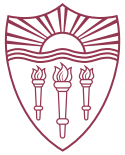


Transforming Unstructured **Historical** and **Geographic Data** into **Spatio-Temporal Knowledge Graphs**

Basel Shbita

PhD Thesis Defense

05/01/2024

Committee Members:

Craig A. Knoblock, Cyrus Shahabi, John P. Wilson, Jay Pujara, and Yao-Yi Chiang

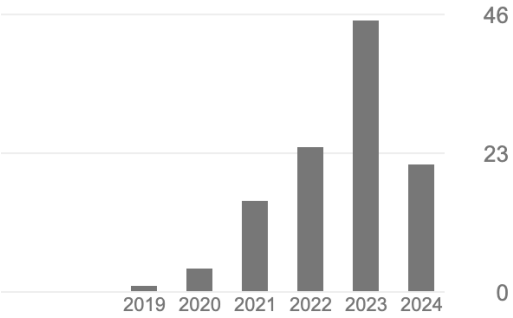
Agenda

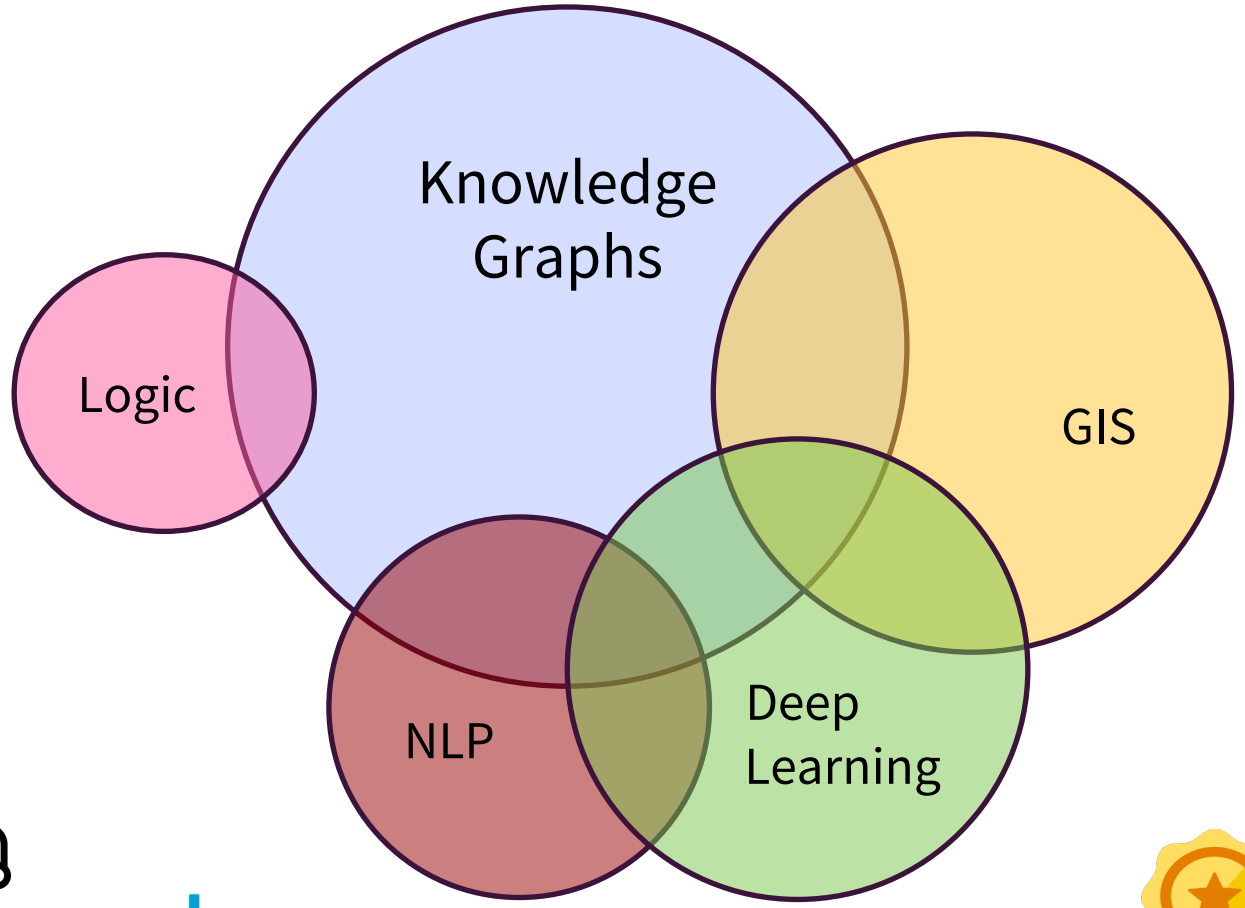
- Basel's PhD Journey ⟸
- Intro
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

My PhD Journey

Cited by

	All	Since 2019
Citations	110	110
h-index	6	6
i10-index	4	4





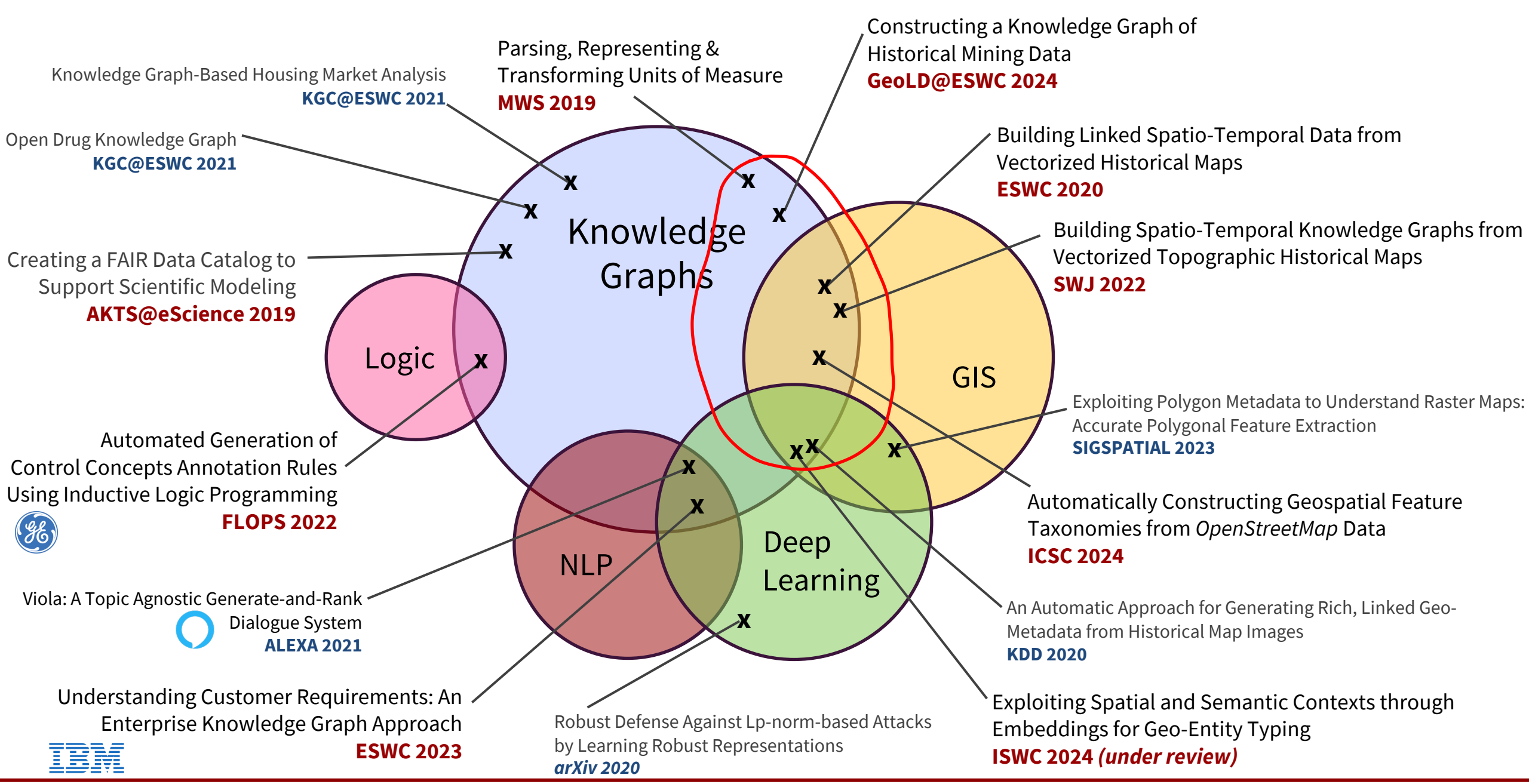
I take different kinds of data sources then instill them with semantics

ISWC 2022, 23, 24
WWW 2023
KDD 2024

GE Research 2021
IBM Research 2022

alexa semifinalists 2021

University Outstanding Teaching Assistant Award 2022
Highest Achievement

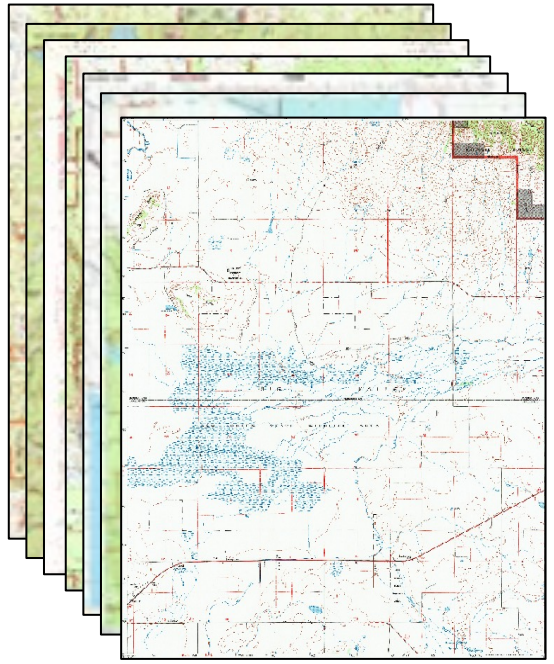

Knowledge Graph-Based Housing Market Analysis
KGC@ESWC 2021
Parsing, Representing & Transforming Units of Measure
MWS 2019
Constructing a Knowledge Graph of Historical Mining Data
GeoLD@ESWC 2024
Open Drug Knowledge Graph
KGC@ESWC 2021
Building Linked Spatio-Temporal Data from Vectorized Historical Maps
ESWC 2020
Creating a FAIR Data Catalog to Support Scientific Modeling
AKTS@eScience 2019
Building Spatio-Temporal Knowledge Graphs from Vectorized Topographic Historical Maps
SWJ 2022
Knowledge Graphs
Logic
GIS
Exploiting Polygon Metadata to Understand Raster Maps: Accurate Polygonal Feature Extraction
SIGSPATIAL 2023
Automated Generation of Control Concepts Annotation Rules Using Inductive Logic Programming
FLOPS 2022
Automatically Constructing Geospatial Feature Taxonomies from OpenStreetMap Data
ICSC 2024
NLP
Deep Learning
Viola: A Topic Agnostic Generate-and-Rank Dialogue System
ALEXA 2021
An Automatic Approach for Generating Rich, Linked Geo-Metadata from Historical Map Images
KDD 2020
Understanding Customer Requirements: An Enterprise Knowledge Graph Approach
ESWC 2023
Robust Defense Against Lp-norm-based Attacks by Learning Robust Representations
arXiv 2020
Exploiting Spatial and Semantic Contexts through Embeddings for Geo-Entity Typing
ISWC 2024 (under review)

Agenda

- Basel's PhD Journey
- Intro 
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

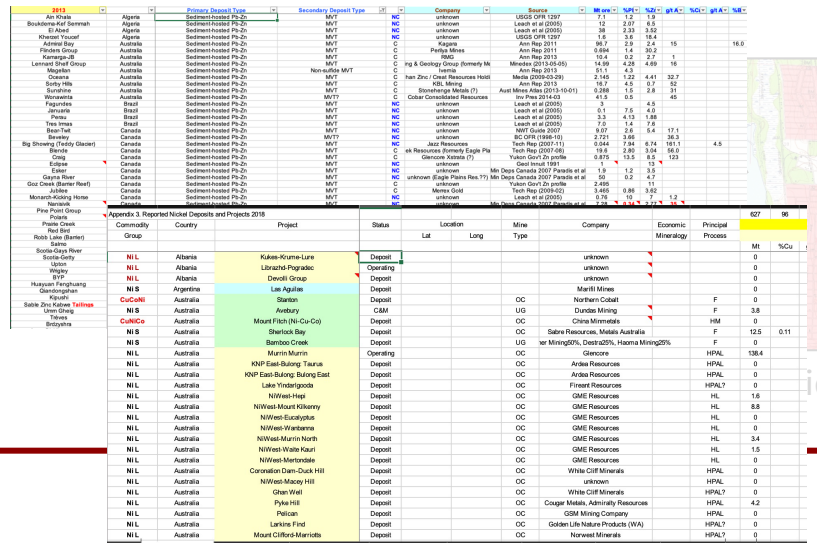
Historical & Geo Data

- Rich sources of information
  - understanding human & environmental systems
  - describing human & natural activities
- Labor-intensive to analyze across time & space
  - e.g., railroad change, decline of wetlands, geo-related characteristics
- Often require grounding & additional contextual information
  - e.g., demographics, geology, stratigraphy, other

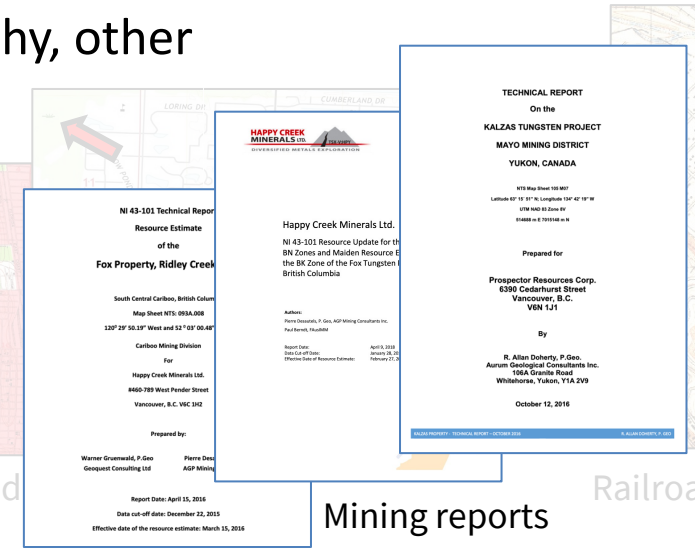


Historical topographic maps

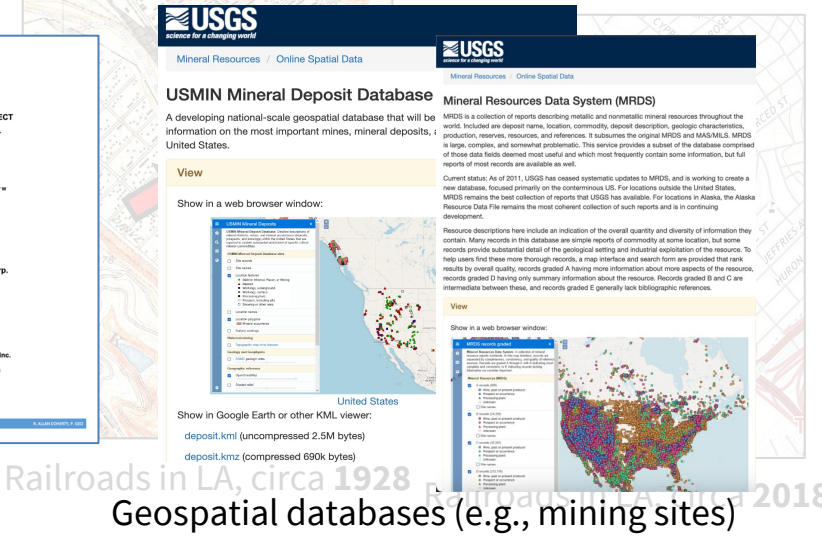
Web



Tables with geo-data



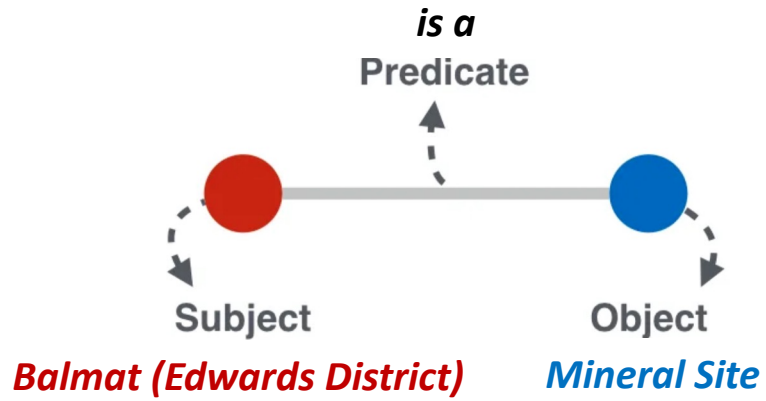
Mining reports



Railroads in LA, circa 1928 Railroads in LA, circa 2018

Geospatial databases (e.g., mining sites)

