Federated Analysis of FAIR data

César Bernabé and
Daphne Wijnbergen

HemaFAIR

Funded by
the European Union

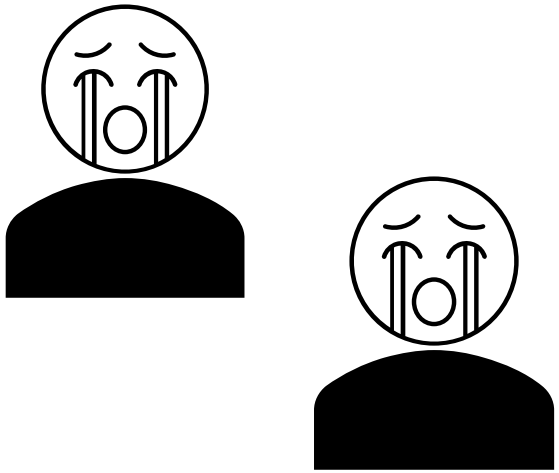
If you need to book a flight, how would you search for the best deal?

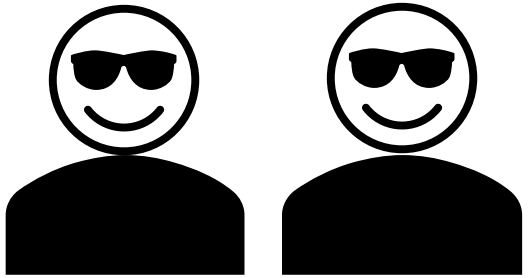
- **A)** Visit multiple airline websites separately
- **B)** Use a flight comparison tool like Skyscanner, Expedia
- **C)** Ask a travel agent
- **D)** Just book with the first airline you know

HemaFAIR

Funded by the European Union

Example





HemaFAIR

Funded by the European Union

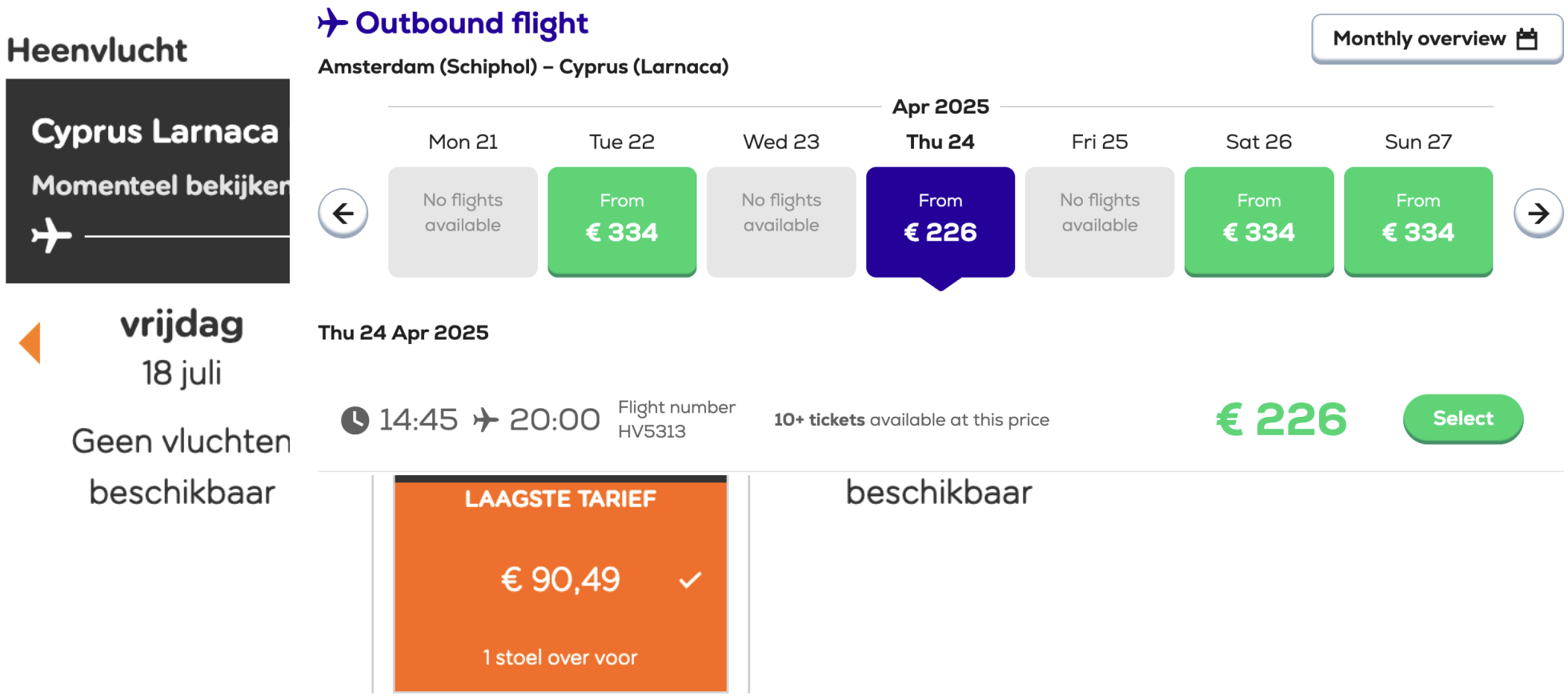
Example

- **Several airlines: AMS – LCA**
  - Easyjet
  - Transavia
  - Sky Express
  - Austrian Airways
  - Aegean
  - AirSerbia
  - …
- **Strategy one:** check one by one, download data and put it all in an Excel sheet

