An Efficient Sampling Method for Characterizing Points of Interests on Maps

Team 1

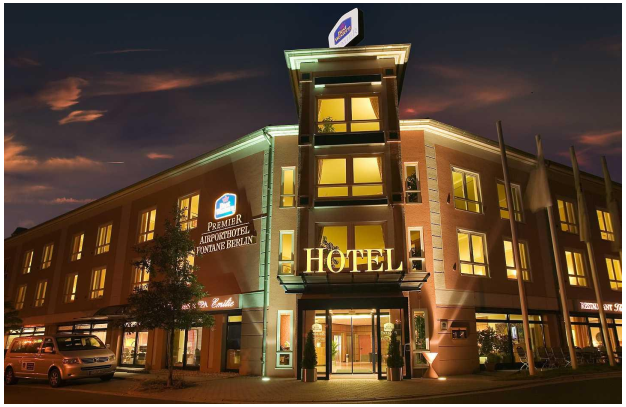
Qiuyi Hong, Caidan Liu, Zhenhua Li,
Jiaqi Liu, Shaowei Gong

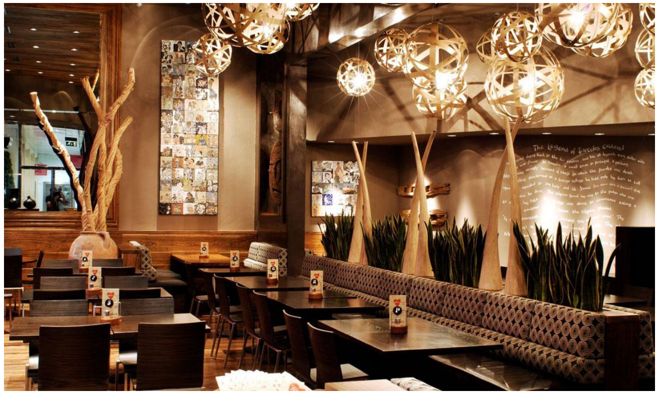
Outline

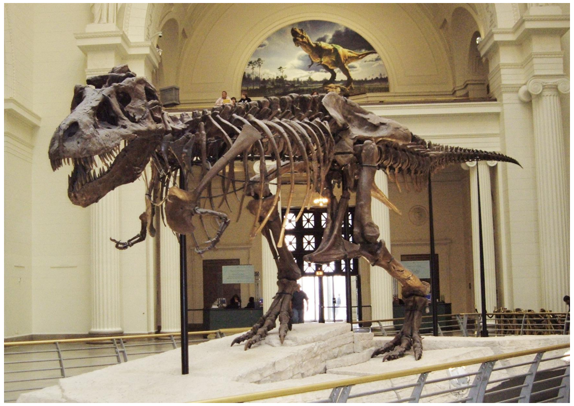
- Background and formulated problem
- Challenges
- Our methods (i.e., RRZI and RRZIC)
- Experiments and Applications
- Conclusions

Note: this slide is from the conference presentation by Yanhua Li

Points of Interests





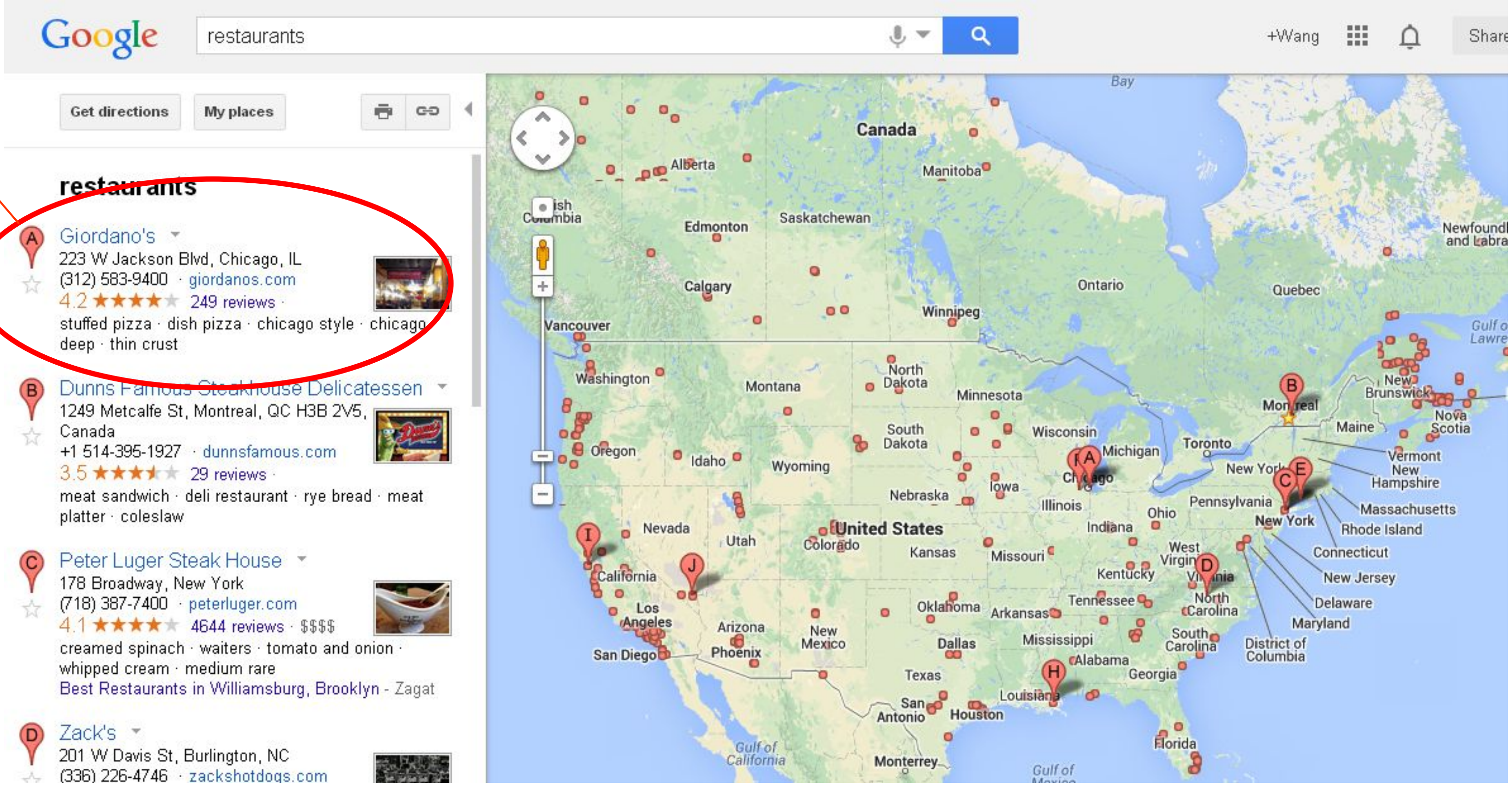


Note: this slide is from the conference presentation by Yanhua Li

Background

- Google Maps: keyword “restaurant”

A PoI: location, rating, flavor, reviews, …



restaurants

Get directions · My places

restaurants

A Giordano's
223 W Jackson Blvd, Chicago, IL
(312) 583-9400 · giordanos.com
4.2 ★★★★★ 249 reviews ·
stuffed pizza · dish pizza · chicago style · chicago deep · thin crust

B Dunns Famous Steakhouse Delicatessen
1249 Metcalfe St, Montreal, QC H3B 2V5, Canada
+1 514-395-1927 · dunnsfamous.com
3.5 ★★★★★ 29 reviews ·
meat sandwich · deli restaurant · rye bread · meat platter · coleslaw

C Peter Luger Steak House
178 Broadway, New York
(718) 387-7400 · peterluger.com
4.1 ★★★★★ 4644 reviews · $$$$
creamed spinach · waiters · tomato and onion · whipped cream · medium rare
Best Restaurants in Williamsburg, Brooklyn - Zagat

