



Centre for Computational
Personalised Medicine
International Research Foundation

*We create computational technologies
for optimised healthcare*

Navigating the challenges of processing distributed medical data

Maciej Malawski, Piotr Nowakowski
Sano Centre for Computational Medicine

The Polish Open Science Conference 2024
Krakow, Apr 11 2024



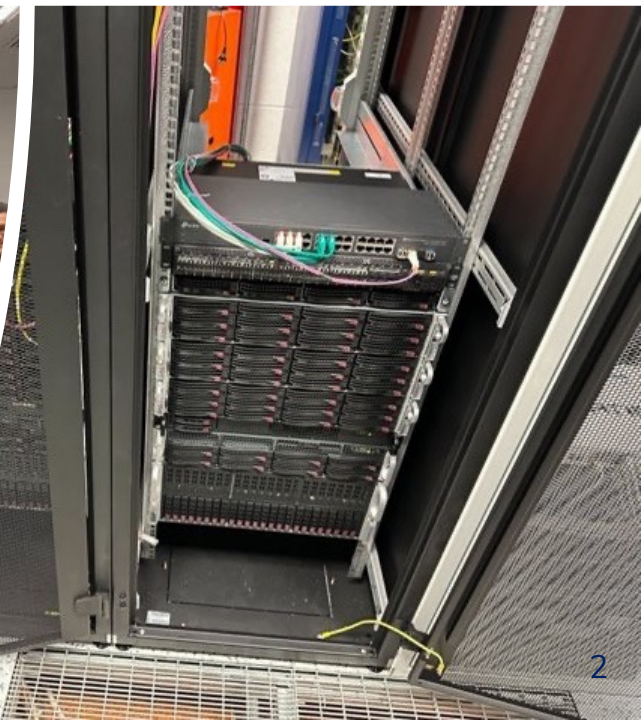
Sano – Centre for Computational Medicine

New research institution in
Krakow

European Centre of Excellence

5 research Teams

90 People



Working with medical data: challenges and solutions sano

- Technological progress, coupled with ongoing digitization of various aspects of social life results in ever increasing quantities of digital data – the ability to process such data paves the way to further civilizational advancements.
- Ongoing digitization is particularly evident in the healthcare domain – with systems such as **e-recepty** (e-prescriptions), **e-skierowania** (e-referrals) or **Internetowe Konto Pacjenta** (Online Patient Account) providing evidence that modern IT solutions may substantially affect the work of healthcare providers and render benefits to the patient.
- We are still at an early stage of the e-health revolution – and we face new challenges related to further digitization of healthcare services, as well as development of AI-based systems to further assist doctors in planning and administering treatment.



Legal context

The legal framework related to using and sharing medical data is being revolutionized with the adoption of new legislation – including the **EU Data Governance Act** and **Data Act**, as well as the **EHDS** initiative, together with national projects

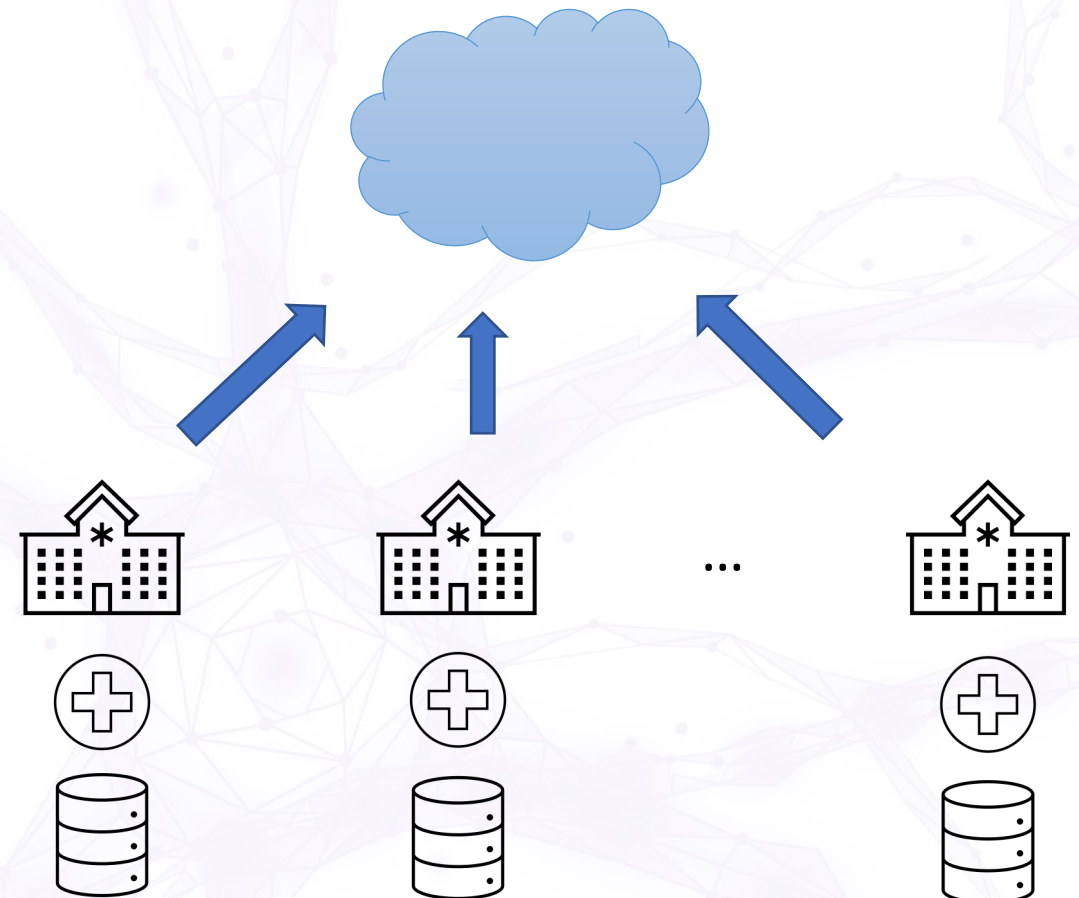
Key objectives include:

