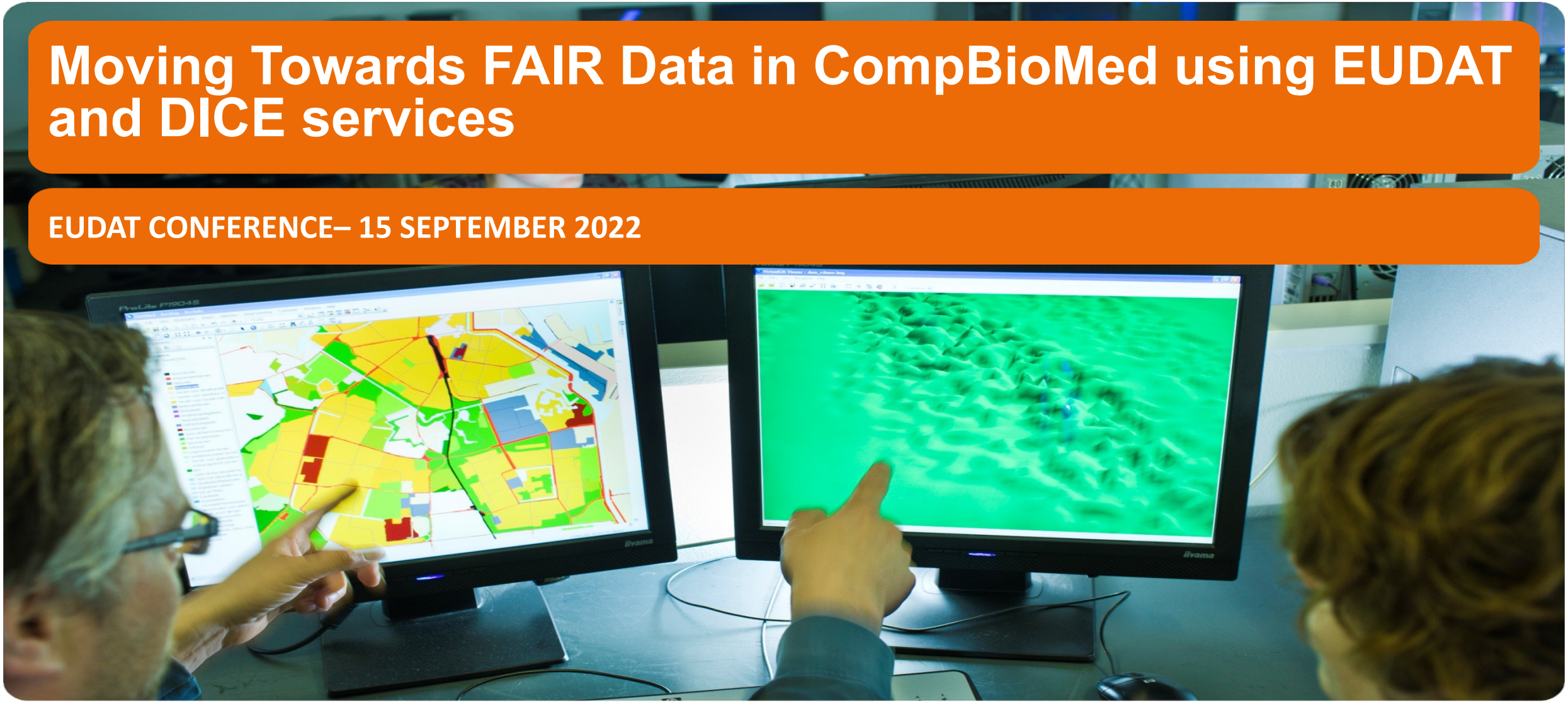
Moving Towards FAIR Data in CompBioMed using EUDAT and DICE services

EUDAT CONFERENCE– 15 SEPTEMBER 2022



Narges Zarrabi (SURF)

Alastair Smith (UCL)

