UNIVERSITY OF CAMBRIDGE

Needle in a Haystack: Searching for Approximate k-Nearest Neighbours in High-Dimensional Data

Liang Wang*, Ville Hyvönen, Teemu Pitkänen, Sotiris Tasoulis, Teemu Roos, and Jukka Corander

University of Cambridge*, UK University of Helsinki, Finland

Not Only Tall, But Also Very Fat

- Data grow in both volume and dimensionality.
- Due to the technology advances and modelling techniques.
  - Advances in measuring and monitoring tools.
  - Advances in computation and storage technologies.
  - DNA, stock market, language models: inherently HD models.
- Why do high-dimensional data matter?
  - It is hard to tell what information matters in the beginning.
  - Save everything and leave this problem later or someone else.

Searching Needle(s) in a Haystack

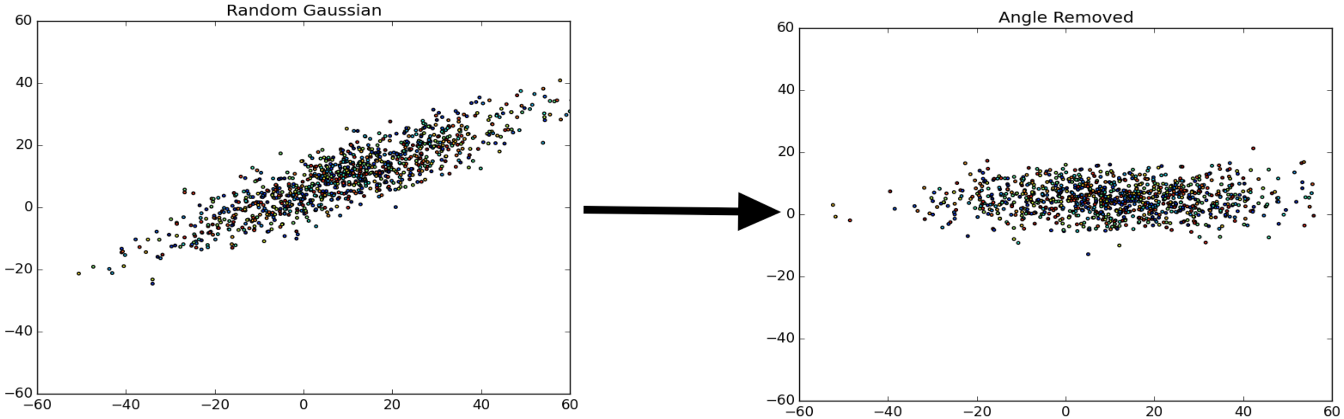
- Searching is among the most important operations.
  - E.g., Computer vision, pattern recognition, natural language processing, online recommenders, and etc.

- Searching is difficult in high-dimensional data. Why?
  - “Under rather general conditions, given a query point, the distance between the nearest and farthest points does not increase as fast as dimensionality.”
  - k-NN quickly becomes unstable in high-dimensional spaces.

K. Beyer, et al. "When is “nearest neighbor” meaningful?." *Database Theory - ICDT’99*. Springer, 1999.

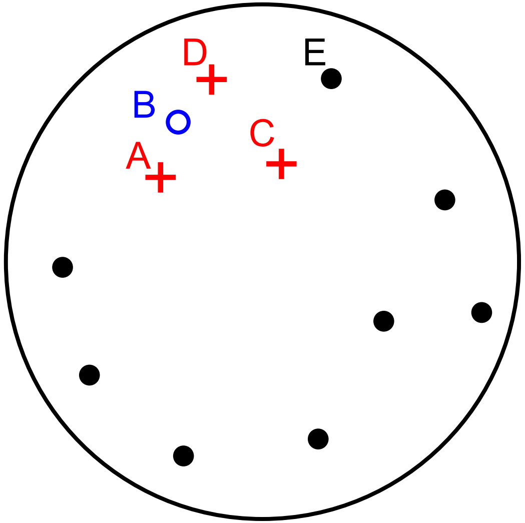
Key Technique - Approximation

- Approximate the original data set with another one of lower dimensionality by “tolerating some error”, i.e., Dimensionality reduction - e.g., SVD, Random Forest, and etc.



Key Technique - Approximation

- Approximate the exact search results with a “roughly” good ones, especially useful for time-constrained applications.

