USC Viterbi
School of Engineering

Harvesting Geospatial Knowledge from Social Metadata

Suradej Intagorn

Anon Plangprasopchok

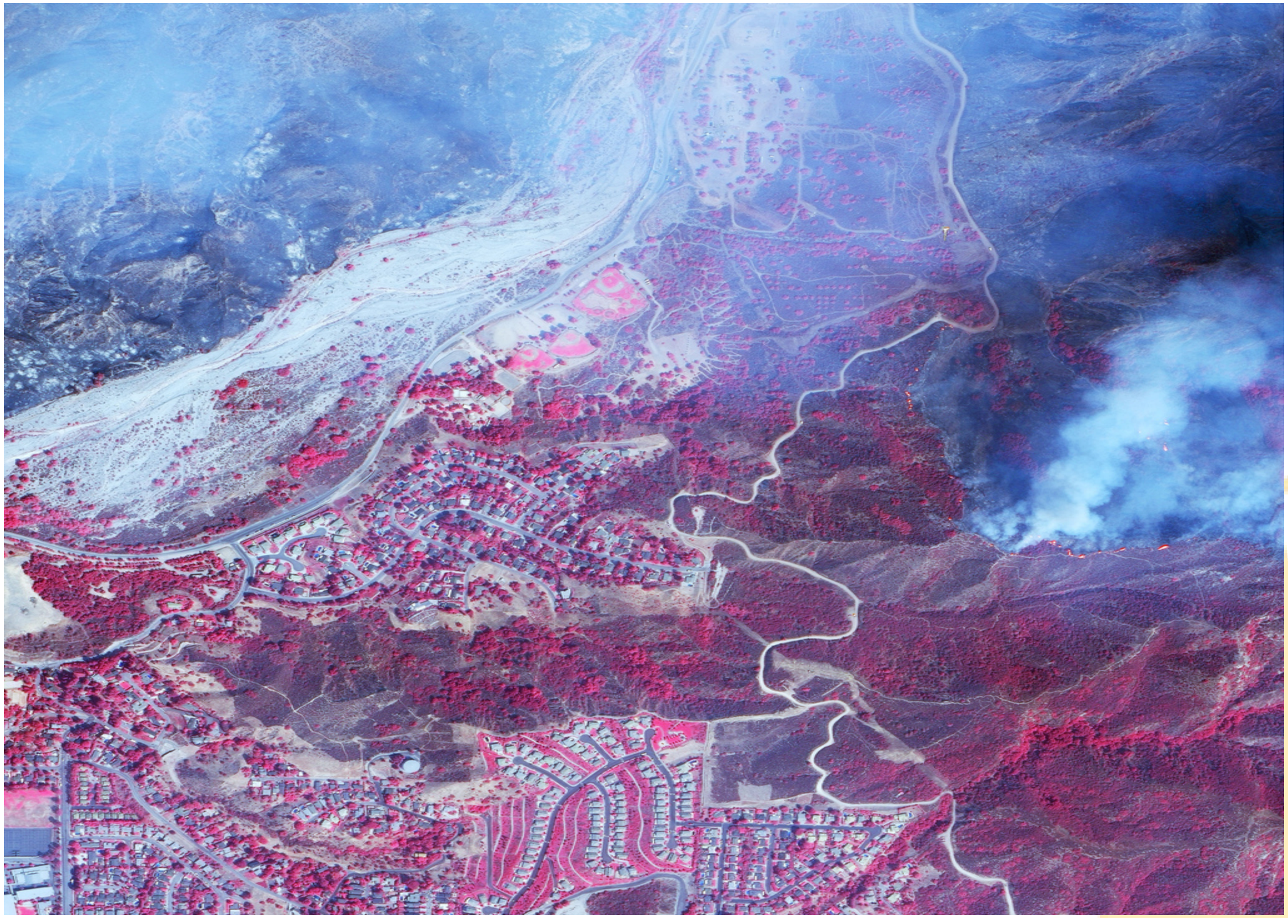
Kristina Lerman

USC Information Sciences Institute

USC

USC Viterbi
School of Engineering

Motivation



Tags:

station
fire
los
angeles
california
satellite
image
digitalglobe
Los Angeles
USA

USC

Credit: DigitalGlobe

Goals

Problem

- How to extract ***geospatial knowledge*** from user contributed annotations (metadata)?

Advantages

- Cheaper and Up-to-date than knowledge produced by a group of experts
- Supplement to exist knowledge created by experts

