Requisiti da parte delle comunità mediche e life science e loro impatto sull'infrastruttura distribuita

Barbara Martelli

On behalf of “DataCloud team”

Outline

- Panoramica dei progette life-science in ambito computing
  - I progetti
  - I finanziamenti
  - Il personale
- Requisiti/obiettivi dei progetti life-science
  - Mapping su WP DataCloud
  - Mapping su tecnologie DataCloud
- Conclusioni e domande aperte

Progetti life science con coinvolgimento INFN in ambito computing

Evoluzione della piattaforma Alleanza Contro il Cancro (verso HBD-DataCloud)

- Patient, image, sample and sequencing information is registered to the ACC LIMS platform
- Images and sequences data is uploaded to the OneData platform deployed for ACC
- Data collected on the platform are of the following types:
  - 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
- Fine grained authorization
- A unique barcode is automatically generated by the system (ACC LIMS) and is used to name uploaded files, ensuring their traceability throughout the project lifecycle
- Validation and processing of the uploaded data
- Reporting for ACC studies
- More than 80 researchers have been registered up to now on the platform and have registered more than 2800 patients for 5 projects currently ongoing, with genomics data, radiomics data or both.

The amount of “live” data available on the platform is 5,5 TB (plus backups and remote archive copies).

Dimensioni medie dei file genomici

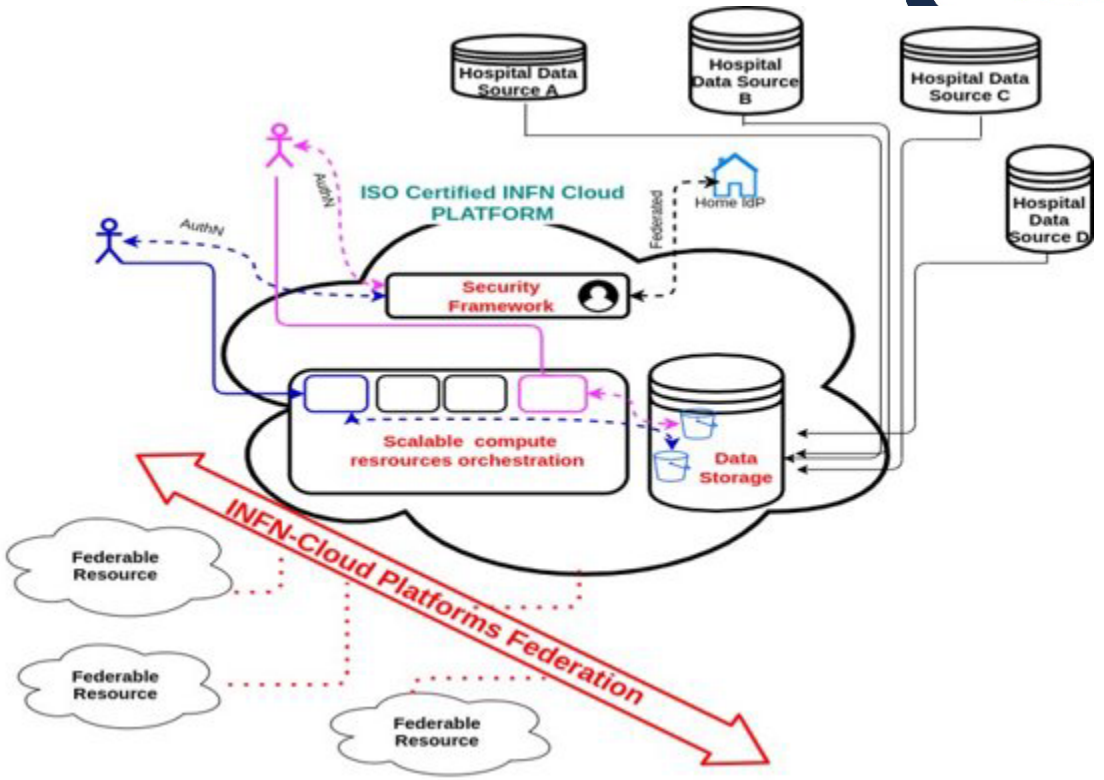
- BAM (Binary Alignment/Map):
  - BAM memorizza le *sequencing reads*, cioè il mapping tra I frammenti di DNA “letti” e la loro posizione in un genoma di riferimento
  - La dimensione del BAM dipende dalla *sequencing deph* che viene chiamato anche *coverage*
  - In un dataset Whole Genome Sequencing (WGS) con “coverage” di 30x e compresso, un file BAM puo’ occupare dai 50 ai 200 GB
- VCF (Variant Call Format):
  - VCF memorizza le informazioni sulle varianti genetiche come Single Nucleotide Polymorphisms (SNPs) e piccole *insertions/deletions* (indels).
  - La dimensione del VCF dipende dal numero di varianti identificate e dal livello di dettaglio
  - In caso di WGS di genoma umano, la dimensione puo’ variare da alcuni MB a decine/centinaia di GB
- MAR (Multiple Alignment Format):
  - MAR memorizza diversi *sequence alignments*, cioè allineamenti di DNA o sequenze proteiche
  - La dimensione dipende dal numero di sequenze allineate
  - Per il genoma umano la dimensione va da alcuni GB ad alcuni TB (GB)

Health Big Data

Creazione di una *Piattaforma Integrata e Federata* nazionale HBD-DataCloud con l'obiettivo di garantire connettività, accesso e condivisione dati alle reti di IRCCS partecipanti (ACC, RIN, Cardio, IDEA) e fornire servizi e capacità di analisi dati avanzate.

