Pivot Selection Techniques

Proximity Searching in Metric Spaces

by Benjamin Bustos, Gonzalo Navarro and Edgar Chávez

Catarina Moreira

Outline

- Introduction
- Pivots and Metric Spaces
- Pivots in Nearest Neighbor Search
- Techniques for Pivot Selection
- Conclusions

Introduction

- Search for information can be performed in two different ways:
  - **Exact Search**: the search is performed by looking for an element whose identifier exactly corresponds to a defined search key
  - **Approximate Search**: the search is performed by finding an item in the dataset which is surrounded by a radius and is sufficiently close to the defined key search
- However, searching in higher dimensional data:
  - Exact and Approximate search will suffer from the curse of dimensionality
  - Dilemma: either reduce the number of features making the search less discriminative and unreliable, or reduce the quality of the results through approximation search.

Introduction

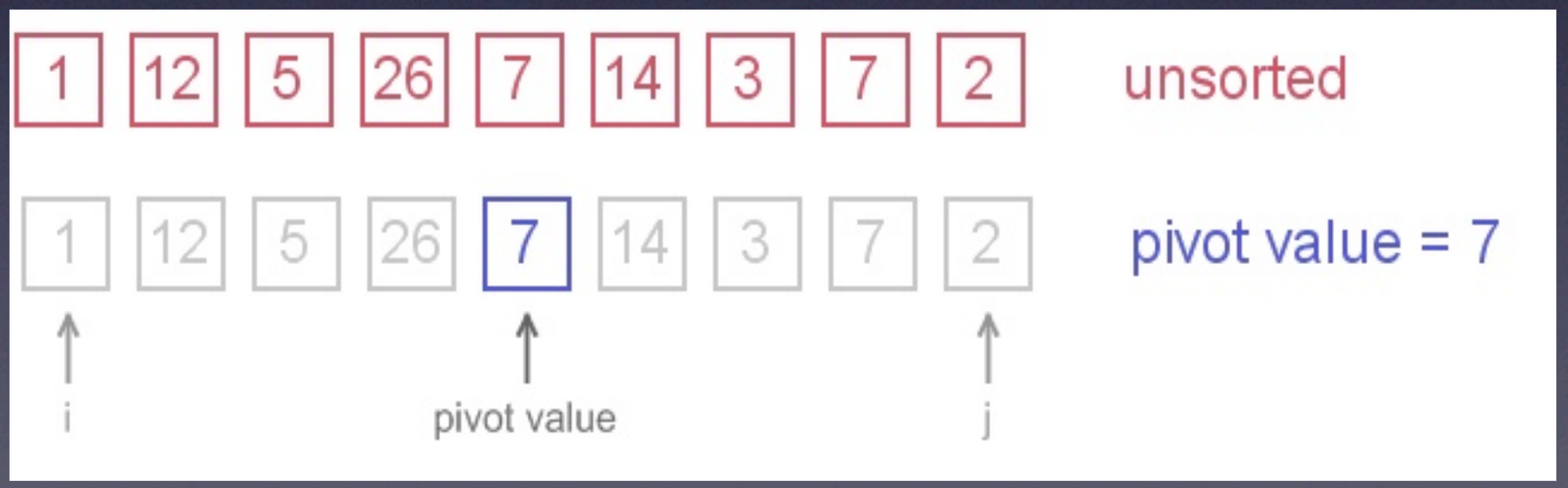
- To answer queries, approximate search relies in indexing methods which minimize the number of comparisons between the query and the items of the dataset
- Pivot methods are indexing structures used in many algorithms of the literature
  - Algorithms using such structures require big amounts of memory for larger datasets
  - The choice of the pivots drastically improves the approximate search.
  - They also suffer from the curse of dimensionality
- In this paper, the authors explore 3 different approaches to select pivots

Outline

- ~~Introduction~~
- Pivots and Metric Spaces
- Pivots in Nearest Neighbor Search
- Techniques for Pivot Selection
- Conclusions

What are Pivots?

