Infrastrutture cloud sicure e federate - l'esperienza di INFN

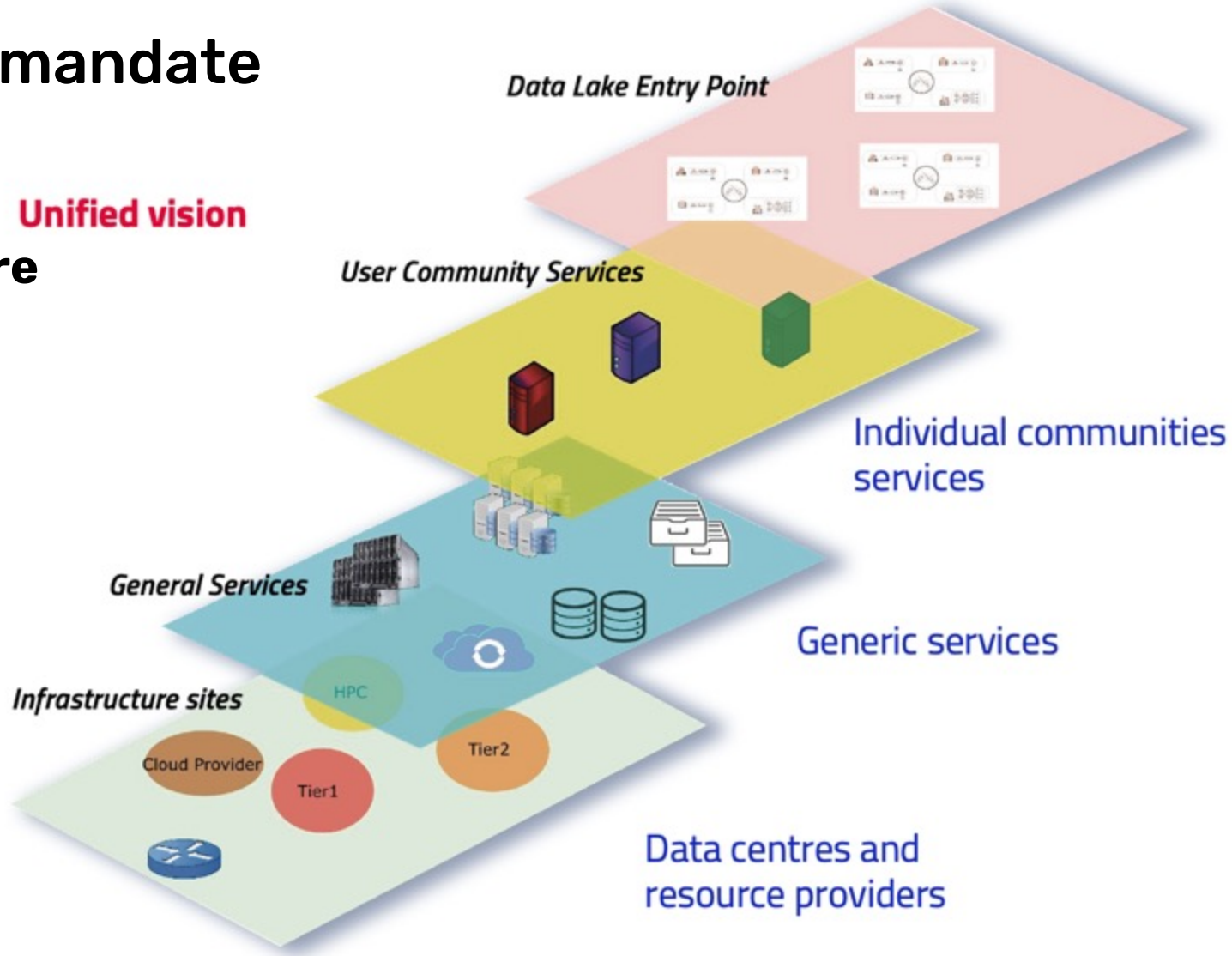
Barbara Martelli, Giacinto Donvito

On Behalf of INFN DataCloud group

INFN

The INFN DataCloud Project – mandate

- The DataCloud Project manages all **core activities related to computing @ INFN and its projects**
  - Development, implementation & management of the INFN Datalake architecture.
  - Development of ISO-Certified solutions mainly for clinical and omics data management
  - Support to users and to the management and operation of all INFN sites (both Grid and Cloud paradigms).
  - Development of new services.
- Focus on **Integration** of resources, methods, people, solutions.
- Modular architecture based on **service composition**.



