Understanding the Semantics of Places in Gazetteers via Spatial Analysis

Rui Zhu

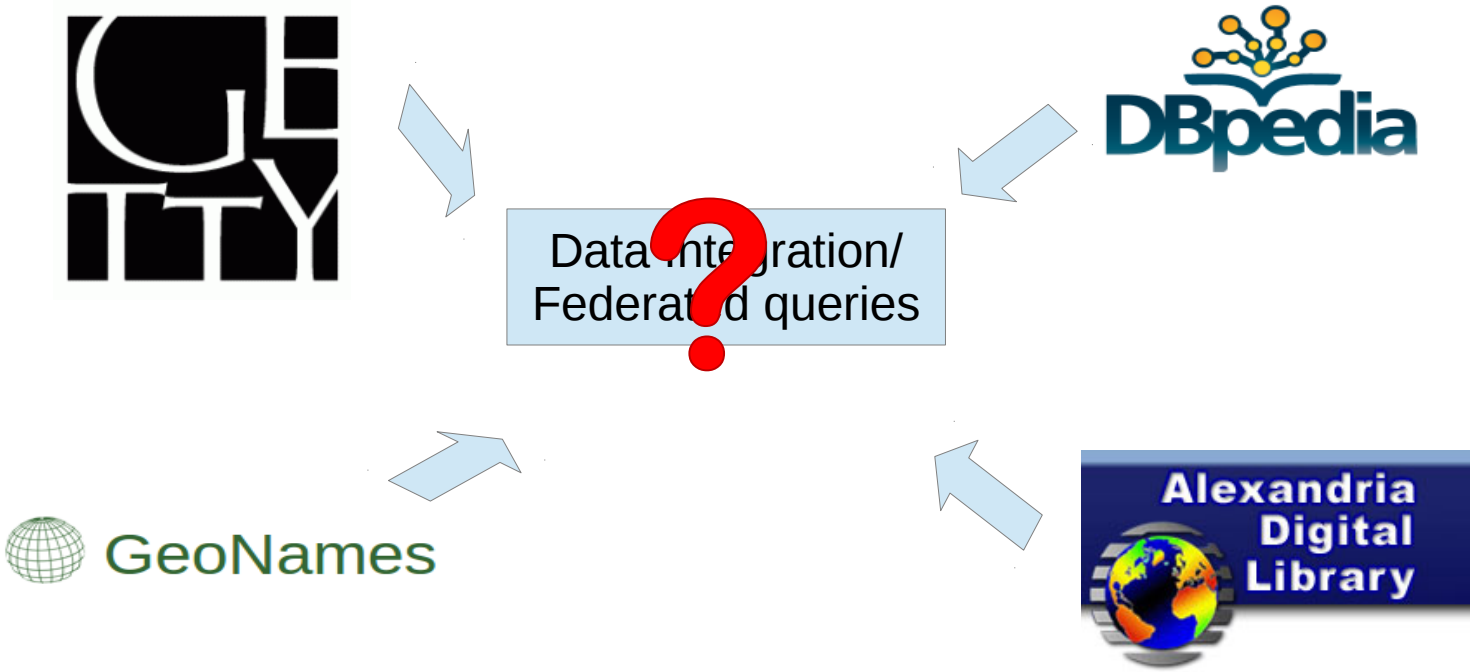
STKO Lab
Department of Geography
University of California, Santa Barbara

