Knowledge Graphs and Spatiotemporal Data

OKN Vocamp, Jan. 2020

Gengchen Mai **Ling Cai**

STKO Lab, University of California Santa Barbara

Knowledge graphs

A **knowledge graph (KG)** is a data repository that stores real-world knowledge under some schema, e.g., an ontology.

- **Directed multi-graphs**
  - Nodes: entities
  - Edges: relationships between entities with relation types as labels
  - Statements: (subject–>predicate–>object)

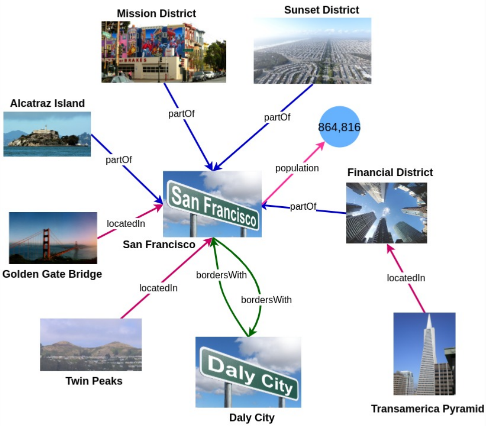


Figure 1: An Example of a KG

Knowledge Graphs

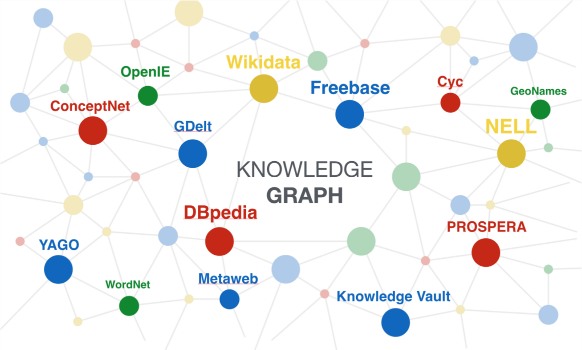


Figure 2: Figure From https://medium.com/@sderymail/challenges-of-knowledge-graph-part-1-d9ffe9e35214

- Knowledge graphs can be linked based on alignment techniques.
  - (dbr:Place, owl:**equivalentClass**, schma-org:Place)
  - (dbr:Santa_Barbara,_California, owl:**sameAs**, freebase:Santa_Barbara,_California)

Applications of Knowledge Graphs

- Cross-domain Research

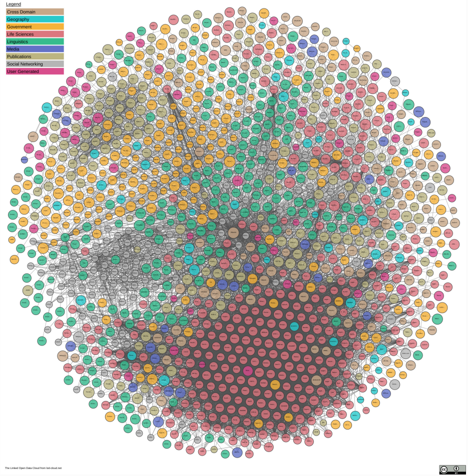


Figure 3: Linked Open Cloud

Applications of Knowledge Graphs

- Question Answering Systems, e.g., Apple Siri, Bing Search.