KGs

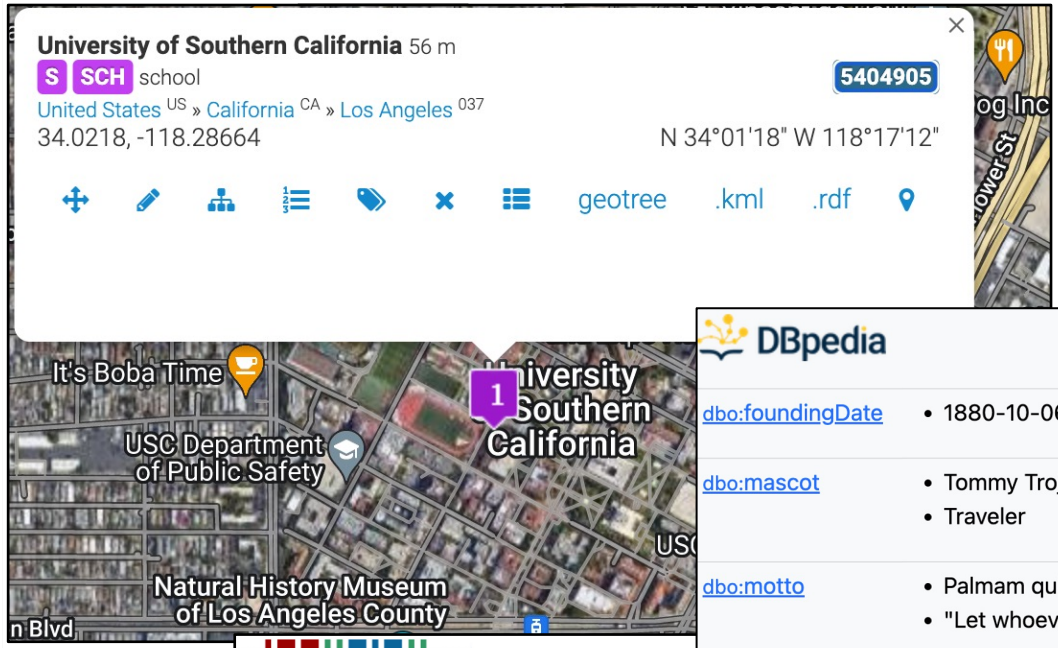


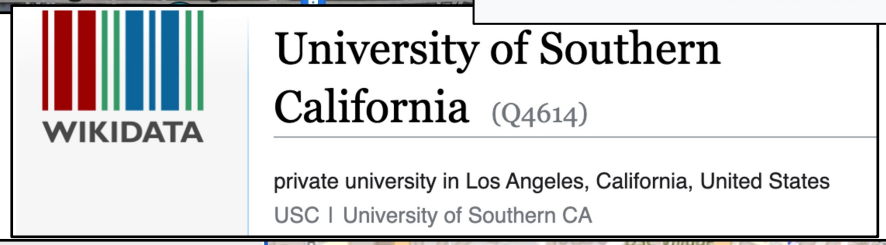
- Knowledge Graph (KG)?
  - Graphs are natural way to encode data
  - Using semantic concepts & relationships
    - Semantic Network = **Knowledge Graph**
- Why use KG?
  - Combine expressivity, interoperability, & standardization in the Semantic Web stack
- Semantic Web?
  - Extension of WWW, enabling the Web of Data (aka "Linked Data")
  - Encoding of semantics with the data
  - Linked Open Data principles // FAIR

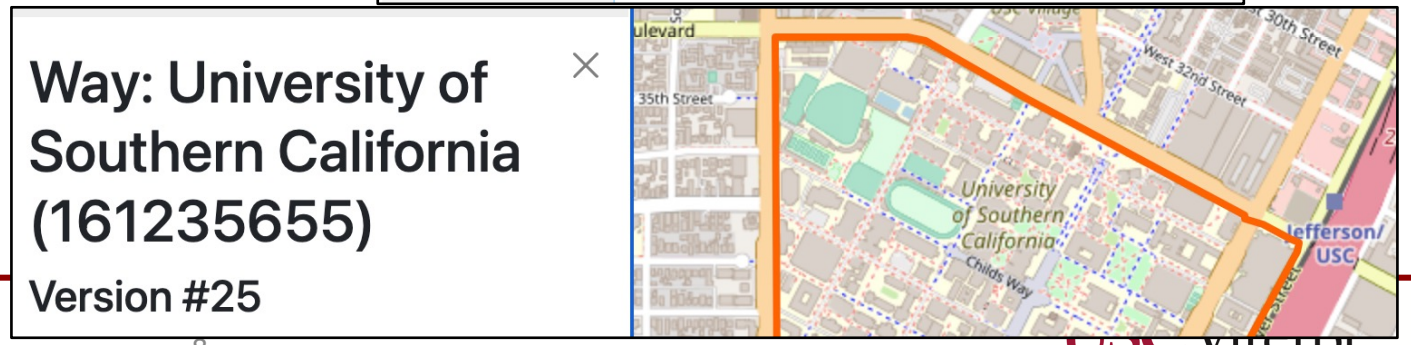
Geo & Spatio-Temporal KGs

- Spatio-Temporal KGs
  - Contextual (what)
  - **Spatial** (where)
  - **Temporal** (when)

- Geo-semantics
  - Representation, annotation, & reasoning
  - Modeling & ontology development
  - Integration & interoperability







Intro: Spatio-Temporal KGs

- So, what's so special about them?
  - Spatial analysis
  - Temporal analysis
  - Spatio-temporal aggregations
  - Geographic QA
  - Environmental & social science
  - Urban planning
  - Transportation
  - etc...

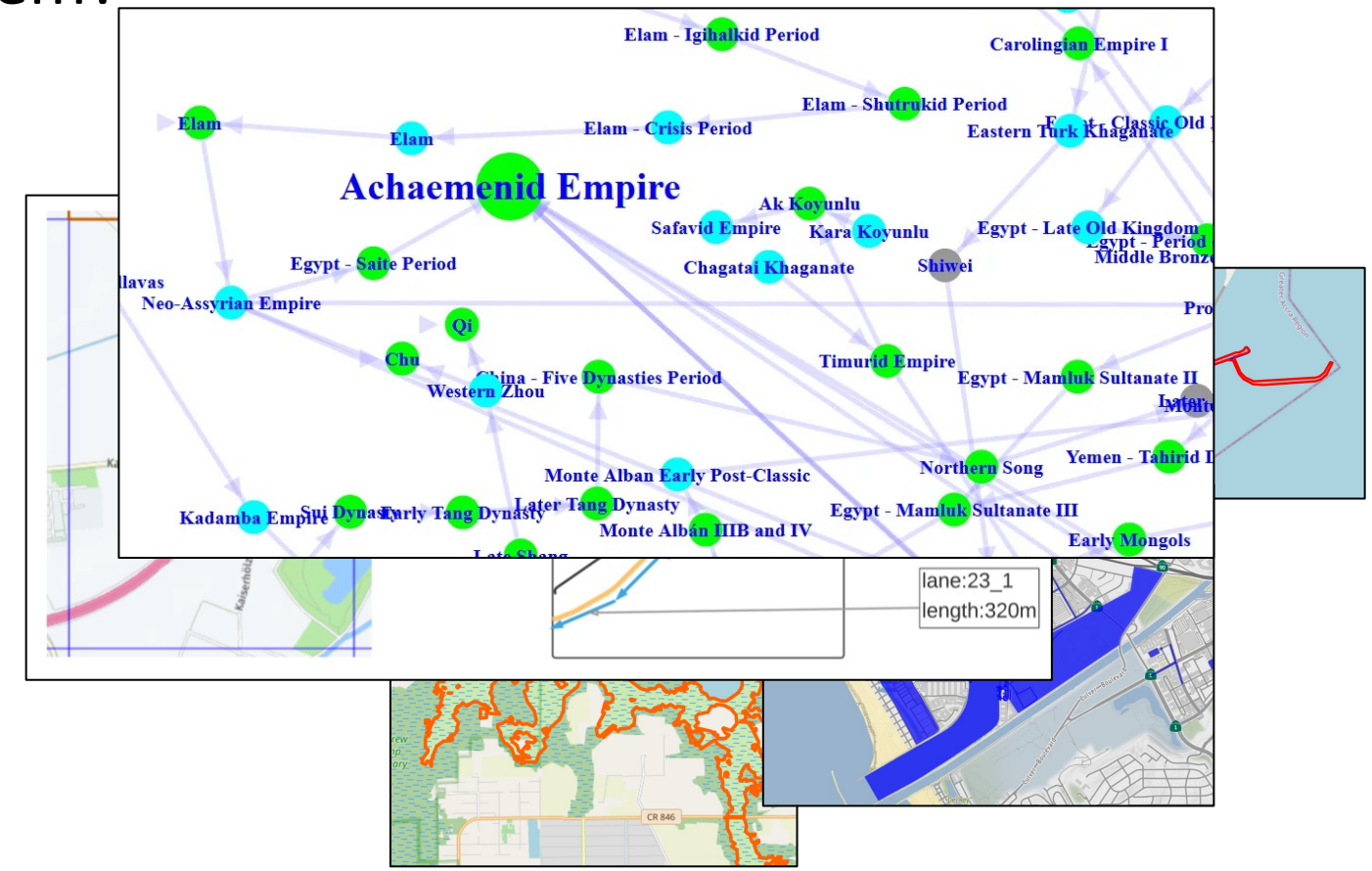


figure from *Szwoch, G. (2019). Combining road network data from openstreetmap with an authoritative database. Journal of Transportation Engineering, Part A: Systems, 145(2), 04018085.*
figure from *https://terminusdb.com/blog/human-history-knowledge-graph/*

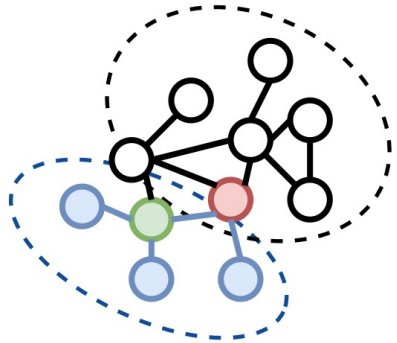
Agenda

- Basel's PhD Journey
- Intro
- Thesis Overview 
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

Research Problem

- How can we transform & link unstructured digitized & historical geo-data into structured, semantic, & queryable spatio-temporal KGs?

- Objectives:
  - Automated KG construction from various historical geo-data sources
  - Contextual geo-entity recognition (ER) & semantic typing
  - Semantic enrichment by linking (EL) to additional sources on the web
  - Adherence to Semantic Web principles
    - shared, accessible, visualized, standardized across-domains, & scalable for easy use by downstream tasks for easy analysis & expressive integration



Thesis statement

This thesis provides tools & techniques for the automated understanding & transformation of unstructured geospatial & historical data from heterogeneous sources into a standardized representation of expressive & interoperable spatio-temporal knowledge graphs. I also present methodologies that both integrate the data with other sources on the web and leverage web knowledge for enhanced data analysis.

Building Spatio-Temporal KGs from Digitized Maps

Embedding Geo-Entities for Semantic Typing

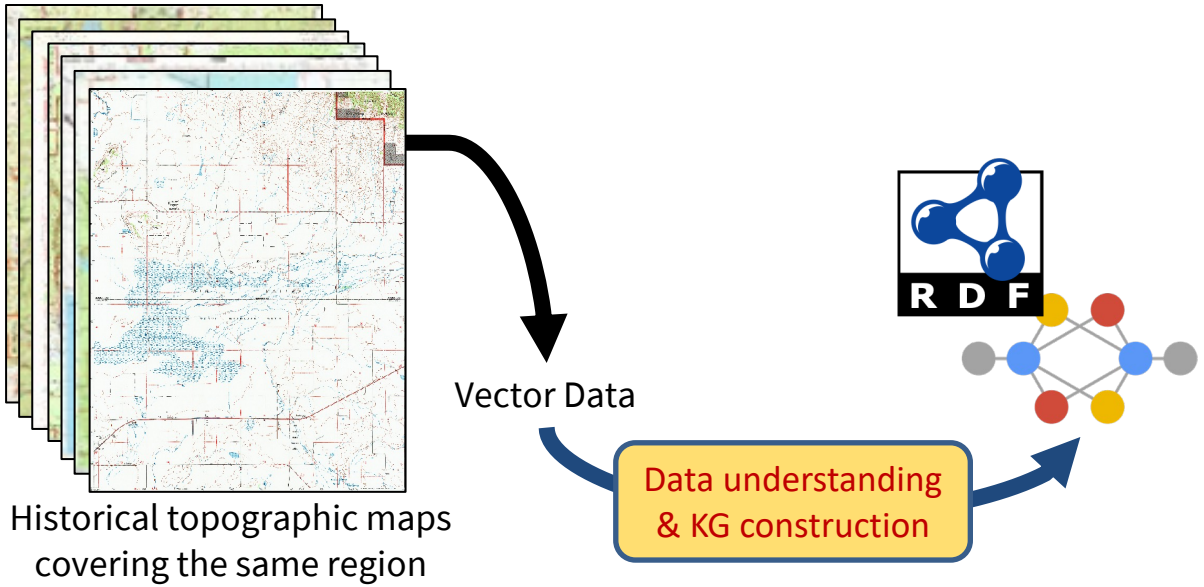
From Digitized Reports to Spatio-Temporal KGs

Agenda

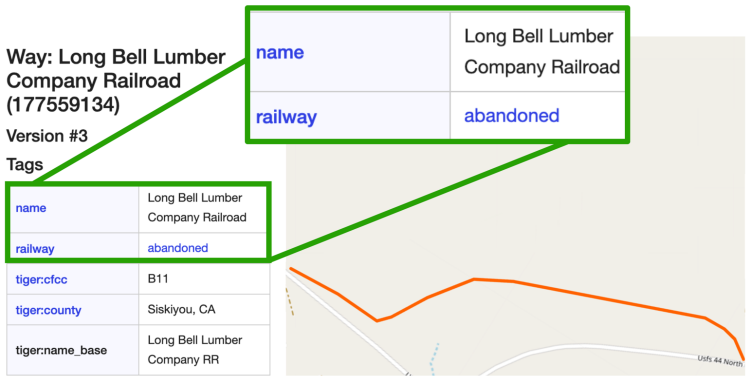
- Basel's PhD Journey
- Intro
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

Building Spatio-Temporal KGs from Digitized Maps

- Goals
  - Automatically integrate, represent, relate & interlink geospatial data from overlapping digitized resources
    - Line & Polygon features (e.g. railroads, wetlands)
    - Geo-semantic representation following LD & SW principles
  - User-assisted (known type) geo-entity linking (EL)

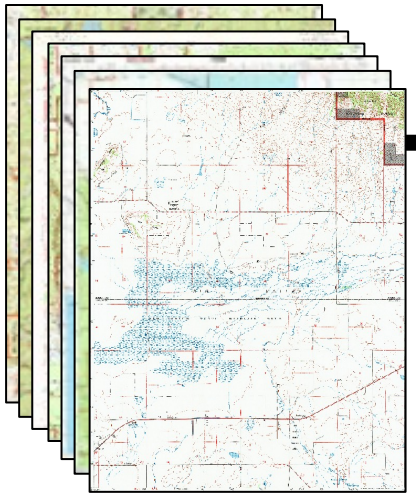


Building Spatio-Temporal KGs from Digitized Maps



"can you show me a subset of instances tagged as abandoned?"