- Establishing centralized data registries and common EU-wide interchange formats,
- Safeguarding security of healthcare records,
- Ensuring patients remain in control of their data, enabling physicians to access it when required,
- Open possibilities for so-called secondary use of medical data (including for research)



Problem – accessing required data for research and development purposes

- Silos
 - Local data in hospitals
- Data heterogeneity
 - Various types of data, modalities, device configurations
 - Non uniform and not identically distributed data
- Metadata issues
- Slow adoption of central repositories – varying by country (Poland, Germany, Finland)





Centre for Computational
Personalised Medicine
International Research Foundation

*We create computational technologies
for optimised healthcare*

Example 1

Federated Learning

Federated Learning - Definition

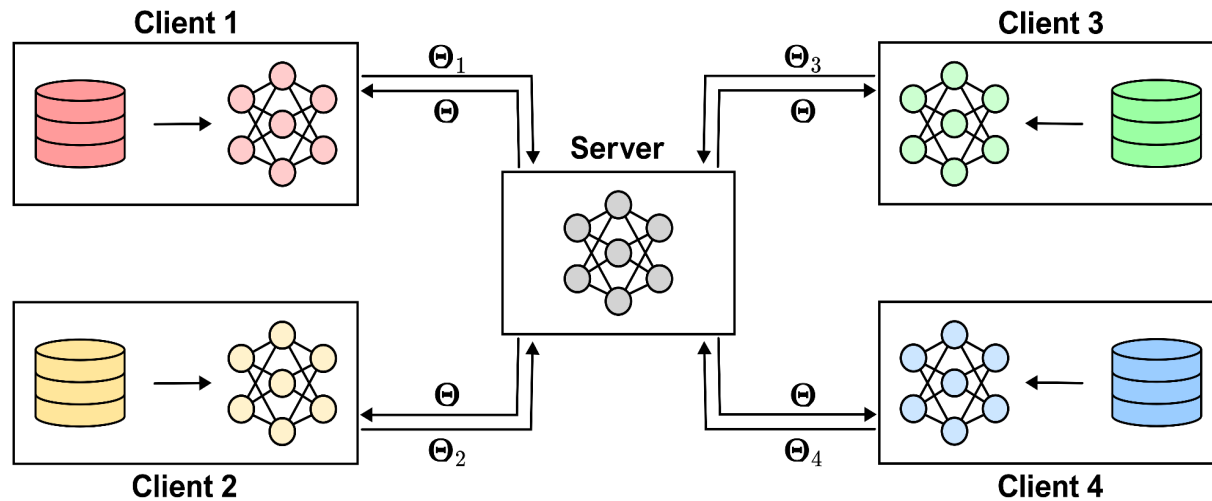


Figure 1: Federated learning scheme involving four institutions (Θ_c – parameters of client c local model, Θ – parameters of the global model).

