A Big Data-Oriented Dynamic R-tree Forest (DRF) For Spatial Object Query

Liming Zhang, Manqi Li, Chengbi Liu
{lzhang22, mli14, cliu19}@gmu.edu
Advisor: Dr. Andreas Züfle

GEORGE MASON UNIVERSITY

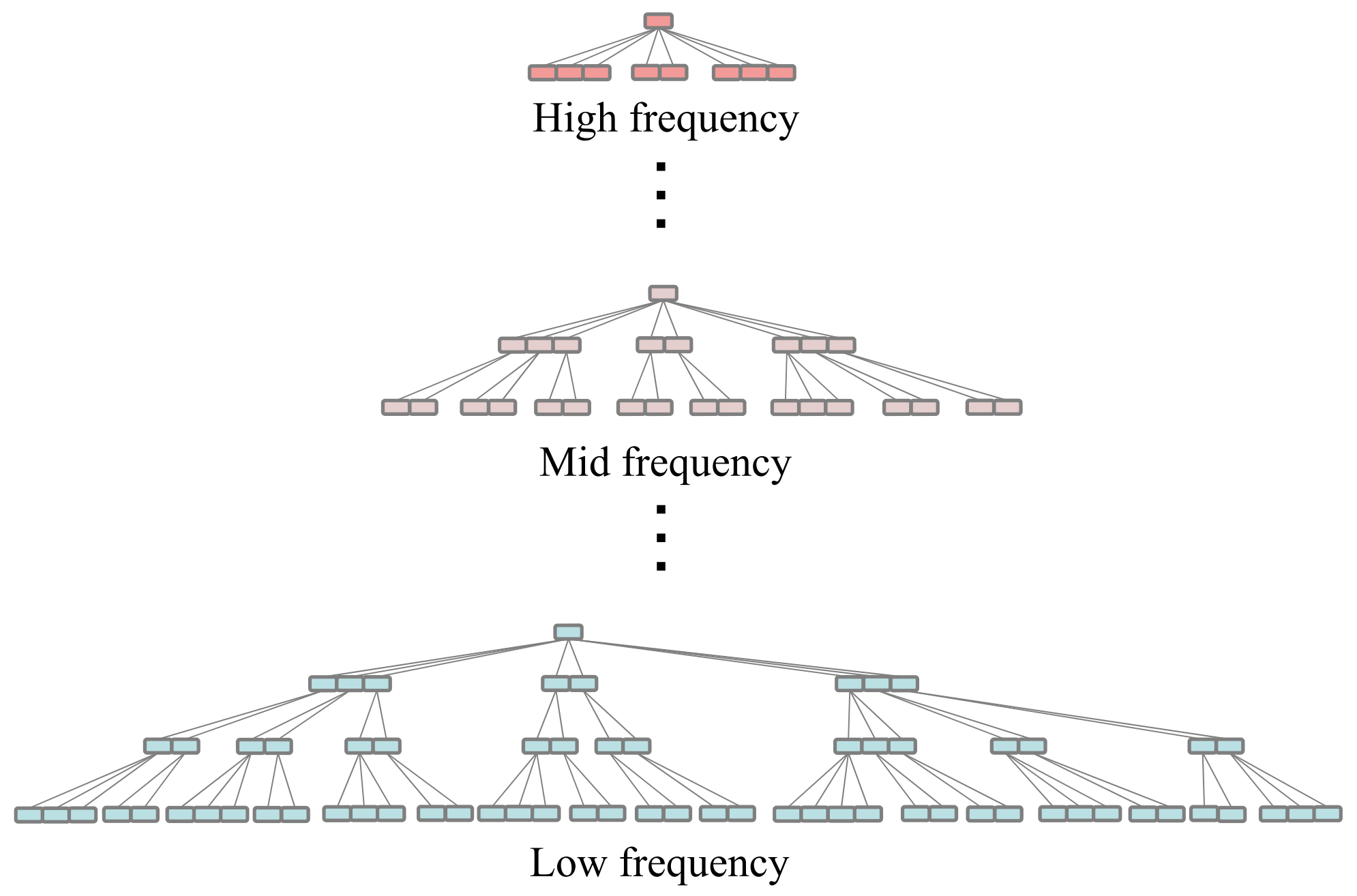
Department of Geography and GeoInformation Science, George Mason University, Fairfax, VA 22030

