Data and Model objects within the six-dimensional VHT framework

Marco Viceconti Alma Mater Studiorum - University of Bologna

Breakout session 2 – 16-5-2023

EDITH

EDITH is a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement n° 101083771.

The VHT is a knowledge space

- The VHT is a systematic digital representation of the available knowledge on the human pathophysiology
- Knowledge can be represented in many way: an effective one is to use *factual statements* and *causal relations* between them
- In the VHT factual statements are captured into **data objects**, and causal relations into **model objects**

Truth in a Bayesian sense

- Any knowledge representation must deal with the issue of truth content; how reliable are the statements and the relations that compose my knowledge?
- In the VHT, we consider the truth content of some knowledge within a Bayesian inference framework, so each statement or relation has a certain probability of being true

The dominant axes of human physiology

- We can organise the knowledge about human pathophysiology in many ways, but some information axes provide particularly powerful organisational drives:
  - Space
  - Time
  - Individuality

The anatomical space

- The space is that of the human body as represented in anatomy. But where descriptive anatomy traditionally provides a qualitative, descriptive (semantic) representation of the spatial organisation of the body, we need a quantitative, universal representation of the anatomical space.
- This anatomical space is the average body geometry of all human beings; the average human body.
- For each type of data, which **transformation functions** that map the data from the individual anatomical space to this average anatomical space we need?

The environment

- The halo around the body represents the continuous mass and energy exchange between the body and the environment.
- What do we call these exchanges?

The time axis

- The individual pathophysiology changes during the lifespan
- We can imagine a time axis that ranges from zero (birth) to one (death)
- How do we **scale the date** a dataset was collected to this axis?

Individuality

- The human pathophysiology varies widely between individuals
- On the one hand, the VHT should contain quantitative knowledge about the pathophysiology of many individuals
- On the other, in several situations, such knowledge needs to be represented (and in some cases is provided) only as an average of a cluster of individuals
- How do we map such **clustering rules**?

