Data-Intensive Geospatial Semantics

Krzysztof Janowicz

University of California, Santa Barbara

Ontology Summit 2012: session-06 - Thu 2012.02.16

Univ. of California Santa Barbara GEOGRAPHY

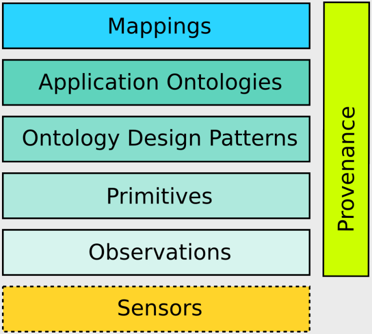
The Three V's of Big Geo-Data

- **Volume**: The **size** of the involved data, their **multi-dimensional** nature, as well as their **inter-linkage** which creates a global graph. E.g., Volunteered Geographic Information, Location-based Social Networks, sensor networks, high resolution remote sensing data, complex transportation simulations.

- **Variety**: The number of heterogeneous **sources** and **type** of data is increasing as well. Combining social media with authoritative sources and integrating different formats such as video, audio, photo, and text allows a more **holistic** analysis but raises new issues in data integration.

- **Velocity**: Big Data is not only about large amounts of data but also the **speed** at which data is **created** and **updated**. A rapidly increasing number of data sources deliver near real-time data which poses new challenges for stream reasoning and rule systems, which data do we **keep**?

ONTOLOGICAL FRAMEWORK TO SUPPORT **VARIETY**



- **Observation-driven** ontology engineering to foster semantic **heterogeneity**
- If ontologies are too lightweight to **restrict meaning**, we can still anchor them in **observations** (and thus provide provenance)
- May require a **symbol grounding** level for observation procedures

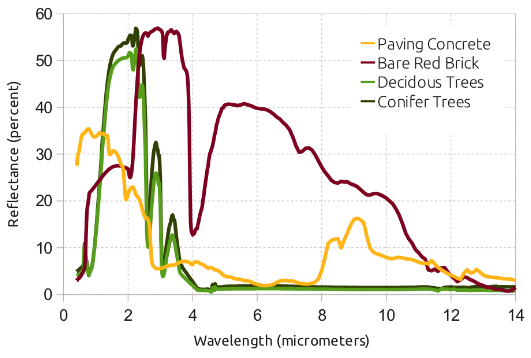
- Ontologies should be about **communication** and not about replacing **numerical models**; please **do not** try to develop an **universal ontology** for rivers, mountains, forests, and so forth

SAMPLING BIG GEO-DATA

Shopping
Specialty
Retail
Hotels
Travel
Bakeries
Beauty
Airports
Restaurants
Music
Salad
Food
Sports
Industry
Stores
Performance
Seafood
Automotive
Convenience
Movie
Electronics
Delis
Accessories
Books
Snacks
Foods
American
Coffee
Health
Groceries
Italian
Outdoors
Subs
Live
Video
Pizza
Magazines
Pub
Home
Farm
Museums
Clubs
Bars
Sandwiches
Furnishings
Steak
Improvement
Entertainment
Drink
Manufacturing
Services
Galleries
Department
Computers
Apparel
Mexican
Theaters
Shops
Recreation
Hamburgers
Cafes

- **Spatial data**: 20,765/3,247,409 POI and 64/71 types (**OpenStreetMap**)
- **Temporal data**: 440,939 check-ins by 35,745 users to 150,300 POI of 408 different types (**Whrrl**); from more than 3 million check-ins per day
- **Thematic data**: 218,760 geo-referenced Wikipedia articles and 287,210 geo-referenced travel blog entries.