HemaFAIR

Funded by the European Union

Example



HemaFAIR

Funded by the European Union

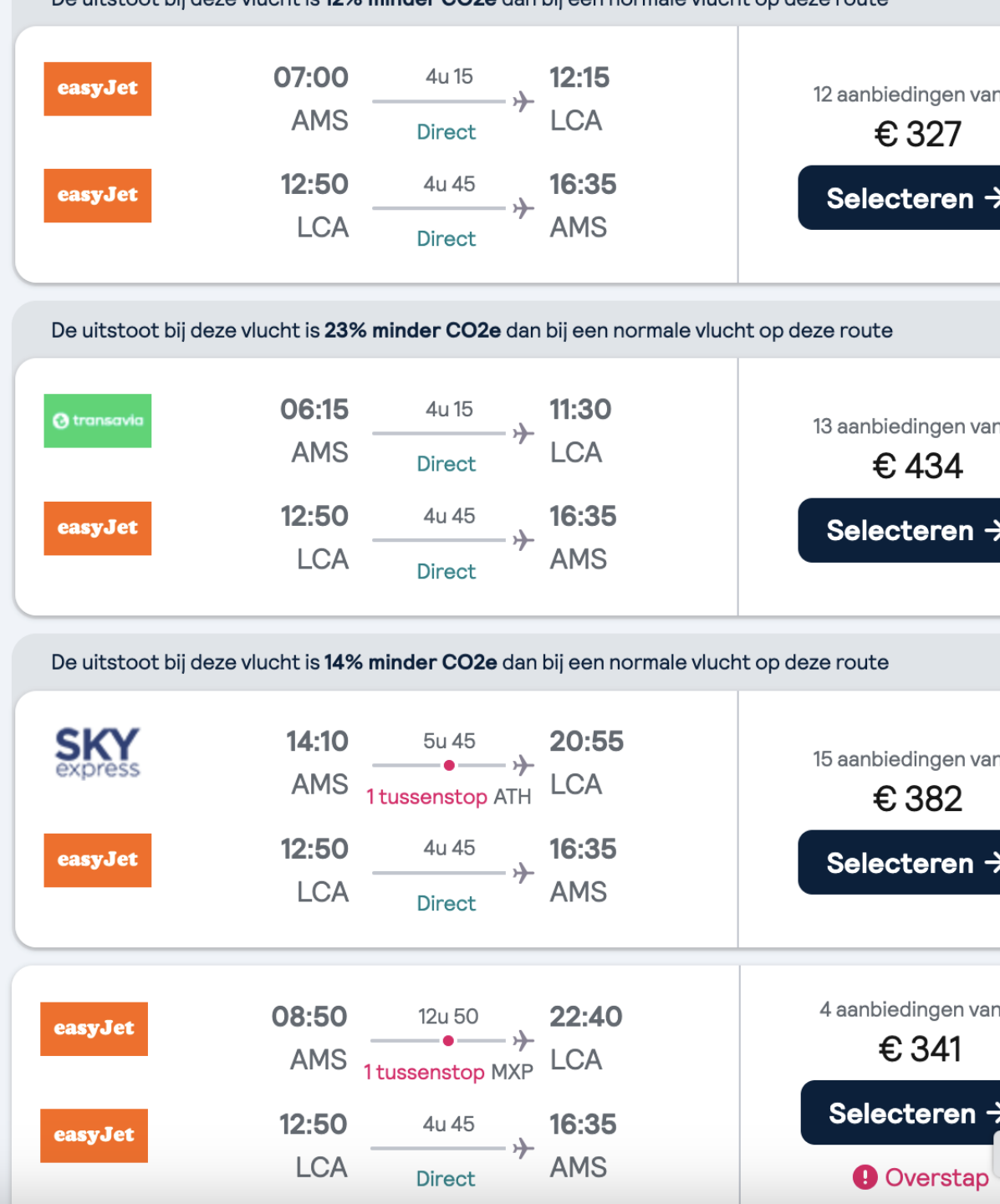
Example

- **Several airlines: AMS – LCA**
  - Easyjet
  - Transavia
  - Sky Express
  - Austrian Airways
  - Aegean
  - AirSerbia
  - …
- **Strategy one:** check one by one, download data and put it all in an Excel sheet
- **Strategy two:** use a price comparison service (which performs federated queries!)