Overview of Approach

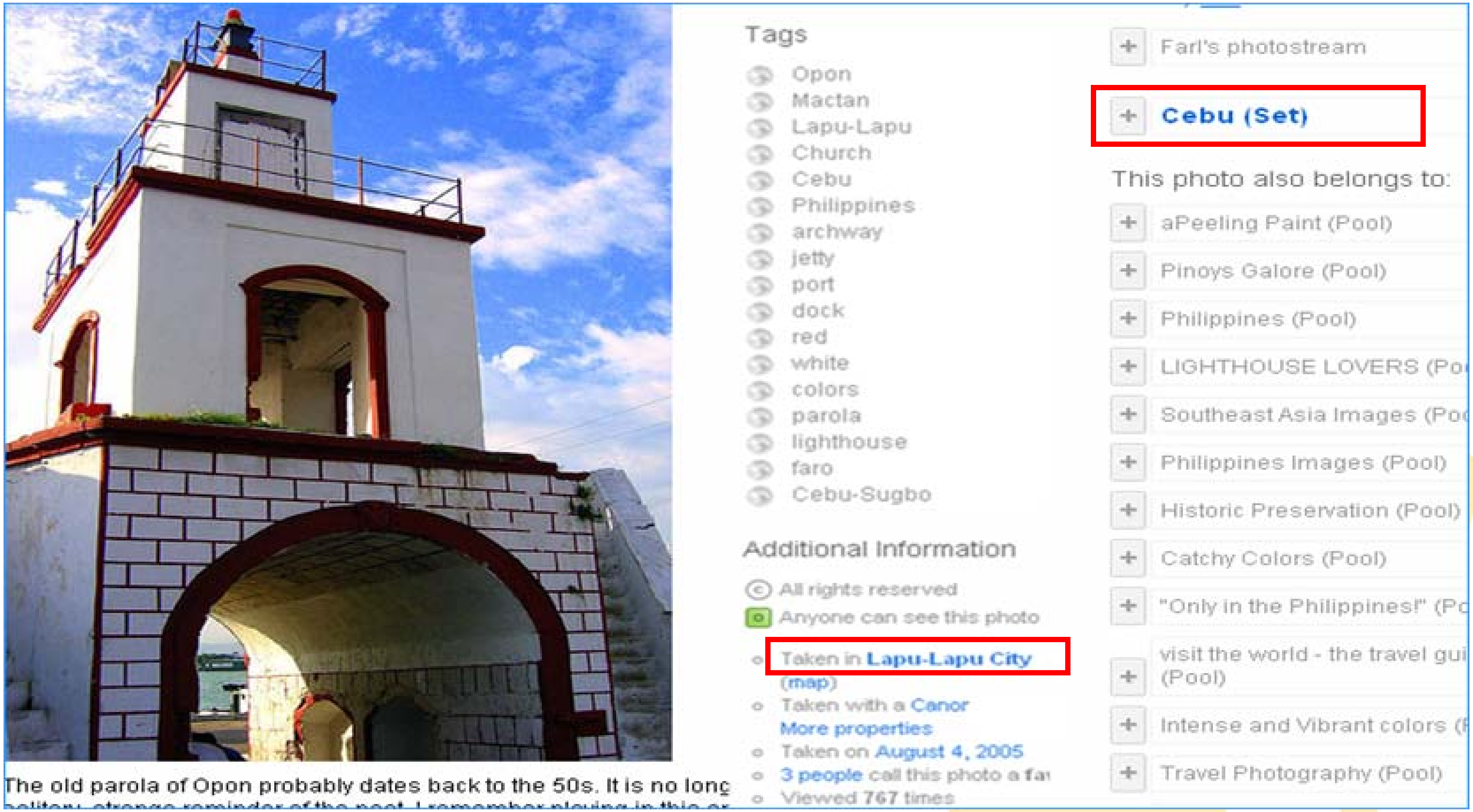
Combine geospatial annotations from many users to learn more complete representation of places & their relations

- tags,
- geo-tags
- set names (to be explained)

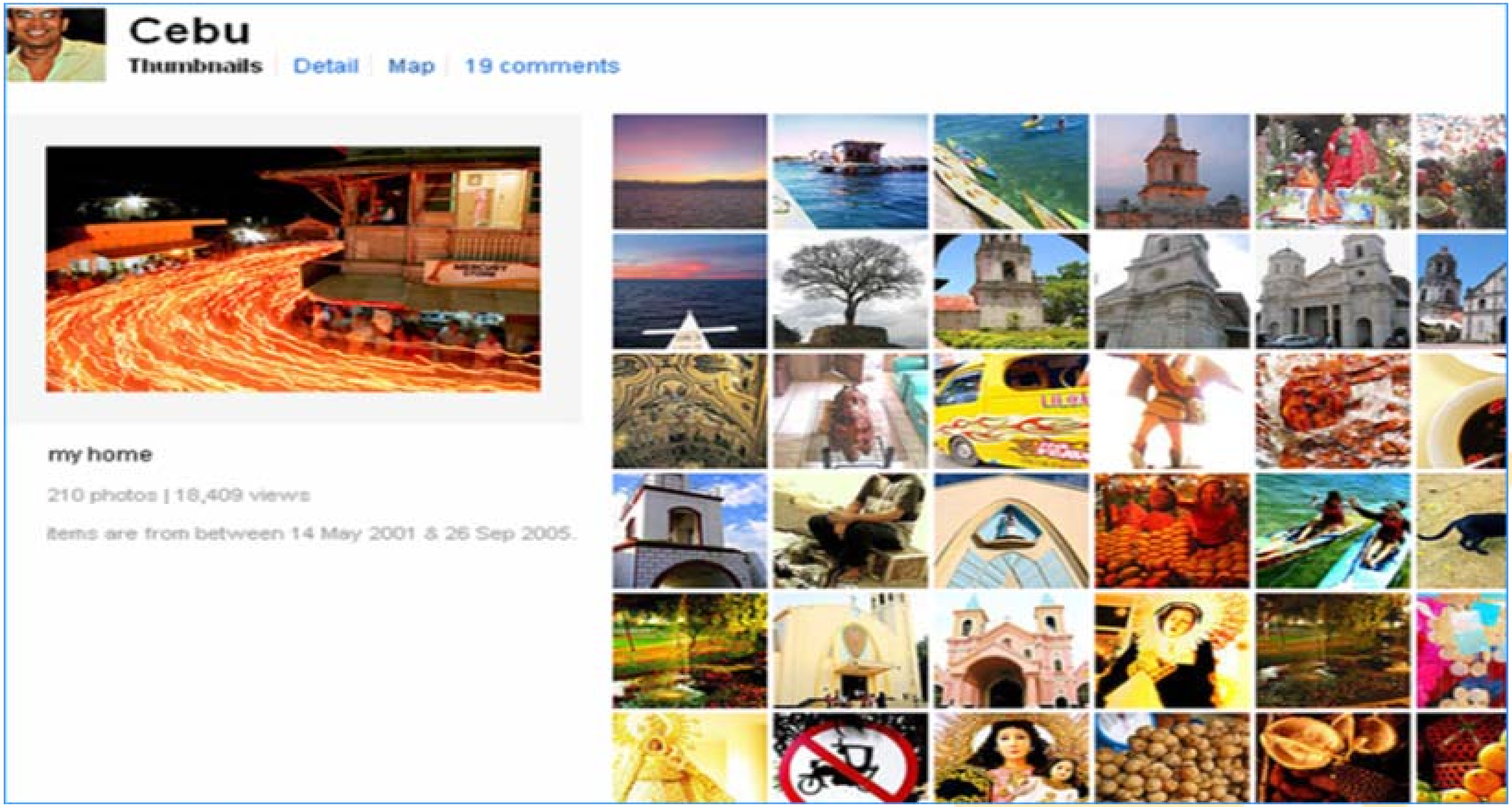
Challenges:

1. Place names are often ambiguous
2. Noisy data points
3. Sparseness

Types of Social Meta Data

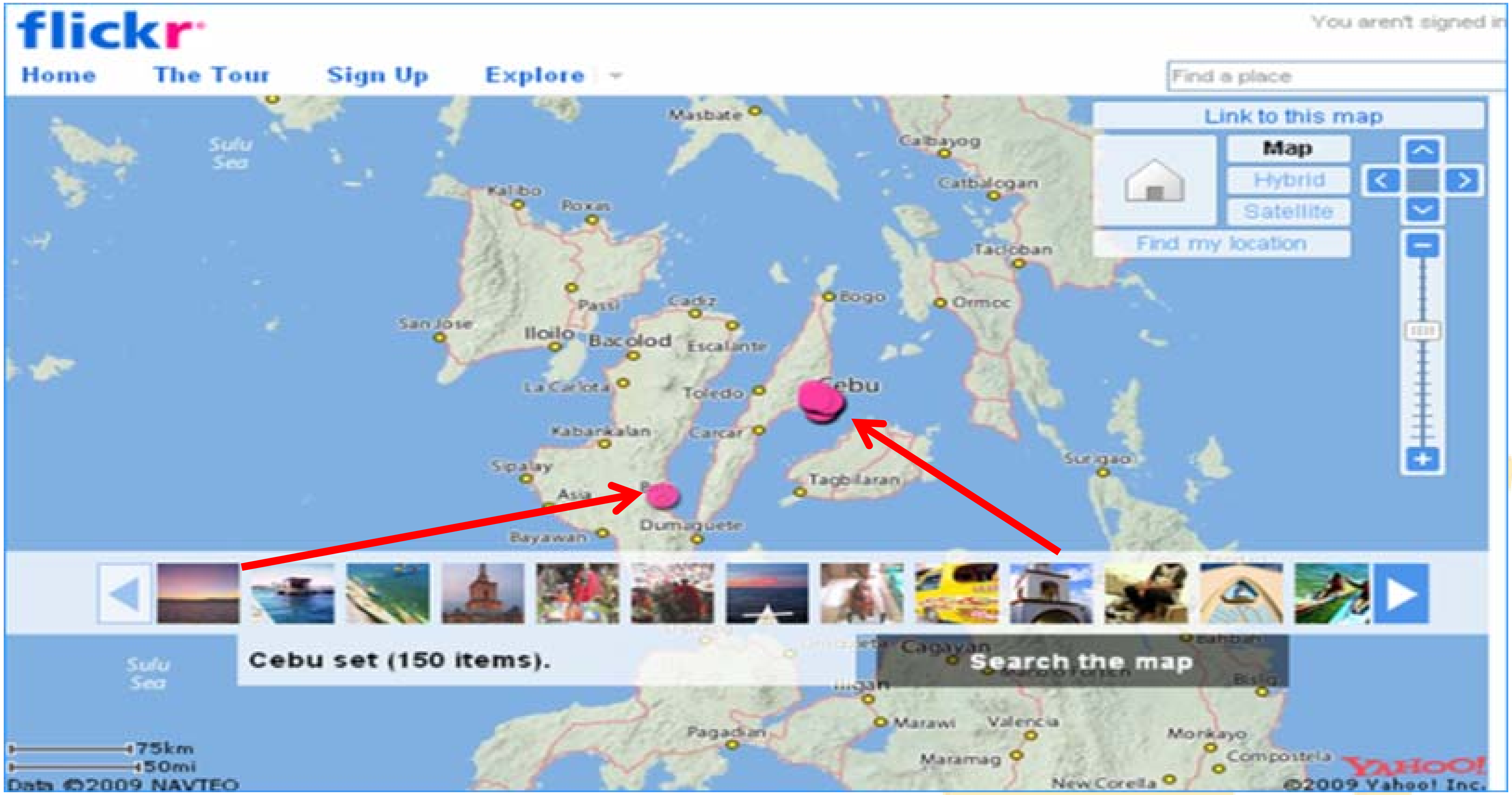


Types of Social Meta Data



USC Viterbi School of Engineering

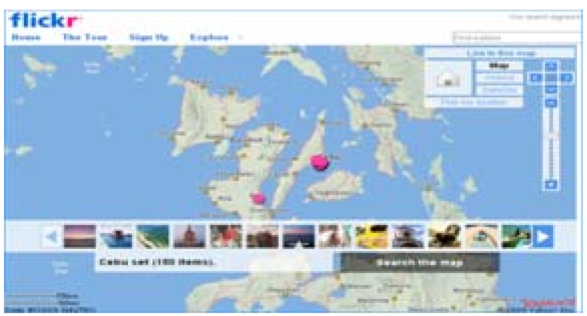
Types of Social meta data



USC

Summary of Social Metadata

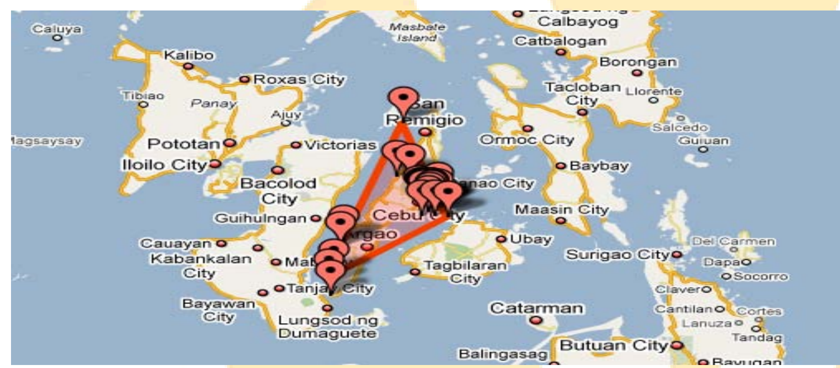
Photo
Photo ID
User ID
Set ID
Set Name
Latitude
Longitude



- Each photo is just a point in spatial domain.