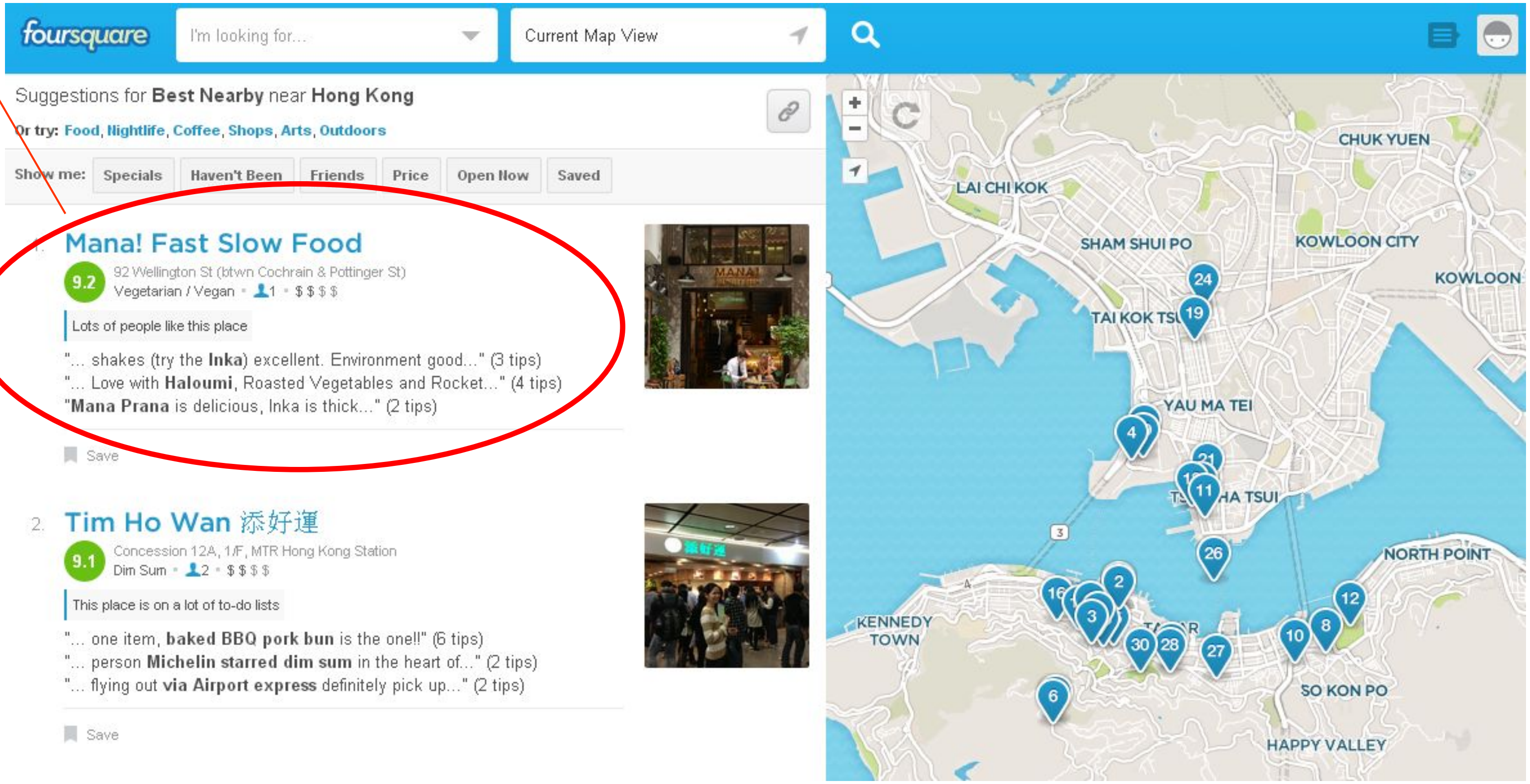
D Zack's
201 W Davis St, Burlington, NC
(336) 226-4746 · zackshotdogs.com

Note: this slide is from the conference presentation by Yanhua Li

Background

- Foursquare: food, nightlife, coffee, shopping, sights, arts, outdoors, …

A PoI:
category
location,
rating,
Reviews,
#check-ins
…



Suggestions for **Best Nearby** near **Hong Kong**

Or try: Food, Nightlife, Coffee, Shops, Arts, Outdoors

Show me: Specials | Haven't Been | Friends | Price | Open Now | Saved

1. Mana! Fast Slow Food
9.2 92 Wellington St (btwn Cochrain & Pottinger St)
Vegetarian / Vegan · 1 · $ $ $ $
Lots of people like this place
"... shakes (try the **Inka**) excellent. Environment good..." (3 tips)
"... Love with **Haloumi**, Roasted Vegetables and Rocket..." (4 tips)
"**Mana Prana** is delicious, Inka is thick..." (2 tips)
Save

2. Tim Ho Wan 添好運
9.1 Concession 12A, 1/F, MTR Hong Kong Station
Dim Sum · 2 · $ $ $ $
This place is on a lot of to-do lists
"... one item, **baked BBQ pork bun** is the one!!" (6 tips)
"... person **Michelin starred dim sum** in the heart of..." (2 tips)
"... flying out **via Airport express** definitely pick up..." (2 tips)
Save

Note: this slide is from the conference presentation by Yanhua Li

Formulated Problem

- Objective 1
  - ➢ Sum aggregate

$$f_s(\mathsf{P}) = \sum_{p \in \mathsf{P}} f(p)$$

Example 1:

f(*p*) is the number of rooms a hotel *p* has,

$f_s(P)$ is the total number of rooms in the area of interest

Example 2:

$f(p)=1$

$f_s(P)$ is the total number of hotels in the area of interest

Note: this slide is from the conference presentation by Yanhua Li

Formulated Problem

- Objective 2
  - ➢ Average aggregate

$$f_s(\mathrm{P}) = \frac{1}{|\mathrm{P}|}\sum_{p\in\mathrm{P}} f(p)$$

Example:
f(*p*) is the average price of a hotel *p*,
$f_s(P)$ is the average price of hotels in the area of interest

Note: this slide is from the conference presentation by Yanhua Li

Formulated Problem

- Objective 3
  - PoI distribution

$$\theta_j = \frac{1}{|\mathrm{P}|} \sum_{p \in \mathrm{P}} 1(L(p) = l_j), \quad j = 1, 2, \ldots$$

Example:

L(*p*) is the star rating of *p*

θ is the star rating distribution of hotels in the area of interest

Note: this slide is from the conference presentation by Yanhua Li

Formulated Problem

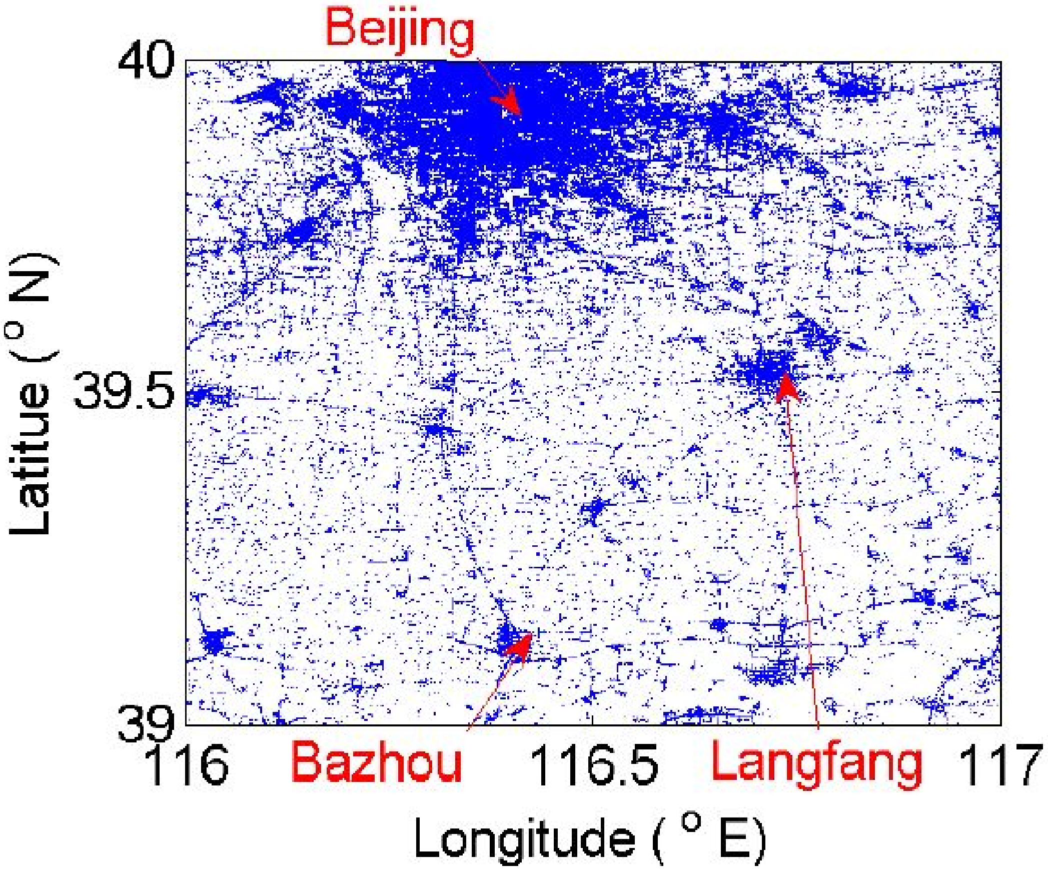
- We focus on designing efficient **sampling** methods to estimate the above statistics, since it is costly to collect PoIs within a large area.

  For example, to collect PoIs within 14 cities in Foursquare, Li et al. spent almost two months using 40 machines in parallel.

Note: this slide is from the conference presentation by Yanhua Li

Challenges

- The underlying distribution of PoI is unknown

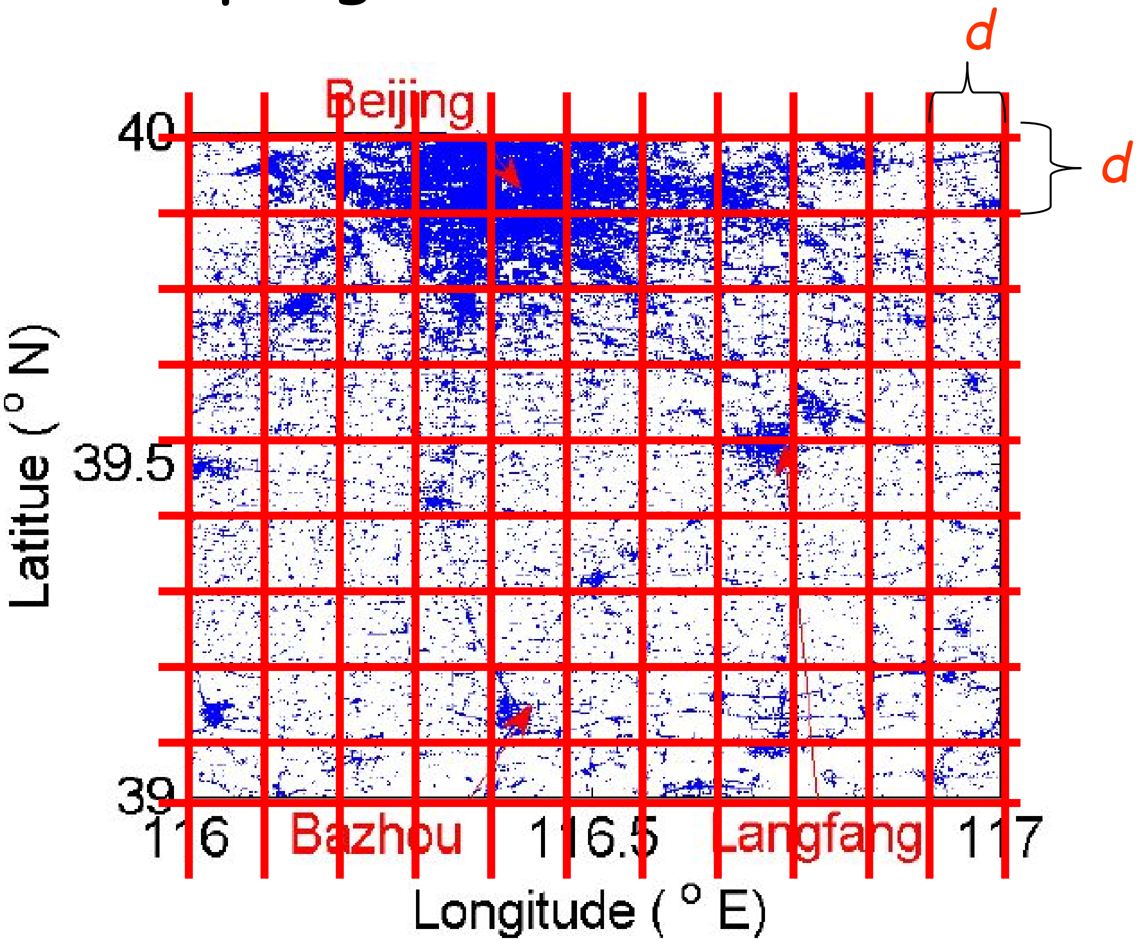


Note: this slide is from the conference presentation by Yanhua Li

Challenges

- Straightforward sampling method

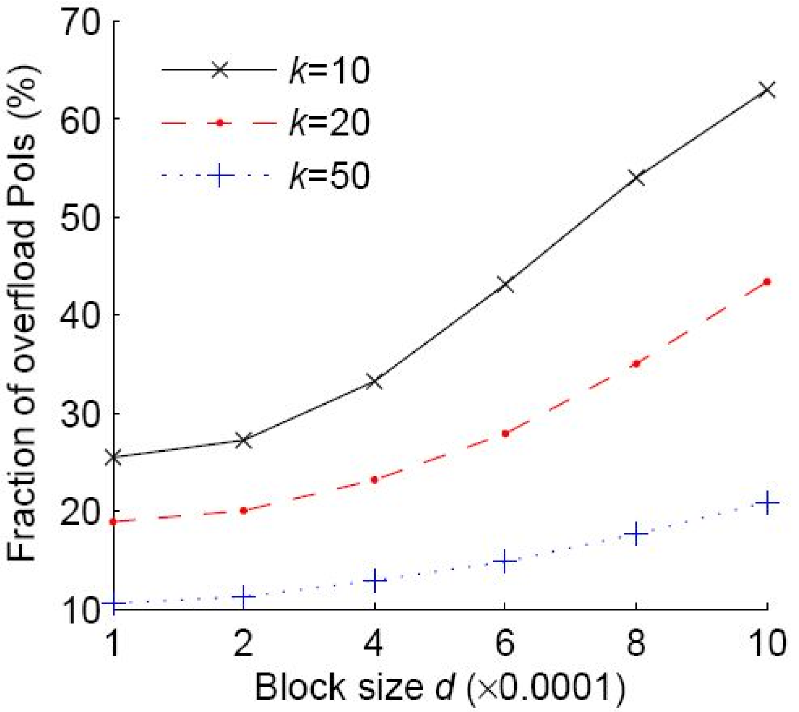
1. Split the region into small sub-regions evenly
2. Random sample sub-regions uniformly

