Data Mining

DATA UNDERSTANDING & PREPARATION

Matteo Golfarelli

Political Spectrum

A political spectrum is a system of classifying different political positions upon one or more geometric axes that symbolize independent political dimensions

Let's 10 political/social questions be given e.g. "*Do you agree with the idea of repatriating African refugees?*". Each to be scored [0,..,10]

Each party have a different position (score) about each of the questions, thus its overall position is defined by a 10-dimensional point.

PCA can be used to provide a simplified view of voter orientations

Visualizing the political spectrum

Open the file behavior.csv

Attribute	Value
Q1	Score [0,..,10] for question Q1
…	…
Q10	Score [0,..,10] for question Q10
Class	Political orientation Left/Center/Right

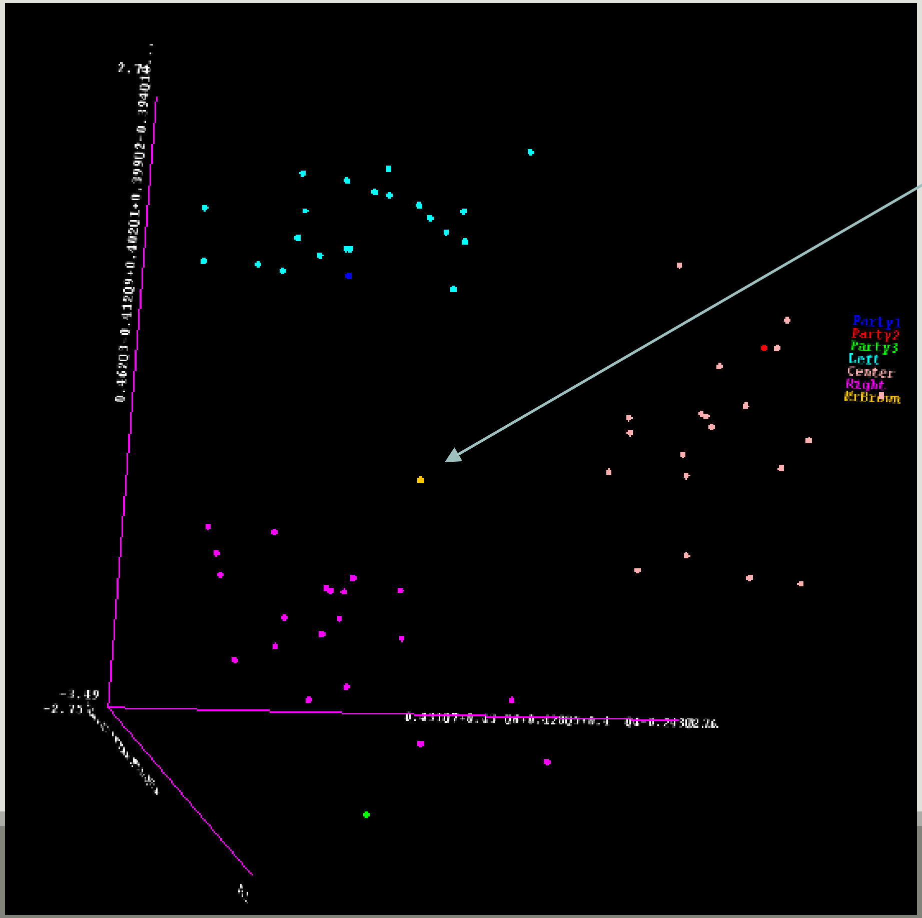
Analyze data in 10 dimensions (2D and 3D visualizations)
Apply PCA (select attribute panel) and tune covered variance so that to obtain 2D and a 3D spaces

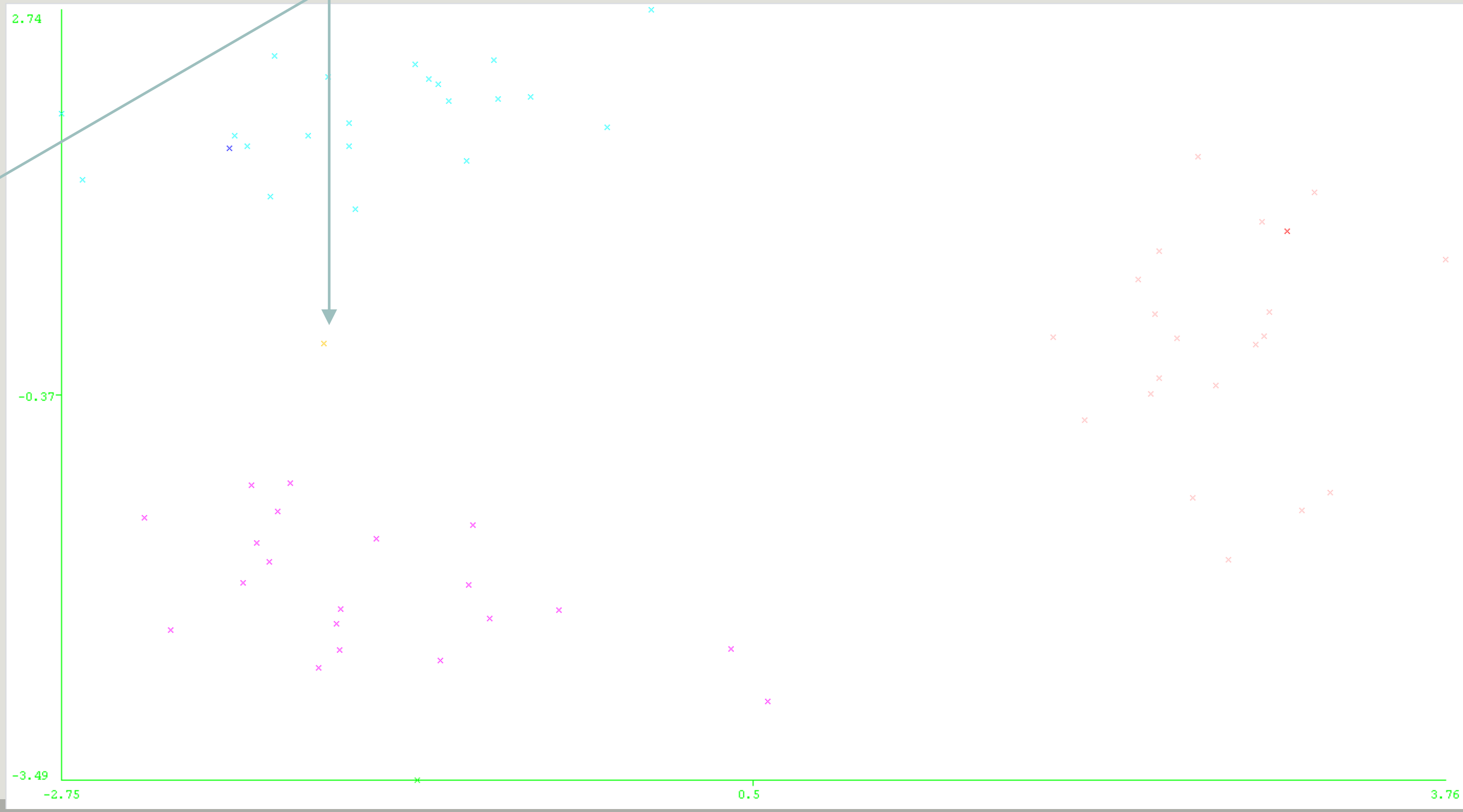
Which party MrBrown should vote for?

Visualizing the political spectrum

VarianceCovered has been set to 0.7

MrBrown





Customer Retention

Customer retention, churn analysis, Dropout analysis are synonims for predictive analyses carried out by organizations & companies to avoid losing customers.

DropOut analysis has relevance to a wide range of organizations, including (but not limited to):

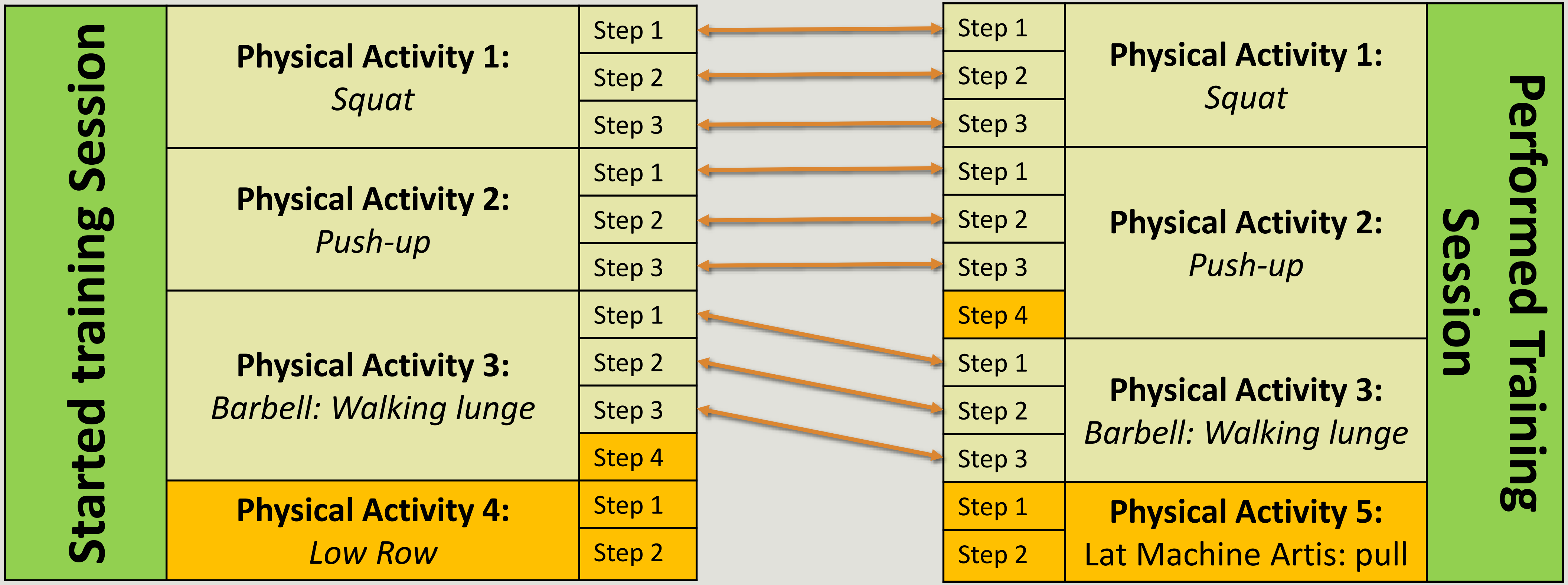
- Companies that rely on repeat business with clients to build long-term profitability
- Organizations with products or services based around a subscription or renewable contract
- Educational establishments

