GeoRich Data Management and Mining

Gao Cong (丛高)
Nanyang Technological University

Geo-Textual Data

- Social media
  - Facebook
  - Twitter
  - Weibo, etc.
- Location-based services
  - Foursquare
  - Yelp
  - Flickr, etc.

Geo-textual data grows in an unprecedented scale!

- Twitter – **500 million** tweets daily[1] (3.1% contains location[2])
- Foursquare – **9 million** check-in daily[3]

1. http://www.internetlivestats.com/twitter-statistics/
2. Sloan L, Morgan J. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PloS one*, 10(11), 2015.
3. https://foursquare.com/about

Geo-Textual Data

- Components of Geo-Textual Data:
  - Text
  - Geographical location
  - Time
- Example: Geo-tagged Tweets

Geo-textual Data – Sources

- Static geo-textual data
  - Web pages with location
  - Online business directories
    - E.g., Google My Business
  - POI data in Location-based social networks
    - E.g., 65 million POIs at Foursquare

- Streaming geo-textual data
  - Geo-tagged micro-blog posts
    - E.g., 10 million geotagged Tweets per day
  - Photos with tags and geo-location in social photo sharing websites
    - E.g., Flickr, Instagram
  - Check-in information at POIs in location-based social networks (e.g., Foursquare, Facebook places)
    - E.g., Foursquare had 7 million check-in on 3rd Oct 2015

Smart Nation Applications

GeoSpatial Data Mining

- POI recommendation & prediction
- Interactive exploration geospatial data
- Knowledge graph for locations
- Trajectory representation and similarity
- Speed, travel time, route prediction
- Region search, (e.g., burst region)
- Region exploration (topic, crowdness)

Querying and indexing spatio-temporal data

- Snapshot queries (OLTP, OLAP)
- Continuous queries

Distributed streaming systems

- Distributed load balance
  Distributed materialized view
- Index &query optimizer
- Machine learning techniques

Big static/streaming geo-spatial + X (e.g., text, temporal) data

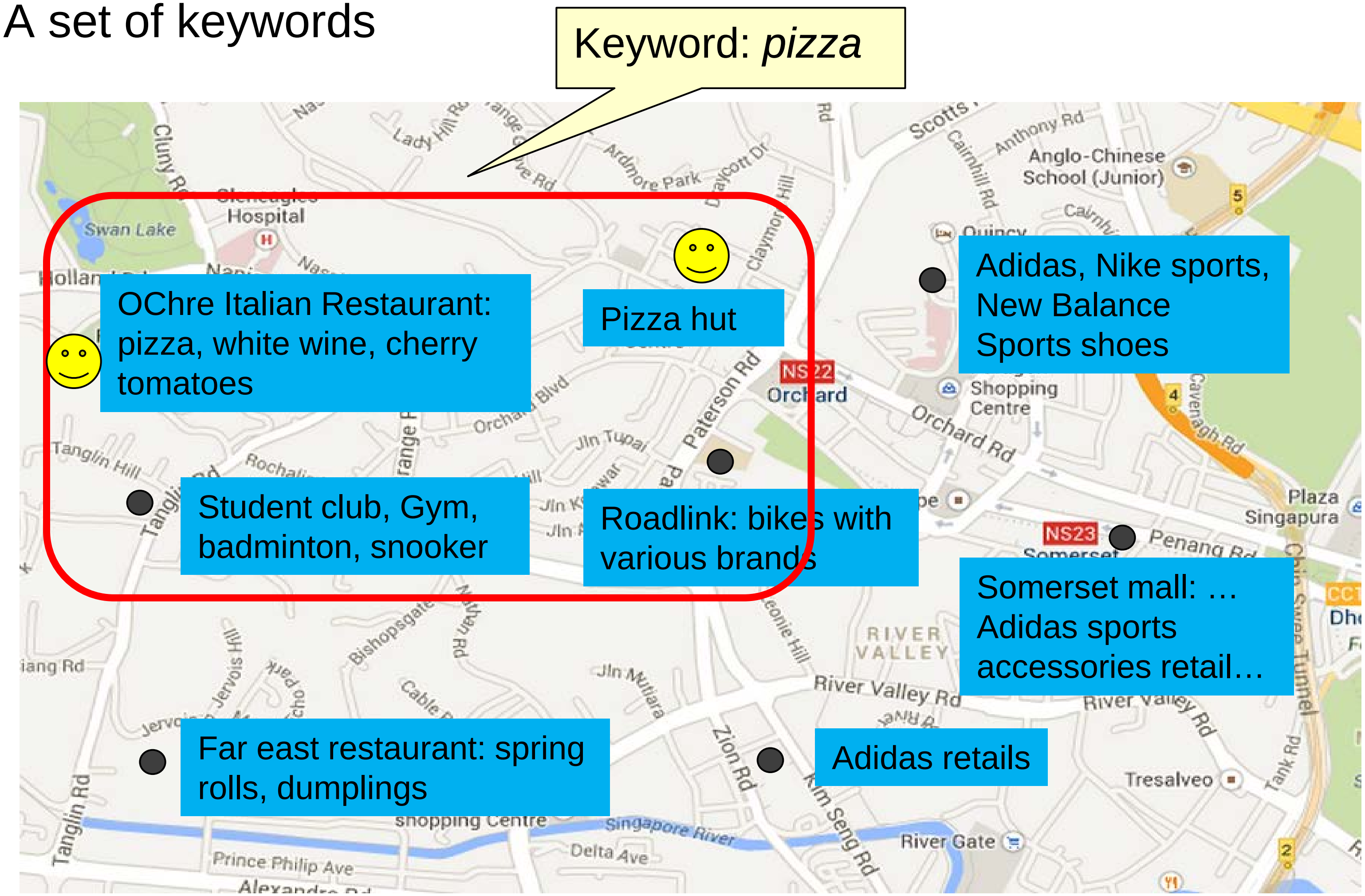
foursquare yelp Grab Google places twitter Instagram Fast beautiful photo sharing Hi, stranger

Outline

- Spatial-textual Data Management
  - Spatial keyword queries on static geo-textual data
    - **Standard queries**: spatial DB queries + IR queries
      - Boolean range query
      - Top-*k* *k*NN queries
    - Beyond single object result granularity
    - Index structures
  - Querying geo-textual streams
- Location recognition and linking (WWW'16)
- POI recommendation (VLDB'17)
- Exploring geospatial data mining (SIGMOD'18)
- Region Level geospatial data mining
  - Topic exploration in regions (SIGMOD'16, VLDB J'19)
  - Similar region search (KDD'18)

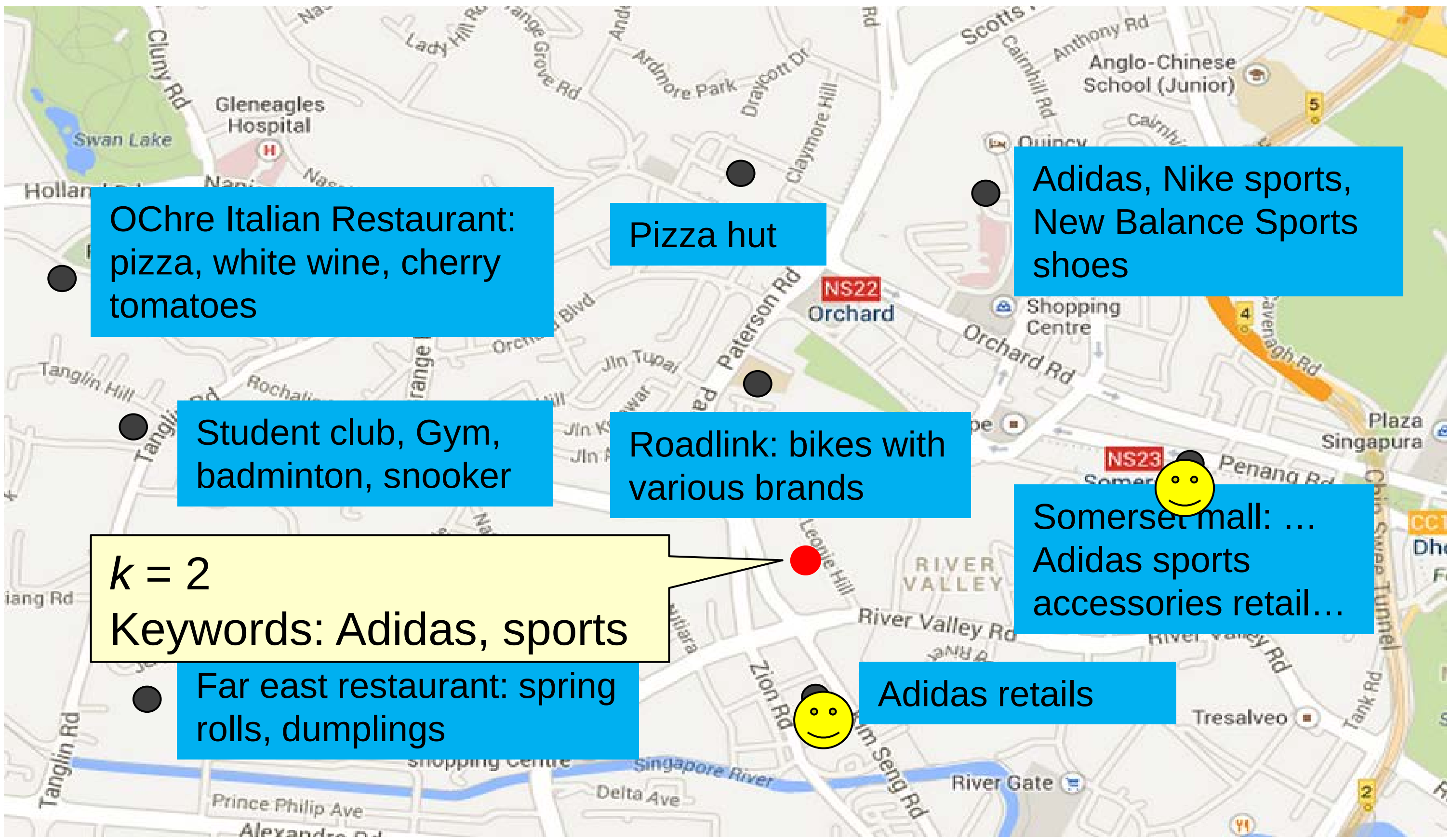
Boolean Range Query

- A query region
- A set of keywords



Top-*k* *k*NN Query (TkQ)

- A query location
- A set of keywords
- Ranking criteria: A combination of spatial proximity and text relevancy



Top-*k* Spatial Keyword Query

- Objects: $p = \langle \lambda, \psi \rangle$ (location, text description)
- Query: $q = \langle \lambda, \psi, k \rangle$ (location, keywords, # of objects)

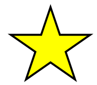
- Ranking function

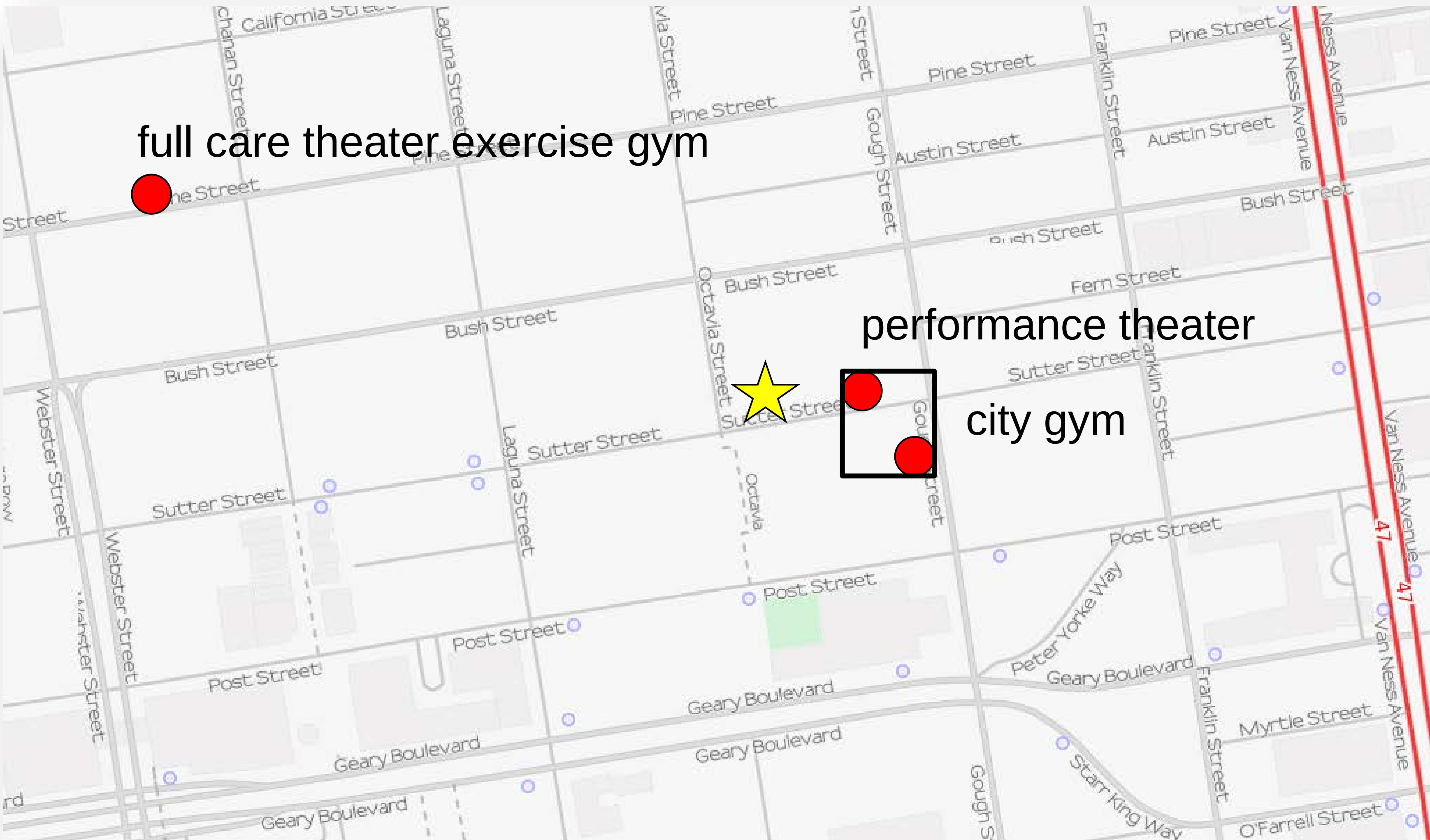
$$rank_q(p) = \alpha \frac{\| q.\lambda, p.\lambda \|}{\max D} + (1-\alpha)(1 - \frac{tr_{q.\psi}(p.\psi)}{\max P}) \qquad 0 \le \alpha \le 1$$