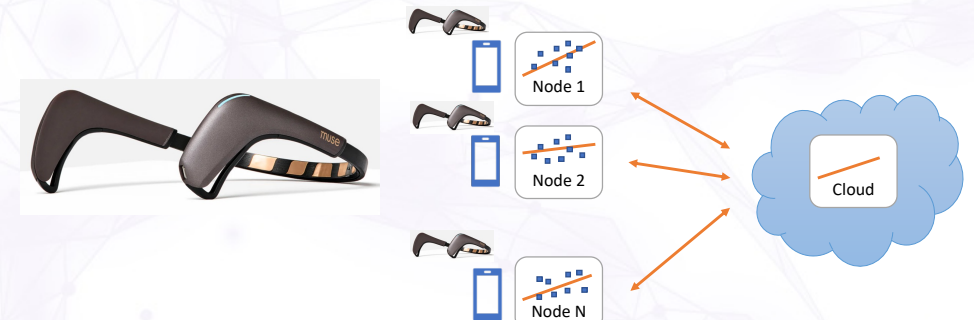
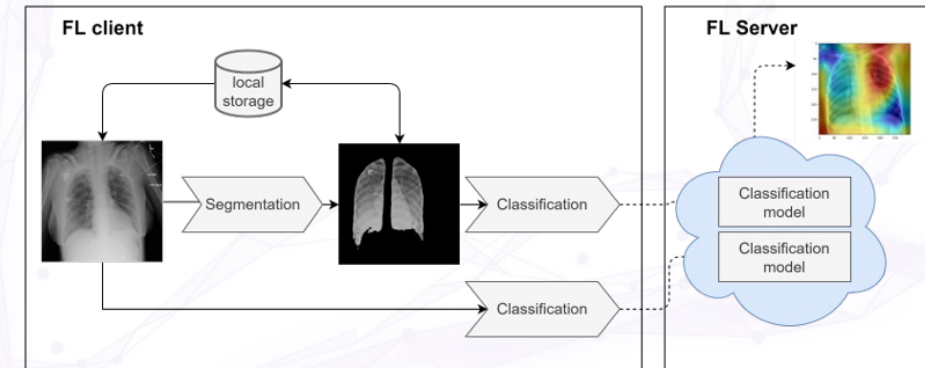
Global aggregation:

$$\Theta^{r+1} = \sum_{c=1}^C w_c \Theta_c^r,$$

C – the number of selected clients, Θ_c^r – the parameters of client's c local model at federated round r , w_c – the aggregation weight.

Examples of Federated Learning projects at Sano

- Large-scale analysis of Chest X-Ray data (600 000 images) from multiple sources
 - Classification, segmentation
- Edge / cloud computing scenario
 - EEG signal from mobile sensors
 - Pilot project on dementia patients
- Transcriptomics data
 - Disease type classification based on genomics variants
 - NearData EU project
- Brain MRI data analysis:
 - Data translation of MRI data (between T1 and T2 weighted images)
 - Estimation of Brain Microstructural Parameters
- Experiments using
 - HPC - Cyfronet: Prometheus, Ares, Athena (GPU)
 - Cloud: Google Cloud Platform (10 GPU Instances)
- Software:
 - Flower.dev – Open Source
 - Pytorch



Data translation in MRI

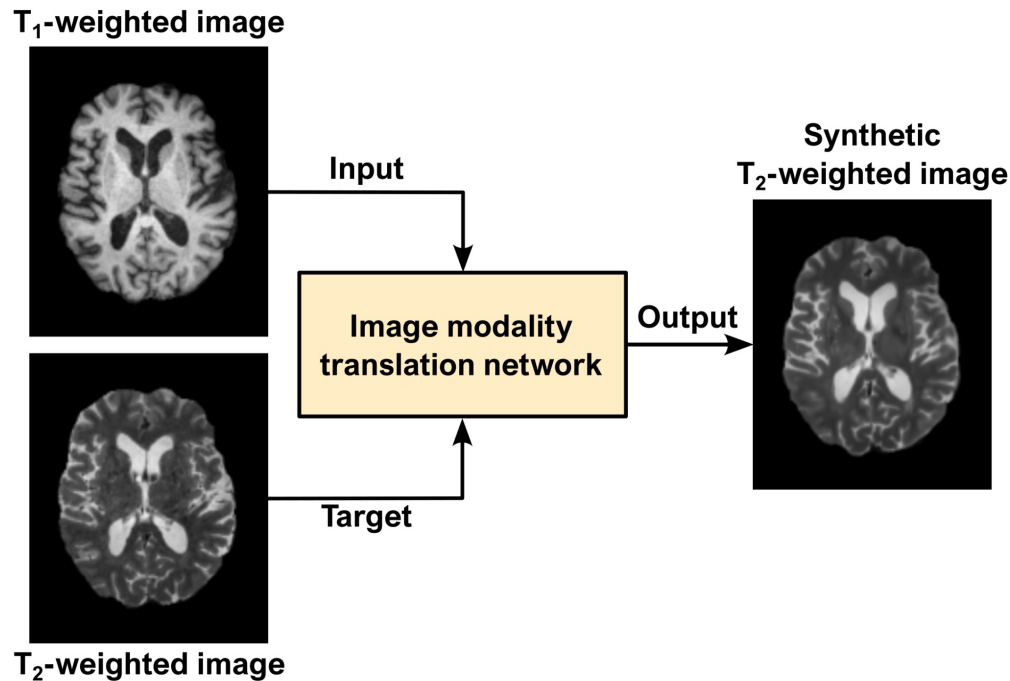


Figure 4: Visual representation of the translation process.