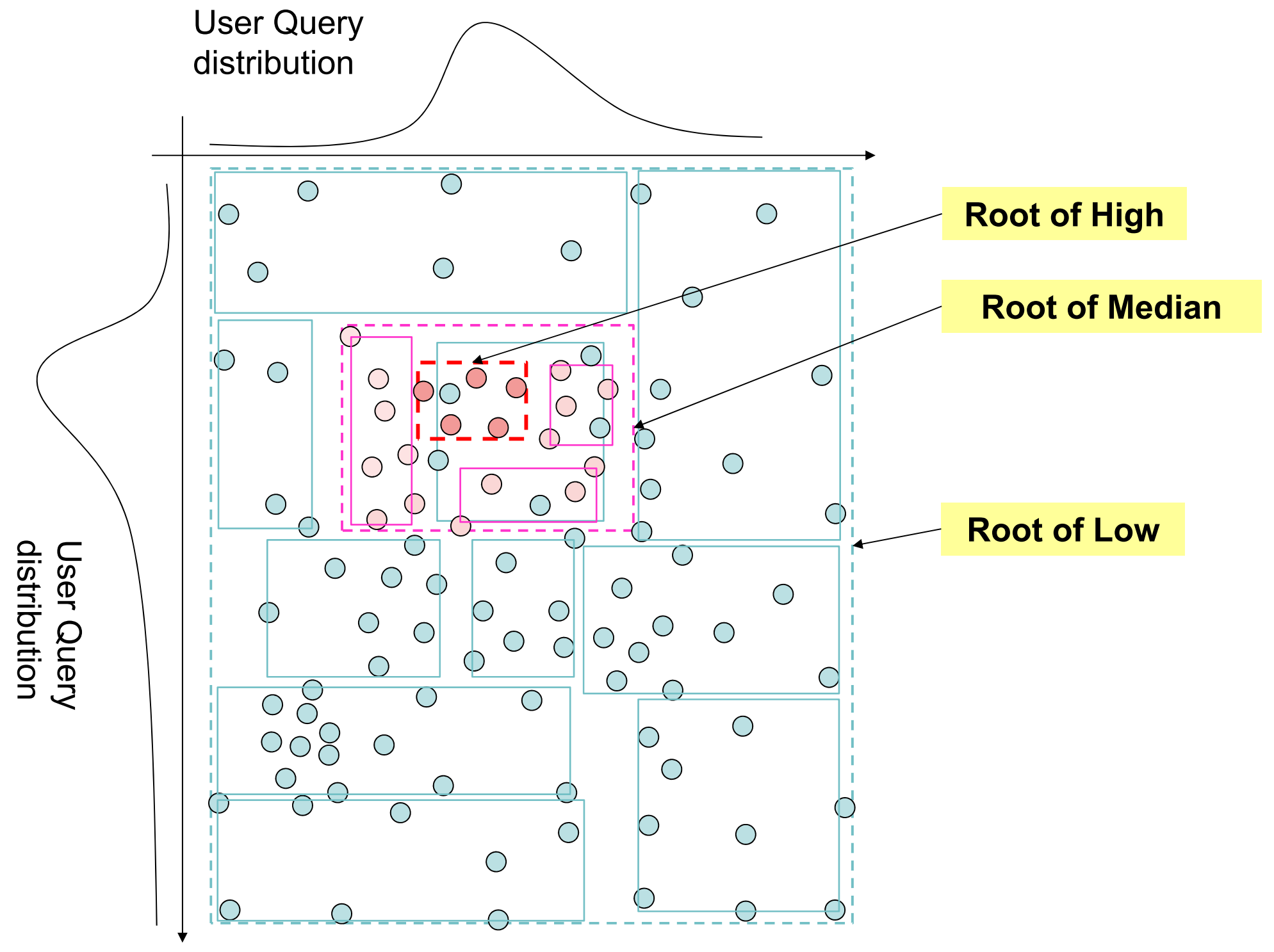
Motivation

- ❑ In the Big Data era, 95% users are interested in 5% of data, intuitively.
- ❑ For spatial data, the mostly used query algorithms, for example R-tree, have potential to be optimized by users' behavior nowadays.
- ❑ We intend to incorporate users behavioral pattern as a third dimension into the traditional R-tree model for improving spatial query efficiency.

Innovation

To improve the efficiency of current R-tree, we propose a Dynamic R-tree Forest (DRF) algorithm. The mostly used R-tree structure was invented in 1984 by Antonin Guttman aiming to optimize the efficiency of single-user query. Basically, he made an assumption that users had a uniform behavior when performing spatial queries. Based on this assumption, Guttman designed R-tree to be balanced and different data are not given preference as being queried. However, when we deal with queries from millions of users simultaneously, it is necessary to consider users behavioral pattern, and optimize the system holistically. In our research, we made an assumption that users behavior is not uniform. Using this information, we can improve the pruning efficiency of R-trees so that we can have quicker queries. There are other research groups had also done some work in this area. They used unbalanced R-trees. However, considering that the traditional R-tree is a balanced tree structure, we propose a different method to use a forest of balanced R-trees.





DRF Algorithm

DRF algorithm is a forest of balanced R-trees. The indexing of data records the information of query counts. The forest has a structure to dynamically assign each data point to different R-trees based on this count information of queries. When data has a query count above certain threshold, it migrates to another R-tree, which is for higher-query-count data. Each tree is still a balanced tree, which optimizes query inside itself. Through this dynamic promotion mechanism, data with higher query counts are separated from other data with lower query counts. For implementations, we have used an array-based data structure.

