INDIANA UNIVERSITY

Beyond Co-occurrence:
Discovering and Visualizing Tag Relationships from Geo-spatial and Temporal Similarities

Haipeng Zhang, Mohammed Korayem, Erkang You
and David Crandall
School of Informatics and Computing,
Indiana University

Online Photo Sharing and Tagging

- More than 5 billion photos on Flickr
- Meta data: taken time, owner, upload time…
- Text tags -> describe, organize and share photos
- Camera/mobile phone with GPS -> geo location of photo

Taken time: 2007.8.17
Text tags: {snow zoo leopard potterparkzoo}
Geo location: 42.7179 -84.529

- Study tag relationships to extract knowledge and build services (tag recommender systems, search engines)

Flickr Tag Attributes and Our Intuition

- Much previous research on **tag relationships** was based on **tag co-occurrences**
- Other than co-occurrences, **geo and temporal patterns of tags** might also help measure tag similarities
- Reveal **tag semantics** based on geo/temporal similarities by **clustering tags and visualizing clusters**
- Give a sense **why** tags are similar

Related Work

- Clustering tags based on co-occurrences
  - Tag suggestion: [Garg08] [Sigurbjörnsson08] [Liu09]
  - Tag clustering: [Shepitsen08] [Begelman06]
- Temporal and geo-spatial properties of tags
  - Burst detection, finding place/event tags: [Rattenbury07] [Moxley09]
  - Cluster photos based on geotags and find representative text tags: [Crandall09] [Kennedy07]
- Visualizing tag clusters
  - Tag cloud: [Kaser07], tag evolving over time through animations: [Dubinko07]
- Spatial clustering and co-location pattern mining
  - Spatial clustering: [Ng94], co-location pattern mining: [Xiao08] [Huang06]
- **Studies of query logs, tweets and news articles**
  - **Temporal patterns of words in news articles, word semantics**: [Radinsky11]
  - Temporal patterns in search logs: [Vlachos04] [Chien05]
  - Geo patterns in search logs: [Backstrom08]
  - **Geo and temporal patterns in search logs, similar queries**: [Mohebbi11]
  - **Temporal patterns in tweets and news articles, dynamics of attentions**: [Yang11]

Baseline Tag Similarity Measures Based on Co-occurrences

- Raw **tag co-occurrences** on photos

Tag A	Tag B	co_occur(A,B)
newyorkcity	nyc	228173
newyorkcity	brooklyn	38378
indiana	university	10824

- **Mutual information** between tag A and tag B, based on co-occurrences [Begelman06]

$$I(A,B) = \log\left(\frac{p(A,B)}{p(A)p(B)}\right)$$

Tag Similarity Measures Based on Geo and Temporal Tag Usage

- Extract **geo**/**temporal**/**motion** vectors from tag usage data to represent every tag
- Measure the **geo similarity** between two tags by the **squared Euclidean distance** between their corresponding **geo vectors**
- Compute the temporal and the motion similarities in a similar fashion

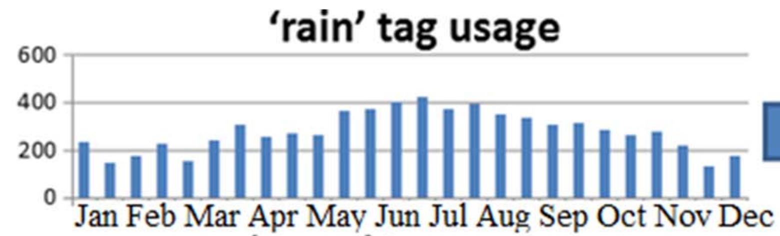
Data Set

- Metadata of a set of photos from North America, until the end of 2009, downloaded through Flickr API
- Over **30M geo-tagged** photos
- **Top 2000 tags** from this dataset (ranked by number of unique users)

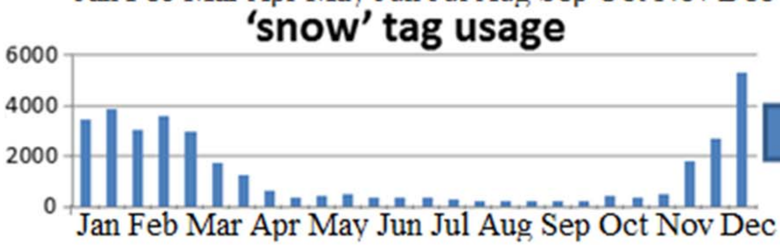
sunset	night	red	flower	river	
beach	snow	bridge	green	white	
water	blue	trees	nature	reflection	…
sky	clouds	lake	california	city	
tree	park	flowers	winter	newyork	

Extract Temporal Vectors

- Divide the usage data of a tag into *k* *i*-day periods (bins), ignoring the year; each period(bin) records # of unique users with the tag
- Form a *k*-D vector accordingly and normalize it