Proposed architecture for ICSC and TeRABIT projects

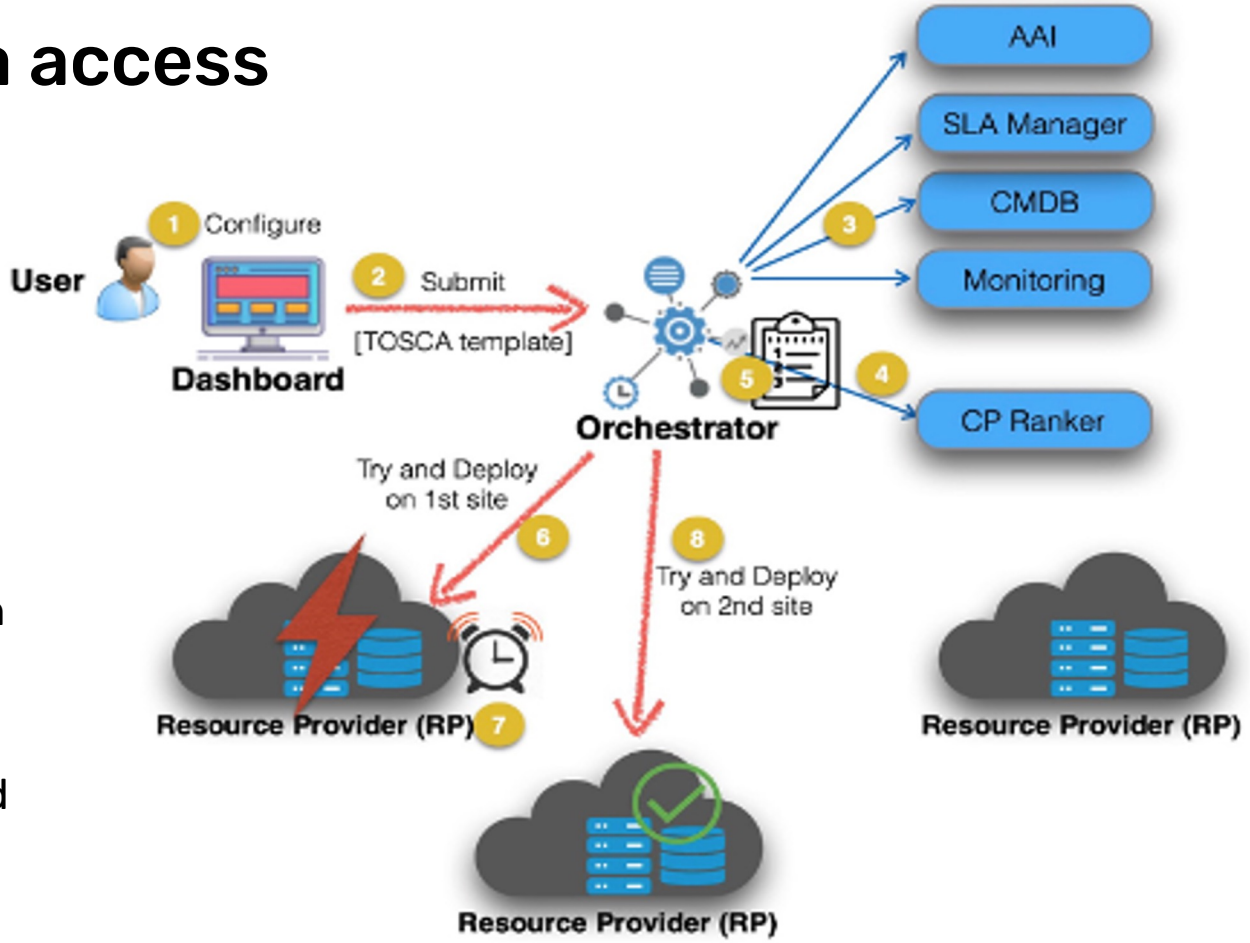
Federation of computing and data access

The INDIGO PaaS Orchestrator

enables the federation of distributed and heterogeneous compute environments, including Clouds, Docker orchestration platforms (such as Kubernetes), HPC systems

Key features:

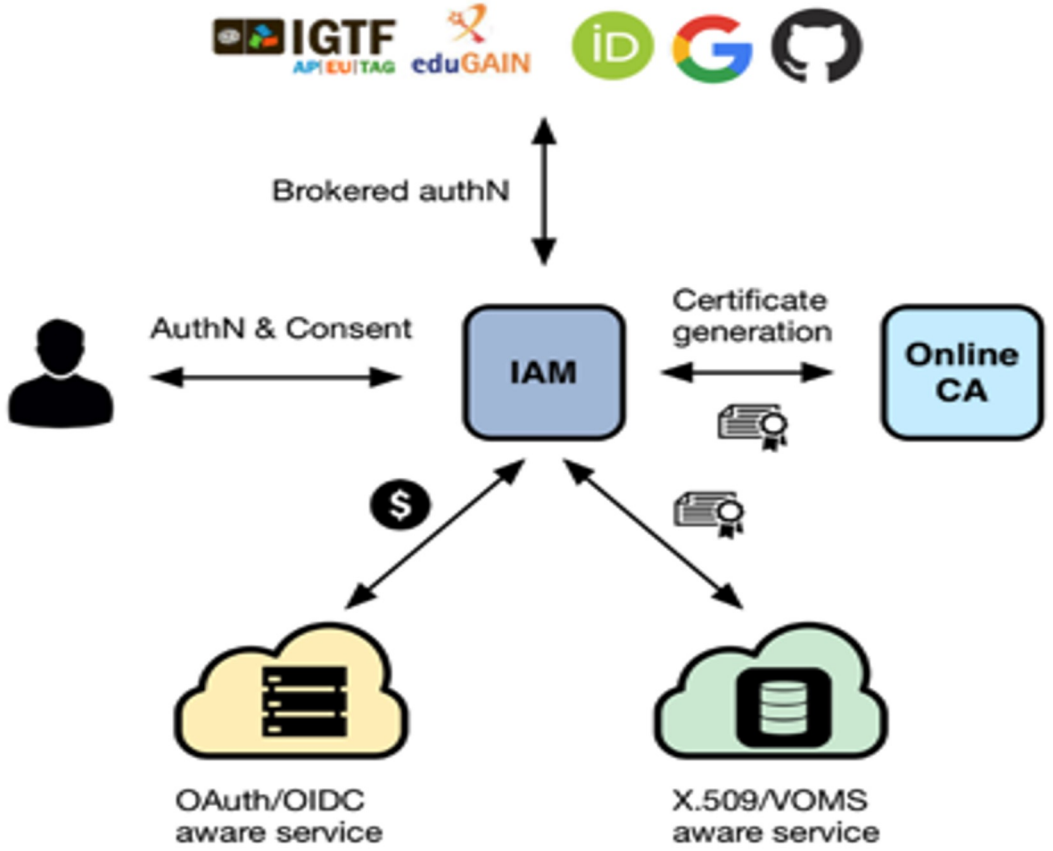
- Smart scheduling, supporting the automated selection of the best provider, based on compute or storage requirements vs. provider capabilities. This takes into account criteria such as resource quotas, SLAs, monitoring data, support for specialized hardware, and data location.
- Support for hybrid deployments and network orchestration.
- Client interfaces for advanced users in the form of REST APIs, Command-Line Interfaces (CLI), as well as for regular users with a simple to use web interface