- Each set is a collection of points representing a place

- They can be aggregated to get better representation of the place.



Approach

Recognizing Places

Disambiguating Places

Noise Filtering

Representing Places

Geographic subsumption

GeoNames

USC Viterbi School of Engineering

Approach

Recognizing Places

↓

Disambiguating Places

↓

Noise Filtering

↓

Representing Places

↓

Geographic subsumption

GeoNames

USC

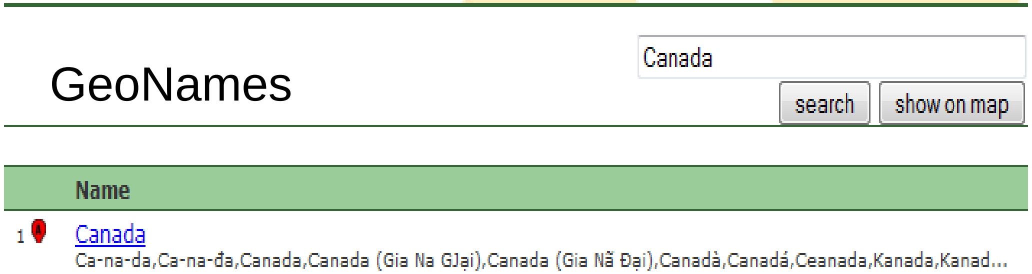
Recognizing Place-related Set Names

Assumption

- Set names refer to places
- But, need to filter out non-place set name

Current approach

- Use reference set from GeoNames.org