- Il nucleo della piattaforma cloud integrata federata HBD-DataCloud è il tenant ACC su EPIC Cloud
- Iniziata la fase di *validazione prospettica della Piattaforma Cloud:* partenza da un progetto scientifico multi-centrico per ciascuna Rete
  - generazione di dati clinico-scientifici presso gli IRCCS
  - ingestione dati nella HBD-DataCloud
  - condivisione dati da parte degli IRCCS afferenti

- Modello integrabile con altre iniziative come ICSC e TeRABIT



- Possibilità per gli IRCCS di generare e memorizzare i dati in locale
- Condivisione su HBD-DataCloud solo di metadati, o sottoinsiemi di dati approvati dai comitati etici, sulla base di specifici progetti scientifici ed in accordo a regole condivise di governance del dato

Accordo di collaborazione scientifica INFN - IRCCS Sant'Orsola

- Creazione di una Piattaforma di Genomica Computazionale che funga da infrastruttura di supporto ai ricercatori nello sviluppo di progetti di ricerca
  - a partire dalle soluzioni già in uso per la Piattaforma di Genomica Computazionale di IRCCS Sant'Orosla e dalla piattaforma di analisi dati genomici sviluppata da INFN - ACC
- Trasferimento dei dati genomici secondo standard di privacy by-design e by-default, automazione e integrità; studio e applicazione di soluzioni basate su GPU (Graphical Processing Units) in metodi di analisi genomica, con l'obiettivo di migliorare la resa computazionale e la scalabilità
- Studio e progettazione di piattaforme Cloud federate ed integrate per la gestione e analisi di dati omici;
- Adattamento di pipeline di calcolo ad architetture di tipo Cloud e DataLake basate su microservizi;
- Promozione dello sviluppo tecnologico e la ricerca nelle discipline genomiche
- Definizione di una modalità stabile di formazione del personale che offra un percorso di certificazione validato per l'utilizzo delle tecnologie e l'elaborazione e interpretazione dei dati
- Esplorazione di scenari di integrazione con progetti analoghi condotti a livello nazionale ed europeo (p.es. 1Million Genomes, Health Big Data, Alma Health DB, etc.) e con iniziative nell'ambito del progetto nazionale di ripresa e resilienza (PNRR)

Accordo di collaborazione scientifica con IFOM (in fase di definizione)

- Tematiche tecnologiche:
  - Utilizzo di Jupyter notebooks integrati con la piattaforma cloud (accesso a GPU, risorse elastiche, connessione a container Singularity, storage S3);
  - Istanziazione di ambienti di calcolo/analisi on demand (batch system Slurm, Jupyter, Conda/pip, R, conversione di ambienti Conda in container Singularity);
  - Estensione delle funzionalità di Galaxy (personalizzazione dei tool e delle configurazioni in base al profilo utente, integrazione di Nexflow per rilasciare pipeline ad utenti meno esperti);
  - Aggiungere alla piattaforma cloud funzionalità per consentire la riproducibilità/tracciabilità delle analisi e la condivisione dei dati in base a utenti/gruppi autorizzati;
  - Integrare librerie e tool per analisi avanzate (alphafold, cellpose, gestione dati di transcrittomica e single cell imaging con tiledb, visualizzazione con Browser Genomici come igv).

DARE (DigitAl lifelong pRevEntion)

- **Duration: 48 months**
- Hub and 3 Spokes
  - Hub: coordinated by Unibo, set up in the form of a **non-profit foundation**, with the initial involvement of 16 partners.
  - Spoke 1: Enabling Factors and Technologies for Lifelong Digital Prevention → North
    - The solution provider
  - Spoke 2: Community-based Digital Primary Prevention → South
    - Target: general population
  - Spoke 3: Digitally-enabled Secondary and Tertiary Prevention → Center
    - Target: patients
- 40 pilot studies, mostly based on spokes 2 and 3.

