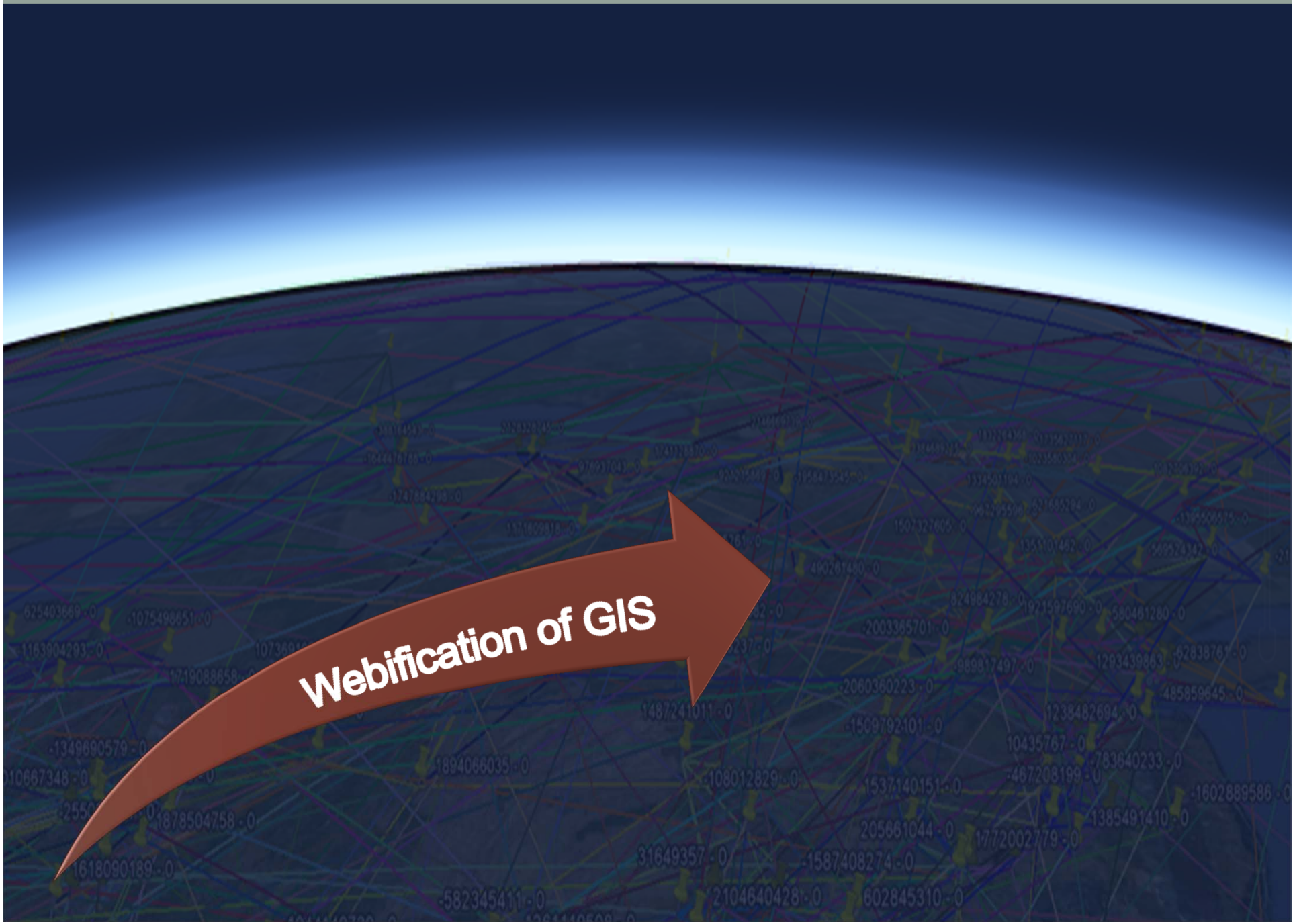
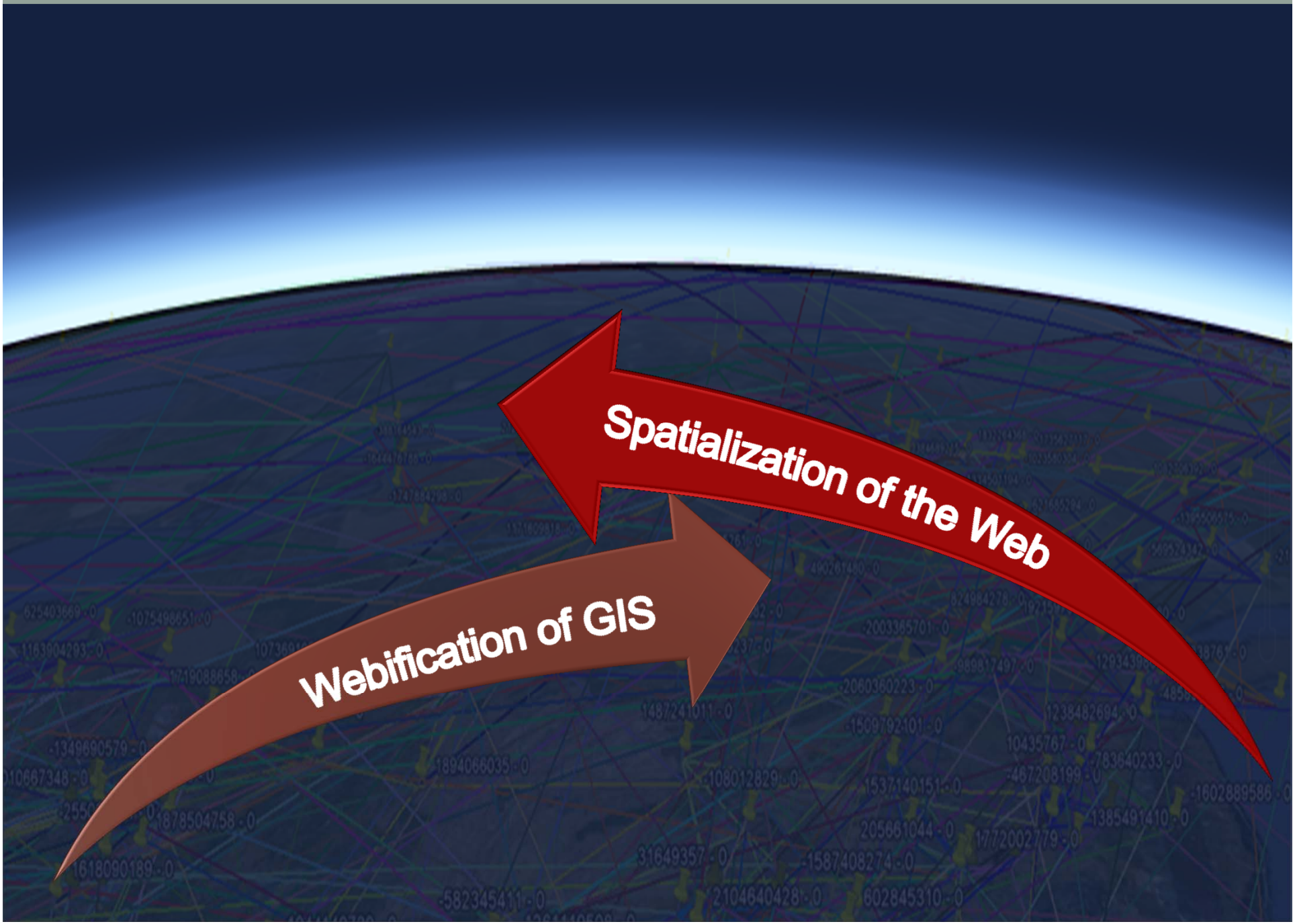
# The Spatialization of the Web

