Geographic Knowledge Graph

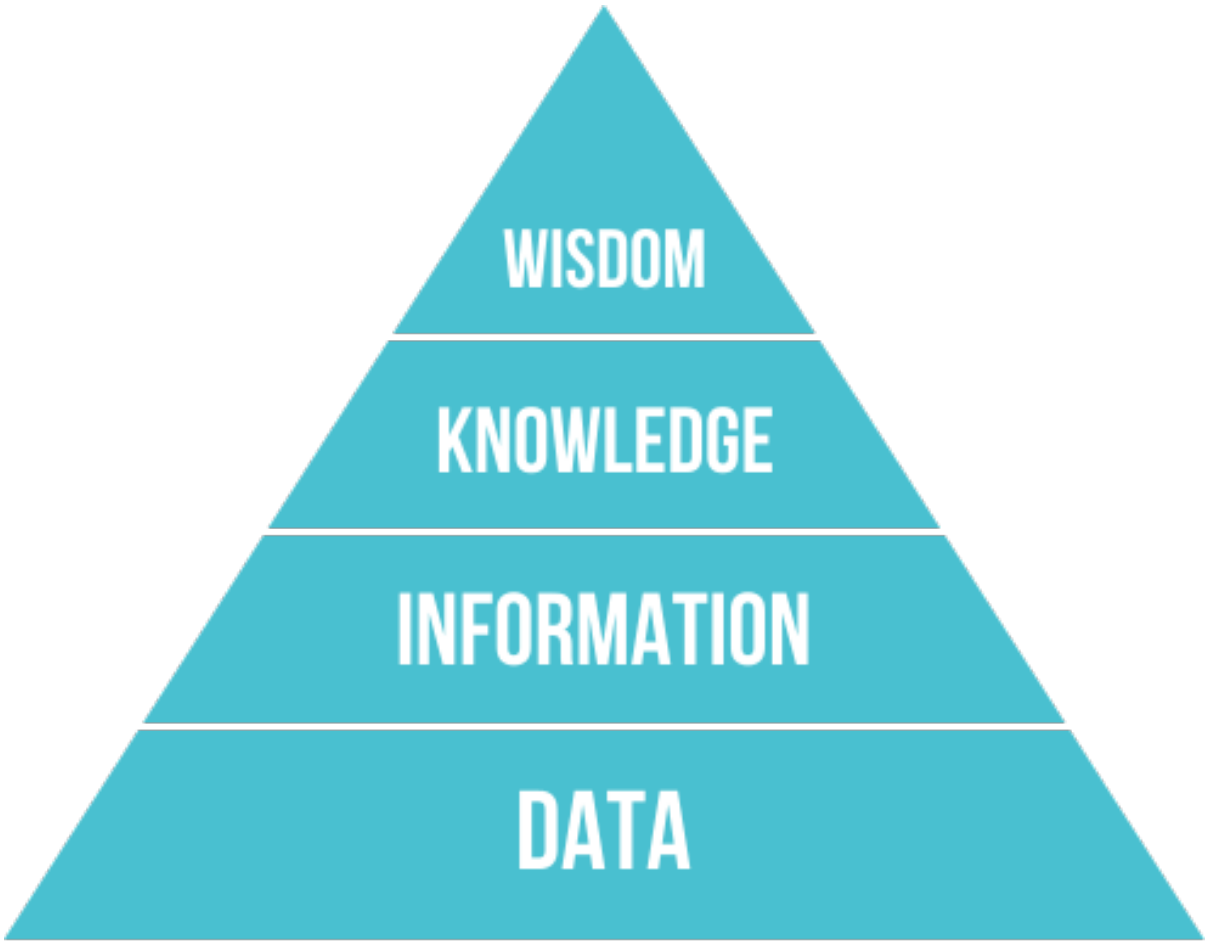
GEOG 176C

Gengchen Mai

STKO Lab, Dept. of Geography, UC Santa Barbara, CA, USA

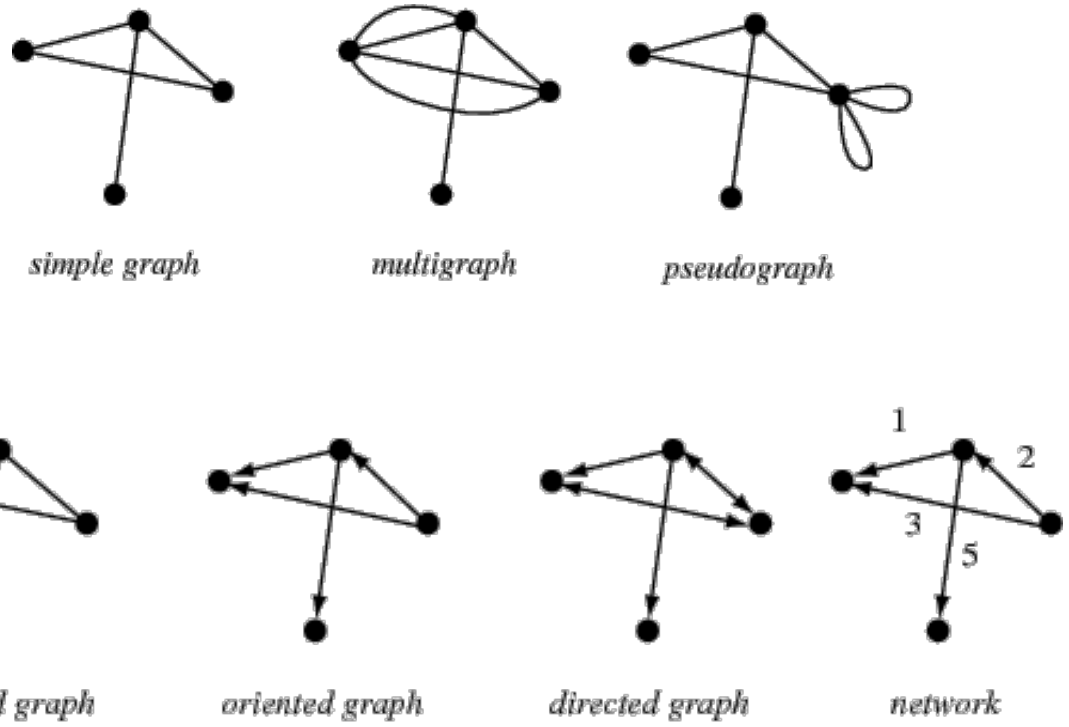
May 21st, 2018

Knowledge



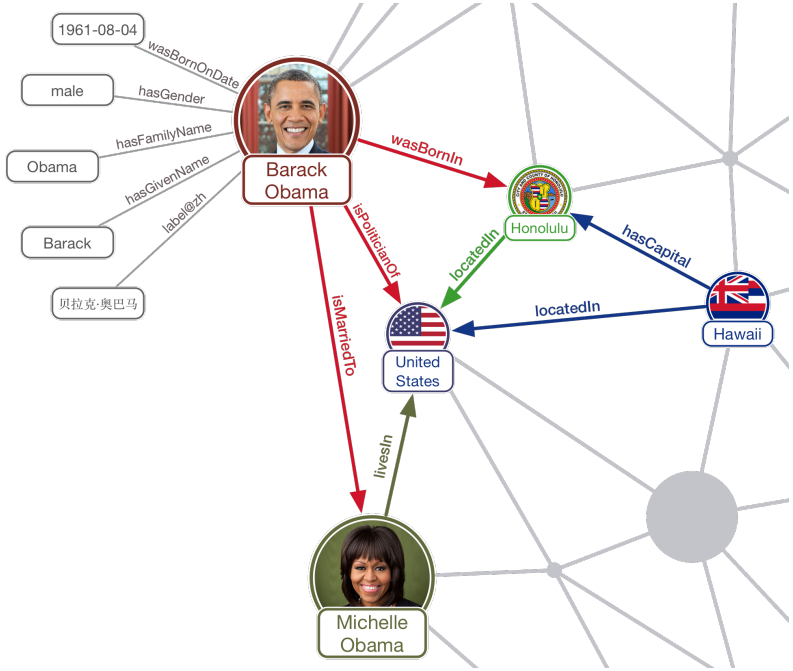
- Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject. (*Oxford Dictionary*)

Graph



- A graph is an ordered pair G = (V, E) ; where V is a set of vertices (nodes) and E is a set of edges (arcs) which are 2-element subsets of V (every edge is connected to two vertices)

Knowledge Graph

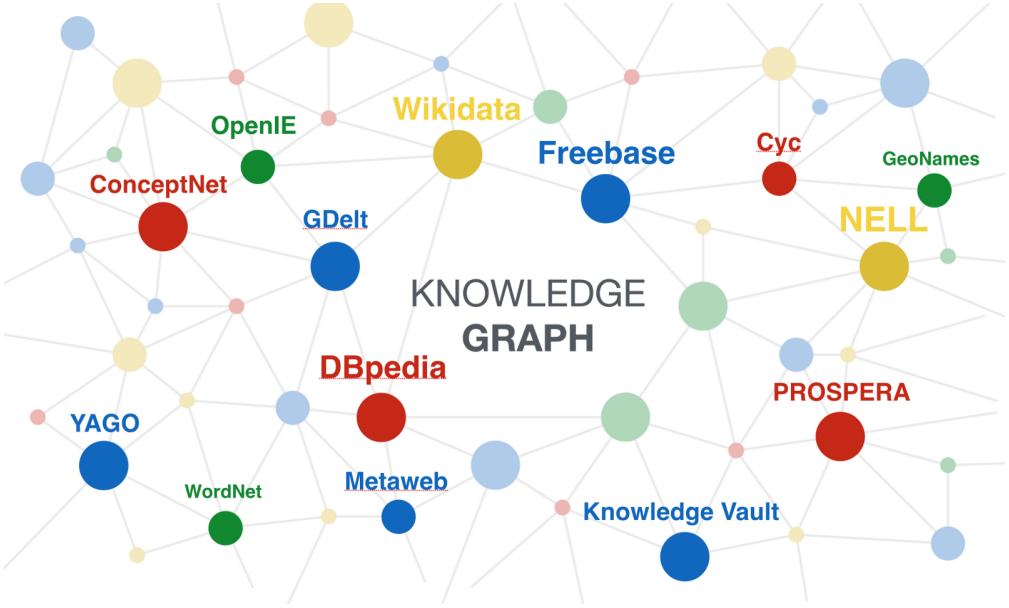


- A KG is typically organized in the form of a graph, e.g., a directed multi-relational graph, such that the nodes represent (real-world) entities and edges represent their relations.
- A knowledge graph (KG) is a data repository that describes entities and their relationships across domains according to some schema, e.g., an ontology

Knowledge Graph

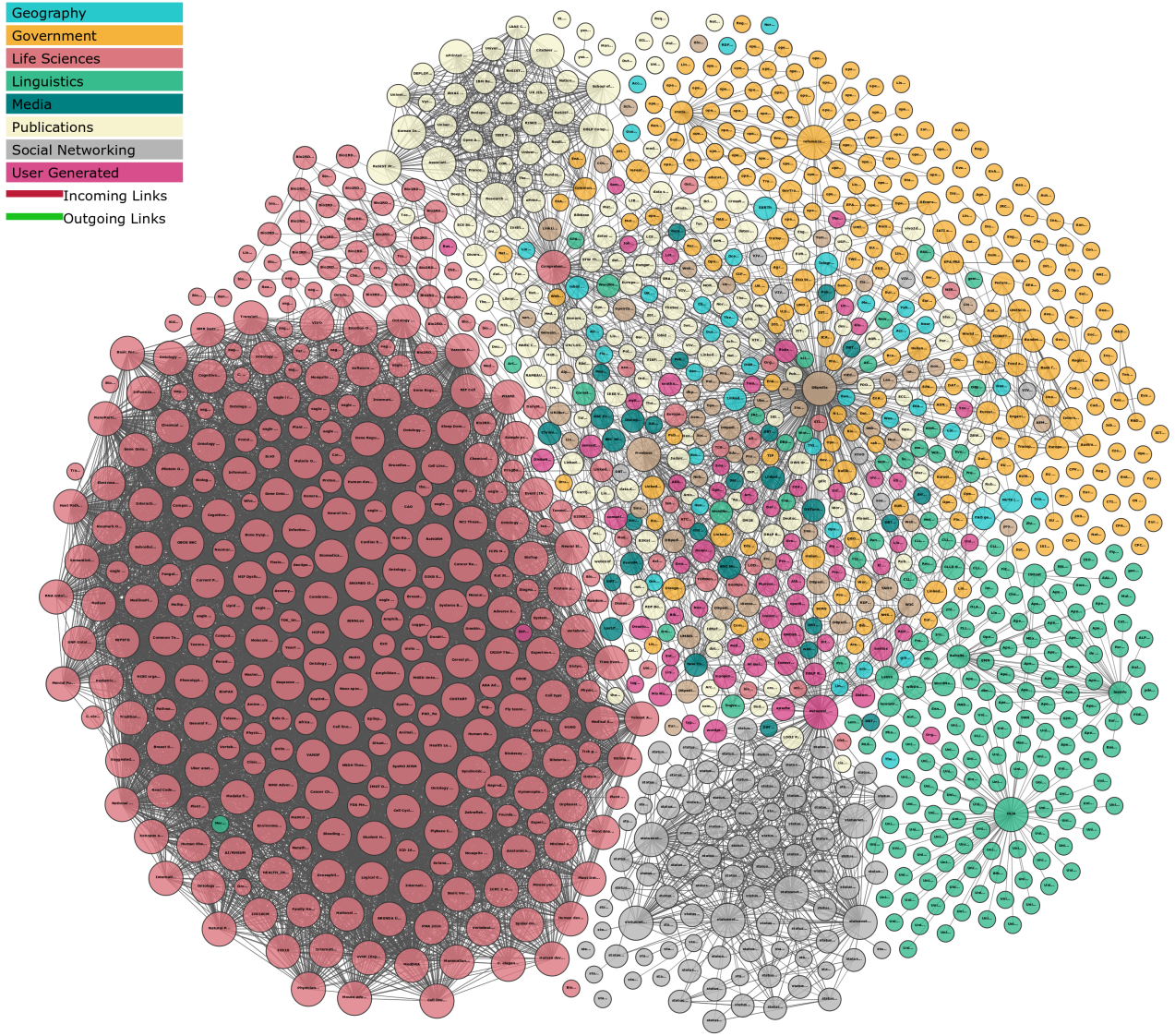
- A collection of triples/statement in the form of (*subject - predicate - object*)
  - (dbr:Santa_Barbara,_California , dbo:isPartOf , dbr:California)
- Example: Google Knowledge Graph, MicrosoftâĂŹs Satori, and Freebase to KGs based on W3C technologies such as DBpedia, YAGO, and Wikidata.