Prediction can be based on historical data modeling behavior of custormers that are willing to/not willing to leave the company

Historical data do not necessarily esplicitly model the customer behavior and typically a large effort must be devoted to make information explicit.

The Gym Case Study

A gym chain stores the information about every customer training session



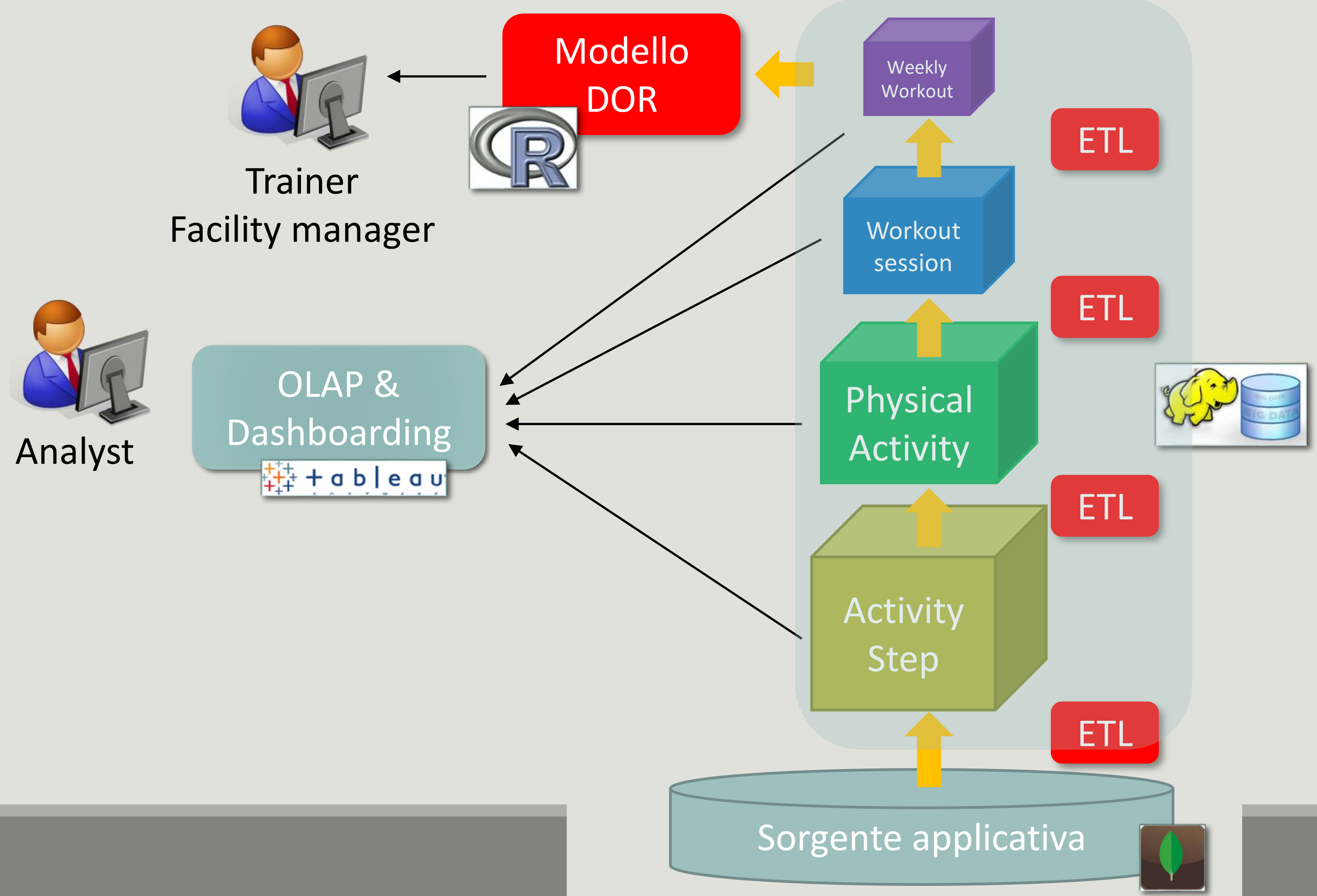
Terminology

- Physical Activity: an exercise (e.g. *Squat, Crunch*) that can be repeted several times
- Step: the execution of a Physical activity. It can be characterized by a weight, speed, duration or number of repetitions.
- Session: each time a customer enters the gym and performs a training
- Each customer has been assigned 1 or more training Workout Programs. During a session she picks and executes/performs one of the programs. The sequence of physical activities composing such program are called Assigned Workout Session (AWS). The user could perform physical activities that differ from the once in the AWS, thus the actual sequence of Physical Activities is called Performed Training Session (PWS).
- MOVE: a Physical Fitness Metric that sums up the effort required/consumed by the physical activity. It is possible to compute the number of moves related to a step, a physical activity or a Session

Acronyms

- AWP: set of training program assigned to a user
- PWP: set of training program performed by a user. It cand differ from AWP
- AWS: an exercise (e.g. *Squat, Crunch*) that can be repeted several times
- PWP: an exercise (e.g. *Squat, Crunch*) that can be repeted several times

A Big Data Architecture

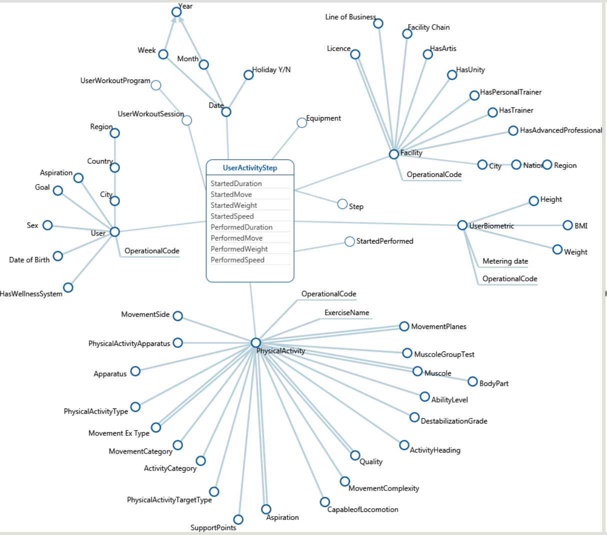


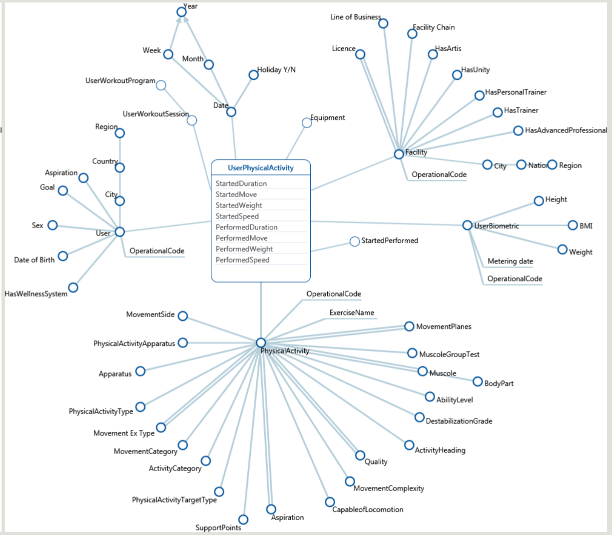
- What are the most used StrengthLoad exercises among women in different countries?
- Which application is most used by a certain type of user / In a certain country / in a certain type of facility?