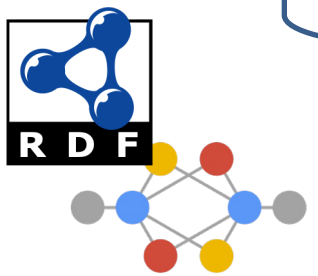
"railroad segments that are present in 1962 but are not present in 2001"



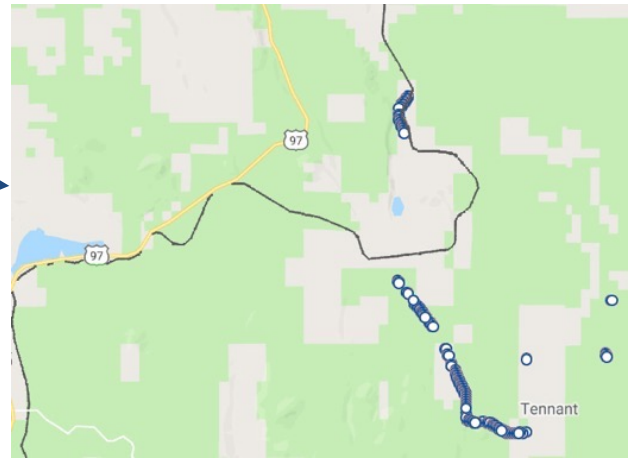
Historical topographic maps covering the same region

Vector Data

Data understanding & KG construction

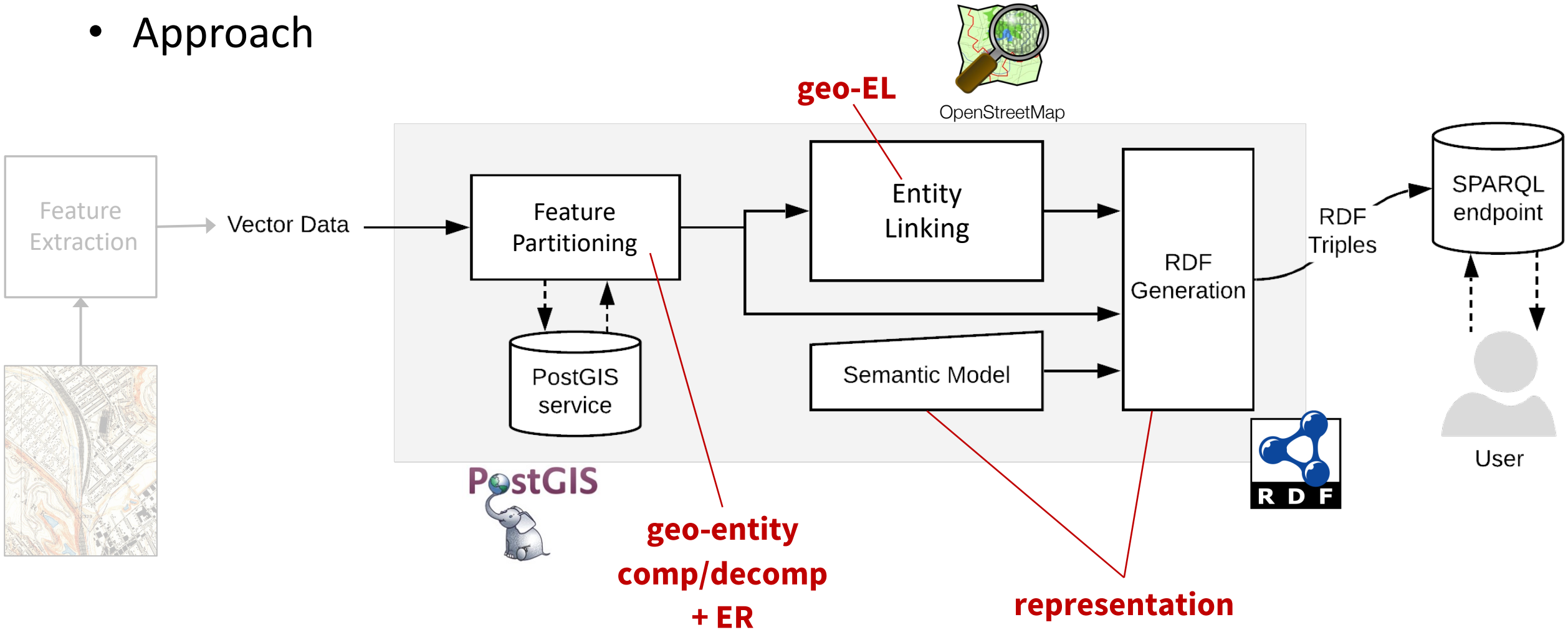


SPARQL endpoint



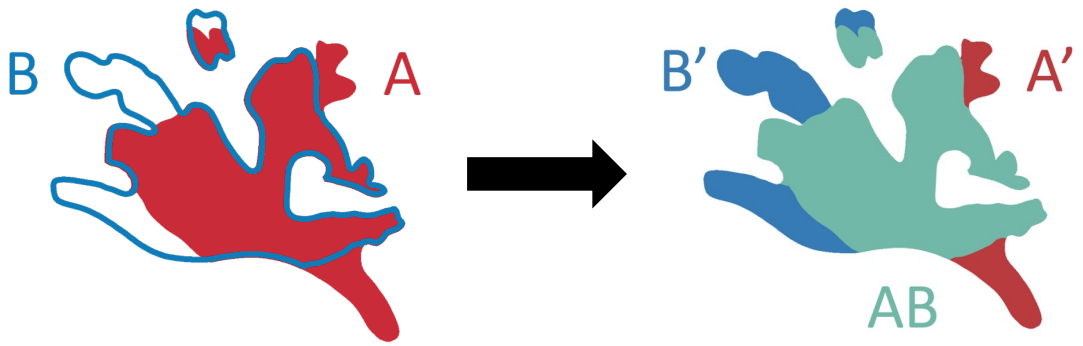
Building Spatio-Temporal KGs from Digitized Maps

- Approach



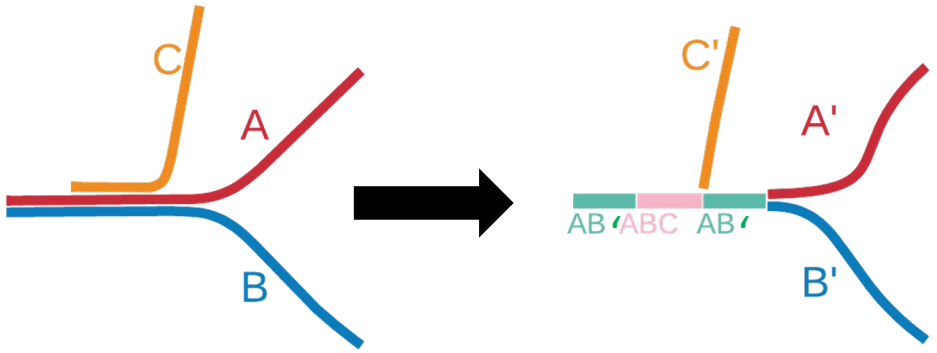
Building Spatio-Temporal KGs from Digitized Maps

- Approach
  - Step 1. Feature Partitioning
    - Generate building block geometries (i.e. geo entities) to represent the topographic features from different map sheets
      - Represent common & distinct parts (changes) of the geo-features
      - Allow incremental additions over time
    - Create a DAG of building-block geometries (nodes) and their relations (edges)

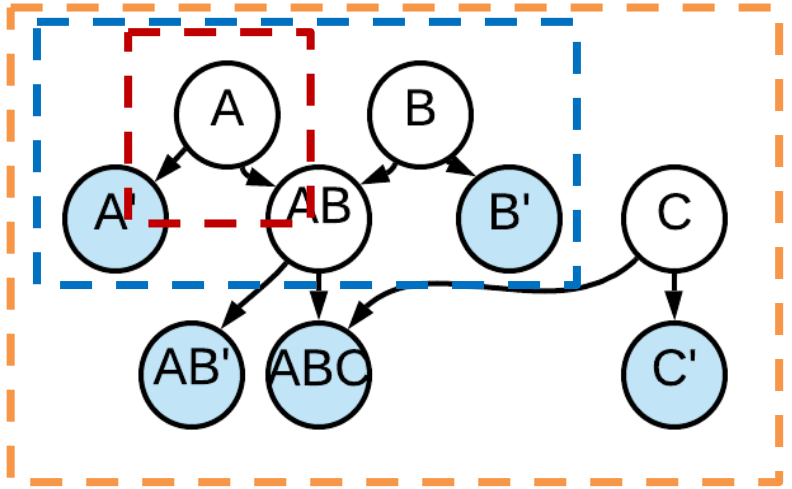


Building Spatio-Temporal KGs from Digitized Maps

- Approach
  - Step 1. Feature Partitioning

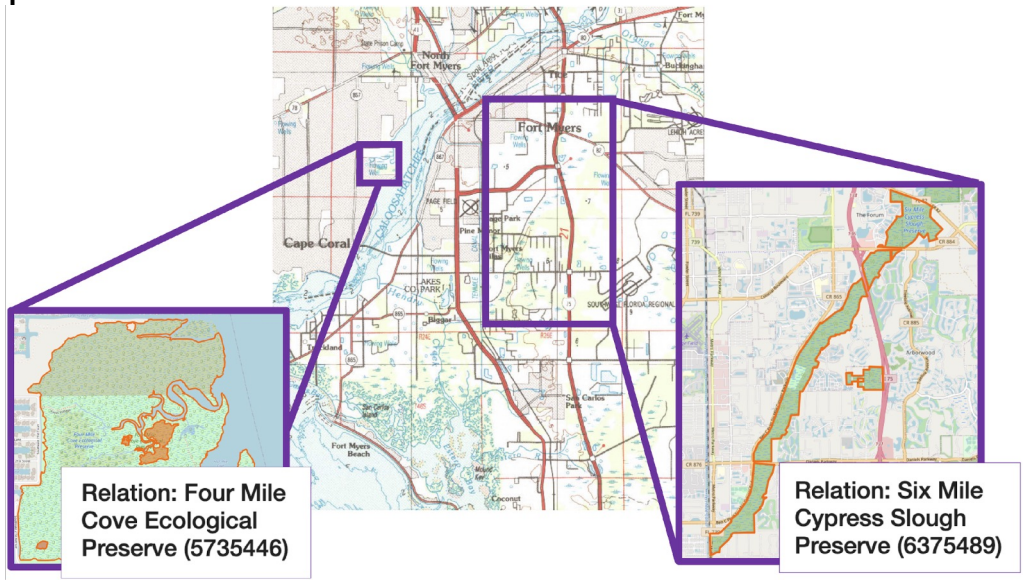


foreach $i \in \mathcal{M}$ **do**
  foreach $k \in \mathcal{L}$ **do**
    $\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$
    $\mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha;$
  end
  $\mathcal{F}_\delta = \mathcal{F}_i \setminus (\bigcup_{j \in \mathcal{L}} F_j);$
end



Building Spatio-Temporal KGs from Digitized Maps

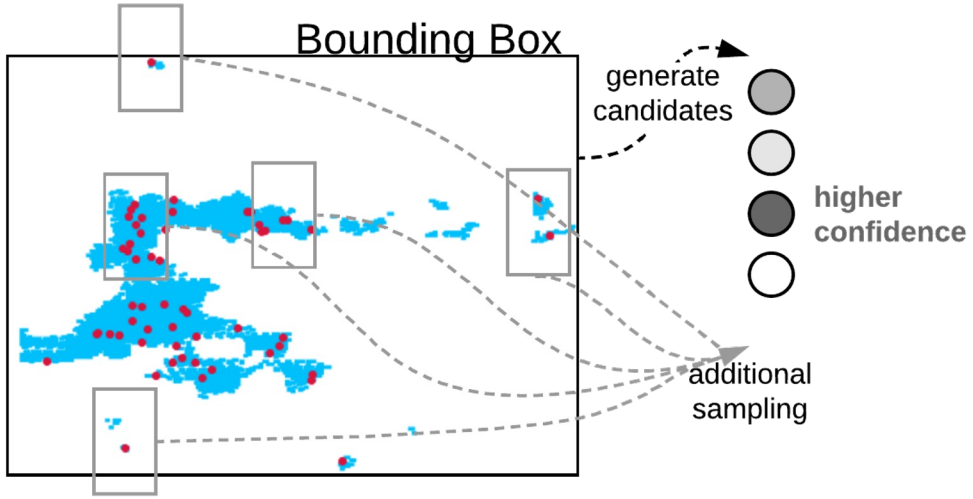
- Approach
  - Step 2. Geo-entity Linking
    - Link the generated entities to LoD (OSM, geoNames, LinkedGeoData, Wikidata)
    - Sampling
      - Reverse geocoding for initial filtering (geo feature type is known)
      - Determine confidence by frequency of (random) samples