HemaFAIR

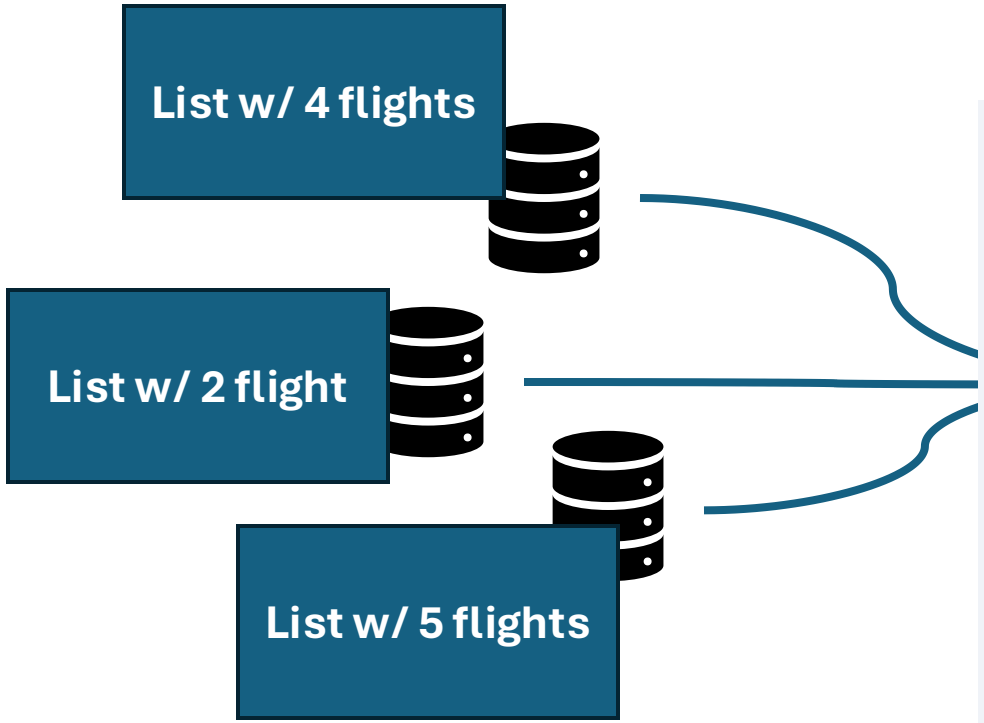
Funded by the European Union

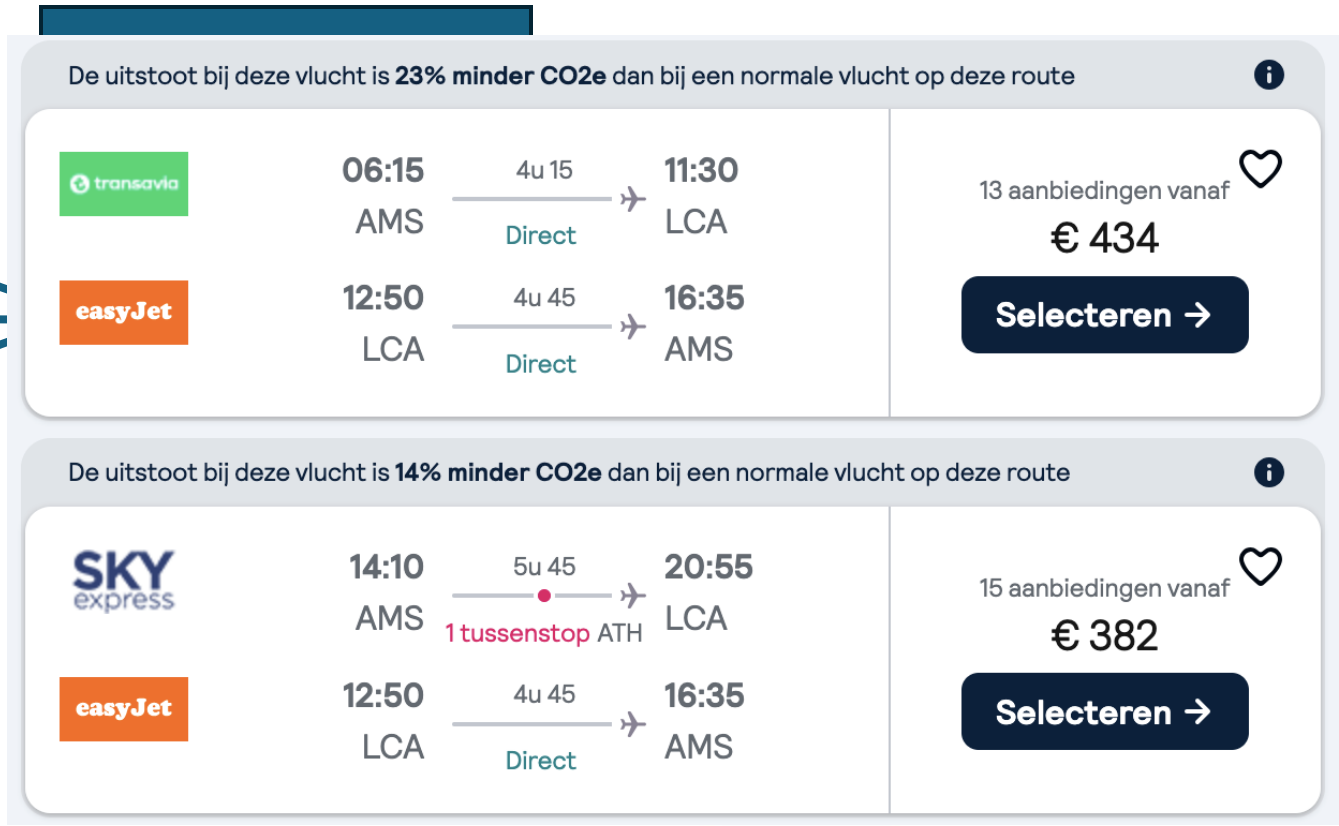
Example: Skyscanner



HemaFAIR

Example: HemaSKY





HemaFAIR

Funded by the European Union

Example

- **The point is:** for César and Daphne, the whole search as perceived as an unique search
  - Prices were retrieved in real time
  - No need to copy data to a local repository
  - Results presented harmonised and unified
  - Results combined to be more valuable (e.g. flight in with Transavia and fly out with EasyJet)

HemaFAIR

Funded by the European Union

Summary

1. A query is sent to multiple independent databases.
   - Algorithms can be sent to databases instead of queries
2. Each database processes the request and returns results.
3. Results are combined and presented in a unified format.
4. Data remains in its original location, ensuring security.

HemaFAIR

Funded by the European Union

What about rare diseases research?

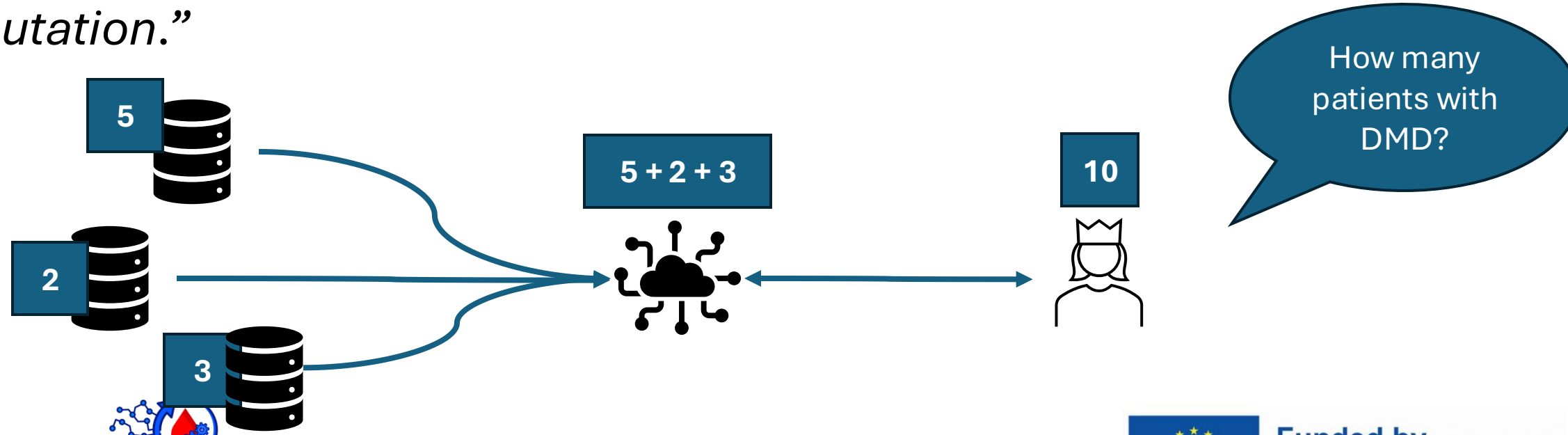
- **Patient Cohort Identification**
  “Retrieve anonymized counts of patients diagnosed with Duchenne Muscular Dystrophy (DMD) across multiple hospitals, filtering by age group and presence of a specific genetic mutation.”