  - Distance: $\| q.\lambda, p.\lambda \|$
  - Text relevancy: $tr_{q.\psi}(p.\psi)$
    - Probability of generating the keywords in the query from the language models of the documents

- Generalizes the *k*NN query and text retrieval

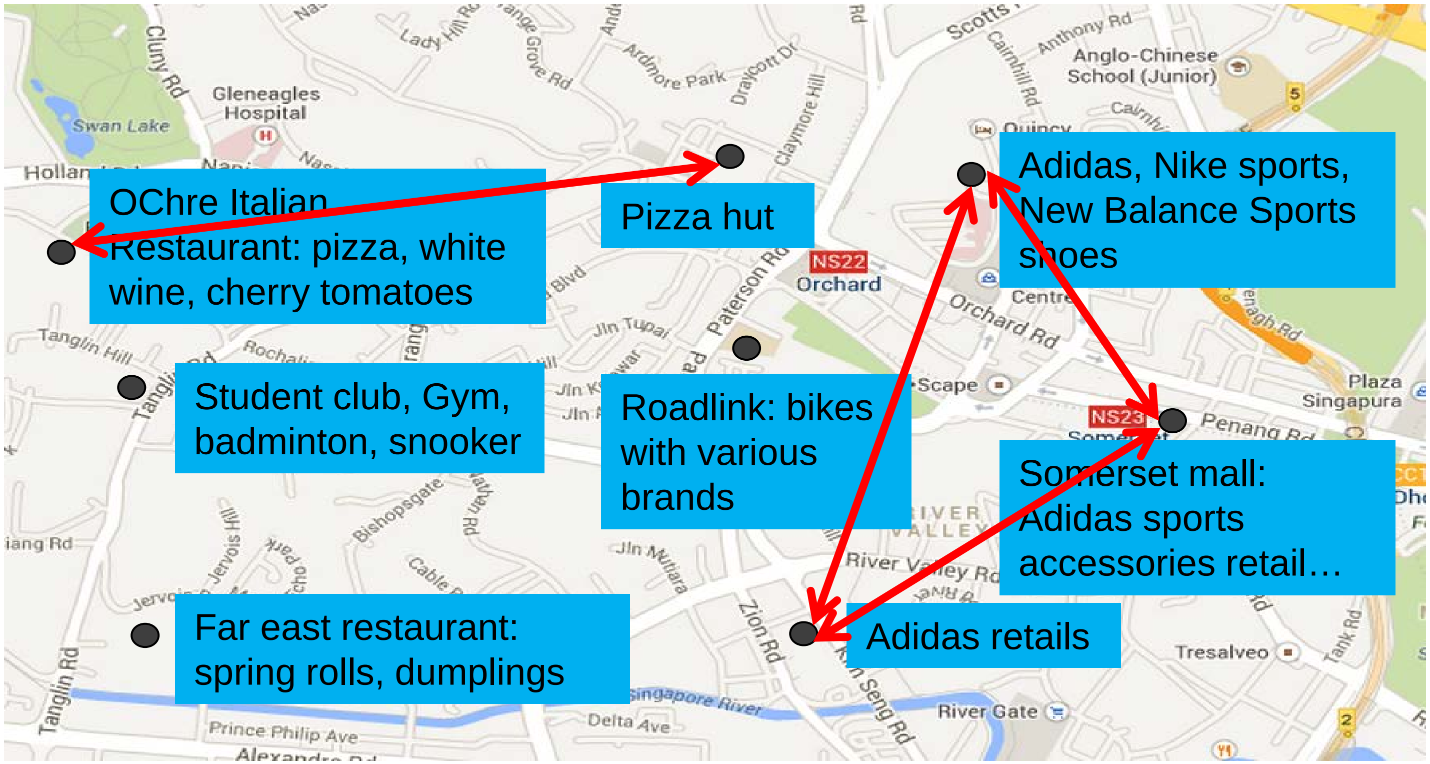
The Collective Spatial Keyword Query

- Query location: 
- Query keywords: theater, gym



Spatio-Textual Similarity Join

- Text Similarity Threshold T_{text}
- Spatial Distance Threshold $T_{distance}$
- Objective: Retrieve all pairs of geo-textual objects (o_i, o_j) s.t.
  - (1) TextSim (o_i, o_j) ≥ T_{text}
  - (2) Distance (o_i, o_j) ≤ $T_{distance}$

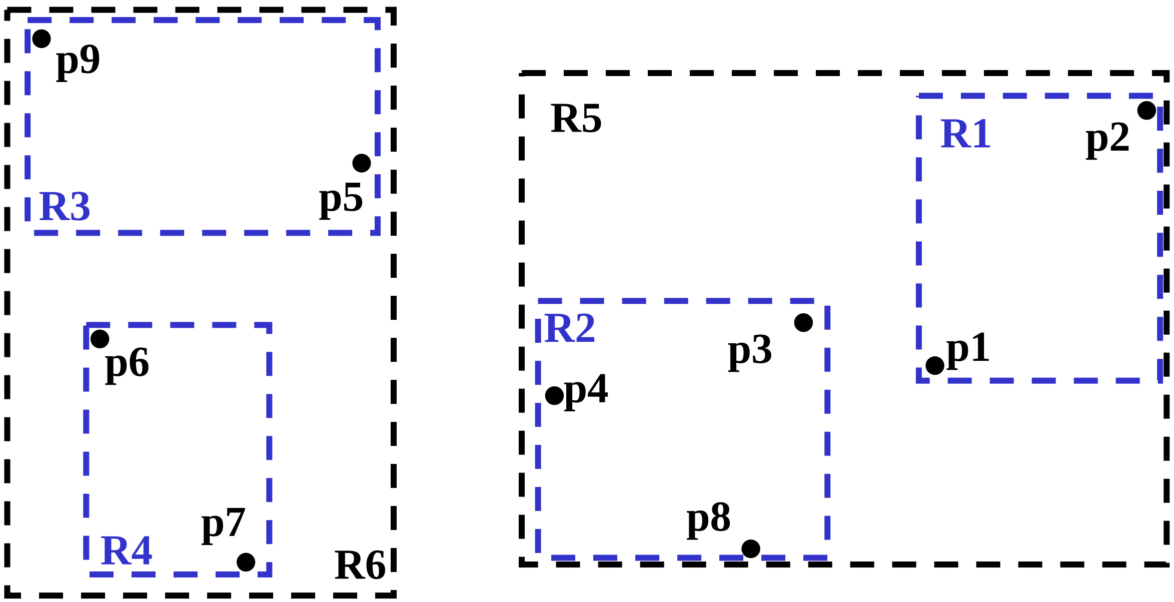


Bouros et al. *Spatio-Textual Similarity Join*, PVLDB'12

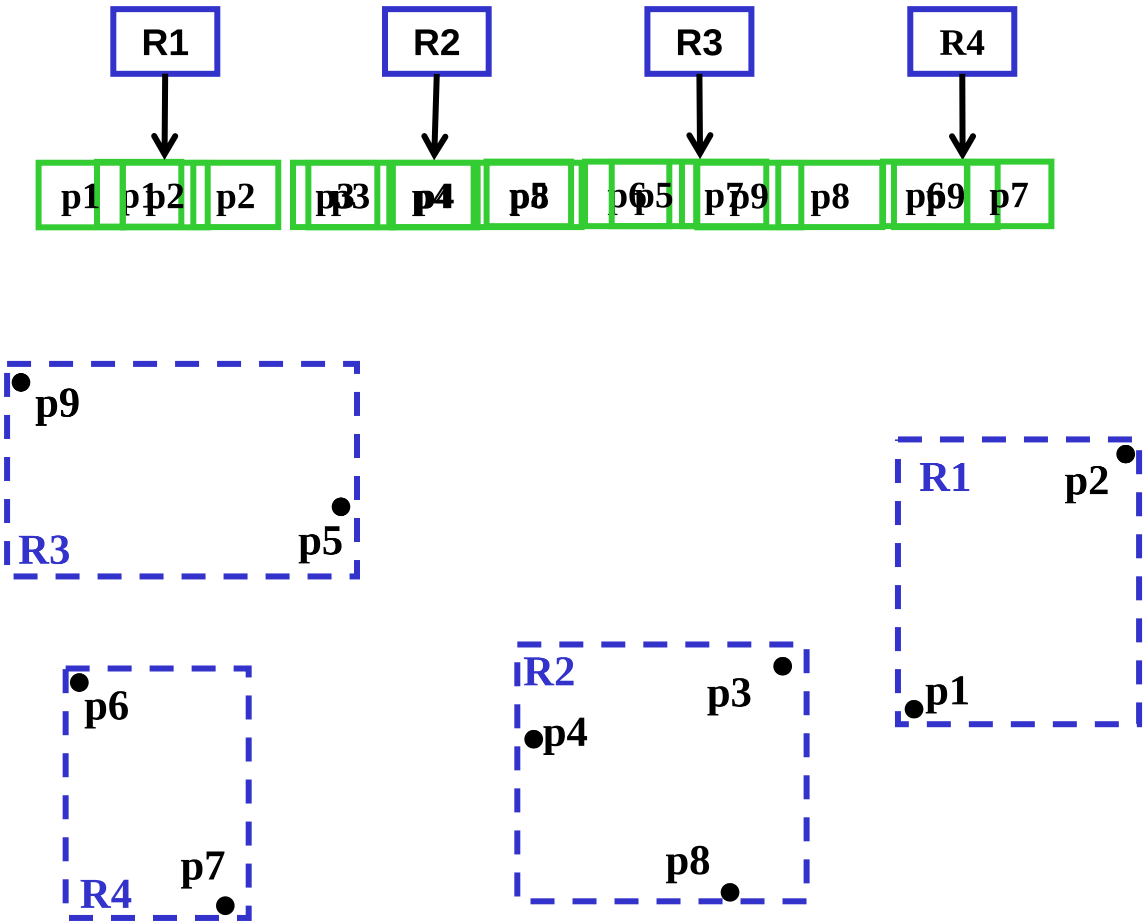
Classification for Geo-textual Indices

- Based on Spatial Indexing Scheme
  - R-tree based indices
  - Grid based indices
  - Space Filling Curve (SFC) based indices
- Based on Textual Indexing Scheme
  - Inverted File based indices
  - Signature file (Bitmap) based indices
- Based on Combination Scheme
  - Spatial-first
  - Text-first
  - Tightly combined (hybrid index)