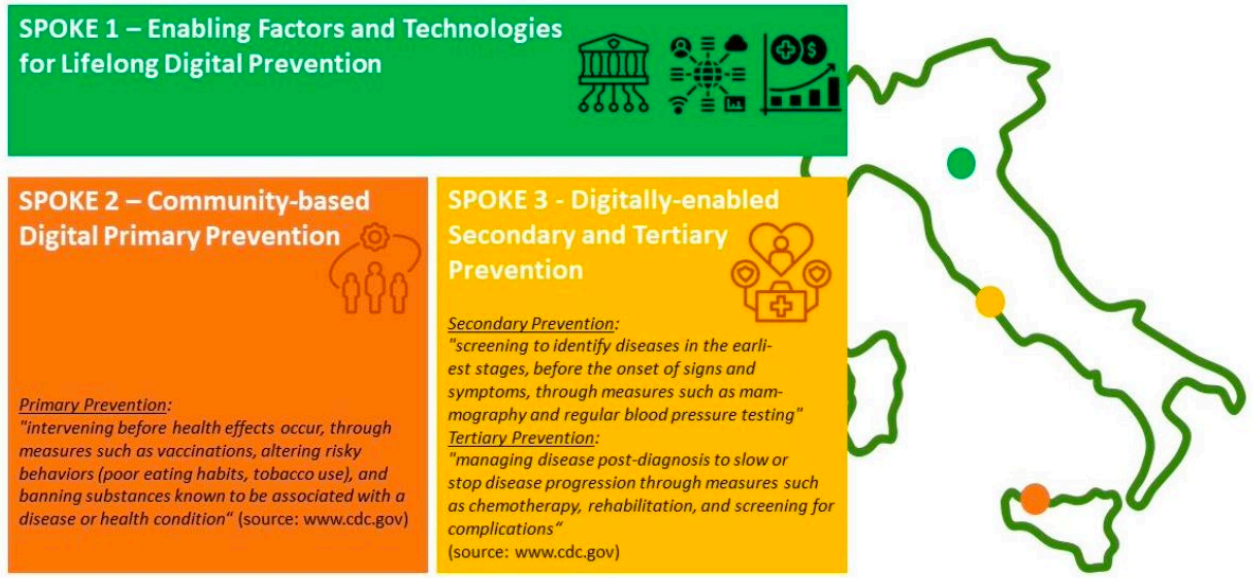
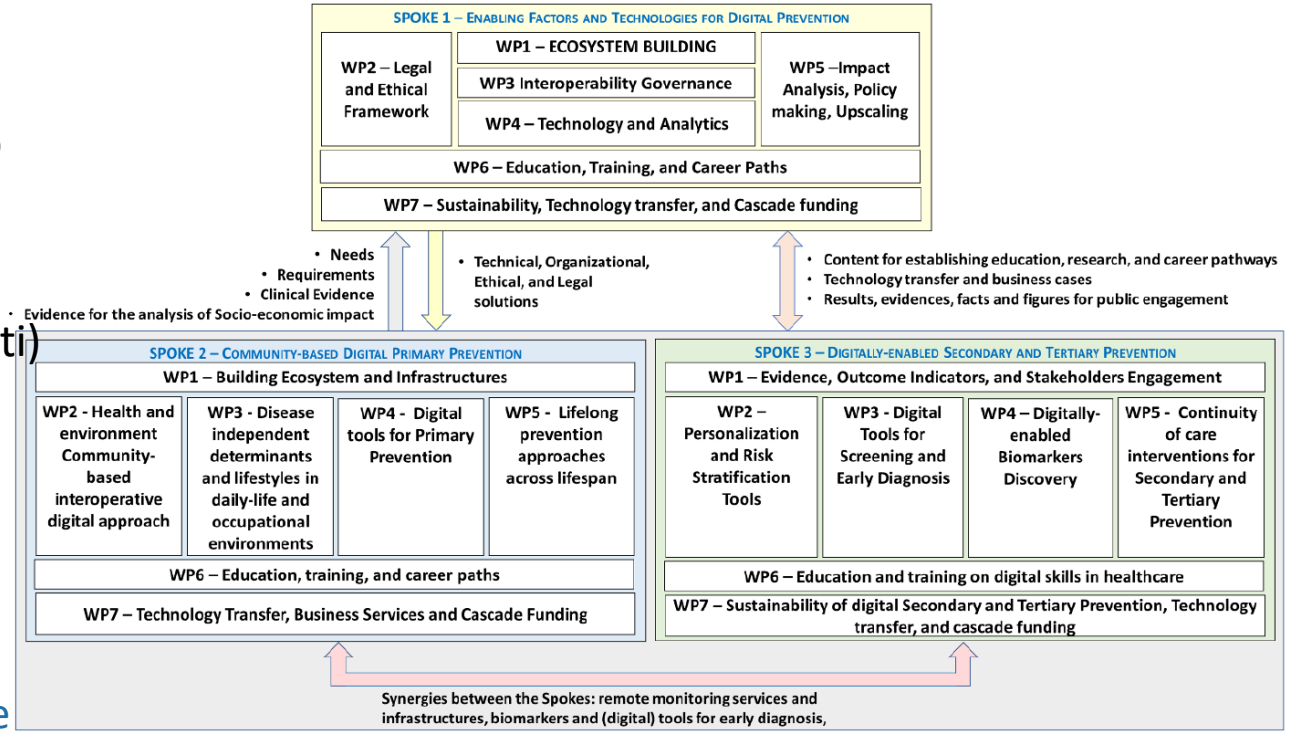


Figure 3- Logical and geographical organization of the 3 spokes

DARE – Ruolo INFN

Partecipazione a Spoke 1

- Lowering and breaking down barriers for adopting innovative, sustainable, high quality, and effective digitally enabled solutions for prevention. The overall objective is to co-create with stakeholders a personalized prevention roadmap for future healthcare that incorporates digital solutions along the entire prevention path.
- Partecipazione a WP2 Legal and Ethical Framework (Foggetti)
  - Leadership del Task 2.3 «Artificial Intelligence regulation and applicable norms» (Foggetti)
- Leadership WP 3 Interobility and Governance (Martelli)
  - Interoperabilità a tutti i livelli (ferro, data entry, data format, API), tenendo conto anche della normativa privacy
- Partecipazione a WP4 Technology and analytics
  - Leadership Task 4.1 Task 4.1 - HPC, HPDA, Cloud, and Edge computing (Chierici)
- Partecipazione a WP6 – Education, Training, and Career Paths



DARE: esposizione CV

- Barbara Martelli, CNAF, 3PM/anno - WP3, WP4
- Cristina Vistoli, CNAF, 1 PM/anno – WP6
- Giacinto Donvito, Bari, 1 PM/anno – WP3, WP4
- Davide Salomoni, CNAF, 1 PM/anno -> da sostituire
- Andrea Chierici, CNAF, 3 PM/anno – WP4 (T4.1 leader)
- Stefano Nicotri, Bari, 2 PM/anno