Proposed architecture for ICSC and TeRABIT projects

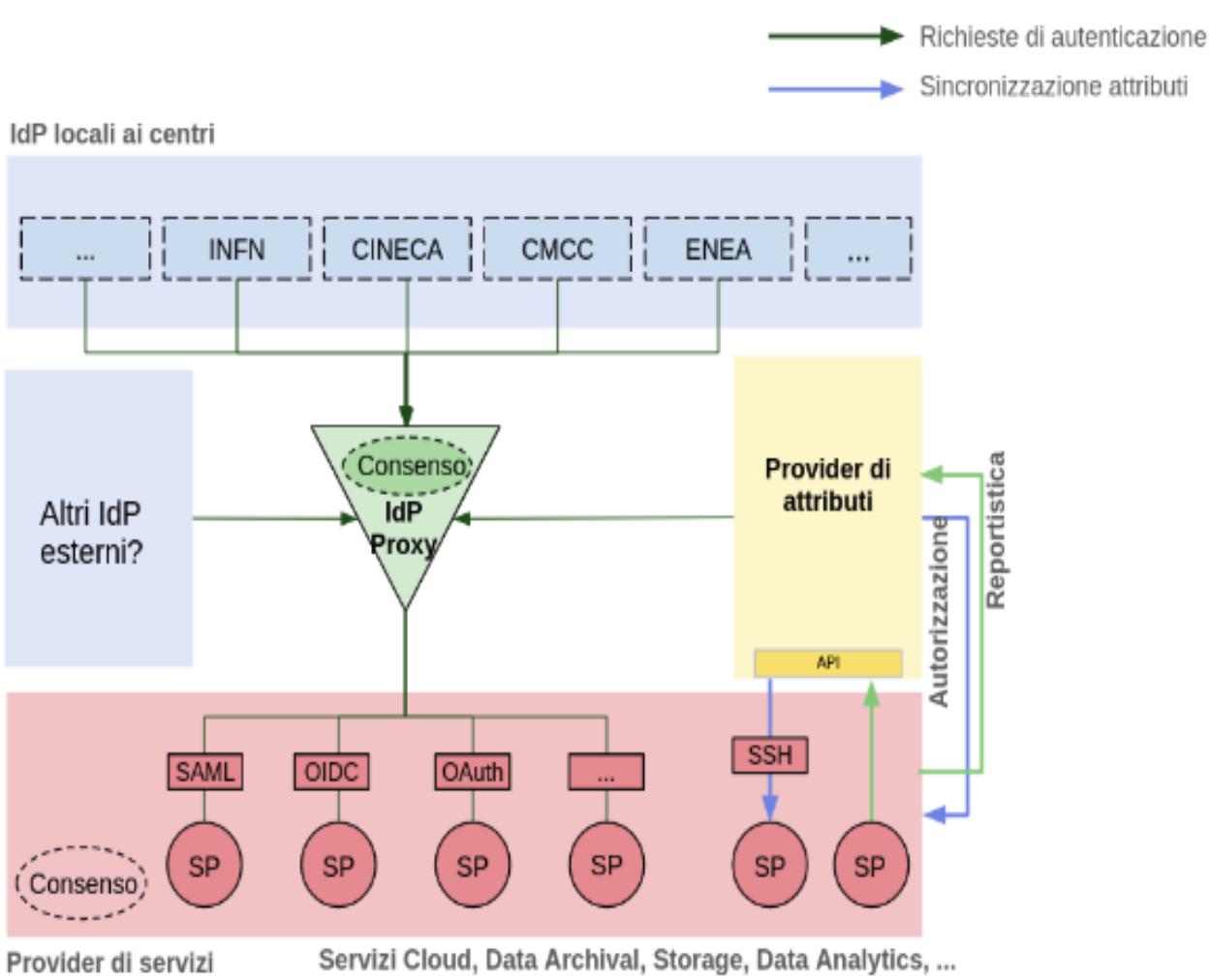
Federated Authentication

- Authentication and Authorization managed through the opensource, standard based INDIGO-IAM
- OpenID Connect and OAuth protocols through INDIGO-IAM are used in all DataCloud infrastructure
- Entity wanting to federate deploy
  - IdP (Identity Provider)
  - SP (Service Provider)