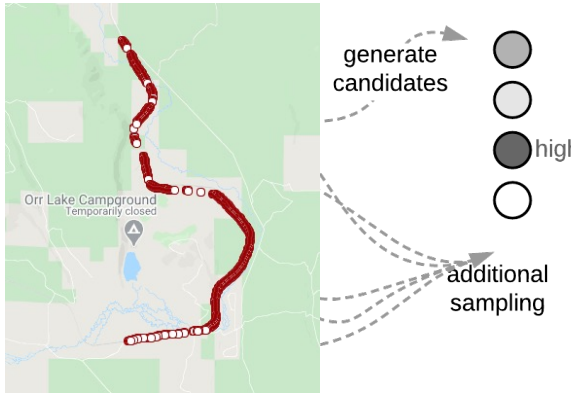


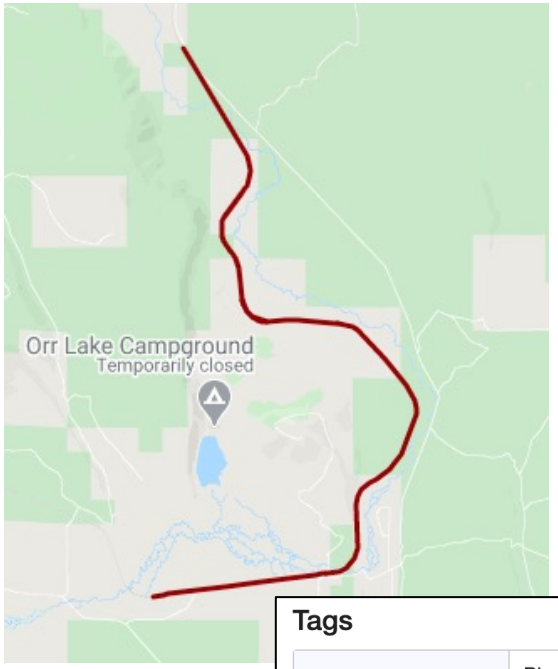
Building Spatio-Temporal KGs from Digitized Maps

- Approach
  - Step 2. Geo-entity Linking

B_s = bounding box wrapping s;
$\mathcal{L}$ = reverse-geocoding(B_s, T);
for 1...N **do**
  e = randomly sample a `Point` in segment s;
  E = reverse-geocoding(e, T);
  $\mathcal{L}$.add(E);
end







Way: Black Butte Subdivision (322131253)

Version #6

Tags

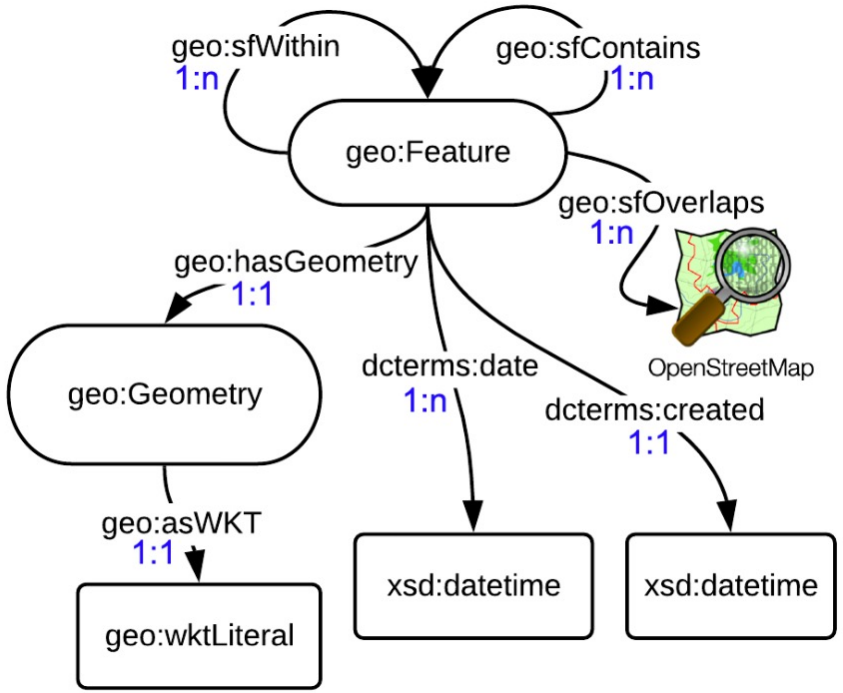
name	Black Butte Subdivision
operator	Union Pacific Railroad;Amtrak
owner	Union Pacific Railroad
railway	rail

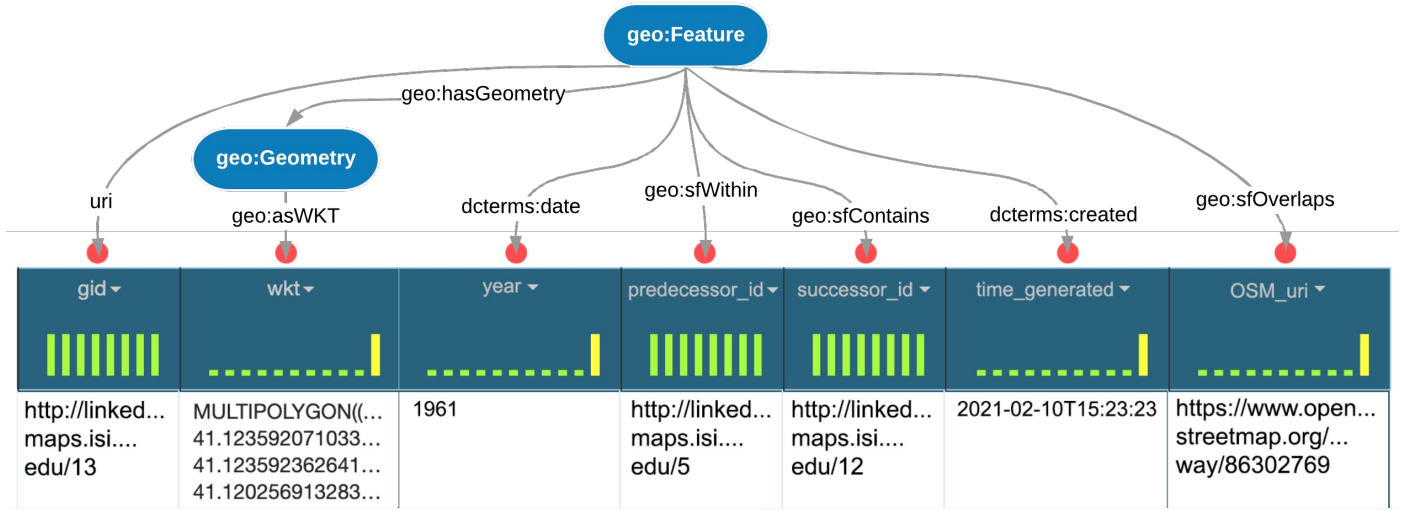
Building Spatio-Temporal KGs from Digitized Maps

- Approach
  - Step 3. Semantic Modeling (Data Representation)
    - Transform & materialize the data (construct KG)
      - Follows linked data principles
      - Provide a useful semantic representation supporting downstream tasks
    - Construct a meaningful semantic model
      - Hierarchically-driven
      - Follows W3C & OGC standards (GeoSPARQL)

Building Spatio-Temporal KGs from Digitized Maps

- Approach
  - Step 3. Semantic Modeling (Data Representation)






```
SELECT ?f ?wkt WHERE {
  ?f a geo:Feature ;
    geo:hasGeometry [ geo:asWKT ?wkt ] ;
    dcterms:date 1962^^xsd:gYear .
  FILTER NOT EXISTS { ?f geo:sfContains _:_ }
  MINUS { ?f dcterms:date 2001^^xsd:gYear . } }
```

Fig. 16. Query feature geometries present in 1962 but not in 2001

Building Spatio-Temporal KGs from Digitized Maps

- Evaluation
  - Feature Partitioning
  - Geo EL
  - RDF Query Performance

Data: 2 x railroad, 3 x wetland (3-7 sheets)

Geo-entity linking results; Area is in square kilometers

		Area	Precision	Recall	F_1
Railroads	CA-baseline	420.39	0.193	1.000	0.323
	CA		0.800	0.750	**0.774**
	CO-baseline	132.01	0.455	1.000	0.625
	CO		0.833	1.000	**0.909**
Wetlands	CA-baseline	224.05	0.556	1.000	0.714
	CA		1.000	1.000	**1.000**
	FL-baseline	27493.98	0.263	1.000	**0.417**
	FL		0.758	0.272	0.400
	TX*	16.62	-	-	-

Partitioning statistics for **CA railroads**

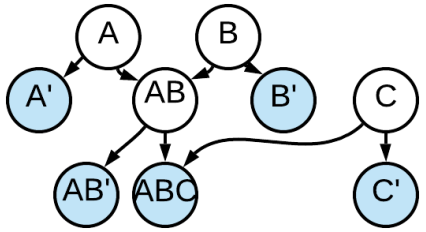
Year	# vecs	Runtime (s)	# nodes
1954	2382	<1	1
1962	2322	36	5
1988	11134	1047	11
1984	11868	581	24
1950	11076	1332	43
2001	497	145	57
1958	1860	222	85

Partitioning statistics for **CA wetlands**

Year	# vecs	Runtime (s)	# nodes
1961	12	<1	1
1993	17	<1	5
1990	27	6	11
2018	9	6	24

Partitioning statistics for **FL wetlands**

Year	# vecs	Runtime (s)	# nodes
1987	184	<1	1
1956	531	180	5
2020	5322	1139	13



Query time ranges **10-50 [ms]**

No significant change with respect to

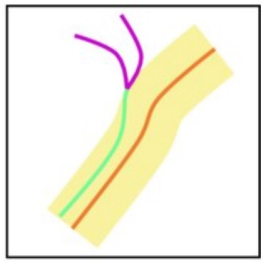
- # of map editions we process
- Complexity of the query we compose

Building Spatio-Temporal KGs from Digitized Maps

- Related Work
  - Transforming geospatial vector data into RDF (Kyzirakos 2014, Usery 2012)
    - Do not address geo-entity linking or entity semantics
  - Contextualizing geospatial data (Vaisman 2019, Smeros 2016)
    - Do not address Linking unlabeled geo entities
  - Geospatial change analysis (Perez 2015, Kauppinen 2014)
    - Do not address incremental process of change over time

Building Spatio-Temporal KGs from Digitized Maps

- Takeaways
  - Paradigm for geospatial data integration on the web
    - Unsupervised
    - Hierarchy-driven & incremental semantic model for simple & efficient querying
      - Follows LD & SW principles
      - Does not require re-generation of data
      - URIs are preserved
    - Linked to Web
      - Fuels further discovery & enrichment
  - Still, many challenges exist
    - Complexity of changes in original topographic maps
    - Quality & level of detail
    - Crowdsourcing: availability, granularity (e.g., mud vs. wetland)
    - User-informed knowledge (target schema)



https://linked-maps.isi.edu

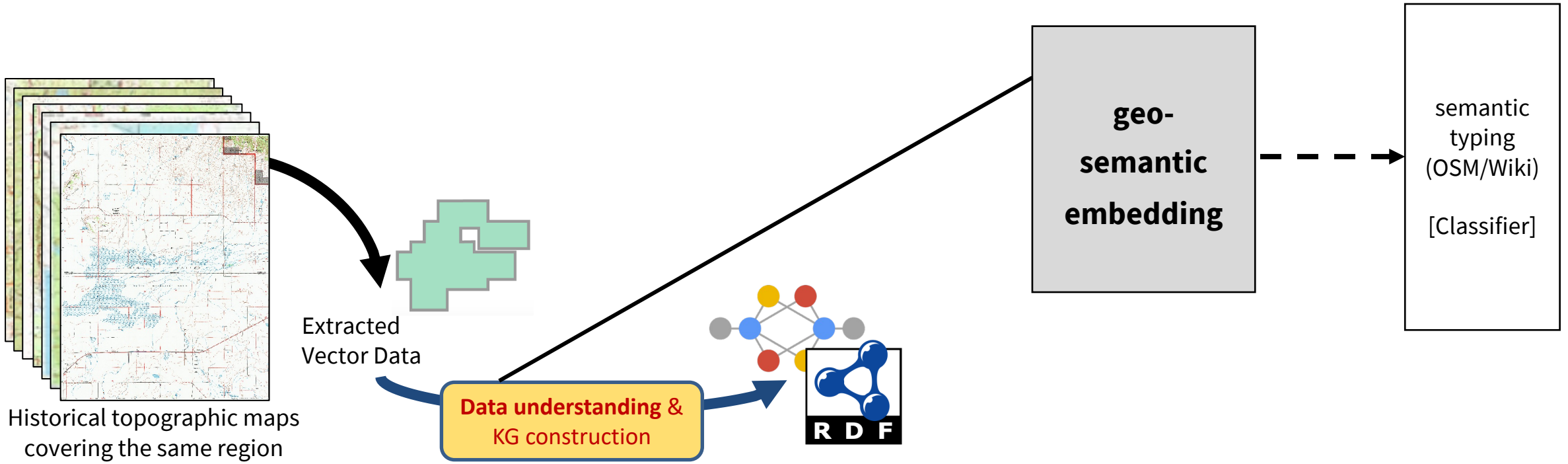
https://github.com/usc-isi-i2/linked-maps

Agenda

- Basel's PhD Journey
- Intro
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

Embedding Geo-Entities for Semantic Typing

- Goals:
  - Embed geospatial data into high-dimensional vector space
    - Preserve its semantic meaning & relationships between entities
  - Develop techniques for semantic typing/labeling of geospatial entities



Embedding Geo-Entities for Semantic Typing

waterway

?

- waterway
  - canal [5748]
  - dam [2425]
  - ditch [37754]
  - drain__waterway [7044]
  - river__waterway [16925]
  - stream [799887]

water

?

- water
  - basin [1870]
  - lake [3214]
  - pond [6835]
  - reservoir [4176]
  - river__water [1570]

building

?

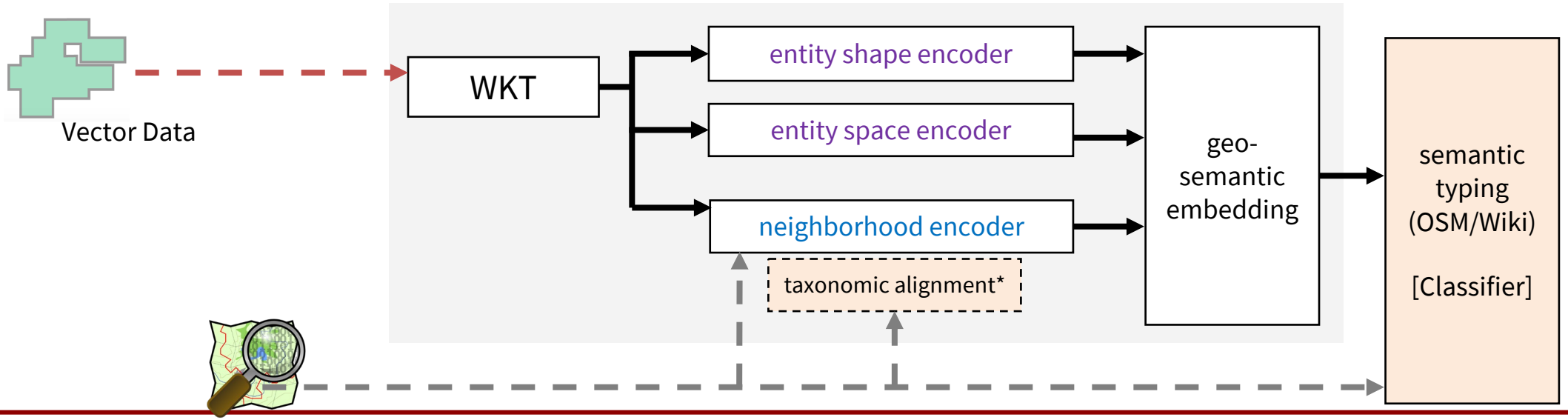
- building
  - apartments [31601]
  - commercial [3875]
  - house [118030]
  - industrial [3223]
  - residential [19763]
  - retail__building [4109]
  - school [2400]
  - warehouse [1098]

"Everything is related to everything else.
But near things are more related than distant things."
-Waldo R. Tobler

