A RESEARCH TALK

PART I: EXTENDING A METAMODEL FOR FORMALIZATION OF DATA WAREHOUSE REQUIREMENTS

Natalija Kozmina, Laila Niedrite, Janis Zemnickis

University of Latvia, Faculty of Computing

Introduction

- We focus our research on applying **demand-driven** (more precisely, **user-driven**) methodology to construct a **DW conceptual model**

- We interpret **DW information requirements** gained from interviews as indicators

- **Indicator definition** from BABOK® Guide:
  An **indicator** identifies a specific numerical measurement for a goal, impact, output, activity, or input. Each factor of interest has at least one indicator to measure it properly, but some may require several.

Background and Questions of Interest

Current Situation

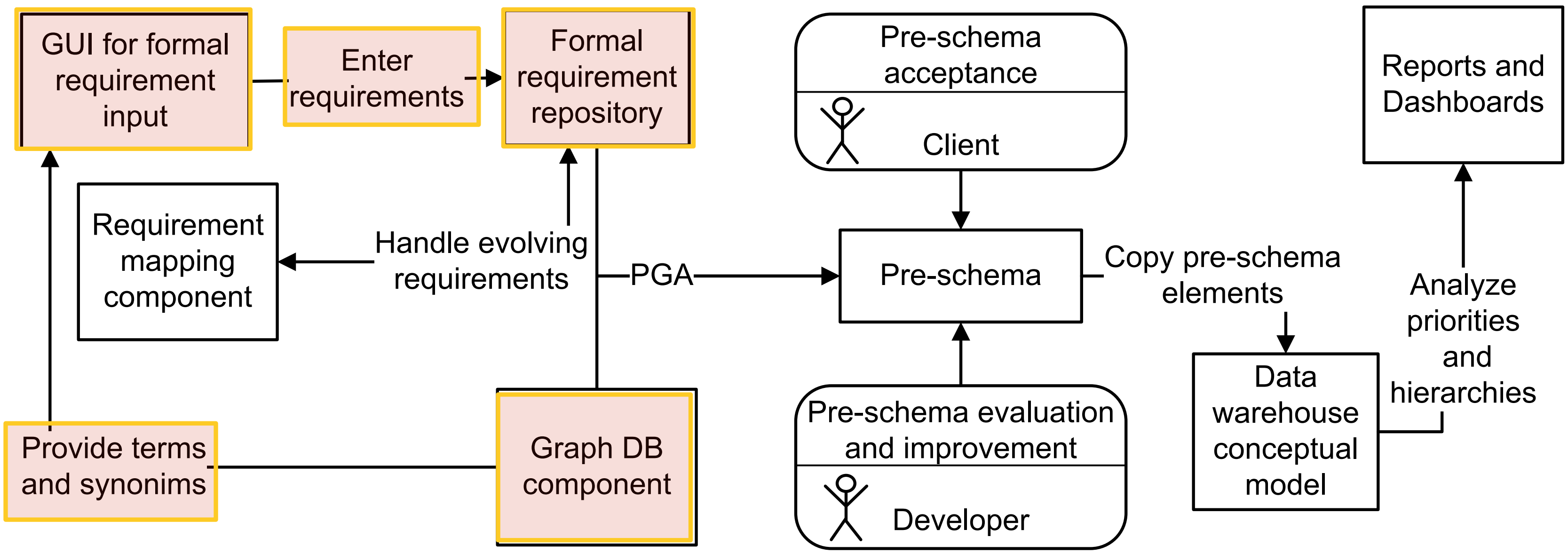
- DW of the University of Latvia accumulates data to reflect diverse indicators
  - Student enrolment, strategic indicators, staff workload, user activity in CMS, etc.
- Regular demand from client's side for DW reports
  - Over 150 reports and growing
- Interest in dashboards

MicroStrategy

Questions

- How to structure and systematize DW information requirements?
- Which reports and schema elements to incorporate into dashboards?

Deriving a Conceptual Model of a DW from Information Requirements



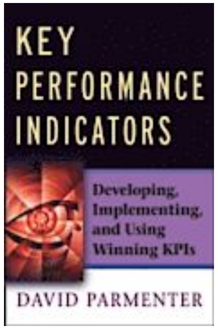
Requirements in natural language → **Requirement formalization metamodel** → Formalized requirements (indicators)

Requirement Formalization Metamodel

Initial version of the metamodel

Based on over **330 indicators** from business field

- Measurement perspectives: customer focus, environment & community, employee satisfaction, finance, internal process, and learning & growth
- **Source:** Indicator database from *"Key Performance Indicators: Developing, Implementing, and Using Winning KPIs"* by Parmenter, D. (2010)



Extended version of the metamodel

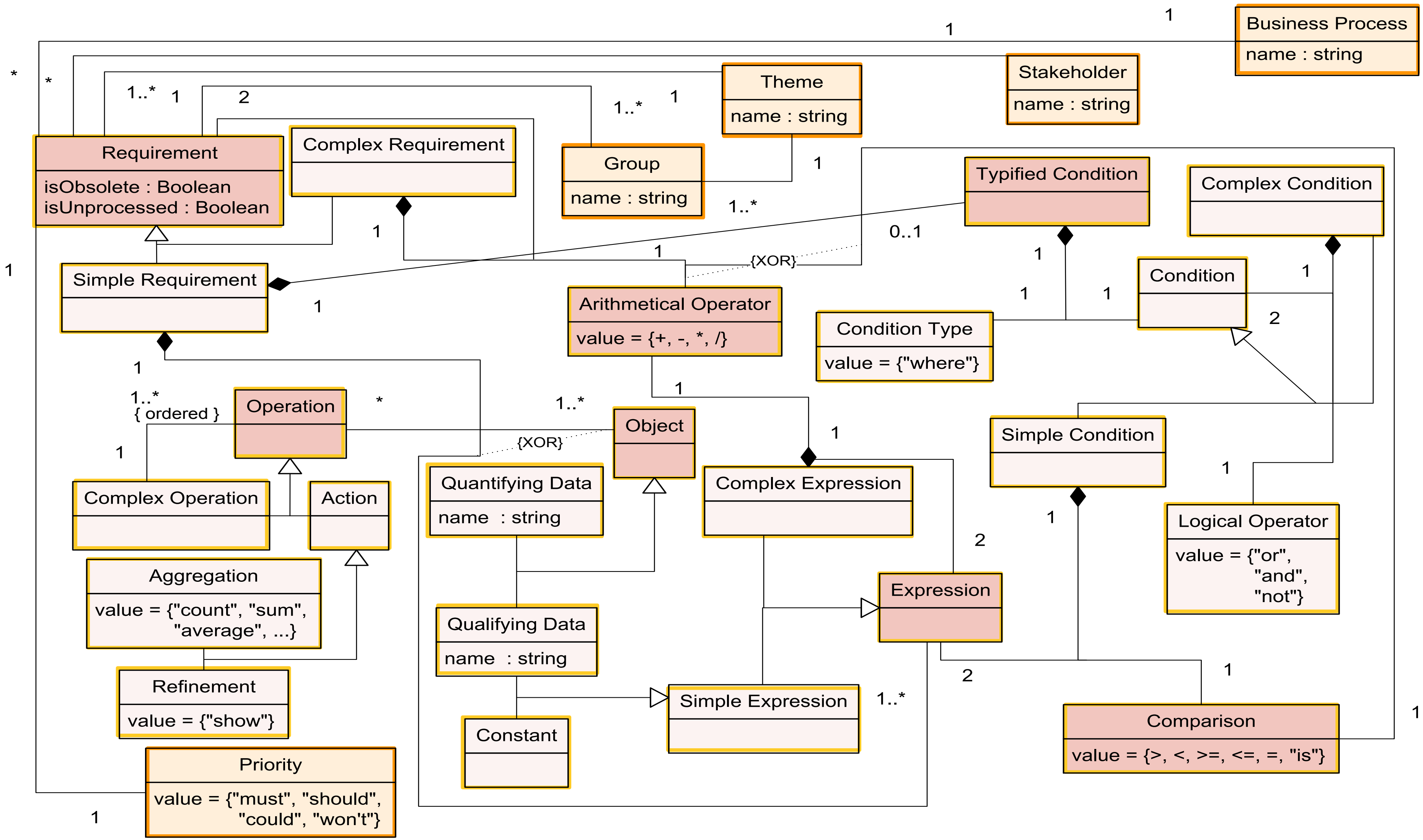
Based on over **150 indicators** from the real DW project of the University of Latvia