Federated Authentication next steps

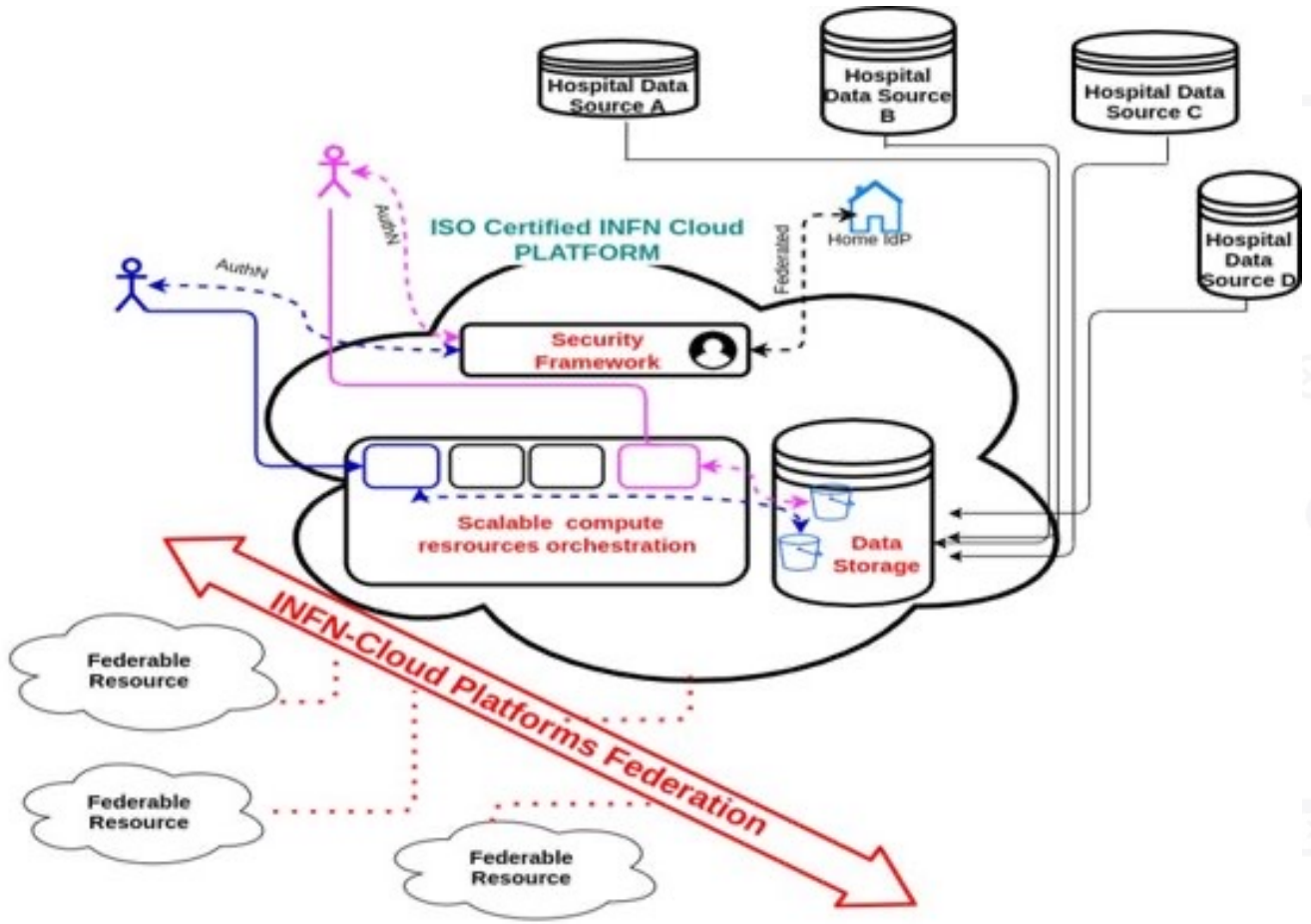
- Work plans on contributing to the OpenID Connect Federation
- Include different entities in the same IdP realizing a central IdP Proxy, to which entities can connect
- IdP Proxy redirects authentication requests, identifies and authenticates users (possibly including users without a proper affiliation), validates user profiles and handles a general access policy.
- Connected to the IdP Proxy, there will be an Attribute Provider, whose scope will be to manage e.g. user projects and groups, as well as handle policies that are specific to the resource providers connected to the IdP Proxy.



