



# Federated Data Analysis

---

## Case Study

Miroslav Puskaric (HLRS)

15.3.2022



ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

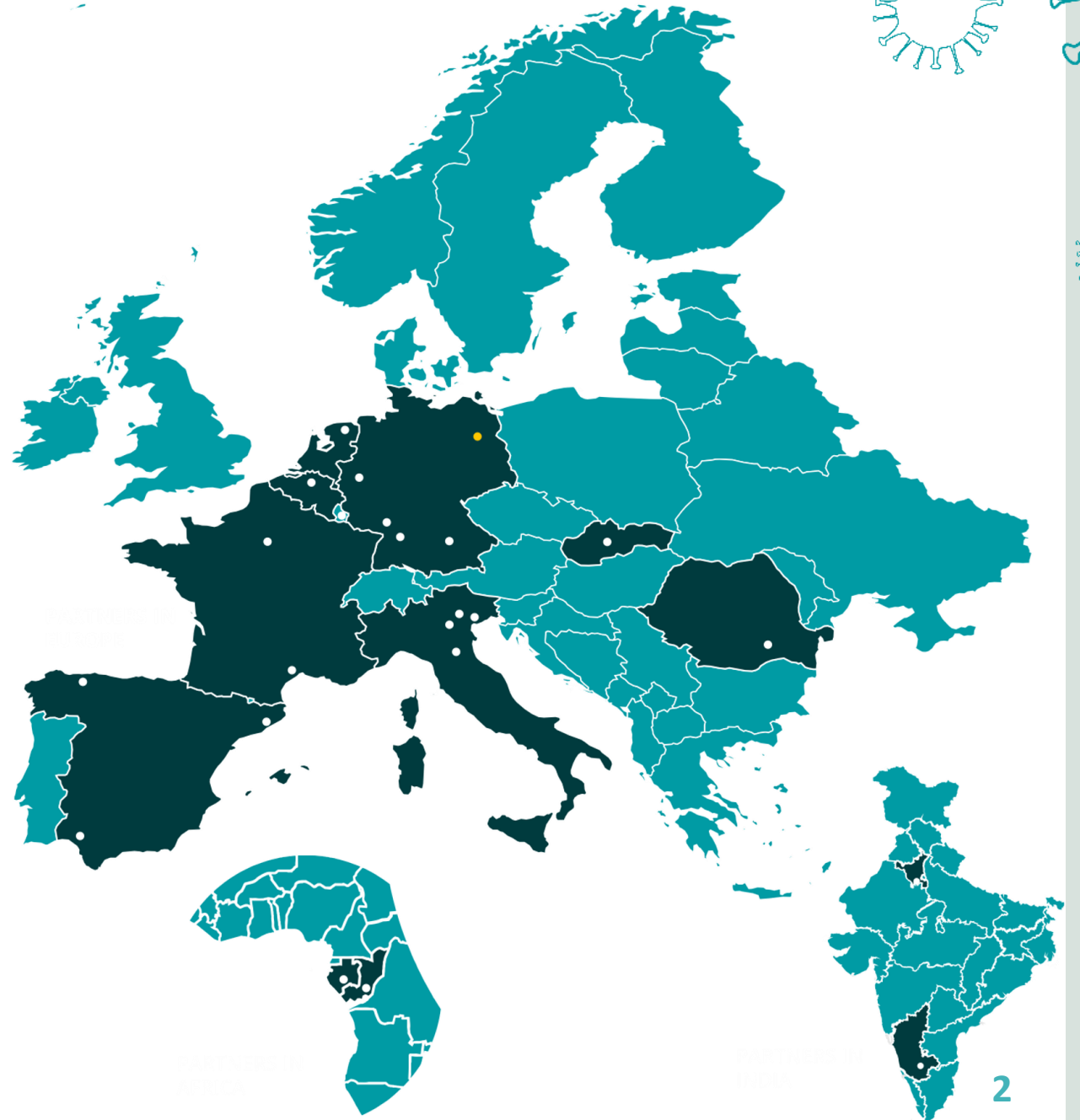


## Objective

To create a new pan-European cohort to rapidly advance the knowledge on the COVID-19 infection

