gaia-x

Health Data Space Event

4 APRIL 2022

#1 Genomics, Imaging and Clinical Data for Cancer Care and Rare Diseases

Chapter 2a: Breakout Session

Chapter 2a: Breakout Use cases

- **#1 Genomics, Imaging and Clinical Data for Cancer Care and Rare Diseases**

Welcome and Opening

- **Dr. Andrea Derix,** Sr. Global Program Head, Bayer AG

‚experiments of nature' as translational medicine model

- Mechanistic data from human genetics can in retrospect identify the in vivo targets of approved drugs
- New ways to study human biology
  - Study gemline genetic variation
  - Develop genotype-phenotype dose-response curves

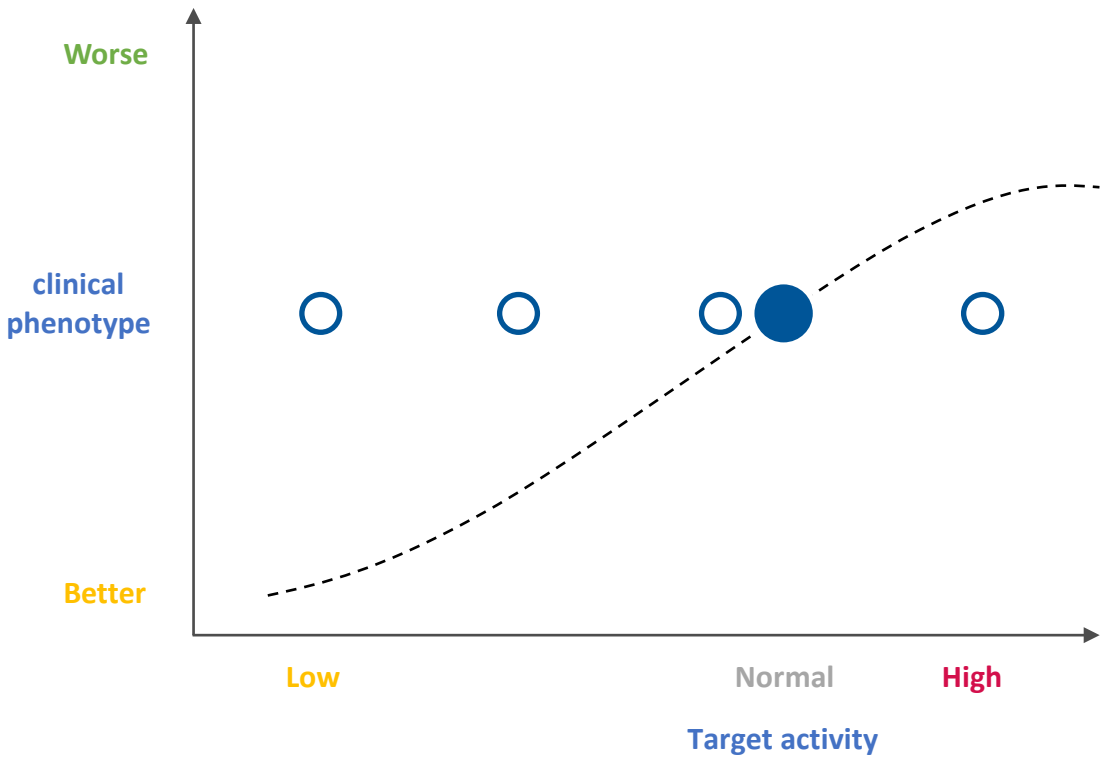
Table 1. Natural selection. Examples of approved drugs with causal support from tissue-specific human autoimmunity or human genetics. *SOST*, sclerostin; *IL6R*, interleukin 6 receptor; *PCSK9*, proprotein convertase subtilisin/kexin type 9; *GLP1R*, glucagon-like peptide–1 receptor; *IL23A*, interleukin 23α; *IL12B*, interleukin 12β.

Clinical disorder	**Tissue-specific autoimmunity**	**Drug**
Encephalitis	DRD2	Antipsychotics
Narcolepsy	Orexin neurons	Suvorexant
Diabetes	Pancreatic beta cells	Insulin
Myasthenia gravis	Acetylcholine receptors	Acetylcholinesterase inhibitors
Thrombotic thrombocytopenic purpura	ADAMTS13 on platelets	Caplacizumab*
Clinical disorder	**Gene(s)**	**Drug(s)**
Schizophrenia	*DRD2*	Antipsychotics
Sclerosteosis and van Buchem disease	*SOST*	Romosozumab*
Rheumatoid arthritis	*IL6R*	Tocilizumab
High cholesterol	*PCSK9*	Evolocumab and alirocumab
Diabetes	*GLP1R*	Incretin mimetics
Psoriasis	*IL23A*, *IL23R*, and *IL12B*	Ustekinumab
Atopic dermatitis	*IL4RA*	Dupilumab*

*Not yet approved.

Analysis of the genetic dose-response relationship can help anticipate clinical consequences of drug target modulation

Slide Source: Daniel Freitag, PhD



„Nature‘s multi-dose randomized trial“:
genetically lower target activity improves clinical phenotype

What are key data supporting the validity of this approach?

~2 fold higher probability to advance from **phase I to approval,** for drug targets with genetic support compared to those without
Source: Systematic analysis of drug development programs (industry-wide) by GSK scientists, Nat Genet. 2015 Aug;47(8):856-60

~2 fold higher success rate in Phase II, for targets with genetic support compared to those without
Source: Review of AstraZeneca pipeline, Nat Rev Drug Discov. 2014 Jun;13(6):419-31

What are key limitations to consider?

Genetic variants lead to **life-long changes** in target activity, which may differ from **short-term** modulation, or intervention at one specific time point (e.g. **later in life**) with a drug

Not all target modulations possible with a drug (e.g. tissue selectivity or bi-specific antibody actions) can be captured reliably by genetic variant effects

Data showing validity of the concept of genetic target validation

The support of human genetic evidence for approved drug indications

Matthew R Nelson[1], Hannah Tipney[2], Jeffery L Painter[1], Judong Shen[1], Paola Nicoletti[3], Yufeng Shen[3,4], Aris Floratos[3,4], Pak Chung Sham[5,6], Mulin Jun Li[6,7], Junwen Wang[6,7], Lon R Cardon[8], John C Whittaker[2] & Philippe Sanseau[2]

Nat Genet. 2015 Aug;47(8):856-60

Table 1 The relative value of genetic support for the probability that a target-indication pair progresses along the drug development pipeline, based on historical drug trial information

	p(progress|genetic support)/(progress|no genetic support)		
Progression	GWASdb and OMIM	GWASdb	OMIM
Phase I to phase II	1.2 (1.1–1.3)	1.2 (1.1–1.3)	1.2 (1.1–1.3)
Phase II to phase III	1.5 (1.3–1.7)	1.4 (1.2–1.7)	1.6 (1.3–1.9)
Phase III to approval	1.1 (1.0–1.2)	1.0 (0.8–1.2)	1.1 (0.9–1.3)
Phase I to phase III	1.8 (1.5–2.1)	1.8 (1.4–2.1)	1.9 (1.5–2.3)
Phase I to approval	2.0 (1.6–2.4)	1.8 (1.3–2.3)	2.2 (1.6–2.8)

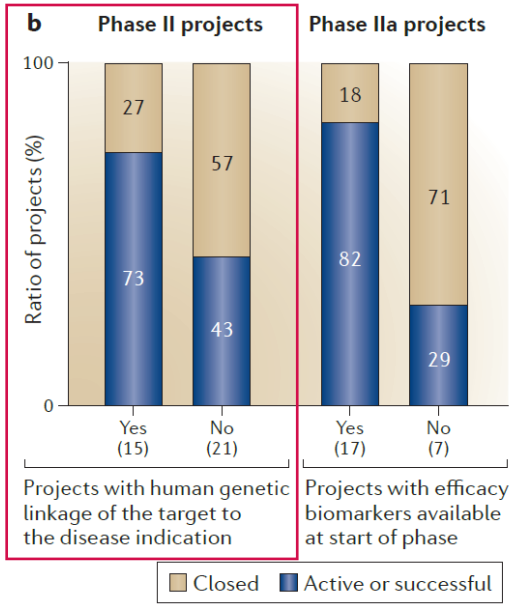
Values give the ratio of the probability of a target-indication pair progressing given genetic support to the probability of progressing without genetic support; 95% confidence intervals are given in parentheses.

OUTLOOK

Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework

David Cook, Dearg Brown, Robert Alexander, Ruth March, Paul Morgan, Gemma Satterthwaite and Menelas N. Pangalos

Nat Rev Drug Discov. 2014 Jun;13(6):419-31



Beyond 1M genome project (Elixir)

gaia-x

- **Juan Arenas Marquez,** Head of ELIXIR Project Management Office

1+MG initiative

Juan Arenas (ELIXIR)

04/04/2022

Accessing genomic data at scale across borders

Long-term strategy: cross-border access to genomic data, implementation of genomics-based health
Thematic working groups (12) & National Mirror Groups
Use Cases working groups: cancer, infectious diseases, rare diseases, common complex diseases, GoE

3 years — 2023

Maturity Model
ELSI toolkits
Technical recommendations and guidelines

1+MG trust framework
- ELSI
- Data and Quality
- Infrastructure
- Maturity model for genomic medicine

Genome of Europe: Data generation, supported by B1MG and upcoming project

4 years — 2026

EUROPEAN HEALTH DATA SPACE

Digital Europe: Genomic Data Infrastructure
- Deployment and operations of the 1+MG infrastructure
- Legal model, governance and business plan
- Advanced use cases:
- New technologies and solutions
- Stakeholders engagement

"The 1+ Million Genomes initiative has the potential to improve disease prevention, allow for more personalised treatments and provide new impactful research"

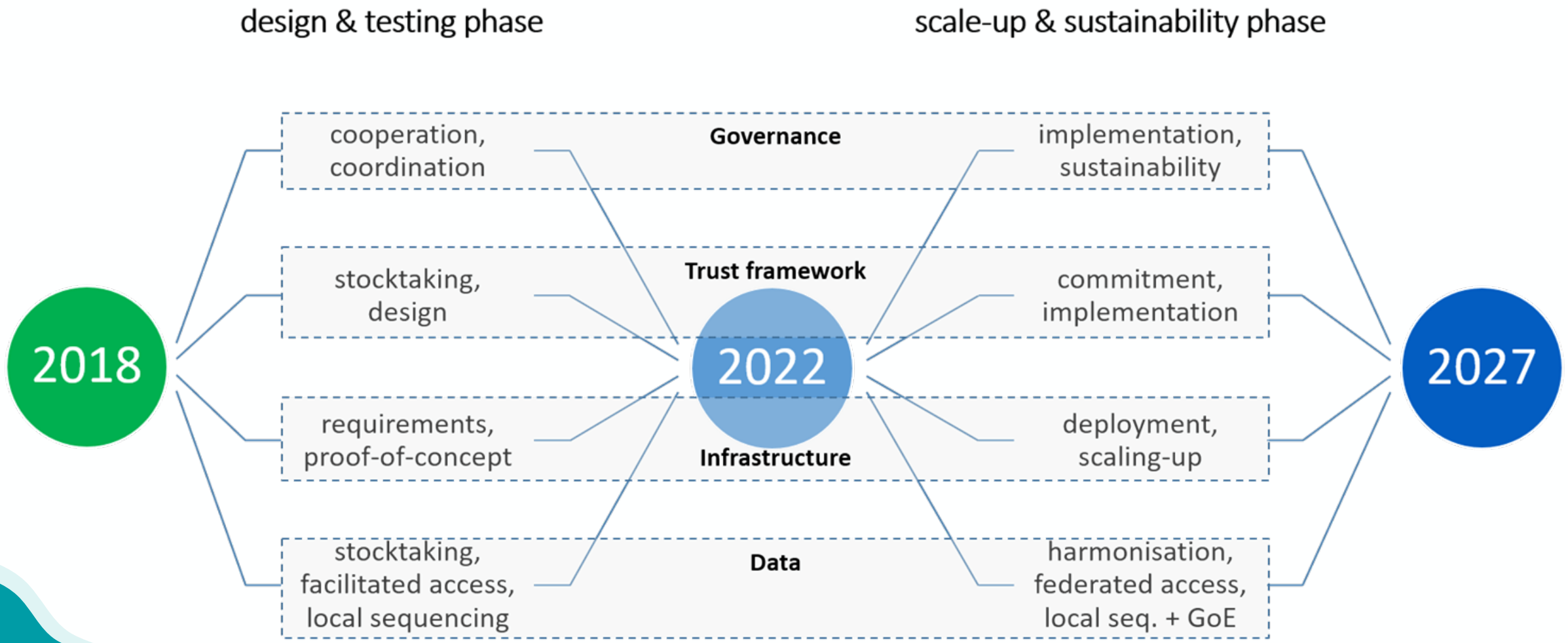
From the 1-million-genome EC web page

Austria
Belgium
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
Estonia
Finland
Germany
Greece
Hungary
Italy
Latvia
Lithuania
Luxembourg
Malta
Netherlands
Norway
Portugal
Slovenia
Spain
Sweden
UK

updated on 18 December 2020

European Commission

1+MG Roadmap 2018-2027



B1MG

1+MG Trust Framework

ELSI

- Transparency and consent policy
- Incidental findings
- Communication of results
- Special (vulnerable) subjects and groups
- Roles for research data sharing
- Legal basis for secondary use
- Data protection impact assessment
- Data governance