HemaFAIR

Funded by the European Union

What about rare diseases research?

- **Patient Cohort Identification**
  "Retrieve anonymized counts of patients diagnosed with Duchenne Muscular Dystrophy (DMD) across multiple hospitals, filtering by age group and presence of a specific genetic mutation."



HemaFAIR

Funded by the European Union

What are the implications?

- There must be an agreement on minimal infrastructure, common data terms and legal interfaces to allow for federated querying
- For instance, to query for "**Duchenne muscular dystrophy**"
  - Different languages: Distrofia Muscular de Duchenne, Μυική Δυστροφία Duchenne
  - Different means to capture information: "X-linked muscular dystrophy with abnormal dystrophin", "Duchenne and Becker muscular dystrophy", SNOMED:240048000, ORPHA:262

HemaFAIR

Funded by the European Union

Does FAIR data enable federated?

- **FAIR Data** facilitates federated analysis, as:
  - Relevant datasets are **found** by machines
  - **Access conditions** are explicit
    - Machines now if they can send and receive queries
  - Results can be **interoperated** as data is harmonised
  - Results are presented using relevant formats for **reuse**

HemaFAIR

Funded by the European Union

Part II

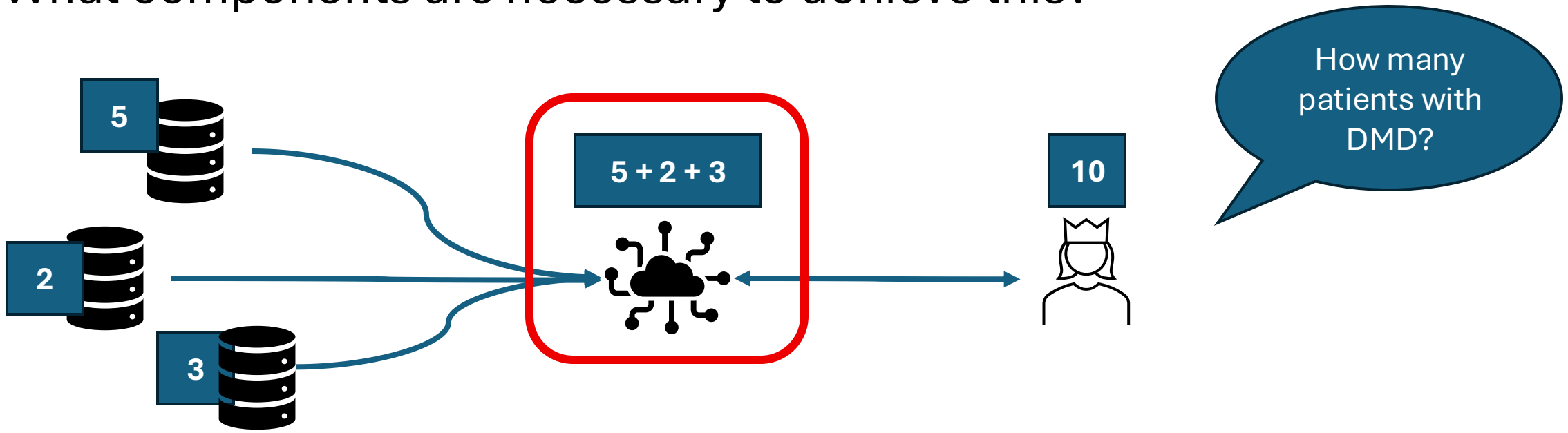
Infrastructure for federated analysis

HemaFAIR

Funded by
the European Union

Infrastructure for federated analysis

- What is happening here?
- What components are necessary to achieve this?