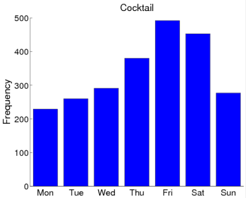
Introducing Semantic Signatures

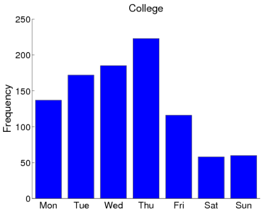


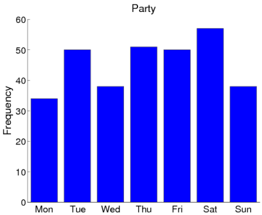
- Combine **numerical** (statistical) **models** and data with ontologies to derive **local** (personal) **primitives** (reifications)
- **Analogy** to **Spectral Signatures** used in Remote Sensing
- Multiple spectral **bands** → multiple **semantic bands**

Semantic Signatures – Temporal Bands

- **When** you are is **what** you are

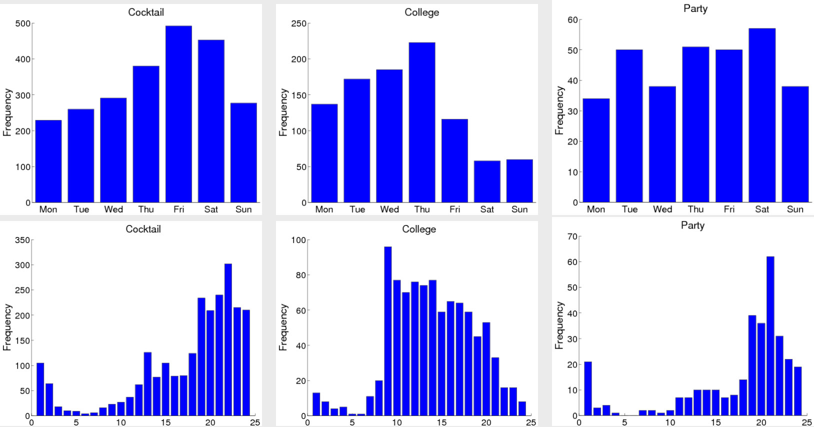






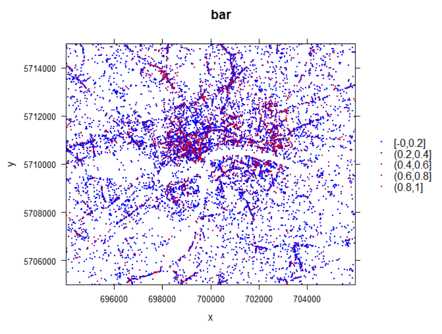
- Locations **types** and **log-in patterns** from the Location-based Social Network *Whrrl*
- **Day-band** from Semantic Signatures
- **Local Reifications (Primitives)**: e.g., **Weekend** vs. **Workday**
- We used them to **automatically compute** missing **types** of Whrrl POIs

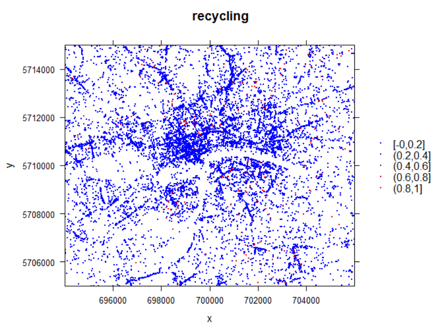
Semantic Signatures – Day and Hour-based Bands



- **Multiple semantic bands** may be required to distinguish between feature types. We can add a **hour-band** in addition to the **day-band**