Grant McKenzie
Florian Straub
Ben Adams


Webification of GIS


Spatialization of the Web
Webification of GIS

# The «modern» Web…

# The «modern» Web…

- …a Web of Prosumers

- The change to Web 2.0 is a *change towards a Web of participation*, where an ever increasing number of users simultaneously acts as producer and consumer of content.

- In 2015 more then 70 % of the WWW`s content is going to be user generated!

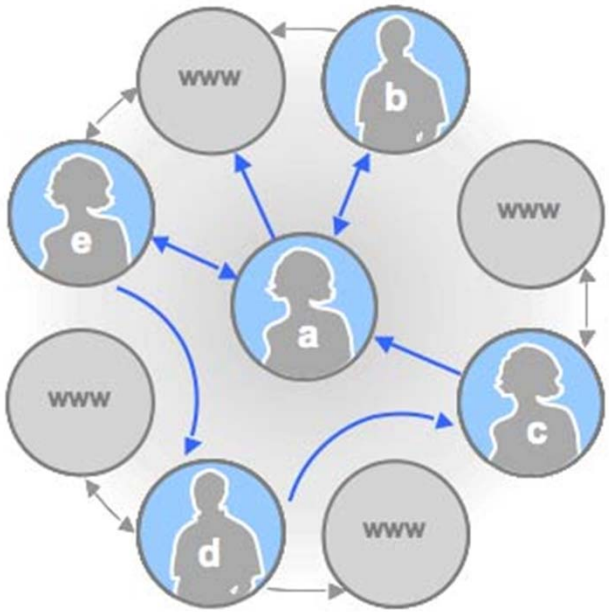
# The «modern» Web…

## …a social Web

The *change from* 1:n publication of “cold” (often factual) information with a globally comprehensible semantics for an unspecified audience

*to* personal n:m communication of “warm” personal content in Social Networking and the Social Web with a more restricted contextual scope and validity.

personal queries for information are often directed to friends in SN platforms instead of querying centralized anonymous search engines.

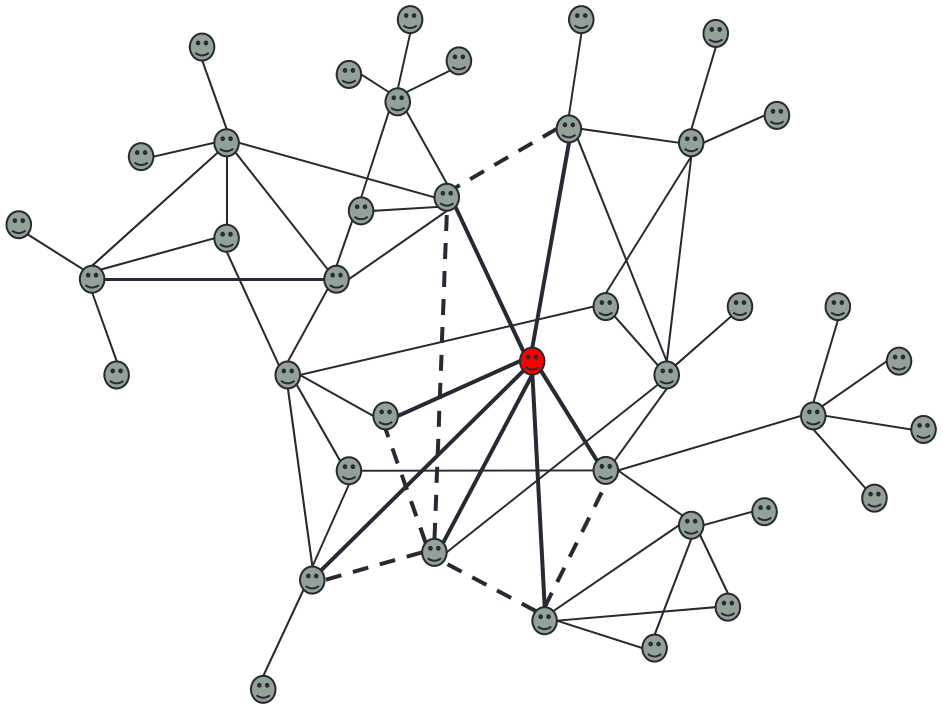


# The «modern» Web…
## The Small World Phenomenon

„*six degrees of separation*“ in the human social network
[Milgram1967]