#	TITLE	Lead	Start	End
WP1	Ecosystem Building	Stefania Boccia (UCSC)	M01	M48
WP2	Legal and Ethical Framework	Matilde Ratti (UNIBO)	M01	M48
WP3	Interoperability Governance	Barbara Martelli (INFN)	M01	M48
WP4	Technology and Analytics	Antonella Carbonaro (UNIBO)	M01	M48
WP5	Impact Analysis, Policymaking, Upscaling	Vincenzo Atella (UNIROMA2)	M01	M48
WP6	Education, Training, and Career Paths	Angela Montanari (UNIBO)	M01	M48
WP7	Sustainability, Technology Transfer, and Cascade Funding	Giuseppe Pirlo (UNIBA)	M01	M48

ICSC - Spoke 8

- Ruolo INFN:
  - WP3 leader “Integrated data flow between clinics and HPC centres”
    - Estensione della certificazione ISO 27k alle regioni DataCloud di Catania e Bari
  - Partecipante a tutti i WP
- Alcuni requisiti emersi dalle discussioni Spoke 8
  - Richiesta precisa di sustained bandwidth di 50Gbps (collegamenti con Terabit?)
  - Parabrics su GPU
  - Pipeline standardizzate
  - Utilizzo di standard, si parte da quelli identificati dal sistema dei dati sanitari attualmente oggetto di riforma
    - OMOP, HL7 FHIR, HL7 CDA (per scambiare dati in europa), DICOM (radiomica), VCF (Variant Calling Format), MAF (Mutation Annotation Format)

ICSC Spoke 8: esposizione CV

Pablo Cirrone, LNS, 1 PM / anno, WP4

Alessandra Retico, Pisa, 1-2-2 PM, WP5

Cristina Vistoli, CNAF, 3 PM / anno, WP3/WP5

Francesco Romano, Catania, 3 PM / anno, WP2/WP5

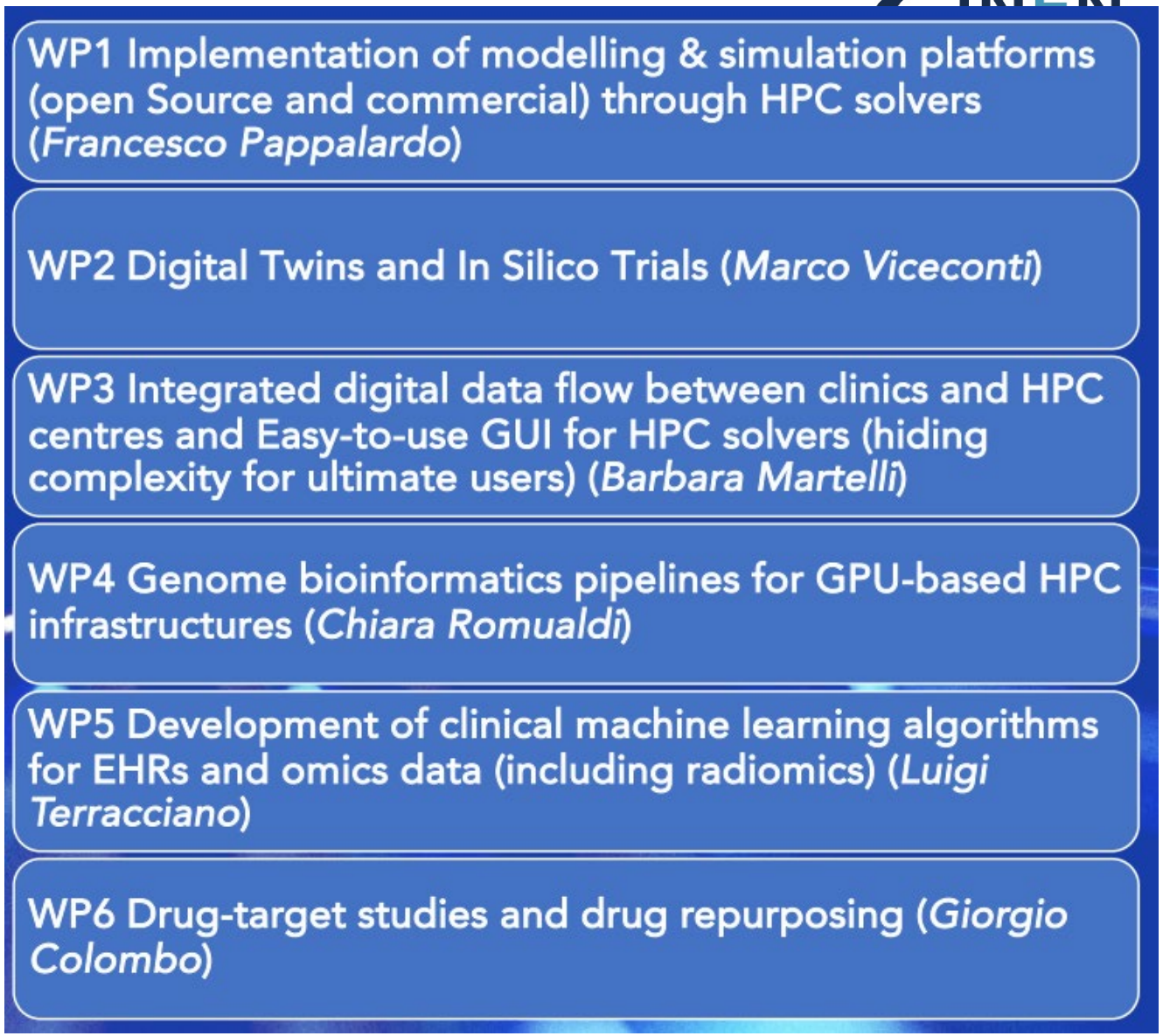
Paolo Cardarelli, Ferrara, 3 PM / anno, WP2/WP5