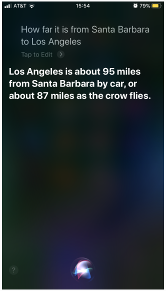


Figure 4: Siri

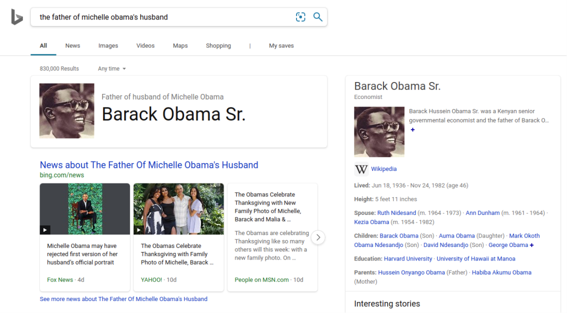


Figure 5: Bing search

Applications of Knowledge Graphs

- Information Retrieval, e.g. Google Knowledge Graph

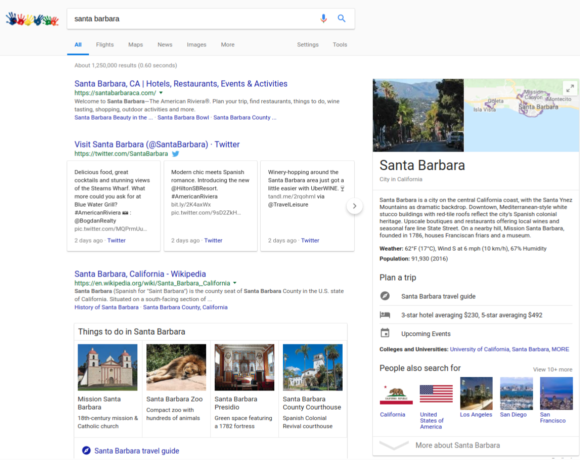


Figure 6: Google Knowledge Graph

Spatiotemporal Data in Knowledge graphs

Geographic Information

- **Geographic Information of Entities**
  - Coordinate information
    - (Santa Barbara -> coordinateLocation -> **(34°25’33"N, 119°42’51"W)**);
  - Topological relations
    - (Santa Barbara -> **partOf** -> California) ;
- **Other Geospatial-Related Statements**
  - (**France** -> memberOf -> **European Union**);
  - (**Washington, D.C.** -> hasPopulation -> 672,228);
  - (**Los Angeles** -> twinnedAdministrativeBody -> **Berlin**);

Spatiotemporal Data in Knowledge graphs

Temporal Information

- **Temporal Scope of a Statement**
  - (Poland –> memberOf –> Warsaw Pact, **[1955, 1991]**);
  - (Washington, D.C. -> hasPopulation -> 672,228, **2015**); ...
- **Time as Literals**
  - (Barack Obama –> dateOfBirth –> **4 August 1961**);
  - (Santa Barbara –> inception –> **1847**); ...
- **Transaction Time**
  - (Fernando Torres–> playFor–>Chelsea, [2011,2015), [**09/02/2017**])

Why do spatiotemporal data matter?

- **Examples**:
  - **Geographic question**: Find the cities in California which the longest river in California flowed through?
    - Find the longest river in California.
    - Spatial operations are imposed over the river and all the cities in California.
  - **Temporal query**: (?Person) (?Person –> workLocation –> New York City) $\wedge$ (?Person –> positionHeld –> President of the United States)
    - Find candidates that satisfy both statements.
    - Check the temporal scoping of the two statements.
  - ...

Knowledge Graph Embeddings

- **Basic idea:** encode entities and relations as latent low-dimensional vectors, where each dimension represents one latent feature.
  - Take TransE as an example:
    - Given a statement (Santa Barbara–>partOf–>California), |**Santa Barbara**+**partOf**-**California**|=0

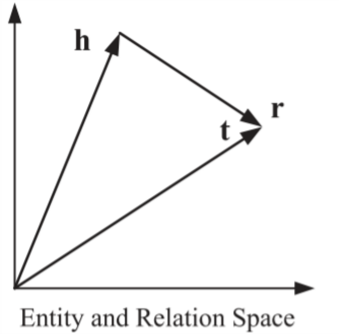
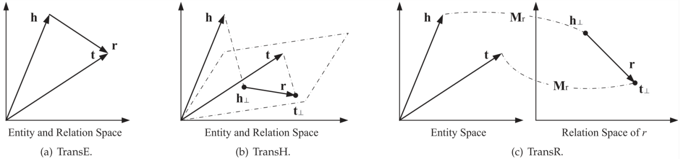


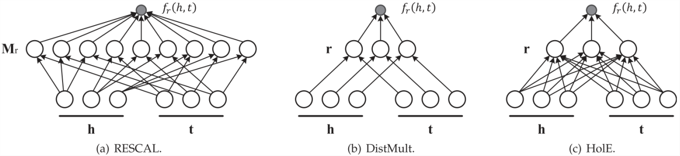
Figure 7: Knowledge Graph Embedding- TransE

Knowledge Graph Embeddings

- **Translation-based models** (e.g. TransE, TransH, and TransR)



- **Semantic matching models** (e.g. RESCAL, DisMult, and HolE)