- Simple name matching with set name



Approach

Recognizing Places

GeoNames

Disambiguating Places

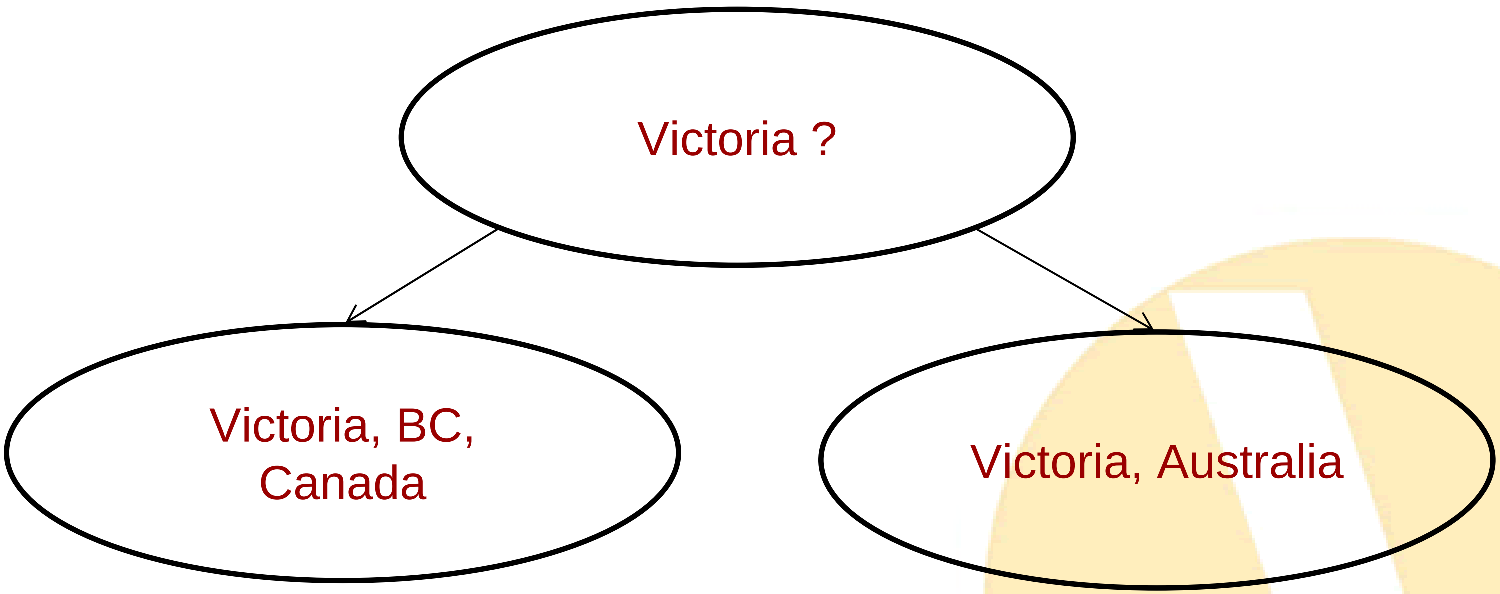
Noise Filtering

Representing Places

Geographic subsumption

Challenges: Ambiguity

Place names are often ambiguous



USC Viterbi School of Engineering

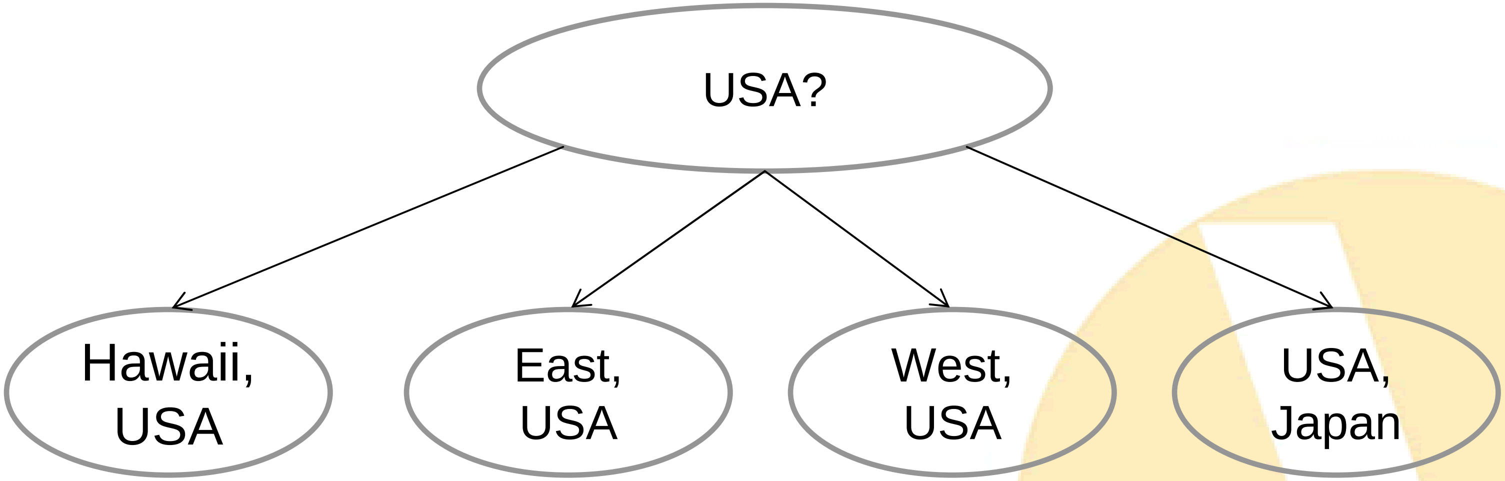
Disambiguating Places

Problem: Place may be also non-contiguous

USC

Disambiguating Places

Two "USA" places: (1) United States; (2) Usa, Japan



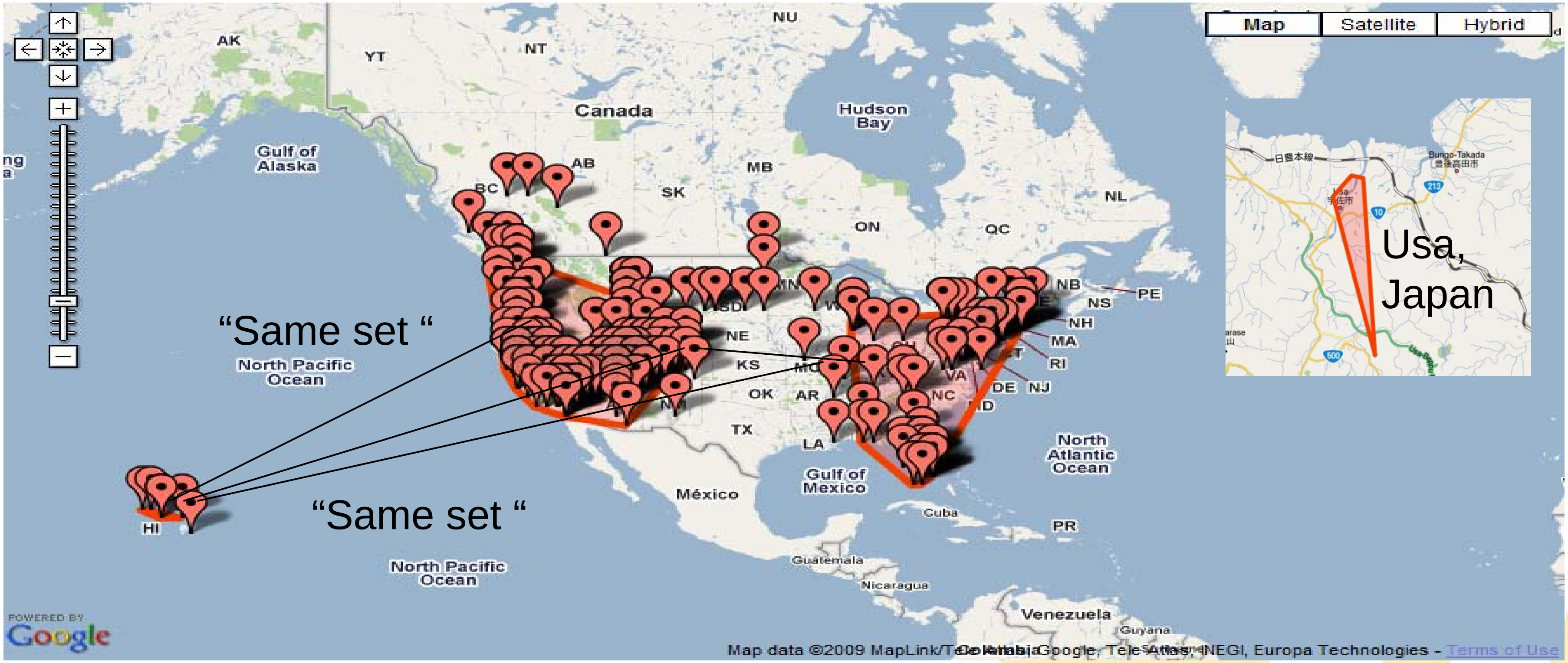
Disambiguating Places

Assumption: ambiguous places can be disambiguated by their locations

How?