Knowledge Graph



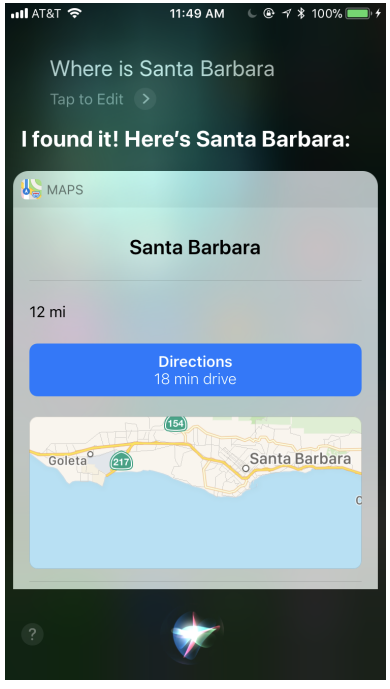
- These data repository can be linked with each other based on ontology alignment (owl:equivalentClass) and instance level alignment (owl:sameAs)
  - (dbr:Place, owl:equivalentClass, schma-org:Place)
  - (dbr:Santa_Barbara,_California, owl:sameAs, freebase:Santa_Barbara,_California)

Knowledge Graph & Linked Data Cloud



Applications of Knowledge Graph

- Questions and Answering, e.g. Apple Siri, Google Duplex

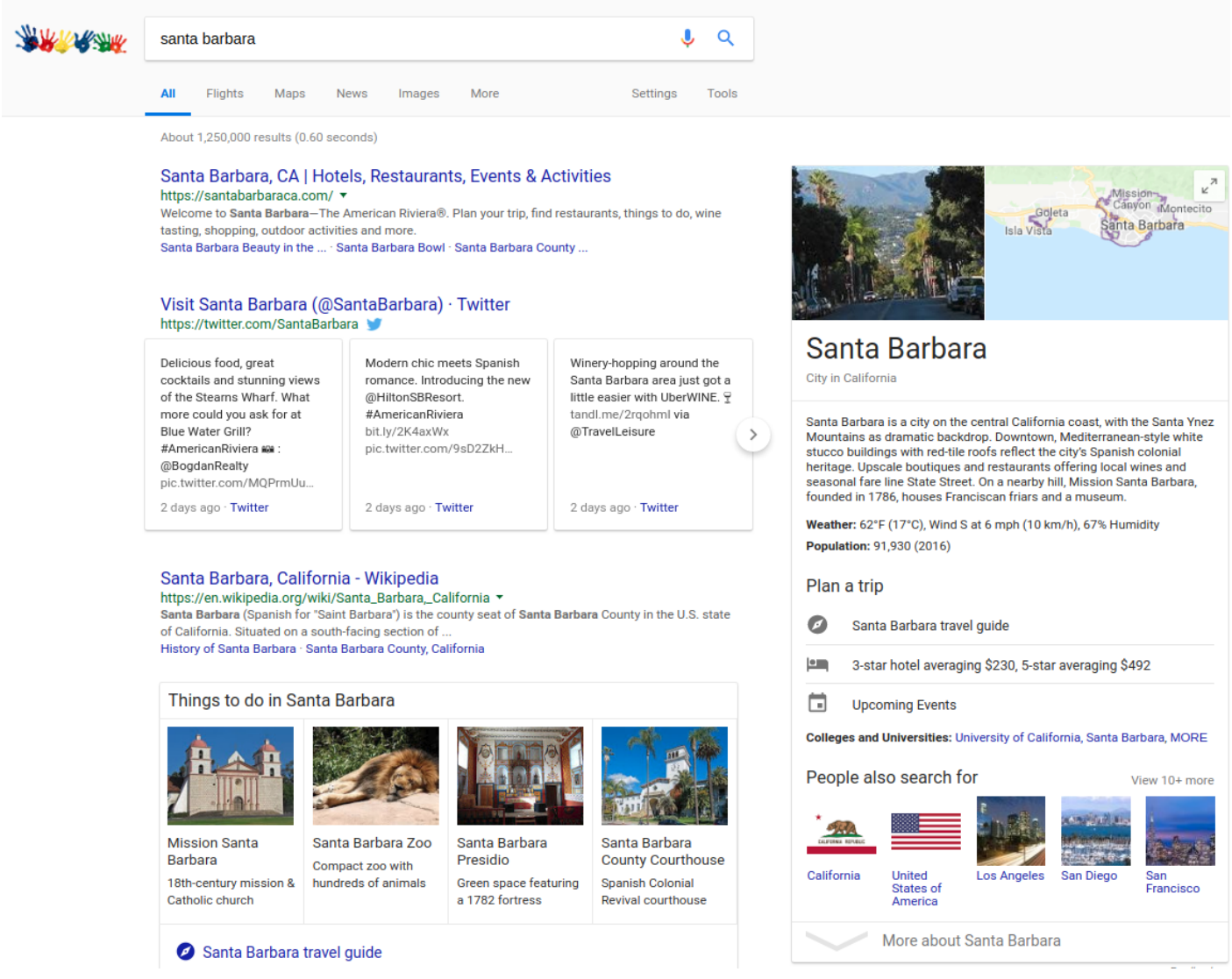


Apple Siri

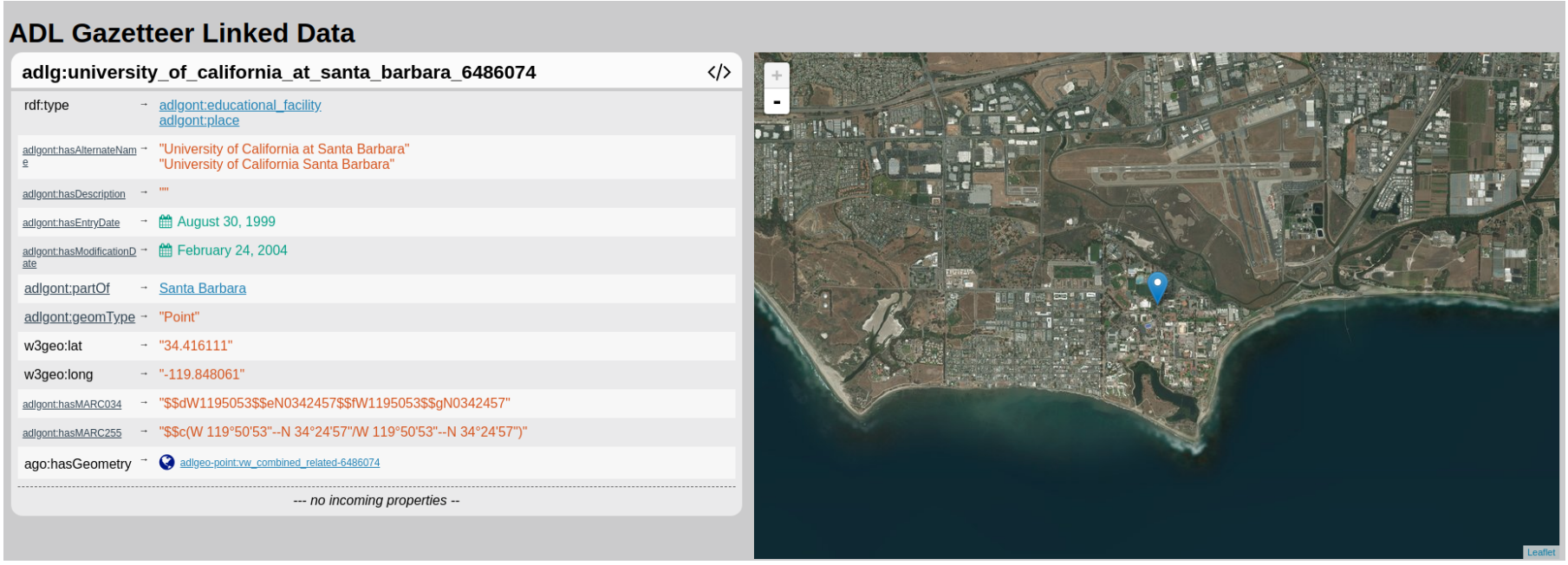
Google Duplex

Applications of Knowledge Graph

- Information Retrieval, e.g. Google Knowledge Graph



Geographic Knowledge Graph



- A **geographic knowledge graphn** is a KG which encodes geographic information.

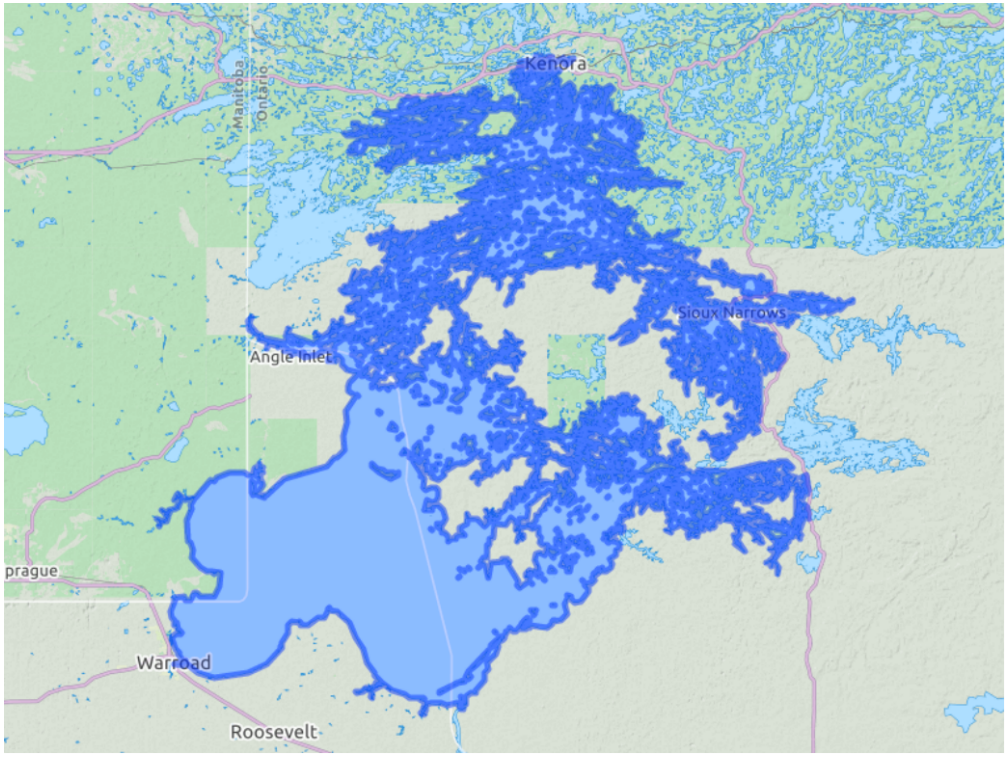
Geographic Knowledge Graph

- Specific problems for geographic knowledge graphs [1]:
  - **How to meaningfully encode geographic information (complex geometries) in a KG? [7]**
  - **How to meaningfully visualize geographic knowledge graphs? [6]**
  - **How to enable GIS computations on geographic knowledge graphs on the fly?**

[1] Some works are done by other STKO members

Representations of Complex Geometries in Geographic KG

- How to serialize complex geometries in geographic knowledge graphs?