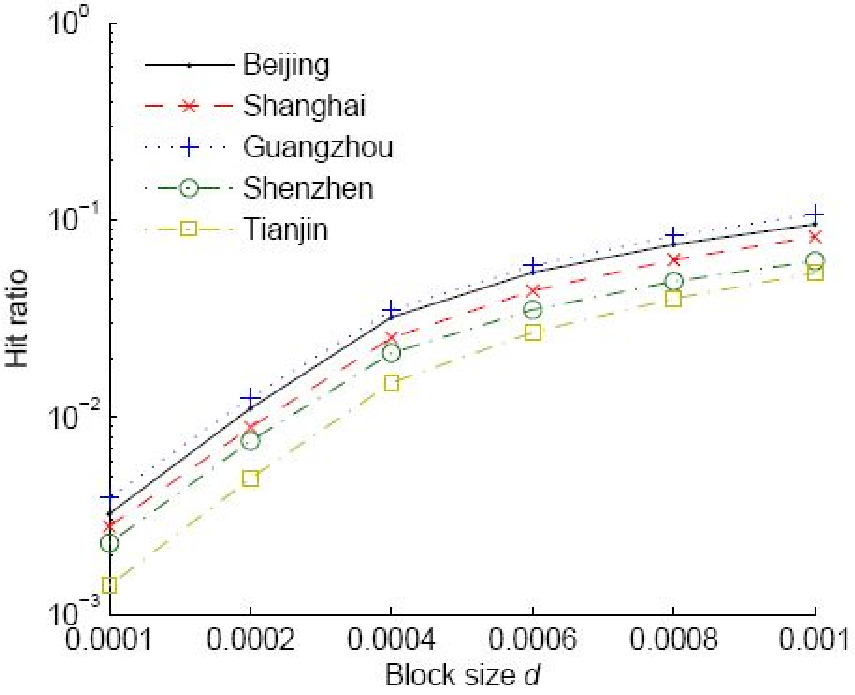


Note: this slide is from the conference presentation by Yanhua Li

Challenges

- Drawbacks of straightforward sampling method
  - A sub-region may include a large fraction of PoIs
  - Many empty sub-regions for small *d*

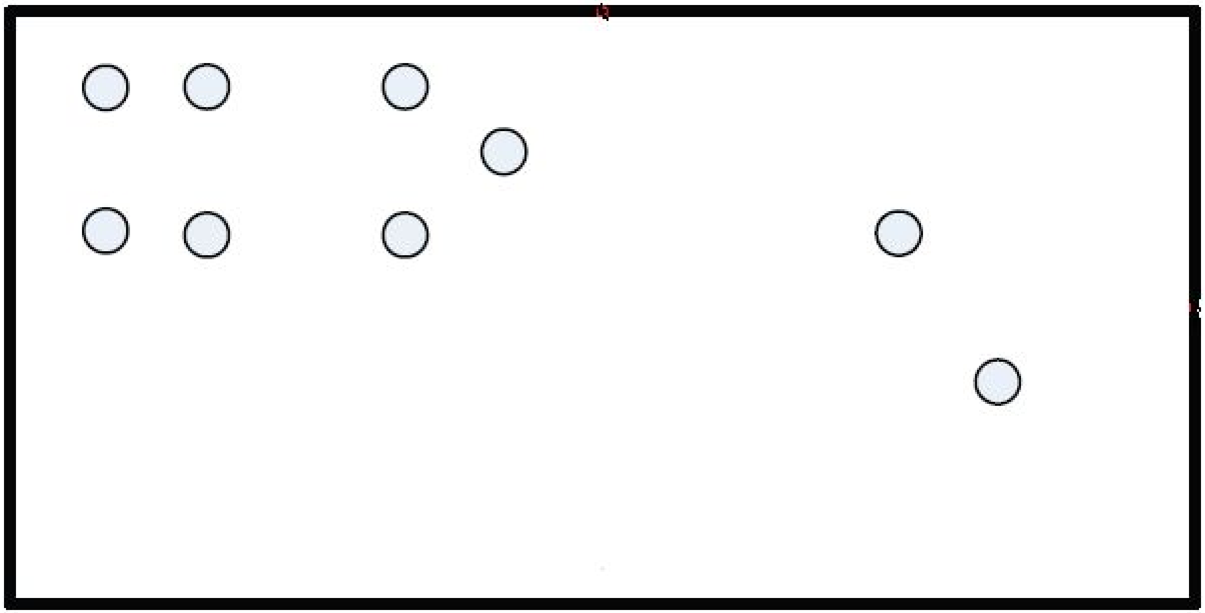




Note: this slide is from the conference presentation by Yanhua Li

Our method: Random Region Zoom-in on Maps

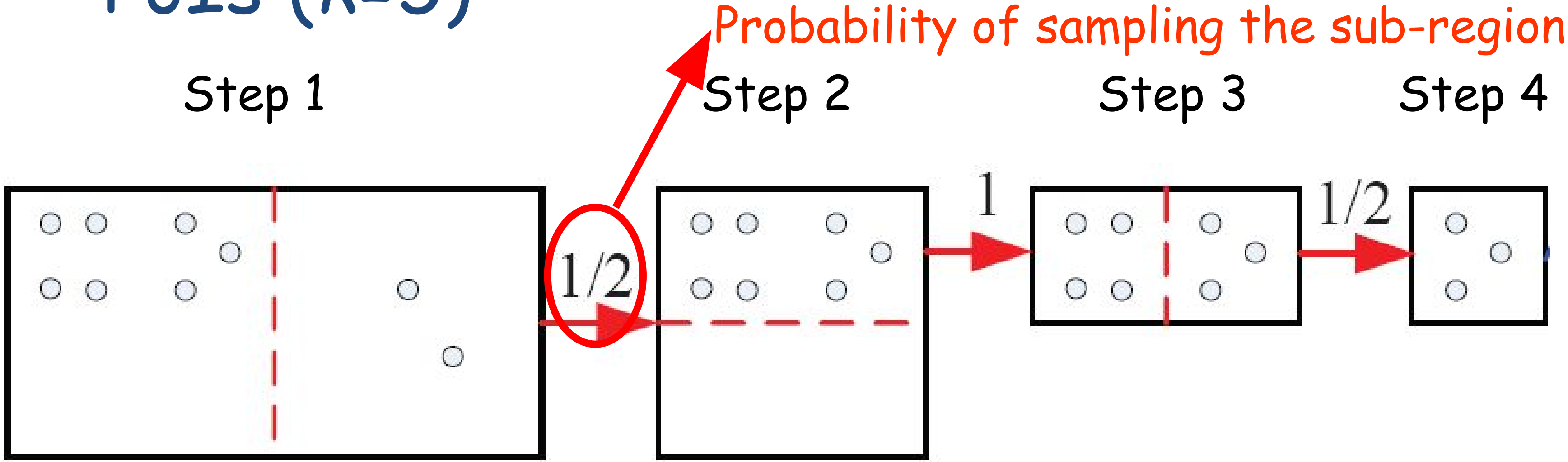
- RRZI(*A*)
  - ➢ Input: *A*, the area of interest
  - ➢ Output: a random sub-region Q with PoIs less than *k and* *τ(Q,A)*