- Indicator groups: student enrolment, strategic indicators, staff workload, user activity in CMS, staff/ student publications, etc.
- **Source:** Indicators for existing reports developed with MicroStrategy tools

Case Study & Findings

- We analyzed sentences that express indicators in natural language with an aim to discover **common patterns**
- **Initial version of the metamodel:**
  A set of **principles** was worked out that serves to **translate** the informal requirements in natural language to a state that is **compatible** with the requirement formalization metamodel
  - *calls* → count (call), *number of visits* → count (visit),
  - *listing of customers* → show customers,
    total income → sum (income), etc.
- **Extended version of the metamodel:**
  Indicators were **reformulated**, and **checked for compliance** with the initial metamodel – around **14%** of indicators did not comply
- As a result, new **classes** and **relationships** between classes were added to the requirement formalization metamodel

Requirement Formalization Metamodel



An Example of a Formalized Indicator

- **An indicator in natural language:** *"The ratio of master level graduates in the University of Latvia in 2016, who are employers, has to be 10% of master level graduates in the University of Latvia in 2015"*
- **A formalized indicator:**
  *"((count (graduate) where level='master' and year='2016' and status='employer') / (count (graduate) where level='master' and year='2016')) = (10% * (count (graduate) where level='master' and year='2015'))"*

- If there are such components as "%", "percent", "percentage", or **"ratio"**, then it is substituted by **division of partial quantity by total quantity**
- A component to be measured is treated as an aggregated number of all its occurrences: **"graduates" → "count (graduate)"**

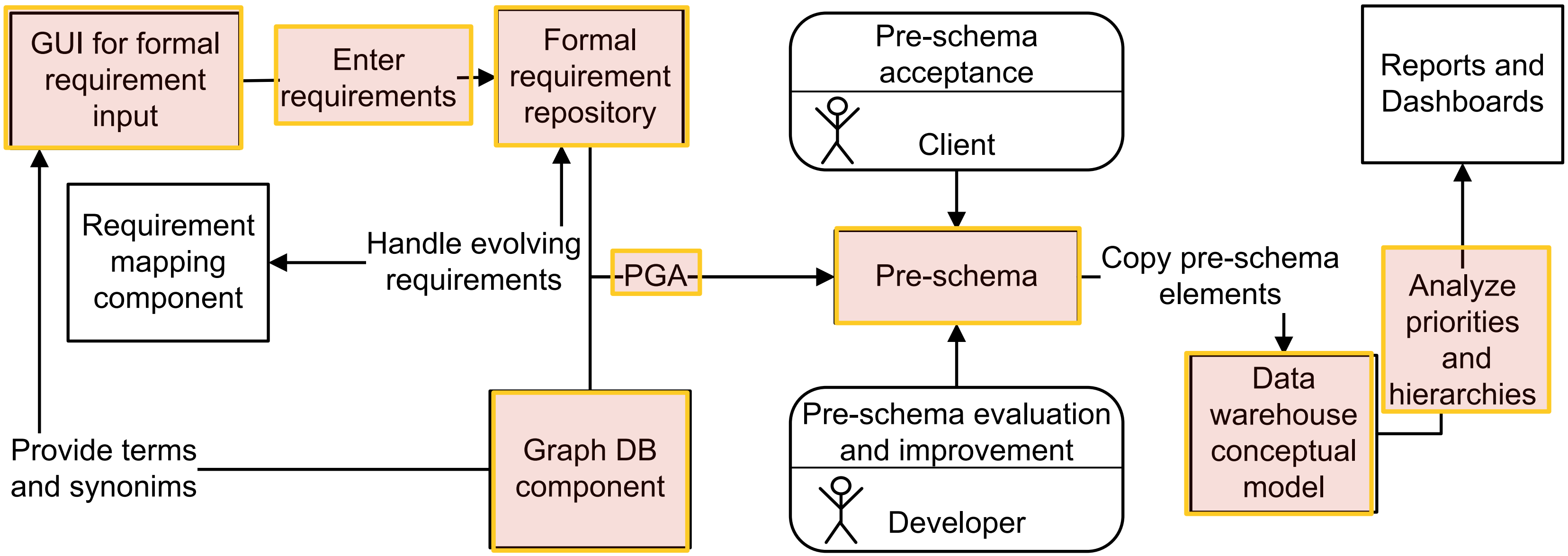
- **"Has to"** was interpreted as a request for equality → **"=" sign**
- **"10%"** is a **simple requirement** that consists of a single constant value
- Now it is possible also to **evaluate the ratio** (e.g. "has to be 10% …")

Prioritization Technique

- Preferred approach – **MoSCoW** Analysis
  - A fast and straightforward approach with precisely defined priority values
  - Doesn't require complex calculations during re-prioritisation process
  - Suitable for a small group of decision-makers

- **Priority values** in MoSCoW (from BABOK® Guide)
  - **"Must"**: *must be satisfied* in the final solution for it to be considered a success
  - **"Should"**: *should be included* in the solution if it is possible
  - **"Could"**: *desirable* but not necessary
  - **"Won't"**: will not be implemented in a given release, but *may be considered*

How requirement priority values are propagated to pre-schema elements?



- A pre-schema generation algorithm (**PGA**) can **map** elements of formalized requirements to DW schema elements
- If a **schema element** (e.g. a Study Program attribute) has **multiple priority values** (e.g. must, could), then the one with the **higher value** is assigned (i.e. must)

Which elements of the accepted pre-schema to incorporate into dashboards?

- Detect **schema elements with highest priorities**
- Check if any of these elements build up **data hierarchies**

- Examples of formalized requirements (with high priorities):
  - **R1:** show **course** count (user session occurrence) where user role = “student”
  - **R2:** show **course category** count (user session occurrence) where user role = “student”
  - **R1** → **R2** is a requirement hierarchy example, because schema elements form a hierarchy **Course** → **Course Category**

- A dashboard report would include the **R2** requirement

Dashboard Example

Dashboard

Darbinieki un studenti: Akadēmiskā darbā strādājošo un studentu kopskaits pa fakultātēm (Dashboard)

Laiks	Pamatstruktūrvienība	Metrikas	Akad. darbā strādājošo skaits	PLE	Studentu skaits	Studentu skaits / PLE	Studentu skaits / Akad. darba strādājošo skaits
09.2016	Bioloģijas fakultāte		36	4,6669	369	79,07	10,25
	Datorikas fakultāte		35	6,0535	861	142,23	24,60
	Ekonomikas un vadības fakultāte		68	13,3511	2 000	149,80	29,41
	Fizikas un matemātikas fakultāte		83	14,1429	582	41,15	7,01
	Ģeogrāfijas un Zemes zinātņu fakultāte		64	10,1033	538	53,25	8,41
	Humanitāro zinātņu fakultāte		109	21,5343	1 398	64,92	12,83
	Juridiskā fakultāte		39	8,96	1 367	152,57	35,05
	Ķīmijas fakultāte		34	4,9142	449	91,37	13,21
	Medicīnas fakultāte		54	9,9909	1 009	100,99	18,69
	Pedagoģijas, psiholoģijas un mākslas fakultāte		98	21,2919	1 614	75,80	16,47
	Sociālo zinātņu fakultāte		41	7,3176	1 127	154,01	27,49
	Teoloģijas fakultāte		13	1,8416	137	74,39	10,54
	Vēstures un filozofijas fakultāte		25	3,4524	373	108,04	14,92
Kopā			699	127,6206	11 824	99,05	17,60