Multipart polygon representing the Lake of the Woods [7]

- GeoSPARQL
- Awesemantic-Geo [7]

GeoSPARQL

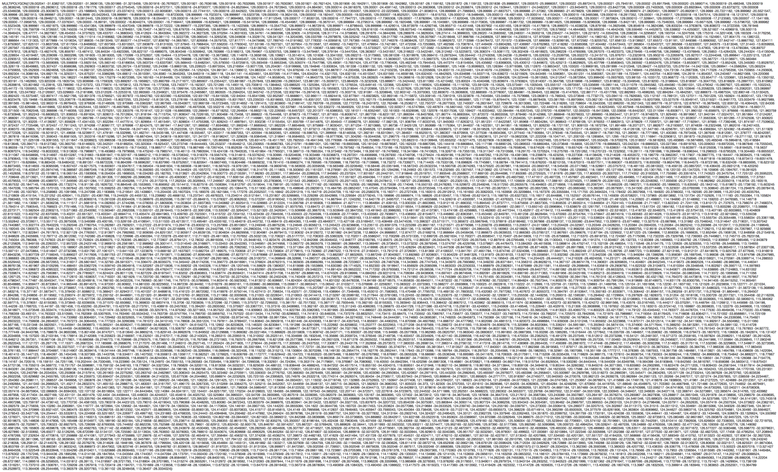
```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix geo: <http://www.opengis.net/ont/geosparql#> .

@prefix ex: <http://www.example.org/POI#> .

@prefix sf: <http://www.opengis.net/ont/sf#> .


ex:NationalMall a ex:Park;

        rdfs:label "National␣Mall";

        geo:hasGeometry ex:NMPoly .

ex:NMPoly a sf:Polygon;

        geo:asWKT "POLYGON((-77.050125␣38.892086,␣-77.039482␣38.892036,␣-77.039482␣
            ↪ 38.895393,␣-77.033669␣38.895508,␣-77.033585␣38.892052,␣-77.031906␣
            ↪ 38.892086,␣-77.031883␣38.887474,␣-77.050232␣38.887142,␣-77.050125␣
            ↪ 38.892086␣))"^^geo:wktLiteral.
```

- Store the entire geometry and the coordinate system in a single RDF literal, thus eliminating any issues brought on by embedding complex structures as RDF

GeoSPARQL

- pros:
  - storing serialized geometry data within RDF triples;
  - supporting coordinate reference systems;
  - maintaining the distinction between entities and their geometric representation;
  - enabling geospatial queries on linked geographic data;
- cons:
  - Challenges associated with the storage and transmission of large WKT strings;
  - Timely execution of SPARQL queries that make use of geospatial functions.

GeoSPARQL



The Well-Known Text string for the geometry of Western Australia taken from Open Street Map

Awesemantic-Geo

base uri geometry ID

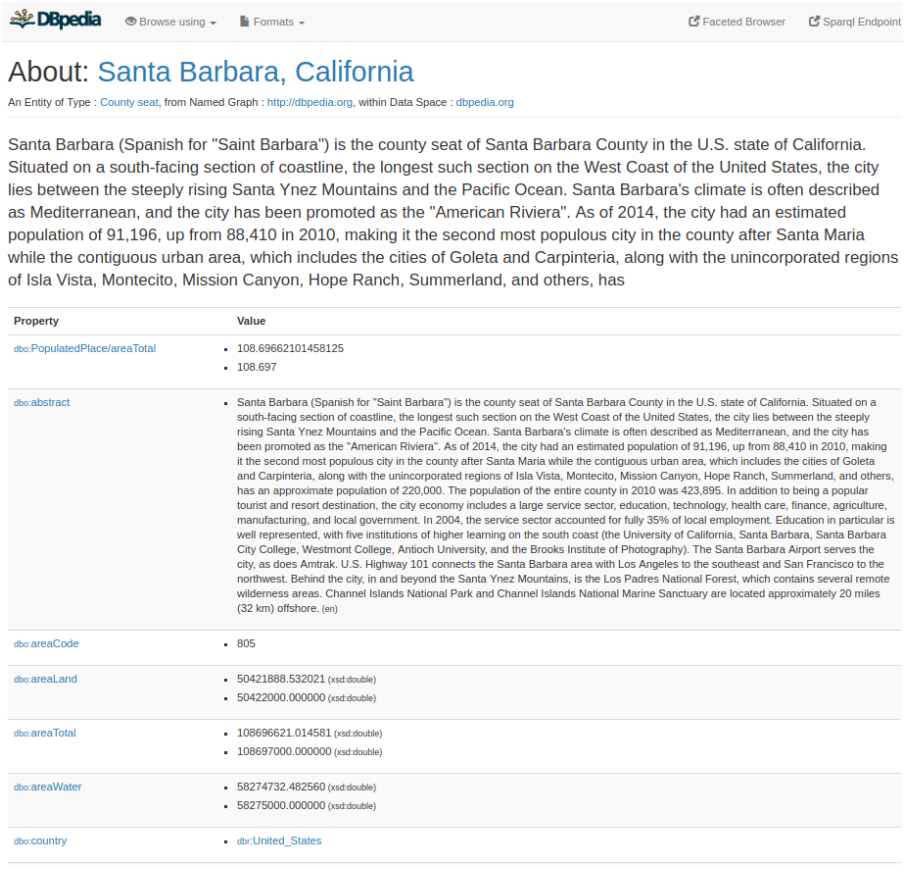
http://ex.co/geometry/polygon/12345#113.05281,-38.11945/153.30671,-11.15957

geometry type WGS84 bounding box coordinates

Encoding scheme used to mint URIs representing a geometry [7]

- Using dereferencable Uniform Resource Identifiers (URIs) to represent the geometry as a resource;
- Including important information (geometry type; bounding box) in the URIs;
- Can not get raw geometries information using SPARQL.

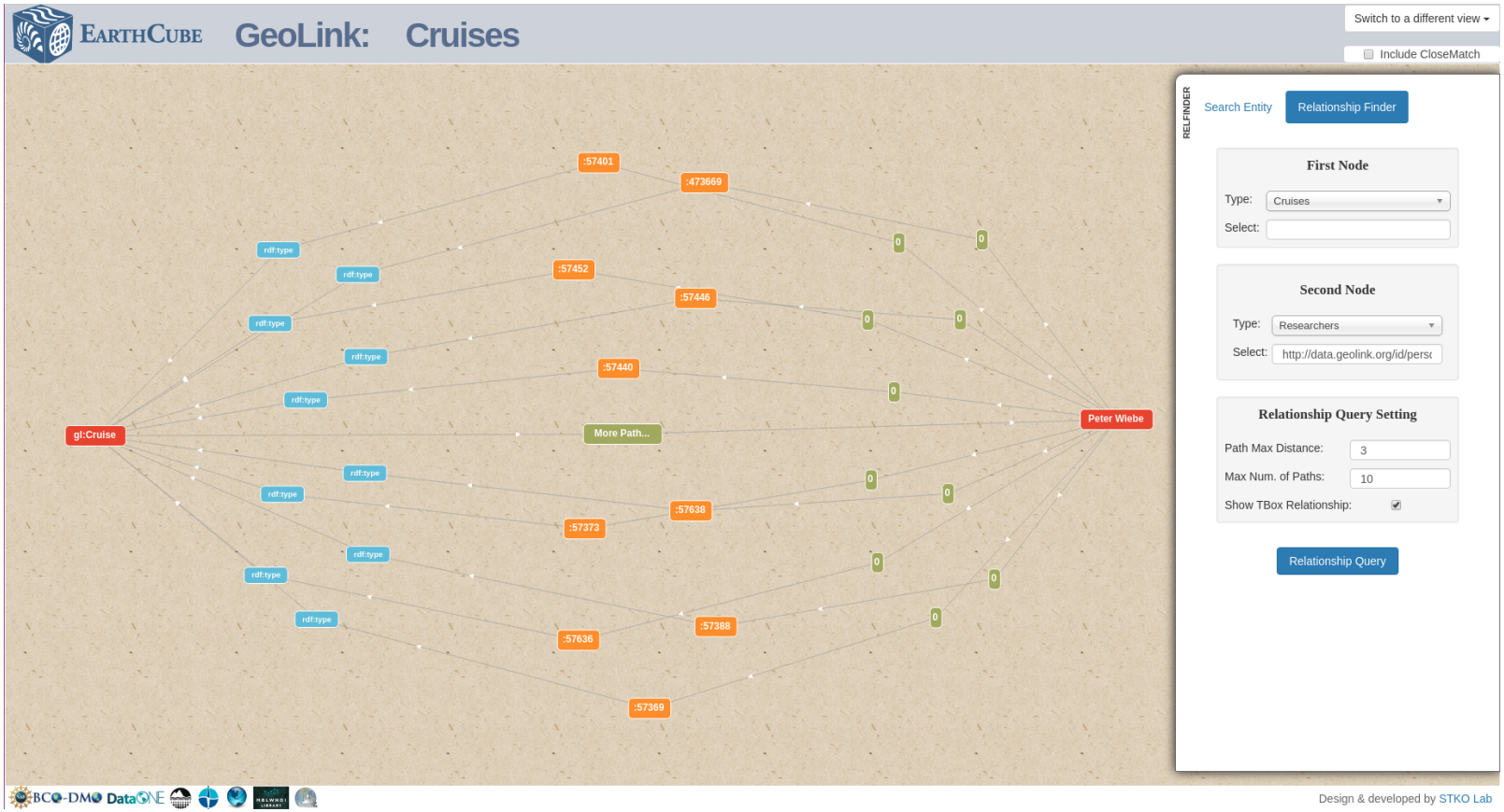
Visualization of Geographic Knowledge Graph



- Geographic Knowledge Graphs explicitly encode the semantic of geographic entities.
- **Spatial information:** can be easily visualized by maps.
- **Semantic information:** can not be easily visualized.

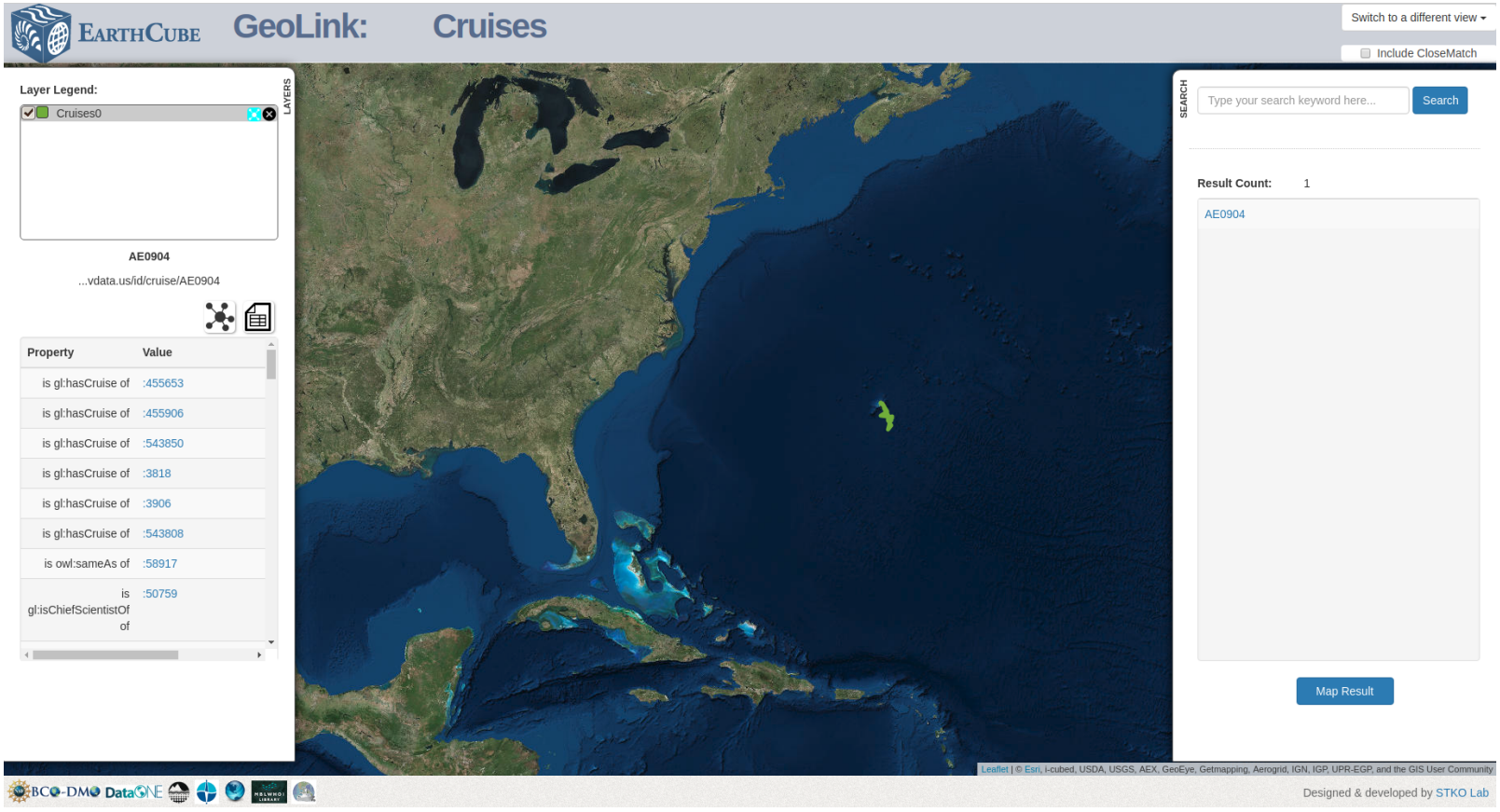
Visualize Geographic KG in Multiple Ways

