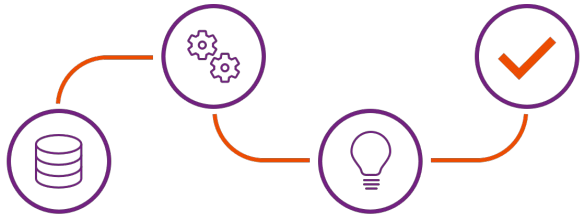
SEMANTiCS
Karlsruhe 2019



SIMPLE-ML

Towards a Framework for Semantic Data Analytics Workflows

Simon Gottschalk[1], Nicolas Tempelmeier[1], Günter Kniesel[2], Vasileios Iosifidis[1], Besnik Fetahu[1], Elena Demidova[1]

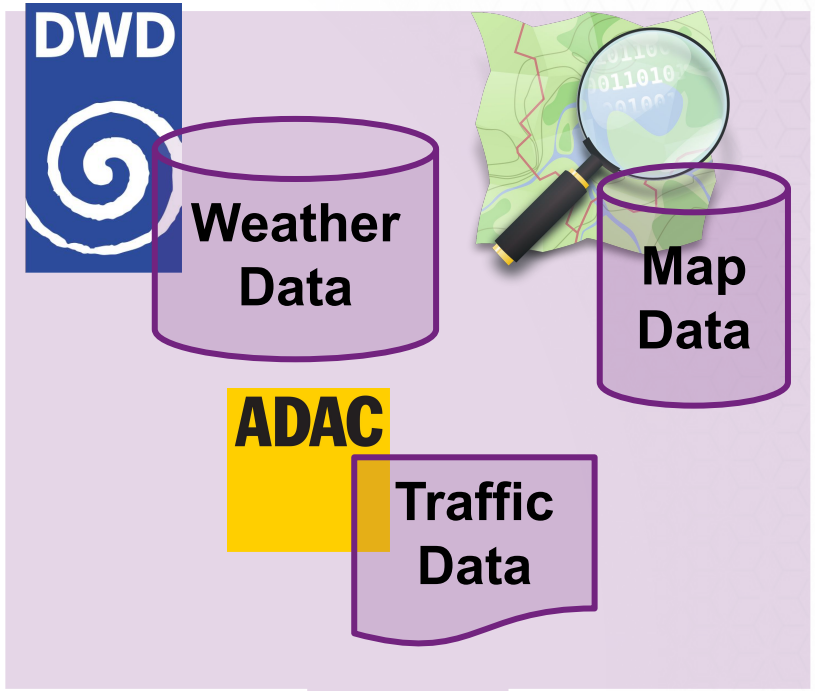
[1] L3S Research Center, Leibniz Universität Hannover

[2] Smart Data Analytics Group (SDA), Universität Bonn

Gefördert vom Bundesministerium für Bildung und Forschung

Motivation & Goal

- Data analytics demands
  - Data science expertise
  - Knowledge about data access, data integration, feature extraction, …
- Many concepts and operations are domain-specific
  - Map matching: connect position data with street segments
- Semantics can support these tasks and simplify data analytics for non-expert users

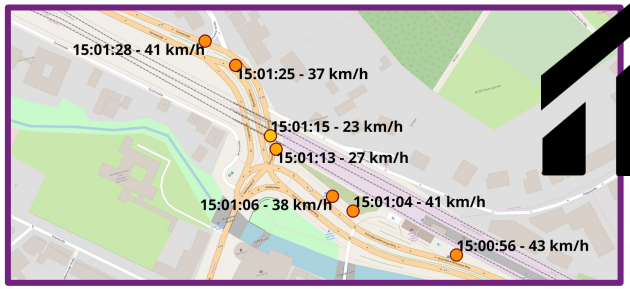


Feature Extraction

Map Matching

Data Materialisation

?