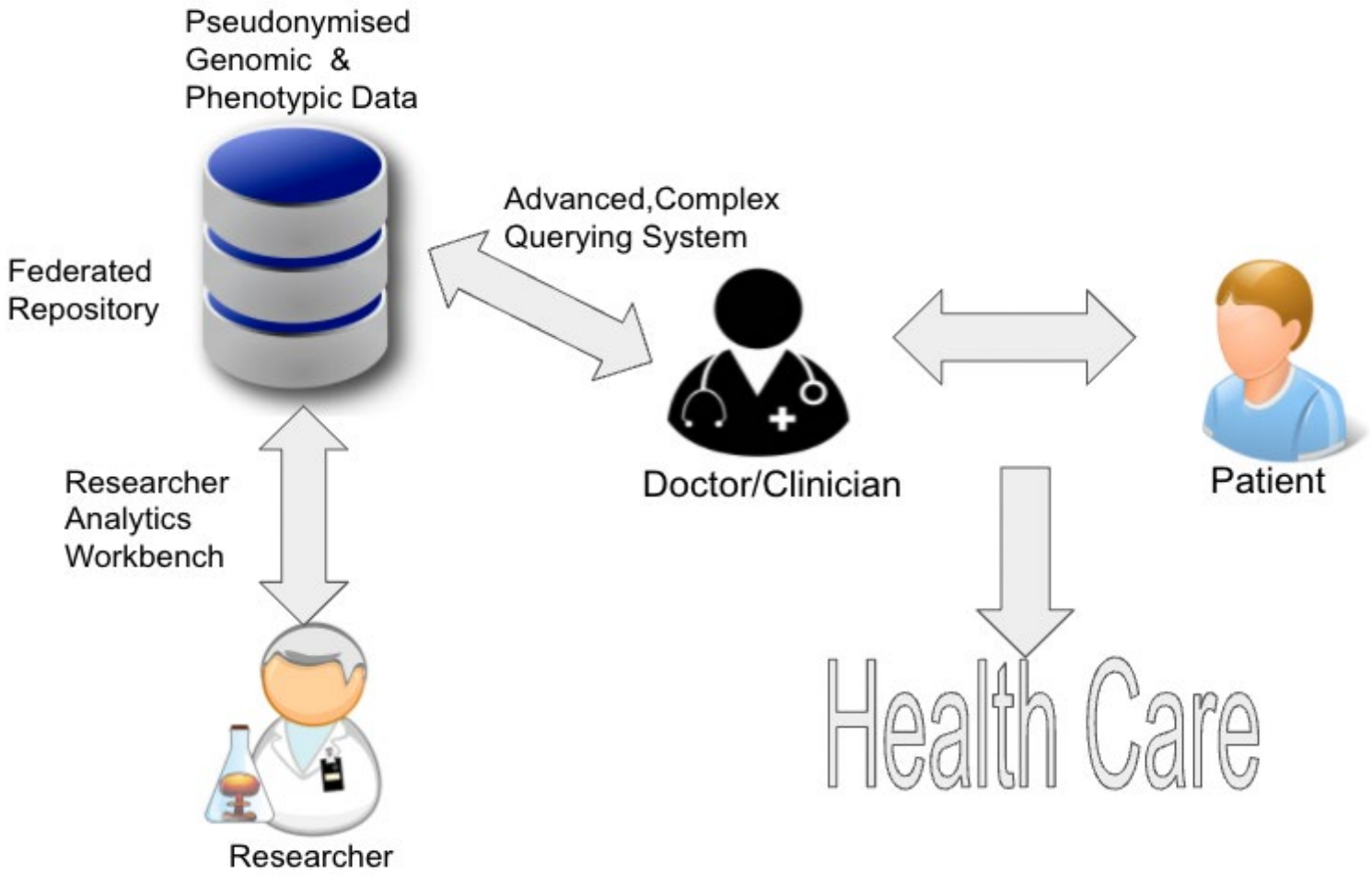
Quality & Data Standards

- Quality metrics for sequencing
- Sequencing practices for WGS & WEs
- Best practices in sharing and linking phenotypic and genomic data
- Minimal metadata models

Infrastructure

- Infrastructure standards stack
- Reference implementations
- PoC using synthetic data

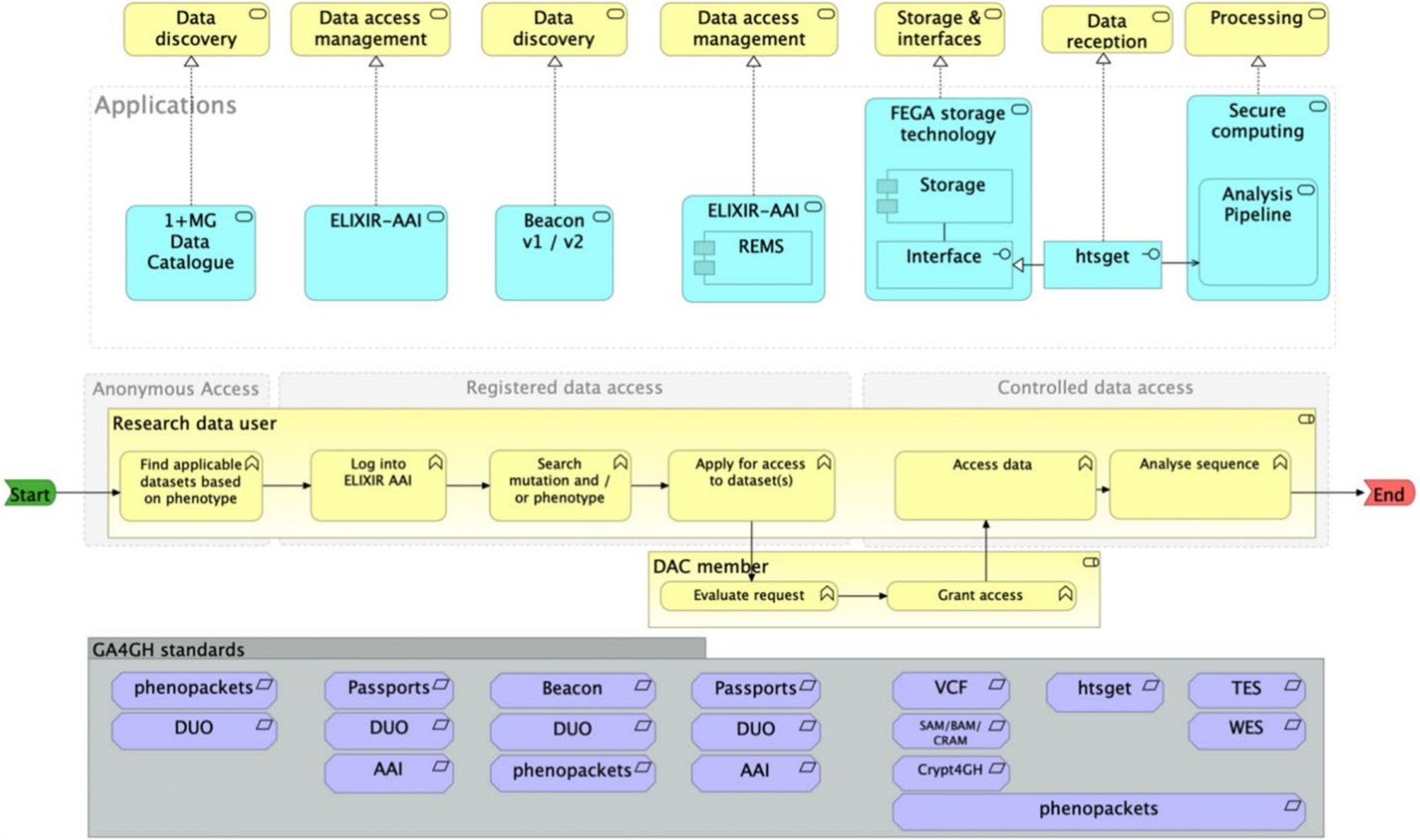
Scenario 3.a : Research repository & Clinical Reference (PoC - Rare Disease)



✓ Demonstrated technical feasibility of genomic and phenotypic data access across borders (ES, FI, SE) using synthetic data

Seeking for similar variants or phenotypes (monogenic disease)

Federated Genomomic Data Infrastructure



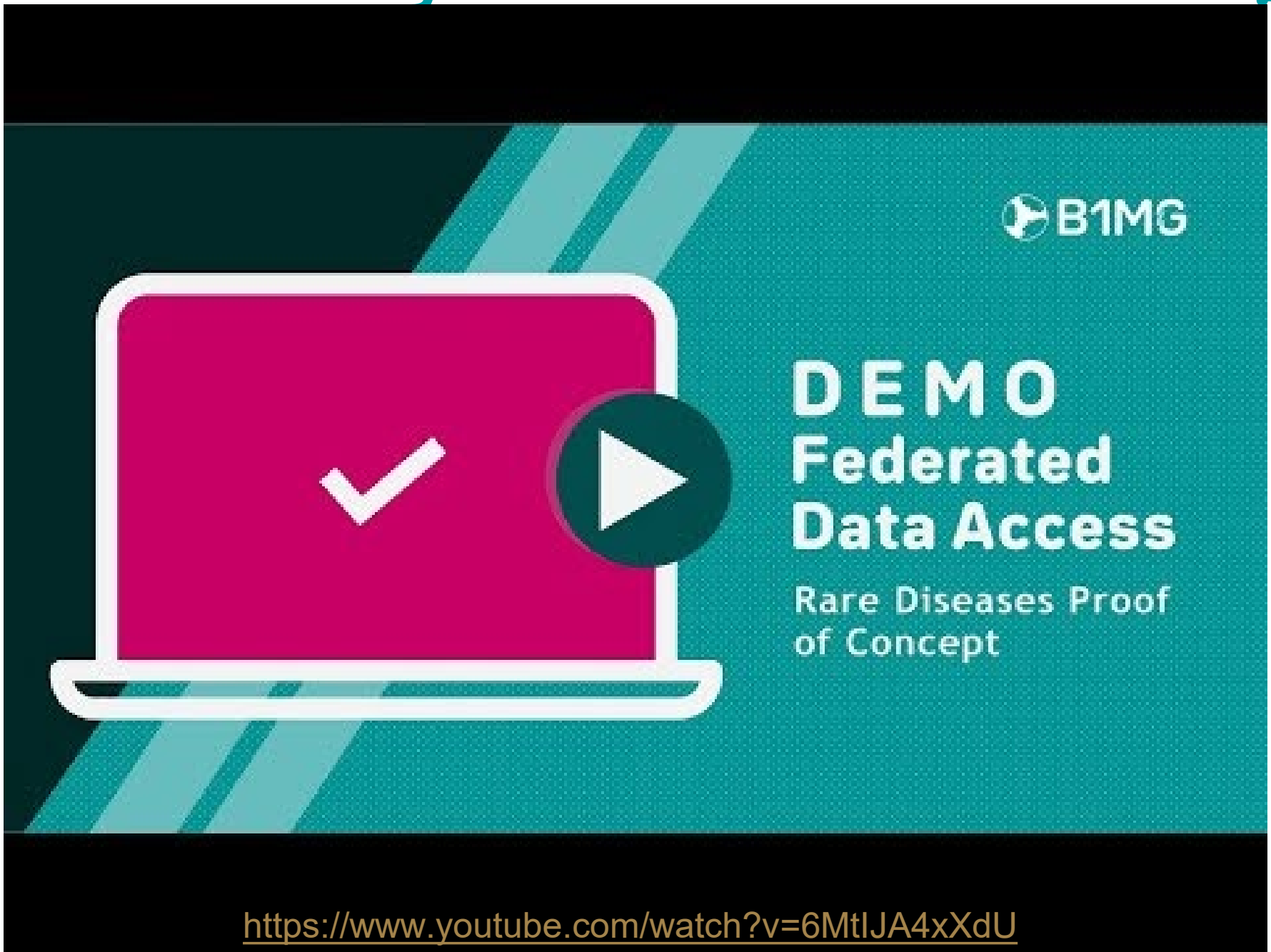
Functionalities

Open source Reference Implementations

Research Workflow

Global Standards (GA4GH)

RD 1+MG PoC by B1MG



https://www.youtube.com/watch?v=6MtIJA4xXdU

Genomic Data Infrastructure

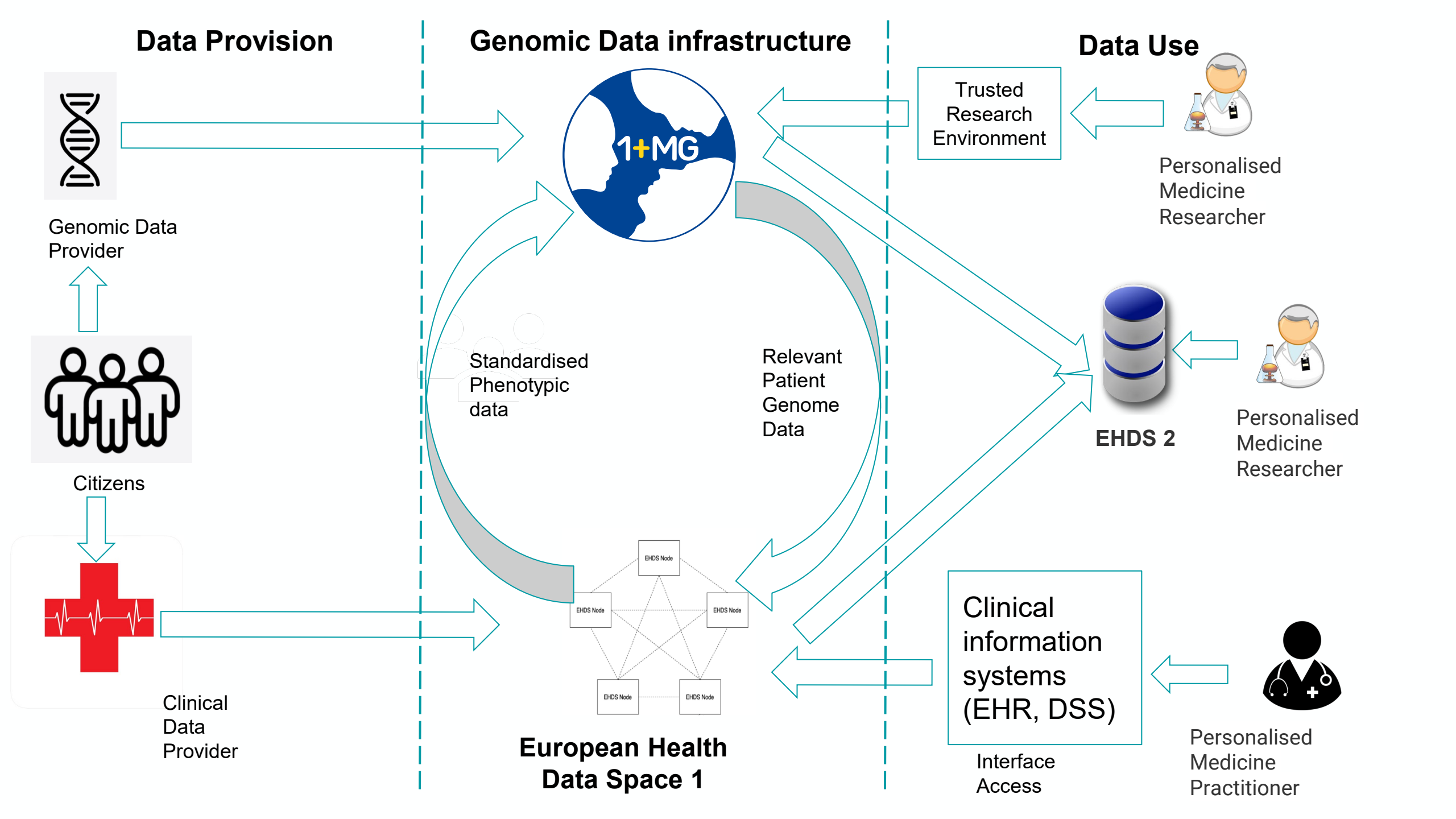
Digital Europe

Deployments and operation of the 1+MG infrastructure

- +13 countries fully operational (production ready) by the end of the project (2026)
  - Members states involvement
    - National funding to be aligned with the project (50% cofund)
    - IT services provisioning institutions nominated by each country to establish the national gateway and to support the deployment of the national infrastructure. **Note:** National infrastructure (centralised, distributed or federated) to be deployed by each MS
    - Data generation activities in each country (GoE + Research + Clinical Data)
    - Capacity building: 1+MG Trust Framework components
    - National helpdesk
  - Central services
    - Helpdesk: Support to users and technical teams
    - Authorisation and access systems
    - Accessible genomes dashboard