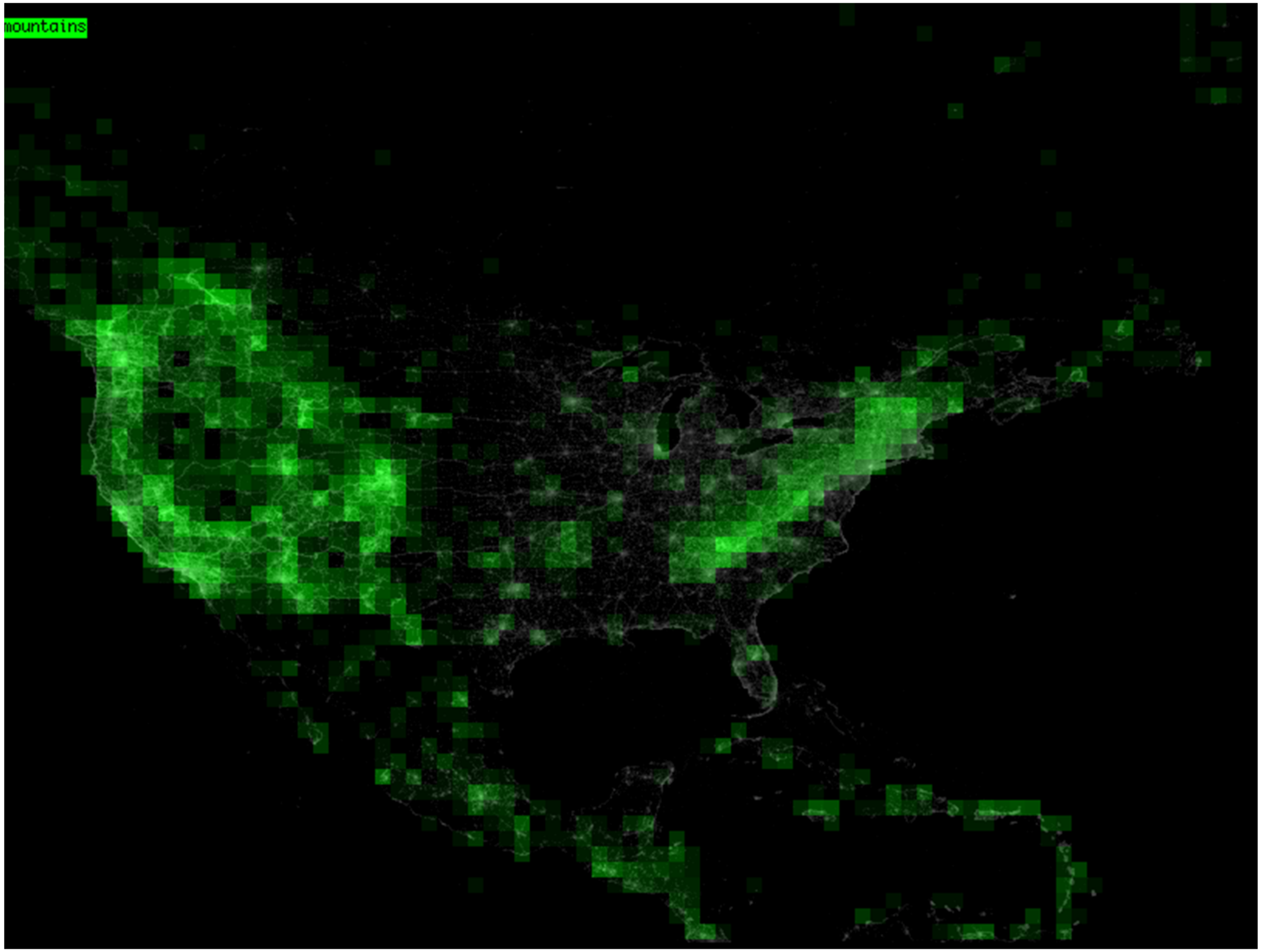
26D vector → [233, 147, ..., 175] → **Normalized 26D vector** [0.032,0.02 , ...,0.024]



26D vector → [3419, 3882, ..., 5304] → **Normalized 26D vector** [0.097, 0.11, ..., 0.148]