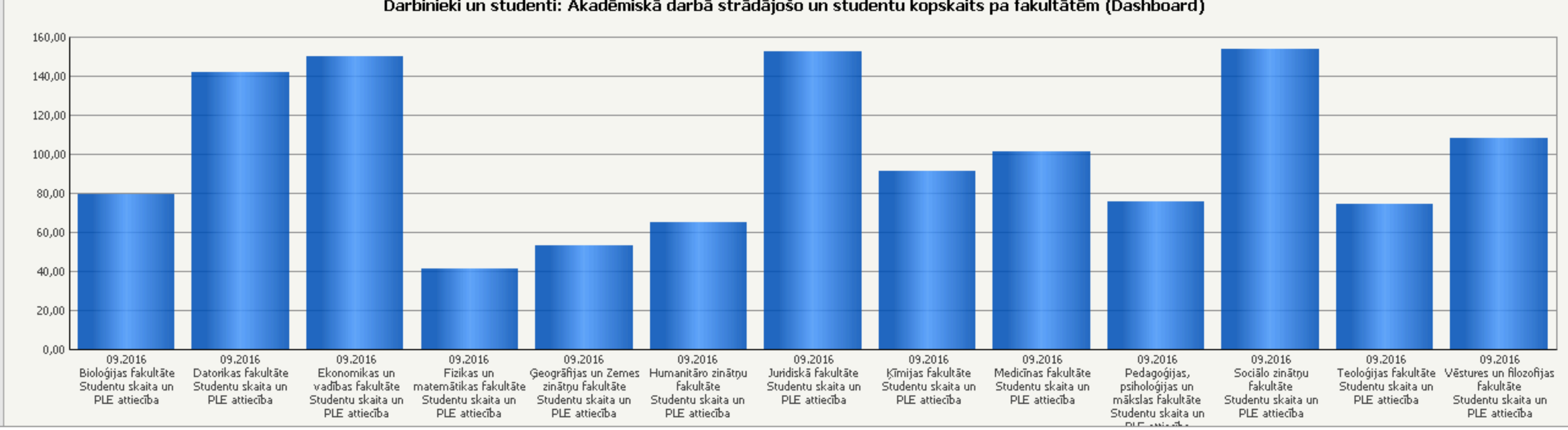
Studentu skaita un akad. darba strādājošo skaita attiecības bāzes vērtība ir 10,4.
Indikatoru krāsu nozīme:
Zaļš : [7,4; 13,4];
Dzeltens : [4,4; 7,4) U (13,4; 16,4];
Brūns : (-bezgalība; 4,4) U (16,4; bezgalība).

Studentu un akad. darba strādājošo skaita attie...

nav nulles vērtībā

Laiks
05.2016
06.2016
07.2016
08.2016
09.2016

Darbinieki un studenti: Akadēmiskā darbā strādājošo un studentu kopskaits pa fakultātēm (Dashboard)



Summary

- A case study was conducted to test the existing requirement formalization metamodel on a set of **over 150 indicators** for a real currently operating **DW project** of the University of Latvia
- Due to a specific structure of requirements that contain an **evaluation of ratios**, the metamodel had to be restructured and extended with some additional classes like **themes**, **grouping**, **business processes**, **stakeholders**, and requirement **priorities**
- **MoSCoW analysis** was chosen as the most suitable requirement prioritization technique
- Application of **priorities** was discussed in the context of **dashboard** and **report** development

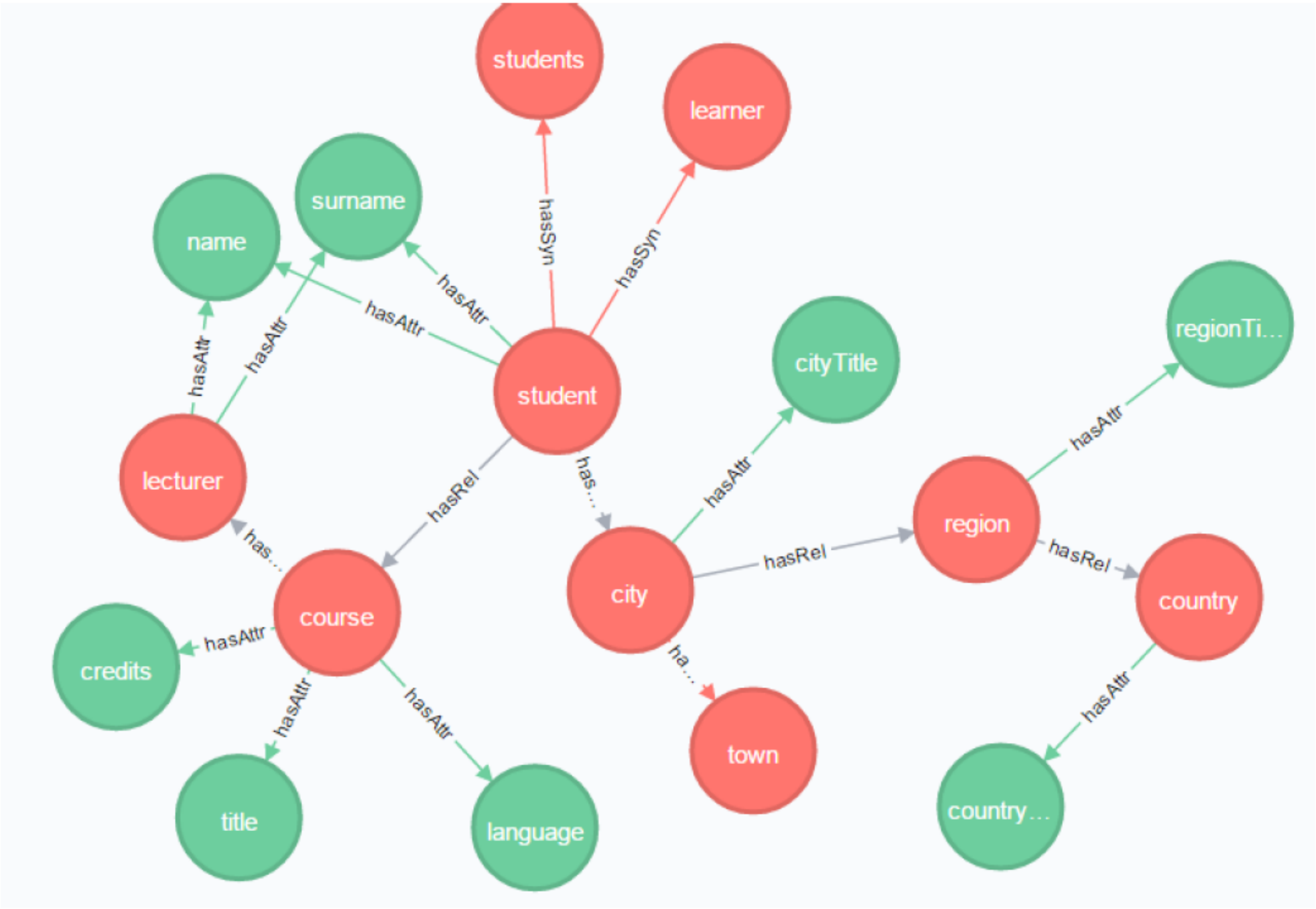
Details on Technical Implementation of iReq

iReq as a GUI for formalized requirement input

- A **web-based** tool with **responsive** design
- iReq is written in **PHP** (Laravel framework)
- Requirement input: **HTML, CSS, JavaScript** (Bootstrap, jQuery libraries)
- Data are stored in **MariaDB**
- **Neo4j** for the glossary