## 37 partners

From 15 European / non-EU countries



# Data Management in ORCHESTRA

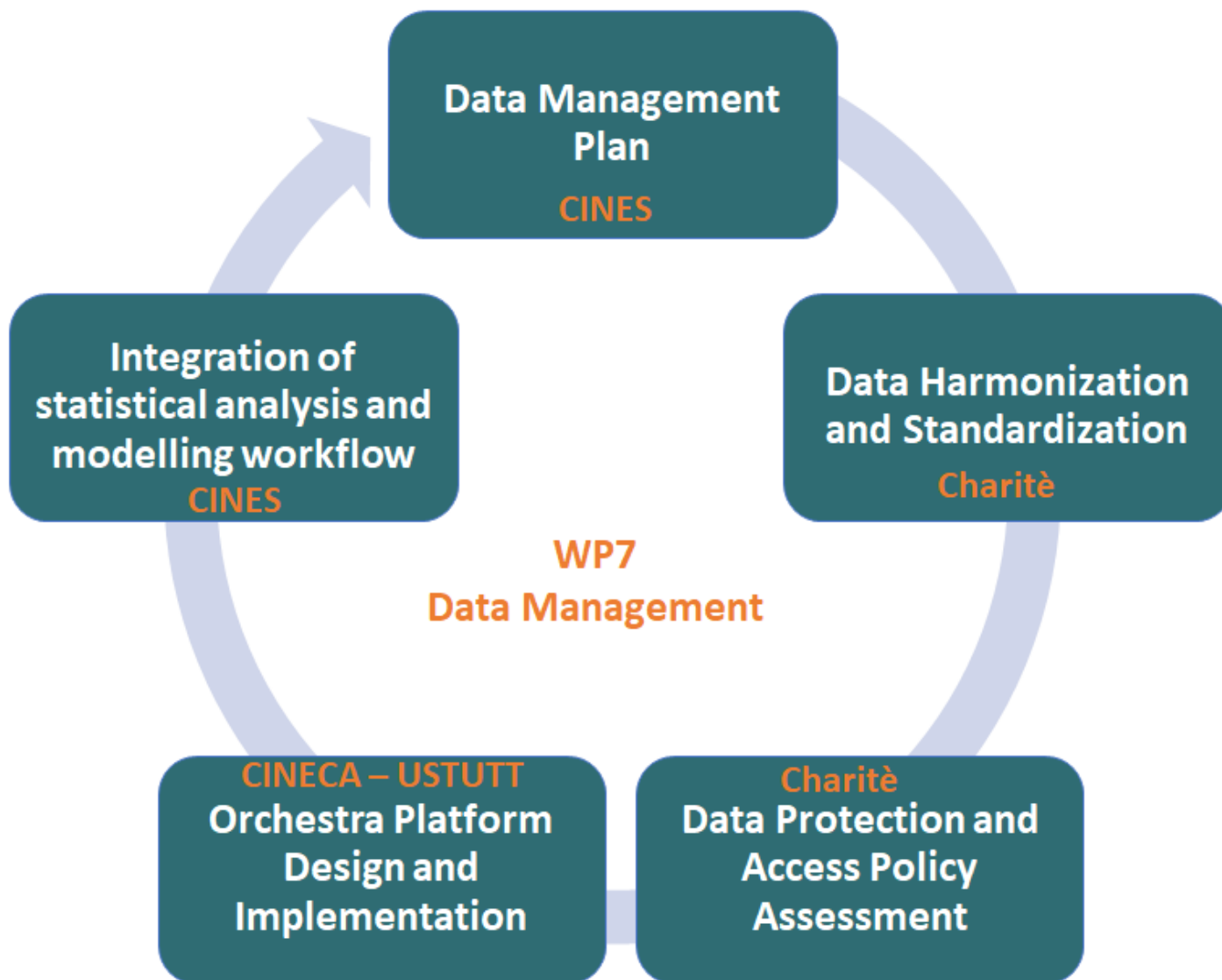
## WP7 overview

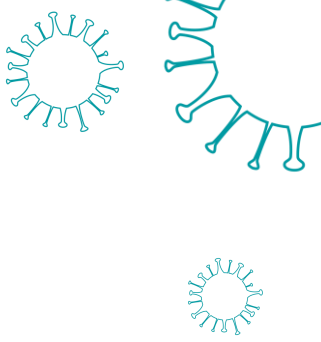
### Partners involved

#### 3 HPC centers:

- **CINECA** - Italy
- **CINES** - France
- **HLRS** - Germany

- BIH/**Charité**
- HMGU
- UNIVR





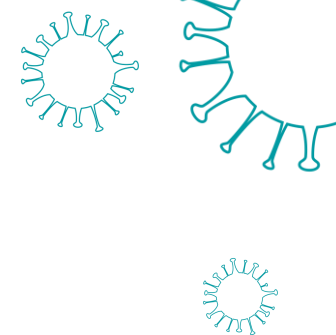
# Data Management in ORCHESTRA Challenges

