L 9: Approximate Nearest Neighbors

Feb 5, 2025

Jeff M. Phillips

Large Set Data $P = \{p_1, p_2, \dots p_n\} \in \mathcal{X}$

Given a distance $D : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$

Nearest Neighbor for a query $q \in \mathcal{X}$

$$\phi_P(q) = \underset{p \in P}{\arg\min}\, D(p, q)$$

k-Nearest Neighbors $\phi_{P,k}(q)$

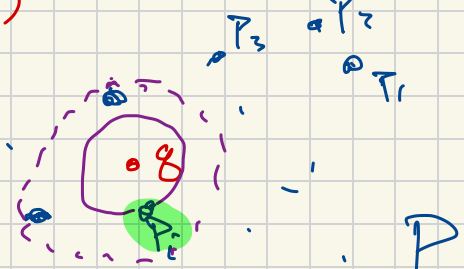
query q, $k \geq 1$

$p_i = \phi_P(q)$

Set $S \subset P$ s.t. $|S| = k$

no point $p' \in P \setminus S$ has

$D(q, p') < D(q, s)$ so $s \in S$



Algo for NN

$|P| = n$

1. Baseline: check all $O(|P|)$, return smallest

Pre-process: Spend $O(n)$ time

$O(n \log n)$ time

near-linear

Build Data structure $A(P)$

Then can ask NN queries of $A(P)$ quickly.

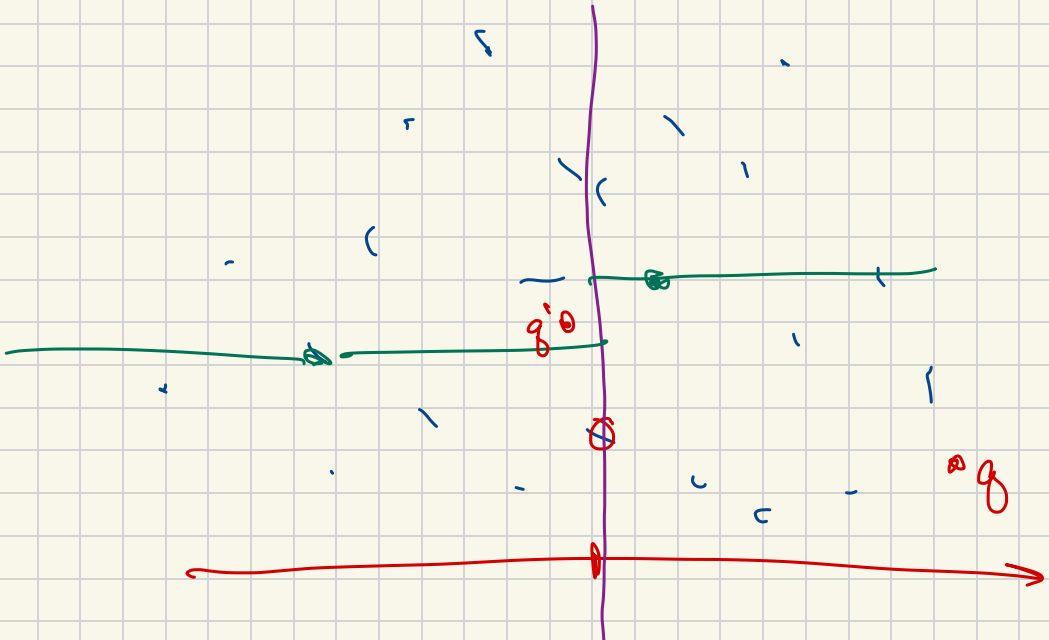
Ex: if PCR: Build binary tree, Queries $O(\log n)$