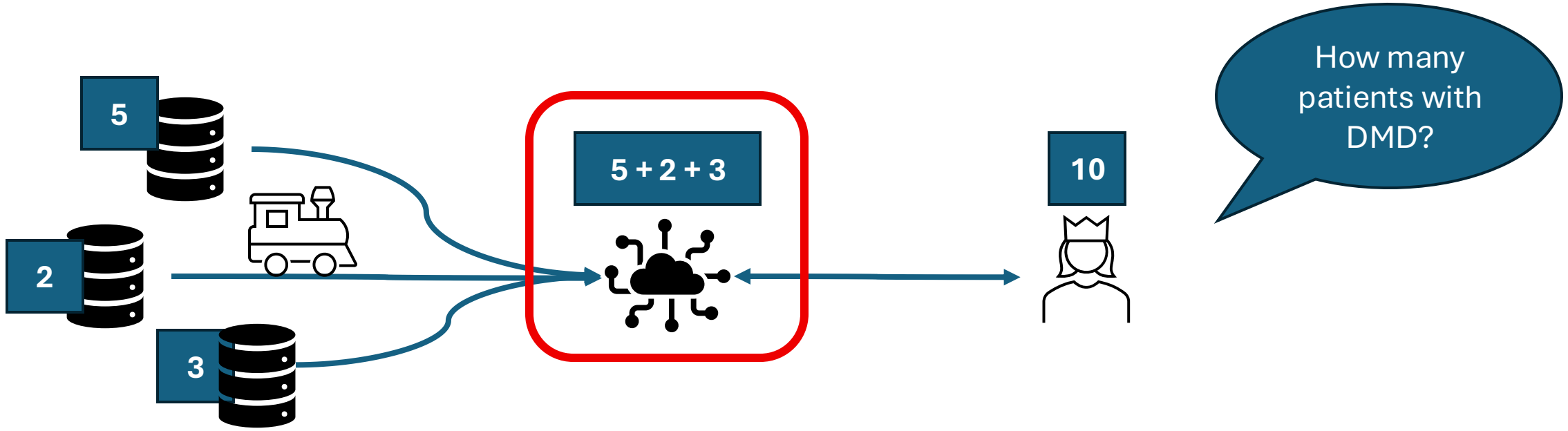


HemaFAIR

Funded by the European Union

Data Train Analogy

- Questions and answers being moved around will be represented as trains



HemaFAIR

Funded by the European Union

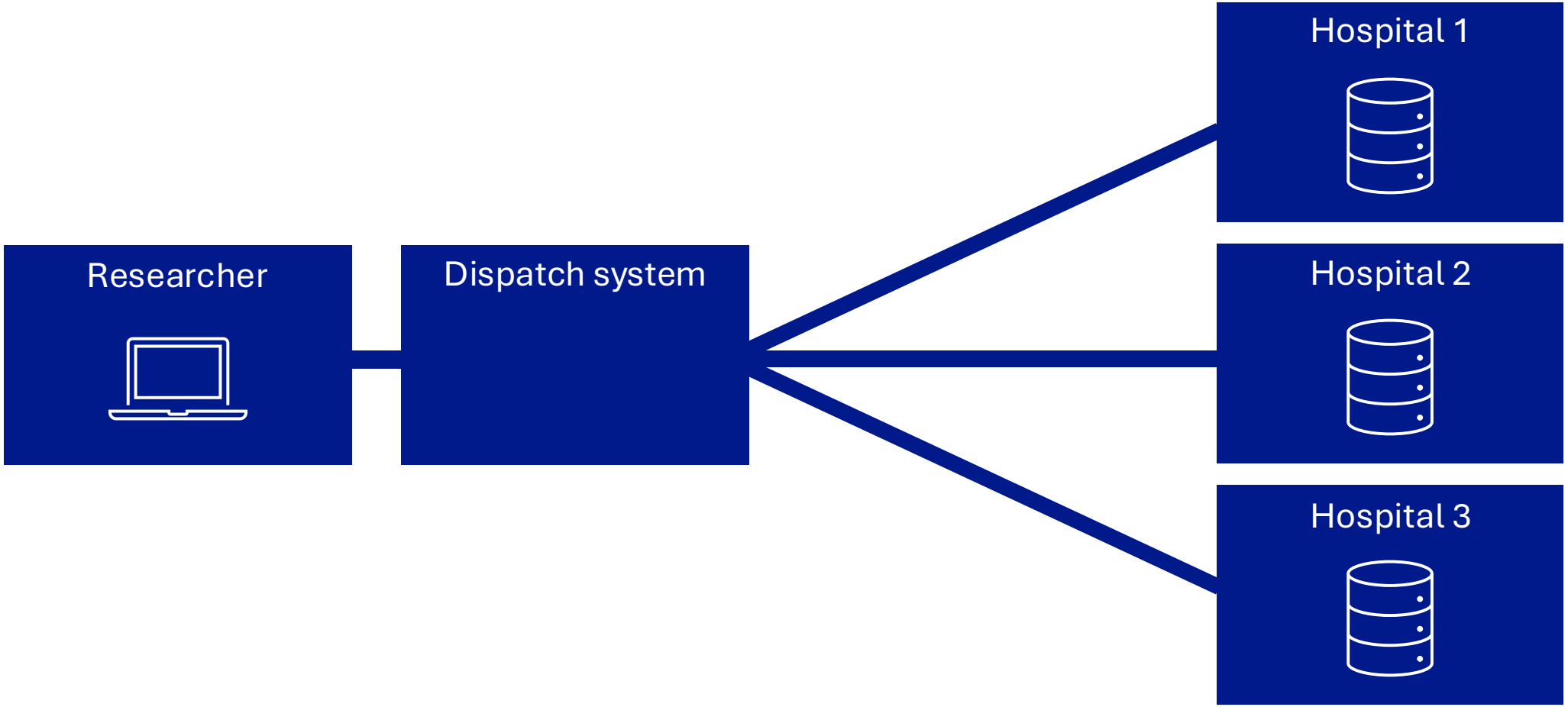
Starting point

- Three hospitals have data about patients with DMD
- We want to ask the question: how many patients are there with DMD?