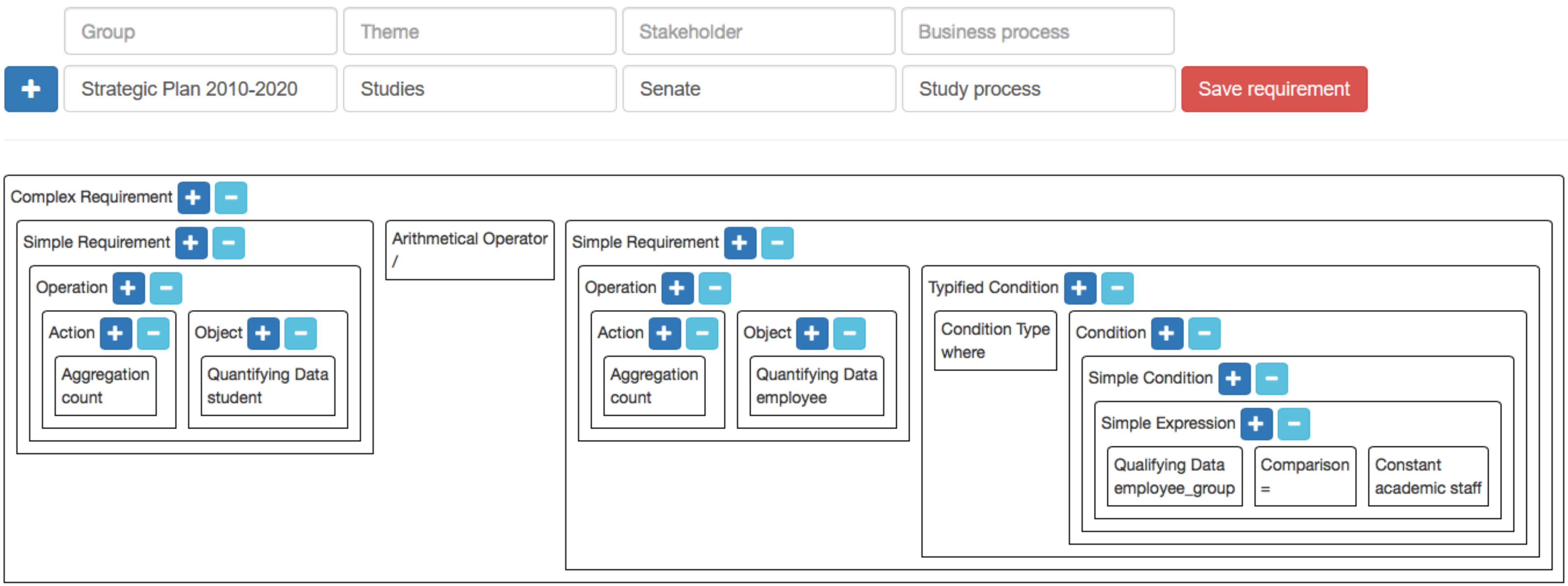
An Example of Glossary as Graph DB

Source DB structure → .CSV → Neo4j Graph DB



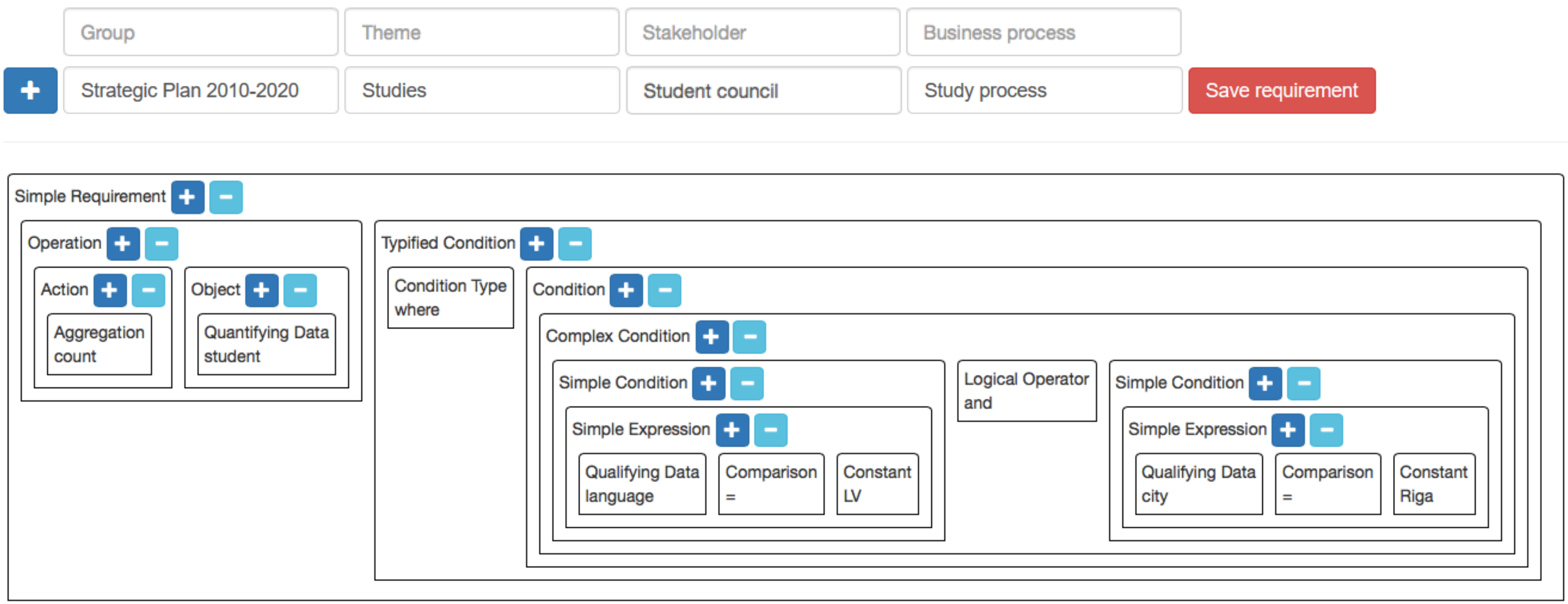
A Requirement Example in iReq Tool

- An example requirement from the Strategic Plan (2010-2020)
- *“Show information on student and academic staff ratio”*



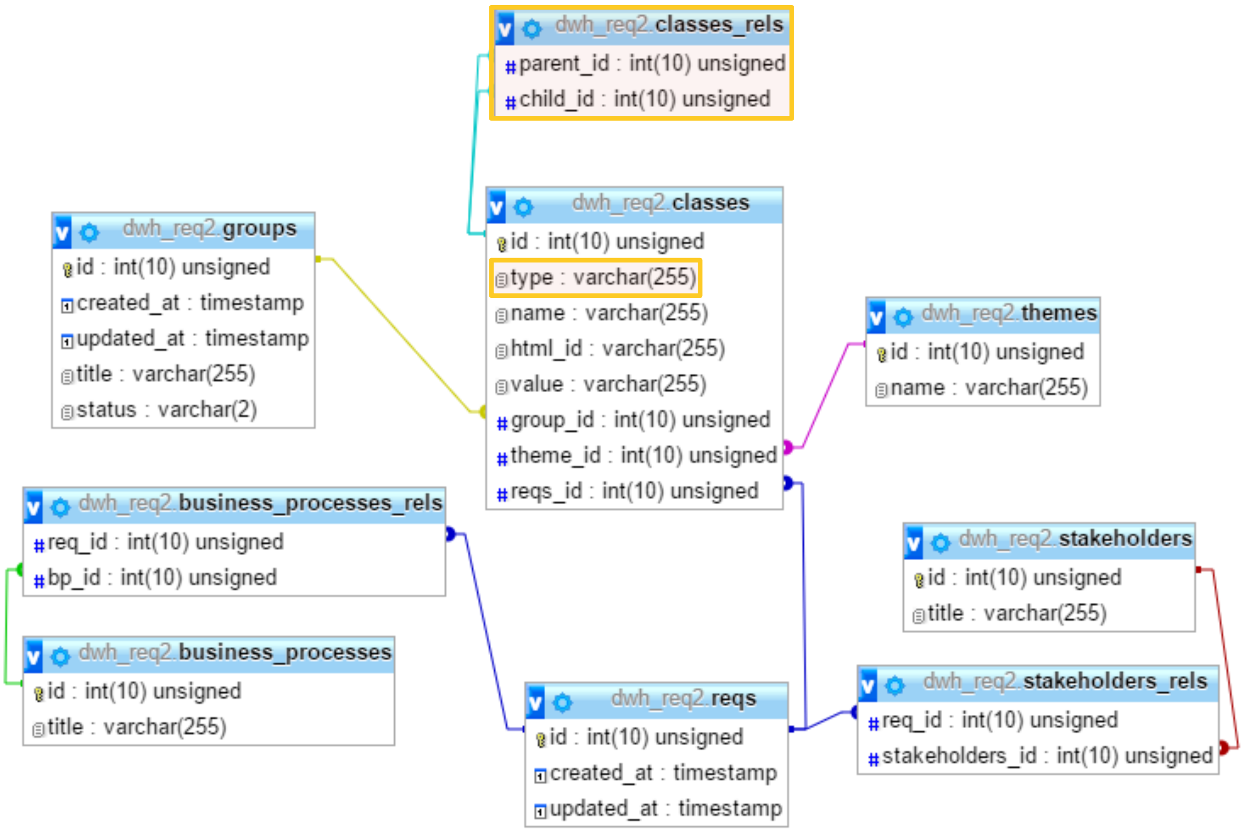
Another Example in iReq Tool

- An example requirement from the Student Council
- *“Show information on students from Riga that attend lecturs held in Latvian”*



ER Model of iReq Requirements Repository

- Table *classes* stores data on all the elements of requirements
- *classes.type* - Action, Simple Condition, Quantifying data, etc.