**Structural characteristics:**
type of graph that is characterized by characteristically **large clustering coefficient** and a **small diameter**

**Algorithmic characteristic:**
Routing without global knowledge of the network but with local knowledge only is possible.
→ **decentralized Search**



**Can be found in:**

- human social network
- US Power-Grid
- Neural Network
- link-graph of the Web

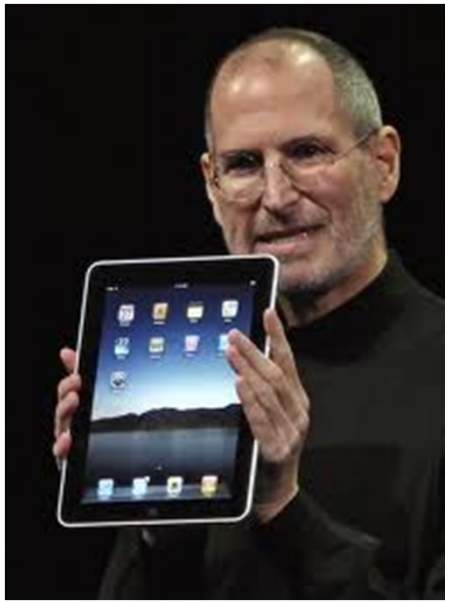
# The Small World Phenomenon

http://oracleofbacon.org/

# The «modern» Web…

## …a mobile and ubiquitous Web

New classes of mobile devices like smartphones or net-books allow for an *ever deeper and smoother integration of the Web into our everyday lives* and blur the borderlines between virtual and real world.

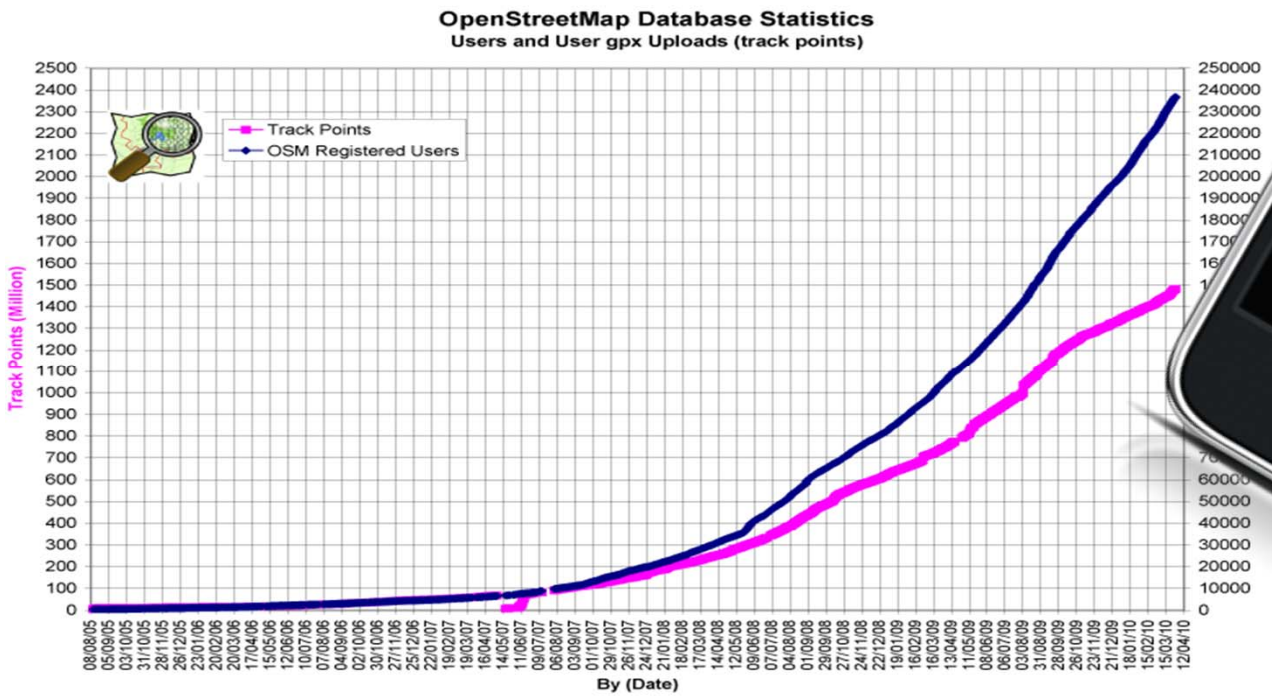


# The «modern» Web…

…a geospatial Web

Geo-enabled Devices (e.g. smartphones with GPS)