Spatially Explicit Model

- **Spatially Explicit Model** (Goodchild et al., 2004): A model is said to be spatially explicit when it differentiates behaviors and predictions according to spatial location
- What makes a model spatially explicit? (Goodchild et al., 2001)
  - **The invariance test**: the results are **not invariant** under **relocation** of the studied phenomena
  - **The representation test**: contain **spatial representations** of the studied phenomena in their implementations (e.g., coordinates, spatial relations, place names, and so on)
  - **The formulation test**: use **spatial concepts** in their formulations, e.g. the notion of a neighborhood
  - **The outcome test**: the spatial structures/forms of inputs and outcomes of the model differ

Spatially Explicit Machine Learning Model

- **Spatially Explicit Machine Learning Model**: Improve the performance of current state-of-the-art machine learning models by using **spatial thinking and principles** such as:
  - **spatial variability**
  - **distance decay effect**
  - **map projection**
- Examples:
  - Geographic Question Answering
  - Geographic Knowledge Graph Summarization
  - Location Encoding

Geographic Question Answering

- Due to **missing information** and **logical inconsistency**, it is likely to receive **no answer** for questions given a knowledge graph.
- This challenge is commonly handled by **query relaxation/rewriting** based on **knowledge graph embedding**.
- Examples:
  - What is the weather like in **Montecito**? (**missing information**)
    - After **rewriting**: What is the weather like in **Santa Barbara**?
  - Which city spans Texas and Colorado? (**logical inconsistency**)
    - After **relaxation**: Which city locates in Texas?
- The relaxation of geo-queries should consider **spatial proximity** and **place hierachy**.

Query Relaxation Based on Knowledge Graph Embeddings

- What is the American drama films directed by Tim Burton, one of whose star actors was born in New York?

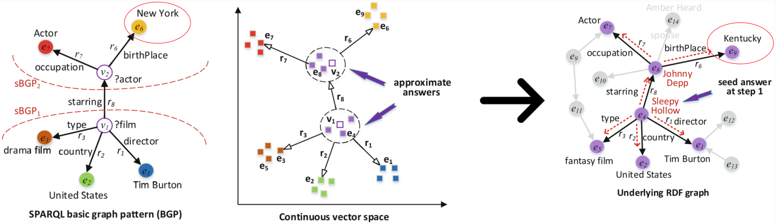


Figure 8: M. Wang et al., 2018

Workflow

Question

↓

Query

↓

Query relaxation / rewriting

↓

Relaxed query

↓

Relaxed Question + Answer

Q: In which computer hardware company located in Cupertino is/was Steve Jobs a board member?

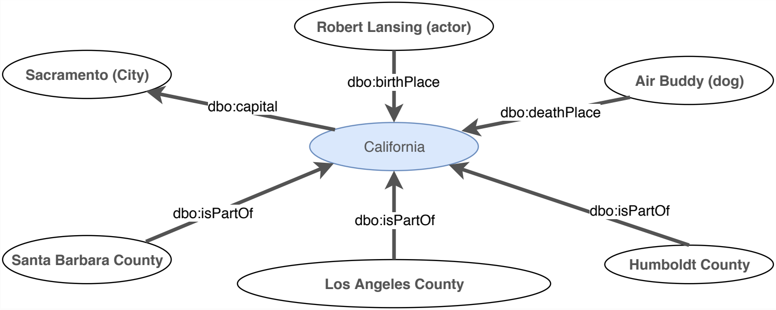
```
SELECT ?v
WHERE {
?v dbo:locationCity dbr:Cupertino,_California .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs dbo:board ?v .
}
```


```
SELECT ?v
WHERE {
?v dbo:locationCity dbr:California .
?v dbo:industry dbr:Computer_hardware .
dbr:Steve_Jobs dbo:board ?v .
}
```

Q: In which computer hardware company located in California is/was Steve Jobs a board member?
A: Apple Inc.