Embedding Geo-Entities for Semantic Typing

- Approach:
  - Method to embed:
    - Geometric attributes (shape)
    - Spatial attributes (area, length)
    - Neighborhood context (nearby geo-entities)

    to generate a representation that can learn & infer properties about geo-entities



Embedding Geo-Entities for Sem

- Data

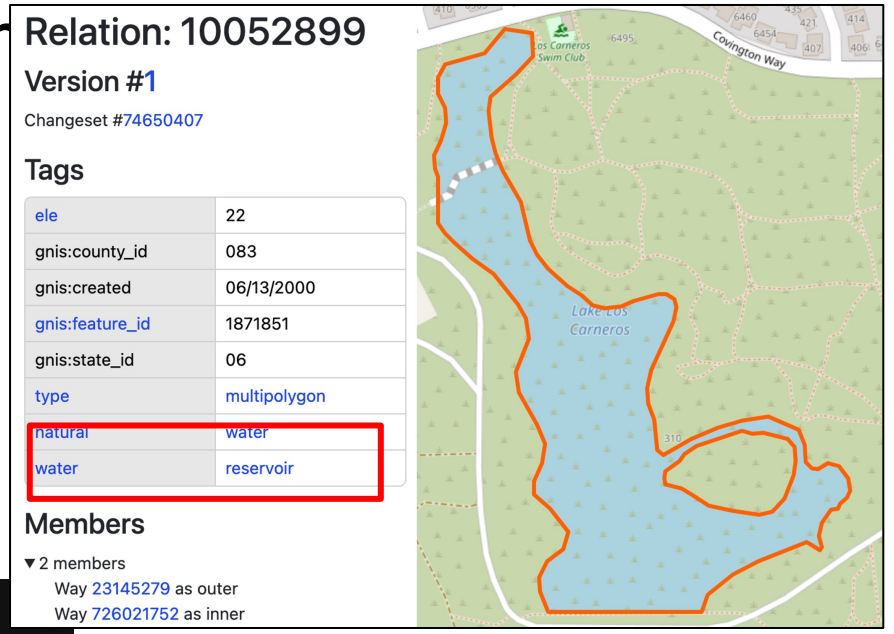
OpenStreetMap


Nodes - dots used to mark locations


Ways - connected line of nodes


Relation - used to create more complex shapes

CA OSM Snapshot



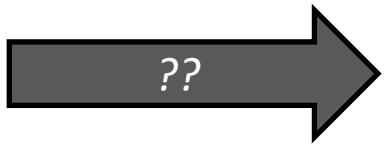
	index	
0	node_tagged	1000170
1	node_untagged	133590505
2	way_untagged	4029438
3	way_tagged	9017322
4	relation_untagged	84613
5	relation_tagged	68462

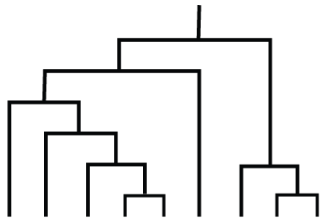
	shp_type	count
0	Polygon	4876318
1	LineString	4176529
2	Point	1000170
3	MultiPolygon	12694
4	GeometryCollection	1364

Embedding Geo-Entities for Semantic Typing

- but, OSM data is
  - Inconsistent across regions
  - Varying-granularity
  - Noisy

OpenStreetMap data/dump





WIKIDATA

sidewalk (Q177749)

pedestrian path along the side of a road

pavement | footpath | footway | platform

Statements

subclass of	thoroughfare

Statements

subclass of	public space
	line construction
	axis of communication
	geographical feature

OSMonto - An Ontology of OpenStreetMap Tags

Mihai Codescu*, Gregor Horsinka*, Oliver Kutz**
Till Mossakowski*,**, Rafaela Rau*

* DFKI GmbH Bremen, Germany
** Research Center on Spatial Cognition,
SFB/TR 8, University of Bremen, Germany

OUTDATED

Embedding Geo-Entities for Semantic Typing

- Auxiliary (Appendix A):
  - Generate a lightweight taxonomy from OSM tag data

OpenStreetMap data/dump

```
<way id="232250107" visible="true" vers
2019-05-06T23:22:23Z" user="Enock4seth"
 <nd ref="5058536215"/>
 <nd ref="1797433673"/>
 <nd ref="4992821222"/>
 <tag k="highway" v="tertiary"/>
 <tag k="name" v="Nana Kana Street"/>
</way>
<way id="244376453" visible="true" vers
2015-04-02T14:55:17Z" user="sidneys" ui
 <nd ref="2517024878"/>
 <nd ref="2517024879"/>
 <nd ref="2517024880"/>
 <nd ref="2517024881"/>
 <nd ref="2517024878"/>
 <tag k="building" v="industrial"/>
</way>
<way id="244376454" visible="true" vers
2015-04-02T13:43:25Z" user="sidneys" ui
```

construct base terminology

frequent non-informative

infrequent informative

```
{'apartments',
 'building',
 'driveway',
 'highway',
 'house',
 'residential',
 'service'}
```

count parent-child relations

path frequency

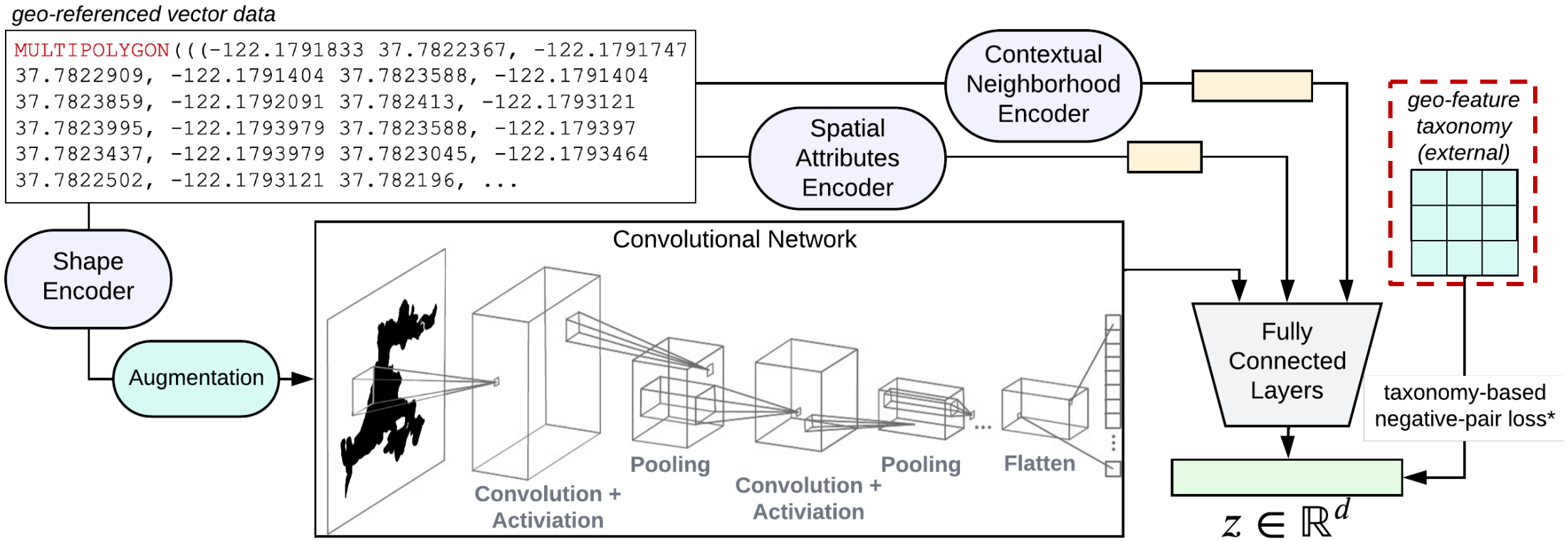
parent	child	counter
building	house	15
highway	service	14
building	residential	33
highway	residential	22
building	apartments	2
service	driveway	5

build taxonomy

conflict resolution

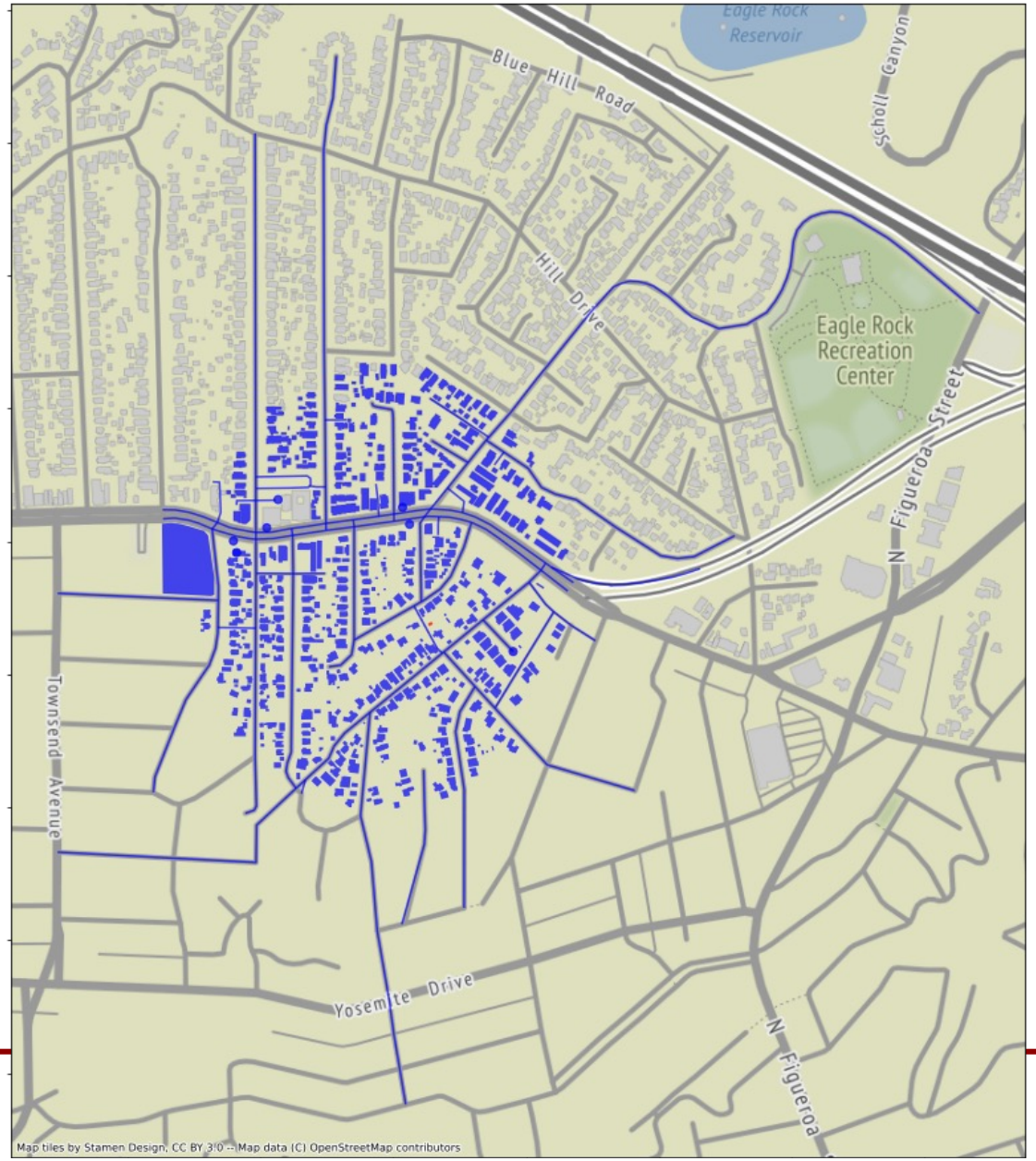
```
OSM
├── building
│   ├── apartments
│   ├── house
│   └── residential__building
└── highway
    ├── residential__highway
    └── service
        └── driveway
```

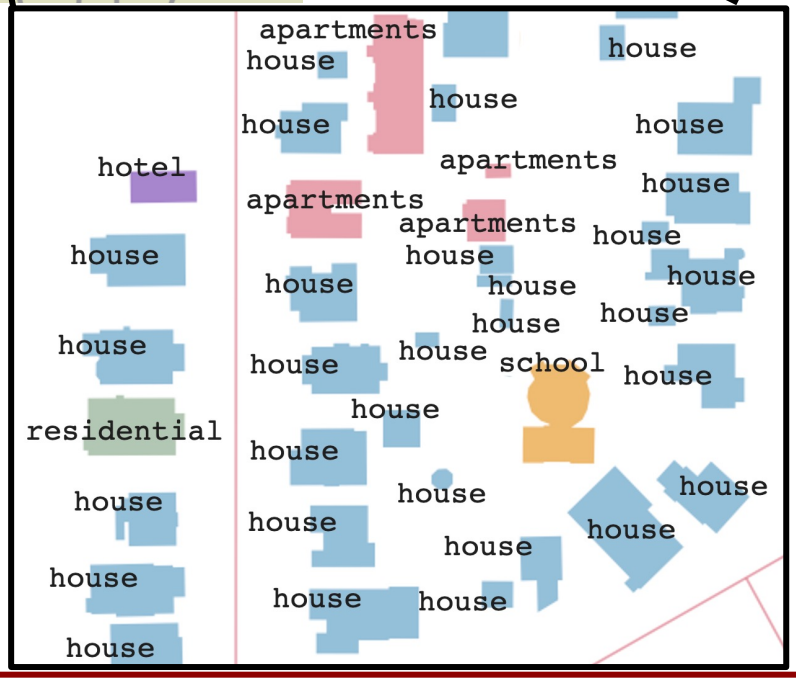
Embedding Geo-Entities for Semantic Typing



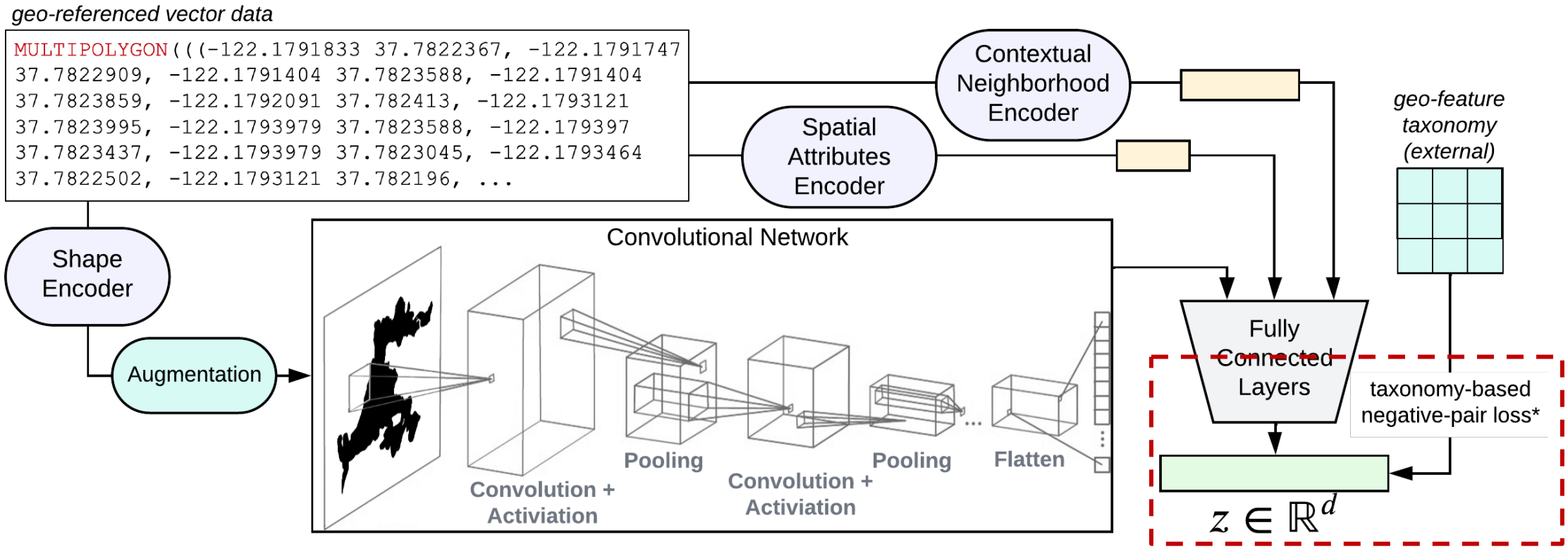
Embedding Geo-Entities for Semantic Typing

house	414
residential__building	57
primary	29
apartments__building	28
residential__highway	21
service	11
commercial__building	8
retail__building	6
hotel__tourism	6
platform	4
alley	3
hotel__building	2
school__building	2
tertiary	2
industrial__building	1
steps	1
warehouse	1
restaurant	1
turning_circle	1
motorway_link	1
place_of_worship	1
driveway	1
school__amenity	1