Proposed architecture for ICSC and TeRABIT projects

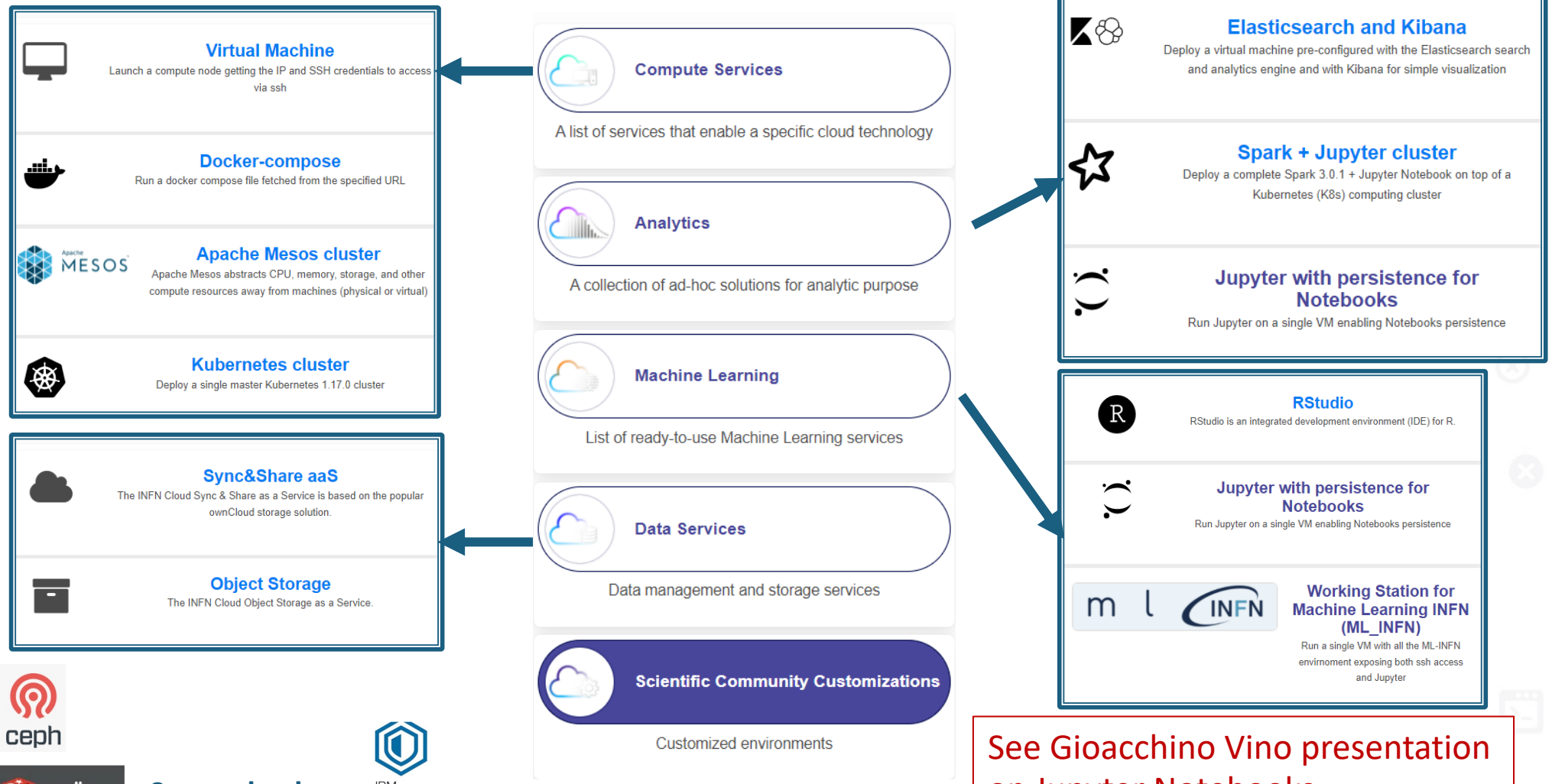
The ISO-certified partition of INFN DataCloud: EPIC Cloud

- The Enhanced PrIvacy and Compliance (EPIC) Cloud is a region of INFN DataCloud certified ISO/IEC 27001 27017 27018
- Currently located at CNAF (Bologna) is being extended to INFN Bari and INFN Catania sites
- **EPIC Cloud**: a reference Cloud implementation for the treatment of sensitive data at INFN
- Some technical and organizational measures adopted:
  - Multifactor authentication
  - Multiple backups
  - Fine grained audit logs
  - Segregation of duties
  - IP whitelists



EPIC Cloud extension

- The extension of the ISO-certified region will be completed by June ‘24
- The extension will bring the full capabilities of DataCloud to life science communities



Secure storage: MINIO, ceph

In-memory data store: redis

Secure backup: IBM Spectrum Protect

Secret management: HashiCorp

PaaS Orchestrator: Indigo IM

Selectable storage QoS levels: fast (SSD), normal (HDD), archive (tape-backed), remote replicas