Spatially Explicit Knowledge Graph Embedding

- **TransGeo**: to assign **larger weights** to geographical triples in an entity context, and these weights are modeled using a **distance decay function**



Evaluation

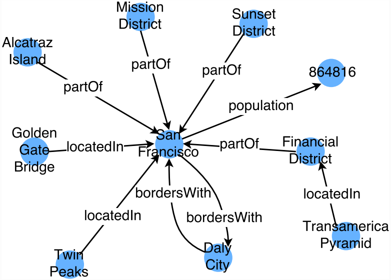
- **Link prediction**: Given h, r, to predict the correct t
- **Answer prediction by relaxation/rewriting**: The rank of the correct answer in the queried answer ranking list

Table 1: Two evaluation tasks for different KG embedding models

	Link Prediction				Query Relaxation	
	MRR		HIT@10		MRR	HIT@10
	Raw	Filter	Raw	Filter		
TransE Model	**0.122**	0.149	30.00%	34.00%	0.008	5% (1 out of 20)
Wang et al. (2018)	0.113	0.154	27.20%	30.50%	0.000	0% (0 out of 20)
$TransGeo_{regular}$	0.094	0.129	28.50%	33.40%	0.098	25% (5 out of 20)
$TransGeo_{unweighted}$	0.108	0.152	30.80%	37.80%	0.043	15% (3 out of 20)
TransGeo	0.104	**0.159**	**32.40%**	**42.10%**	**0.109**	**30% (6 out of 20)**

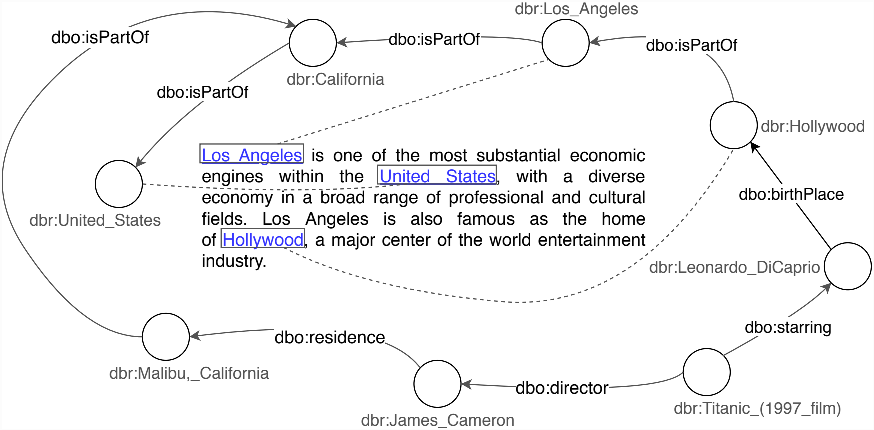
Geo Knowledge Graph Summarization

- Summarization
  - Identify the underlying structure and meaning of the original Geographic KG using a digest graph

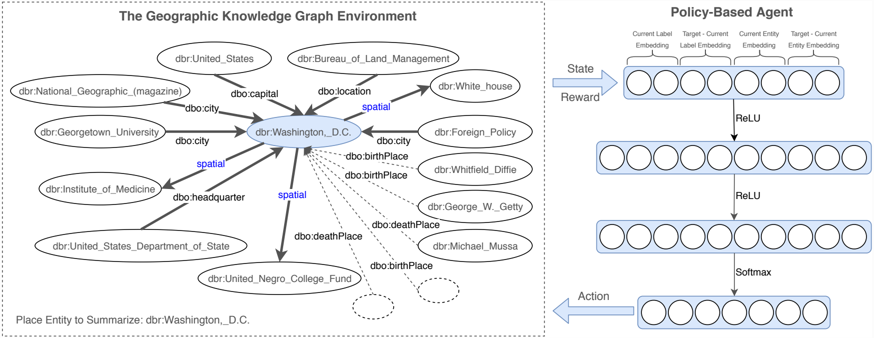


- **Question**: How can we leverage both **top-down** knowledge (e.g., considering **spatial component explicitly**) and **bottom-up** approaches (e.g., **machine learning**) to help summarize geo KGs by taking into account the balance between **commonality** and **variability**?