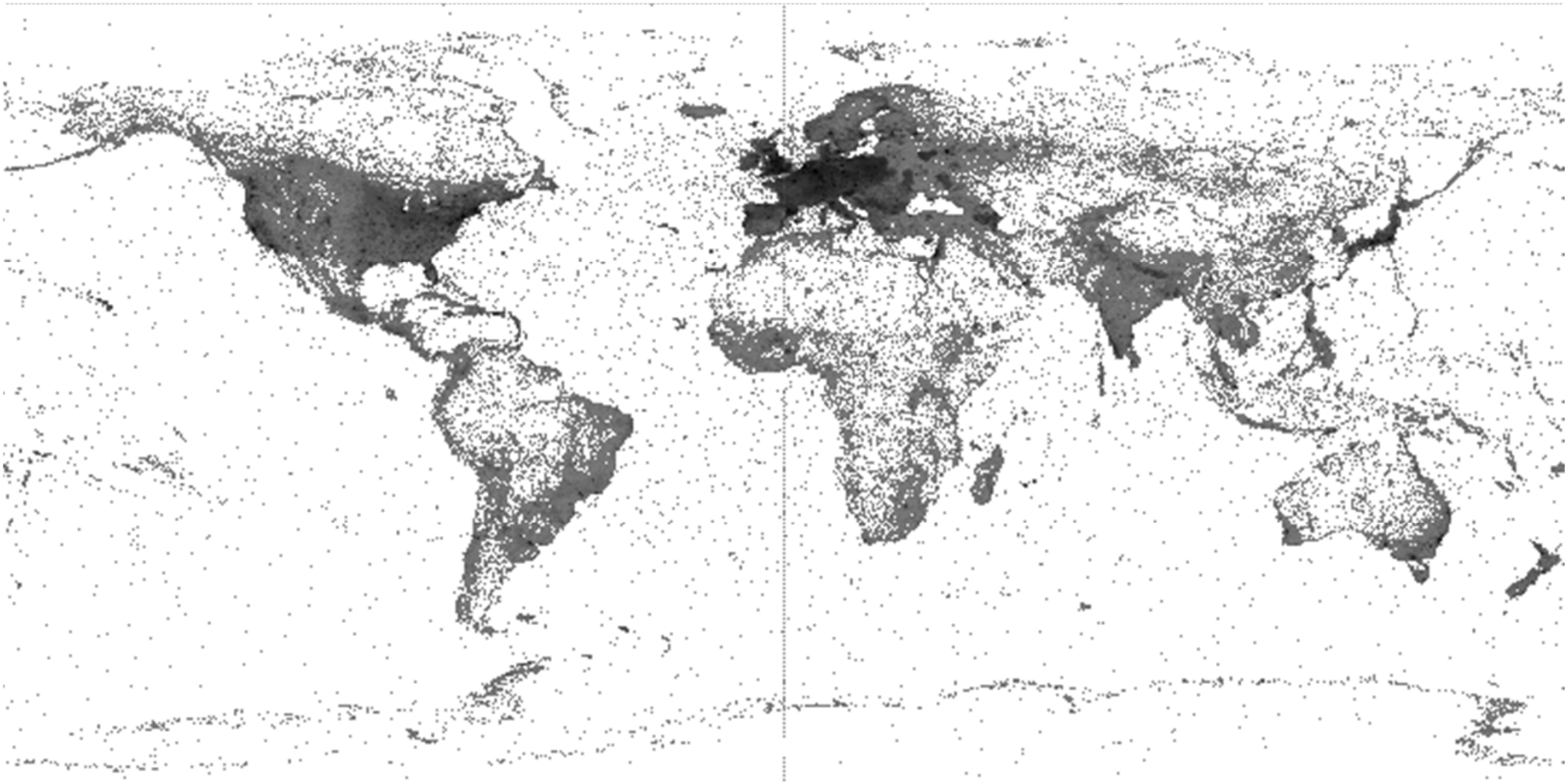
User Generated Geo-Content and Collaborative Geo-Applications





*[Image sources: wiki.openstreetmap.org and apple.com (modified)]*

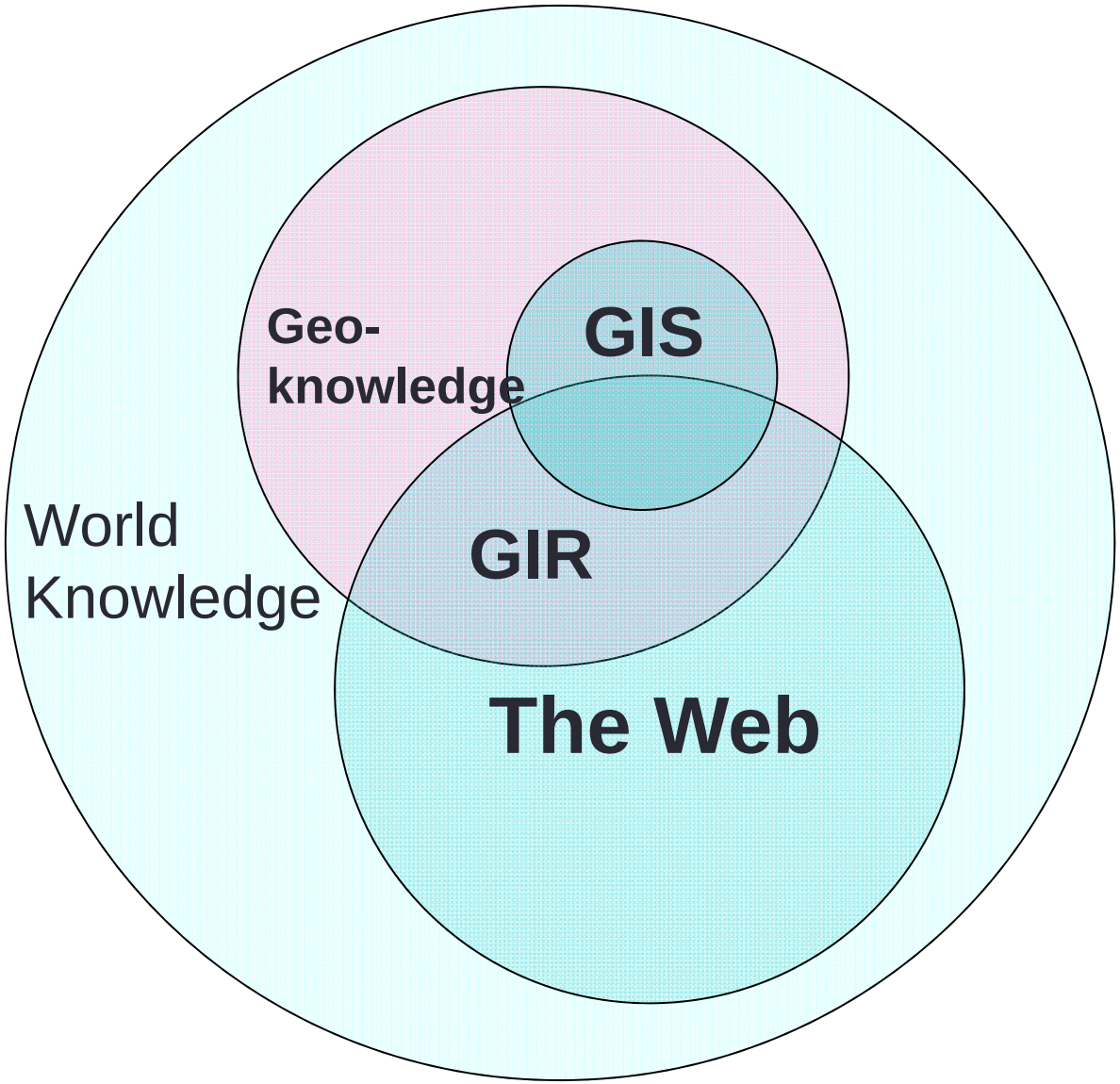
# Spatial distribution of Wikipedia articles



