Semantic geographic knowledge on a world scale – interlinking OpenStreetMap and knowledge graphs

Prof. Dr. Elena Demidova

GISRUK Online Seminar Series 2021

25 March 2021

Introduction

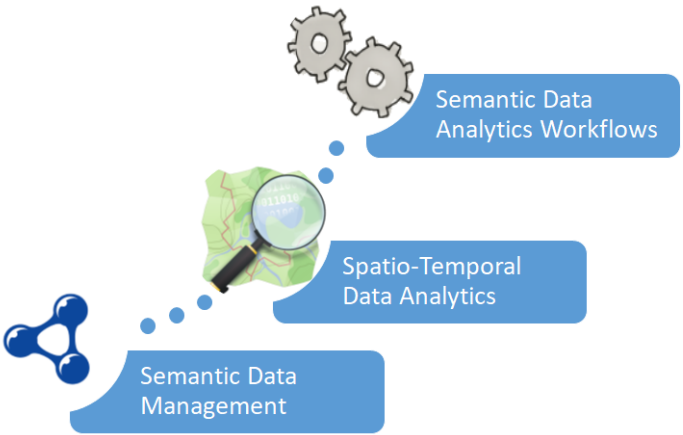
Prof. Dr. Elena Demidova

- Professor of Computer Science
  Data Science & Intelligent Systems (DSIS)
  Computer Science Institute
  University of Bonn

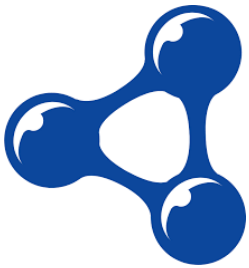
  `http://dsis.iai.uni-bonn.de`
- Research areas: Computer Science → Artificial Intelligence; Semantic Data Management and Analytics

Vision: Semantic Information Spaces for Data Analytics

- Semantic data management
  - Knowledge graphs and semantic technologies as a core
  - Integration of open spatio-temporal, semantic and contextual data
  - Semantic technologies for data governance
- Spatio-temporal data analytics
  - Robust data analytics methods
  - Applications to mobility, smart cities
- Semantic data analytics workflows
  - Intuitive, robust and reusable



Semantic Data Management



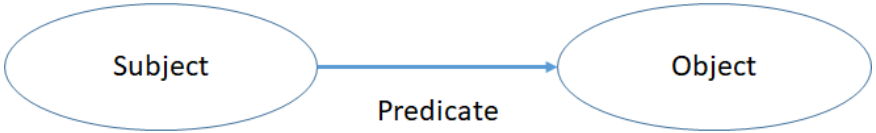
Knowledge Graphs

Knowledge graphs, i.e. semantic / RDF knowledge bases are increasingly important in data analytics, e.g. in

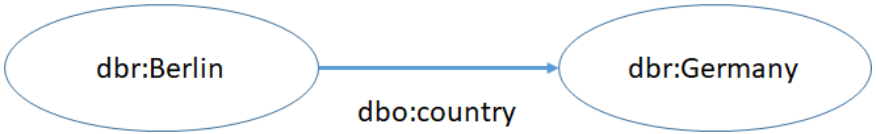
- Search engines, recommender systems, dialog systems, ...
- Popular knowledge graphs:
  - Google knowledge graph in 2016: 570+ million entities
  - Wikidata: > 60 million data items in 2019
  - DBpedia: > 4.5 million data items in English Version (2019)

RDF Knowledge Graphs

RDF = Resource Description Framework
Standard data model for data exchange on the Web (W3C standard)
A fact (triple) consists of a node-edge-node



Example:

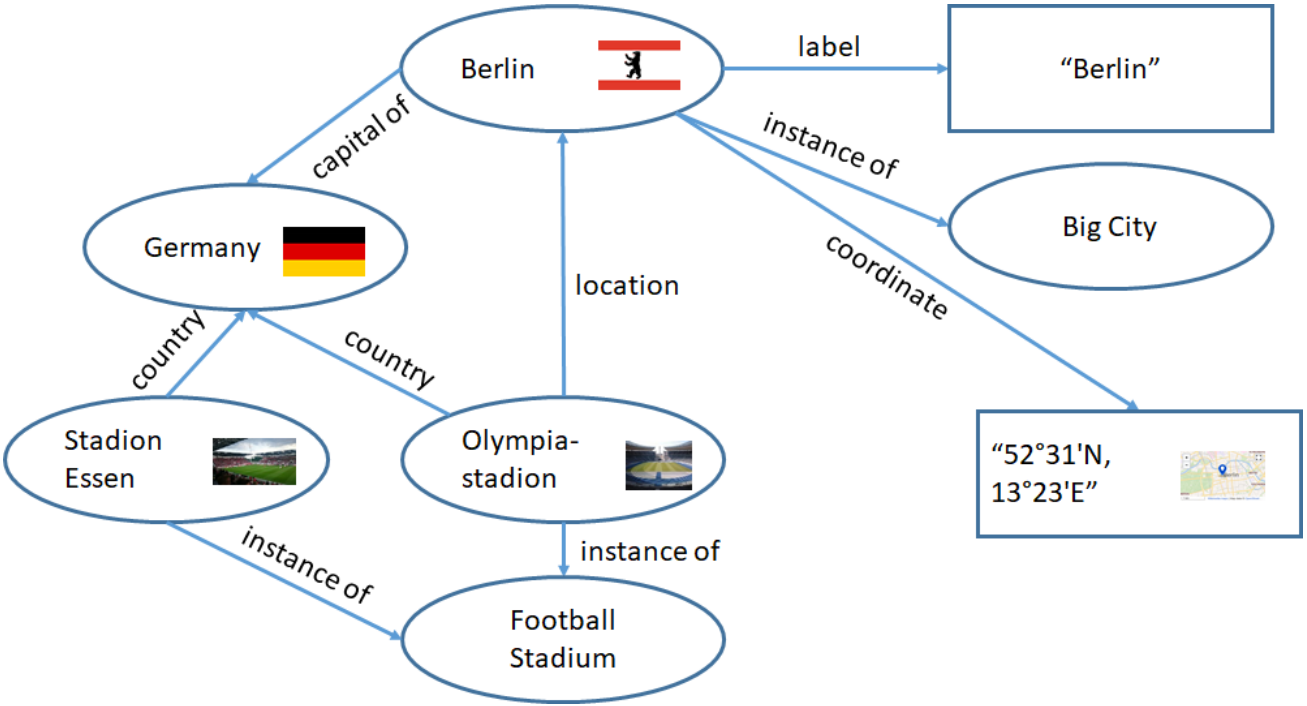


URLs as unique identifiers of nodes and edges

- dbo: http://dbpedia.org/ontology/
- dbr: http://dbpedia.org/resource/

Semantics defined using ontologies: dbo="DBpedia ontologie“

RDF Knowledge Graph - an Example



EventKG Knowledge Graph for Events

Problem: Existing knowledge graphs are focused on entities; events are underrepresented

EventKG knowledge graph → a multilingual reference dataset for events and temporal relations

- Extraction, interlinking and fusion of events from
  - Wikidata, DBpedia, YAGO, Wikipedia
- > 1.3 million events, > 4.5 million temporal relations, 15 languages
- Knowledge graph completion: events and event relations