Note: this slide is from the conference presentation by Yanhua Li

Our method: Random Region Zoom-in on Maps

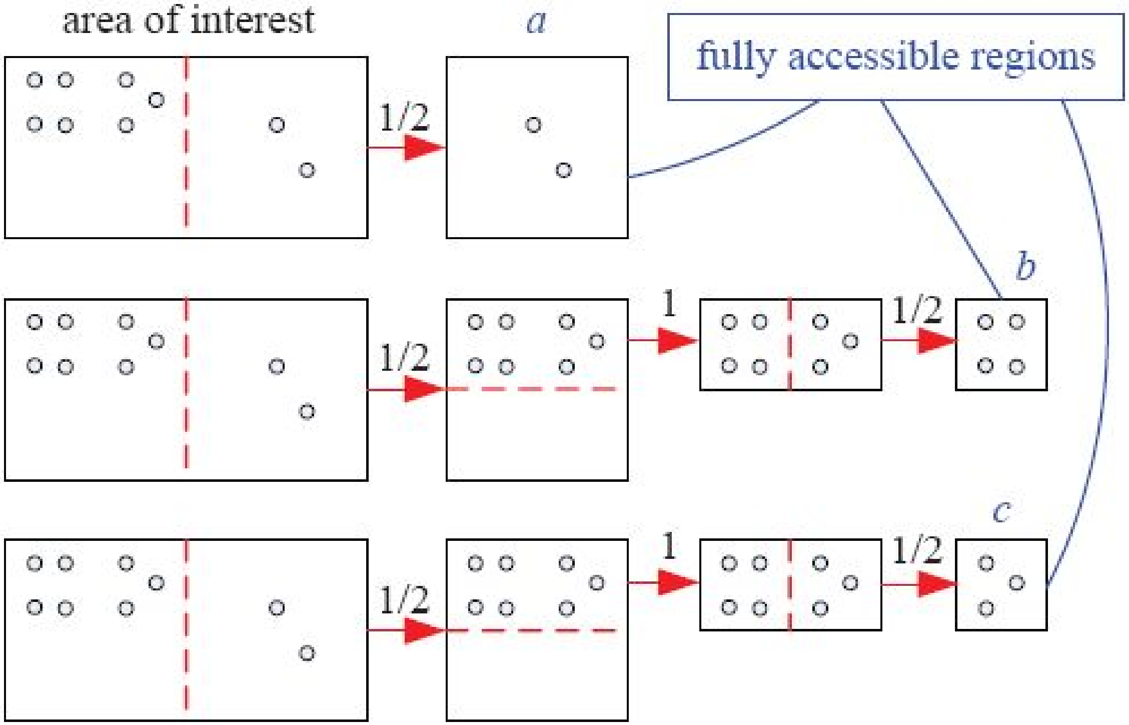
- RRZI(*A*): At each step, RRZI divides the current queried region into two sub-regions and randomly selects a non-empty sub-region to zoom-in when it contains more than or equal to *k* PoIs (*k*=5)



Note: this slide is from the conference presentation by Yanhua Li

Our method: Random Region Zoom-in on Maps

- RRZI(*A*): probability of sampling a sub-region with PoIs less than 5
p(*a*)=1/2, p(*b*)=1/4, p(*c*)=1/4



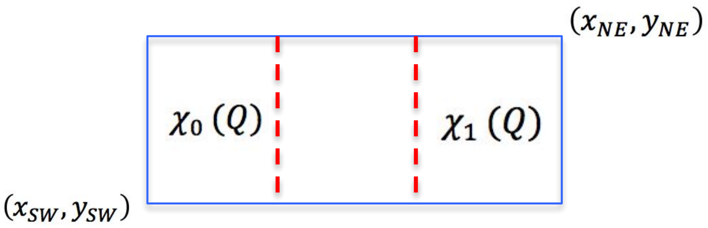
Note: this slide is from the conference presentation by Yanhua Li

Our method: Random Region Zoom-in on Maps

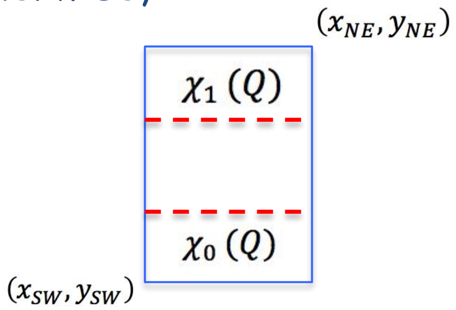
RRZI(A): three critical questions

- To divide Q into two non-overlapping regions Q_0 and Q_1

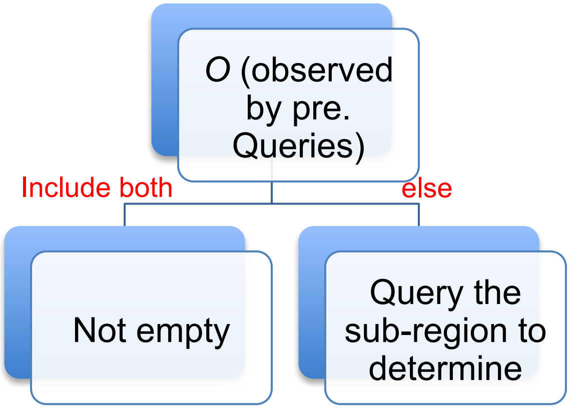
  If $|x_{SW} - x_{NE}| \geq \beta(x_{SW} - x_{NE})|y_{SW} - y_{NE}|$



Otherwise,



- To determine whether $\chi_0\,(Q)$ and $\chi_1\,(Q)$ are empty regions or not using a minimum number of queries.


O (observed by pre. Queries)
Include both
else
Not empty
Query the sub-region to determine

- Does RRZI sample PoIs uniformly? If not, how to remove the sampling bias?

  No. Use counter

Our method: Random Region Zoom-in on Maps

RRZI(A): Estimates the sum aggregate

$$\tilde{f}_s(\mathbb{P}) = \frac{1}{m}\sum_{i=1}^{m}\sum_{p\in P(r_i)}\frac{f(p)}{\tau(r_i,\mathbb{A})}.$$

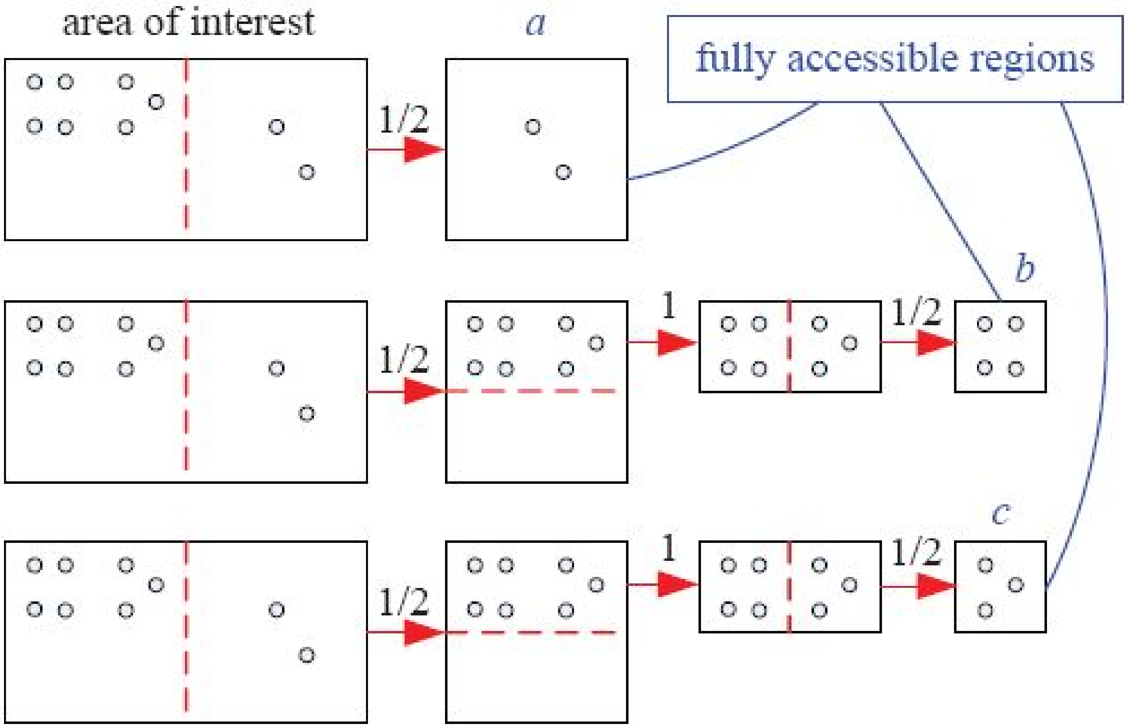
Note:

- m: number of sampled fully accessible regions
- $T(r_i,A)$: the probability of sampling a fully accessible region Q from A.
- $P(r_i)$: set of PoIs within a region r_i
- $f(p)$: the function related with p, e.g. # within p, constant 1, unit price, etc.