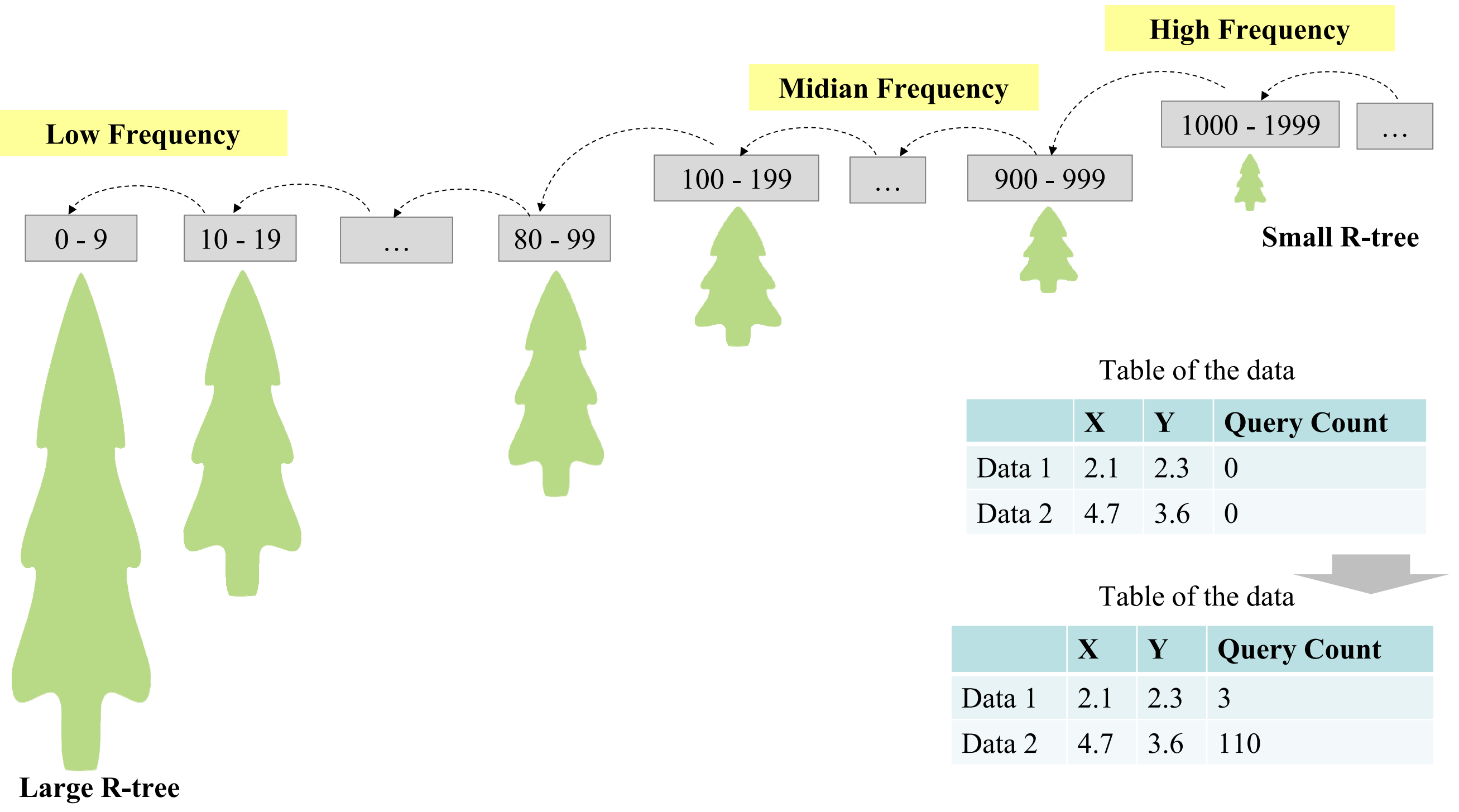


Table of the data

	X	Y	Query Count
Data 1	2.1	2.3	0
Data 2	4.7	3.6	0

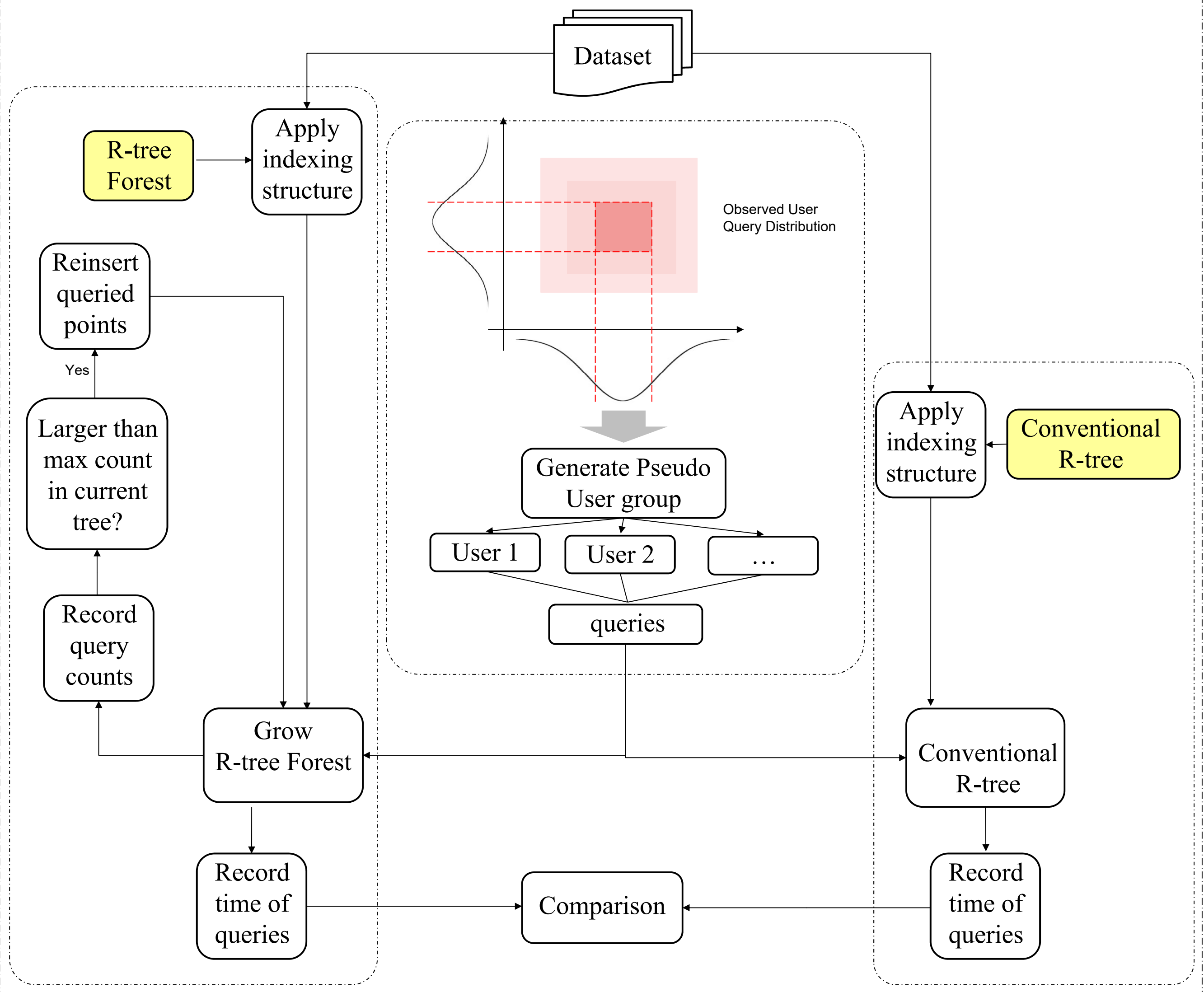
Table of the data

	X	Y	Query Count
Data 1	2.1	2.3	3
Data 2	4.7	3.6	110

Framework

Our algorithm will be implemented to compare our approach with conventional R-tree:

