	Samsung Medical Center – AI Research Center (multi-modal clinical data analysis)	Multi-modal biomedical AI	FL can integrate **multi-site, multi-modal biomedical data** (imaging, genomics, clinical records) to improve AI models, while keeping data siloed at each institution.
HORIZON-HLTH-2025-03-IND-03: [two-stage] Facilitating the conduct of multinational clinical studies of orphan devices and/or of highly innovative (“breakthrough”) devices	Asan Medical Center – Clinical AI Lab (multi-hospital clinical trial data collaboration)	Clinical research / MedTech	FL allows pooling of limited patient data from multiple countries to evaluate novel medical devices, aiding robust analysis without direct data exchange.

Connection of FL with Trials (Horizon calls)

Regulatory Source	Core Randomisation Requirements	How Federated Learning (FL) Meets Them
ICH E9 (§1.2, §3.3)	Unbiased, reproducible, concealed allocation Pre-specified blocks & stratification	Central IRT keeps full randomisation list; FL touches only **analysis layer**, so concealment & reproducibility remain intact
FDA (2023 Covariate-Adjustment Guide)	Stratify on key prognostic factors before randomization Store allocation table in 21 CFR Part 11-compliant system with audit trail	Same IRT captures stratification factors (e.g., baseline NAS, BMI) & audit trail; FL clients use those factors locally for covariate-adjusted models—no data export needed
EMA / EU CTR 536/2014 & ICH E8 R1	Documented algorithm & traceable randomisation in Trial Master File (TMF) Decentralised elements allowed if auditable	Sponsor retains master RNG seed & algo in eTMF; sites fetch next code via secure web-service—independent of FL vs central analytics, **fully traceable**
AIFA “Linee guida sperimentazioni cliniche”, 2024	Clear SOPs; forbid deterministic site allocation Electronic systems must reside in audited EU datacentres	IRT hosted in Azure EU West; API call from each site’s EMR logs allocation for inspectors—FL analysis does not alter this workflow

Horizon Call (ID & short title)	Trial / Translation Focus	How FL Speeds Trial Conduction
DISEASE-01 – RCTs for phage therapy against antibiotic-resistant infections	Multi-centre **randomised controlled trial** testing safety & efficacy of personalised phage cocktails	Securely pools patient-level microbiology + clinical outcomes across hospitals without data transfer. Real-time interim analyses & adaptive randomisation from federated models shorten decision cycles. Eases ethics/ GDPR hurdles, so more sites can join quickly → faster recruitment.
TOOL-05 – Boosting translation of biotech research into innovative health therapies	Early **Phase I/II first-in-human trials** of novel biotech (gene, RNA, nanobody, etc.) therapies with omics biomarkers	Federated integration of multi-omics, imaging, and safety data lets sponsors refine dose/biomarker signals sooner. Supports cross-site adaptive-design analytics while IP-sensitive data stay local. Fewer data-sharing contracts → quicker site activation.
TOOL-01 – Enhancing cell therapies with synthetic biology	Proof-of-concept/Phase I trials of engineered cell therapies; parallel manufacturing QC studies	FL links manufacturing-site QC data with clinical-site safety read-outs to predict batch performance early. Enables federated immune-profiling models that detect adverse events sooner → safer, more efficient escalation.
IND-03 – Multinational clinical studies of orphan / "breakthrough" devices	Prospective, multi-country **clinical performance studies** for high-innovation medical devices	Device-generated imaging/sensor data stay on-premise; federated analysis builds global performance models without moving large files. Harmonises endpoints across jurisdictions, reducing regulatory review loops. Accelerates CE/FDA evidence generation for small-population indications.
DISEASE-07 – Tackling high-burden & under-researched conditions	Pragmatic or hybrid trials/implementation research in rare or neglected diseases	FL joins scattered EHR, registry and -omics data to reach statistical power without centralising sensitive records. Enables continuous evidence updates, making mid-study protocol refinements faster.



UBEP Federated Learning Research Roadmap 2025–2030
Challenges
Drivacy & Security
• Differential privacy
• Secure aggregation
Data Heterogeneity
• Transfer learning
• Meta-analysis
Interpretabilitty & Trust
• Explainable AI
• Federated model interpretability
Regulatory & Ethical Compliance
• Consent-aware governance
Methodological Imovations
Advanced harmonization techniques
Transfer learning
• Meta-analysis
Transfer learning
• Meta-analysis
Consent-aware governance frameworks
Statistical Advances
Differential privacy
Meta-analysis
Multi-omics integration
Communication
Scalable systems
Bioinformatics Solutions
Secure aggregation
Advanced harmonisation techniques
Federated model interpretability
Standard ontologies
Scalable systems

https://www.unipd-ubep.it

Machine Learning & AI
Corrado Lanera
Luca Vedovelli
Daniele Sabbatini
Daniele Gasparini
Pinto Katende Jonathan
Shinto Pulickal

Clinical epidemiology
Giulia Lorenzoni
Ileana Baldi
Gloria Brigiari
Maria Vittoria Vescovo
Konstantina Thaleia Pilali
Honoria Ocagli

Biostatistics
Ester Rosa
Stefania Lando
Jirenia Olani Kedida
Ajsi Kanapari

Project management
Erica Bauducco

Medical informatics
Cinzia Anna Maria Papappiccolo
Godwill Ilunga
Mbulay Onesime
Ali Aqsa