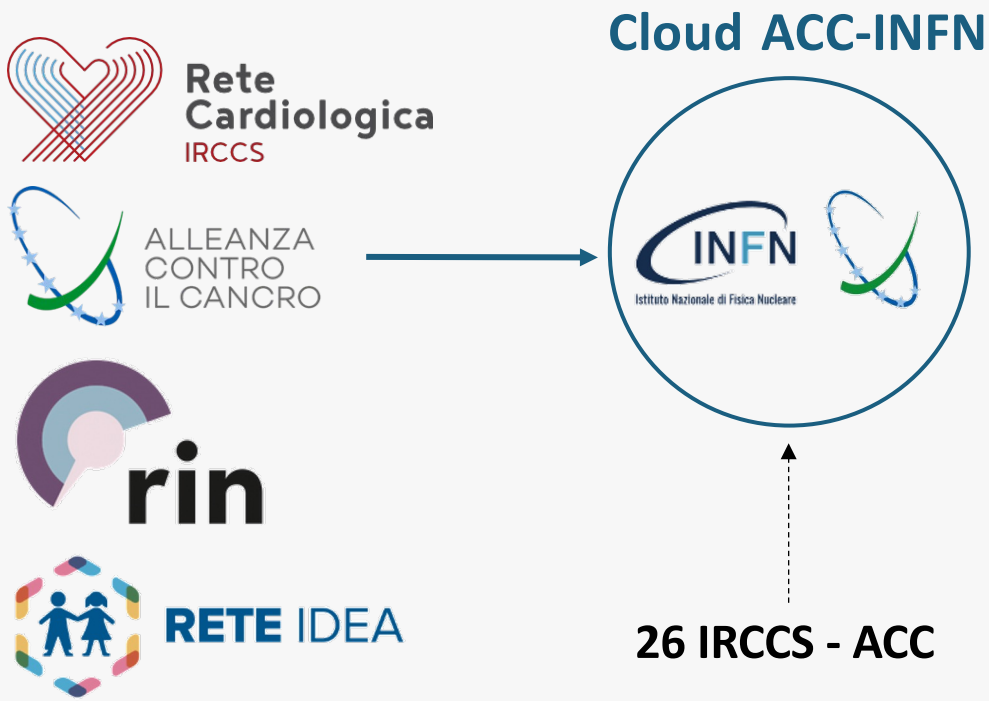
See Gioacchino Vino presentation on Jupyter Notebooks

An example of EPIC Cloud exploitation: the Health Big Data project

HealthBigData
The Italian Hub.

IRCCS-networks Platform | Federated Platform

The nucleus of the HBD federation is the ACC – INFN DataCloud platform



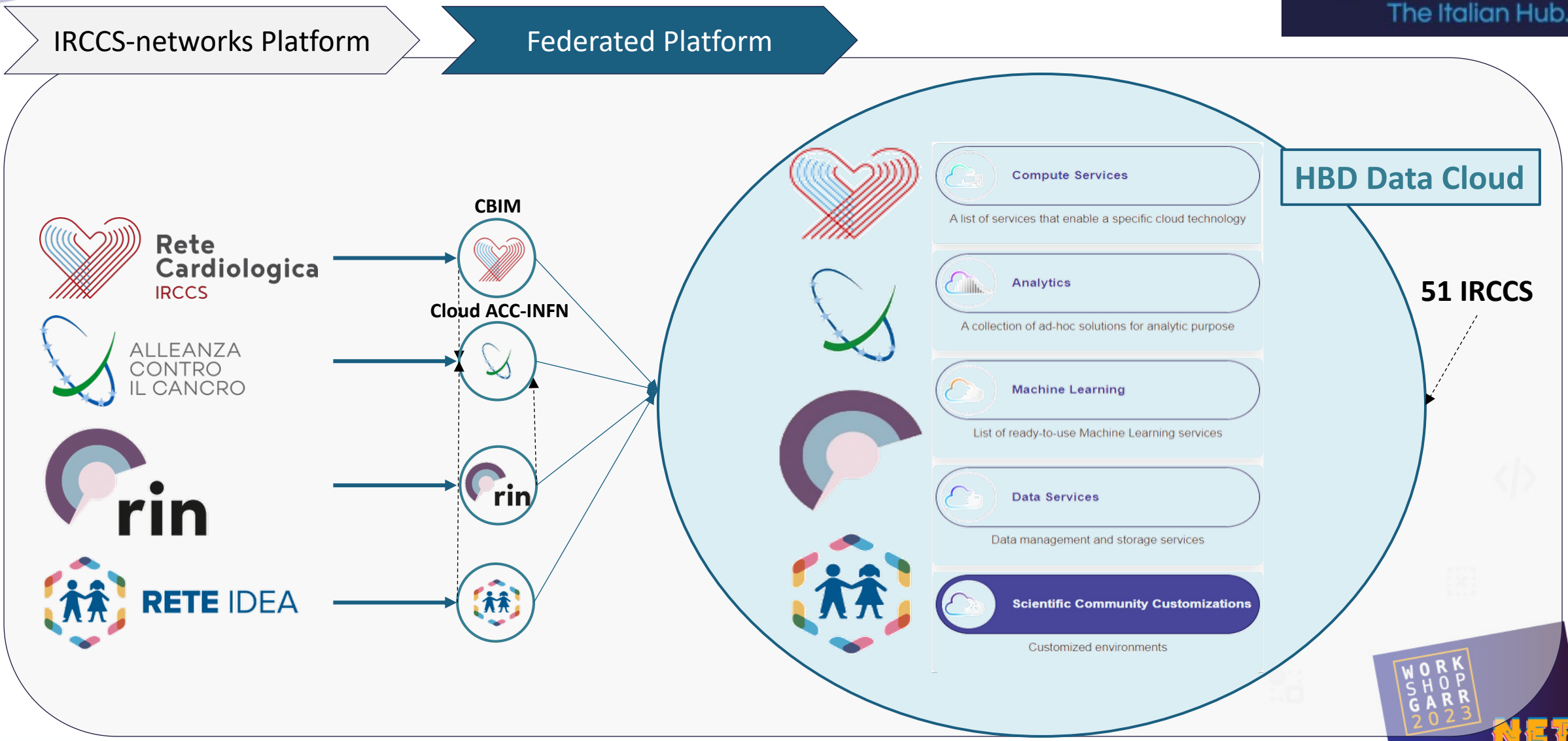
> 80 researchers
> 2800 patients for 5 projects
5,5 TB of "live" data (plus backups and remote archive copies).