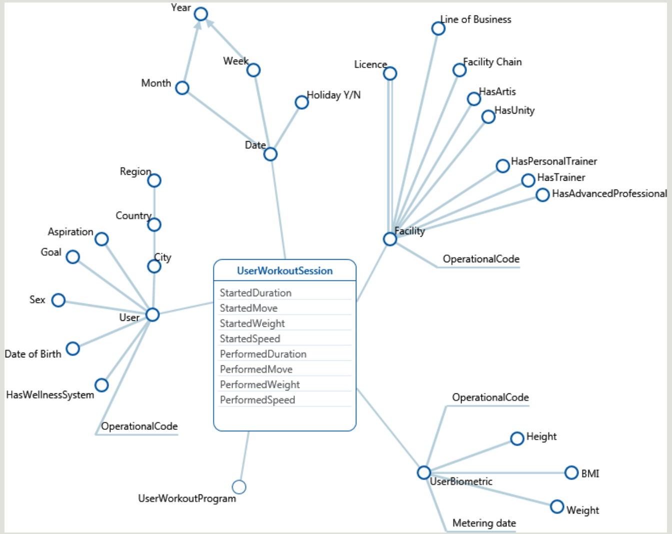
Data cubes

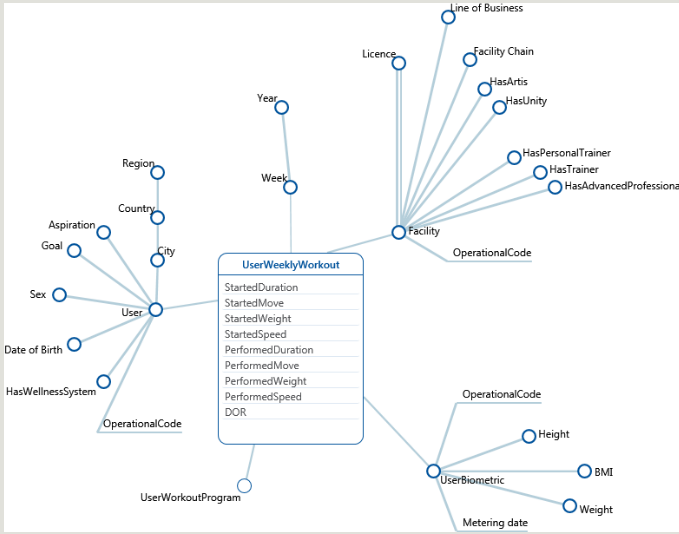
4 cubes store the raw data at different granularity levels:

- For each step:
- For each physical activity
- For each session
- For each week





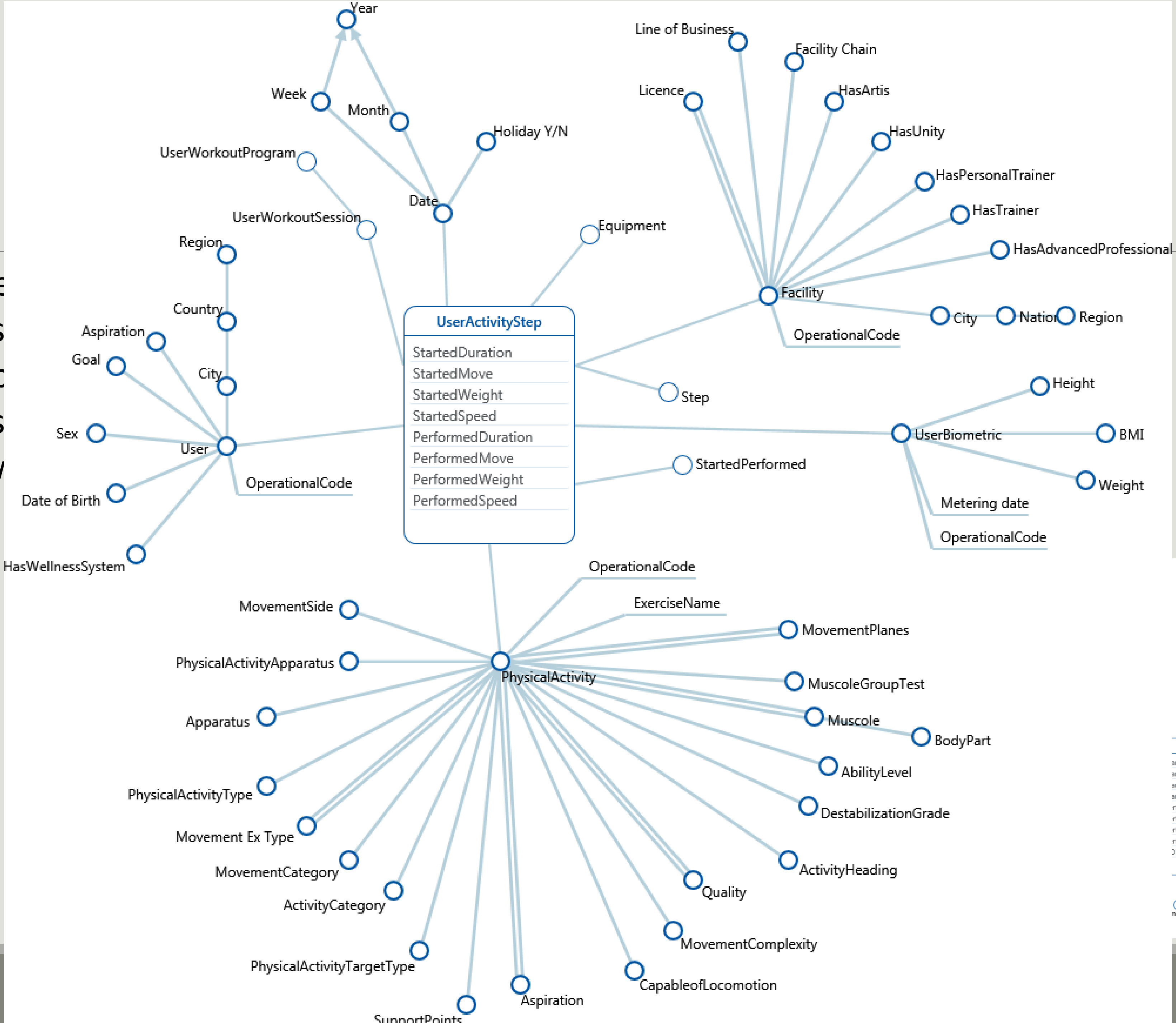


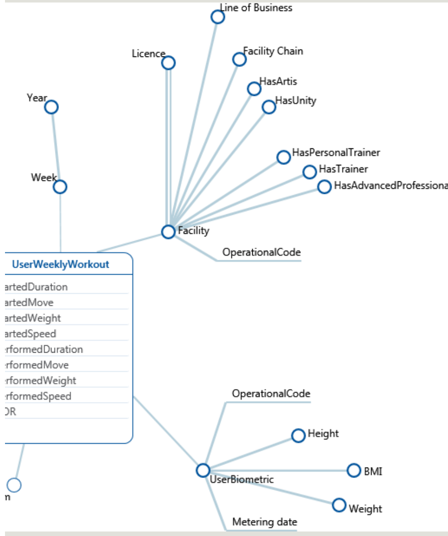


Data

4 cubes store

- For each s
- For each p
- For each s
- For each v

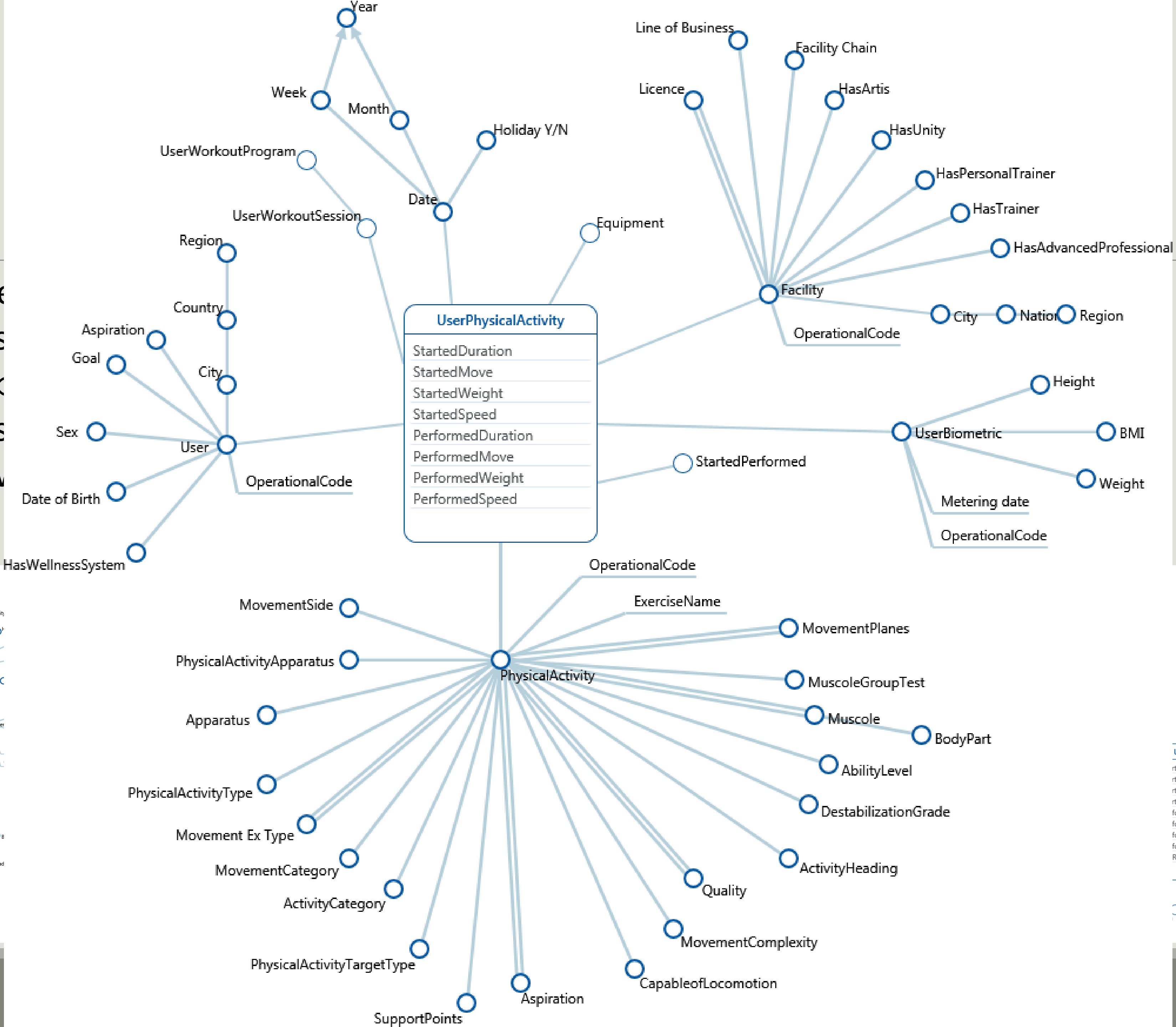


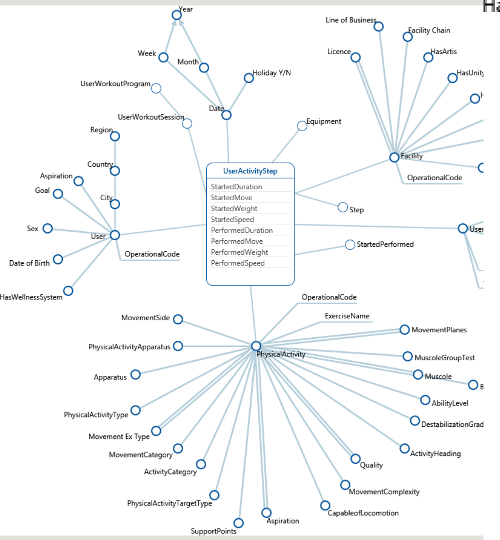


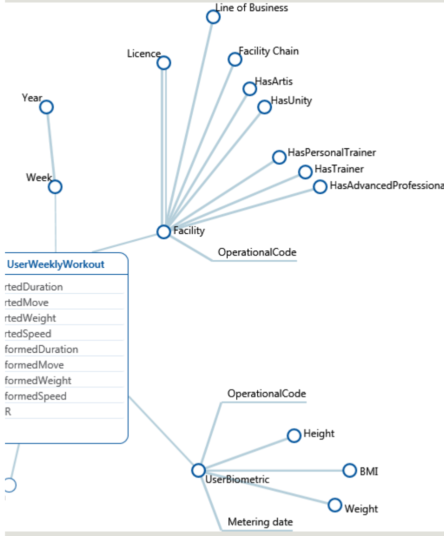
Data

4 cubes store

- For each s
- For each p
- For each s
- For each v



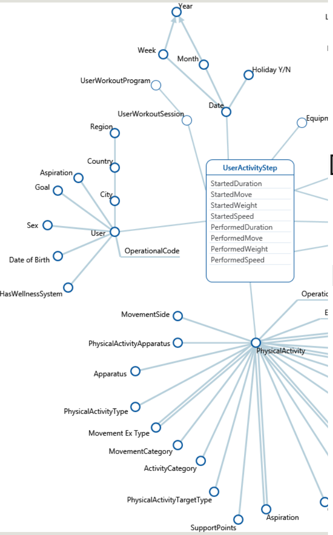


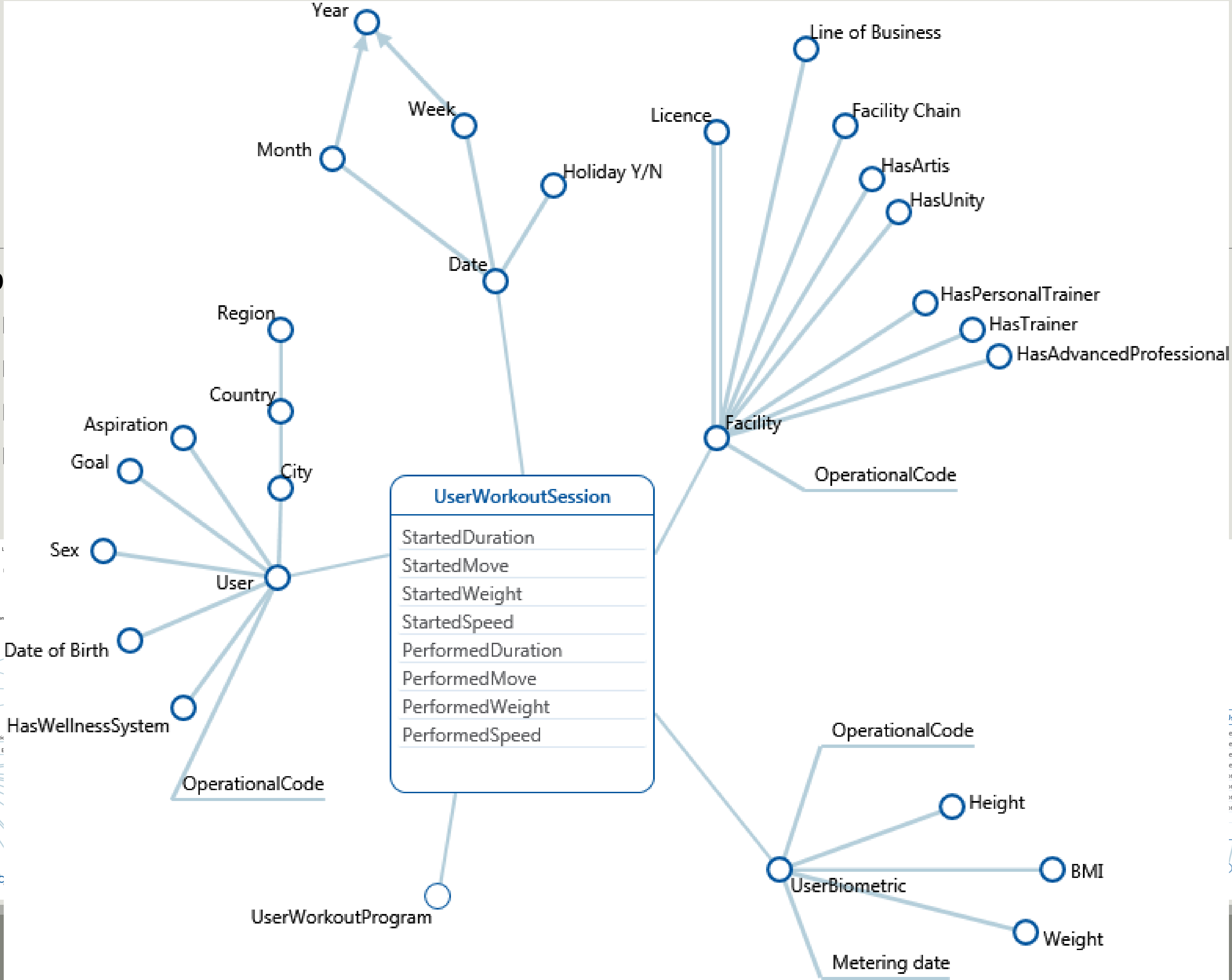


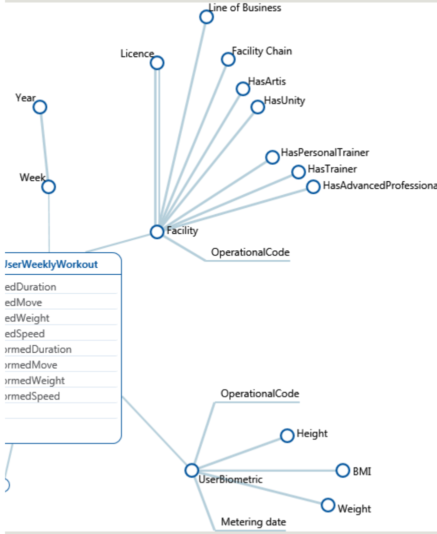
D

4 cub

- Fo
- Fo
- Fo
- Fo



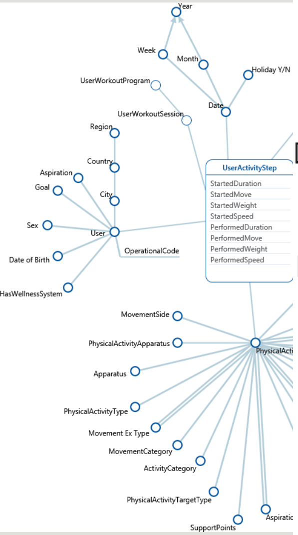


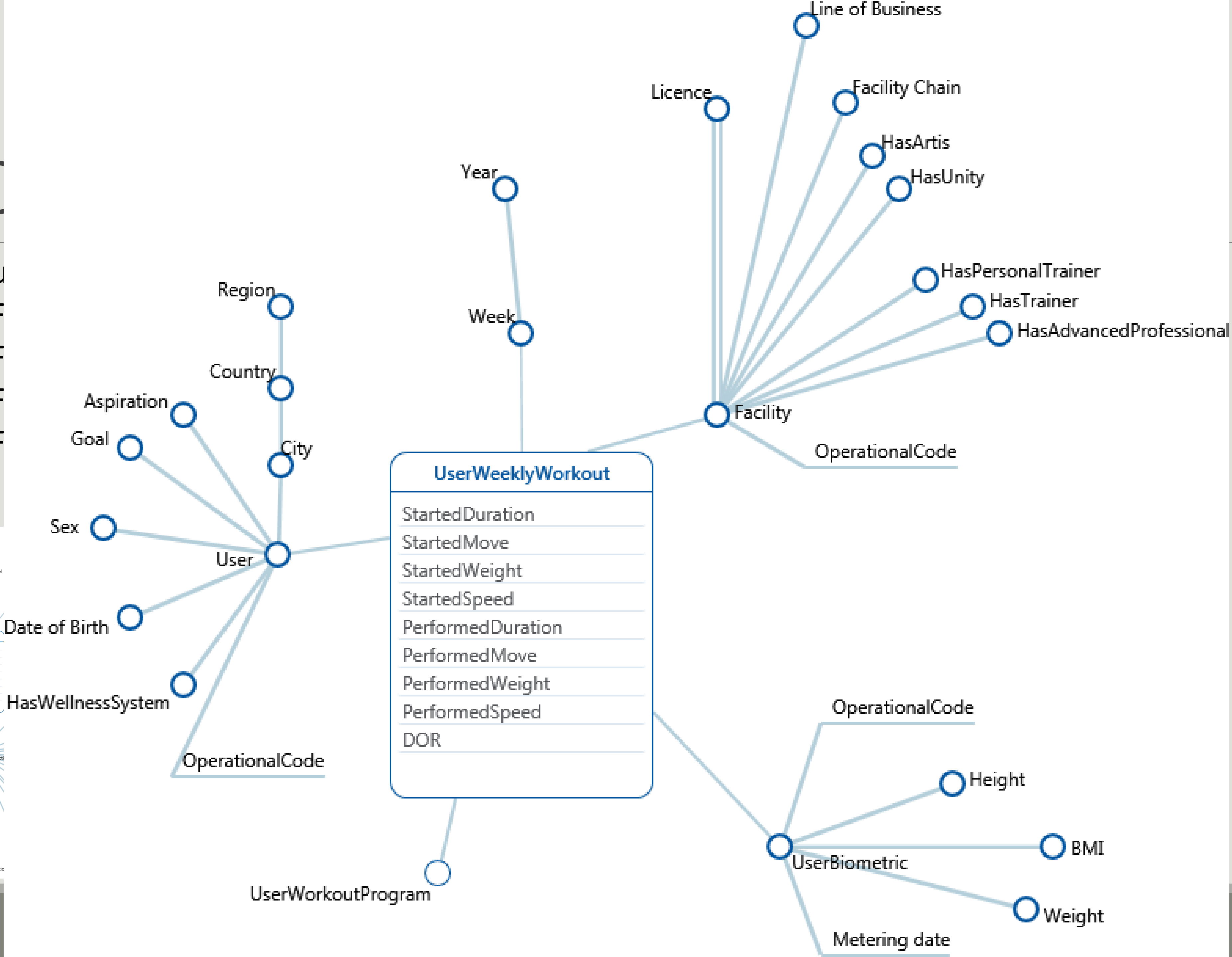


D

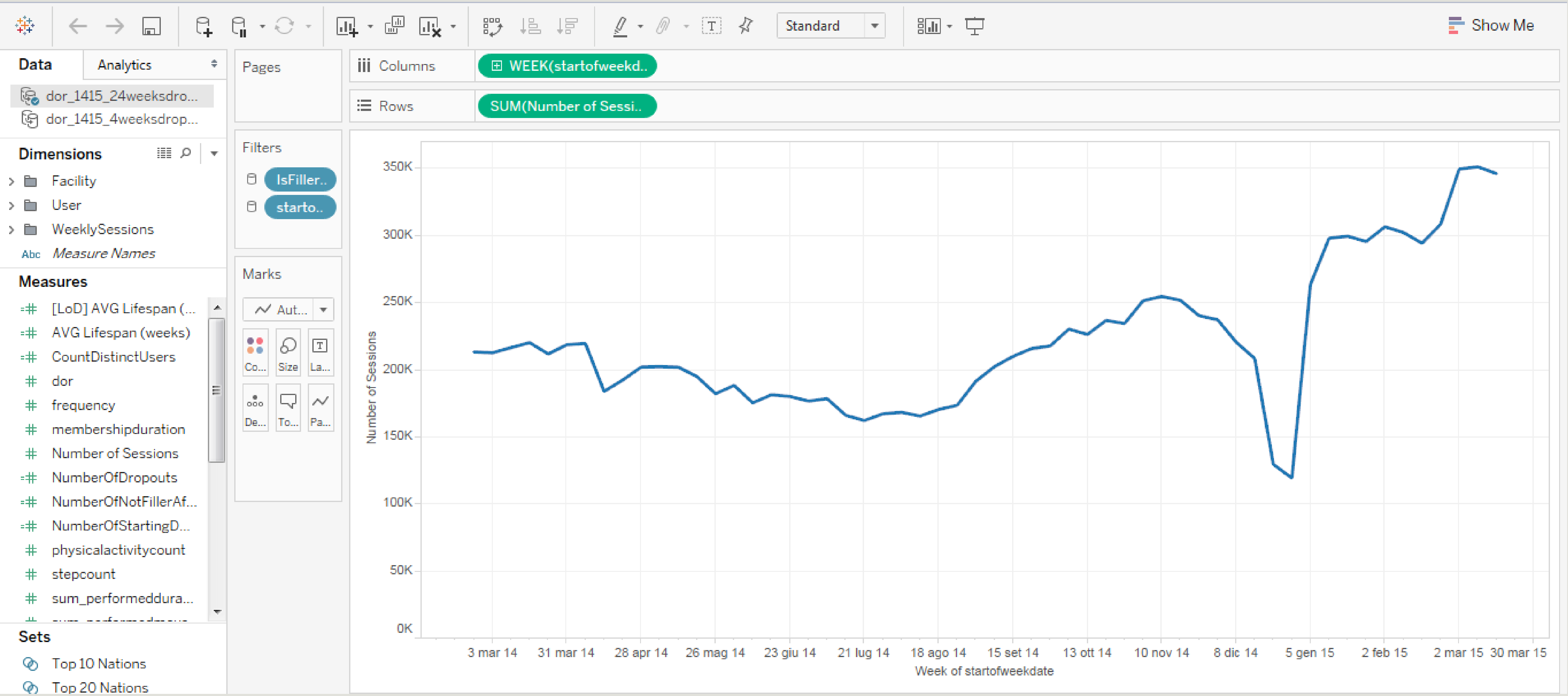
4 cu

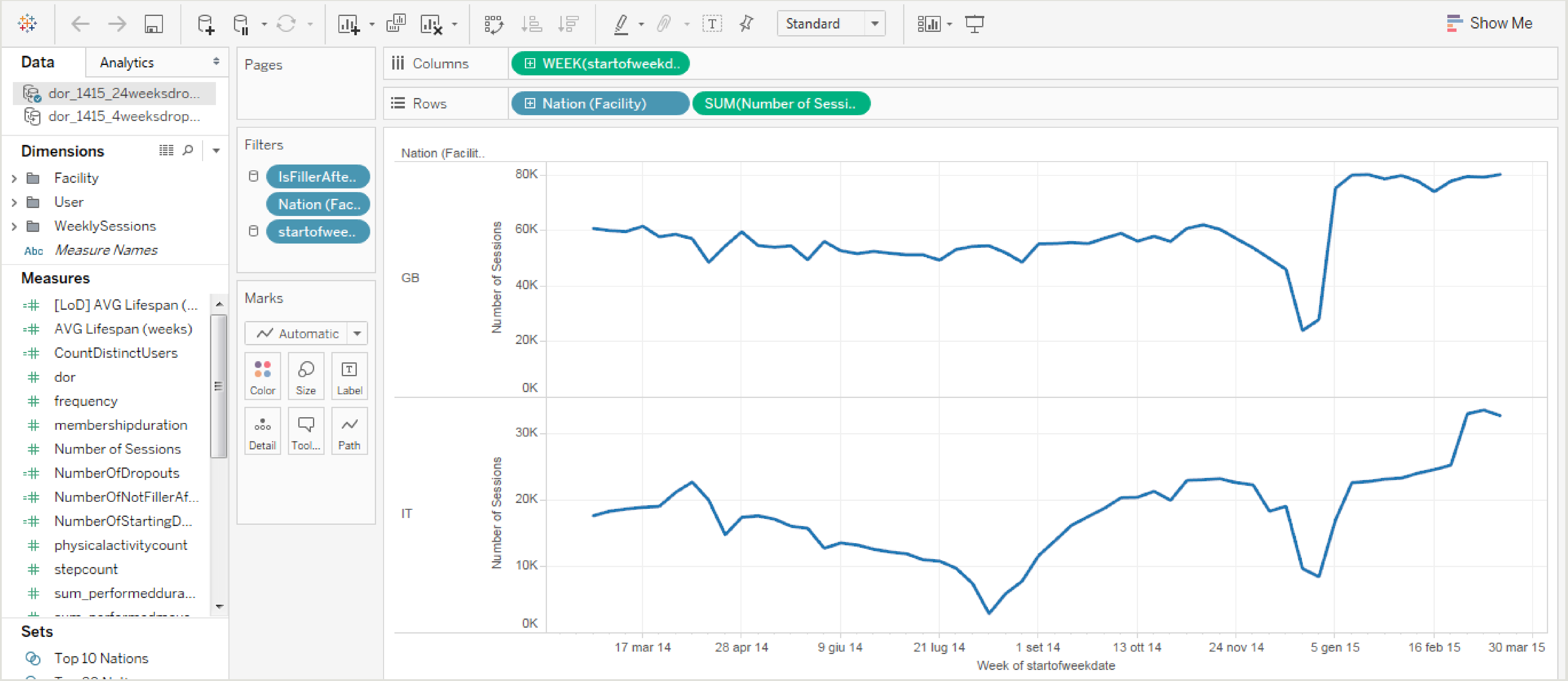
- F
- F
- F
- F





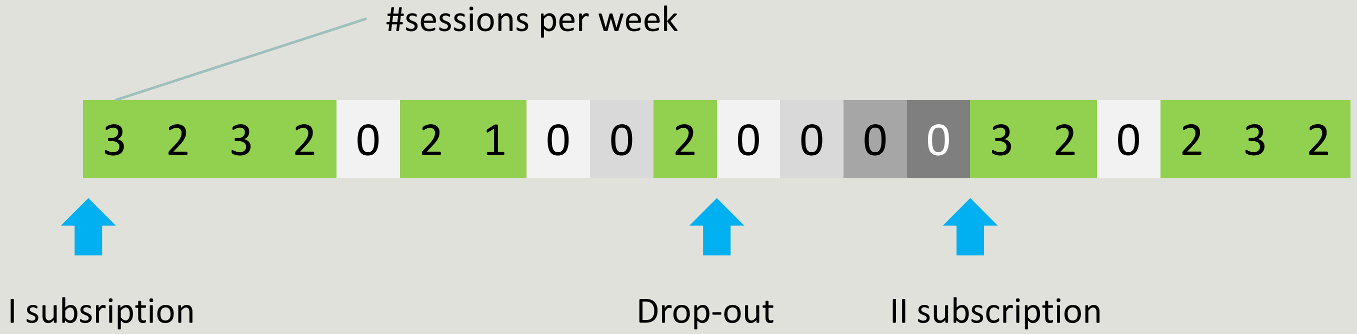
Line of Business
Licence
Facility Chain
HasArtis
HasUnity
Year
Week
HasPersonalTrainer
HasTrainer
HasAdvancedProfessional
Region
Country
Aspiration
Goal
City
Facility
OperationalCode