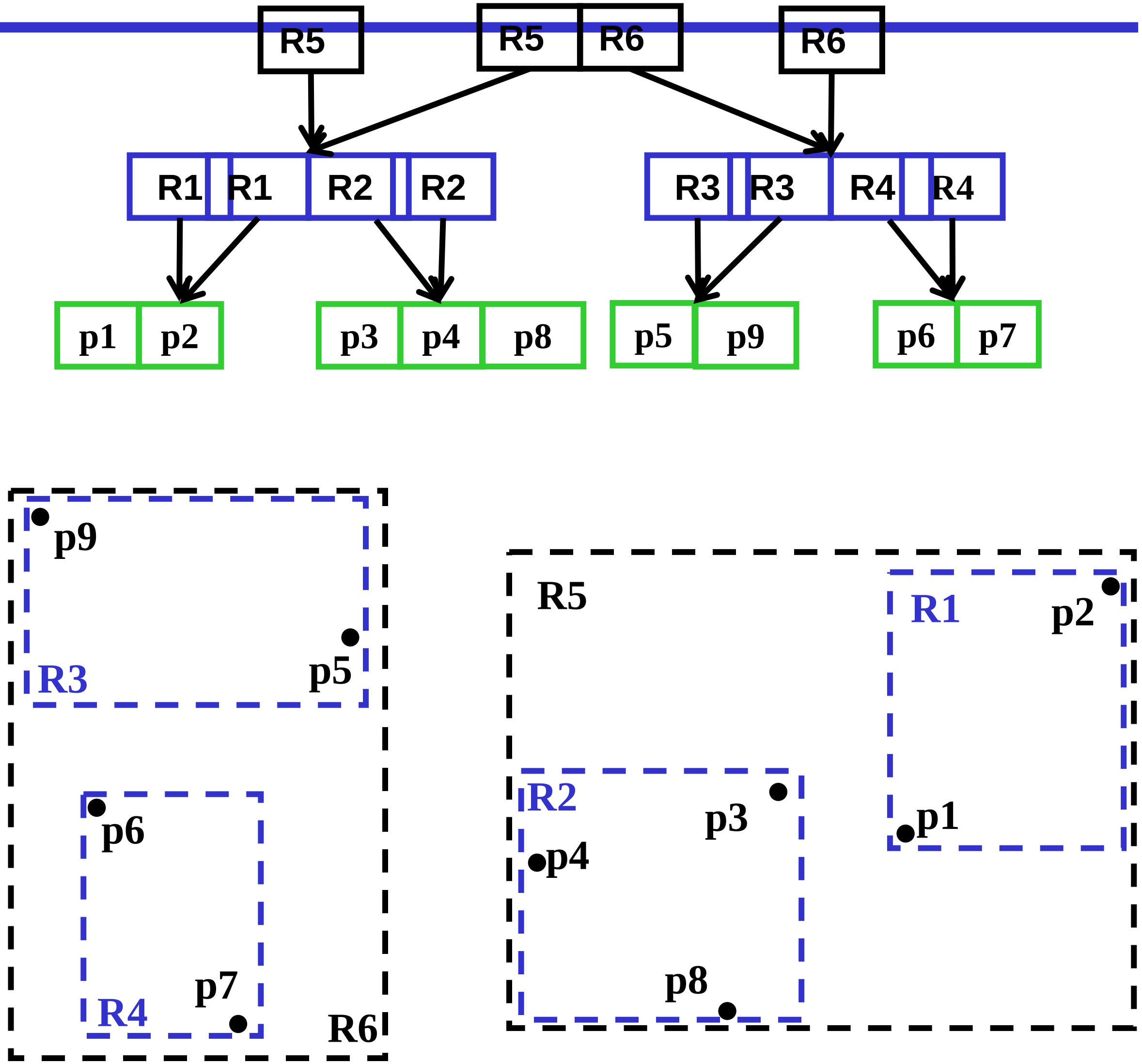
Example R-tree



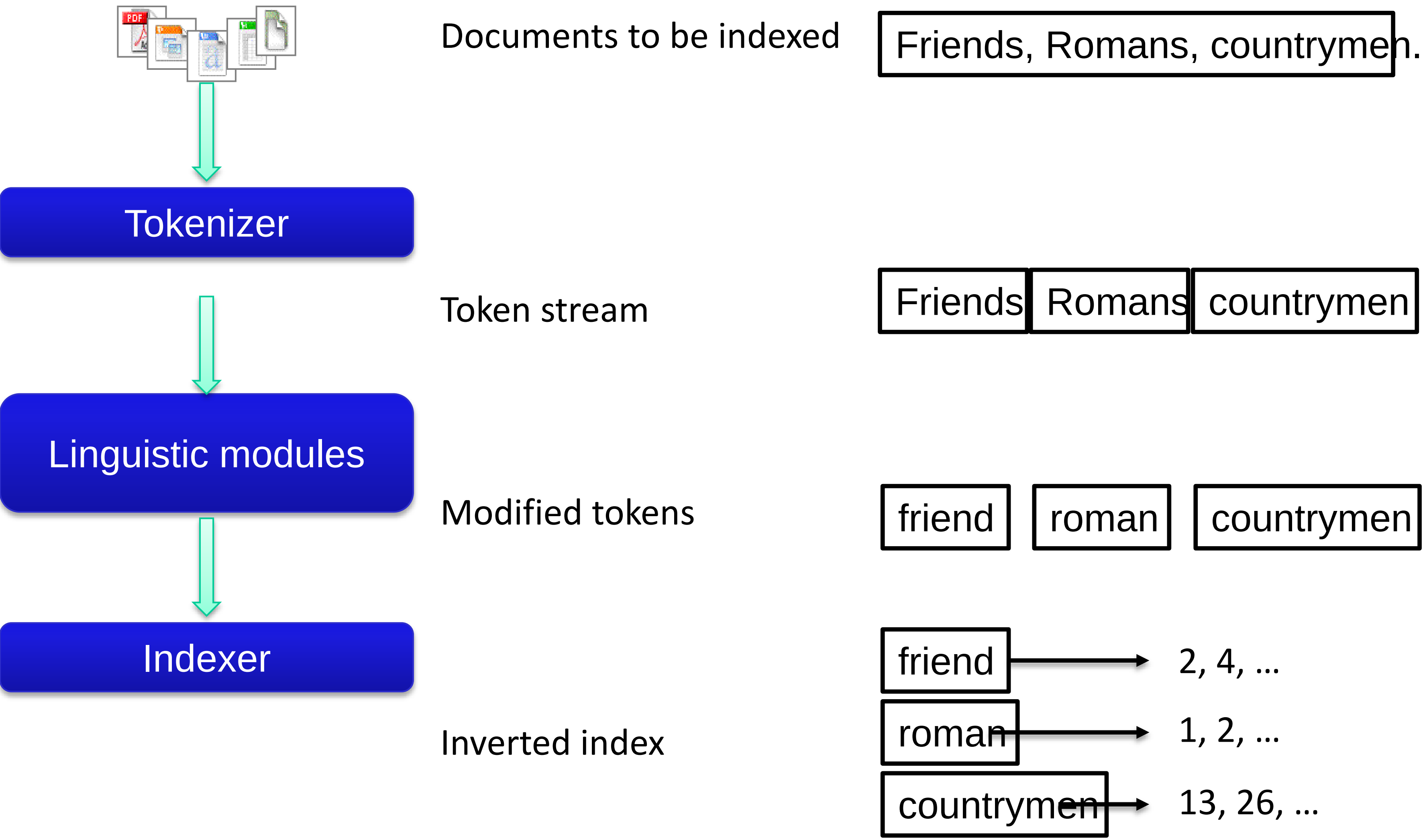
Example R-tree



Example R-tree



Recall the basic indexing pipeline



Acknowledgement: slide from website of “Introduction to Information Retrieval”

Key steps: sort by terms, then by docIDs → doc. freq, postings

Term	Doc #
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2

100 million terms to sort

Term → TermIDs (4 bytes)

Term	Doc #
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

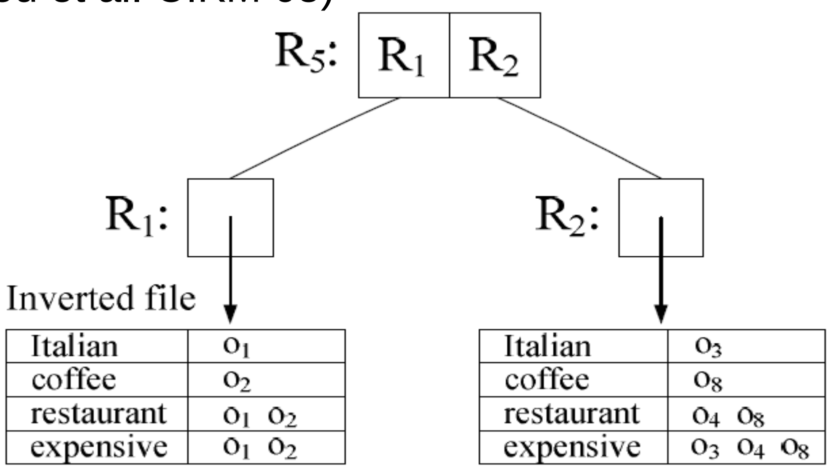
term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
i	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2

Acknowledgement: slide from website of “Introduction to Information Retrieval”

Geo-textual Indices

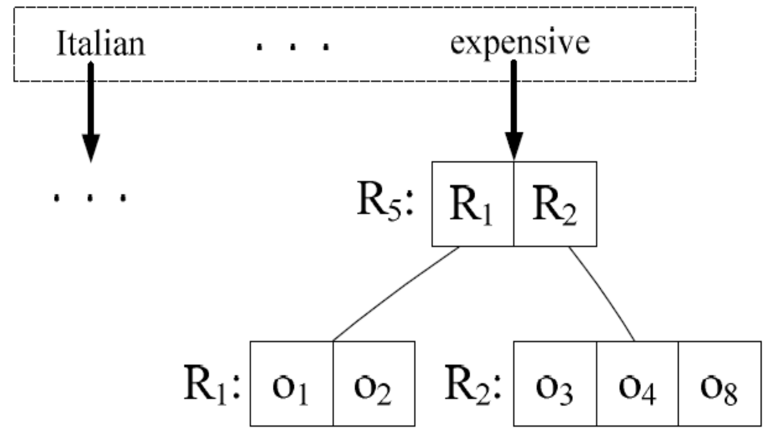
R-tree + Inverted file (R-IF)

(Zhou et al. CIKM 05)



Inverted file + R-tree (IF-R)

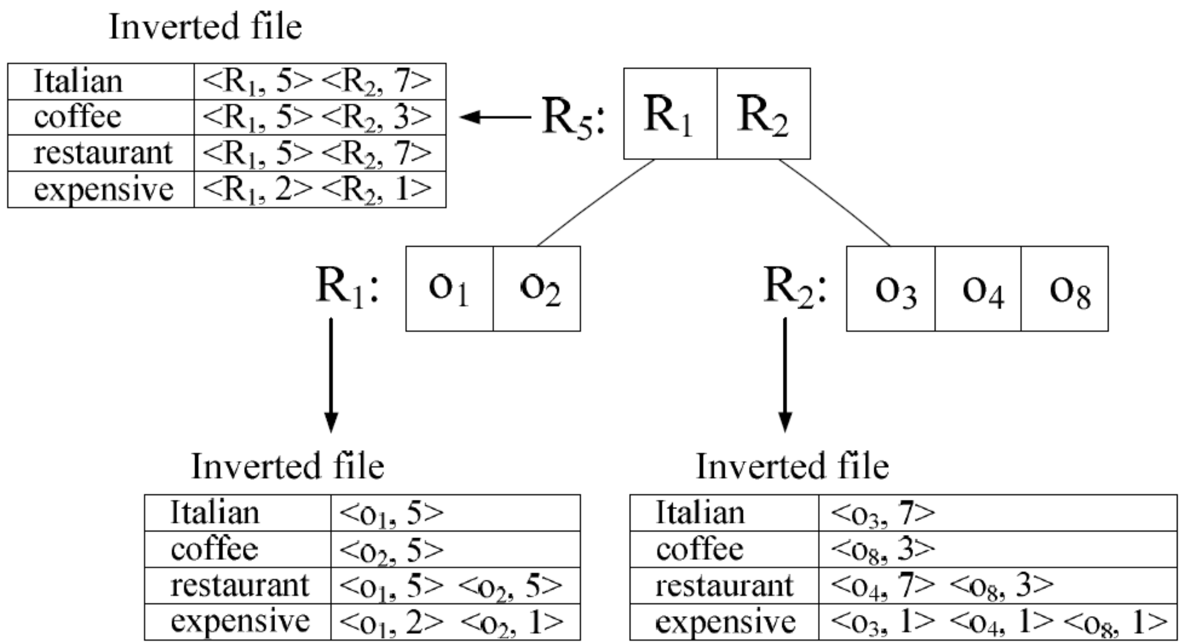
(Zhou et al. CIKM 05)



Index:	R-IF	IF-R
Spatial:	R*-tree	R*-tree
Textual:	Inverted file	Inverted file
Combination:	Spatial-first	Text-first
Query Types:	BkQ, BRQ	BkQ, BRQ

Geo-textual Indices

IR-tree and its variants (WIBR-tree, CDIR-tree)
(Cong et al. VLDB 09, Li et al. TKDE, Wu et al. TKDE 11)



IR-tree / CDIR-tree

Spatial:	R-tree
Textual:	Inverted file
Combination:	Hybrid
Query Types:	BkQ, BRQ, TkQ

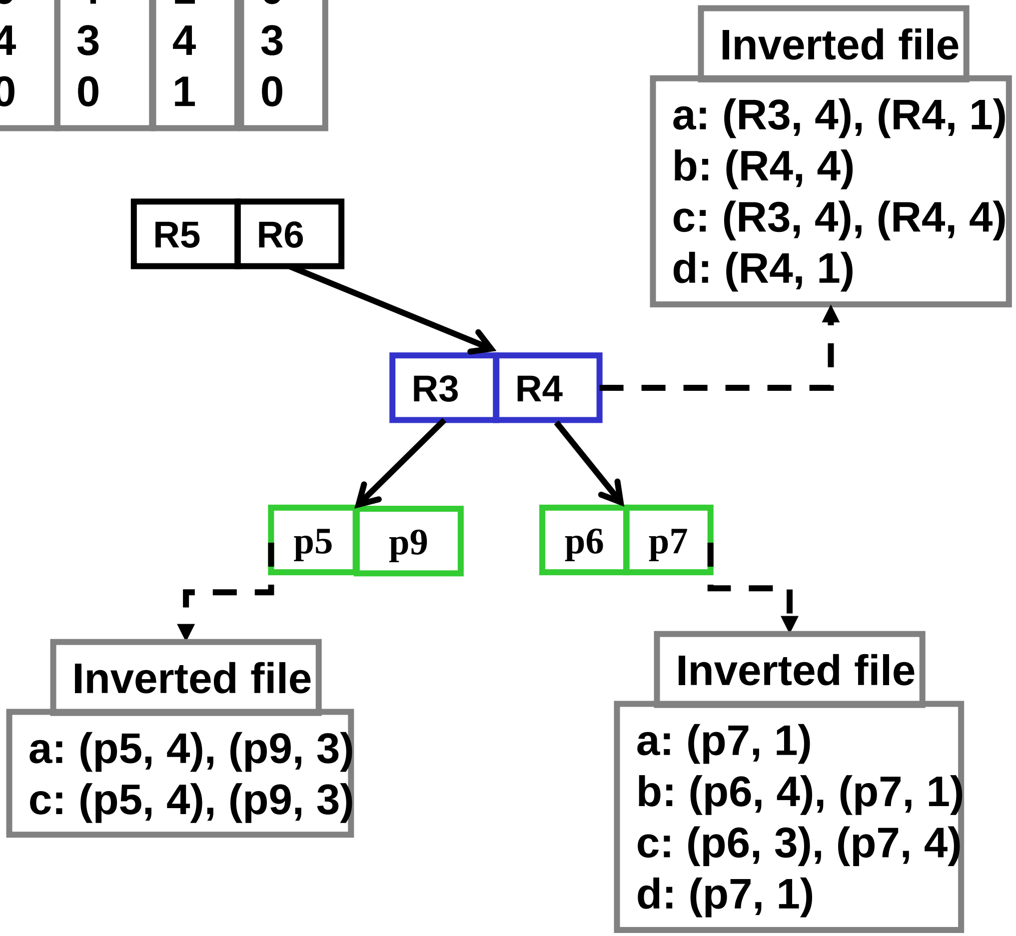
WIBR-tree

Spatial:	R-tree
Textual:	Inverted Bitmap
Combination:	Hybrid
Query Types:	BkQ, BRQ

Example of IR-tree index

Object descriptions

	p5	p6	p7	p9
a	4	0	1	3
b	0	4	1	0
c	4	3	4	3
d	0	0	1	0



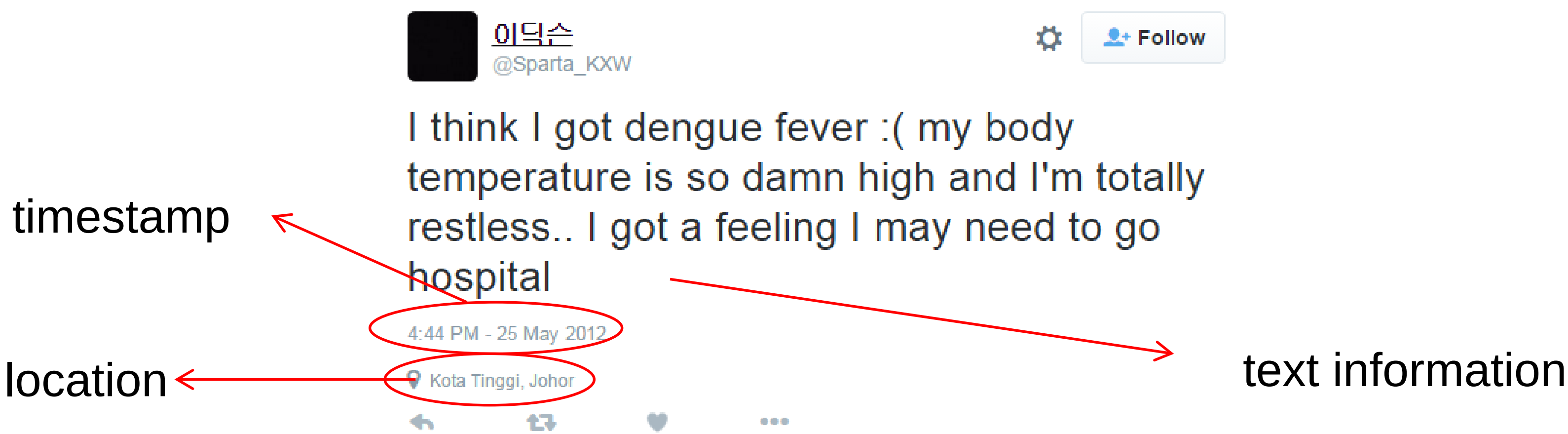
IR-Tree

- Each node of the IR-tree records a summary of the location information and the textual content of its sub-tree.