# Spatial distribution of Facebook`s 800 million people and there social graph

# (Geographic) Information Retrieval

# (Geographic) Information Retrieval



# (Geographic) Information Retrieval

**GIS** is concerned with exact spatial representations and complex analysis at the level of the individual spatial object or field.
→ Users are experts, information is structured and unambiguous!

**GIR** is concerned with retrieving geo-referenced information resources that **may be relevant** to a geographic query region.
→ Unstructured and ambiguous information, everyday applications!

*Similar to the difference between search engines and relational database systems!*

# (Geographic) Information Retrieval

## Current motivation of GIR:

about 80 % of all decisions of an individual are related to a direct or indirect spatial reference

about 20 % of all Web-searches (not distinguishing between mobile and desktop access) have a geographic context.

Find geo-specific resources on the Web

*<topic keywords, spatial relation, location>*

Something ***related_to*** **Somewhere**

*related_to* = in, near, within x km, north_of ..etc.

# Natural IR Principles

**Cognitive process of searching for information:**

1. asking yourself (WHAT) → formulate your "information need"

2. (WHO) to ask for this information
   ←→ (WHAT) the topic
   ←→ (WHERE) the geographic context
   ←→ (WHEN) the temporal context

3. (WHERE) the aforementioned source (WHO) is

**WHO←→WHERE←→WHAT←→WHEN**

In the real world, we discover that people can successfully find relevant information for questions by just asking the "right" people through their social network.

# (Geographic) Information Retrieval

**Geo-Tagging** = Geo-Parsing + Geo-Coding

**Geo-Parsing**

Geographic context embedded in natural language descriptions

Recognising geographic references with Named Entity Recognition (NER) from Natural Language Processing (NLP).

**Geo-Coding** (grounding)

Attaching a unique quantitative location (footprint) to Named Entities

# Geo-Parsing

Place names can be recognised with gazetteers

Some types of true geographic reference:

- the name of the place
- an address (address fragment)
- a postcode / zip code
- a phone number

Some types of false geographic reference:

- Personal names (e.g. Jack London)
- Business name (e.g. York Hotel)
- Street names (e.g. Oxford Street)
- Common words that are also places

→ distinguishing between false and true geo-references

# Geo-Coding

Geo-coding (grounding) the genuine geo-references

From a Named Entity to a footprint

(e.g. with a gazetteer)

NE(“Santa Barbara”)

# Existing Gazetteers

- Alexandria Digital Library (ADL) Gazetteer
  - ~6 million entries
  - Has tried to standardize the format, description, and distribution of gazetteer data.
  - Has a published, detailed schema.
  - Basis for OGC standard.
- GeoNames website
  - Integrates information from multiple sources.
  - Publishes OWL ontology.
  - ~6 million entries
- Google / Bing / Yahoo…
  - All have proprietary gazetteers available through public APIs

# Spatio-temporal Indexing

Typical strategy is to have separate indexes.

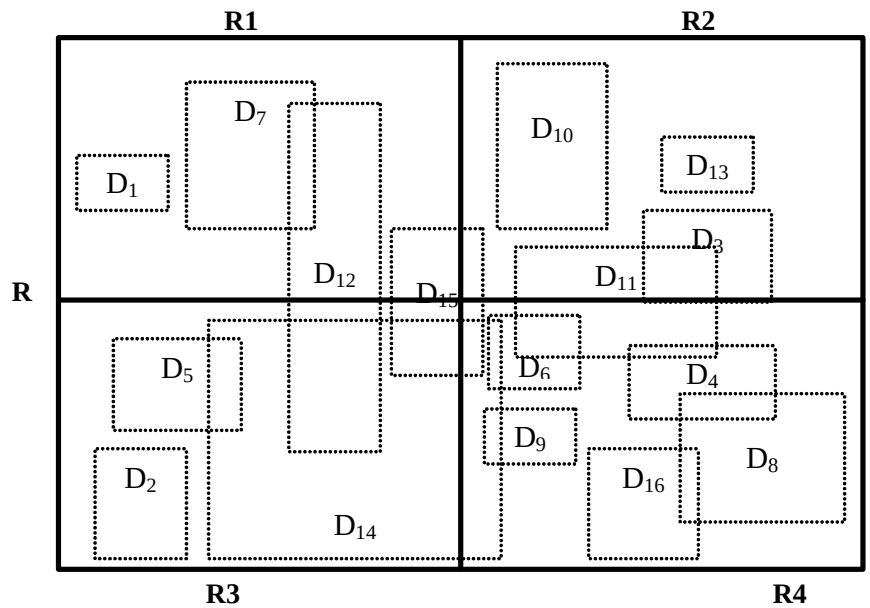
- Inverted index for text.
- R-tree for footprints.