UserWeeklyWorkout
StartedDuration
StartedMove
StartedWeight
StartedSpeed
PerformedDuration
PerformedMove
PerformedWeight
PerformedSpeed
DOR
Sex
User
Date of Birth
HasWellnessSystem
OperationalCode
OperationalCode
Height
BMI
UserBiometric
Weight
UserWorkoutProgram
Metering date


Show Me
Data
Analytics
dor_1415_24weeksdro...
dor_1415_4weeksdrop...
Dimensions
Facility
User
WeeklySessions
Measure Names
Measures
[LoD] AVG Lifespan (...
AVG Lifespan (weeks)
CountDistinctUsers
dor
frequency
membershipduration
Number of Sessions
NumberOfDropouts
NumberOfNotFillerAf...
NumberOfStartingD...
physicalactivitycount
stepcount
sum_performeddura...
Sets
Top 10 Nations
Top 20 Nations
Pages
Filters
IsFiller..
starto..
Marks
Aut...
Columns
WEEK(startofweekd..
Rows
SUM(Number of Sessi..
Standard
Number of Sessions
Week of startofweekdate
350K
300K
250K
200K
150K
100K
50K
0K
3 mar 14
31 mar 14
28 apr 14
26 mag 14
23 giu 14
21 lug 14
18 ago 14
15 set 14
13 ott 14
10 nov 14
8 dic 14
5 gen 15
2 feb 15
2 mar 15
30 mar 15


Show Me
Standard
Data
Analytics