- Graph View



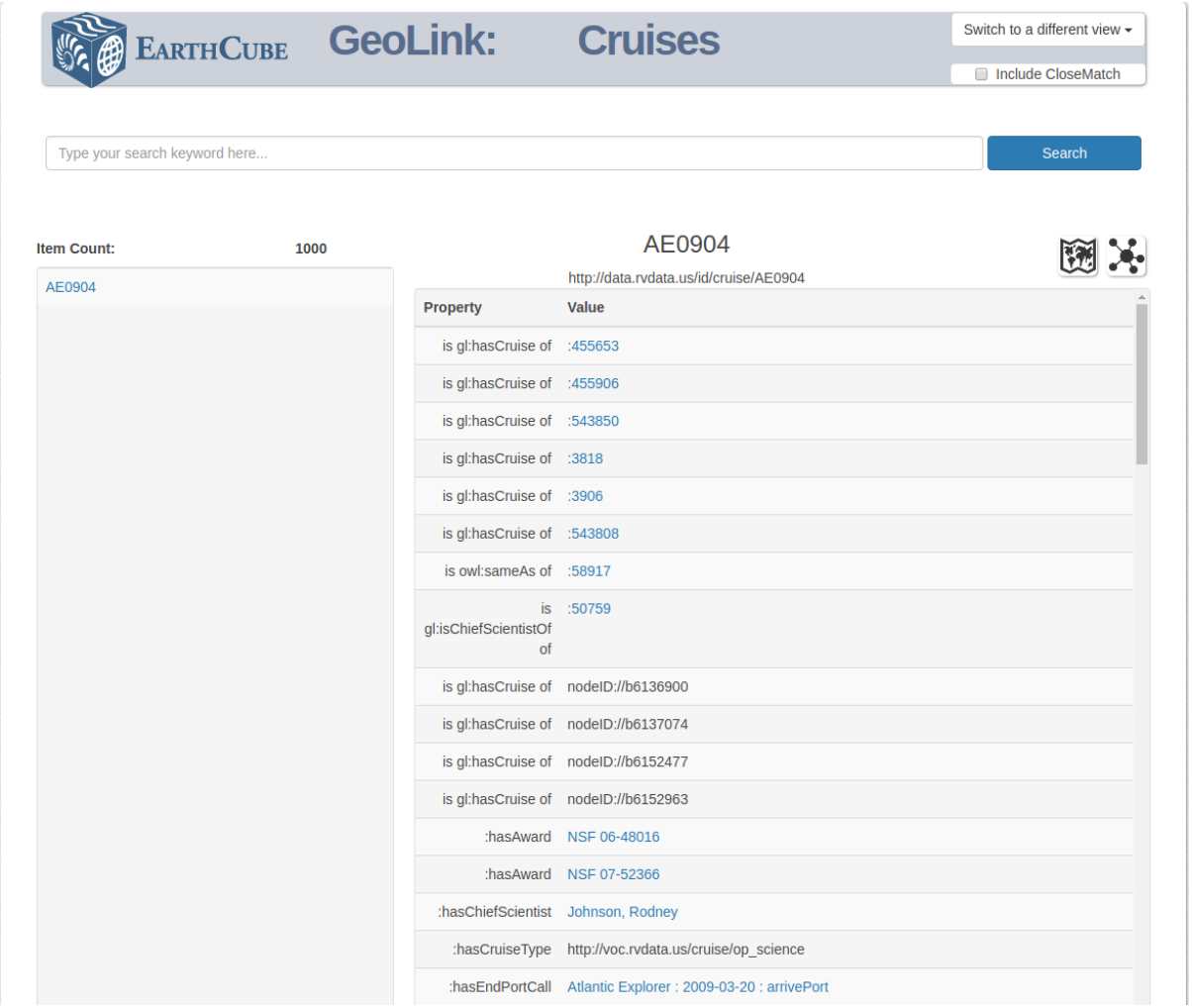
Visualize Geographic KG in Multiple Ways

- Map View



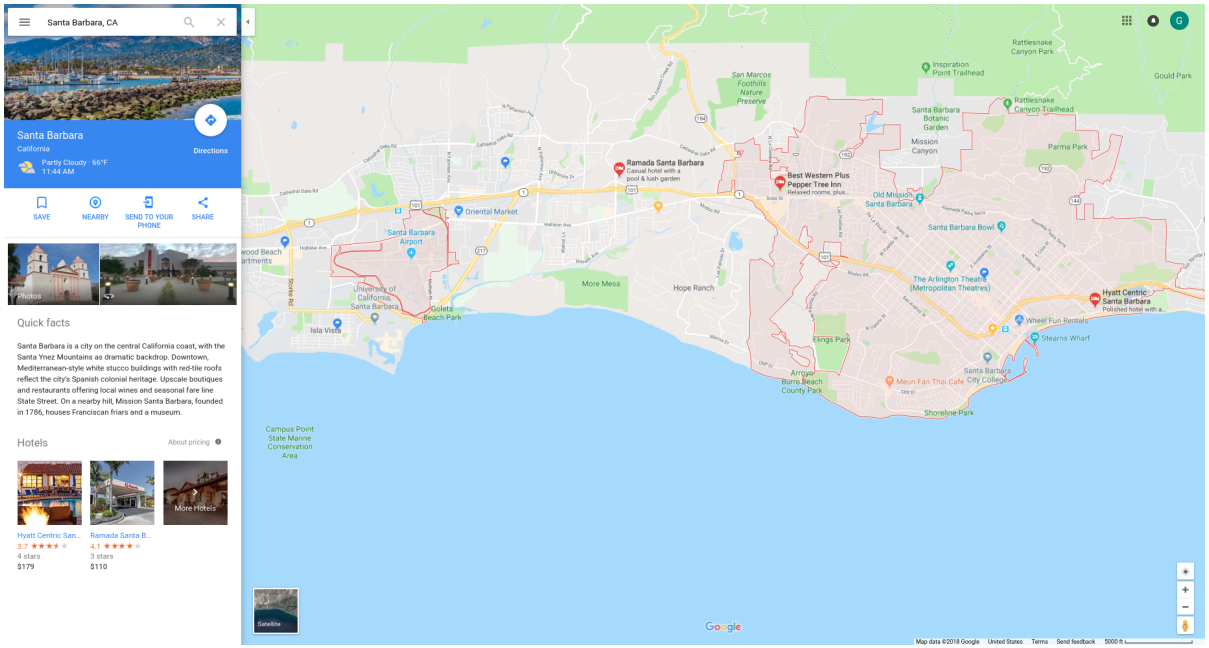
Visualize Geographic KG in Multiple Ways

- Table View



A Semantically Enriched Visualization

- **Maps**: extensively used to visualize GI and spatial relationships.



  - Difficult to directly express non-spatial relationships (semantic similarity) using such maps.

A Semantically Enriched Visualization

A Semantically Enriched Visualization: An **analogy** of thematic maps to visualize the distribution of geographic features in a semantic space.

- Points: Geographic Coordinates $\rightarrow$ Locations in the Semantic Space
- Polygons: Administration Regions/Continents $\rightarrow$ Semantic Continents

A Semantically Enriched Visualization

A Semantically Enriched Visualization:

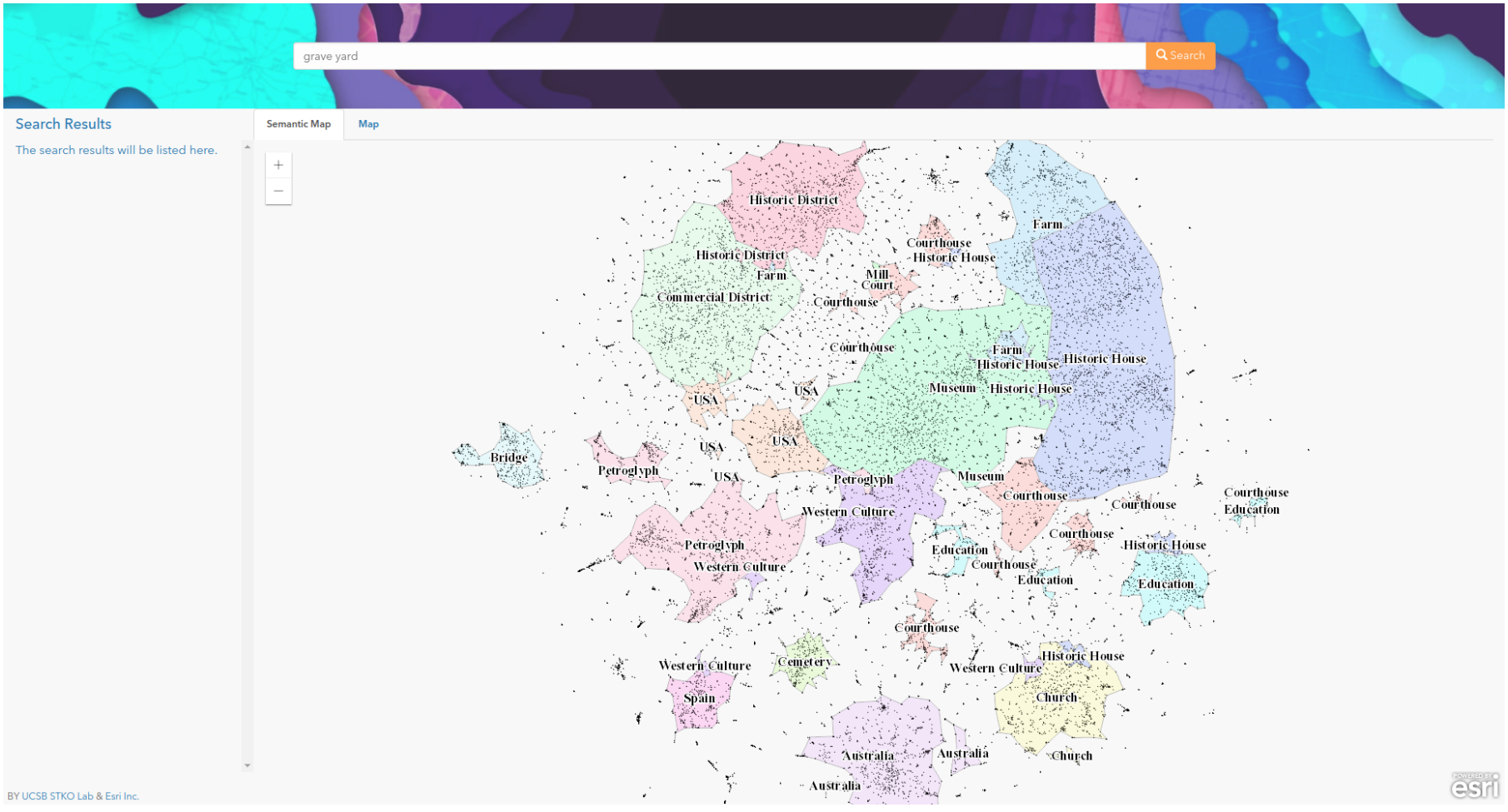
- Semantically similar entities are clusters within the same region;
- The distance between geographic features represents how similar they are.

A Semantically Enriched Visualization

- In this work, a **semantically enriched** geospatial data **visualization** and **searching** framework are presented.
- We evaluate it using a subset of places from DBpedia.

