Enriching Top-down Geo-ontologies Using Bottom-up Knowledge Mined from Linked Open Data

Yingjie Hu and Krzysztof Janowicz
STKO Lab, Department of Geography, U.C. Santa Barbara

Outline

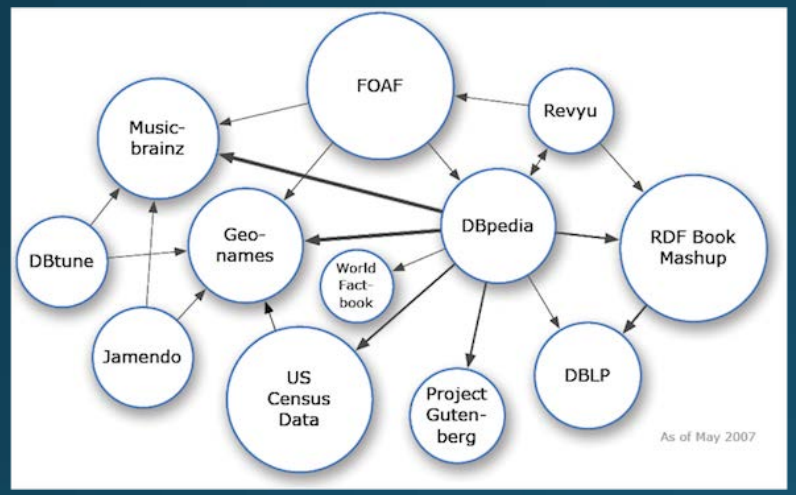
- Introduction
- Workflow
- Experiment
- Conclusions

Introduction

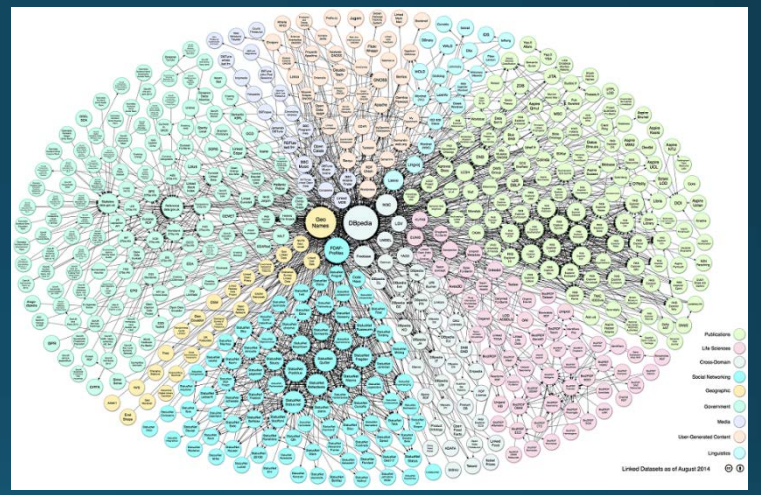
- Geo-ontologies play important roles in GIScience
  - Enhance semantic interoperability
  - Improve geographic information retrieval
  - Support spatial decision making
  - …
- A top-down approach for developing geo-ontologies
  - Pros: captures valuable expert knowledge; provides concise and meaningful terms
  - Cons: the derived ontology may be biased towards the opinions of the participating experts; or may be incomplete