dor_1415_24weeksdro...
dor_1415_4weeksdrop...
Dimensions
Facility
User
WeeklySessions
Measure Names
Measures
[LoD] AVG Lifespan (...
AVG Lifespan (weeks)
CountDistinctUsers
dor
frequency
membershipduration
Number of Sessions
NumberOfDropouts
NumberOfNotFillerAf...
NumberOfStartingD...
physicalactivitycount
stepcount
sum_performeddura...
Sets
Top 10 Nations
Pages
Filters
IsFillerAfte..
Nation (Fac..
startofwee..
Marks
Automatic
Color
Size
Label
Detail
Tool...
Path
Columns
WEEK(startofweekd..
Rows
Nation (Facility)
SUM(Number of Sessi..
Nation (Facilit..
GB
IT
Number of Sessions
80K
60K
40K
20K
0K
30K
20K
10K
0K
17 mar 14
28 apr 14
9 giu 14
21 lug 14
1 set 14
13 ott 14
24 nov 14
5 gen 15
16 feb 15
30 mar 15
Week of startofweekdate

The Gym Case Study

A drop-out take place if the customer does not enter the the gym for 4 consecutive weeks



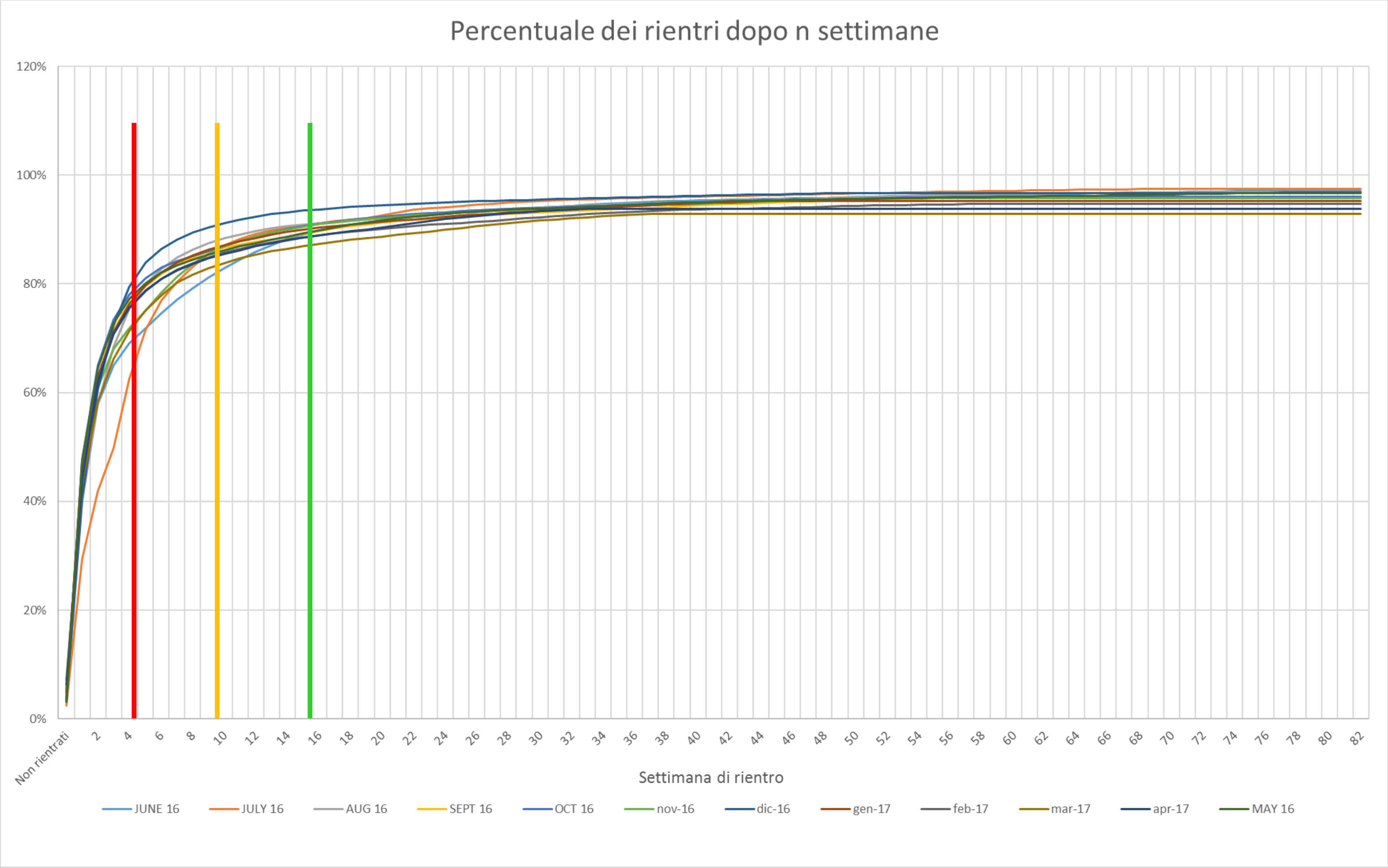
The same concept can be computed on a 6-month period basis to disregard seasonal behaviors

The previous definition is confirmed by data?