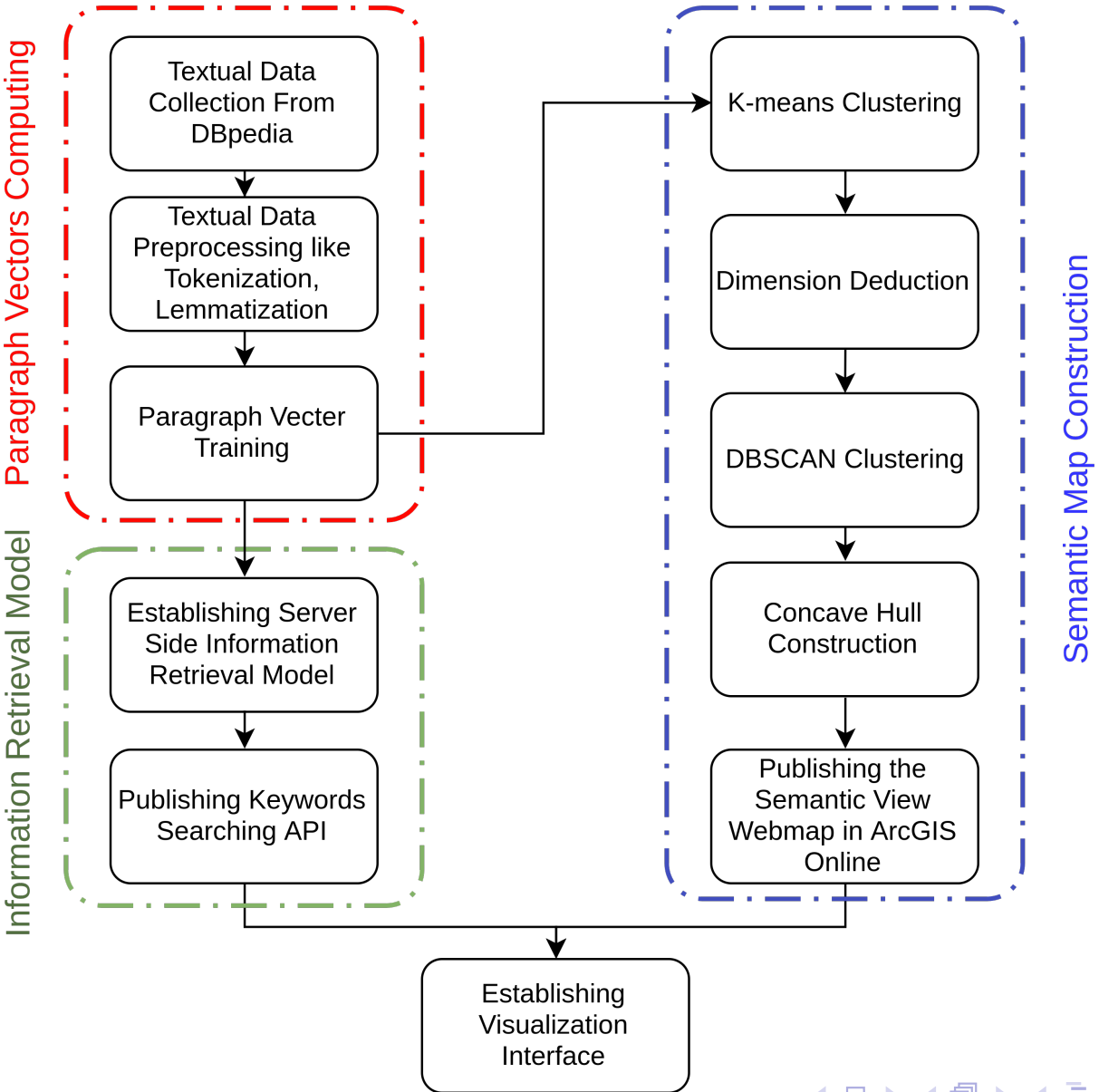
- Multiple techniques:
  - Paragraph Vector
  - Spatial Clustering
  - Concave Hull Construction
  - Information Retrieval (IR) Model

A Semantically Enriched Visualization



A semantically enriched visualization resembles cartographic layouts

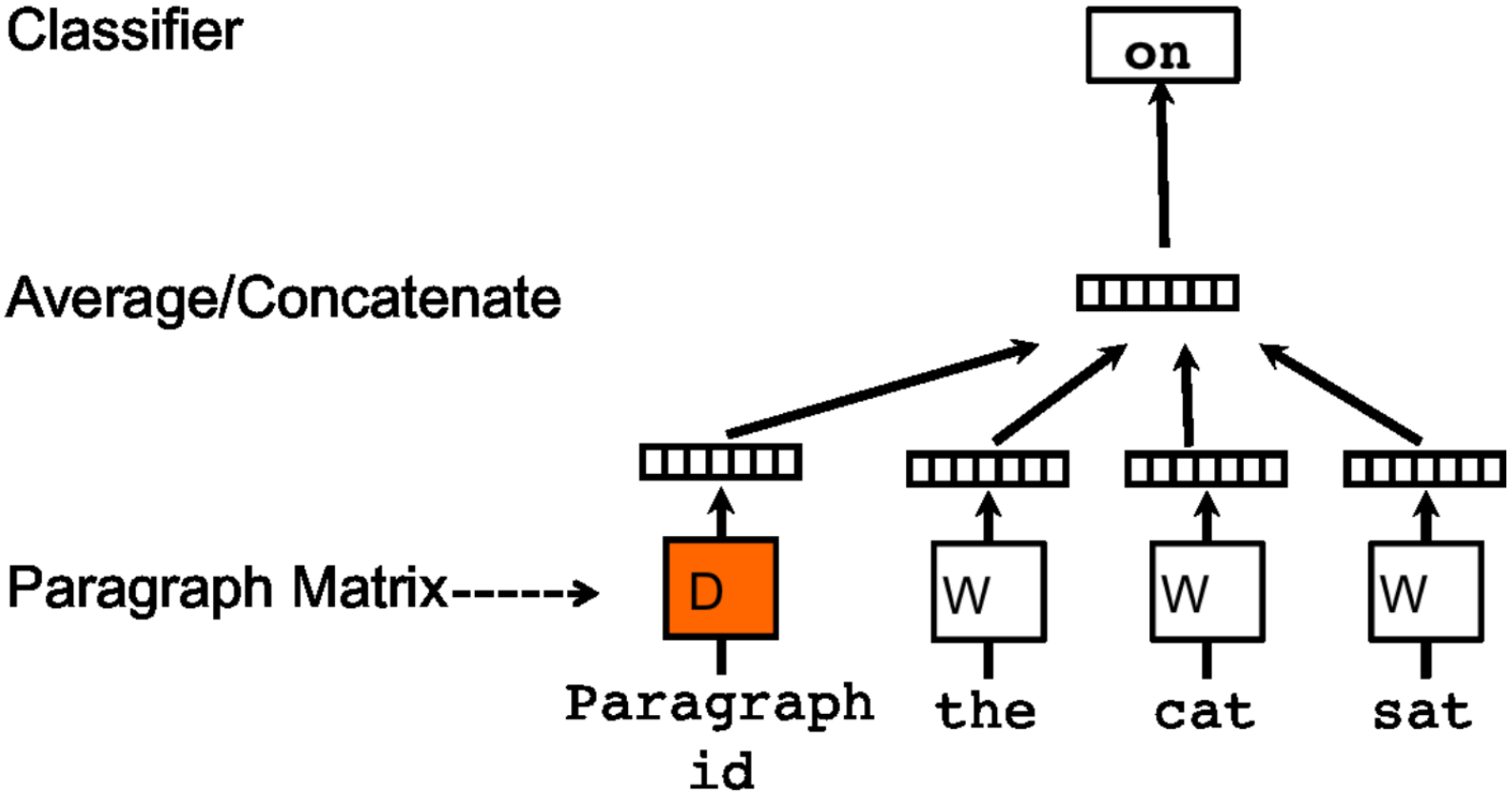
The Workflow



Paragraph Vectors Computing

- **Paragraph Vector** (or called **Doc2Vec**) is a representation learning method proposed by Natural Language Processing community.
- **Idea**: Give a collection of documents, Doc2Vec learns a high-dimensional continuous vector (embedding) for each document.
- The **cosine similarity** between the learned document vectors represents the **semantic similarity** between their corresponding documents.

Paragraph Vectors Computing



The two-layer neural network architectural of Doc2Vec

Paragraph Vectors Computing

Outputs of Doc2Vec:

- Embeddings of documents;
- Embeddings of word tokens in the document corpus.

Paragraph Vectors Computing

- **Data Source:** All entities typed dbo:HistoricalPlace in DBpedia (21010 places)
  - Each historic place has an abstract, comments, images, and geographic coordinates.
- **Method:** Doc2Vec Model (PVDM [5])
  - **Textual data collection**: Treat each place as a document whose content is its abstract and comments
  - **Textual data preprocessing**: tokenization and lemmalization
  - **Paragraph vector training**: embedding dimention $K = 300$; window size $N = 10$; learning rate $\alpha = 0.025$

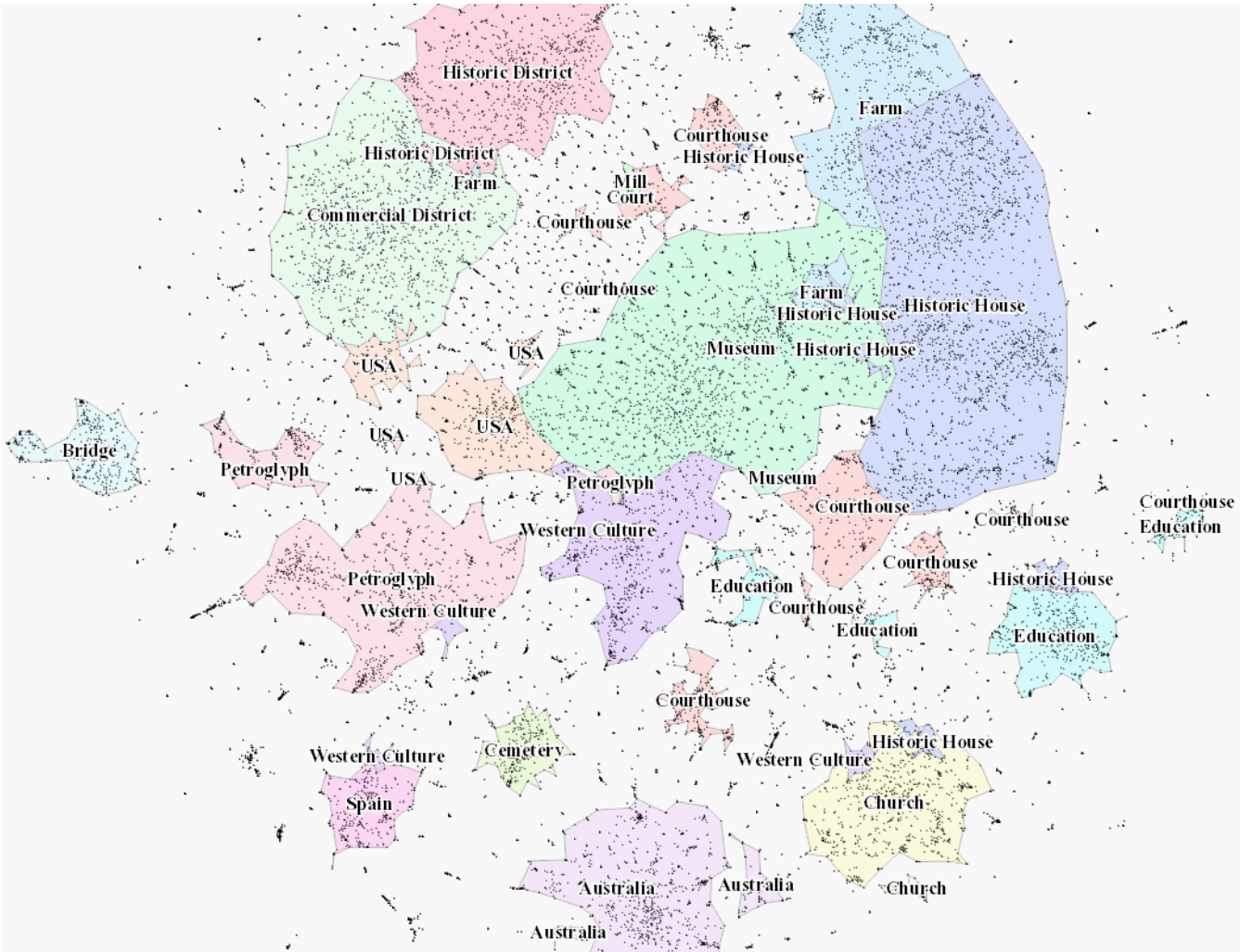
Information Retrieval Model

- **Place Embeddings**: the learned embedding of each historic place from Doc2Vec.
- **Query Embeddings**:
  - Utilize the *Doc2Vec.infer_vector()* function from gensim's Doc2Vec package
  - The TF-IDF score weighted embedding based on word embeddings of query word tokens
  - **The simple average of the query tokens' embeddings after stop words removal**
- **Semantic Similarity Score Function**: the cosine similarity between the query embedding and place embeddings
- An API [2] is provided for the semantic searching functionality among *DBpedia* historic places.

[2] http://stko-testing.geog.ucsb.edu:3050/semantic/search?searchText=grave%20yard.