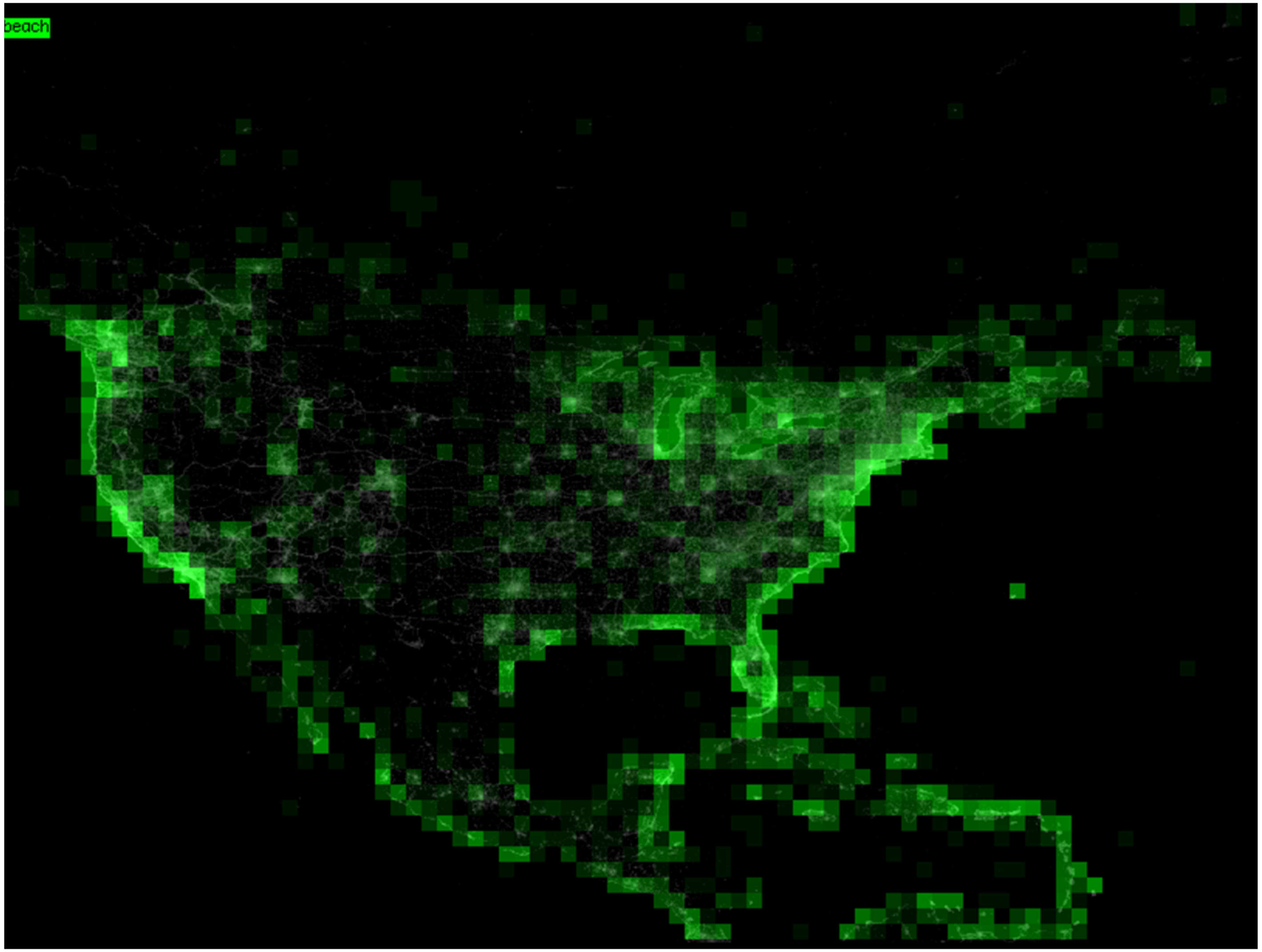
Extract Geo Vectors

- Heat map for the tag usage of '**mountains**'