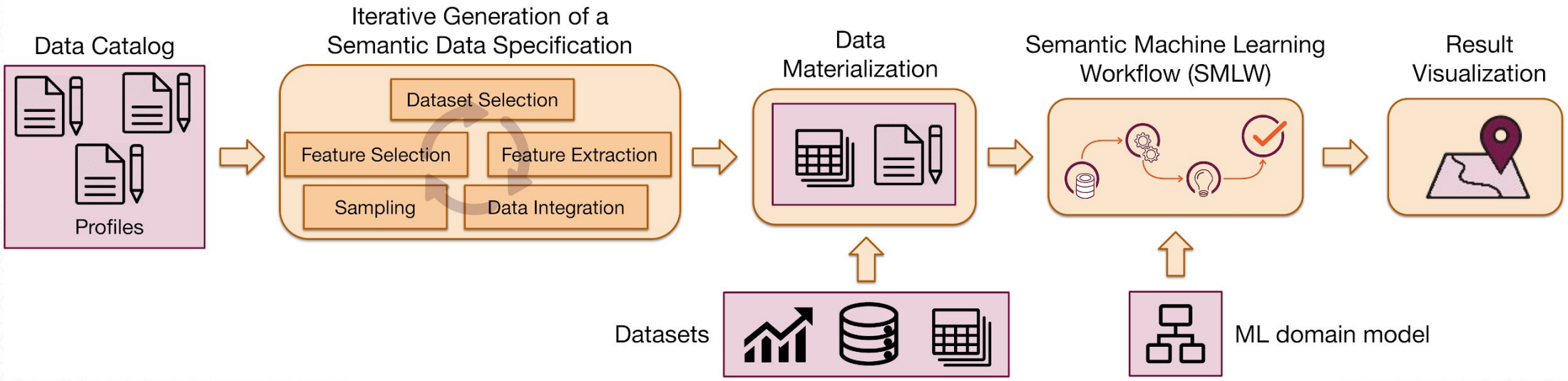
Traffic speed prediction

Adopt semantic technologies to support the
- efficient creation,
- configuration
- and reusability