Deployments and operation of the 1+MG infrastructure

- Use cases, gap analysis and evaluation of new standards and technologies
  - Use cases and users support(*Clinicians, Innovators and Researchers*)
    - GoE
    - Cancer
    - Infectious diseases
    - Rare diseases
  - Evaluate new standards and technologies to support Use Cases
    - Distributed analysis
    - Federated learning
  - Interoperability with other initiatives
    - European Health Data Space 1/2
    - GAIA-X ?


Data Provision
Genomic Data infrastructure
Data Use
1+MG
Trusted Research Environment
Genomic Data Provider
Personalised Medicine Researcher
Citizens
Standardised Phenotypic data
Relevant Patient Genome Data
EHDS 2
Personalised Medicine Researcher
EHDS Node
EHDS Node
EHDS Node
EHDS Node
EHDS Node
Clinical information systems (EHR, DSS)
Clinical Data Provider
European Health Data Space 1
Interface Access
Personalised Medicine Practitioner

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

juan.arenas@elixir-europe.org

b1mg-coordination@elixir-europe.org

Thank you!

HPC in pediatric oncology

gaia-x

- **Dr. Patrick Kemmeren,** Principal Investigator & head Big Data Core
  Princess Máxima Center for Pediatric Oncology

Princess máxima center
pediatric oncology

Clinical genomics cloud use cases

Patrick Kemmeren, PhD
Principal Investigator & head Big Data Core

Atos

Our mission at the Princess Máxima Center

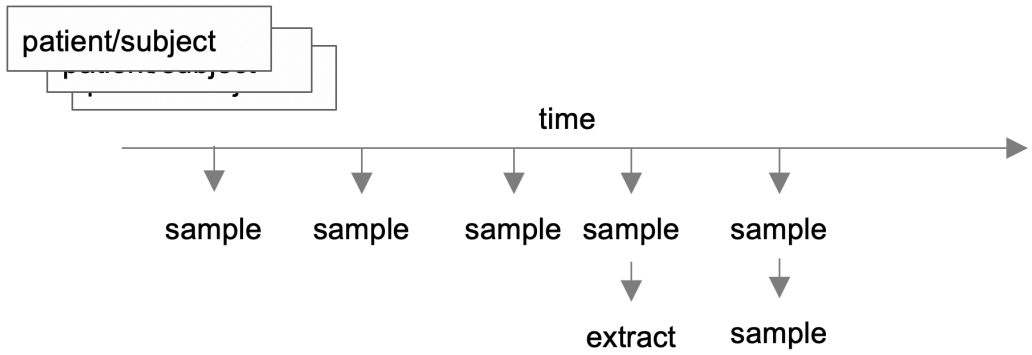
Childhood cancer survival rate has increased the last decades to around 80% today

Still major cause of child death in developed countries

Our aim: 100% survival with better quality of life

Sequencing for precision medicine & biobanking

Collecting of material & data, including DNA(/RNA) sequencing of tumor and healthy tissue of all children treated at the Princess Máxima Center



Clinical translation of research results

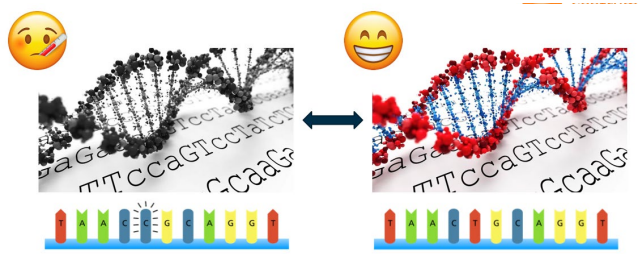
Princess máxima center
pediatric oncology

Clinical research

Healthcare

Diagnostics

Preclinical research



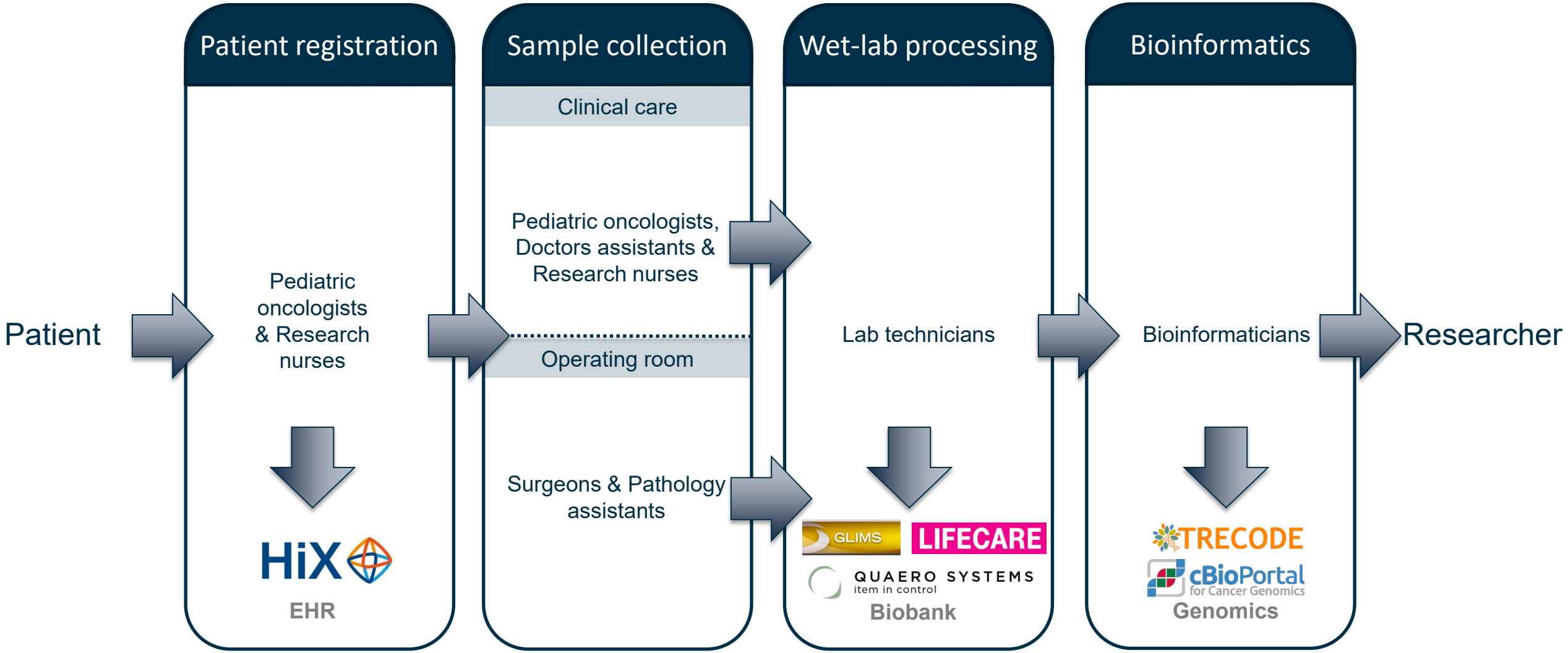
Compare DNA sequence

Sample ID	# of mapped reads	Detected fusion	Fusion detected (rec=reciprocol)	Known fusion
PMABM000AOP	67 Mln	NAB2--STAT6	YES	NAB2-STAT6
PMABM000AOO	77 Mln	NAB2--STAT6	YES	NAB2-STAT6
PMABM000ANZ	76 Mln	NUP98--KDM5A	YES (rec)	NUP98-NSD1
PMABM000AOA	75 Mln	NUP98--KDM5A	YES (rec)	NUP98-KDM5A
PMABM000ANV	68 Mln	KMT2A--MLLT1	YES	MLL-ENL (KMT2A-MLLT1)
PMABM000ANY	67 Mln	POLD3--KMT2A	YES	MLL-AF10 (KMT2A-MLLT10)
PMABM000ANX	67 Mln	KMT2A--AFDN	YES	MLL-AF6 (KMT2A-AFDN)
PMABM000ANU	73 Mln	KMT2A--MLLT3	YES	MLL-AF9 (KMT2A-MLLT3)
PMABM000ANW	71 Mln	AFF1--KMT2A	YES (rec)	MLL-AF4 (KMT2A-AFF1)
PMABM000AON	77 Mln	NAB2--STAT6	YES	NAB2-STAT6
PMABM000AOC	78 Mln	RBM15--MKL1	YES	RBM15-MKL1
PMABM000AOB	121 Mln	KMT2A--MLLT3	NO YES (rec)	CBFA2T3-GLIS2 KMT2A--MLLT3
PMABM000AOY	90 Mln	NAB2--STAT6	YES	NAB2-STAT6
PMABM000AOT	70 Mln	EWSR1--FLI1	YES (rec)	EWSR1--FLI1
PMABM000AOU	80 Mln	SS18--SSX2	YES	??
PMABM000AOZ	29 Mln	no	NO	EWSR1--FLI1
PMABM000AOW	42 Mln	FUS--CREB3L2	YES	FUS--CREB3L2
PMABM000ANR	62 Mln	RUNX1--RUNX1T1	YES (rec)	AML1-ETO (RUNX1--RUNX1T1)
PMABM000ANP	67 Mln	KMT2A--AFF1	YES (rec)	MLL-AF4 (KMT2A-AFF1)
PMABM000ANN	63 Mln	BCR--ABL1	YES (rec)	BCR--ABL1
PMABM000ANM	77 Mln	BCR--ABL1	YES (rec)	BCR--ABL1
PMABM000ANQ	76 Mln	RUNX1--ETV6	YES (rec)	TEL-AML1 (ETV6-RUNX1)
PMABM000ANS	81 Mln	CBFB--MYH11	YES	CBFB-MYH11 Type A
PMABM000ANO	73 Mln	KMT2A--AFF1	YES	MLL-AF4 (KMT2A-AFF1)
PMABM000ANT	78 Mln	CBFB--MYH11	YES	CBFB-MYH11 Type B

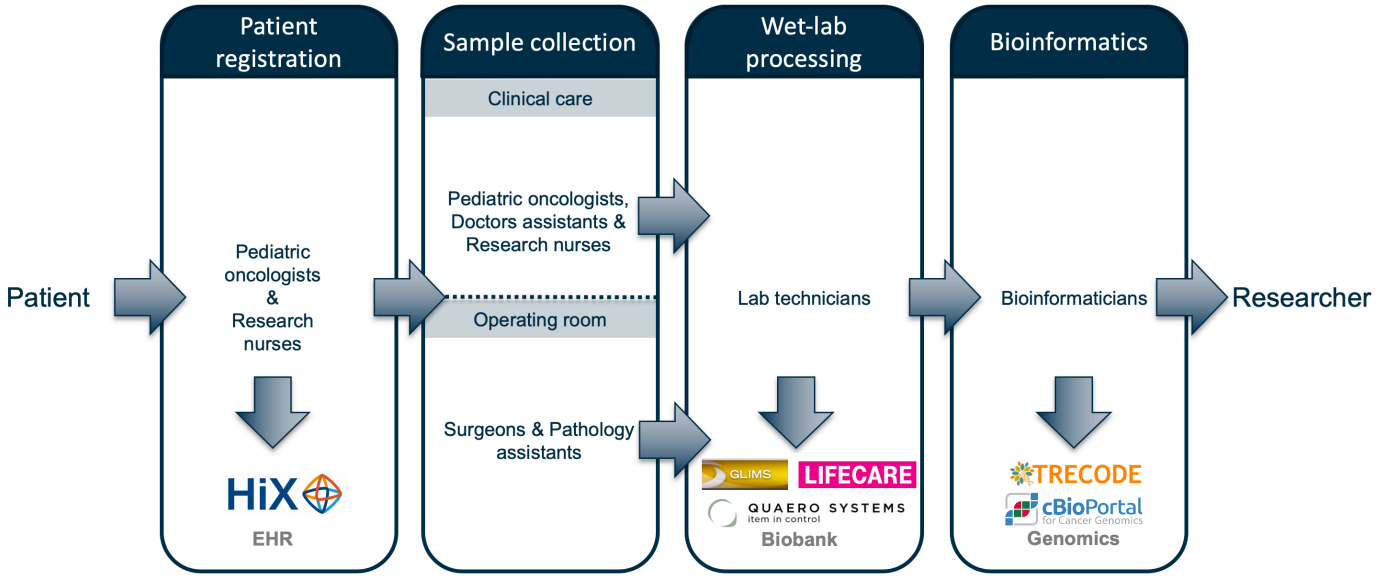
List of abnormalities in tumor DNA:
Single base substitutions, gene fusions, copy number changes