- Data translation is an area of research focused on generating images within and across medical imaging modalities.
- It simplifies clinical workflow by replacing infeasible imaging procedures due to time, labor or expense constraints.

- J. Fiszler, D. Ciupek, M. Malawski, T. Pięciak. *Image-to-image multi-contrast data synthesis from heterogeneous sources using a federated learning concept.*

Visual results

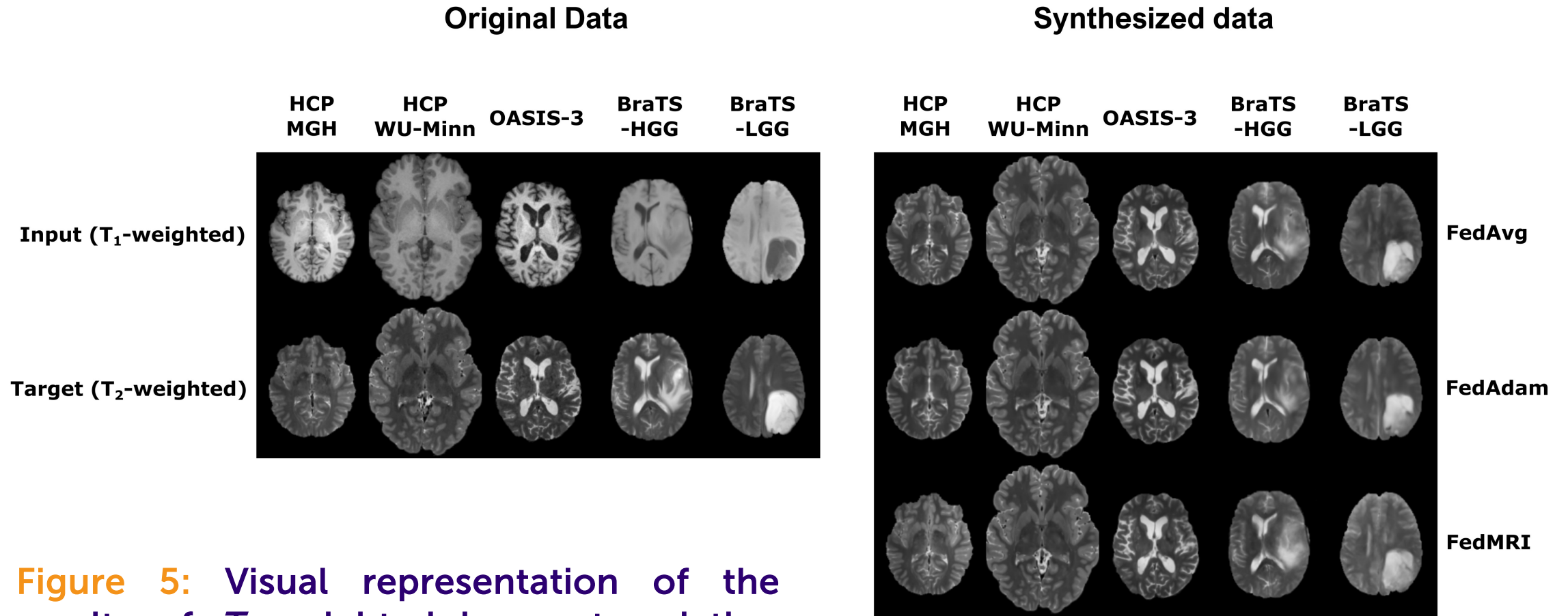


Figure 5: Visual representation of the results of T_2 -weighted image translation after applying selected federated learning algorithms

J. Fiszler, D. Ciupek, M. Malawski, T. Pięciak. *Image-to-image multi-contrast data synthesis from heterogeneous sources using a federated learning concept.*

Example 2

Open Data in In Silico Trials

Exploiting medical data for research: *in silico* clinical trials



- InSilicoWorld is an international collaboration which aims to accelerate the uptake of modelling and simulation technologies.
- The project investigates a broad range of solutions targeting various medical specialities (endocrinology, orthopaedics, infectiology, neurology, oncology, cardiology) and diseases (osteoporosis, dynapenia-sarcopenia, tuberculosis, multiple sclerosis, mammary carcinoma, arterial stenosis, etc.)
- Sano participates in the project through its Extreme Scale Data and Computing team.