Classic Idea Applies $P \subset \mathbb{R}^d$ d, small

✓ d = 2, 3

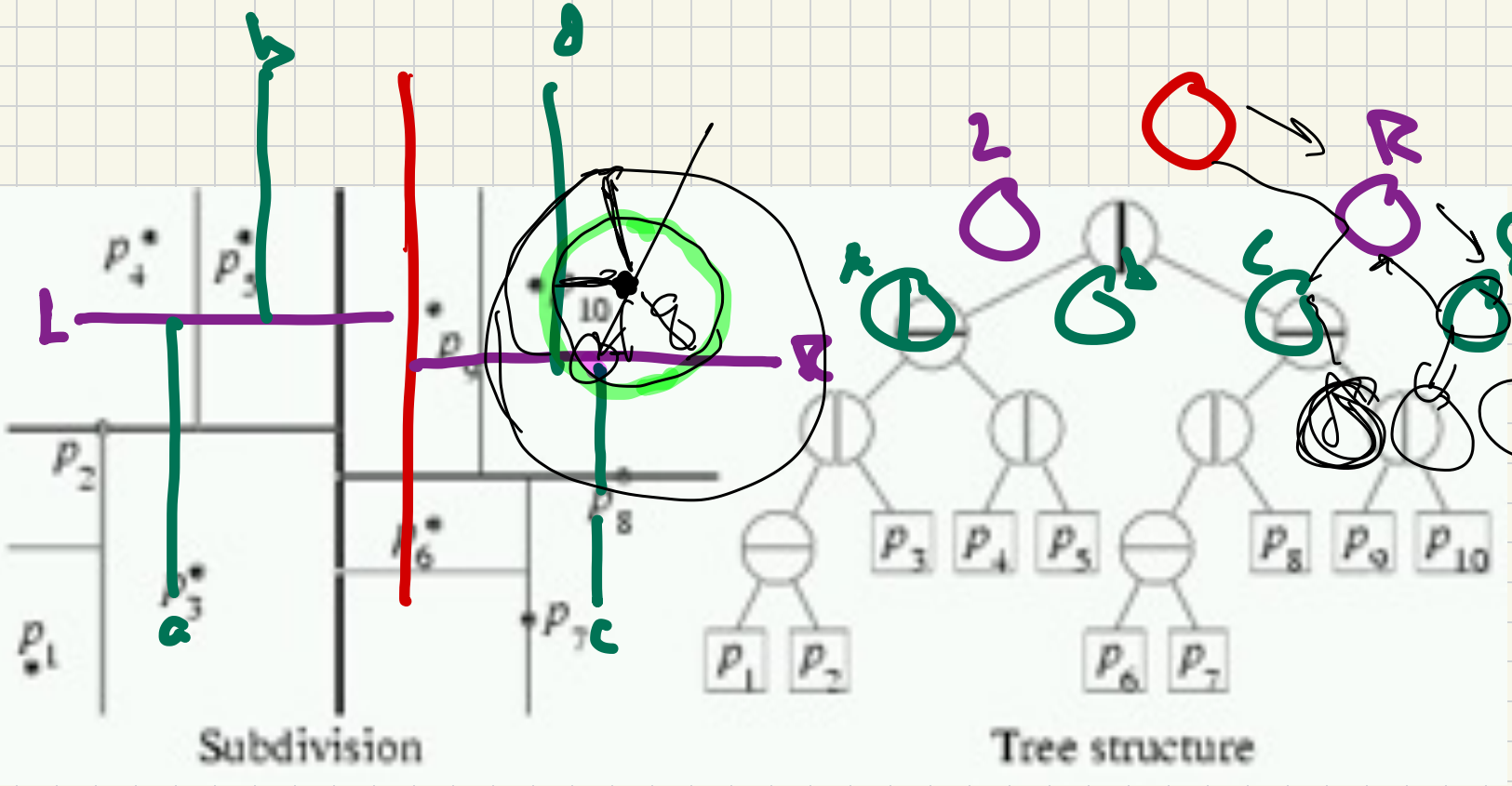
or d = 4 ... 12

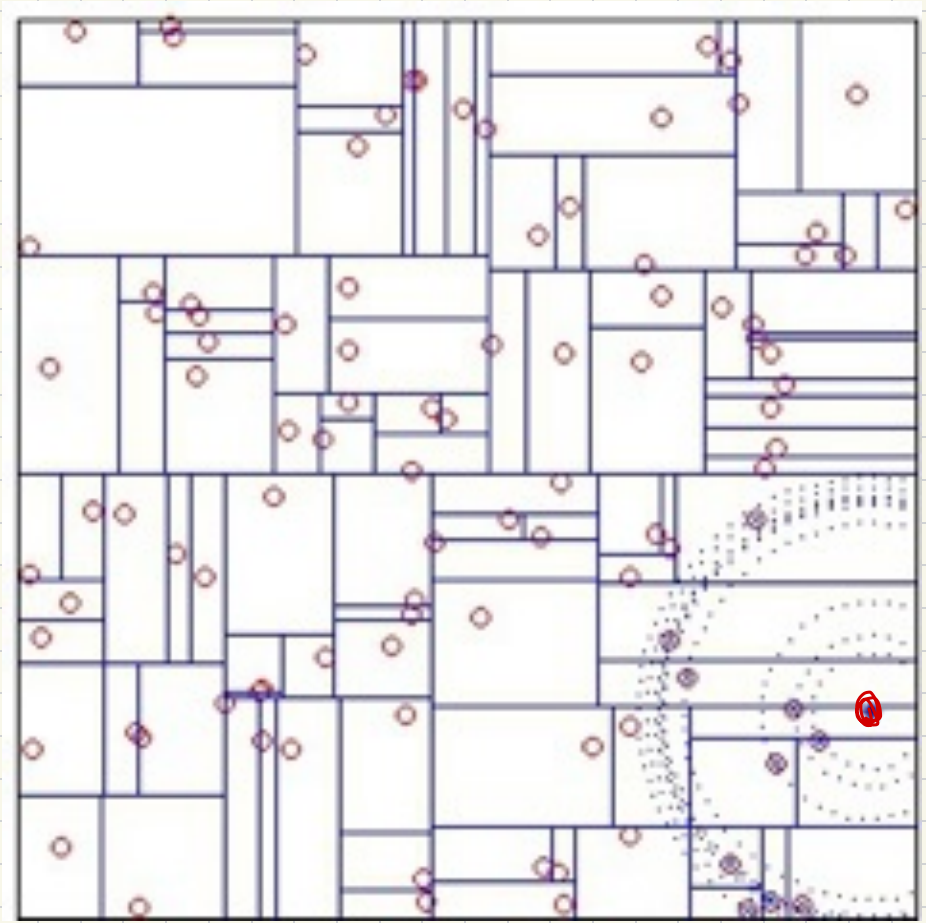
→ Heirarchy on Data extending binary kd-tree idea



kd-tree

kd-tree


L
p4
p5
p2
p3
p1
Q
p6
p7
C
p8
10
R
Subdivision
2
p1
p2
p3
p4
p5
p6
p7
p8
p9
p10
Tree structure