Diagnosis

Diagnostic & biobanking flow



Diagnostic & biobanking flow



Diagnostics
WES
RNA-seq
DNA methylation

Research
WGS
RNA-seq
DNA methylation

Genomics data management

Infrastructure

Princess máxima center
pediatric oncology

End User interfaces

cBioPortal for Cancer Genomics

i2b2 tranSMART FOUNDATION

R2: Genomics analysis and visualization platform

Curation / Morphing

SCRIPTING LANGUAGE

LIMS / Metadata

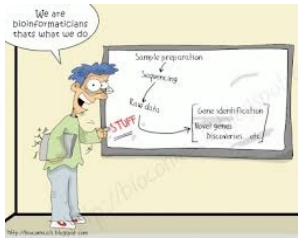
TRECODE

API calls

Storage and workflow infrastructure

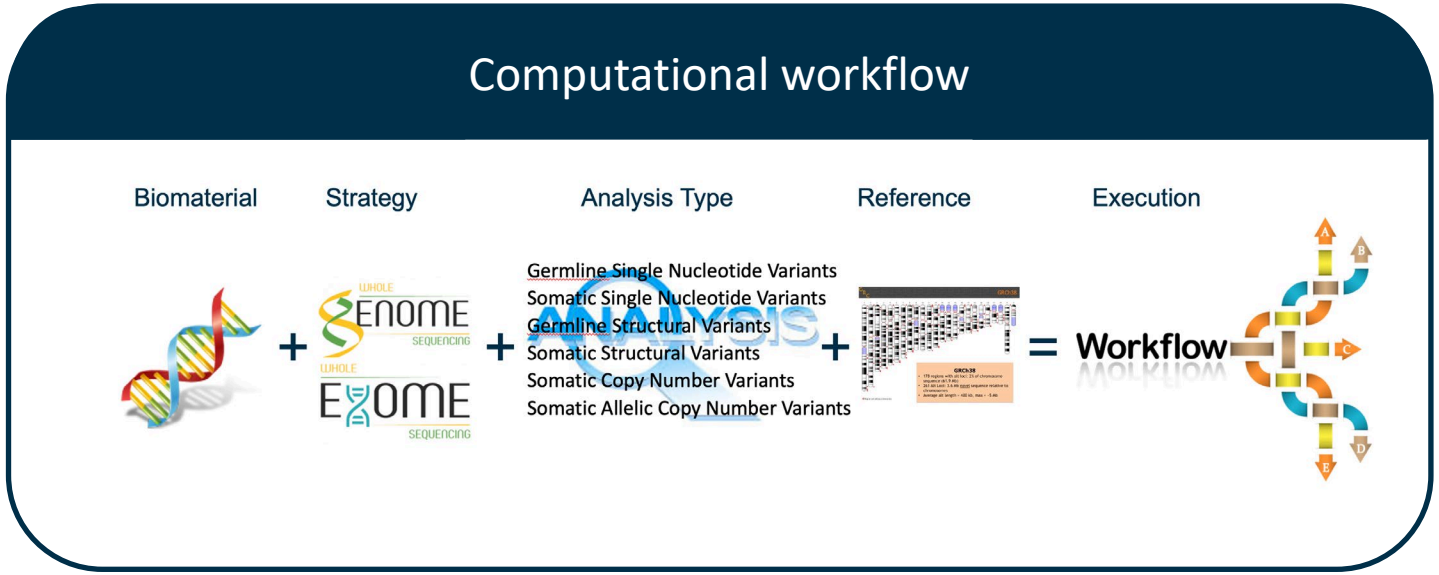
{wdl}

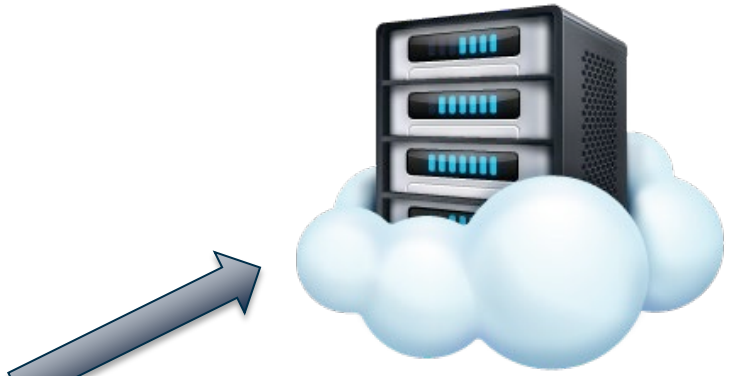
PetaGene
transparent lossless compression



Bioinformatic analyses

Automate computational workflows





HPC cloud

local HPC

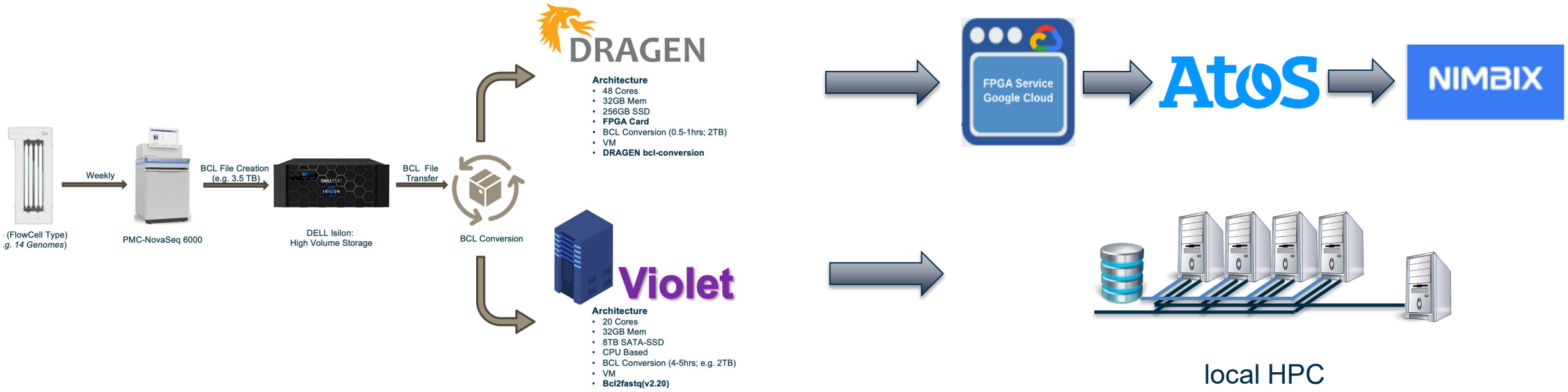
Transferring workflows to the cloud

Use cases/scenarios

In case our local HPC is down, we would like to have a scale-out solution for the diagnostic pipelines (WES +RNA-seq)

We would like to reanalyze the whole cohort of WGS data once every year

We would like to run an internationally harmonized pipeline on ours and collaborators data



Transferring workflows to the cloud

Use cases/scenarios - challenges

Princess máxima center
pediatric oncology

Cost effective

Turnaround time

Transferability

Authentication

Legal/GDPR

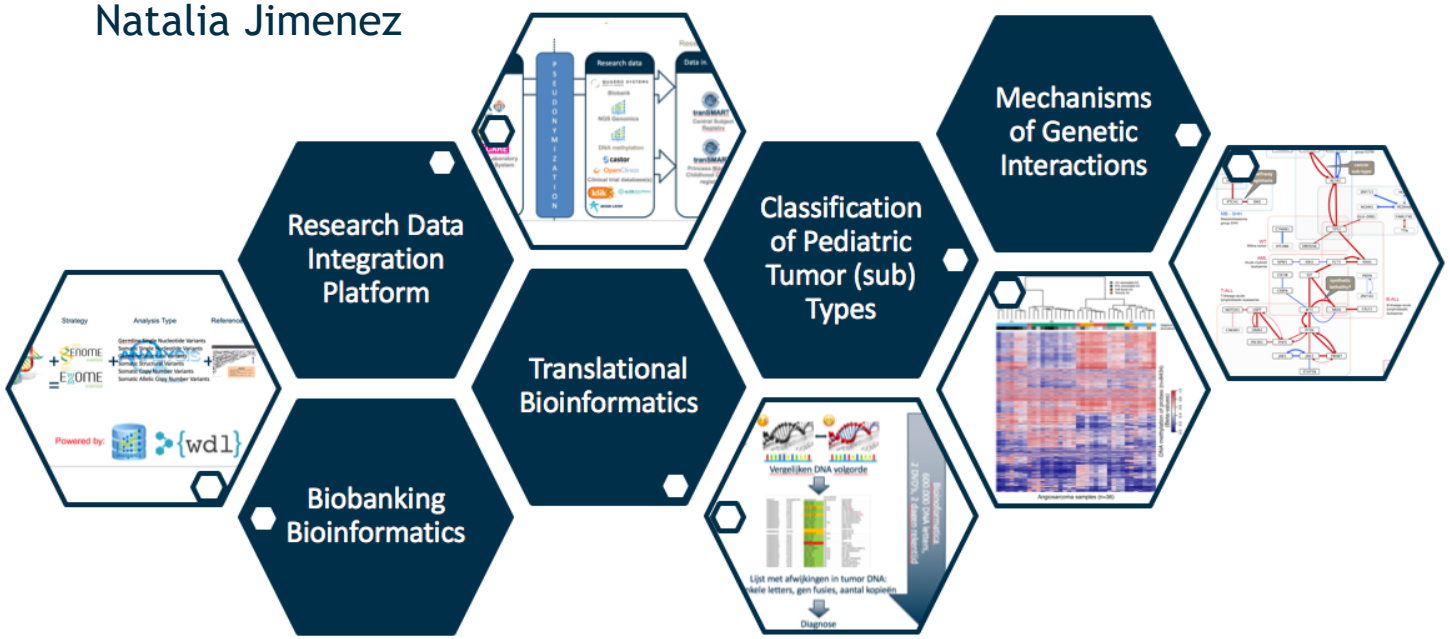
Acknowledgements

Kemmeren group
Jayne Hehir-Kwa
Hinri Kerstens
Ianthe van Belzen
Joanna von Berg
Puck Veen
Richard Gremmen
Ramon van Amerongen
Moniek Schippers

IDT
Paul van Dijk

Big Data Core
Jayne Hehir-Kwa
Eugène Verwiel
Douwe van der Leest
Hinri Kerstens
Shashi Badloe
Alex Janse
John Baker-Hernandez
Sam de Vos
Jet Zoon
Ellen de Jong
Chris van Run
Arianna Tonazolli
Laurene Picandet

Lex de Weille
Erwin Dijkstra
Luc Keppens
Natalia Jimenez



https://research.prinsesmaximacentrum.nl/en/core-facilities/big-data-core

Funding

AI application in omics studies: what we need

- **J.M. Christille, PhD,** Director , Fondazione Clément Fillietroz-ONLUS (OAVdA)

GAIA-X Health Data Space

AI application in omics studies: what we need

J.M. Christille, PhD, Director ,
Fondazione Clément Fillietroz-ONLUS (OAVdA)

Osservatorio Astronomico della Regione Autonoma Valle d'Aosta Planetario di Lignan
Observatoire Astronomique de la Région Autonome Vallée d'Aoste Planétarium de Lignan

Osservatorio Astronomico della Regione Autonoma Valle d'Aosta (OAVdA)

Scientific Research

Individual research projects and undergraduates/graduates/ Ph.D tutoring in Astrophysics

Technology Transfer

Applying solutions developed to study the Universe in other fields: manufacturing, ICT, etc.

Public Outreach and Education

Disseminating astronomy and astrophysics to the general public; supporting teachers in their work.

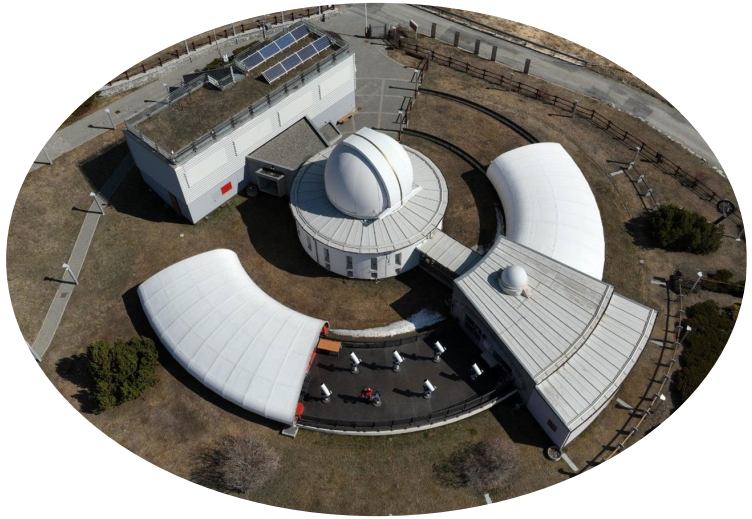
Founding members:

Selected as institutional body by:

Other main funding institutions and agencies:

FONDAZIONE CLÉMENT FILLIETROZ - ONLUS

Astronomical Observatory of the Autonomous Region of the Aosta Valley



+  = ?