DiCE Data Infrastructure Capacity for EOSC

CompBioMed

SURF is the collaborative organisation for IT
in Dutch education and research

DiCE Data Infrastructure Capacity for EOSC

EOSC-hub

FAIRSFAIR Fostering Fair Data Practices in Europe

EUDAT Collaborative Data Infrastructure

CompBioMed

EOSC Future

Consultancy

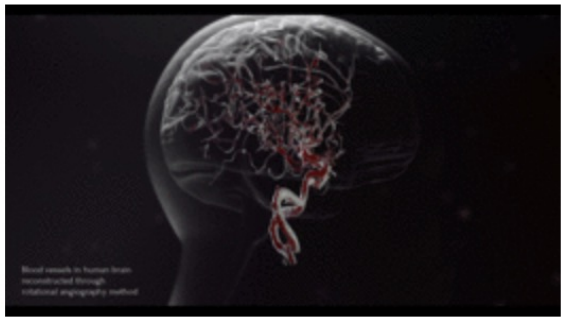
Training

Knowledge Exchange

SURF

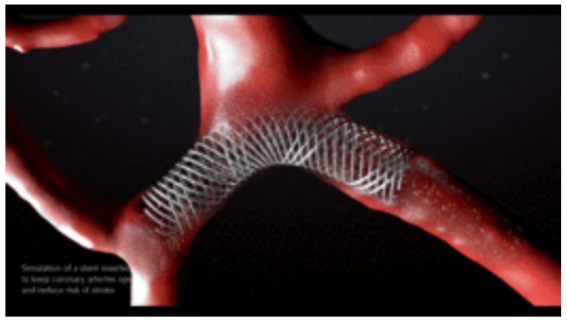
Computational BioMedicine

Academic Users



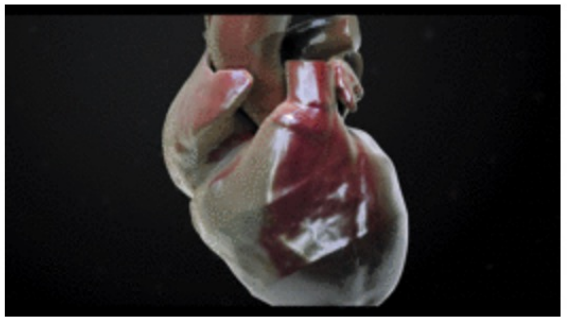
In this section you will find links relevant to Academic Users including user case studies, and information from our Academic Partners.

Industrial Users



In this section you will find links relevant to Industrial Users including user case studies, and information from our Industrial Partners.

Clinical Users



In this section you will find links relevant to Clinical Users including user case studies, and information from our Partners working with medical institutions.

General Public

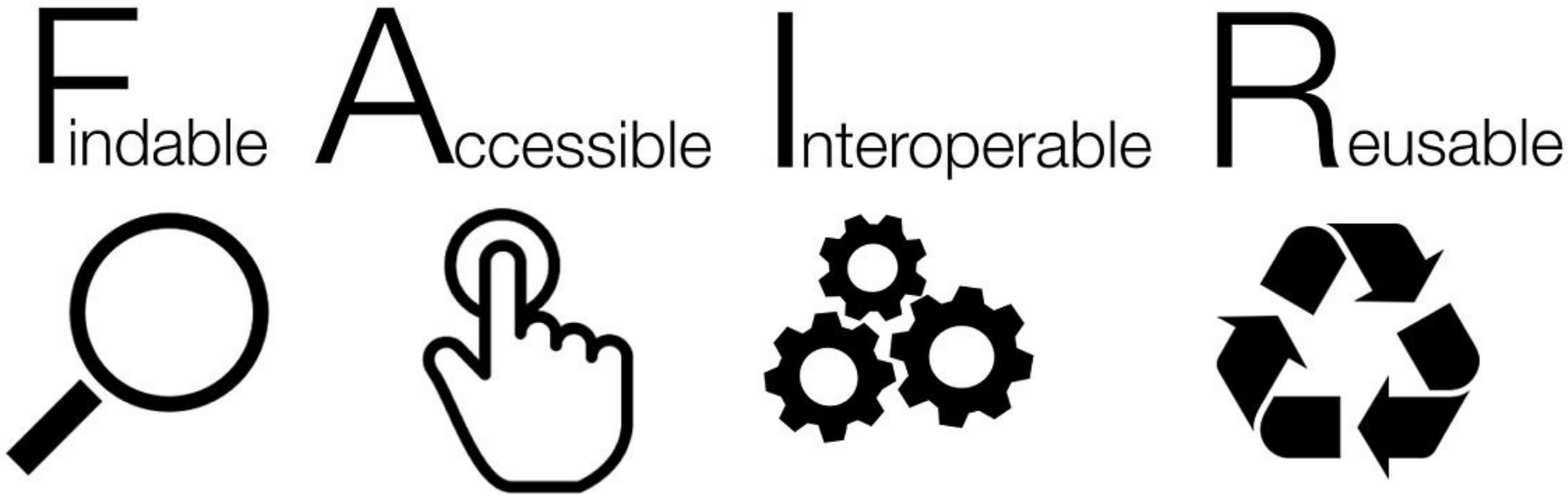
For those from the general public and media who are interested in our project and what we are planning follow this link and the relevant links on the page.