Approximate Nearest Neighbor

ε error tolerance $\varepsilon \in (0.05, 1)$

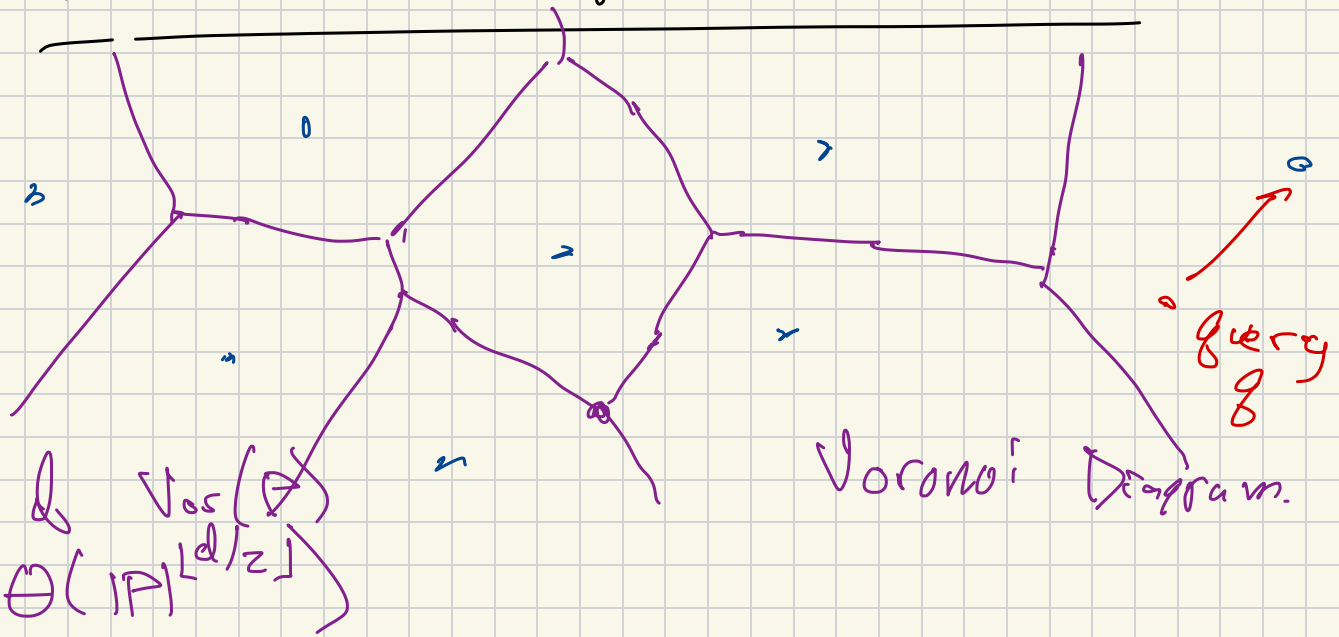
$$D(\phi_P(q), q) = V$$

or with some $p' \in P$ so

$$D(q, p') \leq (1+\varepsilon) V$$

then p' ε-approx NN

Nearest Neighbor Problem

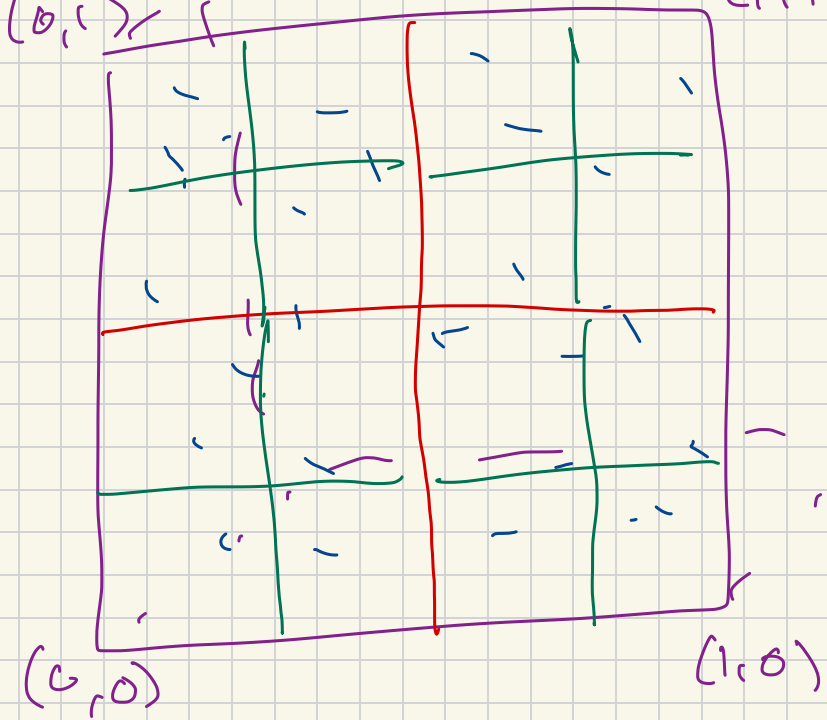


Size & Vor(P)
$\Theta(|P|^{\lfloor d/2 \rfloor})$

Voronoi Diagram.