1. Construct graph of geo points
2. Then, finding connected components

Learning Non-Contiguous Places



Disambiguating Places

Construct Graph *G1 = (V, E)*,

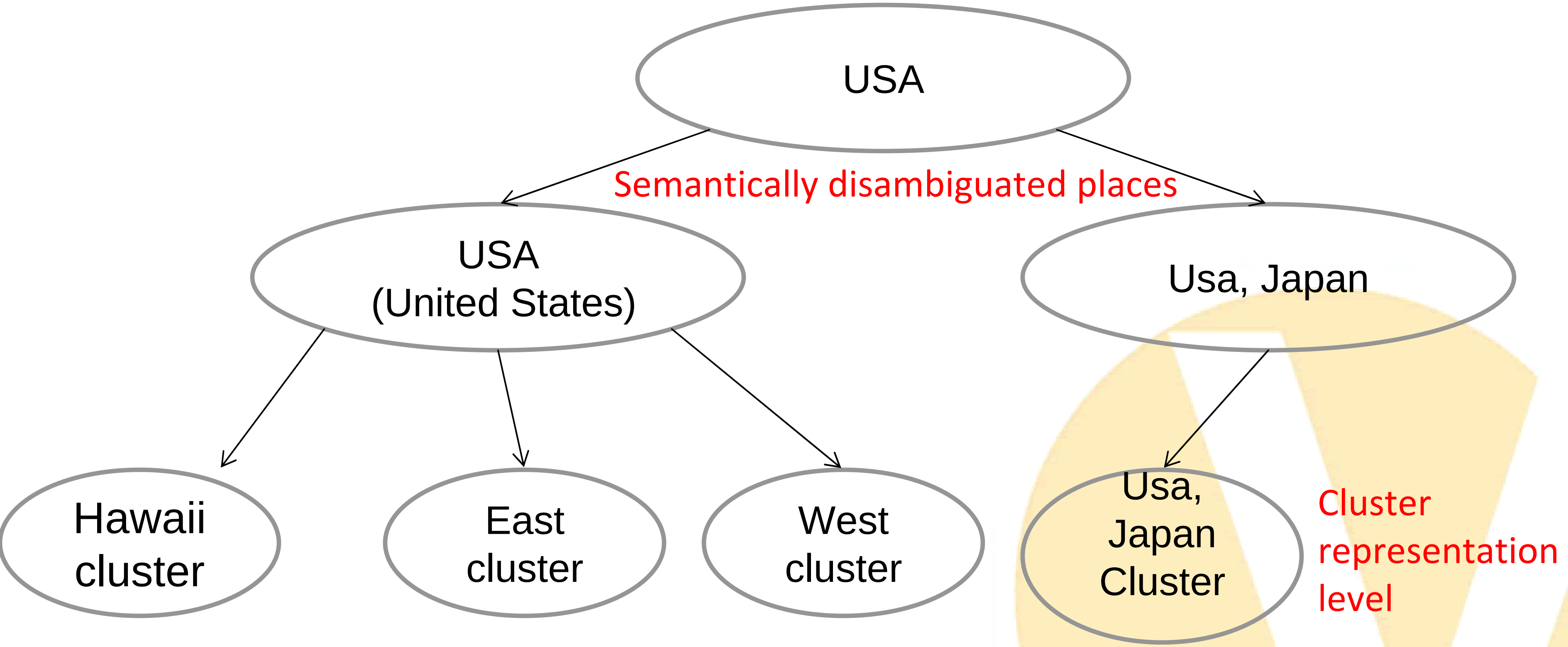
Vertices, *V*, are points corresponding to geo-tagged photos

Edges, *E*, are the edges between vertices

When is the edge between two photos created?