HemaFAIR

Funded by the European Union

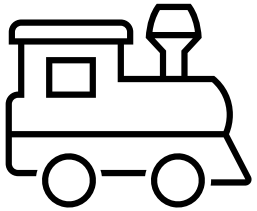
Starting point



HemaFAIR

Funded by the European Union

Trains

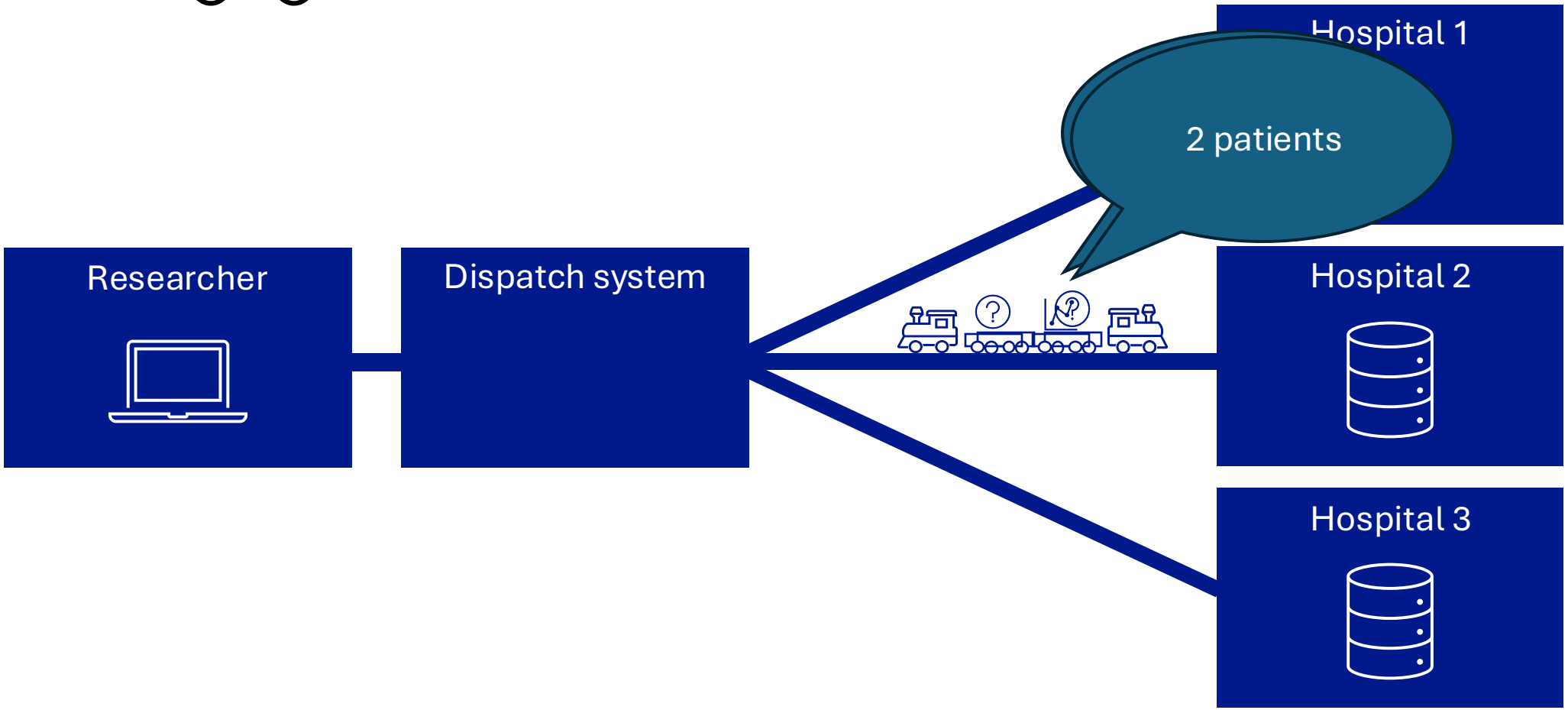


- We can use the analogy of trains
- These trains carry a question to a hospital, and carry a result back to the researcher

HemaFAIR

Funded by the European Union

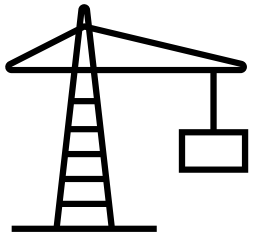
Trains



HemaFAIR

Funded by the European Union

Interaction mechanism

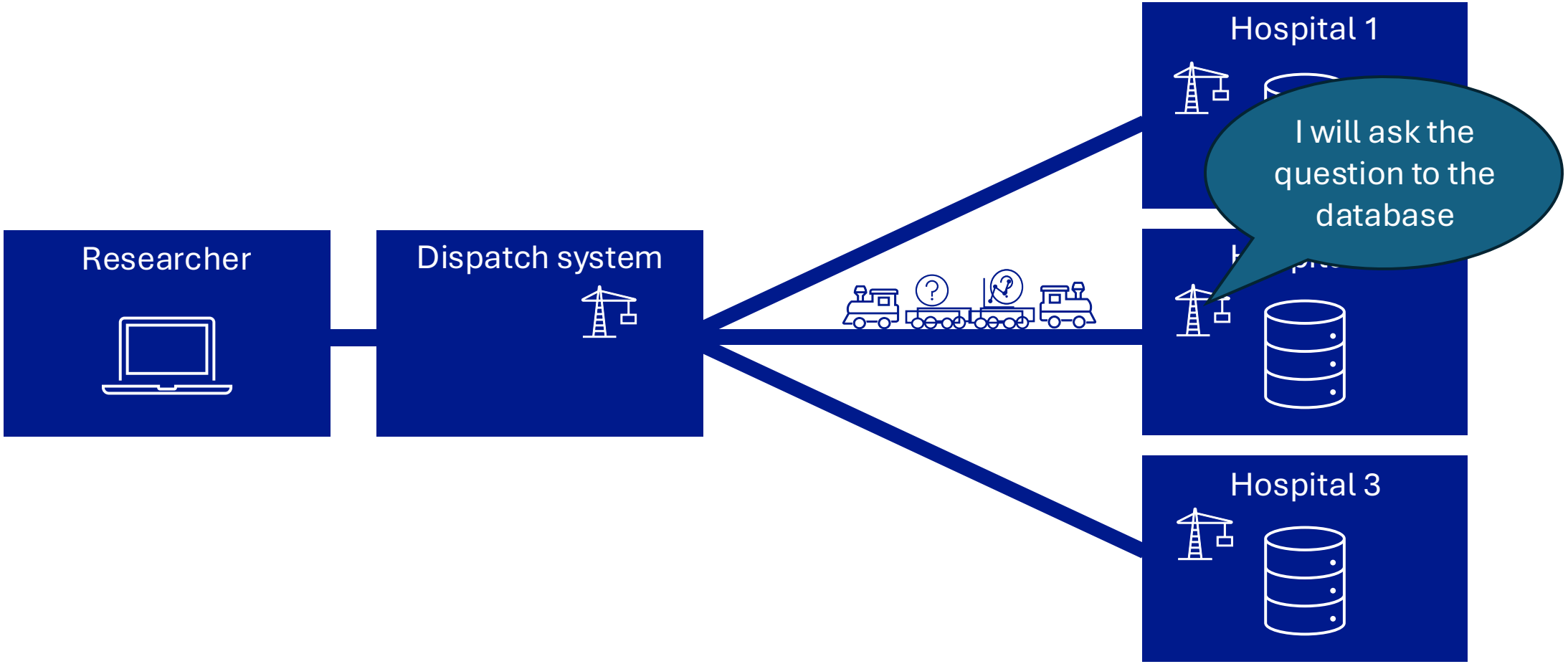


- A train can’t connect to a database directly
- An interaction mechanism should exist between the train and the data
- This mechanisms needs to perform several actions:
  - Obtain the question from the train
  - Run this question on the database/file
  - Return the answer to the train
  - And more..

