# SPATIAL SIMILARITY DISCOVERY USING K-NEAREST NEIGHBORS (KNN)

**Authors:** Farhan Ali, Trinish Nepal, Riyan Saleem **Group:** 6

## Problem Statement (What & Why?)

1. Problem Statement (What & Why?)
Goal: Identify the most similar locations to a given target point using geographic coordinates.
Why is this interesting?

- Many real-world problems (urban planning, resource allocation, geospatial analysis) depend on finding similar nearby locations.
- Spatial similarity is non-trivial because distance directly influences relevance.

Problem Complexity:

- Continuous spatial data (latitude & longitude)
- Distance-based similarity computation
- Sensitivity to scale and choice of distance metric

## Dataset Description

The dataset (challenge.csv) consists of numerical attributes including latitude (N) and longitude (E). All attributes are numeric and contain no missing values. The final record in the dataset is treated as a query instance, while all preceding records serve as candidate locations for similarity comparison

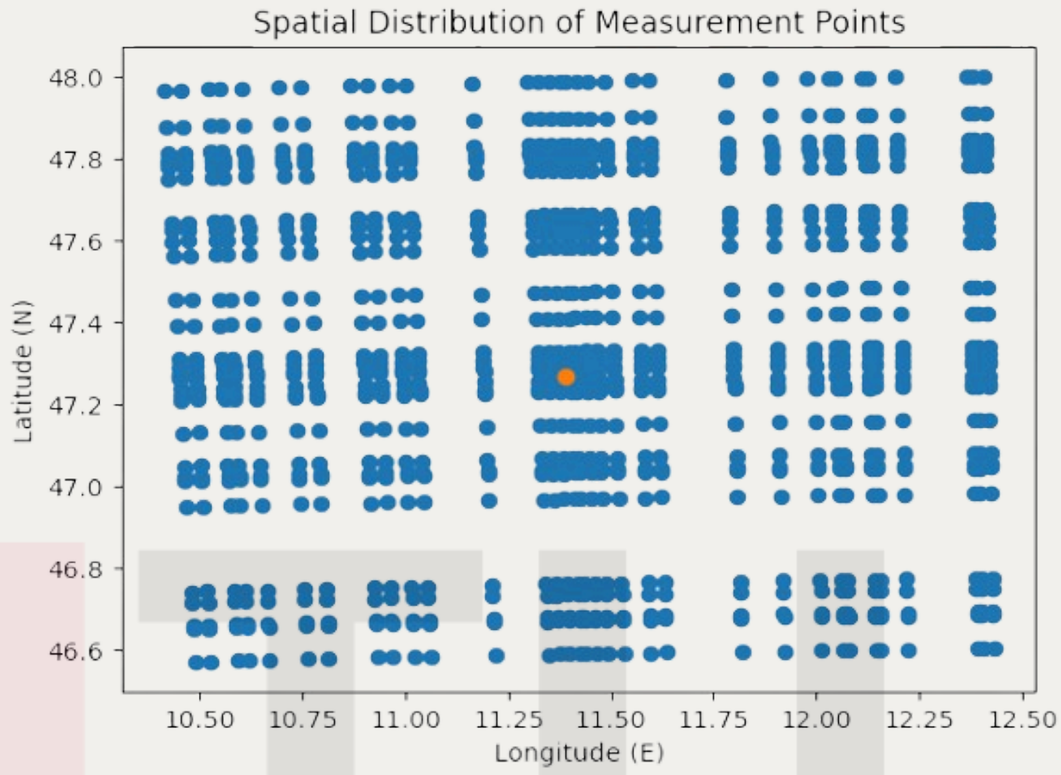


Figure 1: Spatial distribution of all locations with the target instance and its k-nearest neighbors highlighted.

## KDD Pipeline

| Data Selection | Preprocessing | Transformation | Data Mining | Interpretation |
|---|---|---|---|---|
| Load CSV | Feature Selection | Coordinate Extraction | KNN Similarity | Visual and Distance Analysis |

## Methodology & Rationale

We apply the K-Nearest Neighbors (KNN) algorithm, an instance-based learning method well-suited for similarity discovery. Euclidean distance is computed using latitude and longitude features, and the k = 10 closest locations are retrieved. The method is implemented using the NearestNeighbors module from scikit-learn