Data Management Challenges of Research Communninties

- **More efficient data access, sharing and transfer**
  - *Intensive data-sharing and transfer*
  - *Restricted data-sharing and transfer*
- **Preserving research data**
  - *Storage, backup and archiving large data, synchronizing data over distributed places*
  - *data provenance*
- **Accessible research Data**
  - *Making data accessible to research communities, PIDs*
  - *Publishing data with domain specific metadata*
  - *Linking published data to processed and raw data*
- **Findable research data**
  - *A major challenges scientific communities is to discover data from research data collections and repositories*

Main Challenge to make data FAIR

- Lack of an encompassing solution for publishing data and/or metadata
- Technical knowledge and awareness for producing FAIR data

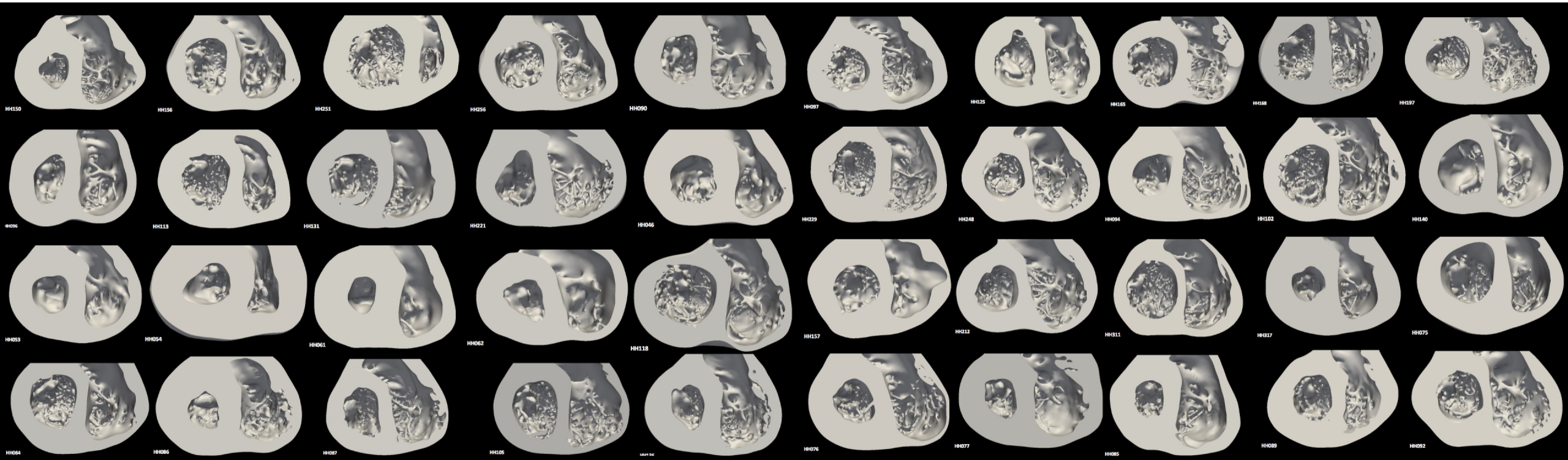


Example research use case with Alya application

In-Silico Human Clinical Trial for Cardiac Safety Assessment of Drugs

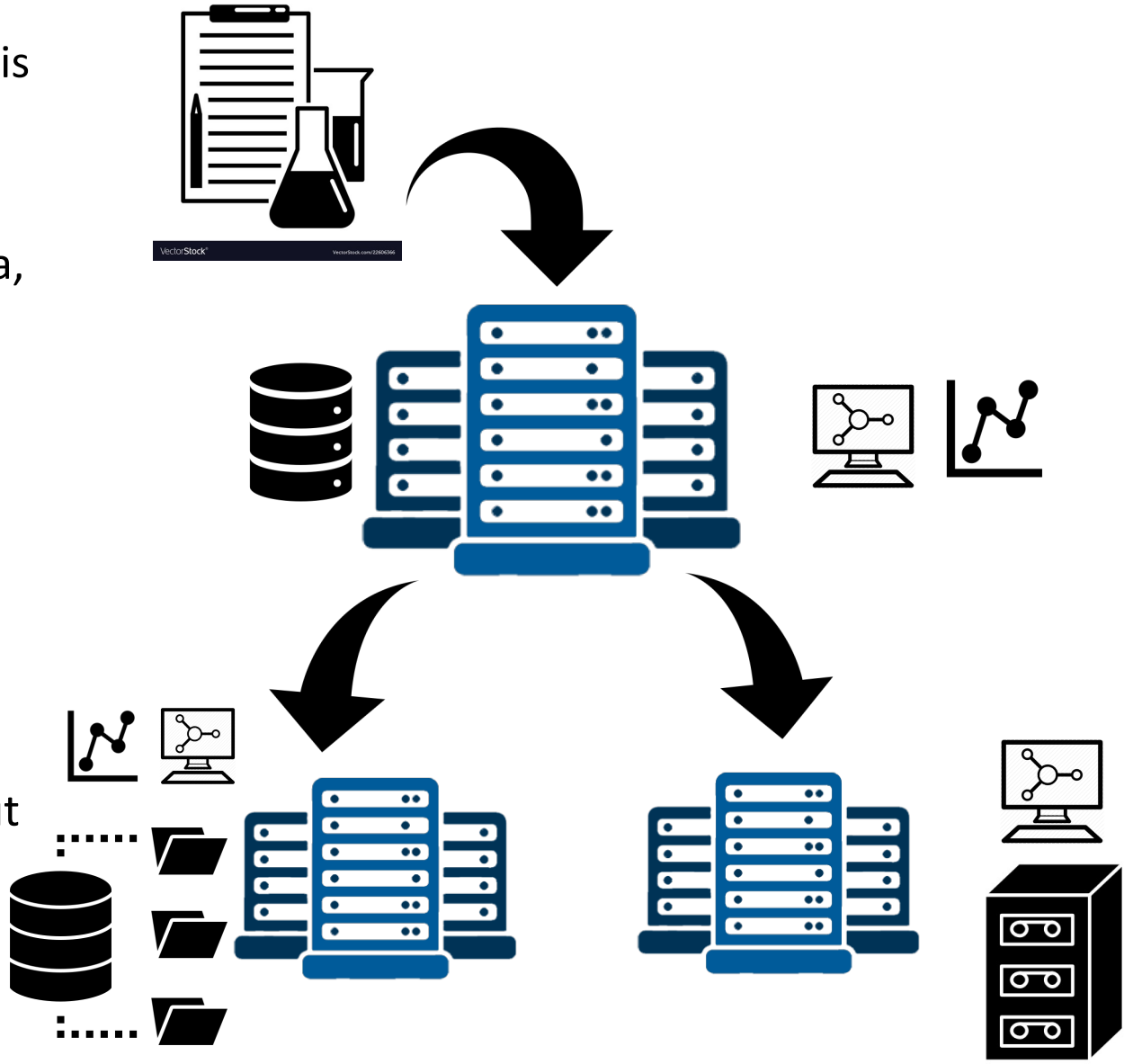
- Why do drugs may produce pro-arrhythmic effects on some people and not others?
- Can we reproduce this observed behaviour to create a normal human in-silico population?

- Alya is a simulation code for high performance computational mechanics
- Solves coupled multiphysics problems