$$dist(v_i,v_j) < \tau \ \text{ or } \ s_{v_i} = s_{vj}.$$

Multi Cluster Representation



Approach

Recognizing Places

GeoNames

Disambiguating Places

Noise Filtering

Representing Places

Geographic subsumption

Noise Filtering

Heuristic: Cluster that contains photos from < k users is noise

USC Viterbi
School of Engineering

Approach

Recognizing Places

GeoNames

Disambiguating Places

Noise Filtering

Representing Places

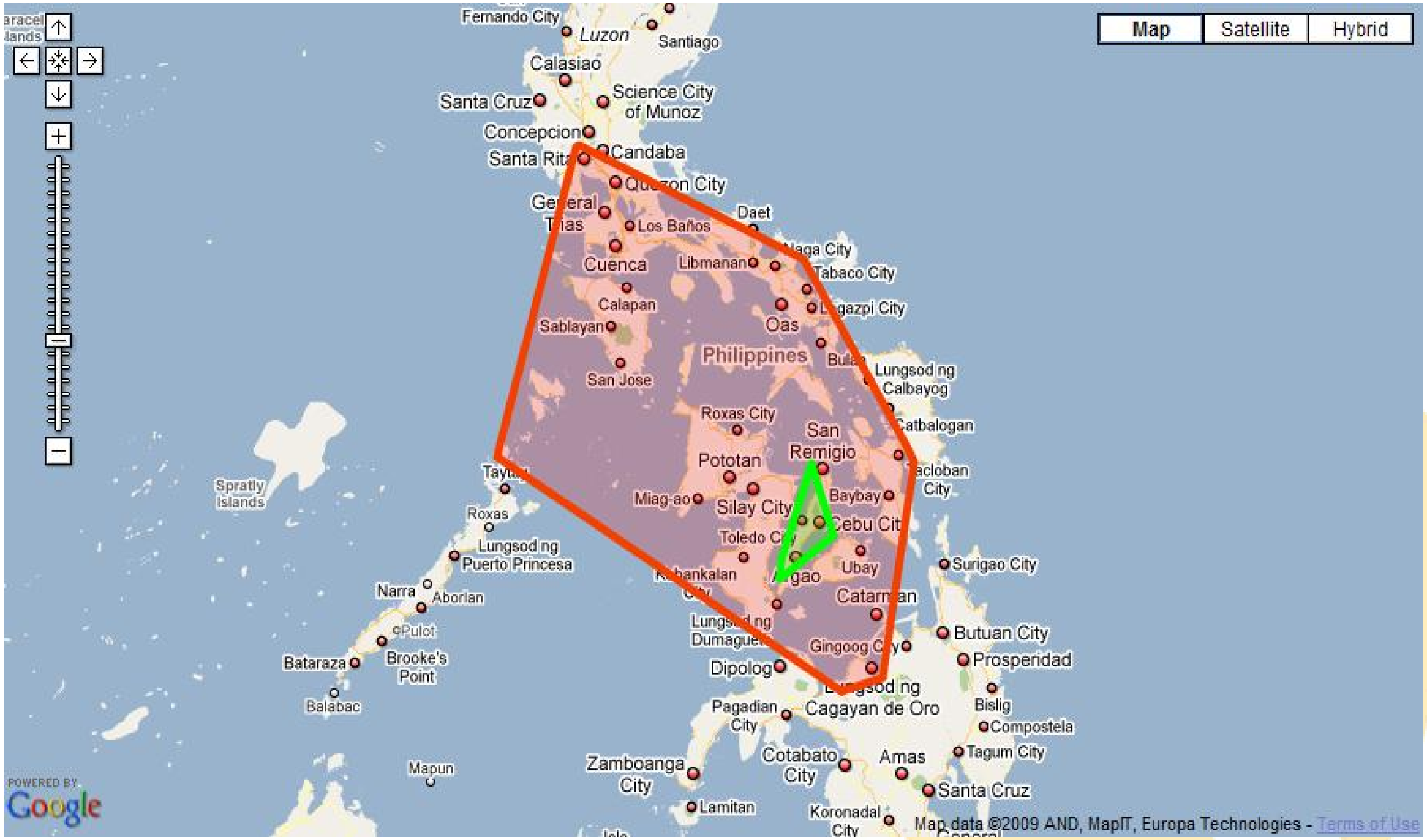
Geographic subsumption

USC

Representing Places

Multiple Convex Hulls

Learning relations between places



Geographic Subsumption

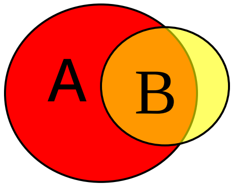
Based on probabilistic subsumption

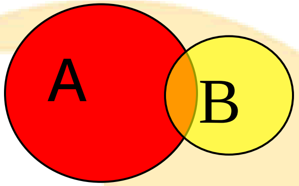
(Sanderson and Croft, 1999; Schmitz, 2006)

place *A* subsumes place *B* if

$p(A|B) >= t$ and $p(B|A) < t$

, where t is a predefined threshold ranging from 0 to 1





$p(A|B) = Area(A \cap B)/Area(B)$

$p(B|A) = Area(A \cap B)/Area(A)$

Probabilistic Subsumption on Tags

Probabilistic subsumption on tags (Schmitz, 2006)

place *A* subsumes place *B* if

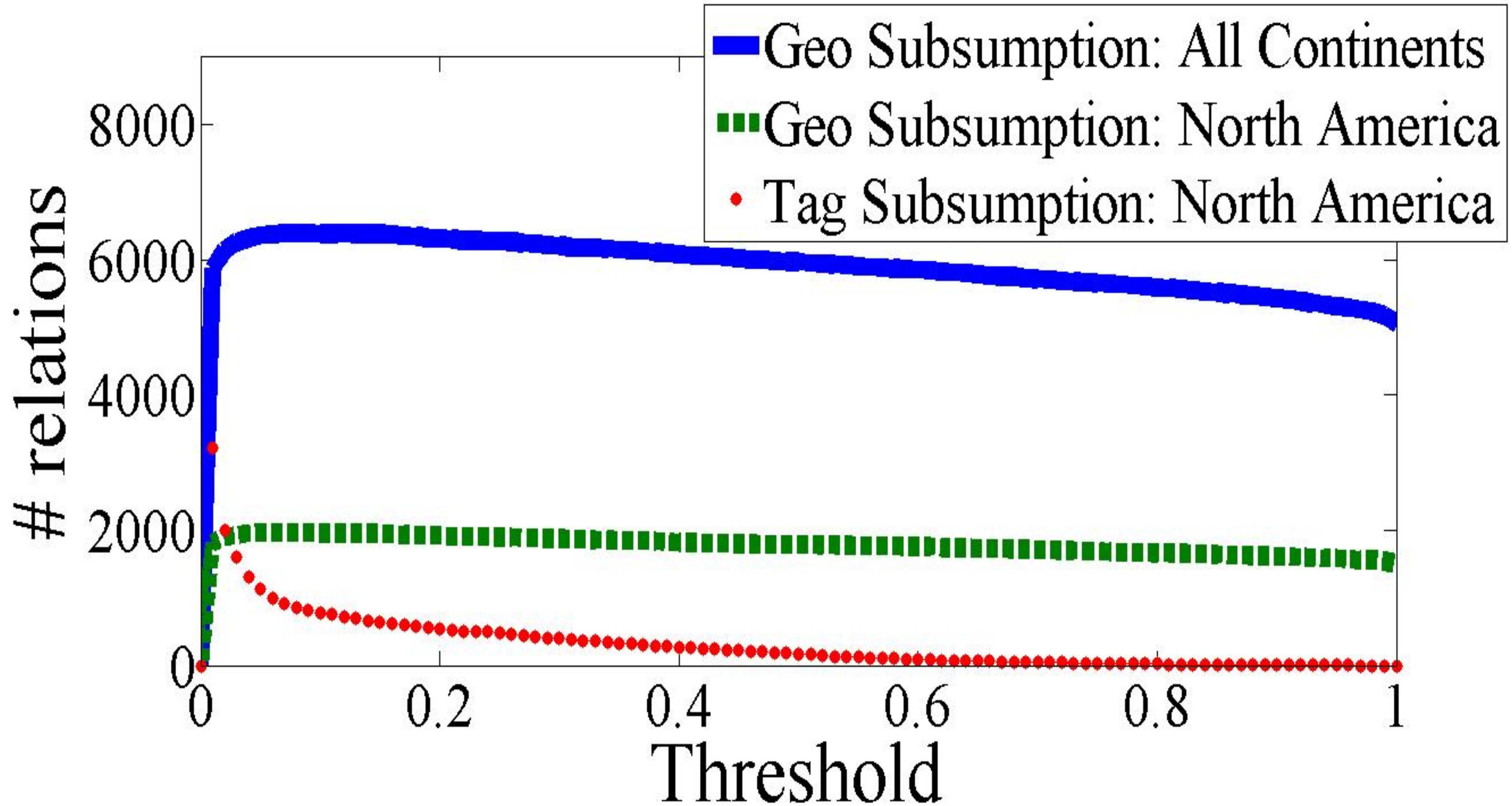
p(A|B) >= t and p(B|A) < t

p(A|B) = *Frequency*(*A*,*B*)/*Frequency*(*B*)

p(B|A) = *Frequency*(*A*,*B*)/*Frequency*(*A*)