Outline

- Motivations
- Data sets
- Methods
- Results & Discussions
- Future work

Motivations

- A wide variety of gazetteers

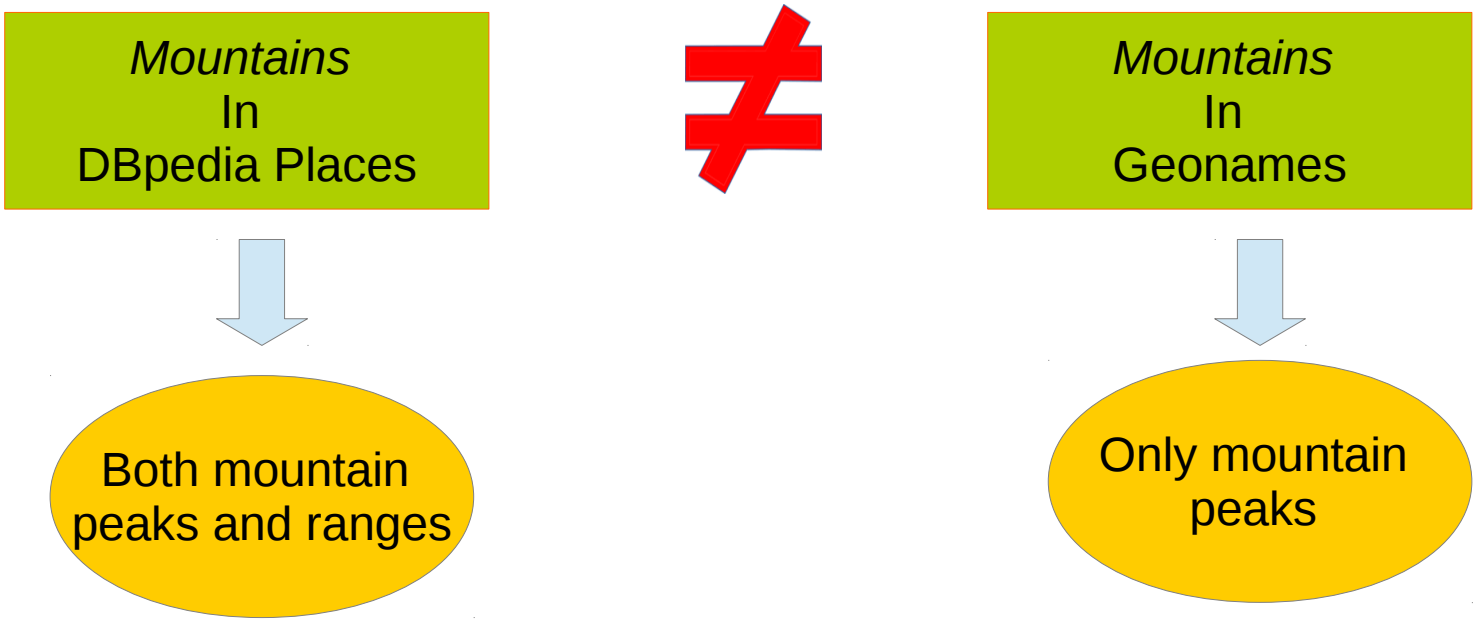


Motivations (cont.)

- Traditional techniques for integration/ alignment:
  - Expert guess
  - String similarity measures; e.g. Levenshtein distance
  - Network similarity measures. e.g. Structure equivalence

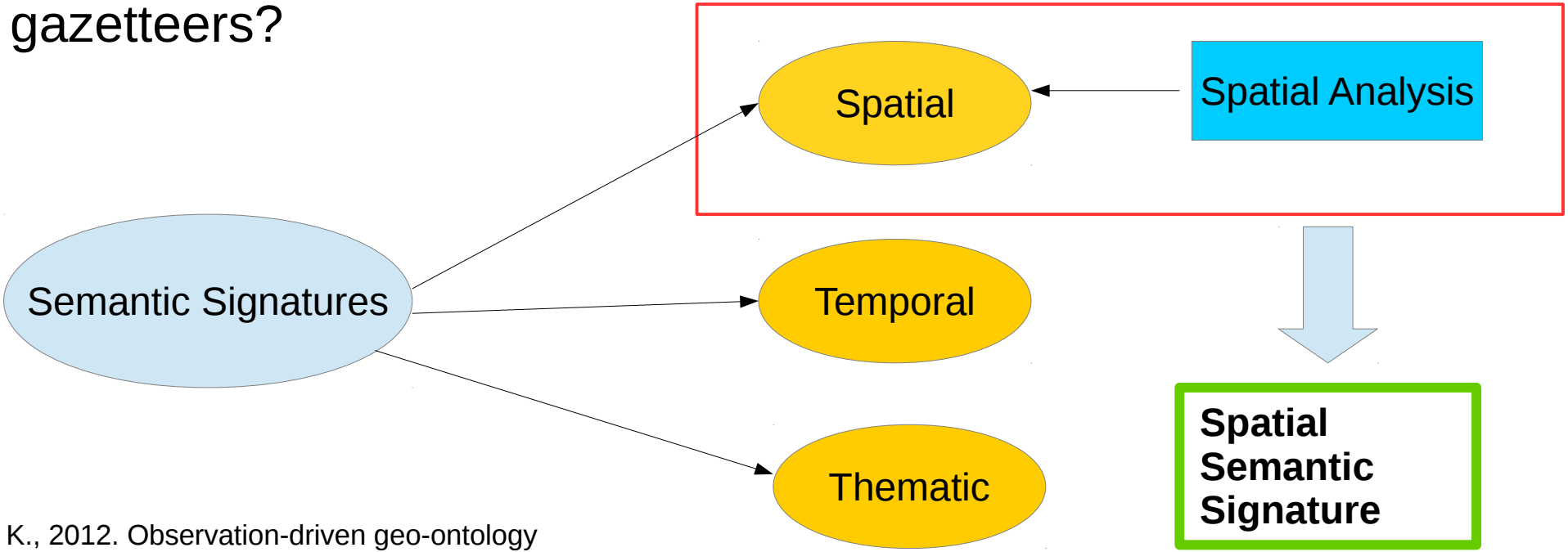
Motivations (cont.)