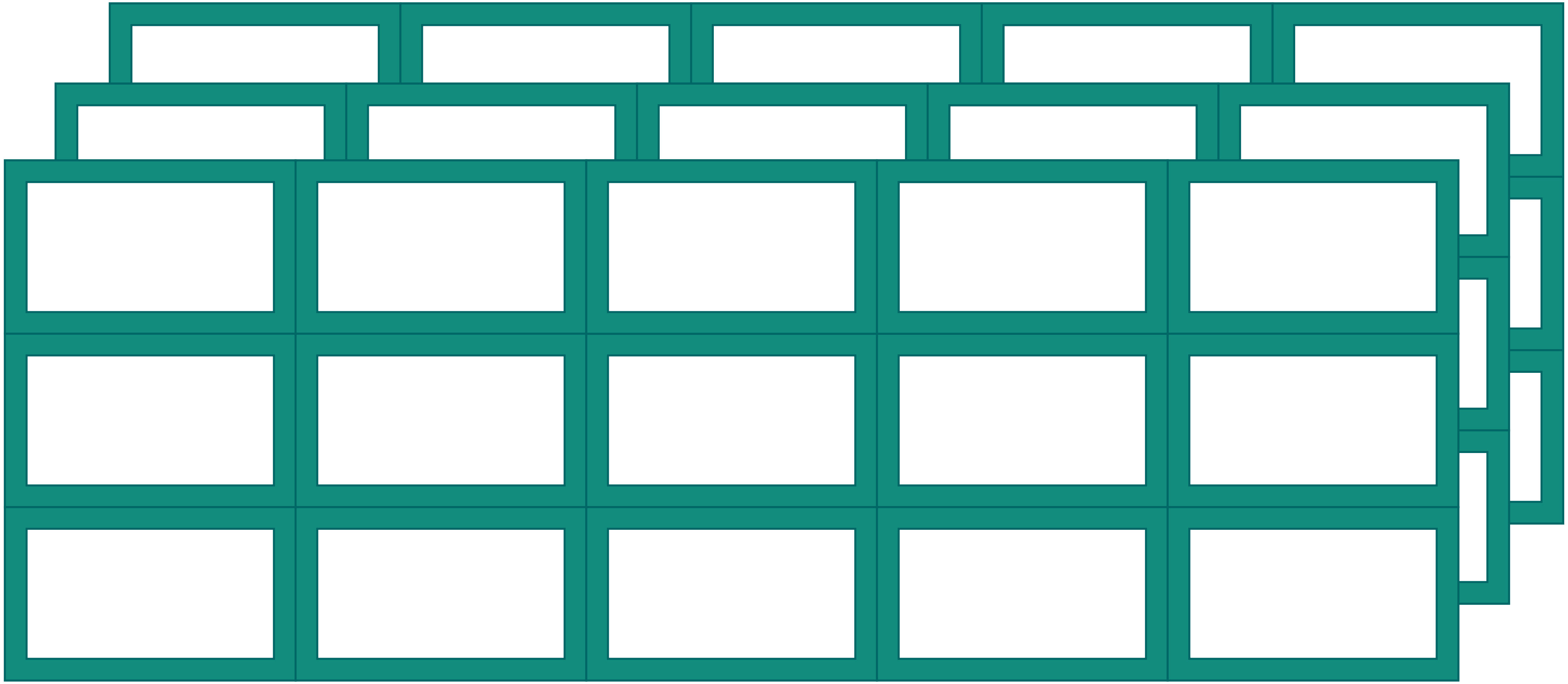
The diseased state /1

- “Health is an adaptive state unique to each person. This subjective state must be distinguished from the objective state of disease. The experience of health and illness (or poor health) can occur both in the absence and presence of objective disease. Given that the subjective experience of health, as well as the finding of objective disease in the community, follow a Pareto distribution, the following questions arise: What are the processes that allow the emergence of four observable states—(1) subjective health in the absence of objective disease, (2) subjective health in the presence of objective disease, (3) illness in the absence of objective disease, and (4) illness in the presence of objective disease?” Stumberg *et al*, 2019.

The diseased state /2

- We can represent the human body as a closed system in a state defined at each instant by the values assumed by its state variables.
- Diagnosis, the identification of a disease, is not only characterised by symptoms (unusual combinations of values for some state variables) but also by some causal relationships.
- Syndromes are clusters of symptoms that might be associated with different diseases (differential diagnosis) or lack any causal explanation.
- But being healthy (or diseased) is also a subjective construct of the patient
- How do we represent the diseased state in the VHT?