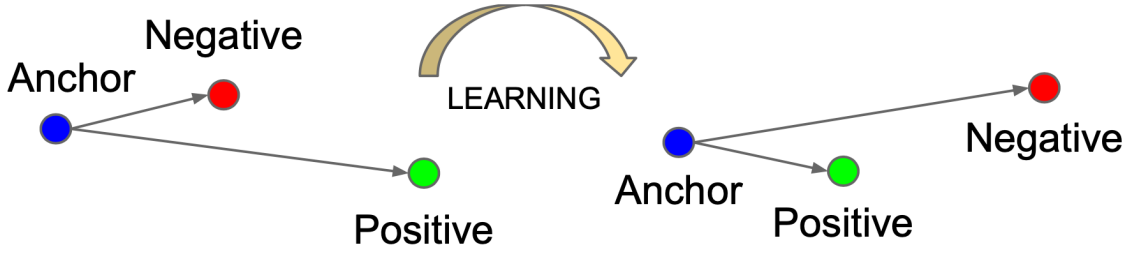




Embedding Geo-Entities for Semantic Typing

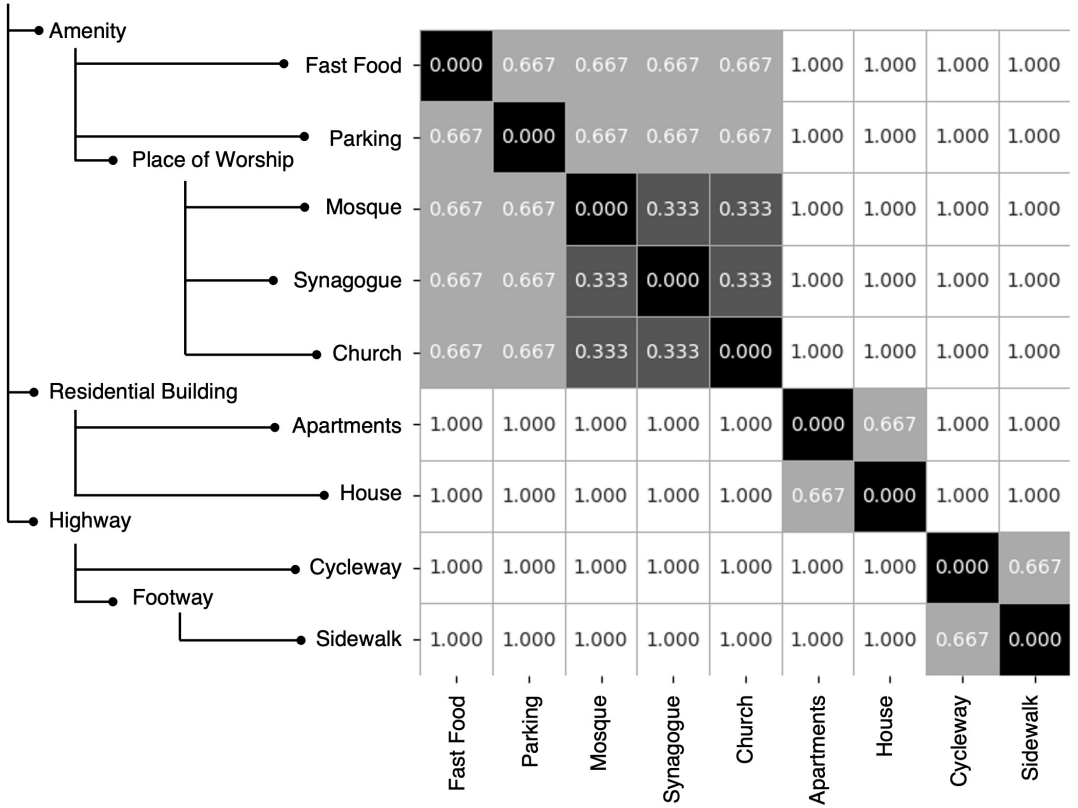


Embedding Geo-Entities for Semantic Typing



$$L_q = -\log \frac{\exp(sim(e_q, e_+)/\tau)}{\sum_{i=0}^{K} \exp(sim(e_q, e_i) \cdot w_{q,i}/\tau)}$$

$$w_{i,j} = \frac{d_{tree} - d_{i,j}}{d_{tree}}$$



	Fast Food	Parking	Mosque	Synagogue	Church	Apartments	House	Cycleway	Sidewalk
Fast Food	0.000	0.667	0.667	0.667	0.667	1.000	1.000	1.000	1.000
Parking	0.667	0.000	0.667	0.667	0.667	1.000	1.000	1.000	1.000
Mosque	0.667	0.667	0.000	0.333	0.333	1.000	1.000	1.000	1.000
Synagogue	0.667	0.667	0.333	0.000	0.333	1.000	1.000	1.000	1.000
Church	0.667	0.667	0.333	0.333	0.000	1.000	1.000	1.000	1.000
Apartments	1.000	1.000	1.000	1.000	1.000	0.000	0.667	1.000	1.000
House	1.000	1.000	1.000	1.000	1.000	0.667	0.000	1.000	1.000
Cycleway	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.000	0.667
Sidewalk	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.667	0.000

Embedding Geo-Entities for Semantic Typing

- Evaluation
  - 8-fold SVC on embeddings

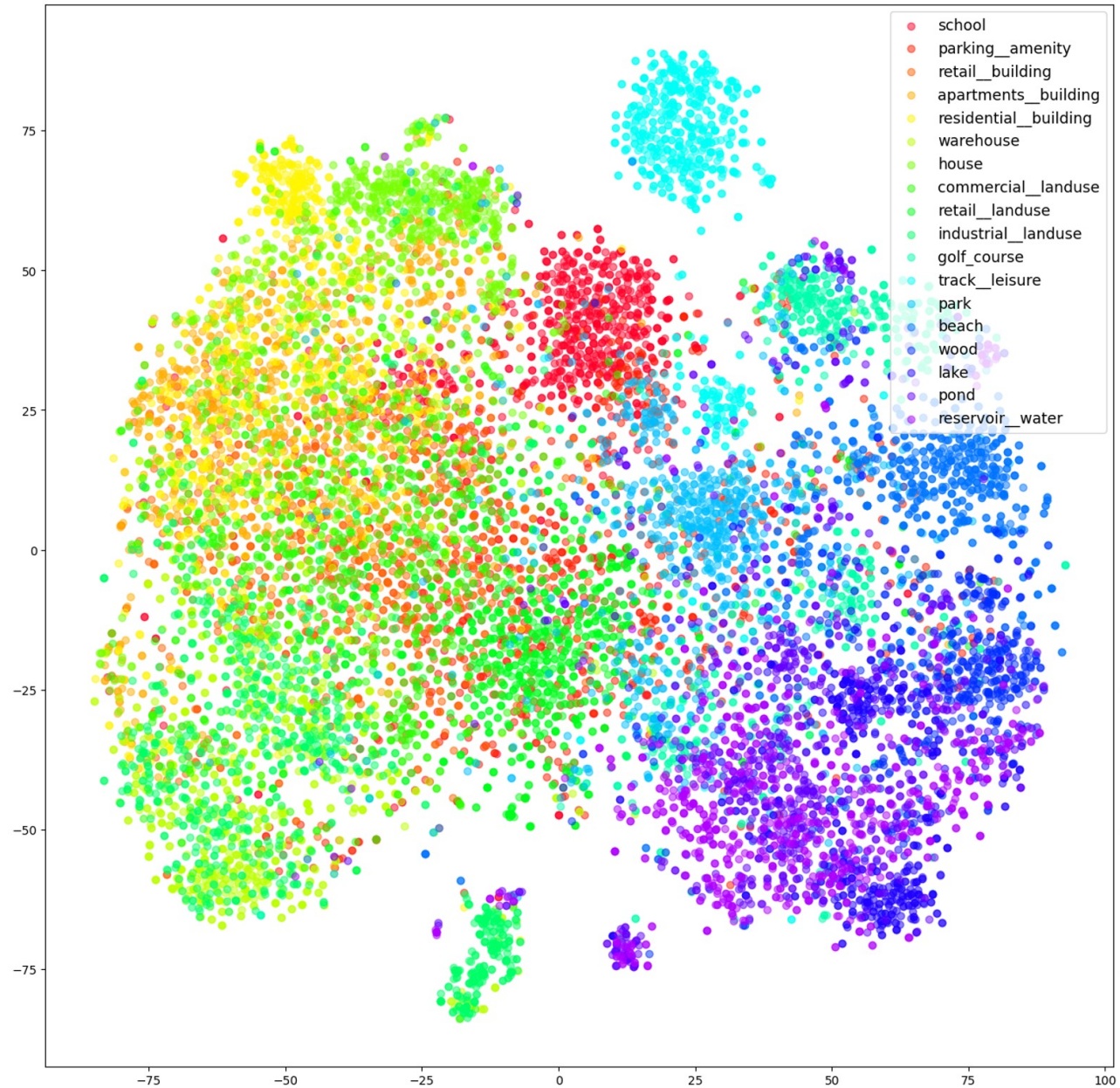
Data: 2k+ instances → 11 WD classes

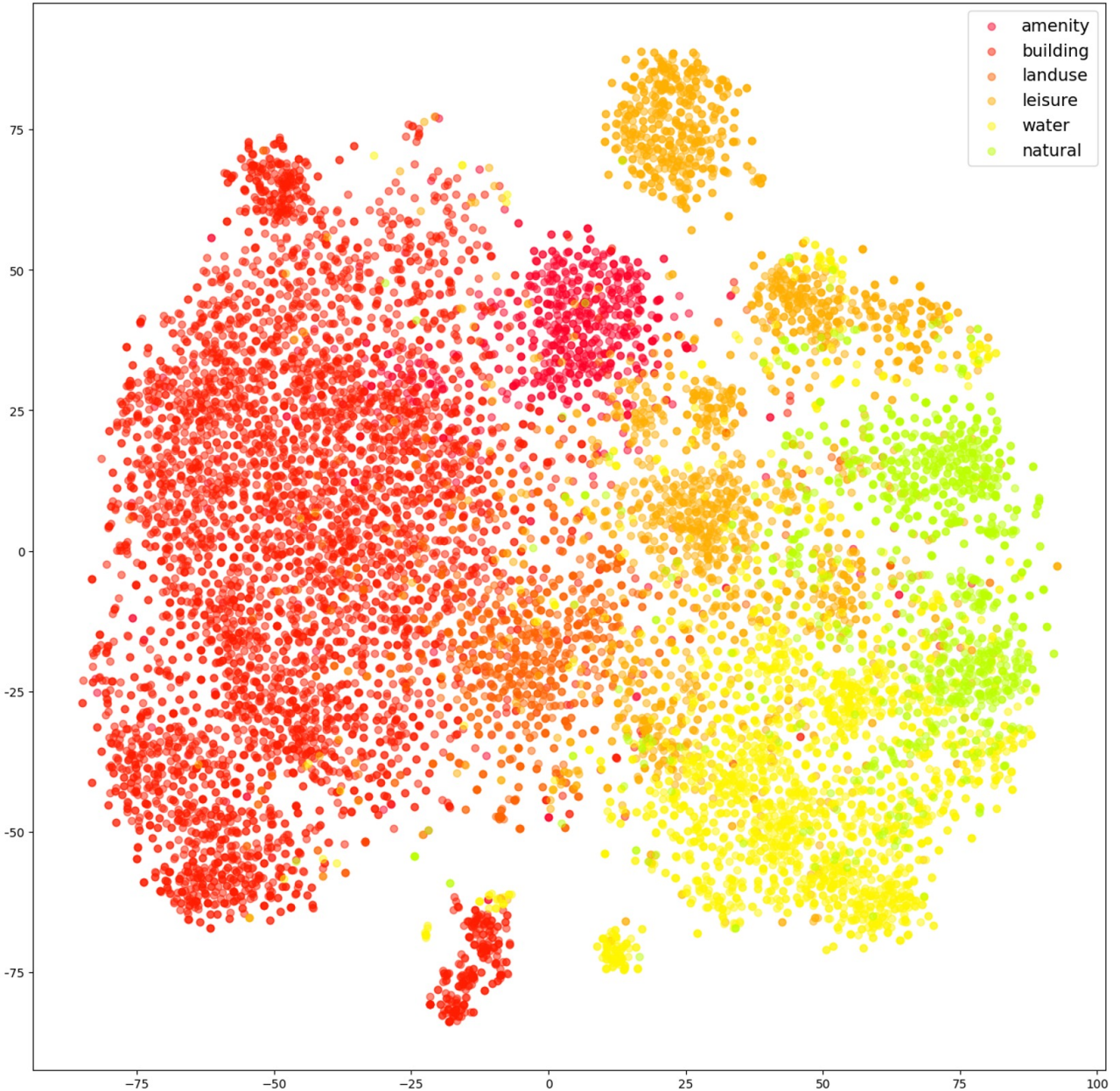
16k+ instances → 18 OSM tags

Training: 200k CA OSM dump (2.3 tags avg)

4 settings

	Setting	WD-2k Precision	WD-2k Recall	WD-2k F_1	OSM-16k Precision	OSM-16k Recall	OSM-16k F_1
1	$\text{Ours}_{\texttt{shape}}$	0.497	0.506	0.501	0.473	0.512	0.492
2	$\text{Ours}_{\texttt{shape+spatial}}$	0.506	0.545	0.525	0.491	0.536	0.513
3	$\text{Ours}_{\texttt{full}}$	**0.850**	0.823	0.836	**0.877**	0.725	0.794
4	$\text{Ours}_{\texttt{full w/taxonomy}}$	0.849	**0.852**	**0.850**	0.858	**0.854**	**0.856**
	GPT-3.5-Turbo	0.198	0.209	0.121	0.145	0.063	0.026
SotA	GeoVectors	0.819	0.834	0.826	0.833	0.815	0.824


school
parking__amenity
retail__building
apartments__building
residential__building
warehouse
house
commercial__landuse
retail__landuse
industrial__landuse
golf_course
track__leisure
park
beach
wood
lake
pond
reservoir__water