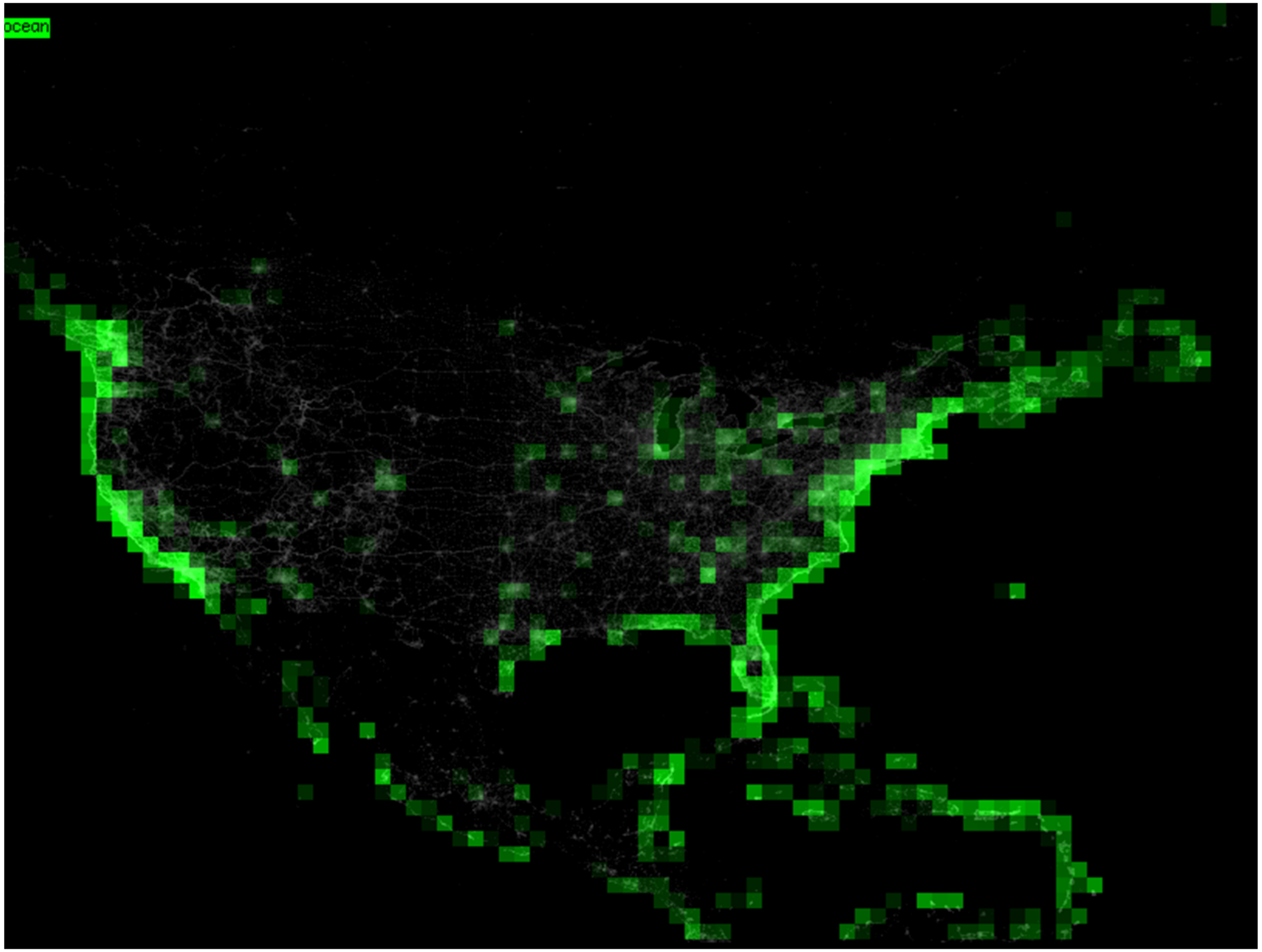
Extract Geo Vectors

- Heat map for the tag usage of ‘**beach**’



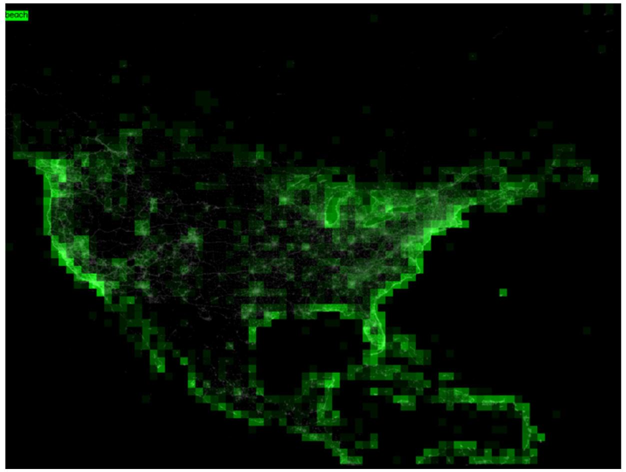
Extract Geo Vectors

- Heat map for the tag usage of '**ocean**'



Extract Geo Vectors

- Divide North America into *m*n* *g*-deg by *g*-deg geo bins
- In the *m*n* tag usage matrix, record the usage (# of unique users) of a particular tag in the corresponding geo bins
- Convert the matrix into an *m*n*-D vector and normalize it



60 by 80 tag usage matrix for tag 'beach', bin size 1-deg by 1-deg

4800-D usage vector

Extract Motion Vectors

- Extract motion vectors to capture the **movement of tags**, e.g. species migration
- Divide the data into k i-day periods
- For each i-day period, build an $m*n$-D geo vector
- Concatenate the k geo vectors into a $k*m*n$-D motion vector and normalize it

Clustering Tags and Ranking Clusters

- Cluster 2000 tags into 50 clusters, using 5 tag similarity measurements: **geo**, **temporal**, **motion**, **raw co-occurrences** and **mutual information** respectively
- Cluster **geo/temporal/motion** vectors using **k-means** [MacQueen67]
- Partition **raw co-occurrences** and **mutual information** tag graphs by **KMETIS** [Begelman06][Karypis96]
- **Rank geo, temporal** and **motion clusters** by average **second moment**, which measures the **peakiness** of their distributions

  a vector v's peakiness: second_moment(v)=$v \cdot v$
- Sampling twice from a dist and getting the same value

Evaluation using MTurk

- No objective ground truth; ask for subjective opinions from users
- Qualified **Amazon Mechanical Turk (MTurk)** users judged the geo/temporal relevancy of the clusters, given the tags within clusters
- **MTurk:** a crowdsourcing Internet marketplace, users get paid to finish tasks; in our case, each question answered by 20 users
- The **geo/temporal/motion clusters** have more **geo/temporal signals**

Metric	Geographically relevant rate (# geo relevant clusters/50)	Temporally relevant rate (# temp relevant clusters/50)
Geo clusters	**58%**	
Temporal clusters		**26%**
Motion clusters	**60%**	10%
Raw co-occurrence clusters	22%	2%
Mutual information clusters	22%	12%

Evaluation using MTurk

- Clusters with high average second moment values are more likely to be judged as 'relevant'.