→ Access spatial index with query footprint.

→ Access text index with query terms.

→ Merge results and find the intersection.

| | |
|---|---|
| **Term1** | D1, D2, D23, … |
| Term2 | D9, D11, D100, … |
| Term3 | D27, D85, .. |



# Tobler`s First Law of Geography

*"All things are related,*
*but nearby things are more related than distant things"*

→ "nearby things are more similar than distant things"

[W.R.Tobler, 1970. *A computer movie simulating urban growth in the Detroit region.* Economic Geography46]

## Formalization in Geostatistics

–variogram, covariogram
–measuring how similarity decreases with distance
–spatial autocorrelation
–IDW and Kriging

## The First Law of Cognitive Geography

"People think closer things are more similar" [Montello and Fabrikant]

# Spatio-temporal Ranking

Spatial similarity can indicate relevance

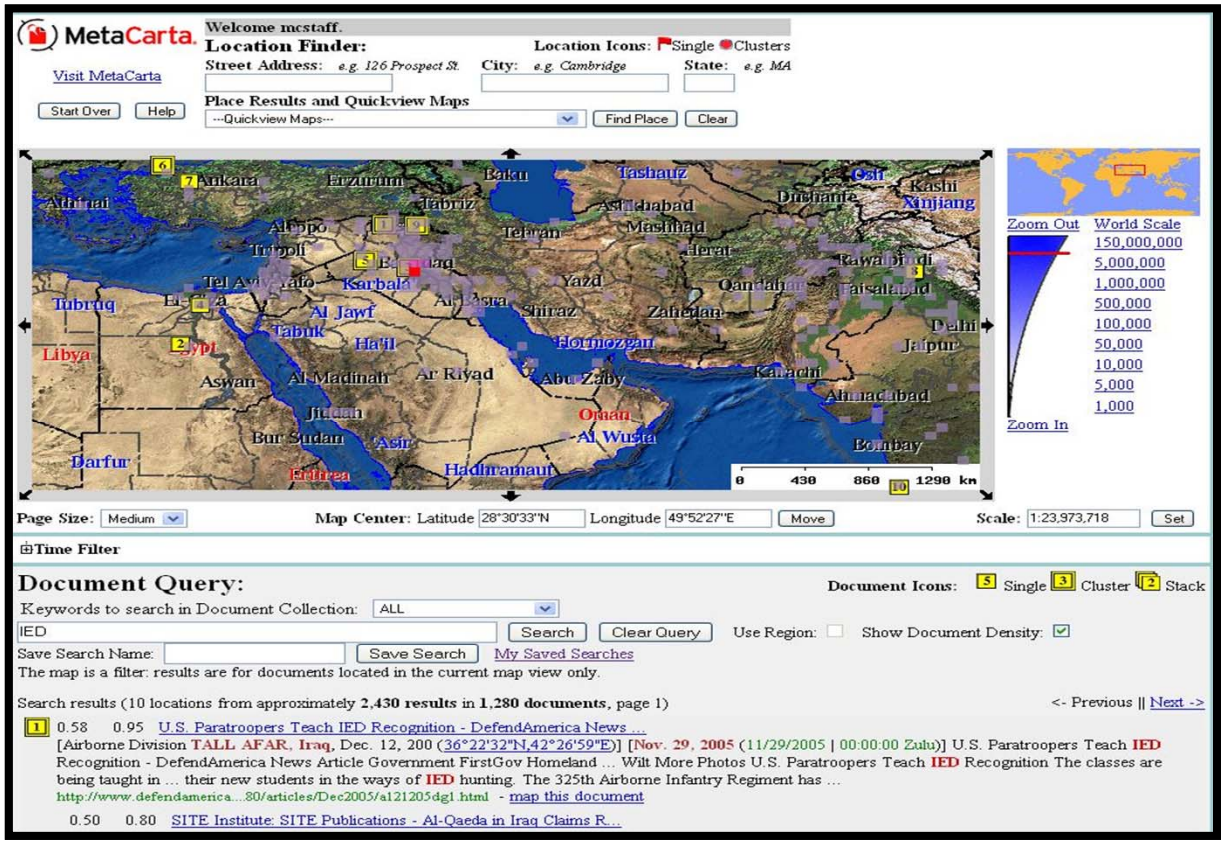
- Documents whose spatial content is more similar to the spatial content of the query should appear first.

But we need to consider both, the

- Thematic relevance
- Geographic relevance
  Geometric (e.g. distance) and non-geometric (e.g. topology)

State of the art consists of doing a linear combination.