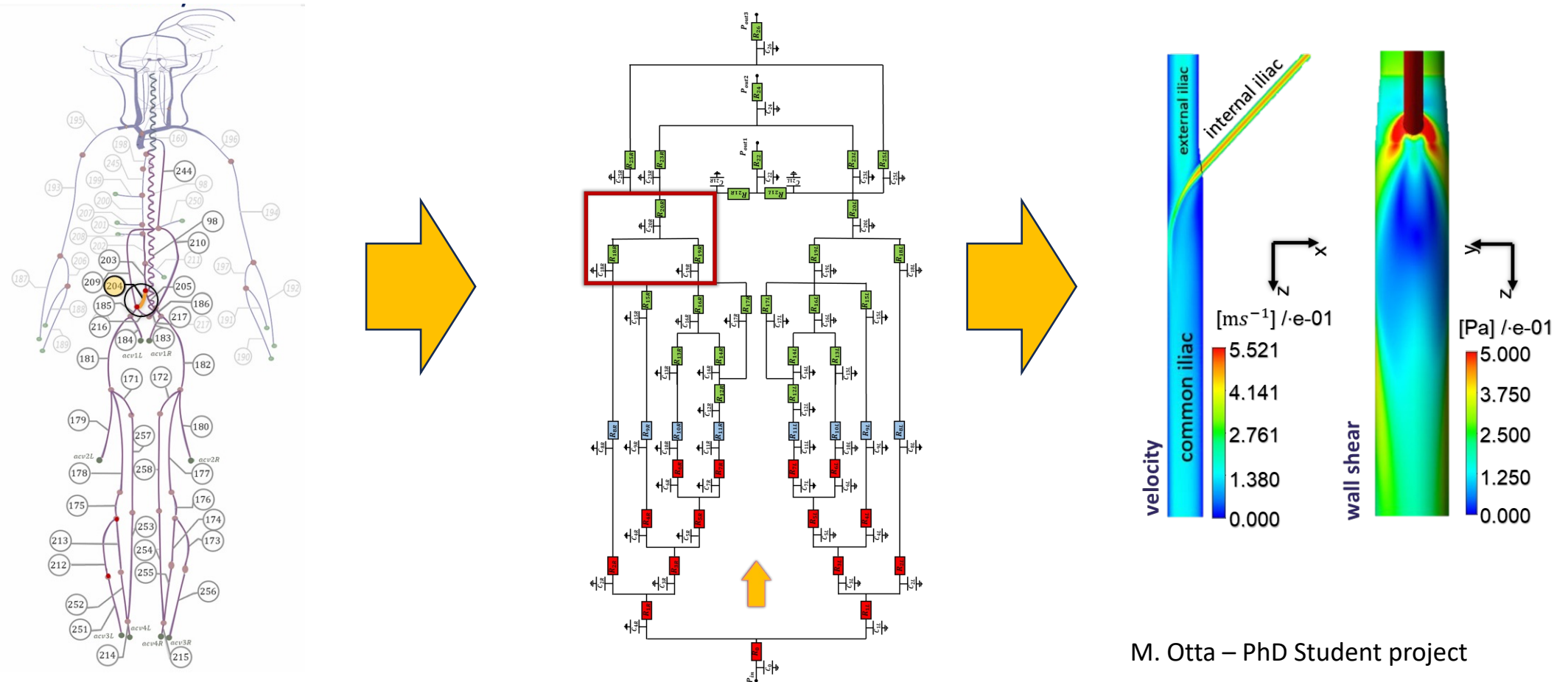
Our goal: to develop an advanced, easy to use simulation environment enabling *repeatability*, *replicability* and *reproducibility* of simulation results; and efficient access and usage of computation and storage resources.

SciProg lead: Piotr Nowakowski



Digital Twin / Virtual Human Twin (VHT)

- Digital, mathematical representation of human body
- VHT is an infrastructure that makes it easier to develop and validate digital twins.



In silico clinical trials – the ISW project



Integrating application models with HPC: the Model Execution Environment



- ✪ **Patient** – virtual space to run different calculations on patient data.
- ✪ **Pipeline** – set of steps which should be executed on patient data.
- ✪ **Pipeline step** – specification of how the model should be run on HPC and what input data is required.
- ✪ **Model** – set of scripts/source code stored in a git repository. Git repository gives us the ability to store development history and run different model versions.

The screenshot displays the Model Execution Environment interface, divided into three main sections: Patient, Pipeline, and Model.

Patient View: Shows a patient case named 'mktest A patient case in your cohort'. The 'Current pipelines' table lists one pipeline:

Id	Name
1	test Demo with number generation (manual pipeline)

The 'mktest patient inputs' section is currently empty.