- Pivots are items from the collection which are chosen to index the dataset
- Choosing the best pivots is a very challenging problem
- QuickSort is an example of an algorithm which uses pivot as an indexing structure



Metric Spaces

- Metric Spaces are defined as a pair *(E, d)* with $E \subset U$, where *U* is the universe of the elements in the space, *E* is a finite subset of *U* which represents the dataset and *d* is a function $U \times U \rightarrow \mathfrak{R}$ which obeys to the following properties:

- Positivity ..... $\forall x, y \in U, d(x, y) \geq 0$
- Symmetry ..... $\forall x, y \in U, d(x, y) = d(y, x)$
- Reflexivity ..... $\forall x \in U, d(x, x) = 0$
- Strict Positivity ..... $\forall x, y \in U, x \neq y \Rightarrow d(x, y) > 0$
- Triangular Inequality ..... $\forall x, y, z \in U, d(x, z) \leq d(x, y) + d(y, z)$

Outline

- ~~Introduction~~
- ~~Pivots and Metric Spaces~~
- Pivots in Nearest Neighbor Search
- Techniques for Pivot Selection
- Conclusions

Nearest Neighbor Search

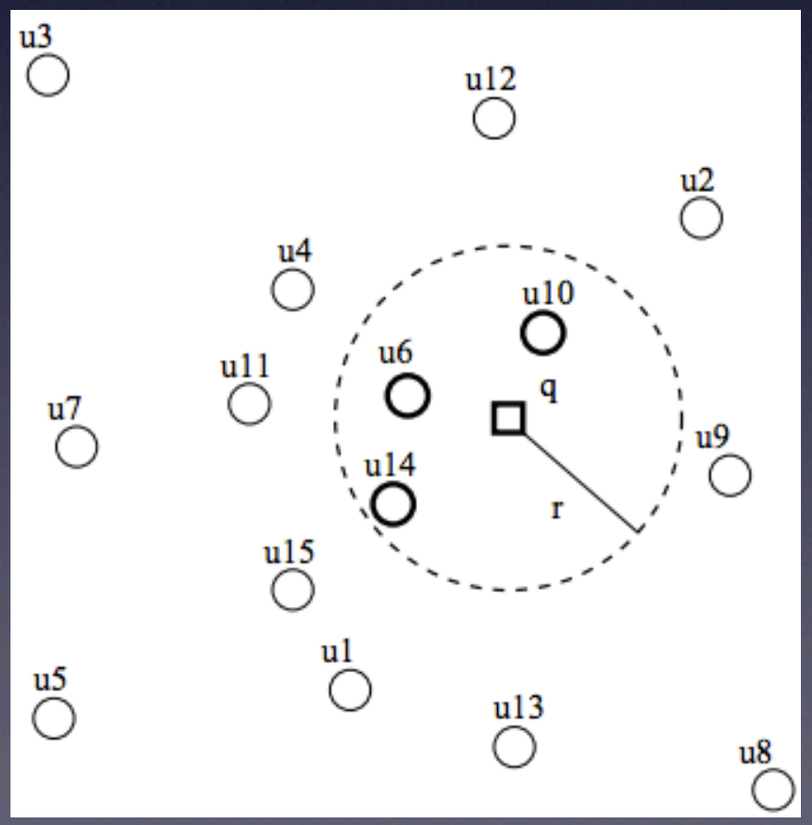
- It is an optimization problem for finding closest points in metric spaces.
- Given a metric space *(E, d)*, a query $q \in U$ and a radius $r > 0$, $r \in \Re$, the approximate search $(q, r)_d$ over the metric space *(E, d)* is defined as:

$$(q,r)_d = \{x \in E, d(q,x) \leq r\}$$

$$L_2((x_1,..,x_n),(y_1,..,y_n)) = \sqrt{(\sum_{i=1}^{n} |x_i - y_i|^2)}$$

Nearest Neighbor Search

- Given a metric space *(E, d)*, a query $q \in U$ and a radius $r > 0$, $r \in \Re$, the approximate search $(q, r)_d$ over the metric space *(E, d)* is defined as:



In the figure, the result of the query $(q, r)_d$ is the set of elements: u6 - u10 -u14

Pivots in NN Search

- Given a query *(q, r)* and a set of *k* pivots $p_i \in E$, through the triangular inequality we obtain:

$$d(p_i, x) \leq d(p_i, q) + d(q, x) \qquad d(p_i, q) \leq d(p_i, x) + d(q, x)$$

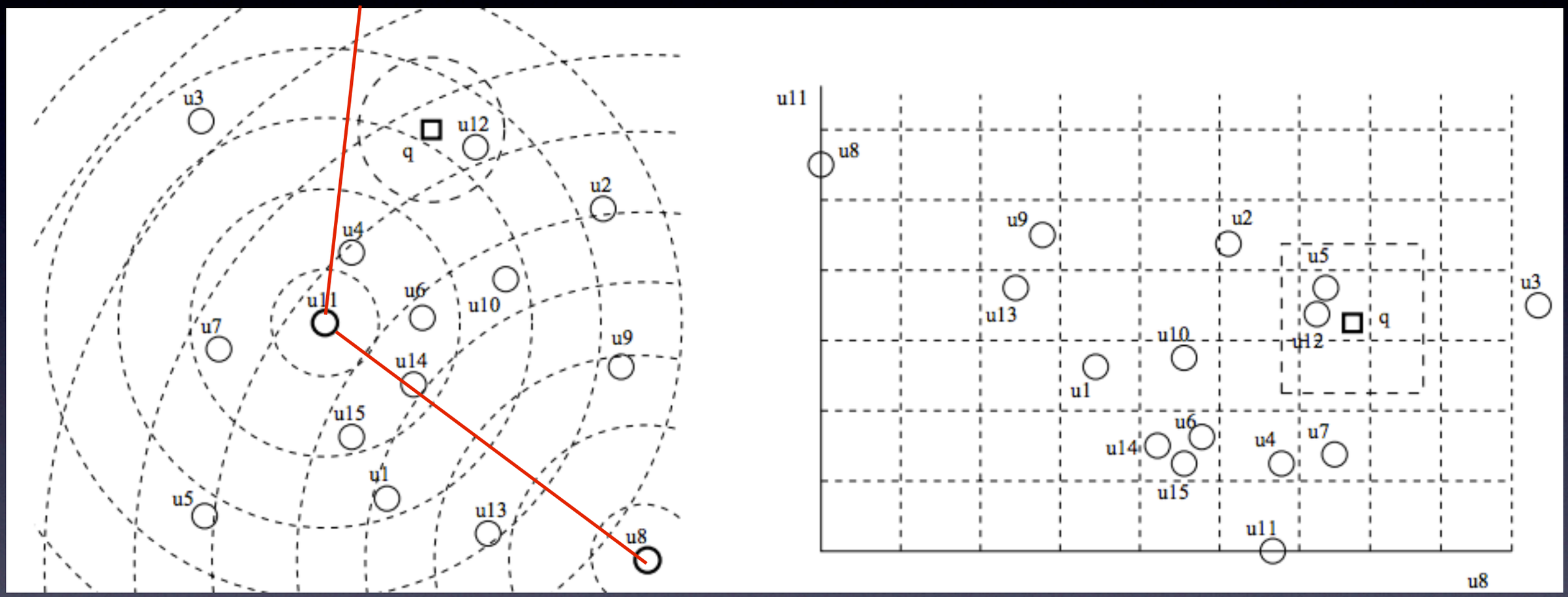
- So, the distance between *q* and *x* is given by:

$$d(q, x) \geq \left| d(p_i, x) - d(p_i, q) \right|$$

- It follows that all items which are not inside the radius r can be discarded (items which do not obey to the equation):

$$\left| d(q, p_i) - d(x, p_i) \right| \leq r, \forall i = 1..k$$

Pivots in NN Search



Mapping of a metric space into a vector space with function *d* using two pivots (u8 - u11)

Good Pivots

- Good Pivots have the following properties:
  - They are far away between each other, i.e., the mean distance between pivots is higher than the mean distance between random elements of the metric space
  - They are far away from the rest of the elements of the metric space (outliers)

Outline

- ~~Introduction~~
- ~~Pivots and Metric Spaces~~
- ~~Pivots in Nearest Neighbor Search~~
- Techniques for Pivot Selection
- Conclusions

Techniques for Pivot Selection

- In the literature, it has already been shown that the choice of the pivots can drastically affect the quality of the search process.
- In their work, the authors have proposed 3 methods to select pivots:
  - Selection of *N* Random Groups
  - Incremental Selection
  - Local Optimum Selection

Selection of *N* Random Groups

N groups of *k* pivots are chosen at random among the elements of the database, and *D* is calculated for each of these groups of pivots. The group that has the maximum *D* value is selected.

Incremental Selection

A pivot p_1 is selected from a sample of *N* elements of the database, such that, that pivot alone has the maximum u_D value (efficiency criterion). Then, a second pivot p_2 is chosen from another sample of *N* elements of the dataset, such that $\{p_1, p_2\}$ have the maximum u_D value, considering p_1 fixed. The process is repeated until all *k* pivots are selected.