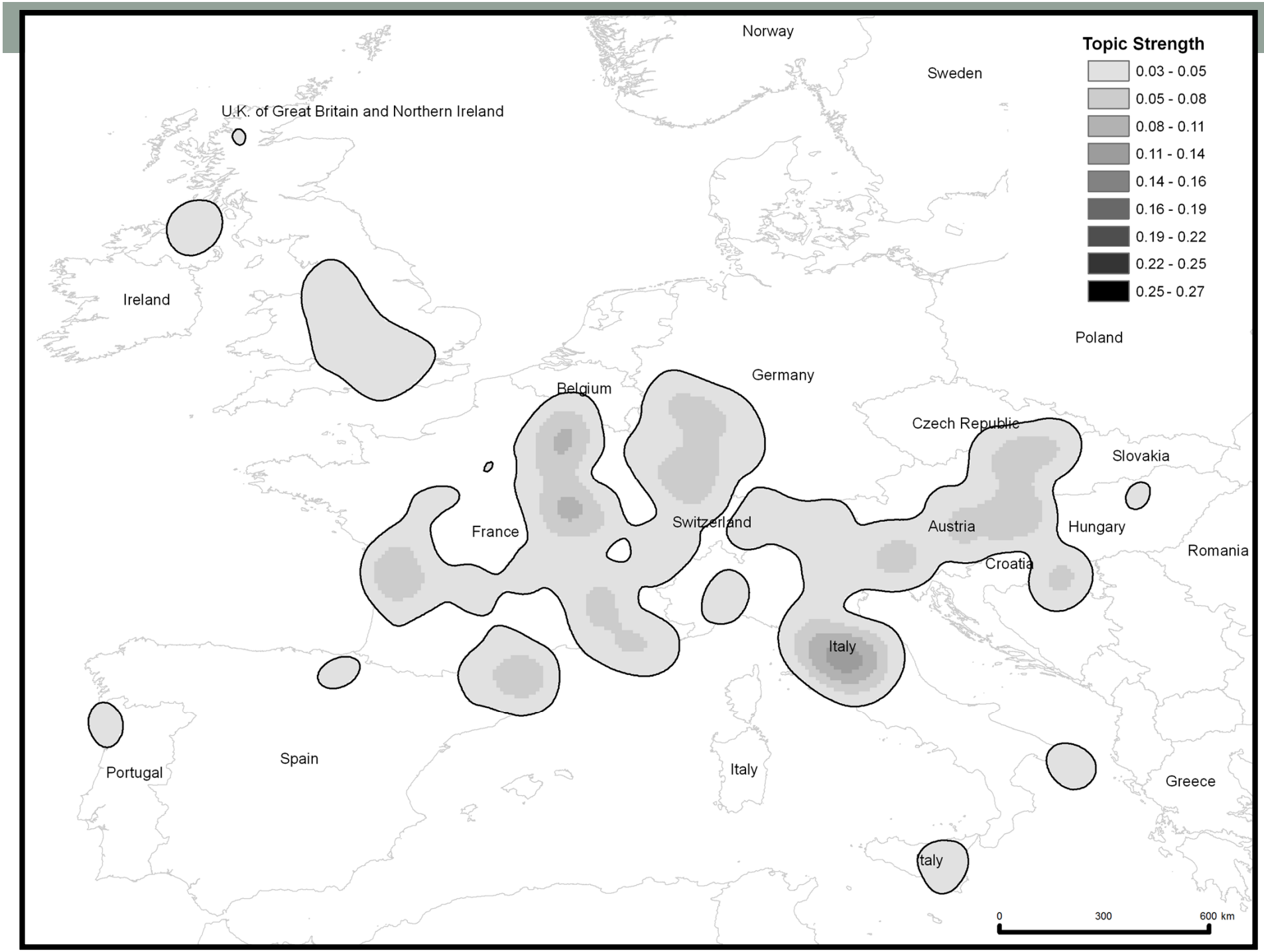
- The MetaCarta system
  - Pioneer system addressing all aspects given in this talk.
  - Conducts geo-parsing and geo-coding of text documents, and sends back possible location references with relative strength scores.
  - Uses NLP to find possible location references.
  - Contains a gazetteer of ~14 million entries.