Is an absence of 4 weeks representative of a customer abandonment?

The Gym Case Study

A drop-out take place if the customer does not enter the the gym for 4 consecutive weeks



The sam … regard seasonal behavior

The prev

Is an abse

The Gym Case Study: Basic Assumption

Goal: classify users as willing or not willing to drop-out depending on their behavior in the gym.

The practitioner who is about to leave the gym is training poorly

The Gym Case Study: Basic Assumption

Goal: classify users as willing or not willing to drop-out depending on their behavior in the gym.

How can we characterize the user behaviors?

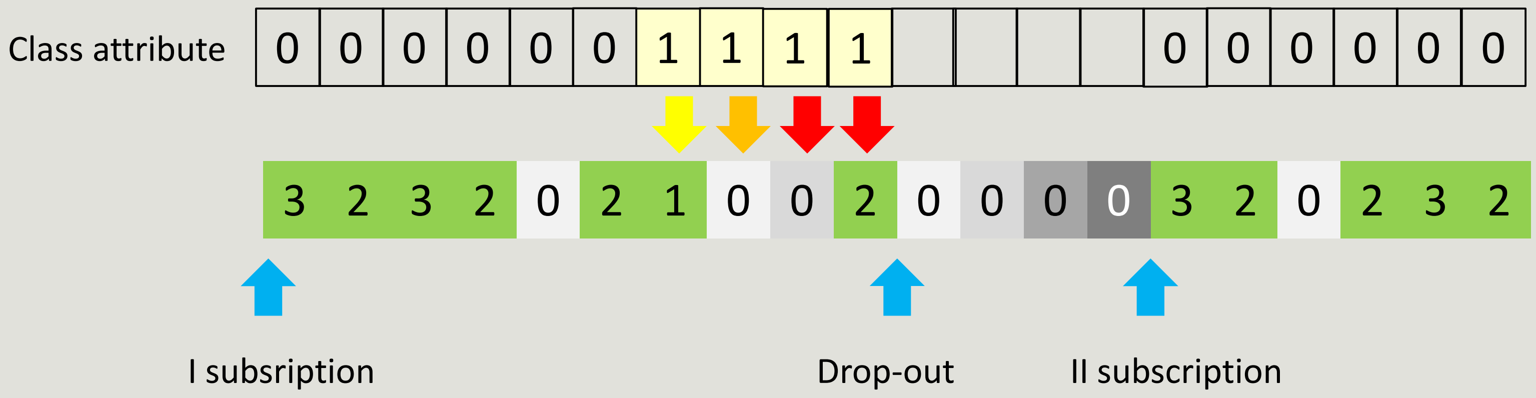
How long does it last?

The Gym Case Study: Basic Assumption

Goal: classify users as willing or not willing to drop-out depending on their behavior in the gym.

How can we characterize the user behaviors?

How long does it last?



Capturing the user behavior

Does the user train regularly?

Does the user respect the workout assigned to him?

We must quantify such qualitative questions through KPIs

- **Compliance**: quantifies the adherence of the performed workout to the assigned one
- **Regularity**: quantifies the regularity of the training sessions with reference to the prescribed one.

Some remarks

Behavior can be characterized at different granularity levels (steps, physical activities, sessions, weeks)

- It is not easy to understand which is the best granularity level: a very detailed one could be blurred by noise of unintersting details. A coarse one could not capture the behavioral changes.

Time plays a major role in understanding the user behavior

- This implies considering the sequence of workouts rather than the single workout
- Sequence mining is not trivial and reduce the number of techniques to be adopted

Compliance: Current State

Before our project, compliance was computed through two KPIs. The first one is based on **MOVE**:

$$Compliance_{MOVE}(AWS, PWS) = \frac{\sum_{i \in PWS} MOVE_i}{\sum_{j \in AWS} MOVE_j}$$

where *AWS* and *PWS* are the sets of training sessions assigned and performed by the user, respectively.

- Does not evaluate regularity
- A user could perform completely different PAs from the assigned ones retaining a compliance = 1
- Compensations are possible
- Compliance can be > 1

The second one is based on the **number of sessions** included in the training program:

$$Compliance_{WS}(AWS, PWS) = \frac{|PWS|}{\frac{AWS\ per\ week}{7} \cdot \#days\ from\ the\ AWS\ begin}$$

Compliance: desiderata

- Ranging in [0,..1]
- Being 1 only if the user performs all and only the assigned exercise
- Compensations are not allowed
- Making possible to understand which exercises are overdone/leftover

Session Evaluation Example

PA	MUSCLE	ASSIGNED (MOVE)	PERFORMED (MOVE)
1	M1	40	30
2	M2	85	85
3	M1	0	40

$$Correct = \min(Assigned, Performed)$$

$$Leftover = \max(0, Assigned - Correct)$$

$$Overdone = \max(0, Performed - Correct)$$

PA	MUSCLE	ASSIGNED (MOVE)	PERFORMED (MOVE)	LEFTOVER	CORRECT	OVERDONE
1	M1	40	30	10	30	0
2	M2	85	85	0	85	0
3	M1	0	40	0	0	40

Compliance & Deviation

Compliance: describes ***how much*** the user has adhered to the AWS

$$Compliance = \frac{Correct}{Leftover + Correct + Overdone}$$

Deviation: describes ***how*** the user deviates from the AWS

- *Leftover*: Occurs when a user does less than the assigned
- *Overdone*: Occurs when a user does more than the assigned

$$Deviation = Overdone - Leftover$$

Compliance & Deviation: PA

PA	MUSCLE	STARTED (MOVE)	PERFORMED (MOVE)	LEFTOVER	CORRECT	OVERDONE
1	M1	40	30	10	30	0
2	M2	85	85	0	85	0
3	M1	0	40	0	0	40

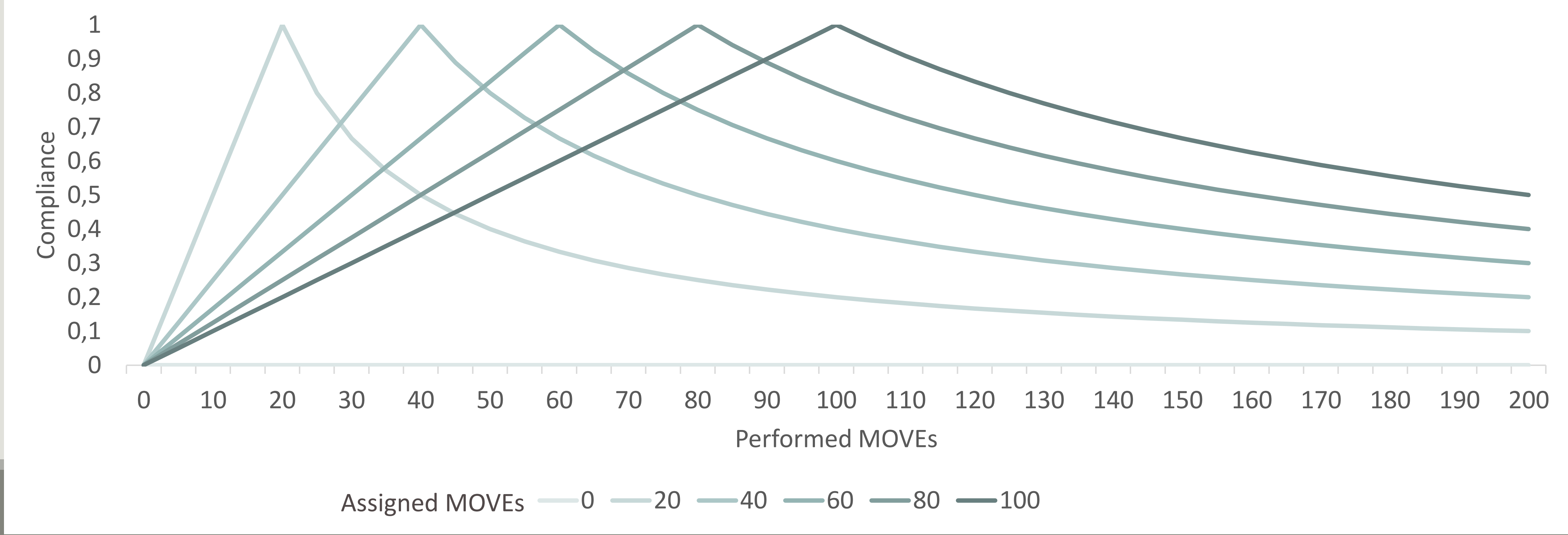
Overdone - Leftover

PA	COMPLIANCE	Deviation
1	0.75	
2	1	
3	0	

$$Compliance = \frac{Correct}{Leftover + Correct + Overdone}$$

Compliance definition

$$Compliance = \frac{Correct}{Leftover + Correct + Overdone} = \begin{cases} \dfrac{Assigned}{Performed} & if\, Assigned < Performed \\ \dfrac{Performed}{Assigned} & else \end{cases}$$