- Thus, it is able to prune the search space by simultaneously making use of both spatial proximity and text relevancy.
  - For example, for TKQ, Object *O* is enclosed in the rectangle of node *N.*
    - Theorem: We can estimate an upper bound ranking score for all the objects enclosed in N by making using of the summary information
  - Algorithm: Apply best-first search

Motivation of publish/subscribe

- Streaming geo-textual data (e.g., geo-tagged tweets) often has the quickest first-hand reports of:
  - Breaking news
    - E.g., Osama Bin Laden's death[1]
  - Disasters
    - E.g., Bomb blast in Mumbai in Nov. 2008[3], flooding of Red River Valley in Mar 2009[2]
  - Public Health – Disease Outbreaks
    - E.g., Norovirus outbreak at universities[3], influenza epidemic 2009[3]



1. Hu, Mengdie, et al. *Breaking News on Twitter.* CHI, 2012. 2751-2754.
2. Atefeh, Farzindar, et al. *A Survey of Techniques for Event Detection in Twitter.* Computational Intelligence, 2015. Vol. 31
3. Krieck, Manuela, et al. *A New Age of Public Health: Identifying Disease Outbreaks by Analyzing Tweets.* Proceedings of Health Web-Science Workshop, 2011.

Applications of Publish/Subscribe

- Applications
  - Location-based services, e.g., Location-aware event, Local news subscription, Location-based E-coupon
    - location-based and keyword-based requirements
    - Real-time requirement (instant feeding)
  - Annotation of Points-of-Interest (POIs) with social media feeds: Bridge dynamic (streaming) world and offline world
- Challenges:

> - High arrival rate of geo-textual objects.
>   - Over 10 million new tweets with coordinates per day [1,2]
>   - Over 100 million new tweets with semantic locations per day [1,2]
> - A large number of subscription queries.

1. Leetaru, Kalev, et al. *Mapping the global Twitter heartbeat: The geography of Twitter*. *First Monday* 18.5 (2013).
2. http://www.internetlivestats.com/twitter-statistics/#sources

Boolean Range Subscription Query

- Boolean Range Subscription (BRS) Query

  $q = (\psi , r)$

  - ψ : a set of keywords connected by AND or OR semantics

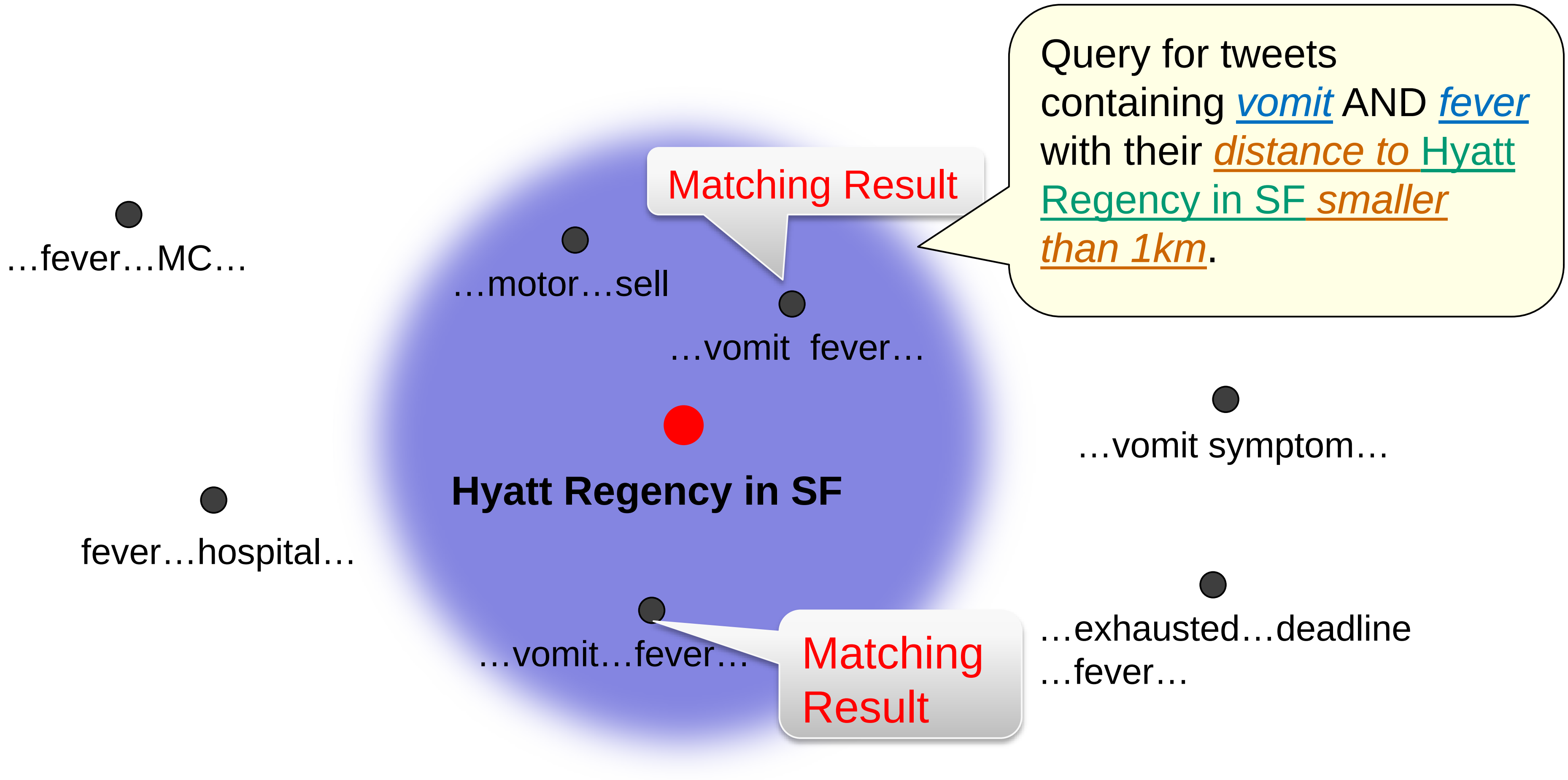
    (dengue AND fever, vomit OR poisoning)

  - r : the query region (within 1 km from Hyatt Regency in SF)

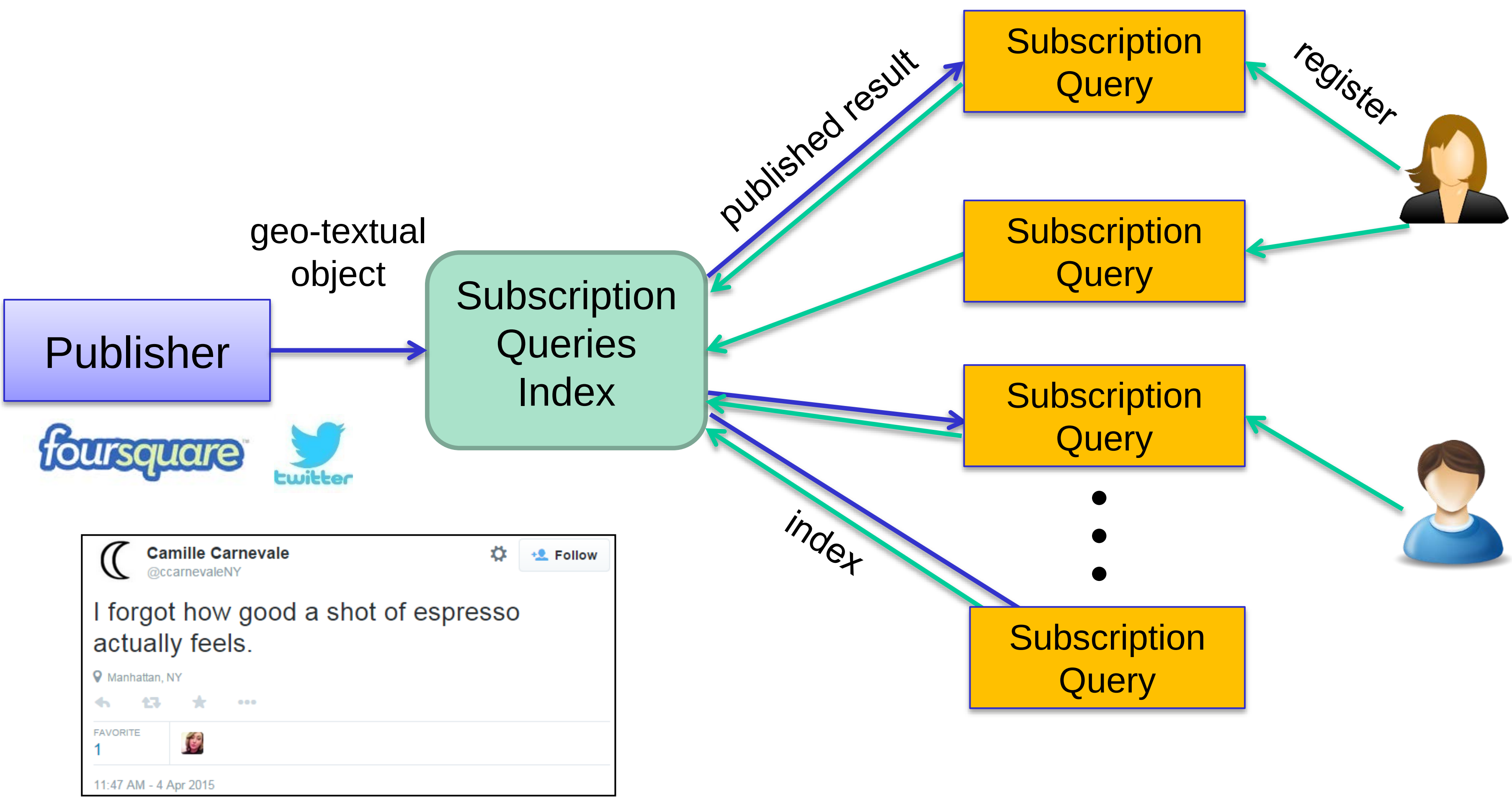
- To trigger the action of "pushing", the following conditions should be satisfied:
  - The Boolean expression, as indicated in ψ, should be satisfied by the object terms.
  - The location of object should be within the query region r.

Boolean Range Subscription Query

- Problem: answering a stream of BRS queries in real time on a stream of geo-textual objects continuously.



Location-Aware Publish/Subscribe Model



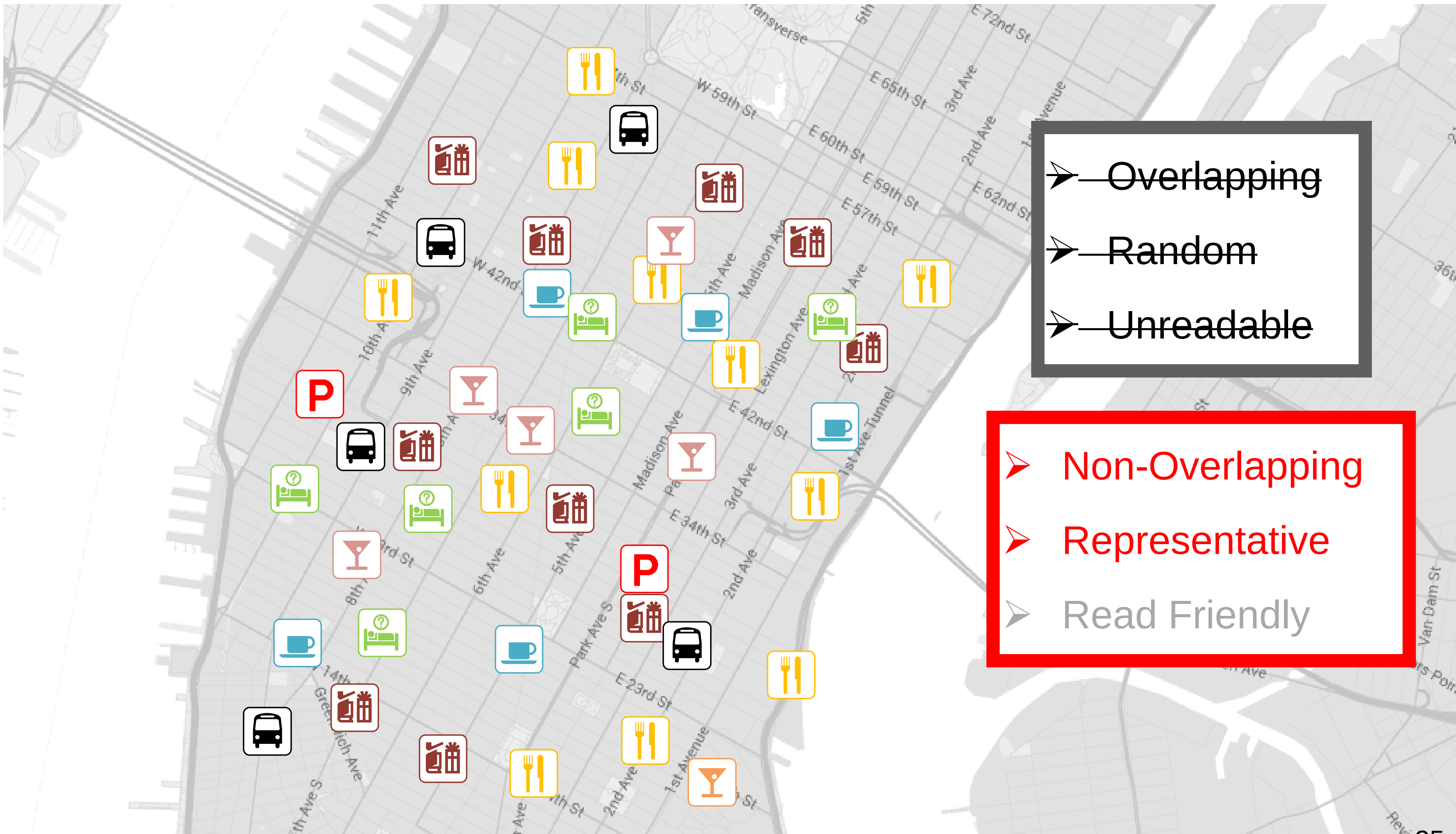
Outline

- Spatial-textual Data Management
  - Spatial keyword queries on static geo-textual data
    - **Standard queries**: spatial DB queries + IR queries
      - Boolean range query
      - Top-*k* *k*NN queries
    - Beyond single object result granularity
    - Index structures
  - Querying geo-textual streams
- Exploring geospatial data mining (SIGMOD'18)
- Location recognition and linking (WWW'16)
- POI recommendation (VLDB'17)
- Region Level geospatial data mining
  - Topic exploration in regions (SIGMOD'16, VLDB J'19)
  - Similar region search (KDD'18)