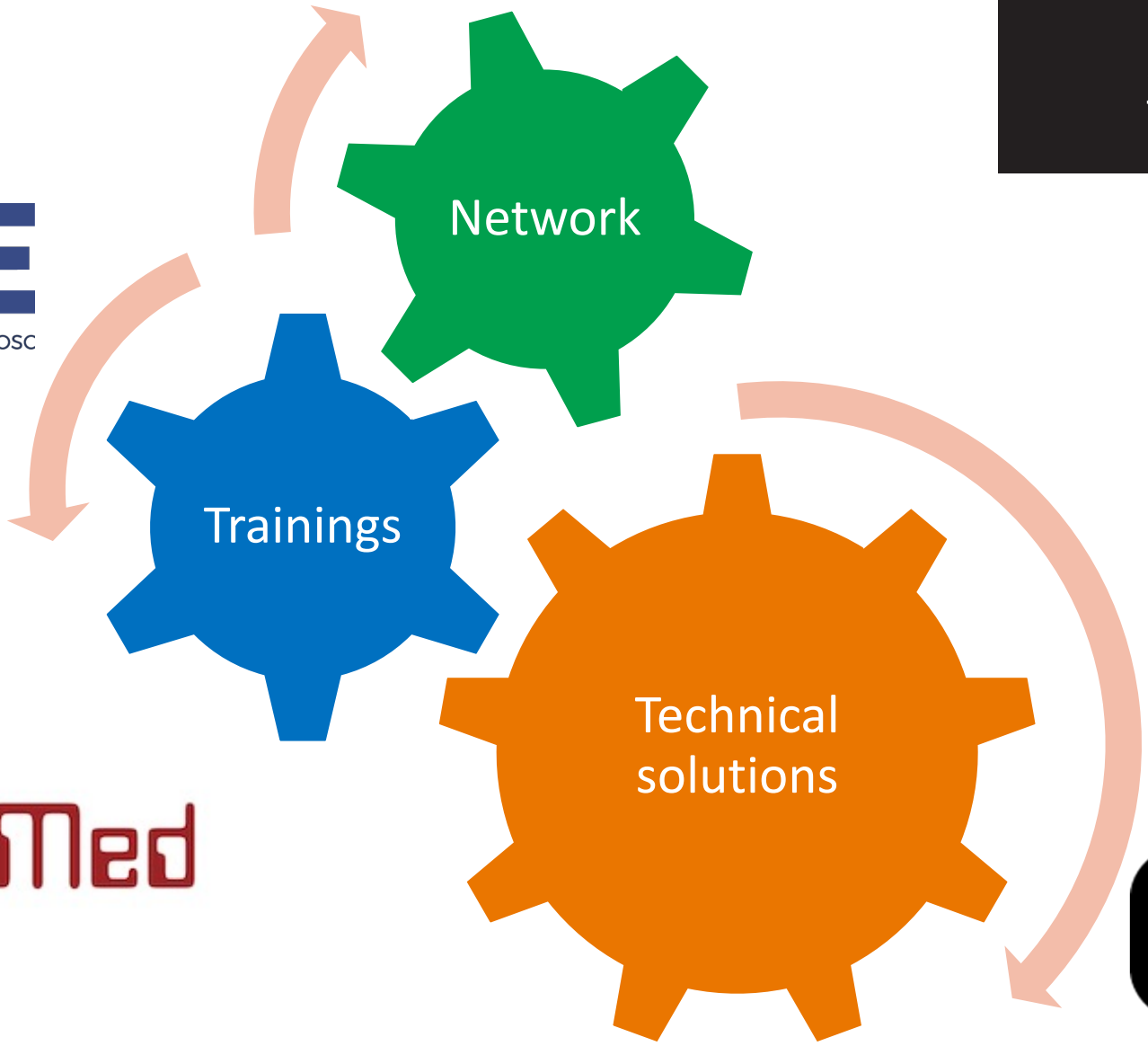


Workflow using Alya Application

- **Step 1: Data creation and transfer:** The raw data is collected at a lab (i.e. ESRF in France), stored locally on tapes. And a copy transferred to BSC.
- **Step 2: Data pre-processing:** Pre-process the data, manual and automated steps for image stitching, segmentation and meshing.
- **Step 3: Data replication:** Replicate data from BSC to other HPC centers such as SURF
- **Step 4: Data Processing and analysis:** run simulations on HPC and analyze output data
- **Step 5: Data publication and preservation:** transferring data to tape archive or publish output data in other repositories



DICE & CompBioMed Collaboration



EUDADT services used in CompBioMed

Service	*Description*	*Resources Needed*	*Provider*
B2SHARE	Data Repository for data publication. Metadata schema can be implemented in this repository. Integration with B2FIND for harvesting data and facilitating findability of the data.	50 TB	UCL
B2HANDLE	Tool required to make persistent identifiers (PIDs) for the data to facilitate findability of the data. The PIDs will potentially be used in B2SAFE and B2SHARE.	1 prefix 10000 PIDs	SURF
B2SAFE	Data staging and safe replication of research data between HPC centers in CompBioMed. The archival storage on tape facilitates long-term preservation of the data.	50 TB 50 TB	SURF BSC