of robust data analytics workflows

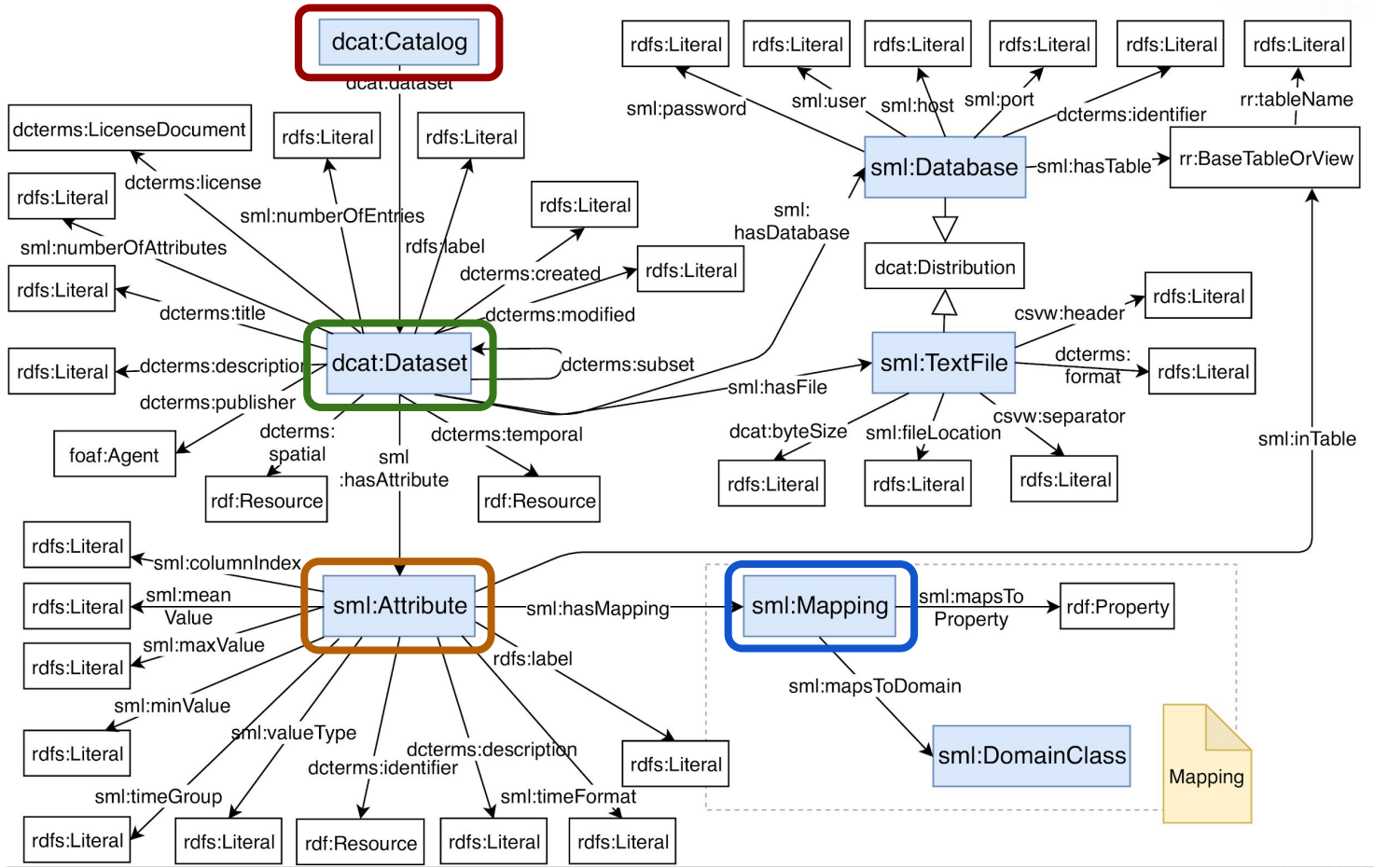
Semantic Data Analytics Workflow

- A semantic data specification is generated based on a data catalog and user-selected operations
  - Dataset selection
  - Feature selection and extraction
  - Dataset integration
  - Sampling
- The actual data is materialised later
- Potential subsequent machine learning and result visualisation steps make use of semantic models