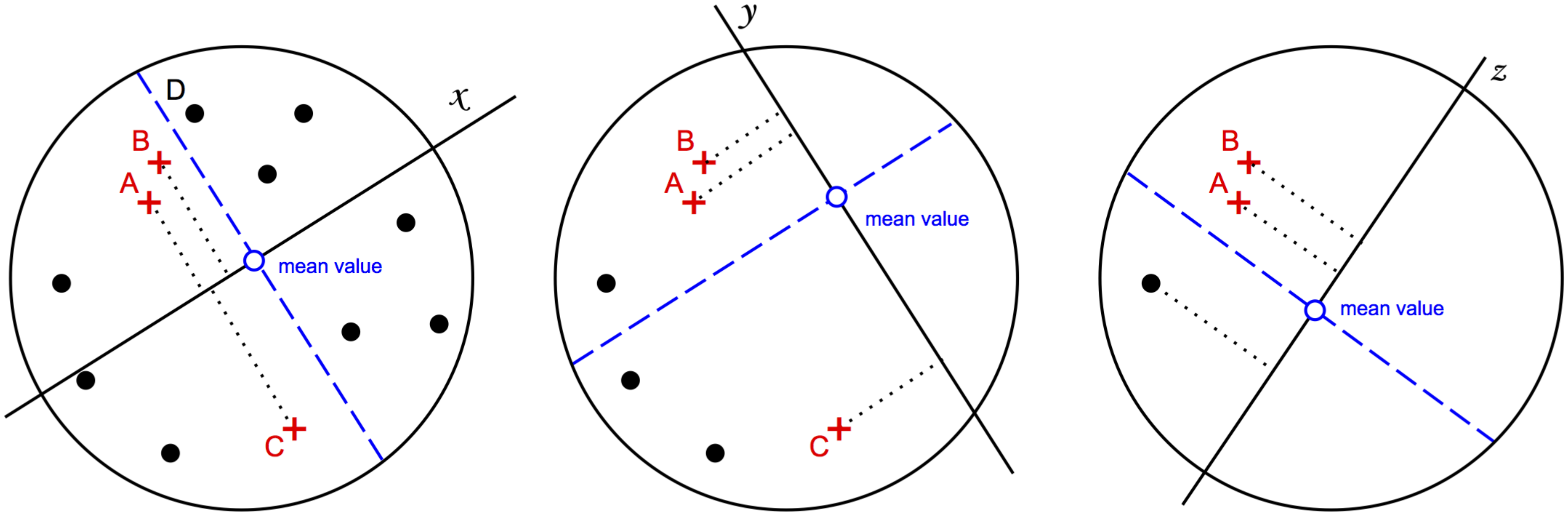


For example, B’s 3-nearest neighbours are A, C and D. Instead of returning the exact result, we can return A, C, and E if our application can tolerate certain level of error.