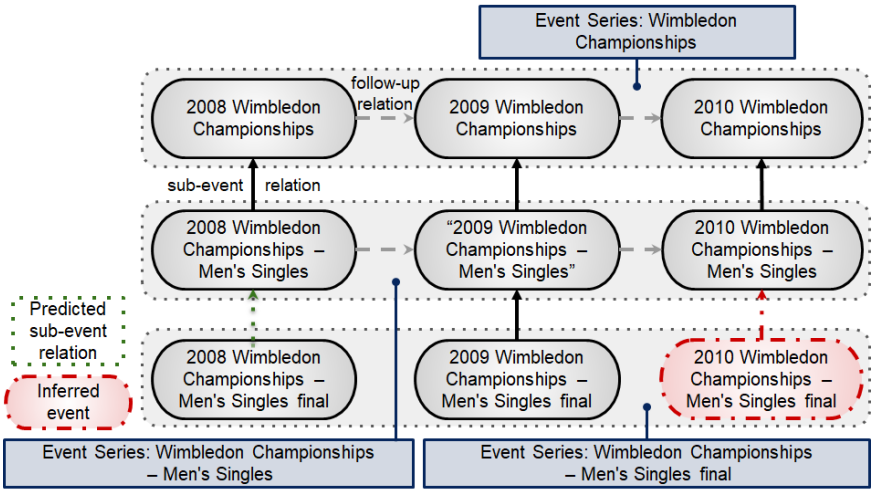
`http://eventkg.l3s.uni-hannover.de/`

S. Gottschalk and E. Demidova. EventKG – the Hub of Event Knowledge on the Web – and Biographical Timeline Generation. Semantic Web 10(6):1039-1070 (2019)

★ CLEOPATRA. Marie Skłodowska-Curie ITN, 2019-2022.

Knowledge Graph Completion: Event Series

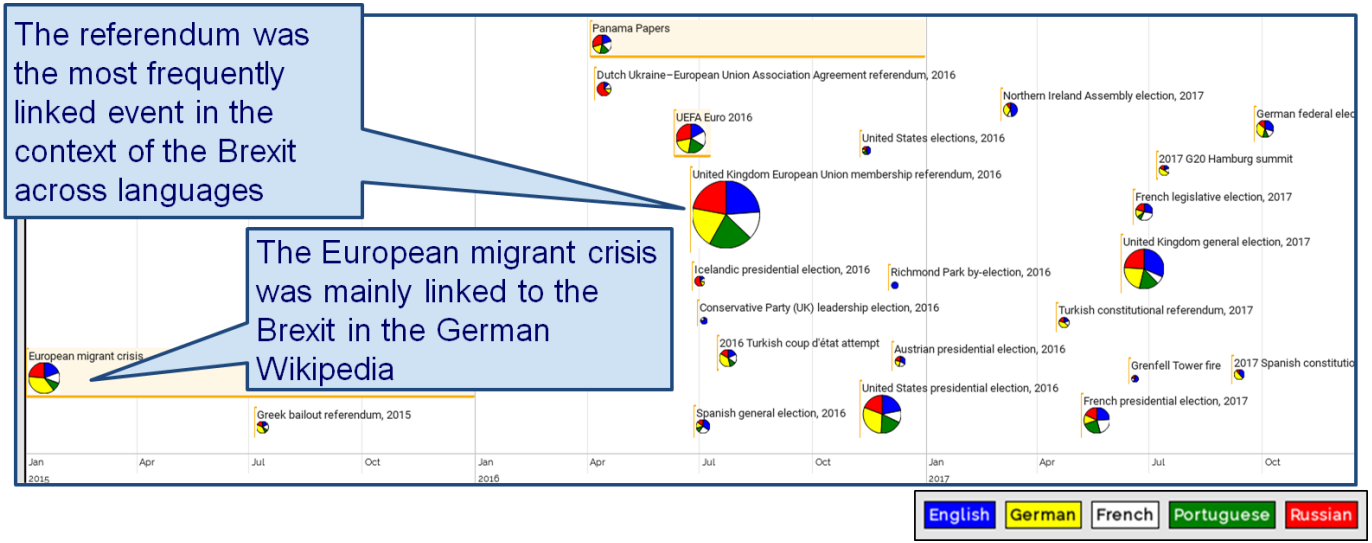
- Event series: sequence of topically related events that occur repeatedly in a similar form
  - US presidential elections 2004, 2008, 2012, ...
  - Wimbledon Championships 2017, 2018, 2019, ...
- Goal: Predict sub-event relations, infer missing events without relying on external knowledge



S. Gottschalk and E. Demidova. (2019) HapPenIng: Happen, Predict, Infer - event series completion in a knowledge graph. In ISWC 2019.