Patient, image, sample and sequencing information is registered to the ACC LIMS platform.
Images and sequences data is uploaded to the platform. Data collected are:

- genomics: germline and tumor samples, DNA and RNA, in BAM or VCF format;
- radiomics: radiomic features belonging to different families: morphological, textural, statistical, in DICOM format.

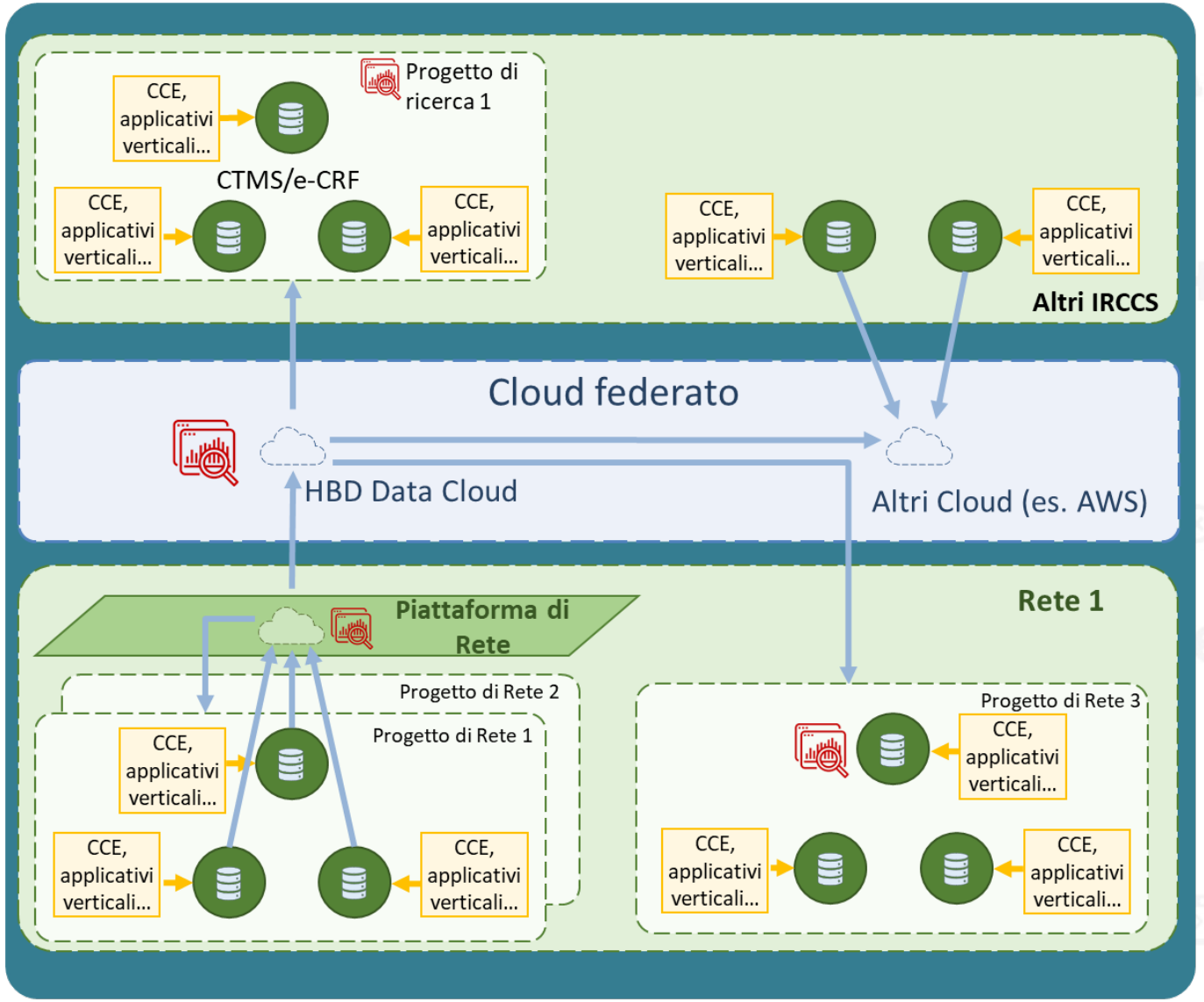
See Arnaud Ceol presentation at BITS on 23 June 2023
https://bioinformatics.it/bits2023