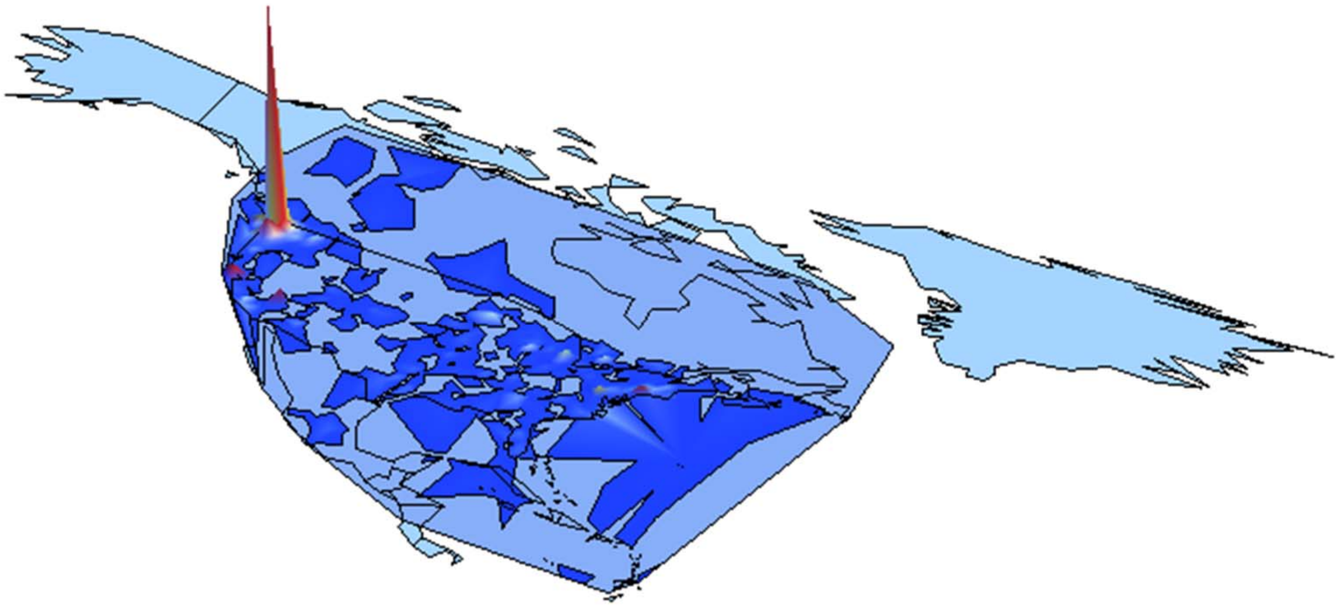
Metric	# of relev. clusters in top 10 results
Geo clusters	**9** clusters are **geo** relevant
Temporal clusters	**7** clusters are **temporally** relevant
Motion clusters	**9** clusters are **geo** relevant

- Average second moment is an indicator of geo/temporal relevancy

Visualizations

- Geographically relevant geo clusters

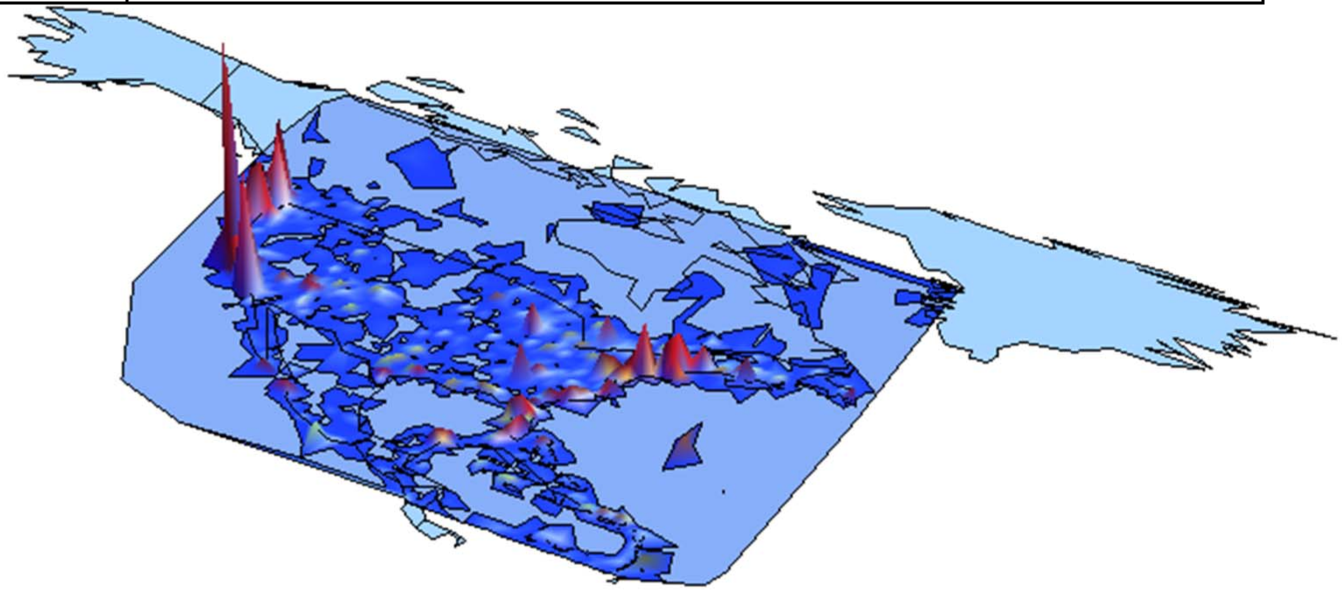
rank	6
tags	seattle needle pugetsound spaceneedle wa sound fremont northwest