Pipeline View: Shows the details of the selected pipeline 'test Demo with number generation (manual pipeline)'. A step named 'Generate numbers' is highlighted with a green box. The 'Step' details show:

Start time	15 Jul 10:05
Revision	68f5542cb88904046bd54ca350efa225132
Execution time	00h 00m 35s
Outputs	stdout, stderr
Status	Completed

The 'Saved parameter values' table shows:

Parameter	Value
Numbers count	10
Grant	plgrimage3
Model version	master

Model View: Shows the 'demo-steps' repository. A table of commits is visible:

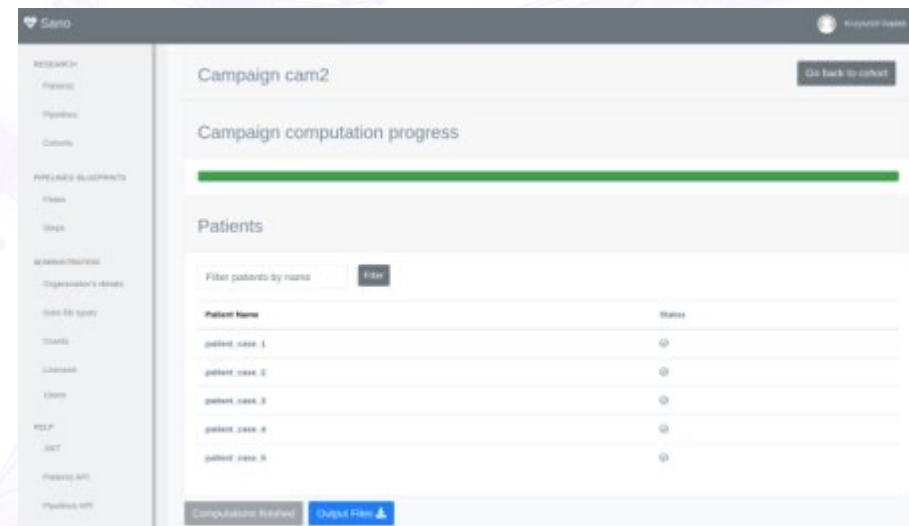
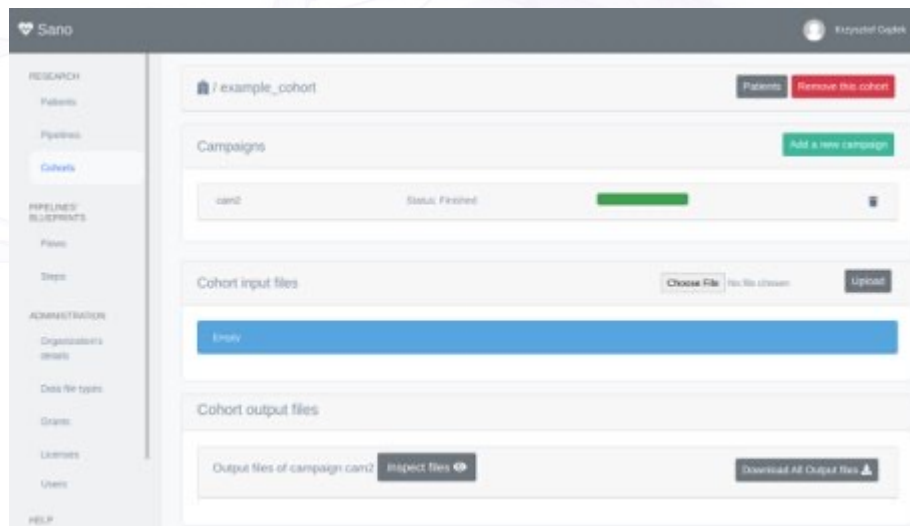
Name	Last commit	Last update
generate_input_job_batch.py...	Changing execution time. It should not L...	8 months ago
generate_input.py	Generate only 5 random numbers	1 year ago
_launch_job_batch.py	Changing execution time. It should not L...	8 months ago
_launch.py	Added parentheses to prints in pytho...	1 year ago
_visualization_batch	Generate git annotation with 5 frames pe...	1 year ago