Cross-Lingual Timelines with EventKG

- Zoom out: Cross-lingual popularity of events in a timeline
- Example: Brexit



A pie chart → linked event; size → language-independent event popularity; slice area → ratio of event popularity in a language context

S. Gottschalk and E. Demidova. EventKG+TL: Creating cross-lingual timelines from an event-centric knowledge graph. In ESWC'18 Satellite Events.

Knowledge Graph Completion: VGI Link Prediction

Goal: Link prediction for knowledge graphs and VGI (Volunteered Geographic Information)

Wikidata Knowledge Graph

- Large-scale knowledge graph: 56+ million entities in 2019
- Community project
- $>$ 6.4 million geographic entities (09/2018)

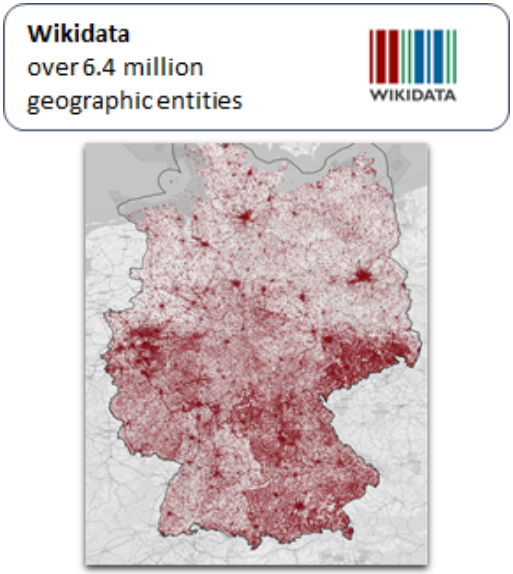
OpenStreetMap (OSM)

- Open world map (ObDL licence), from 2004
- Currently $>$ 5.8 million user
- Information is provided through volunteers (VGI)

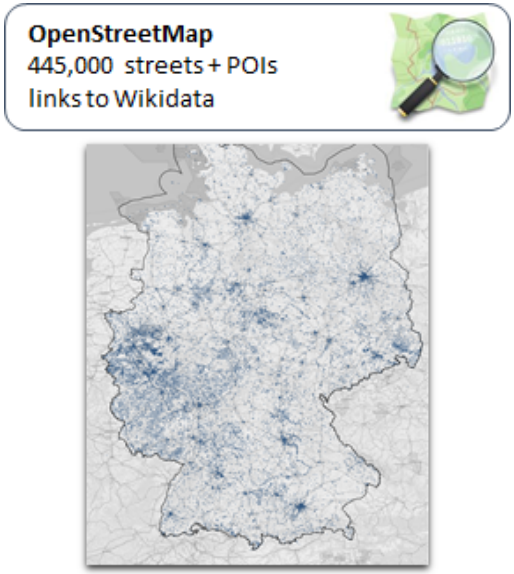
★ WorldKG. World-Scale Completion of Geographic Knowledge (DFG, 2020-2022)

Knowledge Graph Completion: VGI Link Prediction

Problem: Only about 0.01% of OSM nodes possess Wikidata links



All Wikidata entities



OSM links to Wikidata entities

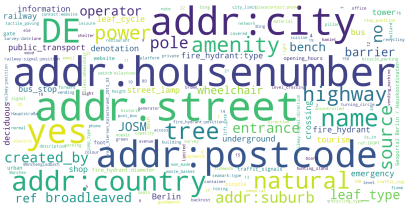
N. Tempelmeier, and E. Demidova. Linking OpenStreetMap with knowledge graphs – Link discovery for schema-agnostic volunteered geographic information. Future Gener. Comput. Syst. 116: 349-364 (2021)

Knowledge Graph Completion: VGI Link Prediction

Goal: Predict identity links between spatial entities in knowledge graphs and OSM nodes

Challenges:

- OpenStreetMap data is highly heterogeneous
- Equivalent properties have different representations in KGs and OSM



Frequent tags: OSM-DE

Wikidata representation of Berlin

Subject	Predicate	Object
$Q64$	*name*	*Berlin*
$Q64$	*instance of*	*Big City*
$Q64$	*coordinate*	$5231'N,\ 1323'E$
$Q64$	*capital of*	*Germany*