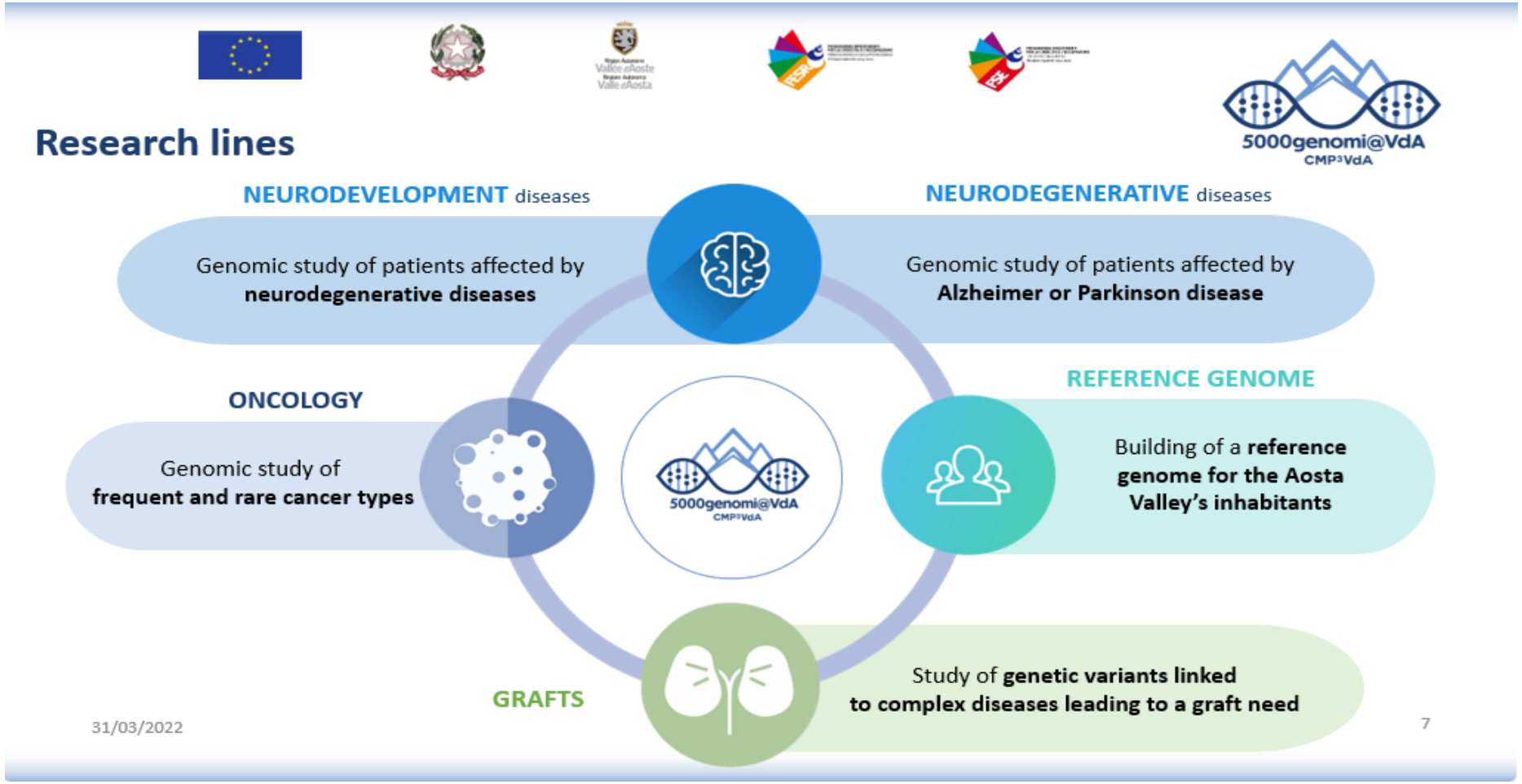
Big Data generators

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

doi:10.1371/journal.pbio.1002195.t001

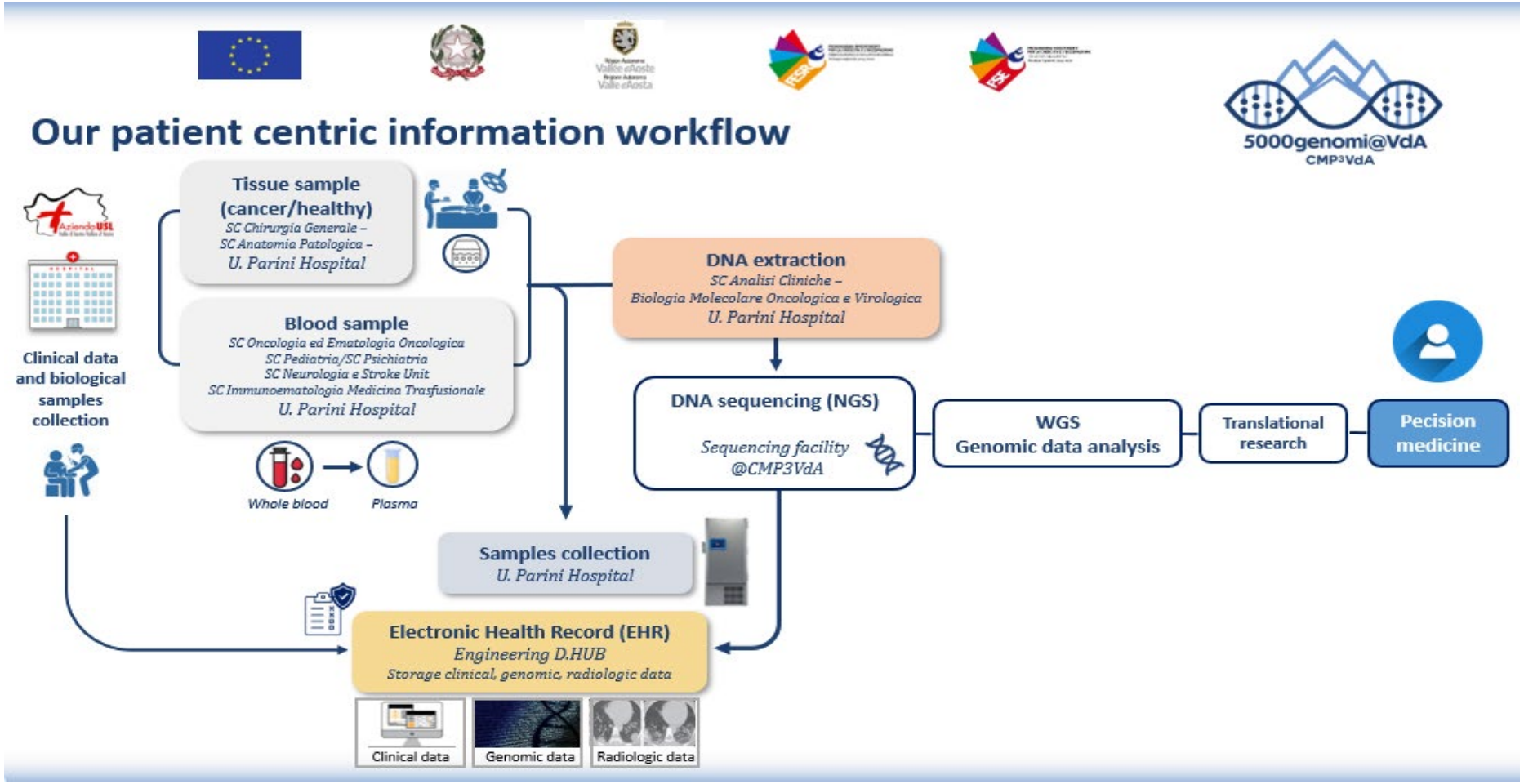
Credit: Stephens ZD, Lee SY, Faghri F, Campbell RH et al. (2015) Big Data: Astronomical or Genomical? PLoS Biol 13(7)

CMP³VdA



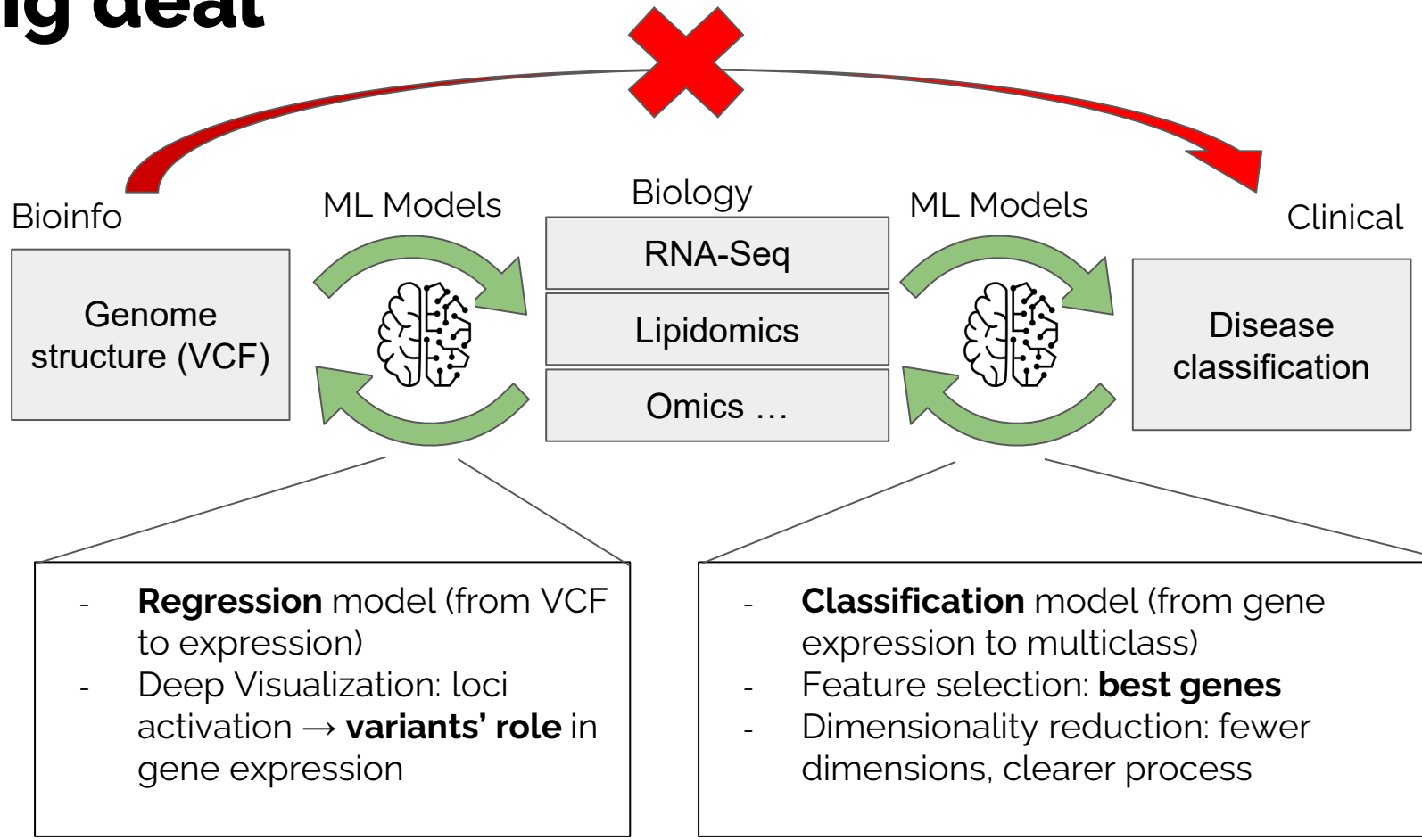
Credit: CMP3VdA - 5000genomi@VdA consortium

CMP³VdA



Credit: CMP3VdA - 5000genomi@VdA consortium

AI: a big deal



A multi-omics approach is needed!

Federated AI: a bigger deal

In order to get well trained AI models, achieving Data Access on federated platforms is not the only problem:

- ➔ Biological and Data Acquisition Standards must be defined to guarantee reproducibility
- ➔ HPC Platforms enabling technologies for data driven solution must be released in open access for research purposes

A radical change in the researchers' mindset is needed together with strong support from national health infrastructures.
A growing open access and diffused data system is the only way to drive cutting edge research.

Thanks!

Any questions?

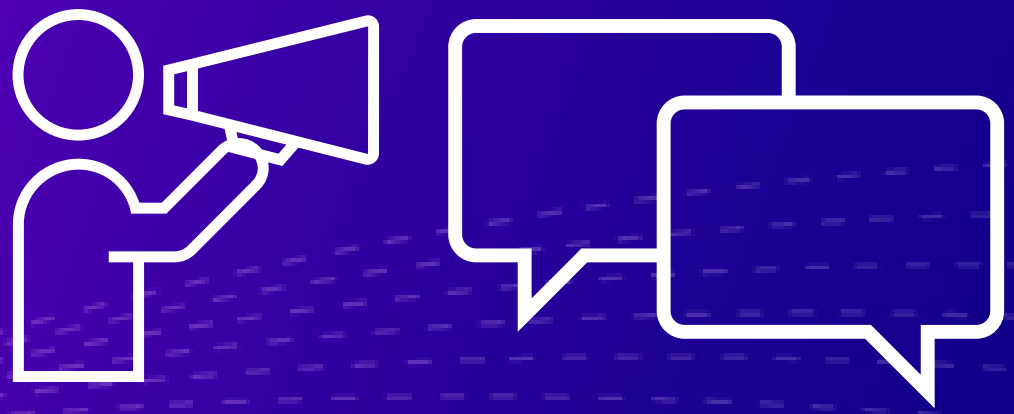
You can find me at **direttore@oavda.it**