Example:

$$f_s(\mathrm{P}) = \frac{1}{3}(\frac{2}{1/2}+\frac{4}{1/4}+\frac{3}{1/4})=\frac{32}{3}\approx 11$$

- RRZI(*A*): probability of sampling a sub-region with PoIs less than 5 p(*a*)=1/2, p(*b*)=1/4, p(*c*)=1/4

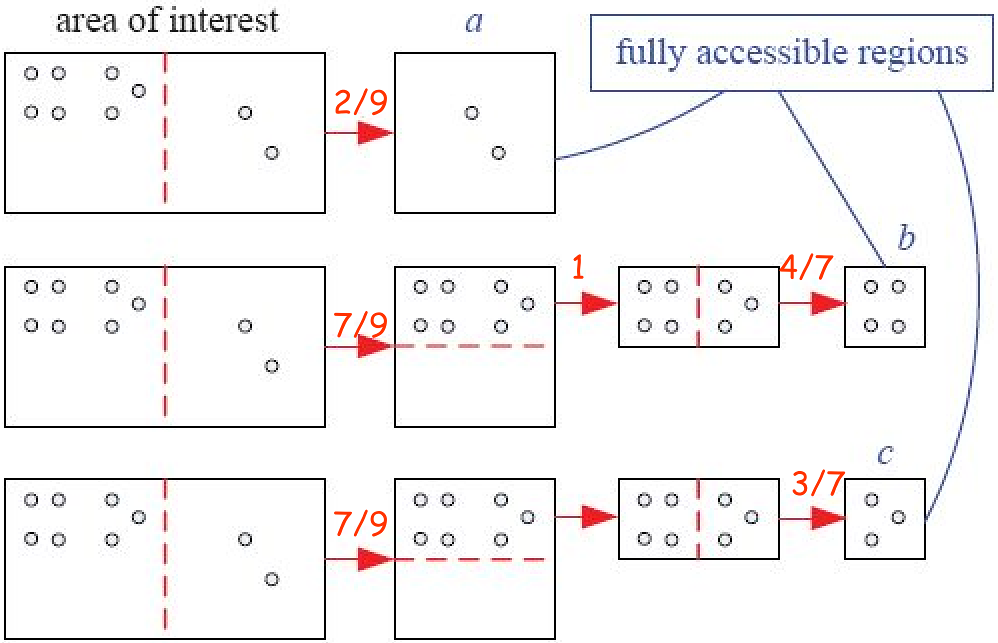


Note: this slide is from the conference presentation by Yanhua Li

Random Region Zoom-in on Maps With Count Information

- RRZIC(*A*): Sample sub-regions with probability proportional to the number of PoIs.

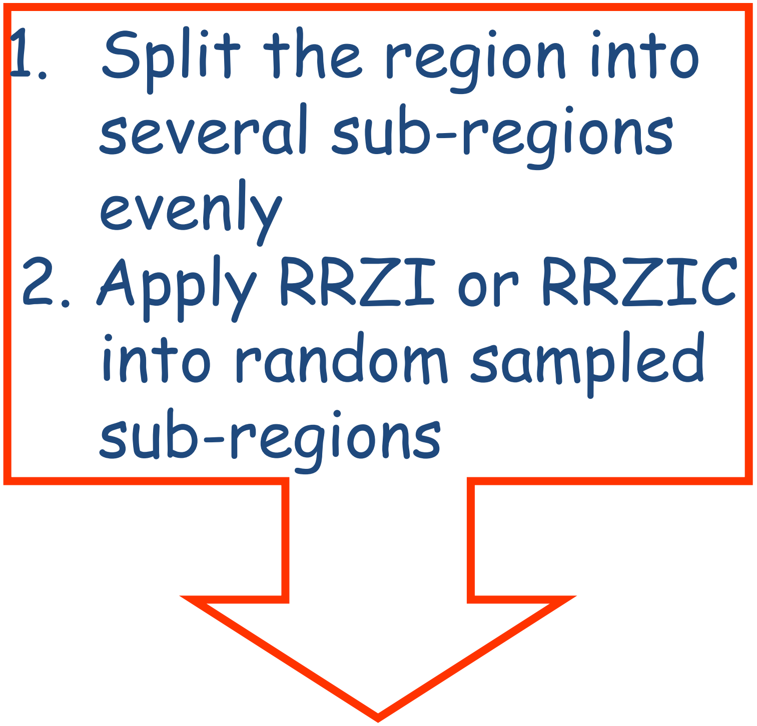
$$p(a)=2/9, p(b)=4/9, p(c)=3/9$$



Note: this slide is from the conference presentation by Yanhua Li

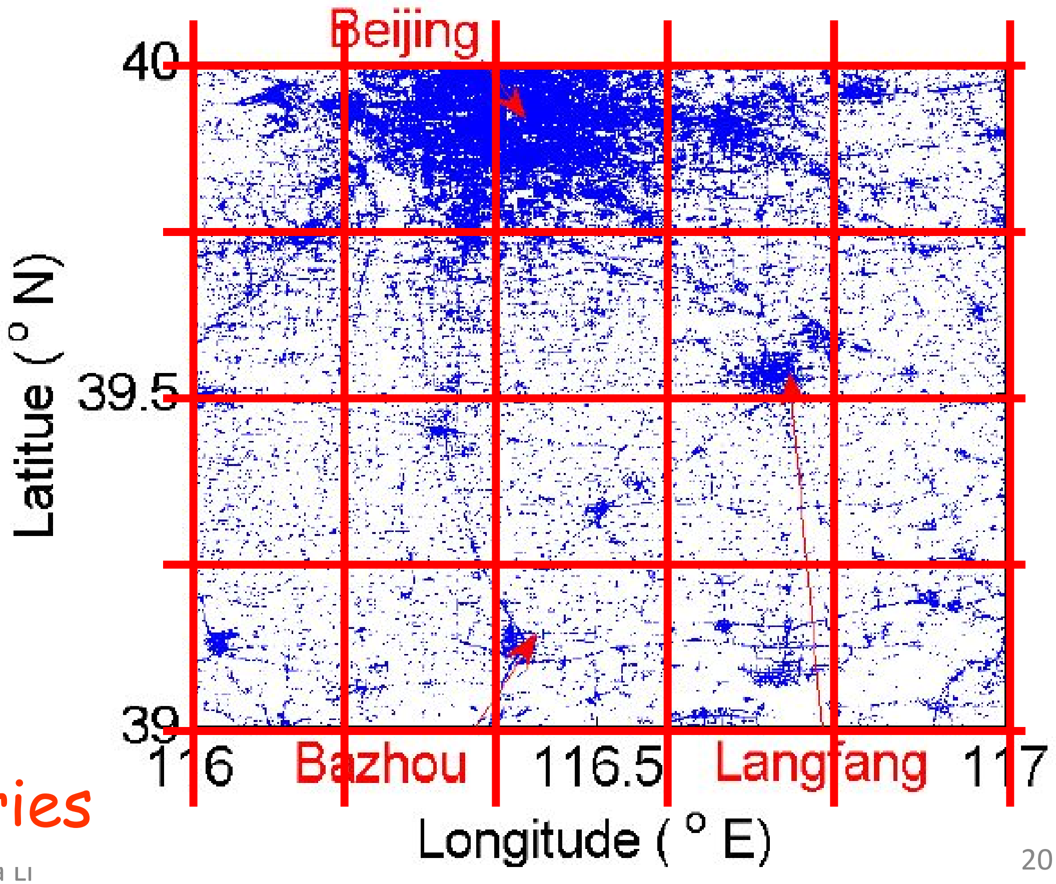
Our method: Mix Methods

- Mix methods: It's not necessary to apply RRZI and RRZIC into the entire area directly.