Summarization Example



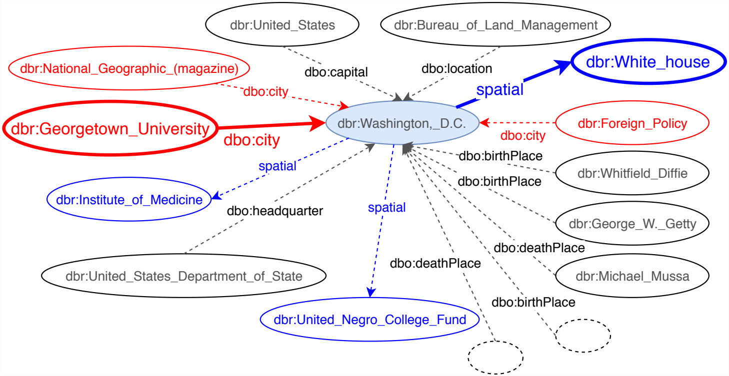
Reinforcement Learning Framework



- The process starts with only **one node**
- The **agent** analyzes the **original graph structure** and the **Wikipedia summary**
- The agent iteratively adds **new relations and nodes** to the graph until the graph conveys information comparable to the Wikipedia summary

Markov Decision Process

- Actions
  - 534 relations + 1 special **spatial relation**



Result

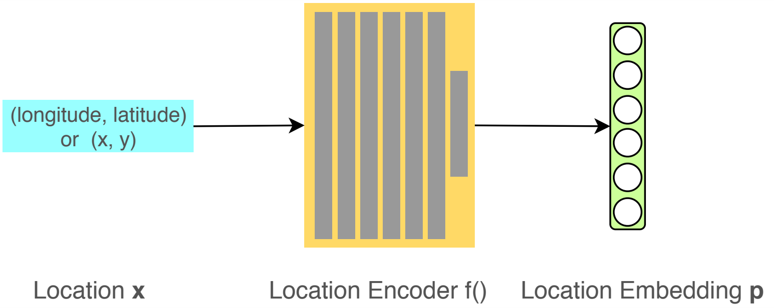
- RL-based models improve the cosine similarity (the summary graph is comparable to the Wikipedia abstract)

	RL (nonspatial-normal)	RL (spatial-normal)	RL (nonspatial-maxmin)	RL (spatial-maxmin)	RL (spatial-maxmin-pr)
Entity Embedding	0.0307	0.0496	0.0523	0.0732	**0.0760**
Word Embdding	0.1659	0.2527	0.2444	0.3025	**0.3159**

- **The spatially explicit model can perform twice as good as non-spatial models**

Location Encoding

- More direct approach?
- A **general-purpose representation model for space** is particular useful to design **spatially explicit models** for multiple tasks
- Advantage:
  - Preserve **spatial proximity** and **directions**
  - Easy to **generalize to unseen locations**
  - Avoid **explicit pairwise distance computation** which is unnecessarily expensive



Grid Cell

- Nobel winning Neuroscience research shows that **grid cells** in mammals provide a **multi-scale periodic representation** that functions as a metric for **coding space**.
- Grid cells are critical for integrating self-motion (path integration, or so-called dead-reckoning).

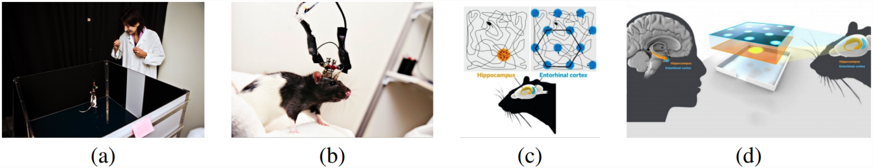


Figure 10: Figure from R. Gao et al., (2019)

Grid Cell

- Blair et al. (2007) show that the **multi-scale periodic representation of grid cells** can be simulated by summing three **cosine grating functions** oriented 60° apart.

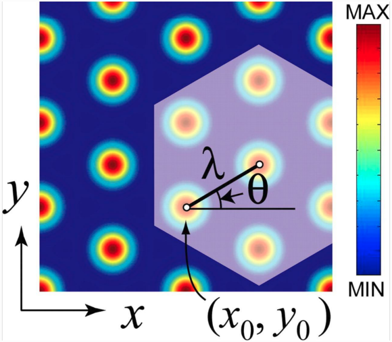


Figure 11: Figure from Blair et al. (2007)

- Encode locations with **multi-scale periodic representations** by using **3 sinusoidal functions**.