- Research data (e.g. patient records) are highly sensitive, since they include information about:
  - Individual health record, such as symptoms, pre-existing conditions
  - Data from samples such as blood, nasopharyngeal swabs, ...
  - Genomic data
- Research data are distributed among various sites in different countries
  - Data might exist in different formats
  - Different legal regulations may exist
- Project activities coordination of a network of 37 partners from 15 countries

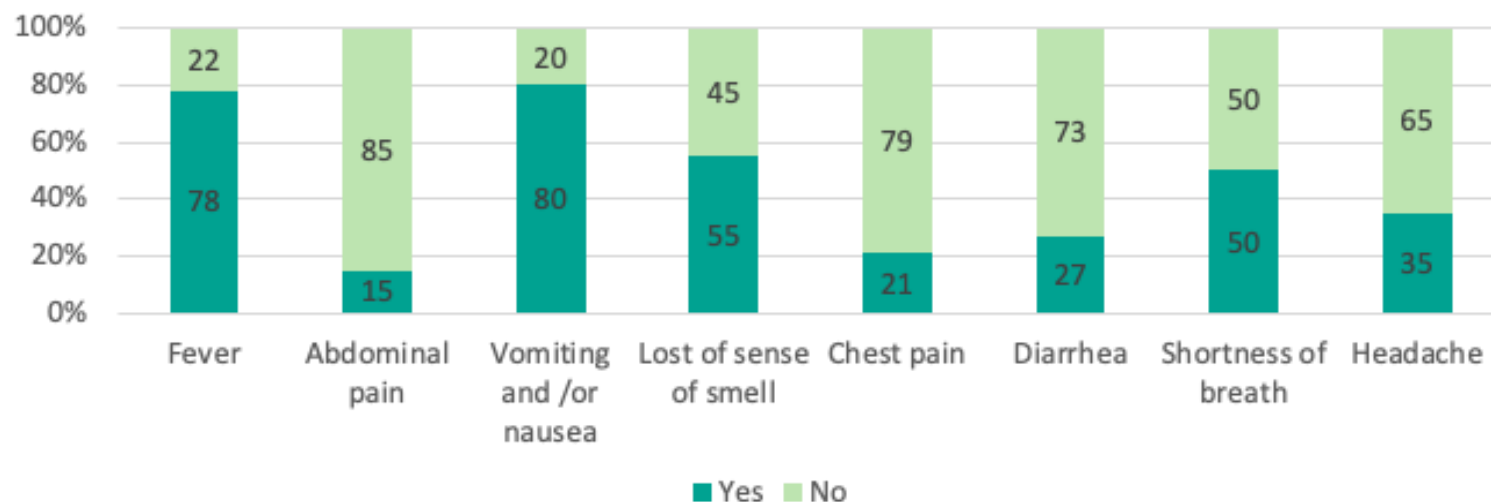
The ORCHESTRA project aims for data management that can balance the need for data sharing across borders for collective analysis and the protection of sensitive personal data.

# Data Management in ORCHESTRA

## Example of a dataset



Patient	Fever	Abdominal pain	Vomiting and/or nausea	Loss of sense of smell	Chest pain	Diarrhea	Shortness of breath	Headache
Patient 1	Yes	Yes	Yes	No	No	No	Yes	Yes
Patient 2	No	Yes	No	No	No	Yes	Yes	Yes
Patient 3	Yes	No	No	Yes	Yes	No	Yes	Yes
Patient 4	Yes	No	Yes	No	Yes	Yes	Yes	Yes
Patient 5	No	No	No	Yes	No	No	No	No