ISW: *in silico* clinical trials

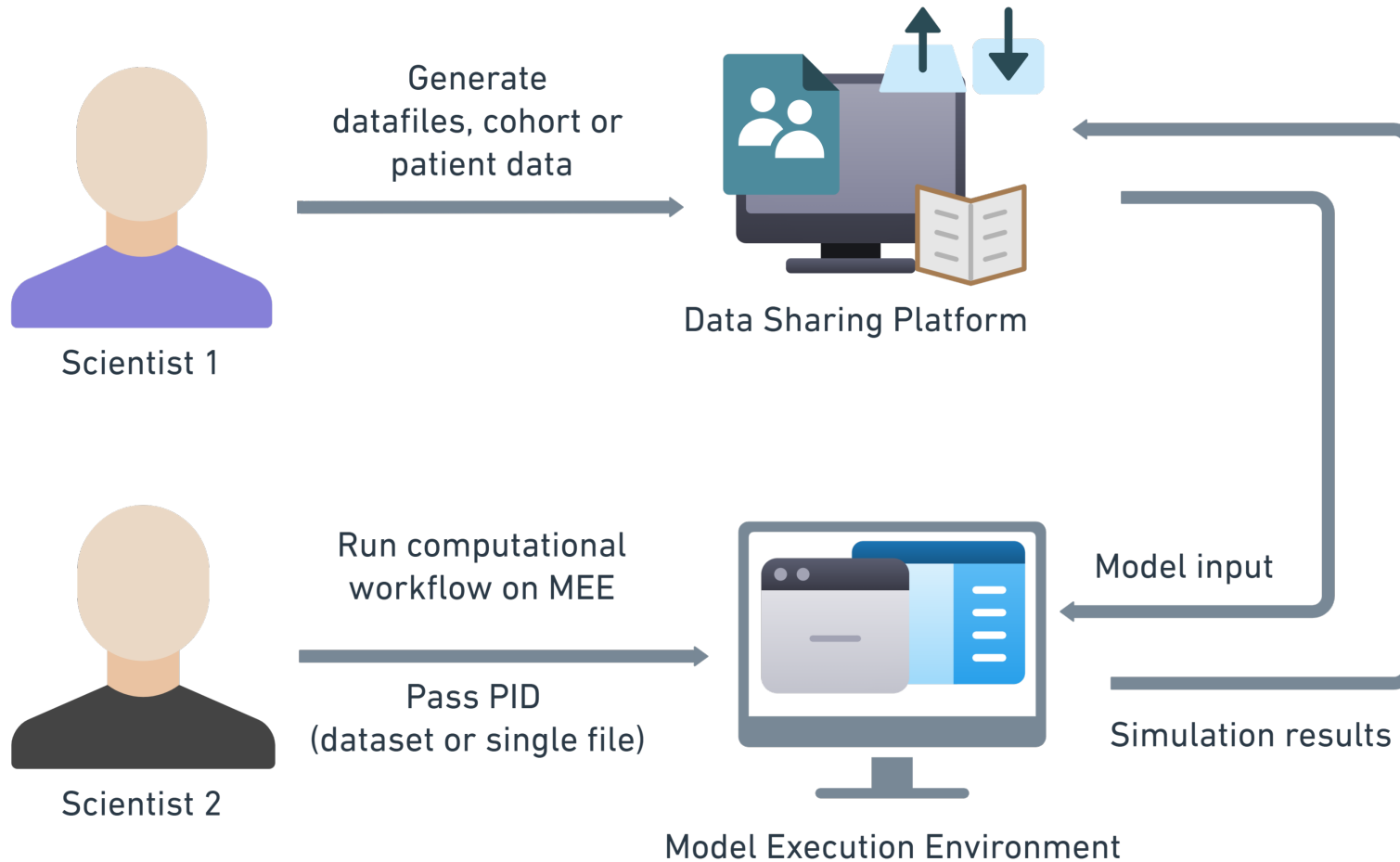


Cohort and campaign management

- The main feature, ensuring efficiency of ISW simulations in MEE are patient cohorts. The user is able to run multiple computational units (**Pipelines**) associated with specific patients with a single click of a button. Such bulk processing is known as a **Campaign**. The screenshots present a sample cohort of 5 patients and results of a simple campaign involving these patients. Every pipeline produces some output files, which are zipped into a single archive and exported for download.



Accessing and publishing medical data in the ISW environment

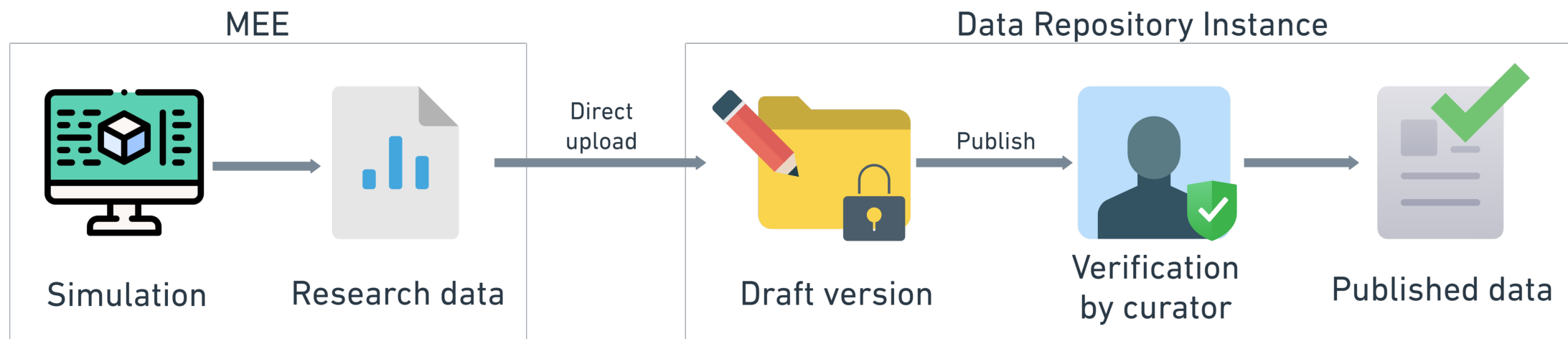


- Data and model levels
- Data as required model input
- Sharing simulation results