Introduction

- Linked Open Data (LOD) cloud: a fast evolving data resource



2007

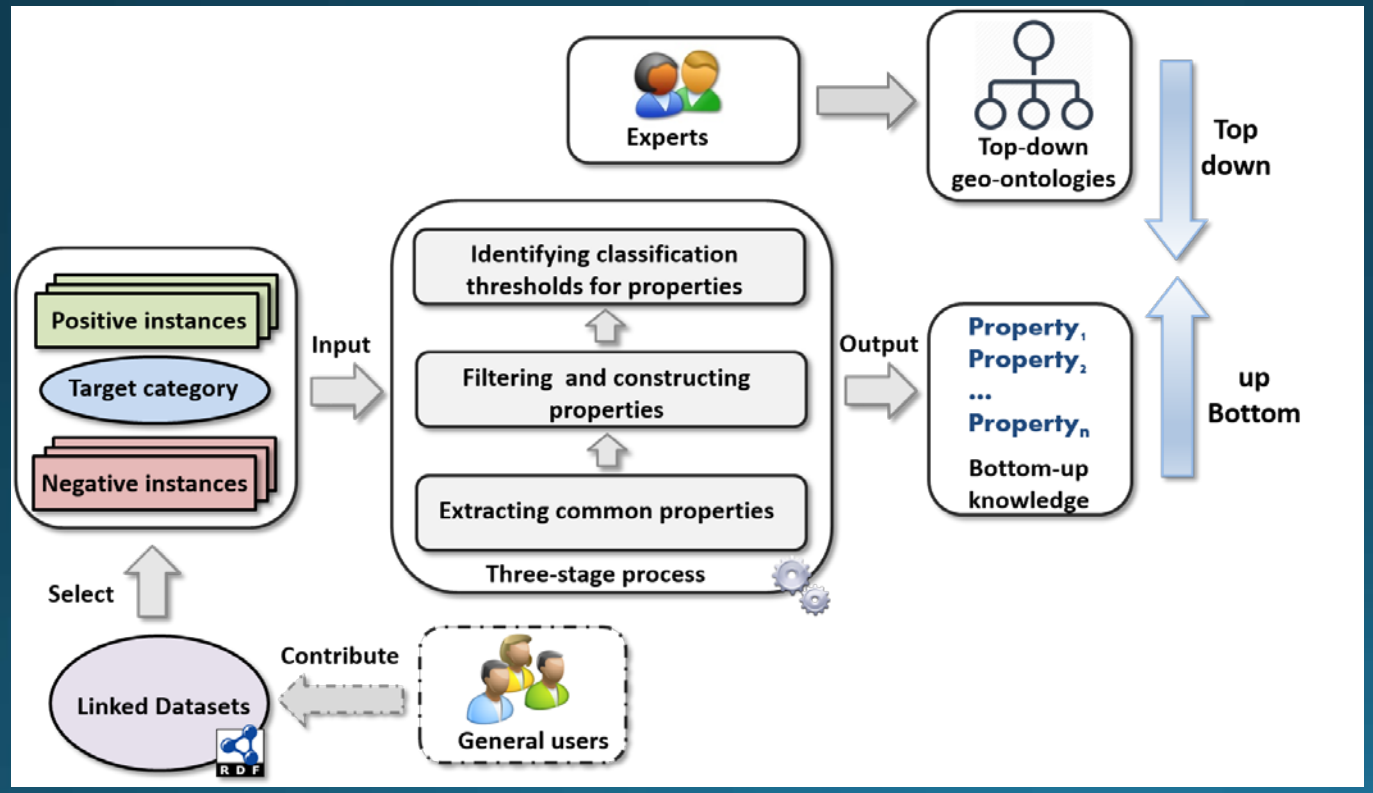


2014

- Merits of the LOD cloud:
  - A rich amount of data from both authorities and the general public
  - A lot of data are about geographic places: DBpedia, Geonames, LinkedGeoData, …
  - Data are structured using Resource Description Framework (RDF)