Applications

- KG related tasks:
  - Geographic Question Answering
  - Geographic Knowledge Graph Summarization
- Other tasks:
  - Air Pollution Forecasting
  - **Location-Aware Image Classification**

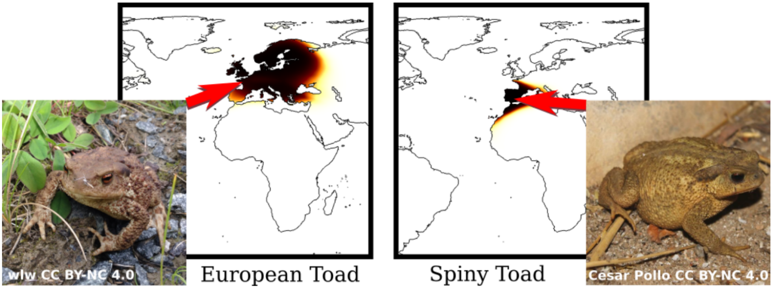


Figure 12: Figure from Mac Aodha et al. (2019)

Location-Aware Image Classification

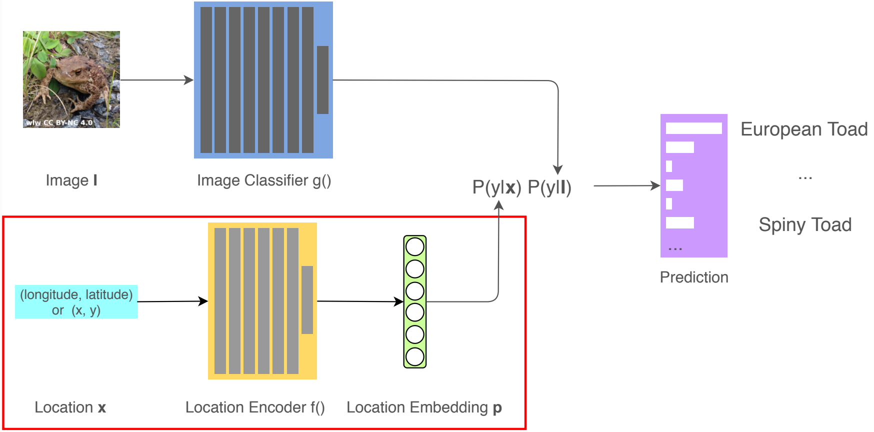


Figure 13: Location-Aware Image Classification

Evaluation

	BirdSnap†	NABirds†
No Prior (i.e. uniform)	70.07	76.08
Nearest Neighbor (num)	77.76	79.99
Nearest Neighbor (spatial)	77.98	80.79
Adaptive Kernel (Berg et al., 2014)	78.65	81.11
tile (Tang et al., 2015) (location only)	77.19	79.58
wrap (Mac Aodha et al., 2019) (location only)	78.65	81.15
grid (λ_{min}=0.0001, λ_{max}=360, S = 64)	**79.44**	81.28
theory (λ_{min}=0.0001, λ_{max}=360, S = 64)	79.35	**81.59**

Figure 14: Evaluation Result for Location Aware Image Classification

Summary

- Knowledge graphs play important roles in **data storage**, **data sharing**, **data synthesis**, **semantic search**, **cross-domain studies**, etc.
- **Spatiotemporal data** are abundant within and beyond knowledge graphs.
- **Spatially explicit models** are needed for the advancement of spatial data science.

References

- Gengchen Mai, Bo Yan, Krzysztof Janowicz, Rui Zhu. *Relaxing Unanswerable Geographic Questions Using A Spatially Explicit Knowledge Graph Embedding Model*, In: Proceedings of AGILE 2019, June 17 - 20, 2019, Limassol, Cyprus. * **1st Best Full Paper Award**
- Bo Yan, Krzysztof Janowicz, Gengchen Mai, Rui Zhu. *A Spatially-Explicit Reinforcement Learning Model for Geographic Knowledge Graph Summarization*. Transactions in GIS, 23(2019), 620-640. DOI:10.1111/tgis.12547
- Gengchen Mai, Krzysztof Janowicz, Bo Yan, Rui Zhu, Ling Cai, Ni Lao. *Multi-Scale Representation Learning for Spatial Feature Distributions using Grid Cells*, In: Proceedings of ICLR 2020, Apr. 26 - 30, 2020, Addis Ababa, ETHIOPIA . * **Spotlight Paper**