When we want to look into the spatial dataset…

- Overlapping
- Random
- Unreadable

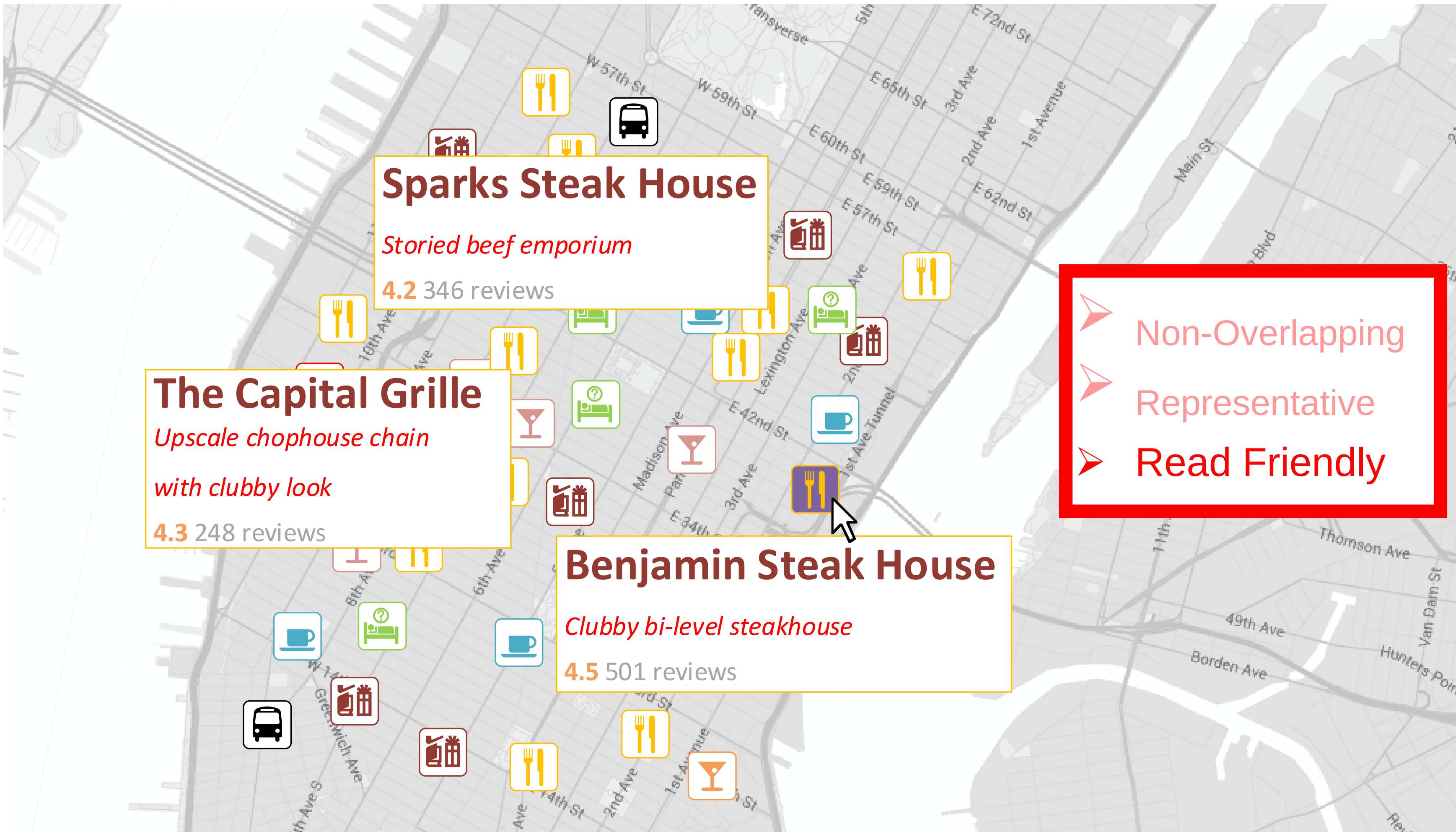
Can we do better?



- ~~Overlapping~~
- ~~Random~~
- ~~Unreadable~~

- Non-Overlapping
- Representative
- Read Friendly

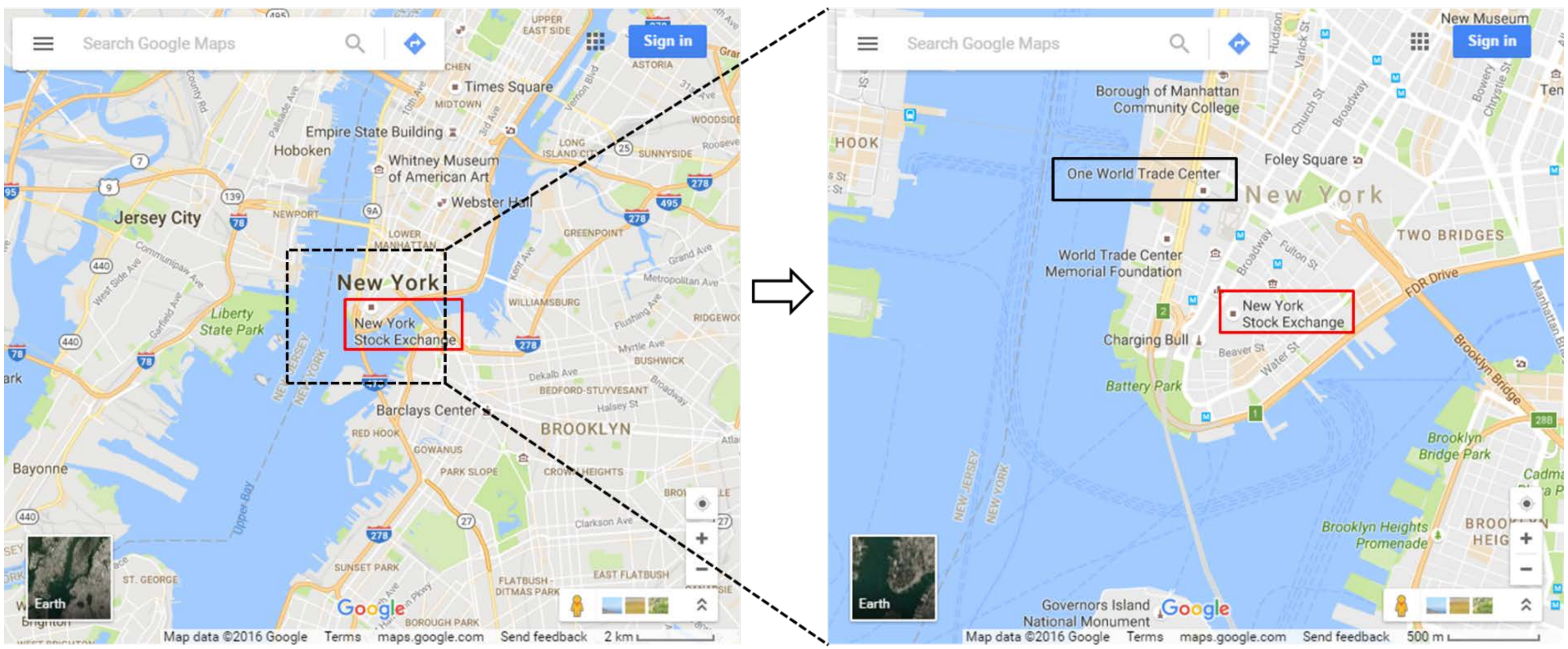
Can we do better?



- Non-Overlapping
- Representative
- Read Friendly

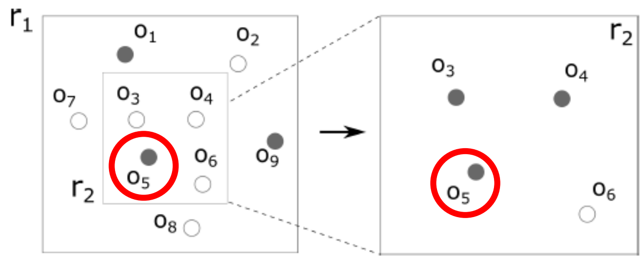
Issues of Map Exploration

- Consistency of interactive operations

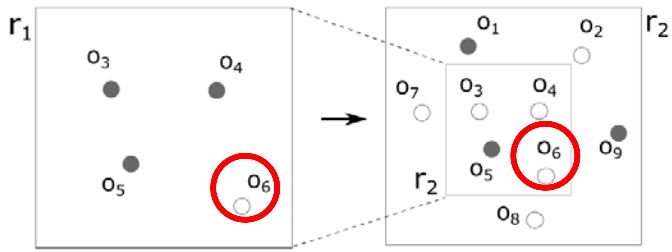


Three types of Interactive Operations

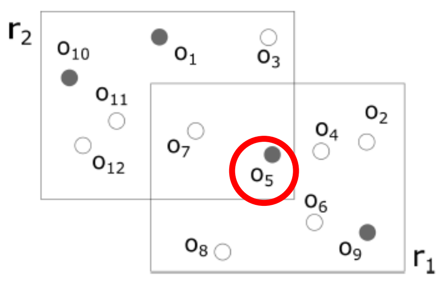
- Zoom-in



- Zoom-out



- Panning



Features of this problem

1. Zooming/Movement Consistency
   - Objects should not appear/disappear oddly.
2. Visibility Constraint
   - Objects shown should not be too close to each other.
3. Representativeness
   - The selected objects should represent as most of the dataset as possible.
4. Similarity Metrics
   - General. We aim to support various types of geospatial data. (Documents, Images, …)
5. Online
   - The user can choose a random region on the map, and the system should immediately respond.

Problem Definition

- Given a set of geospatial objects *O*, each object o in *O* is represented by a triple o=<λ,ω,A>.
    1. $o.\lambda$ is the location where o is posted.
    2. $o.\omega$ is the weight of the object (normalized in [0, 1]).
    3. $o.A$ is a set of attributes of the object.
        - For geo-tagged tweet, $o.A$ can be the textual content, post user, timestamp, …
        - The similarity between two objects $Sim(o_i, o_j)$ is computed from the attributes $o.A$.

Problem Definition

- Representative Score
  - The similarity between two objects. $Sim(o_i, o_j)$
  - The similarity between an object o and a set of objects *S*.

$$Sim(o, S) = \max_{o' \in S} Sim(o, o')$$

An object *o* can be represented by other objects *S*
= the similarity between *o* and some object in *S* is high

  - Extend *o* to all the objects *O*

$$Score(S) = Sim(O, S) = \frac{1}{|O|} \sum_{o \in O} o.\omega \times Sim(o, S)$$

Problem Definition

- Select a subset of objects *S* of size *k* from a geospatial object set *O*.
  1. Meeting the map constraints.
     - Visibility Constraint
       $dist(o_i, o_j) \geq \theta$ for any $o_i, o_j \in S$
  2. The selected objects should represent *S* as much as possible.
     - $Sim(O, S)$ is maximized

- We denote it by Spatial Object Selection (SOS) Problem

Problem Definition

- Select a subset of objects *S* from a geospatial object set *G*, where
  - $G \subseteq O$ is the set of **candidate geospatial objects**,
  - $D \subseteq O$ be the set of geospatial objects that are always visible.
  - $|S \cup D| = k$.

1. Meeting the map constraints.
   - Visibility Constraint

   $dist(o_i, o_j) \geq \theta$ for any $o_i, o_j \in S \cup D$
2. The selected objects should represent S as much as possible.
   - $Sim(O, S \cup D)$ is maximized

- We denote it by Interactive Spatial Object Selection (ISOS) Problem

Solution Overview

1. We have developed an interactive visualized exploration system for geospatial data, which took representativeness, visibility, zooming consistency, and panning consistency into consideration.
2. We propose SOS problem to select *k* representative objects, and prove it is NP-hard.
3. We enhance the efficiency for large dataset by a sampling technique with theoretical guarantee.
4. We propose ISOS problem to support navigation operations.

Solution Overview

- We prove that the objective function is submodular.
  - We propose a greedy solution to solve the SOS/ISOS problem.
    1. Initially, *S* is empty.
    2. In each round of iteration, we choose the object that increases the maximum marginal RP score.
    3. It is repeated until *k* objects are selected.
- We prove that the approximate ratio is 1/*8*.
- Optimization.
  - We utilize “lazy forward” to speed up, and the time complexity is reduced.

Solution Overview (cont.)

- For the ISOS problem, we notice some computation can be done before the user’s next operation.
  - We propose to use prefetching technique to do the pre-computation.

- Sampling Extension.
  - When the size of dataset is too large, we can accelerate the selection by working on sampled data.
  - We prove the RP score can be bounded by sampling specific number of objects.