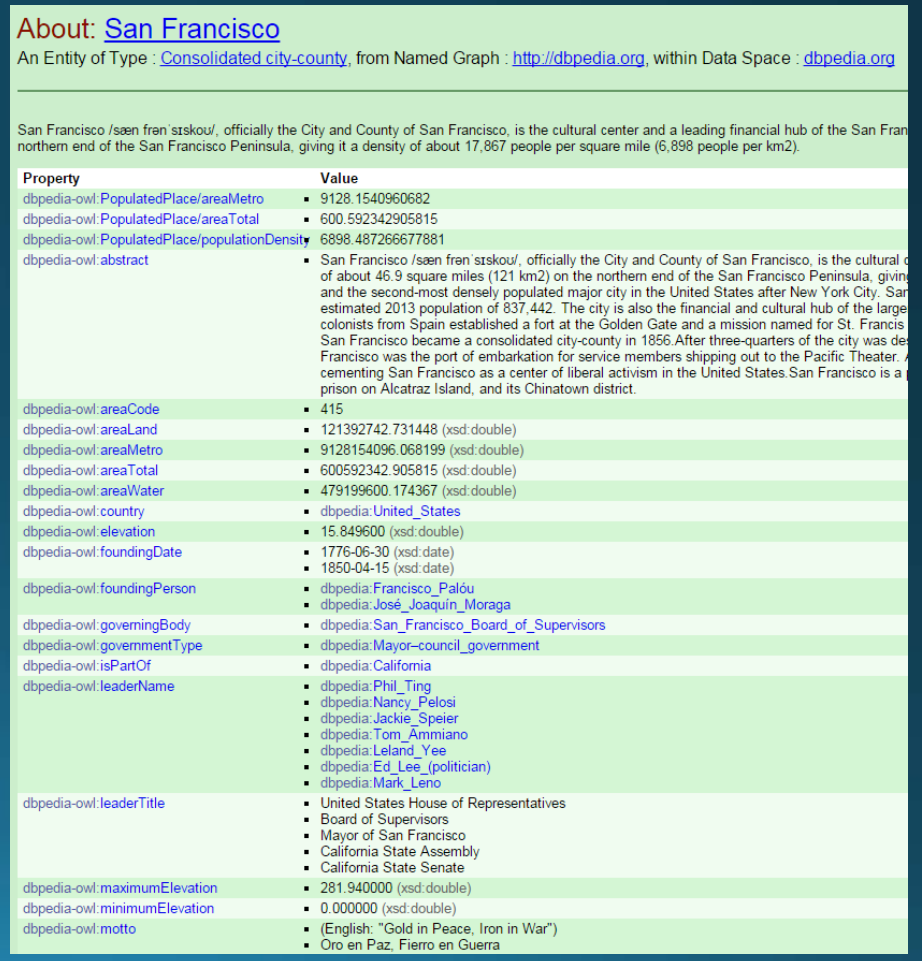
Workflow

- A workflow for extracting bottom-up knowledge
  - A concept learning approach



Workflow

- Three-stage process for extracting knowledge
  - 1. Extracting common properties
    - Properties only in positive instances
    - Properties shared by both instances
  - 2. Filtering and constructing properties
    - Filter out irrelevant properties, e.g., leaderTitle
    - Construct potentially relevant properties, e.g., population density
  - 3. Identifying distinguishable properties and classification thresholds



About: San Francisco

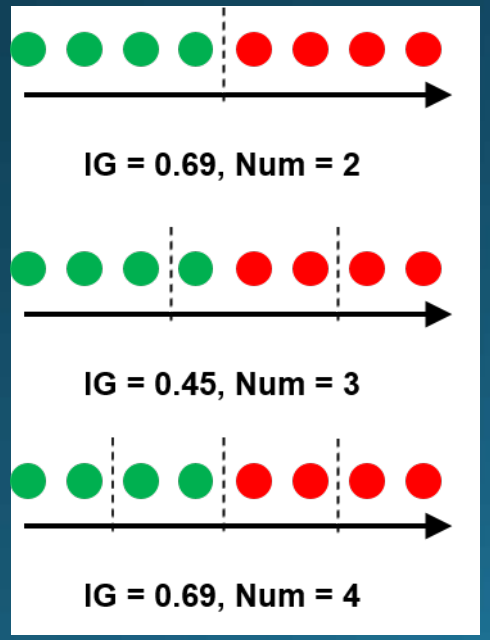
An Entity of Type : Consolidated city-county, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

San Francisco /sæn frənˈsɪskoʊ/, officially the City and County of San Francisco, is the cultural center and a leading financial hub of the San Fran
northern end of the San Francisco Peninsula, giving it a density of about 17,867 people per square mile (6,898 people per km2).