www.oavda.it

linktr.ee/OAVdA

Q&A

gaia-x



Carecol - gastric cancer data collaborative

gaia-x

- **Andrea Pescino,** Founding Partner, Stratejai

Data Collaborative for Gastric Cancer

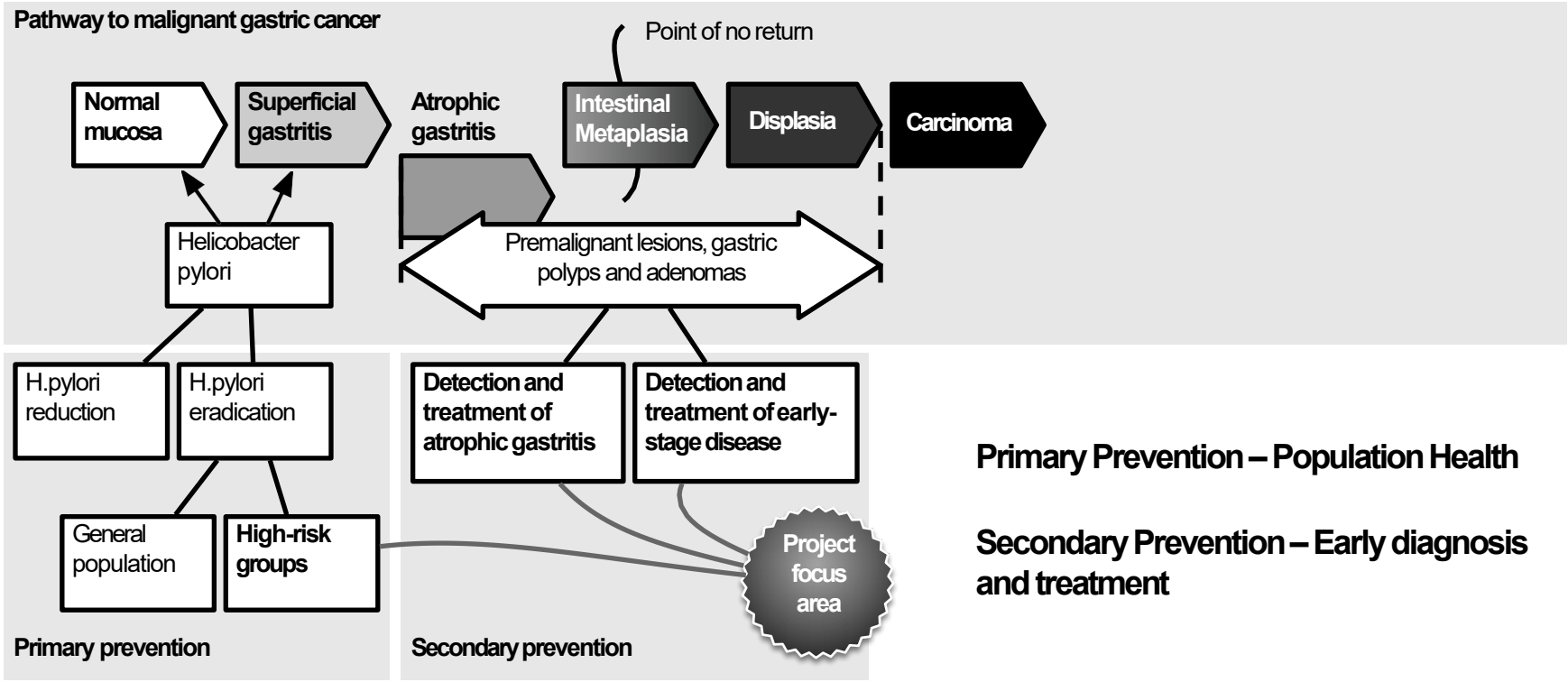
Carecol

High level technology brief

Carecol –Cancer Research Data Collaborative use case of gastric cancer

Gastric cancer affects 1 million new patients every year.

Our project focuses on the detection of patients in the earliest stages of the disease, in order to start treatment well before the formation of carcinomas

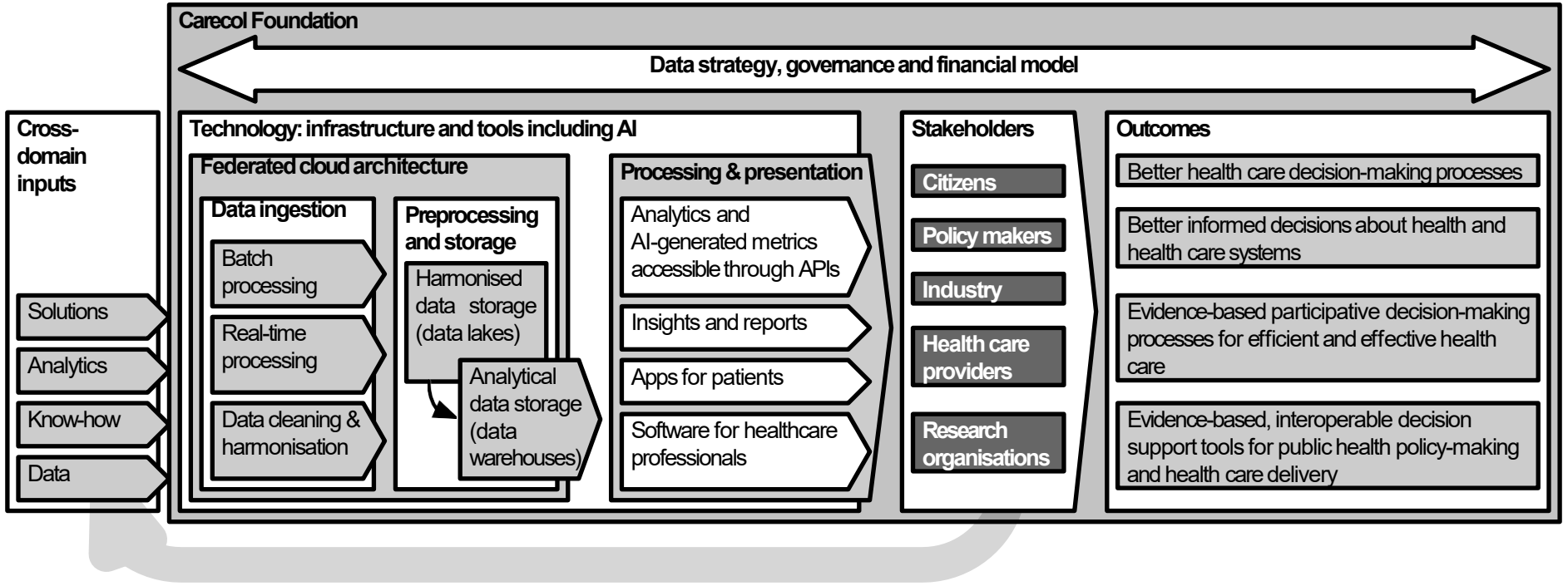


Primary Prevention – Population Health

Secondary Prevention – Early diagnosis and treatment

Carecol Project objective

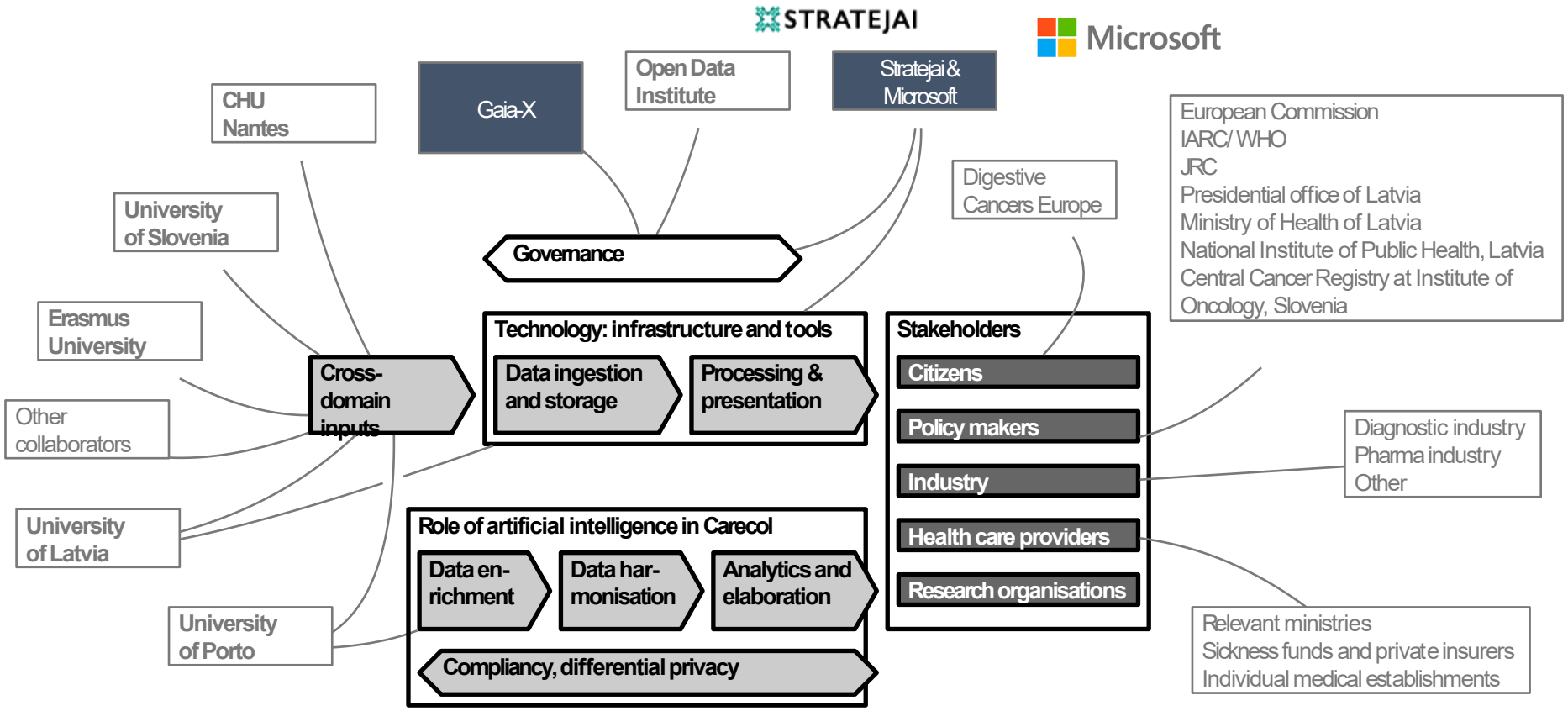
The project's objective is a **technically and clinically validated, data-agnostic**, data-driven decision-support tool that combines various data sets on gastric cancer for better health care delivery and policy making



High Quality Data Foundation – Platform for further value development

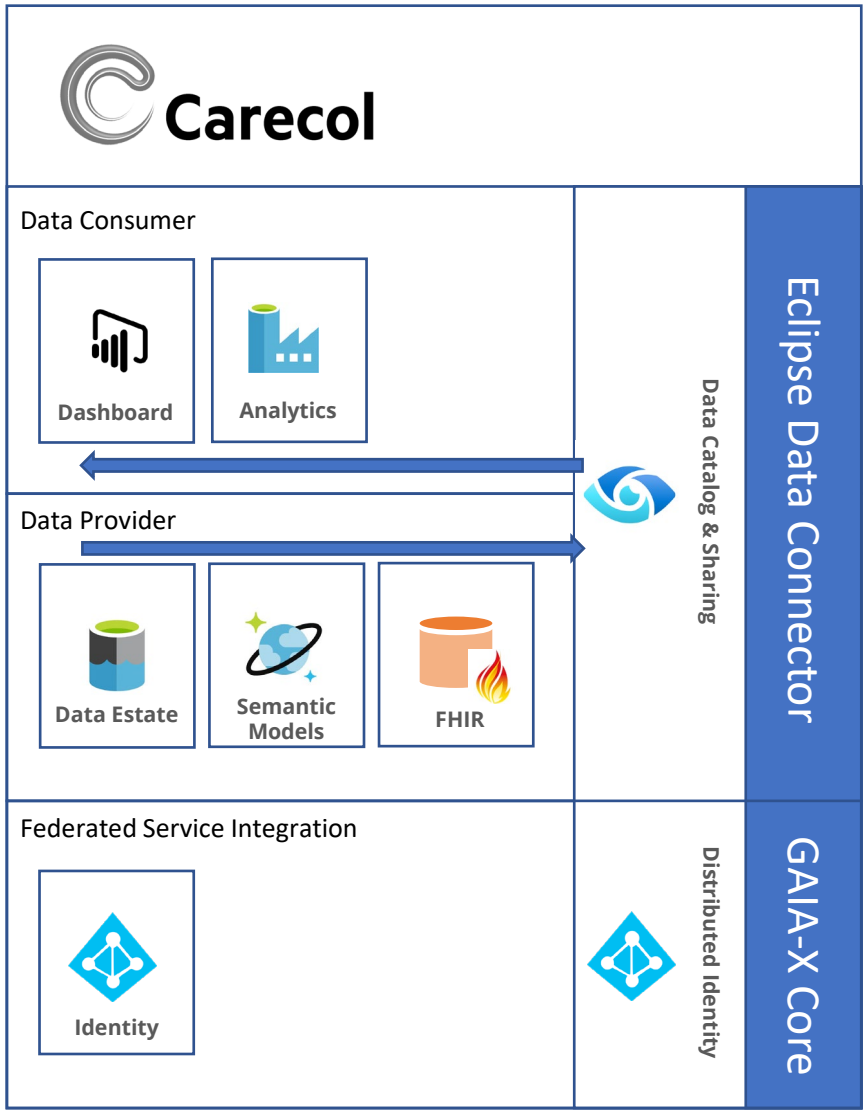
Ecosystem

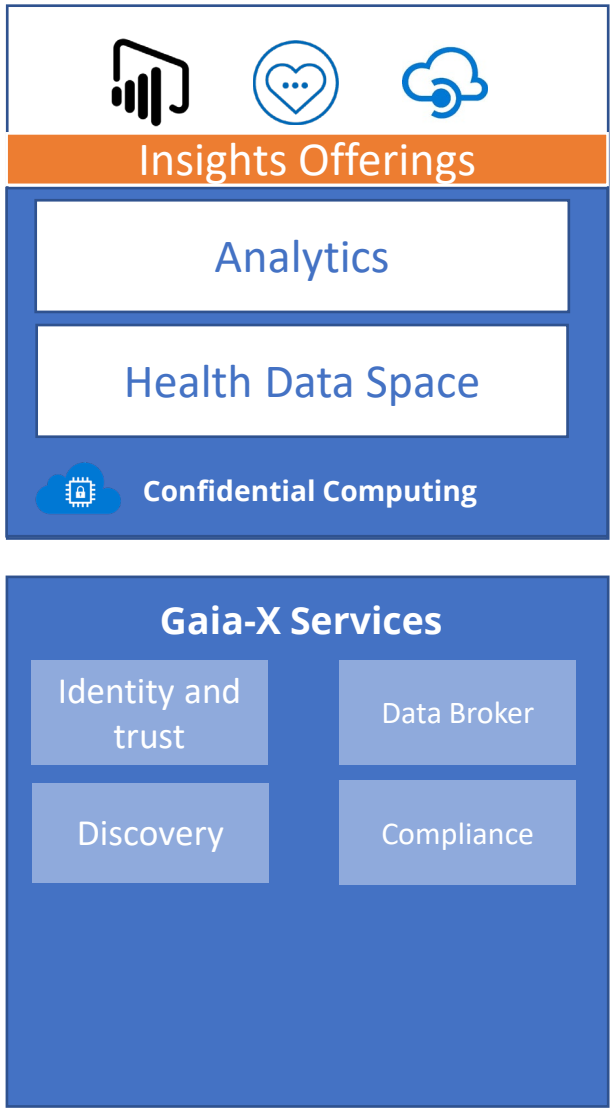
The entire ecosystem is represented in the project: some as consortium partners, some as supporting organisations, and others as potential paying users of the Carecol's outputs