Davide Salomoni, CNAF, 1 PM / anno, WP1/WP3 -> da sostituire

Giacinto Donvito, Bari, 1 PM / anno, WP1/WP3

Barbara Martelli, CNAF, 3 PM / anno, WP1/WP3

Gaia Pupillo, LNL, 3 PM / anno, WP6



ELIXIR-Italy, DICE

- Partecipazione ad ELIXIR-Italy come parte dell' Italian Joint Research Unit.
- INFN supporta ELIXIR da diversi anni e contribuisce sia all'offerta di risorse, sia allo sviluppo di soluzioni innovative come Laniakea per il deployment automatico di ambienti virtuali Galaxy per le scienze della vita

- DICE in fase di conclusione
  - Partecipazione al datathlon di fine maggio ad Amsterdam con EPIC e Laniakea

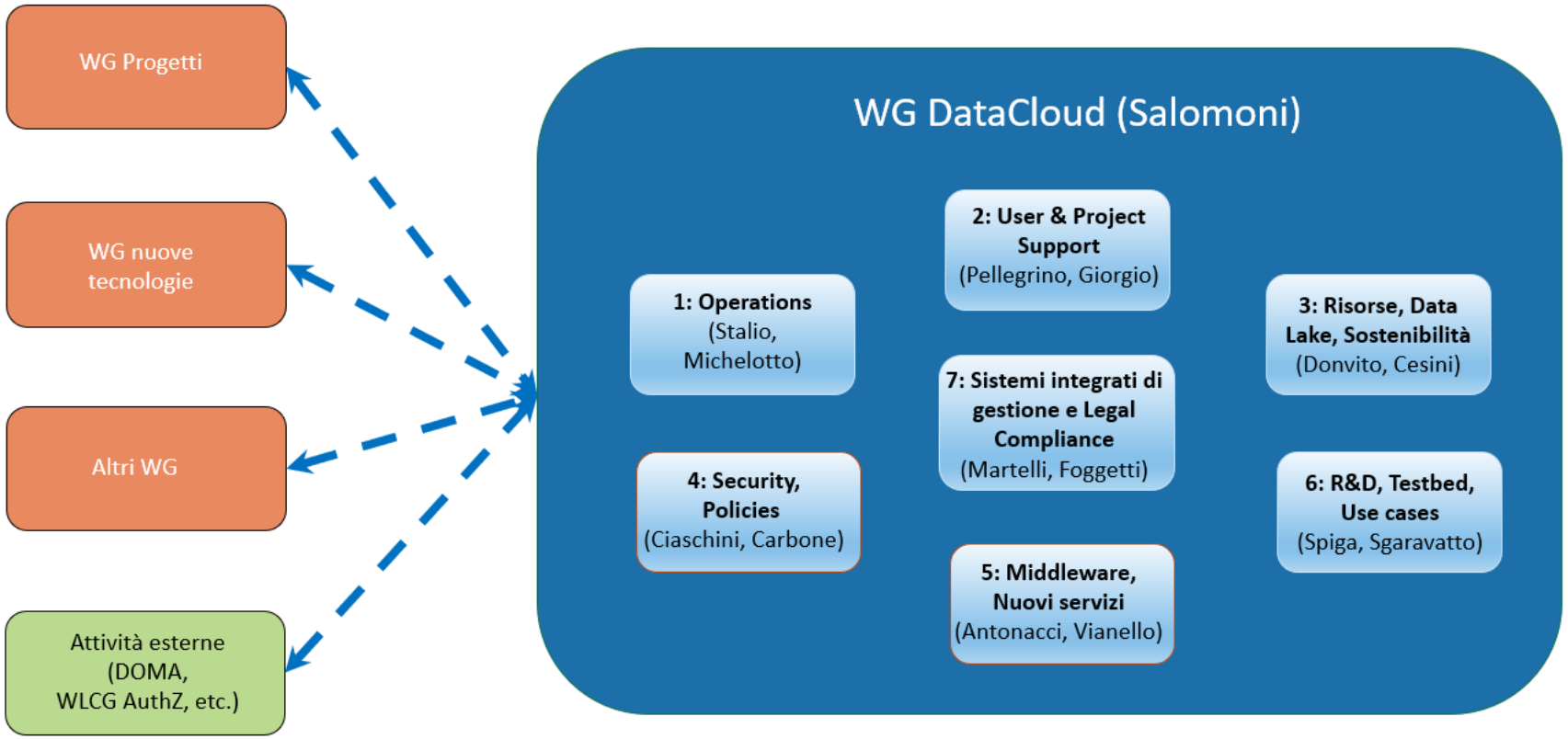
Finanziamenti per HW

- 120 keuro da ACC (accordo 2019-2021)
- 465 keuro/anno da HBD fino al ‘29
- ~330 keuro da Sant’Orsola da rendere fruibili entro metà ‘25
- Spoke 8 dovrà negoziare le risorse passando dal resource allocation board di ICSC
- 2.288.keuro da DARE diviso tra CNAF (~1800k) e Bari (~450k) – il progetto finisce a dicembre ’26
- è in corso la redazione di un accordo di ricerca con IFOM