Data Federation in CompBioMed

BSC

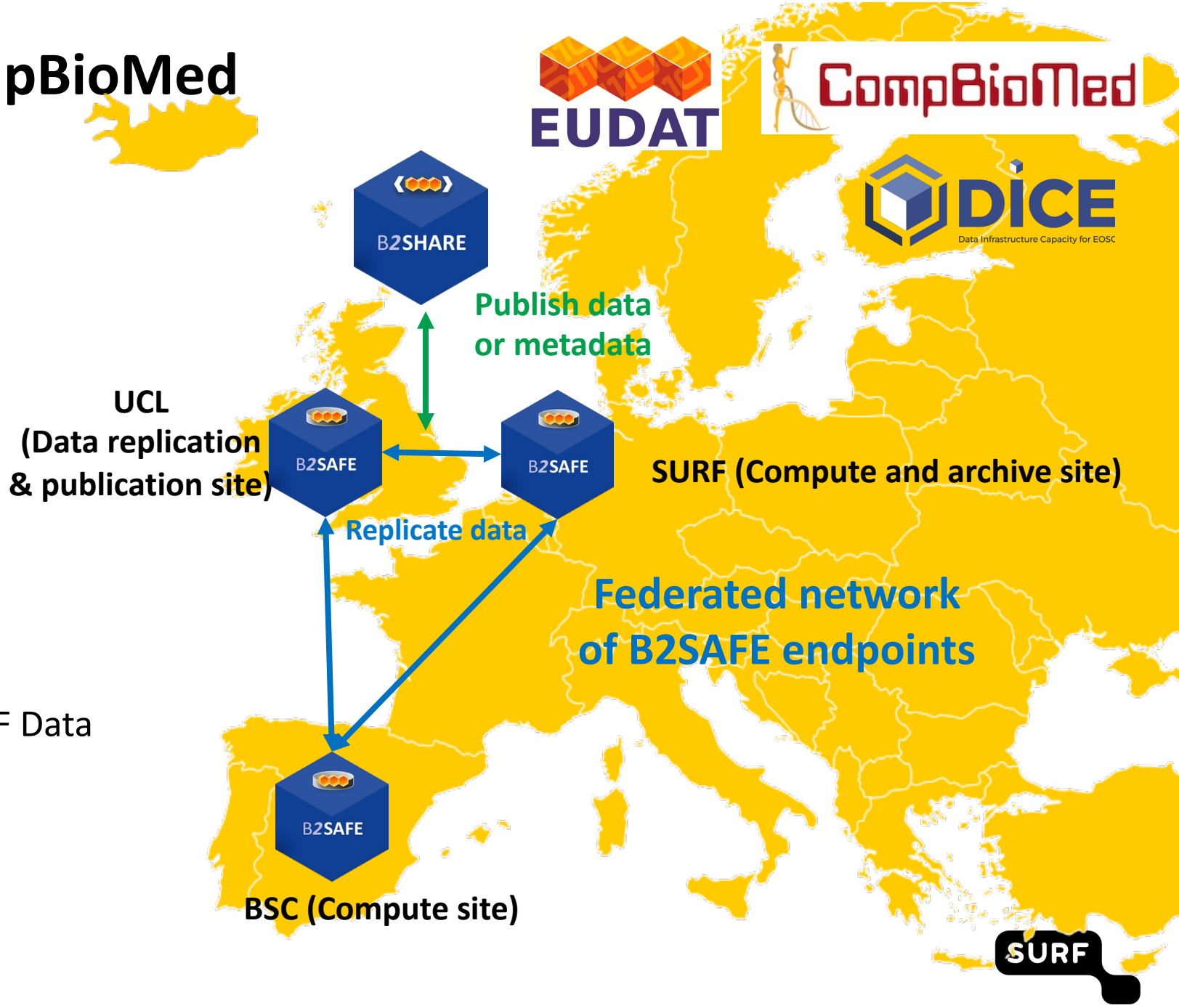
- Compute site
- B2SAFE endpoint

SURF

- Compute site
- B2SAFE endpoint
- Data Archiving site
- Possibility to publish data in SURF Data repository