Compliance & Deviation: Muscle

PA	MUSCLE	ASSIGNED (MOVE)	PERFORMED (MOVE)	LEFTOVER	CORRECT	OVERDONE
1	M1	40	30	10	30	0
2	M2	85	85	0	85	0
3	M1	0	40	0	0	40

Grouping (sum) per *muscle*

MUSCLE	ASSIGNED (MOVE)	PERFORMED (MOVE)	LEFTOVER	CORRECT	OVERDONE
M1	40	70	0	40	30
M2	85	85	0	85	0

MUSCLE	COMPLIANCE	DEVIATION
M1	0.57	→
M2	1	•

Regularity

Regularity: it is computed at the week granularity

- Differently from compliance, it can be > 1

$$Regularity = \frac{\#\ PWS\ in\ the\ current\ week}{\#\ AWS\ per\ week} \in [0, +\infty)$$

Week	1/2015	2/2015	3/2015	4/2015	5/2015	6/2015	7/2015	8/2015	9/2015	10/2015
# Performed Workouts	2	2	1	3	3	0	4	1	2	2
# Assigned Workouts	2	2	2	2	2	2	2	2	2	2
Regularity	1.0	1.0	0.5	1.5	1.5	0.0	2.0	0.5	1.0	1.0

Capturing the user behavior along time

Compliance and Regularity quantify the user behavior at a specific time (state), but they do not campure the behavioral changes along time. Possibile solultions are

- Analyze sequences of user behavior states
- Include in the user behavior state the behavioral changes along time, and analyze a single state at a time

To include behavioral changes into the current user state the *"variation concept"* can be adopted

$$ComplianceVar = \frac{Compliance}{AVG(Compliance, 4\ previous\ weeks)}$$

$$RegularityVar = \frac{Regularity}{AVG(Regularity, 4\ previous\ weeks)}$$

Capturing the user behavior along time

Compliance and Regularity quantify the user behavior at a specific time (state), but they do not capture the behavioral changes along time. Possibile solultions are

- Analyze sequences of user behavior states
- Include in the user behavior state the behavioral changes along time, and analyze a single state at a time

Alternatively a rolling average along time captures a long term behavior rather than a punctual one:

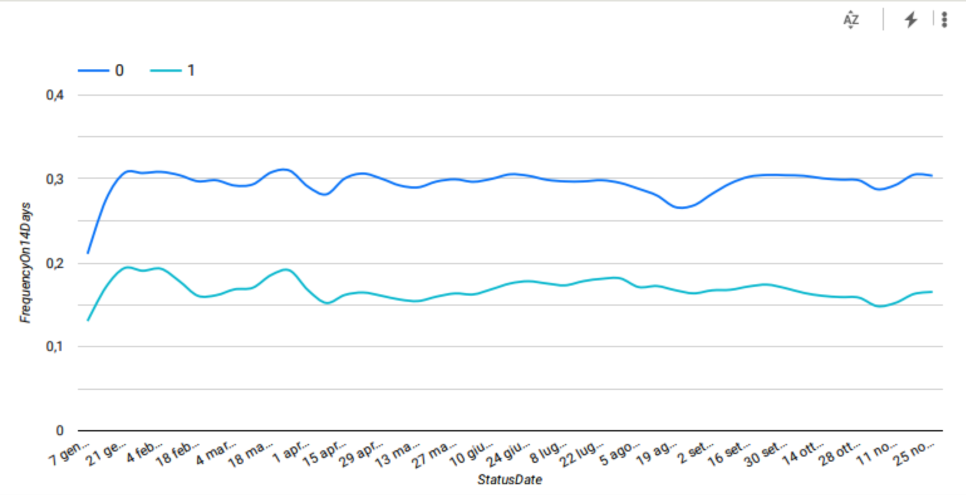
$$ComplianceRoll2 = AVG(Compliance, 2\ previous\ weeks)$$

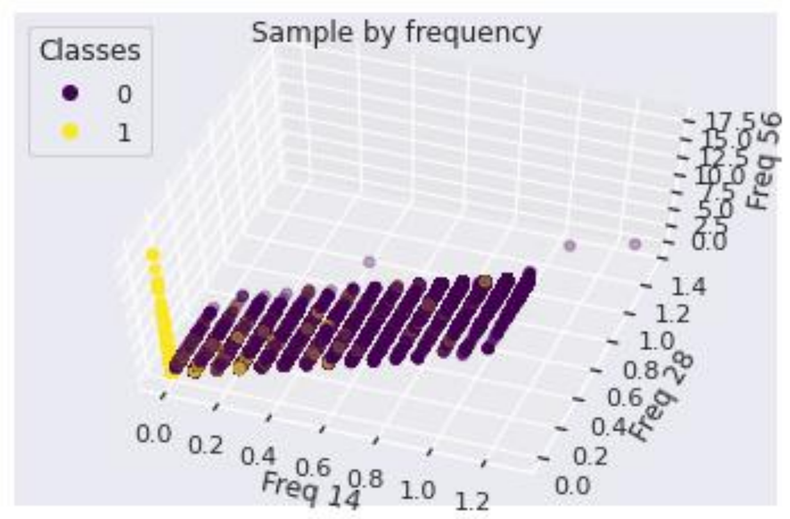
$$RegularityVar = AVG(Regularity, 2\ previous\ weeks)$$

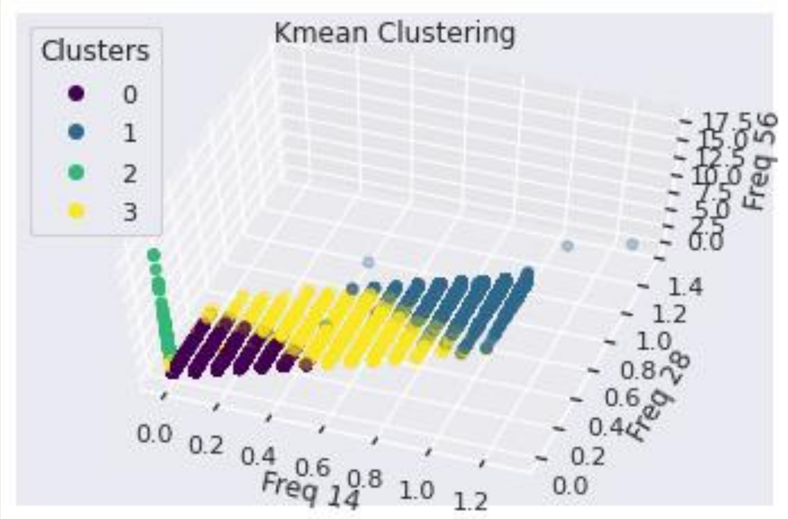
The proper duration along time must be tested on data: longer periods capture more stable behaviors and hide spot ones but require more time to emphasize behavioral change

Basic Hypothesis Testing

Do dropping users have a lower frequency?







DropoutIndex=1
in Cluster=0 : 64.43%
in Cluster=3 : 27.5%
in Cluster=1 : 5.67%
in Cluster=2 : 2.41%

DropoutIndex=0
in Cluster=1 : 37.47%
in Cluster=0 : 32.41%
in Cluster=3 : 30.12%
in Cluster=2 : 0.0%

The User Weekly State

At the end of each week a user state can be defined. Historical data can be labeled with a DOR (True/False) attribute to train the model

Name	Description
started_startofweekdate	State week
id_user	User ID
id_facility	Facility ID
count_session	Weekly performed sessions
count_step	Weekly performed steps
count_physicalactivity	Weekly performed PA
sum_assignedmove	Weekly assigned move
sum_performedduration	Weekly performed minutes
sum_performedweight	Weekly performed KGs
avg_session_length	AVG performed minutes per session
pa_compliance	Compliance computed at the PA level
muscle_compliance	Compliance computed at the muscle level
execise_quality_compliance	Compliance computed at the step level
Month	Month of the year
user_sex	Male/Female
user_age	Age of the user

Name	Description
facility_nation	IT for all the tuples
facility_hasunity	True if the Facility has the top level Technogym console
facility_hasartis	True if the Facility has the top level Technogym product line
WeeksSinceLastSession	#of weeks passed from the previous session
WeeksToNextSession	#of weeks before the next session
weeks_since_membership	# of consecutive weeks of memberships
id_membership	Membership ID
cum_count_session	#of session since the begin of the membership
weeks_to_drop	# of weeks before drop out
count_session_rolling_sum2	Sum of sessions performed in the last 2 weeks
count_session_rolling_sum8	Sum of sessions performed in the last 8 weeks
rolling_frequency2	AVG frequency in the last 2 weeks
pa_compliance_rolling_avg2	AVG frequency in the last 8 weeks
pa_compliance_rolling_avg8	AVG PA-level compliance in the last 8 weeks
muscle_compliance_rolling_avg2	AVG muscle -level compliance in the last 2 weeks
muscle_compliance_rolling_avg8	AVG muscle -level compliance in the last 8 weeks
DropFlag	FALSE/TRUE

The Gym Case Study

Open the Gym.arff file and create a model to identify the user that are willing to drop out.

The users will be reported to the gym manager so that an action to retain the customers can be carried out.

- The number of False Positive must be also minimized in order to minimize the gym manager effort