Collaboration

Collaboration scenario



- Expose obtained data as draft version of dataset, ready for publication
- Automatically gather interesting results and extend collections stored in the data repository

Integration of data repositories

- Data repository instance can be configured for MEE organization (**admin**)

Data repository configuration

Enable data repository integration

The data repository integration allow your organization to use liquid tags for accessing remote resources from supported repositories. Users will be able to access files and datasets with the API token, if present in user profile. More details can be found in help section: [Data Repository Manual](#)

Data Repository Type

zenodo ✓

dataverse

zenodo

https://sandbox.zenodo.org ✓

Your organization is integrated with data repository:

Data repository type: zenodo, URL: https://sandbox.zenodo.org

Data repository token:

..... ✓

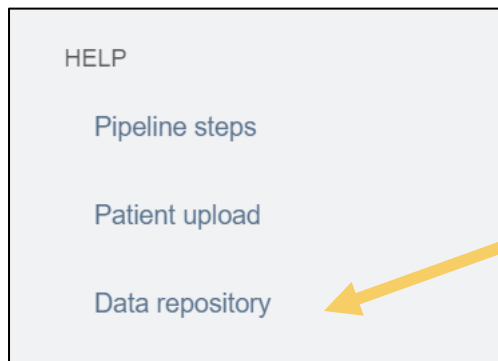
Data repository API token is used to access data on the organization's data repository instance. It is required for the usage of stage in and stage out there. More details in: [Data Repository Manual](#)

[Update token](#)

- Data repository access token can be configured for specific organization membership (**user**)

Accessing and publishing data

- Easy workflow adaptation to external data usage during your simulation
 - Download single file
 - Download dataset
 - Upload file to dataset or record



User guide for advanced tags

```
22 echo Download dataset from dataverse
23 echo -----START-----
24 dataset_doi={% value_of dataset_doi %}
25 file_doi={% value_of file_doi %}
26 {% dataverse_dataset_stage_in $dataset_doi %}
27 echo "Downloaded dataset"
28 {% dataverse_file_stage_in $file_doi %}
29 echo "Downloaded file"
30 echo -----END-----
31
32 find . -type f -name "*.txt" -exec cat {} + > merged.out
33
34
35 echo Uploading results
36 echo -----START-----
37 {% dataverse_file_stage_out merged.out $dataset_doi %}
38
39 {% dataverse_file_stage_out merged.out $dataset_doi {"description":"My
description.", "directoryLabel":"dataverse/subdir1", "categories":["Data", "Dummy
File"], "restrict":"false", "tabIngest":"false", "jshfdbv":"sjhfd"} %}
40
41 echo -----END-----
42 echo Finish
43
```

Usage inside step script

Future work: rule-based data sharing

- Initial research and work on rule-based sharing model
- Potential difficulties in model automation

How to verify the data content to avoid **data abuse**?

Dataset versioning – permanent access?

How to use existing solutions?

The screenshot shows a file management interface. On the left, a sidebar titled 'Files' indicates a 'Restricted' status with the message: 'The record is publicly accessible, but files are restricted to users with access.' Below this, there is a section for 'Request access' with the instruction: 'If you would like to request access to these files, please fill out the form below.' The main content area shows '2 results found' and a 'Sort by' dropdown set to 'Newest'. Two file entries are listed:

- User access:** 'MEE test datafiles for UISS-COVID19', opened 26 days ago by Zaj. It has 'Accept' and 'Decline' buttons.
- Guest access:** 'Test restricted', opened 1 month ago by karol.zajac30@gmail.com. It also has 'Accept' and 'Decline' buttons.



Request



Verification



Acceptance



Collection update

Conclusions: Wide spectrum of data access modes



Type of data

Examples

Solutions

Examples of projects

Public datasets

Large Genomics databases

Big Data / Cloud solutions

Transcriptomics Atlas Pipeline in NearData Project

Private data

Medical imaging

Federated Learning

Applications to MRI Data

Research results

In-silico modeling and simulation data

Open Repositories

Model Execution Environment in InSilicoWorld Project

Maciej Malawski

Director

m.malawski@sano.science.org

Piotr Nowakowski

Head of Scientific Programmers Team

p.powakowski@sano.science.org

Thank you!

<https://sano.science>



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533 and from the International Research Agendas Programme of the Foundation for Polish Science No MAB PLUS/2019/13.



Republic
of Poland



European Union
European Regional
Development Fund



Minister of National Education
Republic of Poland

The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796.

Sano Centre for Computational Medicine, Krakow, Poland
www.sano.science