Semantic Similarity Map Construction

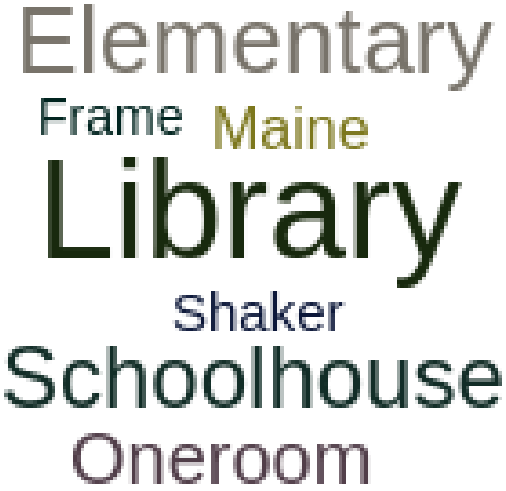
Spatialization: how to construct an overview of the semantic distribution of geographic entities such that it follows a cartographic tradition (*semantic continent*).



Semantic Similarity Map Construction

K-means clustering: group these place embeddings into different clusters;

- Try #(*clusters*) from 2 to 30 and compute silhouette coefficient [8] of the clustering results;
- #(*clusters*) = 16 gives the highest silhouette coefficient;
- The descriptions of places in each cluster are combined as one document;
- Word clouds are produced from 10 word with highest TF-IDF score;
- Each cluster is named according to the its top 10 words.



The word cloud for *Education* cluster

Semantic Similarity Map Construction

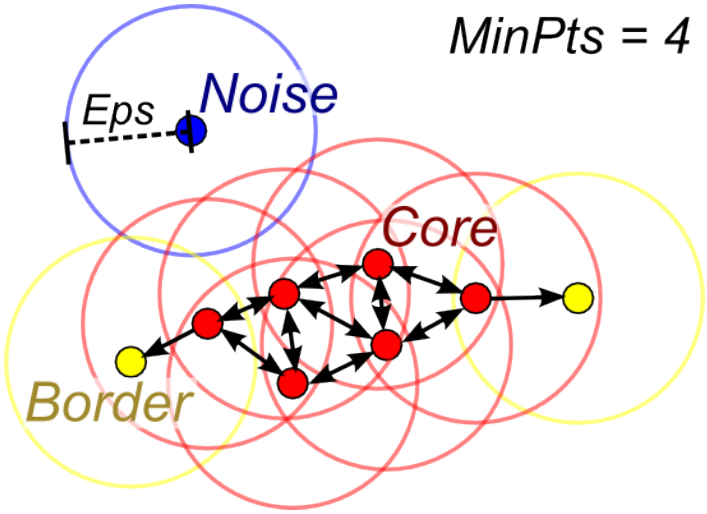
Dimension reduction: to visualize the semantic distribution of geographic entities in a 2-dimensional space

- Different dimension deduction methods including PCA and t-SNE are experimented;
- t-SNE performs best and the clusters derived from k-means are still well separated.

Semantic Similarity Map Construction

DBSCAN:

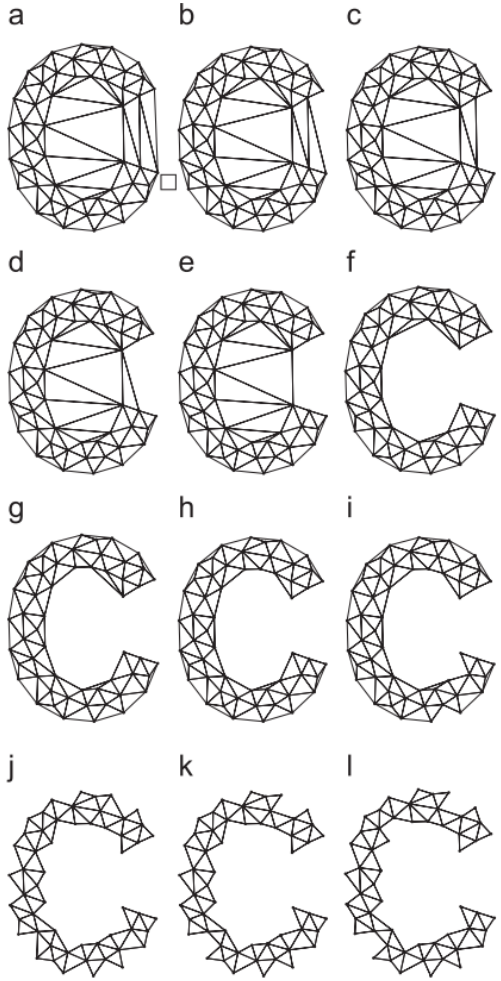
- Although t-SNE produces a good dimension reduction result, some points are far away from their cluster centroids and scattered in the 2D space.
- We apply DBSCAN [3] to each projected k-means cluster to extract the "core" parts of them.
- Visual interpretation are used to select the parameter combination for DBSCAN. ($Eps = 1.1$ and $MinPts = 6$)



Semantic Similarity Map Construction

Concave Hull Construction: *Chi-shape* algorithm [2]

- It first constructs a Delaunay triangulation;
- It erodes the boundary by deleting boundary's edges until the longest edge less than a threshold.
- A normalized length parameter $\lambda_p \in [1, 100]$ controls this threshold;
- To get optimal λ_p, a fitness score function [1] is used to balance the *complexity* and *emptyness* of the resulting concave hull.



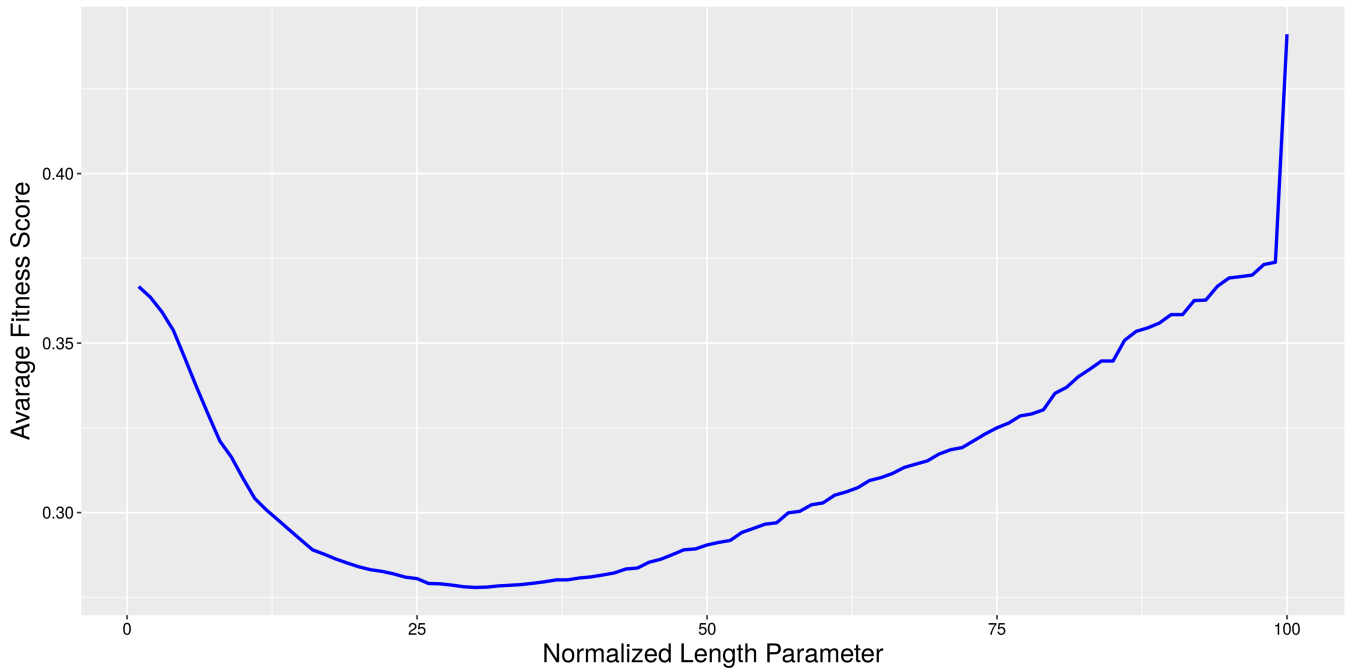
$$\phi(P, D) = Emptiness(P, D) + C * Complexity(P) \quad (1)$$

P: the derived simple polygon; D: the Delaunay triangulation of the corresponding point cluster.

Semantic Similarity Map Construction

Concave Hull Construction:

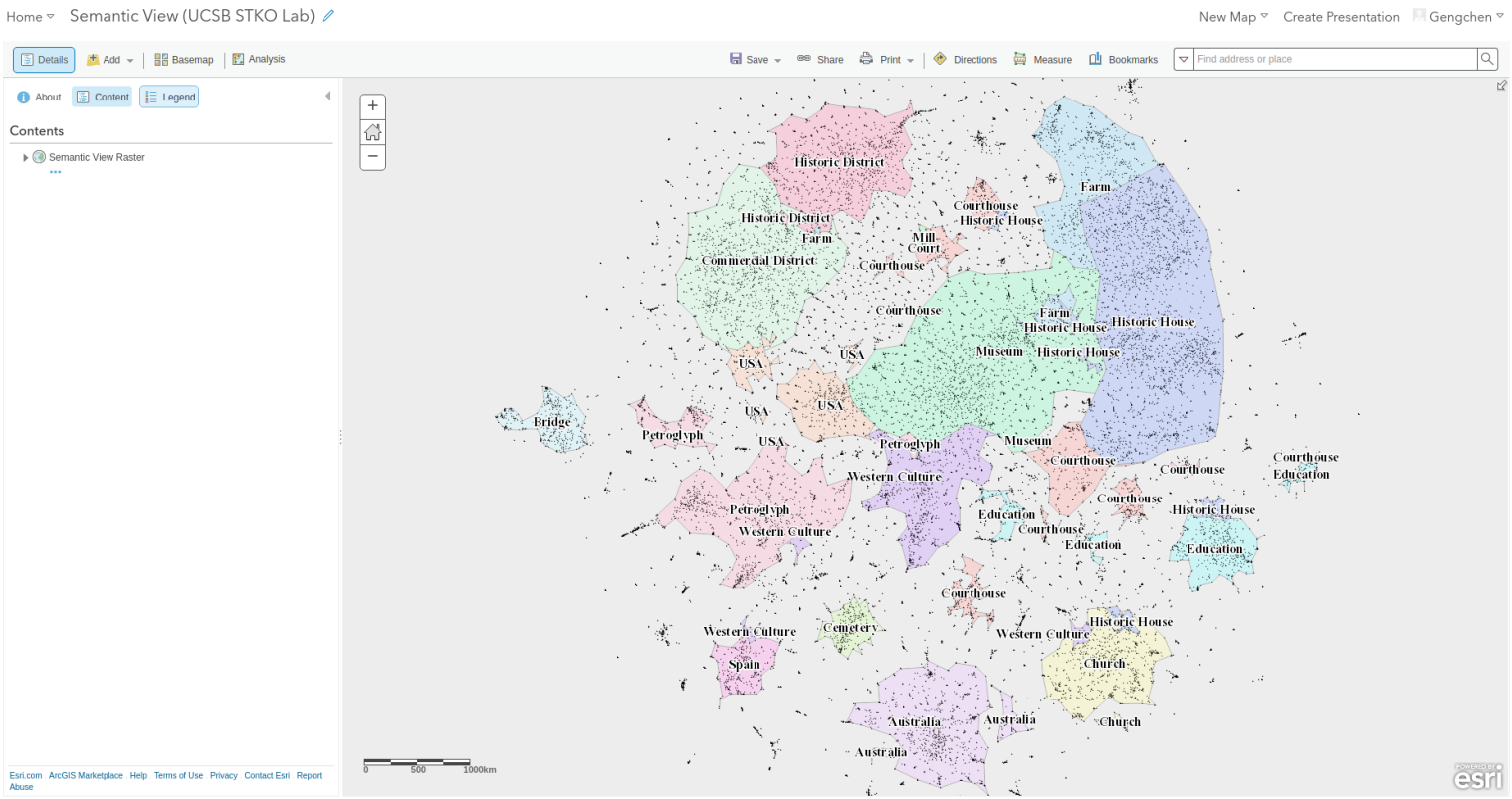
- We iterate λ_p from 1 to 100 and compute the average fitness score of all point clusters produced by DBSCAN;
- The optimal λ_p with the lowest average fitness score is 30.



The average fitness score for different λ_p among all DBSCAN clusters.