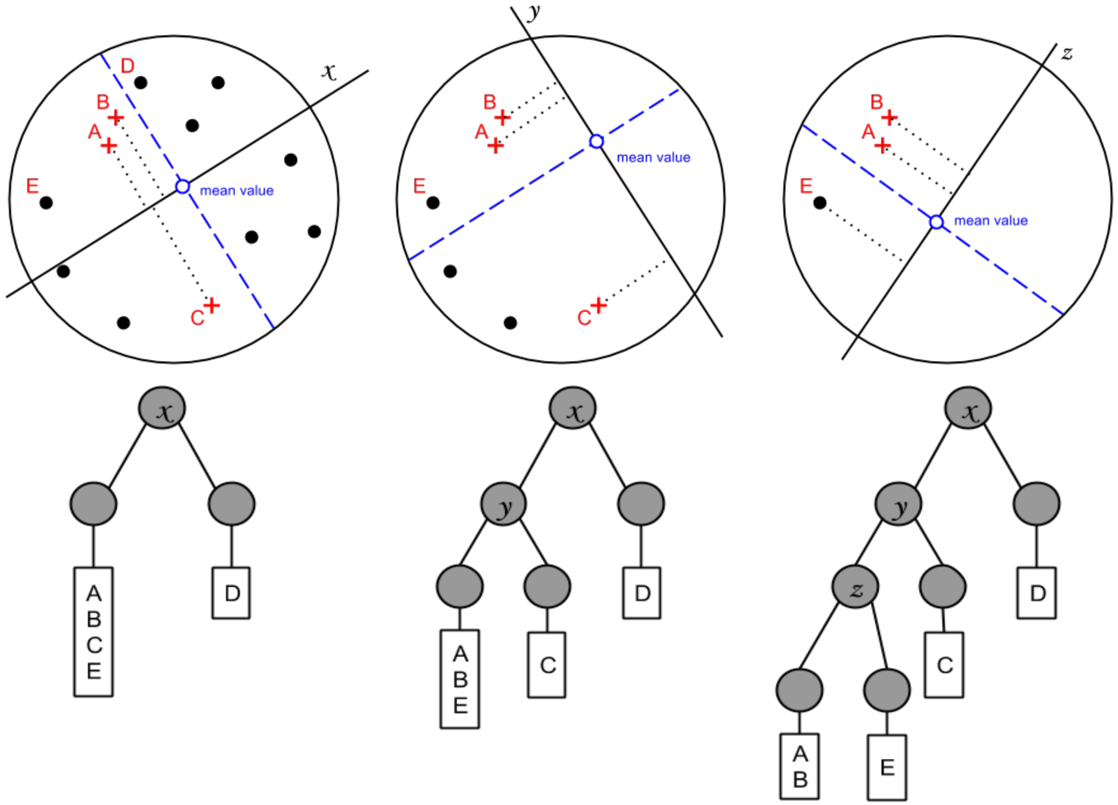
By so doing, we are usually able to gain a significant improvement on searching efficiency.

Random Projection

- Essentially, it is all about clustering - similar points should be grouped together, i.e., in a cluster.



Classic Random-Projection Tree



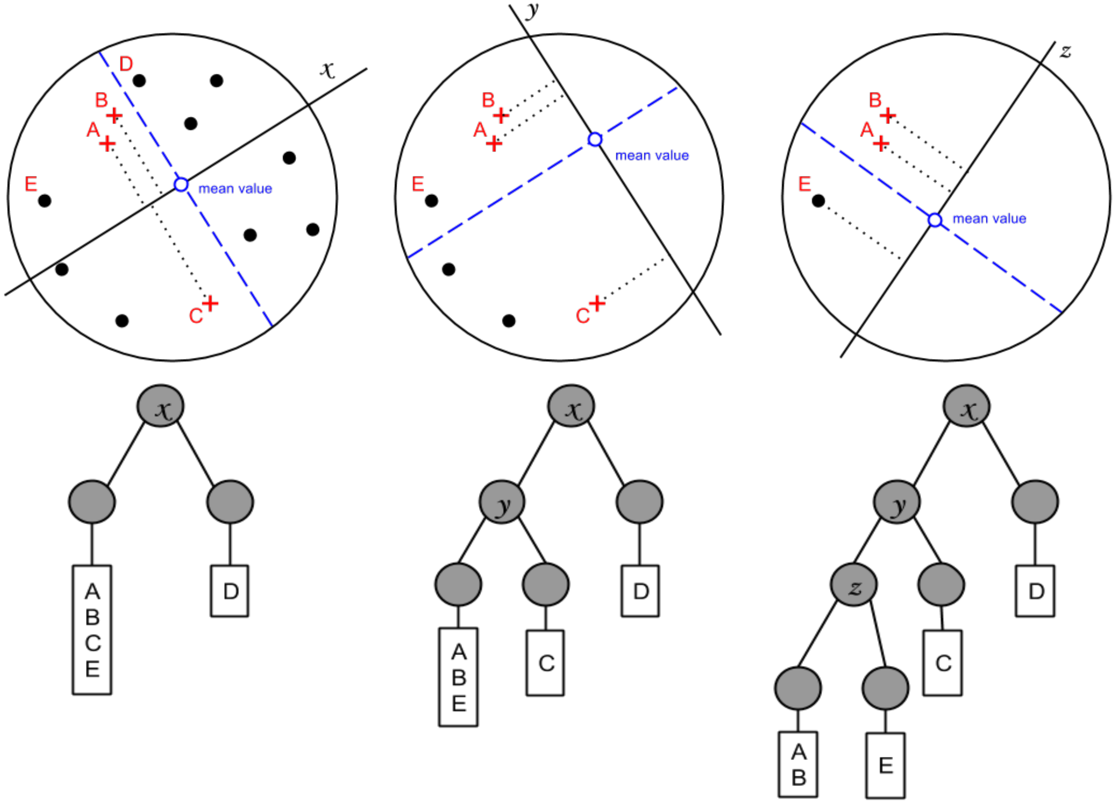
In every step, the problem space will be divided into half, then solved separately. It is a typical divide and conquer technique.