- However, their geo-ontologies/ typing schema are different!



Motivations (cont.)

- How to understand such semantic heterogeneity among gazetteers?



Janowicz, K., 2012. Observation-driven geo-ontology engineering. Transactions in GIS 16 (3), 351–374.

Data Sets

	DBpedia Places	GeoNames	Getty Thesaurus of Geographic Names (TGN)
	Extracted from DBpedia articles	Formal geographical database that contains over eight million place names	Focus on places that are culturally or historically significant
Number of feature types	72	234	285

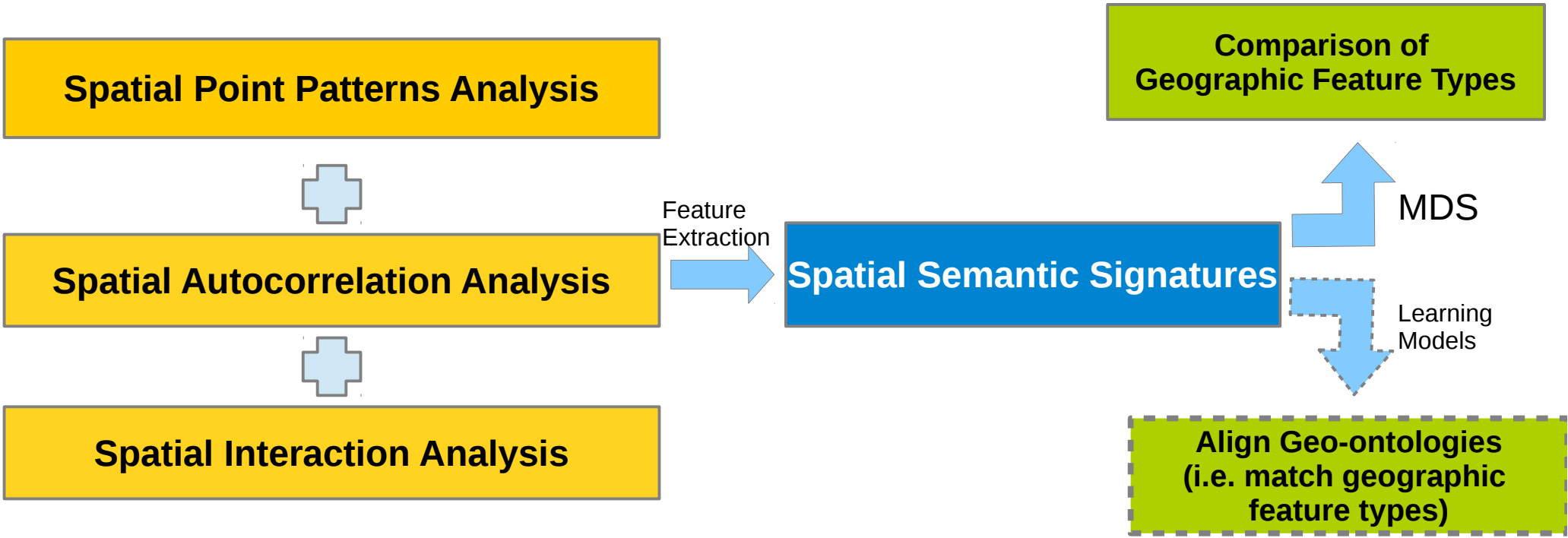
Data Sets (cont.)

- Common information
  - Toponyms/Place names
  - Geographic feature type
  - Spatial footprints

Dams in GeoNames

Methods

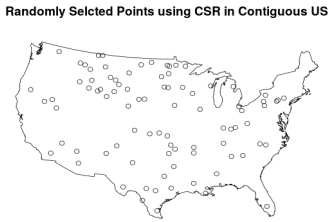
- Overview:



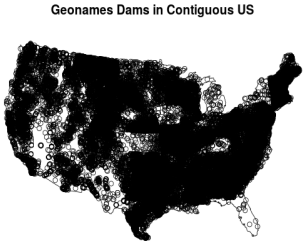
Methods (cont.)