Vision for the VHT: a 6D scaffold

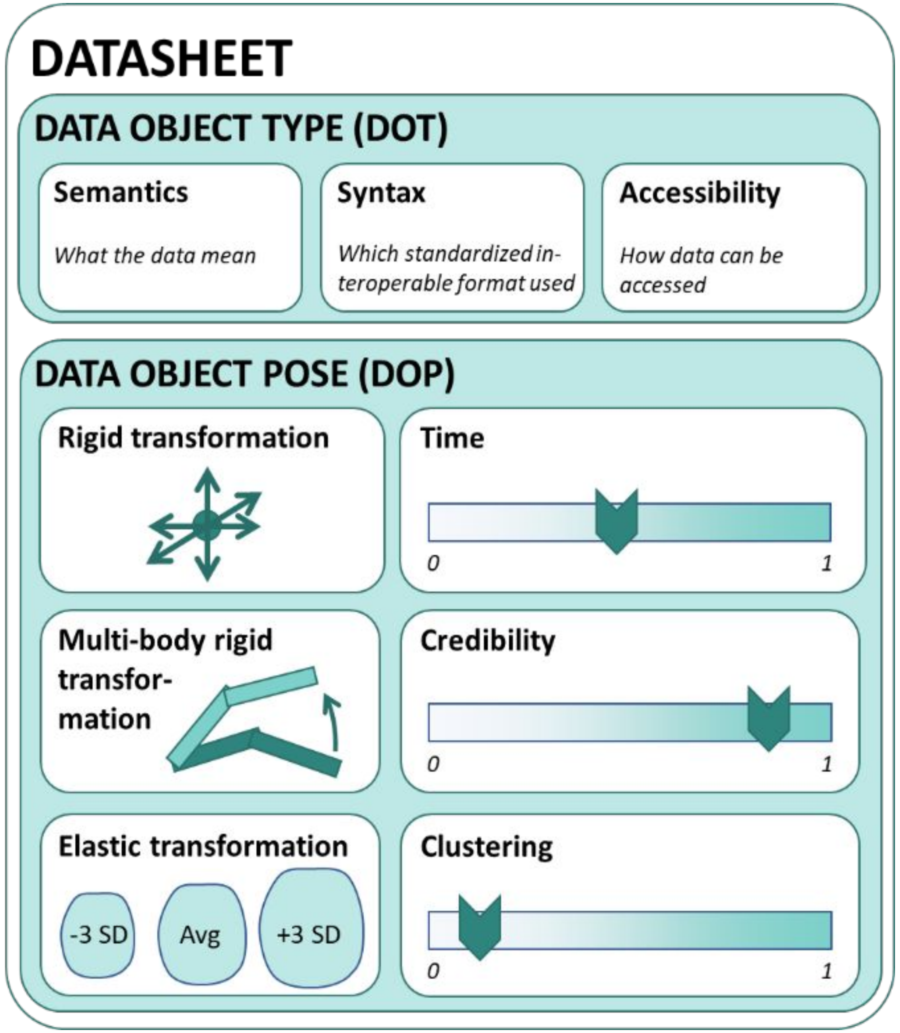


On the scaffold we store ….

- **Data objects**
  - Data Object Type: a dynamically expandible ontology of all data types supported by the VHT framework
  - Data Object Pose: the position and orientation of the data object in a 6D semantic space
- **Model objects**
  - Remote execution procedures that predict certain data objects when provided in input with some other data objects

The six dimensions

- Every data object is annotated with a data object type (DOT) and a data object pose (DOP)
- DOT: makes possible the automatic association between model objects and data objects
- DOP: provide a standardised representation of the data objects over:
  - Space (anatomical space of the average human body)
  - Time (human life span from birth to death)
  - Clustering (from individual to average *Homo Sapiens Sapiens*
  - Credibility
- Credibility = 0 is a data object with evidence of credibility; credibility = 1 is a data object certified for medical use by a competent authority. The Community of Practice defines intermediate steps.
- All other info is added as optional metadata



DATASHEET

DATA OBJECT TYPE (DOT)

Semantics — *What the data mean*

Syntax — *Which standardized interoperable format used*

Accessibility — *How data can be accessed*

DATA OBJECT POSE (DOP)

Rigid transformation

Time

Multi-body rigid transformation

Credibility

Elastic transformation

Clustering

Normalisations

Time

0 - Birth 1 - death

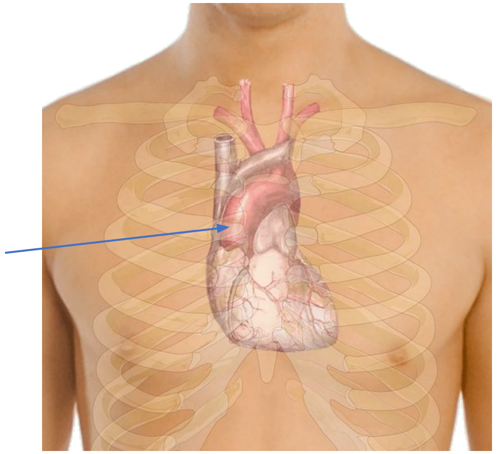
Credibility

0 - none 1 - max

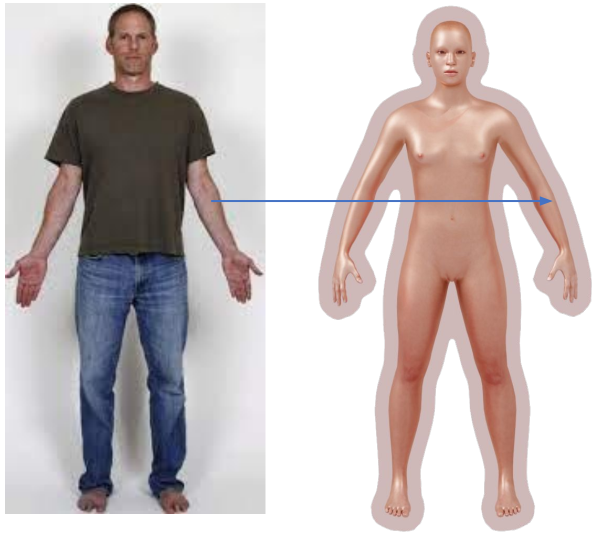
Clustering

0 - none 1 - all

Space



Aortic blood pressure



Non-spatially localised data: conventional anatomical locations

Spatially localised data: spatial transformation

the degree of clustering k is defined as k = 1/x, where x is the number of clusters: the homo sapiens sapiens cluster has k = 1; male-female clustering has k = 0.5; and individual datasets have k = 0, assuming an infinite number of human beings

