**Motivation**
- Geo-Spatial data is being collected on a massive scale.
- Approximate analysis is fast and often effective for this data.

**Hybrid R-Tree**
- Top levels are memory only nodes which are persisted in main memory for fast access.
- Lower levels are disk nodes which are persisted on disk and loaded into main memory when needed.

**Baseline Algorithms**
- **Query First** — Calculate \( S = P \cap Q \). Repeatably extract a sample from \( S \) upon request.
- **Sample First** — Upon Request, pick a point \( p \) where \( p \in P \). If \( p \in Q \), report \( p \), otherwise repeat.

**Data Structures**
- **R-tree** — A Hybrid R-tree implementation based off of Hilbert R-tree.
- **RS-tree** — Single R-tree \( T \) with a sample buffer \( s \) attached to each internal node. For each internal node \( u \in T \) with sample buffer \( s \), we sample uniformly from the children of \( u \), \( p(x = y \mid x \in s \land p \in P(u)) = \frac{1}{|P(u)|} \).
- **LS-tree** — a collection of R-trees where each R-tree indexes a set of samples from the original data set. The sample rates for these sets of samples form a geometric series.

**Individual Nodes of RS-tree**
- Once all local samples are observed during a query, children within MBB are added to frontier.

**Comparison of various sampling algorithms**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Query Time</th>
<th>Update Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>QueryFirst</td>
<td>( r(N) + q )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>SampleFirst</td>
<td>( kn/q )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>RandomPath</td>
<td>( k \log N )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>RandomShuffle</td>
<td>( kn/q )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>LS-tree</td>
<td>( \sum_{u \in P} \log q/k + \frac{\ell}{\log 2} )</td>
<td>( \log N )</td>
</tr>
<tr>
<td>RS-tree</td>
<td>( r(kn/q) + k )</td>
<td>( \log N )</td>
</tr>
</tbody>
</table>

**Full analysis of algorithms, including I/O cost can be found in the paper.**

**Building Index Experiments**

**Sample Query Experiments**

- **(a) GEO data, \( q = 250,000 \)**
- **(b) OSM data, \( q = 2.2 \) million**