- Spatial Point Pattern Analysis

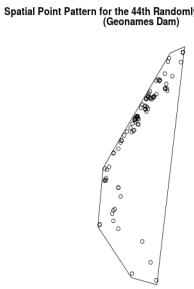
  Sampling for local analysis



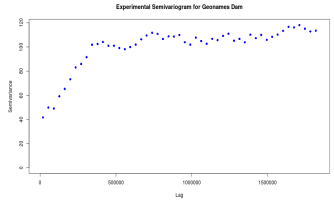
Generate random points (Complete Spatial Randomness)

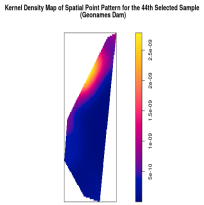


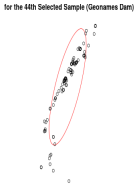
Points for one place type



Select nearest 100 neighbors for each random points







Conduct spatial point pattern analysis on these 100 neighbors

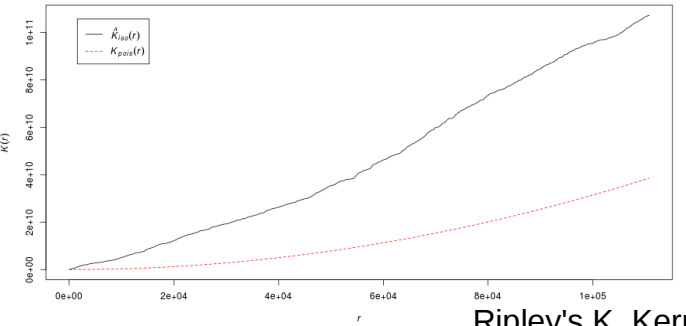
Average the statistics over all random points

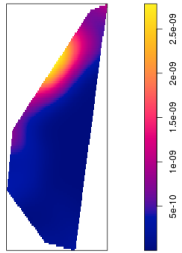
Methods (cont.)

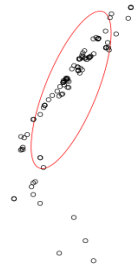
- Spatial Point Pattern Analysis (cont.)

  Spatial Semantic Signatures (Local)

  - **Intensity of point patterns**
  - **Distance to nearest neighbor**
  - Ripley's K (i.e. **range** and **mean deviation from the theoretical values**)
  - Kernel density estimation (i.e. **bandwidth** and **range**)
  - Standard deviation ellipse (i.e. **rotation**, **std. along x-axis** and **y-axis**)







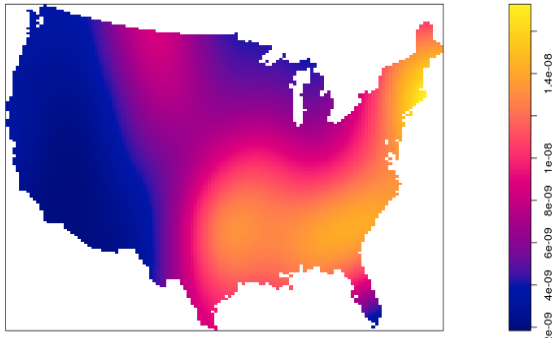
Ripley's K, Kernel density estimateion and Standard deviation ellipse for Dams in GeoNames

Methods (cont.)

- Spatial Point Pattern Analysis (cont.)

  Spatial Semantic Signatures (Global)

  - **Overall intensity of point patterns**
  - Kernel density estimation (i.e. **bandwidth** and **range**)