Outline

- Spatial-textual Data Management
  - Spatial keyword queries on static geo-textual data
    - **Standard queries**: spatial DB queries + IR queries
      - Boolean range query
      - Top-*k* *k*NN queries
    - Beyond single object result granularity
    - Index structures
  - Querying geo-textual streams
- Exploring geospatial data mining (SIGMOD'18)
- Location recognition and linking (WWW'16)
- POI recommendation (VLDB'17)
- Region Level geospatial data mining
  - Topic exploration in regions (SIGMOD'16, VLDB J'19)
  - Similar region search (KDD'18)

Fine-grained location recognition and linking

- Where
  - Countries
  - States / cities
  - Fine-grained locations
    - {Street, road, hotels, shopping mall…}
- When
- What

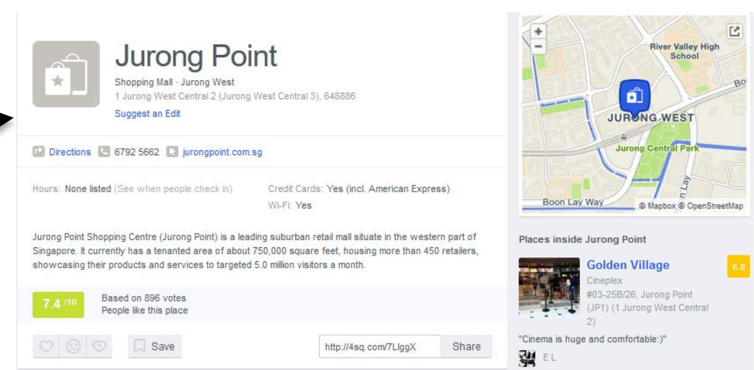
“see you later at **popular @ jp**”

Recognition

Linking

POPULAR OPENS YOUR MIND

JURONGPOINT



Tweets → recognition and linking → Foursquare

- Fine-grained location]
  - Focused geographic entity: district, area, street, road…
  - Specific point location: hotel, landmark, school, mall…

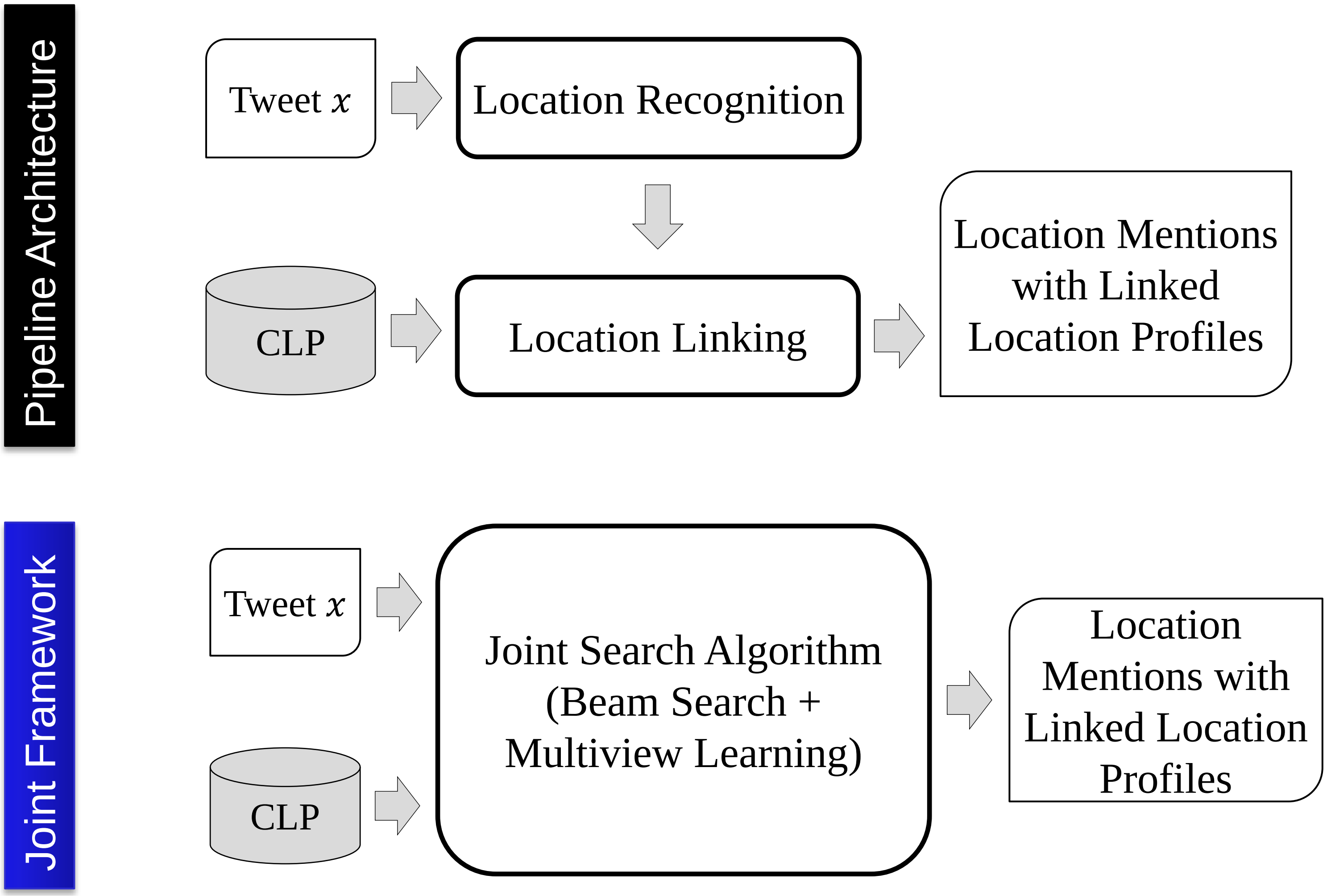
- Source document: tweets
  - Informal writing, incomplete name, nickname …
  - Ambiguity: *popular*, *mac*

- Well-defined locations:
  - **Location profile**: name, category, address, geo-coordinates ….

Example location profiles in Foursquare

Distinct location profiles for each branch with the ***same name***

Pipeline Architecture vs Joint Framework



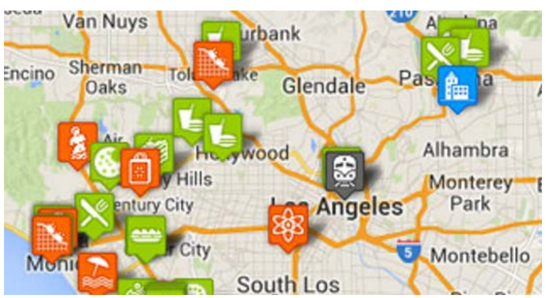
CLP: collection of location profiles from Foursquare

Introduction

- POI recommendation:
  - Given a set of POIs $\mathcal{L}$, and a set of users $\mathcal{U}$ each associated with a set of visited POIs $\mathcal{L}^u$ of user u, POI recommendation is to recommend for each user $u \in \mathcal{U}$ new POIs, i.e., $\mathcal{L}/\mathcal{L}^u$, that are likely to be visited.

  - POI recommendation helps users exploring new places and enrich their experiences.

A large number of POIs

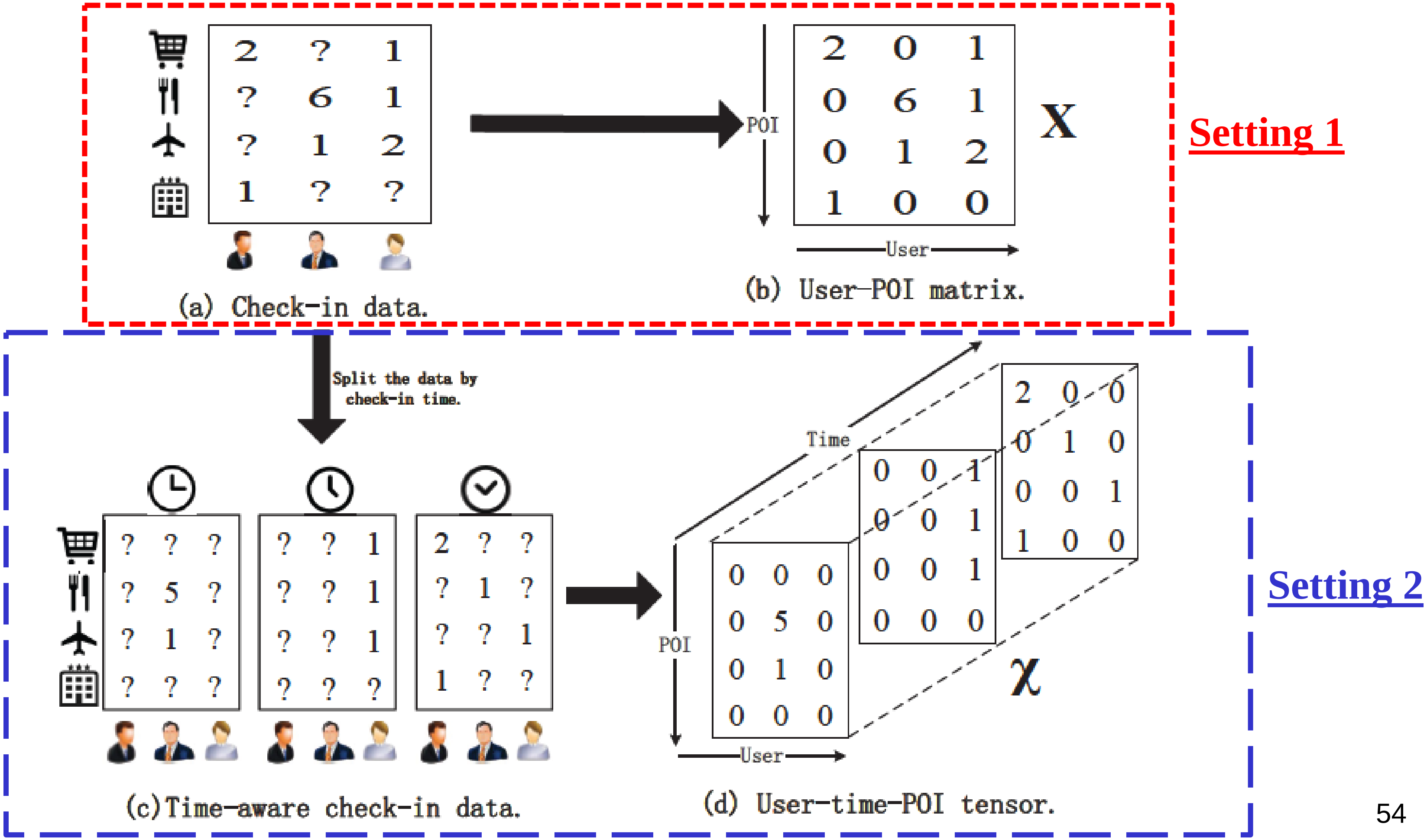


POI recommendation

Users with different interests

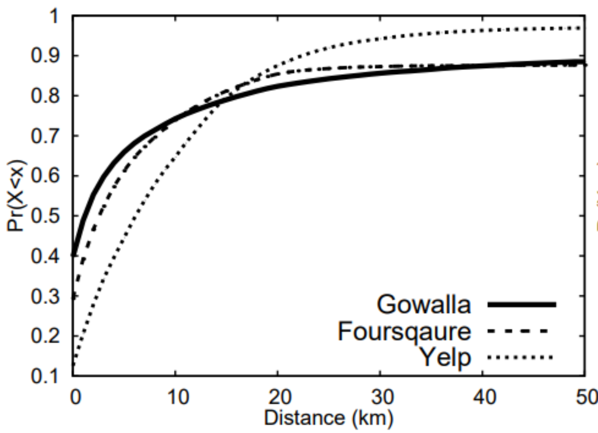
Problem settings

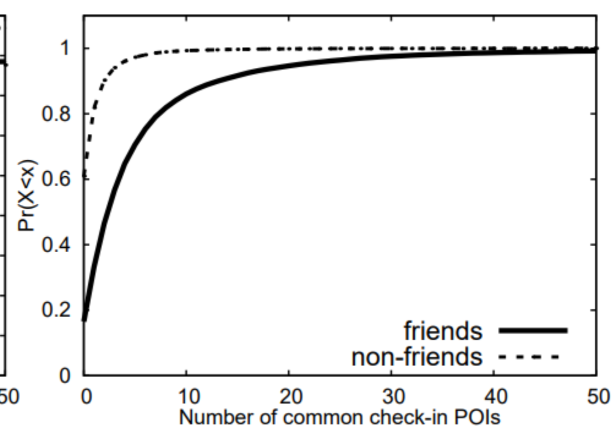
- Setting 1: POI recommendation (given *u*, recommending *l*)
- Setting 2: Context-aware POI recommendation
  - Given *u* and *t* , recommending *l*
  - Given the current location, recommend the next location.



Introduction

- Challenges of POI recommendation:
  - **Data scarcity problem**: each user has only visited a very small portion of all the POIs (usually around 0.1%, worse than traditional recommendation problems, e.g., 1.2% for Netflix data).
  - **Rich context**: user's mobility preference is affected by many types of context information, e.g., geographical distance, time and social relations.