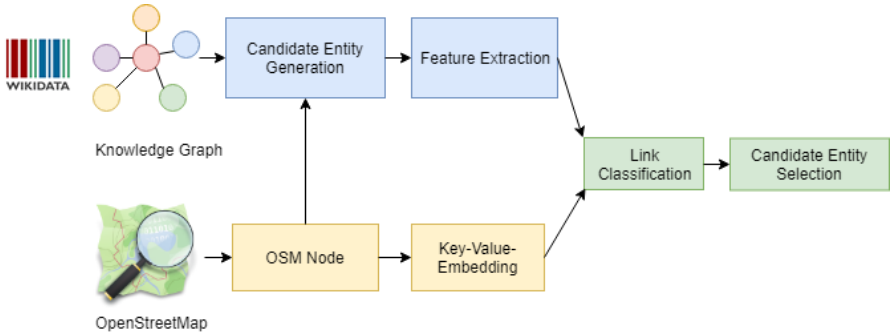
OSM representation of Berlin

Key	Value
i	240109189
l	$52.5170365, 13.3888599$
`name`	*Berlin*
`place`	*city*
`capital`	*yes*

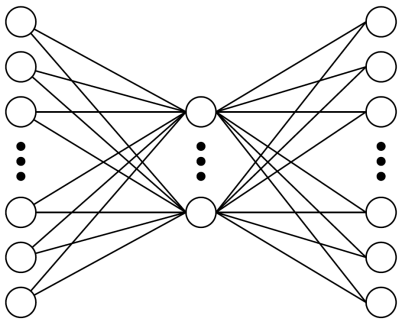
Knowledge Graph Completion: VGI Link Prediction

Link prediction approach:

- Classification model: learns if a knowledge graph entity and an OSM node represent the same real-world entity
- KG features: entity type, popularity, spatial distance
- OSM node embedding: captures semantic similarity of the nodes. ML-approach inspired by the skip gram model for word embeddings (i.e. prediction of context words)



OSM2KG pipeline



Embedding architecture

Learning OSM Node Embedding

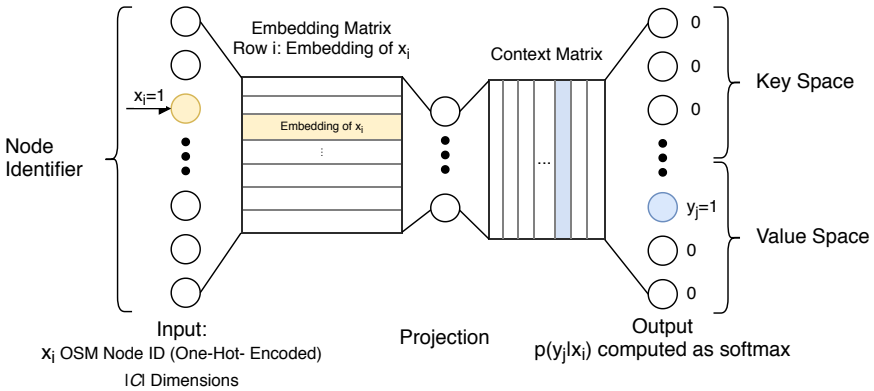
OSM nodes: $n = \langle id, T \rangle, n \in \mathcal{C}$, T: key-value pairs
Goal: OSM node embedding to encode semantic similarity
Architecture: feedforward neural network with one hidden layer
Learning goal: minimisation of the cross-entropy using stochastic gradient descent

$$\mathcal{L}_\theta = -\sum_{n \in \mathcal{C}} \sum_{\langle k,v \rangle \in n.T} \log p(k|n.id) + \log p(v|n.id)$$

OSM node embedding: a row in the node embedding matrix

Key (k)	**Value** (v)
`name`	*Berlin*
`place`	*city*

Key (k)	**Value** (v)
`name`	*Hannover*
`place`	*city*



Experimental Results

Results

- OSM2KG outperforms the baselines with respect to recall and F_1
- Prediction is more difficult in the areas with higher node density
- OSM2KG provides a compact representation - significant reduction of memory consumption compared to TF-IDF

Table 5: Macro averages for precision, recall and F1 score [%], best scores are bold. Statistically significant (according to paired t-tests with $p < 0.05$) F1 score results of OSM2KG compared to all baselines and OSM2KG-TFIDF are marked with *.