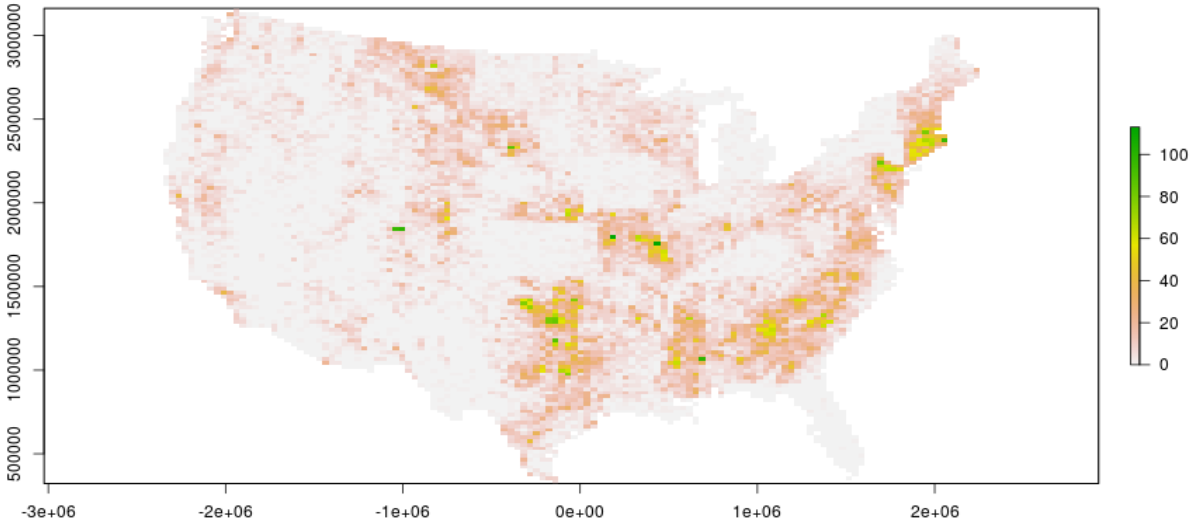
Dams in GeoNames

Methods (cont.)

- Spatial Autocorrelation Analysis

  Conversion: Point data → Raster Map

Dams in GeoNames



Cell size : 36 km * 22.2 km

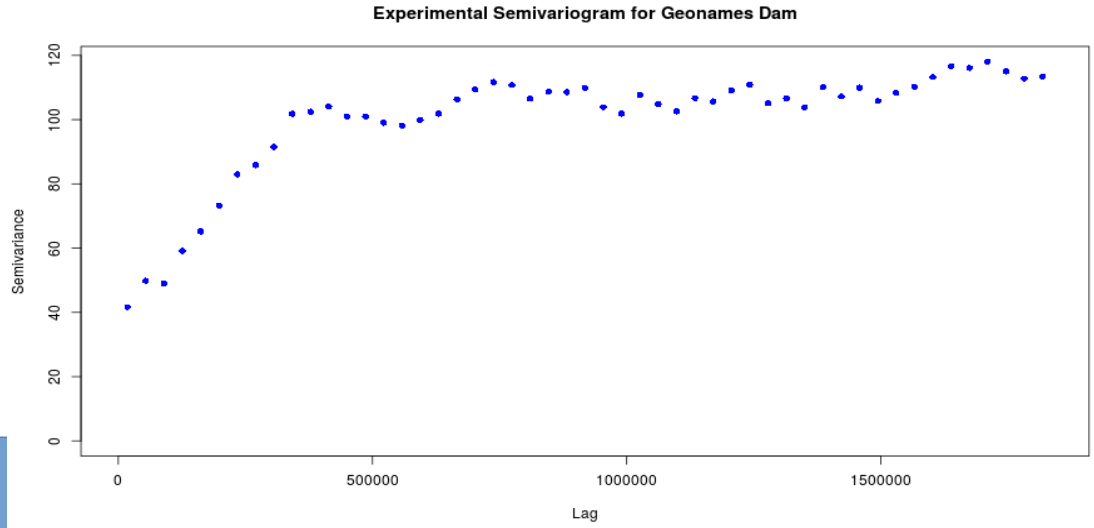
Cell value: number of instances falling in the cell

Methods (cont.)

- Spatial Autocorrelation Analysis (cont.)

  Spatial Semantic Signatures

  - **Global Moran's I**
  - Sample Semivariogram (i.e. **semivariances at first, median and last lag distances**).



Methods (cont.)

- Spatial Interaction with other geographic features

  Spatial Semantic Signatures

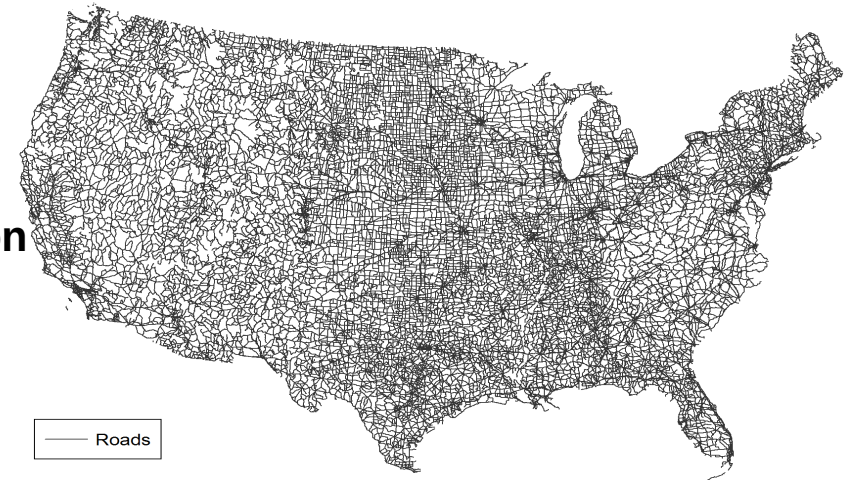
Population (LandScan2014)