Personale

- 5 persone dedicate ad EPIC CNAF (progetti DARE e ICSC-Spoke8)
- 1 persona dedicata a EPIC Bari (progetto DARE) + molto effort da anni speso in progetti life science sia dal punto di vista tecnico che legale
- 2 persone dedicate a EPIC Catania (progetto ICSC-Spoke8) + effort del personale che già gestisce un SGSI 27001
- Vari assegni di ricerca acquisibili tramite le singole convenzioni di ricerca (Sant'Orsola 1 AR fino al 2025, HBD 2 AR fino al 2029, IFOM da concordare)
- Varie tesi in corso: Ana Velimirović su blockchain (demo domani), Georgii Iardukhin su AlphaFold as a Service su INFN Cloud

Requisiti/obiettivi progettuali Mapping su DataCloud



Da dove partono le comunità life science

Data sharing and browsing
CBio portal http://cbioportal.org
Genomic Data Commons https://gdc.cancer.gov/

Soluzioni proprietarie
Clara Parabricks (analisi genomica su GPU)
https://www.nvidia.com/it-it/clara/genomics/

Soluzioni home-made
Difficilmente riutilizzabili

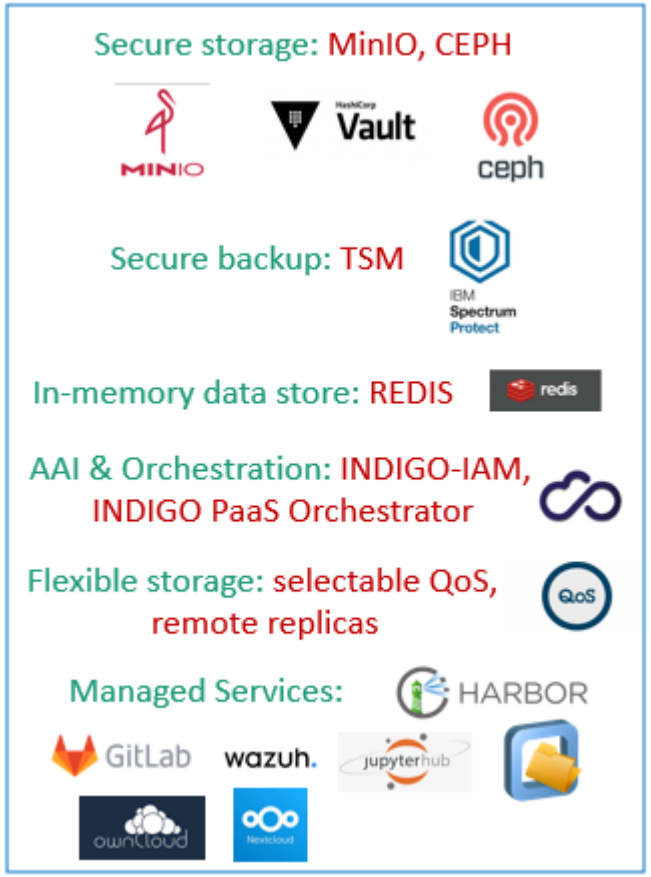
Soluzioni monolitiche

Soluzioni verticali
RedCap https://www.project-redcap.org/
LIMS (Laboratory Information Management System)
MTB Molecular Tumor Board

Workflow managers/pipeline processing/environment managers
GATK https://gatk.broadinstitute.org/hc/en-us
XNat https://xnat.org/
Galaxy https://elixir-europe.org/communities/galaxy
Snakemake https://snakemake.readthedocs.io/en/stable/
Anaconda https://www.anaconda.com/

Cosa DataCloud può offrire

- PaaS Orchestrator (OpenID-Connect Authentication, multi-tenancy, secrets management, dynamic view of servicecatalog)
- INDIGO IAM
- HPC Bubbles
- Data Management Service – plug in your own catalog
- Laniakea
- VPN as a Service
- Scalabilità/elasticità delle architetture (in alternativa alle soluzioni monolitiche)