Visualizations

- Geographically relevant geo clusters

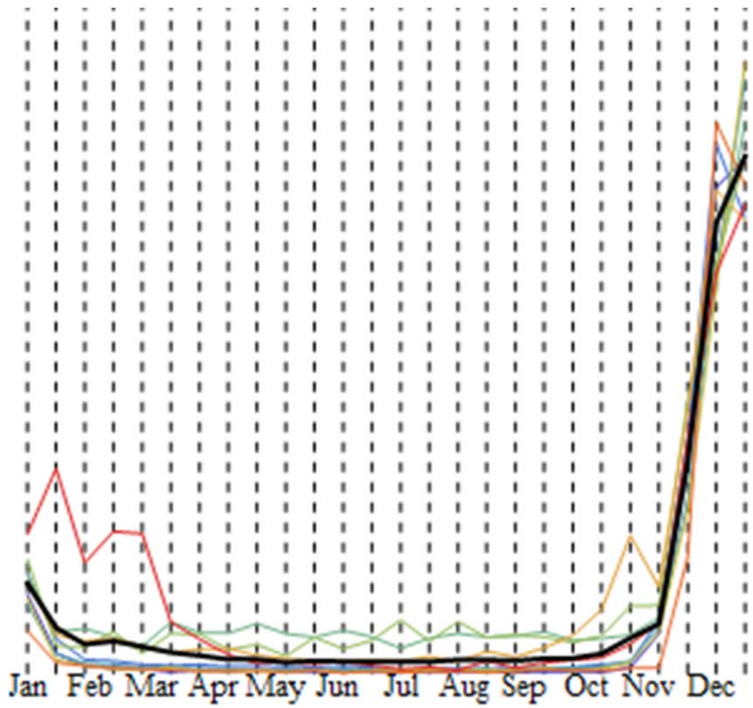
rank	28
tags	seaweed ocean waves pacific wave starfish sea seal coast pacificocean tide cliff cliffs otter jellyfish aquarium whale cove monterey



Visualizations

- Temporally relevant temporal clusters

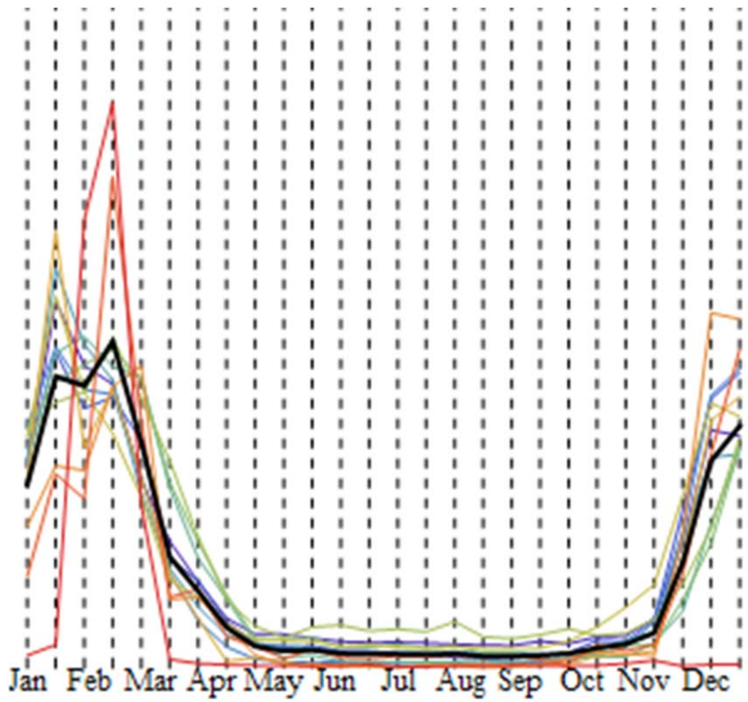
rank	7
tags	christmastree christmaslights christmas ornament holidays xmas decorations december snowman



Visualizations

- Temporally relevant temporal clusters

rank	12
tags	ice snow winter frozen snowboarding skiing ski cold icicles snowstorm blizzard february