Future Work for iReq GUI

- Perform **more GUI testing** of the iReq tool to improve it (e.g. add informal description of requirements)

- Provide an option for entering formalized requirements manually as **input expressions** in order to parse with some natural language processing component (e.g. **Xtext**, **SpaCy**) and **save(retrieve)** them **into(from) a database** correctly

- Make collected requirements fully or partially **reusable**


THANK YOU

PART II:
DATA MODELLING FOR DYNAMIC MONITORING OF VITAL SIGNS - CHALLENGES AND PERSPECTIVES

Natalija Kozmina*, Emil Syundyukov*, Aleksejs Kozmins^

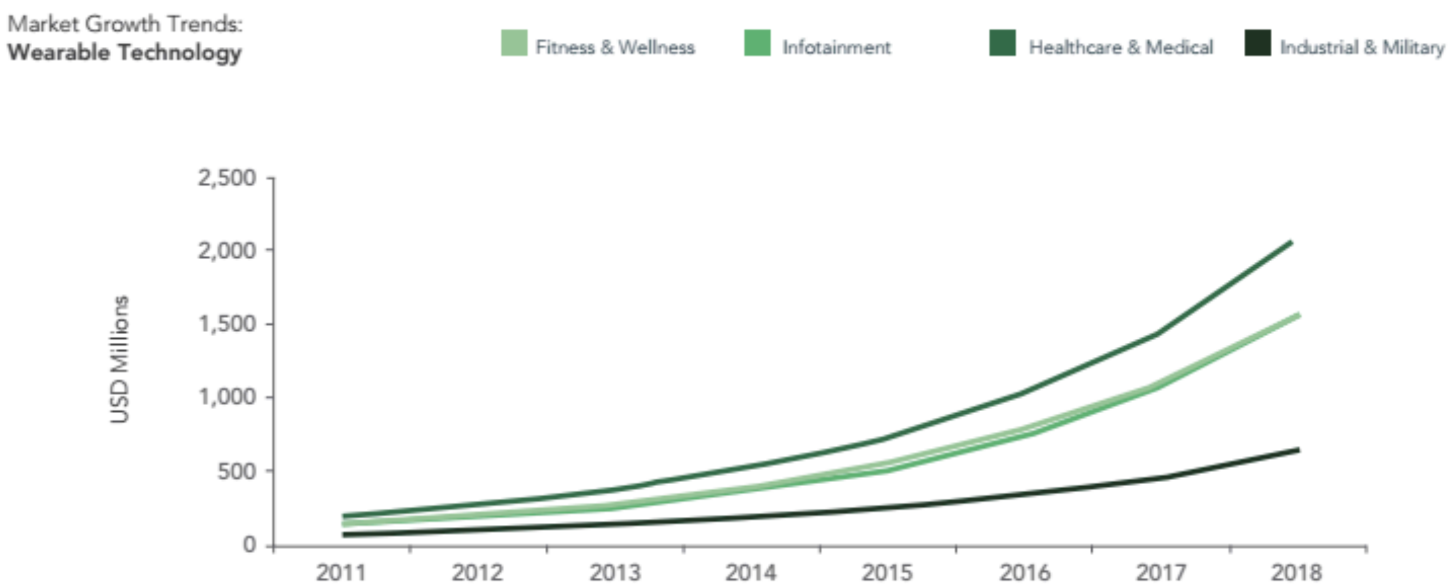
*University of Latvia, Faculty of Computing

^Accenture Latvia

Introduction

- An **injury of the knee joint** (fractures, dislocations, ligament tears) is one of the most common regardless of the age
- A **rehabilitation routine** is aimed to minimize **swelling**, return the **range of movement**, strengthen **leg muscles** by taking into account **limitations** set by a physiotherapist
- A **body sensor network** of wireless sensors attached to a patient provides a promising method to collect **clinically relevant information** about knee function in everyday life

IoT – A Possible Solution?



Care Connect

Past

The global wearable technology market stood at $750 million in 2012.

The healthcare and medical segment accounted for about 35.1 percent of the overall wearable technology market in 2012

Present

Predictions are a compound annual growth rate of 40.8 percent from 2012 to 2018.

Future

The global market will reach $5.8 billion in 2018.

The major ramp up in device sales and revenue will happen towards the latter end of the review period between 2017 and 2018.

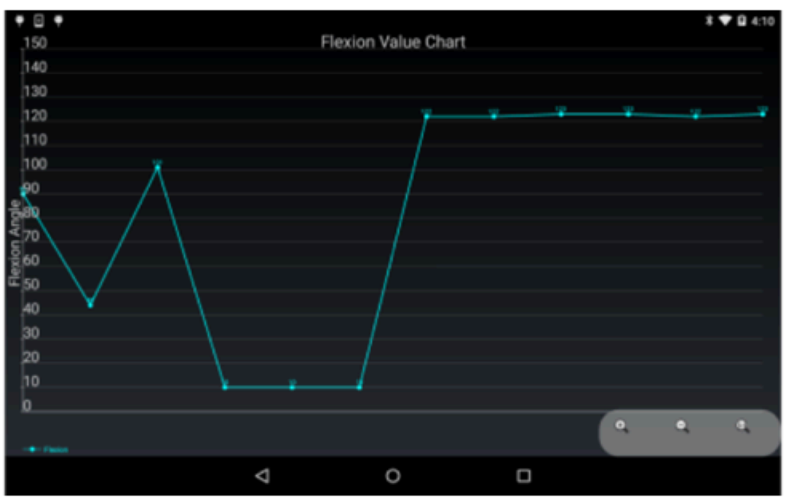
Existing Similar Solutions: Riablo

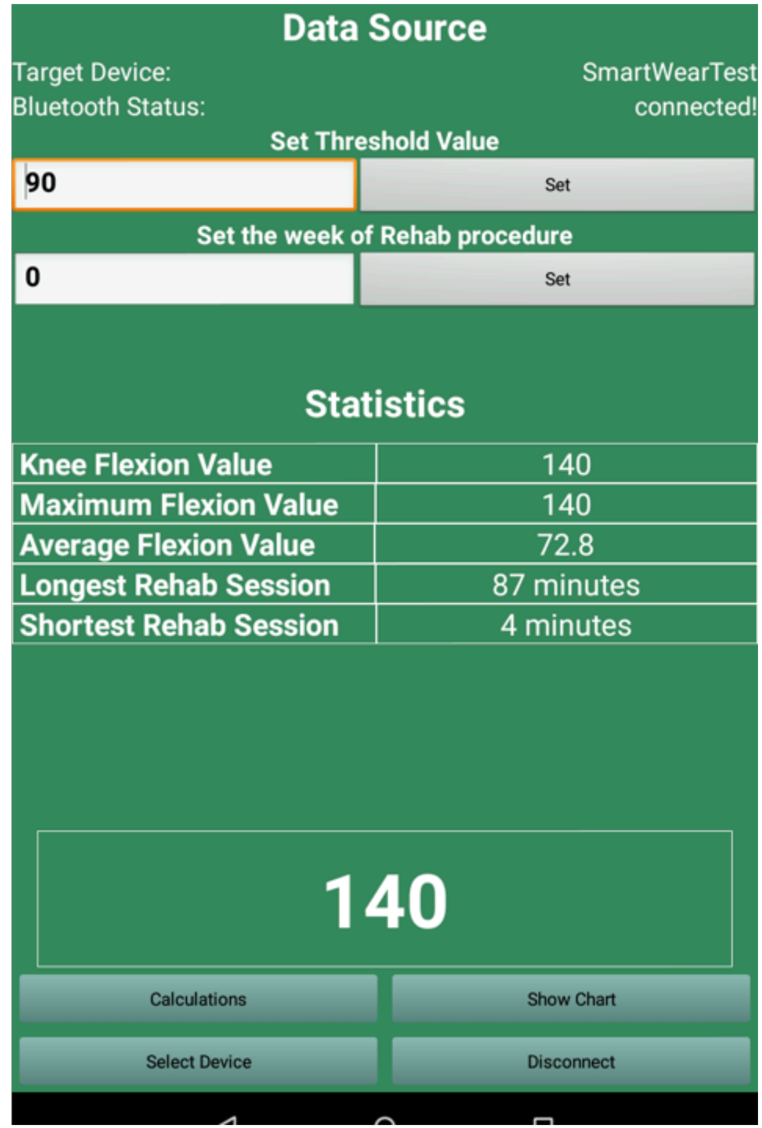
- **CoRehab:** “Controlled exercises from clinics to home”
- Wearable sensor systems to collect **biofeedback** during **exercise sessions** in form of a **video game**

Wearable Sensor System

- A **wearable sensor system** for data acquisition and analysis
- A framework to **assess** patients' state of health, monitor **dynamics** in real-time, and perform **historical** data analysis
- To calculate the knee joint flexion/extension angle, a network consisting of **four 3-axial accelerometers** is used

Mobile App to Gather Data





The Main Research Question

- ***RQ:* Is there a data modelling approach to enhance both real-time and historical data analysis?**

- **Search in Google Scholar:**
  *("data model") AND ("post-traumatic" OR "post-operative") AND -"post-traumatic stress disorder" AND rehab**

- **Results: 72** sources → **no relevant work** on data modelling for rehabilitation procedures

Which Data Model to Choose?