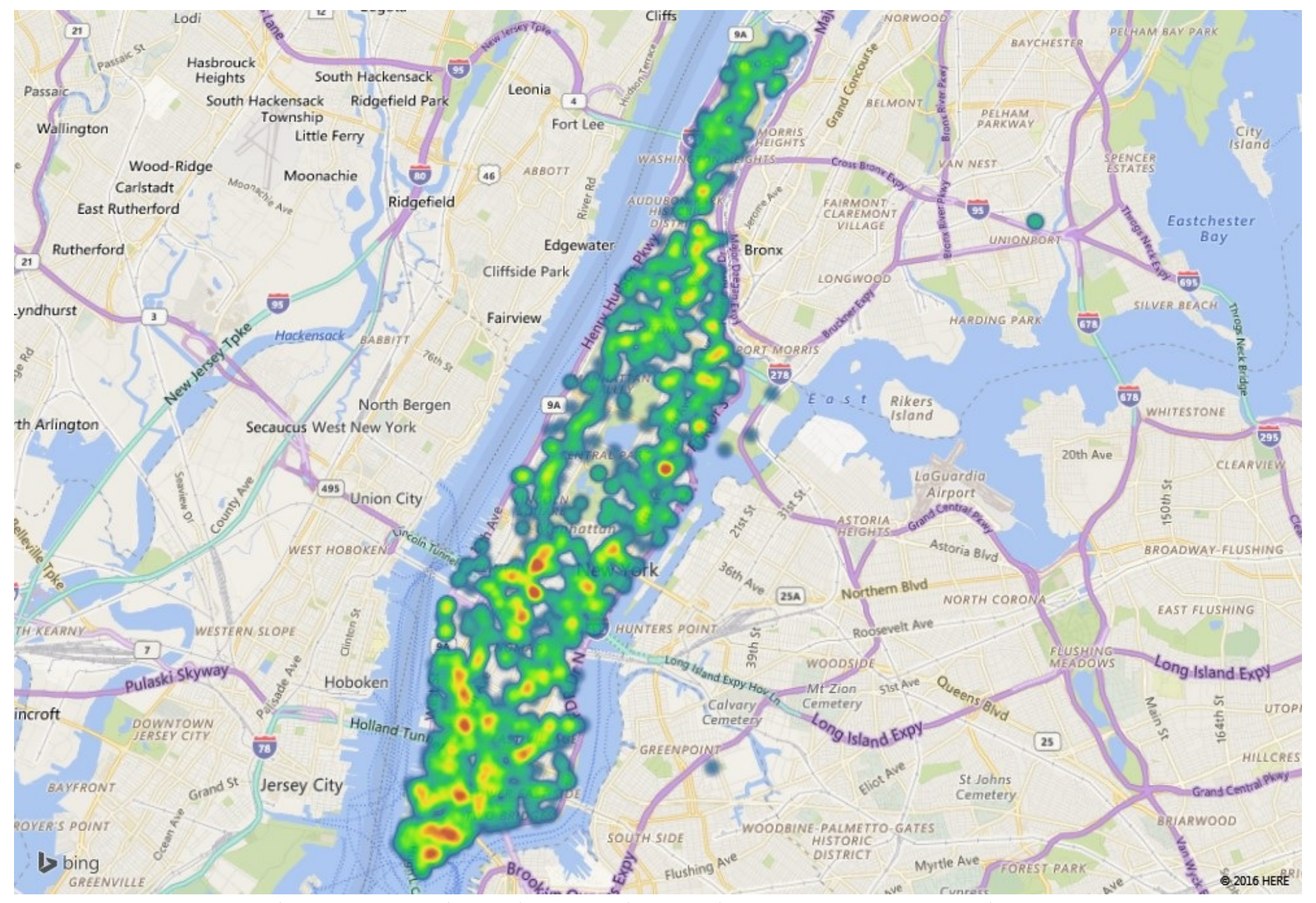
- ❑ a group of pseudo user queries are generated, according to certain spatial distribution
- ❑ a real dataset is used to compare performances between our algorithm and conventional R-tree



Potential Applications

Leveraging the fruitful information Big Data has provided us, our method could be applied to various real-life scenarios, so as to optimize the query speed and results. Potential applications include:

- ❑ point of interest retrieval on maps
- ❑ natural disasters and humanitarian missions
- ❑ restaurant and hotel recommendation systems
- ❑ location-explicit and user-targeted advertising



A heatmap showing Point Of Interest at Manhattan.

Specifically, we pay our attentions to the use of social media, a prevailing big data phenomenon in the Web 2.0 era. Twitter, a popular social media in the form of microblogging, has earned our particular interest because of:

- ❑ large volume and capacity: 317 million monthly active users and 500 million tweets per day;
- ❑ easy availability and low latency: Twitter APIs provides free and easy access to the real-time and historical Twitter data;
- ❑ diversity of its data types (e.g., text, image, video, and URL) and message topics (e.g., sports, health, and politics).

In terms of our DRF model, Twitter has high potential to contribute to the generalization of the nonuniformly distributed users behavior. In the practical sense, the concept 'frequency' used in the DRF model is then substantiated by real data. For example, we may translate the DRF model's 'frequency' to 'popularity' of the restaurants in Manhattan New York, so that the popular restaurants will be higher in the hierarchy in the DRF model and their query speed will be enhanced. This 'popularity' is defined by various characteristics of the corresponding Twitter data:

- ❑ Explicit characteristics include the number of tweets, check-ins, and mentions, as well as authors' opinions.
- ❑ Implicit characteristics refer to such information that are relevant but need to be further mined. For instance,
  - the distance of paired user-to-user relationship
  - power of influence of the individual users
  - geographical distance between the users and the location of the mentioned restaurants.