1. Split the region into several sub-regions evenly
2. Apply RRZI or RRZIC into random sampled sub-regions

Reduce the number of queries



Note: this slide is from the conference presentation by Yanhua Li

Measure the effect of Sampling

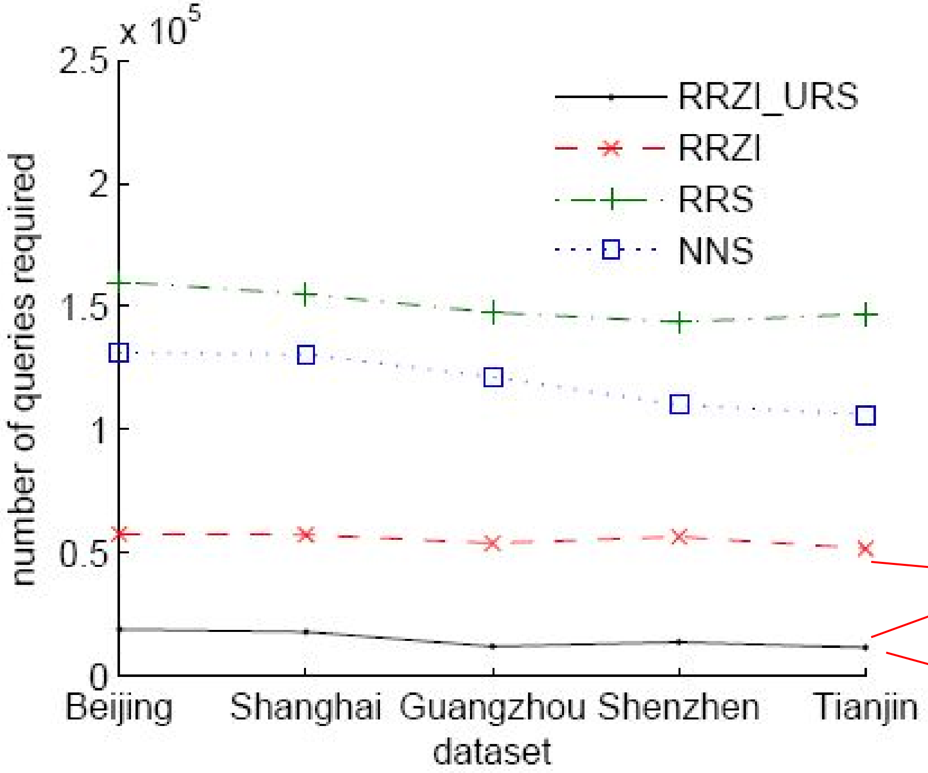
- NRMSE(normalized root mean square error): Eliminate the effects of unit and scale of data

- Control either the number of queries or error(NRMSE)

Note: this slide is from the conference presentation by Yanhua Li

Experimental Results

- The number of queries required to obtain an estimate of the number of PoIs with NRMSE less than 0.1



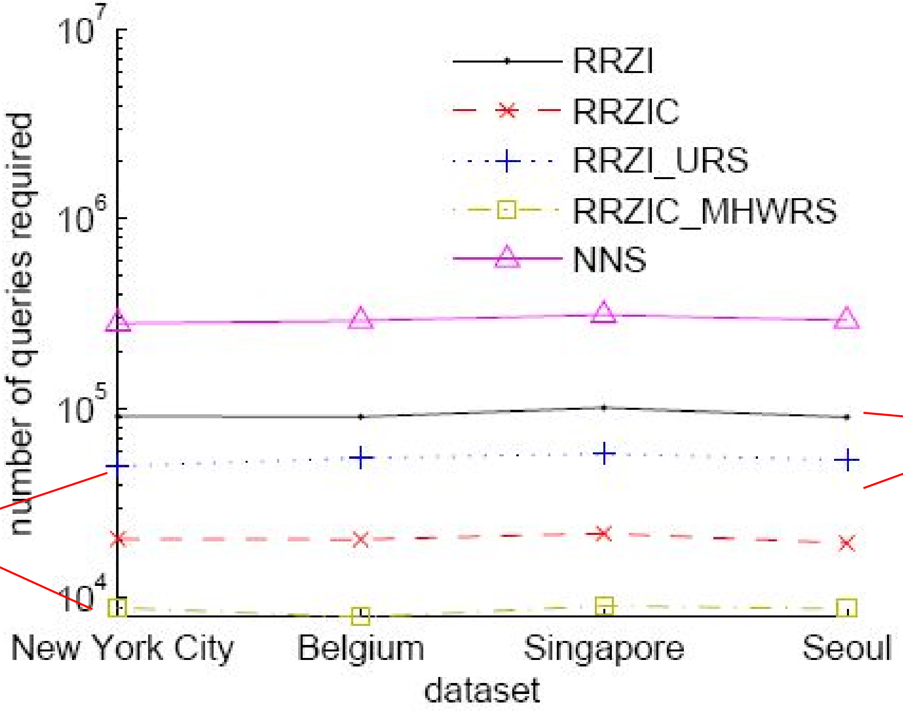
our methods

mix method

Note: this slide is from the conference presentation by Yanhua Li

Experimental Results

- The number of queries required to obtain an estimate of the average number of Foursquare check-ins with NRMSE less than 0.1



mix methods

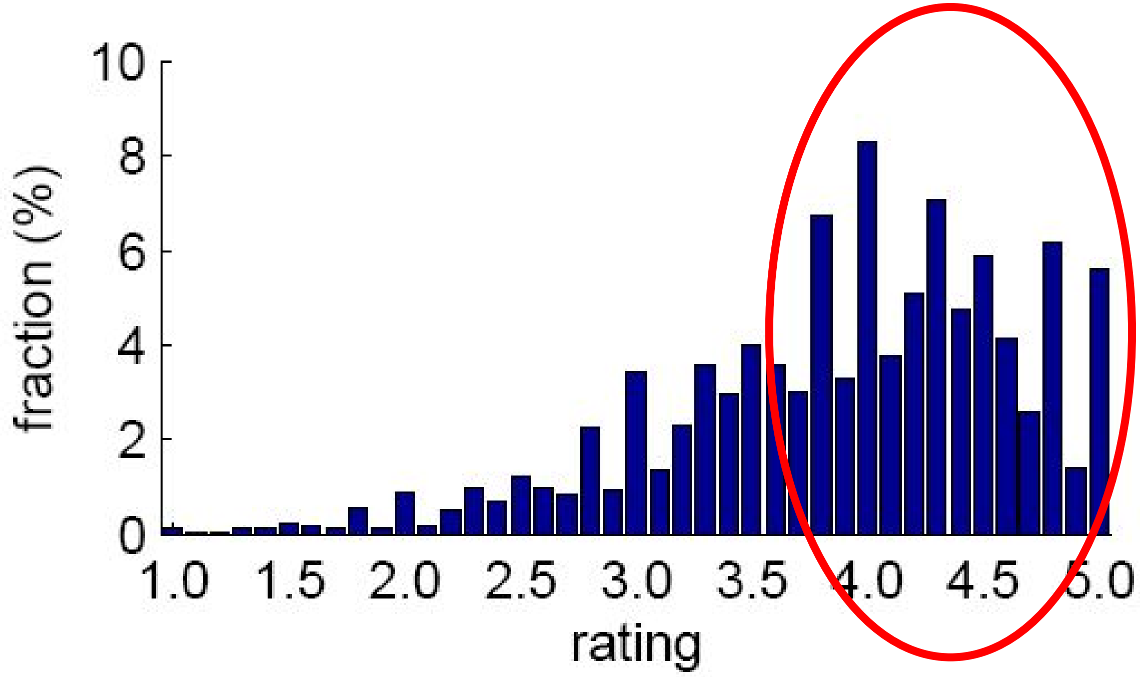
our methods not using PoI count information