The split point can be mean value, median, or other more complicated statistics.

The leaf node is a cluster of points which are close to each other.

UNIVERSITY OF CAMBRIDGE

Issues of Classic RP-Tree



In general, the accuracy is not very high even for a data set of medium dimensionalities.

The accuracy is impacted by two kinds of misclassifications: i.e., B and D; A and C.

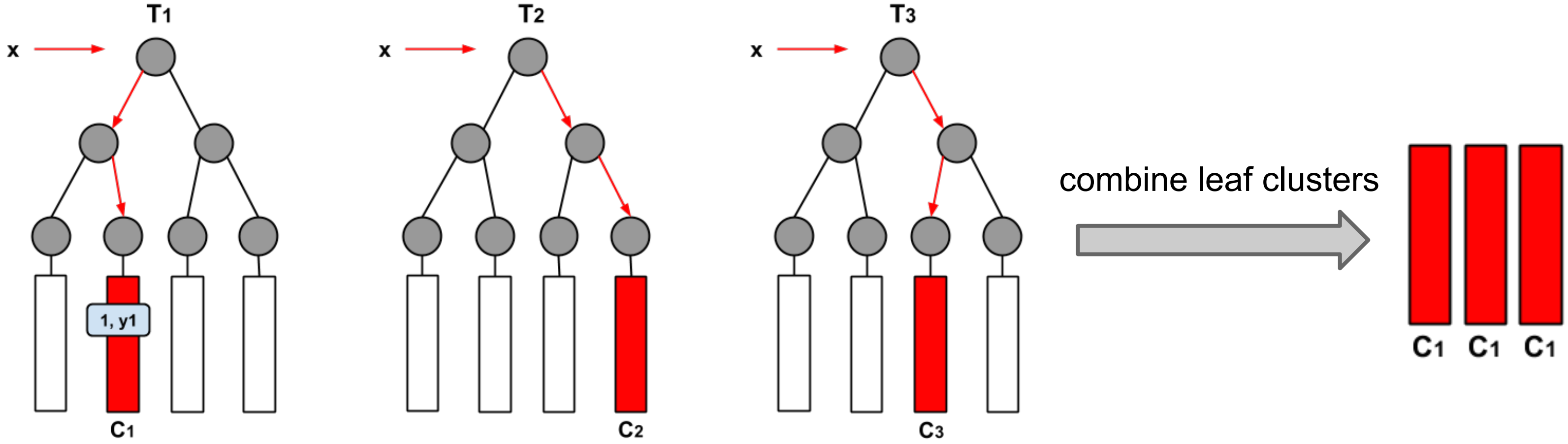
The process of Index building has only limited parallelism, so not very efficient in practice.

Index size is big due to storing high-dimensional vectors in the intermediate nodes.

UNIVERSITY OF CAMBRIDGE

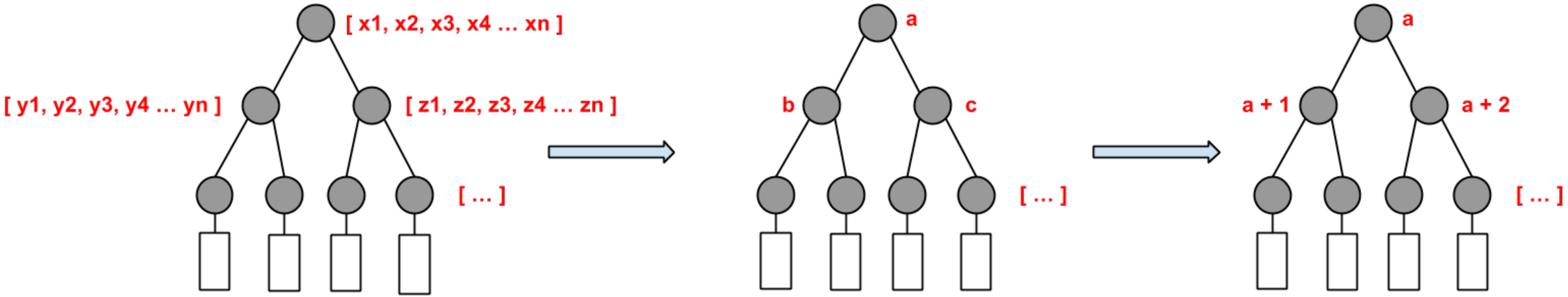
MRPT - Improve Accuracy

- Increase either leaf size or # of trees, but which is better?