Users tend to travel short distance

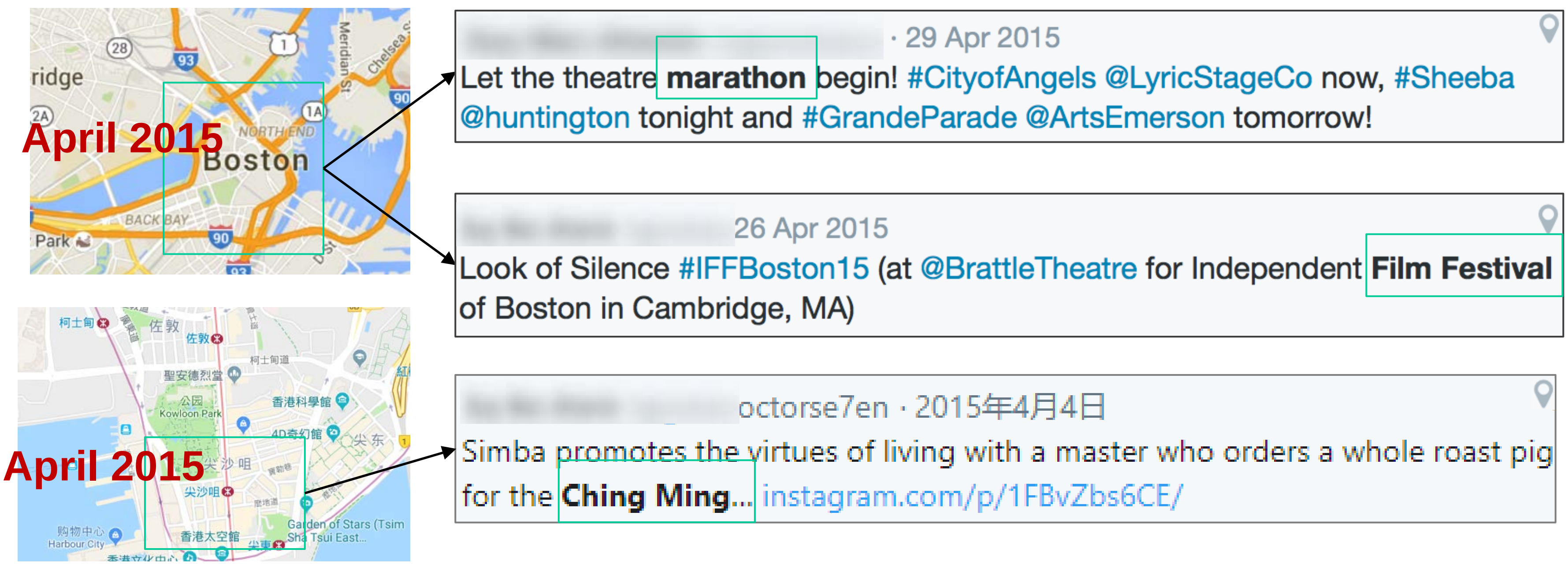
Users behave similarly to their friends

Outline

- Spatial-textual Data Management
  - Spatial keyword queries on static geo-textual data
    - **Standard queries**: spatial DB queries + IR queries
      - Boolean range query
      - Top-*k* *k*NN queries
    - Beyond single object result granularity
    - Index structures
  - Querying geo-textual streams
- Exploring geospatial data mining (SIGMOD'18)
- Location recognition and linking (WWW'16)
- POI recommendation (VLDB'17)
- Region Level geospatial data mining
  - Topic exploration in regions (SIGMOD'16, VLDB J'19)
  - Similar region search (KDD'18)

Motivation

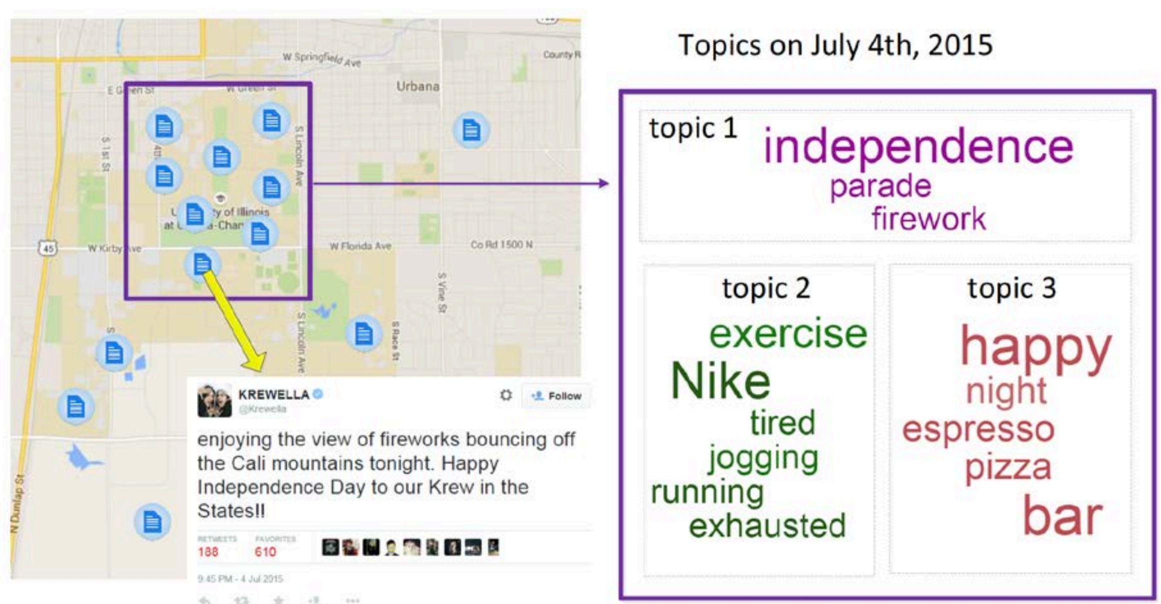
- Spatiotemporal collections, e.g., Twitter, Facebook
  - **Multiple attributes**: location (latitude, longitude), time and text
  - **Large**: 320M monthly active users, 500M tweets/day in Twitter
  - **Informative**: topics/events described by tweets
  - Topics are different in different spatiotemporal dimensions



Topic Exploration in Spatio-Temporal Data

The topic exploration problem:

- Given a collection of spatio-temporal documents D;
- **Input**: query rectangle region R and timespan $[t_b, t_e]$;
- **Output**: K topics in the region and timespan.

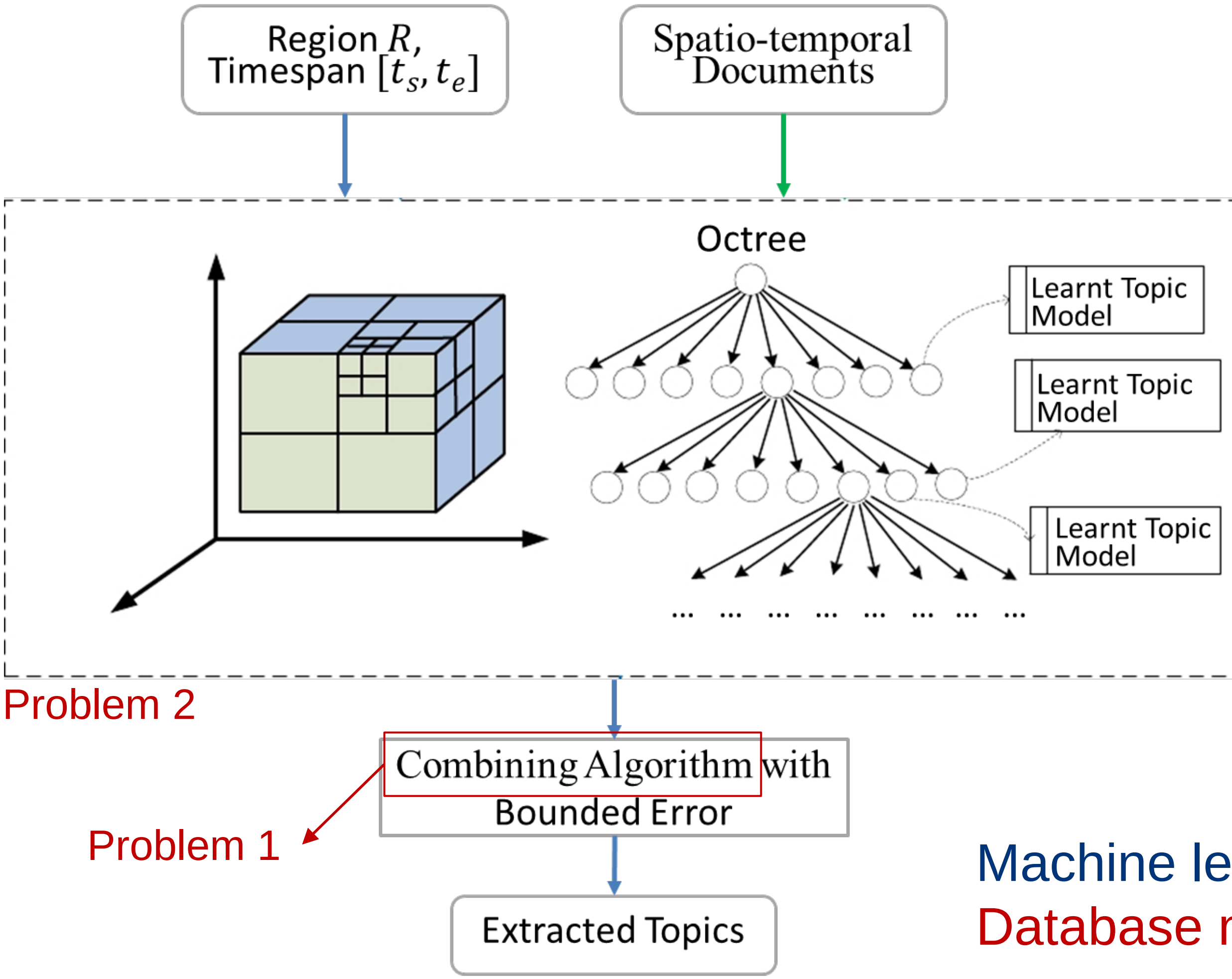


Challenges

Efficiency issue:

- Training topic models online is time consuming
  - The complexity of training LDA is $O(|W|KI)$
  - 3-months tweets (250M) in NYC → 13.85 hrs training time
  - In real scale, the number of tweets in a user specified region and timespan could be large
- User could consider to modify the query in an exploratory manner, and we could not train topic models for each query offline in advance.

Framework

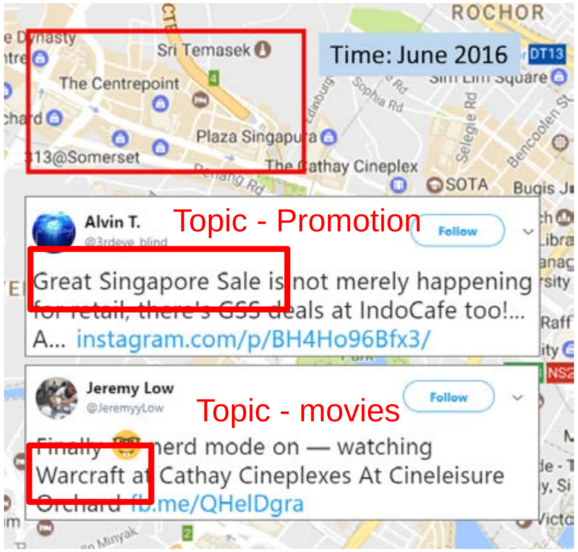


Main idea: Organize the documents into a *hierarchy structure* and *pre-train* models on selected cells to accelerate online training

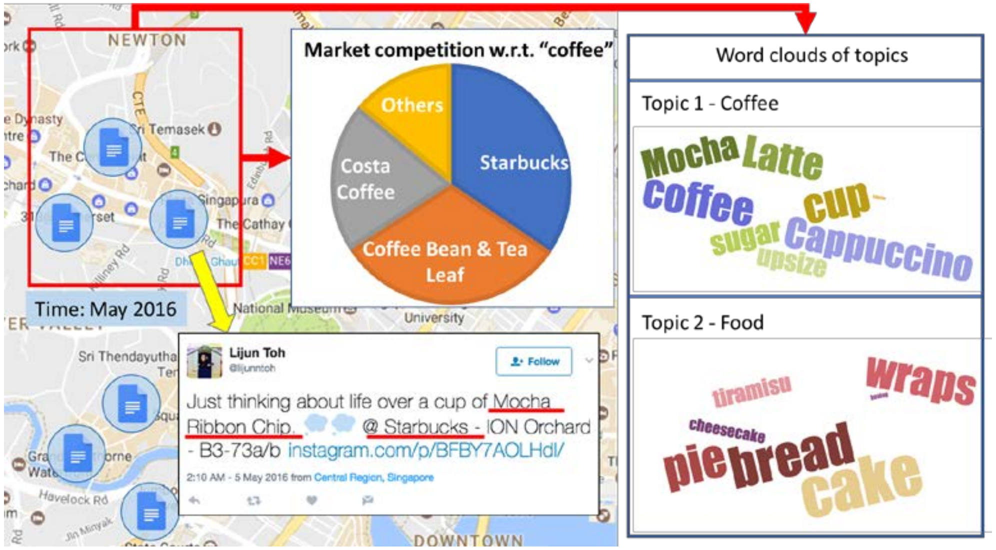
Machine learning techniques + Database management principles

Extension

- Applications
  - The high volume of data is difficult for users to consume



Summarization



Business analytics

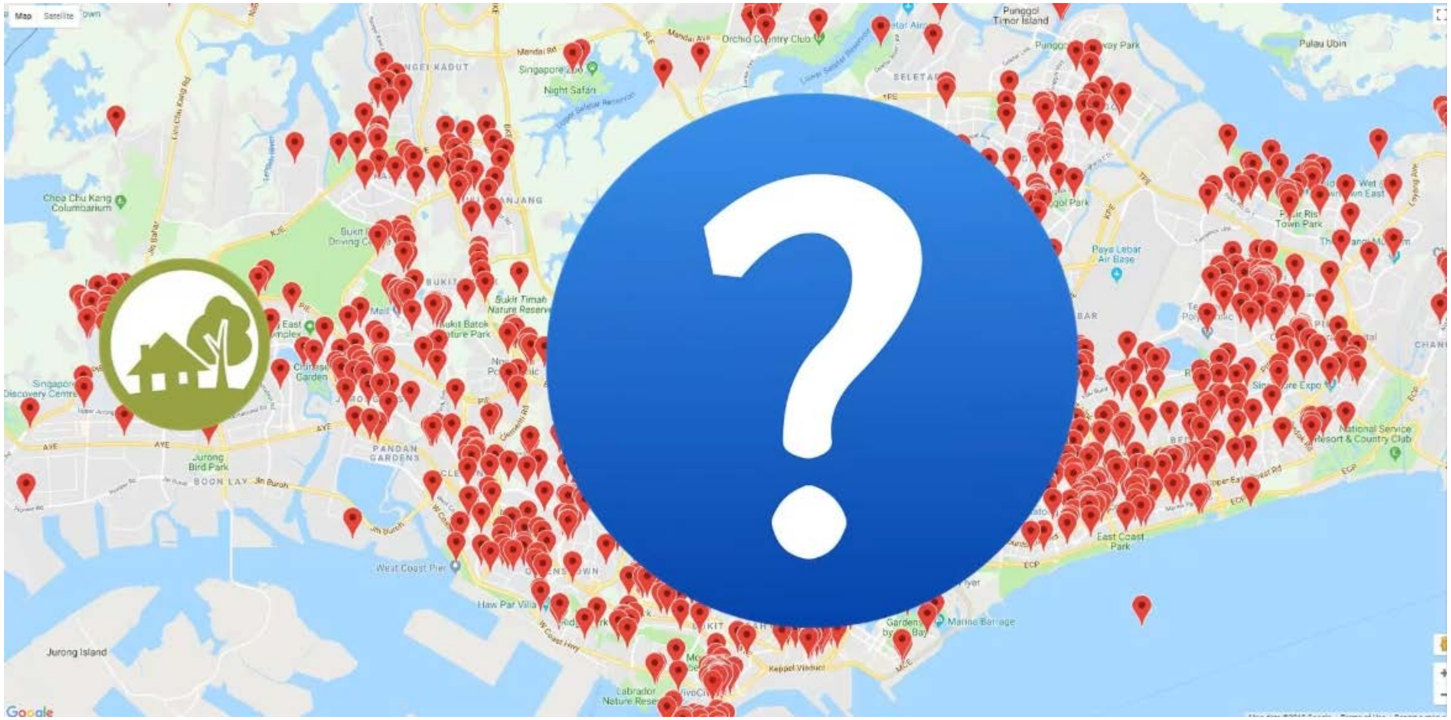
User may want to submit spatiotemporal query *in an exploratory manner* to view topics in different regions and time spans.