(a) Link prediction performance on the Wikidata datasets

Approach	Wikidata-OSM-FR			Wikidata-OSM-DE			Wikidata-OSM-IT			Average		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
BM25	45.22	42.59	43.86	47.28	41.60	44.26	44.49	41.67	43.04	45.66	41.95	43.72
SPOTLIGHT	65.17	32.26	43.15	69.79	51.03	58.95	54.79	26.89	36.08	63.25	36.73	46.06
GEO-DIST	74.46	74.46	74.46	62.16	62.16	62.16	72.80	72.80	72.80	69.81	69.81	69.81
LGD	**100.00**	44.09	61.20	**100.00**	47.46	64.37	**100.00**	43.59	60.71	**100.00**	45.05	62.09
LGD-SUPER	**100.00**	53.25	69.50	**100.00**	55.34	71.25	**100.00**	53.79	69.95	**100.00**	54.13	70.23
YAGO2GEO	63.66	44.98	52.71	64.48	48.61	55.43	58.40	47.36	52.30	62.18	46.98	53.48
YAGO2GEO-SUPER	78.49	47.38	59.09	73.49	48.96	58.76	72.25	48.73	58.20	74.74	48.36	58.69
LIMES/WOMBAT	74.03	17.50	28.31	78.54	17.01	27.97	65.28	17.22	27.25	72.62	17.25	27.84
OSM2KG-TFIDF	95.06	90.60	92.77	93.67	86.37	89.87	93.98	87.07	90.39	94.24	88.01	91.01
OSM2KG	95.51	**91.90**	**93.67***	93.98	**88.29**	**91.05***	94.39	**88.68**	**91.45***	94.62	**89.63**	**92.05**

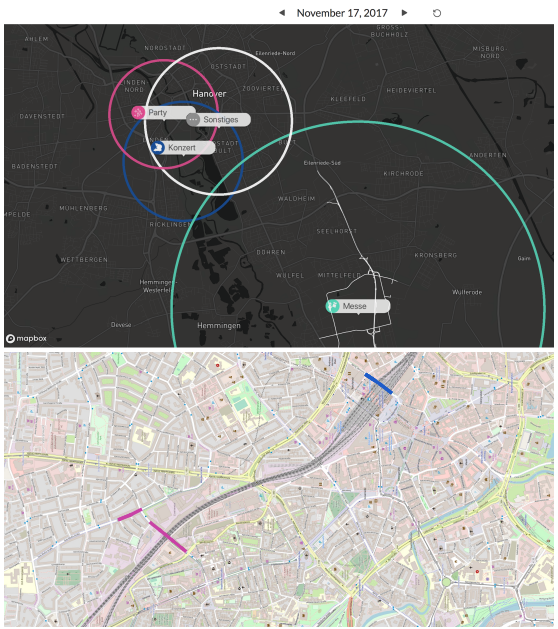
Spatio-Temporal Data Analytics

Spatio-Temporal Data Analytics

Selected research questions in the current projects:

- Prediction of spatial impact for planned special events
- Detection of structural dependencies in road networks
- Detection of dangerous places
- Demand forecast for charging facilities for electric vehicles

→ Require integration of a variety of heterogeneous data sources (events, traffic, maps, weather, accidents, etc.)



★ Data4UrbanMobility (BMBF, 2017-2020). ★ Campaneo (BMWi, 2019-2022).
★ d-E-mand (BMWi, 2020-2022). ★ smashHit (EU H2020, 2020-2022).