Data Catalog & Dataset Profiles

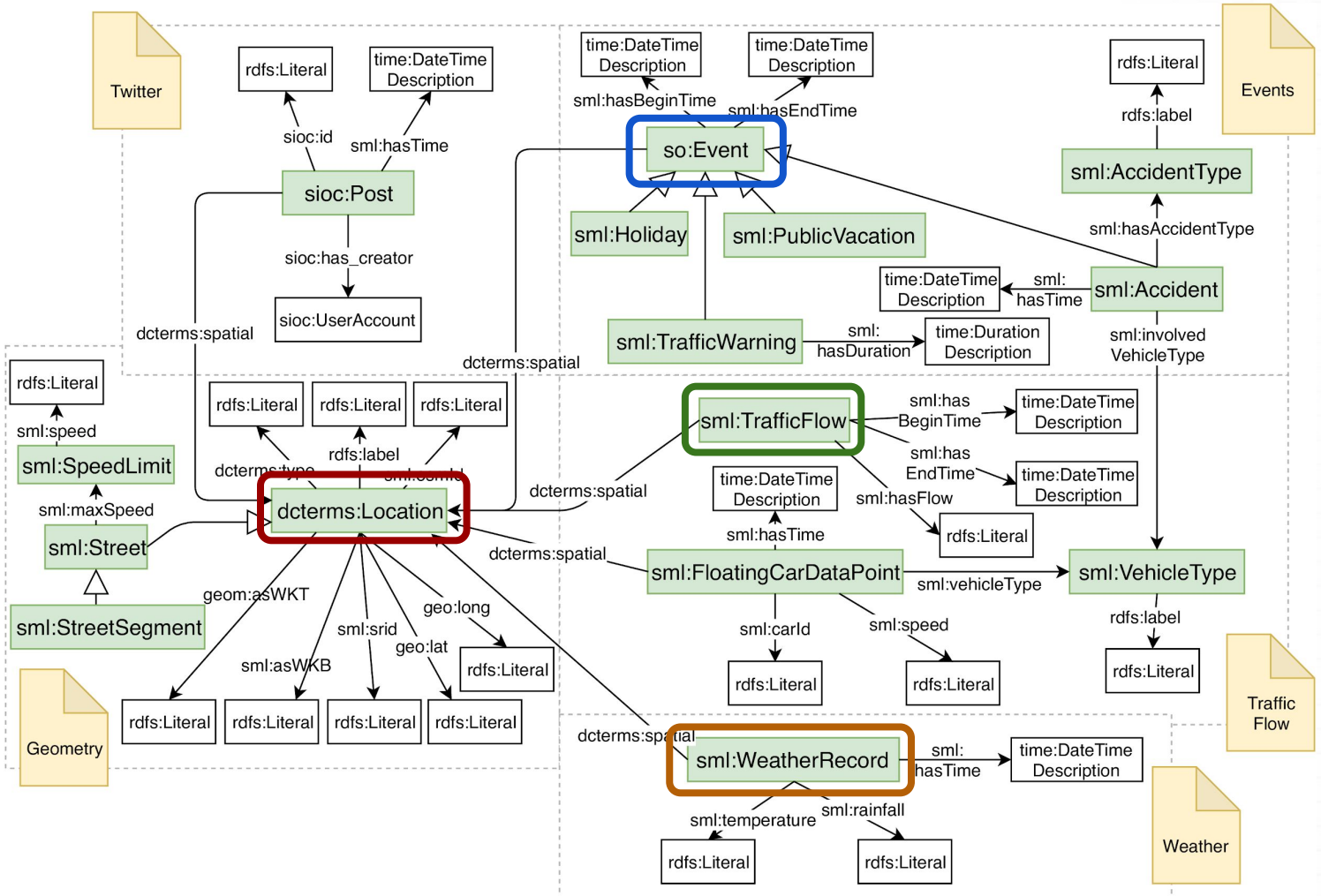
- A ***dataset profile*** represents dataset characteristics
  - statistics, license, data access information, …
- A domain-specific ***data catalog*** contains dataset profiles
- The **attributes** of a dataset are **mapped** to a domain model



Data Catalog Schema

Domain Model — Example: Mobility

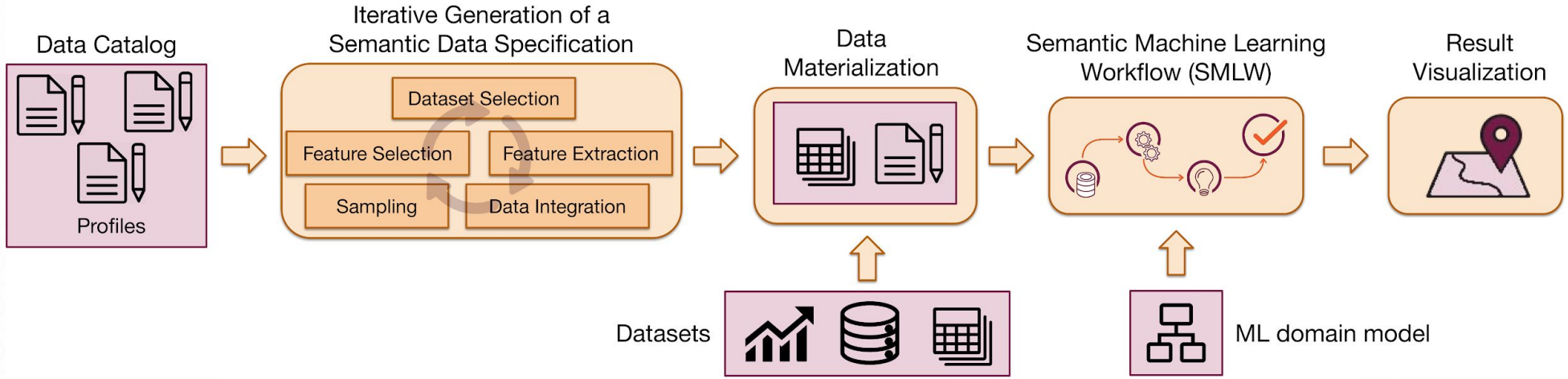
- description of relevant concepts, their properties and relations in the specific application domain
- Example: Mobility
  - Locations
    - Position representations
  - Traffic statistics
    - Floating car data, …
  - Events
    - Accidents, warnings, …
  - Weather
    - Rainfall, temperature
  - …
- Reuse of existing vocabularies if possible