Population for each feature point

Road Segment (Digital Chart of the World)

Distance to nearest segment for each feature point

- **Mimimum**
- **Maximum**
- **Mean**
- **Standard deviation**



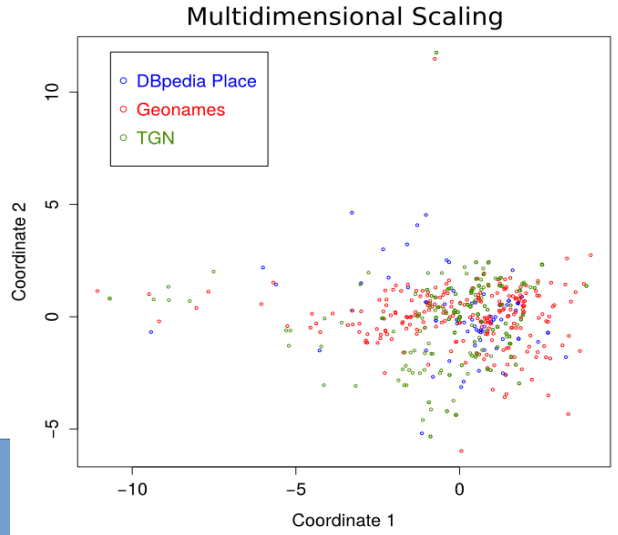
Methods (cont.)

24 features for spatial semantic signature

Transform high-dimensional feature space to 2-dimensional (Multidimensional scaling)

