W ostatnich latach badania nad poszukiwaniem podobnych obiektów w przestrzeniach wielowymiarowych nabrały niespodziewanego rozpędu. Mają one szerokie zastosowanie w dziedzinach takich jak przetwarzanie dużych danych, odzyskiwanie informacji, uczenie maszynowe, kryptografia czy bioinformatyka. Algorytmy te są często używane jako teoretyczny poligon badawczy do oceny jakości rozwiązania. W ramach tego projektu chcemy rozwijać trzy obiecujące kierunki badań w tej dziedzinie:

- **Dolne granice:**

  W informatyce, dolne granice używane są jako wyznacznik tego ile minimalnie zasobów należy zużyć aby rozwiązać dany problem. Dolne granice służą głównie do pokazania że dany algorytm jest optymalny. Poprawienie go wymaga zmiany pewnych istotnych założeń. Na przykład, po 10 latach badań udowodniono, że drzewa decyzyjne rozmiaru $O(n^\rho)$, które gwarantują aproksymacje rzędu $O(\log_\rho \log d)$ są optymalne dla metryki l_∞. Nie istnieje ponadto żadna struktura danych, której udałoby się pokonać tę granice. W ramach badań w tym kierunku, chcielibyśmy odpowiedzieć na następujące pytania: Jakie są dolne granice dla innych metryk. Których założeń należy się pozbyć żeby istotnie pokonać tą granice? Czy możemy dodać dodatkowe gwarancje do rozwiązania i zapewnić optymalność granicy? Najważniejszym celem jest analiza kompromisów pomiędzy spodziewanym czasem odpowiedzi na zapytanie a zużyciem pamięci.

- **Usunięcie problemu wyników fałszywie negatywnych:**

  Wyniki fałszywie negatywne to sytuacja gdy program niepoprawnie sklasyfikuje poprawną odpowiedź jako błędną. Niedawno pokazaliśmy, że przy niewielkim koszcie, jesteśmy w stanie rozwiązać ten problem w algorytmie poszukiwania najbliższych sąsiadów w przestrzeniach wielowymiarowych. Teraz chcemy przenieść nasze rozumowanie na inne problemy, które maja zastosowanie w kryptografii, uczeniu maszynowym oraz w bazach danych.

- **Stochastyczne LSH**

  Zapytania stochastyczne to strategia użycia szybkiej pamięci (ang. cache) w celu przyśpieszenia działania algorytmu. Wykorzystują one właściwość, że niektóre zapytania pojawiają się znacząco częściej niż inne. To dodatkowe założenie pozwoliło polepszyć czasy działania w niektórych problemach NP-trudnych (np. set cover). Tutaj, chcielibyśmy przeanalizować różne algorytmy (głównie LSH) w tym modelu i znaleźć warunki na to jaki wpływ na czas odpowiedzi ma rozmiar cache.

Sukces projektu będzie znaczącym wkładem w rozwój algorytmów wyszukiwania najbliższych sąsiadów w przestrzeniach wielowymiarowych. Wierzymy, że postępy w tej dziedzinie przełożą się na rozwój takich dziedzin jak przetwarzania dużych danych, rozpoznawanie obrazów, odzyskiwanie danych, bioinformatyce a nawet w kryptografii. W ramach tego projektu postaramy się udowodnić optymalność niektórych algorytmów. W dłuższej perspektywie nasze wyniki przełożą się na praktyczne zastosowania w obliczeniach na chmurach, systemach rekomendacyjnych oraz reklamie internetowej.