Over the last few years, research on similarity search in high dimensional spaces has become intensely dynamic. It is caused by a recent application of this techniques in the fields like big data, information retrieval, machine learning, cryptography, bioinformatics, etc. The algorithms that solve high dimensional similarity search are often used as a theoretical test-bed to solve a specified problem in the aforementioned fields. In this project we want to work on the following three challenges:

- **Lower bounds:** In computer science, lower bounds are used as a roadmap that shows what is the minimal amount of resources needed to solve a problem. Inevitably, lower bounds are used to show that some algorithms are optimal and in order to improve them, we need to relax some significant constraint. For example, after 10 years of research it has been finally showed that the decision trees of size $O(n^\rho)$, that achieve approximation $O(\log_\rho \log d)$ are optimal in the metric l_∞ and we cannot overcome this bound with any other data structure. What are the lower bounds in different metrics? Which assumptions we relax if we want to break these bonds. Can we add additional guarantees and achieve similar lower bounds? Here, we would like to answer these questions, especially what is the trade-off between the expected query time and memory usage.

- **Eradicate False Negatives:** Due to the high cost of an exact solution, we have to use an approximate algorithm. When one considers such an algorithm, inevitably some errors may occur. When the algorithm states that the correct answer is incorrect we say it has the problem with false negatives. Recently, we showed that with some cost we can eliminate this problem in high dimensional, nearest neighbor search. Now, we want to apply our reasoning to some similar high dimensional problems. Those guarantees may be of interest in the cryptography, machine learning and databases.

- **Stochastic LSH:** Stochastic models exploit the fact that queries/ data come from some random distribution, e.g., some queryies are more frequent than the others. These additional assumption to the model can significantly improve response time in some NP-hard problems (e.g., set cover). Surprisingly, no one has theoretically analysed how this strategy would affect response times in high dimensional similarity search. Here, we want to analyse popular algorithms (mainly LSH) in the stochastic models.

The success of the project will significantly contribute to the similarity search in high dimensional spaces research. The progress in this area implies progress in big data, computer vision, information retrieval, bioinformatics and even cryptography. In these fields, the similarity search is used regularly. We hope that within this project we will be able to prove the optimality of some algorithms and consequently determine the new paths for further research. In the long run, our results will have an important practical application in cloud computing, recommending systems and internet advertisement.