Semantic Similarity Map Construction

Publishing the Semantic View Webmap in ArcGIS Online:

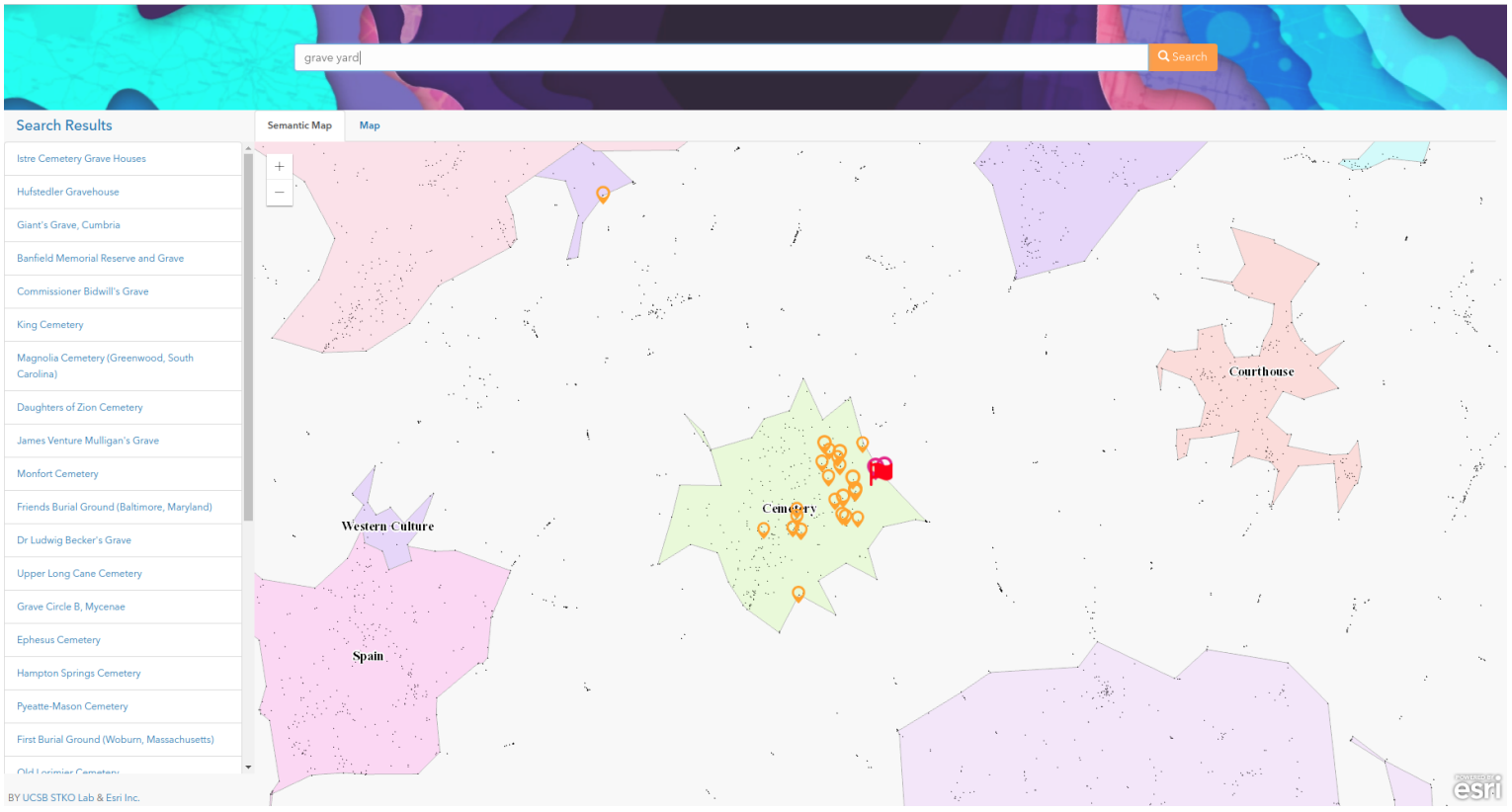


[3].

[3] http://www.arcgis.com/home/item.html?id=7e15f98399ff4788a502fd04320bdafc

Result

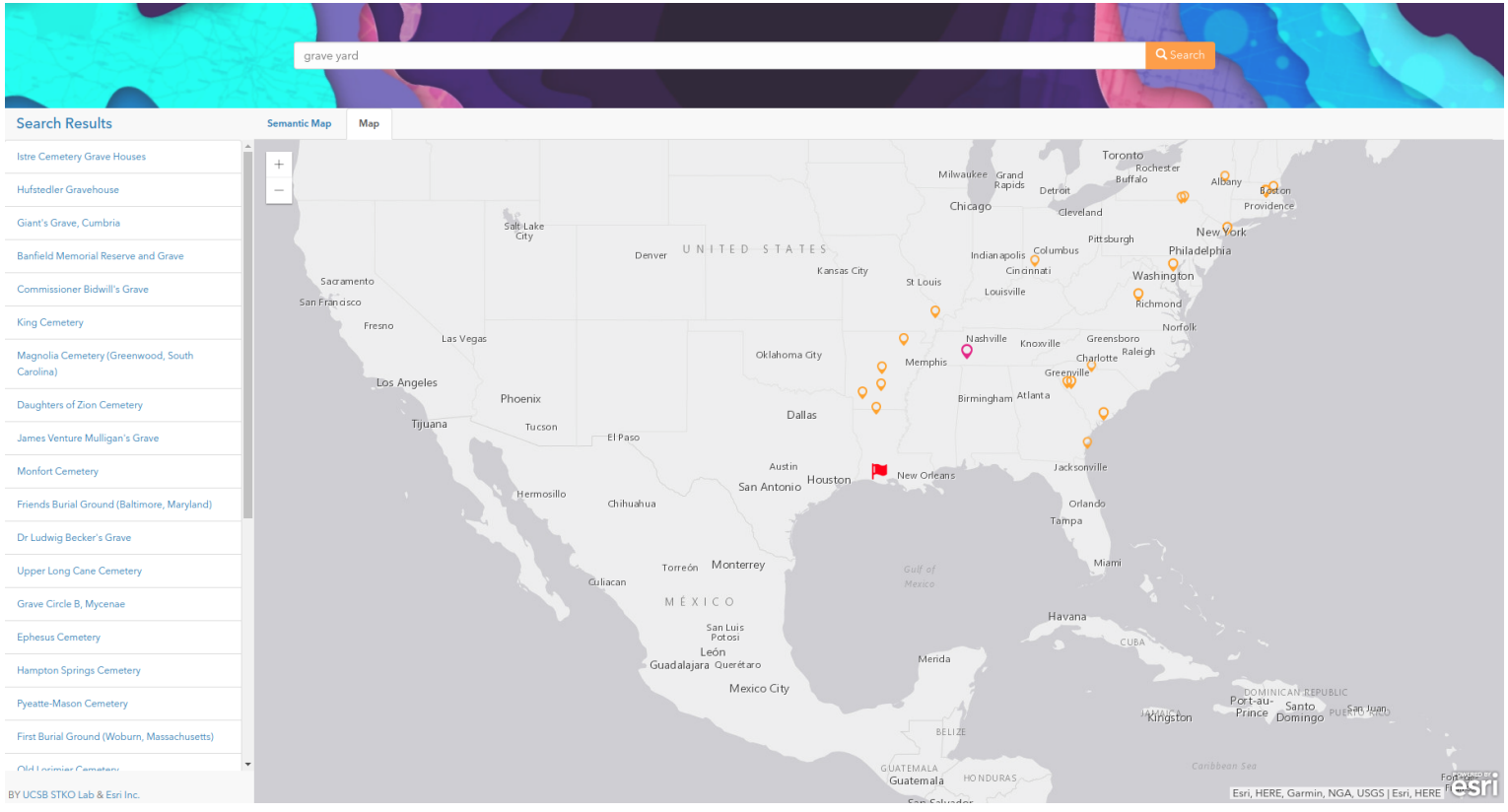
We have deployed a web-based user interface[4] to showcase the functionality using the historical places dataset.



the search result of "grave yard" in the semantic space

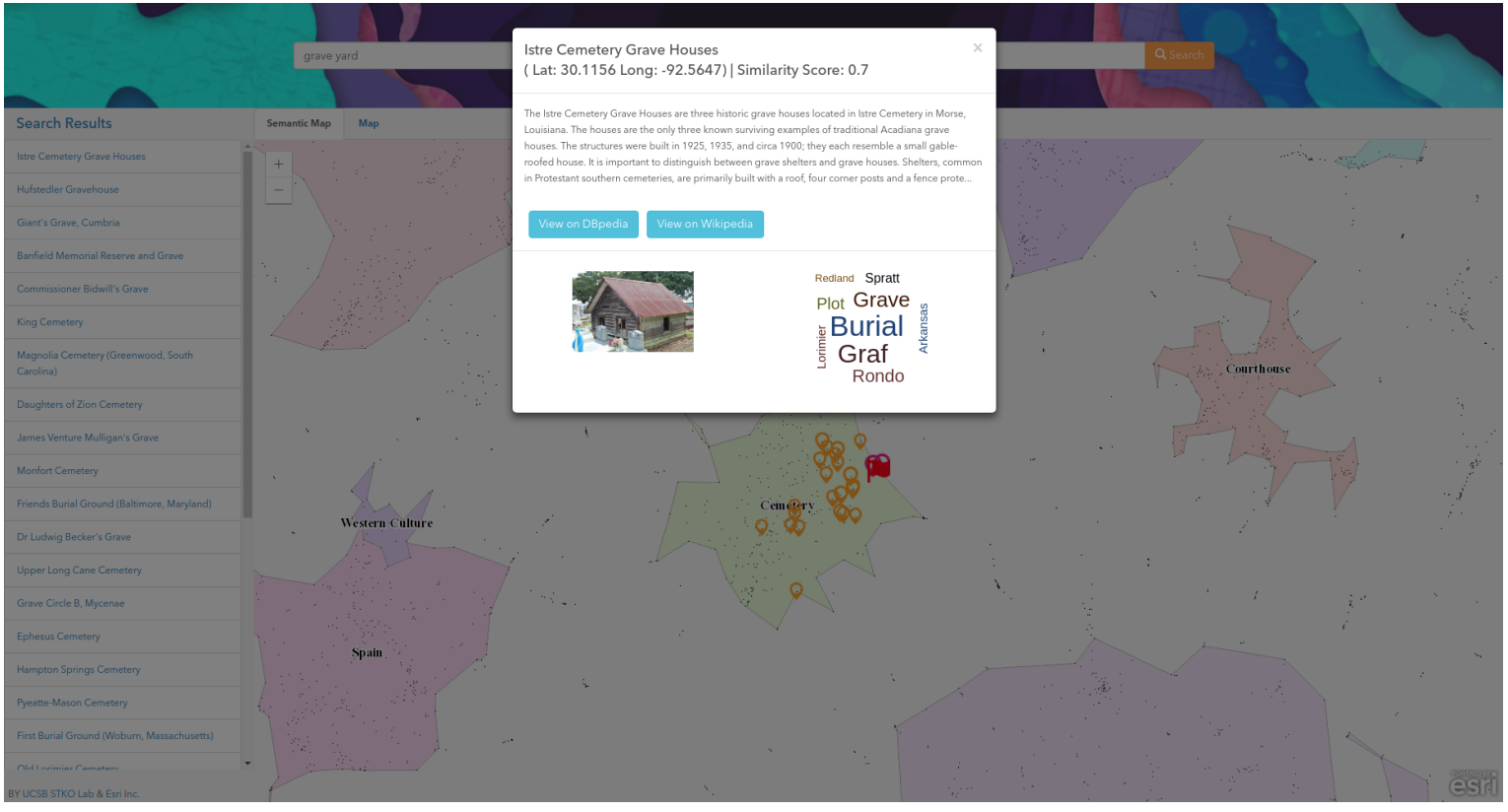
4 http://stko-testing.geog.ucsb.edu:3050/

Result



the search result of “grave yard” in the geographic space

Result



The pop-up window shows some basic information for

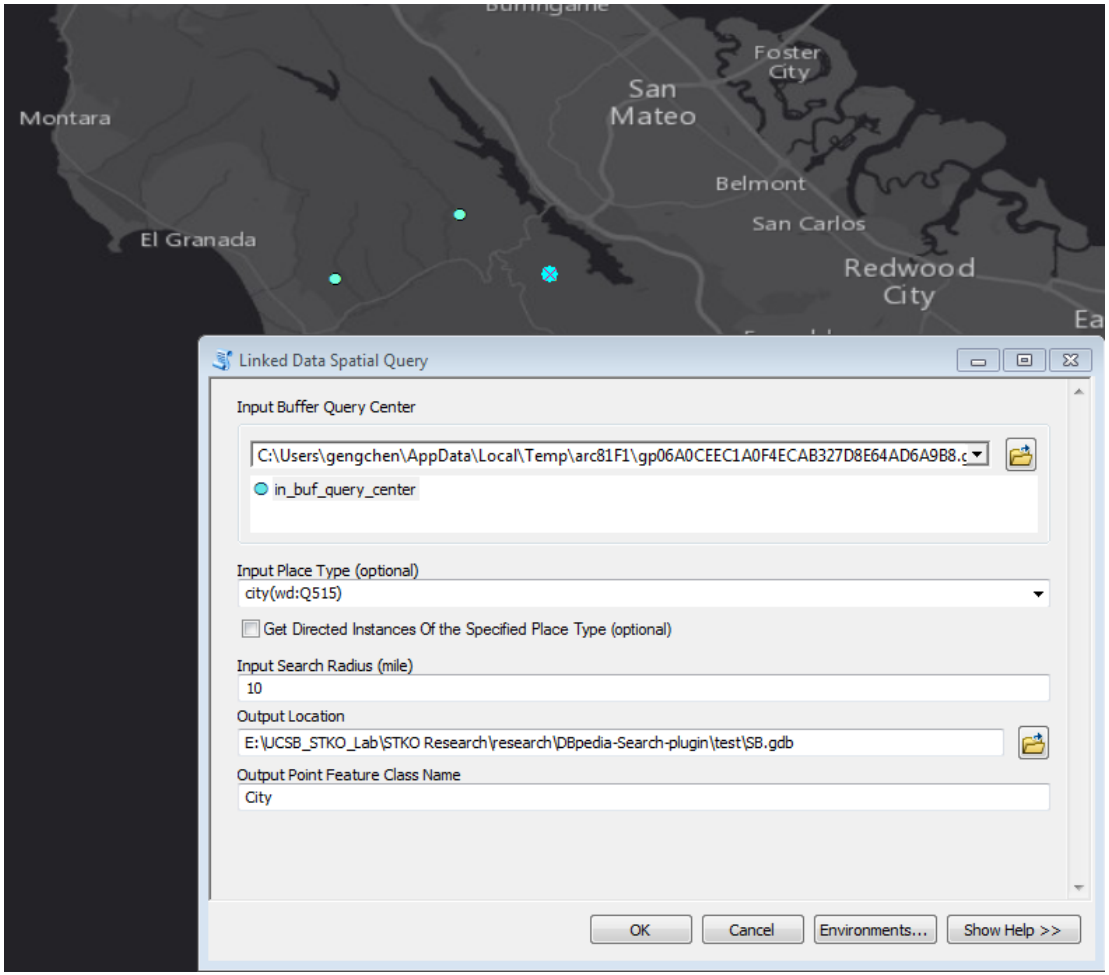
`dbo:Istre_Cemetery_Grave_Houses`.