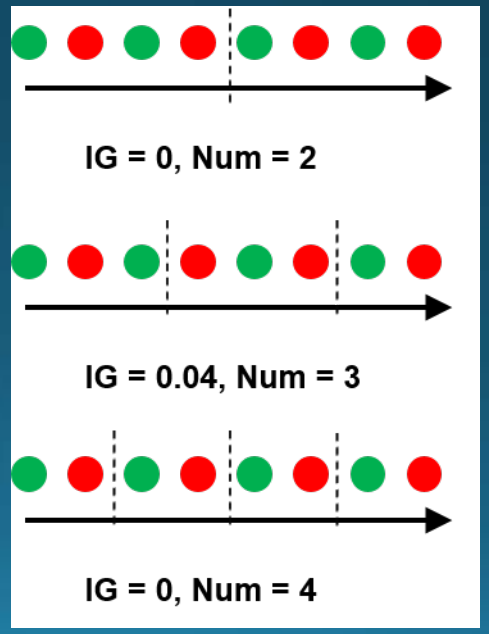
Property	Value
dbpedia-owl:PopulatedPlace/areaMetro	9128.1540960682
dbpedia-owl:PopulatedPlace/areaTotal	600.592342905815
dbpedia-owl:PopulatedPlace/populationDensity	6898.487266677881
dbpedia-owl:abstract	San Francisco /sæn frənˈsɪskoʊ/, officially the City and County of San Francisco, is the cultural c of about 46.9 square miles (121 km2) on the northern end of the San Francisco Peninsula, giving and the second-most densely populated major city in the United States after New York City. San estimated 2013 population of 837,442. The city is also the financial and cultural hub of the large colonists from Spain established a fort at the Golden Gate and a mission named for St. Francis San Francisco became a consolidated city-county in 1856.After three-quarters of the city was de Francisco was the port of embarkation for service members shipping out to the Pacific Theater. cementing San Francisco as a center of liberal activism in the United States.San Francisco is a prison on Alcatraz Island, and its Chinatown district.
dbpedia-owl:areaCode	415
dbpedia-owl:areaLand	121392742.731448 (xsd:double)
dbpedia-owl:areaMetro	9128154096.068199 (xsd:double)
dbpedia-owl:areaTotal	600592342.905815 (xsd:double)
dbpedia-owl:areaWater	479199600.174367 (xsd:double)
dbpedia-owl:country	dbpedia:United_States
dbpedia-owl:elevation	15.849600 (xsd:double)
dbpedia-owl:foundingDate	1776-06-30 (xsd:date) 1850-04-15 (xsd:date)
dbpedia-owl:foundingPerson	dbpedia:Francisco_Palóu dbpedia:José_Joaquín_Moraga
dbpedia-owl:governingBody	dbpedia:San_Francisco_Board_of_Supervisors
dbpedia-owl:governmentType	dbpedia:Mayor–council_government
dbpedia-owl:isPartOf	dbpedia:California
dbpedia-owl:leaderName	dbpedia:Phil_Ting dbpedia:Nancy_Pelosi dbpedia:Jackie_Speier dbpedia:Tom_Ammiano dbpedia:Leland_Yee dbpedia:Ed_Lee_(politician) dbpedia:Mark_Leno
dbpedia-owl:leaderTitle	United States House of Representatives Board of Supervisors Mayor of San Francisco California State Assembly California State Senate
dbpedia-owl:maximumElevation	281.940000 (xsd:double)
dbpedia-owl:minimumElevation	0.000000 (xsd:double)
dbpedia-owl:motto	(English: "Gold in Peace, Iron in War") Oro en Paz, Fierro en Guerra

Workflow

- Three-stage process for extracting knowledge
  - 3. Identifying distinguishable properties and classification thresholds
    - Segment instances in a property into an increasing numbers of groups
    - Calculate entropy for each segmentation $entropy(X) = -\sum_{i=\{pos,neg\}} P(x_i)\log P(x_i)$
    - Information gain before and after the property has been segmented $IG = entropy_b(X) - entropy_a(X)$