- **Relational data model?**
  Requires significant time and effort for adaptation

- Our **check-list** for the required data model:
  - Transactional data processing
  - Historical tracking and analytical processing feature
  - Flexibility in restructuring of the stored data according to conceptual changes
  - Effort- and time-saving in development and support

- **Data Vault (DV)** model is designed for solving the problems of flexibility and performance + a permanent system of records ("all data, all the time")

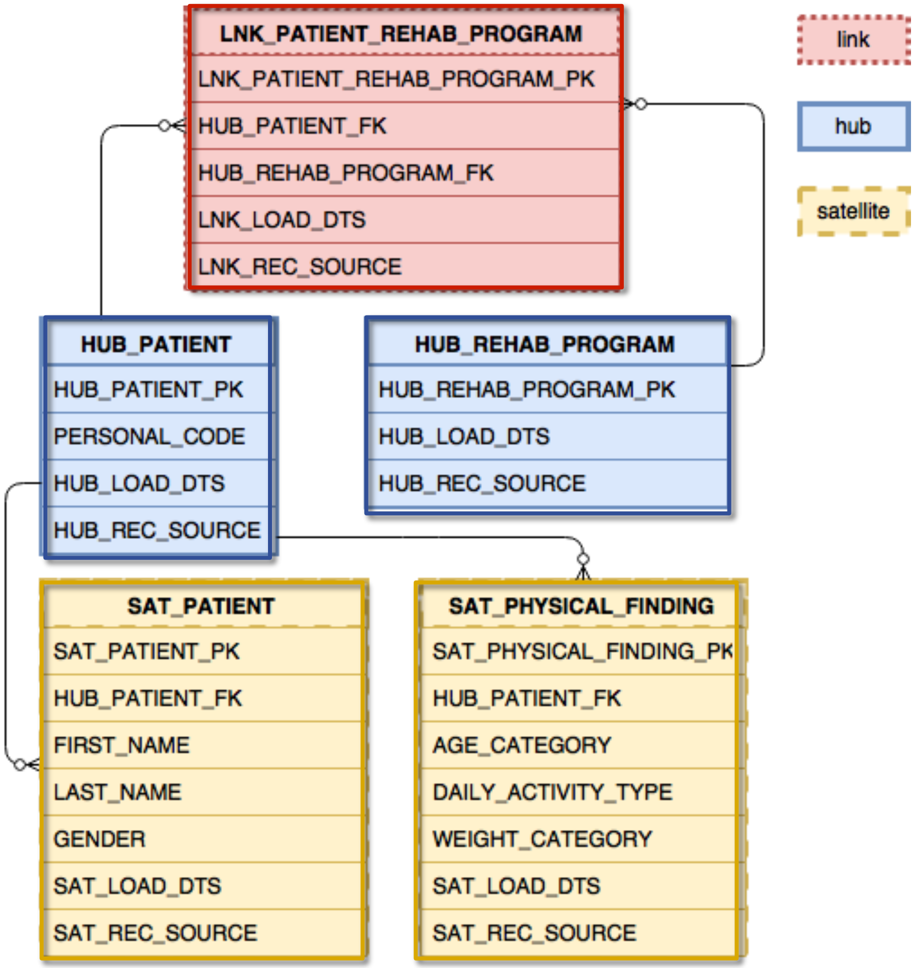
Data Vault Structure – The Main Goals

① Maximize resilience to change in the business environment when storing historical data

② Accommodate data regardless of their quality and of their conformity to standard and business rules

③ Enable parallel loading so that very large implementations can scale out without the need of major redesign

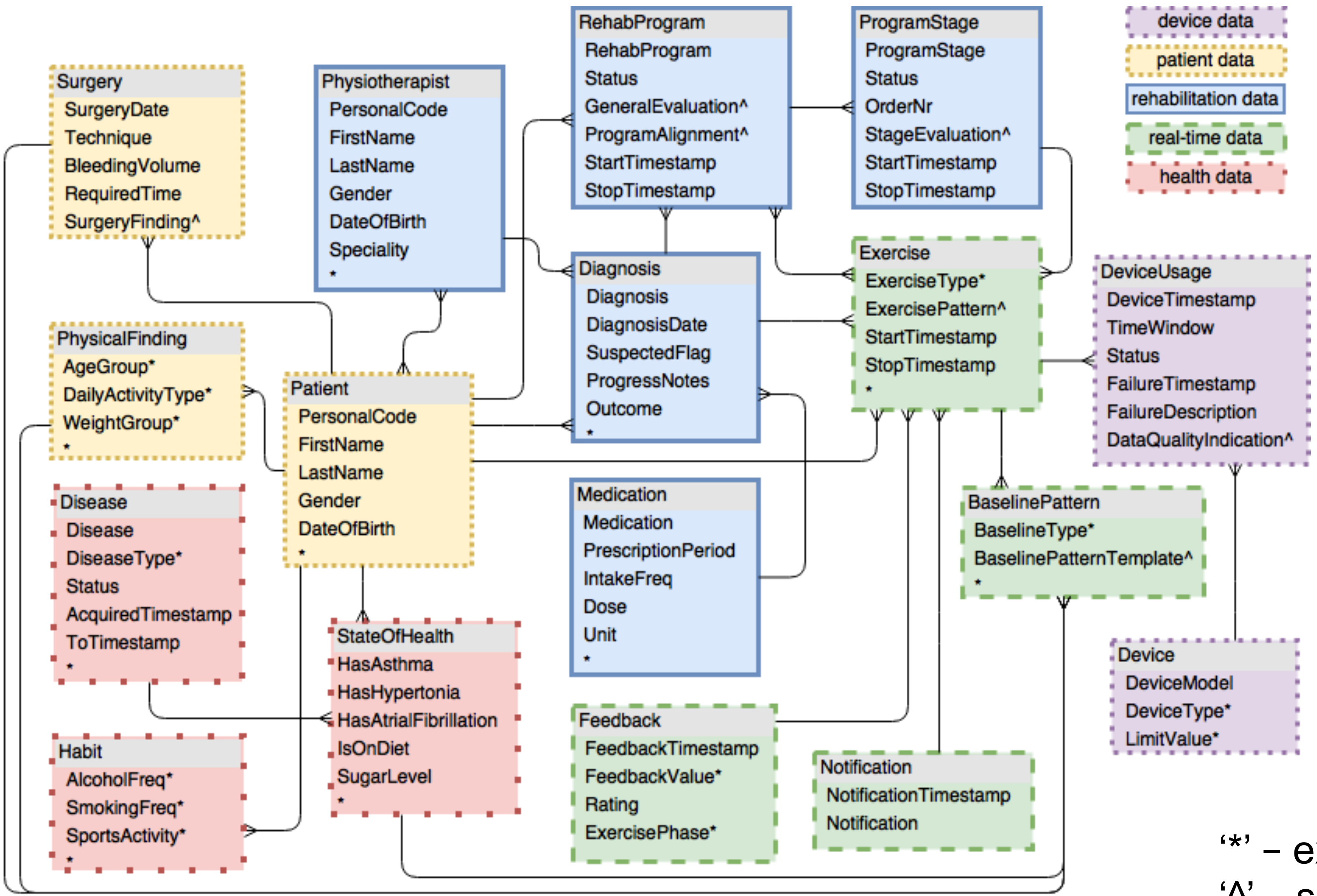
Data Vault Structure – An Example



- **Hubs** are unique lists of *business keys* that are used to track and identify key information
- **Links** define *relations* between objects
- **Satellites** contain *descriptive attributes* of the objects

