Summary of Changes

Editor Comments

Most of the review comments have been satisfactorily addressed. There are still some concerns on the query accuracy and experimental result. Please address them in a further revision.

Our thanks to the editors and reviewers for their helpful feedback. This second revision includes:

- An addition to our Random Forest benchmark that provides results from a synthetic dataset of equal size to the original dataset (*Synthetic - 100%*). This allows a direct comparison between the raw dataset and synthetic data constructed from feature distributions.
- Clarifications in the text regarding query accuracy and the associated trade-off space. We now include more discussion on false positives vs. false negatives, which should help clear up any confusion on the trade-offs being made.

Our responses and actions taken are highlighted below.

Reviewer: 1

In this revision, all my previous comments have been addressed satisfactorily.

Thank you very much, and we appreciate all of your previous suggestions!

Reviewer: 2

I still have some concern about Table 4. To clarify the results of Synthetic better, Table 4 should have a row of Synthetic-100%, not just Synthetic-10% and Synthetic-20%.

This is an excellent point, and we have added the suggested *Synthetic - 100%* result to Table 4. This demonstrates the fidelity limitations of the sketch; while model accuracy is comparable with the others, we can observe a slight increase in error which is likely due to over-fitting on training samples that do not improve the expressiveness of the model.

Reviewer: 3

The authors have mostly addressed my comments. However, I am still confused about how the query accuracy is guaranteed.

If I understand correctly, SIFT stores statistical information for each sketchlet (corresponds to a geohash bounding box, right?). This is different from R-tree, which stores accurate information for each tree node. My question is that, a query may overlap with several sketchlets, and some points are in these overlapped sketchlets, but not in the query range. The statistical information in SIFT is computed in the sketchlets level. The points not in the query range are used to compute the statistics of these sketchlets, which are used in query processing, and this causes the error. I am still not clear how this problem can be fixed in your proposed method. I think this problem also happens for the other dimensions.

When a query overlaps multiple sketchlets (or in other words, multiple geohashes), it will be forwarded to relevant machines in the system using our distributed prefix tree. Once the queries arrive at their destination(s), data points that do not match will be eliminated via a tree traversal; since full-resolution geohashes are stored in the sketchlets, spatial information is reported with high accuracy. However, you are correct in that we do not maintain the original (x, y) or (lat, lon) points like an R-tree would, and therefore even fine-grained geohash bounding boxes may still produce false positive matches. In general, these situations are rare, but arise due to the memory-accuracy trade-offs made in Synopsis.

A related issue is the accuracy of other features/dimensions; since our quantization scheme assigns data points to variable-sized bins, range queries can overlap with bins that they do not fully cover (shown in Figure 4). However, when query results are transmitted to the user, they contain the ranges and distributions of the data points. We leverage this information to provide measures of accuracy and ensure that synthetic datasets only contain data within the ranges specified. Additionally, if a user desires assurance that their results do not contain any values outside the query range, vertices that do not overlap entirely can be pruned.

This boils down to two trade-offs managed by Synopsis to deal with voluminous datasets: accuracy vs. memory, and false positives vs. false negatives. To clarify this, we have added more detail to the discussion in Section 3.5.