Example domain model (mobility)

Benefits of Semantics in Data Analytics

- The data catalog with domain-specific dataset profiles is used …
  - … to support the user in generating a data specification
    - extraction of domain-specific features
    - suggest semantically meaningful joins
  - … to materialise the data
    - data access information in the catalog schema
  - … to configure the machine-learning workflow
    - type checking based on the domain model
  - … to visualise results
    - domain-specific visualisations (e.g. rainfall vs. temperature)

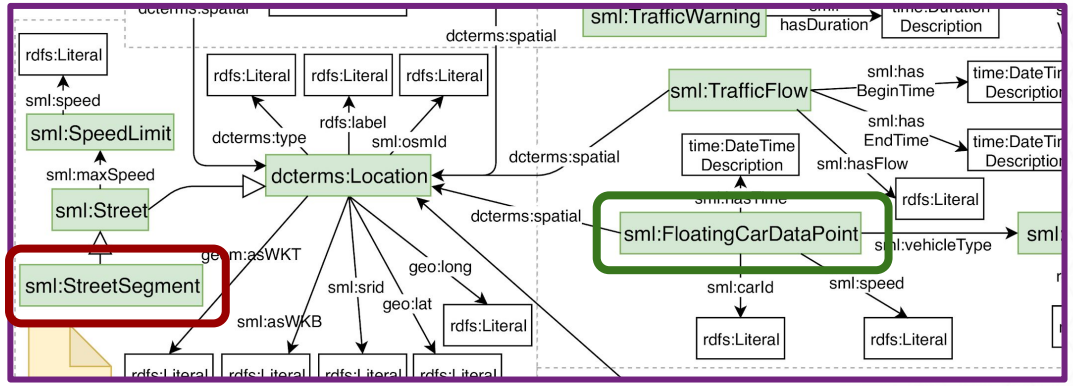


Example Data Catalog in the Mobility Domain

```
sml:SimpleMLCatalog        a dcat:Catalog ;
                           dcat:dataset sml:FCDDataset .

sml:FCDDataset             a dcat:Dataset ;
                           dcterms:title "Floating Car Data" ;
                           dcterms:temporal [
                              so:startDate "2017-08-01"^^xsd:date ;
                              so:endDate "2017-12-31"^^xsd:date ] ;
                           sml:hasAttribute sml:FCDDatasetAttribute1 .

sml:FCDDatasetFile         a sml:TextFile ;
                           dcterms:format "text/comma-separated-values" ;
                           csvw:separator ";" .

sml:FCDDatasetAttribute1   a sml:Attribute ;
                           rdfs:label "vehicle id"@en ;
                           sml:columnNumber "0"^^xsd:integer ;
                           sml:hasMapping [
                              sml:mapsToProperty sml:carId ;
                              sml:mapsToDomain sml:FloatingCarDataPoint ] .
```

Example: Traffic Speed Prediction (Mobility Domain)

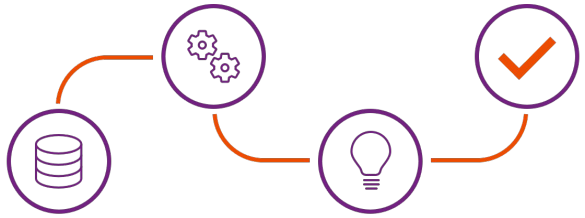
- Task: Traffic speed prediction for a specific street segment at a given time
- The user selects two datasets:
  - Floating car data (historic positions and speed of vehicles)
  - OpenStreetMap (map data with street segments)
- The user selects and extracts a set of features
- Data materialisation based on the resulting data specification