Outline

- Spatial-textual Data Management
  - Spatial keyword queries on static geo-textual data
    - **Standard queries**: spatial DB queries + IR queries
      - Boolean range query
      - Top-*k* *k*NN queries
    - Beyond single object result granularity
    - Index structures
  - Querying geo-textual streams
- Exploring geospatial data mining (SIGMOD’18)
- Location recognition and linking (WWW’16)
- POI recommendation (VLDB’17)
- Region Level geospatial data mining
  - Topic exploration in regions (SIGMOD’16, VLDB J’19)
  - Similar region search (KDD’18)

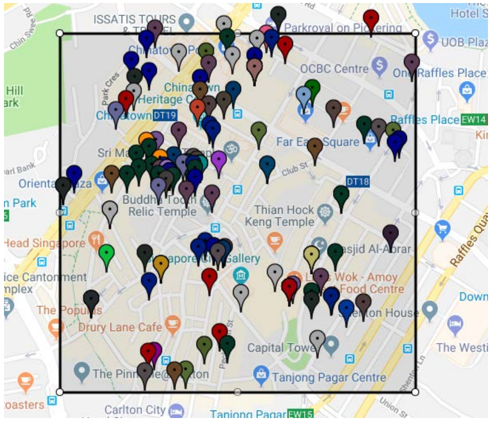
Motivation

- Expansion and complication of urban cities
  - People are only familiar with a small area (where they live).
  - Difficult and expensive to explore new areas.



Motivation

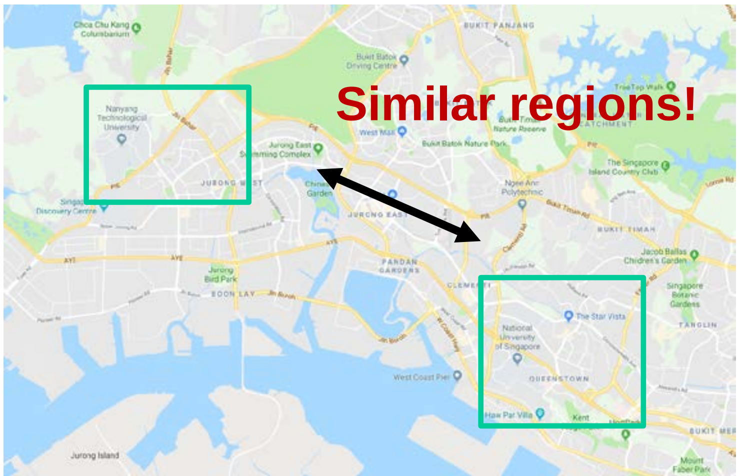
- Similar Region Search (SRS)
  - **Definition**: Given a query region, finding other similar regions.
  - **Region**: a rectangular geographical space that contain a set of POIs.
    - Each POI is associated with an attribute vector (e.g., category).

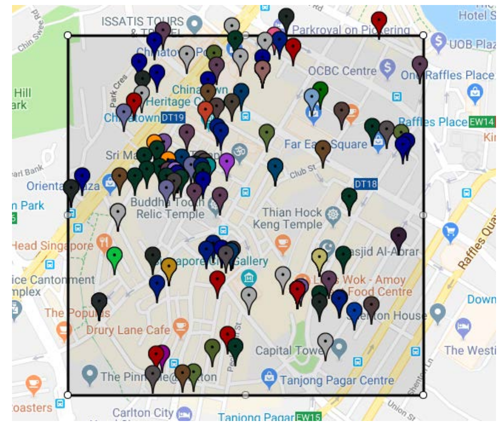


  - **Applications**: easier exploration on geographical space
    - City planning.
    - Business site selection.
    - Improving location-based services (e.g., POI/region recommendations).

Motivation

- A novel application: Similar Region Search (SRS)



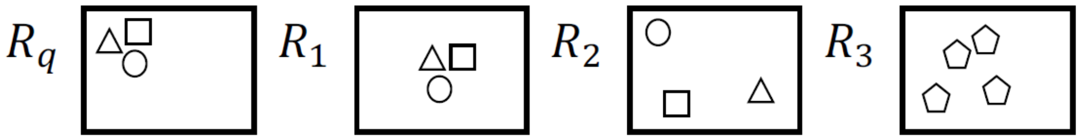


- Formal definition:
  - Given: a geo space P; a set of POIs $\mathcal{O}$ on P; a query region R_q.
  - Retrieve: a set of regions $\mathcal{R}$, such that

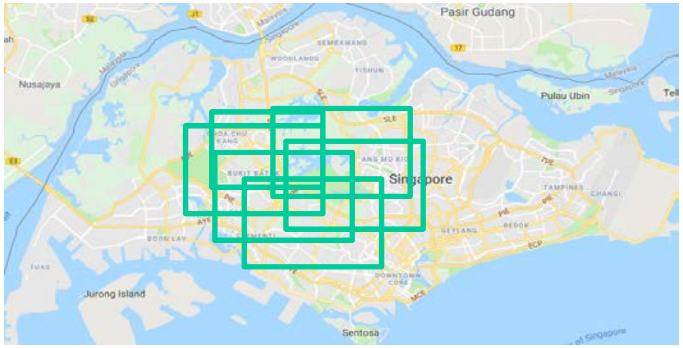
$$sim(R_q, R_i) \geq sim(R_q, R_j), \forall R_i \in \mathcal{R}, \forall R_j \notin \mathcal{R}.$$

Motivation

- Challenges
  - **C1**: How to **represent a region** to define region similarity.
    - Region ≠ Bag of POIs.
    - POIs in a region have influence to each other.
    - Considering both POI category & their spatial relations.



  - **C2**: How to **efficiently search** over large geographical space.
    - Billions of ways to place a region in a city with various sizes and scales.



Contributions

- To support similar region search:

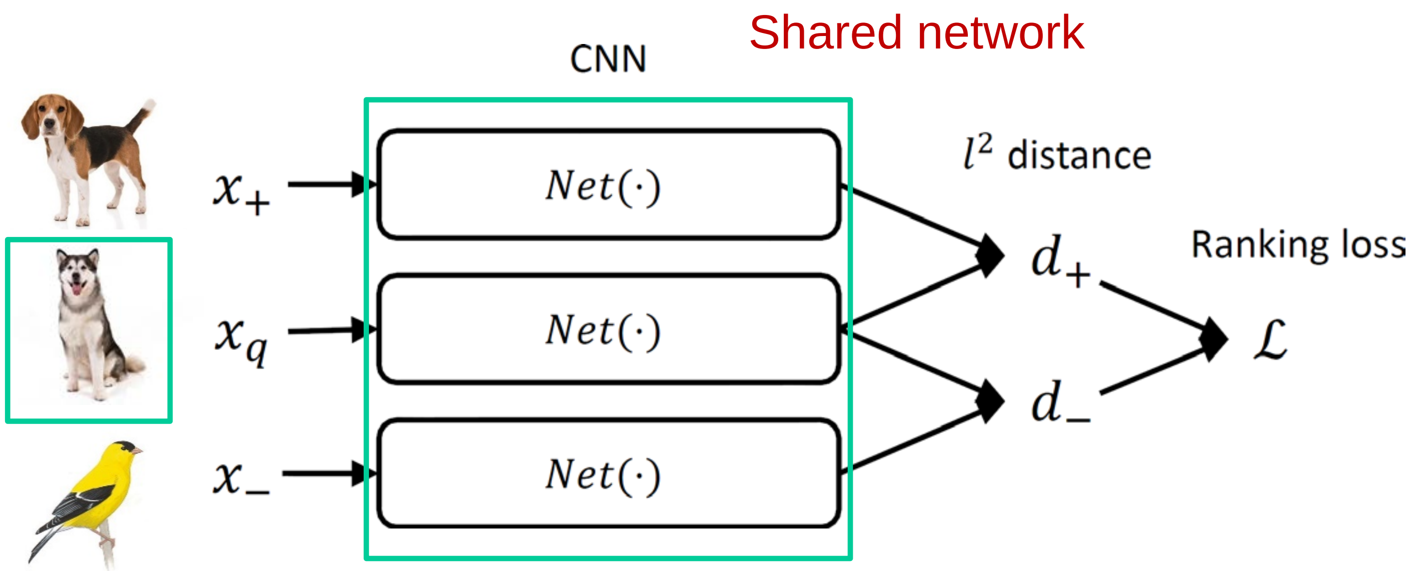
  - **Solving C1**: We propose a **deep metric learning model** to learn region similarity, which consider both POI category and their spatial relations.

  - **Solving C2**: We propose an **efficient search algorithm** (45× faster than the best baseline).

  - **Solving C2**: We propose an approximation method that can make a trade-off between accuracy and efficiency.

Preliminary

- Deep metric learning method: triplet network
  - A shared Convolutional Neural Network.



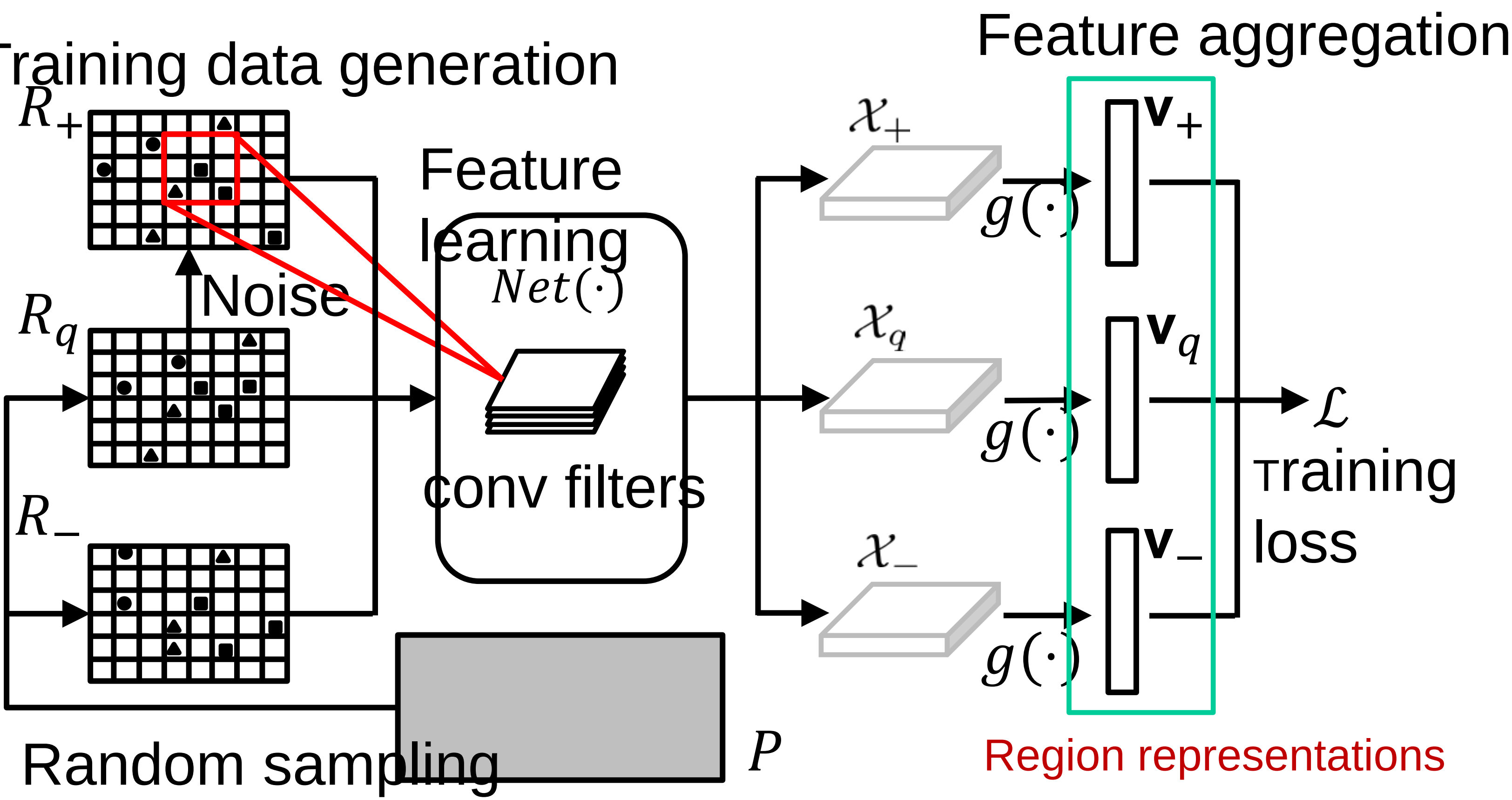
$$d_+ = ||Net(x_q) - Net(x_+)||_2 \qquad d_- = ||Net(x_q) - Net(x_-)||_2$$

$$\boxed{\mathcal{L} = \sum_{(x_q, x_+, x_-)} max\{0, d_+ - d_- + \delta\} + \lambda ||Net(\cdot)||_2}$$

Contrastive loss

Deep Metric Learning for Regions

- Learning from self-similarity
  - $sim(R_1, R_2) = \frac{1}{1+||\mathbf{v}_1-\mathbf{v}_2||_2}$.



Experiments

- Case study: finding top-5 similar regions:

Query: Bukit Timah Nature Reserve

Our method **Baseline**

Smart Nation Applications

GeoSpatial Data Mining

- POI recommendation & prediction
- Interactive exploration geospatial data
- Knowledge graph for locations
- Trajectory representation and similarity
- Speed, travel time, route prediction
- Region search, (e.g., burst region)
- Region exploration (topic, crowdness)

Querying and indexing spatio-temporal data

- Snapshot queries (OLTP, OLAP)
- Continuous queries

Distributed streaming systems

- Distributed load balance
  Distributed materialized view
- Index &query optimizer
- Machine learning techniques

Big static/streaming geo-spatial + X (e.g., text, temporal) data

foursquare yelp Grab Google places twitter Hi, stranger Instagram Fast beautiful photo sharing

NANYANG TECHNOLOGICAL UNIVERSITY

Acknowledgement to my students and collaborators: Tao Guo, Xiucheng Li, Yiding Liu, Di Yao, Kaiqi Zhao.

Thank You !

Q & A?

Demo URL: http://spatialkeyword.sce.ntu.edu.sg/index.html#