***Frequency*(*A*,*B*): how many times tags A and B co-appear in the data set;**
***Frequency*(*A*): how many times tag A appears in the data set.**

Result: # of Induced Relations



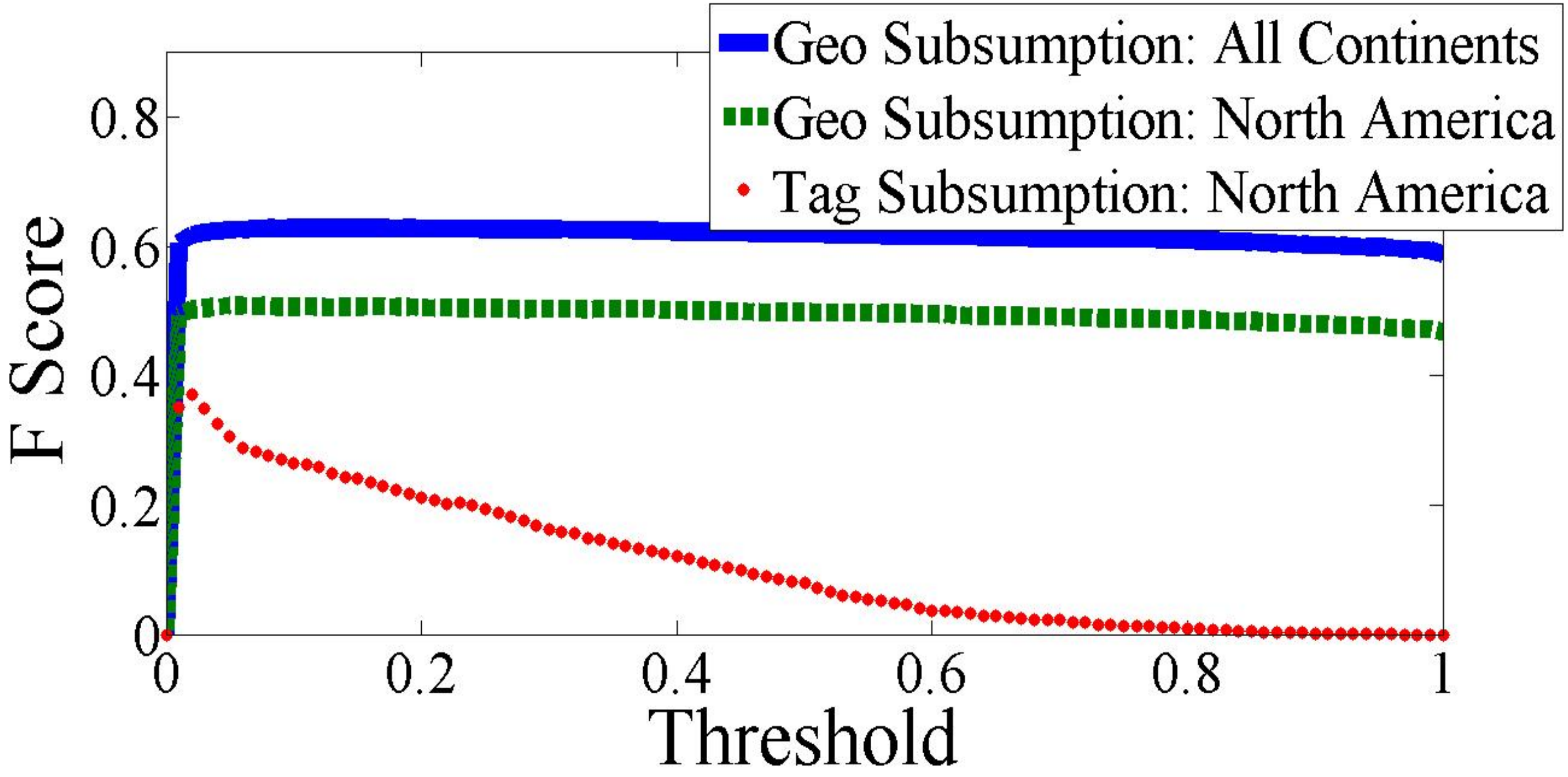
Ground Truth

Evaluate the quality of learned relations against the reference set from GeoNames.org

GeoNames

Result: Quality of the Learned Relations

F-score = (2*Precision*Recall)/(Precision + Recall)



Examples of Novel Relations

Child	Parent		Child	Parent
anaheim	la		disneyland resort	disneyland
ballard	puget sound		disneyland	la
brandywine park	wilmington		golden gate bridge	san francisco bay
bronx	new york city		pearl harbor	oahu
bronx zoo	new york city		times square	new york city
coney island	new york city		university of washington	puget sound

Related Work

- Inducing conceptual hierarchies from tags' frequencies [Schmitz, 2006; Brooks & Montanez, 2006, Mika, 2007]
- Inducing conceptual hierarchies from collection – set relations [Plangprasopchok & Lerman, 2009]
- Identifying "place tags" from geopoints [Rattenbury & Naaman, 2009]

- Web-A-Where [Toni: please shortly mention it]
- Li et al. [Toni: please shortly mention it]
- Spirit tagger?

Conclusion

Our approach can:

- learn more high quality relations than the baseline
- learn novel relations that do not exist in expert-created knowledge bases
- leverage user generated knowledge to aid decision makers

USC Viterbi
School of Engineering

Thank you

USC