Quad Tree

Assumes $P \subset [0,1]^d$

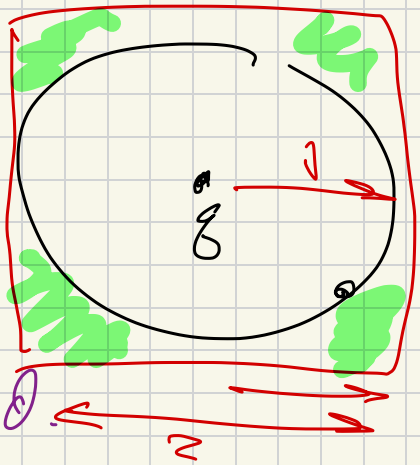


Library

ANN

Challenges in High Dimension

Using Euclidean Distances



Approximating ○ with □

Rectangle

pass in high dimension

$$\text{Vol}(\text{Ball}, \text{rad}=1) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \approx \frac{\pi^{d/2}}{(d/2)!}$$

→ shrinks as $d \to \infty$

$$\text{Vol}([0,1]^d \text{ box}) = 2^d$$

→ grows exponentially in d.

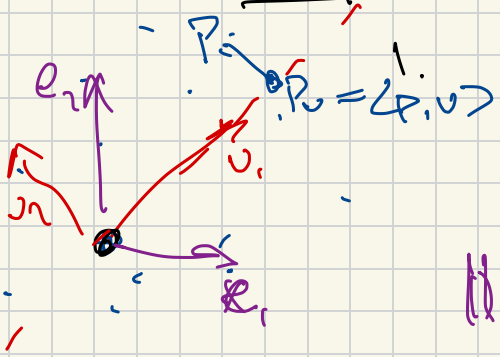
Locality Sensitive Hashing

hash function

$D_{Euc}(p,q) = \|p-q\|$

2 parts

Choose $u \sim Unif(S^{d-1})$



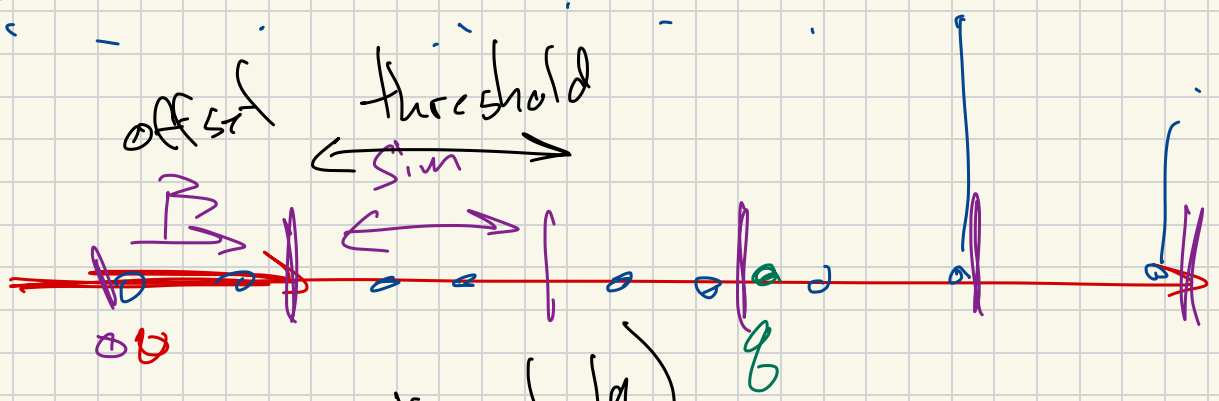
All of P project $P_u = \{\langle p_1, u\rangle, \langle p_2, u\rangle \ldots$