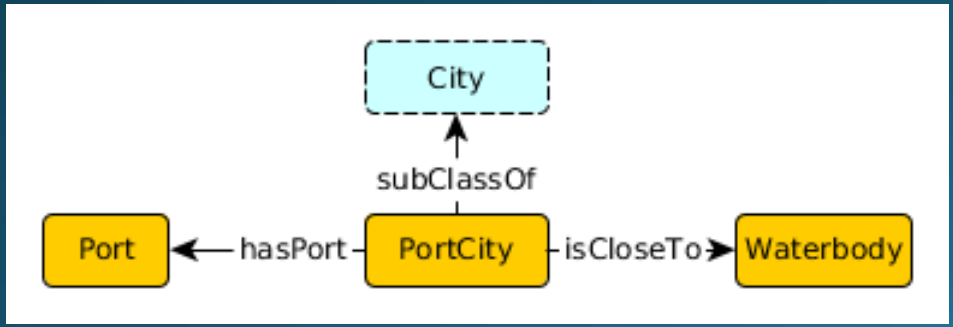
(A) a property with clear cut

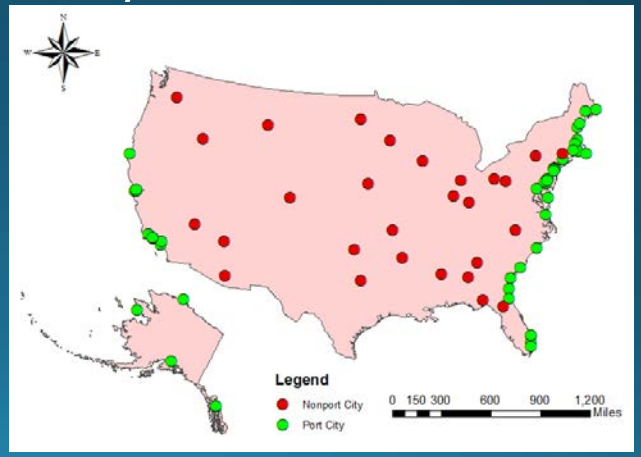


(B) a property with mixed instances

Experiment

- An example geographic concept (*port city*) and a possible top-down ontology
- A sample dataset from DBpedia
  - Target category: Port cities and towns of the United States Atlantic coast and Port cities and towns of the United States Pacific coast