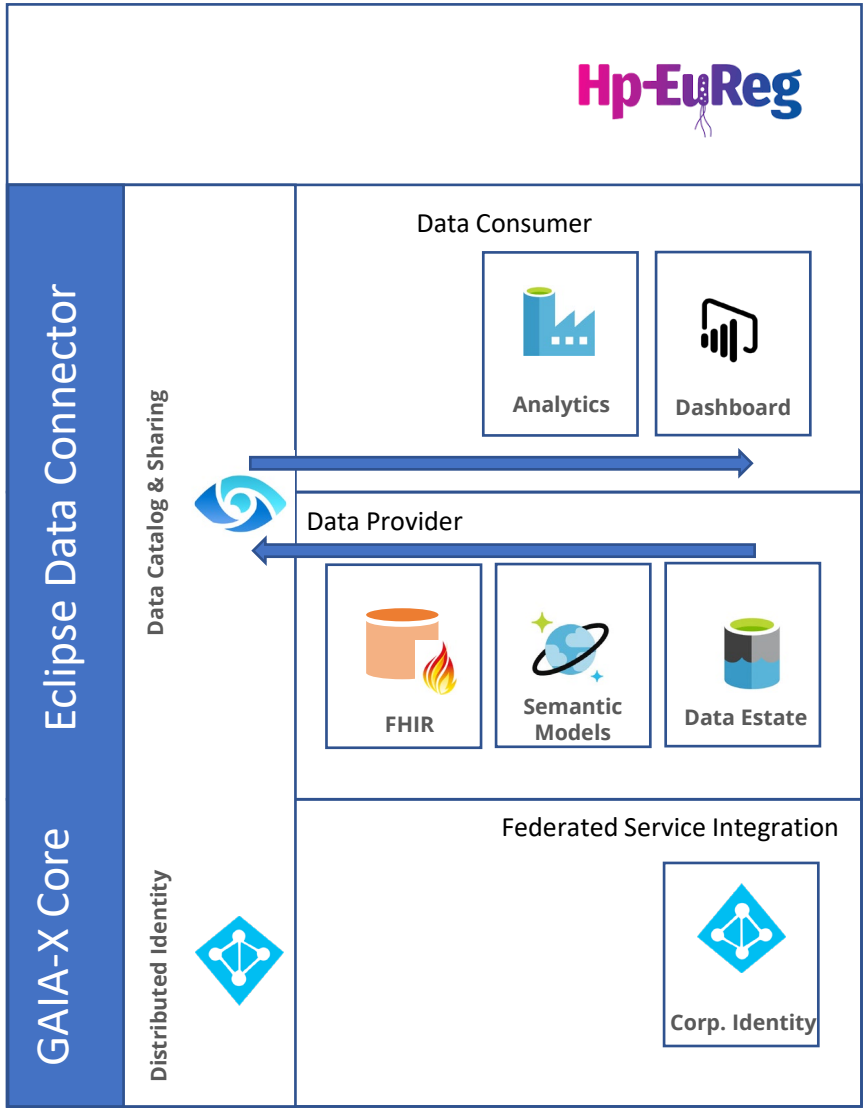


Gaia-X North Star Architecture

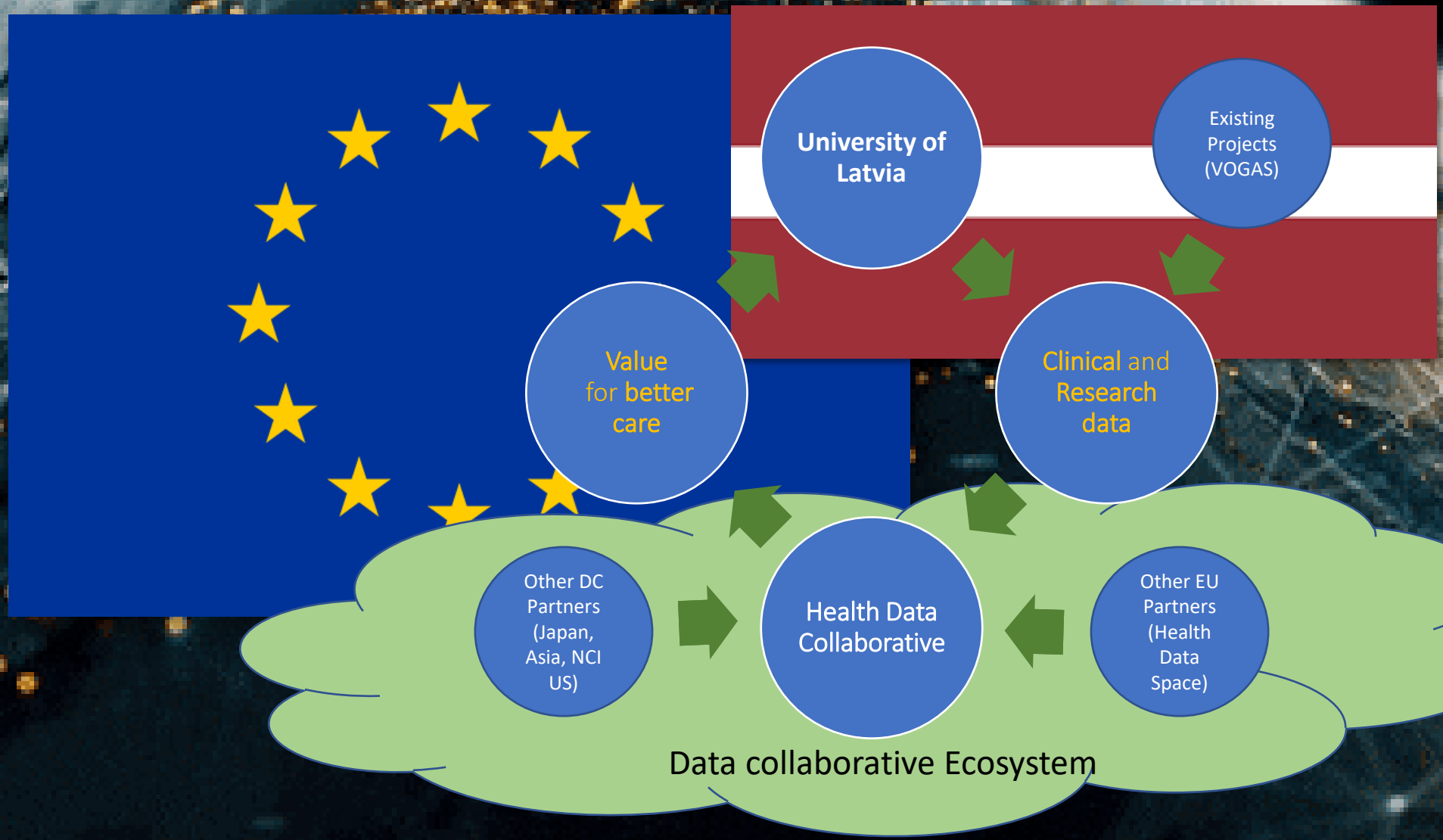
Open Standards







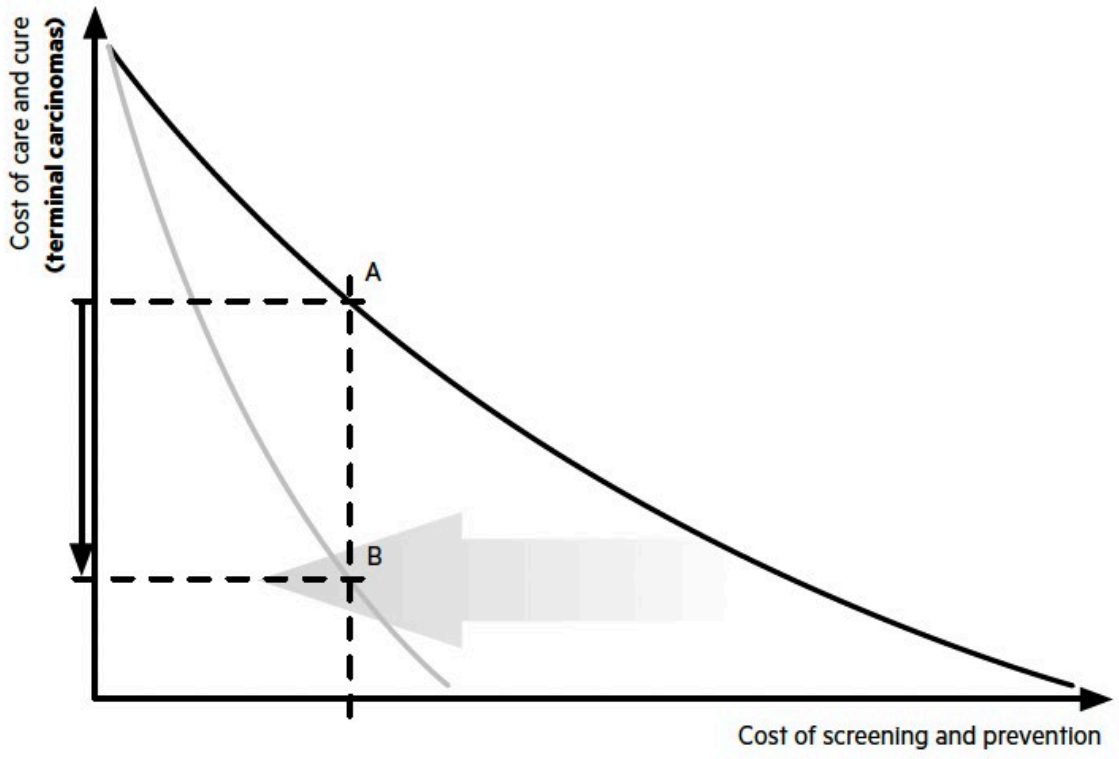
Design of the Data Collaborative have been inspired by the ODI Initiative



Carecol –Impact

Carecol helps lower the incidence of terminal gastric cancer for a given prevention and screening budget. This directly benefits health outcomes and quality of life.

Full implementation of Carecol could save up to **20 000 lives per year** in Europe with associated healthcare savings of **€ 4 bn**



The heavy black curve represents the current situation. **Carecol shifts the curve to the left** and lowers the incidence of terminal gastric cancer (from A to B in the figure) for a given screening budget, by offering the possibility of early prevention strategies.

The shape, scale and relative positions of the curves are arbitrary and only serve to illustrate the point being made

Use cases

Actual Use case map

High-Risk groups identification for *H.pylori*

Data

Multicentric randomized data of *H.pylori* eradication and pepsinogen testing

Impact

Risk Score

High Risk Groups identification

Early diagnosis of critical precancerous lesions – Endoscopy Images

Data

Endoscopies High-res Images and videos
Pathology images

Impact

Lesions
Risk Identification
Improve in-patient care

Real-time breath analysis for VOCs (Volatile organic compounds) Detection

Data

Volative Organic Compounds data from Breath Analyzers

Impact

Research Enablement

Disruptive Diagnostic approach

applicable less applicable

de.NBI Cloud: Not quite a Data Space (yet)

gaia-x

- **Harald Wagener**
  Group Lead Cloud and IT
  Center for Digital Health
  BIH@Charité

de.NBI Cloud: Overview

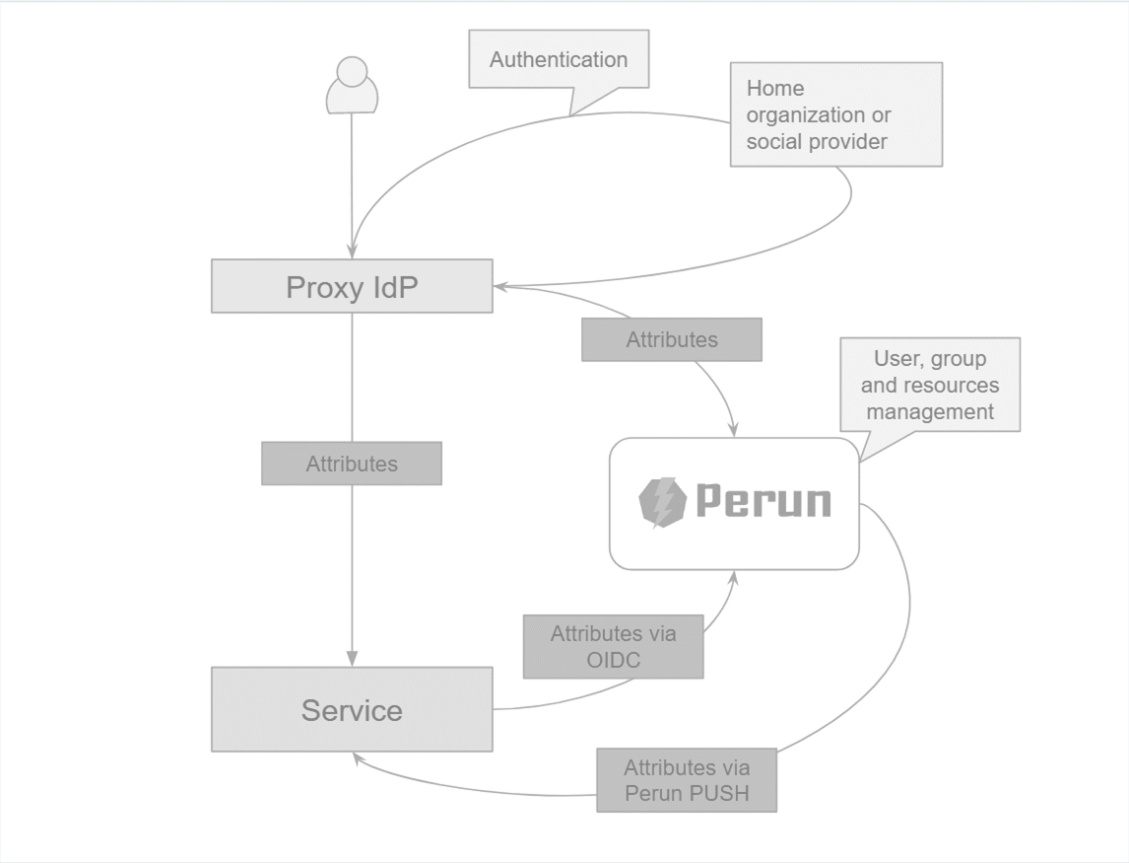
Academic Cloud Infrastructure for Bioinformatics Research

- Services, Training, Cloud Infrastructure
- Funded by BMBF
- Since 2022 led by FZJ
- Seven local cloud centers
- German ELIXIR node since 2016

de.NBI Cloud: Data Space?