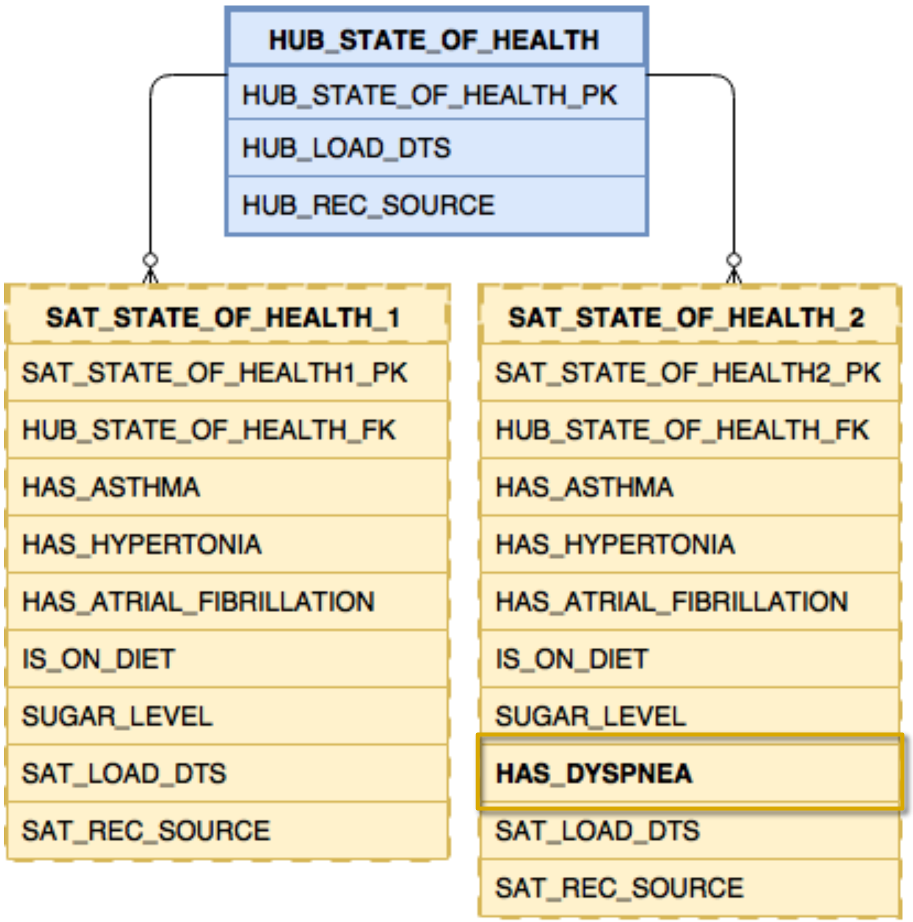
- **LOAD_DTS** – load date timestamp
- **REC_SOURCE** – record source

Structural Flexibility of the Data Model



‘*’ – extensibility
‘^’ – sub-process

Maintaining the Change History



- A **new attribute** is added
- The whole **history** of changes is being preserved
- Applying **multi-instanced** approach to represent satellites of the same hub

- Which **satellite** to use?
  - Time of creation
  - Treatment program

Starry Vault – Remarks

East European Conference on Advances in Databases and Information Systems

ADBIS 2016: Advances in Databases and Information Systems pp 137-151 | Cite as

Starry Vault: Automating Multidimensional Modeling from Data Vaults

Authors | Authors and affiliations

Matteo Golfarelli, Simone Graziani, Stefano Rizzi

- "…a data vault is **not suitable for direct multidimensional querying** both for performance reasons (it is not optimized for OLAP workloads) and because it is hardly supported by OLAP front-ends"
- "…our **future work** on this topic will be mainly focused on **investigating ad hoc techniques to support the data scientist in discovering a multidimensional structure** even in situations in which the source data are **poorly-structured or schemaless**, as is the case for document databases."

Starry Vault – Example and FD

- An **order** is made by 1 **customer** and a customer belongs to 1 **class**
- A customer normally issues several orders, each normally including several lines

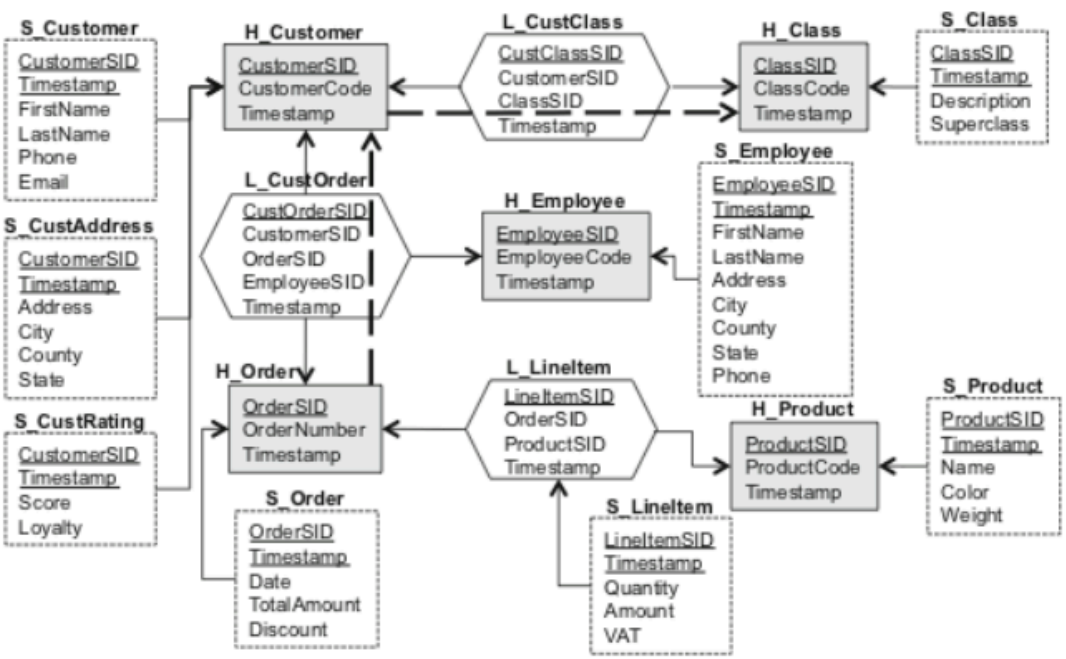
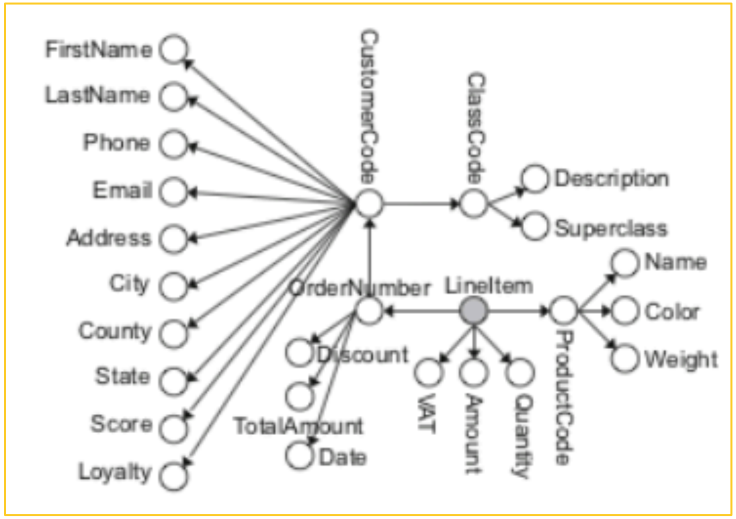