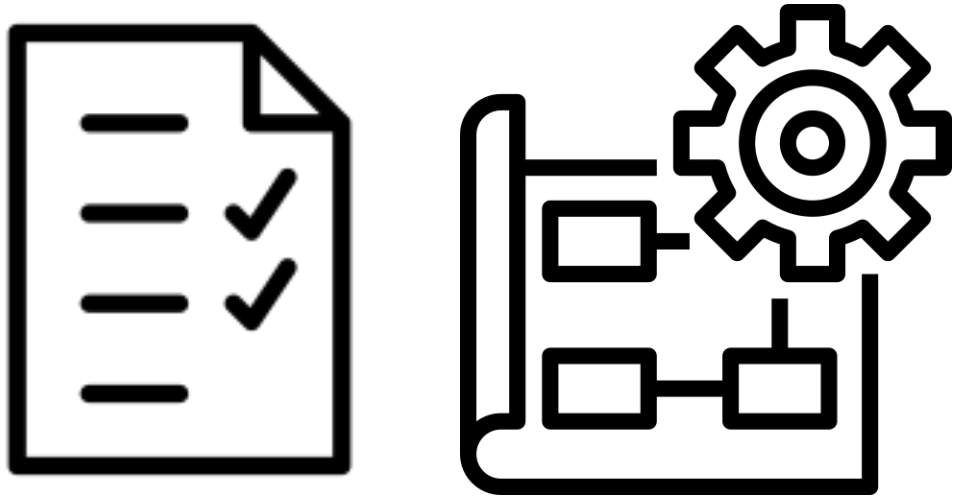
UCL

- Data publication site
- B2SAFE endpoint



Workplan and technical task descriptios

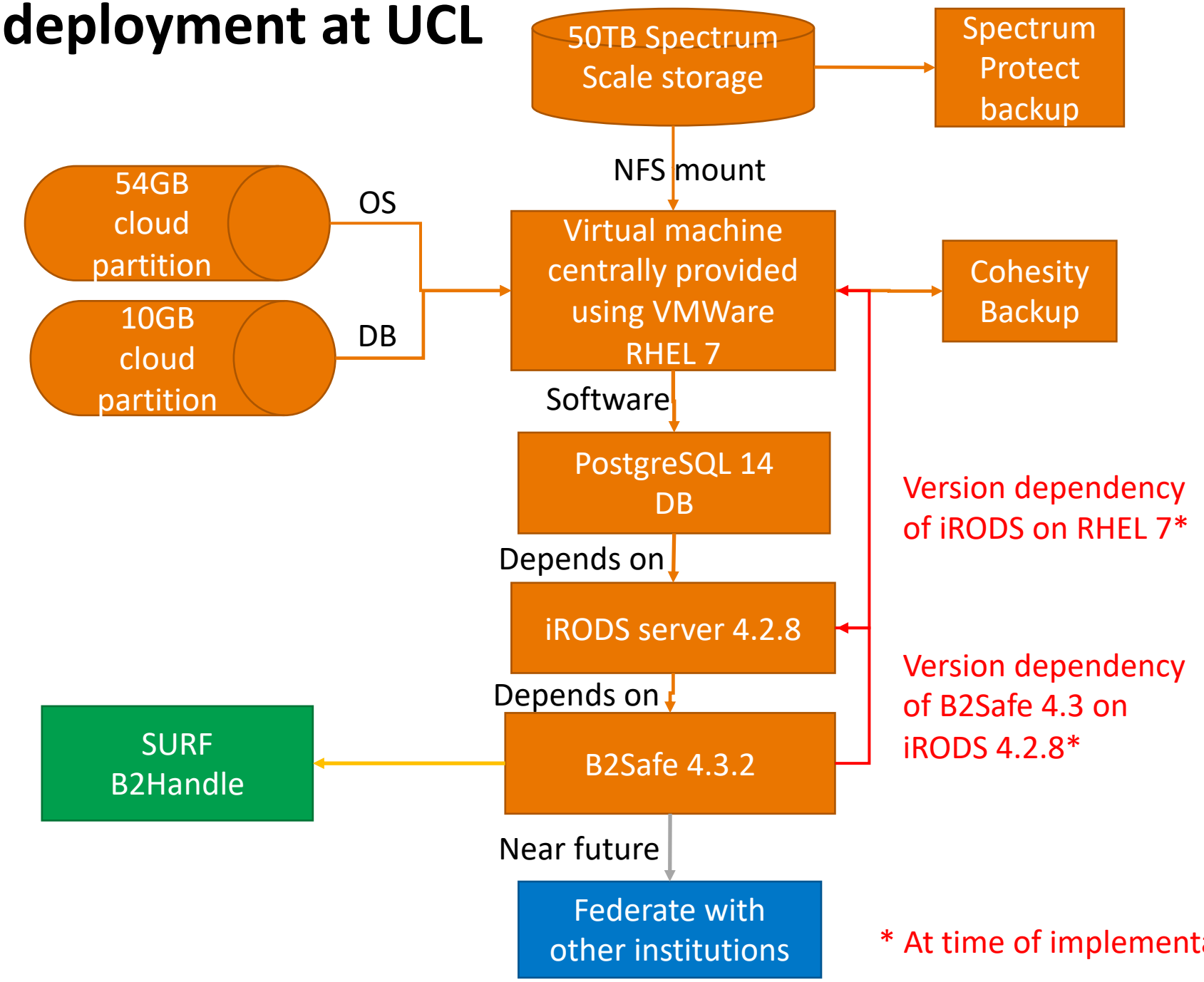
- Technical workplan
- In process of deploying, configuring and federating services
- We get the technical support through the DICE collaboration



Detailed technical tasks

- BSC (Compute site)
  - Deployment of B2SAFE and B2Handle
  - Federation with SURF B2SAFE endpoints
  - Allocation of storage in B2SAFE and data uploaded
  - Performance tests and actual replication
- SURF (Compute and archive site)
  - Providing B2Handle service for BSC and UCL
  - Federation with BSC B2SAFE endpoint
  - Federation with UCL B2SAFE endpoint
  - Allocation of storage in B2SAFE and tape storage
    - Monitor integration of B2SAFE-B2SHARE
- UCL (Data publication site)
  - Deployment of B2SAFE and B2Handle for making PIDs
  - Federation with SURF B2SAFE endpoint
    - Deployment of B2SHARE data repository
    - Integration B2SHARE-B2FIND