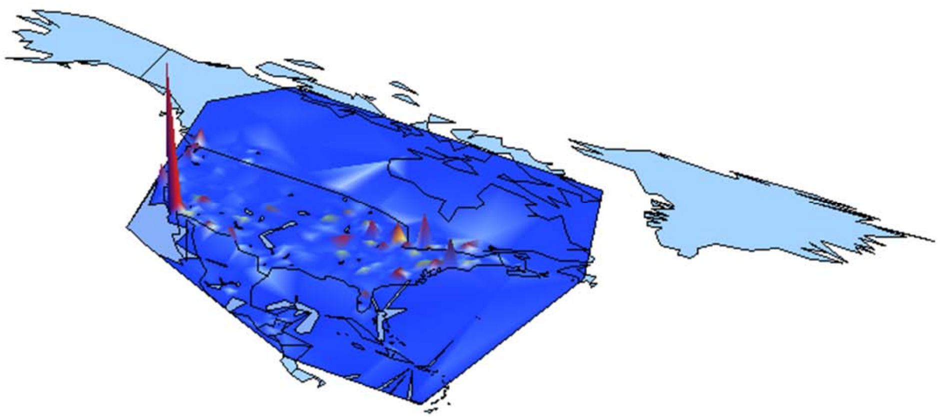
Visualization and Evaluation

- Wanted to see what happened when people were shown the visualizations
- Gave visualizations to users when they were judging the relevancy just as possible references; asked them to judge base on tags

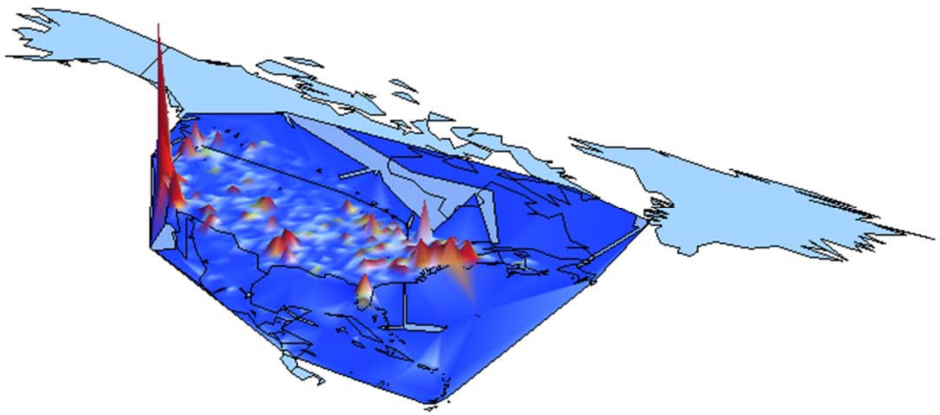
Metric	Geo relevant rate	Temporally relevant rate
Geo clusters	58% -> **(62% if with visualizations)**	
Temporal clusters		26% -> **(38% if with visualizations)**

Visualization and Evaluation

- Cases in which people changed their minds, after they saw the visualizations
- (**without vis**.) not geo relevant. -> (**with vis**.) geo relevant



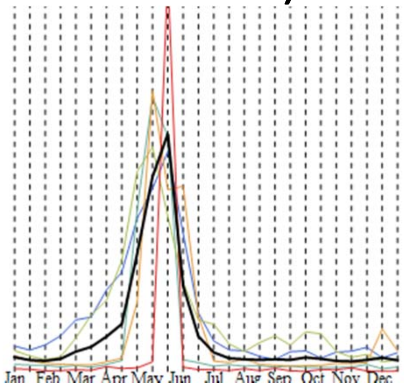
diego sandiego polarbear border



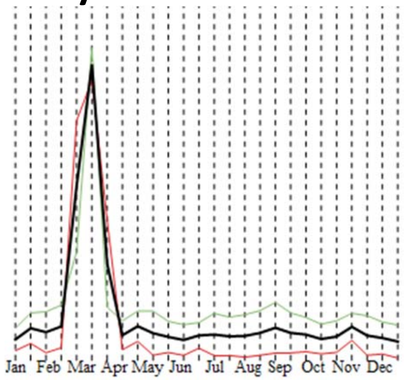
wine grapes vines barrel cows winery vineyard cattle ranch

Visualization and Evaluation

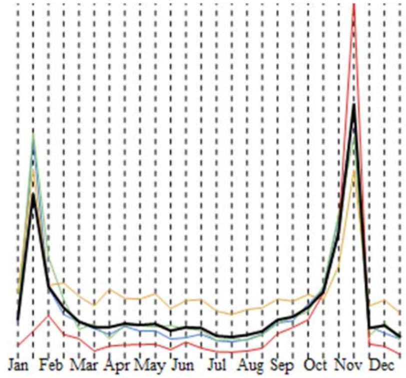
- (without visualizations) not temporally relevant -> (with visualizations) temporally relevant



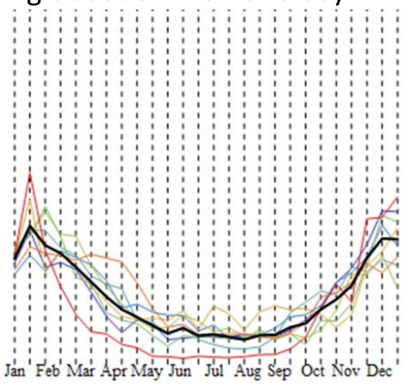
iris may dandelion graduation memorialday