A Deep Integration of Geographic Linked Data with GIS

- From a GIS perspective, Linked Data seems almost like a one-way street.
- Considerations when integrate Linked Data with GIS:
  - How GIS and its users should interact with Linked Data?
  - How these key benefits of Linked Data can be maintained during conversion into GIS data formats and analysis without having to flatten the data back to a tabular format?
  - How to utilize the ontologies used to semantically lift Linked Data instead of merely relying on strings?

ESRI Linked Data Connector

- Buffer Search on Geographic KG:
  - Find cities around the search center



ESRI Linked Data Connector

- Property Enrichment for Geographic Entities (Utilize Semantic Web reasoning and ontologies to extract additional properties by using subsumption reasoning and (inverse) partonomical relations as examples):
  - People that died in San Francisco / Bay Area?

Table

SanMateoCounty_wikiURL_subDivisionIRI_DBpediaIRI_is_deathPlace_Of

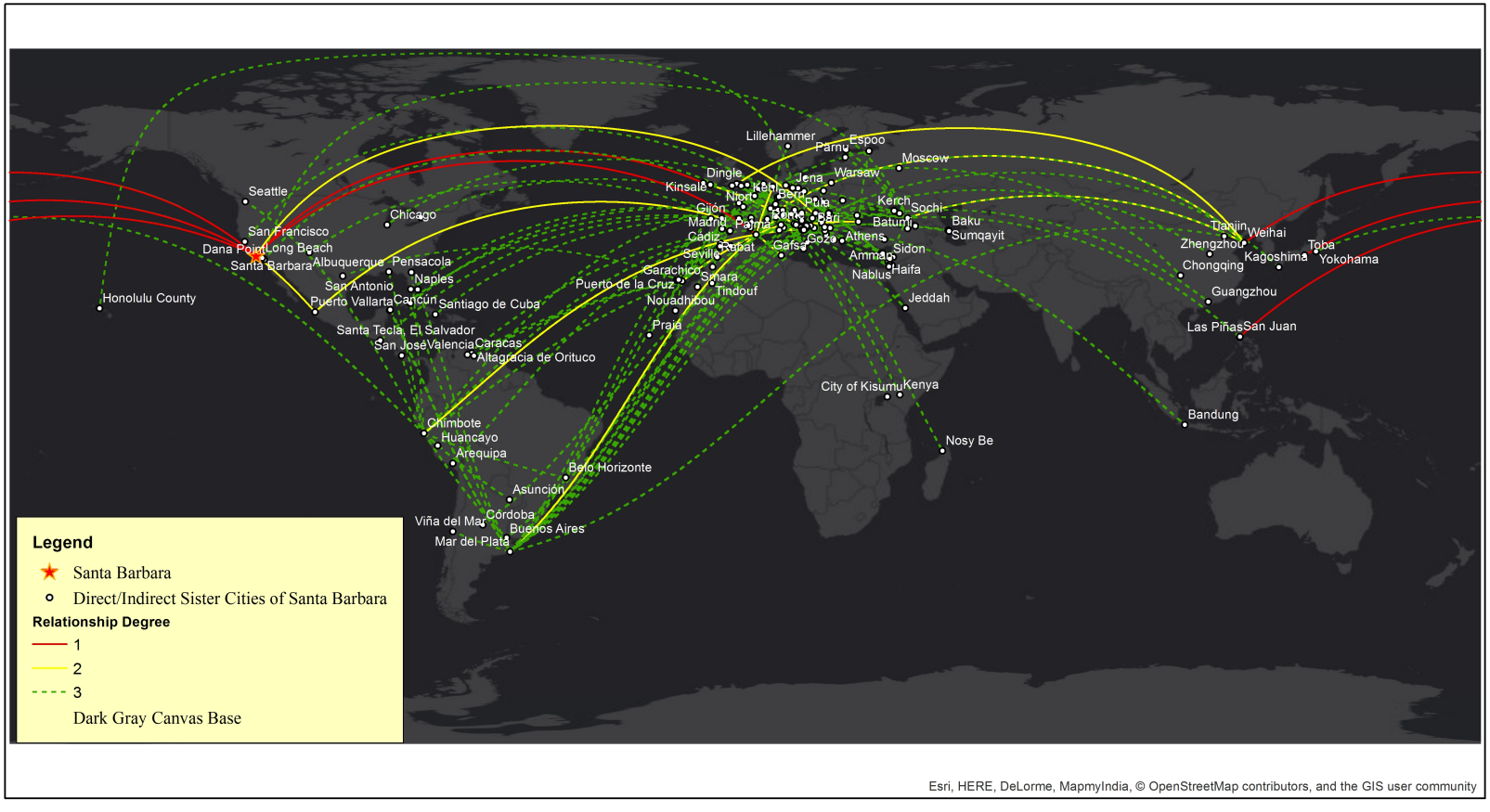
OBJECTID *	DBpediaIRI *	is_deathPlace_Of
1	http://dbpedia.org/resource/Brisbane,_California	http://dbpedia.org/resource/Lafayette_Thomas
2	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Robert_Cottle
3	http://dbpedia.org/resource/Pacifica,_California	http://dbpedia.org/resource/Christian_Theodore_Pedersen
4	http://dbpedia.org/resource/Menlo_Park,_California	http://dbpedia.org/resource/Charles_N._Felton
5	http://dbpedia.org/resource/Millbrae,_California	http://dbpedia.org/resource/Albert_Johnson_(jockey)
6	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Greta_Johansson
7	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Wallie_Herzer
8	http://dbpedia.org/resource/South_San_Francisco,_California	http://dbpedia.org/resource/John_L._Wasserman
9	http://dbpedia.org/resource/Menlo_Park,_California	http://dbpedia.org/resource/Nan_Wood_Graham
10	http://dbpedia.org/resource/La_Honda,_California	http://dbpedia.org/resource/Ben_Keith
11	http://dbpedia.org/resource/Burlingame,_California	http://dbpedia.org/resource/Michael_T._Gottlieb
12	http://dbpedia.org/resource/Burlingame,_California	http://dbpedia.org/resource/Lurline_Matson_Roth
13	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Bill_Werle
14	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Con_Dempsey
15	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Morris_Kirksey
16	http://dbpedia.org/resource/Burlingame,_California	http://dbpedia.org/resource/Dick_Jones_(baseball)
17	http://dbpedia.org/resource/Daly_City,_California	http://dbpedia.org/resource/Babe_Pinelli
18	http://dbpedia.org/resource/Pacifica,_California	http://dbpedia.org/resource/Mel_Stewart
19	http://dbpedia.org/resource/Menlo_Park,_California	http://dbpedia.org/resource/Cuckoo_Christensen
20	http://dbpedia.org/resource/Daly_City,_California	http://dbpedia.org/resource/Henry_Liu
21	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Wagner_Jorgensen
22	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Ray_Apolskis
23	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Bob_Garber
24	http://dbpedia.org/resource/Atherton,_California	http://dbpedia.org/resource/John_R._Beckett
25	http://dbpedia.org/resource/Hillsborough,_California	http://dbpedia.org/resource/Eugenio_Lopez,_Jr.
26	http://dbpedia.org/resource/San_Carlos,_California	http://dbpedia.org/resource/Omar_Ahmad_(politician)
27	http://dbpedia.org/resource/Millbrae,_California	http://dbpedia.org/resource/Ronnie_Montrose
28	http://dbpedia.org/resource/Menlo_Park,_California	http://dbpedia.org/resource/Thomas_A._Bailey
29	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Robert_R._Barry
30	http://dbpedia.org/resource/Atherton,_California	http://dbpedia.org/resource/Rajeev_Motwani
31	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Allan_Nevins
32	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Jim_Davenport
33	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Bert_Cole
34	http://dbpedia.org/resource/Portola_Valley,_California	http://dbpedia.org/resource/Albert_H._Bowker
35	http://dbpedia.org/resource/Menlo_Park,_California	http://dbpedia.org/resource/Abraham_Maslow
36	http://dbpedia.org/resource/Pacifica,_California	http://dbpedia.org/resource/Ragnar_Hasselgren
37	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Wally_Jay
38	http://dbpedia.org/resource/San_Bruno,_California	http://dbpedia.org/resource/Mosheim_Feaster
39	http://dbpedia.org/resource/San_Carlos,_California	http://dbpedia.org/resource/Tom_Tennant
40	http://dbpedia.org/resource/Atherton,_California	http://dbpedia.org/resource/Samuel_M._Shortridge
41	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Marjorie_Deanne
42	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Harry_Cheek
43	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/Ray_Medeiros
44	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Bob_Hoskins_(American_football)
45	http://dbpedia.org/resource/Burlingame,_California	http://dbpedia.org/resource/Jennison_Heaton
46	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Helen_Levitov_Sobell
47	http://dbpedia.org/resource/Redwood_City,_California	http://dbpedia.org/resource/Chang-Lin_Tien
48	http://dbpedia.org/resource/San_Mateo,_California	http://dbpedia.org/resource/David_Breeden
49	http://dbpedia.org/resource/Daly_City,_California	http://dbpedia.org/resource/Jeanne_Manford
50	http://dbpedia.org/resource/Portola_Valley,_California	http://dbpedia.org/resource/Emile_Despres
51	http://dbpedia.org/resource/Belmont,_California	http://dbpedia.org/resource/Julio_Bonetti
52	http://dbpedia.org/resource/Woodside,_California	http://dbpedia.org/resource/Shirley_Temple

1 (0 out of 174 Selected)

SanMateoCounty_wikiURL_subDivisionIRI_DBpediaIRI_is_deathPlace_Of

ESRI Linked Data Connector

- Relationship Finder between geospatial entities:
  - Which cities are 3-degree sister cities of Santa Barbara?



References I

[1] Akdag, Fatih, Eick, Christoph F, & Chen, Guoning. 2014. Creating polygon models for spatial clusters. *Pages 493–499 of: International Symposium on Methodologies for Intelligent Systems*. Springer.

[2] Duckham, Matt, Kulik, Lars, Worboys, Mike, & Galton, Antony. 2008. Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognition*, **41**(10), 3224–3236.

[3] Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, Xu, Xiaowei, *et al.* . 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Pages 226–231 of: Kdd*, vol. 96.

References II

[4] Janowicz, Krzysztof, Hu, Yingjie, McKenzie, Grant, Gao, Song, Regalia, Blake, Mai, Gengchen, Zhu, Rui, Adams, Benjamin, & Taylor, Kerry. 2016. Moon landing or safari? A study of systematic errors and their causes in geographic linked data. *Pages 275–290 of: International Conference on Geographic Information Science*. Springer.

[5] Le, Quoc, & Mikolov, Tomas. 2014. Distributed representations of sentences and documents. *Pages 1188–1196 of: International Conference on Machine Learning*.

[6] Mai, Gengchen, Janowicz, Krzysztof, Prasad, Sathya, & Yan, Bo. 2018. Visualizing The Semantic Similarity of Geographic Features.

[7] Regalia, Blake, Janowicz, Krzysztof, & McKenzie, Grant. 2017. Revisiting the Representation of and Need for Raw Geometries on the Linked Data Web. *In: LDOW@ WWW*.

References III

[8] Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, **20**, 53–65.