The next step: the federation of all IRCCS-networks



HBD from a life science researcher point of view

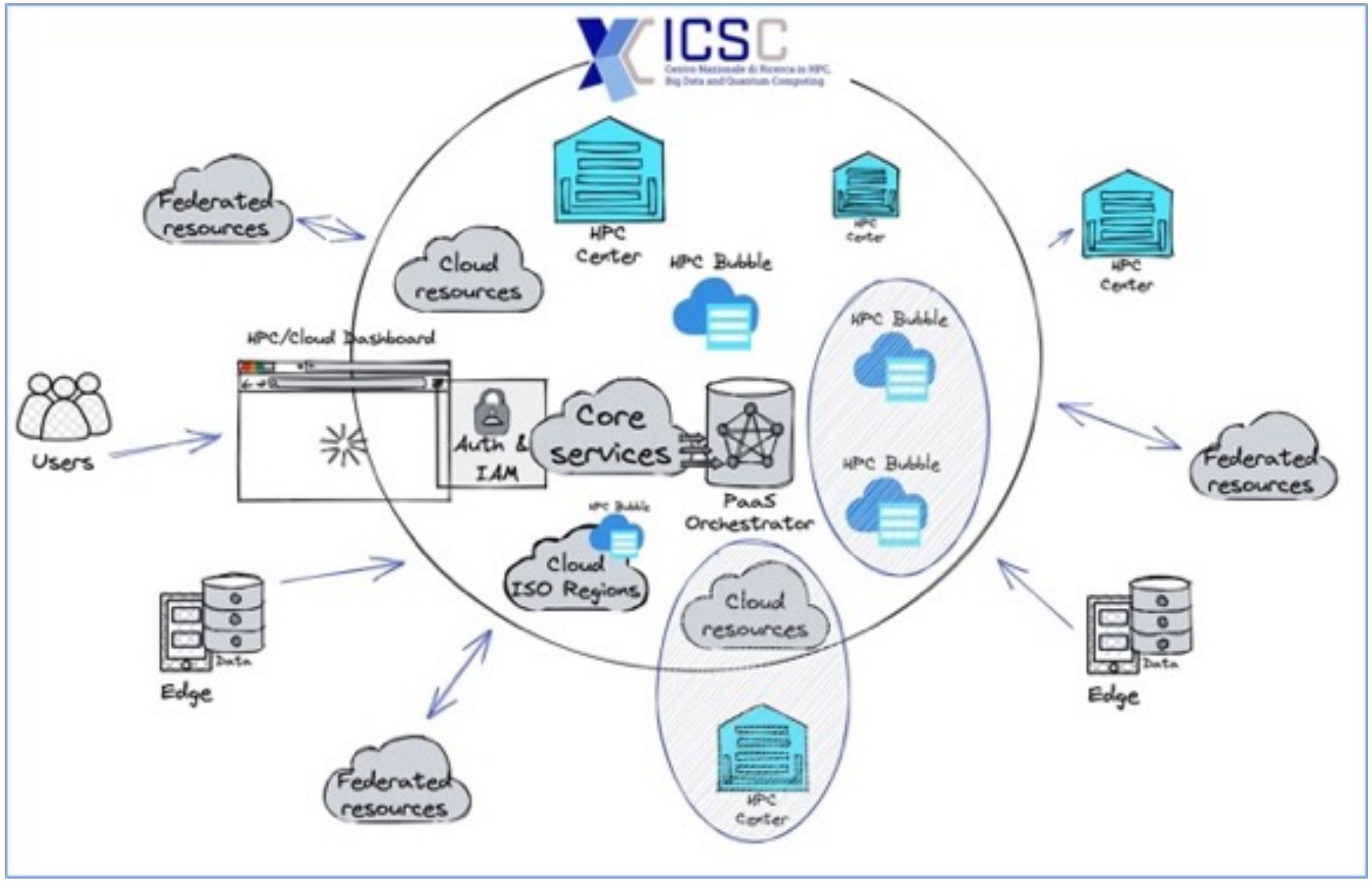
- Extraction/ingestion of
  - information from clinical documents with NLP techniques
  - Omics, signals, imaging and clinical data (CCE and CTMS)
- Data management with INFN Cloud tools like RUCIO and FTS
- Health analytics
- Federation with external cloud providers
- Secure HPC Bubbles available for machine learning and genomic pipelines



HBD Federation with other initiatives at national and international level

- Collaboration with UNCAN.eu on the Implementation and management of a European Federated Cancer Research Data Hub
  - The blueprint will be presented next week in Brussels
- ImmunoPortale project (with AIFA)
- STRIDES project (with National Institute of Health, US)
- Other synergies are being explored with
  - Centro Nazionale terapia genica e farmaci con tecnologia RNA
  - Partenariati estesi sull'intelligenza artificiale e su Diagnostica e terapie innovative nella medicina di precisione
  - Centro Nazionale di Ricerca HPC (ICSC)

The overall architecture: the *continuum* from Edge, to Cloud, to HPC



Thanks!

giacinto.donvito@ba.infn.it
barbara.martelli@cnaf.infn.it
cloud@lists.infn.it