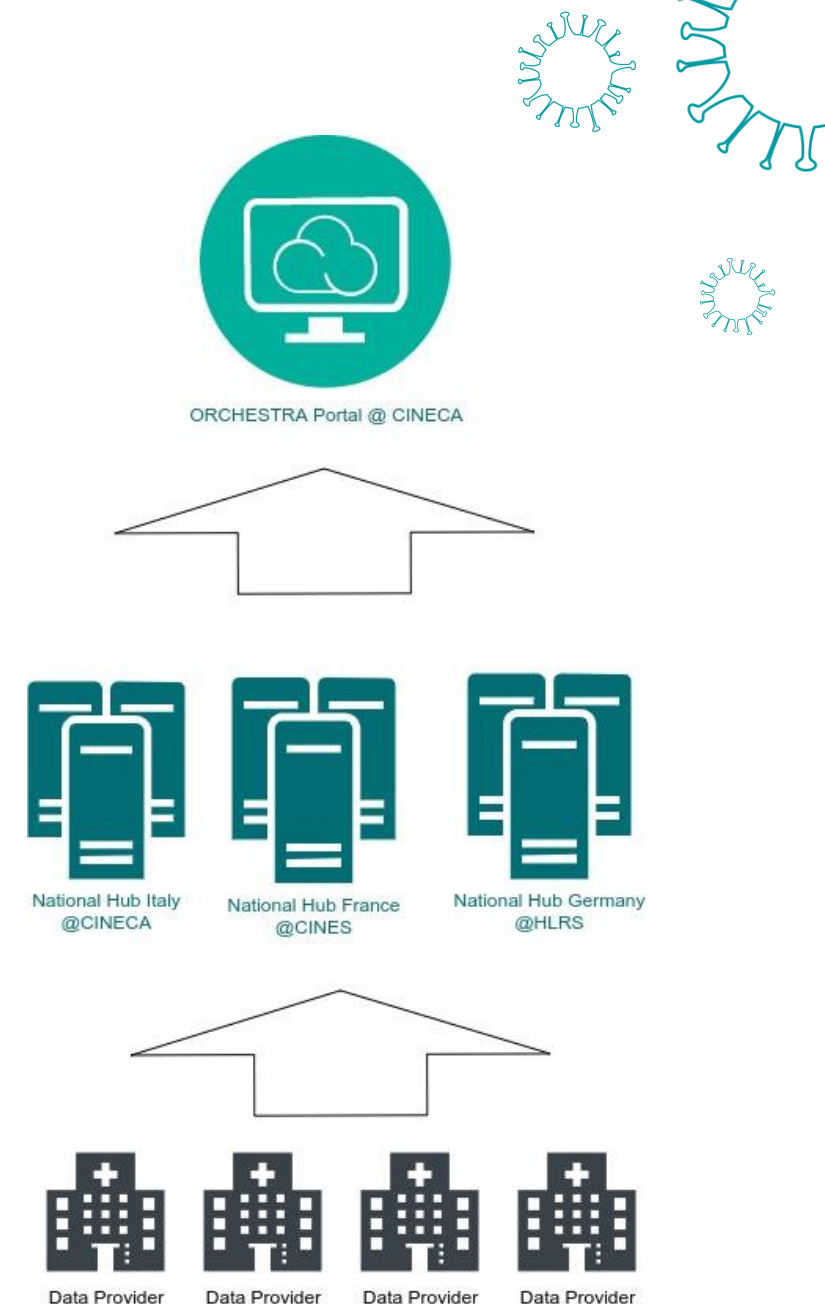


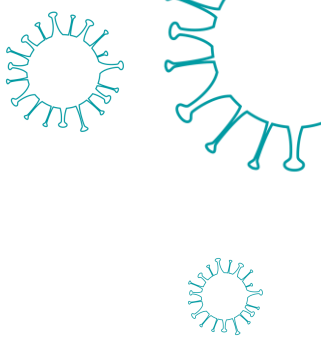
- **Cohorts:**
  - COVID-19 patients
  - General population
  - Fragile patient groups
  - Healthcare workers
- **Biobanking**
- **Genomics**
- **Virus-host interactions**

# Data Management in ORCHESTRA

## Architecture model: 3 layers

- Data Providers
  - Hold research data necessary for the studies
  - Can also securely provide data for analysis purposes
  - Example: university hospitals
- National Hubs
  - Storage brokers in Italy (CINECA), France (CINES) and Germany (HLRS)
  - Centralize and share the data on a national level
  - Secure provision of harmonized and pseudonymized/anonymized data for analysis purposes
- ORCHESTRA Portal
  - Deployed at CINECA
  - Pan-european portal for sharing aggregated data
  - Also envisioned as a central node for data analysis