Local Optimum Selection

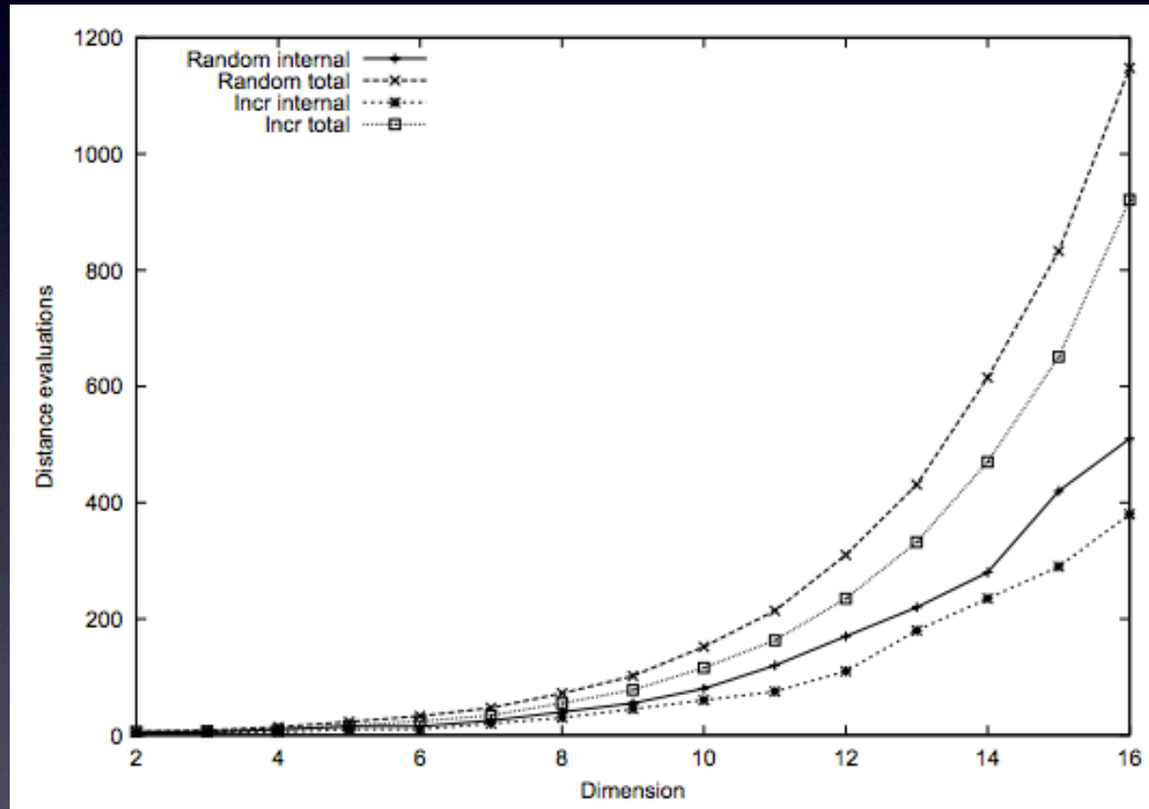
A group of k pivots are chosen at random among the elements of the database. Compute the matrix:

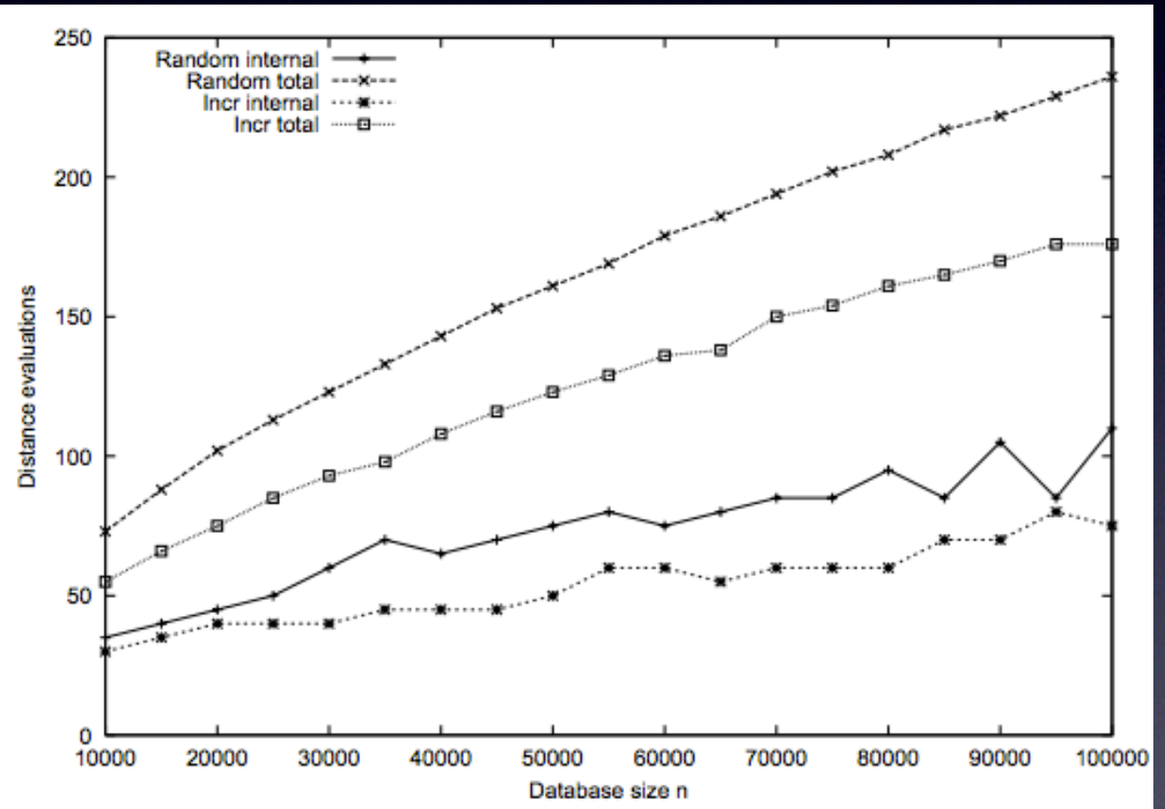
$$M(r,j) = \left| d(a_r, p_j) - d(a'_r, p_j) \right|, r = 1..A$$

Sum each row of the matrix and check which rows have the highest scores

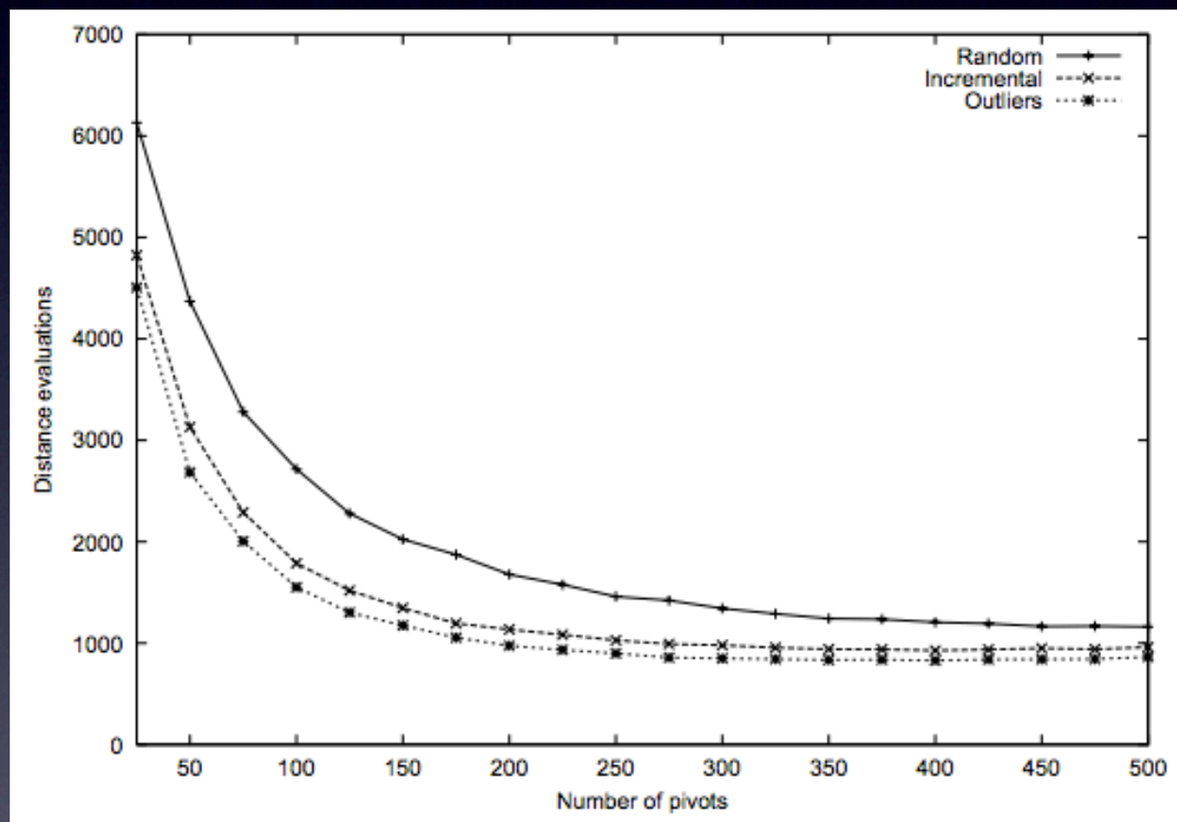
The indexes of the rows with the best scores will correspond to the pivots