Challenges and Data Analytics Methods

Challenges

- Large variety of heterogeneous data sources (traffic, map, weather, accidents, etc.)
- Temporal gaps: irregular data collection
- Spatial gaps: missing geographic coverage
- Legal gaps: missing consent for data analytics

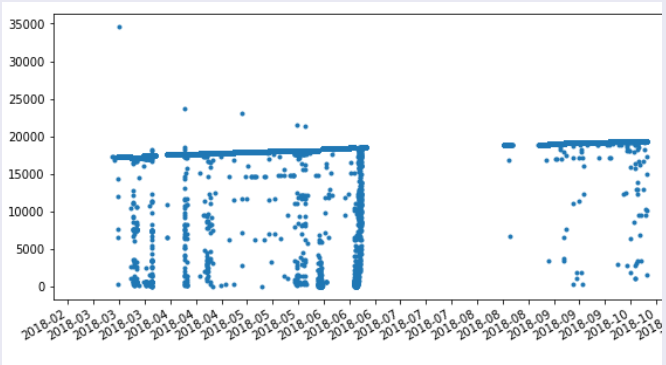
Data analytics methods

- Semantic data integration
- Robust data analytics
- Spatial knowledge transfer
- Transparent data analysis (data use traceability)

Data Quality Problems: Sparsity

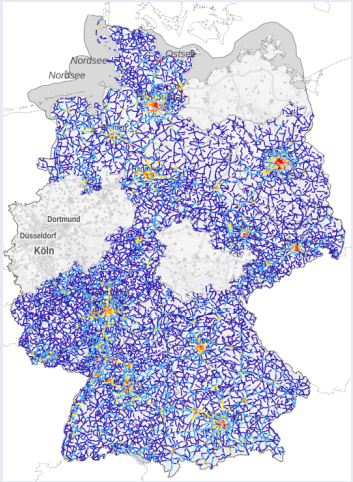
- Mobility data is typically sparse and incomplete

Irregular data acquisition



Traffic speed records: number of data points (i.e. street segments) captured in Hannover, Brunswick and Wolfsburg every 15 minutes 02/2018-10/2018 in the D4UM project

Lack of regional coverage



Missing regions in the accident atlas `unfallatlas.statistikportal.de`

Robust Data Analytics

- Idea: learning outlier patterns and correlations to identify meaningful dependencies
- Role of knowledge graphs: semantic data integration, contextual information to identify influence factors

Supervised learning

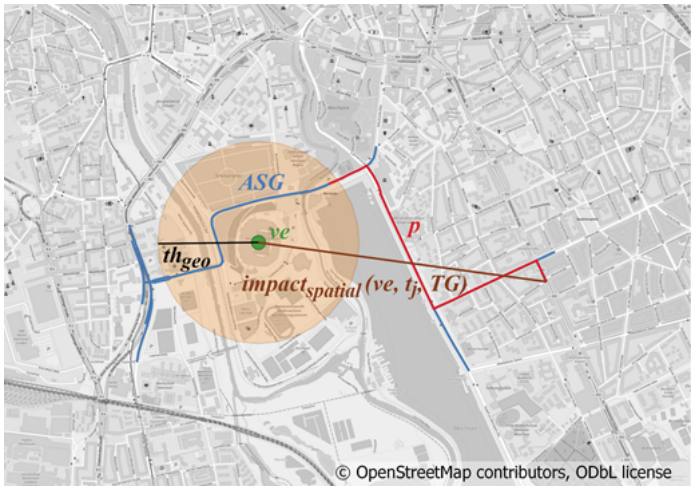
- Spatio-temporal outlier patterns to be determined are defined manually
- s. Analysis of the impact of planned special events

Unsupervised learning

- Spatio-temporal outlier patterns are determined automatically by algorithms (e.g. spatio-temporal clustering)
- s. Detection of structural dependencies

Prediction of Spatial Impact for Planned Special Events

- Traffic data: speed measurements (ca. every 15 minutes)
- OSM data and contextual information in knowledge graphs: road type, venue, popularity, number of participants
- Method: load prediction using outlier patterns and regression



- Define outlier-based load patterns: affected subgraph ASG in t_j.
- Start near the venue v_e (threshold $th_{geo} = 500m$).
- Predict the longest affected subgraph p.