MRPT - Improve Index Size

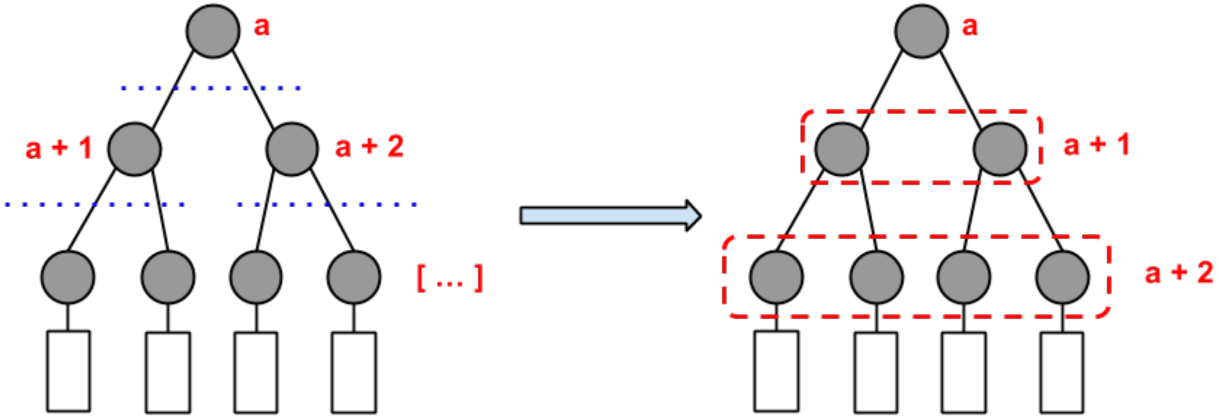
- We do not need to store the actual vector at each node.
- Instead, we can use a random seed to generate on the fly.


[x1, x2, x3, x4 ... xn]
[y1, y2, y3, y4 ... yn]
[z1, z2, z3, z4 ... zn]
[...]
a
b
c
[...]
a
a + 1
a + 2
[...]

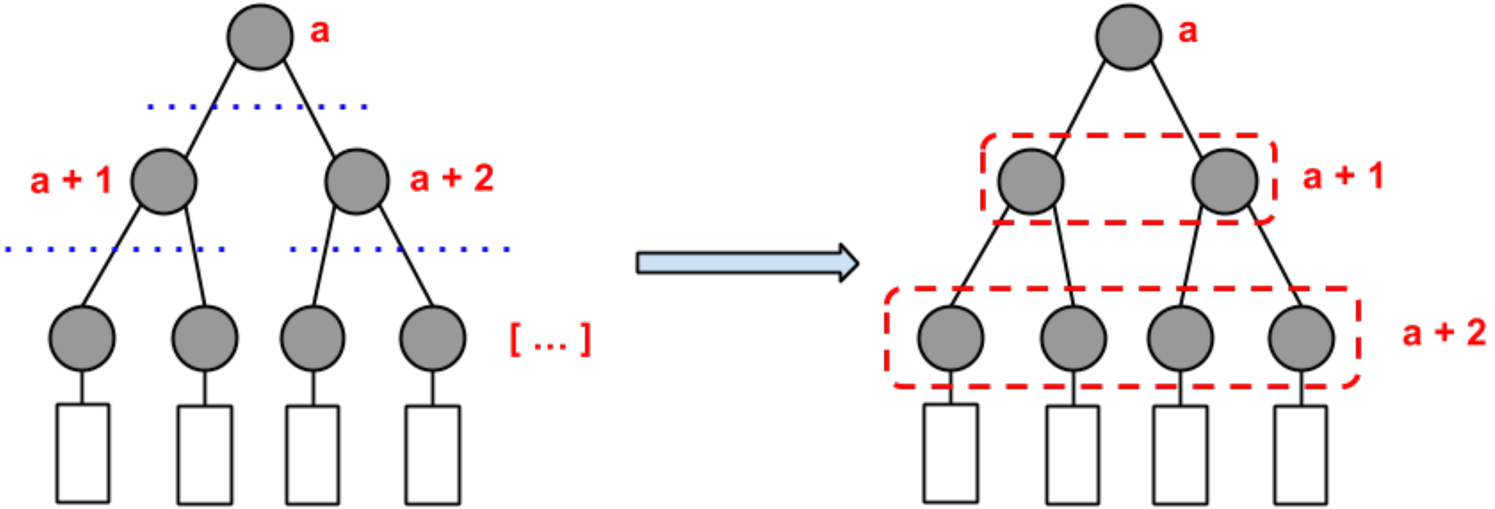
In a leaf cluster, only the indices of vectors in the original data set are stored.