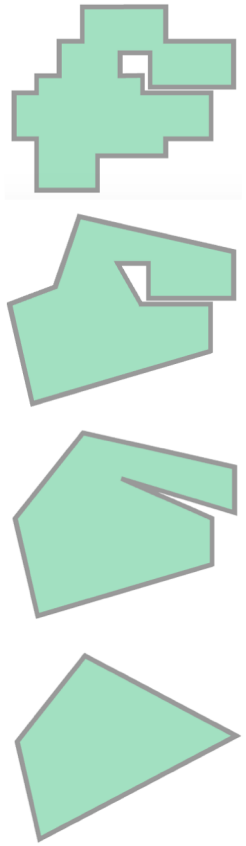

amenity
building
landuse
leisure
water
natural

Embedding Geo-Entities for Semantic Typing

- Related Work
  - ML for Geospatial Classification (Castelluccio 2015, Klemmer 2023, Kaczmarek 2023, Xu 2022, Yan 2021)
    - Employ CNNs, GNNs, and GCNs for: building footprints & urban land-use classification
    - Do not address the incorporation of external (open) knowledge
  - Geospatial Embedding Techniques (Tempelmeier 2021, Jenkins 2019, Li 2022)
    - Develop unsupervised embedding such as *GeoVectors* & *SpaBert*
    - Do not address shape or explicit spatial data for enhanced geo-entity representation
  - OSM Embedding (Woźniak 2021)
    - Proposes embedding method for OSM regions using hexagonal grids
    - Does not address individual entities

Embedding Geo-Entities for Semantic Typing

- Takeaways
  - Method for geo-referenced entity embedding on the web
    - self-supervised
    - leverages geometric, spatial, & semantic contexts
    - weighted contrastive learning
    - enables seamless semantic typing for integration on the web
      - fuels further discovery & enrichment
  - Still, many challenges exist
    - availability
    - granularity
    - quality



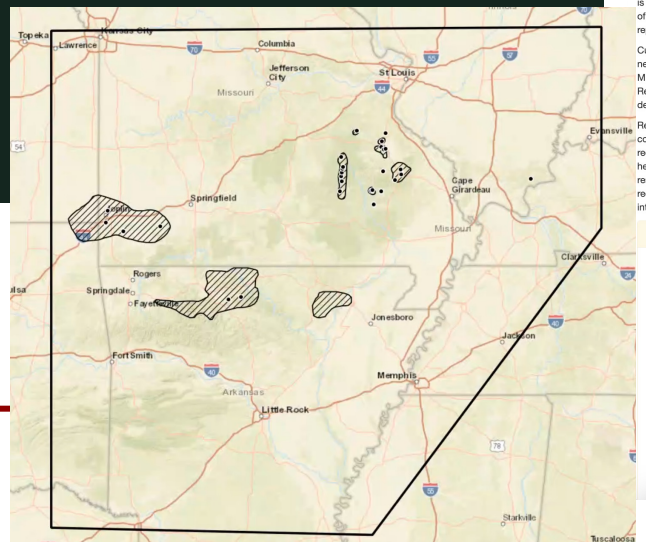
Agenda

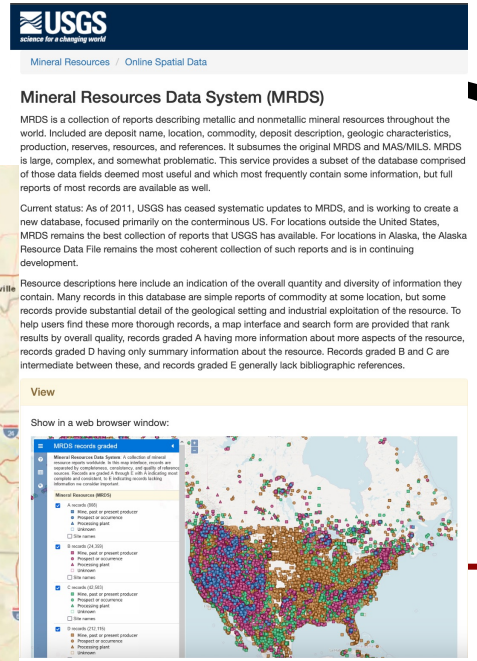
- Basel's PhD Journey
- Intro
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions

From Digitized Reports to Spatio-Temporal KGs

- Goals:
  - Integrating geo-referenced textual & historical data with quantitative information into a comprehensive, dynamic, & spatio-temporal KG
    - capture data & entity semantics, entity resolution, & accurate data modeling
  - Demonstrate via a KG of historical mining data

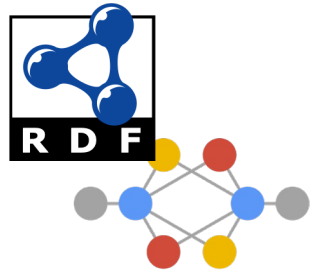

```
{
  "MineralSite": [
    {
      "source_id": "https://w3id.org/usgs/z/4530692/2P29BJHV",
      "record_id": 1,
      "name": "NI 43-101 Technical Report for the Lantinen Koillismaa Project in Europe, Finland dated
March 2017",
      "location_info": {
        "location": "POINT(28.17472 65.94611)",
        "crs": "WGS84",
        "country": "Finland",
        "state_or_province": "Central Finland"
      },
```



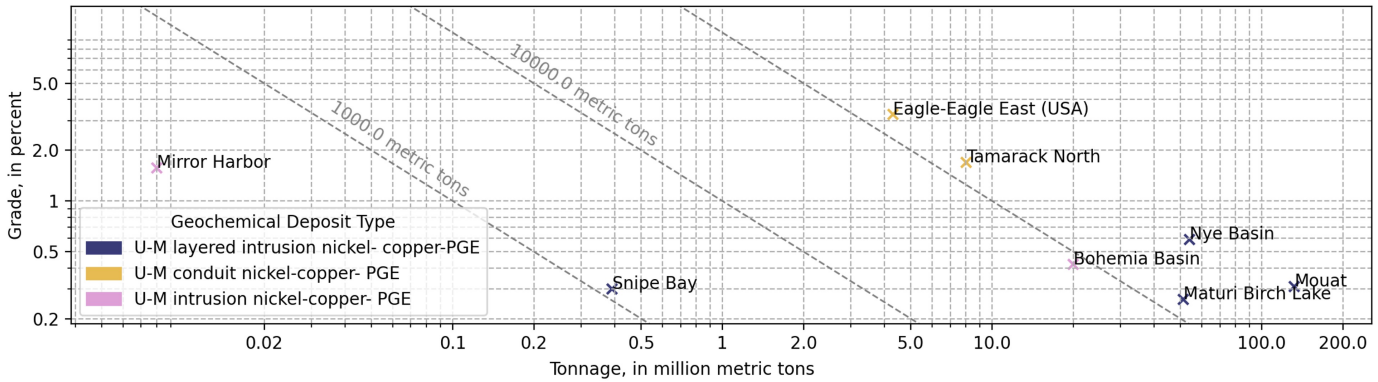
Extracted Textual & Vector Data

Data understanding & KG construction



From Digitized Reports to Spatio-Temporal KGs

- For a given commodity/deposit type/location/time-range:
  - Construct grade and tonnage models from the data on existing mines
  - Compile rich mineral site data



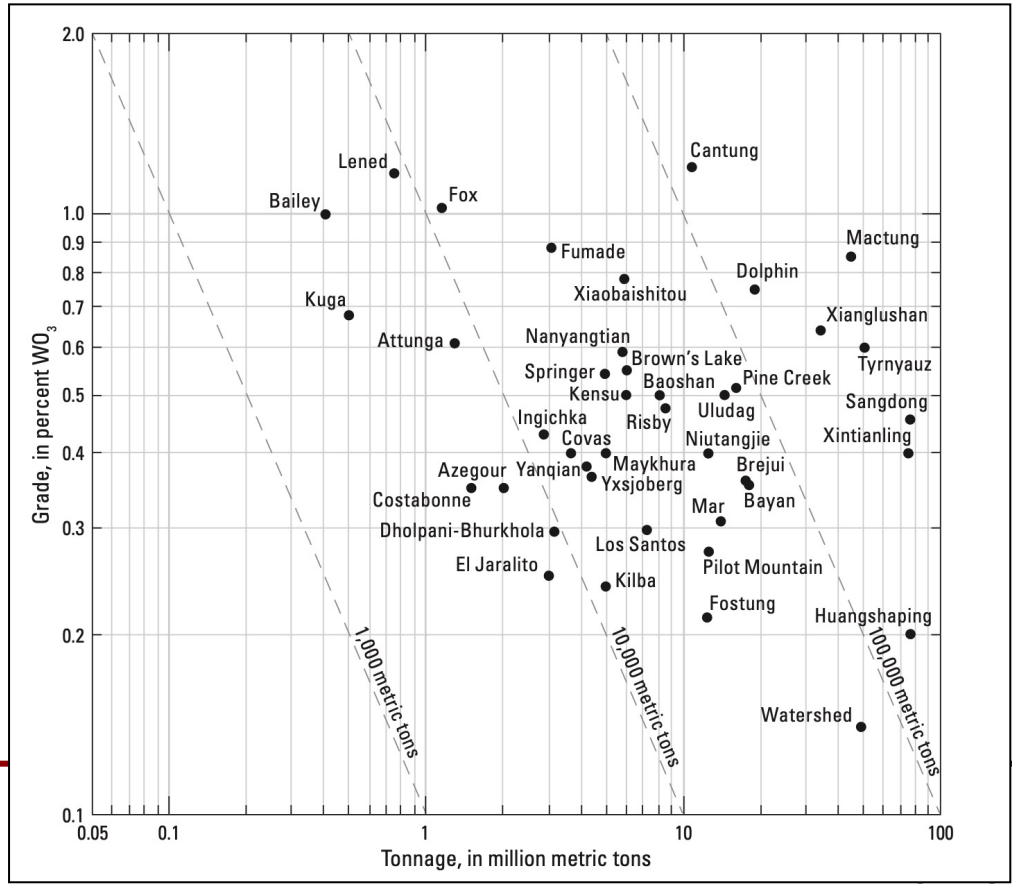
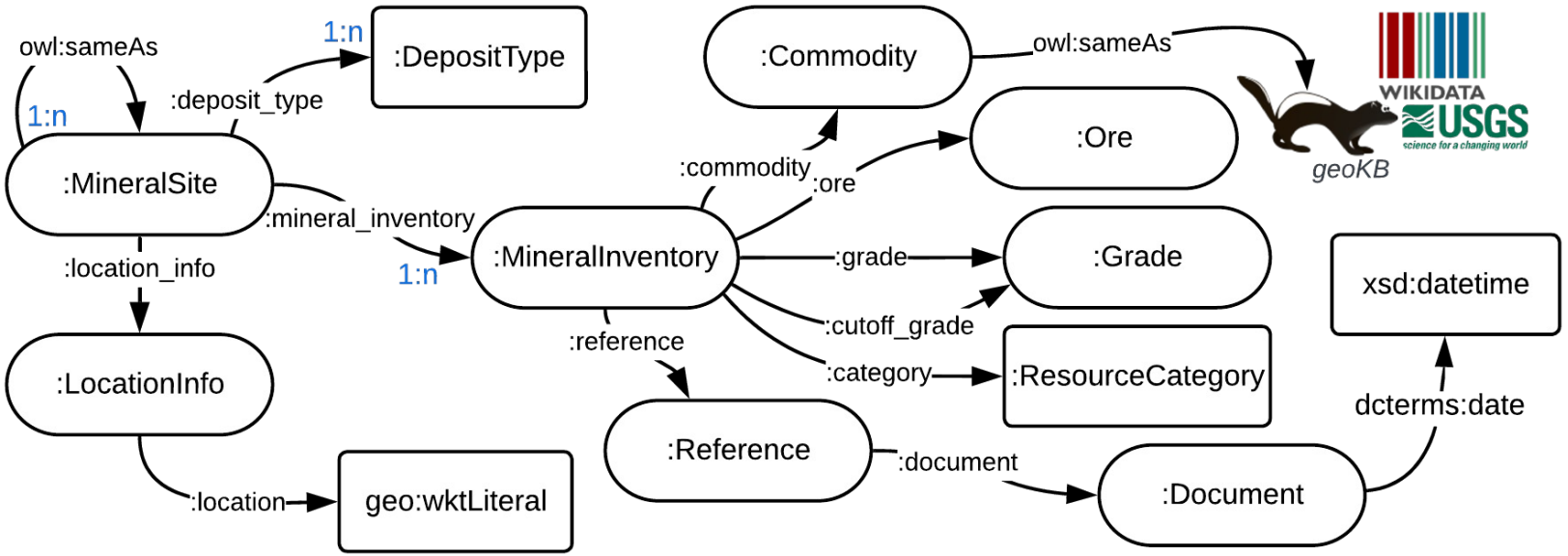


Figure from Green, C. J., Lederer, G. W., Parks, H. L., & Zientek, M. L. (2020). Grade and tonnage model for tungsten skarn deposits—2020 update (No. 2020-5085). US Geological Survey

From Digitized Reports to Spatio-Temporal KGs

- Approach
  - Step 1. Semantic Modeling & URI assignment (Data Representation)
    - Transform & materialize the data (construct KG)
      - Generate entities (URIs) based on unique identifiers
      - Provide a useful semantic representation supporting downstream tasks
    - Construct a meaningful semantic model
      - Follows W3C & OGC standards (GeoSPARQL)



From Digitized Reports to Spatio-Temporal KGs

- Approach
  - Step 2. Entity Linking
    - Link the generated entities to a domain data-rich vocab (i.e., GeoKB)
      - Determine similarity by textual similarity (i.e., Jaccard)
      - Directly within SPARQL

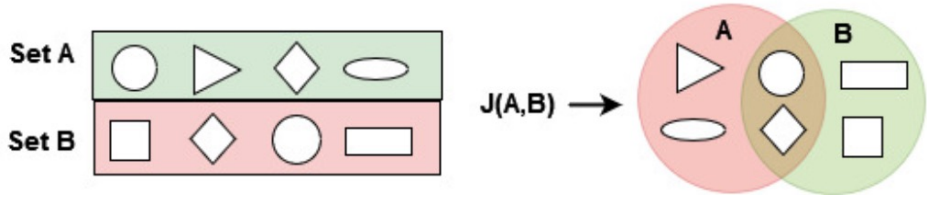
```
"Clay, Fire (Refractory)"
```

		entity.value
0	uri	https://geokb.wikibase.cloud/entity/Q413
1	uri	https://geokb.wikibase.cloud/entity/Q424
2	uri	https://geokb.wikibase.cloud/entity/Q423
3	uri	https://geokb.wikibase.cloud/entity/Q421
4	uri	https://geokb.wikibase.cloud/entity/Q162319
5	uri	https://geokb.wikibase.cloud/entity/Q425
6	uri	https://geokb.wikibase.cloud/entity/Q426
7	uri	https://geokb.wikibase.cloud/entity/Q428
8	uri	https://geokb.wikibase.cloud/entity/Q427
9	uri	https://geokb.wikibase.cloud/entity/Q429

	entityLabel.xml:lang	entityLabel.type	entityLabel.value
0	en	literal	high alumina clay aluminum
1	en	literal	bloating material clay
2	en	literal	brick clay
3	en	literal	clay
4	en	literal	Clay
5	en	literal	bentonite clay
6	en	literal	chlorite clay
7	en	literal	fire (refractory) clay
8	en	literal	fullers earth clay
9	en	literal	glauconite clay

```
1  SELECT ?entity ?entityLabel WHERE {
2    ?entity rdfs:label ?entityLabel.
3    ?entity gkbt:P1 gkbi:Q406. # instance of mineral commodity
4    FILTER(CONTAINS(LCASE(?entityLabel), "nickel")) }
```