# Data Management in ORCHESTRA

## FAIR data principles

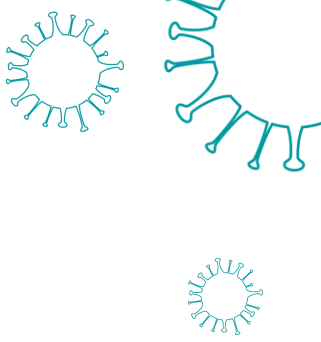
- **Findable**
  - ORCHESTRA platform repository
    - Metadata management
- **Accessible**
  - If allowed by the data protection legislation
    - Anonymized data
    - Aggregated data
    - Cohort statistics
- **Interoperable**
  - Standardized dataset for collecting cohort data
    - GECCO [1] dataset as basis for electronic case report form (eCRF)
    - HL7 FHIR [2] standard to define interoperable, machine-readable data formats
  - Semantic interoperability
    - SNOMED CT [3] collection of medical terminology
    - LOINC [4] terminology for medical laboratory observations
- **Reusable**
  - Data quality assurance
  - Information about analysis workflows

[1] <https://pubmed.ncbi.nlm.nih.gov/33349259/>

[2] <https://www.hl7.org/>

[3] <https://www.snomed.org/>

[4] <https://loinc.org/>



# Federated Data Analysis

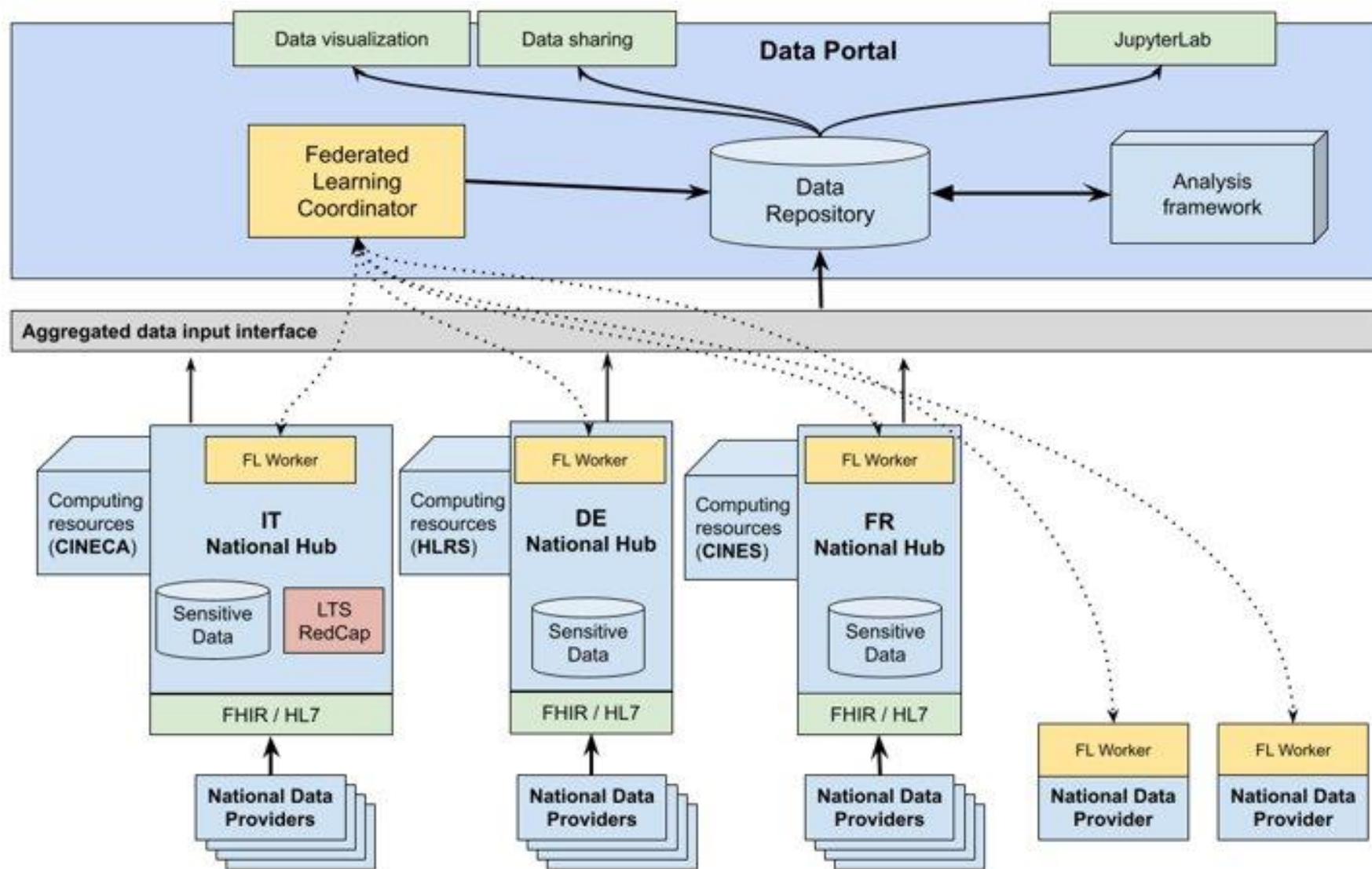
## The need for federated data analysis

- Cases where data cannot be moved to a national hub or to the ORCHESTRA portal due to
  - Legal regulations
  - Missing informed consent
  - Large data volume (e.g. genomics)
- Motivation
  - No data centralization
  - No exchange of sensitive data
  - To bring the algorithms to the data

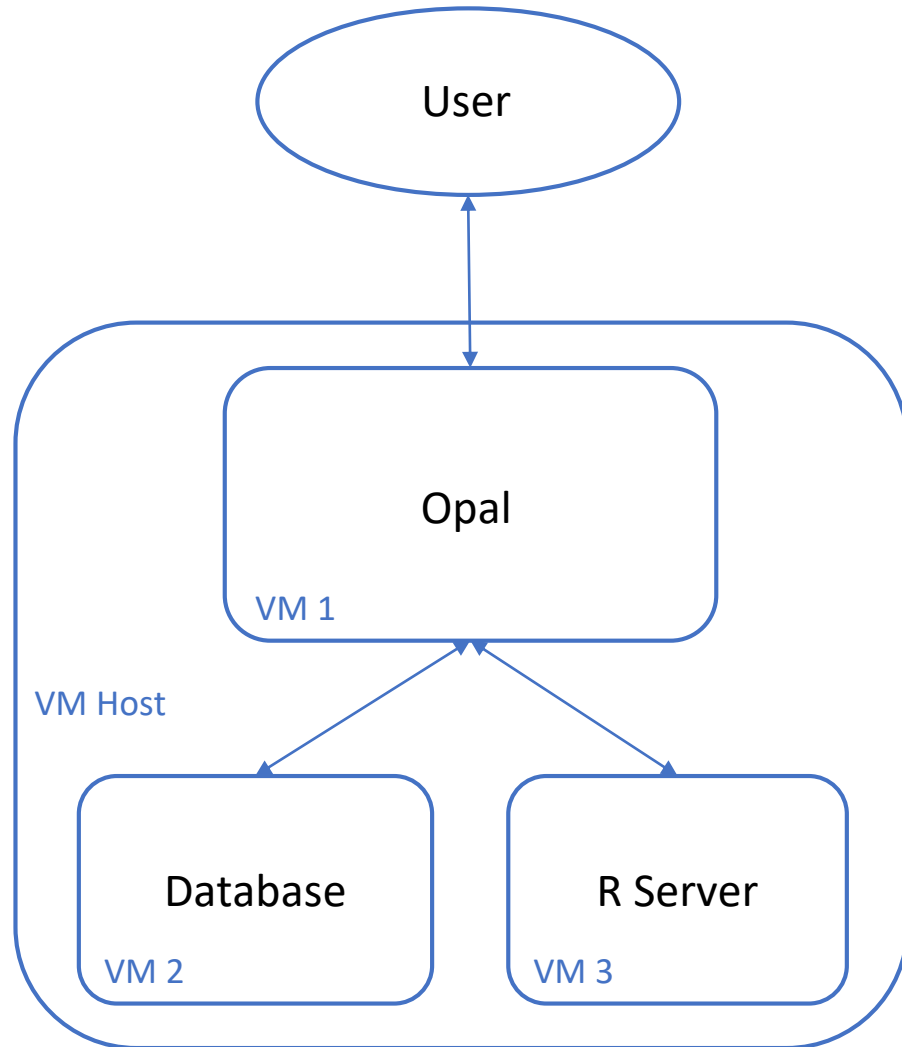
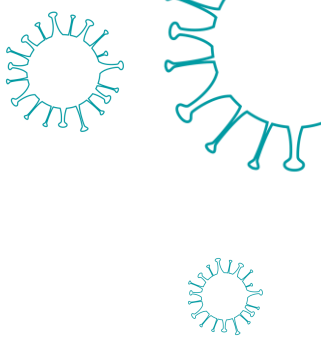
Collectively analyze data with the concept of federated data analysis