Semantic Signatures – Spatial-Semantic Band

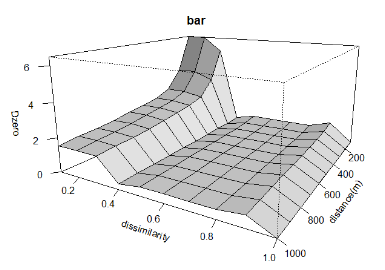


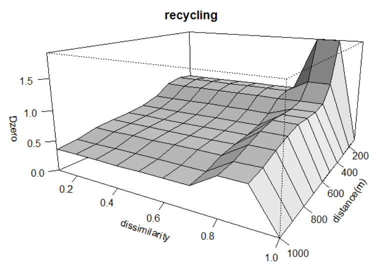


- POIs plotted by **similarity** to bar and recycling in OSM data, London, UK
- **Local Reifications (Primitives)**: e.g., **Uniform** and **Clumped**
- **Bars** (and similar features) tend to **clump** together
- **Recycling** (and similar features) are rather **uniformly** distributed

Semantic Signatures – Spatial-Semantic Band

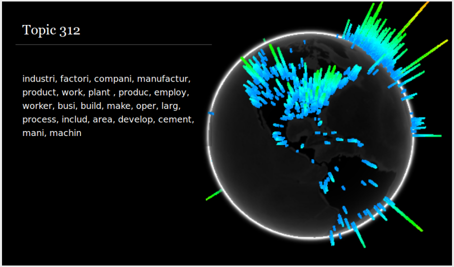
- **Where** you are is **what** you are





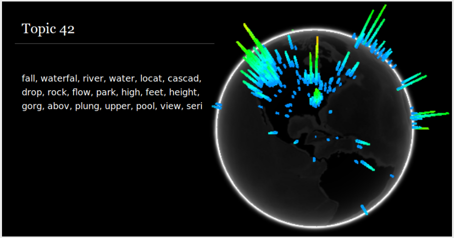
- D_{zero} **measures** the **likeliness** of features of a certain **type** to co-occur within a specific **semantic and spatial range**.
- User support: generate **recommendations**, and **clean up** data based on **type likelihood**. 'How likely is a recycling center directly next to an existing one?'

Semantic Signatures – Thematic Band



Topic 312

industri, factori, compani, manufactur, product, work, plant , produc, employ, worker, busi, build, make, oper, larg, process, includ, area, develop, cement, mani, machin

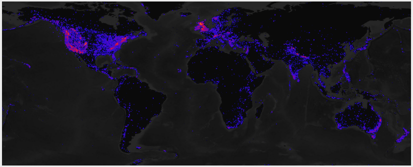


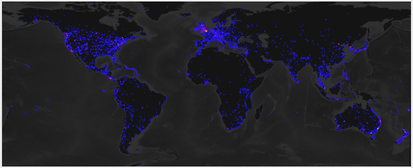
Topic 42

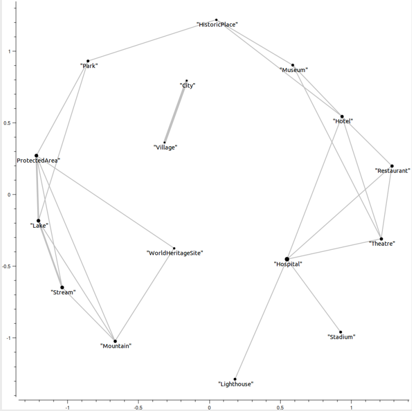
fall, waterfal, river, water, locat, cascad, drop, rock, flow, park, high, feet, height, gorg, abov, plung, upper, pool, view, seri

- A **thematic band** can be computed out of unstructured text using latent Dirichlet allocation (LDA); data source Wikipedia and Travelblogs
- Non-georeferenced plain text is often still **geo-indicative**
- Different **types** (taken from **DBPedia**) of geographic features have different, **diagnostic** topics associated to them (out of **500** topics)