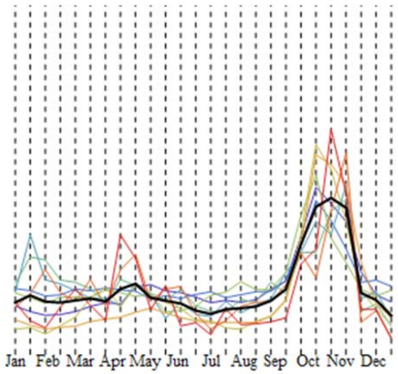
irish march



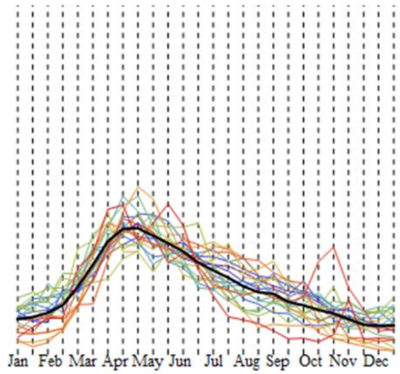
obama barackobama president election



scarf jacket hockey skating basketball footprints branches frost



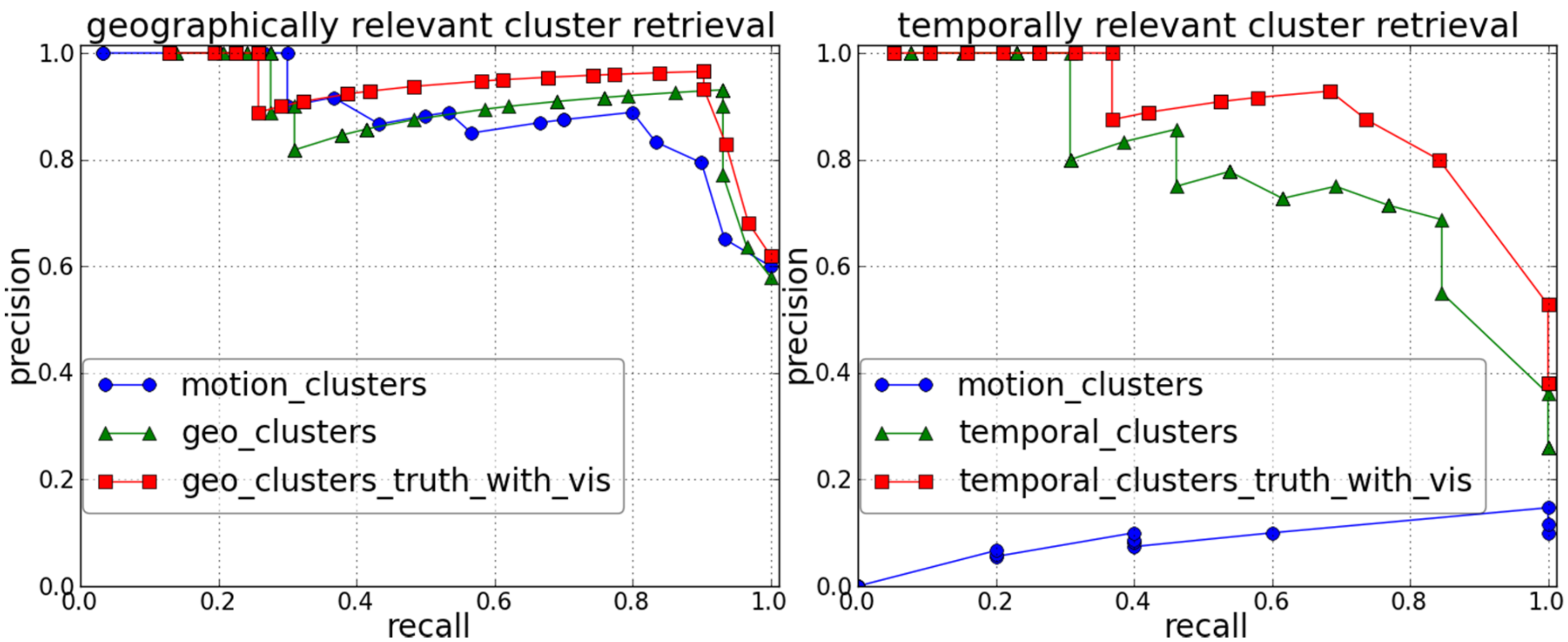
leaf colors change politics colours maple leaves rally marathon



flowers petals flower nest floral turtles osprey bud violet bloom peacock robin strawberry kite pollen wildflower iflickr wildflowers baseball ladybug poppy

Second Moment and Retrieval

- **Threshold average second moment** values to **retrieve geo/temporally relevant clusters** from geo/temporal/motion clusters



- Red curves show that when the ground truth is from the users given the visualizations, the retrieval performance is better

Conclusions

- We measured the semantic similarity of tags by comparing geo, temporal and geo-temporal patterns of use
  - Clustered tags using the proposed measurement
  - Visualized the geo and temporal clusters
- Evaluated the clusters using MTurk
  - Clusters have high quality semantics
  - Visualizations might be able to help users understand the geo-temporal semantics
  - Second moment is a simple measurement for selecting geo/temp. relevant clusters
- Future direction
  - Flexible framework that selects number of tags and clusters automatically with scalable temporal and geo bin sizes
  - Tag suggestion systems

Questions

Thank you!