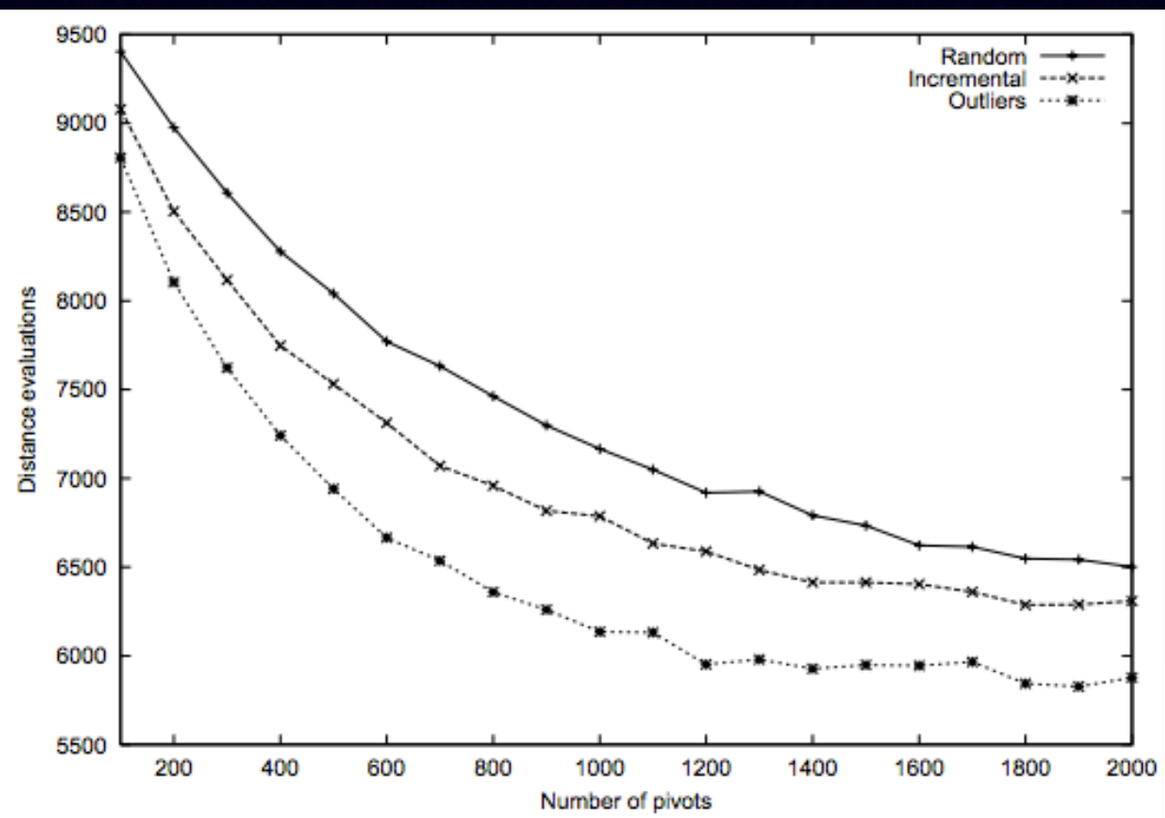
Comparison Between Random and Incremental Selections





Comparison Between Incremental and Optimum Selections





Outline

- ~~Introduction~~
- ~~Pivots and Metric Spaces~~
- ~~Pivots in Nearest Neighbor Search~~
- ~~Techniques for Pivot Selection~~
- Conclusions

Conclusions

- Pivot methods are effective indexing structures
- Nearest Neighbor Search using such structures can improve the search results, if pivots are carefully chosen
- Experiments showed that the incremental pivot selection method is the most effective one
- Despite the good results, these indexing structures still suffer from the curse of dimensionality