Semantic Signatures – Thematic Band

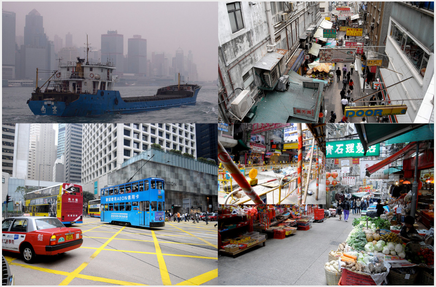


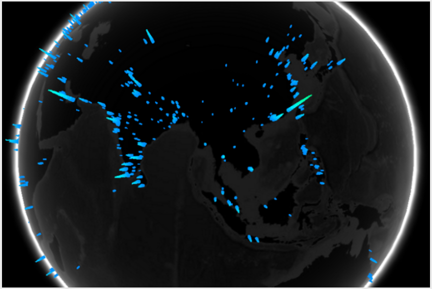




- **City** topics: **204**>**450**>**104**>282>267>497>443>484>277>97>**...**
  Town topics: 425>**450**>419>367>**104**>429>266>69>**204**>308>**...**
  Mountain topics: **27**>110>5>172>208>459>232>398>453>183>**...**

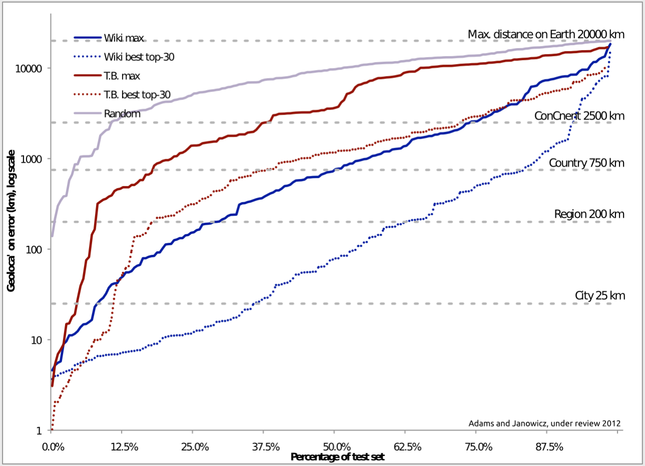
DiaLoc – New Data Sources For Old Questions





- Use **plain text**, not low-level image analysis as data source to estimate geographic locations from images
- *'market food street narrow dense populated asia economy air conditioning smog fog humid warm building construction skyscrapers skyline shipping export channel harbor transportation tram city advertisement'*

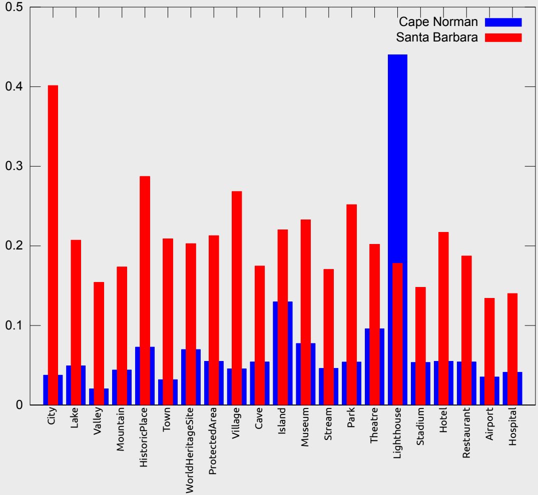
DiaLoc – Location Estimation



Adams and Janowicz, under review 2012

- In about **70%** of all queries, DiaLoc excludes **99.9%** of the **land-surface** of the Earth.

DiaLoc – Geographic Feature Types



- Semantic signatures used to infer which **type of feature** is described.

Summary and Lessons Learned

- **Big** (Geo)-Data requires small, local theories (**microtheories**)
- Develop geo-ontology design **patterns**, not domain ontologies
- Developed patterns in a **community** process (GeoVoCamp)
- Mine and learn ontological **primitives** from **observation** data
- **Semantic signatures** as one methodology to learn primitives
- Make the domain expert the **knowledge engineer**
- **Late** assignment of classes (e.g., *River*)
- Ontology **alignment** and mapping to connect **local theories**

- What is the **80-20** rule of geospatial semantics?
- **Cross-media validation** using different semantic bands