HemaFAIR

Funded by the European Union

Interaction mechanism



HemaFAIR

Funded by the European Union

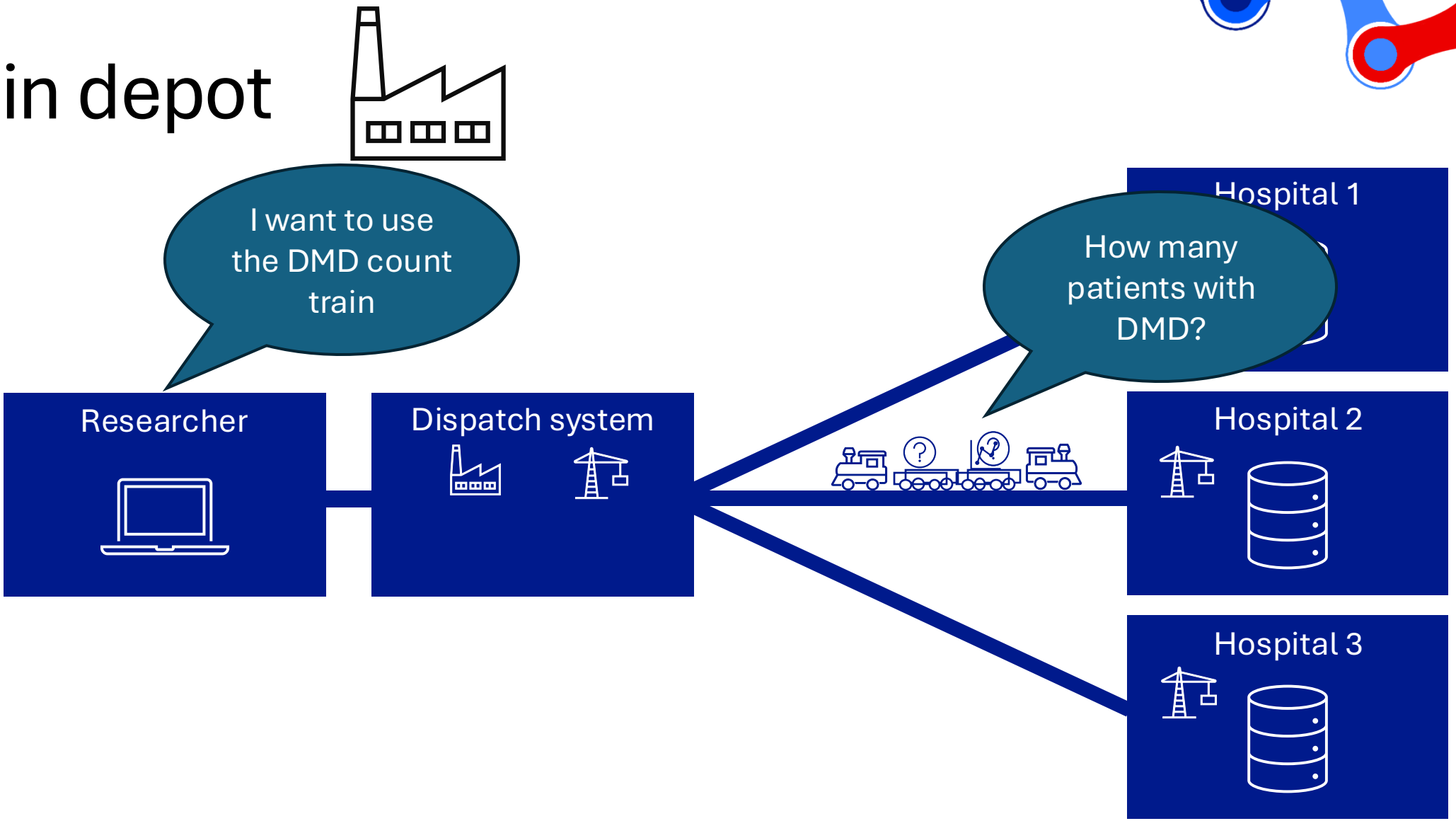
Train depot

- Trains with algorithms that can be send out should be stored somewhere

HemaFAIR

Funded by the European Union

Train depot



HemaFAIR

Funded by the European Union

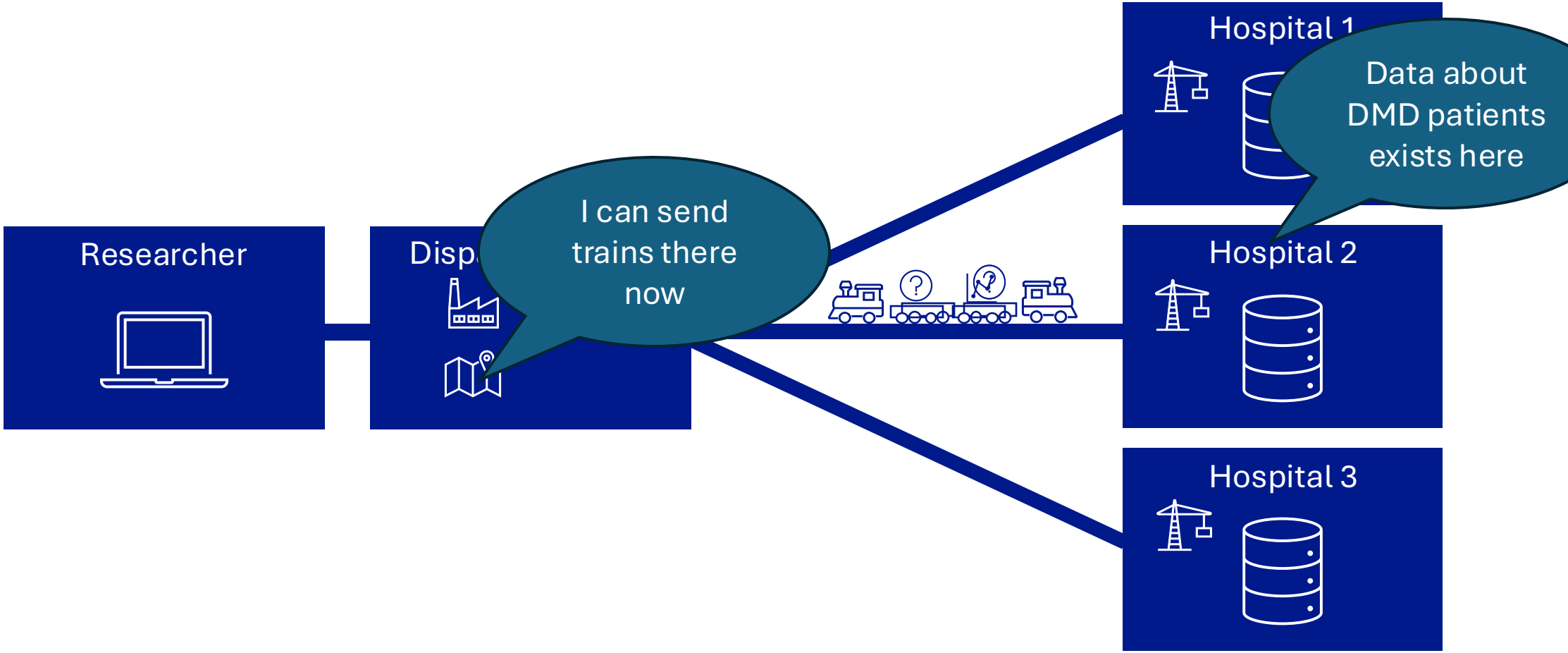
Railway map

- The trains need to know which hospitals can be visited
  - Know that the data exists.
  - Know that it can ask a question about the data
- This is one part where FAIR is important, as FAIR Data Points can point towards the data and describe it with metadata

HemaFAIR

Funded by the European Union

Railway map



HemaFAIR

Funded by the European Union

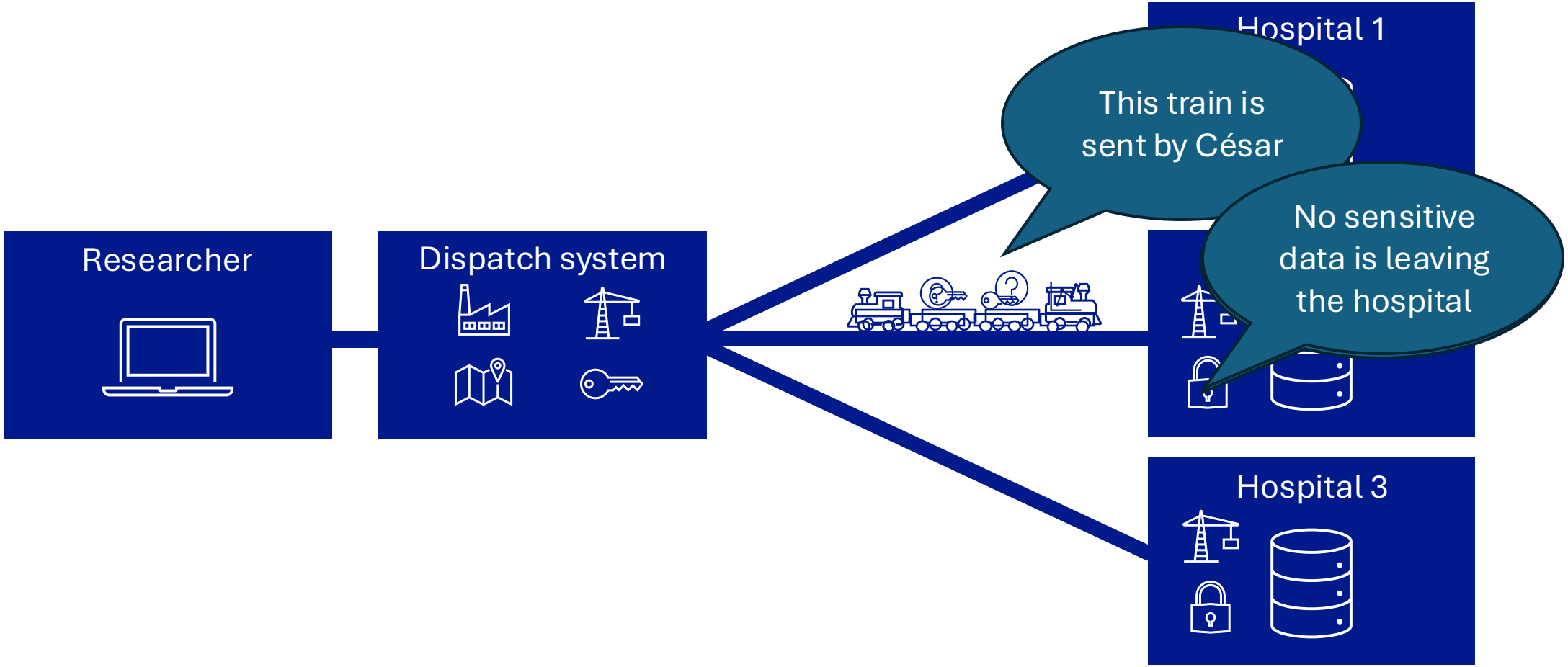
Authentication and validation

- Privacy is the main concern, which is why we do federated analysis
  - The hospital should verify the identity of a researcher
  - The hospital should verify that no sensitive data leaves the hospital

HemaFAIR

Funded by the European Union