our methods using PoI count information

Note: this slide is from the conference presentation by Yanhua Li

Real application on Google maps

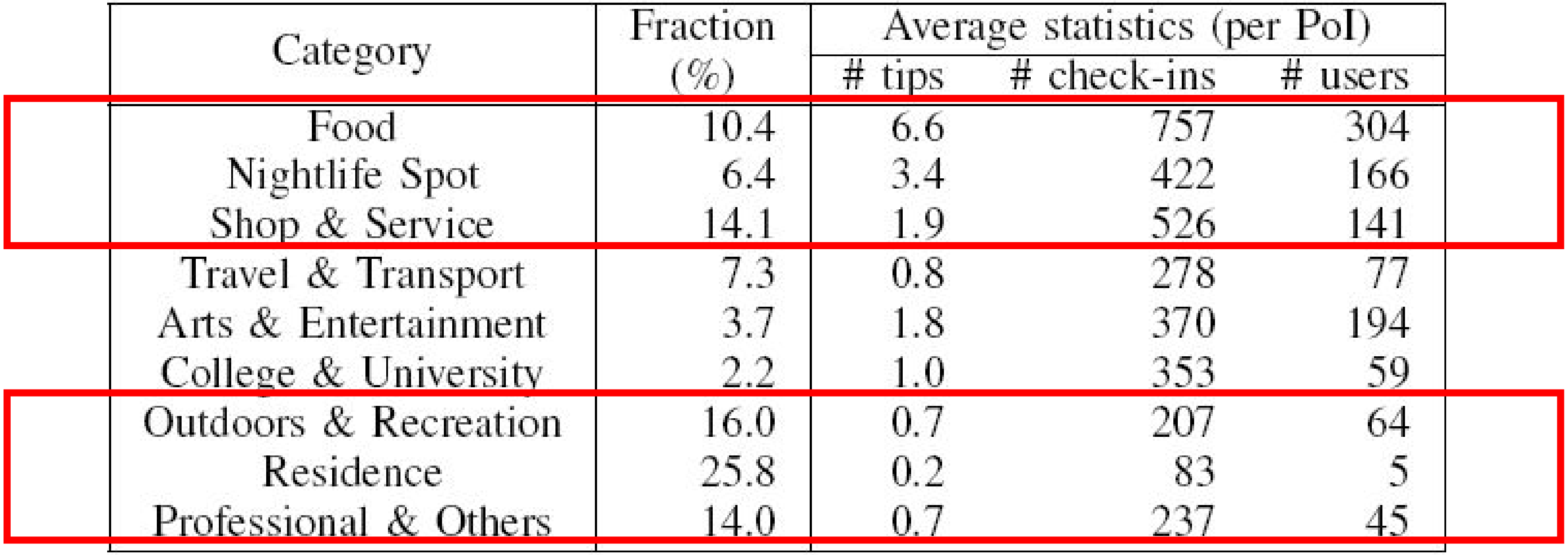
- Rating distribution of food-type PoIs within US.



Note: this slide is from the conference presentation by Yanhua Li

Real application on Foursquare

- Statistics of PoIs in US

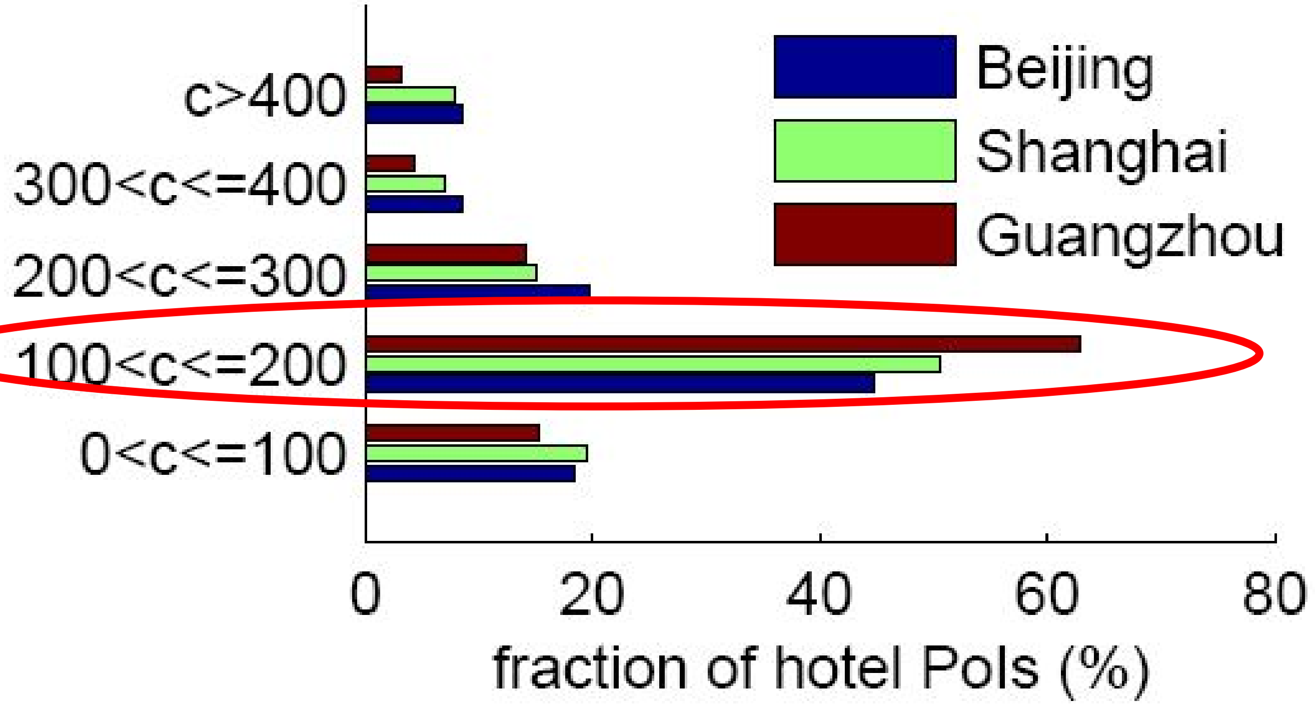


Category	Fraction (%)	Average statistics (per PoI)		
		# tips	# check-ins	# users
Food	10.4	6.6	757	304
Nightlife Spot	6.4	3.4	422	166
Shop & Service	14.1	1.9	526	141
Travel & Transport	7.3	0.8	278	77
Arts & Entertainment	3.7	1.8	370	194
College & University	2.2	1.0	353	59
Outdoors & Recreation	16.0	0.7	207	64
Residence	25.8	0.2	83	5
Professional & Others	14.0	0.7	237	45

Note: this slide is from the conference presentation by Yanhua Li

Real application on Baidu maps

- Distribution of hotel-type PoIs' prices per room per night.



Note: this slide is from the conference presentation by Yanhua Li

Conclusions

- Random zoom-in methods are efficient
- Mix methods are more efficient
- Methods (e.g., RRZIC) using PoI count information are more accurate.

Note: this slide is from the conference presentation by Yanhua Li

Thanks !