Model objects

- The Model Object Type (MOT) must contain the list of DOTs from its input and output set. It must also contain all the metadata to ensure traceability, such as version, author, etc.
- Location in the anatomical space, Age, and clustering does not apply to models. So models are organised in the VHT space only along the Credibility Axis. However, for consistency, and to reuse user interfaces to search resources, we can conventionally place model objects in the anatomical space and on the Age axis at a coordinate that is the average of its output's coordinates.
- One thing we need to specify in the metadata of the Model Objects is what happens when the model has run. I see two scenarios:
  - I do not trust my model so much, so before the outputs are published in the VHT, I want to inspect them. This would require the output to be placed in a sandbox of the modeller, where they can revise, accept and publish the resulting outputs
  - Second, the output is published automatically in the VHT as soon as computed.

Execution of model objects

- Each model object should be annotated with all the execution requirements, including the expected computational cost.

- We can automate the polling for new valid input sets and the model's execution with those new inputs (eager execution). When a new valid input set appears, the model's execution with that input is put in a queue, and there it stays until someone offers a computational resource that fits its requirements.

- On the other hand, we can offer computational resources with certain characteristics and reserve them for specific groups of model objects. So, I can expose my computational resource to run only my models, the models that produce the output I need, any cardiovascular model, or any model whose inputs are replicated in the database near my cores.

- A brokering software would match demand and offer, send the container of the model / replicate the input data in the right location, and start the execution. The output would then be handled according to the options for that model.

Models brokering

- The same logic used for the data objects should apply:
  1. Someone defines a Model Object Type by specifying the output set the model should predict; I do not specify the input set, as different models may predict the same output using different inputs.
  2. Someone submits a request for models of that MOT.
  3. Someone posts a new model of that MOT, with a sub-MOT specifying the input set. The input set is added to the polling list; if valid input sets exist, the model is put in the queue for execution for each input available.
  4. Someone provides computational resources for those executions; the model runs, and the predicted outputs are added to the VHT.

Workflows

- Complex problems are more easily modelled as orchestration of different models.
- If the models are strongly coupled in their execution, exposing them as monolithic executables is usually better. Otherwise, orchestrations can be exposed as workflows.
- A workflow represents an orchestration with two instruction sets: a data flow and a control flow. If the control flow is complex, these workflows should also be exposed as monolithic executables that use any workflow manager to run.
- The exception is when the complex workflow produces intermediate results that are worth storing as data objects in the VHT; in this case, the workflow manager should be integrated into the VHT, at least for the data management
- If the control flow is a Directed Acyclic Graph (one model after the other), the workflow may be left implicit in the data structure, thanks to the VHT automatic execution.

Some points to discuss /1

1. How do we represent the 6D scaffold to those searching for specific knowledge?
2. It all starts we the creation of a Data Object Type. How can a user ask for the addition of a new DOT?
3. If a DOT exists but is empty (no data) in some cells in the scaffold (for example, no data of that type for women over 50, how can a user request the community to produce such missing data?
4. Which spatial transformation functions do we need?

Some points to discuss /2

1. How do we represent the boundary conditions? What do we call them?
2. How do we represent Clustering? The Clustering value is not unique: male-female or children-adults have both k=0.5 but represent very different clusters.
3. How do we represent the diseased state in the VHT?
4. Should MOTs also have a 6D pose?
5. What is the minimum metadata set to automate models' execution?
6. Do we need to integrate workflow managers? How?

http://www.edith-csa.eu

Deliverables available under tab ‘dissemination/material’

Indication of interest via de contact form

EDITH

EDITH is a coordination and support action funded by the Digital Europe program of the European Commission under grant agreement n° 101083771.