Authentication and validation

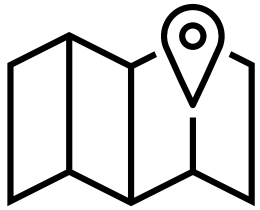


HemaFAIR

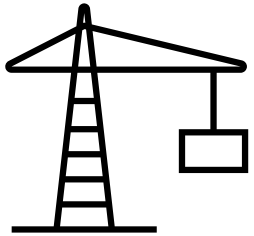
Funded by the European Union

What can FAIR metadata do for federated analysis?

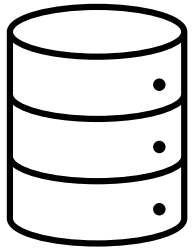
- Advertise the data
  - E.g. There is data for x here
  - E.g. This data has y access conditions



- Explain to the hospital how it can interact with trains
  - E.g. The train contains a SPARQL/FHIR/SQL query
  - E.g. The train needs x GB of memory to run



- Enable harmonization of data so that the same train works for multiple hospitals



HemaFAIR

Funded by the European Union

Conclusion

- Several components are needed for federated analysis
- The FAIR principles can enable several of these components

Analysist	Dispatch system	Hospital
• User Interface to dispatch system	• Interaction mechanism • Train depot • Railway map • Authorization system	• Data storage • Data interaction mechanism • Validation and authorization mechanism

HemaFAIR

Funded by the European Union