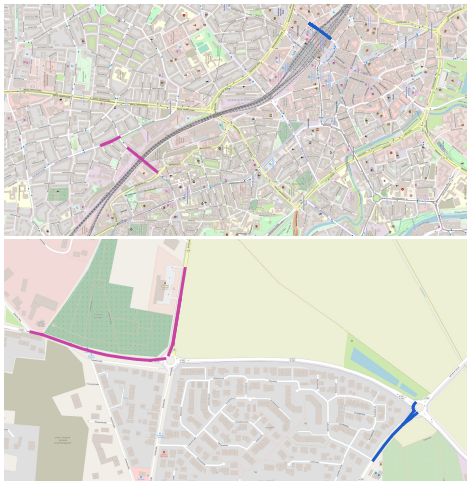
N. Tempelmeier, et al., Crosstown traffic - supervised prediction of impact of planned special events on urban traffic. GeoInformatica 24(2): 339-370 (2020).

Detection of Structural Dependencies

Goal: Detection of structural dependencies in road networks
Intuition: Co-occurrence of outlier patterns in traffic data on nearby road segments can indicate structural dependencies

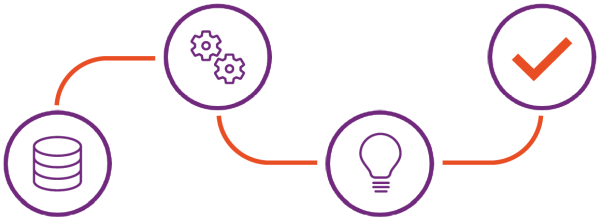
Unsupervised learning method:

- Outlier detection in traffic data
- Outlier-based clustering of road segments (region growing method)
- Calculation of mutual information between clusters in spatio-temporal proximity



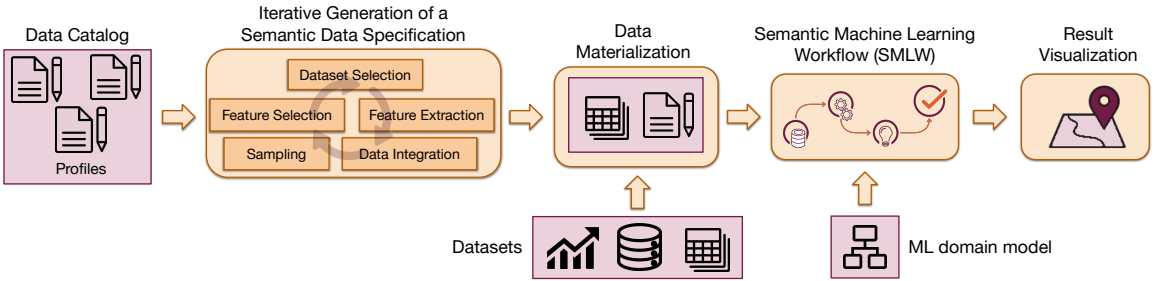
N. Tempelmeier, et al. ST-Discovery: Data-driven discovery of structural dependencies in urban road networks. In ACM SIGSPATIAL 2019.
N. Tempelmeier, et al. Mining topological dependencies of recurrent congestion in road networks. ISPRS International Journal of Geo-Information, 10(4), 2021.

Semantic Data Analytics Workflows



Semantic Data Analytics Workflows

- Goal: Increasing the usability, efficiency, robustness, explainability and reusability of ML workflows
- Adoption of semantic information in all ML workflow components
  - Semantic data profiles
  - Domain-specific semantic data models
  - Semantic integration of heterogeneous data sources in domain-specific knowledge graphs
- Application domains: mobility and logistics



S. Gottschalk, et al. Simple-ML: Towards a framework for semantic data analytics workflows. In Proc. of SEMANTiCS 2019.

★ Simple-ML. Big data ML workflows made easy (BMBF, 2019-2022)

Thank you for your attention! Questions?

Prof. Dr. Elena Demidova
Data Science & Intelligent Systems
Computer Science Institute
University of Bonn

email: demidova@cs.uni-bonn.de
web: `http://dsis.iai.uni-bonn.de`