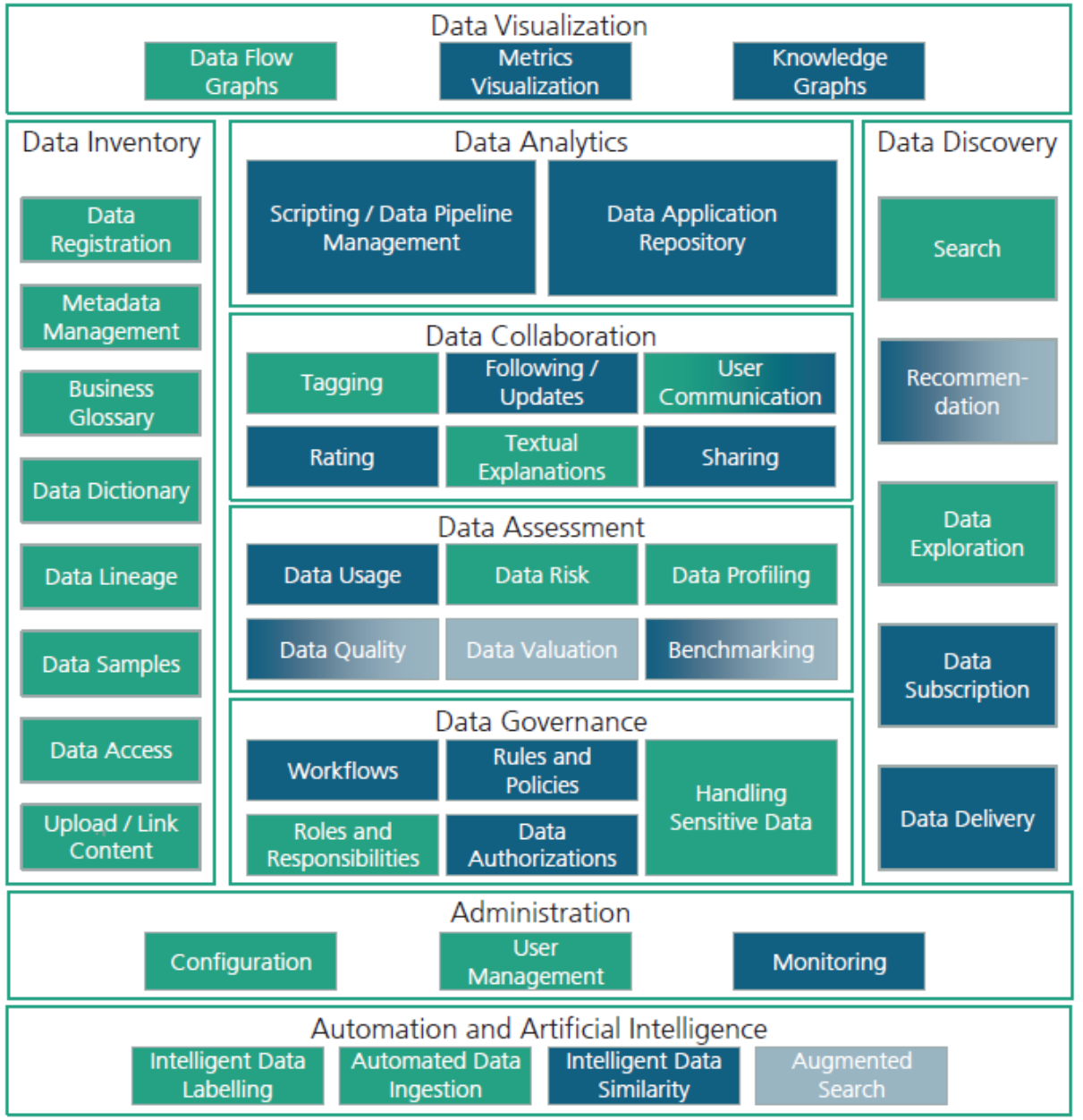
Requisiti comuni a tutti gli use case

- Compliance GDPR e legislativa WP7
  - Certificazione utile a dimostrare ai comitati etici degli ospedali che è legittimo utilizzare risorse di calcolo esterne per gestire i dati omici e clinici
  - Aiuto per DPIA
  - Aiuto per redazione template di accordi che consentano l'invio dei dati dagli IRCCS a INFN
- Modernizzazione degli stack applicativi WP5, WP6
  - Verso architetture cloud ready
  - Verso maggiore interoperabilità (adozione di standard a tutti i livelli)
  - Verso la possibilità di federarsi in uno o più datalake, mantenendo il controllo sulla governance dei dati

Data Catalog

- Attività in corso in HBD WG4
- Analizzata la letteratura, identificate le funzionalità necessarie, identificati alcune offerte proprietarie (AWS Glue, Microsoft Purview, Google Cloud Data Catalog)
- Identificata soluzione open-source: Apache Atlas

→ connettere Rucio ad Apache Atlas? WP6



[Fraunhofer data catalog report 2021]

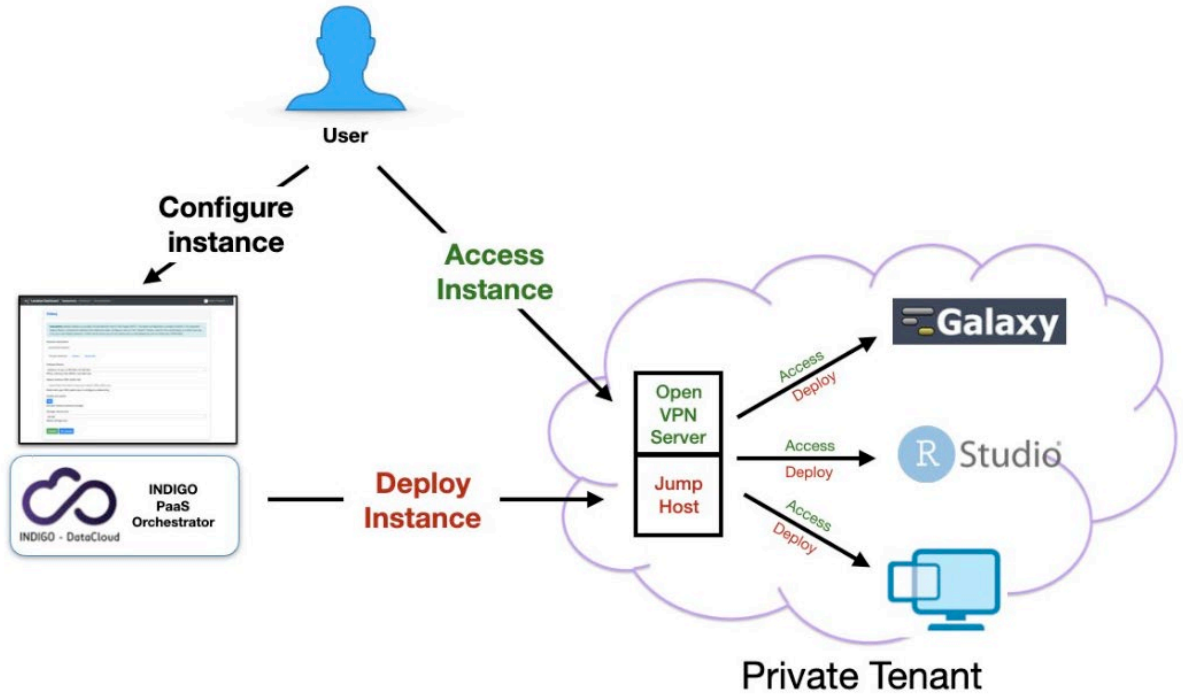
Core Data Catalog Capability
Extended Data Catalog Capability
Add-on Data Catalog Capability