# Federated Data Analysis Architecture



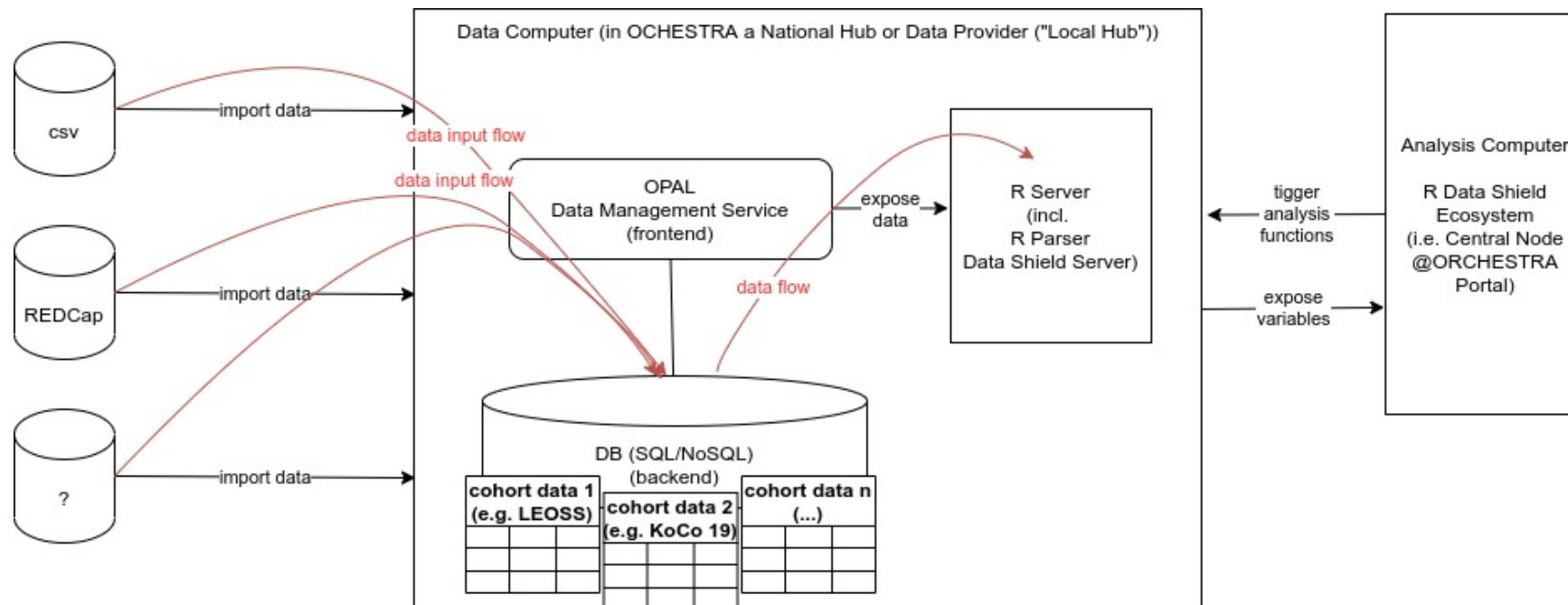
# Federated Data Analysis Architecture



- Hardware
  - Dedicated virtual machine host
  - 3 virtual machines
- Software
  - Opal - <https://www.obiba.org/pages/products/opal/>
  - MongoDB - <https://www.mongodb.com/>
  - R - <https://www.r-project.org/>
    - Rserve package - <https://www.rforge.net/Rserve/>
    - Rock API - <https://www.obiba.org/pages/products/rock/>
    - DataSHIELD - <https://www.datashield.org/>