Floating Car Data Point				Street Segment	
Type	**Speed**	**Time (day)**	**Time (hour)**	**Type**	**Max Speed**
Truck	74	Sunday	23	motorway link	80
Car	84	Sunday	16	motorway	*none*
Truck	17	February	8	secondary	70

Example instances

SEMANTiCS
Karlsruhe 2019



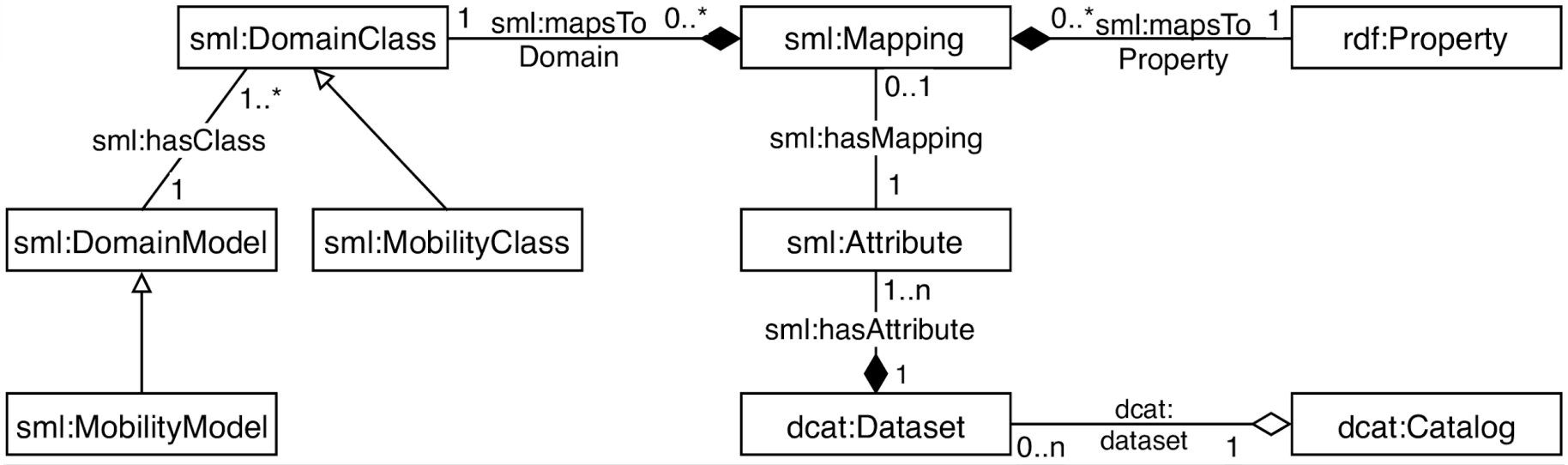
SIMPLE-ML

Towards a Framework for Semantic Data Analytics Workflows

Thank you!

Questions?

Forschungszentrum L3S Research Center
SMART DATA ANALYTICS FROM DATA TO KNOWLEDGE
AIM
PROJEKTIONISTEN®
Gefördert vom Bundesministerium für Bildung und Forschung


sml:DomainClass
1
sml:mapsTo
Domain
0..*
sml:Mapping
0..*
sml:mapsTo
Property
1
rdf:Property
1..*
sml:hasClass
1
0..1
sml:hasMapping
1
sml:DomainModel
sml:MobilityClass
sml:Attribute
1..n
sml:hasAttribute
1
sml:MobilityModel
dcat:Dataset
0..n
dcat:
dataset
1
dcat:Catalog

Example Query: Attribute Selection

```
SELECT
   ?columnNumber ?attrName ?mapProperty ?mapDomain
WHERE {
   sml:FCDDataset sml:hasAttribute ?attribute .
   ?attribute dcterms:identifier ?attrName .
   ?attribute sml:columnNumber ?columnNumber .
   OPTIONAL {
      ?attribute sml:hasMapping [
            sml:mapsToProperty ?mapProperty ;
            sml:mapsToDomain ?mapDomain ;
         ] .
   }
}
```