$$\|q-p\|^2 = \sum_{j=1}^{d} (p_j - q_j)^2$$

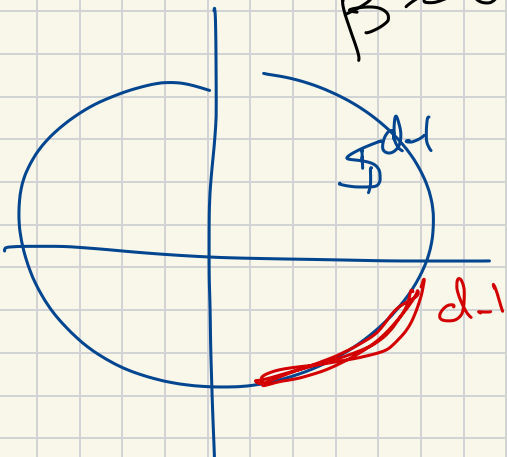
$$\|q-p\|^2 = \sum_{j=1}^{d} \langle (p-q), u_j \rangle^2$$

$$E_{u \sim \ldots}\left[\langle p-q, u\rangle^2\right] = \frac{1}{d}\|p-q\|^2$$

FALCONN



$$b \sim \mathrm{Unif}(0, \text{threshold})$$



$$\mathrm{Sim}(p, q) = \max\{0, 1 - \|p - q\|\}$$

LSH

- For NN queries, need right threshold

  Binary search on threshold w/ $O(1)$ queries

  get $O(1)$ points.

What preprocessing + query time?

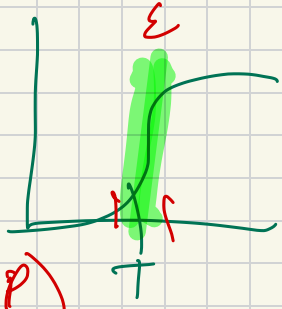
t = hash functions

b in each bin superhash

r bins/super hash query each

$\tilde{O}(n \cdot r)$

$\tilde{O}(r)$



if $\varepsilon = 1$

$r = n^{1/t}$ ~~$= O(n^p)$~~

$n = 10$ billion $\Rightarrow n^{1/t} = 10$

Graph Descent NN

Data P, build graph on P

k-NN graph

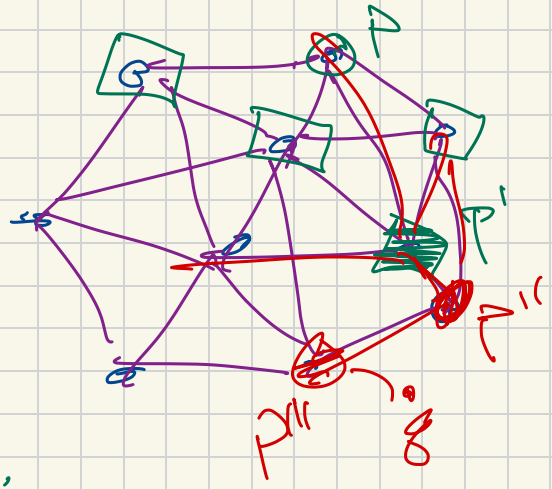
edge p_i to p_j if $p_j \in \Phi_{P,k}(p_i)$

On query q

start at any $p \in P$

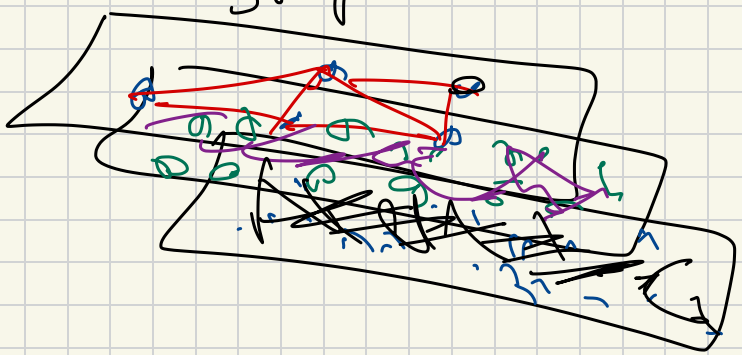
1. check if any neighbor of p is closer to q.
2. Move to closest → goto 1

   or (if not) stop & return p.



- Maintain k nearest neighbors
  always check all Neighborhoods
  until can't improve.

- HNSW Build hierarchy over graph



DiskANN

FAISS

Pinecone