  - Positive instances: 49 cities which have been classified into these two categories by Wikipedia users
  - Negative instances: 29 inland U.S. cities randomly selected





Experiment

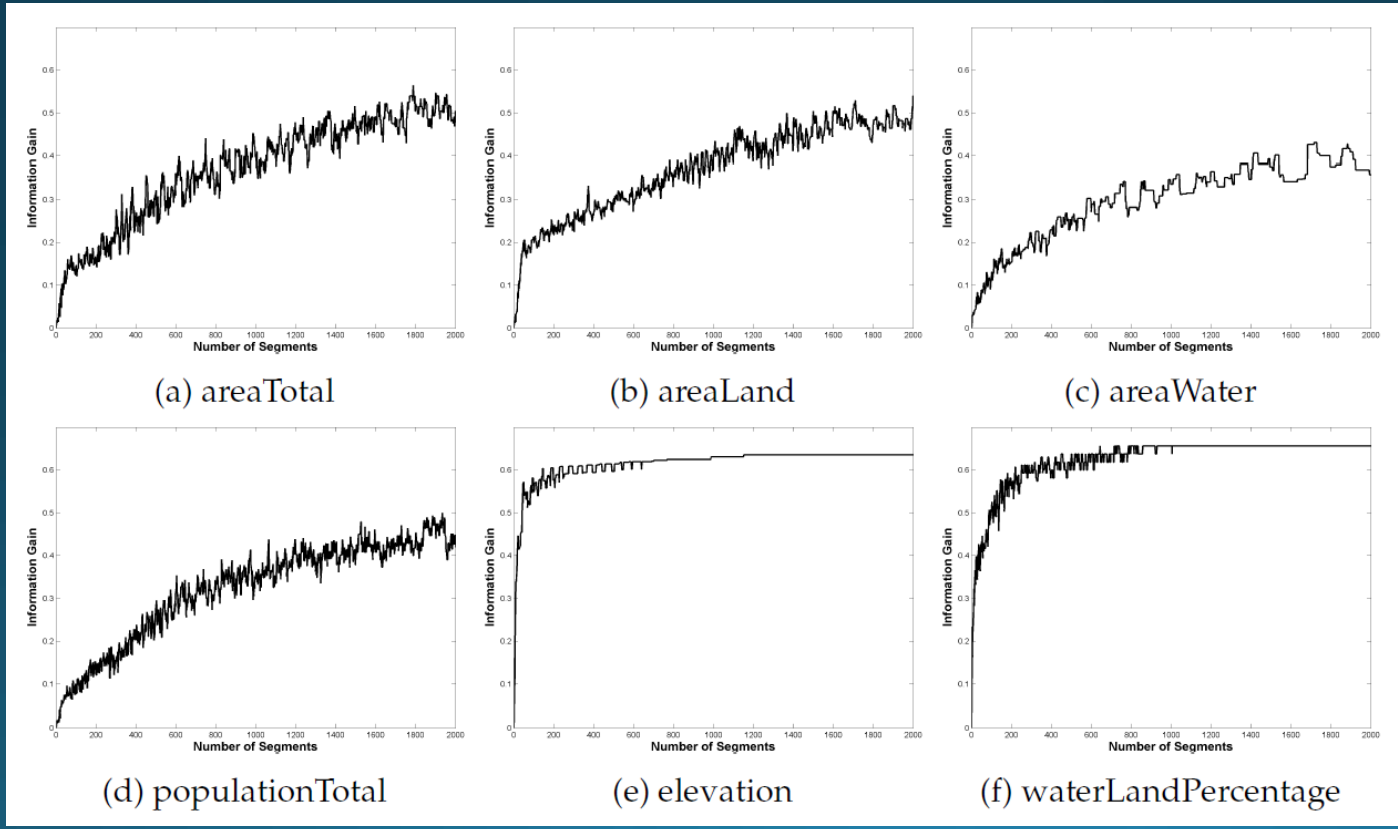
- A Java program developed to identify common properties
  - Properties shared by at least 95% of positive instances and no more than 5% of negative instances: is dbpedia-owl:homeport of. dbpedia:Ship
  - Properties shared by at least 95% of both positive and negative instances