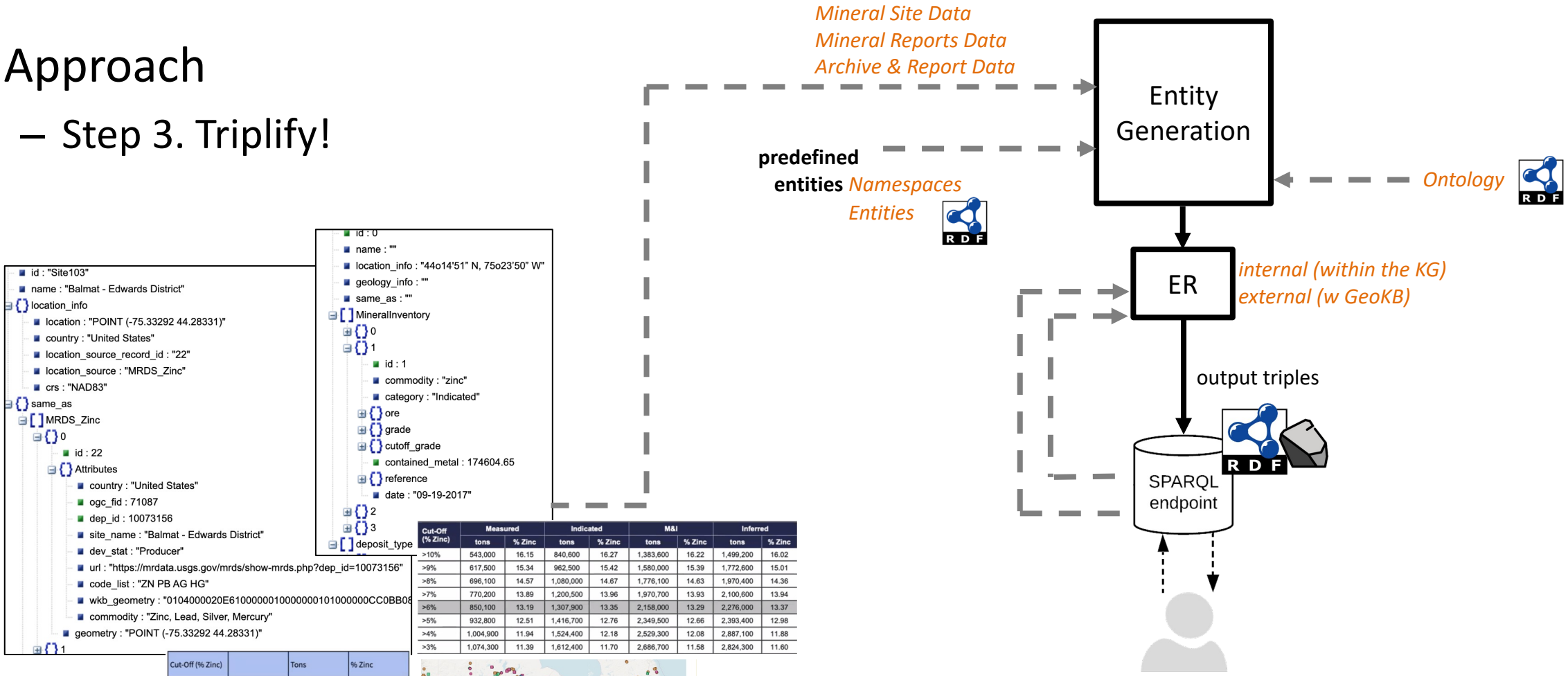
Jaccard



USC Viterbi School of Engineering

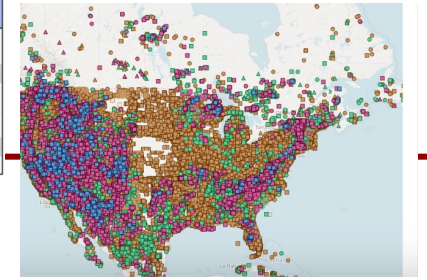
From Digitized Reports to Spatio-Temporal KGs

- Approach
  - Step 3. Triplify!



Cut-Off (% Zinc)	Measured		Indicated		M&I		Inferred	
	tons	% Zinc	tons	% Zinc	tons	% Zinc	tons	% Zinc
>10%	543,000	16.15	840,600	16.27	1,383,600	16.22	1,499,200	16.02
>9%	617,500	15.34	962,500	15.42	1,580,000	15.39	1,772,600	15.01
>8%	696,100	14.57	1,080,000	14.67	1,776,100	14.63	1,970,400	14.36
>7%	770,200	13.89	1,200,500	13.96	1,970,700	13.93	2,100,600	13.94
>6%	850,100	13.19	1,307,900	13.35	2,158,000	13.29	2,276,000	13.37
>5%	932,800	12.51	1,416,700	12.76	2,349,500	12.66	2,393,400	12.98
>4%	1,004,900	11.94	1,524,400	12.18	2,529,300	12.08	2,887,100	11.88
>3%	1,074,300	11.39	1,612,400	11.70	2,686,700	11.58	2,824,300	11.60

Cut-Off (% Zinc)		Tons	% Zinc
>10%	Measured	543,000	16.15
	Indicated	840,600	16.27
	M&I	1,383,600	16.22
	Inferred	1,499,200	16.02
>9%	Measured	617,500	15.34
	Indicated	962,500	15.42
	M&I	1,580,000	15.39
	Inferred	1,772,600	15.01



From Digitized Reports to Spatio-Temporal KGs

- Evaluation
  - Data completeness (SHACL)
  - EL
  - RDF Query Performance

Data: 2.4m triples // 135 commodities // focus on 2 critical mineral: nickel, zinc

Characteristic	Count
Total Triples	2,397,708
Distinct Classes	16
Instances (Non-literals)	226,267
Geospatial Instances	2,884
Blank Nodes	1,518,981

Method	MRR	Hits@1	Hits@3	Hits@5
String search, then Jaro	0.557	0.459	0.659	0.659
String search, then Jaccard	0.648	0.637	0.659	0.659
Instance search, then Jaro	0.801	0.689	0.926	0.956
Instance search, then Jaccard (proposed)	**0.940**	**0.904**	**0.978**	**0.978**

```
1 ?ms :location_info/:location ?loc_wkt .
2 FILTER(geof:distance(?loc_wkt, "POINT(-118.57 47.56)"^^geo:wktLiteral, unit:mile) < 500)
```

Query Constraint Type	Avg	Min	Max
Textual	450	369	649
Temporal/Numeric	438	388	607
Spatial	708	501	811

From Digitized Reports to Spatio-Temporal KGs

- Related Work
  - General geo KBs (Zhu 2017, Brodaric 2020)
    - Mostly encompasses conceptual knowledge & data
    - Does not address: quantitative data integration
  - GeoKGs related to mineral data (Qun 2023)
    - Tailored for geochemical data
    - Does not address: quantitative data integration
  - Information extraction for geo KGs (Wang 2018)
    - Focus is on the data extraction
    - Does not address: data integration & entity linking

From Digitized Reports to Spatio-Temporal KGs

- Takeaways
  - Paradigm for geospatial data integration from digitized textual & geo-refenced archive data
  - Method to identify & retrieve instances of a given type from a publicly available KG
    - Automated
    - Incremental semantic model for simple & efficient querying
      - Follows LD & SW principles
    - Linked to Web
      - Fuels further discovery & enrichment
  - Still, many challenges exist
    - Quality of extracted data
    - Encoding domain knowledge

Putting it togg

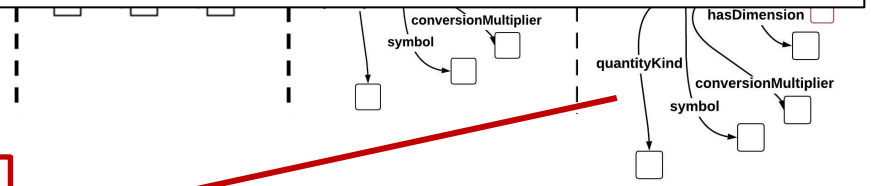
Grade	Assays					
	Cu (%)	Ni (%)	S (%)	Au (g/t)	Pt (g/t)	
Concentrate	7.16-10.1	1.66-2.20	18.4-21.5	0.65-1.28	1.17-1.59	

2nd aggregation: **total** grade & tonnage computation

1st aggregation: tonnage computation

```
    ?total_contained_indicated ?total_contained_inferre
    (?total_tonnage_measured + ?total_tonnage_indicated
    (?total_contained_measured + ?total_contained_indic
    (IF(?total_tonnage > 0, ?total_contained_metal / ?t
WHERE {
  {
    SELECT ?ms ?ms_name ?deposit_name ?country ?loc_wkt
           (SUM(?tonnage_measured) AS ?total_tonnage_meas
           (SUM(?tonnage_indicated) AS ?total_tonnage_ind
           (SUM(?tonnage_inferred) AS ?total_tonnage_infe
           (SUM(?contained_measured) AS ?total_contained_
           (SUM(?contained_indicated) AS ?total_contained
           (SUM(?contained_inferred) AS ?total_contained_
    WHERE {
        ?ms :deposit_type [ rdfs:label ?deposit_name ] .
        ?ms :mineral_inventory ?mi .
        OPTIONAL { ?ms rdfs:label|:name ?ms_name . }
        ?ms :location_info/:location ?loc_wkt .
        ?mi :category ?mi_cat .
        ?mi :sample/:collection_info/:collection_date/:date ?date ] .
        FILTER(geof:sfWithin(?loc_wkt, "POLYGON(...)"^^geo:wktLiteral))
        FILTER(geof:distance(?loc_wkt, "POINT(-118.57 47.56)"^^geo:wktLiteral, unit:mile) < 500)
        FILTER(?date >= "2000"^^xsd:gYear && ?date <= "2010"^^xsd:gYear) .
        ?mi :ore [ :ore_value ?ore_val_raw; :ore_unit ?ore_unit] .
        ?mi :grade [ :grade_value ?grade_val; :grade_unit ?grade_unit] .
        BIND(IF(bound(?ore_val_raw), ?ore_val_raw, 0) AS ?ore_val_pre)
        BIND(IF(?ore_unit = <http://data.nasa.gov/qudt/owl/unit#MetricTon>, ?ore_val_pre / 1e6, ?ore_val_pre)) AS ?ore_val)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "measured"), ?ore_val, 0) AS ?tonnage_measured)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "indicated"), ?ore_val, 0) AS ?tonnage_indicated)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "inferred"), ?ore_val, 0) AS ?tonnage_inferred)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "measured") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_measured)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "indicated") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_indicated)
        BIND(IF(CONTAINS(LCASE(STR(?mi_cat)), "inferred") && ?grade_val > 0, ?ore_val * ?grade_val, 0) AS ?contained_inferred)
    }
    GROUP BY ?ms ?ms_name ?deposit_name ?loc_wkt }
```

```
qudt:hasDimension: "",
qudt:abbreviation": "g tonne-1",
..
qudt:hasPart: [
{
    qudt:hasDimension: "M",
    qudt:quantityKind: "http://data.nasa.gov/qudt/owl/unit#Gram",
    qudt:conversionMultiplier: 0.001,
    qudt:conversionOffset: 0.0,
    qudt:symbol: "g"
  },
  {
    ccut:exponent: "-1",
    qudt:hasDimension: "M",
    qudt:quantityKind: "http://data.nasa.gov/qudt/owl/unit#MetricTon",
    qudt:conversionMultiplier: 1000.0,
    qudt:conversionOffset: 0.0,
    qudt:symbol: "t"
  }
]
```



Agenda

- Basel's PhD Journey
- Intro
- Thesis Overview
- Approach:
  - Building Spatio-Temporal KGs from Digitized Maps
  - Embedding Geo-Entities for Semantic Typing
  - From Digitized Reports to Spatio-Temporal KGs
- Conclusions & Future Directions ⟸

Conclusions

- Presented my thesis, based on 3 main contributions
  - Paradigm for automatic transformation of historical maps into dynamic spatio-temporal KGs
    - evaluated via change analysis of 2 different topographic features over time
  - Approach for accurate geo-entity embedding, classification & integration
    - evaluated via semantic typing of geo-referenced digitized instances to two different Open KBs
  - Method for the construction of a spatio-temporal KG from geo-referenced spatial entities in archive reports
    - demonstrated & evaluated on different tasks in the domain of historical mining data
- Presented 2 additional (auxiliary) contributions
  - Method & tool to generate geo-feature taxonomies from public OSM data
  - Approach & tool to identify, ontologize & transform units of measurement for quantitative data

Future Directions

- Advanced data modeling
  - More modalities
  - More data (e.g., rapidly changing geographies)
  - Hyperparameter optimization
- Enhanced embedding techniques
  - Utilize subword information and deep learning attention mechanisms
  - Expand integration of textual data
- KG expansion
  - Extend KG linkage to cover a broader range of knowledge bases & LOD
  - Apply & integrate with additional domains like archaeology & environmental sciences
- Dynamic semantic modeling
  - Create more sophisticated & evolving semantic models for accurate representation across multiple domains

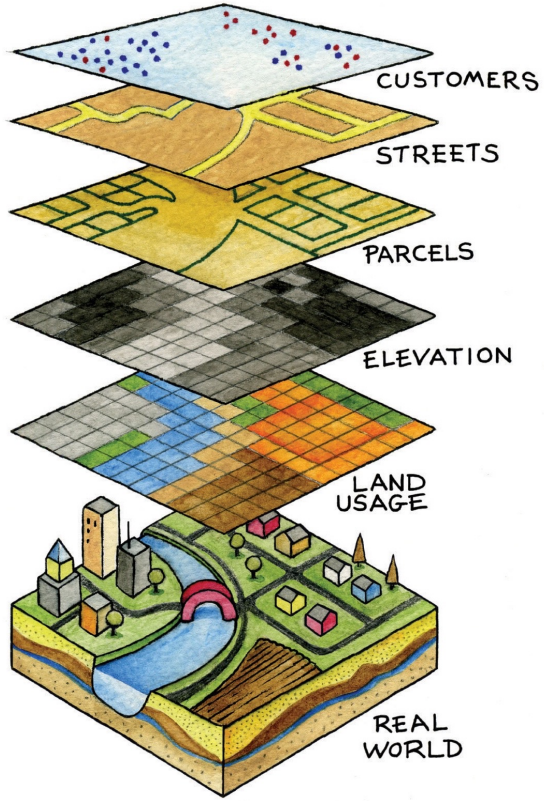


figure from *Essentials of Geographic Information Systems, Ch 7, Saylor Academy, 2012*

Final Remarks

- Thanks to my collaborators!
  - Craig A. Knoblock (advisor)
  - Cyrus Shahabi, John P. Wilson, Jay Pujara, Yao-Yi Chiang (committee)
  - Pedro Szekely, Filip Ilievski, Muhammad Rostami, Jon May (USC/ISI collaborators)
  - Johannes H. Uhl, Stefan Leyk (University of Colorado Boulder)
  - Anna Lisa Gentile, Pengyuan Li, Guang-Jie Ren (IBM Research)
  - Abha Moitra (GE Research)
  - Shui Hu (Amazon Research)
  - USC/ISI colleagues & friends

Thank you for listening!

Questions?