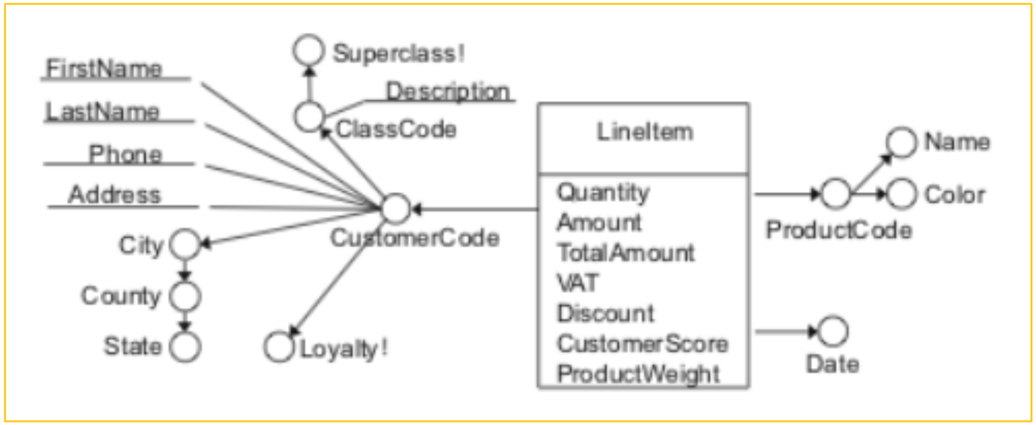
Fig. 1. A sale data vault. Grey boxes, hexagons, and dashed boxes represent hubs, links, and satellites, respectively; additional FDs are shown with thick dashed arrows

From FDs to an MD-Schema

- The goal is to detect the **FDs** holding between **hubs** related by a **link**, which can be achieved by detecting the AFDs (TANE) involving the **foreign keys** in that **link**

- **Draft md-schemata** of the fact L_LineItem **(left)**
- The **enriched md-schemata** of fact S_LineItem **(right)**





Technical Implementation Challenges

In our use-case, **change management** should be provided not only at the development stage, but also to users

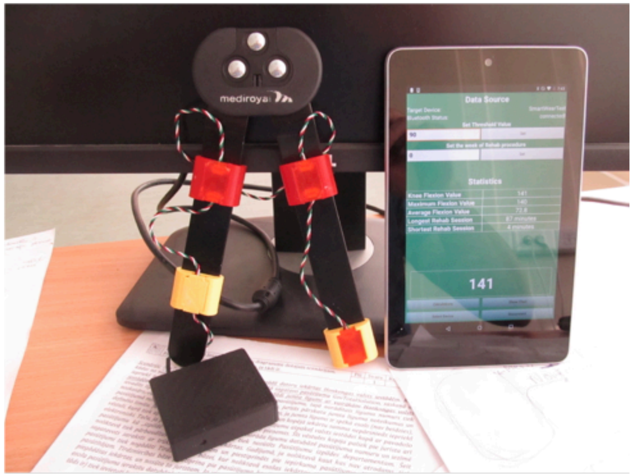
↓

① **GUI** for the transformations both in data structures and mappings

② Subject-oriented **DDL** and **DML** should let operate freely with conceptual objects

③ An additional **analytical layer** over DV data for reporting is needed + **DML for querying** over DV objects

Conclusions



- **A prototype:** wearable device + a mobile app for health data acquisition to collect, store, visualize data, and communicate via notifications

- The **overall requirement** for the data model is to give **maximum simplicity** and **flexibility** to maintain:
  - Changes in the **structure** of all entities and **inter-component** relations
  - **History** of all changes made
  - **Analytical** queries

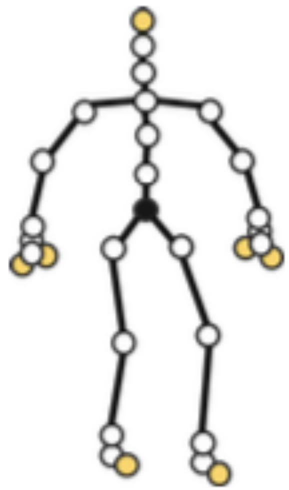
- **Data Vault (DV)** could be adapted to frequent changes in information requirements

Future Work – Empirical Studies

- **Abnormal values** in certain **individuals/cohorts** of patients to prevent the risk of gaining a **repeated trauma**
- **Positioning** of the sensor nodes for **more precise** harvesting of vital signs
- **Supplementary** sensors such as **gyroscopes** and more **advanced signal processing** to boost performance
- Tracing the wearable **device workability** in real-time by recognizing bad data in real-time during exercise sessions

DISA
Laboratory of Data Intensive Systems and Applications

Motion Capture Data and Fundamental Operations (disa.fi.muni.cz)



Skeleton pose
with 31 tracked joints

Motion Capture Data

Spatio-temporal data captured in a 3D space, e.g., 31 body joints & 120 frames per second. It is an interesting research challenge to process such complex data automatically.

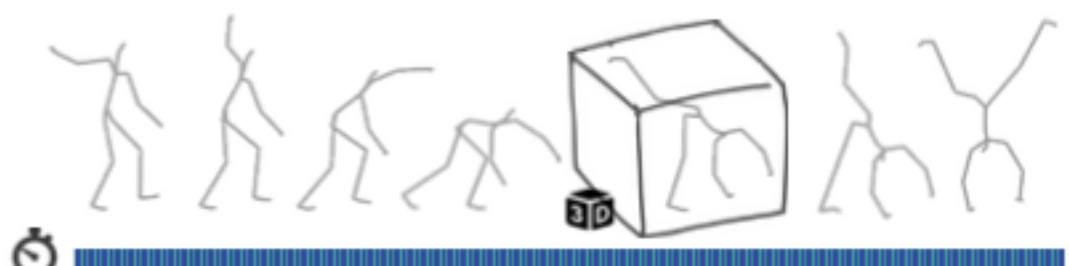
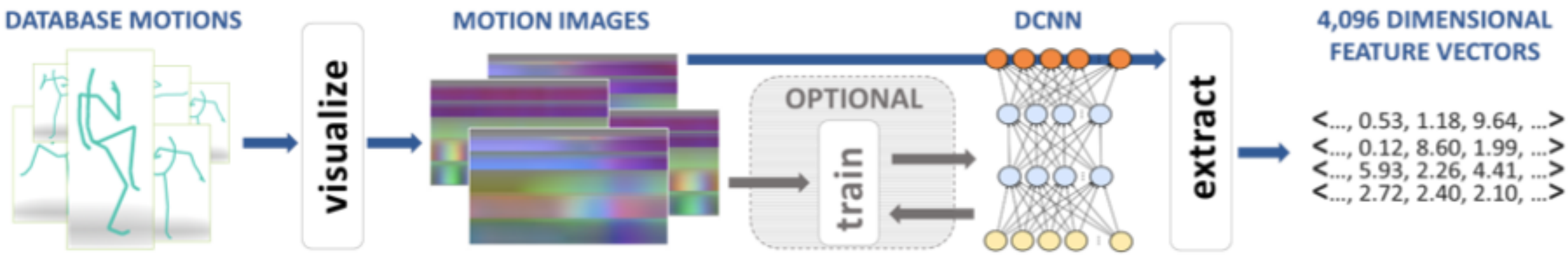


Illustration of a short cartwheel motion sequence
5 seconds of 120 Hz mocap data represent 55,800 float numbers

Terminology

Motion Capture Data and Fundamental Operations (disa.fi.muni.cz)

DISA
Laboratory of Data Intensive Systems and Applications



Transforming motions into the features
Normalized body joints are encoded within the RGB image and fixed-size 4,096D features are extracted. The similarity is compared using the Euclidean distance.

5 Levers to Reduce Healthcare Costs

Right living

- Patient health management
- Health education
- Prescription adherence

Right care

- Physician communication
- Clinical decision support
- Disease/case management
- Public health monitoring
- Safety detection

Right provider

- Resource/finance optimization
- Performance quality measurement

Right value

- Consumer decision making
- Patient information exchange
- Fraud prevention

Right innovation

- Trial operations improvement
- Avoid repeating failures
- Unmet needs identification
- Post-approval risk detection

PatientsLikeMe

People like you share symptoms, treatment info, and health outcomes.

PatientsLikeMe turns that into millions of data points about disease…

…and aggregates and organizes the data to reveal new insights.

We share back what we've learned with everyone – that's our **give data, get data** philosophy.

Then, we share the patient experience with the industry so they can develop better products, services, and care.


THANK YOU