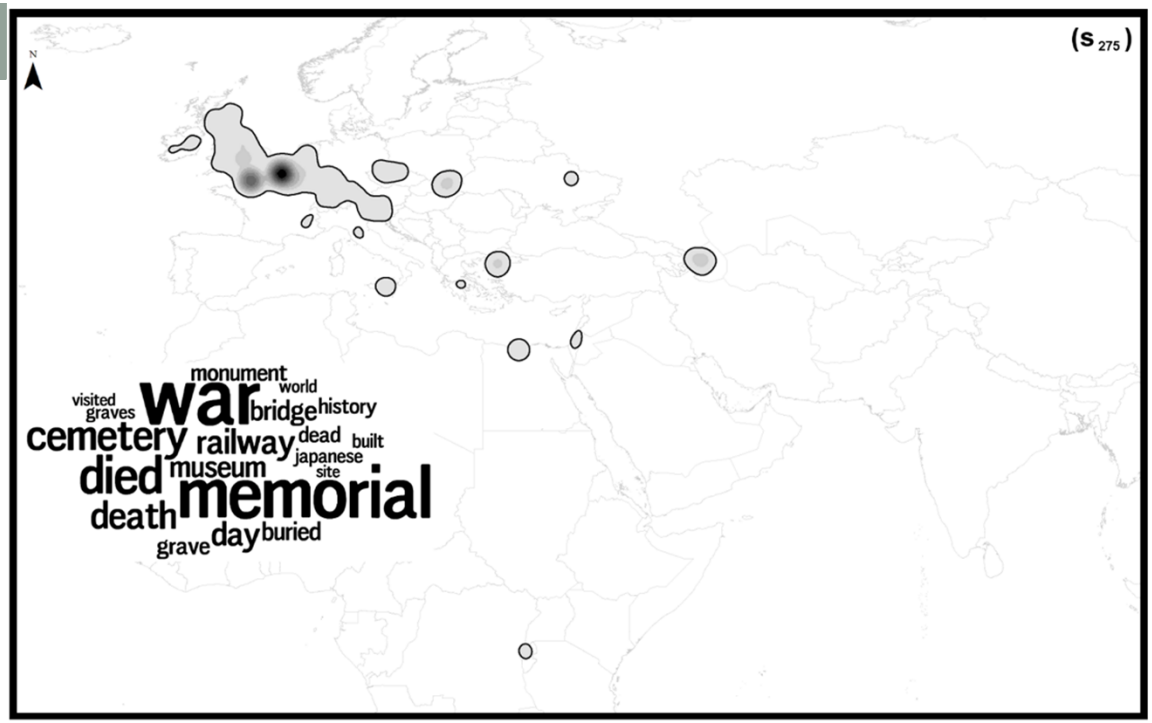


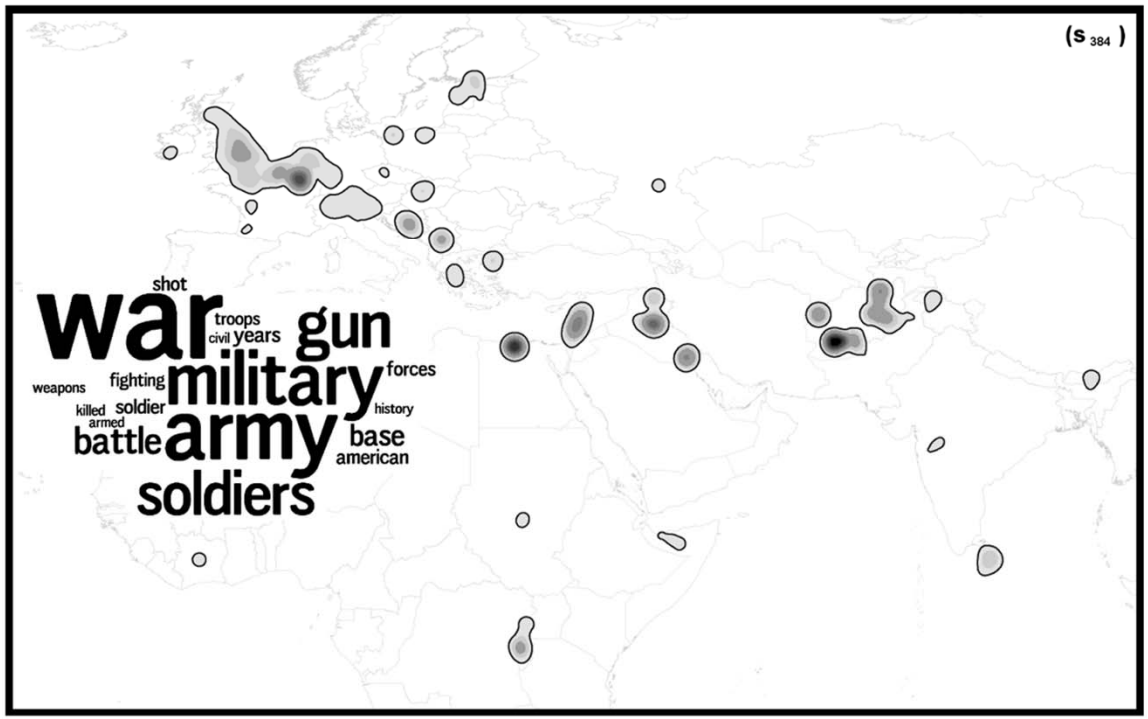
"… identify the themes that are most closely associated with places around the world."

- ~ 275,000 Travelblogs
  - Each blog tagged to one specific location
  - Approached each location as a corpus
- Divided the world up into ¼ degree grid cells
  - Approx. 10,000 locations
- Latent Dirichlet allocation
  - 200 Topics


hike
top
climb
trail
mountain
rock
steep
desert
dune
sand
star
camel
festival
dance
people
celebration
parade
crowd
london
ben
england
pub
walk
incredible
absolutely
lovely
beautiful
amazing
night
camp
tent
campsite
river
bridge
water
crossing
shop
market
sell
buy
street
hong
kong
bread
cheese
breakfast
food
potato
eat
wine
bottle
taste
glass
tree
garden
forest
flower
plant
green
sea
whale
dolphin
penguin
seal
island
bird
watch
water
peru
lima
arequipa
local
bus
canyon
monday
weekend
saturday
friday
sunday
week
thursday
wednesday
tuesday
morning
night
walk
beach
water
room
house
kitchen
apartment
live
door
floor
bed
bathroom
bedroom
window
japan
japanese
tokyo
police
prison
guard
people
security
happen
story
danger


Topic Strength
0.03 - 0.05
0.05 - 0.08
0.08 - 0.11
0.11 - 0.14
0.14 - 0.16
0.16 - 0.19
0.19 - 0.22
0.22 - 0.25
0.25 - 0.27
Norway
Sweden
U.K. of Great Britain and Northern Ireland
Ireland
Poland
Germany
Belgium
Czech Republic
Slovakia
France
Switzerland
Austria
Hungary
Romania
Croatia
Italy
Spain
Portugal
Italy
Greece
Italy
0
300
600 km


(s 275 )
N
monument
world
visited
graves
war
bridge
history
cemetery
railway
dead
built
japanese
died
museum
site
memorial
death
grave
day
buried


(s 384 )
shot
war
troops
civil
years
gun
military
forces
weapons
fighting
history
killed
soldier
armed
battle
army
base
american
soldiers


Topic Dominance
Both Human & Physical Characteristics
Physical Characteristics
Human Characteristics
0 175 350 km

# Current Challenges in Geographic IR

## Improve “conventional GIR” components and methods

- Geo-tagging, spatio-textual indexing and geo-relevance ranking.
- Improved understanding of spatial natural language terminology.
- Better user interfaces for exploration of GIR results.

## Integration of geographical with temporal aspects

- Everything we do happens in space and time!

## Creation of place ontologies with world-wide coverage

- e.g. Fuzzy regions and intra-urban placenames

# What this means to you…

- You are the next generation of “spatial thinkers” and GIS experts.
- Understand not only how GIS has been influenced by the modern web, but also how the modern web has been influenced by GIS.
- Huge advancements in the area of Information retrieval. Spatial science has a big role to play
- Realize that GIS is not separate from your everyday lives.