B2SAFE deployment at UCL

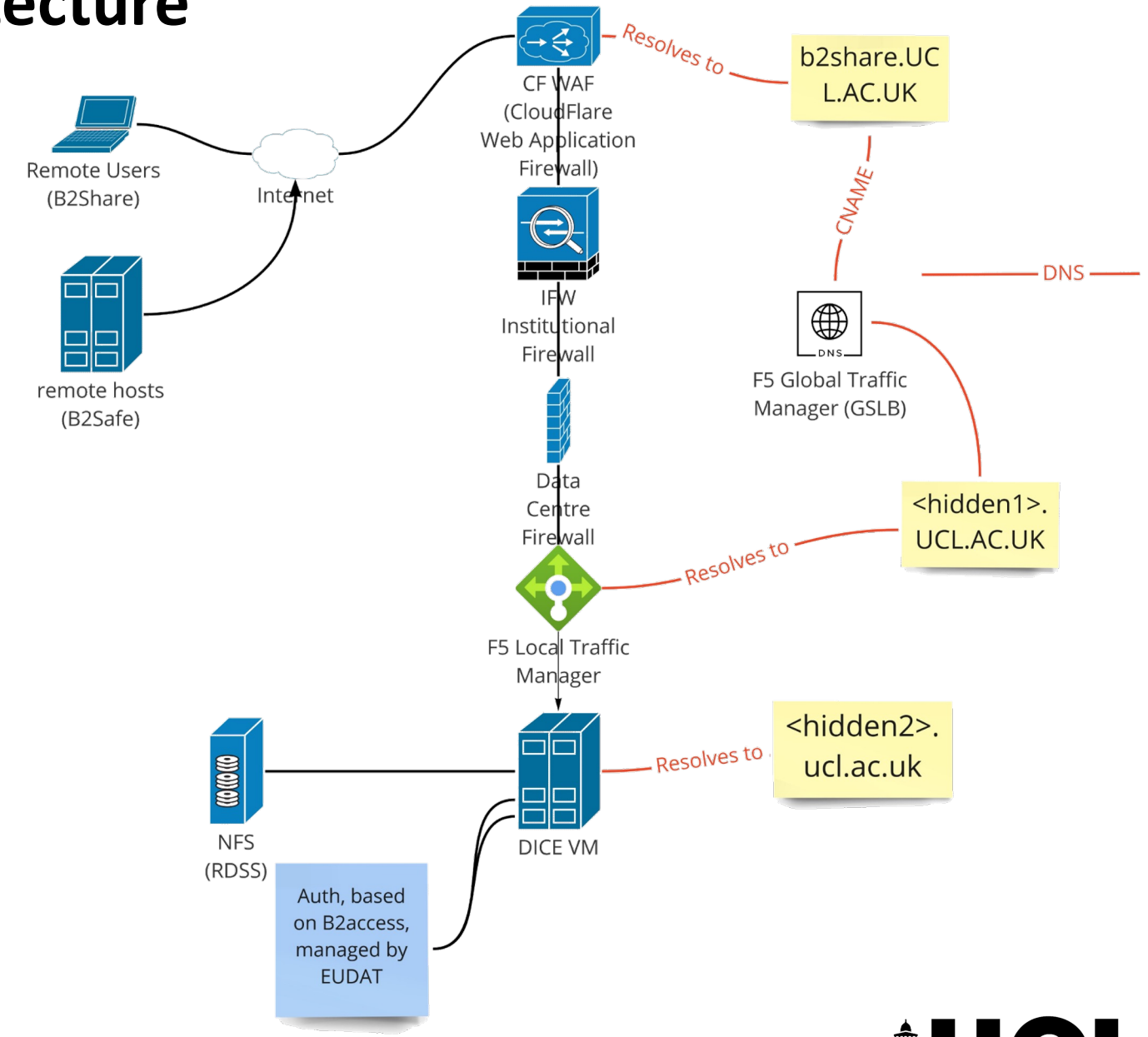


* At time of implementation

UCL

UCL Planned Network Architecture

- Fulfilled UCL Cyber security minimum requirements related to vulnerability scans on the VM
- A Virtual IP has been created for this project
- An F5 (a load balancer that is also used to configure Virtual IPs), has been configured to allow the required ports through.
- DNS name assigned to Virtual IP
- Firewall on the VM itself have been opened up
- To Do: Open institutional firewall for SURF for the federation to be made



Difficulties and Improvement points

- Usage and installation information spread over different sites.
- Documentation is incomplete
- Yum repo for B2Safe has broken dependency
- Yum repo for B2Safe appears to be down for several weeks.
- Limited support
- Local UCL policies on using latest enterprise OS for security and maintenance purposes (currently RHEL 8).
- iRODS is not compatible with latest OSs (roughly three years lag).
- B2Safe lags behind latest iRODS

UCL

Technical knowledge and awareness

- DICE Roadshow webinar
- Training on Data Management and Publication

Thank you!

UCL

- James A J Wilson
- Alastair Smith
- Peter Coveney
- Emily Lumley

BSC

- Nadia Tonello
- Simon Carroll
- Jazmin Aguado

Cineca

- Debora Testi
- Michele Carpene

SURF

- Marco Verdicchio
- Robert Verkerk
- Sara Ramezani
- Mark van de Sanden
- Claudia Behnke
- Claudio Cacciari