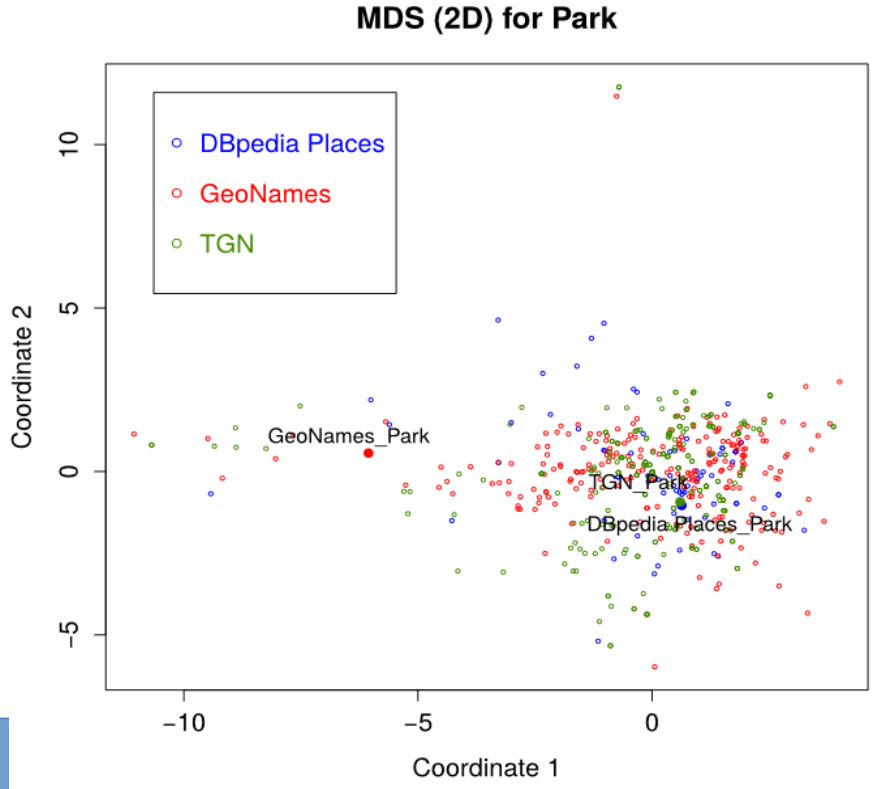
Visualize it on a 2D map

Spatial Point Patterns		Spatial Autocorrelations	Spatial Interactions with Other Geographic Features
Local	Intensity	Global Moran' I	Population Value (min)
	Mean Distance to Nearest Neighbor		Population Value (max)
	Kernel Density (range)		Population Value (mean)
	Kernel Density (band width)		Population Value (standard deviation)
	Ripley's K (range)		
	Ripley's K (mean deviation)		
	Standard Deviation Ellipse (rotation)	Semivariogram (nugget)	Shortest Distance to Road (min)
	Standard Deviation Ellipse (std. along x-axis)		Shortest Distance to Road (max)
	Standard Deviation Ellipse (std. along y-axis)	Semivariogram (median distance lag)	Shortest Distance to Road (mean)
Global	Intensity		Shortest Distance to Road (standard deviation)
	Kernel Density (range)	Semivariogram (last distance lag)	
	Kernel Density (band width)		