Federated Cloud Infrastructure

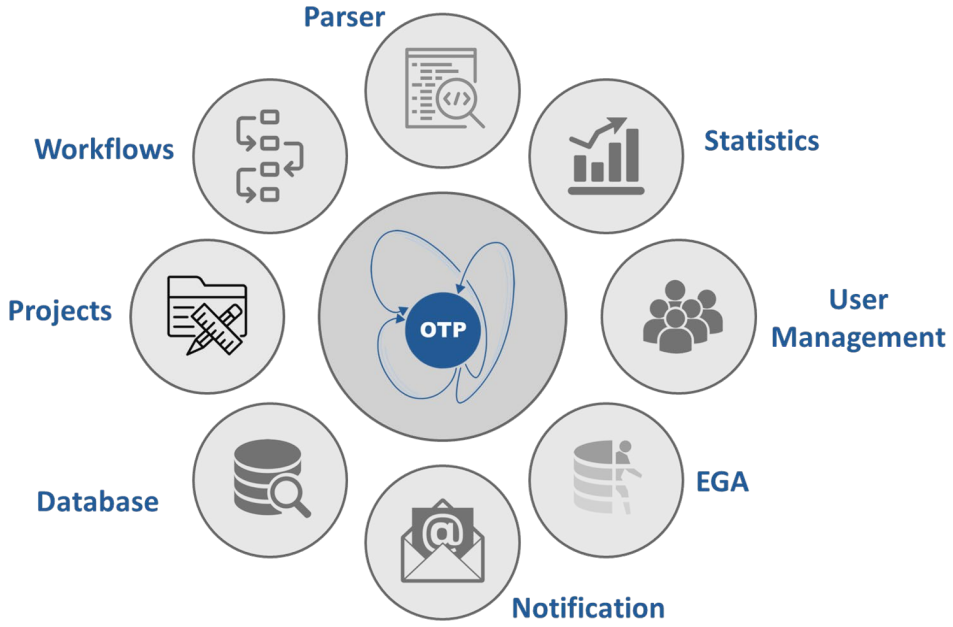
- Independent nodes
- Standardized Services
- Distributed Identity Management
- Decentralized Project Management



de.NBI Cloud Use Case: One Touch Pipeline

Platform developed at DKFZ and Charité

- processing, management, and analysis of sensitive cancer genomics data
- Big data sets (>100GB)
- Highly sensitive and protected by GDPR
- Decentralized due to nature of projects and data



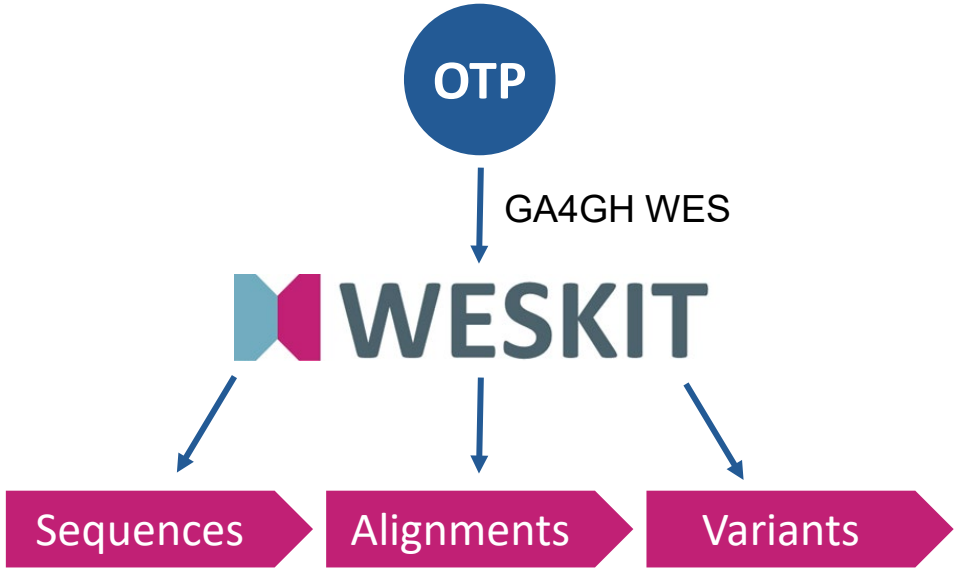
de.NBI Cloud Use Case: One Touch Pipeline

OTP: Platform developed at DKFZ and Charité

Sharing is required

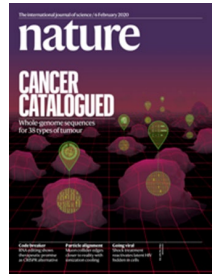
- Collaboration with experienced partners
- Combining Data from rare disease entities
- Population Studies

Solution: Compute To Data



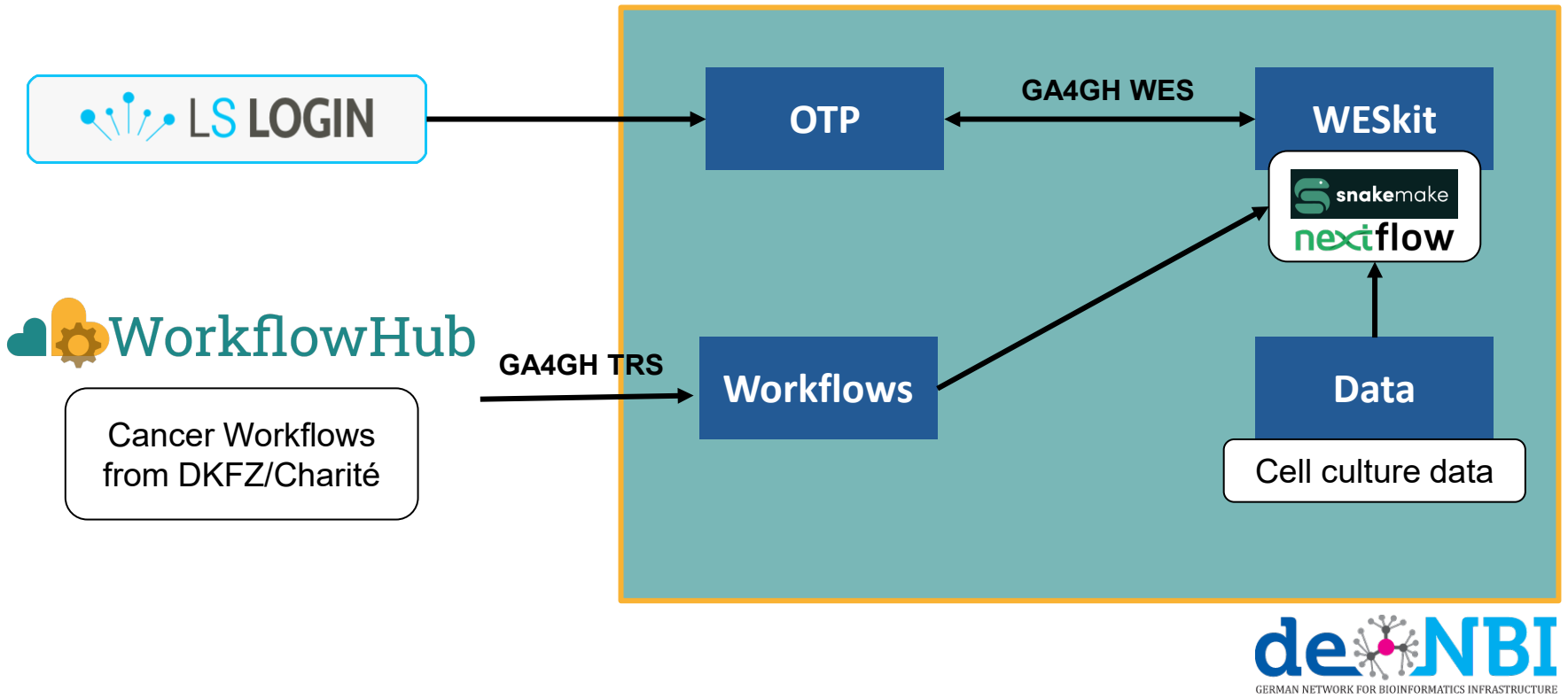
snakemake

nextflow



The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium., Campbell, P.J., Getz, G. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578,** 82–93 (2020).

de.NBI Cloud Use Case: One Touch Pipeline

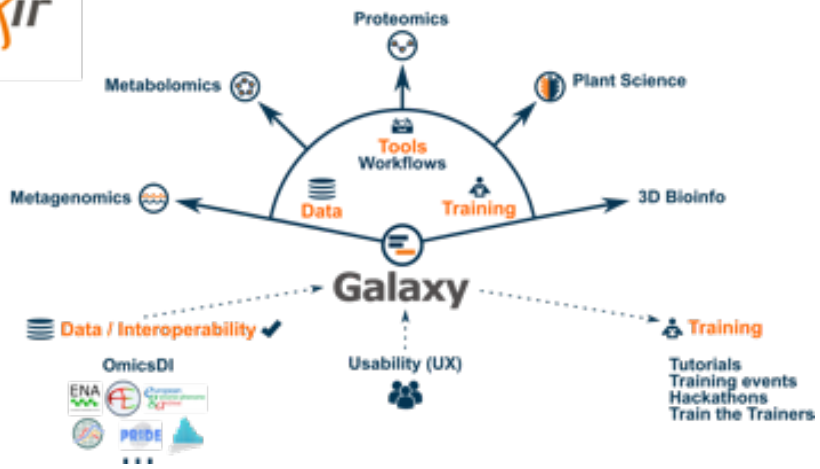


de.NBI Cloud Use Case: GALAXY

Open Source platform for FAIR data analysis

- galaxyproject.org
  - **Use established Tools**
    and combine them with workflows
  - **Run Analyses**
    in interactive environments
  - **Manage Data**
    (data sharing, publishing results, workflows, and visualizations)
  - **Ensure Reproducibility**

- usegalaxy.eu (de.NBI Cloud Freiburg)
  2020: 4.5M jobs, 46k workflow runs



de.NBI Cloud Use Case: GALAXY

Web based UI for building and running data analysis workflows

”Over 20 workflows related to the COVID-19 pandemic have been deployed in Galaxy from tools built with conda and containers and made available […]. These workflows address SARS-CoV-2 genome assembly and analysis, evolutionary analysis of viral genomes, proteomics, direct RNA sequencing and cheminformatics screening for millions of compounds that could inhibit the SARS-CoV-2 protease. Of note, analyses are regularly updated by running the workflows again as new data becomes available.”

Quote: EOSC-Life: Building a digital space for the life sciences D2.1 – Cloud implementation of exemplary workflows

galaxyproject.org usegalaxy.eu

de.NBI Cloud Challenges

- Sustainability
- Regulations for Sensitive Data Hosting and Processing
- Limits to Commercial Use

de.NBI Cloud and Gaia-X?

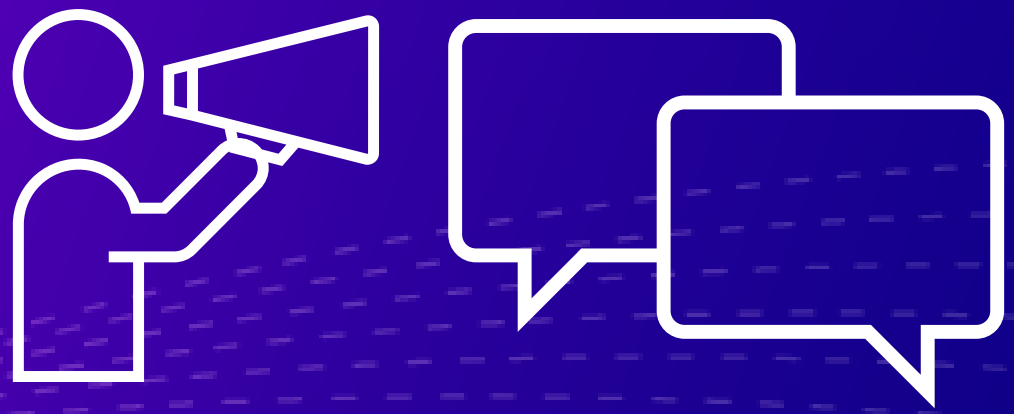
- Proven capability and experience running federated infrastructure
- Clinical Cloud Infrastructure within HEALTH-X project at Charité
- HEALTH-X: patient centered approach to data donation and consent
- Gaia-X as bridge between de.NBI Cloud and other Healthcare actors
- Open up use to partners beyond de.NBI/EOSC-Life/ELIXIR
- Showcase for ELIXIR how to integrate with Gaia-X
- Gaia-X labels for de.NBI nodes?

Thank you!

harald.wagener@bih-charite.de

Q&A

gaia-x



Session Summary

gaia-x