MRPT - Improve Efficiency

- Current algorithm can be parallelised to some extent, especially when moving towards leaves.
- Can we do better? By maximising the parallelism?



MRPT - Improve Efficiency



Blue dotted lines are critical boundaries. The computations in the child-branches cannot proceed without finishing the computation in the parent node.

There is no critical boundary. All the projections can be done in just one matrix multiplication. Therefore, the parallelism can be maximised.

Almost Done, Let's Conclude

- High-dimensional data sets are quite common in practical applications. Efficient and accurate searching is difficult.

- MRPT is a compact data structure which provides approximate k-NN search for high-dimensional big data sets.

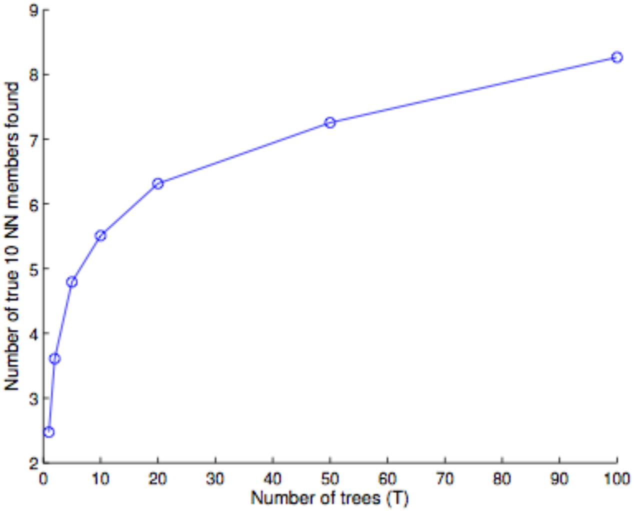
- MRPT optimises the index size, searching accuracy, searching efficiency, and parallelism of a building process.

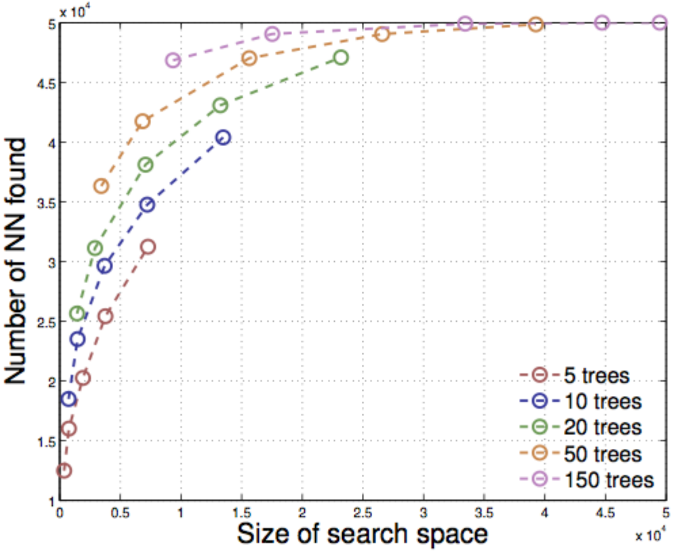
Thank you. Questions?

UNIVERSITY OF CAMBRIDGE

MRPT - Improve Accuracy

- Increase either leaf size or # of trees, but which is better?





Finally, A Concrete Application of MRPT