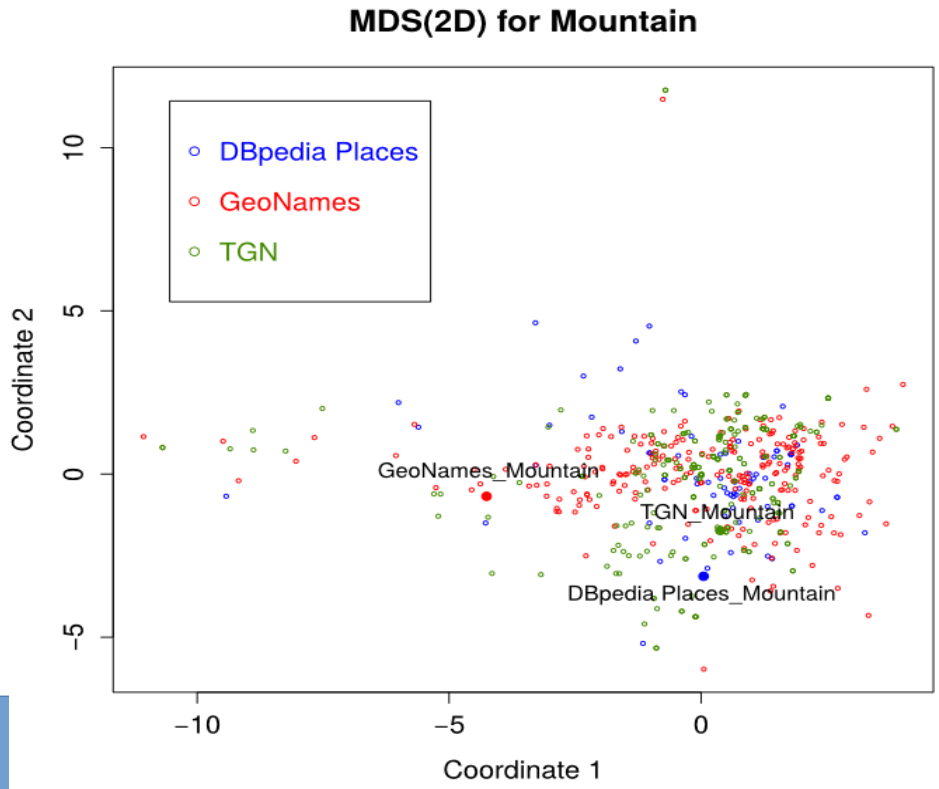
Results and Discussions

- Same names and similar spatial patterns



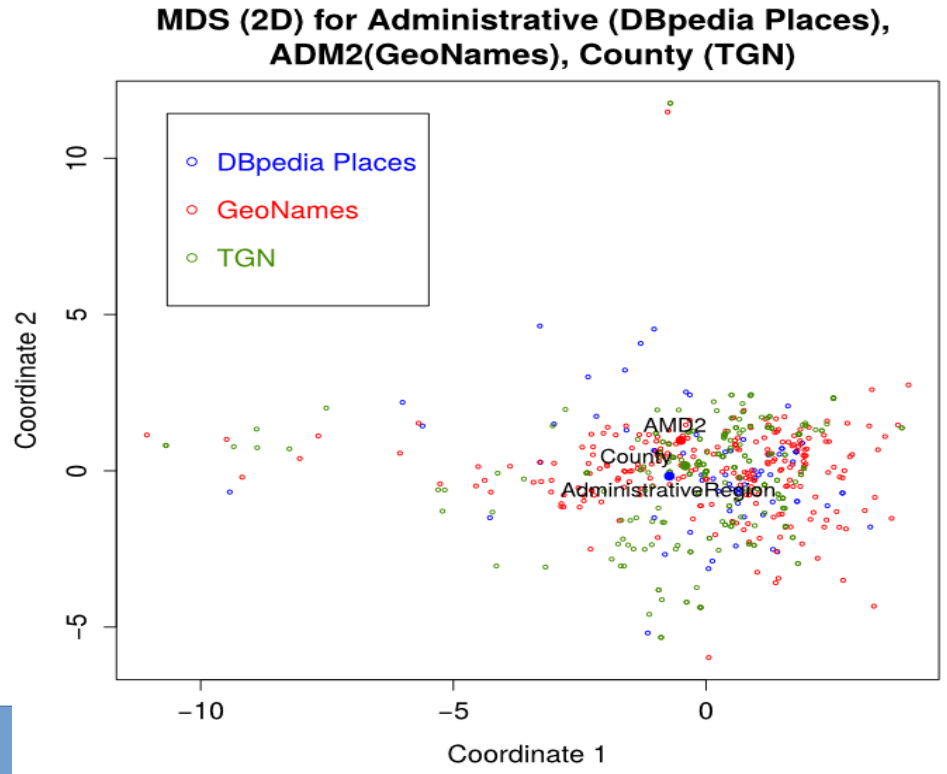
Results and Discussions (cont.)

- Same names but different spatial pattern



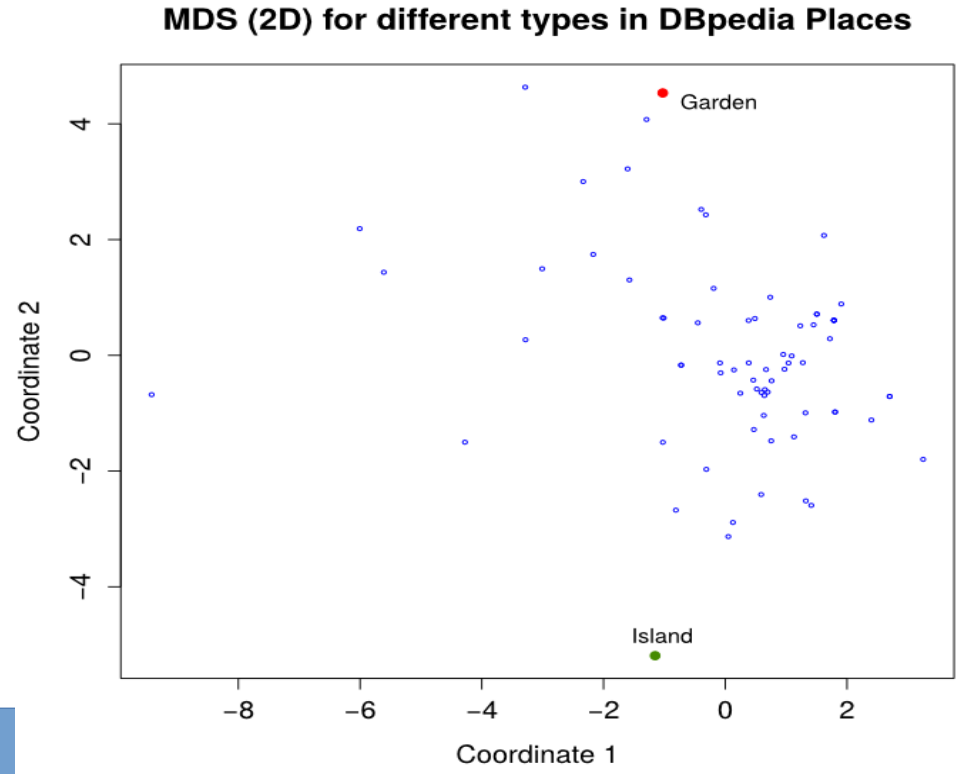
Results and Discussions (cont.)

- Different names but similar spatial pattern



Results and Discussions (cont.)

- Different names and different spatial pattern



Future work

- Derive additional statistical features to represent the spatial semantic signature (e.g. statistics for co-occurrence, topological relations);
- Quantify the dissimilarity/similarity of place types using such spatial semantic signatures (e.g. supervised/ unsupervised learning algorithms);
- Combine spatial signatures with previously studied temporal and thematic signatures;
- Integrate this study (bottom-up) with classical top-down knowledge engineering.

Special thanks to:

Yingjie Hu, Krzysztof Janowicz and Grant McKenzie

Questions and/or comments?