# Federated Data Analysis Architecture

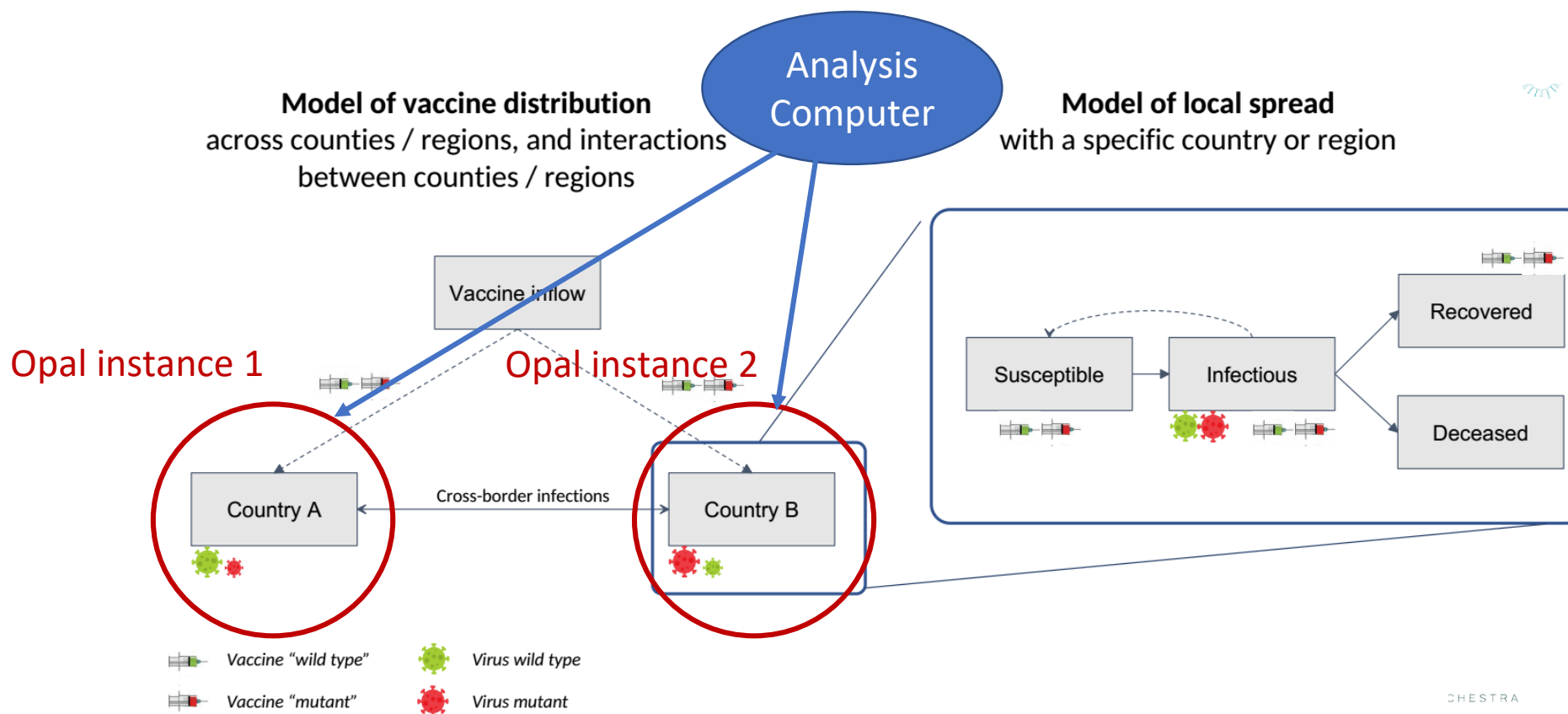
- Data from a remote source are de-identified and ingested into the Opal data warehouse
  - Opal manages database and R server
- These data are then exposed to the R server with DataSHIELD packages installed
- Only meta information such as variable names are exposed to the data analyst
- R server can then:
  - Trigger an analysis function on the data
  - Expose the non-disclosive results to the data analyst
- Data analyst can perform an analysis over multiple Opal instances

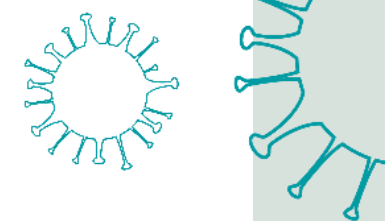


# Federated Data Analysis

## Analysis scenarios

- Intervention effects on infection process and economy
- Impact of socio-economic and environmental factors
- Impact of the pandemic on socio-economic factors
- Impact of delayed services to fragile population and relative costs
- Modelling of vaccine trials and strategies





**Thank you for your attention**

**For more information please visit**  
**<https://orchestra-cohort.eu/>**