- Filtering irrelevant properties and constructing new properties
  - Filtering out irrelevant properties, e.g., names of the celebrities…
  - Constructing a new property, waterLandPercentage

dbpedia-owl:areaTotal	dbpedia-owl:areaLand	dbpedia-owl:areaWater
dbpedia-owl:populationTotal	dbpedia-owl:elevation	waterLandPercentage

Experiment

- Examining the information gain for each property under different numbers of segmentations

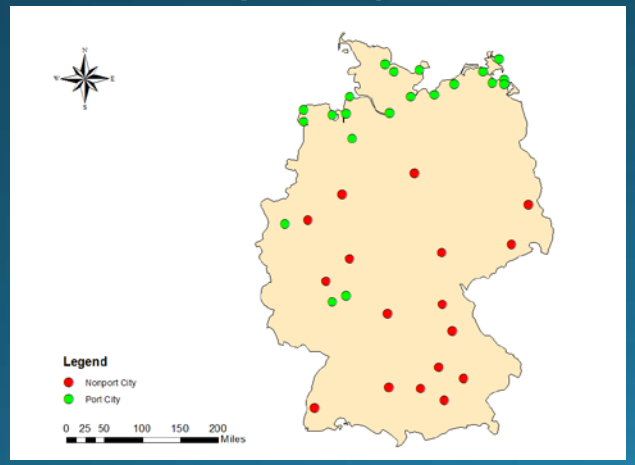


Experiment

- Aggregate the values of positive instances to derive thresholds

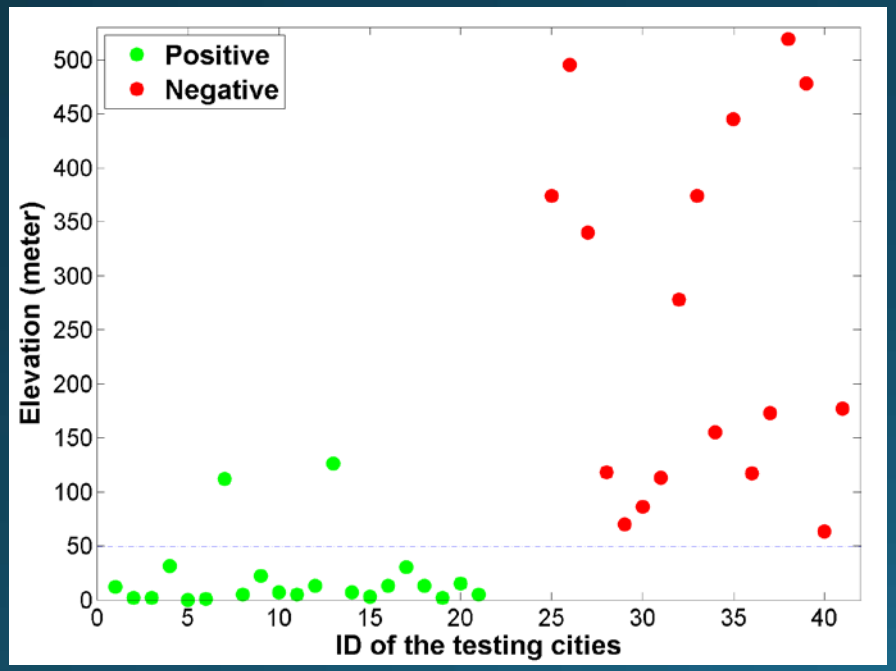
$$elevation < 49.36$$
$$waterLandPercentage > 11.79\%$$

- Evaluation: does the extracted knowledge make sense?
  - An unseen dataset from DBpedia to test the extracted knowledge
  - 38 cities from Germany (21 positive and 17 negative)

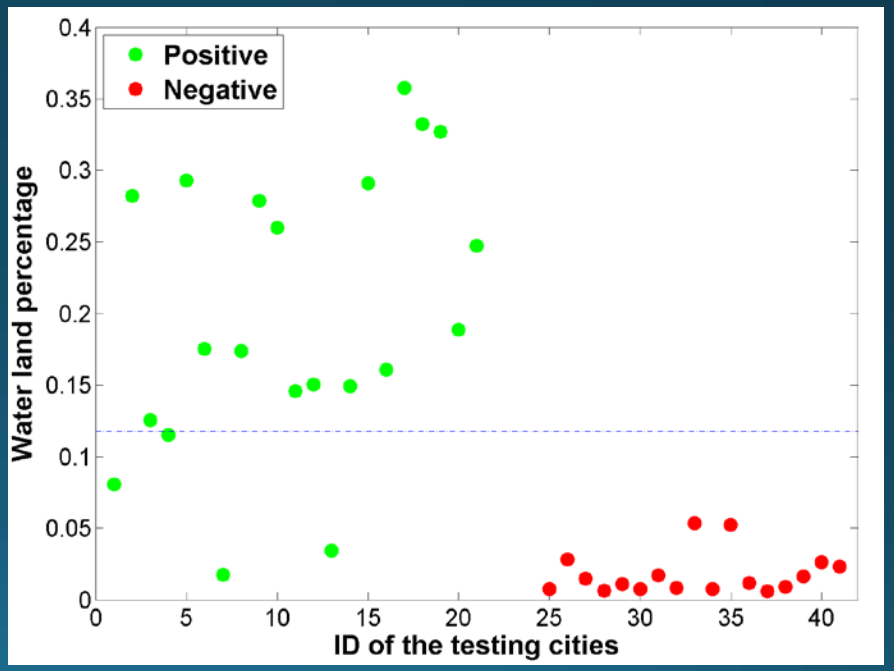


Experiment

- Evaluating the extracted knowledge using unseen cities



(A) Elevation



(B) WaterLandPercentage

Conclusions and Future work

- Top-down geo-ontologies capture valuable expert knowledge but may be biased or incomplete
- The rich amount of data from the LOD cloud provide a resource to mine geographic knowledge
- This study presents a preliminary framework to extract bottom-up knowledge from Linked Datasets
- Limitations and future work:
  - The selection of positive and negative instances
  - Regional variability of geographic concepts

Thank you!

Yingjie Hu
yingjiehu@umail.ucsb.edu
http://geog.ucsb.edu/~hu