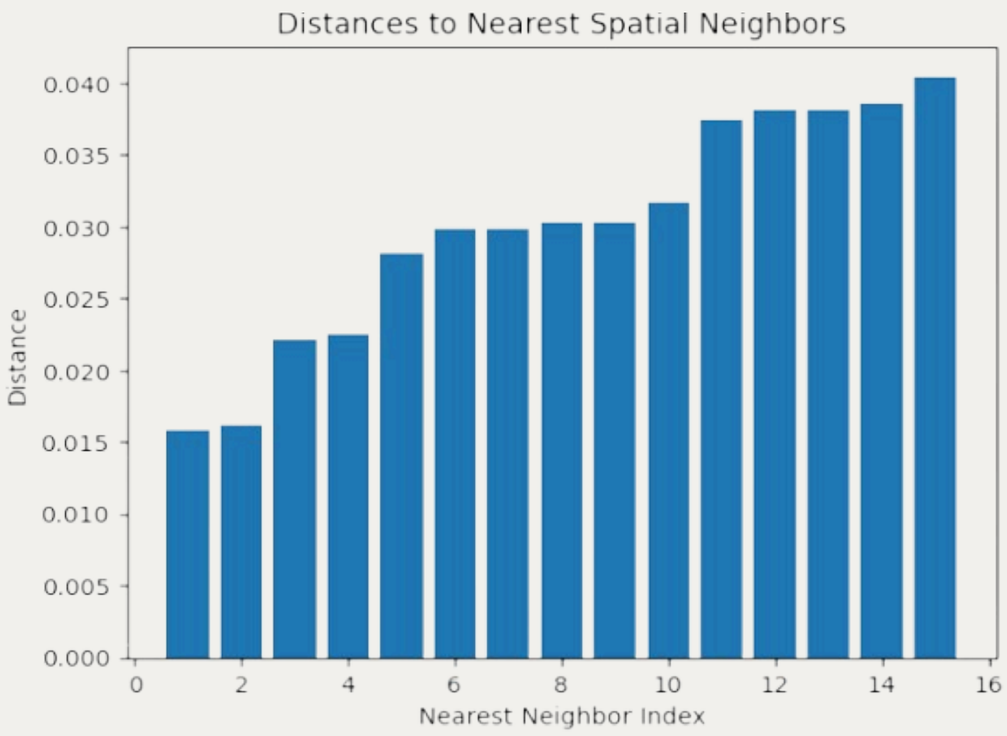


Figure 2: Distance-based ranking of the retrieved nearest neighbors.

## Baseline & Evaluation Strategy

A trivial random neighbor selection is used as a baseline. Since the task focuses on unsupervised similarity discovery rather than prediction, evaluation is performed using distance-based ranking and visual inspection instead of accuracy metrics or train-test splitting

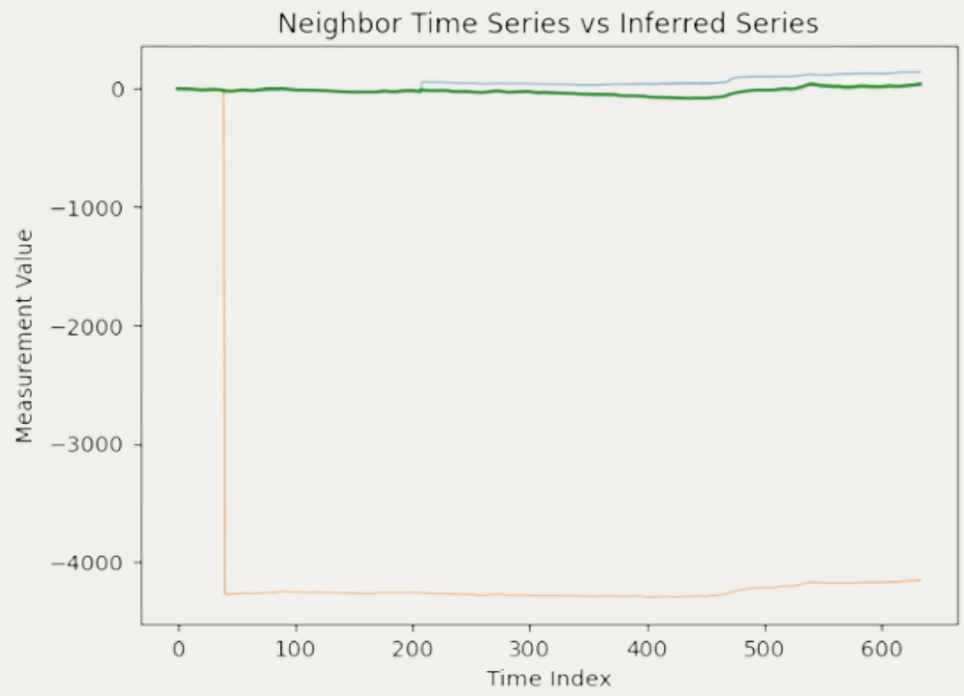


Figure 3: Comparison of spatial proximity among retrieved neighbors.

## Key Findings & Takeaways

- Spatial distance is a strong indicator of similarity.
- KNN provides a simple yet effective baseline for spatial similarity tasks.
- Interpretability and reproducibility are key strengths of the approach.

## Results & Discussion

The results show that compared to the random baseline, the selected neighbors exhibit significantly smaller spatial distances, validating the effectiveness of distance-based similarity in this context

## Limitations & Future Work

KNN is sensitive to noise and does not generalize beyond stored instances. Future work includes feature scaling, weighted distance metrics, incorporation of additional attributes, and comparison with clustering methods such as DBSCAN.