Messa in sicurezza dei tool

- Analisi della sostenibilità del ciclo di vita del software, incluse patch di sicurezza -> definita draft policy per adozione open source in EPIC WP5
- Analisi architetturale sia a livello di singolo sw, sia a livello di piattaforma integrata per il singolo progetto (intero contenuto di un tenant EPIC) Tutti i WP
  - Threat analysis
  - Valutazione se policy EPIC sono applicabili o se sono necessari cambiamenti architetturali o di singole tecnologie
  - Gap analysis: identificazione delle azioni necessarie per mettere in sicurezza il SW
- Analisi vulnerabilità WP4
- Applicazione best practices di sicurezza WP1, WP2, WP4
- Manutenzione del sw su EPIC WP1

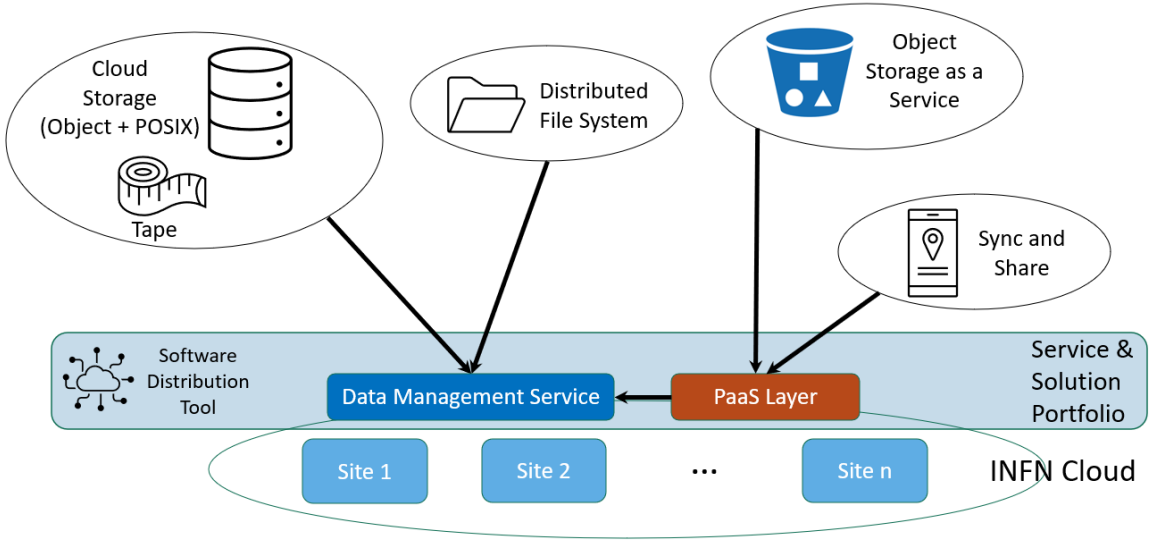
Rendere scalabili e cloud-ready le soluzioni verticali in uso presso le varie comunità

- Deployment di tenant EPIC in modo veloce e scalabile
  - distillando dall'infrastruttura Laniakea le parti che forniscono segregazione dei tenant, deployment VPN server di tenant, cifratura del file system ed utilizzandola come ambiente per applicazioni diverse da Galaxy WP4, WP5, WP6
  - Distillando dalle policy/procedure di EPIC Cloud quelle relative alla gestione delle applicazioni (al modello SaaS per intenderci) e adattandole alle applicazioni verticali di volta in volta individuate WP4, WP7



Deployment di cloud federate presso gli IRCCS

- Servizio che consente agli IRCCS di federarsi con INFN Cloud pur mantenendo i dati e le infrastrutture IT on premises (simile ad Amazon Outposts). Utile quando i requisiti includono:
  - bassa latenza
  - elaborazione dati locale
  - residenza dei dati
- La gestione del servizio presso l'edge può essere fatta
  - da personale dell'IRCCS, in questo caso è necessario fornire ricette Ansible per il deployment, documentazione e policy/procedure di gestione sicura
  - da personale INFN Cloud, in questo caso l'effort dovrà essere finanziato nell'ambito di un progetto di ricerca con AR o simili



Conclusioni e domande aperte

- Le comunità life science avranno necessità di memorizzare ed archiviare grandi moli di dati nel prossimo futuro -> i tool e l'esperienza delle comunità della fisica che affrontano questi problemi da decenni saranno strategiche per trovare soluzioni open source, scalabili, vendor neutral ed interoperabili
- Necessario molto supporto da DPO, Harmony, ufficio legale INFN, siamo strutturati per averlo?
  - Necessario template per accesso risorse INFN da parte di esterni
- Possibilità di ospitare dati genomici pubblici su risorse INFN, abbiamo le policy per farlo?
- Necessario fare scelte architetturali cruciali: framework di AuthN/Z interno ai tenant, data management tools, strumenti di security